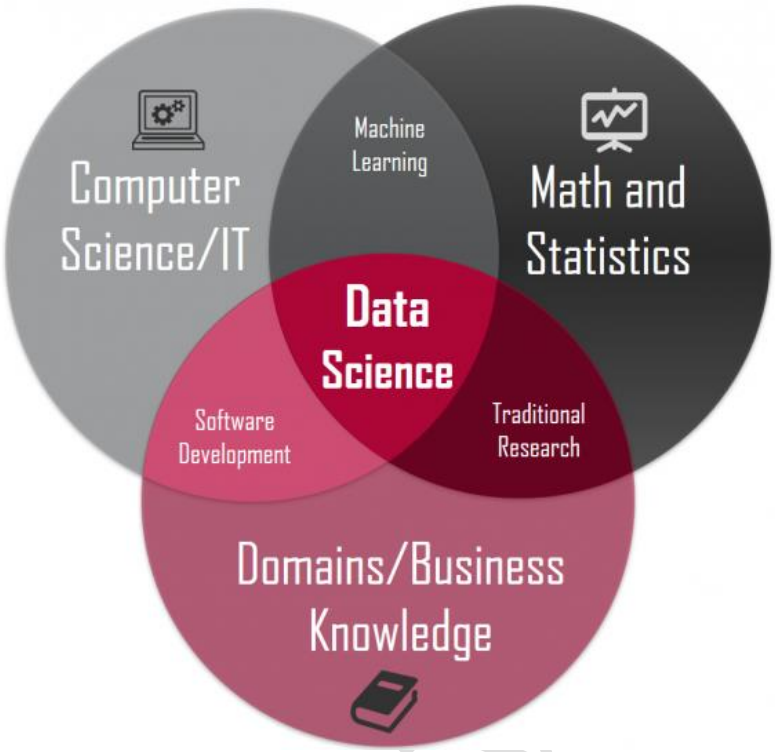
Unit-1

Outline:

1. **What is Data Science?**
2. **What is Big Data?**
    a. **4 V's of Big Data**
3. **What is Data Science not?**
    a. **Not Machine Learning**
    b. **Not Statistics**
    c. **Not Big Data**
4. **Big Data vs. Data Science**
5. **Few Example case studies**
6. **Data Science Process**
7. **What are the responsibilities of Data Scientists?**
8. **Kinds of Data**

1) What is Data Science?

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into tangible business value.



The main aspect of data science is discovering new results from data. People are exploring at a granular level to understand and mine complex inferences, behaviors, and trends. It's about uncovering hidden information that may be able to help companies make smarter choices for their business, For example:

- Data mines in Netflix are used to look for movie viewing patterns to better understand user's interests and to make decisions on the Netflix series they should produce.
- Target tries to find the major customer segments in its customer base and their shopping behaviors, which helps them to guide messaging to other market groups.
- Proctor and Gamble looks towards time series models to help them to understand the future demand and plan production levels.

So how does the data scientist mine all this information? It begins with **data exploration.** When a data scientist is given a challenging question, they become a detective. They will start to investigate leads, and then try to understand characteristics or patterns in the data. This means they need a lot of analytical creativity.

Then a data scientist can use quantitative techniques to dive a little deeper, such as synthetic control experiments, time series forecasting, segmentation, and inferential models. The purpose of these is to use data to piece together a better understanding of the information.

The use of data-driven insight is what helps to provide strategic guidance. This means that a data scientist works a lot like a consultant, guiding businesses on how they should respond to their findings. Data science will then give you a **data product**. **Data products are a technical asset that:**

1. **Uses data like input.**
2. **Processes the data to get an algorithmically-generated result.**

One of the classic examples of a data product is an engine which takes in user data, and then creates a personalized recommendation based upon that data. The following are some **examples of data products**:

- The recommendation engine that Amazon uses suggests new items to its users, which is determined by their algorithms. Spotify recommends new music. Netflix recommends new movies.
- The spam filter in Gmail is a data product. This is a behind the scenes algorithm that processes the incoming mail and decides whether or not it is junk.
- The computer vision that is used for self-driving cars is also a data product. Machine learning algorithms can recognize pedestrians, traffic lights, other cars, and so on.

Data products work differently than data insights. **Data insights help to provide some advice to help a business executive make smarter decisions**. **Data products are a technical functionality that encompasses the algorithm, and it is designed to work into the main applications**.

The data scientist plays one of the most central roles in coming up with the data product. This means that they have to build out algorithms and test, refine, and

technically deploy it into a production system. The data scientist also works as a technical developer by creating assets that become leverage on a wide scale.

2) What is Big Data?

According to Gartner, the definition of Big Data –

"Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

4 V's of Big data:

1) Variety

Variety of Big Data refers to structured, unstructured, and semistructured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

Variety is one the most interesting developments in technology as more and more information is digitized. Traditional data types (structured data) include things on a bank statement like date, amount, and time. These are things that fit neatly in a relational database.

Structured data is augmented by unstructured data, which is where things like Twitter feeds, audio files, MRI images, web pages, web logs are put — anything that can be captured and stored but doesn't have a meta model (a set of rules to frame a concept or idea — it defines a class of information and how to express it) that neatly defines it.

Unstructured data is a fundamental concept in big data. The best way to understand unstructured data is by comparing it to structured data. Think of structured data as data that is well defined in a set of rules. For example, money will always be numbers and have at least two decimal points; names are expressed as text; and dates follow a specific pattern.

With unstructured data, on the other hand, there are no rules. A picture, a voice recording, a tweet — they all can be different but express ideas and thoughts based on human understanding. One of the goals of big data is to use technology to take this unstructured data and make sense of it.

The definition of big data depends on whether the data can be ingested, processed, and examined in a time that meets a particular business's requirements. For one company or system, big data may be 50TB; for another, it may be 10PB.

2) Velocity

Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

3) Volume

Volume is one of the characteristics of big data. We already know that Big Data indicates huge 'volumes' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data are stored in data warehouses. Thus comes to the end of characteristics of big data.

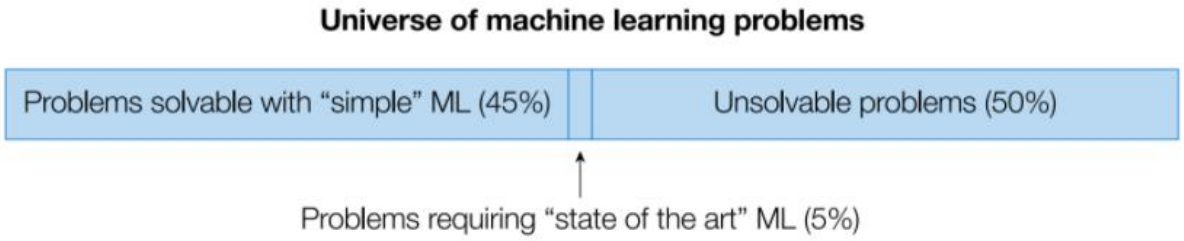
4) Veracity

Veracity refers to the trustworthiness of the data. Can the manager rely on the fact that the data is representative? Every good manager knows that there are inherent discrepancies in all the data collected.

3) What is Data Science not?

a) Data science is not (just) machine Learning

Making good machine learning based predictions can be an important part of data science, but the truly hard elements of data science involve also involve collecting the data to begin with, defining the problem you're trying to solve (and frequently, re-defining it many times based upon improved understanding of the problem over time), and then interpreting and understanding the results, and knowing what actions to take based upon this. Furthermore, machine learning as a field is typically concerned primarily about the development of new algorithms. Deep learning methods, for instance (typically defined as methods based upon multi-layer neural networks, though honestly the right definition of "deep learning" is as hard as the right definition of "data science", more on that in a later lecture) have come to dominate much recent work in machine learning, and the focus here is often on advanced (and typically quite complex) algorithms that can squeeze out improved performance on extremely challenging tasks. The reality is that for many data science problems. simple machine learning algorithms suffice to attain sufficiently good performance (by whatever metric you want to define performance, but I simply mean that they effectively solve the problem). I'm partial to drawing this picture for people when they ask about how data science compares to machine learning research:



A (rhetorical) breakdown of the universe of machine learning problems

The numbers here are all just examples (specifically the solvable/unsolvable ratio), but the point it gets at is important. There are many data science problems one would like to be able to solve, but in a large number of these cases, there is simply no way to solve the problem given the available data. For the set of problems that are solvable with some kind of machine learning, the vast majority will be solvable at least to a level of sufficient performance, using relatively simple models. The 5% of remaining problems is an important one, because they often consist of the most "interesting" problems from a research standpoint (think problems like speech recognition, natural language understanding, computer vision), but they are often not indicative of the types of problems one encounters in "most" data science applications.

b) Data science is not (just) statistics

So if data science is not machine learning, perhaps simple Statistics (that is, the discipline of Statistics) is a better fit? After all, "analyzing data computationally" ... "to understand phenomena in the real world"? This sounds an awful lot like statistics. And I think the fit here is frankly much better than for the standard definition of machine learning. Indeed, the Statistics department at CMU just renamed itself to "Statistics and Data Science".

There are, however, two primary distinctions that are worth making about between data science and Statistics as it is commonly practiced in an academic setting. The first is that historically, the academic field of statistics has tended more towards the theoretical aspects of data analysis than the practical aspects. David Donoho has an excellent article

on this subject, 50 years of data science (http://courses.csail.mit.edu/3.8.337/201.5/docs/50YearsDataScience.pdf), which doubles as an alternative view of data science from a more statistics-centric standpoint, which does an excellent job of covering the distinctions between traditional statistics. Second, from a historical context, data science has evolved from computer science as much as it has from statistics: topics like data scraping, and data processing more generally, are core to data science, typically are steeped more in the historical context of computer science, and are unlikely to appear in many statistics courses.

The last difference, of course, is that statisticians use R, while data scientists use Python. This is non-negotiable. (Since my sarcasm may not come across correctly in written form, yes, I'm saying this tongue-in-cheek. but there absolutely is a grain of truth here: R, or more specifically, the vast set of libraries that have been developed for R. are far superior for running advanced statistical algorithms; but Python is a much nicer language for collecting and processing data, especially if you consider the set of external libraries it supports).

c) Data science is not (just) big data

Lastly. there is still a contingent that equates data science with the rise of big data (again we're likely dealing with a term that needs its own set of notes defining it. for the purposes of these notes. let's assume that big data just refers to data that can't easily fit in memory on a single machine). And while it's absolutely true that some data science work really does use vast amounts of data to build models or gain insights, this is frankly the exception rather than the rule. Most data science can work just fine using the (right set of) data that still fits into memory on a single machine. It's useful to know the techniques needed to address big data challenges. but don't create more work for yourself if you don't have to.

4) Big Data vs. Data Science:

Here are the main differences between data science and big data.

- Organizations have to gather big data to help improve their efficiency, enhance competitiveness, and understand new markets. Data science provides the mechanisms or tools to understand and use big data quickly.
- There is no limit to how much valuable data that can be collected. To use this data, the important information for business decisions has to be extracted. This is where data science is needed.
- People characterize big data by its volume, velocity, and variety, which is often referred to as the 3Vs. Data science provides the techniques and methods to look at the data that is characterized by the 3Vs.
- Big data provides a business with the possibility for better performance. However, finding that information in big data to utilize its potential to enhance performance is a challenge. Data science will use experimental and theoretical approaches as well as inductive and deductive reasoning. It will unearth the hidden insight from all the complexity of unstructured data, which will support organizations to notice the potential of all the big data.
- Big data analysts perform mining of helpful information from large sets of data. Data scientists use statistical methods and machine learning algorithms to train computers to find information without a lot of programming, and to make predictions based upon big data. Because of this, it's important that you don't confuse data science with big data analytics.
- Big data relates to technology, such as Hive, Java, Hadoop, and so on, and is a distributed computing, and software and analytics tool. This is different than data science, which looks at the strategy for business decisions, data structures and statistics, data dissemination using math, and all other methods mentioned earlier.

Through the differences between data science and big data, you can see that data science is included as part of the concept of big data. Data science is an important part of a lot of different application areas. Data science uses big data to find the most helpful insights through predictive

analysis. The results are then used to make better choices. Data science is included as a part of big data, but big data is not included in data science.

The following is to help show the fundamental differences:

- Meaning:
  - Big Data:
    - Large volumes of data that can't be handled using a normal database program.
    - Characterized by velocity, volume, and variety.
  - Data Science:
    - Data focused scientific activity.
    - Similar in nature to data mining.
    - Harnesses the potential of big data to support business decisions.
    - Includes approaches to process big data.
- Concept:
  - Big Data:
    - Includes all formats and types of data.
    - Diverse data types are generated from several different sources.
  - Data Science:
    - Helps organizations make decisions.
    - Provides techniques to help extract insights and information to create large datasets.
    - A specialized approach that involves scientific programming tools, technique and models to process big data.
- Basis of Formation:
  - Big Data:
    - Data is generated from system logs.
    - Data is created in organizations — emails, spreadsheets, DB, transactions, and so on.
    - Online discussion forums.

        - Video and audio streams that include live feeds.
        - Electronic devices — RFID, sensors, and so on. • Internet traffic and users.
    - Data Science:
        - Working apps are made by programming developed models.
        - It captures complex patterns from big data and developed models.
        - It is related to data analysis, preparation, and filtering.
        - Applies scientific methods to find the knowledge in big data.
- Application Areas:
    - Big Data:
        - Security and law enforcement.
        - Research and development.
        - Commerce.
        - Sports and health.
        - Performance optimization.
        - Optimizing business processes.
        - Telecommunications.
        - Financial services.
    - Data Science:
        - Web development.
        - Fraud and risk detection.
        - Image and speech recognition.
        - Search recommenders.
        - Digital advertisements.
        - Internet search.
        - Other miscellaneous areas and utilities.
- Approach
    - Big Data:
        - To understand the market and to gain new customers.
        - To find sustainability.

- To establish realistic ROI and metrics.
- To leverage datasets for the advantage of the business.
- To gain competitiveness.
- To develop business agility.

- Data Science:
  - Data Visualization and prediction.
  - Data destroy, preserve, publishing, processing, preparation, or acquisition.
  - Programming skills, like NoSQL, SQL, and Hadoop platforms.
  - State-of-the-art algorithms and techniques for data mining.
  - Involves the extensive use of statistics, mathematics, and other tools.

4) Applications / Case studies of Data Science

Using data science, companies have become intelligent enough to push & sell products as per customers purchasing power & interest. Here's how they are ruling our hearts and minds:

Internet Search

When we speak of search, we think 'Google'. Right? But there are many other search engines like Yahoo, Bing, Ask, AOL, Duckduckgo etc. All these search engines (including Google) make use of data science algorithms to deliver the best result for our searched query in fraction of seconds. Considering the fact that, Google processes more than 20 petabytes of data everyday. Had there been no data science, Google wouldn't have been the 'Google' we know today.

Digital Advertisements (Targeted Advertising and re-targeting)

If you thought Search would have been the biggest application of data science and machine learning, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various

websites to the digital bill boards at the airports – almost all of them are decided by using data science algorithms.

This is the reason why digital ads have been able to get a lot higher CTR than traditional advertisements. They can be targeted based on user's past behaviour. This is the reason why I see ads of analytics trainings while my friend sees ad of apparels in the same place at the same time.

Recommender Systems:
Who can forget the suggestions about similar products on Amazon? They not only help you find relevant products from billions of products available with them, but also adds a lot to the user experience.

A lot of companies have fervidly used this engine / system to promote their products / suggestions in accordance with user's interest and relevance of information. Internet giants like Amazon, Twitter, Google Play, Netflix, Linkedin, imdb and many more uses this system to improve user experience. The recommendations are made based on previous search results for a user.

Image Recognition:
You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. Similarly, while using whatsapp web, you scan a barcode in your web browser using your mobile phone. In addition, Google provides you the option to search for images by uploading them. It uses image recognition and provides related search results.

Speech Recognition:
Some of the best example of speech recognition products are Google Voice, Siri, Cortana etc. Using speech recognition feature, even if you aren't in a position to type a message, your life wouldn't stop. Simply speak out the message and it will be converted to text. However, at times,

you would realize, speech recognition doesn't perform accurately. Just for laugh, check out this hilarious video(1:30 mins) and the conversation between Cortana & Satya Nadela (CEO, Microsoft)

Gaming:
EA Sports, Zynga, Sony, Nintendo, Activision-Blizzard have led gaming experience to the next level using data science. Games are now designed using machine learning algorithms which improve / upgrade themselves as the player moves up to a higher level. In motion gaming also, your opponent (computer) analyzes your previous moves and accordingly shapes up its game.

Price Comparison Websites:
At a basic level, these websites are being driven by lots and lots of data which is fetched using APIs and RSS Feeds. If you have ever used these websites, you would know, the convenience of comparing the price of a product from multiple vendors at one place. PriceGrabber, PriceRunner, Junglee, Shopzilla, DealTime are some examples of price comparison websites. Now a days, price comparison website can be found in almost every domain such as technology, hospitality, automobiles, durables, apparels etc.

Airline Route Planning:
Airline Industry across the world is known to bear heavy losses. Except a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air fuel prices and need to offer heavy discounts to customers has further made the situation worse. It wasn't for long when airlines companies started using data science to identify the strategic areas of improvements. Now using data science, the airline companies can:

Predict flight delay
Decide which class of airplanes to buy

Whether to directly land at the destination, or take a halt in between (For example: A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)
Effectively drive customer loyalty programs
Southwest Airlines, Alaska Airlines are among the top companies who've embraced data science to bring changes in their way of working.

Fraud and Risk Detection:
One of the first applications of data science originated from Finance discipline. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paper work while sanctioning loans. They decided to bring in data science practices in order to rescue them out of losses. Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

Delivery logistics:
Who says data science has limited applications? Logistic companies like DHL, FedEx, UPS, Kuhne+Nagel have used data science to improve their operational efficiency. Using data science, these companies have discovered the best routes to ship, the best suited time to deliver, the best mode of transport to choose thus leading to cost efficiency, and many more to mention. Further more, the data that these companies generate using the GPS installed, provides them a lots of possibilities to explore using data science.

Miscellaneous
Apart from the applications mentioned above, data science is also used in Marketing, Finance, Human Resources, Health Care, Government Policies and every possible industry where data gets generated. Using data science, the marketing departments of companies decide which products are best for Up selling and cross selling, based on the behavioral data

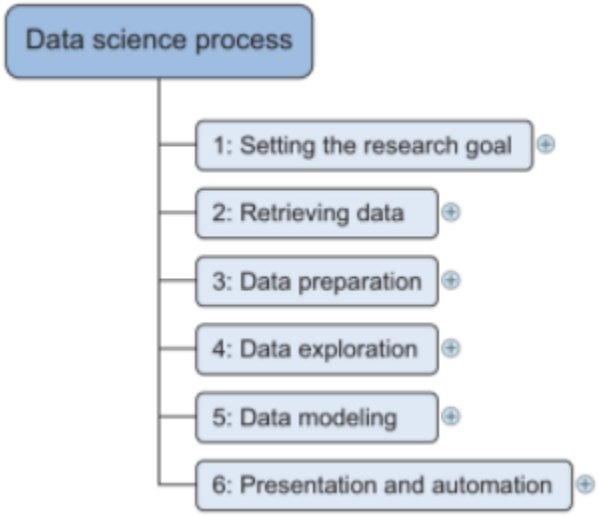
from customers. In addition, predicting the wallet share of a customer, which customer is likely to churn, which customer should be pitched for high value product and many other questions can be easily answered by data science. Finance (Credit Risk, Fraud), Human Resources (which employees are most likely to leave, employees performance, decide employees bonus) and many other tasks are easily accomplished using data science in these disciplines.

5) Data science process.

The data science process typically consists of six steps, as you can see in the mind map in figure.

1. Setting the research goal

Data science is mostly applied in the context of an organization. When the business asks you to perform a data science project, you'll first prepare a project charter. This charter contains information such as what you're going to research, how the company benefits from that, what data and resources you need, a timetable, and deliverables.



2. Retrieving data

The second step is to collect data. You've stated in the project charter which data you need and where you can find it. In this step you ensure that you can use the data in your program, which means checking the existence of, quality, and access to the data. Data can also be delivered by third-party companies and takes many forms ranging from Excel spreadsheets to different types of databases.

3. Data preparation

Data collection is an error-prone process; in this phase you enhance the quality of the data and prepare it for use in subsequent steps. This phase consists of three sub- phases: data cleansing removes false values from a data source and inconsistencies across data sources, data integration enriches data sources by combining information from multiple data sources, and data transformation ensures that the data is in a suitable format for use in your models.

4. Data exploration

Data exploration is concerned with building a deeper understanding of your data. You try to understand how variables interact with each other, the distribution of the data, and whether there are outliers. To achieve this you mainly use descriptive statistics, visual techniques, and simple modeling. This step often goes by the abbreviation EDA, for Exploratory Data Analysis.

5. Data modeling or model building

In this phase you use models, domain knowledge, and insights about the data you found in the previous steps to answer the research question. You select a technique from the fields of statistics, machine learning, operations research, and so on. Building a model is an iterative process that involves selecting the variables for the model, executing the model, and model diagnostics.

6. Presentation and automation

Finally, you present the results to your business. These results can take many forms, ranging from presentations to research reports. Sometimes you'll need to automate the execution of the process because the business will want to use the insights you gained in another project or enable an operational process to use the outcome from your model.

AN ITERATIVE PROCESS The previous description of the data science process gives you the impression that you walk through this process in a linear way, but in reality you often have to step back and rework certain findings. For instance, you might find outliers in the data exploration phase that point to data import errors. As part of the data science process you gain incremental insights, which may lead to new questions. To prevent rework, make sure that you scope the business question clearly and thoroughly at the start.

6) What are the responsibilities of Data Scientists?

A data scientist may have to:

- Recommend the most cost-effective changes that should be made to existing strategies and procedures.
- Communicate findings and predictions to IT and management departments through effective reports and visualizations of data.
- Come up with new algorithms to figure out problems and create new tools to automate work.
- Devise data-driven solutions to challenges that are most pressing.
- Examine and explore data from several different angles to find hidden opportunities, weaknesses, and trends.
- Thoroughly prune and clean data to get rid of the irrelevant information.
- Employ sophisticated analytics programs, statistical methods, and machine learning to get data ready for use in a prescriptive and predictive modeling.
- Extract data from several external and internal sources.
- Conduct undirected research and create open-ended questions.

Different companies will have a different idea of data scientist tasks. There are some businesses that will treat their data scientists like glorified data analysts, or combine the duties with data engineering. There are others that need top-

level analytics experts that are skilled in intense data visualizations and machine learning.

As data scientists reach new experience levels or change jobs, the responsibilities they face will change as well. For example, a person that works alone for a mid-sized company may spend most of their day cleaning and munging data. High-level employees that are a part of a business that offers database services could have to create new products or structure big data projects on an almost daily basis.

7) Kinds of Data:

In data science and big data you'll come across many different types of data, and each of them tends to require different tools and techniques. The main categories of data are these:

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

Let's explore all these interesting data types.

1. Structured data

Structured data is data that depends on a data model and resides in a fixed field within a record. As such, it's often easy to store structured data in tables within databases or Excel files (figure 1.1). SQL, or Structured Query Language, is the preferred way to manage and query data that resides in databases. You may also come across structured data that might give you a hard time storing it in a traditional relational database. Hierarchical data such as a family tree is one such example.

The world isn't made up of structured data, though; it's imposed upon it by humans and machines. More often, data comes unstructured.

	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Figure 1.1 An Excel table is an example of structured data.

2. Unstructured data

Unstructured data is data that isn't easy to fit into a data model because the content is context-specific or varying. One example of unstructured data is your regular email (figure 1.2). Although email contains structured elements such as the sender, title, and body text, it's a challenge to find the number of people who have written an email complaint about a specific employee because so many ways exist to refer to a person, for example. The thousands of different languages and dialects out there further complicate this. A human-written email, as shown in figure 1.2, is also a perfect example of natural language data.

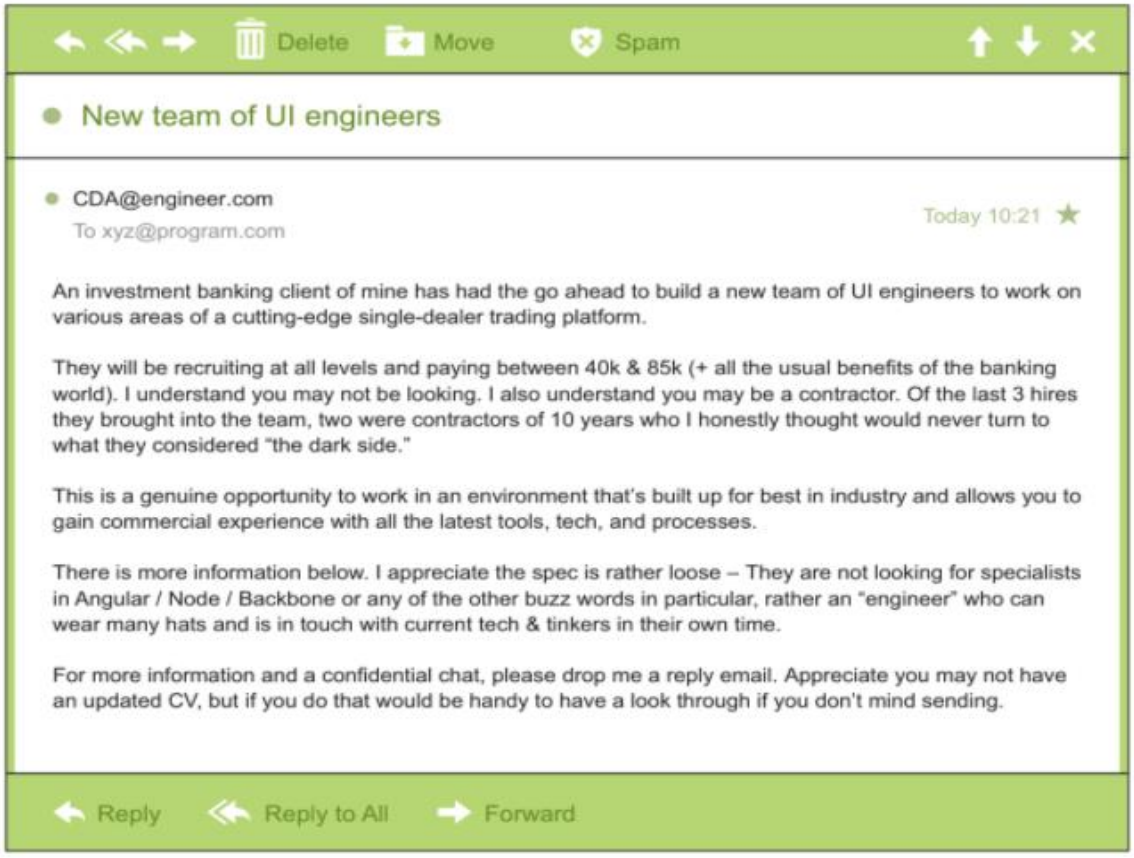


Figure 1.2 Email is simultaneously an example of unstructured data and natural language data.

3. Natural language

Natural language is a special type of unstructured data; it's challenging to process because it requires knowledge of specific data science techniques and linguistics.

The natural language processing community has had success in entity recognition, topic recognition, summarization, text completion, and sentiment analysis, but models trained in one domain don't generalize well to other domains. Even state-of-the-art techniques aren't able to decipher the meaning of every piece of text. This shouldn't be a surprise though: humans struggle with natural language as well. It's ambiguous by nature. The concept of meaning itself is questionable here. Have two people listen to the same conversation. Will they get the same meaning? The meaning of the same words can vary when coming from someone upset or joyous.

4. Machine-generated data

Machine-generated data is information that's automatically created by a computer, process, application, or other machine without human intervention. Machine-generated data is becoming a major data resource and will continue to do so. Wikibon has forecast that the market value of the industrial Internet (a term coined by Frost & Sullivan to refer to the integration of complex physical machinery with networked sensors and software) will be approximately $540 billion in 2020. IDC (International Data Corporation) has estimated there will be 26 times more connected things than people in 2020. This network is commonly referred to as the internet of things. The analysis of machine data relies on highly scalable tools, due to its high volume and speed. Examples of machine data are web server logs, call detail records, network event logs, and telemetry (figure 1.3).

The machine data shown in figure 1.3 would fit nicely in a classic table-structure database. This isn't the best approach for highly interconnected or "networked" data, where the relationships between entities have a valuable role to play.

```
CSIPERF:TXCOMMIT;313236
2014-11-28 11:36:13, Info                  CSI    00000153 Creating NT transaction (seq
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000154 Created NT transaction (seq 69)
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info                  CSI    00000155@2014/11/28:10:36:13.471
Beginning NT transaction commit...
2014-11-28 11:36:13, Info                  CSI    00000156@2014/11/28:10:36:13.705 CSI perf
trace:
CSIPERF:TXCOMMIT;273983
2014-11-28 11:36:13, Info                  CSI    00000157 Creating NT transaction (seq
70), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000158 Created NT transaction (seq 70)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:13, Info                  CSI    00000159@2014/11/28:10:36:13.764
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                  CSI    0000015a@2014/11/28:10:36:14.094 CSI perf
trace:
CSIPERF:TXCOMMIT;386259
2014-11-28 11:36:14, Info                  CSI    0000015b Creating NT transaction (seq
71), objectname [6]"(null)"
2014-11-28 11:36:14, Info                  CSI    0000015c Created NT transaction (seq 71)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:14, Info                  CSI    0000015d@2014/11/28:10:36:14.106
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                  CSI    0000015e@2014/11/28:10:36:14.428 CSI perf
trace:
CSIPERF:TXCOMMIT;375581
```

Figure 1.3 Example of machine-generated data

5.Graph-based or network data

"Graph data" can be a confusing term because any data can be shown in a graph. "Graph" in this case points to mathematical graph theory. In graph theory, a graph is a mathematical structure to model pair-wise relationships between objects. Graph or network data is, in short, data that focuses on the relationship or adjacency of objects. The graph structures use nodes, edges, and properties to represent and store graphical data. Graph-based data is a natural way to represent social networks, and its structure allows you to calculate specific metrics such as the influence of a person and the shortest path between two people.

Examples of graph-based data can be found on many social media websites (figure 1.4). For instance, on LinkedIn you can see who you know at which company. Your follower list on Twitter is another example of graph-based data. The power and sophistication comes from multiple,

overlapping graphs of the same nodes. For example, imagine the connecting edges here to show "friends" on Facebook. Imagine another graph with the same people which connects business colleagues via LinkedIn. Imagine a third graph based on movie interests on Netflix. Overlapping the three different-looking graphs makes more interesting questions possible.

Graph databases are used to store graph-based data and are queried with specialized query languages such as SPARQL. Graph data poses its challenges, but for a computer interpreting additive and image data, it can be even more difficult.

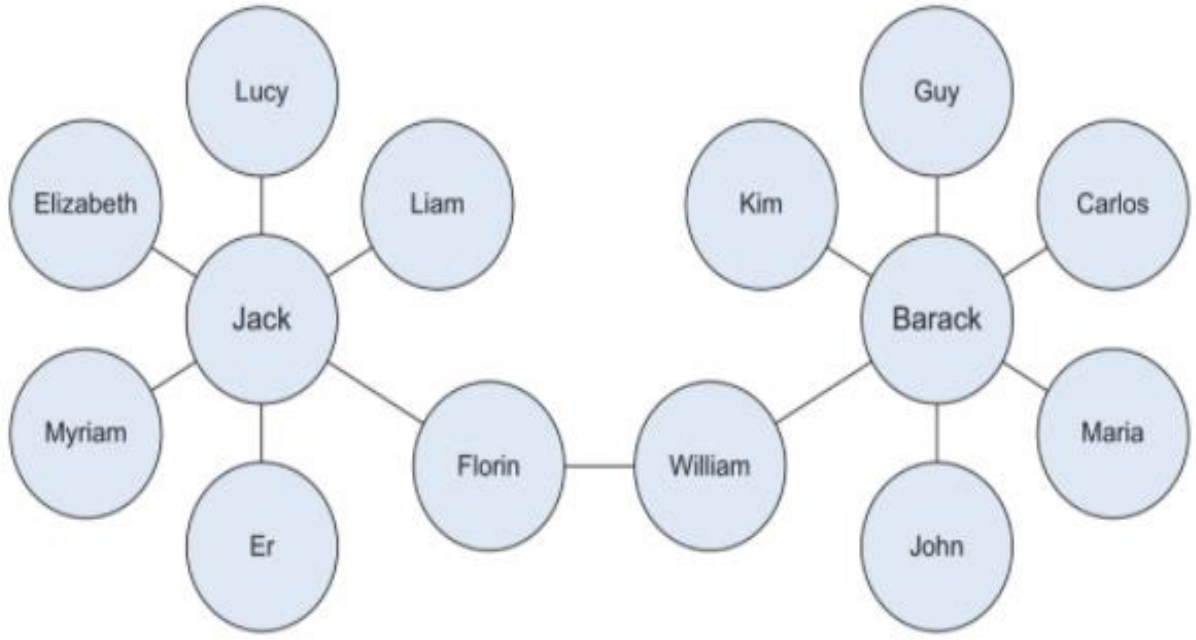


Figure 1.4 Friends in a social network are an example of graph-based data.

6. Audio, image, and video

Audio, image, and video are data types that pose specific challenges to a data scientist. Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers. MLBAM (Major League Baseball Advanced Media) announced in 2014 that they'll increase video capture to approximately 7 TB per game for the purpose

of live, in-game analytics. High-speed cameras at stadiums will capture ball and athlete movements to calculate in real time, for example, the path taken by a defender relative to two baselines.

Recently a company called DeepMind succeeded at creating an algorithm that's capable of learning how to play video games. This algorithm takes the video screen as input and learns to interpret everything via a complex process of deep learning. It's a remarkable feat that prompted Google to buy the company for their own Artificial Intelligence (AI) development plans. The learning algorithm takes in data as it's produced by the computer game; it's streaming data.

7. Streaming data

While streaming data can take almost any of the previous forms, it has an extra property. The data flows into the system when an event happens instead of being loaded into a data store in a batch. Although this isn't really a different type of data, we treat it here as such because you need to adapt your process to deal with this type of information.

Examples are the "What's trending" on Twitter, live sporting or music events, and the stock market.