Indian Institute of Foreign Trade

Day 1: Introduction to Data Analytics*

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Dr. Tanujit Chakraborty
ctanujit@gmail.com
https://www.ctanujit.org/DA.html

Course on Data Analytics for MBA (IB) Students.

---

* In 1962, John Tukey described a field called "data analysis", which resembles modern data science.

Recent Trends and Buzzwords

1. Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

2. Data science is the study of the generalizable extraction of knowledge from data, yet the keyword is science.

3. Machine learning is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed.

4. Artificial Intelligence research is defined as the study of intelligent agents: any device that perceives its environment and takes actions that maximize its chance of success at some goal.

5. Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.

Topic 1 : Statistics

Statistics: View Points

"Statistics is the universal tool of inductive inference, research in natural and social sciences, and technological applications. Statistics must have a clearly defined purpose, one aspect of which is scientific advance and the other, human welfare and national development"

- Professor P C Mahalanobis.

"All knowledge is, in final analysis, History.
All sciences are, in the abstract, Mathematics.
All judgements are, in their rationale, Statistics."

- Professor C R Rao.

- Role of Statistics:
  1. Making inference from samples
  2. Development of new methods for complex data sets
  3. Quantification of uncertainty and variability
- Two Views of Statistics:
  1. Statistics as a Mathematical Science
  2. Statistics as a Data Science

Statistics : Present

- Data : Large bodies of data with complex data structures are generated from computers, sensors, manufacturing industries, etc.
- Models : Non/Semiparametric models but in complex probability spaces / high-dimensional functional spaces (e.g., deep neural net, reinforcement learning, decision trees, etc.).
- Emphases : Making predictions, causation, algorithmic convergence.
- Data are necessary and at the core of Statistical Learning, Data Science & Machine Learning.
- Statistics : Not only has strong interactions with Probability but also other parts of Data Science (Machine Learning, Artificial Intelligence, etc.).

Statistics: Past Vs. Present

- Traditional Problems in Applied Statistics:

  - Well formulated question that we would like to answer.
  - Expensive to gather data and/or expensive to do computation.
  - Create specially designed experiments to collect high quality data.

- Current Situation : Information Revolution

  - Improvements in computers and data storage devices.
  - Powerful data capturing devices.
  - Lots of data with potentially valuable information available.

What is the Difference?

- Data characteristics:

  - Size
  - Dimensionality
  - Complexity
  - Messy
  - Secondary sources

- Focus on generalization performance :

  - Prediction on new data
  - Action in new circumstances
  - Complex models needed for good generalization

- Computational considerations :

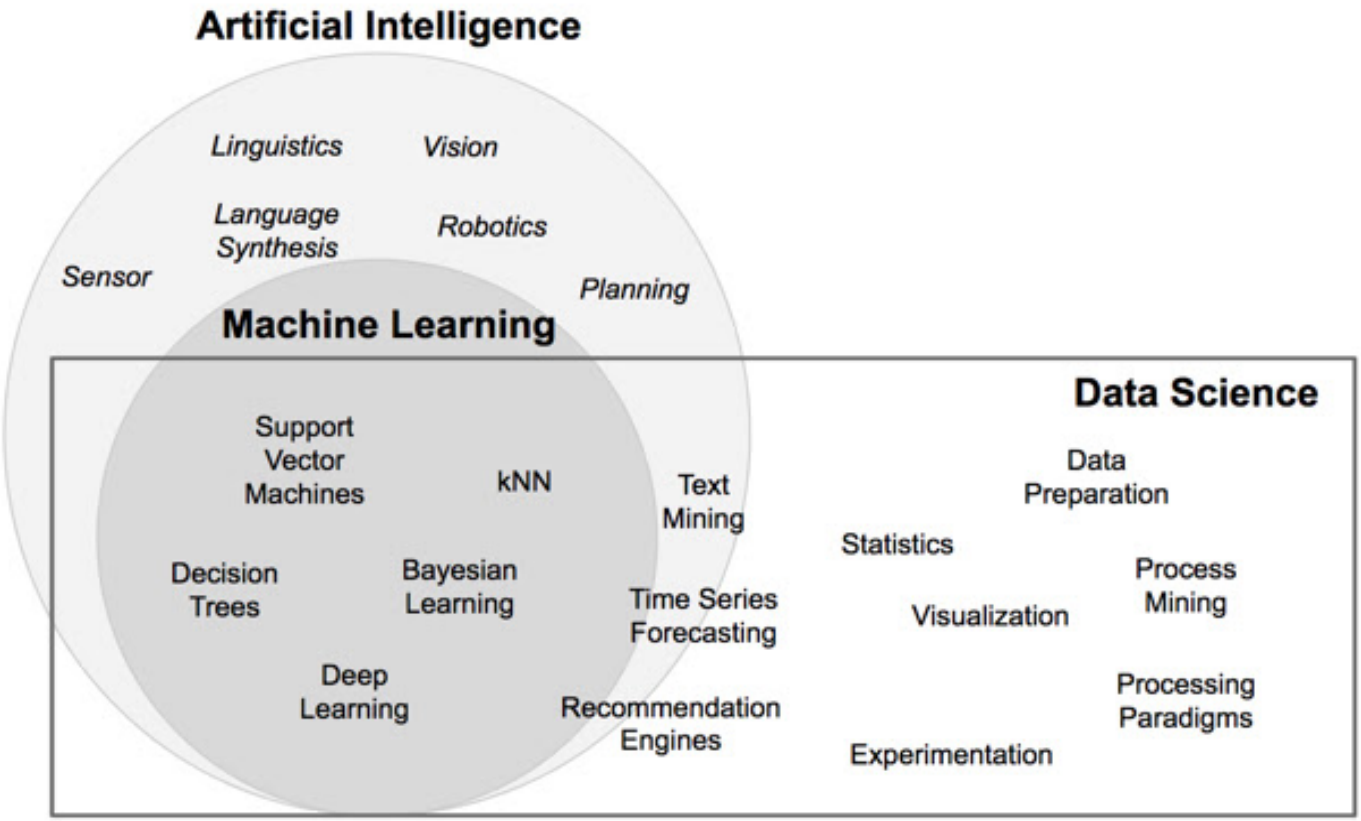
  - Large scale and complex systems

Topic 2 : Data Science

What is Data Science?

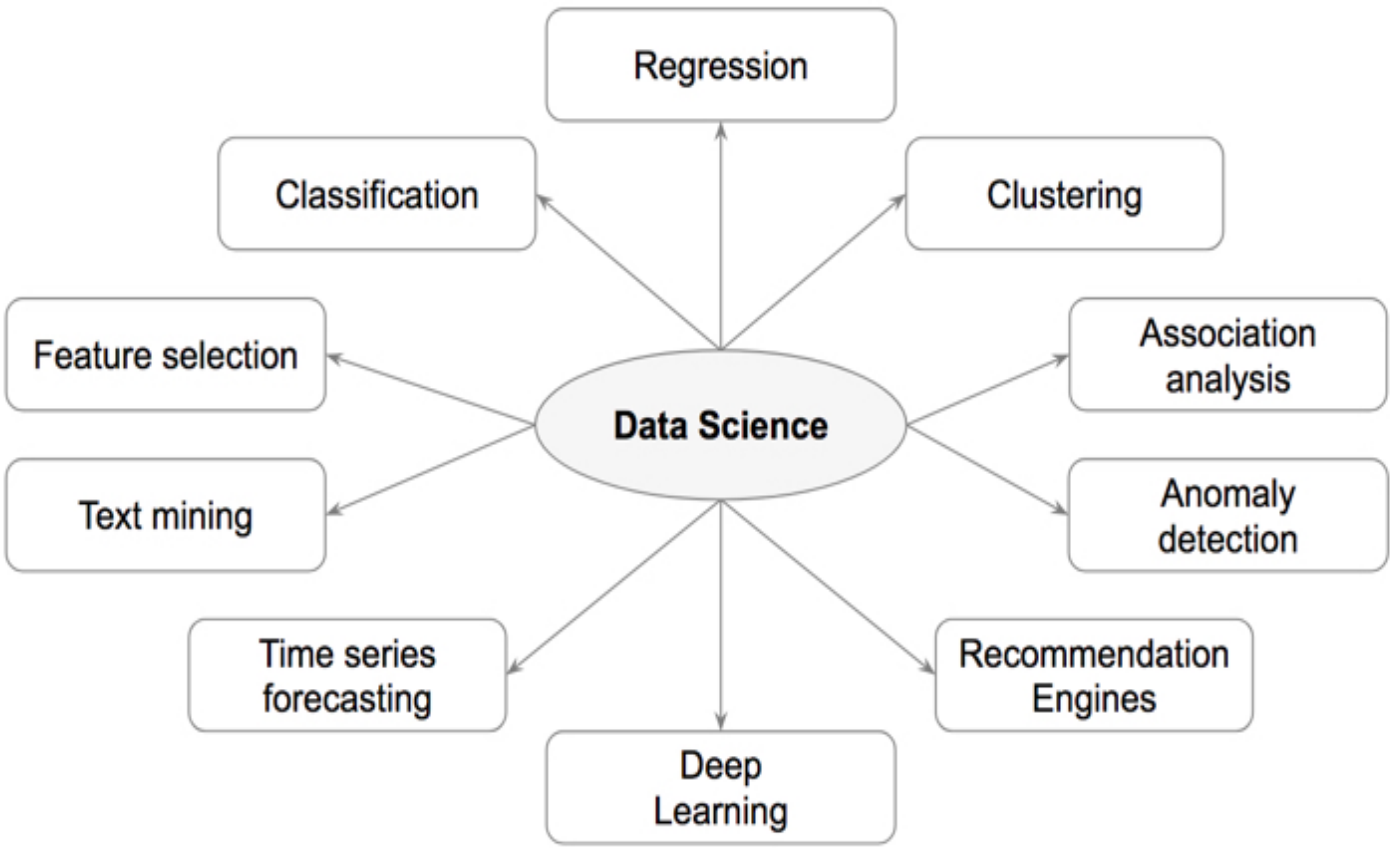
"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."

- Clive Humby, UK Mathematician and Architect of Tesco's Clubcard.



Types of Data Science?

"When you're fundraising, it's AI.
When you're hiring, it's ML.
When you're implementing, it's Linear Regression.
When you're debugging, it's printf()."
- Baron Schwartz, Founder and CEO of VividCortex, 2017.



Role of data: Present

The World is Data Rich

Astronomy

Social Networks

Healthcare

Banking

Genomics

Weather measurements

Data in Data Analytics

Basic Definitions:

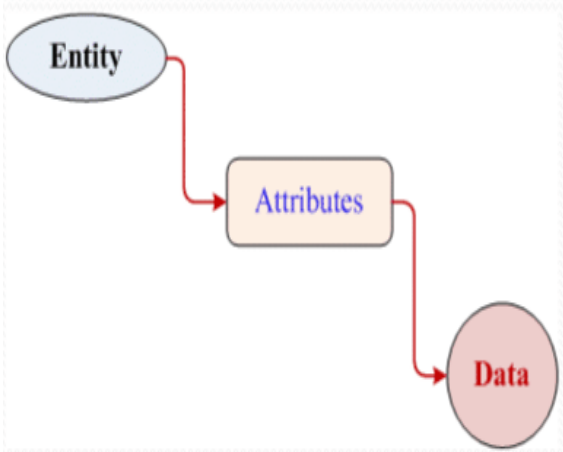
Entity: A particular thing is called entity or object.

Attribute: An attribute is a measurable or observable property of an entity.

Data: A measurement of an attribute is called data.

Note: Data defines an entity and Computer can manage all type of data (e.g., audio, video, text, etc.). In general, there are many types of data that can be used to measure the properties of an entity.

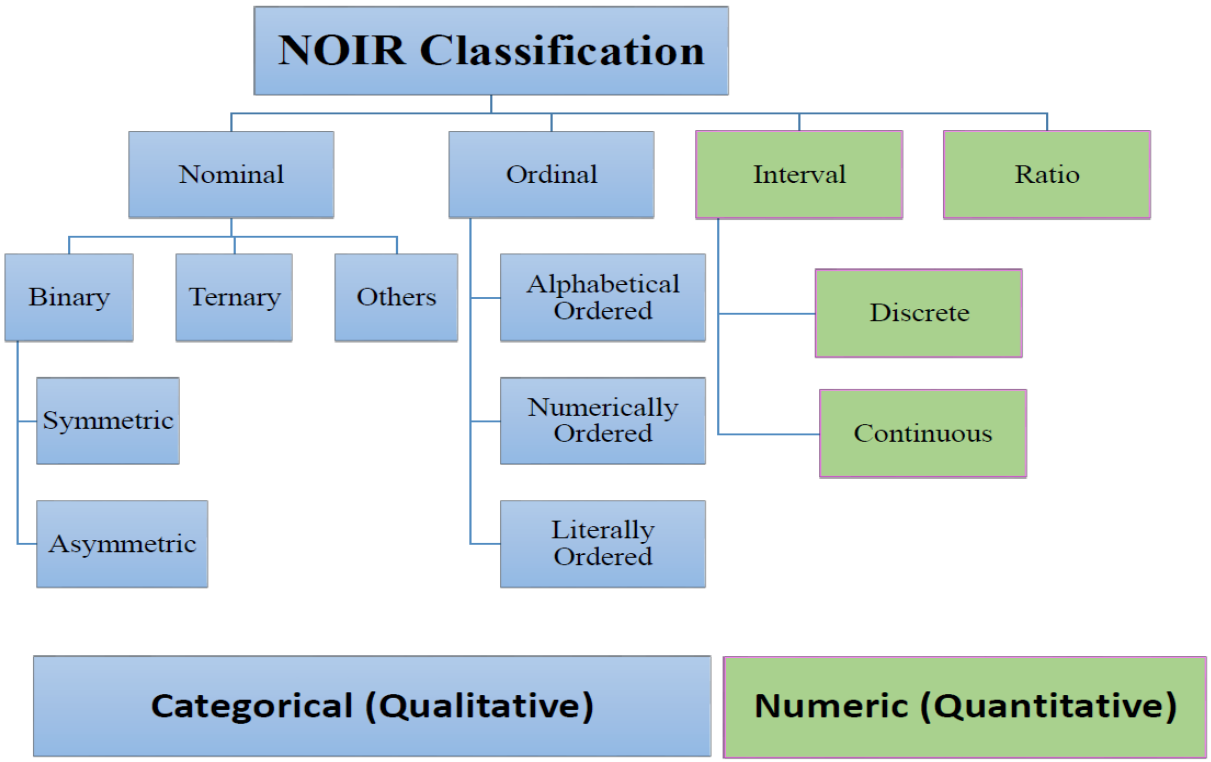
Scale: A good understanding of data scales (also called scales of measurement) is important. Depending on the scales of measurement, different techniques are followed to derive hitherto unknown knowledge in the form of patterns, associations, anomalies or similarities from a volume of data.



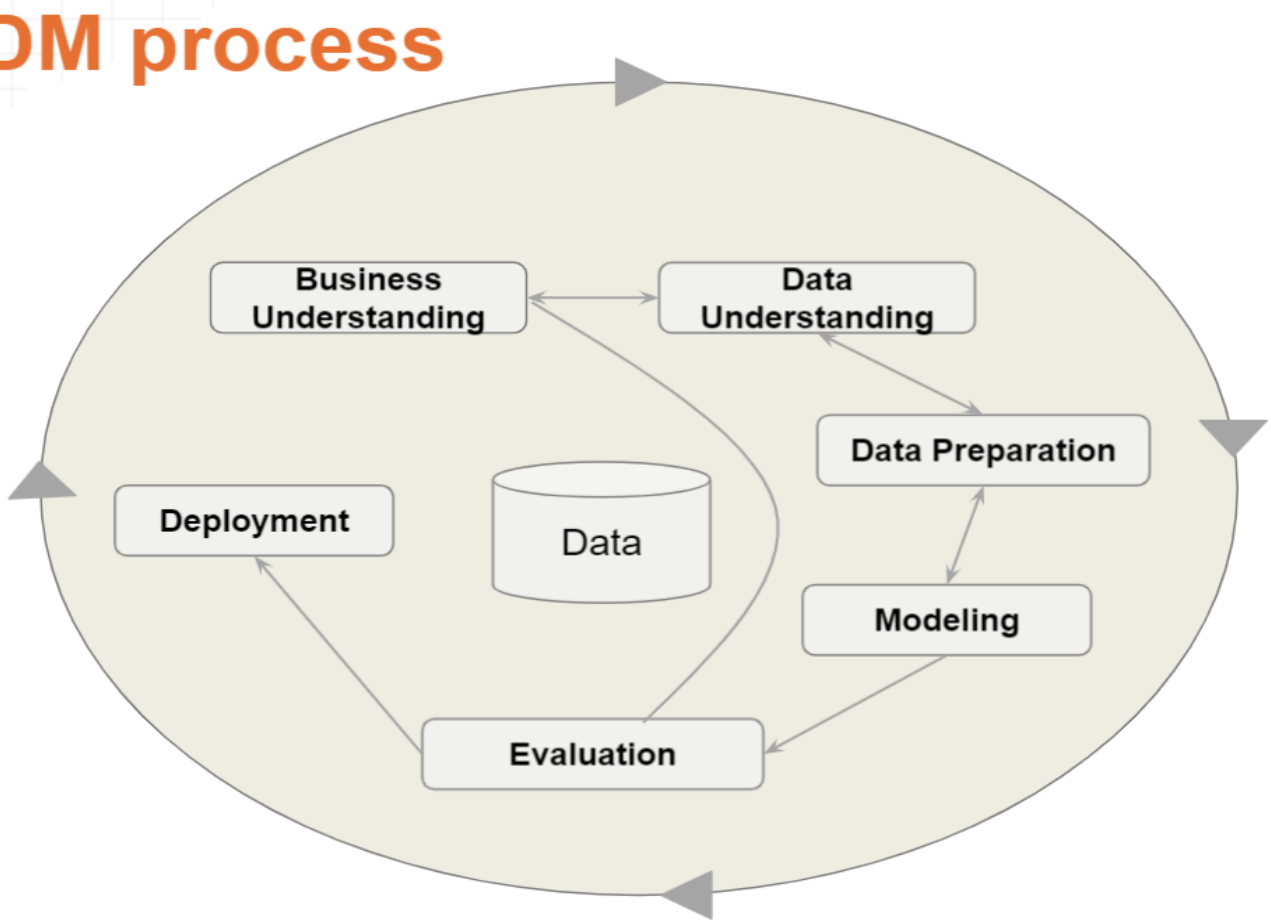
NAME	AGE	GENDER	SALARY	EMPLOYER
:				
:				
ABCD	34	F	40000	XYZ
:				
:				

NOIR: Scales of Measurement

- The NOIR scale is the fundamental building block on which the extended data types are built.
- Further, nominal (Blood groups, Attendance) and ordinal (Shirt size) are collectively referred to as categorical or qualitative data. Whereas, interval (weight, temperature) and ratio (Sound intensity in Decibel) data are collectively referred to as quantitative or numeric data.

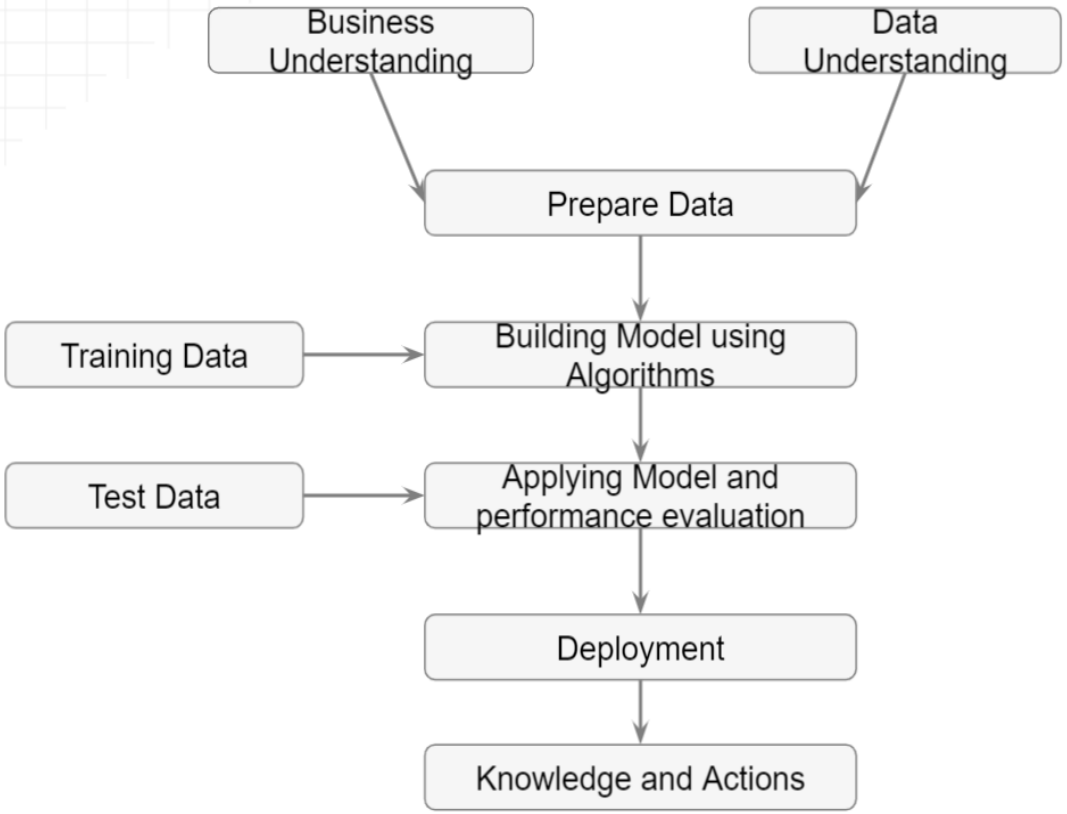


Data Mining Process



Data Mining Process

Process



1. Prior Knowledge
2. Preparation
3. Modeling
4. Application
5. Knowledge

1. Prior Knowledge

Gaining information on

- Objective of the problem.
- Subject area of the problem.
- Data.

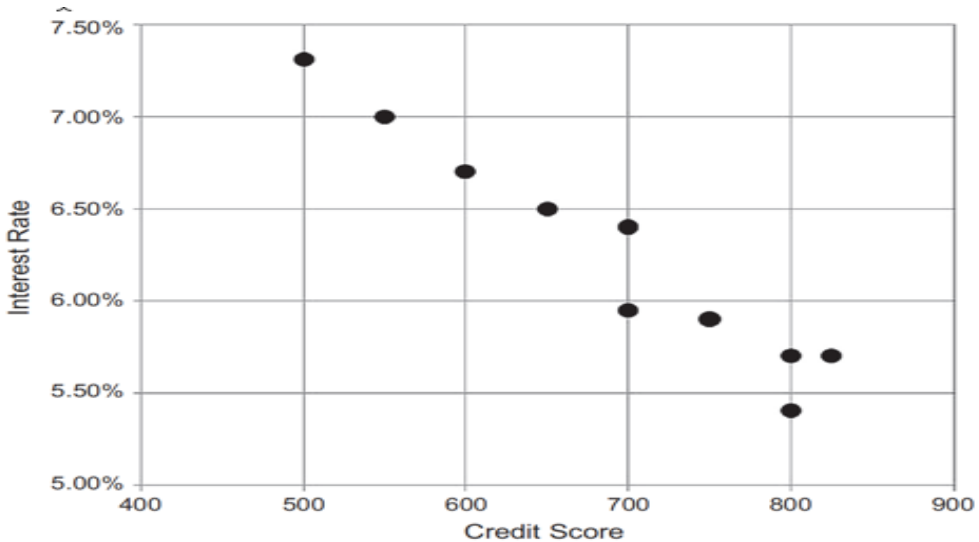
Table 2.1 Data Set

Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03	700	5.95%
04	700	6.40%
05	800	5.40%
06	800	5.70%
07	750	5.90%
08	550	7.00%
09	650	6.50%
10	825	5.70%

2. Data Preparation

Gaining information on

- Data Exploration and Data quality.
- Handling missing values and Outliers.
- Data type conversion.
- Transformation, Feature selection and Sampling.



3. Modeling

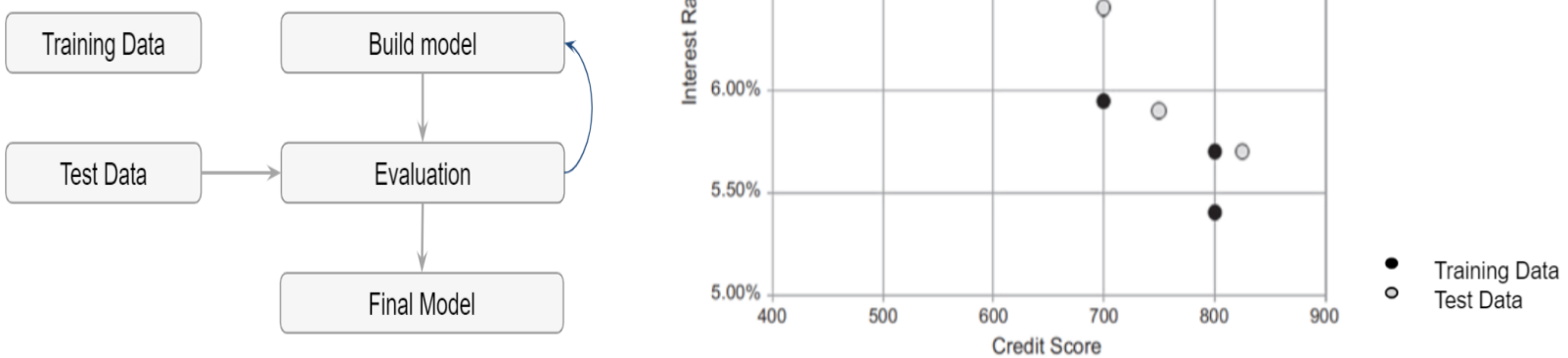
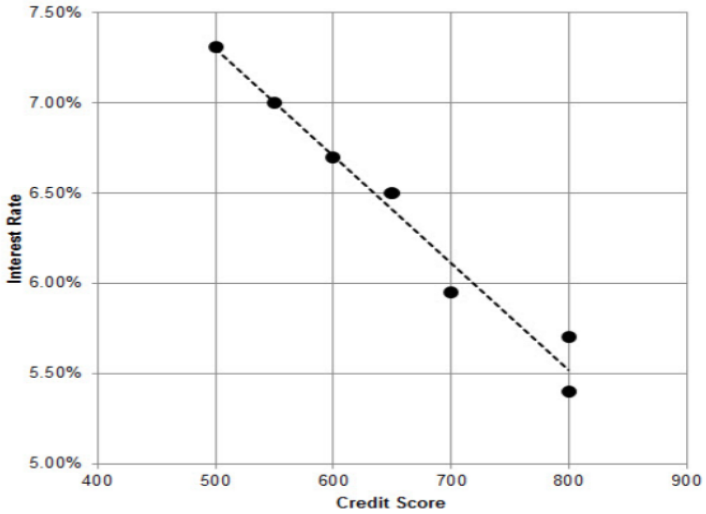


Figure: Splitting data into training and test data sets (right).

3. Modeling



$$y = 0.1 + \frac{6}{100,000}x$$

Table 2.5 Evaluation of Test Data Set

Borrower	Credit Score (X)	Interest Rate (Y)	Model Predicted (Y)	Model Error
04	700	6.40%	6.11%	-0.29%
07	750	5.90%	5.81%	-0.09%
10	825	5.70%	5.37%	-0.33%

Figure: Evaluation of test dataset (right).

Application and Knowledge

4. Application:

- Product readiness.
- Technical integration.
- Model response time.
- Remodeling.
- Assimilation.

5. Knowledge:

- Posterior knowledge.

Objectives of Data Exploration:

- Understanding data.
- Data preparation and Data mining tasks.
- Interpreting data mining results.

Data exploration

Roadmap:

- Organize the data set.
- Find the central point for each attribute (central tendency).
- Understand the spread of the attributes (dispersion).
- Visualize the distribution of each attributes (shapes).
- Pivot the data.
- Watch out for outliers.
- Understanding the relationship between attributes.
- Visualize the relationship between attributes.
- Visualization high dimensional data sets.
- For more details, read Kotu, V., Deshpande, B. (2014). Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann.

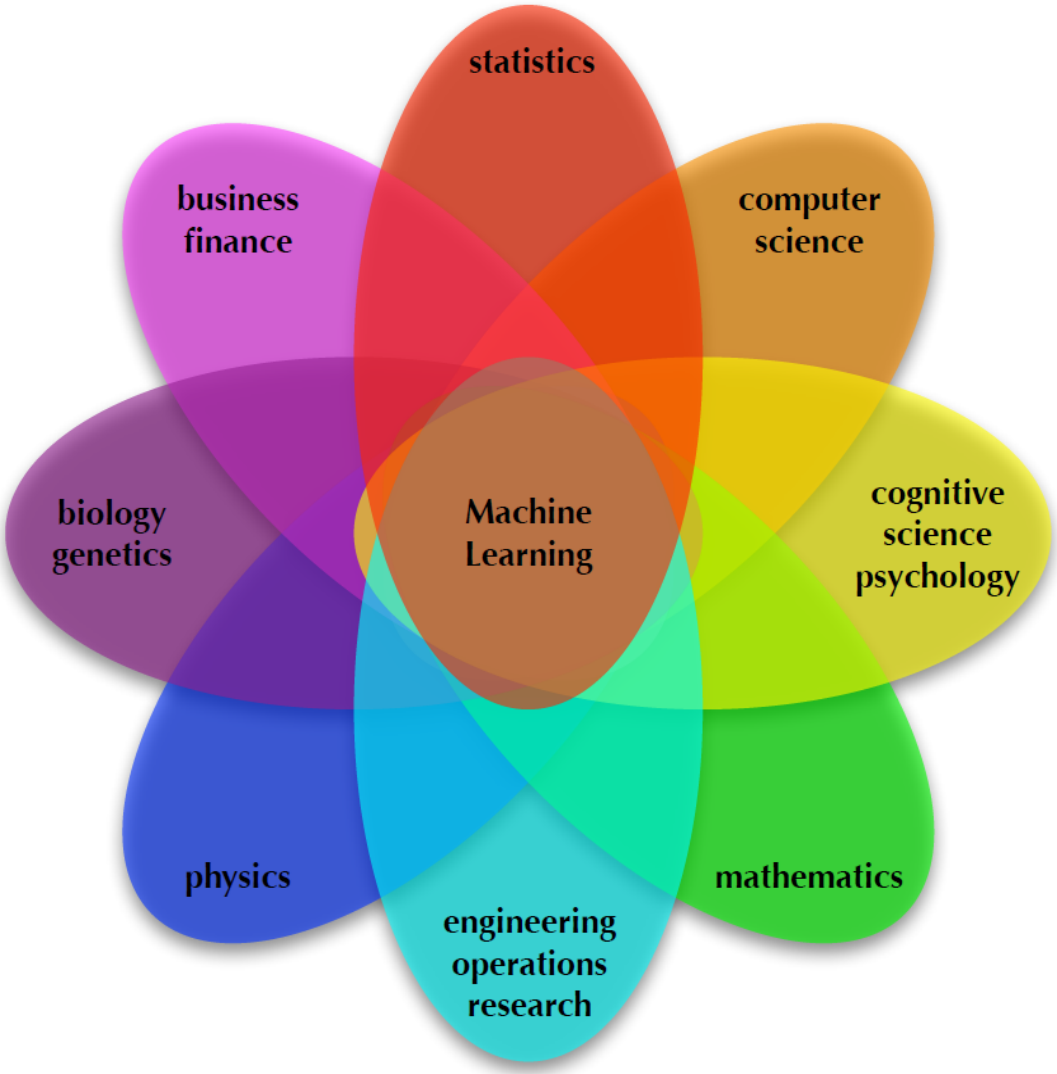
Overview of Data Science Tools

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, Density based, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.

Topic 3 : Machine Learning

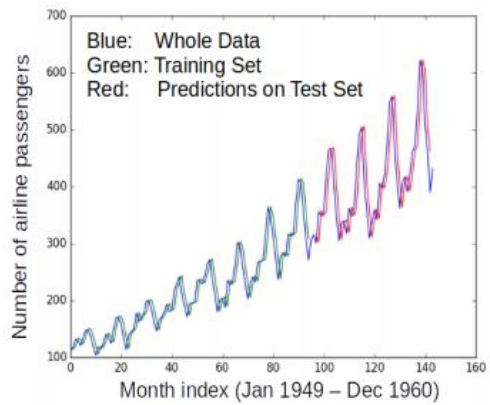
What is Machine Learning?

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.



Introduction to Machine Learning

- Designing algorithms that ingest data and learn a model of the data.
- The learned model can be used to
  1. Detect patterns/structures/themes/trends etc. in the data
  2. Make predictions about future data and make decisions



- Modern ML algorithms are heavily "data-driven".
- Optimize a performance criterion using example data or past experience.

Types of Machine Learning

- Unsupervised Learning:

  - Uncover structure hidden in 'unlabelled' data.
  - Given network of social interactions, find communities.
  - Given shopping habits for people using loyalty cards: find groups of 'similar' shoppers.
  - Given expression measurements of 1000s of genes for 1000s of patients, find groups of functionally similar genes.
  - Goal: Hypothesis generation, visualization.

- Supervised Learning:

  - A database of examples along with 'labels' (task-specific).
  - Given expression measurements of 1000s of genes for 1000s of patients along with an indicator of absence or presence of a specific cancer, predict if the cancer is present for a new patient.
  - Given network of social interactions along with their browsing habits, predict what news might users find interesting.
  - Goal: Prediction on new examples.

Types of Machine Learning

- Semi-supervised Learning:

  - A database of examples, only a small subset of which are labelled.

- Multi-task Learning:

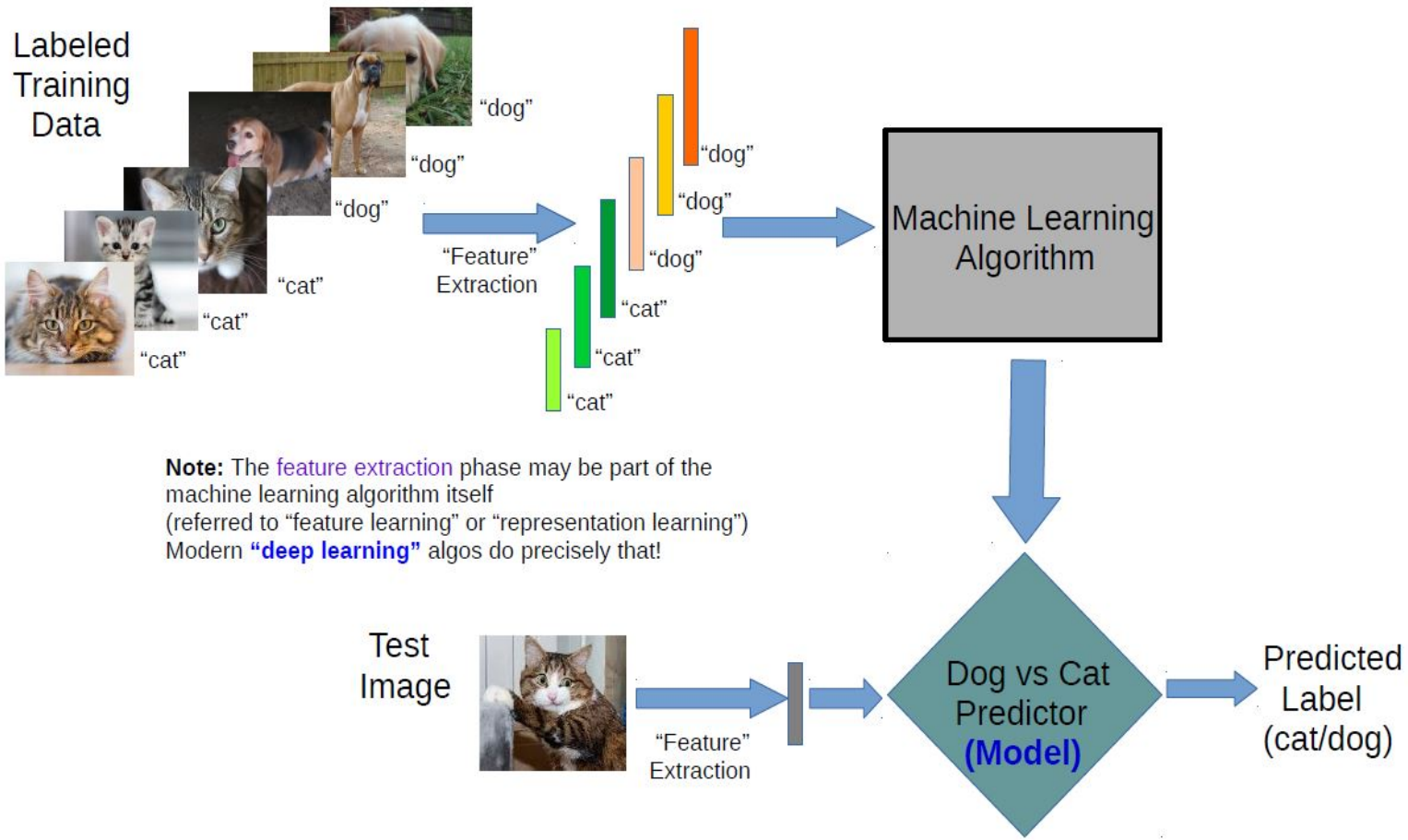
  - A database of examples, each of which has multiple labels corresponding to different prediction tasks.

- Reinforcement Learning:

  - An agent acting in an environment, given rewards for performing appropriate actions, learns to maximize its reward.

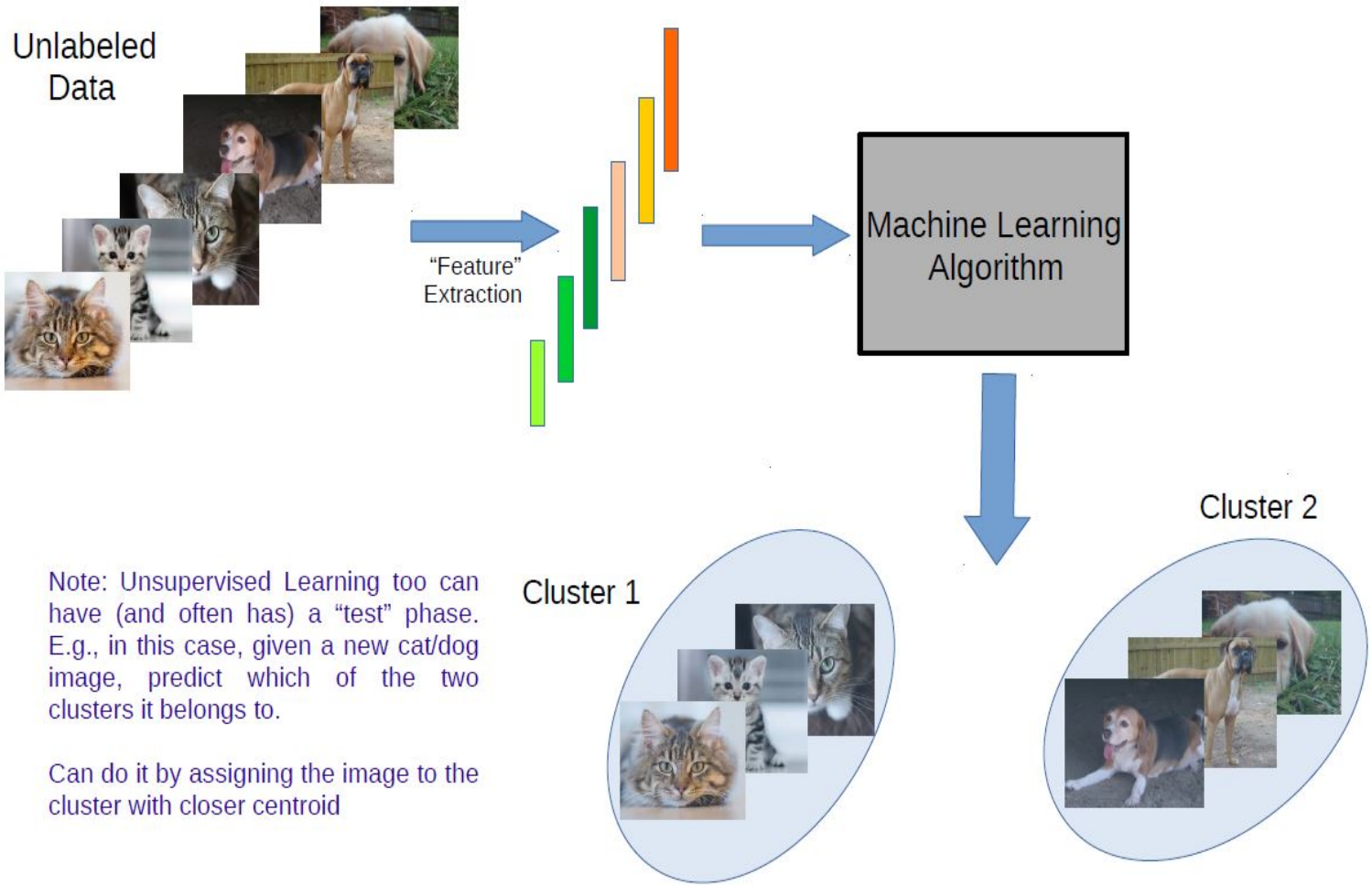
A Typical Supervised Learning Workflow

Supervised Learning: Predicting patterns in the data



A Typical Unsupervised Learning Workflow

Unsupervised Learning: Discovering patterns in the data

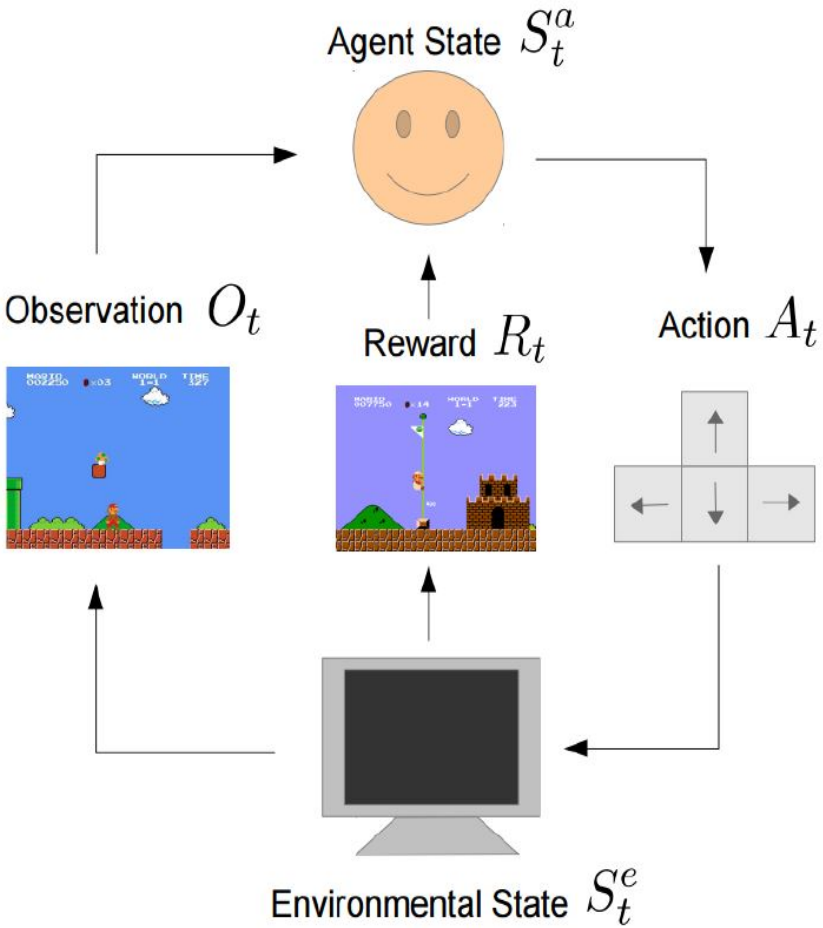


Note: Unsupervised Learning too can have (and often has) a "test" phase. E.g., in this case, given a new cat/dog image, predict which of the two clusters it belongs to.

Can do it by assigning the image to the cluster with closer centroid

A Typical Reinforcement Learning Workflow

Reinforcement Learning: Learning a "policy" by performing actions and getting rewards (e.g, robot controls, beating games)



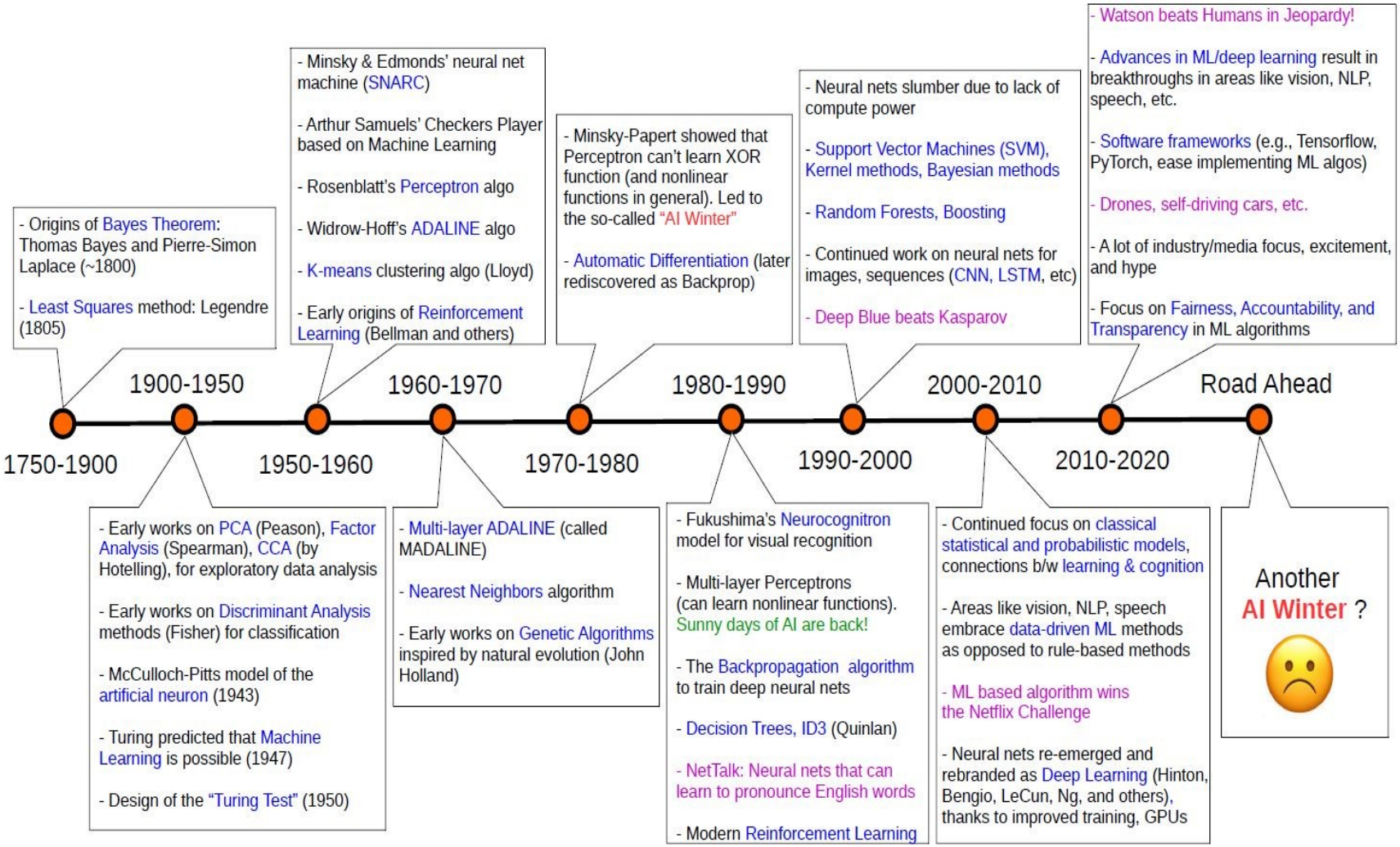
Agent's goal is to learn a policy for some task

Agent does the following repeatedly

- Senses/observes the environment
- Takes an action based on its current policy
- Receives a reward for that action
- Updates its policy

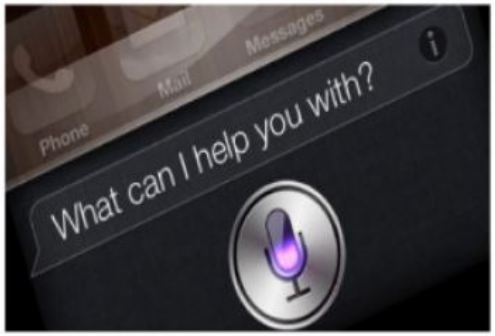
There IS supervision, not explicit (as in Supervised Learning) but rather implicit (feedback based)

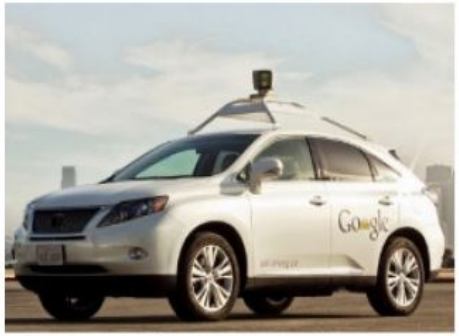
Machine Learning: A Brief Timeline



Machine Learning in the real-world

Broadly applicable in many domains (e.g., internet, robotics, healthcare and biology, computer vision, NLP, databases, computer systems, finance, etc.).






Predictive Policing

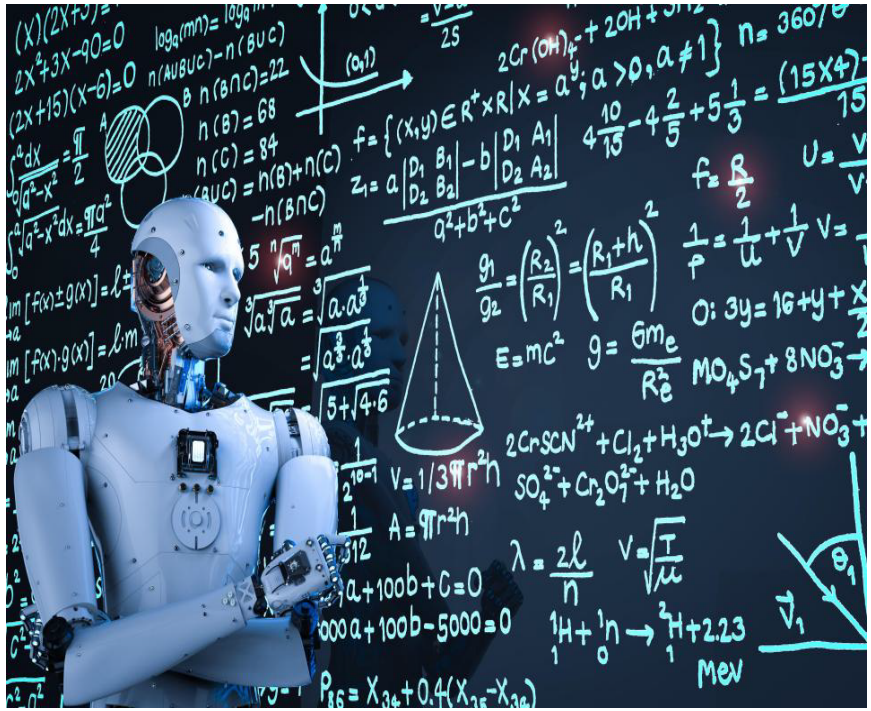

Online Fraud Detection

Topic 4 : Artificial intelligence

A first look at Artificial Intelligence

- What is Artificial Intelligence?
- What are the main challenges?
- What are the applications of AI?
- What are the issues raised by AI?

- On September 1955, a project was proposed by McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon introducing formally for the first time the term "Artificial Intelligence".



Misconception of AI

- AI is about electronic device able to mimic human thinking:
  1. Artificial Intelligence.
  2. One famous class of AI algorithms are called neural networks.
  3. Android are close to humans in shape so they must think like human.
- Most AI algorithms do not aim at reproducing human reasoning.
- "The study and design of intelligent agents" where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success - Frequent definition of AI.
- "In from three to eight years we will have a machine with the general intelligence of an average human being." - Marvin Minsky (1970, Life Magazine).

AI is not human intelligence

"What often happens is that an engineer has an idea of how the brain works (in his opinion) and then designs a machine that behaves that way. This new machine may in fact work very well. But, I must warn you that it does not tell us anything about how the brain actually works, nor is it necessary to ever really know that, in order to make a computer very capable. It is not necessary to understand the way birds flap their wings and how the feathers are designed in order to make a flying machine [...] ***It is therefore not necessary to imitate the behavior of Nature in detail in order to engineer a device which can in many respects surpass Nature's abilities.***"

- Richard Feynman (1999).

AI technology - Autonomous cars

- Originates from 1920 (NY)
- First use of neural networks to control autonomous cars (1989)
- Four US states allow self-driving cars (2013)
- First known fatal accident (May 2016)
- Singapore launched the first self-driving taxi service (Aug. 2016)
- A Arizona pedestrian was killed by an Uber self-driving car (March 2018).

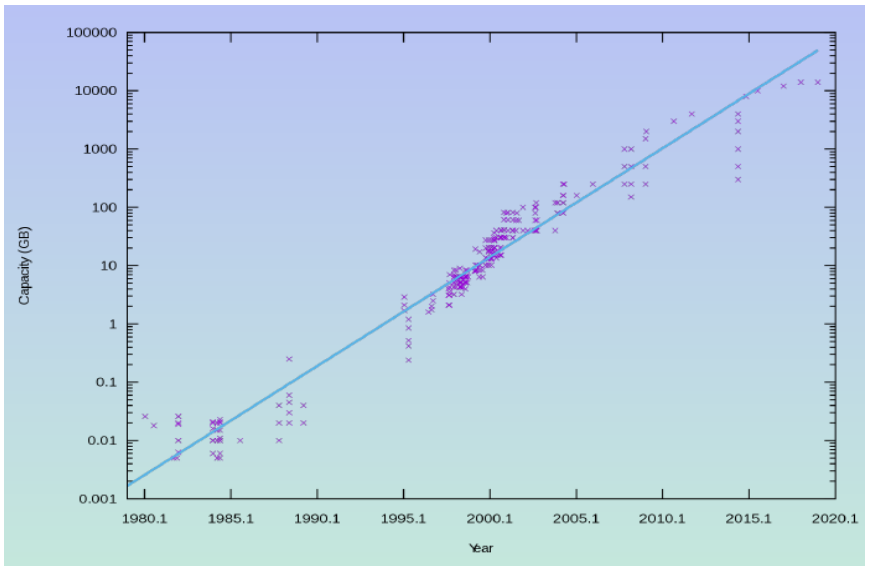
AI technology - virtual assistant / chatbot

- Voice recognition tool "Harpy" masters about 1000 words (1970s, CMU, US Defense).
- System capable of analyzing entire word sequences (1980).
- Siri was the first modern digital virtual assistant installed on a smartphone (2011).
- Watson won the TV show Jeopardy! (2011).

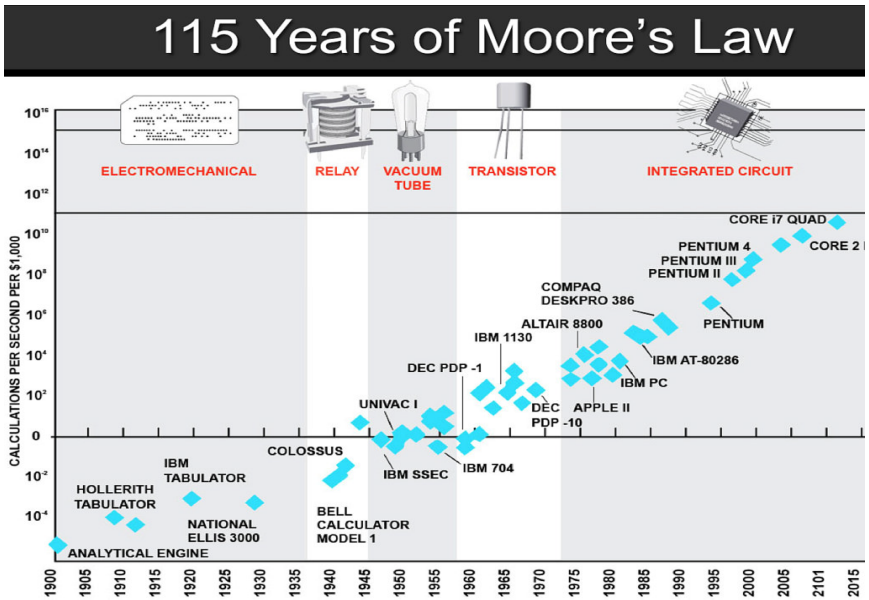
Topic 5 : Big Data

Storage and Processing capacities

- Kryder's Law is the assumption that disk drive density, also known as areal density, will double every thirteen months. The implication of Kryder's Law is that as areal density improves, storage will become cheaper.
- Moore's Law refers to Moore's perception that the number of transistors on a microchip doubles every two years. Moore's Law states that we can expect the speed and capability of our computers to increase every couple of years, and we will pay less for them.



Storage capacity (Kryder's law)



Processor capacity (Moore's law)

How large your data is?

- What is the maximum file size you have dealt so far? (Movies/files/streaming video that you have used)
- What is the maximum download speed you get? (To retrieve data stored in distant locations?)
- How fast your computation is? (How much time to just transfer from you, process and get result?)
- "Every day, we create 2.5 quintillion bytes of data in 2020" (So much that 90% of the data in the world today has been created in the last two years alone).

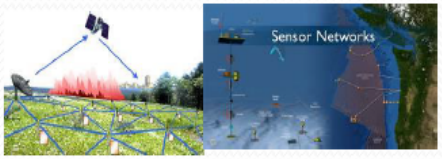
Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Data Source: Examples

Social media and networks
(All of us are generating data)

Scientific instruments
(Collecting all sorts of data)

Mobile devices
(Tracking all objects all the time)



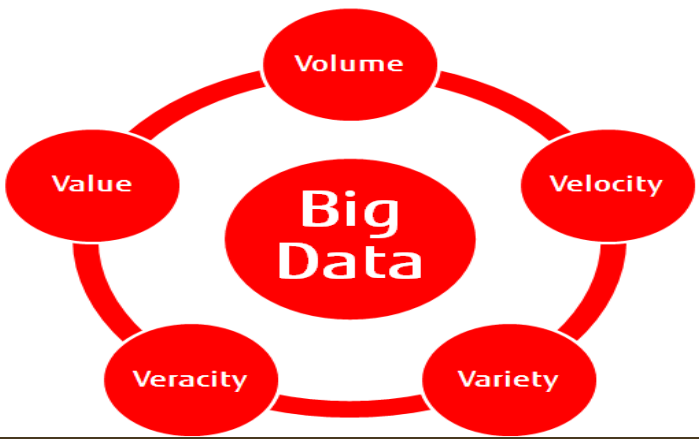
Sensor technology and networks
(Measuring all kinds of data)

Now data is Big data!

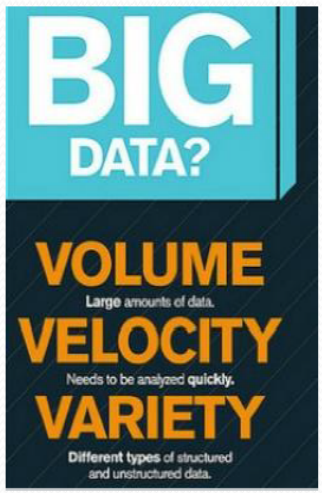
"Big data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it" - Standard definition.

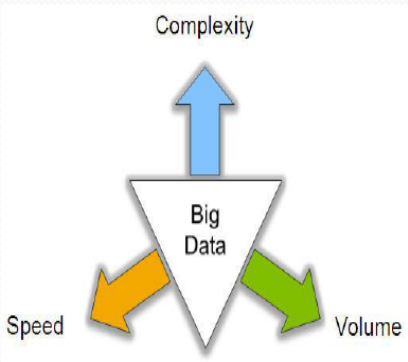
Difficulties related to (Big) data::

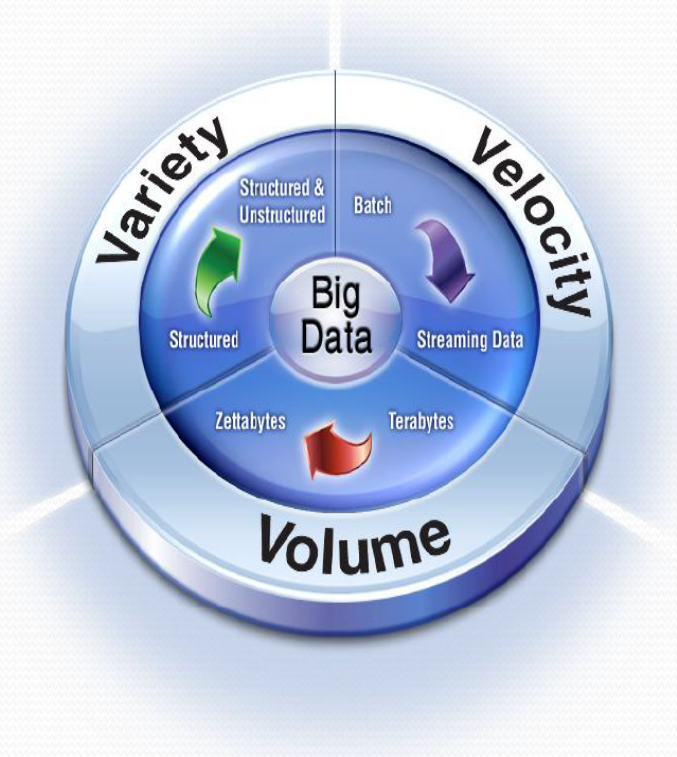
- The prediction must be accurate: difficult for some tasks like image classification, video captioning...
- The prediction must be quick: online recommendation should not take minutes.
- Data must be stored and accessible easily.
- It may be difficult to access all data at the same time. Data may come sequentially.



Characteristics of Big data: V3







Big data: V3

Volume:

- Volume of data that needs to be processed is increasing rapidly.
- Need more storage capacity.
- Need more computation facility.
- Need more tools and techniques.

Variety:

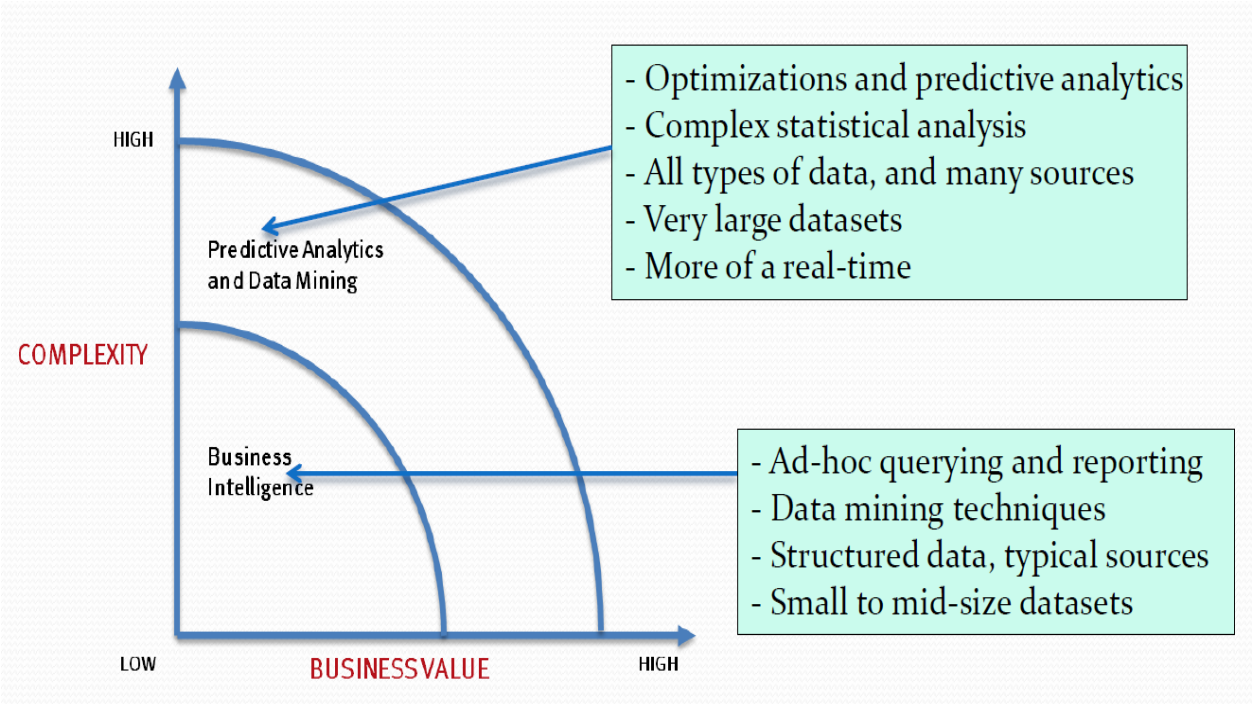
- Various formats, types, and structures.
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dimensional arrays, etc.
- A single application can be generating/collecting many types of data.

Velocity:

- Data is being generated fast and need to be processed fast.
- For time sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.
- Analyze 500 million daily call detail records in real-time to predict customer churn faster.

Big data vs. Small data

Big data is more real time in nature than traditional applications...



Challenges ahead…

- The Bottleneck is in technology: New architecture, algorithms, techniques are needed.
- Also in technical skills: Experts in using the new technology and dealing with Big data
- Who are the major players in the world of Big data?
- Ethical issues: Tay ("thinking about you") was an AI released by Microsoft via Twitter in 2016. It was shut down when the bot began to post in inflammatory and offensive tweets, only 16 hours after its launch.

Far from terminator

- Stephen Hawking BBC, Dec 2 2014
  The development of full artificial intelligence could spell the end of the human race. We cannot quite know what will happen if a machine exceeds our own intelligence, so we can't know if we'll be infinitely helped by it, or ignored by it and sidelined, or conceivably destroyed by it.

Textbook and References