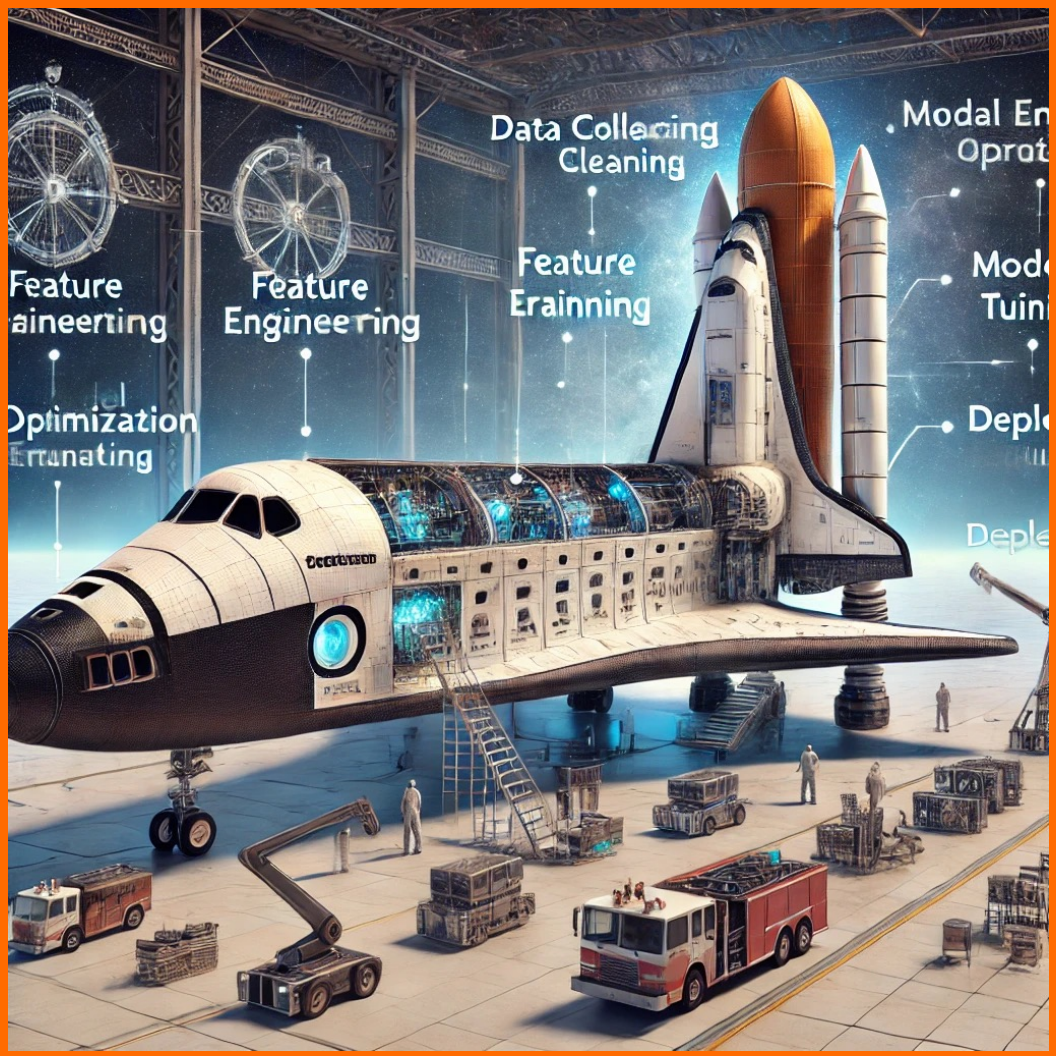
Introduction
Foundations of Machine Learning

Farshid Alizadeh-Shabdiz, PhD, MBA

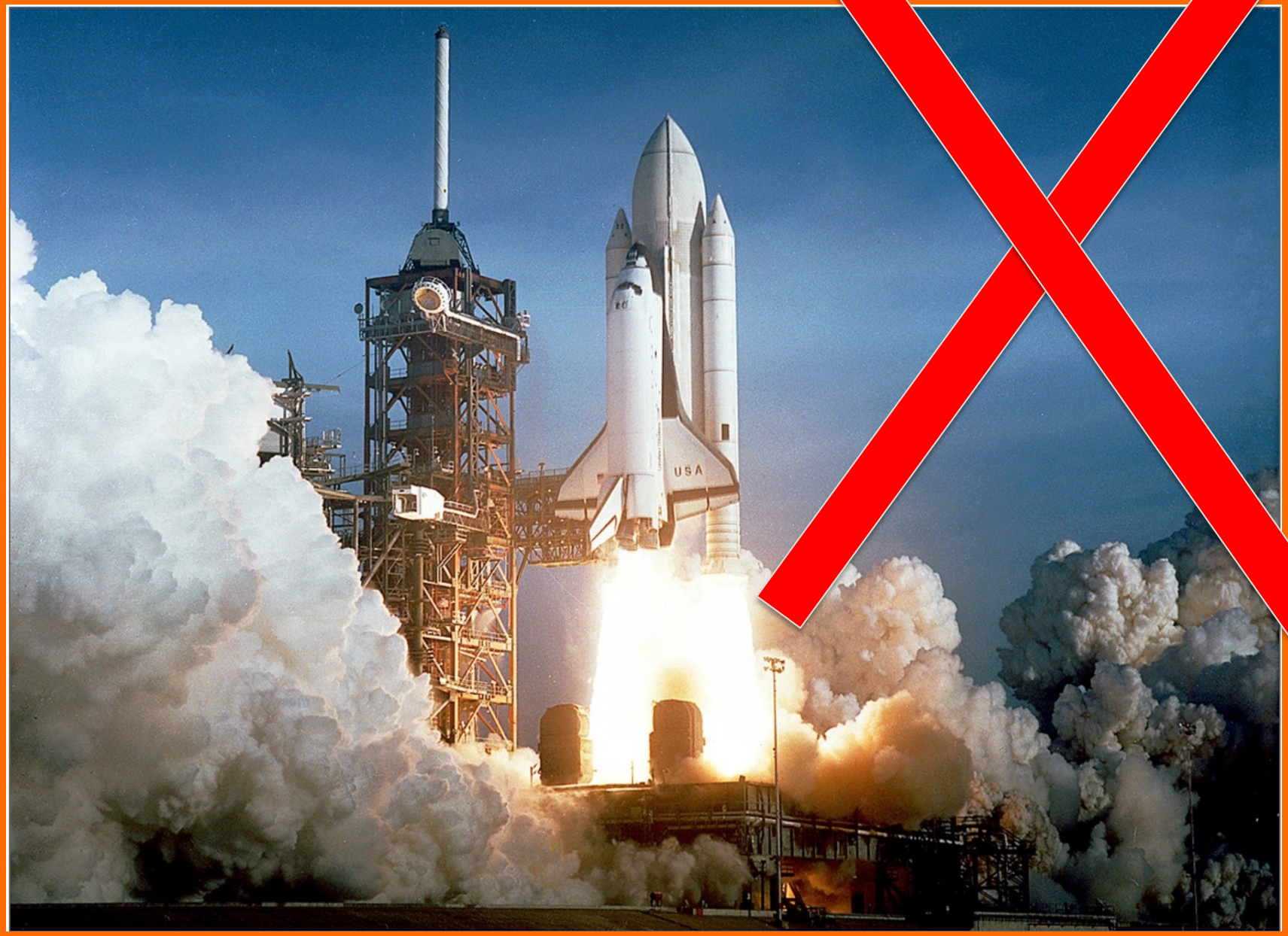
Fall 2024

What Are we covering in this course?

Statistics - which concerns data; their collection, analysis, and interpretation.



What is NOT? Advanced Machine Learning



Examples of The Problems That Statistical Machine Learning Tackle

FiveThirtyEight –
Started as New York Times blog

- An American website that focuses on opinion poll analysis, politics, economics, and sports blogging.

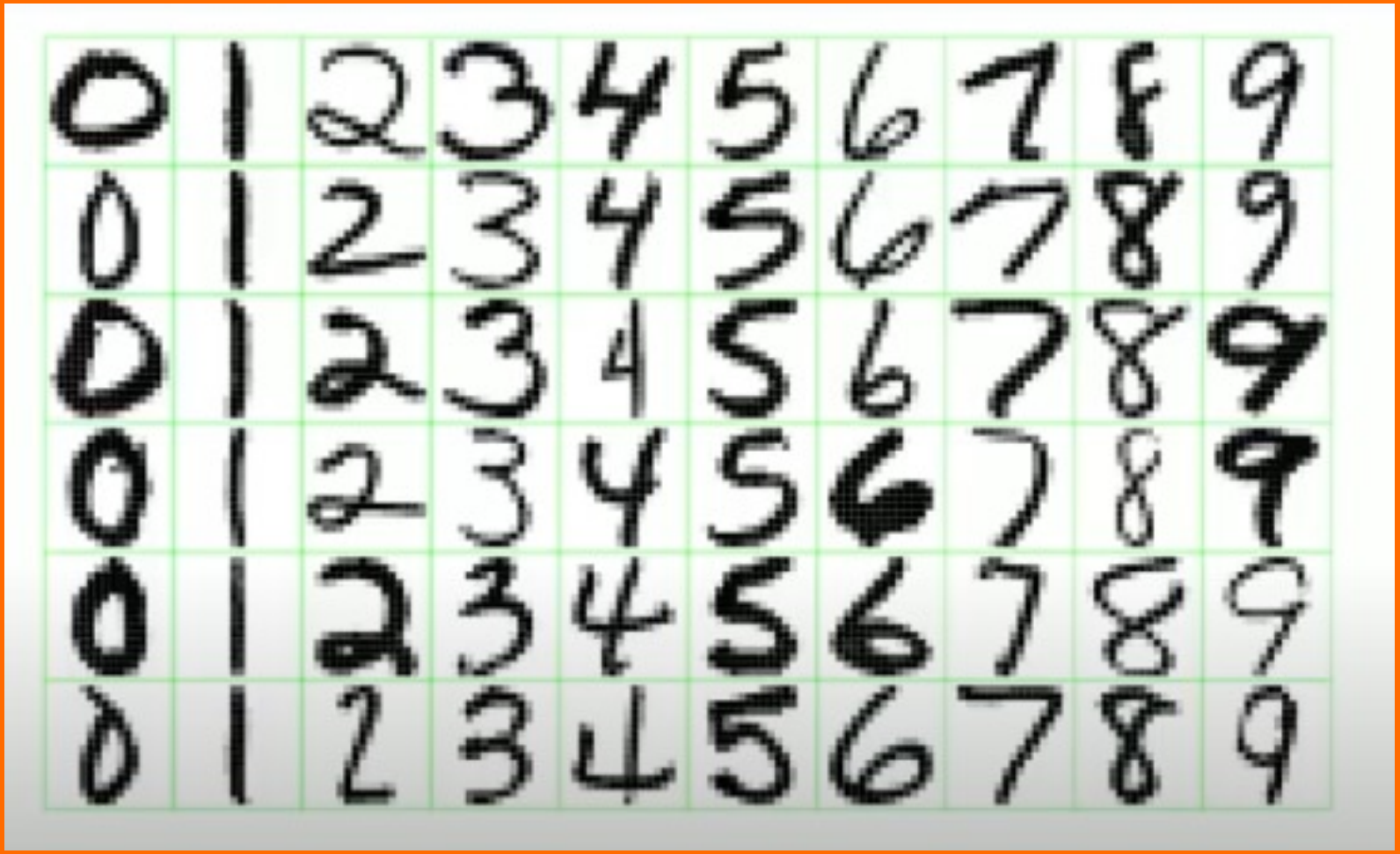
- Successfully predicted 2012 Senate and presidential outcome.
- Got acquired by ESPN and later by ABC.

Identify the risk factors of prostate cancer

- 97 patients collected by Stanford's physician
- Scatter plot matrix of pair of variables

Identify handwritten numbers

- Example is MNIST dataset with 60,000 samples.



Statistical Learning Problems
Identify the risk factors of heart disease

- Framingham heart study (FHS)

"FHS **findings** have informed the understanding of how cardiovascular health affects the rest of the body. The **study** found high blood pressure and high blood cholesterol to be **major risk** factors for cardiovascular disease."

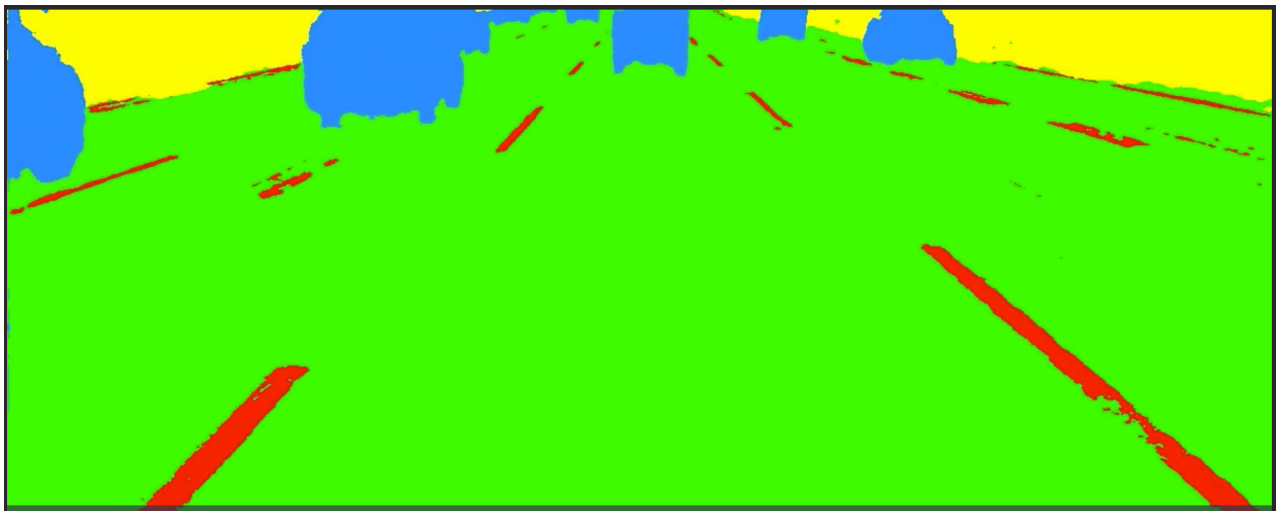
- Initial number of patients 5200
- Resulted to a risk score to estimate 10-year cardiovascular risk, showed risk factors as
  - Age, total cholesterol, smoking, and systolic blood pressure
  - Very small correlation with HDL cholesterol

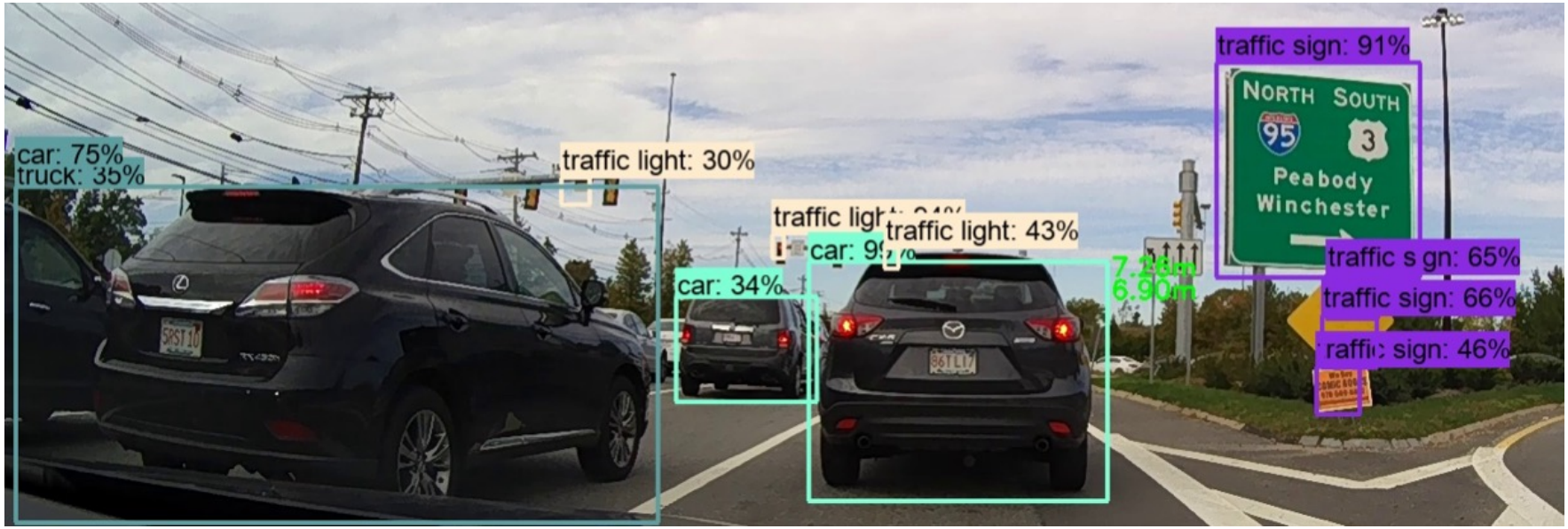
More complex problems

- Detect and read handwriting
- Assessment of online review of items
- Detect cancer images
- Spam detection

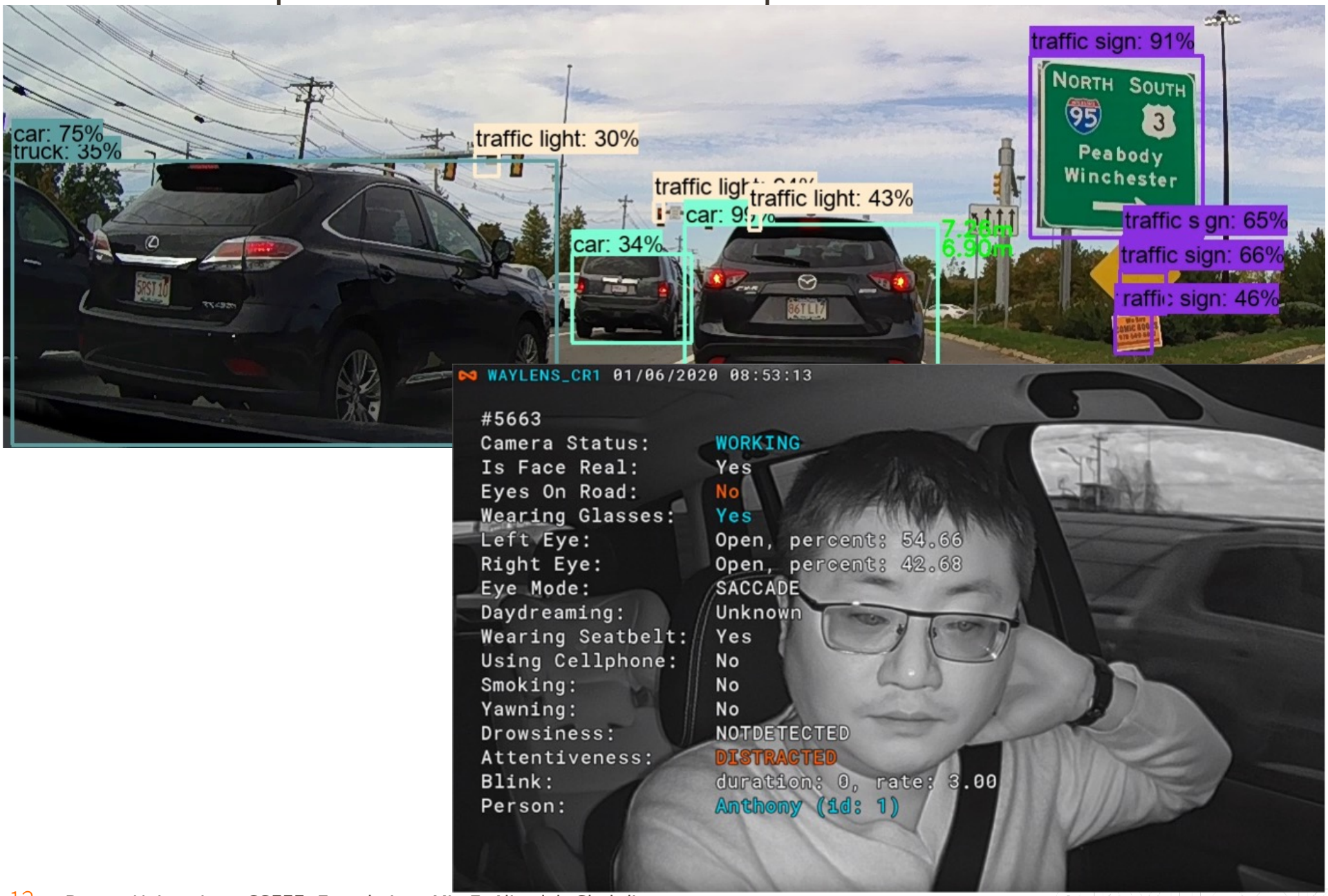
Face Recognition or Mask Detection

- Image segmentation
- Object detection





Computer Vision Object Detection



Social Networks and Ad Companies

- Google
- Facebook
- Twitter
- Netflix

What is Data Analytics?

Learning from data

Summarizing data

Presenting data

Modeling data + its precision and uncertainty

Data Analytics

- Visualization
  - Scatter plot
  - Box plot
  - Etc
- Pair wise correlation
- Analysis

Data or Algorithm

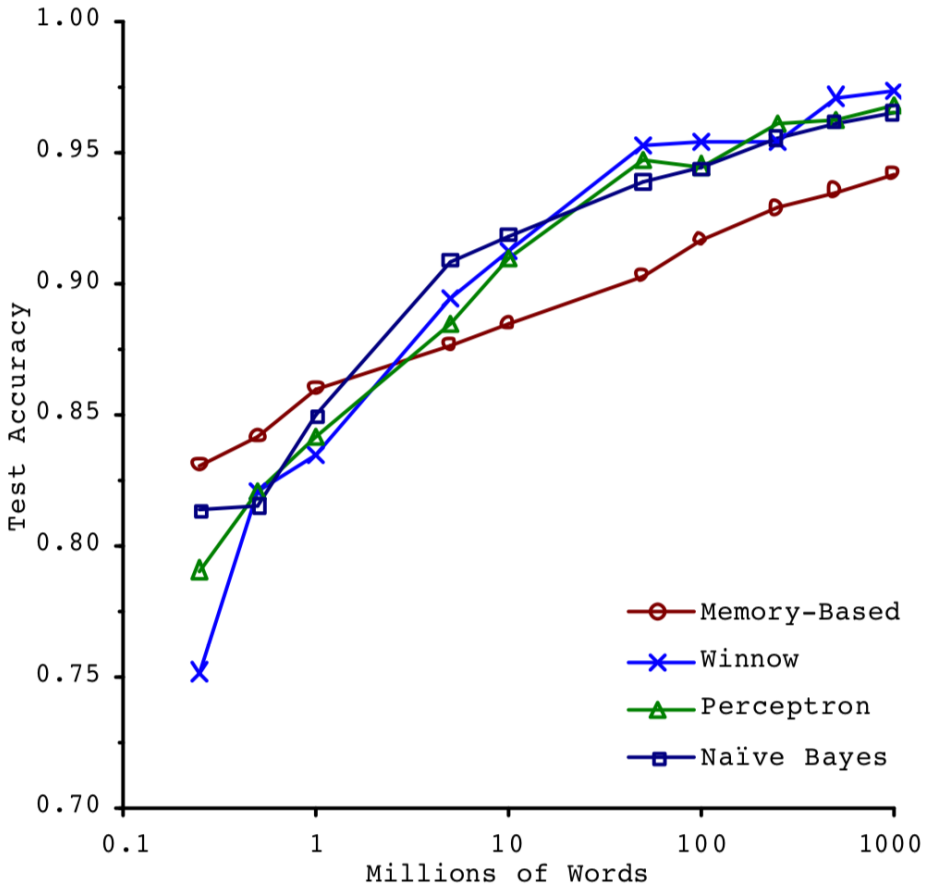


Figure 1. Learning Curves for Confusion Set Disambiguation

Michele Banko, Eric Brill, "Scaling a very very large Corpora for natural language disambiguation", Microsoft Research Lab, 2001

Data Traps

- Over-fitting
- Under-fitting

Administration

Philosophy of the Course

- Build foundation of data analytics
- Learn ideas behind the techniques
  - Helps to apply them correctly – know how and when to use them
- Being able to assess algorithms and help you to go for the simple ones. Ocham's razor!
- Learn basics to be able to learn on your own, since this is a dynamic field and extremely active – so everyday new ideas come out

Introduce yourself

- Name
- MS major & Bachelor major
- A sentence about yourself
- Hobby

Course Material

- Course material on Blackboard
  - Notes and material are the main source
  - Short write-ups & video tutorials
  - Text book
    - An Introduction to Statistical Learning with Applications in R.
      - by: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. (2013) Springer. Available for free online.
  - So many sources on the web

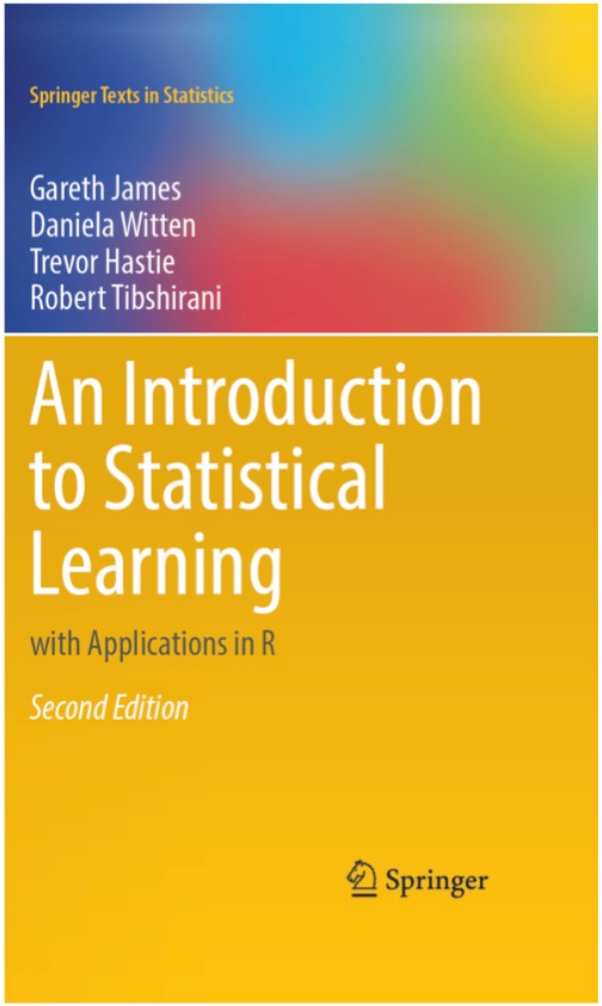
- Academic code of conduct

Text Book

An Introduction to Statistical Learning with Applications in R.

- By: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. (2013) Springer.

The book has been made available online at https://statlearning.com



Additional Books

- Teetor, P. (2019). **R cookbook.** Sebastopol, CA: O'Reilly. ISBN -13: 978-1492040682

  The book has been made available online at https://rc2e.com and

- Chang, W. (2021). R graphics cookbook. Sebastopol, CA: O'Reilly. ISBN 9781491978573

  The book has been made available online at https://r-graphics.org

Admin

- All the material will be posted on Blackboard
- There will be take home and in-class quizzes
- Assignments are due before the next class
  - No late assignment will be accepted
- Office hours: by appointment
- TA
  - **Kunal Vishwa Sivakumar** kvishwa@bu.edu – section A4
  - **Saya Atchibay** sayokit@bu.edu – **section A3**
  - TAs office hours will be announced by TA

Grading

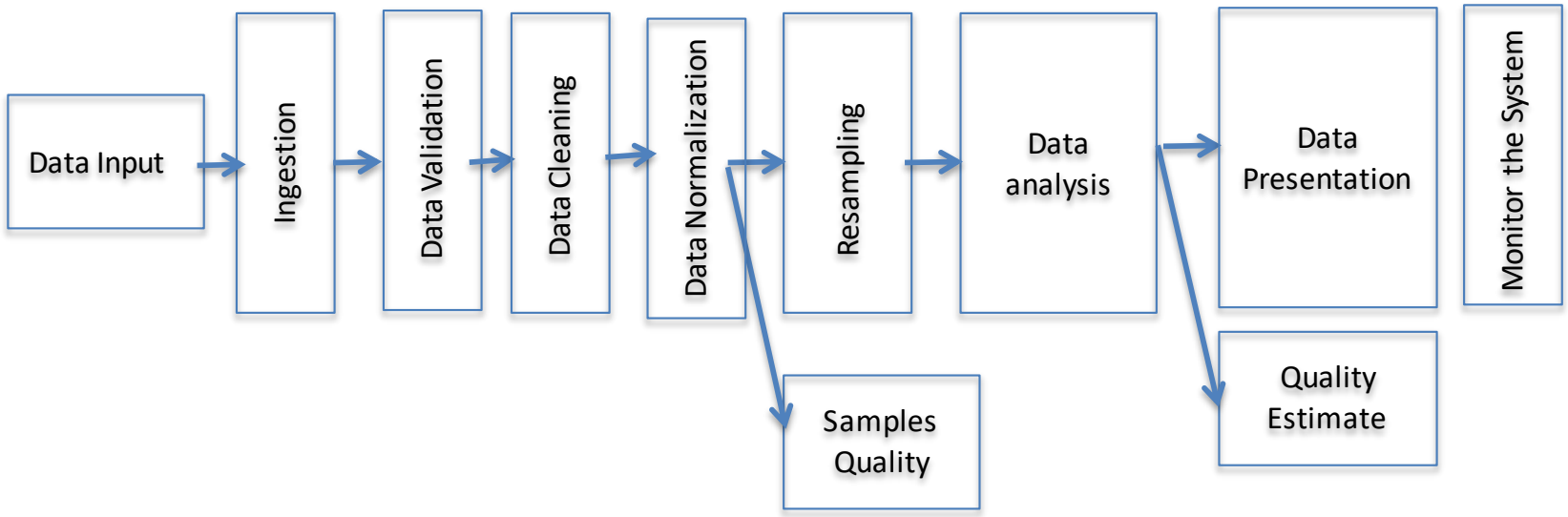
- Grading
  - Assignments – 25%
  - Quizzes – 30%
    - Take home and in-class quizzes
    - One page of cheat sheet
    - One quiz will be dropped
  - Term project – 15 %
    - To apply what you have learned to a semi-real project
    - You choose a dataset
    - Projects are individual
  - Final exam – 30%
    - From the entire material
    - Closed book, closed notes, and closed laptop
    - Five pages of cheat sheet

Why R Programming?

- Learn through practice
- R vs Python
  - R is simple and quick for prototyping
  - Python is little bit more work, but still easy. More suitable for production than R

Data Analytics

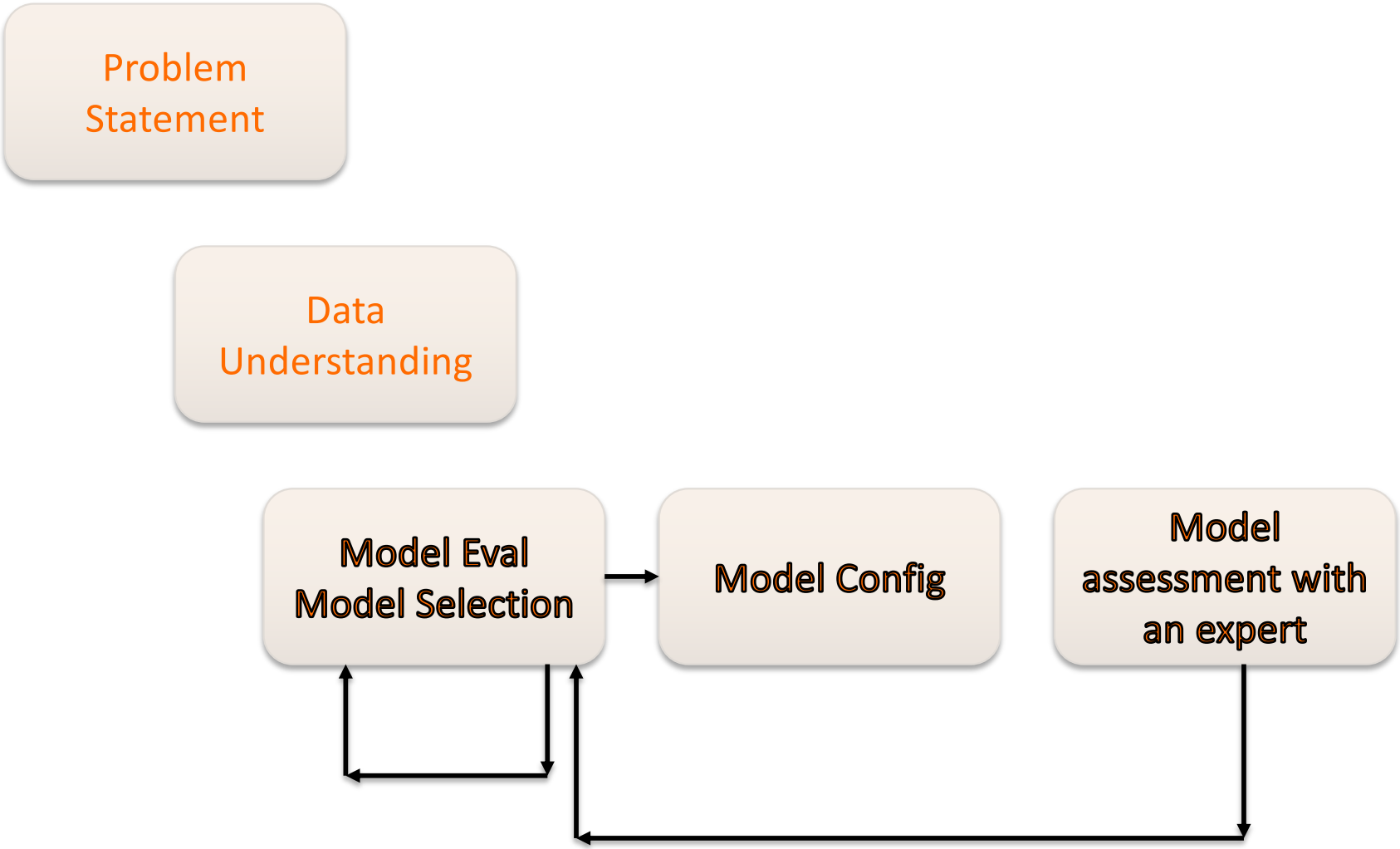
Flow Chart of Data Analytics System



Process of Dev of an Algorithm

1. Data Understanding
2. Problem Framing
3. Data Preparation
4. Model Evaluation and selection
5. Model Configuration
6. Model assessment with experts – testing and presenting the model to experts
7. Model deployment

Algorithm Process



Supervised Learning Problem

- Y: Outcome measurement - it is also called dependent variable, response, and target.
- X: Vector of p predictor measurements – it is also called inputs, regressors, features, or independent variables

Supervised Learning Problem

- Suppose we observe Y_i and $X_i = (X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, n$
- We would like to extract relationship between *Y* and *X*'s.
- We will model the relationship as

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

- Where *f* is an unknown function and ε is a random error with mean zero.

Objective of Statistical Learning

- Prediction: Estimate f from data to predict general population from the seen sample set
- Inference:
  - Understand how each input(s) affect the output
  - Understand relationship between Y and each predictor
  - Understand relationship between inputs
- Note: also quality of estimate

Example: Predicting Median House Price

- Predicting median house price based on house features and location
- Probably want to understand which factors have the biggest effect on the response and how big the effect is.
- E.G. assessing impact of number of bedrooms or water view on the house value.

How to Estimate f?
Parametric vs non-parametric

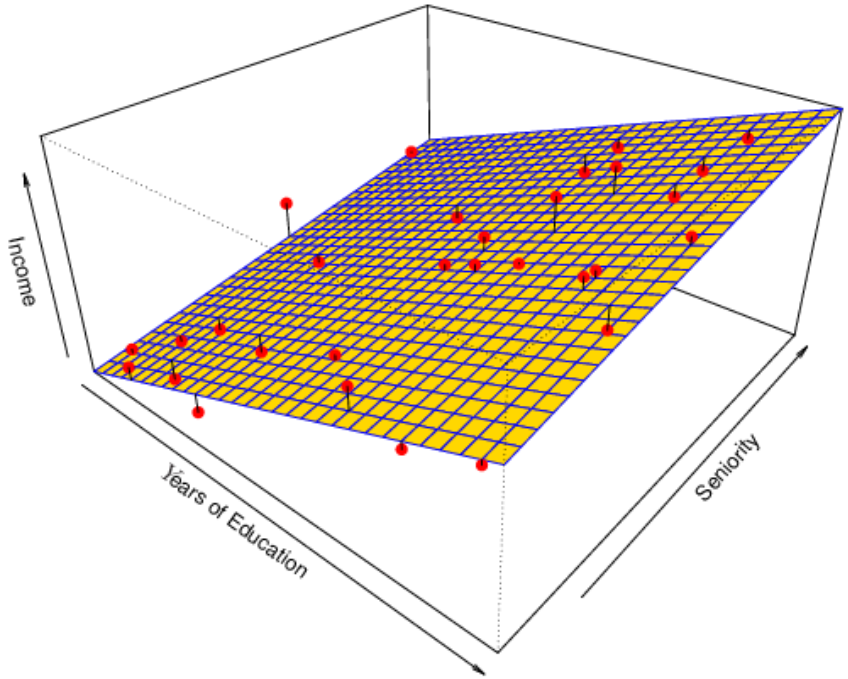
- Observation of some data points
- Model
  - Parametric – assume a model for *f(x), e.g. linear model*

$$f(\mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

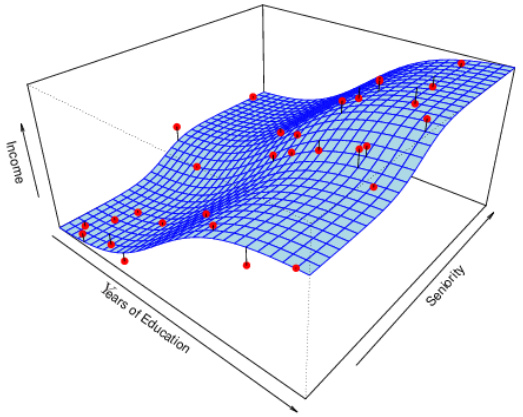
  - Non-parametric
- Using the training data , estimate the unknown parameters of the model, e.g. $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

Example of a Parametric Estimation

- Income vs Education & Seniority



$$f = \beta_0 + \beta_1 \times Education + \beta_2 \times Seniority$$



Non-Parametric Methods

- No explicit assumptions about the functional form of *f*.

- Advantages: They accurately fit a wider range of possible shapes of models.
- Disadvantages: A very large number of observations is required to obtain an accurate estimate of the model.
- Level of smoothness has to get selected

Unsupervised Learning

- No Y: No Output measurement
- Objective: finding of samples that behave similarly
- Assessment is harder than supervised learning
- Why unsupervised learning
  - Unlabeled data is readily available, and labeling is expensive
  - Learn from data – there is no desired output

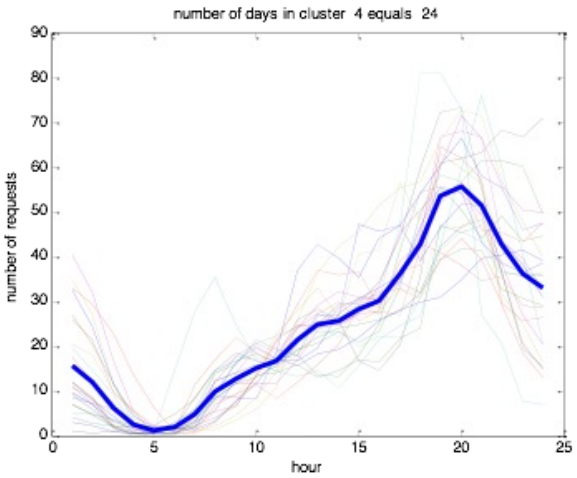
Note: It can be used as pre-processing step for supervised learning

Example of Unsupervised Learning

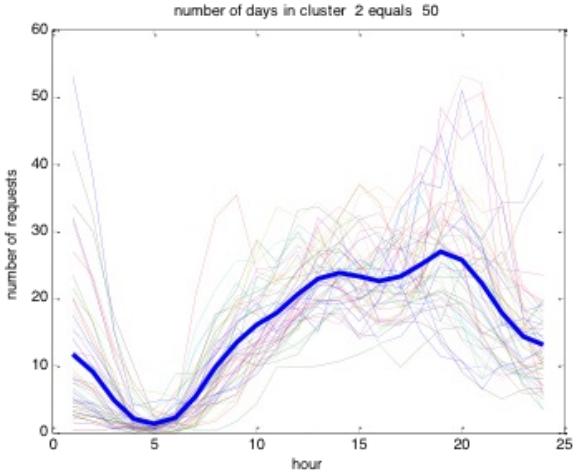
- Extracting traffic patterns around schools in Boston Metro area

Location Based Behavior Analysis

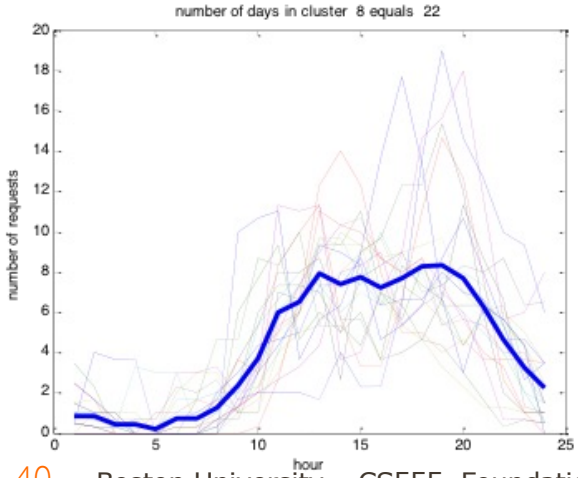
Friday & Sat



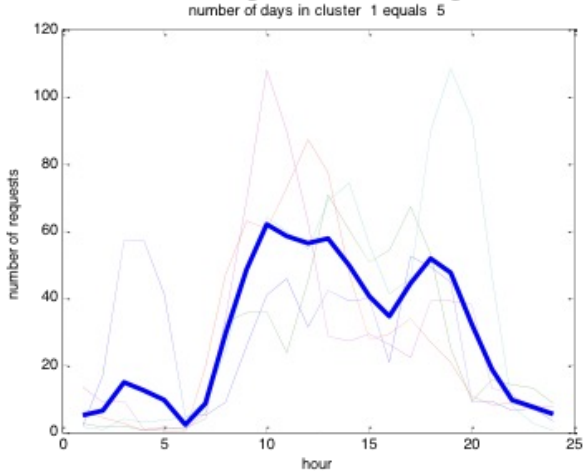
Normal weekday



Sunday - Suburb



Crazy Monday



Why bother with Parametric?

- It results to simpler model and more interpretable
- It might be more accurate – more complex model sometimes performs worse (counter intuitive !)

Supervised Learning – Regression vs Classification

- Supervised learning problems can be further divided into
  - Regression
  - Classification
- Regression covers situations where Y is numerical. e.g.
  - Predicting the risk of heart attack in 10 years.
  - Predicting the value of a given house based on various inputs.
- Classification covers situations where Y is categorical e.g.
  - Will a patient have cancer or not?
  - Is this email a SPAM or not?

High Level View of Data Analysis