CHAPTER

5

Data Preparation

The fuel for AI

In late 2015, IBM announced the acquisition of the Weather Company—which included the weather data, forecast information, website, app, and various intellectual property. The price tag was over an estimated $2 billion.[1]

What was the rationale for this? Was IBM getting into the weather business?

The deal was actually great for the company. Data is the fuel for AI, and weather data has broad applications. IBM folded the Weather Company assets into the Watson AI platform. The forecasting segment, called WSI, was likely the most important asset. The business included license revenue from over 5,000 companies in industries like airlines and utilities.

In terms of the data assets, they included three billion weather forecast reference points. There was also infrastructure for data collection from over 40 million smartphones and 50,000 airplane flights per day.

IBM was not the only mega tech suitor that was interested. Google had already made a bid for the Weather Company but it was rebuffed. It's also important to note that IBM had a partnership with the company.

[1] www.wsj.com/articles/ibm-to-buy-weather-co-s-digital-data-assets-1446039939

T. Taulli, *Implementing AI Systems*, https://doi.org/10.1007/978-1-4842-6385-3_5

Here's what John Kelly, the Senior Vice President at IBM Solutions Portfolio and Research, said about the deal: "The Weather Company's extremely high-volume data platform, coupled with IBM's global cloud and the advanced cognitive computing capabilities of Watson, will be unsurpassed in the Internet of Things, providing our clients significant competitive advantage as they link their business and sensor data with weather and other pertinent information in real time. This powerful cloud platform will position IBM to arm entire industries with deep multimodal insights that will help enterprises gain clarity and take action from the oceans of data being generated around them."[2]

Here are some of the ways weather data and IoT can be combined for powerful results:

- Models can use data from social media sentiment and transformation flows to help retailers and distributors improve their supply chains.
- Weather data can help airlines save millions of dollars with lower fuel consumption, reduced delays, and less airport congestion.

But the acquisition of the Weather Company was not the only deal. IBM had already struck arrangements with companies like Twitter, Apple, Medtronic, and Johnson & Johnson for data sources. There was also the $700 million purchase of Merge Health, which owned a large dataset.

These moves from IBM highlight the strategic importance of data for AI. It's also a key reason why data-rich companies like Netflix, Google, Facebook, and Microsoft are leaders in the industry.

In this chapter, we will take a look at data and how to use it effectively for an AI project.

The Data Explosion

The growth of data is staggering. Based on research from IDC, more than 59 zettabytes of data will be created in 2020 and the growth will be 26% per year until 2024 (a zettabyte is 1,000,000,000,000,000,000,000 bytes).[3] To put things into perspective, the next three years will have about the same amount of data as the past 30! IDC actually increased its forecast because of the impact of the COVID-19 virus as there was an acceleration of the usage of

[2] https://business.weather.com/news/ibm-plans-to-acquire-the-weather-companys-product-and-technology-businesses-extends-power-of-watson-to-the-internet-of-things

[3] www.datanami.com/2020/05/19/global-datasphere-to-hit-59-zettabytes-in-2020-alone-idc-projects/

work-from-home technologies, which rely heavily on data-intensive video streaming. But another factor is the significant increases in metadata and sensor data, such as for the Internet of Things.

When boiling things down, there are four main types of data. They include the following:

- *Structured data*: This is the data that is stored in relational databases and spreadsheets. Because there is usually labeling, this type of data is fairly easy to work with. There are a myriad of BI (business intelligence) tools that can glean insights or create visualizations from this information. And yes, when working with an AI model, structured data certainly makes the process much easier. But the reality is that there is usually not enough of this data. The rule of thumb is that about 20% of datasets are structured.
- *Unstructured data*: This is the majority of the information available for AI models. Examples of unstructured data include images, emails, videos, text files, satellite images, and social media messages. In terms of storage, there are next-generation systems like NoSQL databases that can better handle this type of information. But when it comes to AI models, one of the biggest challenges is finding ways to prepare and analyze unstructured data.
- *Semi-structured data*: As the name implies, this is a blend of structured and unstructured data. In a dataset, the majority of the items are usually unstructured. An example of semi-structured data is XML (Extensible Markup Language), which maps information on a document using techniques like JSON (JavaScript Object Notation). This helps to facilitate the exchange of information using APIs (application programming interfaces).
- *Time-series data*: This can be either structured or unstructured. But for the most part, time-series data involves interactions, like the "customer journey" on the Web and mobile devices. This type of data has become much more important for AI models, especially with the growth of the IoT.

With the growth in data, there has emerged the megatrend of big data. Gartner analyst Doug Laney coined this term in 2001 and it caught the attention of the corporate world. There were even startups that began to call

themselves big data providers and many of them were able to raise substantial amounts of venture capital.

According to Laney, big data has three key attributes: volume, variety, and velocity. They are known as the three Vs.

Let's take a look at each:

- *Volume*: This is about the enormous scale of the data, which is usually unstructured. There is no bright-line definition for the amount, but the volume probably is in the tens of terabytes. In the early days of big data, the handling of large amounts of data was a major challenge. Yet with the innovations in cloud computing, the process has become much easier.
- *Variety*: This is the diversity of the data, which is a blend of structured, semi-structured, and unstructured data. There is a diversity of sources and uses, say from IoT, social media, and so on.
- *Velocity*: This is the speed at which the data is generated. The sources are often the mega tech platforms like YouTube, Facebook, and Salesforce.com. They create enormous amounts of data (often this is user-generated). To allow for this, these companies invest significant amounts in building data centers across the world (this is often a major topic on earnings calls for Wall Street). When it comes to the three Vs, velocity is usually the most challenging.

After Laney came up with his framework, more Vs have been created. This is an indication of the growth and complexity of big data. Here are some of the other Vs:

- *Visualization*: This involves the creation of graphs and charts to better understand data.
- *Value*: This is about the usefulness or effectiveness of the data. Generally, this means that the data needs to come from a trusted source.
- *Variability*: This describes the inevitable change of data over time. This is particularly the case with channels like social media.
- *Veracity*: This is the accuracy of the data. Data sets usually have lots of problems and there needs to be much wrangling.

The bottom line: Big data is a big topic! And yes, it is constantly evolving. But this category is essential for having a strong AI project.

The Database Market

In the early 1960s, corporations started to look at mainframe computers as an effective way to manage operations like payroll and financials. But the computer languages were too complicated for non-technical people and did not handle the kinds of functions that businesses needed, such as with currencies and reporting. This is why COBOL (Common Business Oriented Language) was created. The language had English-like commands as well as the ability to ingest large amounts of data. The use of the COBOL language would even lead to the development of databases, which helped to power platforms like SABRE, which allowed American Airlines to manage reservations.

Yet the database technology remained quite rudimentary for quite a while. The system was based usually on batch processing since access to the data was from tape storage.

But in 1970, an innovation emerged that would transform the database market. This came from an IBM computer scientist, Edger Codd, who published a paper called "A Relational Model of Data for Large Shared Data Banks." In it, he set forth the fundamentals of the relational database. At the heart of this was an easy scripting language called SQL or Structured Query Language. It included English-like commands to do CRUD (create, read, update, and delete) operations. This was possible because the relational database was organized into various tables that had connections to each other with primary and foreign keys.

The main relationships include

- *One-to-one*: The is where a row in a table is linked to one row in another table. An example of this is a driver's license, which is a unique number that can allow for a unique reference to a customer.
- *One-to-many*: This is a relationship in which one row in a table is linked to two or more other tables. For example, a table for customers could link to various tables that have purchase orders.
- *Many-to-many*: This is a fairly complex relationship, where more than one table is linked to more than two other tables. This is the case when multiple reports have different authors.

With these types of relationships, it was possible to create highly sophisticated database systems. But interestingly enough, Codd's innovation did not get much attention. IBM did not even think it had much potential! The company believed that relational databases were not scalable for mission-critical corporate environments. Instead, IBM continued to focus on its proprietary database technologies—and this would prove to be a huge mistake.

Instead, it was entrepreneur Larry Ellison who saw the advantages of relational databases and he formed a company called Oracle in 1977 to capitalize on the opportunity. From the start, growth was strong. The emergence of the PC was a major catalyst. Computers were no longer just for large enterprises. Even small businesses saw the benefits of this technology.

The database industry would see a myriad of startups pop up, such as Sybase, Ashton-Tate, Paradox, and FoxPro. But by the end of the 1980s, Oracle would be the dominate player in the industry and this continues today. The company's market value is over $170 billion.

Despite this, the relational database is far from perfect. First of all, there is the problem of data sprawl, which is where many different databases proliferate across an organization. Because of this, it can be difficult to centralize the data, say for AI projects.

Next, relational databases were not built for many kinds of modern-day scenarios. That is, they tend to underperform when handling high-velocity data, unstructured data, and cloud platforms. It's also challenging to develop applications on relational databases because the systems can be quite rigid and inflexible.

Something else: Relational databases are not cheap. Not only are the licenses hefty, but there are the ongoing costs for maintenance and upgrades. In a world of big data, a relational database is often uneconomical.

Now this is not to imply that relational databases are dinosaurs. The technology will remain a major part of the IT world. But new approaches have emerged, such as NoSQL and cloud-based databases, which we will cover next.

Next-Generation Databases

Until the late 1990s, much of the development for databases came from large software vendors. They had the resources, engineers, and customer bases to be successful. When it comes to databases, customers want a vendor with a strong foundation.

But there was a problem: the database software vendors were not innovating enough. As a result, programmers started to develop their own open source projects. With the rapid growth of the Internet, it was becoming possible to

get quick and wide distribution for software. This would be a game changer for the database market.

A big part of the development was for data warehouses, which can store huge amounts of information. This technology got much attention because of the rise of large Internet platforms.

One of the pioneers of open source data warehouses is Doug Cutting. First he created Lucene, which was a sophisticated text searching application. But as usage increased significantly, he realized he needed a better database infrastructure. This led him to develop Hadoop with Mike Cafarella. Cutting based Hadoop on a paper he read from Google that set forth the framework for a massive file system. He expanded on this greatly and developed MapReduce, which made it possible to process huge amounts of data across various services. The system would then merge these to allow for the creation of reports.

As should be no surprise, the early adopters of Hadoop were companies like Yahoo!, Facebook, and Twitter. They had an urgent need for better ways to handle the firehouse of data on their platforms. And for the most part, Hadoop was an effective solution. The technology made it easier for these companies to run analytics across their datasets so as to improve user engagement.

Yet Hadoop had its limitations. Venture-backed startups like Hortonworks would see this as an opportunity to build systems such as YARN on the Hadoop platform. This allowed for enterprise-level features like online data, interactive SQL, and in-memory processing. The upshot was that Hadoop saw rapid adoption across the corporate world.

Of course other open source data warehouse projects emerged. One of the most popular turned out to be Apache Spark, which has become especially useful for AI. This system deals with the limitations of MapReduce by allowing lower latency with the distribution of data.

The founder of the Spark project is Matei Zaharia, who created the technology while at UC Berkeley in 2009. He would go on to launch Databricks, which has raised a whopping $897 million. The company operates a cloud-based platform to work with the Spark system, such as by using Python-like notebooks.

Besides rapid innovation in the data warehouse market, there has also been much change with the structures for databases. This is highlighted with the development of NoSQL systems.

Essentially, this means that the database is non-relational. The data is stored as a free-form document. Some of the key benefits include more flexibility to create sophisticated AI models at massive scale but also to better handle unstructured and semi-structured data. What's more, NoSQL databases generally have lower costs.

By far, the most popular version of this technology is MongoDB. The company was originally a traditional software venture but transitioned to the open source model in 2009. It was certainly the right move. At the time, there was significant demand for innovative databases, such as for smartphone apps and cloud computing.

In late 2017, MongoDB went public in a highly successful IPO, raising $192 million. The valuation was at $1.1 billion. However, within a few years, the market capitalization would hit a hefty $11.4 billion. Since inception, MongoDB has logged more than 110 million downloads and there are over 18,400 customers across more than 100 countries.[4]

Mega tech operators like Amazon.com, Google, Alibaba, and Microsoft have also moved aggressively into the database market, especially with cloud-based offerings. Part of this has been for their internal AI efforts. But these companies also see online databases as a lucrative opportunity.

In the meantime, numerous startups are raising capital to get a piece of the growing market. Perhaps the highest-profile one is Snowflake. Back in 2012, Thierry Cruanes, Benoit Dageville, and Marcin Zukowski co-founded the company. They had experience in the data warehouse industry, having worked at companies like Oracle, IBM, and Google.

Snowflake is a fairly easy system to use. All you need to do is fill out a form to create a sophisticated database. The underlying architecture also is well designed. By separating the storage and compute functions, there is much scaling and the platform is multi-cloud. This is essential for AI applications.

In early 2020, Snowflake raised $479 million at a $12.4 billion valuation from investors like Salesforce.com (NYSE:CRM), Sequoia, Altimeter Capital, and Dragoneer Investment Group. The company would then pull off the biggest software IPO ever, raising 3.4 billion. The market value was at over 67 billion.

Among the cloud-based systems, there has been another interesting technology: the data lake. This allows for ingesting huge databases of structured and unstructured data, which often means little need for reformatting. A data lake is also ideal for AI because they work seamlessly with Apache Spark, TensorFlow, and other analytics platforms.

And finally, there are feature stores, which are databases for hyperscale AI companies like Twitter, Facebook, and Airbnb. This is the most cutting-edge technology for handling data in AI models.

The feature store allows for the storage and processing of enormous amounts of features for AI models. The pioneer of this technology is Uber. In 2015, the company was having severe challenges with the hyper growth. Data was

[4] www.mongodb.com/company

scattered across silos and it was nearly impossible to create useful AI models. So Uber set forth to create a proprietary platform called Michelangelo to streamline and accelerate model creation. The result was that the company was able to develop thousands of models, which helped to greatly improve the app.

The Data Challenge

A study from Anaconda, which included 2,360 data scientists, students, and academics/researchers, found that only about 34% of the time spent on AI is on model selection, training, scoring, and deployment. The rest is about the data. Figure 5-1 is a break-down of this.

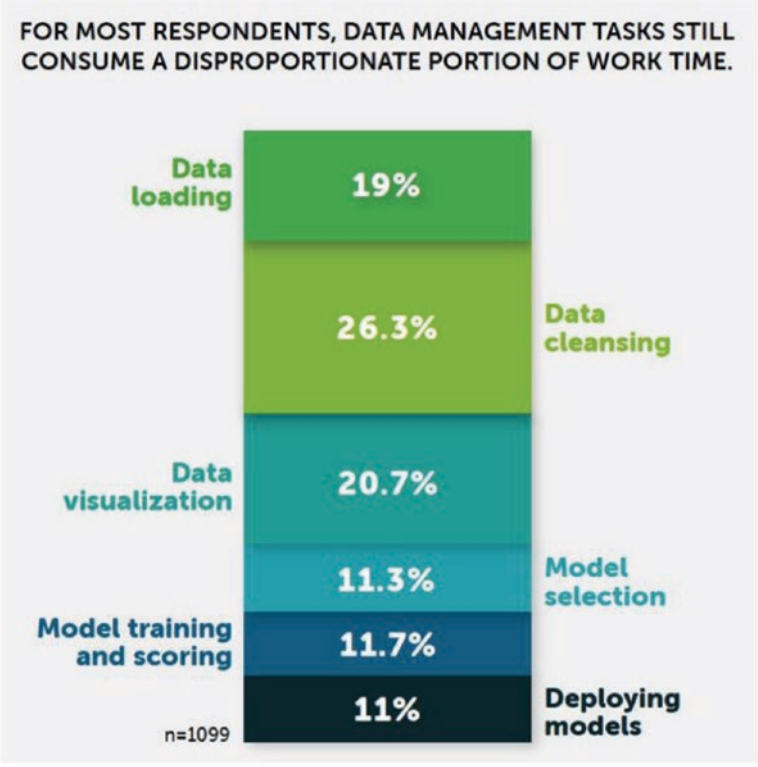

Figure 5-1. In a study from Anaconda, the majority of the time spent on AI is about the data

This may come as a surprise to many, especially business people. After all, the media often focuses on the latest models, such as deep learning. But for these algorithms to have any useful value, there must be high-quality data. And usually huge amounts of it.

Keep in mind that data is often one of the biggest reasons for failed AI models. Consider a report from Appen. It highlights that 40% of executives believe that their organizations do not have enough data or lack the necessary management systems.[5]

Here are some other studies that show issues about data:

- *Gartner*: "Organizations are aware that without sufficient data—or if the situation encountered does not match past data—AI falters. Others know that the more complex the situation, the more likely the situation will not match the AI's existing data, leading to AI failures."[6]
- *Accenture*: From a study of over 1,100 executives, 48% said they have data quality problems and 36% indicated insufficient training data.[7]
- *Deloitte*: "Most executives are not comfortable accessing or using data. Fully 67% of those surveyed (who are senior managers or higher) say they are not comfortable accessing or using data from their tools and resources. The proportion is significant even at companies with strong data-driven cultures, where 37% of respondents still express discomfort."[8]
- *Deloitte*: "Far fewer (18%) have taken advantage of unstructured data (such as product images or customer audio files) or comments from social media."[9]

If you want to have a successful AI implementation, there needs to be a commitment to a strong data strategy. It will build the right foundation. In the rest of the chapter, we'll take a look at some of the strategies to get the most out of the data sets.

---

[5] www.businesswire.com/news/home/20200623005242/en/Appen%E2%80%99s-Annual-State-AI-Report-Finds-Skyrocketing

[6] www.gartner.com/smarterwithgartner/3-barriers-to-ai-adoption/

[7] www.accenture.com/_acnmedia/pdf-73/accenture-strategy-ai-momentum-mind-set-exec-summary-pov.pdf

[8] www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html

[9] www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html

Data Collection

The first step in the data process is to see what datasets are available to your organization. This may take some time as data is often spread across different silos. To this end, it's a good idea to talk to managers across different departments. Some of the data sets will be internally created, say from the website, mobile apps, and IoT systems, but there may also be licenses to third-party sources.

But note that there are numerous publicly available data sets that can be extremely useful. For example, governments often provide information for free. There are also sites like Kaggle.com that have plentiful sources.

Interestingly enough, when going through this process, you may miss certain types of data. After all, what about using information from customer surveys? Social media interactions? What about recorded phone calls with customers and prospects? So think expansively when considering data sources.

Even if there is enough data for a model, there could be another issue: corporate politics. Data can be the subject of intense turf wars! This is why it is essential to have an executive sponsor who can deal with these problems.

Data Evaluation

After you have put together your data assets, you can then do some preliminary assessments. This is important because you do not want to spend much time on the data process only to realize that the sources are not good candidates for AI.

First of all, you want to see if the data is relevant. Does it have the kinds of features that will help provide useful outcomes for your model? True, this can be tough to evaluate in the early stages of a project. But an experienced data scientist should be able to come up with a good answer.

Next, is the data timely or is it too old? In some cases, data can become obsolete quickly. If so, you want to make sure you have access to ongoing data streams, or else the model will probably not work.

Another thing to keep in mind is the target audience. Is the population representative? Or is it skewed? For example, if there are a large number of males, then the model may lead to outcomes that are biased.

During the early stages, you also should get a legal review done—that is, does the organization have the right to use the data for the particular purpose of AI? If the information has personally identifiable items, then there may be onerous restrictions or even bans.

Finally, you want to see if your organization has the right IT infrastructure for the data. Can you handle the volumes with your storage systems or data warehouses? If not, what will it take to build such a system? What are the costs?

Data Wrangling

Data wrangling or data preparation is the process of cleaning up, restructuring, and improving a raw data sets. This is an essential step before you can perform an AI model.

"Predictions by a machine learning model are only as good as the data on which the model is based," said Rosaria Silipo, who is a Ph.D. and Principal Data Scientist at KNIME. "And often the data is crap. Missing values, lack of standardization, empty or almost empty attributes, irrelevant attributes, errors in data collection … you name it! Those are all possible causes for low-quality input data. Often, information is in the data; it is just hidden beneath it. Some data cleaning and appropriate data transformation can bring the hidden information up to the surface. The problem in a data lab is that you hire a number of data scientists, and all they want to do is try and tune machine learning models to gain that .01% more in accuracy. Somehow, the data cleaning is the least attractive part in the job description. So, you might think that hiring a data engineer, less specialized than a data scientist, would solve the problem. And this is partially true. However, often it is hard to get data engineers willing to work on data blending and data wrangling tasks who, at the same time, are experts in complex data science scripts. Some help here might come from tools relying on a graphical user interface rather than on scripts, on drag-and-drop rather than on code. The easier the tool, the easier it will be to find data engineers up to the challenge."[10]

An example of a data-wrangling tool is Trifacta, which was launched in 2012. The product grew out of a project from researchers at UC Berkeley and Stanford who wanted to find much better ways to streamline model development.

Trifacta can work in any environment, whether on a desktop, the cloud, or a huge data lake. Here are the key workflows for the platform that mimic what a typical data scientist would use:

[10]From the author's interview with Rosaria Silipo on June 12, 2020.

- *Discovery*: This helps explore the usefulness and value of the data for the given project. Trifacta will also detect initial problems, like outliers, and provide a general distribution (such as with visualizations).
- *Structure*: This will help put the data set into a structure that works for models. Note that even well-structured datasets are often not enough.
- *Clean up*: This is the process of dealing with issues like duplications, missing values, and so on. But there is a focus on standardizing the information.
- *Enrichment*: With this, you can augment the data by bringing in other internal data sets or third-party databases. This can be complicated, requiring joins, unions, and other derivations.
- *Validation*: This tests the data for consistency and quality. The system also handles values across multiple dimensions.
- *Publishing*: This provides for the delivery of the data, such as loading it into certain analytics packages or even archiving it.

A case study of Trifacta is GlaxoSmithKline, which is a leading pharmaceutical company that has operations in more than 150 countries. It conducts large numbers of clinical trials that require sophisticated data management, such as with the preparation, collection, and distribution.

But one of the issues is searching for data, which is extremely complex. Using the traditional approach, the process could take weeks to months. So GlaxoSmithKline was missing out on opportunities for innovation.

With Trifacta, the company was able to allow clinical researchers to do their own wrangling of the R&D data, and this greatly reduced the turnaround times. A big part of this was allowing for rich visualizations.

But data searching and exploration are not the only benefits. Going forward, GlaxoSmithKline will be in a much better position to use the information to build advanced AI models. This should further accelerate the drug development process and provide for new discoveries.

However, data tools like Trifacta cannot do everything. There is still a need to have some level of human intervention in the data wrangling process.

So then, here are some of the common techniques:

- *Outliers*: This is where some data is significantly outside the general range of the distribution. This could mean that some of the items are really not representative or there could be errors. Then again, there are scenarios where you actively look for outliers, such as for fraud detection (which involves low-probability events).
- *Standardization*: Data may have inconsistent formats. For example, "California" may be abbreviated as "CA." For purposes of a model, this can degrade the outcome. This is why there should be standardization for how data is expressed or labeled.
- *Duplications*: This is common with data sets. And yes, you want to root them out.
- *Creation*: With a certain data type, you can create a new one. For example, if you have dates of births, you can calculate the ages by subtracting the years from the current year.
- *Conversation table*: This is a system that translates data from one standard to another. An example is where you have information expressed in the decimal system and you want to convert it to the metric system.
- *Binning*: You may not necessarily need to be too granular for the data. Hey, is it really important whether a person is 25 or 27? In most cases, the answer is no. Instead, you can group the data with ranges, such as between 20 to 40 and so on.
- *One-hot encoding*: To see how this works, let's take an example. Suppose you have a data set that has three types of iPhones: iPhone 11 Pro, iPhone 11, and iPhone SE. You could convert them into numbers, where iPhone 11 Pro is 1, iPhone SE is 2, and so on. But when this is processed in a model, the AI may consider iPhone SE better because it has a higher number! But with one-hot encoding, you can come up with a neutral classification system, such as is_iPhone11Pro, is_iPhone11, and is_iPhoneSE. Thus, for each row of data, you would put 1 for the phone that is in use and 0 for the rest.
- *Validation rules*: You can put rules in places to improve the quality of the data. For example, if an age is a negative number, it can be flagged.

- *Missing data*: One approach is to use an average for this. True, it is not perfect, but it can help smooth out some of the gaps. Although, if there is a large amount of missing data, it might be advisable to not use this information.

Data Labeling

Creating labels for data, which is known as data annotation, is often required because the raw data set is not in a workable form.

"Labeling the data consistently at scale can be both a technological and business challenge," said Alyssa Simpson Rochwerger, who is the VP of AI and Data at Appen. "Labeling large volumes of high-quality data can be expensive, and it's easy to spend a lot of money and get back bad data. It's easy to label 100 or 1,000 examples, but it's an entirely different matter to label 100k or 1M or 100M items to achieve the accuracy and precision a business use case requires. Take, for example, a large financial services company we work with that builds a product that allows you to take a picture of a receipt and 'magically' it is transcribed and uploaded into the expense software. Behind the scenes, that requires lots of humans to transcribe receipts since the AI model requires such high precision. Can you imagine if there was a mistake? The company would look silly! Right now, they have built a model that allows for them to send 10% of the data to a model with high confidence, and they are expanding that over time using high quality training data."[11]

However, the data labeling is not a "set and forget it" process. There usually needs to be ongoing refreshing in order to have success with the models. "It's not enough to gather and label training data at one point in time and never come back to it," said Rochwerger. "It's critical to think through a constant feedback loop and training data pipeline to measure the performance of the model over time as well as retrain and address low accuracy or low performant areas. For my example about receipts, when the expense product first launched, the team only took into consideration US-based receipts as in US dollars, with the date format in the US and English language. However, some customers were uploading receipts from international business trips and the product could not handle those use cases."

As you saw earlier in this book, companies like Facebook have used semi-structured data, like tags on Instagram, to help accelerate the process. But unfortunately, this is usually not an option for most data sets.

Even if the data set is manageable, you still may not want to do the labeling in-house. Keep in mind that this process can be intricate and complicated. You will likely need to purchase some software tools but also have some people

[11] From the author's interview with Alyssa Simpson Rochwerger on June 5, 2020.

for quality assurance. If you do not set things up properly, the data could easily be corrupted.

Another option is to use a crowdsourcing option, which can certainly be effective. With this approach, you will specify how you want the data set labeled and then the third-party provider will recruit the people for the project.

One of the top labeling outsourcers is Appen. The company has actually been around since 1996 and is publicly traded on the Australian stock exchange. During 2019, revenues jumped by 47% to $536 million and adjusted earnings came to $101 million. This is a clear sign that the data labeling industry is seeing tremendous growth.

For its training data services, Appen has more than one million contractors in more than 130 countries who speak 180 languages and dialects. But Appen also has created its own AI platform that helps to automate the data labeling process, which is built for quality, accuracy, and speed.

For a case study of a customer, take a look at Zefr. The company develops technology for YouTube ad targeting and has customers like Target, Netflix, Adidas, and Honda.

But when Zefr internally crowdsourced the data labeling, it could only handle about 15,000 videos per month. This was simply not enough for the scale of Zefr. By using Appen's contractors, the company was able to increase the output to about 100,000 per month. Because of this, Zefr was able to have much richer and effective data to train models for more accurate video recommendations.

"For data labeling, human-in-the-loop training data provides the highest quality," said Wilson Pang, who is the Chief Technology Officer at Appen. "People involved in the labeling, training, testing, and tuning will ensure a higher rate of accuracy and success of a project."[12]

Note According to a study from research firm Cognilytica, the market for outsourcing data labeling is expected to go from $150 million in 2018 to $1 billion by 2023.[13]

---

[12]From the author's interview with Wilson Pang on June 2, 2020.
[13]www.ft.com/content/56dde36c-aa40-11e9-984c-fac8325aaa04

Simulation

While data is growing at a dramatic pace, there is a nagging issue: a lot of the data is not useful or good quality. This is particularly troubling for advanced AI use cases, such as with self-driving cars. Even if a company collects data from cars that drive millions of miles, there will be many important scenarios that are missed.

What to do then? Does this mean something like autonomous vehicles are impossible? Not necessarily, although this technology has certainly proven extremely challenging.

Data scientists are finding creative ways to deal with the data shortages. One approach is to use simulations.

"Sometimes data is difficult to acquire for the conditions you want the AI system to find," said Paul Pilotte, who is the AI technical marketing lead at MathWorks. "For example, you could have a hydraulic pump with failure conditions like a worn bearing or a seal leak that you want to find. These conditions rarely happen and are destructive, making it very difficult to get failure data to train an AI model. That's where simulation comes in. You can use a model of the pump and run simulations to produce signals representing failure behavior, signals that can be used to train an AI model to detect the future occurrence of it on real systems in the field. The combination of automated tools to label data and simulation to generate synthetic data are key tools to help teams create the labeled data needed for AI systems."[14]

A company that has been on the leading edge of simulation is Waymo, which is Google's self-driving unit. Researchers from Waymo have created SurfelGAN, which uses texture mapping to come up with richer scenes from camera data. While simulators are not necessarily new, this one is generally more versatile because it can create data with a myriad of distances and angles. The system also does this with fairly low needs for computations. This means that SurfelGAN can work in real time. And as the name implies, this AI uses a generative adversarial network (GAN), which is useful in data creation. SurfelGAN also does not need labeled data.[15]

How Much Data Do You Need?

In many cases, you will need a large amount of data for an AI project. There is Hughes Phenomenon, which indicates that the more features you add to the model, the higher the accuracy.

---

[14]This is from the author's interview with Paul Pilotte on May 21, 2020.

[15]https://venturebeat.com/2020/05/20/waymo-is-using-ai-to-simulate-autonomous-vehicle-camera-data/

"One of the major challenges in machine learning is the data efficiency problem," said Ryan Sinnet, who is the CTO and co-founder of Miso Robotics. "While machines can often learn to be more accurate than humans, it takes machines a lot more practice than humans. You may be able to learn to recognize a new exotic vegetable from a handful of pictures whereas a machine may require several thousand pictures."

But there are definitely exceptions. In fact, a model can have too much data, which will result in a degrading of the effectiveness. This is known as the curse of dimensionality.

Thus, it's a good idea to not have any preconceived notions about how much data is needed. The fact is that some models may actually need very little amounts of data. According to Andrew Ng, who is the CEO of Landing AI and the former head of Google Brain, some projects may require only a mere 100 data points.

Consider the new field of study emerging in AI known as small data. It means that a model can be effectively and efficiently trained on small data sets.

An example of this is Google Research's new model, called Entities as Experts (EAE). It includes an assortment of people, organizations, times, figures, and so on. As for the model, it has been shown to solve complex natural language challenges without the need to use entity-specific knowledge.

For example, EAE was used to analyze Wikipedia posts that had more than 17 million entity mentions. But the system only needed to keep about one million of them to provide effective results.[16]

Data Problems

In New Zealand in March of 2019, a shooter live-streamed on Facebook his horrific killings of 50 people in two mosques. The online connection was not cut off until 29 minutes after the attack started. This meant that there were millions of views.

How was this possible? Part of this was due to bad actors who attempted to foil the Facebook system. But the underlying AI technology was ineffectual as well, primarily because of issues with the data.

In a blog post, Facebook's VP of Product Management, Guy Rosen, described this as follows: "AI systems are based on 'training data,' which means you need many thousands of examples of content in order to train a system that can detect certain types of text, imagery, or video. This approach has worked very well for areas such as nudity, terrorist propaganda, and also graphic violence

[16] https://venturebeat.com/2020/04/20/googles-entities-as-experts-ai-answers-text-based-questions-with-less-data/

where there is a large number of examples we can use to train our systems. However, this particular video did not trigger our automatic detection systems. To achieve that, we will need to provide our systems with large volumes of data of this specific kind of content, something which is difficult as these events are thankfully rare. Another challenge is to automatically discern this content from visually similar, innocuous content, for example if thousands of videos from live-streamed video games are flagged by our systems, our reviewers could miss the important real-world videos where we could alert first responders to get help on the ground."[17]

The lack of useful data has been a major issue with AI projects. Yet researchers and data scientists are finding creative ways to deal with this. Actually, this is one of the most important areas in the AI field and we'll likely see more innovations in the years ahead.

More Data Concepts

The data field definitely has lots of jargon. And it would not be possible to cover everything! But to end this chapter, let's take a look at a few other terms that you will likely encounter:

- *Categorical data*: This is data that does not have a numerical meaning. Instead, it is based on text, like a description of a group or category (say race or gender). But you can assign numbers to each of the elements.
- *Data type*: This is the kind of data a variable represents. Some examples include Booleans (true/false values), strings, integers, and floats (a number that has a decimal point).
- *Feature*: This is a column of data.
- *Instance*: This is a row of data.
- *Metadata*: This is data that describes other data. An example is a video file, which has metadata like size, date of the upload, comments, topic, and so on. Note that this type of data can be extremely useful in an AI model.
- *Ordinal data*: This is a mix of numerical and categorical data. An example is a five-star rating system on an app like Yelp.
- *Transactional data*: This is data generated from actions on ERPs and other enterprise systems.

[17] https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/

Conclusion

You can easily get bogged down with data. This is why you need to be practical. There is no such thing as a perfect data set because there will always be issues and challenges. But as seen in this chapter, there are many best practices and automation tools that can streamline the process, which can get to effective outcomes quicker.

In the next chapter, we will take a look at AI model building.

Key Takeaways

- Data is essential for success with AI. It's a key reason why data-rich companies like Netflix, Google, Facebook, and Microsoft have been so successful with this technology.
- Data is growing at a rapid pace. According to IDC, the amount forecasted for the next three years will be the equivalent of what has been created for the past three decades. The COVID-19 pandemic is a major catalyst for growth, as there has been an accelerated shift to data-intensive technologies like videoconferencing.
- There are four main types of data: structured data, unstructured data, semi-structured data, and time-series data.
    - *Structured data*: This is information stored in spreadsheets and databases. For the most part, this type of information is fairly easy to use in AI models. However, there is generally less structured data available to work with.
    - *Unstructured data*: Examples include images, emails, videos, text files, satellite images, and social media messages. The majority of data is unstructured.
    - *Semi-structured data*: This is a blend of structured and unstructured data.
    - *Time-series data*: This shows interactions of data, such as with the "customer journey." The data can be structured or unstructured.

- Big data has three key attributes: volume, variety, and velocity, which are known as the three Vs.
    - *Volume*: This describes the huge scale of the data, at least in the tens of terabytes.
    - *Variety*: This shows the diversity of data, say with structured, unstructured, and semi-structured sources.
    - *Velocity*: This focuses on the speed that the data is generated. This is perhaps the most challenging part of big data.
- Over the years, other Vs have emerged: visualization of the data, value (how effective the information is), veracity (whether the sources can be trusted), and so on.
- A relational database has several features that make it much easier to work with data. At the heart of this technology is SQL or Structured Query Language, which is an English-like scripting system that helps to create tables, read them, make updates, and handle deletions. But relational databases also allow for making relationships among the tables.
- Yet relational databases certainly have their downsides, such as data sprawl and difficulties handling modern-day use cases like unstructured data and high-velocity data.
- Over the years, there have emerged next-generation databases to deal with the problems. One technology is the data warehouse, which efficiently handles large amounts of data. But there has also been the development of NoSQL databases. They use a document model, which provides a high degree of flexibility.
- The data lake has also become important. This makes it possible to have massive storage of data of any format.
- Then there is the feature store. This is a next-generation database that is built specifically for AI and ML.
- When it comes to developing AI projects, the data preparation is usually the most time-consuming. Despite this, many companies still do not devote enough resources to this part of the process. Yet problems with data are often the main reason why an AI project fails.

- Data collection is the first step in the data preparation process. You need to get an inventory of the assets of the organization. Then, you can look elsewhere for data sources, which may involve paying for licenses to third-party databases. But there are also many freely available sources, such as from the government.
- The next step is to evaluate the data. This involves seeing if the data is relevant, timely, and representative of the population. There should also be a legal review to see if the data can be used and if the IT infrastructure is able to handle the storage and processing.
- Data wrangling is a critical step. It is about improving the quality of the data set, such as dealing with missing items, duplications, and so on. There are automation tools that can help with this process. But someone with data science expertise needs to provide assistance.
- Labeling data is tedious and time consuming. Now there are automation tools to help out but they usually have limitations. Because of this, many companies will still use people for labeling. There are a variety of companies that provide this service, such as by using crowdsourcing.
- While many AI applications need enormous amounts data, there are still use cases where the needs are far less. Even a dataset of only 100 items can be enough.