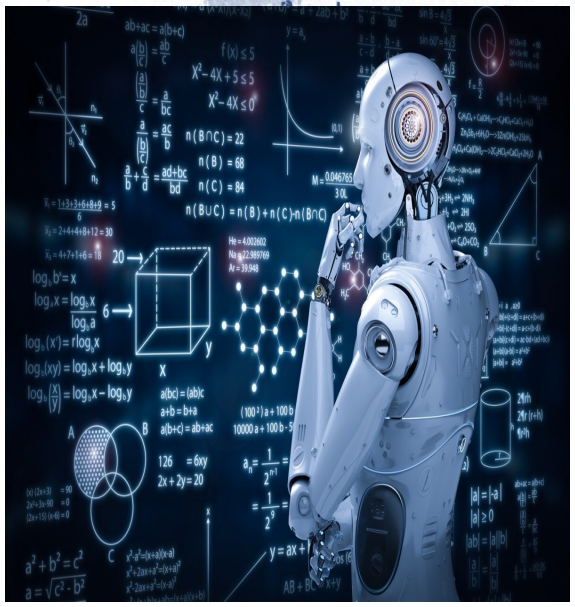
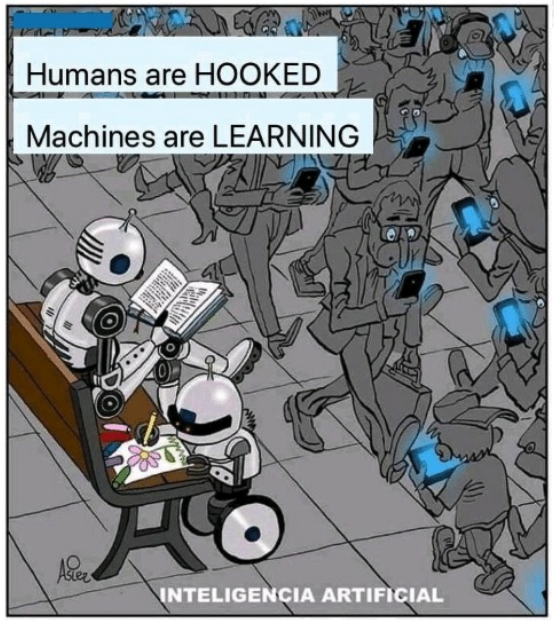
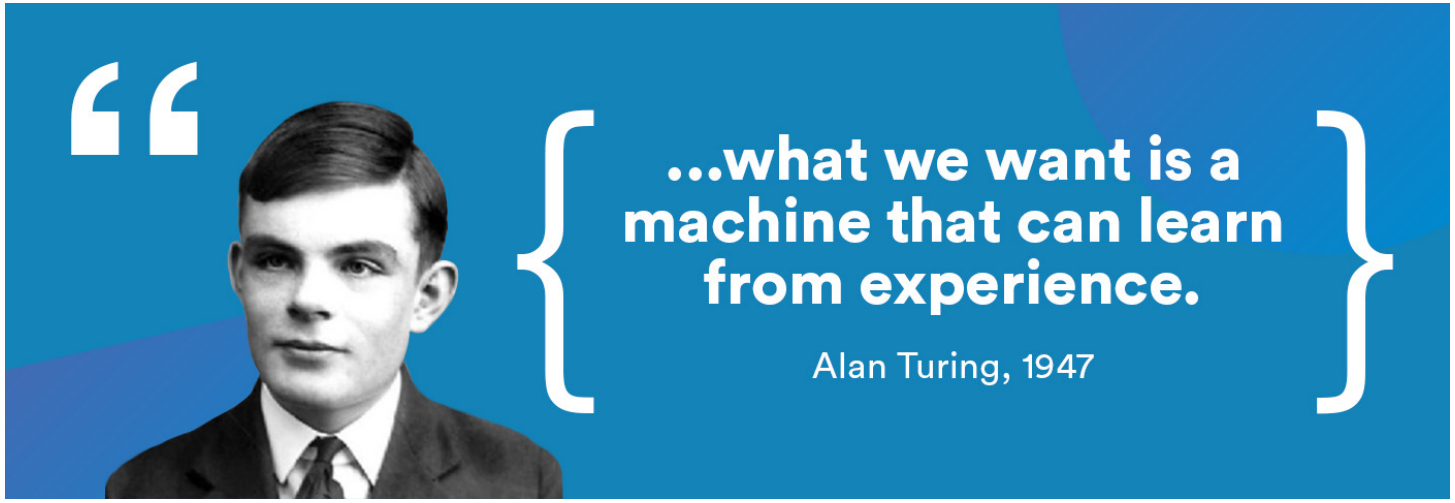
Introduction to Machine Learning

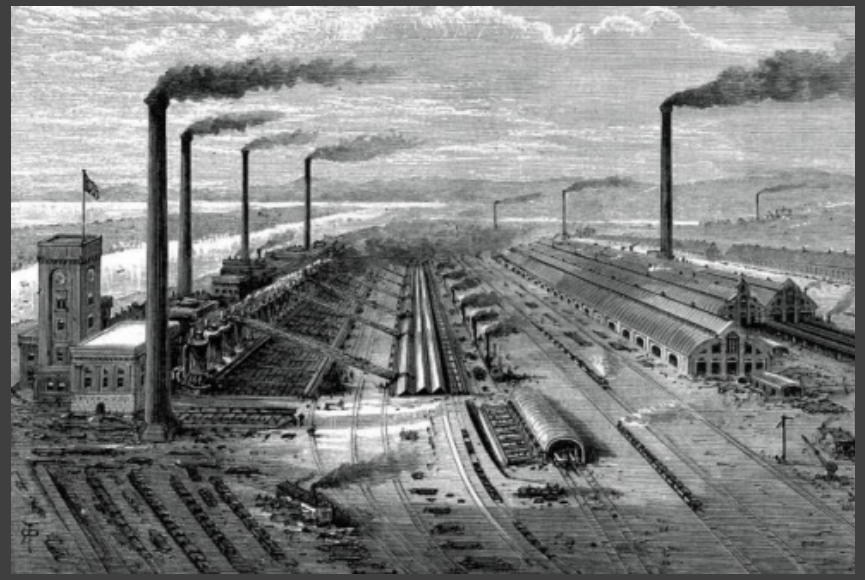


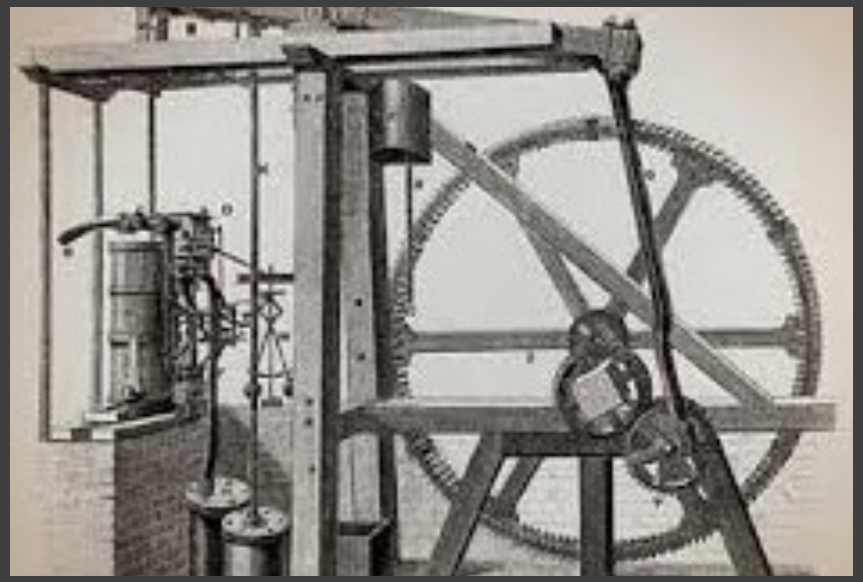
Dr. Srijith P K
Computer Science and Engineering
IIT Hyderabad





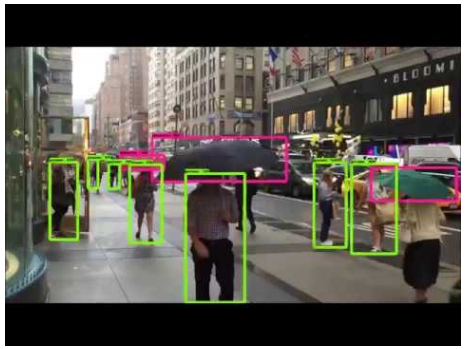
- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)
- •“Machine learning is the next Internet” (Tony Tether, Director, DARPA)
- “AI is one of the most important things humanity is working on. It is more profound than, I don't know, electricity or fire.” - Sundar Pichai (Google CEO)

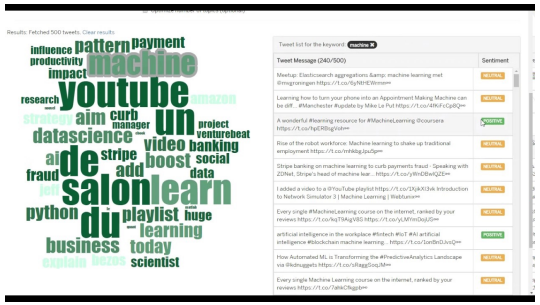





MW

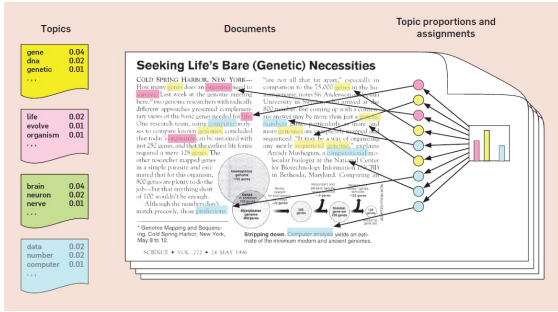

I'm originally from Dublin but now live in Edinburgh.
मैं डबलिन से मूल हूँ, लेकिन अब एडिनबुरग में रहते हैं
Machine Translation
Source Input
Target Output




influence
productivity
impact
pattern
payment
machine
research
youtube
amazon
aim
curb
manager
un
project
venturebeat
datascience
video
banking
boost
social
de
stripe
add
data
salon
learn
fraud
python
du
playlist
huge
learning
business
today
scientist

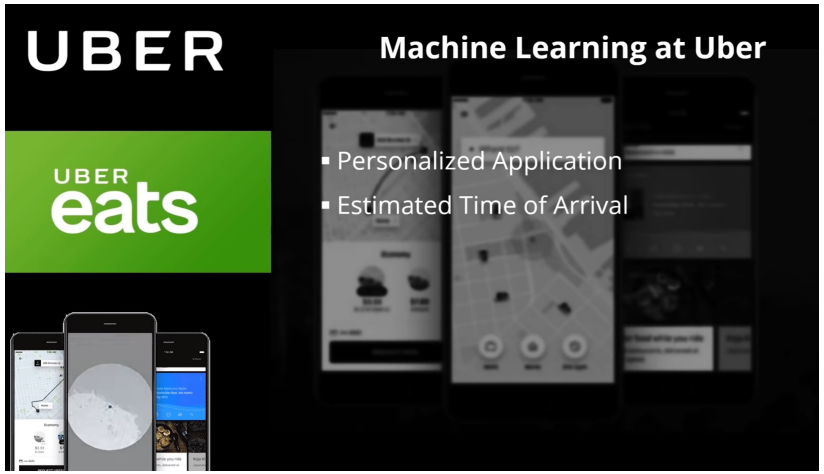

Internet of Things
Open Data
Smart Agriculture
Smart Retail
Smart Home
Smart Mobility
Education
SMART CITY
Smart Health
Smart Grid/
Smart Energy
Smart Government

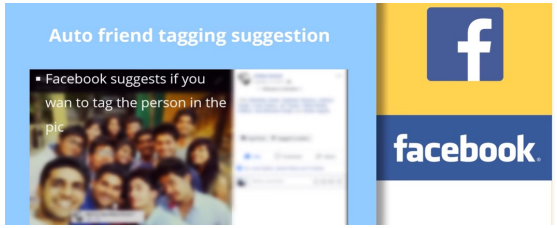

You
Tube
Be
Meetup

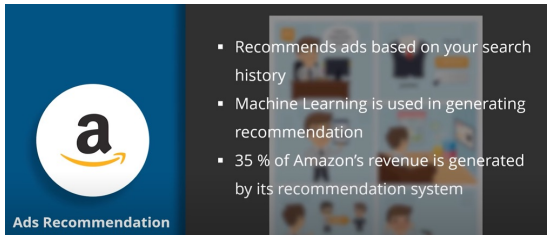

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
life 0.02
evolve 0.01
organism 0.01
brain 0.04
neuron 0.02
nerve 0.01
data 0.02
number 0.02
computer 0.01
Seeking Life's Bare (Genetic) Necessities


Recommender Systems

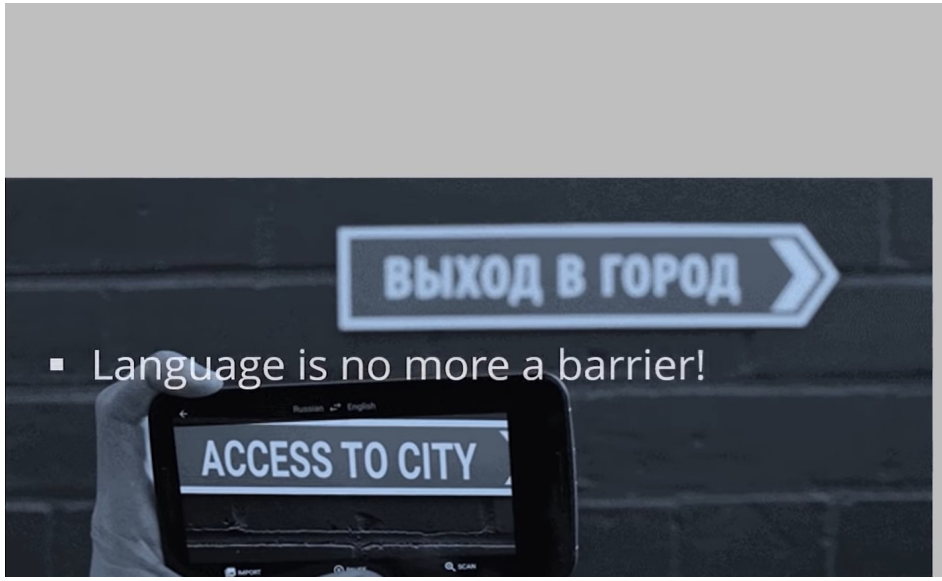

Self-driving cars
Autonomous driving

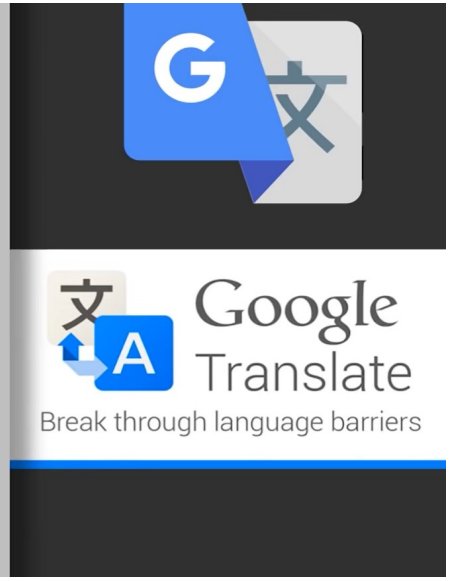

UBER
Machine Learning at Uber
UBER eats
Personalized Application
Estimated Time of Arrival

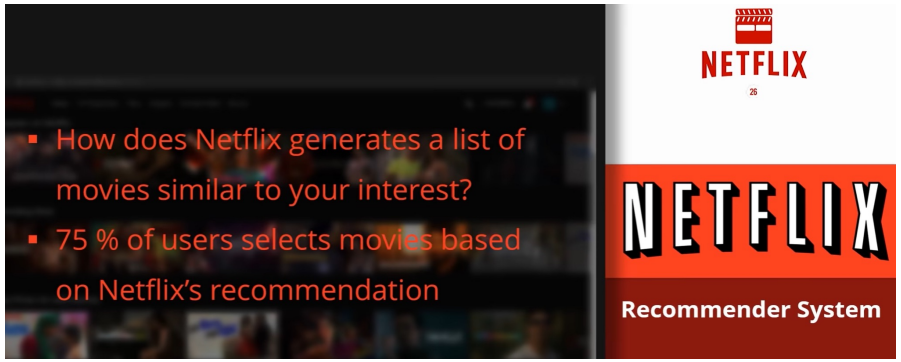

Auto friend tagging suggestion
Facebook suggests if you wan to tag the person in the pic
facebook

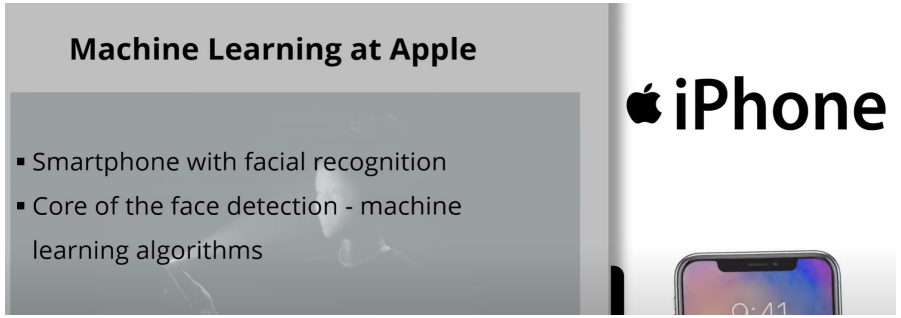

Recommends ads based on your search history
Machine Learning is used in generating recommendation
35 % of Amazon's revenue is generated by its recommendation system
Ads Recommendation


TESLA
SCANNING
Driverless Cars!
Tesla's AI is driven by Nvidia's H/W focusing mainly on unsupervised learning
Crowdsources data from all of its vehicles and its drivers - internal & external sensors


ВЫХОД В ГОРОД
Language is no more a barrier!
ACCESS TO CITY


Google Translate
Break through language barriers


How does Netflix generates a list of movies similar to your interest?
75 % of users selects movies based on Netflix's recommendation
NETFLIX
NETFLIX
Recommender System


Machine Learning at Apple
Smartphone with facial recognition
Core of the face detection - machine learning algorithms
iPhone

Big data Era

- Google: 3.5 billion search queries every day.2.
- Facebook: 350 million photos are uploaded to Facebook each day. Facebook generates 4 petabytes of data every day.
- Every day, 306.4 billion emails are sent, and 5 million Tweets are made in Twitter:
- Astronomy: Satellite data is in hundreds of PB.

- **1.7MB of data** is created every second by every person during 2020.
- In the last two years alone, the astonishing **90%** of the world's data has been created.
- **2.5 quintillion bytes** of data are produced by humans every day.
- **463 exabytes** of data will be generated each day by humans as of 2025.
- **95 million** photos and videos are shared every day on Instagram.
- By the end of 2020, **44 zettabytes** will make up the entire digital universe.
- Every day, **306.4 billion emails** are sent, and **5 million Tweets** are made.

https://techjury.net/blog/how-much-data-is-created-every-day/#gref

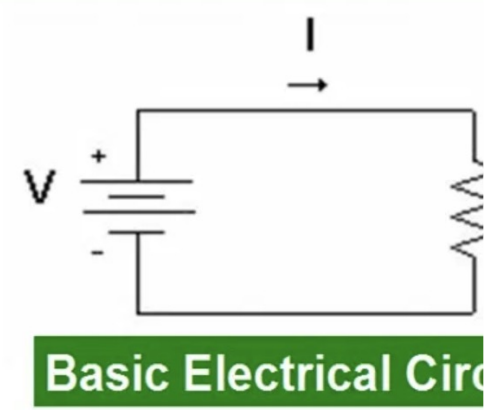
Why machine learning : Data every where !

"We are drowning in information and starving for knowledge." – John Naisbitt.

- This deluge of data calls for automated methods of data analysis, which is what machine learning provides.
- Defined as a "set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" - Kevin Murphy (Machine Learning : A Probabilistic Perspective)

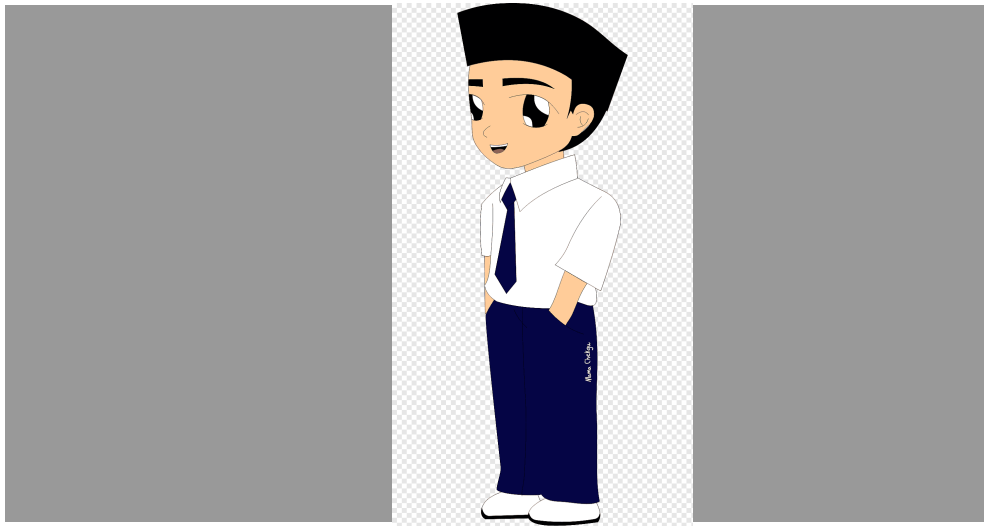
Whats Machine Learning ?

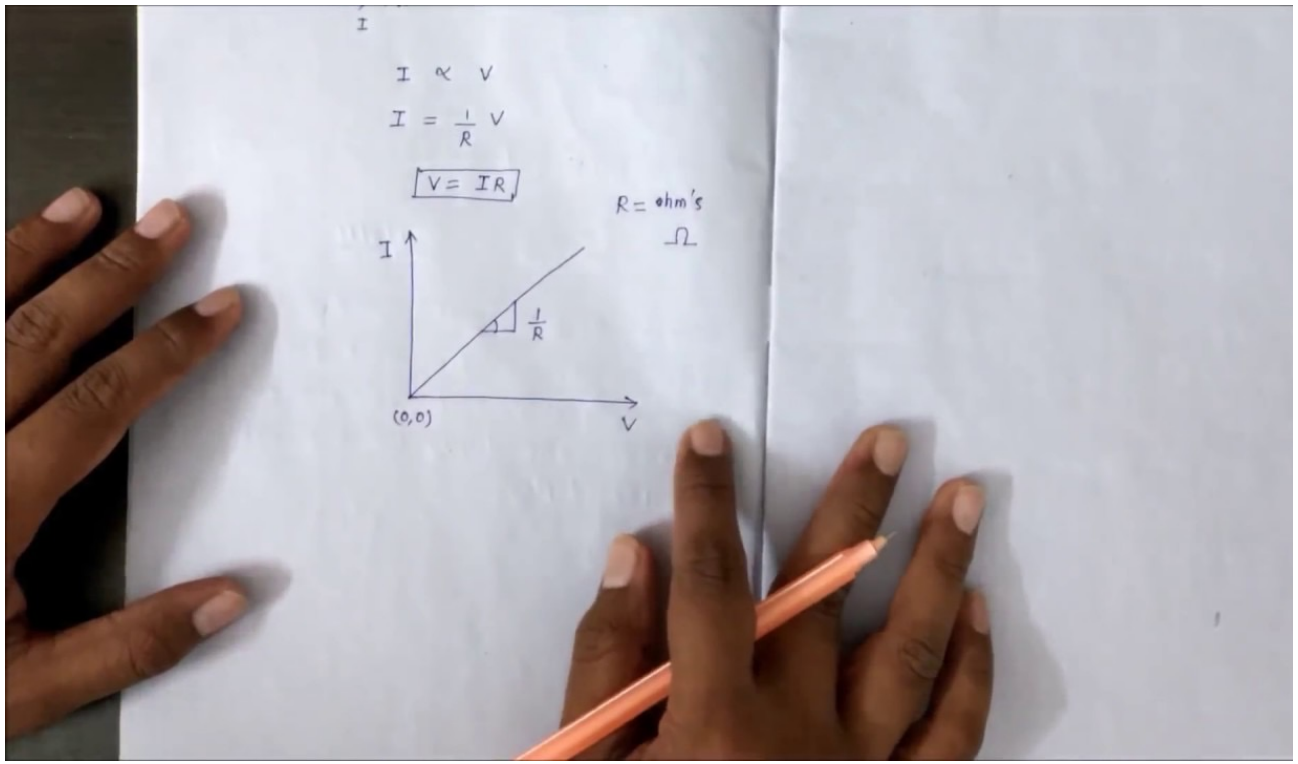
Georg Simon Ohm



Basic Electrical Circ

$$I = \frac{V}{R}$$





Estimate Resistance

- Connect the variable voltage supply to both the ends of the rheostat.Connect the ammeter in series of the rheostat.Connect the voltmeter in parallel of the rheostat.Start measuring the voltage and current as you move the rheostat moving hand from minimum position to the maximum position in the steps of constant increase in current.

Potential difference V (in volt)	0.5	1.0	1.5	2.0	2.5
Current I (in ampere)	0.2	0.4	0.6	0.8	1.0

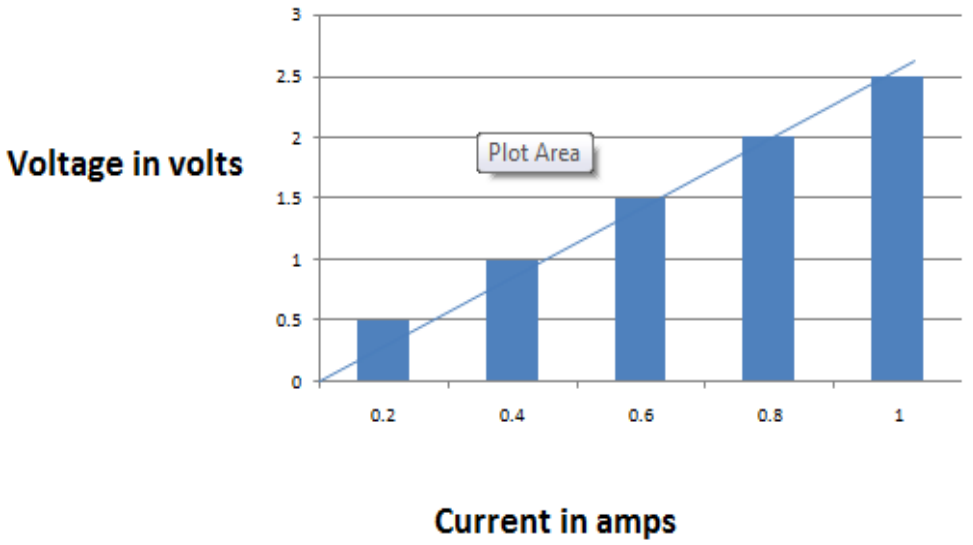
Estimating Resistance is machine learning !

Potential difference V (in volt)	0.5	1.0	1.5	2.0	2.5
Current I (in ampere)	0.2	0.4	0.6	0.8	1.0

- A currentvoltage characteristic or IV curve (currentvoltage curve) is a relationship, typically represented as a chart or graph, between the electric current through a circuit, device, or material, and the corresponding voltage, or potential difference across it. In the graph, the voltage is plotted along the y-axis and the current is plotted along the x-axis.

V = R I

resistance is 2.5 ohms!



Estimating Resistance is machine learning !

Potential difference V (in volt)	0.5	1.0	1.5	2.0	2.5
Current I (in ampere)	0.2	0.4	0.6	0.8	1.0

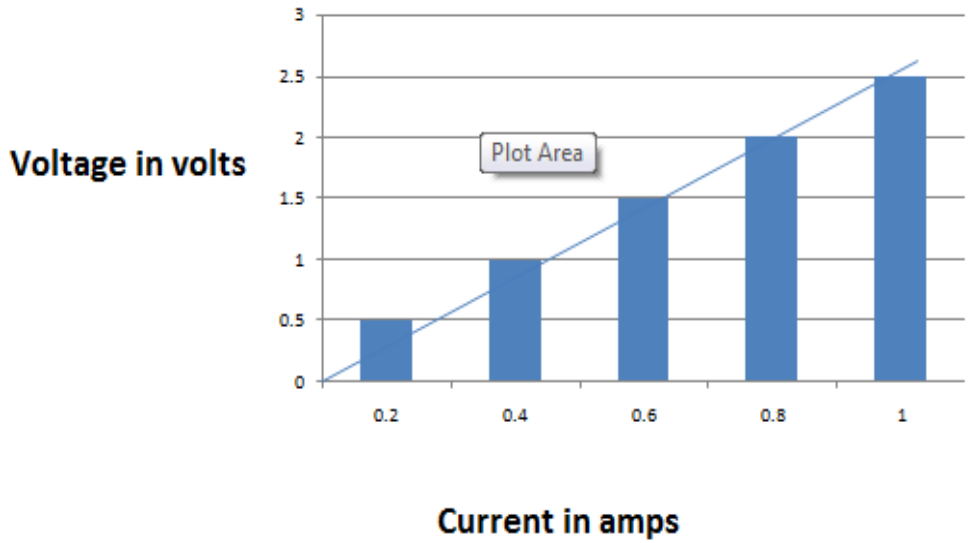
- A currentvoltage characteristic or IV curve (currentvoltage curve) is a relationship, typically represented as a chart or graph, between the electric current through a circuit, device, or material, and the corresponding voltage, or potential difference across it. In the graph, the voltage is plotted along the y-axis and the current is plotted along the x-axis.

V = R I

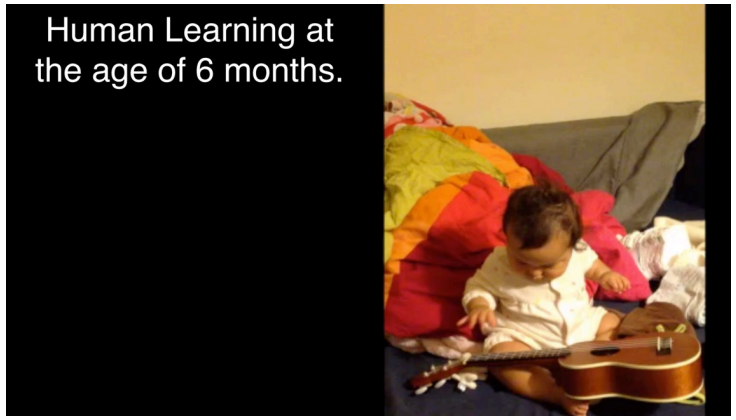
Output

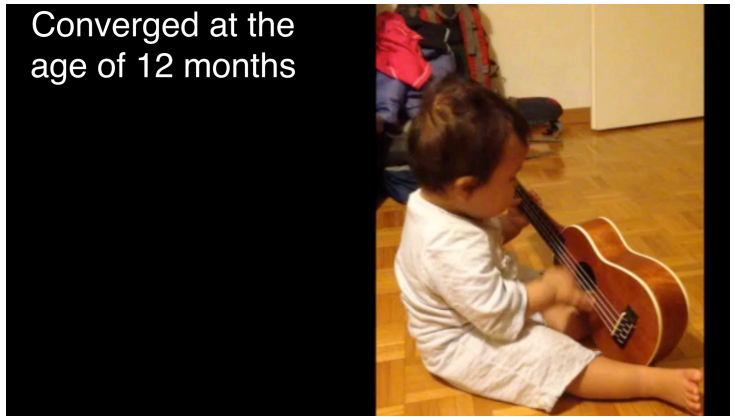
Input

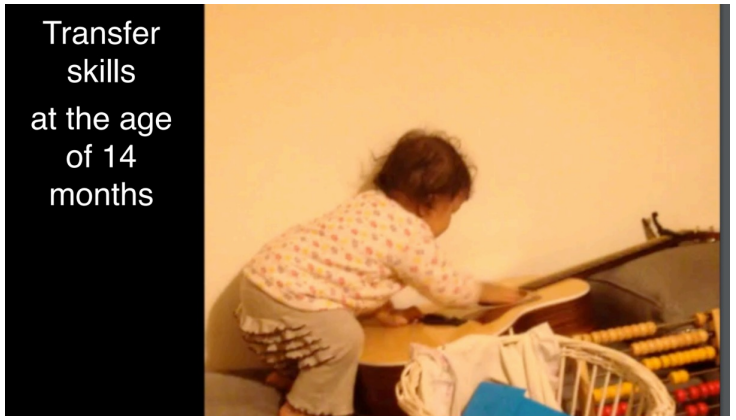
y = m x



What’s machine learning ?







- Algorithms which let machines learn like humans from observations !
- To discover the fundamental principles of learning from data and use them to develop algorithms that can learn like living beings !
- Programming computers to optimize a performance criterion using example data or past experience(Ethem Alpay din, Machine Learning, 2010)
- How do we create computer programs that improve with experience? (Tom Mitchel)

Machine learning

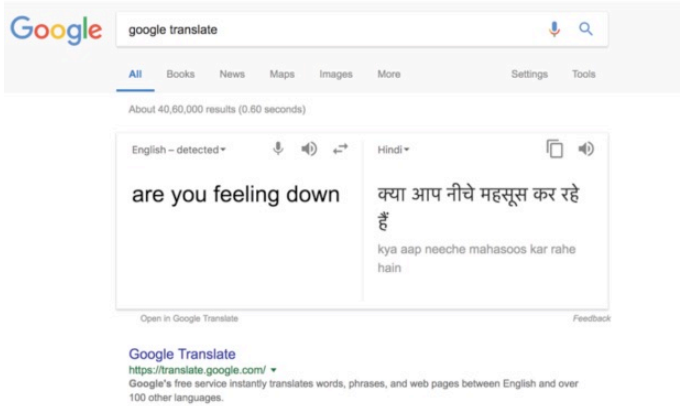
Traditional Programming

Machine Learning

Machine Learning, Data Mining, Knowledge Discovery, Artificial Intelligence, Statistical Learning, Pattern Recognition, Computational Learning

ML based applications

-Machine learning has become prominent approach to solve problems in AI domains like computer vision, language and speech processing



Face detection

Viola-Jones method.

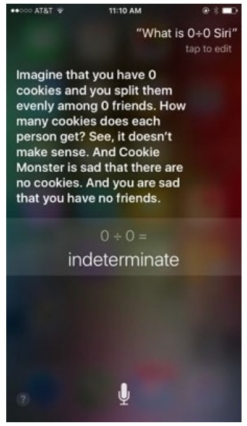
Jeopardy! (2011): Humans vs. IBM Watson

By Rosemaryetoufee (Own work), via Wikimedia Commons

Natural Language Understanding and information extraction!

Speech recognition

- Virtual assistants: Siri (Apple), Echo (Amazon), Google Now, Cor- tana (Microsoft).
- "They" helps get things done: send an email, make an appointment, find a restaurant, tell you the weather and more.
- Leverage deep neural networks to handle **speech recognition** and **natural language understanding**.



ML based applications

- Machine learning has become prominent approach to solve problems in AI domains like computer vision, language and speech processing
- Early approaches to AI was based on logic but applications has to face a lot of uncertain situations and has to perform well on unseen situations.
- Machine learning focused on developing algorithms which could perform well on future unseen data (generalization performance) which differentiates it from statistics

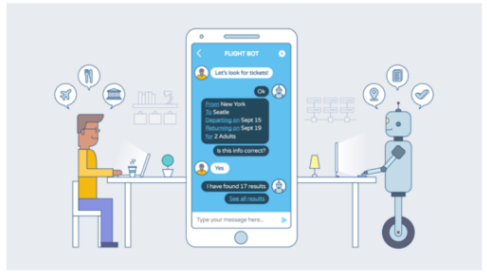
Go (2016): Lee Sedol versus Google AlphaGo

AlphaGo

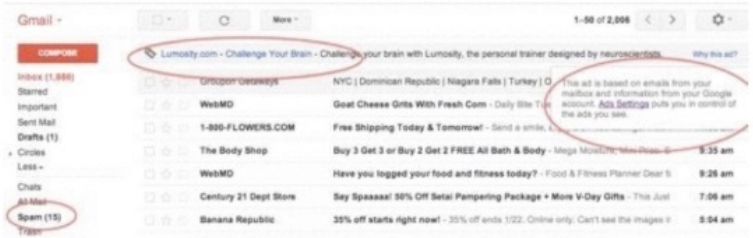
(Left) By LG Electronics, via Wikimedia Commons (Right) By Google DeepMind, via Wikimedia Commons

Deep Learning, reinforcement learning, and search algorithms!

Chat bots



https://botkit.ai



Autonomous driving

Machine learning is interdisciplinary

- Science (Astronomy, neuroscience, medical imaging, bio-informatics)
- Environment (energy, climate, weather, resources)
- Retail (Intelligent stock control, demographic store placement)
- Manufacturing (Intelligent control, automated monitoring, detection methods)
- Security (Intelligent smoke alarms, fraud detection)
- Marketing (promotions, ...)
- Management (Scheduling, timetabling)
- Finance (credit scoring, risk analysis...)
- Web data (information retrieval, information extraction, ...)

Overview of Machine learning

Supervised learning

- Predict an output y when given an input **x**
- For categorical y : classification.
- For real-valued y : regression.

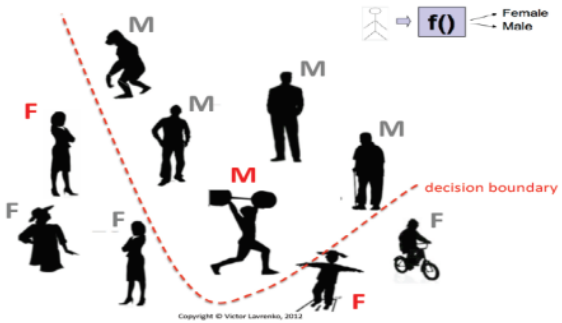
Unsupervised learning

- Create an internal representation of the input, e.g. clustering, dimensionality reduction
- This is important in machine learning as getting labels is often difficult and expensive

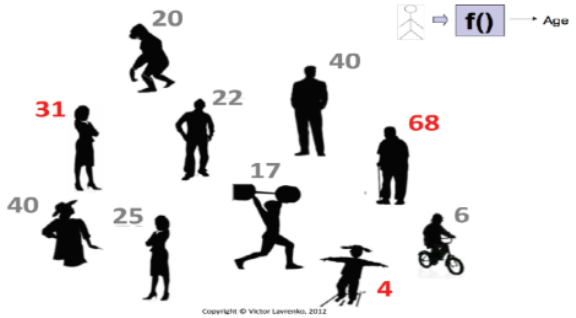
Other settings of ML

- Reinforcement learning (learning from “rewards”)
- Semi-supervised learning (combines supervised + unsupervised)
- Active learning, online learning, Transfer learning, multi-task learning, Structured prediction

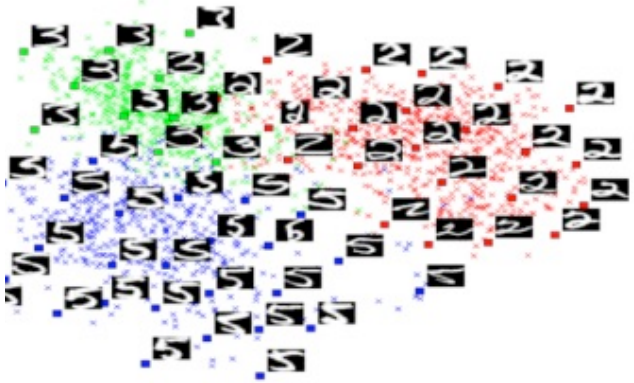
Classification (Supervised Learning)



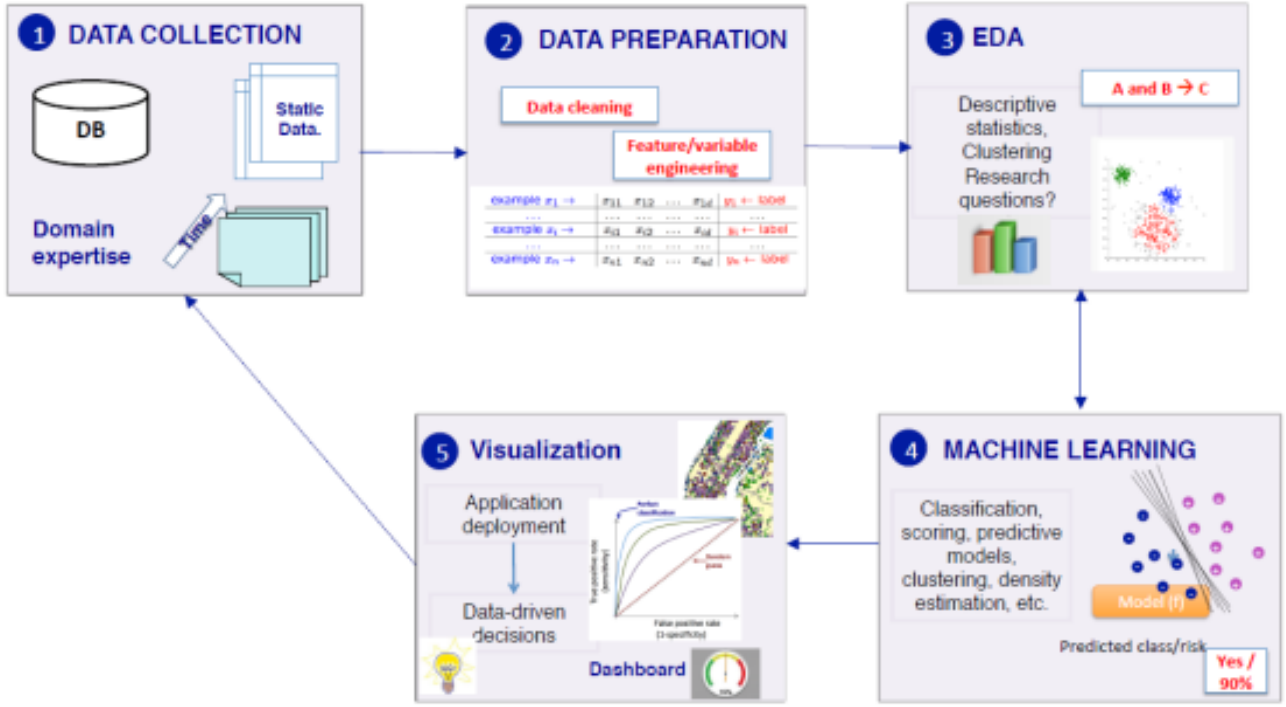
Regression (Supervised Learning)



Clustering (Unsupervised Learning)



ML in practice



- Understanding domain, prior knowledge, and goals
- Data integration, selection, cleaning, pre-processing, etc.
- Learning models
- Interpreting results
- Consolidating and deploying discovered knowledge
- Loop

Supervised learning (classification)

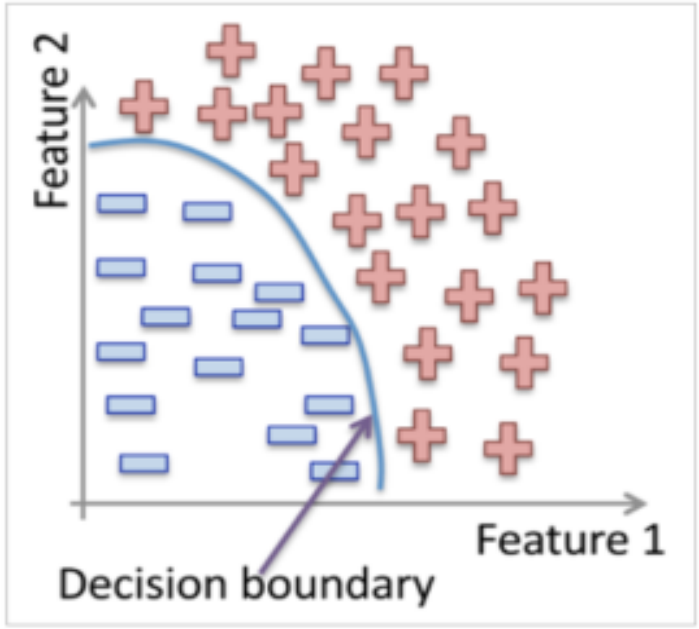
Input $\mathbf{x} \in \mathcal{R}^d$ and output y a label. Learn a function

$$f : \mathbf{x} \rightarrow y$$

Input

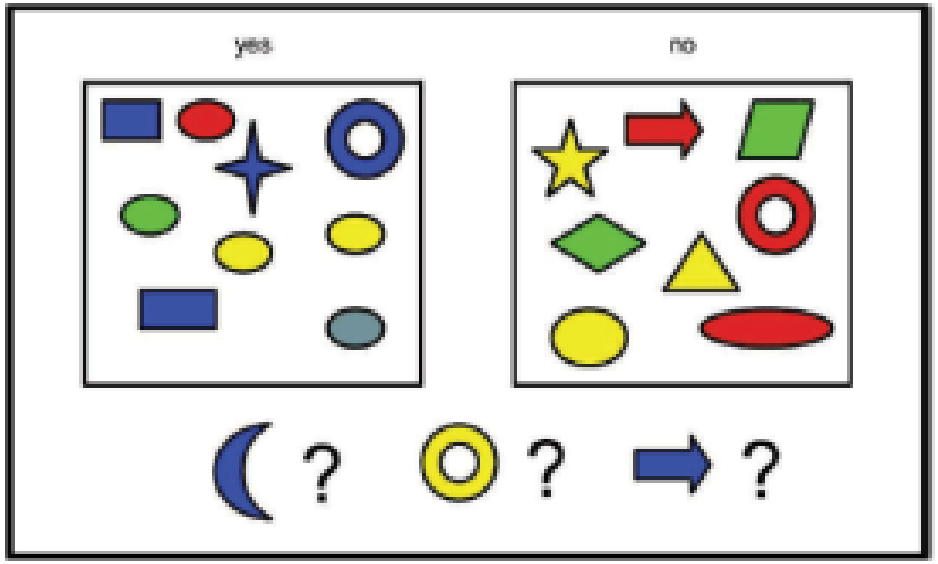
Output

example $x_1 \rightarrow$	x_{11}	x_{12}	. . .	x_{1d}	$y_1 \leftarrow$ label
. . .	. . .	. . .	. . .	. . .	. . .
example $x_i \rightarrow$	x_{i1}	x_{i2}	. . .	x_{id}	$y_i \leftarrow$ label
. . .	. . .	. . .	. . .	. . .	. . .
example $x_n \rightarrow$	x_{n1}	x_{n2}	. . .	x_{nd}	$y_n \leftarrow$ label



fruit	length	width	weight	label
fruit 1	165	38	172	Banana
fruit 2	218	39	230	Banana
fruit 3	76	80	145	Orange
fruit 4	145	35	150	Banana
fruit 5	90	88	160	Orange
...				
fruit n	...	...	...	...

Supervised learning (Classification)



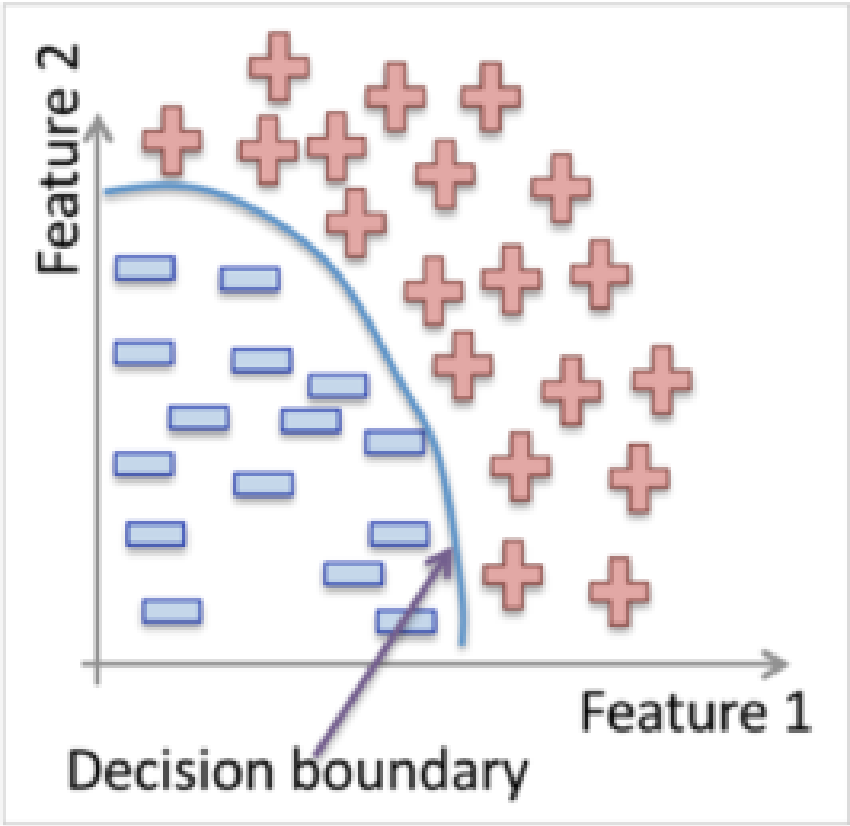
D features (attributes)

N cases

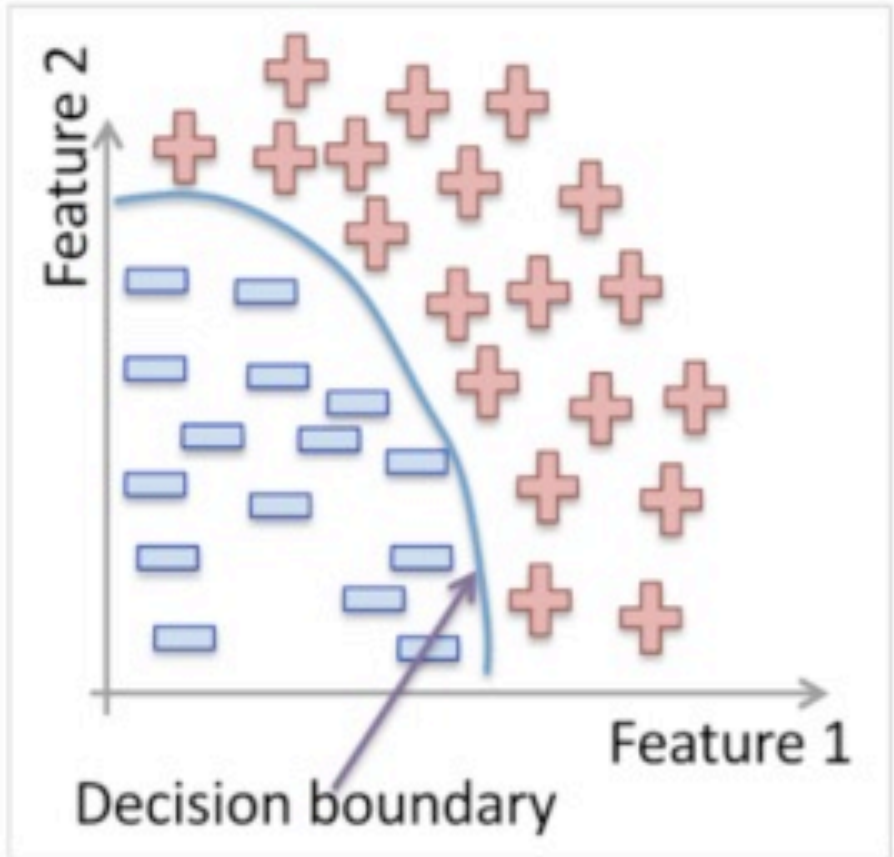
Color	Shape	Size (cm)
Blue	Square	10
Red	Ellipse	2.4
Red	Ellipse	20.7

Label
1
1
0

Supervised learning



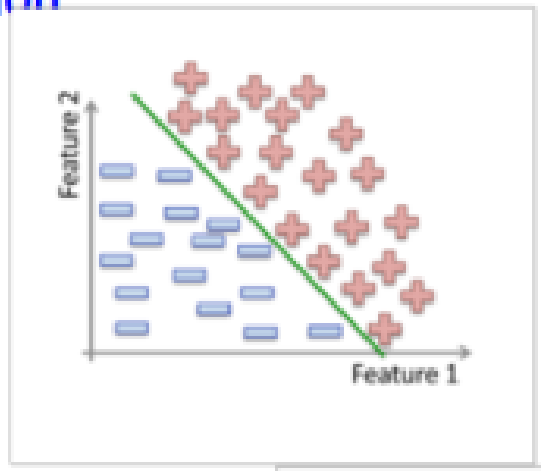
Supervised learning

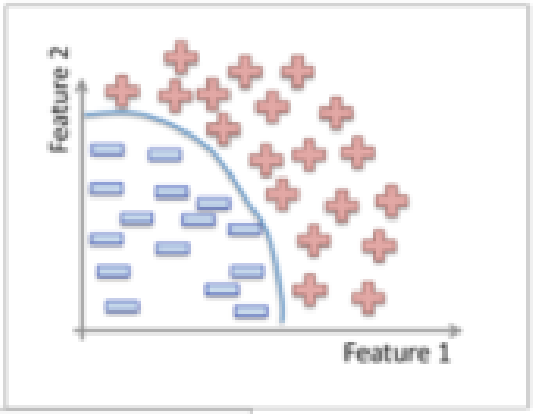


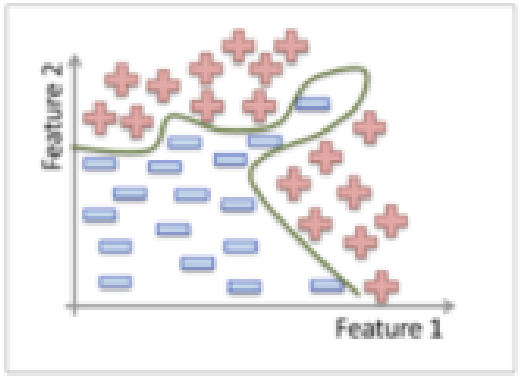
Methods: Support Vector Machines, neural networks, decision trees, K-nearest neighbors, naive Bayes, etc.

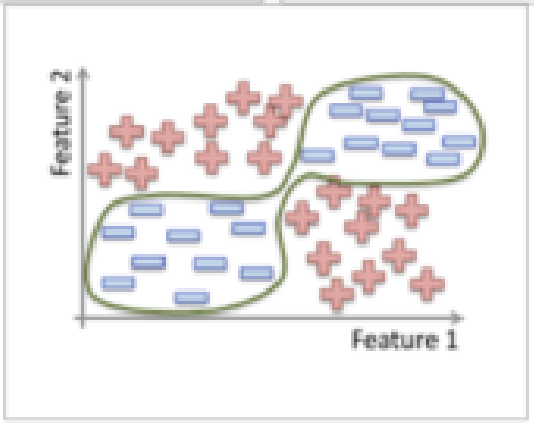
Supervised learning

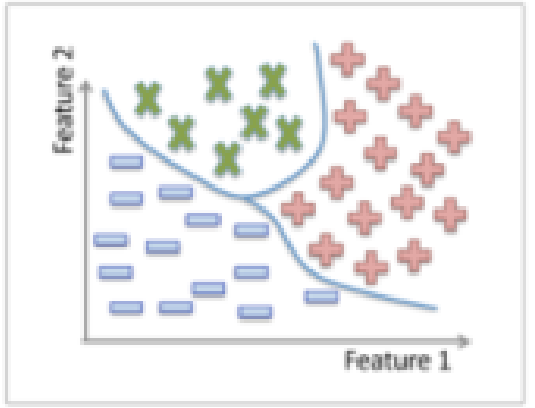
Classification:



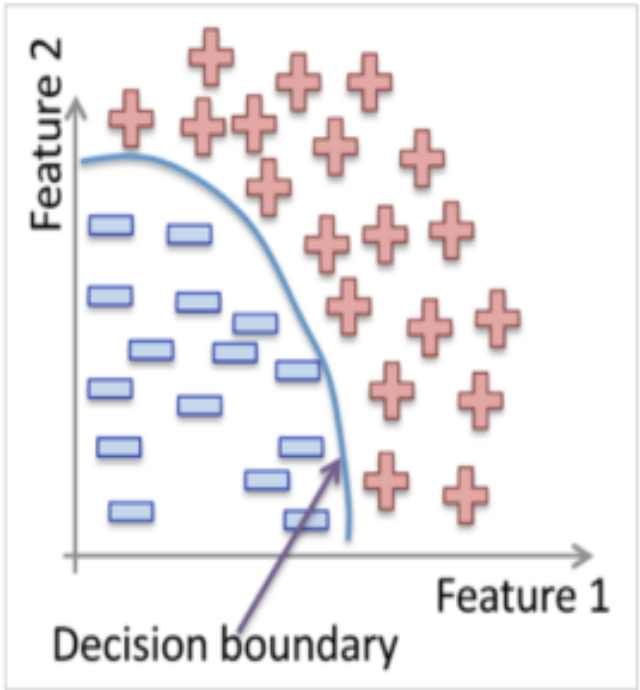




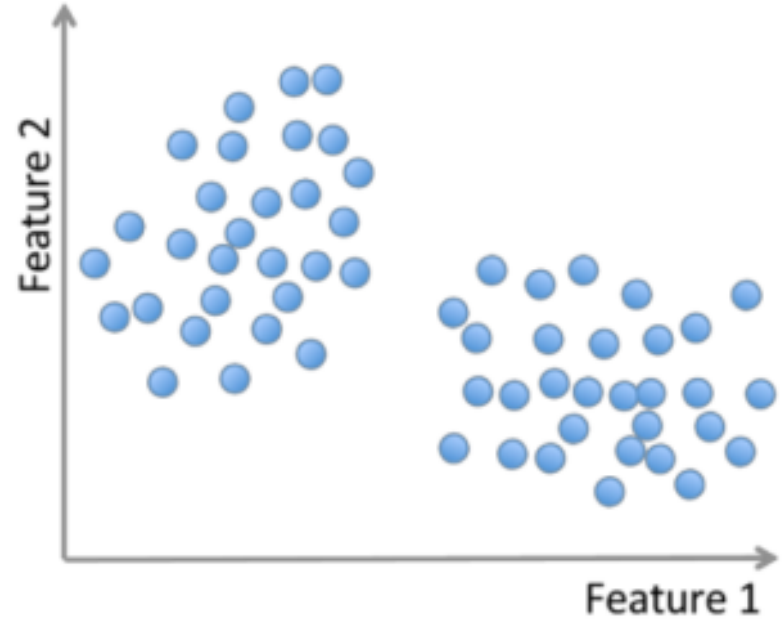




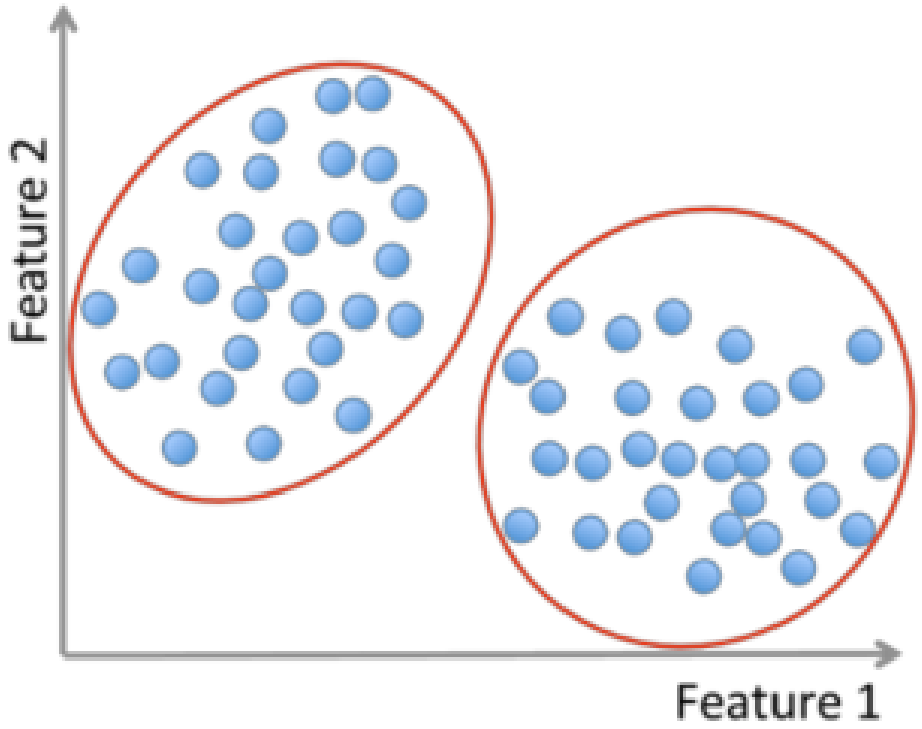
Supervised learning



Unsupervised learning



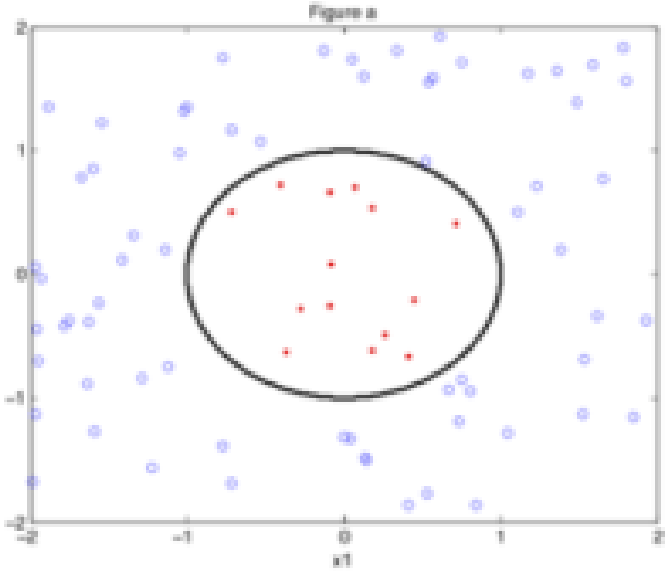
Unsupervised learning

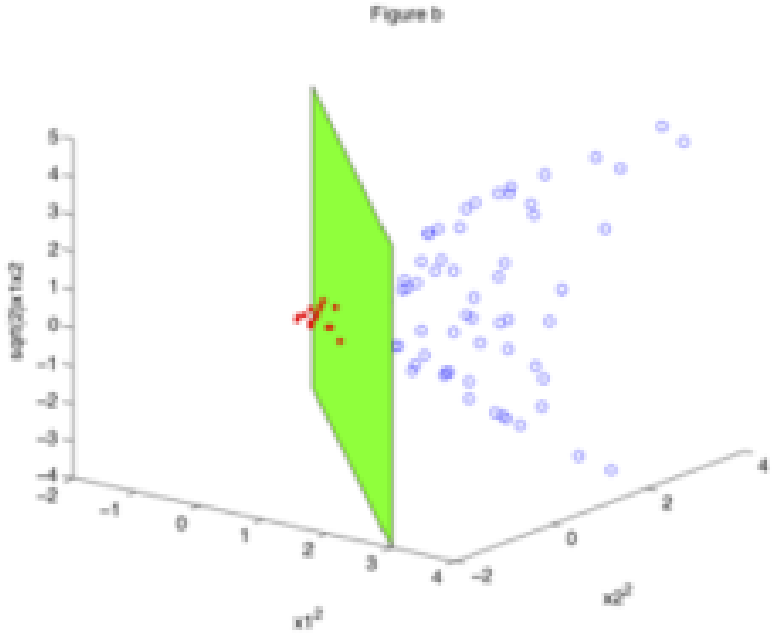


Methods: K-means, gaussian mixtures, hierarchical clustering, spectral clustering, etc.

Supervised learning

Non linear classification





K-nearest neighbors

- Not every ML method builds a model!
- Our first ML method: KNN.
- Main idea: Uses the **similarity** between examples.
- Assumption: Two similar examples should have same labels.
- Assumes all examples (instances) are points in the d dimensional space R^d.

K-nearest neighbors

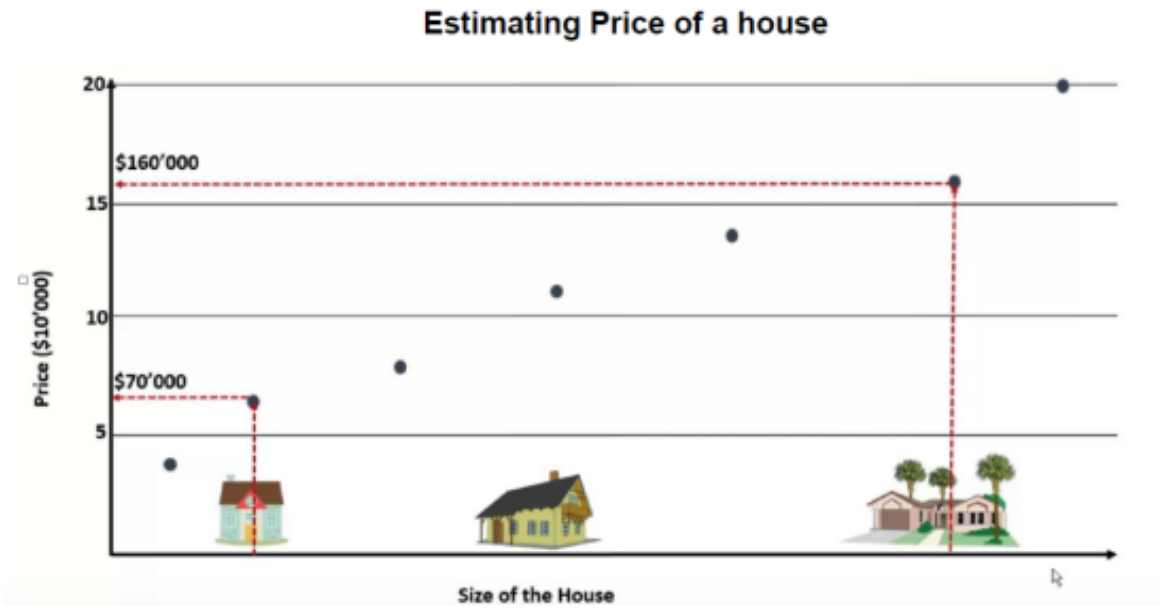
3-NN. Credit: Introduction to Statistical Learning.

Supervised learning (Regression)

Input $\mathbf{x} \in \mathcal{R}^d$ and output y a real value. Learn a function

$$f : \mathbf{x} \rightarrow y$$

Output



Goal is to learn a function which maps inputs to outputs so that it will predict well on future data points – Generalization performance

Input

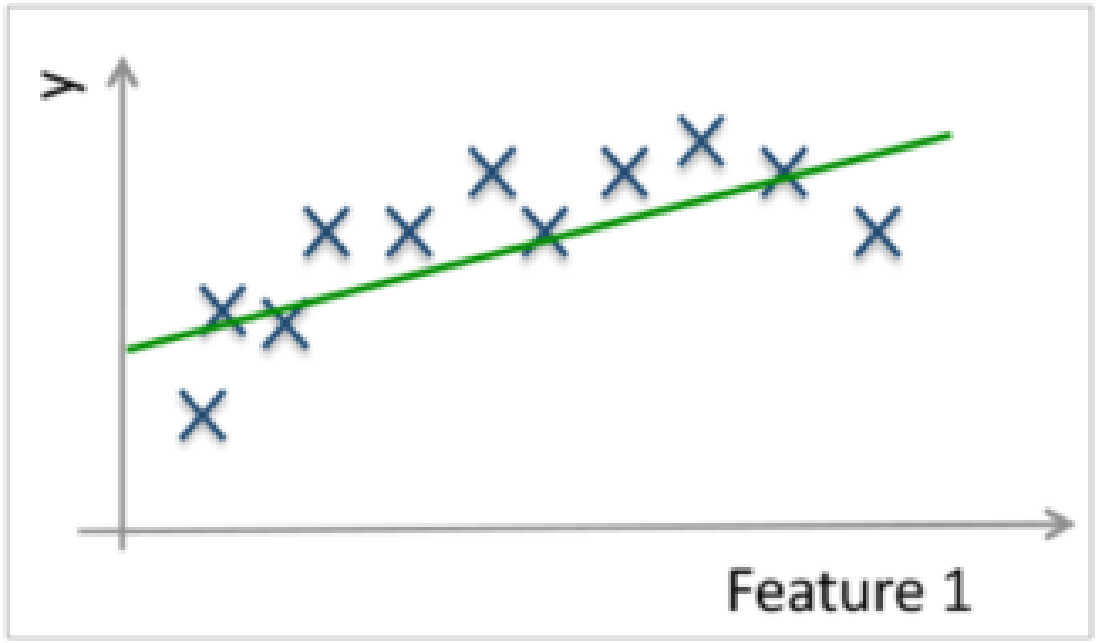
Supervised learning

Regression:

Example: Income in function of age, weight of the fruit in function of its length.

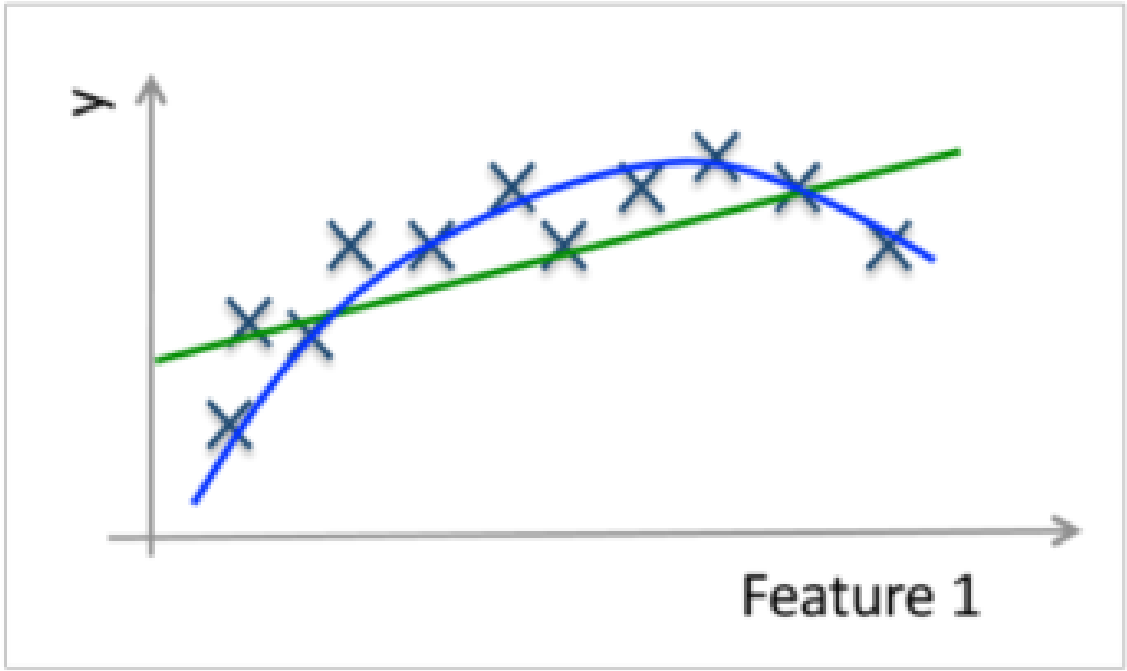
Supervised learning

Regression:



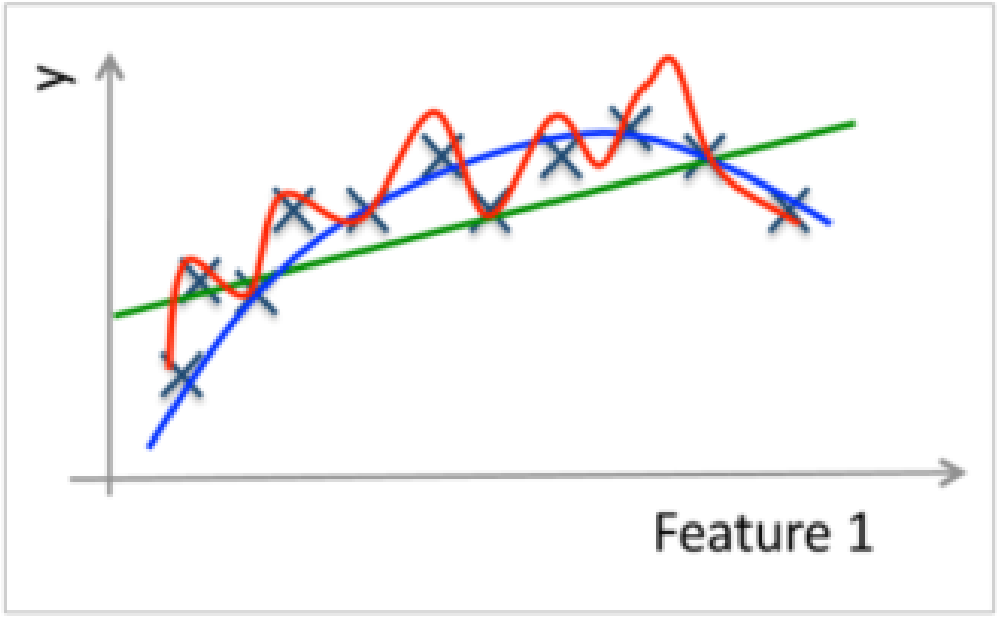
Supervised learning

Regression:



Supervised learning

Regression:

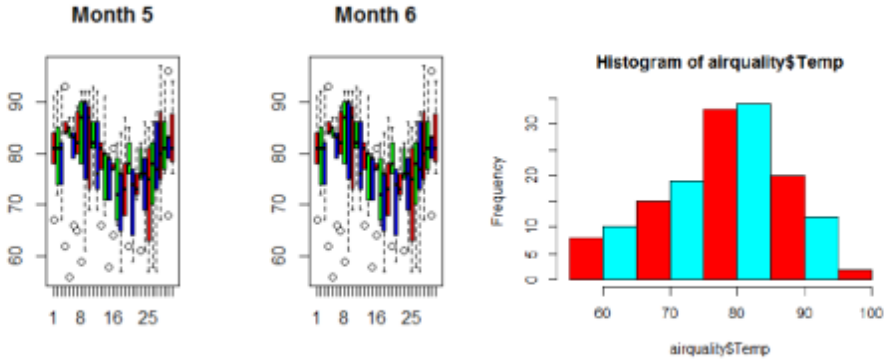


Linear regression

- Temp=w1.Solar.R +w2.Ozone + w3.Wind + error.
- Temperature of house depends on ozone, wind and solar radiations
- linear regression helps to discover relation between dependent and independent variables

Airquality data

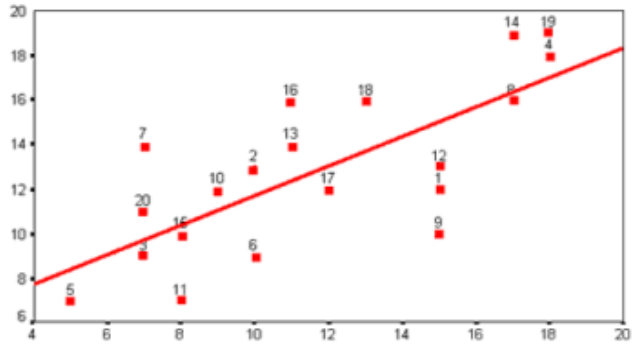
	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10



https://www.edvancer.in/step-step-guide-to-execute-linear-regression-r/

Linear Regression

- Observations need not lie on a line
  - Observations are not generated by a linear line
  - Observations are noisy, due to measurement errors



Linear regression

- Learn a function which maps input to output f : X -> Y
- Consider a Linear function

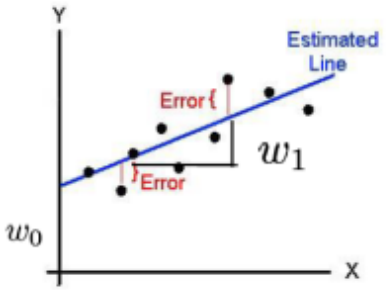
Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

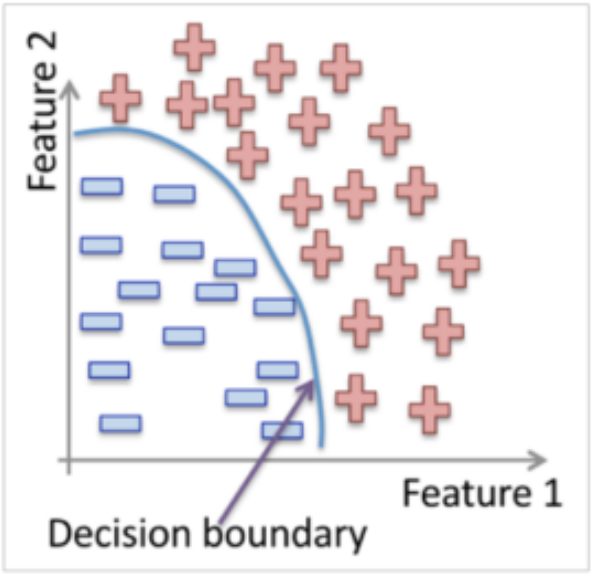
$$\hat{Y}_i = w_0 + w_1 X_i$$



Learn the function which passes through as many points as possible : Minimize the **Least Squares Error**

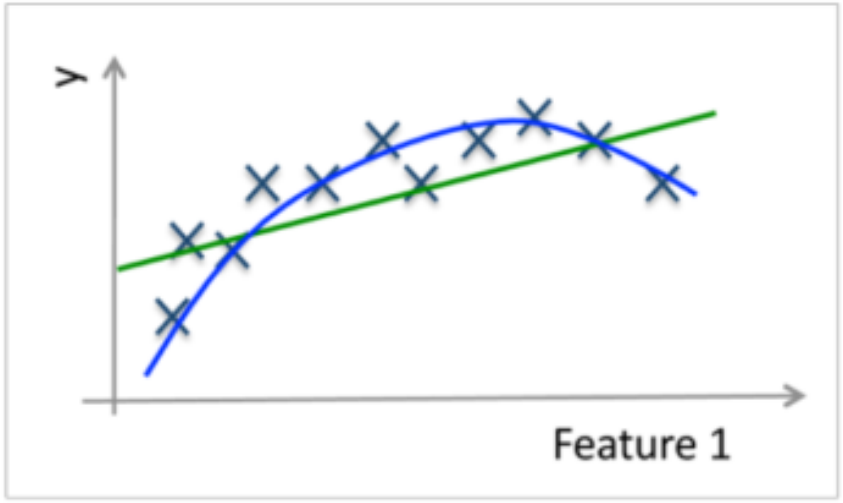
$$E(w) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad = \frac{1}{2}\sum_{i=1}^{N}(y_i - X_i^\top w)^2$$

Machine learning algorithms

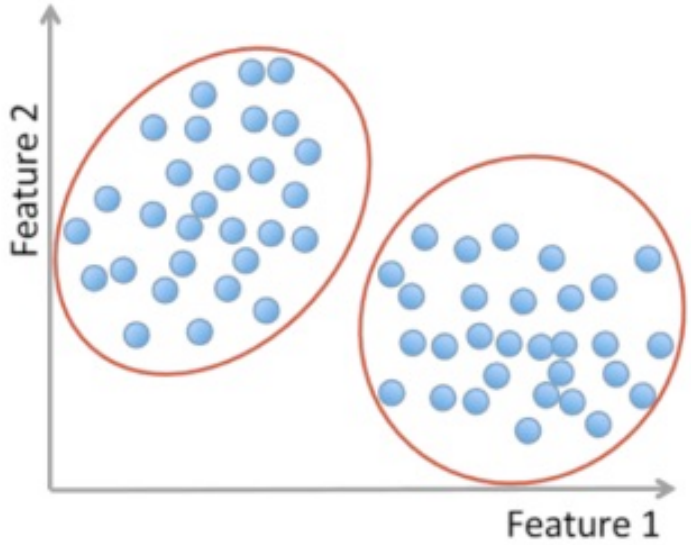


Methods: Support Vector Machines, neural networks, decision trees, K-nearest neighbors, naive Bayes, etc.

Regression:

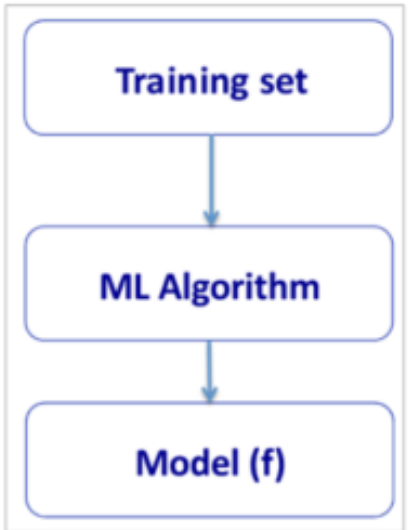


Method : Linear regression, support vector regression, gaussian process regression

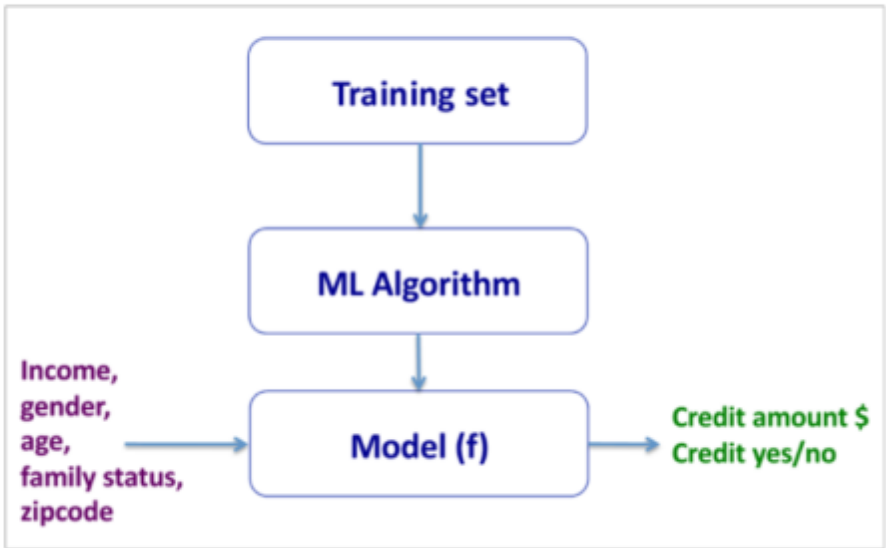


Methods: K-means, gaussian mixtures, hierarchical clustering, spectral clustering, etc.

Training and Testing

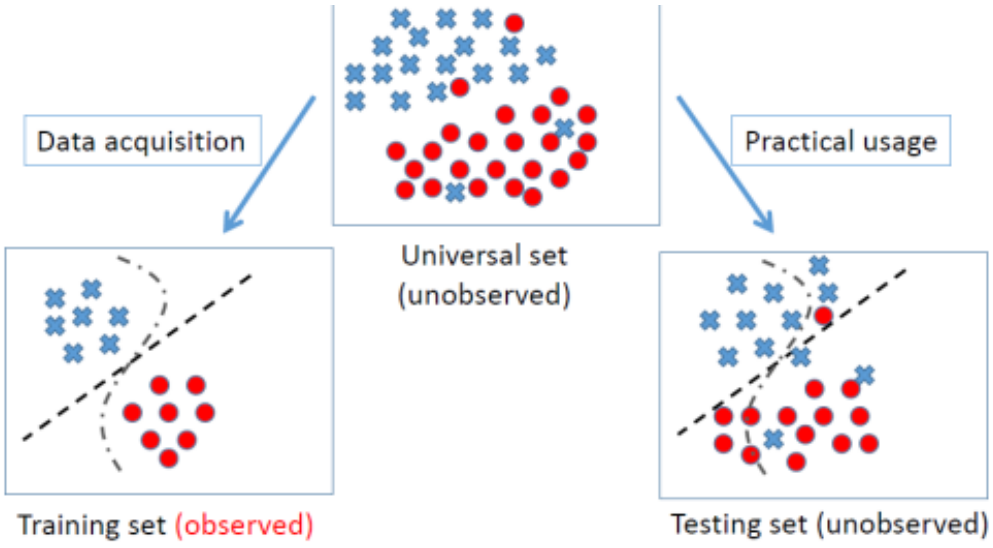


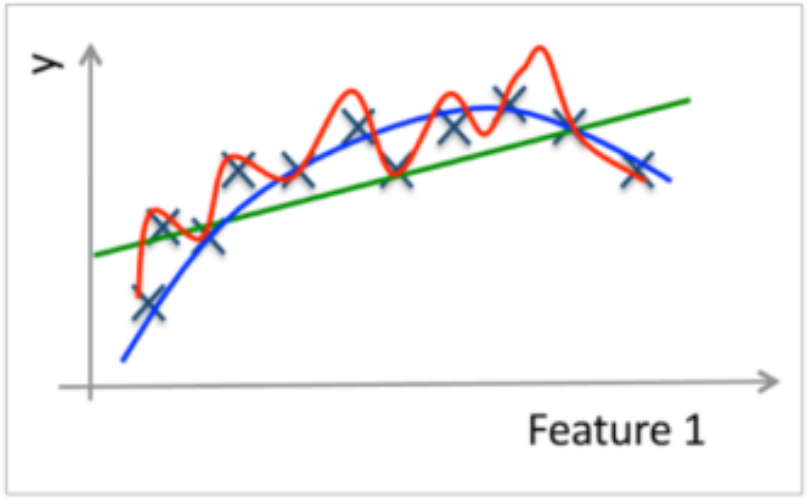
Training and Testing



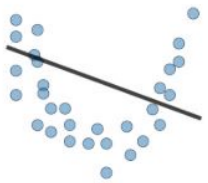
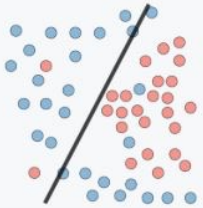
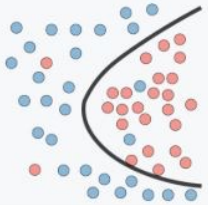
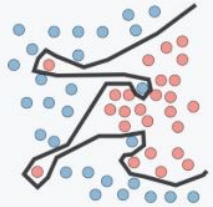
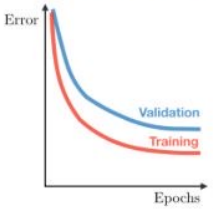
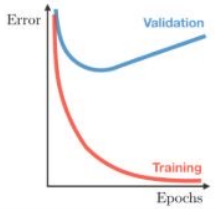
Training and Testing ML models

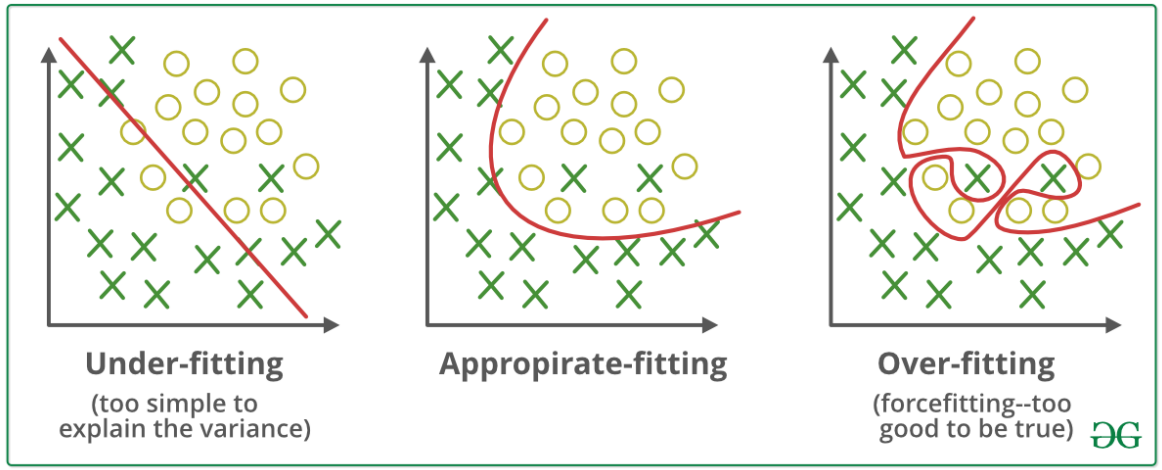
ML models should be able to predict well (generalization ability)
Model with minimum error on train data need not be the correct model

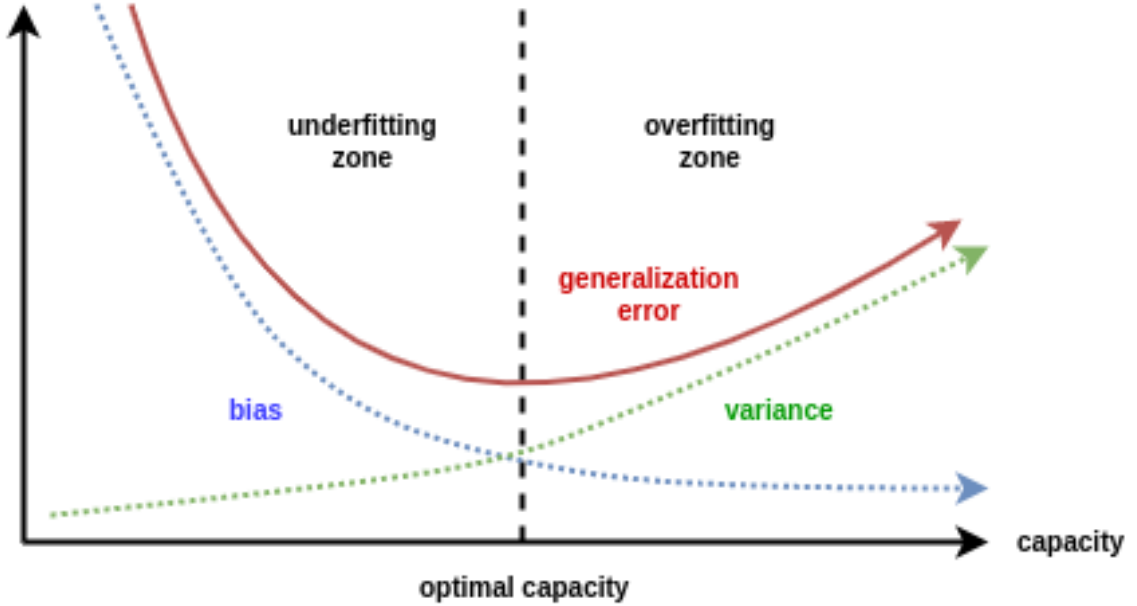




Generalization ability and overfitting

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data





Train, Validation and Test

Example: Split the data randomly into 60% for training, 20% for validation and 20% for testing.

Train, Validation and Test

Example: Split the data randomly into 60% for training, 20% for validation and 20% for testing.

1. Training set is a set of examples used for learning a model (e.g., a classification model).
2. Validation set is a set of examples that cannot be used for learning the model but can help tune model parameters (e.g., selecting K in K-NN). Validation helps control overfitting.
3. Test set is used to assess the performance of the final model and provide an estimation of the test error.

Note: Never use the test set in any way to further tune the parameters or revise the model.

Evaluation metrics

Classification

		Actual Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct
Sensitivity (Recall)	TP / (TP + FN)	The percentage of positive cases that were predicted as positive
Specificity	TN / (TN + FP)	The percentage of negative cases that were predicted as negative

Regression

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Machine learning

- Main references:
- Mitchell, Tom. Machine Learning. New York, NY: McGraw-Hill, 1997
- Alpaydin, Ethem Introduction to Machine Learning. MIT Press, 2014.
- Bishop, Christopher M. Pattern Recognition and Machine Learning. Springer, 2006.
- Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. Cambridge, Mass. MIT Press
- Hastie, T., R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York, NY: Springer, 2001

ML Resources

- MOOCs
  - Coursera, EdX, Udacity
- Conferences/Journals
  - JMLR, Machine Learning, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, Annals of Statistics
  - ICML, NIPS, KDD, IJCAI, AAAI, ICDM

ML Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- Statlib: http://lib.stat.cmu.edu/
- Kaggle
- Many more…