INFO 371
Introduction to Data Science

Spring 2018

What is a statistical model?

The gender gap in 100-meter dash times

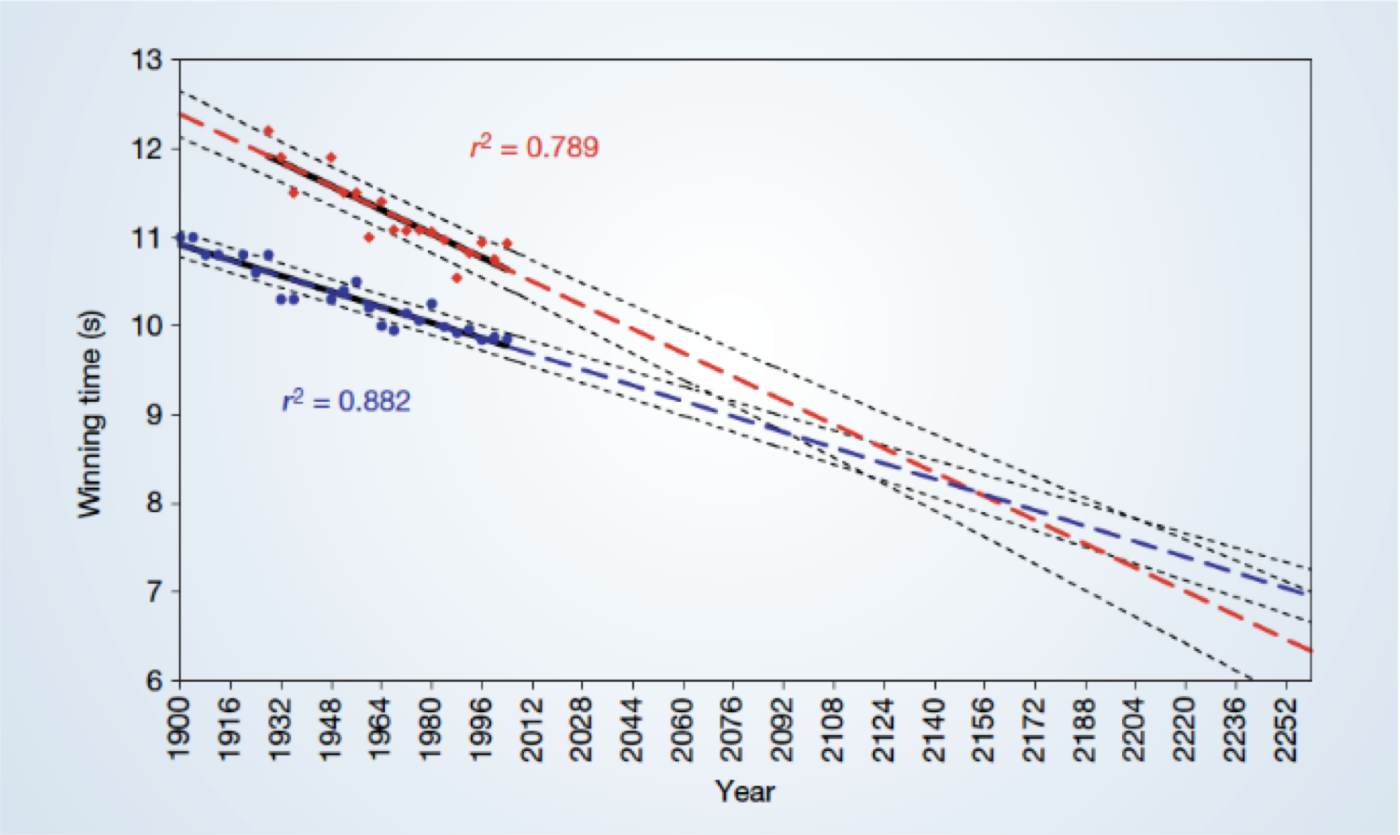


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

http://callingbullshit.org/case_studies/case_study_gender_gap_running.html

Good model?

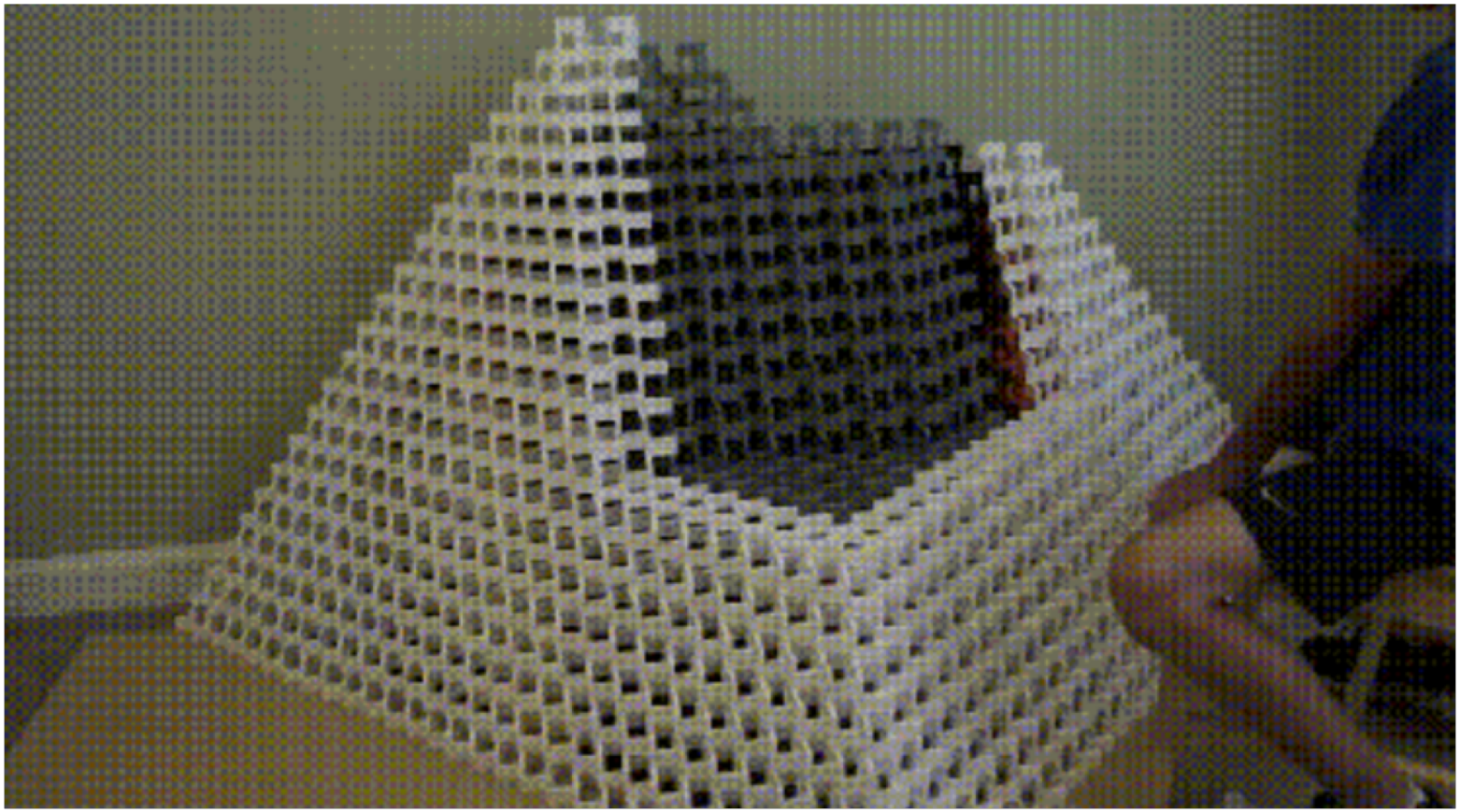
- Let's assume we have a linear model to predict income from education
- Suppose there are 100k observations to fit the model
- Is this a "good" model?
- Why/why not?

Features →

Observations ↓

	Age	Education	Income
0	35	8	30942
1	23	8	37323
2	58	8	49381
3	41	5	31680
4	35	13	81147
5	43	9	38682
6	35	8	34632
7	56	7	14394
8	62	11	22243

Is the model “right”?



http://proteas.microlab.ntua.gr/ksiop/phd_funny/index.html

Is the model “right”?

- “All models are wrong but some are useful.”

George Box
1919 – 2013

Statistical Models

- “All models are limited by the validity of the assumptions on which they ride.”
- “Assumptions behind models are rarely articulated, let alone defended.”

David Freedman (1938 – 2008)

Be careful!

- Our R^2 increases as we
  - Add complexity
  - Iterate on features
  - Try different models
  - User different datasets
- **Good fit is not the same a good model!**

Determining model accuracy

- Remember, in supervised learning, we have labels!
- So, how do we evaluate performance?
  - Different cost functions. Examples:
    - Sum of absolute residuals
    - Sum of residuals (less sensitive to outliers)
    - Sum of cubed residuals (more sensitive to outliers)
    - Sum of weighted residuals
  - It all depends on context and what you are trying to achieve

Regression recap

- What is the formula? (the model?)
- What is the evaluation criteria? (what do we generally minimize?)
- What is the optimization? (how do we approach this minimization?)

The problem

- Suppose you work at LinkedIn. Your boss asks you to create a model that will predict people's salaries based on their age and years of education
- Assume the requirements for a linear model are more or less fulfilled, so you create a linear regression model based on education along first

Regression – an alternate view

- Two paradigms:
  - Statistics/econometrics
    - Explaining relationships
    - Understanding causality
    - Answering: Does more education lead to higher wages?
  - ML/Computer Science
    - Predictions
    - Looking at more generalizable patterns
    - Answering: How much would I earn if I left the university?
- Prediction doesn't rely on any causal underpinnings
  - Econometrics more likely to focus on inferences; ML on predictions

DOI:10.1145/2347736.2347755

Tapping into the "folk knowledge" needed to advance machine learning applications.

BY PEDRO DOMINGOS

A Few Useful Things to Know About Machine Learning

MACHINE LEARNING SYSTEMS automatically learn programs from data. This is often a very attractive alternative to manually constructing them, and in the last decade the use of machine learning has spread rapidly throughout computer science and beyond. Machine learning is used in Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other applications. A recent report from the McKinsey Global Institute asserts that machine learning (a.k.a. data mining or predictive analytics) will be the driver of the next big wave of innovation.[15] Several fine textbooks are available to interested practitioners and researchers (for example, Mitchell[16] and Witten et al.[24]). However, much of the "folk knowledge" that is needed to successfully develop machine learning applications is not readily available in them. As a result, many machine learning projects take much longer than necessary or wind up producing less-than-ideal results. Yet much of this folk knowledge is fairly easy to communicate. This is the purpose of this article.

» key insights

- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples. This is often feasible and cost-effective where manual programming is not. As more data becomes available, more ambitious problems can be tackled.
- Machine learning is widely used in computer science and other fields. However, developing successful machine learning applications requires a substantial amount of "black art" that is difficult to find in textbooks.
- This article summarizes 12 key lessons that machine learning researchers and practitioners have learned. These include pitfalls to avoid, important issues to focus on, and answers to common questions.

Predicting Student Dropout in Higher Education

Lovenoor (Lavi) Aulck (*laulck@uw.edu*)
Nishant Velagapudi, Joshua Blumenstock, Jevin West

International Conference on Machine Learning's #Data4Good
June 24, 2016

Learning Objectives

1. Introduction to ML jargon
2. Distinguish between supervised and unsupervised learning
3. Appreciate overfitting
4. Be able to describe the roles that representation, evaluation and optimization play in ML

Outline

- **Introduction to Machine Learning**
- Supervised vs. Unsupervised Learning
- Key Issues in (Supervised) Machine Learning
- Philosophical Interlude

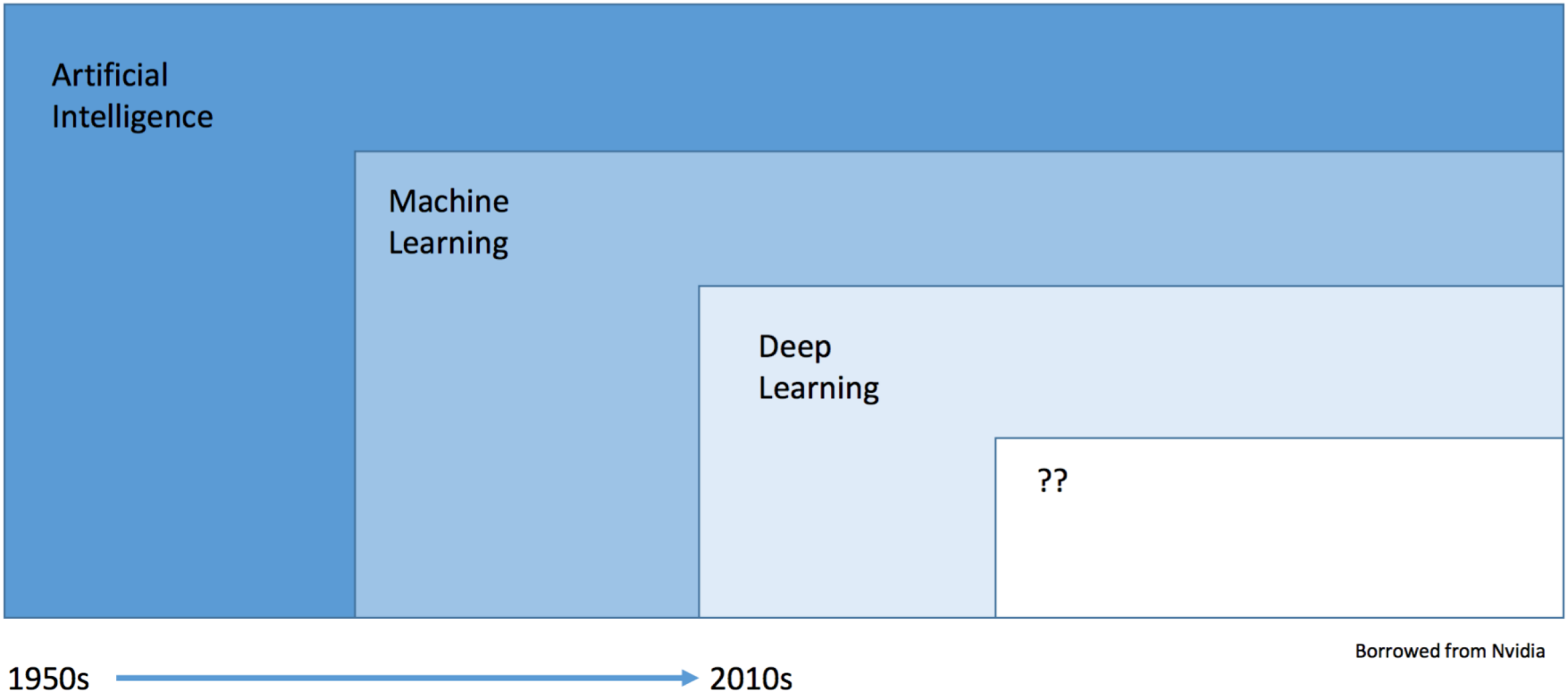
*These slides are developed based on lectures that Josh Blumenstock gave for INFX 574.

Machine Learning?

Machine Learning: Introduction

- **Machine learning** is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions.

What is ML?



Borrowed from Nvidia

Machine Learning: Context

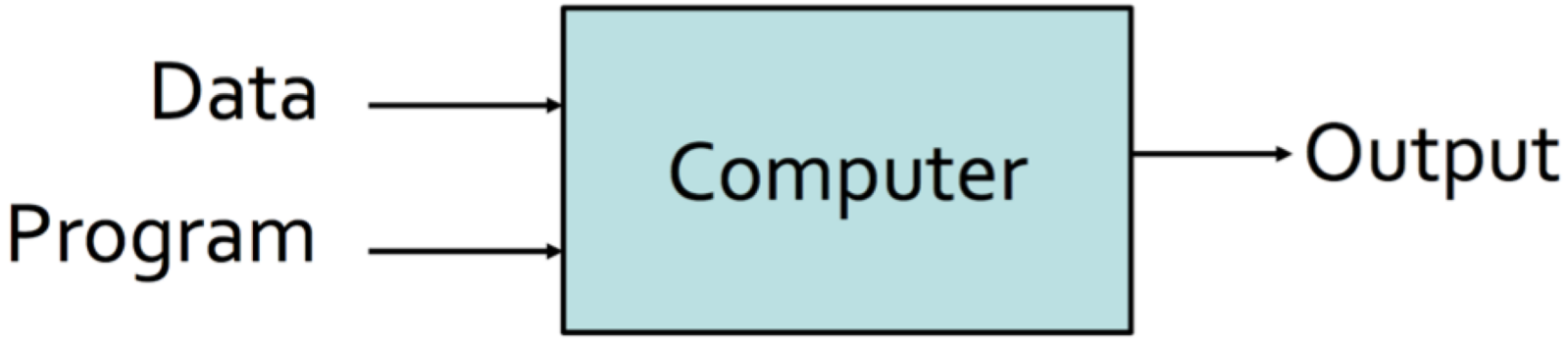
- Traditionally:
  - Computer Science
    - Artificial Intelligence: “the study and design of intelligent/rational agents”
      - Machine Learning: “Learning without explicitly programming” (Samuels 1959) - includes robotics, computer vision, agency, cognition)
        - » Data Mining

Machine Learning: Key Concepts

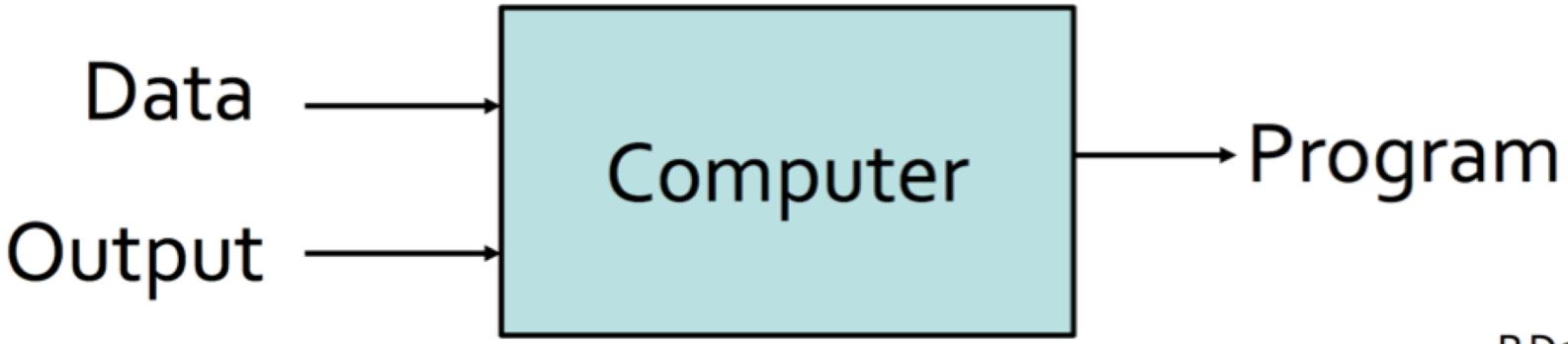
- Representation
- Evaluation
- Optimization
- Supervised Learning
- Unsupervised Learning
- Overfitting
- Generalization

What is Machine Learning?

- Traditional Programming



- Machine Learning



P.Domingos

What is Machine Learning?

- “Learning is more like farming, which lets nature do most of the work. Farmers combine seeds with nutrients to grow crops. Learners combine knowledge with data to grow programs.” –Pedro Domingos, CACM

ML in a Nutshell

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

- Tens of thousands of ML algorithms
- Every ML algorithm has three primary components
  - **Representation** (i.e., model)
  - **Evaluation** (i.e., an objective function)
  - **Optimization** (e.g., search)

P. Domingos

Representation/Model

- "Choosing a representation for a learner is tantamount to choosing the set of classifiers that it can possibly learn. This set is called the *hypothesis space* of the learner."
  - Decision trees
  - Instance-based
  - Neural Networks
  - Support Vector Machines
  - Probabilistic (graphical) models
  - Model Ensembles
  - Etc.
- "As we will see, some choices in a ML project may be even more important than the choice of learner"

P. Domingos

Evaluation

- Is our model effective?
  - Accuracy
  - Precision and recall
  - Squared error
  - Likelihood
  - Cost/Utility
  - Margin
  - Entropy, K-L divergence, etc.

Optimization/Search

- How to improve?
  - Combinatorial optimization (discrete)
    - E.g., Greedy search
  - Convex optimization (continuous)
    - E.g., Gradient descent
  - Constrained optimation
    - E.g., Linear Programming

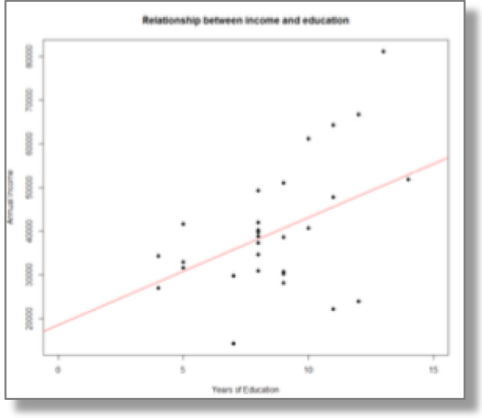
Outline

- Introduction to Machine Learning
- **Supervised vs. Unsupervised Learning**
- Key Issues in (Supervised) Machine Learning
- Philosophical Interlude

Supervised vs. Unsupervised

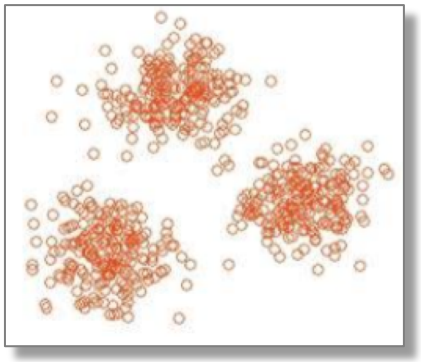
- Generally, two “flavors” of machine learning (semi-supervised also exists which combines the two)
- Key distinction:
  - Whether or not you know the “right” answer

Supervised Learning

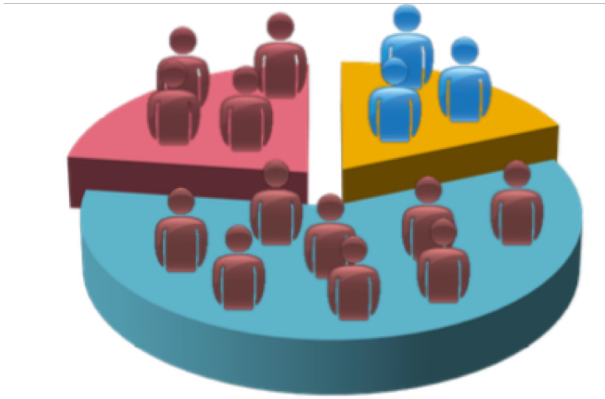


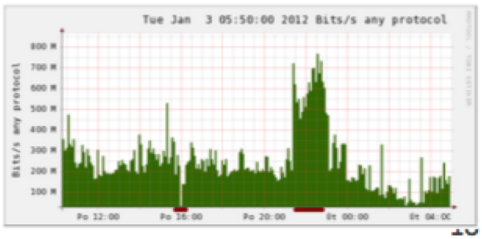
- We know the “right answer” for some values
  - Goal is typically to model the relationship between input (independent) variables and know output (dependent) variables
- Examples
  - Classifying hand-written numbers from images
  - Identifying presidential candidates based on their speeches
  - Determining whether a student will graduate from UW
  - Disease classification
  - Credit scoring
  - Etc.
- Methods
  - Decision Trees
  - Linear models (regression, logistic regression, SVM)
  - Naïve Bayes
  - Ensemble methods
  - Neural networks

Unsupervised Learning



- We don't now the "right answer", the right groupings, or "ground truth"
  - Goal is typically to discover underlying structure in the data
  - Often more explanatory than supervised learning, which is more directed
- Examples
  - Market segmentation, disease classification
  - Visualizing complex data
  - Network clustering
- Methods
  - K-means and hierarchical clustering
  - Principal Component Analysis (PCA)
  - SVD, NMA, LDA





Other approaches to ML

- Semi-supervised learning
  - We have some labeled instances
- Reinforcement learning
  - Learning by interacting with an environment
  - Rewards from sequence of actions

Pop quiz

- Of the following examples, which would you address using an unsupervised learning algorithm?
  - Given email labeled as spam/not spam, learn a spam filter
  - Given a set of news articles found on the web, group them into a set of articles about the same story
  - Given a database of customer data, automatically discover market segments and group customers into different market segments
  - Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not
  - Given phone records of individuals and survey data about their income, predict the incomes of new subscribers