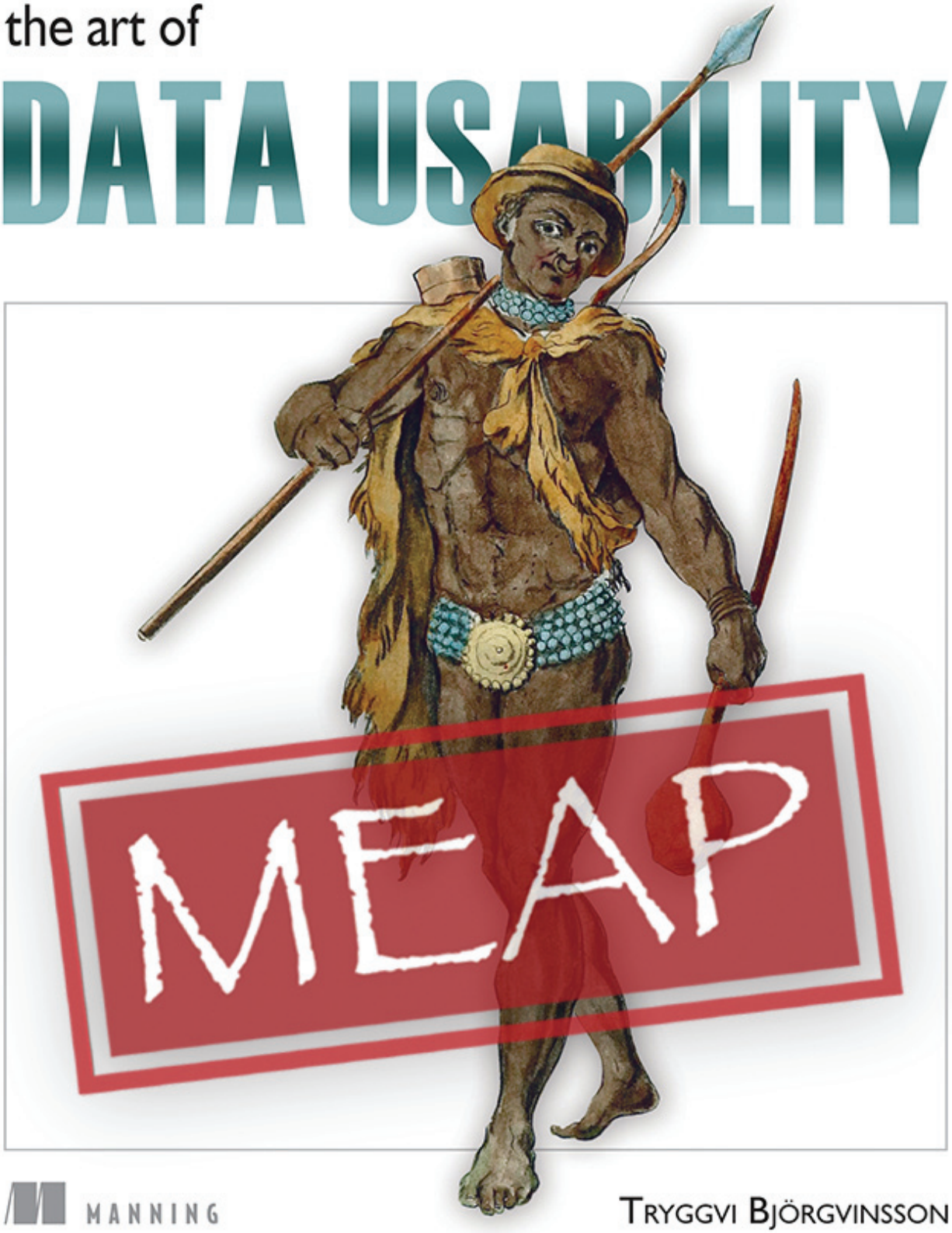
the art of

DATA USABILITY



MANNING

TRYGGVI BJÖRGVINSSON

MEAP Edition
Manning Early Access Program
The Art of Data Usability
Version 6

For more information on this and other Manning titles go to
www.manning.com

welcome

Thank you for purchasing the MEAP of *The Art of Data Usability*. I'm very excited about this book and I hope it will help you hone your data skills from the first page.

I think it's safe to say that data science has exploded in popularity the last few years. That's actually not surprising because the power that data science wields enables us make better decisions; decisions based on facts and observations. It helps us analyze and innovate faster, and understand the „smørgåsbord“ of data we have access to.

Even with all this data being generated around us and the clear willingness to make the most of it, too little thought is put into making the data useful. We're kind of throwing it all against the wall to see what sticks; hoping that users of the data can, with some amount of magical data fairy dust, just make it work. In the end, we're all working with data to help someone and to maximize the opportunities of our data, we need to focus on data usability.

You'll quickly see that data usability is all about quality. The book is founded in tried and true quality management methods that shape the way we work, and how we can iteratively make data more useful but there's one benefit that data has over most other subjects of quality management: data is most often digital so we can automate a lot of the work.

The book goes through the different phases of a data project life cycle: project design, data management, collection, processing, dissemination, and closing the project. You'll see how data usability can be applied to each of these phases, through example attributes, discussions about those attributes, and code examples, written in Python, to show examples of quality controls in practice. You don't need really need any prior experience to get through the book and understand its concepts, but beginner knowledge of programming helps a lot, so does experience of data science and „data munging“.

You now have the third edition of the MEAP in your hands. This version includes changes to the previous chapters to better focus the book and explain how I envision that the book may help you become a data usability expert. It’s not only revisions though, it’s also two new chapters – chapters 5 and 6 – the first two chapters that follow the same structure as the rest of the book’s chapters (the future chapters).

These two chapters form the third part of the book, about managing the data and tackle two important usability attributes, availability and recoverability. Without revealing too much, the quality level of both of these attributes are increased with backups and ironically, while preparing this third edition of the MEAP my own computer hard disk failed. Luckily, I had cold backups so recoverability was in good shape, but it took me a

long time to get everything up and running and availability suffered. Talk about living your book!

In previous editions of the MEAP I claimed that one of the keys to data usability is to listen to your users. This is at its core what you're always trying to achieve and do when working towards usability and quality. It's what the book's all about! Your feedback is key to making that happen. If you have any comments, questions, or suggestions about the book, please be sure to post them to the Author Online forum. I'll be reading and responding to your posts, and incorporating it into the book to make it as useful as possible.

Thank you again for purchasing the MEAP!

— Tryggvi Björgvinsson

brief contents

1 Learning from the past

This chapter covers:

- How we improve data quality
- The Data-Information-Knowledge-Wisdom hierarchy
- What to expect from the book

Should it be possible to convict you as a criminal because you don't have enough knowledge? Are you a criminal for making a wrong prediction? Those questions arose in a regional court case in Italy, where science itself was put on trial following a big earthquake that destroyed the city of L'Aquila and killed 309 people in April 2009. The tragedy of those deaths was made even worse in 2012 when six scientists and a government official, all members of the National Commission for the Forecast and Prevention of Major Risks, were accused of giving falsely reassuring statements prior to the quake.

It took the judge only four hours to reach a verdict. All seven individuals were found *guilty of multiple manslaughter* for inadequately warning residents by giving inaccurate, incomplete, and contradictory information. The convicted individuals were left confused. They didn't really know what they had been convicted for but they knew they had just been sentenced to six years in prison and barred from ever holding public office again.

Scientists around the world were appalled and afraid this would give a bad precedence. Never again would knowledge relating to uncertainty be shared with the public. Scientific progress would grind to a halt in more fields than just seismology. Thankfully, the conviction was overturned two years later. All of the convicted were

cleared in an appeal that was confirmed by the Italian Supreme Court in November 2015.

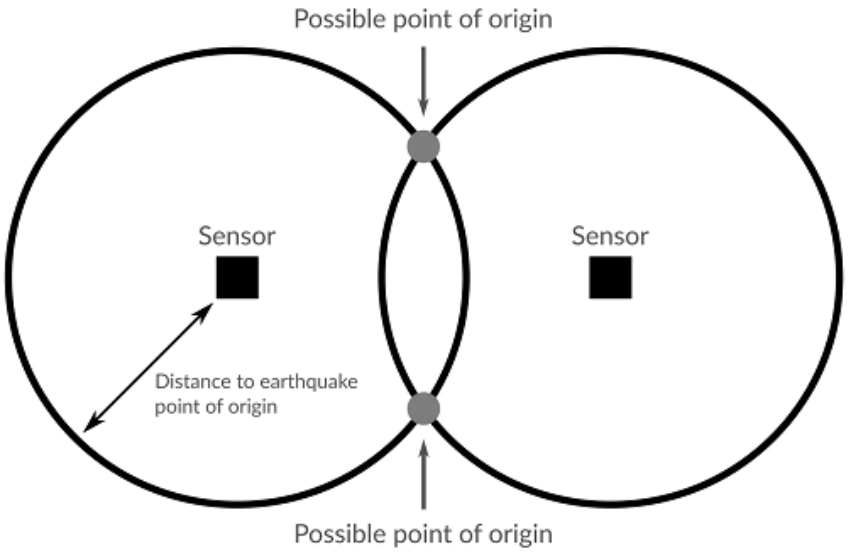
Being able to make correct predictions is a result of a seemingly endless amount of small insights and improvements, made over the course of many years, each step putting us in a better position to make better predictions. But, all of these incremental improvements only allow us to make better predictions, not speak in certainties.

1.1 Basing improvements on past insights

Like all of us, scientists can be wrong or make mistakes. There's another interesting example of scientists being inaccurate from January 1973. What makes this example different is how the mistake was used as a basis for improvements instead of as grounds for a court case. This event took place in Iceland and much like with L'Aquila, it began with earthquake sensors detecting increased seismic activity. Earthquake sensors are expensive and they were even more expensive back in 1973. You couldn't put as many of them as you wanted wherever you wanted. This created a problem back in those days because the amount of sensors matters and the cost limits how many sensors are put into use. It turned out, in the 1973 event, that this limitation actually risked more than 5000 human lives.

It all started when increased seismic activity was detected by two earthquake sensors in south Iceland. The way these sensors work is that they record earthquake waves that travel through the Earth's crust from the point of origin to the sensor. By analyzing the time it takes to reach different earthquake sensors, you can predict the point of origin (figure 1.1). The analysis of the data from the two earthquake sensors in 1973 indicated two possible points of origin: an active earthquake area called Veiðivötn, or a volcano in an archipelago off the south coast of Iceland called Vestmannaeyjar.

Figure 1.1. Two possible points of origin with two sensors



Veiðivötn are a more active seismic area than Vestmannaeyjar, so the scientists predicted that Veiðivötn was the point of origin. Their focus was therefore on monitoring the Veiðivötn area closely, but as you can probably guess, they were monitoring the wrong area. The volcano on the inhabited island in the Vestmannaeyjar archipelago erupted.

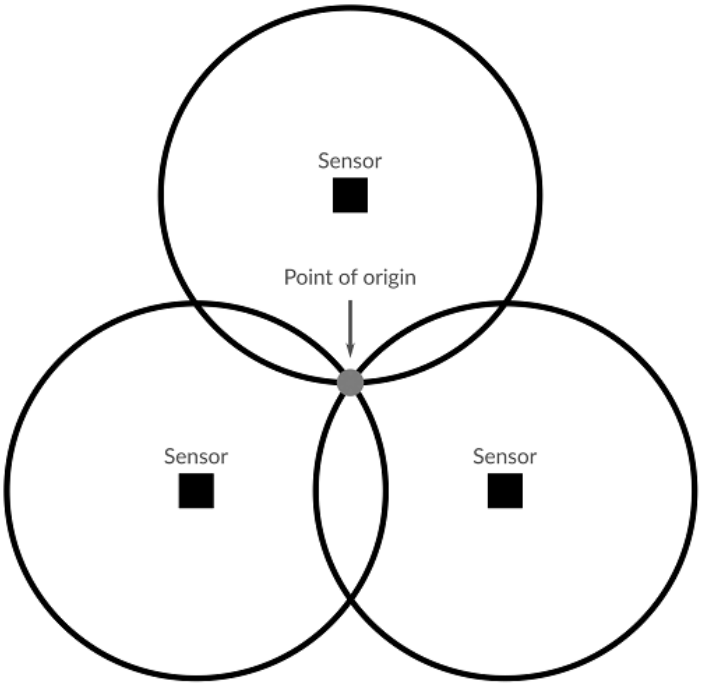
NOTE **Historical rescue efforts**

The Vestmannaeyjar eruption threatened the more than 5200 inhabitants who lived on the small 5.2 square mile island where the volcano erupted. Even if the majority of inhabitants on the island had gone to sleep when it started erupting around midnight, they were all able to leave their homes and reach safety ashore Iceland.

The rescue efforts were very successful and the efforts also mark the first time in history that lava flow was changed with human intervention. To save the prosperous harbor of Vestmannaeyjar, sea water was hosed on the flowing lava and the lava current was redirected away from the harbor.

In contrast to the L'Aquila trial, the scientists weren't put on trial. Instead they analyzed their predictions, and realized that they could improve their predictions. The problem they had wasn't that the data was wrong, it was incomplete. They needed more earthquake sensors (figure 1.2) to improve the quality of their predictions.

Figure 1.2. An extra sensor identifies the correct point of origin



Consider these two different approaches to getting better. The approach with the L'Aquila trial was to punish for lack of quality, an encouragement equivalent to when Homer Simpson told his kids, *"You tried your best and you failed miserably. The lesson is, never try."*

To improve predictions, the approach in the Vestmannaeyjar eruption is more likely to be fruitful; analyzing the mistake is a much better approach to improve predictions. You look at what you've done and what data you have. From that, you try to figure out how you can improve what you do in the future (in this case, by adding more sensors). This approach increases the quality by encouraging improvements instead of pushing people away from them.

Why am I telling you this? What does this have to do with being *usable*? Well, it's because the key ingredient of *data usability* is *quality*. The secret behind the art of *data usability* is to improve *data quality*; I'd even go so far as saying that these two terms can be used interchangeably. The difference to me is just that the term *data usability* reminds us of what we're trying to achieve (useful data) and *data quality* tells us how we're going to achieve it (by improving quality). That's why I'll devote the next chapter to *data quality*, the underpinning of *data usability*.

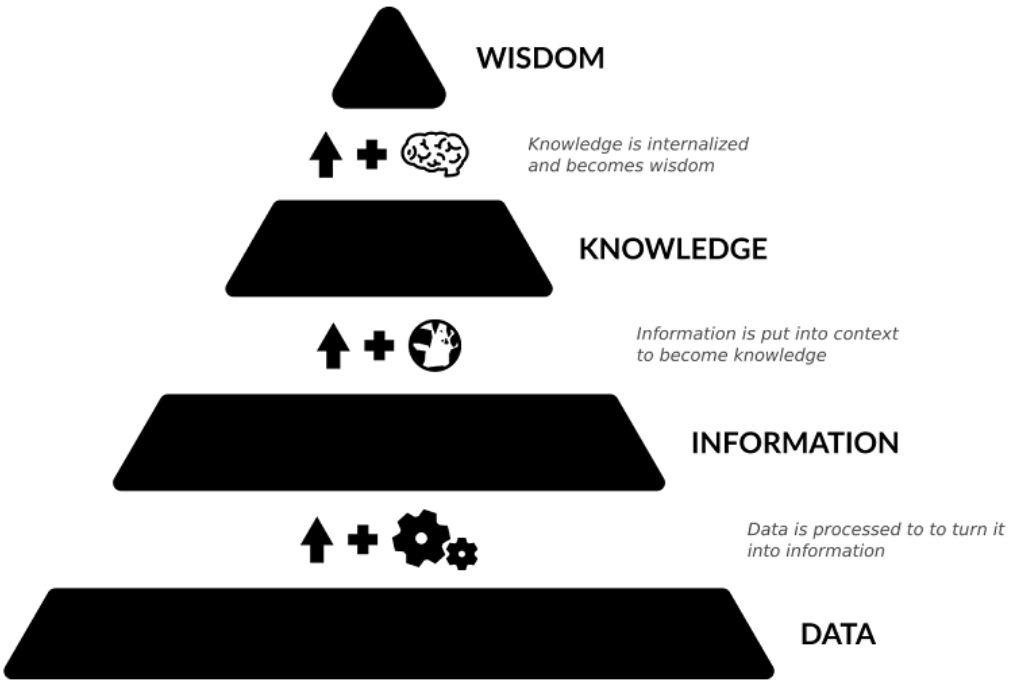
1.2 Understanding the DIKW hierarchy

Data, information, knowledge, and wisdom are the four levels/models of what is

known as the DIKW hierarchy (the name comes from the initials of each level). This hierarchy is most often depicted as a pyramid (figure 1.1) because each level is a foundation for the next; you need data to produce information, you need information to generate knowledge, and you need knowledge to gain wisdom.

Thinking of the model as a pyramid or a flow helps visualize the relationship between these four models. Data comprises records of facts or observations. Processing the data changes it into information. Analyzing the information and putting it into context generates knowledge. By attaining and really understanding the knowledge, it can then be applied to situations in which it becomes wisdom.

Figure 1.3. The DIKW pyramid shows how each model is a foundation for the next model



When we think of data usability, we need to think about it on all those levels. Most often we skip the wisdom level because that's based on human judgment and our work is to improve everything leading up to that application of judgment.

1.2.1 Finding the path to wisdom

Even if the entire next chapter is about data quality, let's get one misunderstanding about data quality out of the way right away because the term can confuse you (another reason to use *data usability*instead). For many people the term data quality means clean and correct data; you could call it the reward from wrestling bad data. As you will see in this book, that's not the whole story. It's actually far from it. You have to think about a lot more than just fixing bad data. The scientists monitoring the seismic

activity in January 1973 may have gotten 100% correct data but their predictions were still wrong. They needed that third sensor to improve their predictions. The data lacked quality because they needed more data sources.

The thing is, even if we're working with data we're not aiming only for data quality, we're aiming for more. In addition to ensuring the quality of the data itself, we also need to ensure the quality of transformed data and the interpretations of data. Ultimately, we're trying to generate good knowledge. If the knowledge we have is of good quality, we'll be able to make better decisions or you could say we become wiser. That's what we're trying to achieve, so how do we get from data to wisdom?

Great wisdom is built on facts in the form of data, but you don't get wise by just looking up facts. There's an intricate journey that transforms facts a couple of times before we can call them wisdom. The results of each of those transformations are known by different names and each result is the foundation for the next transformation. Let's go through each stage of this journey from facts to wisdom.

1.2.2 Data

If great wisdom is built on facts, then the path to more and better wisdom must start with the collection and recording of facts, or as we'll most often refer to it, *data*. Data, generally speaking, is nothing more than a recording of facts or observations.

Each fact or observation is known by many names. It is referred to by some as a datum, by others as a data point, sometimes as a record; in statistics, it's the variable. What term is used can depend on the situation or just personal preference.

TIP **Data should actually be a plural word because data is a collection, but it is also correct to use data as a singular word because it can be treated as a mass noun, just like information. The plural data is mostly tied to scientific community, whereas singular data is used more in everyday conversations. When I think of data, I always think of it as a dataset, not data in the theoretical sense, so I'll be using it as a mass noun, in the singular form.**

Data isn't a natural phenomenon. Data is man-made. We collect data from either a single source or from multiple sources, or we create something that generates data for us.

There can be multiple reasons for collecting data. In most cases, we systematically collect data to answer a question. It might be an implicit or explicit question. If you're collecting and recording facts to answer a question, it's called *captured* data. If you're generating data as a byproduct of a functionality, like making a web server log the IP address, time, path, browser, and other data points for each request, it's called *exhaust* data. Exhaust data is usually generated for future questions that might arise.

Regardless of how or why we record data, the raw collection of facts is rarely useful to us as it is. We have to process the data to produce a usable form of data, commonly referred to as *information*.

1.2.3 Information

Let's say you're going on vacation and because you're fascinated with small guitar-like instruments you're interested in either visiting Hawaii (to learn more about the ukulele) or Portugal (to learn about the ukulele's predecessor the machete). Because you can't make up your mind about where to go, you decide to go to the place with the highest recorded temperature (just because you need some sort of a tie-breaker). You find the highest recorded temperature for both and write those facts down into a very small data collection, shown in table 1.1.

Table 1.1. Data: highest recorded temperatures

Place	Max temperature	Record date
Hawaii	100	Apr. 27, 1931
Portugal	47.4	1.8.2003

By the looks of it, the highest recorded temperature in Hawaii is about twice as warm as the record in Portugal. But wait! Don't book your flight yet. You haven't processed the data. You've only recorded the facts. Take a closer look at the records. Don't you think those dates look a little off? Why are they different? That's because the two countries have different conventions. You're going to have to convert the dates into the same format.

But there's more. Portugal records temperature in degrees Celsius but the USA uses degrees Fahrenheit. You're going to have to convert those as well. For that, you have to decide what convention you'll use.

You decide to go with Celsius because you like SI units (even if 100 is a nice round number), and you decide to go with the ISO 8601 date format because it's better to order dates in that format (if you ever extend your data). Then you sit down to do your magic (and end up with something like table 1.2).

Table 1.2. Information: highest recorded temperatures

Place	Max temperature (°C)	Record date
Hawaii	37.8	1931-04-27
Portugal	47.4	2003-08-01

Now you see that there isn't that much of a difference between the warmest temperature of Hawaii and Portugal, just a little less than 10 degrees Celsius. You understand the data now. You've turned your data into information and it looks like you'll be going to Portugal.

Information is processed data. It's similar to data, and many people use the two terms interchangeably. They're not wrong. Information is still data. The difference lies in whether you've performed steps in the data collection such as validating the data, aggregating data points into a more understandable set of points, converting everything to the same units or transforming the raw source data into something more useful.

Often, you convert data into information automatically.

So, the next time you happen to participate in a quiz, and the quiz master asks what the record-highest temperature in Hawaii is, you'll be able to answer correctly but you might have to go to a lot of quizzes to ever be asked that question. That's the problem with a piece of information: by itself, a single piece of information isn't much; it's just an understandable data point. Unless your life goal is to become a living encyclopedia, you need something else. You need to turn the information into knowledge by putting it in context with how you understand the world.

1.2.4 Knowledge

Knowledge is the sum of information and the capability to understand the world around oneself. You don't pick your vacation spot based on the highest recorded temperature, especially not if one of those temperatures were recorded in the 1930s. You know that you should use more recent data, for example the average temperature over the past few years. That information is more relevant to you when you pick the vacation spot, but watch out, turning information into knowledge can be very dangerous.

You may not think about tabletop games like Monopoly, Sorry, or Snakes & Ladders as a particularly dangerous hobby. After all, you just sit around a table with your friends and throw dice. You wouldn't expect that to be related to worshiping Satan or planning a prison riot but that was the reality of role-playing games back in the 1980s and 1990s. Role-playing games are a special form of tabletop games.

At its core, a role-playing game is about a small group of people who collectively create a story together, and instead of writing the story down, they experience it together. The usual way to play such games is for a small group of friends to come together around a table, with papers, pencils, dice, and a lot of imagination. Players create characters for themselves (which they describe on the paper), and those characters are the heroes of the story who come together as a group to face an adventure that is bigger than each of them. The players role-play those heroes when they face terrifying things such as battles with monsters, dangerous cliffs, or social interactions.

The reaction to this form of play was unusual. Parents who saw their kids play such games believed the kids might start to worship the devil. A prison banned role-playing games because it was believed that a role-playing group was forming a gang. Perhaps the weirdest response came from the Secret Service when they raided a prominent producer of role-playing games and themes for stories, Steve Jackson Games, in an attempt to rid the world of computer criminals.

Leading up to the raid and after the raid, the Secret Service collected lots of data and information about a new upcoming world threat, computer criminals. Among the incriminating data for Steve Jackson Games was:

- One of the computer criminal they monitored, *Urvile*, liked to play role-playing games from Steve Jackson Games
- After apprehending Urvile, lots of detailed written scenarios around Steve Jackson

Games, including an attempt at breaking encryption on a super-computer, were discovered close to records of actual computer intrusions

- Urvile used two books from Steve Jackson Games, one about high tech and the other about special ops
- One of Steve Jackson Games employees was a computer criminal
- Steve Jackson Games, ran a bulletin board (an early Internet social network) frequented by computer criminals
- An upcoming (role-playing) book from Steve Jackson Games, Cyberpunk, looked like a manual for computer crime

Clearly the small company was an elaborately hidden school for computer criminals. The Secret Service collected a lot of data against Steve Jackson Games, but they completely missed the context. Computer criminals have hobbies. Those hobbies may be related to computers, like with the case of Urvile who created role-playing adventures for fellow players where they had to deal with special ops in high tech situations. Steve Jackson Games created many different template books for adventures, including books on magic and space in which they tried to set out general guidelines and rules for such role-playing adventures. That doesn't mean that Steve Jackson Games was the real-world Hogwarts nor did they have aspirations of becoming our future Starfleet Academy. It's a company that creates games for people to have a fun time together.

At the time, computers were getting more popular so role-playing adventures around computers seemed like a logical next step. Unfortunately, the person hired to write those adventures knew a lot about computers and intrusions (based on personal experience), but all that knowledge was probably why Steve Jackson Games hired that person, not because they were looking for a teacher for their *student criminals*.

This huge misunderstanding is an interesting example of how data was turned into information and then interpreted based on a wrong world-view or misunderstanding. The information was turned into the wrong knowledge, which explains the weird decision to raid a company.

Data is processed and becomes information. That information can be put into context to generate knowledge, and the Secret Service agents failed to put their information into the appropriate context. This is what we're trying to improve when we talk about data quality; We're trying to improve the quality of not only data, but also the information and the knowledge.

If we do that, we'll have a solid foundation for our wisdom, which is what we get when we turn our knowledge into wisdom by applying judgment.

1.2.5 Wisdom

To gain wisdom is human. It requires ethics and a sense of right or wrong to pass judgment. *Wisdom* is being able to do the right thing based on knowledge, apparently almost without thinking.

Imagine you're a mobile app designer. A client comes to you with an idea for an application called *Make it happyn*. The gist is that users of the app can go through images of things that may or may not make them happy. For each image, the user can press a button to indicate that it will make the user happy or press a different button if not. Friends of this user can then tap into the data generated to find a gift that would make this user happy and avoid gifts that would make the user unhappy.

Your client wants this app to be really simple for the user. The interface should only show the image and two buttons, red and green. No text, just beautiful simplicity.

With all the design *knowledge* you have acquired, you know that a rather large portion of males (and a few females) suffer from red-green color blindness. They will be unable to see the difference between red and green. You also know of cultural problems with the color red, where in many countries it is associated with something bad; in others such as China, it is associated with happiness or joy.

You can apply your knowledge to the client's request and make the judgment that this is not the best design, because it excludes too many people. Instead you suggest the use of icons to symbolize happiness or bad feelings. It turns out your client is just as happy to have the users of the app press 😀 or ☹. This application of your knowledge transforms it into wisdom. You knew that the original design request wasn't the right thing to do.

That leaves us with the task of improving quality of data, information, and knowledge.

Don't worry, I'll show you how to accomplish it.

1.3 What to expect from this book

It's difficult to teach data usability in a book because data projects vary in tools and techniques, the subject of the data, the size of the project or the team, geographical distribution, and legal environment. The fact is, data is everywhere; it's all around us, and we take different approaches to try to understand and use it.

This book is not about **using** data.

- It's not going to teach you how to clean and process data, even though there are chapters on data collection and processing.
- It's not going to teach you how to use R, Pandas, nor anything else to extract information and knowledge out of a big dataset.

This book is about **making data usable**.

- It's about all the other things you need to think about if you want other people to get the most out of your data.
- It's about what those users need or want. You may be thinking that they just want to get the data but you have to think about how they want it.
  - Do they want the data in a standardized format?
  - Does size of the dataset matter?

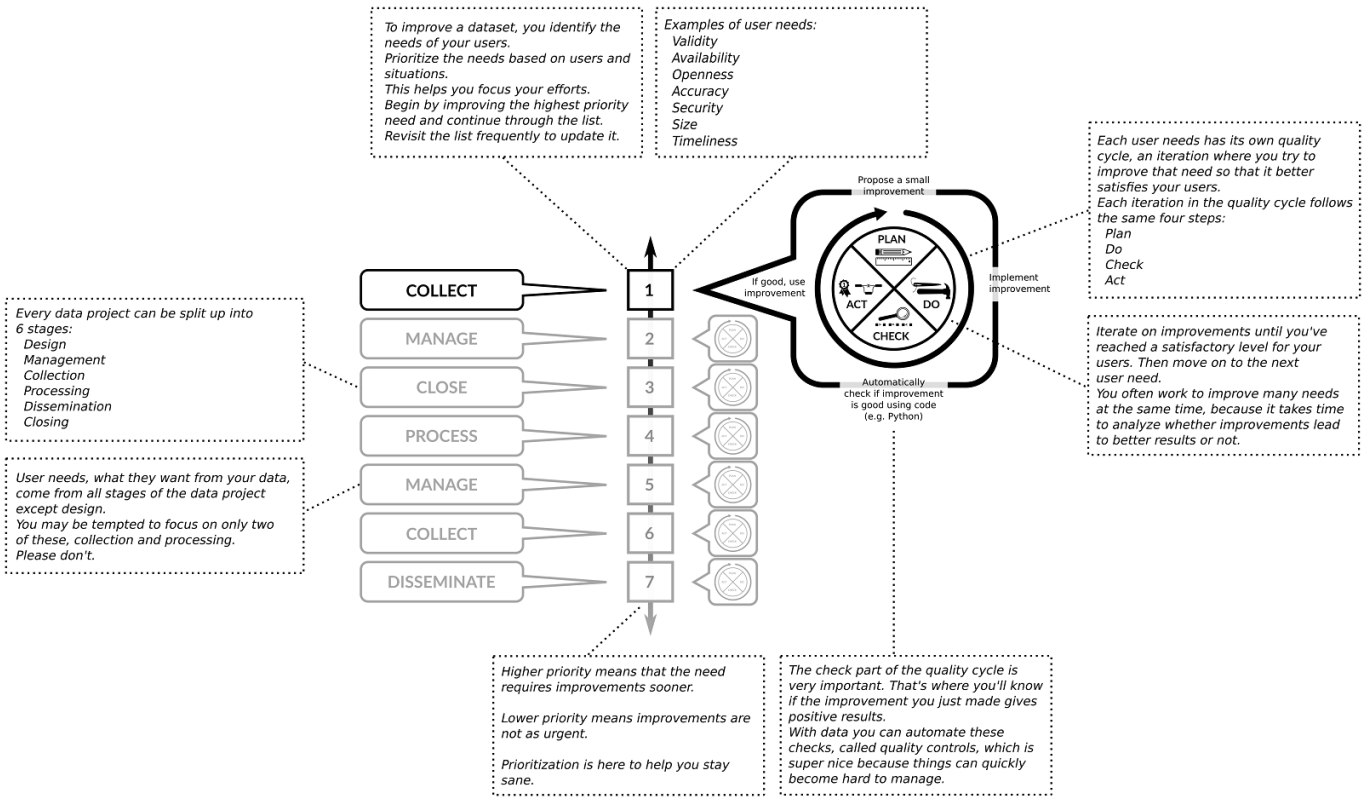
- Is it even important that the data is correct or are your users only looking for rough numbers?
- This book is about how you **know** that your data is what users want and what you can do to meet their needs.

1.3.1 Overview of the usability work

Figure 1.3 shows an overview of how to work with usability; How you identify and prioritize needs from users across all stages of a data project, how you set up a quality cycle for each need and what you do in each step of the quality cycle.

It's a busy figure but that's why you have this book. It will familiarize you with the foundations of usability and quality, and provide tips and tricks for you in the process. In each of the chapters in the book I'll show you this figure and highlight what part of it we'll be working with in that chapter.

Figure 1.4. Overview of how to work with usability



I assume you have some experience with working with data but you're more likely not to have a lot of experience with quality. Quality is the foundation of data usability, but there are many definitions for the term *quality*. My understanding, gained through

experience, may not be the same as others you might consult. Therefore, it's important that we're on the same page throughout the book (pun intended), so we'll discuss this important foundation in the next chapter. It introduces the notion of quality and the main process you will use throughout all the different usability improvements in your data projects.

Alright! Onward to fun!

1.4 Summary

In this chapter, you've learned the following:

- Data usability is all about improving data quality, in fact, these terms can be used interchangeably.
- Data is a collection of facts or observations and it's made by humans, directly or indirectly.
- We collect or generate data points to answer questions we have or may have in the future. To improve the future, we must analyze the past.
- Data must be processed to become usable. Processed data is known as information.
- The act of taking the information and connecting it with other relevant information and situations turns it into knowledge. Knowledge is information in context.
- By understanding and internalizing the knowledge, we can apply it to new situations, almost without thinking. This internalized knowledge is referred to as wisdom.
- Data usability is about more than just clean and correct data, there are other attributes, and other forms of data (information and knowledge) we must consider as well. Data usability is relevant in all stages of a data project.