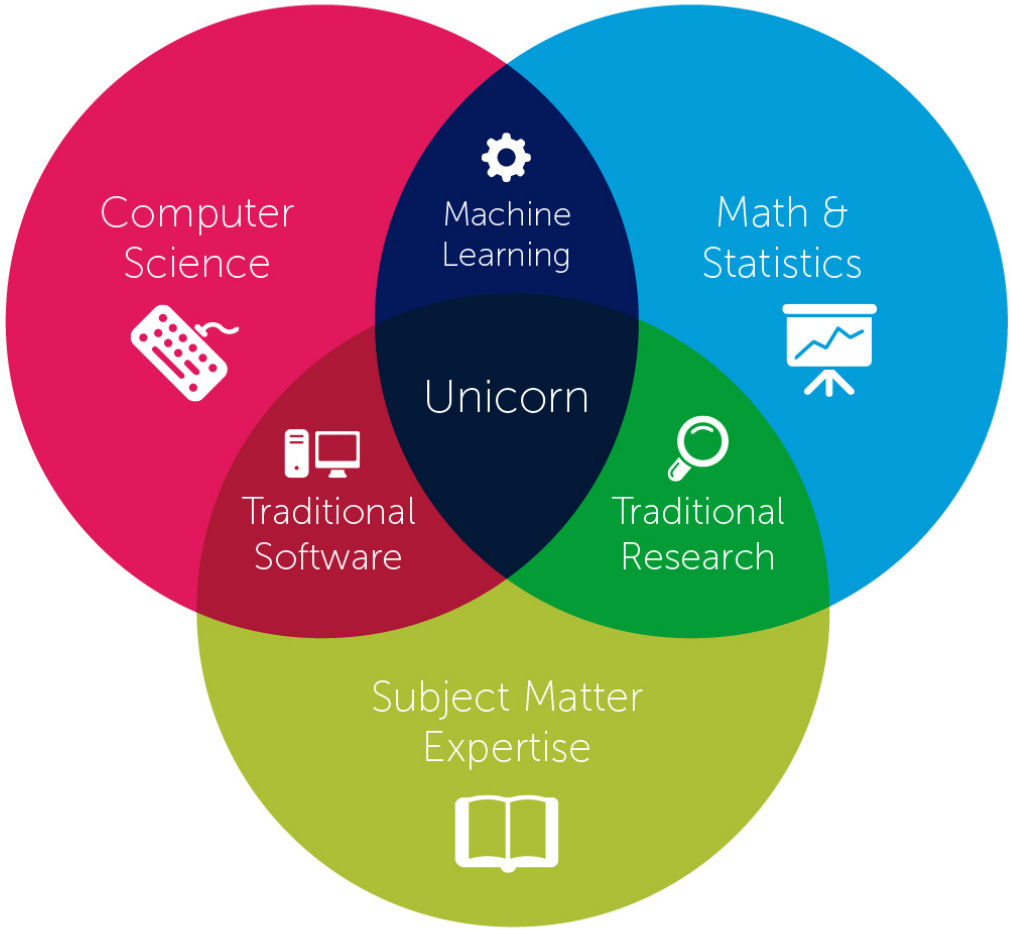
Day 1: Introduction to Statistical Learning

Lucas Leemann

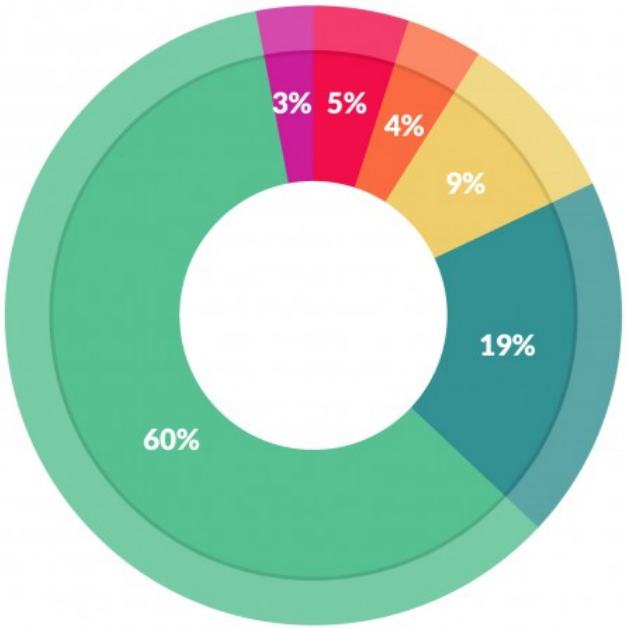
Essex Summer School

Introduction to Statistical Learning

Data Science



Reality



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

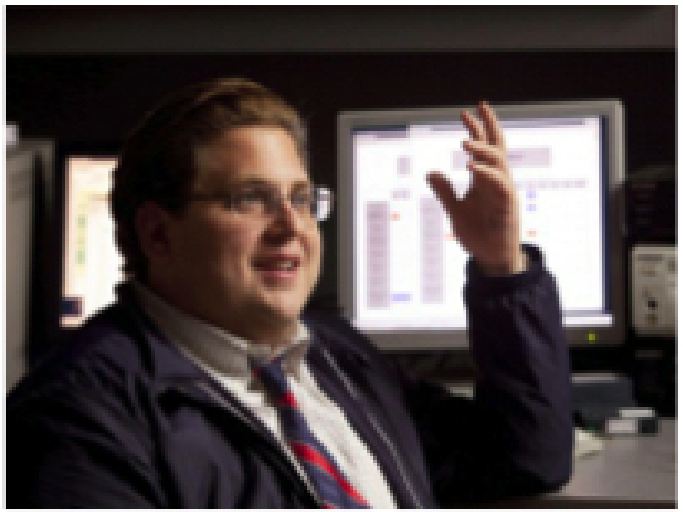
Source: `http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#4a79a76c7f75`

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil



"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?" Hal Varian (Chief Economist at Google, 2009).

Machine Learning Problems

- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten post code.
- Establish the relationship between salary and demographic variables in population based on survey data.
- Identify best model to predict vote choice.

The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy – find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working (simpler methods often perform as well as fancier ones!).
- This is an exciting research area, having important applications in science, industry and policy.
- Statistical learning is a fundamental ingredient in the training of a modern data scientist.

The Netflix prize

- competition started in October 2006. Training data is ratings for 18,000 movies by 400,000 Netflix customers, each rating between 1 and 5.
- training data is very sparse – about 98% missing.
- objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data.
- Netflix's original algorithm achieved a root MSE of 0.953. The first team to achieve a 10% improvement wins one million dollars.
- is this a supervised or unsupervised problem?

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 **leaders.**

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43

Check Ezra Klein's interview with Danah Boyd Link to Podcast

Statistical Learning versus Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap – both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on large scale applications and prediction accuracy.
  - Statistical learning emphasizes models and their interpretability, and precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization".
- Machine learning as a general label.

Statistical Learning vs Quantitative Methods

Quantitative Methods

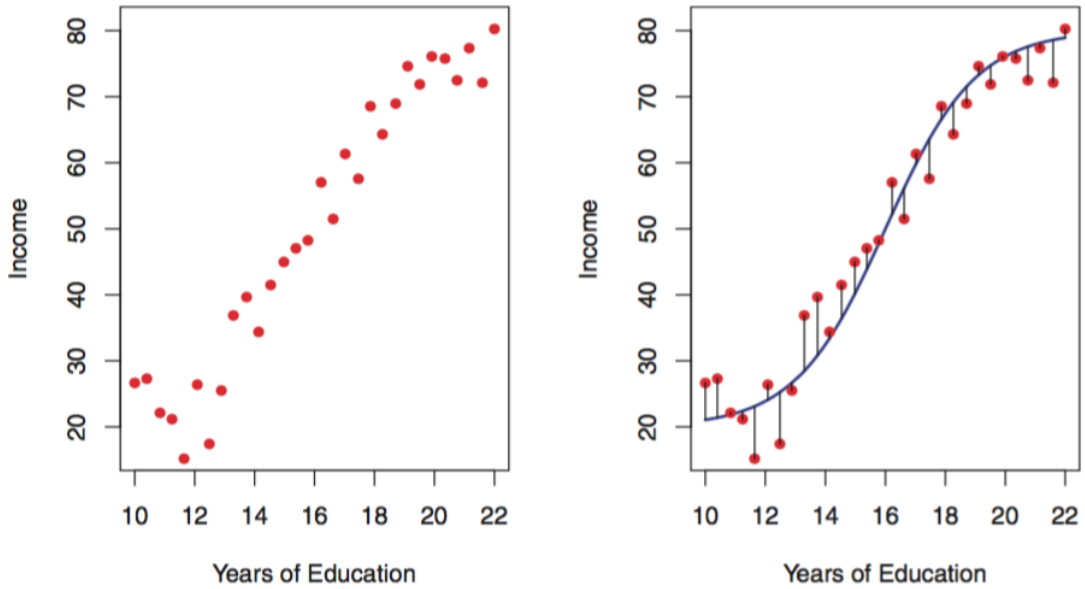
Statistical applications in social sciences with the aim to test theoretically derived hypotheses. The goal is to refute the theoretical implication and thereby show that the theory is wrong.

Statistical Learning (supervised)

Statistical applications in any field of human endeavor with the aim to create an automated/algorithmic prediction procedure. The goal is often to produce as good predictions as possible but sometimes may also be on finding causal factors.

Fundamental Problem

Example



(James et al. 2013: 17)

$$Y = f(X) + \varepsilon$$

$f(X)$

- We use training data to estimate $\hat{f}(X)$.
- This allows us to predict Y when we know X, i.e. $\hat{Y} = \hat{f}(X)$
- The error has two parts, the reducible and the irreducible part:

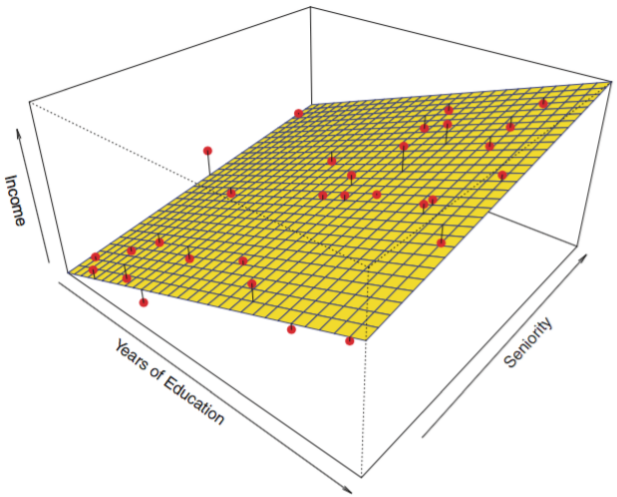
$$\begin{aligned} E[Y - \hat{Y}]^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\textit{reducible}} + \underbrace{Var(\varepsilon)}_{\textit{irreducible}} \end{aligned}$$

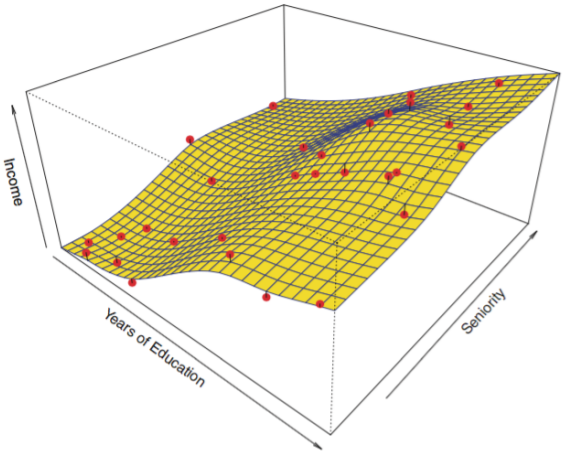
- Irreducible: Because truly random, infinitely many unmodeled causes, treatment heterogeneity
- Various ways to estimate $f(X)$ and we often just rely on simple linear models: $f(X) = \beta_0 + \beta_1 X$

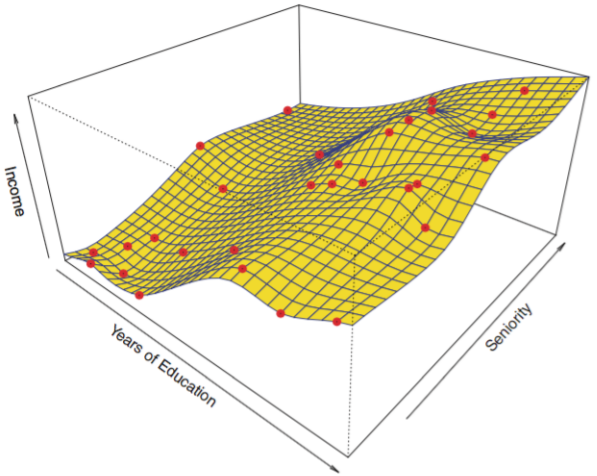
How Do We Estimate $f(X)$?

- We will use training data, $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, to estimate $\hat{f}$, s.t. $Y \approx \hat{f}(X)$.
- Parametric methods:
  1. Functional form assumption, e.g. linear model: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$
  2. Estimation: A way to get at $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$, e.g. ordinary squares.
     - Parametric because we do not estimate $f()$ but rather its components β_0, β_1,..., β_p.
- Non-parametric methods:
  1. No functional form assumptions, but e.g. splines
  2. Very flexible (can be an advantage as well as a disadvantage)
     - Requires usually much more data than parametric approaches.

Example







(James et al. 2013: 22-24)

$\rightarrow$ Trade-off between model accuracy and interpretability

Assessing Model Accuracy

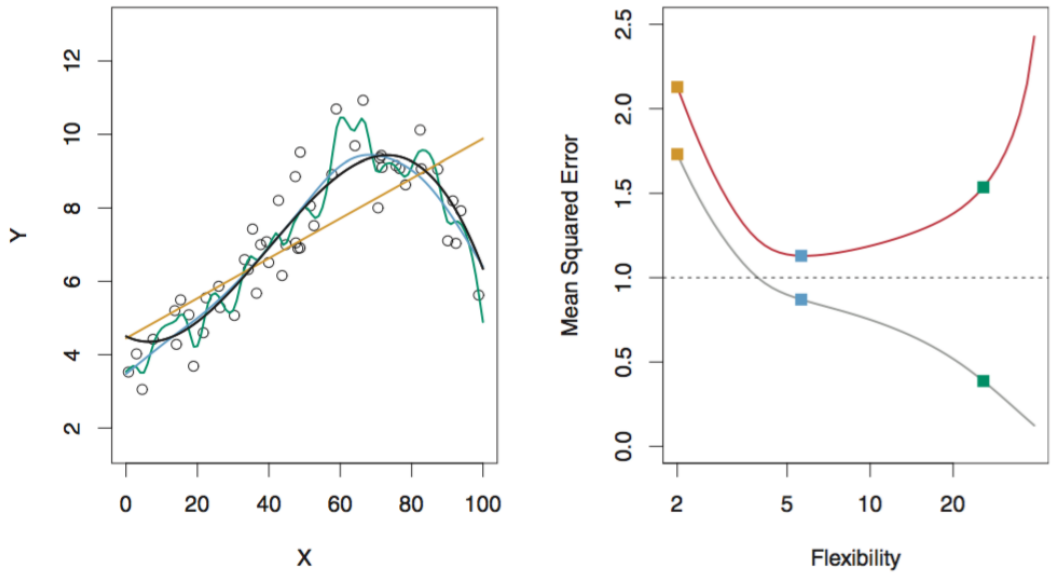
- In order to be able and select the best approach for a *specific* problem, we need to evaluate performance.
- For prediction problems (continuous outcomes) we can look at the *mean squared error*:

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

- We determine $\hat{f}(x)$ on the training dataset and then generate MSE based on the test data.

Variance-Bias Tradeoff 1

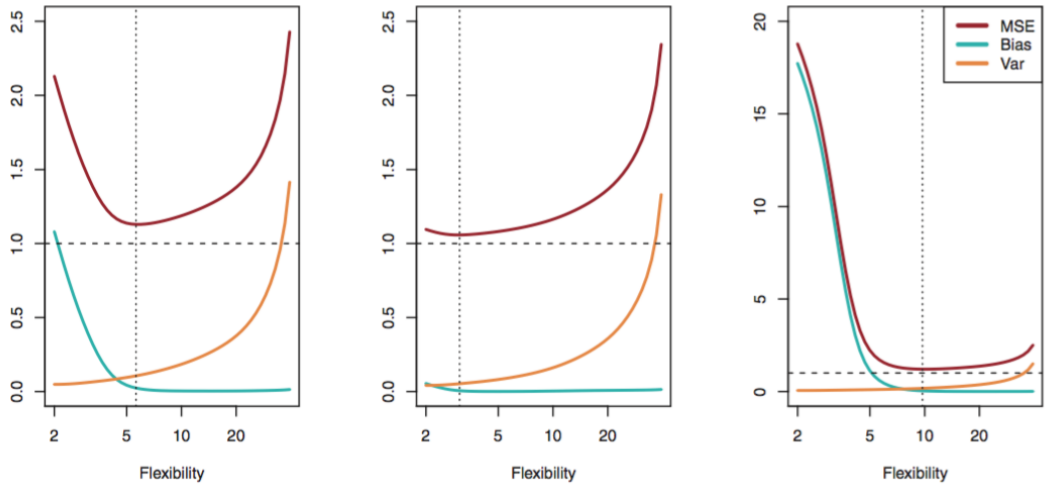
- If we chose models only based on training MSE, we end up with bad predictions.
- The problem is known as *over-fitting*:



(James et al. 2013: 22-24)

Variance-Bias Tradeoff 2

- test MSE $= \text{Var}(\hat{f}(X)) + [\text{Bias}(\hat{f}(X))]^2 + \text{Var}(\varepsilon)$
- The V-B tradeoff exists because there are two opposite principles at work:
  - Bias: As the model becomes less complex, the bias increases.
  - Variance: As the model becomes more complex, the variance increases.



(James et al. 2013: 36)

Classification Problem

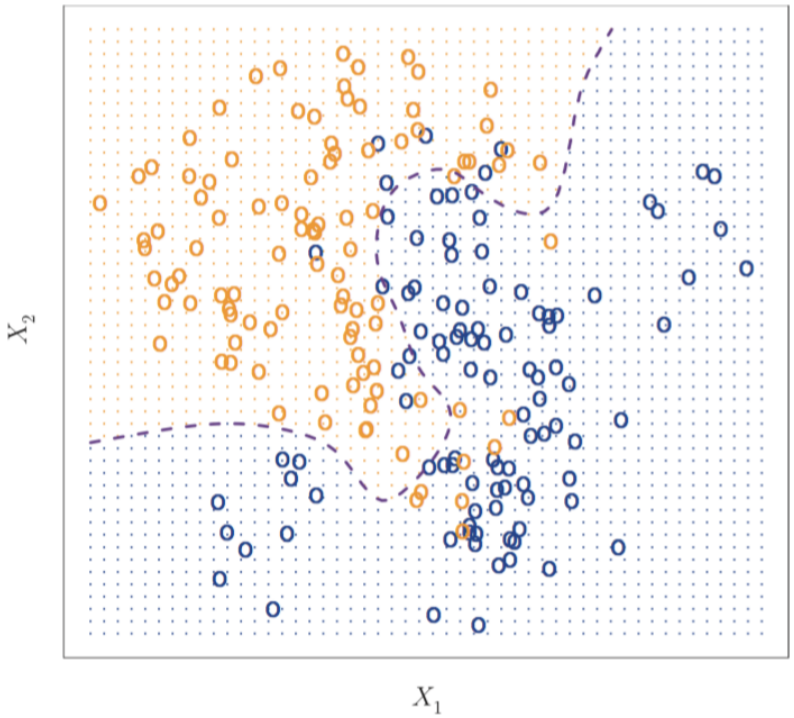
Classification

- When Y is not continuous but qualitative, we have a classification problem.
- The goal is to predict the correct *class* of an observations based on its X.
- We assess the quality of classification via the *error rate*:

$$\text{Error rate} = \frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \hat{y}_i\right)$$

- We prefer the classification that minimizes the error rate in the test data.

Classification: Naive Bayes

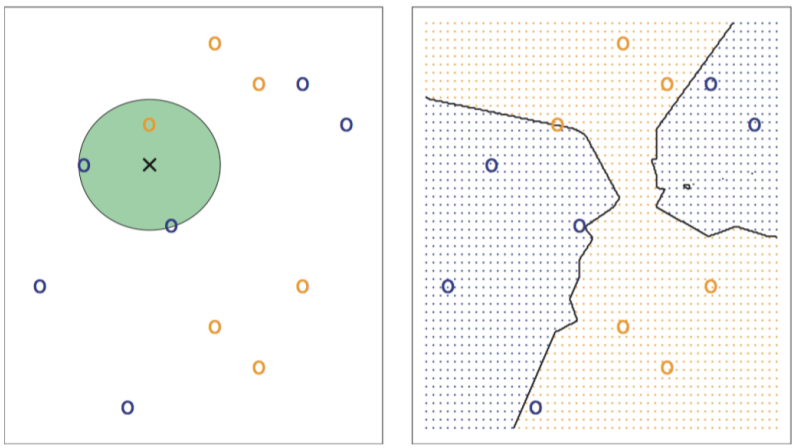


(James et al. 2013: 38)

Classification: K Nearest Neighbor

- Alternative: We look at the K nearest neighbors (based on x_0) and base our classification on them.
- We assign the class for which this quantity is largest:

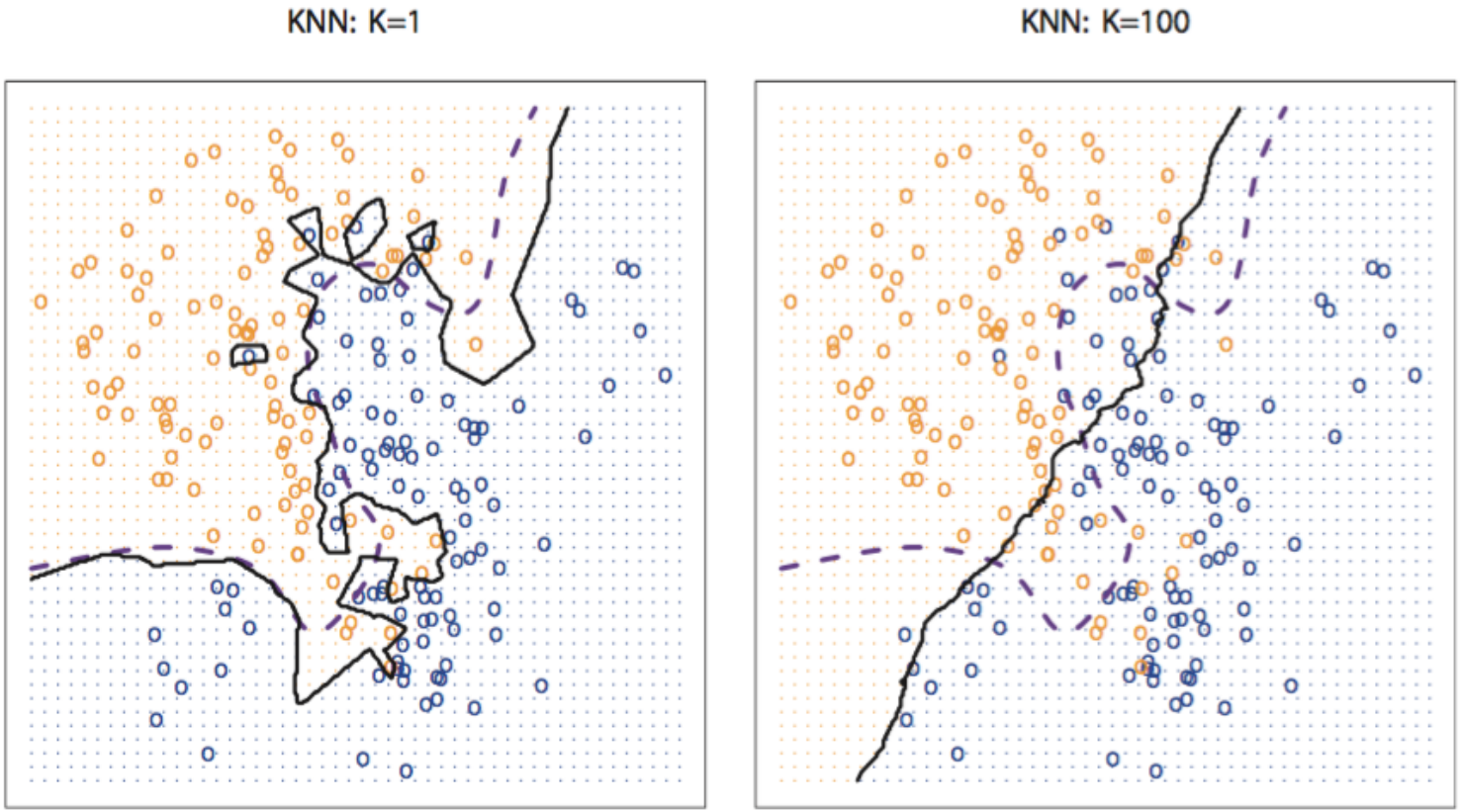
$$P(Y = j|X = x_0) \quad = \quad \frac{1}{K} \sum_{i \in \mathcal{N}_0} I\left(y_i \in j\right)$$



(James et al. 2013: 40)

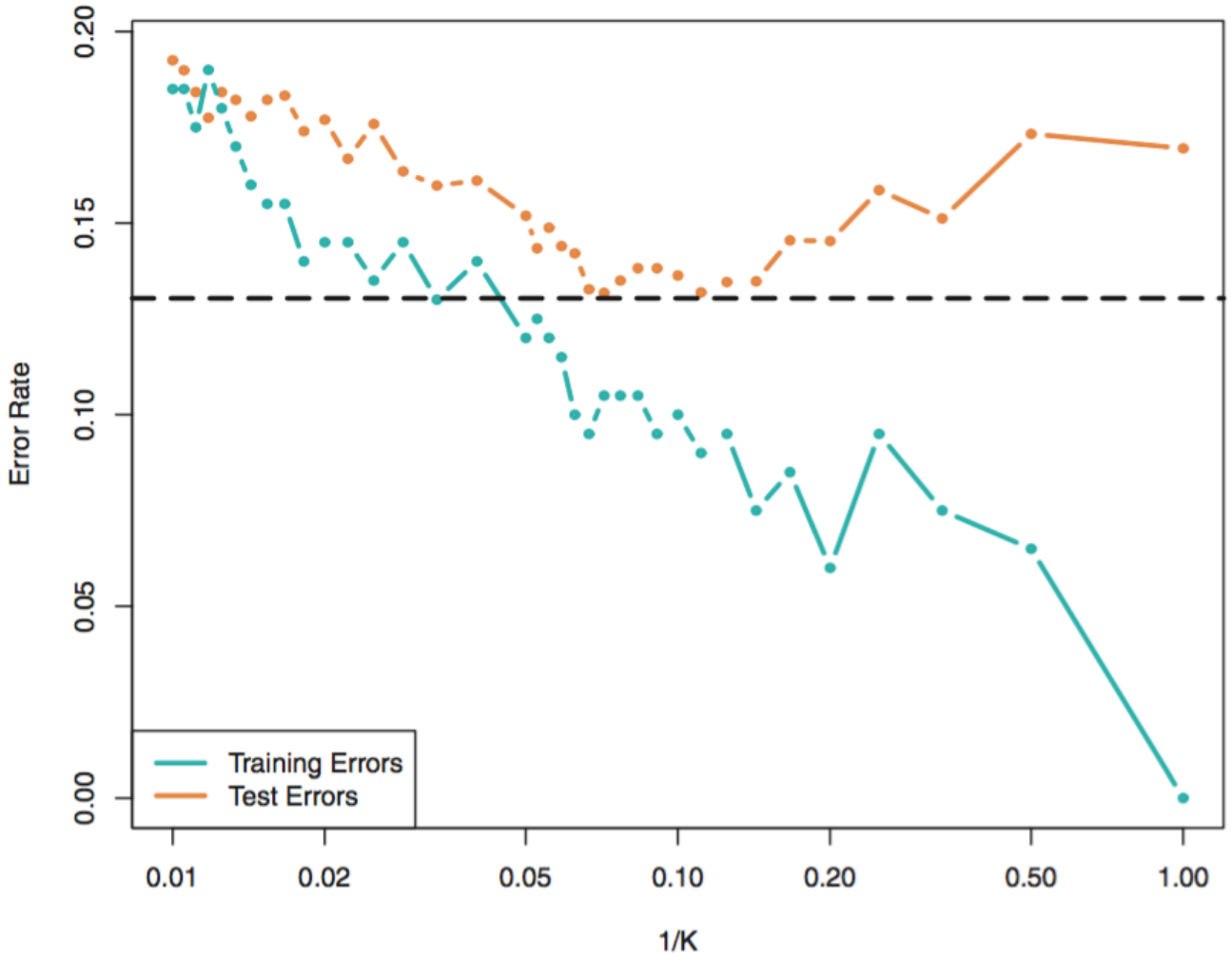
Classification: K Nearest Neighbor 2

The choice of K matters:



(James et al. 2013: 41)

KNN and the V-B tradeoff



(James et al. 2013: 42)

Lab

- Introduction to RStudio
- Rstudio computer game...`library(BetaBit)`
- All labs: `philippbroniecki.github.io/ML2017.io/`