KIOS
Research and Innovation Center of Excellence

MSc on Intelligent Critical Infrastructure Systems

Machine Learning
Lecture 1

Christos Kyrkou
Research Lecturer
KIOS Research and Innovation Center of Excellence
University of Cyprus

Imperial College London

University of Cyprus

funded by:

Course Outline

- **Main instructor:** Prof. Marios Polycarpou
  - Professor
  - mpolycar@ucy.ac.cy

- **Co-instructor:** Dr. Christos Kyrkou
  - Research Lecturer
  - kyrkou.christos@ucy.ac.cy

- **Co-instructor:** Dr. Kleanthis Malialis
  - Research Associate
  - malialis.kleanthis@ucy.ac.cy

- **Teaching Assistant:** Rafaella Elia
  - PhD Candidate
  - elia.rafaella@ucy.ac.cy

- **Course Website:**

  https://www.msccis.ucy.ac.cy/ece-805-course/

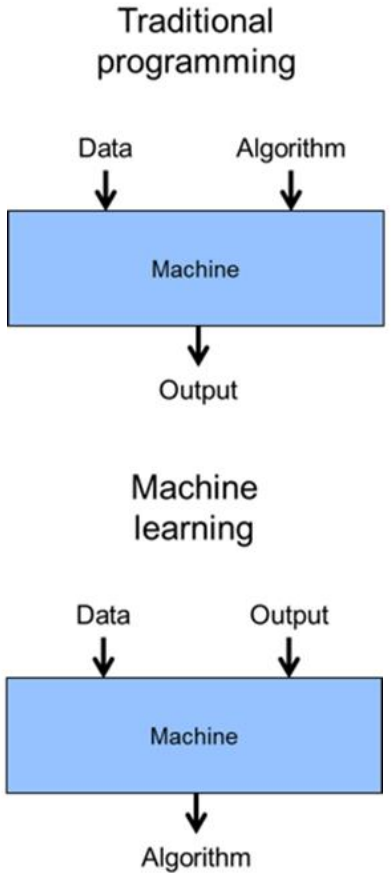
- **Course Syllabus:**
  - Lectures, Tutorials
  - Instructor, Teaching Assistants
  - Course Objectives
  - Course Outline
  - References
  - Prerequisites
  - Course Evaluation
  - Academic Honesty

Course Outline

- **Week 1**
  - Introduction & Preliminaries
- **Week 2**
  - Linear regression
  - Logistic regression, Regularisation, SVMs
- **Week 3**
  - Neural Networks and Deep Learning
- **Week 4**
  - Feature engineering and Evaluation
  - Online learning
- **Week 5**
  - Unsupervised Learning
- **Week 6**
  - Reinforcement learning
- **Week 7**
  - Monitoring and Control

What is Machine Learning?

- Early definition of machine learning:
  - "Field of study that gives computers the ability to learn without being explicitly programmed."
    - Arthur Samuel (1959): Computer pioneer who wrote first self-learning program, which played checkers – learned from "experience"
- ML Paradigm:
  - Specify some **goal** on the behavior of a desirable program
  - Write a rough skeleton of the code that identifies a subset of program space to search
  - Use the computational resources at our disposal to **search this space for a program that works**.



https://karpathy.medium.com/software-2-0-a64152b37c35

A very brief history of Machine Learning

- 1950s:
  - Samuel's checker player
- 1960s:
  - Neural networks: Perceptron
  - Pattern recognition
  - Minsky and Papert prove limitations of Perceptron
- 1970s:
  - Symbolic concept induction
  - Expert systems and the knowledge acquisition bottleneck
- 1980s:
  - Advanced decision tree and rule learning
  - Resurgence of neural networks (connectionism, backpropagation)
  - Learning and planning and problem solving
- 1990s:
  - Data mining
  - Text learning
  - Reinforcement learning (RL)
  - Ensembles: Bagging, Boosting, and Stacking
- 2000s:
  - Support vector machines & kernel methods
  - Sequence Learning
  - E-mail management
  - Personalized assistants that learn
  - Learning in robotics and vision
- 2010s:
  - Deep learning systems
  - Learning for big data
  - Multi-task & lifelong learning
- 2020s-Beyond:
  - Self Supervised Learning
  - Automated Machine Learning
  - Foundation Models /Large Language Models
  - Learning to reason/Understanding how the physical world operates

Motivation for Machine Learning

Cloud Data

Internet Of Things

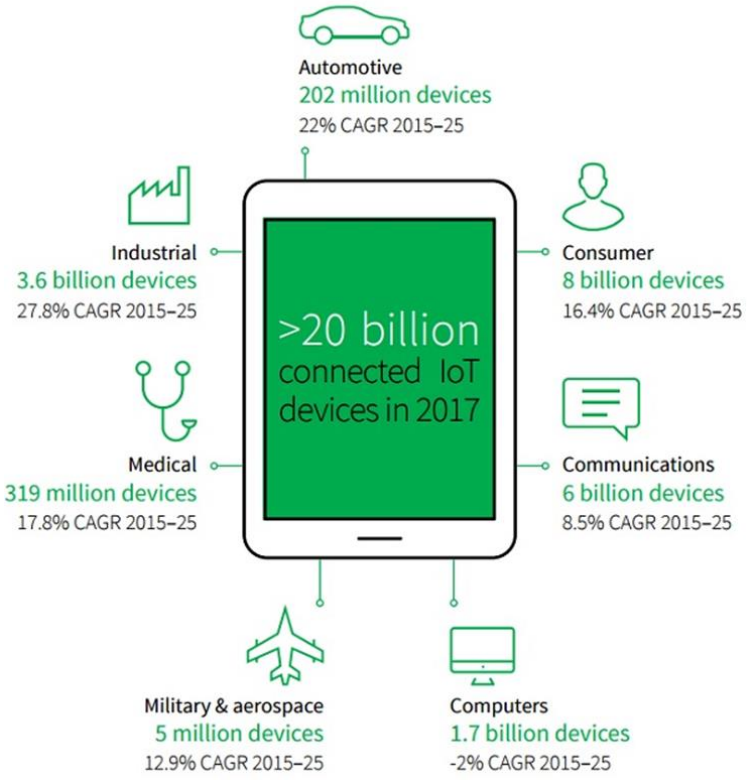
Social Media

"Over *2.5 quintillion bytes* of data are created every single day, and it's only going to grow from there. By 2020, it's estimated that 1.7MB of data will be created every second for every person on earth."

… A lot of data are being generated

Motivation for Machine Learning

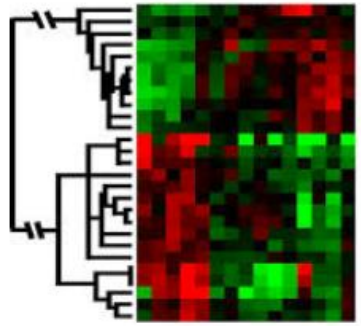
- Sensor technology
  - wealth of sensors
  - new generation sensors
- Information & Communication Technology (ICT)
  - store, process and transmit data collected by sensors
- Internet of Things (IoT)
  - sensor enabled devices connected to the internet and able to communicate with each other
- Big Data
  - sensor technology and ICT enabled the collection of extremely large data sets
  - contribute to the better perception of complex systems



Why use learning?

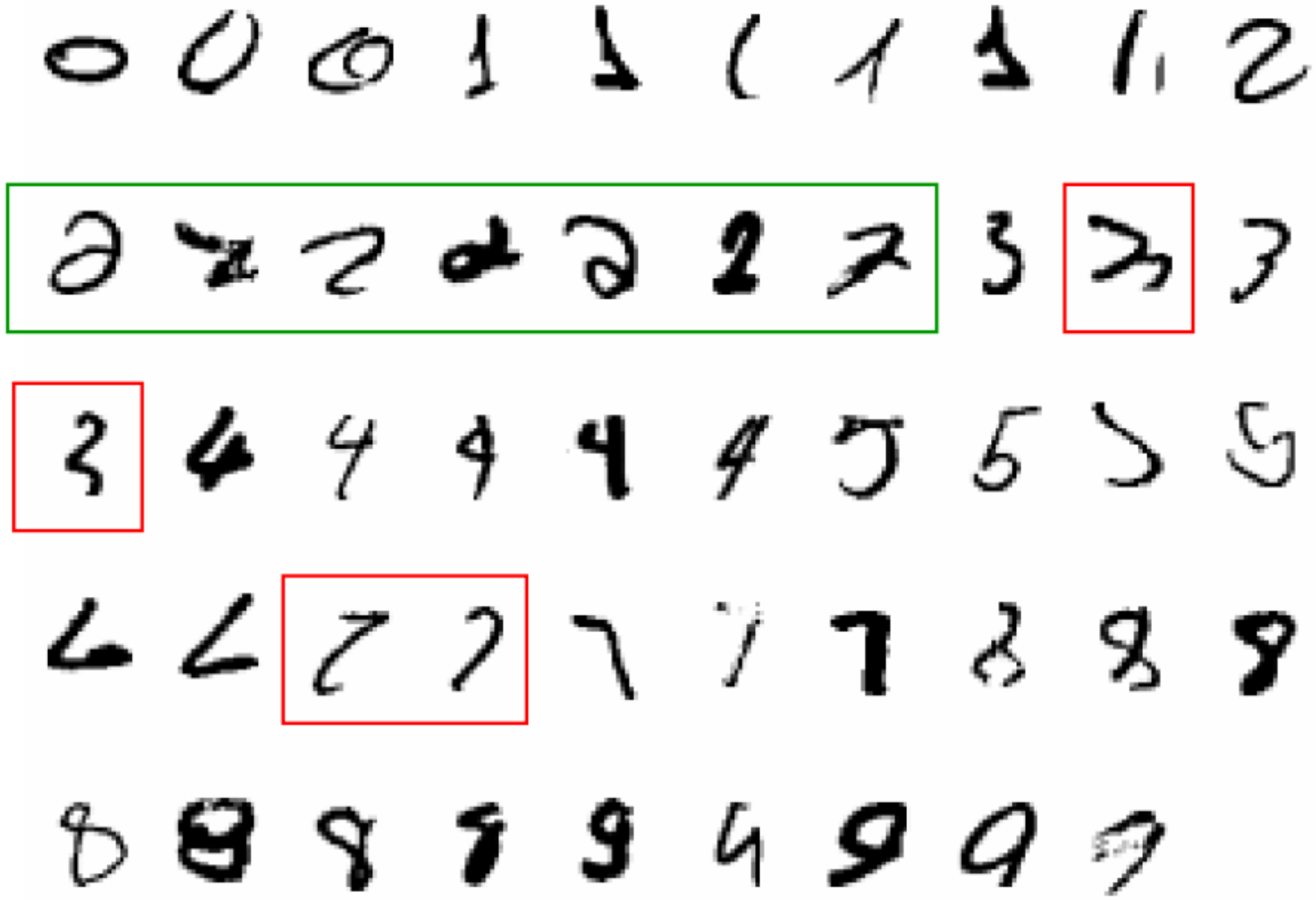
- We typically use machine learning when the function/rule we want the system to apply is unknown to us, and we cannot “think” about it.
  - Human expertise does not exist (navigating on Mars)
  - Humans can’t explain their expertise (speech recognition)
  - Models must be customized (personalized medicine)
  - Models are based on huge amounts of data (genomics)



- Learning isn’t always useful:
  - There is no need to “learn” to calculate the payroll

An example: What makes a 2?



Slide Credit: Geoffrey Hinton

Tasks best solved with machine learning

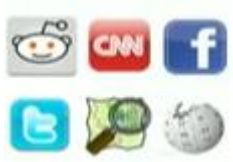
- Recognizing patterns
  - Facial identities or facial expressions
  - Handwritten or spoken words
  - Medical Images
- Generating Patterns
  - Generating images or motion sequences
- Recognizing anomalies
  - Unusual credit card transactions
  - Unusual patterns or sensor readings
- Prediction
  - Future stock prices or currency exchange rates.

Machine **Learning**

- Learning is at the core of:
  - Understanding High Level Cognition
  - Performing knowledge intensive inferences
  - Building adaptive intelligent systems
  - Dealing with messy, real-world data
  - Analytics
- Learning has multiple purposes
  - Knowledge Acquisition
  - Integration of various knowledge sources to ensure robust behavior
  - Adaptation (human, systems)
  - Decision Making (Predictions)

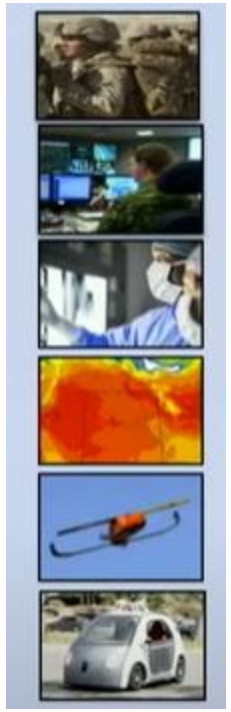
Machine Learning

Sensors



Sources

Modern Computing

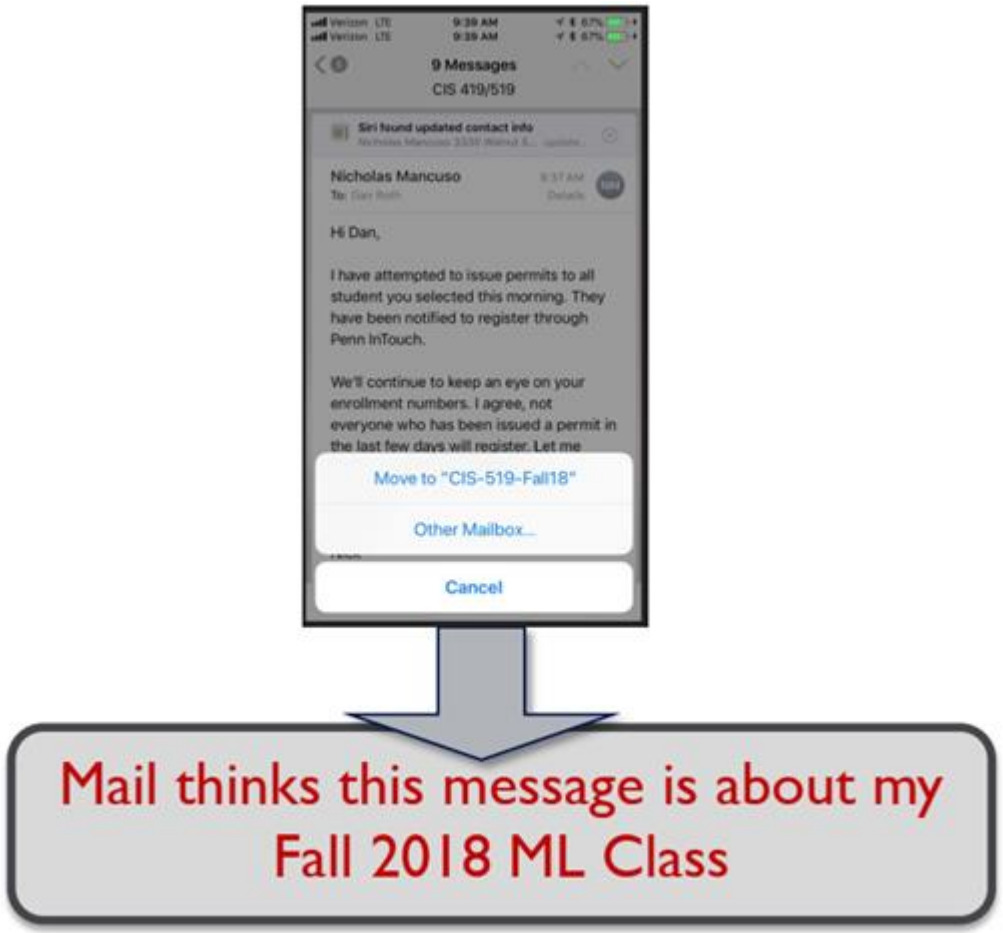


Users/Market

Credit: MIT RES.LL-005

Learning = Generalization

- H.Simon - “Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time”
- Generalization
  - The ability to perform a task in a situation which has never been encountered before



Why Study Machine Learning?

- Computer systems with new capabilities.
  - AI
  - Understand human and biological learning
  - Understanding hidden structures within data
- Time is right.
  - Initial algorithms and theory in place
  - Growing amounts of on-line data
  - Computational power available
  - Necessity: many things we want to do cannot be done by “programming”
  - (Think about all the examples given earlier)

Why Study Machine Learning?

- Learning techniques will be a basis for applications that involve systems that interact with the messy real world
- Learning algorithms are ready for use in applications today
- Prospects for broader future applications make for exciting fundamental research and development opportunities
- Many unresolved issues – Theory and Systems
  - While learning is hot, there are many things we don't know how to do well
- Very active field
- What to learn
  - The fundamental paradigms
  - Some of the important algorithmic ideas
  - Modelling

Why Study Machine Learning?

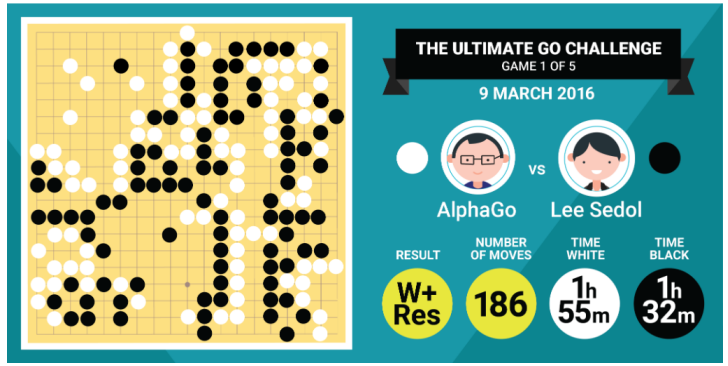
- "A breakthrough in machine learning would be worth ten Microsofts"
  - Bill Gates, Chairman, Microsoft
- "Machine learning is the next Internet"
  - Tony Tether, Former Director, DARPA
- Machine learning is the hot new thing"
  - John Hennessy, President, Stanford
- "Web rankings today are mostly a matter of machine learning"
  - Prabhakar Raghavan, Dir. Research, Yahoo
- "Machine learning is going to result in a real revolution"
  - Greg Papadopoulos, CTO, Sun
- "Machine learning is today's discontinuity"
  - Jerry Yang, CEO, Yahoo

Fourth Industrial Revolution

- What is a technological revolution?
- 1st Industrial Revolution (1760) [hand to machines, automation]
- 2nd Industrial Revolution (1900) [electrification, telegraph, transportation]
- 3rd Industrial Revolution (1960) [digital revolution]
- 4th Industrial Revolution (2000) – term coined in 2015 by Klaus Schwab, executive director of the World Economic Forum.

→ Combines hardware, software, and biology (cyber-physical systems), and emphasizes advances in communication and connectivity. It is expected to be marked by breakthroughs in fields such as **machine learning**, robotics, nanotechnology, biotechnology, the internet of things (IoT), wireless technologies (5G), 3D printing and fully autonomous vehicles.

Machine Learning is Everywhere?



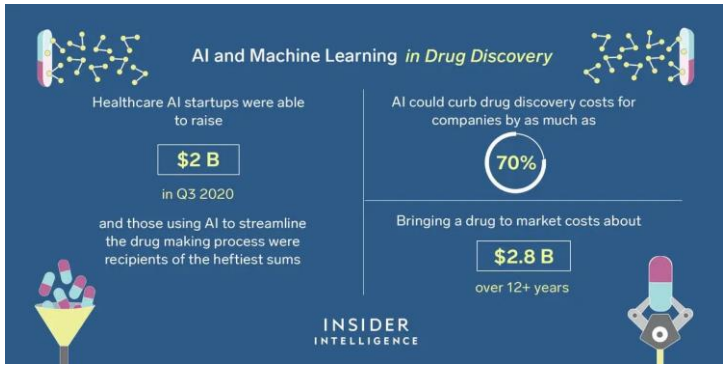
AlphaGo

Over 75% of what people watch comes from a recommendation

Recommendation Systems

Character Recognition

Assisted Driving



Drug Discovery

Finance

"Hey Siri/Google" Voice Assistants

How is Machine Learning related to Intelligent Critical Infrastructure Systems?

Critical Infrastructure Systems (CIS)

Heterogeneous - Interdependent - Interconnected

Risks, Faults, Attacks

Information Communication Technologies (ICT)

Sensors, Actuators

Big Data, Internet of Things

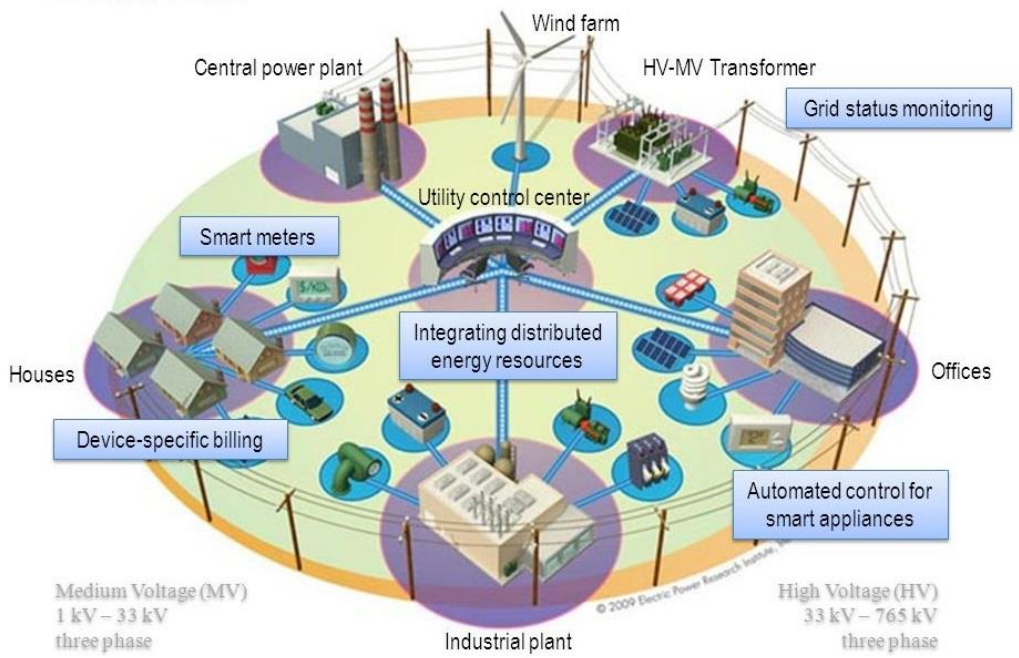
Intelligent Systems and Networks

Monitoring, Control

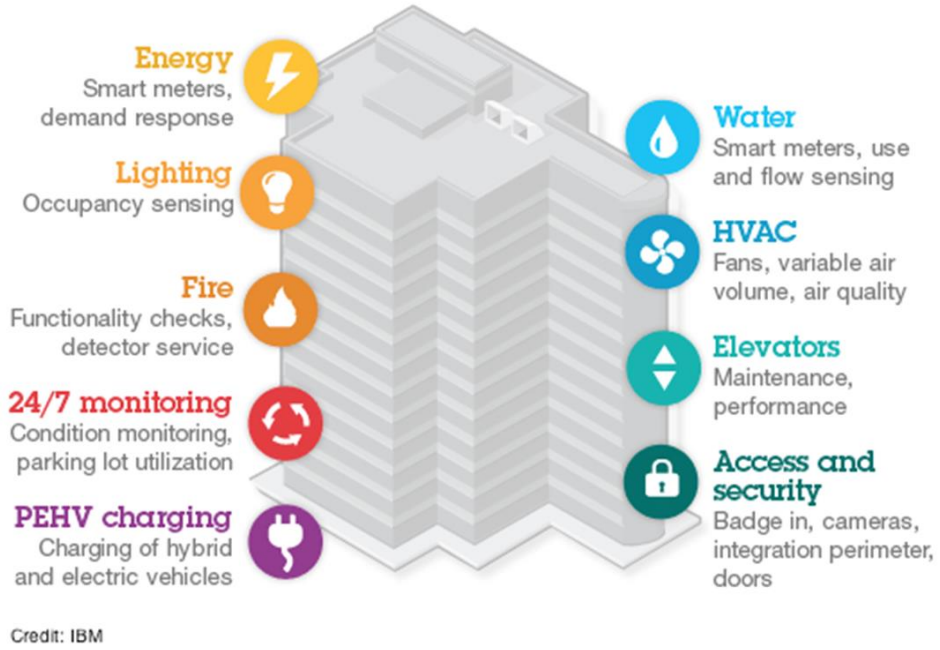
Management, Security

Big Data → Smart Decisions

How is Machine Learning related to Intelligent Critical Infrastructure Systems?



Smart Grids



Smart Buildings

How is Machine Learning related to Intelligent Critical Infrastructure Systems?

Water Systems

Revenue & Water Losses → Water Quality → Energy Consumption → Safety & Security

Water loss: seven things you need to know about an invisible global problem

A staggering 46bn litres of drinking water are lost globally every day. What can consumers, business and governments do?

▲ Iraqis fill drinking water and wash clothes at a broken water pipeline in a Shia district of Sadr City, Baghdad. Photograph: Karim Kadim/AP

China water contamination affects 2.4m after oil leak

12 April 2014 Share

Queuing up to buy bottled water on Friday - shops are now reported to have sold out

Guardian sustainable business

Solutions to the water energy nexus remain elusive

Volkswagen and Coca-Cola among businesses acting on interconnected issues of water and energy, but the voice of power companies and energy producers is largely absent

Giulio Boccaletti

Wed 10 Sep 2014 13.09 BST

49 3

▲ A section of Lake Oroville, California. The drought-stricken state uses around a fifth of its electricity for water-related purposes. Photograph: Justin Sullivan/Getty Images

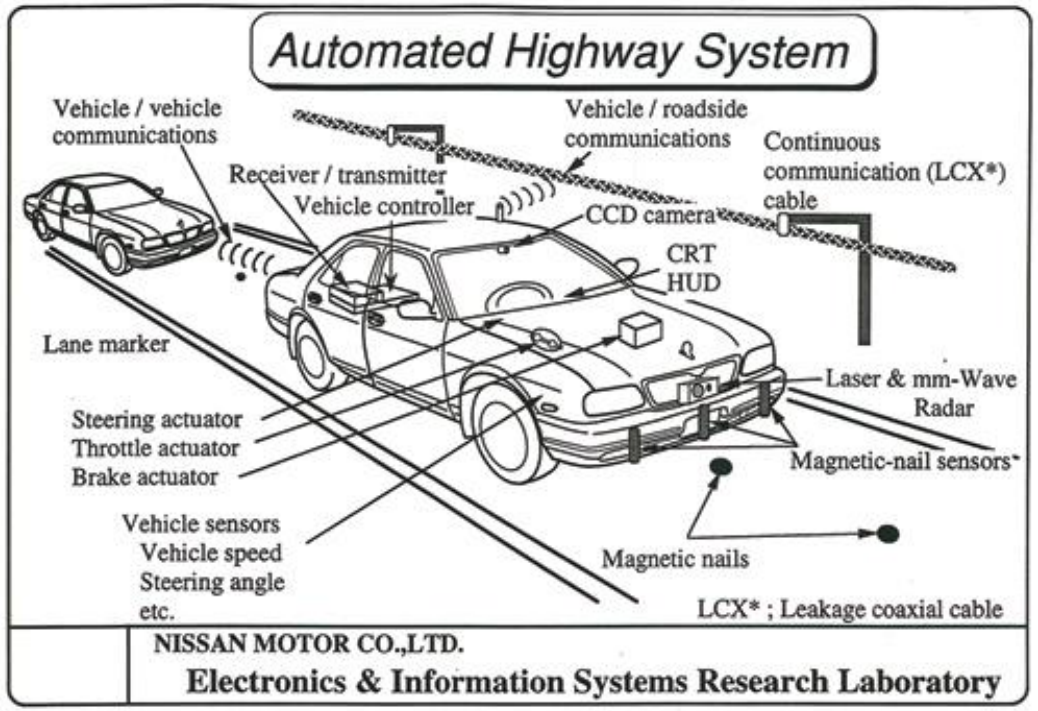
US news

Cyber-attack claims at US water facility

FBI and Homeland Security to investigate shutdown of a water pump suspected to be work of foreign hackers

▲ The US Department of Homeland Security and the FBI are to investigate claims that hackers retrieved control of a software at water utility shutting down its pump. Photograph: Jeff Gentner/AP

How is Machine Learning related to Intelligent Critical Infrastructure Systems?

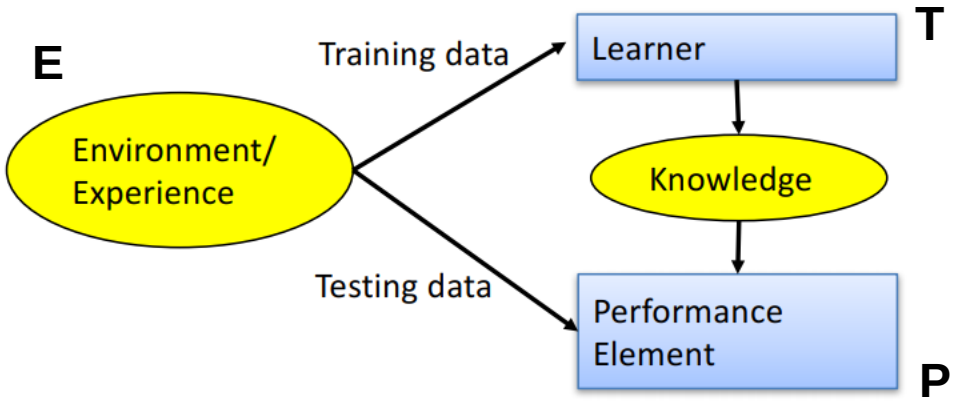


Intelligent Transportation

Multi-Agent Formation

Framing a Machine Learning Problem

- A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E** (Mitchell, 1997).

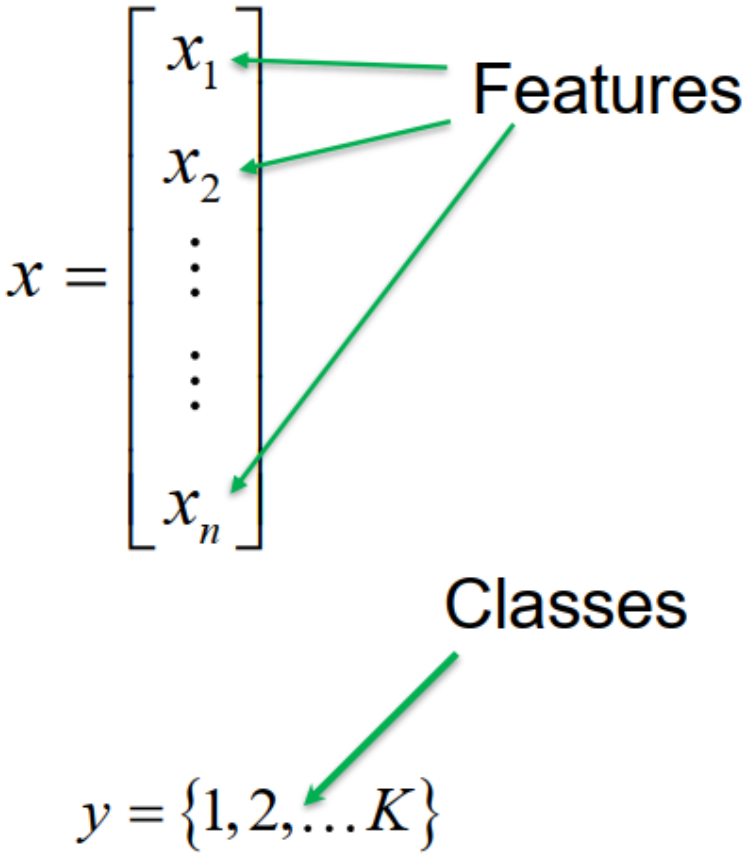


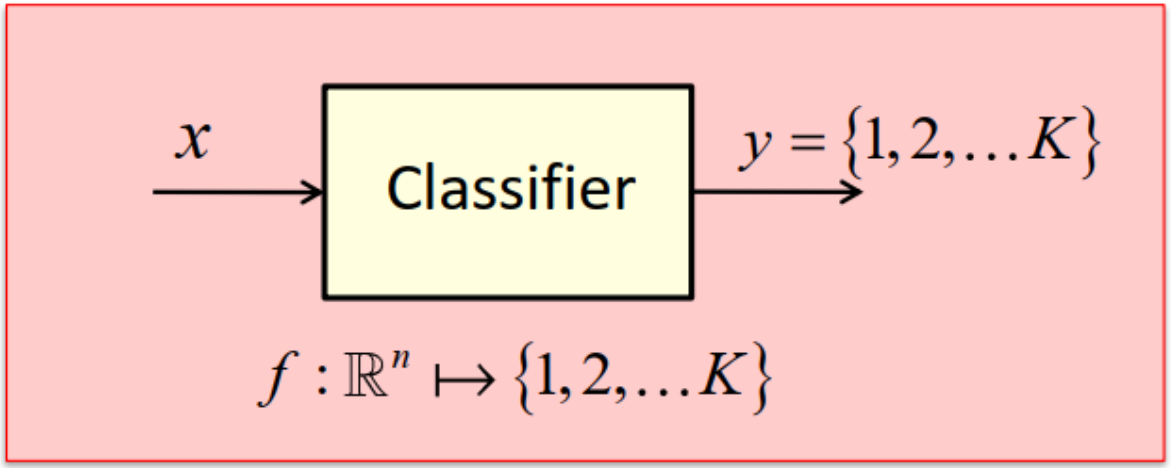
T.M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

The task T

- **Classification**
- **Regression**
- Examples:
  - Monitoring and anomaly detection
  - Transcription
  - Speech recognition
  - Machine translation
  - Navigation and Control

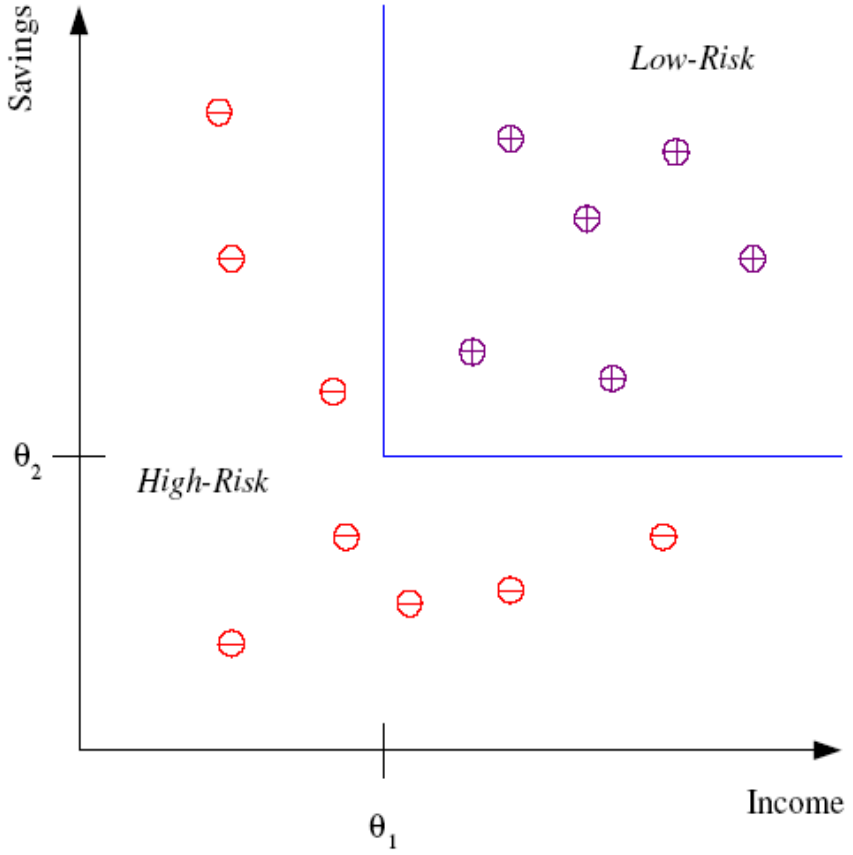
The Task T - Classification





Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their income and savings



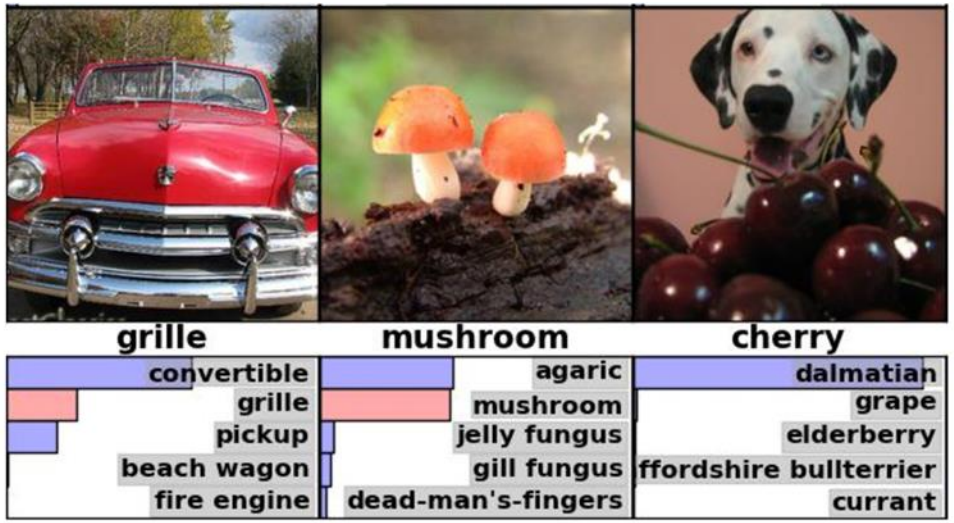
Discriminant: IF *income* $> \theta_1$ AND *savings* $> \theta_2$
THEN low-risk ELSE high-risk

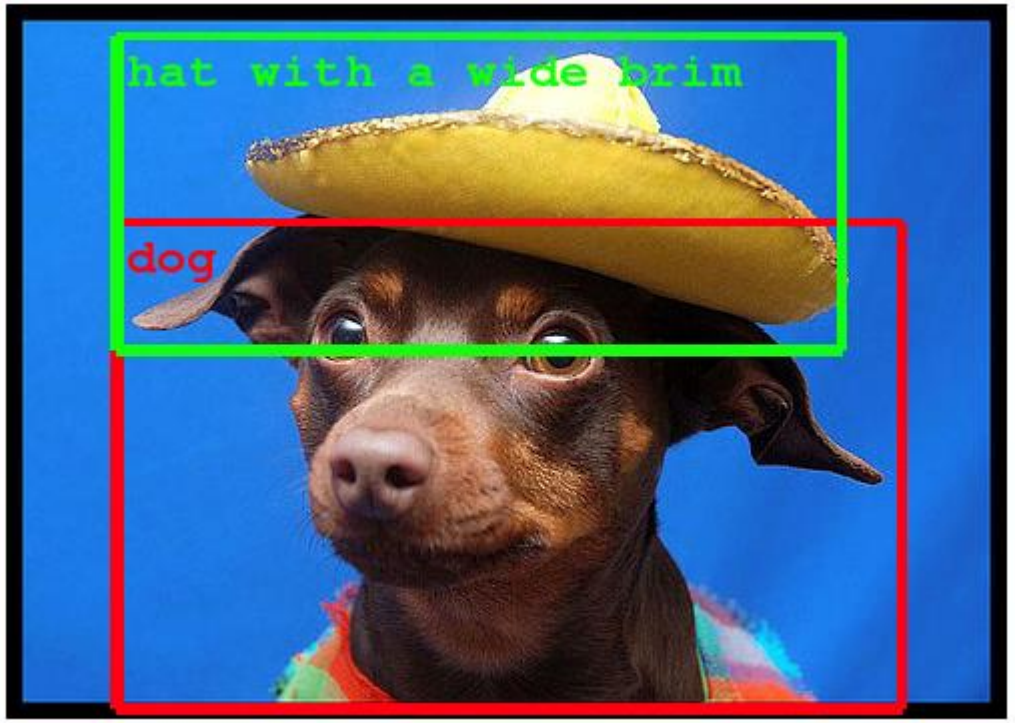
Classification: Applications

- Aka Pattern recognition
  - Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
  - Character recognition: Different handwriting styles.
  - Speech recognition: Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
  - Medical diagnosis: From symptoms to illnesses
  - ...

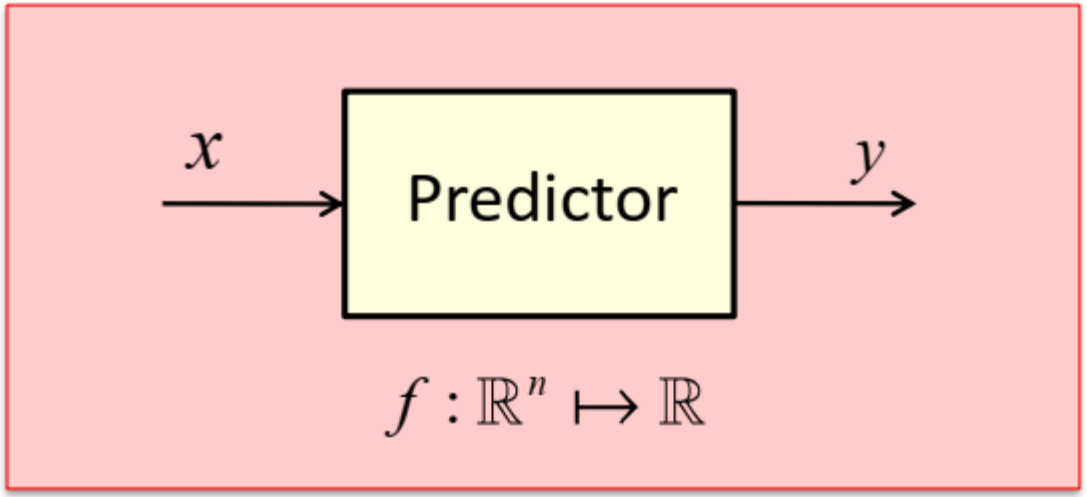
Classification

- Object recognition
- https://ai.googleblog.com/2014/09/building-deeper-understanding-of-images.html





The Task T - Regression



Regression

- Example: Price of a used car

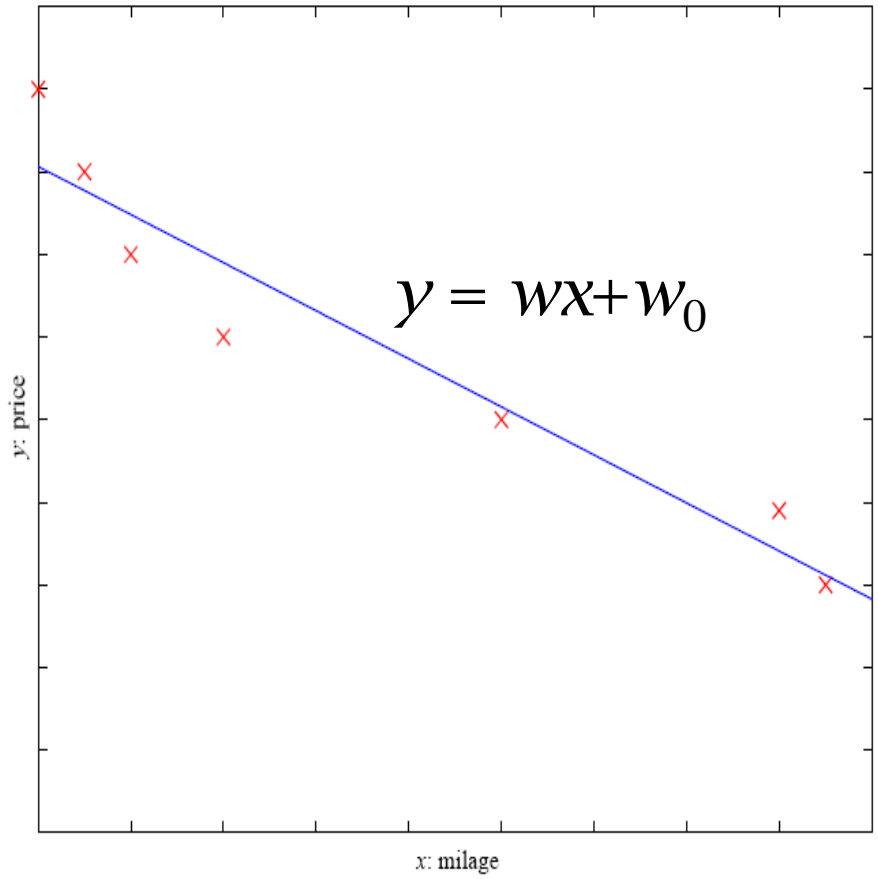
  x : car attributes

  y : price

  $$y = g(x \mid \theta)$$

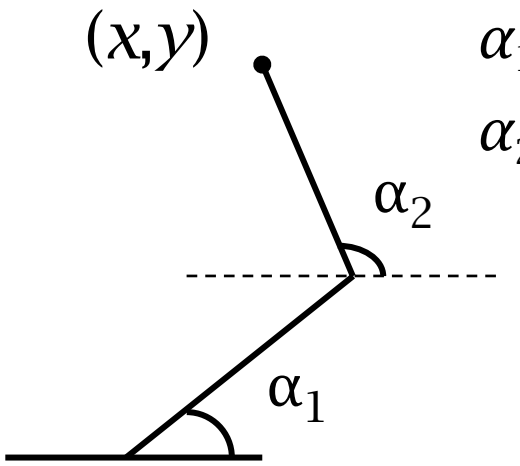
  g ( ) model,

  θ parameters



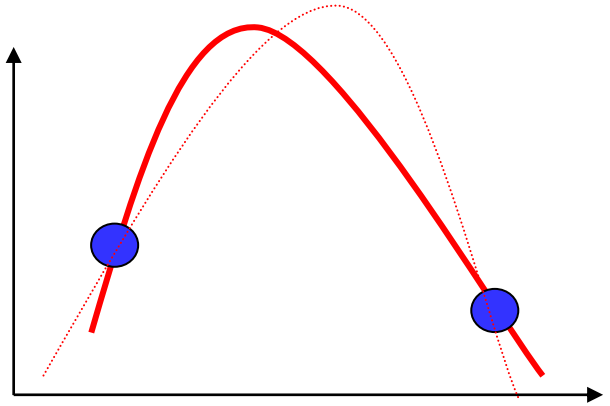
Regression Applications

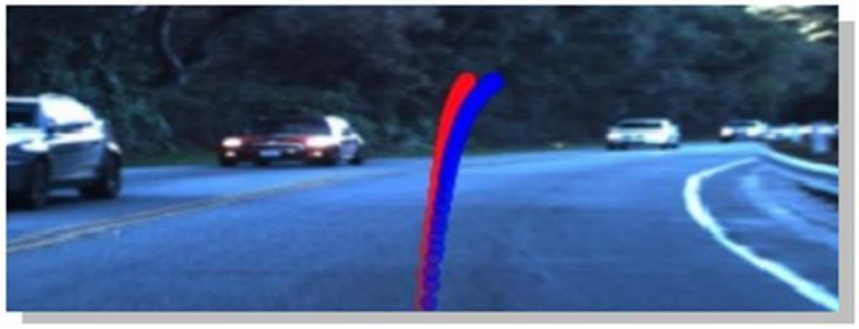
- Navigating a car: Angle of the steering wheel (CMU NavLab)
- Kinematics of a robot arm



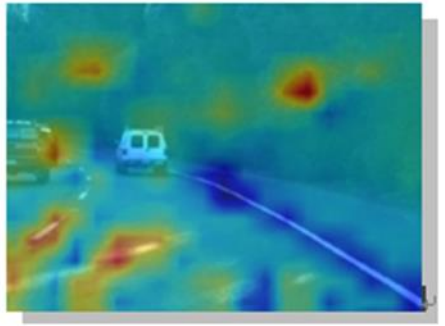
$$\alpha_1 = g_1(x,y)$$
$$\alpha_2 = g_2(x,y)$$

Response surface design





Steering angle



Attention Map

Regression

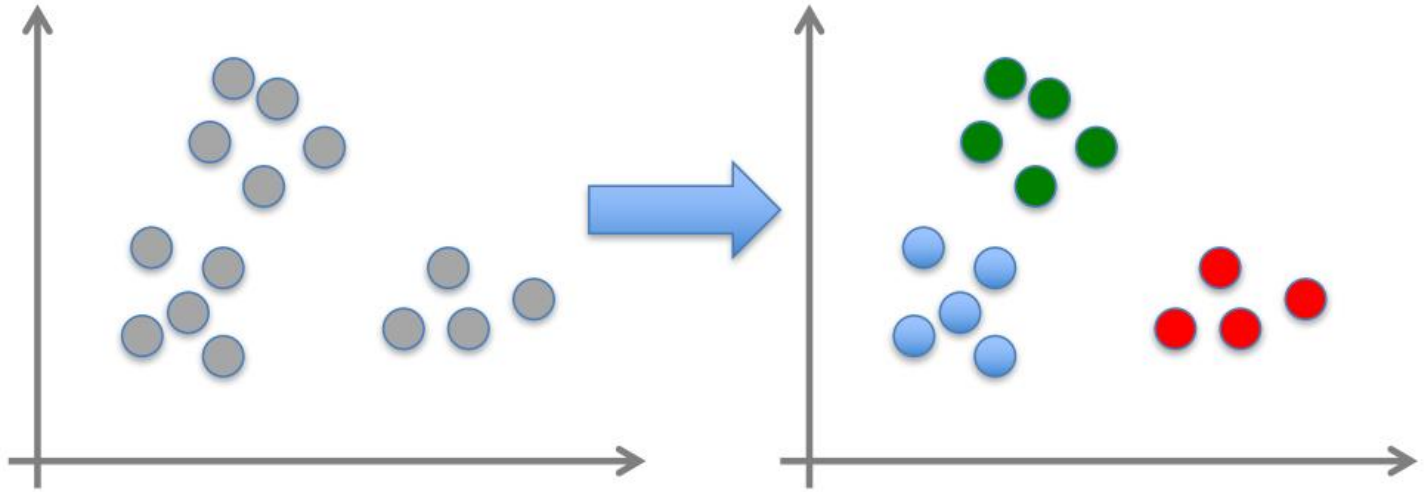
- Colorize B&W images automatically
- https://tinyclouds.org/colorize/

Supervised Learning: Uses

- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

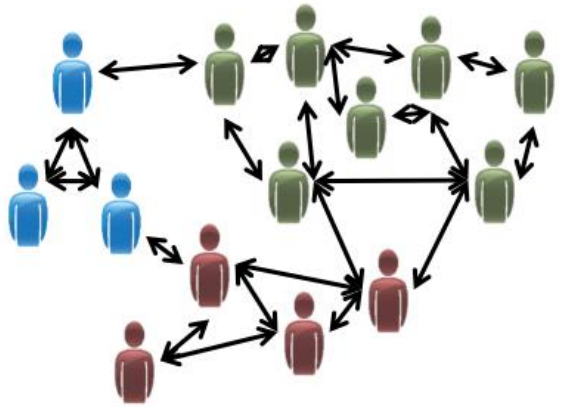
Unsupervised Learning

- Learning “what normally happens”
- Given $x_1, x_2, \ldots, x_n$ (without any labels/targets)
- Output hidden structure behind the x’s
  - e.g., clustering:
    - grouping similar instances



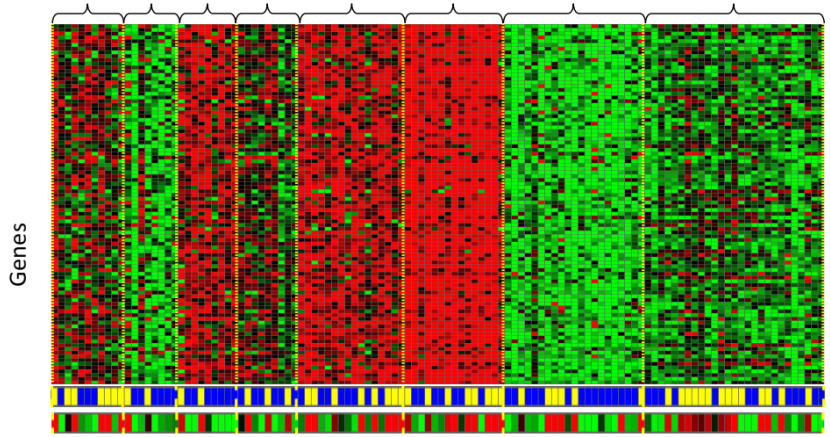
Unsupervised Learning

Organize computing clusters



Social network analysis

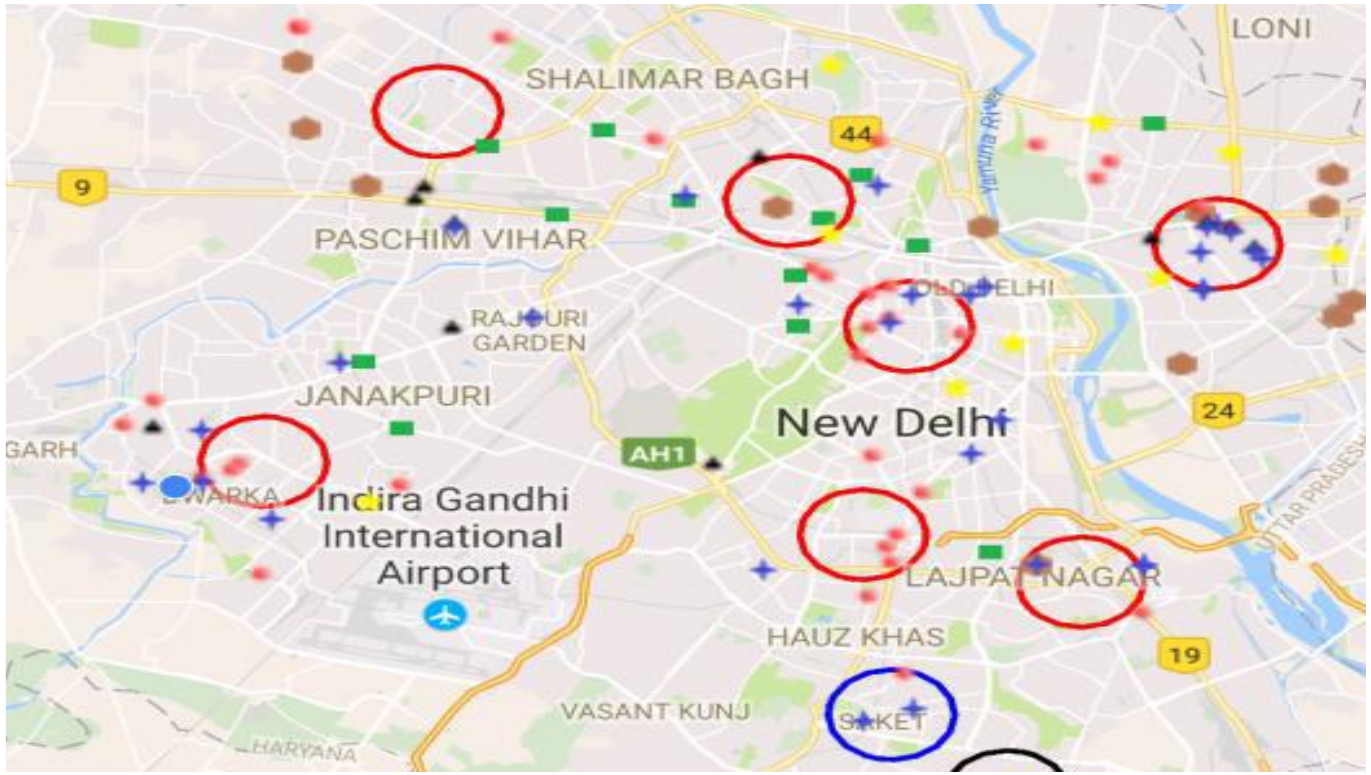
Market segmentation



Genomics application: group individuals by genetic similarity

Clustering

- Crime prediction using k-means clustering
- http://www.grdjournals.com/uploads/article/GRDJE/V02/I05/0176/GRDJEV02I050176.pdf

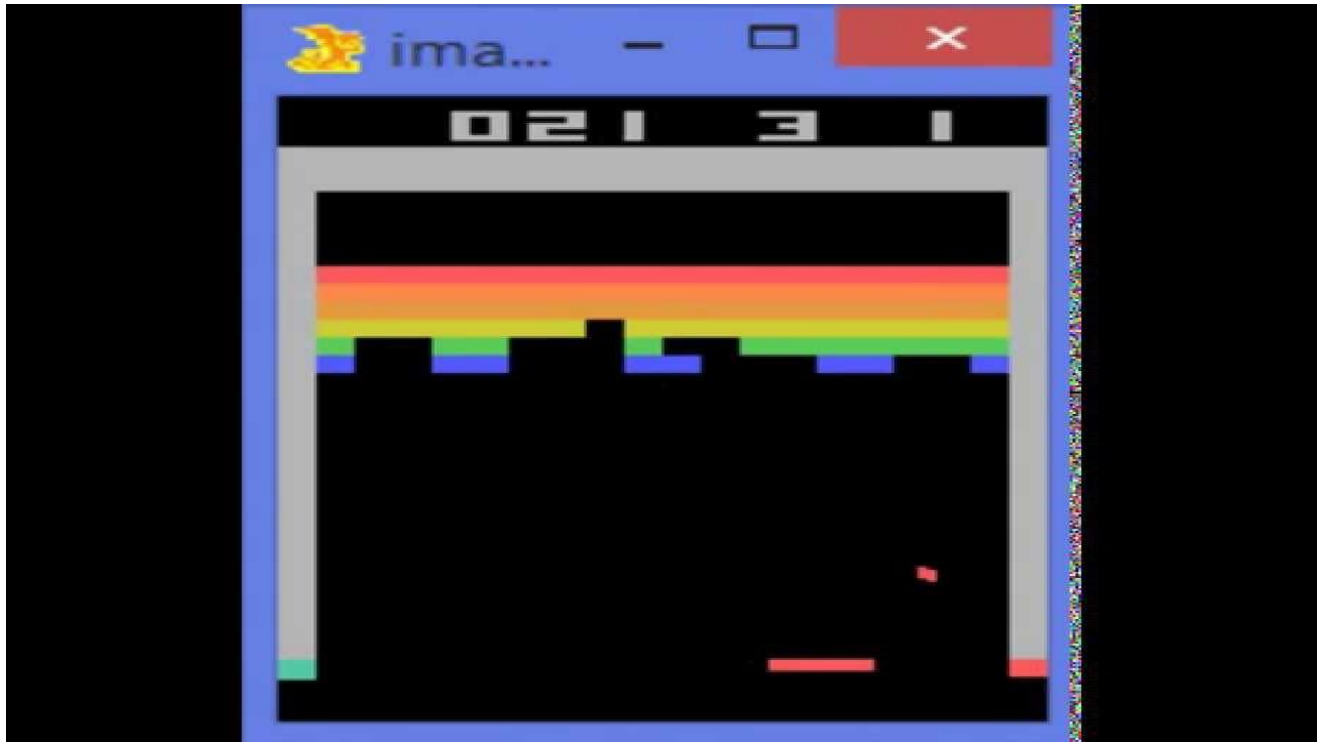


Reinforcement Learning

- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from states → actions that tells you what to do in a given state
- Learning a policy: A sequence of outputs
- No supervised output but delayed reward
- Examples:
  - Credit assignment problem
  - Game playing
  - Robot in a maze
  - Balance a pole on your hand

Learning to play games

- Learning to play Break Out
  - https://www.youtube.com/watch?v=V1eYniJ0Rnk
- May not seem useful at first
  - Easy playground for developing algorithms which can then be applied to real-world problems



The Performance Measure P

- Error rate; error properties
- Training set; Test set
- Different performance measures may be used
  - Accuracy
  - Precision and Recall
  - Squared Error
  - Likelihood
  - Cost/Utility
  - Margin
  - Entropy
  - KL-divergence

The Experience E

- What type of dataset is available?
- We generally assume that the training and test examples are independently drawn from the same overall distribution of data
  - We call this “i.i.d.” which stands for “independent and identically distributed”
- A search through a space of hypotheses (representations of functions) for one that best fits a set of training data.

Training set (labels known)

Test set (labels unknown)

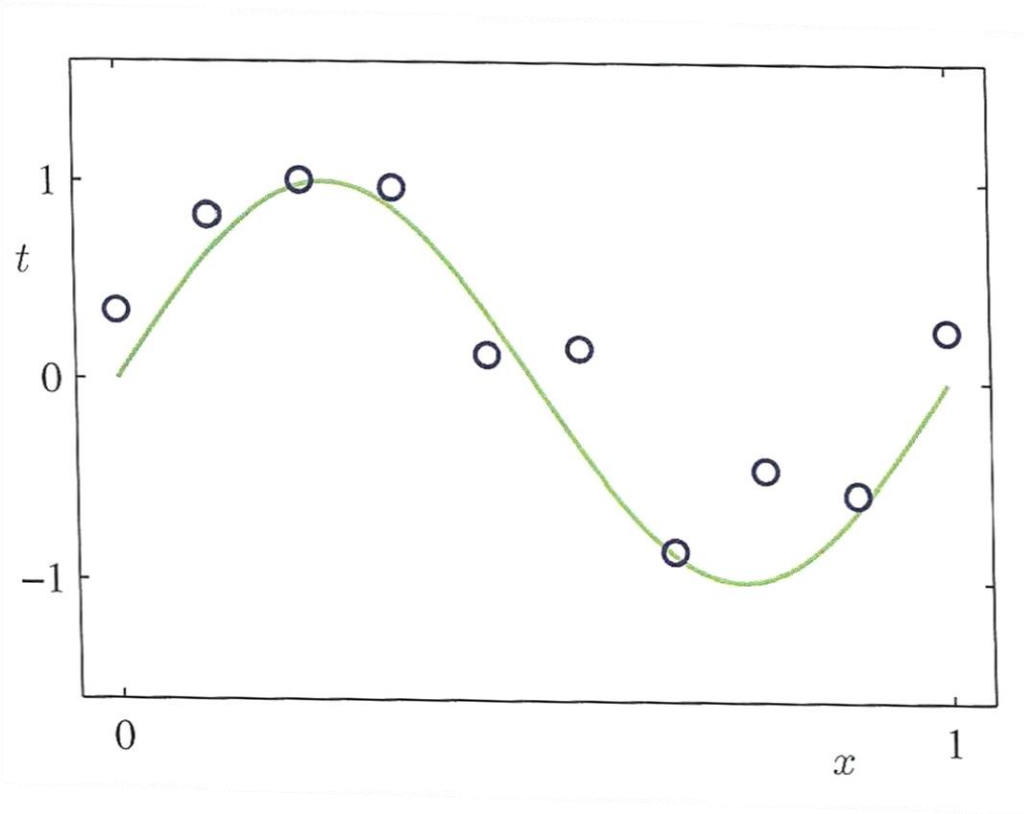
Machine Learning – Simple Example

Dataset:

$$f(x) = \sin(2\pi x)$$

$$y = f(x) + \varepsilon$$

$$x = (x_1, x_2, \ldots x_N) \qquad N = 10$$



Machine Learning – Simple Example

- Adaptive Approximation Model: polynomial

$$\hat{f}(x;\theta) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_M x^M = \sum_{j=0}^{M} \theta_j x^j$$

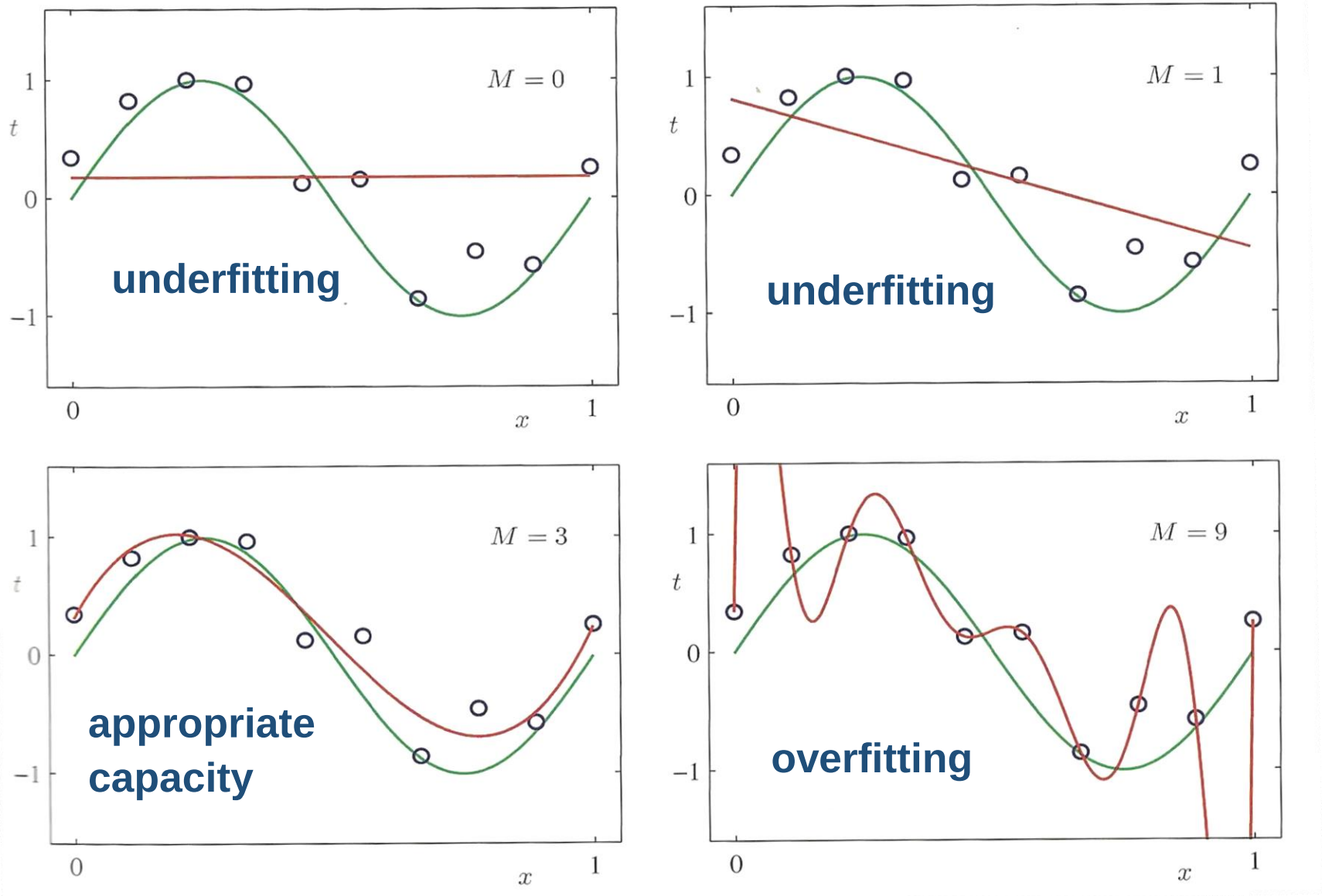
- Learning: minimizing a cost function (or error function)
  - Empirical Risk Minimization: Measure performance on a known set of training data (the "empirical" risk)

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i,\theta), y_i) \qquad \frac{1}{n}\sum_{i=1}^{n} \left(\hat{f}(x_n,\theta) - y_n\right)^2$$

$$\mathrm{R(f)} = \mathrm{E[L(f(x,\theta),y)]} = \int \mathrm{L(f(x,\theta),y)} dP(x,y)$$

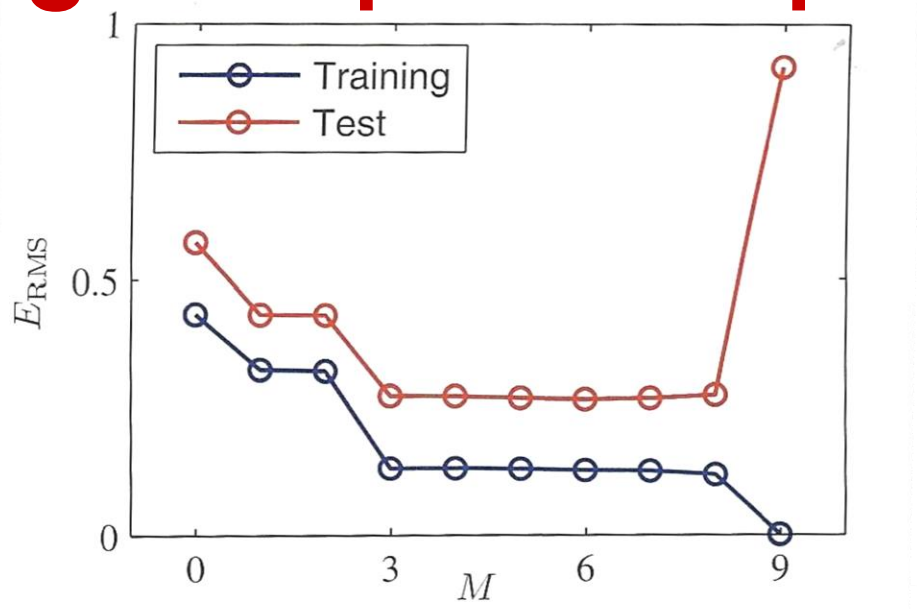
$$\hat{f} = \underset{f \in H}{\operatorname{argmin}} R_{emp}(f)$$

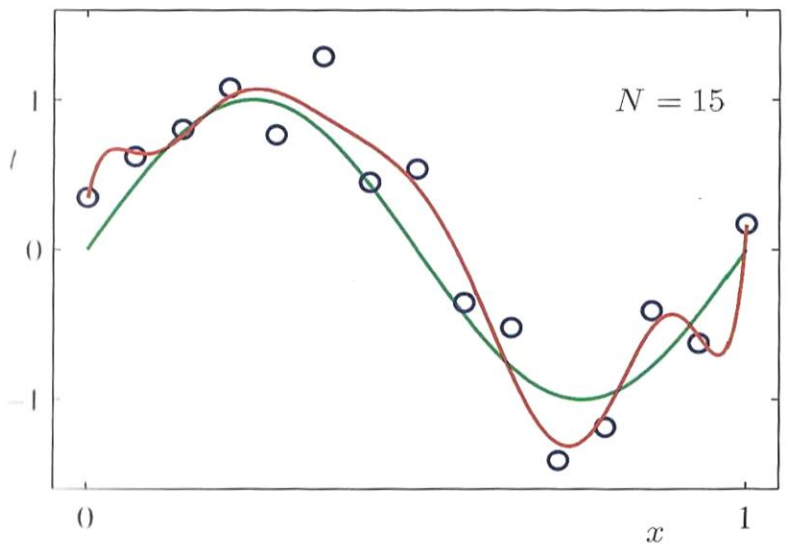
Machine Learning – Simple Example

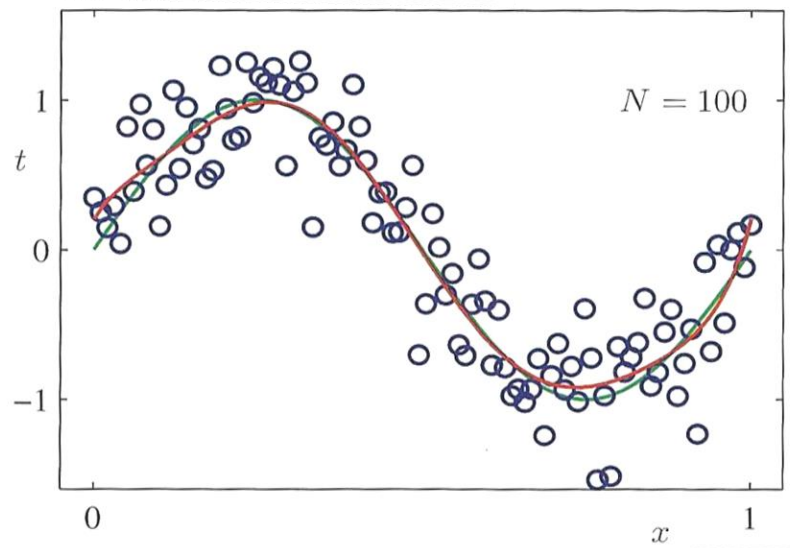


Machine Learning – Simple Example

underfitting and
overfitting depends
on both M and N







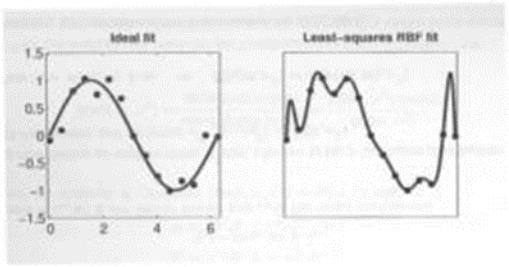
Underfitting and Overfitting

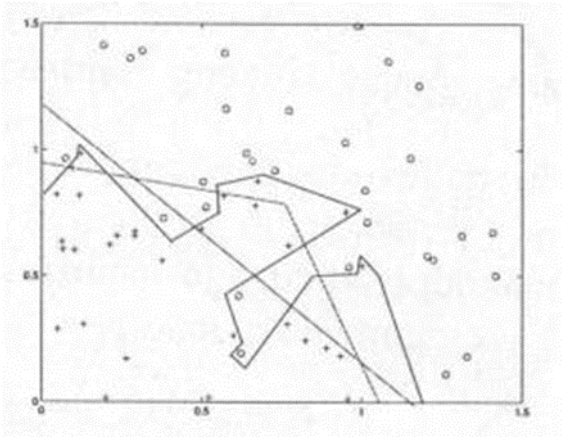
- Underfitting
  - Model is too small to fit the data
  - In other words, the approximation error is big

- Overfitting
  - Artificially good agreement with the data
  - Empirical risk is small only for given data, and gets big for other data
  - Estimation error gets big.

Need for Restriction on Model

- It is always possible to construct a function that fits the given data exactly
- But is it reasonable? Is it desirable?





- Need a PRINCIPLED way of restricting the functions a learning algorithm will construct.

No Free lunch Principle

- Learning does not take place in a vacuum
- Without assumptions on how the past and future are related, prediction is impossible
- Without restriction on the possible phenomenon (model, function class) generalization is impossible.

- Data will never replace knowledge

Capacity, Overfitting and Underfitting

- **Occam's Razor** – is a principle that states that among competing hypothesis that explain known observations equally well, we should choose the simplest one. Named after William of Ockham (1287-1347).
- **Statistical Learning Theory** provides various means of quantifying model capacity. The Vapnik-Chervonenkis dimension (VC dimension) is the most well known. However, it is not very practical with advanced machine learning algorithms
  - Unfortunately, while these complexity measures have become broadly useful tools in statistical theory, they turn out to be powerless (as straightforwardly applied) for explaining why deep neural networks generalize.

Capacity, Overfitting and Underfitting

- The ability to perform well on previously unobserved inputs is called ***generalization***. This is a key challenge in machine learning.
- Training set vs. test set. Typically, the generalization error (test error) is measuring the performance on a test set.
- The expected test error is greater than or equal to the expected training error

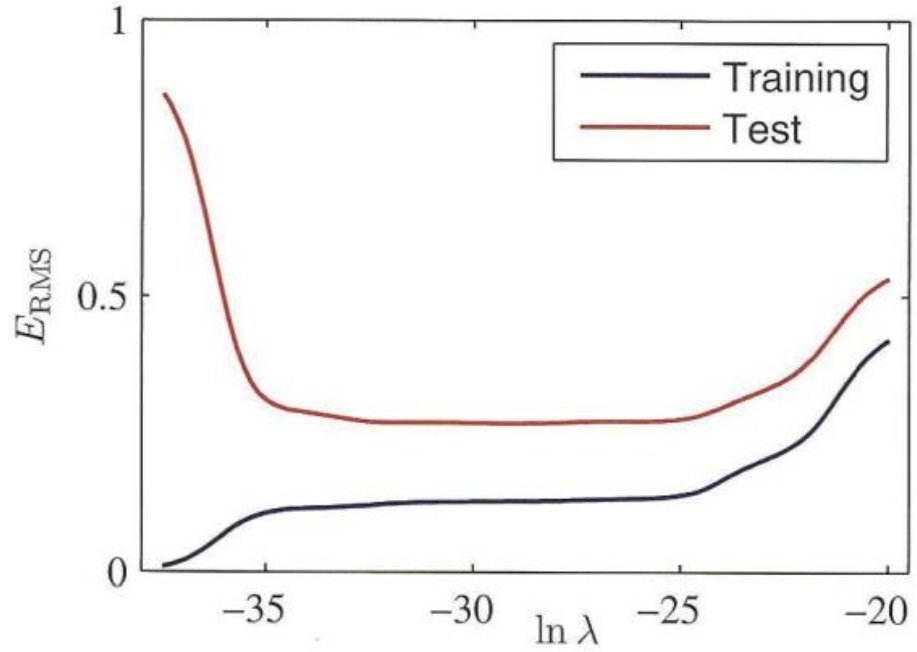
The performance of a machine learning algorithm is based on both:

1. Making the training error small
2. Making the gap between the training and test error small

Regularizarion

- Structural risk minimization (SRM)

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{f}(x_n, \theta) - y_n\right)^2 + \frac{\lambda}{2}\|\theta\|^2$$



Hyperparameters and Validation Sets

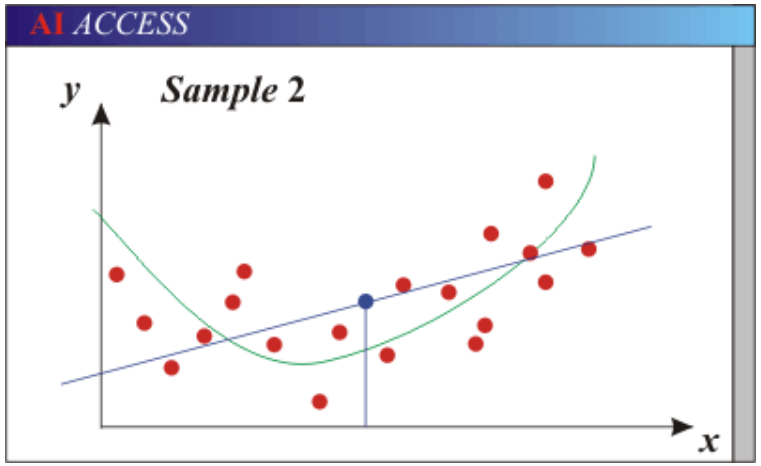
- **Hyperparameters** – parameters that affect the performance of the algorithm but they are not adapted by the learning algorithm itself; e.g., the degree of the polynomial is a capacity hyperparameter; the value of λ in regularization is another hyperparameter.
- **Validation Set** – is used to select the hyperparameters. Split the training set into two disjoint subsets; one subset of data is used to learn the parameters, the other subset (validation set) is used to estimate the generalization error in order to update the hyperparameters; e.g.; 80% of data is used for training and 20% for validation.
- Three types of data: Training set, Validation set, Test set.

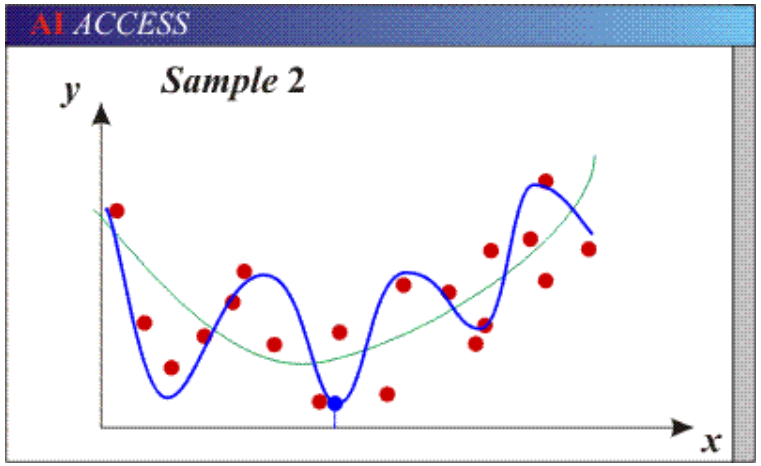
  (for example, 70% training set, 15% validation and 15% testing set)

Generalization

- Components of generalization error
  - **Bias**: how much the average model over all training sets differ from the true model?
  - **Variance**: how much models estimated from different training sets differ from each other
  - Error due to inaccurate assumptions/simplifications made by the model
- Underfitting: model is too “simple” to represent all the relevant class characteristics
  - High bias and low variance
  - High training error and high test error
- Overfitting: model is too “complex” and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error

Bias-Variance Trade-off





- Models with too few parameters are inaccurate because of a large bias (not enough flexibility).
- Models with too many parameters are inaccurate because of a large variance (too much sensitivity to the sample).

Bias-Variance Trade-off

E(MSE) = bias^2 + variance + noise^2

$$\mathrm{E}_{D,\varepsilon}\left[\left(y-\hat{f}\left(x;D\right)\right)^2\right]=\left(\mathrm{Bias}_D\left[\hat{f}\left(x;D\right)\right]\right)^2+\mathrm{Var}_D\left[\hat{f}\left(x;D\right)\right]+\sigma^2$$

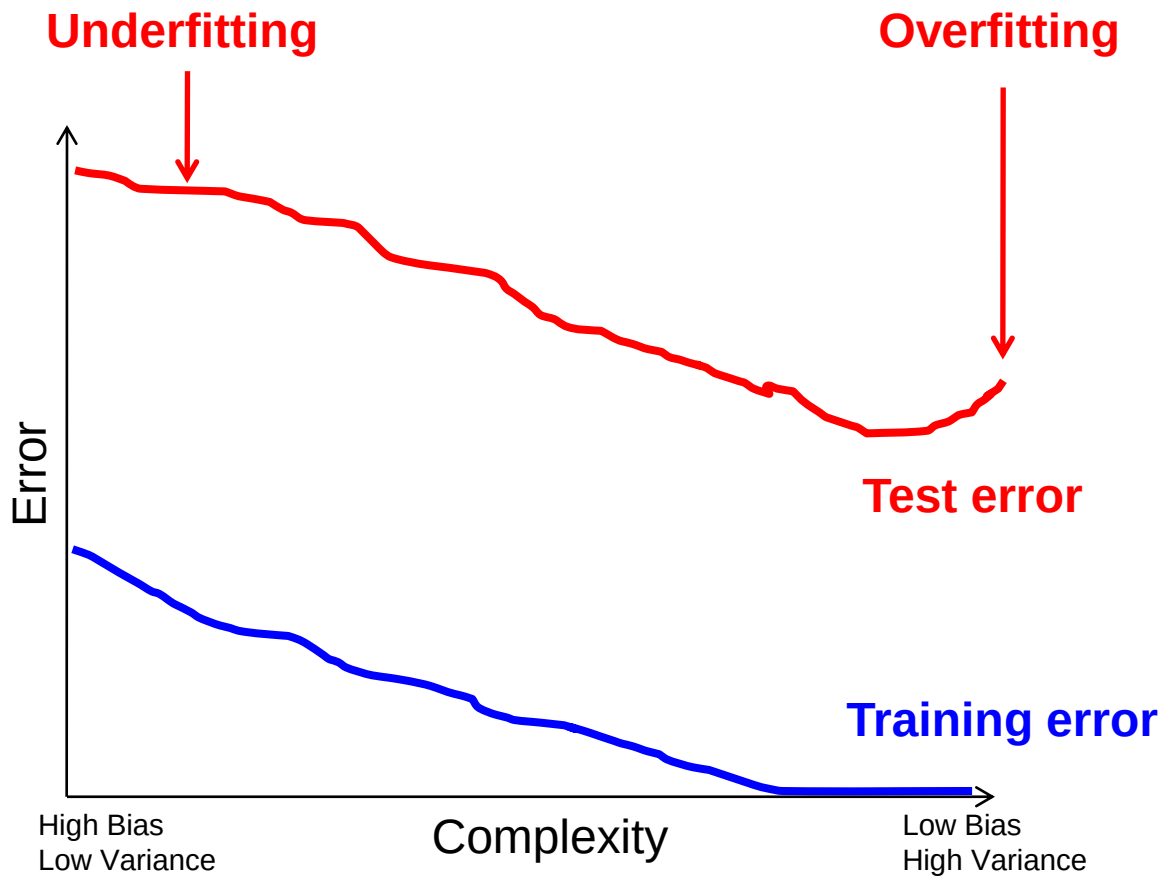
Error due to incorrect assumptions

Error due to variance of training samples

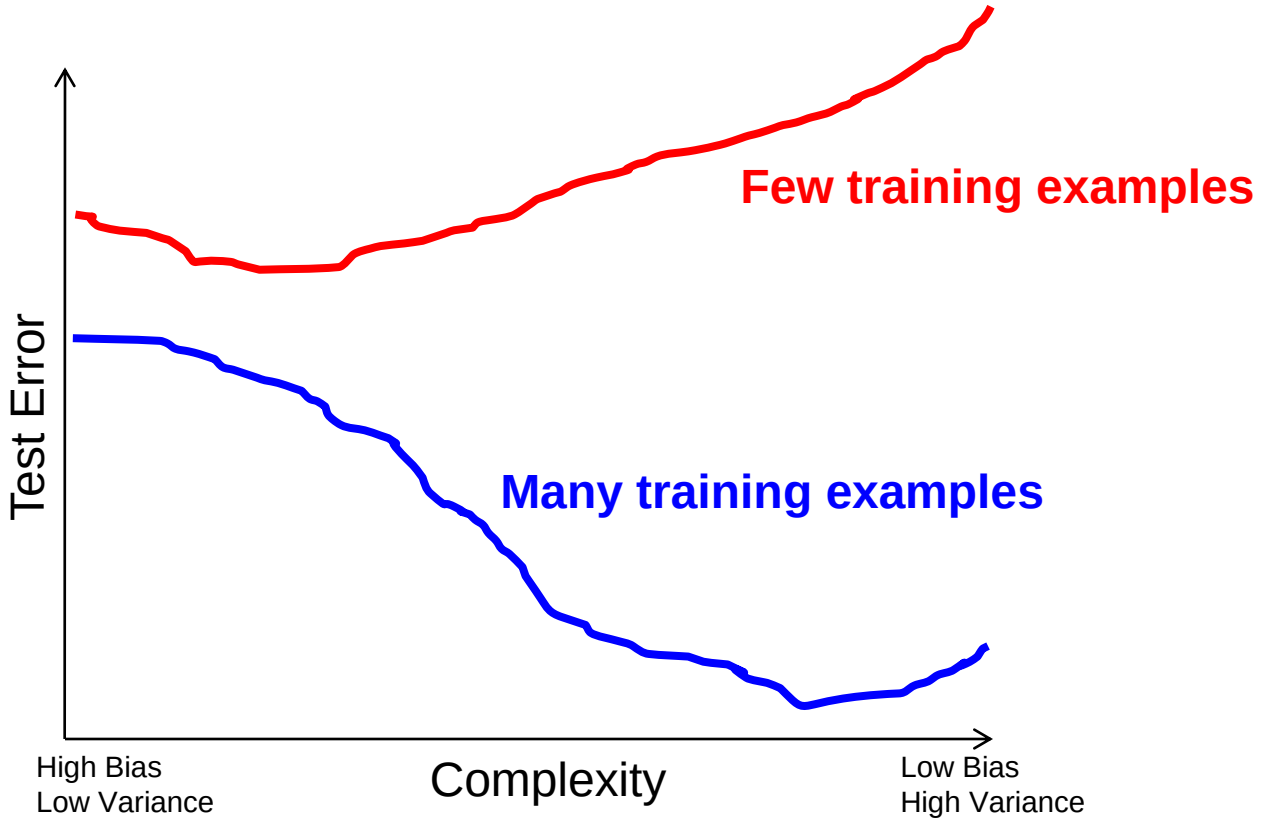
Unavoidable error

- See the following for explanations of bias-variance (also Bishop's "Neural Networks" book):
- http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture4/BiasVariance.pdf

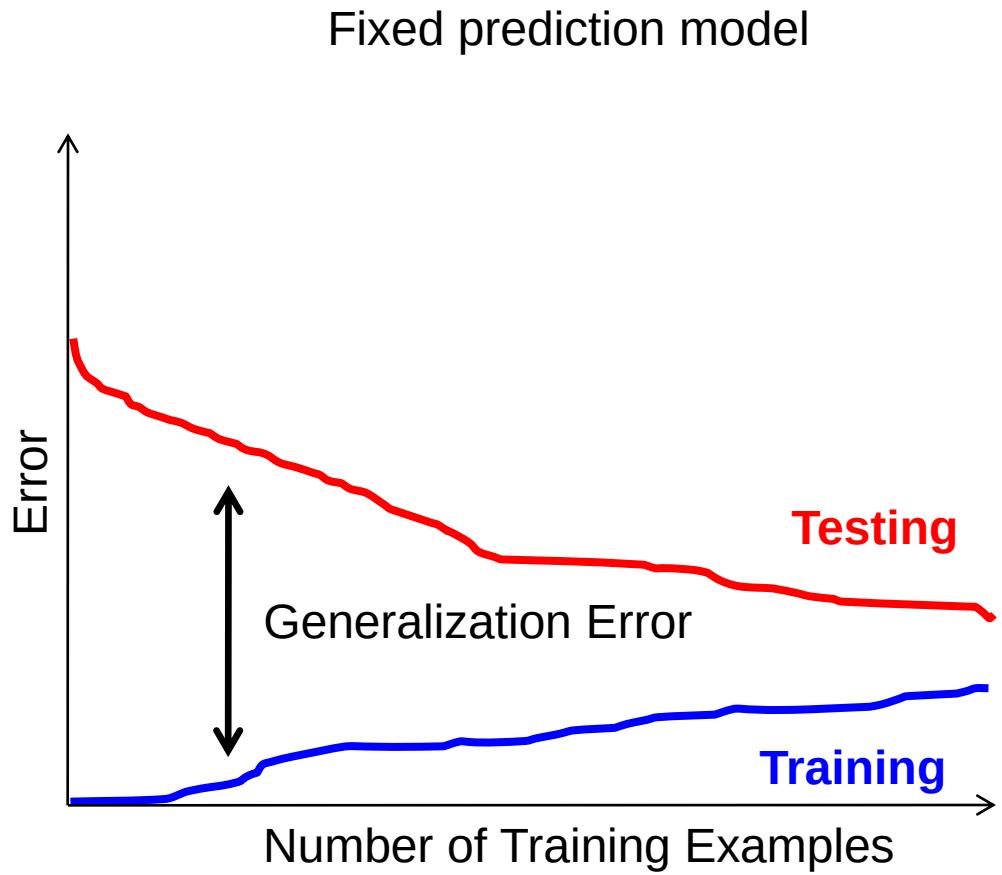
Bias-variance tradeoff



Bias-variance tradeoff



Effect of Training Size



Batch Learning vs. Online Learning

- **Batch Learning** – machine learning methods that are based on learning on the entire training data set.
- **Online Learning** – machine learning methods that are based on data becoming available in sequential order and is used to update the parameters at each step. Data does not need to be stored after it is used to update the adjustable parameters
- Something in between batch learning and online learning is to use a window of data (not the entire data) to update the parameters.

Core Machine Learning Concepts

- Tens of thousands of machine learning algorithms
  - Hundreds new/variations every year

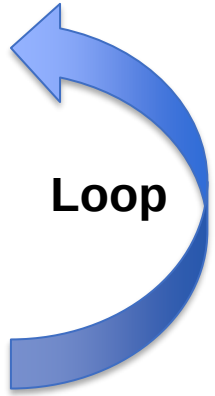
- **Objective function:** encodes the right loss for the problem
- **Parameterization:** makes assumptions that fit the problem
- **Regularization:** right level of regularization for amount of training data
- **Training algorithm:** can find parameters that maximize objective on training set
- **Inference algorithm:** can solve for objective function in evaluation

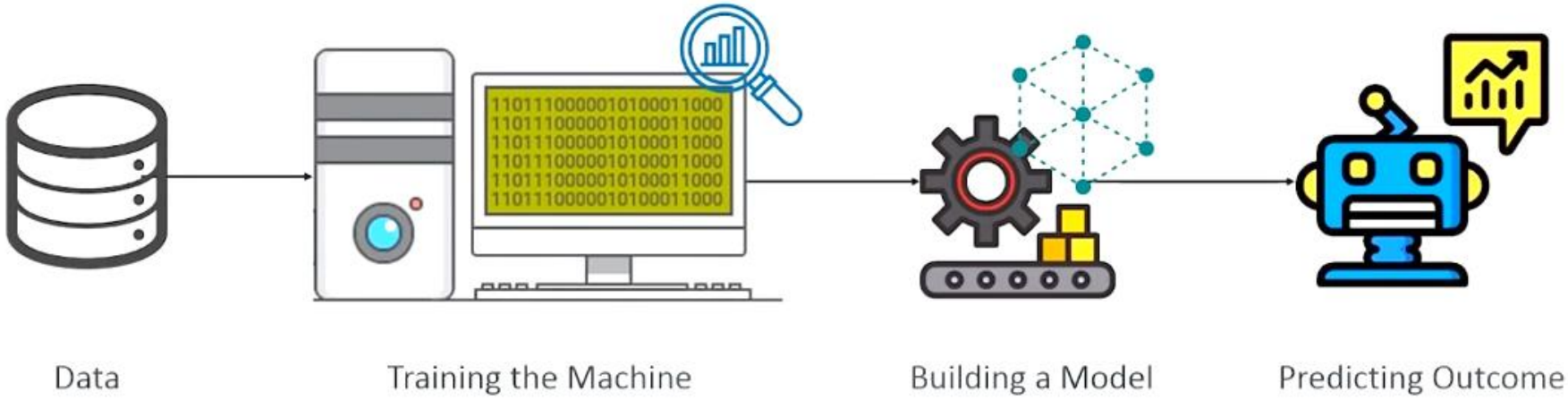
Key Issues in Machine Learning

- Modelling
  - How to formulate application problems as machine learning problems ?
  - How to represent the data?
  - Learning Protocols (where is the data & labels coming from?)
- Representation
  - What functions should we learn (hypothesis spaces) ?
  - How to map raw input to an instance space?
  - Any rigorous way to find these? Any general approach?
- Algorithms
  - What are good algorithms?
  - How do we define success?
- Generalization vs. over fitting
- The computational problem

Machine Learning in Practice

- Machine Learning involves building a **Predictive Model** that can be used to find a **Solution** for a **Problem Statement**
  - Understand domain, prior knowledge, and goals
  - Data integration, selection, cleaning, pre-processing, etc.
  - Learning Model - Training
  - Interpret results
  - Consolidate and deploy discovered knowledge





Fairness, Accountability, and Transparency

- Finally, it is important to remember that when you deploy machine learning systems you are not merely optimizing a predictive model—you are typically providing a tool that will be used to (partially or fully) automate decisions.
- These technical systems can impact the lives of individuals subject to the resulting decisions.
- The leap from considering predictions to decisions raises not only new technical questions, but also a slew of ethical questions that must be carefully considered.
- Often, the various mechanisms by which a model's predictions become coupled to its training data are unaccounted for in the modelling process.

Next…

- Understand aspects of optimization methods for learning the model
  - How to find the best model?
  - Navigate the model hypothesis space.