



CHAPTER 1

INTRODUCTION TO DATA SCIENCE

Kwankamon Dittakan, Ph.D.
College of Computing
Prince of Songkla University
Phuket, Thailand

CONTENT

1. What is data science?
2. Why data science?
3. Data science vs Computer science
4. Properties of data
5. Categories of data
6. Data science process
7. Hierarchy of needs
8. Applications of data science

1. WHAT IS DATA SCIENCE?



1. WHAT IS DATA SCIENCE?

1. **Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.[Data science is related to data mining, machine learning and big data (Wikipedia).
2. **Data Science** is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. But how is this different from what statisticians have been doing for years? (edureka).
3. **Data science** combines multiple fields including statistics, scientific methods, and data analysis to extract value from data. Those who practice data science are called data scientists, and they combine a range of skills to analyze data collected from the web, smartphones, customers, sensors, and other source (Oracle).

1. WHAT IS DATA SCIENCE?

4. **Data science** continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand that they must advance past the traditional skills of analyzing large amounts of data, data mining, and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process (Berkeley).
5. **Data science** is an essential part of any industry today, given the massive amounts of data that are produced. Data science is one of the most debated topics in the industries these days. Its popularity has grown over the years, and companies have started implementing data science techniques to grow their business and increase customer satisfaction. In this article, we'll learn what data science is, and how you can become a data scientist (simplilearn).

2. WHY DATA SCIENCE?

Data science or data-driven science enables better decision making, predictive analysis, and pattern discovery:

- ❖ Find the leading cause of a problem by asking the right questions
- ❖ Perform exploratory study on the data
- ❖ Model the data using various algorithms
- ❖ Communicate and visualize the results via graphs, dashboards, etc.

EXAMPLE 1: AUTOMATIC IMAGE CAPTIONING



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



"girl in pink dress is jumping in air."



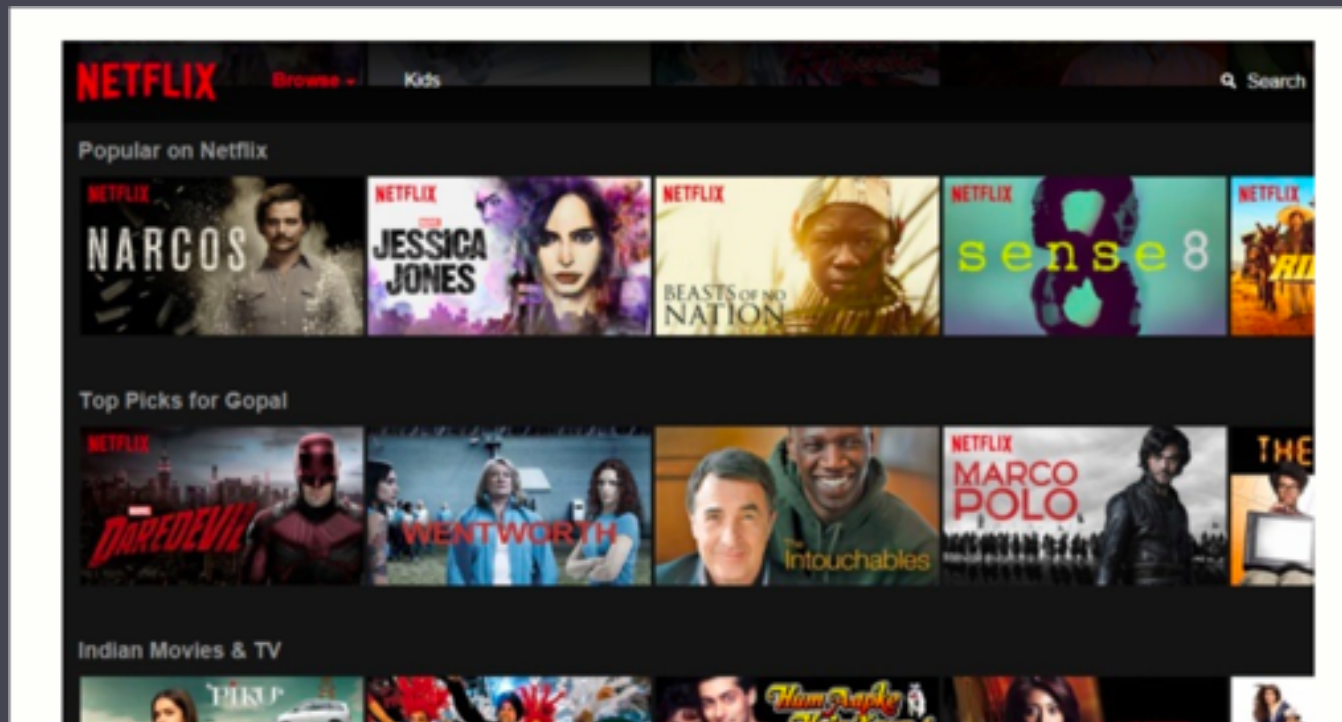
"black and white dog jumps over bar."



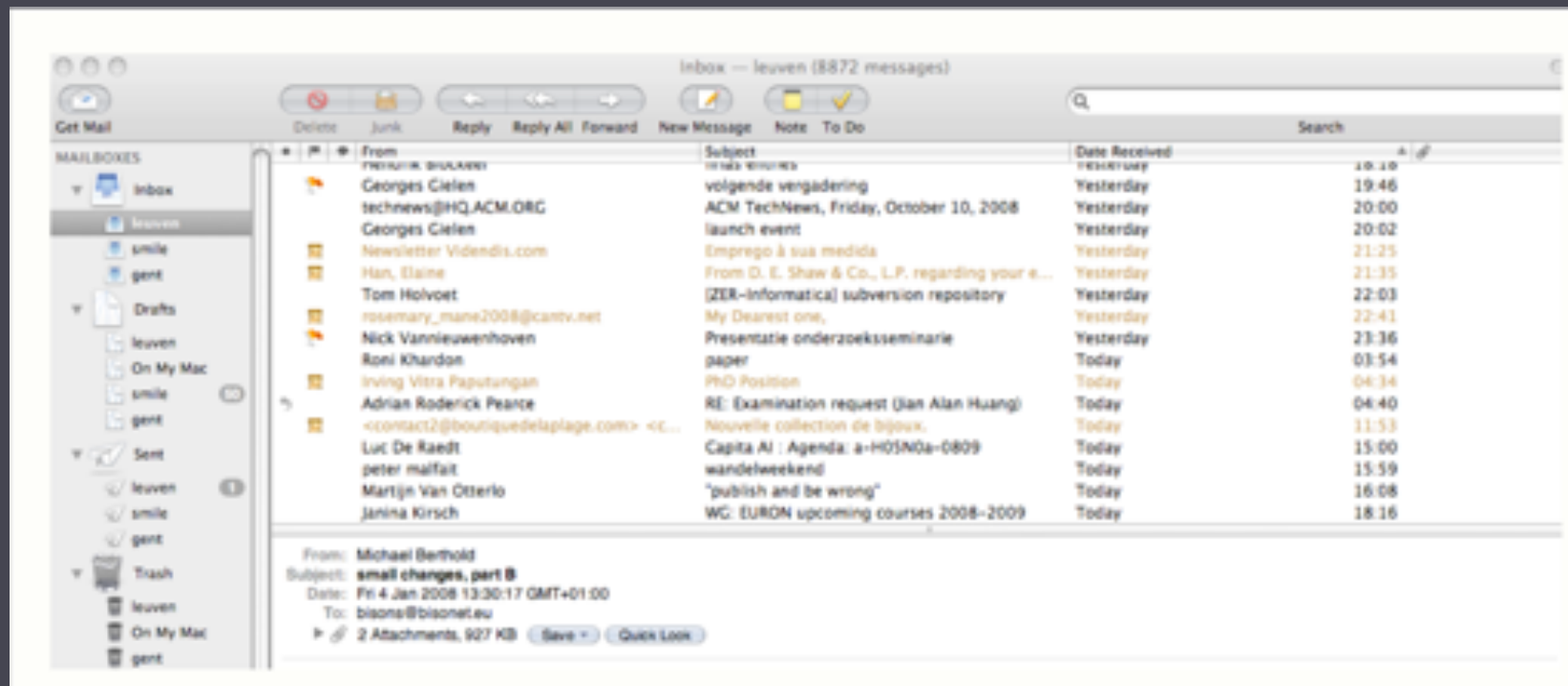
"young girl in pink shirt is swinging on swing."

Automatic Image Caption Generation
Sample taken from [Andrej Karpathy, Li Fei-Fei](#)

EXAMPLE 2: PRODUCT RECOMMENDATION



EXAMPLE 3: SPAM DETECTION



3. COMPUTER SCIENCE VS DATA SCIENCE

Computer science is the main branch whereas data science is a branch of computer science.

- ❖ *Computer science* deals with algorithms with more focus on software engineering and development. Computer Science is the study of the theory and practice of how computers work. When we earn a degree in Computer Science, we learn programming, software, operating systems, algorithms and everything needed to run a computer,
- ❖ *Data science* is an advanced discipline that teaches students how to analyze and find patterns in large amounts of data. In data science, data is gathered (or mined) and analyzed for any valuable insights, trends or patterns. Data Science degrees focus on mathematical concepts and understanding, such as calculus and statistics. Other subjects such as machine learning, deep learning, data visualization and databases are also covered.

3. COMPUTER SCIENCE VS DATA SCIENCE

Some typical computer science-related job duties include:

- ❖ Testing, documenting and debugging code
- ❖ Creating or modifying software and mobile apps
- ❖ Designing components of an application and integrating them into a larger overall product
- ❖ Collaborating with a team of programmers to build and optimize code

In general, computer science jobs revolve around building, modifying and digging into the inner workings of software applications.

3. COMPUTER SCIENCE VS DATA SCIENCE

Some typical data science-related job duties include:

- ❖ Collecting, “cleaning” and organizing data sets
- ❖ Building data models
- ❖ Asking and answering questions with large scale data analysis
- ❖ Creating data visualizations and presenting findings to stakeholders

Data science jobs are a bit more abstract—often their work revolves around attempting to improve a process or answer an unknown by pulling together huge amounts of information from multiple sources and analyzing it.

4. PROPERTIES OF DATA

1. Structured vs. Unstructured Data
2. Quantitative vs. Categorical Data
3. Big Data vs. Small Data

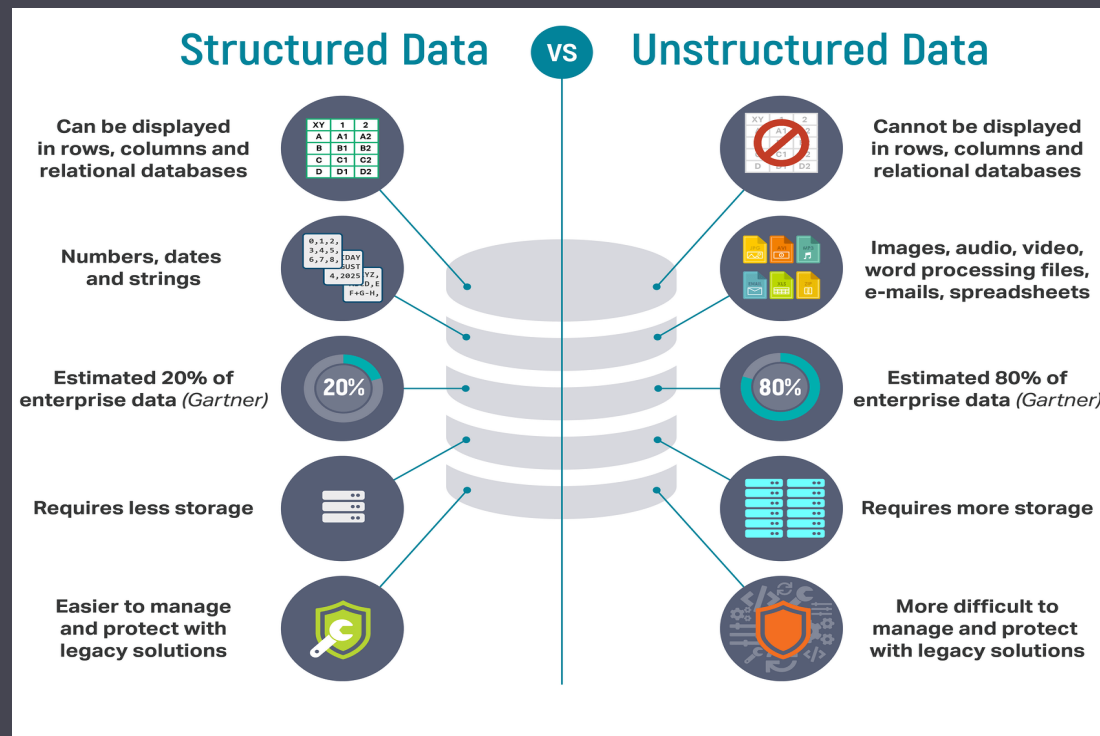
STRUCTURED VS. UNSTRUCTURED DATA

- ❖ *Structured data* is comprised of clearly defined data types whose pattern makes them easily searchable. Structured data, data is often represented by a matrix, where the rows of the matrix represent distinct items or records, and the columns represent distinct properties of these items such as the tables in a database or spreadsheet program
- ❖ *Unstructured data* – “everything else” – is comprised of data that is usually not as easily searchable, including formats like audio, video, and social media postings

STRUCTURED VS. UNSTRUCTURED DATA

- ❖ Structured data is clearly defined and searchable types of data, while unstructured data is usually stored in its native format.
- ❖ Structured data is quantitative, while unstructured data is qualitative.
- ❖ Structured data is often stored in database or data warehouses, while unstructured data is stored in data lakes.
- ❖ Structured data is easy to search and analyze, while unstructured data requires more work to process and understand.
- ❖ Structured data exists in predefined formats, while unstructured data is in a variety of formats.

STRUCTURED VS. UNSTRUCTURED DATA



QUANTITATIVE VS. CATEGORICAL DATA

- ❖ *Quantitative data* consists of numerical values, like height and weight. Such data can be incorporated directly into algebraic formulas and mathematical models, or displayed in conventional graphs and charts.
- ❖ *Categorical data* consists of labels describing the properties of the objects under investigation, like gender, hair colour, and occupation. This descriptive information can be every bit as precise and meaningful as numerical data, but it cannot be worked with using the same techniques.

BIG DATA VS. SMALL DATA

- ❖ *Small Data:* It can be defined as small datasets that are capable of impacting decisions in the present. Anything that is currently ongoing and whose data can be accumulated in an Excel file. Small Data is also helpful in making decisions, but does not aim to impact the business to a great extent, rather for a short span of time. Small data can be described as small datasets that are capable of having an influence on current decisions.
- ❖ *Big Data:* It can be represented as large chunks of structured and unstructured data. The amount of data stored is immense. It is therefore important for analysts to thoroughly dig the whole thing into making it relevant and useful to make proper business decisions. In short, datasets that are really huge and complex that conventional data processing techniques can not manage them are known as big data.

BIG DATA VS. SMALL DATA

Feature	Small Data	Big Data
Technology	Traditional	Modern
Collection	Generally, it is obtained in an organized manner than is inserted into the database	The Big Data collection is done by using pipelines having queues like AWS Kinesis or Google Pub / Sub to balance high-speed data
Volume	Data in the range of tens or hundreds of Gigabytes	Size of Data is more than Terabytes
Analysis Areas	Data marts(Analysts)	Clusters(Data Scientists), Data marts(Analysts)
Quality	Contains less noise as data is less collected in a controlled manner	Usually, the quality of data is not guaranteed
Processing	It requires batch-oriented processing pipelines	It has both batch and stream processing pipelines
Database	SQL	NoSQL
Velocity	A regulated and constant flow of data, data aggregation is slow	Data arrives at extremely high speeds, large volumes of data aggregation in a short time
Structure	Structured data in tabular format with fixed schema(Relational)	Numerous variety of data set including tabular data, text, audio, images, video, logs, JSON etc.(Non Relational)
Scalability	They are usually vertically scaled	They are mostly based on horizontally scaling architectures, which gives more versatility at a lower cost

BIG DATA VS. SMALL DATA

Feature	Small Data	Big Data
Query Language	only Sequel	Python, R, Java, Sequel
Hardware	A single server is sufficient	Requires more than one server
Value	Business Intelligence, analysis and reporting	Complex data mining techniques for pattern finding, recommendation, prediction etc.
Optimization	Data can be optimized manually(human powered)	Requires machine learning techniques for data optimization
Storage	Storage within enterprises, local servers etc.	Usually requires distributed storage systems on cloud or in external file systems
People	Data Analysts, Database Administrators and Data Engineers	Data Scientists, Data Analysts, Database Administrators and Data Engineers
Security	Security practices for Small Data include user privileges, data encryption, hashing, etc.	Securing Big Data systems are much more complicated. Best security practices include data encryption, cluster network isolation, strong access control protocols etc.
Nomenclature	Database, Data Warehouse, Data Mart	Data Lake
Infrastructure	Predictable resource allocation, mostly vertically scalable hardware.	More agile infrastructure with horizontally scalable hardware

5. CATEGORIES OF DATA

In data science and big data, the user may 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

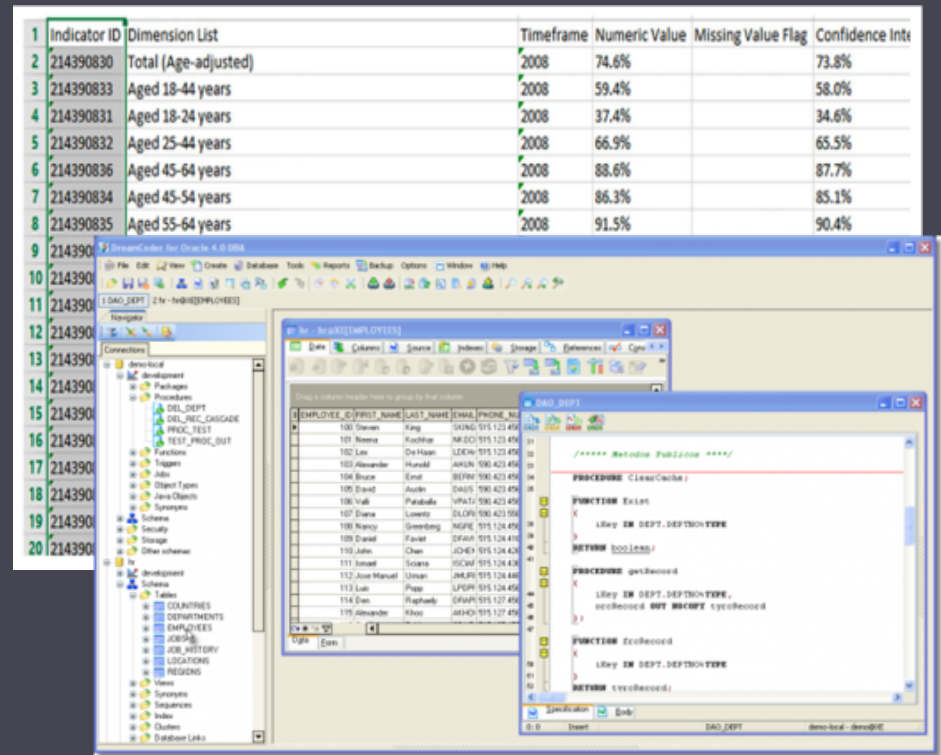
STRUCTURED DATA

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.

SQL, or Structured Query Language, is the preferred way to manage and query data that resides in databases.

User may also come across structured data that might give a hard time storing it in a traditional relational database. Hierarchical data such as a family tree is one such example.

Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Int
214390830	Total (Age-adjusted)	2008	74.6%		73.8%
214390833	Aged 18-44 years	2008	59.4%		58.0%
214390831	Aged 18-24 years	2008	37.4%		34.6%
214390832	Aged 25-44 years	2008	66.9%		65.5%
214390836	Aged 45-64 years	2008	88.6%		87.7%
214390834	Aged 45-54 years	2008	86.3%		85.1%
214390835	Aged 55-64 years	2008	91.5%		90.4%



The screenshot shows the Microsoft Access 2007 interface. The main window displays a table named 'EMPLOYEE' with the following data:

EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMBER
100	Steven	King	SKING	515 121 456
101	Neena	Kochhar	NKCH	515 122 456
102	Lex	DeHaan	LDEH	515 122 456
103	Alexander	Hunold	AHUN	508 422 456
104	Bruce	Ernst	BERN	508 422 456
105	Daniel	Farrel	DFAR	508 422 456
106	Valli	Pataballa	VPAT	508 422 456
107	Diana	Lorent	DLOR	508 422 456
108	Nancy	Greenberg	NGRE	515 124 456
109	Daniel	Forest	DFOR	515 124 456
110	John	Chen	JCH	515 124 456
111	Ismael	Scorcia	ISCOR	515 124 456
112	Jose Manuel	Uribe	JMUR	515 124 456
113	Luke	Platt	LPLA	515 124 456
114	Den	Raphaely	DRAP	515 127 456
115	Alexander	Kuhn	AKUH	515 127 456

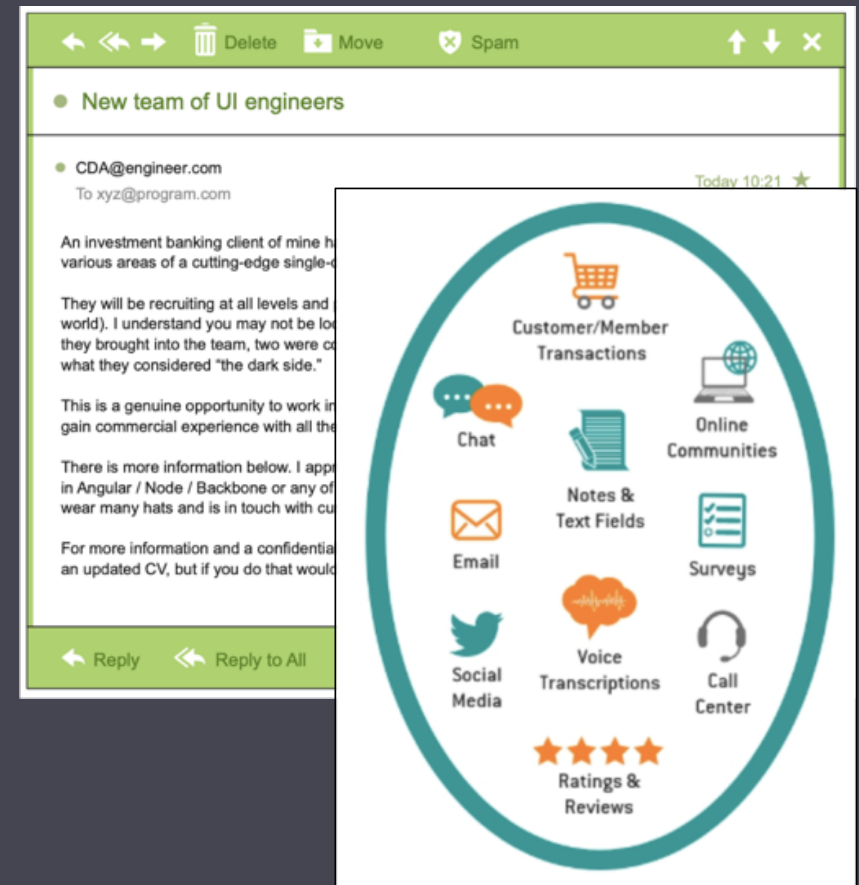
The query 'DAO_DEPT' contains the following SQL code:

```
FUNCTION getRecord  
    IF NOT ISNULL(DEPT.DEPTNAME)  
        RETURN DEPT.DEPTNAME  
    RETURN NULL  
END FUNCTION  
FUNCTION getRecord  
    IF NOT ISNULL(DEPT.DEPTNAME)  
        RETURN DEPT.DEPTNAME  
    RETURN NULL  
END FUNCTION
```


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. 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.



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.

MACHINE-GENERATED DATA

- ❖ Machine-generated data is information that's automatically created by a computer, process, application, or another machine without human intervention.
- ❖ Machine-generated data is becoming a major data resource and will continue to do so.
- ❖ 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.

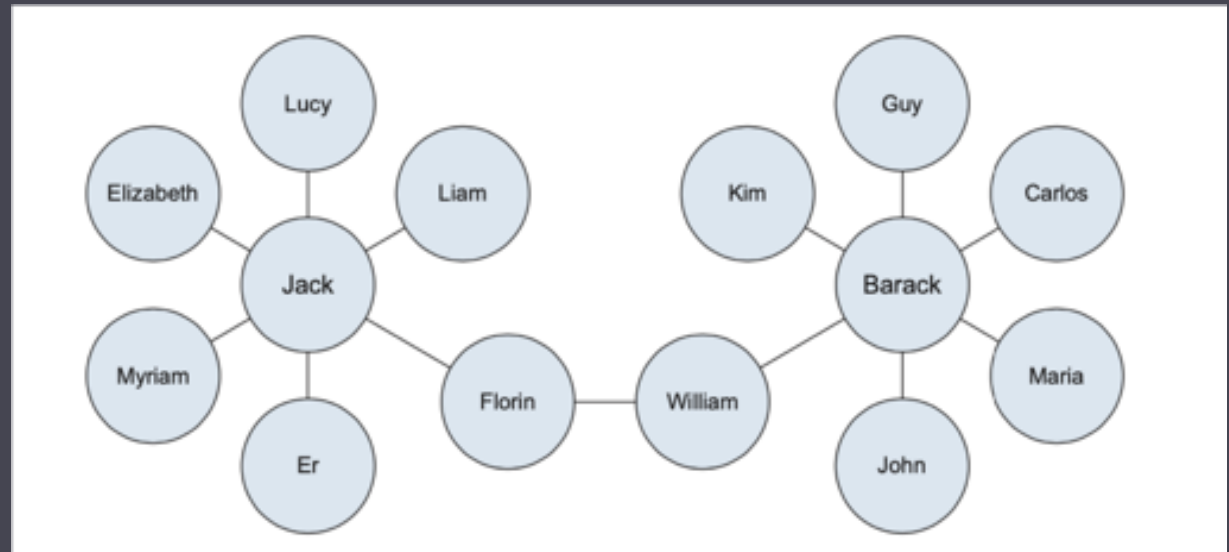
```
CSIPERF:TXCOMMIT:313236
2014-11-28 11:36:13, Info
2014-11-28 11:36:13, Info
result 0x00000000, handle 0x4e54
2014-11-28 11:36:13, Info
Beginning NT transaction commit...
2014-11-28 11:36:13, Info
trace:
CSIPERF:TXCOMMIT:273983
2014-11-28 11:36:13, Info
2014-11-28 11:36:13, Info
result 0x00000000,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

214.1.211.251 - - [15/Apr/2011:09:40:17 -0700] "GET /global.asa HTTP/1.0" 404 315 "-" "Mozilla/5.0"
214.1.211.251 - - [15/Apr/2011:09:40:17 -0700] "GET /-root HTTP/1.0" 404 318 "-" "Mozilla/5.0"
214.1.211.251 - - [15/Apr/2011:09:40:18 -0700] "GET /-apache HTTP/1.0" 404 312 "-" "Mozilla/5.0"
219.167.17.173 - - [17/Apr/2011:17:55:40 -0700] "POST /sony/mnr HTTP/1.1" 200 130 "-" "PS3"
218.41.54.67 - - [17/Apr/2011:18:20:18 -0700] "POST /sony/mnr HTTP/1.1" 200 130 "-" "PS3"
10.132.93.114 - - [18/Apr/2011:11:05:39 -0700] "POST /sony/mnr HTTP/1.1" 200 61 "-" "Ledu"
10.132.93.114 - - [18/Apr/2011:11:07:07 -0700] "POST /sony/mnr HTTP/1.1" 200 61 "-" "Ledu"
10.132.93.114 - - [18/Apr/2011:11:13:52 -0700] "POST /sony/mnr HTTP/1.1" 200 61 "-" "Ledu"
218.41.54.67 - - [20/Apr/2011:17:42:37 -0700] "POST /sony/mnr HTTP/1.1" 200 100 "-" "PS3"
60.34.131.229 - - [20/Apr/2011:18:22:32 -0700] "POST /sony/mnr HTTP/1.1" 200 100 "-" "PS3"
202.213.251.245 - - [21/Apr/2011:21:16:45 -0700] "POST /sony/mnr HTTP/1.1" 200 100 "-" "PS3"
202.213.251.245 - - [21/Apr/2011:21:24:43 -0700] "POST /sony/mnr HTTP/1.1" 200 100 "-" "PS3"
178.202.110.92 - - [22/Apr/2011:18:59:05 -0700] "GET / HTTP/1.1" 200 315 "-" "Mozilla/5.0"
178.202.110.92 - - [22/Apr/2011:18:59:05 -0700] "GET /favicon.ico HTTP/1.1" 404 333 "-" "Mozilla/5.0"
178.202.110.92 - - [22/Apr/2011:18:59:05 -0700] "GET /favicon.ico HTTP/1.1" 404 333 "-" "Mozilla/5.0"
178.202.110.92 - - [22/Apr/2011:18:59:07 -0700] "GET /access-navigator-media HTTP/1.1" 200 200 "-" "Mozilla/5.0"
178.202.110.92 - - [22/Apr/2011:19:05:00 -0700] "GET /admin/cdr/counter.txt HTTP/1.1" 404 315 "-" "Mozilla/5.0"
178.202.110.92 - - [22/Apr/2011:19:05:41 -0700] "GET //help/readme.nsf?OpenAbout HTTP/1.1" 404 315 "-" "Mozilla/5.0"
178.202.110.92 - - [22/Apr/2011:19:05:54 -0700] "GET /catinfo?A HTTP/1.1" 404 329 "-" "Mozilla/5.0"
178.202.110.92 - - [22/Apr/2011:19:06:08 -0700] "GET /errors-navigator-media HTTP/1.1" 200 200 "-" "Mozilla/5.0"
178.202.110.92 - - [22/Apr/2011:19:27:04 -0700] "GET / HTTP/1.1" 200 315 "-" "Mozilla/5.0"
```

GRAPH DATA

- ❖ “Graph data” 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 users to calculate specific metrics such as the influence of a person and the shortest path between two people.

GRAPH DATA



- ❖ Examples of graph-based data can be found on many social media websites.
- ❖ 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 come 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.

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.

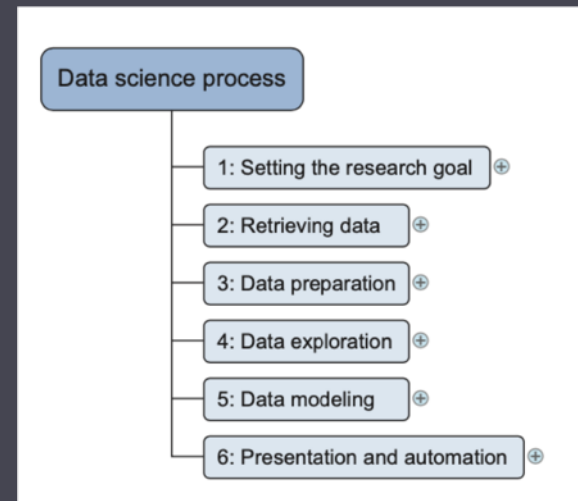
STREAMING DATA

- ❖ 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.

6. DATA SCIENCE PROCESS

The data science process typically consists of six steps:

1. Setting the research goal
2. Retrieving data
3. Data preparation
4. Data exploration
5. Data modeling
6. Presentation and automation



DATA SCIENCE PROCESS

1. **Setting the research goal:** prepare a project objective.

Example: what user is going to research, how the company benefits from that, what data and resources needs, a timetable, and deliverables.

2. **Retrieving data:** collect the data according to the project objective.

3. **Data preparation:** an error-prone process

(i) data cleansing (ii) data integration, and (iii) data transformation.

DATA SCIENCE PROCESS

4. **Data exploration:** building a deeper understanding of the data.

The descriptive statistics may be applied, visual techniques, and simple modelling. This step often goes by the abbreviation EDA, for Exploratory Data Analysis.

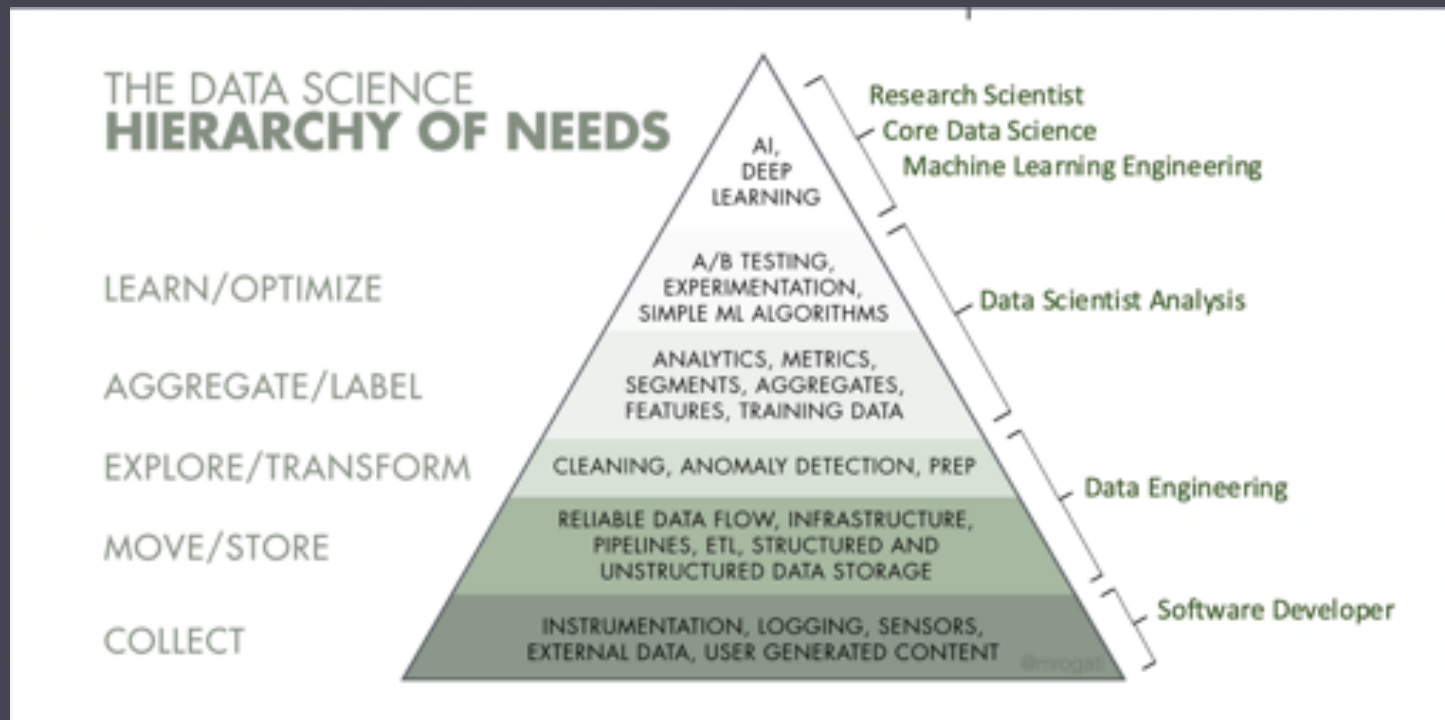
5. **Data modelling :** uses models, domain knowledge, and insights about the data founded in the previous steps to answer the research question.

Statistics, machine learning, operations research, and so on.

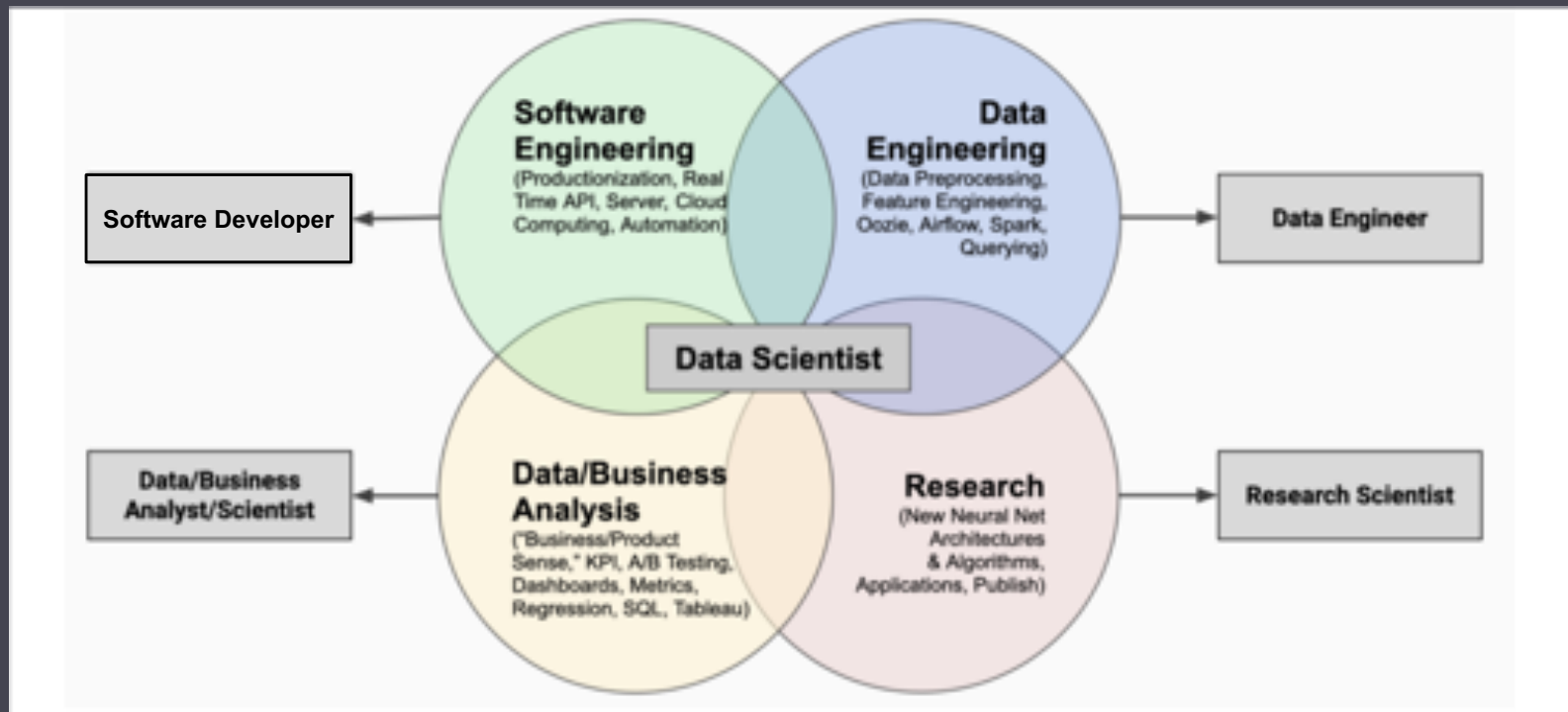
6. **Presentation and automation:** present the results to the business.

These results can take many forms, ranging from presentations to research reports.

7. HIERARCHY OF NEEDS



DIFFERENT DATA SCIENTISTS



8. APPLICATIONS OF DATA SCIENCE

Data science has found its applications in almost every industry.

- ❖ **Healthcare:** Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.
- ❖ **Gaming:** Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.
- ❖ **Image Recognition:** Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.
- ❖ **Recommendation Systems:** Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms.
- ❖ **Logistics:** Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.
- ❖ **Fraud Detection:** Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.