Sorbonne AD

Day 1: Business Data Analytics*

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Dr. Tanujit Chakraborty
Assistant Professor (Machine Learning) @ Sorbonne.
https://www.ctanujit.org/MDA.html

Quote of the day

"It is very easy to be a teacher, but very difficult to be a student.

A good student has to learn many concepts, perform in examinations, loyal to his / her teacher and others."

Why this course?

Let data drive decisions, not the **HIghest Paid Person's Opinion**.

What is Data?

"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.

Example:

City	Morning temperature	Evening temperature
Austin	62	90.7
Boston	41	48.0
Chicago	51	57.2
Denver	45	52.5

Role of data: Present

Role of data: Present

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

- Sherlock Holmes, in A Scandal in Bohemia.

The World is Data Rich

Astronomy

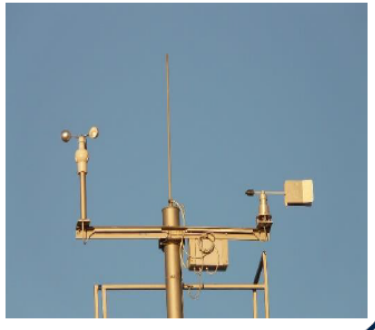
Social Networks

Healthcare

Banking

Genomics

Weather measurements



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 key word 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."

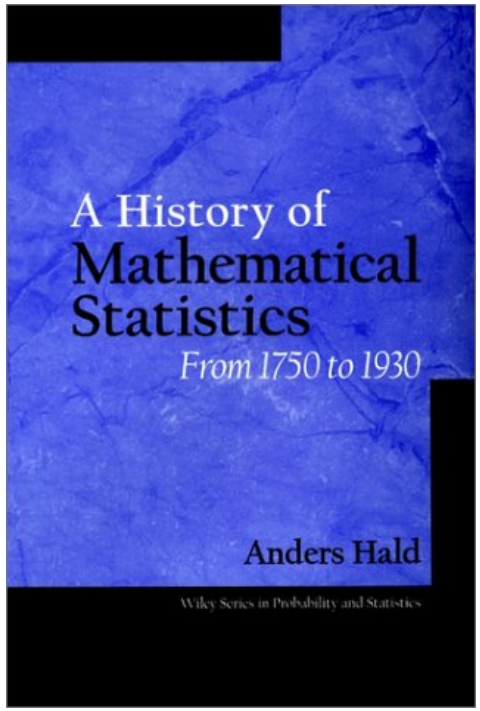
- Professor C R Rao.

- 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: View Points

In 1925, R A Fisher published Statistical Methods for Research Workers in the Biological Monographs and Manuals Series by the publisher Oliver and Boyd of Edinburgh in Scotland. In the first few pages of the book Fisher gives an Introduction to statistics and its methods. We give below a part of this Introduction:-

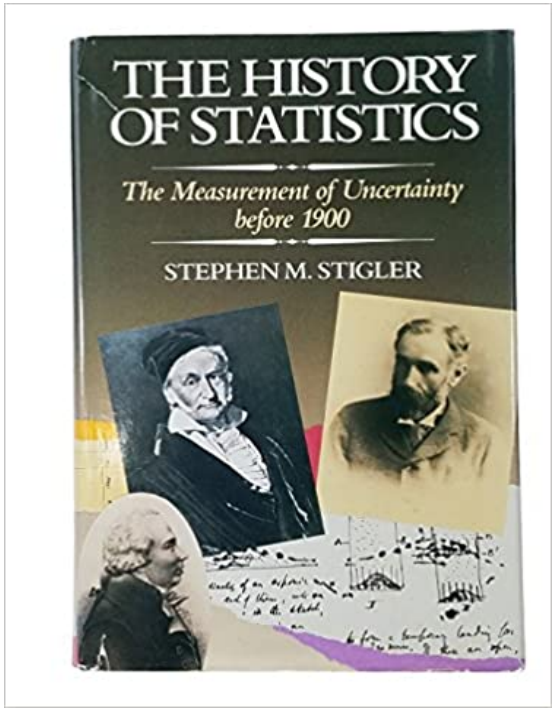
The science of Statistics is essentially a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data. As in other mathematical studies, the same formula is equally relevant to widely different groups of subject-matter. Consequently the unity of the different applications has usually been overlooked, the more naturally because the development of the underlying mathematical theory has been much neglected. Statistics may be regarded as (i) the study of populations, (ii) as the study of variation, (iii) as the study of methods of the reduction of data.



History of Statistics : Pre 1900

- Probability and its application to Gambling & Astronomy

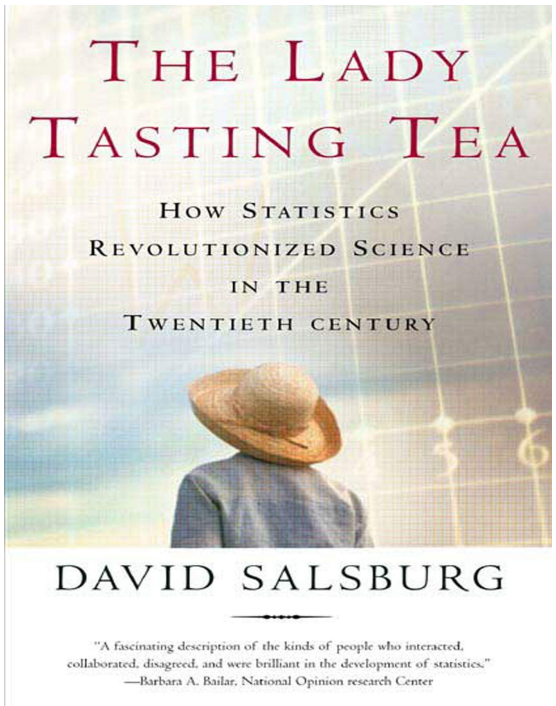
  - Jacob Bernoulli (Prob. distribution, Law of large numbers, etc.)
  - Pierre-Simon Laplace (Double exponential, Transformation, etc.)
  - Thomas Bayes (Bayes' theorem, etc.)
  - Gauss and Legendre (Least Square Method)
  - Francis Galton (Correlation & Regression)
  - Karl Pearson (χ^2 test, distribution, etc.)
  - and many others.



History of Statistics : 1900 - 1980

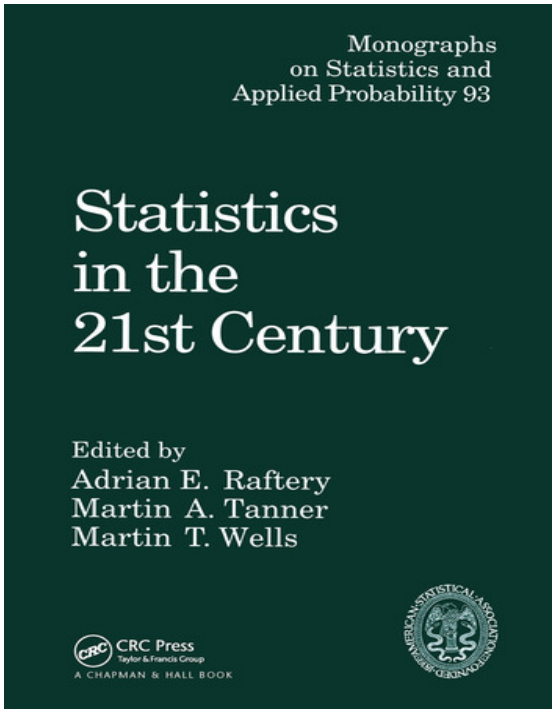
- Development of Statistics and its application to Agriculture, Economics, Geology, Medical, Technology, Clinical Trials, etc.

  - Ronald Fisher (Discriminant analysis, Likelihood, ANOVA & DOE, etc.)
  - Jerzy Neyman, Egon Sharpe Pearson & Wald (Decision theory, Optimality, etc.)
  - Lehmann, Hotelling, Anderson & Tukey (Multivariate & Inferential Statistics, etc.)
  - Box, Cox, Jenkin and Blackwell (Time Series)
  - Shewhart, Deming, Taguchi & Juran (SQC)
  - Efron, Breiman, Friedman, Cramer (Modern Statistical Tools)
  - PC Mahalanobis (Mahalanobis Distance), C. R. Rao (Linear Models, Multivariate Analysis, Orthogonal arrays, etc.), etc.



History of Statistics : Upto 1980

- Parametric Models : One Sample, two sample, linear models, survival data, Estimation, Testing of Hypothesis.
- Probability distributions were believed to generate data (e.g., Gaussian, Logistic, Poisson, Exponential, etc.).
- Semiparametric & Nonparametric Models : Dropping assumptions on population, dependence and errors.
- Emphases on Optimality in various ways : Bayes optimality, Decision theory, minimax and unbiasedness.
- Exact distributional (t, F) approaches and asymptotic methods (samples size $\rightarrow \infty$ viewed as approximation).



Statistics : 1980 - 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 : Present (2021)

- Probability : Has moved to the center of Mathematics and having strong interactions with Statistical Physics and Theoretical Computer Science.
- Statistics : Not only has strong interactions with Probability but also other parts of Data Science (Machine Learning, Artificial Intelligence, etc.).
- Computational : Computing skills are essential, the construction of fast training algorithms and computation time.
- Applications : Strong interactions with substantive fields in all areas. Applications of statistical methods in almost all fields are evident. Statistics became a key technology driven by data (**“Data is the new oil"**).

Statistics: Past Vs. Present

- 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.

"Everybody needs data literacy, because data is everywhere. It's the new currency, it's the language of the business. We need to be able to speak that."

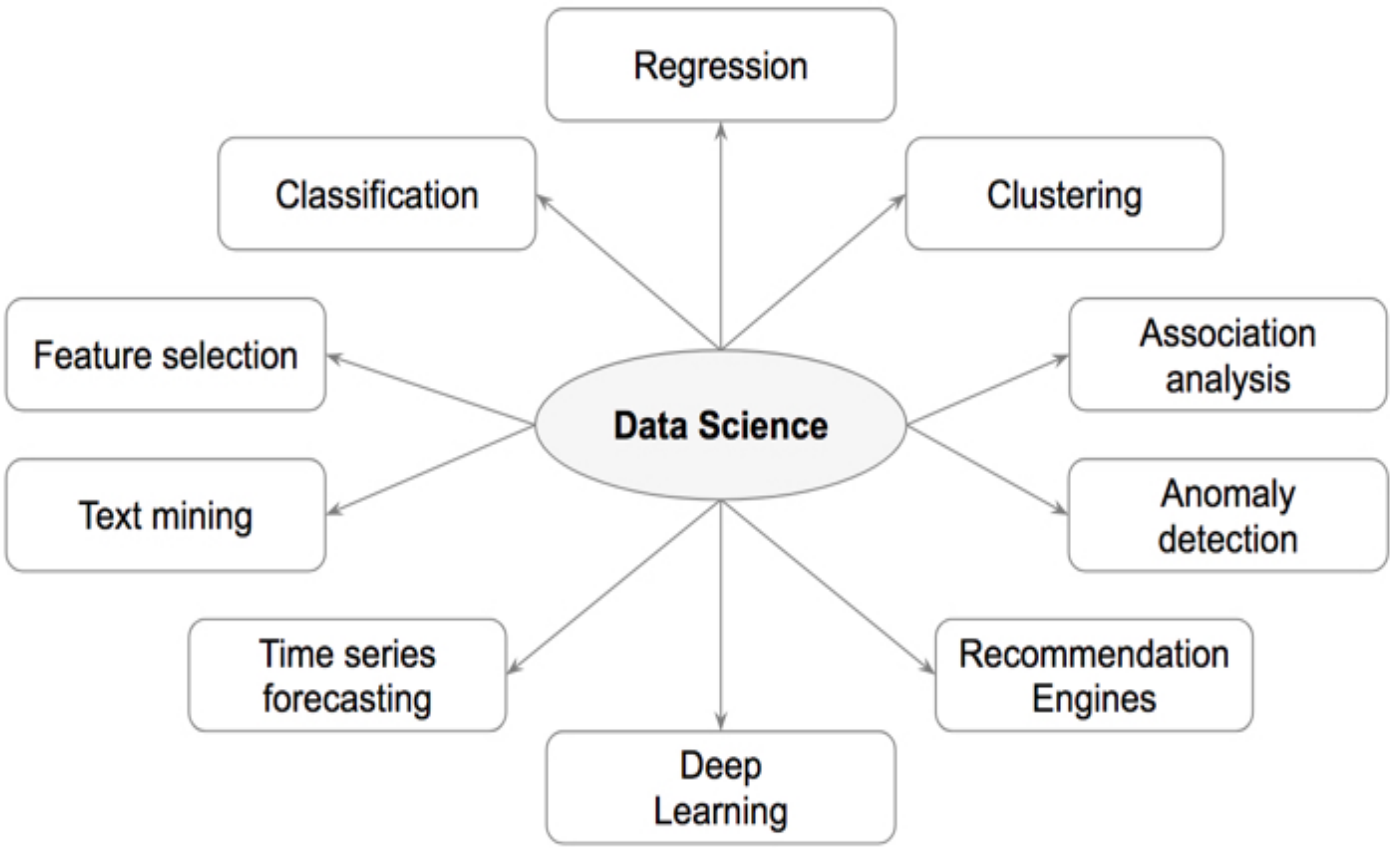
- MIT Sloan School of Management

What is Data Science?

- *Interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from complex data in various forms, either structured or unstructured.*
- *A concept to unify Statistics, Data Analysis and their related methods in order to "understand and analyze actual phenomena" with data.*
- *Employs techniques and theories drawn from many fields within the broad areas of Mathematics, Statistics, Information Science, and Computer Science, in particular from the subdomains of Machine learning, classification, cluster analysis, data mining, databases, and visualization.*
- *Fourth paradigm of Science (empirical, theoretical, computational and data-driven)*

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.



Data in Data Analytics

Basic Definitions:

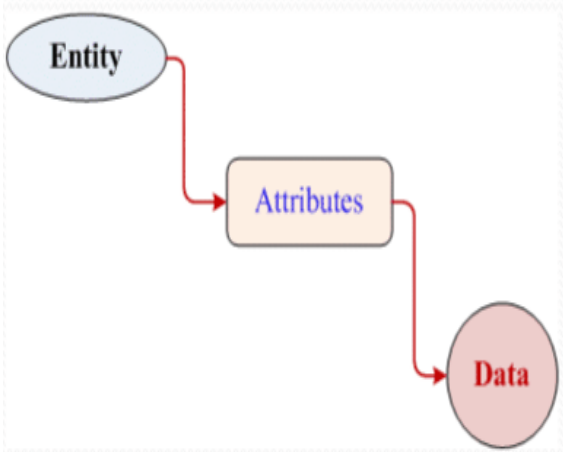
Entity: A particular thing is called entity or object.

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.

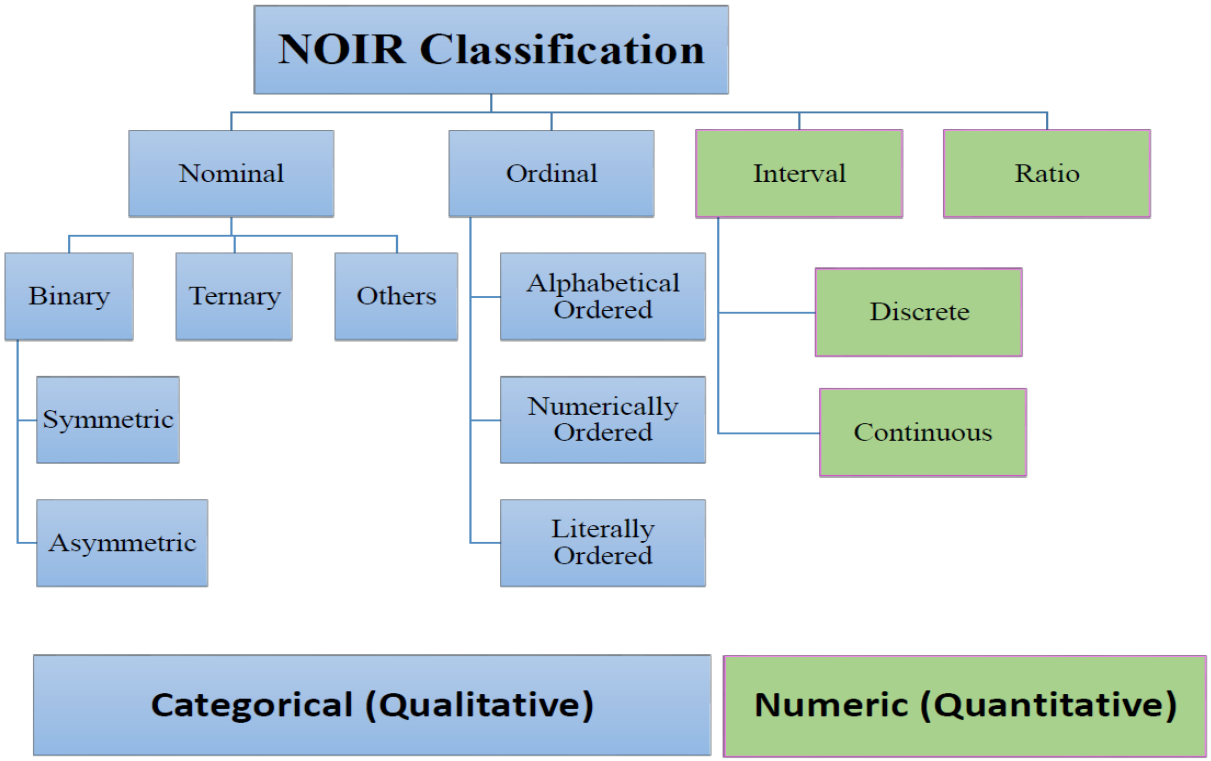
Scale: A good understanding of data scales (also called scales of measurement) is important. Depending on the scales of measurement, different techniques are followed to derive hitherto unknown knowledge in the form of patterns, associations, anomalies or similarities from a volume of data.



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 Cube

Concept of data cube:

A multidimensional data model views data in the form of a cube. A data cube is characterized with two things:

- Dimension: The perspective or entities with respect to which an organization wants to keep record.
- Fact: The actual values in the record.

Example: Rainfall data of Meteorological Department.

- Time (Year, Season, Month, Week, Day, etc.)
- Location (Country, Region, State, etc.)

2-D view of rainfall data

- In this 2-D representation, the rainfall for “North-East” region are shown with respect to different months for a period of years...

Reagion: North-East

Year \ Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2005												
2006												
2007												
2008												
2009												
2010												

View of 3-D rainfall data

- Suppose, we want to represent data according to times (Year, Month) as well as regions of a country say East, West, North, North-East, etc.

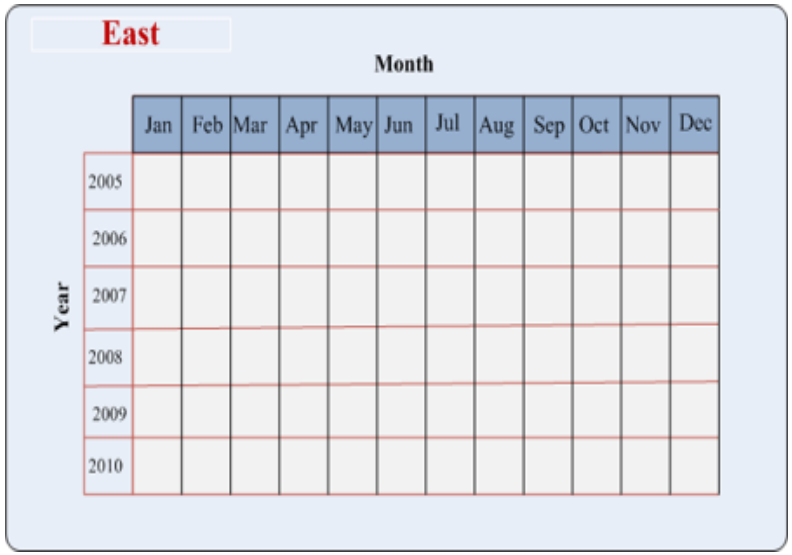


Figure: 2-D view of rainfall data

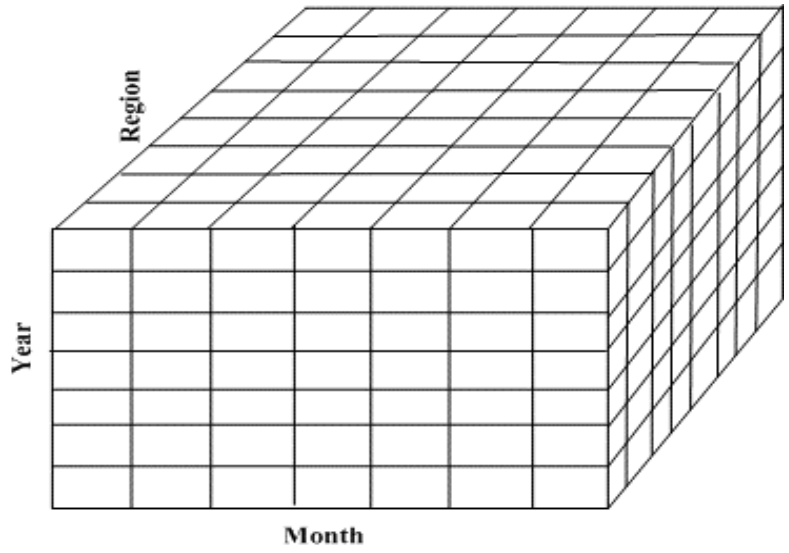
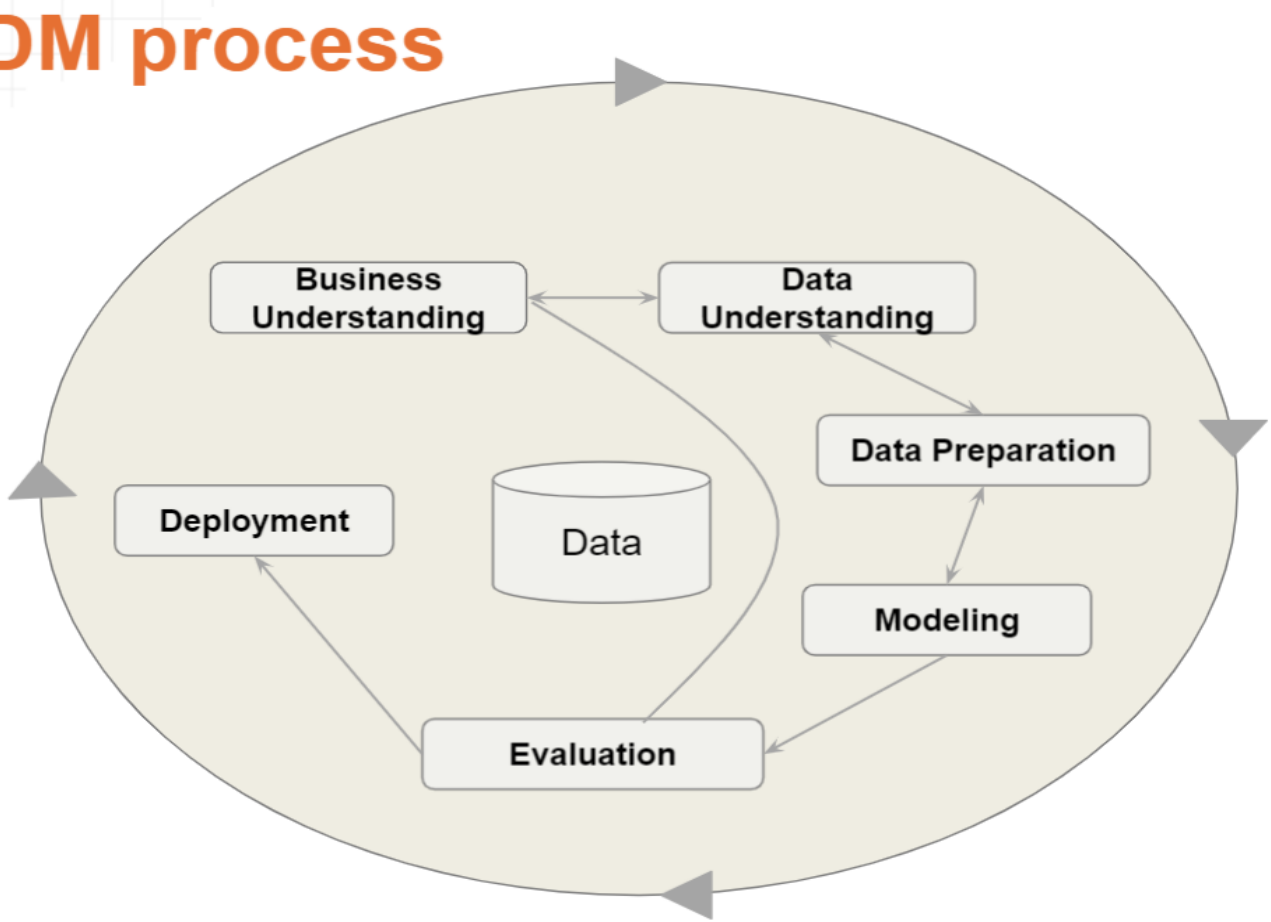


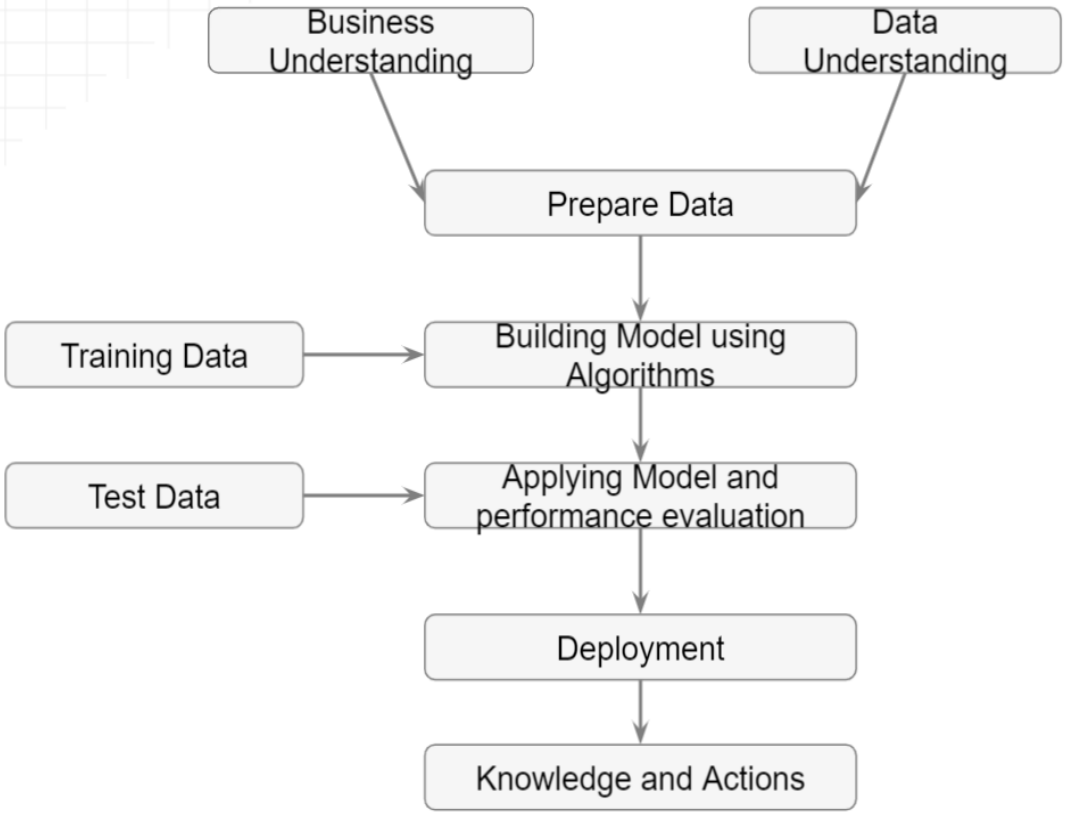
Figure: 3-D view of rainfall data

Data Mining Process



Data Mining Process

Process



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

Gaining information on

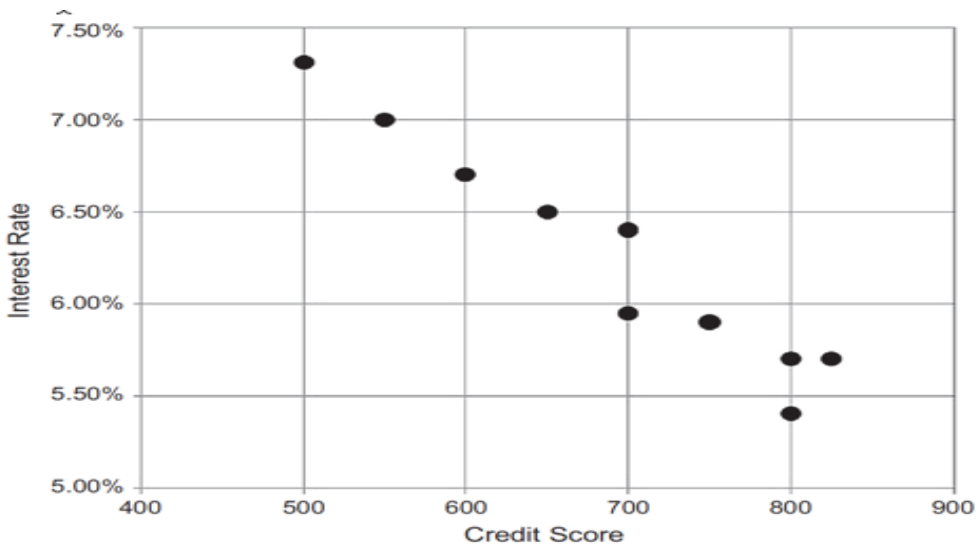
- Objective of the problem.
- Subject area of the problem.
- Data.

Table 2.1 Data Set

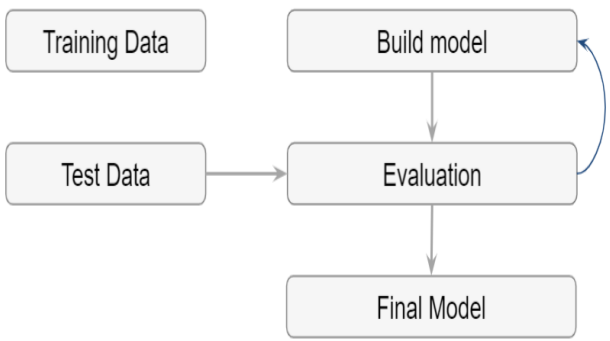
Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03	700	5.95%
04	700	6.40%
05	800	5.40%
06	800	5.70%
07	750	5.90%
08	550	7.00%
09	650	6.50%
10	825	5.70%

Gaining information on

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



3. Modeling



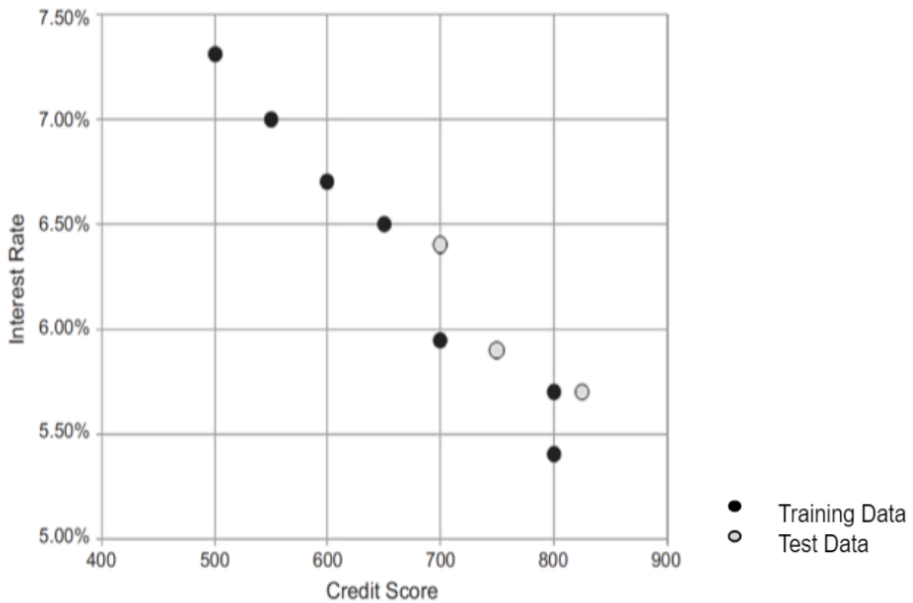
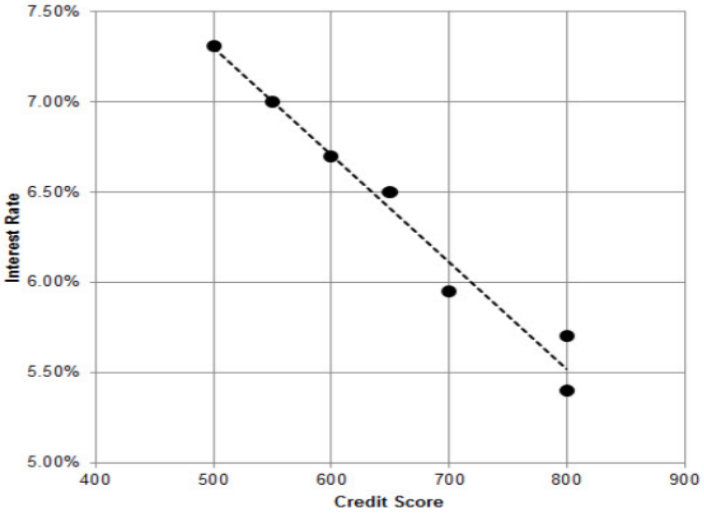


Figure: Splitting data into training and test data sets (right).

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

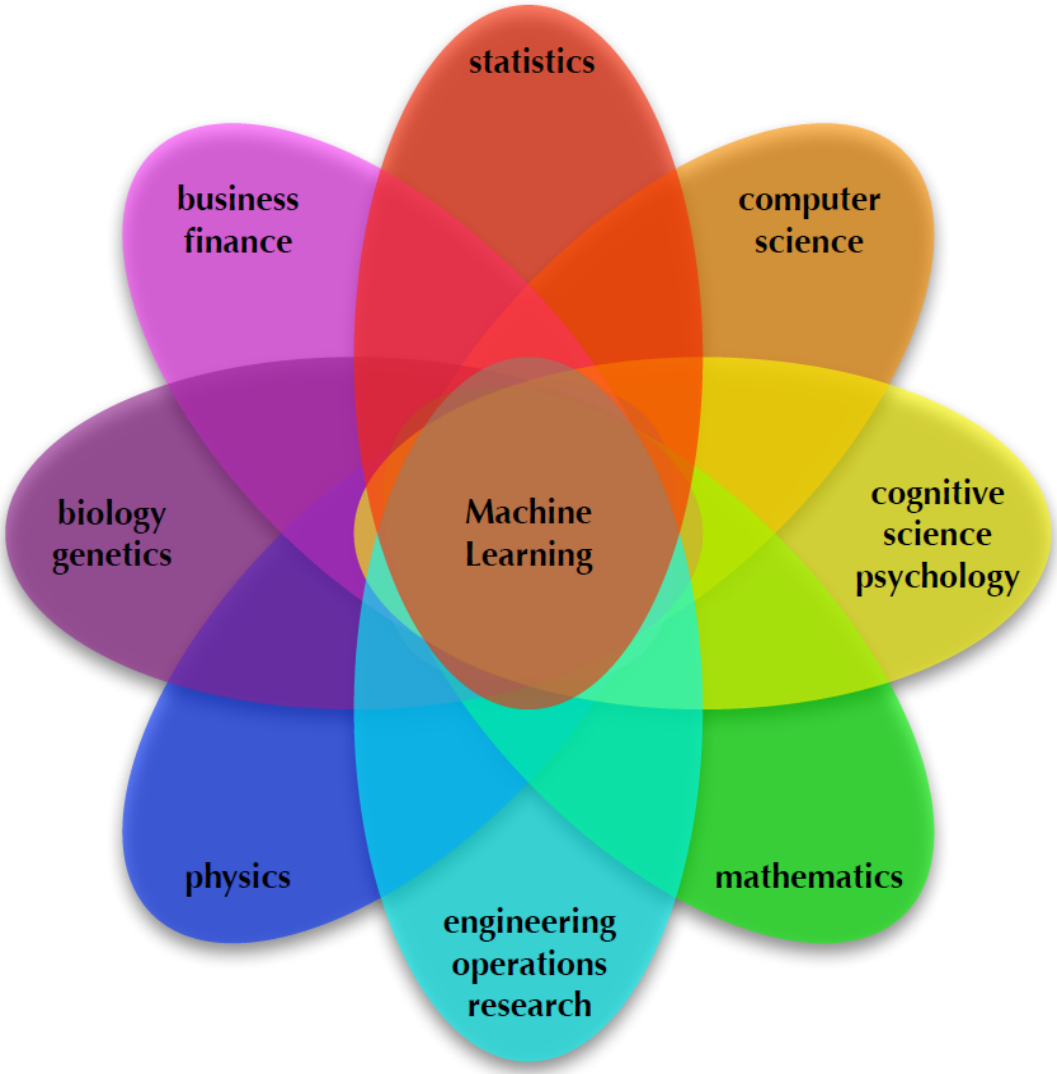
Roadmap:

- 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.

Topic 3 : Machine Learning

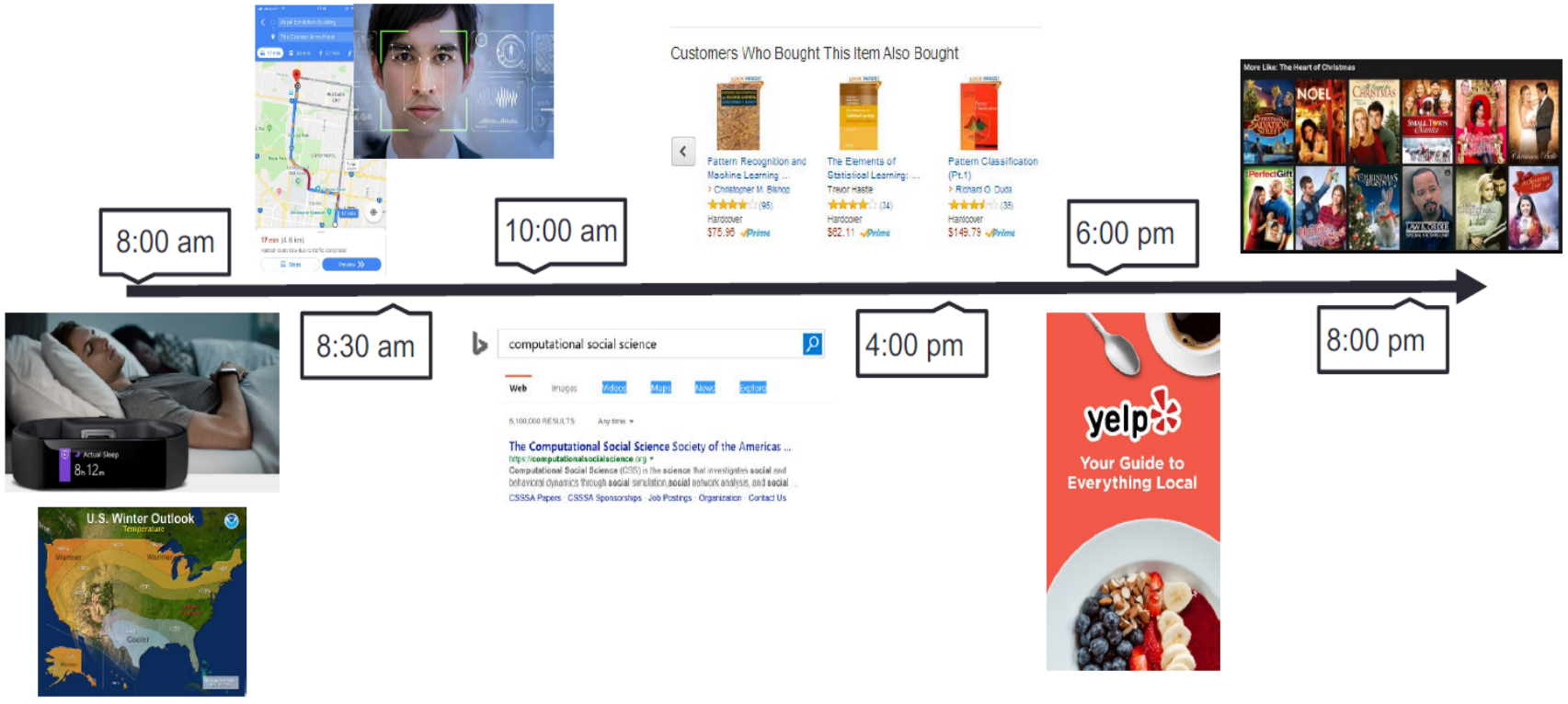
What is Machine Learning?

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



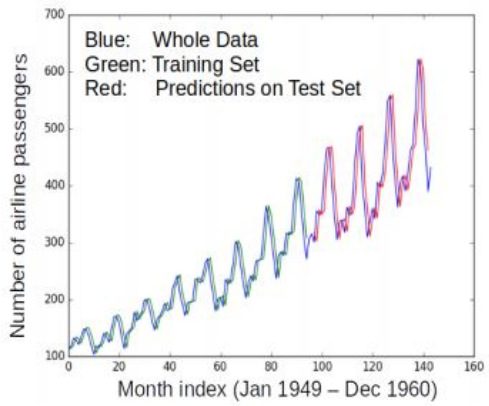
ML techniques are impacting our life

A day in our life with ML techniques...



Introduction to Machine Learning

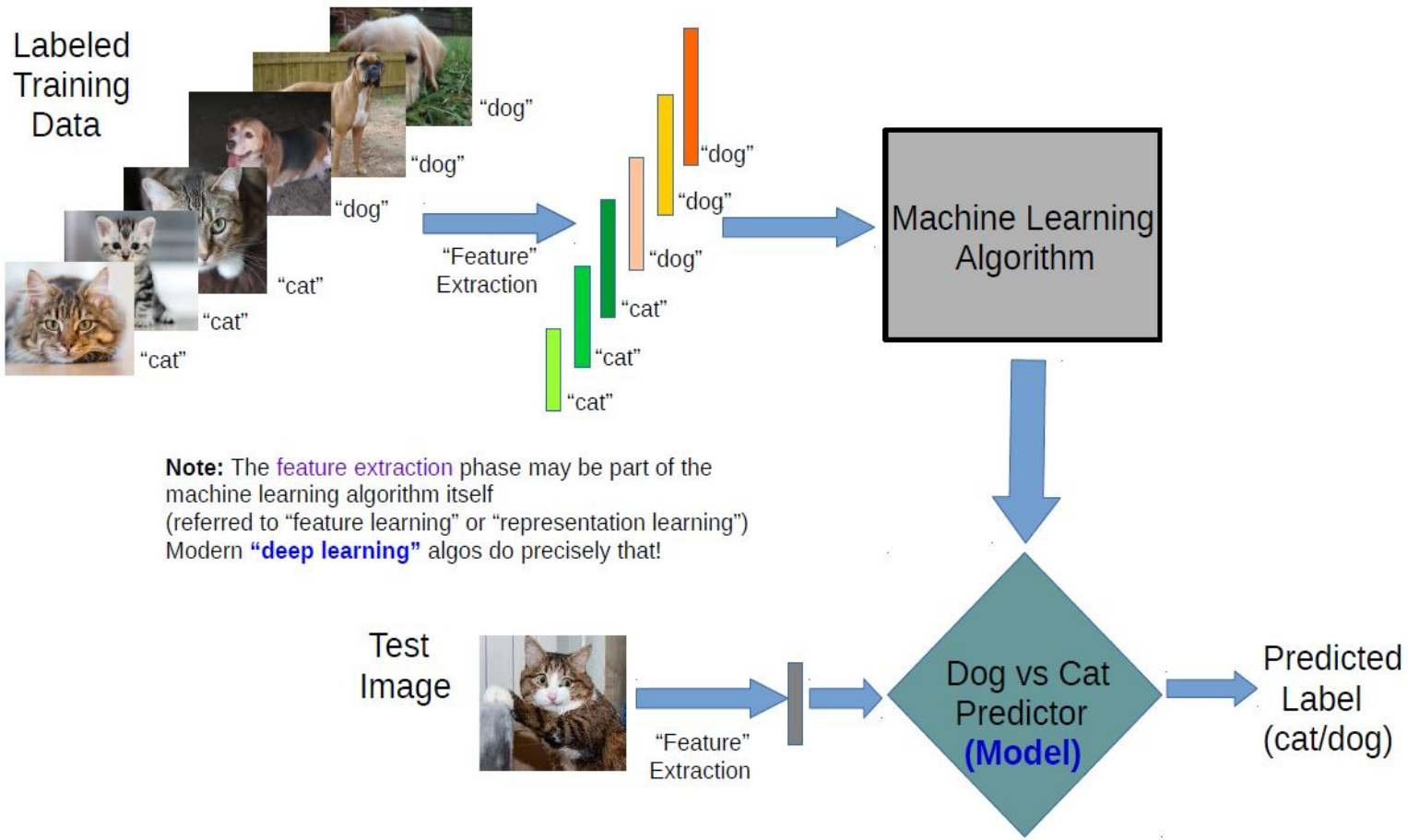
- Designing algorithms that ingest data and learn a model of the data.
- The learned model can be used to
  1. Detect patterns/structures/themes/trends etc. in the data
  2. Make predictions about future data and make decisions



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

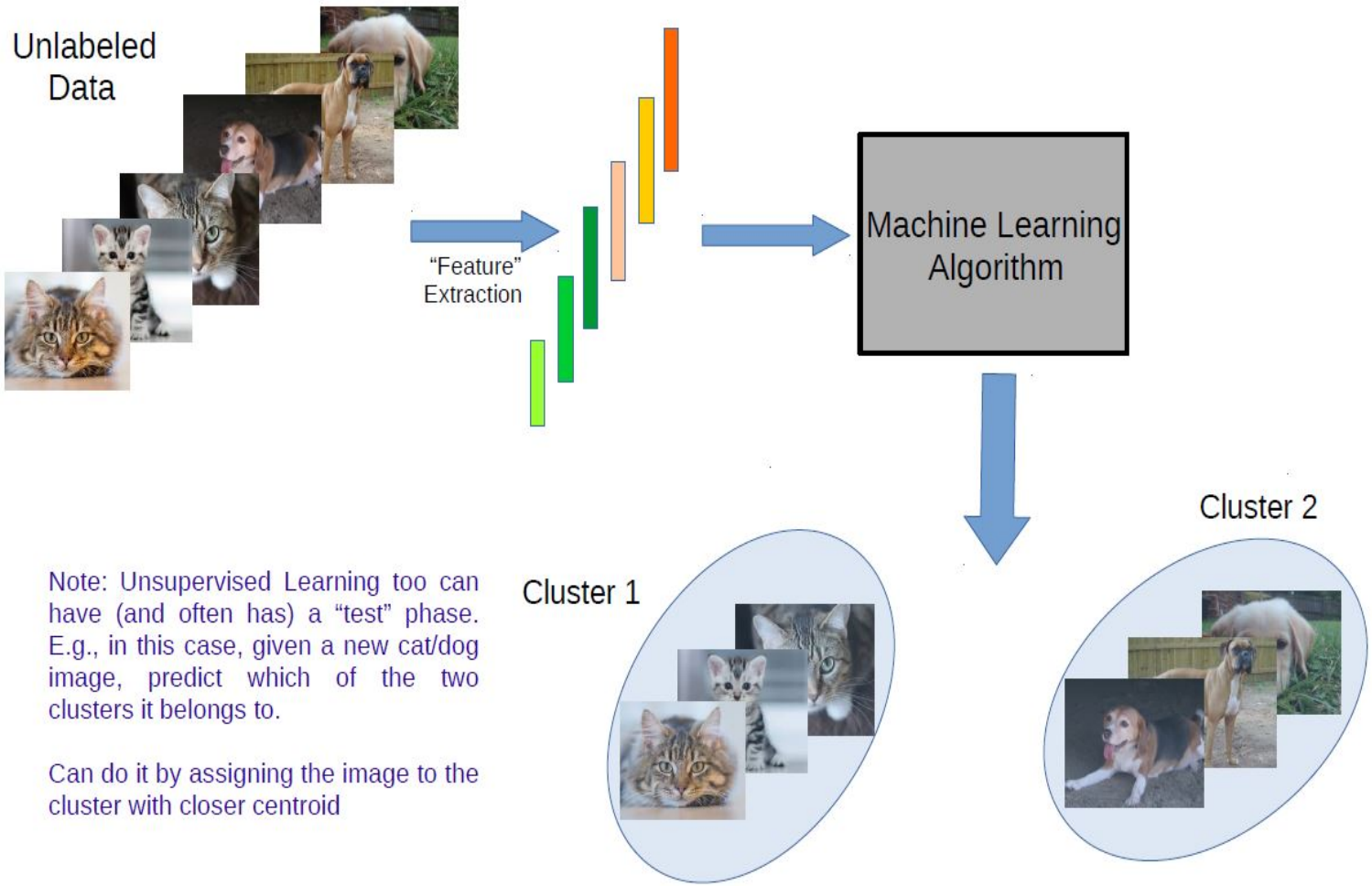
A Typical Supervised Learning Workflow

Supervised Learning: Predicting patterns in the data



A Typical Unsupervised Learning Workflow

Unsupervised Learning: Discovering patterns in the data

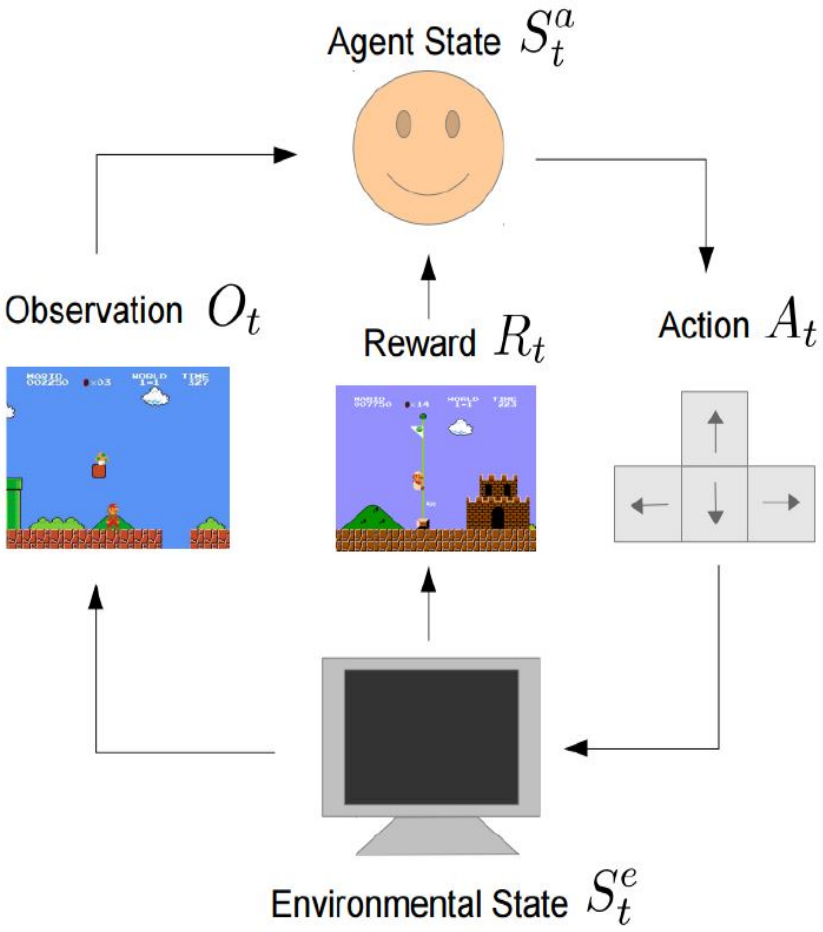


Note: Unsupervised Learning too can have (and often has) a "test" phase. E.g., in this case, given a new cat/dog image, predict which of the two clusters it belongs to.

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)



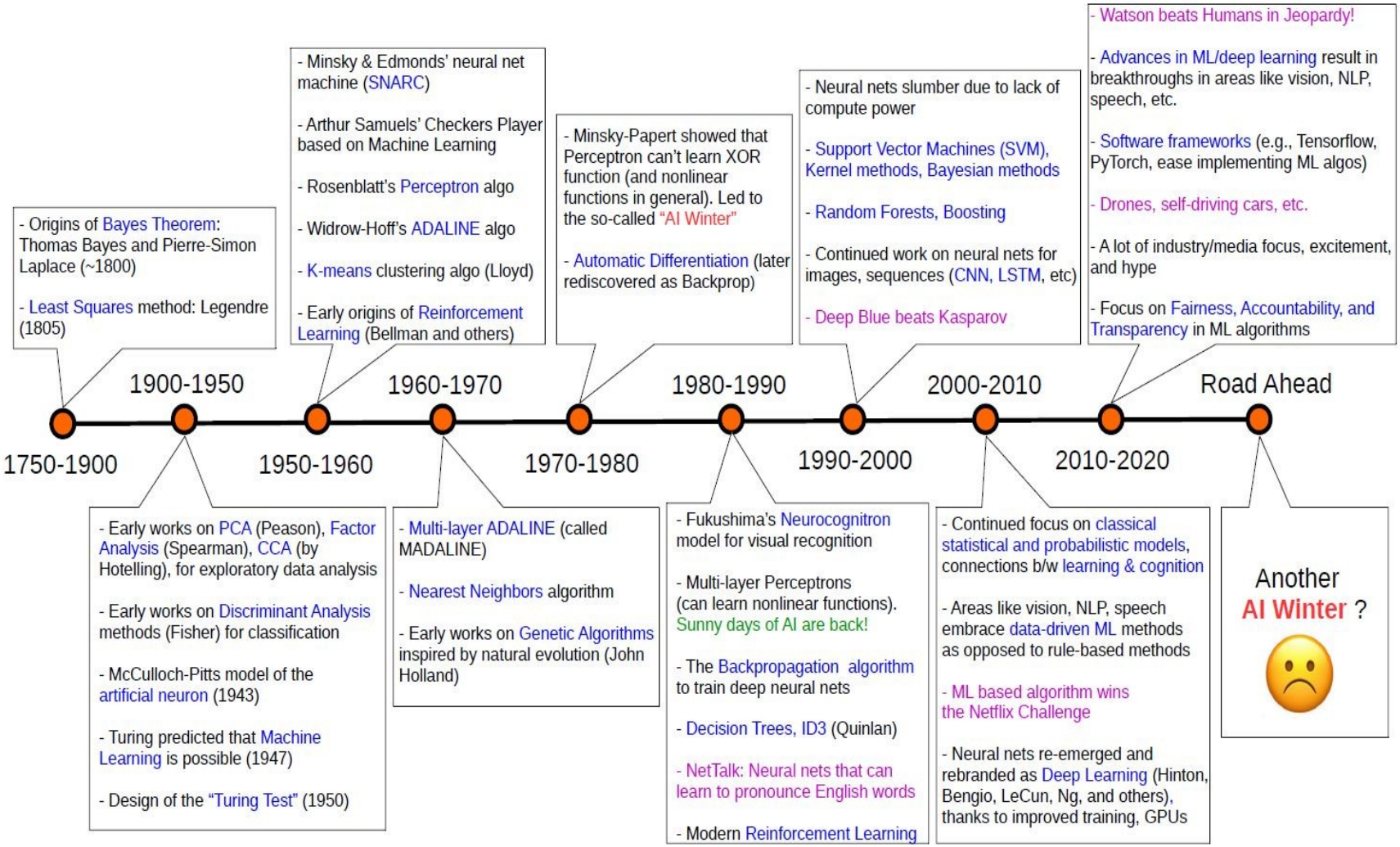
Agent's goal is to learn a policy for some task

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



Developments of Neural Nets

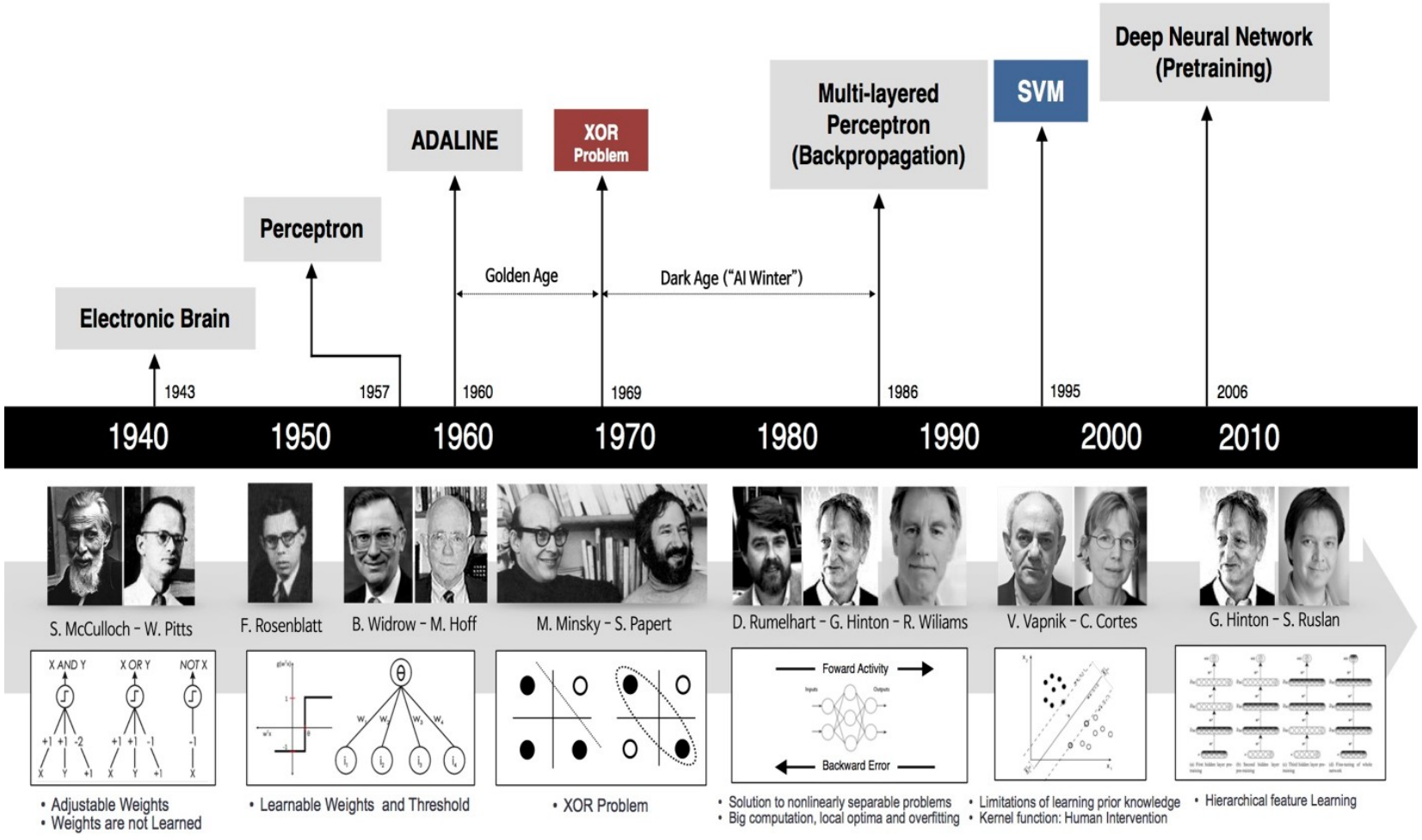
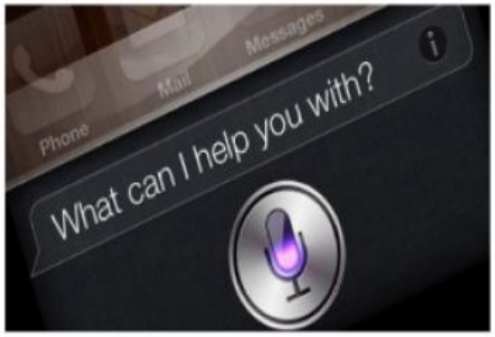


Fig: Developments of Neural Network Models

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.).

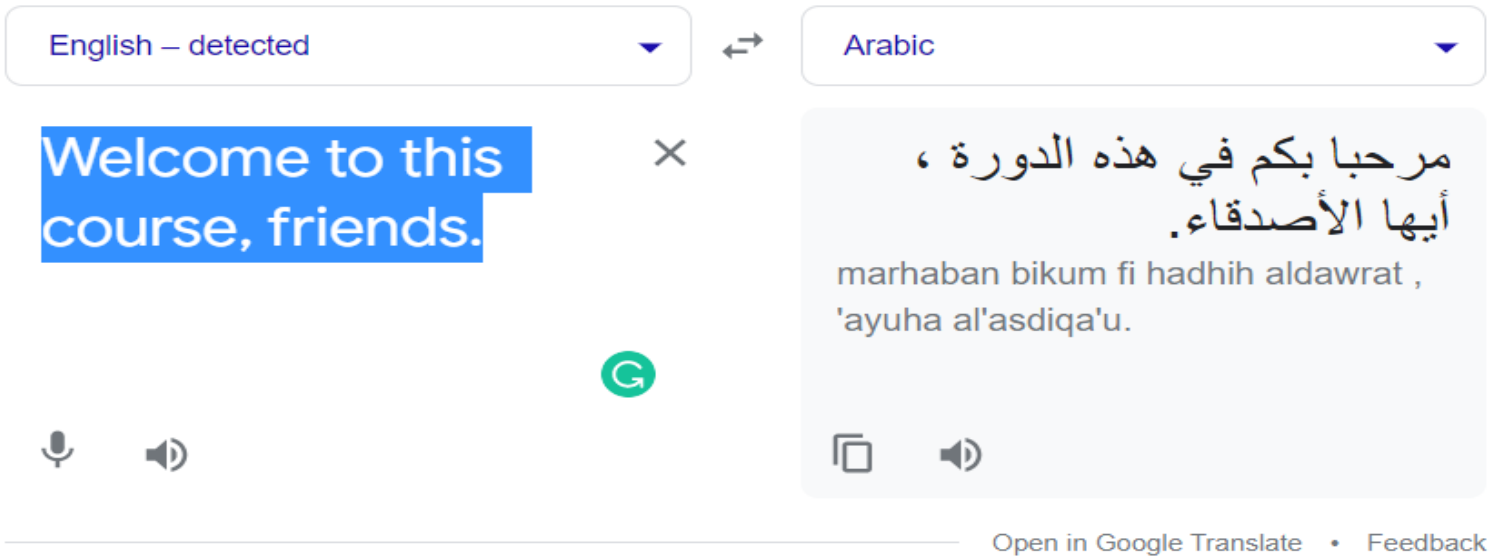



Predictive Policing


Online Fraud Detection

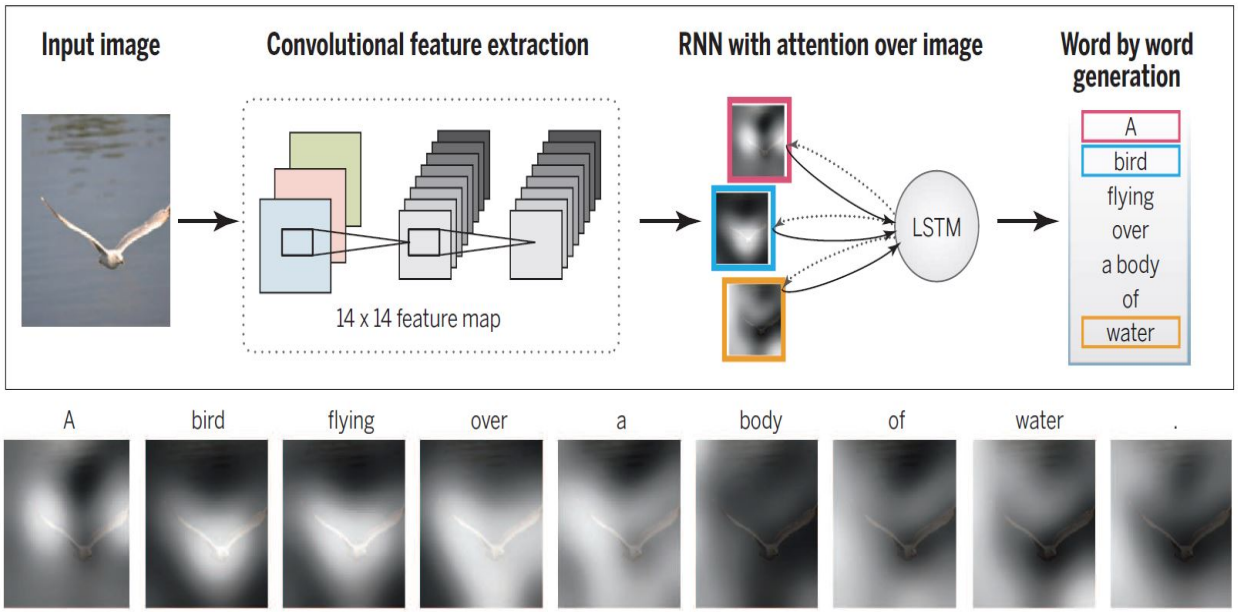
Machine Learning helps NLP

ML algorithm can learn to translate text in the domain of Natural Language Processing



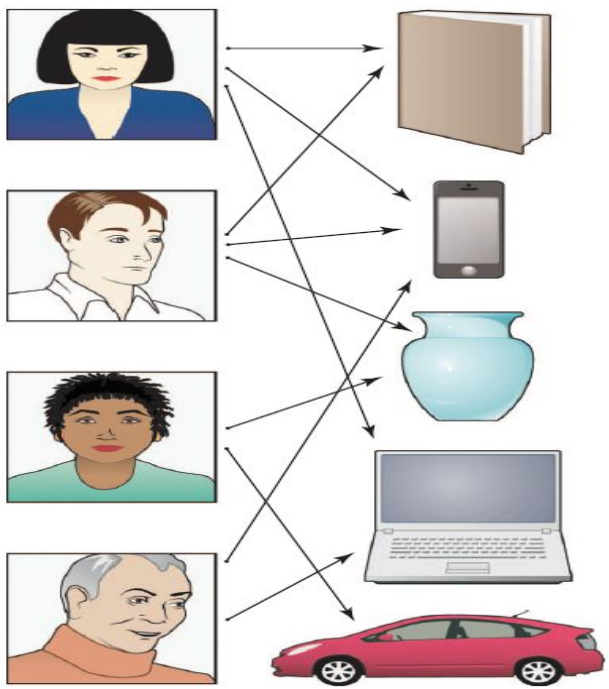
Machine Learning helps Computer Vision

- Automatic generation of text captions for images:
  A convolutional neural network is trained to interpret images, and its output is then used by a recurrent neural network trained to generate a text caption.
- The sequence at the bottom shows the word-by-word focus of the network on different parts of input image while it generates the caption word-by-word.



ML helps Recommendation systems

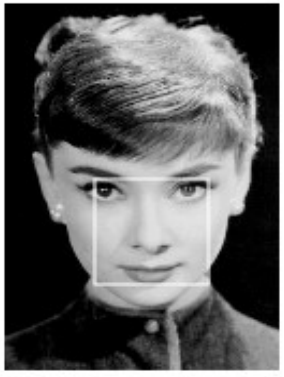
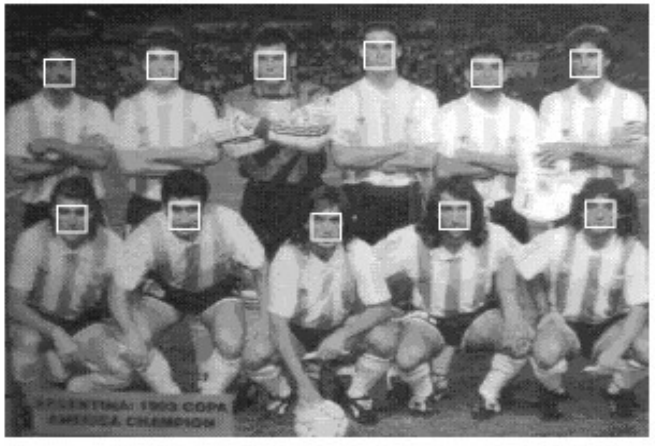
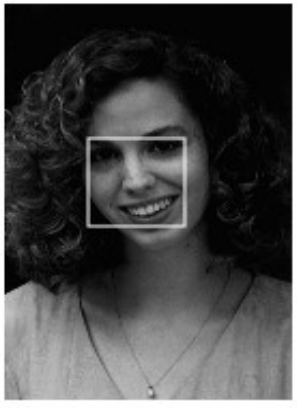
- **A recommendation system** is a machine-learning system that is based on data that indicate links between a set of a users (e.g., people) and a set of items (e.g., products).
- A link between a user and a product means that the user has indicated an interest in the product in some fashion (perhaps by purchasing that item in the past).
- The machine-learning problem is to suggest other items to a given user that he or she may also be interested in, based on the data across all users.



ML helps Image Restoration

ML algorithms can generate high-quality new images using old images.

Machine Learning helps Image Recognition

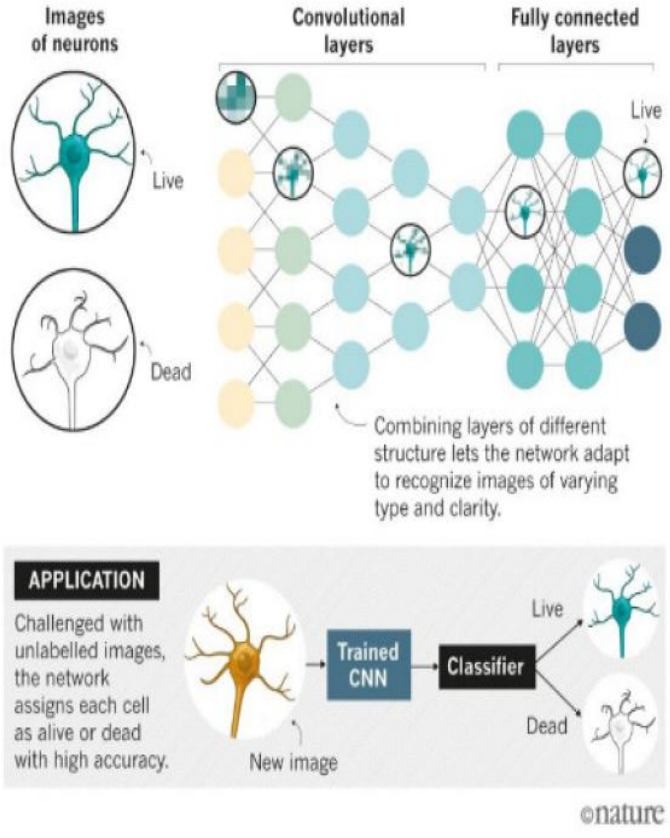









ML helps Many Other Areas...

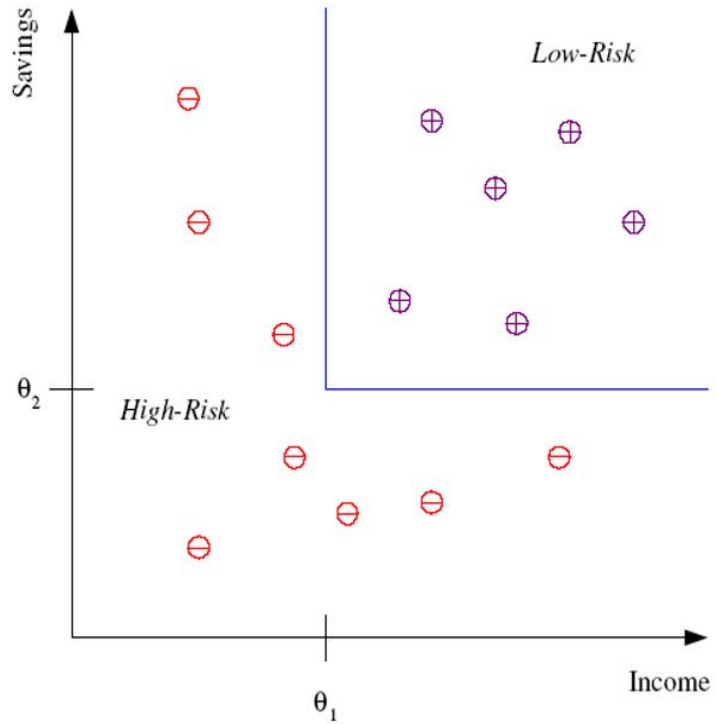
Biology



Finance

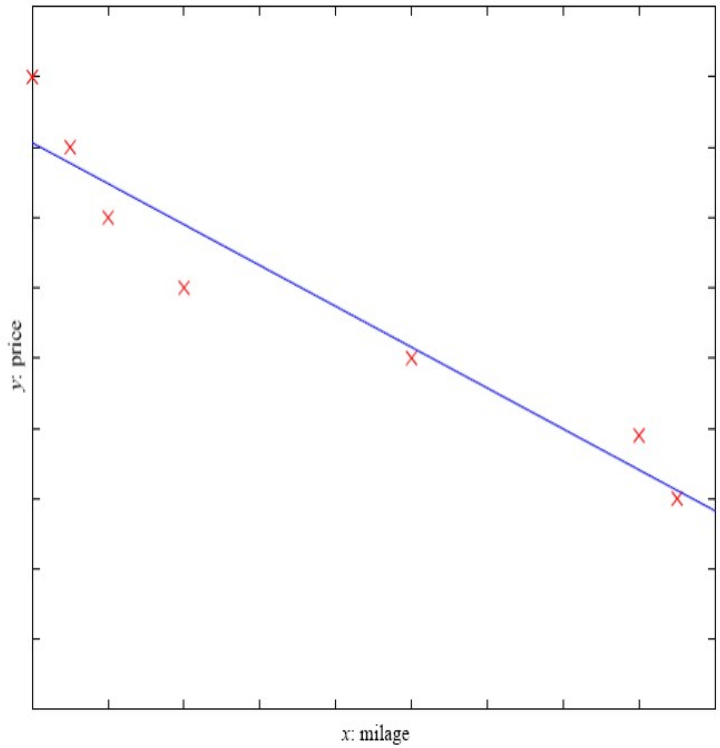
Supervised Learning: Classification

- Example: Credit scoring.
- Differentiating between low-risk and high-risk customers from their income and savings.
- Discriminant: IF Income $> \theta_1$ AND Savings $> \theta_2$ THEN low-risk ELSE high-risk.
- Classification: Learn a linear/nonlinear separator (the “model") using training data consisting of input-output pairs (each output is discrete-valued “label" of the corresponding input).
- Use it to predict the labels for new “test" inputs.
- Other Applications: Image Recognition, Spam Detection, Medical Diagnosis.



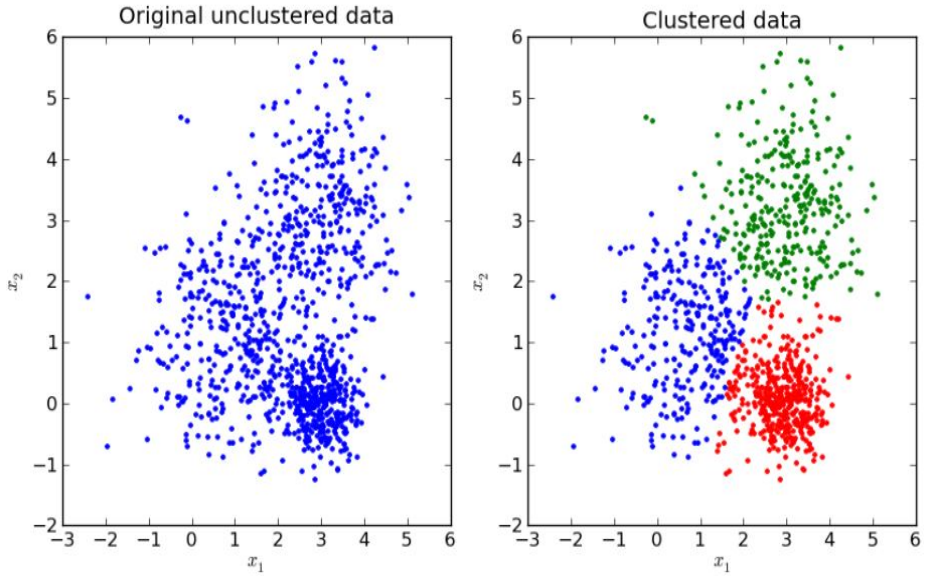
Supervised Learning: Regression

- Example: Price of a used car.
- X : car attributes; Y : price and $Y = f(X, \theta)$
- f( ) is the model and θ is the model parameters.
- Regression: Learn a line/curve (the "model") using training data consisting of Input-output pairs (each output is a real-valued number).
- Use it to predict the outputs for new "test" inputs.
- Other Applications: Price Estimation, Process Improvement, Weather Forecasting.

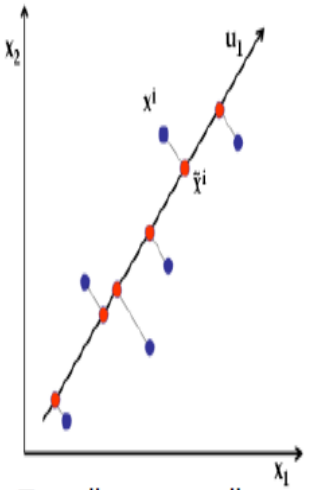


Unsupervised Learning: Clustering

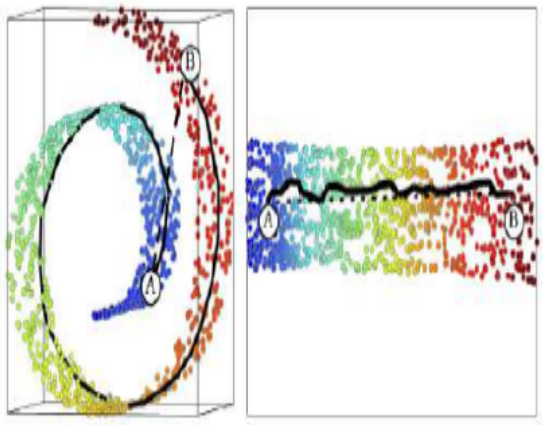
- Given: Training data in form of unlabeled instances $\{x_1, x_2, ..., x_N\}$
- Goal: Learn the intrinsic latent structure that summarizes/explains data
- Clustering: Learn the grouping structure for a given set of unlabeled inputs.
- Homogeneous groups as latent structure: Clustering
- Other Applications: Topic Modelling, Image Segmentation, Social Networking.



- Given: Training data in form of unlabeled instances $\{x_1, x_2, \ldots, x_N\}$
- Dimensionality Reduction: Learn a Low-dimensional representation for a given set of high-dimensional inputs
- Note: DR also comes in supervised flavors (supervised DR)
- Other Applications: facial recognition, computer vision and image compression.



Two-dim to one-dim linear projection



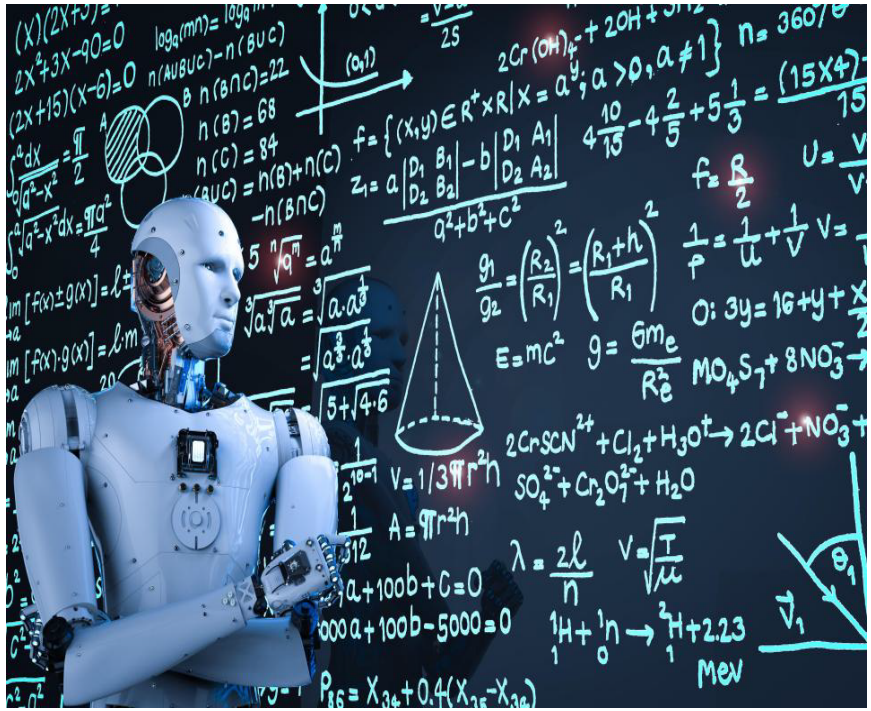
Three-dim to two-dim nonlinear projection (a.k.a. manifold learning)

Topic 4 : Artificial intelligence

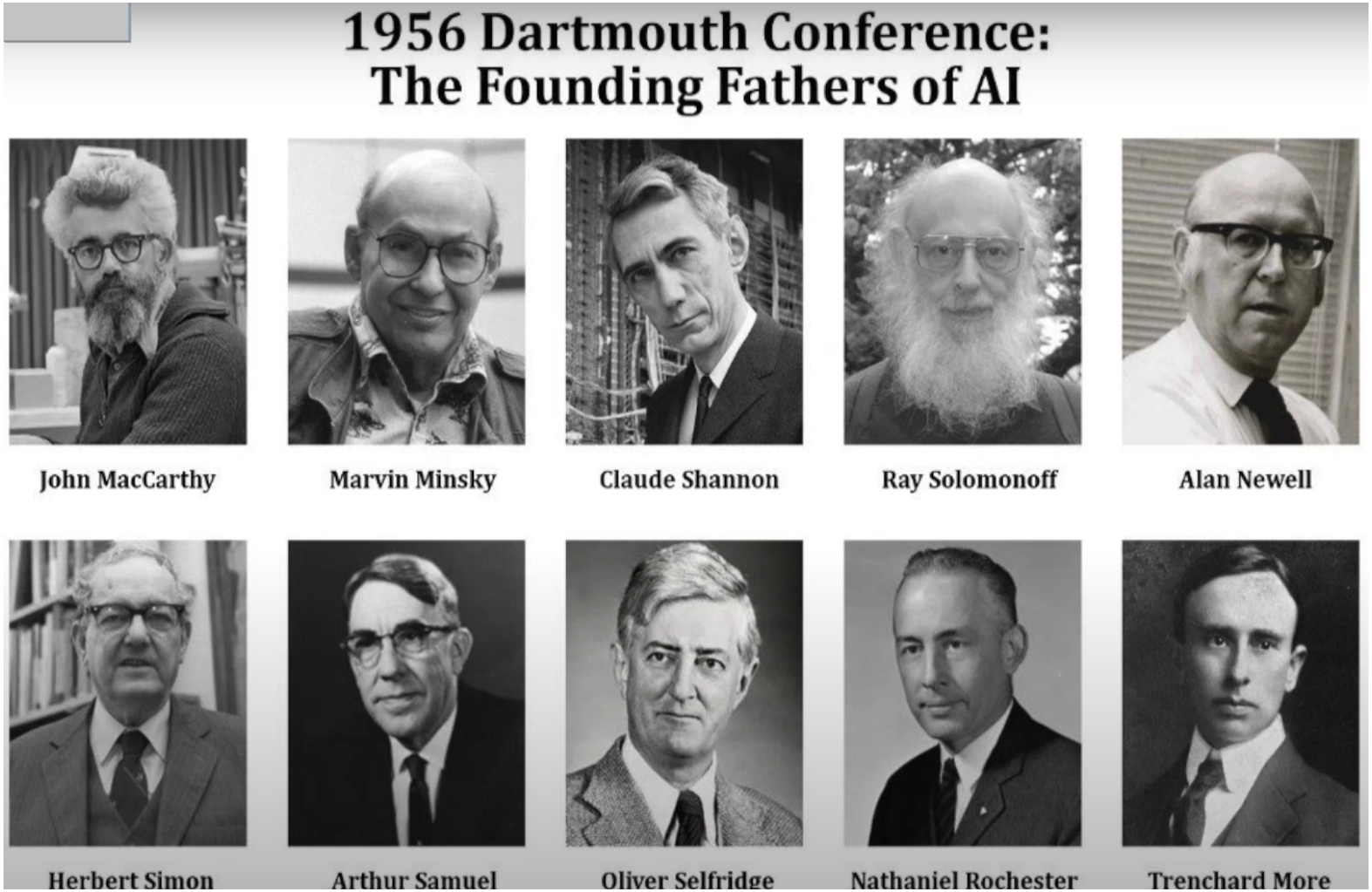
A first look at Artificial Intelligence

- 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".



Genesis of AI



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

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).

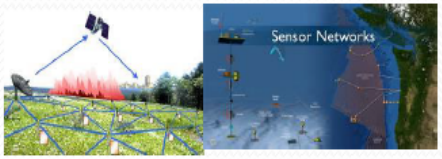
Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

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)



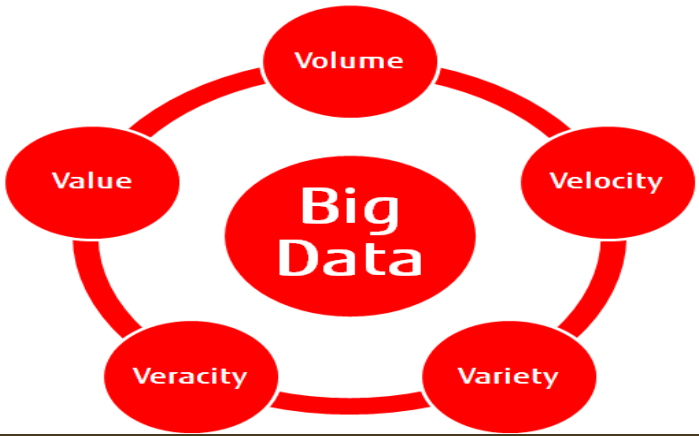
Sensor technology and networks
(Measuring all kinds of data)

Now data is Big data!

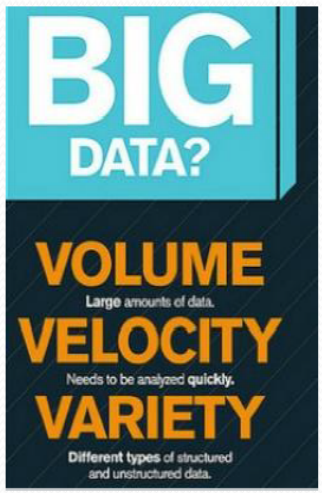
"Big data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it" - Standard definition.

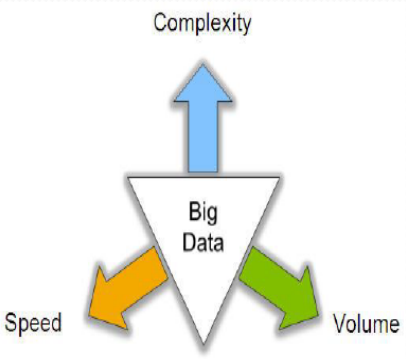
Difficulties related to (Big) data:

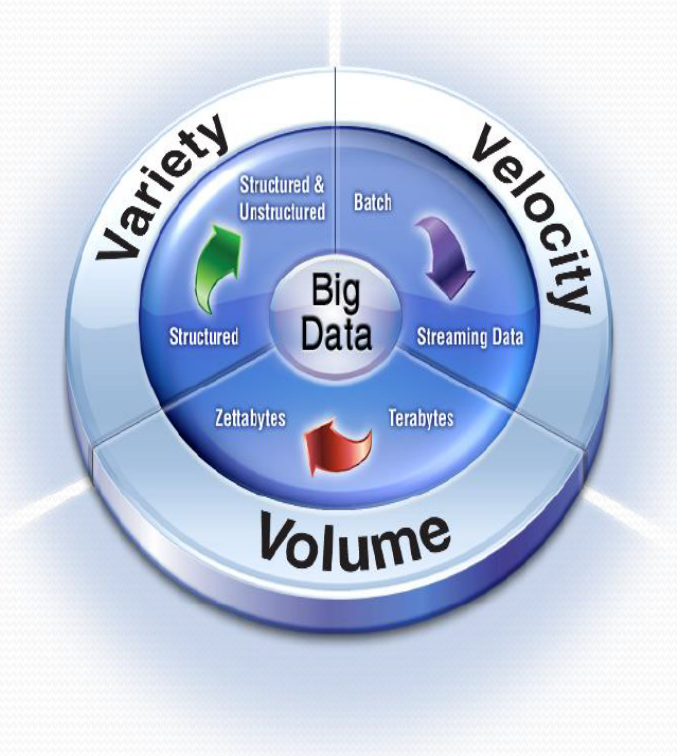
- The prediction must be accurate: difficult for some tasks like image classification, video captioning...
- The prediction must be quick: online recommendation should not take minutes.
- Data must be stored and accessible easily.
- It may be difficult to access all data at the same time. Data may come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



Characteristics of Big data: 3Vs



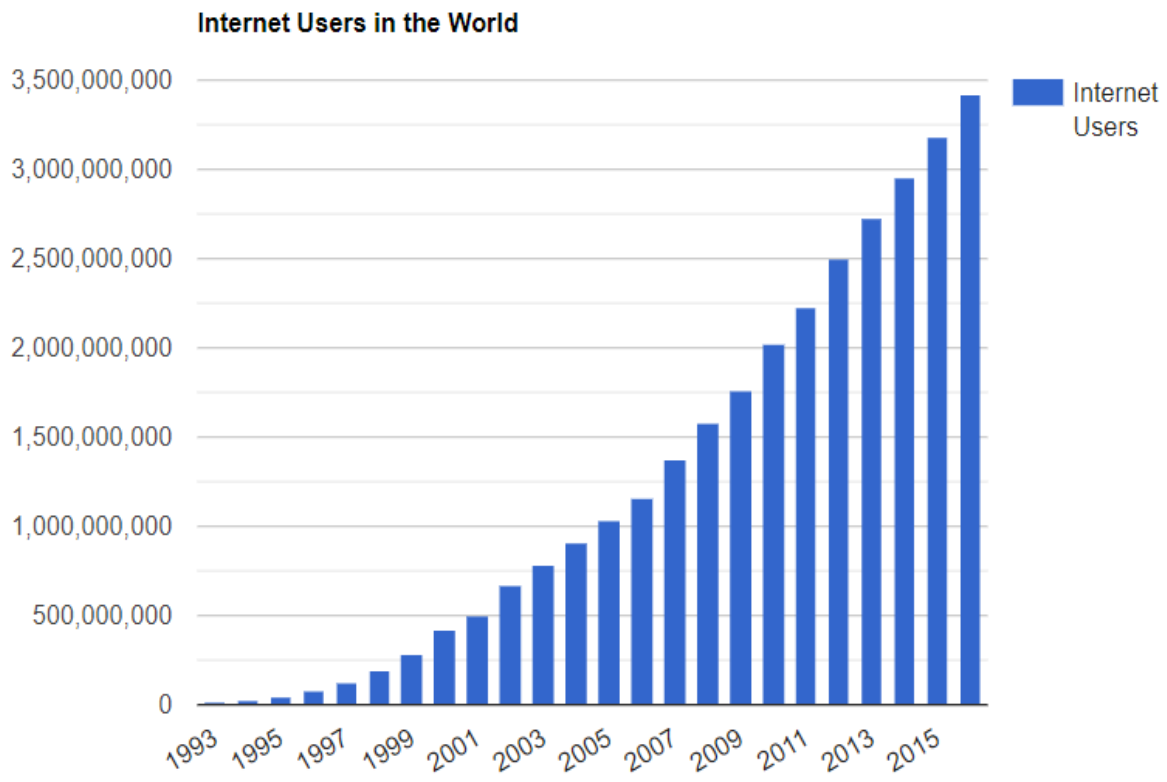




Big data: V for Volume

Volume:

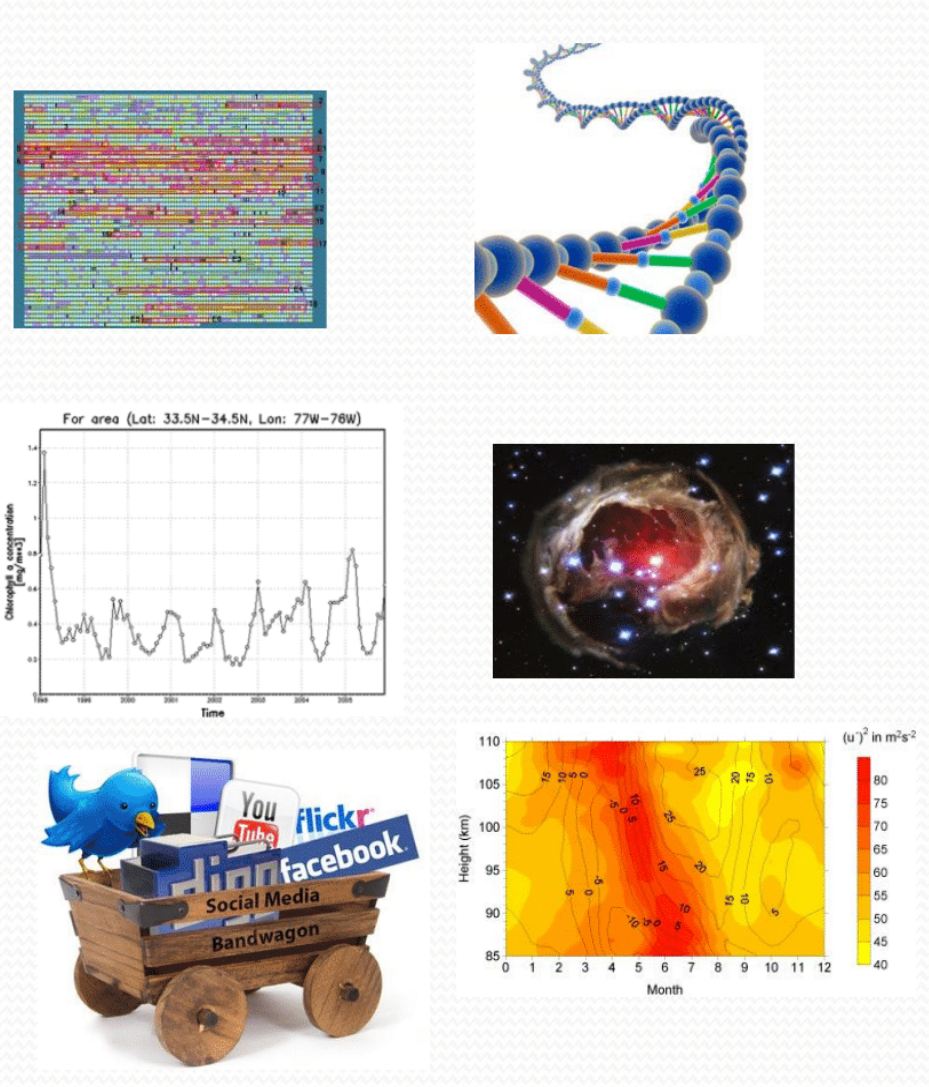
- Volume of data that needs to be processed is increasing rapidly.
- Need more storage capacity.
- Need more computation facility.
- Need more tools and techniques.



Big data: V for Variety

Variety:

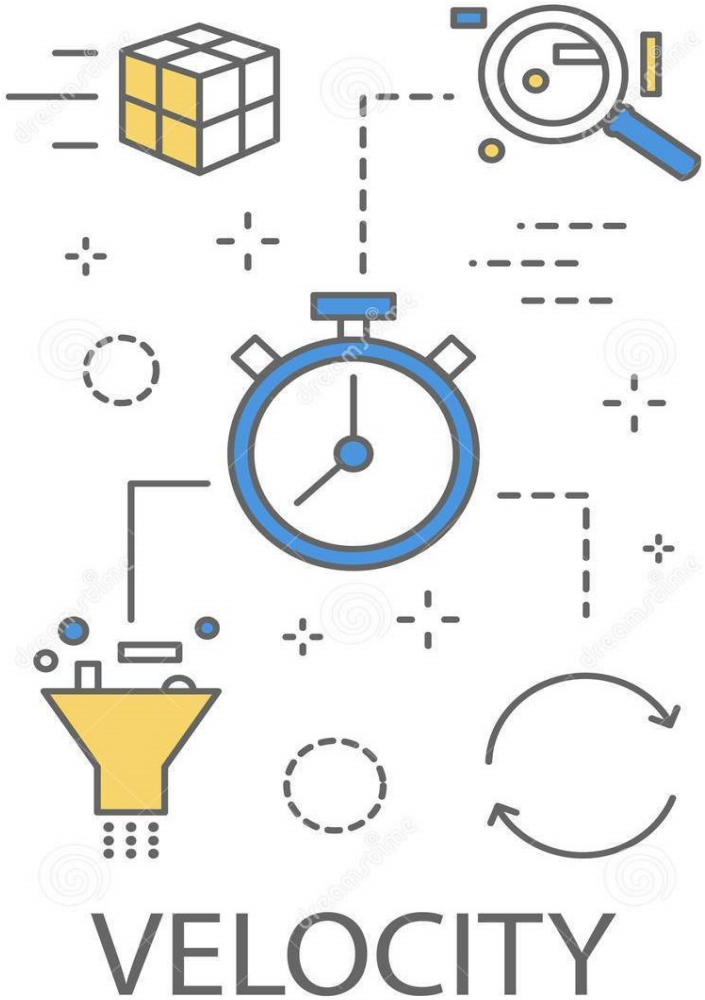
- Various formats, types, and structures.
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dimensional arrays, etc.
- A single application can be generating or collecting many types of data.
- To extract knowledge, all these types of data need to be linked together.



Big data: V for Velocity

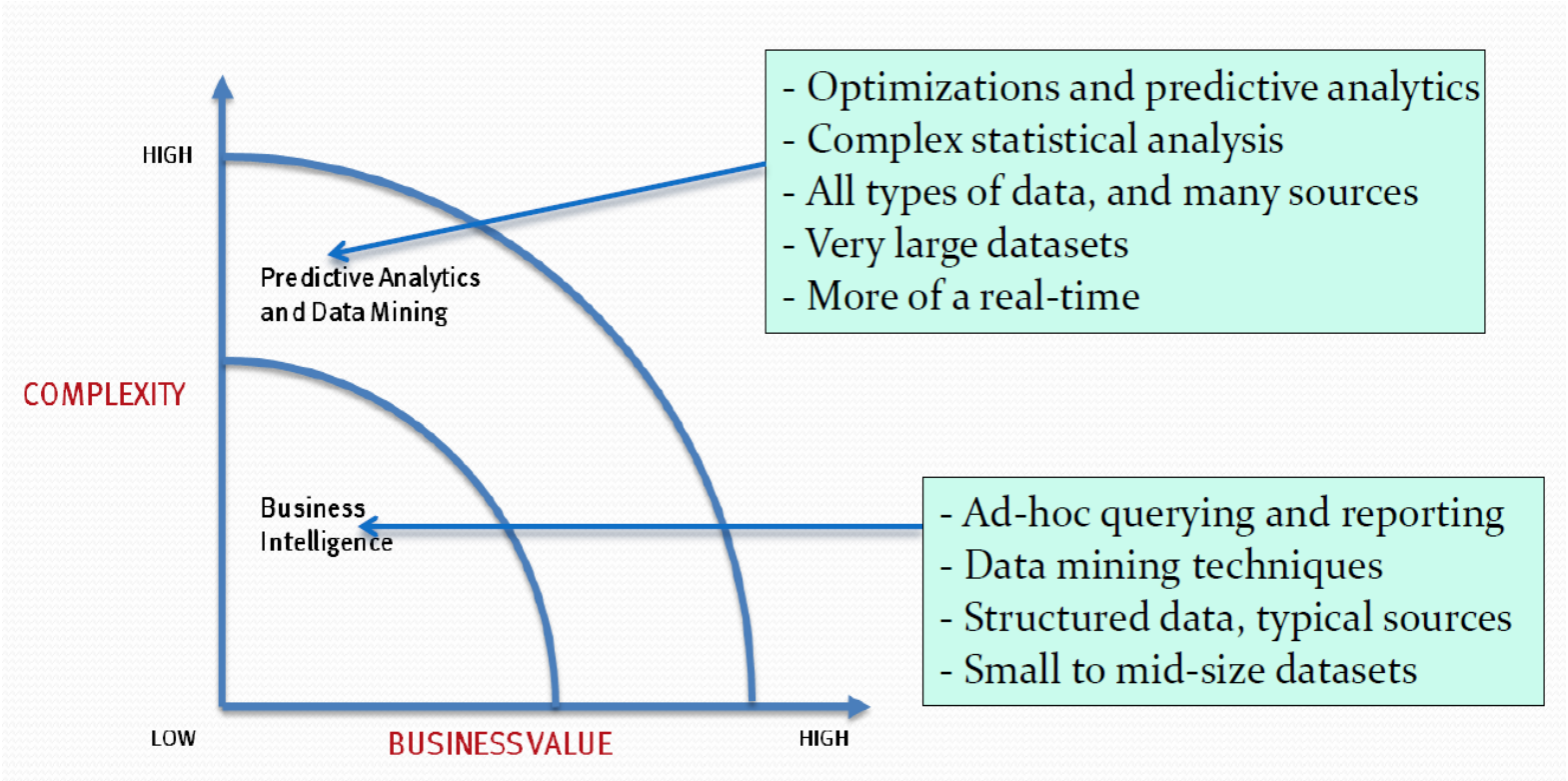
Velocity:

- 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.
- Scrutinize 5 million trade events created each day to identify potential fraud.



Big data vs. Small data

Big data is more real time in nature than traditional applications (Massively parallel processing, scale out architectures are well suited for big data 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.

Still 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.

Now we are stepping into risk-sensitive

Shifting from Performance Driven to Risk Sensitive (Explainable AI)...

Human

Questions of the day

What type of data are involved in the following applications?

- Weather forecasting
- Mobile usage of all customers of a service provider
- Anomaly (e.g. fraud) detection in a bank organization
- Person categorization, that is, identifying a human
- Air traffic control in an airport
- Streaming data from all flying air crafts of Boeing

End of Session

Caution: *"Prediction is very difficult, especially if it's about the future"*
- Niels Bohr, Father of Quantum.