Data Mining and Business Intelligence

Lecture 1: Introduction

Jing Peng
University of Connecticut

1/23/20


WELCOME


LETS GET TO
KNOW EACH
OTHER

Background

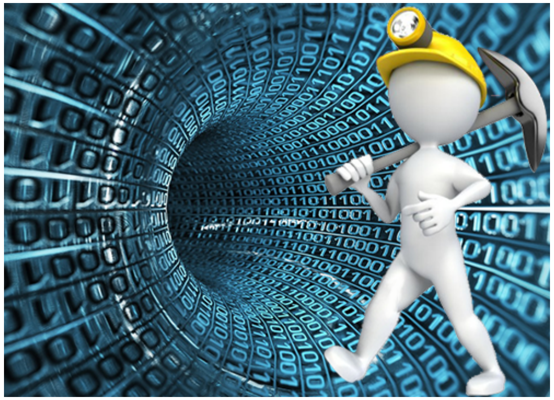
- Background Survey:

  https://uconn.co1.qualtrics.com/jfe/form/SV_3OwmdgtxzoVezFb

Course Introduction

Course Structure

- Data Mining
  - Overview
  - Data Structure, Data Reduction, and Data Acquisition
  - SAS Enterprise Miner
  - Parsing and Quantifying Text
  - Text Mining Application



- Time Series
  - Basics and SAS TSFS
  - Diagnostics
  - Simple Forecasting Models
  - ARIMA Models
  - Regressors
  - Application

Tools

- SAS Enterprise Miner (Data/Text Mining)
- SAS Time Series Forecasting System
- R (used in illustrative examples and Twitter data collection)

Assessment

- Assignments (Individual, 40%)
  - Conceptual questions (one)
  - Developing predictive models on real-world data (two)
- Team Project (Group, 30%)
- Exam (Individual, 25%)
- Participation (Individual, 5%)

Academic Integrity

- Please cite materials (e.g., papers, code, and links) you used in your writeups
- Do not share your solutions or answers with others
- Be honest and fair in peer evaluation

Office Hours

- My office hour
  - Before or after class Wed/Thu (by appointment)

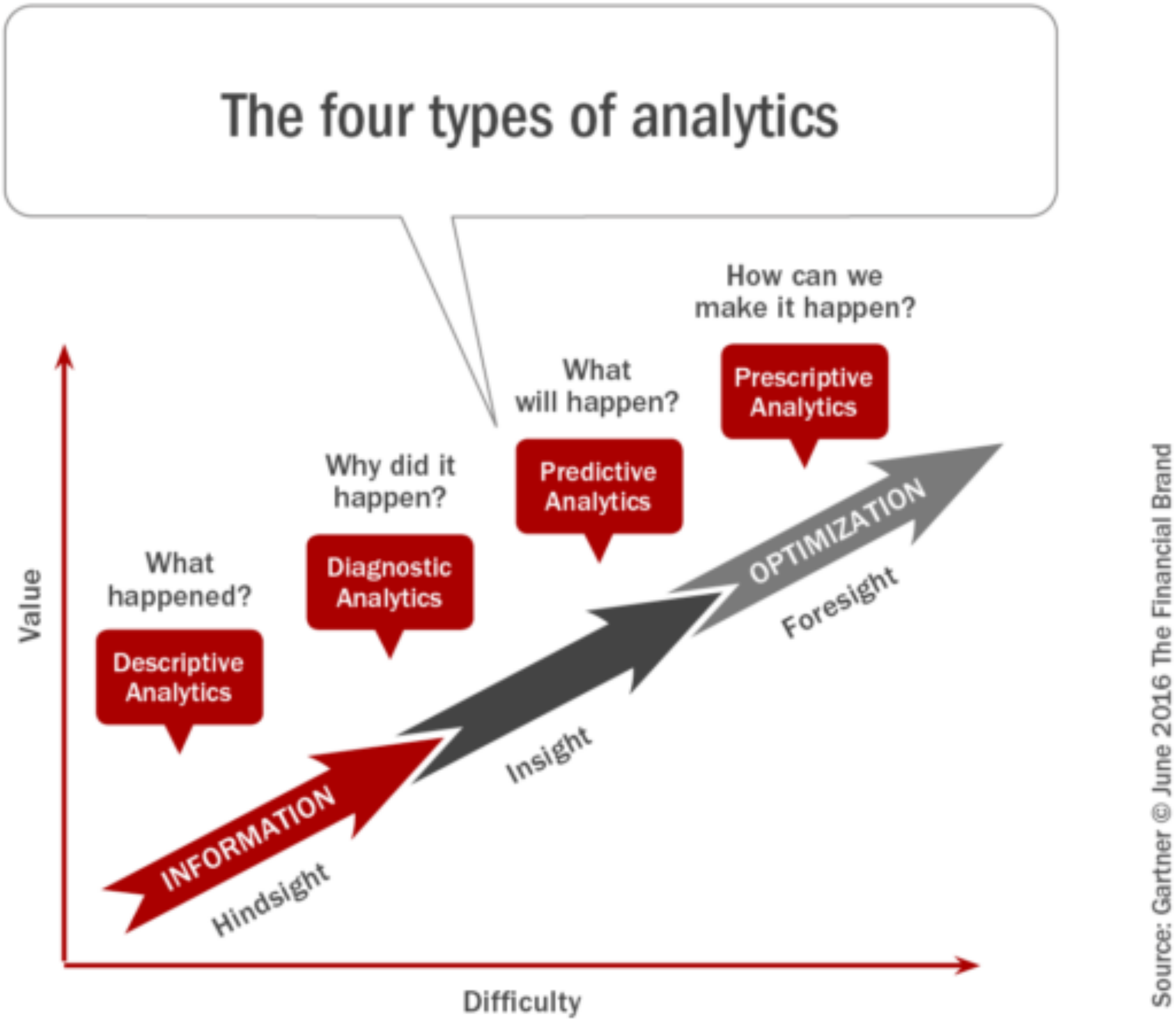
- Grader: Hao Li
  - Email: hao.5.li@uconn.edu
  - Office hour: TBA

Business Analytics

Business Analytics

- **Business analytics** refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning (Wikipedia)
  - Exploring data to find novel patterns
  - Explaining why a certain result occurred
  - **Forecasting future outcome (the focus of this class)**
  - Identify causal relationship between input and output
  - Experimenting to test decisions

Analytics Value Escalator



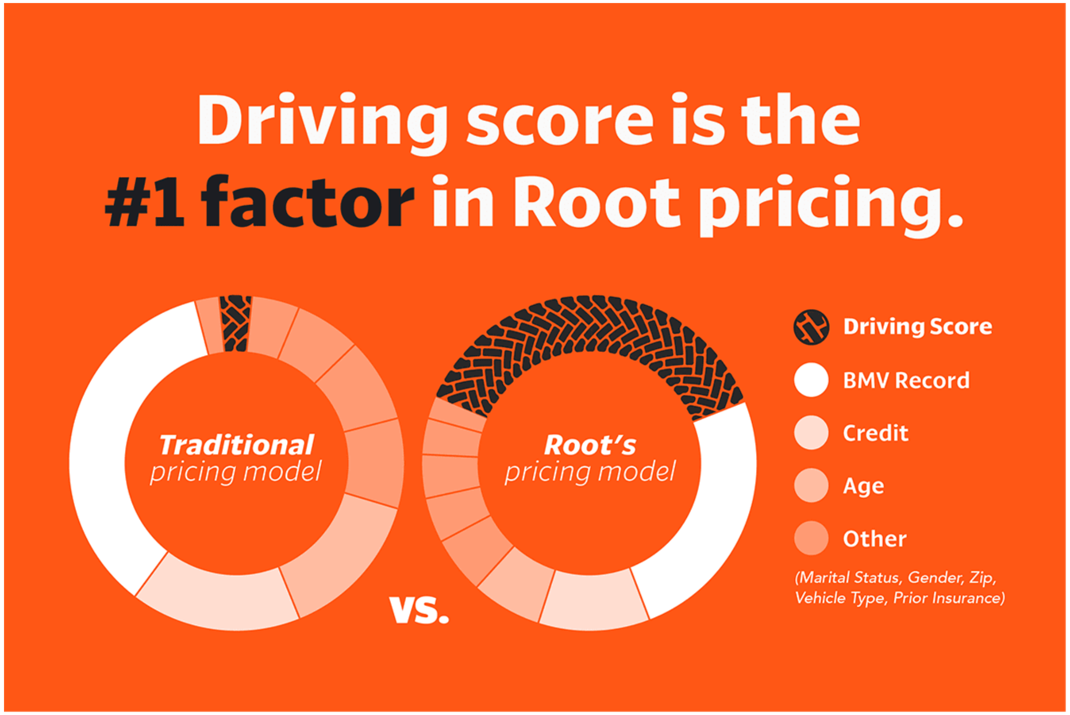
Predictive vs. Prescriptive Analytics

	Predictive	Prescriptive
Focus	Prediction accuracy on the dependent variable	Effect of an intervention on the dependent variable
Implication	Planning & automation	Policy-making
Finding	Correlation	Causality
Model Selection	Performance on validation set	Model fit & assumptions

Two Examples Using Analytics

- Do you really need high speed internet?
  - https://www.wsj.com/graphics/faster-internet-not-worth-it/

- How to lower auto insurance premium?



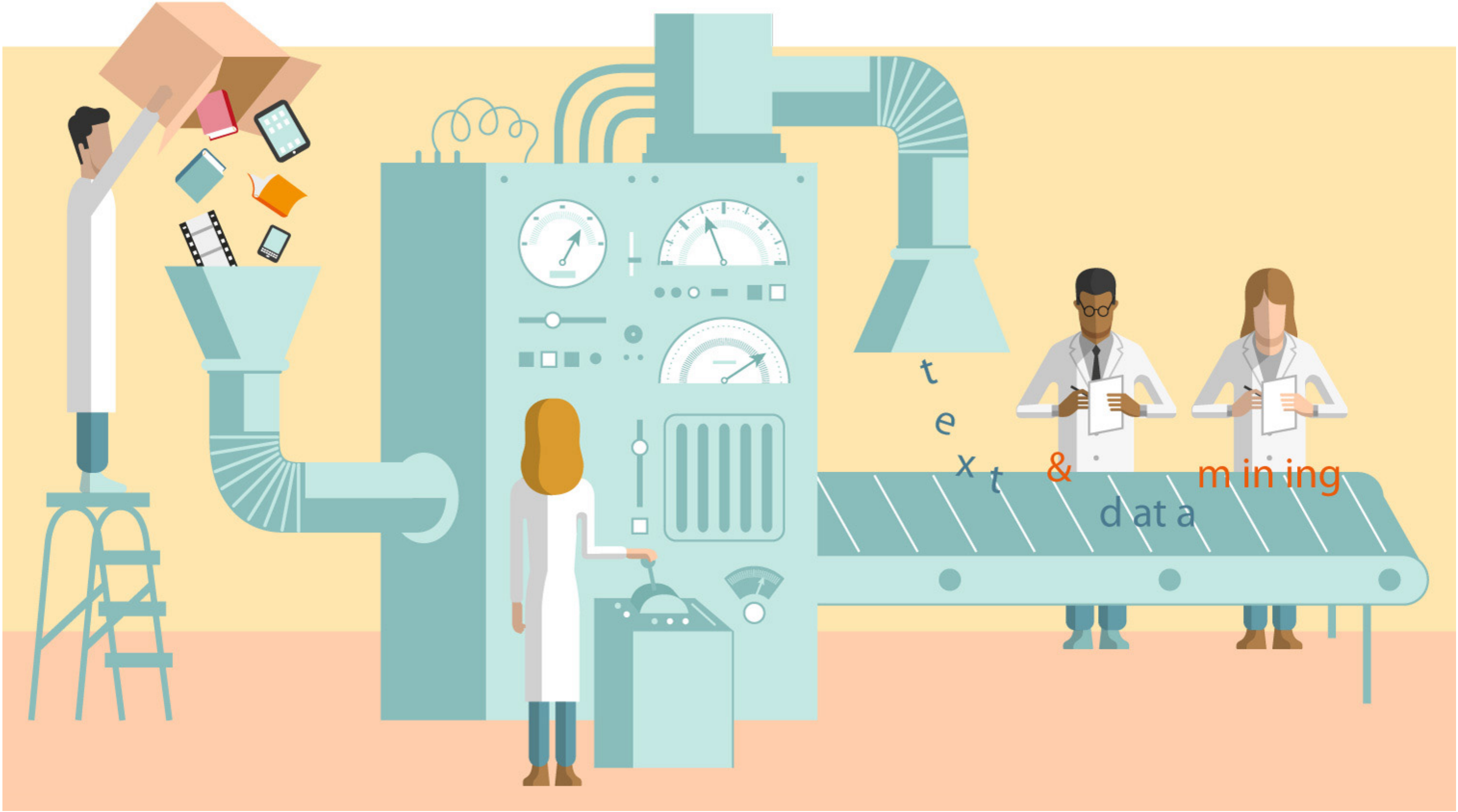
Which types of analytics are used in the two examples?

Source: https://www.joinroot.com/blog/how-root-prices-car-insurance/

Data Mining Overview

Data Mining

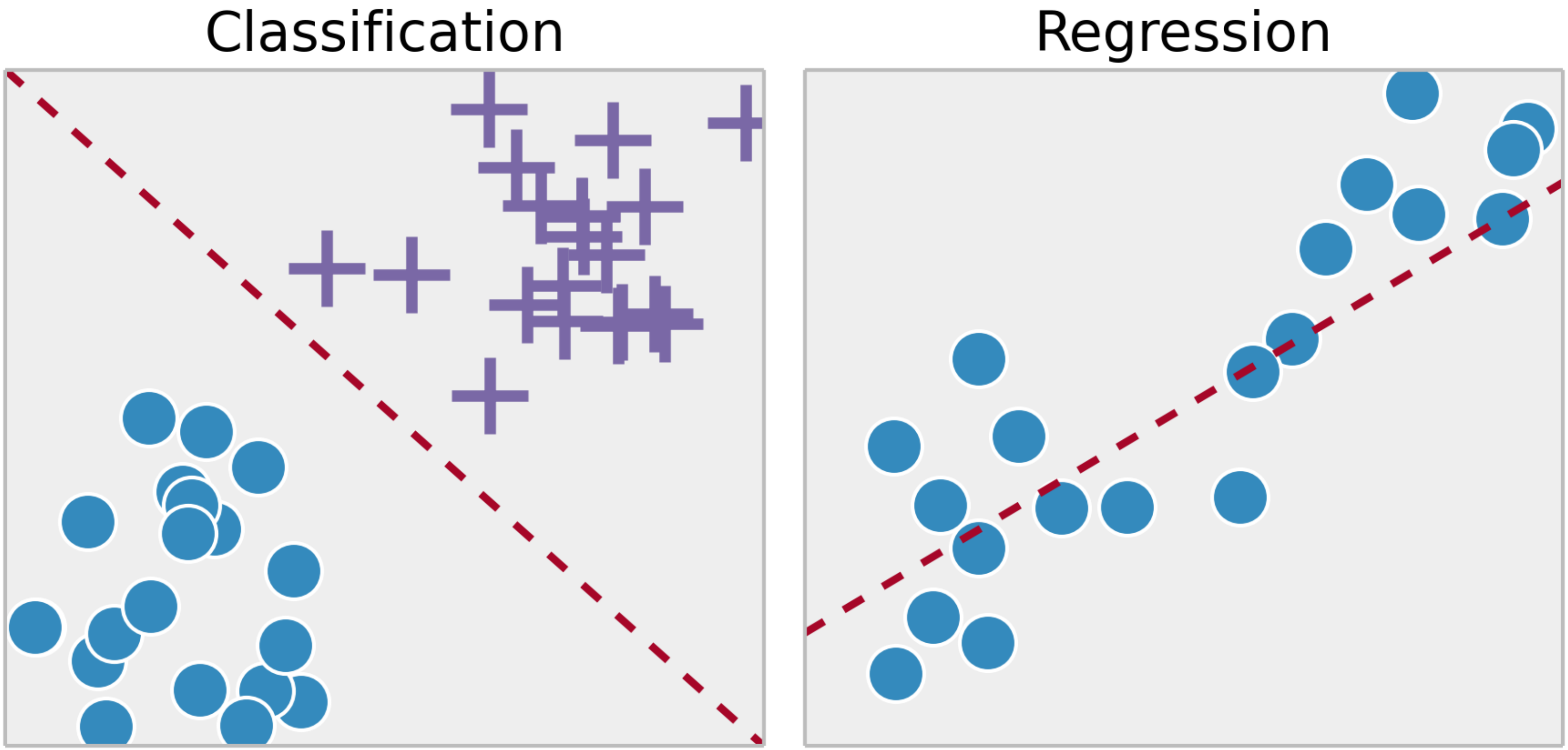
- Data mining is an iterative process of creating **predictive** and **descriptive** models, by uncovering previously unknown trends and patterns in vast amounts of data, in order to support decision making.



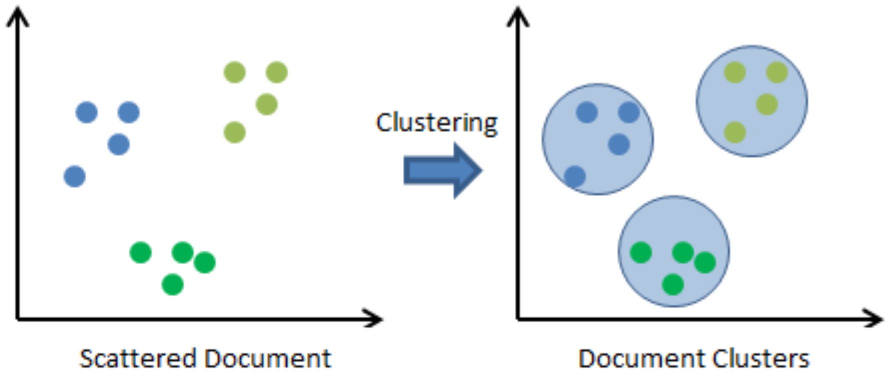
Machine Learning Algorithms for Data Mining

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Association Rule Mining
- Semi-Supervised Learning

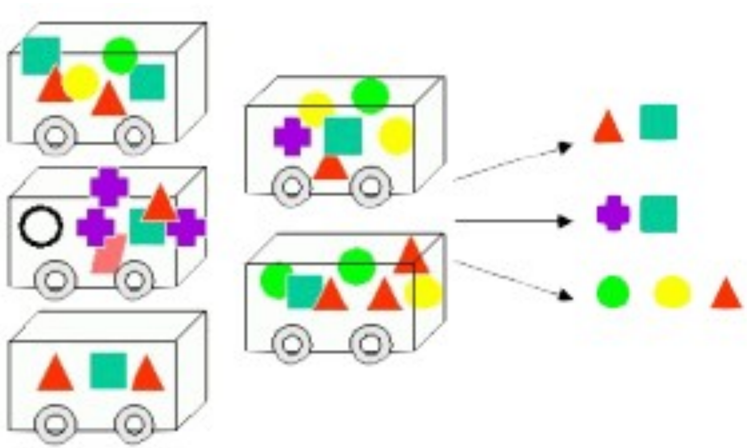
Supervised Learning



Unsupervised Learning

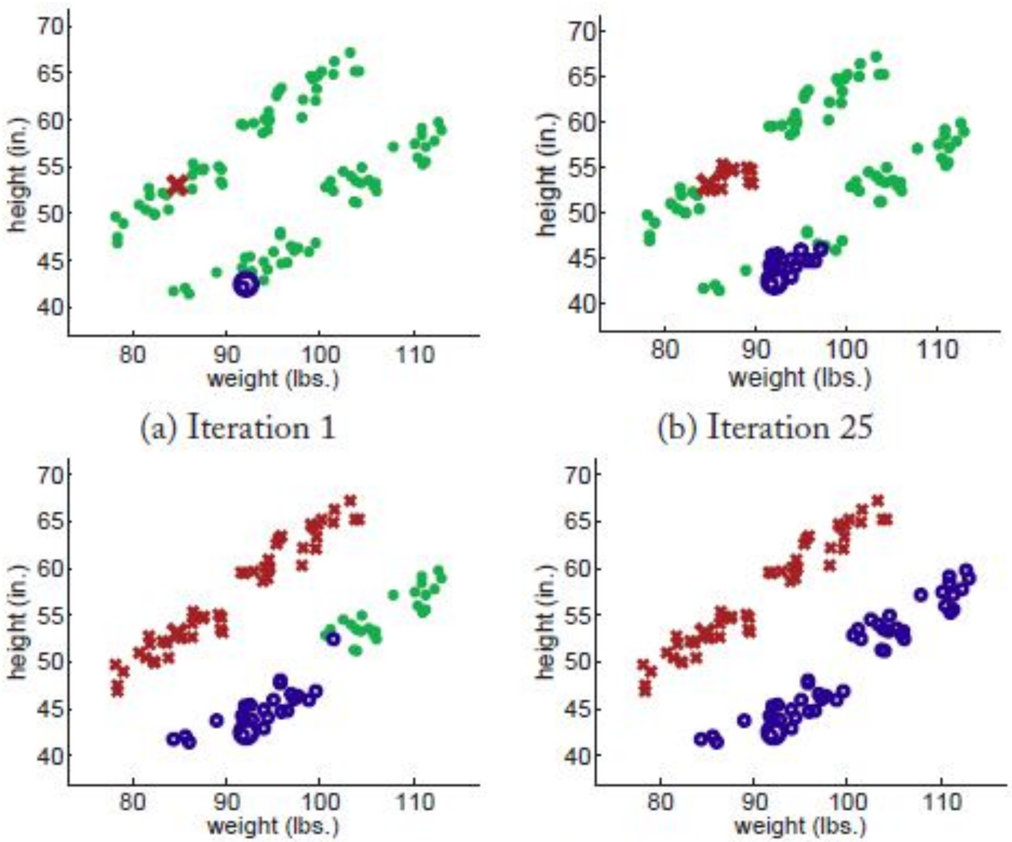


Clustering



Association Rules Mining

Semi-supervised Learning



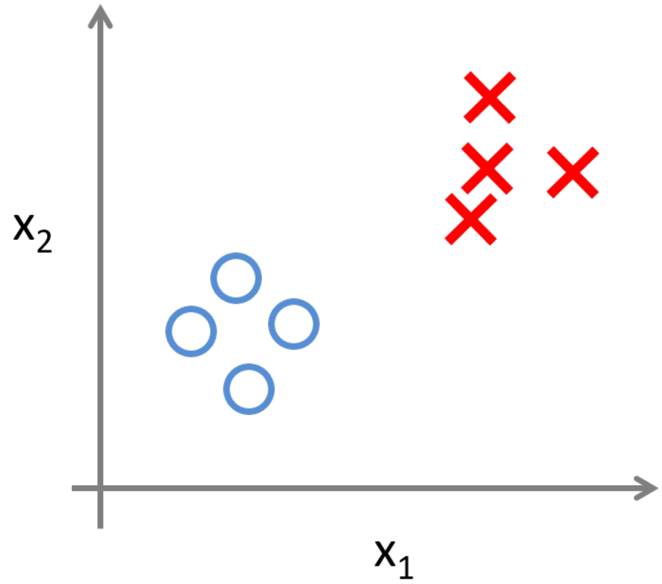
(a) Iteration 1 (b) Iteration 25 (c) Iteration 74 (d) Final labeling of all instances

Propagating 1-nearest-neighbor applied to the 100-little-green-alien data.

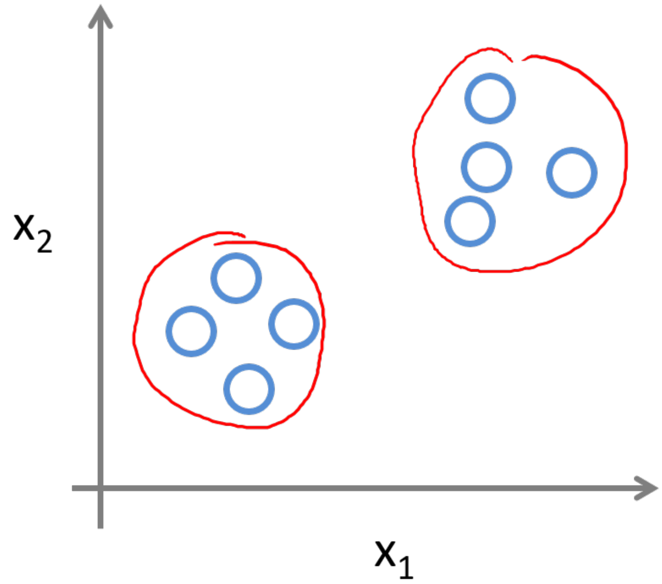
Source: http://cs.calstatela.edu/wiki/index.php/Courses/CS_461/Fall_2011/Bayesian_reasoning

Supervised vs. Unsupervised Learning

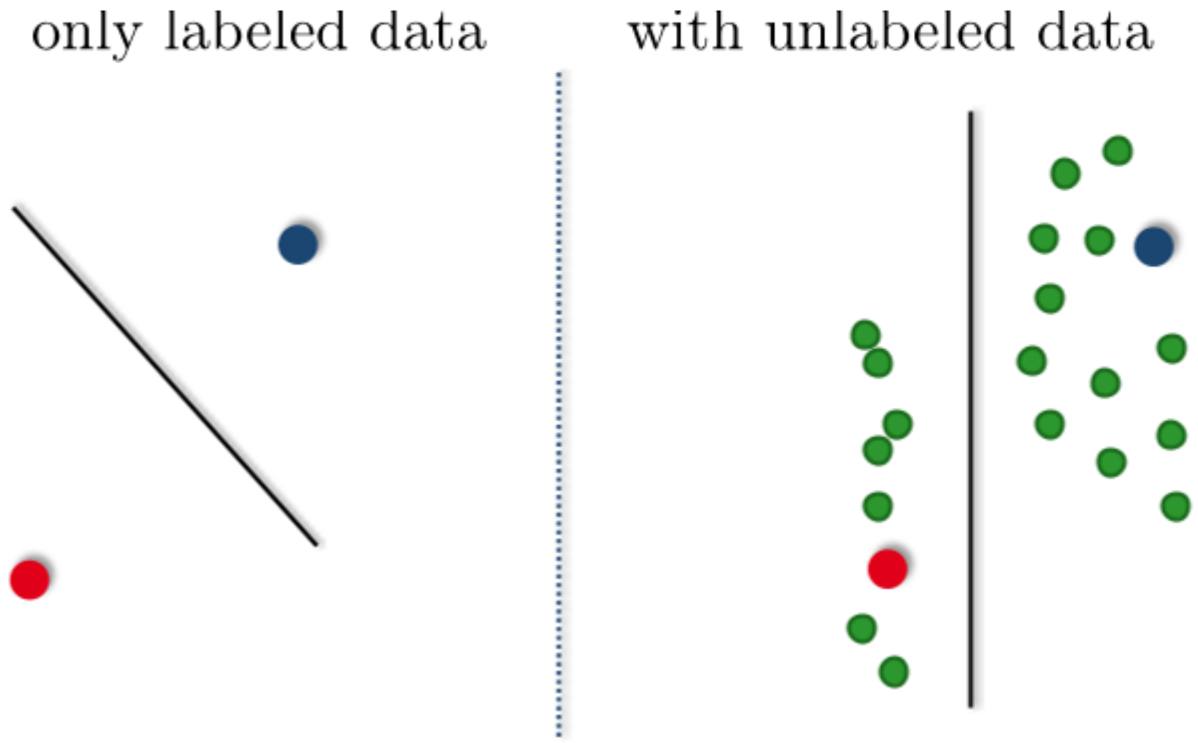
Supervised Learning



Unsupervised Learning



Supervised vs. Semi-Supervised Learning



Applications of Supervised Learning

Classification

- Rain or not
- Default or not
- Buy or not
- Spam detection
- Fraud detection

Regression

- Precipitation
- Default amount
- Sales
- Revenue forecasting
- Stock price

Data Mining vs. Statistics

- Statistics
  - User driven, data is often collected for specific purpose
  - There exist underlying theory about certain relationships in data
  - Use statistical methods to test the theory and/or hypotheses
- Data Mining
  - Data driven, data are often observational and collected for some other purposes
  - Often no pre-existing theory
  - Use statistics, machine learning, and other techniques to examine data and uncover unknown relationships

Key Assumptions for Data Mining

- Past behavior is a good predictor of future behavior
- Data are available for use
- Data contain what you want to predict

Model Evaluation
(Classification)

Confusion Matrix

Predicted Class	True Outcome : Patients have Disease A	
	Positive (Patients have disease A)	Negative (Patients do not have disease A)
Positive (Patients have disease A)	True Positives	False Positives (Patients wrongly identified to have disease A)
Negative (Patients do not have disease A)	False Negatives (Patients have been left out from treatment for Disease)	True Negatives

Evaluation Metrics

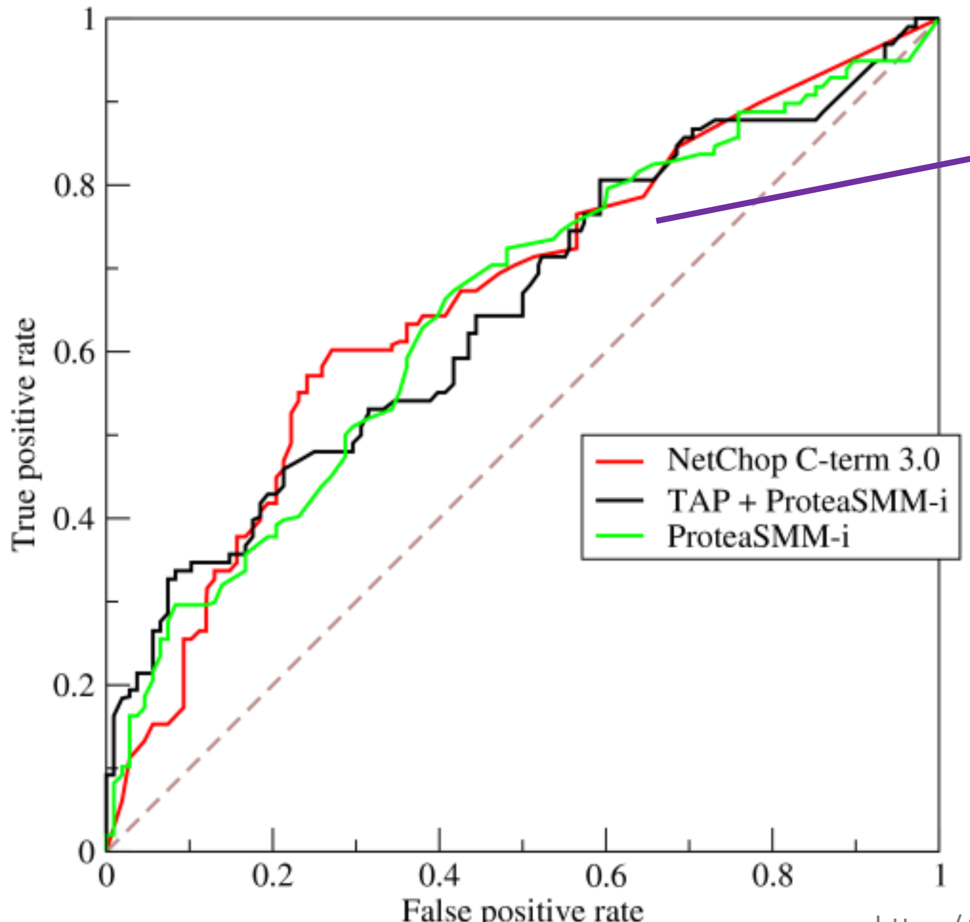
		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Type I error: Rejection of a true null hypothesis (significance level α)

https://en.wikipedia.org/wiki/Confusion_matrix

Receiver Operating Characteristic (ROC)

- ROC curve: how the performance of a binary classifier, measured as False Positive Rate (x-axis) and True Positive Rate (y-axis), vary with cutoff threshold

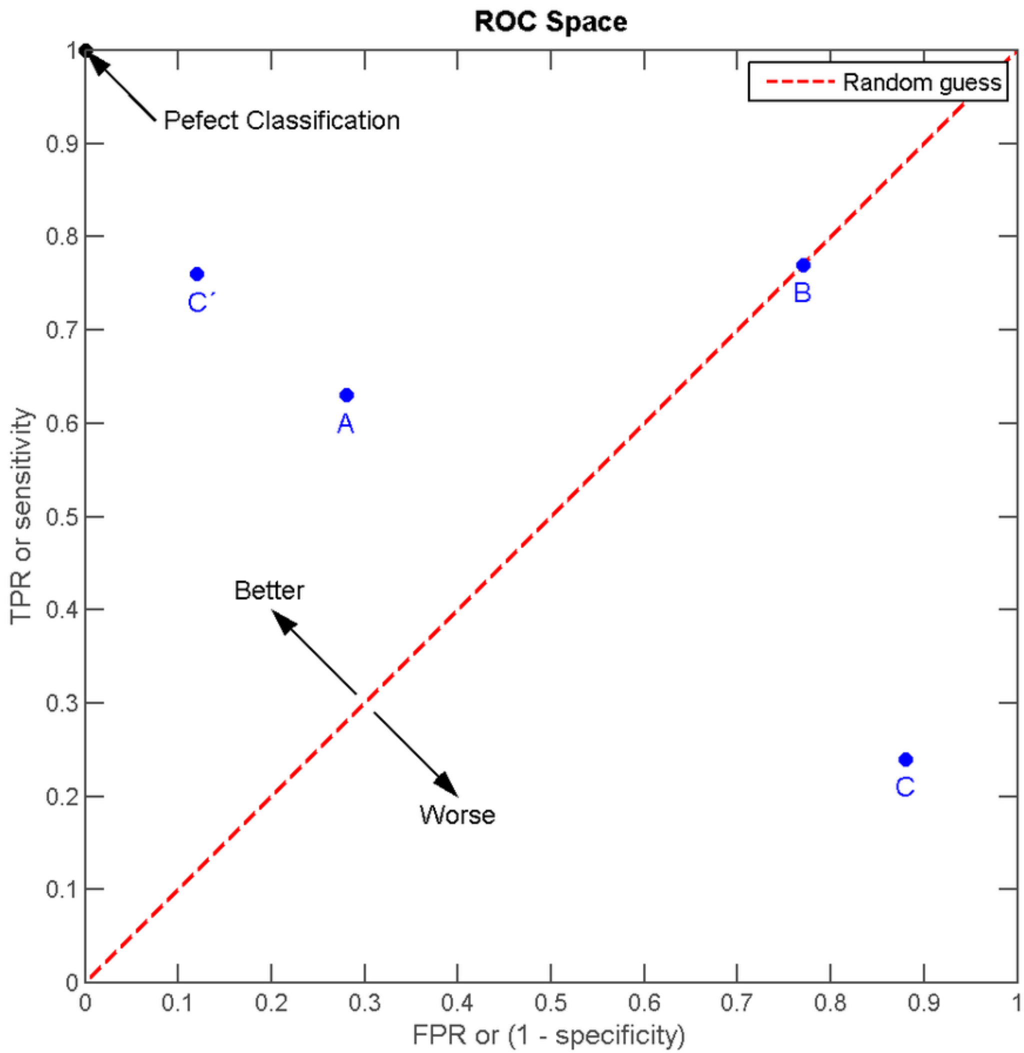


Area Under Curve (AUC)

AUC (referred to as ROC in SAS) equals to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Receiver Operating Characteristic (ROC)



Suppose the data are balanced, what are the accuracy and AUC of random guess?

What if the data are highly imbalanced?

- Does random guess still give an accuracy of 50%?
- Does random guess still give an AUC of 0.5?
- How about predicting every instance as the majority?

ROC vs. Accuracy

- Accuracy
  - Classification performance at a given threshold
  - Useful when discrete decisions need to be make
- ROC
  - Overall performance at all possible thresholds
  - Uses information about the ranking of the predictions
  - Helpful when outcome is highly imbalanced

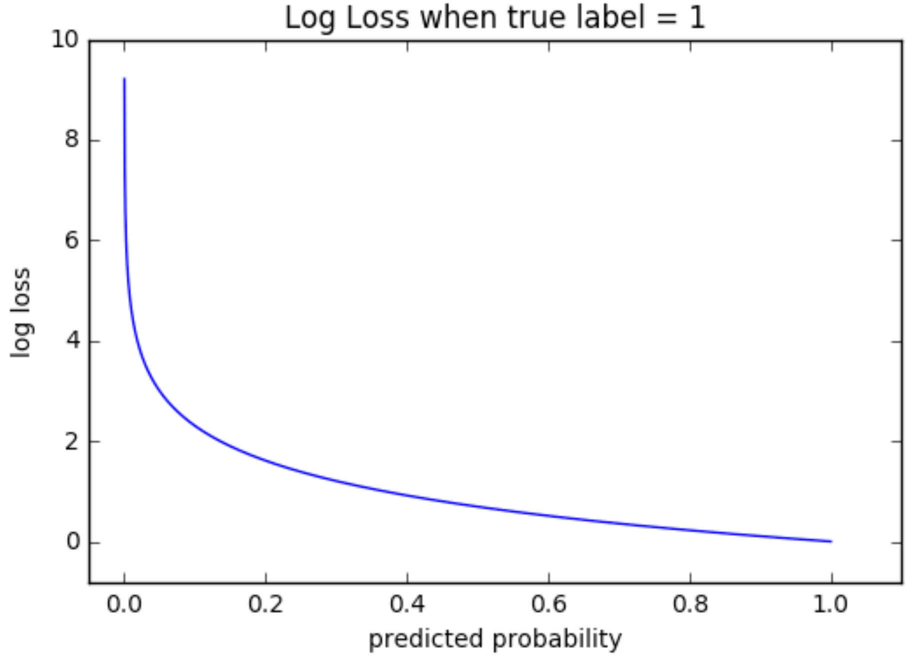
- What if these two are not consistent?

Log Loss

- Log loss (cross-entropy loss) for binary classification:

$$-\sum_{i=1}^{N} y_i \ln p_i + (1 - y_i) \ln(1 - p_i)$$

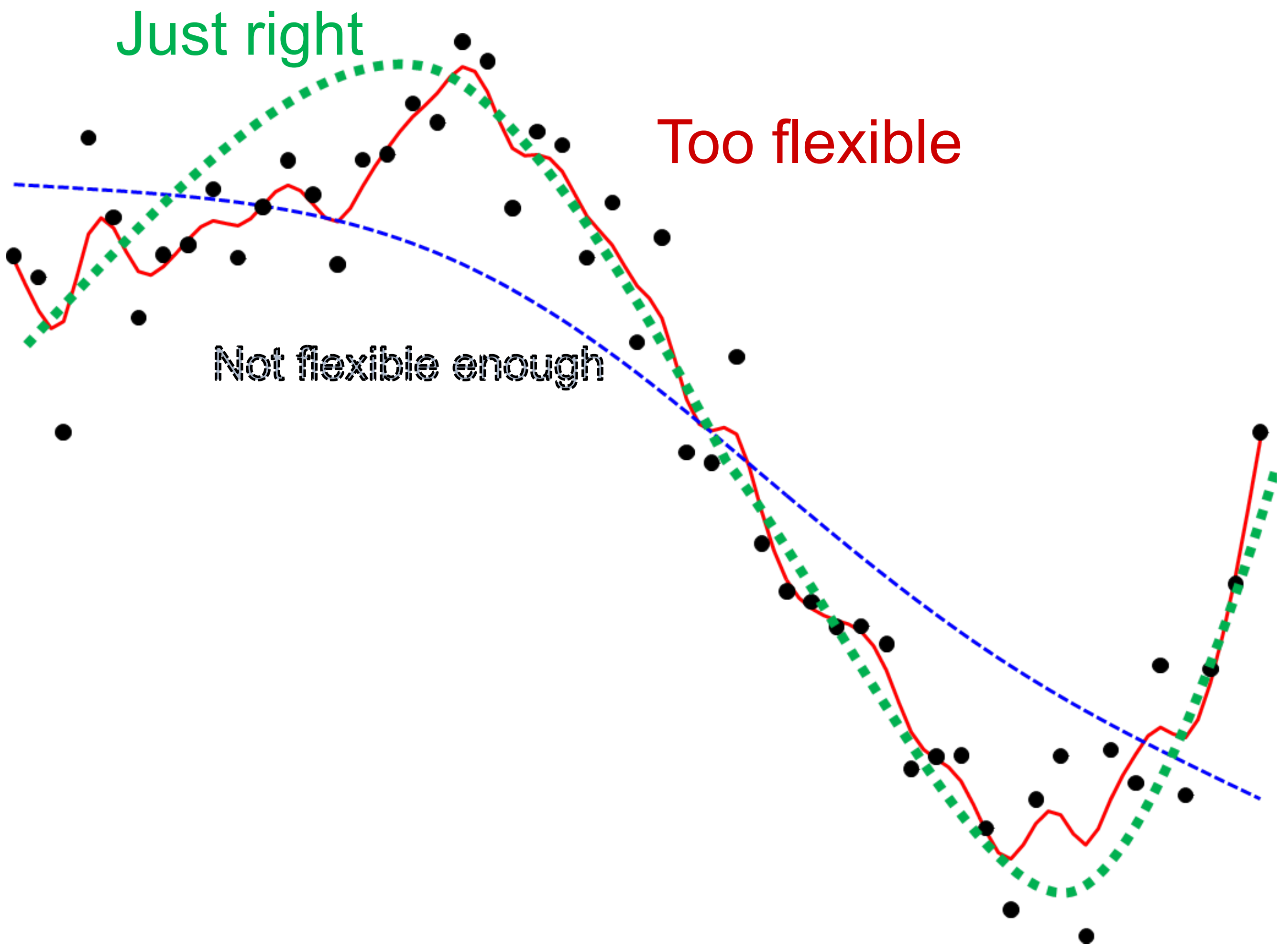
- Log loss function heavily penalizes predictions that are confident and wrong!



Source: http://wiki.fast.ai/index.php/Log_Loss

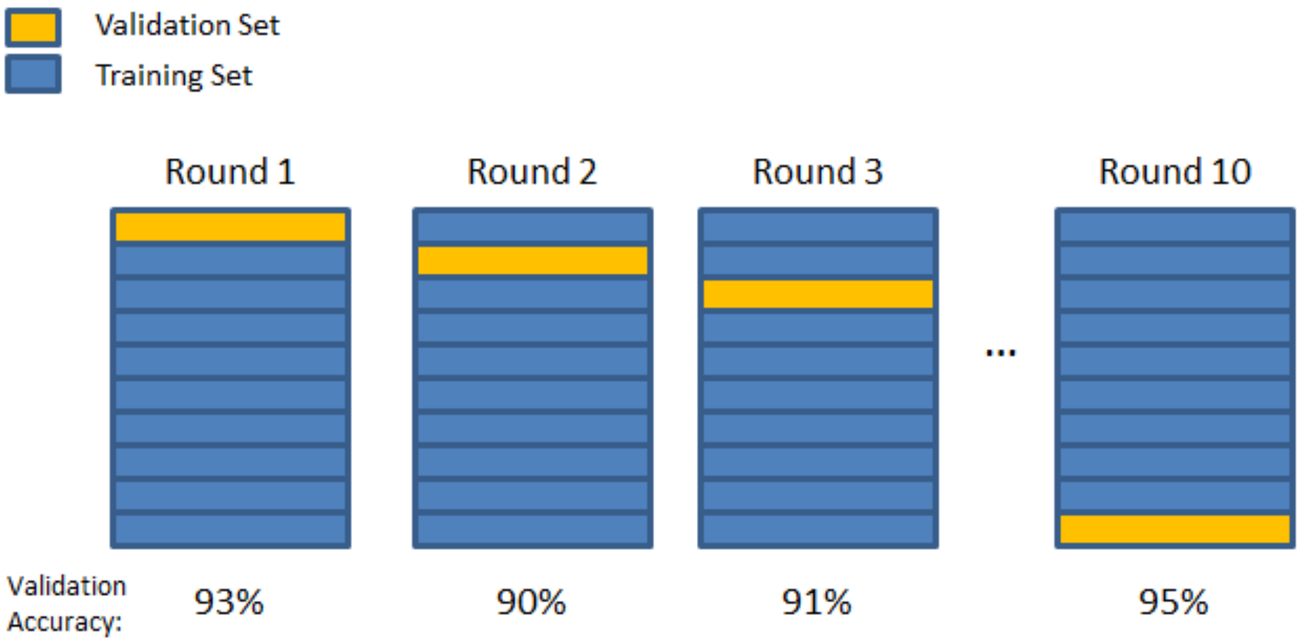
Model Selection

Model Complexity



Honest Assessment

Cross Validation



Final Accuracy = Average(Round 1, Round 2, ...)

Repeated Cross-Validation

- Repeat the Cross-Validation process multiple times and then take the average
  - The group assignments are different across cross-validations
  - Can deliver smaller bias than standard cross-validation
  - May construct confidence intervals in a bootstrap manner

Kim, J.-H. 2009. "Estimating Classification Error Rate: Repeated Cross-Validation, Repeated Hold-out and Bootstrap," *Computational statistics & data analysis* (53:11), pp. 3735-3745

Access to SAS

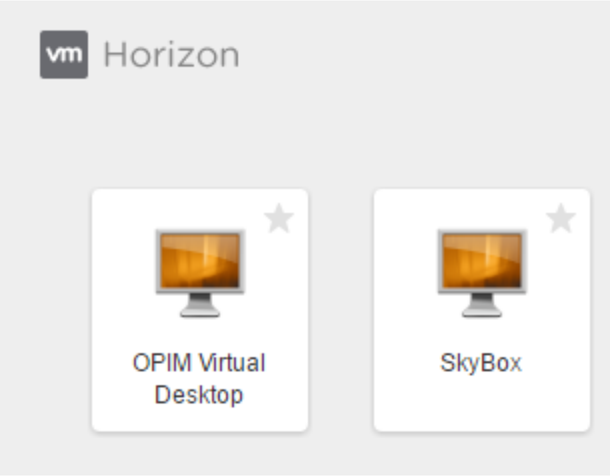
SAS Options

Option	Available Component	Note
Installing SAS on PC	TSFS	**Recommended**, if you have >20GB free disk space
Skybox or AnyWare	TSFS	Remote access, fast, need to save files to P drive
OPIM Virtual Desktop	TSFS and Enterprise Miner	Remote access, slow, need to save files to P drive

* If you use Skybox or OPIM Virtual desktop, you may want to install the client to access virtual machines.

Virtual Machines

- Skybox
  - Fast
  - University level
  - No SAS Enterprise Miner



- OPIM Virtual Desktop (OVD)
  - Slow and frequent log-in
  - department level
  - Include all software needed

- All data stored within the virtual machine (Skybox or OVD) will be wiped upon restart. Store files in P drive if you use the virtual machine

Virtual Machines Client

- Install VMWare client
  - https://my.vmware.com/en/web/vmware/info/slug/desktop_end_user_computing/vmware_horizon_clients/4_0#win64

- Connect to Server
  - https://confluence.uconn.edu/busnit/opim-virtual-desktop/converting-to-the-new-opim-virtual-desktop
  - New server: horizon.uconn.edu

- If you can’t see “OPIM Virtual Desktop”, contact IT department
  - (860) 486-5450
  - help@business.uconn.edu