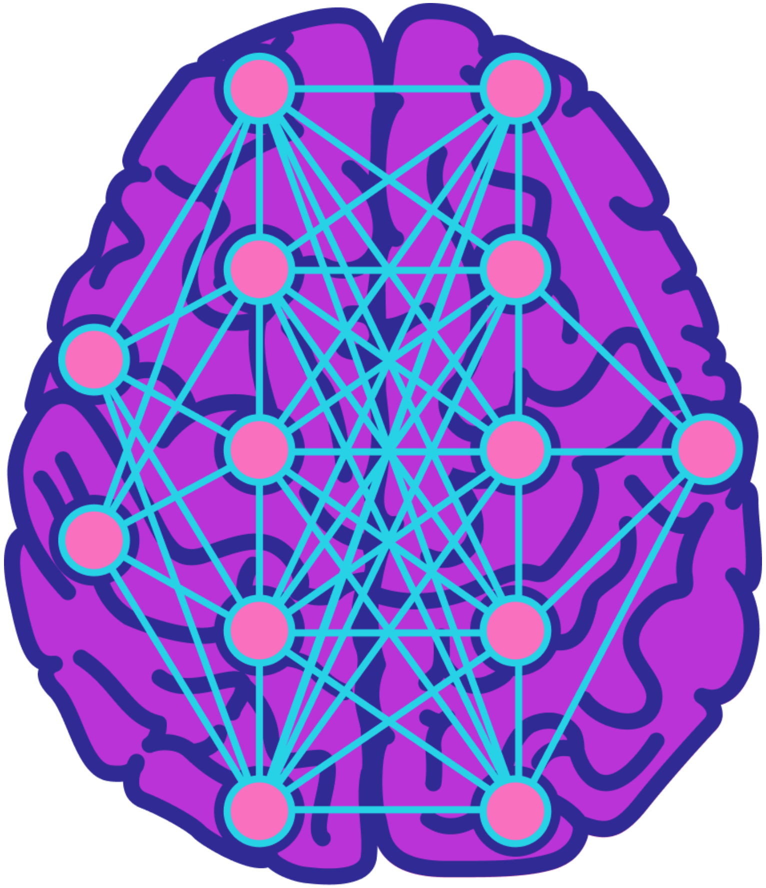
DS 4400

Machine Learning and Data Mining I

Alina Oprea
Associate Professor, CCIS
Northeastern University

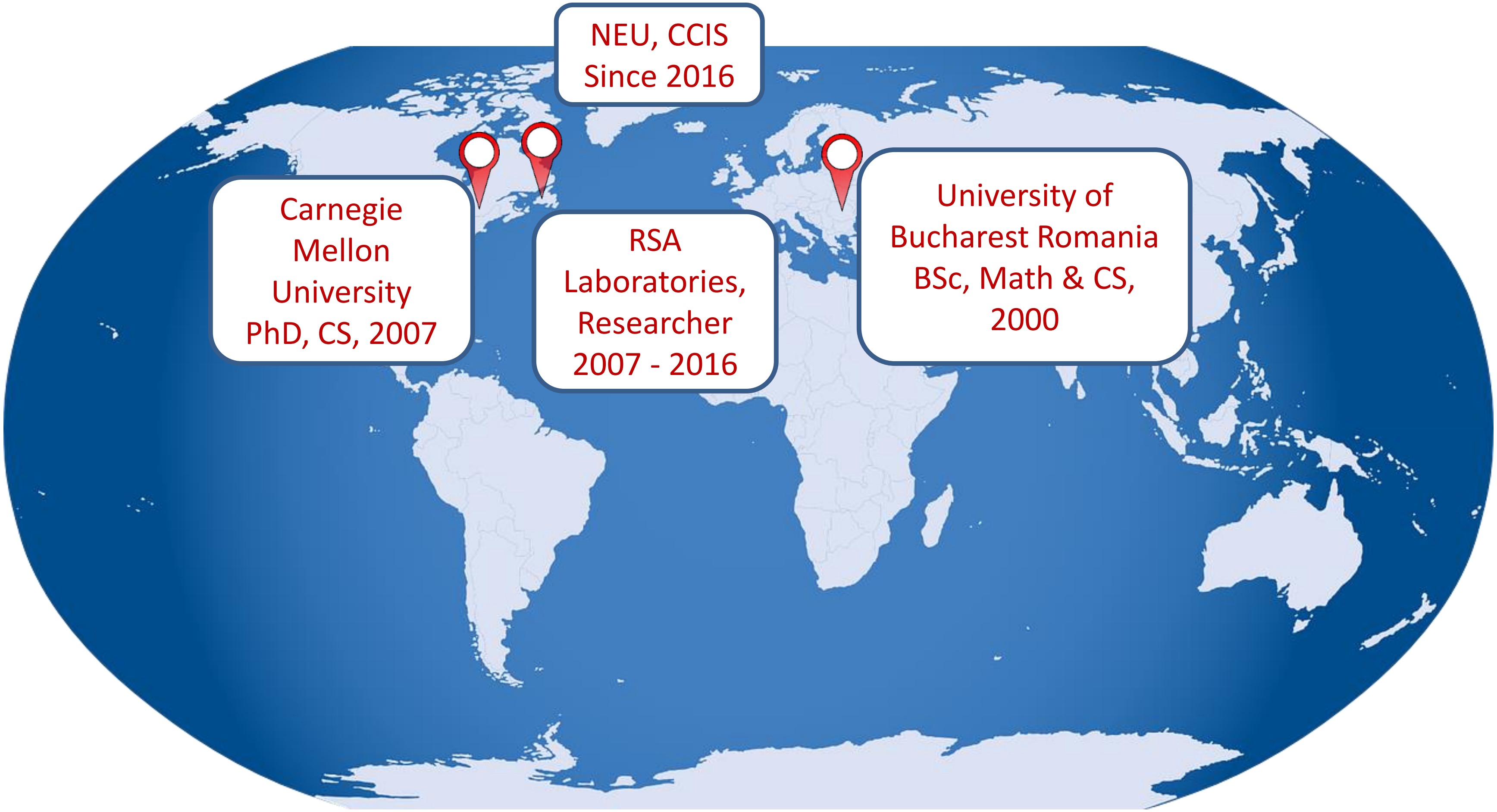
September 6 2018

Welcome to DS 4400!



Machine Learning and Data Mining I

Introductions



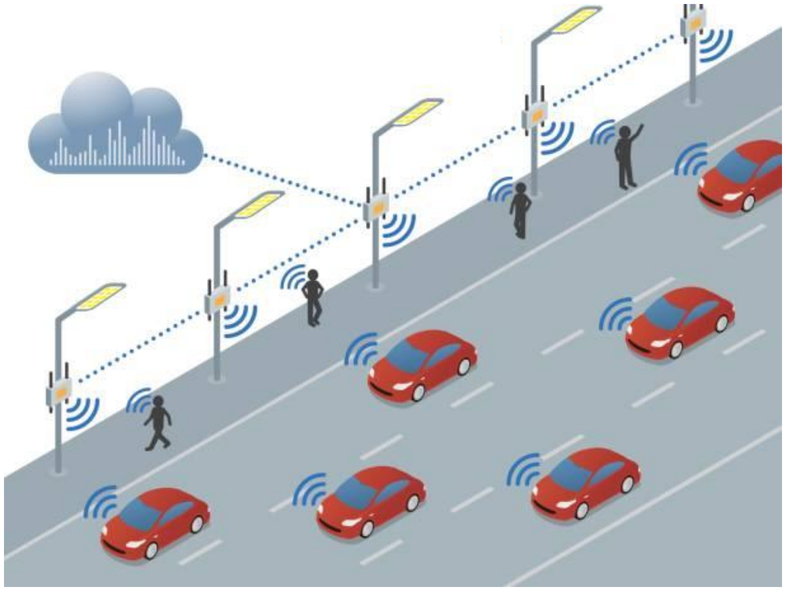
Background

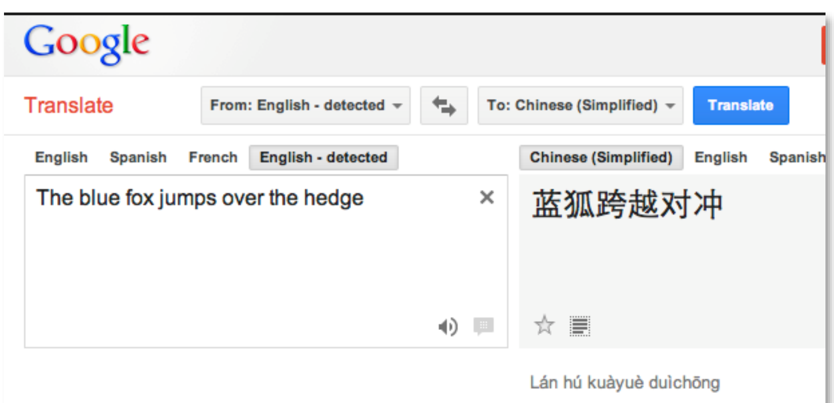
- Ph.D. at CMU
  - Research in storage security & cryptographic file systems
- RSA Laboratories
  - Cloud security, applied cryptography
  - Security analytics (ML in security)
- NEU CCIS – since Fall 2016
  - ML for security applications (threat detection, IoT, fuzzing)
  - Adversarial ML

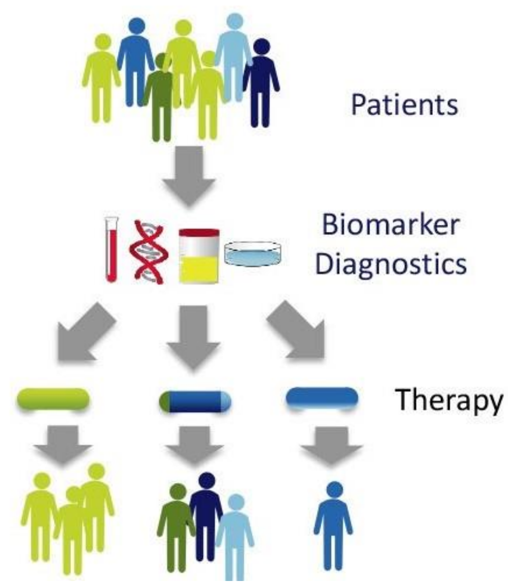
Class Introductions

- Enrollment of 13

Machine learning is everywhere



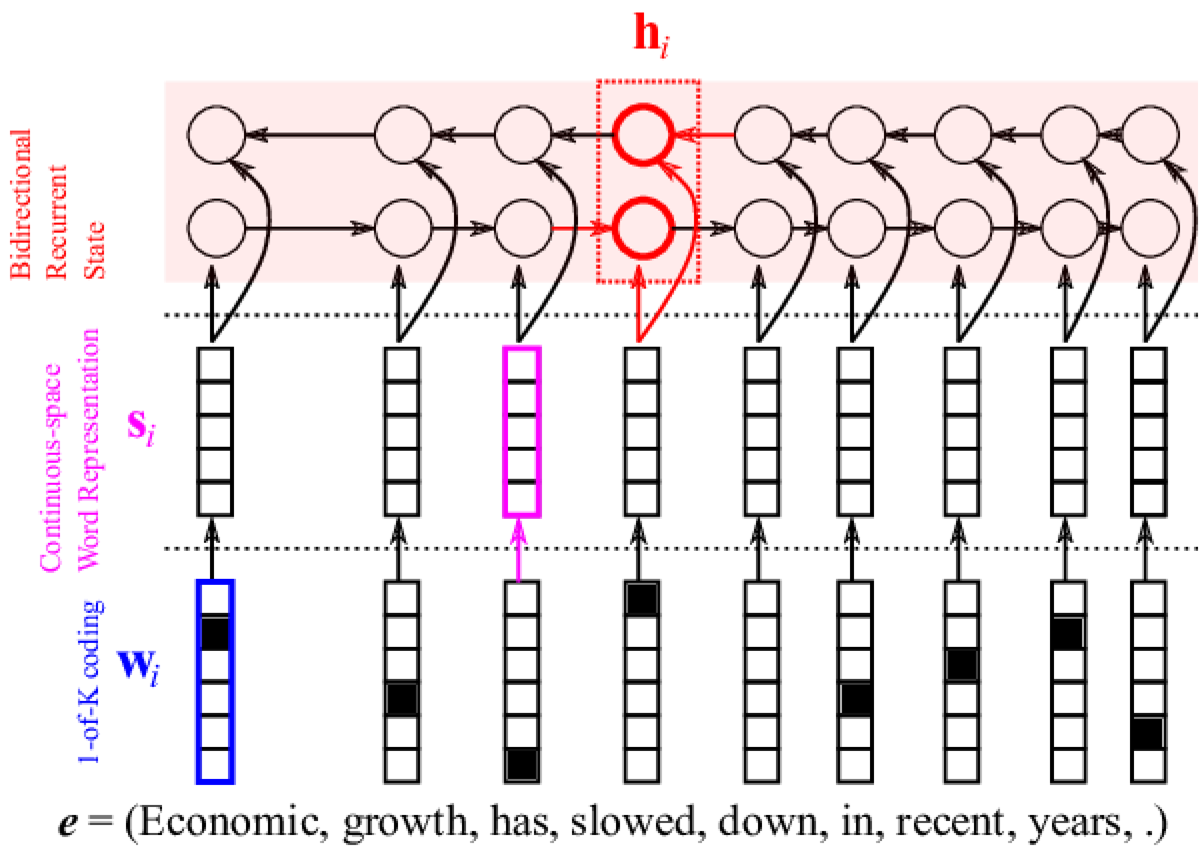




DS-4400

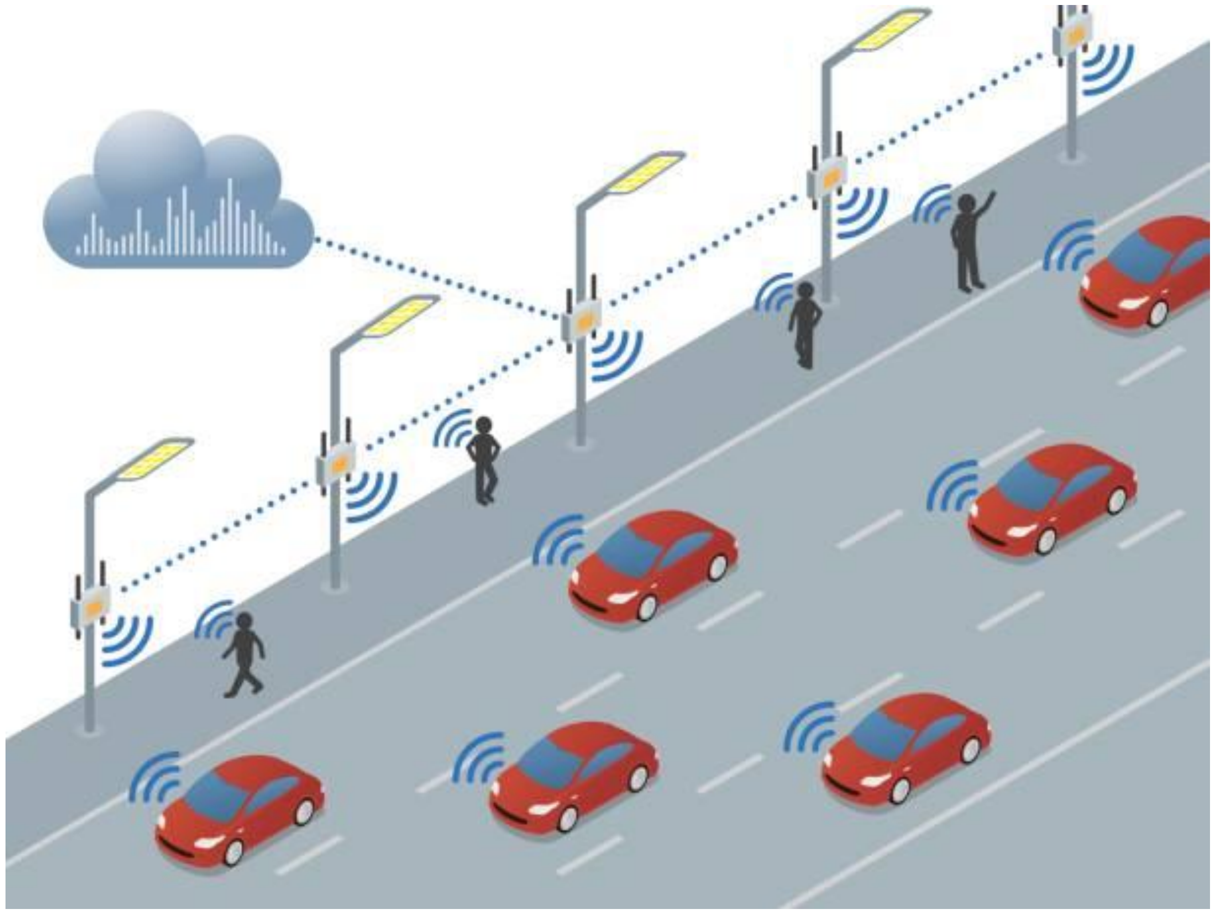
- What is *machine learning*?
  - The science of teaching machines how to learn
  - Design predictive algorithms that learn from data
  - Replace humans in critical tasks
  - Subset of AI
- Machine learning very successful in:
  - Machine translation
  - Precision medicine
  - Recommendation systems
  - Self-driving cars
- Why the hype?
  - Availability: data created/reproduced in 2010 reached 1,200 exabytes
  - Reduced cost of storage
  - Computational power (cloud, multi-core CPUs, GPUs)

Natural Language Processing (NLP)



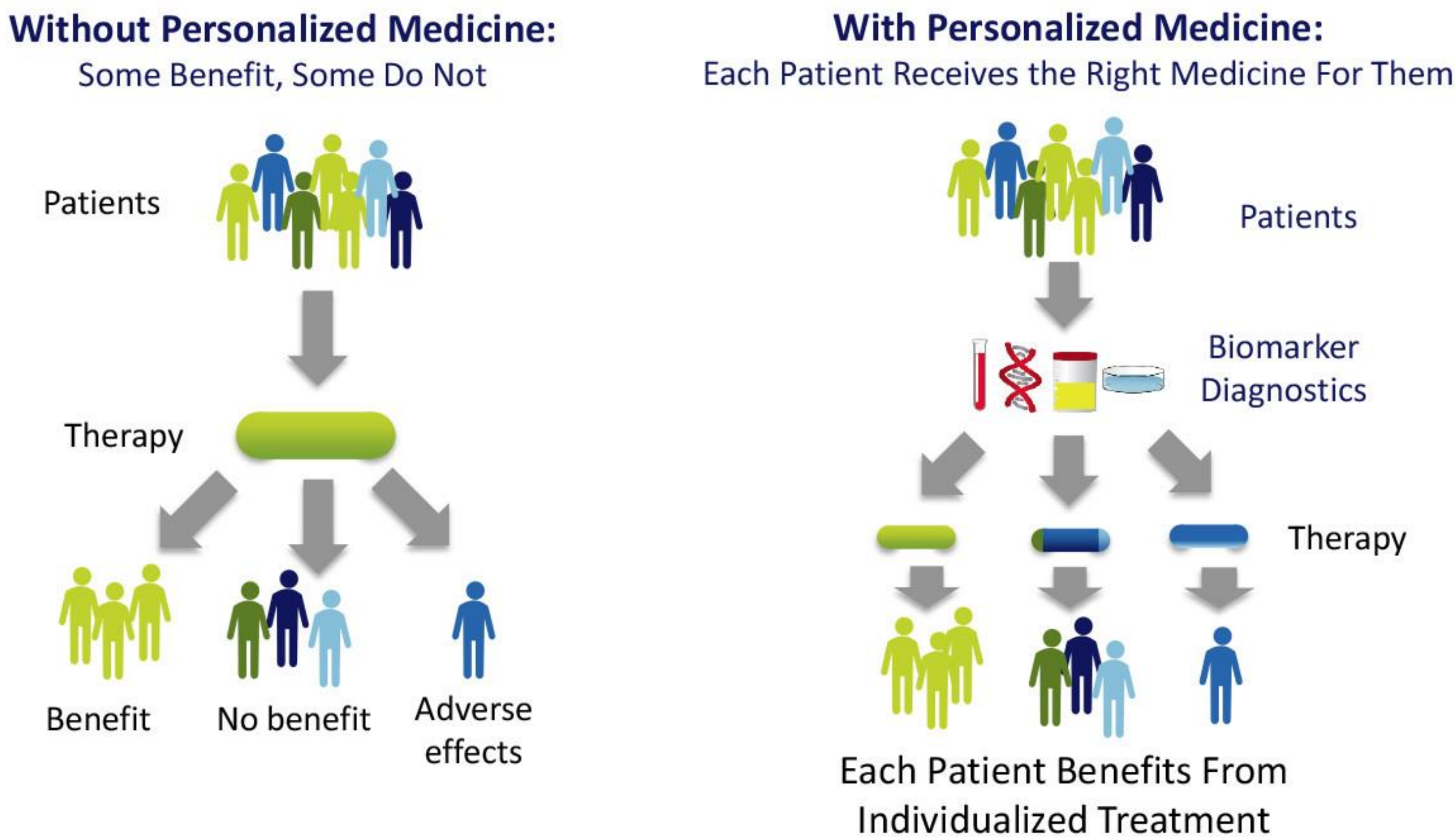
- Understand language semantics
- Real-time translation, speech recognition

Autonomous vehicles



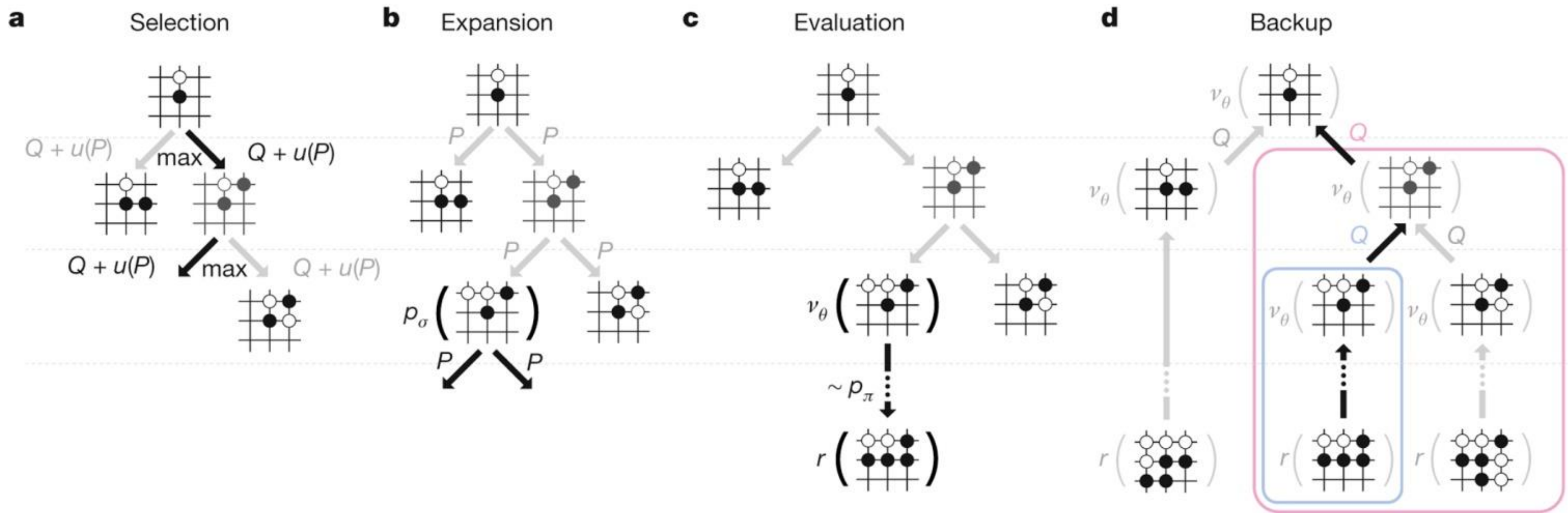
- Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication
- Assist drivers in making decisions to increase safety

Personalized medicine



- Treatment adjusted to individual patients
- Predictive models using a variety of features related to patient history and genetics

Playing games



Reinforcement learning

- AlphaGo
- Chess

DS-4400 Course objectives

- Become familiar with machine learning tasks
  - Supervised learning vs unsupervised learning
  - Classification vs Regression vs Clustering
- Study most well-known algorithms and understand to which problem they apply
  - Regression (linear regression)
  - Classification (SVM, decision trees, neural networks)
  - Clustering (k-means)
- Learn to apply ML algorithms to real datasets
  - Using existing packages in R and Python
- Learn about security challenges of ML
  - Introduction to adversarial ML

http://www.ccs.neu.edu/home/alina/classes/Fall2018/

Class Outline

- Introduction – 1 week
  - Probability and linear algebra review
- Supervised learning - 5 weeks
  - Linear regression
  - Classification (logistic regression, LDA, kNN, decision trees, random forest, SVM, Naïve Bayes)
  - Model selection, regularization, cross validation
- Neural networks and deep learning – 1.5 weeks
  - Back-propagation, gradient descent
  - NN architectures
- Unsupervised learning – 2.5 weeks
  - Dimensionality reduction (PCA)
  - Clustering (k-means, hierarchical)
- Adversarial ML – 1 week
  - Security of ML at testing and training time

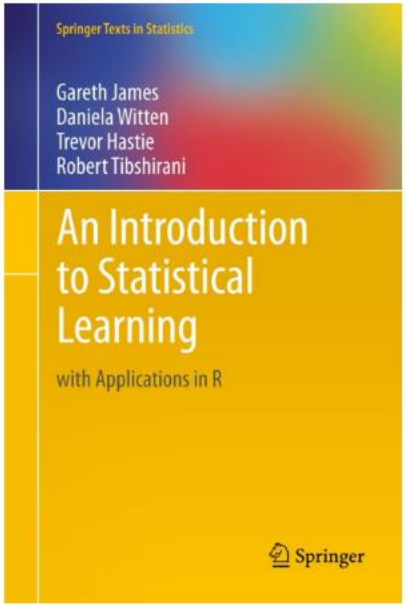
Textbook

An Introduction to Statistical Learning

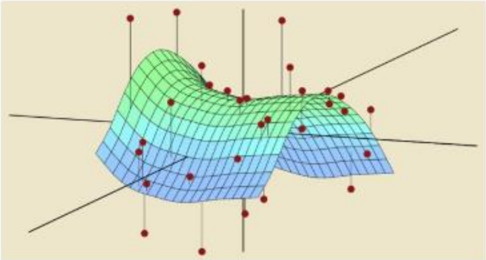
with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

- Home
- About this Book
- R Code for Labs
- Data Sets and Figures
- ISLR Package
- Get the Book
- Author Bios
- Errata



Download the book PDF
(corrected 7th printing)



Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Start anytime in self-paced mode.

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students

Policies

- Instructors
  - Alina Oprea
  - TA: Anand Lad
- Schedule
  - Tue 11:45am – 1:25pm, Thu 2:50-4:30pm; Ryder Hall 158
  - Office hours:
    - Alina: Thu 4:30 – 6:00 pm (ISEC 625)
    - Anand: Tue 2-3pm (ISEC 605)
- Your responsibilities
  - Please be on time and attend classes
  - Participate in interactive discussion
  - Submit assignments/ programming projects on time
- Late days for assignments
  - 5 total late days, after that loose 20% for every late day
  - Assignments are due at 11:59pm on the specified date
- Respect university code of conduct
  - No collaboration on homework / programming projects
  - http://www.northeastern.edu/osccr/academic-integrity-policy/

Grading

- Assignments – 20%
  - 4-5 assignments based on studied material in class, including programming exercises
  - Language: R or Python; Jupyter notebooks
- Final project – 25%
  - Select your own project based on public dataset
  - Submit short project proposal and milestone
  - Presentation at end of class (10 min) and report
- Exams – 50%
  - Midterm – 25%
  - Final exam – 25%
- Class participation – 5%
  - Participate in class discussion and on Piazza

Outline

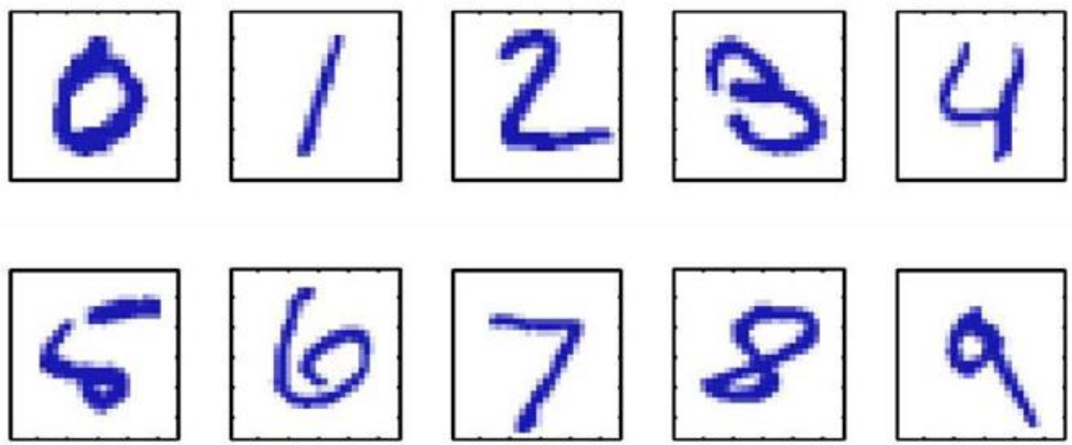
- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
- Bias-Variance Tradeoff
- Occam's Razor

Slides adapted from
- A. Zisserman, University of Oxford, UK
- S. Ullman, T. Poggio, D. Harari, D. Zysman, D Seibert, MIT
- D. Sontag, MIT
- Figures from "An Introduction to Statistical Learning", James et al.

Introduction

- What is Machine Learning?
  - Subset of AI
  - Design algorithms that learn from real data and can automate critical tasks
- When can it be applied?
  - It cannot solve any problem!
  - When task can be expressed as learning task
  - When high-quality data is available
    - Labeled data (by human experts) is preferable!
  - When some error is acceptable (can rarely achieve 100% accuracy)
    - Example: recommendation system, advertisement engine

Example 1
Handwritten digit recognition



Images are 28 x 28 pixels

Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \to \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Predict the digit

Multi-class classifier

Supervised Learning: Overview

Hypothesis space

Functions $\mathcal{F}$

$f : \mathcal{X} \rightarrow \mathcal{Y}$

Training data

$\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}$

Training

LEARNING

$\boxed{\text{find } \hat{f} \in \mathcal{F} \quad \text{s.t. } y_i \approx \hat{f}(x_i)}$

Learning machine

New data

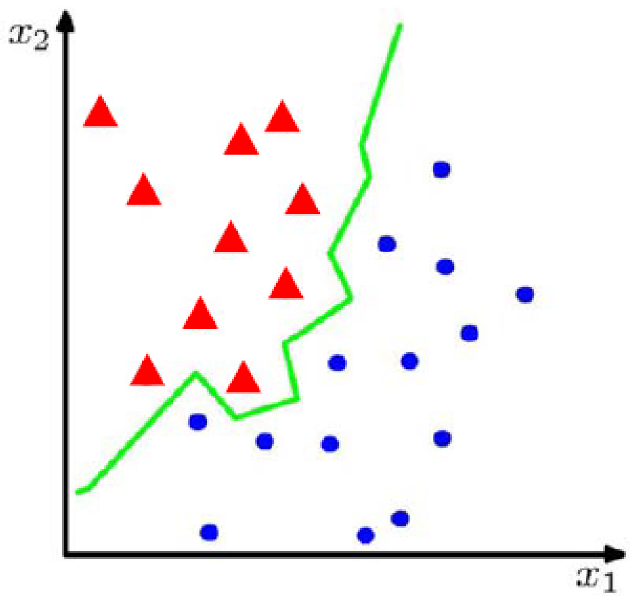
Testing

PREDICTION

$\boxed{y = \hat{f}(x)} \Leftarrow \boxed{x}$

$\hat{f}$ model

Classification



Binary

- Suppose we are given a training set of N observations

$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N), x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

- Classification problem is to estimate f(x) from this data such that

$$f(x_i) = y_i$$

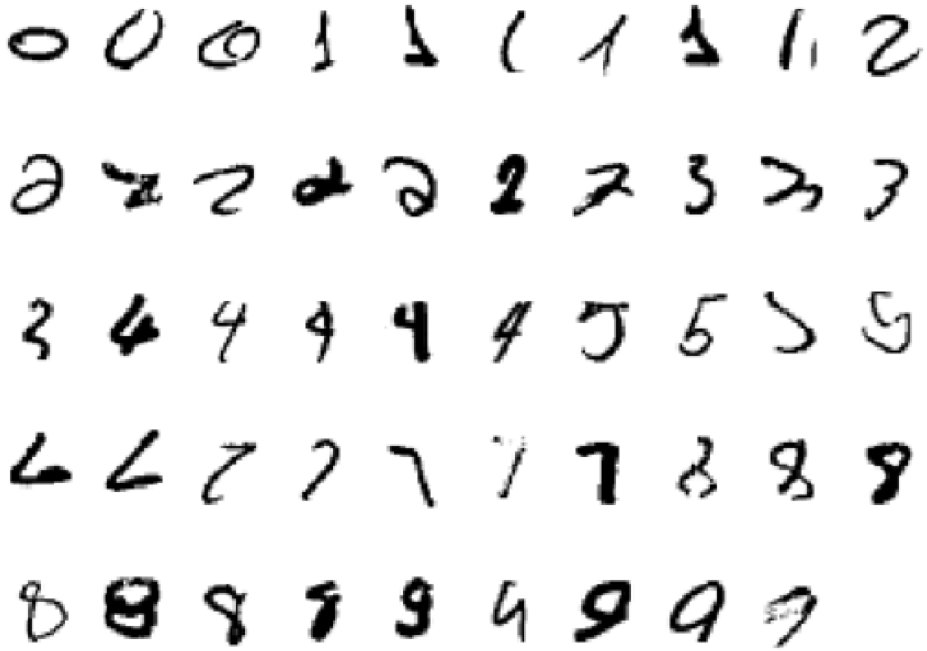
Extended to multi-label classification

- handwritten digit recognition

Model the problem

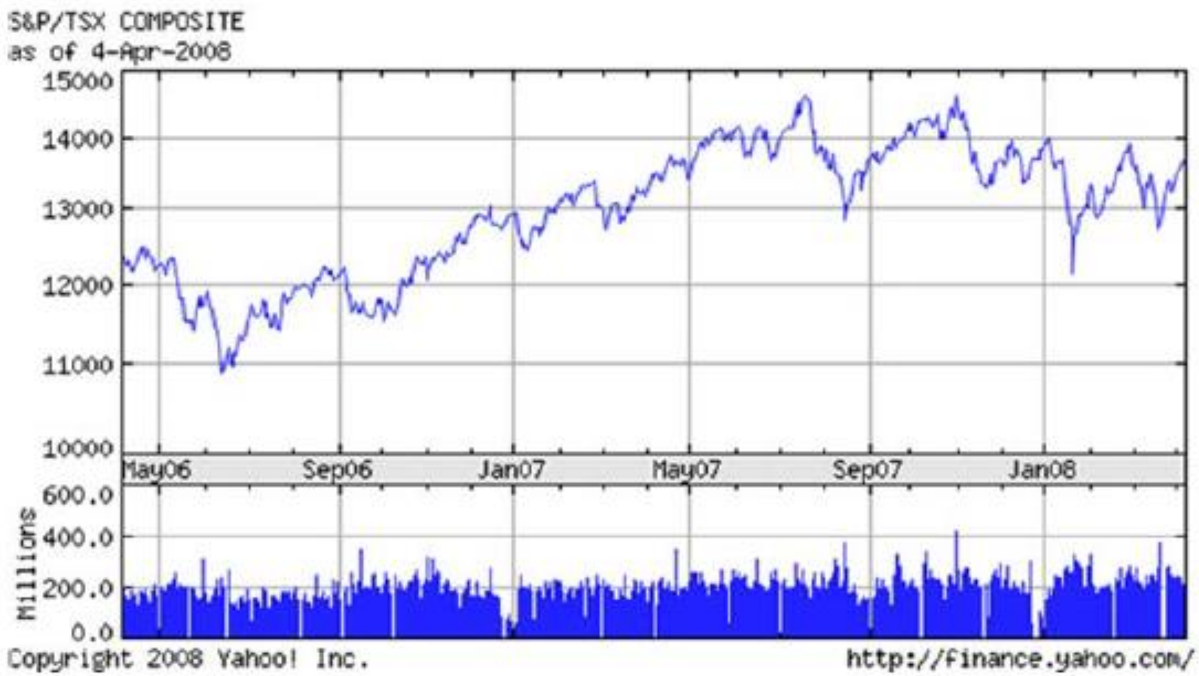
As a supervised classification problem

Start with training data, e.g. 6000 examples of each digit



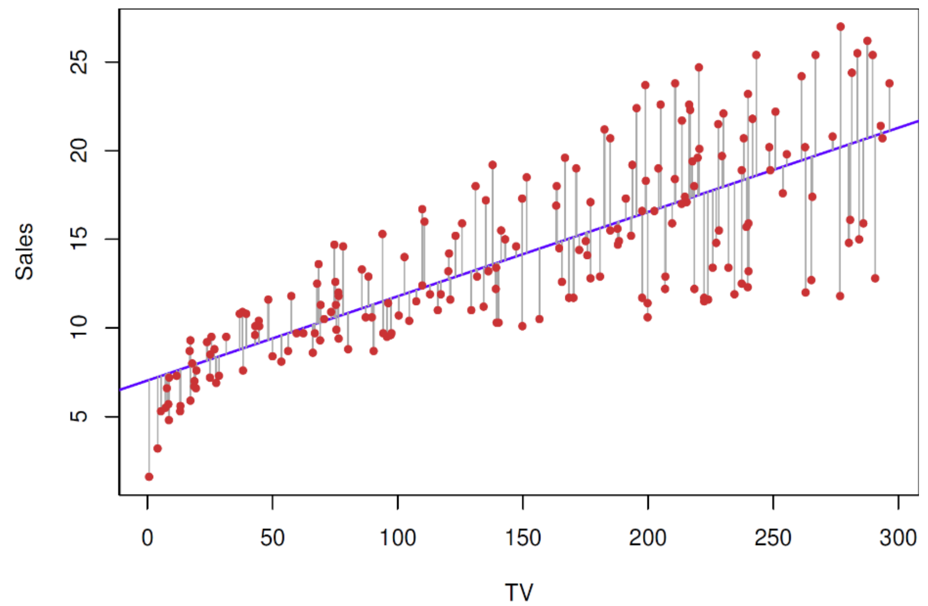
- Can achieve testing error of 0.4%
- One of first commercial and widely used ML systems (for zip codes & checks)

Example 2
Stock market prediction



- Task is to predict stock price at future date
- This is a regression task, as the output is continuous

Regression



Linear regression
1 dimension

- Suppose we are given a training set of N observations

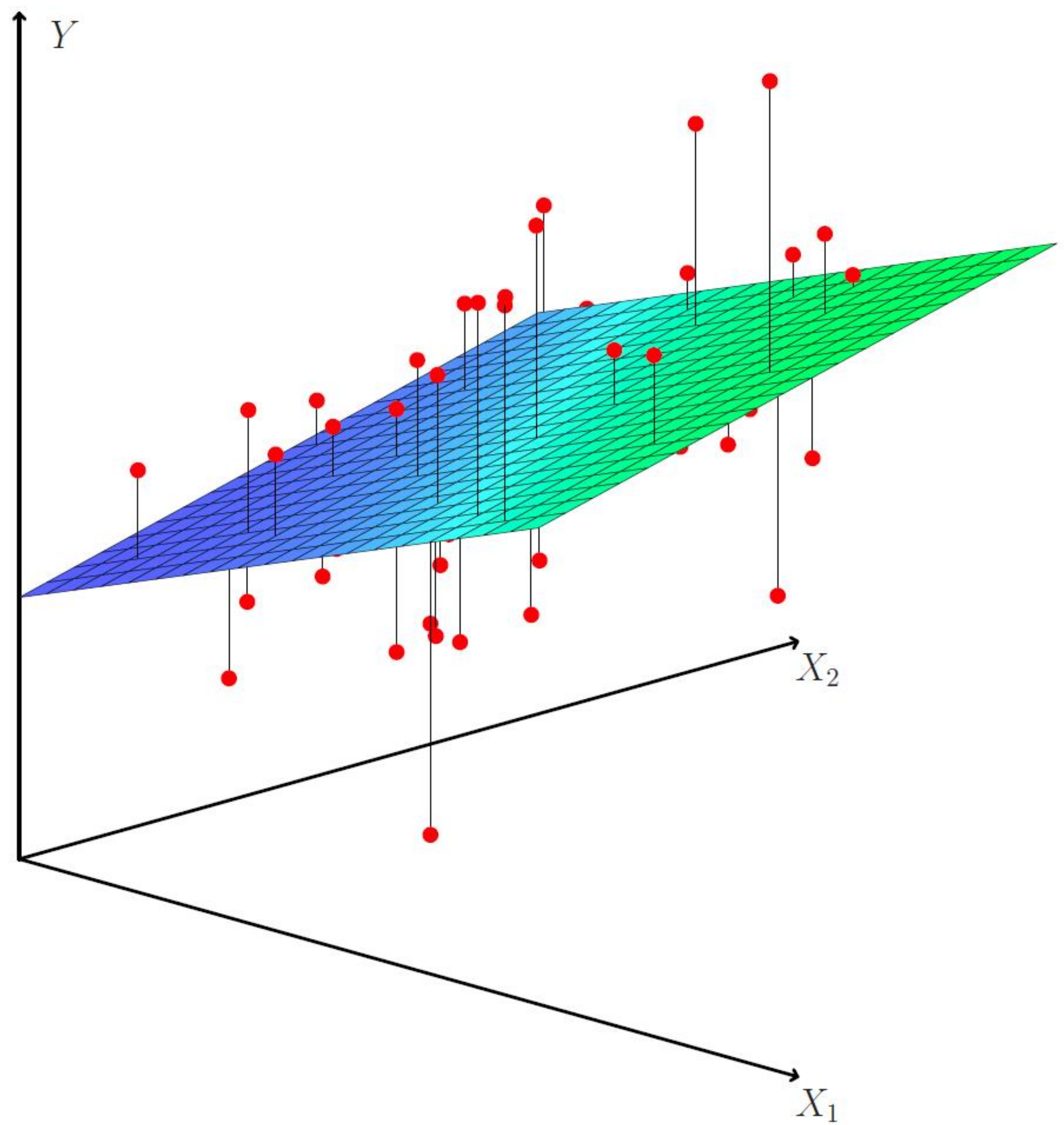
$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N)$$

- Regression problem is to estimate y(x) from this data

$x_i = (x_{i1}, \ldots x_{id})$ - d predictors (features)

y_i - response variable

Multi-dimensional linear regression



Wage Prediction

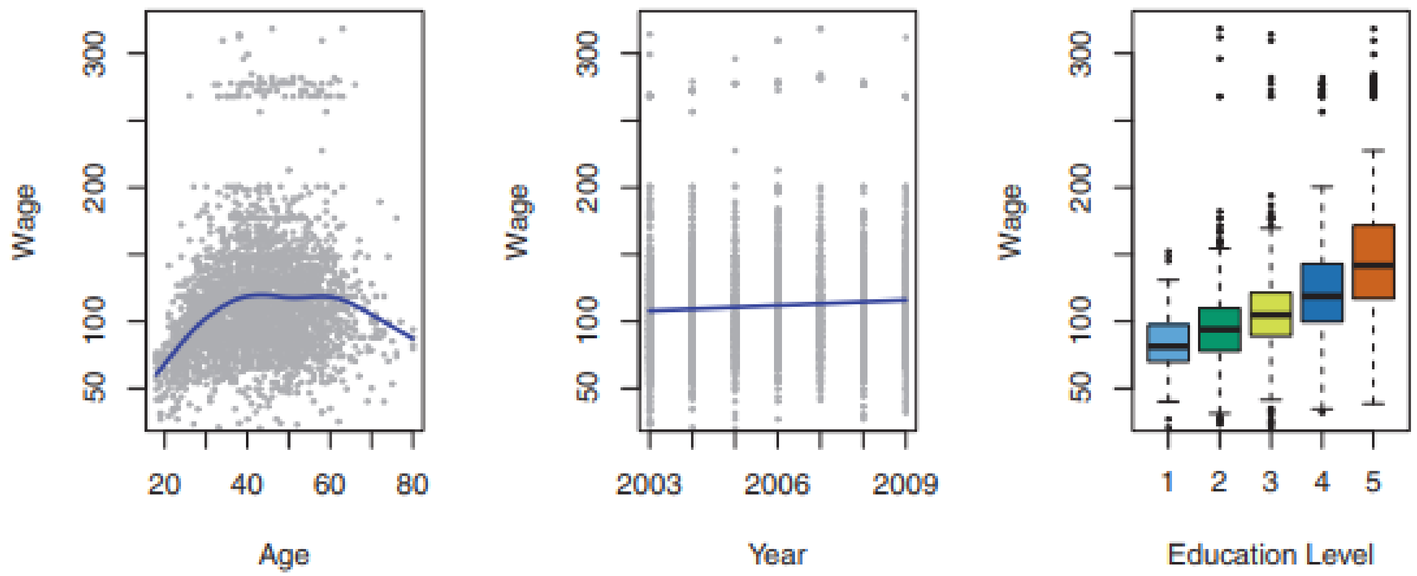
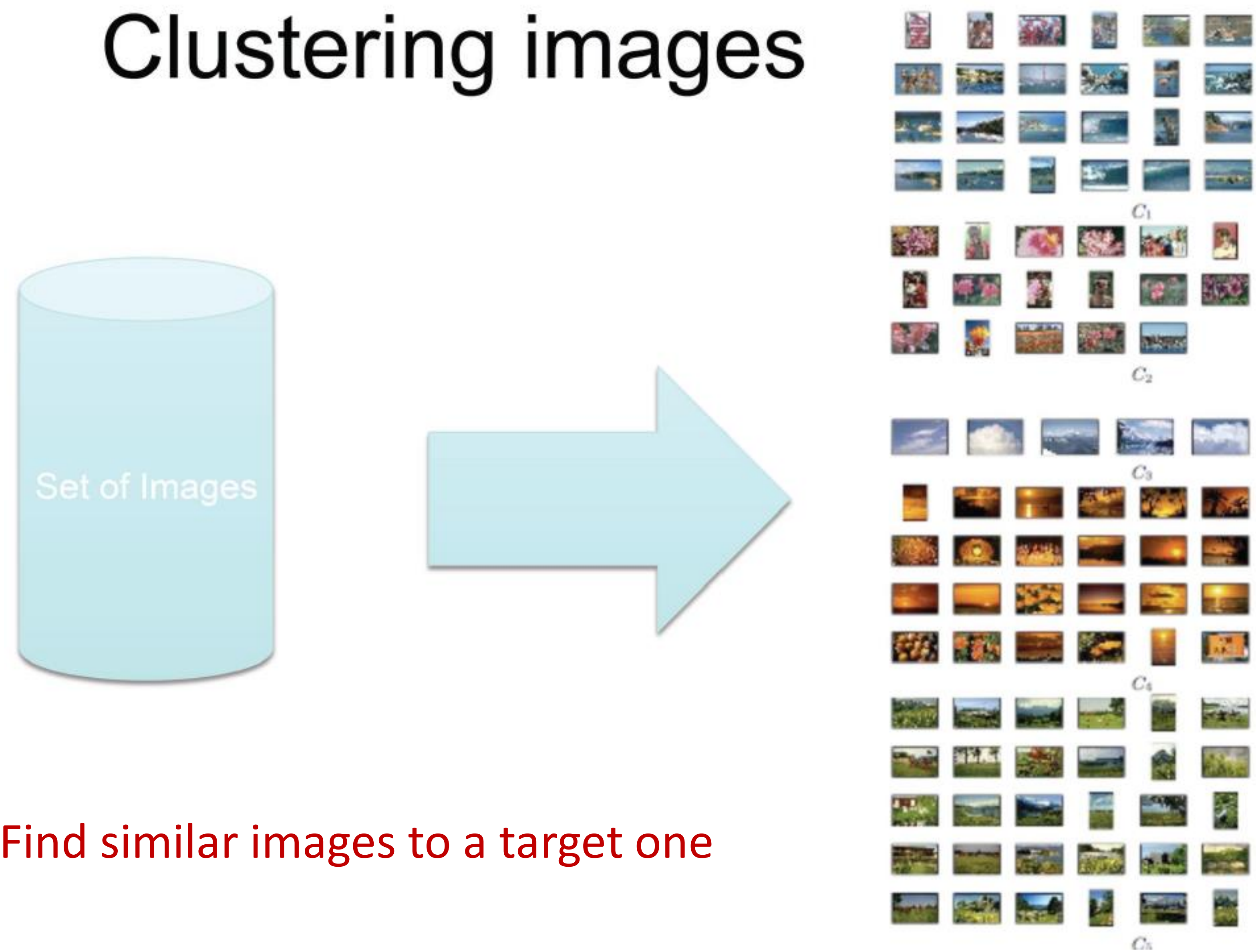


FIGURE 1.1. Wage *data, which contains income survey information for males from the central Atlantic region of the United States.* Left: wage *as a function of* age. *On average,* wage *increases with* age *until about 60 years of age, at which point it begins to decline.* Center: wage *as a function of* year. *There is a slow but steady increase of approximately* \$10,000 *in the average* wage *between* 2003 *and* 2009. Right: *Boxplots displaying* wage *as a function of* education, *with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* wage *increases with the level of education.*

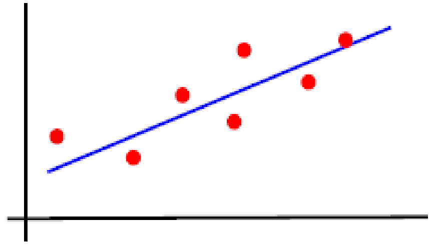
Example 3: image search



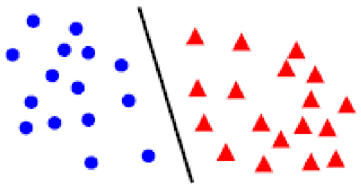
Find similar images to a target one

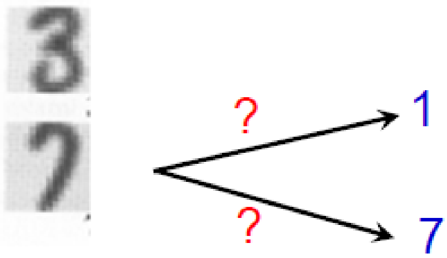
Three canonical learning problems

1. Regression - supervised
   - estimate parameters, e.g. of weight vs height



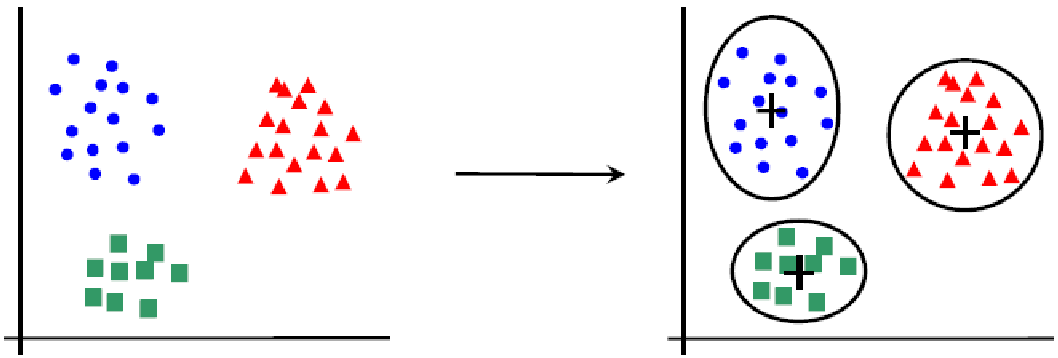
2. Classification - supervised
   - estimate class, e.g. handwritten digit classification



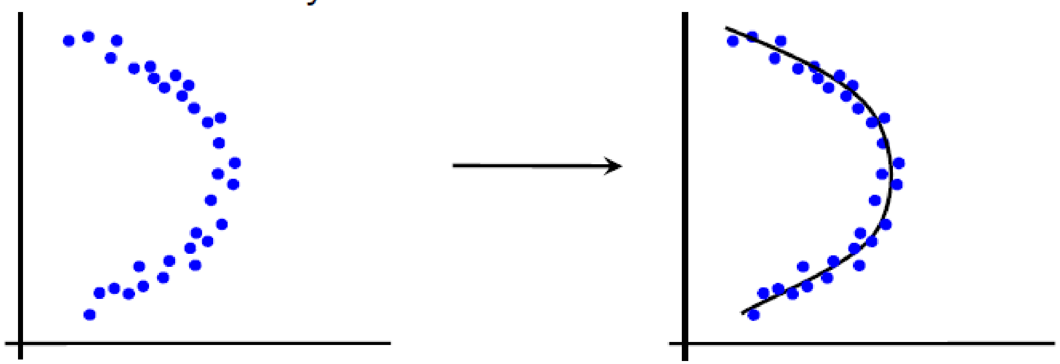


3. Unsupervised learning – model the data

- clustering



- dimensionality reduction



Terminology

- Hypothesis space $H = \{f: X \to Y\}$
- Training data $\mathrm{D} = (x_{\mathrm{i}}, y_{\mathrm{i}}) \in X \times Y$
- Features: $x_i \in X$
- Labels $y_{\mathrm{i}} \in Y$
  - Classification: discrete $y_{\mathrm{i}} \in$ {-1,1}
  - Regression: $y_{\mathrm{i}} \in \mathrm{R}$
- Loss function: $L(f, D)$
  - Measures how well f fits training data
- Training algorithm: Find hypothesis $\hat{f}: X \to Y$
  - $\hat{f} = \underset{f \in H}{\operatorname{argmin}} \; L(f, D)$

Learning f

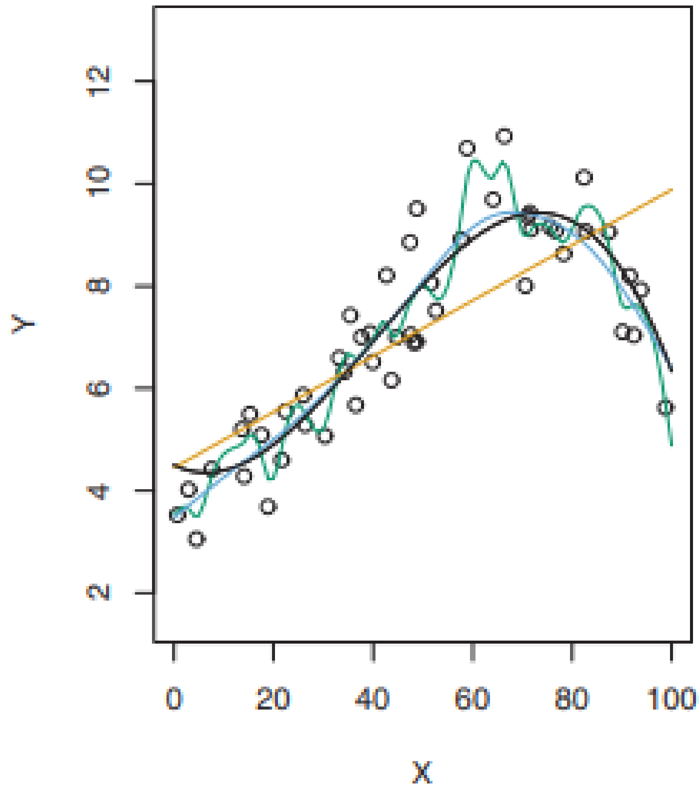
- Estimate f from training data
  - Classification error defined as:

$$1/\mathrm{N} \sum_{i=1}^{N} [y_i \neq f(x_i)]$$

- Real goal
  - Classify well new testing data
- Variance
  - Amount by which f would change if we estimated it using a different training data set
  - More complex models result in higher variance
- Bias
  - Error introduced by approximating a real-life problem by a much simpler model
  - E.g., assume linear model (linear regression)
  - More complex models result in lower bias

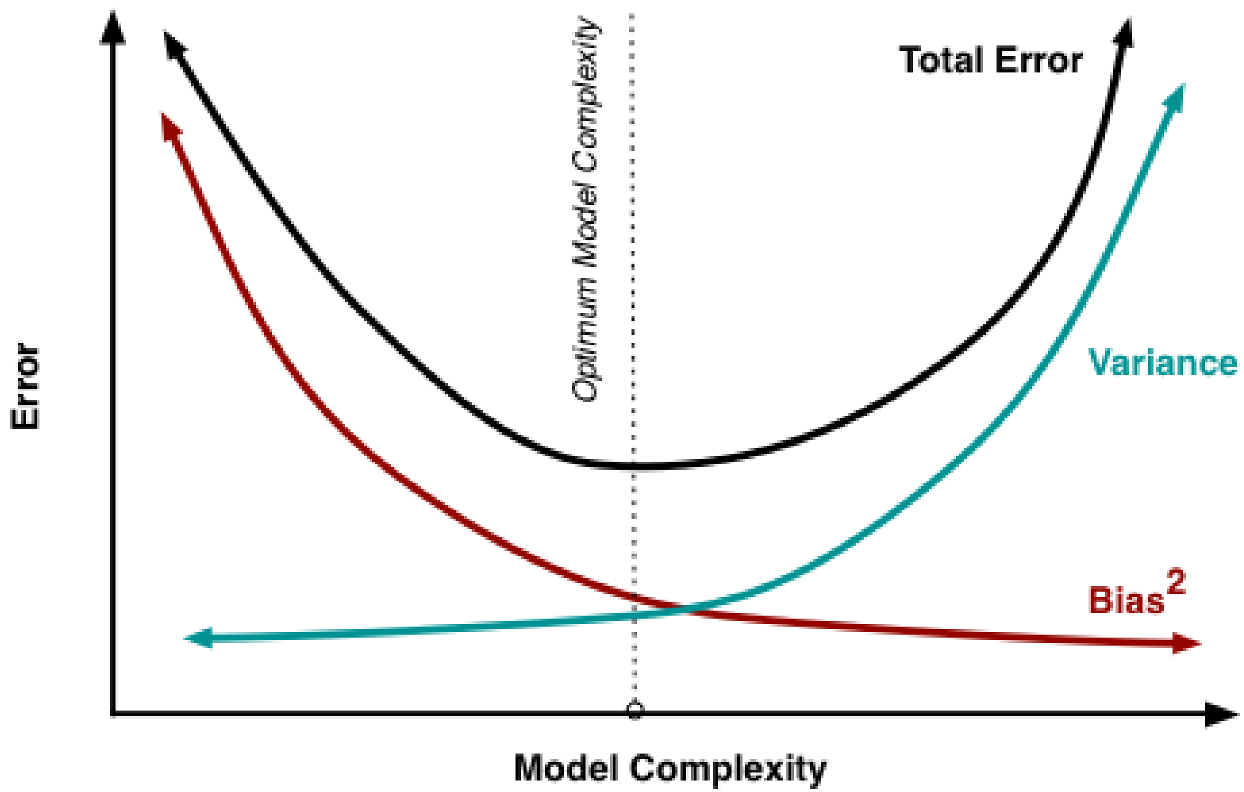
Bias-Variance tradeoff

Bias



- True data – Black
- Linear model – Orange (High Bias, Low Variance)
- Other models – Green and Blue (High variance, Low Bias)

Bias-Variance Tradeoff



Generalization

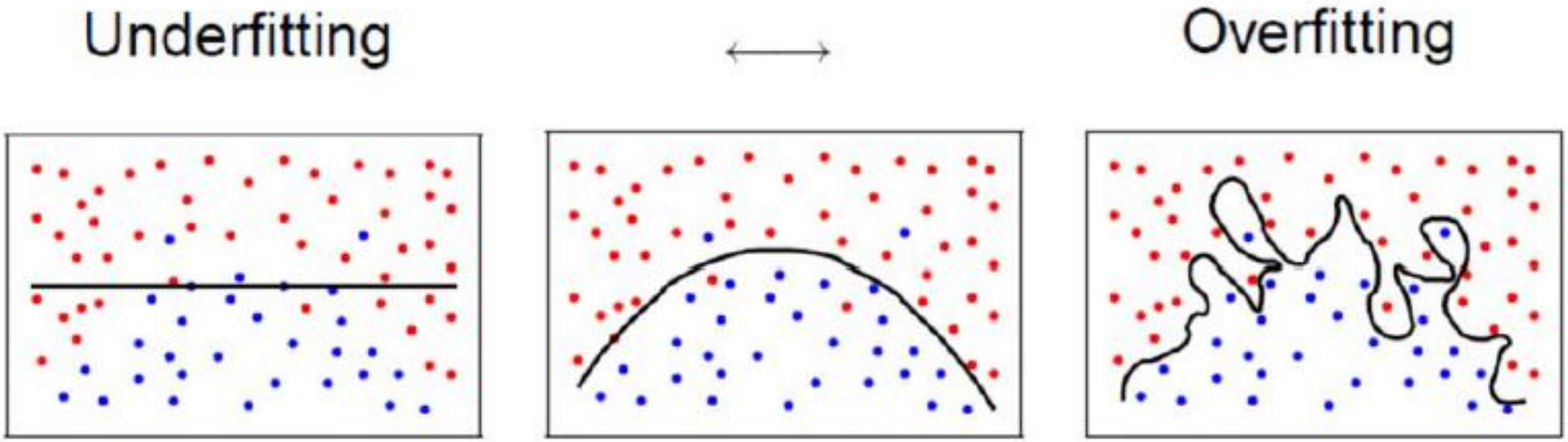
- The real aim of supervised learning is to do well on test data that is not known during learning

- Choosing the values for the parameters that minimize the loss function on the training data is not necessarily the best policy

- We want the learning machine to model the true regularities in the data and to ignore the noise in the data.

- Risk of *overfitting* model to training data
  - Could result in poor accuracy on new testing data

Generalization Problem in Classification



• Again, need to control the complexity of the (discriminant) function

Occam's Razor

- William of Occam: Monk living in the 14^{th} century
- Principle of parsimony:

"One should not increase, beyond what is necessary, the number of entities required to explain anything"

- When many solutions are available for a given problem, we should select the simplest one
- But what do we mean by simple?
- We will use prior knowledge of the problem to solve to define what is a simple solution

Key insights

- ML is a subset of AI designing learning algorithms
- Learning tasks are supervised (e.g., classification and regression) or unsupervised (e.g., clustering)
  - Supervised learning uses labeled training data
- Learning the “best” model is challenging
  - Select hypothesis space and loss function
  - Design algorithm to min loss function
  - Bias-Variance tradeoff
  - Need to generalize on new, unseen test data
  - Occam’s razor (prefer simplest model with good performance)