Chapter 2

Data Science Is Transdisciplinary

Data science's primary progenitors are statistics, operations research, and computing, but the sciences, humanities, and social sciences are also all part of its story for these reasons:

1. **New application areas.** Data science adds valuable techniques to a large and growing number of application domains, building on their unique data and preexisting capabilities. The combined capabilities of data science and domain-specific know-how can solve important scientific and social problems and have commercial value.
2. **Advancing data science.** Some domains possess advanced methods for dealing with their data. Data science benefits by incorporating these methods and then making them available for wider use. For example, the petabyte-scale datasets generated by experiments in physics, astronomy, and biology led to inventing new data science techniques.
3. **Building coalitions.** Data science operates in a societal context. Making sure we "get it right" requires partnerships which must include viewpoints from non-STEM (science, technology, engineering, and mathematics) domains such as sociology, law, economics, philosophy, and political policy. Good solutions can have great societal benefit, while poor ones can cause harm.

This chapter's title includes the term *transdisciplinary* to emphasize that data science has been able to achieve its theoretical, methodological, and practical results by combining the approaches of different disciplines to create a new field.[6] This combination is needed not only for data science's core disciplines, but also for its application areas and the fields that influence its proper use.

[6] *Multidisciplinary* is when different fields separately contribute their approaches to a problem. *Interdisciplinary* is when the approaches interact. *Transdisciplinary* is when a new field emerges from that interaction.

https://doi.org/10.1017/9781009272230.004 Published online by Cambridge University Press

2.1 New Application Areas

We use two approaches to illustrate data science's broad applicability. First, we discuss data science's relevance to each economic sector. Second, we consider its relevance to academic research areas.

For the first approach, we divide the entire economy into major buckets, using the US Bureau of Economic Analysis's GDP Report as a guide (US Bureau of Economic Analysis, n.d.). Despite the data's US-centricity, the categories of economic activity are probably representative of most economies, and an analysis shows existing and growing data science roles in each one. Table 2.1 lists the categories and a few example data science applications for each.

While this book provides detailed examples from many of the areas, we have inevitably omitted some. Among others, we do not devote much attention to data science's many uses in national defense-related topics such as logistics, guidance and targeting, decision support, maintenance, or cybersecurity. We also don't discuss uses in precision agriculture, factory automation, building management, optimizing social service delivery, and so on.

For the second approach (relevance to academic research areas), we consider the role of data science in the various university education and research disciplines. Below, we discuss science, social science, engineering, and the humanities.

2.1.1 Sciences

The sciences have been a major source of data science use cases. As data becomes easier to use, scientific models have become more complex and highly tuned to real-world inputs. In some cases, data science has automated the process of creating models.

As an example, efforts to combat the COVID-19 pandemic show data science's increasing role in biomedical and social applications. Vaccine development and deployment could not have happened as quickly without pre-existing infrastructure, tools, and data ready to be used in a new situation:

- Genetic and protein structure databases and other tools for simulating structure facilitated the rapid decoding and promulgation of the underlying SARS-CoV-2 genetic structure.
- Tools were in place to manipulate this genetic data.
- Large-scale data management technologies were used to rapidly create, locate, manage, and monitor well-structured clinical trials.
- Logistics data and algorithms helped plan and control the complex supply chain for vaccinating large populations.

https://doi.org/10.1017/9781009272230.004 Published online by Cambridge University Press

Table 2.1 *Components of the economy and data science applicability*.*

Sector of US economy	Areas of data science applicability
Agriculture, forestry, and fishing	Precision cultivation and harvesting, fishery management, quality control, risk reduction
Mining	Predicting resource location, risk management, pricing
Utilities	Fault detection, optimized energy sources, production automation, predictive maintenance
Construction	Scheduling, logistics, optimized design and materials use
Manufacturing – durable goods	Quality control, production scheduling, automated design and manufacturing, customer support
Manufacturing – non-durable goods	Risk management, logistics, demand prediction, pricing
Wholesale trade	Inventory management, demand forecasting, logistics
Retail trade	Merchandising, advertising, pricing, upsell, loyalty programs, inventory management
Transportation and warehousing	Optimized routing, storage, pricing, tracing, safety monitoring, semi- and fully automated vehicles, yield management
Information	Advertising, audience engagement, content moderation, translation services
Finance and insurance	Risk assessment, portfolio construction, security, regulatory monitoring
Real estate, rental, and leasing	Construction, maintenance, property management, service automation
Professional, scientific, and technical services	Mapping and surveying automation, new software development tools, data-driven marketing, data-driven science
Management of companies and enterprises	Decision support of all forms, improved communications
Other administrative services	Employee hiring and scheduling, automated transcription, security monitoring, credit scoring
Education, health, and social assistance	Personalized education, remote health monitoring, disease diagnosis, social service delivery, fraud detection
Arts, entertainment, recreation, and food services	Personalization, pricing, upsell opportunities, automation, immersive experiences
Other highly diverse services	Fault diagnosis, dating services, locating civic needs, fund-raising
US national government – defense	Logistics, guidance and targeting, decision support, maintenance, readiness, cybersecurity, wargaming
US national government – non-defense	Tax audit, civic outreach, societal and economic monitoring
US state and local government	Maintenance operations, educational programs, criminal justice system, monitoring service fairness

* The first column divides the breadth of economic output into buckets. These come from the Bureau of Economic Research, which sources them from the North American Industry Classification System (NAICS) (US Census Bureau, n.d.). The second column is an eclectic list of data-science-enabled applications, either existing or soon to be likely.

https://doi.org/10.1017/9781009272230.004 Published online by Cambridge University Press

On the other hand, our COVID-19 experience showed some data science weaknesses: in particular, limitations in our ability to draw conclusions from public health monitoring and prediction. We give examples of these and other weaknesses throughout this book.

Aspirationally, Turing Award winner Jim Gray proposed a new model for scientific research, which he termed *the Fourth Paradigm*, in a talk at a 2007 National Academies meeting. In it, he talked about how science can increasingly benefit from new tools and techniques for data capture, analysis, and communication/publication. Gray was first and foremost a computer scientist specializing in databases, but he became involved in projects at the borders of astrophysics, mapping, and computer science. This led to his advising scientists in many disciplines on their growing data-related problems. As captured and edited by colleagues at Microsoft Research (see Gray, 2009), Gray said:

> The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science.

Science is increasingly moving in this direction, particularly with the rapid growth in machine learning capabilities. As an example, scientists trained a neural network on thousands of molecules with known antibacterial properties. They then applied that network to a dataset of over 6,000 compounds with potential antibiotic activity. This approach quickly uncovered a potential new drug, Halicin, for treating certain antibiotic-resistant bacteria (Stokes et al., 2020). While this data-centric approach is capable of screening far more than 6,000 compounds, and preliminary laboratory studies showed positive results, this particular drug may well encounter roadblocks on the path to approval (Rees, 2020).

2.1.2 Social Sciences

Much of the social sciences involves gathering and analyzing data. In economics, this was traditionally confined to the subfield of econometrics, which Samuelson explained as enabling one to "apply the tools of statistics ... to sift through mountains of data to extract simple relationships" (Samuelson & Nordhaus, 2009). However, today, data science more broadly impacts all of economics. The related area of finance has been at the forefront of using large datasets and sophisticated predictive models in diverse applications. We present specific examples from economics and finance in Section 6.5.

https://doi.org/10.1017/9781009272230.004 Published online by Cambridge University Press

In broader society, governments have always needed information about their populations. For example, at the most basic level, they need to count their people and collect taxes. They collect a great deal of other information as well; for instance, labor statistics, transportation statistics, and health data. From the other side, voters are interested in the results of pre-election polls. Polling agencies and social media platforms also collect a great deal of political and sociological data. More and more, information gathering underlies societal systems. We discuss some examples of political and governmental data science in Section 6.6.

In the future, perhaps all economic transactions will be digital, and physical currency will no longer be used. (China's early 2020s experimentation with a national digital currency may foreshadow this.) Economists may be able to measure and track all financial flows, and offer new mechanisms for economic governance. In a very ambitious proposal, data scientists and economists hypothesize that data science results could dynamically set tax policies to more efficiently balance revenue and equity objectives, as discussed in Section 6.5 (Zheng et al., 2020). A related, much simpler, present-day example is congestion pricing, in which highway tolls vary to reduce congestion and motivate use when there is extra capacity.

2.1.3 Engineering

Engineering disciplines abound with data science use cases, in both the design of new products and services, and the efficient operation of complex systems. For the former, data science and machine learning are the basis of GitHub Copilot, which assists programmers (Chen et al., 2021). Google researchers have developed a system that learns to do the physical layout of devices on computer chips (Mirhoseini et al., 2021). Civil engineers are likewise researching similar approaches to design far larger structures made of steel rather than silicon. Many engineering challenges, such as speech recognition, as discussed in Section 4.2 and Section 5.2, were practically unsolvable until we applied data-driven approaches to them.

In the domain of engineering operations, data, often directly provided by embedded instrumentation, can predict maintenance needs, provide early warnings of failures, and optimize system operations for applications, including cars, power grids, rails, naval propulsion systems, jet engines, and many more. Neural networks can classify work orders and optimize environmental and power systems. Continual anomaly detection is at the heart of many computer security systems, whether they are looking for operational failures or intrusions.

2.1.4 Humanities

In the humanities, applications have been slower to come into common use. The earliest days of computers did see occasional projects to connect digitized data with literature, art, and history. These early, one-of-a-kind, efforts all focused on single works. As large web-based archives of books and images were built, interest grew substantially in applying data science tools and techniques. Circa 2010, academicians began to recognize the opportunities (Kirschenbaum, 2012), and the US National Endowment for the Humanities launched its Digital Humanities Initiative. In 2009, the National Science Foundation (where co-author Jeannette was working at the time), in partnership with the US National Endowment for the Humanities, the Joint Information Systems Committee of the UK, and the Social Sciences Research Council of Canada, launched a "Digging into Data" challenge. It asked the question: "What could you do with a million books?"[7] It continues to this day, with 17 additional international partners.

Leveraging the millions of Google-scanned books, Google Research gave out awards in the digital humanities (Orwant, 2010), initially focusing on text analysis. J.-B. Michel et al. (including co-author Peter) used the Ngram viewer's frequency count of words or phrases in millions of books to gather insight on many societal or linguistic changes, ranging from the impact of censorship on books published to the rate at which irregularly conjugated English verbs change to become regularly conjugated (Michel et al., 2011).

Matt Connelly, circa 2015, created the History Lab Project at Columbia, which maintains the world's largest collection of declassified documents and lets researchers analyze them using data science techniques (Connelly et al., 2021). Legal scholars are analyzing the tens of millions of online court judgments released by China since 2014, to better understand the Chinese legal system and its rulings and consequences (Liebman et al., 2020). Many projects now geocode large numbers of records, placing people, historical events, or statistical measures on a map to provide insight or bring history to life. The Smithsonian Collections Search Center is an online catalog with 17.2 million records: "records relating to areas for Art & Design, History & Culture, and Science & Technology with over 6.6 million images, videos, audio files, podcasts, blog posts and electronic journals" (Smithsonian Institution, n.d.).

We could discuss many more applications in virtually every domain. We conclude this section by observing that data science's applications are broad and growing, while requiring the fusion of data science and discipline-specific capabilities to achieve their goals.

[7] A million books may not sound like a lot, but reading a book a day from birth to age 100 (give or take a day for leap years) amounts to only 36,525 books.

2.2 Advancing Data Science

As other fields address their data-related problems, they often end up making their own important contributions to data science. Necessity being the mother of invention, if a new problem requires a new capability, the field may be rapidly advanced.

A great example of a new capability coming out of physics is the World Wide Web. Sir Tim Berners-Lee created what became the World Wide Web to help the greater CERN supercollider research community communicate and collaborate. It turned out this was such a good idea that it rapidly caught on elsewhere and became a fundamental pillar of both data science and our everyday lives. As another example, many important technologies for efficient pattern matching came from computational biologists' need to match nucleic acid sequences. Finally, social scientists' need for census data is forcing consideration about making aggregations of data broadly and accurately available while still preserving privacy. We have a new understanding of privacy-preserving data aggregation – regrettably, one showing its difficulty.

Soon after the World Wide Web became available to the general public, online advertising became a big business. Advertising fees paid for the growth of many well-known internet services and also brought huge attention and money to data science. Deciding what ad to show to a viewer is a data science problem: there is data on what the current user is searching for and the pages they have been browsing, and there is data on what similar users have done in the past. Economists contributed by introducing data scientists to algorithmic game theory. Among other things, this helps determine what auction style is best for selling ad space, balancing interests of consumers, sellers, and web publishers.

2.3 Building Coalitions

Data scientists need to partner with many other disciplines to ensure their resulting work will be maximally beneficial, societally acceptable, or perhaps even legal. Here are six examples:

- **Philosophers** can help frame ethical considerations about data science work, including issues of privacy, free will, and fairness. These will be considered throughout this book, starting in the next section, as we lay out an ethical framework.
- **Lawyers**, **politicians**, and **political scientists** can help with the legal and policy issues relating to data stewardship and the governance of data-intensive applications. Public opinion can be strongly influenced by information publishing and recommendation systems.

- **Designers** and **psychologists** can help data scientists present aesthetically pleasing and accessible results that users find easy to understand and use.
- **Economists** (in microeconomics, behavioral economics, and more) bring economic modes of analysis to data science problems that can help achieve efficiency or fairness.
- **Sociologists** provide insights for new ways to study human behavior and social relationships using data from large digitally connected social networks.
- **Journalists** can aid data science's goal of explanation so it emphasizes truthfulness and is valuable to users and society. Computing a number or displaying a bar graph is not enough; the results need to tell a coherent and truthful story.