Predictive Modeling

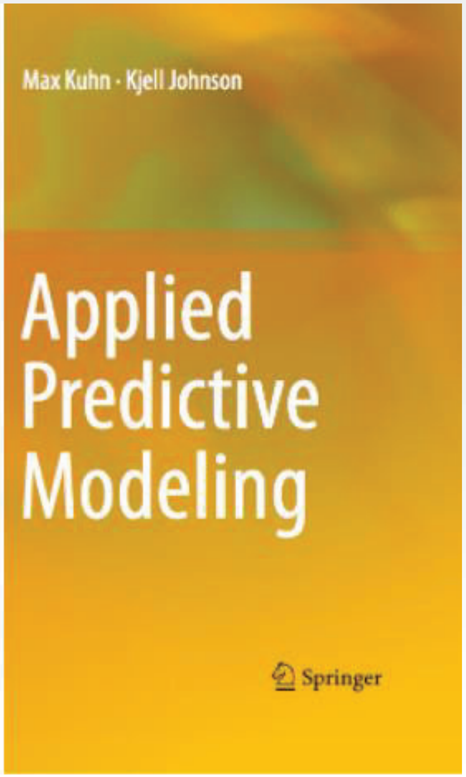
Chapter 1: Introduction to Predictive Modeling

STA 6543

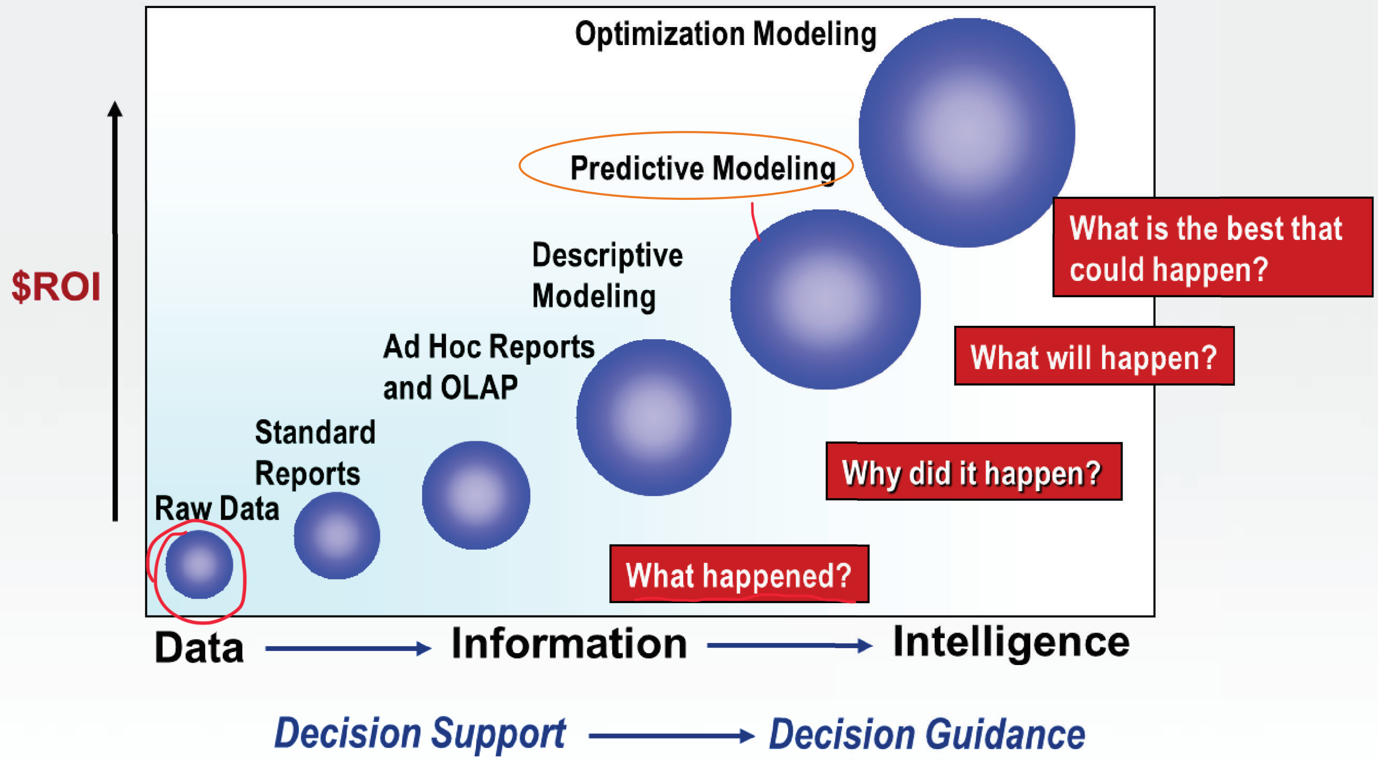
The University of Texas at San Antonio

Textbook information

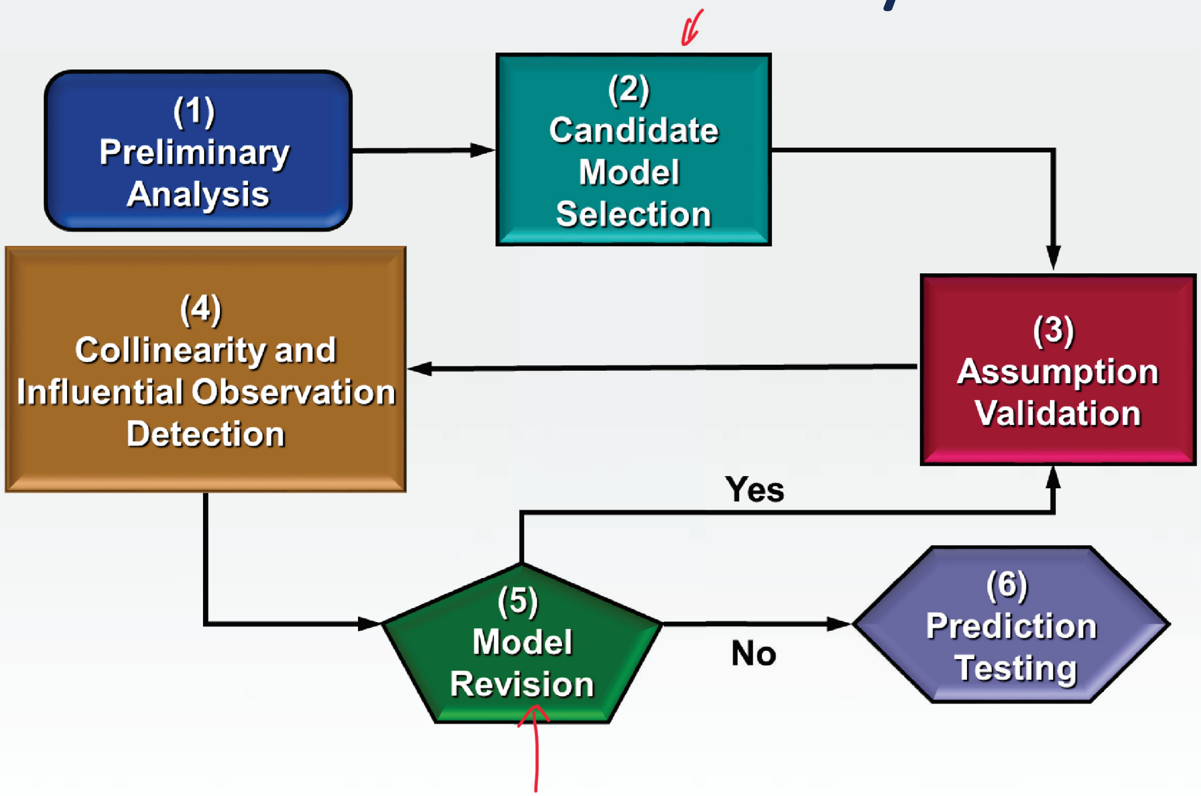
- Applied Predictive Modeling
  http://appliedpredictivemodeling.com/
- The AppliedPredictiveModeling R package contains many of the data sets used here and R scripts to reproduce the analyses in each chapter of the book.



Overview of statistical data analytics



Overview of statistical data analytics



What is predictive modeling?

- What Is predictive modeling?
  - Definition of predictive modeling
  - The trade-off between prediction accuracy and model interpretability
  - Supervised vs. unsupervised learning
  - Regression vs. classification problems
- Introduction to R (Lab 1)

What is predictive modeling?

Chapter 1

What is predictive modeling?

- Predictive modeling: the process of developing a *mathematical tool or model* that generates an *accurate* prediction.

Examples

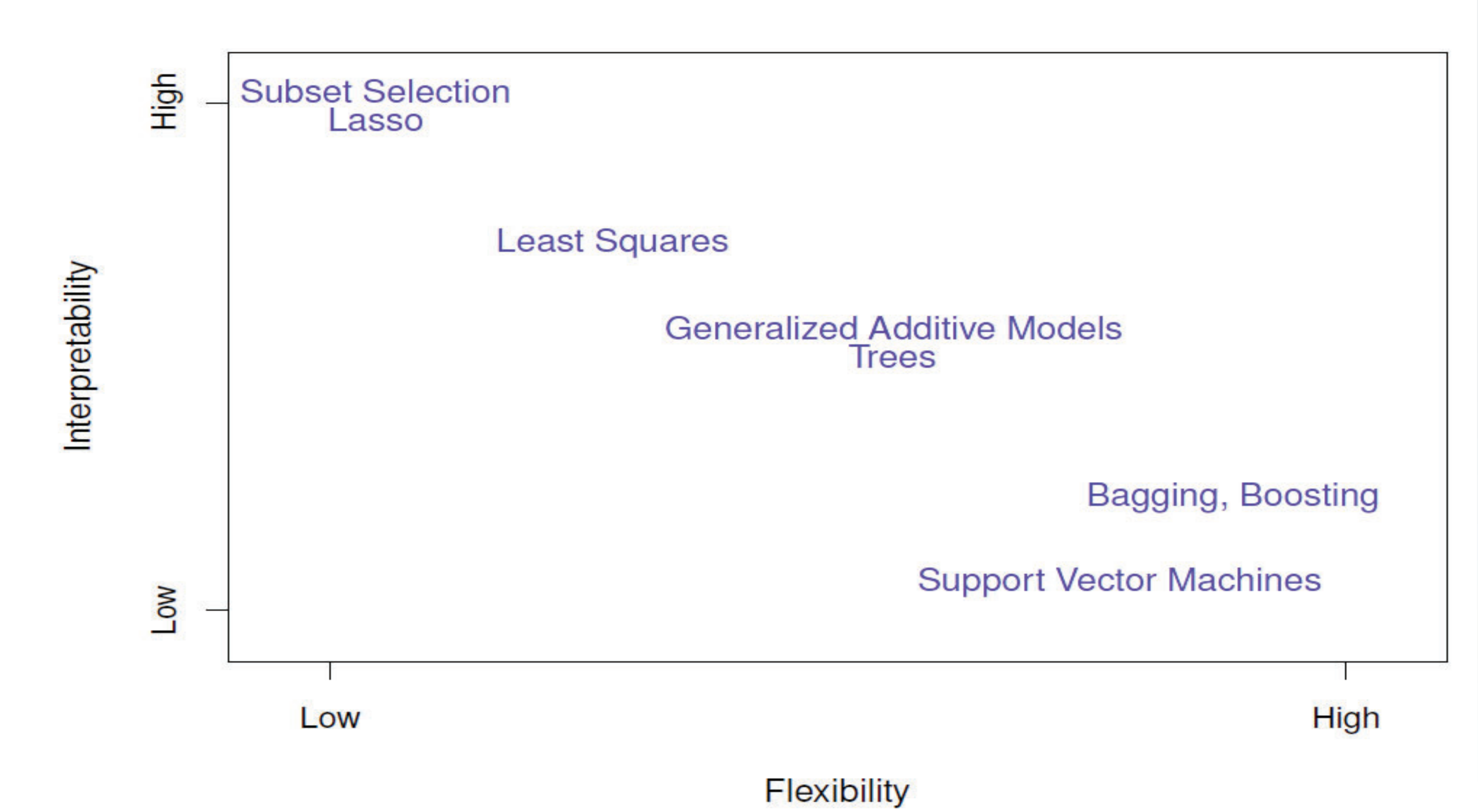
- How many copies will this book sell? (regression)
- How much will my house sell for in the current market? (regression)
- Will this customer move their business to a different company?
- Does a patient have a specific disease?
- Should I sell this stock?
- Is an e-mail spam?
- Will this patient respond to this therapy?

classification

Some tradeoffs

- Prediction accuracy versus interpretability.
  - Linear models are easy to interpret; non-parametric models are not.
- Good fit versus over-fit or under-fit.
  - How do we know when the fit is just right?
- Parsimony versus black-box.
  - We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

Tradeoff between flexibility and interpretability



Terminology

training data

- The term *sample* often refers to a subset of data points, such as the training set sample.
- The *training set* consists of the data used to develop models while the *test* or *validation* sets are used solely for evaluating the performance of a final set of candidate models.

features

- The *predictors (X)*, *independent variables*, *attributes*, or *descriptors* are the data used as input for the prediction equation.
- *Outcome*, *dependent variable*, *target*, *class*, or *response (Y)* refer to the outcome event or quantity that is being predicted.

Terminology

- *Continuous* data have natural, numeric scales.
- *Categorical* data, otherwise known as *nominal*, *attribute*, or *discrete* data, take on specific values that have no scale
- *Model building*, *model training*, and *parameter estimation* all refer to the process of using data to determine values of model equations.

1. Prediction

- Consider

unknown

$$Y = f(X) + \epsilon,$$

- If we can produce a good estimate for *f* (and the variance of *ε* is not too large) we can make accurate predictions for the response, Y, based on a new value of X.

We may consider a complex relationship for 'f'

'black box'

Example: direct mailing prediction

- Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
- Don’t care too much about each individual characteristic.
- Just want to know: For a given individual should I send out a mailing?

2. Inference

- Alternatively, we may also be interested in the type of relationship between Y and the X's.
- For example,
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?

Example: housing inference

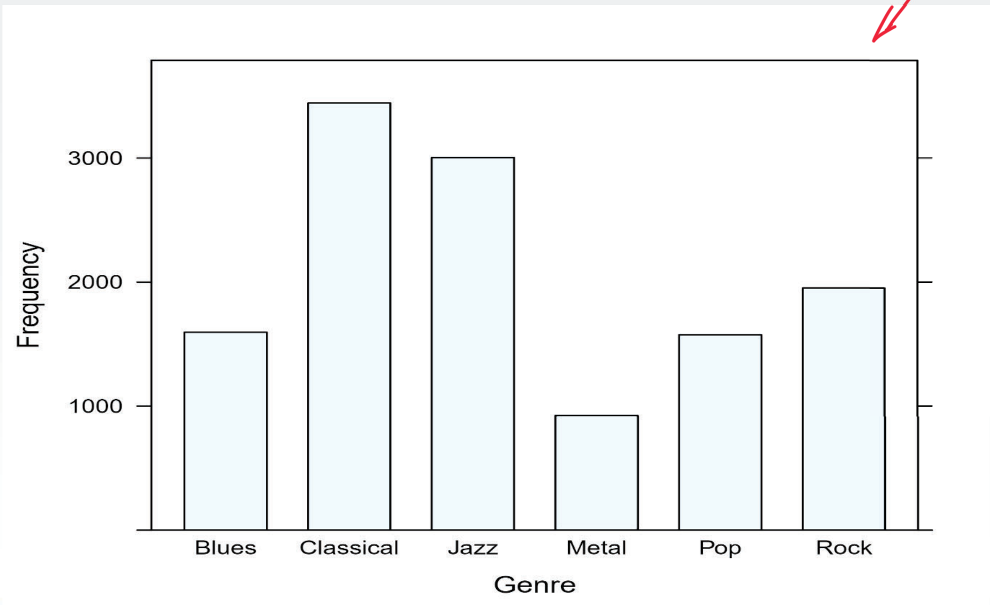
- Wish to predict median house price based on 14 variables.
- Probably want to understand which factors have the biggest effect on the response and how big the effect is.
- For example how much impact does a river view have on the house value etc.

The objectives of this course

- Foundational principles for *building predictive models*
- Intuitive explanations of many commonly used predictive modeling methods for both classification and regression problems
- Principles and steps for validating a predictive model
- Computer code to perform the necessary foundational work to build and validate predictive models

Example: music genre

- The objective was to use the predictors to *classify* music samples into the appropriate music genre.

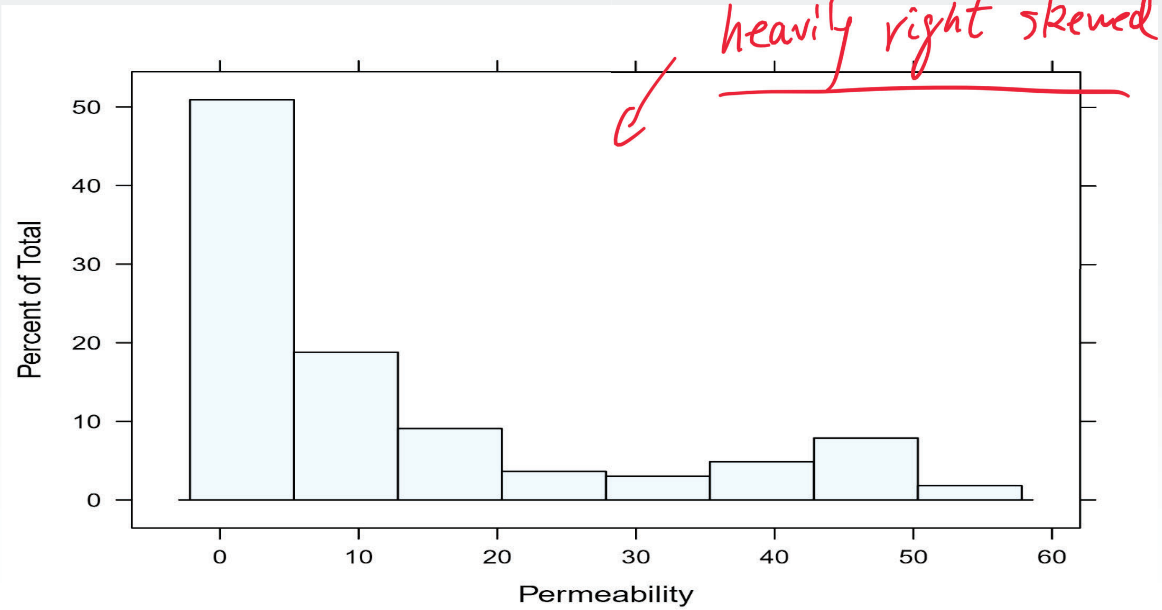


unbalanced issue for the response variable.

Example: permeability

- The objective was to use the predictors to *model* compounds' permeability

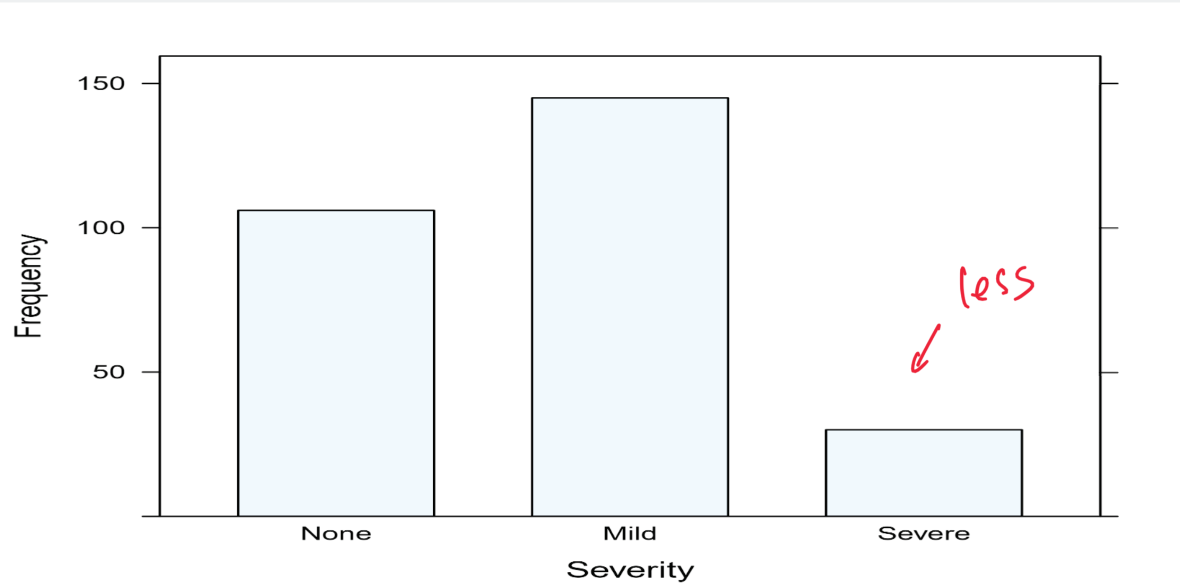
heavily right skewed (continous)



Box-cox

Example: hepatic injury

- The objective was to build a *predictive model* for hepatic injury so that other compounds can be screened for the likelihood of causing hepatic injury.



less

unbalanced ?

Table 1.1: A comparison of several characteristics of the example data sets

$\hat{\beta} = (X'X)^{-1}X'y$

Data characteristic	Music genre	Grant applications	Hepatic injury	Fraud detection	Permeability	Chemical manufacturing
	Data set					
Dimensions	$n > p$				$n < p$	
# Samples	12,495	8,707	281	204	165	177
# Predictors	191	249	376	20	1,107	57
					Lasso	
Response characteristics						
Categorical or continuous	Categorical	Categorical	Categorical	Categorical	Continuous	Continuous
Balanced/symmetric		×		×		×
Unbalanced/skewed	×		×	×		
Independent			×		×	
Predictor Characteristics						
Continuous	×	×	×	×		×
Count	×	×	×			×
Categorical		×	×	×	×	×
Correlated/associated	×	×	×	×	×	×
Different scales	×	×	×	×		×
Missing values		×				×
Sparse					×	

$n < p$

$X'X$ is not invertible

Supervised vs. unsupervised learning

- We can divide all learning problems into *Supervised and Unsupervised* situations
- Supervised learning:
  - Supervised Learning is where both the predictors X and the response Y are observed.
  - We wish to accurately predict unseen test cases, understand which inputs affect the outcome and how, and assess the quality of our prediction and inferences.
  - Most of this course deal with supervised learning.
- Unsupervised learning:
  - Only the predictors X are observed.
  - Use the predictors to groups of samples that behave similarly.
  - Difficult to know how well you are doing, but can be useful as a pre-processing step for supervised learning.

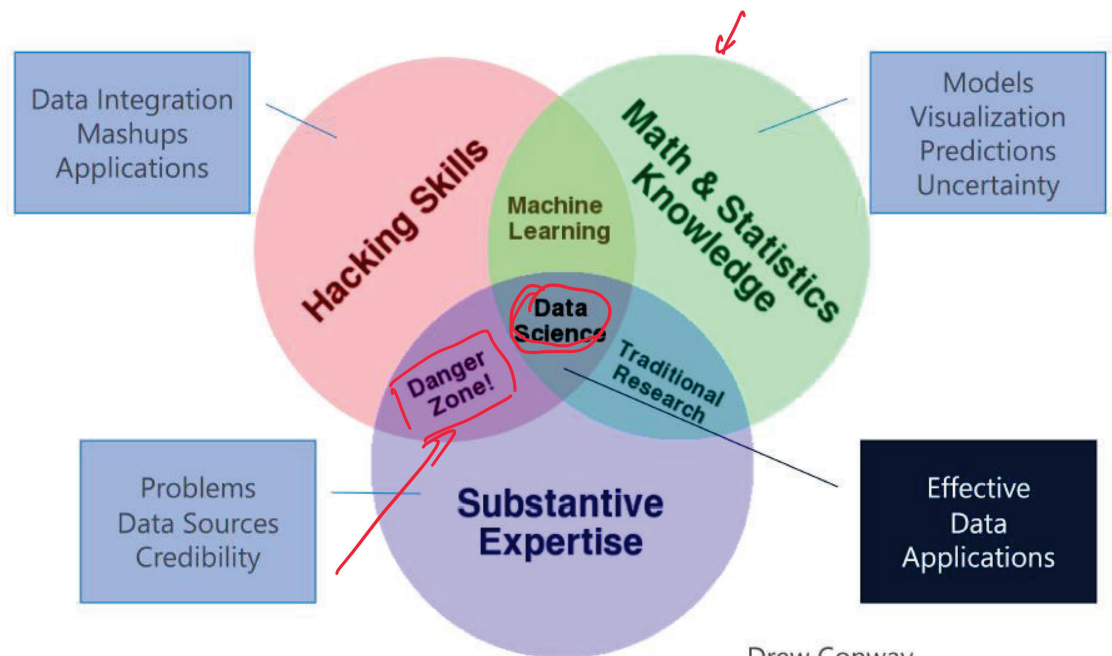
Regression vs. classification

- Supervised learning problems can be further divided into regression and classification problems.
- Regression covers situations where Y is continuous/numerical. e.g.
  - Predicting the value of the Dow in 6 months.
  - Predicting the value of a given house based on various inputs.
- Classification covers situations where Y is categorical e.g.
  - Will the Dow be up (U) or down (D) in 6 months?
  - Is this email a SPAM or not?

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 finance.
- Statistical learning is a fundamental ingredient in the training of a modern *data scientist*.

Data scientist core skills



Drew Conway
http://www.dataists.com/2010/09/the-data-science-venn-diagram/

Some resources

- 11 Clever Methods of Overfitting and how to avoid them
- Big Idea To Avoid Overfitting: Reusable Holdout to Preserve Validity in Adaptive Data Analysis
- 21 Must-Know Data Science Interview Questions and Answers
- 21 Must-Know Data Science Interview Questions and Answers, part 2