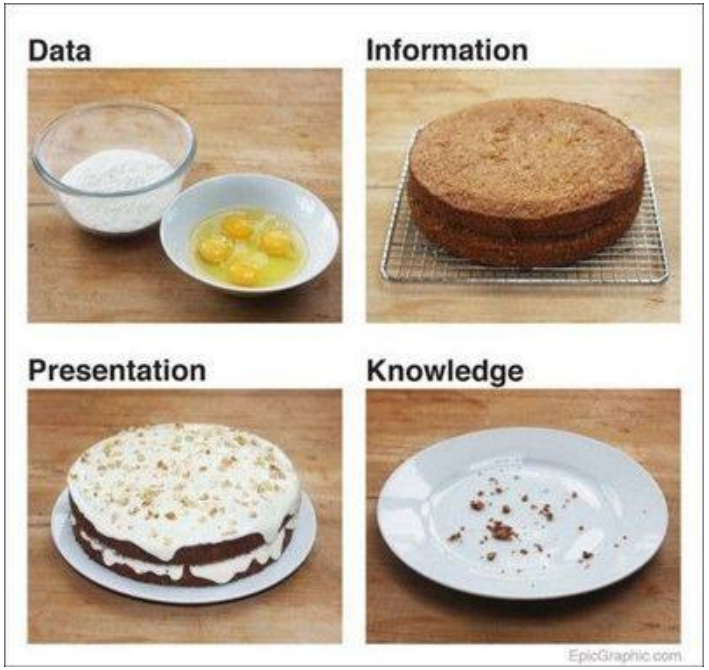
DI501 Introduction to Data Informatics

- Course instructor: Assoc.Prof.Dr. Tugba Taskaya Temizel
  - E-mail: ttemizel@metu.edu.tr
- Teaching assistant: Mert Mecit mert.mecit@metu.edu.tr
- Background Requirements:
  - Basic programming skills.
- Python will be used.
  - Online tutorials will be provided.
- All the reading materials and lecture notes will be available in ODTUClass.
- Please attend actively to the forum discussions!
- Assessment: Assignments and Quizzes (40%), Final Exam (30%), Final Project (30%)
- The course will be conducted online.
- Almost in every two weeks, there will be an online quiz during the official lecture hours: Mondays (9:40-12:30)
- All your quiz grades will be considered as one single assignment.

COURSE OUTLINE

- Week 1: Introduction to Data Science: Basic Concepts
- Week 2: Platforms & Workbenches
- Week 3: Understanding Data: Exploratory Data Analysis
- Week 4: Understanding Data: Descriptive Analytics & Visualization
- Week 5: Data Acquisition and Preprocessing: Part I
- Week 6: Data Acquisition and Preprocessing: Part II
- Week 7: Practical Machine Learning: Part I
- Week 8: Practical Machine Learning (ML): Part II
- Week 9: Model Evaluation and Performance Metrics
- Week 10: Data Storytelling
- Week 11: Current Topics in Data Science: Big Data, ML Advance Topics & Data Mining
- Week 12: Data Driven Organizations, Ethics and Legal Issues in Data Science



DI501

Introduction to Data Informatics

Lecture 1

Data Informatics

Terms

- Data analytics refers to analysis of the data in some way using quantitative and qualitative techniques to be able to explore for trends and patterns in the data.
- Informatics is defined as a collaborative activity that involves *people*, *processes*, and *technologies* to produce and use trusted data for better decision making.

https://bok.ahima.org/doc?oid=302313#.YVrkbZpBwuU

What is Data Science?

- The Science [and Art] of …
  - Discovering what we do not know from data
  - Obtaining predictive, actionable insight from data
  - Creating data products that have business impact now
  - Communicating relevant business stories from data
  - Building confidence in decisions that drive business value

A good example for creative data analysis is Lipton Ice Tea Sociology Project- Felis 2017 Creative Use of Data Winner: https://www.youtube.com/watch?v=bdiLYiWdh-8

What is Data Science?

- Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from data
  - usually from large data sets
- Example patterns:
  - groups of customers exhibiting similar behavior and tastes
  - products frequently bought together
  - strange or abnormal events
  - models that can be used for prediction/classification
  - …

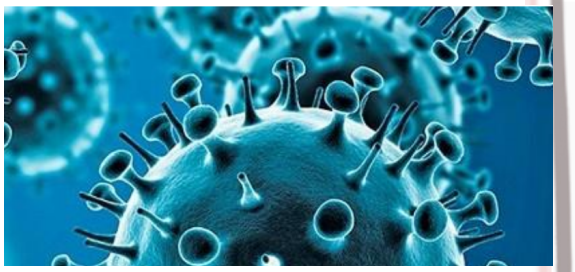
Non-obvious Patterns

- It is generally not worth the time and effort of using data science to discover patterns that can be created easily by human experts in their minds
- We need use data science when there is large amount of data and patterns are complex to discover and extract manually
  - Humans can define rules that check up to two, three features
  - They start to struggle to handle the interactions between too many features
- Data science is often applied to look for patterns among tens, hundreds, thousands, and, in extreme cases, millions of features

Useful Patterns

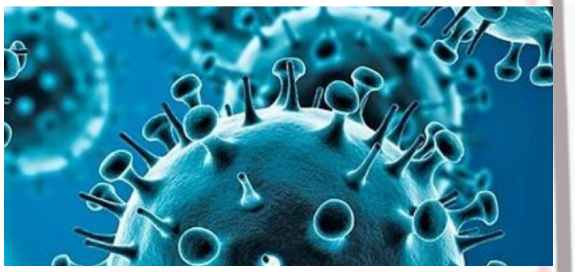
- A useful pattern gives us insight into the problem that enables us to do something to help solve a problem
- What we want the extracted patterns to give us is actionable insight
  - “insight” term emphasizes that the pattern should give us relevant information about the problem that isn’t obvious
  - “actionable” term emphasizes that the insight we get should also be something that we have the capacity to use in some way

Example #1: Retail Store



- Another outbreak for infectious disease like Covid!
- What should retail store executives do?
  - Will there be unusual local demand for products?
  - They might be able to anticipate unusual demand for products and rush stock to the stores ahead of any potential outbreak.
  - Can you think of examples for information that is obvious and/ or not useful?
    - More toilet papers– Obvious. Would we need data science to discover this?
    - It would be more valuable to discover patterns due to the outbreak that were not obvious.
  - What had happened when Covid-19 outbreak struck?

Example #1: Retail Store



- Non obvious items at the beginning of the first outbreak: cologne, masks (even different types of masks), gloves, face masks
- Even more non obvious items: bread makers, kitchen utensils, puzzles, books (due to a possible lockdown)
- These are also useful patterns, which can be used for taking actions.
  - Even the quantities matter!
- Not useful pattern:
  - People don't visit markets after 21:00.
    - Due to lockdown, it is not allowed. You cannot do anything to improve your customer numbers!

Example #2: Telecommunication

- Customers switching from one company to another is called churn.
- Many cell phone customers leave when their contracts expire and it is getting increasingly difficult to acquire new customers.
  - Since the cell phone market is now saturated and companies try to attract each other's customers while retaining their own.
- Churn is expensive for companies!
  - One company must spend on incentives to attract a customer while another company loses revenue when the customer departs.
  - Attracting new customers is much more expensive than retaining existing ones!!
  - Preventing churn is important but how to use the budget effectively?
  - Predicting Customer Churn may help to decide which customers should be offered the special retention deal prior to the expiration of their contracts.

DIKW Pyramid

- To elaborate on the concept, listen to Jennifer Rowley (2007):
  - "Typically information is defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge."
- DIKW: Data, Information, Knowledge, Wisdom

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of information science*, *33*(2), 163-180.

DIKW Pyramid

Data

- Data is just a set of signals or symbols created through some abstractions or measurements
- Data is unprocessed facts and figures without any added interpretation or analysis.
  - "The list of students”
  - Example: server logs, user behaviour events, or any other data set.
- It’s unorganized and unprocessed. It’s inert.
- If we don’t know what it means, it’s useless.
- **Etymology**: "Data" comes from a singular Latin word, *datum*, which originally meant "something given." Its early usage dates back to the 1600s. Over time "data" has become the plural of *datum*.

DIKW Pyramid

Information

- Information is data that have been processed, structured, or contextualized so that it is meaningful to humans
- Information is data that has been interpreted so that it has meaning for the user.
  - "The list of students taking DI501 course"
  - "The grades of students taking DI501 course"
  - A list of dates — data — is meaningless without the information that makes the dates relevant (dates of holiday or birthday).
  - The history of temperature readings all over the world for the past 100 years is data. If this data is organized and analyzed to find that global temperature is rising, then that is information.
  - The number of visitors to a website by country is an example of data.
    - Finding out that traffic from the U.S. is increasing while that from Australia is decreasing is meaningful information.

DIKW Pyramid

Information

- **Etymology**: "Information" is an older word that dates back to the 1300s and has Old French and Middle English origins.
  - It has always referred to "the act of informing," usually in regard to education, instruction, or other knowledge communication.
- Information is data with meaning.

DIKW Pyramid

Knowledge

- Knowledge is a combination of information, experience and insight that may benefit the individual or the organization.
  - "If the student takes higher than 85 from both midterm and the assignments, it is highly likely that the student will get AA from this course at the end"
- Knowledge is a mental structure, made from accumulated learning and systematic analysis of Information

DIKW Pyramid

Knowledge

- Explicit knowledge:
  - can be easily passed on to others.
  - Most forms of explicit knowledge can be stored in certain media such as in encyclopedias and textbooks.
- Tacit knowledge:
  - difficult to pass on to another person just by writing it down such as ability to speak a language, bake bread, program a computer or use complicated machinery requires additional pieces of knowledge (such as that gained through experience)

DIKW Pyramid

Wisdom

- Although commonly included as a level in DIKW, "there is limited reference to wisdom" in discussions of the model.
- Cleveland described wisdom simply as "*integrated knowledge—information made super-useful*".
- Other authors have characterized wisdom as "*knowing the right things to do*" and "*the ability to make sound judgments and decisions apparently without thought*.

Cleveland, Harlan (December 1982). "Information as a Resource". The Futurist: 34–39.

DIKW Pyramid

A Dummy Example

- **Data**: Temperature, humidity level and rainfall intensity measurements over one-minute time intervals
- **Information**: The temperature dropped 5 degrees, the humidity went up by 5% in one hour and then it started raining at 3 pm.
- **Knowledge**: A quick increase in the humidity, accompanied by a temperature drop caused by lower pressure areas, will likely make the atmosphere unable to hold the moisture and rain.
- **Wisdom**: Based on the observations and maths model, we can predict why and when it will rain in the future, and we can do it so fast and systematically that it won't require a lot of analysis. We already have an understanding of all the interactions that happen between evaporation, air currents, temperature gradients, changes, and raining.

Ref: https://towardsdatascience.com/rootstrap-dikw-model-32cef9ae6dfb

DIKW Pyramid

A flow diagram of the DIKW hierarchy

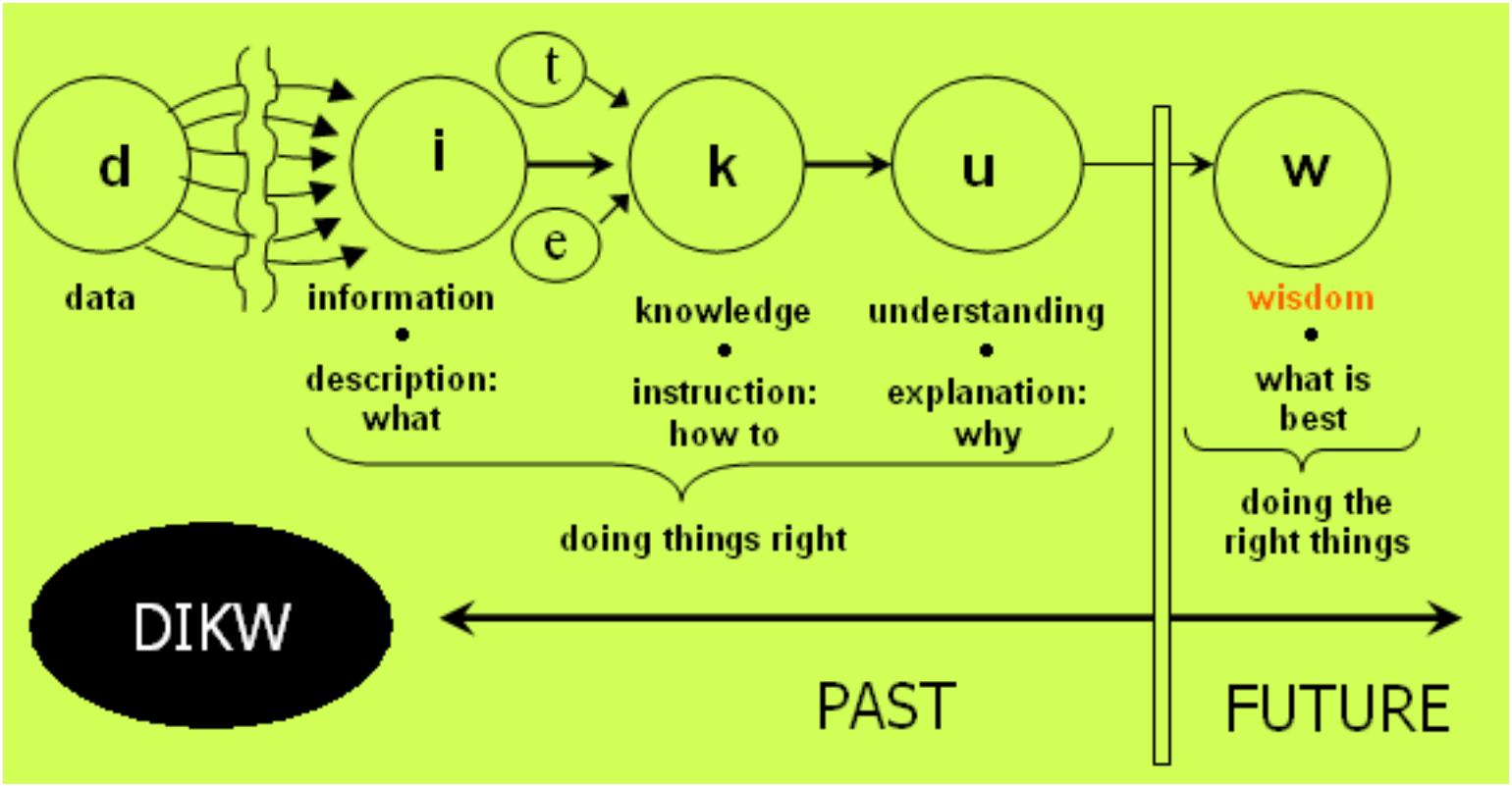


Image courtesy: https://en.m.wikipedia.org/wiki/File:DIKW_(1).png

Some Application Areas of Data Science

- Marketing: targeted marketing, online advertising, recommendations for cross-selling
- Retailers: increasing sales, price optimization, supply-chain management
- Customer Relationship Management: analysis of customer behavior to retain existing customers and maximize expected customer value
- Finance: credit scoring, fraud detection
- Manufacturing: supply-chain management, increased productivity, predictive maintenance
- …

Data Science Fueled Enterprises

- Google: Search Engine, Google Ads, YouTube, …
  - Everything they do is data driven
- Amazon: Selling products
  - Increase sales by product recommendations
- Facebook: Facebook, Instagram, WhatsApp
  - Companies love using Facebook as an advertising medium because they know a lot about you
- …

Data Driven Decision Making

- From the business point of view, the ultimate goal of data science can be seen as improving decision making
- Data driven decision making (DDD) refers to the practice of basing decisions on the analysis of data, rather than purely on intuition
  - Decisions for which “discoveries” need to be made within data
  - Decisions that repeat, especially at massive scale, and so decision-making can benefit from even small increases in decision-making accuracy based on data analysis

Data Driven Decision Making

- The more data-driven a firm is, the more productive it is
- There is also a causal correlation between DDD and return on assets, return on equity, asset utilization, market value, etc.

Data Science for Businesses

- Today, vast amounts of data available and companies in almost every industry are focused on exploiting data for competitive advantage
  - Manual analysis is not feasible due to the volume and variety of data
- Computers have become far more powerful, networking has become ubiquitous, and we have algorithms for broader and deeper analyses
- These have given rise to the increasingly widespread business application of data science
- if your competitors are relying on data-driven decision-making and you aren't, they will surpass you and steal your market share
  - You will simply be forced out of business

Altavista vs. Google Example

- AltaVista was a Web search engine established in 1995
- It became one of the most-used early search engines, but lost ground to Google and was purchased by Yahoo! in 2003
- On July 8, 2013, the service was shut down by Yahoo! and since then the domain has redirected to Yahoo!'s own search site.

Altavista vs. Google Example

- AltaVista solution:
  - It sent "crawlers" to extract the text from all the pages on the web.
  - The crawlers brought the text back to AltaVista.
  - AltaVista indexed all the text.
  - AltaVista then presented the results as an ordered list of web pages, with the pages that had the most frequent mentions of the term at the top.
  - This is a straightforward computer science solution, though at the time, they solved some very difficult scaling problems.

Altavista vs. Google Example

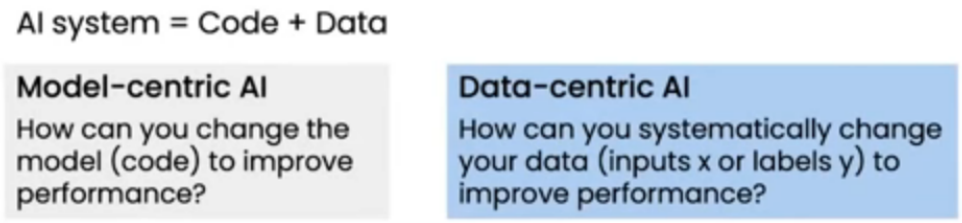
- Google solution:
  - In the late 1990s the founders of Google invented a different way to do search.
  - They combined math, statistics, data engineering, advanced computation, and the hacker spirit to create a search engine that displaced AltaVista.
  - The algorithm is known as PageRank.
  - PageRank looks not only at the words on the page but the hyperlinks as well.
  - PageRank assumes that an inbound hyperlink is an indicator that some other person thought the current page was important enough to put a link to it on their own page.

Data and Data Science Capability as Strategic Assets

- Data and the capability to extract useful knowledge from data should be regarded as key strategic assets
- As with all assets, it is often necessary to make investments
  - The best data science team can yield little value without the appropriate data
  - The right data often cannot substantially improve decisions without suitable data science talent
- Often, we don't have exactly the right data to best make decisions and/or the right talent to best support making decisions from the data

Data and Data Science Capability as Strategic Assets

- Popular terms in 2021:
  - Data-centric AI: There is a current trend shifting from model centric AI to data centric AI.

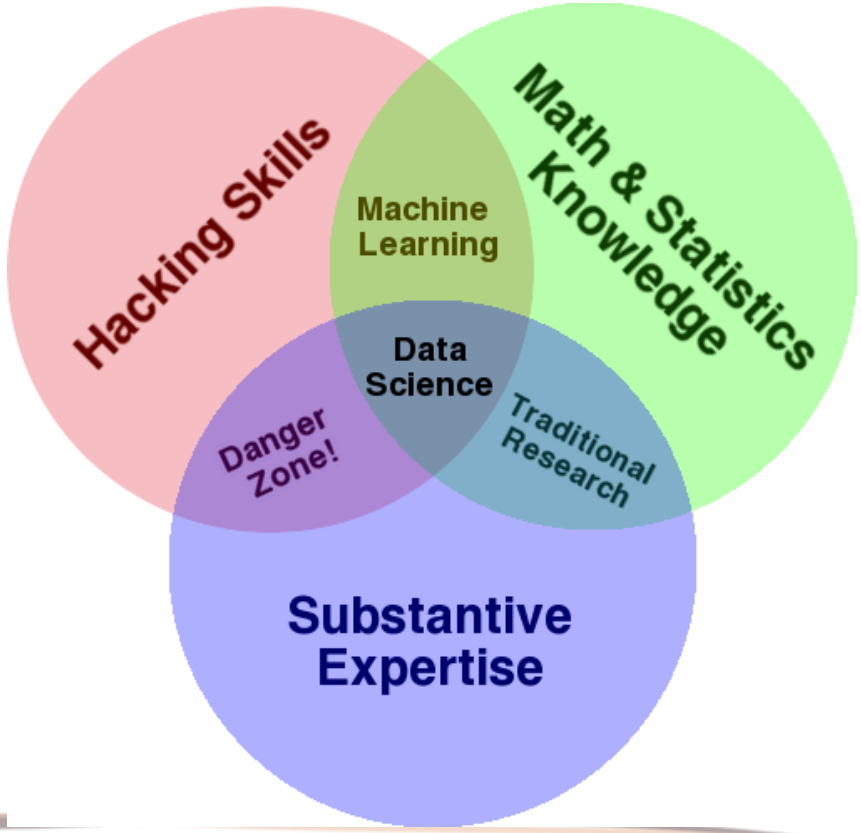


    - Attempt to improve your datasets by changing/enhancing them.
  - DataOps:
    - applies to the entire data lifecycle from data preparation to reporting, and recognizes the interconnected nature of the data analytics team and information technology operations.

Ref: A Chat with Andrew on MLOps: From Model-centric to Data-centric AI

Interdisciplinary Nature of Data Science

- Drew Conway's Venn diagram of data science from 2010



Ref: http://www.drewconway.com/zia/?p=2378

Interdisciplinary Nature

Hacking Skills are necessary for working with massive amount of electronic data that must be acquired, cleared and manipulated.

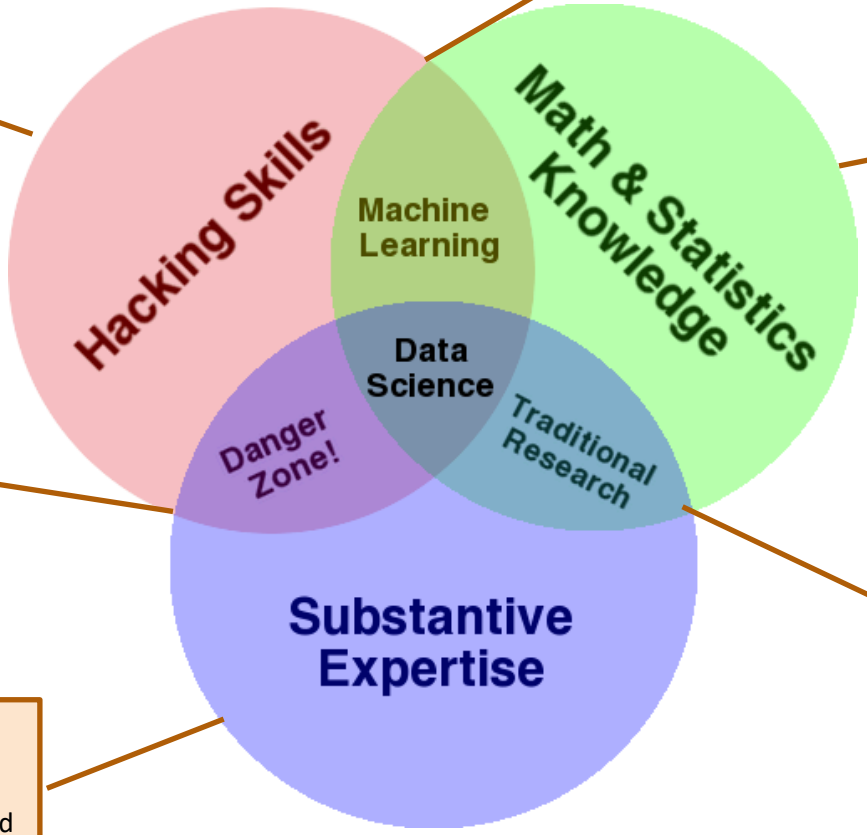
Machine learning stems from combining skills with math and statistics knowledge. It requires tuning parameters and model selection.

Math and Statistics Knowledge is fundamental in order to extract insight from data properly. It is the basis of all the methods.

Danger zone: Hacking skills combined with substantive expertise without rigorous method can result in incorrect analysis.

Traditional Research requires one to know the domain related traditional research methods (statistics and math knowledge).

Substantive Expertise in a scientific field is crucial for generating motivating question and hypotheses and interpreting result.



Ref: https://www.datasciencecentral.com/profiles/blogs/the-story-of-big-data-data-science-amp-data-mining

Example

Computational social science uses large-scale demographic, behavioural and network data to investigate human activity and relationships

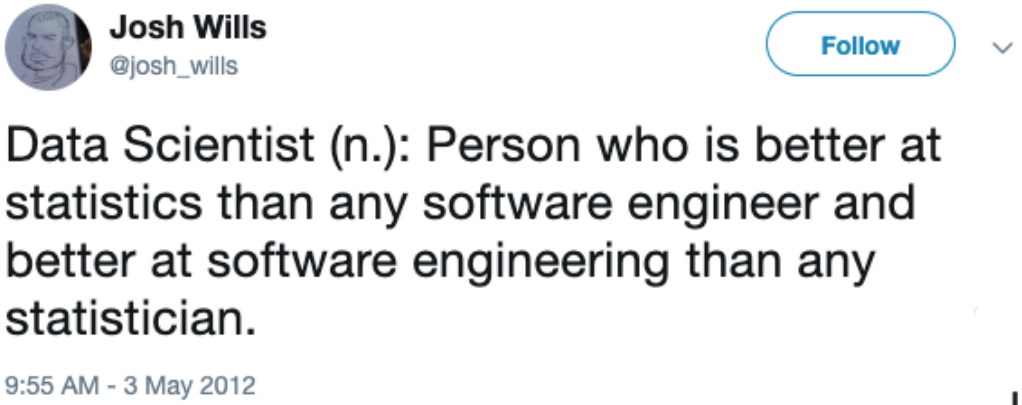
A data scientist working in this domain needs to have

- substantive expertise in the domain: needs to understand individual and collective human behaviour
- knowledge and expertise on math&statistics, machine learning
- knows how to use these models effectively (tuning parameters, model selection etc.)

Personality prediction using Facebook likes

Ref: http://www.drewconway.com/zia/?p=2378

What Data Scientists Do…



What Data Scientists Do…

- *A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the **big data lifecycle**, till the delivery of an **expected scientific** and **business value** to science or industry*

NIST SP 1500-1 NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions, September 2015 [online] http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

Others Skills?

- Data science not just knowing some programming languages, math, statistics and have “domain knowledge”.
- […] The *resolution to Business / Organizations problems through mathematics, programming and the scientific method* that involves the creation of hypotheses, experiments and tests through the analysis of data and the generation of predictive models.
- It is responsible for *transforming these problems into well-posed questions* that can also respond to the initial hypothesis in a creative way.
- It must also include the *effective communication* of the results obtained and *how the solution adds value to the Business / Organization.*

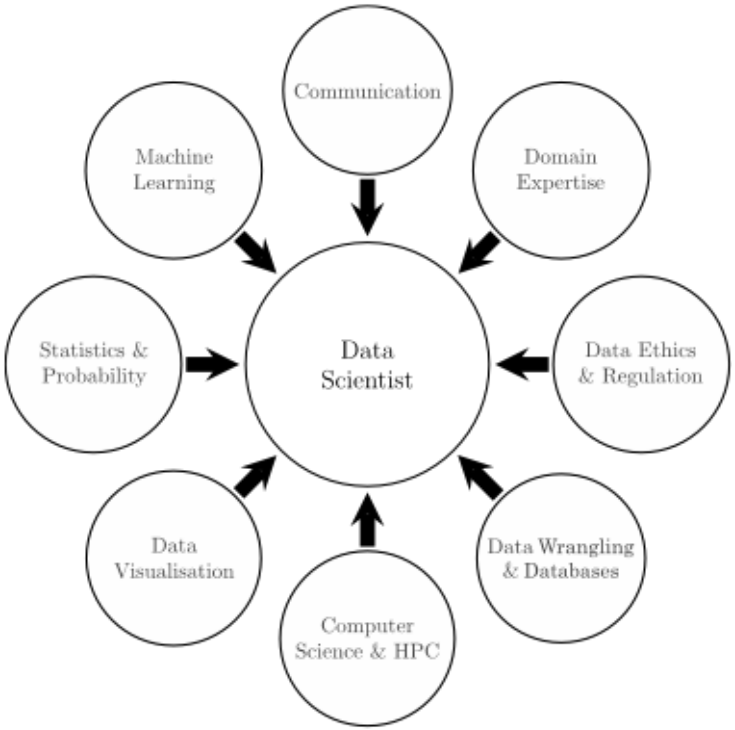
What Data Scientists Do…

- Depends on the level of seniority and the industry in particular:
  - Internet companies, manufacturing companies etc.
- A data scientist
  - is someone who knows how to extract meaning from and interpret data, which requires both tools and methods from statistics and machine learning
  - spends a lot of time in the process of collecting, cleaning, and munging data (data wrangling), because data is never clean.
    - This process requires persistence, statistics, and software engineering skills—skills that are also necessary for understanding biases in the data, and for debugging logging output from code

What Data Scientists Do…

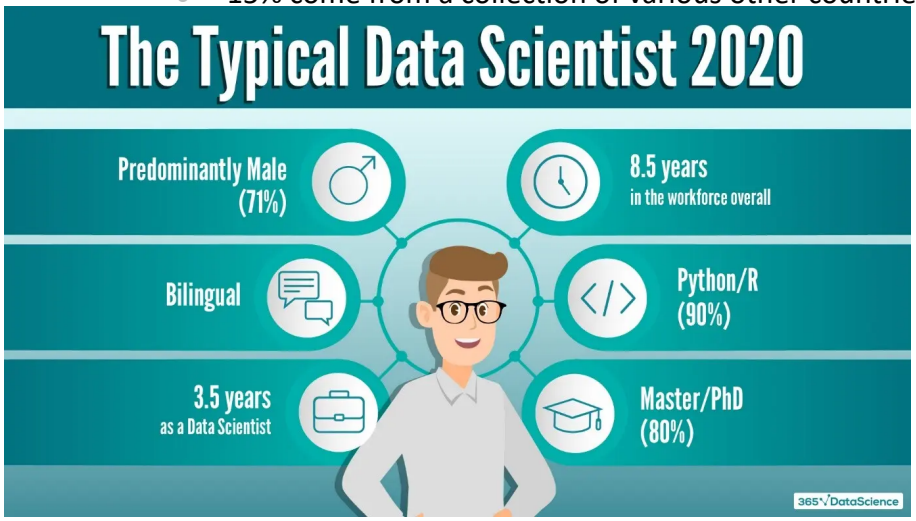
- A chief data scientist should be
  - setting the data strategy of the company
    - setting everything up from the engineering and infrastructure for collecting data and logging, to privacy concerns, to deciding what data will be user-facing, how data is going to be used to make decisions, and how it's going to be built back into the product.
  - managing a team of engineers, scientists, and analysts
  - communicating with leadership across the company, including the CEO, CTO, and product leadership
  - concerning with patenting innovative solutions and setting research goals

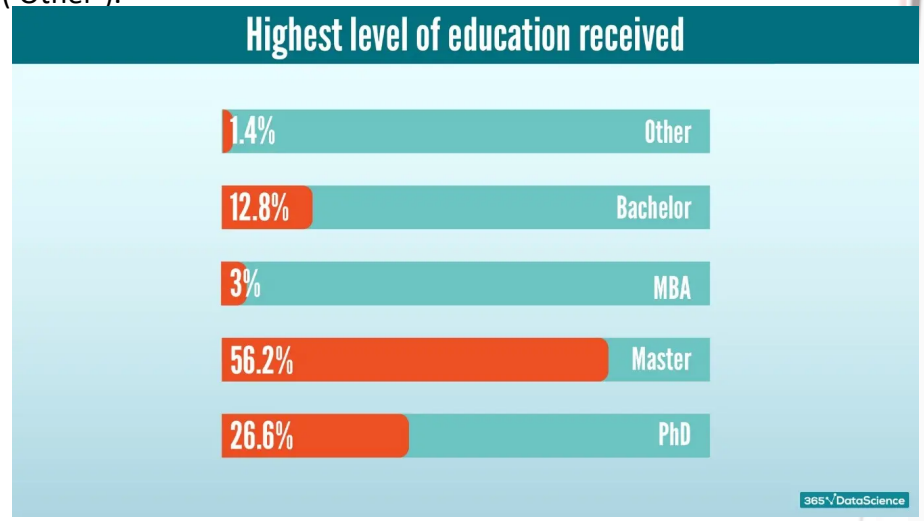
Essential Skillset for Data Scientists



What Data Scientists Do…

- The data for this report is based on the publicly available information in the LinkedIn profiles of 1,001 professionals, currently employed as data scientists. The sample includes junior, experts, and senior data scientists.
- Location
  - 40% of the data comprises data scientists currently employed in the United States;
  - 30% are data scientists in the UK;
  - 15% are currently in India;
  - 15% come from a collection of various other countries ('Other').

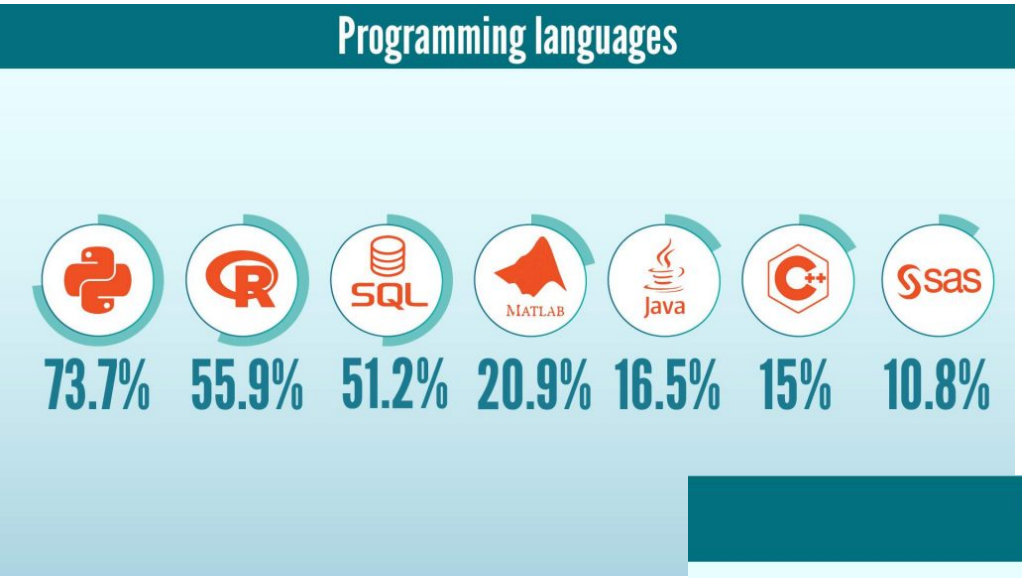


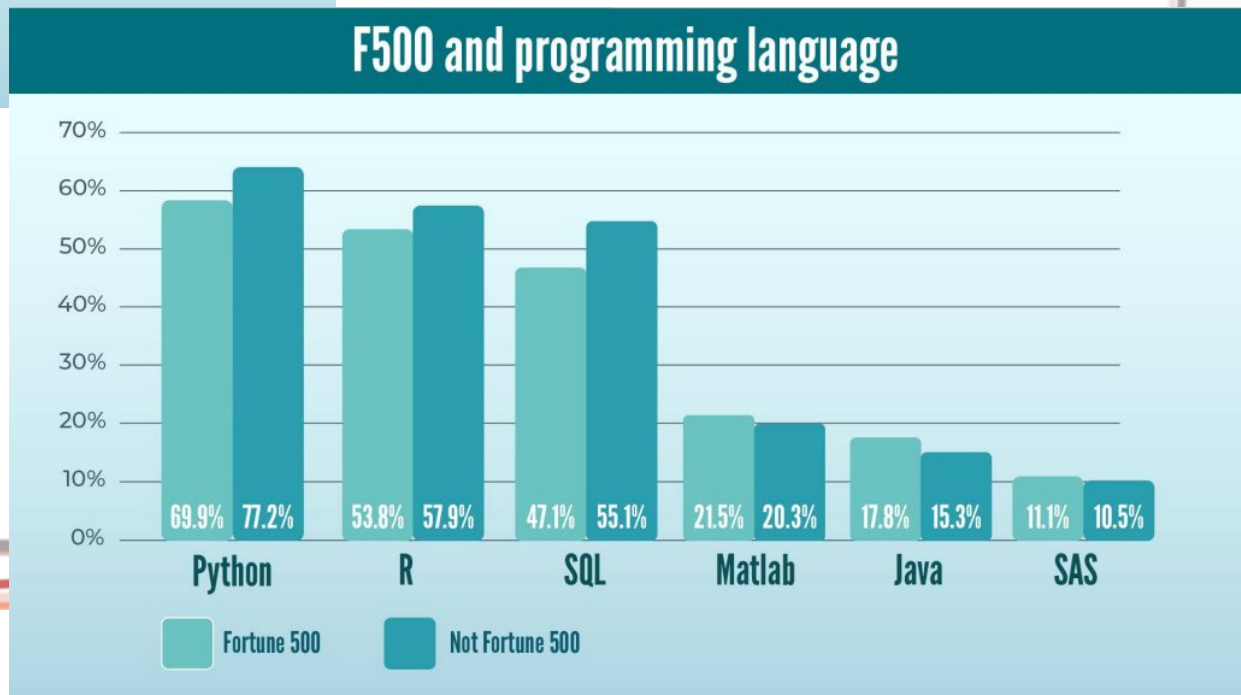


Regarding programming languages, in 2018, 50% of data scientists were using Python or R.

Ref: https://365datascience.com/career-advice/career-guides/become-data-scientist-2020/

What Data Scientists Do… Cont..



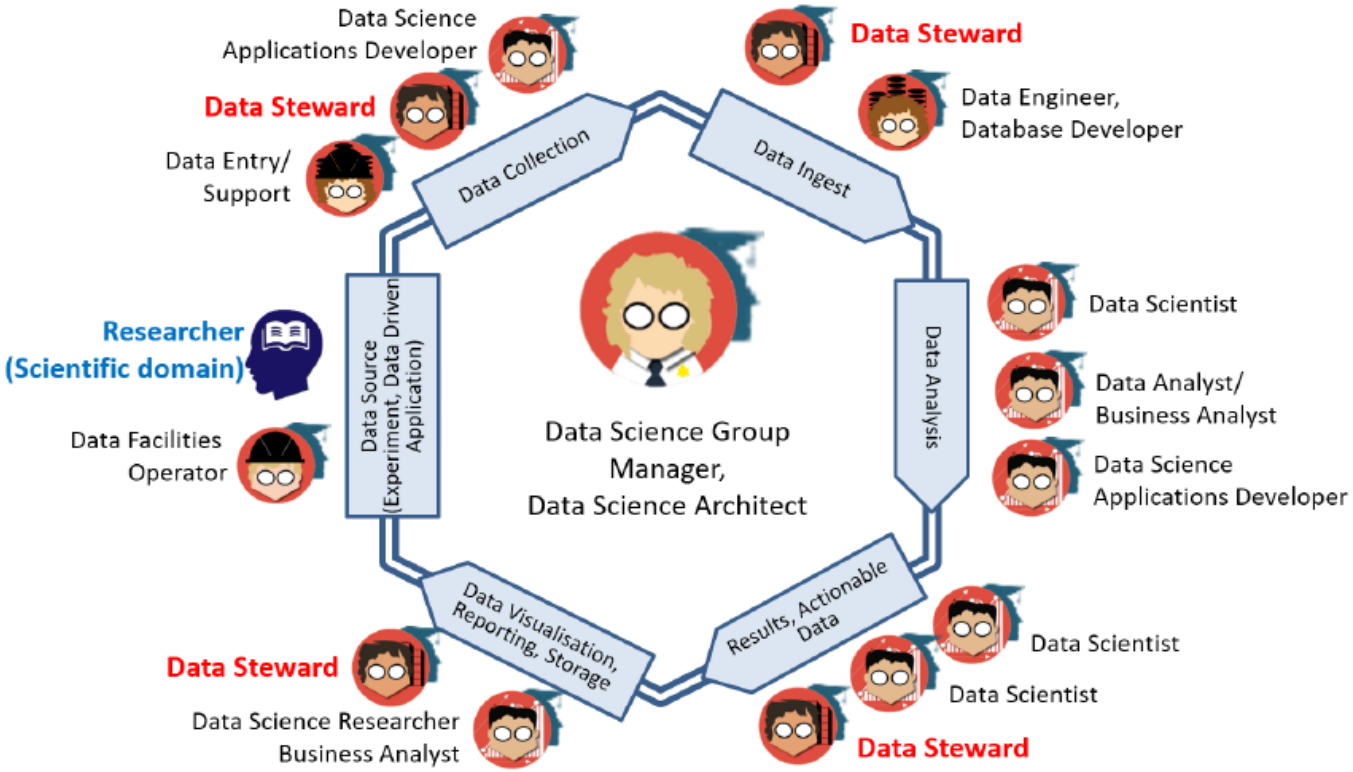


https://365datascience.com/career-advice/career-guides/become-data-scientist-2020/

Data Science Professional Profiles

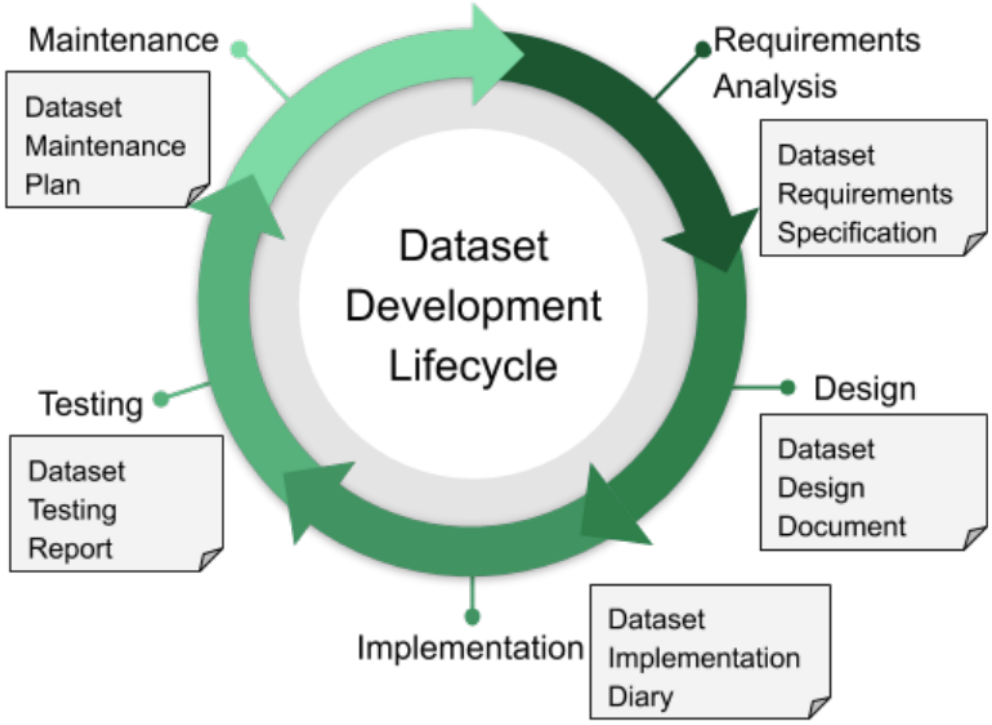
- Data Science Architects: They design and maintain the architecture of Data Science applications and facilities.
  - Create relevant data models and processes workflows.
- Data Stewards: Plan, implement and manage (research) data input, storage, search, presentation; creates data model for domain specific data; support and advice domain scientists/ researchers.
  - Create data model for domain specific data,
  - Support and advice domain scientists/researchers during the whole research cycle and data management lifecycle.
- Data Analyst: Although sometimes used as an alternative name for data scientist, they are generally regarded as junior roles(entry level) conducting data handling, modelling and reporting techniques.
- Data engineers: They deal with data pipelining and performance optimization issues. They need to deal with creating and integrating APIs.

Data Science Team composition



This cycle shows operational aspects of both data and modelling phases.

Dataset Development Lifecycle



This cycle shows the lifecycle of dataset creation.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Who has to know about Data Science?

- You have to know if you want to apply it
- Others? It is important to understand data science even if you never intend to apply it yourself
- Data-analytic thinking enables you to evaluate proposals for data science projects
  - For example, if there is a proposal to improve a particular business application by extracting knowledge from data, you should be able to assess the proposal systematically and decide whether it is sound or flawed
  - This does not mean that you will be able to tell whether it will actually succeed but you should be able to spot obvious flaws, unrealistic assumptions, and missing pieces

Data Science Projects

- Every project in this field should be at least:
  - **Reproducible**: Necessary for making easy to test other's work and analysis.
  - **Fallible**: Data science and science don't look for the truth, they look for knowledge, so every project can be substituted or improved in the future, no solution is the ultimate solution.
  - **Collaborative**: Data scientists don't exist alone, they need a team, this team will make things possible for developing intelligent solutions.
  - **Creative**: Most of what data scientists do is new research, new approaches or take on different solutions, so their environment should be very creative and easy to work.
  - **Compliant to regulations:** Right now there are a lot of regulations in science, not that much in data science, but there will be more in the future.
    - It is important that the projects we are building are aware of these different types of regulations so we can create a clean and acceptable solution for the problems.

Myths about Data Science

- Data Science offers many advantages but it also has some limitations

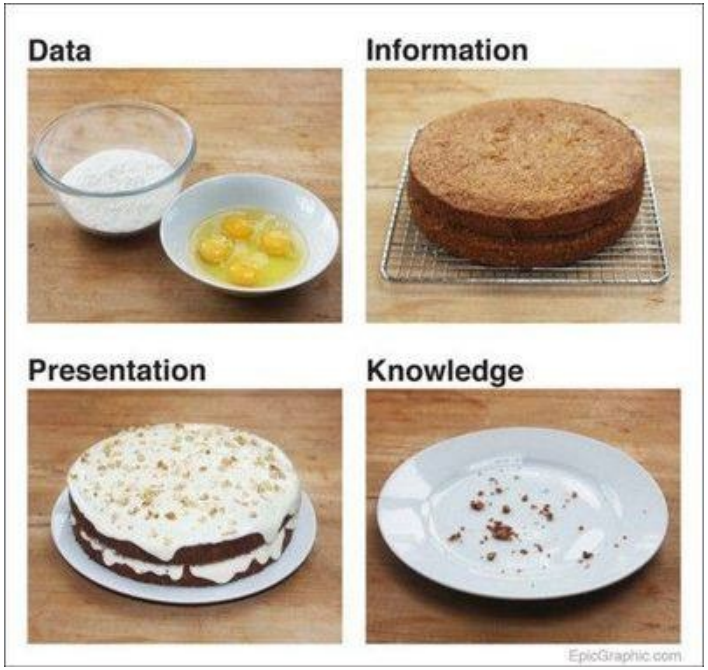
- Myth: Data science is an autonomous process that uses data to provide solutions to our problems
- Reality: It requires skilled human oversight throughout the different stages of the process

- Myth: Every project needs Big Data and uses Deep Learning
- Reality: In general, having more data helps, but having the right data is the more important requirement

Myths about Data Science

- Myth: Modern data science software is easy to use, and so data science is easy to do
- Reality: DS software is more user-friendly but doing data science properly requires both appropriate domain knowledge and the expertise regarding the properties of the data and the assumptions underpinning the different algorithms

- Myth: Data science pays for itself quickly
- Reality: This depends on the context. Adopting data science can require significant investment in terms of developing data infrastructure and hiring staff with data science expertise.



DI501

Week 1: Databases

What types of data can be analyzed?

Data: Research vs. Production

Research	Production
Clean	Messy
Historic	Constantly shifting
Mostly historic data	Historical+streaming data
	Biased, and you don't know how biased
	Privacy+regulatory concerns

Stanford CS 329S: Machine Learning Systems Design course notes, https://docs.google.com/presentation/d/1XKBXoLW3ED3FcvDRK3jfQeIlLCK2QTnsoZeopR6ziRQ/edit#slide=id.gb3e5cf682f_0_557

What types of data can be analyzed? Ordered/sequence data

- In many data mining tasks the order and timing of events contains important information
  - *Credit card usage profile* (10.4 €0, 11.4 €1000, 12.4 €1500, ..)
  - *Travel plan* (Road E75 for 100km, Road 24 for 25km, Road 313 for 5km)
  - *Process monitoring* (Warning X at 1am, Crash Y at 2am,...)
  - *Web sequence* < {Homepage} {Electronics} {Digital Cameras} {Canon Digital Camera} {Shopping Cart} {Order Confirmation} {Return to Shopping} >
  - *Sequence of books checked out at a library* <{Introduction to Data Mining} {Fellowship of the Ring} {The Two Towers, Return of the King}>

What types of data can be analyzed?

- Database data
- Transactional data
- Data warehouse data
- Data streams
- Ordered/sequence data
- Graph or networked data
- Spatial data
- Text data
- Multimedia data
- WWW

What types of data can be analyzed?

Databases vs. Data Warehouses: Purposes

Databases	Data Warehouses
Provides real time information Online transaction processing (OLTP) systems	Contain entire historical data Online analytical processing systems (OLAP)
Used for running the business Cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing, banking etc	Used for getting insights about how to run the business. Serve users or knowledge workers in the role of data analysis and decision making

What types of data can be analyzed?

Databases

- Relational Databases (SQL databases, RDBMS): use the Structured Query Language programming language to support their processes.
  - SQL databases rely on relations, or tables, and leverage common characteristics or patterns within the data to categorize and store information.
  - **Examples**: Oracle, MySQL, Google's F1 (a distributed relational database)

What types of data can be analyzed?

Data Warehouse

- A data warehouse as a storehouse, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema.
- A data warehouse gives the option to analyze data from different sources under the same roof.
- To facilitate decision-making and multi-dimensional views, data warehouses are usually modeled by a multi-dimensional data structure.

What types of data can be analyzed?

Databases

- Relational Databases (SQL databases): data is ordered in a structure with arranged rows and columns known as a table (relation). The rows are also referred as tuples whereas the columns are referred as attributes.

employee_id	first_name	last_name	address
1	John	Doe	New York
2	Benjamin	Button	Chicago
3	Mycroft	Holmes	London

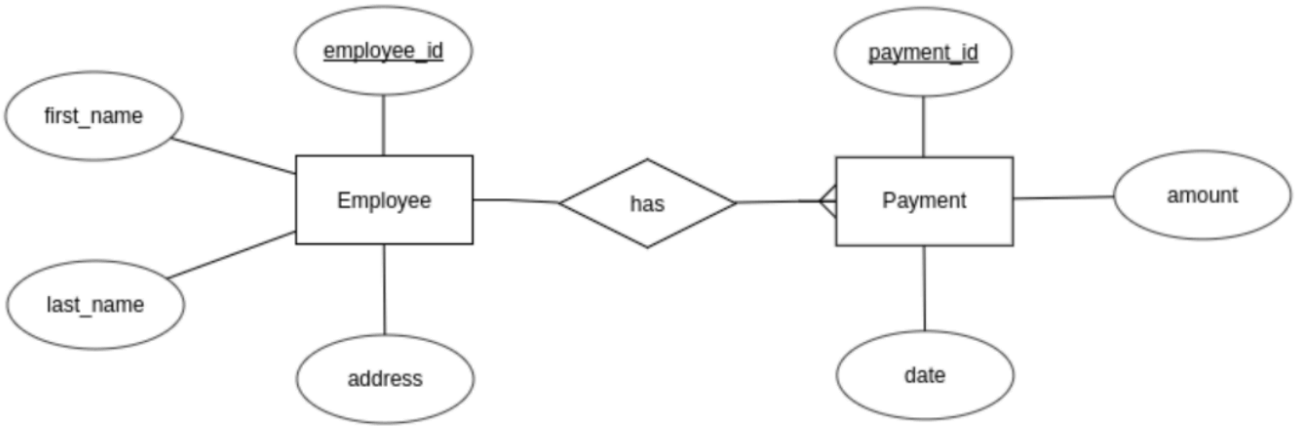
Employees: Table

The attribute *employee_id* is used to uniquely identify a tuple, as its value is distinct for each of the tuples. This is known as the *primary key*.

What types of data can be analyzed?

Databases

- Relational Databases (SQL databases): There can be several interrelated entities which can be represented by a entity-relationship diagram as follows.



Employee and Payment: Relationship Representation

What types of data can be analyzed?

Databases

- Relational Databases (SQL databases): In RDBMS tables, the relationships are represented by *foreign keys*.

employee_id	first_name	last_name	address
1	John	Doe	New York
2	Benjamin	Button	Chicago
3	Mycroft	Holmes	London

FOREIGN KEY

payment_id	employee_id	amount	date
1	1	50,000	01/12/2017
2	1	20,000	01/13/2017
3	2	75,000	01/14/2017
4	3	40,000	01/15/2017
5	3	20,000	01/17/2017
6	3	25,000	01/18/2017

With this representation, we can identify which user is associated with each of the payment. And the data can be retrieved using the query language, related to the RDBMS.

What types of data can be analyzed?

Databases

- If data does not conform to some schema or it is difficult to do, an alternative solution, **NoSQL** can be used.
- They still exhibit the same characteristics with SQL databases (durable, resilient, persistent, replicated, distributed, or performant)
  - They do not enforce schemas (or enforce only loose ones)
  - Extremely popular with big data because writes are too fast.
- Example applications: search engine databases

What types of data can be analyzed?

Databases: NoSQL

Brief History of NoSQL Databases

1998- Carlo Strozzi use the term NoSQL for his lightweight, open-source relational database

2000- Graph database Neo4j is launched

2004- Google BigTable is launched

2005- CouchDB is launched

2007- The research paper on Amazon Dynamo is released

2008- Facebooks open sources the Cassandra project

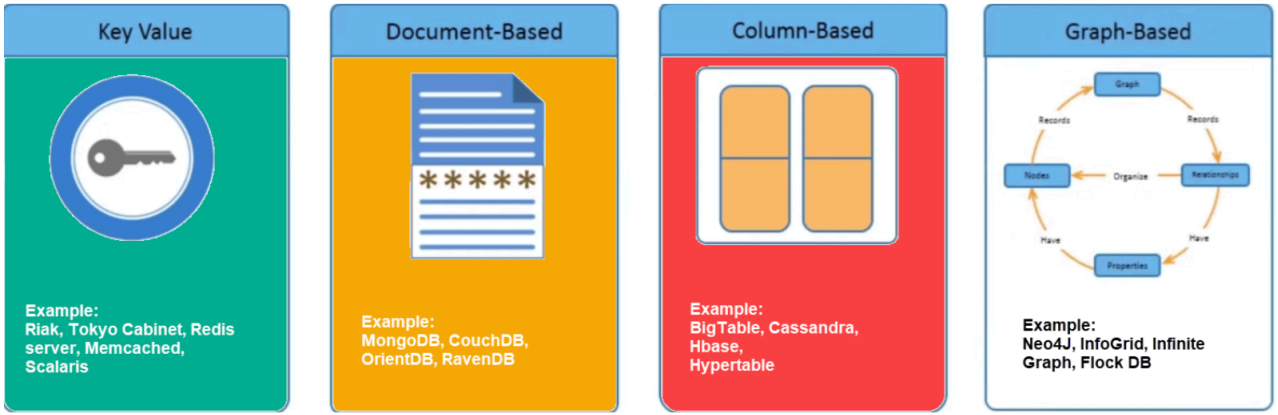
2009- The term NoSQL was reintroduced

Ref: https://www.guru99.com/nosql-tutorial.html

What types of data can be analyzed?

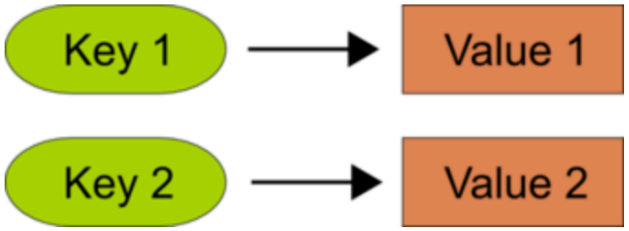
Databases

- Nonrelational Databases (NoSQL): do not use tables to store or categorize data. These databases can handle a wide variety of data types and models, and primarily leverage JSON documents, or complete and readily readable entities or data sets.
  - NoSQL databases are known for their ease-of-use and flexibility, as well as their scalability and performance.



What types of data can be analyzed? Databases

- Nonrelational Databases (NoSQL) types:
  - **Key-value store databases,** like Redis and Memcached, are also known as associative array databases. Here, each value or piece of data is assigned a specific key used to store and categorize information. These values can include a variety of data types, such as numbers, JSON, XML, HTML, PHP, images, videos, lists or even other key-value data.



Key-value stores can be considered as the most primary and the simplest version of all the databases.

It just has a one-way mapping from the key to the value to store data (like a production-scale hashmap).

What types of data can be analyzed? Databases

- Nonrelational Databases (NoSQL) types:
  - **Key-value store databases:** When defining a key there are three main components to specify. They are *prefix*, *identifiers,* and *suffix*. In general, any RDBMS table can be represented in a key-value schema as follows.

employee_id	first_name	last_name	address
1	John	Doe	New York
2	Benjamin	Button	Chicago
3	Mycroft	Holmes	London

Employees: Table

Prefix *identifiers* *suffix*

$table_name:$primary_key_value:$attribute_name = $value

For this table, the key-value schema can be defined as follows:

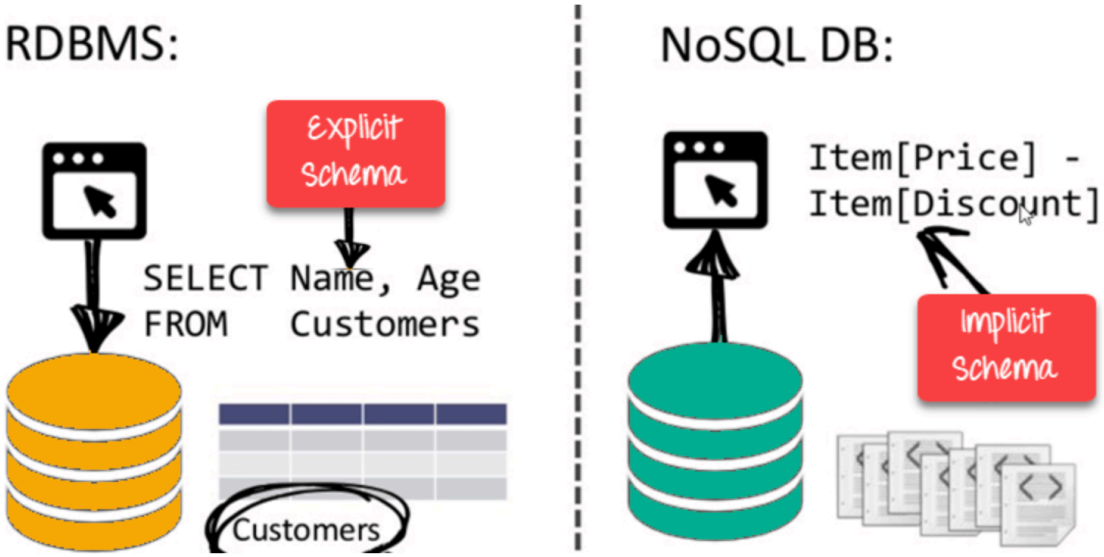
employee:$employee_id:$attribute_name = $value

employee:1:first_name = "John"
employee:1:last_name = "Doe"
employee:1:address = "New York“
employee:2:first_name = "Benjamin"
employee:2:last_name = "Button"
employee:2:address = "Chicago“
employee:3:first_name = "Mycroft"
employee:3:last_name = "Holmes"
employee:3:address = "London"

What types of data can be analyzed?

Databases

- Nonrelational Databases (NoSQL) types:
  - **Key-value store databases vs RDBMS**



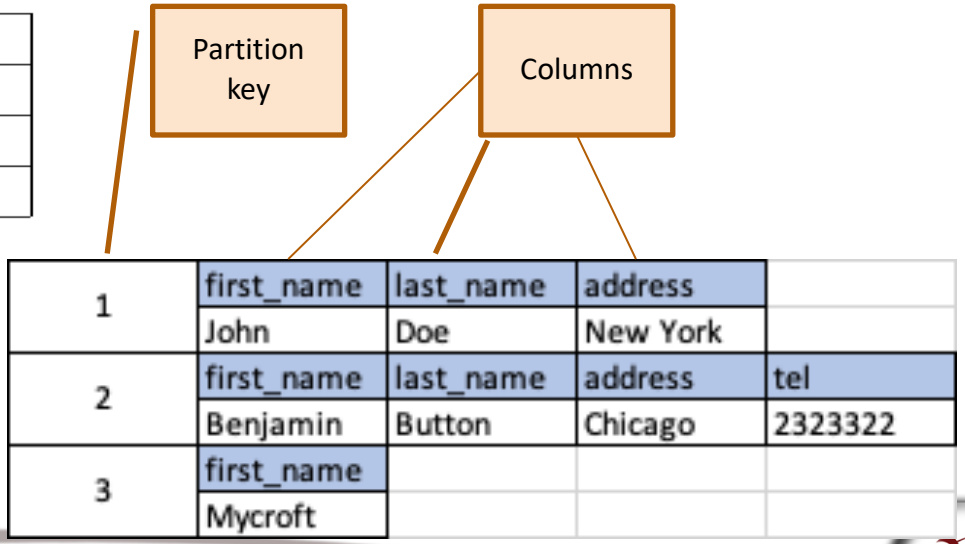
NoSQL is Schema-Free

What types of data can be analyzed? Databases

- Nonrelational Databases (NoSQL) types:
  - **Wide-column store (columnar) databases** are also nonrelational databases that leverage a dynamic data storage style.
    - Wide-column store databases are sometimes treated as a subset of key-value store databases, but they do include certain characteristics of traditional SQL databases.
    - Wide-column store databases include column families similar to that seen within relational database tables.

employee_id	first_name	last_name	address
1	John	Doe	New York
2	Benjamin	Button	Chicago
3	Mycroft	Holmes	London

Employees: Table



What types of data can be analyzed?

Databases

- **Wide-column store databases software:** MonetDB, C-Store, Apache Cassandra, Scylla, Apache HBase, Google BigTable, and Microsoft Azure Cosmos DB
- With a wide-column database, developers can simply add elements to a new column, without impacting existing columns or the data they hold.
- They are highly scalable because the data is stored in individual columns which can be sharded or partitioned across multiple servers.
- They don't have a defined table schema, which leaves them flexible to have certain columns only apply to certain records.
- Disadvantage: Updating or deleting a specific tuple for multiple attributes can be inefficient.
  - Therefore they are preferred for OLAP but not OLTP.

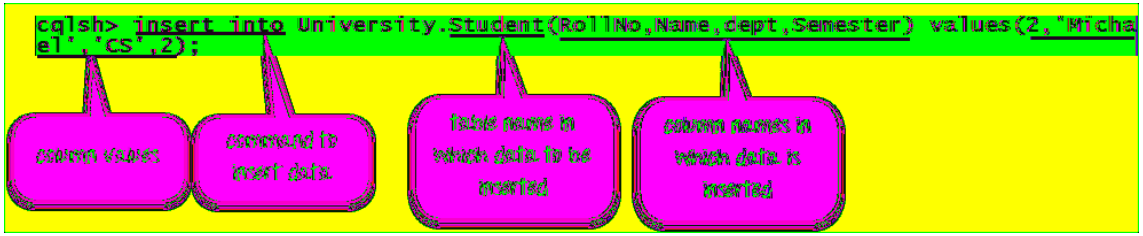
Ref: https://www.scylladb.com/glossary/wide-column-database/

What types of data can be analyzed? Databases

- **Wide-column store databases Cassandra example:**

Insert into KeyspaceName.TableName(ColumnName1, ColumnName2, ColumnName3 . . . .)
values (Column1Value, Column2Value, Column3Value . . . .)

Here is the snapshot of the executed Cassandra Insert into table query that will insert one record in Cassandra table 'Student'.



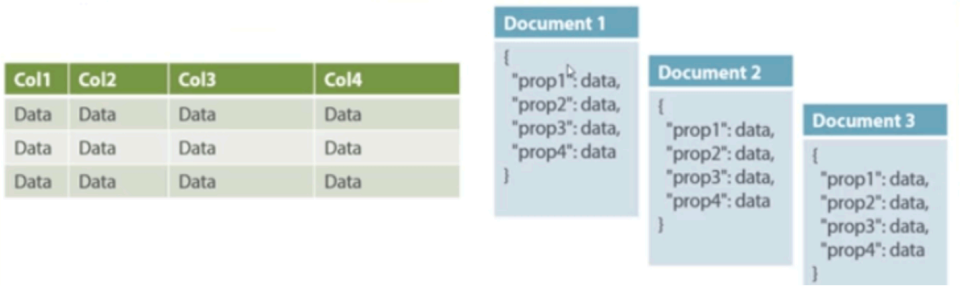
Insert into University.Student(RollNo,Name,dept,Semester) values(2,'Michael','CS', 2);

There are following limitations in Cassandra query language (CQL).
1.CQL does not support aggregation queries like max, min, avg
2.CQL does not support group by, having queries.
3.CQL does not support joins.
4.CQL does not support OR queries.
5.CQL does not support wildcard queries.
6.CQL does not support Union, Intersection queries.

Ref: https://www.guru99.com/cassandra-query-language-cql-insert-update-delete-read-data.html

What types of data can be analyzed? Databases

- Nonrelational Databases (NoSQL) types:
  - **Document store databases,** including MongoDB and Couchbase, utilize JSON, BSON, and XML document file types to store data within a more flexible, unstructured schema.
    - Documents can include all different types of data, making these types of databases ideal for semi-structured, or completely unstructured data.



Relational Vs. Document

A Json file example:

```
{
    "empid": "SJ011MS",
    "personal": {
        "name":"Smith Jones",
        "gender": "Male",
        "age": 28,
        "address":
        {
            "streetaddress": "7 24th Street",
            "city": "New York",
            "state": "NY",
            "postalcode": "10038"
        }
    },
    "profile": {
        "designation": "Deputy General",
        "department": "Finance"
    }
}
```

www.kodingmadesimple.com

What types of data can be analyzed? Databases

- Nonrelational Databases (NoSQL) types:
  - **Document store databases: Example**

Figure illustrates a MongoDB database structure where, the customer database has a collection of usernames and courses documents.

The "_id" field is a primary key, a unique identifier type named ObjectId which is created by the application when the document is created.

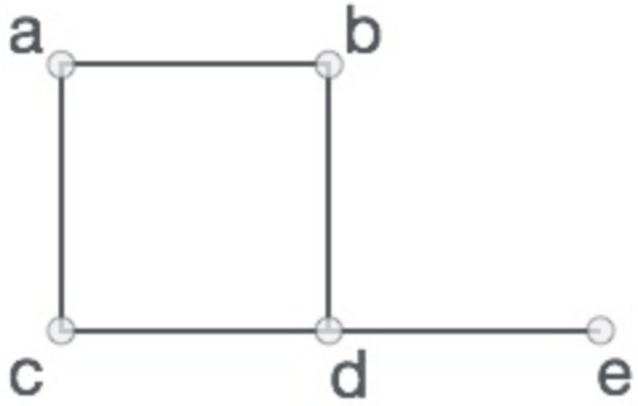
Fig 4: MongoDB database structure

Ref: https://dev.to/nyashanziramasanga/document-store-database-mongodb-part-3-4226

What types of data can be analyzed?

Graph or network data

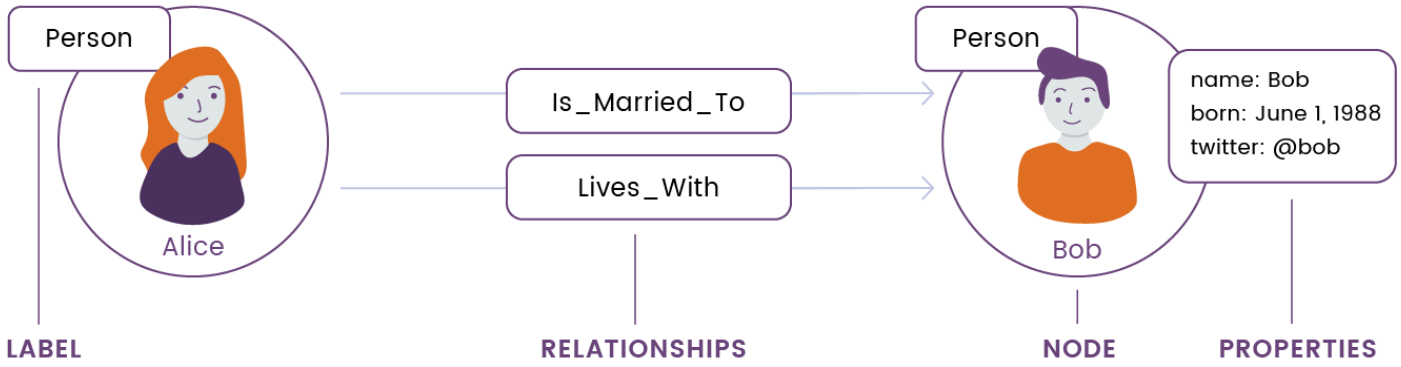
- A graph is a pictorial representation of a set of objects where some pairs of objects are connected by links. The interconnected objects are represented by points termed as vertices, and the links that connect the vertices are called edges.



What types of data can be analyzed?

Graph or network data

- Nonrelational Databases (NoSQL) types:
  - **Graph Based:** The graph data structure might seem unusual, but it's simple and natural. Here's an example of a simple graph data model in Neo4j database:



This graph contains two nodes (Alice and Bob) that are connected by relationships. Both nodes share the same label, *Person*. In the graph, only Bob's node has properties, but in Neo4j every node and relationship can have properties.

What types of data can be analyzed?

Graph or network data

- Some applications:
  - Security (such as detecting cyber threats, organized crime family)
  - Hyperlink analysis (such as finding the high quality web pages)

What types of data can be analyzed?

Graph or netw

- Some applications:
- Recommendation systems (such as in which advertisements you would be interested in Facebook, Instagram)
- User attribute and behavior understanding (such as how people will react to certain events)
- Figure right shows an example for an identity graph.

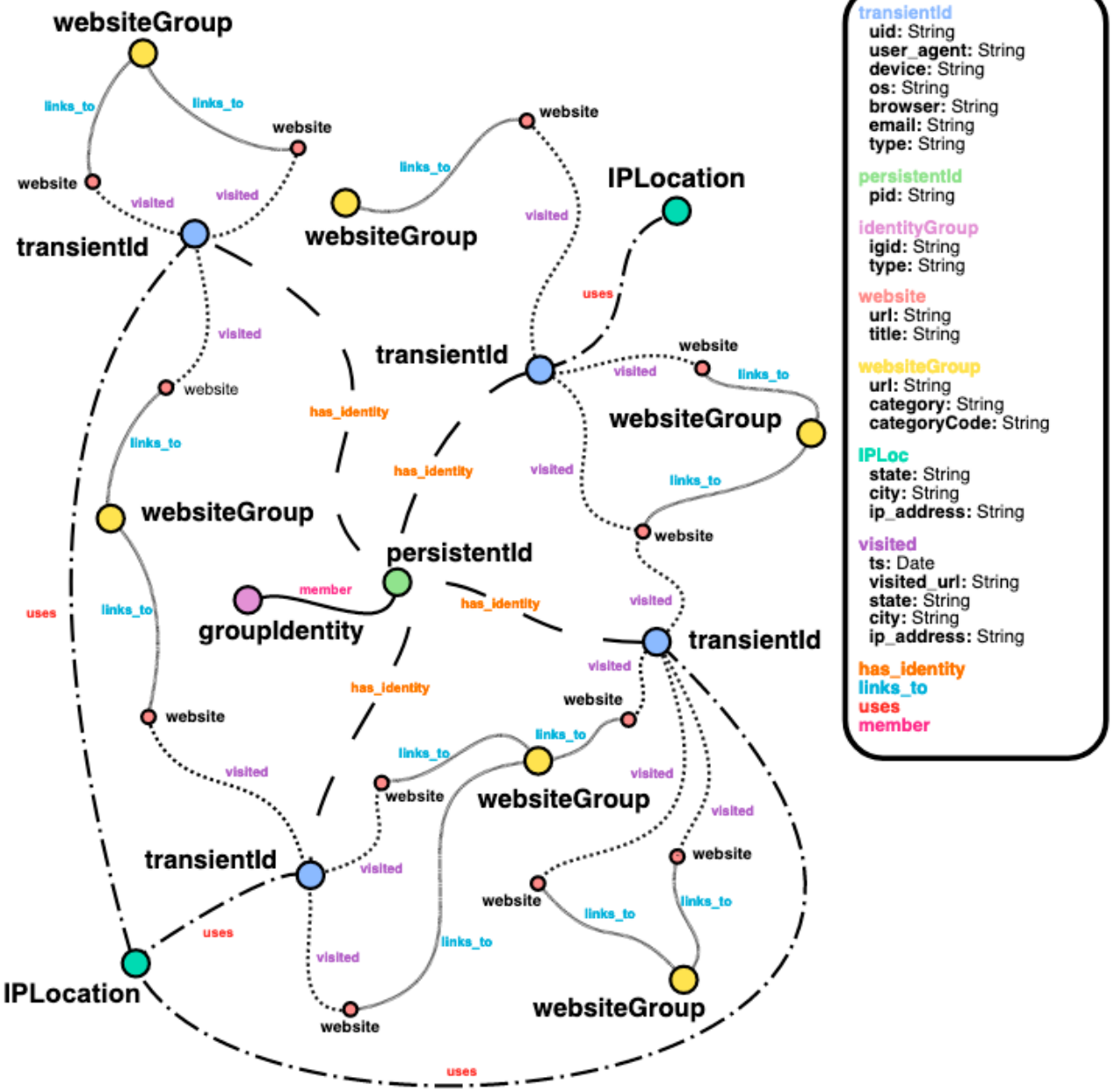


Image ref: https://aws.amazon.com/neptune/identity-graphs-on-aws/?pg=ln&sec=uc

What types of data can be analyzed?

Spatial Databases

- Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning.
- Spatial objects must have spatial coordinates (latitude and longitude).
  - It has geometry as points, lines and polygons.
- Spatial databases store spatial relationships between its objects.
  - Example: you can search for an object that spatially intersects another.

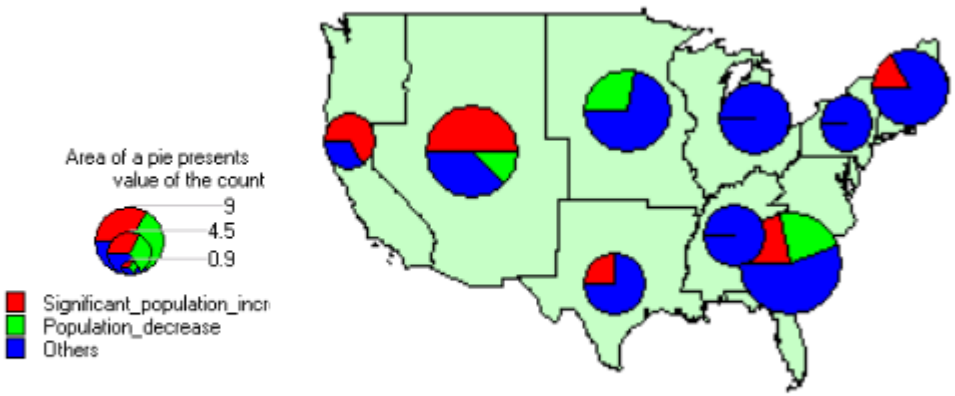


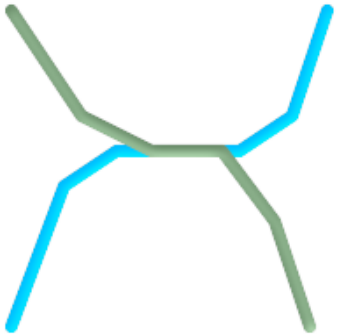
Figure 1.6: Visualization of spatial OLAP (from GeoMiner system)

What types of data can be analyzed?

Spatial Databases

- There are popular databases such as PostgreSQL used for spatial querying.

Using intersection matrix patterns, specific spatial relationships can be evaluated in a more succinct way.
The ST_Relate and the ST_RelateMatch functions can be used to test intersection matrix patterns.
Note that the intersection matrix pattern specifying two lines intersecting in a line is '1*1***1**':



```
-- Find road segments that intersect in a line
SELECT a.id
FROM roads a, roads b
WHERE a.id != b.id
      AND a.geom && b.geom
      AND ST_Relate(a.geom, b.geom, '1*1***1**');
```

https://postgis.net/docs/manual-dev/using_postgis_query.html

What types of data can be analyzed?

Transactional Databases

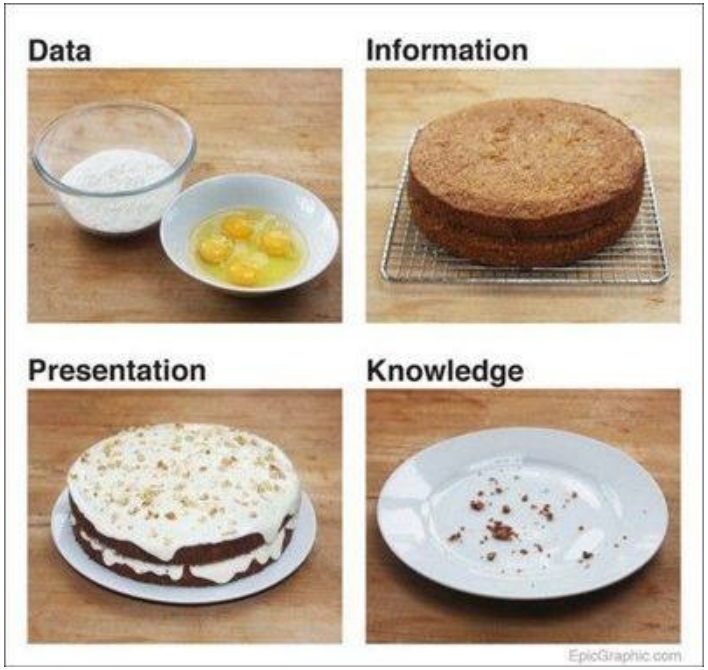
- A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items.
  - Transactional databases are row-stores, which means that data is stored on disk as rows, rather than columns.
- Since relational databases do not allow nested tables (i.e. a set as attribute value), transactions are usually stored in flat files or stored in two normalized transaction tables, one for the transactions and one for the transaction items.

Rentals

transactionID	date	time	customerID	itemList
T12345	99/09/06	19:38	C1234	{I2, I6, I10, I45 …}
…				

Figure 1.5: Fragment of a transaction database for the rentals at OurVideoStore.

TID	Transaction time	Transaction
T_1	2015/3/07 09:30	a, c, e
T_2	2015/3/07 10:20	b, d
T_3	2015/3/09 19:35	a, b, c, e
T_4	2015/3/09 20:20	c, d
T_5	2015/3/10 10:00	b, c, e
T_6	2015/3/10 13:45	b, d
T_7	2015/3/11 09:10	a, c, d, e
T_8	2015/3/11 9:44	b, c, e
T_9	2015/3/11 16:10	a, c, d
T_{10}	2015/3/12 10:35	a, b, c, d, e



DI501

Week 1: Related Fields and Big Data

Related Fields

Data Mining and Machine Learning

- Many of the elements of data science have been developed in related fields such as machine learning and data mining
  - In fact, the terms data science, machine learning, and data mining are often used interchangeably
- The commonality across these disciplines is a focus on improving decision making through the analysis of data
- Although data science borrows from these other fields, it is broader in scope

Related Fields

Data Mining and Machine Learning

- Machine learning (ML) focuses on the design and evaluation of algorithms for extracting patterns from data
- Data mining generally deals with the analysis of structured data and often implies an emphasis on commercial applications
- Data science takes all of these considerations into account but also takes up other challenges, such as
  - the capturing, cleaning, and transforming of unstructured data such as social media and web data
  - the use of big-data technologies to store and process big, unstructured data sets, and
  - questions related to data ethics and regulation

Related Fields

Data Mining

- Data mining turns a large collection of data into knowledge (interesting patterns)
- Data Mining, also popularly known as Knowledge Discovery in Databases(KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.
- While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Related Fields

Data Mining

- Looking for patterns in data
  - Items X,Y,Z are bought together frequently
  - People who like movie X also like movie Y
  - Patients who respond well to medicines X and Y also respond well to medicine Z
  - Students going to the same university are frequently online friends
  - Wealthier people are moving from cities to suburbs

Related Fields

Data Mining

- Looking for patterns in data
  - Items X,Y,Z are bought together frequently
  - People v
  - Patients
    respond
  - Student
    online friends
  - Wealthier people are moving from cities to suburbs

Frequent Itemsets
Association Rules
Specialized techniques for networks, text, multimedia

Related Fields

Machine Learning

- Using data to build models and make predictions (a subset of Artificial Intelligence)
  - Customers who are women over age 20 are likely to respond to an advertisement
  - Students with good grades are predicted to do well on the SAT
  - The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

Related Fields

Machine Learning

- Using data to b[illegible] predictions
  - Customers w[illegible] 20 are likely to respond to a
  - Students wit[illegible]cted to do well on the SAT
  - The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

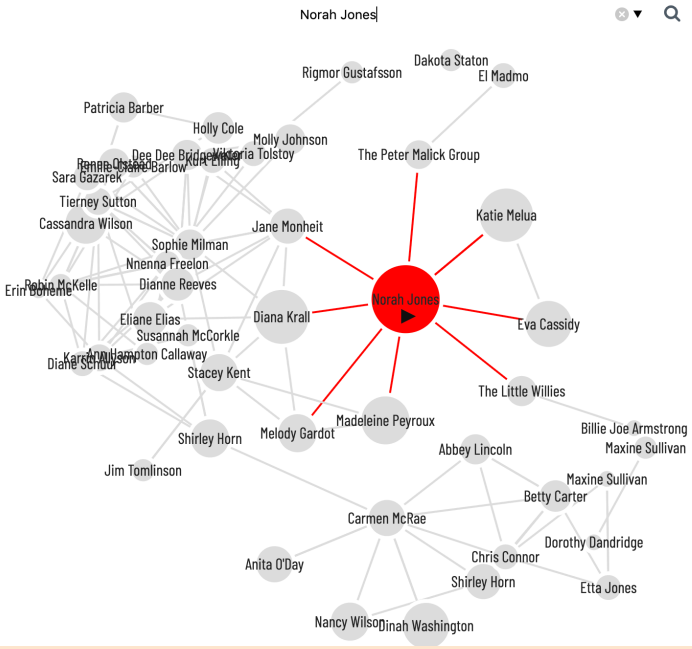
Regression
Classification
Clustering

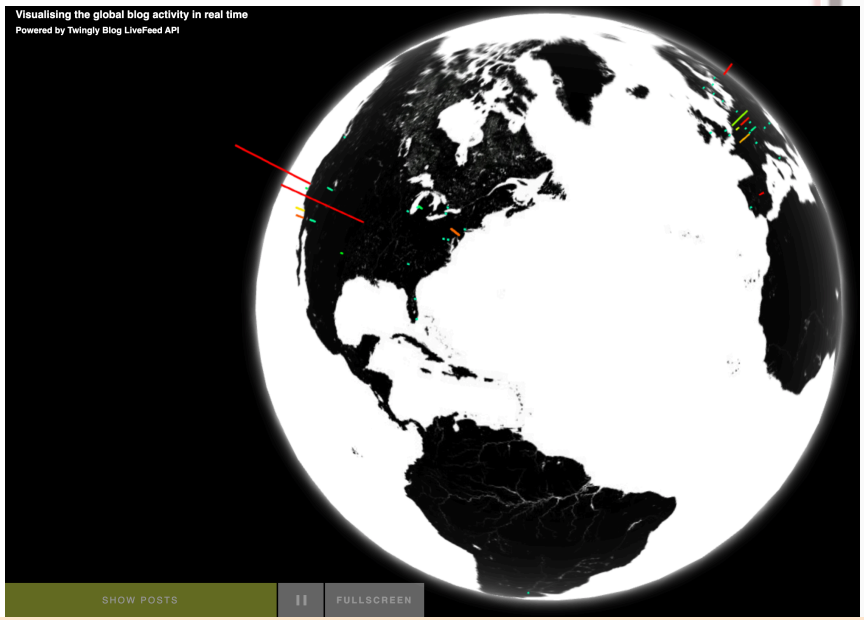
Roughly: Basic data analysis and data mining give answers from the available data, while machine learning uses the available data to make predictions about missing or future data.

Related Fields

Data Visualization: Fancy ones

- “A picture is worth a thousand words”





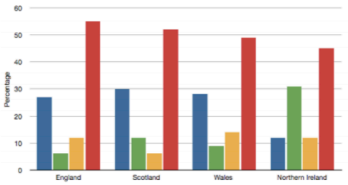
Liveplasma is a music and movie visualization app that aims to help you discover other musicians or movies you might enjoy. Type in the name of a band, artist, movie, director or actor and liveplasma will show you related people, bands or movies.

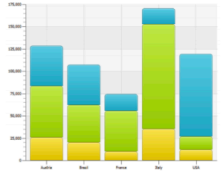
The Twingly Screensaver visualizes the blogosphere worldwide in real time. You get a continuous feed of blog activity straight to your screen.

Related Fields

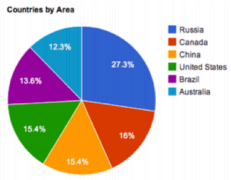
Data Visualization: Basic ones

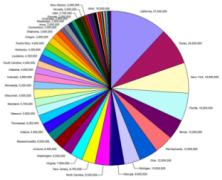
- Don’t underestimate the power of basic visualizations
  - Bar charts



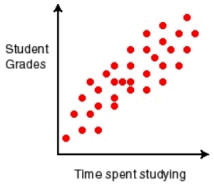


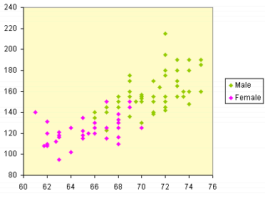
  - Pie charts



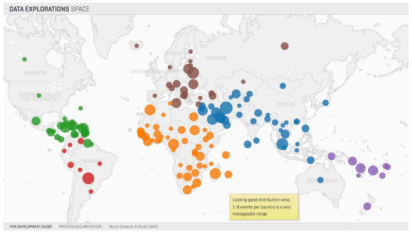


  - Scatterplots



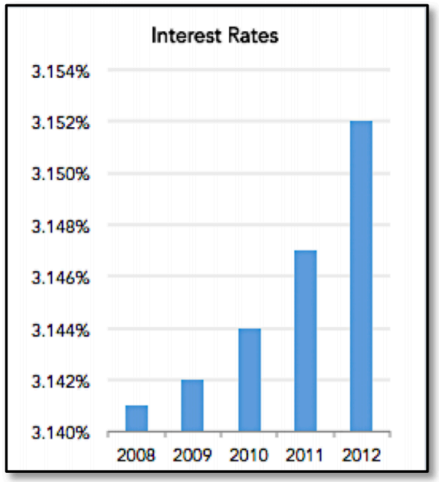


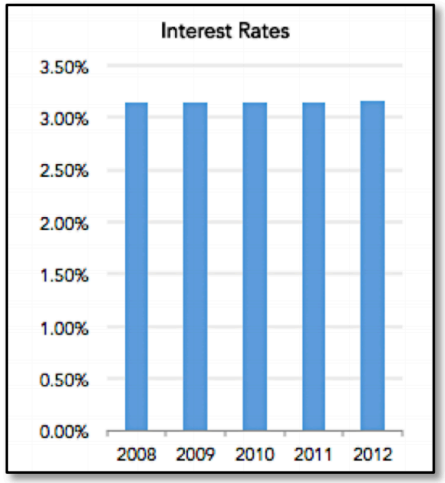
  - Maps

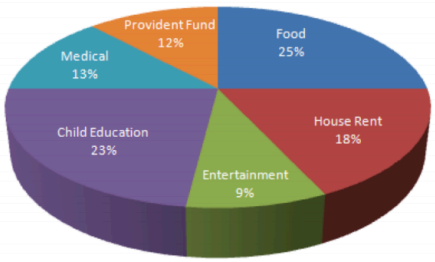


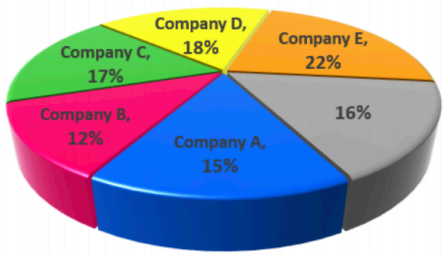
Related Fields

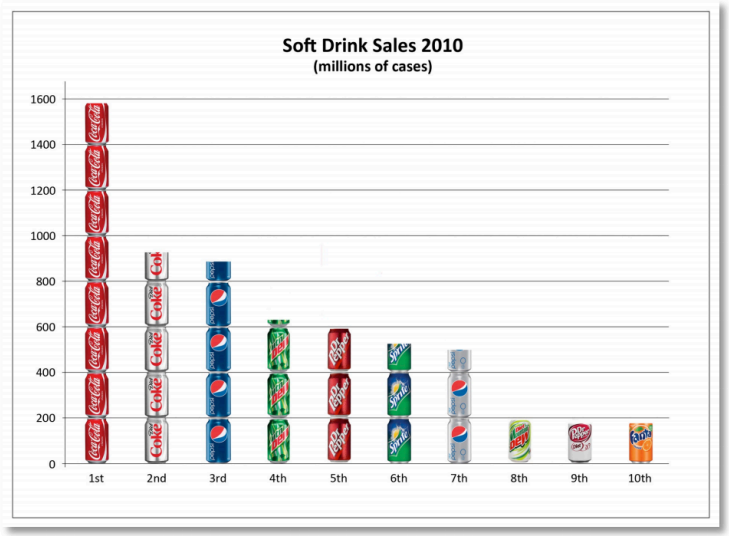
Data Visualization: Misleading ones











Related Fields

Data Engineering

- Data engineering is the aspect of data science that focuses on practical applications of data collection and analysis.
- Data engineers focus on the applications and harvesting of big data. Their role doesn't include a great deal of analysis or experimental design.

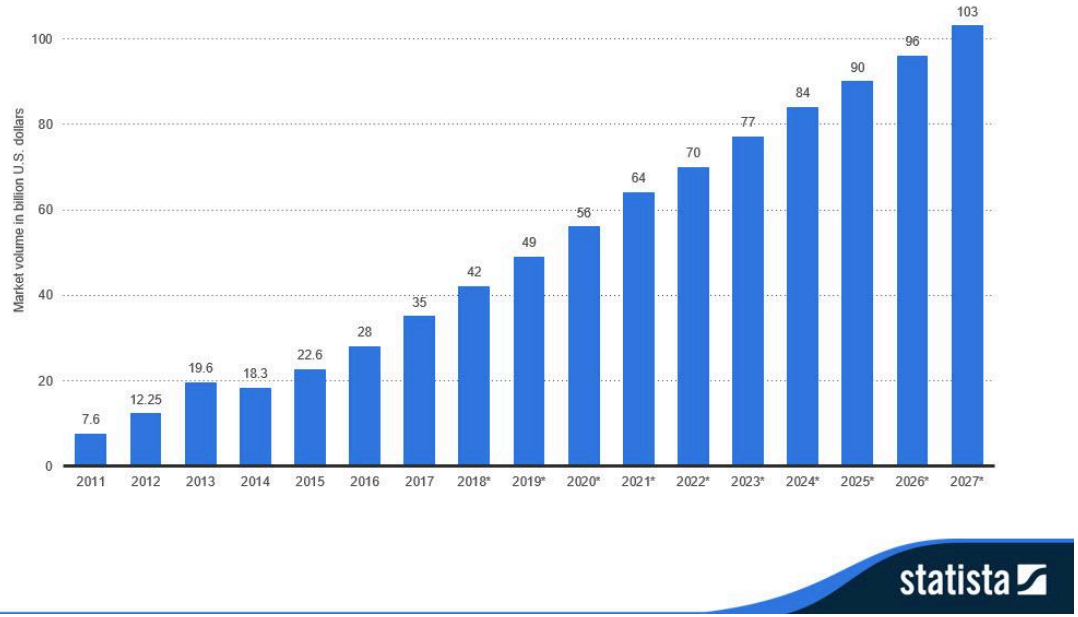
What is Big Data?

- Complete works of William Shakespeare
  - 5 megabytes
- Average individual
  - 50 gigabytes (10,000 Shakespeares)
- USA Library of Congress
  - 10 terabytes (2 million Shakespeares)
- Uploaded to Facebook daily
  - 1 petabyte (200 million Shakespeares)
- Produced by humanity daily
  - 2.5 exabytes (500 trillion Shakespeares)

What is Big Data?

Forecast Revenue Big Data Market Worldwide 2011-2027

Big Data Market Size Revenue Forecast Worldwide From 2011 To 2027 (in billion U.S. dollars)



The Hadoop and Big Data Market are projected to grow from $17.1B in 2017 to $99.31B in 2022 attaining a 28.5% CAGR (Compound Annual Growth Rate). The greatest period of projected growth is in 2021 and 2022 when the market is projected to jump $30B in value in one year.

What is Big Data?

- Big data 7 V's
  - Volume
    - Scale
    - Sources
  - Variety
    - Relational
    - NoSQL
  - Velocity
    - Operational
    - Analytical
  - Variability
  - Veracity
  - Visualization
  - Value

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

What is Big Data?

- Big data 7 V's
  - Volume
    - Scale
    - Sources
  - Variety
    - Relational
    - NoSQL
  - Velocity
    - Operational
    - Analytical
  - Value

Quantities of data that reach almost incomprehensible proportions.
As far back as 2016, Facebook had 2.5 trillion posts and more than 250 billion images.

What is Big Data?

- Big data 7 V's
  - Volume
    - Scale
    - Sources
  - Variety
    - Relational
    - NoSQL
  - Velocity
    - Operational
    - Analytical
  - Variability
  - Veracity
  - Visualization

Different types of data: images, video, text, audio recordings
Both structured and unstructured

What is Big Data?

- Big data 7 V's
  - Volume
    - Scale
    - Sources
  - Variety
    - Relational
    - NoSQL
  - Velocity
    - Operational
    - Analytical
  - Variability
  - Veracity
  - Visualization
  - Value

Velocity is the measure of how fast the data is coming in. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.

What is Big Data?

- Big data 7 V's
  - Volume
    - Scale
    - Sources
  - Variety
    - Relatio
    - NoSQL
  - Velocity
    - Operat
    - Analyt
  - Variability
  - Veracity
  - Visualization
  - Value

Variability is different from variety. A coffee shop may offer 6 different blends of coffee, but if you get the same blend every day and it tastes different every day, that is variability. The same is true of data, if the meaning or structure is constantly changing it can have a huge impact on your data homogenization.
If you change variables, your model will also change.

What is Big Data?

- Big data 7 V's
  - Volume
    - Scale
    - Sources
  - Variety
    - Relational
    - NoSQL
  - Velocity
    - Operat
    - Analyt
  - Variability
  - Veracity
  - Visualization
  - Value

Veracity is all about making sure the data is accurate, which requires processes to keep the bad data from accumulating in your systems.

What is Big Data?

- Big data 7 V's
  - Volume
    - Scale
    - Sources
  - Variety
    - Relational
    - NoSQL
  - Velocity
    - Operat
    - Analyt
  - Variability
  - Veracity
  - Visualization
  - Value

Visualization is critical in today's world. Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.

What is Big Data?

- Big data 7 V's
  - Volume
    - Scale
    - Sources
  - Variety
    - Relational
    - NoSQL
  - Velocity
    - Operational
    - Analytical
  - Variability
  - Veracity
  - Visualization
  - Value

You want to be sure your organization is getting value from the data.

Conclusion

- Data science is a multidisciplinary field that combines the latest innovations in advanced analytics, including machine learning and artificial intelligence, with high-performance computing and visualizations.
- Data scientists are expected to have a variety of skills.
- Data science is a fast and developing domain.
  - Be prepared to keep track of the advancements in the domain. Otherwise, you will be out of the league.

Conclusion

- Data informatics involve people, processes, and technologies related to data.
  - Getting more complex, bigger and challenging.
- You still need to specialize in data informatics:
  - Machine learning experts, data engineers, etc. all have common and different experiences and interests.

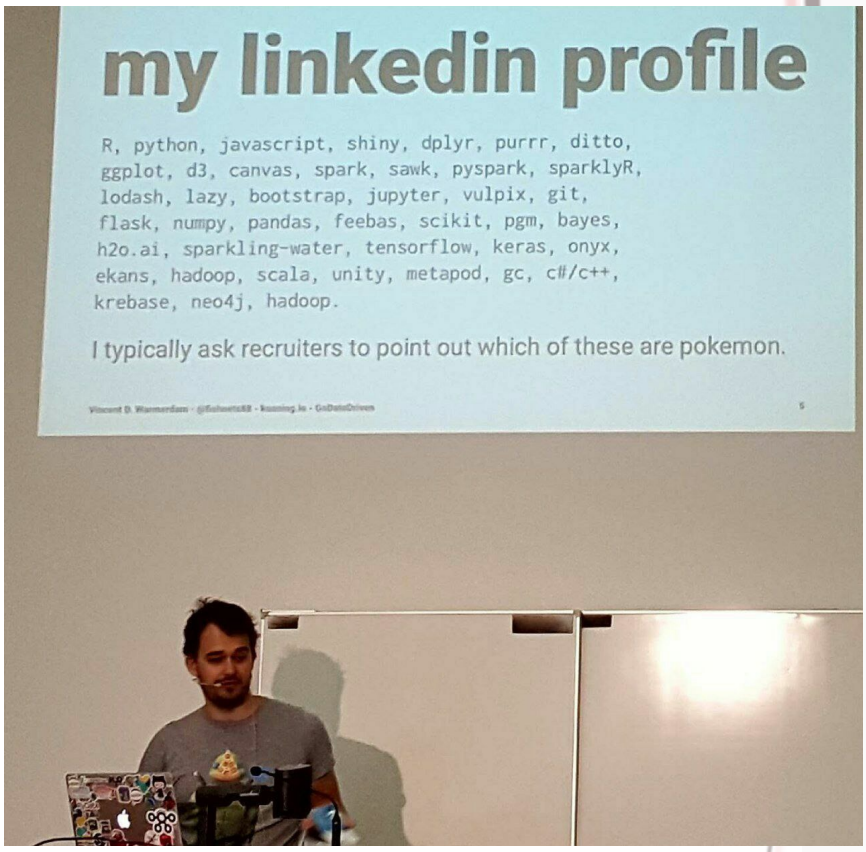


Image courtesy: https://i.redd.it/bka1gb843z7z.jpg