MACHINE LEARNING OVERVIEW

17th August, 2016

By

Satyajit Roy

Machine Learning – A Definition

Definition: 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.

A Few Quotes

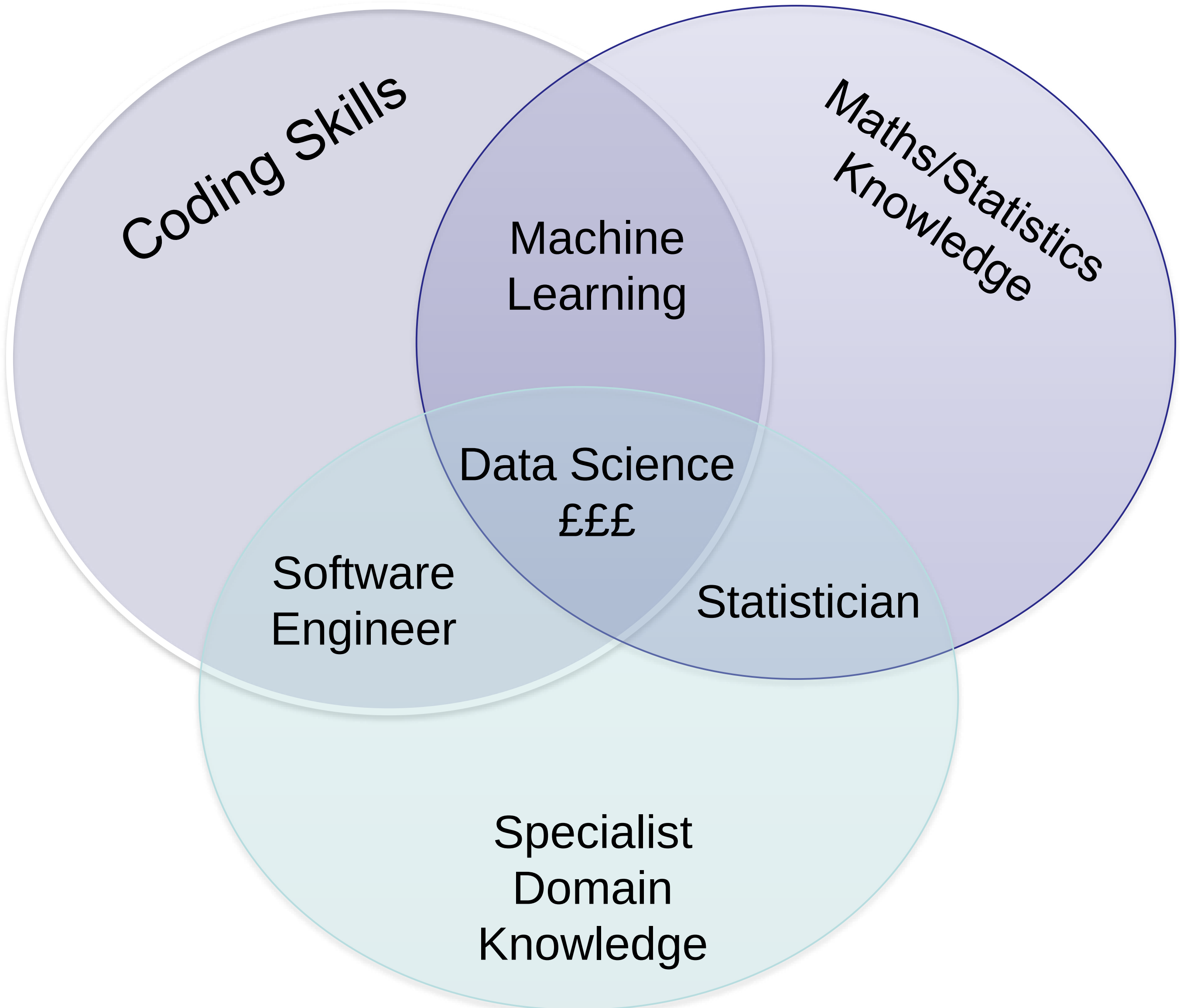
- "A breakthrough in machine learning would be worth ten Microsofts" **(Bill Gates, Chairman, Microsoft)**
- "Machine learning is the next Internet" (**Tony Tether, Director, DARPA)**
- "Machine learning is the hot new thing" **(John Hennessy, President, Stanford)**
- "Web rankings today are mostly a 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)**

What is Machine Learning ? (More Details)

- Machine Learning
  - Study of algorithms that
  - improve their performance
  - at some task
  - with experience

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

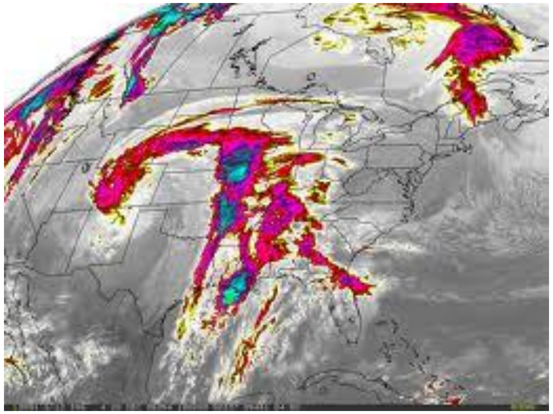
Machine Learning and Fields



Learning from Data

The world is driven by data.

- Germany’s climate research center generates 10 petabytes per year
- Google processes 24 petabytes per day
- The Large Hadron Collider produces 60 gigabytes per minute (~12 DVDs)
- There are over 50m credit card transactions a day in the US alone.

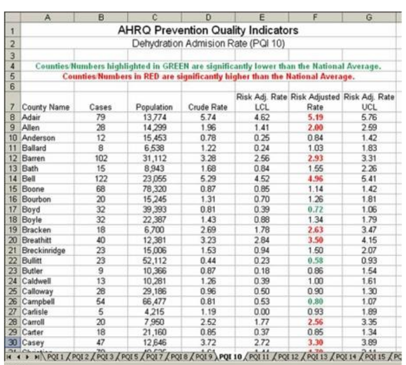


Learning from Data

- **Data** is recorded from some real-world phenomenon.

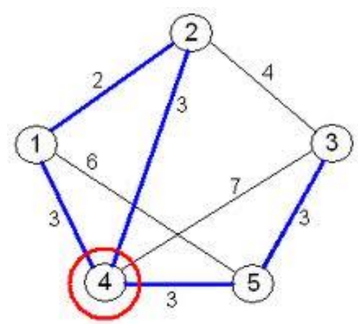
- What might we want to do with that data?
- **Prediction**
- **-** what can we **predict** about this phenomenon?

- **Description**
- **-** how can we **describe/understand** this phenomenon in a new way?



	A	B	C	D	E	F	G
1	AHRQ Prevention Quality Indicators						
2	Dehydration Admision Rate (PQI 10)						
3							
4	Counties/Numbers highlighted in GREEN are significantly lower than the National Average.						
5	Counties/Numbers in RED are significantly higher than the National Average.						
6							
7	County Name	Cases	Population	Crude Rate	Risk Adj. Rate LCL	Risk Adjusted Rate	Risk Adj. Rate UCL
8	Adair	79	13,774	5.74	4.62	5.19	5.76
9	Allen	28	14,299	1.96	1.41	2.00	2.59
10	Anderson	12	15,453	0.78	0.25	0.84	1.42
11	Ballard	8	6,538	1.22	0.24	1.03	1.83
12	Barren	102	31,112	3.28	2.56	2.93	3.31
13	Bath	15	8,943	1.68	0.84	1.55	2.26
14	Bell	122	23,055	5.29	4.52	4.96	5.41
15	Boone	68	78,320	0.87	0.85	1.14	1.42
16	Bourbon	20	15,245	1.31	0.70	1.26	1.81
17	Boyd	32	39,393	0.81	0.39	0.72	1.06
18	Boyle	32	22,387	1.43	0.88	1.34	1.79
19	Bracken	18	6,700	2.69	1.78	2.63	3.47
20	Breathitt	40	12,381	3.23	2.84	3.50	4.15
21	Breckinridge	23	15,006	1.53	0.94	1.50	2.07
22	Bullitt	23	52,112	0.44	0.23	0.58	0.93
23	Butler	9	10,366	0.87	0.18	0.86	1.54
24	Caldwell	13	10,281	1.26	0.39	1.00	1.61
25	Calloway	28	29,186	0.96	0.50	0.90	1.30
26	Campbell	54	66,477	0.81	0.53	0.80	1.07
27	Carlisle	5	4,215	1.19	0.00	0.93	1.89
28	Carroll	20	7,950	2.52	1.77	2.56	3.35
29	Carter	18	21,160	0.85	0.37	0.85	1.34
30	Casey	47	12,646	3.72	2.72	3.30	3.89

PQI 1 / PQI 2 / PQI 3 / PQI 5 / PQI 7 / PQI 8 / PQI 9 / PQI 10 / PQI 11 / PQI 12 / PQI 13 / PQI 14 / PQI 15



Learning from Data

How can we extract knowledge from data to help humans take decisions?

How can we automate decisions from data?

How can we adapt systems dynamically to enable better user experiences?

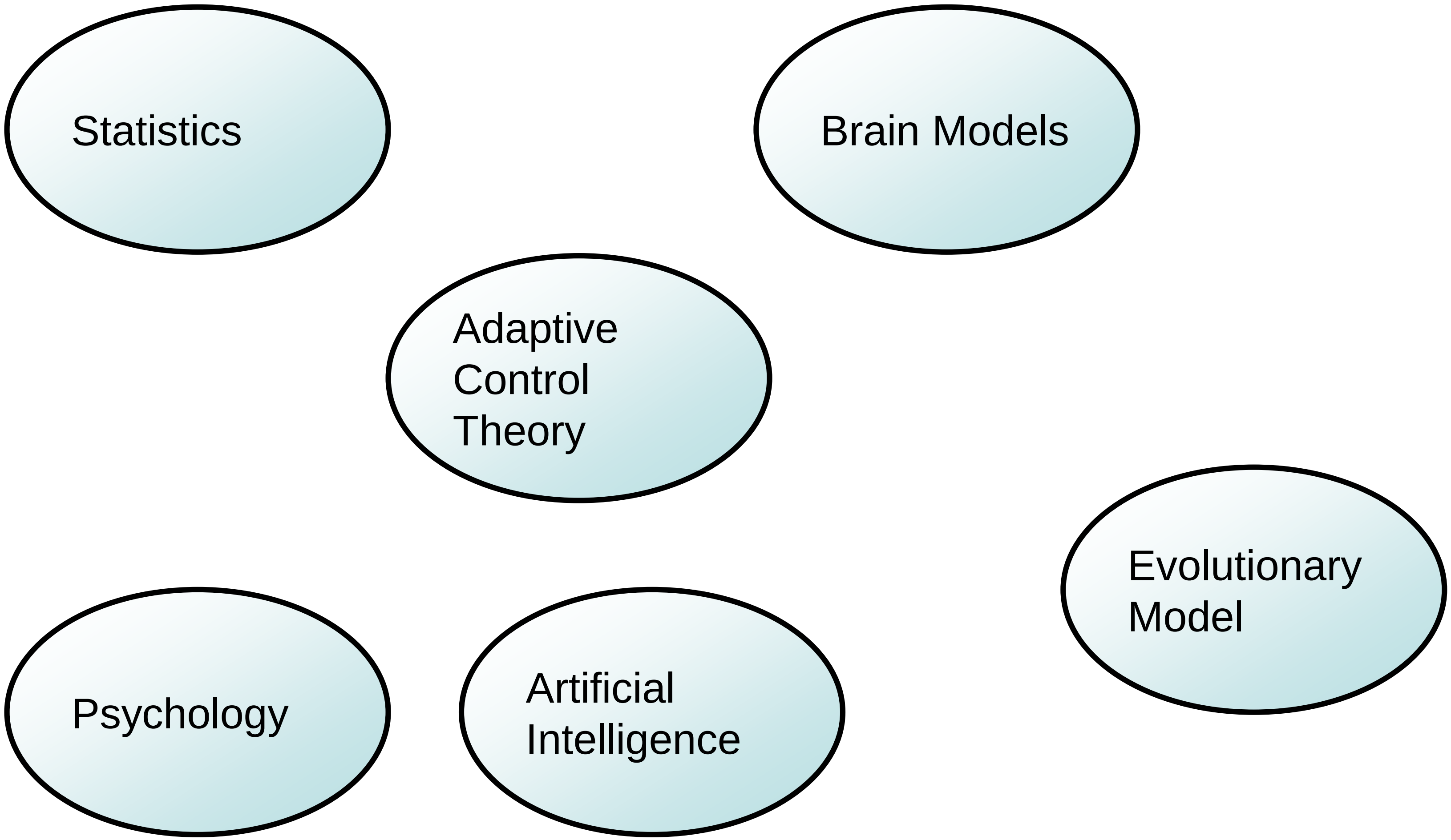
Write code to explicitly do the above tasks

Write code to make the computer ***learn*** how to do the tasks

Why is Machine Learning Important?

- Some tasks cannot be defined well, except by examples (e.g., recognizing people)
- Relationships and correlations can be hidden within large amounts of data. Machine Learning/Data Mining may be able to find these relationships
- Human designers often produce machines that do not work as well as desired in the environments in which they are
- The amount of knowledge available about certain tasks might be too large for explicit encoding by humans (e.g., medical diagnostic)
- Environments change over time.
- New knowledge about tasks is constantly being discovered by humans. It may be difficult to continuously re-design systems "by hand"

Areas of Influence for Machine Learning



Areas of Influence for Machine Learning

- **Statistics:** How best to use samples drawn from unknown probability distributions to help decide from which distribution some new sample is drawn?

- **Brain Models:** Non-linear elements with weighted inputs (Artificial Neural Networks) have been suggested as simple models of biological neurons.
- **Adaptive Control Theory:** How to deal with controlling a process having unknown parameters that must be estimated during operation?
- **Psychology:** How to model human performance on various learning tasks?
- **Artificial Intelligence:** How to write algorithms to acquire the knowledge humans are able to acquire, at least, as well as humans?
- **Evolutionary Models:** How to model certain aspects of biological evolution to improve the performance of computer programs?

Machine Learning in Practice

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

The machine learning framework

$$y = f(\mathbf{x})$$

output prediction function features

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- **Testing:** apply f to a never before seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

Machine Learning Steps

Training

Training Images

Image Features → Training → Learned model

Training Labels → Training

Testing

Test Image

Image Features → Learned model → Prediction

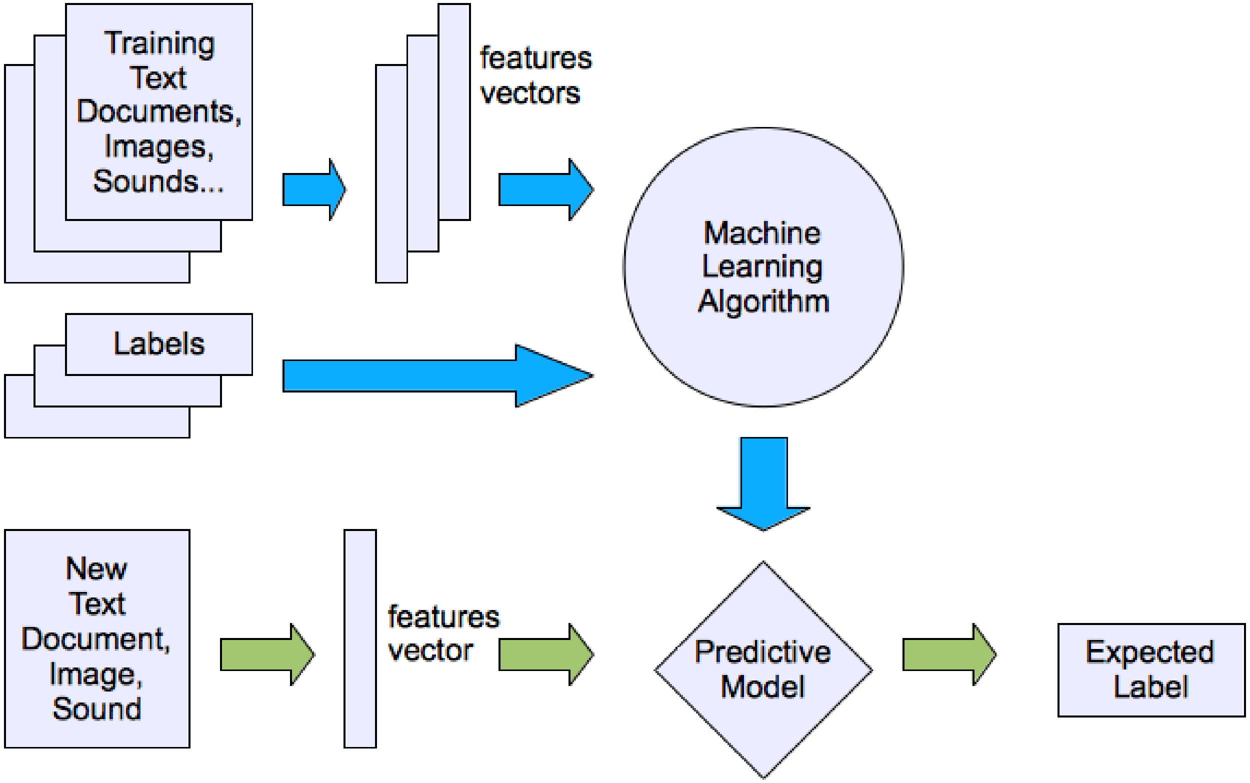
Learning Methods

	Supervised Learning	Unsupervised Learning
Continuous	**Labelled Data** (Regression)	**Un-labelled Data** (Dimensionality Reduction)
Discrete	**Labelled Data** (Classification or Categorization)	**Un-labelled Data** (Clustering)

- Other Learning methods
  - Reinforcement learning: e.g., game-playing agent
  - Learning to rank, e.g., document ranking in Web search
  - And many others….

Supervised Learning

- Learning a mapping (or a Predictive Model) from a set of inputs to a target variable
  - Classification: target variable is discrete (e.g., spam email)
  - Regression: target variable is real-valued (e.g., stock market)
  - Prediction



Supervised Learning

examples

label

label$_1$

label$_3$

label$_4$

label$_5$

labeled examples

Supervised learning: given labeled examples

Supervised Learning

label

label_1

label_3

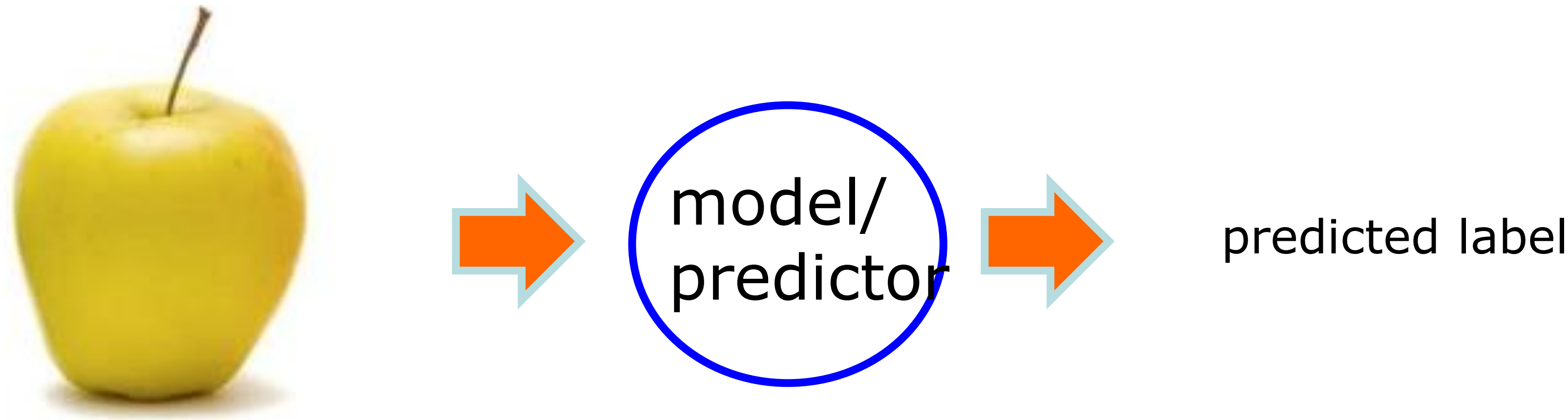
model/
predictor

label_4

label_5

Supervised learning: given labeled examples

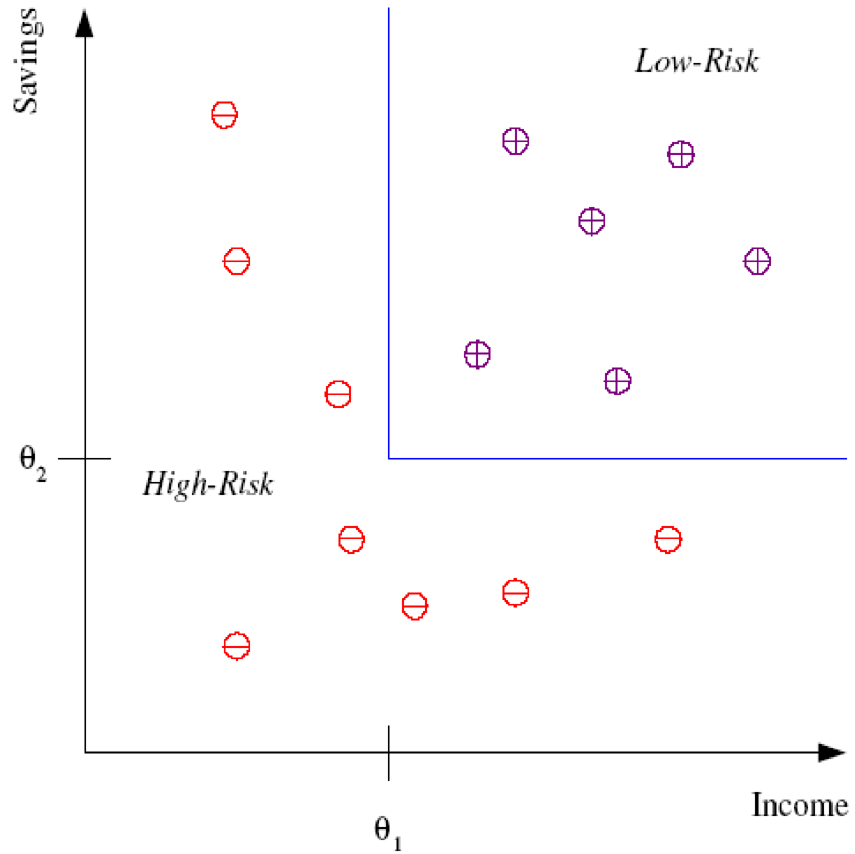
Supervised Learning



Supervised learning: learn to predict new example

Supervised Learning – Classification Example

Differentiate between low-risk and high-risk customers from their *income* and *savings*



Supervised Learning - Regression

label

-4.5

10.1

3.2

4.3

Regression: label is real-valued

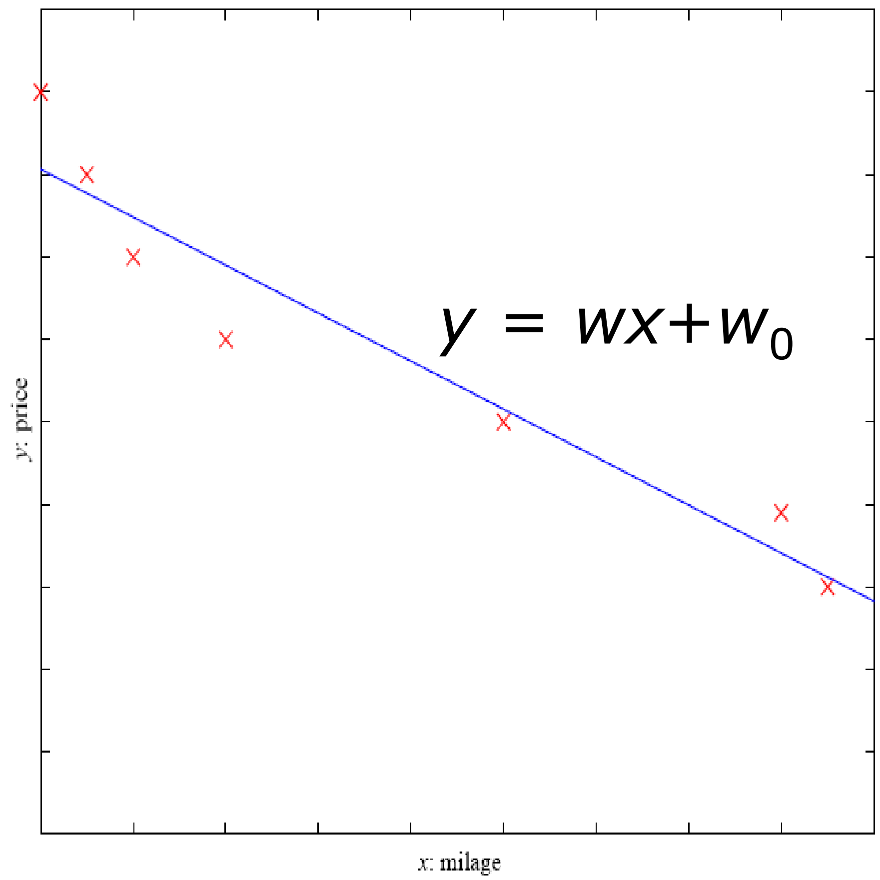
Supervised learning: given labeled examples

Regression Example

Price of a used car

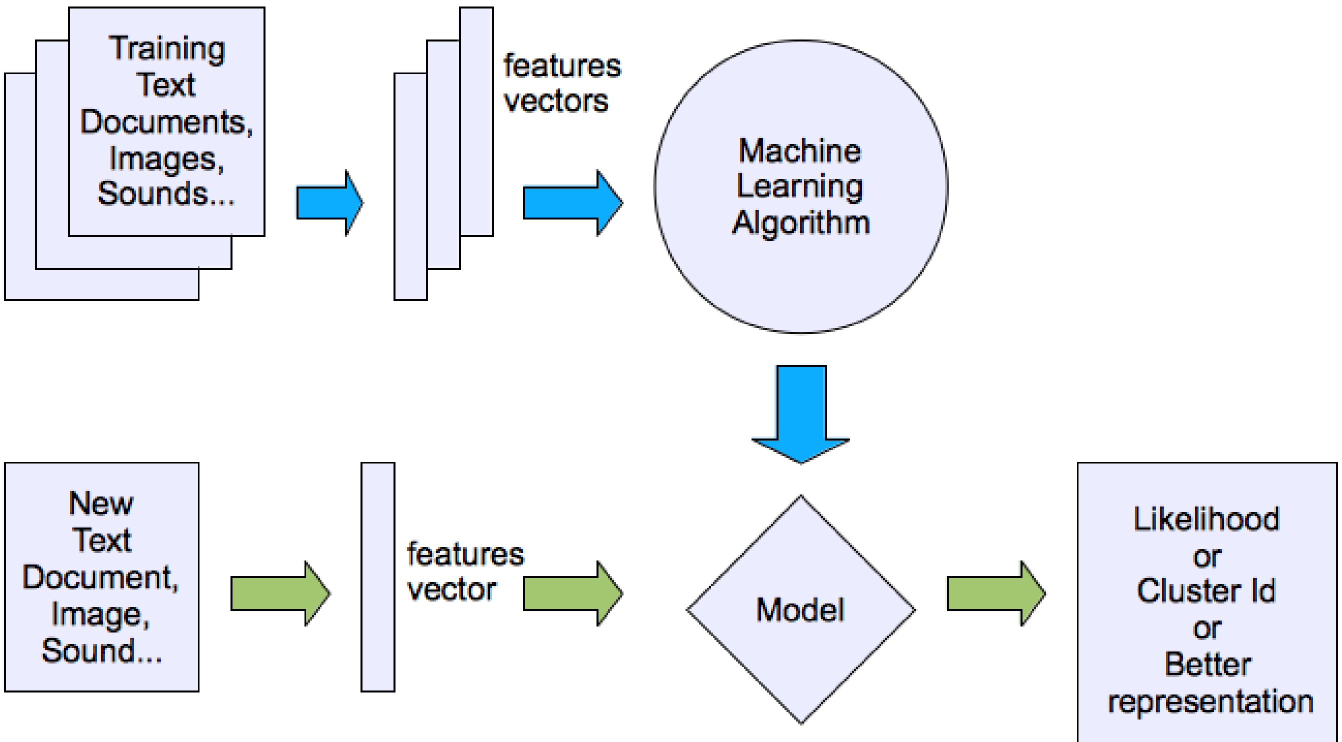
x : car attributes
(e.g. mileage)

y : price



Unsupervised Learning

- No target variable provided
- Clustering: grouping similar instances
  - grouping data into K groups
  - Dimensionality reduction
    - Principal component analysis (PCA)
    - Factor analysis



Unsupervised Learning

Un-supervised learning: given data, i.e. examples, but no labels

Un-supervised Learning - Applications

- Learn clusters / groups without any label
- Customer segmentation (i.e. grouping)
- Image compression
- Dimensionality reduction

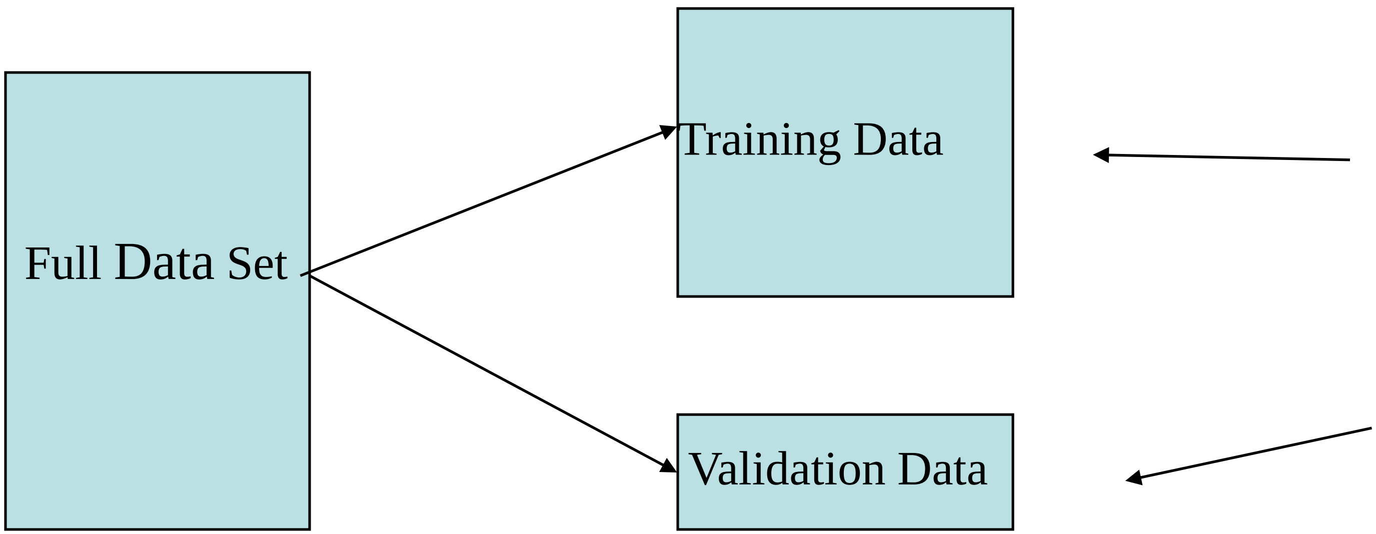
Reinforcement Learning

left, right, straight, left, left, left, straight	**GOOD**
left, straight, straight, left, right, straight, straight	**BAD**

left, right, straight, left, left, left, straight	**18.5**
left, straight, straight, left, right, straight, straight	**-3**

Given a *sequence* of examples/states and a *reward* after completing that sequence, learn to predict the action to take in for an individual example/state

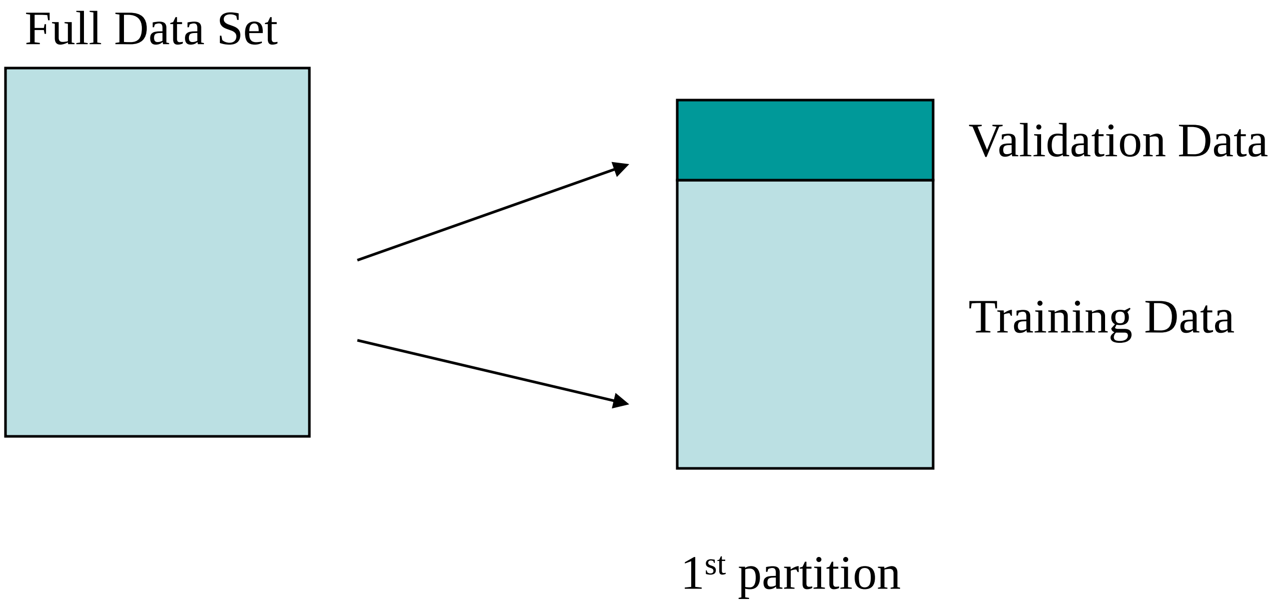
Training and Validation Data



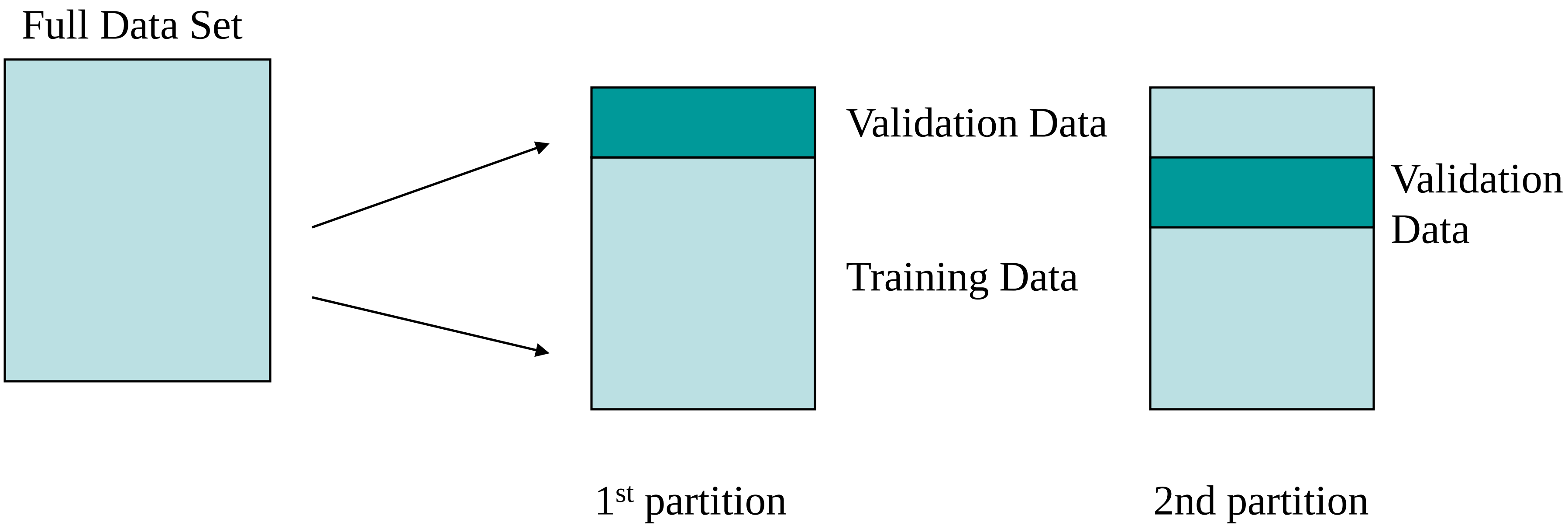
Idea: train each model on the "training data"

and then test each model's accuracy on the validation data

Disjoint Validation Data Sets



Disjoint Validation Data Sets



Issues in Machine Learning (i.e. Generalization)

- What algorithms are available for learning a concept? How well do they perform?
- How much training data is sufficient to learn a concept with high confidence?
- When is it useful to use prior knowledge?
- Are some training examples more useful than others?
- What are best tasks for a system to learn?
- What is the best way for a system to represent its knowledge?

Generalization

Training set (labels known)

Test set (labels unknown)

- How well does a learned model generalize from the data it was trained on to a new test set?

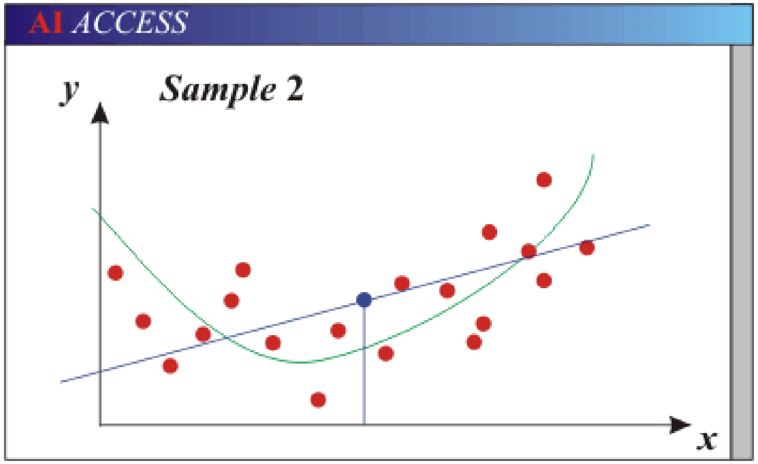
Generalization

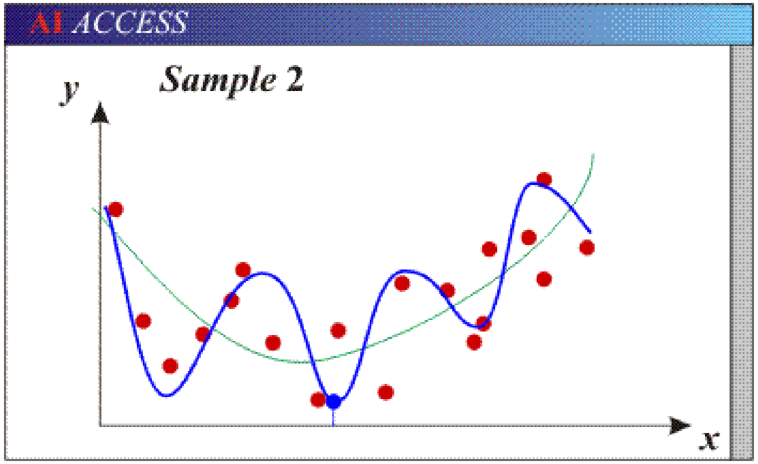
- Components of generalization error
  - **Bias:** how much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - **Variance:** how much models estimated from different training sets differ from each other

- **Under fitting:** model is too "simple" to represent all the relevant class characteristics
  - High bias and low variance
  - High training error and high test error

- **Over fitting:** 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 Tradeoff





- 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).

Next Presentation

Application of Machine Learning on Networking and Communication Systems