DSC5103 Statistics

Session 1. Introduction

Tong Wang
NUS Business School

About me

- Associate Professor in Dept. of Analytics & Operations, joined NUS in 2008
- Office: Biz 1 (Mochtar Riady Building) 8-68
- Tel: 6516-1356
- Email: *tong.wang@nus.edu.sg*
- Web: *http://www.bschool.nus.edu.sg/staff/bizwt/*

- Education
  - 2008, **Ph.D** in Decision Sciences, INSEAD, France/Singapore
  - 2003, **M.Phil** in Systems Engineering, CUHK, Hong Kong
  - 2001, **B.Eng** in Industrial Engineering, SJTU, Shanghai

- Research
  - Retailing Operations, Supply Chain Management, Pricing and Revenue Management
  - Bayesian learning, joint estimation and optimization, sensors
  - Football Analytics

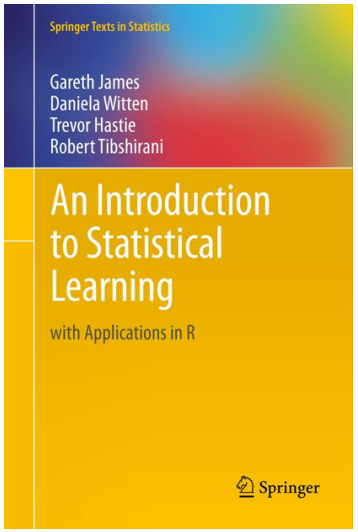
- Teaching
  - Data Analytics (BBA, MSBA, PhD), Dynamic Pricing and Revenue Management

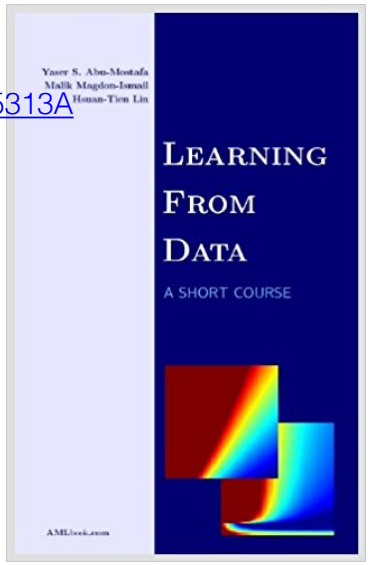
Admin Matters

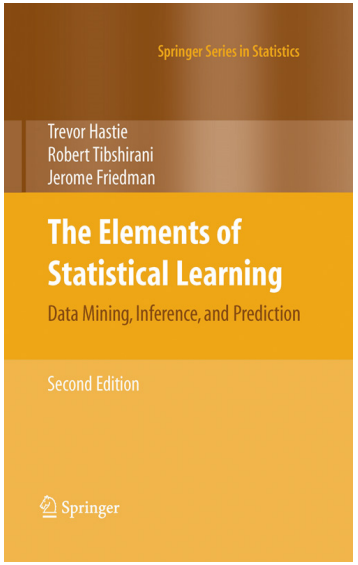
- Coverage: modern statistical learning
  - Classic statistics and R programming as prerequisites

- Textbook
  - *An Introduction to Statistical Learning --- with Applications in R*
    - http://www-bcf.usc.edu/~gareth/ISL/index.html
    - Youtube Videos
  - References:
    - *Learning From Data*
      - http://amlbook.com/
      - https://www.youtube.com/playlist?list=PLD63A284B7615313A
    - *Elements of Statistical Learning*
      - http://statweb.stanford.edu/~tibs/ElemStatLearn/

- Assessment
  - Class Participation (individual) 10%
  - Assignment (group) * 6 30%
  - In-class Test (individual) 40%
  - Final project (group) 20%





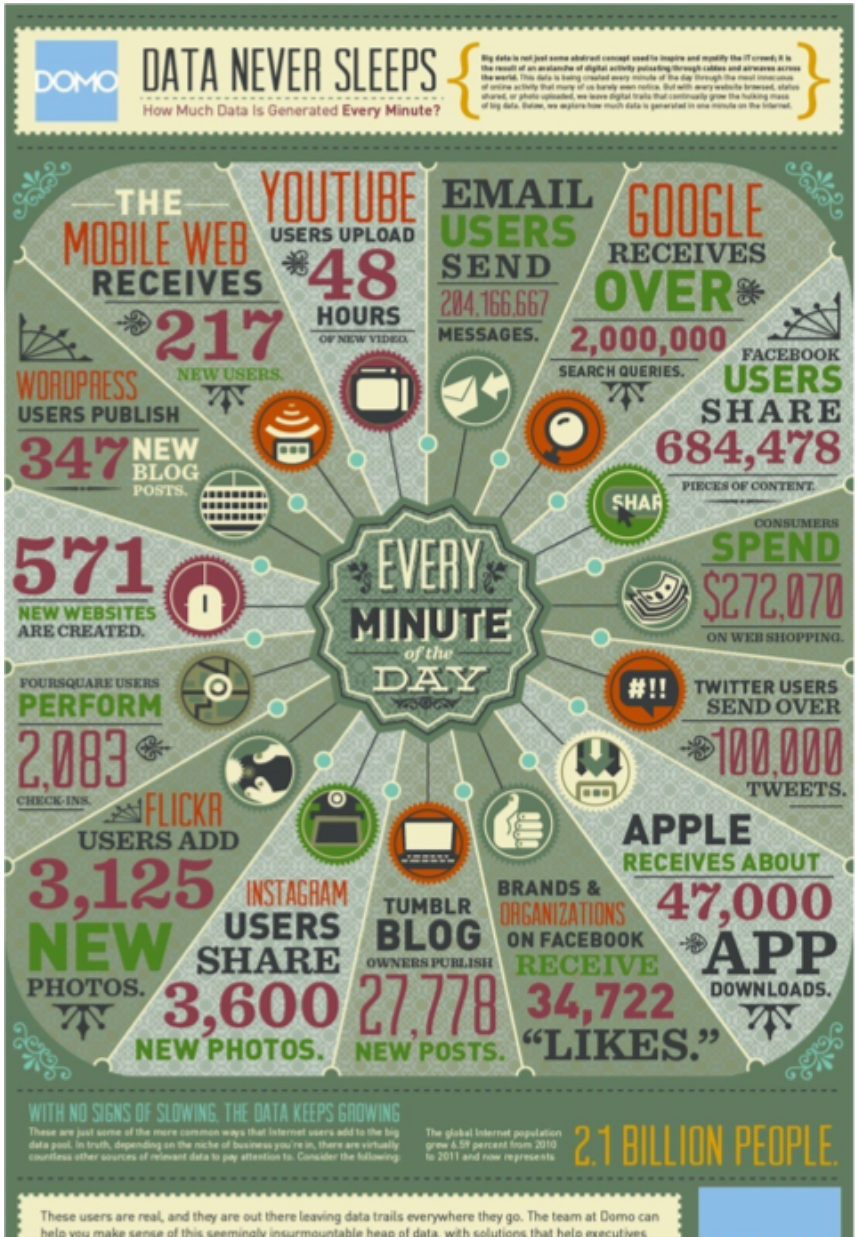


Admin Matters

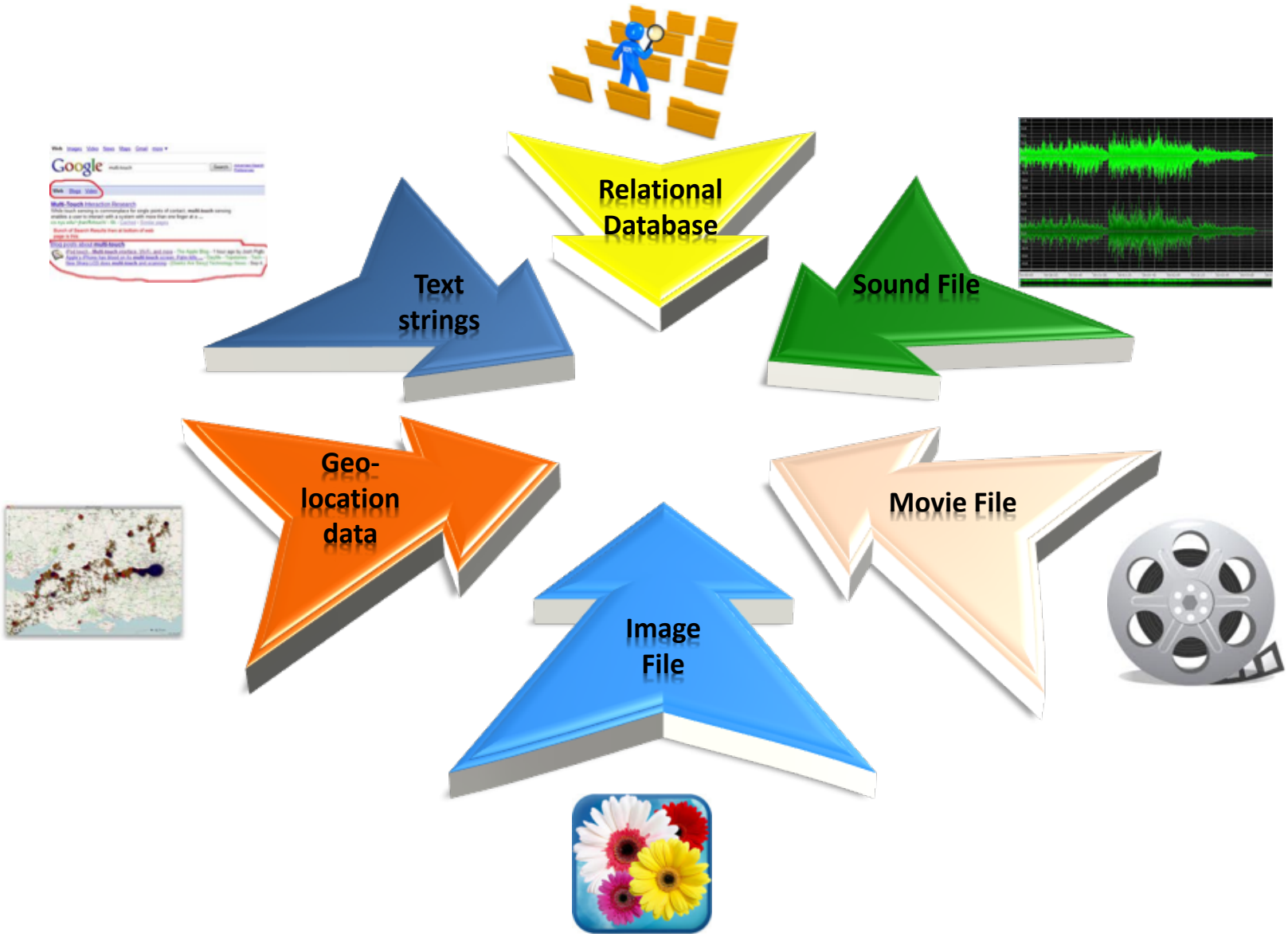
- The IVLE Platform: https://ivle.nus.edu.sg
  - Follow ***Lesson Plan*** for weekly plan and relevant file download
  - Upload assignments/projects using ***Files***
  - Check ***Gradebook*** for grades/comments
  - ***Web Lecture*** for video recordings
  - ***Forum*** for Q&A and discussion
    - NO EMAIL!!
  - ***Anonymous Feedback***

OVERVIEW

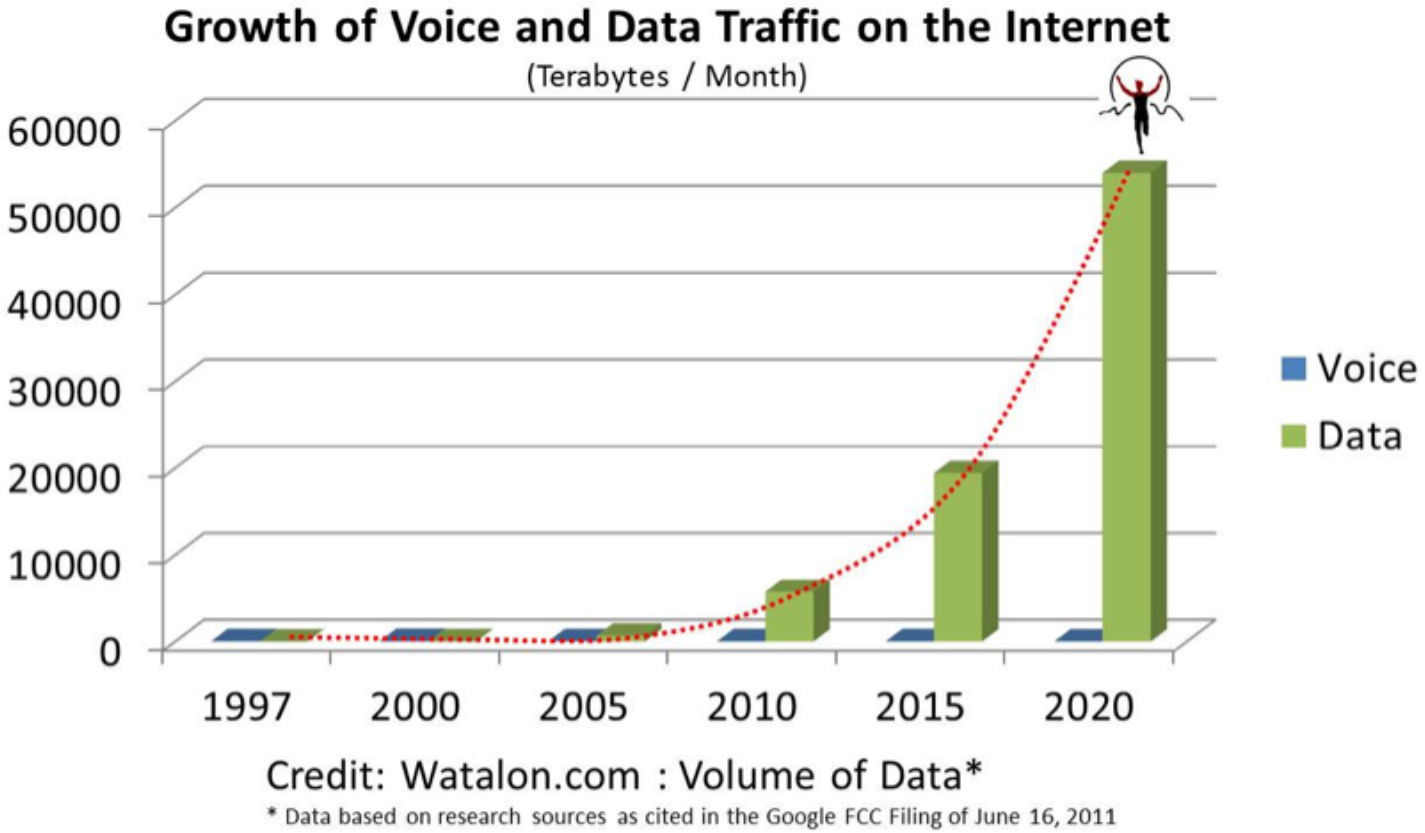
The Age of Big Data - Volume



The Age of Big Data - Variety



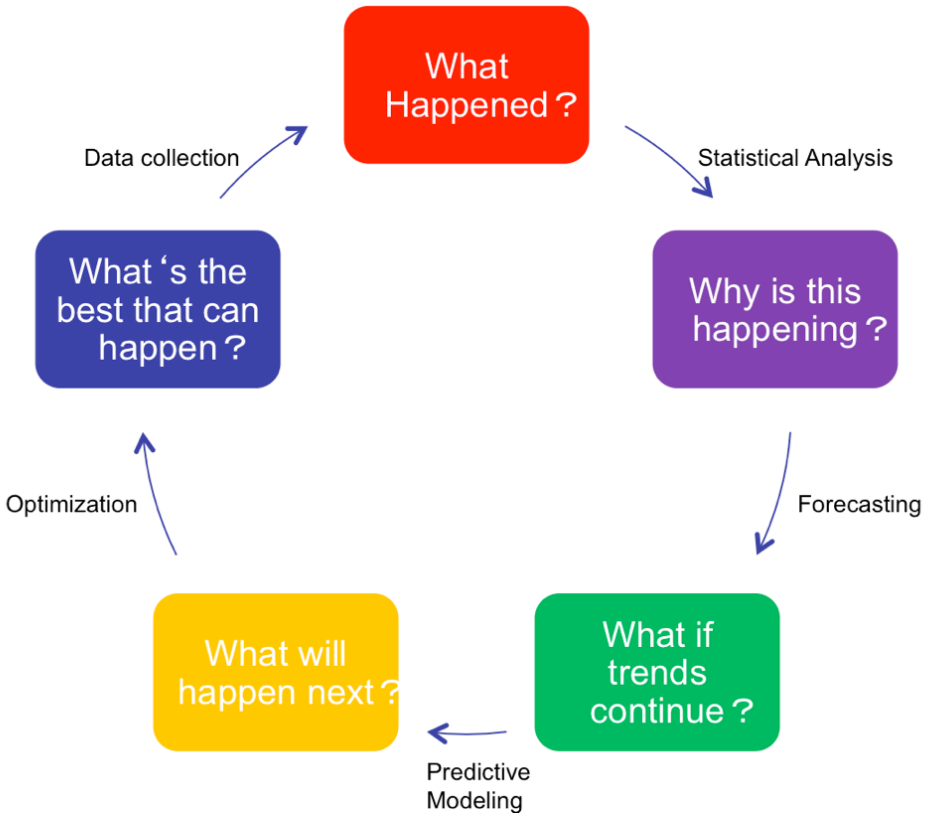
The Age of Big Data - Velocity

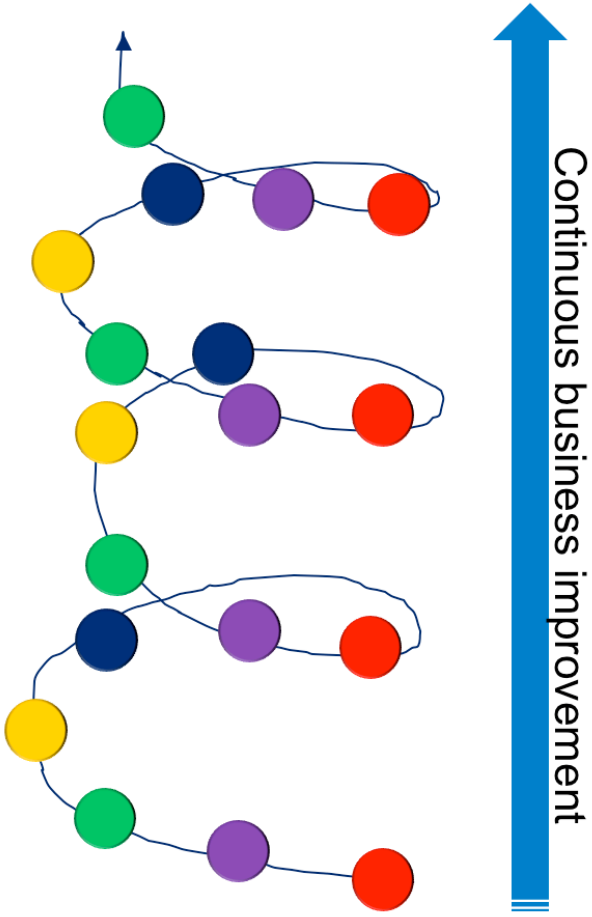


Business Analytics

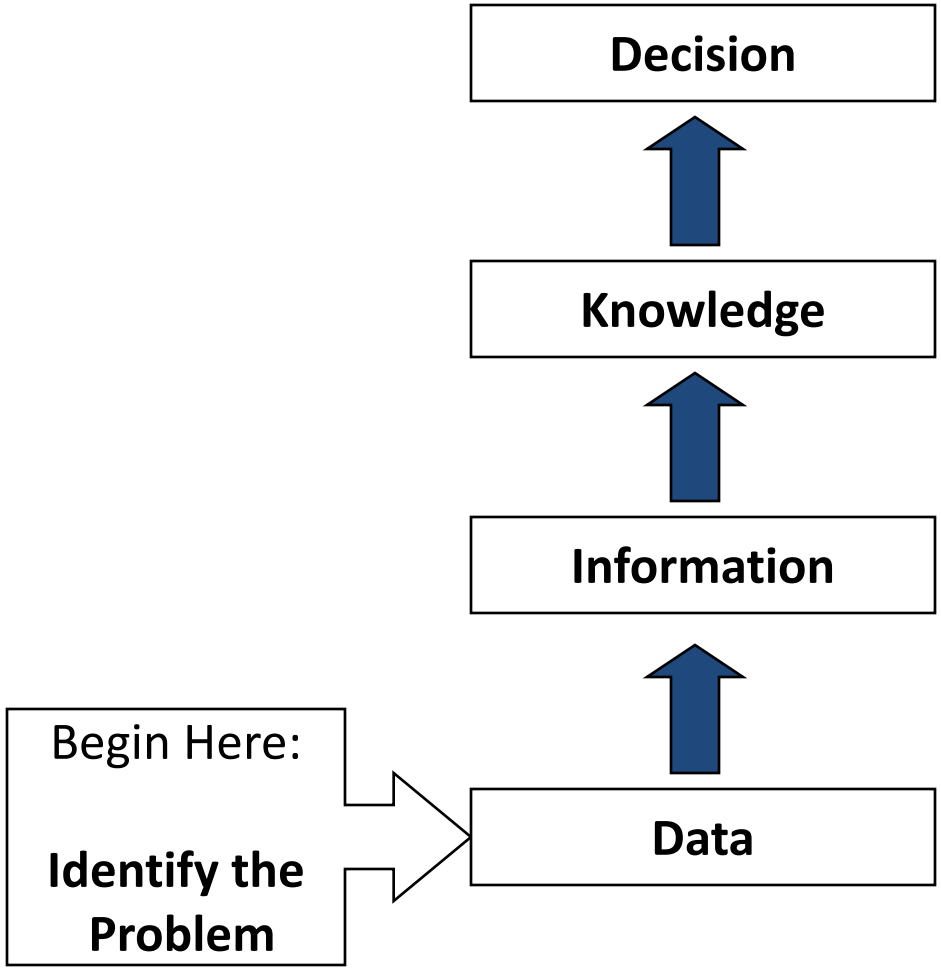
- Wikipedia
  - Business analytics (BA) refers to the skills, technologies, applications and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.

Business Analytics Workflow



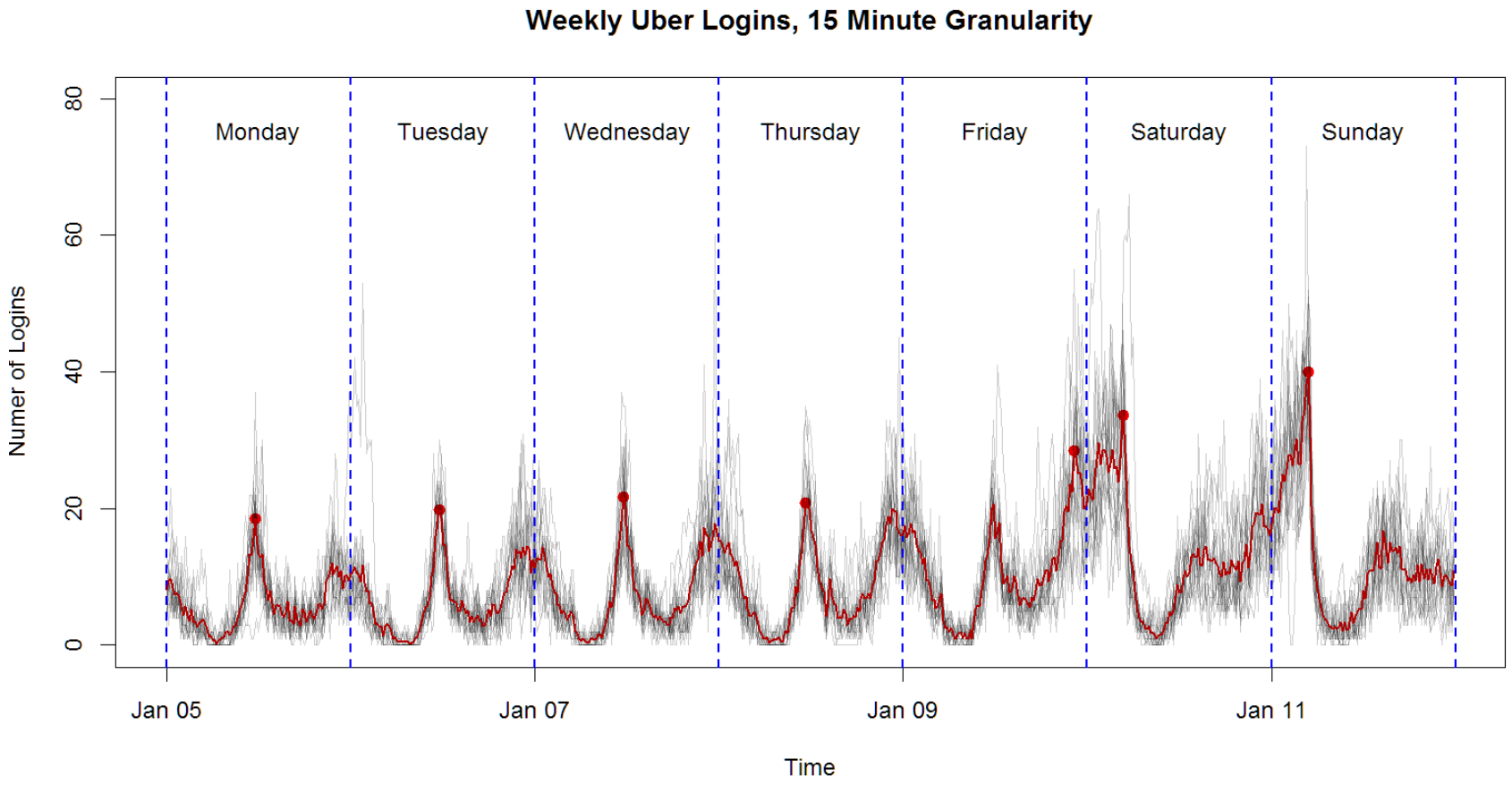


Data-Driven Decision Making Process

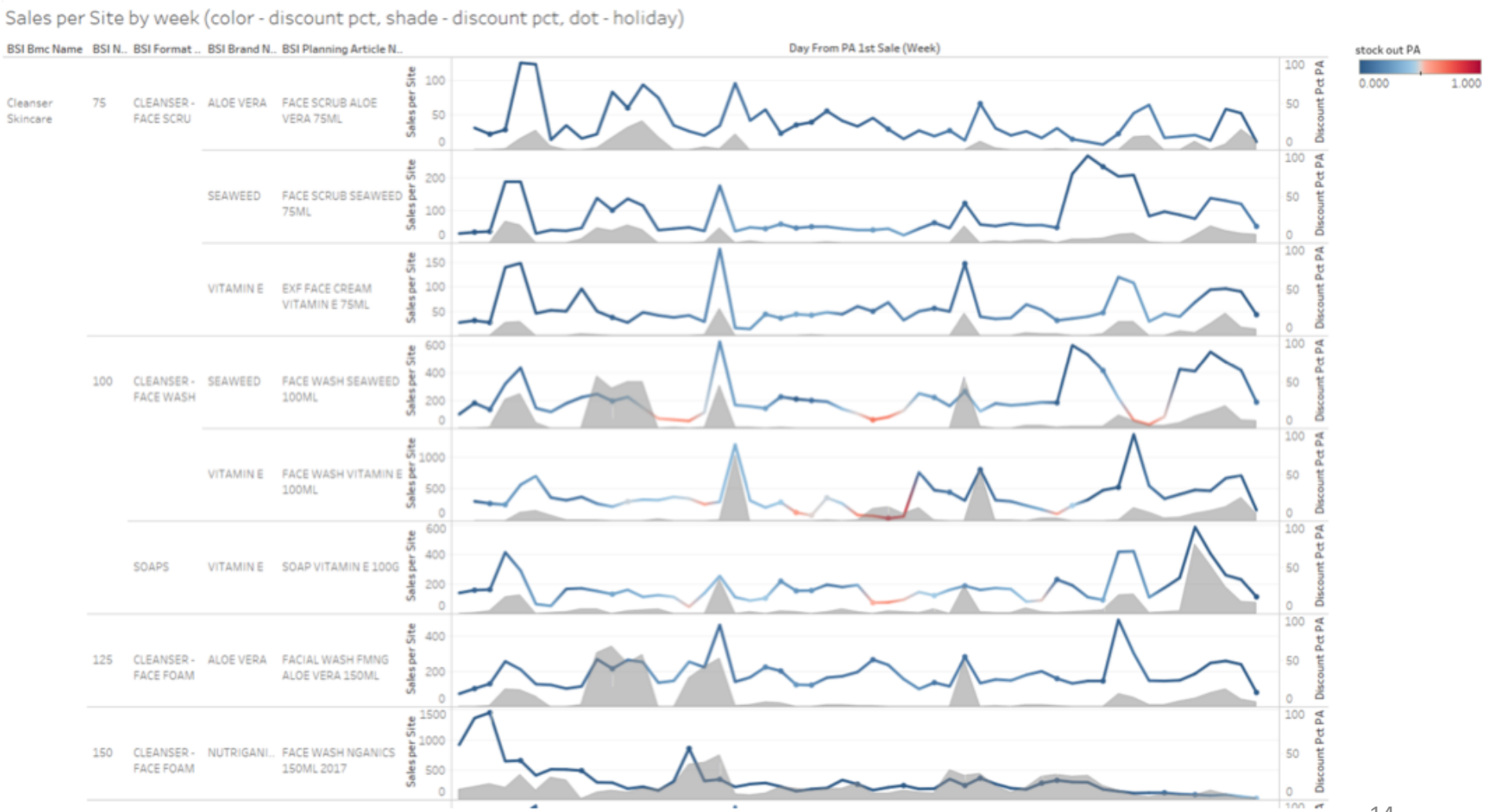


Example 1: Target's prediction

Example 2: Uber's demand

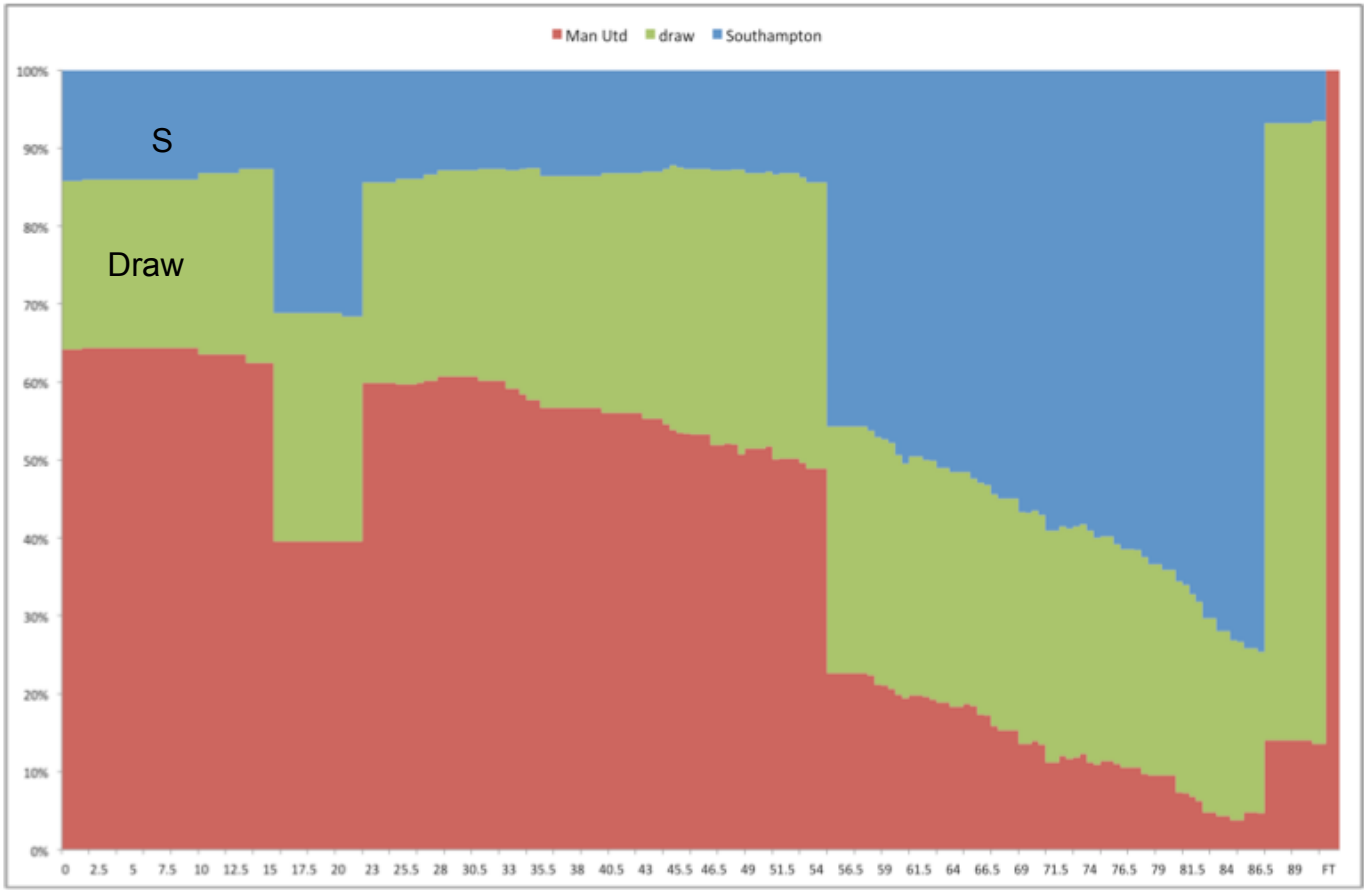


Example 3: Retail Operations



Example 4: Soccer gambling

Southampton 2 - 3 Man Utd (EPL 2012-09-02)



min 0

min 90

Time	Goal
15:43	Southampton
22:43	Man Utd
54:58	Southampton
86:53	Man Utd
91:09	Man Utd

Learning

- Learning (from data)
  - Based on observations of external evidences
  - Data-driven
  - Examples:
    - Cognitive process: recognize a tree
    - Physics: F = m*a
    - Statistics: Weight ~ Height
  - Exercise: credit rating

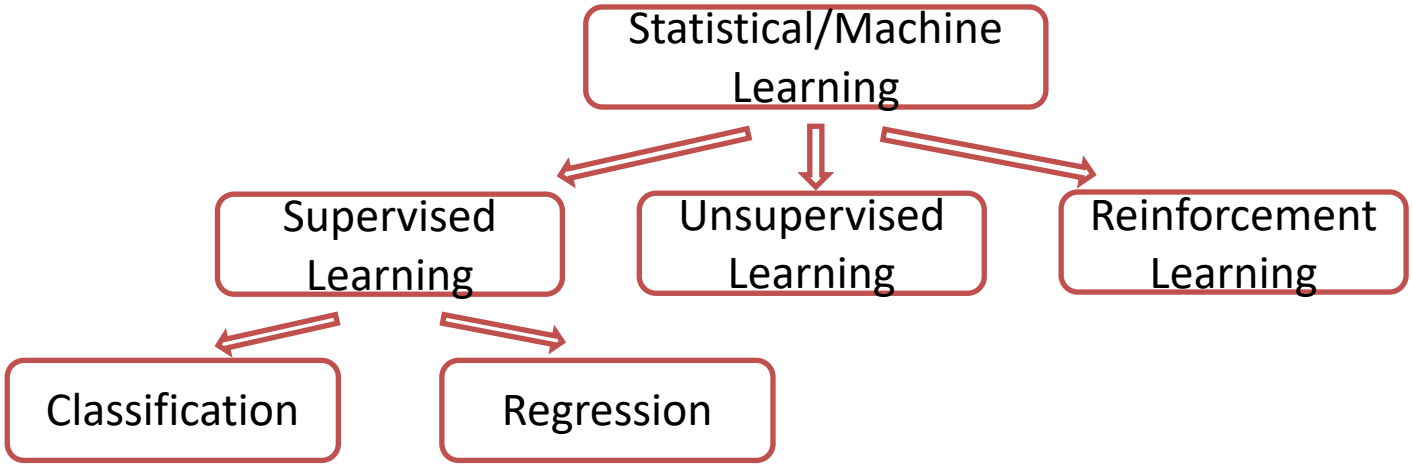
    default~age+education+income

design

- Modeling (from specifications)
  - Based on knowledge of internal mechanisms
  - Model-driven
  - Examples:
    - Probability theories: P(Head) = 0.5
    - Math: 11 x 12 = 132
    - University ranking

score=f(data)

Statistical/Machine Learning



- Reinforcement learning
  - Partial / ambiguous / delayed feedback
  - E.g., AlphaGo, http://www.nature.com/nature/journal/v518/n7540/full/nature14236.html#videos

Supervised Learning

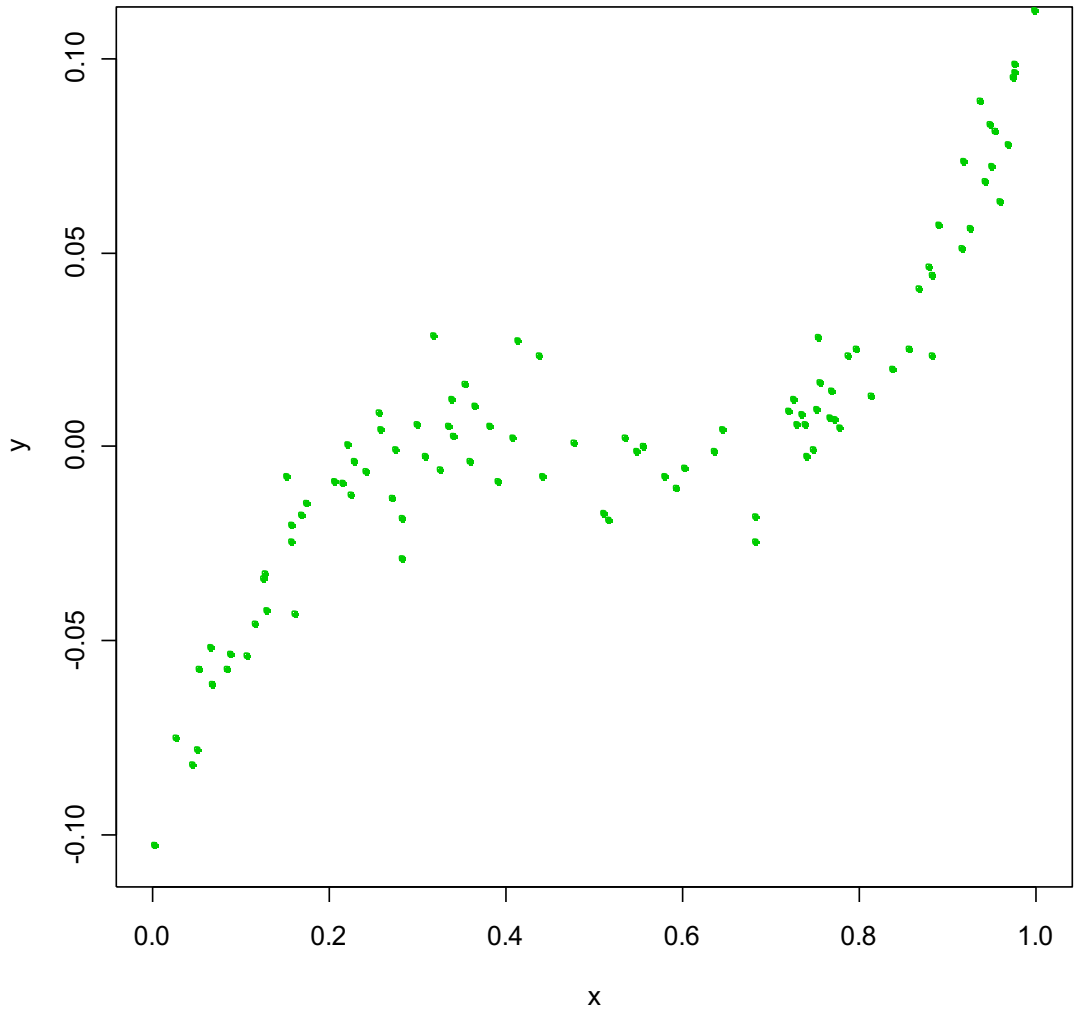
- We are interested in Y
  - outcome, dependent variable, response, target
- There is a vector of X
  - inputs, independent variables, features, covariates, predictors
- We have data $(x_1, y_1), \ldots, (x_N, y_N)$
  - observations, samples, instances, data points

Supervised Learning

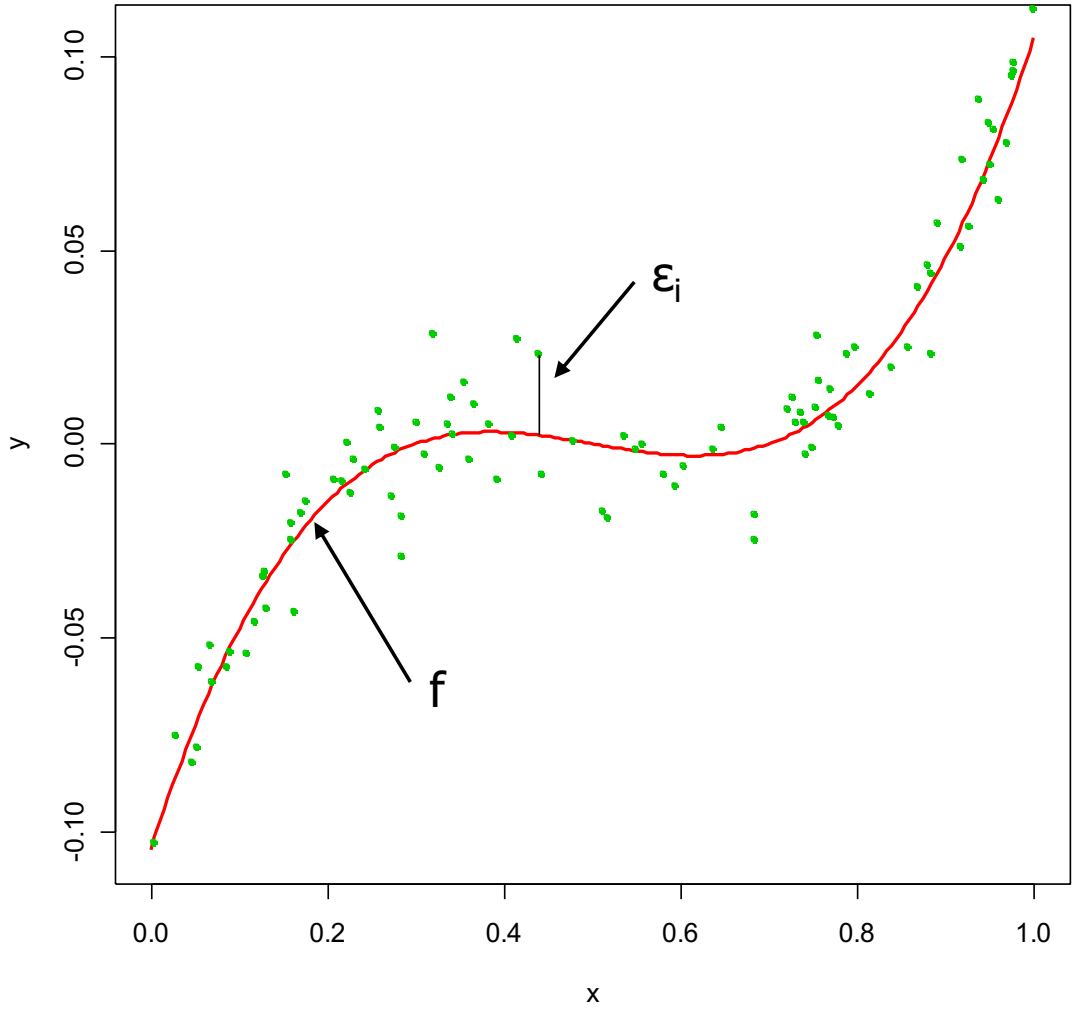
- We believe there is some relationship $Y = f(X) + \varepsilon$
  - $f()$: systematic information that X provides about Y
  - ε: the error term

- Based on the data, we'd like to be
  - Descriptive: understand how and how much X affects the outcome, isolate the effect of $f()$ and the effect of ε
  - Predictive: accurately predict unseen Y' if we have a new X'
  - Assess the quality of our predictions and inferences

A Simple Example

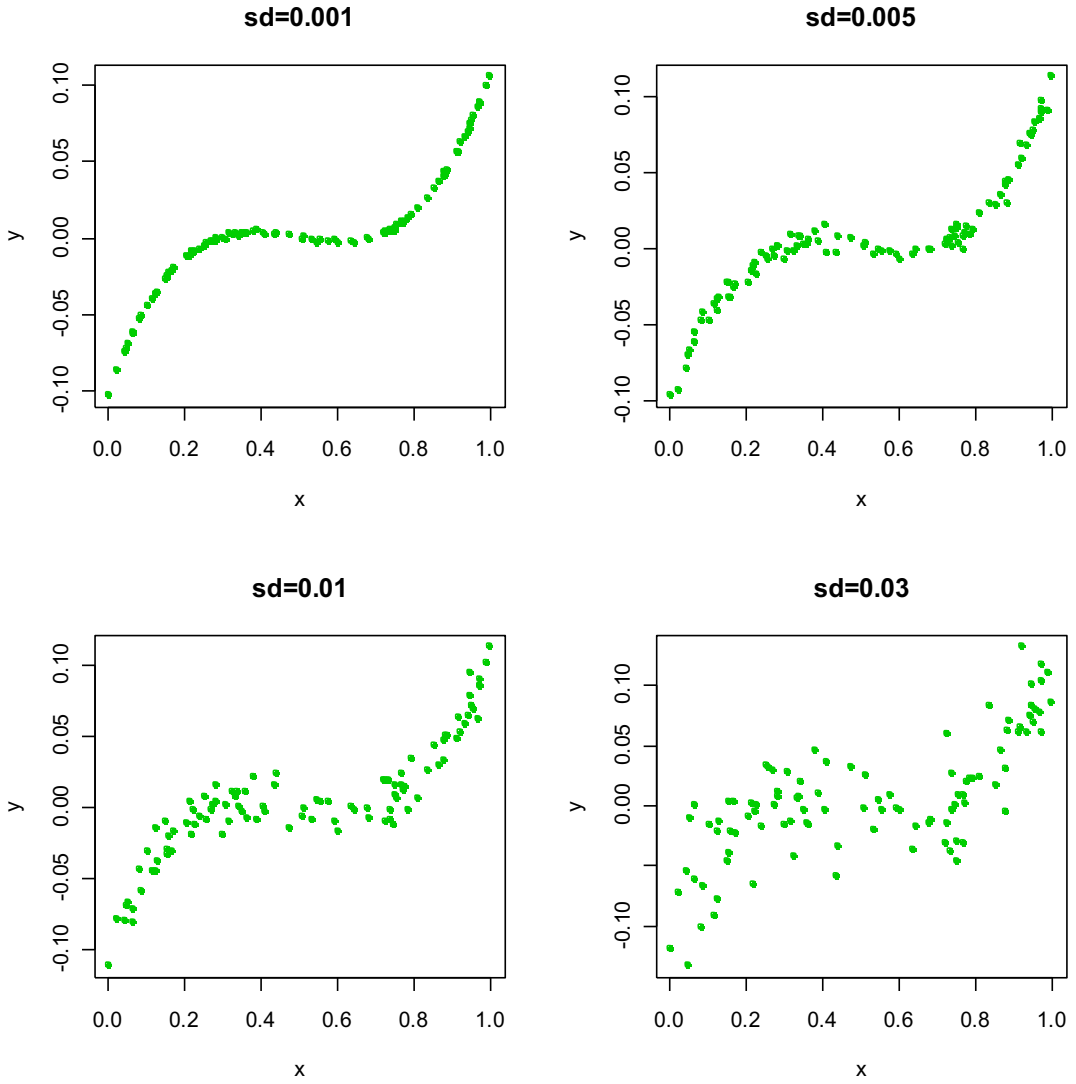


A Simple Example

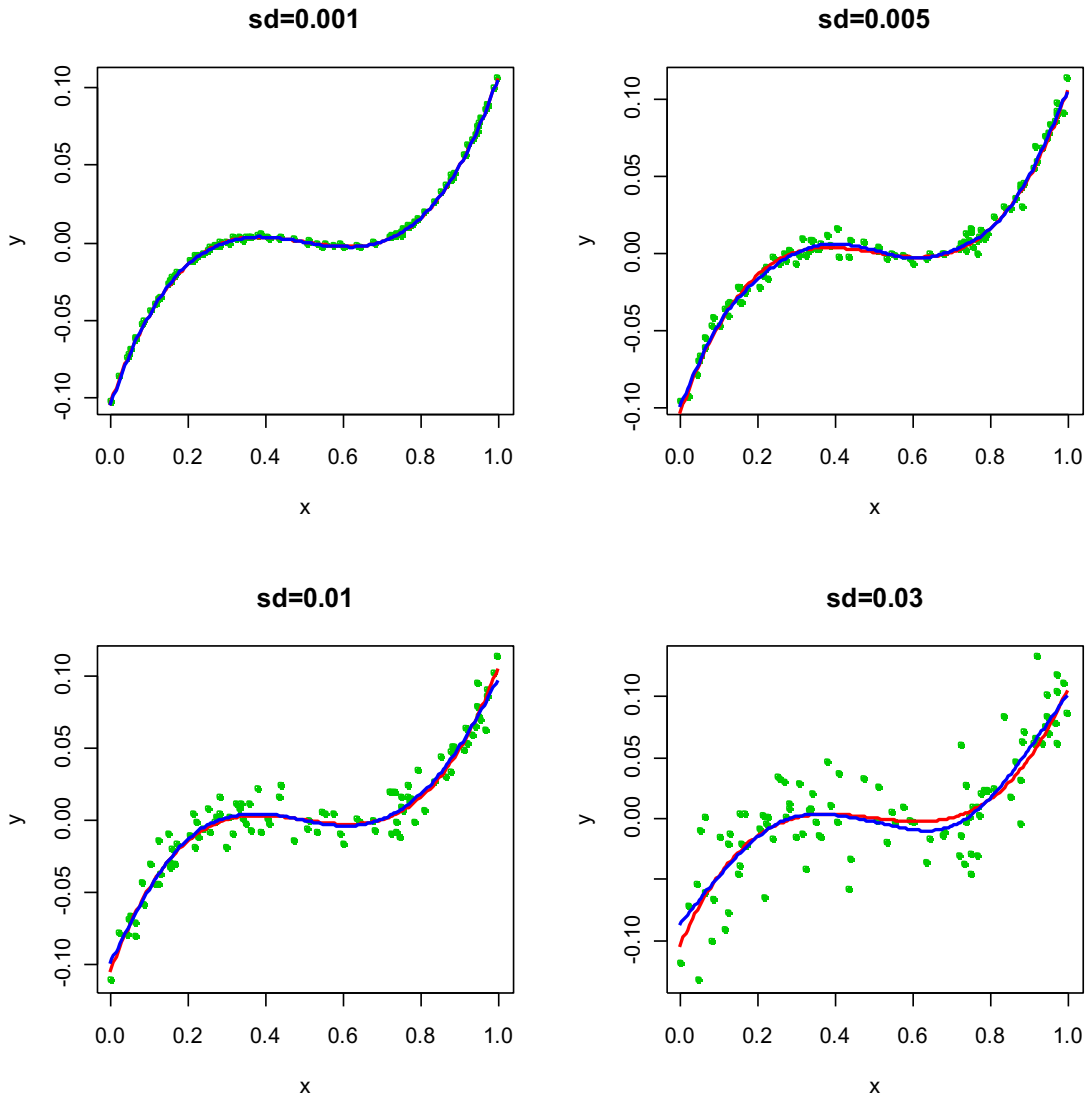


Different Variabilities

- The difficulty of estimating f will depend on the variability of the ε's.



Different Estimates For f



Regression vs. Classification

- Supervised learning problems can be further divided into regression and classification problems

- Regression: *Y* is numerical
  - Demand of a product, stock price

- Classification: *Y* is categorical
  - Whether a coupon be redeemed
  - Will Man Utd beat Southampton?
  - Color identification

- How about discrete *Y*?
  - The ~~score~~ of Man Utd vs. Southampton

    total number of goals

binomial
binary {0,1},{T,F}

multinomial {W,LD}
{R,G,B}

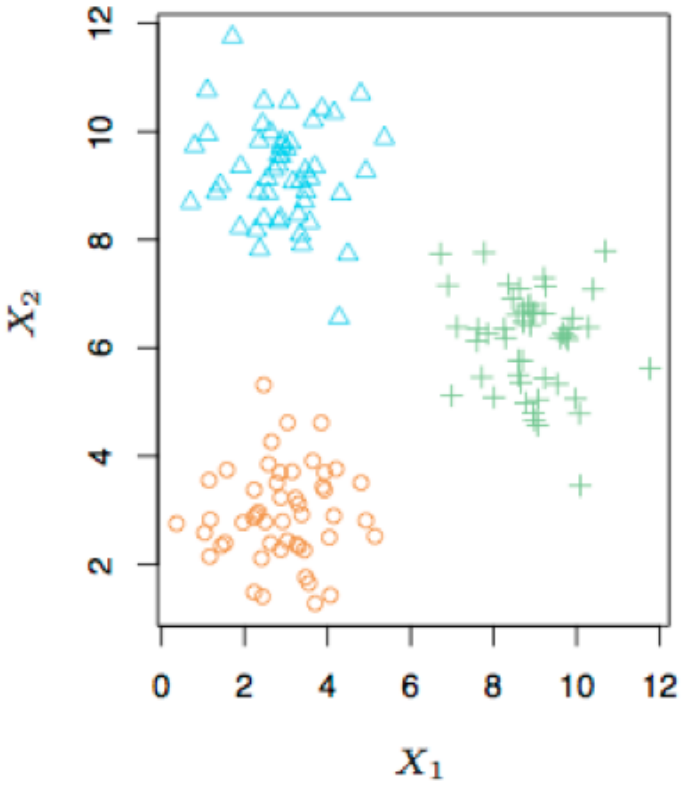
Static/Active/Online Learning

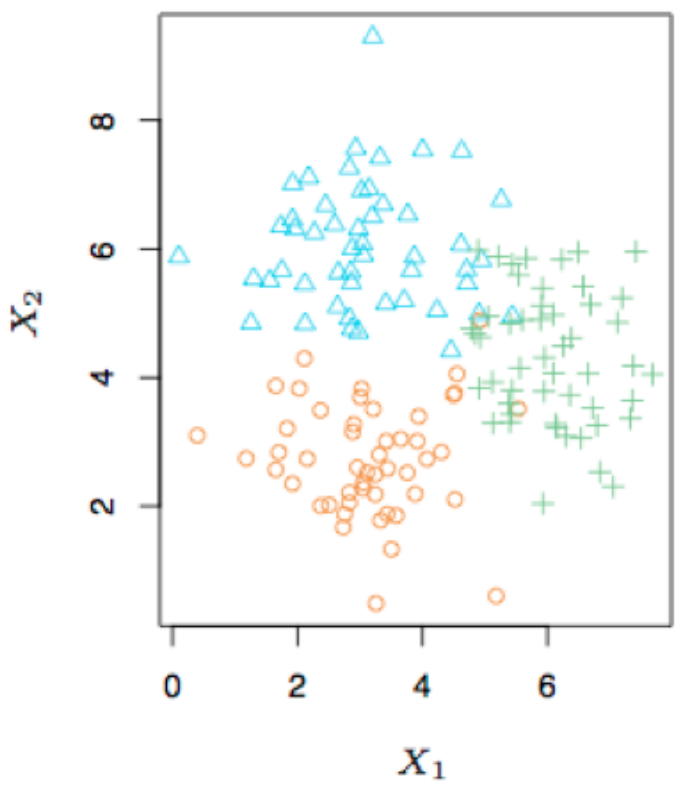
- In terms of how data is acquired
  - Passive (static) learning: a given/static dataset (X, Y)
  - Active learning: can actively choose X first then observe the resulted Y
  - Online learning: data points (X_i, Y_i) are observed sequentially
    - Learning => decision => data => learning => decision => data …
    - Exploration vs. Exploitation

Unsupervised Learning

- There is only X, no Y
- Objective: to represent X in a more compact way (dimension reduction)
  - Categorize samples and group with similar characteristics (Clustering)
  - Get rid of redundant features (PCA)
  - Construct features that explain most variations (PCA)
  - Both are about reducing dimensionality (rows vs columns)
- Useful as a pre-processing step for supervised learning

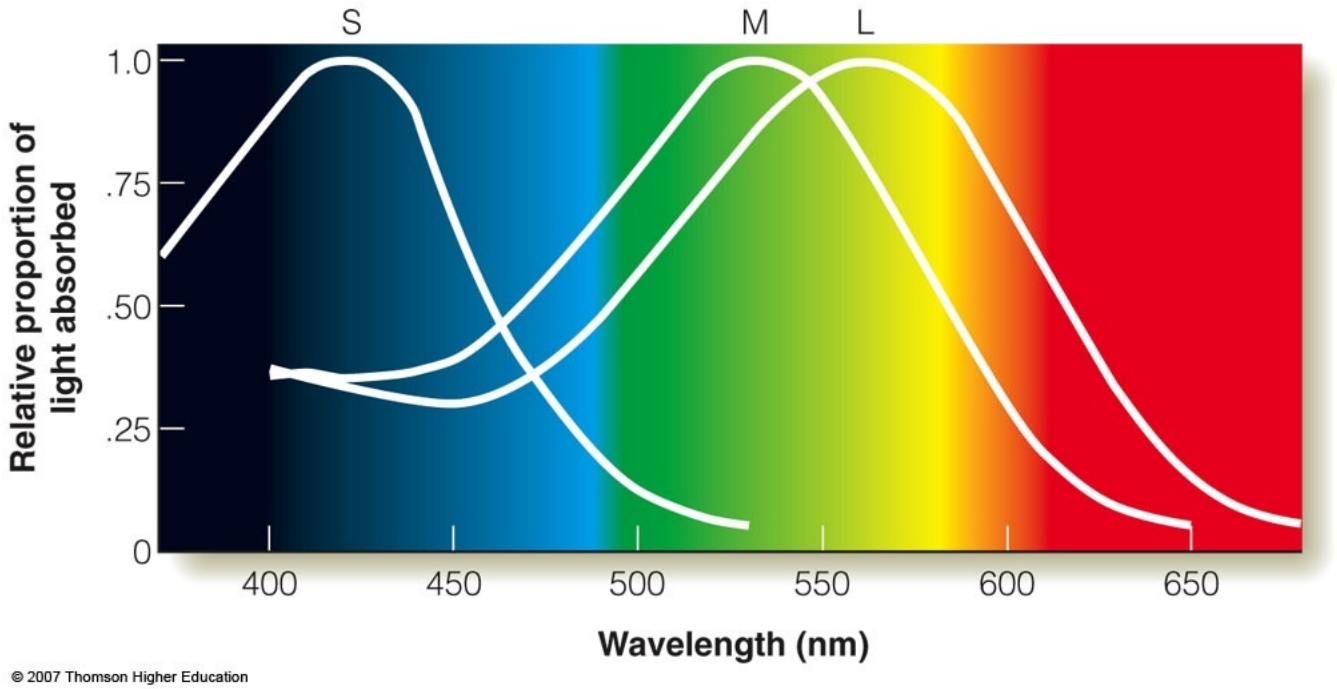
A Simple Clustering Example





Color Categorization

- Is it true that no one could see the color blue until modern times?



Statistical Learning vs. Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence, Computer Science
- Statistical learning arose as a subfield of Statistics

- Machine learning has a greater emphasis on large scale applications and prediction accuracy
- Statistical learning emphasizes structures/models and their interpretability, and precision and uncertainty

- Very vague boundary, much overlap

Philosophy

- Building blocks of successful BA applications
  - Problem, Data, Tools

- We will focus on tools in the course
  - To understand the ideas behind the various techniques
  - To accurately assess the performance of a method
  - To know how and when to use them

  - To get familiar with the standard workflow of BA projects

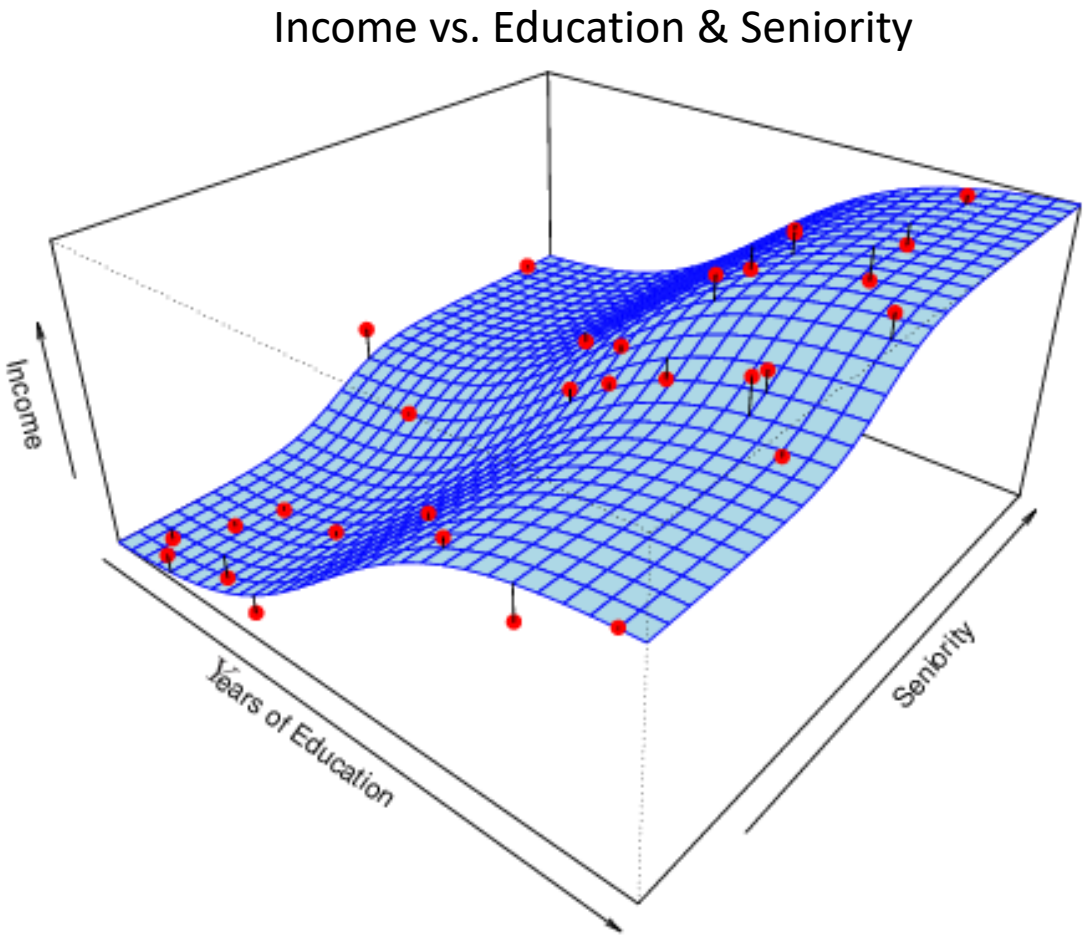
Overview of the tools

- Supervised learning toolbox
  - Regression and generalizations
    - Linear regression
    - Generalized linear models: logistic regression, Poisson regression
    - Regularized regression: ridge regression, Lasso
    - Polynomial regression
  - Tree-based methods
    - Decision tree
    - Bagging, Random Forest, Boosting
  - ~~Support Vector Machine~~
  - ~~Neural Network~~
- Unsupervised learning toolbox
  - K-means clustering, Gaussian Mixture Models, hierarchical clustering
  - Principal Components Analysis

Resampling Methods:
- Cross Validation
- Bootstrap

ASSESSING MODELS

How to estimate *f()*?



$$income = f(Education, Seniority) + \varepsilon$$

Parametric Methods

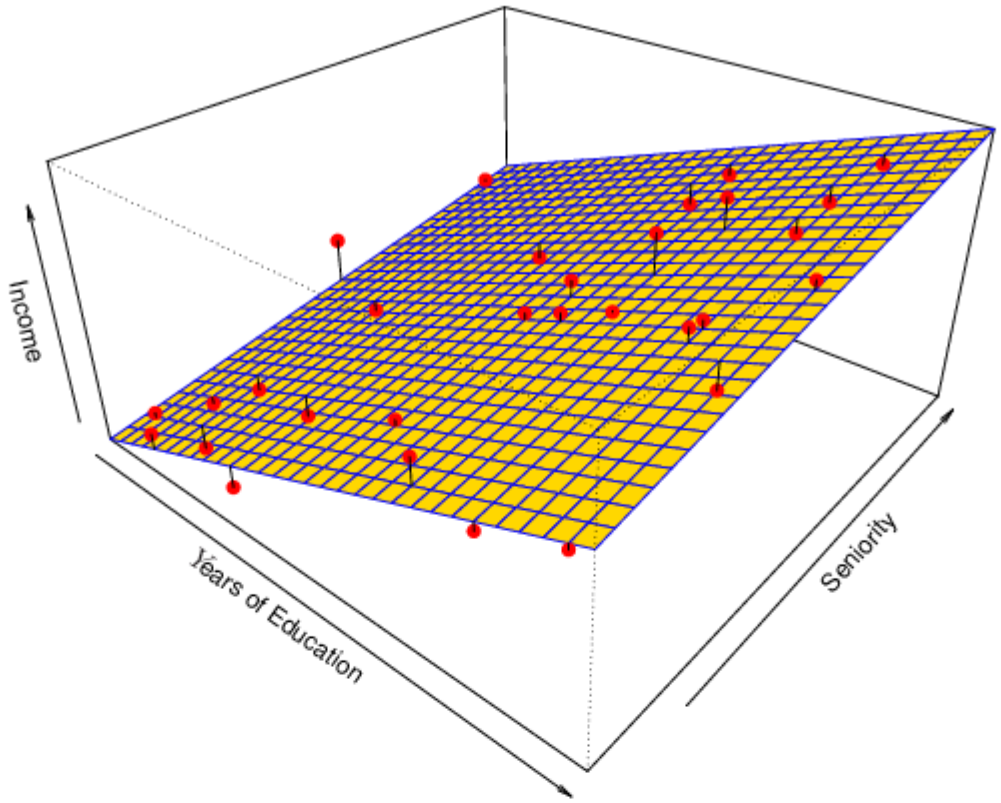
- It reduces the problem of estimating ***f()*** down to one of estimating a set of parameters.

- They involve a two-step model based approach
  1. Make some assumptions about the functional form of ***f()***, i.e., come up with a model. The most common example is a linear model

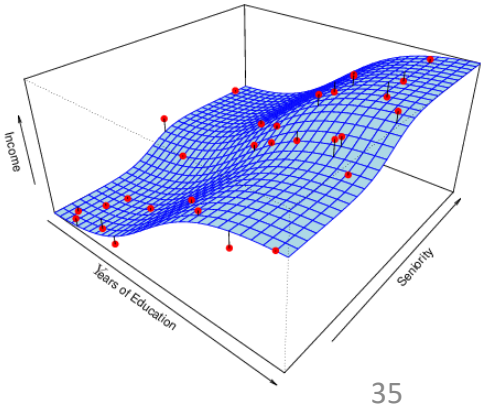
$$f(X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

  2. Use the training data to fit the model, i.e., estimate ***f()*** or equivalently the unknown parameters such as $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

Example: A Linear Regression Estimate



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

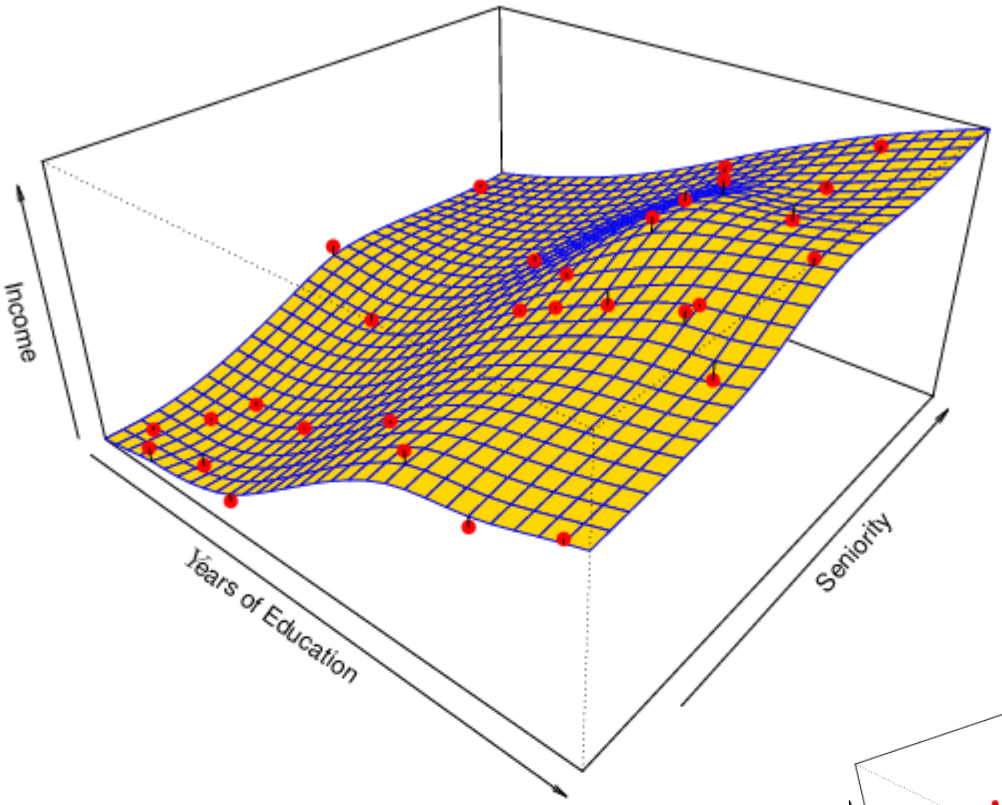


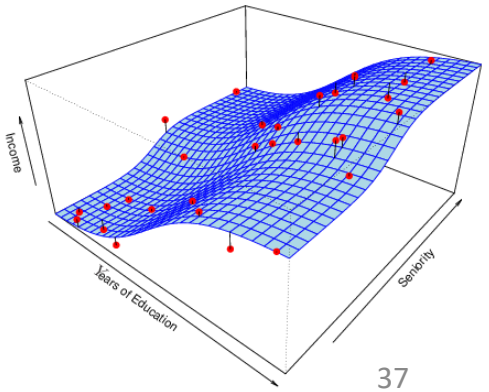
Non-parametric Methods

- They do not make explicit assumptions about the functional form of *f()*.
  - Advantages: They accurately fit a wider range of possible shapes of *f()*.
  - Disadvantages: A very large number of observations is required to obtain an accurate estimate of *f()*.

Example: A Smooth Thin-Plate Spline Estimate

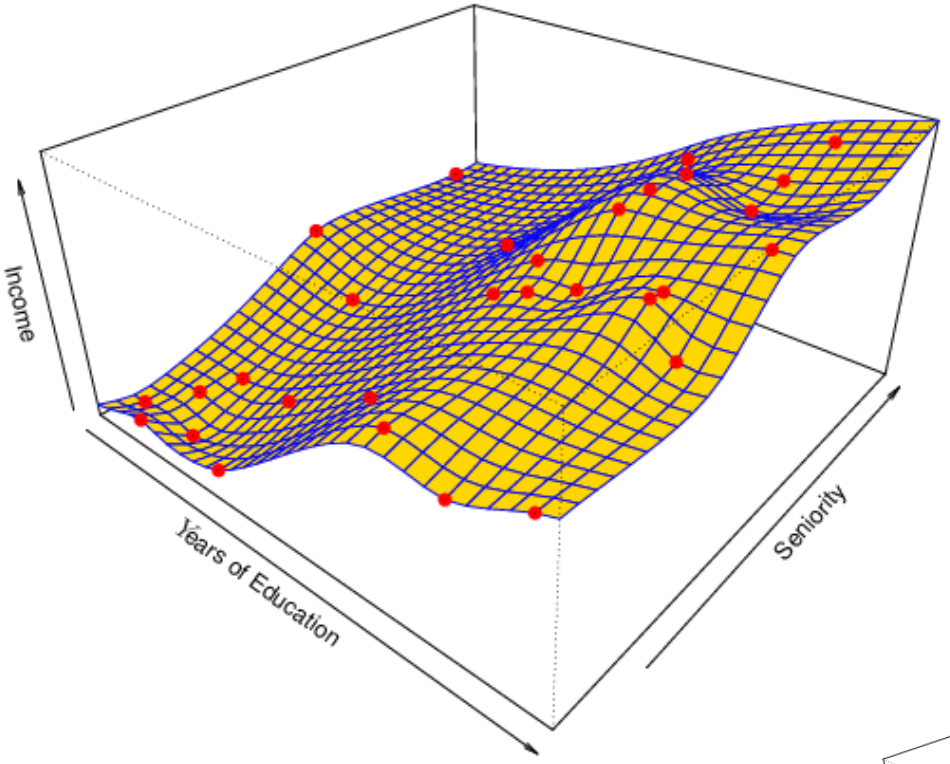
- Non-linear regression methods are more flexible and can potentially provide more accurate estimates.

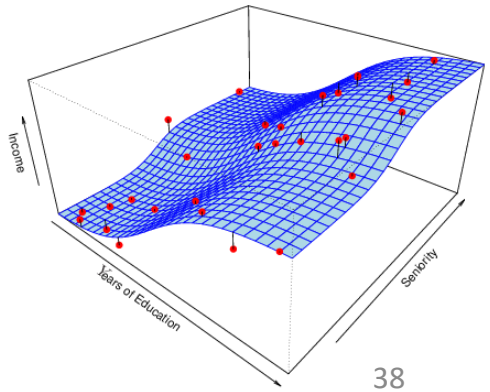




Example: A Rough Thin-Plate Spline Estimate

- Non-linear regression methods can also be too flexible and produce poor estimates for f().

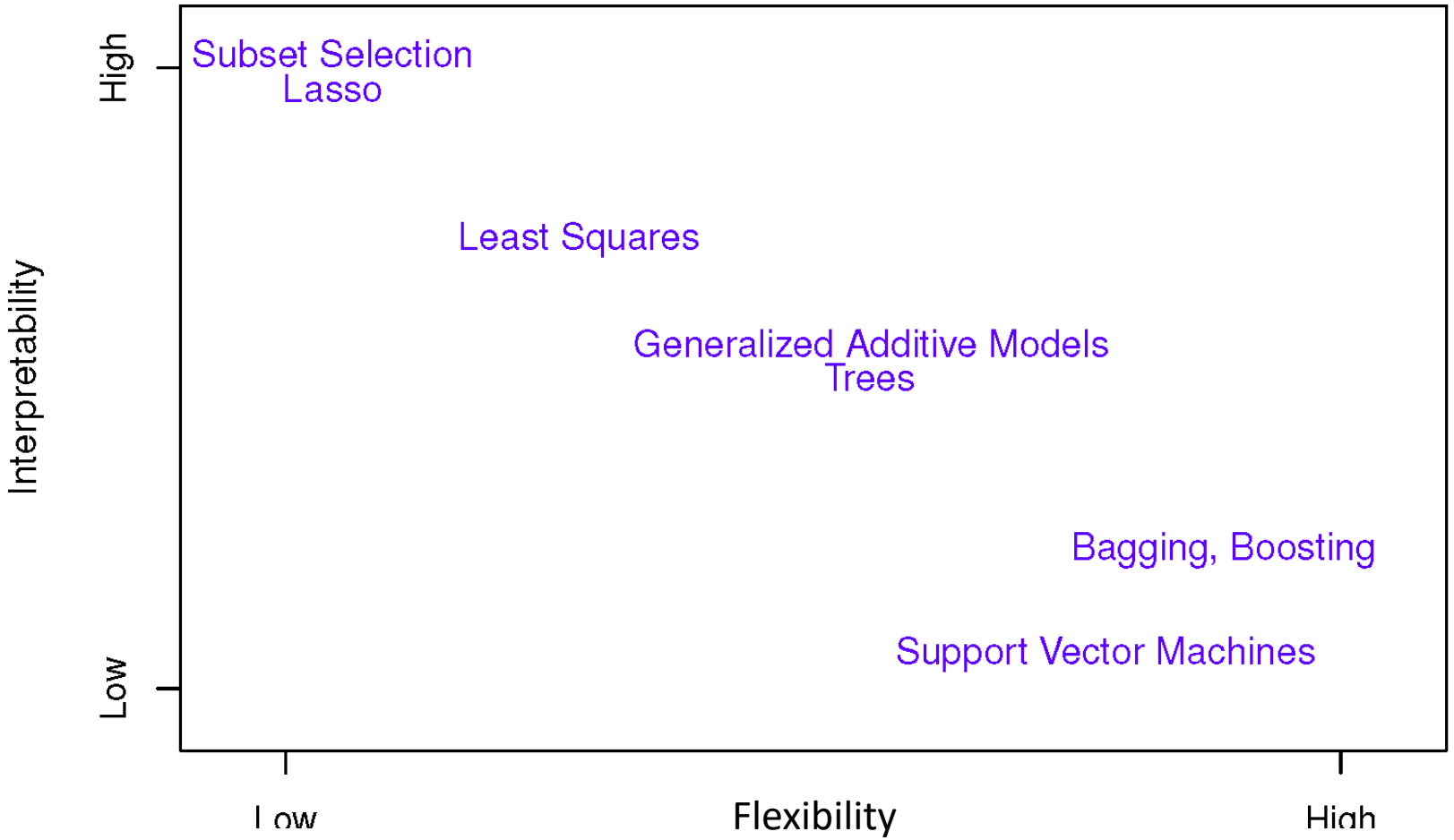




Prediction Accuracy vs. Model Interpretability

- Why not just use a more flexible method if it fits the data better?
  - Interpretability
    - For example, in a linear model, β_j is the average increase in Y for a one unit increase in X_j holding all other variables constant.
  - Parsimony
    - prefer a simpler model involving fewer variables over a black-box predictor involving them all.
  - Risk of over-fitting
    - Better fit in sample may not necessarily translate to better fit out of sample.

Flexibility vs. Interpretability



Assessing Model Accuracy in Regression

- Mean Squared Error (MSE)

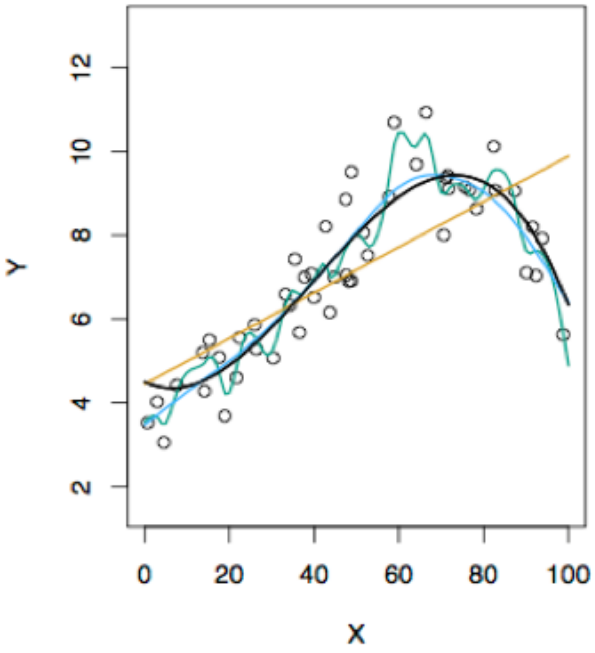
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

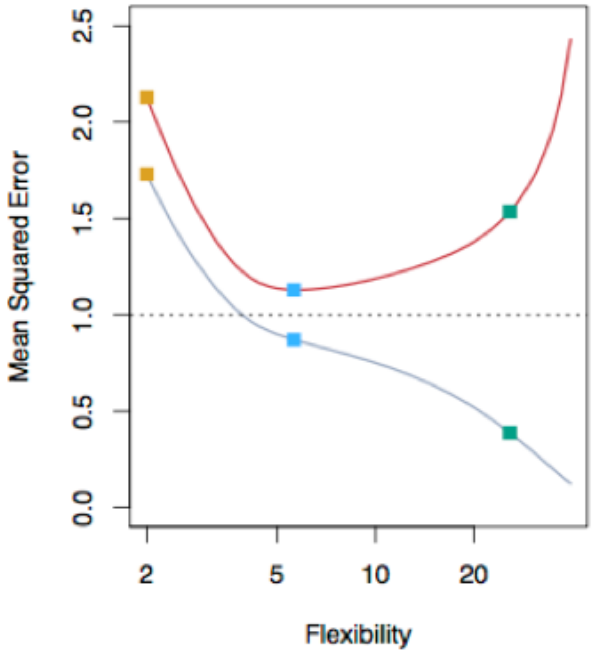
  - Where y_i are the true data and $f(x_i)$ is the prediction

- MSE of the **training** dataset (in sample) can be arbitrarily small, as long as we increase model flexibility

- What is really important is the out-of-sample performance
  - MSE of a **test** dataset

Examples with Different Levels of Flexibility





LEFT

Black: Truth

Orange: Linear Estimate

Blue: smoothing spline

Green: smoothing spline (more flexible)

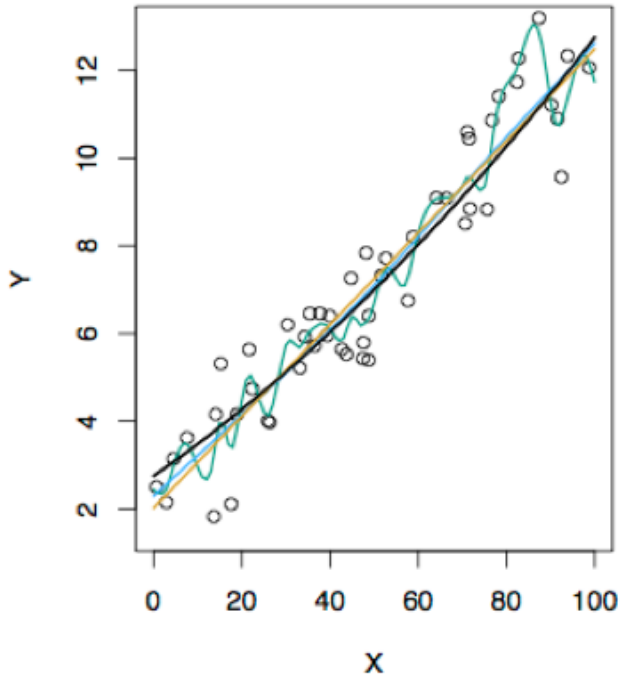
RIGHT

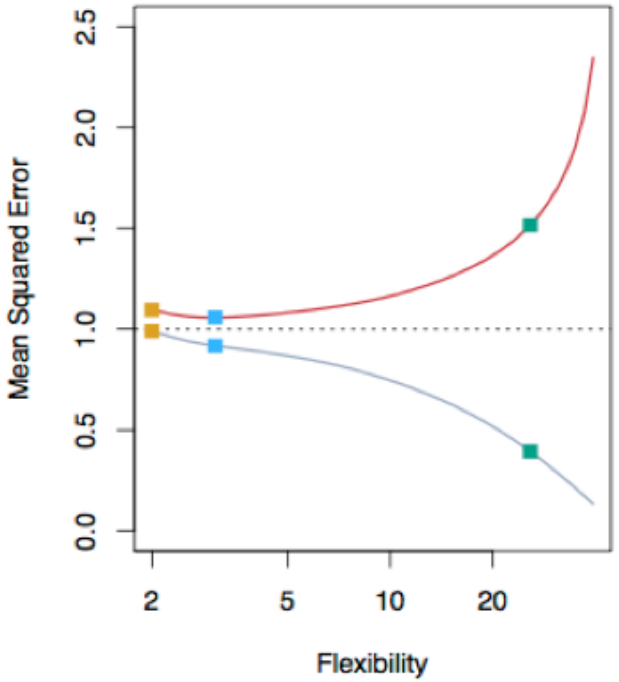
RED: Test MES

Grey: Training MSE

Dashed: Minimum possible test MSE (irreducible error)

Examples with Different Levels of Flexibility





LEFT

Black: Truth

Orange: Linear Estimate

Blue: smoothing spline

Green: smoothing spline (more flexible)

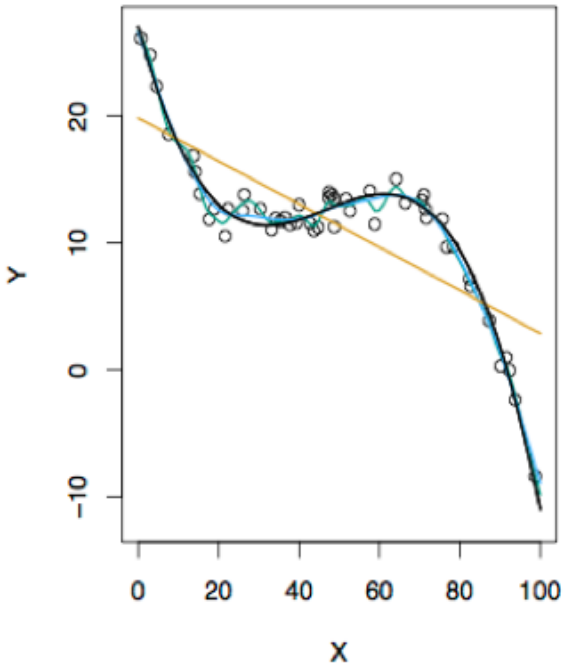
RIGHT

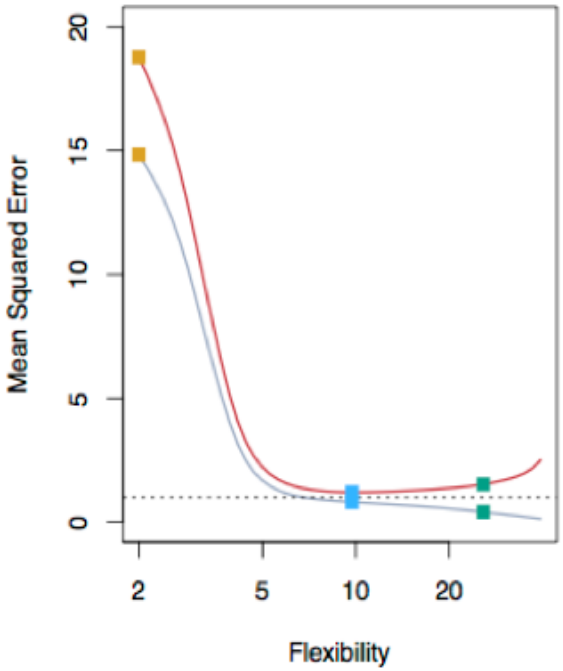
RED: Test MES

Grey: Training MSE

Dashed: Minimum possible test MSE (irreducible error)

Examples with Different Levels of Flexibility





LEFT

Black: Truth

Orange: Linear Estimate

Blue: smoothing spline

Green: smoothing spline (more flexible)

RIGHT

RED: Test MES

Grey: Training MSE

Dashed: Minimum possible test MSE (irreducible error)

The Bias-Variance Tradeoff

- Test versus training MSE illustrates a key tradeoff that governs the choice of statistical learning methods.
  - **Bias**: the error that is introduced by modeling a complicated problem by a simpler model. more flexibility => less bias

$$\text{Bias}[\hat{f}(x)] = \text{E}[\hat{f}(x)] - f(x)$$

  - **Variance**: how much your estimate for ***f()*** would change by if you had a different training data set. more flexibility => more variance

$$\text{Var}[\hat{f}(x)]$$

The Bias-Variance Tradeoff

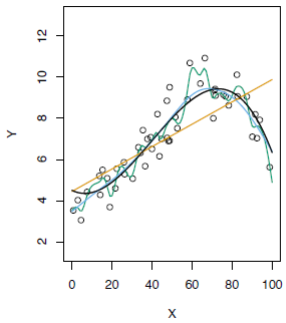
- Consider a test dataset x_t, y_t, the test MSE will be

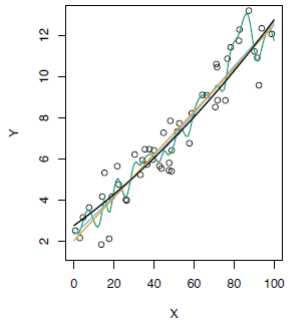
$$\mathrm{E}[(y_t - \hat{f}(x_t))^2] = \mathrm{Bias}^2[\hat{f}(x_t)] + \mathrm{Var}[\hat{f}(x_t)] + \mathrm{Var}[\epsilon]$$

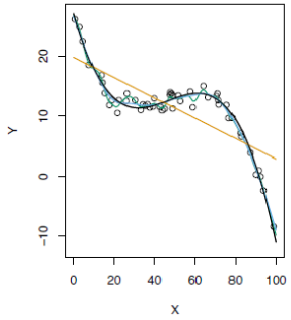
  - decomposed into: bias, variance, and irreducible error

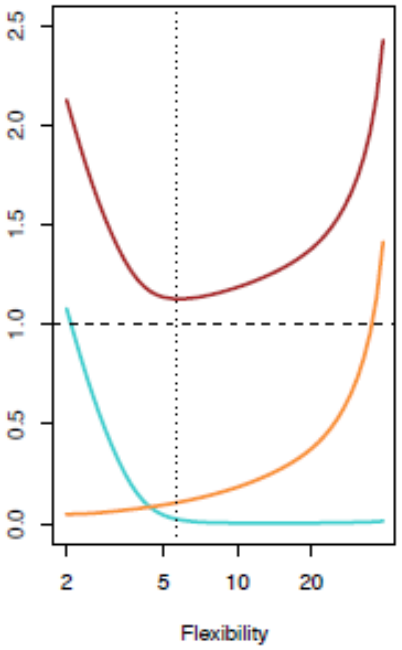
- The two competing forces govern test MSE
  - As a method gets more flexible, the bias will decrease and the variance will increase.
  - But the expected test MSE may go up or down!
  - Choosing the flexibility based on average test error amounts to a bias-variance trade-off.

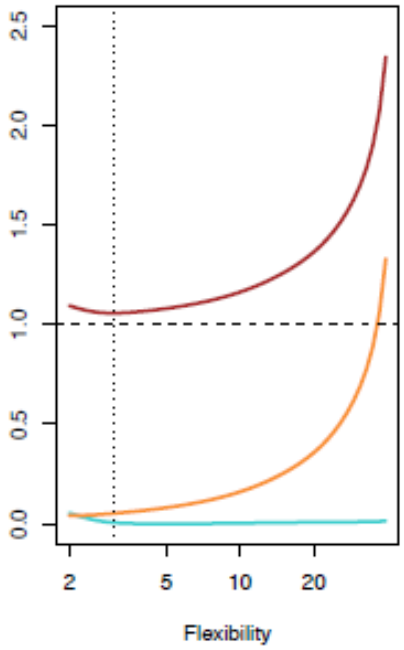
Test MSE, Bias, and Variance

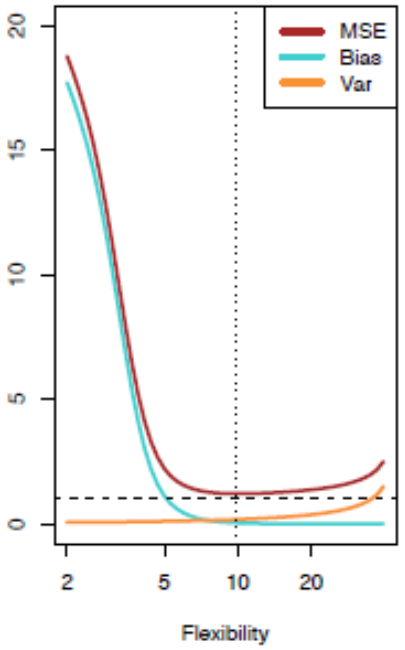












The Classification Setting

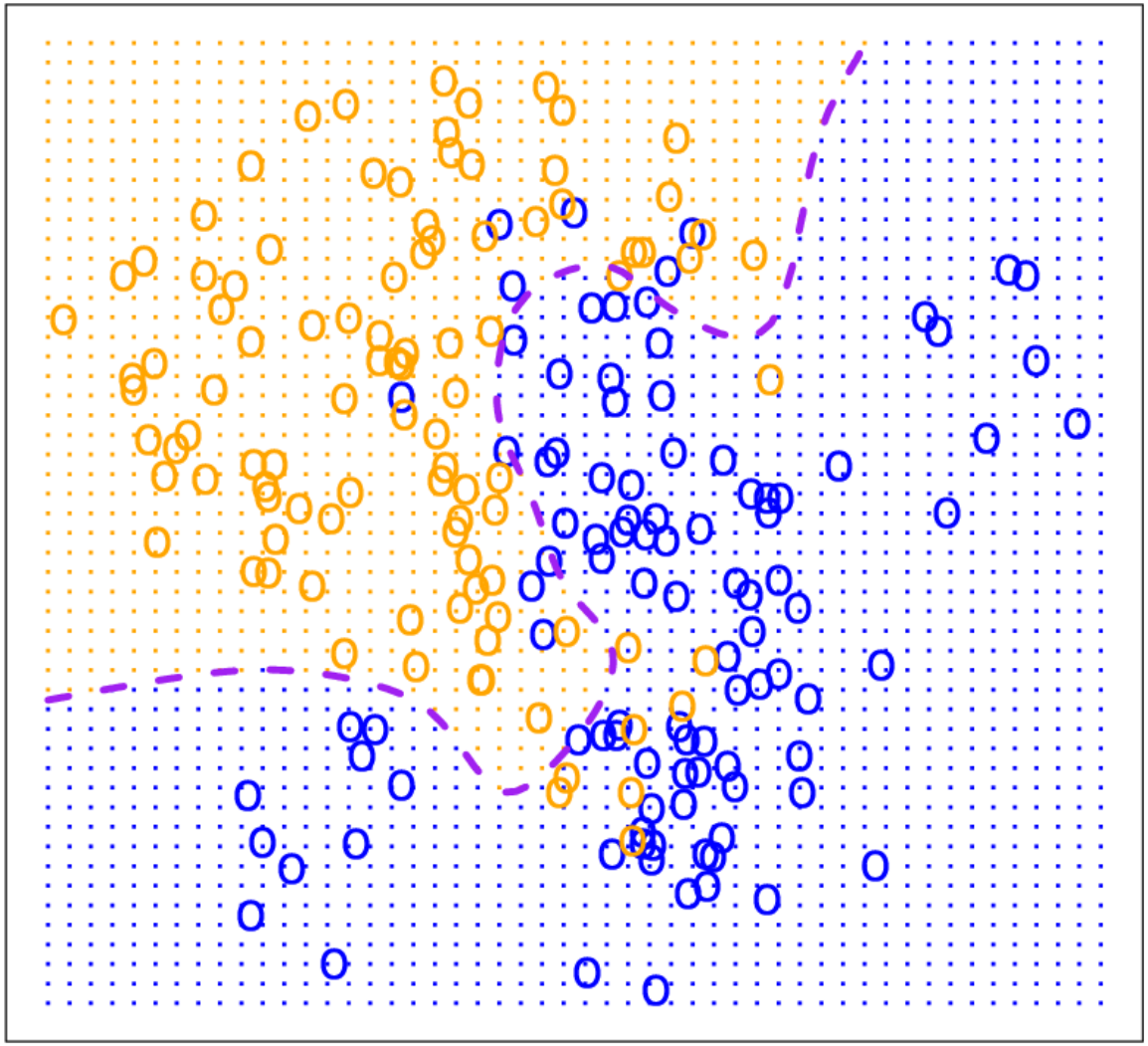
- For a regression problem, we used the MSE to assess the accuracy of the statistical learning method
- For a classification problem we can use the error rate i.e.

$$\text{ErrorRate} = \frac{1}{n}\sum_{i=1}^{n} I\{y_i \neq \hat{y}_i\}$$

- *I()* is an indicator function, which will give 1 if the condition is correct, otherwise it gives a 0.

- Thus the error rate represents the fraction of incorrect classifications, or misclassifications, hence a.k.a. **misclassification rate**.

Bayes Optimal Classifier



A Fundamental Picture

- In general training errors will always decline.
- Test errors will decline at first (as reductions in bias dominate) but will then start to increase again (as increases in variance dominate).
- More flexible/complicated is not always better!