Big Data in the Psychological Sciences

Wilma A. Bainbridge

Big Data in the Psychological Sciences

Cutting-edge computational tools like artificial intelligence, data scraping, and online experiments are leading to new discoveries about the human mind. However, these new methods can be intimidating to many students. This textbook demonstrates how Big Data is transforming the field of psychology, in an approachable and engaging way that is geared toward undergraduate students without any computational training. Each chapter covers a hot topic, such as social networks, smart devices, mobile apps, and computational linguistics. Students are introduced to the types of Big Data one can collect, the methods for analyzing such data, and the psychological theories we can address. Each chapter also includes discussion of real-world applications and ethical issues. Supplementary resources include an instructor manual with assignment questions and sample answers, figures and tables, and varied resources for students such as interactive class exercises, experiment demos, articles, and tools.

Wilma A. Bainbridge is an associate professor in the Department of Psychology at the University of Chicago. She has won the Association for Psychological Sciences Rising Stars Award (2023), an Alfred P. Sloan Fellowship in Neuroscience (2024), and the American Psychological Association's Distinguished Scientific Award for Early Career Contributions to Psychology (2025). Her research has garnered attention from outlets such as CNN, *Vox*, and *Wired*. She has previously edited two books on vision and memory, and her "Big Data in Psychology" class has earned a Curricular Innovation Award from the University of Chicago.

"From social media to sensors to AI, this book offers a brilliant tour of how the Big Data revolution is reshaping psychology. Accessible, inspiring, and grounded in real research problems, it walks students through everything from hands-on skills like web scraping, to big-picture theory testing, and even thoughtful discussions of ethics – all presented with incredible clarity by one of the field's most inspiring new voices."

Timothy Brady, University of California San Diego

"Exceptionally timely and comprehensive, Bainbridge's textbook deserves a place in every curriculum for behavioral methods. The chapters – enhanced with interactive features and thought-provoking ethical questions – are so engaging that they make me want to teach the course. And whether or not you work with Big Data, this is essential reading for all."

Marvin M. Chun, Yale University

"Combining conceptual depth and accessible writing, Bainbridge offers a timely contribution with a comprehensive overview of the field, covering definitions of big data in psychology and expertly navigating its key sources, methods, and analytical approaches. It addresses both foundational topics, such as neuroimaging tools and statistical techniques, as well as emerging and contemporary discussions, including natural language processing, the development of large language models, and their applications in psychological research. It will resonate with a wide audience, from curious undergraduates to seasoned researchers looking to deepen their understanding of big data and its potential to reshape the psychological sciences."

Nemanja Vaci, University of Sheffield

Big Data in the Psychological Sciences

Wilma A. Bainbridge
University of Chicago

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009343589

DOI: 10.1017/9781009343602

When citing this work, please include a reference to the DOI 10.1017/9781009343602

First published 2026

Cover image: FrankRamspott / DigitalVision Vectors / Getty Images.

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-34358-9 Hardback
ISBN 978-1-009-34357-2 Paperback

Additional resources for this publication at www.cambridge.org/bainbridge

Big Data in the Psychological Sciences

Cutting-edge computational tools like artificial intelligence, data scraping, and online experiments are leading to new discoveries about the human mind. However, these new methods can be intimidating to many students. This textbook demonstrates how Big Data is transforming the field of psychology, in an approachable and engaging way that is geared toward undergraduate students without any computational training. Each chapter covers a hot topic, such as social networks, smart devices, mobile apps, and computational linguistics. Students are introduced to the types of Big Data one can collect, the methods for analyzing such data, and the psychological theories we can address. Each chapter also includes discussion of real-world applications and ethical issues. Supplementary resources include an instructor manual with assignment questions and sample answers, figures and tables, and varied resources for students such as interactive class exercises, experiment demos, articles, and tools.

Wilma A. Bainbridge is an associate professor in the Department of Psychology at the University of Chicago. She has won the Association for Psychological Sciences Rising Stars Award (2023), an Alfred P. Sloan Fellowship in Neuroscience (2024), and the American Psychological Association's Distinguished Scientific Award for Early Career Contributions to Psychology (2025). Her research has garnered attention from outlets such as CNN, *Vox*, and *Wired*. She has previously edited two books on vision and memory, and her "Big Data in Psychology" class has earned a Curricular Innovation Award from the University of Chicago.

"From social media to sensors to AI, this book offers a brilliant tour of how the Big Data revolution is reshaping psychology. Accessible, inspiring, and grounded in real research problems, it walks students through everything from hands-on skills like web scraping, to big-picture theory testing, and even thoughtful discussions of ethics – all presented with incredible clarity by one of the field's most inspiring new voices."

Timothy Brady, University of California San Diego

"Exceptionally timely and comprehensive, Bainbridge's textbook deserves a place in every curriculum for behavioral methods. The chapters – enhanced with interactive features and thought-provoking ethical questions – are so engaging that they make me want to teach the course. And whether or not you work with Big Data, this is essential reading for all."

Marvin M. Chun, Yale University

"Combining conceptual depth and accessible writing, Bainbridge offers a timely contribution with a comprehensive overview of the field, covering definitions of big data in psychology and expertly navigating its key sources, methods, and analytical approaches. It addresses both foundational topics, such as neuroimaging tools and statistical techniques, as well as emerging and contemporary discussions, including natural language processing, the development of large language models, and their applications in psychological research. It will resonate with a wide audience, from curious undergraduates to seasoned researchers looking to deepen their understanding of big data and its potential to reshape the psychological sciences."

Nemanja Vaci, University of Sheffield

Big Data in the Psychological Sciences

Wilma A. Bainbridge
University of Chicago

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009343589

DOI: 10.1017/9781009343602

When citing this work, please include a reference to the DOI 10.1017/9781009343602

First published 2026

Cover image: FrankRamspott / DigitalVision Vectors / Getty Images.

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-34358-9 Hardback
ISBN 978-1-009-34357-2 Paperback

Additional resources for this publication at www.cambridge.org/bainbridge

Thank you to the "RAV4" – Robert, Ally, and Vicky – for being a loving and supportive family! It's hard to believe I began this book when still pregnant with Ally and Vicky and am finishing it as they are running around and chatting our ears off.

Thank you to my mom Erika, dad William, and sister Connie – finally I get to write book dedications for you rather than the other way around!

And thank you to the wonderful Brain Bridge Lab and my department at the University of Chicago – your support has really helped me flourish and think Big.

Brief Contents

Detailed Contents

Preface

Learn how to see. Realize that everything connects to everything else.

Leonardo da Vinci (1452–1519)

We live in a world where we are all constantly generating data – in our interactions with our phones, social media apps, games, websites, fitness trackers, and more. This data is commonly referred to as "Big Data" because its scale is so large that it cannot be analyzed manually. Such Big Data serves as a useful means to understand human cognition – showing us how people see, feel, respond, remember, interact, and make decisions with these different tools. We can also look at these cognitive processes across different groups of people – across countries, cultures, ages, and experiences – as well as across species. As a result, Big Data ways of thinking and analysis have become incredibly important tools to psychologists, across fields. Psychologists are now running online experiments that can gather data from thousands of participants, running machine learning models that can decode patterns from thousands of datapoints, or analyzing brain data from thousands of subregions.

As a result, psychology as a field is at a major transition point. Familiarity with advanced statistical analyses and computer programming is becoming increasingly essential to keep up with the state of the art. However, the idea of wrangling Big Data can be incredibly daunting to people entering the field, especially given that most undergraduate psychology curricula do not require computational or advanced statistical coursework. The main goal of this textbook is to make these new directions in Big Data accessible and meaningful to any psychology student – without the need of training in computer science or statistics. By reading this textbook, you'll gain basic fluency and familiarity with the important topics in the field, so you can decide what topics you want to pursue more deeply. Students who are already familiar with computational methods will learn ways in which these methods can be applied to answer a myriad of psychological questions. As a result, the book will lightly touch upon a wide range of topics, including experimental design, web programming, data scraping, artificial intelligence, different methods in brain imaging, computational linguistics, network science, wearables, user interface design, crowdsourcing, and representative sampling.

To my knowledge, this is the first undergraduate textbook on Big Data in psychology. It was inspired by a course I created in Spring 2020 as a new assistant professor, and I've seen these sorts of courses start to grow in the last few years. Because this is such a new topic, this textbook and course is really for almost anyone. Familiarity with psychology is helpful (e.g., how experiments are run and what are some of the key topics of inquiry), and at some points I will bring up simple statistical concepts (e.g., *p*-values), although knowledge there is not

required. Each chapter focuses on a different angle of how Big Data interfaces with psychology, and includes sections on ethical questions related to the topic and its real-world applicability. Each section also includes thought-provoking questions that can be discussed as a class and an assignment that's relatively open-ended and should engage the students in thinking deeply about that topic. The chapters can be covered in pretty much any order, but the book is generally divided into two parts: 1) how to rethink psychology experiments from a Big Data angle (Chapters 1–7), and 2) various sources of Big Data to enrich the study of psychology (Chapters 8–12).

In conjunction with the Big Data theme, I also want to make this course follow the principles I preach in terms of modernizing psychological research. As a result, this book is paired with an interactive online resource (www.cambridge.org/bainbridge) that includes videos, demonstrations, links, and additional resources that will be constantly updated. This way, you will still have access to the latest developments in the field even after the publication of this book. I also maintain a public data repository on the Open Science Framework of Big Data student projects that came out of the course that I teach at the University of Chicago (https://osf.io/hz843), and I am happy to link to such a repository from anyone else using this book.

Now go forth, and think big!

1 What Is Big Data?

Introduction

The amount of data generated every day is insane. Each day, we create approximately 403 million terabytes of data (or 403 exabytes) (Duarte, 2024). This is about how much data can be stored by 4 billion phones (those of about half the world population) – and just in one day! In that same day, about 300 billion emails are sent, 8.5 billion searches are done on Google, 1.6 billion swipes are made on Tinder, 1.4 billion hours of video are streamed, and $638 million is spent on Amazon. You as an individual are contributing to a lot of this growing collection of data. As you commute to your classes, your map app tracks your movement behavior and may take note of any specific locations you visit. Your phone or watch tracks your steps and sleep patterns. As you look up web pages on your phone, these pages track your browsing and click behavior. And as you scroll through and post on social media, these apps track how you engage with posts, through measures like viewing time and click behavior. We are constantly surrounded by and creating Big Data. This Big Data can be messy and tricky to sift through, but within it are potential insights about the human mind waiting to be discovered.

In this introductory chapter, we will establish definitions of the central themes of this book, to guide you as you read the rest of the book. First, we will talk more about how data has changed in the last few decades (Section 1.1) and then provide a definition of Big Data (Section 1.2). We will then define Psychology within the context of this book (Section 1.3). With these two definitions in hand, we will discuss how Big Data and Psychology interact (Section 1.4) and why now is the perfect time to study this interaction (Section 1.5). Finally, we will wrap up this chapter with a guide on how to use this book and its online resources (Section 1.6).

1.1 Moore's Law

Our data has gotten so *Big* thanks to the exponential growth in processing and storage power over the past handful of decades. This is reflected by **Moore's Law**, which was proposed by Intel cofounder Gordon Moore in 1965 (Moore, 1965). This law predicts that the number of transistors (one of the key components in computer chips) that can be packed into a given

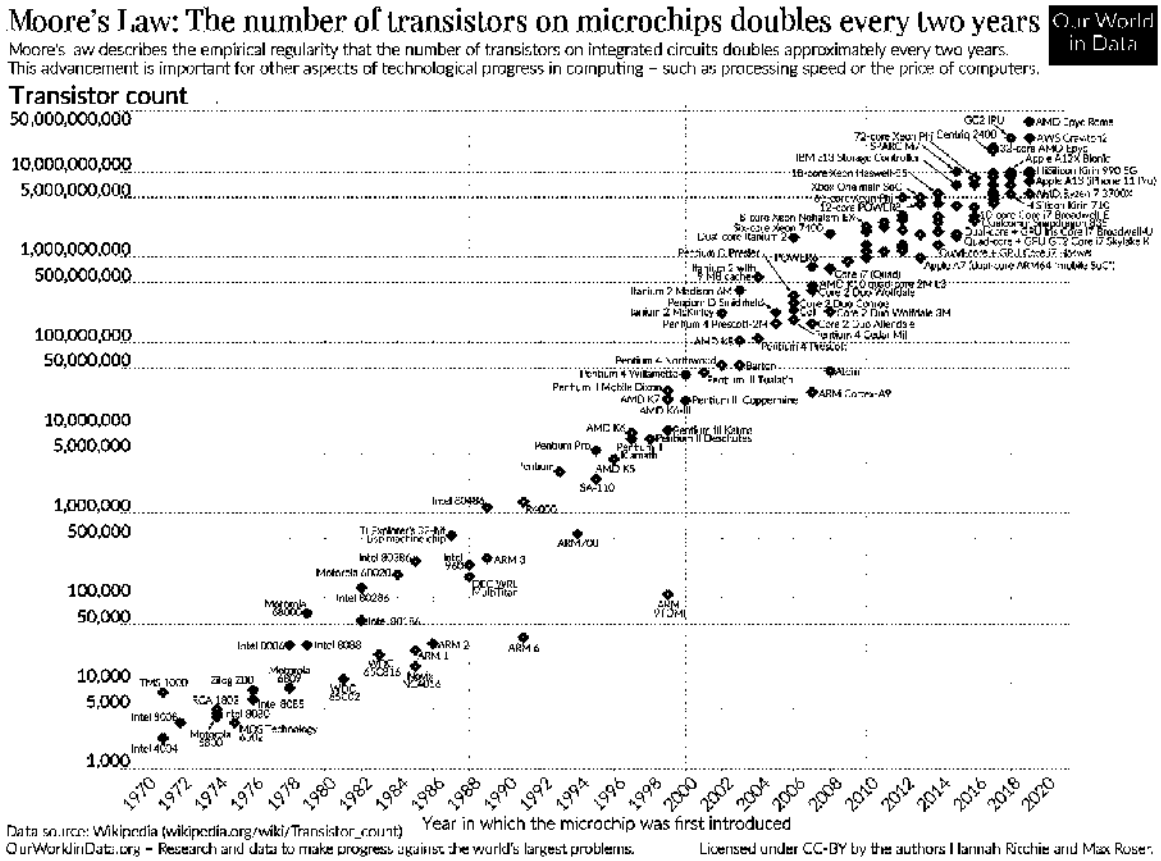


Figure 1.1 A depiction of Moore's law, showing it still holds in 2020. Moore's law posed in 1965 predicts that the number of transistors we can fit in a circuit will double every two years – resulting in exponential growth in our computing capabilities. Note that the y-axis here is an exponential scale (1,000 and 5,000 at the bottom are spaced as closely as 10 trillion and 50 trillion at the top), so indeed, we are keeping up with this law!

unit of space will double roughly every two years. Remarkably, this prediction of exponential increase in computing power has held true for 60 years (see Figure 1.1), although some scientists forecast that we will reach the limit of feasibility within the next few years (Kumar, 2015; Waldrop, 2016). We can feel the effects of Moore's law by looking at how the size of storage devices has drastically changed over our lifetimes.

Discussion Question

What did the size of data (e.g., devices, files, performance) look like when you were a child versus now? Does it feel like there has been exponential growth in that time? What sorts of innovations enabled that growth?

To understand what makes a set of data Big Data, let's first discuss how data is measured. The building blocks of the data and processes in our computers are 0s (off) and 1s (on), and a single digit is called a **bit**. Because of this 1/0 building block, instead of data being measured in our decimal base-10 system, data sizes are measured in **binary**, or a base-2 system, where the only digits possible are 0 and 1. When you want to count in higher numbers in binary beyond 1, you use additional digits (that are still limited to 0 and 1). So the numbers 0, 1, 2, 3, 4, and 5 in decimal are represented as 00, 01, 10, 11, 100, and 101 in binary. While these building blocks seem simple, they can combine to form the complex data we interact with on our computers – just as letters can combine to create the complexities of language. Because of the binary system, powers of 2 end up being important to the measurement of data. A set of eight bits ($2\times2\times2$) is called a **byte**. A byte can be used to represent a single character of text. For example, in the most common character encoding standard for computing called **ASCII (American Standard Code for Information Interchange)**, the letter A is represented by the byte containing the bits 0100 0001, while a space is represented by 0010 0000. Above the level of the byte, the naming of the counting system resembles that of the metric system. A set of 1,024 bytes is called a kilobyte (KB), like how 1,000 meters is a kilometer (but because we are operating in binary, it is a multiple of 2, or 2^{10}). A set of 1,024 kilobytes is called a megabyte (MB). A set of 1,024 megabytes is called a gigabyte (GB). After that, we have terabytes (TB), petabytes (PB), exabytes, and zettabytes. So, for example there are 8,000,000 bits (1s or 0s) in one megabyte of data. When we talk about data transfer speeds (like how fast your internet is), the measures tend to be in bits per second (instead of bytes per second). So early internet modems would have a download speed of 28.8 Kbps, or around 28,800 bits per second.

The rapid change in computing sizes is quite drastic when we look at the history of data storage across personal computing (Figure 1.2). Back in 1956, IMB shipped its first hard drive. It was the size of two refrigerators and could hold 5 MB of data – the equivalent of about one song. In the 1970s, some consumers were starting to get their own computers, and the most common way to store and transfer files was through floppy disks. These could only hold about 100 KB in early years, and 1.44 MB in later years, the equivalent of a few text documents or pictures. However, this medium became so ubiquitous that many pieces of software still use an icon of a floppy disk as their "save" icon. Once software became more advanced, users needed more and more of these disks – for example it took seven floppy disks to install an early version of Adobe Photoshop (Adobe, Inc., 2013).

In the late 1980s, a more advanced data storage method emerged – the CD-ROM (compact disc read-only memory). These could hold as much as 900 MB – about one-third of a movie. However, as their "read-only" name implies, these needed special devices called CD burners to write data to the CD-ROM, and most CDs could not be rewritten once data was saved onto it. In the mid-1990s, we moved onto DVDs (digital video disks), which could store closer to 5 GB (about 1–2 movies) but faced similar shortcomings as CD-ROMs. The early 2000s saw the first USB (universal serial bus) flash drives, which were smaller and more convenient but could only store about 10 MB at first. The early 2000s also saw the explosion of the internet, and **cloud storage** – saving data to online servers – started to grow. This really took off as internet speeds became faster, and websites emerged dedicated to hosting large amounts of data – YouTube for video started in 2005, Dropbox for files started in 2007, and Flickr for photos started in 2004.

Figure 1.2 Photos of various types of older data storage. **(a)** The shipping of IBM's first hard drive, holding 5 MB and taking up the size of two refrigerators. **(b)** A 3.5-inch floppy disk, the main external data storage format in the 1980s and 1990s. **(c)** A CD-ROM inserted into a laptop's CD drive. These were commonly used in the 1990s and 2000s for data storage and were the main medium for holding music albums.
Source: (a) Michael de Groot / Flickr. (b) Pablo Jeffs Munizaga – Fototrekking / Moment / Getty Images.
(c) EThamPhoto / The Image Bank / Getty Images.

In the 2010s, storage amounts and data transfer speeds increased by many magnitudes. Personal computers could hold close to a TB of data, and mobile technologies emerged enabling people to generate vast amounts of data through the photos and videos they're taking. In the 2020s, higher mobile data speeds and faster personal computer processors are allowing us to interact with vast amounts of data and at faster rates – supporting growths in avenues like online gaming, video streaming, and real-time artificial intelligence (AI) on board many of our systems.

1.2 How Do We Define Big Data?

Owing to this constant growth in technology, the definition of Big Data is a moving target. In the 1990s, even a low-resolution video would be considered Big Data, while in the 2020s, Big Data is more in the order of libraries of thousands of movies, representing petabytes of data (or even more). A fundamental aspect of what makes data qualify as Big Data is that it is so large that we cannot process it by hand – we can't manually input, clean, or analyze the data. This is data that is also often so large that we cannot process it with basic programs on our computer (e.g., Microsoft Excel), but usually have to use code or bespoke tools. For our purposes of studying human cognition, Big Data tends to be more naturalistic – recorded from real people or dynamic behaviors – so it may be unstructured or more noise- or error-prone compared to smaller datasets. Data can be Big for several different reasons. It may have very high temporal sampling – for example, taking a measurement every handful of milliseconds. It may have high spatial sampling instead – for example, collecting data across all the intersections in a city. It may have high participant sampling – collecting data from a large and diverse set of people. Or, it may have high stimulus sampling – capturing data from lots of sources (e.g., images, videos, news articles, products) or tasks. All of these encompass examples of Big Data.

So here, we will loosely define Big Data as: *unstructured, naturalistic human data requiring complex analytical methods*. For some of the exercises and examples we discuss in the book, we may not use the massive amounts of space or computing power that traditionally make up Big Data. But the concepts that you learn here should translate to bigger sets.

1.3 How Do We Define Psychology?

It is also important that we define what psychology is within the framework of this book. Broadly, psychology is *the empirical study of the mind, brain, and behavior*. For the vast majority of this book, we will focus on a more quantitative and research-based approach to psychology, where psychologists conduct experiments that aim to provide broad insight, using falsifiable questions and hypotheses. This is in contrast with some psychologists who use more *qualitative* approaches, like revealing new insights using interviews or observations, or using therapeutic techniques like talk therapy to help improve mental health. A central aspect of the psychology we will discuss here is that it is composed of research questions that are testable and falsifiable. **Falsifiable questions** are those where you can obtain evidence to prove that question wrong. For example, one active area of debate is the degree to which we can falsify questions in **evolutionary psychology** – the study of the mind, brain, and behavior through the perspective of evolution (Gannon, 2002). We cannot go back in time and run experiments on our ancestors. We also cannot measure how much of people's current behaviors are a result of evolution versus more recent societal norms. There are some creative scientific methods to test evolutionary hypotheses by looking at other animal species or running computational models. However, we will generally avoid tackling unfalsifiable topics in the current book.

We are also interested in questions that are **generalizable** – that give us insight into the thinking of a group of people, and can allow us to make predictions in the future about different events. For example, a question like "how did people feel about Argentina winning the 2022 World Cup?" is a question that measures emotions and behavior, but is so highly

specific that it does not really teach us about the human mind. Thus, we would not characterize this as a psychological question. A more generalizable psychological question might be something like "how does sentiment in social interactions change directly after major sporting events?" Sometimes when dealing with Big Data, we may accidentally take an overly narrow scope, due to the data that we have available (like data from a specific app, Chapter 8). But we should always try to focus on a big-picture question about the mind, and any limitations to the data we are collecting to answer this question.

There are many different branches to the field of psychology. When you think of a "psychologist," your mind may first go to **clinical or abnormal psychology** – the study of atypical behavior and mental health, with the goals of diagnosis and treatment. Closely related is counseling psychology, which is the practice of helping people through therapy and counseling. While this book will discuss some research on abnormal psychology through the lens of understanding the underlying roots of an impairment, we will not discuss in depth therapies or treatments of individuals. **Industrial psychology** is the study of the mind within the workplace, and how to optimize people's effectiveness at work. As this field is more applications-focused, we will not discuss this at length in this book, nor other more applied fields like forensic psychology, school psychology, and health psychology. The major focus of this book will be **cognitive psychology**, the branch of psychology dedicated to the scientific study of our internal mental processes. This encompasses a broad range of processes, including sensation, perception, action, memory, reading, speaking, emotion, decision-making, morality, imagination, and others. In fact, a "psychologist" can be a researcher with a laboratory that runs experiments to study these processes (this is the type of psychologist I am). Related to cognitive psychology, we will sometimes discuss the brain, through a lens of **neuropsychology** – looking at how the brain and mind interact. We will also bring up many examples from **developmental psychology** – the study of the development of the mind across the lifespan (from infants through aging) – and **social psychology** – the study of the interactions of multiple minds.

1.4 How Do Big Data and Psychology Interact?

A large proportion of Big Data out in the world just *happens* – you record a video and post it on social media, and now there are several new megabytes of data on your phone, on a server belonging to that social media site, and being downloaded to other people's phones. In this way, much of Big Data is just passively accumulated as we perform tasks with our phones, computers, and the internet. Another major slice of Big Data is being actively collected by companies, where they are testing how they can improve your experience, how you navigate their app, and how they can improve engagement and purchasing. However, the data being generated out in the world also serves as incredibly rich records of human behavior that can give insight into questions on almost any topic of psychology.

Discussion Question

What are some ways you can envision Big Data might be changing the types of questions we can ask or answer in psychology?

An important skill for you to nurture will be in identifying these intersections of Big Data and psychology. What is a psychological question you want to answer, and how could Big Data answer that question? Could there be a preexisting dataset out there that answers the question for you, or could Big Data help you collect that data in some way? For example, a few years ago, I was curious how older memories (2+ year-old memories) might be represented in the brain. This is hard to test in the laboratory because I would need to have participants study some images and then come back two years later. But then it dawned on me that people are constantly capturing their memories on social media, dating back to many years prior. So, I collaborated with the app 1 Second Everyday to recruit users who had recorded years of their memories, and then I scanned their brains while they viewed these older memories. Long story short, we found patterns in the brain reflective of the age of a memory (Bainbridge & Baker, 2022; see Section 8.4.2). As you look through data in your daily life, think to yourself – what does this reflect about the human mind and can it show us something new? And, are there ways in which Big Data technologies are influencing how we think or interact? For example, an active area of current research is how social media may be impacting feelings of isolation and depression (Section 12.3.2). Overall, a major part of this class will be thinking creatively and with an open mind on how we can use data to answer questions.

1.5 Why Study This Now?

Computation and psychology are both at points of incredible transition right now. On the technological side, we are generating more data than ever, but tools to process this data are also starting to become more accessible to the average person. There are notably five main changes that have occurred with computing technologies that have enabled data to become so big. First, as we have discussed, there have been drastic improvements in cheap, large data storage in small form factors. This means that the average person has on their phone or computer tens of thousands of files, documents, images, videos, and pieces of software. This also means there are places where we can easily save our big datasets. Because these storage devices are getting smaller, we can have large amounts of storage in small devices, like phones. Meanwhile, cloud storage allows people to maintain massive amounts of data that they can access with a multitude of devices. Second, there have been major improvements in faster and cheaper processing power, such as the explosion in parallel processing graphics chips. For example, the average processor in a consumer computer can make about 150 billion calculations per second. This allows us to analyze big datasets relatively rapidly, and even in real time as we acquire it. Third, sensor technology and speed has also improved – most people have high-quality cameras in their phones, and may have devices (like fitness trackers) that can record movement, heart rate, elevation, skin conductance, and other measures. This allows us to obtain big physiological data, which can reveal underlying information about one's cognitive state (Chapter 9). Fourth, the wide spread of high-speed internet both in homes and out in the world is allowing more people to form communities, creating large amounts of data generated by people's interactions online. Fifth, our algorithms are getting better and smarter – we are able to compress data more efficiently and analyze data more effectively with tools like artificial

intelligence. The combination of these five computational improvements has led to an explosion of data produced by and accessible to the average person.

Big Data is also more important than ever for psychology research. Psychology has always been a multidisciplinary field, straddling social science and biological science programs at many universities. For example, psychology has clear links to neuroscience and experimentation, but also has implications for therapeutic practice and philosophy. However, recently, psychology as a field has begun to undergo a transition, with greater emphasis focused on experiments and complex analyses. Many exciting discoveries are coming about thanks to Big Data innovations, such as online experiments, artificial intelligence, and rich physiological data. With these innovations, researchers have been able to revisit classical psychological questions with a Big Data lens that allows them to assess their applicability across more diverse samples or make computational models that can predict people's behaviors. These innovations have also been saving psychologists a lot of time – making it faster to collect and analyze data. These changes go hand in hand with a new global scientific community that is developing, based around sharing data and code openly, in reaction to a "replication crisis" that emerged around unreplicable findings in small-scale experiments (see Section 2.1.3). So now is the perfect time to learn about these changes in psychology, to ride its waves as it moves into these new approaches.

Discussion Question

What topics relating to Big Data and psychology are you particularly excited to learn about in this book, and in your class?

1.6 How to Use This Book

As we just discussed, psychology is changing. If you want to go into psychological research for your career, professors and laboratories are now increasingly looking for candidates with experience in programming and statistics. Outside of academia, many jobs after college geared toward psychology majors – such as user experience design or data science jobs – also require these skills. For those of you wanting to practice clinical psychology, counseling, or go into education, it is still helpful to be up-to-date with the latest research and techniques (e.g., how is artificial intelligence changing the diagnosis of neuropsychological disorders?). And I would argue that some of these topics we will discuss in this book can help improve your daily life. I know for me personally, I've coded data scraping tools to find the best flights for a vacation, used generative AI to make a personalized storybook for my kids, or analyzed my fitness tracker data to get a sense of whether a diet is working. Knowing what is possible with data can change how you look at and use data in your daily life. In this book, we will also touch on some of the hot-button topics that have erupted in the news and the legal sphere as a result of Big Data – how do we deal with AI-generated fake information? How do we navigate the privacy risks created by the data recorded in many mobile apps and websites?

It can be intimidating jumping into learning about data and computer programming if this is your first foray into the topic. My number one goal is to demystify these topics and make you comfortable talking about them and thinking about them. As a student starting along this journey, it can feel like there's a big gap between you and your image of a computer scientist who may have been hacking computers since they were in elementary school. It can feel like you just aren't meant to be someone who codes or does complex math. But really these thoughts are a part of a mystical (but inaccurate!) aura that has surrounded computation. I'd liken computer programming to something like learning a foreign language or training for your first 5 km run. Most of the time, your goal isn't to become completely fluent or a record-breaking marathon runner. Usually, it's that you find these skills useful and enjoy the process of getting there. You also usually aren't worried about how you compare to the pros – you don't feel bad comparing your Spanish skills to those of a native speaker (and often they are impressed that you are trying!), or feel bad watching Olympics runners beat your time. In the same way, a seasoned software developer won't be quizzing you on the latest Python functions. You will also find that gaining these skills can enrich your daily life – you can now navigate a little around Madrid with your newfound Spanish skills, or be able to run to catch a bus without getting winded. Similarly here, you'll have moments where you may wish you could do something on your computer in an automated way and then realize there may be a way to use your skills learned here to do that!

At the same time, this book is not going to teach you programming or statistics from the ground up. It's the first step in learning the lingo and giving you the lay of the land, so you can then decide where you want to do a deep dive in future classes or explorations (e.g., do I want to study more neuroscience? Or web design? Or graph theory?). The online resources with this book will provide some stepping stones for doing these deep dives. With that foreign language metaphor, this book is your travel guidebook to help you decide where you want to study abroad. Then once you've picked a country, you can start focusing on learning its language. With this book, I want you to become fluent in the topics of new technologies being widely used in psychological research. I want you to have an increased level of agency over your own data and how it is used by companies and researchers. And, I want you to practice thinking creatively about psychological research questions and how we can answer them.

This book can be read from front to back or you can skip around sections as needed. In these first two chapters, I introduce what Big Data (Chapter 1) and small data (Chapter 2) are and how they compare to each other. Then for the rest of the first half of the book, I will give you the building blocks for running Big Data studies – looking at the participants (Chapter 3), the stimuli (Chapter 4), and the experiments themselves (Chapter 5). Once we are armed with our Big Data, we can then analyze it using artificial intelligence (Chapter 6) or human crowdsourcing (Chapter 7). In the latter half of the book, we will delve into different topics that are changing as a result of Big Data, and so these chapters are a bit more standalone. We will talk about software developments with apps and games (Chapter 8), as well as hardware innovations and physiological sensing (Chapter 9). We will talk about Big Data in neuroscience (Chapter 10), language and natural language processing (Chapter 11), and wrap up with social interactions and graph theory (Chapter 12).

With the exception of this chapter, each chapter will end with four key sections. In "Big Ethical Questions," we will talk about the ethical implications of the topic discussed in the chapter. These topics are sure to spark interesting discussion, especially because the ethical implications of these new methods are still being actively addressed in science and society. This section is then followed by a section on "Applications of the Chapter." While most of this book takes the framework of theory-driven psychology – where we are conducting experiments for the sake of understanding the mind, not creating a product – in these sections, we will discuss how the chapter's topics can be applied to impact the real world. Each chapter then ends with a Chapter Summary that reminds the reader of the major points, and Further Reading which suggests further sources to explore if you are interested in going beyond the pages of this book. There will be discussion questions laced throughout the chapters, as well as a sample homework assignment at the end of each chapter. The companion Teacher's Guide will include additional discussion questions, exercises, and demonstrations for each topic.

Importantly, data, computation, and the internet are always changing. While this book is written at a static point in time (2022–2025!), there is an accompanying online resource (www.cambridge.org/bainbridge) that will be updated as technologies change in the world. If you read anything in this book that seems outdated, check out the online resource to see if there is a new version of that information. The online resource also has interactive demos and programming tools to let you learn more about programming and test out online experiments.

With that, let's proceed to Chapter 2 to discuss what "small data" is and how that differs from Big Data in the context of psychological research.

CHAPTER SUMMARY

In this chapter, we introduced the concepts of Big Data, psychology, and how now is the perfect time to study them and their interactions.

1. Here, we define Big Data as unstructured, naturalistic human data requiring complex analytical methods.
2. We define psychology as the empirical study of the mind, brain, and behavior. This book mainly focuses on quantitative experiment-based psychology.
3. With major improvements in our technological capabilities over the past few decades and changes in the landscape of psychology, now is the perfect time to study how Big Data can be used in psychology research.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter.

- **Read about and watch an original video describing the world's first hard drive, developed by IBM in 1956:** Seeley, C. (2014, October 28). History snapshot: 1956 – the world's first moving head hard disk drive. Data Clinic Ltd. News. www.dataclinic.co.uk/history-snapshot-1956-the-worlds-first-moving-head-hard-disk-drive

- **A review of how realism in our studies can actually show differences in the brain:** Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*, 506–519.

ASSIGNMENT

The purpose of this assignment is to learn more about your experience with Big Data and provide you a sense of the data you generate.

Total Points: 50

1. **Fill out the class survey.** Your professor will provide a link. (20 points)
   Let's look at how much data you are generating just from your phone!
2. Locate where your phone describes your storage usage. **Answer:**
   a. How much storage are you using for images? (1 point)
   b. How much storage are you using for videos? (1 point)
   c. How much storage are you using for music? (1 point)
   d. How much storage are you using for apps/applications? (1 point)
3. Let's get a rough estimate of how much data you are generating a day with your phone camera.
   a. Add together your answers from 2a and 2b and **report that number here**. (2 points)
   b. Get an estimate of how long you have had your phone – find the date of the first photo you took. Then search on Google "how many days between [that date] and today" and it should return you the number of days. **Report that date of the first photo and the number of days since then.** (2 points)
   c. Divide your answer in 3a by your answer in 3b and **report that number here.** This tells you about how much data you are generating with your camera per day. (4 points)
   d. **How many bytes of data is that?** (2 points)
   e. One byte is the amount of data used to type one character (e.g., "A"). A novel contains about 500,000 characters. **How many books worth of data is that?** You're likely creating the equivalent of books of information a day! (3 points)
4. Let's see how long you are using your phone for.
   a. First, guess: **How much screen time do you think you use a day?** (2 points)
   Now, locate where your phone describes your screen time usage (this might be under a "Screen Time" setting or a "Digital Wellbeing" setting).
   b. **On average, how much screen time do you actually use a day? How does this compare to your guess?** (4 points)
   c. Based on your screen time report, **on average how much screen time do you spend on social media a week?** (2 points)
   d. You use on average 500 MB of data per hour by browsing social media. **How many GB (or MB) of social media data are you viewing per week?** (5 points)

As you can see, we interact with massive amounts of data in our daily lives!

REFERENCES

Adobe, Inc. (2013, August 1). Did you know that Photoshop 3.0 was the last version of Adobe Photoshop to be sold on the floppy disc? Facebook. www.facebook.com/photo?fbid = 10151614431968871&set = a.468676338870

Bainbridge, W. A., & Baker, C. I. (2022). Multidimensional memory topography in the medial parietal cortex identified from neuroimaging of thousands of daily memory videos. *Nature Communications*, *13*(1), 6508.

Duarte, F. (2024). Amount of data created daily. Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day

Gannon, L. (2002). A critique of evolutionary psychology. *Psychology, Evolution & Gender*, *4*, 173–218.

Kumar, S. (2015). *Fundamental limits to Moore's law*. arXiv:1511.05956.

Moore, G. E. (1965). Cramming more components into integrated circuits. *Electronics*, *38*(8).

Waldrop, M. M. (2016, February 9). The chips are down for Moore's law. *Nature* [news feature]. www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338

2 What Is Small Data?

Introduction

In order to learn about Big Data, you first need to understand its counterpoint, "small data." Small data isn't often called this, because data from most psychology studies fits under this umbrella, and many times its scale can suit our purposes just fine. Thus, a definition of small data would be any data that isn't Big Data. While it is incredibly common, solely using small data severely limits the takeaways we can get from psychological research. In this chapter, I will discuss the limitations of small data, as well as the limitations of Big Data. You will see how the two can work in synthesis to pinpoint the rich phenomena occurring in our minds and brains.

Specifically, first to understand the benefits we gain from Big Data, we will go through a few example small data experiments (Section 2.1.1) and see what they are lacking (Section 2.1.2 and Section 2.1.3) and how they can be made bigger in scale (Section 2.1.4). We will then discuss some limitations to Big Data experiments (Section 2.2), including new issues they introduce (Section 2.2.1) and the limitations that will always be present with any study (Section 2.2.2). We will then discuss how experiments can be dichotomized into being hypothesis-driven or data-driven (Section 2.3), as well as how Big Data studies can be characterized as deep or wide (Section 2.4). We will discuss the ethical issues that can come about from the multiple analyses run with Big Data (Section 2.5). We will then discuss applications of the topics in the chapter, such as examples of famous data-driven discoveries (Section 2.6.1) and medical applications of deep and wide data (Section 2.6.2).

2.1 Turning a Small Data Experiment into a Big Data Experiment

Let us first begin with an example of a small data experiment and think about how we can make it bigger and broader.

2.1.1 A Case Study

There is a famous effect in psychology called the **own-age effect** (Anastasi & Rhodes, 2005), where people tend to remember faces close to themselves in age better than faces farther in age.

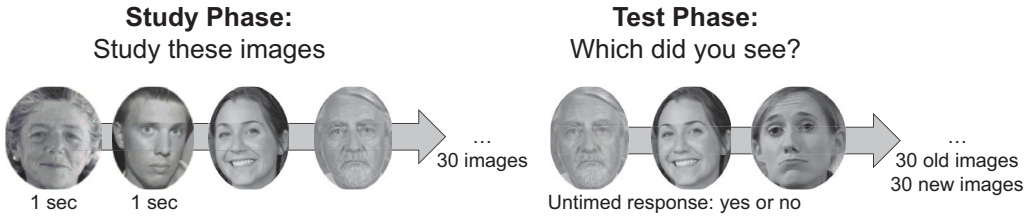


Figure 2.1 The experimental methods for our own-age effect experiment. We have participants first study thirty face images for 1 second at a time. Half of the face images are from older adults and half are from younger adults. We then test them where we show them thirty of the face images they saw, randomly mixed up with thirty new face images they didn't see. For each face they have to respond if they saw it before ("yes") or not ("no"). Our main research question is whether participants have different levels of memory accuracy based on the match between their own age and the age of the face images.

You may have experienced this before, where you may have an easier time recognizing your classmates than professors on campus. (There are many memory effects driven by the similarity of a face to your own – there is also famously an own-race effect; Chiroro & Valentine, 1995). Let's say we are in a traditional psychology lab, and we are running an experiment to test the own-age effect. Our experimental methods look something like what is in Figure 2.1.

The idea is to recruit fellow psychology students on campus and run them through a face memory test on the computer in the lab (as most psychology experiments are done!). In that memory test, we will show a series of thirty faces (like in Figure 2.1), where half are college-aged, while the other half are older adults. We will then test to see if there is a significant difference in memory for those two groups of faces. After running twenty participants, we find a significant effect – indeed the own-age effect holds true!

Discussion Question

What prevents us from generalizing these results to saying that the own-age effect occurs for all observers and all faces?

2.1.2 What Does a Small Data Experiment Miss?

The previous case study was an example of a typical psychology experiment. However, there are many aspects of it that prevent us from generalizing to all observers and all faces. Specifically, the participants, the stimuli (the face images), and the experiment itself are all constrained and artificial in some way.

Small Participants: First, the number and scale of the participants is "small" – can we really make generalizations about humans as a whole from an experiment run with twenty students at a specific university? For example, would these effects replicate for people who are frequently exposed to faces of other ages – like in cultures where young adults tend to live with older

generations? In Chapter 3, we discuss more about the problems with using small college samples in a large proportion of psychology studies, and what we can do about it in the field.

Small Stimuli: Second, the images are also very small in scale. Like the issue we have with participants, can thirty faces really capture the rich variance of human faces out in the world? If you look at the paradigm (Figure 2.1), all these faces are very homogenous. They are all front-facing white people of moderate attractiveness with an oval cropped around their face so you cannot see much of their hair or clothing. This can sometimes be intentional – researchers often want to control for factors they're not interested in, so that those cannot be alternate explanations of their effect. For example, you don't want to think there is an effect of age on memory when it's actually the clothes the models are wearing (maybe clothes from a few decades ago are more memorable than clothes from today!). But, too much control will limit our ability to make generalizations across different lighting, viewpoints, and facial expressions – it doesn't let us make confident predictions about memory out in the real world. And, by only doing research on constrained demographics (e.g., all white people), we aren't studying the rich variation in human experience. In order to generalize to the real world, we need images that better capture the diversity we observe in that world. In Chapter 4, we talk more about how to think about and create more representative stimulus sets.

Small Experiment: Even in just the way they are conducted, experiments are much smaller in scale than the real world. They don't capture what it's like to meet a moving, emotive, multisensory human being, and try to encode them into memory. Experiments tend to be brief (usually 30 minutes to an hour) and constrained to a two-dimensional computer screen, with a few seconds to see each face. This is not at all what it's like to meet a face in reality – you see them out situated in the real world, and you may spend hours interacting with them. Perhaps the dynamic, moving aspects of a face can contribute to your memory for that face, and that would be completely ignored by the experiment. Or perhaps seeing faces in the three-dimensional world is fundamentally different for memory than seeing them on a flat, two-dimensional screen in an experiment. (Although seeing faces in two dimensions may be becoming more natural, as virtual meetings are becoming more common.) Also, because faces are so dynamic, it's unlikely in the real world that you will ever see the exact same view of a face again; you can never take the exact same photograph twice. The second time you see a person, their facial muscles will be engaged in a slightly different way, the lighting will hit their face differently, or they may have a slightly different glow to them. This is completely different from an experiment which shows you the exact same photograph twice.

2.1.3 A Second Case Study and the Replication Crisis

Let's examine another sample experiment. Within the field of social psychology, one phenomenon that has been proposed is the phenomenon of **social priming**. The idea with social priming is that when you are made to think of a social category, you automatically think about related behaviors and stereotypes and start to subtly behave in a similar way. This was first demonstrated in a study by Bargh and colleagues (1996) across a series of experiments. For example, in one experiment, thirty psychology class undergraduates from New York University were asked to complete a task where they had to take a set of five words and create a grammatically

correct four-word sentence as quickly as possible. They did this for thirty sentences in total. Unbeknownst to those participants, half of them received words specifically related to being elderly – *old*, *grey*, *sentimental*, *bingo*, *wrinkle* – while the other half received neutral words. The idea was that the elderly-related words might prime them to think about elderly individuals and act in a similar way. An experimenter then secretly timed how long it took participants to exit the hallway leaving the testing room. The researchers found that participants primed to think about being elderly had a significantly slower walking speed (8.3 seconds to travel the hallway) than participants given a neutral prime (7.3 seconds), confirming their hypothesis. Participants reported not being aware of this elderly manipulation, or a change in their behavior, suggesting these social priming effects could happen unconsciously.

Discussion Question

What factors in this experiment might prevent us from generalizing more broadly?

This experiment uses a relatively small number of participants (fifteen in the elderly prime condition) and stimuli (thirty), making the robustness of the effect unclear (though the original experimenters do actually replicate this effect in a second thirty-participant experiment). The participants come from a very specific sample – psychology undergraduates in the New York area – who are unrepresentative of the world population. The words are also not validated as conjuring an image of "the elderly" in an objective way. The study does a fairly good job at using a naturalistic task (i.e., measuring walking time). However, there could be modern improvements on how it is measured, rather than relying on an experimenter's timing skills, which could introduce a subtle bias that accounts for the 1-second difference between conditions. As a result of these critiques and others, Doyen and colleagues (2012) ran a larger-scale replication of the experiment. They ran 120 participants (albeit also from a fairly specific sample – Belgian French-speaking undergraduate students). They used elderly word stimuli that were first confirmed by a separate set of eighty participants as representing old age. The experimenters then used infrared sensors to precisely measure the amount of time it took to traverse the hallway. With this "bigger" experiment, researchers found no difference in walking speed between their two participant conditions.

Around the same time (in the early 2010s), many psychological findings were unsuccessfully replicated. Researchers were failing to find clear evidence for many social psychological phenomena that had become well-accepted – in addition to social priming, there was now evidence against ideas like ego depletion (the idea that willpower is a finite resource) and power posing (that standing in a certain way will increase your confidence) among others. This launched a "**replication crisis**" across the field of psychology bringing into question the quality of the research in the field. One event that ignited this crisis was when a paper was published in one of the most revered social psychology journals (*Journal of Personality and Social Psychology*) claiming evidence that people can see the future (use "precognition"; Bem, 2011). Researchers realized that a combination of poor research practices as well as publication pressures in the field (see Section 4.4) was overinflating the reporting of supposed "results" across many papers.

At this major breaking point, hundreds of researchers as part of the Open Science Framework launched an effort to attempt to reproduce a hundred findings in psychology. Shockingly, only thirty-six were successfully replicated (Open Science Collaboration, 2015). This served as a reality check for psychologists – we need to run experiments with larger samples, more generalizable experiments, and better statistical measures. We also often should run multiple replication experiments to confirm our effects really hold, and aren't just occurring due to chance.

2.1.4 Making an Experiment Big

Even if I have convinced you that traditional psychology experiments are often unnatural simulations of the real world, how can we improve upon this? How can we make our experiments "bigger"?

> **Discussion Question**
>
> What would a Big Data version of the example face experiment look like? What would you change?

We need to think about how we can improve upon the three points mentioned earlier: the participants, the stimuli, and the experiment. For the participants, can we recruit more people, and more widely? In Chapter 5, we will talk about how to conduct online experiments, which lets you reach thousands of people, with more diversity than the average college campus. For the stimuli, we can also strive to collect image sets that are larger, more natural, and more diverse (refer to Chapter 4 to learn how we can do that!). For making the experiment more naturalistic, there is the difficult balance of wanting scientific control but also generalizability. If you want to keep it as a computerized task, what if you test people on memory for a face across different photographs of that person, rather than memory for a specific image? (It turns out recognizing an unfamiliar person from different photographs is a very difficult task! See Jenkins et al., 2011). New technologies are also making it easier to conduct experiments in more dynamic, three-dimensional environments like virtual reality, or even out in the real world (e.g., Snow & Culham, 2021; Võ et al., 2019). If we are able to expand our experiments out in these three ways, then we have a more generalizable study in these three ways as well. We can know things about face memory across a wider range of observers and faces being observed, and we can try to make predictions about behavior out in the real world.

For example, in our lab, we were curious about people's memory for faces more generally than the own-age effect. So, we generated a large database with demographics matching the United States (see Chapter 4 for more information). We then had over 800 diverse individuals engage in a face memory experiment online (Bainbridge et al., 2013). In this experiment, people viewed a stream of face images and pressed a key when they recognized a repeat from earlier (called a **continuous recognition task**) – a little like the experience of walking through a crowd and recognizing some people as you pass them. We also ran a version of the experiment where we tested people's memory for faces across different viewpoints and facial expressions,

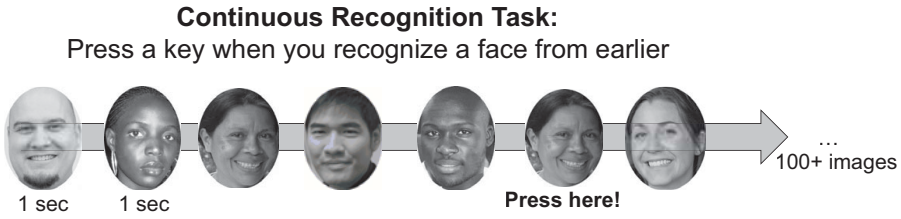


Figure 2.2 The experimental methods for our more "Big Data" face experiment. People still view face images for 1 second at a time. However, now they are continuously seeing images and indicating their memory as part of a "continuous recognition task," akin to the experience of walking through a stream of people and sometimes recognizing someone. We now also have a much more diverse range of faces, so we can test many phenomena, such as the own-age effect or the own-race effect, as well as measure the intrinsic memorability of a given face image. We would also test this task with many diverse participants, so we can look at more generalizable effects of the viewer, too.

not just face images (Bainbridge, 2017). So, our experiments looked a little more like Figure 2.2. What we found in the end was that there are certain faces that are remembered very well by most people, and some faces that are remembered very poorly. In other words, faces have an intrinsic *memorability*. We have recently extended these findings to images more generally, and tested them out in the real world – for example, discovering that we can even make predictions about what paintings people will remember in a freeform visit to an art museum (Davis & Bainbridge, 2023).

2.2 Limitations of Big Data

It may seem obvious that we'd want to aspire for diverse, generalizable experiments as described. However, there are some obstacles presented by Big Data experiments that are important to consider.

2.2.1 Problems with Big Experiments

One of the major cons of Big Data-style experiments is that they can introduce noise into our data. First, sometimes the images can vary too much. If each image is different along many dimensions (age, gender, attractiveness, facial expression, lighting, angle, hairstyle, eyebrow width, etc.), then how can we pinpoint which specific dimension is causing the effect we're studying? And maybe many (or even a majority!) of these dimensions might be things we don't care about, like lighting. Similarly, if our participants vary too much, we can have the same problem – everyone may act in a unique way that prevents us from finding generalizable effects. And participants may vary in ways that are not interesting to us – for example, in running an online experiment, what if the participant's screen monitor resolution, or what they ate for breakfast that specific day, or the length of their fingers influenced how quickly they pressed keys on a task? Small experiments let us directly test our effect of interest, without having all these

extraneous factors in the way because we control for them as much as possible. So, sometimes, we have to design our Big experiments in a way that they're not *too* big. At the same time, Big Data experiments let us simultaneously look at the contributions of many factors – for example, we can analyze how gender, age, and race all contribute to face memory, at the same time.

2.2.2 Imperfect Experiments

Even when we want to run Big Data experiments, there will always be limitations that make it impossible to run a perfect experiment. There will inevitably still be biases with the stimuli and participants in the experiment, because you cannot make everyone in the world participate in the experiment, or make everyone agree to be photographed as a face in the experiment. There are some factors limiting who can be a participant in your experiment – participants must have some familiarity with how to read a computer, and they have to have free time and interest in participating over other things they could be doing. Similarly, only a subset of people would be okay being photographed for the study, and any set of natural photographs will likely have an over-representation of happy or neutral facial expressions (over angry or sad).

There are also some other practical limitations with Big Data. Sometimes the data is so big that we are limited by the processing power, storage, or internet speeds that support us saving and analyzing the data. For example, one person's MRI brain data can take up 1 terabyte of space, which is more than the amount of space many computers come with (in 2025). It can also take half a day to download this data for just one person! So, it can be difficult to analyze data from hundreds, let alone dozens, of participants. Large-scale experiments can also be very costly with time and money. Using the same example of an MRI experiment (which is on the upper end of what psychology experiments cost), one participant usually lies in the scanner for about 2 hours, and it will cost the researchers around US$1,000 to the scanner center for that time. So, an experiment with 100 participants would end up costing $100,000 and take 200 hours of the researchers' time to just collect the data. We are also still limited in our analytical techniques for Big Data. When dealing with very big, naturalistic data, we often don't look at just a single measure or statistic. But, at the same time, our statistical tools and artificial intelligence are not yet able to fully interpret natural human behavior. For example, let's say we wanted to look at face memory in the real world, and recorded participants' view as they navigate through a party, using some sort of head-mounted camera. It's not clear how we would analyze these data – how to turn the conversations with people, the amount of time looking at them, the thoughts related to them, etc. – into numbers in order to make conclusions about what influences someone's memory of a person. So, our analytical techniques are limited (and in fact, they are dependent on the study of psychology to guide us on how to analyze such complex human behaviors).

2.3 Hypothesis-Driven versus Data-Driven Research

A majority of psychology experiments can be characterized as **hypothesis-driven research**. These are experiments where the researchers have one or a few key research questions. They also tend

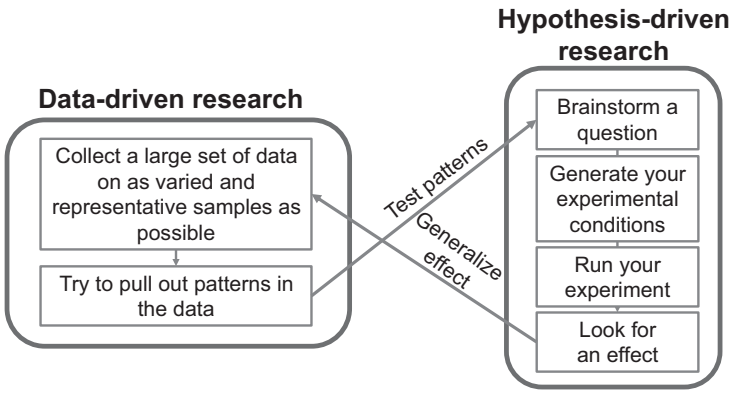


Figure 2.3 The series of steps you take when conducting data-driven research and hypothesis-driven research, and the way in which they interact. In data-driven research, you collect a large, representative set of data and identify patterns in the data. You can then take those patterns and test them in controlled, hypothesis-driven experiments to determine the specific mechanisms driving the effect. To conduct hypothesis-driven research, you would brainstorm your research question, create experimental conditions to answer that question, and run your experiment. If you identify an effect in a hypothesis-driven study, then it can be helpful to test whether the effect generalizes by running a more naturalistic, data-driven study.

to have a clear idea of the different alternatives of the results (in other words, **hypotheses** about the results) and what that means for the bigger picture question. The general pipeline for running a hypothesis-driven study follows the flow of the right side of Figure 2.3. One of the most important first steps is deciding on the main question. So, taking the first case study example we covered, let's say your question is: "Are people better at remembering faces close to them in age than faces far from them in age?" Your two experimental hypotheses would be something like: 1) yes, faces closer in age are remembered better, or 2) no, there is no difference in memory based on age of the face. You will then design your experiment, selecting experimental conditions that let you pinpoint that question. **Experimental conditions** are the different ways you divide up your experiment to answer your research question. For example, for this study, we may have different participant groups (older and younger people) as well as different stimulus sets (older and younger faces). So, we would have four conditions: young participants viewing young faces, young participants viewing older faces, older participants viewing young faces, and older participants viewing older faces. In experimental design terms, participant age is a **between-subjects** factor, because the condition changes from subject to subject (in other words, between the subjects). In contrast, face image age is a **within-subjects** factor, because within a single subject, they will see faces from both older and younger face conditions. Importantly, the conditions and factors that we intentionally change or control as experimenters (like the gender of participants and faces) are called **independent variables (IVs)**. It's because these are variables that we set, and so they are not dependent on other things in the experiment – they stand on their own. Once you have designed your experiment with its conditions, you then run the experiment (i.e., having your participants do a memory task with

the images you choose). The data that comes out of your experiment are **dependent variables** (DV), because they are dependent on the IVs and the experiment that you run. Finally, you can then run statistical tests on your data to directly answer the research question you set out to test. For a statistical test, you will typically define your **statistical hypotheses** – these are subtly different from experimental hypotheses in that they specify the different possible results of your statistical test. With most traditional statistical tests (called parametric statistics; see more discussion in Chapter 3), you would have a **null hypothesis** reflecting the hypothesis that there is no effect for a given statistical comparison. In this case, the null hypothesis would be that there is no difference in memory performance across our different participant-image age conditions. You would also define an **alternate hypothesis** reflecting the hypothesis that there is a significant effect. Here, the alternate hypothesis would be that there *is* a difference in memory across these conditions. So, we run our test and see which hypothesis we have more evidence for – can we reject this null hypothesis or fail to do so? Specifically, you would directly test whether memory performance (your DV) is higher for same-age conditions (young participants / faces and old participants / faces) than different-age conditions. And, voila, you have your answer! (So far, the research seems to point to yes: Chiroro & Valentine, 1995).

This hypothesis-centric way of thinking often goes hand in hand with small data research, because you want to run experiments that specifically target your question of interest. You *can* run these in a Big Data way (e.g., with thousands of participants and thousands of images), but in the end, the only factor you'll want to differ is the one you're interested in (age), and you'll want other factors like race, gender, attractiveness, etc. to be the same across your different conditions. This is because you want to be certain that your experimental manipulation has a direct influence on the effect you observe (in other words, you want your IV to directly influence your DV). If you do not control for other factors, you risk having **confounds** that could explain the link between your IVs and DVs. A confound is another variable that can account for the relationship between your IV(s) and DV(s) that you are not intentionally manipulating. In other words, they can be alternate explanations for your results. When designing an experiment, you want to make sure you can avoid these confounds, or at least can take them into account in some way. For example, let's say we look at monthly data across a year and find a correlation between ice cream sales and drowning deaths: when ice cream is popular, drowning is more common (Mumford & Anjum, 2013). Does this imply that ice cream causes people to drown? That would be ridiculous (and unfortunate)!

Discussion Question

What are some confounds that could explain a relationship between ice cream sales and drowning deaths?

One of the big confounds here is weather! During the summertime when the weather is nice, people want to go out swimming. They certainly have a much higher risk of drowning if they're swimming in a lake than if they're staying home bundled up by a fire during the wintertime. During the summer, people are also probably going out to buy ice cream to cool off from the hot weather, so you would see both high ice cream sales and increased drownings. On the other

hand, during the winter, the dessert of choice might be something more like a slice of apple pie or a mug of hot chocolate, and you would be unlikely to be out swimming. So, we would say that weather here is a confound in this relationship between drownings and ice cream sales. When designing a hypothesis-based study, it's important to keep in mind all potential confounds, and sometimes it can be impossible to control for all of them.

The counterpoint to hypothesis-driven research is **data-driven research**. The idea is that you collect tons of data, trying to gain samples that are as representative and varied as possible (Figure 2.3), and this is the approach often taken when using Big Data. Then, you use statistical methods to try and pull out patterns from the data that can help answer some questions. For example, you could collect memory test data for a wide range of faces and participants, and then see if people generally tend to remember faces closer to their own age best. This type of research is also sometimes called **exploratory research**, because you can explore around the data and look for different patterns without necessarily having a hypothesis from the beginning. What's great about data-driven research is that you generate big datasets that can help answer many questions. So, these databases can be multiuse – you could look at questions about memory and age, but also memory and attractiveness, or memory and face shape. You can also look at how these different factors all work together to form the big picture (i.e., what combinations of features influence the memorability of a face?). However, because these data tend to be collected without controlling much in the experiments, you run a higher risk of having more confounds that can explain your effects. However, because you often aren't relying on a single research question or statistical test, one confound may be less impactful on the use of the dataset overall. However, if you are not careful, data-driven research has some other big risks that can result in low-quality science (see Section 2.5 on data fishing).

Overall, there are pros and cons to both hypothesis-driven and data-driven research, with the key points summarized in Table 2.1. But ultimately, science benefits most when we do both, because they serve as an interconnected loop. Data-driven studies let us discover new, unexpected effects that can emerge from large or naturalistic data. Hypothesis-driven studies then let us take these effects and pinpoint the reasons behind these effects and link them to broader theories about human cognition. A lot of psychology reasonably takes the hypothesis-driven approach as a result. But to broaden our perspective on what questions to ask and what blinders we might have on in the field, I would argue the data-driven approach is just as important – and this is what will be the focus of this textbook.

Table 2.1 Comparison of the pros and cons of hypothesis-driven and data-driven research

Hypothesis-driven research	Data-driven research
Usually small data	Usually Big Data
Can isolate specific effects	Can be more naturalistic
Pulling out data based on theory-driven questions	Pulling out questions based on diverse datasets
Larger effect sizes	Statistical significance more likely
Have to be wary of confounds	Have to be wary of data fishing

2.4 Deep Data versus Wide Data

When designing a Big Data study, there are two dimensions along which it can be Big – deep or wide. A **deep data** study is one where you collect lots of data for a smaller number of individuals (so you're getting a deep look at a few people). Some examples include sensor data like a fitness watch recording frequently and over long periods of time (Chapter 9), or software-based data recording lots of samples over time (like on your phone; Chapter 8). There are also some experiments that focus on running the same participants many times over a series of sessions. These repeated measurements can give rich information about individuals, letting us look at the influence on cognition of things that vary like the time of day, attention fluctuations over time, and complex behaviors. One issue with deep data, though, is that it can risk being invasive of participants' privacy because you're learning so much about specific people. For example, if you look at one person's measurements from a fitness watch over months, you would learn all about their sleeping and exercise habits. It can also be tedious for participants to collect and provide all this data, especially if it's a study where they have to come in for multiple sessions. So, it can be hard to recruit participants for deep studies.

A **wide data** study is one where you collect relatively small amounts of data from a large number of participants at a single time. Some examples include data from an online experiment across thousands of people, or a snapshot of rich data from an app or piece of software at a single point (like all the tweets for a topic on a single day). What's great about wide data is that it can give diverse information across a large, representative sample of people. However, you often cannot capture very complex behaviors that vary over time or an interaction.

Some researchers characterize Big Data along three dimensions – being deep, wide, and long. In this case, deep data would still involve multiple measurements (like we see with sensor data). However, wide data now instead reflects collecting data across multiple variables or measures (e.g., with a battery of questionnaires). Finally, long data would involve collecting data from many people. Regardless of how you characterize the dimensions of Big Data, studies can be any combination of deep, wide, and long (e.g., collecting tons of data from many people), although this can be hard to achieve, so scientists may need to pick one dimension along which to specialize. When looking at a study, it's worth thinking how it falls on these different measures of size.

2.5 Big Ethical Questions

When you have a huge experiment, it can be easy to go fishing around for significant effects. This is something called **data fishing**, **data dredging**, or p-hacking. Since you have so much data, it seems like one of the benefits is that you should be able to look at many different effects in your data at once, right? Well, yes, you can do this to some degree, but you also need to think about how statistical tests are conducted.

For a majority of standard statistical tests that compare your data to a distribution (like t-tests, ANOVAs, regressions, etc.), you aim to estimate a ***p*-value**. What this *p*-value represents is the probability you would observe something as extreme as your results if the null hypothesis

were true. Recall that the null hypothesis reflects the hypothesis that there is no effect in your data. However, even if this null hypothesis were true, there is noise in our measurements and people's behaviors, so we would still sometimes observe a difference between our conditions "by chance." When we run a statistical test, we are looking at what the distribution of data would look like if the null hypothesis were true (and there was no effect). We are then seeing where our observed data falls in this distribution – how likely is it to occur given this null distribution? We calculate our *p*-value as the proportion of data in the null distribution that is equal to or more extreme than our observed value. So, for example, a *p*-value of 0.03 indicates there's only a 3 percent chance you would happen to observe these results just by random chance. That seems pretty low, and as a field, we've currently accepted a cut-off of 5 percent ($p < 0.05$) to be how we determine what we'll take to be a significant finding or not. Another way to phrase this is that in our field, we have accepted a 5 percent **false positive rate.** This is the rate of falsely saying something shows an effect when it does not. You may have heard this term used to refer to the rate of a medical test falsely saying you have an illness when you do not – same idea.

While this 5 percent chance of a false positive seems rare, when you're dealing with Big Data, you're doing many statistical tests – maybe hundreds of tests (e.g., for a psychological battery), or even up to hundreds of thousands of tests (e.g., for the case of MRI brain data). And so, in the realm of hundreds of thousands of tests, even with this seemingly strict false positive rate, we will get about 5,000 tests (5 percent of 100,000) that come out as "significant" just by chance! So we need to think carefully about how we define significance with data-driven research since we are doing many tests, inflating the chance that we find a false positive in at least one of these tests. In order to circumvent these issues, we do something called **multiple comparisons correction**, which is a group of statistical methods that let us calculate an adjusted *p*-value threshold for our study that takes into account the many tests that we are doing. While we won't go into these methods in detail, some example methods include **Bonferroni correction** and **false discovery rate correction**. Bonferroni correction corrects for the rate of false positives across all of the statistical tests that you perform. It does this by calculating an adjusted threshold for "significance" (called the alpha level, or α), based on the number of tests you are running. So, if you run ten tests, your alpha level would be $p < 0.005$ instead of $p < 0.05$. False discovery rate correction is a more liberal method that corrects for the proportion of false positives among all results initially labeled as significant – in other words, calculating an alpha level so that we are okay with 5 percent of our discoveries being false positives.

Let's look at an example study that ran many statistical tests. In Moore et al.'s 2006 study "Thongs, flip flops, and unintended pregnancy," the researchers wanted to investigate if there were some lifestyle factors that were related to unintended pregnancies. They conducted a 50+ question survey with 126 women who were currently or recently pregnant, and conducted 362 statistical tests to analyze their data. They found some surprising results: unintended pregnancy was associated with preferences for yoga, beaches, thongs, Doritos, contact lens, and text messaging. They also found that baby boys were more common if mothers preferred trucks, beef, and boys, while baby girls were more common if mothers preferred cars, chicken, and girls. So does that mean if you see your friend texting their friends during some beach yoga while tucking into a bag of Doritos, that you should encourage them to be vigilant with their

contraception? No, because if you think back to what we just discussed with running many statistical tests, we would expect about 18 of their 362 tests to come out as significant just by chance given our *p*-value threshold of 0.05! So even if there are no meaningful relationships between any of these factors and unintended pregnancy, just because of random noise in measurement, participant behaviors, and the environment, it would be unsurprising to find some relationships that come out as statistically "significant" but aren't real.

So one of the risks of Big Data is that it's easy to run many, many statistical tests until you find something significant. Because of all of the rich data you have, it's tempting to test many different questions. There are also big pressures in the scientific world to publish significant results, so researchers may be tempted to focus on these "significant" results without accounting for the number of statistical tests that they're running. In other words, you may be tempted to fish around for a result in your big sea of data. There are three main ways to make sure you are not doing data fishing with your own data. One way is to perform multiple comparisons correction across all the tests you run. A second way is to decide your analyses and hypotheses in advance before seeing your data (called preregistration; see Section 4.4) – so in other words, running a combined hypothesis- and data-driven study. A third way is to replicate any findings you discover across multiple datasets, analyses, and/or labs to be sure that what you're finding is real, rather than something that emerges just by chance.

There is also the question of the **effect sizes** of the results you end up finding. While we often care about statistical significance in our data, we also care about how strong the effects are in the differences that we are measuring. Effect size is often quantified as the proportion of the **signal** of interest to the level of **noise**. So, for example, for a t-test, the measure of effect size is the difference between the conditions' averages (the "signal"), divided by the standard deviation pooled across the two conditions (the "noise"). You can have a significant effect that's a weak effect or a strong effect. For example, let's say you're looking at whether an intervention in the classroom results in a difference in test scores on a test with a maximum of a hundred points. You could get a significant effect where the intervention results in a one-point increase. While this would mean the intervention likely worked (because the effect is significant), it didn't work very well (the effect is weak)! If the intervention instead resulted in a significant thirty-point increase, we would say this is a strong effect! And, if you have a nonsignificant effect with a thirty-point increase, that would mean our results aren't strong enough (e.g., there may be too much noise), so we cannot be confident that this thirty-point increase didn't just happen by chance. Because of its large sample sizes, Big Data can be prone to identifying significant but weak effects – effects that would only be detectable when you have thousands of people. Therefore, even if you find a significant result, consider what the result means. If it is a meaningful, strong psychological effect, ideally we would even see it occur at the level of a smaller sample, and even at the level of the individual.

2.6 Applications of the Chapter

In this chapter, we discussed characterizing research in a few different ways – for example, hypothesis-driven versus data-driven or deep versus wide data. These ways of thinking about

data have promoted discoveries beyond the field of psychology, and have guided recent advancements in the medical field.

2.6.1 Data-Driven Discoveries

We mentioned how data-driven research can result in new questions or effects that we may not have conceived of if we only stuck to pre-existing theories and hypothesis-driven experiments. There are in fact many exciting scientific discoveries that came about thanks to people trying out many different things. One of the most famous examples in psychology is the discovery by Professors David Hubel and Torsten Wiesel that led to their Nobel Prize win in 1981. They were trying to see what information was coded in neurons in the occipital lobe (the early visual regions of the brain), by recording directly from cats' brains while showing them different images (see Chapter 10 to learn more about neuronal recording). They were struggling to find any specific image that would cause these neurons to spike. Back in the day, they were using a slide projector, and suddenly when they were swapping out the slides, they heard the neuron they were recording from start to fire. After playing with the slide, they discovered that this neuron was sensitive to the edge of the slide when it was shown at a specific angle. This led to our current understanding of the visual system in the brain, where neurons are sensitive to edges oriented at specific angles. You may have heard about similar fortuitous "eureka!" moments throughout the sciences. As the classic example, around 246 BC, Greek scientist Archimedes realized how to calculate volume and density while taking a bath, and purportedly ran through the streets shouting "eureka!" In 1820, Dr. Hans Christian Oersted noticed a compass move when he placed it near an early battery he was creating – resulting in the discovery that electrical currents generate a magnetic field. Percy Spencer invented the microwave in 1946 when he noticed the chocolate in his pocket melted when he was testing out a new vacuum tube. These discoveries may have never happened without the experimenters just trying out different things. Big Data can encourage such exploration, which can lead to exciting discoveries.

2.6.2 Medical Applications of Deep and Wide Research

Deep and wide methods have had some wide-reaching applications in the clinical realm. Deep data has helped form the field of **precision medicine**, where healthcare workers can make honed, personalized predictions of health outcomes based on genetics, environment, lifestyle, and sensor measures. Big Data lets us create **predictive models** that take these different factors and then make guesses about outcomes for a single person (see Chapter 6). Precision medicine goes hand in hand with preventative medicine and telehealth, where people can wear sensors and use apps to remotely track and communicate symptoms before they develop into a full-blown condition. For example, researchers are working on apps to help identify early stages of Alzheimer's disease (Konig et al., 2018) and apps to help elderly individuals develop memory strategies (Martin et al., 2022).

Wide data is key in letting us learn about diseases: It lets one see global trends in disease, identify rare groups at particular risk, and find hidden links to a cause or cure (Heggie, 2019). Wide data played an important role in identifying the symptoms early on in the COVID-19

pandemic, when it wasn't clear what symptoms were being caused by the virus. Researchers conducted a large-scale wide symptom study where anyone could enter their symptoms online, and they ended up receiving information from 4.4 million participants (Menni et al., 2020). As a result, they were one of the first groups to identify a loss of smell or taste as one of the symptoms of COVID-19. They also found some other interesting trends: for example, for the first wave of the pandemic, one out of twenty participants had symptoms that lasted more than 8 weeks, and longer COVID was correlated with having more different symptoms in the first week. They also found that during the pandemic lockdowns, 20 percent of participants had an increase in alcohol consumption, and an average weight gain of 4.6 lbs.

CHAPTER SUMMARY

In this chapter, we discussed the small data experiments traditionally utilized in psychology research and showed how they compare to the Big Data experiments that are becoming increasingly popular. Here are some of the main takeaways:

1. The key to making a small data experiment "Big" is expanding its participants, stimuli, and paradigms to be more naturalistic and representative of the real world. However, you can almost never make a perfectly representative experiment.
2. There are different benefits to hypothesis-driven research versus data-driven research, and both are necessary for the progression of psychology as an innovative and rigorous field.
3. Big Data can be characterized by two key dimensions – its depth (how many measures you collect per individual) and its width (how many individuals you record from).
4. With the large amount of data you can get from a Big Data study, we must be cautious of not "fishing" for effects without accounting for all of the statistical tests that we are conducting.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter.

- **Learn about how Big Data is causing big strides in the understanding of disease:** Heggie, J. (2019, January 8). How can big data beat disease? *National Geographic*. https://tinyurl.com/ykkabh53
- **A cautionary tale on how too many statistical tests can lead to effects that may not be real:** Moore, R. P., Galvin, S. L., & Imseis, H. M. (2006). Thongs, flip-flops, and unintended pregnancy: The seduction of $p < 0.05$. *MAHEC Online Journal of Research*, *1*, 1.
- **Dive deeper into the statistics used to correct for multiple comparisons:** Lindquist, M. A., & Meija, A. (2015). Zen and the art of multiple comparisons. *Psychosomatic Medicine*, *77*, 114–125.

ASSIGNMENT

The purpose of this assignment is to get you thinking about Big Data and how to build out Big Data experiments. Please submit your response in a way so that it is clear what questions and sub-questions you are responding to.

Total Points: 50

1. Pick two psychology papers describing an experiment on a topic that sounds interesting to you. (Do not use a review paper.) They can come from either:
   i) a psychology class you are currently taking, or took in the past;
   ii) a lab you are currently working in; or
   iii) any "Open Access" articles from the most recent year of the journal *Psychological Science* (https://journals.sagepub.com/home/pss).

   Please choose at least one paper that you would consider a "small data" experiment. Provide the citation (titles, authors, year, journal) and abstract of the two papers here: (2 points)

   We will now look at these papers with the frameworks we discussed in this chapter.
2. **For paper 1, answer the following questions**. If the study includes multiple experiments, answer for the first or main experiment:
   a. Is this a **"small data"** or a **"Big Data"** experiment? How do you know? How small/big is the sample size (number of participants)? How small/big is the experiment itself (e.g., number of conditions, stimuli, outcome measures)? (4 points)
   b. Is this a **hypothesis-driven** or a **data-driven** experiment? How can you tell? (3 points)
      i. If this is a hypothesis-driven experiment, what is their hypothesis?
      ii. If this is a data-driven experiment, what new hypotheses come out of their data? How did they avoid p-hacking / data fishing?
   c. Is the data **deep** or **wide** (or both)? How do you know? (2 points)
3. **For paper 1, we will do some more brainstorming on Big Data. (Note that the next part of the question has two options – for a given paper, answer for either small data or Big Data, not both.)**

   If this is a "small data" experiment, we will think up how to make it into a Big Data experiment. Answer these questions:
   a. How **naturalistic** versus **artificial** is their experiment? What are ways in which the stimuli, experiment, or participants are not *representative* of reality? What are ways in which they are? (3 points)
   b. How can we improve the **representativeness** of the study? Use your creativity to brainstorm how you would make this into a "Big Data" experiment. How would you change the experimental paradigm, participant recruiting, the stimuli, or the measurement techniques to capture bigger, more diverse, more naturalistic, and more representative data? (5 points)
   c. What **limitations** could you envision with these changes? These can be limitations in terms of feasibility/practicality (e.g., how much time or money does your change add)? In what ways is your version still not fully representative? (3 points)

   If this is a "Big Data" experiment, we will see how it improves upon small data studies. Answer these questions:
   a. What would the **"small data"** version of the experiment have looked like? (4 points)
   b. **Why** did the experimenters decide to take this "Big Data" approach? What innovations did they apply to make it "Big Data"? (4 points)

c. What **limitations** still exist with their approach? In what ways are the data still not fully representative of real people / images / cognitive processes? What additional improvements could you envision, and how feasible are they (e.g., how much time or money does your change add?) (3 points)

4. Answer questions 2 and 3 for **paper 2** below. (20 points)
5. What are some ideas implemented by paper 1 that could be useful for paper 2 in making their experiments more representative in terms of participants, stimuli, or paradigms? Similarly, what are some ideas implemented by paper 2 that could be useful for paper 1? (5 points)

REFERENCES

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*, 1043–1047.

Bainbridge, W. A. (2017). The memorability of people: Intrinsic memorability across transformations of a person's face. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 706–716.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*, 1323.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 48, 879–894.

Davis, T., & Bainbridge, W. A. (2023). Memory for artwork is predictable. *Proceedings of the National Academy of Sciences USA*, *12*, e2302389120.

Doyen, S., Klein, O., Phcion, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Heggie, J. (2019, January 8). How can big data beat disease? *National Geographic*. https://tinyurl.com/ykkabh53

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323.

Konig, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., & Robert, P. H. (2018). Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, *15*, 120–129.

Martin, C. B., Hong, B., Newsome, R. N., Savel, K., Meade, M. E., Xia, A., Honey, C. J., & Barense, M. D. (2022). A smartphone intervention that enhances real-world memory and promotes differentiation of hippocampal activity in older adults. *Proceedings of the National Academy of Sciences USA*, *119*, e2214285119.

Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., El-Sayed Moustafa, J. S., Visconti, A., Hysi, P., Bowyer, R. C. E., Mangino, M., Falchi, M., Wolf, J., Ourselin, S., Chan, A. T., Steves, C. J., & Spector, T. D. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine*, *26*, 1037–1040.

Moore, R. P., Galvin, S. L., & Imseis, H. M. (2006). Thongs, flip-flops, and unintended pregnancy: The seduction of $p < 0.05$. *MAHEC Online Journal of Research*, *1*, 1.

Mumford, S., & Anjum, R. L. (2013, November 15). Correlation is not causation. Oxford University Press blog. https://blog.oup.com/2013/11/correlation-is-not-causation

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*, 506–519.

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210.

3 Big Participant Samples

Introduction

Many students first gain their interest in experimental psychology as a guinea pig – by volunteering in an experiment on campus. It can be exciting to apply the knowledge you learn from class to guess the main manipulation in a behavioral experiment, or see inside your brain after an MRI study. It can also be a great way to earn money (at least I know I funded all my college vacations by being a "professional subject"!), or you may be required to participate in studies as part of passing a course. This built-in participant sample is also fantastic for researchers: They have a pool of eager young students who are readily available during weekday hours, attending class right next to your laboratory, and often attentive and enthusiastic about each new experiment. However, while it seems like running university experiments with college students is a win-win situation for both the researchers and the students, this convenient choice actually causes concerning damage to the field of psychology as a whole.

In this chapter, we will cover the problems with current norms in the participants we recruit for psychology experiments and how to solve some of these problems by taking a Big Data approach. First, we will go through how small data studies recruit their participants (Section 3.1). We will then talk about how the average college sample differs from adults worldwide (Section 3.2), individuals from smaller societies (Section 3.3), other industrialized nations (Section 3.4), and others even within the same country (Section 3.5). The issues boil down to a difference between our sample and population (Section 3.6), and we will discuss how we can move toward more representative groups using Big Data (Section 3.7). However, we will never be able to make a perfect sample (Section 3.8), and sometimes we may want to intentionally restrict the people we recruit (Section 3.9). The chapter will finish with a look at the big ethical questions surrounding participant recruitment (Section 3.10) and imbalances in the demographics of psychology researchers themselves (Section 3.11).

Big Data in the Psychological Sciences

Wilma A. Bainbridge

Big Data in the Psychological Sciences

Cutting-edge computational tools like artificial intelligence, data scraping, and online experiments are leading to new discoveries about the human mind. However, these new methods can be intimidating to many students. This textbook demonstrates how Big Data is transforming the field of psychology, in an approachable and engaging way that is geared toward undergraduate students without any computational training. Each chapter covers a hot topic, such as social networks, smart devices, mobile apps, and computational linguistics. Students are introduced to the types of Big Data one can collect, the methods for analyzing such data, and the psychological theories we can address. Each chapter also includes discussion of real-world applications and ethical issues. Supplementary resources include an instructor manual with assignment questions and sample answers, figures and tables, and varied resources for students such as interactive class exercises, experiment demos, articles, and tools.

Wilma A. Bainbridge is an associate professor in the Department of Psychology at the University of Chicago. She has won the Association for Psychological Sciences Rising Stars Award (2023), an Alfred P. Sloan Fellowship in Neuroscience (2024), and the American Psychological Association's Distinguished Scientific Award for Early Career Contributions to Psychology (2025). Her research has garnered attention from outlets such as CNN, *Vox*, and *Wired*. She has previously edited two books on vision and memory, and her "Big Data in Psychology" class has earned a Curricular Innovation Award from the University of Chicago.

"From social media to sensors to AI, this book offers a brilliant tour of how the Big Data revolution is reshaping psychology. Accessible, inspiring, and grounded in real research problems, it walks students through everything from hands-on skills like web scraping, to big-picture theory testing, and even thoughtful discussions of ethics – all presented with incredible clarity by one of the field's most inspiring new voices."

Timothy Brady, University of California San Diego

"Exceptionally timely and comprehensive, Bainbridge's textbook deserves a place in every curriculum for behavioral methods. The chapters – enhanced with interactive features and thought-provoking ethical questions – are so engaging that they make me want to teach the course. And whether or not you work with Big Data, this is essential reading for all."

Marvin M. Chun, Yale University

"Combining conceptual depth and accessible writing, Bainbridge offers a timely contribution with a comprehensive overview of the field, covering definitions of big data in psychology and expertly navigating its key sources, methods, and analytical approaches. It addresses both foundational topics, such as neuroimaging tools and statistical techniques, as well as emerging and contemporary discussions, including natural language processing, the development of large language models, and their applications in psychological research. It will resonate with a wide audience, from curious undergraduates to seasoned researchers looking to deepen their understanding of big data and its potential to reshape the psychological sciences."

Nemanja Vaci, University of Sheffield

Big Data in the Psychological Sciences

Wilma A. Bainbridge
University of Chicago

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009343589

DOI: 10.1017/9781009343602

When citing this work, please include a reference to the DOI 10.1017/9781009343602

First published 2026

Cover image: FrankRamspott / DigitalVision Vectors / Getty Images.

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-34358-9 Hardback
ISBN 978-1-009-34357-2 Paperback

Additional resources for this publication at www.cambridge.org/bainbridge

Big Data in the Psychological Sciences

Cutting-edge computational tools like artificial intelligence, data scraping, and online experiments are leading to new discoveries about the human mind. However, these new methods can be intimidating to many students. This textbook demonstrates how Big Data is transforming the field of psychology, in an approachable and engaging way that is geared toward undergraduate students without any computational training. Each chapter covers a hot topic, such as social networks, smart devices, mobile apps, and computational linguistics. Students are introduced to the types of Big Data one can collect, the methods for analyzing such data, and the psychological theories we can address. Each chapter also includes discussion of real-world applications and ethical issues. Supplementary resources include an instructor manual with assignment questions and sample answers, figures and tables, and varied resources for students such as interactive class exercises, experiment demos, articles, and tools.

Wilma A. Bainbridge is an associate professor in the Department of Psychology at the University of Chicago. She has won the Association for Psychological Sciences Rising Stars Award (2023), an Alfred P. Sloan Fellowship in Neuroscience (2024), and the American Psychological Association's Distinguished Scientific Award for Early Career Contributions to Psychology (2025). Her research has garnered attention from outlets such as CNN, *Vox*, and *Wired*. She has previously edited two books on vision and memory, and her "Big Data in Psychology" class has earned a Curricular Innovation Award from the University of Chicago.

"From social media to sensors to AI, this book offers a brilliant tour of how the Big Data revolution is reshaping psychology. Accessible, inspiring, and grounded in real research problems, it walks students through everything from hands-on skills like web scraping, to big-picture theory testing, and even thoughtful discussions of ethics – all presented with incredible clarity by one of the field's most inspiring new voices."

Timothy Brady, University of California San Diego

"Exceptionally timely and comprehensive, Bainbridge's textbook deserves a place in every curriculum for behavioral methods. The chapters – enhanced with interactive features and thought-provoking ethical questions – are so engaging that they make me want to teach the course. And whether or not you work with Big Data, this is essential reading for all."

Marvin M. Chun, Yale University

"Combining conceptual depth and accessible writing, Bainbridge offers a timely contribution with a comprehensive overview of the field, covering definitions of big data in psychology and expertly navigating its key sources, methods, and analytical approaches. It addresses both foundational topics, such as neuroimaging tools and statistical techniques, as well as emerging and contemporary discussions, including natural language processing, the development of large language models, and their applications in psychological research. It will resonate with a wide audience, from curious undergraduates to seasoned researchers looking to deepen their understanding of big data and its potential to reshape the psychological sciences."

Nemanja Vaci, University of Sheffield

Big Data in the Psychological Sciences

Wilma A. Bainbridge
University of Chicago

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009343589

DOI: 10.1017/9781009343602

When citing this work, please include a reference to the DOI 10.1017/9781009343602

First published 2026

Cover image: FrankRamspott / DigitalVision Vectors / Getty Images.

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-34358-9 Hardback
ISBN 978-1-009-34357-2 Paperback

Additional resources for this publication at www.cambridge.org/bainbridge

Thank you to the "RAV4" – Robert, Ally, and Vicky – for being a loving and supportive family! It's hard to believe I began this book when still pregnant with Ally and Vicky and am finishing it as they are running around and chatting our ears off.

Thank you to my mom Erika, dad William, and sister Connie – finally I get to write book dedications for you rather than the other way around!

And thank you to the wonderful Brain Bridge Lab and my department at the University of Chicago – your support has really helped me flourish and think Big.

Brief Contents

Detailed Contents

Preface

Learn how to see. Realize that everything connects to everything else.

Leonardo da Vinci (1452–1519)

We live in a world where we are all constantly generating data – in our interactions with our phones, social media apps, games, websites, fitness trackers, and more. This data is commonly referred to as "Big Data" because its scale is so large that it cannot be analyzed manually. Such Big Data serves as a useful means to understand human cognition – showing us how people see, feel, respond, remember, interact, and make decisions with these different tools. We can also look at these cognitive processes across different groups of people – across countries, cultures, ages, and experiences – as well as across species. As a result, Big Data ways of thinking and analysis have become incredibly important tools to psychologists, across fields. Psychologists are now running online experiments that can gather data from thousands of participants, running machine learning models that can decode patterns from thousands of datapoints, or analyzing brain data from thousands of subregions.

As a result, psychology as a field is at a major transition point. Familiarity with advanced statistical analyses and computer programming is becoming increasingly essential to keep up with the state of the art. However, the idea of wrangling Big Data can be incredibly daunting to people entering the field, especially given that most undergraduate psychology curricula do not require computational or advanced statistical coursework. The main goal of this textbook is to make these new directions in Big Data accessible and meaningful to any psychology student – without the need of training in computer science or statistics. By reading this textbook, you'll gain basic fluency and familiarity with the important topics in the field, so you can decide what topics you want to pursue more deeply. Students who are already familiar with computational methods will learn ways in which these methods can be applied to answer a myriad of psychological questions. As a result, the book will lightly touch upon a wide range of topics, including experimental design, web programming, data scraping, artificial intelligence, different methods in brain imaging, computational linguistics, network science, wearables, user interface design, crowdsourcing, and representative sampling.

To my knowledge, this is the first undergraduate textbook on Big Data in psychology. It was inspired by a course I created in Spring 2020 as a new assistant professor, and I've seen these sorts of courses start to grow in the last few years. Because this is such a new topic, this textbook and course is really for almost anyone. Familiarity with psychology is helpful (e.g., how experiments are run and what are some of the key topics of inquiry), and at some points I will bring up simple statistical concepts (e.g., *p*-values), although knowledge there is not

required. Each chapter focuses on a different angle of how Big Data interfaces with psychology, and includes sections on ethical questions related to the topic and its real-world applicability. Each section also includes thought-provoking questions that can be discussed as a class and an assignment that's relatively open-ended and should engage the students in thinking deeply about that topic. The chapters can be covered in pretty much any order, but the book is generally divided into two parts: 1) how to rethink psychology experiments from a Big Data angle (Chapters 1–7), and 2) various sources of Big Data to enrich the study of psychology (Chapters 8–12).

In conjunction with the Big Data theme, I also want to make this course follow the principles I preach in terms of modernizing psychological research. As a result, this book is paired with an interactive online resource (www.cambridge.org/bainbridge) that includes videos, demonstrations, links, and additional resources that will be constantly updated. This way, you will still have access to the latest developments in the field even after the publication of this book. I also maintain a public data repository on the Open Science Framework of Big Data student projects that came out of the course that I teach at the University of Chicago (https://osf.io/hz843), and I am happy to link to such a repository from anyone else using this book.

Now go forth, and think big!

1 What Is Big Data?

Introduction

The amount of data generated every day is insane. Each day, we create approximately 403 million terabytes of data (or 403 exabytes) (Duarte, 2024). This is about how much data can be stored by 4 billion phones (those of about half the world population) – and just in one day! In that same day, about 300 billion emails are sent, 8.5 billion searches are done on Google, 1.6 billion swipes are made on Tinder, 1.4 billion hours of video are streamed, and $638 million is spent on Amazon. You as an individual are contributing to a lot of this growing collection of data. As you commute to your classes, your map app tracks your movement behavior and may take note of any specific locations you visit. Your phone or watch tracks your steps and sleep patterns. As you look up web pages on your phone, these pages track your browsing and click behavior. And as you scroll through and post on social media, these apps track how you engage with posts, through measures like viewing time and click behavior. We are constantly surrounded by and creating Big Data. This Big Data can be messy and tricky to sift through, but within it are potential insights about the human mind waiting to be discovered.

In this introductory chapter, we will establish definitions of the central themes of this book, to guide you as you read the rest of the book. First, we will talk more about how data has changed in the last few decades (Section 1.1) and then provide a definition of Big Data (Section 1.2). We will then define Psychology within the context of this book (Section 1.3). With these two definitions in hand, we will discuss how Big Data and Psychology interact (Section 1.4) and why now is the perfect time to study this interaction (Section 1.5). Finally, we will wrap up this chapter with a guide on how to use this book and its online resources (Section 1.6).

1.1 Moore's Law

Our data has gotten so *Big* thanks to the exponential growth in processing and storage power over the past handful of decades. This is reflected by **Moore's Law**, which was proposed by Intel cofounder Gordon Moore in 1965 (Moore, 1965). This law predicts that the number of transistors (one of the key components in computer chips) that can be packed into a given

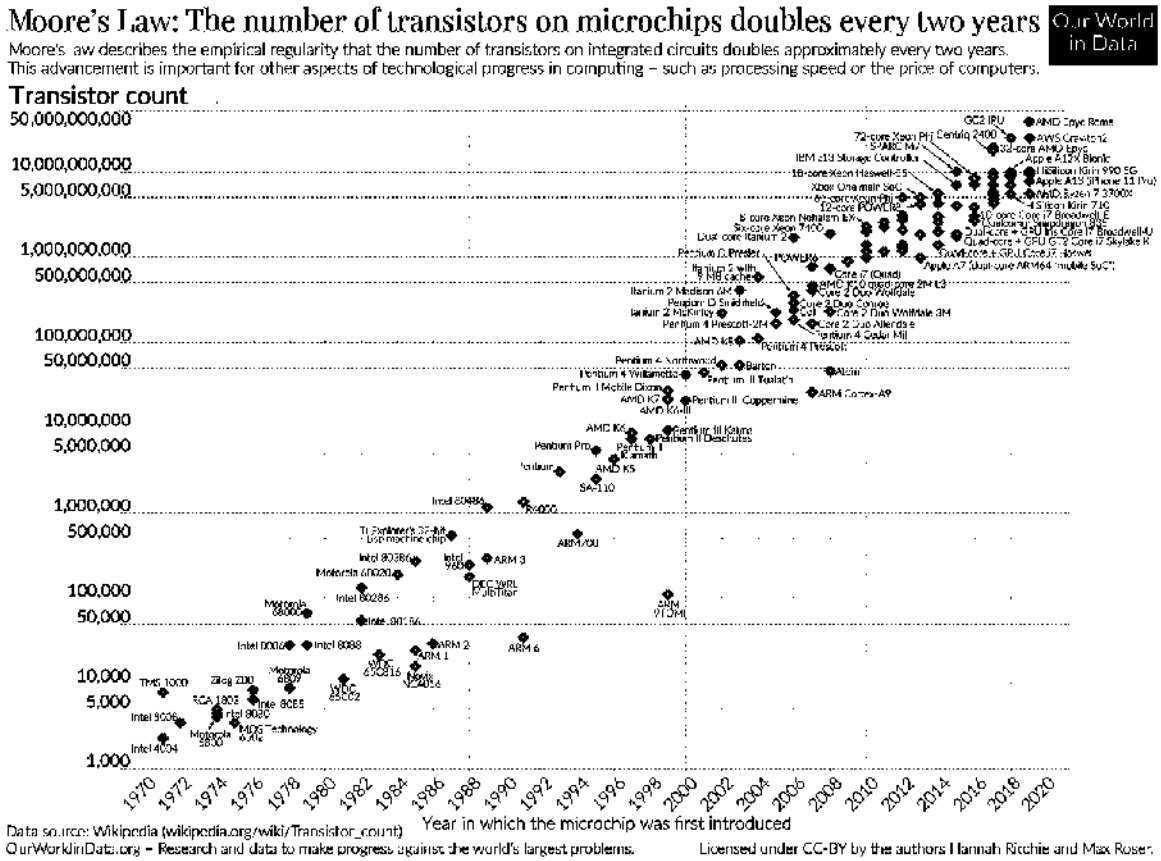


Figure 1.1 A depiction of Moore's law, showing it still holds in 2020. Moore's law posed in 1965 predicts that the number of transistors we can fit in a circuit will double every two years – resulting in exponential growth in our computing capabilities. Note that the y-axis here is an exponential scale (1,000 and 5,000 at the bottom are spaced as closely as 10 trillion and 50 trillion at the top), so indeed, we are keeping up with this law!

unit of space will double roughly every two years. Remarkably, this prediction of exponential increase in computing power has held true for 60 years (see Figure 1.1), although some scientists forecast that we will reach the limit of feasibility within the next few years (Kumar, 2015; Waldrop, 2016). We can feel the effects of Moore's law by looking at how the size of storage devices has drastically changed over our lifetimes.

Discussion Question

What did the size of data (e.g., devices, files, performance) look like when you were a child versus now? Does it feel like there has been exponential growth in that time? What sorts of innovations enabled that growth?

To understand what makes a set of data Big Data, let's first discuss how data is measured. The building blocks of the data and processes in our computers are 0s (off) and 1s (on), and a single digit is called a **bit**. Because of this 1/0 building block, instead of data being measured in our decimal base-10 system, data sizes are measured in **binary**, or a base-2 system, where the only digits possible are 0 and 1. When you want to count in higher numbers in binary beyond 1, you use additional digits (that are still limited to 0 and 1). So the numbers 0, 1, 2, 3, 4, and 5 in decimal are represented as 00, 01, 10, 11, 100, and 101 in binary. While these building blocks seem simple, they can combine to form the complex data we interact with on our computers – just as letters can combine to create the complexities of language. Because of the binary system, powers of 2 end up being important to the measurement of data. A set of eight bits ($2\times2\times2$) is called a **byte**. A byte can be used to represent a single character of text. For example, in the most common character encoding standard for computing called **ASCII (American Standard Code for Information Interchange)**, the letter A is represented by the byte containing the bits 0100 0001, while a space is represented by 0010 0000. Above the level of the byte, the naming of the counting system resembles that of the metric system. A set of 1,024 bytes is called a kilobyte (KB), like how 1,000 meters is a kilometer (but because we are operating in binary, it is a multiple of 2, or 2^{10}). A set of 1,024 kilobytes is called a megabyte (MB). A set of 1,024 megabytes is called a gigabyte (GB). After that, we have terabytes (TB), petabytes (PB), exabytes, and zettabytes. So, for example there are 8,000,000 bits (1s or 0s) in one megabyte of data. When we talk about data transfer speeds (like how fast your internet is), the measures tend to be in bits per second (instead of bytes per second). So early internet modems would have a download speed of 28.8 Kbps, or around 28,800 bits per second.

The rapid change in computing sizes is quite drastic when we look at the history of data storage across personal computing (Figure 1.2). Back in 1956, IMB shipped its first hard drive. It was the size of two refrigerators and could hold 5 MB of data – the equivalent of about one song. In the 1970s, some consumers were starting to get their own computers, and the most common way to store and transfer files was through floppy disks. These could only hold about 100 KB in early years, and 1.44 MB in later years, the equivalent of a few text documents or pictures. However, this medium became so ubiquitous that many pieces of software still use an icon of a floppy disk as their "save" icon. Once software became more advanced, users needed more and more of these disks – for example it took seven floppy disks to install an early version of Adobe Photoshop (Adobe, Inc., 2013).

In the late 1980s, a more advanced data storage method emerged – the CD-ROM (compact disc read-only memory). These could hold as much as 900 MB – about one-third of a movie. However, as their "read-only" name implies, these needed special devices called CD burners to write data to the CD-ROM, and most CDs could not be rewritten once data was saved onto it. In the mid-1990s, we moved onto DVDs (digital video disks), which could store closer to 5 GB (about 1–2 movies) but faced similar shortcomings as CD-ROMs. The early 2000s saw the first USB (universal serial bus) flash drives, which were smaller and more convenient but could only store about 10 MB at first. The early 2000s also saw the explosion of the internet, and **cloud storage** – saving data to online servers – started to grow. This really took off as internet speeds became faster, and websites emerged dedicated to hosting large amounts of data – YouTube for video started in 2005, Dropbox for files started in 2007, and Flickr for photos started in 2004.

Figure 1.2 Photos of various types of older data storage. **(a)** The shipping of IBM's first hard drive, holding 5 MB and taking up the size of two refrigerators. **(b)** A 3.5-inch floppy disk, the main external data storage format in the 1980s and 1990s. **(c)** A CD-ROM inserted into a laptop's CD drive. These were commonly used in the 1990s and 2000s for data storage and were the main medium for holding music albums.
Source: (a) Michael de Groot / Flickr. (b) Pablo Jeffs Munizaga – Fototrekking / Moment / Getty Images.
(c) EThamPhoto / The Image Bank / Getty Images.

In the 2010s, storage amounts and data transfer speeds increased by many magnitudes. Personal computers could hold close to a TB of data, and mobile technologies emerged enabling people to generate vast amounts of data through the photos and videos they're taking. In the 2020s, higher mobile data speeds and faster personal computer processors are allowing us to interact with vast amounts of data and at faster rates – supporting growths in avenues like online gaming, video streaming, and real-time artificial intelligence (AI) on board many of our systems.

1.2 How Do We Define Big Data?

Owing to this constant growth in technology, the definition of Big Data is a moving target. In the 1990s, even a low-resolution video would be considered Big Data, while in the 2020s, Big Data is more in the order of libraries of thousands of movies, representing petabytes of data (or even more). A fundamental aspect of what makes data qualify as Big Data is that it is so large that we cannot process it by hand – we can't manually input, clean, or analyze the data. This is data that is also often so large that we cannot process it with basic programs on our computer (e.g., Microsoft Excel), but usually have to use code or bespoke tools. For our purposes of studying human cognition, Big Data tends to be more naturalistic – recorded from real people or dynamic behaviors – so it may be unstructured or more noise- or error-prone compared to smaller datasets. Data can be Big for several different reasons. It may have very high temporal sampling – for example, taking a measurement every handful of milliseconds. It may have high spatial sampling instead – for example, collecting data across all the intersections in a city. It may have high participant sampling – collecting data from a large and diverse set of people. Or, it may have high stimulus sampling – capturing data from lots of sources (e.g., images, videos, news articles, products) or tasks. All of these encompass examples of Big Data.

So here, we will loosely define Big Data as: *unstructured, naturalistic human data requiring complex analytical methods*. For some of the exercises and examples we discuss in the book, we may not use the massive amounts of space or computing power that traditionally make up Big Data. But the concepts that you learn here should translate to bigger sets.

1.3 How Do We Define Psychology?

It is also important that we define what psychology is within the framework of this book. Broadly, psychology is *the empirical study of the mind, brain, and behavior*. For the vast majority of this book, we will focus on a more quantitative and research-based approach to psychology, where psychologists conduct experiments that aim to provide broad insight, using falsifiable questions and hypotheses. This is in contrast with some psychologists who use more *qualitative* approaches, like revealing new insights using interviews or observations, or using therapeutic techniques like talk therapy to help improve mental health. A central aspect of the psychology we will discuss here is that it is composed of research questions that are testable and falsifiable. **Falsifiable questions** are those where you can obtain evidence to prove that question wrong. For example, one active area of debate is the degree to which we can falsify questions in **evolutionary psychology** – the study of the mind, brain, and behavior through the perspective of evolution (Gannon, 2002). We cannot go back in time and run experiments on our ancestors. We also cannot measure how much of people's current behaviors are a result of evolution versus more recent societal norms. There are some creative scientific methods to test evolutionary hypotheses by looking at other animal species or running computational models. However, we will generally avoid tackling unfalsifiable topics in the current book.

We are also interested in questions that are **generalizable** – that give us insight into the thinking of a group of people, and can allow us to make predictions in the future about different events. For example, a question like "how did people feel about Argentina winning the 2022 World Cup?" is a question that measures emotions and behavior, but is so highly

specific that it does not really teach us about the human mind. Thus, we would not characterize this as a psychological question. A more generalizable psychological question might be something like "how does sentiment in social interactions change directly after major sporting events?" Sometimes when dealing with Big Data, we may accidentally take an overly narrow scope, due to the data that we have available (like data from a specific app, Chapter 8). But we should always try to focus on a big-picture question about the mind, and any limitations to the data we are collecting to answer this question.

There are many different branches to the field of psychology. When you think of a "psychologist," your mind may first go to **clinical or abnormal psychology** – the study of atypical behavior and mental health, with the goals of diagnosis and treatment. Closely related is counseling psychology, which is the practice of helping people through therapy and counseling. While this book will discuss some research on abnormal psychology through the lens of understanding the underlying roots of an impairment, we will not discuss in depth therapies or treatments of individuals. **Industrial psychology** is the study of the mind within the workplace, and how to optimize people's effectiveness at work. As this field is more applications-focused, we will not discuss this at length in this book, nor other more applied fields like forensic psychology, school psychology, and health psychology. The major focus of this book will be **cognitive psychology**, the branch of psychology dedicated to the scientific study of our internal mental processes. This encompasses a broad range of processes, including sensation, perception, action, memory, reading, speaking, emotion, decision-making, morality, imagination, and others. In fact, a "psychologist" can be a researcher with a laboratory that runs experiments to study these processes (this is the type of psychologist I am). Related to cognitive psychology, we will sometimes discuss the brain, through a lens of **neuropsychology** – looking at how the brain and mind interact. We will also bring up many examples from **developmental psychology** – the study of the development of the mind across the lifespan (from infants through aging) – and **social psychology** – the study of the interactions of multiple minds.

1.4 How Do Big Data and Psychology Interact?

A large proportion of Big Data out in the world just *happens* – you record a video and post it on social media, and now there are several new megabytes of data on your phone, on a server belonging to that social media site, and being downloaded to other people's phones. In this way, much of Big Data is just passively accumulated as we perform tasks with our phones, computers, and the internet. Another major slice of Big Data is being actively collected by companies, where they are testing how they can improve your experience, how you navigate their app, and how they can improve engagement and purchasing. However, the data being generated out in the world also serves as incredibly rich records of human behavior that can give insight into questions on almost any topic of psychology.

Discussion Question

What are some ways you can envision Big Data might be changing the types of questions we can ask or answer in psychology?

An important skill for you to nurture will be in identifying these intersections of Big Data and psychology. What is a psychological question you want to answer, and how could Big Data answer that question? Could there be a preexisting dataset out there that answers the question for you, or could Big Data help you collect that data in some way? For example, a few years ago, I was curious how older memories (2+ year-old memories) might be represented in the brain. This is hard to test in the laboratory because I would need to have participants study some images and then come back two years later. But then it dawned on me that people are constantly capturing their memories on social media, dating back to many years prior. So, I collaborated with the app 1 Second Everyday to recruit users who had recorded years of their memories, and then I scanned their brains while they viewed these older memories. Long story short, we found patterns in the brain reflective of the age of a memory (Bainbridge & Baker, 2022; see Section 8.4.2). As you look through data in your daily life, think to yourself – what does this reflect about the human mind and can it show us something new? And, are there ways in which Big Data technologies are influencing how we think or interact? For example, an active area of current research is how social media may be impacting feelings of isolation and depression (Section 12.3.2). Overall, a major part of this class will be thinking creatively and with an open mind on how we can use data to answer questions.

1.5 Why Study This Now?

Computation and psychology are both at points of incredible transition right now. On the technological side, we are generating more data than ever, but tools to process this data are also starting to become more accessible to the average person. There are notably five main changes that have occurred with computing technologies that have enabled data to become so big. First, as we have discussed, there have been drastic improvements in cheap, large data storage in small form factors. This means that the average person has on their phone or computer tens of thousands of files, documents, images, videos, and pieces of software. This also means there are places where we can easily save our big datasets. Because these storage devices are getting smaller, we can have large amounts of storage in small devices, like phones. Meanwhile, cloud storage allows people to maintain massive amounts of data that they can access with a multitude of devices. Second, there have been major improvements in faster and cheaper processing power, such as the explosion in parallel processing graphics chips. For example, the average processor in a consumer computer can make about 150 billion calculations per second. This allows us to analyze big datasets relatively rapidly, and even in real time as we acquire it. Third, sensor technology and speed has also improved – most people have high-quality cameras in their phones, and may have devices (like fitness trackers) that can record movement, heart rate, elevation, skin conductance, and other measures. This allows us to obtain big physiological data, which can reveal underlying information about one's cognitive state (Chapter 9). Fourth, the wide spread of high-speed internet both in homes and out in the world is allowing more people to form communities, creating large amounts of data generated by people's interactions online. Fifth, our algorithms are getting better and smarter – we are able to compress data more efficiently and analyze data more effectively with tools like artificial

intelligence. The combination of these five computational improvements has led to an explosion of data produced by and accessible to the average person.

Big Data is also more important than ever for psychology research. Psychology has always been a multidisciplinary field, straddling social science and biological science programs at many universities. For example, psychology has clear links to neuroscience and experimentation, but also has implications for therapeutic practice and philosophy. However, recently, psychology as a field has begun to undergo a transition, with greater emphasis focused on experiments and complex analyses. Many exciting discoveries are coming about thanks to Big Data innovations, such as online experiments, artificial intelligence, and rich physiological data. With these innovations, researchers have been able to revisit classical psychological questions with a Big Data lens that allows them to assess their applicability across more diverse samples or make computational models that can predict people's behaviors. These innovations have also been saving psychologists a lot of time – making it faster to collect and analyze data. These changes go hand in hand with a new global scientific community that is developing, based around sharing data and code openly, in reaction to a "replication crisis" that emerged around unreplicable findings in small-scale experiments (see Section 2.1.3). So now is the perfect time to learn about these changes in psychology, to ride its waves as it moves into these new approaches.

Discussion Question

What topics relating to Big Data and psychology are you particularly excited to learn about in this book, and in your class?

1.6 How to Use This Book

As we just discussed, psychology is changing. If you want to go into psychological research for your career, professors and laboratories are now increasingly looking for candidates with experience in programming and statistics. Outside of academia, many jobs after college geared toward psychology majors – such as user experience design or data science jobs – also require these skills. For those of you wanting to practice clinical psychology, counseling, or go into education, it is still helpful to be up-to-date with the latest research and techniques (e.g., how is artificial intelligence changing the diagnosis of neuropsychological disorders?). And I would argue that some of these topics we will discuss in this book can help improve your daily life. I know for me personally, I've coded data scraping tools to find the best flights for a vacation, used generative AI to make a personalized storybook for my kids, or analyzed my fitness tracker data to get a sense of whether a diet is working. Knowing what is possible with data can change how you look at and use data in your daily life. In this book, we will also touch on some of the hot-button topics that have erupted in the news and the legal sphere as a result of Big Data – how do we deal with AI-generated fake information? How do we navigate the privacy risks created by the data recorded in many mobile apps and websites?

It can be intimidating jumping into learning about data and computer programming if this is your first foray into the topic. My number one goal is to demystify these topics and make you comfortable talking about them and thinking about them. As a student starting along this journey, it can feel like there's a big gap between you and your image of a computer scientist who may have been hacking computers since they were in elementary school. It can feel like you just aren't meant to be someone who codes or does complex math. But really these thoughts are a part of a mystical (but inaccurate!) aura that has surrounded computation. I'd liken computer programming to something like learning a foreign language or training for your first 5 km run. Most of the time, your goal isn't to become completely fluent or a record-breaking marathon runner. Usually, it's that you find these skills useful and enjoy the process of getting there. You also usually aren't worried about how you compare to the pros – you don't feel bad comparing your Spanish skills to those of a native speaker (and often they are impressed that you are trying!), or feel bad watching Olympics runners beat your time. In the same way, a seasoned software developer won't be quizzing you on the latest Python functions. You will also find that gaining these skills can enrich your daily life – you can now navigate a little around Madrid with your newfound Spanish skills, or be able to run to catch a bus without getting winded. Similarly here, you'll have moments where you may wish you could do something on your computer in an automated way and then realize there may be a way to use your skills learned here to do that!

At the same time, this book is not going to teach you programming or statistics from the ground up. It's the first step in learning the lingo and giving you the lay of the land, so you can then decide where you want to do a deep dive in future classes or explorations (e.g., do I want to study more neuroscience? Or web design? Or graph theory?). The online resources with this book will provide some stepping stones for doing these deep dives. With that foreign language metaphor, this book is your travel guidebook to help you decide where you want to study abroad. Then once you've picked a country, you can start focusing on learning its language. With this book, I want you to become fluent in the topics of new technologies being widely used in psychological research. I want you to have an increased level of agency over your own data and how it is used by companies and researchers. And, I want you to practice thinking creatively about psychological research questions and how we can answer them.

This book can be read from front to back or you can skip around sections as needed. In these first two chapters, I introduce what Big Data (Chapter 1) and small data (Chapter 2) are and how they compare to each other. Then for the rest of the first half of the book, I will give you the building blocks for running Big Data studies – looking at the participants (Chapter 3), the stimuli (Chapter 4), and the experiments themselves (Chapter 5). Once we are armed with our Big Data, we can then analyze it using artificial intelligence (Chapter 6) or human crowdsourcing (Chapter 7). In the latter half of the book, we will delve into different topics that are changing as a result of Big Data, and so these chapters are a bit more standalone. We will talk about software developments with apps and games (Chapter 8), as well as hardware innovations and physiological sensing (Chapter 9). We will talk about Big Data in neuroscience (Chapter 10), language and natural language processing (Chapter 11), and wrap up with social interactions and graph theory (Chapter 12).

With the exception of this chapter, each chapter will end with four key sections. In "Big Ethical Questions," we will talk about the ethical implications of the topic discussed in the chapter. These topics are sure to spark interesting discussion, especially because the ethical implications of these new methods are still being actively addressed in science and society. This section is then followed by a section on "Applications of the Chapter." While most of this book takes the framework of theory-driven psychology – where we are conducting experiments for the sake of understanding the mind, not creating a product – in these sections, we will discuss how the chapter's topics can be applied to impact the real world. Each chapter then ends with a Chapter Summary that reminds the reader of the major points, and Further Reading which suggests further sources to explore if you are interested in going beyond the pages of this book. There will be discussion questions laced throughout the chapters, as well as a sample homework assignment at the end of each chapter. The companion Teacher's Guide will include additional discussion questions, exercises, and demonstrations for each topic.

Importantly, data, computation, and the internet are always changing. While this book is written at a static point in time (2022–2025!), there is an accompanying online resource (www.cambridge.org/bainbridge) that will be updated as technologies change in the world. If you read anything in this book that seems outdated, check out the online resource to see if there is a new version of that information. The online resource also has interactive demos and programming tools to let you learn more about programming and test out online experiments.

With that, let's proceed to Chapter 2 to discuss what "small data" is and how that differs from Big Data in the context of psychological research.

CHAPTER SUMMARY

In this chapter, we introduced the concepts of Big Data, psychology, and how now is the perfect time to study them and their interactions.

1. Here, we define Big Data as unstructured, naturalistic human data requiring complex analytical methods.
2. We define psychology as the empirical study of the mind, brain, and behavior. This book mainly focuses on quantitative experiment-based psychology.
3. With major improvements in our technological capabilities over the past few decades and changes in the landscape of psychology, now is the perfect time to study how Big Data can be used in psychology research.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter.

- **Read about and watch an original video describing the world's first hard drive, developed by IBM in 1956:** Seeley, C. (2014, October 28). History snapshot: 1956 – the world's first moving head hard disk drive. Data Clinic Ltd. News. www.dataclinic.co.uk/history-snapshot-1956-the-worlds-first-moving-head-hard-disk-drive

- **A review of how realism in our studies can actually show differences in the brain:** Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*, 506–519.

ASSIGNMENT

The purpose of this assignment is to learn more about your experience with Big Data and provide you a sense of the data you generate.

Total Points: 50

1. **Fill out the class survey.** Your professor will provide a link. (20 points)
   Let's look at how much data you are generating just from your phone!
2. Locate where your phone describes your storage usage. **Answer:**
   a. How much storage are you using for images? (1 point)
   b. How much storage are you using for videos? (1 point)
   c. How much storage are you using for music? (1 point)
   d. How much storage are you using for apps/applications? (1 point)
3. Let's get a rough estimate of how much data you are generating a day with your phone camera.
   a. Add together your answers from 2a and 2b and **report that number here**. (2 points)
   b. Get an estimate of how long you have had your phone – find the date of the first photo you took. Then search on Google "how many days between [that date] and today" and it should return you the number of days. **Report that date of the first photo and the number of days since then.** (2 points)
   c. Divide your answer in 3a by your answer in 3b and **report that number here.** This tells you about how much data you are generating with your camera per day. (4 points)
   d. **How many bytes of data is that?** (2 points)
   e. One byte is the amount of data used to type one character (e.g., "A"). A novel contains about 500,000 characters. **How many books worth of data is that?** You're likely creating the equivalent of books of information a day! (3 points)
4. Let's see how long you are using your phone for.
   a. First, guess: **How much screen time do you think you use a day?** (2 points)
   Now, locate where your phone describes your screen time usage (this might be under a "Screen Time" setting or a "Digital Wellbeing" setting).
   b. **On average, how much screen time do you actually use a day? How does this compare to your guess?** (4 points)
   c. Based on your screen time report, **on average how much screen time do you spend on social media a week?** (2 points)
   d. You use on average 500 MB of data per hour by browsing social media. **How many GB (or MB) of social media data are you viewing per week?** (5 points)

As you can see, we interact with massive amounts of data in our daily lives!

REFERENCES

Adobe, Inc. (2013, August 1). Did you know that Photoshop 3.0 was the last version of Adobe Photoshop to be sold on the floppy disc? Facebook. www.facebook.com/photo?fbid = 10151614431968871&set = a.468676338870

Bainbridge, W. A., & Baker, C. I. (2022). Multidimensional memory topography in the medial parietal cortex identified from neuroimaging of thousands of daily memory videos. *Nature Communications*, *13*(1), 6508.

Duarte, F. (2024). Amount of data created daily. Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day

Gannon, L. (2002). A critique of evolutionary psychology. *Psychology, Evolution & Gender*, *4*, 173–218.

Kumar, S. (2015). *Fundamental limits to Moore's law*. arXiv:1511.05956.

Moore, G. E. (1965). Cramming more components into integrated circuits. *Electronics*, *38*(8).

Waldrop, M. M. (2016, February 9). The chips are down for Moore's law. *Nature* [news feature]. www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338

2 What Is Small Data?

Introduction

In order to learn about Big Data, you first need to understand its counterpoint, "small data." Small data isn't often called this, because data from most psychology studies fits under this umbrella, and many times its scale can suit our purposes just fine. Thus, a definition of small data would be any data that isn't Big Data. While it is incredibly common, solely using small data severely limits the takeaways we can get from psychological research. In this chapter, I will discuss the limitations of small data, as well as the limitations of Big Data. You will see how the two can work in synthesis to pinpoint the rich phenomena occurring in our minds and brains.

Specifically, first to understand the benefits we gain from Big Data, we will go through a few example small data experiments (Section 2.1.1) and see what they are lacking (Section 2.1.2 and Section 2.1.3) and how they can be made bigger in scale (Section 2.1.4). We will then discuss some limitations to Big Data experiments (Section 2.2), including new issues they introduce (Section 2.2.1) and the limitations that will always be present with any study (Section 2.2.2). We will then discuss how experiments can be dichotomized into being hypothesis-driven or data-driven (Section 2.3), as well as how Big Data studies can be characterized as deep or wide (Section 2.4). We will discuss the ethical issues that can come about from the multiple analyses run with Big Data (Section 2.5). We will then discuss applications of the topics in the chapter, such as examples of famous data-driven discoveries (Section 2.6.1) and medical applications of deep and wide data (Section 2.6.2).

2.1 Turning a Small Data Experiment into a Big Data Experiment

Let us first begin with an example of a small data experiment and think about how we can make it bigger and broader.

2.1.1 A Case Study

There is a famous effect in psychology called the **own-age effect** (Anastasi & Rhodes, 2005), where people tend to remember faces close to themselves in age better than faces farther in age.

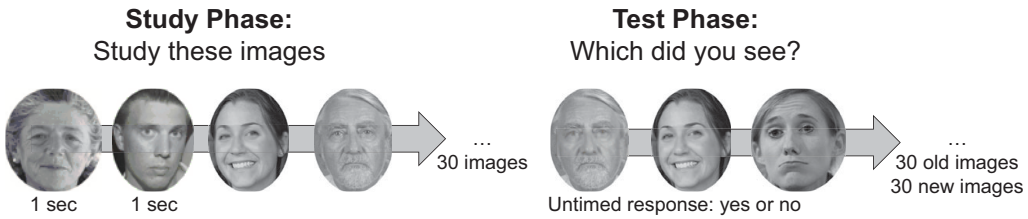


Figure 2.1 The experimental methods for our own-age effect experiment. We have participants first study thirty face images for 1 second at a time. Half of the face images are from older adults and half are from younger adults. We then test them where we show them thirty of the face images they saw, randomly mixed up with thirty new face images they didn't see. For each face they have to respond if they saw it before ("yes") or not ("no"). Our main research question is whether participants have different levels of memory accuracy based on the match between their own age and the age of the face images.

You may have experienced this before, where you may have an easier time recognizing your classmates than professors on campus. (There are many memory effects driven by the similarity of a face to your own – there is also famously an own-race effect; Chiroro & Valentine, 1995). Let's say we are in a traditional psychology lab, and we are running an experiment to test the own-age effect. Our experimental methods look something like what is in Figure 2.1.

The idea is to recruit fellow psychology students on campus and run them through a face memory test on the computer in the lab (as most psychology experiments are done!). In that memory test, we will show a series of thirty faces (like in Figure 2.1), where half are college-aged, while the other half are older adults. We will then test to see if there is a significant difference in memory for those two groups of faces. After running twenty participants, we find a significant effect – indeed the own-age effect holds true!

Discussion Question

What prevents us from generalizing these results to saying that the own-age effect occurs for all observers and all faces?

2.1.2 What Does a Small Data Experiment Miss?

The previous case study was an example of a typical psychology experiment. However, there are many aspects of it that prevent us from generalizing to all observers and all faces. Specifically, the participants, the stimuli (the face images), and the experiment itself are all constrained and artificial in some way.

Small Participants: First, the number and scale of the participants is "small" – can we really make generalizations about humans as a whole from an experiment run with twenty students at a specific university? For example, would these effects replicate for people who are frequently exposed to faces of other ages – like in cultures where young adults tend to live with older

generations? In Chapter 3, we discuss more about the problems with using small college samples in a large proportion of psychology studies, and what we can do about it in the field.

Small Stimuli: Second, the images are also very small in scale. Like the issue we have with participants, can thirty faces really capture the rich variance of human faces out in the world? If you look at the paradigm (Figure 2.1), all these faces are very homogenous. They are all front-facing white people of moderate attractiveness with an oval cropped around their face so you cannot see much of their hair or clothing. This can sometimes be intentional – researchers often want to control for factors they're not interested in, so that those cannot be alternate explanations of their effect. For example, you don't want to think there is an effect of age on memory when it's actually the clothes the models are wearing (maybe clothes from a few decades ago are more memorable than clothes from today!). But, too much control will limit our ability to make generalizations across different lighting, viewpoints, and facial expressions – it doesn't let us make confident predictions about memory out in the real world. And, by only doing research on constrained demographics (e.g., all white people), we aren't studying the rich variation in human experience. In order to generalize to the real world, we need images that better capture the diversity we observe in that world. In Chapter 4, we talk more about how to think about and create more representative stimulus sets.

Small Experiment: Even in just the way they are conducted, experiments are much smaller in scale than the real world. They don't capture what it's like to meet a moving, emotive, multisensory human being, and try to encode them into memory. Experiments tend to be brief (usually 30 minutes to an hour) and constrained to a two-dimensional computer screen, with a few seconds to see each face. This is not at all what it's like to meet a face in reality – you see them out situated in the real world, and you may spend hours interacting with them. Perhaps the dynamic, moving aspects of a face can contribute to your memory for that face, and that would be completely ignored by the experiment. Or perhaps seeing faces in the three-dimensional world is fundamentally different for memory than seeing them on a flat, two-dimensional screen in an experiment. (Although seeing faces in two dimensions may be becoming more natural, as virtual meetings are becoming more common.) Also, because faces are so dynamic, it's unlikely in the real world that you will ever see the exact same view of a face again; you can never take the exact same photograph twice. The second time you see a person, their facial muscles will be engaged in a slightly different way, the lighting will hit their face differently, or they may have a slightly different glow to them. This is completely different from an experiment which shows you the exact same photograph twice.

2.1.3 A Second Case Study and the Replication Crisis

Let's examine another sample experiment. Within the field of social psychology, one phenomenon that has been proposed is the phenomenon of **social priming**. The idea with social priming is that when you are made to think of a social category, you automatically think about related behaviors and stereotypes and start to subtly behave in a similar way. This was first demonstrated in a study by Bargh and colleagues (1996) across a series of experiments. For example, in one experiment, thirty psychology class undergraduates from New York University were asked to complete a task where they had to take a set of five words and create a grammatically

correct four-word sentence as quickly as possible. They did this for thirty sentences in total. Unbeknownst to those participants, half of them received words specifically related to being elderly – *old*, *grey*, *sentimental*, *bingo*, *wrinkle* – while the other half received neutral words. The idea was that the elderly-related words might prime them to think about elderly individuals and act in a similar way. An experimenter then secretly timed how long it took participants to exit the hallway leaving the testing room. The researchers found that participants primed to think about being elderly had a significantly slower walking speed (8.3 seconds to travel the hallway) than participants given a neutral prime (7.3 seconds), confirming their hypothesis. Participants reported not being aware of this elderly manipulation, or a change in their behavior, suggesting these social priming effects could happen unconsciously.

Discussion Question

What factors in this experiment might prevent us from generalizing more broadly?

This experiment uses a relatively small number of participants (fifteen in the elderly prime condition) and stimuli (thirty), making the robustness of the effect unclear (though the original experimenters do actually replicate this effect in a second thirty-participant experiment). The participants come from a very specific sample – psychology undergraduates in the New York area – who are unrepresentative of the world population. The words are also not validated as conjuring an image of "the elderly" in an objective way. The study does a fairly good job at using a naturalistic task (i.e., measuring walking time). However, there could be modern improvements on how it is measured, rather than relying on an experimenter's timing skills, which could introduce a subtle bias that accounts for the 1-second difference between conditions. As a result of these critiques and others, Doyen and colleagues (2012) ran a larger-scale replication of the experiment. They ran 120 participants (albeit also from a fairly specific sample – Belgian French-speaking undergraduate students). They used elderly word stimuli that were first confirmed by a separate set of eighty participants as representing old age. The experimenters then used infrared sensors to precisely measure the amount of time it took to traverse the hallway. With this "bigger" experiment, researchers found no difference in walking speed between their two participant conditions.

Around the same time (in the early 2010s), many psychological findings were unsuccessfully replicated. Researchers were failing to find clear evidence for many social psychological phenomena that had become well-accepted – in addition to social priming, there was now evidence against ideas like ego depletion (the idea that willpower is a finite resource) and power posing (that standing in a certain way will increase your confidence) among others. This launched a "**replication crisis**" across the field of psychology bringing into question the quality of the research in the field. One event that ignited this crisis was when a paper was published in one of the most revered social psychology journals (*Journal of Personality and Social Psychology*) claiming evidence that people can see the future (use "precognition"; Bem, 2011). Researchers realized that a combination of poor research practices as well as publication pressures in the field (see Section 4.4) was overinflating the reporting of supposed "results" across many papers.

At this major breaking point, hundreds of researchers as part of the Open Science Framework launched an effort to attempt to reproduce a hundred findings in psychology. Shockingly, only thirty-six were successfully replicated (Open Science Collaboration, 2015). This served as a reality check for psychologists – we need to run experiments with larger samples, more generalizable experiments, and better statistical measures. We also often should run multiple replication experiments to confirm our effects really hold, and aren't just occurring due to chance.

2.1.4 Making an Experiment Big

Even if I have convinced you that traditional psychology experiments are often unnatural simulations of the real world, how can we improve upon this? How can we make our experiments "bigger"?

> **Discussion Question**
>
> What would a Big Data version of the example face experiment look like? What would you change?

We need to think about how we can improve upon the three points mentioned earlier: the participants, the stimuli, and the experiment. For the participants, can we recruit more people, and more widely? In Chapter 5, we will talk about how to conduct online experiments, which lets you reach thousands of people, with more diversity than the average college campus. For the stimuli, we can also strive to collect image sets that are larger, more natural, and more diverse (refer to Chapter 4 to learn how we can do that!). For making the experiment more naturalistic, there is the difficult balance of wanting scientific control but also generalizability. If you want to keep it as a computerized task, what if you test people on memory for a face across different photographs of that person, rather than memory for a specific image? (It turns out recognizing an unfamiliar person from different photographs is a very difficult task! See Jenkins et al., 2011). New technologies are also making it easier to conduct experiments in more dynamic, three-dimensional environments like virtual reality, or even out in the real world (e.g., Snow & Culham, 2021; Võ et al., 2019). If we are able to expand our experiments out in these three ways, then we have a more generalizable study in these three ways as well. We can know things about face memory across a wider range of observers and faces being observed, and we can try to make predictions about behavior out in the real world.

For example, in our lab, we were curious about people's memory for faces more generally than the own-age effect. So, we generated a large database with demographics matching the United States (see Chapter 4 for more information). We then had over 800 diverse individuals engage in a face memory experiment online (Bainbridge et al., 2013). In this experiment, people viewed a stream of face images and pressed a key when they recognized a repeat from earlier (called a **continuous recognition task**) – a little like the experience of walking through a crowd and recognizing some people as you pass them. We also ran a version of the experiment where we tested people's memory for faces across different viewpoints and facial expressions,

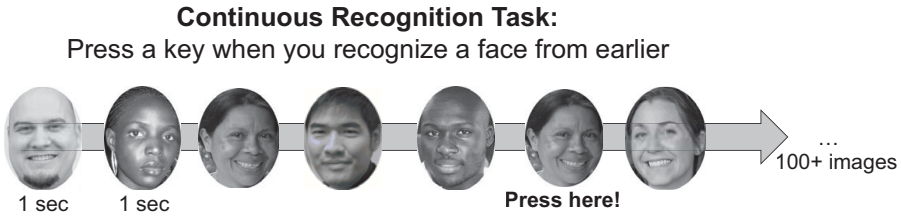


Figure 2.2 The experimental methods for our more "Big Data" face experiment. People still view face images for 1 second at a time. However, now they are continuously seeing images and indicating their memory as part of a "continuous recognition task," akin to the experience of walking through a stream of people and sometimes recognizing someone. We now also have a much more diverse range of faces, so we can test many phenomena, such as the own-age effect or the own-race effect, as well as measure the intrinsic memorability of a given face image. We would also test this task with many diverse participants, so we can look at more generalizable effects of the viewer, too.

not just face images (Bainbridge, 2017). So, our experiments looked a little more like Figure 2.2. What we found in the end was that there are certain faces that are remembered very well by most people, and some faces that are remembered very poorly. In other words, faces have an intrinsic *memorability*. We have recently extended these findings to images more generally, and tested them out in the real world – for example, discovering that we can even make predictions about what paintings people will remember in a freeform visit to an art museum (Davis & Bainbridge, 2023).

2.2 Limitations of Big Data

It may seem obvious that we'd want to aspire for diverse, generalizable experiments as described. However, there are some obstacles presented by Big Data experiments that are important to consider.

2.2.1 Problems with Big Experiments

One of the major cons of Big Data-style experiments is that they can introduce noise into our data. First, sometimes the images can vary too much. If each image is different along many dimensions (age, gender, attractiveness, facial expression, lighting, angle, hairstyle, eyebrow width, etc.), then how can we pinpoint which specific dimension is causing the effect we're studying? And maybe many (or even a majority!) of these dimensions might be things we don't care about, like lighting. Similarly, if our participants vary too much, we can have the same problem – everyone may act in a unique way that prevents us from finding generalizable effects. And participants may vary in ways that are not interesting to us – for example, in running an online experiment, what if the participant's screen monitor resolution, or what they ate for breakfast that specific day, or the length of their fingers influenced how quickly they pressed keys on a task? Small experiments let us directly test our effect of interest, without having all these

extraneous factors in the way because we control for them as much as possible. So, sometimes, we have to design our Big experiments in a way that they're not *too* big. At the same time, Big Data experiments let us simultaneously look at the contributions of many factors – for example, we can analyze how gender, age, and race all contribute to face memory, at the same time.

2.2.2 Imperfect Experiments

Even when we want to run Big Data experiments, there will always be limitations that make it impossible to run a perfect experiment. There will inevitably still be biases with the stimuli and participants in the experiment, because you cannot make everyone in the world participate in the experiment, or make everyone agree to be photographed as a face in the experiment. There are some factors limiting who can be a participant in your experiment – participants must have some familiarity with how to read a computer, and they have to have free time and interest in participating over other things they could be doing. Similarly, only a subset of people would be okay being photographed for the study, and any set of natural photographs will likely have an over-representation of happy or neutral facial expressions (over angry or sad).

There are also some other practical limitations with Big Data. Sometimes the data is so big that we are limited by the processing power, storage, or internet speeds that support us saving and analyzing the data. For example, one person's MRI brain data can take up 1 terabyte of space, which is more than the amount of space many computers come with (in 2025). It can also take half a day to download this data for just one person! So, it can be difficult to analyze data from hundreds, let alone dozens, of participants. Large-scale experiments can also be very costly with time and money. Using the same example of an MRI experiment (which is on the upper end of what psychology experiments cost), one participant usually lies in the scanner for about 2 hours, and it will cost the researchers around US$1,000 to the scanner center for that time. So, an experiment with 100 participants would end up costing $100,000 and take 200 hours of the researchers' time to just collect the data. We are also still limited in our analytical techniques for Big Data. When dealing with very big, naturalistic data, we often don't look at just a single measure or statistic. But, at the same time, our statistical tools and artificial intelligence are not yet able to fully interpret natural human behavior. For example, let's say we wanted to look at face memory in the real world, and recorded participants' view as they navigate through a party, using some sort of head-mounted camera. It's not clear how we would analyze these data – how to turn the conversations with people, the amount of time looking at them, the thoughts related to them, etc. – into numbers in order to make conclusions about what influences someone's memory of a person. So, our analytical techniques are limited (and in fact, they are dependent on the study of psychology to guide us on how to analyze such complex human behaviors).

2.3 Hypothesis-Driven versus Data-Driven Research

A majority of psychology experiments can be characterized as **hypothesis-driven research**. These are experiments where the researchers have one or a few key research questions. They also tend

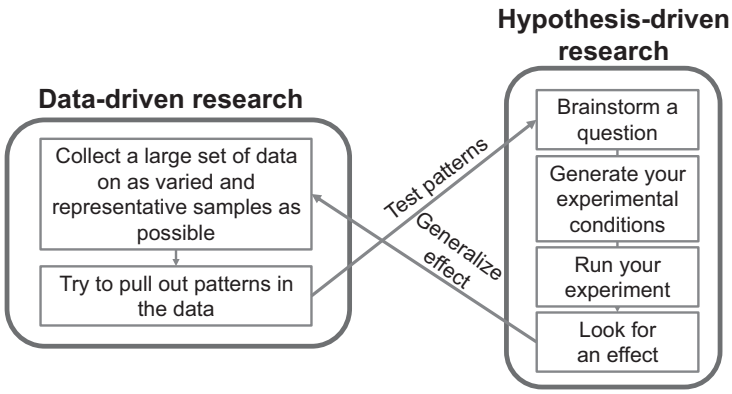


Figure 2.3 The series of steps you take when conducting data-driven research and hypothesis-driven research, and the way in which they interact. In data-driven research, you collect a large, representative set of data and identify patterns in the data. You can then take those patterns and test them in controlled, hypothesis-driven experiments to determine the specific mechanisms driving the effect. To conduct hypothesis-driven research, you would brainstorm your research question, create experimental conditions to answer that question, and run your experiment. If you identify an effect in a hypothesis-driven study, then it can be helpful to test whether the effect generalizes by running a more naturalistic, data-driven study.

to have a clear idea of the different alternatives of the results (in other words, **hypotheses** about the results) and what that means for the bigger picture question. The general pipeline for running a hypothesis-driven study follows the flow of the right side of Figure 2.3. One of the most important first steps is deciding on the main question. So, taking the first case study example we covered, let's say your question is: "Are people better at remembering faces close to them in age than faces far from them in age?" Your two experimental hypotheses would be something like: 1) yes, faces closer in age are remembered better, or 2) no, there is no difference in memory based on age of the face. You will then design your experiment, selecting experimental conditions that let you pinpoint that question. **Experimental conditions** are the different ways you divide up your experiment to answer your research question. For example, for this study, we may have different participant groups (older and younger people) as well as different stimulus sets (older and younger faces). So, we would have four conditions: young participants viewing young faces, young participants viewing older faces, older participants viewing young faces, and older participants viewing older faces. In experimental design terms, participant age is a **between-subjects** factor, because the condition changes from subject to subject (in other words, between the subjects). In contrast, face image age is a **within-subjects** factor, because within a single subject, they will see faces from both older and younger face conditions. Importantly, the conditions and factors that we intentionally change or control as experimenters (like the gender of participants and faces) are called **independent variables (IVs)**. It's because these are variables that we set, and so they are not dependent on other things in the experiment – they stand on their own. Once you have designed your experiment with its conditions, you then run the experiment (i.e., having your participants do a memory task with

the images you choose). The data that comes out of your experiment are **dependent variables** (DV), because they are dependent on the IVs and the experiment that you run. Finally, you can then run statistical tests on your data to directly answer the research question you set out to test. For a statistical test, you will typically define your **statistical hypotheses** – these are subtly different from experimental hypotheses in that they specify the different possible results of your statistical test. With most traditional statistical tests (called parametric statistics; see more discussion in Chapter 3), you would have a **null hypothesis** reflecting the hypothesis that there is no effect for a given statistical comparison. In this case, the null hypothesis would be that there is no difference in memory performance across our different participant-image age conditions. You would also define an **alternate hypothesis** reflecting the hypothesis that there is a significant effect. Here, the alternate hypothesis would be that there *is* a difference in memory across these conditions. So, we run our test and see which hypothesis we have more evidence for – can we reject this null hypothesis or fail to do so? Specifically, you would directly test whether memory performance (your DV) is higher for same-age conditions (young participants / faces and old participants / faces) than different-age conditions. And, voila, you have your answer! (So far, the research seems to point to yes: Chiroro & Valentine, 1995).

This hypothesis-centric way of thinking often goes hand in hand with small data research, because you want to run experiments that specifically target your question of interest. You *can* run these in a Big Data way (e.g., with thousands of participants and thousands of images), but in the end, the only factor you'll want to differ is the one you're interested in (age), and you'll want other factors like race, gender, attractiveness, etc. to be the same across your different conditions. This is because you want to be certain that your experimental manipulation has a direct influence on the effect you observe (in other words, you want your IV to directly influence your DV). If you do not control for other factors, you risk having **confounds** that could explain the link between your IVs and DVs. A confound is another variable that can account for the relationship between your IV(s) and DV(s) that you are not intentionally manipulating. In other words, they can be alternate explanations for your results. When designing an experiment, you want to make sure you can avoid these confounds, or at least can take them into account in some way. For example, let's say we look at monthly data across a year and find a correlation between ice cream sales and drowning deaths: when ice cream is popular, drowning is more common (Mumford & Anjum, 2013). Does this imply that ice cream causes people to drown? That would be ridiculous (and unfortunate)!

Discussion Question

What are some confounds that could explain a relationship between ice cream sales and drowning deaths?

One of the big confounds here is weather! During the summertime when the weather is nice, people want to go out swimming. They certainly have a much higher risk of drowning if they're swimming in a lake than if they're staying home bundled up by a fire during the wintertime. During the summer, people are also probably going out to buy ice cream to cool off from the hot weather, so you would see both high ice cream sales and increased drownings. On the other

hand, during the winter, the dessert of choice might be something more like a slice of apple pie or a mug of hot chocolate, and you would be unlikely to be out swimming. So, we would say that weather here is a confound in this relationship between drownings and ice cream sales. When designing a hypothesis-based study, it's important to keep in mind all potential confounds, and sometimes it can be impossible to control for all of them.

The counterpoint to hypothesis-driven research is **data-driven research**. The idea is that you collect tons of data, trying to gain samples that are as representative and varied as possible (Figure 2.3), and this is the approach often taken when using Big Data. Then, you use statistical methods to try and pull out patterns from the data that can help answer some questions. For example, you could collect memory test data for a wide range of faces and participants, and then see if people generally tend to remember faces closer to their own age best. This type of research is also sometimes called **exploratory research**, because you can explore around the data and look for different patterns without necessarily having a hypothesis from the beginning. What's great about data-driven research is that you generate big datasets that can help answer many questions. So, these databases can be multiuse – you could look at questions about memory and age, but also memory and attractiveness, or memory and face shape. You can also look at how these different factors all work together to form the big picture (i.e., what combinations of features influence the memorability of a face?). However, because these data tend to be collected without controlling much in the experiments, you run a higher risk of having more confounds that can explain your effects. However, because you often aren't relying on a single research question or statistical test, one confound may be less impactful on the use of the dataset overall. However, if you are not careful, data-driven research has some other big risks that can result in low-quality science (see Section 2.5 on data fishing).

Overall, there are pros and cons to both hypothesis-driven and data-driven research, with the key points summarized in Table 2.1. But ultimately, science benefits most when we do both, because they serve as an interconnected loop. Data-driven studies let us discover new, unexpected effects that can emerge from large or naturalistic data. Hypothesis-driven studies then let us take these effects and pinpoint the reasons behind these effects and link them to broader theories about human cognition. A lot of psychology reasonably takes the hypothesis-driven approach as a result. But to broaden our perspective on what questions to ask and what blinders we might have on in the field, I would argue the data-driven approach is just as important – and this is what will be the focus of this textbook.

Table 2.1 Comparison of the pros and cons of hypothesis-driven and data-driven research

Hypothesis-driven research	Data-driven research
Usually small data	Usually Big Data
Can isolate specific effects	Can be more naturalistic
Pulling out data based on theory-driven questions	Pulling out questions based on diverse datasets
Larger effect sizes	Statistical significance more likely
Have to be wary of confounds	Have to be wary of data fishing

2.4 Deep Data versus Wide Data

When designing a Big Data study, there are two dimensions along which it can be Big – deep or wide. A **deep data** study is one where you collect lots of data for a smaller number of individuals (so you're getting a deep look at a few people). Some examples include sensor data like a fitness watch recording frequently and over long periods of time (Chapter 9), or software-based data recording lots of samples over time (like on your phone; Chapter 8). There are also some experiments that focus on running the same participants many times over a series of sessions. These repeated measurements can give rich information about individuals, letting us look at the influence on cognition of things that vary like the time of day, attention fluctuations over time, and complex behaviors. One issue with deep data, though, is that it can risk being invasive of participants' privacy because you're learning so much about specific people. For example, if you look at one person's measurements from a fitness watch over months, you would learn all about their sleeping and exercise habits. It can also be tedious for participants to collect and provide all this data, especially if it's a study where they have to come in for multiple sessions. So, it can be hard to recruit participants for deep studies.

A **wide data** study is one where you collect relatively small amounts of data from a large number of participants at a single time. Some examples include data from an online experiment across thousands of people, or a snapshot of rich data from an app or piece of software at a single point (like all the tweets for a topic on a single day). What's great about wide data is that it can give diverse information across a large, representative sample of people. However, you often cannot capture very complex behaviors that vary over time or an interaction.

Some researchers characterize Big Data along three dimensions – being deep, wide, and long. In this case, deep data would still involve multiple measurements (like we see with sensor data). However, wide data now instead reflects collecting data across multiple variables or measures (e.g., with a battery of questionnaires). Finally, long data would involve collecting data from many people. Regardless of how you characterize the dimensions of Big Data, studies can be any combination of deep, wide, and long (e.g., collecting tons of data from many people), although this can be hard to achieve, so scientists may need to pick one dimension along which to specialize. When looking at a study, it's worth thinking how it falls on these different measures of size.

2.5 Big Ethical Questions

When you have a huge experiment, it can be easy to go fishing around for significant effects. This is something called **data fishing**, **data dredging**, or p-hacking. Since you have so much data, it seems like one of the benefits is that you should be able to look at many different effects in your data at once, right? Well, yes, you can do this to some degree, but you also need to think about how statistical tests are conducted.

For a majority of standard statistical tests that compare your data to a distribution (like t-tests, ANOVAs, regressions, etc.), you aim to estimate a ***p*-value**. What this *p*-value represents is the probability you would observe something as extreme as your results if the null hypothesis

were true. Recall that the null hypothesis reflects the hypothesis that there is no effect in your data. However, even if this null hypothesis were true, there is noise in our measurements and people's behaviors, so we would still sometimes observe a difference between our conditions "by chance." When we run a statistical test, we are looking at what the distribution of data would look like if the null hypothesis were true (and there was no effect). We are then seeing where our observed data falls in this distribution – how likely is it to occur given this null distribution? We calculate our *p*-value as the proportion of data in the null distribution that is equal to or more extreme than our observed value. So, for example, a *p*-value of 0.03 indicates there's only a 3 percent chance you would happen to observe these results just by random chance. That seems pretty low, and as a field, we've currently accepted a cut-off of 5 percent ($p < 0.05$) to be how we determine what we'll take to be a significant finding or not. Another way to phrase this is that in our field, we have accepted a 5 percent **false positive rate.** This is the rate of falsely saying something shows an effect when it does not. You may have heard this term used to refer to the rate of a medical test falsely saying you have an illness when you do not – same idea.

While this 5 percent chance of a false positive seems rare, when you're dealing with Big Data, you're doing many statistical tests – maybe hundreds of tests (e.g., for a psychological battery), or even up to hundreds of thousands of tests (e.g., for the case of MRI brain data). And so, in the realm of hundreds of thousands of tests, even with this seemingly strict false positive rate, we will get about 5,000 tests (5 percent of 100,000) that come out as "significant" just by chance! So we need to think carefully about how we define significance with data-driven research since we are doing many tests, inflating the chance that we find a false positive in at least one of these tests. In order to circumvent these issues, we do something called **multiple comparisons correction**, which is a group of statistical methods that let us calculate an adjusted *p*-value threshold for our study that takes into account the many tests that we are doing. While we won't go into these methods in detail, some example methods include **Bonferroni correction** and **false discovery rate correction**. Bonferroni correction corrects for the rate of false positives across all of the statistical tests that you perform. It does this by calculating an adjusted threshold for "significance" (called the alpha level, or α), based on the number of tests you are running. So, if you run ten tests, your alpha level would be $p < 0.005$ instead of $p < 0.05$. False discovery rate correction is a more liberal method that corrects for the proportion of false positives among all results initially labeled as significant – in other words, calculating an alpha level so that we are okay with 5 percent of our discoveries being false positives.

Let's look at an example study that ran many statistical tests. In Moore et al.'s 2006 study "Thongs, flip flops, and unintended pregnancy," the researchers wanted to investigate if there were some lifestyle factors that were related to unintended pregnancies. They conducted a 50+ question survey with 126 women who were currently or recently pregnant, and conducted 362 statistical tests to analyze their data. They found some surprising results: unintended pregnancy was associated with preferences for yoga, beaches, thongs, Doritos, contact lens, and text messaging. They also found that baby boys were more common if mothers preferred trucks, beef, and boys, while baby girls were more common if mothers preferred cars, chicken, and girls. So does that mean if you see your friend texting their friends during some beach yoga while tucking into a bag of Doritos, that you should encourage them to be vigilant with their

contraception? No, because if you think back to what we just discussed with running many statistical tests, we would expect about 18 of their 362 tests to come out as significant just by chance given our *p*-value threshold of 0.05! So even if there are no meaningful relationships between any of these factors and unintended pregnancy, just because of random noise in measurement, participant behaviors, and the environment, it would be unsurprising to find some relationships that come out as statistically "significant" but aren't real.

So one of the risks of Big Data is that it's easy to run many, many statistical tests until you find something significant. Because of all of the rich data you have, it's tempting to test many different questions. There are also big pressures in the scientific world to publish significant results, so researchers may be tempted to focus on these "significant" results without accounting for the number of statistical tests that they're running. In other words, you may be tempted to fish around for a result in your big sea of data. There are three main ways to make sure you are not doing data fishing with your own data. One way is to perform multiple comparisons correction across all the tests you run. A second way is to decide your analyses and hypotheses in advance before seeing your data (called preregistration; see Section 4.4) – so in other words, running a combined hypothesis- and data-driven study. A third way is to replicate any findings you discover across multiple datasets, analyses, and/or labs to be sure that what you're finding is real, rather than something that emerges just by chance.

There is also the question of the **effect sizes** of the results you end up finding. While we often care about statistical significance in our data, we also care about how strong the effects are in the differences that we are measuring. Effect size is often quantified as the proportion of the **signal** of interest to the level of **noise**. So, for example, for a t-test, the measure of effect size is the difference between the conditions' averages (the "signal"), divided by the standard deviation pooled across the two conditions (the "noise"). You can have a significant effect that's a weak effect or a strong effect. For example, let's say you're looking at whether an intervention in the classroom results in a difference in test scores on a test with a maximum of a hundred points. You could get a significant effect where the intervention results in a one-point increase. While this would mean the intervention likely worked (because the effect is significant), it didn't work very well (the effect is weak)! If the intervention instead resulted in a significant thirty-point increase, we would say this is a strong effect! And, if you have a nonsignificant effect with a thirty-point increase, that would mean our results aren't strong enough (e.g., there may be too much noise), so we cannot be confident that this thirty-point increase didn't just happen by chance. Because of its large sample sizes, Big Data can be prone to identifying significant but weak effects – effects that would only be detectable when you have thousands of people. Therefore, even if you find a significant result, consider what the result means. If it is a meaningful, strong psychological effect, ideally we would even see it occur at the level of a smaller sample, and even at the level of the individual.

2.6 Applications of the Chapter

In this chapter, we discussed characterizing research in a few different ways – for example, hypothesis-driven versus data-driven or deep versus wide data. These ways of thinking about

data have promoted discoveries beyond the field of psychology, and have guided recent advancements in the medical field.

2.6.1 Data-Driven Discoveries

We mentioned how data-driven research can result in new questions or effects that we may not have conceived of if we only stuck to pre-existing theories and hypothesis-driven experiments. There are in fact many exciting scientific discoveries that came about thanks to people trying out many different things. One of the most famous examples in psychology is the discovery by Professors David Hubel and Torsten Wiesel that led to their Nobel Prize win in 1981. They were trying to see what information was coded in neurons in the occipital lobe (the early visual regions of the brain), by recording directly from cats' brains while showing them different images (see Chapter 10 to learn more about neuronal recording). They were struggling to find any specific image that would cause these neurons to spike. Back in the day, they were using a slide projector, and suddenly when they were swapping out the slides, they heard the neuron they were recording from start to fire. After playing with the slide, they discovered that this neuron was sensitive to the edge of the slide when it was shown at a specific angle. This led to our current understanding of the visual system in the brain, where neurons are sensitive to edges oriented at specific angles. You may have heard about similar fortuitous "eureka!" moments throughout the sciences. As the classic example, around 246 BC, Greek scientist Archimedes realized how to calculate volume and density while taking a bath, and purportedly ran through the streets shouting "eureka!" In 1820, Dr. Hans Christian Oersted noticed a compass move when he placed it near an early battery he was creating – resulting in the discovery that electrical currents generate a magnetic field. Percy Spencer invented the microwave in 1946 when he noticed the chocolate in his pocket melted when he was testing out a new vacuum tube. These discoveries may have never happened without the experimenters just trying out different things. Big Data can encourage such exploration, which can lead to exciting discoveries.

2.6.2 Medical Applications of Deep and Wide Research

Deep and wide methods have had some wide-reaching applications in the clinical realm. Deep data has helped form the field of **precision medicine**, where healthcare workers can make honed, personalized predictions of health outcomes based on genetics, environment, lifestyle, and sensor measures. Big Data lets us create **predictive models** that take these different factors and then make guesses about outcomes for a single person (see Chapter 6). Precision medicine goes hand in hand with preventative medicine and telehealth, where people can wear sensors and use apps to remotely track and communicate symptoms before they develop into a full-blown condition. For example, researchers are working on apps to help identify early stages of Alzheimer's disease (Konig et al., 2018) and apps to help elderly individuals develop memory strategies (Martin et al., 2022).

Wide data is key in letting us learn about diseases: It lets one see global trends in disease, identify rare groups at particular risk, and find hidden links to a cause or cure (Heggie, 2019). Wide data played an important role in identifying the symptoms early on in the COVID-19

pandemic, when it wasn't clear what symptoms were being caused by the virus. Researchers conducted a large-scale wide symptom study where anyone could enter their symptoms online, and they ended up receiving information from 4.4 million participants (Menni et al., 2020). As a result, they were one of the first groups to identify a loss of smell or taste as one of the symptoms of COVID-19. They also found some other interesting trends: for example, for the first wave of the pandemic, one out of twenty participants had symptoms that lasted more than 8 weeks, and longer COVID was correlated with having more different symptoms in the first week. They also found that during the pandemic lockdowns, 20 percent of participants had an increase in alcohol consumption, and an average weight gain of 4.6 lbs.

CHAPTER SUMMARY

In this chapter, we discussed the small data experiments traditionally utilized in psychology research and showed how they compare to the Big Data experiments that are becoming increasingly popular. Here are some of the main takeaways:

1. The key to making a small data experiment "Big" is expanding its participants, stimuli, and paradigms to be more naturalistic and representative of the real world. However, you can almost never make a perfectly representative experiment.
2. There are different benefits to hypothesis-driven research versus data-driven research, and both are necessary for the progression of psychology as an innovative and rigorous field.
3. Big Data can be characterized by two key dimensions – its depth (how many measures you collect per individual) and its width (how many individuals you record from).
4. With the large amount of data you can get from a Big Data study, we must be cautious of not "fishing" for effects without accounting for all of the statistical tests that we are conducting.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter.

- **Learn about how Big Data is causing big strides in the understanding of disease:** Heggie, J. (2019, January 8). How can big data beat disease? *National Geographic*. https://tinyurl.com/ykkabh53
- **A cautionary tale on how too many statistical tests can lead to effects that may not be real:** Moore, R. P., Galvin, S. L., & Imseis, H. M. (2006). Thongs, flip-flops, and unintended pregnancy: The seduction of $p < 0.05$. *MAHEC Online Journal of Research*, *1*, 1.
- **Dive deeper into the statistics used to correct for multiple comparisons:** Lindquist, M. A., & Meija, A. (2015). Zen and the art of multiple comparisons. *Psychosomatic Medicine*, *77*, 114–125.

ASSIGNMENT

The purpose of this assignment is to get you thinking about Big Data and how to build out Big Data experiments. Please submit your response in a way so that it is clear what questions and sub-questions you are responding to.

Total Points: 50

1. Pick two psychology papers describing an experiment on a topic that sounds interesting to you. (Do not use a review paper.) They can come from either:
   i) a psychology class you are currently taking, or took in the past;
   ii) a lab you are currently working in; or
   iii) any "Open Access" articles from the most recent year of the journal *Psychological Science* (https://journals.sagepub.com/home/pss).
   Please choose at least one paper that you would consider a "small data" experiment. Provide the citation (titles, authors, year, journal) and abstract of the two papers here: (2 points)

   We will now look at these papers with the frameworks we discussed in this chapter.
2. **For paper 1, answer the following questions**. If the study includes multiple experiments, answer for the first or main experiment:
   a. Is this a **"small data"** or a **"Big Data"** experiment? How do you know? How small/big is the sample size (number of participants)? How small/big is the experiment itself (e.g., number of conditions, stimuli, outcome measures)? (4 points)
   b. Is this a **hypothesis-driven** or a **data-driven** experiment? How can you tell? (3 points)
      i. If this is a hypothesis-driven experiment, what is their hypothesis?
      ii. If this is a data-driven experiment, what new hypotheses come out of their data? How did they avoid p-hacking / data fishing?
   c. Is the data **deep** or **wide** (or both)? How do you know? (2 points)
3. **For paper 1, we will do some more brainstorming on Big Data. (Note that the next part of the question has two options – for a given paper, answer for either small data or Big Data, not both.)**

   If this is a "small data" experiment, we will think up how to make it into a Big Data experiment. Answer these questions:
   a. How **naturalistic** versus **artificial** is their experiment? What are ways in which the stimuli, experiment, or participants are not *representative* of reality? What are ways in which they are? (3 points)
   b. How can we improve the **representativeness** of the study? Use your creativity to brainstorm how you would make this into a "Big Data" experiment. How would you change the experimental paradigm, participant recruiting, the stimuli, or the measurement techniques to capture bigger, more diverse, more naturalistic, and more representative data? (5 points)
   c. What **limitations** could you envision with these changes? These can be limitations in terms of feasibility/practicality (e.g., how much time or money does your change add)? In what ways is your version still not fully representative? (3 points)

   If this is a "Big Data" experiment, we will see how it improves upon small data studies. Answer these questions:
   a. What would the **"small data"** version of the experiment have looked like? (4 points)
   b. **Why** did the experimenters decide to take this "Big Data" approach? What innovations did they apply to make it "Big Data"? (4 points)

c. What **limitations** still exist with their approach? In what ways are the data still not fully representative of real people / images / cognitive processes? What additional improvements could you envision, and how feasible are they (e.g., how much time or money does your change add?) (3 points)

4. Answer questions 2 and 3 for **paper 2** below. (20 points)
5. What are some ideas implemented by paper 1 that could be useful for paper 2 in making their experiments more representative in terms of participants, stimuli, or paradigms? Similarly, what are some ideas implemented by paper 2 that could be useful for paper 1? (5 points)

REFERENCES

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*, 1043–1047.

Bainbridge, W. A. (2017). The memorability of people: Intrinsic memorability across transformations of a person's face. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 706–716.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*, 1323.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 48, 879–894.

Davis, T., & Bainbridge, W. A. (2023). Memory for artwork is predictable. *Proceedings of the National Academy of Sciences USA*, *12*, e2302389120.

Doyen, S., Klein, O., Phcion, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Heggie, J. (2019, January 8). How can big data beat disease? *National Geographic.* https://tinyurl.com/ykkabh53

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323.

Konig, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., & Robert, P. H. (2018). Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, *15*, 120–129.

Martin, C. B., Hong, B., Newsome, R. N., Savel, K., Meade, M. E., Xia, A., Honey, C. J., & Barense, M. D. (2022). A smartphone intervention that enhances real-world memory and promotes differentiation of hippocampal activity in older adults. *Proceedings of the National Academy of Sciences USA*, *119*, e2214285119.

Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., El-Sayed Moustafa, J. S., Visconti, A., Hysi, P., Bowyer, R. C. E., Mangino, M., Falchi, M., Wolf, J., Ourselin, S., Chan, A. T., Steves, C. J., & Spector, T. D. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine*, *26*, 1037–1040.

Moore, R. P., Galvin, S. L., & Imseis, H. M. (2006). Thongs, flip-flops, and unintended pregnancy: The seduction of $p < 0.05$. *MAHEC Online Journal of Research*, *1*, 1.

Mumford, S., & Anjum, R. L. (2013, November 15). Correlation is not causation. Oxford University Press blog. https://blog.oup.com/2013/11/correlation-is-not-causation

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*, 506–519.

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210.

3 Big Participant Samples

Introduction

Many students first gain their interest in experimental psychology as a guinea pig – by volunteering in an experiment on campus. It can be exciting to apply the knowledge you learn from class to guess the main manipulation in a behavioral experiment, or see inside your brain after an MRI study. It can also be a great way to earn money (at least I know I funded all my college vacations by being a "professional subject"!), or you may be required to participate in studies as part of passing a course. This built-in participant sample is also fantastic for researchers: They have a pool of eager young students who are readily available during weekday hours, attending class right next to your laboratory, and often attentive and enthusiastic about each new experiment. However, while it seems like running university experiments with college students is a win-win situation for both the researchers and the students, this convenient choice actually causes concerning damage to the field of psychology as a whole.

In this chapter, we will cover the problems with current norms in the participants we recruit for psychology experiments and how to solve some of these problems by taking a Big Data approach. First, we will go through how small data studies recruit their participants (Section 3.1). We will then talk about how the average college sample differs from adults worldwide (Section 3.2), individuals from smaller societies (Section 3.3), other industrialized nations (Section 3.4), and others even within the same country (Section 3.5). The issues boil down to a difference between our sample and population (Section 3.6), and we will discuss how we can move toward more representative groups using Big Data (Section 3.7). However, we will never be able to make a perfect sample (Section 3.8), and sometimes we may want to intentionally restrict the people we recruit (Section 3.9). The chapter will finish with a look at the big ethical questions surrounding participant recruitment (Section 3.10) and imbalances in the demographics of psychology researchers themselves (Section 3.11).

Big Data in the Psychological Sciences

Wilma A. Bainbridge

Big Data in the Psychological Sciences

Cutting-edge computational tools like artificial intelligence, data scraping, and online experiments are leading to new discoveries about the human mind. However, these new methods can be intimidating to many students. This textbook demonstrates how Big Data is transforming the field of psychology, in an approachable and engaging way that is geared toward undergraduate students without any computational training. Each chapter covers a hot topic, such as social networks, smart devices, mobile apps, and computational linguistics. Students are introduced to the types of Big Data one can collect, the methods for analyzing such data, and the psychological theories we can address. Each chapter also includes discussion of real-world applications and ethical issues. Supplementary resources include an instructor manual with assignment questions and sample answers, figures and tables, and varied resources for students such as interactive class exercises, experiment demos, articles, and tools.

Wilma A. Bainbridge is an associate professor in the Department of Psychology at the University of Chicago. She has won the Association for Psychological Sciences Rising Stars Award (2023), an Alfred P. Sloan Fellowship in Neuroscience (2024), and the American Psychological Association's Distinguished Scientific Award for Early Career Contributions to Psychology (2025). Her research has garnered attention from outlets such as CNN, *Vox*, and *Wired*. She has previously edited two books on vision and memory, and her "Big Data in Psychology" class has earned a Curricular Innovation Award from the University of Chicago.

"From social media to sensors to AI, this book offers a brilliant tour of how the Big Data revolution is reshaping psychology. Accessible, inspiring, and grounded in real research problems, it walks students through everything from hands-on skills like web scraping, to big-picture theory testing, and even thoughtful discussions of ethics – all presented with incredible clarity by one of the field's most inspiring new voices."

Timothy Brady, University of California San Diego

"Exceptionally timely and comprehensive, Bainbridge's textbook deserves a place in every curriculum for behavioral methods. The chapters – enhanced with interactive features and thought-provoking ethical questions – are so engaging that they make me want to teach the course. And whether or not you work with Big Data, this is essential reading for all."

Marvin M. Chun, Yale University

"Combining conceptual depth and accessible writing, Bainbridge offers a timely contribution with a comprehensive overview of the field, covering definitions of big data in psychology and expertly navigating its key sources, methods, and analytical approaches. It addresses both foundational topics, such as neuroimaging tools and statistical techniques, as well as emerging and contemporary discussions, including natural language processing, the development of large language models, and their applications in psychological research. It will resonate with a wide audience, from curious undergraduates to seasoned researchers looking to deepen their understanding of big data and its potential to reshape the psychological sciences."

Nemanja Vaci, University of Sheffield

Big Data in the Psychological Sciences

Wilma A. Bainbridge
University of Chicago

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009343589

DOI: 10.1017/9781009343602

When citing this work, please include a reference to the DOI 10.1017/9781009343602

First published 2026

Cover image: FrankRamspott / DigitalVision Vectors / Getty Images.

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-34358-9 Hardback
ISBN 978-1-009-34357-2 Paperback

Additional resources for this publication at www.cambridge.org/bainbridge

Big Data in the Psychological Sciences

Cutting-edge computational tools like artificial intelligence, data scraping, and online experiments are leading to new discoveries about the human mind. However, these new methods can be intimidating to many students. This textbook demonstrates how Big Data is transforming the field of psychology, in an approachable and engaging way that is geared toward undergraduate students without any computational training. Each chapter covers a hot topic, such as social networks, smart devices, mobile apps, and computational linguistics. Students are introduced to the types of Big Data one can collect, the methods for analyzing such data, and the psychological theories we can address. Each chapter also includes discussion of real-world applications and ethical issues. Supplementary resources include an instructor manual with assignment questions and sample answers, figures and tables, and varied resources for students such as interactive class exercises, experiment demos, articles, and tools.

Wilma A. Bainbridge is an associate professor in the Department of Psychology at the University of Chicago. She has won the Association for Psychological Sciences Rising Stars Award (2023), an Alfred P. Sloan Fellowship in Neuroscience (2024), and the American Psychological Association's Distinguished Scientific Award for Early Career Contributions to Psychology (2025). Her research has garnered attention from outlets such as CNN, *Vox*, and *Wired*. She has previously edited two books on vision and memory, and her "Big Data in Psychology" class has earned a Curricular Innovation Award from the University of Chicago.

"From social media to sensors to AI, this book offers a brilliant tour of how the Big Data revolution is reshaping psychology. Accessible, inspiring, and grounded in real research problems, it walks students through everything from hands-on skills like web scraping, to big-picture theory testing, and even thoughtful discussions of ethics – all presented with incredible clarity by one of the field's most inspiring new voices."

Timothy Brady, University of California San Diego

"Exceptionally timely and comprehensive, Bainbridge's textbook deserves a place in every curriculum for behavioral methods. The chapters – enhanced with interactive features and thought-provoking ethical questions – are so engaging that they make me want to teach the course. And whether or not you work with Big Data, this is essential reading for all."

Marvin M. Chun, Yale University

"Combining conceptual depth and accessible writing, Bainbridge offers a timely contribution with a comprehensive overview of the field, covering definitions of big data in psychology and expertly navigating its key sources, methods, and analytical approaches. It addresses both foundational topics, such as neuroimaging tools and statistical techniques, as well as emerging and contemporary discussions, including natural language processing, the development of large language models, and their applications in psychological research. It will resonate with a wide audience, from curious undergraduates to seasoned researchers looking to deepen their understanding of big data and its potential to reshape the psychological sciences."

Nemanja Vaci, University of Sheffield

Big Data in the Psychological Sciences

Wilma A. Bainbridge
University of Chicago

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009343589

DOI: 10.1017/9781009343602

When citing this work, please include a reference to the DOI 10.1017/9781009343602

First published 2026

Cover image: FrankRamspott / DigitalVision Vectors / Getty Images.

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-34358-9 Hardback
ISBN 978-1-009-34357-2 Paperback

Additional resources for this publication at www.cambridge.org/bainbridge

Thank you to the "RAV4" – Robert, Ally, and Vicky – for being a loving and supportive family! It's hard to believe I began this book when still pregnant with Ally and Vicky and am finishing it as they are running around and chatting our ears off.

Thank you to my mom Erika, dad William, and sister Connie – finally I get to write book dedications for you rather than the other way around!

And thank you to the wonderful Brain Bridge Lab and my department at the University of Chicago – your support has really helped me flourish and think Big.

Brief Contents

Detailed Contents

Preface

Learn how to see. Realize that everything connects to everything else.

Leonardo da Vinci (1452–1519)

We live in a world where we are all constantly generating data – in our interactions with our phones, social media apps, games, websites, fitness trackers, and more. This data is commonly referred to as "Big Data" because its scale is so large that it cannot be analyzed manually. Such Big Data serves as a useful means to understand human cognition – showing us how people see, feel, respond, remember, interact, and make decisions with these different tools. We can also look at these cognitive processes across different groups of people – across countries, cultures, ages, and experiences – as well as across species. As a result, Big Data ways of thinking and analysis have become incredibly important tools to psychologists, across fields. Psychologists are now running online experiments that can gather data from thousands of participants, running machine learning models that can decode patterns from thousands of datapoints, or analyzing brain data from thousands of subregions.

As a result, psychology as a field is at a major transition point. Familiarity with advanced statistical analyses and computer programming is becoming increasingly essential to keep up with the state of the art. However, the idea of wrangling Big Data can be incredibly daunting to people entering the field, especially given that most undergraduate psychology curricula do not require computational or advanced statistical coursework. The main goal of this textbook is to make these new directions in Big Data accessible and meaningful to any psychology student – without the need of training in computer science or statistics. By reading this textbook, you'll gain basic fluency and familiarity with the important topics in the field, so you can decide what topics you want to pursue more deeply. Students who are already familiar with computational methods will learn ways in which these methods can be applied to answer a myriad of psychological questions. As a result, the book will lightly touch upon a wide range of topics, including experimental design, web programming, data scraping, artificial intelligence, different methods in brain imaging, computational linguistics, network science, wearables, user interface design, crowdsourcing, and representative sampling.

To my knowledge, this is the first undergraduate textbook on Big Data in psychology. It was inspired by a course I created in Spring 2020 as a new assistant professor, and I've seen these sorts of courses start to grow in the last few years. Because this is such a new topic, this textbook and course is really for almost anyone. Familiarity with psychology is helpful (e.g., how experiments are run and what are some of the key topics of inquiry), and at some points I will bring up simple statistical concepts (e.g., *p*-values), although knowledge there is not

required. Each chapter focuses on a different angle of how Big Data interfaces with psychology, and includes sections on ethical questions related to the topic and its real-world applicability. Each section also includes thought-provoking questions that can be discussed as a class and an assignment that's relatively open-ended and should engage the students in thinking deeply about that topic. The chapters can be covered in pretty much any order, but the book is generally divided into two parts: 1) how to rethink psychology experiments from a Big Data angle (Chapters 1–7), and 2) various sources of Big Data to enrich the study of psychology (Chapters 8–12).

In conjunction with the Big Data theme, I also want to make this course follow the principles I preach in terms of modernizing psychological research. As a result, this book is paired with an interactive online resource (www.cambridge.org/bainbridge) that includes videos, demonstrations, links, and additional resources that will be constantly updated. This way, you will still have access to the latest developments in the field even after the publication of this book. I also maintain a public data repository on the Open Science Framework of Big Data student projects that came out of the course that I teach at the University of Chicago (https://osf.io/hz843), and I am happy to link to such a repository from anyone else using this book.

Now go forth, and think big!

1 What Is Big Data?

Introduction

The amount of data generated every day is insane. Each day, we create approximately 403 million terabytes of data (or 403 exabytes) (Duarte, 2024). This is about how much data can be stored by 4 billion phones (those of about half the world population) – and just in one day! In that same day, about 300 billion emails are sent, 8.5 billion searches are done on Google, 1.6 billion swipes are made on Tinder, 1.4 billion hours of video are streamed, and $638 million is spent on Amazon. You as an individual are contributing to a lot of this growing collection of data. As you commute to your classes, your map app tracks your movement behavior and may take note of any specific locations you visit. Your phone or watch tracks your steps and sleep patterns. As you look up web pages on your phone, these pages track your browsing and click behavior. And as you scroll through and post on social media, these apps track how you engage with posts, through measures like viewing time and click behavior. We are constantly surrounded by and creating Big Data. This Big Data can be messy and tricky to sift through, but within it are potential insights about the human mind waiting to be discovered.

In this introductory chapter, we will establish definitions of the central themes of this book, to guide you as you read the rest of the book. First, we will talk more about how data has changed in the last few decades (Section 1.1) and then provide a definition of Big Data (Section 1.2). We will then define Psychology within the context of this book (Section 1.3). With these two definitions in hand, we will discuss how Big Data and Psychology interact (Section 1.4) and why now is the perfect time to study this interaction (Section 1.5). Finally, we will wrap up this chapter with a guide on how to use this book and its online resources (Section 1.6).

1.1 Moore's Law

Our data has gotten so *Big* thanks to the exponential growth in processing and storage power over the past handful of decades. This is reflected by **Moore's Law**, which was proposed by Intel cofounder Gordon Moore in 1965 (Moore, 1965). This law predicts that the number of transistors (one of the key components in computer chips) that can be packed into a given

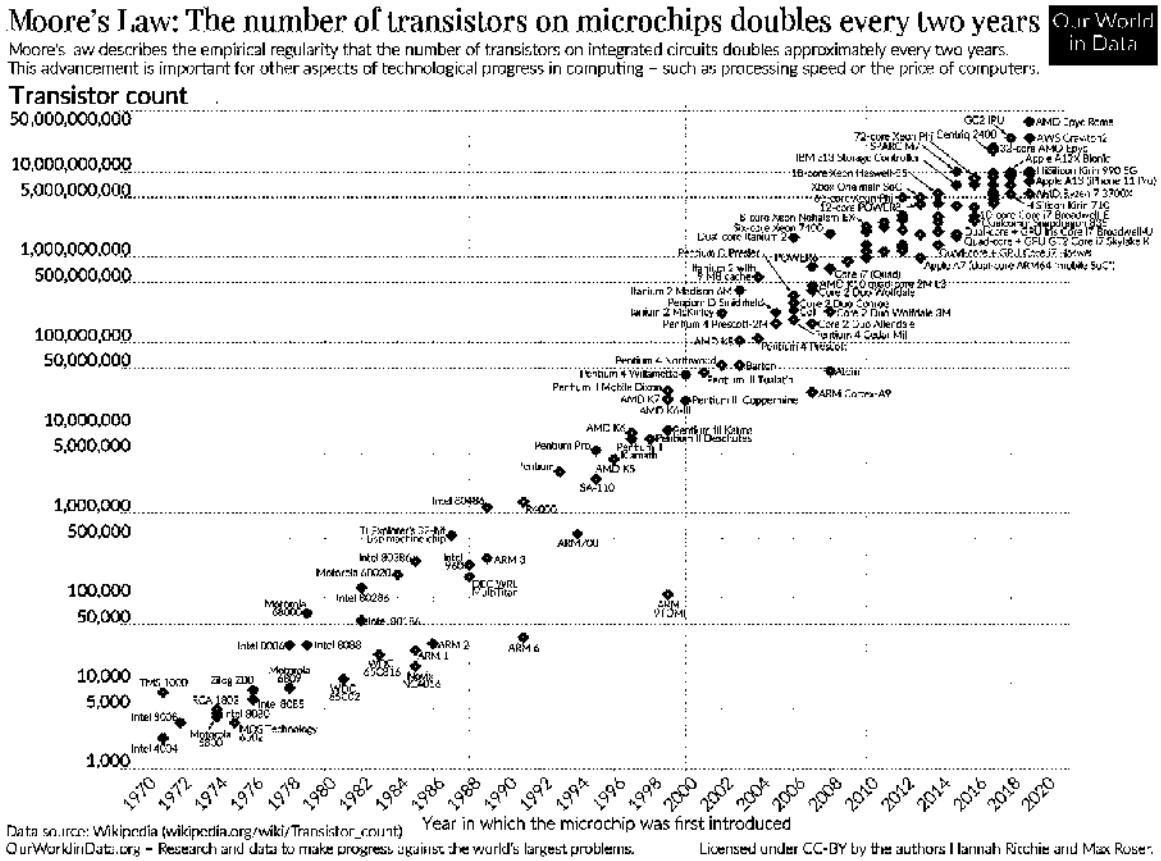


Figure 1.1 A depiction of Moore's law, showing it still holds in 2020. Moore's law posed in 1965 predicts that the number of transistors we can fit in a circuit will double every two years – resulting in exponential growth in our computing capabilities. Note that the y-axis here is an exponential scale (1,000 and 5,000 at the bottom are spaced as closely as 10 trillion and 50 trillion at the top), so indeed, we are keeping up with this law!

unit of space will double roughly every two years. Remarkably, this prediction of exponential increase in computing power has held true for 60 years (see Figure 1.1), although some scientists forecast that we will reach the limit of feasibility within the next few years (Kumar, 2015; Waldrop, 2016). We can feel the effects of Moore's law by looking at how the size of storage devices has drastically changed over our lifetimes.

Discussion Question

What did the size of data (e.g., devices, files, performance) look like when you were a child versus now? Does it feel like there has been exponential growth in that time? What sorts of innovations enabled that growth?

To understand what makes a set of data Big Data, let's first discuss how data is measured. The building blocks of the data and processes in our computers are 0s (off) and 1s (on), and a single digit is called a **bit**. Because of this 1/0 building block, instead of data being measured in our decimal base-10 system, data sizes are measured in **binary**, or a base-2 system, where the only digits possible are 0 and 1. When you want to count in higher numbers in binary beyond 1, you use additional digits (that are still limited to 0 and 1). So the numbers 0, 1, 2, 3, 4, and 5 in decimal are represented as 00, 01, 10, 11, 100, and 101 in binary. While these building blocks seem simple, they can combine to form the complex data we interact with on our computers – just as letters can combine to create the complexities of language. Because of the binary system, powers of 2 end up being important to the measurement of data. A set of eight bits ($2\times2\times2$) is called a **byte**. A byte can be used to represent a single character of text. For example, in the most common character encoding standard for computing called **ASCII (American Standard Code for Information Interchange)**, the letter A is represented by the byte containing the bits 0100 0001, while a space is represented by 0010 0000. Above the level of the byte, the naming of the counting system resembles that of the metric system. A set of 1,024 bytes is called a kilobyte (KB), like how 1,000 meters is a kilometer (but because we are operating in binary, it is a multiple of 2, or 2^{10}). A set of 1,024 kilobytes is called a megabyte (MB). A set of 1,024 megabytes is called a gigabyte (GB). After that, we have terabytes (TB), petabytes (PB), exabytes, and zettabytes. So, for example there are 8,000,000 bits (1s or 0s) in one megabyte of data. When we talk about data transfer speeds (like how fast your internet is), the measures tend to be in bits per second (instead of bytes per second). So early internet modems would have a download speed of 28.8 Kbps, or around 28,800 bits per second.

The rapid change in computing sizes is quite drastic when we look at the history of data storage across personal computing (Figure 1.2). Back in 1956, IMB shipped its first hard drive. It was the size of two refrigerators and could hold 5 MB of data – the equivalent of about one song. In the 1970s, some consumers were starting to get their own computers, and the most common way to store and transfer files was through floppy disks. These could only hold about 100 KB in early years, and 1.44 MB in later years, the equivalent of a few text documents or pictures. However, this medium became so ubiquitous that many pieces of software still use an icon of a floppy disk as their "save" icon. Once software became more advanced, users needed more and more of these disks – for example it took seven floppy disks to install an early version of Adobe Photoshop (Adobe, Inc., 2013).

In the late 1980s, a more advanced data storage method emerged – the CD-ROM (compact disc read-only memory). These could hold as much as 900 MB – about one-third of a movie. However, as their "read-only" name implies, these needed special devices called CD burners to write data to the CD-ROM, and most CDs could not be rewritten once data was saved onto it. In the mid-1990s, we moved onto DVDs (digital video disks), which could store closer to 5 GB (about 1–2 movies) but faced similar shortcomings as CD-ROMs. The early 2000s saw the first USB (universal serial bus) flash drives, which were smaller and more convenient but could only store about 10 MB at first. The early 2000s also saw the explosion of the internet, and **cloud storage** – saving data to online servers – started to grow. This really took off as internet speeds became faster, and websites emerged dedicated to hosting large amounts of data – YouTube for video started in 2005, Dropbox for files started in 2007, and Flickr for photos started in 2004.

Figure 1.2 Photos of various types of older data storage. **(a)** The shipping of IBM's first hard drive, holding 5 MB and taking up the size of two refrigerators. **(b)** A 3.5-inch floppy disk, the main external data storage format in the 1980s and 1990s. **(c)** A CD-ROM inserted into a laptop's CD drive. These were commonly used in the 1990s and 2000s for data storage and were the main medium for holding music albums.
Source: (a) Michael de Groot / Flickr. (b) Pablo Jeffs Munizaga – Fototrekking / Moment / Getty Images.
(c) EThamPhoto / The Image Bank / Getty Images.

In the 2010s, storage amounts and data transfer speeds increased by many magnitudes. Personal computers could hold close to a TB of data, and mobile technologies emerged enabling people to generate vast amounts of data through the photos and videos they're taking. In the 2020s, higher mobile data speeds and faster personal computer processors are allowing us to interact with vast amounts of data and at faster rates – supporting growths in avenues like online gaming, video streaming, and real-time artificial intelligence (AI) on board many of our systems.

1.2 How Do We Define Big Data?

Owing to this constant growth in technology, the definition of Big Data is a moving target. In the 1990s, even a low-resolution video would be considered Big Data, while in the 2020s, Big Data is more in the order of libraries of thousands of movies, representing petabytes of data (or even more). A fundamental aspect of what makes data qualify as Big Data is that it is so large that we cannot process it by hand – we can't manually input, clean, or analyze the data. This is data that is also often so large that we cannot process it with basic programs on our computer (e.g., Microsoft Excel), but usually have to use code or bespoke tools. For our purposes of studying human cognition, Big Data tends to be more naturalistic – recorded from real people or dynamic behaviors – so it may be unstructured or more noise- or error-prone compared to smaller datasets. Data can be Big for several different reasons. It may have very high temporal sampling – for example, taking a measurement every handful of milliseconds. It may have high spatial sampling instead – for example, collecting data across all the intersections in a city. It may have high participant sampling – collecting data from a large and diverse set of people. Or, it may have high stimulus sampling – capturing data from lots of sources (e.g., images, videos, news articles, products) or tasks. All of these encompass examples of Big Data.

So here, we will loosely define Big Data as: *unstructured, naturalistic human data requiring complex analytical methods*. For some of the exercises and examples we discuss in the book, we may not use the massive amounts of space or computing power that traditionally make up Big Data. But the concepts that you learn here should translate to bigger sets.

1.3 How Do We Define Psychology?

It is also important that we define what psychology is within the framework of this book. Broadly, psychology is *the empirical study of the mind, brain, and behavior*. For the vast majority of this book, we will focus on a more quantitative and research-based approach to psychology, where psychologists conduct experiments that aim to provide broad insight, using falsifiable questions and hypotheses. This is in contrast with some psychologists who use more *qualitative* approaches, like revealing new insights using interviews or observations, or using therapeutic techniques like talk therapy to help improve mental health. A central aspect of the psychology we will discuss here is that it is composed of research questions that are testable and falsifiable. **Falsifiable questions** are those where you can obtain evidence to prove that question wrong. For example, one active area of debate is the degree to which we can falsify questions in **evolutionary psychology** – the study of the mind, brain, and behavior through the perspective of evolution (Gannon, 2002). We cannot go back in time and run experiments on our ancestors. We also cannot measure how much of people's current behaviors are a result of evolution versus more recent societal norms. There are some creative scientific methods to test evolutionary hypotheses by looking at other animal species or running computational models. However, we will generally avoid tackling unfalsifiable topics in the current book.

We are also interested in questions that are **generalizable** – that give us insight into the thinking of a group of people, and can allow us to make predictions in the future about different events. For example, a question like "how did people feel about Argentina winning the 2022 World Cup?" is a question that measures emotions and behavior, but is so highly

specific that it does not really teach us about the human mind. Thus, we would not characterize this as a psychological question. A more generalizable psychological question might be something like "how does sentiment in social interactions change directly after major sporting events?" Sometimes when dealing with Big Data, we may accidentally take an overly narrow scope, due to the data that we have available (like data from a specific app, Chapter 8). But we should always try to focus on a big-picture question about the mind, and any limitations to the data we are collecting to answer this question.

There are many different branches to the field of psychology. When you think of a "psychologist," your mind may first go to **clinical or abnormal psychology** – the study of atypical behavior and mental health, with the goals of diagnosis and treatment. Closely related is counseling psychology, which is the practice of helping people through therapy and counseling. While this book will discuss some research on abnormal psychology through the lens of understanding the underlying roots of an impairment, we will not discuss in depth therapies or treatments of individuals. **Industrial psychology** is the study of the mind within the workplace, and how to optimize people's effectiveness at work. As this field is more applications-focused, we will not discuss this at length in this book, nor other more applied fields like forensic psychology, school psychology, and health psychology. The major focus of this book will be **cognitive psychology**, the branch of psychology dedicated to the scientific study of our internal mental processes. This encompasses a broad range of processes, including sensation, perception, action, memory, reading, speaking, emotion, decision-making, morality, imagination, and others. In fact, a "psychologist" can be a researcher with a laboratory that runs experiments to study these processes (this is the type of psychologist I am). Related to cognitive psychology, we will sometimes discuss the brain, through a lens of **neuropsychology** – looking at how the brain and mind interact. We will also bring up many examples from **developmental psychology** – the study of the development of the mind across the lifespan (from infants through aging) – and **social psychology** – the study of the interactions of multiple minds.

1.4 How Do Big Data and Psychology Interact?

A large proportion of Big Data out in the world just *happens* – you record a video and post it on social media, and now there are several new megabytes of data on your phone, on a server belonging to that social media site, and being downloaded to other people's phones. In this way, much of Big Data is just passively accumulated as we perform tasks with our phones, computers, and the internet. Another major slice of Big Data is being actively collected by companies, where they are testing how they can improve your experience, how you navigate their app, and how they can improve engagement and purchasing. However, the data being generated out in the world also serves as incredibly rich records of human behavior that can give insight into questions on almost any topic of psychology.

Discussion Question

What are some ways you can envision Big Data might be changing the types of questions we can ask or answer in psychology?

An important skill for you to nurture will be in identifying these intersections of Big Data and psychology. What is a psychological question you want to answer, and how could Big Data answer that question? Could there be a preexisting dataset out there that answers the question for you, or could Big Data help you collect that data in some way? For example, a few years ago, I was curious how older memories (2+ year-old memories) might be represented in the brain. This is hard to test in the laboratory because I would need to have participants study some images and then come back two years later. But then it dawned on me that people are constantly capturing their memories on social media, dating back to many years prior. So, I collaborated with the app 1 Second Everyday to recruit users who had recorded years of their memories, and then I scanned their brains while they viewed these older memories. Long story short, we found patterns in the brain reflective of the age of a memory (Bainbridge & Baker, 2022; see Section 8.4.2). As you look through data in your daily life, think to yourself – what does this reflect about the human mind and can it show us something new? And, are there ways in which Big Data technologies are influencing how we think or interact? For example, an active area of current research is how social media may be impacting feelings of isolation and depression (Section 12.3.2). Overall, a major part of this class will be thinking creatively and with an open mind on how we can use data to answer questions.

1.5 Why Study This Now?

Computation and psychology are both at points of incredible transition right now. On the technological side, we are generating more data than ever, but tools to process this data are also starting to become more accessible to the average person. There are notably five main changes that have occurred with computing technologies that have enabled data to become so big. First, as we have discussed, there have been drastic improvements in cheap, large data storage in small form factors. This means that the average person has on their phone or computer tens of thousands of files, documents, images, videos, and pieces of software. This also means there are places where we can easily save our big datasets. Because these storage devices are getting smaller, we can have large amounts of storage in small devices, like phones. Meanwhile, cloud storage allows people to maintain massive amounts of data that they can access with a multitude of devices. Second, there have been major improvements in faster and cheaper processing power, such as the explosion in parallel processing graphics chips. For example, the average processor in a consumer computer can make about 150 billion calculations per second. This allows us to analyze big datasets relatively rapidly, and even in real time as we acquire it. Third, sensor technology and speed has also improved – most people have high-quality cameras in their phones, and may have devices (like fitness trackers) that can record movement, heart rate, elevation, skin conductance, and other measures. This allows us to obtain big physiological data, which can reveal underlying information about one's cognitive state (Chapter 9). Fourth, the wide spread of high-speed internet both in homes and out in the world is allowing more people to form communities, creating large amounts of data generated by people's interactions online. Fifth, our algorithms are getting better and smarter – we are able to compress data more efficiently and analyze data more effectively with tools like artificial

intelligence. The combination of these five computational improvements has led to an explosion of data produced by and accessible to the average person.

Big Data is also more important than ever for psychology research. Psychology has always been a multidisciplinary field, straddling social science and biological science programs at many universities. For example, psychology has clear links to neuroscience and experimentation, but also has implications for therapeutic practice and philosophy. However, recently, psychology as a field has begun to undergo a transition, with greater emphasis focused on experiments and complex analyses. Many exciting discoveries are coming about thanks to Big Data innovations, such as online experiments, artificial intelligence, and rich physiological data. With these innovations, researchers have been able to revisit classical psychological questions with a Big Data lens that allows them to assess their applicability across more diverse samples or make computational models that can predict people's behaviors. These innovations have also been saving psychologists a lot of time – making it faster to collect and analyze data. These changes go hand in hand with a new global scientific community that is developing, based around sharing data and code openly, in reaction to a "replication crisis" that emerged around unreplicable findings in small-scale experiments (see Section 2.1.3). So now is the perfect time to learn about these changes in psychology, to ride its waves as it moves into these new approaches.

Discussion Question

What topics relating to Big Data and psychology are you particularly excited to learn about in this book, and in your class?

1.6 How to Use This Book

As we just discussed, psychology is changing. If you want to go into psychological research for your career, professors and laboratories are now increasingly looking for candidates with experience in programming and statistics. Outside of academia, many jobs after college geared toward psychology majors – such as user experience design or data science jobs – also require these skills. For those of you wanting to practice clinical psychology, counseling, or go into education, it is still helpful to be up-to-date with the latest research and techniques (e.g., how is artificial intelligence changing the diagnosis of neuropsychological disorders?). And I would argue that some of these topics we will discuss in this book can help improve your daily life. I know for me personally, I've coded data scraping tools to find the best flights for a vacation, used generative AI to make a personalized storybook for my kids, or analyzed my fitness tracker data to get a sense of whether a diet is working. Knowing what is possible with data can change how you look at and use data in your daily life. In this book, we will also touch on some of the hot-button topics that have erupted in the news and the legal sphere as a result of Big Data – how do we deal with AI-generated fake information? How do we navigate the privacy risks created by the data recorded in many mobile apps and websites?

It can be intimidating jumping into learning about data and computer programming if this is your first foray into the topic. My number one goal is to demystify these topics and make you comfortable talking about them and thinking about them. As a student starting along this journey, it can feel like there's a big gap between you and your image of a computer scientist who may have been hacking computers since they were in elementary school. It can feel like you just aren't meant to be someone who codes or does complex math. But really these thoughts are a part of a mystical (but inaccurate!) aura that has surrounded computation. I'd liken computer programming to something like learning a foreign language or training for your first 5 km run. Most of the time, your goal isn't to become completely fluent or a record-breaking marathon runner. Usually, it's that you find these skills useful and enjoy the process of getting there. You also usually aren't worried about how you compare to the pros – you don't feel bad comparing your Spanish skills to those of a native speaker (and often they are impressed that you are trying!), or feel bad watching Olympics runners beat your time. In the same way, a seasoned software developer won't be quizzing you on the latest Python functions. You will also find that gaining these skills can enrich your daily life – you can now navigate a little around Madrid with your newfound Spanish skills, or be able to run to catch a bus without getting winded. Similarly here, you'll have moments where you may wish you could do something on your computer in an automated way and then realize there may be a way to use your skills learned here to do that!

At the same time, this book is not going to teach you programming or statistics from the ground up. It's the first step in learning the lingo and giving you the lay of the land, so you can then decide where you want to do a deep dive in future classes or explorations (e.g., do I want to study more neuroscience? Or web design? Or graph theory?). The online resources with this book will provide some stepping stones for doing these deep dives. With that foreign language metaphor, this book is your travel guidebook to help you decide where you want to study abroad. Then once you've picked a country, you can start focusing on learning its language. With this book, I want you to become fluent in the topics of new technologies being widely used in psychological research. I want you to have an increased level of agency over your own data and how it is used by companies and researchers. And, I want you to practice thinking creatively about psychological research questions and how we can answer them.

This book can be read from front to back or you can skip around sections as needed. In these first two chapters, I introduce what Big Data (Chapter 1) and small data (Chapter 2) are and how they compare to each other. Then for the rest of the first half of the book, I will give you the building blocks for running Big Data studies – looking at the participants (Chapter 3), the stimuli (Chapter 4), and the experiments themselves (Chapter 5). Once we are armed with our Big Data, we can then analyze it using artificial intelligence (Chapter 6) or human crowdsourcing (Chapter 7). In the latter half of the book, we will delve into different topics that are changing as a result of Big Data, and so these chapters are a bit more standalone. We will talk about software developments with apps and games (Chapter 8), as well as hardware innovations and physiological sensing (Chapter 9). We will talk about Big Data in neuroscience (Chapter 10), language and natural language processing (Chapter 11), and wrap up with social interactions and graph theory (Chapter 12).

With the exception of this chapter, each chapter will end with four key sections. In "Big Ethical Questions," we will talk about the ethical implications of the topic discussed in the chapter. These topics are sure to spark interesting discussion, especially because the ethical implications of these new methods are still being actively addressed in science and society. This section is then followed by a section on "Applications of the Chapter." While most of this book takes the framework of theory-driven psychology – where we are conducting experiments for the sake of understanding the mind, not creating a product – in these sections, we will discuss how the chapter's topics can be applied to impact the real world. Each chapter then ends with a Chapter Summary that reminds the reader of the major points, and Further Reading which suggests further sources to explore if you are interested in going beyond the pages of this book. There will be discussion questions laced throughout the chapters, as well as a sample homework assignment at the end of each chapter. The companion Teacher's Guide will include additional discussion questions, exercises, and demonstrations for each topic.

Importantly, data, computation, and the internet are always changing. While this book is written at a static point in time (2022–2025!), there is an accompanying online resource (www.cambridge.org/bainbridge) that will be updated as technologies change in the world. If you read anything in this book that seems outdated, check out the online resource to see if there is a new version of that information. The online resource also has interactive demos and programming tools to let you learn more about programming and test out online experiments.

With that, let's proceed to Chapter 2 to discuss what "small data" is and how that differs from Big Data in the context of psychological research.

CHAPTER SUMMARY

In this chapter, we introduced the concepts of Big Data, psychology, and how now is the perfect time to study them and their interactions.

1. Here, we define Big Data as unstructured, naturalistic human data requiring complex analytical methods.
2. We define psychology as the empirical study of the mind, brain, and behavior. This book mainly focuses on quantitative experiment-based psychology.
3. With major improvements in our technological capabilities over the past few decades and changes in the landscape of psychology, now is the perfect time to study how Big Data can be used in psychology research.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter.

- **Read about and watch an original video describing the world's first hard drive, developed by IBM in 1956:** Seeley, C. (2014, October 28). History snapshot: 1956 – the world's first moving head hard disk drive. Data Clinic Ltd. News. www.dataclinic.co.uk/history-snapshot-1956-the-worlds-first-moving-head-hard-disk-drive

- **A review of how realism in our studies can actually show differences in the brain:** Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*, 506–519.

ASSIGNMENT

The purpose of this assignment is to learn more about your experience with Big Data and provide you a sense of the data you generate.

Total Points: 50

1. **Fill out the class survey.** Your professor will provide a link. (20 points)
   Let's look at how much data you are generating just from your phone!
2. Locate where your phone describes your storage usage. **Answer:**
   a. How much storage are you using for images? (1 point)
   b. How much storage are you using for videos? (1 point)
   c. How much storage are you using for music? (1 point)
   d. How much storage are you using for apps/applications? (1 point)
3. Let's get a rough estimate of how much data you are generating a day with your phone camera.
   a. Add together your answers from 2a and 2b and **report that number here**. (2 points)
   b. Get an estimate of how long you have had your phone – find the date of the first photo you took. Then search on Google "how many days between [that date] and today" and it should return you the number of days. **Report that date of the first photo and the number of days since then.** (2 points)
   c. Divide your answer in 3a by your answer in 3b and **report that number here.** This tells you about how much data you are generating with your camera per day. (4 points)
   d. **How many bytes of data is that?** (2 points)
   e. One byte is the amount of data used to type one character (e.g., "A"). A novel contains about 500,000 characters. **How many books worth of data is that?** You're likely creating the equivalent of books of information a day! (3 points)
4. Let's see how long you are using your phone for.
   a. First, guess: **How much screen time do you think you use a day?** (2 points)
   Now, locate where your phone describes your screen time usage (this might be under a "Screen Time" setting or a "Digital Wellbeing" setting).
   b. **On average, how much screen time do you actually use a day? How does this compare to your guess?** (4 points)
   c. Based on your screen time report, **on average how much screen time do you spend on social media a week?** (2 points)
   d. You use on average 500 MB of data per hour by browsing social media. **How many GB (or MB) of social media data are you viewing per week?** (5 points)

As you can see, we interact with massive amounts of data in our daily lives!

REFERENCES

Adobe, Inc. (2013, August 1). Did you know that Photoshop 3.0 was the last version of Adobe Photoshop to be sold on the floppy disc? Facebook. www.facebook.com/photo?fbid = 10151614431968871&set = a.468676338870

Bainbridge, W. A., & Baker, C. I. (2022). Multidimensional memory topography in the medial parietal cortex identified from neuroimaging of thousands of daily memory videos. *Nature Communications*, *13*(1), 6508.

Duarte, F. (2024). Amount of data created daily. Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day

Gannon, L. (2002). A critique of evolutionary psychology. *Psychology, Evolution & Gender*, *4*, 173–218.

Kumar, S. (2015). *Fundamental limits to Moore's law*. arXiv:1511.05956.

Moore, G. E. (1965). Cramming more components into integrated circuits. *Electronics*, *38*(8).

Waldrop, M. M. (2016, February 9). The chips are down for Moore's law. *Nature* [news feature]. www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338

2 What Is Small Data?

Introduction

In order to learn about Big Data, you first need to understand its counterpoint, "small data." Small data isn't often called this, because data from most psychology studies fits under this umbrella, and many times its scale can suit our purposes just fine. Thus, a definition of small data would be any data that isn't Big Data. While it is incredibly common, solely using small data severely limits the takeaways we can get from psychological research. In this chapter, I will discuss the limitations of small data, as well as the limitations of Big Data. You will see how the two can work in synthesis to pinpoint the rich phenomena occurring in our minds and brains.

Specifically, first to understand the benefits we gain from Big Data, we will go through a few example small data experiments (Section 2.1.1) and see what they are lacking (Section 2.1.2 and Section 2.1.3) and how they can be made bigger in scale (Section 2.1.4). We will then discuss some limitations to Big Data experiments (Section 2.2), including new issues they introduce (Section 2.2.1) and the limitations that will always be present with any study (Section 2.2.2). We will then discuss how experiments can be dichotomized into being hypothesis-driven or data-driven (Section 2.3), as well as how Big Data studies can be characterized as deep or wide (Section 2.4). We will discuss the ethical issues that can come about from the multiple analyses run with Big Data (Section 2.5). We will then discuss applications of the topics in the chapter, such as examples of famous data-driven discoveries (Section 2.6.1) and medical applications of deep and wide data (Section 2.6.2).

2.1 Turning a Small Data Experiment into a Big Data Experiment

Let us first begin with an example of a small data experiment and think about how we can make it bigger and broader.

2.1.1 A Case Study

There is a famous effect in psychology called the **own-age effect** (Anastasi & Rhodes, 2005), where people tend to remember faces close to themselves in age better than faces farther in age.

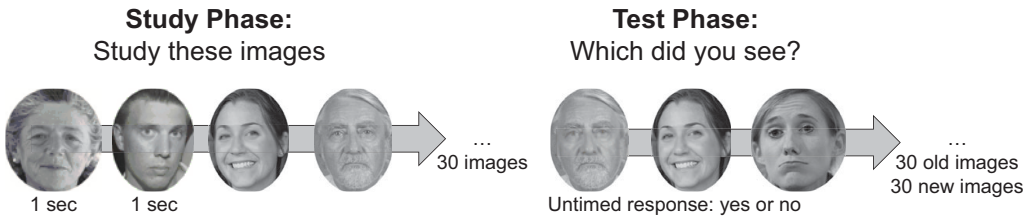


Figure 2.1 The experimental methods for our own-age effect experiment. We have participants first study thirty face images for 1 second at a time. Half of the face images are from older adults and half are from younger adults. We then test them where we show them thirty of the face images they saw, randomly mixed up with thirty new face images they didn't see. For each face they have to respond if they saw it before ("yes") or not ("no"). Our main research question is whether participants have different levels of memory accuracy based on the match between their own age and the age of the face images.

You may have experienced this before, where you may have an easier time recognizing your classmates than professors on campus. (There are many memory effects driven by the similarity of a face to your own – there is also famously an own-race effect; Chiroro & Valentine, 1995). Let's say we are in a traditional psychology lab, and we are running an experiment to test the own-age effect. Our experimental methods look something like what is in Figure 2.1.

The idea is to recruit fellow psychology students on campus and run them through a face memory test on the computer in the lab (as most psychology experiments are done!). In that memory test, we will show a series of thirty faces (like in Figure 2.1), where half are college-aged, while the other half are older adults. We will then test to see if there is a significant difference in memory for those two groups of faces. After running twenty participants, we find a significant effect – indeed the own-age effect holds true!

Discussion Question

What prevents us from generalizing these results to saying that the own-age effect occurs for all observers and all faces?

2.1.2 What Does a Small Data Experiment Miss?

The previous case study was an example of a typical psychology experiment. However, there are many aspects of it that prevent us from generalizing to all observers and all faces. Specifically, the participants, the stimuli (the face images), and the experiment itself are all constrained and artificial in some way.

Small Participants: First, the number and scale of the participants is "small" – can we really make generalizations about humans as a whole from an experiment run with twenty students at a specific university? For example, would these effects replicate for people who are frequently exposed to faces of other ages – like in cultures where young adults tend to live with older

generations? In Chapter 3, we discuss more about the problems with using small college samples in a large proportion of psychology studies, and what we can do about it in the field.

Small Stimuli: Second, the images are also very small in scale. Like the issue we have with participants, can thirty faces really capture the rich variance of human faces out in the world? If you look at the paradigm (Figure 2.1), all these faces are very homogenous. They are all front-facing white people of moderate attractiveness with an oval cropped around their face so you cannot see much of their hair or clothing. This can sometimes be intentional – researchers often want to control for factors they're not interested in, so that those cannot be alternate explanations of their effect. For example, you don't want to think there is an effect of age on memory when it's actually the clothes the models are wearing (maybe clothes from a few decades ago are more memorable than clothes from today!). But, too much control will limit our ability to make generalizations across different lighting, viewpoints, and facial expressions – it doesn't let us make confident predictions about memory out in the real world. And, by only doing research on constrained demographics (e.g., all white people), we aren't studying the rich variation in human experience. In order to generalize to the real world, we need images that better capture the diversity we observe in that world. In Chapter 4, we talk more about how to think about and create more representative stimulus sets.

Small Experiment: Even in just the way they are conducted, experiments are much smaller in scale than the real world. They don't capture what it's like to meet a moving, emotive, multisensory human being, and try to encode them into memory. Experiments tend to be brief (usually 30 minutes to an hour) and constrained to a two-dimensional computer screen, with a few seconds to see each face. This is not at all what it's like to meet a face in reality – you see them out situated in the real world, and you may spend hours interacting with them. Perhaps the dynamic, moving aspects of a face can contribute to your memory for that face, and that would be completely ignored by the experiment. Or perhaps seeing faces in the three-dimensional world is fundamentally different for memory than seeing them on a flat, two-dimensional screen in an experiment. (Although seeing faces in two dimensions may be becoming more natural, as virtual meetings are becoming more common.) Also, because faces are so dynamic, it's unlikely in the real world that you will ever see the exact same view of a face again; you can never take the exact same photograph twice. The second time you see a person, their facial muscles will be engaged in a slightly different way, the lighting will hit their face differently, or they may have a slightly different glow to them. This is completely different from an experiment which shows you the exact same photograph twice.

2.1.3 A Second Case Study and the Replication Crisis

Let's examine another sample experiment. Within the field of social psychology, one phenomenon that has been proposed is the phenomenon of **social priming**. The idea with social priming is that when you are made to think of a social category, you automatically think about related behaviors and stereotypes and start to subtly behave in a similar way. This was first demonstrated in a study by Bargh and colleagues (1996) across a series of experiments. For example, in one experiment, thirty psychology class undergraduates from New York University were asked to complete a task where they had to take a set of five words and create a grammatically

correct four-word sentence as quickly as possible. They did this for thirty sentences in total. Unbeknownst to those participants, half of them received words specifically related to being elderly – *old*, *grey*, *sentimental*, *bingo*, *wrinkle* – while the other half received neutral words. The idea was that the elderly-related words might prime them to think about elderly individuals and act in a similar way. An experimenter then secretly timed how long it took participants to exit the hallway leaving the testing room. The researchers found that participants primed to think about being elderly had a significantly slower walking speed (8.3 seconds to travel the hallway) than participants given a neutral prime (7.3 seconds), confirming their hypothesis. Participants reported not being aware of this elderly manipulation, or a change in their behavior, suggesting these social priming effects could happen unconsciously.

Discussion Question

What factors in this experiment might prevent us from generalizing more broadly?

This experiment uses a relatively small number of participants (fifteen in the elderly prime condition) and stimuli (thirty), making the robustness of the effect unclear (though the original experimenters do actually replicate this effect in a second thirty-participant experiment). The participants come from a very specific sample – psychology undergraduates in the New York area – who are unrepresentative of the world population. The words are also not validated as conjuring an image of "the elderly" in an objective way. The study does a fairly good job at using a naturalistic task (i.e., measuring walking time). However, there could be modern improvements on how it is measured, rather than relying on an experimenter's timing skills, which could introduce a subtle bias that accounts for the 1-second difference between conditions. As a result of these critiques and others, Doyen and colleagues (2012) ran a larger-scale replication of the experiment. They ran 120 participants (albeit also from a fairly specific sample – Belgian French-speaking undergraduate students). They used elderly word stimuli that were first confirmed by a separate set of eighty participants as representing old age. The experimenters then used infrared sensors to precisely measure the amount of time it took to traverse the hallway. With this "bigger" experiment, researchers found no difference in walking speed between their two participant conditions.

Around the same time (in the early 2010s), many psychological findings were unsuccessfully replicated. Researchers were failing to find clear evidence for many social psychological phenomena that had become well-accepted – in addition to social priming, there was now evidence against ideas like ego depletion (the idea that willpower is a finite resource) and power posing (that standing in a certain way will increase your confidence) among others. This launched a "**replication crisis**" across the field of psychology bringing into question the quality of the research in the field. One event that ignited this crisis was when a paper was published in one of the most revered social psychology journals (*Journal of Personality and Social Psychology*) claiming evidence that people can see the future (use "precognition"; Bem, 2011). Researchers realized that a combination of poor research practices as well as publication pressures in the field (see Section 4.4) was overinflating the reporting of supposed "results" across many papers.

At this major breaking point, hundreds of researchers as part of the Open Science Framework launched an effort to attempt to reproduce a hundred findings in psychology. Shockingly, only thirty-six were successfully replicated (Open Science Collaboration, 2015). This served as a reality check for psychologists – we need to run experiments with larger samples, more generalizable experiments, and better statistical measures. We also often should run multiple replication experiments to confirm our effects really hold, and aren't just occurring due to chance.

2.1.4 Making an Experiment Big

Even if I have convinced you that traditional psychology experiments are often unnatural simulations of the real world, how can we improve upon this? How can we make our experiments "bigger"?

> **Discussion Question**
>
> What would a Big Data version of the example face experiment look like? What would you change?

We need to think about how we can improve upon the three points mentioned earlier: the participants, the stimuli, and the experiment. For the participants, can we recruit more people, and more widely? In Chapter 5, we will talk about how to conduct online experiments, which lets you reach thousands of people, with more diversity than the average college campus. For the stimuli, we can also strive to collect image sets that are larger, more natural, and more diverse (refer to Chapter 4 to learn how we can do that!). For making the experiment more naturalistic, there is the difficult balance of wanting scientific control but also generalizability. If you want to keep it as a computerized task, what if you test people on memory for a face across different photographs of that person, rather than memory for a specific image? (It turns out recognizing an unfamiliar person from different photographs is a very difficult task! See Jenkins et al., 2011). New technologies are also making it easier to conduct experiments in more dynamic, three-dimensional environments like virtual reality, or even out in the real world (e.g., Snow & Culham, 2021; Võ et al., 2019). If we are able to expand our experiments out in these three ways, then we have a more generalizable study in these three ways as well. We can know things about face memory across a wider range of observers and faces being observed, and we can try to make predictions about behavior out in the real world.

For example, in our lab, we were curious about people's memory for faces more generally than the own-age effect. So, we generated a large database with demographics matching the United States (see Chapter 4 for more information). We then had over 800 diverse individuals engage in a face memory experiment online (Bainbridge et al., 2013). In this experiment, people viewed a stream of face images and pressed a key when they recognized a repeat from earlier (called a **continuous recognition task**) – a little like the experience of walking through a crowd and recognizing some people as you pass them. We also ran a version of the experiment where we tested people's memory for faces across different viewpoints and facial expressions,

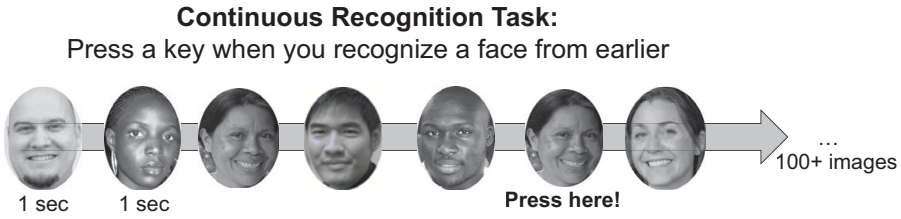


Figure 2.2 The experimental methods for our more "Big Data" face experiment. People still view face images for 1 second at a time. However, now they are continuously seeing images and indicating their memory as part of a "continuous recognition task," akin to the experience of walking through a stream of people and sometimes recognizing someone. We now also have a much more diverse range of faces, so we can test many phenomena, such as the own-age effect or the own-race effect, as well as measure the intrinsic memorability of a given face image. We would also test this task with many diverse participants, so we can look at more generalizable effects of the viewer, too.

not just face images (Bainbridge, 2017). So, our experiments looked a little more like Figure 2.2. What we found in the end was that there are certain faces that are remembered very well by most people, and some faces that are remembered very poorly. In other words, faces have an intrinsic *memorability*. We have recently extended these findings to images more generally, and tested them out in the real world – for example, discovering that we can even make predictions about what paintings people will remember in a freeform visit to an art museum (Davis & Bainbridge, 2023).

2.2 Limitations of Big Data

It may seem obvious that we'd want to aspire for diverse, generalizable experiments as described. However, there are some obstacles presented by Big Data experiments that are important to consider.

2.2.1 Problems with Big Experiments

One of the major cons of Big Data-style experiments is that they can introduce noise into our data. First, sometimes the images can vary too much. If each image is different along many dimensions (age, gender, attractiveness, facial expression, lighting, angle, hairstyle, eyebrow width, etc.), then how can we pinpoint which specific dimension is causing the effect we're studying? And maybe many (or even a majority!) of these dimensions might be things we don't care about, like lighting. Similarly, if our participants vary too much, we can have the same problem – everyone may act in a unique way that prevents us from finding generalizable effects. And participants may vary in ways that are not interesting to us – for example, in running an online experiment, what if the participant's screen monitor resolution, or what they ate for breakfast that specific day, or the length of their fingers influenced how quickly they pressed keys on a task? Small experiments let us directly test our effect of interest, without having all these

extraneous factors in the way because we control for them as much as possible. So, sometimes, we have to design our Big experiments in a way that they're not *too* big. At the same time, Big Data experiments let us simultaneously look at the contributions of many factors – for example, we can analyze how gender, age, and race all contribute to face memory, at the same time.

2.2.2 Imperfect Experiments

Even when we want to run Big Data experiments, there will always be limitations that make it impossible to run a perfect experiment. There will inevitably still be biases with the stimuli and participants in the experiment, because you cannot make everyone in the world participate in the experiment, or make everyone agree to be photographed as a face in the experiment. There are some factors limiting who can be a participant in your experiment – participants must have some familiarity with how to read a computer, and they have to have free time and interest in participating over other things they could be doing. Similarly, only a subset of people would be okay being photographed for the study, and any set of natural photographs will likely have an over-representation of happy or neutral facial expressions (over angry or sad).

There are also some other practical limitations with Big Data. Sometimes the data is so big that we are limited by the processing power, storage, or internet speeds that support us saving and analyzing the data. For example, one person's MRI brain data can take up 1 terabyte of space, which is more than the amount of space many computers come with (in 2025). It can also take half a day to download this data for just one person! So, it can be difficult to analyze data from hundreds, let alone dozens, of participants. Large-scale experiments can also be very costly with time and money. Using the same example of an MRI experiment (which is on the upper end of what psychology experiments cost), one participant usually lies in the scanner for about 2 hours, and it will cost the researchers around US$1,000 to the scanner center for that time. So, an experiment with 100 participants would end up costing $100,000 and take 200 hours of the researchers' time to just collect the data. We are also still limited in our analytical techniques for Big Data. When dealing with very big, naturalistic data, we often don't look at just a single measure or statistic. But, at the same time, our statistical tools and artificial intelligence are not yet able to fully interpret natural human behavior. For example, let's say we wanted to look at face memory in the real world, and recorded participants' view as they navigate through a party, using some sort of head-mounted camera. It's not clear how we would analyze these data – how to turn the conversations with people, the amount of time looking at them, the thoughts related to them, etc. – into numbers in order to make conclusions about what influences someone's memory of a person. So, our analytical techniques are limited (and in fact, they are dependent on the study of psychology to guide us on how to analyze such complex human behaviors).

2.3 Hypothesis-Driven versus Data-Driven Research

A majority of psychology experiments can be characterized as **hypothesis-driven research**. These are experiments where the researchers have one or a few key research questions. They also tend

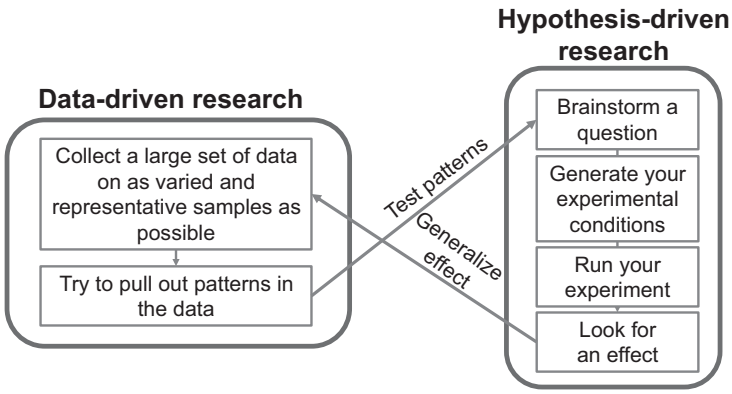


Figure 2.3 The series of steps you take when conducting data-driven research and hypothesis-driven research, and the way in which they interact. In data-driven research, you collect a large, representative set of data and identify patterns in the data. You can then take those patterns and test them in controlled, hypothesis-driven experiments to determine the specific mechanisms driving the effect. To conduct hypothesis-driven research, you would brainstorm your research question, create experimental conditions to answer that question, and run your experiment. If you identify an effect in a hypothesis-driven study, then it can be helpful to test whether the effect generalizes by running a more naturalistic, data-driven study.

to have a clear idea of the different alternatives of the results (in other words, **hypotheses** about the results) and what that means for the bigger picture question. The general pipeline for running a hypothesis-driven study follows the flow of the right side of Figure 2.3. One of the most important first steps is deciding on the main question. So, taking the first case study example we covered, let's say your question is: "Are people better at remembering faces close to them in age than faces far from them in age?" Your two experimental hypotheses would be something like: 1) yes, faces closer in age are remembered better, or 2) no, there is no difference in memory based on age of the face. You will then design your experiment, selecting experimental conditions that let you pinpoint that question. **Experimental conditions** are the different ways you divide up your experiment to answer your research question. For example, for this study, we may have different participant groups (older and younger people) as well as different stimulus sets (older and younger faces). So, we would have four conditions: young participants viewing young faces, young participants viewing older faces, older participants viewing young faces, and older participants viewing older faces. In experimental design terms, participant age is a **between-subjects** factor, because the condition changes from subject to subject (in other words, between the subjects). In contrast, face image age is a **within-subjects** factor, because within a single subject, they will see faces from both older and younger face conditions. Importantly, the conditions and factors that we intentionally change or control as experimenters (like the gender of participants and faces) are called **independent variables (IVs)**. It's because these are variables that we set, and so they are not dependent on other things in the experiment – they stand on their own. Once you have designed your experiment with its conditions, you then run the experiment (i.e., having your participants do a memory task with

the images you choose). The data that comes out of your experiment are **dependent variables** (DV), because they are dependent on the IVs and the experiment that you run. Finally, you can then run statistical tests on your data to directly answer the research question you set out to test. For a statistical test, you will typically define your **statistical hypotheses** – these are subtly different from experimental hypotheses in that they specify the different possible results of your statistical test. With most traditional statistical tests (called parametric statistics; see more discussion in Chapter 3), you would have a **null hypothesis** reflecting the hypothesis that there is no effect for a given statistical comparison. In this case, the null hypothesis would be that there is no difference in memory performance across our different participant-image age conditions. You would also define an **alternate hypothesis** reflecting the hypothesis that there is a significant effect. Here, the alternate hypothesis would be that there *is* a difference in memory across these conditions. So, we run our test and see which hypothesis we have more evidence for – can we reject this null hypothesis or fail to do so? Specifically, you would directly test whether memory performance (your DV) is higher for same-age conditions (young participants / faces and old participants / faces) than different-age conditions. And, voila, you have your answer! (So far, the research seems to point to yes: Chiroro & Valentine, 1995).

This hypothesis-centric way of thinking often goes hand in hand with small data research, because you want to run experiments that specifically target your question of interest. You *can* run these in a Big Data way (e.g., with thousands of participants and thousands of images), but in the end, the only factor you'll want to differ is the one you're interested in (age), and you'll want other factors like race, gender, attractiveness, etc. to be the same across your different conditions. This is because you want to be certain that your experimental manipulation has a direct influence on the effect you observe (in other words, you want your IV to directly influence your DV). If you do not control for other factors, you risk having **confounds** that could explain the link between your IVs and DVs. A confound is another variable that can account for the relationship between your IV(s) and DV(s) that you are not intentionally manipulating. In other words, they can be alternate explanations for your results. When designing an experiment, you want to make sure you can avoid these confounds, or at least can take them into account in some way. For example, let's say we look at monthly data across a year and find a correlation between ice cream sales and drowning deaths: when ice cream is popular, drowning is more common (Mumford & Anjum, 2013). Does this imply that ice cream causes people to drown? That would be ridiculous (and unfortunate)!

Discussion Question

What are some confounds that could explain a relationship between ice cream sales and drowning deaths?

One of the big confounds here is weather! During the summertime when the weather is nice, people want to go out swimming. They certainly have a much higher risk of drowning if they're swimming in a lake than if they're staying home bundled up by a fire during the wintertime. During the summer, people are also probably going out to buy ice cream to cool off from the hot weather, so you would see both high ice cream sales and increased drownings. On the other

hand, during the winter, the dessert of choice might be something more like a slice of apple pie or a mug of hot chocolate, and you would be unlikely to be out swimming. So, we would say that weather here is a confound in this relationship between drownings and ice cream sales. When designing a hypothesis-based study, it's important to keep in mind all potential confounds, and sometimes it can be impossible to control for all of them.

The counterpoint to hypothesis-driven research is **data-driven research**. The idea is that you collect tons of data, trying to gain samples that are as representative and varied as possible (Figure 2.3), and this is the approach often taken when using Big Data. Then, you use statistical methods to try and pull out patterns from the data that can help answer some questions. For example, you could collect memory test data for a wide range of faces and participants, and then see if people generally tend to remember faces closer to their own age best. This type of research is also sometimes called **exploratory research**, because you can explore around the data and look for different patterns without necessarily having a hypothesis from the beginning. What's great about data-driven research is that you generate big datasets that can help answer many questions. So, these databases can be multiuse – you could look at questions about memory and age, but also memory and attractiveness, or memory and face shape. You can also look at how these different factors all work together to form the big picture (i.e., what combinations of features influence the memorability of a face?). However, because these data tend to be collected without controlling much in the experiments, you run a higher risk of having more confounds that can explain your effects. However, because you often aren't relying on a single research question or statistical test, one confound may be less impactful on the use of the dataset overall. However, if you are not careful, data-driven research has some other big risks that can result in low-quality science (see Section 2.5 on data fishing).

Overall, there are pros and cons to both hypothesis-driven and data-driven research, with the key points summarized in Table 2.1. But ultimately, science benefits most when we do both, because they serve as an interconnected loop. Data-driven studies let us discover new, unexpected effects that can emerge from large or naturalistic data. Hypothesis-driven studies then let us take these effects and pinpoint the reasons behind these effects and link them to broader theories about human cognition. A lot of psychology reasonably takes the hypothesis-driven approach as a result. But to broaden our perspective on what questions to ask and what blinders we might have on in the field, I would argue the data-driven approach is just as important – and this is what will be the focus of this textbook.

Table 2.1 Comparison of the pros and cons of hypothesis-driven and data-driven research

Hypothesis-driven research	Data-driven research
Usually small data	Usually Big Data
Can isolate specific effects	Can be more naturalistic
Pulling out data based on theory-driven questions	Pulling out questions based on diverse datasets
Larger effect sizes	Statistical significance more likely
Have to be wary of confounds	Have to be wary of data fishing

2.4 Deep Data versus Wide Data

When designing a Big Data study, there are two dimensions along which it can be Big – deep or wide. A **deep data** study is one where you collect lots of data for a smaller number of individuals (so you're getting a deep look at a few people). Some examples include sensor data like a fitness watch recording frequently and over long periods of time (Chapter 9), or software-based data recording lots of samples over time (like on your phone; Chapter 8). There are also some experiments that focus on running the same participants many times over a series of sessions. These repeated measurements can give rich information about individuals, letting us look at the influence on cognition of things that vary like the time of day, attention fluctuations over time, and complex behaviors. One issue with deep data, though, is that it can risk being invasive of participants' privacy because you're learning so much about specific people. For example, if you look at one person's measurements from a fitness watch over months, you would learn all about their sleeping and exercise habits. It can also be tedious for participants to collect and provide all this data, especially if it's a study where they have to come in for multiple sessions. So, it can be hard to recruit participants for deep studies.

A **wide data** study is one where you collect relatively small amounts of data from a large number of participants at a single time. Some examples include data from an online experiment across thousands of people, or a snapshot of rich data from an app or piece of software at a single point (like all the tweets for a topic on a single day). What's great about wide data is that it can give diverse information across a large, representative sample of people. However, you often cannot capture very complex behaviors that vary over time or an interaction.

Some researchers characterize Big Data along three dimensions – being deep, wide, and long. In this case, deep data would still involve multiple measurements (like we see with sensor data). However, wide data now instead reflects collecting data across multiple variables or measures (e.g., with a battery of questionnaires). Finally, long data would involve collecting data from many people. Regardless of how you characterize the dimensions of Big Data, studies can be any combination of deep, wide, and long (e.g., collecting tons of data from many people), although this can be hard to achieve, so scientists may need to pick one dimension along which to specialize. When looking at a study, it's worth thinking how it falls on these different measures of size.

2.5 Big Ethical Questions

When you have a huge experiment, it can be easy to go fishing around for significant effects. This is something called **data fishing**, **data dredging**, or p-hacking. Since you have so much data, it seems like one of the benefits is that you should be able to look at many different effects in your data at once, right? Well, yes, you can do this to some degree, but you also need to think about how statistical tests are conducted.

For a majority of standard statistical tests that compare your data to a distribution (like t-tests, ANOVAs, regressions, etc.), you aim to estimate a ***p*-value**. What this *p*-value represents is the probability you would observe something as extreme as your results if the null hypothesis

were true. Recall that the null hypothesis reflects the hypothesis that there is no effect in your data. However, even if this null hypothesis were true, there is noise in our measurements and people's behaviors, so we would still sometimes observe a difference between our conditions "by chance." When we run a statistical test, we are looking at what the distribution of data would look like if the null hypothesis were true (and there was no effect). We are then seeing where our observed data falls in this distribution – how likely is it to occur given this null distribution? We calculate our *p*-value as the proportion of data in the null distribution that is equal to or more extreme than our observed value. So, for example, a *p*-value of 0.03 indicates there's only a 3 percent chance you would happen to observe these results just by random chance. That seems pretty low, and as a field, we've currently accepted a cut-off of 5 percent ($p < 0.05$) to be how we determine what we'll take to be a significant finding or not. Another way to phrase this is that in our field, we have accepted a 5 percent **false positive rate.** This is the rate of falsely saying something shows an effect when it does not. You may have heard this term used to refer to the rate of a medical test falsely saying you have an illness when you do not – same idea.

While this 5 percent chance of a false positive seems rare, when you're dealing with Big Data, you're doing many statistical tests – maybe hundreds of tests (e.g., for a psychological battery), or even up to hundreds of thousands of tests (e.g., for the case of MRI brain data). And so, in the realm of hundreds of thousands of tests, even with this seemingly strict false positive rate, we will get about 5,000 tests (5 percent of 100,000) that come out as "significant" just by chance! So we need to think carefully about how we define significance with data-driven research since we are doing many tests, inflating the chance that we find a false positive in at least one of these tests. In order to circumvent these issues, we do something called **multiple comparisons correction**, which is a group of statistical methods that let us calculate an adjusted *p*-value threshold for our study that takes into account the many tests that we are doing. While we won't go into these methods in detail, some example methods include **Bonferroni correction** and **false discovery rate correction**. Bonferroni correction corrects for the rate of false positives across all of the statistical tests that you perform. It does this by calculating an adjusted threshold for "significance" (called the alpha level, or α), based on the number of tests you are running. So, if you run ten tests, your alpha level would be $p < 0.005$ instead of $p < 0.05$. False discovery rate correction is a more liberal method that corrects for the proportion of false positives among all results initially labeled as significant – in other words, calculating an alpha level so that we are okay with 5 percent of our discoveries being false positives.

Let's look at an example study that ran many statistical tests. In Moore et al.'s 2006 study "Thongs, flip flops, and unintended pregnancy," the researchers wanted to investigate if there were some lifestyle factors that were related to unintended pregnancies. They conducted a 50+ question survey with 126 women who were currently or recently pregnant, and conducted 362 statistical tests to analyze their data. They found some surprising results: unintended pregnancy was associated with preferences for yoga, beaches, thongs, Doritos, contact lens, and text messaging. They also found that baby boys were more common if mothers preferred trucks, beef, and boys, while baby girls were more common if mothers preferred cars, chicken, and girls. So does that mean if you see your friend texting their friends during some beach yoga while tucking into a bag of Doritos, that you should encourage them to be vigilant with their

contraception? No, because if you think back to what we just discussed with running many statistical tests, we would expect about 18 of their 362 tests to come out as significant just by chance given our *p*-value threshold of 0.05! So even if there are no meaningful relationships between any of these factors and unintended pregnancy, just because of random noise in measurement, participant behaviors, and the environment, it would be unsurprising to find some relationships that come out as statistically "significant" but aren't real.

So one of the risks of Big Data is that it's easy to run many, many statistical tests until you find something significant. Because of all of the rich data you have, it's tempting to test many different questions. There are also big pressures in the scientific world to publish significant results, so researchers may be tempted to focus on these "significant" results without accounting for the number of statistical tests that they're running. In other words, you may be tempted to fish around for a result in your big sea of data. There are three main ways to make sure you are not doing data fishing with your own data. One way is to perform multiple comparisons correction across all the tests you run. A second way is to decide your analyses and hypotheses in advance before seeing your data (called preregistration; see Section 4.4) – so in other words, running a combined hypothesis- and data-driven study. A third way is to replicate any findings you discover across multiple datasets, analyses, and/or labs to be sure that what you're finding is real, rather than something that emerges just by chance.

There is also the question of the **effect sizes** of the results you end up finding. While we often care about statistical significance in our data, we also care about how strong the effects are in the differences that we are measuring. Effect size is often quantified as the proportion of the **signal** of interest to the level of **noise**. So, for example, for a t-test, the measure of effect size is the difference between the conditions' averages (the "signal"), divided by the standard deviation pooled across the two conditions (the "noise"). You can have a significant effect that's a weak effect or a strong effect. For example, let's say you're looking at whether an intervention in the classroom results in a difference in test scores on a test with a maximum of a hundred points. You could get a significant effect where the intervention results in a one-point increase. While this would mean the intervention likely worked (because the effect is significant), it didn't work very well (the effect is weak)! If the intervention instead resulted in a significant thirty-point increase, we would say this is a strong effect! And, if you have a nonsignificant effect with a thirty-point increase, that would mean our results aren't strong enough (e.g., there may be too much noise), so we cannot be confident that this thirty-point increase didn't just happen by chance. Because of its large sample sizes, Big Data can be prone to identifying significant but weak effects – effects that would only be detectable when you have thousands of people. Therefore, even if you find a significant result, consider what the result means. If it is a meaningful, strong psychological effect, ideally we would even see it occur at the level of a smaller sample, and even at the level of the individual.

2.6 Applications of the Chapter

In this chapter, we discussed characterizing research in a few different ways – for example, hypothesis-driven versus data-driven or deep versus wide data. These ways of thinking about

data have promoted discoveries beyond the field of psychology, and have guided recent advancements in the medical field.

2.6.1 Data-Driven Discoveries

We mentioned how data-driven research can result in new questions or effects that we may not have conceived of if we only stuck to pre-existing theories and hypothesis-driven experiments. There are in fact many exciting scientific discoveries that came about thanks to people trying out many different things. One of the most famous examples in psychology is the discovery by Professors David Hubel and Torsten Wiesel that led to their Nobel Prize win in 1981. They were trying to see what information was coded in neurons in the occipital lobe (the early visual regions of the brain), by recording directly from cats' brains while showing them different images (see Chapter 10 to learn more about neuronal recording). They were struggling to find any specific image that would cause these neurons to spike. Back in the day, they were using a slide projector, and suddenly when they were swapping out the slides, they heard the neuron they were recording from start to fire. After playing with the slide, they discovered that this neuron was sensitive to the edge of the slide when it was shown at a specific angle. This led to our current understanding of the visual system in the brain, where neurons are sensitive to edges oriented at specific angles. You may have heard about similar fortuitous "eureka!" moments throughout the sciences. As the classic example, around 246 BC, Greek scientist Archimedes realized how to calculate volume and density while taking a bath, and purportedly ran through the streets shouting "eureka!" In 1820, Dr. Hans Christian Oersted noticed a compass move when he placed it near an early battery he was creating – resulting in the discovery that electrical currents generate a magnetic field. Percy Spencer invented the microwave in 1946 when he noticed the chocolate in his pocket melted when he was testing out a new vacuum tube. These discoveries may have never happened without the experimenters just trying out different things. Big Data can encourage such exploration, which can lead to exciting discoveries.

2.6.2 Medical Applications of Deep and Wide Research

Deep and wide methods have had some wide-reaching applications in the clinical realm. Deep data has helped form the field of **precision medicine**, where healthcare workers can make honed, personalized predictions of health outcomes based on genetics, environment, lifestyle, and sensor measures. Big Data lets us create **predictive models** that take these different factors and then make guesses about outcomes for a single person (see Chapter 6). Precision medicine goes hand in hand with preventative medicine and telehealth, where people can wear sensors and use apps to remotely track and communicate symptoms before they develop into a full-blown condition. For example, researchers are working on apps to help identify early stages of Alzheimer's disease (Konig et al., 2018) and apps to help elderly individuals develop memory strategies (Martin et al., 2022).

Wide data is key in letting us learn about diseases: It lets one see global trends in disease, identify rare groups at particular risk, and find hidden links to a cause or cure (Heggie, 2019). Wide data played an important role in identifying the symptoms early on in the COVID-19

pandemic, when it wasn't clear what symptoms were being caused by the virus. Researchers conducted a large-scale wide symptom study where anyone could enter their symptoms online, and they ended up receiving information from 4.4 million participants (Menni et al., 2020). As a result, they were one of the first groups to identify a loss of smell or taste as one of the symptoms of COVID-19. They also found some other interesting trends: for example, for the first wave of the pandemic, one out of twenty participants had symptoms that lasted more than 8 weeks, and longer COVID was correlated with having more different symptoms in the first week. They also found that during the pandemic lockdowns, 20 percent of participants had an increase in alcohol consumption, and an average weight gain of 4.6 lbs.

CHAPTER SUMMARY

In this chapter, we discussed the small data experiments traditionally utilized in psychology research and showed how they compare to the Big Data experiments that are becoming increasingly popular. Here are some of the main takeaways:

1. The key to making a small data experiment "Big" is expanding its participants, stimuli, and paradigms to be more naturalistic and representative of the real world. However, you can almost never make a perfectly representative experiment.
2. There are different benefits to hypothesis-driven research versus data-driven research, and both are necessary for the progression of psychology as an innovative and rigorous field.
3. Big Data can be characterized by two key dimensions – its depth (how many measures you collect per individual) and its width (how many individuals you record from).
4. With the large amount of data you can get from a Big Data study, we must be cautious of not "fishing" for effects without accounting for all of the statistical tests that we are conducting.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter.

- **Learn about how Big Data is causing big strides in the understanding of disease:** Heggie, J. (2019, January 8). How can big data beat disease? *National Geographic*. https://tinyurl.com/ykkabh53
- **A cautionary tale on how too many statistical tests can lead to effects that may not be real:** Moore, R. P., Galvin, S. L., & Imseis, H. M. (2006). Thongs, flip-flops, and unintended pregnancy: The seduction of $p < 0.05$. *MAHEC Online Journal of Research*, *1*, 1.
- **Dive deeper into the statistics used to correct for multiple comparisons:** Lindquist, M. A., & Meija, A. (2015). Zen and the art of multiple comparisons. *Psychosomatic Medicine*, *77*, 114–125.

ASSIGNMENT

The purpose of this assignment is to get you thinking about Big Data and how to build out Big Data experiments. Please submit your response in a way so that it is clear what questions and sub-questions you are responding to.

Total Points: 50

1. Pick two psychology papers describing an experiment on a topic that sounds interesting to you. (Do not use a review paper.) They can come from either:
   i) a psychology class you are currently taking, or took in the past;
   ii) a lab you are currently working in; or
   iii) any "Open Access" articles from the most recent year of the journal *Psychological Science* (https://journals.sagepub.com/home/pss).
   Please choose at least one paper that you would consider a "small data" experiment. Provide the citation (titles, authors, year, journal) and abstract of the two papers here: (2 points)

   We will now look at these papers with the frameworks we discussed in this chapter.
2. **For paper 1, answer the following questions**. If the study includes multiple experiments, answer for the first or main experiment:
   a. Is this a **"small data"** or a **"Big Data"** experiment? How do you know? How small/big is the sample size (number of participants)? How small/big is the experiment itself (e.g., number of conditions, stimuli, outcome measures)? (4 points)
   b. Is this a **hypothesis-driven** or a **data-driven** experiment? How can you tell? (3 points)
      i. If this is a hypothesis-driven experiment, what is their hypothesis?
      ii. If this is a data-driven experiment, what new hypotheses come out of their data? How did they avoid p-hacking / data fishing?
   c. Is the data **deep** or **wide** (or both)? How do you know? (2 points)
3. **For paper 1, we will do some more brainstorming on Big Data. (Note that the next part of the question has two options – for a given paper, answer for either small data or Big Data, not both.)**

   If this is a "small data" experiment, we will think up how to make it into a Big Data experiment. Answer these questions:
   a. How **naturalistic** versus **artificial** is their experiment? What are ways in which the stimuli, experiment, or participants are not *representative* of reality? What are ways in which they are? (3 points)
   b. How can we improve the **representativeness** of the study? Use your creativity to brainstorm how you would make this into a "Big Data" experiment. How would you change the experimental paradigm, participant recruiting, the stimuli, or the measurement techniques to capture bigger, more diverse, more naturalistic, and more representative data? (5 points)
   c. What **limitations** could you envision with these changes? These can be limitations in terms of feasibility/practicality (e.g., how much time or money does your change add)? In what ways is your version still not fully representative? (3 points)

   If this is a "Big Data" experiment, we will see how it improves upon small data studies. Answer these questions:
   a. What would the **"small data"** version of the experiment have looked like? (4 points)
   b. **Why** did the experimenters decide to take this "Big Data" approach? What innovations did they apply to make it "Big Data"? (4 points)

c. What **limitations** still exist with their approach? In what ways are the data still not fully representative of real people / images / cognitive processes? What additional improvements could you envision, and how feasible are they (e.g., how much time or money does your change add?) (3 points)

4. Answer questions 2 and 3 for **paper 2** below. (20 points)
5. What are some ideas implemented by paper 1 that could be useful for paper 2 in making their experiments more representative in terms of participants, stimuli, or paradigms? Similarly, what are some ideas implemented by paper 2 that could be useful for paper 1? (5 points)

REFERENCES

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*, 1043–1047.

Bainbridge, W. A. (2017). The memorability of people: Intrinsic memorability across transformations of a person's face. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 706–716.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*, 1323.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 48, 879–894.

Davis, T., & Bainbridge, W. A. (2023). Memory for artwork is predictable. *Proceedings of the National Academy of Sciences USA*, *12*, e2302389120.

Doyen, S., Klein, O., Phcion, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Heggie, J. (2019, January 8). How can big data beat disease? *National Geographic.* https://tinyurl.com/ykkabh53

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323.

Konig, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., & Robert, P. H. (2018). Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, *15*, 120–129.

Martin, C. B., Hong, B., Newsome, R. N., Savel, K., Meade, M. E., Xia, A., Honey, C. J., & Barense, M. D. (2022). A smartphone intervention that enhances real-world memory and promotes differentiation of hippocampal activity in older adults. *Proceedings of the National Academy of Sciences USA*, *119*, e2214285119.

Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., El-Sayed Moustafa, J. S., Visconti, A., Hysi, P., Bowyer, R. C. E., Mangino, M., Falchi, M., Wolf, J., Ourselin, S., Chan, A. T., Steves, C. J., & Spector, T. D. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine*, *26*, 1037–1040.

Moore, R. P., Galvin, S. L., & Imseis, H. M. (2006). Thongs, flip-flops, and unintended pregnancy: The seduction of $p < 0.05$. *MAHEC Online Journal of Research*, *1*, 1.

Mumford, S., & Anjum, R. L. (2013, November 15). Correlation is not causation. Oxford University Press blog. https://blog.oup.com/2013/11/correlation-is-not-causation

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*, 506–519.

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210.

3 Big Participant Samples

Introduction

Many students first gain their interest in experimental psychology as a guinea pig – by volunteering in an experiment on campus. It can be exciting to apply the knowledge you learn from class to guess the main manipulation in a behavioral experiment, or see inside your brain after an MRI study. It can also be a great way to earn money (at least I know I funded all my college vacations by being a "professional subject"!), or you may be required to participate in studies as part of passing a course. This built-in participant sample is also fantastic for researchers: They have a pool of eager young students who are readily available during weekday hours, attending class right next to your laboratory, and often attentive and enthusiastic about each new experiment. However, while it seems like running university experiments with college students is a win-win situation for both the researchers and the students, this convenient choice actually causes concerning damage to the field of psychology as a whole.

In this chapter, we will cover the problems with current norms in the participants we recruit for psychology experiments and how to solve some of these problems by taking a Big Data approach. First, we will go through how small data studies recruit their participants (Section 3.1). We will then talk about how the average college sample differs from adults worldwide (Section 3.2), individuals from smaller societies (Section 3.3), other industrialized nations (Section 3.4), and others even within the same country (Section 3.5). The issues boil down to a difference between our sample and population (Section 3.6), and we will discuss how we can move toward more representative groups using Big Data (Section 3.7). However, we will never be able to make a perfect sample (Section 3.8), and sometimes we may want to intentionally restrict the people we recruit (Section 3.9). The chapter will finish with a look at the big ethical questions surrounding participant recruitment (Section 3.10) and imbalances in the demographics of psychology researchers themselves (Section 3.11).

Big Data in the Psychological Sciences

Wilma A. Bainbridge

Big Data in the Psychological Sciences

Cutting-edge computational tools like artificial intelligence, data scraping, and online experiments are leading to new discoveries about the human mind. However, these new methods can be intimidating to many students. This textbook demonstrates how Big Data is transforming the field of psychology, in an approachable and engaging way that is geared toward undergraduate students without any computational training. Each chapter covers a hot topic, such as social networks, smart devices, mobile apps, and computational linguistics. Students are introduced to the types of Big Data one can collect, the methods for analyzing such data, and the psychological theories we can address. Each chapter also includes discussion of real-world applications and ethical issues. Supplementary resources include an instructor manual with assignment questions and sample answers, figures and tables, and varied resources for students such as interactive class exercises, experiment demos, articles, and tools.

Wilma A. Bainbridge is an associate professor in the Department of Psychology at the University of Chicago. She has won the Association for Psychological Sciences Rising Stars Award (2023), an Alfred P. Sloan Fellowship in Neuroscience (2024), and the American Psychological Association's Distinguished Scientific Award for Early Career Contributions to Psychology (2025). Her research has garnered attention from outlets such as CNN, *Vox*, and *Wired*. She has previously edited two books on vision and memory, and her "Big Data in Psychology" class has earned a Curricular Innovation Award from the University of Chicago.

"From social media to sensors to AI, this book offers a brilliant tour of how the Big Data revolution is reshaping psychology. Accessible, inspiring, and grounded in real research problems, it walks students through everything from hands-on skills like web scraping, to big-picture theory testing, and even thoughtful discussions of ethics – all presented with incredible clarity by one of the field's most inspiring new voices."

Timothy Brady, University of California San Diego

"Exceptionally timely and comprehensive, Bainbridge's textbook deserves a place in every curriculum for behavioral methods. The chapters – enhanced with interactive features and thought-provoking ethical questions – are so engaging that they make me want to teach the course. And whether or not you work with Big Data, this is essential reading for all."

Marvin M. Chun, Yale University

"Combining conceptual depth and accessible writing, Bainbridge offers a timely contribution with a comprehensive overview of the field, covering definitions of big data in psychology and expertly navigating its key sources, methods, and analytical approaches. It addresses both foundational topics, such as neuroimaging tools and statistical techniques, as well as emerging and contemporary discussions, including natural language processing, the development of large language models, and their applications in psychological research. It will resonate with a wide audience, from curious undergraduates to seasoned researchers looking to deepen their understanding of big data and its potential to reshape the psychological sciences."

Nemanja Vaci, University of Sheffield

Big Data in the Psychological Sciences

Wilma A. Bainbridge
University of Chicago

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009343589

DOI: 10.1017/9781009343602

When citing this work, please include a reference to the DOI 10.1017/9781009343602

First published 2026

Cover image: FrankRamspott / DigitalVision Vectors / Getty Images.

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-34358-9 Hardback
ISBN 978-1-009-34357-2 Paperback

Additional resources for this publication at www.cambridge.org/bainbridge

Big Data in the Psychological Sciences

Cutting-edge computational tools like artificial intelligence, data scraping, and online experiments are leading to new discoveries about the human mind. However, these new methods can be intimidating to many students. This textbook demonstrates how Big Data is transforming the field of psychology, in an approachable and engaging way that is geared toward undergraduate students without any computational training. Each chapter covers a hot topic, such as social networks, smart devices, mobile apps, and computational linguistics. Students are introduced to the types of Big Data one can collect, the methods for analyzing such data, and the psychological theories we can address. Each chapter also includes discussion of real-world applications and ethical issues. Supplementary resources include an instructor manual with assignment questions and sample answers, figures and tables, and varied resources for students such as interactive class exercises, experiment demos, articles, and tools.

Wilma A. Bainbridge is an associate professor in the Department of Psychology at the University of Chicago. She has won the Association for Psychological Sciences Rising Stars Award (2023), an Alfred P. Sloan Fellowship in Neuroscience (2024), and the American Psychological Association's Distinguished Scientific Award for Early Career Contributions to Psychology (2025). Her research has garnered attention from outlets such as CNN, *Vox*, and *Wired*. She has previously edited two books on vision and memory, and her "Big Data in Psychology" class has earned a Curricular Innovation Award from the University of Chicago.

"From social media to sensors to AI, this book offers a brilliant tour of how the Big Data revolution is reshaping psychology. Accessible, inspiring, and grounded in real research problems, it walks students through everything from hands-on skills like web scraping, to big-picture theory testing, and even thoughtful discussions of ethics – all presented with incredible clarity by one of the field's most inspiring new voices."

Timothy Brady, University of California San Diego

"Exceptionally timely and comprehensive, Bainbridge's textbook deserves a place in every curriculum for behavioral methods. The chapters – enhanced with interactive features and thought-provoking ethical questions – are so engaging that they make me want to teach the course. And whether or not you work with Big Data, this is essential reading for all."

Marvin M. Chun, Yale University

"Combining conceptual depth and accessible writing, Bainbridge offers a timely contribution with a comprehensive overview of the field, covering definitions of big data in psychology and expertly navigating its key sources, methods, and analytical approaches. It addresses both foundational topics, such as neuroimaging tools and statistical techniques, as well as emerging and contemporary discussions, including natural language processing, the development of large language models, and their applications in psychological research. It will resonate with a wide audience, from curious undergraduates to seasoned researchers looking to deepen their understanding of big data and its potential to reshape the psychological sciences."

Nemanja Vaci, University of Sheffield

Big Data in the Psychological Sciences

Wilma A. Bainbridge
University of Chicago

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009343589

DOI: 10.1017/9781009343602

When citing this work, please include a reference to the DOI 10.1017/9781009343602

First published 2026

Cover image: FrankRamspott / DigitalVision Vectors / Getty Images.

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-34358-9 Hardback
ISBN 978-1-009-34357-2 Paperback

Additional resources for this publication at www.cambridge.org/bainbridge

Thank you to the "RAV4" – Robert, Ally, and Vicky – for being a loving and supportive family! It's hard to believe I began this book when still pregnant with Ally and Vicky and am finishing it as they are running around and chatting our ears off.

Thank you to my mom Erika, dad William, and sister Connie – finally I get to write book dedications for you rather than the other way around!

And thank you to the wonderful Brain Bridge Lab and my department at the University of Chicago – your support has really helped me flourish and think Big.

Brief Contents

Detailed Contents

Preface

Learn how to see. Realize that everything connects to everything else.

Leonardo da Vinci (1452–1519)

We live in a world where we are all constantly generating data – in our interactions with our phones, social media apps, games, websites, fitness trackers, and more. This data is commonly referred to as "Big Data" because its scale is so large that it cannot be analyzed manually. Such Big Data serves as a useful means to understand human cognition – showing us how people see, feel, respond, remember, interact, and make decisions with these different tools. We can also look at these cognitive processes across different groups of people – across countries, cultures, ages, and experiences – as well as across species. As a result, Big Data ways of thinking and analysis have become incredibly important tools to psychologists, across fields. Psychologists are now running online experiments that can gather data from thousands of participants, running machine learning models that can decode patterns from thousands of datapoints, or analyzing brain data from thousands of subregions.

As a result, psychology as a field is at a major transition point. Familiarity with advanced statistical analyses and computer programming is becoming increasingly essential to keep up with the state of the art. However, the idea of wrangling Big Data can be incredibly daunting to people entering the field, especially given that most undergraduate psychology curricula do not require computational or advanced statistical coursework. The main goal of this textbook is to make these new directions in Big Data accessible and meaningful to any psychology student – without the need of training in computer science or statistics. By reading this textbook, you'll gain basic fluency and familiarity with the important topics in the field, so you can decide what topics you want to pursue more deeply. Students who are already familiar with computational methods will learn ways in which these methods can be applied to answer a myriad of psychological questions. As a result, the book will lightly touch upon a wide range of topics, including experimental design, web programming, data scraping, artificial intelligence, different methods in brain imaging, computational linguistics, network science, wearables, user interface design, crowdsourcing, and representative sampling.

To my knowledge, this is the first undergraduate textbook on Big Data in psychology. It was inspired by a course I created in Spring 2020 as a new assistant professor, and I've seen these sorts of courses start to grow in the last few years. Because this is such a new topic, this textbook and course is really for almost anyone. Familiarity with psychology is helpful (e.g., how experiments are run and what are some of the key topics of inquiry), and at some points I will bring up simple statistical concepts (e.g., *p*-values), although knowledge there is not

required. Each chapter focuses on a different angle of how Big Data interfaces with psychology, and includes sections on ethical questions related to the topic and its real-world applicability. Each section also includes thought-provoking questions that can be discussed as a class and an assignment that's relatively open-ended and should engage the students in thinking deeply about that topic. The chapters can be covered in pretty much any order, but the book is generally divided into two parts: 1) how to rethink psychology experiments from a Big Data angle (Chapters 1–7), and 2) various sources of Big Data to enrich the study of psychology (Chapters 8–12).

In conjunction with the Big Data theme, I also want to make this course follow the principles I preach in terms of modernizing psychological research. As a result, this book is paired with an interactive online resource (www.cambridge.org/bainbridge) that includes videos, demonstrations, links, and additional resources that will be constantly updated. This way, you will still have access to the latest developments in the field even after the publication of this book. I also maintain a public data repository on the Open Science Framework of Big Data student projects that came out of the course that I teach at the University of Chicago (https://osf.io/hz843), and I am happy to link to such a repository from anyone else using this book.

Now go forth, and think big!

1 What Is Big Data?

Introduction

The amount of data generated every day is insane. Each day, we create approximately 403 million terabytes of data (or 403 exabytes) (Duarte, 2024). This is about how much data can be stored by 4 billion phones (those of about half the world population) – and just in one day! In that same day, about 300 billion emails are sent, 8.5 billion searches are done on Google, 1.6 billion swipes are made on Tinder, 1.4 billion hours of video are streamed, and $638 million is spent on Amazon. You as an individual are contributing to a lot of this growing collection of data. As you commute to your classes, your map app tracks your movement behavior and may take note of any specific locations you visit. Your phone or watch tracks your steps and sleep patterns. As you look up web pages on your phone, these pages track your browsing and click behavior. And as you scroll through and post on social media, these apps track how you engage with posts, through measures like viewing time and click behavior. We are constantly surrounded by and creating Big Data. This Big Data can be messy and tricky to sift through, but within it are potential insights about the human mind waiting to be discovered.

In this introductory chapter, we will establish definitions of the central themes of this book, to guide you as you read the rest of the book. First, we will talk more about how data has changed in the last few decades (Section 1.1) and then provide a definition of Big Data (Section 1.2). We will then define Psychology within the context of this book (Section 1.3). With these two definitions in hand, we will discuss how Big Data and Psychology interact (Section 1.4) and why now is the perfect time to study this interaction (Section 1.5). Finally, we will wrap up this chapter with a guide on how to use this book and its online resources (Section 1.6).

1.1 Moore's Law

Our data has gotten so *Big* thanks to the exponential growth in processing and storage power over the past handful of decades. This is reflected by **Moore's Law**, which was proposed by Intel cofounder Gordon Moore in 1965 (Moore, 1965). This law predicts that the number of transistors (one of the key components in computer chips) that can be packed into a given

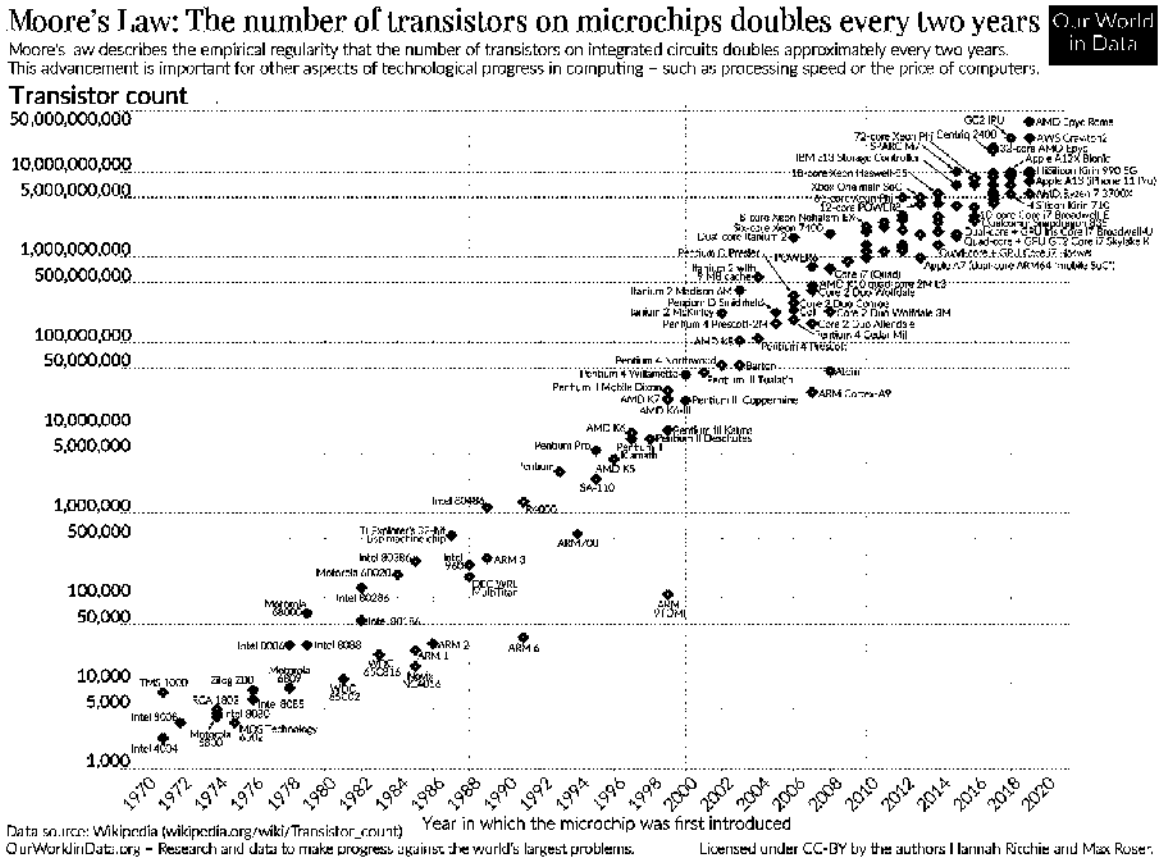


Figure 1.1 A depiction of Moore's law, showing it still holds in 2020. Moore's law posed in 1965 predicts that the number of transistors we can fit in a circuit will double every two years – resulting in exponential growth in our computing capabilities. Note that the y-axis here is an exponential scale (1,000 and 5,000 at the bottom are spaced as closely as 10 trillion and 50 trillion at the top), so indeed, we are keeping up with this law!

unit of space will double roughly every two years. Remarkably, this prediction of exponential increase in computing power has held true for 60 years (see Figure 1.1), although some scientists forecast that we will reach the limit of feasibility within the next few years (Kumar, 2015; Waldrop, 2016). We can feel the effects of Moore's law by looking at how the size of storage devices has drastically changed over our lifetimes.

Discussion Question

What did the size of data (e.g., devices, files, performance) look like when you were a child versus now? Does it feel like there has been exponential growth in that time? What sorts of innovations enabled that growth?

To understand what makes a set of data Big Data, let's first discuss how data is measured. The building blocks of the data and processes in our computers are 0s (off) and 1s (on), and a single digit is called a **bit**. Because of this 1/0 building block, instead of data being measured in our decimal base-10 system, data sizes are measured in **binary**, or a base-2 system, where the only digits possible are 0 and 1. When you want to count in higher numbers in binary beyond 1, you use additional digits (that are still limited to 0 and 1). So the numbers 0, 1, 2, 3, 4, and 5 in decimal are represented as 00, 01, 10, 11, 100, and 101 in binary. While these building blocks seem simple, they can combine to form the complex data we interact with on our computers – just as letters can combine to create the complexities of language. Because of the binary system, powers of 2 end up being important to the measurement of data. A set of eight bits ($2\times2\times2$) is called a **byte**. A byte can be used to represent a single character of text. For example, in the most common character encoding standard for computing called **ASCII (American Standard Code for Information Interchange)**, the letter A is represented by the byte containing the bits 0100 0001, while a space is represented by 0010 0000. Above the level of the byte, the naming of the counting system resembles that of the metric system. A set of 1,024 bytes is called a kilobyte (KB), like how 1,000 meters is a kilometer (but because we are operating in binary, it is a multiple of 2, or 2^{10}). A set of 1,024 kilobytes is called a megabyte (MB). A set of 1,024 megabytes is called a gigabyte (GB). After that, we have terabytes (TB), petabytes (PB), exabytes, and zettabytes. So, for example there are 8,000,000 bits (1s or 0s) in one megabyte of data. When we talk about data transfer speeds (like how fast your internet is), the measures tend to be in bits per second (instead of bytes per second). So early internet modems would have a download speed of 28.8 Kbps, or around 28,800 bits per second.

The rapid change in computing sizes is quite drastic when we look at the history of data storage across personal computing (Figure 1.2). Back in 1956, IMB shipped its first hard drive. It was the size of two refrigerators and could hold 5 MB of data – the equivalent of about one song. In the 1970s, some consumers were starting to get their own computers, and the most common way to store and transfer files was through floppy disks. These could only hold about 100 KB in early years, and 1.44 MB in later years, the equivalent of a few text documents or pictures. However, this medium became so ubiquitous that many pieces of software still use an icon of a floppy disk as their "save" icon. Once software became more advanced, users needed more and more of these disks – for example it took seven floppy disks to install an early version of Adobe Photoshop (Adobe, Inc., 2013).

In the late 1980s, a more advanced data storage method emerged – the CD-ROM (compact disc read-only memory). These could hold as much as 900 MB – about one-third of a movie. However, as their "read-only" name implies, these needed special devices called CD burners to write data to the CD-ROM, and most CDs could not be rewritten once data was saved onto it. In the mid-1990s, we moved onto DVDs (digital video disks), which could store closer to 5 GB (about 1–2 movies) but faced similar shortcomings as CD-ROMs. The early 2000s saw the first USB (universal serial bus) flash drives, which were smaller and more convenient but could only store about 10 MB at first. The early 2000s also saw the explosion of the internet, and **cloud storage** – saving data to online servers – started to grow. This really took off as internet speeds became faster, and websites emerged dedicated to hosting large amounts of data – YouTube for video started in 2005, Dropbox for files started in 2007, and Flickr for photos started in 2004.

Figure 1.2 Photos of various types of older data storage. **(a)** The shipping of IBM's first hard drive, holding 5 MB and taking up the size of two refrigerators. **(b)** A 3.5-inch floppy disk, the main external data storage format in the 1980s and 1990s. **(c)** A CD-ROM inserted into a laptop's CD drive. These were commonly used in the 1990s and 2000s for data storage and were the main medium for holding music albums.
Source: (a) Michael de Groot / Flickr. (b) Pablo Jeffs Munizaga – Fototrekking / Moment / Getty Images.
(c) EThamPhoto / The Image Bank / Getty Images.

In the 2010s, storage amounts and data transfer speeds increased by many magnitudes. Personal computers could hold close to a TB of data, and mobile technologies emerged enabling people to generate vast amounts of data through the photos and videos they're taking. In the 2020s, higher mobile data speeds and faster personal computer processors are allowing us to interact with vast amounts of data and at faster rates – supporting growths in avenues like online gaming, video streaming, and real-time artificial intelligence (AI) on board many of our systems.

1.2 How Do We Define Big Data?

Owing to this constant growth in technology, the definition of Big Data is a moving target. In the 1990s, even a low-resolution video would be considered Big Data, while in the 2020s, Big Data is more in the order of libraries of thousands of movies, representing petabytes of data (or even more). A fundamental aspect of what makes data qualify as Big Data is that it is so large that we cannot process it by hand – we can't manually input, clean, or analyze the data. This is data that is also often so large that we cannot process it with basic programs on our computer (e.g., Microsoft Excel), but usually have to use code or bespoke tools. For our purposes of studying human cognition, Big Data tends to be more naturalistic – recorded from real people or dynamic behaviors – so it may be unstructured or more noise- or error-prone compared to smaller datasets. Data can be Big for several different reasons. It may have very high temporal sampling – for example, taking a measurement every handful of milliseconds. It may have high spatial sampling instead – for example, collecting data across all the intersections in a city. It may have high participant sampling – collecting data from a large and diverse set of people. Or, it may have high stimulus sampling – capturing data from lots of sources (e.g., images, videos, news articles, products) or tasks. All of these encompass examples of Big Data.

So here, we will loosely define Big Data as: *unstructured, naturalistic human data requiring complex analytical methods*. For some of the exercises and examples we discuss in the book, we may not use the massive amounts of space or computing power that traditionally make up Big Data. But the concepts that you learn here should translate to bigger sets.

1.3 How Do We Define Psychology?

It is also important that we define what psychology is within the framework of this book. Broadly, psychology is *the empirical study of the mind, brain, and behavior*. For the vast majority of this book, we will focus on a more quantitative and research-based approach to psychology, where psychologists conduct experiments that aim to provide broad insight, using falsifiable questions and hypotheses. This is in contrast with some psychologists who use more *qualitative* approaches, like revealing new insights using interviews or observations, or using therapeutic techniques like talk therapy to help improve mental health. A central aspect of the psychology we will discuss here is that it is composed of research questions that are testable and falsifiable. **Falsifiable questions** are those where you can obtain evidence to prove that question wrong. For example, one active area of debate is the degree to which we can falsify questions in **evolutionary psychology** – the study of the mind, brain, and behavior through the perspective of evolution (Gannon, 2002). We cannot go back in time and run experiments on our ancestors. We also cannot measure how much of people's current behaviors are a result of evolution versus more recent societal norms. There are some creative scientific methods to test evolutionary hypotheses by looking at other animal species or running computational models. However, we will generally avoid tackling unfalsifiable topics in the current book.

We are also interested in questions that are **generalizable** – that give us insight into the thinking of a group of people, and can allow us to make predictions in the future about different events. For example, a question like "how did people feel about Argentina winning the 2022 World Cup?" is a question that measures emotions and behavior, but is so highly

specific that it does not really teach us about the human mind. Thus, we would not characterize this as a psychological question. A more generalizable psychological question might be something like "how does sentiment in social interactions change directly after major sporting events?" Sometimes when dealing with Big Data, we may accidentally take an overly narrow scope, due to the data that we have available (like data from a specific app, Chapter 8). But we should always try to focus on a big-picture question about the mind, and any limitations to the data we are collecting to answer this question.

There are many different branches to the field of psychology. When you think of a "psychologist," your mind may first go to **clinical or abnormal psychology** – the study of atypical behavior and mental health, with the goals of diagnosis and treatment. Closely related is counseling psychology, which is the practice of helping people through therapy and counseling. While this book will discuss some research on abnormal psychology through the lens of understanding the underlying roots of an impairment, we will not discuss in depth therapies or treatments of individuals. **Industrial psychology** is the study of the mind within the workplace, and how to optimize people's effectiveness at work. As this field is more applications-focused, we will not discuss this at length in this book, nor other more applied fields like forensic psychology, school psychology, and health psychology. The major focus of this book will be **cognitive psychology**, the branch of psychology dedicated to the scientific study of our internal mental processes. This encompasses a broad range of processes, including sensation, perception, action, memory, reading, speaking, emotion, decision-making, morality, imagination, and others. In fact, a "psychologist" can be a researcher with a laboratory that runs experiments to study these processes (this is the type of psychologist I am). Related to cognitive psychology, we will sometimes discuss the brain, through a lens of **neuropsychology** – looking at how the brain and mind interact. We will also bring up many examples from **developmental psychology** – the study of the development of the mind across the lifespan (from infants through aging) – and **social psychology** – the study of the interactions of multiple minds.

1.4 How Do Big Data and Psychology Interact?

A large proportion of Big Data out in the world just *happens* – you record a video and post it on social media, and now there are several new megabytes of data on your phone, on a server belonging to that social media site, and being downloaded to other people's phones. In this way, much of Big Data is just passively accumulated as we perform tasks with our phones, computers, and the internet. Another major slice of Big Data is being actively collected by companies, where they are testing how they can improve your experience, how you navigate their app, and how they can improve engagement and purchasing. However, the data being generated out in the world also serves as incredibly rich records of human behavior that can give insight into questions on almost any topic of psychology.

Discussion Question

What are some ways you can envision Big Data might be changing the types of questions we can ask or answer in psychology?

An important skill for you to nurture will be in identifying these intersections of Big Data and psychology. What is a psychological question you want to answer, and how could Big Data answer that question? Could there be a preexisting dataset out there that answers the question for you, or could Big Data help you collect that data in some way? For example, a few years ago, I was curious how older memories (2+ year-old memories) might be represented in the brain. This is hard to test in the laboratory because I would need to have participants study some images and then come back two years later. But then it dawned on me that people are constantly capturing their memories on social media, dating back to many years prior. So, I collaborated with the app 1 Second Everyday to recruit users who had recorded years of their memories, and then I scanned their brains while they viewed these older memories. Long story short, we found patterns in the brain reflective of the age of a memory (Bainbridge & Baker, 2022; see Section 8.4.2). As you look through data in your daily life, think to yourself – what does this reflect about the human mind and can it show us something new? And, are there ways in which Big Data technologies are influencing how we think or interact? For example, an active area of current research is how social media may be impacting feelings of isolation and depression (Section 12.3.2). Overall, a major part of this class will be thinking creatively and with an open mind on how we can use data to answer questions.

1.5 Why Study This Now?

Computation and psychology are both at points of incredible transition right now. On the technological side, we are generating more data than ever, but tools to process this data are also starting to become more accessible to the average person. There are notably five main changes that have occurred with computing technologies that have enabled data to become so big. First, as we have discussed, there have been drastic improvements in cheap, large data storage in small form factors. This means that the average person has on their phone or computer tens of thousands of files, documents, images, videos, and pieces of software. This also means there are places where we can easily save our big datasets. Because these storage devices are getting smaller, we can have large amounts of storage in small devices, like phones. Meanwhile, cloud storage allows people to maintain massive amounts of data that they can access with a multitude of devices. Second, there have been major improvements in faster and cheaper processing power, such as the explosion in parallel processing graphics chips. For example, the average processor in a consumer computer can make about 150 billion calculations per second. This allows us to analyze big datasets relatively rapidly, and even in real time as we acquire it. Third, sensor technology and speed has also improved – most people have high-quality cameras in their phones, and may have devices (like fitness trackers) that can record movement, heart rate, elevation, skin conductance, and other measures. This allows us to obtain big physiological data, which can reveal underlying information about one's cognitive state (Chapter 9). Fourth, the wide spread of high-speed internet both in homes and out in the world is allowing more people to form communities, creating large amounts of data generated by people's interactions online. Fifth, our algorithms are getting better and smarter – we are able to compress data more efficiently and analyze data more effectively with tools like artificial

intelligence. The combination of these five computational improvements has led to an explosion of data produced by and accessible to the average person.

Big Data is also more important than ever for psychology research. Psychology has always been a multidisciplinary field, straddling social science and biological science programs at many universities. For example, psychology has clear links to neuroscience and experimentation, but also has implications for therapeutic practice and philosophy. However, recently, psychology as a field has begun to undergo a transition, with greater emphasis focused on experiments and complex analyses. Many exciting discoveries are coming about thanks to Big Data innovations, such as online experiments, artificial intelligence, and rich physiological data. With these innovations, researchers have been able to revisit classical psychological questions with a Big Data lens that allows them to assess their applicability across more diverse samples or make computational models that can predict people's behaviors. These innovations have also been saving psychologists a lot of time – making it faster to collect and analyze data. These changes go hand in hand with a new global scientific community that is developing, based around sharing data and code openly, in reaction to a "replication crisis" that emerged around unreplicable findings in small-scale experiments (see Section 2.1.3). So now is the perfect time to learn about these changes in psychology, to ride its waves as it moves into these new approaches.

Discussion Question

What topics relating to Big Data and psychology are you particularly excited to learn about in this book, and in your class?

1.6 How to Use This Book

As we just discussed, psychology is changing. If you want to go into psychological research for your career, professors and laboratories are now increasingly looking for candidates with experience in programming and statistics. Outside of academia, many jobs after college geared toward psychology majors – such as user experience design or data science jobs – also require these skills. For those of you wanting to practice clinical psychology, counseling, or go into education, it is still helpful to be up-to-date with the latest research and techniques (e.g., how is artificial intelligence changing the diagnosis of neuropsychological disorders?). And I would argue that some of these topics we will discuss in this book can help improve your daily life. I know for me personally, I've coded data scraping tools to find the best flights for a vacation, used generative AI to make a personalized storybook for my kids, or analyzed my fitness tracker data to get a sense of whether a diet is working. Knowing what is possible with data can change how you look at and use data in your daily life. In this book, we will also touch on some of the hot-button topics that have erupted in the news and the legal sphere as a result of Big Data – how do we deal with AI-generated fake information? How do we navigate the privacy risks created by the data recorded in many mobile apps and websites?

It can be intimidating jumping into learning about data and computer programming if this is your first foray into the topic. My number one goal is to demystify these topics and make you comfortable talking about them and thinking about them. As a student starting along this journey, it can feel like there's a big gap between you and your image of a computer scientist who may have been hacking computers since they were in elementary school. It can feel like you just aren't meant to be someone who codes or does complex math. But really these thoughts are a part of a mystical (but inaccurate!) aura that has surrounded computation. I'd liken computer programming to something like learning a foreign language or training for your first 5 km run. Most of the time, your goal isn't to become completely fluent or a record-breaking marathon runner. Usually, it's that you find these skills useful and enjoy the process of getting there. You also usually aren't worried about how you compare to the pros – you don't feel bad comparing your Spanish skills to those of a native speaker (and often they are impressed that you are trying!), or feel bad watching Olympics runners beat your time. In the same way, a seasoned software developer won't be quizzing you on the latest Python functions. You will also find that gaining these skills can enrich your daily life – you can now navigate a little around Madrid with your newfound Spanish skills, or be able to run to catch a bus without getting winded. Similarly here, you'll have moments where you may wish you could do something on your computer in an automated way and then realize there may be a way to use your skills learned here to do that!

At the same time, this book is not going to teach you programming or statistics from the ground up. It's the first step in learning the lingo and giving you the lay of the land, so you can then decide where you want to do a deep dive in future classes or explorations (e.g., do I want to study more neuroscience? Or web design? Or graph theory?). The online resources with this book will provide some stepping stones for doing these deep dives. With that foreign language metaphor, this book is your travel guidebook to help you decide where you want to study abroad. Then once you've picked a country, you can start focusing on learning its language. With this book, I want you to become fluent in the topics of new technologies being widely used in psychological research. I want you to have an increased level of agency over your own data and how it is used by companies and researchers. And, I want you to practice thinking creatively about psychological research questions and how we can answer them.

This book can be read from front to back or you can skip around sections as needed. In these first two chapters, I introduce what Big Data (Chapter 1) and small data (Chapter 2) are and how they compare to each other. Then for the rest of the first half of the book, I will give you the building blocks for running Big Data studies – looking at the participants (Chapter 3), the stimuli (Chapter 4), and the experiments themselves (Chapter 5). Once we are armed with our Big Data, we can then analyze it using artificial intelligence (Chapter 6) or human crowdsourcing (Chapter 7). In the latter half of the book, we will delve into different topics that are changing as a result of Big Data, and so these chapters are a bit more standalone. We will talk about software developments with apps and games (Chapter 8), as well as hardware innovations and physiological sensing (Chapter 9). We will talk about Big Data in neuroscience (Chapter 10), language and natural language processing (Chapter 11), and wrap up with social interactions and graph theory (Chapter 12).

With the exception of this chapter, each chapter will end with four key sections. In "Big Ethical Questions," we will talk about the ethical implications of the topic discussed in the chapter. These topics are sure to spark interesting discussion, especially because the ethical implications of these new methods are still being actively addressed in science and society. This section is then followed by a section on "Applications of the Chapter." While most of this book takes the framework of theory-driven psychology – where we are conducting experiments for the sake of understanding the mind, not creating a product – in these sections, we will discuss how the chapter's topics can be applied to impact the real world. Each chapter then ends with a Chapter Summary that reminds the reader of the major points, and Further Reading which suggests further sources to explore if you are interested in going beyond the pages of this book. There will be discussion questions laced throughout the chapters, as well as a sample homework assignment at the end of each chapter. The companion Teacher's Guide will include additional discussion questions, exercises, and demonstrations for each topic.

Importantly, data, computation, and the internet are always changing. While this book is written at a static point in time (2022–2025!), there is an accompanying online resource (www.cambridge.org/bainbridge) that will be updated as technologies change in the world. If you read anything in this book that seems outdated, check out the online resource to see if there is a new version of that information. The online resource also has interactive demos and programming tools to let you learn more about programming and test out online experiments.

With that, let's proceed to Chapter 2 to discuss what "small data" is and how that differs from Big Data in the context of psychological research.

CHAPTER SUMMARY

In this chapter, we introduced the concepts of Big Data, psychology, and how now is the perfect time to study them and their interactions.

1. Here, we define Big Data as unstructured, naturalistic human data requiring complex analytical methods.
2. We define psychology as the empirical study of the mind, brain, and behavior. This book mainly focuses on quantitative experiment-based psychology.
3. With major improvements in our technological capabilities over the past few decades and changes in the landscape of psychology, now is the perfect time to study how Big Data can be used in psychology research.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter.

- **Read about and watch an original video describing the world's first hard drive, developed by IBM in 1956:** Seeley, C. (2014, October 28). History snapshot: 1956 – the world's first moving head hard disk drive. Data Clinic Ltd. News. www.dataclinic.co.uk/history-snapshot-1956-the-worlds-first-moving-head-hard-disk-drive

- **A review of how realism in our studies can actually show differences in the brain:** Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*, 506–519.

ASSIGNMENT

The purpose of this assignment is to learn more about your experience with Big Data and provide you a sense of the data you generate.

Total Points: 50

1. **Fill out the class survey.** Your professor will provide a link. (20 points)
   Let's look at how much data you are generating just from your phone!
2. Locate where your phone describes your storage usage. **Answer:**
   a. How much storage are you using for images? (1 point)
   b. How much storage are you using for videos? (1 point)
   c. How much storage are you using for music? (1 point)
   d. How much storage are you using for apps/applications? (1 point)
3. Let's get a rough estimate of how much data you are generating a day with your phone camera.
   a. Add together your answers from 2a and 2b and **report that number here**. (2 points)
   b. Get an estimate of how long you have had your phone – find the date of the first photo you took. Then search on Google "how many days between [that date] and today" and it should return you the number of days. **Report that date of the first photo and the number of days since then.** (2 points)
   c. Divide your answer in 3a by your answer in 3b and **report that number here.** This tells you about how much data you are generating with your camera per day. (4 points)
   d. **How many bytes of data is that?** (2 points)
   e. One byte is the amount of data used to type one character (e.g., "A"). A novel contains about 500,000 characters. **How many books worth of data is that?** You're likely creating the equivalent of books of information a day! (3 points)
4. Let's see how long you are using your phone for.
   a. First, guess: **How much screen time do you think you use a day?** (2 points)
   Now, locate where your phone describes your screen time usage (this might be under a "Screen Time" setting or a "Digital Wellbeing" setting).
   b. **On average, how much screen time do you actually use a day? How does this compare to your guess?** (4 points)
   c. Based on your screen time report, **on average how much screen time do you spend on social media a week?** (2 points)
   d. You use on average 500 MB of data per hour by browsing social media. **How many GB (or MB) of social media data are you viewing per week?** (5 points)

As you can see, we interact with massive amounts of data in our daily lives!

REFERENCES

Adobe, Inc. (2013, August 1). Did you know that Photoshop 3.0 was the last version of Adobe Photoshop to be sold on the floppy disc? Facebook. www.facebook.com/photo?fbid = 10151614431968871&set = a.468676338870

Bainbridge, W. A., & Baker, C. I. (2022). Multidimensional memory topography in the medial parietal cortex identified from neuroimaging of thousands of daily memory videos. *Nature Communications*, *13*(1), 6508.

Duarte, F. (2024). Amount of data created daily. Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day

Gannon, L. (2002). A critique of evolutionary psychology. *Psychology, Evolution & Gender*, *4*, 173–218.

Kumar, S. (2015). *Fundamental limits to Moore's law*. arXiv:1511.05956.

Moore, G. E. (1965). Cramming more components into integrated circuits. *Electronics*, *38*(8).

Waldrop, M. M. (2016, February 9). The chips are down for Moore's law. *Nature* [news feature]. www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338

2 What Is Small Data?

Introduction

In order to learn about Big Data, you first need to understand its counterpoint, "small data." Small data isn't often called this, because data from most psychology studies fits under this umbrella, and many times its scale can suit our purposes just fine. Thus, a definition of small data would be any data that isn't Big Data. While it is incredibly common, solely using small data severely limits the takeaways we can get from psychological research. In this chapter, I will discuss the limitations of small data, as well as the limitations of Big Data. You will see how the two can work in synthesis to pinpoint the rich phenomena occurring in our minds and brains.

Specifically, first to understand the benefits we gain from Big Data, we will go through a few example small data experiments (Section 2.1.1) and see what they are lacking (Section 2.1.2 and Section 2.1.3) and how they can be made bigger in scale (Section 2.1.4). We will then discuss some limitations to Big Data experiments (Section 2.2), including new issues they introduce (Section 2.2.1) and the limitations that will always be present with any study (Section 2.2.2). We will then discuss how experiments can be dichotomized into being hypothesis-driven or data-driven (Section 2.3), as well as how Big Data studies can be characterized as deep or wide (Section 2.4). We will discuss the ethical issues that can come about from the multiple analyses run with Big Data (Section 2.5). We will then discuss applications of the topics in the chapter, such as examples of famous data-driven discoveries (Section 2.6.1) and medical applications of deep and wide data (Section 2.6.2).

2.1 Turning a Small Data Experiment into a Big Data Experiment

Let us first begin with an example of a small data experiment and think about how we can make it bigger and broader.

2.1.1 A Case Study

There is a famous effect in psychology called the **own-age effect** (Anastasi & Rhodes, 2005), where people tend to remember faces close to themselves in age better than faces farther in age.

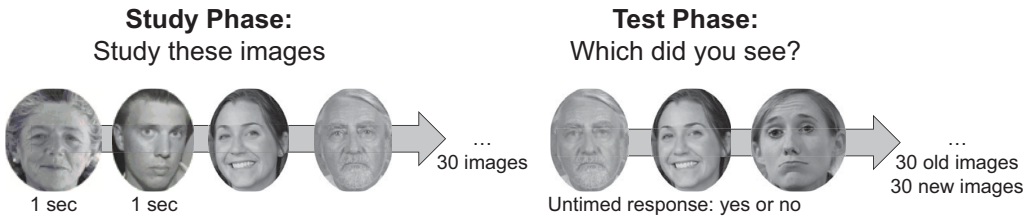


Figure 2.1 The experimental methods for our own-age effect experiment. We have participants first study thirty face images for 1 second at a time. Half of the face images are from older adults and half are from younger adults. We then test them where we show them thirty of the face images they saw, randomly mixed up with thirty new face images they didn't see. For each face they have to respond if they saw it before ("yes") or not ("no"). Our main research question is whether participants have different levels of memory accuracy based on the match between their own age and the age of the face images.

You may have experienced this before, where you may have an easier time recognizing your classmates than professors on campus. (There are many memory effects driven by the similarity of a face to your own – there is also famously an own-race effect; Chiroro & Valentine, 1995). Let's say we are in a traditional psychology lab, and we are running an experiment to test the own-age effect. Our experimental methods look something like what is in Figure 2.1.

The idea is to recruit fellow psychology students on campus and run them through a face memory test on the computer in the lab (as most psychology experiments are done!). In that memory test, we will show a series of thirty faces (like in Figure 2.1), where half are college-aged, while the other half are older adults. We will then test to see if there is a significant difference in memory for those two groups of faces. After running twenty participants, we find a significant effect – indeed the own-age effect holds true!

Discussion Question

What prevents us from generalizing these results to saying that the own-age effect occurs for all observers and all faces?

2.1.2 What Does a Small Data Experiment Miss?

The previous case study was an example of a typical psychology experiment. However, there are many aspects of it that prevent us from generalizing to all observers and all faces. Specifically, the participants, the stimuli (the face images), and the experiment itself are all constrained and artificial in some way.

Small Participants: First, the number and scale of the participants is "small" – can we really make generalizations about humans as a whole from an experiment run with twenty students at a specific university? For example, would these effects replicate for people who are frequently exposed to faces of other ages – like in cultures where young adults tend to live with older

generations? In Chapter 3, we discuss more about the problems with using small college samples in a large proportion of psychology studies, and what we can do about it in the field.

Small Stimuli: Second, the images are also very small in scale. Like the issue we have with participants, can thirty faces really capture the rich variance of human faces out in the world? If you look at the paradigm (Figure 2.1), all these faces are very homogenous. They are all front-facing white people of moderate attractiveness with an oval cropped around their face so you cannot see much of their hair or clothing. This can sometimes be intentional – researchers often want to control for factors they're not interested in, so that those cannot be alternate explanations of their effect. For example, you don't want to think there is an effect of age on memory when it's actually the clothes the models are wearing (maybe clothes from a few decades ago are more memorable than clothes from today!). But, too much control will limit our ability to make generalizations across different lighting, viewpoints, and facial expressions – it doesn't let us make confident predictions about memory out in the real world. And, by only doing research on constrained demographics (e.g., all white people), we aren't studying the rich variation in human experience. In order to generalize to the real world, we need images that better capture the diversity we observe in that world. In Chapter 4, we talk more about how to think about and create more representative stimulus sets.

Small Experiment: Even in just the way they are conducted, experiments are much smaller in scale than the real world. They don't capture what it's like to meet a moving, emotive, multisensory human being, and try to encode them into memory. Experiments tend to be brief (usually 30 minutes to an hour) and constrained to a two-dimensional computer screen, with a few seconds to see each face. This is not at all what it's like to meet a face in reality – you see them out situated in the real world, and you may spend hours interacting with them. Perhaps the dynamic, moving aspects of a face can contribute to your memory for that face, and that would be completely ignored by the experiment. Or perhaps seeing faces in the three-dimensional world is fundamentally different for memory than seeing them on a flat, two-dimensional screen in an experiment. (Although seeing faces in two dimensions may be becoming more natural, as virtual meetings are becoming more common.) Also, because faces are so dynamic, it's unlikely in the real world that you will ever see the exact same view of a face again; you can never take the exact same photograph twice. The second time you see a person, their facial muscles will be engaged in a slightly different way, the lighting will hit their face differently, or they may have a slightly different glow to them. This is completely different from an experiment which shows you the exact same photograph twice.

2.1.3 A Second Case Study and the Replication Crisis

Let's examine another sample experiment. Within the field of social psychology, one phenomenon that has been proposed is the phenomenon of **social priming**. The idea with social priming is that when you are made to think of a social category, you automatically think about related behaviors and stereotypes and start to subtly behave in a similar way. This was first demonstrated in a study by Bargh and colleagues (1996) across a series of experiments. For example, in one experiment, thirty psychology class undergraduates from New York University were asked to complete a task where they had to take a set of five words and create a grammatically

correct four-word sentence as quickly as possible. They did this for thirty sentences in total. Unbeknownst to those participants, half of them received words specifically related to being elderly – *old*, *grey*, *sentimental*, *bingo*, *wrinkle* – while the other half received neutral words. The idea was that the elderly-related words might prime them to think about elderly individuals and act in a similar way. An experimenter then secretly timed how long it took participants to exit the hallway leaving the testing room. The researchers found that participants primed to think about being elderly had a significantly slower walking speed (8.3 seconds to travel the hallway) than participants given a neutral prime (7.3 seconds), confirming their hypothesis. Participants reported not being aware of this elderly manipulation, or a change in their behavior, suggesting these social priming effects could happen unconsciously.

Discussion Question

What factors in this experiment might prevent us from generalizing more broadly?

This experiment uses a relatively small number of participants (fifteen in the elderly prime condition) and stimuli (thirty), making the robustness of the effect unclear (though the original experimenters do actually replicate this effect in a second thirty-participant experiment). The participants come from a very specific sample – psychology undergraduates in the New York area – who are unrepresentative of the world population. The words are also not validated as conjuring an image of "the elderly" in an objective way. The study does a fairly good job at using a naturalistic task (i.e., measuring walking time). However, there could be modern improvements on how it is measured, rather than relying on an experimenter's timing skills, which could introduce a subtle bias that accounts for the 1-second difference between conditions. As a result of these critiques and others, Doyen and colleagues (2012) ran a larger-scale replication of the experiment. They ran 120 participants (albeit also from a fairly specific sample – Belgian French-speaking undergraduate students). They used elderly word stimuli that were first confirmed by a separate set of eighty participants as representing old age. The experimenters then used infrared sensors to precisely measure the amount of time it took to traverse the hallway. With this "bigger" experiment, researchers found no difference in walking speed between their two participant conditions.

Around the same time (in the early 2010s), many psychological findings were unsuccessfully replicated. Researchers were failing to find clear evidence for many social psychological phenomena that had become well-accepted – in addition to social priming, there was now evidence against ideas like ego depletion (the idea that willpower is a finite resource) and power posing (that standing in a certain way will increase your confidence) among others. This launched a "**replication crisis**" across the field of psychology bringing into question the quality of the research in the field. One event that ignited this crisis was when a paper was published in one of the most revered social psychology journals (*Journal of Personality and Social Psychology*) claiming evidence that people can see the future (use "precognition"; Bem, 2011). Researchers realized that a combination of poor research practices as well as publication pressures in the field (see Section 4.4) was overinflating the reporting of supposed "results" across many papers.

At this major breaking point, hundreds of researchers as part of the Open Science Framework launched an effort to attempt to reproduce a hundred findings in psychology. Shockingly, only thirty-six were successfully replicated (Open Science Collaboration, 2015). This served as a reality check for psychologists – we need to run experiments with larger samples, more generalizable experiments, and better statistical measures. We also often should run multiple replication experiments to confirm our effects really hold, and aren't just occurring due to chance.

2.1.4 Making an Experiment Big

Even if I have convinced you that traditional psychology experiments are often unnatural simulations of the real world, how can we improve upon this? How can we make our experiments "bigger"?

> **Discussion Question**
>
> What would a Big Data version of the example face experiment look like? What would you change?

We need to think about how we can improve upon the three points mentioned earlier: the participants, the stimuli, and the experiment. For the participants, can we recruit more people, and more widely? In Chapter 5, we will talk about how to conduct online experiments, which lets you reach thousands of people, with more diversity than the average college campus. For the stimuli, we can also strive to collect image sets that are larger, more natural, and more diverse (refer to Chapter 4 to learn how we can do that!). For making the experiment more naturalistic, there is the difficult balance of wanting scientific control but also generalizability. If you want to keep it as a computerized task, what if you test people on memory for a face across different photographs of that person, rather than memory for a specific image? (It turns out recognizing an unfamiliar person from different photographs is a very difficult task! See Jenkins et al., 2011). New technologies are also making it easier to conduct experiments in more dynamic, three-dimensional environments like virtual reality, or even out in the real world (e.g., Snow & Culham, 2021; Võ et al., 2019). If we are able to expand our experiments out in these three ways, then we have a more generalizable study in these three ways as well. We can know things about face memory across a wider range of observers and faces being observed, and we can try to make predictions about behavior out in the real world.

For example, in our lab, we were curious about people's memory for faces more generally than the own-age effect. So, we generated a large database with demographics matching the United States (see Chapter 4 for more information). We then had over 800 diverse individuals engage in a face memory experiment online (Bainbridge et al., 2013). In this experiment, people viewed a stream of face images and pressed a key when they recognized a repeat from earlier (called a **continuous recognition task**) – a little like the experience of walking through a crowd and recognizing some people as you pass them. We also ran a version of the experiment where we tested people's memory for faces across different viewpoints and facial expressions,

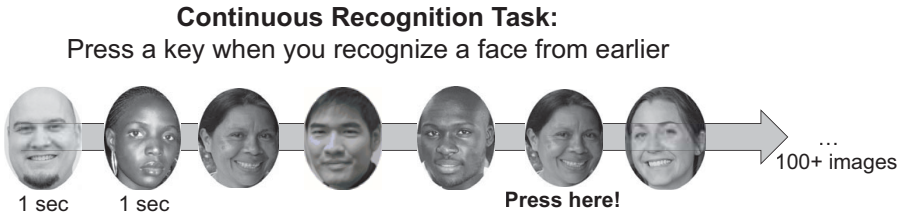


Figure 2.2 The experimental methods for our more "Big Data" face experiment. People still view face images for 1 second at a time. However, now they are continuously seeing images and indicating their memory as part of a "continuous recognition task," akin to the experience of walking through a stream of people and sometimes recognizing someone. We now also have a much more diverse range of faces, so we can test many phenomena, such as the own-age effect or the own-race effect, as well as measure the intrinsic memorability of a given face image. We would also test this task with many diverse participants, so we can look at more generalizable effects of the viewer, too.

not just face images (Bainbridge, 2017). So, our experiments looked a little more like Figure 2.2. What we found in the end was that there are certain faces that are remembered very well by most people, and some faces that are remembered very poorly. In other words, faces have an intrinsic *memorability*. We have recently extended these findings to images more generally, and tested them out in the real world – for example, discovering that we can even make predictions about what paintings people will remember in a freeform visit to an art museum (Davis & Bainbridge, 2023).

2.2 Limitations of Big Data

It may seem obvious that we'd want to aspire for diverse, generalizable experiments as described. However, there are some obstacles presented by Big Data experiments that are important to consider.

2.2.1 Problems with Big Experiments

One of the major cons of Big Data-style experiments is that they can introduce noise into our data. First, sometimes the images can vary too much. If each image is different along many dimensions (age, gender, attractiveness, facial expression, lighting, angle, hairstyle, eyebrow width, etc.), then how can we pinpoint which specific dimension is causing the effect we're studying? And maybe many (or even a majority!) of these dimensions might be things we don't care about, like lighting. Similarly, if our participants vary too much, we can have the same problem – everyone may act in a unique way that prevents us from finding generalizable effects. And participants may vary in ways that are not interesting to us – for example, in running an online experiment, what if the participant's screen monitor resolution, or what they ate for breakfast that specific day, or the length of their fingers influenced how quickly they pressed keys on a task? Small experiments let us directly test our effect of interest, without having all these

extraneous factors in the way because we control for them as much as possible. So, sometimes, we have to design our Big experiments in a way that they're not *too* big. At the same time, Big Data experiments let us simultaneously look at the contributions of many factors – for example, we can analyze how gender, age, and race all contribute to face memory, at the same time.

2.2.2 Imperfect Experiments

Even when we want to run Big Data experiments, there will always be limitations that make it impossible to run a perfect experiment. There will inevitably still be biases with the stimuli and participants in the experiment, because you cannot make everyone in the world participate in the experiment, or make everyone agree to be photographed as a face in the experiment. There are some factors limiting who can be a participant in your experiment – participants must have some familiarity with how to read a computer, and they have to have free time and interest in participating over other things they could be doing. Similarly, only a subset of people would be okay being photographed for the study, and any set of natural photographs will likely have an over-representation of happy or neutral facial expressions (over angry or sad).

There are also some other practical limitations with Big Data. Sometimes the data is so big that we are limited by the processing power, storage, or internet speeds that support us saving and analyzing the data. For example, one person's MRI brain data can take up 1 terabyte of space, which is more than the amount of space many computers come with (in 2025). It can also take half a day to download this data for just one person! So, it can be difficult to analyze data from hundreds, let alone dozens, of participants. Large-scale experiments can also be very costly with time and money. Using the same example of an MRI experiment (which is on the upper end of what psychology experiments cost), one participant usually lies in the scanner for about 2 hours, and it will cost the researchers around US$1,000 to the scanner center for that time. So, an experiment with 100 participants would end up costing $100,000 and take 200 hours of the researchers' time to just collect the data. We are also still limited in our analytical techniques for Big Data. When dealing with very big, naturalistic data, we often don't look at just a single measure or statistic. But, at the same time, our statistical tools and artificial intelligence are not yet able to fully interpret natural human behavior. For example, let's say we wanted to look at face memory in the real world, and recorded participants' view as they navigate through a party, using some sort of head-mounted camera. It's not clear how we would analyze these data – how to turn the conversations with people, the amount of time looking at them, the thoughts related to them, etc. – into numbers in order to make conclusions about what influences someone's memory of a person. So, our analytical techniques are limited (and in fact, they are dependent on the study of psychology to guide us on how to analyze such complex human behaviors).

2.3 Hypothesis-Driven versus Data-Driven Research

A majority of psychology experiments can be characterized as **hypothesis-driven research**. These are experiments where the researchers have one or a few key research questions. They also tend

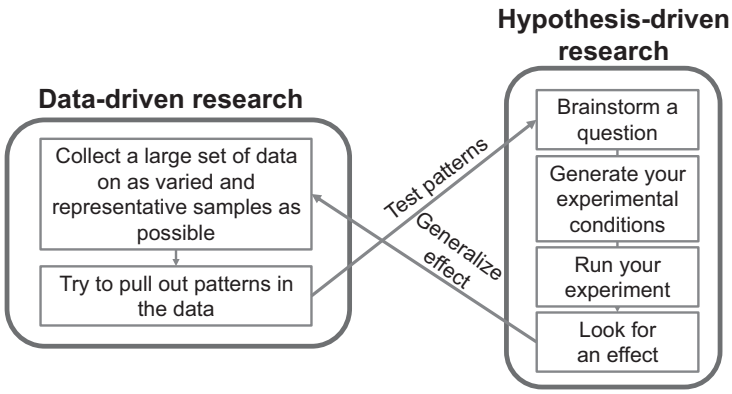


Figure 2.3 The series of steps you take when conducting data-driven research and hypothesis-driven research, and the way in which they interact. In data-driven research, you collect a large, representative set of data and identify patterns in the data. You can then take those patterns and test them in controlled, hypothesis-driven experiments to determine the specific mechanisms driving the effect. To conduct hypothesis-driven research, you would brainstorm your research question, create experimental conditions to answer that question, and run your experiment. If you identify an effect in a hypothesis-driven study, then it can be helpful to test whether the effect generalizes by running a more naturalistic, data-driven study.

to have a clear idea of the different alternatives of the results (in other words, **hypotheses** about the results) and what that means for the bigger picture question. The general pipeline for running a hypothesis-driven study follows the flow of the right side of Figure 2.3. One of the most important first steps is deciding on the main question. So, taking the first case study example we covered, let's say your question is: "Are people better at remembering faces close to them in age than faces far from them in age?" Your two experimental hypotheses would be something like: 1) yes, faces closer in age are remembered better, or 2) no, there is no difference in memory based on age of the face. You will then design your experiment, selecting experimental conditions that let you pinpoint that question. **Experimental conditions** are the different ways you divide up your experiment to answer your research question. For example, for this study, we may have different participant groups (older and younger people) as well as different stimulus sets (older and younger faces). So, we would have four conditions: young participants viewing young faces, young participants viewing older faces, older participants viewing young faces, and older participants viewing older faces. In experimental design terms, participant age is a **between-subjects** factor, because the condition changes from subject to subject (in other words, between the subjects). In contrast, face image age is a **within-subjects** factor, because within a single subject, they will see faces from both older and younger face conditions. Importantly, the conditions and factors that we intentionally change or control as experimenters (like the gender of participants and faces) are called **independent variables (IVs)**. It's because these are variables that we set, and so they are not dependent on other things in the experiment – they stand on their own. Once you have designed your experiment with its conditions, you then run the experiment (i.e., having your participants do a memory task with

the images you choose). The data that comes out of your experiment are **dependent variables** (DV), because they are dependent on the IVs and the experiment that you run. Finally, you can then run statistical tests on your data to directly answer the research question you set out to test. For a statistical test, you will typically define your **statistical hypotheses** – these are subtly different from experimental hypotheses in that they specify the different possible results of your statistical test. With most traditional statistical tests (called parametric statistics; see more discussion in Chapter 3), you would have a **null hypothesis** reflecting the hypothesis that there is no effect for a given statistical comparison. In this case, the null hypothesis would be that there is no difference in memory performance across our different participant-image age conditions. You would also define an **alternate hypothesis** reflecting the hypothesis that there is a significant effect. Here, the alternate hypothesis would be that there *is* a difference in memory across these conditions. So, we run our test and see which hypothesis we have more evidence for – can we reject this null hypothesis or fail to do so? Specifically, you would directly test whether memory performance (your DV) is higher for same-age conditions (young participants / faces and old participants / faces) than different-age conditions. And, voila, you have your answer! (So far, the research seems to point to yes: Chiroro & Valentine, 1995).

This hypothesis-centric way of thinking often goes hand in hand with small data research, because you want to run experiments that specifically target your question of interest. You *can* run these in a Big Data way (e.g., with thousands of participants and thousands of images), but in the end, the only factor you'll want to differ is the one you're interested in (age), and you'll want other factors like race, gender, attractiveness, etc. to be the same across your different conditions. This is because you want to be certain that your experimental manipulation has a direct influence on the effect you observe (in other words, you want your IV to directly influence your DV). If you do not control for other factors, you risk having **confounds** that could explain the link between your IVs and DVs. A confound is another variable that can account for the relationship between your IV(s) and DV(s) that you are not intentionally manipulating. In other words, they can be alternate explanations for your results. When designing an experiment, you want to make sure you can avoid these confounds, or at least can take them into account in some way. For example, let's say we look at monthly data across a year and find a correlation between ice cream sales and drowning deaths: when ice cream is popular, drowning is more common (Mumford & Anjum, 2013). Does this imply that ice cream causes people to drown? That would be ridiculous (and unfortunate)!

Discussion Question

What are some confounds that could explain a relationship between ice cream sales and drowning deaths?

One of the big confounds here is weather! During the summertime when the weather is nice, people want to go out swimming. They certainly have a much higher risk of drowning if they're swimming in a lake than if they're staying home bundled up by a fire during the wintertime. During the summer, people are also probably going out to buy ice cream to cool off from the hot weather, so you would see both high ice cream sales and increased drownings. On the other

hand, during the winter, the dessert of choice might be something more like a slice of apple pie or a mug of hot chocolate, and you would be unlikely to be out swimming. So, we would say that weather here is a confound in this relationship between drownings and ice cream sales. When designing a hypothesis-based study, it's important to keep in mind all potential confounds, and sometimes it can be impossible to control for all of them.

The counterpoint to hypothesis-driven research is **data-driven research**. The idea is that you collect tons of data, trying to gain samples that are as representative and varied as possible (Figure 2.3), and this is the approach often taken when using Big Data. Then, you use statistical methods to try and pull out patterns from the data that can help answer some questions. For example, you could collect memory test data for a wide range of faces and participants, and then see if people generally tend to remember faces closer to their own age best. This type of research is also sometimes called **exploratory research**, because you can explore around the data and look for different patterns without necessarily having a hypothesis from the beginning. What's great about data-driven research is that you generate big datasets that can help answer many questions. So, these databases can be multiuse – you could look at questions about memory and age, but also memory and attractiveness, or memory and face shape. You can also look at how these different factors all work together to form the big picture (i.e., what combinations of features influence the memorability of a face?). However, because these data tend to be collected without controlling much in the experiments, you run a higher risk of having more confounds that can explain your effects. However, because you often aren't relying on a single research question or statistical test, one confound may be less impactful on the use of the dataset overall. However, if you are not careful, data-driven research has some other big risks that can result in low-quality science (see Section 2.5 on data fishing).

Overall, there are pros and cons to both hypothesis-driven and data-driven research, with the key points summarized in Table 2.1. But ultimately, science benefits most when we do both, because they serve as an interconnected loop. Data-driven studies let us discover new, unexpected effects that can emerge from large or naturalistic data. Hypothesis-driven studies then let us take these effects and pinpoint the reasons behind these effects and link them to broader theories about human cognition. A lot of psychology reasonably takes the hypothesis-driven approach as a result. But to broaden our perspective on what questions to ask and what blinders we might have on in the field, I would argue the data-driven approach is just as important – and this is what will be the focus of this textbook.

Table 2.1 Comparison of the pros and cons of hypothesis-driven and data-driven research

Hypothesis-driven research	Data-driven research
Usually small data	Usually Big Data
Can isolate specific effects	Can be more naturalistic
Pulling out data based on theory-driven questions	Pulling out questions based on diverse datasets
Larger effect sizes	Statistical significance more likely
Have to be wary of confounds	Have to be wary of data fishing

2.4 Deep Data versus Wide Data

When designing a Big Data study, there are two dimensions along which it can be Big – deep or wide. A **deep data** study is one where you collect lots of data for a smaller number of individuals (so you're getting a deep look at a few people). Some examples include sensor data like a fitness watch recording frequently and over long periods of time (Chapter 9), or software-based data recording lots of samples over time (like on your phone; Chapter 8). There are also some experiments that focus on running the same participants many times over a series of sessions. These repeated measurements can give rich information about individuals, letting us look at the influence on cognition of things that vary like the time of day, attention fluctuations over time, and complex behaviors. One issue with deep data, though, is that it can risk being invasive of participants' privacy because you're learning so much about specific people. For example, if you look at one person's measurements from a fitness watch over months, you would learn all about their sleeping and exercise habits. It can also be tedious for participants to collect and provide all this data, especially if it's a study where they have to come in for multiple sessions. So, it can be hard to recruit participants for deep studies.

A **wide data** study is one where you collect relatively small amounts of data from a large number of participants at a single time. Some examples include data from an online experiment across thousands of people, or a snapshot of rich data from an app or piece of software at a single point (like all the tweets for a topic on a single day). What's great about wide data is that it can give diverse information across a large, representative sample of people. However, you often cannot capture very complex behaviors that vary over time or an interaction.

Some researchers characterize Big Data along three dimensions – being deep, wide, and long. In this case, deep data would still involve multiple measurements (like we see with sensor data). However, wide data now instead reflects collecting data across multiple variables or measures (e.g., with a battery of questionnaires). Finally, long data would involve collecting data from many people. Regardless of how you characterize the dimensions of Big Data, studies can be any combination of deep, wide, and long (e.g., collecting tons of data from many people), although this can be hard to achieve, so scientists may need to pick one dimension along which to specialize. When looking at a study, it's worth thinking how it falls on these different measures of size.

2.5 Big Ethical Questions

When you have a huge experiment, it can be easy to go fishing around for significant effects. This is something called **data fishing**, **data dredging**, or p-hacking. Since you have so much data, it seems like one of the benefits is that you should be able to look at many different effects in your data at once, right? Well, yes, you can do this to some degree, but you also need to think about how statistical tests are conducted.

For a majority of standard statistical tests that compare your data to a distribution (like t-tests, ANOVAs, regressions, etc.), you aim to estimate a ***p*-value**. What this *p*-value represents is the probability you would observe something as extreme as your results if the null hypothesis

were true. Recall that the null hypothesis reflects the hypothesis that there is no effect in your data. However, even if this null hypothesis were true, there is noise in our measurements and people's behaviors, so we would still sometimes observe a difference between our conditions "by chance." When we run a statistical test, we are looking at what the distribution of data would look like if the null hypothesis were true (and there was no effect). We are then seeing where our observed data falls in this distribution – how likely is it to occur given this null distribution? We calculate our *p*-value as the proportion of data in the null distribution that is equal to or more extreme than our observed value. So, for example, a *p*-value of 0.03 indicates there's only a 3 percent chance you would happen to observe these results just by random chance. That seems pretty low, and as a field, we've currently accepted a cut-off of 5 percent ($p < 0.05$) to be how we determine what we'll take to be a significant finding or not. Another way to phrase this is that in our field, we have accepted a 5 percent **false positive rate.** This is the rate of falsely saying something shows an effect when it does not. You may have heard this term used to refer to the rate of a medical test falsely saying you have an illness when you do not – same idea.

While this 5 percent chance of a false positive seems rare, when you're dealing with Big Data, you're doing many statistical tests – maybe hundreds of tests (e.g., for a psychological battery), or even up to hundreds of thousands of tests (e.g., for the case of MRI brain data). And so, in the realm of hundreds of thousands of tests, even with this seemingly strict false positive rate, we will get about 5,000 tests (5 percent of 100,000) that come out as "significant" just by chance! So we need to think carefully about how we define significance with data-driven research since we are doing many tests, inflating the chance that we find a false positive in at least one of these tests. In order to circumvent these issues, we do something called **multiple comparisons correction**, which is a group of statistical methods that let us calculate an adjusted *p*-value threshold for our study that takes into account the many tests that we are doing. While we won't go into these methods in detail, some example methods include **Bonferroni correction** and **false discovery rate correction**. Bonferroni correction corrects for the rate of false positives across all of the statistical tests that you perform. It does this by calculating an adjusted threshold for "significance" (called the alpha level, or α), based on the number of tests you are running. So, if you run ten tests, your alpha level would be $p < 0.005$ instead of $p < 0.05$. False discovery rate correction is a more liberal method that corrects for the proportion of false positives among all results initially labeled as significant – in other words, calculating an alpha level so that we are okay with 5 percent of our discoveries being false positives.

Let's look at an example study that ran many statistical tests. In Moore et al.'s 2006 study "Thongs, flip flops, and unintended pregnancy," the researchers wanted to investigate if there were some lifestyle factors that were related to unintended pregnancies. They conducted a 50+ question survey with 126 women who were currently or recently pregnant, and conducted 362 statistical tests to analyze their data. They found some surprising results: unintended pregnancy was associated with preferences for yoga, beaches, thongs, Doritos, contact lens, and text messaging. They also found that baby boys were more common if mothers preferred trucks, beef, and boys, while baby girls were more common if mothers preferred cars, chicken, and girls. So does that mean if you see your friend texting their friends during some beach yoga while tucking into a bag of Doritos, that you should encourage them to be vigilant with their

contraception? No, because if you think back to what we just discussed with running many statistical tests, we would expect about 18 of their 362 tests to come out as significant just by chance given our *p*-value threshold of 0.05! So even if there are no meaningful relationships between any of these factors and unintended pregnancy, just because of random noise in measurement, participant behaviors, and the environment, it would be unsurprising to find some relationships that come out as statistically "significant" but aren't real.

So one of the risks of Big Data is that it's easy to run many, many statistical tests until you find something significant. Because of all of the rich data you have, it's tempting to test many different questions. There are also big pressures in the scientific world to publish significant results, so researchers may be tempted to focus on these "significant" results without accounting for the number of statistical tests that they're running. In other words, you may be tempted to fish around for a result in your big sea of data. There are three main ways to make sure you are not doing data fishing with your own data. One way is to perform multiple comparisons correction across all the tests you run. A second way is to decide your analyses and hypotheses in advance before seeing your data (called preregistration; see Section 4.4) – so in other words, running a combined hypothesis- and data-driven study. A third way is to replicate any findings you discover across multiple datasets, analyses, and/or labs to be sure that what you're finding is real, rather than something that emerges just by chance.

There is also the question of the **effect sizes** of the results you end up finding. While we often care about statistical significance in our data, we also care about how strong the effects are in the differences that we are measuring. Effect size is often quantified as the proportion of the **signal** of interest to the level of **noise**. So, for example, for a t-test, the measure of effect size is the difference between the conditions' averages (the "signal"), divided by the standard deviation pooled across the two conditions (the "noise"). You can have a significant effect that's a weak effect or a strong effect. For example, let's say you're looking at whether an intervention in the classroom results in a difference in test scores on a test with a maximum of a hundred points. You could get a significant effect where the intervention results in a one-point increase. While this would mean the intervention likely worked (because the effect is significant), it didn't work very well (the effect is weak)! If the intervention instead resulted in a significant thirty-point increase, we would say this is a strong effect! And, if you have a nonsignificant effect with a thirty-point increase, that would mean our results aren't strong enough (e.g., there may be too much noise), so we cannot be confident that this thirty-point increase didn't just happen by chance. Because of its large sample sizes, Big Data can be prone to identifying significant but weak effects – effects that would only be detectable when you have thousands of people. Therefore, even if you find a significant result, consider what the result means. If it is a meaningful, strong psychological effect, ideally we would even see it occur at the level of a smaller sample, and even at the level of the individual.

2.6 Applications of the Chapter

In this chapter, we discussed characterizing research in a few different ways – for example, hypothesis-driven versus data-driven or deep versus wide data. These ways of thinking about

data have promoted discoveries beyond the field of psychology, and have guided recent advancements in the medical field.

2.6.1 Data-Driven Discoveries

We mentioned how data-driven research can result in new questions or effects that we may not have conceived of if we only stuck to pre-existing theories and hypothesis-driven experiments. There are in fact many exciting scientific discoveries that came about thanks to people trying out many different things. One of the most famous examples in psychology is the discovery by Professors David Hubel and Torsten Wiesel that led to their Nobel Prize win in 1981. They were trying to see what information was coded in neurons in the occipital lobe (the early visual regions of the brain), by recording directly from cats' brains while showing them different images (see Chapter 10 to learn more about neuronal recording). They were struggling to find any specific image that would cause these neurons to spike. Back in the day, they were using a slide projector, and suddenly when they were swapping out the slides, they heard the neuron they were recording from start to fire. After playing with the slide, they discovered that this neuron was sensitive to the edge of the slide when it was shown at a specific angle. This led to our current understanding of the visual system in the brain, where neurons are sensitive to edges oriented at specific angles. You may have heard about similar fortuitous "eureka!" moments throughout the sciences. As the classic example, around 246 BC, Greek scientist Archimedes realized how to calculate volume and density while taking a bath, and purportedly ran through the streets shouting "eureka!" In 1820, Dr. Hans Christian Oersted noticed a compass move when he placed it near an early battery he was creating – resulting in the discovery that electrical currents generate a magnetic field. Percy Spencer invented the microwave in 1946 when he noticed the chocolate in his pocket melted when he was testing out a new vacuum tube. These discoveries may have never happened without the experimenters just trying out different things. Big Data can encourage such exploration, which can lead to exciting discoveries.

2.6.2 Medical Applications of Deep and Wide Research

Deep and wide methods have had some wide-reaching applications in the clinical realm. Deep data has helped form the field of **precision medicine**, where healthcare workers can make honed, personalized predictions of health outcomes based on genetics, environment, lifestyle, and sensor measures. Big Data lets us create **predictive models** that take these different factors and then make guesses about outcomes for a single person (see Chapter 6). Precision medicine goes hand in hand with preventative medicine and telehealth, where people can wear sensors and use apps to remotely track and communicate symptoms before they develop into a full-blown condition. For example, researchers are working on apps to help identify early stages of Alzheimer's disease (Konig et al., 2018) and apps to help elderly individuals develop memory strategies (Martin et al., 2022).

Wide data is key in letting us learn about diseases: It lets one see global trends in disease, identify rare groups at particular risk, and find hidden links to a cause or cure (Heggie, 2019). Wide data played an important role in identifying the symptoms early on in the COVID-19

pandemic, when it wasn't clear what symptoms were being caused by the virus. Researchers conducted a large-scale wide symptom study where anyone could enter their symptoms online, and they ended up receiving information from 4.4 million participants (Menni et al., 2020). As a result, they were one of the first groups to identify a loss of smell or taste as one of the symptoms of COVID-19. They also found some other interesting trends: for example, for the first wave of the pandemic, one out of twenty participants had symptoms that lasted more than 8 weeks, and longer COVID was correlated with having more different symptoms in the first week. They also found that during the pandemic lockdowns, 20 percent of participants had an increase in alcohol consumption, and an average weight gain of 4.6 lbs.

CHAPTER SUMMARY

In this chapter, we discussed the small data experiments traditionally utilized in psychology research and showed how they compare to the Big Data experiments that are becoming increasingly popular. Here are some of the main takeaways:

1. The key to making a small data experiment "Big" is expanding its participants, stimuli, and paradigms to be more naturalistic and representative of the real world. However, you can almost never make a perfectly representative experiment.
2. There are different benefits to hypothesis-driven research versus data-driven research, and both are necessary for the progression of psychology as an innovative and rigorous field.
3. Big Data can be characterized by two key dimensions – its depth (how many measures you collect per individual) and its width (how many individuals you record from).
4. With the large amount of data you can get from a Big Data study, we must be cautious of not "fishing" for effects without accounting for all of the statistical tests that we are conducting.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter.

- **Learn about how Big Data is causing big strides in the understanding of disease:** Heggie, J. (2019, January 8). How can big data beat disease? *National Geographic*. https://tinyurl.com/ykkabh53
- **A cautionary tale on how too many statistical tests can lead to effects that may not be real:** Moore, R. P., Galvin, S. L., & Imseis, H. M. (2006). Thongs, flip-flops, and unintended pregnancy: The seduction of $p < 0.05$. *MAHEC Online Journal of Research*, *1*, 1.
- **Dive deeper into the statistics used to correct for multiple comparisons:** Lindquist, M. A., & Meija, A. (2015). Zen and the art of multiple comparisons. *Psychosomatic Medicine*, *77*, 114–125.

ASSIGNMENT

The purpose of this assignment is to get you thinking about Big Data and how to build out Big Data experiments. Please submit your response in a way so that it is clear what questions and sub-questions you are responding to.

Total Points: 50

1. Pick two psychology papers describing an experiment on a topic that sounds interesting to you. (Do not use a review paper.) They can come from either:
   i) a psychology class you are currently taking, or took in the past;
   ii) a lab you are currently working in; or
   iii) any "Open Access" articles from the most recent year of the journal *Psychological Science* (https://journals.sagepub.com/home/pss).
   Please choose at least one paper that you would consider a "small data" experiment. Provide the citation (titles, authors, year, journal) and abstract of the two papers here: (2 points)

   We will now look at these papers with the frameworks we discussed in this chapter.
2. **For paper 1, answer the following questions**. If the study includes multiple experiments, answer for the first or main experiment:
   a. Is this a **"small data"** or a **"Big Data"** experiment? How do you know? How small/big is the sample size (number of participants)? How small/big is the experiment itself (e.g., number of conditions, stimuli, outcome measures)? (4 points)
   b. Is this a **hypothesis-driven** or a **data-driven** experiment? How can you tell? (3 points)
      i. If this is a hypothesis-driven experiment, what is their hypothesis?
      ii. If this is a data-driven experiment, what new hypotheses come out of their data? How did they avoid p-hacking / data fishing?
   c. Is the data **deep** or **wide** (or both)? How do you know? (2 points)
3. **For paper 1, we will do some more brainstorming on Big Data. (Note that the next part of the question has two options – for a given paper, answer for either small data or Big Data, not both.)**

   If this is a "small data" experiment, we will think up how to make it into a Big Data experiment. Answer these questions:
   a. How **naturalistic** versus **artificial** is their experiment? What are ways in which the stimuli, experiment, or participants are not *representative* of reality? What are ways in which they are? (3 points)
   b. How can we improve the **representativeness** of the study? Use your creativity to brainstorm how you would make this into a "Big Data" experiment. How would you change the experimental paradigm, participant recruiting, the stimuli, or the measurement techniques to capture bigger, more diverse, more naturalistic, and more representative data? (5 points)
   c. What **limitations** could you envision with these changes? These can be limitations in terms of feasibility/practicality (e.g., how much time or money does your change add)? In what ways is your version still not fully representative? (3 points)

   If this is a "Big Data" experiment, we will see how it improves upon small data studies. Answer these questions:
   a. What would the **"small data"** version of the experiment have looked like? (4 points)
   b. **Why** did the experimenters decide to take this "Big Data" approach? What innovations did they apply to make it "Big Data"? (4 points)

c. What **limitations** still exist with their approach? In what ways are the data still not fully representative of real people / images / cognitive processes? What additional improvements could you envision, and how feasible are they (e.g., how much time or money does your change add?) (3 points)

4. Answer questions 2 and 3 for **paper 2** below. (20 points)
5. What are some ideas implemented by paper 1 that could be useful for paper 2 in making their experiments more representative in terms of participants, stimuli, or paradigms? Similarly, what are some ideas implemented by paper 2 that could be useful for paper 1? (5 points)

REFERENCES

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*, 1043–1047.

Bainbridge, W. A. (2017). The memorability of people: Intrinsic memorability across transformations of a person's face. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 706–716.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*, 1323.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 48, 879–894.

Davis, T., & Bainbridge, W. A. (2023). Memory for artwork is predictable. *Proceedings of the National Academy of Sciences USA*, *12*, e2302389120.

Doyen, S., Klein, O., Phcion, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Heggie, J. (2019, January 8). How can big data beat disease? *National Geographic.* https://tinyurl.com/ykkabh53

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323.

Konig, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., & Robert, P. H. (2018). Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, *15*, 120–129.

Martin, C. B., Hong, B., Newsome, R. N., Savel, K., Meade, M. E., Xia, A., Honey, C. J., & Barense, M. D. (2022). A smartphone intervention that enhances real-world memory and promotes differentiation of hippocampal activity in older adults. *Proceedings of the National Academy of Sciences USA*, *119*, e2214285119.

Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., El-Sayed Moustafa, J. S., Visconti, A., Hysi, P., Bowyer, R. C. E., Mangino, M., Falchi, M., Wolf, J., Ourselin, S., Chan, A. T., Steves, C. J., & Spector, T. D. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine*, *26*, 1037–1040.

Moore, R. P., Galvin, S. L., & Imseis, H. M. (2006). Thongs, flip-flops, and unintended pregnancy: The seduction of $p < 0.05$. *MAHEC Online Journal of Research*, *1*, 1.

Mumford, S., & Anjum, R. L. (2013, November 15). Correlation is not causation. Oxford University Press blog. https://blog.oup.com/2013/11/correlation-is-not-causation

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Sciences*, *25*, 506–519.

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210.

12 Big Social Interactions

Introduction

A lot of the psychological questions we have discussed have focused on the mind and brain of the individual. However, humans are a social species, and social interactions shape our ways of thinking from birth to old age. In the last two decades, these interactions have transformed, as the internet has become an essential thread to how we connect with people. We interact with people we do and don't know over social media sites. We vet potential partners from an online profile and virtual interactions. We join niche online communities that don't exist in our local neighborhood. As a result, social psychologists have been busy examining the nature of these new types of relationships and interactions. Excitingly, the digital nature of this new social world means that we can quantify it using our Big Data tools.

In this chapter, we will briefly discuss some psychological principles of social networks (Section 12.1). We will then discuss methods for quantifying social networks based in network theory (Section 12.2), including how to create social network graphs (Section 12.2.1) and how to measure them (Section 12.2.2). We will look at effects of the small-world phenomenon (Section 12.2.3) and social ties (Section 12.2.4). We will then look at the case of online social networks (Section 12.3) and what can be learned about you from social media (Section 12.3.1). We will look at their effects on psychology (Section 12.3.2), and discuss broader findings on social network representations in the brain (Section 12.4). We will talk about ethical questions (Section 12.5) related to privacy concerns on social media (Section 12.5.1) and AI-based social interactions (Section 12.5.2). Finally, we will discuss how network theory can reveal biases in social groups (Section 12.6).

12.1 Psychology of Social Networks

When you think of the phrase *social network*, what might first come to mind are all of the social networking websites and apps that we belong to. However, a **social network** applies to any social structure connected by interpersonal relationships.

Discussion Question

What social networks are you a part of? (Hint: Think beyond just online social network sites.)

You belong to a multitude of social networks that may differ in the size of the network and types of personal interactions. Your family and groups of friends are social networks. Your university campus or a job you may work for form social networks. You belong to geographic social networks of your local community. You may also belong to social networks joined by interests, such as extracurriculars, sports teams, political views, religious groups, or hobbies. Finally, you may also belong to different types of virtual networks – professional networks (e.g., LinkedIn, Slack, Microsoft Teams) and more social networks (e.g., Instagram, Reddit, Facebook, TikTok). You will belong to some social groups for life, while for others you might even change membership from day to day. For example, you may find yourself fuming at different people and aligning more in your stances on traffic laws based on whether you're a pedestrian, bicyclist, or driver that day.

A large portion of the field of social psychology is dedicated to understanding these social groups and their interactions. One common observation is that people act differently to those perceived to be within their social group (an **in-group member**) versus those outside their social group (an **out-group member**; Brewer, 1979). Across a range of metrics, people will show preferences for in-group members: They will evaluate them more highly and share more resources with them. Individuals are more likely to align their beliefs with their in-group's **social norms** (i.e., the informal rules that determine acceptable behavior; Marques et al., 1998). People also often experience the **out-group homogeneity effect**, where they perceive that their own group is more diverse, while out-groups are more homogenous (Judd & Park, 1988). If you think about different cliques you might have been part of in high school, you can envision these differences. The "school nerds" may feel they are all more intelligent and deeper thinkers than the "athletes" who all look the same to them. The "athletes" may feel they are more socially adjusted and popular than the "nerds" who all seem quiet. This sort of thinking unfortunately does not leave us when we graduate from high school, and can result in the cases of discrimination, prejudice, and bias we see across society. Unfortunately, children pick up these in-group biases early on – developing a preference for their in-groups around 6 years of age, and a distaste for out-groups around 8 years of age (Buttelmann & Böhm, 2014). One hypothesized reason is that there may be an evolutionary mechanism to these biases, as it may be an effective strategy to bind together human tribes (Fu et al., 2012).

12.2 Network Theory

At first, it may seem unclear how one can turn the complex subjective data of human relationships into quantifiable measures. However, we can turn these sets of relationships into networks that we can quantify, using the principles of **network theory**. Network theory is a form

of **graph theory** – a method where we can define relationships across items by creating graphs of individuals and their relationships. Network theory can be applied to many fields such as linguistics, neuroscience, biology, and computer science, but it also works well for quantifying social networks. It provides a method for visualizing that social network, but also a framework for computing measures about it.

12.2.1 Turning Relationships into Networks

You may have seen a social network graph if you have geeked out over a complex television show, book series, comic series, or movie in recent years. For example, Professor Andrew Beveridge (2016), currently at Macalester College, published the social network data and graph from the popular book series *A Song of Ice and Fire* (also known as the TV series *Game of Thrones*), and this has become a hugely popular dataset for teaching students how to conduct network analyses because of the complex relationship dynamics across its characters.

Here, to illustrate an example social network, let's invent a love triangle across three people: Bob, Sandy, and Ash. Bob has had a crush on Sandy ever since he met her at the local university café. He then spotted her in his Big Data class and has been asking for advice on the assignments, gradually working up the courage to ask her on a nonhomework-related date. Unfortunately Sandy's been busy playing for the school volleyball team and has been completely oblivious to Bob's interest. At the same time, Ash has been sitting in the back of class, getting absorbed by Bob's endearing mannerisms, and has started to develop a crush of their own.

How do we quantify the interactions across these three people? Refer to Figure 12.1 as we form the graph of our three-person network. A network starts with two key components. First, we have **vertices** (also called nodes or points), which define each of the people in our network (Bob, Sandy, and Ash). These are often drawn as circles or dots in our graph. Second, we have **edges** (also called links, lines, or ties) that represent connections between our vertices, or in

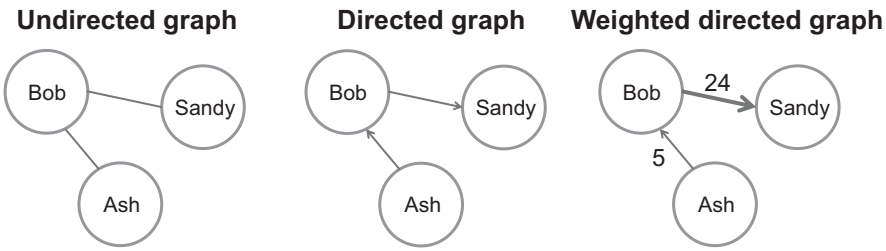


Figure 12.1 Social network graphs illustrating the relationships between Bob, Sandy, and Ash. Each person is a node, and their relationships are shown as edges between the nodes. An undirected graph shows symmetrical relationships (e.g., friendships). A directed graph shows directional relationships (e.g., Ash has a crush on Bob, Bob has a crush on Sandy). A weighted graph shows relationships with different strengths, measured by something like number of text messages sent from one person to the other (e.g., five from Ash to Bob, twenty-four from Bob to Sandy).

other words, the relationships between people. These are drawn as lines or arrows between our circles. If edges have arrows, then they can reflect directionality of a relationship, and this is called a **directed graph**. For example, our love triangle might be best depicted as a directed graph representing the directions of these crushes, with an arrow from Ash to Bob, and an arrow from Bob to Sandy. If there is no directionality in the relationships (i.e., the relationships between nodes are symmetrical), then the edges would just be lines, and this is called an **undirected graph**. For example, you might use an undirected graph to represent a friends group, because friendships are often symmetrical – if you're friends with someone, they are also friends with you. Graphs can also be **weighted graphs**, if you assign a value to the edges. Higher values will often represent stronger connections and can be drawn as thicker lines on the graph. For example, perhaps the intensity of Bob's crush is stronger than that of Ash's, so the edge between Bob and Sandy would be stronger. The thickness of the edge could also be represented by a measure, like how many texts Bob has sent Sandy (e.g., twenty-four texts) versus how many Ash has sent Bob (e.g., five texts).

The edges in a graph can reflect any sort of measure that you can quantify between two nodes (i.e., two people), allowing you to flexibly represent a range of types of relationships within a graph. However, you should use the same measure for all of your edges within a single graph. The measure that you choose for your edges determines whether your graph will be directed or not. For example, you could use any of these measures (or more) to determine the edge between two people:

1. *Communication behavior.* Examples: Number of texts sent from one person to another; number of social media posts tagging the other person; number of likes for the other person on social media; length of phone calls between two people.
2. *Co-occurrences*. Examples: Proportion of time spent in the same location; number of times two fictional characters are in the same scene; number of events attended by both people.
3. *Categorical (i.e., what categories do people fall into?)*. Examples: Are these two people siblings, cousins, friends, or acquaintances?
4. *Similarity*. Examples: How similar are these two people from a personality questionnaire? What is the level of genetic similarity between these two people? How similar is the way they talk (quantified by something like word embeddings in Section 11.2.3)?
5. *Subjective ratings*. Examples: Ratings of how strong your relationship is with the other person; third-party ratings of how related two people are.

Discussion Question

For each of the above edge measures, what type of graph would you draw (undirected or directed? Weighted or unweighted?)

Therefore, edges do not only have to be the relationship itself (i.e., Person A and Person B are friends). They can use other measures capturing people's interactions or the flow of information or goods between people to serve as a proxy for their relationship.

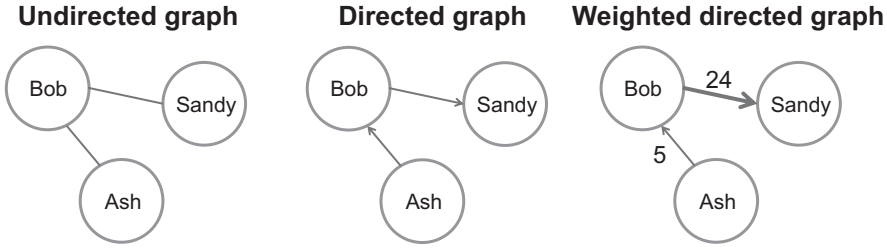
12.2.2 Quantifying Graphs

Now that we know how to draw graphs, how can we turn them into numbers? Graphs can be represented as a **matrix**, which allows us to use various matrix math methods and linear algebra to quantify measures about these matrices. Essentially, the rows and columns of the matrix are each node, in the same order along the rows and columns (Figure 12.2). Each cell in the matrix can then represent a pair of people based on its row and column. In this cell is where you would put in the edge value. For example, the cell in the Sandy row and the Bob column represents the relationship between Sandy and Bob, and you could put their number of text messages in that cell. If your graph is unweighted, then you will put 1s in cells where there is a line, and 0s in cells where there is no line. If your graph is weighted, then you will place the corresponding edge value instead. If your graph is undirected, then your matrix will be symmetrical along the diagonal – because Bob and Sandy are as related to each other as Sandy and Bob. However, if your matrix is directed (e.g., Bob likes Sandy but not vice versa), then your matrix would be asymmetrical (the Bob–Sandy cell would be 1, while the Sandy–Bob cell would be 0). Finally, one other feature of the matrix is that you should leave the diagonal blank (or filled with Not-A-Number/NaNs in your statistical package), because most of the time you cannot really measure a person's relationship to themself.

Once you have your network in a numerical form, there are many features you can quantify for a network! Many of these you can gain an intuition for by viewing the graph, too, if you have a small enough dataset. First off, you can get a sense of how big a network is, or its **network size**. This is quantified as the total number of nodes (in our toy case, three) or total number of edges (two). You can then calculate the **network density**, or the degree to which everyone is connected to each other. This is calculated as:

$$\text{network density} = (\#\text{ of edges})/(\#\text{ of possible edges based on }\#\text{ of nodes})$$

If we take our simple network example from Figure 12.2, among three nodes, the maximum number of possible edges would be three edges (if all people were connected to each other).



	Bob	Sandy	Ash
Bob	–	1	1
Sandy	1	–	0
Ash	1	0	–

	Bob	Sandy	Ash
Bob	–	1	0
Sandy	0	–	0
Ash	1	0	–

	Bob	Sandy	Ash
Bob	–	24	0
Sandy	0	–	0
Ash	5	0	–

Figure 12.2 The corresponding matrices representing the graphs from Figure 12.1.

However, this network only has two edges, so the network density is two-thirds, or 66.67 percent.

You can also quantify features for individual nodes in a network. **Degree centrality** is the number of edges that a specific node has. The more edges a node has, the more "central" that node is because it is more connected with other people. You can use this measure to identify key players in a social network – for example, Bob has the highest degree centrality (two) in our example and thus may be the main character of our story. You can also quantify the relationships between two people, by finding their **shortest path** – how many edges minimum does it take to travel from one person to another? For example, the shortest path from Ash to Sandy is two. If it is a weighted graph, your measure of shortest path might depend on what the edges represent – but you would choose the path that might have the lowest or highest weights to traverse. The number of edges along the shortest path between two nodes is often quantified as **degrees**. For example, if we are neighbors in a network, we are one degree apart. However, if it takes two hops to travel between two people (like Ash and Sandy), then they are two degrees apart. For the network as a whole, you can then calculate its **network diameter**, which is the largest of the shortest paths between all pairs of nodes (in this case, it's just two).

With these network metrics in hand, you can answer many psychological questions. For example, do certain groups have smaller network diameters and higher network densities than others? What personality traits cause one to have a higher degree centrality? Can length of the shortest path between two people predict how they would interact? In Sections 12.2.3 and 12.2.4, we go over a few psychological phenomena that have been illuminated by network theory.

12.2.3 Small-World Phenomenon

From these social relationship graphs, we can often observe a psychological effect called the **small-world phenomenon**. What this phenomenon observes is that most social networks are rich in short paths – there are a small number of hops needed to connect any two nodes. You may have heard of the "Six Degrees of Kevin Bacon" – it is almost impossible to find an actor or actress who is more than six degrees or connections away from actor Kevin Bacon (P. Reynolds, 1999; try it out here: https://oracleofbacon.org). This phenomenon was impressively illustrated in a real-world study by Travers and Milgram in 1969. In this study, 296 people in Nebraska and Boston were asked to send a piece of snail mail to someone they knew with the goal of getting it closer to a target individual they didn't know in the suburbs of Massachusetts. Whomever received a letter was also instructed to send it along to get it closer to this target person. So, in other words, people were making a chain of letters across the United States to get a letter to a specific person. For this specific target person, they knew some details like his name, occupation, university, military service dates, and his wife's name. But this experiment predated the internet, so you couldn't look up anything about the person or their friends. Even so, sixty-four letter chains (29 percent of participants) successfully made it to this target, with an average number of only five steps between the starting person and the target. So, even in a large social network like the United States (population 200 million at the time), the network diameter is actually quite small. There were also some key players (with high degree centrality) in this experiment who served as the bridges across small networks in

the larger network of the US: 48 percent of the chains passed through the same three people before making it to the target.

12.2.4 Social Ties

Another way in which graph theory has been helpful has been in quantifying the impact of relationships between people and other life factors. There are two main categories of people's social connections: strong ties and weak ties. **Strong ties** are those with deep affinity, like the connections you will have with family members, good friends, or even colleagues you work closely with. **Weak ties** represent acquaintance relationships, like classmates or other people from a common group. We have already discussed different edge metrics you could use to categorize relationships as being strong or weak, like amount of communication. While you might think that more strong ties are better for a person's social life, weak ties are incredibly important because they serve as **bridges** between small worlds, and can provide conduits of knowledge, opinions, and opportunities. In other words, you can gain a large benefit to branching out beyond your immediate group (i.e., developing weak ties outside of the "small worlds" you belong to), because you can combat the in-group/out-group biases we have discussed, and gain diverse perspectives. It's thought that the three key players in the Travers and Milgram mailing experiment (1969) were so well-connected because of their ability to serve as network bridges.

There is strong empirical evidence supporting the benefit of weak ties. Individuals with more weak ties have been shown to find jobs with higher compensation and satisfaction (Granovetter, 1973). Big Data analyses of online social networking platforms have helped confirm how this happens. The career networking website LinkedIn has a section on its site called "People You May Know" that suggests potential connections to its users. In a study by Rajkumar and colleagues (2022), they analyzed the strength of ties suggested in "People You May Know" for 20 million LinkedIn users, and saw how these ties linked to finding a new job. They discovered that moderately weak ties were the most helpful for finding new jobs (over strong ties!), especially for remote jobs and jobs in the tech industry. Another study, conducted by Gee et al. (2017), looked at 6 million Facebook users who were helped by a Facebook friend in finding their latest job. In this study, they quantified the strength of the tie between two people using various measures: number of mutual friends, number of wall posts with each other, and number of photos they are tagged in together. The researchers found that 90 percent of successful job recruitments were facilitated by weak ties, although at the level of specific relationships, individual strong ties had the highest chance of helping someone find a job. In other words, you had the best chances of finding a job if you had high-quality strong ties, but a high quantity of weak ties.

Discussion Question

What is your balance of strong and weak ties in your own social networks? Can you tell who the bridges are within your networks?

12.3 Online Social Networks

As illustrated in the previous examples about social ties and job recruitment success, you can learn a lot about society-level trends by analyzing the Big Data present in online social networks. In order to know what Big Data analyses we can conduct, it's first helpful to think about the scope of data that exists online about each of us.

Discussion Question

What information do online social networks know about you?

The type of data will depend on the specific social media platform, but as a whole we provide an outstanding amount of information about ourselves on social media. We give basic demographic information about ourselves like our marital status, life stage, gender, sexual orientation, and ethnicity and race. We communicate our socioeconomic status with information about our occupation, career stage, and possibly even income range. We publicly post about our views on politics and religion. We engage in groups related to our favorite hobbies, bands, books, TV shows, movies, and franchises. We show who are the people (and animals) important to us through posts, payments, and photos. One can also glean information on our changes in behavior and opinions over time through social media. We display our financial interests through ad engagement, purchasing behavior, and posts. As a result, these sites can see what your needs are in the moment – will you be traveling soon? What is your mood? Are you seeking a certain type of information? See Section 8.2 for more discussion on what data websites can record about you and how. As a result, companies and researchers can learn a lot about you.

12.3.1 What Can We Learn about You from Social Media?

From all of this data, companies and researchers can build models from large samples of publicly available profiles, that can then make predictions about traits and behavior of specific individuals. Refer to Chapter 6 for ideas on how we could build such models, using artificial intelligence. This idea in fact forms the basis of how many social media platforms operate. For example, TikTok likely uses models of many people's interactions with videos to predict which ones you will engage in most, given your past history of engagement. Such personality models can be incredibly useful for advertisers, job recruiters, and political campaigns to make targeted, effective advertisements. However, these models also can help deepen our understanding of human personality and behavior. Such big social media data allows us to examine fundamental questions in social psychology about networks, relationships, and interactions, in a naturalistic setting.

These models of behavior and personality are so impressive, that social media might sometimes know more about you than your friends or family do. In fact, your pattern of Facebook likes has been shown to be better a predictor of your personality than personality

ratings made by friends and family (Youyou et al., 2015). Your Facebook network structure also reveals who your current romantic partner likely is, even if you are not "dating" or "married" on each other's profiles. What is even scarier is that your network structure is also significantly predictive of your likelihood to break up with that person in the next two months (Backstrom & Kleinberg, 2014). In general, the researchers find that the more a couple's mutual friends are connected within their social network, the less likely a couple is to break up. In fact, Facebook filed a "Predicting Life Changes" patent in 2018 that allows it to use profile data to predict when users will graduate, marry, and die (Cuthbertson, 2018).

Even if you are careful about privacy and have no information on your profile, AI and Big Data analyses can allow one to predict a lot about you just based on who your friends are. Your occupation and income are predictable from your Twitter/X connections (Aletras & Chamberlain, 2018). Researchers can even make pretty accurate predictions of what you would tweet if you did have a profile, based on the behavior of your Twitter connections (Bagrow et al., 2019). Companies can then monetize these predictions – they can even make successful predictions of your chances of buying a specific item or clicking on an ad based on what your friends have done (Goel & Goldstein, 2014).

Beyond the level of the individual, Big Data can also allow us to identify large-scale trends happening across people, like new fads or thought movements. Researchers have built models that can predict what photos on Facebook will be reshared (Cheng et al., 2014) and which tweets will be retweeted (Tan et al., 2014; visit our online resources to quantify your messages before you post them!). Many studies suggest you can even predict important political decisions based on sentiment expressed on social media. For example, one 2017 study used machine learning to examine 23 million tweets related to Brexit (the referendum for the withdrawal of the UK from the European Union) (Amador Diaz Lopez et al., 2017). They found that the sentiments captured by their model accurately captured the trends found in separate online polling data, and that their model was able to predict the UK leaving the EU.

12.3.2 Effects of Social Media on Psychology

In addition to social media providing a means to analyze human psychology, its use has also resulted in some interesting *impacts* on human psychology, launching a subfield of psychology research studying these impacts. Social media has become an integral part of many people's lives, with us meeting new partners, joining new communities, and strengthening real-world friendships through its use. So far, this subfield is still nascent, and there is little consensus on the overall impact of social media use on psychological measures.

Some studies have identified positive impacts of social media use. For example, spending time on your own Facebook profile has been shown to enhance self-esteem (Gonzales & Hancock, 2011). (As a sidebar: you may have noticed Facebook has been frequently studied in these experiments; this is likely because Facebook was one of the first big social media platforms that coincided with the rise in Big Data analytics.) Social media use can ease symptoms of depression when used to stay in touch with people (Bessière et al., 2010). It can also decrease feelings of isolation in older adults (Hogeboom et al., 2010). And, generally,

active, interactive use of social media can increase mood, well-being, and life satisfaction (Oh et al., 2014). The root mechanism behind all of these positive effects is that having an ability to more easily interact with other people can improve well-being. A majority of Americans (53 percent) report that they have between one and four close friends, with 8 percent reporting no close friends (Goddard, 2023). Similarly, the average Brit has 3.7 close friends, with 9 percent reporting no close friends (M. Reynolds, 2023). Thus, social media platforms may help to fulfill our desires to have close interactions with people, even when our friends are not available in person.

At the same time, researchers have also identified many negative effects of social media. While it may promote more interactions, social media can also promote feelings of loneliness and isolation due to a lower quality of interactions (Clark et al., 2018). In fact, an experimental study showed that social media use can have a direct causal role on depression (Hunt et al., 2018). Also, because social media is often used as a substitute for in-person interactions – acting as **social compensation** – it can promote addictive behavior and decreased involvement in real-life communities (Kuss & Griffiths, 2011). Finally, using social media apps in the 30 minutes before going to bed has been shown to decrease the quality of sleep (Levenson et al., 2017). Overall, social media is becoming an almost unavoidable aspect of our daily lives. However, it is helpful to be conscious of how we use it, what information we make available through it, and how we can avoid these negative impacts of overuse.

12.4 Social Networks in the Brain

As covered in Chapter 10, the brain also provides extensive Big Data that can give us insight about different types of psychological processes. Recent innovations in neuroimaging methods and analyses have allowed us to look at how we represent social networks and interactions in the brain.

In one study, participants interacted with characters in a roleplaying-game-like story while they were being scanned by an MRI machine (Tavares et al., 2015). The researchers found that the hippocampus, a region often linked with spatial maps and memory, also contained a map of the social network of these characters. This map was generally organized along two axes – power (how dominant/submissive a character is) and affiliation (how friendly the character is) – and participants with better social skills showed a higher link between the hippocampus's activity and locations in this social space. A related study by Parkinson et al. (2014) had participants view: 1) images of close and far distances (e.g., an object within your reach or many feet away); 2) prompts of recent and distant time points (e.g., *a few seconds from now* versus *years later*); and 3) images of people with close or far social relationships (e.g., friends versus acquaintances) (Figure 12.3). They found that a region in the brain (called the right inferior parietal lobule) had very similar patterns for representing spatial, temporal, and social distances. These studies provide evidence that we may represent social networks almost like how we represent physical maps.

Several studies have also looked at how group membership influences brain measures. We can use a measure called **intersubject correlations** where we see when and where the patterns

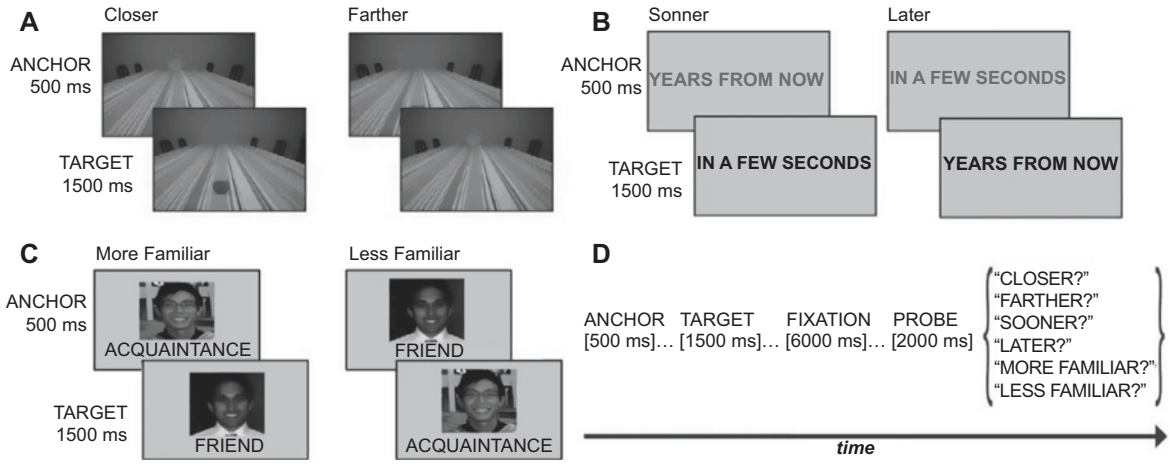


Figure 12.3 A summary of the different conditions tested in the fMRI experiment by Parkinson and colleagues (2014). They tested people on judging the distance between an anchor and target stimulus. These stimuli could vary in physical distance (A), temporal distance (B), or social distance (C). D shows the timing of a trial – participants would see the anchor stimulus, then the target stimulus, and after a fixation delay, they would have to answer a question about the target in comparison to the anchor (like if it was closer). They found that all three types of distances are represented in a region of the brain called the right inferior parietal lobule.

in the brain of two people are similar (e.g., correlated) or dissimilar (see also Section 10.3 on hyperscanning). This allows us to see how group membership influences underlying cognitive processes across people. For example, when people on opposite ends of the political spectrum (conservatives and liberals) view the same political video, their brain responses become polarized (Leong et al., 2020). In other words, in- and out-group ways of thinking can drive thinking that is even reflected by differences in the brain. At the same time, interacting more with others can cause our brains to become more similar. For example, the brain patterns of teachers and students to a lesson become more similar the better a student learns (Nguyen et al., 2022). I'm hopeful that after reading this book, you, me, and your professor will have strong neural synchrony!

12.5 Big Ethical Questions

The boom of social media has also presented new ethical challenges in terms of preserving privacy, and having bots that imitate interacting humans.

12.5.1 Too Much Information (on Social Media)

Because we often supply a lot of information about ourselves to social media platforms, these sites and apps are likely the most common source of privacy leaks across all of the methods and measures we are discussing in this book. This information can be used for ethically ambiguous targeted advertising. For example, in one newsworthy anecdote, a man learned that his teenage

daughter was pregnant after she received targeted coupons and ads for cribs and baby clothes in the mail, due to machine learning-based predictions from her shopping behavior at Target (Duhigg, 2012). However, even more nefariously, this information could influence health insurance or hiring decisions.

One fascinating example comes from a project analyzing public Venmo posts by data scientist Hang Do Thi Duc (2017). Venmo is a social payment app where basic information about people's transactions are public by default (showing the sender, recipient, and description, but often not the purchase price). Users have generally converged on using emojis to describe the nature of their transactions. Figure 12.4 shows examples of the types of profile information one might observe.

Profile 1:

Profile 2:

Jan 24 1:16pm	Sushi
Jan 24 4:24pm	Tempura
Jan 31 12:43pm	Ramen
Feb 14 1:46pm	Curry
Feb 14 4:38pm	Ramen

Figure 12.4 The publicly available information from two made-up Venmo profiles (but inspired by real profiles in *Public by Default*, Do Thi Duc, 2017). Profile 1 shows the emojis associated with each transaction. Profile 2 shows the date, time, and comment associated with transactions with a given food truck. What private information could be compromised about these two people based on these public profiles?

Discussion Question

What can you tell about each of these individual profiles in Figure 12.4?

You can learn a lot about a person based on their posting history, even if it consists of singular emojis. For example, you can see whether someone may have unhealthy eating habits, which could influence health insurers' decisions on how much of a premium to charge (Profile 1 in Figure 12.4). By identifying regular lunch transactions, you could predict exactly where and when to find a person (Profile 2 in Figure 12.4). A tax agency could also identify if someone is earning unreported income if there are repeated transactions from a business to that individual.

You can even learn information about a person that they do not disclose on their profile at all. In one study, Wang and Kosinski (2018) created a deep learning model (Section 6.6) that can predict someone's sexual orientation from a photograph of that person. To train this model, they downloaded and inputted 130,000 photos and the corresponding sexual orientation of the profiles from an online dating app. To test the model, they then showed it photos that were not part of that training set. The researchers found that with just five images of an

individual, they could predict a man's sexual orientation with 91 percent success, and 83 percent success for a woman. In contrast, humans only had a 57 percent accuracy for men, and 58 percent accuracy for women, and this accuracy didn't differ much by the sexual orientation of the observer. Thus, even by posting photos online, websites could make guesses about you or disclose information you would like to keep private, without your consent.

As of the time of publication of this book, there is still active discussion of whether laws can protect our privacy and redistribute the profitability of our private information into our own hands. The 2016 General Data Protection Regulation in the European Union has made great strides in protecting privacy. It requires that websites obtain clear consent from users to process their personal data, it gives users easier access to their own data, and it enforces the right to object to the use of personal data. Many countries have followed suit with similar laws, including Japan, Brazil, and Sri Lanka. However, there is also the complementary question of whether we should allow people to have complete anonymity online – or should we allow some limitations to privacy to protect public safety (e.g., preventing crimes).

12.5.2 Fake Social Interactions

Given the important sway that social media has on trends, decisions, political movements, and votes, there is great power in being able to control that sway. As a result, methods in AI (Chapter 6), crowdsourcing (Chapter 7), and natural language processing (Chapter 11) have enabled methods to make it appear as if a trend is popular. For example, if a product has many positive comments, likes, and is recirculated, you may think it is popular and well-loved and that you should buy it too. As a result, companies will invest large amounts of money in developing bots that can act as an army of people, boosting their products in the social media landscape. There are even some services where you can crowdsource out social media engagement, by purchasing likes, upvotes, comments, and reviews from real people, so their interactions are indistinguishable from honest consumers.

For example, I mentioned in Section 12.3.1 that tweets can be predictive of major political decisions, like Brexit. Knowing this, campaigns have utilized bots to try to inflate the popularity of their candidates on social media. In a study of the 2016 US presidential election between Hillary Clinton and Donald Trump, they found that 19 percent of tweets about the election were generated by bots (Bessi & Ferrara, 2016). They identified bots by finding profiles that lacked personalization to their profile, had a random username, had no geographic data, had nonstop activity around the clock, and those that posted more retweets than original posts. Bots were identified for both camps, and interestingly, most of the Trump bots posted positive sentiments about Trump, while most of the Clinton bots posted negative sentiments about Trump (rather than positive sentiments about Clinton). Celebrities may also use bots to boost their popularity. A 2019 study by the Institute of Contemporary Music Performance analyzed the celebrities with the most fake followers (the study has been since taken down, but see coverage: Hickman, 2019). They found that 42 percent of the Instagram and Twitter followers of soccer/football star Cristiano Ronaldo were fake. Similarly, 46 percent of Ariana Grande's, 46 percent of Taylor Swift's, and 43 percent of Kim Kardashian's followers are all bots. There is still no foolproof method to identify these bots, so as a consumer, it can be hard to know

what is real. It is thus important to keep these possibilities in mind when making a decision using solely information you glean from social media.

12.6 Applications of the Chapter

The analysis of social networks and graphs can also help reveal potential biases within a group of people. Looking at small worlds can help one quantify cliques and biases that could be happening across communities, universities, and companies. As an example, we can look at how biases might be perpetuated within universities. As we mentioned in Section 3.11, the demographic distribution of faculty conducting research is unrepresentative (i.e., mostly white males), which may result in biases in the sorts of research that is conducted. One of the reasons behind such biases is that faculty hiring may often occur from the same social networks, potentially amplifying any imbalances that exist. For example, one study analyzing 19,000 US faculty job hires across computer science, business, and history found that PhD programs from 25 percent of universities (specifically, highly ranked schools like Harvard, MIT, and Yale) produced on average 79 percent of the hires; see Figure 12.5 (Clauset et al., 2015). This suggests that academia is still very hierarchical, with a strong emphasis on university prestige (although see Miuccio et al., 2017).

It's worth thinking critically about the social networks that you are a part of, and those that you are studying. This can help you strategize how to recruit diverse participants for studies (Chapter 3). It may even be interesting to see if social network measures might account for any relationships you find in your studies. But also in general, thinking about your own network in these ways may help you broaden it (and find those weak ties and social bridges that can help you after graduation!)

CHAPTER SUMMARY

In this chapter, we learned about social network theory and several important phenomena studied in social psychology. We talked about online social media, what information you reveal in your profile, and how it causes psychological effects. Finally, we discussed our current neuroscientific understanding of social networks, and important ethical implications to keep in mind. Here are some of the main takeaways of the chapter:

1. A social network can be depicted as a graph of nodes and edges. Edges can be determined by a range of possible measures between two people. From this network graph, you can quantify several metrics about the network that let you learn about the group, its people, and their interactions.
2. Most social networks are "small worlds," where the network diameter is fairly small. Small worlds are then often connected by people who act as bridges, with many weak ties to people.
3. A lot of information can be predicted about you based on your social media profile, or the profiles of the people you are connected with.

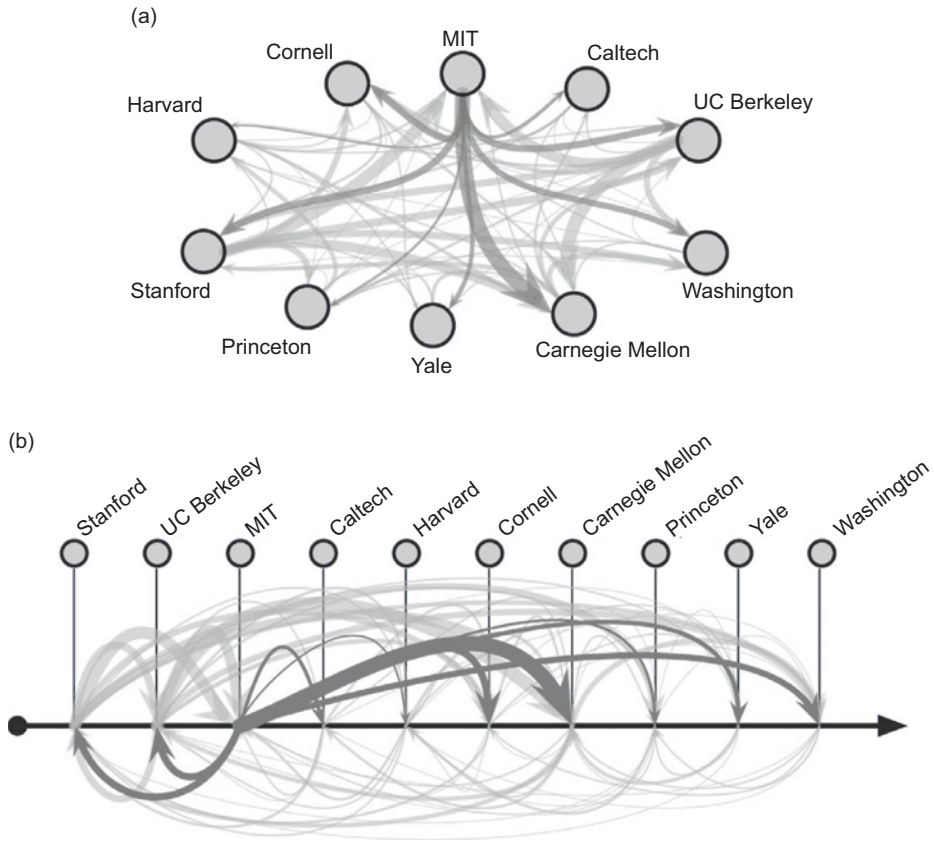


Figure 12.5 (a) A directed weighted graph showing number of computer science faculty hired from different PhD programs, with the base of the arrow representing the PhD program and the head of the arrow showing where they were hired. As you can see, there is a lot of interchange across these top PhD programs, and they also constitute a large portion of the hires at other American universities. (b) The graph data sorted on a line represents "prestige" based on how many outgoing faculty the university has versus incoming faculty from other institutions.
Source: Figure from Clauset et al., 2015.

4. Social networks may be represented similarly in the brain to networks of spatial distances and networks of time.

FURTHER READING

Here are some key resources to learn more about the topics discussed in this chapter:

- **Explore the public Venmo profiles on *Public by Default* and see what you can infer about these profiles:** Do Thi Duc, H. (2017). Public by default. https://publicbydefault.fyi
- **See how your brain gets more similar to your teachers' brains as you learn:** Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., & Hasson, U. (2022). Teacher–student neural

coupling during teaching and learning. *Social Cognitive and Affective Neuroscience*, *17*(4), 367–376.

- **Learn about network theory through the lens of *Game of Thrones* / *A Song of Ice and Fire*:** https://networkofthrones.com (Beveridge, A., & Shan, J. (2016). Network of thrones. *Math Horizons*, *23*(4), 18–22.)
- **Test whether you can find someone more than six degrees away from Kevin Bacon!** Reynolds, P. (1999). The oracle of Bacon. https://oracleofbacon.org

ASSIGNMENT

The purpose of this assignment is to give you experience applying basic concepts in network theory to your own social networks.

Total Points: 50

Let's look at a social network relevant to you.

1. **Decide on the main players for a social network relevant to you, and list them here.** This can be: your own social network made up of a collection of family, friends, and colleagues; the social network for a class or group you are in; or the social network in a book, movie, game, or TV series you enjoy. They will serve as the nodes of a network. There should be at least five nodes (and we recommend having fewer than twenty for simplicity). (4 points)
2. **List two measures** you can use to create the connections (or edges) between all pairs of individuals in the network. You can refer to some examples described in the chapter. **Are these edges directed or undirected?** (3 points)
3. Pick one of the two measures. Create a square matrix where each row/column is one of the nodes (people). In each cell, put the edge weight with the measure you selected for quantifying connections. **Put the matrix here.** (8 points)
4. **Draw out the network based on this matrix (with nodes and edges) and show it here.** (6 points)
5. **Answer these questions about the characteristics of your network:** (5 points)
   a. What is the **network density**?
   b. What is the **network diameter**?
   c. Who are the three people with the highest **degree centrality**?
   d. What are some examples of **weak ties** in the network? What are some examples of **strong ties** in the network?

   We will now look at how to extract information about someone based on their social network.
6. **List five types of information** you would be able to list for each node (e.g., their favorite sport). You don't need to actually list them for each node. (5 points)
7. **Choose a node with at least two connections.** (We will call this the "target node.")
   a. For each connected node (e.g., its "neighbors"), write out the responses for those five types of information (e.g., soccer). (5 points)
   b. Let's see to what degree we can predict something about a node based on its neighbors. Let's make a "predicted profile" for this target node. To create your predicted profile,

write down the modal response across the neighbors for each of the five types of information. The modal response is the most frequent response. If there is no mode, use a random number (e.g., roll a die) to pick a response from one of the neighbors. If the type of information is a number, instead calculate the median. **Write down here the predicted profile** for the five types of information. (5 points)

c. **Now, write down the "ground truth"** – the actual values for the five types of information for that target profile. (5 points)

d. How do the predictions compare to the actual profile (in what ways are they similar or dissimilar)? What types of information were more predictive, and which were less? (4 points)

REFERENCES

Aletras, N., & Chamberlain, B. P. (2018). Predicting Twitter user socioeconomic attributes with network and language information. *Proceedings of the 29th on Hypertext and Social Media*, 20–24.

Amador Diaz Lopez, J. C., Collignon-Delmar, S., Benoit, K., & Matsuo, A. (2017). Predicting the Brexit vote by tracking and classifying public opinion using Twitter data. *Statistics, Politics and Policy*, *8*(1), 85–104.

Backstrom, L., & Kleinberg, J. (2014, February). Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on Facebook. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 831–841.

Bagrow, J. P., Liu, X., & Mitchell, L. (2019). Information flow reveals prediction limits in online social activity. *Nature Human Behaviour*, *3*, 122–128.

Bessi, A., & Ferrara, E. (2016, November 7). Social bots distort the 2016 US presidential election online discussion. *First Monday*, *21*(11).

Bessière, K., Pressman, S., Kiesler, S., & Kraut, R. (2010). Effects of internet use on health and depression: A longitudinal study. *Journal of Medical Internet Research*, *12*(1), e6.

Beveridge, A., & Shan, J. (2016). Network of thrones. *Math Horizons*, *23*(4), 18–22.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*(2), 307–324.

Buttelmann, D., & Böhm, R. (2014). The ontogeny of the motivation that underlies in-group bias. *Psychological Science*, *25*(4), 921–927.

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014, April). Can cascades be predicted? *Proceedings of the 23rd International Conference on World Wide Web*, 925–936.

Clark, J. L., Algoe, S. B., & Green, M. C. (2018). Social network sites and well-being: The role of social connection. *Current Directions in Psychological Science*, *27*(1), 32–37.

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, *1*, e1400005.

Cuthbertson, A. (2018, June 26). Facebook patent predicts when you'll die. *The Independent*. www.independent.co.uk/tech/facebook-patent-predict-die-death-prediction-algorithm-personal-data-privacy-a8417771.html

Do Thi Duc, H. (2017). Public by default. https://publicbydefault.fyi

Duhigg, C. (2012, February 16). How companies learn your secrets. *The New York Times Magazine*. www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., & Nowak, M. A. (2012). Evolution of in-group favoritism. *Scientific Reports*, *2*(1), 460.

Gee, L. K., Jones, J., & Burke, M. (2017). Social networks and labor markets: How strong ties relate to job finding on Facebook's social network. *Journal of Labor Economics*, *35*(2), 485–518.

Goddard, I. (2023, October 12). What does friendship look like in America? Pew Research Center. https://tinyurl.com/27ra6px3

Goel, S., & Goldstein, D. G. (2014). Predicting individual behavior with social networks. *Marketing Science*, *33*(1), 82–93.

Gonzales, A. L., & Hancock, J. T. (2011). Mirror, mirror on my Facebook wall: Effects of exposure to Facebook on self-esteem. *Cyberpsychology, Behavior, and Social Networking*, *14*(1–2), 79–83.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Hickman, A. (2019, August 13). Taylor Swift, Kardashians and Neymar among celebs with most fake followers, finds new study. *PR Week*. https://tinyurl.com/2v4h9rtd

Hogeboom, D. L., McDermott, R. J., Perrin, K. M., Osman, H., & Bell-Ellison, B. A. (2010). Internet use and social networking among middle aged and older adults. *Educational Gerontology*, *36*(2), 93–111.

Hunt, M. G., Marx, R., Lipson, C., & Young, J. (2018). No more FOMO: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology*, *37*(10), 751–768.

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, *54*(5), 778–788.

Kuss, D. J., & Griffiths, M. D. (2011). Online social networking and addiction: A review of the psychological literature. *International Journal of Environmental Research and Public Health*, *8*(9), 3528–3552.

Leong, Y. C., Chen, J., Willer, R., & Zaki, J. (2020). Conservative and liberal attitudes drive polarized neural responses to political content. *Proceedings of the National Academy of Sciences USA*, *117*(44), 27731–27739.

Levenson, J. C., Shensa, A., Sidani, J. E., Colditz, J. B., & Primack, B. A. (2017). Social media use before bed and sleep disturbance among young adults in the United States: A nationally representative study. *Sleep*, *40*(9), zsx113.

Marques, J., Abrams, D., Paez, D., & Martinez-Taboada, C. (1998). The role of categorization and in-group norms in judgments of groups and their members. *Journal of Personality and Social Psychology*, *75*(4), 976–988.

Miuccio, M., Liu, K. Y., Lau, H., & Peters, M. A. (2017). Six-fold over-representation of graduates from prestigious universities does not necessitate unmeritocratic selection in the faculty hiring process. *PLoS One*, *12*(10), e0185900.

Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., & Hasson, U. (2022). Teacher–student neural coupling during teaching and learning. *Social Cognitive and Affective Neuroscience*, *17*(4), 367–376.

Oh, H. J., Ozkaya, E., & LaRose, R. (2014). How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. *Computers in Human Behavior*, *30*, 69–78.

Parkinson, C., Liu, S., & Wheatley, T. (2014). A common cortical metric for spatial, temporal, and social distance. *Journal of Neuroscience*, *34*(5), 1979–1987.

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., & Aral, S. (2022). A causal test of the strength of weak ties. *Science*, *377*(6612), 1304–1310.

Reynolds, M. (2023, October 5). Brits have less than 4 true friends, new study reveals. *Daily Express*. www.express.co.uk/news/uk/1820417/british-people-friends-study

Reynolds, P. (1999). The oracle of Bacon. https://oracleofbacon.org

Tan, C., Lee, L., & Pang, B. (2014). *The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter*. arXiv:1405.1438.

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A map for social navigation in the human brain. *Neuron*, *87*(1), 231–243.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, *32*, 425–443.

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040.

Index