Recent Trends in Statistics and Machine Learning

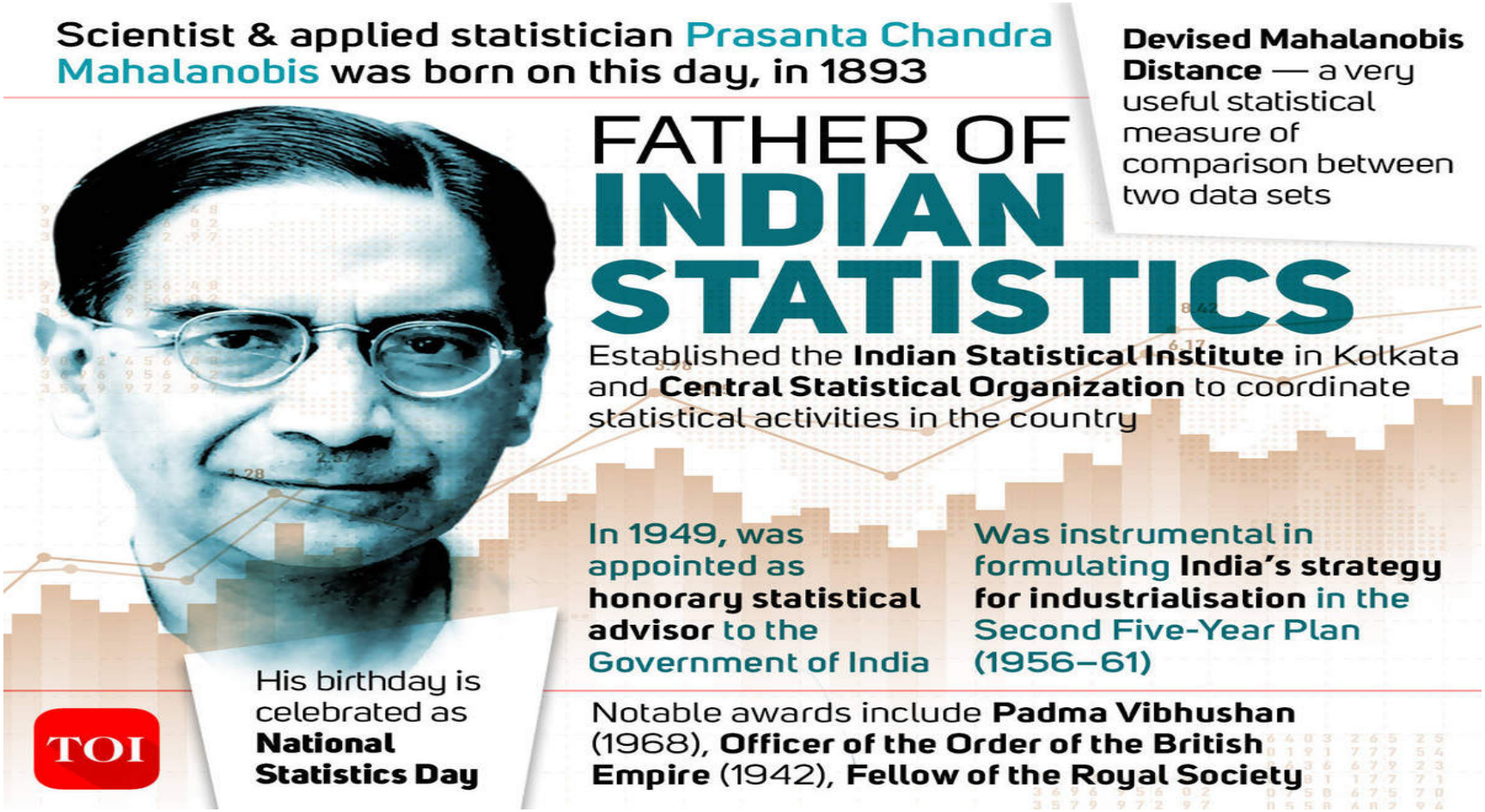
by

Dr. Tanujit Chakraborty

Ph.D. from Indian Statistical Institute, Kolkata.
Currently a Postdoctoral Fellow at IIIT Bangalore.
Visiting Faculty at XIM University, Bhubaneswar.

Webpage : www.ctanujit.org/

Prof. Prasanta Chandra Mahalanobis (29 Jun 1893 - 28 Jun 1972)

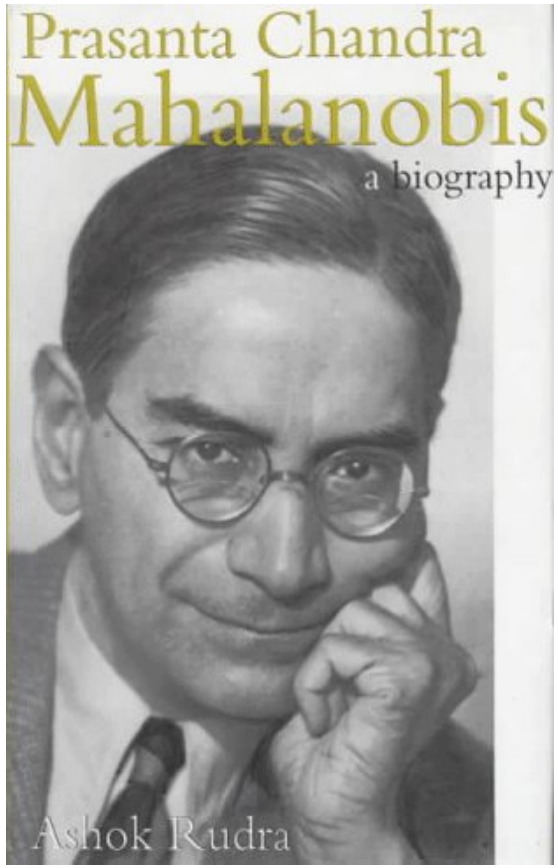


Historical Facts

- "Prasanta Chandra Mahalanobis (Professor) was a physicist by training, a statistician by instinct and an economist by conviction."
  - Professor C R Rao.

- Mahalanobis met Nelson Annandale (Director of GSI) at the 1920 Indian Science Congress. Annandale asked Mahalanobis to analyze anthropometric measurements of Anglo-Indians in Calcutta. Mahalanobis distance (1922) is a device that compares two different populations.

- His main contributions to statistical theory and applications are multivariate methods in taxonomy (Mahalanobis distance), optimum design of large scale sample surveys, and use of econometric models in planning.



Recent Trends & Buzzwords

- 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.
- 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.
- Machine learning is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed.
- Data science is the study of the generalizable extraction of knowledge from data, yet the key word is science.
- Statistics is the study of the collection, analysis, interpretation, presentation and organization of data.

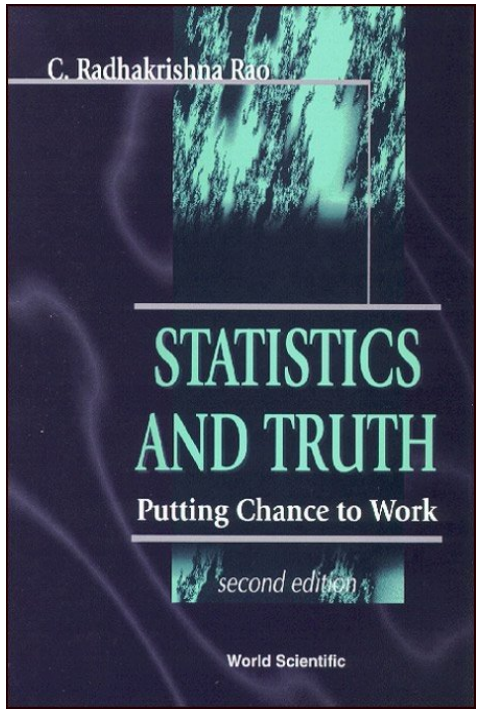
Reality & Truth

- 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, Professor Emeritus at Pennsylvania State University, Jan 1987.

- 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, Nov 2017.

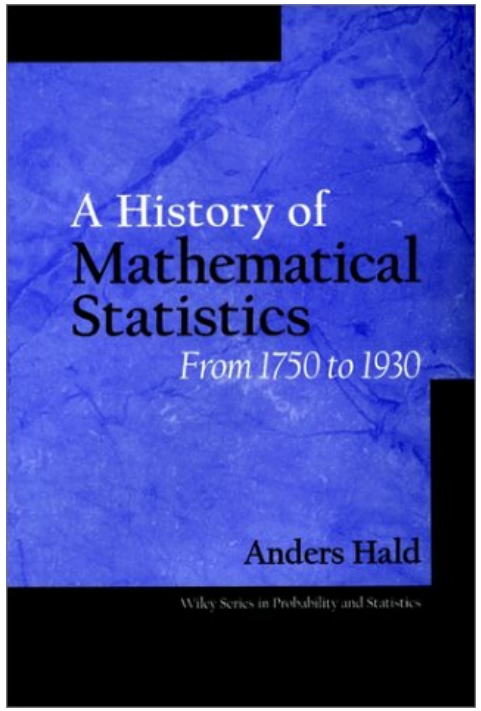


Statistics: View Points

"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 develop ment"

- Professor P C Mahalanobis.

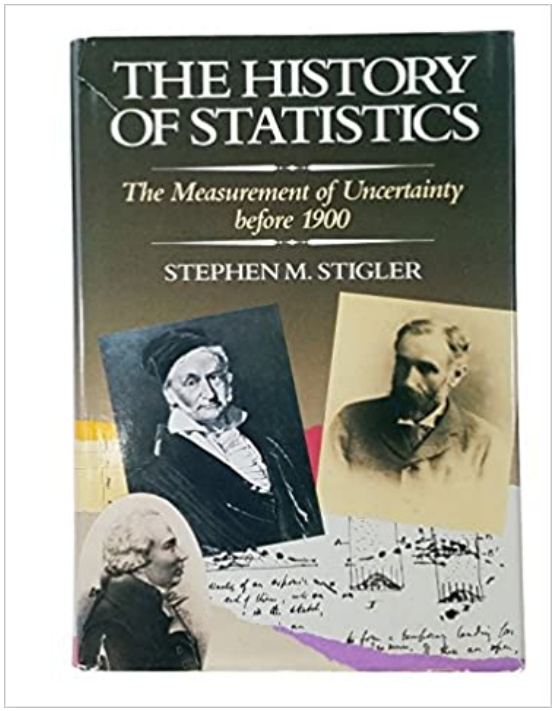


- Role of Statistics:
  1. Making inference from samples
  2. Development of new methods for complex data sets
  3. Quantification of uncertainty and variability
- Two Views of Statistics:
  1. Statistics as a Mathematical Science
  2. Statistics as a Data Science

History of Statistics : Pre 1900

- Probability and its application to Gambling & Astronomy

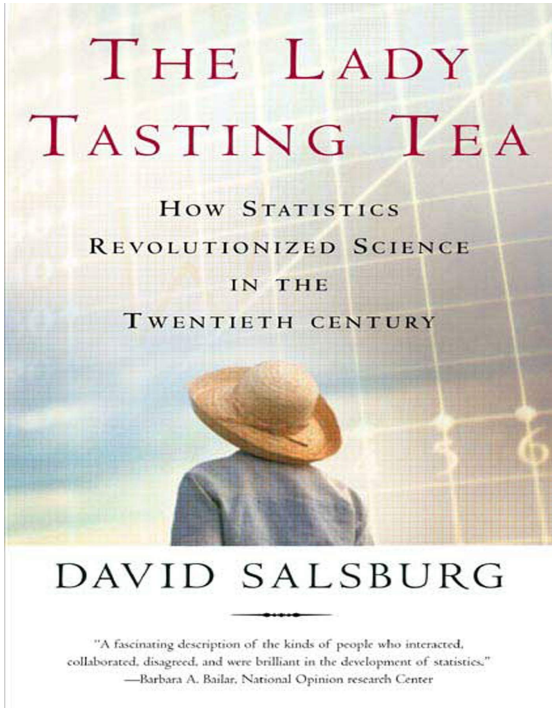
  - Jacob Bernoulli (Prob. distribution, Law of large numbers, etc.)
  - Pierre-Simon Laplace (Double exponential, Transformation, etc.)
  - Thomas Bayes (Bayes' theorem, etc.)
  - Gauss and Legendre (Least Square Method)
  - Francis Galton (Correlation & Regression)
  - Karl Pearson (χ^2 test, distribution,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



History of Statistics : 1900 - 1980

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

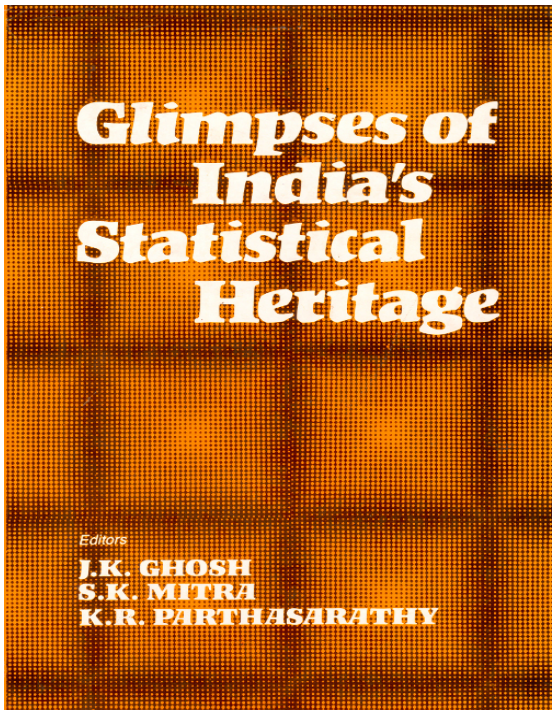
  - Ronald Fisher (Discriminant analysis, Likelihood, ANOVA & DOE, etc.)
  - Jerzy Neyman, Egon Sharpe Pearson & Wald (Decision theory, Optimality, etc.)
  - Lehmann, Hotelling, Anderson & Tukey (Multivariate & Inferential Statistics, etc.)
  - Box, Cox, Jenkin and Blackwell (Time Series)
  - Shewhart, Deming, Taguchi & Juran (SQC)
  - Efron, Breiman, Friedman, Cramer (Modern Statistical Tools)
  - and many others.



History of Statistics : Contributions from India

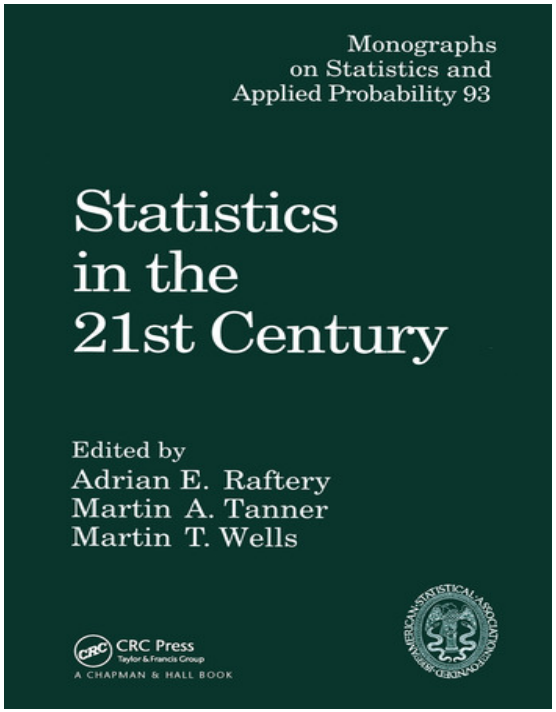
- In a speech at ISI, Professor R A Fisher once pointed out that more than half the qualified statisticians working in the world were Indians, for quite some time. Most of them were Professor Rao's students.

  - (Professor) PC Mahalanobis (Mahalanobis Distance, Founder of ISI.)
  - C. R. Rao (Linear Models, Multivariate Analysis, Orthogonal arrays, etc.)
  - J K Ghosh (Bayesian & High-dimensional Statistics, etc.)
  - Samarendra Nath Roy (Multivariate Statistics)
  - Raj Chandra Bose (Combinatorial Design)
  - Debabrata Basu (Statistical Inference)
  - and many others.



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).



Few Famous Quotations : Beyond Normality

- "Normality is a myth; there never was, and there never will be a normal distribution"

  - Roy C. Geary (1947; Biometrika, vol. 34, 248).

- "Everybody believes in the exponential law of errors (i.e. the normal distribution), the experimenters, because they think it can be proved by mathematicians; and the mathematicians, because they believe that it has been established by observations"

  - E.T. Whittaker and G. Robinson (1967).

- "... the statisticians knows ... that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false he can often derive results which match to a useful approximation, those found in real world"

  - George W. Box (1976, Jour. Amer. Stat. Asso., vol. 71, 791-799).

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 : 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, 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"**).

Role of data: Present

The World is Data Rich

Astronomy

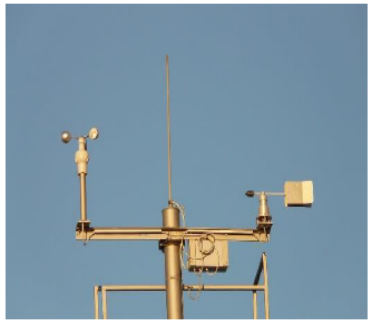
Social Networks

Healthcare

Banking

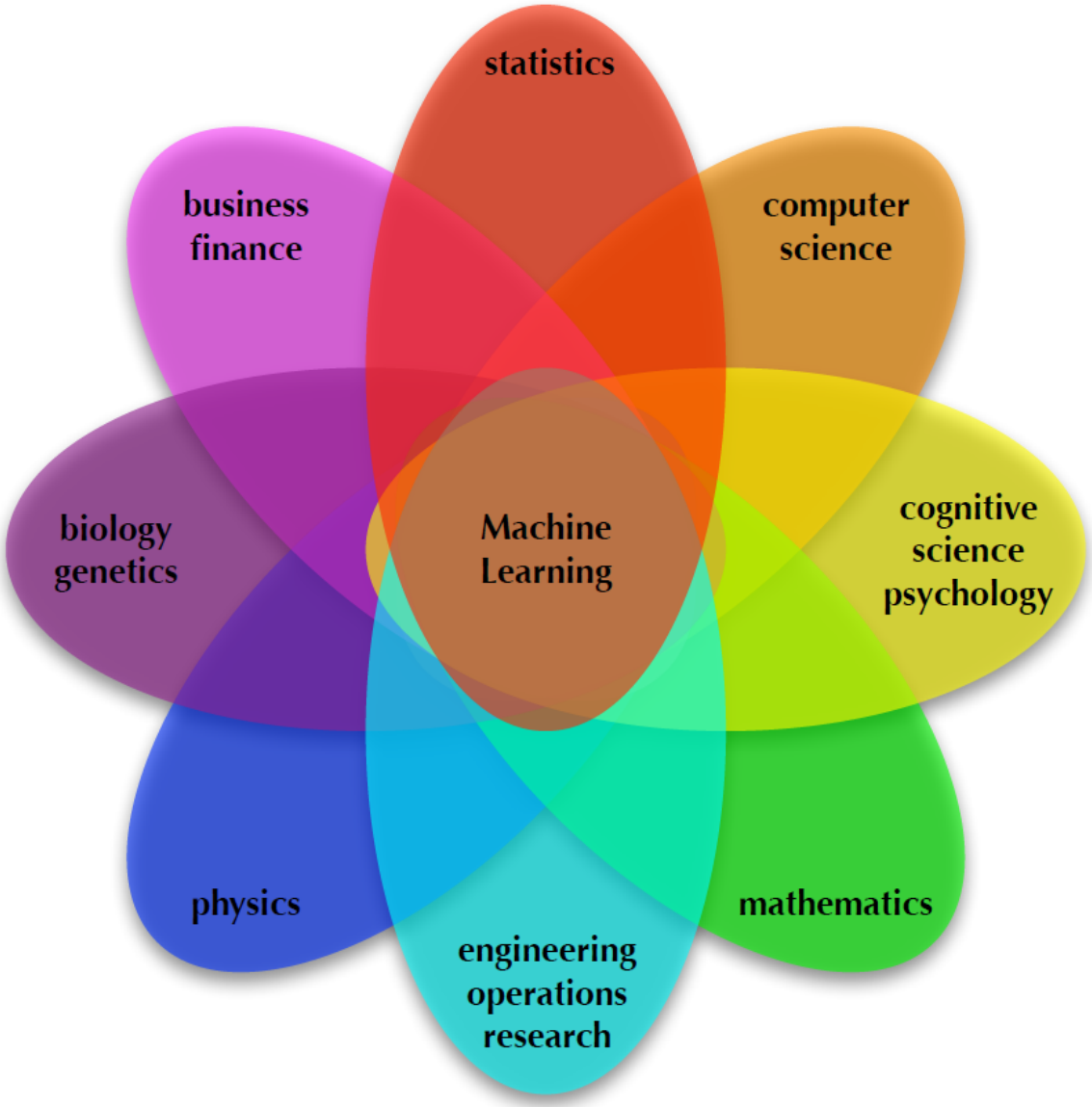
Genomics

Weather measurements



What is Machine Learning?

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



What is the Difference?

- Traditional Problems in Applied Statistics:

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

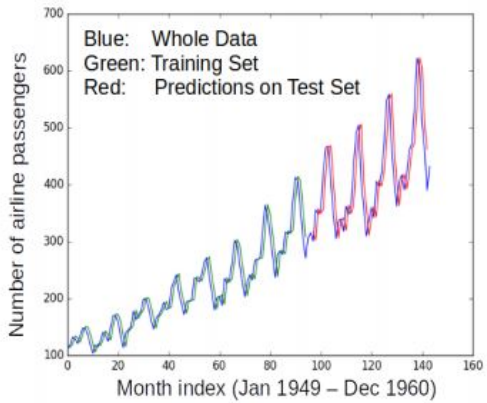
- Current Situation : Information Revolution

  - 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

- 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:

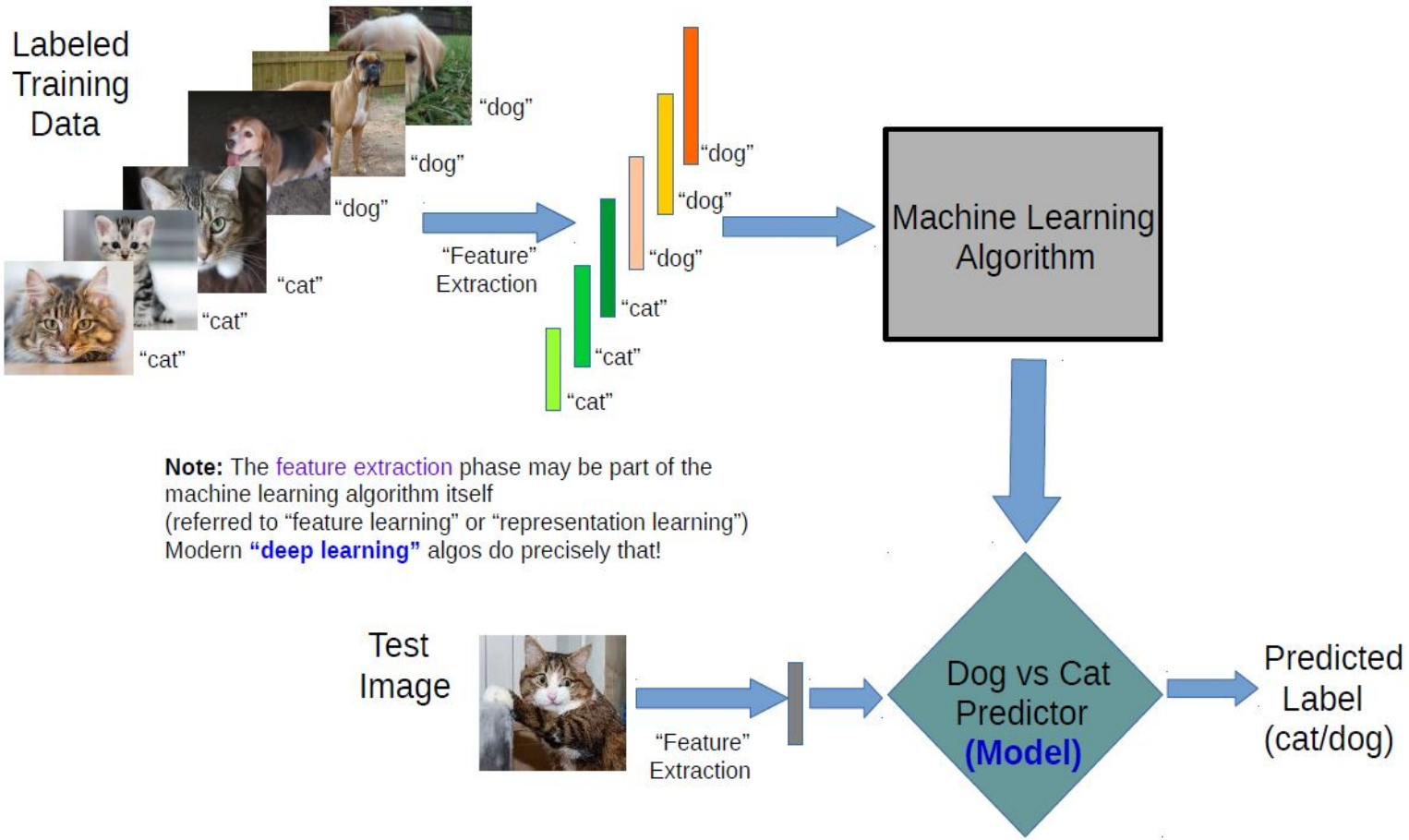
  - A database of examples, each of which has multiple labels corresponding to different prediction tasks.

- Reinforcement Learning:

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

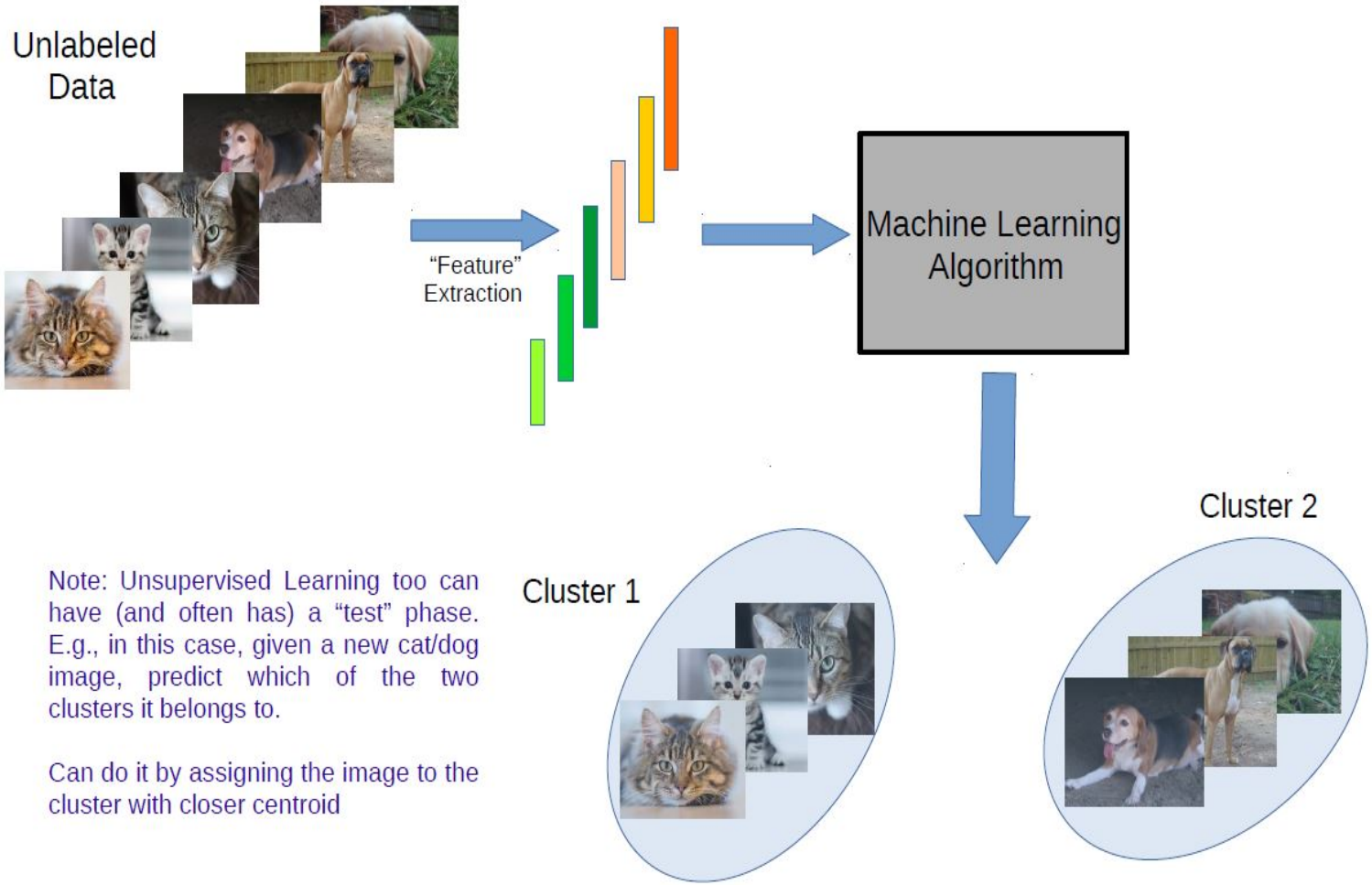
A Typical Supervised Learning Workflow (for Classification)

Supervised Learning: Predicting patterns in the data



A Typical Unsupervised Learning Workflow (for Clustering)

Unsupervised Learning: Discovering patterns in the data

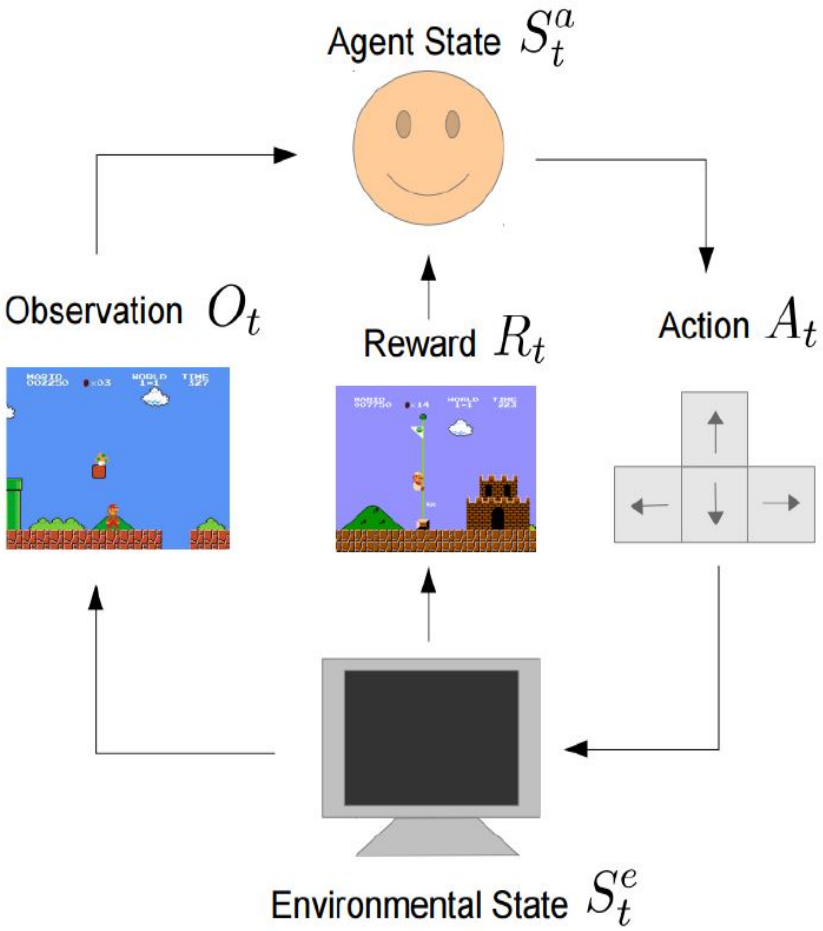


Note: Unsupervised Learning too can have (and often has) a "test" phase. E.g., in this case, given a new cat/dog image, predict which of the two clusters it belongs to.

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)



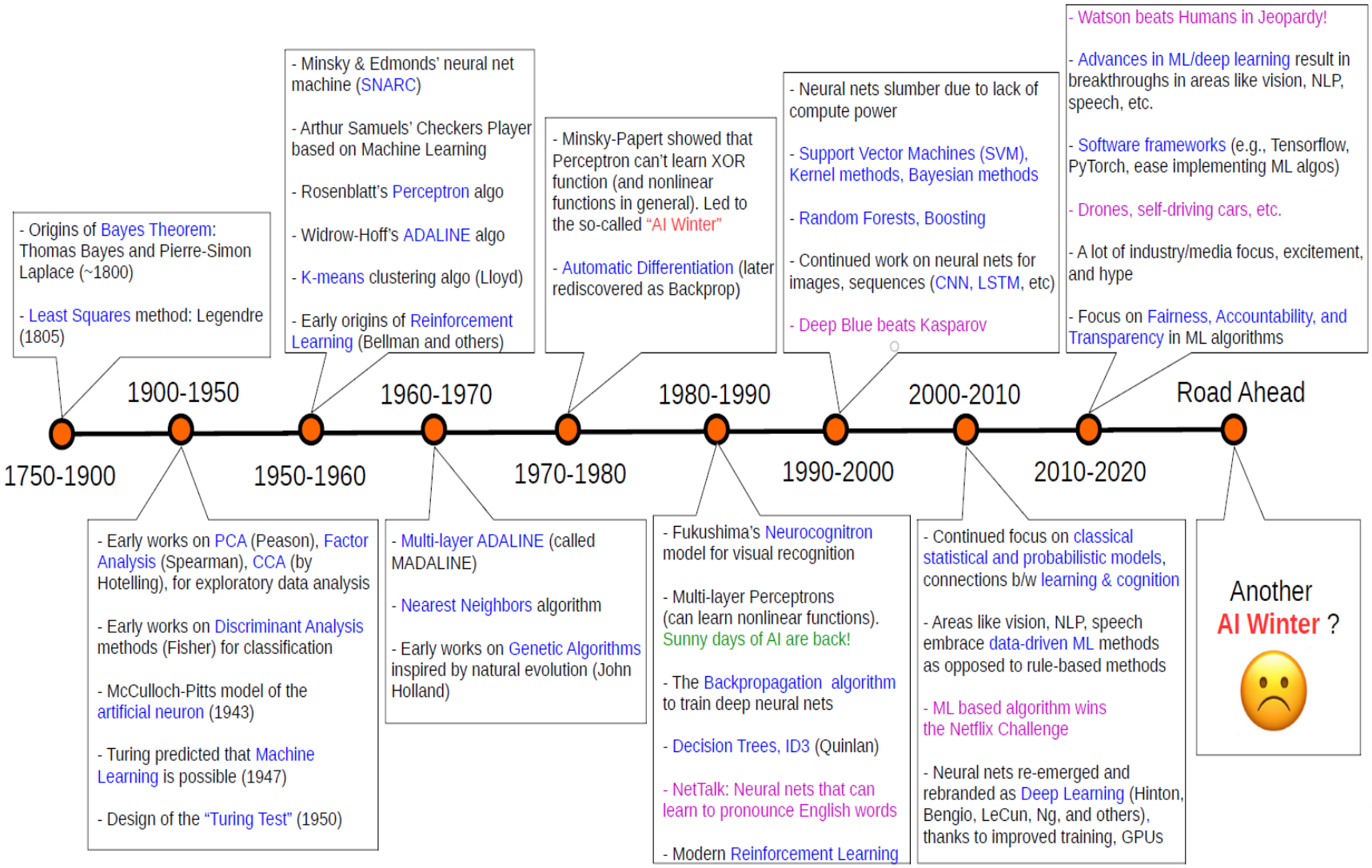
Agent's goal is to learn a policy for some task

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

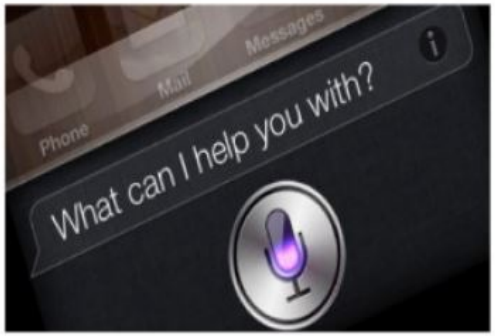
There IS supervision, not explicit (as in Supervised Learning) but rather implicit (feedback based)

Parallel Progress in Statistics & Machine Learning

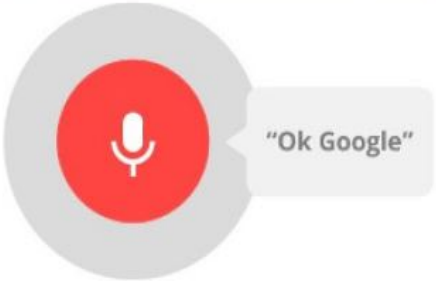


Machine Learning in the real-world

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



Recommendation systems




Predictive Policing


Online Fraud Detection

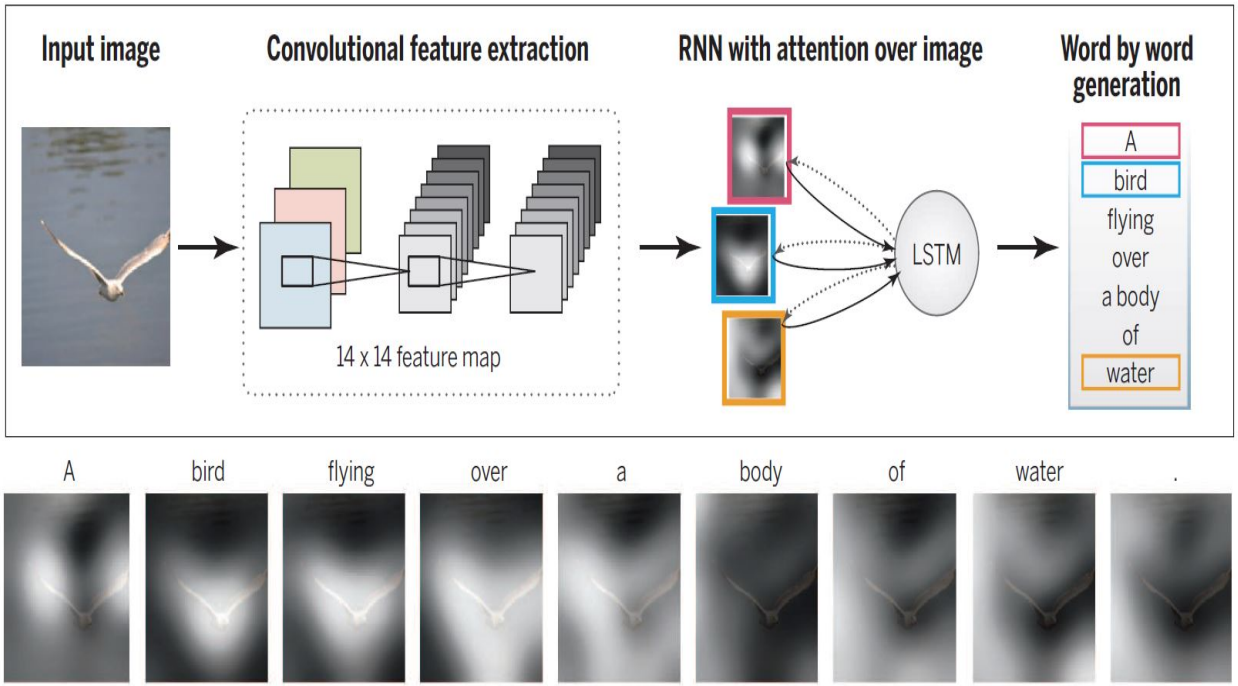
Machine Learning helps Natural Language Processing

ML algorithm can learn to translate text

ML algorithms can learn to translate speech in real time

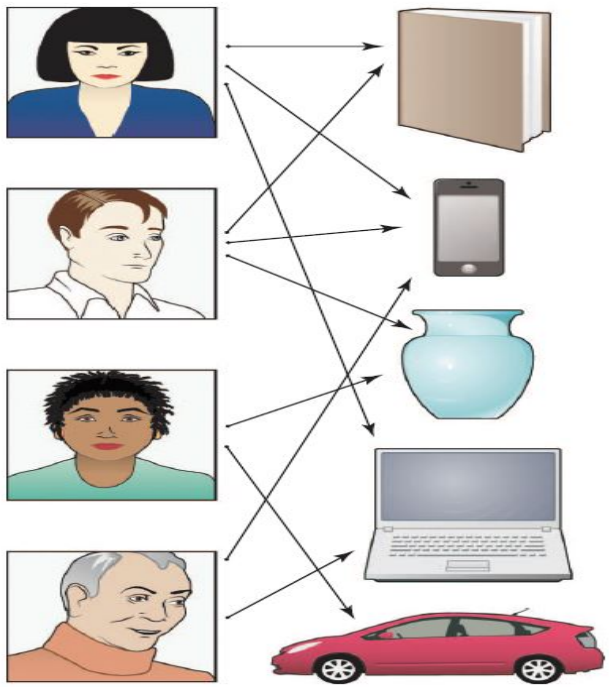
Machine Learning helps Computer Vision

- 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.



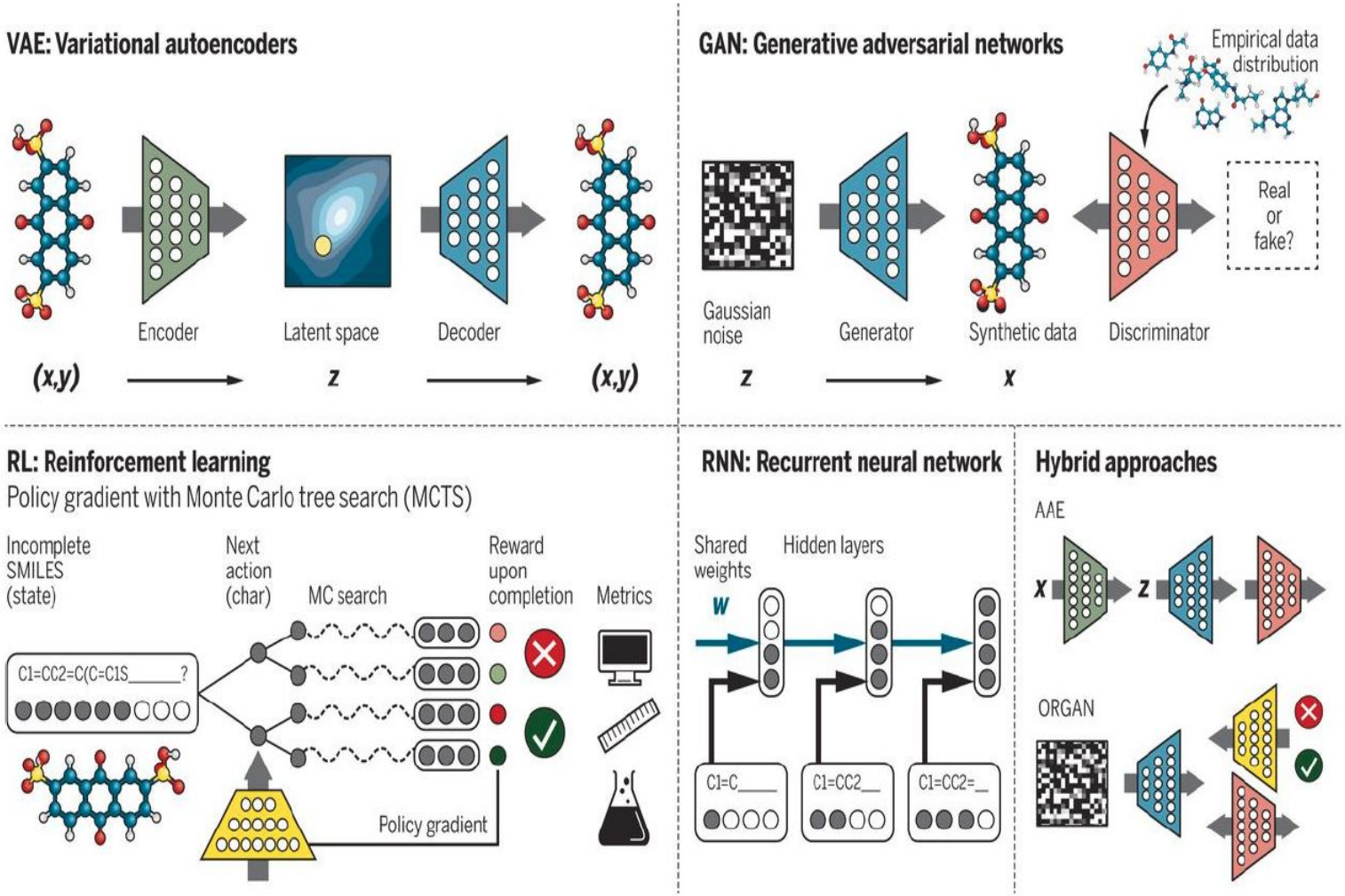
Machine Learning helps Recommendation systems

- **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.



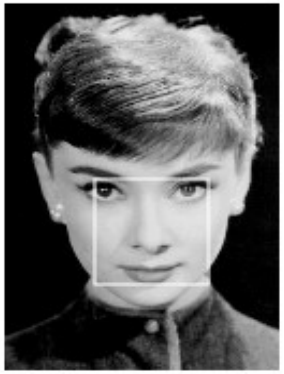
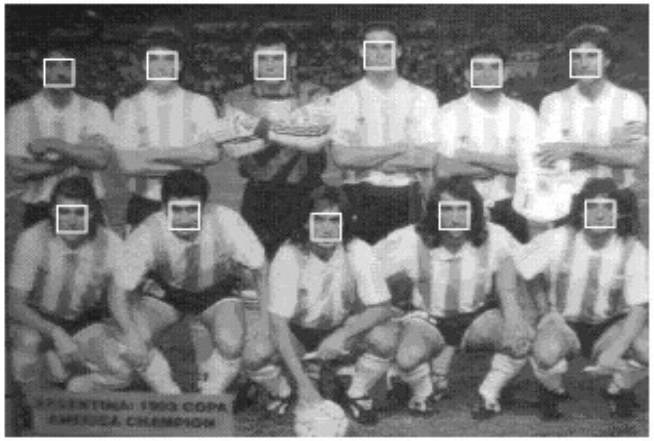
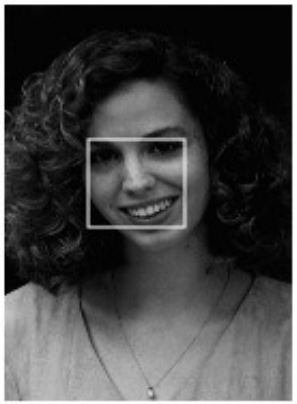
Machine Learning helps Chemistry

ML algorithms can understand properties of molecules and learn to synthesize new molecules[1].



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

Machine Learning helps Image Recognition

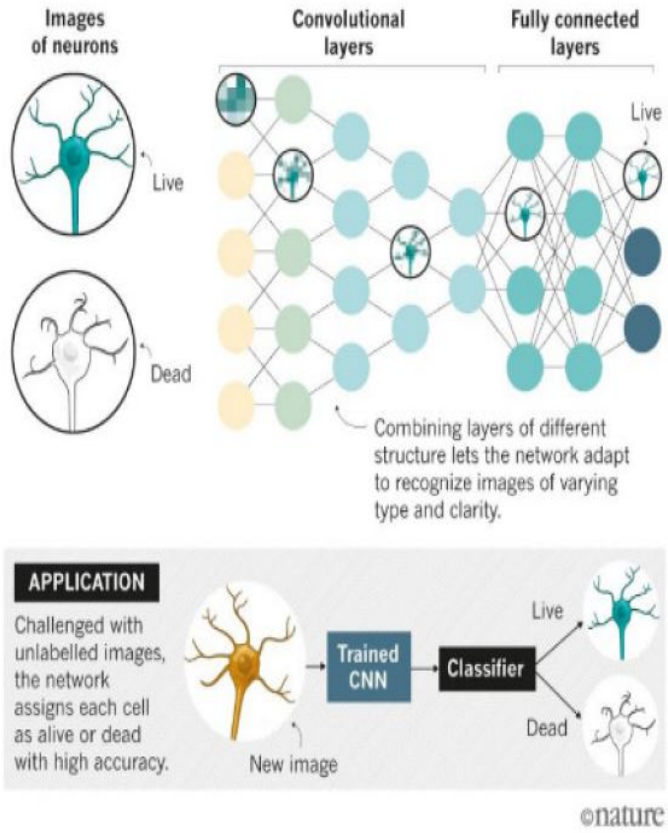









Machine Learning helps Many Other Areas...

Biology



Finance

Take Home

- 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.

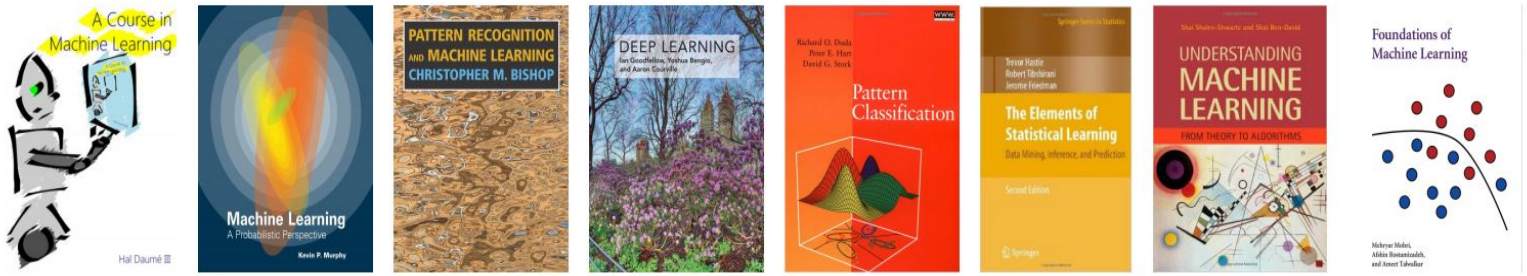
- Statistical Thinking:

  - 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.

Textbook and References

Start with these three books

Then you can start reading these books



Essential Video Lectures for UG & PG Statistics Students:
https://www.ctanujit.org/video-letures.html


thank you