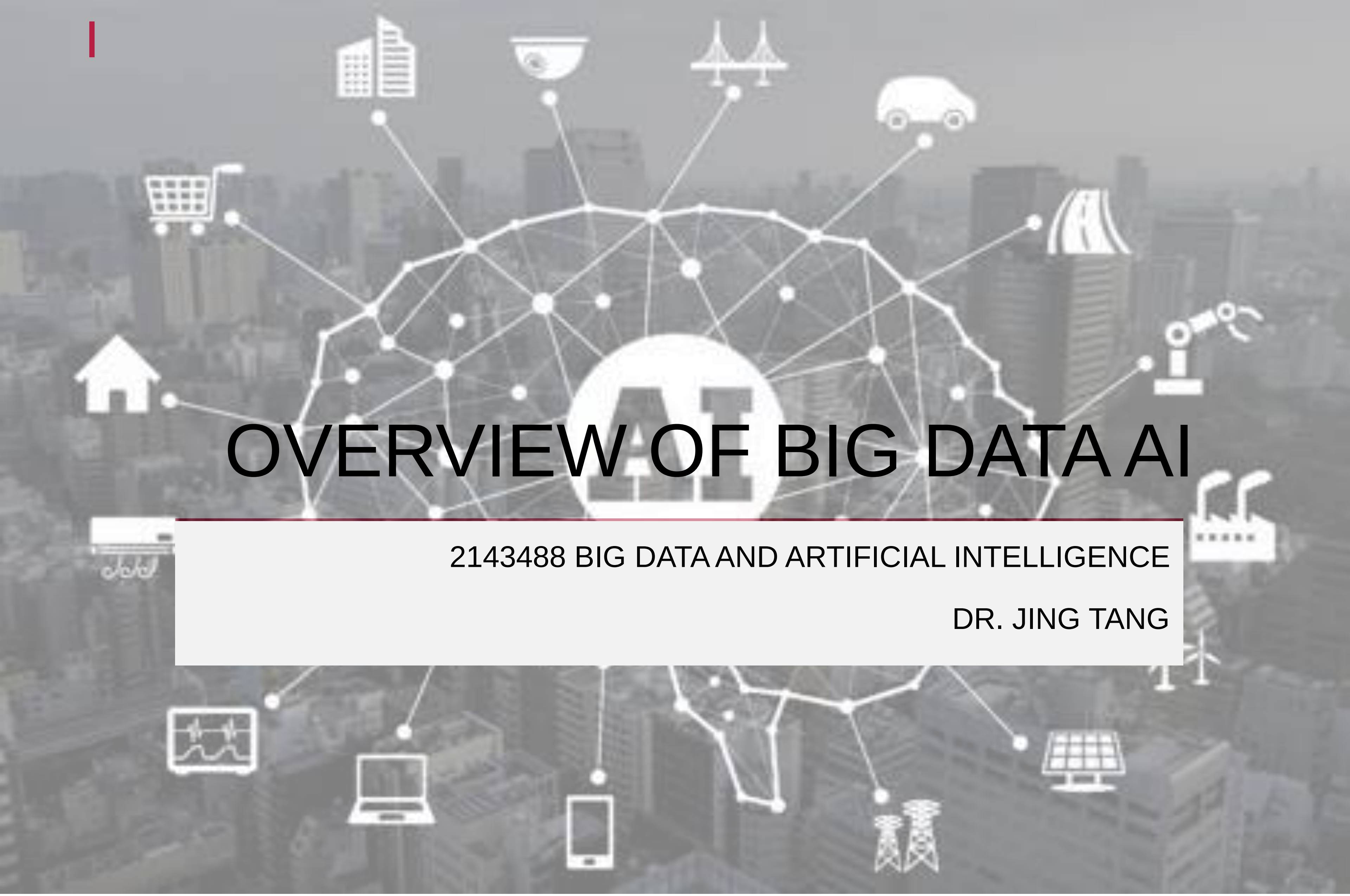


OVERVIEW OF BIG DATA AI

2143488 BIG DATA AND ARTIFICIAL INTELLIGENCE

DR. JING TANG

DATA ALL AROUND

- Data management is not just a software issue.
  - 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: 2.5 B *(2012)* → 4.5B *(2019)*

Data: 1.2 ZB *(2010)* → 2.8 ZB *(2012)* → 40 ZB *(2020)*

(Source: EMC) ***1 ZB = 250 B DVDs = Netfilx * 3000***

About 200 M years to download all data

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

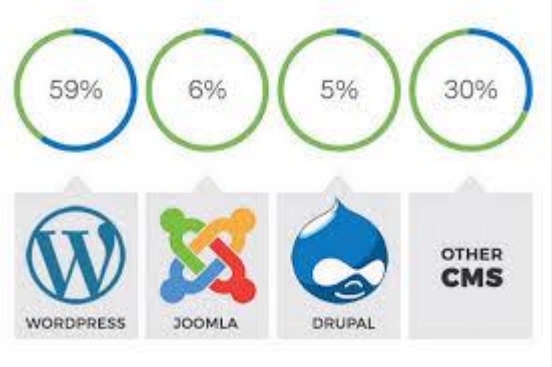
80% of data is unstructured!

STRUCTURED 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

- Its usage is rapidly expanding, but still challenge
- Machine-generated: Satellite; Science; Photo & Video; Radar & Sonar
- Human-generated: Text; Social media; Mobile; Website
- Content Management System (CMS)



BIG DATA

- Big Data is any data that is expensive to manage and hard to extract value from

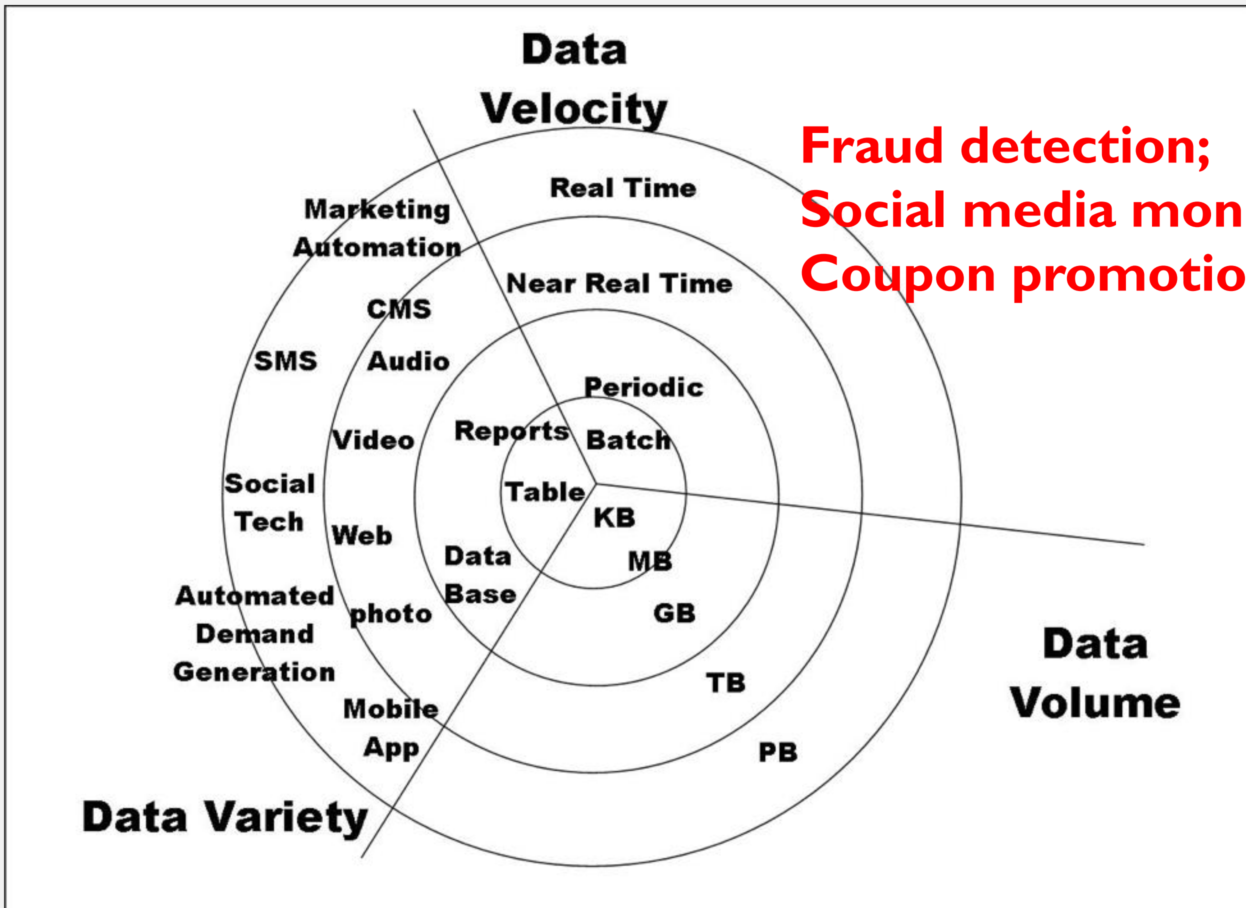
90% created in last 2 years *(IBM,2017)*

Short useful lifespan

Difficult to integrate

- 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



Fraud detection;
Social media monitoring;
Coupon promotion at POS

DEVELOPMENT OF MOBILE COMMUNICATION NETWORK

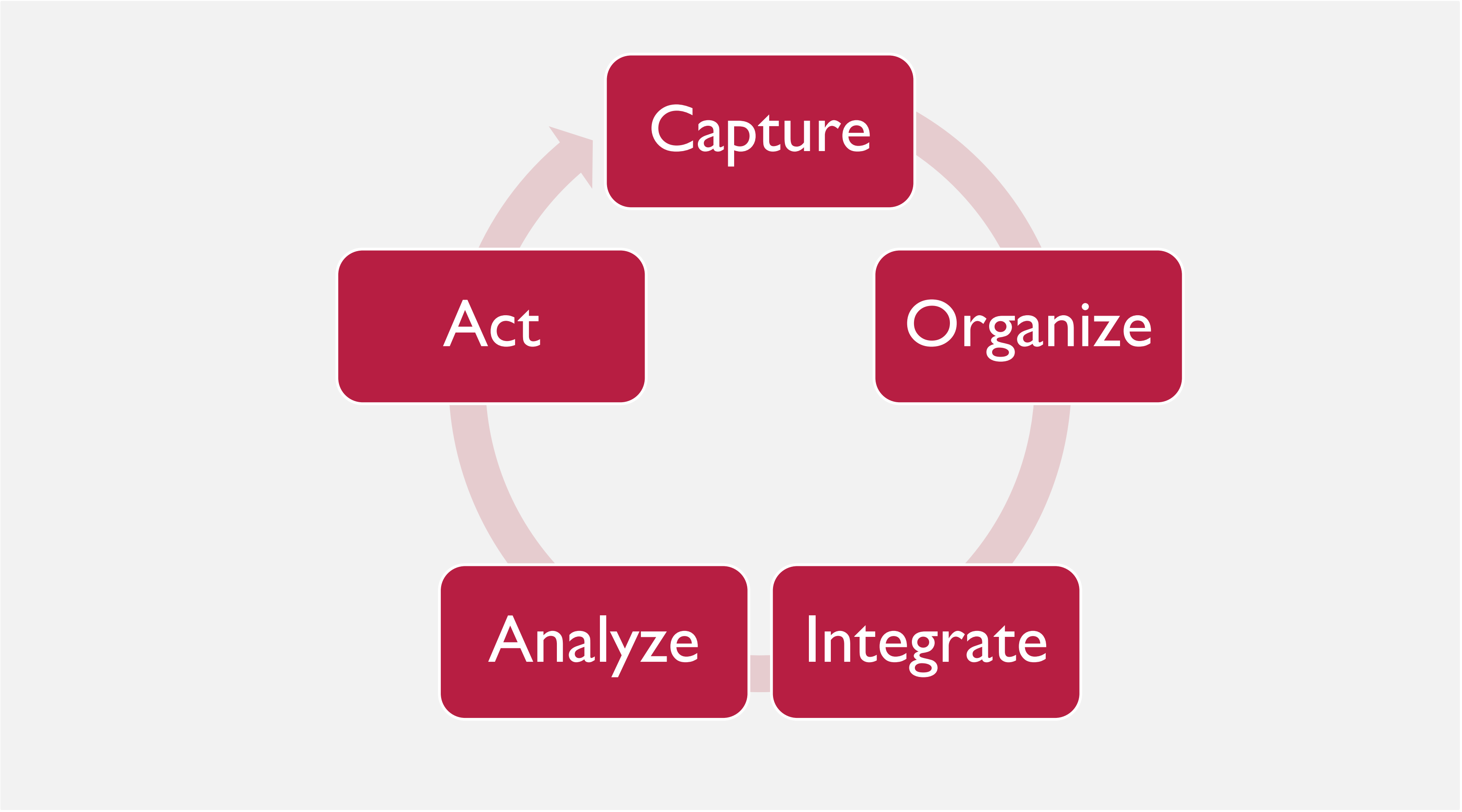
<table>
<tr><th></th><th colspan="4">1G</th><th colspan="3">2G</th><th colspan="2">3G</th><th colspan="2">4G</th><th>5G</th></tr>
<tr><td>Technology standard</td><td>AMPS</td><td>NMT</td><td>TACS</td><td>C-Netz</td><td>GSM</td><td>D-AMPS</td><td>IS-95 A</td><td>UMTS</td><td>CDMA2000</td><td colspan="2">LTE</td><td>NR</td></tr>
<tr><td>Digital or not?</td><td colspan="4">Analogue</td><td colspan="3">Digital</td><td colspan="2">Digital</td><td colspan="2">Digital</td><td>Digital</td></tr>
<tr><td>Launch year (approx.)</td><td colspan="4">~1980</td><td colspan="3">~1990</td><td colspan="2">~2000</td><td colspan="2">~2010</td><td>~2020</td></tr>
<tr><td rowspan="3">Enhancements</td><td colspan="4" rowspan="3"></td><td>GPRS</td><td rowspan="3"></td><td rowspan="3">IS-95 B</td><td>HSPA</td><td>EVDO Rev. 0</td><td colspan="2">LTE-Advanced</td><td rowspan="3"></td></tr>
<tr><td rowspan="2">EDGE</td><td rowspan="2">HSPA+</td><td>EVDO Rev. A</td><td colspan="2" rowspan="2">LTE-Pro</td></tr>
<tr><td>EVDO Rev. B</td></tr>
<tr><td>Services</td><td colspan="4">Voice only</td><td colspan="8">Voice + SMS + Data (Mobile Internet)</td></tr>
<tr><td rowspan="7">Peak download speeds</td><td colspan="4" rowspan="7">-</td><td rowspan="2">GPRS</td><td colspan="2" rowspan="2">171.2 kbps</td><td>UMTS</td><td>2 Mbps</td><td rowspan="2">LTE</td><td rowspan="2">300 Mbps</td><td rowspan="7">10 Gbps</td></tr>
<tr><td>HSPA</td><td>14.4 Mbps</td></tr>
<tr><td>EDGE</td><td colspan="2">384 kbps</td><td>HSPA+</td><td>42 Mbps</td><td rowspan="2">LTE-A</td><td rowspan="2">1 Gbps</td></tr>
<tr><td>IS-95 A</td><td colspan="2">14.4 kbps</td><td>CDMA2000</td><td>153 kbps</td></tr>
<tr><td rowspan="3">IS-95 B</td><td colspan="2" rowspan="3">115 kbps</td><td>EVDO 0</td><td>2.4 Mbps</td><td rowspan="3">LTE-Pro</td><td rowspan="3">3Gbps</td></tr>
<tr><td>EVDO A</td><td>3.1 Mbps</td></tr>
<tr><td>EVDO B</td><td>14.7 Mbps</td></tr>
</table>

WHAT TO DO WITH THESE DATA?

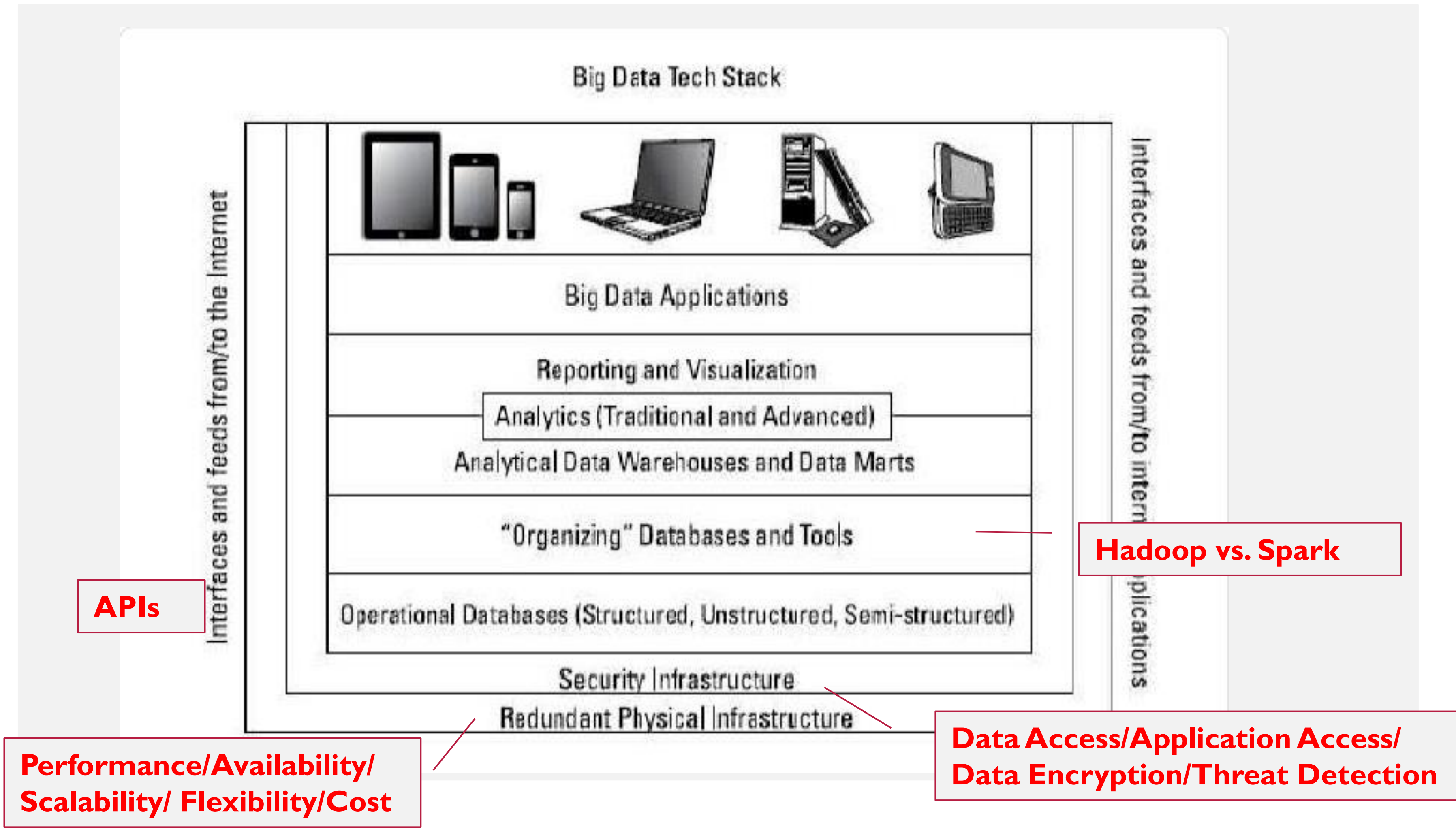
- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

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)*

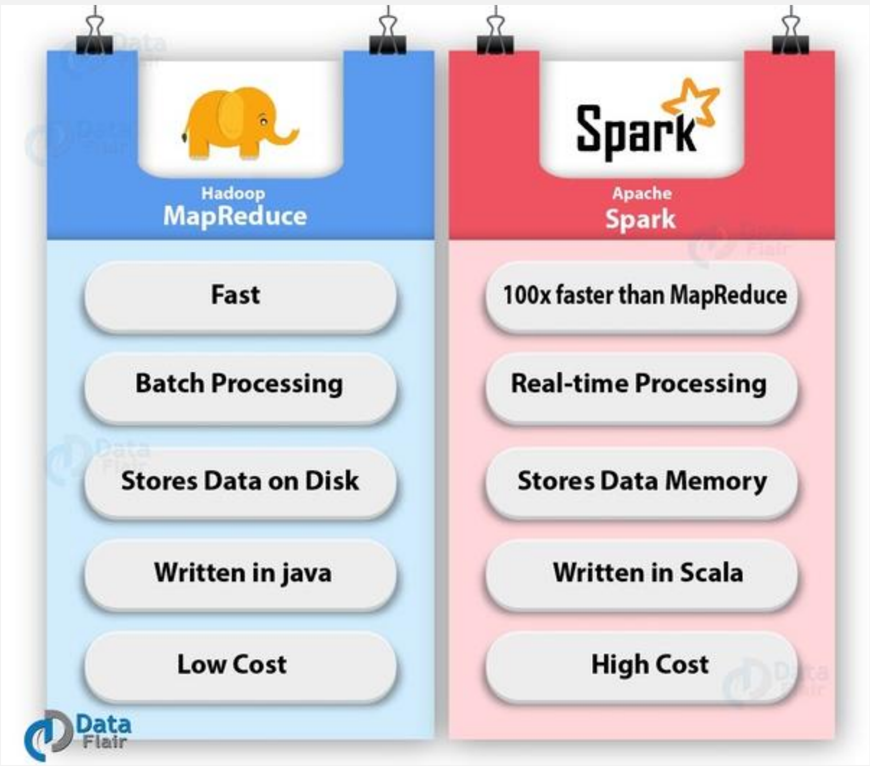
CYCLE OF BIG DATA MANAGEMENT

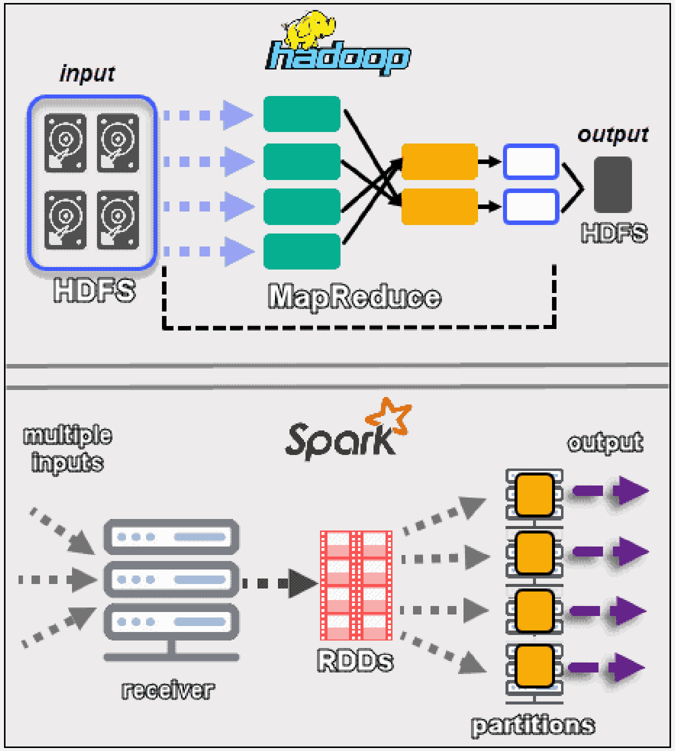


ARCHITECTURAL FOUNDATION OF BIG DATA

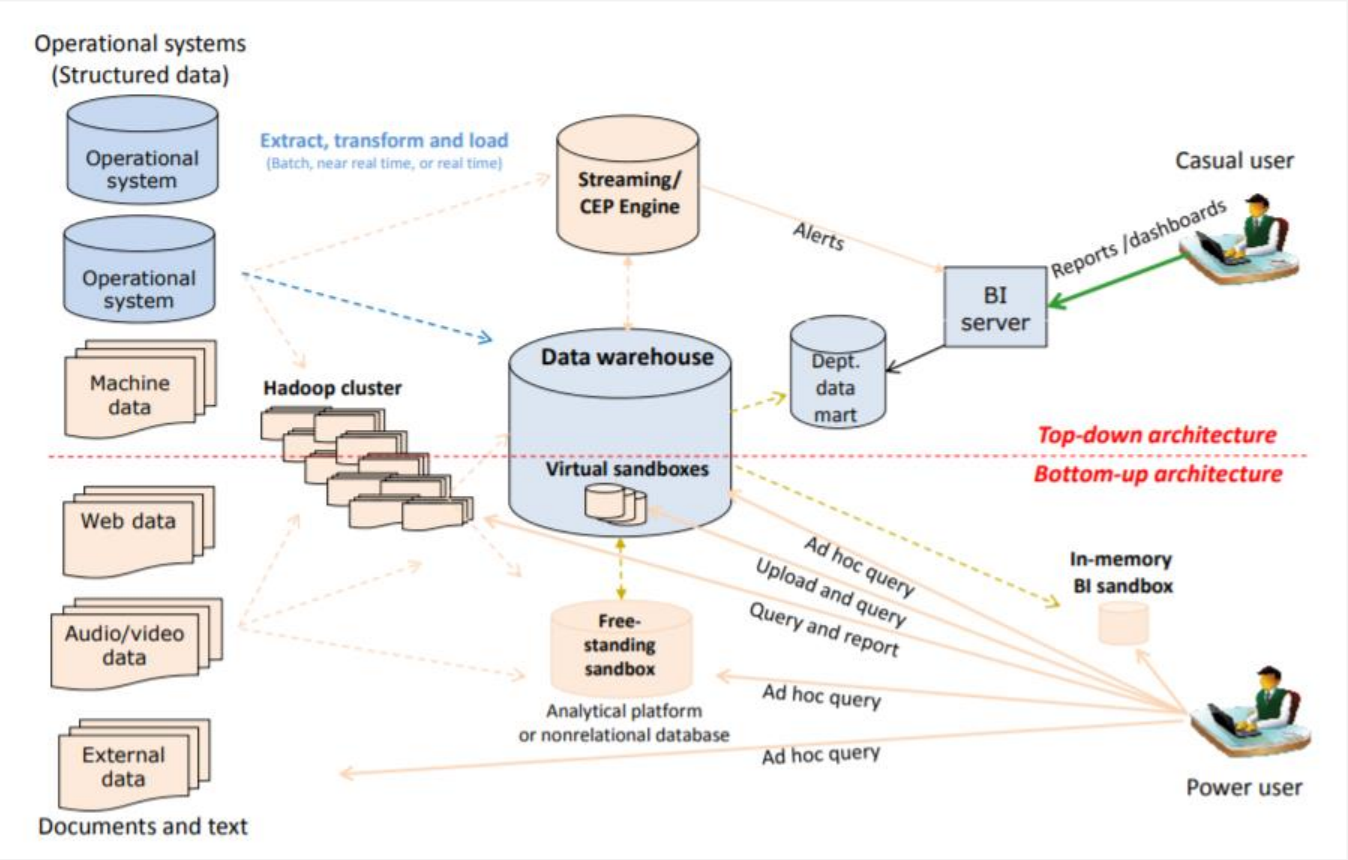


HADOOP VS. SPARK





DATA WAREHOUSES & MARTS



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

- Reporting and Dashboard
- Visualization
- Analytics and Advanced Analytics

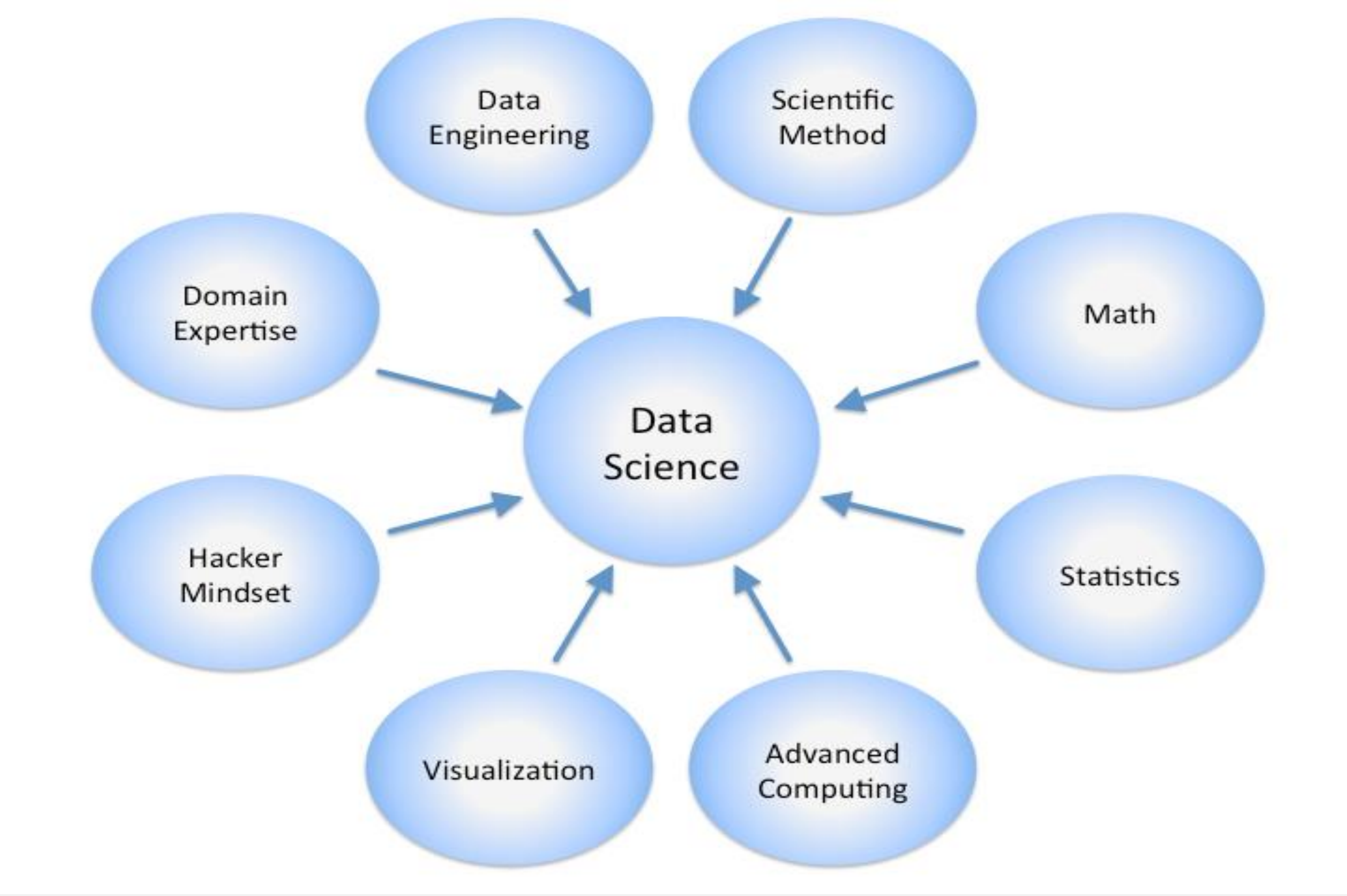
WHAT IS DATA SCIENCE?

- 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

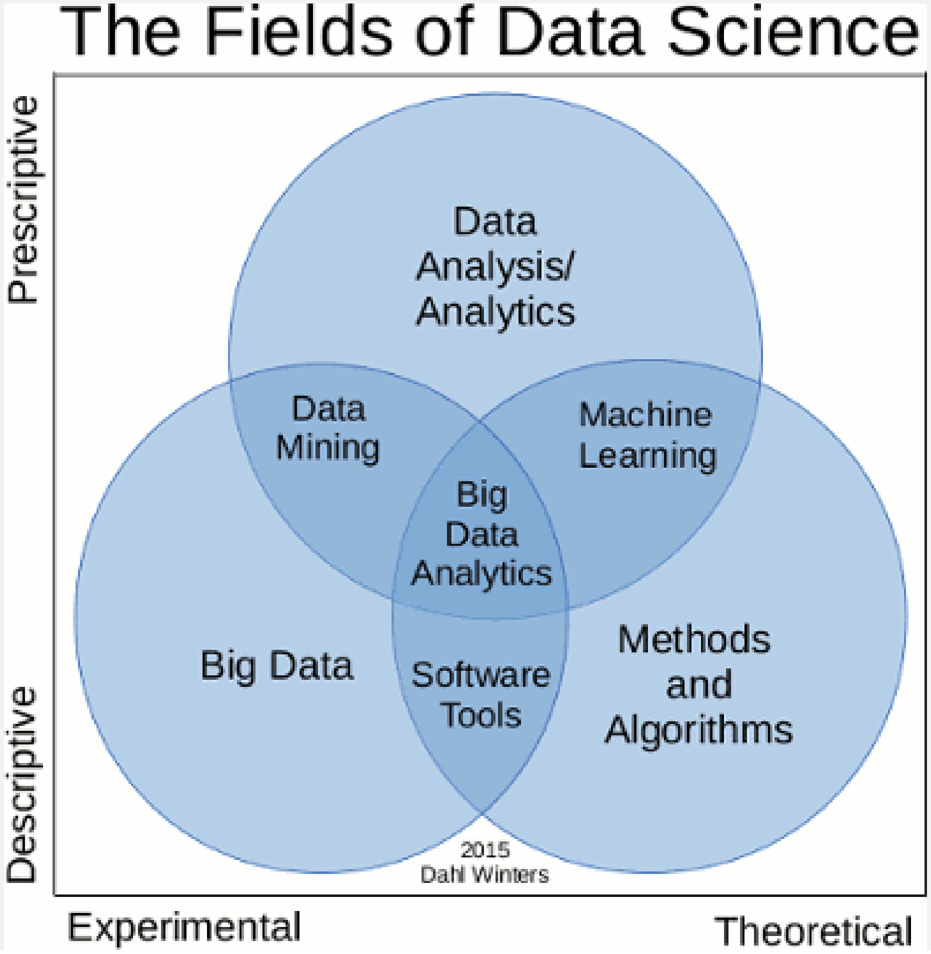
WHAT IS DATA SCIENCE?

- 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



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…

NEED A TEAM FOR BIG DATA

- 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.
- 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

Preprocess (Collect, Move, Clean, Integrate) → Analysis & Visualization → Prediction & Decision

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 (2)

- **Artificial intelligence** aims to improve machine behavior in tackling such complex tasks.
- 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


All-new
echo
"Alexa, turn it up."


TESLA


EASY INVEST
ROBO ADVISOR

DEFINITIONS OF AI:

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

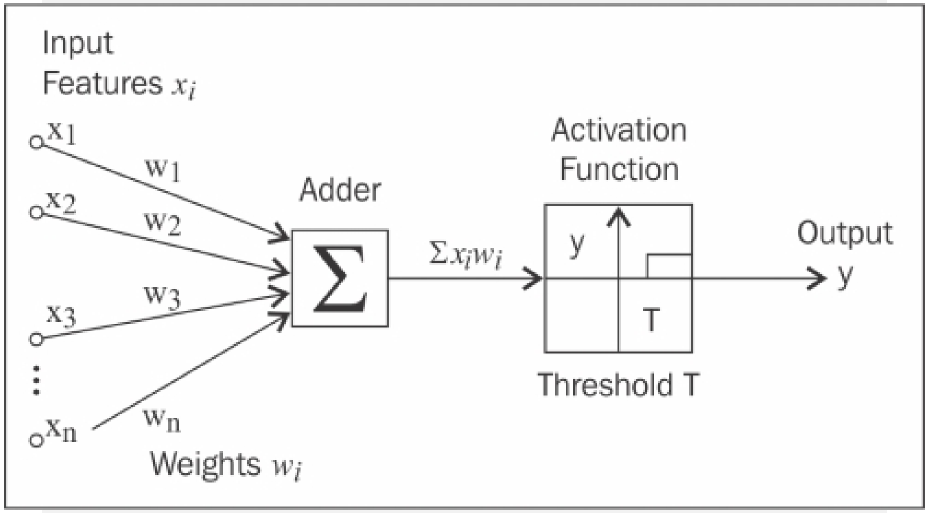
- 1923 *Karel Capek* used the word “Robot” in English.
- The first work that is now generally recognized as AI was done by *Warren McCulloch and Walter Pitts (1943).*
- They drew on three sources:
  1) knowledge of the basic physiology and function of neurons in the brain
  2) a formal analysis of propositional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

1943 1952 1955 1956 1966 1969 1973 1979 1980 1986 1987 1995 2001

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.

1943 1952 1955 1956 1966 1969 1973 1979 1980 1986 1987 1995 2001

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)

1943 1952 1955 1956 1966 1969 1973 1979 1980 1986 1987 1995 2001

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 <u>simple examples</u>.
- 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 a 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.

1943 1952 1955 1956 1966 1969 1973 1979 1980 1986 1987 1995 2001

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.

1943 1952 1955 1956 1966 1969 1973 1979 1980 1986 1987 1995 2001

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)

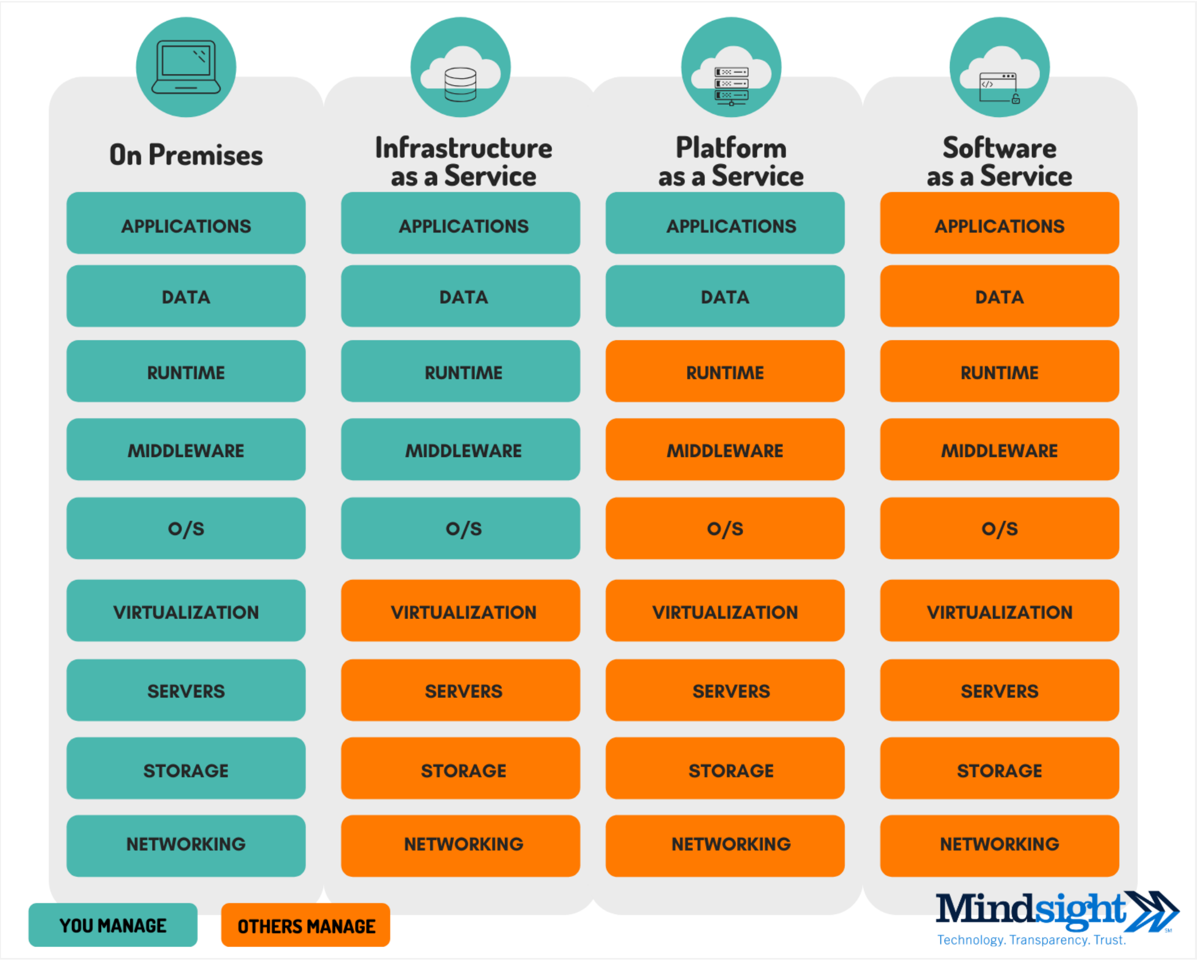
- It makes more sense to worry about the data and be less picky about what algorithm to apply.
- The performance of filling in holes in a photograph algorithm was poor when they used a collection of only ten thousand photos but crossed a threshold into excellent performance when they grew the collection 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:

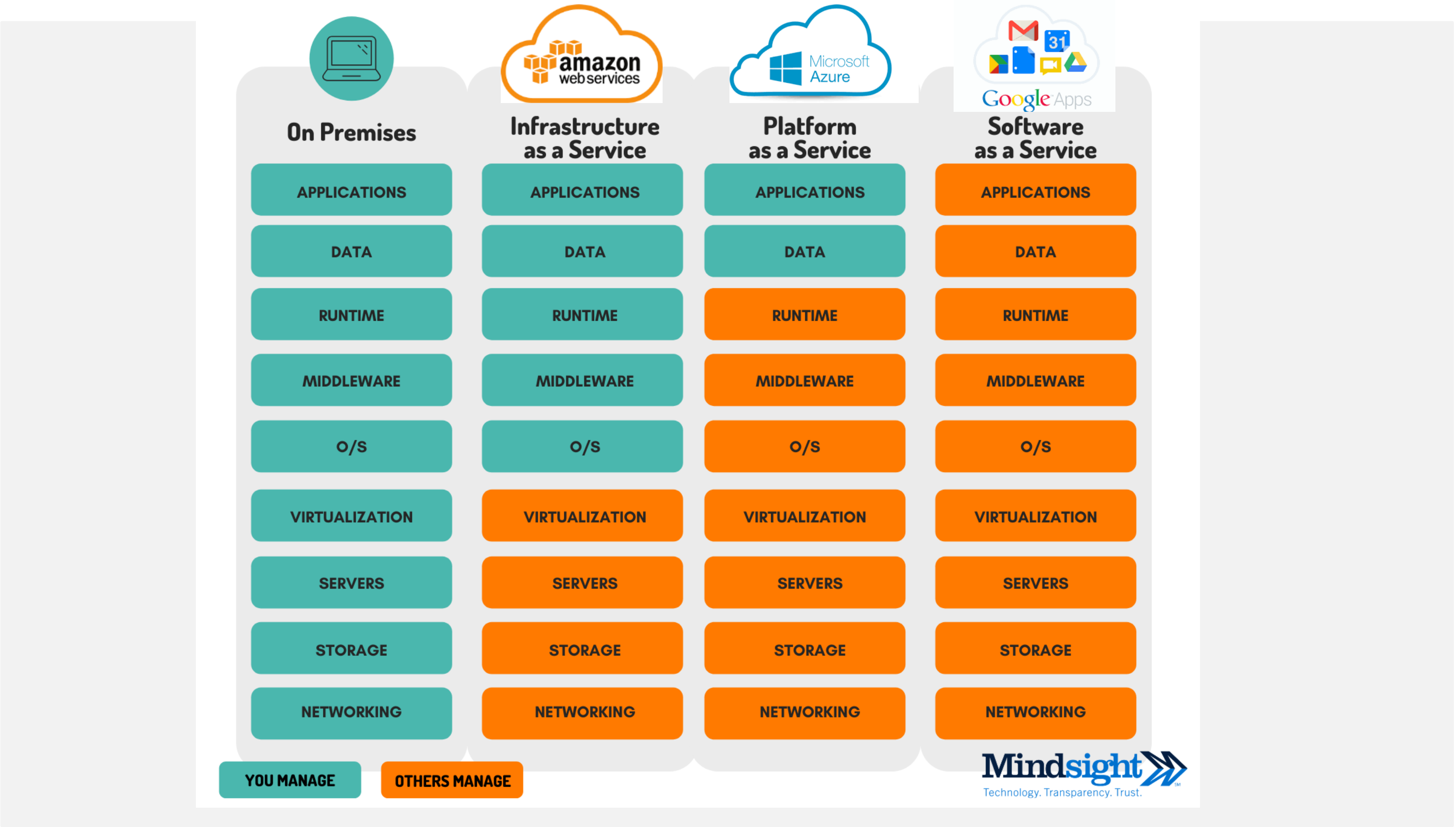
Tell me the name of representative **cloud AI service** the following companies provide

- Amazon Web Service
- IBM Watson
- Microsoft Azure
- Google Cloud Platform
- Alibaba Cloud
- Line Clova

IAAS VS. PAAS VS. SAAS

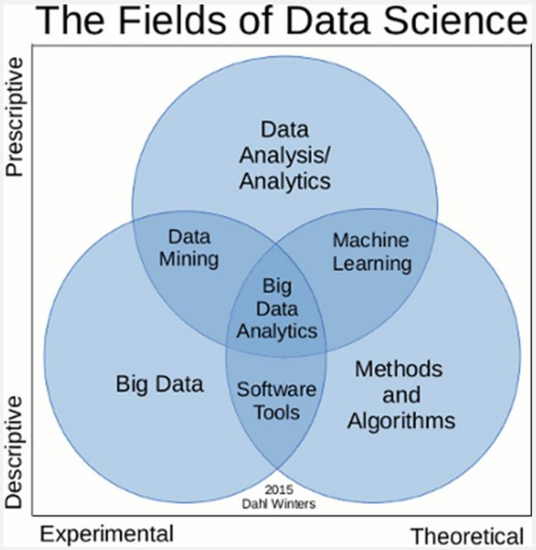


IAAS VS. PAAS VS. SAAS

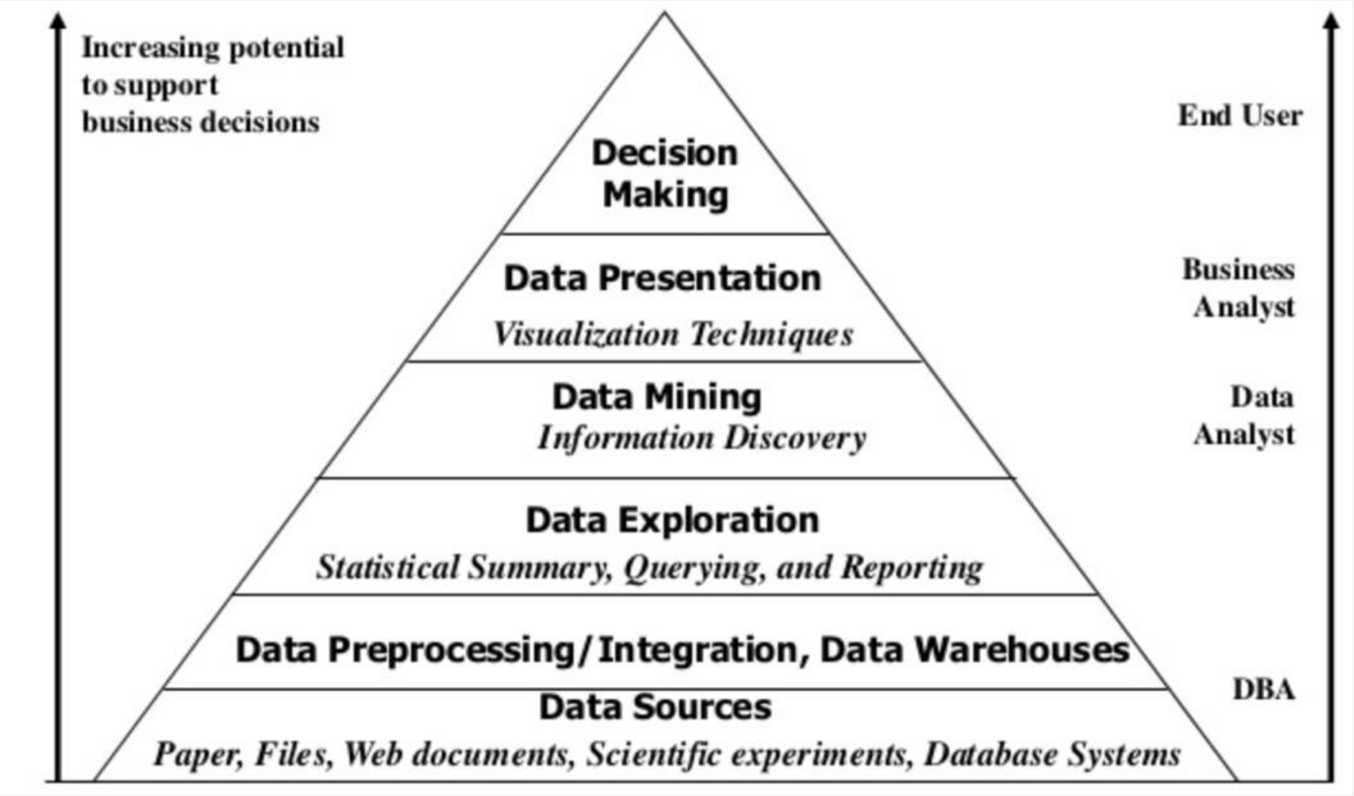


DATA MINING

- 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



Explaining data

Spotting data

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

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
  - Shopping History
  - Pregnancy?

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
- 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
- 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?
- Determine next steps

6. DEPLOYMENT

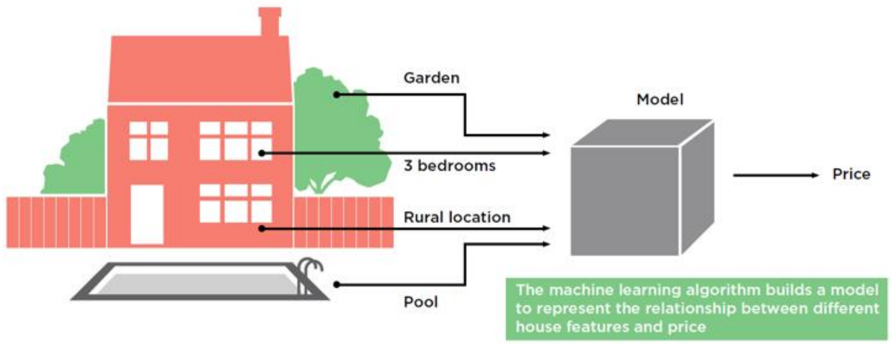
- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project

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



- 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 **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 data sets**.

TYPE OF PROBLEMS

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

REGRESSION

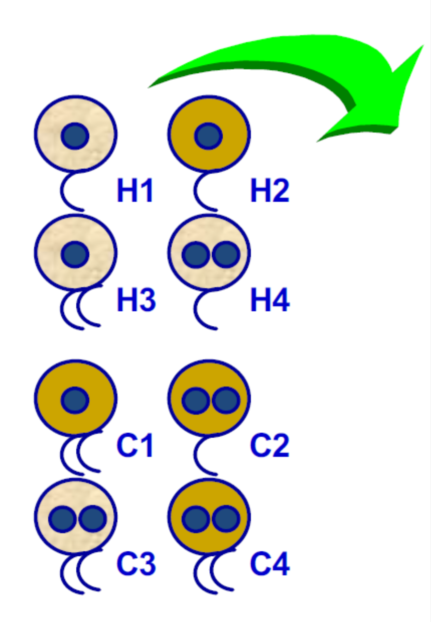
- Data is given a real value rather than a label attribute.
- Machine learning predicts **values** for new data.
  - House prices based on the historical local market data
  - Price of a stock or a market over time
  - Population increases in next 10 years

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
  - Handwriting recognition
  - 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.

CLASSIFICATION

Label



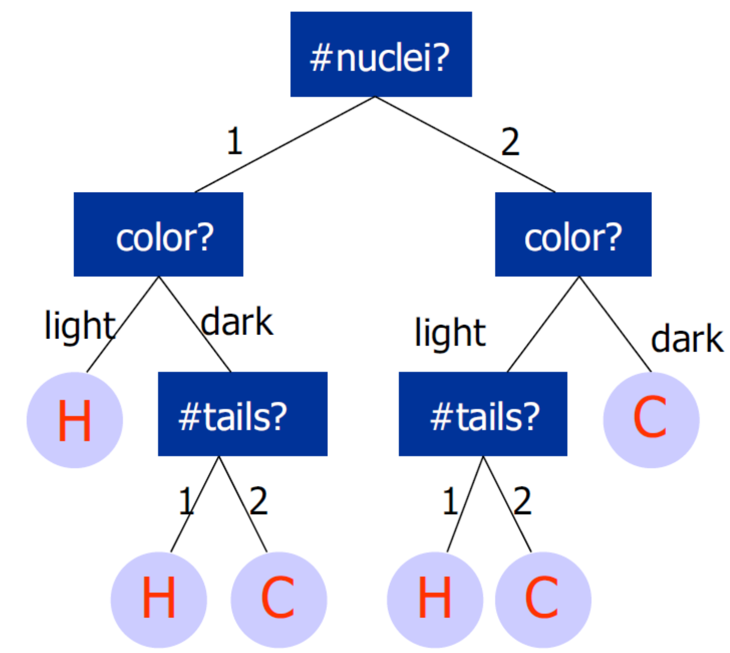
ID	color	#nuclei	#tails	status
H1	light	1	1	healthy
H2	dark	1	1	healthy
H3	light	1	2	healthy
H4	light	2	1	healthy
C1	dark	1	2	cancerous
C2	dark	2	1	cancerous
C3	light	2	2	cancerous
C4	dark	2	2	cancerous

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



CLUSTERING

- **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.

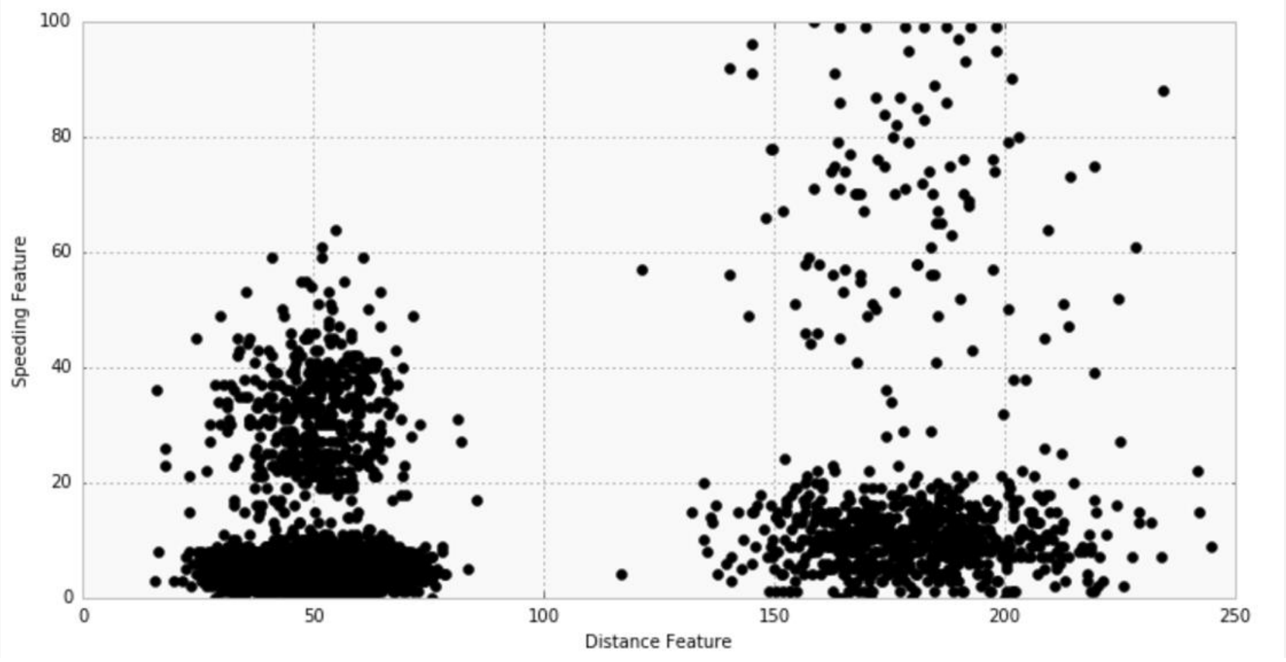
CLUSTERING

- 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

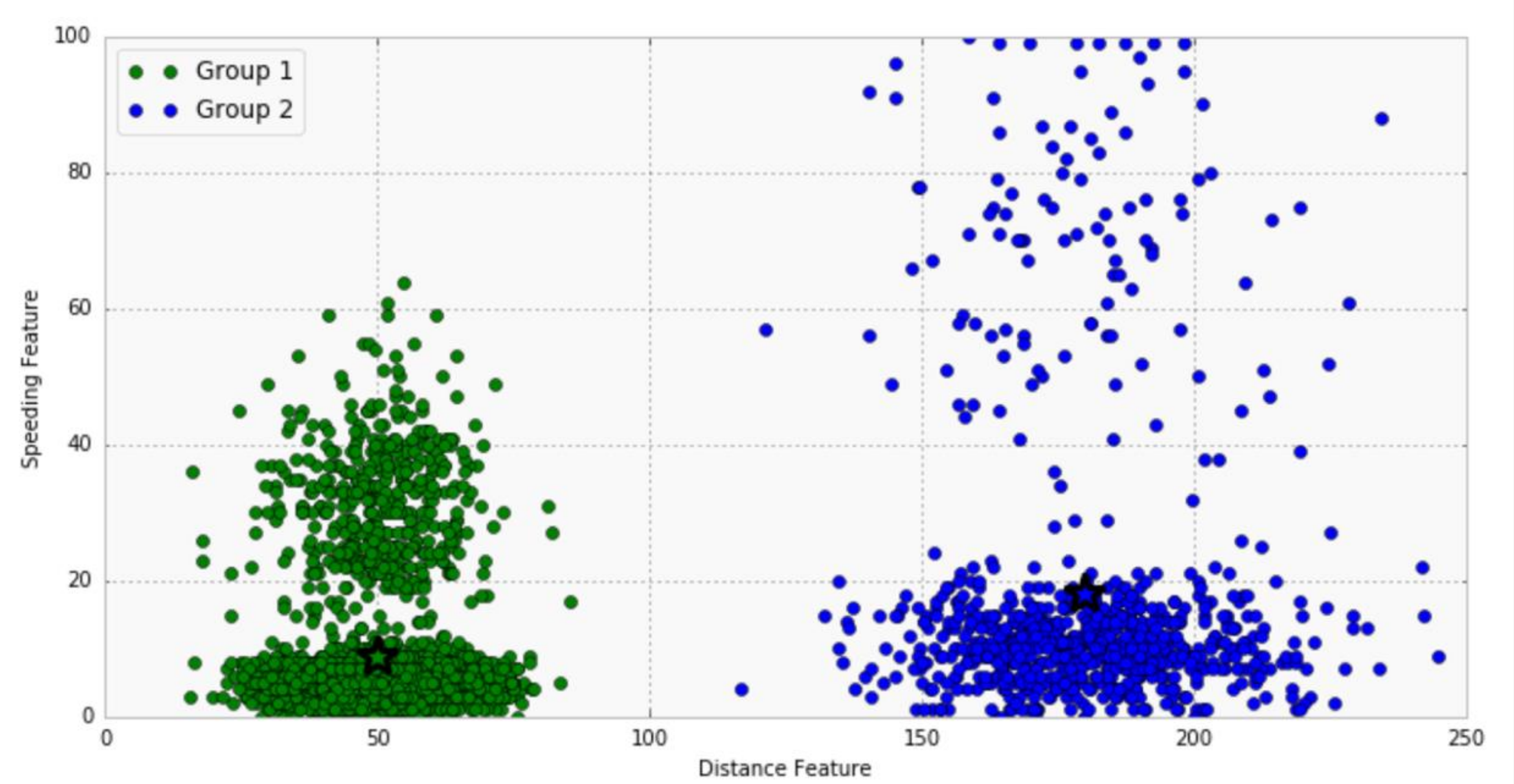
CLUSTERING

- 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.



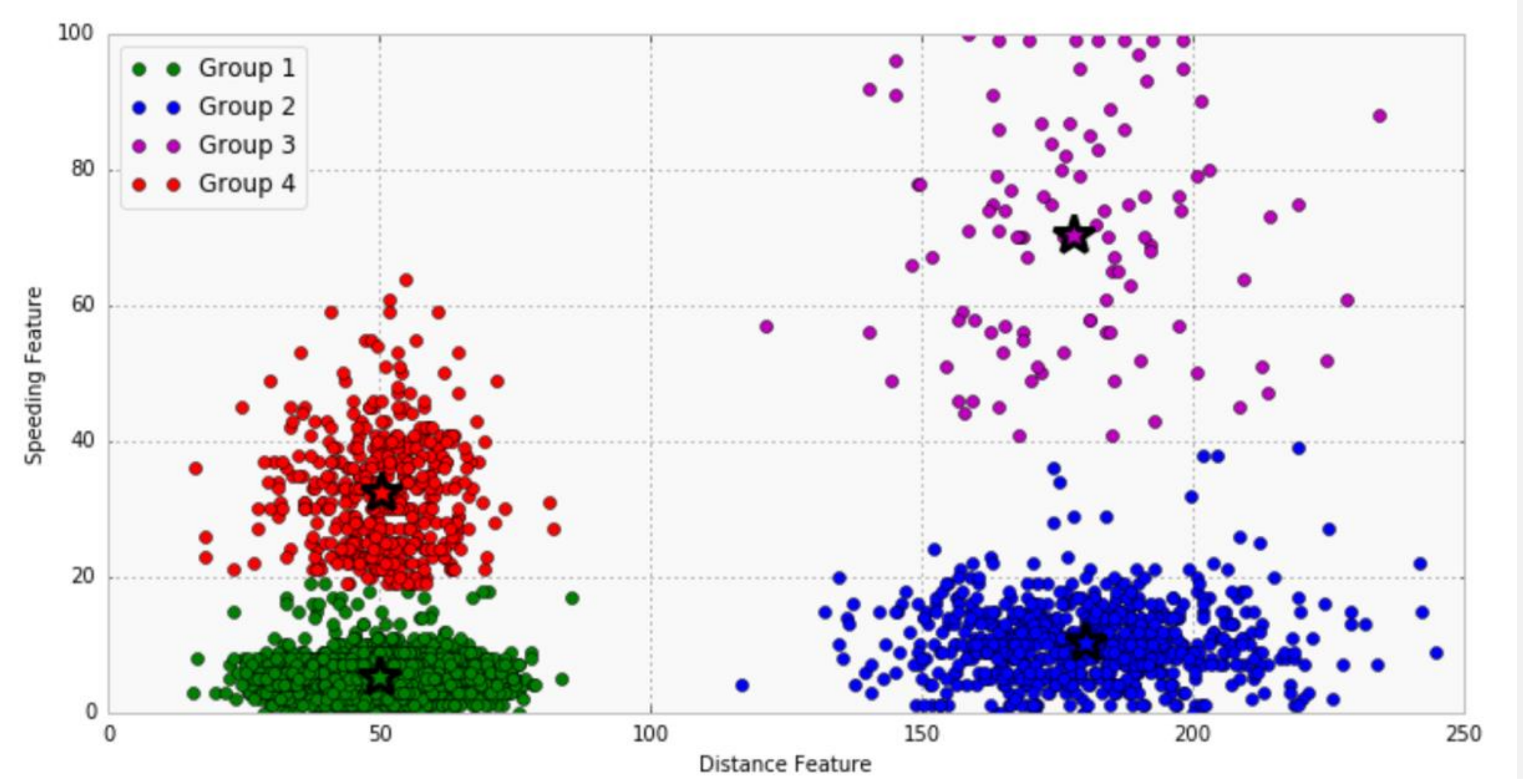
CLUSTERING

- K-means clustering (K = 2)



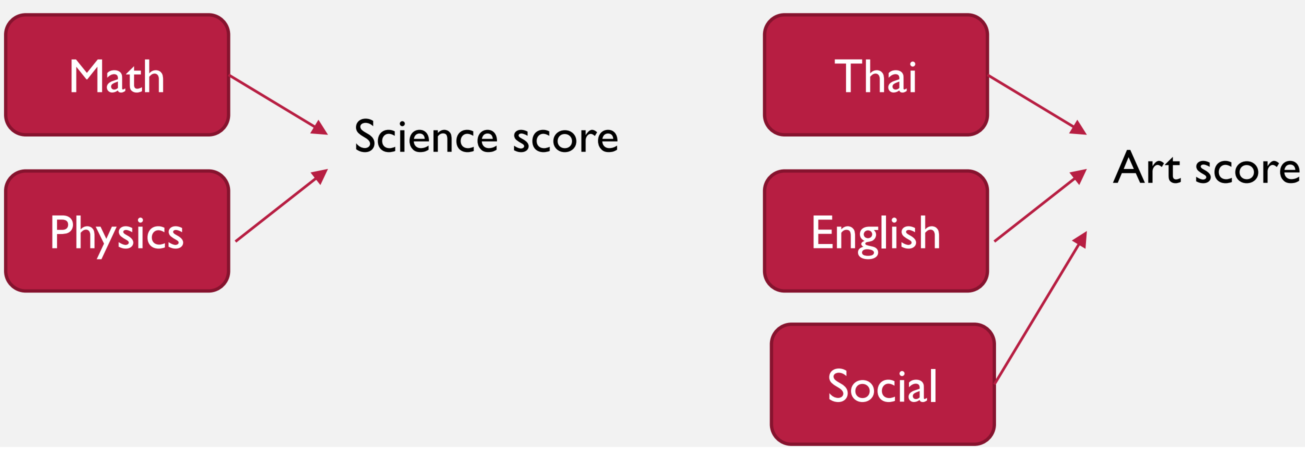
CLUSTERING

- 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

- 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

GAME

TYPES OF MACHINE LEARNING

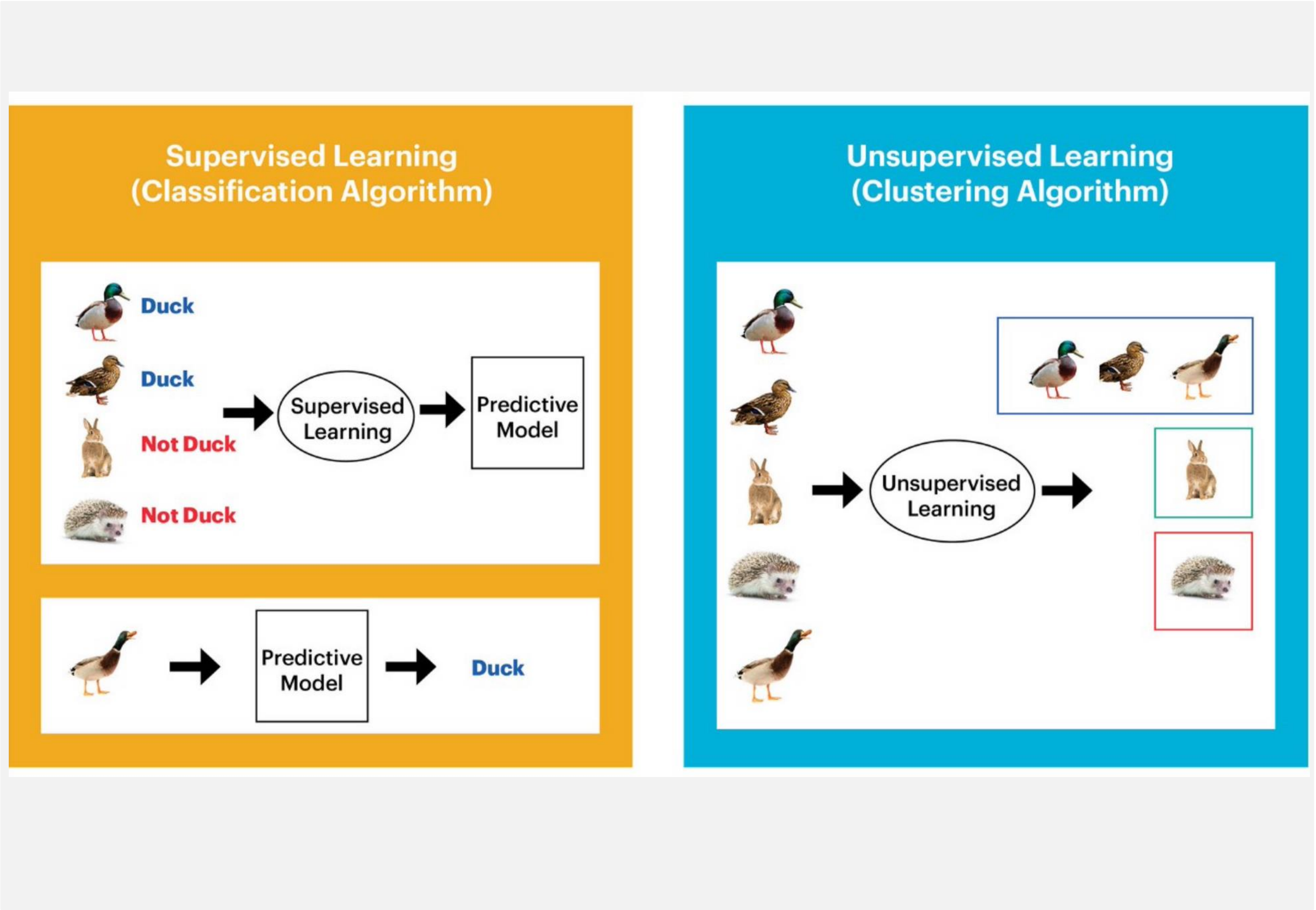
1) Supervised learning
2) Unsupervised learning
3) Semi-supervised learning
4) Reinforcement learning

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.
- **Clustering, data reduction**, **similarity matching,** and **association learning** problems can be solved with an unsupervised approach.


Supervised Learning
(Classification Algorithm)
Duck
Duck
Not Duck
Not Duck
Supervised Learning
Predictive Model
Predictive Model
Duck
Unsupervised Learning
(Clustering Algorithm)
Unsupervised Learning

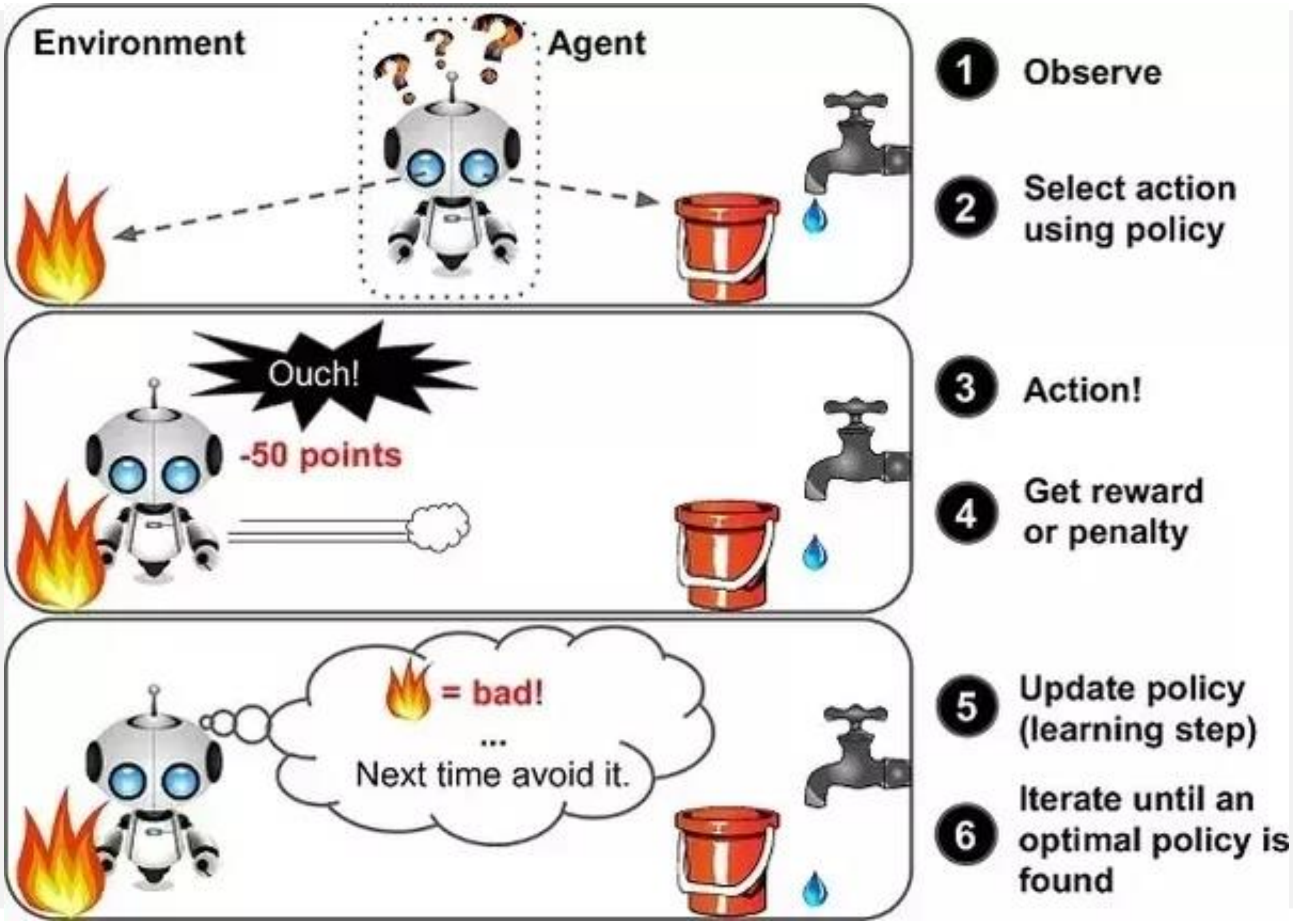
SEMI-SUPERVISED LEARNING

- It uses mainly **unlabeled** and a small amount of **labeled** input data.
- Using a small amount of labeled data can greatly increase the efficiency of unsupervised learning tasks.
- The model must learn the structure to organize the data as well as make predictions in **classification** and **regression** problems.

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

REINFORCEMENT LEARNING



MACHINE LEARNING VS STATISTICAL MODELS

- 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.

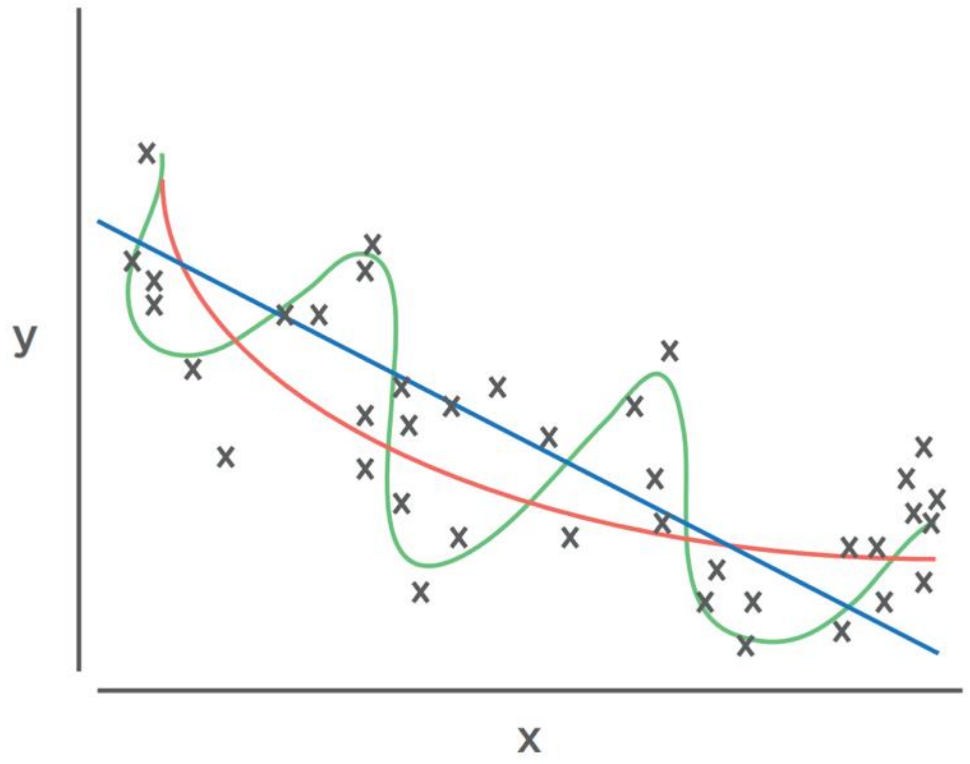
MACHINE LEARNING VS. STATISTICS

- 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

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)