

Week 3

# Concepts in Machine Learning

[fredhutch.io](http://fredhutch.io)

Fred Hutchinson Cancer Research Center

# Week 3 Learning Objectives

By the end of today's class, you should...

## CRISP-DM

- Review what each step generally entails

## Classification

- Recognize the difference between binary and multiclass classifiers
- Understand the difference between hard and soft classifiers

## Evaluating Hard Classification Models

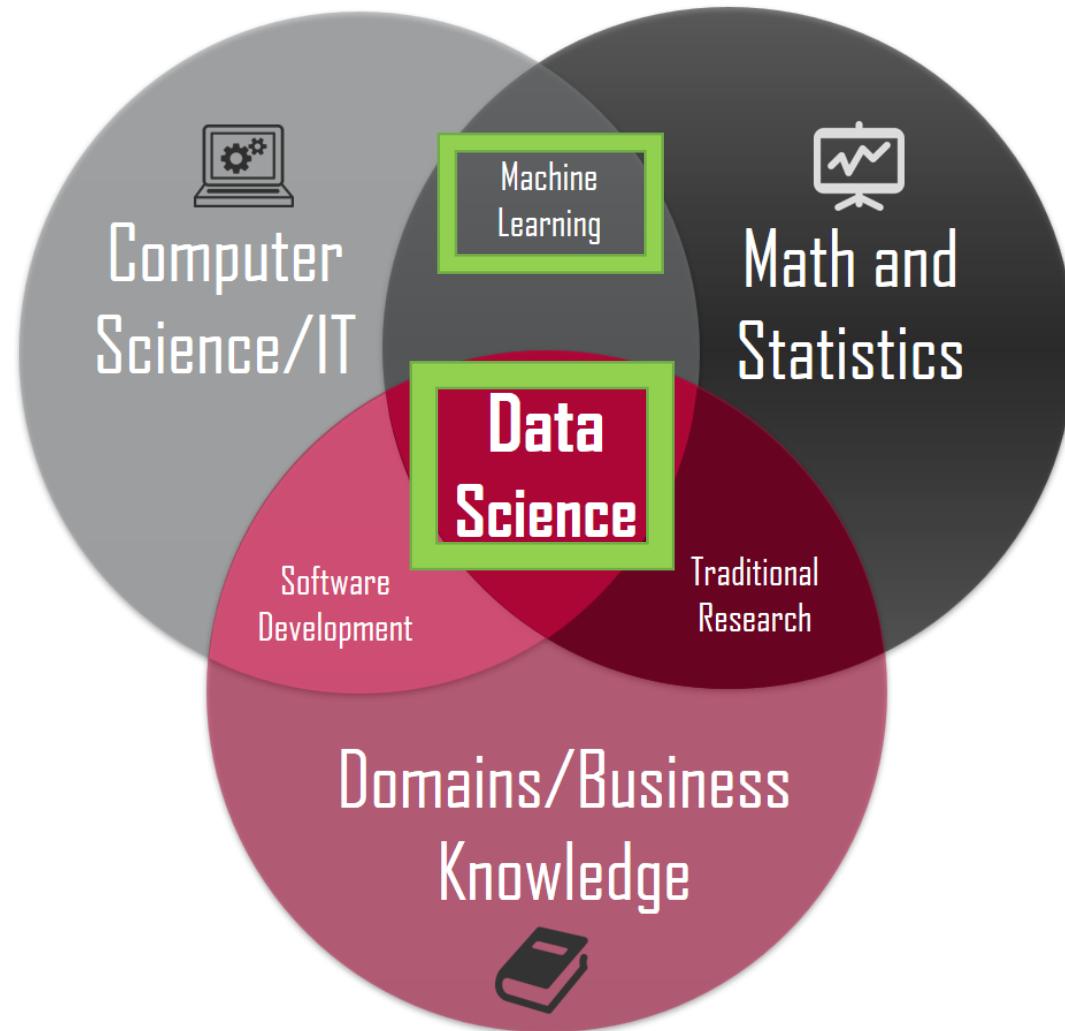
- Know how to interpret the quadrants of a confusion matrix (TP, FN, FP, TN)
- Understand how those quadrants form useful sets of ratios

## Evaluating Soft Classification Models

- Recognize common proper scoring rules for soft classifiers (especially log-loss)
- Understand how thresholding can convert a soft classifier to a hard classifier
- Recognize how to evaluate models at different thresholds using ROC curves and AUC

# Definitions

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.  
[\(https://expertsystem.com/machine-learning-definition/\)](https://expertsystem.com/machine-learning-definition/)
- Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. [\(https://en.wikipedia.org/wiki/Machine\\_learning\)](https://en.wikipedia.org/wiki/Machine_learning)
- Problem + Data + Algorithm(self-adjusting) + Compute ==> Insight



# An Imperfect Analogy: Cabinet Making



(not technically a cabinet)

# Capable Cabinet Maker...

- Inspects and understands raw materials
- Uses the tools thoughtfully to shape and join materials
- Chooses approach and tools based on materials and goals
- Applies thoughtfulness born of experience



...but how does this relate  
to machine learning?

# Capable Cabinet Maker ML Practitioner...

- Inspects and understands ~~raw materials~~ **data**
- Uses the tools thoughtfully to shape and join ~~materials~~ **data**
- Chooses approach and tools based on ~~materials~~ **data** and goals
- Applies thoughtfulness born of experience



...and the **tools**?

# SAT Style Analogy

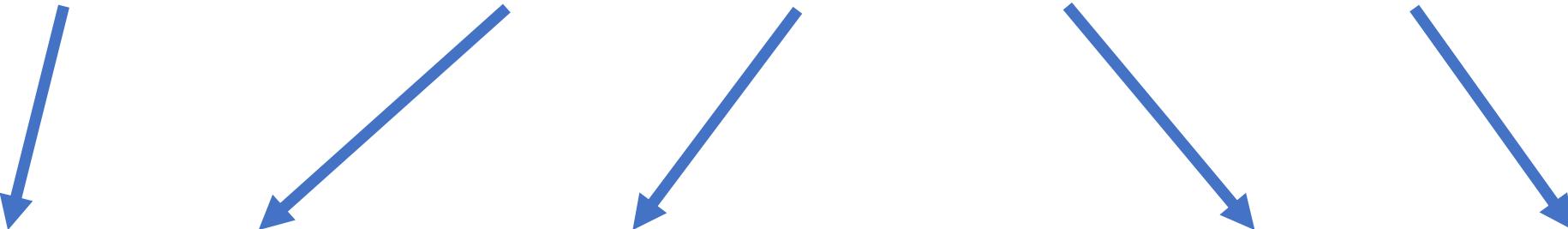
Tools : Cabinet Making

as

Algorithms : Machine Learning

# An Imperfect (Extended) Analogy

- Storage Need + Raw Materials + Tools + Work ==> Cabinet



- Problem + Data + Algorithm(self-adjusting) + Compute ==> Insight

# Brief Aside: Experimental Design

- Difficult to master or even do well
- Close interplay between
  - Goals
  - Methods
  - Data
  - Execution
- Requires thoughtful approach and broad understanding

# Brief Aside: Experimental Design

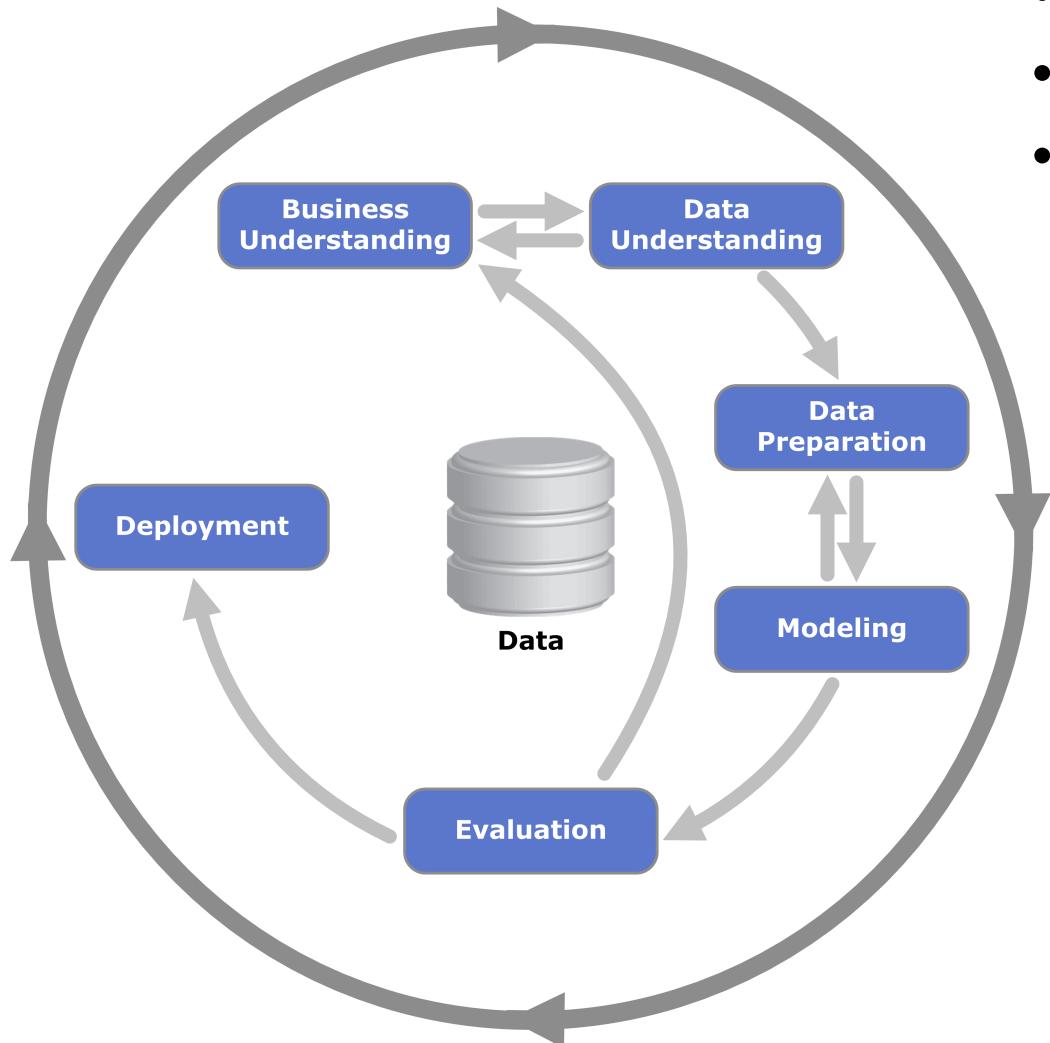
- Difficult to master or even do well
- Close interplay between
  - Goals
  - Methods
  - Data
  - Execution
- Requires thoughtful approach and broad understanding

Because machine learning shares so many of these characteristics, it's helpful to follow a process

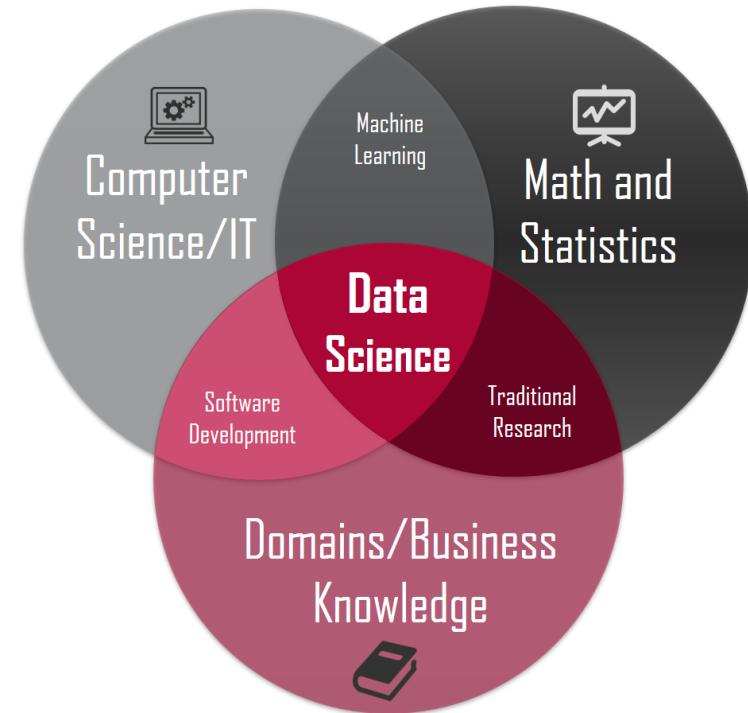
Experts in machine learning often are guided by the steps laid out in...

# CRISP-DM

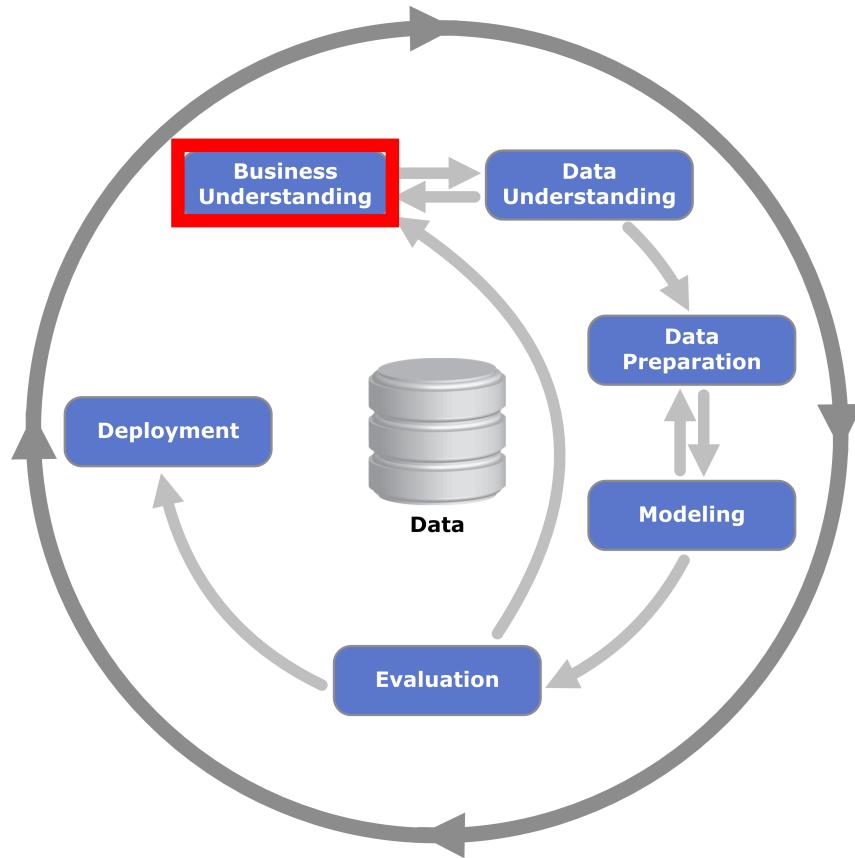
## Cross-industry standard process for data mining



- Cyclical
- Iterative
- Connecting all 3 areas of the classic notion of “data science”

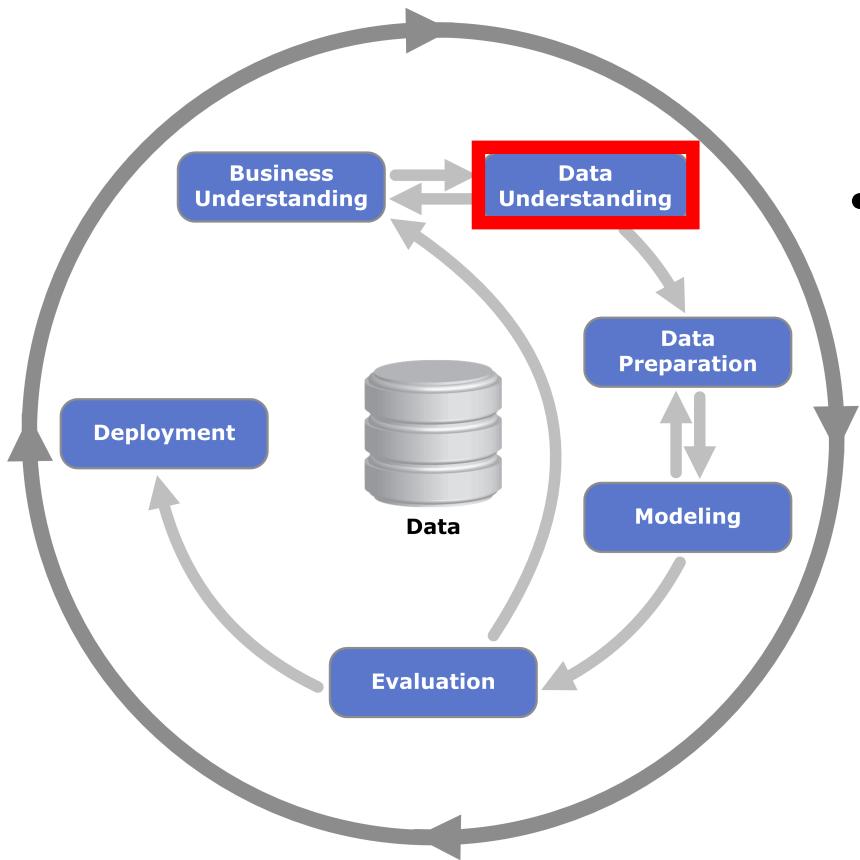


# Business/Scientific Understanding



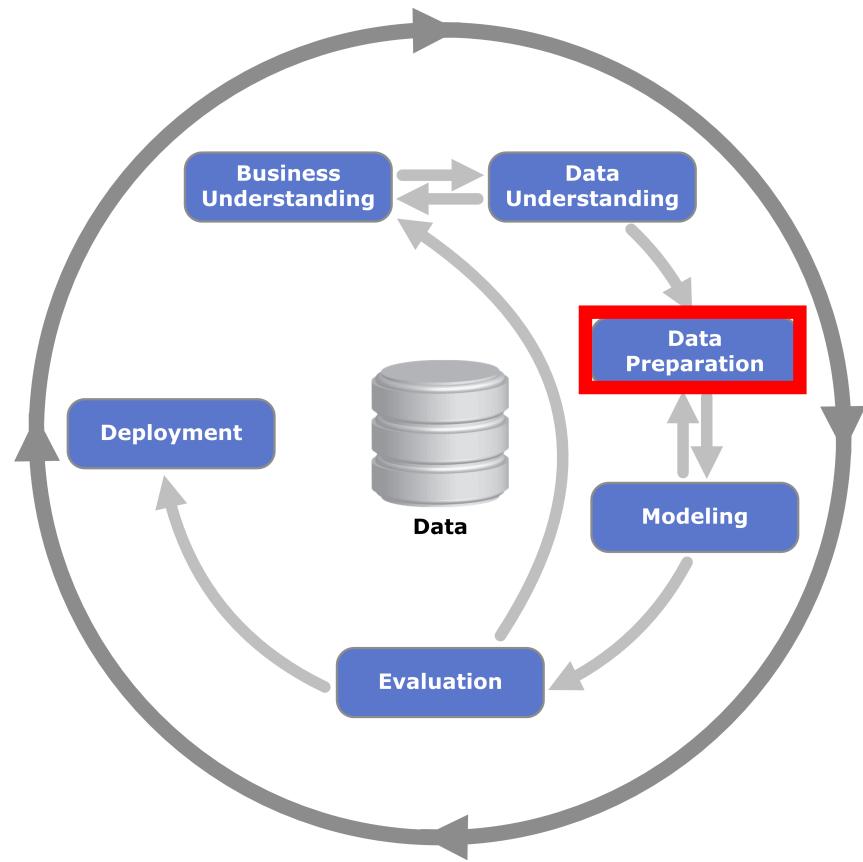
- Scientific domain knowledge
- What do we already know or believe?
- What are our research aims?
- How do we think we should proceed initially?
- Nearly every choice within any of the subsequent steps should refer back to this step!

# Data Understanding



- Exploratory Data Analysis (EDA)
  - Data structure
  - Data quality
  - First insights
  - Interesting subsets
  - Form hypotheses for hidden information

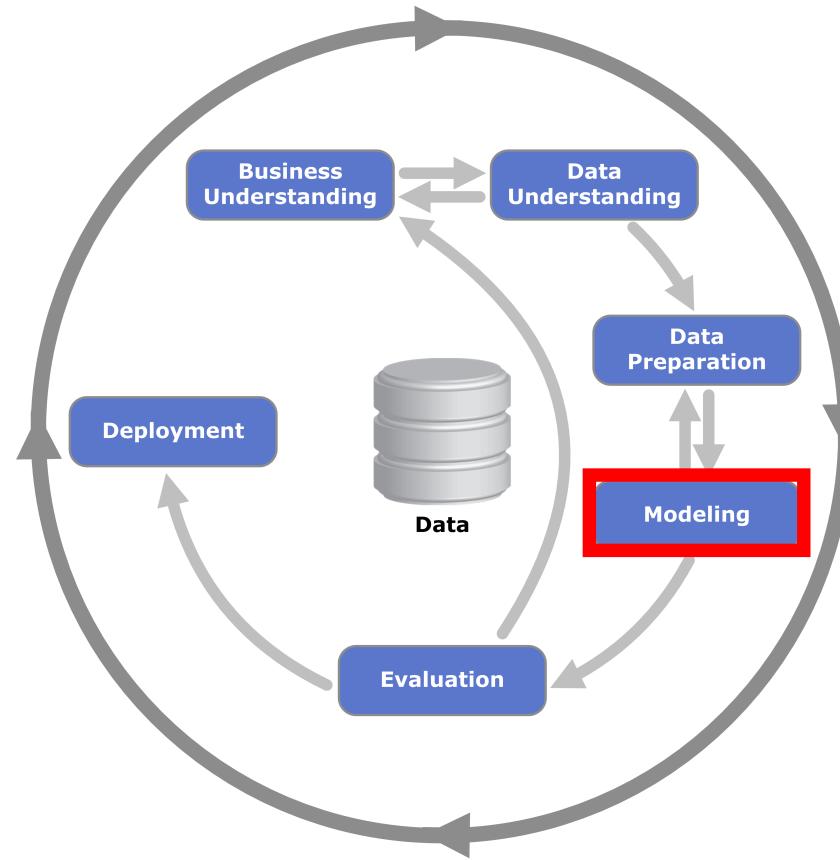
# Data Preparation



- Data acquisition
- Data selection
- Data integration and formatting
- Data cleaning
- Data transformation and enrichment

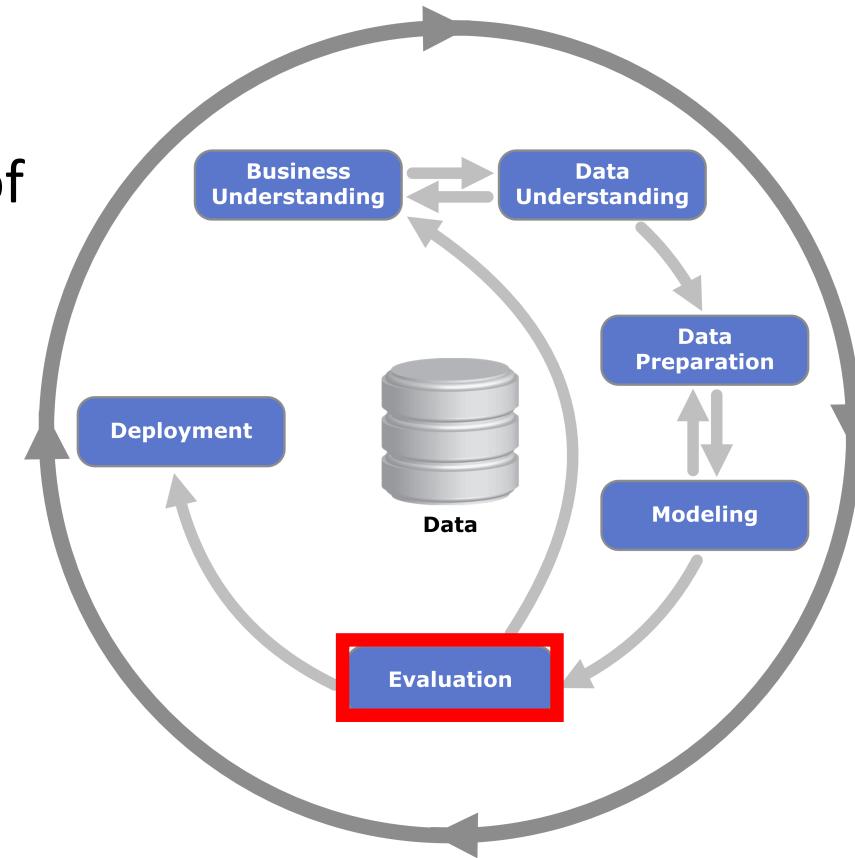
# Modeling

- Selection of appropriate modeling technique
- Splitting of the dataset into training and testing subsets
- Examination of alternative algorithms and parameter settings
- Fine tuning of the model settings



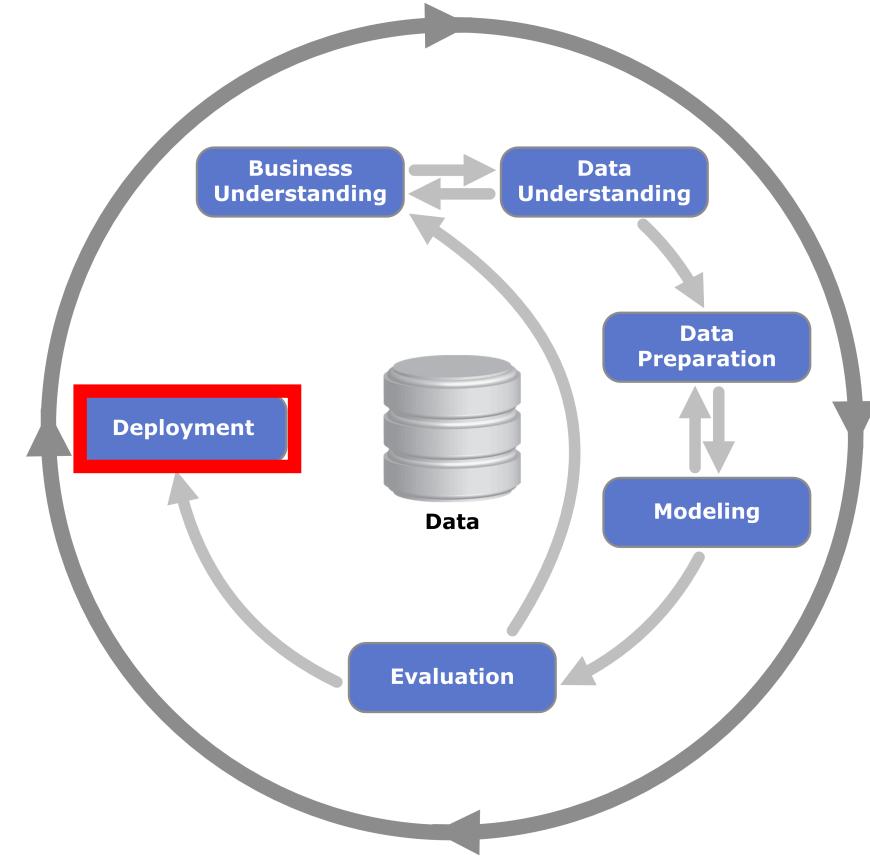
# Evaluation

- Evaluation of the model in the context of the scientific success criteria
- Performance relative to TEST data with chosen loss function
- Balancing tradeoff between bias and variance



# Deployment

- Will be specific to each problem space
  - At FHCRC could relate to grants or publications



# Machine Learning Paradigms

---



## par·a·digm

/'perə,dīm/

*noun*

1. a typical example or pattern of something; a model.  
"there is a new paradigm for public art in this country"

Similar:

model

pattern

example

standard

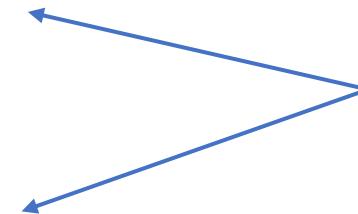
prototype

archetype



# 3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning



We'll focus on these

- Reinforcement Learning
- Semi-Supervised Learning

# 3 or 4 Machine Learning Paradigms

- Supervised Learning

- Unsupervised Learning

- Reinforcement Learning

- Semi-Supervised Learning

Data “Prediction”



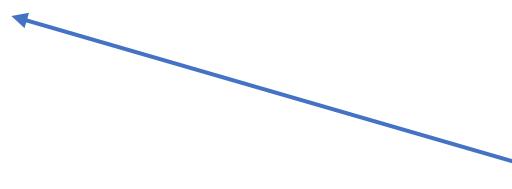
# 3 or 4 Machine Learning Paradigms

- Supervised Learning

- Unsupervised Learning

- Reinforcement Learning

- Semi-Supervised Learning

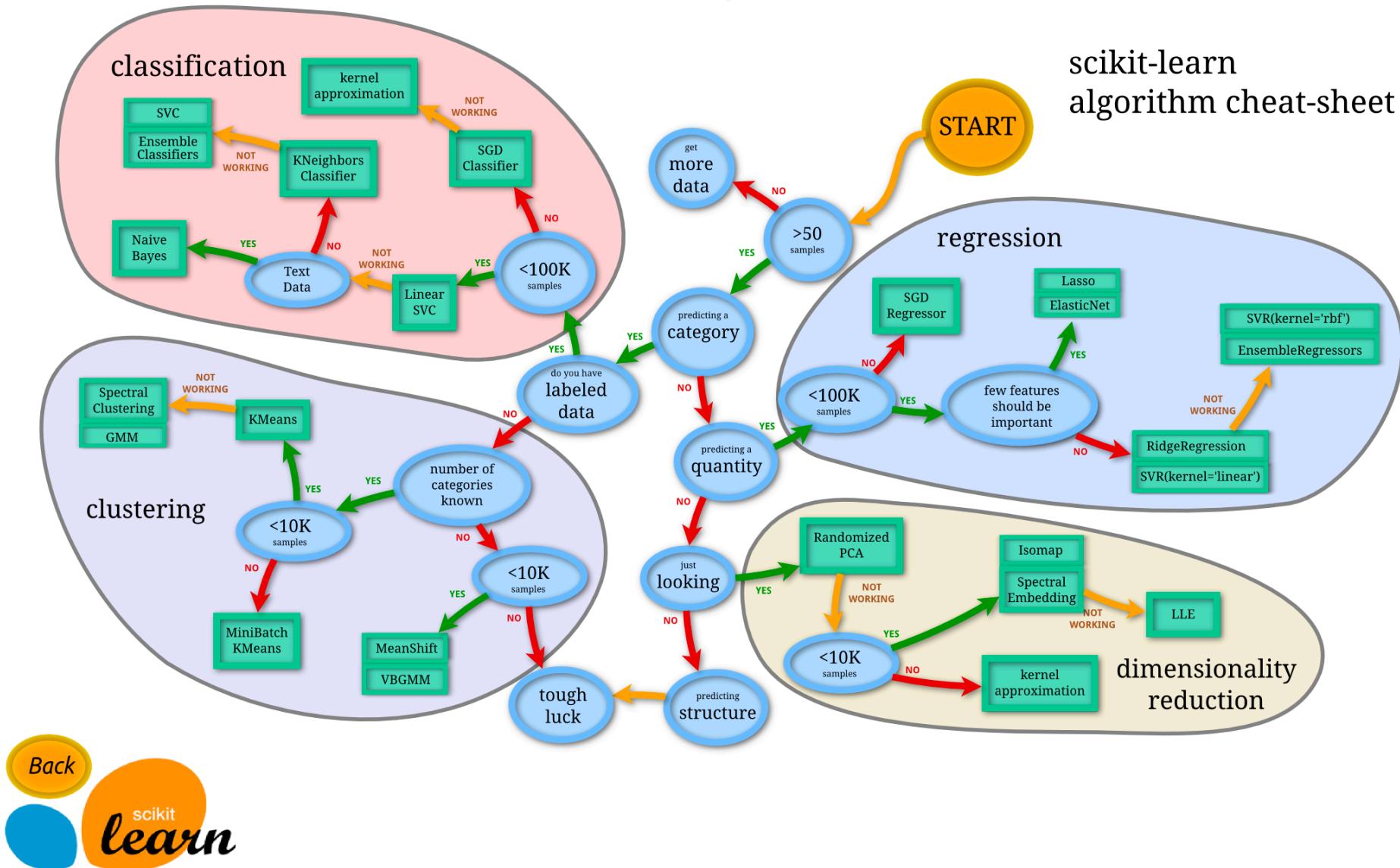


Data “Expression”

# Supervised Problems: Categorical Problems

- Can we state our outcome as a choice of A **vs.** B? (could be any number of categories)
  - Yes, today we're interested in categories

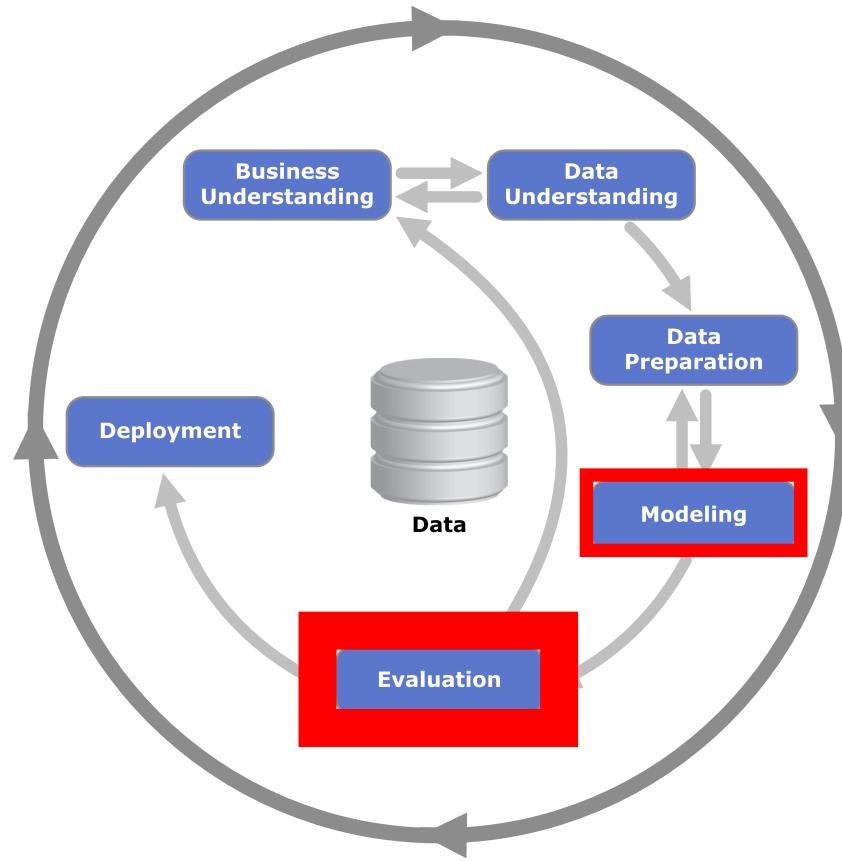
# A Nifty Chart



# A Nifty Chart

[https://scikit-learn.org/stable/tutorial/machine learning map/](https://scikit-learn.org/stable/tutorial/machine_learning_map/)

# Modeling & Evaluation

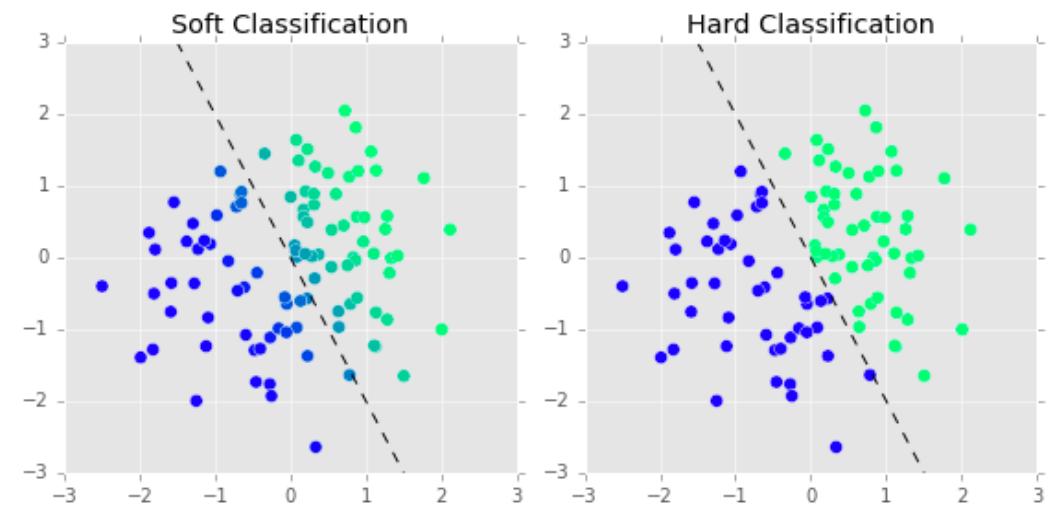


# Classification - Binary vs Multiclass

- Classification → Categorical Response
  - red, yellow, or green → Multiclass (class A or B or C or...)
  - dead or alive → Binary (A or B, A or  $\neg$ A)
  - HIV- or HIV+ → Binary (- or +, 0 or 1)
- Binary response → Only two values: 0 or 1
  - Most common case
  - Response 0 → negative class
  - Response 1 → positive class

# Classification - Soft vs Hard

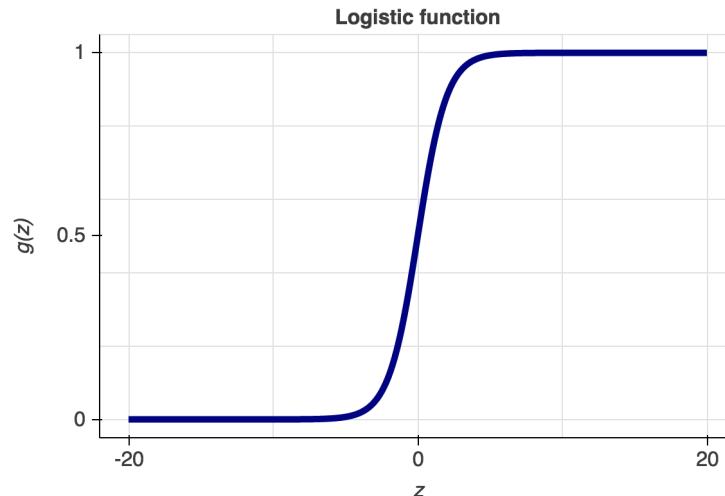
- Soft Classification:
  - Attempt to estimate probabilistic information
- Hard Classification:
  - Only attempt to estimate class membership



# Soft Classification

- Model attempts to estimate the probability that an observation belongs to a class.
  - For binary:  $P(y = 1 | X)$
  - Conditional probability
    - If we change data  $X$ , then estimate of probability changes
- Basic example is Logistic Regression

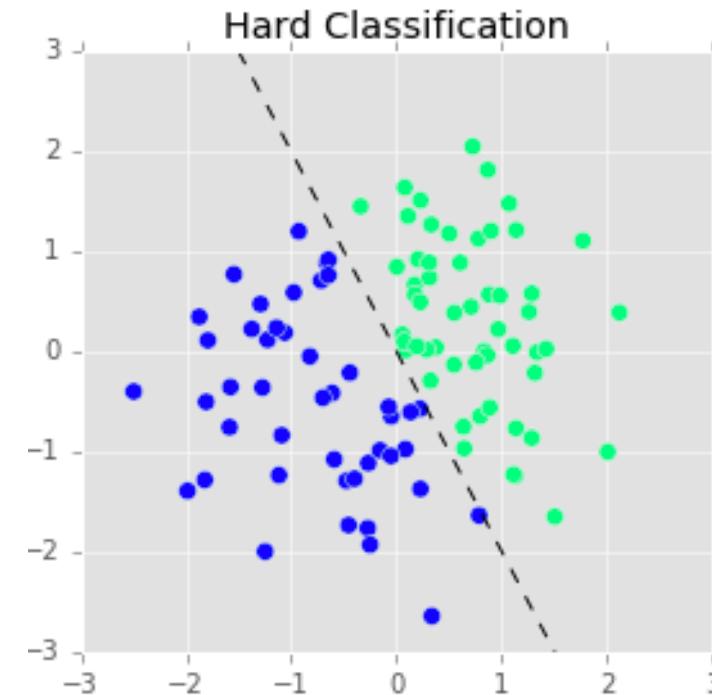
$$P(y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}}$$

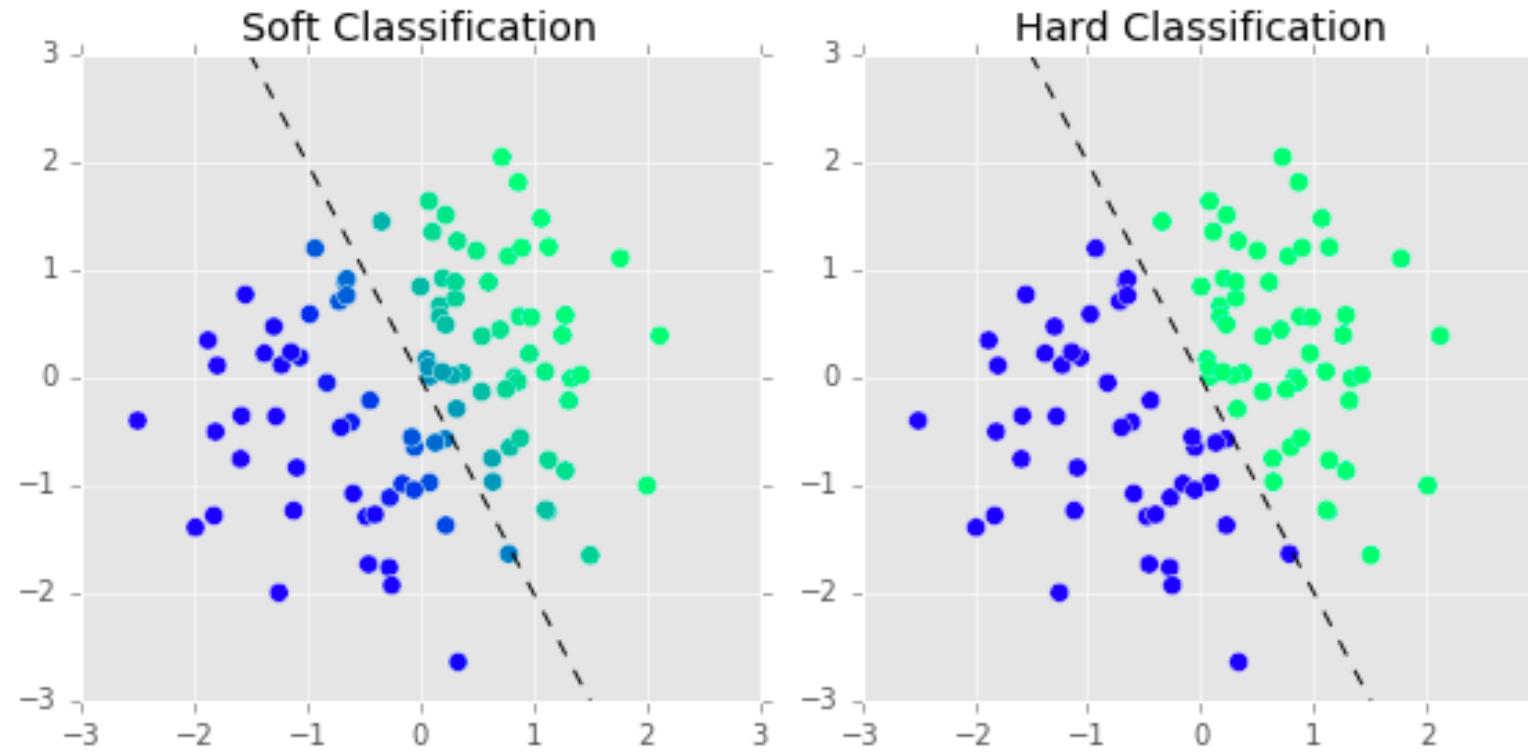


Note:  
 $0 < y < 1$

# Hard Classification

- Model only attempts to estimate class membership
- Predictions are **not** probabilities
- Predictions are either zeros or ones

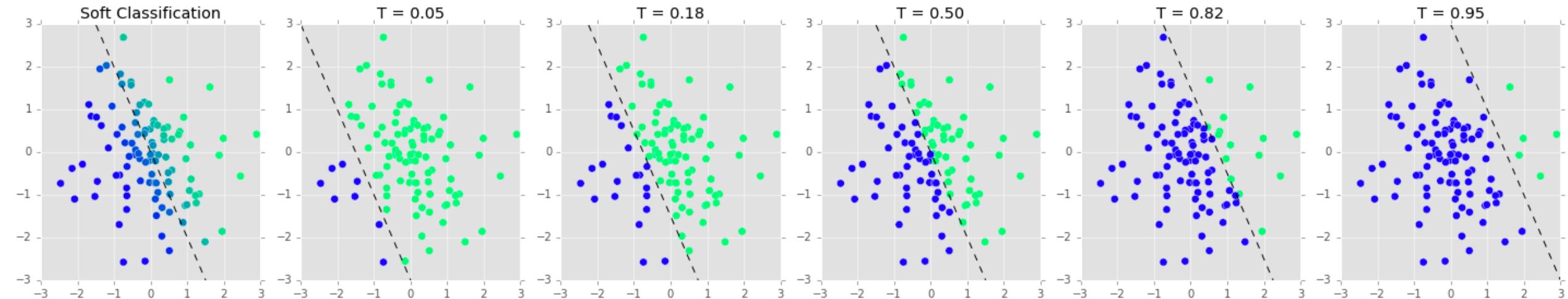




- Common misconception: **All** classification models are hard classifiers
  - **Not true!**
- There **are** some models strictly designed to be hard classifiers
  - For instance SVMs (Support Vector Machines)
- Often, soft classification is a better strategy!

# Relationship: Soft and Hard Classifiers

- You can always turn a soft classifier into a hard classifier!
  - For binary soft classifiers,  
this is called **Thresholding**
  - Pick a **cutoff** or **threshold T**
- If  $P(y = 1 | X) < T$  then classify  $\hat{y} = 0$
- If  $P(y = 1 | X) \geq T$  then classify  $\hat{y} = 1$



# Evaluating Classification Models: Two Basic Approaches

- Evaluate probabilities directly

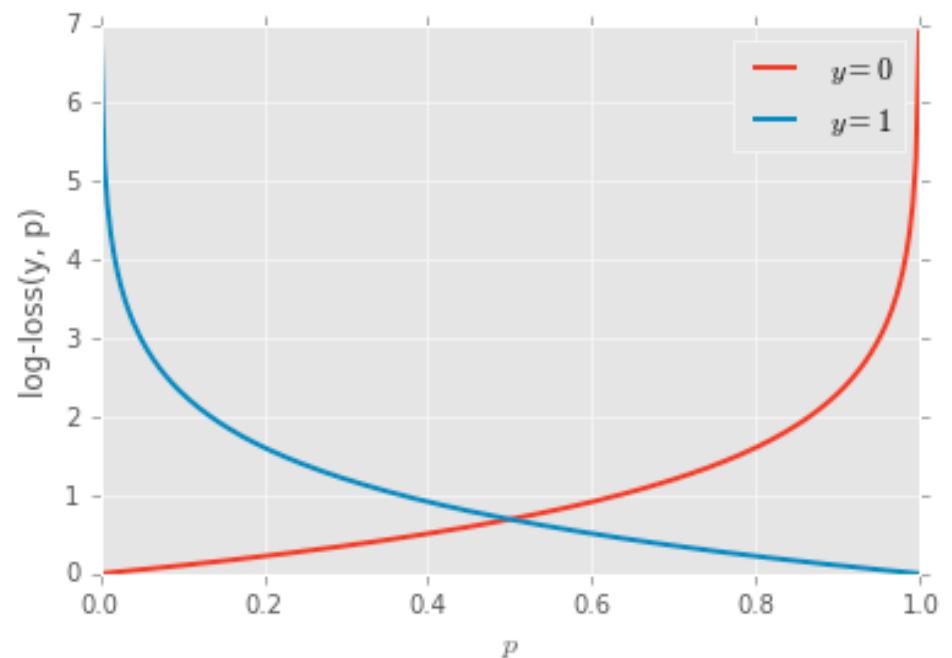
or

- Evaluate ordering of probabilities
  - i.e. something predicted as more probable than something else actually **is** more probable

# Evaluating Classification Models: Soft

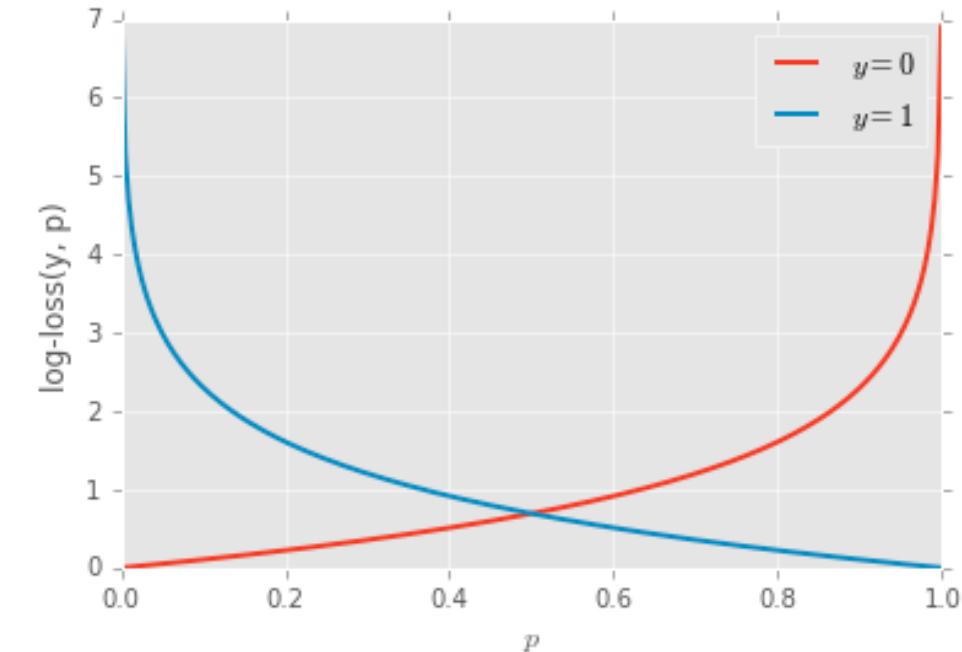
- Soft classification models are best evaluated and compared on the basis of their predicted probabilities (*not* by converting them to hard classifiers and evaluating the result)
- The gold standard metric for evaluating soft classification models is the log-loss

$$\text{log-loss}(y, p) = - \sum_i y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$



# Evaluating Classification Models: Soft

- Math reference:  $\log(0.2) = -1.389$  and  $\log(0.8) = -0.223$
- $\text{log-loss}(0, 0.2) = 0.223$  (confidently correct)
- $\text{log-loss}(0, 0.8) = 1.389$  (confidently wrong)
- $\text{log-loss}(1, 0.2) = 1.389$  (confidently wrong)
- $\text{log-loss}(1, 0.8) = 0.223$  (confidently correct)



$$\text{log-loss}(y, p) = - \sum_i y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

# Evaluating Classification Models: Hard

- When evaluating hard classifications, there are only four possibilities for each data point:
  - True Negative:** Actual 0, Predicted 0.
  - False Positive:** Actual 0, Predicted 1.
  - False Negative:** Actual 1, Predicted 0.
  - True Positive:** Actual 1, Predicted 1.
- Arranged in a grid → Confusion Matrix
  - Many hard classification evaluation metrics can be derived from this table

	<b>Predicted 1</b>	<b>Predicted 0</b>
<b>Actual 1</b>	True Positive	False Negative
<b>Actual 0</b>	False Positive	True Negative

# Accuracy

- The simplest and most natural thing to do is compute the proportion of predictions that we got correct
- Lots of problems here

$$\text{Accuracy} = \frac{\# \text{ True Positives} + \# \text{ True Negatives}}{\text{Total } \# \text{ of Data Points}}$$

or

$$\text{Accuracy} = 1 - \frac{\# \text{ False Positives} + \# \text{ False Negatives}}{\text{Total } \# \text{ of Data Points}}$$

# False Positive and True Positive Rate

- While the accuracy measures how our predictions perform across all of our dataset, the **false positive rate** and **false negative rate** focus on the positive and negative classes individually.

$$\text{FP Rate} = \frac{\# \text{ False Positives}}{\# \text{ False Positives} + \# \text{ True Negatives}}$$

$$\text{TN Rate} = 1 - \frac{\# \text{ False Positives}}{\# \text{ False Positives} + \# \text{ True Negatives}} = \frac{\# \text{ True Negatives}}{\# \text{ False Positives} + \# \text{ True Negatives}}$$

$$\text{TP Rate} = \frac{\# \text{ True Positives}}{\# \text{ False Negatives} + \# \text{ True Positives}}$$

# Grand Confusion Matrix

- Each is a ratio
- $(TP, TN, FP, FN) / \text{sum of 2 others}$
- Numerator given by label location
- Denominator given by locations spanned by arrow

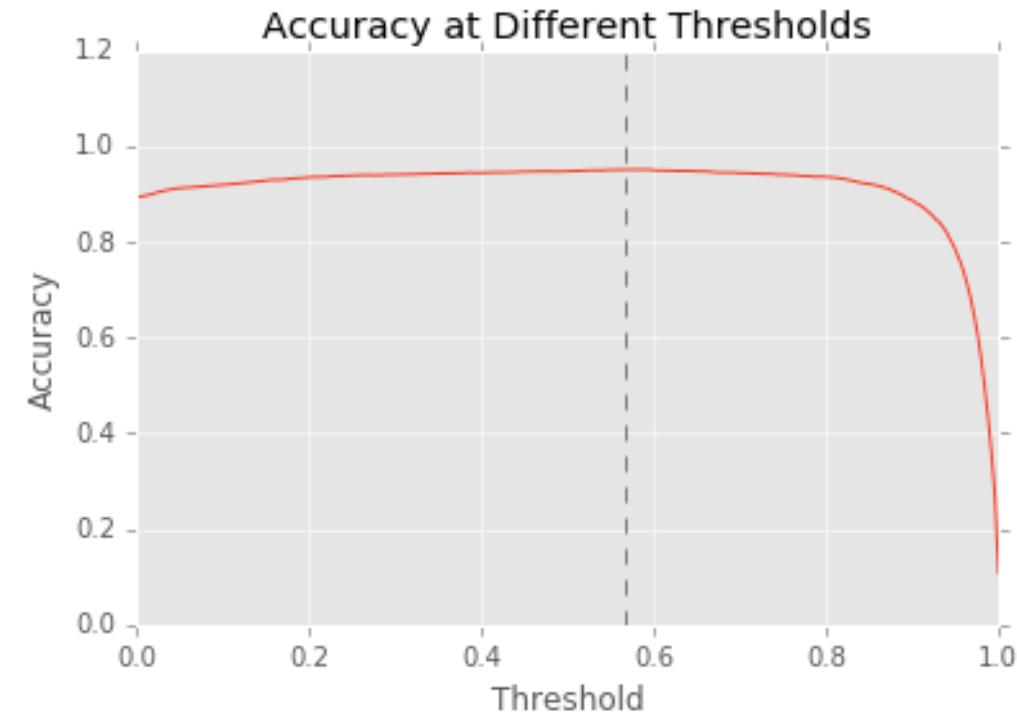
Confusion Matrix	Predicted Positive	Predicted Negative
Actual Positive	<b>True Positive</b> <div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid black; padding: 5px; width: 150px;"> <b>Precision</b> PPV             </div> <div style="border: 1px solid black; padding: 5px; width: 150px;"> <b>Sensitivity</b> TPR <b>Recall</b> <b>Power</b> </div> </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <div style="border: 1px solid black; padding: 5px; width: 150px;"> <b>FNR</b> <math>\beta</math> </div> <div style="border: 1px solid black; padding: 5px; width: 150px;"></div> </div>	<b>False Negative (Type II)</b> <div style="border: 1px solid black; padding: 5px; width: 150px;"> <b>NPV</b> </div>
Actual Negative	<b>False Positive (Type I)</b> <div style="border: 1px solid black; padding: 5px; width: 150px;"> <b>Specificity</b> TNR             </div>	<b>True Negative</b> <div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid black; padding: 5px; width: 150px;"> <b>FPR</b> Fall-out <math>\alpha</math> </div> <div style="border: 1px solid black; padding: 5px; width: 150px;"></div> </div>

# Hard Classification Metrics for Soft Classification Models

- Some metrics that can be used to evaluate both hard and soft classifiers.
  - Since a soft classifier can be made into a hard classifier by thresholding the predicted probabilities at a fixed level...
  - ...any metric designed for hard classification models can, after thresholding, be applied to a soft classification model.
- Interesting to see how each of these metrics behaves as we vary the threshold.
- The following slides have plots of some of these measures for a logistic regression on a testing data set.

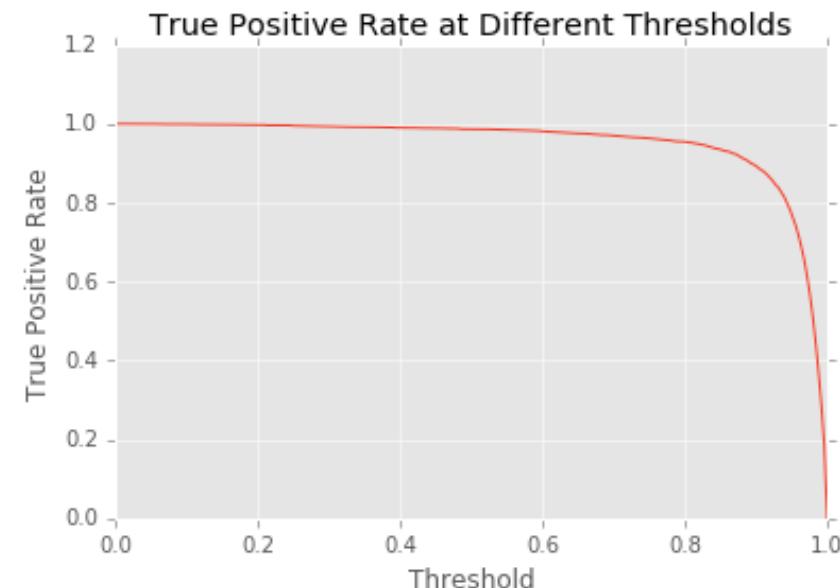
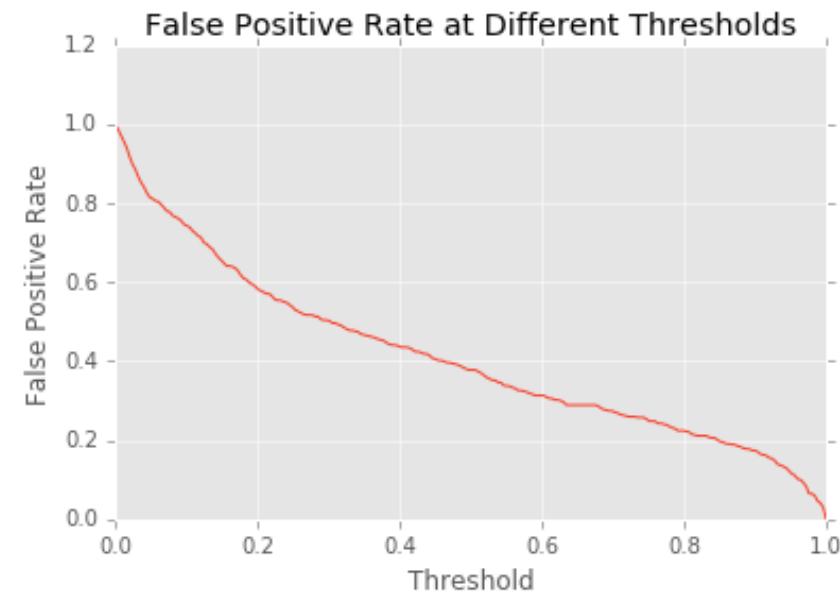
# Thresholds - Accuracy

- Tons of things to talk about here, but the main takeaway is that accuracy is rarely a good metric to evaluate hard classifiers
- Example: HIV test with >99% accuracy?
  - Easy!
  - Classify everyone as HIV-
  - For most populations, that's >99% accurate
- Can be a decent metric in rare cases
  - Balanced dataset
  - Scaling minority class to balance data?
    - Not worth it just to use accuracy



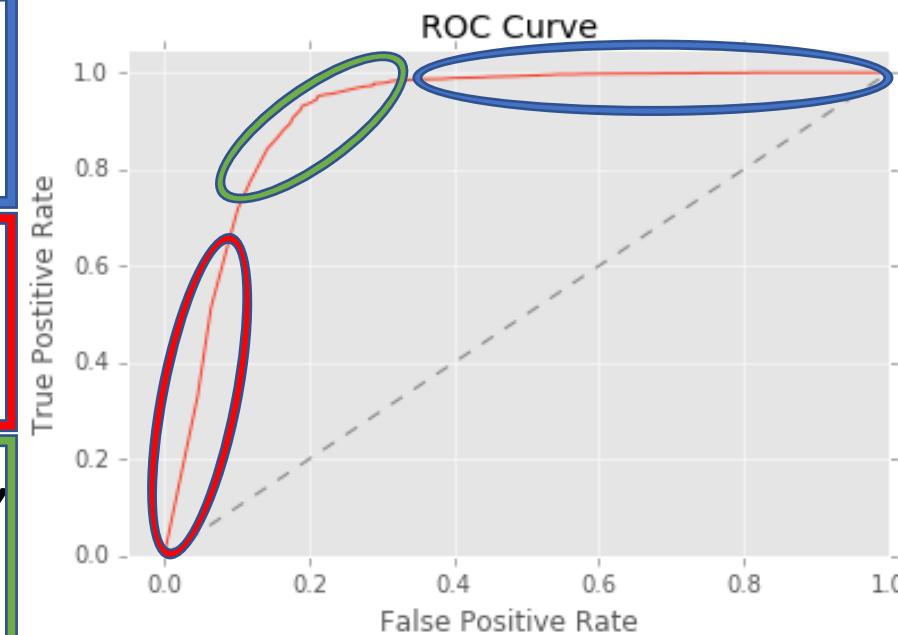
# Thresholds - False Positive and True Positive Rates

- Intuitively, as threshold increases
  - More conservative in predicting positive class
  - FP rate decreases
- At the extremes
  - A threshold of zero classifies everything as positive, so all our negative classes are falsely classified
  - A threshold of one classifies nothing as positive, so all the negative classes are correctly classified.
- We interpret behavior of TP rate curve similarly



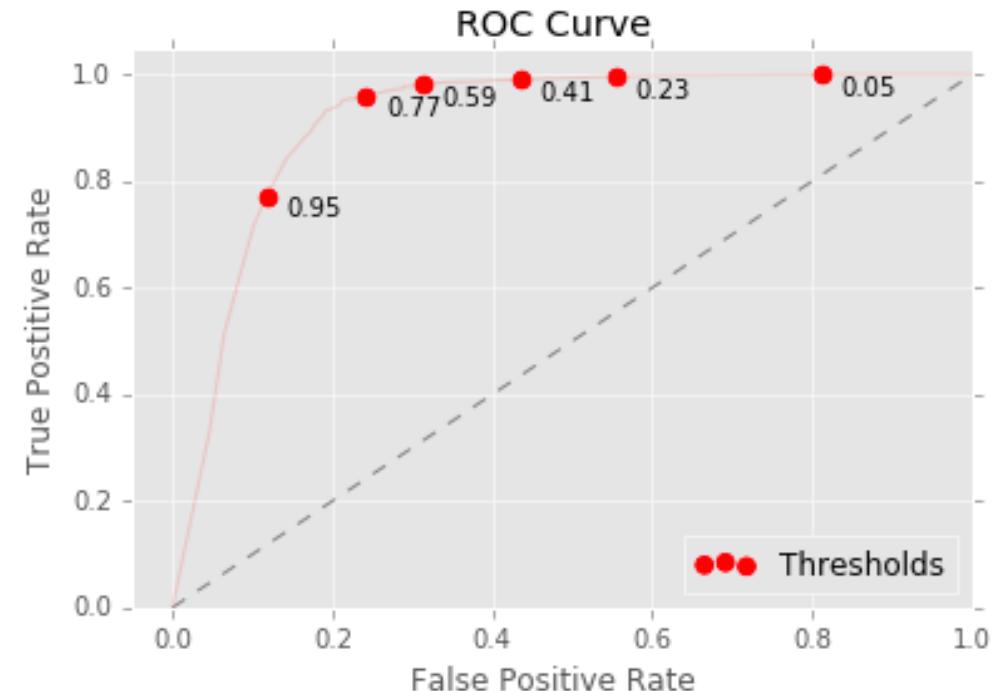
# ROC Curves

- A trade-off when setting the threshold of a soft classifier if a hard classification is needed
  - If the threshold is set too low, then we will be very liberal in what observations we classify as positive, which drives up the false positive rate of the model.
  - If the threshold is set too high, then we will be very conservative in what observations we classify as positive, which will drive down the true positive rate.
  - We suspect that in the middle there is some sweet spot, where we have balanced a trade-off between the false positive and true positive rates.
- An **ROC** (receiver operating characteristic) **curve** expresses this relationship in one picture.



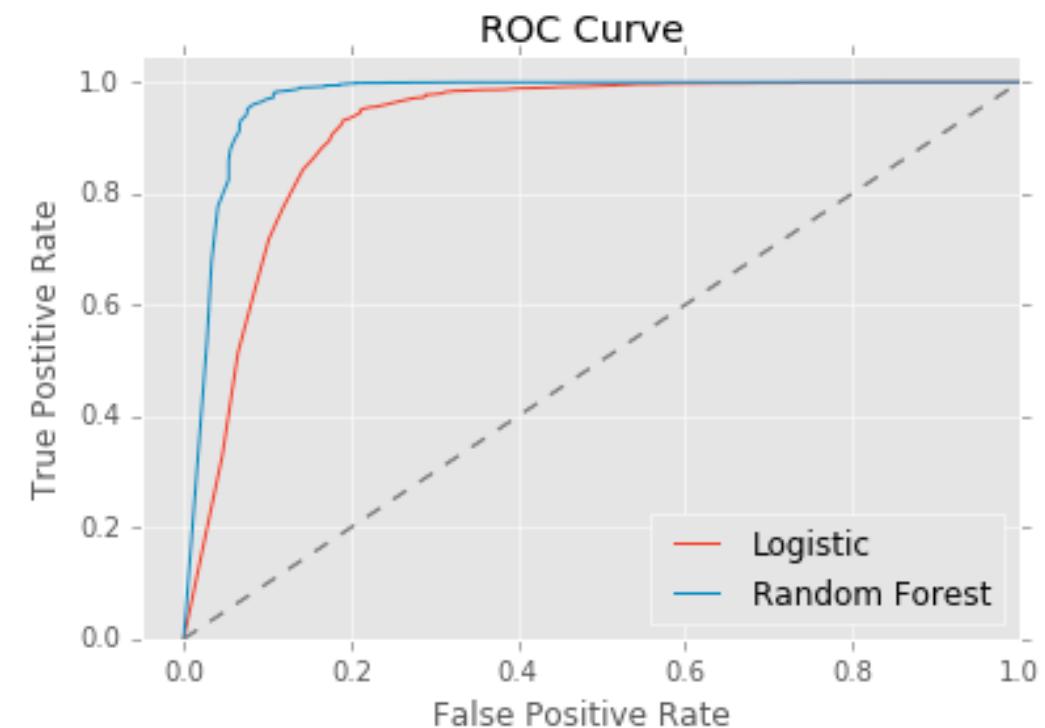
# ROC Curves

- An ROC curve visualizes the trade-offs between false positives and true positives as we vary the threshold on a soft classification model.
- Each choice of the threshold results in **one** point on the ROC curve
- Dashed line between (0, 0) and (1, 1)
  - Traditional to include
  - Represents random class assignment
  - Curve above line
    - Predictive power better than random guess



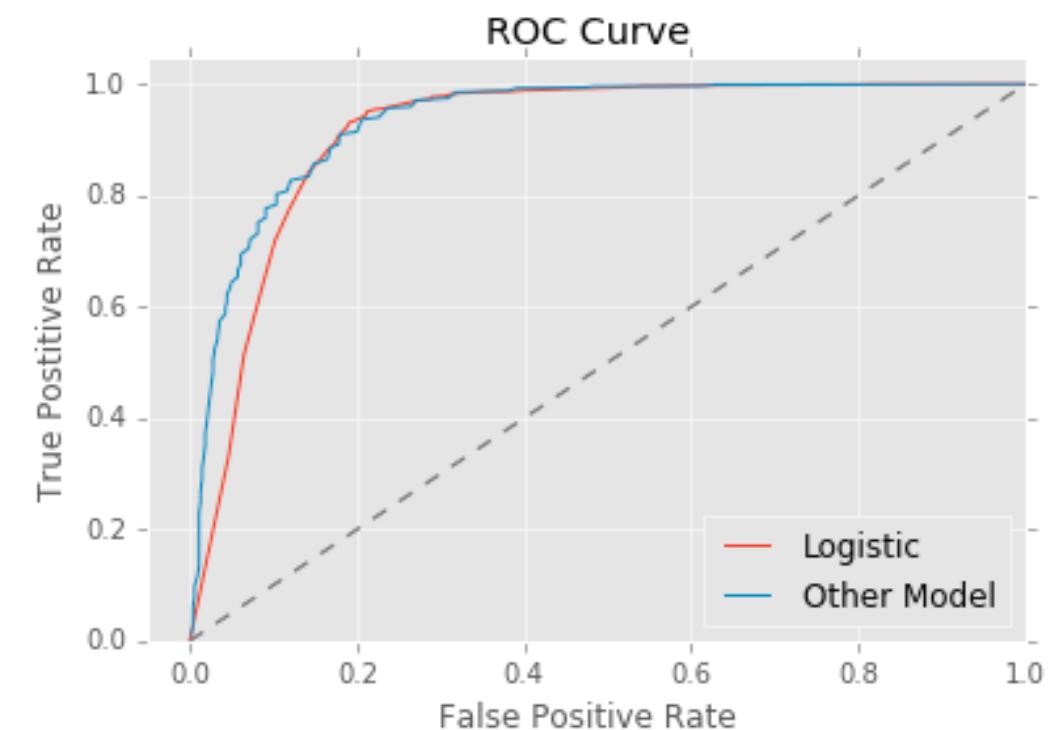
# ROC Curves

- Common to inspect ROC curves of competing classifiers
  - Get a feel for how models make tradeoffs for FP and FN rates
- Here, random forest does a better job of distinguishing the positive from negative classes.



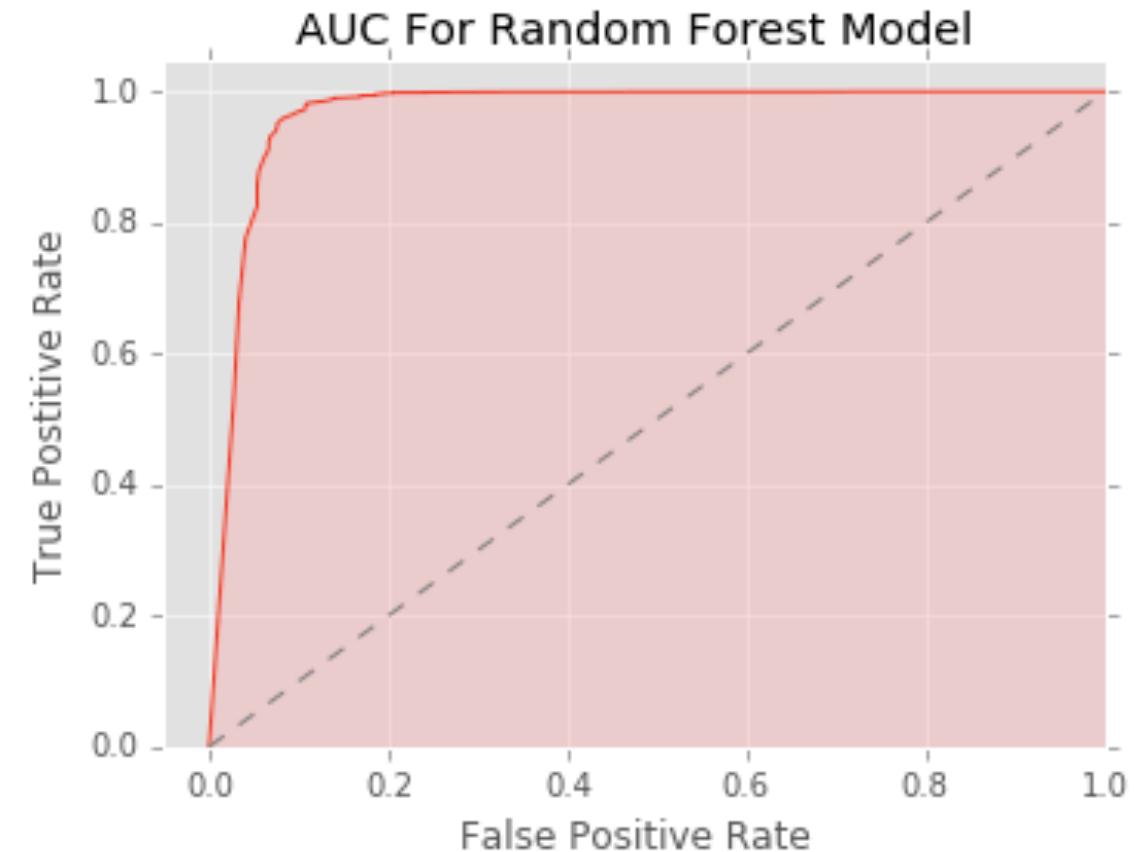
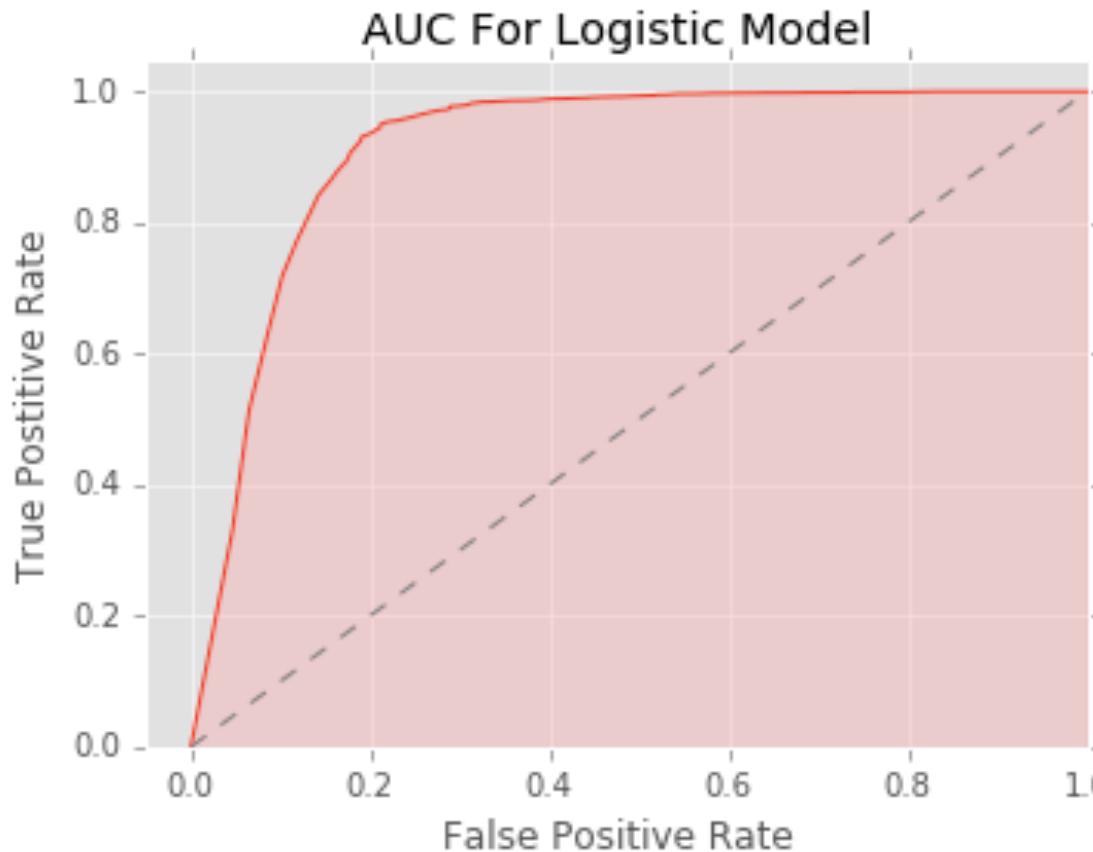
# ROC Curves

- In some cases, the ROC curves for the models being compared will cross
- ROC curves give us no statistical reason to prefer one model over another
  - It depends on the costs of false positive and false negative errors in our decision problem.
  - Profit curves useful in business questions, but each problem space will need its own approach



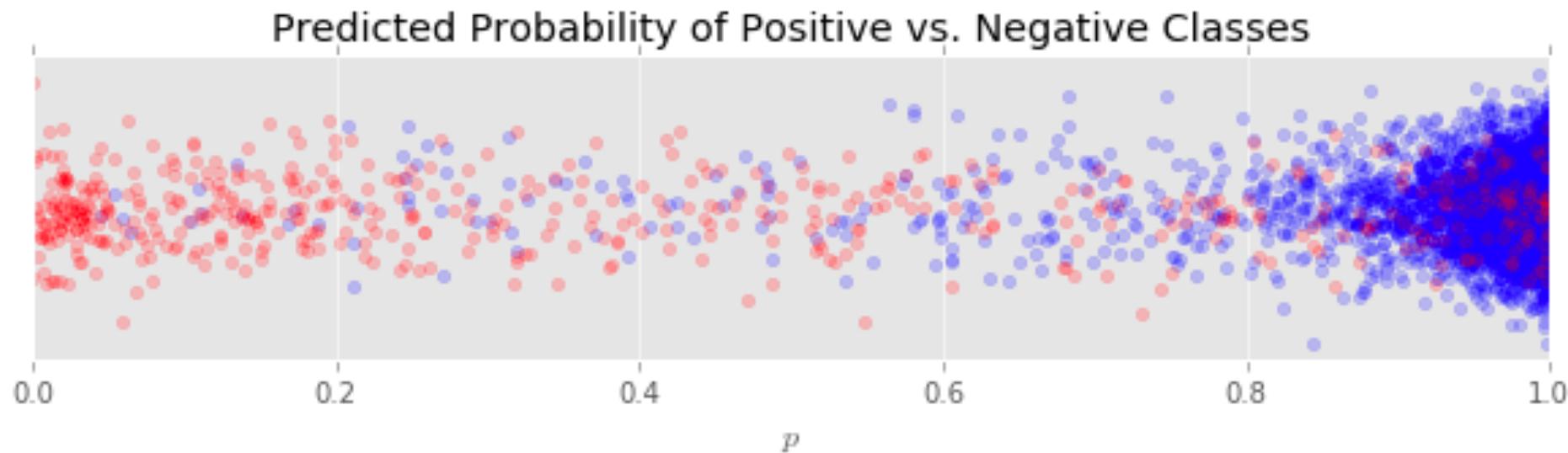
# The AUC

- The **area** under the ROC curve is a numeric measure of a model's overall ability to distinguish between positive and negative classes.
- AUC is a nice measure of model fit, as it averages over **all** thresholds.



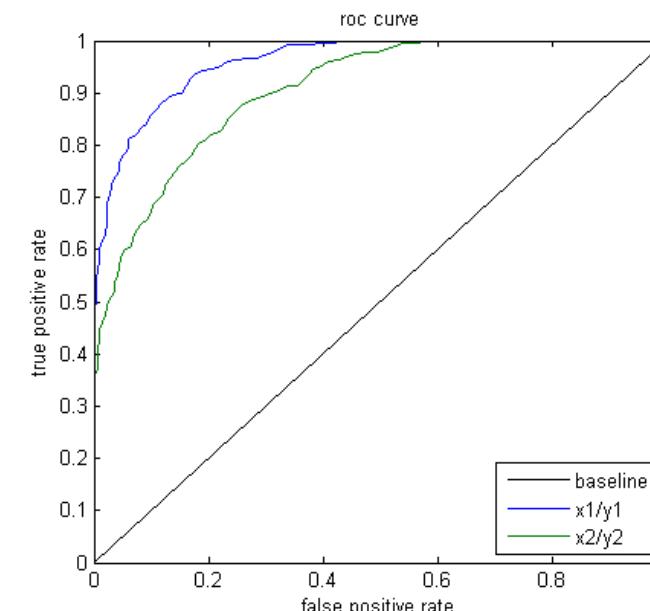
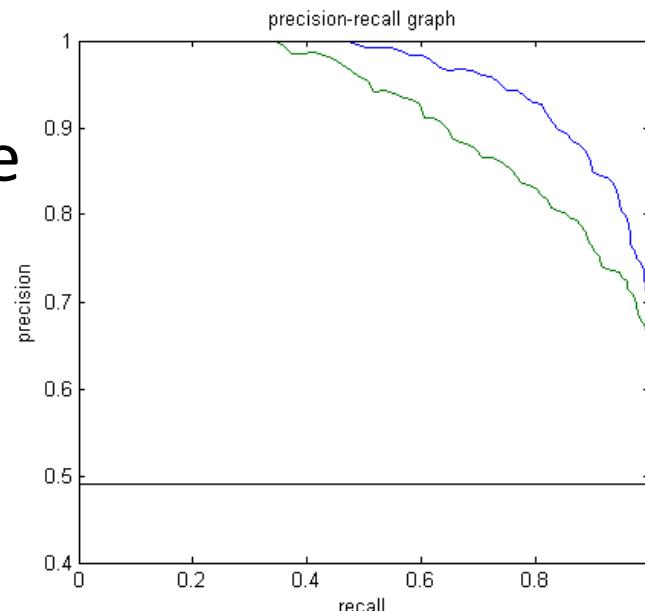
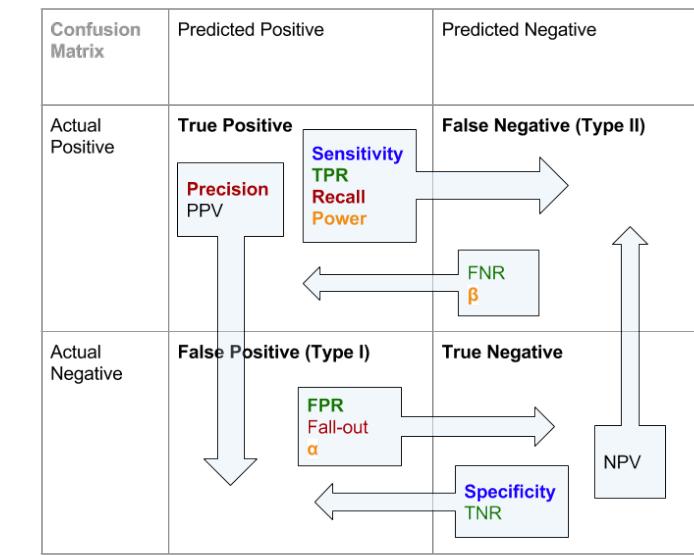
# Interpretation of AUC

- The AUC has a very nice probabilistic interpretation:
  - The probability of our model predictions *ranking* a **randomly chosen positive class** higher than a **random negative class**
- The AUC statistic measures **the chance that a random pair of (red, blue) points is ranked in the correct order**



# Precision-Recall Curve

- Another chart used to describe the trade-off at different thresholds
- Also parameterized by the threshold (like the ROC)
- But plots the precision (the ratio of true positives over predicted positives) against the recall (a.k.a. sensitivity a.k.a. TPR).



- Precision-recall curve
  - x is recall
  - y is precision

- ROC curve
  - x is FPR
  - y is TPR

# Conclusion:

## Discussion - Do You Really Need Hard Classification?

- The most persistent mistake in this area is the *overuse of hard classification models*. Let's be explicit: soft classification models are more flexible in practice, and they should generally be preferred.
- But many business or scientific problems **do** call for hard classification, especially when some decision must be made
  - e.g. we are going to do this or not, we are going to invest in this or that, etc.
- It is important to distinguish in our problem solving process between:
  - Estimating the probability of uncertain events.
  - Using those probabilities to make hard decisions or develop decision rules.
- The best general principle
  - Convert your probabilities into decisions as late into the process as possible.