Chapter 5

The Analysis Rubric

This chapter defines the **Analysis Rubric**, which consists of seven major considerations for determining data science's applicability to a proposed application. While these considerations may not be fully understood at a project's inception, there needs to be a belief that answers will be forthcoming prior to completion. Three of these address requirements-oriented aspects ("for what or why") of data science applications, and three address implementation-oriented aspects ("how to"). The seventh addresses legal, societal, and ethical implications (ELSI[8]). Collectively, these considerations, or Analysis Rubric elements, cover the complex trade-offs needed to achieve practical, valuable, legal, and ethical results.

Implementation-oriented elements

- **Tractable data.** Consider whether data of sufficient integrity, size, quality, and manageability exists or could be obtained.
- **A technical approach.** Consider whether there is a technical approach grounded in data, such as an analysis, a model, or an interactive visualization, that can achieve the desired result.
- **Dependability.** Dependability[9] aggregates the following four considerations. Does the application meet needed privacy protections? Is its security sufficient to thwart attackers who try to break it? Does it resist the abuse of malevolent users? Does it have the resilience to operate correctly in the face of unforeseen circumstances or changes to the world?

Requirements-oriented elements

- **Understandability.** This means the approach must enable others to understand the application. Consider whether the application needs only to provide

[8] The acronym ELSI stands for "ethical, legal, and social implications." It was coined by James Watson in October 1988 as described in *ELSI: Origins and Early History* (Sankar, 2014). We will typically address these issues in a more operationally focused order that begins with legal issues, followed thereafter by societal and ethical issues.

[9] We devoted much effort before settling on *dependability* to aggregate privacy, security, abuse-resistance, and resilience. While *dependability* is often a generic term, this book will consistently use it as a placeholder for these four properties.

https://doi.org/10.1017/9781009272230.007 Published online by Cambridge University Press

conclusions or if it will have to explain "why" it has rendered these conclusions. Will the application need to detail the causal chain underlying its conclusions? Or will it make its underlying data and associated models, software, and techniques transparent and provide **reproducibility** – that is, the ability for analysts or scientists to understand, validate, duplicate, or extend the results?

- **Clear objectives.** Consider whether the application is trying to achieve well-specified objectives that align with what we truly want to happen.
- **Toleration of failures.** Consider both the possible unintended side effects if the objective is not quite right and the possible damage from failing to meet objectives. Many data science approaches only achieve good results probabilistically, so occasional poor results must be acceptable.

Ethical, legal, and societal implications (ELSI) element

- **Ethical, legal, and social issues.** Consider the application holistically with regard to legality, risk, and ethical considerations. Many of the topics under "Dependability" or "Clear objectives" topics are relevant, but this holistic analysis is broader.

Many applications start with a **bottom-up approach**, focusing on implementation-related Analysis Rubric elements relating to data availability, a technical approach providing the necessary results, and techniques to provide needed dependability. This analysis then informs the requirements definition and influences its refinement.

Others require a **top-down approach**, first focusing on the requirements-oriented Analysis Rubric elements relating to understanding, clarity of objectives, and failure tolerance. This analysis then informs the implementation approach and influences its refinement.

Most commonly, the bottom-up and top-down approaches are mixed, and there is iterative flitting back and forth between different considerations. No matter what design approach is used, the ethical, legal, and societal implications must be considered throughout the design and analysis. They cannot be bolted on at the last minute, and they must be carefully reviewed before any effort is declared complete.

The Analysis Rubric is important to this book. It is illustrated in Figure 5.1, which summarizes its considerations in a graphic. The next six sections will make the Analysis Rubric more concrete by demonstrating its application to the six examples of Chapter 4.

5.1 Analyzing Spelling Correction

Spelling correction is a clear example of a really good data science application – as evaluation using the Analysis Rubric shows.

https://doi.org/10.1017/9781009272230.007 Published online by Cambridge University Press

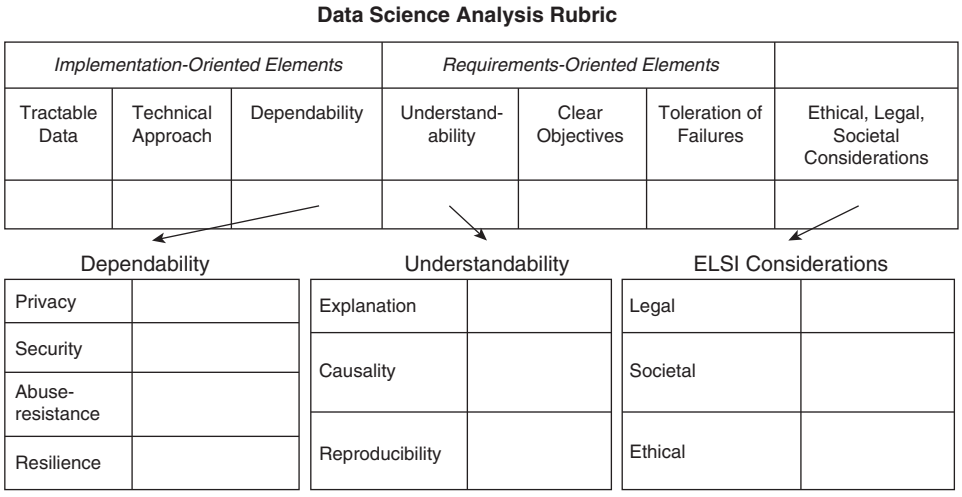

Figure 5.1 This graphic shows the seven top-level elements of the Analysis Rubric and the further breakdown of Dependability, Understandability, and Ethical, legal, and societal considerations.

- **Tractable data.** Anyone can easily collect an appropriate corpus of online text. A company already running a service can easily collect user feedback from spelling suggestions to verify which suggestions are good. There are "only" a few million distinct word tokens in any language, so individual word count data is relatively small. However, multi-word phrase data quickly grows in size – the Google Books Ngram project has a few hundred gigabytes of data for counts of phrases up to five words long.
- **A technical approach.** Section 4.1 outlined an approach to spelling correction in a search engine using word and phrase frequencies in the search corpus, together with user feedback from accepting or rejecting suggestions. The model is relatively simple, and a basic version takes just a few dozen lines of code (Norvig, 2009).
- **Dependability.** Spelling correction relies mostly on non-private data, so privacy and security are not major issues. However, privacy is always tricky, and a system that learns from an individual or institution should not expose confidential information (such as the spelling of code names) to outsiders. Erroneous corrections may occur, but the cost of a spelling error is low. Some care must be taken to prevent an attacker from spamming the spelling corrector with an incorrect spelling (perhaps to promote their brand name).
- **Understandability.** Users don't really care how spelling correctors work. Spelling correctors also don't need to understand a spelling error's root cause. Finally, a spelling corrector's internal operation can be opaque. Neither must its

https://doi.org/10.1017/9781009272230.007 Published online by Cambridge University Press

inner workings be understandable nor must its logic and data be published. This is good, because the specific words each individual user types must be kept secret.

- **Clear objectives.** The clear goal is determining and providing the correct spelling. While a spelling corrector could correct "wheg" to many different words such as "when," "where," or "Whig," the correct spelling is what the user *meant* to type.
- **Toleration of failures.** While a spelling corrector should almost always do the right thing, almost all users are accepting if it does not correct a word's spelling or even guesses a word incorrectly, as long as the failure is plausible. However, even a rare failure that "corrects" words to become profane or otherwise objectionable would be unacceptable.
- **Legal, risk, and ethical issues.** Spelling correction would seem to have no legal issues and minimal risks (as long as it does not inappropriately suggest taboo words). Spelling would not seem to have ethical concerns, though Nicholas Carr questions whether automation of mundane things is harmful to us as humans (Carr, 2008), and Nick Romeo questions the impact of spelling correction, per se (Romeo, 2014). However, Romeo observes this type of concern is old, and references Plato's *Phaedrus*, which poses the question of whether even the written word might reduce people's memories and make them dependent. We think the advantages outweigh any disadvantages, and autocorrection could even teach us to spell.[10]

5.2 Analyzing Speech Recognition

Speech recognition has some similarities to spelling correction, despite it being a much harder technical problem.

- **Tractable data.** Speech recognition's most important data sources are the repositories pairing speech utterances (recorded waveforms) with correct transcriptions on which the system can be trained. These utterances may have been professionally spoken and transcribed or may have been mined from usage. While recognizing speech on behalf of a user, a speech recognition application may also consult repositories relating to local specialized word stores from the user's common vocabulary (such as personal or place names) and more. Just as in spelling correction, the repository may include user-provided corrections, for both improved overall speech recognition accuracy and better individual adaptation.

[10] A system could even remember the "meaningful" spelling corrections it has made for us, and periodically remind us of them, perhaps even teaching us with virtual flashcards.

- **A technical approach.** While data-oriented approaches have long been applied, the quality improvement brought by deep neural network recognizers made them the dominant approach. These systems initially used cloud computing and specialized hardware for the audio-to-text transformation step. This required complex engineering to efficiently transmit waveform data from a personal device to the Cloud and then to receive the resulting text. That round trip requires reliable communication (particularly for dictation), but now even cell phones can do many aspects of speech recognition locally, thereby reducing off-device processing.
- **Dependability.** When speech recognition systems collect utterances either for personalization or for their overall improvement, they must take great pains to protect those utterances. This is particularly difficult when the process uses human transcribers.[11] Depending on the application, speech recognition will make errors, as do even the best human transcriptionists. Many speech recognition systems learn from user feedback, such as user-supplied corrections, which makes them more resilient and adaptable, but also means they must protect against abuse in ways similar to spelling correction.
- **Understandability.** As with spelling correction, speech recognition need not be concerned with explanation, transparent reproducibility, or causality.
- **Clear objectives.** Here, the objectives are a bit more complex than spelling due to the need to consider differential accents, speech impediments, and multilingual speech. Speech recognition systems, if to be of broad utility, must consider word accuracy across an entire population. Furthermore, the objective may have occasional ambiguities, but usually the right answer can be understood from context.
- **Toleration of failures.** Speech recognition's failure tolerance depends on the application. In some cases, failures up to a certain rate are acceptable. However, in some situations, such as transcribing a judicial hearing, controlling a vehicle, or basing emergency response off of a 911 transcription, errors could cause substantial harm, necessitating mitigating techniques (e.g., manual checking). Thus, failure tolerance is application-dependent.
- **Legal, risk, and ethical issues.** Speech recognition creates minor risks except in critical safety applications. There is a privacy issue if speech recordings and/or transcripts are transmitted or stored. Therefore, many applications do speech recognition on-device without retaining recordings. There have also been controversies when humans have listened to speech utterances to do transcriptions to provide more training data. Machine-learned speech recognition models tend to

[11] There have been privacy concerns with speech utterances being sent to outside contractors for manual transcription, so speech recognition providers have had to both increase their disclosure to users and change their approach to manual transcription.

https://doi.org/10.1017/9781009272230.007 Published online by Cambridge University Press

get more data and perform better for majority populations of speakers, and may perform poorly for subpopulations. Supporting these subpopulations is beneficial to all, and the balance of effort to do so is a fairness issue that must be considered.

5.3 Analyzing Music Recommendation

Recommendation systems are a broader data science application than our previous two examples. They have far more diverse uses, employ many more underlying techniques, and, while they focus on prediction, they may do many types of learning to achieve it.

- **Tractable data.** Music recommendations' underlying data sources are quite heterogeneous. Here are some example datasets that a music service would have available:
  - Recommendations and click data histories, indicating what was recommended and what was accepted and/or listened to.
  - A semantic information database about music, musicians, and musical periods. This can be compiled from many sources, ranging from record label data to Wikipedia.
  - Metadata about the music recordings and performances.
  - The music's audio tracks.

There have been millions of separate CD albums (Wikipedia contributors, n.d.). Assuming about 10 million albums at less than 100 megabytes per album (for, say, MP3 files), we have a corpus of about a petabyte of sound. As of 2021, this can be stored, albeit without redundancy, on a mere 100 hard drives!

- **A technical approach.** As described in Section 4.3, there are many approaches to creating successful recommendations, and they can be assembled into an ensemble to gain their collective value. Approaches to broader recommendations vary based on the specific application, but there are usually many available techniques. Often the choice is governed by how they address the objectives and need for understanding.
- **Dependability**
  - Systems that maintain a history of user interaction must very carefully protect it to prevent security or privacy problems. This is not so easy; Arvind Narayanan and Vitaly Shmatikov in the mid-2000s showed that even highly de-identified data from Netflix usage databases could divulge sensitive information (Narayanan & Shmatikov, 2007).
  - Resilience is not a great concern given that recommendations need not always be accepted, though there must be guards against really terrible (or jarring) recommendations that would be greatly disliked by a listener.

- Music (and almost all) recommendation systems must include anti-abuse technologies, as bad actors can trick them into recommending something. Unlike with spelling correction or speech recognition, music recommendation systems may need to adapt to new data quickly, thereby being sensitive to easier-to-mount abuse attacks. This is a very considerable challenge for recommendation systems, and abuse-resistance may significantly impact what technical approaches are feasible.

- **Understandability**
  - A system may be required to explain why it has made certain recommendations to fulfill regulatory, or even contractual/audit, requirements. Even if there are no such requirements, users may still like knowing why a system thought they would like something, and might find that information educational. A system's implementers may also want to know this to help debug these complex systems, which gets harder to do as they combine more and more signals and approaches. When recommendations are based on an ensemble of algorithms operating on data from a huge user population, it may be hard to ferret out the relative contribution of any component. This is especially true if the system uses hard-to-interpret neural networks.
  - With respect to causality, why a user likes a recommendation is a truly complex matter. The ultimate cause may rest in neuroscience and be well beyond the capability of any system and its available data.
  - There is no reason for a music recommendation system to release its underlying data and models to others. There is no scientific result that others need to duplicate, though regulatory frameworks might require such release for other recommendations, e.g., in the realm of investment management. The Netflix data release mentioned under "Dependability" is a cautionary lesson.

- **Clear objectives.** The system's objective is to recommend tracks that a user then plays. However, there are shades of gray that make implementing a recommendation system very tricky indeed:
  - Should it try to diversify a user's listening repertoire and perhaps educate the listener?
  - Should it consider material's royalty costs? Notably, if there were a differential cost of material, as in movies or books, subscription services might penalize recommendations with high licensing costs, while à la carte services might recommend expensive items to make an increased profit.
  - Is a high proportion of accepted recommendations a good enough surrogate for listener satisfaction? Or could listeners – over time – become fatigued and

https://doi.org/10.1017/9781009272230.007 Published online by Cambridge University Press

inattentive to the system, even though they have actively or passively accepted its recommendations?
  - How much should a music recommendation system "throttle" itself? Some say good recommendation systems distract society from more important pursuits by enticing people to instead listen to yet more music. Other recommendation systems have similar concerns, perhaps enticing someone to read yet another novel or watch another cute cat video, or perhaps (more seriously) reinforcing an erroneous "fact."
  - These and many more secondary considerations make music recommendations quite challenging, and Chapter 12 discusses these objective-setting challenges in much more detail.

- **Toleration of failures.** Music recommendations need not be "perfect," and a user may not heed a particular recommendation for many reasons. It probably is not even a goal that every recommendation be accepted, as users may appreciate bold or creative suggestions. However, recommendation systems need to walk the fine line between bold and jarring recommendations.
- **Legal, risk, and ethical issues.** Music recommendation has few legal issues and fewer risks than other domains (although, for example, it is crucial to be careful about recommending obscene lyrics to minors). However, there are many ethical issues relating to the type of recommendations made and their impact on individual listeners, their community, and the creator/artist whose success may be at the mercy of these algorithms. The fourth item under "Clear objectives" could easily be considered also an ethical consideration due to its need to balance rights and harms.

5.4 Analyzing Protein Folding

Protein folding does not provide direct consumer benefits, but rather intermediate results that scientists use to make other discoveries. This makes some Analysis Rubric elements easier to satisfy (e.g., no abuse), but increases others' importance (e.g., reproducibility).

- **Tractable data.** Data-driven approaches to protein folding build on many databases, for example, UniProt, which contains sequence data about millions of proteins, and the Protein Data Bank, which contains a global archive of experimentally determined 3D protein structures (wwPDB consortium, 2019). AlphaFold 2 requires many hundreds of gigabytes of such data as inputs. The extent to which the models can use raw, unprocessed data from these

databases versus needing preprocessing (as in Section 8.2) varies depending on the details of their technical approach.

- **A technical approach.** In contrast to the spelling example, which has very simple models, the recently successful protein folding prediction uses some hard-coded techniques (e.g., for aligning related amino acid sequences) and several connected machine learning models (e.g., transformers (Vaswani et al., 2017)), which have only recently become understood. The technical approach also blends in physical constraints such as the triangle inequality and energy minimization.
- **Dependability.** There are few privacy, security, or abuse-resistance concerns for this application, although organizations may want to maintain confidentiality. Resilience is important, as scientists would prefer the highest-quality results independent of the precise protein being predicted or occasional input data errors.
- **Understandability.** Causality and explanation are not critical, as scientists already understand much of the underlying physics. Also, good structure predictions are more important than the explanation of how they were achieved. On the other hand, reproducibility is important for two reasons. First, others should be able to build on and improve the work. Second, scientists need to compare different protein folding prediction systems to learn their strengths and weaknesses. Even a cursory examination of the release packages of both AlphaFold 2 and RoseTTAFold shows the enormous amount of work data scientists must do so others can reproduce their results.
- **Clear objectives.** Protein folding's objectives are generally clear: accurately predict the correct structure. There is some leeway in how to handle near misses.
- **Toleration of failures.** Errors in protein structure can be tolerated if scientists know their likelihood and can estimate the costs of the extra work they cause. To minimize risk, scientists can often do experimental work to confirm predictions. Reproducibility allows others to find failures before they become more problematic and to suggest fixes.
- **Legal, risk, and ethical issues.** ELSI issues related to protein folding are minimal, though applying that knowledge (e.g., in diagnosing or treating disease) will result in many challenges.

5.5 Analyzing Healthcare Records

As we discussed in Section 4.5, there are immense opportunities to improve human health, but also many complexities. Applying the Analysis Rubric makes this clear.

https://doi.org/10.1017/9781009272230.007 Published online by Cambridge University Press

- **Tractable data.** A vast amount of healthcare record data is kept at medical institutions, testing facilities, insurance companies, and other locations. However, the data is fragmentary, encoded in different ways, and recorded with differing degrees of accuracy. The data must be carefully guarded, due to both data privacy reasons and economic value. As an example of the complexity, a dictionary of different histologic test results alone fills a 388 page document (National Cancer Institute, 2022), and it represents a very small portion of the needed data definitions and standards!
- **A technical approach.** As healthcare records can be used in many different problem domains, technical approaches vary greatly. They center, however, on sophisticated methods informed by critical thinking (Schuemie et al., 2020). There are established best practices, sometimes implemented in standard libraries to reduce the effort in undertaking new research applications.
- **Dependability.** Privacy and security are of foremost concern, due to both ethical needs and legal protections of human healthcare data. As a result, institutions must control their own data carefully and not release it to others without careful safeguards, such as aggregation and anonymization (see more on the latter in Chapter 10 on Dependability). Abuse is not likely, but resilience is very important because errors, even if eventually uncovered by further experimentation, could be very costly.
- **Understandability.** The specific application determines what understanding needs to be provided. If the objective is prediction, as in the HIV-risk example, explanation may be needed. When retrospective, observational studies are used to learn correlations that would catalyze further study in clinical settings, there is a particular need to expose underlying assumptions and perhaps allow reproducibility (see Section 11.3). However, scientific reproducibility is complicated by the complex coding of data and privacy-related limitations on data dissemination. While it is very difficult to show causality from retrospective studies, many will want to use such studies to make healthcare decisions because no better information may be available. For example, by September 2021, the World Health Organization reported over 150 retrospective observational studies on COVID-19 vaccine effectiveness, and, despite limitations, their data was indeed used to inform vaccination policy (Sterne, 2021). (See Section 11.2 for more on causality.)
- **Clear objectives.** These applications usually have clear objectives, although, in the realm of prediction, balancing the likelihood of false positives and false negatives, and setting related thresholds may be difficult. Study design objectives may be open-ended when the goal is creating an interactive analysis platform for gaining insight, though there are statistical risks to this, as discussed in Section 11.4.2.

- **Toleration of failures.** For healthcare records, failure toleration varies. While many observational studies are used to create and hone hypotheses, it is quite important that the hypotheses are of sufficient value to warrant the cost of the ensuing and necessary confirmatory research.
- **Legal, risk, and ethical issues.** Health-related data is significantly regulated, as are study designs involving patient health records. The objectives must take account of compliance with these regulations. There can be great financial and reputational (not to mention safety) risks if data is lost or misused. Ethical issues frequently arise and are best illustrated with questions: Are different elements of society served equitably? If an observational study shows a potential risk to a patient or a population, should that risk be made known even if there is a lack of corroboratory evidence or uncertain potential harm? Should a study, of potentially great value, be undertaken knowing its reproducibility might be in doubt? We present more about these ethical issues throughout Part III.

5.6 Analyzing Predicting COVID-19 Mortality

We will be briefer in this sixth application of the Analysis Rubric, yet still attempt to provide additional color on this important, yet difficult, COVID-19 mortality prediction application.

- **Tractable data.** Data, particularly early in the COVID-19 pandemic, was late in arriving. It was inconsistent across time periods and populations, occasionally erroneous, and very incomplete. COVID-19 modeling certainly did not have location, diagnosis, or health data on individuals or even small subpopulations. Much greater detail (e.g., universal GPS/Bluetooth location tracking and reporting) might have greatly improved modeling capabilities, but would have had unacceptable privacy implications in many societies. Compare and contrast data availability for this application with that for recommendation systems, which have vast amounts of personal, highly detailed click data to make individual recommendations.
- **A technical approach.** There was no shortage of technical approaches, ranging from SEIR models based on disease transmission dynamics to autoregressive machine-learned ones. Some worked well for near-term modeling (e.g., for a period of weeks), but none worked well for longer-term modeling. As suggested above, vastly more data would have made this a very different, and likely more tractable, modeling problem.
- **Dependability.** Privacy and security were not concerns for modelers, but these issues are at the heart of why detailed data was not collected on individuals or

https://doi.org/10.1017/9781009272230.007 Published online by Cambridge University Press

subgroups. Abuse wasn't a problem, mostly because there was no crowd-sourced data. Models were reasonably resilient for near-term projections, but not long-term.

- **Understandability.** Some models were based on disease transmission dynamics. These were explainable and naturally suggested causal relationships. Others were not. To the extent that models would be used to make important public policy decisions, explanation was crucial. The researchers who created the models referenced in the aforementioned Cramer et al. (2021) paper provided for scientific reproducibility.
- **Clear objectives.** The primary objectives were clear. First and foremost, this was to predict the mortality rate from COVID-19, though there was some definitional ambiguity in classifying the cause of death. Secondarily, it would have been excellent if models could have predicted what would happen under different policy assumptions, such as the impact of school openings. While Cramer et al. did not explicitly evaluate model performance for these goals, they are even harder predictions to make, and it is unlikely the models could shed much light on them.
- **Toleration of failures.** Failures are problematic, as they would be expected to have very significant effects on human behavior.
- **Legal, risk, and ethical issues.** There are few, if any, legal issues. The risks of poor forecasts are very real due to their serious impacts on health and welfare.

5.7 The Analysis Rubric in Summary

This chapter illustrated the breadth of considerations for effectively applying data science to a problem.

- **Tractable data** and a **technical approach** are necessary. Implementations must also include a significant focus on **dependability** (**privacy**, **security**, resistance to **abuse**, and **resilience**). These latter issues may be even more complex than what some might consider the "core" data science.
- Data science applications that must provide **understanding** (for **explanation**, determination of **causality**, or release of data and algorithms to enable **reproducibility**) have added implementation complexity. Many data science techniques do not easily support such objectives.
- While in many cases a data science application's **objectives** may seem clear, when considered in depth, they are hard to pin down, especially given possible unintended consequences. Recall the term **objective function** from operations research. Can we develop objectives with the precision connoted by this term?

- Data science techniques frequently work only probabilistically. If an application cannot **tolerate failures**, the challenges may prove insurmountable.
- Finally, data science applications can cause substantial problems, especially if incorrectly specified or implemented. The Analysis Rubric element covering **legal, societal, and ethical issues** must be very carefully weighed.

https://doi.org/10.1017/9781009272230.007 Published online by Cambridge University Press