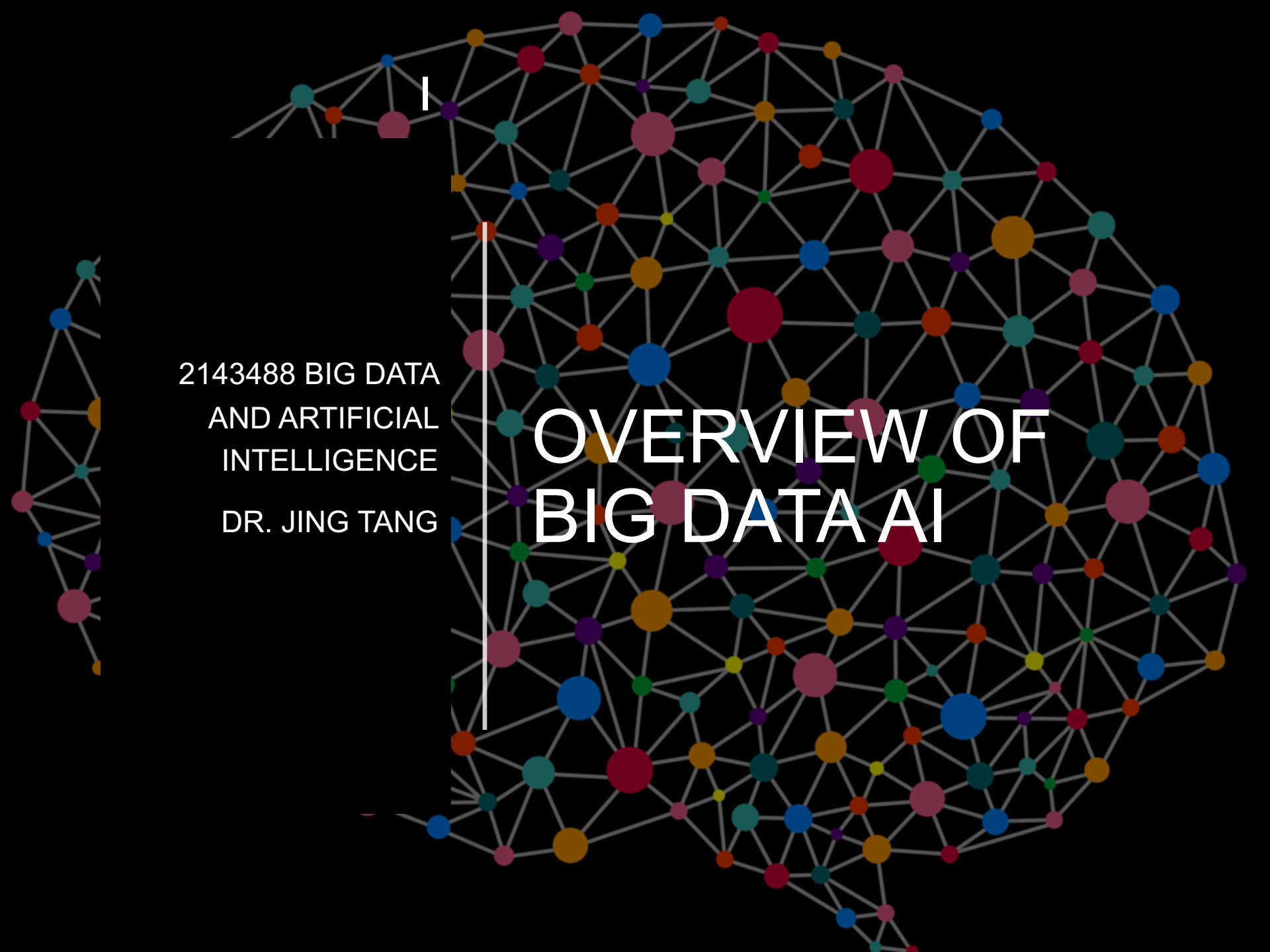


2143488 BIG DATA AND ARTIFICIAL INTELLIGENCE

DR. JING TANG

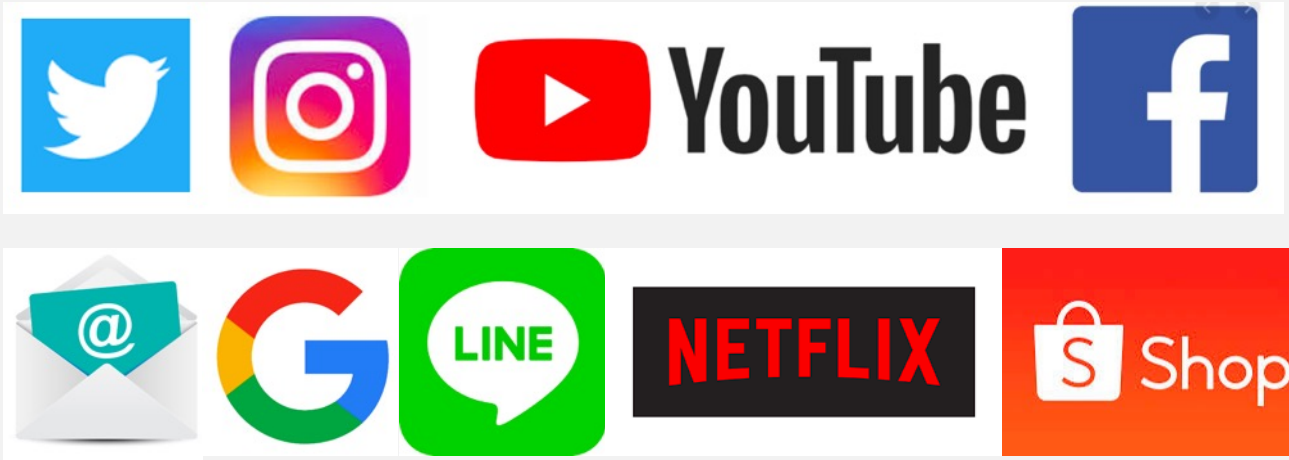
OVERVIEW OF BIG DATA AI

DATA ALL AROUND

- Emerging technologies and reduction in costs from storage to compute have transformed the data landscape
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network
  - Magnified by IoT, etc.

Watch "Big Data in 5 Minutes" on YouTube

HOW MUCH DATA DO **WE** PRODUCE?



Internet users: 2.5B (2012) -> 4.5B (2019 -> 5.07B (2022)

1 Single people: 1.7MB/s (2022)

Data: 1.2ZB (2010) -> 40ZB(2020) => 97ZB (2022) ZB is 10^21 bytes

1 ZB = 160B hrs. 4K Netflix Streaming = 800B hrs. 720p Zoom Meeting

About 2.7M ***years to download all data via AIS 5G***

B, KB, GB, TB, PB, EB , ZB, YB

TYPES OF DATA

- Structured Data: tabular form
  - E.g.: Excel, mySQL
- Semi-structured Data: no formal data model
  - E.g.: XML, HTML, Email, Website pages
- Unstructured Data: no pre-defined data model
  - E.g.: Emails, Image, Photo, Video, Text, Social media, message, Website content

Separate data by columns.
Each row is 1 record.

Database: strictly 1 column 1 feature
Excel: can customize stuff

Can be any format, very customized

email: you can write whatever title, content. but the whole email has from, to, subject, attached files, etc. Email is constructed from different parts. So, email is semi-structured.

If we only consider the content part, then it's unstructured.

80% of data is unstructured

STRUCTURED DATA

- 20% of Enterprise data
- Being produced often in real time, in large volumes
- Machine-generated: Sensor (RFID, GPS, smart meters, medical devices); Web log data; POS; Financial
- Human-generated: Input; Click-stream; Gaming-related
- Relational Database Management System (RDBMS)
  - Data is stored in Tables
  - Each column is an attribute; each row is a record
  - Schema shows a relationship of tables
  - SQL is used to query tables

UNSTRUCTURED DATA

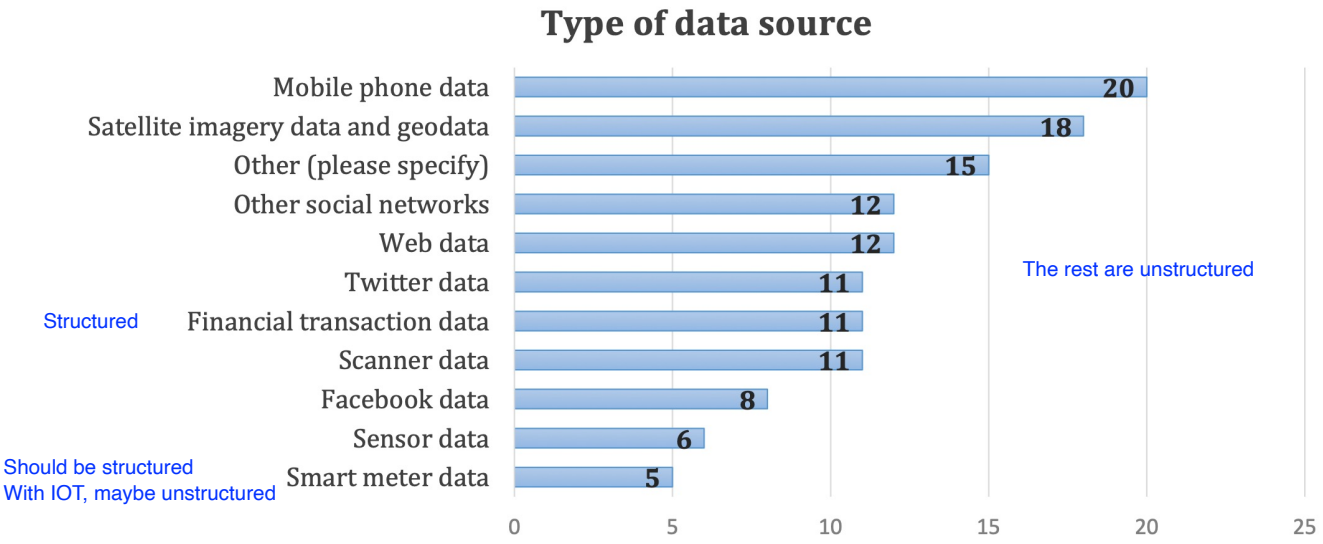
- 80% of Enterprise data
- Requires more storage, and more difficult to manage
- Its usage is rapidly expanding, but still challenge
- Machine-generated: Satellite; Science; Photo & Video; Radar & Sonar
- Human-generated: Text; Social media; Mobile; Website
- NoSQL("not only SQL") Databases
  - Non-tabular databased and store data differently

Different sensors collect different information, but might be in different format across OS's or mobile phone brands.
Mobile phone data is highly unstructured. Free format.

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

BigData
UN Global Working Group

1. Survey of SDG-related Big Data projects

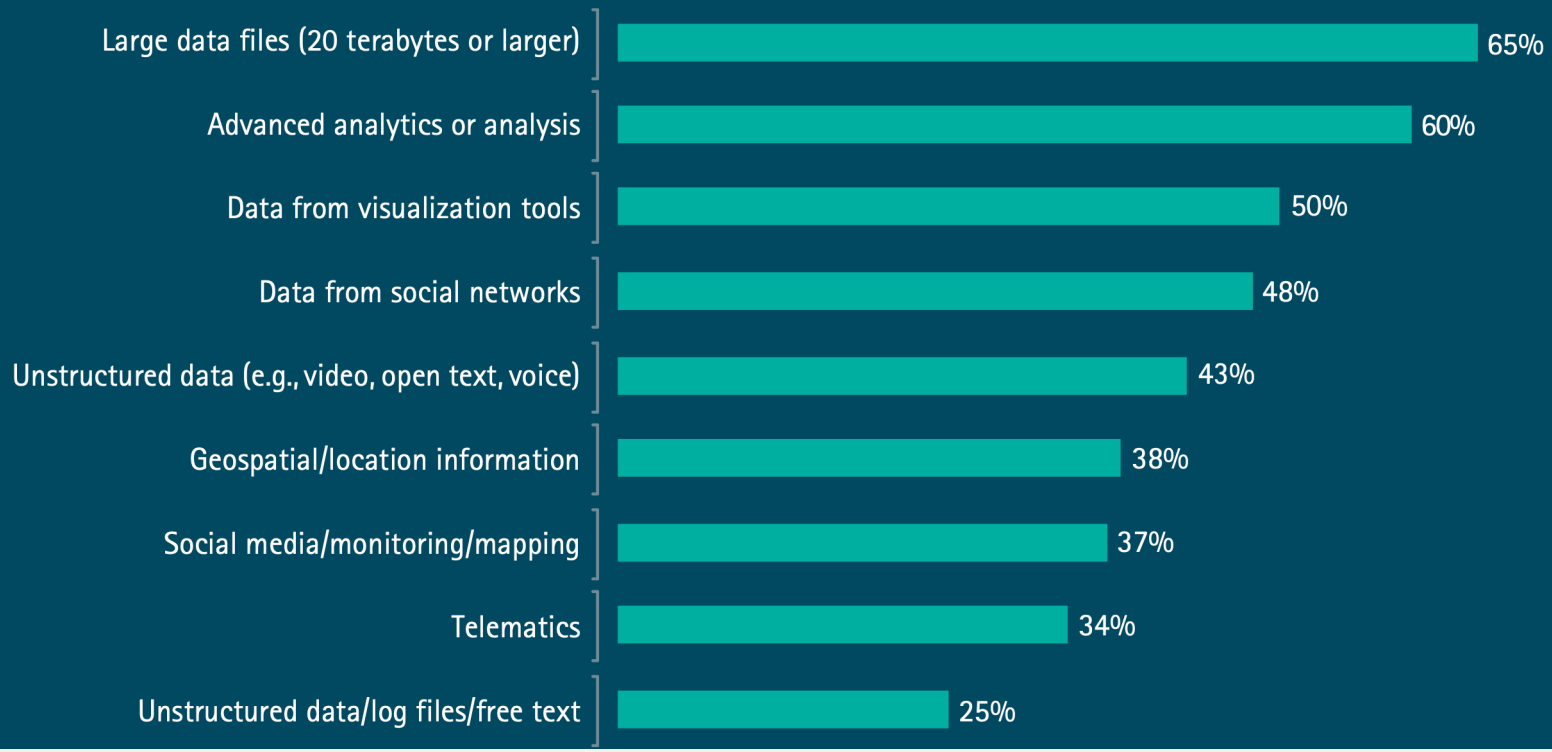


- Mobile phones (20), satellite imagery (18) and social media (11+8) are the most prominent sources
- Otherwise, wide range of sources

Soure: unstats.un.org (2016)

Figure 2: Sources of big data

Which of the following do you consider part of big data (regardless of whether your company uses each)?



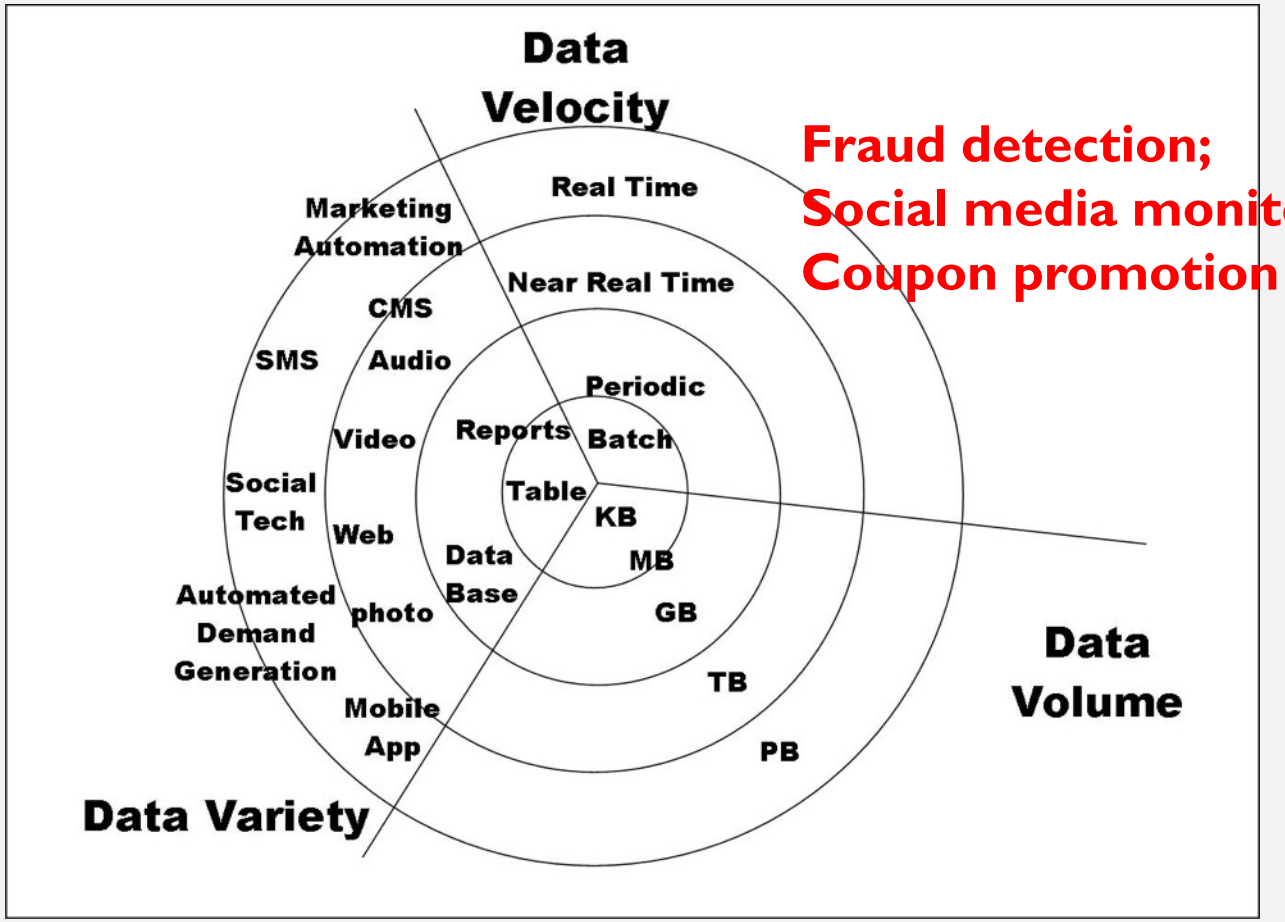
Source: Accenture Big Success with Big Data Survey (2014)

BIG DATA

"5V"

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume — **90% created in last 2 years** *(IBM,2017)*
    Volume grows very quickly. Needs lots of storage and processing power.
  - Velocity — **Streaming data, short useful lifespan**
    Speed. Lots of data generated. Data expires very quickly. Need to make processing fast with acceptable accuracy
  - Variety — **In many forms, difficult to integrate**
    Most of data are unstructured. Maybe data are not in 1 database, need to query many DBs or need to format data
  - Veracity — **Untrusted, uncleaned**
    Data is untrusted, uncleaned. Sensor ex: sensors affected by the internet, or the sensors themselves. Usually errors or some cases where data are missing.
  - Value — **Data is the new oil!**
    Data only makes value when you get insights of things.
- Big Data is the capability to manage a huge volume of disparate data, at **right speed**, and within the **right time frame** to allow real-time analysis and reaction

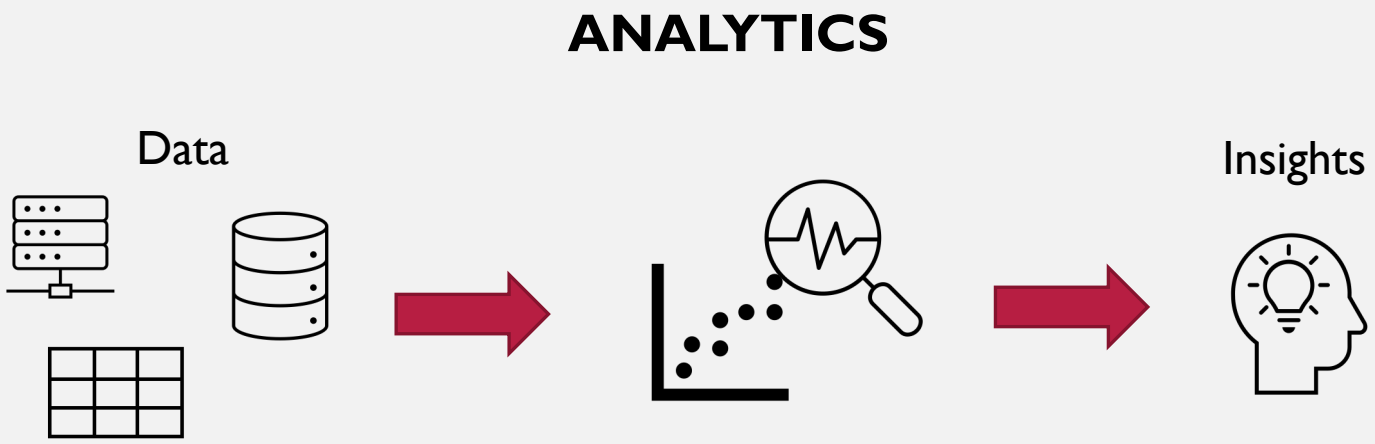
BIG DATA


Data Velocity
Real Time
Near Real Time
Periodic
Marketing Automation
CMS
SMS
Audio
Video
Reports
Batch
Social Tech
Table
KB
Web
Data Base
MB
Automated Demand Generation
photo
GB
TB
Mobile App
PB
Data Variety
Data Volume
Fraud detection;
Social media monitoring;
Coupon promotion at POS

DEVELOPMENT OF MOBILE COMMUNICATION NETWORK

	1G				2G			3G		4G		5G
Technology standard	AMPS	NMT	TACS	C-Netz	GSM	D-AMPS	IS-95 A	UMTS	CDMA2000	LTE		NR
Digital or not?	Analogue				Digital			Digital		Digital		Digital
Launch year (approx.)	~1980				~1990			~2000		~2010		~2020
Enhancements					GPRS			HSPA	EVDO Rev. 0	LTE-Advanced		
					EDGE		IS-95 B	HSPA+	EVDO Rev. A	LTE-Pro		
									EVDO Rev. B			
Services	Voice only				Voice + SMS + Data (Mobile Internet)							
Peak download speeds	-				GPRS	171.2 kbps		UMTS	2 Mbps	LTE	300 Mbps	10 Gbps
								HSPA	14.4 Mbps			
					EDGE	384 kbps		HSPA+	42 Mbps	LTE-A	1 Gbps	
					IS-95 A	14.4 kbps		CDMA2000	153 kbps			
					IS-95 B	115 kbps		EVDO 0	2.4 Mbps	LTE-Pro	3Gbps	
								EVDO A	3.1 Mbps			
								EVDO B	14.7 Mbps			

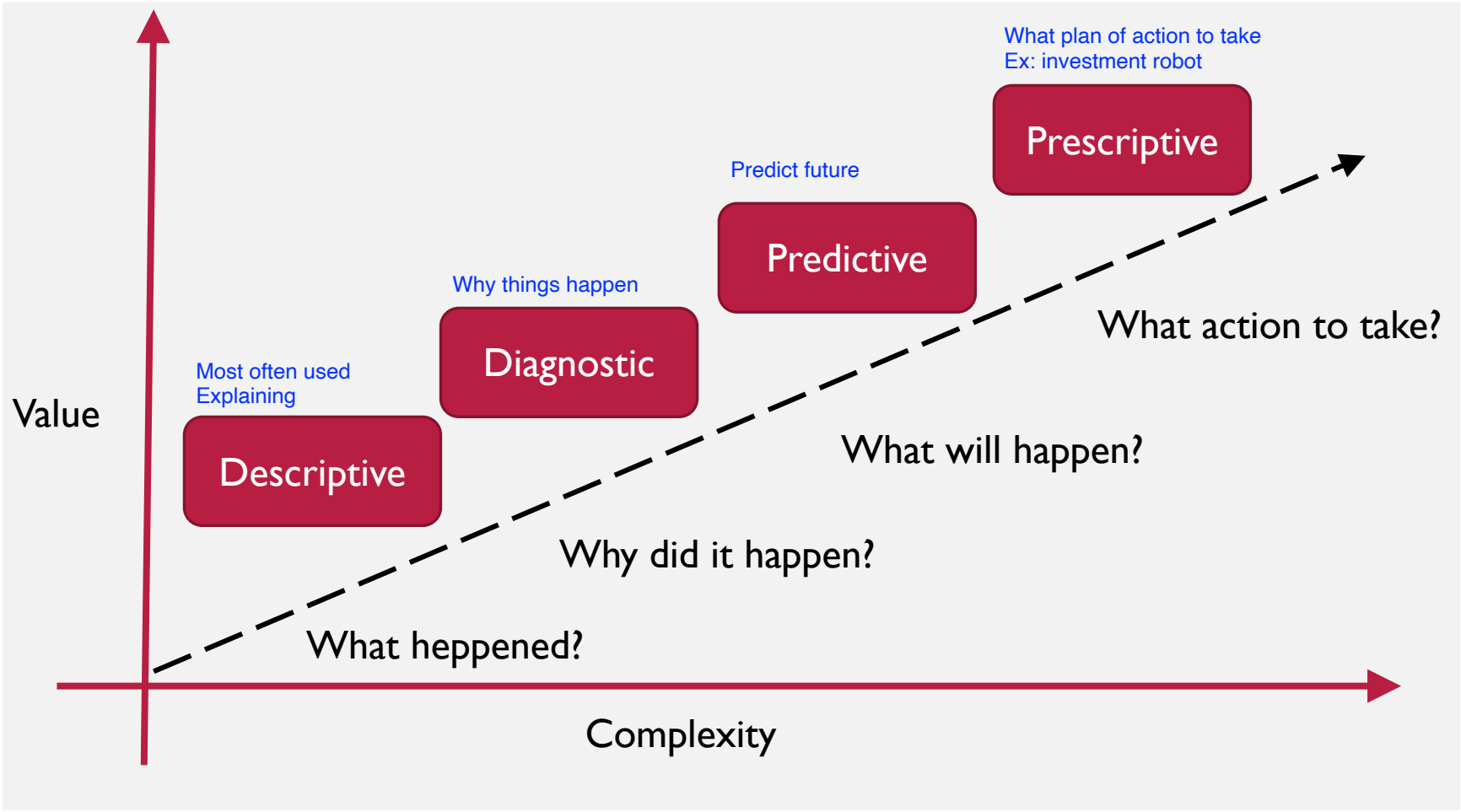
WHAT TO DO WITH THESE DATA?



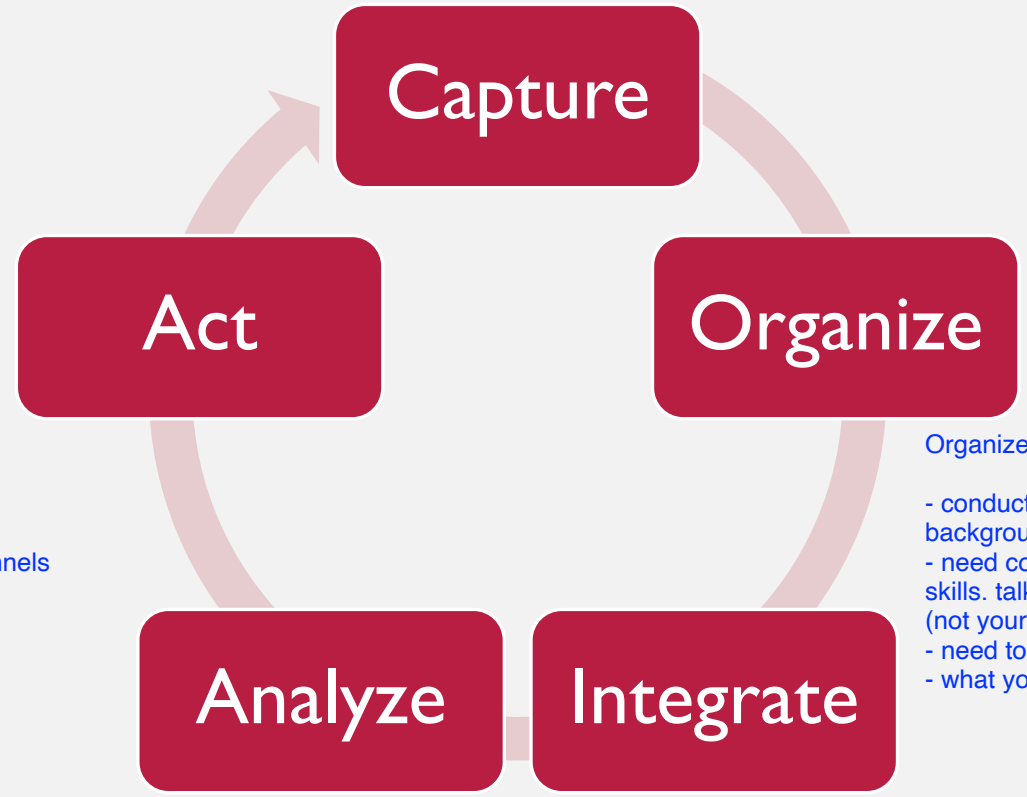
You have lots of data, but they are to messy. And the useful data aren't being analyzed.
Not all business problems can be solved with big data, because they may not be valid.

In 2012, 22% of all data was useful, but only 0.5% was analyzed. Useful data will grow to 37% in 2020. *(Source: The Guardian)*

FOUR TYPES OF BIG DATA ANALYTICS



CYCLE OF BIG DATA MANAGEMENT



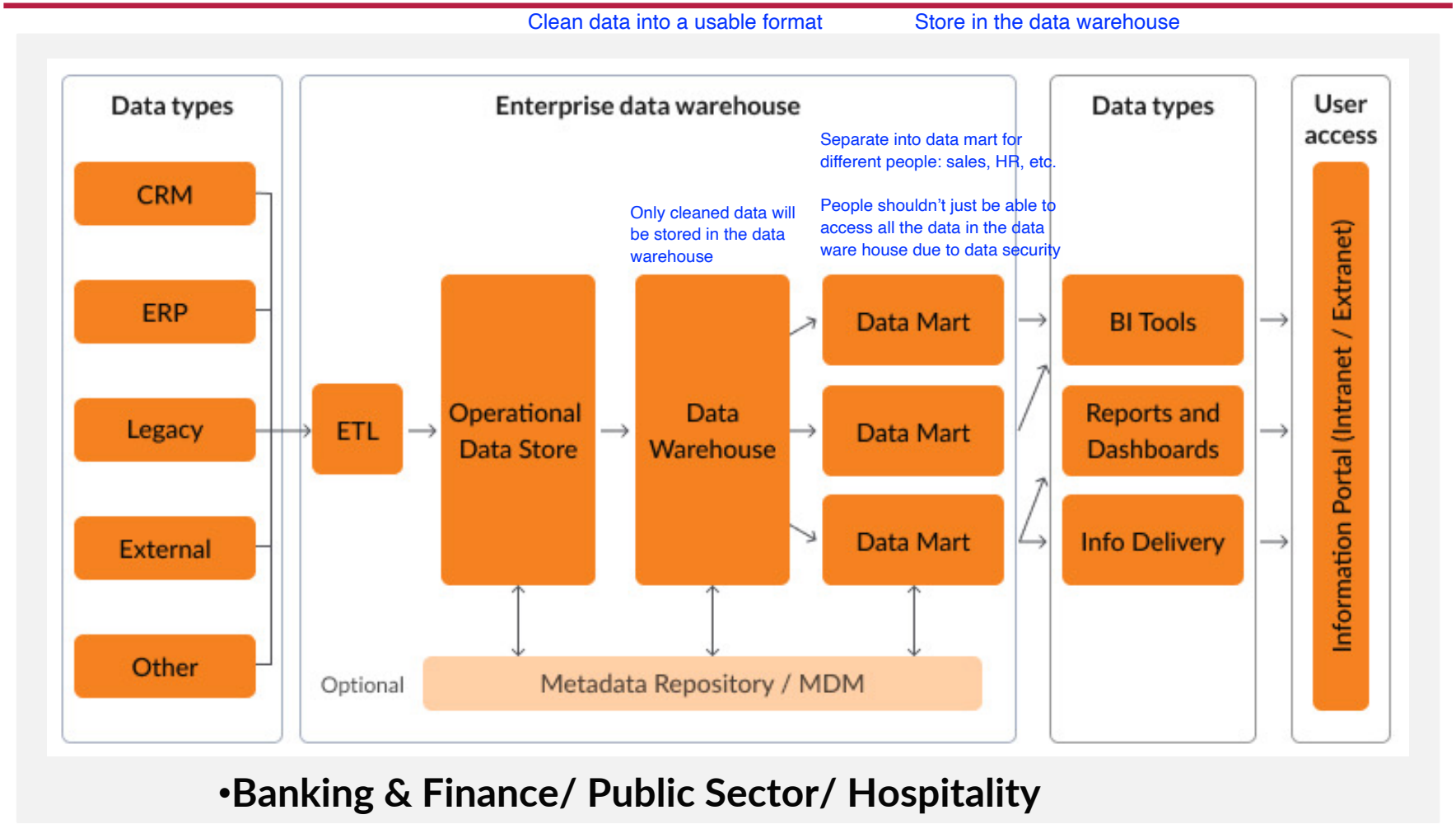
Capture, act

- conducted by business personnels

Organize, integrate, analyze

- conducted by people with engineering backgrounds
- need communication and presentation skills. talk based on what you have found (not your imagination).
- need to be proven/supported by the data
- what you have, why you want to do

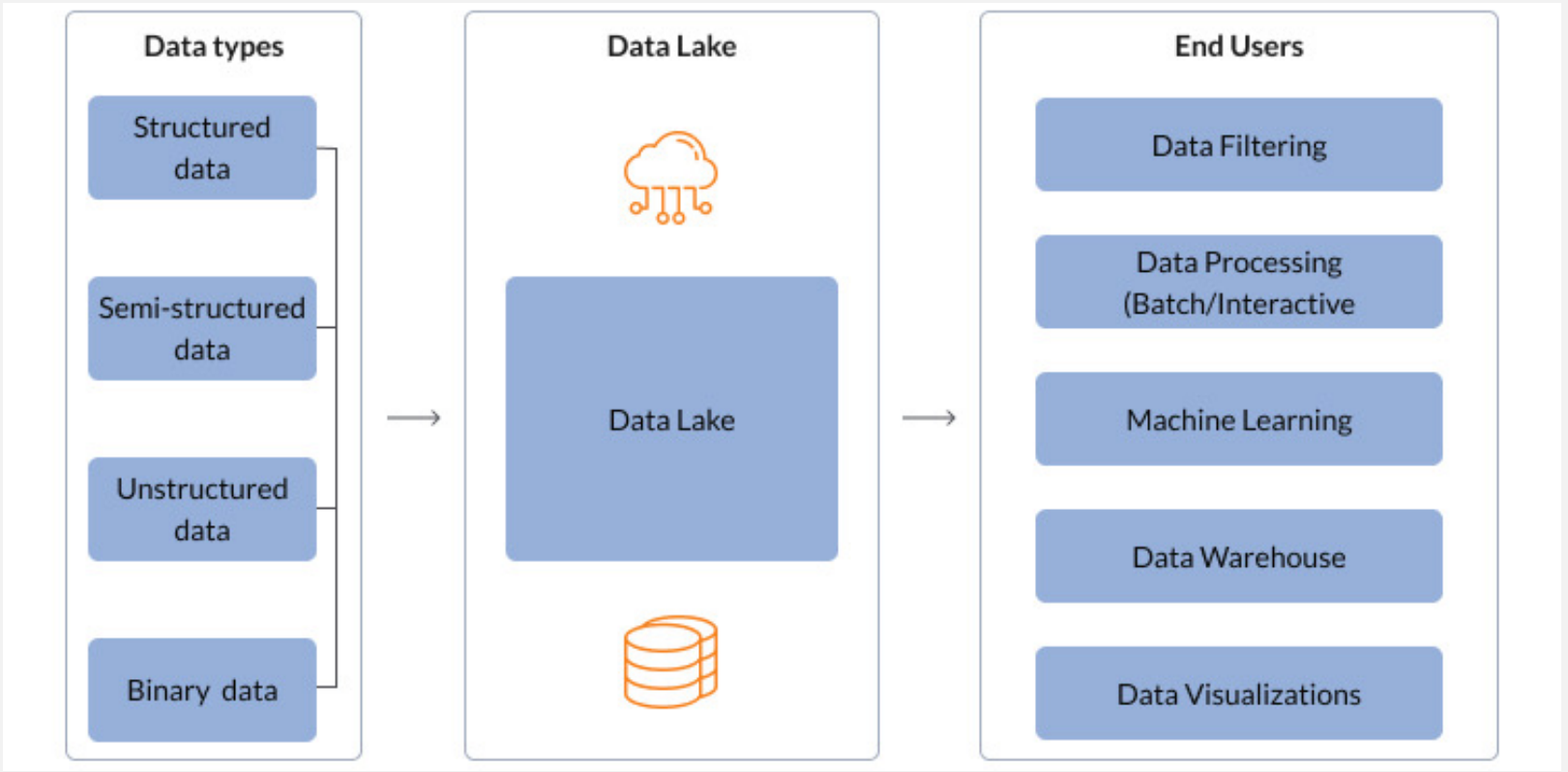
DATA WAREHOUSE
-- ETL (EXTRACT TRANSFORM LOAD)



•Banking & Finance/ Public Sector/ Hospitality

- Since storage is cheap, why not store every kinds of data?
- Some people might want some data that is usually unstructured and can't be found in data warehouses
- Data warehouses need data to be structured and formatted first -> slow
- To overcome this -> data lake
- Install all the data in the data lake (not cleaned yet), and only transform and use the data as needed. The end user will extract the data and clean it before using it
- Data warehouse is easier to use for the end-users
- Data lake used by data scientists, need to clean the data by themselves, project-based
- But if you need to generate some report often, you might need to build a data warehouse
- Big data, AI-related usually utilizes data lake

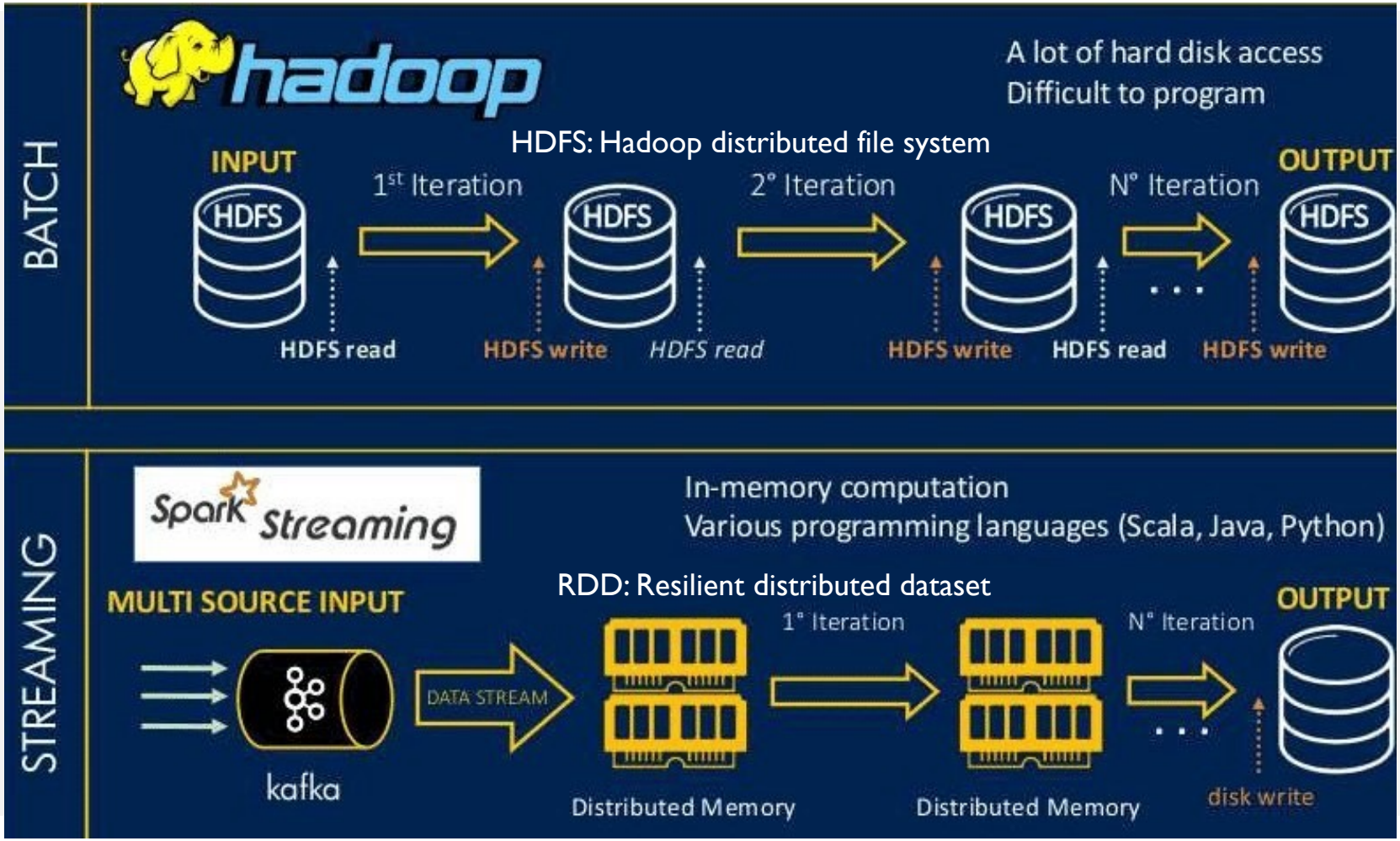
DATA LAKE
-- ELT (EXTRACT LOAD TRANSFORM)



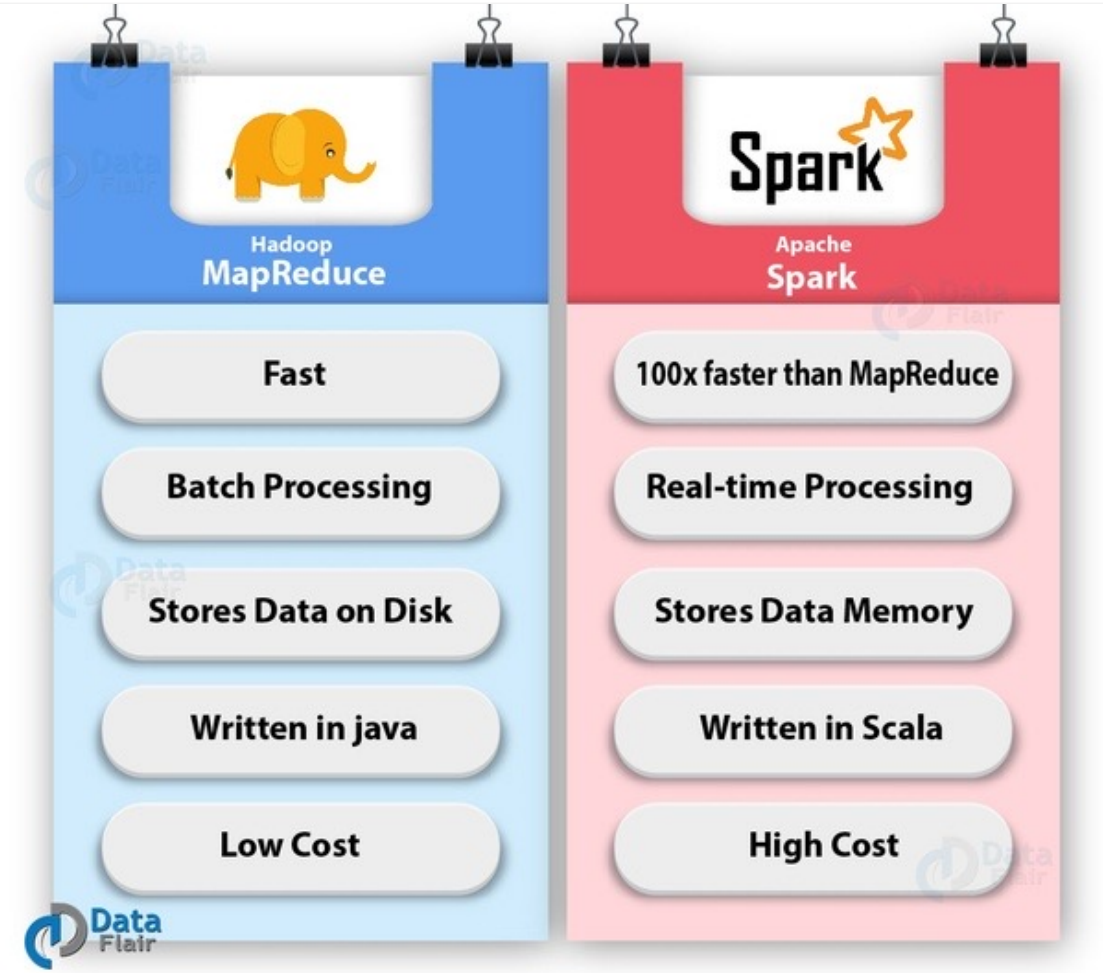
•Healthcare/ Education /Transportation

Hadoop	Map data -> store data in the hard disk Problem wit Hadoop: need to read and load to/from hard disk too often. Slow as it needs to access the hard disk often
Spark	In-memory processing. Much faster than Hadoop. Can do real-time processing. RAM is very big (expensive)

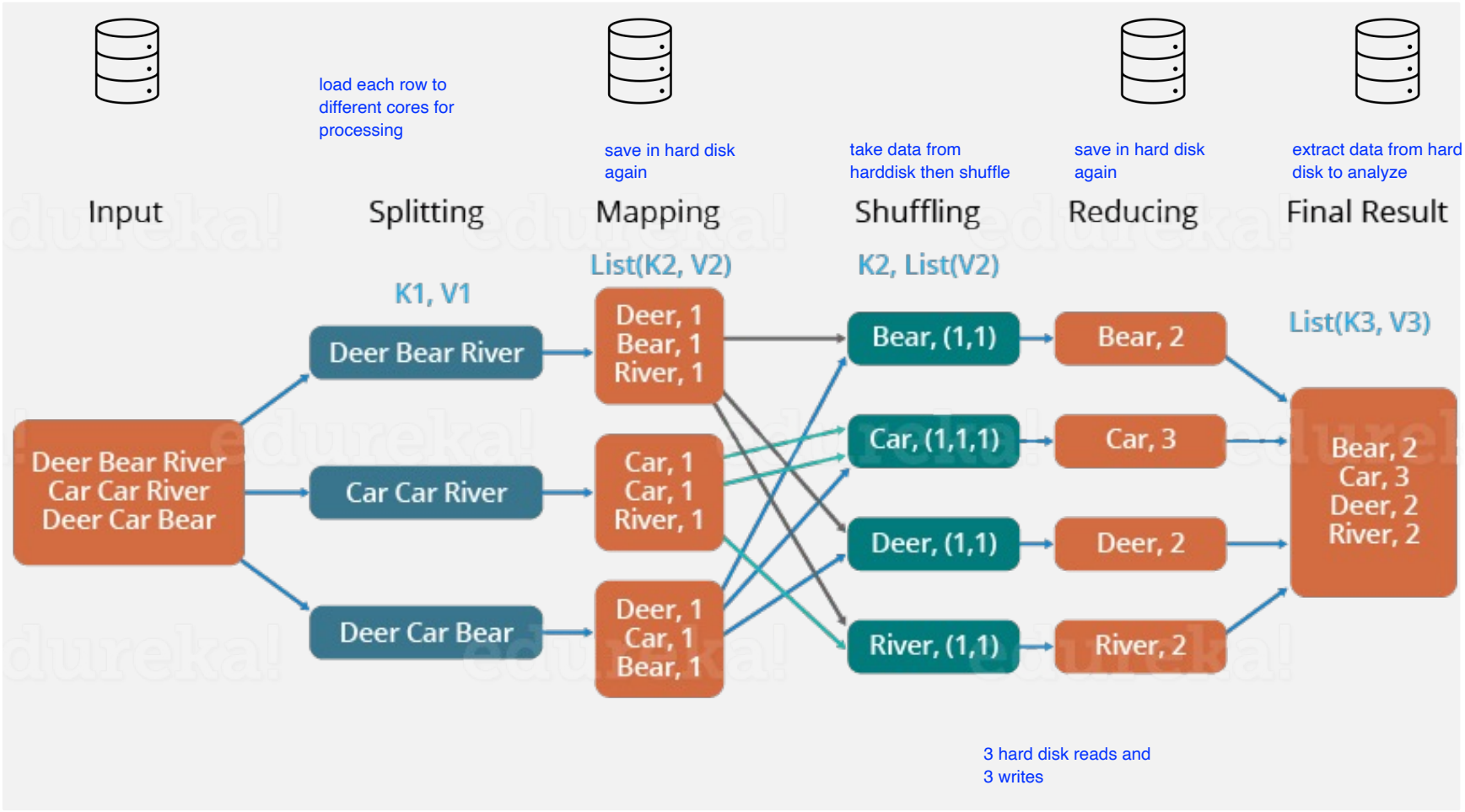
HADOOP VS. SPARK



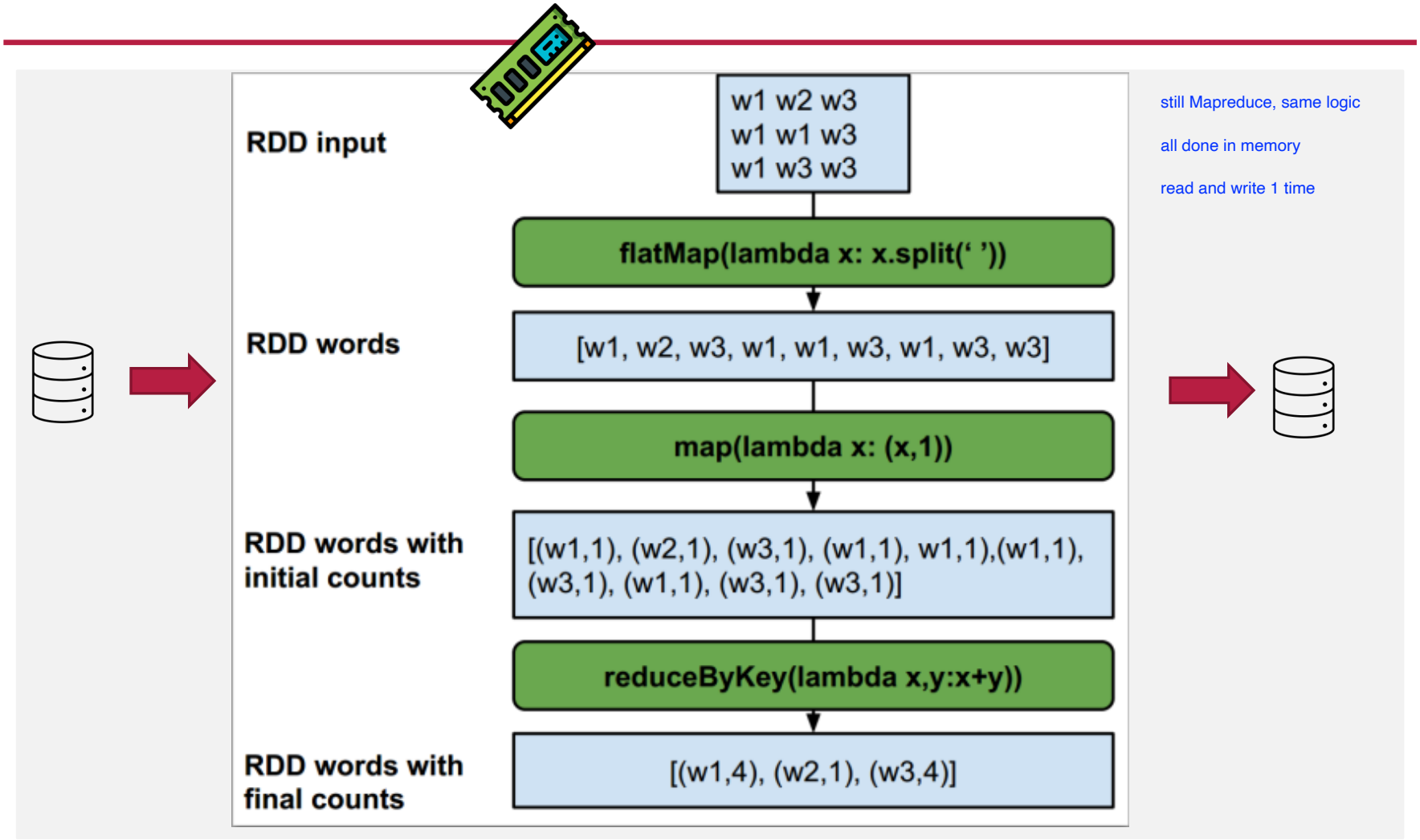
HADOOP VS. SPARK CONT.



WORD COUNT IN HADOOP MAPREDUCE



WORD COUNT IN SPARK



BIG DATA ANALYTICS -- AI/DATA SCIENCE/MACHINE LEARNING

- Reporting and Dashboard
- Visualization
- Analytics and Advanced Analytics

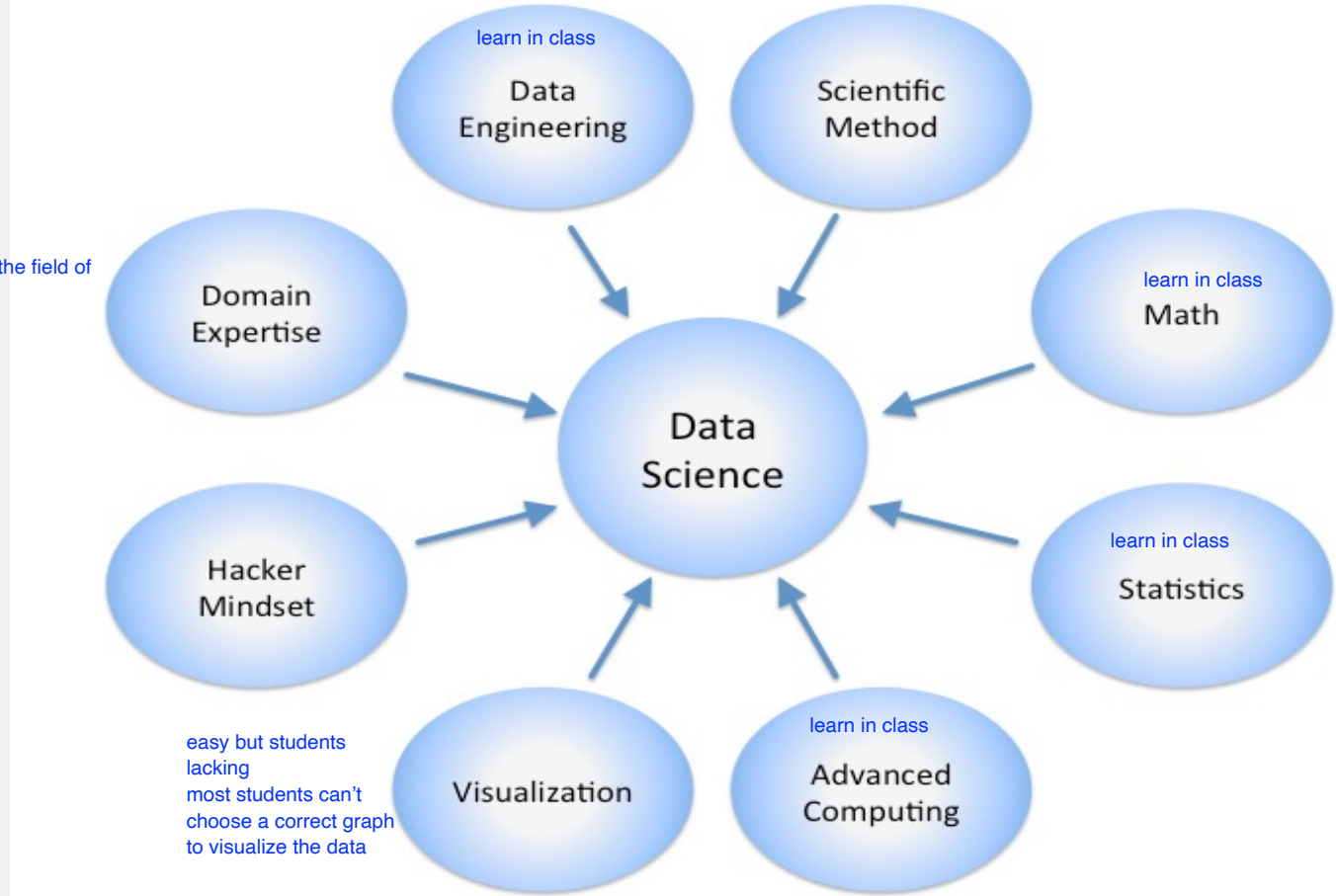
WHAT IS DATA SCIENCE?

- An area that **manages, manipulates, extracts, and interprets knowledge** from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

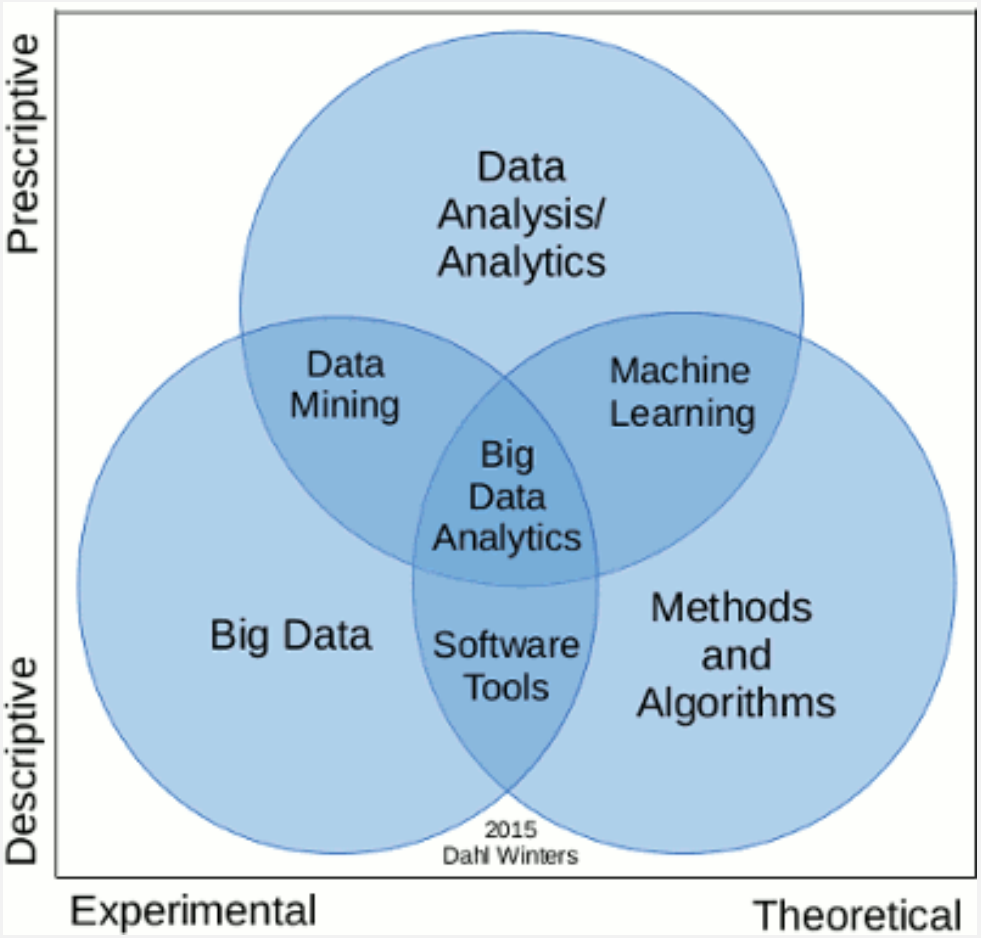
WHAT IS DATA SCIENCE? CONT.

- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

DATA SCIENCE SKILLSETS



FIELDS OF DATA SCIENCE



BIG DATA APPLICATIONS

- **Automotive**: Auto sensors reporting location, problems
- **Communications**: Location-based advertising
- **Shopping**: Sentiment analysis; Marketing
- **Life Science**: Clinical Trials Genomics
- **Entertainment**: Viewers/ advertising effectiveness
- **Utilities**: Smart meter analysis for network capacity
- **Banking**: Fraud analysis; AI-supported Investment
- Election: Obama…

If you want to take a leader role, you need to master all these roles

NEED A TEAM FOR BIG DATA

highest-paying

- **Data Scientist**: who has augmented math and statistics background with programming to analyze data and create applied math models for the real usage.
- **Data Engineer**: who has specialized their skills in creating software solutions around big data. SQL, API, ETL, ELT, Good in software
- **Operation Engineer**: who with an operational or systems engineering background who has specialized their skills in big data operations, understands data, and has learned some programming.

DATA SCIENTISTS

- Data Scientist
  - The Sexiest Job of the 21st Century
- They find stories, extract knowledge.
- They are not reporters

AI AS A TOOL FOR DATA SCIENTIST

- **Computers** are fundamentally well suited to performing mechanical Computations that used Fixed programmed rules.
- **Machines** performs simple task efficiently and reliably, which humans are ill-suited to.
- For more complex problems, things get more difficult. Unlike humans, **computers** have trouble understanding specific situations, and adapting to new situations.

AI AS A TOOL FOR DATA SCIENTIST CONT.

- **Artificial intelligence** aims to improve machine behavior in tackling such Complex.
- AI is used for analyzing the data.

AI AS A DISCIPLINE

- As a discipline, AI is **NOT** primarily bonneted to a knowledge domain, but to a purpose:

  Conceiving artificial systems that are intelligent

- Intelligence involves certain mental activities:
  - Learning
  - Reasoning
  - Understanding
  - Grasping Truths
  - Seeing Relationships
  - Considering Meaning
  - Separating fact from belief

SUBSETS OF AI

- **Machine Learning**: machines learn by themselves
  - Neural Network, Deep learning
- **Evolutionary Computation**: optimization
- **Neutral Language Processing**
- **Image Processing**
- Agent:
  - Software Agent vs Robot Agent

AI IN LIFE


Hey Siri


Netflix
ครองบัลลังก์ทีวี
ด้วยพลังอัจฉริยะแห่ง AI


TESLA


EASY INVEST
ROBO ADVISOR
สร้างพอร์ตกองทุนรวมอัตโนมัติ
ตอบทุกเป้าหมายการลงทุน

DEFINITIONS OF AI:

- **Weak** AI: intelligent actions or reasoning in some situations limited

  - image processing -> for what -> coded differently
  - applies only for a limited use

  - sense or ‘scan’ for things that are like what they already know and classify them accordingly.
- **Strong** AI: mental/thought capabilities equal to (or better than) human Equivalent or better than human
  - use clustering and association to process data.
  - E.g.: Go AI player

WHEN WILL COMPUTERS BECOME TRULY INTELLIGENT?

- To date, all the brains of human intelligence have not been captured and applied together to spawn an intelligent artificial creature.

probably not in the next 5-10 years

- Currently, AI rather seems to focus on lucrative domain specific applications, which do not necessarily require the full extent of AI capabilities.
- There is little doubt among the community that artificial machines will be capable of intelligent thought soon.

HISTORY OF AI

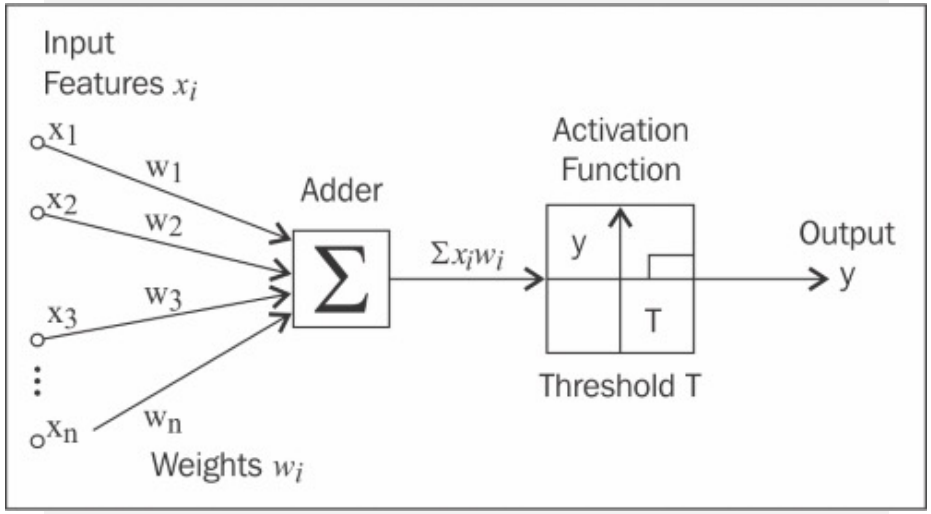
The gestation of artificial intelligence (1943-1955)

- 1923 *Karel Capek* used the word “Robot” in English.
- The first work that is now generally recognized as AI was done by *Warren McCulloch and Walter Pitts (1943).*
- They drew on three sources:
  1) knowledge of the basic physiology and function of neurons in the brain
  2) a formal analysis of propositional logic
  3) Turing’s theory of computation

HISTORY OF AI

The gestation of artificial intelligence (1943-1955)

- They proposed a model of artificial neurons in which each neuron is characterized as being “on” or “off,” with a switch to “on” occurring in response to stimulation by a sufficient number of neighboring neurons



McCulloch, Warren S. and Walter Pitts. 1943. "A Logical Calculus of the
Ideas Immanent in Nervous Activity." Bull. Math. Biophysics, 5, 115-133.

HISTORY OF AI


John McCarthy (1927-2011)

The birth of AI (1956)

- *John McCarthy* organized a workshop in the summer of 1956.
- A remarkable group of ~20 scientist and engineers, including:
  - John McCarthy (LISP language, situation calculus, non-monotonic logics)
  - Marvin Minsky (frames, perceptron, society of minds)
  - **Herbert Simon** (logic theorist, general problem solver, bounded rationality)
  - Allen Newell (logic theorist, general problem solver, the knowledge level)
  - Ray Solomonoff (father of algorithmic probability, algorithmic information theory)
  - Arthur Lee Samuel (first machine learning algorithm for checkers)
  - W. Ross Ashby (pioneer in cybernetics, law of requisite variety)
  - **Claude Shannon** (father of information theory)
  - **John Nash** (father of game theory)

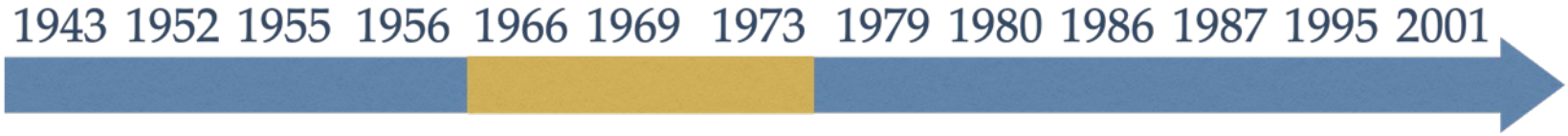
HISTORY OF AI

Early enthusiasm, great expectations (1952-1969)

- Many applications have been developed in this period
- **General Problem Solver (GPS)** was probably the first program to embody the “thinking humanly” approach.
- High-level language **LISP**, which was to become the dominant AI programming language for the next 30 years.
- **Advice Taker**, the first complete AI system designed to use knowledge to search for solutions to problems, and was also designed to accept new axioms in the normal course of operation, thereby allowing it to achieve competence in new areas without being reprogrammed.

GPS: CROSSING THE STREET PROBLEM

- **Orient**: Gather information: amount of traffic, traffic light, sirens? How does traffic work? Do cars stop for lights? Do bikes? Do pedestrians?
- **Plan**: Consider various approaches. What if I jaywalk? What if I walk to the corner and cross?
- **Execute**: Cross the street.
- **Check**: Did I make it safely? Were there any unexpected hazards? Should I keep using this approach or modify it next time?

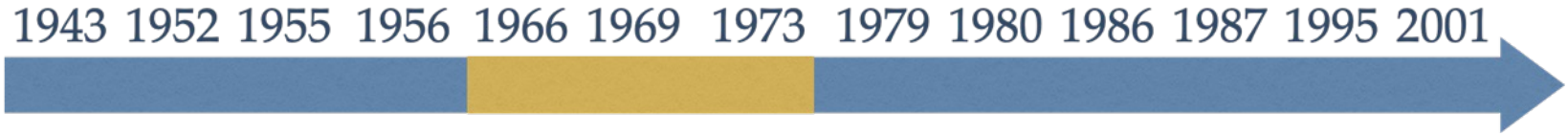


HISTORY OF AI

A dose of reality (1966-1973)

- Researchers' overconfidence was due to the promising performance of early AI systems on simple examples.
- In 1965 *Joseph Weizenbaum* at MIT built **ELIZA**, an interactive problem that carries on a dialogue in English.

Human: Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here?
Human: He says I'm depressed much of the time.
ELIZA: I am sorry to hear you are depressed.
Human: It's true. I'm unhappy.
ELIZA: Do you think coming here will help you not to be unhappy?[8]



HISTORY OF AI

A dose of reality (1966-1973)

- The early systems turned out to fail miserably when tried out on wider selections of problems and on more difficult problems.
- Reasons:
  - Most early programs knew nothing of their subject matter.
  - The early AI programs solved problems by trying out different combinations of steps until the solution was found.
  - There are some fundamental limitations on the basic structures being used to generate intelligent behavior.

HISTORY OF AI

Knowledge-based systems: The key to power? (1969-1979)

- **DENDRAL** (1965) was the first successful knowledge-intensive (chemical-analysis) system. Its expertise derived from large numbers of special-purpose rules.
- With about 450 rules, **MYCIN** as an expert system to identify bacteria causing severe infections, was able to perform considerably better than junior doctors.

HISTORY OF AI

AI becomes an industry (1980-present)

- Overall, the AI industry boomed from a few million dollars in 1980 to billions of dollars.
- In 1982, the first successful commercial expert system, R1/XCON, helped configure orders for new computer systems to save cost for $40 million a year.
- In 1988, hundreds of companies building expert systems, vision systems, robots, and software and hardware specialized for their purposes.

HISTORY OF AI

The return of neural networks (1986-present)

- Reinvention of the back-propagation learning algorithm
- As occurred with the separation of AI and cognitive science, modern neural network research has bifurcated into two fields:
  - creating effective network architectures and algorithms and understanding their mathematical properties
  - careful modeling of the empirical properties of actual neurons and ensembles of neurons

HISTORY OF AI

AI adopts the scientific method (1987-present)

- It is more common to build on existing theories than to propose brand-new ones, to base claims on rigorous theorems or hard experimental evidence rather than on intuition, and to show relevance to real-world applications rather than toy examples
- AI has finally come firmly under the scientific method.
- To be accepted, hypotheses must be subjected to rigorous empirical experiments, and the results must be analyzed statistically for their importance (Cohen, 1995).
- It is now possible to replicate experiments by using shared repositories of test data and code.

HISTORY OF AI

The emergence of intelligent agents (1995-present)

- One of the most important environments for intelligent agents is the Internet.
- AI systems have become so common in Web-based applications that the “-bot” suffix has entered everyday language (e.g., search engines, recommender systems, and Web site aggregators).
- Driving a car, playing chess, or recognizing speech.

HISTORY OF AI

The availability of very large data sets (2001-present)

- It makes more sense to worry about the data and be less picky about what algorithm to apply.
- The performance of filling in holes in a photograph algorithm was poor when they used a collection of only ten thousand photos but crossed a threshold into excellent performance when they grew the collection to two million photos (Hays and Efros, 2007).
- The problem of how to express all the knowledge that a system needs—may be solved in many applications by learning methods rather than hand-coded knowledge engineering, provided the learning algorithms have enough data to go on (Halevy et al., 2009).

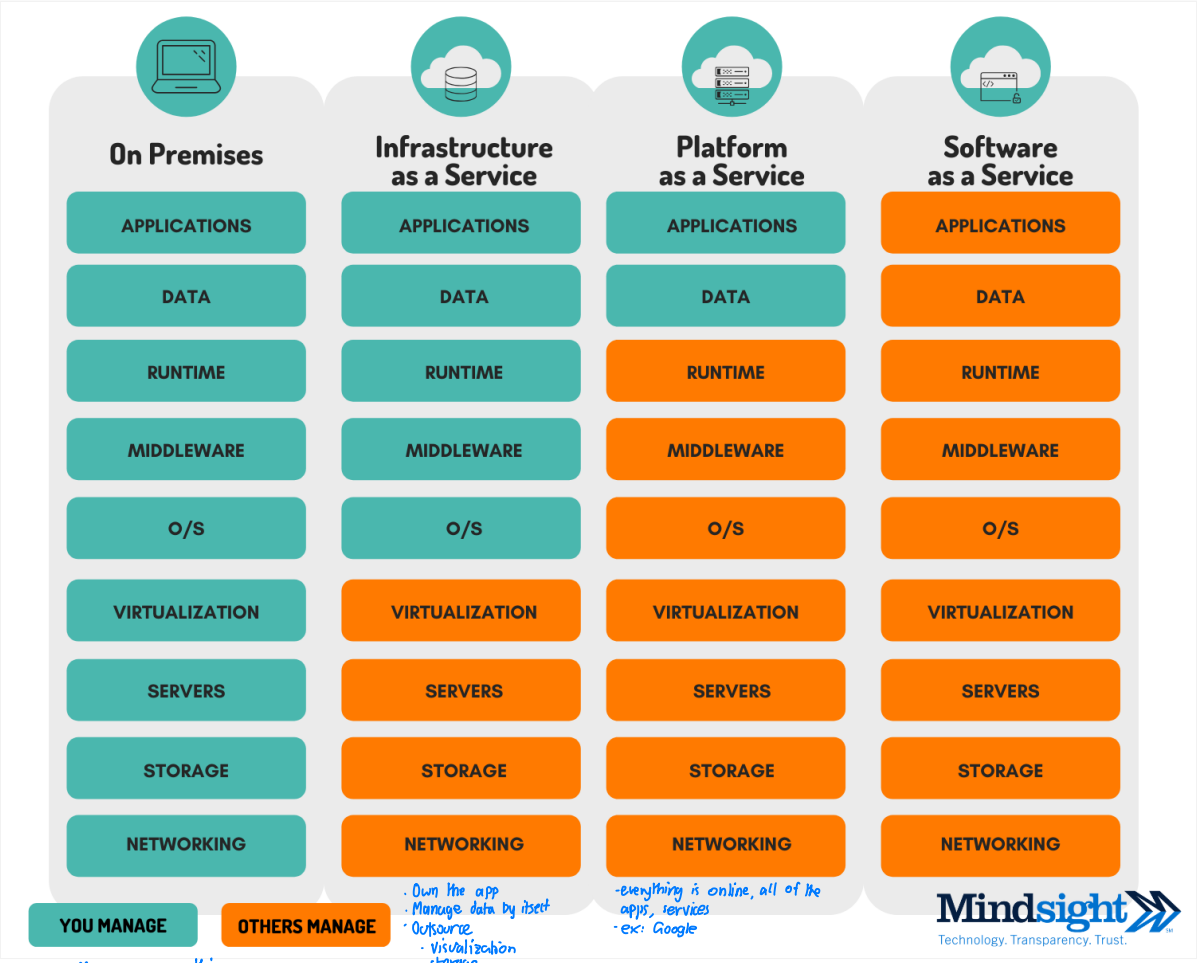
HW1:

Tell me the name of representative **cloud AI service** the following companies provide

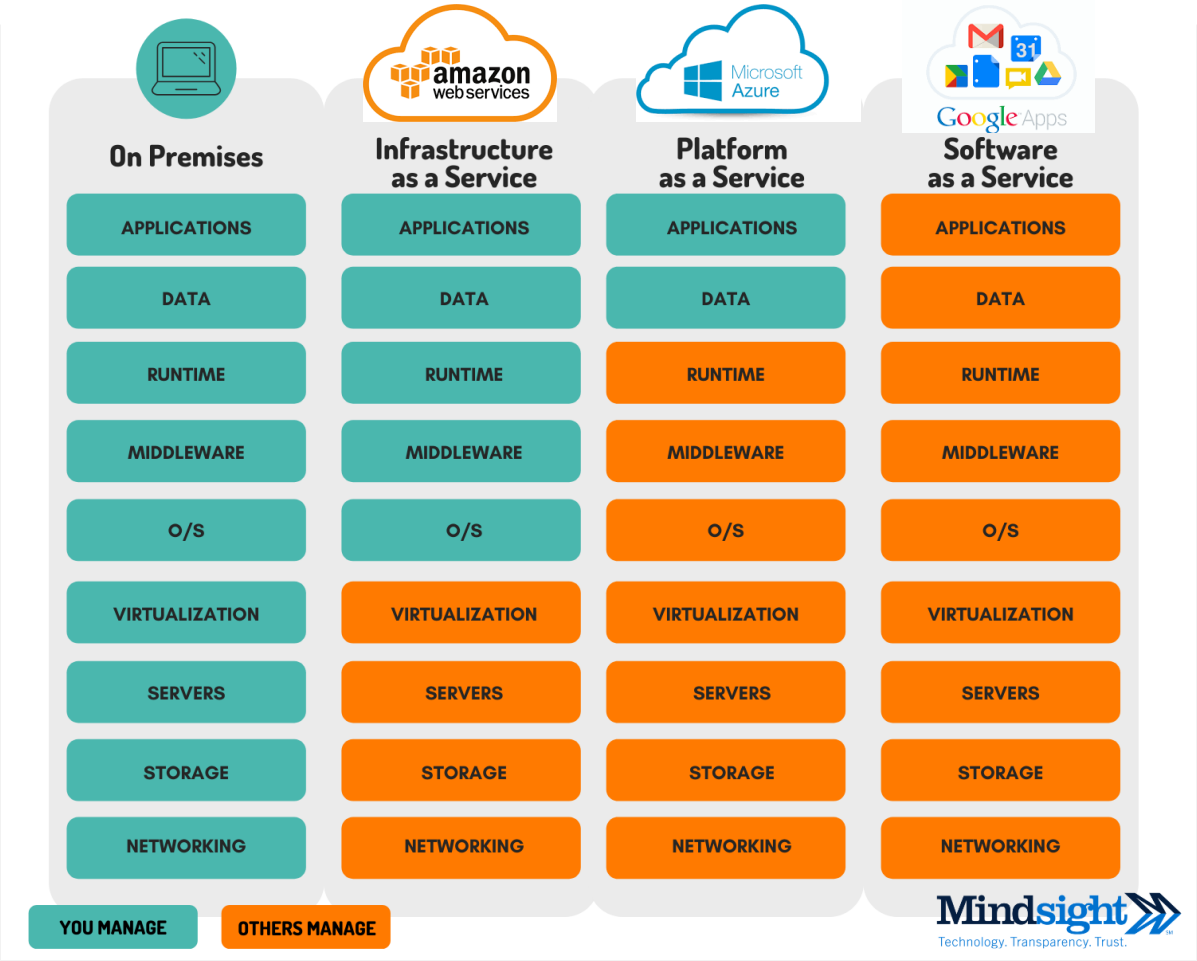
- Amazon
- IBM
- Microsoft
- Google
- Alibaba
- Line

See HW 1

IAAS VS. PAAS VS. SAAS

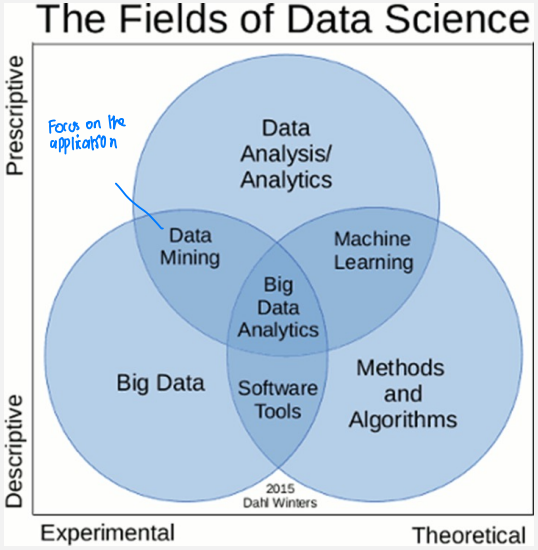


IAAS VS. PAAS VS. SAAS

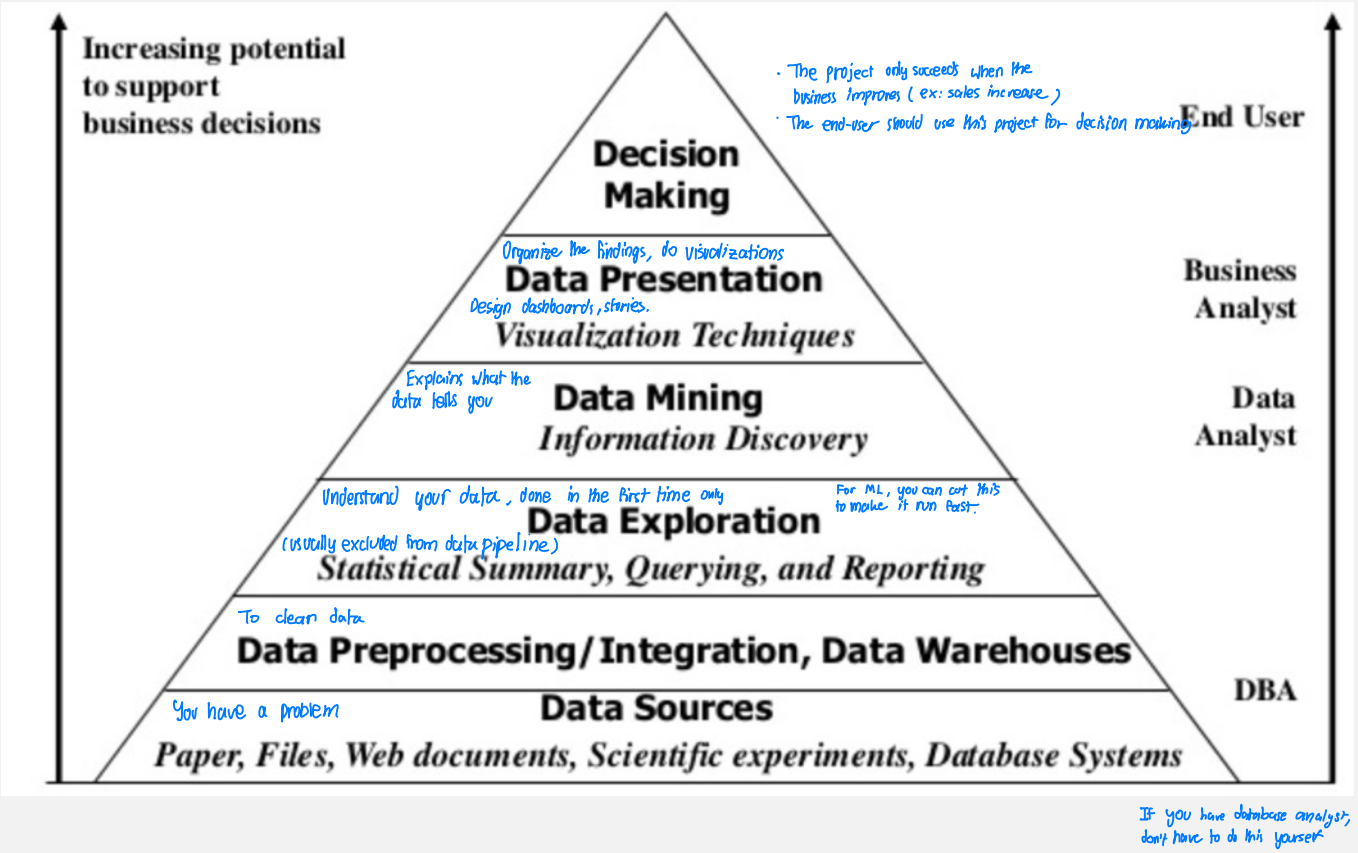


DATA MINING

- Data mining is the computing process of **discovering** patterns in large **data sets** involving methods at the intersection of machine learning, statistics, and database systems.

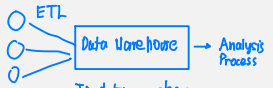


DATA MINING AND BUSINESS INTELLIGENCE



Exploiting Data

Spotting Data

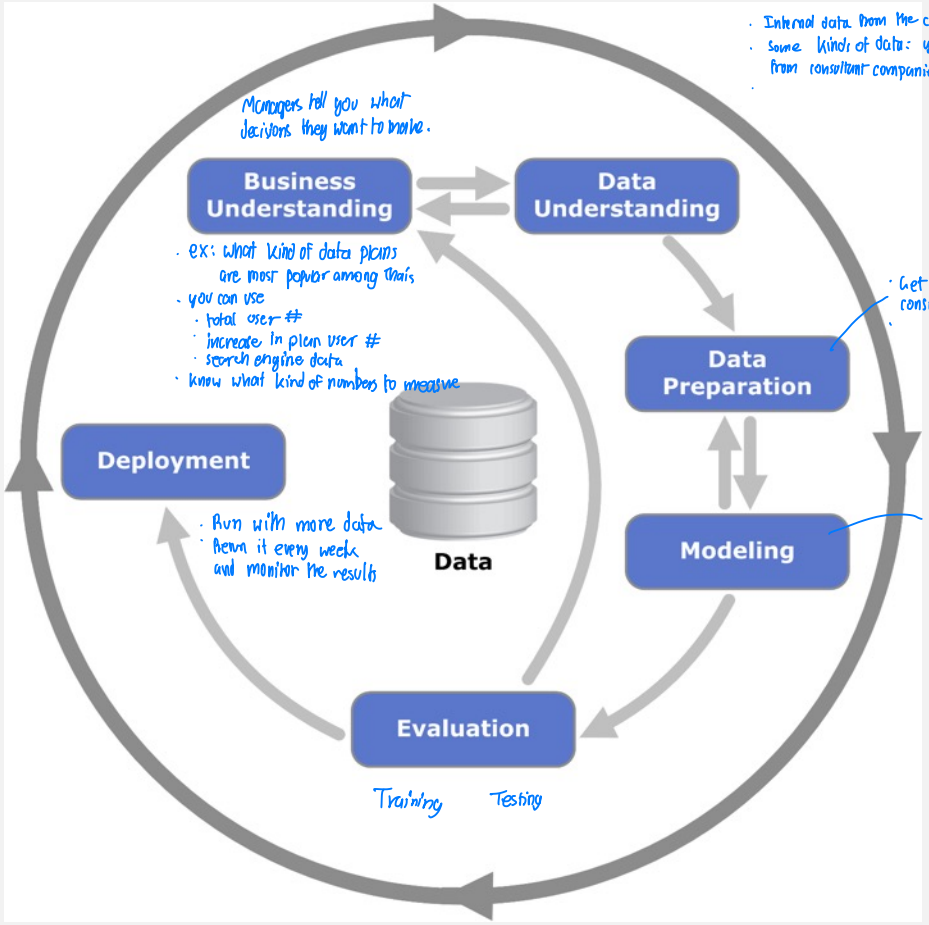


In data warehouse
- limited, but always available
- need to find the source
  - ex: need to pay Google to get search engine data
  - takes time

Data Lake
- ELT
- Maintain it to be updated
- Data are prepared
- Clean data only when it will be used.

Data Warehouse
- ETL
- Organized data
- Transformed data

CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING: CRISP-DM



- Takes the longest in the real world
- Interns don't have access, or was already assigned data and managers explain thoroughly

1. BUSINESS UNDERSTANDING

- What are the desired outputs of the project?
  - Objectives in business terminology
  - Increase catalogue sales to existing customers.
- Assess the current situation
  - Fact-finding about all of the resources, constraints, assumptions, etc.
- Determine data mining goals
  - Objectives in technical terms
  - Predict how many widgets a customer will buy
- Produce project plan

2. DATA UNDERSTANDING

- Describe data
  - Format, quantity (for example, the number of records and fields in each table), the identities of the fields and any other surface features which have been discovered.
  - Evaluate whether the data acquired satisfies your requirements.
- Explore data
  - Simple statistical analyses
  - Distribution of key attributes (for example, the target attribute of a prediction task)
- Verify data quality
  - Complete (cover all the cases required)
  - Correct
  - No missing values
- Data quality report

FROM BUSINESS PROBLEMS TO DATA MINING TASKS

- A critical skill in data science is **the ability to** decompose a data analytics problem into pieces such that each piece matches **a known task** for which **tools** are available.

EXE 1: HOW DOES A SUPERMARKET USE DATA ANALYTICS TO PREDICT PREGNANCIES

- Business understanding:
  - Target marketing towards pregnant women
- Data understanding:
  - Membership : age , date , etc
  - Shopping History
  - Pregnancy?

what kind of data helps identify who are pregnant?
· product purchase : teethers, supplements, milk, etc...
· transaction database

transaction history

· some key parameters might not be readily available in the database
· need to gather by yourself → time consuming
many projects dropped due to this.

3. DATA PREPARATION

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Combine data from multiple sources
- Data reduction
  - Data cute aggregation, attribute selection, data compression, etc.
- Data transformation
  - Generalization, normalization, discretization, etc.

4. MODELING

- Select modeling technique
- Generate test design
  - Training, test and validation datasets
- Build model
  - Parameter settings ex: try different learning rate
- Assess model
  - Model assessment
  - Revised parameter settings

5. EVALUATION

- Evaluate your results
  - Summaries assessment results in terms of business success criteria
  - Whether the project already meets the initial business objectives

ex slow but high accuracy
fast but low accuracy

- Review process
  - Did we correctly build the model?
  - Did we use only the attributes that we are allowed to use and that are available for future analyzes?

social model: < 50%. fine, but need to evaluate whether its model or data problem and how to solve it.

- Determine next steps

6. DEPLOYMENT

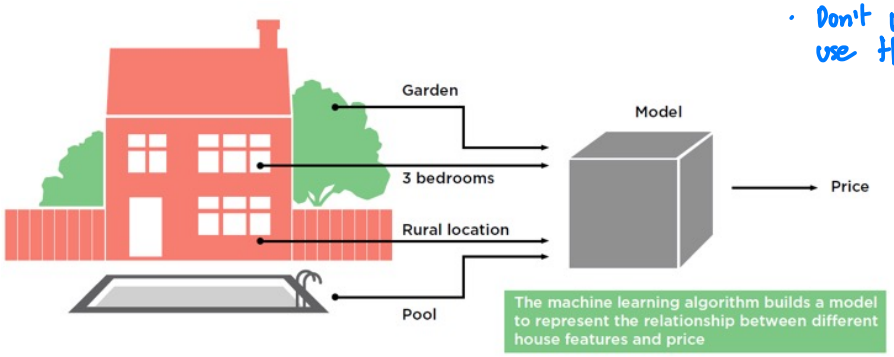
- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project

CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING: CRISP-DM

MACHINE LEARNING

- Machine learning algorithms **learn** to predict outputs based on previous examples of **relationships between input data and outputs** (called training data).
- A model of the relationship between inputs and outputs is **gradually improved** by testing its predictions and correcting when wrong (in testing data).
- Machine learning is a set of computerized techniques for recognizing patterns in data. It's useful to automate this process when the data has many features and is very complex.

EXAMPLE: HOUSE PRICE PREDICTION



- Don't use common sense to make a model, use the data (don't assume)

- There is no simple relationship between size, functions, location and price.
- A machine learning can learn through details of **thousands or millions** of houses for sale to model the relationship between different **factors and price**.

MACHINE LEARNING (2)

- Using these machine learnings bypasses **the need to write specific code** to solve each specific problem.
- Each machine learning can be used to solve lots of different problems by adapting the model to **fit** different **data sets**.

TYPE OF PROBLEMS

1) Regression
2) Clustering
3) Classification
4) Data Reduction
5) Similar Matching
6) Game, etc.

Focus on these 3 types

Exam will ask to identify the problem type.

	Independent Data, Input, Features ↑ x_1, x_2	Dependent Data, output, labeled y
Train	✓	✓
Test	✓	✓ ↔ $\hat{y}$

need numeric y value in your data

$\hat{y}$: numeric ⌶ continuous / discrete →

- feels that you can't use regression for like Likert scale cos you end up with 2.75... what does it mean?
- → score 0 → 100 discrete is more suitable than 1-5 Likert scale, not recommended
- → at least more than 10 discrete val is ok

For all dataset, you must have test and train

REGRESSION

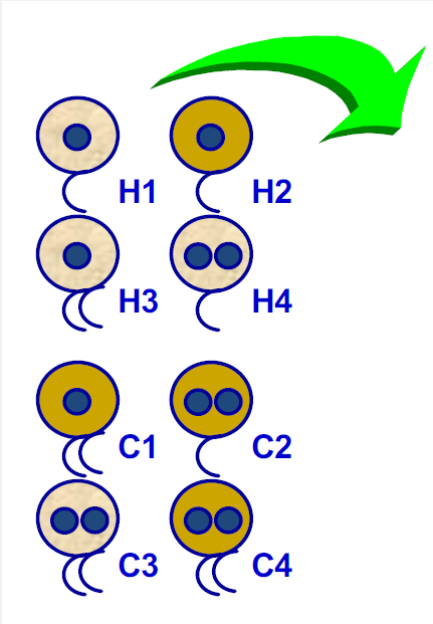
- Data is given a real value rather than a label attribute.
- Machine learning predicts values for new data.
  - House prices based on the historical local market data
  - Price of a stock or a market over time
  - Population increases in next 10 years

Time series can be regression

CLASSIFICATION

- Labeled **data** is used to train the algorithm so it can **predict** the label to attach to **new** unlabeled **data**.
- The algorithm is effectively modeling the differences and similarities between groups or classes.
  - Spam email filtering : spam & unspam
  - Handwriting recognition : limited choice → classification
  - Classification might be used with the house market data to find rural properties, when they haven't been **labeled** as such. Having two or more of a bundles of features like 'farmland', 'near village' or 'own water supply' may together predict whether a property is **rural** or not. classify house prices in different regions

CLASSIFICATION



Label

ID	color	#nuclei	#tails	status
H1	light	1	1	healthy
H2	dark	1	1	healthy
H3	light	1	2	healthy
H4	light	2	1	healthy
C1	dark	1	2	cancerous
C2	dark	2	1	cancerous
C3	light	2	2	cancerous
C4	dark	2	2	cancerous

Descriptive attributes

Color: {dark, light}, #nuclei: {1, 2}, #tails: {1, 2}

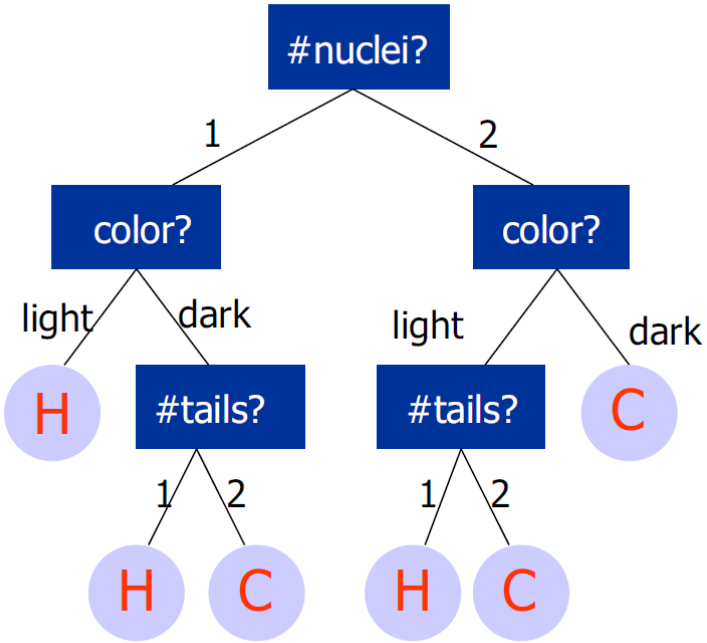
Class attribute

Status {cancerous, healthy}

CLASSIFICATION

ID	color	#nuclei	#tails	status
H1	light	1	1	healthy
H2	dark	1	1	healthy
H3	light	1	2	healthy
H4	light	2	1	healthy
C1	dark	1	2	cancerous
C2	dark	2	1	cancerous
C3	light	2	2	cancerous
C4	dark	2	2	cancerous

Decision tree



CLUSTERING

- Deals with categorical data
- Clustering don't require output
- Only has X, doesn't have y
- separate data into cluster 1,2,3 ...
  then you gotta see what the clustering result means.
- y it provides is based on the similarity of X.

- choose X that the business cares about

- **Data is unlabeled** but can be **divided into** groups based on similarity or other measures of structure within the data.
- The algorithm tries to find the hidden structure of the data.
- Clustering could be used to try and discover new determinants of a house's price.
  - Taking a price range, say $250,000 to $350,000, a clustering algorithm can create a map that groups houses together that share similar features.
  - There might be a group that are small but urban. There might be another group that share period features and gardens.
  - By comparing the groups across different price ranges, the analysis would start to show segmentation in the market and how it changes as prices increase.

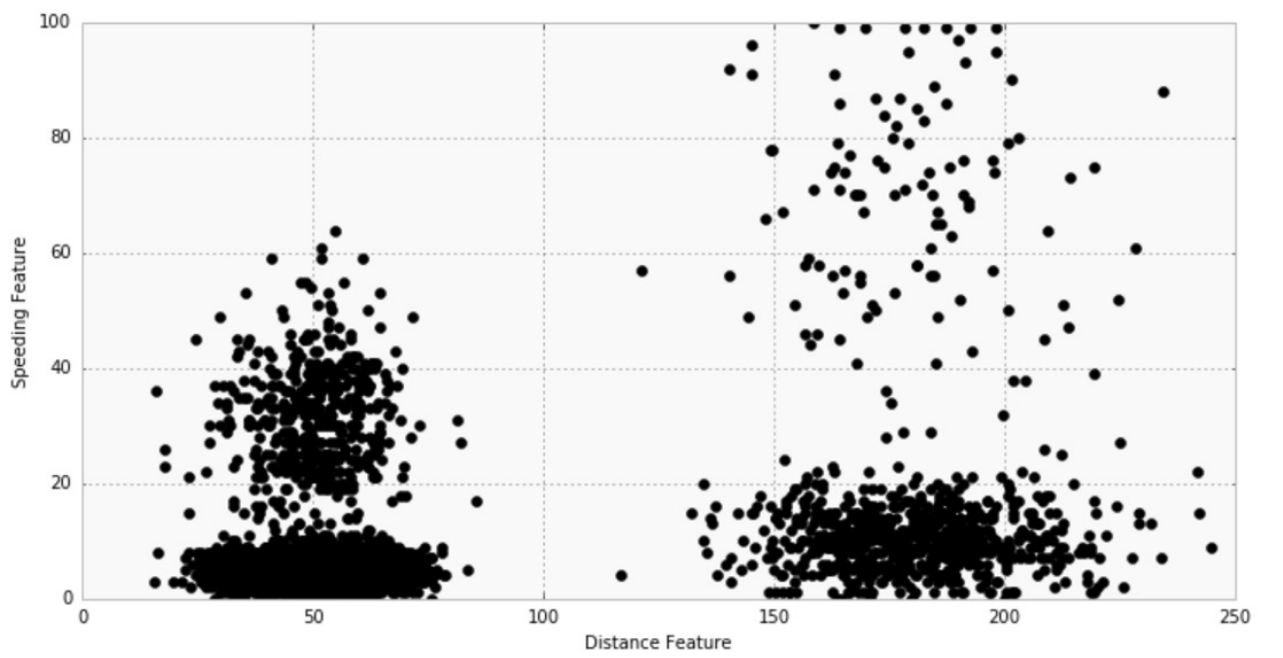
CLUSTERING

- Unlabeled data of drivers

	Driver_ID	Distance_Feature	Speeding_Feature
0	3423311935	71.24	28
1	3423313212	52.53	25
2	3423313724	64.54	27
3	3423311373	55.69	22
4	3423310999	54.58	25

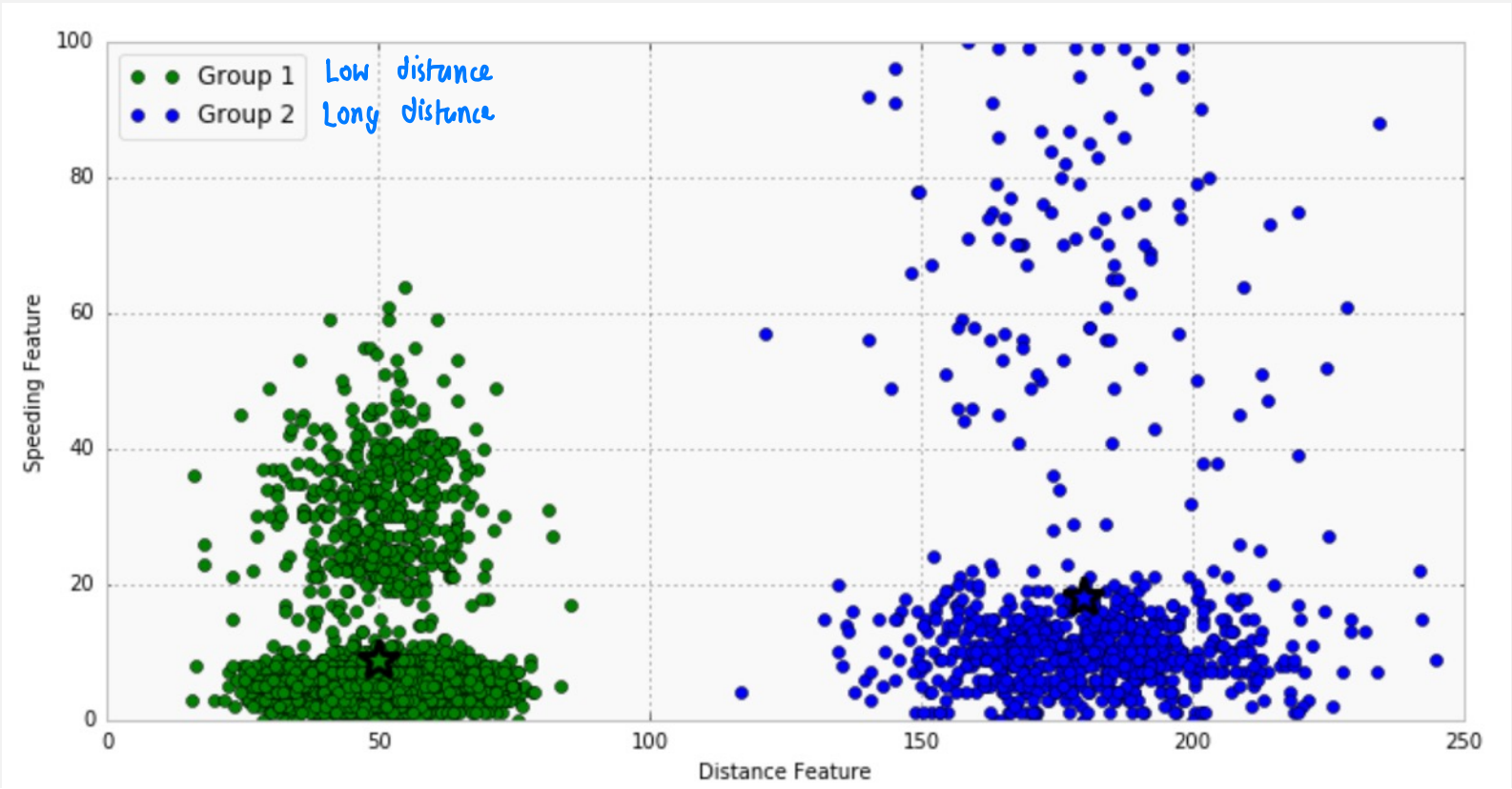
CLUSTERING

- The chart below shows the dataset for 4,000 drivers, with the distance feature on the x-axis and speeding feature on the y-axis.



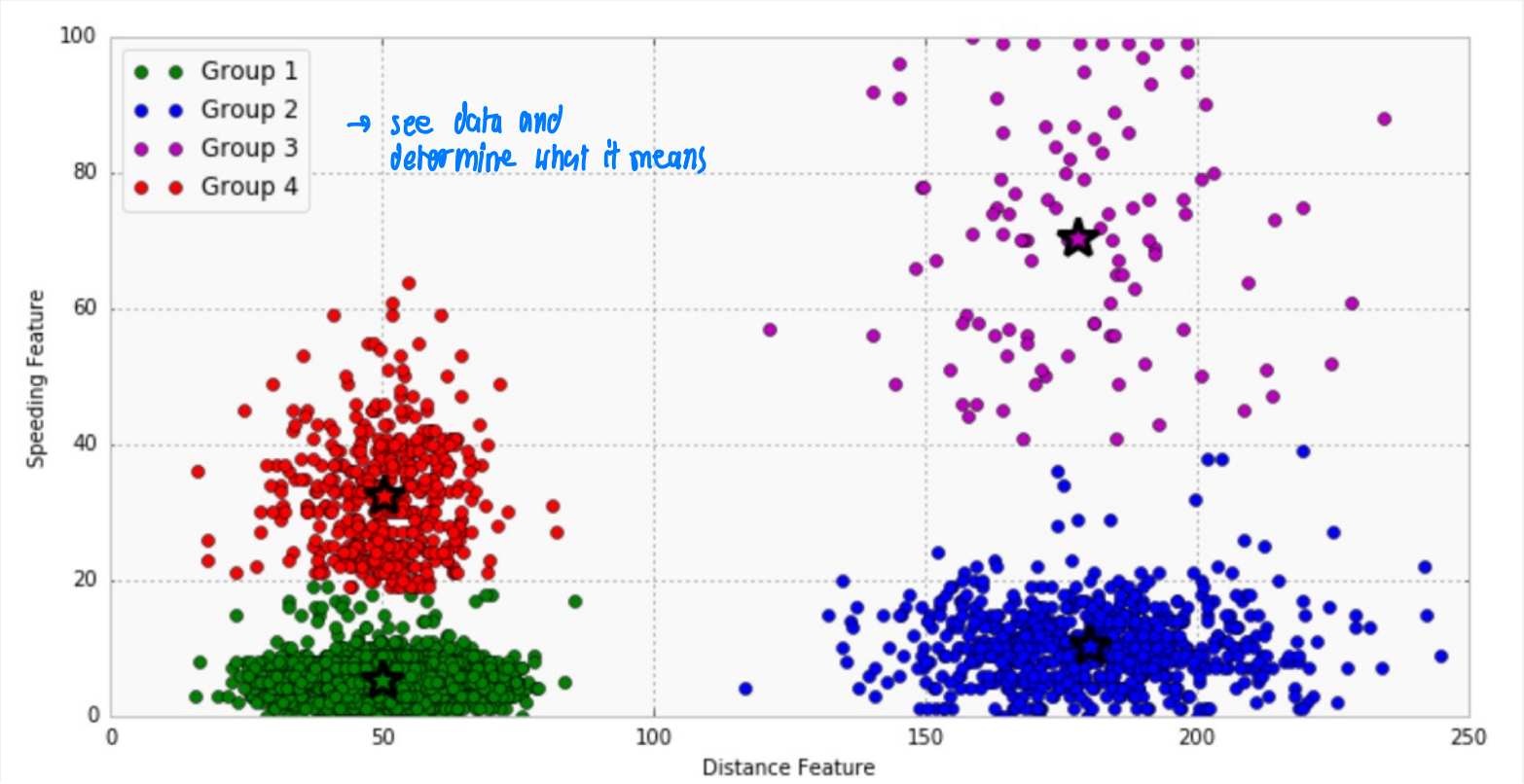
CLUSTERING

- K-means clustering (K = 2)



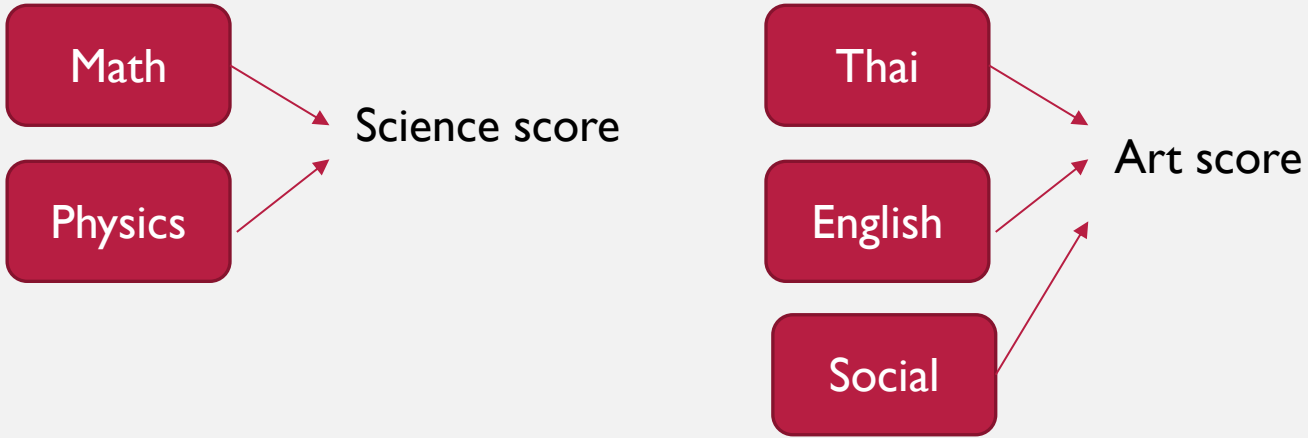
CLUSTERING

- K-means clustering (K = 4)



DATA REDUCTION

- Data reduction attempts to take a **large** set of data and replace it with a **smaller** set of data contains much of the important information in the larger set.
- A **tradeoff** between easiness of processing and loss of information



SIMILARITY MATCHING

ex: Netflix recommends similar movies you watch or people like you watch

- Similarity matching tries to identify similar individuals based on data known about them.
  - IBM likes to find companies similar to their best business customers
  - Product recommendation system

ASSOCIATION LEARNING

- Association learning finds important relations between variables or features in a data set.
  - Groceries stores like to know what get bought together

Ex: diaper & beer often sold together

GAME

TYPES OF MACHINE LEARNING

1) Supervised learning
2) Unsupervised learning
3) Semi-supervised learning
4) Reinforcement learning

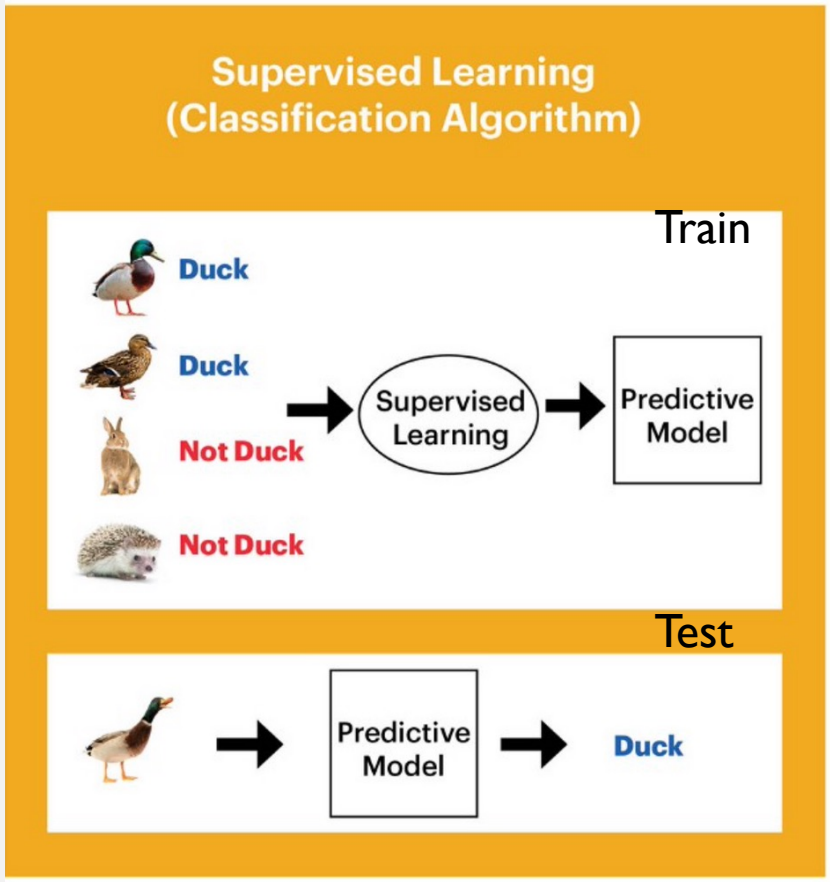
You must have labeled data y

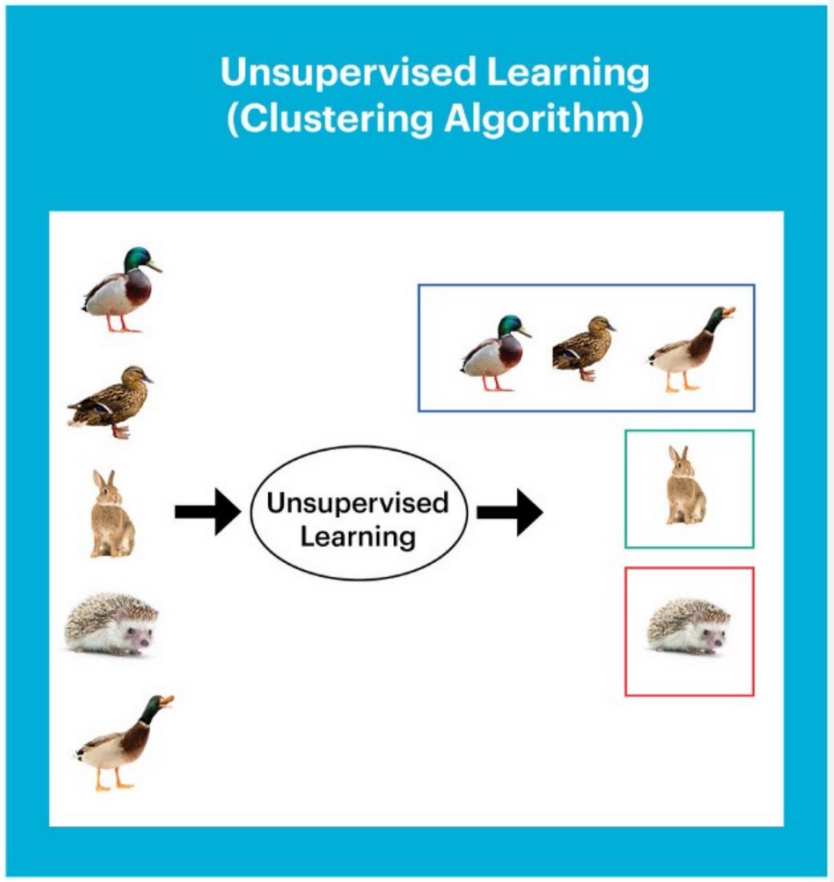
SUPERVISED LEARNING

- This type of learning requires a training data set with **labeled data**, or data with a known output value (e.g., rural/not rural or house price).
- Classification and Regression problems are solved through supervised learning.

UNSUPERVISED LEARNING

- This type of learning techniques **does not** use a **training set** and find patterns or structure in the data by themselves.
- Clusterring, Data reduction, similarity matching and association learning problems can be solved with an unsupervised approach.



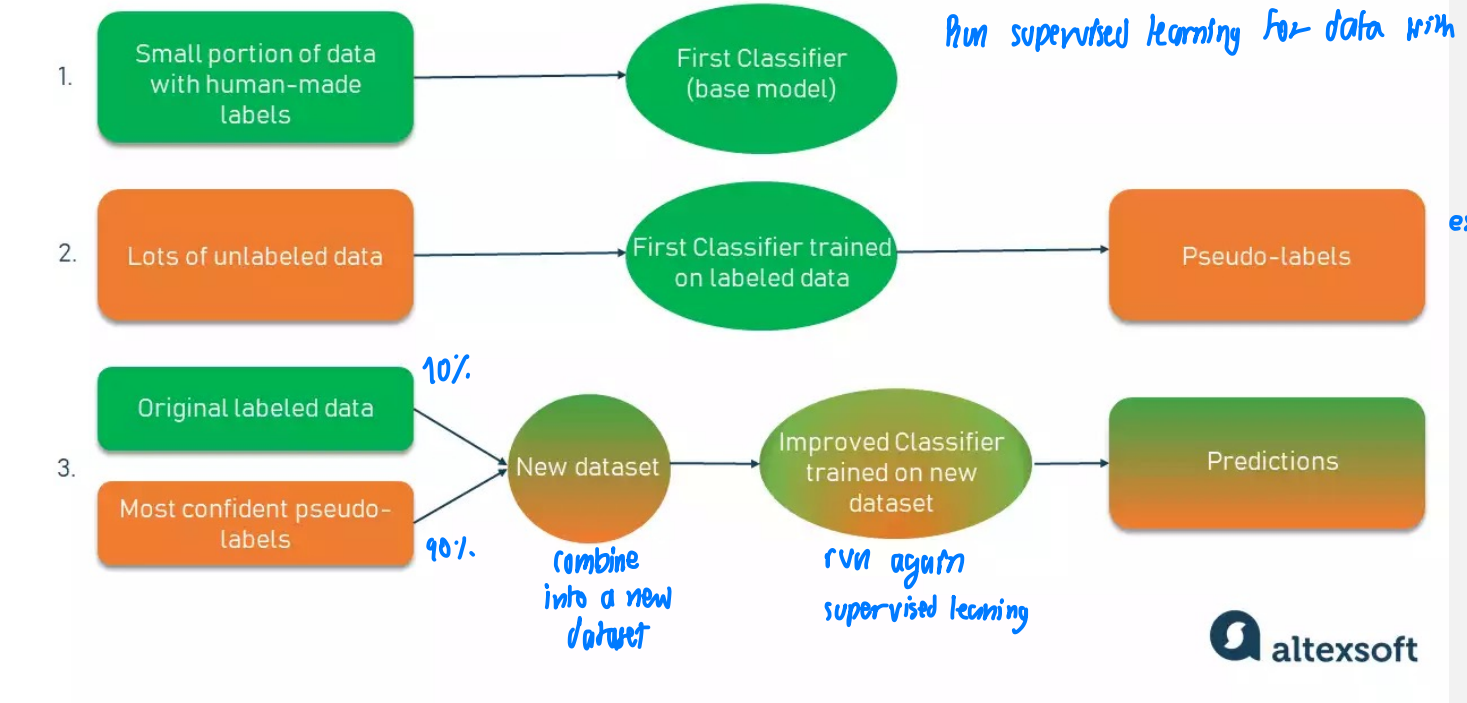


"Train" "Test" used for supervised learning only

No accuracy to measure

SEMI-SUPERVISED LEARNING

- It uses mainly **unlabeled** and a small amount of **labeled** input data.

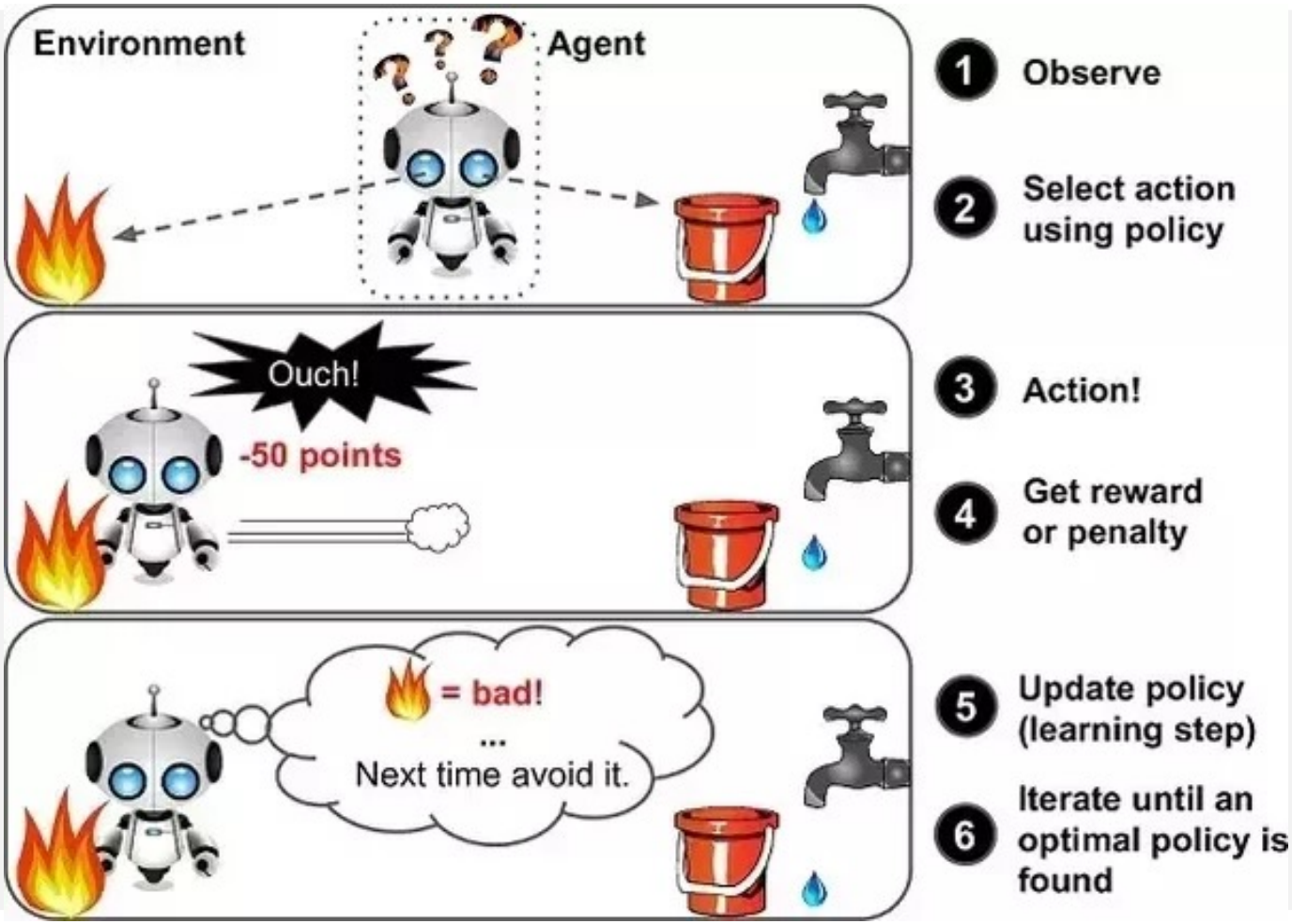


REINFORCEMENT LEARNING

- It uses input data from the **environment** as a stimulus for how the model should react.
- Feedback is not generated through a **training process** like supervised learning but as **rewards** or **penalties** in the environment.
- This type of process is used in **robot** control.
- Agent >>>> Rational agent

- models A, B, C, D
- D has the best performance
- Assign more weight to model D
- Do again → adjust penalties & rewards

REINFORCEMENT LEARNING CONT.



MACHINE LEARNING VS STATISTICAL MODELS

based on human

- Statistical modeling is a formalization of relationships between variables in the data in the form of **mathematical equations**.
- Machine learning is an algorithm that can **learn from data** without relying on rules-based programming.
- Statistics is about sample, population, hypothesis, etc.
- Machine learning is all about predictions, supervised learning, unsupervised learning, etc.

MACHINE LEARNING VS. STATISTICS

- Machine learning requires **no prior assumptions** about the underlying relationships between the variables.
- We can **put all the data** we have into the model, and the algorithm processes the data and **discovers patterns**, using which we can make predictions on the new data set.
- Machine learning treats an algorithm like a **black box**, as long it works. It is generally applied to **high dimensional data sets**, the more data you have, the **more accurate your prediction** is.

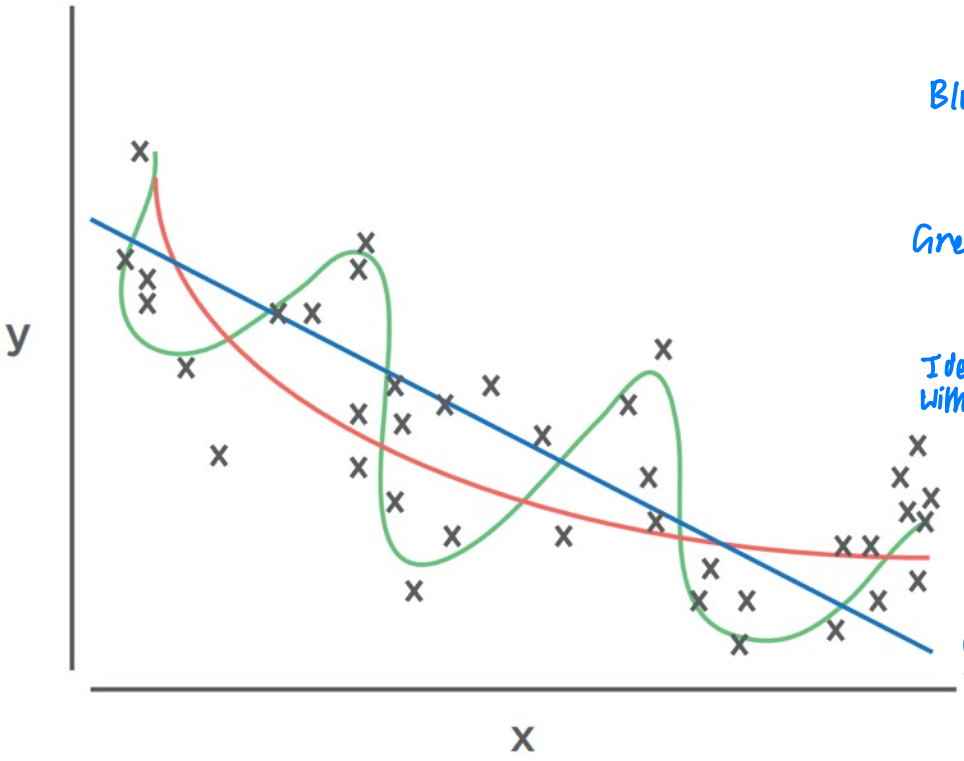
MACHINE LEARNING VS. STATISTICS CONT.

- However, understanding the **association** and knowing their differences enables **machine learners and statisticians** to expand their knowledge and even apply methods **outside their domain of expertise**.
- This is the notion of "**data science**" itself, which aims to bridge the gap. Collaboration and communication between these two-fascinating data-driven disciplines allows us to make better decisions that will ultimately positively affect our way of living.

ASSUMPTIONS AND INDUCTIVE BIAS

- Machine learning algorithms will make assumptions about the ‘best’ function that fits the data.
- It is possible to find **multiple functions** that fit with a given **training data set**.
- To choose one, the machine learning algorithm will need to make assumptions about **what the function being modelled looks like**.

ASSUMPTIONS AND INDUCTIVE BIAS



Several different functions might fit the same data set

X , Y

Blue
- not strong linear correlated
- underfitting
- excluded some functions

Green
- overfitting

Identify underfitting & overfitting issue with accuracy.

	Train	Test
underfit	Low (> 90 %)	Low (< 80 %)
overfit	High (> 98 %)	Low (< 90 %)

rank the model by testing results, not training result.

ASSUMPTIONS AND INDUCTIVE BIAS

- Overfitting
  - The model uses **complex hypotheses** and focuses on irrelevant factors in the **training set** limiting the ability to generalize when faced with **new data**.
- Underfitting
  - The model only considers **simple hypotheses** and therefore excludes the **'real' function**.

HW2: FOR EXE1, WHAT DATA AND MODEL WILL YOU USE? (EXPLAIN)

- what attribute in → transaction
  ↳ purchase . . -

- what model
- no 100% correct answer
- depends on what kind of data can be collected
  ↳ how you plan the project
- submit .PDF file.