Chapter 4

Data Science Applications: Six Examples

In this chapter we present examples of what data science can do. For the technology, healthcare, and science-related examples, we define the problem and then show how to collect data, build a model, and use it to solve the problem.

We start with spelling correction, which is now so common that we hardly notice it. Its models are simple and its objectives are clear. We follow with speech recognition, which also has clear objectives, but requires considerably more complex models. In fact, speech recognition was considered a grand challenge problem in AI for decades. It only became widely practical circa 2012 when vast amounts of data and deep neural networks were applied to it.

Our third example, recommendation systems, may be the single most widespread use of data science. Recommendations have provided extreme value to companies and users alike, but setting their objectives properly is hard and subject to controversy. There are also many implementation challenges, which we will detail throughout this book. Of note, recommendation systems often make use of all six types of conclusions (prediction, recommendation, clustering, classification, transformation, and optimization) that data science offers.

Our fourth example, protein folding, predicts the shape of a protein just from the knowledge of its amino acid sequence. Progress in this grand challenge biochemistry problem was slow until 2020 when a broad ensemble of models using extensive protein databases were successfully applied. This problem differs from previous examples by the diversity and complexity of its modeling and its audience of scientists, not end-users.

Our fifth example is more general and presents the promise of using large quantities of individualized health data to learn about and improve human health. While we show the potential of data science in healthcare, we also illustrate the complexity of gaining meaningful results. Whether due to data quality, privacy, complex models, or the difficulties in determining causality, using healthcare records at scale is difficult.

https://doi.org/10.1017/9781009272230.006 Published online by Cambridge University Press

The sixth and final example in this section is cautionary and relates to predicting mortality during the COVID-19 pandemic. Despite high stakes, high visibility, and a great deal of data, epidemiologists and data scientists have had very limited success in predictions beyond a few weeks.

4.1 Spelling Correction

People make spelling mistakes; about 1% of words in documents are misspelled, as are almost 10% of words in internet search queries. With the help of data science, word processors and search engines can address spelling mistakes in three ways:

- as a **classification** task – misspelled words are identified, perhaps with squiggly red lines;
- as a **recommendation** task – possible corrections are presented to the user to choose from; or
- as a **transformation** task – misspelled words are automatically replaced with corrections.

The **dictionary approach** to spelling correction, introduced in the 1970s, starts with a dictionary of correctly spelled words. Then, for each word in a document or query, check if it is in the dictionary, and, if not, the system can either flag it, make a recommendation, or correct it to the closest dictionary word. In this case, the phrase "dalmation dogs" might just be corrected to "dalmatian dogs," since "dalmatian" is the dictionary word closest to "dalmation" at only one replacement letter away.

However, the dictionary approach has limitations. There are many things a dictionary does not cover. It may not have the latest slang words, nor proper names, so it can't help with "covidiot," "Shia Saide LaBeouf," or "Huawei." A dictionary-based corrector will say that every word in "from me too you" is a proper dictionary word. Of course, "too" should be corrected to "to," based on the context. And finally, a dictionary approach is not sensitive to the fact that different input methods produce different types of errors. For example, typing on a standard keyboard often leads to misordered characters (like "teh caret") while dictation instead often leads to homophone errors ("the carrot").

Because of these limitations, spelling correction has shifted to **data-intensive approaches**. A common data source is a **corpus** (or body of written text) compiled from text published on the Internet. Although published text is not perfect – it may contain its own spelling errors – it has the advantages of showing words in context and of including new words as soon as they appear in the language.

An international company that deals with 100 different languages and multiple sublanguages (e.g., British and American English, formal academic prose, and text messages) might find it tedious to maintain a dictionary for each one. It is much

https://doi.org/10.1017/9781009272230.006 Published online by Cambridge University Press

easier to gather a relevant corpus for each new use case than to hire a team of lexicographers to produce dictionaries for them.

Deciding that "beleif" should be corrected to "belief" and "dalmation" to "dalmatian" is easy because no other words were just one transposition or one substitution away. Other cases are not so clear-cut. What should "teh" be corrected to? It is a transposition away from "the" and a substitution away from "ten," "tea," "peh," and a dozen other words. Deciding which is most likely depends on two factors.

1. What word did the author likely intend? Corpus data can tell us that "the," English's most common word, is very likely and "peh," the 17th letter in the Hebrew alphabet, is rare and thus unlikely. Corpus data can also use the context from surrounding words: "in teh first place" suggests "the," while "teh, Earl Grey, hot" suggests "tea," based on phrase frequency in the corpus.
2. How likely is the user to incorrectly type "teh," given the intended word? The corpus alone can't tell us, but we can create a model that says that similar-sounding syllables, such as "tion" and "tian," are often confused, as are adjacent letters on the keyboard, such as "h" and "n." The model can learn from user interaction. Every time the user accepts a suggested correction, we can add to a new database of [*typo* → *correction*] pairs. We will find that ["teh" → "the"] appears frequently, and ["teh" → "peh"] does not. This database can be personalized for each user. Or it could be anonymized and shared, so everyone contributes to it and gets better spelling correction suggestions.

How fast can we assemble this user-interactions database? Suppose a search engine processes a billion queries a day, and 0.0001% of them mention a new celebrity or product. This would be more than a thousand new examples every day! Lexicographers trying to do this manually could never keep up. Once a model is created, it can be shared by all users of a language with some minor variations; e.g., British and American spellers disagree on "colour" versus "color"; users in one city might refer to "Clark Ave" and in another to "Clarke Ave."

Spelling correction is a good application of data science for several reasons. It shows the usefulness of gathering publicly available data from one source to use for another purpose, as well as the ease of creating a new database from user interactions. Privacy concerns are relatively easy to address because personally identifiable data need not be retained. Making an occasional mistake costs little, because the user is still in charge and can correct it.

4.2 Speech Recognition

Automated speech recognition (ASR), sometimes called speech-to-text, is the task of transforming an audio spectrum to digitized text. It has been of enormous interest

https://doi.org/10.1017/9781009272230.006 Published online by Cambridge University Press

to technologists and the general public for decades due to its many use cases, including taking dictation, automating call centers, enabling hands-free voice interaction with computers or appliances, captioning videos, and motivating great robots in Hollywood movies (Juang & Rabiner, 2005).

Speech recognition is a much harder problem than spelling correction for several reasons.

- Speech is analog, and each person's speech is a little bit different, whereas every typist who transposes the "ie" in "belief" ends up with exactly the same result, "beleif."
- Speech has systematic variation due to different accents, mixed language, speech impediments, random variation due to background noise, and multiple people speaking at once.
- Speech has ambiguity due to homophones as well as the absence of capitalization and punctuation.
- Transcription errors can be distracting to users, but are usually not life-threatening. Risks related to errors could prevent ASR's use in applications where an error would cause great harm.

As with spelling correction, early approaches relied on expert linguists. They wrote rules to define several steps in the speech recognition process pipeline; from acoustic signals to phonemes, from phonemes to words, and from words to sentences. This pipeline let linguists contribute their knowledge about language, but because each component was defined independently, errors propagated between components without correction.

By 1980, the field had shifted to automatically learning a speech model from data. Like spelling correction, it used statistics of past frequencies to analyze novel speech sounds. IBM speech researcher Frederick Jelinek jokingly said, "Anytime a linguist leaves the group the recognition rate goes up" (Jurafsky & Martin, 2009).

For speech recognition, the data are parallel corpora of speech spectra aligned with transcripts of the spoken words. The most valuable data (and most expensive to produce) is aligned word-by-word, but models also use sentence-by-sentence aligned and non-aligned samples.

In the 1980s, only people with a compelling need for the technology were willing to deal with speech recognition systems' idiosyncrasies, which included a tedious training period. Continuing improvements were made year by year, but the real breakthrough occurred in 2009, when Geoffrey Hinton of the University of Toronto and two of his students demonstrated the effectiveness of deep neural networks for speech recognition (Mohamed et al., 2009). Research teams at Microsoft, Google, IBM, and other institutions immediately jumped on this approach, which quickly resulted in a pronounced performance improvement in commercial systems. The

single change of introducing deep neural networks was more effective than all of the previous decade's work (Deng et al., 2013).

At first, the deep networks just replaced and improved individual components in the pipeline. However, by 2015, the whole pipeline had been replaced with an end-to-end neural network. One particular advantage of this approach was that early processing errors were carried forward in such a way that later parts of the network could correct the errors, in contrast to the error chains that linguists' pipelines allowed.

No two speech recordings are identical, but speech recognition systems can generalize, taking a novel audio spectrum and comparing it to its model of previously heard audio spectra, then outputting a transcript – even if the model had never heard that word sequence. With a good microphone, a careful speaker (with an accent sufficiently represented by training data), and minimal background noise, systems (as of 2022) make an error once every 30 or 40 words.

Just as in spelling correction, dynamic training can improve a speech recognition system's quality. Systems can learn to compensate for accents or microphone properties. They learn not only from recorded training data, but also from user-supplied corrections to system errors. As with spelling correction, a single system can be trained to recognize a hundred or more different languages.

We now use speech recognition for captioning videos, controlling home and automotive devices, and doing dictation – whether on computers, smartphones, appliances, or through smart assistants like Siri, Alexa, and Google Assistant. Speech recognition provides greater accessibility to hearing-impaired people, brings people of different cultures together with voice-to-voice translation systems, and perhaps even (someday!) will help call centers provide faster and better customer service.

In all, speech recognition quality has become so good that billions of people use it as an everyday part of their lives, so it now benefits from the virtuous cycle of increasing usage generating more data that improves quality and garners more usage. Like spelling correction, it's a good example of data science but one that is significantly more complex in many dimensions, including data gathering, modeling, breadth of application, and toleration of failures.

4.3 Music Recommendation

One of data science's most widespread uses is in recommendation systems. These systems recommend a user's next song, movie, book, app, or romantic partner. When someone browses a shopping site, the system suggests products they might like. On a news site or social network, they determine what stories to present to users.

Users depend on recommendation systems, because the Web has grown so large that no one can sift through all the available information on their own. Recommendation systems are also crucial to the Web's business model; better

https://doi.org/10.1017/9781009272230.006 Published online by Cambridge University Press

recommendations make for more subscriptions and purchases. Web advertising is in part a recommendation problem – one that must satisfy many different goals.

We focus on **music recommendations** as a representative example. With vast cloud-based music libraries at our disposal, including artists we have never heard of, recommendations truly help us find our way. How do we get recommendations of songs that we actually want to hear? The recommendation system builds a model from three types of data: a song's waveform, a song's metadata (title, artist, genre, composer, date recorded, length, etc.), and listeners' reactions. A "reaction" may be passively listening to the currently playing song, or it may be actively starting, skipping or replaying a song, or rating it with a star or a thumbs down.

A song's waveform can be analyzed for tempo, beat, timbre, and other factors. The system can recommend a song with similar features to songs that the user has previously liked or, for variety, perhaps recommend a contrasting song. The recommendations can be specialized for activity or time of day; perhaps fast, energetic songs for exercising, and slow mellow songs at the end of the day.

Metadata can be applied in many ways and even extended to permit creation of predictive, semantic relations between its entities. For example, a system could scan Wikipedia and other sources to learn that Telemann and Vivaldi lived at about the same time, Haydn taught Beethoven, Ringo Starr was a member of the Beatles, and Ramblin' Jack Elliott covered 24 Bob Dylan songs. If someone likes Telemann, Haydn, Ringo, or Dylan, they probably, respectively, also like Vivaldi, Beethoven, the Beatles, or Elliott.

User reactions help resolve the serious complication that different users should get different recommendations. There is no universally accepted "correct" recommendation in the way that there is a correct spelling of a word. Potentially every user needs a different recommendation model, rather than using a single shared model. This has two key implications:

- The system must ensure the privacy and security of each user's personal data.
- The data will be sparse. A large company with a billion users could gather enough data to build a good spelling correction system in a few days. But that is not nearly enough data to make good music recommendations. They would have billions of user reactions, but only a few for each user. A spelling correction system only has to learn about 100,000 words, but a large music recommendation system has to learn a billion users' preferences for each of a million songs – a quadrillion total preferences.

So, if a system has zero observations of a particular user reacting to a particular song, how can the system decide whether to recommend the song? The key is that it has many examples of *similar* users reacting to *similar* songs. The technique called **collaborative filtering** builds on this idea to examine the songs a listener has liked (or disliked), and compares them to every other listener's reactions. When it finds

a similar history, the songs that the user likes can be recommended to each other listener.

To improve both efficiency and the ability to generalize, a system can group together both similar songs and similar users. A user might belong to groups for "cool jazz" and for "60s female Broadway vocalists." A machine learning system wouldn't know those groups by those names, but rather by their shared collection of numerical and categorical features. The important point is that when some group members agree on a new discovery, it becomes available to all the other group members. Many companies use collaborative filtering, with Amazon and Netflix particularly well known for it.

A second technique to address the "cold start" problem of new users (or new items) is to explicitly pose the recommendation problem as one of **stochastic optimization** or **reinforcement learning**. For example, in the domain of news, many techniques have been developed in the last decade for trying to find the right combination of attributes of the always-changing news stories, users, and their reading histories to maximize the click-through rate (Coenen, 2019; Li et al., 2011). These approaches balance the twin goals of enabling a reader to explore new stories while leveraging popularity.

Music recommendation methods are metaphorically similar to those that underlie quantitative approaches to investment management:

- Momentum investing based on understanding what others are doing is analogous to collaborative filtering.
- Fundamental investing based on knowledge of an investment's business is analogous to using semantic knowledge.
- Technical investing based on raw stock prices and market volumes is analogous to musical signal analysis.

Quantitative investment applications, like recommendation systems, can also learn from their success or failure and feed that back into future decision-making.

Our discussion of music recommendation systems shows that they may make use of all the conclusion types listed in Chapter 1 to make their own recommendation conclusion:

- They may *predict* what a user will like or not like.
- They may *cluster* users or songs into groups, to identify like elements or to do better collaborative filtering.
- They may *classify* songs by genre to make better recommendations or use the genre itself within the user interface.
- They may *optimize* customer satisfaction subject to meeting certain constraints, such as presenting sufficient new material or maximizing revenue.

- They may *transform* signals into a new form. For example, they may transform audio signals into a set of features to allow better comparison of music's sound.

Music recommendations are greatly differentiated from the previous examples by their diversity of models, types of conclusions, and the complexity of setting objectives. (See the case study of Pandora's approach as an example (Dorman, 2018).) Many model systems of this kind will be much larger, harder to maintain, and harder to debug. We will return to recommendations throughout this book due to their broad applicability, their financial importance, the challenges in setting their objectives, and their diversity of techniques they employ. In particular, Section 6.2 and Chapter 7 discuss news recommendations.

4.4 Protein Folding

A human protein is a stringy chain of amino acids connected in a specific order. As postulated by Anfinsen in his 1972 Nobel Lecture (Anfinsen, 1973), the identity and order of the amino acids determines a protein's shape. We also know that shape largely determines how the protein functions and what it does. Protein shapes can often be experimentally determined with techniques like X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy. However, due to this work's time and expense, only about 100,000 protein structures have been determined out of the billions known to exist.

This motivates one of the greatest biochemistry challenges, the **protein folding problem**, which in part aims to predict the 3D structure of proteins directly from sequence data. More easily discovering protein structures would greatly benefit the life sciences, providing better understanding of many diseases and faster drug discovery. To calibrate the problem's scale, humans have 20 different amino acids and a protein is made up of tens to thousands of amino acids, and a protein could be in over 10^{100} possible shapes.

At first, scientists attempted to solve this structure prediction problem by building on fundamental physical principles. These so-called ab initio techniques provided some insight but proved computationally challenging for even the largest computers, peer-to-peer networks, and specialized hardware architectures. Levinthal's paradox contrasts the vast computational difficulty of ab initio modeling against how easily proteins fold in Nature. This disparity ("if Nature can do it . . .") has led many to believe there must be efficient approaches.

Such approaches have begun to appear based on models using (i) machine learning trained on known protein sequences and structures in combination with (ii) underlying physical principles. In 2020, at the 14th Biennial CASP (Critical Assessment of Protein Structure Prediction) competition, DeepMind's AlphaFold 2

https://doi.org/10.1017/9781009272230.006 Published online by Cambridge University Press

software used such an approach and achieved results having similar quality to those using laboratory techniques (Jumper et al., 2021). Combining several types of deep network architectures, the system produces predictions in minutes to hours of GPU processing. While we don't know how well the results will generalize, AlphaFold 2's result is a clear turning point in the protein folding problem's nearly 50-year history.

In July 2021, DeepMind released all of AlphaFold 2's documentation, models, code, and training data for others to scrutinize, learn from, and use its capabilities. Alphabet, DeepMind's parent company, is planning to use its software to determine and then release hundreds of thousands of protein structures, with plans to grow this to over a hundred million. Importantly, also in July 2021, Baek et al. of the University of Washington released RoseTTAFold, a protein folding solution that used similar approaches, adding more credence to this form of modeling (Baek et al., 2021).

As good as these results are, there is still more to do. Time will tell how well data-driven approaches generalize to predicting multiple protein complexes or the shapes that occur from protein interactions. Also, they do not show the dynamic protein changes that might be seen using molecular dynamics (Shaw et al., 2010). As of late 2021, AlphaFold 2 has less accuracy at the sites where proteins bind with ligands. It is also unknown if data-driven approaches can predict protein shapes when there is no training data from evolutionarily related, naturally occurring proteins (Mullard, 2021). Finally, while its accuracy is approaching that of laboratory techniques, it can still be improved.

Nonetheless, AlphaFold 2's success, as well as that of other related work, is a concrete demonstration of data science's ability to leverage decades of experimentally derived data to advance biomedicine and health. The Oxford Protein Informatics Group published an informative summary post with more information on this (Rubiera, 2021).

We have included this example because it is a great example of a scientific application whose results will be used by scientists, not the population at large. It also relies on a very sophisticated collection of machine learning and other models, which have been trained on decades of painstakingly collected data. Even though its generality is not fully known, this protein folding solution will inspire other scientists and demonstrate that data-driven approaches are now a necessity in a scientist's toolkit.

4.5 Healthcare Records

Individual patient healthcare records, including test results, diagnoses, treatments, and their results, are increasingly digitized and available online. This could improve our ability to learn about disease prevalence in populations, different

https://doi.org/10.1017/9781009272230.006 Published online by Cambridge University Press

approaches to treatment, and other ways to improve human health. Vast amounts of data can be brought to bear; large healthcare institutions, such as the Kaiser Permanente Health System and the United States Veterans Administration, have circa 10 million enrollees. National systems have even more – the UK's National Health Service (NHS) has circa 58 million enrollees, whose data is made available for analytics purposes on the OpenSAFELY platform (OpenSAFELY, n.d.). An international collaboration, the Observational Health Data Science and Informatics (OHDSI), aims to make it possible to apply data from an international set of federated data stores. As of 2016, OHDSI had converted over 682 million patient records to a standard and comparable format. This included at least some information about an estimated 200 million patients (Hripcsak et al., 2015). Here are three illustrative results:

- A Kaiser Permanente study used machine learning on nearly four million patient records to train a predictive model to identify patients more at risk for developing HIV, and hence more likely to benefit from taking preventative medication (Marcus et al., 2019). Kaiser, as a very large integrated healthcare system, benefits greatly from its large unified data repository, making Kaiser's studies easier to do than most in the US healthcare system.
- The OHDSI database supported a retrospective study comparing two different common hypertension treatments across over 730,000 patients. It showed that the more highly recommended treatment is associated with more side effects (Hripcsak et al., 2020). This study does not yet prove the standard of care is wrong, but it does alert physicians to watch out for the side effects. It also increases the attention on an ongoing prospective controlled trial comparing the alternatives.
- During the 2020 focus on COVID-19 treatments, OHDSI's database and tools looked at two decades of records relating to the side effects of previous hydroxychloroquine and azithromycin uses. While there had been considerable hope that the combination of these two drugs would help treat COVID-19, the data showed troubling associations with cardiovascular problems, including ones resulting in death (Lane et al., 2020).

More generally, big data systems with many healthcare records enable visualizations and understanding of the natural history of disease and prevailing treatments. Researchers can then "design experiments and inform the generalizability of experimental research" (Hripcsak et al., 2016).

Another approach to using healthcare records at scale is based on crowdsourcing. An excellent use case is post-approval monitoring of new drugs to identify rare side effects that might not show up in more limited pre-approval trials. As an example, the US Centers for Disease Control created the V-Safe system to monitor

https://doi.org/10.1017/9781009272230.006 Published online by Cambridge University Press

post-COVID-19 vaccination side effects, using both text messaging and web technologies to increase coverage (CDC, 2021). With a potential patient base about four orders of magnitude larger than the populations in vaccine clinical trials, V-Safe aimed to provide very rapid information on vaccine reactions and low-prevalence risks. In late summer 2021, V-Safe was augmented to gather highly specific side-effect data from pregnant women, a particularly important subpopulation.

Within a one-month period ending January 13, 2021, 1.6 million US vaccine recipients (representing 10% of the total vaccinations given in the period) completed at least one survey. The rapidly available data allowed publication of both minor and serious reaction rates by mid-February (Gee et al., 2021). The survey showed a very small risk of serious adverse reactions, though about 80% of respondents reported pain at the injection site. The relatively high response rate reduced the risk of selection bias (though it is still possible), but V-Safe, like all crowdsourced systems, could be abused by individuals who register false adverse reactions – possibly to create fear, uncertainty, and doubt. In fact, the broader US vaccine adverse event reporting system has become considerably more politicized in the time of COVID-19 vaccine skepticism, and its data is increasingly being used out of context (Stecuła & Motta, 2021).

Results such as the preceding examples hide the complexity of doing such observational studies. Data needs to be sufficiently standardized, possibly across multiple sites, to be usable in combination. To support data gathering and use, there must be complex data management software that aggregates data while both preserving privacy and allowing for the transparency needed in scientific studies. When data is gathered from multiple sites, due attention must be paid to the risks of fraud. Sites are likely to anonymize or aggregate data prior to release, making validation more difficult.

Enormous attention must be paid to modeling. Observational studies are usually done with a hypothesis in mind, and there is inevitably pressure to show positive results. Even when these pressures are controlled, the complexity of the statistical analyses can lead to hard-to-find errors.

The privacy issues, particularly in systems like OHDSI which need to use cross-site data, are particularly challenging. Tools and methodologies are needed to let researchers evaluate the many possible correlations they identify and develop cost-effective approaches to prioritize the data gathering needed to shed light on causality. However, despite the frequent desire and need to do so, it is extremely difficult (and frequently impossible) to tease out matters of causality in observational studies.

As an example of where something went terribly wrong, well-regarded medical researchers published articles in May 2020 in two of the most prestigious journals

using observational data from then-recent COVID-19 cases (Offord, 2020). One drew conclusions on the risks to hospitalized COVID-19 patients of common blood pressure medications, and the other on the risks of hydroxychloroquine (uncombined with azithromycin) as a treatment. However, the underlying data was not available for peer review, and the articles were eventually shown to be erroneous, if not fraudulent. Both articles were quickly retracted, but real damage happened when organizations briefly suspended research projects and modified patient treatment guidelines based on the work.

Many factors make observational studies challenging, but there are huge opportunities. The near-universal collection of healthcare records can suggest new hypotheses, support post-approval monitoring of new drugs, provide an interactive analysis platform for researchers to explore new ideas, catalyze new approaches for screening or preventing disease, and sometimes answer critical healthcare questions. However, frequently, issues of data quality and availability, modeling complexity, privacy and security, the difficulty in determining causal relationships, and more make these applications difficult. These challenges are discussed in much greater detail throughout Part III.

We've chosen the application of healthcare records at scale to illustrate the enormous opportunities in using observational data, but also the great challenges (e.g., scale, complexity, and potential for harm) to creating truthful insights and conclusions.

4.6 Predicting COVID-19 Mortality in the US

After the early reports of lockdowns and thousands of deaths in China, everyone hungered for predictions about SARS-CoV-2's impact in their own regions. In addition to informing the general public of expected risks, morbidity and mortality predictions could better guide institutions and governments to needed actions. Even better, if models could predict the effect of policy interventions (e.g., masks or levels of quarantine), they could help balance conflicting economic, social, educational, and health objectives. In the US alone, the COVID-19 Forecast Hub (COVID-19 Forecast Hub, n.d.) hosted more than 50 different predictive models, and there were numerous COVID-19 data science efforts elsewhere (von Borzyskowski et al., 2021).

In many ways, mortality prediction might seem a straightforward exercise. There was publicly available aggregate data: the number of COVID-19 positive tests, test positivity as a function of total tests, and number of deaths. There was also a growing understanding of disease transmissibility as it became clear that asymptomatic people spread COVID-19 through virus-laden aerosols. By the late spring 2020, there was initial data on seroprevalence of antibodies to COVID-19, which

could be used to infer how many had been exposed to the disease. There were measures of mobility (as discussed further in Chapter 6) that showed the impact of quarantine regulations, and many more potentially useful features on which to base models.

There was also considerable modeling experience. Previous epidemiological modeling work dates back to Bernoulli's smallpox study in the 1700s (Dietz & Heesterbeek, 2002) and has continued to the present. There are many possible types of models, ranging from susceptible–exposed–infectious–recovered (SEIR) compartmental models (which are interpretable and based on our understanding of disease spread) to machine learning models (some simple, some using many features). Models were illustrated by excellent graphs and charts available in both the scientific literature and the press. These provided exploratory and explanatory insights to data scientists and epidemiologists.

The objective of this predictive modeling was clear, privacy issues were muted (since the data used for modeling is already highly aggregated), and many people would have accepted good predictions without needing an explanation. However, we do acknowledge that Cornell's COVID-19 modeling team emphasized the need for interpretability to increase the acceptance of campus health policies (Frazier, 2022).

However, modeling did not go smoothly for many reasons. First, data was lacking. In many countries there was reasonable data on the number of hospitalizations and deaths but insufficient testing to understand how many people were infected with milder cases. Changing testing availability also made data hard to compare across time intervals; i.e., the number of undiagnosed cases very early in the pandemic was much higher than at many other periods. Mortality was measured in different ways in different jurisdictions (e.g., co-morbidities led to inconsistent policies for attributing death), making that data noisy.

Furthermore, modeling based on reported cases is innately difficult. Infected people may harbor latent disease for a few days or longer before having symptoms, yet still be contagious. In fact, some infected with SARS-Cov-2 never had symptoms but still spread the virus. Viruses mutate, and transmissibility increased during the pandemic.

Popular behavior also changed greatly over time. This was due to changing perception of self-risk, governmental actions, and perhaps even as a direct or indirect result of model predictions on the public.

Finally, the available aggregated data did not match the actual subpopulations that arise due to the cultural affinity, employment, education, or shared circumstances that bind people together. Mortality in these subpopulations, where people differentially interact amongst themselves, can greatly skew societal averages, as happened, for example, in nursing homes at the beginning of the epidemic. Due to

these and many other characteristics of the COVID-19 epidemic, modeling was problematic.

In an excellent comparative review, a community of 229 co-authors wrote a retrospective evaluation of 22 US COVID-19 mortality modeling efforts occurring from May to December 2020 (Cramer et al., 2021). The data on which the models were based varied. All but one used data on prior deaths, many used data on positive cases, and some included data on hospitalizations, demographics, and mobility data. A few models assumed that behavioral patterns might change during the modeling period, but most did not. Some models were based on the characteristics of disease spread (as in the SEIR approach mentioned previously), but most were not.

The paper's methodology compared model predictions to actual data, and also compared how accurate those predictions were to those made by a 23rd model – a naive baseline model with predictions based solely on past deaths. Against that baseline, about one-half of the models did better and about one-half did worse.

Cramer et al. (2021) observed that models with simple data inputs (e.g., positive case and mortality data) were some of the most accurate stand-alone models (which is vaguely depressing to this book's authors). An ensemble model, which equally weighted forecasts from all the available models, gave the best results for one-to-four-weeks predictions, with roughly one-third less error using the metrics of evaluators. While this seems like a very short period of prediction, it is still of value in terms of allocating treatment capacity to needy areas.

Longer-range forecasts had lower accuracy. Four-week-ahead forecasts had roughly twice the error of one-week forecasts; eight- to 20-week horizons had about five to six times higher errors. The longer-range forecasts, if better, would have been very useful in setting policies. Unfortunately, according to Roni Rosenfeld at CMU (one co-author of the Cramer et al. paper): "There were 4 major geo-temporal COVID waves in the US in 2020, and none of them was anticipated by any of the forecasts I have seen (ourselves included)" (R. Rosenfeld, personal communication to Alfred Spector, July 18, 2021).

If we reflect on the relatively poor results of these modeling efforts, they happened due to insufficient and erroneous data, the complexity of the necessary models, the changing nature of the disease, and feedback phenomena catalyzed, in part, by government actions. Rectifying these problems would be very difficult due to the logistics and privacy implications of gathering very fine-grained data and predicting public policy/societal responses. Furthermore, the virus's mutation may have stymied predictions in any case.

We concluded this chapter's examples with COVID-19 mortality prediction to show our humility in the face of very difficult problems. However, we do hold out hope that this data science application could significantly improve with more data and effort.