



MACHINE LEARNING FOR BEGINNERS

SEP 07 2022

MATT ARTHUR

UCR GRADQUANT

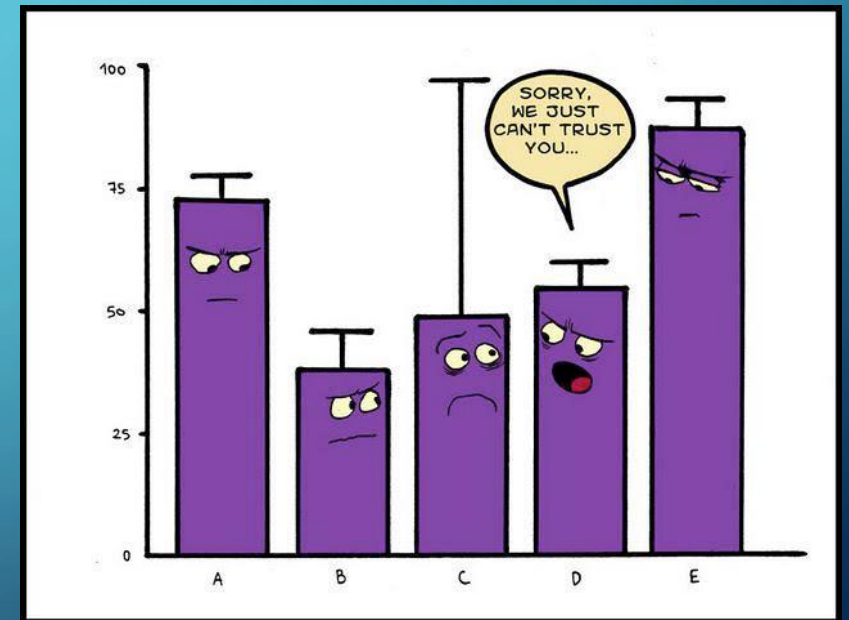
WORKSHOP OUTLINE

- Knowing the difference – AI, ML & Data Science.
- Prepping your data
 - Scaling & Normalization, Visualizing your dataset, Handling missing values.
 - Feature generation & selection, Dimensionality reduction, Class Imbalance, Train-Test Split
- Measures of Performance
 - Overfitting, Bias & Variance, The Confusion Matrix, Precision, Recall & AUC.
- Choosing the Right Algorithm
 - Simple Classifiers, Ensemble methods, Cost-sensitive classifiers.
 - Troubleshooting.

WHAT'S THE DIFFERENCE? – AI/ ML/ DATA SCIENCE

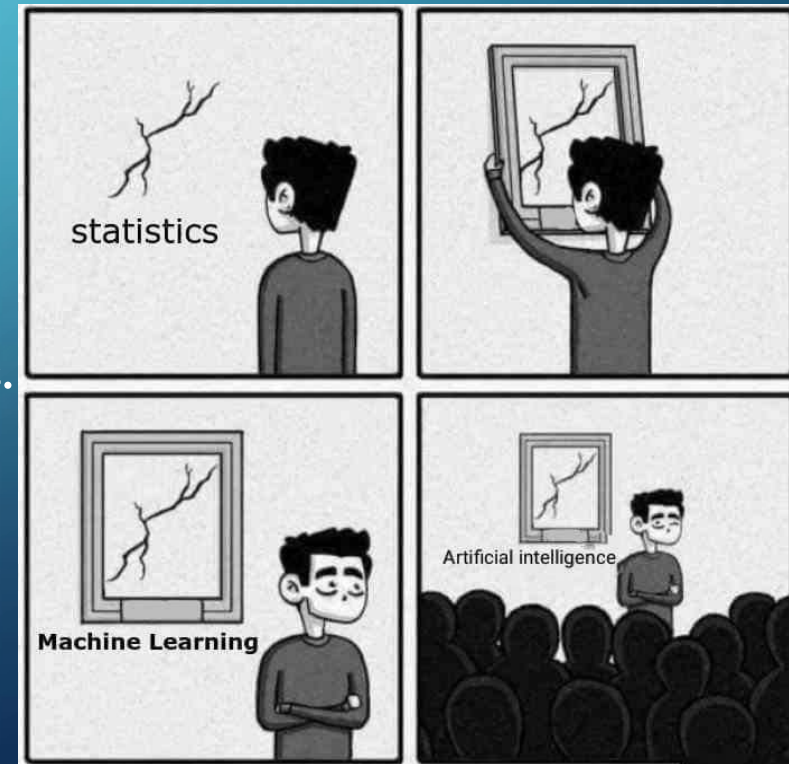
- Data Science

- Draw inferences from data.
- Coding skills needed to manage large datasets.
- Example: Business Analysis, Statistical Inference



WHAT'S THE DIFFERENCE? – AI/ ML/ DATA SCIENCE

- Machine Learning
 - Algorithms that draw inferences from data.
 - Strong reliance on statistical models.
 - Examples: Spam & Fraud detection.
- AI – Artificial Intelligence (Modern version)
 - Make decisions based on data and machine learning algorithms.



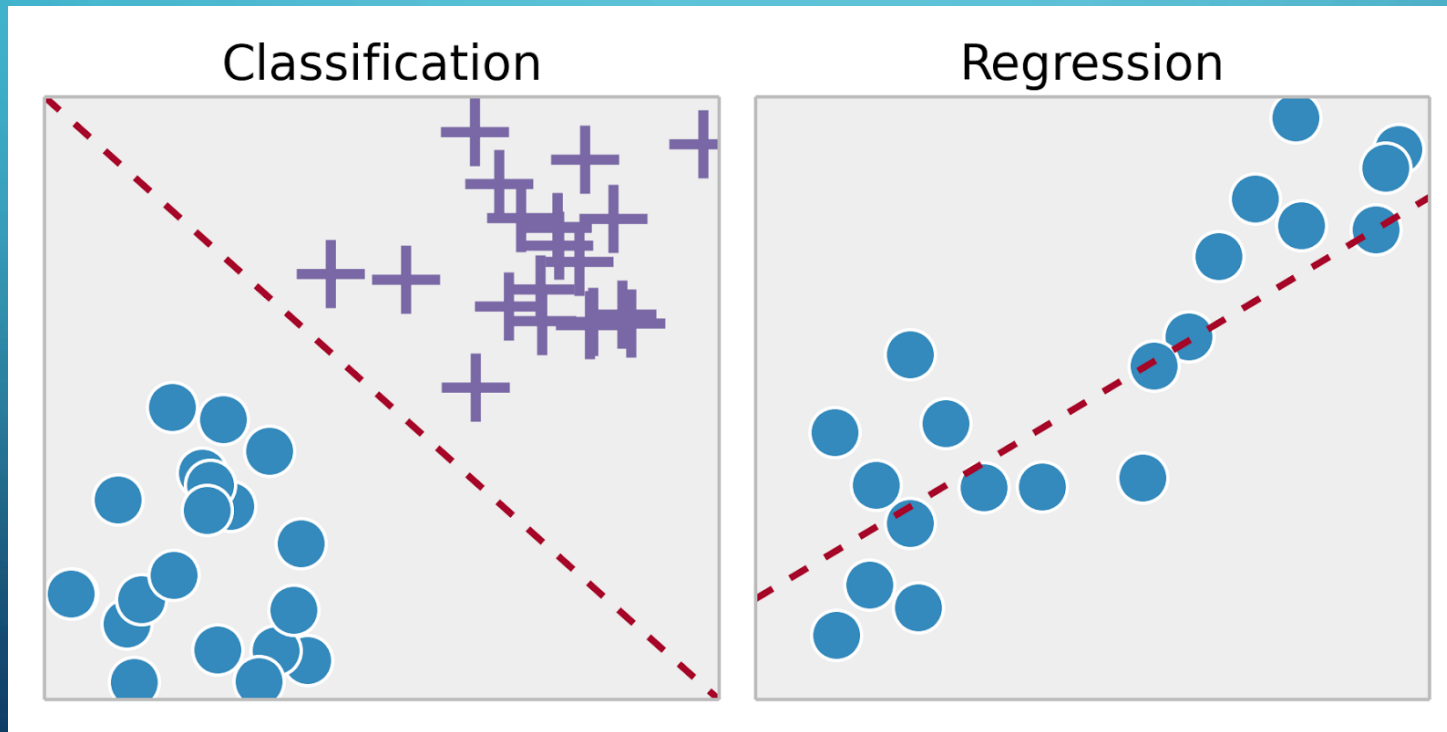
WORKSHOP GOALS

- A barebones introduction to machine learning with R.
 - Alternatives: Python – sklearn/keras, MATLAB, Weka.
- By the end of this workshop you will be able to:
 - Prepare suitable datasets.
 - Understand the choices between algorithms.
 - Evaluate the performance of your models.
 - Assess the limits of what machine-learning can do.



EXERCISE 1 – PREDICT INCOME FROM CENSUS DATA

- Supervised Learning Problem
 - Given sample outputs and for a set of input data, estimate a “fit”.



PREPPING YOUR DATA

- A note on data organization:
 - Keep track of data manipulations in a separate notepad file.
 - After each manipulation, save file as “L#” starting at L1.

TO DO:

1. Download the census dataset
2. Add headers to file and save as “raw_data_L0.xlsx”

raw_data_L1.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
2	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
3	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spou	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-cleane	Not-in-family	White	Male	0	0	40	United-States	<=50K
5	53	Private	234721	11th	7	Married-civ-spou	Handlers-cleane	Husband	Black	Male	0	0	40	United-States	<=50K
6	28	Private	338409	Bachelors	13	Married-civ-spou	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

PREPPING YOUR DATA – DATA TYPES

- 3 broad categories of Data:
 - Record based.
 - Examples: Census data, emails.
 - Time series data.
 - Examples: Stock market, sales forecasts.
 - Graphical data.



PREPPING YOUR DATA – ATTRIBUTE TYPES

- Categorical / Nominal: Raw Text. → "Apple, Banana, Orange"
- Binary: Can have only 2 states. → (1,0) (Male,Female) (+,-)
- Ordinal: A ranked collection. → $C < C+ < B < B+ < A < A+$
- Numeric
 - Discreet: Possible values can be counted. → Flavors at Baskin Robins (32)
 - Continuous: Infinite number of possible values. → Time
 - Interval Scaled: Measured on a scale (Difference).
 - Ratio Scaled: Measured as a fraction (Multiples).

PREPPING YOUR DATA – ATTRIBUTE TYPES

Discrete Num. Ratio

Cont. Num. Ratio

Cont. Num. Interval

Cont. Text

Discrete Num. Ratio

Discret Text

Cont. Num. Ratio

Binary Text

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	age	workclass	fnlwtg	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
2	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
3	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spou	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-cleane	Not-in-family	White	Male	0	0	40	United-States	<=50K
5	53	Private	234721	11th	7	Married-civ-spou	Handlers-cleane	Husband	Black	Male	0	0	40	United-States	<=50K
6	28	Private	338409	Bachelors	13	Married-civ-spou	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

Discrete Text

Cont. Text

Discrete Text

Discrete Text

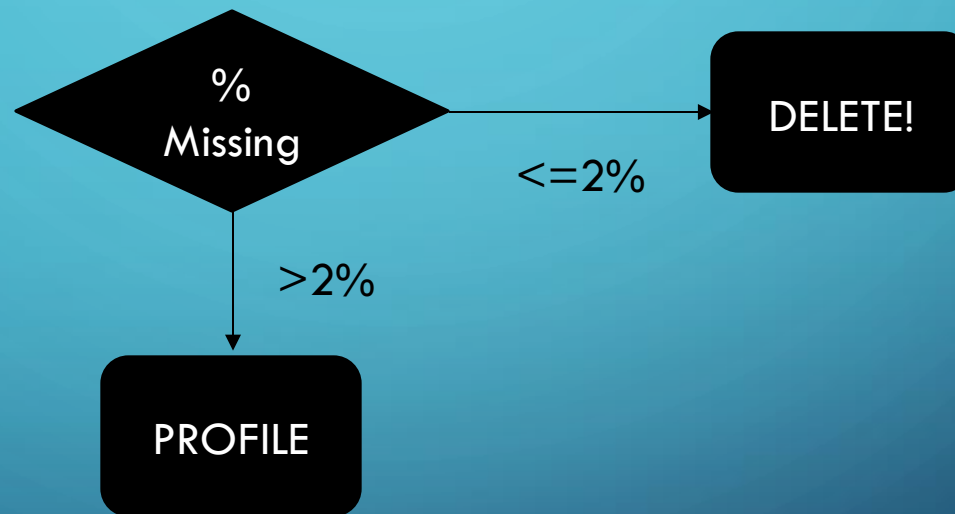
Binary Text

Cont. Num. Ratio

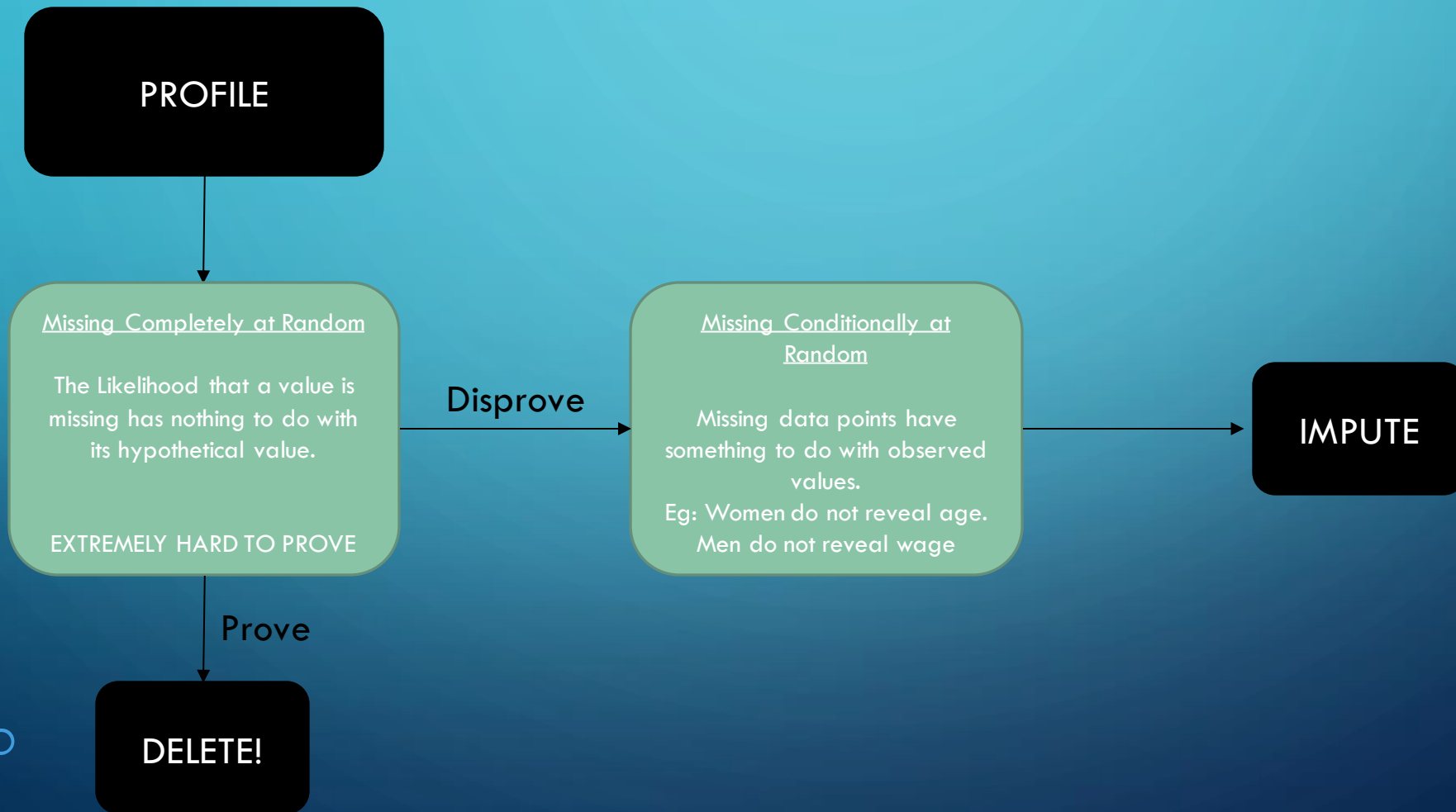
Discrete Text

PREPPING YOUR DATA – MISSING DATA

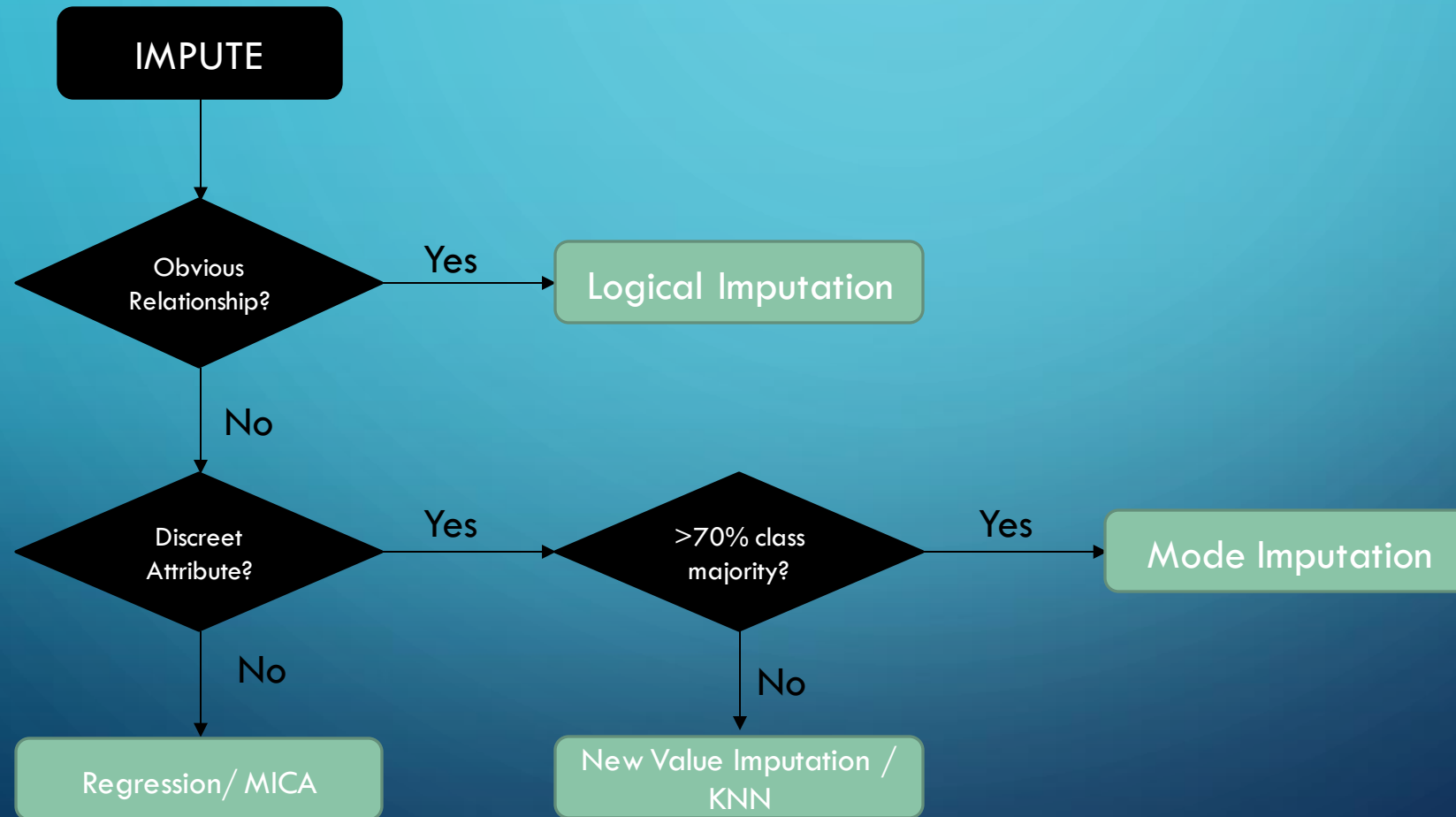
- Missing data is commonly encountered in large datasets.



PREPPING YOUR DATA – MISSING DATA



PREPPING YOUR DATA – IMPUTING MISSING DATA



PREPPING YOUR DATA - VISUALIZATION

TO DO:

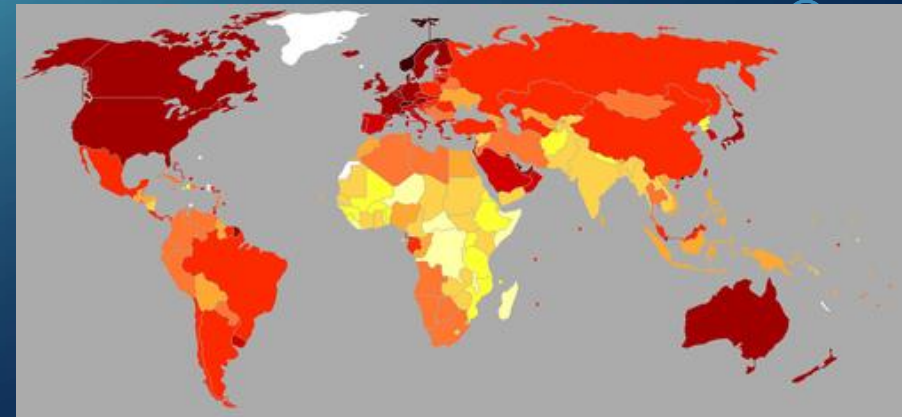
1. Start R/RStudio.
2. Upload your Data File, then invoke "read.csv"
3. `summary(census)`

PREPPING YOUR DATA – FEATURE SELECTION

- Curse of Dimensionality:
 - Too many features ruin algorithm performance.
 - Similar to how too much information clouds your judgement.
- Select features that have an ‘impact’ on your prediction.
 - Hypothesis testing.
 - Common-Sense (People forget this more often than you think)

PREPPING YOUR DATA- FEATURE SELECTION

- Common Sense:
 - 'Education' and 'Education-num' both mean the same thing.
 - 'Education' be blocked into larger subsets.
 - 'Fnlwgt' is assigned based on population projections.
 - Has nothing to do with income.
 - 'Country' has a large number of unique values, mostly US
 - They can be clubbed together as 'Region'.
 - This lets smaller populations have a say in predictions.

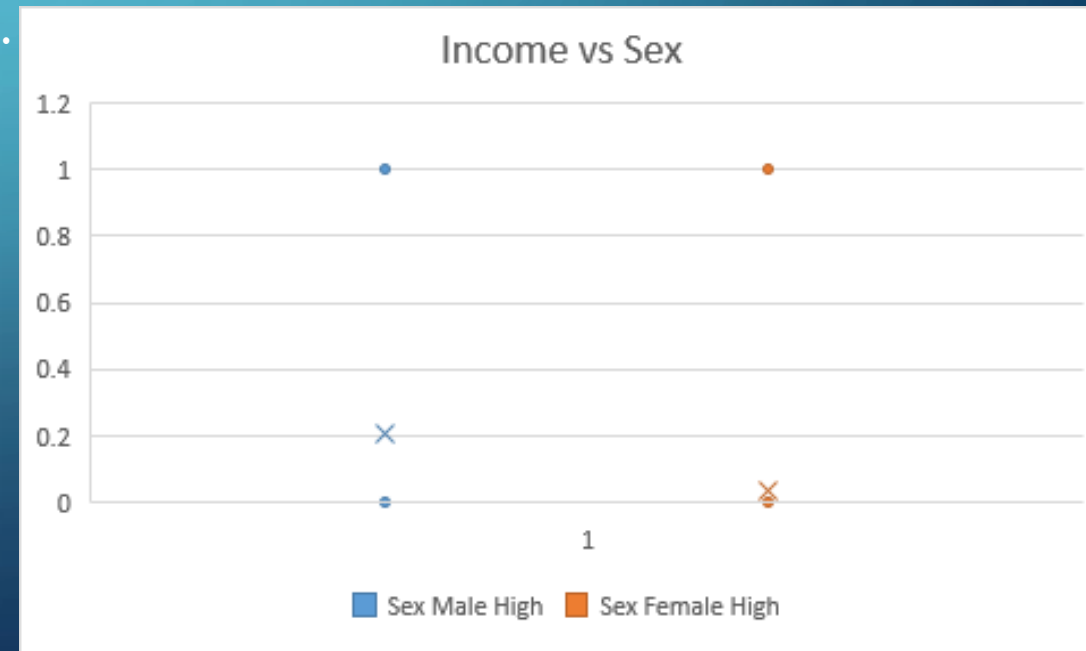


PREPPING YOUR DATA – FEATURE SELECTION

- Common Sense
 - We know that ‘economic sector’ has an impact on income.
 - This can be inferred from the ‘occupation class’.
 - Similar to countries, we can club occupations into ‘Vocational’, ‘Professional’, ‘Service’.
 - Education-num attribute can further be categorized.

PREPPING YOUR DATA – FEATURE SELECTION

- Hypothesis Testing.
 - The Dirty Way: Use a Whisker chart
 - The Professional Way:
 - Perform Pearson chi-squared test on your features.
 - Do a correlation analysis/regression
 - Use stepwise feature evaluation.



PREPPING YOUR DATA – FEATURE SELECTION

- Pearson Chi-Squared Test.
 - Extremely easy to implement.

H0	SEX HAS NOTHING TO WITH INCOME			H0	WORKCLASS HAS NO CONNECTION TO INCOME		
H1	SEX HAS SOMETHING TO DO WITH INCOME			H1	WORKCLASS HAS SOMETHING TO DO WITH INCOME		
1 DOF				1 DOF			
OBSERVED VALUES				OBSERVED VALUES			
	LOW	HIGH			LOW	HIGH	
MALE	14837	6533	21370	PRIVATE	19032	5063	24095
FEMALE	9446	1162	10608	NON PRIVATE	5251	2632	7883
	24283	7695			24283	7695	
CHI SQ	1492.928		$\chi^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(c + d)(b + d)(a + d)}$	CHI SQ	497.8454		
P	0			P	0		

```

=== Attribute Selection on all input data ===

Search Method:
    Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 13 income):
    Chi-squared Ranking Filter

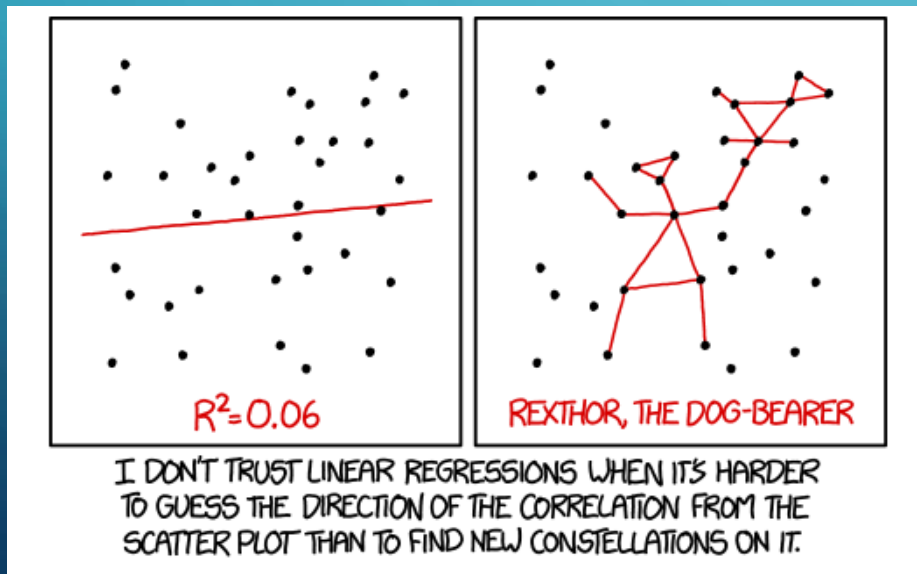
Ranked attributes:
6588.2502  5 relationship
6407.36798 4 marital-status
5443.78475 8 capital-gain
3718.72077 3 Education
3378.69533 1 age
3182.52466 12 Economic Sector
2498.31617 10 hours-per-week
2344.25122 9 capital-loss
1492.92841 7 sex
 549.68448 2 Workclass
 323.00353 6 race
 242.64895 11 Region

Selected attributes: 5,4,8,3,1,12,10,9,7,2,6,11 : 12
    
```

Inconclusive

PREPPING YOUR DATA – FEATURE SELECTION

- Pearson's Correlation Coefficient.
 - Test the linear relationship between 2 variables.
 - CAREFUL! DO NOT USE when linear relationship not apparent. Correlation = 0



```
=== Attribute Selection on all input data ===  
  
Search Method:  
  Attribute ranking.  
  
Attribute Evaluator (supervised, Class (nominal): 13 income):  
  Correlation Ranking Filter  
Ranked attributes:  
0.3312  4 marital-status  
0.2698  5 relationship  
0.2336  1 age  
0.2304 10 hours-per-week  
0.2227  8 capital-gain  
0.2161  7 sex  
0.1947 12 Economic Sector  
0.1494  9 capital-loss  
0.1291  3 Education  
0.114   2 Workclass  
0.0828  6 race  
0.0432 11 Region  
Selected attributes: 4,5,1,10,8,7,12,9,3,2,6,11 : 12
```

PREPPING YOUR DATA – SCALING

- Why & When to Scale.
 - Some algorithms rely on Euclidean distance.
 - For features that vary in scale, higher magnitudes are given more preference.
 - Certain Preprocessing steps will require scaled data. Example: PCA
 - Good Habit: **ALWAYS SCALE** your dataset whether your algorithm requires it or not.
 - This lets you switch between algorithms in the same data-pipeline.

PREPPING YOUR DATA – SCALING

- Scaling Methods:

- Standardization:

$$x' = \frac{x - \bar{x}}{\sigma}$$

X now has $\sigma = 1$
& $\mu = 0$

- Mean Normalization:

$$x' = \frac{x - \bar{x}}{\max - \min}$$

$-1 < X < 1$
& $\mu = 0$

- Min-Max Scaling:

$$x' = \frac{x - \min}{\max - \min}$$

$0 < X < 1$

- Unit Vector:

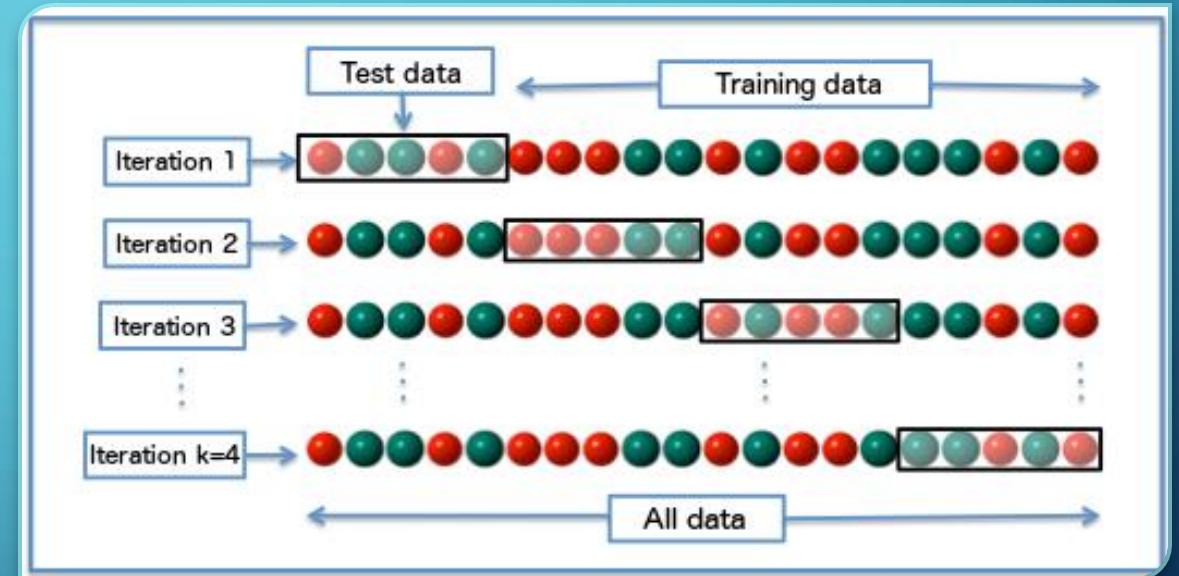
$$x' = \frac{x}{\|x\|}$$

PREPPING YOUR DATA - SCALING

- Min-Max scaling & Unit Vector scaling
 - Works when features have hard boundaries.
- Mean Normalization
 - Great for dimensionality reduction through PCA.
 - Can improve performance of neural networks.
- Can use the `scale()` function to scale numeric data in R

YOUR FIRST MACHINE LEARNING MODEL

- Test-Train Split.
 - Train the model on 1 part of data, test the model on the other part.
 - Suffers from sampling issues.
- Cross-Validate
 - Remove a small slice of data.
 - Train model on rest of the slice, test on the slice.
 - Repeat for all slices.

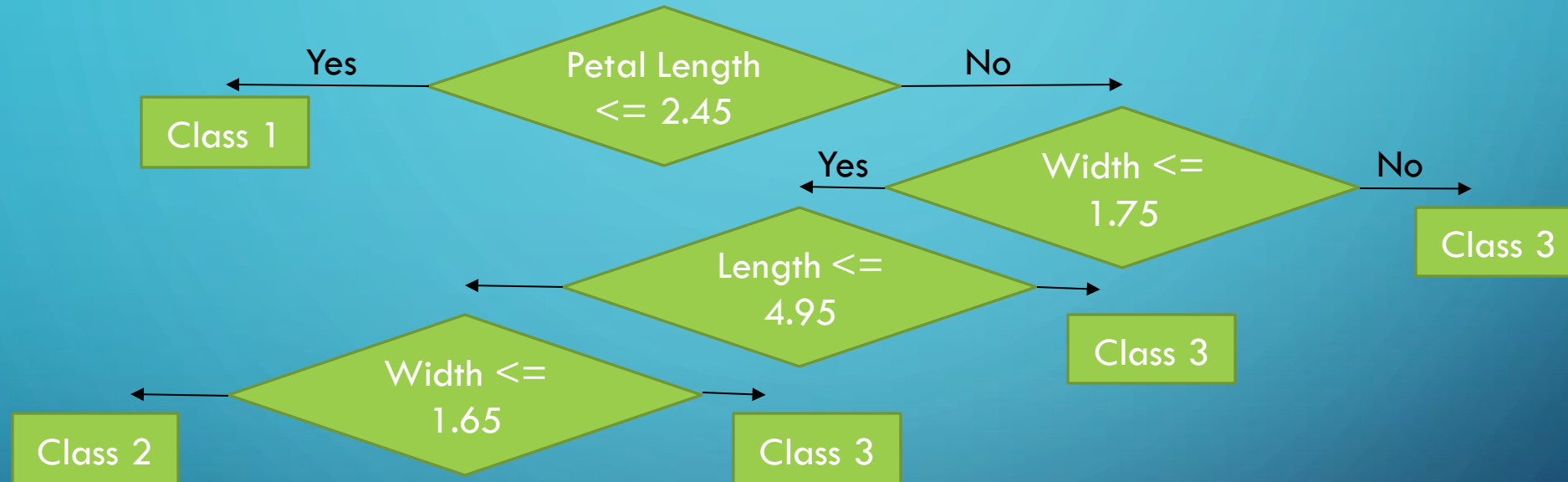


CLASSIFICATION TREES (CART)

- Entropy based decision trees.
 - The go-to algorithm for most classification needs.
 - Tends to overfit data. Beware!



VISUALIZING CART – FISHER'S IRIS DATA



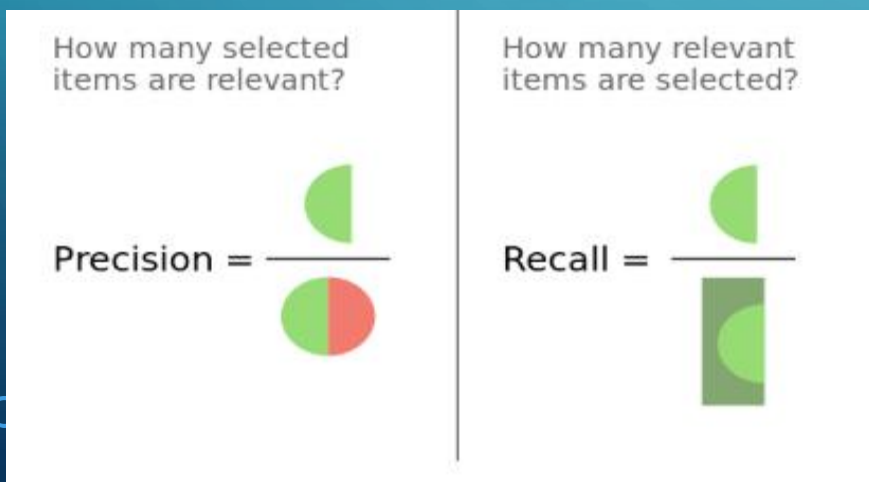
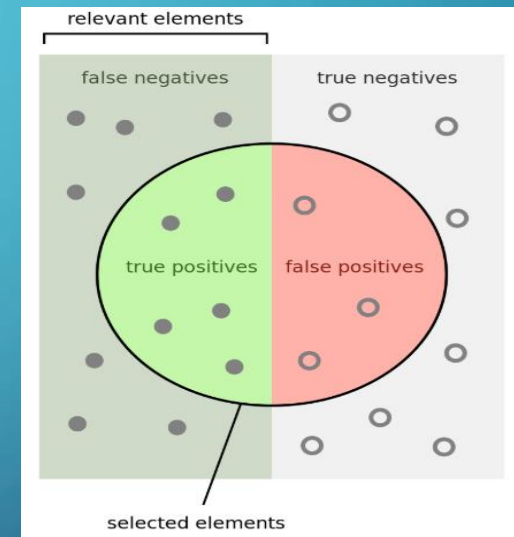
For any dataset, need to determine (1) which variables to use in classification; (2) which thresholds to use; and (3) when to stop splitting

PERFORMANCE – THE CONFUSION MATRIX

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

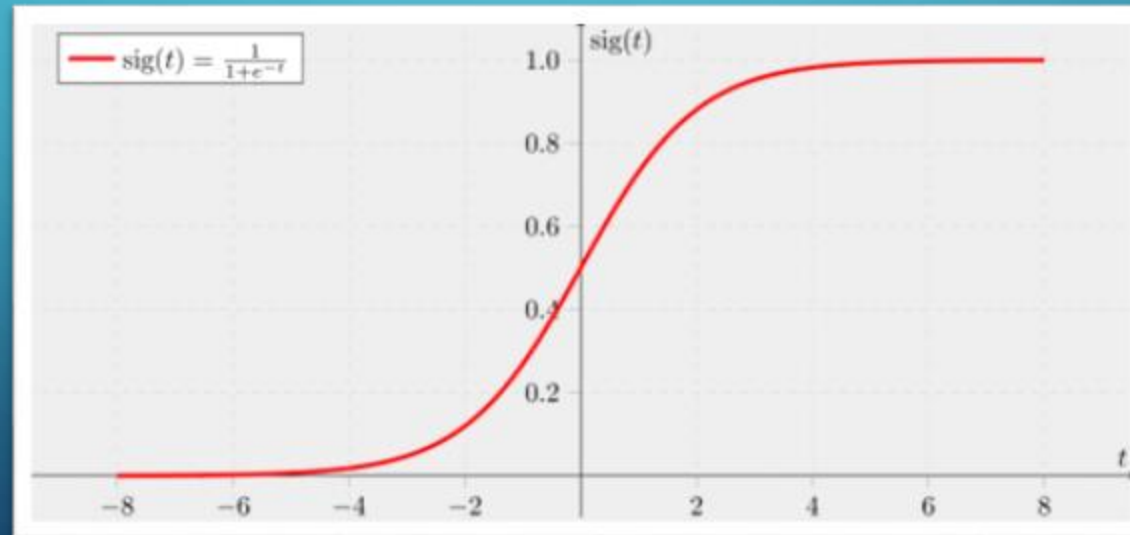
This metric applies to classifiers only.

- $Precision = \frac{TP}{TP+FP}$
- $Recall \text{ or } TPR = \frac{TP}{TP+FN}$
- $Selectivity = \frac{TN}{TN+FP}$
- $Fscore = \frac{2.TP}{2.TP+FP+FN}$
- $FPR = \frac{FP}{TN+FP}$



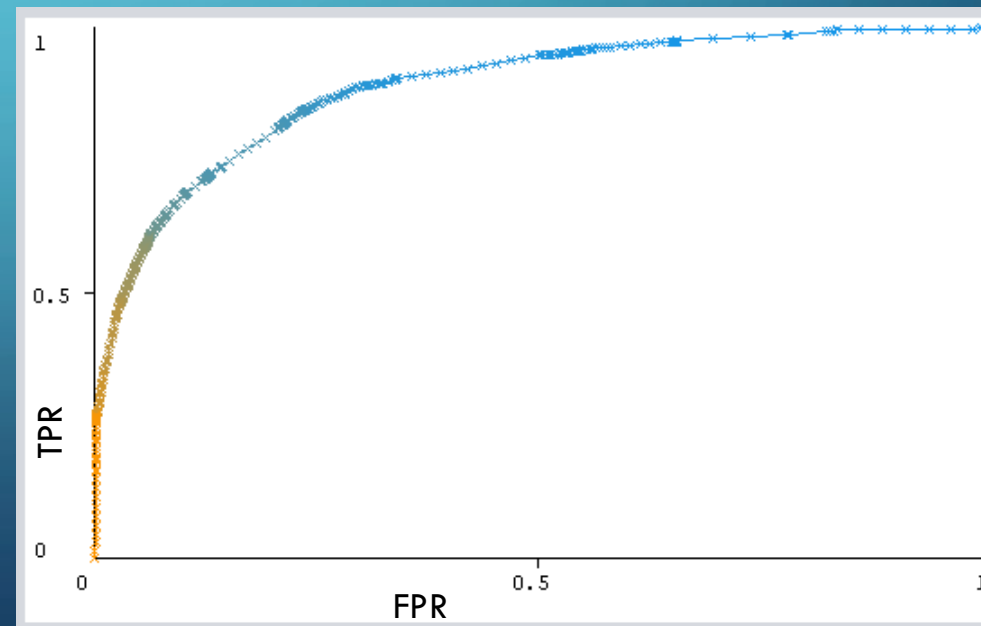
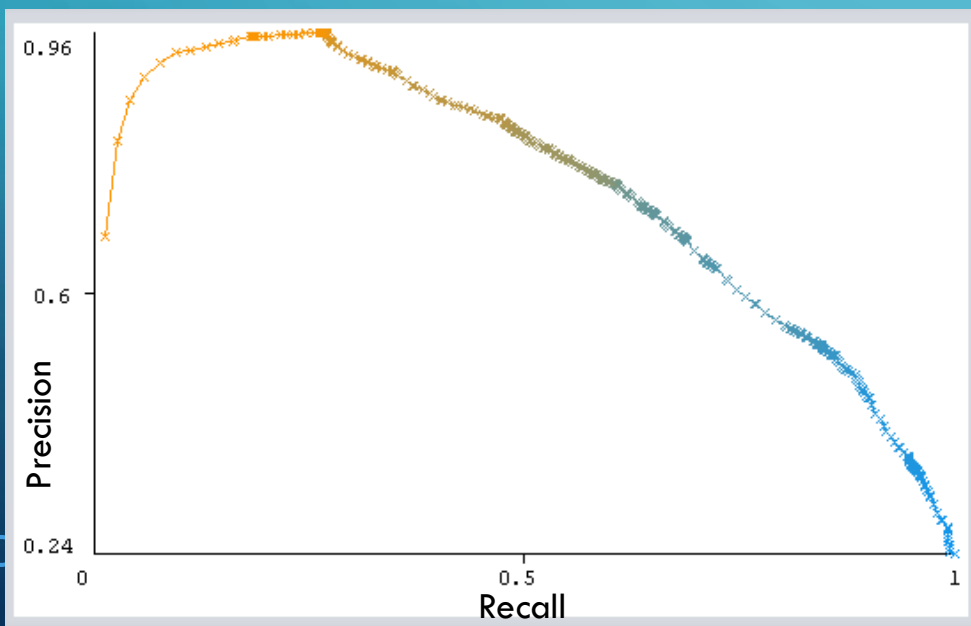
LOGISTIC REGRESSION

- Uses the logistic curve to make predictions.
 - Classification via Regression -> One of the most popular methods.
 - Has been used in the social sciences since early 20th century.



PERFORMANCE – PRC & ROC CURVES

- Classifiers assign a probability to a test sample.
 - The test sample is then assigned a 'class' based on threshold.
 - Varying the threshold changes Precision and Recall.



PERFORMANCE – ROC CURVE

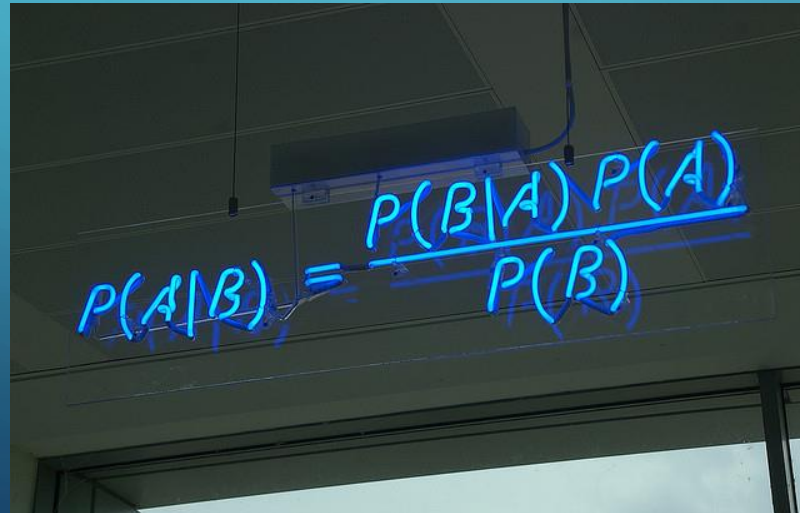
- The Area under the ROC curve (AUC) is a measure of separability.
 - Ranges from 0 to 1.
 - Good models have AUC close to 1.
 - An AUC of 0.5 ($y=x$) means the model cannot tell apart the classes.

PERFORMANCE – PRC CURVE

- PRC curves show the tradeoff between recall and positive predictive value.
 - Especially useful when there is an imbalanced dataset.
 - $y=1-x$ line represents no real gain from the model.
 - AUPRC=1 represents the perfect model.
 - CHOSE ONE! AUPRC OR ROC

THE NAÏVE BAYES CLASSIFIER.

- Classifier builds models based on Bayes Rule.
 - Its 'Naïve' because of the class independence assumption.
 - Also known as the Maximum A Posterior Estimator.
 - Surprisingly resilient to the inter-dependent features.



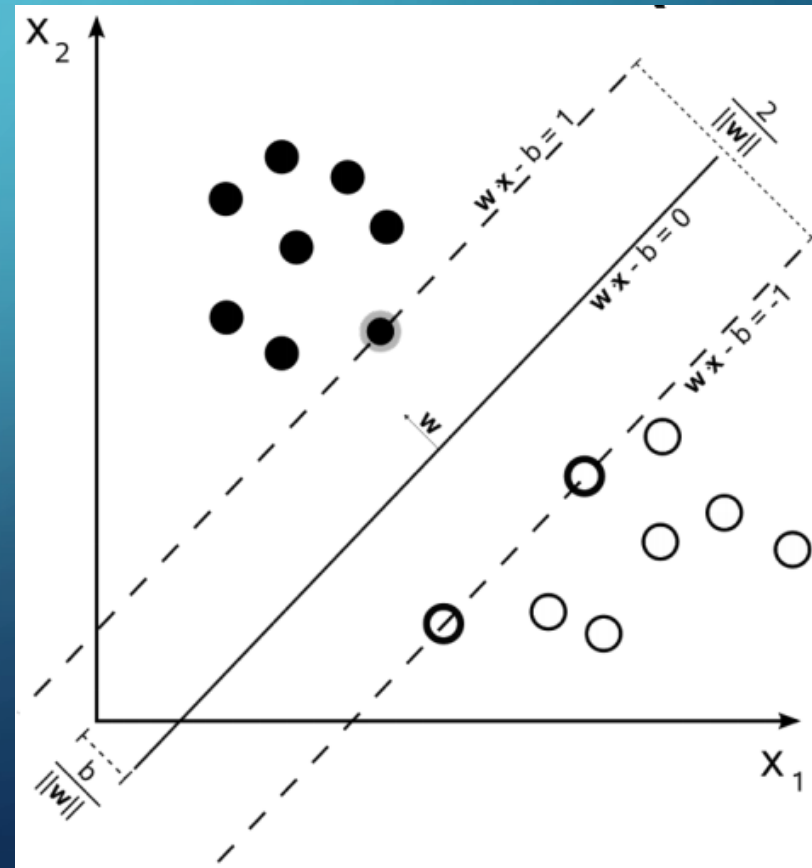
A photograph of a whiteboard with the Bayes' theorem formula written in blue marker. The formula is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The whiteboard is mounted on a wall, and the background is dark.

THE NAÏVE BAYES CLASSIFIER.

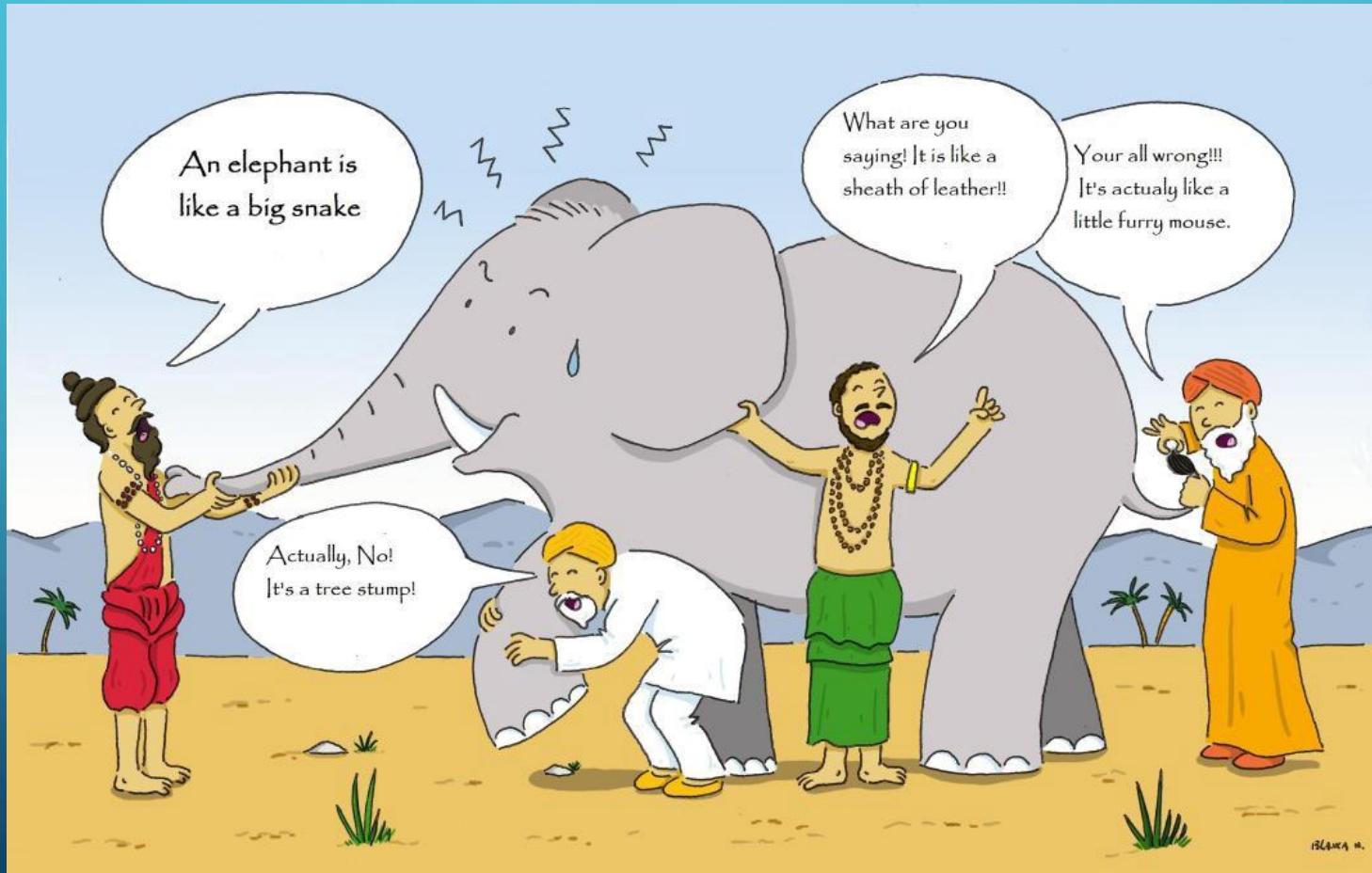
- Y = response variable (class)
- X = vector of features
- We (naively) assume that the components of X are independent, conditional on Y .

SUPPORT VECTOR MACHINES

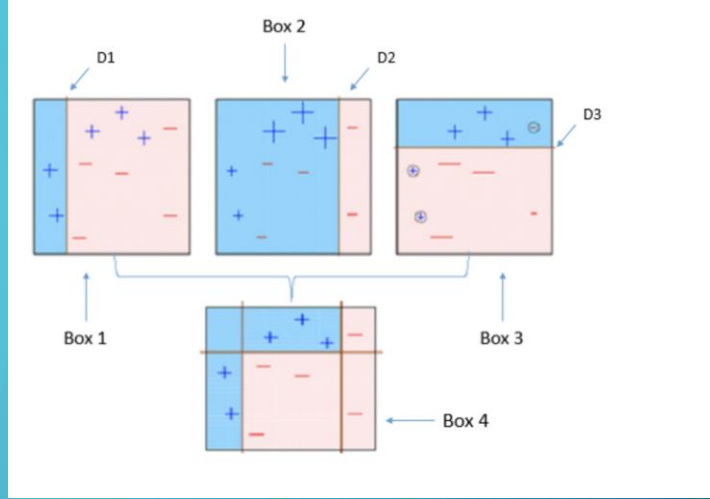
- How is it any different from regular regression?
 - In simple regression, we want to minimize the error rate.
 - In SVR we want the error within 'bounds'.



ENSEMBLE LEARNING



ENSEMBLE LEARNING



- Bagging & Boosting

- Bagging: Retrain the classifier on randomly sampled (bootstrap) data. Let them vote.
 - Low Overfitting
 - But results from each model will be correlated
- Boosting: Let each new model learn from past failures.
 - i.e.--seed samples with observations that were difficult to predict with previous data

ENSEMBLE LEARNING

- Random Forests
 - Random forest tries to break correlation in bagging by randomly selecting a subset of features to as candidates for each split in a classification tree.
 - Typically robust, smaller misclassification errors than single trees/bagging
 - `library(randomForest)` in R

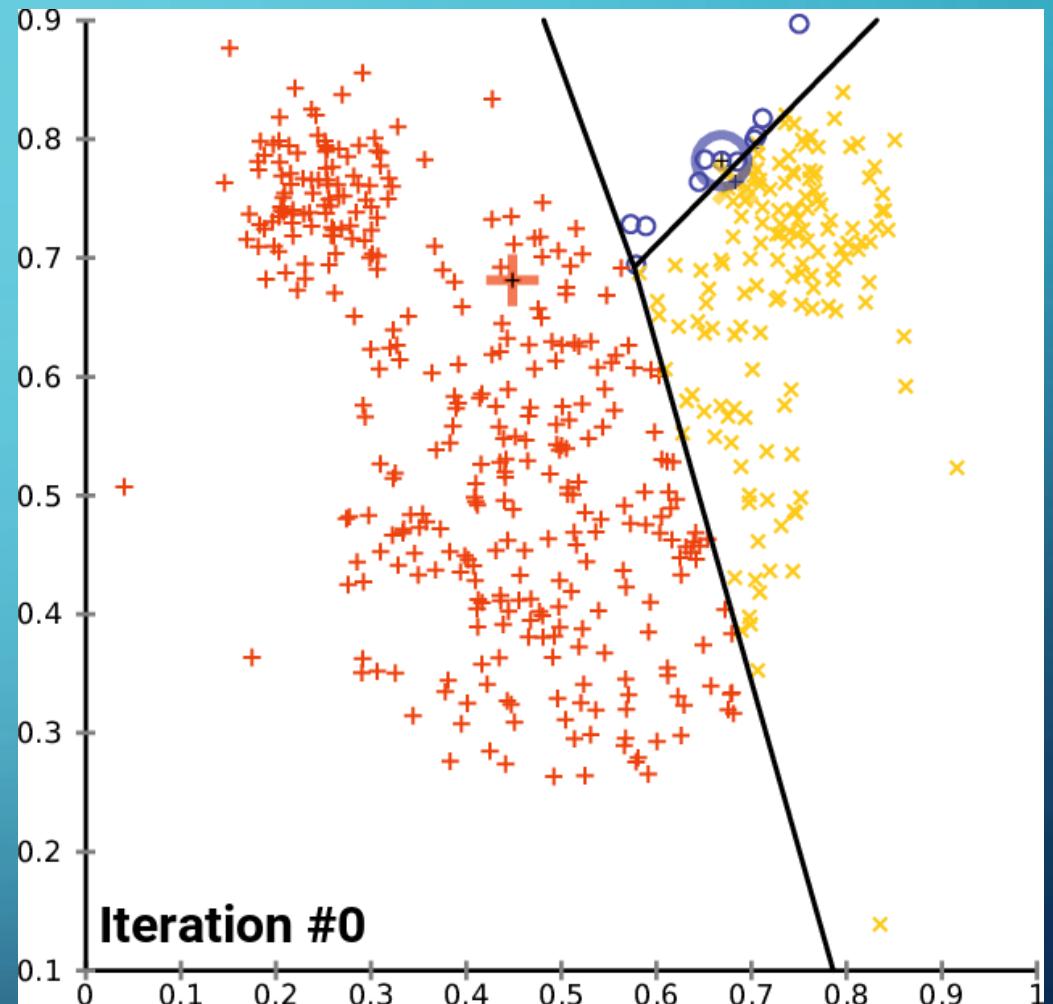
CLASSIFICATION VIA CLUSTERING

- This is an unsupervised learning problem.
 - Finding patterns in data that are not apparent outright.
 - Helps when there is no prior knowledge about the distribution of data.
- Problem: Identify factors that lead to higher incomes
 - Note: This problem is harder than you think 😊



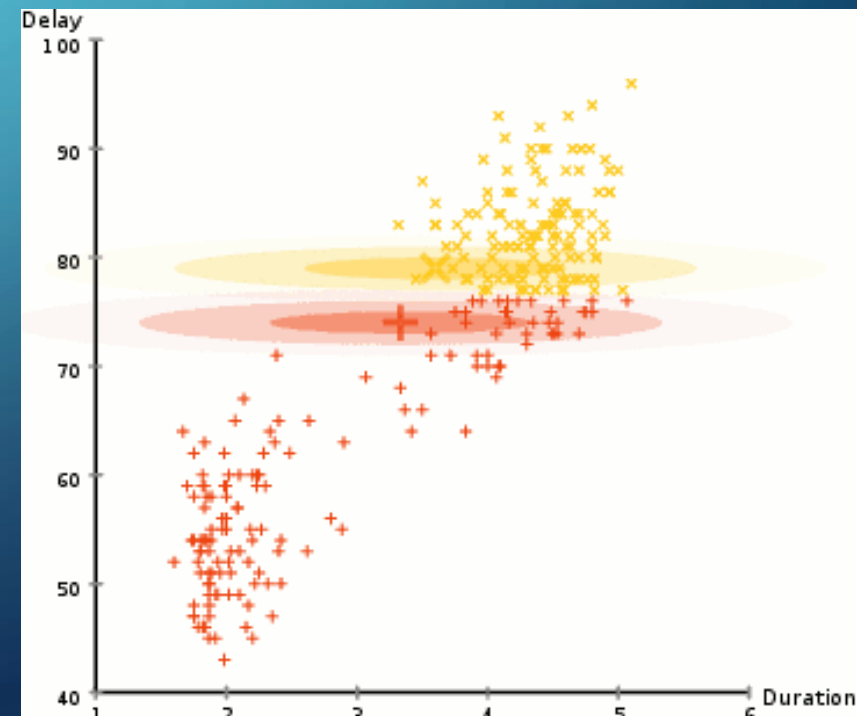
K-MEANS CLUSTERING

- Projects data into a Euclidian domain.
- Partition data into k clusters
 - Assignment is based on Euclidean distance.
 - Pre-set number of clusters in advance.
- Not guaranteed to terminate!
 - Try out multiple random seeds
 - `kmeans()` in R



MODEL-BASED CLUSTERING

- Cluster is modeled as a probability.
- Essentially like estimating means from a mixture model
- Unlike K-means, guaranteed to converge to local optimum using Expectation Maximization (EM)
- `library(Mclust)` in R

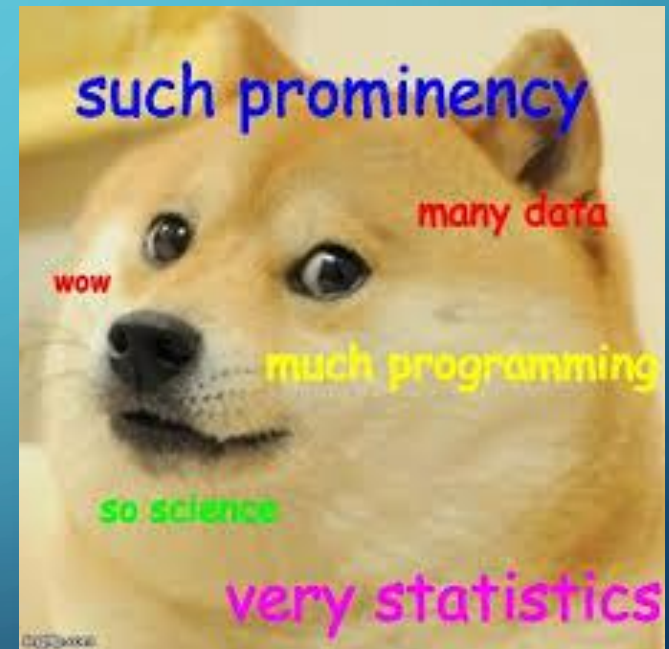


OTHER CLUSTERING METHODS:

- Self-Organizing Maps
- Dendrograms
- Generalizations of kmeans

PHEW, THAT'S A LOT

But we've just scratched the surface



WRAPPING UP

- Machine learning is powerful. But beware!
- No two experiments will give you the same result
 - Different Dataset
 - Different Initial Conditions
 - Inherent Randomness
- Disclaimer – My personal opinion below

Machine Learning should be the last thing you try when every deterministic method has failed

RESOURCES

- <https://towardsdatascience.com/>
- https://www.saedsayad.com/data_mining_map.htm
- Courses at UCR
 - CS235 – DATA MINING TECHNIQUES
 - EE236 – STATE & PARAMETER ESTIMATION THEORY
 - CS226 – BIG DATA MANAGEMENT
 - STAT208 – STATISTICAL DATA MINING METHODS

CREDITS

- Based on a similar presentation delivered by Vashishtha Bhatt (Winter '19)

NEURAL NETWORKS

- Simplest neural network.
 - Each node is connected to every other node in the next layer.
 - They all have the same activation function.
 - `library(nnet)` in R

