

Sorbonne University Abu Dhabi

Day 1: Multivariate Data Analysis*

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Dr. Tanujit Chakraborty, Ph.D. from ISI Kolkata.
Assistant Professor in Statistics at Sorbonne University.
tanujit.chakraborty@sorbonne.ae
<https://www.ctanujit.org/MDA.html>
Course for BSc Mathematics and Data Science Students.

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



Education:

- **2010 - 2013** : B.Sc. in Statistics from **Bidhannagar College**, Kolkata, India.
- **2014 - 2016** : Master of Science from **Indian Statistical Institute**, Kolkata, India. (Specialization: Statistical Quality Control & Operations Research)
- **2016 - 2020** : Ph.D. in Quality, Reliability & Operations Research (QROR) from **Indian Statistical Institute (ISI)**, Kolkata, India.
Title of the Thesis: "Some Nonparametric Hybrid Predictive Models: Asymptotic Properties and Applications".

Awards:

- B.G. Raghavendra Memorial Award from **Operational Research Society of India (ORSI)** in December 2017.
- Best Student Paper Award (Application Category) at the International Conference held at **IIM Ahmedabad** in December 2019.
- Best Paper Award Winner at ACM International Conference on Data Sciences and Management of Data (**ACM CODS-COMAD**) in January 2021.

Motivations: *Primary motivation* comes from the real-world data sets, with a variety of data types, such as business, process efficiency improvement, water quality control, and software defect prediction, among many others. My research works emphasize on the *development of statistical methodologies* that are scalable, robust, accurate, statistically sound, and easily interpretable.

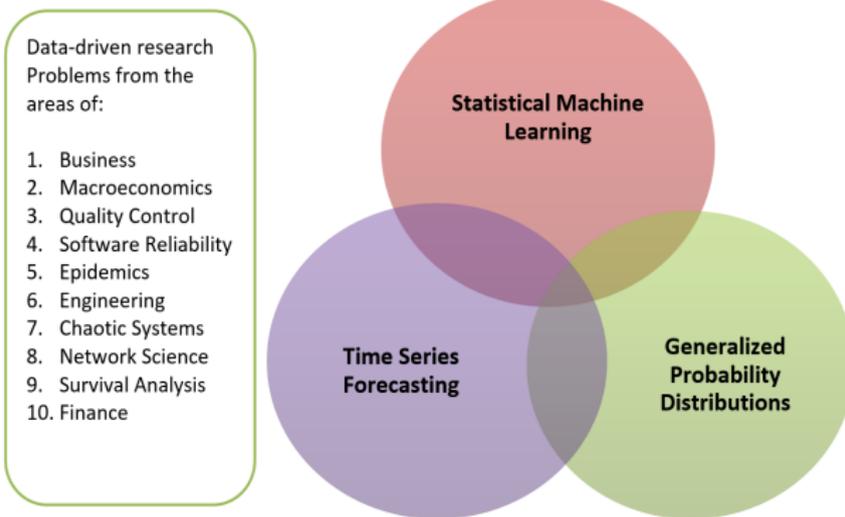


Figure: To know more visit my [Google Scholar Profile](#)



Basic Statistics:

- Descriptive Statistics
- Estimation and Hypothesis Testing
- Analysis of Variance (ANOVA)

Programming Skills:

- Basics of R & RStudio
- Basics of Python
- Basic statistics using R
- Basic statistics using Python

Probability Theory:

- Introductory Probability Theory
- Probability Distributions
- Sampling Distributions
- Bivariate and Multivariate Probability Distributions

Regression Analysis:

- Linear Regression
- Nonlinear Regression
- Shrinkage Methods

Multivariate Analysis:

- SVD & PCA
- Factor Analysis
- Correspondence Analysis
- Discriminant Analysis

Coding & Projects:

- Implementations in R & Python
- Homeworks on Theory & Coding (15%)
- Course Projects (15%)
- Mid-term (20% and Final Exam (50%))

“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.



2018 *This Is What Happens In An Internet Minute*





Astronomy



Social Networks



Healthcare



Banking



Genomics



Weather
measurements





- 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 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:**

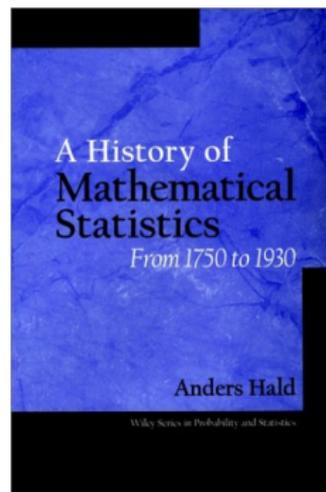
- ① Making inference from samples
- ② Development of new methods for complex data sets
- ③ Quantification of uncertainty and variability

- **Two Views of Statistics:**

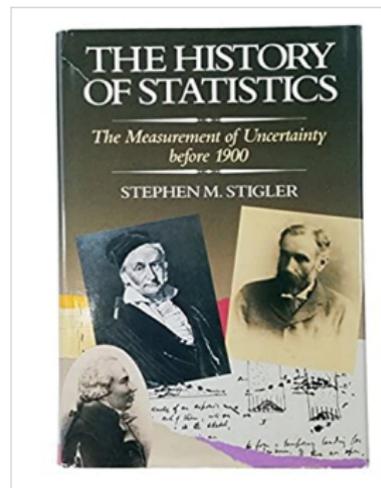
- ① Statistics as a Mathematical Science
- ② Statistics as a Data Science

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:-

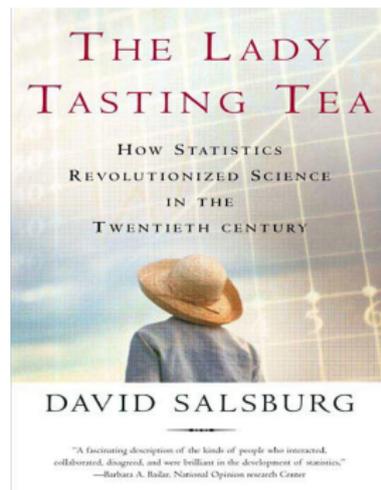
The science of Statistics is essentially a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data. As in other mathematical studies, the same formula is equally relevant to widely different groups of subject-matter. Consequently the unity of the different applications 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.



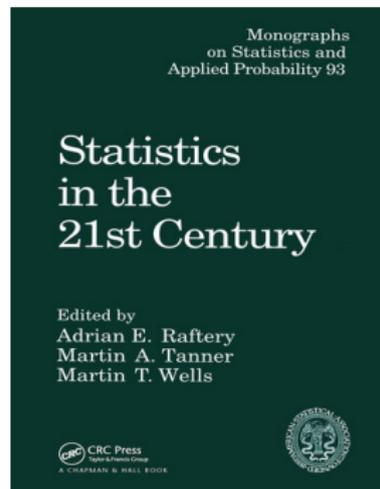
- Probability and its application to Gambling & Astronomy
 - Jacob Bernoulli (Prob. distribution, Law of large numbers, etc.)
 - Pierre-Simon Laplace (Double exponential, Transformation, etc.)
 - Thomas Bayes (Bayes' theorem, etc.)
 - Gauss and Legendre (Least Square Method)
 - Francis Galton (Correlation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



- 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.



- 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).



- 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.).



- **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, construction of fast training algorithms and computation time.
- **Applications** : Strong interactions with substantive fields in all areas. Applications of statistical methods in almost all the fields are evident. Statistics became a key technology driven by data ("**Data is the new oil**").



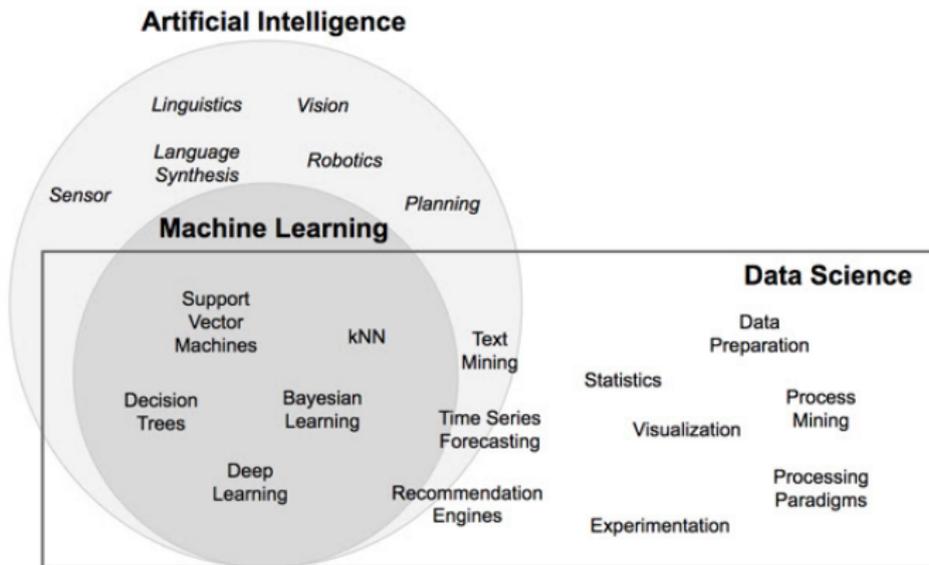
- **Traditional Problems in Applied Statistics:**
 - Well formulated question that we would like to answer.
 - Expensive to gather data and/or expensive to do computation.
 - Create specially designed experiments to collect high quality data.
- **Current Situation :** Information Revolution
 - Improvements in computers and data storage devices.
 - Powerful data capturing devices.
 - Lots of data with potentially valuable information available.

- 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

“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?

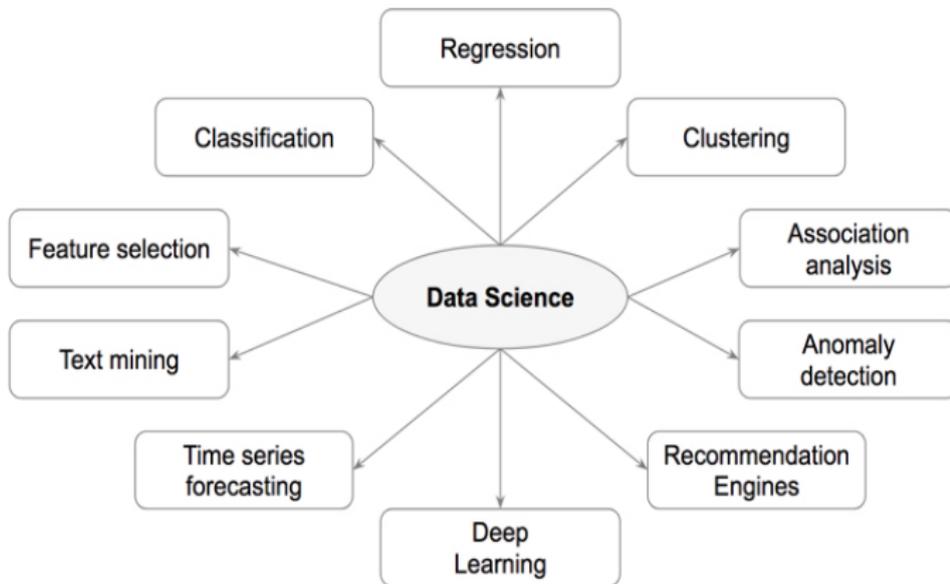
*"When you're fundraising, it's **AI**.*

*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.



Basic Definitions:

Entity: A particular thing is called entity or object.

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

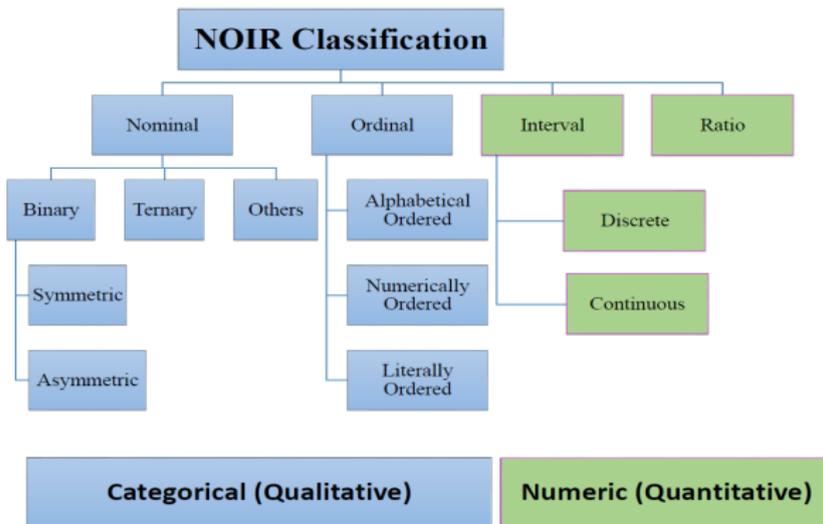
Data: A measurement of an attribute is called data.

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.

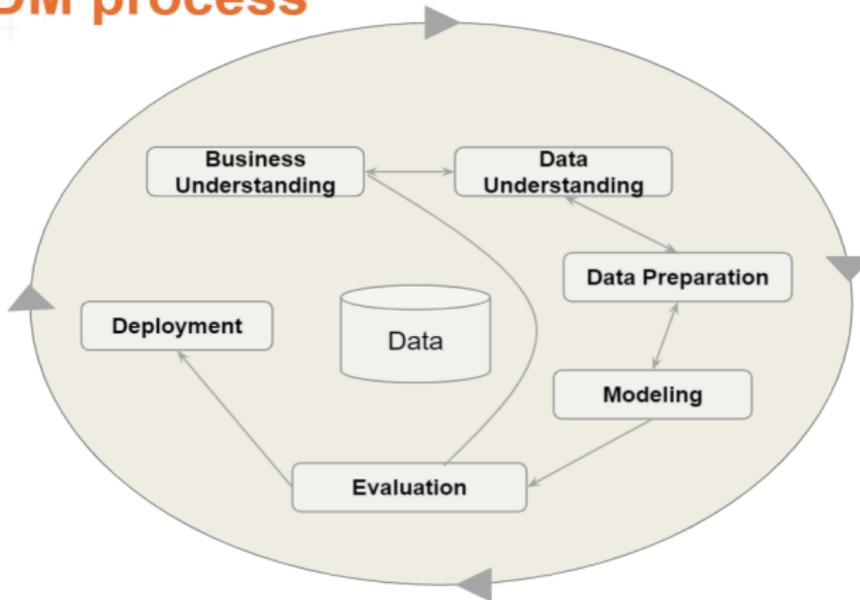
Scale: A good understanding of data scales (also called scales of measurement) is important. Depending on the scales of measurement, different techniques are followed to derive hitherto unknown knowledge in the form of patterns, associations, anomalies or similarities from a volume of data.



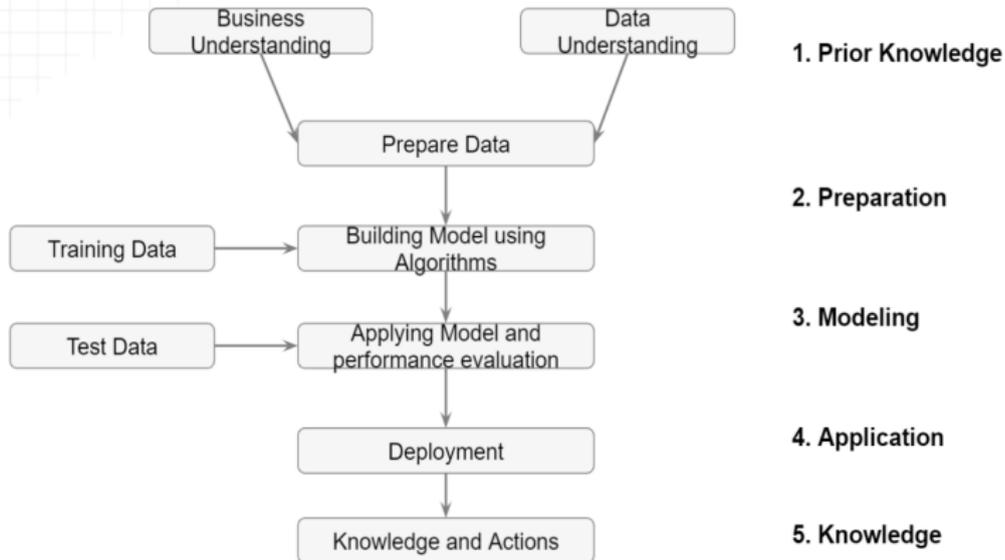
- The **NOIR scale** is the fundamental building block on which the extended data types are built.
- Further, nominal (Blood groups, Attendance) and ordinal (Shirt size) are collectively referred to as **categorical or qualitative data**. Whereas, interval (weight, temperature) and ratio (Sound intensity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



DM process



Process



Gaining information on

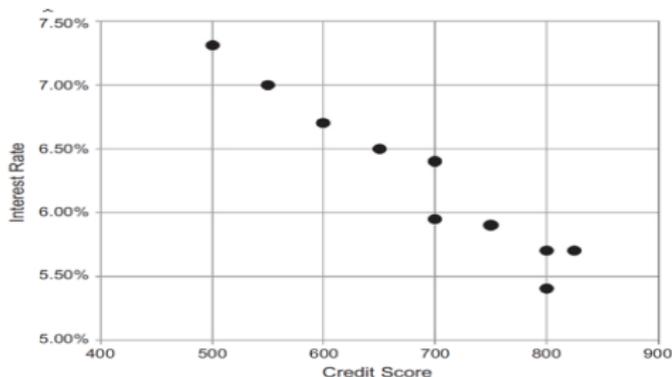
- Objective of the problem.
- Subject area of the problem.
- Data.

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.



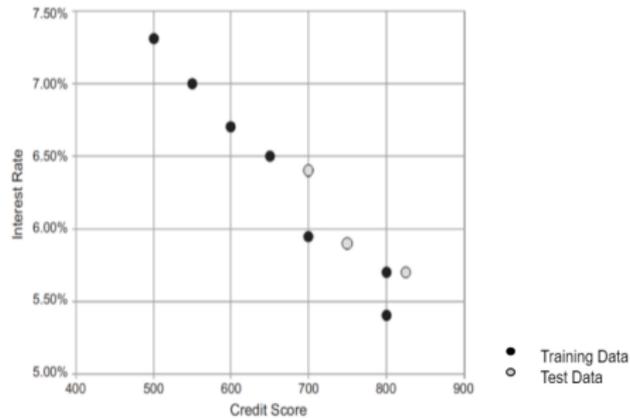
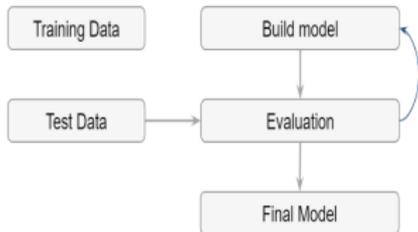
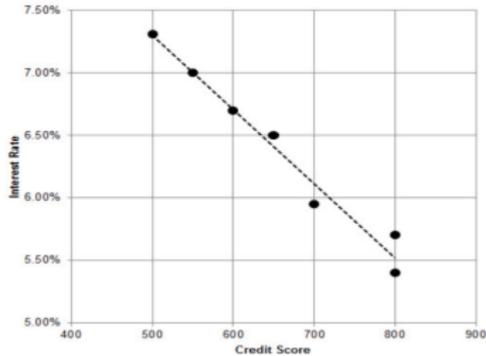


Figure: Splitting data into training and test data sets (right).



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

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).

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.

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.



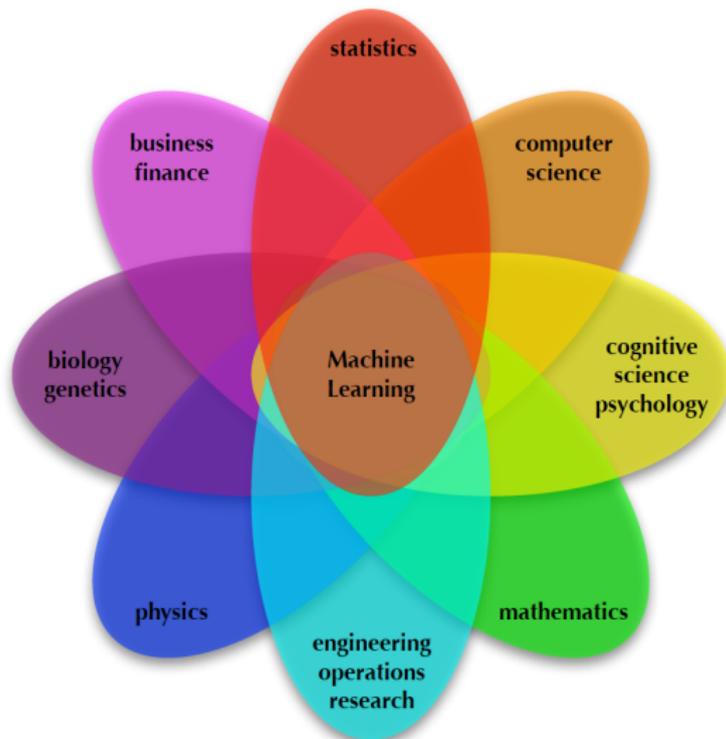
Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by 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 retailer 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.



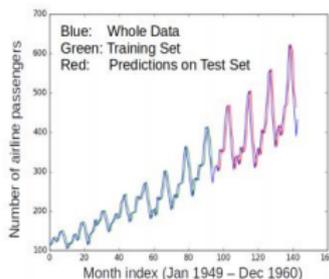


A day in our life with ML techniques...

The diagram illustrates the application of machine learning (ML) techniques throughout a day. A central horizontal arrow points from left to right, with time markers in speech bubbles above it. Various images and screenshots are placed along this timeline to show ML in action.

- 8:00 am:** A navigation app (Google Maps) shows a route. A face recognition interface is overlaid on a photo of a man's face.
- 8:30 am:** A smart alarm clock displays the time 8:12.
- 8:00 am (top):** A weather forecast map of the United States, titled "U.S. Water Outlook".
- 10:00 am:** A screenshot of an Amazon product page for "Pattern Recognition and Machine Learning" by Christopher M. Bishop. Below the product, a section titled "Customers Who Bought This Item Also Bought" recommends other books like "The Elements of Statistical Learning" and "Pattern Classification".
- 4:00 pm:** A screenshot of a search engine result for "computational social science". The top result is from "The Computational Social Science Society of the Americas".
- 6:00 pm:** A screenshot of a movie recommendation interface showing a grid of movie posters, including "NOEL" and "The Best of Christmas".
- 8:00 pm:** A screenshot of the Yelp website, featuring the text "yelp Your Guide to Everything Local" over a background image of a bowl of food.

- 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

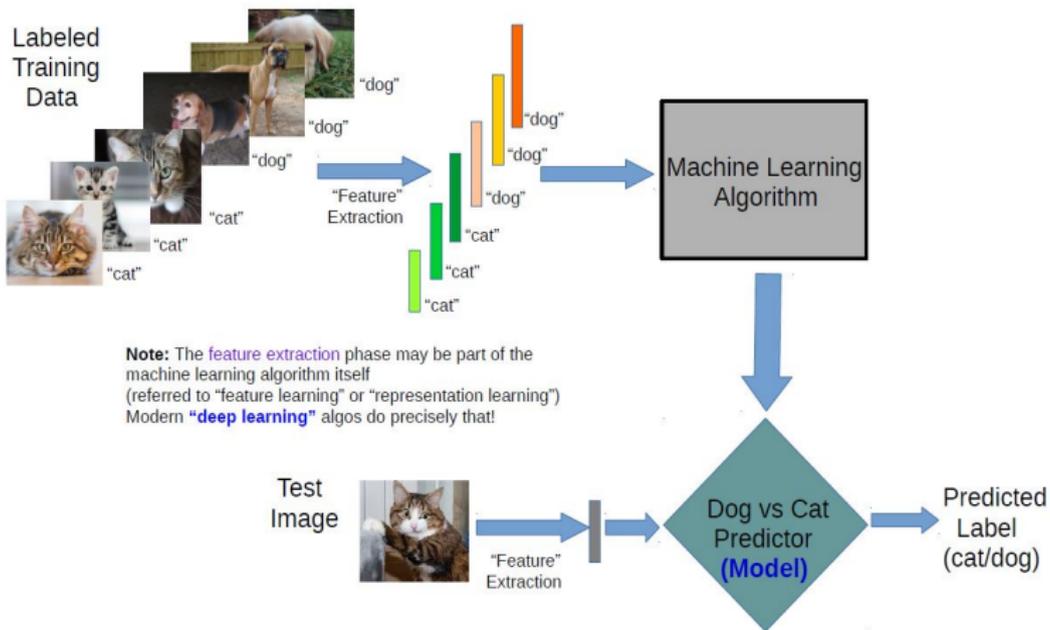


- Modern ML algorithms are heavily **"data-driven"**.
- Optimize a performance criterion using example data or **past experience**.

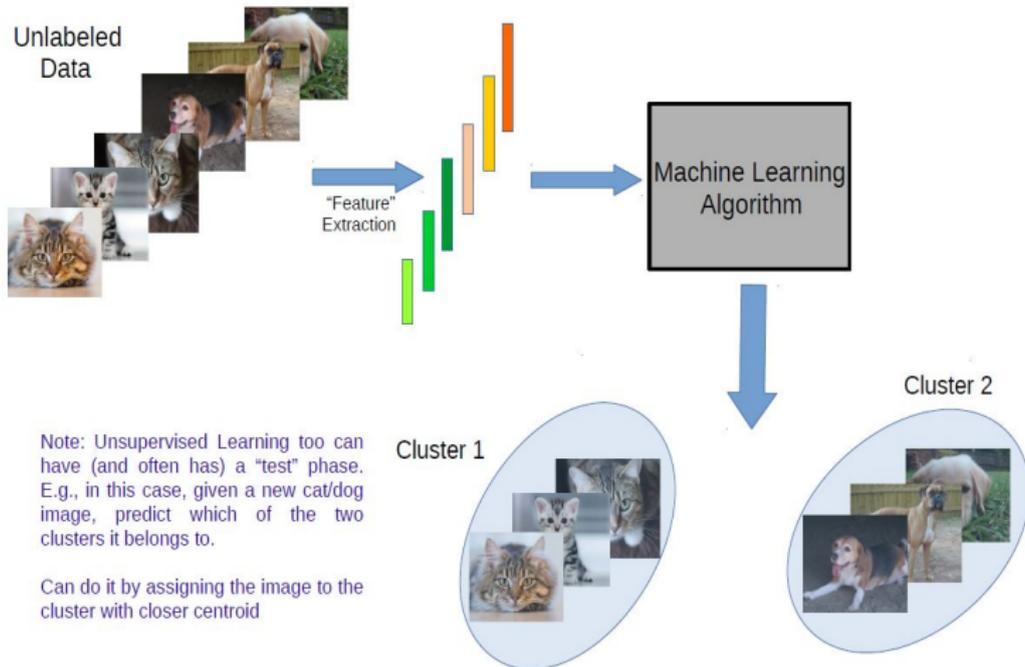
- **Unsupervised Learning:**
 - Uncover structure hidden in 'unlabelled' data.
 - Given network of social interactions, find communities.
 - Given shopping habits for people using loyalty cards: find groups of 'similar' shoppers.
 - Given expression measurements of 1000s of genes for 1000s of patients,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

- **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.

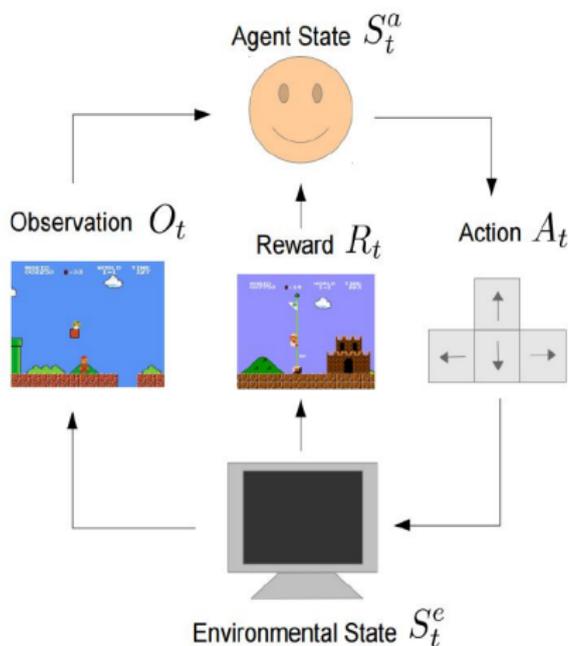
Supervised Learning: Predicting patterns in the data



Unsupervised Learning: Discovering patterns in the data



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

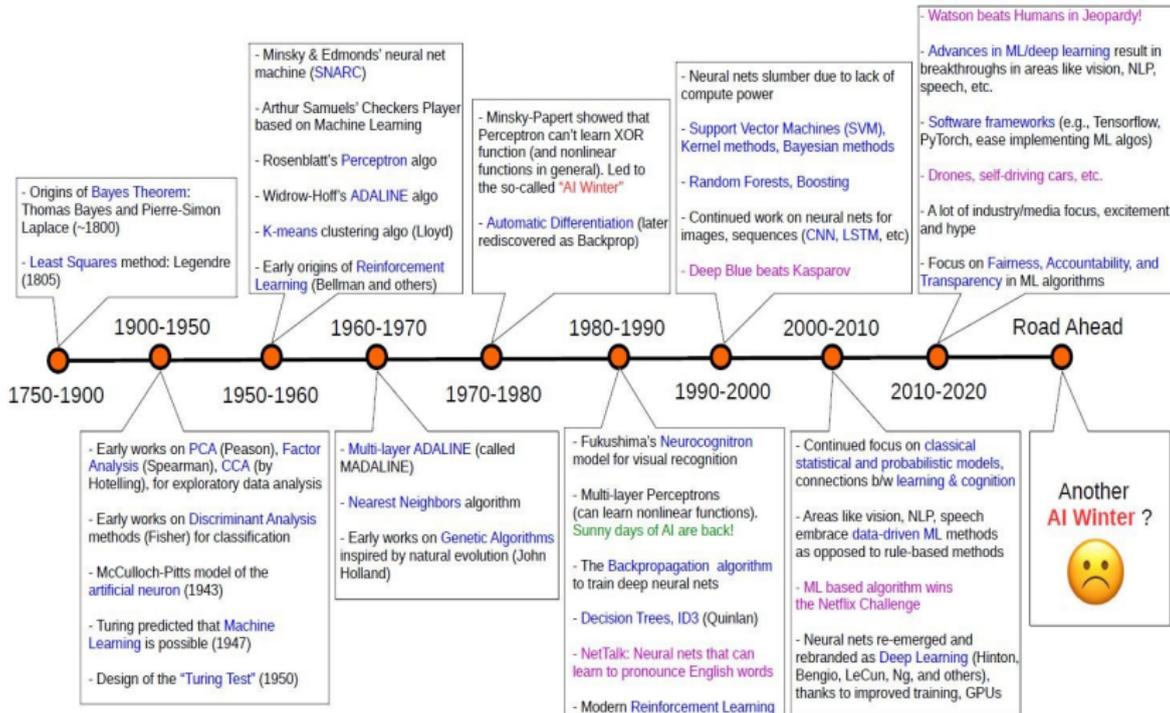


Agent's goal is to learn a policy for some task

Agent does the following repeatedly

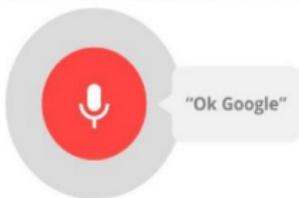
- Senses/observes the environment
- Takes an action based on its current policy
- Receives a reward for that action
- Updates its policy

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





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

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



English ▾   

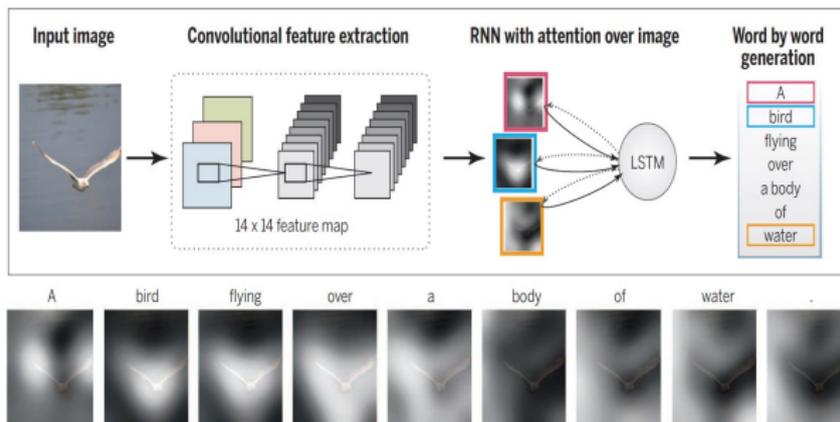
Welcome to this
course Edit

Hindi ▾  

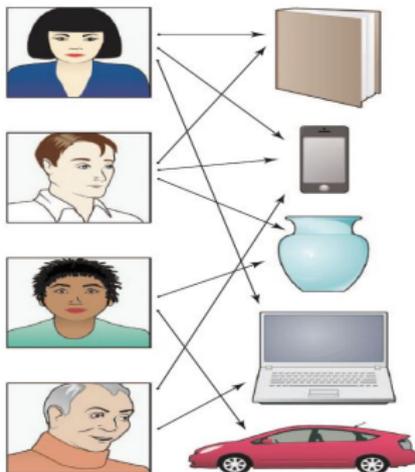
इस कोर्स में आपका स्वागत है
is kors mein aapaka svaagat hai

(even “transliterate”)

- 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.

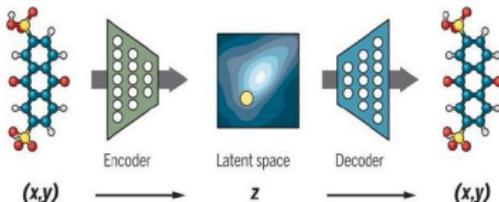


- **A recommendation system** is a machine-learning system that is based on data that indicate links between a set of users (e.g., people) and a set of items (e.g., products).
- A link between a user and a product means that the user has indicated an interest in the product in some fashion (perhaps by 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.

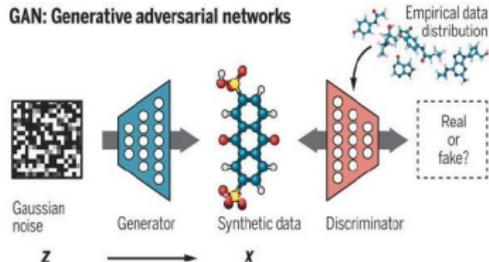


ML algorithms can understand properties of molecules and learn to synthesize new molecules.

VAE: Variational autoencoders

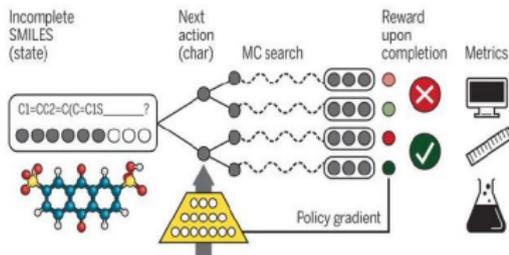


GAN: Generative adversarial networks

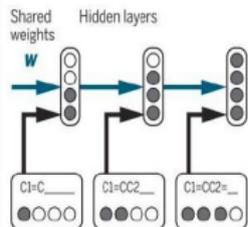


RL: Reinforcement learning

Policy gradient with Monte Carlo tree search (MCTS)



RNN: Recurrent neural network



Hybrid approaches

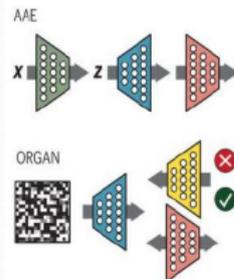
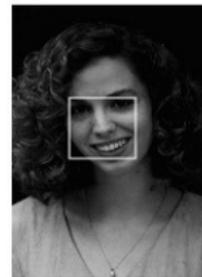


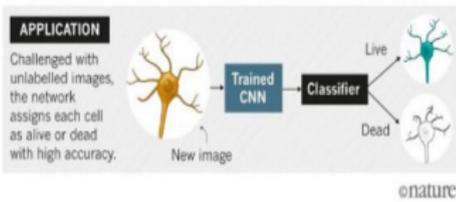
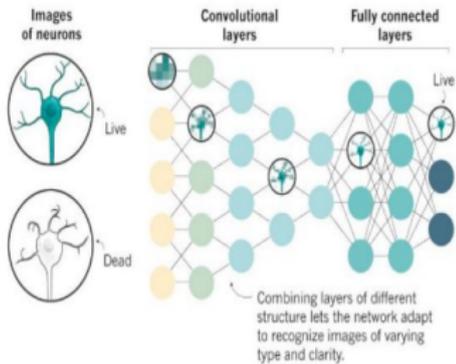
Figure: Inverse molecular design using machine learning: Generative models for matter engineering (Science, 2018)



Machine Learning helps Image Recognition

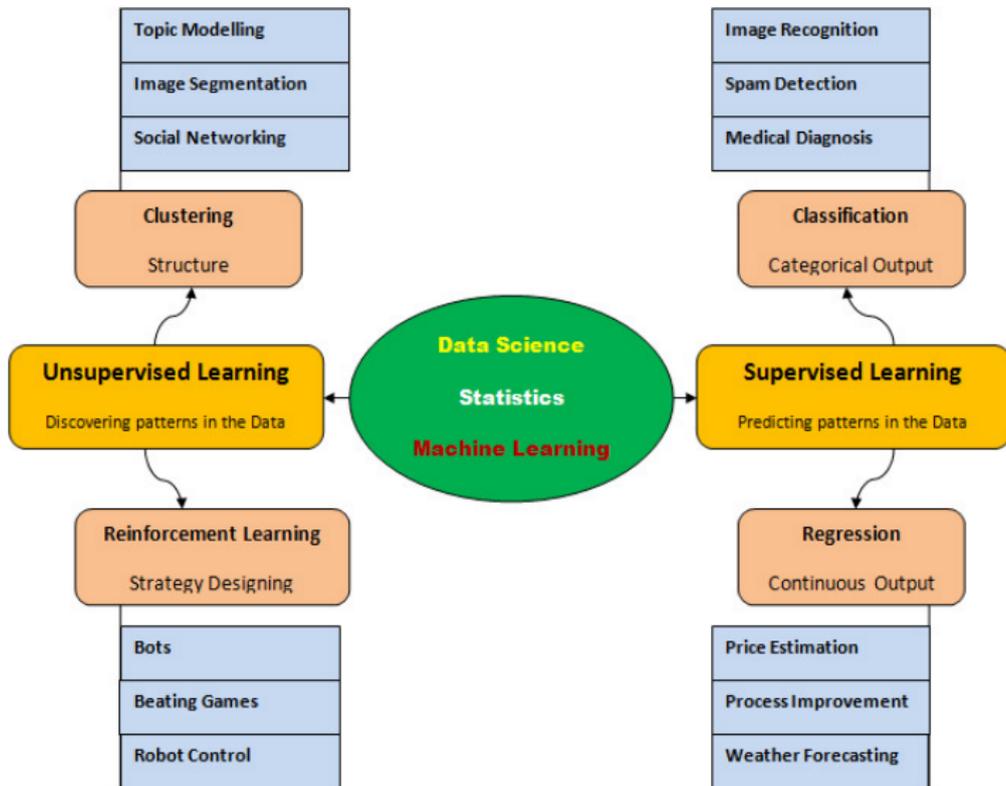


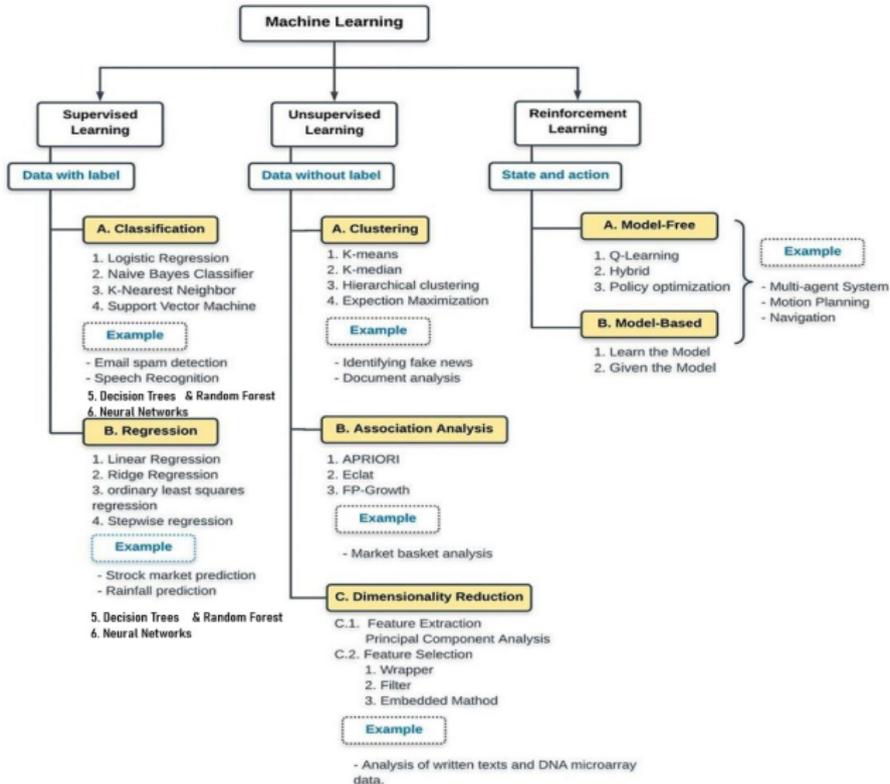
Biology



Finance









- Linear Regression (Galton, 1875).
- Linear Discriminant Analysis (R.A. Fisher, 1936).
- Logistic Regression (Berkson, JASA, 1944).
- k-Nearest Neighbor (Fix and Hodges, 1951).
- Parzen's Density Estimation (E Parzen, AMS, 1962)
- ARIMA Model (Box and Jenkins, 1970).
- Classification and Regression Tree (Breiman et al., 1984).
- Artificial Neural Network (Rumelhart et al., 1985).
- MARS (Friedman, 1991, Annals of Statistics).
- SVM (Cortes and Vapnik, Machine learning, 1995)
- Random forest (Breiman, 2001).
- Deep Convolutional Neural Nets (Krizhevsky, Sutskever, Hinton, NIPS 2012).
- Generative Adversarial Nets (Ian Goodfellow et al., NIPS 2014).
- Deep Learning (LeCun, Bengio, Hinton, Nature 2015).
- Bayesian Deep Neural Network (Yarin Gal, Islam, Zoubin Ghahramani, ICML 2017).

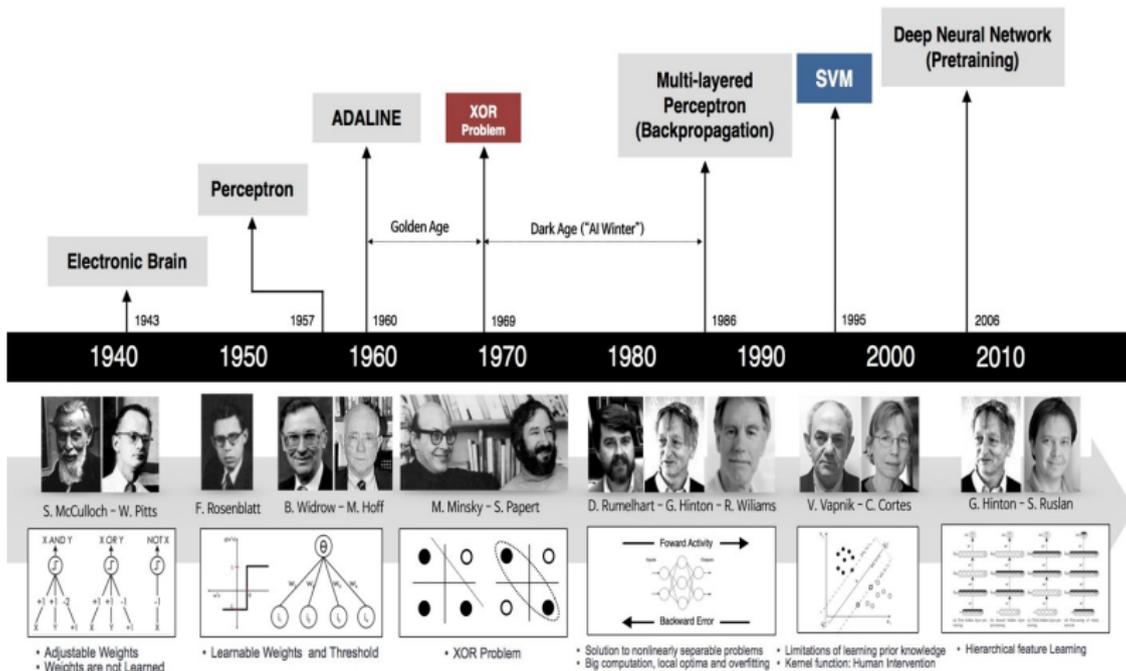
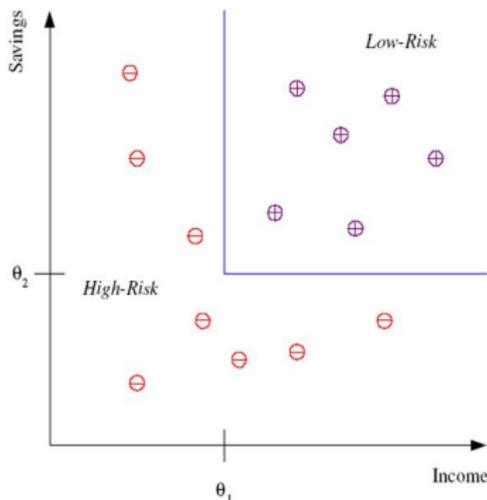
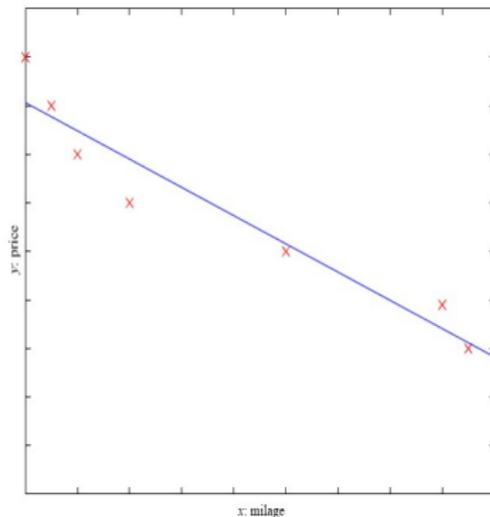


Fig: Developments of Neural Network Models

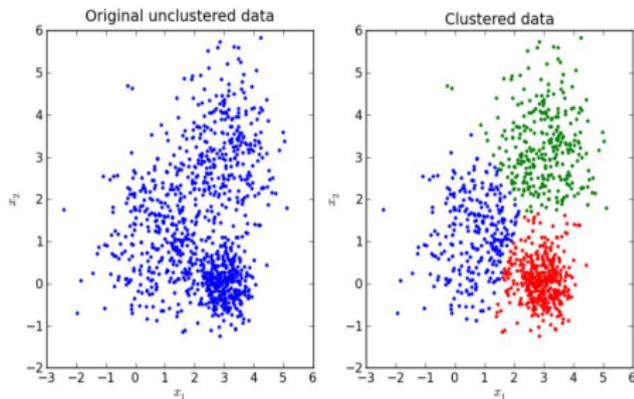
- **Example:** Credit scoring.
- Differentiating between low-risk and high-risk customers from their income and savings.
- **Discriminant:** IF $\text{Income} > \theta_1$ AND $\text{Savings} > \theta_2$ THEN low-risk ELSE high-risk.
- **Classification:** Learn a linear/nonlinear separator (the "model") 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.



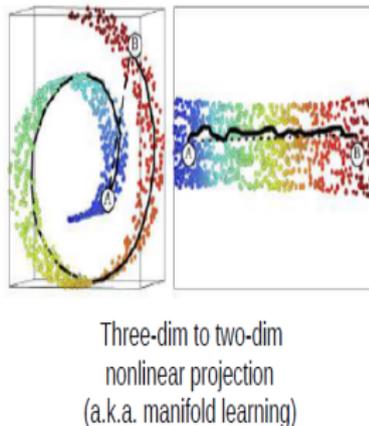
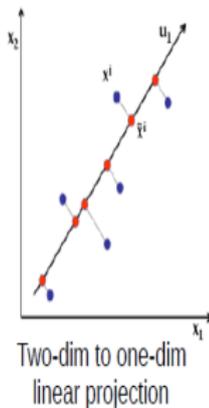
- **Example:** Price of a used car.
- X : car attributes; Y : price and $Y = f(X, \theta)$
- $f()$ is the model and θ is the model parameters.
- **Regression:** Learn a line/curve (the "model") using training data consisting of Input-output pairs (each output is a real-valued number).
- Use it to predict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

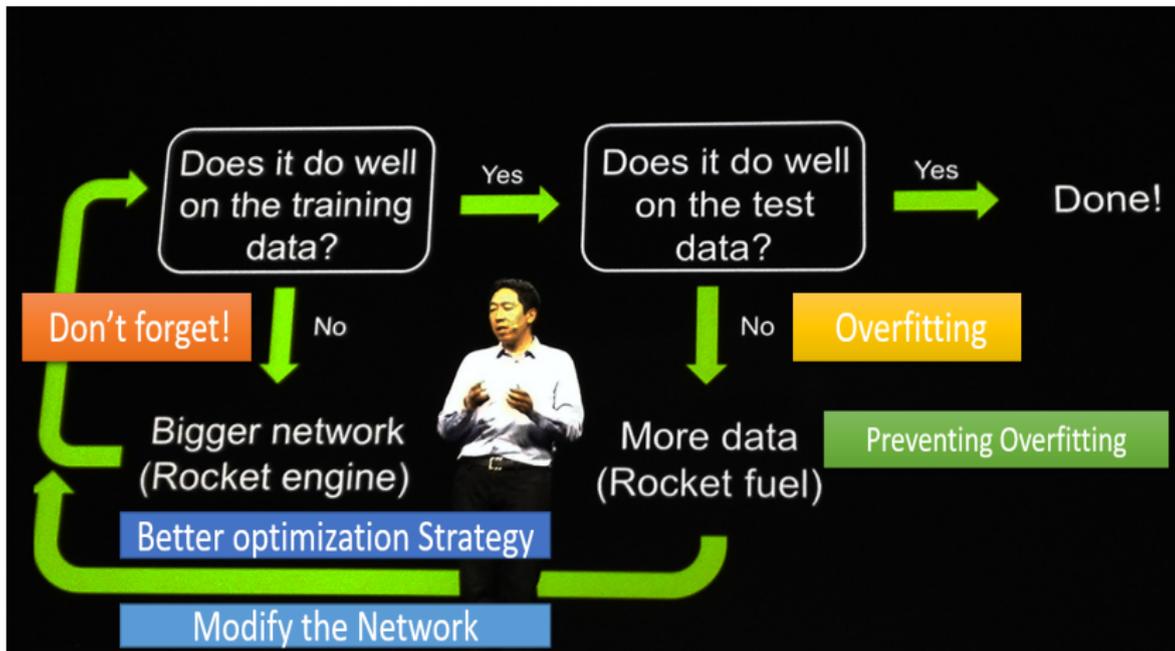


- **Given:** Training data in form of **unlabeled instances**
 $\{x_1, x_2, \dots, x_N\}$
- **Goal:** Learn the intrinsic latent structure that summarizes/explains data
- **Clustering:** Learn the grouping structure for a given set of unlabeled inputs.
- Homogeneous groups as latent structure: **Clustering**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



- **Given:** Training data in form of **unlabeled instances** $\{x_1, x_2, \dots, x_N\}$
- **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.







- An Extract from a Post by DeepMind on December 1, 2021 ([LINK](#)):
 - More than a century ago, [Srinivasa Ramanujan](#) shocked the mathematical world with his extraordinary ability to see remarkable patterns in numbers that no one else could see.
 - The self-taught mathematician from India described his insights as deeply intuitive and spiritual, and patterns often came to him in vivid dreams.
 - These observations captured the tremendous beauty and sheer possibility of the abstract world of pure mathematics.
 - In recent years, we have begun to see AI make breakthroughs in areas involving deep human intuition, and more recently on some of the hardest problems across the sciences.
 - As part of DeepMind's mission to solve intelligence, we explored the potential of machine learning (ML) to recognize mathematical structures and patterns, and help guide mathematicians toward discoveries they may otherwise never have found — demonstrating for the first time that AI can help at the forefront of pure mathematics.



nature

Explore content ▾

About the journal ▾

Publish with us ▾

[nature](#) > [articles](#) > article

Article | [Open Access](#) | [Published: 01 December 2021](#)

Advancing mathematics by guiding human intuition with AI

[Alex Davies](#) , [Petar Veličković](#), [Lars Buesing](#), [Sam Blackwell](#), [Daniel Zheng](#), [Nenad Tomašev](#), [Richard Tanburn](#), [Peter Battaglia](#), [Charles Blundell](#), [András Juhász](#), [Marc Lackenby](#), [Geordie Williamson](#), [Demis Hassabis](#) & [Pushmeet Kohli](#) 

Nature **600**, 70–74 (2021) | [Cite this article](#)

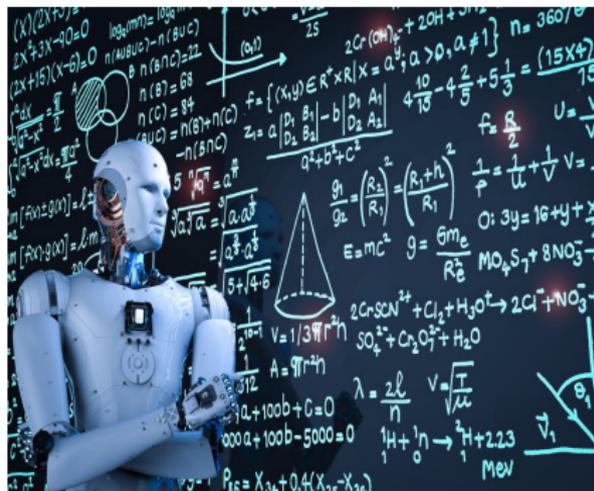
122k Accesses | **1** Citations | **1553** Altmetric | [Metrics](#)

Abstract

The practice of mathematics involves discovering patterns and using these to formulate and prove conjectures, resulting in theorems. Since the 1960s, mathematicians have used computers to assist in the discovery of patterns and formulation of conjectures¹, most

TOPIC 4 : 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".



- 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).

"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).

- 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).



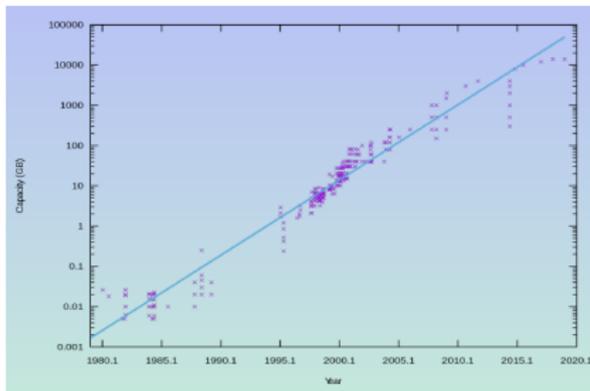
- Voice recognition tool "Harpy" masters about 1000 words (1970s, CMU, US Defense).
- System capable of analyzing entire word sequences (1980).
- Siri was the first modern digital virtual assistant installed on a smartphone (2011).
- Watson won the TV show Jeopardy! (2011).



TOPIC 5 : BIG DATA

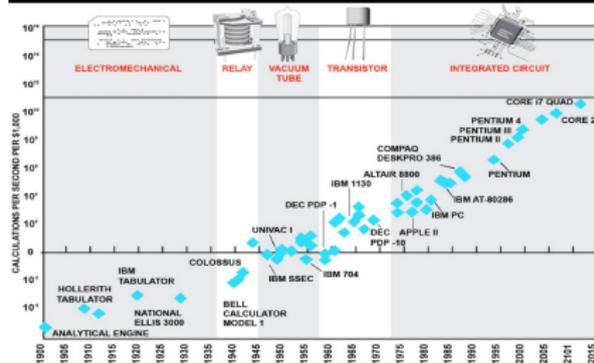
Storage and Processing capacities

- **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)

115 Years of Moore's Law



Processor capacity (Moore's law)

- 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}



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)



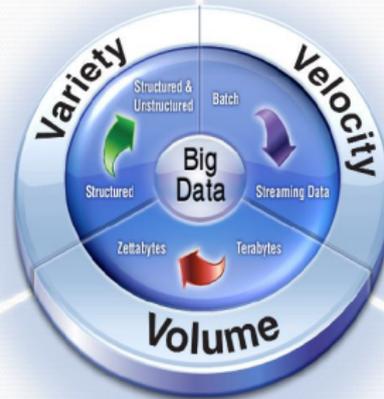
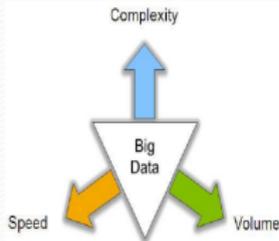
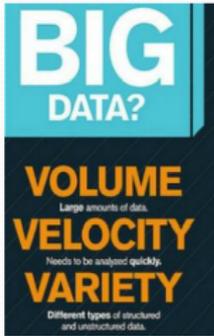
Sensor technology and networks
(Measuring all kinds of data)

“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.





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.

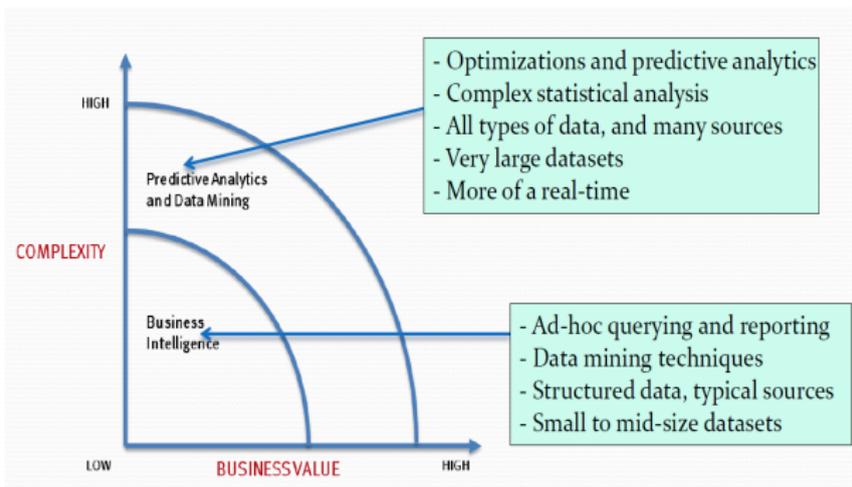
Variety:

- 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.

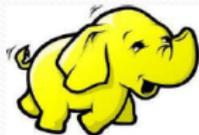
Velocity:

- 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 is more real time in nature than traditional applications...



- 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.



- 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.



Machine Learning:

- Theoretical ML (Explainable & Robust AI)
- Applied ML (Supervised & Unsupervised Problems)
- Deep Learning (Application to CV, IP, IOT, etc.)

Big Data & Analytics:

- Business Analytics (Quality, Reliability, OR Modelling)
- Data Assimilation
- Methods for Big Data
- High-Dimensional Statistics

Other Areas:

- Quantum Computing
- 3D Modelling, AR & VR
- Blockchain & Cryptography
- Robotics



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

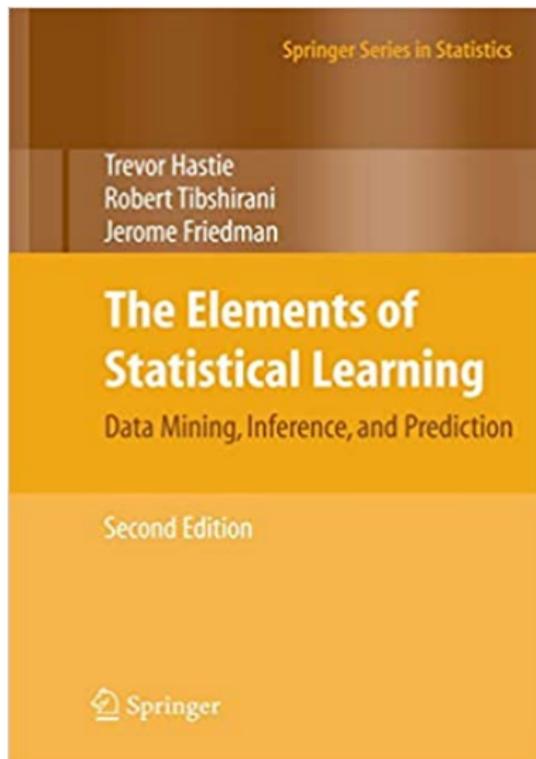
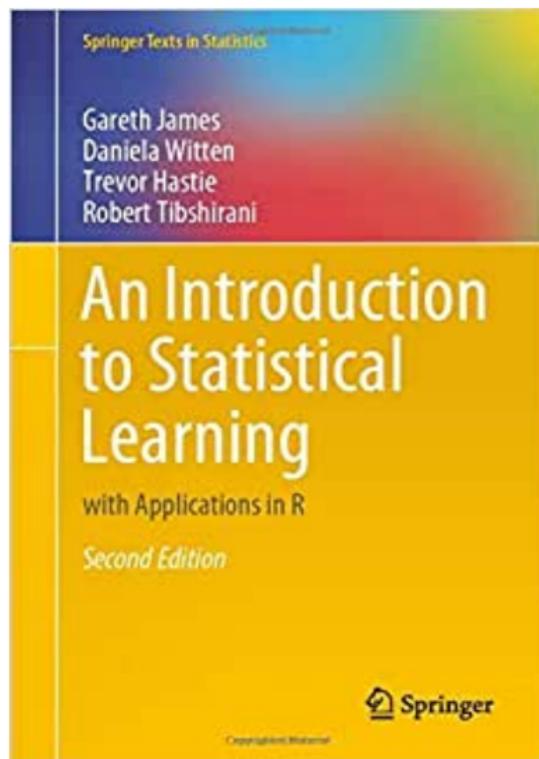


Human



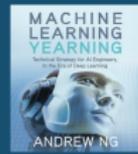
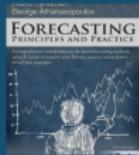
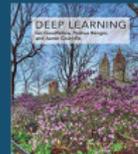
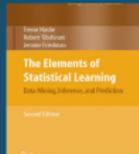
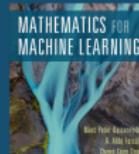
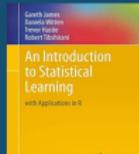


- Emerging Research Areas in Statistics, Data Science & ML:
 - All the areas of Machine Learning.
 - High-Dimensional Small Sample Size Data.
 - Spatio-temporal Data.
 - Fairness in ML and Interpretable Deep Learning.
 - Collaborative research with other disciplines.
- Statistical Thinking:
 - Accepting randomness in formulating models and ideas.
 - Realizing and Analyzing dependence of conclusions on assumptions.
 - Measuring uncertainty in some ways without forgetting dependence on assumptions.





Data Science, Statistics & ML Booklist

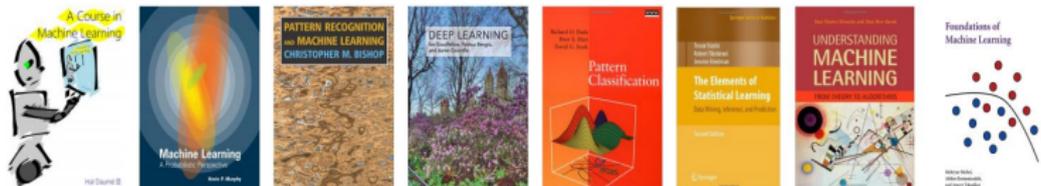


Prepared by Dr. Tanujit Chakraborty

Very important readings...



Advanced reading materials...



Essential Video Lectures for BS & MS Data Science students:

<https://www.ctanujit.org/video-lectures.html>

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



For any other queries mail me at tanujit.chakraborty@sorbonne.ae



*Thank
you!*