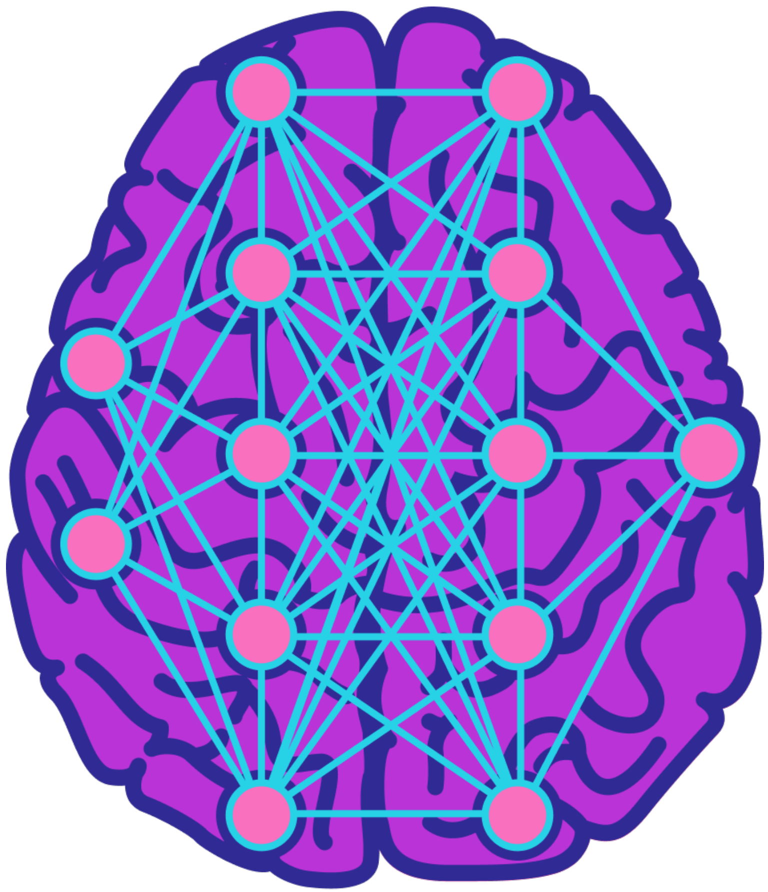
DS 5220

Supervised Machine Learning and Learning Theory

Alina Oprea
Associate Professor, CCIS
Northeastern University

September 4 2019

Welcome to DS 5220!

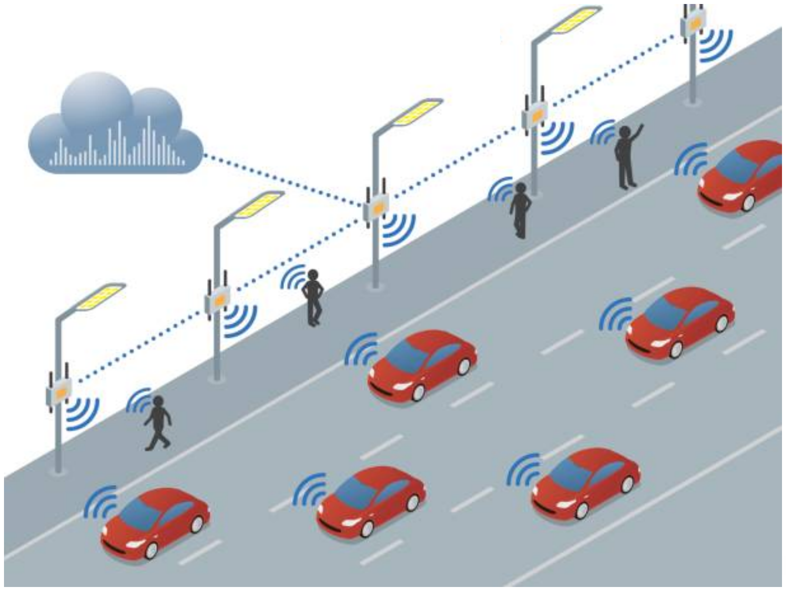


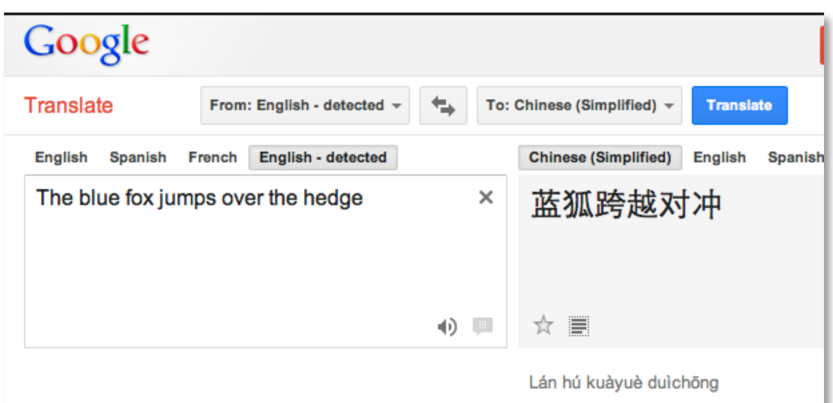
Supervised Machine Learning and Learning Theory

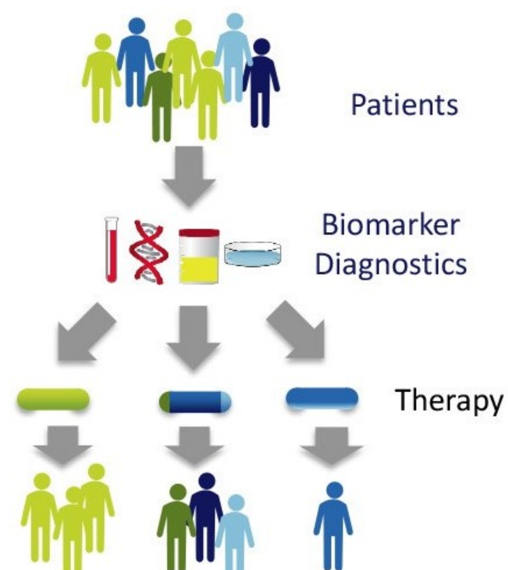
Introduction

- 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 (attack detection, IoT, connected car security)
  - Adversarial ML (study the vulnerabilities of ML in face of attacks and design defenses)

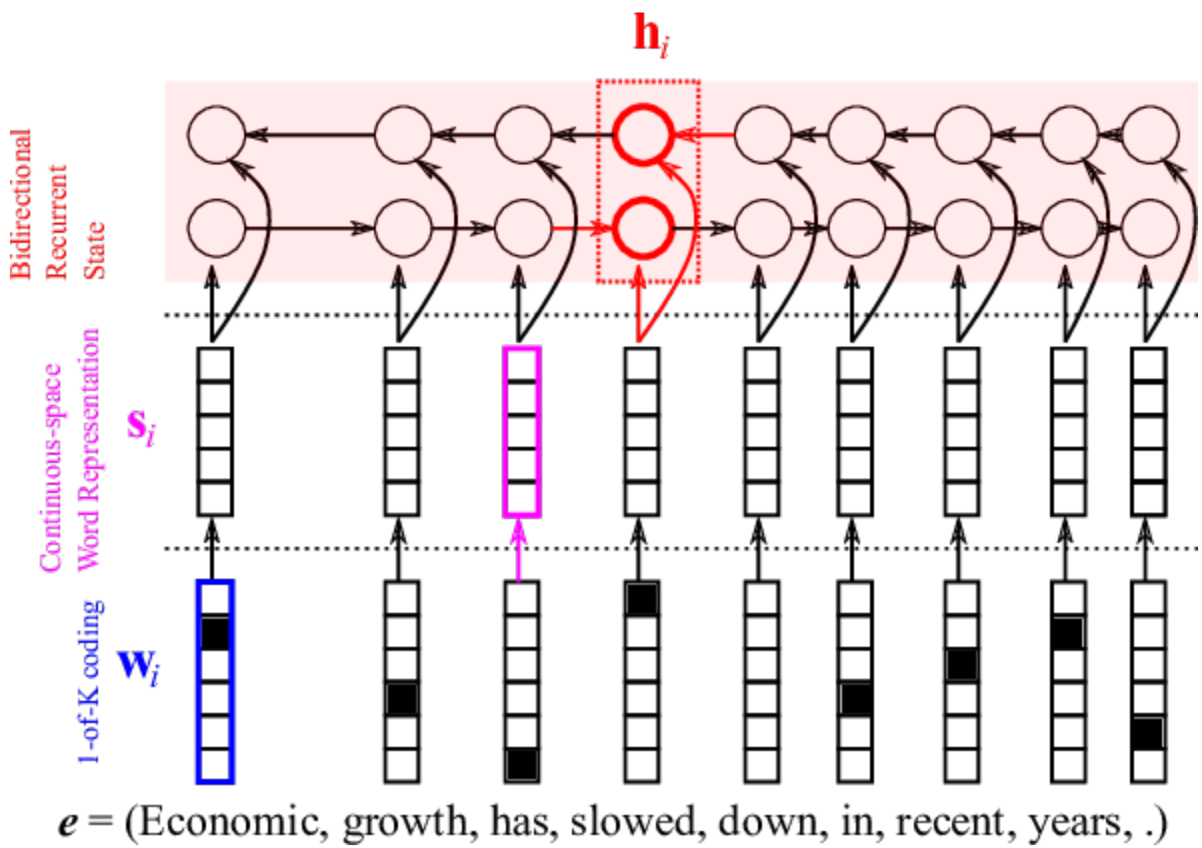
Machine learning is everywhere





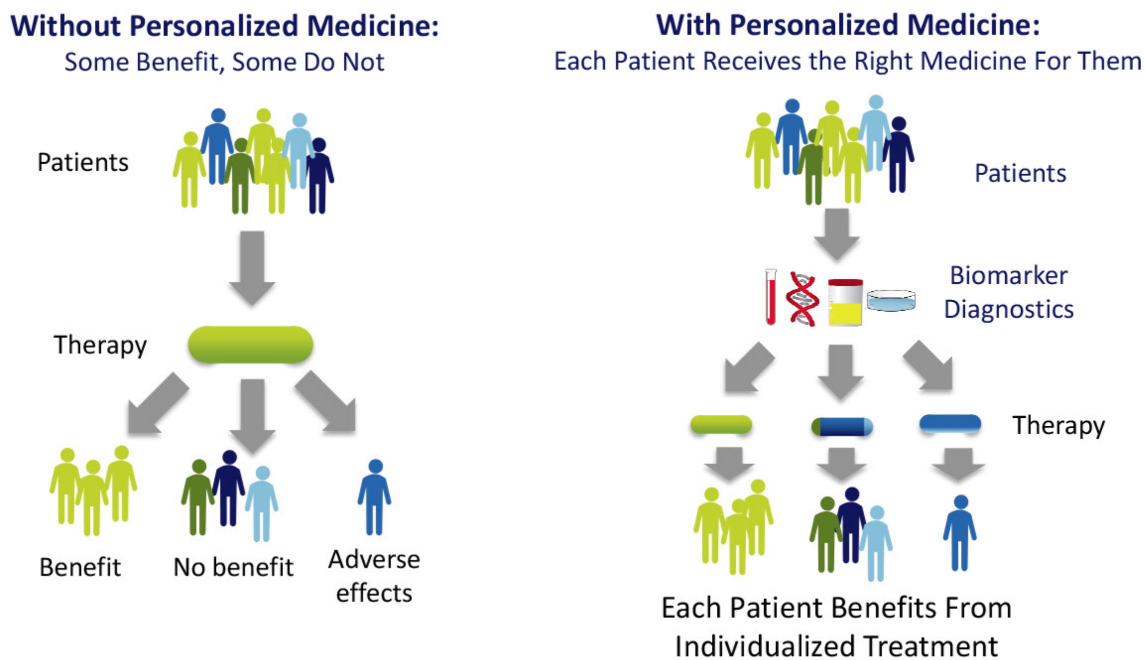


Natural Language Processing (NLP)



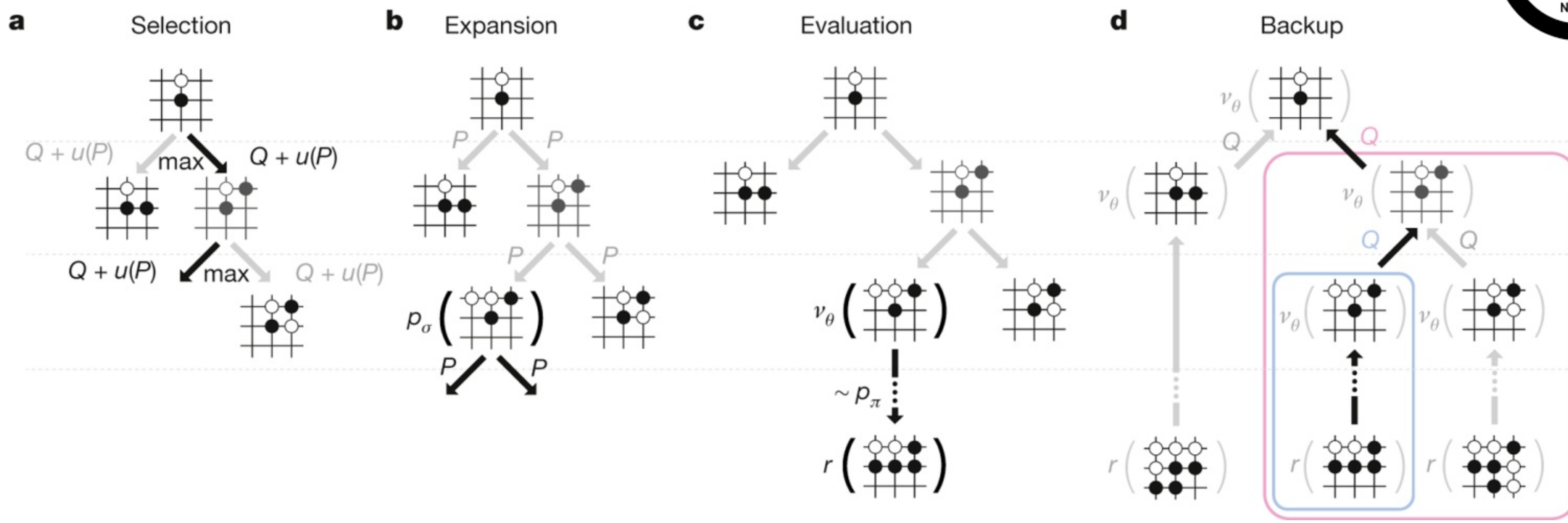
- Understand language semantics
- Real-time translation, speech recognition

Personalized Medicine



- Treatment adjusted to individual patients
- Predictive models using a variety of features
- Better outcome and reduced cost

Playing games



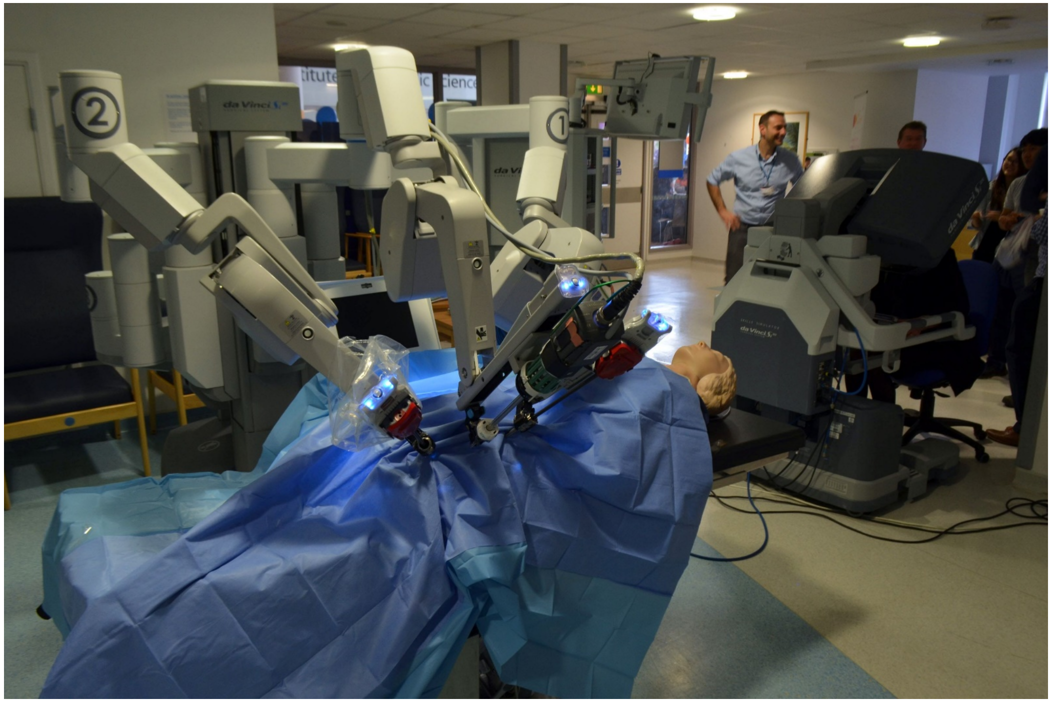
Reinforcement learning

- AlphaGo
- Chess

Fast Forward in the Near Future

AI Transportation in Cities of the Future (10-20 years)

Fast Forward in the Near Future



AI Robots in Medicine of the Future (10-20 years)

Short History

- Legendre and Gauss – linear regression, 1805
  - Astronomy applications
- Bayes and Laplace - Bayes Theorem, 1812
- Markov chains, 1913
- Fisher – linear discriminant analysis, 1936
  - Logistic regression, 1940
- Widrow and Hoff ADELINE neural network, 1959
- Nelder, Wedderburn, generalized linear models, 1970
- "AI winter", limitations of perceptron, 1970
- Breiman, Friedman, Olshen, Stone, decision trees, 1980
- More work on neural networks, 1980
- Cortes and Vapnik , SVM with kernels, 1990
- IBM Deep Blue beats Kasparov at chess, 1996
- Geoffrey Hinton, Deep learning, back propagation, 2006

DS-5220

- 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 Artificial Intelligence (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)

DS-5220 Course objectives

- Become familiar with main machine learning tasks
  - Supervised learning vs unsupervised learning
  - Classification vs Regression
- Study most well-known algorithms
  - Regression (linear regression, spline regression)
  - Classification (SVM, decision trees, Naïve Bayes, ensembles, etc.)
  - Deep learning (different neural network architectures)
- Learn the theory and foundation behind ML algorithms and learn to apply them to real datasets
- Learn about security challenges of ML
  - Introduction to adversarial ML

http://www.ccs.neu.edu/home/alina/classes/Fall2019

Class Outline

- Introduction
  - Probability and linear algebra review
- Regression - 2 weeks
  - Linear regression, polynomial, spline regression
- Classification - 4 weeks
  - Linear classification (logistic regression, LDA)
  - Non-linear models (decision trees, SVM, Naïve Bayes)
  - Ensembles (random forest, AdaBoost)
  - Model selection, regularization, cross validation
- Neural networks and deep learning – 2 weeks
  - Back-propagation, gradient descent
  - NN architectures (feed-forward, convolutional, recurrent)
- Adversarial ML – 1 lecture
  - Security of ML at testing and training time

Resources

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R

- Trevor Hastie, Rob Tibshirani, and Jerry Friedman, Elements of Statistical Learning, Second Edition, Springer, 2009.

- Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

- A. Zhang, Z. Lipton, and A. Smola. Dive into Deep Learning

Policies

- Instructors
  - Alina Oprea
  - TAs: Yuxuan (Ewen) Wang; Christopher Gomes
- Schedule
  - Mon, Wed 2:50-4:30pm
  - West Village H 108
  - Office hours:
    - Alina: Wed 4:30 – 6:00 pm (ISEC 625)
    - Christopher : Monday 5:00-6:00pm (ISEC 605)
    - Ewen: Thursday 5:00-6:00pm (ISEC 605)
- Online resources
  - Slides will be posted after each lecture on public website
  - Piazza for questions and discussion
  - Gradescope for homework and project submission

Policies, cont.

- Your responsibilities
  - Please be on time, attend classes, and take notes
  - Participate in interactive discussion in class
  - 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
  - No need to email for late days, Gradescope shows submission time

Grading

- Assignments – 20%
  - 4-5 assignments and programming exercises based on studied material in class
- Midterm – 25%
- Final exam – 25%
- Final project – 25%
  - Select your own project, public dataset
  - Submit short project proposal and milestone
  - Project pitch and final presentation
  - Project report
- Class participation – 5%
  - Participate in class discussion and on Piazza

Assignments

- Programming exercises, and theory questions
  - Prefer Latex/Word/... write up
- Language
  - Use R or Python
  - Jupyter notebooks recommended
- Submission
  - Submit PDF report in Gradescope
  - Includes all the results, as well as link to code and instructions to run it

Final project

- Goal: work on a larger data science project
  - Build your portfolio and increase your experience
- Requirements
  - Large dataset: at least 10,000 records (public source)
  - Not recommended to collect your own data
  - Pick application of interest, but instructor will also provide potential list of projects
  - Experiment with at least 3 ML models
  - Perform in-depth analysis (which features contribute mostly to prediction, which model performs best)
- Timeline
  - Proposal: mid class; milestone 2-3 weeks after (Instructors will provide early feedback)
  - Final presentation (10 mins) and report (5-6 pages)

Academic Integrity

- Homework is done individually!
- Rules
  - Can discuss with colleagues or instructor
  - Can post and answer questions on Piazza
  - Code cannot be shared with colleagues
  - Cannot use code from the Internet
    - Use python or R packages, but not directly code for ML analysis written by someone else
- NO CHEATHING WILL BE TOLERATED!
- Any cheating will automatically result in grade F and report to the university administration
- http://www.northeastern.edu/osccr/academic-integrity-policy/

Outline

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

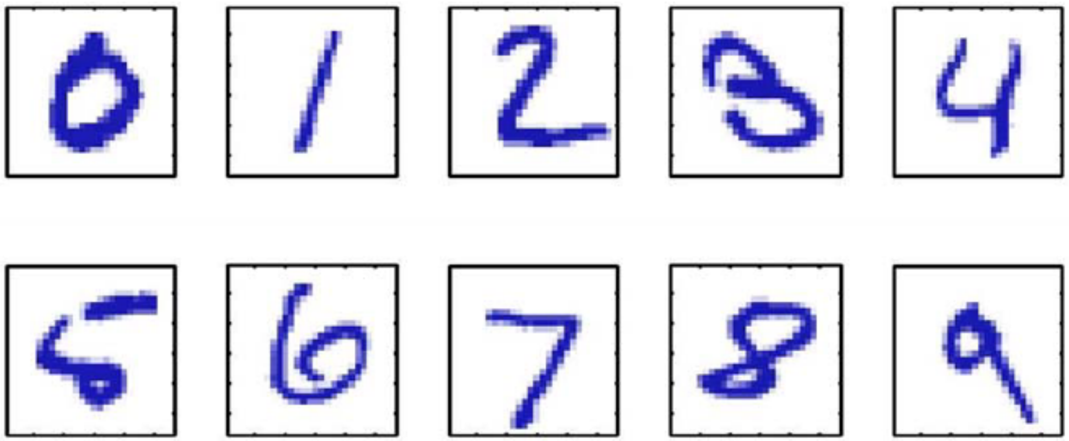
Slides adapted from

- 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 $\mathbf{x} \in \mathbb{R}^{784}$

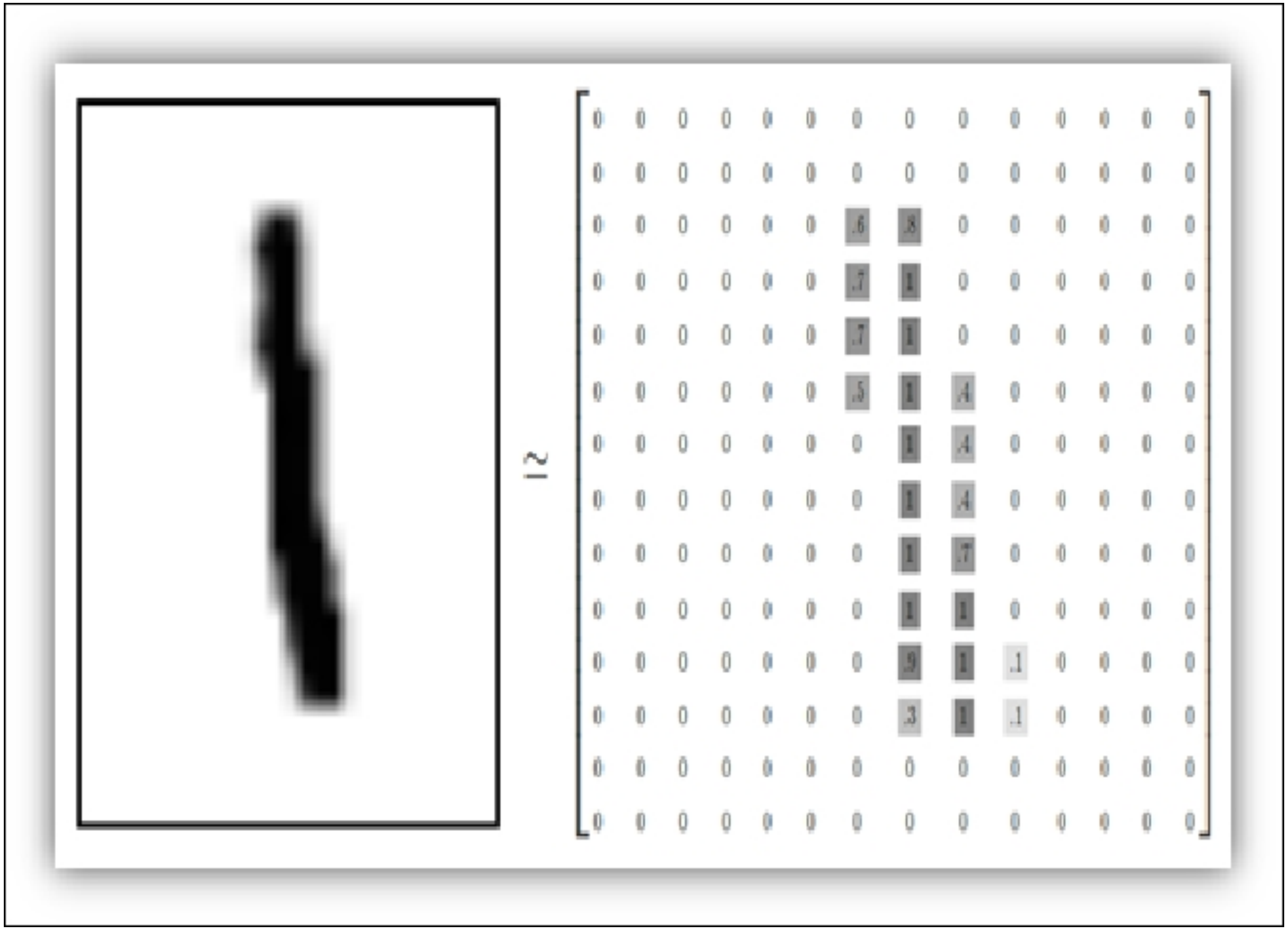
Learn a classifier $f(\mathbf{x})$ such that,

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

MNIST dataset: Predict the digit

Multi-class classifier

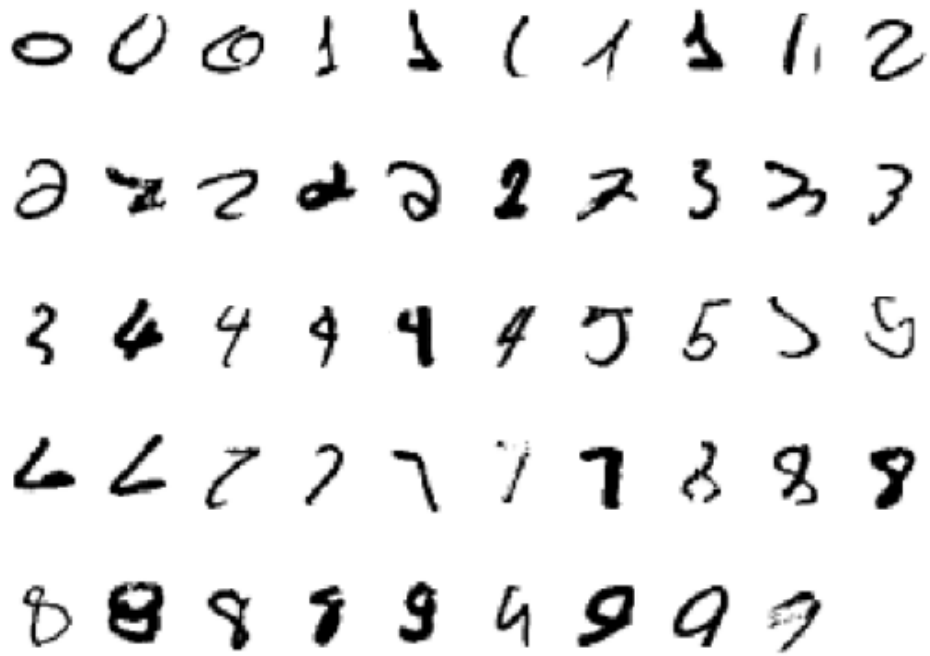
Data Representation



Model the problem

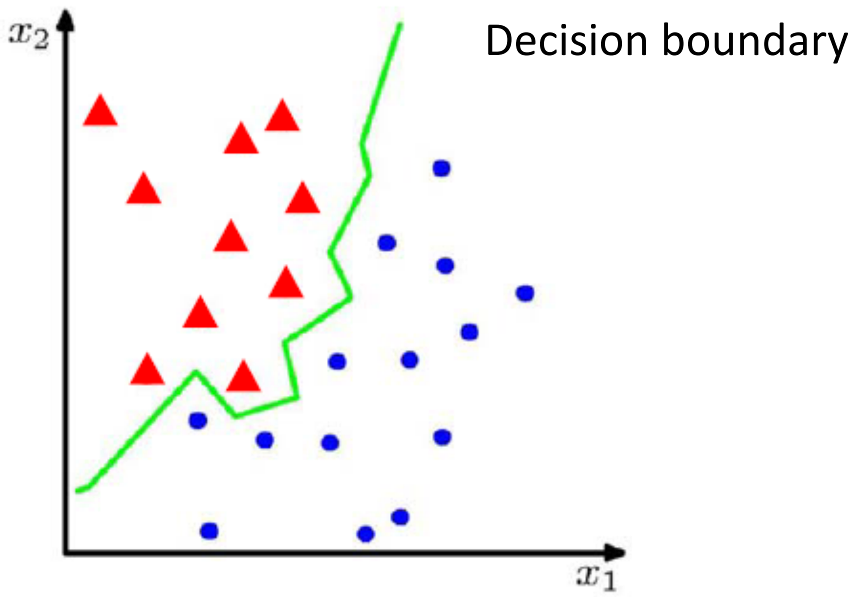
As a supervised classification problem

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)

Classification



- Suppose we are given a training set of N observations

$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N), x_i \in \mathbb{R}^d, y_i \in \{0, 1\}$$

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

$$f(x_i) = y_i$$

Extended to multi-class classification

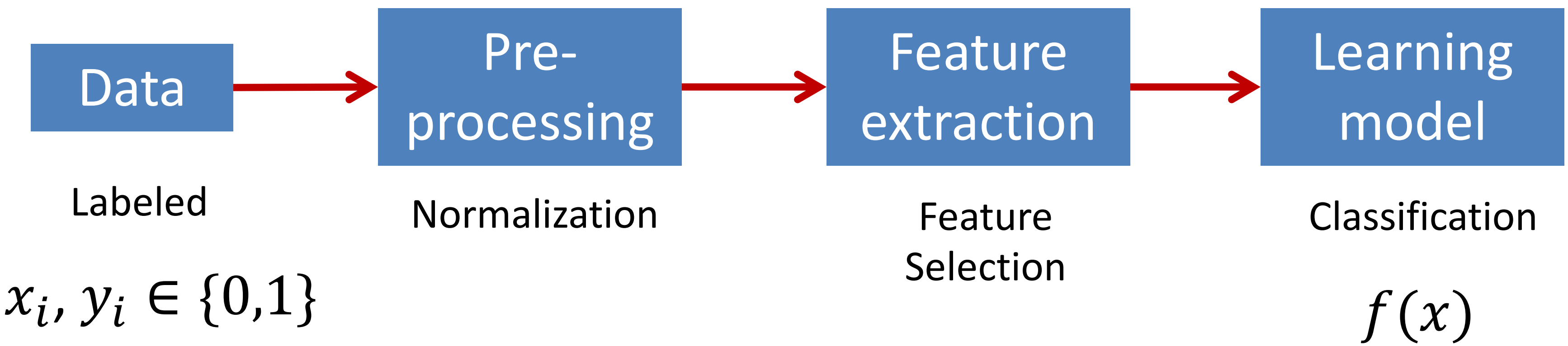
- Handwritten digit recognition: $y_i \in \{0,1, \ldots, 9\}$

Classification

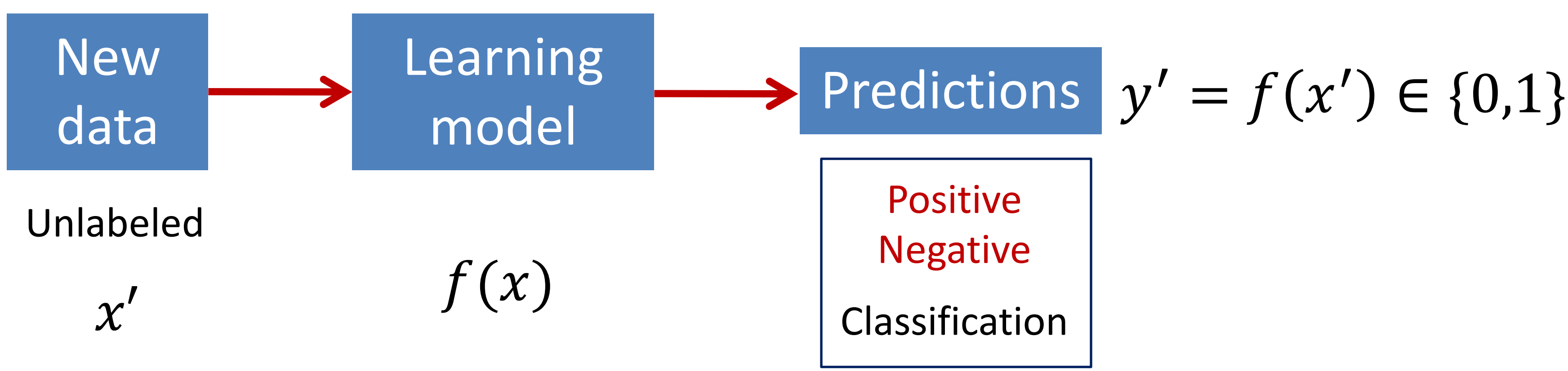
- Training data
  - $x_i = [x_{i,1}, \dots x_{i,d}]$: vector of image pixels (features)
  - Size $d = 28\text{x}28 = 784$
  - y_i: image label
- Models (hypothesis)
  - Example: Linear model (parametric model)
    - $f(x) = wx + b$
  - Classify 1 if $f(x) > \mathrm{T}$; 0 otherwise
- Classification algorithm
  - Training: Learn model parameters w, b to minimize error (number of training examples for which model gives wrong label)
  - Output: "optimal" model
- Testing
  - Apply learned model to new data and generate prediction $f(x)$

Supervised Learning: Classification

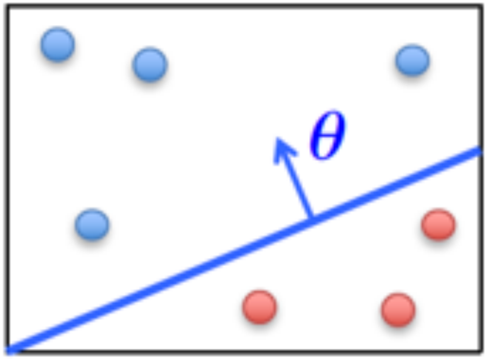
Training



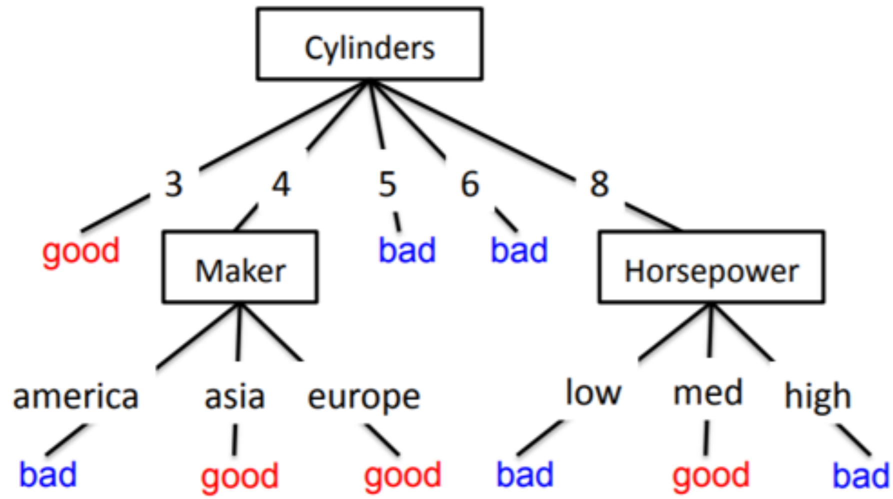
Testing



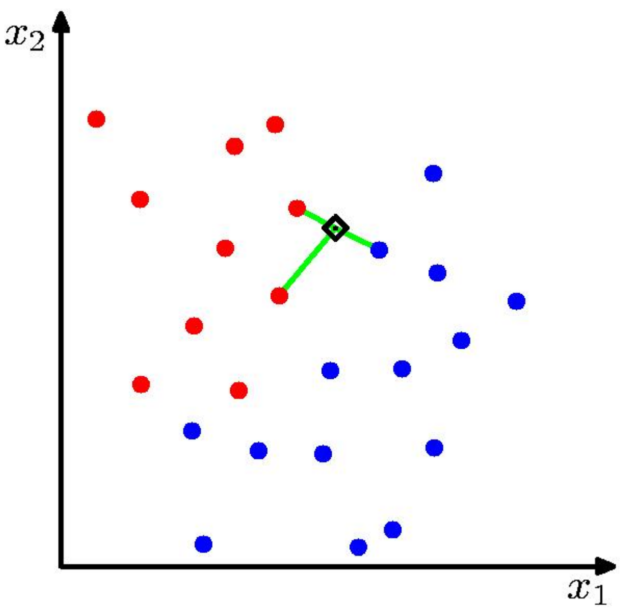
Example Classifiers



Linear classifiers: logistic regression, LDA (parametric)

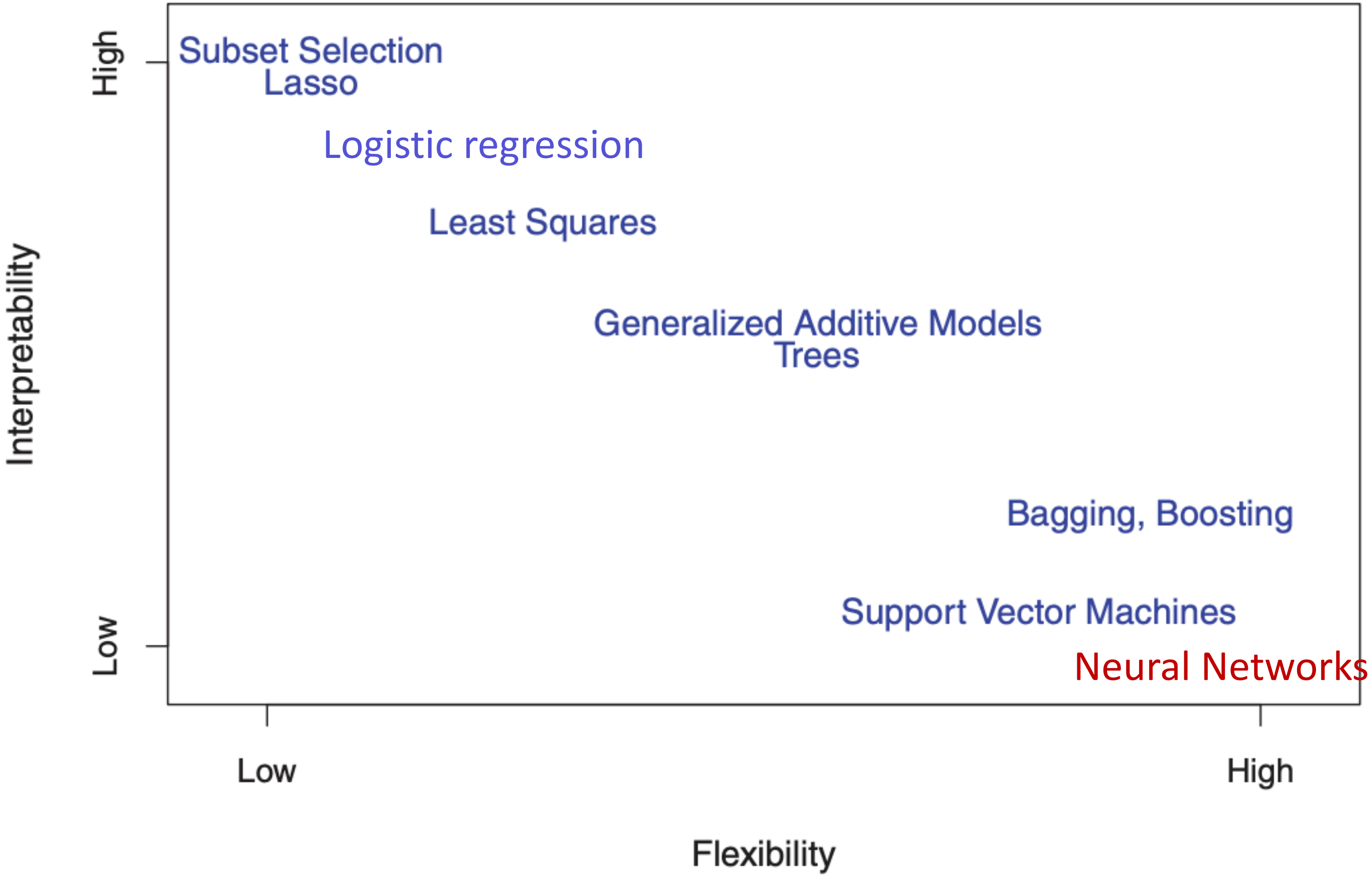


Decision trees (non-parametric)

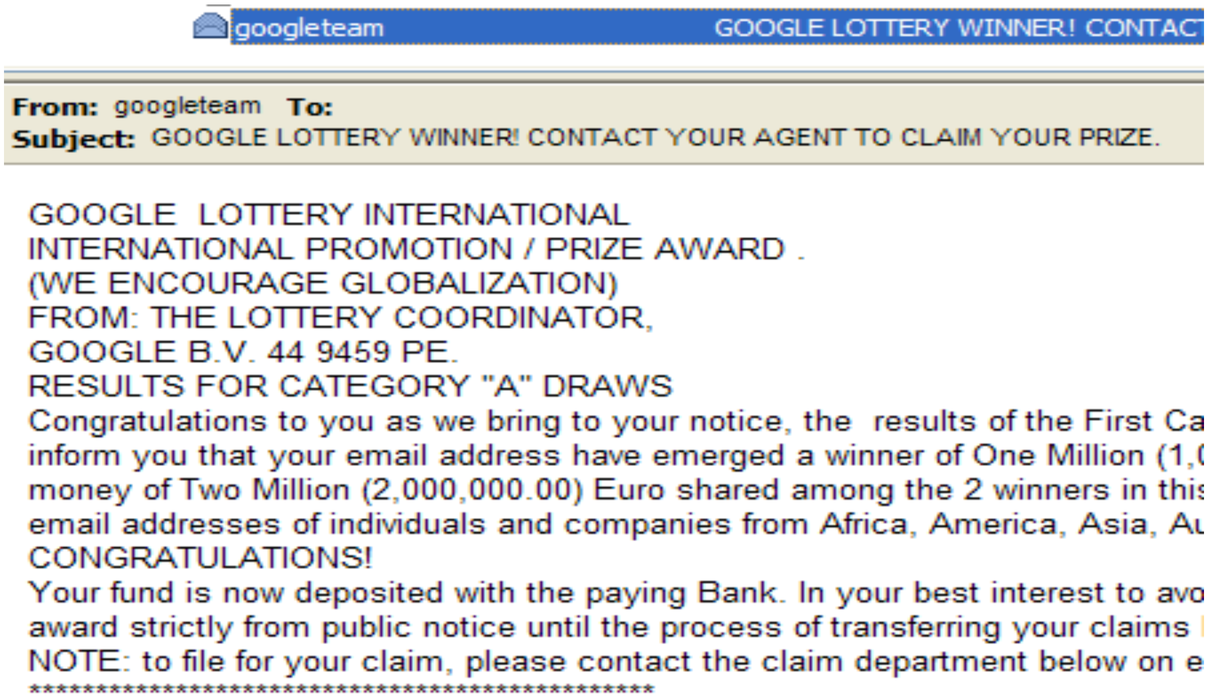


Nearest Neighbors: kNN (non-parametric)

Interpretability



Real-world example: Spam email



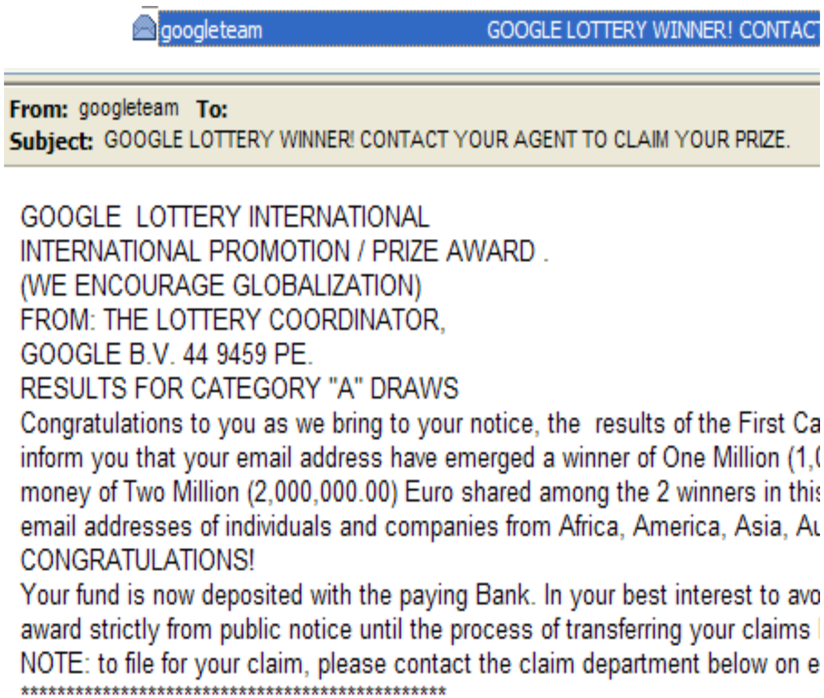
googleteam GOOGLE LOTTERY WINNER! CONTACT

From: googleteam **To:**
Subject: GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

GOOGLE LOTTERY INTERNATIONAL
INTERNATIONAL PROMOTION / PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.
RESULTS FOR CATEGORY "A" DRAWS
Congratulations to you as we bring to your notice, the results of the First Ca
inform you that your email address have emerged a winner of One Million (1,0
money of Two Million (2,000,000.00) Euro shared among the 2 winners in this
email addresses of individuals and companies from Africa, America, Asia, Au
CONGRATULATIONS!
Your fund is now deposited with the paying Bank. In your best interest to avo
award strictly from public notice until the process of transferring your claims
NOTE: to file for your claim, please contact the claim department below on e
****

SPAM email
- Unsolicited
- Advertisement
- Sent to a large number of people

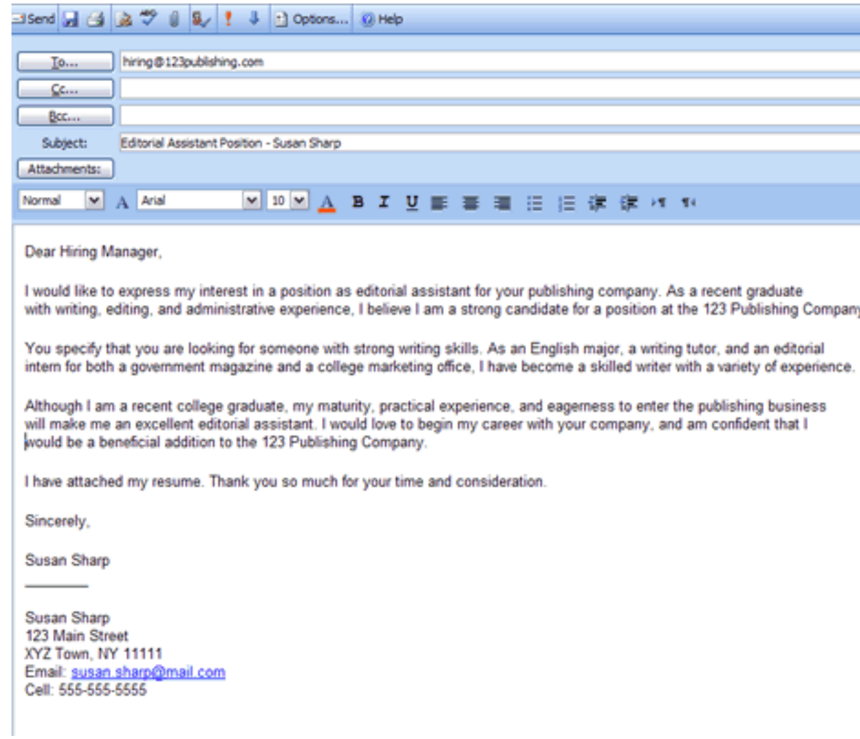
Classifying spam email



googleteam GOOGLE LOTTERY WINNER! CONTAC

From: googleteam To:
Subject: GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

GOOGLE LOTTERY INTERNATIONAL
INTERNATIONAL PROMOTION / PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.
RESULTS FOR CATEGORY "A" DRAWS
Congratulations to you as we bring to your notice, the results of the First Ca
inform you that your email address have emerged a winner of One Million (1,(
money of Two Million (2,000,000.00) Euro shared among the 2 winners in this
email addresses of individuals and companies from Africa, America, Asia, Au
CONGRATULATIONS!
Your fund is now deposited with the paying Bank. In your best interest to avo
award strictly from public notice until the process of transferring your claims
NOTE: to file for your claim, please contact the claim department below on e
****



Send Options... Help

To... hiring@123publishing.com
Cc...
Bcc...
Subject: Editorial Assistant Position - Susan Sharp
Attachments:

Normal Arial 10

Dear Hiring Manager,

I would like to express my interest in a position as editorial assistant for your publishing company. As a recent graduate with writing, editing, and administrative experience, I believe I am a strong candidate for a position at the 123 Publishing Company.

You specify that you are looking for someone with strong writing skills. As an English major, a writing tutor, and an editorial intern for both a government magazine and a college marketing office, I have become a skilled writer with a variety of experience.

Although I am a recent college graduate, my maturity, practical experience, and eagerness to enter the publishing business will make me an excellent editorial assistant. I would love to begin my career with your company, and am confident that I would be a beneficial addition to the 123 Publishing Company.

I have attached my resume. Thank you so much for your time and consideration.

Sincerely,

Susan Sharp

Susan Sharp
123 Main Street
XYZ Town, NY 11111
Email: susan.sharp@mail.com
Cell: 555-555-5555

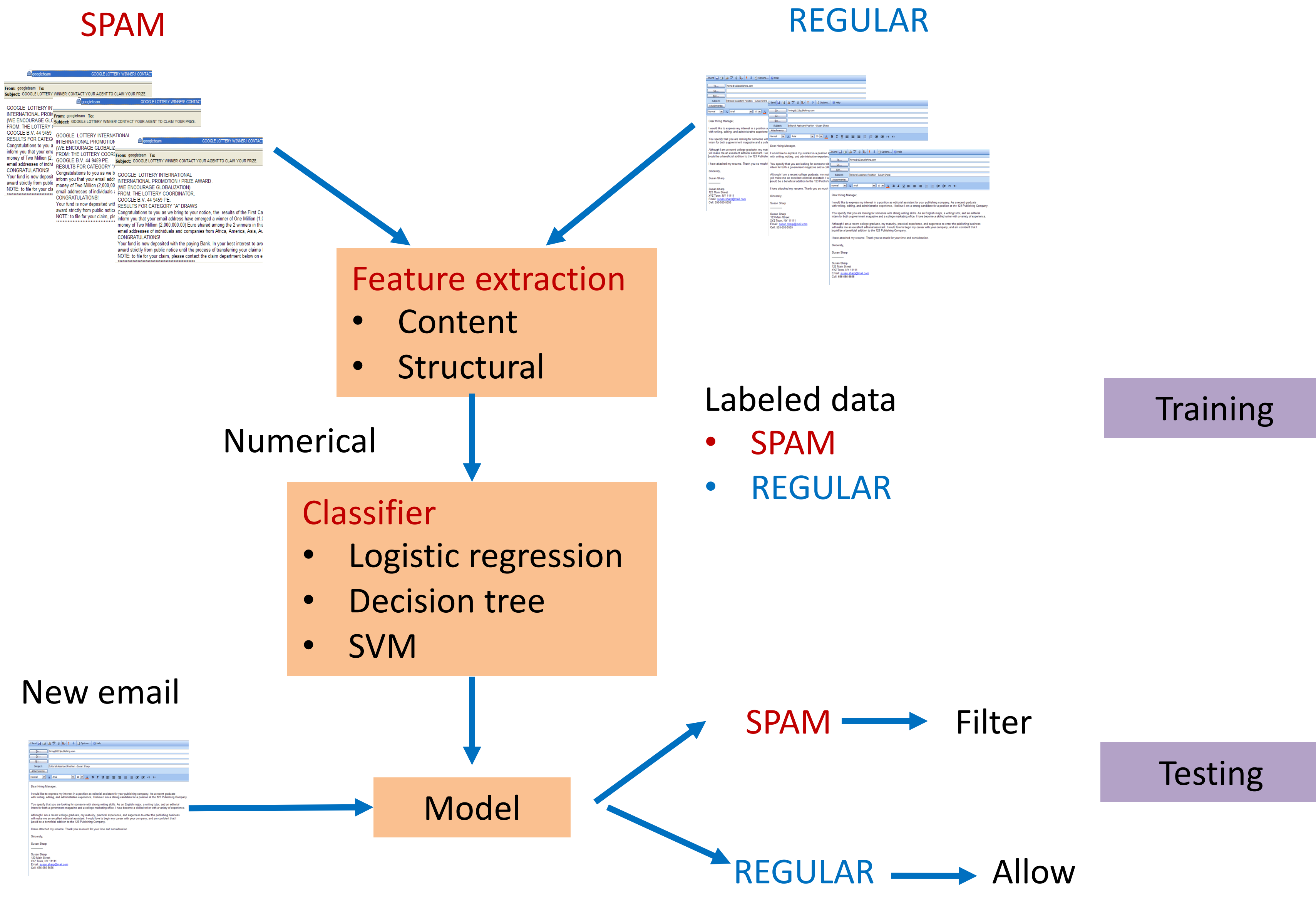
Content-related features

- Use of certain words
- Word frequencies
- Language
- Sentence

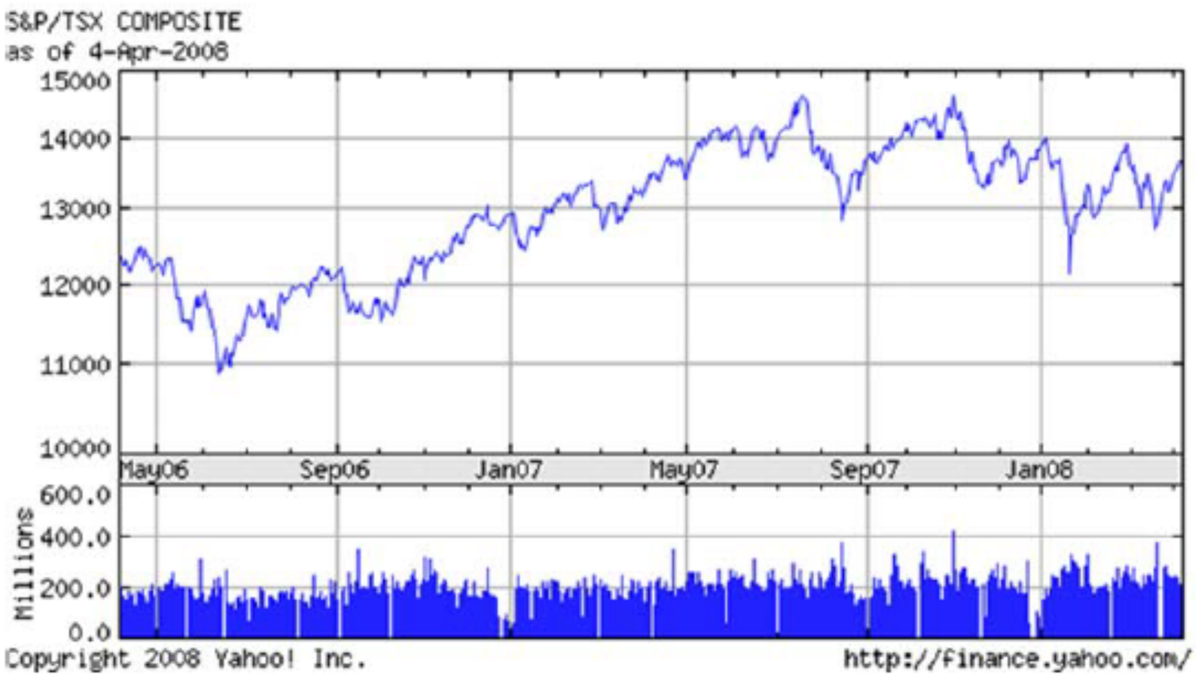
Structural features

- Sender IP address
- IP blacklist
- DNS information
- Email server
- URL links (non-matching)

Classifying spam email

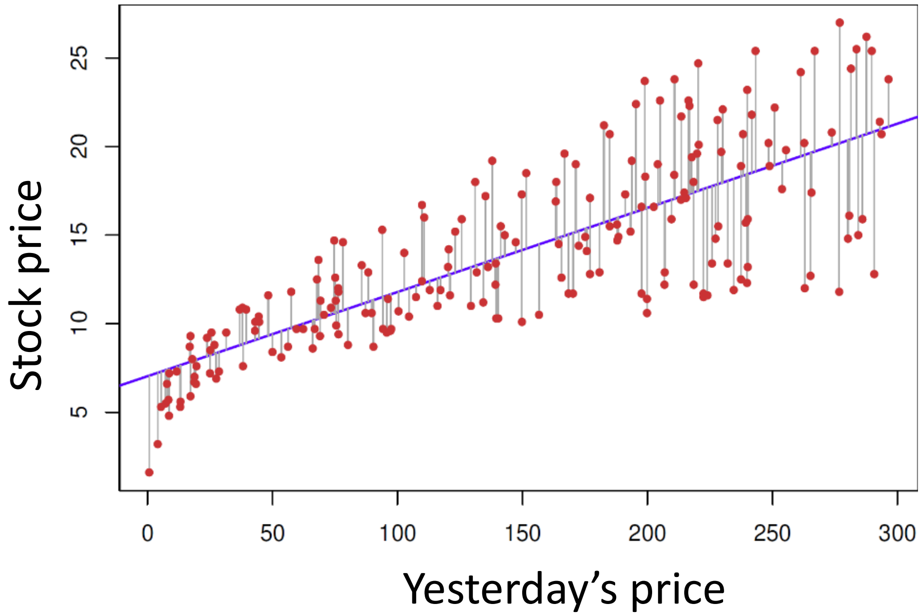


Example 2
Stock market prediction



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

Regression



Linear regression
1 dimension

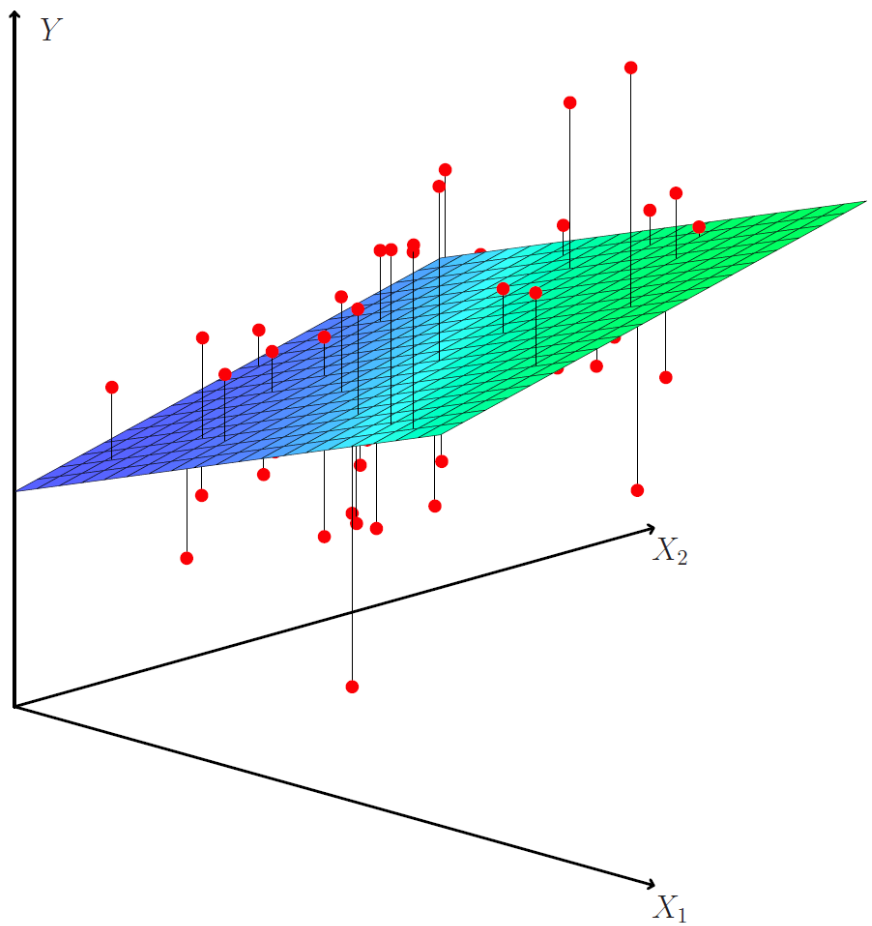
- Suppose we are given a training set of N observations

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

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

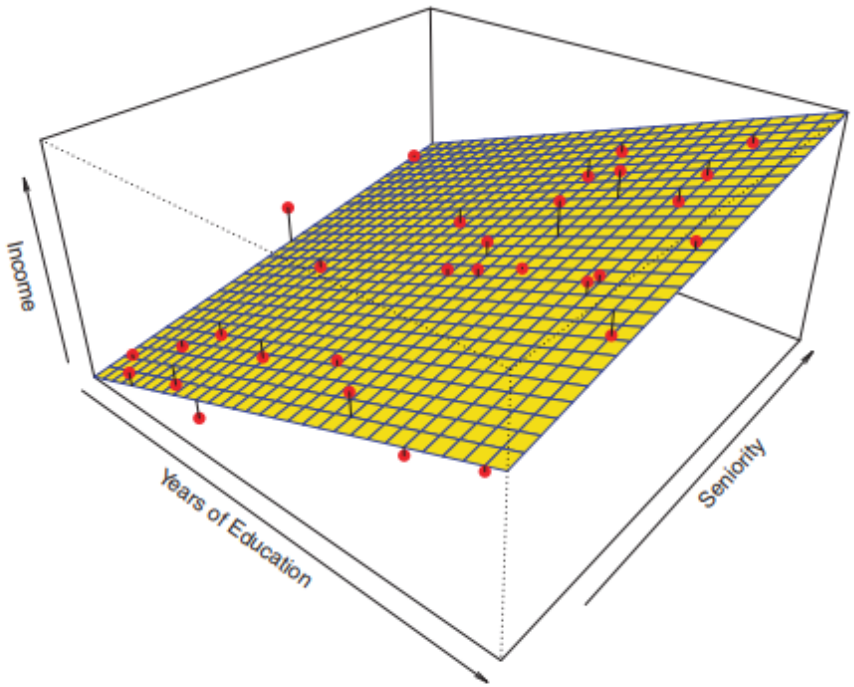
$x_i = (x_{i1}, \ldots, x_{id})$ - d predictors (features)
y_i - response variable, numerical

Multi-dimensional linear regression

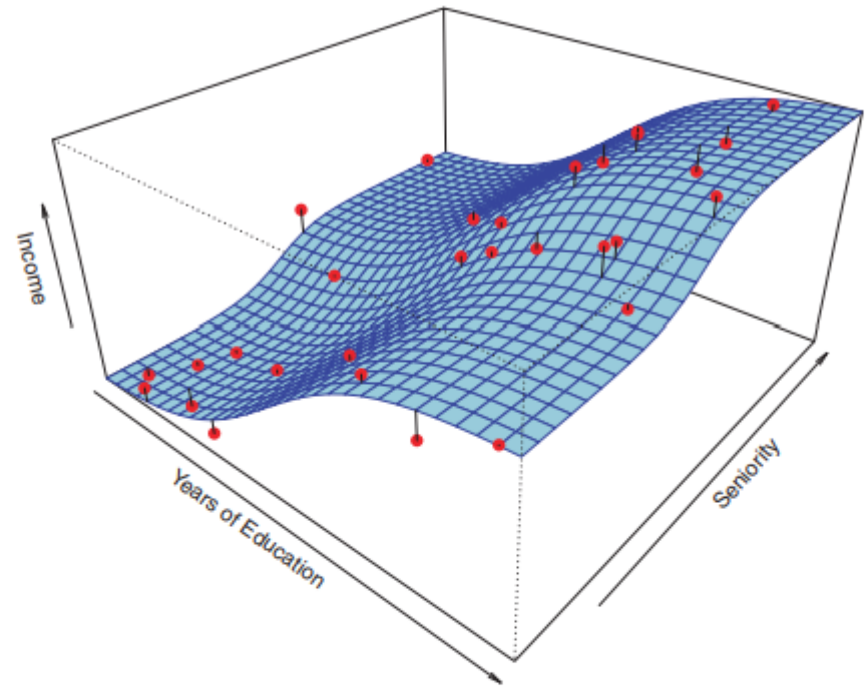


Minimize sum of square error
(error/loss function)

Income Prediction



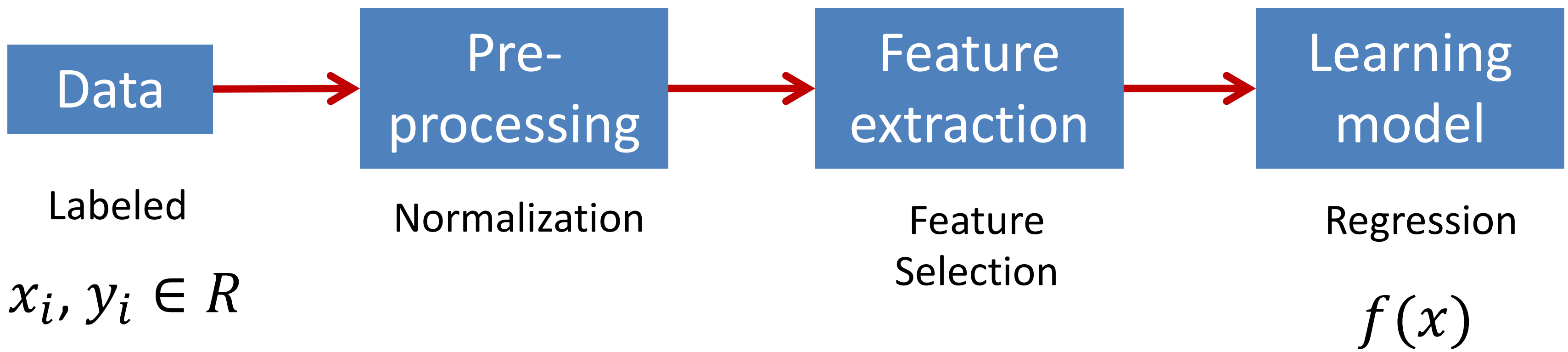
Linear Regression



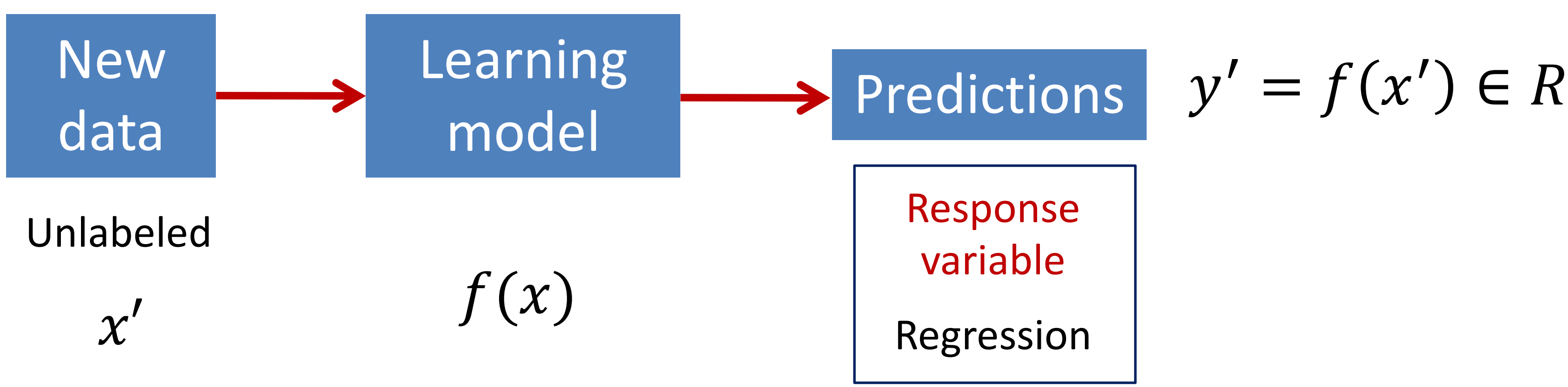
Non-Linear Regression
Polynomial/Spline Regression

Supervised Learning: Regression

Training

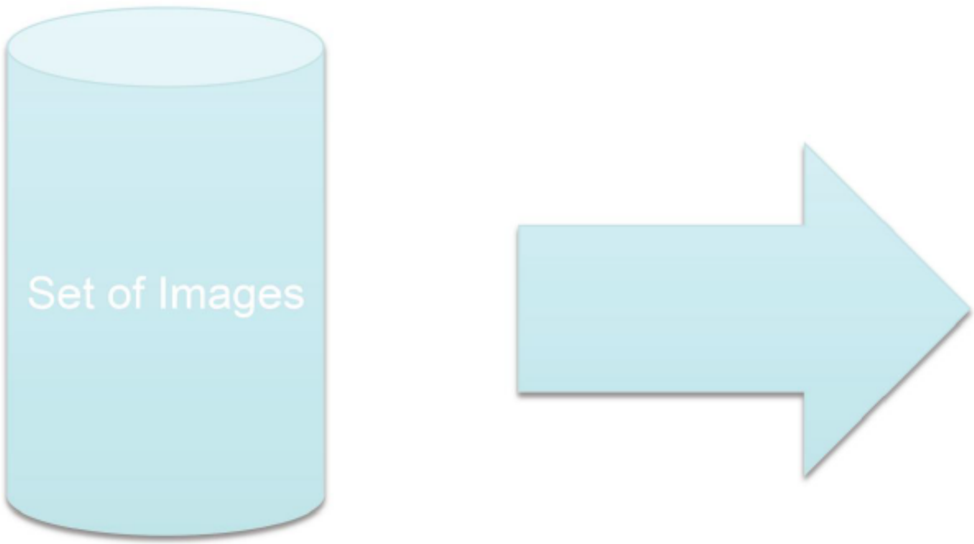


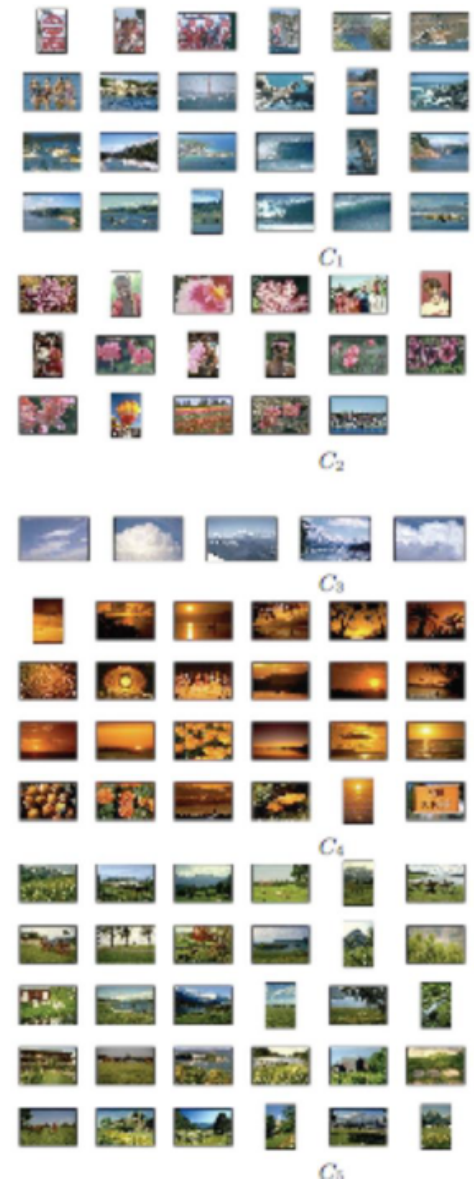
Testing



Example 3: image search

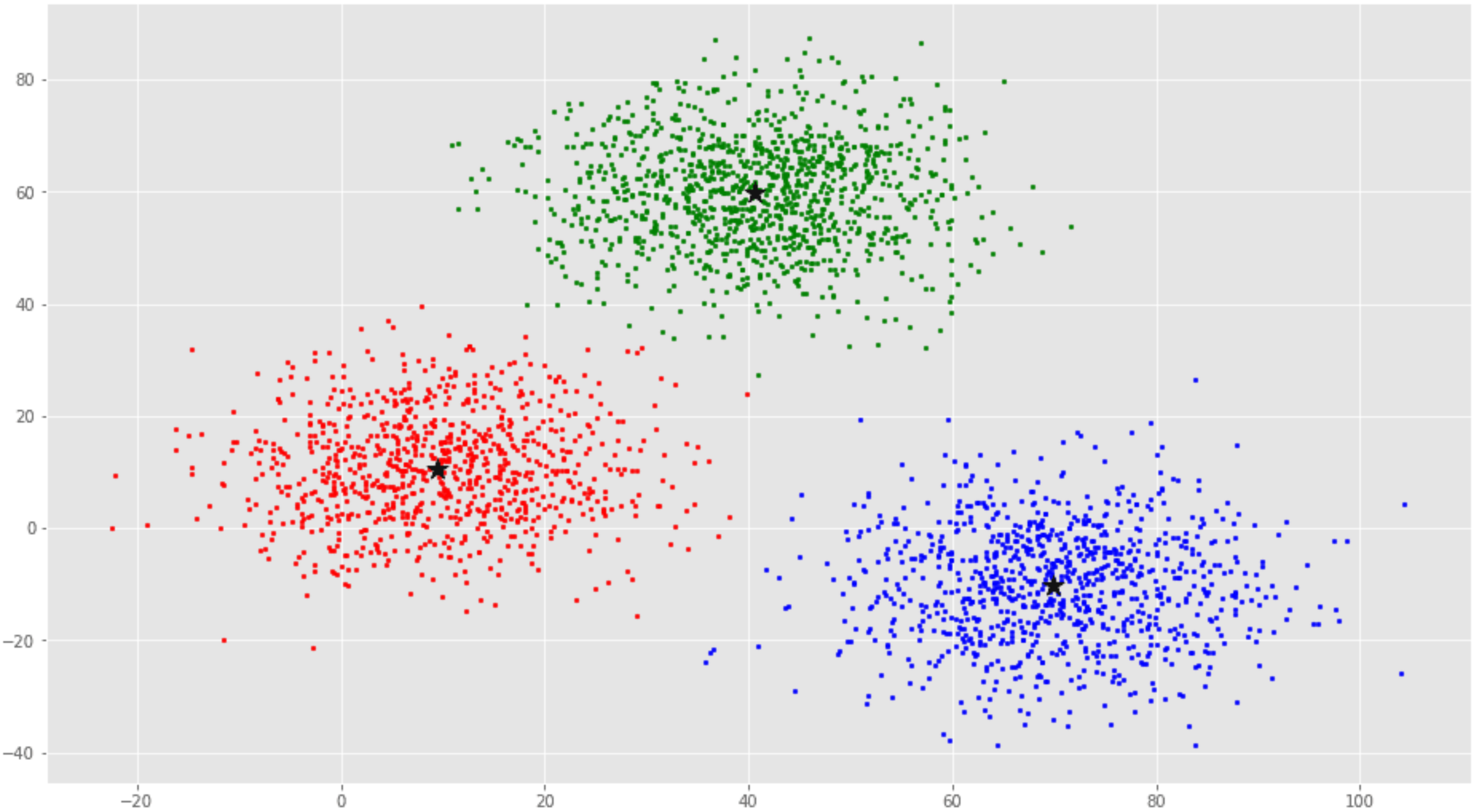
Clustering images





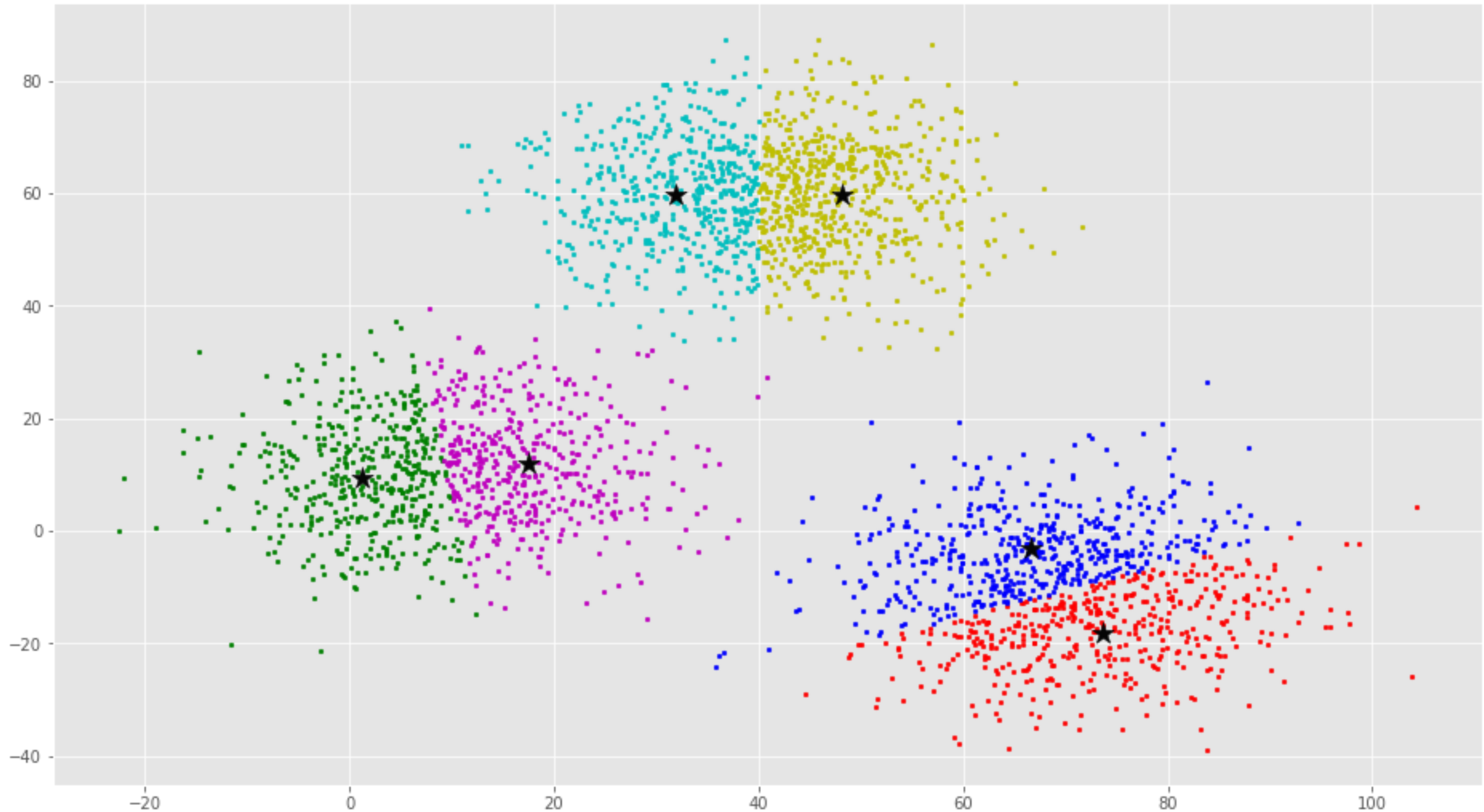
Find similar images to a target one

K-Means Clustering



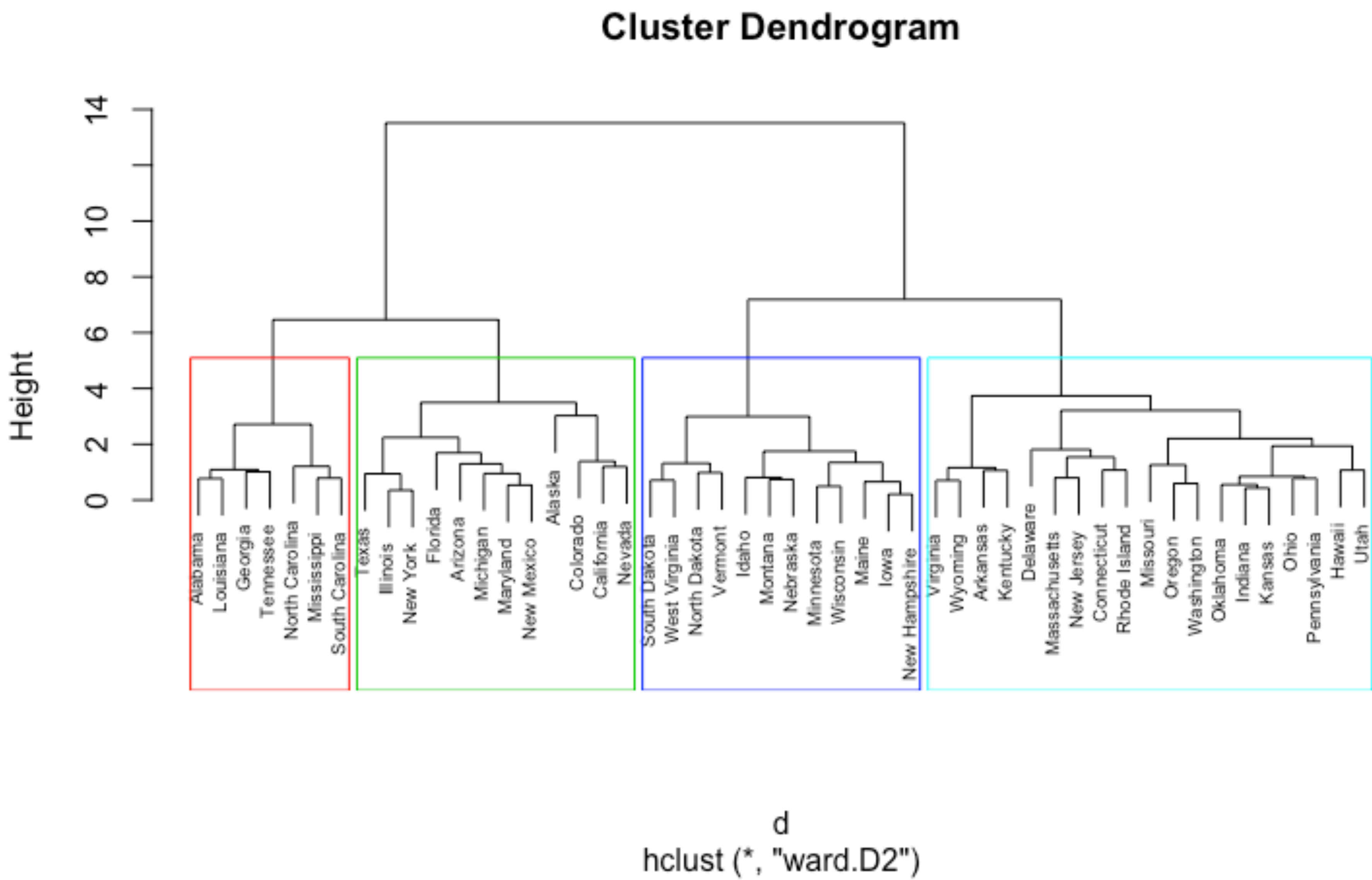
K=3

K-means Clustering



K=6

Hierarchical Clustering



Unsupervised Learning

- Clustering
  – Group similar data points into clusters
  – Example: k-means, hierarchical clustering
- Dimensionality reduction
  – Project the data to lower dimensional space
  – Example: PCA (Principal Component Analysis)
- Feature learning
  – Find feature representations
  – Example: Autoencoders

Supervised Learning Tasks

- Classification
  - Learn to predict class (discrete)
  - Minimize error 1/N $\sum_{i=1}^{N}[y^{(i)} \neq f(x^{(i)})]$
- Regression
  - Learn to predict response variable (numerical)
  - Minimize MSE (Mean Square Error between prediction and actual values)
- Both classification and regression
  - Training and testing phase
  - "Optimal" model is learned in training and applied in testing