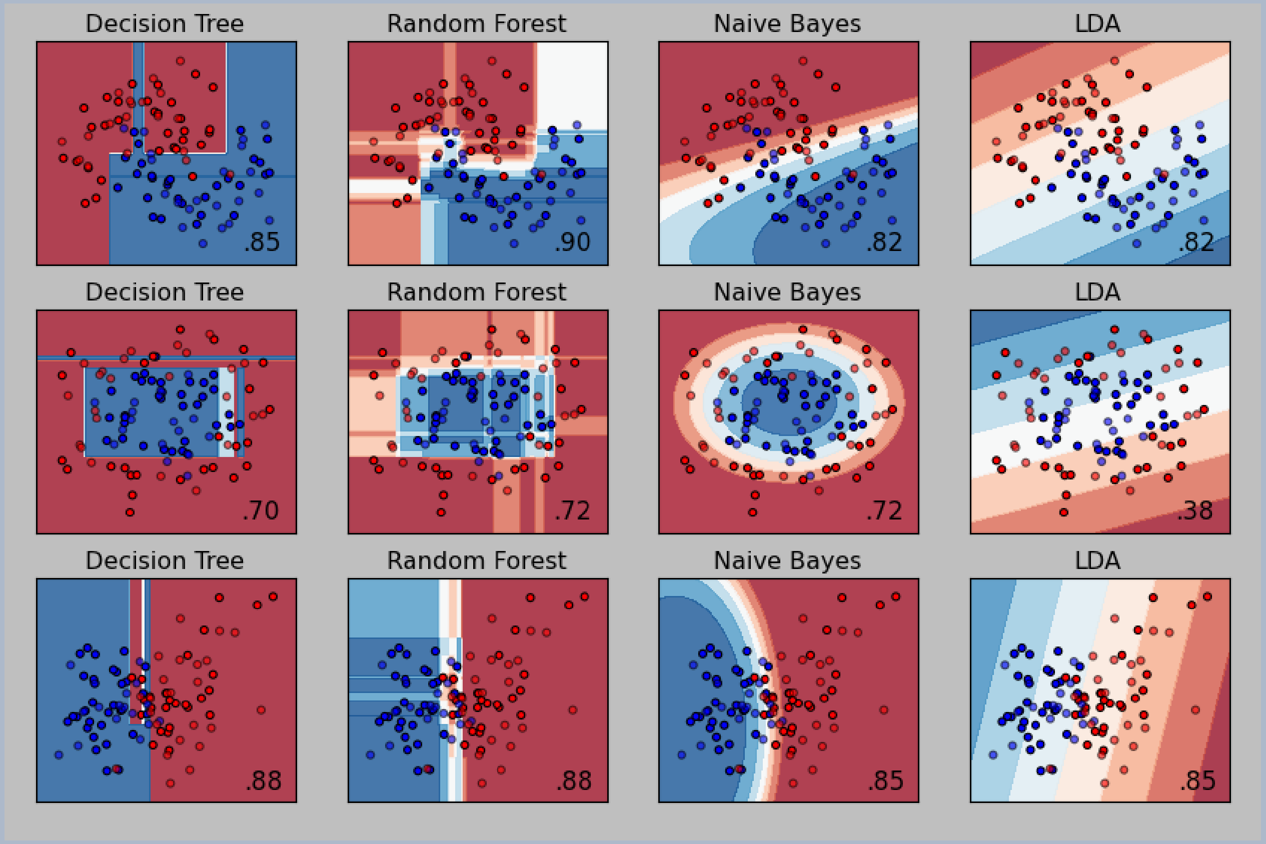
An Introduction to Machine Learning



Ryan Urbanowicz, PhD

Perelman School of Medicine
University of Pennsylvania

Overview

- Fundamentals of Machine Learning (ML)
- Focus: Decision Tree
- Choosing an ML algorithm
- Common ML Pitfalls

Terminology and Definitions

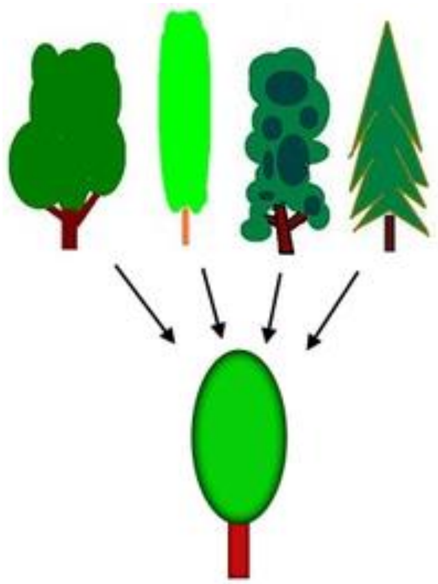
- Instance: an individual or example in data.
  - E.g. A subject/patient in a drug trial.
- Feature: one of the attributes describing an aspect of the instance. E.g. height, weight, age.
- Outcome: In supervised learning, this is endpoint value, a.k.a. the dependent variable, or the target being predicted.
  - Label/Class: Terms used for outcome in classification.
  - In regression, the outcome would be real-valued numbers.
- Model: A representation or simulation of reality. Typically a simplification based on a number of assumptions.

What is Machine Learning (ML)?

- A subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed[1].

[1] *Samuel Arthur – 1959 – ML in Checkers*

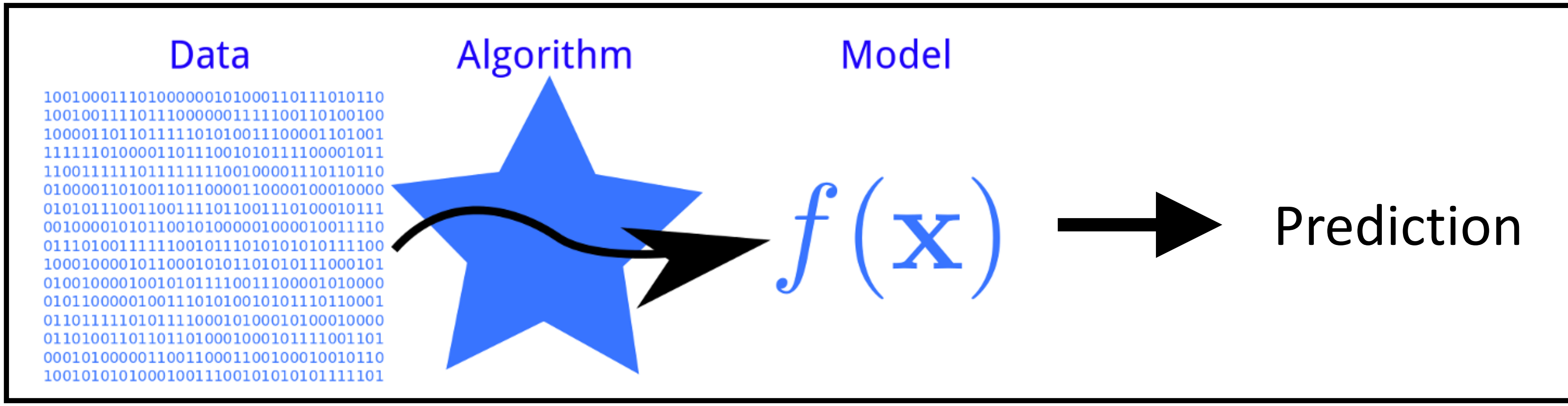
- ML is a general term → many algorithms/methods.
- Big Picture Goal: Learning useful generalizations.



An Important Clarification

- Machine Learning is…
  - Finding patterns or associations that can be used to make predictions.

Example: Predictive Modeling of Outcome



- Mostly NOT
  - Designed to demonstrate causality.
  - At best: associations are candidates for causality.

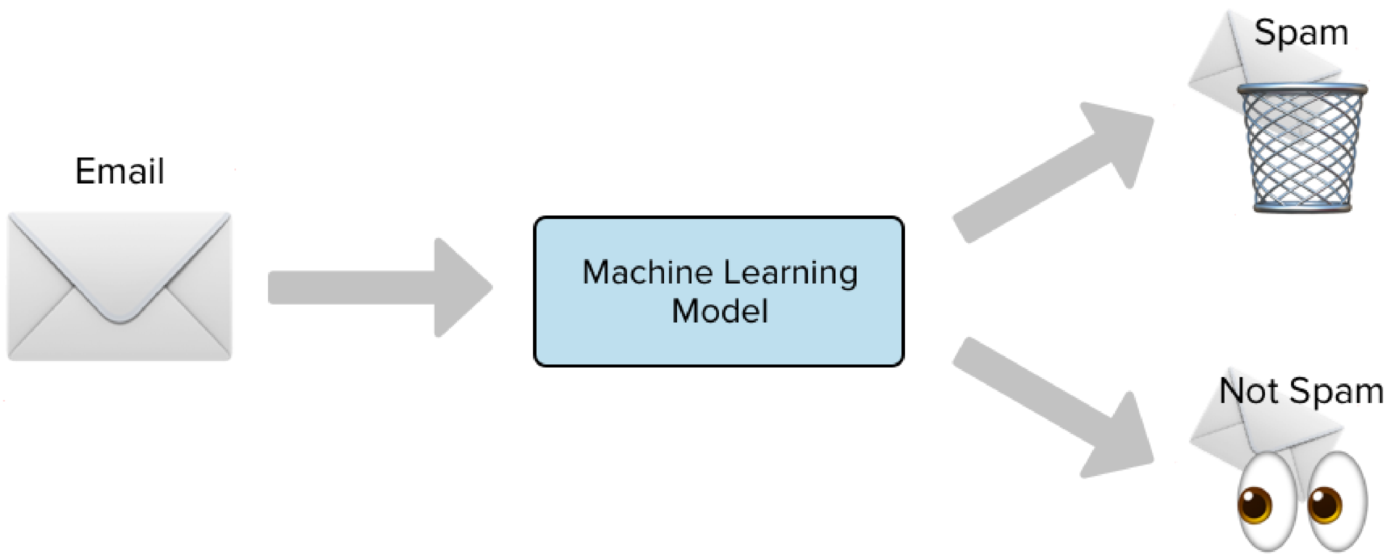
http://phdp.github.io/posts/2013-07-05-dtl.html

Perelman School of Medicine University of Pennsylvania
Ryan J. Urbanowicz, PhD
ryanurbanowicz.com
ryanurb@upenn.edu
@DocUrbs

Example: Email Spam Detection

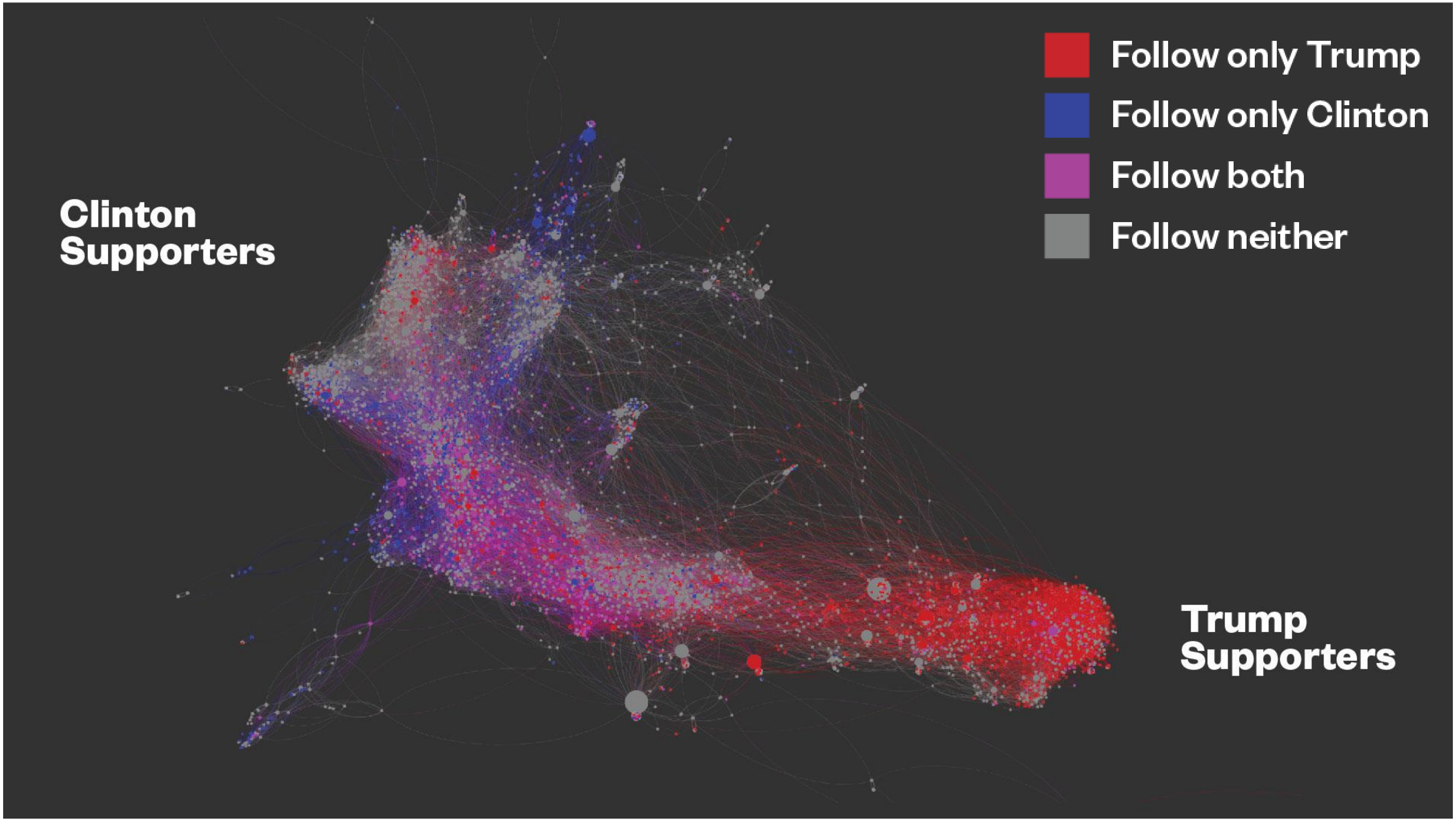
```
From: cheapsales@buystufffromme.com
To: ang@cs.stanford.edu
Subject: Buy now!

Deal of the week! Buy now!
Rolex w4tchs - $100
Med1cine (any kind) - $50
Also low cost M0rgages
available.
```

```
From: Alfred Ng
To: ang@cs.stanford.edu
Subject: Christmas dates?

Hey Andrew,
Was talking to Mom about plans
for Xmas. When do you get off
work. Meet Dec 22?
Alf
```



http://www.holehouse.org/mlclass/11_Machine_Learning_System_Design.html

Example: Community Detection



https://news.vice.com/en_us/article/d3xamx/journalists-and-trump-voters-live-in-separate-online-bubbles-mit-analysis-shows

Example: Association Mining

- Given a set of transactions, find rules that will predict purchase associations among items.

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

market basket transactions

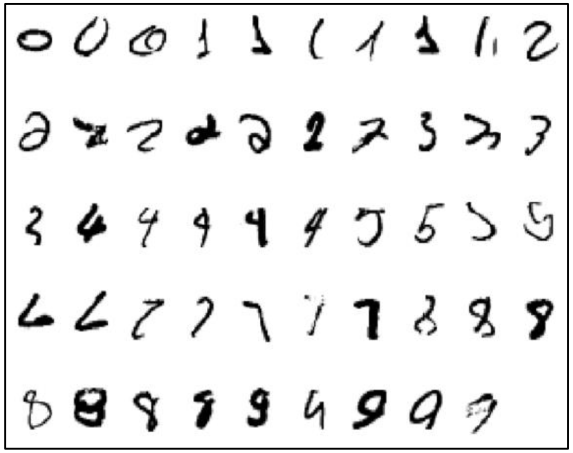
{Diapers, Beer}

{Diapers} → {Beer}

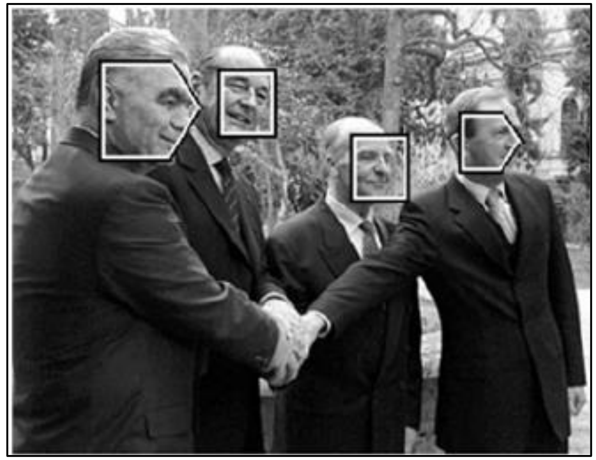
https://www.datacamp.com/community/tutorials/market-basket-analysis-r

Other Examples of Applied ML

Image Classification

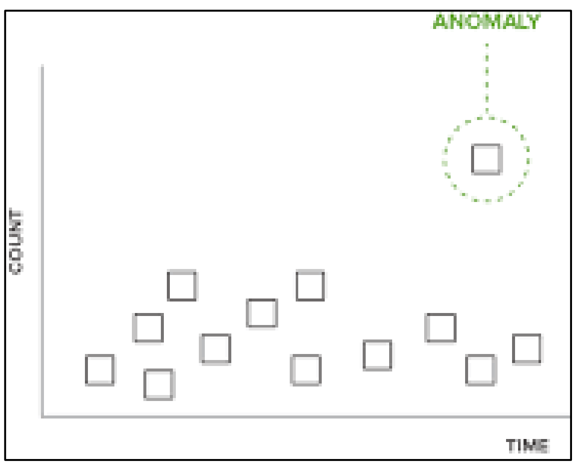


Face Detection



Stock Prediction

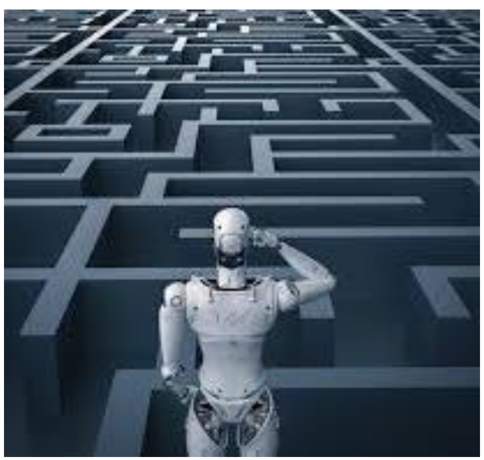
Fraud Detection



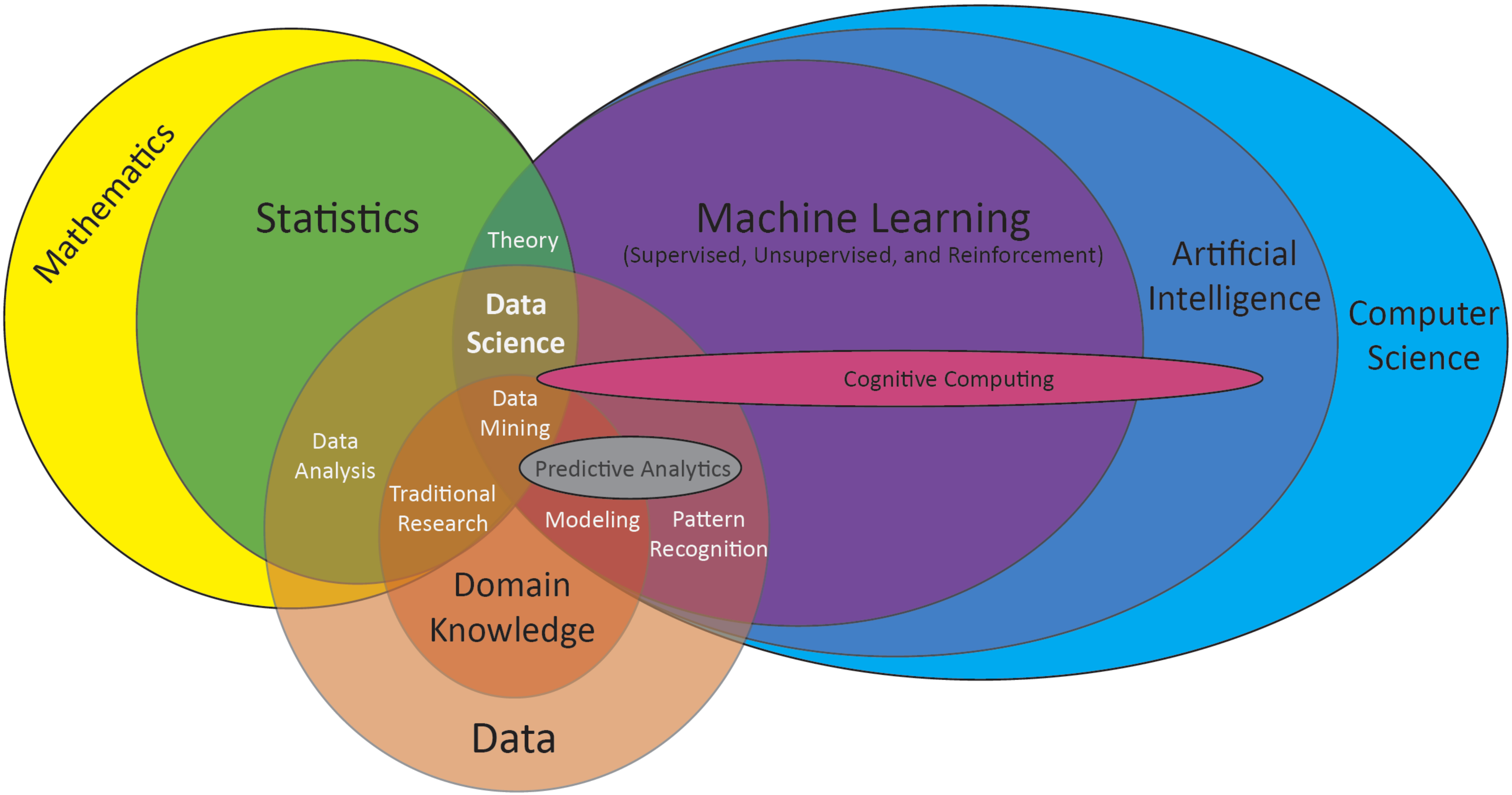
Risk Analysis

EXAMPLE RISK		Probability: Very High	High	Medium	Low	Very Low
Consequence	Very High	Very High	Very High	Very High	High	High
	High	Very High	High	High	Medium	Medium
	Medium	High	High	Medium	Medium	Low
	Low	High	Medium	Medium	Low	Very Low
	Very Low	Medium	Low	Low	Very Low	Very Low

Navigation



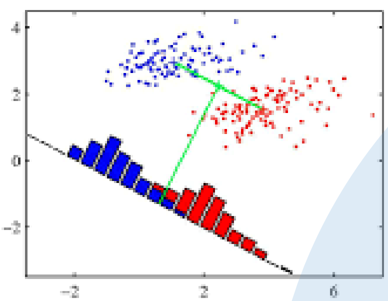
Fields & Terms Related to Machine Learning



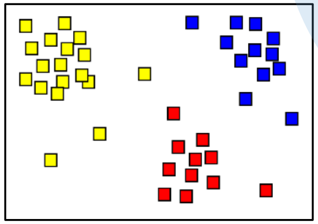
Statistics vs. Machine Learning

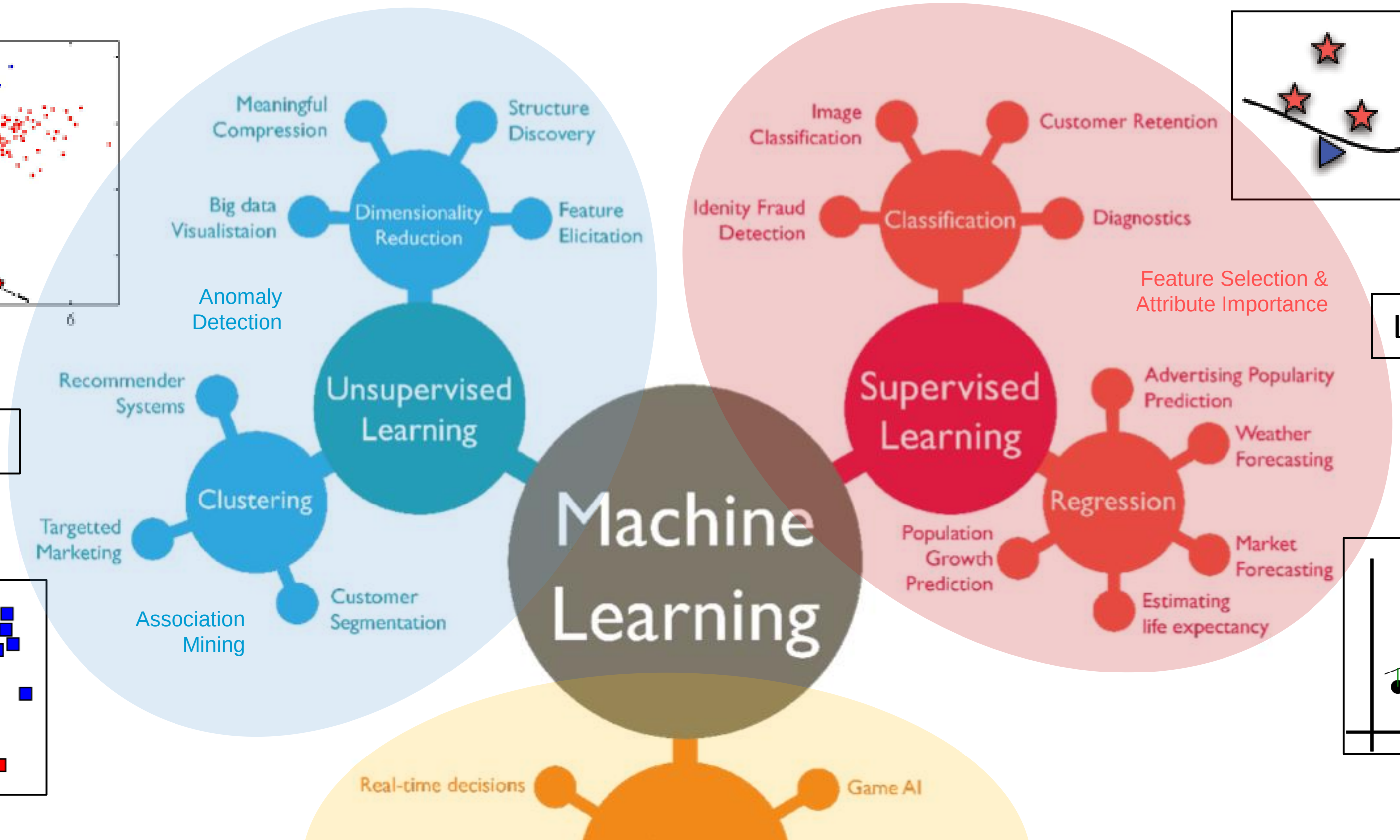
- Largely overlapping fields:
  - Both concerned with learning from data
  - Philosophical difference on ‘focus’ and ‘approach’.
- Statistics:
  - Founded in mathematics
  - Drawing valid conclusions based on analyzing existing data.
    - Making inference about a ‘population’ based on a ‘sample’
    - Tends to focus on fewer variables at once.
    - Precision and uncertainty are measures of model goodness.
- Machine Learning:
  - Founded in computer science
  - Focused on making predictions or seeking patterns (generalization).
    - Often considers a large number of variables at once.
    - Prediction accuracy to measure model goodness.

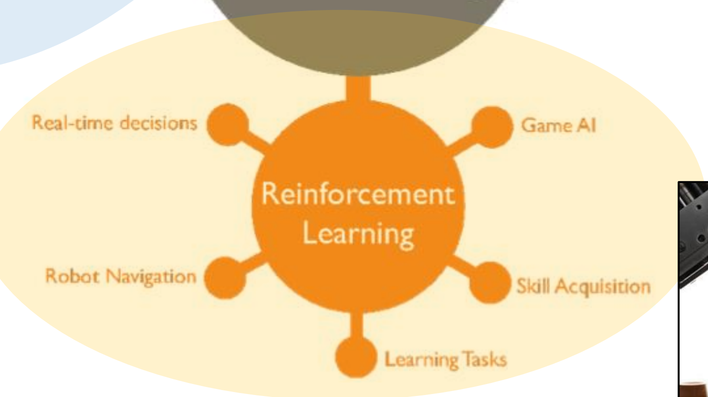
Types of Machine Learning

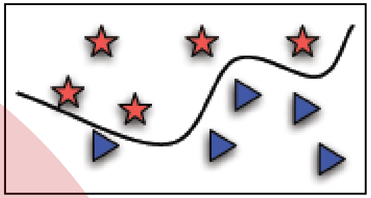


Unlabeled Data

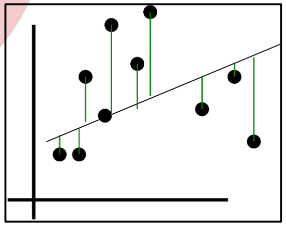


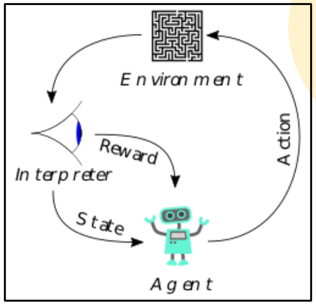






Labeled Data

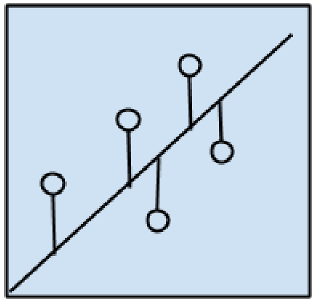




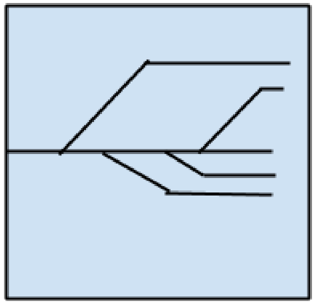
Multi-Step Problems

Adapted from : https://www.pinterest.com/pin/786792997374742269/

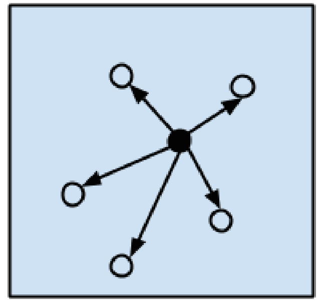
Machine Learning Algorithm Families



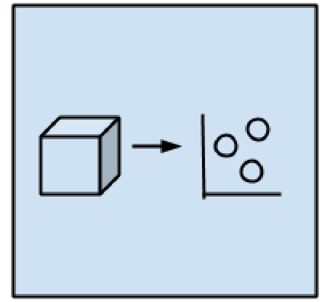
Regression Algorithms



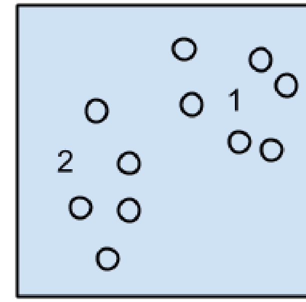
Regularization Algorithms



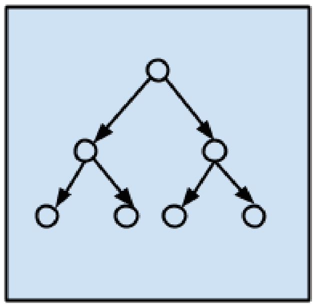
Instance-based Algorithms



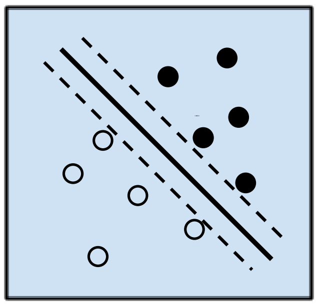
Dimensional Reduction Algorithms



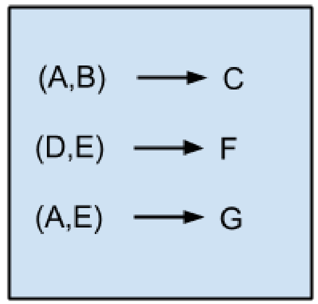
Clustering Algorithms



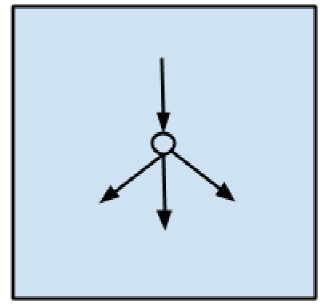
Decision Tree Algorithms



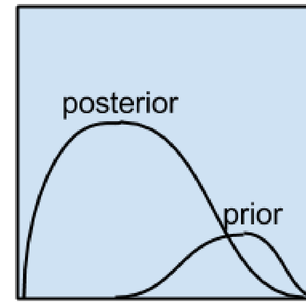
Support Vector Machines



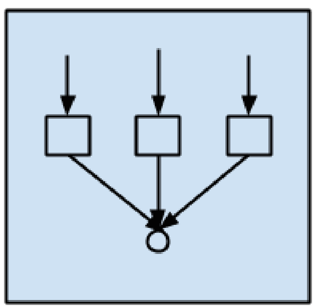
Association Rule Learning Algorithms



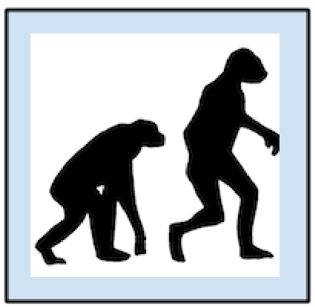
Artificial Neural Network Algorithms



Bayesian Algorithms

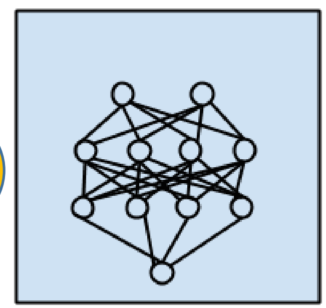


Ensemble Algorithms

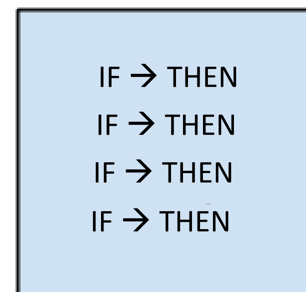


Evolutionary Algorithms

Non-exhaustive list of ML families



Deep Learning Algorithms

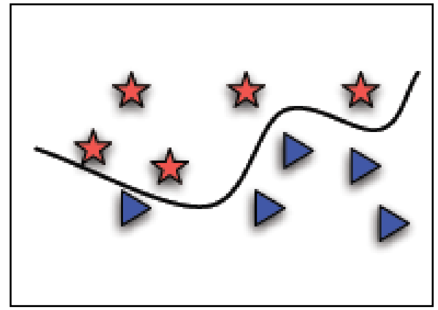


Learning Classifier Systems

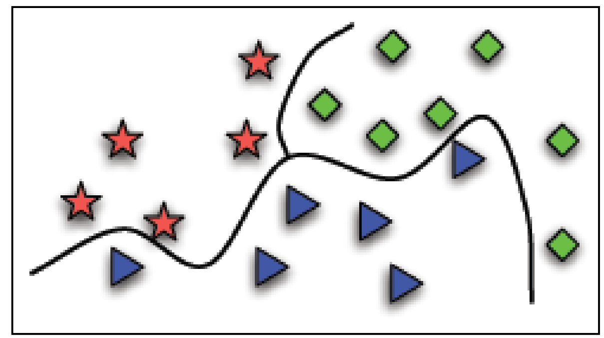
Perelman School of Medicine University of Pennsylvania | Ryan J. Urbanowicz, PhD | ryanurbanowicz.com | ryanurb@upenn.edu | @DocUrbs

Supervised Learning: Prediction

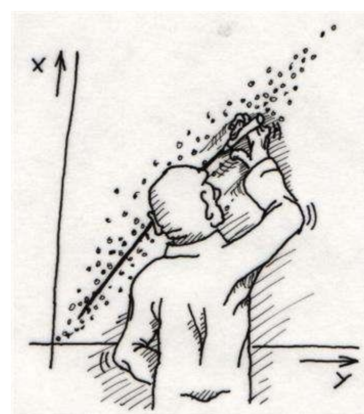
- Binary classification
  - Discriminate between two discrete classes/labels

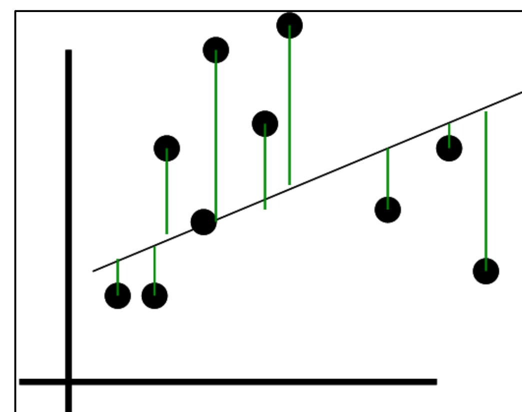


- Multiclass classification
  - Allows for more than 2 discrete classes.
  - E.g. Cancer classes may be healthy, early state, late stage.

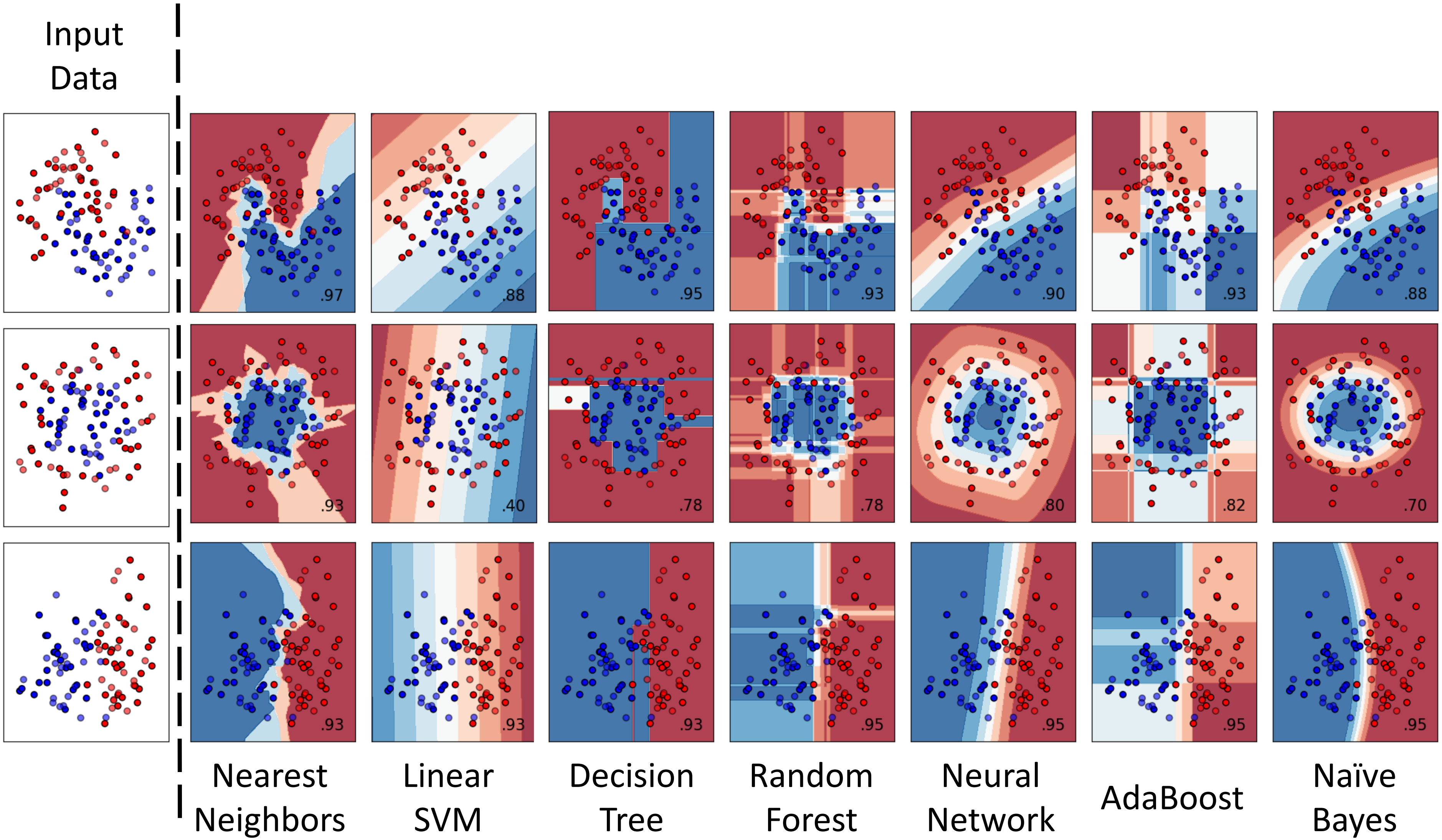


- Regression
  - Estimate a real-valued output variable





Modeling with Machine Learning

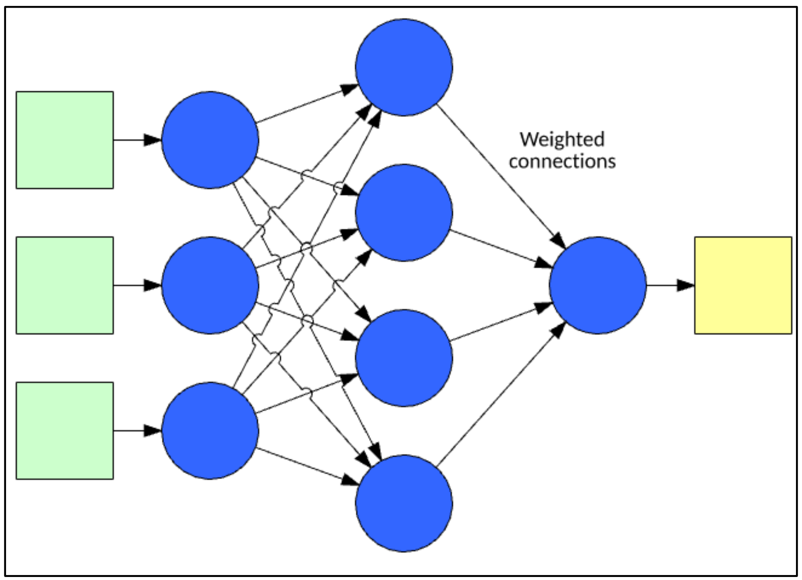


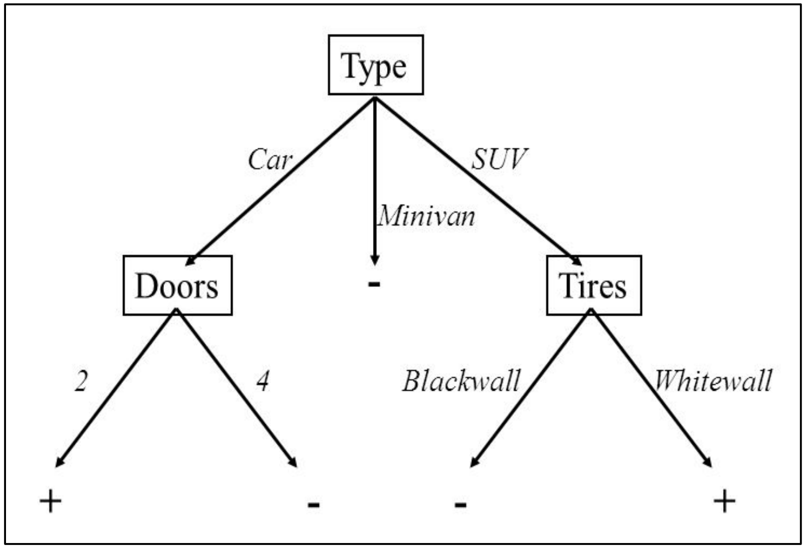
Models/ML: Representation

Population Y intercept; Population Slope Coefficient; Independent Variable; Random Error term; Dependent Variable

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error}}$$

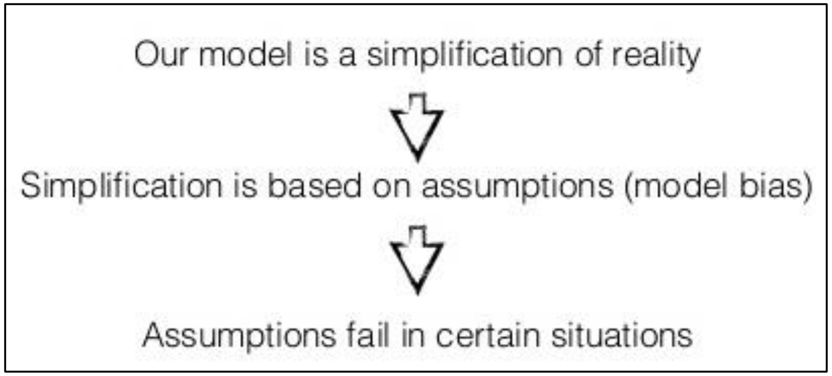
- R1: **IF** the animal has hair
  THEN it is a mammal
- R2: **IF** the animal gives milk
  THEN it is a mammal
- R3: **IF** the animal has feathers
  THEN it is a bird
- R4: **IF** the animal flies
  the animal lays eggs
  THEN it is a bird
- R5: **IF** the animal is a mammal
  the animal eats meat
  THEN it is a carnivore





Models and the NFL

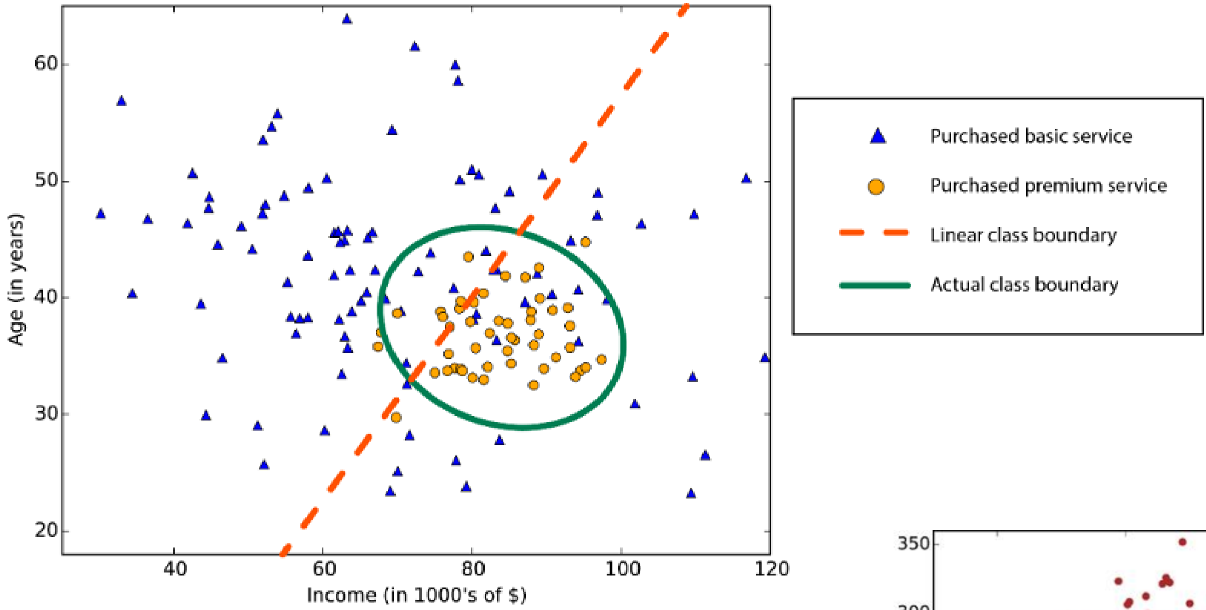
“All models are wrong, but some models are useful” – George Box



- Assumptions that work well in one domain may fail in another.
- No Free Lunch Theorem (NFL):
  - No single algorithm/model can perform optimally across all problems.
- Try:
  - More than one modeling approach
  - Different run parameters
    - “The knobs a data scientist gets to turn when setting up an algorithm to run”
  - Ensemble methods.

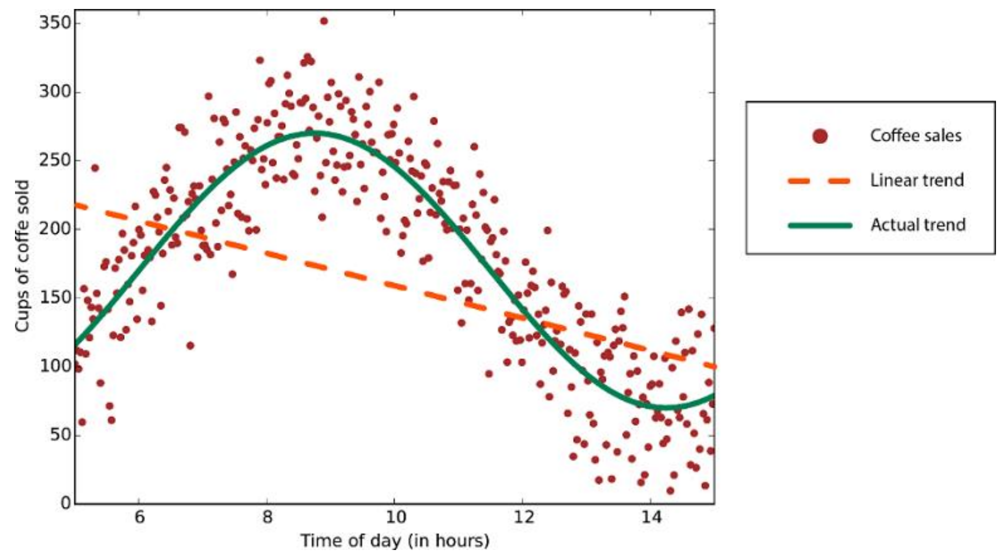
Perelman School of Medicine University of Pennsylvania | Ryan J. Urbanowicz, PhD | ryanurbanowicz.com | ryanurb@upenn.edu | @DocUrbs

Non-Linear Class Boundaries



Linear classification algorithm (e.g. SVM)

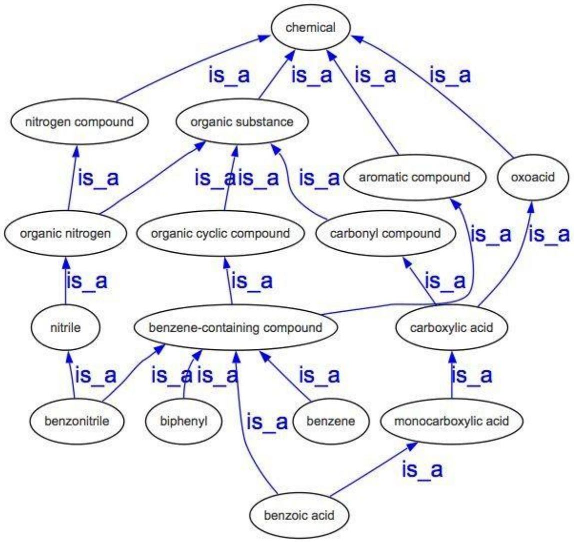
Linear regression algorithm

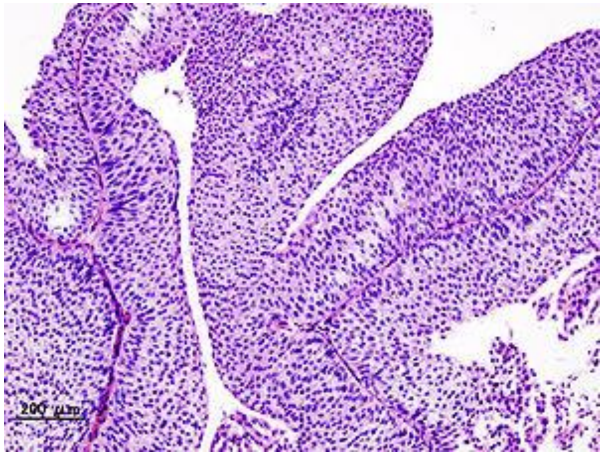


Data: Types

[0, 1, 1, 1, 2, 1, 0, 0]

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$





Feature Extraction/Engineering

Example:

Email Spam Detection

From unstructured text…

…To meaningful features for ML to interrogate.

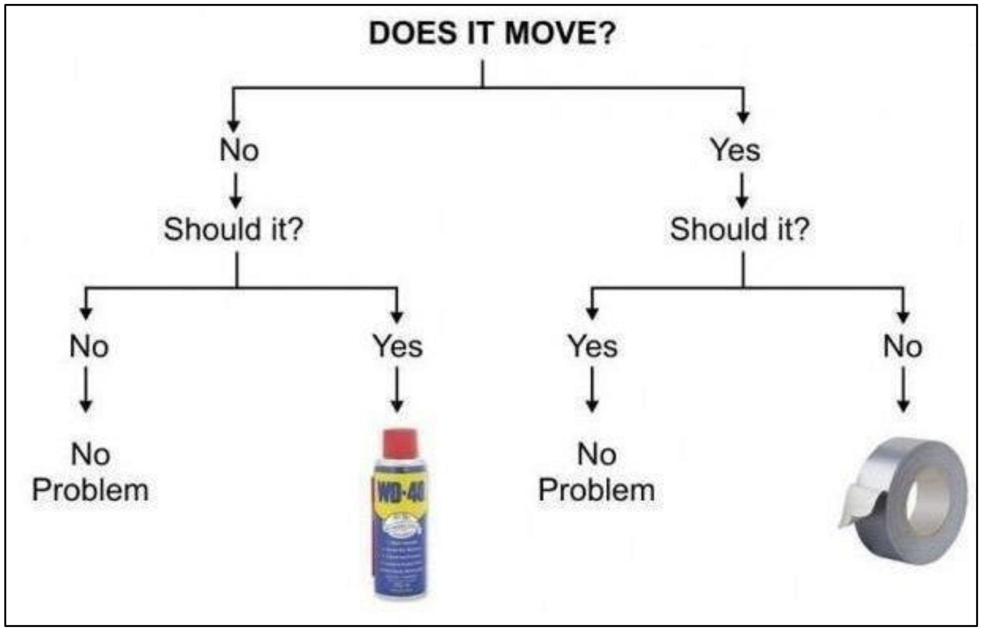
```
From: cheapsales@buystufffromme.com
To: ang@cs.stanford.edu
Subject: Buy now!

Deal of the week! Buy now!
Rolex w4tchs - $100
Med1cine (any kind) - $50
Also low cost M0rgages
available.
```

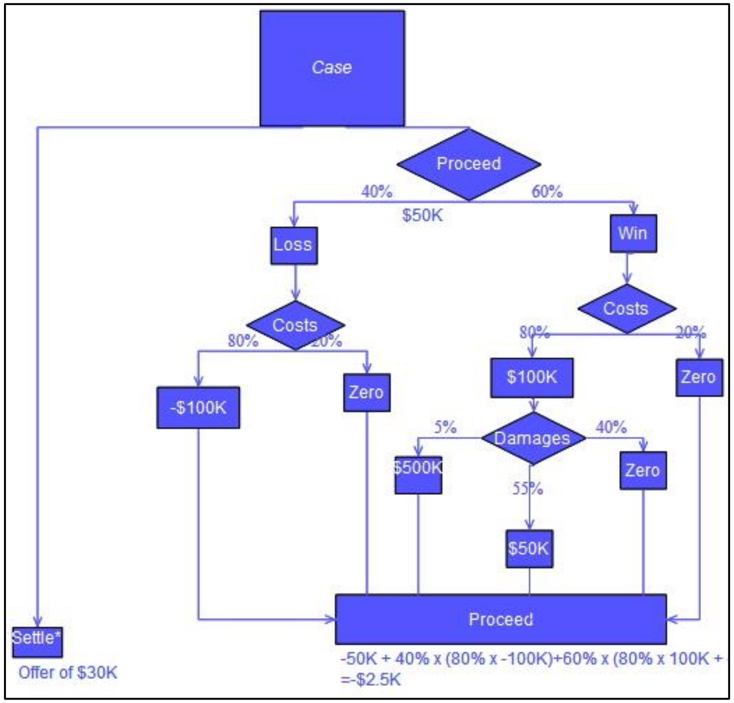
```
From: Alfred Ng
To: ang@cs.stanford.edu
Subject: Christmas dates?

Hey Andrew,
Was talking to Mom about plans
for Xmas. When do you get off
work. Meet Dec 22?
Alf
```

	"money"	"pills"	"Mr."	bad spelling	known-sender	spam?	
	Y	N	Y	Y	N	Y	
	N	N	N	Y	Y	N	
	N	Y	N	N	N	Y	
example	Y	N	N	N	Y	N	label
	N	N	Y	N	Y	N	
	Y	N	N	Y	N	Y	
	N	N	Y	N	N	N	

Decision Tree: What is it?

- A decision support tool: way to present information for decision making and evaluate their consequences (e.g. cost)

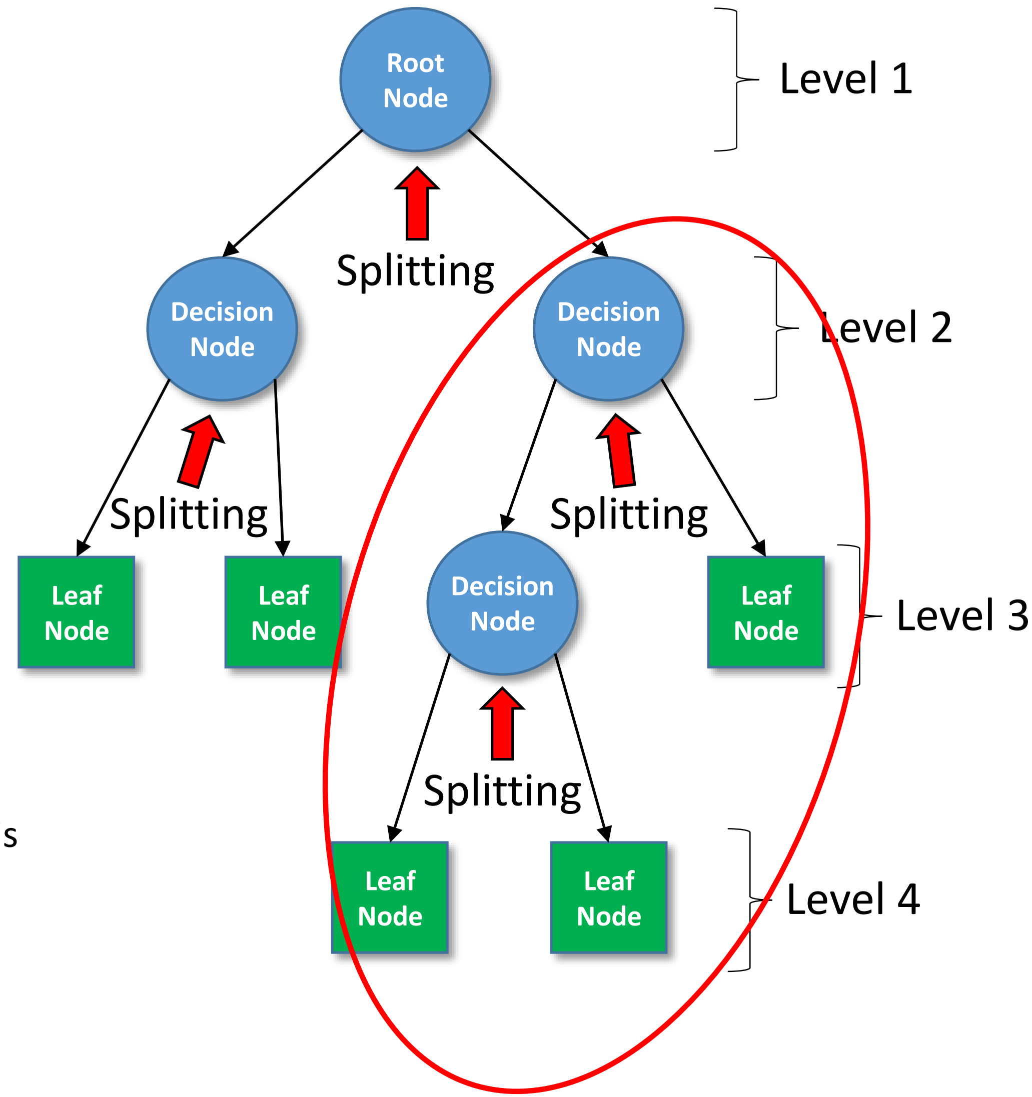




- A supervised, machine learning algorithm to model and predict outcomes

Decision Tree: Terminology

- **Nodes:**
  - **Root:** It represents entire population or sample. Will get divided into two or more homogeneous sets.
  - **Decision:** When a sub-node splits into further sub-nodes, then it is called decision node.
    - (AKA: Sub, internal, split, or chance node)
  - **Leaf:** Nodes that don't split. Gives class or average value.
    - (AKA: Terminal, or outcome node)
  - **Parent and Child:** Parent node splits into offspring nodes.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
- **Levels/Depth:** The number of splits through a given path down the three.

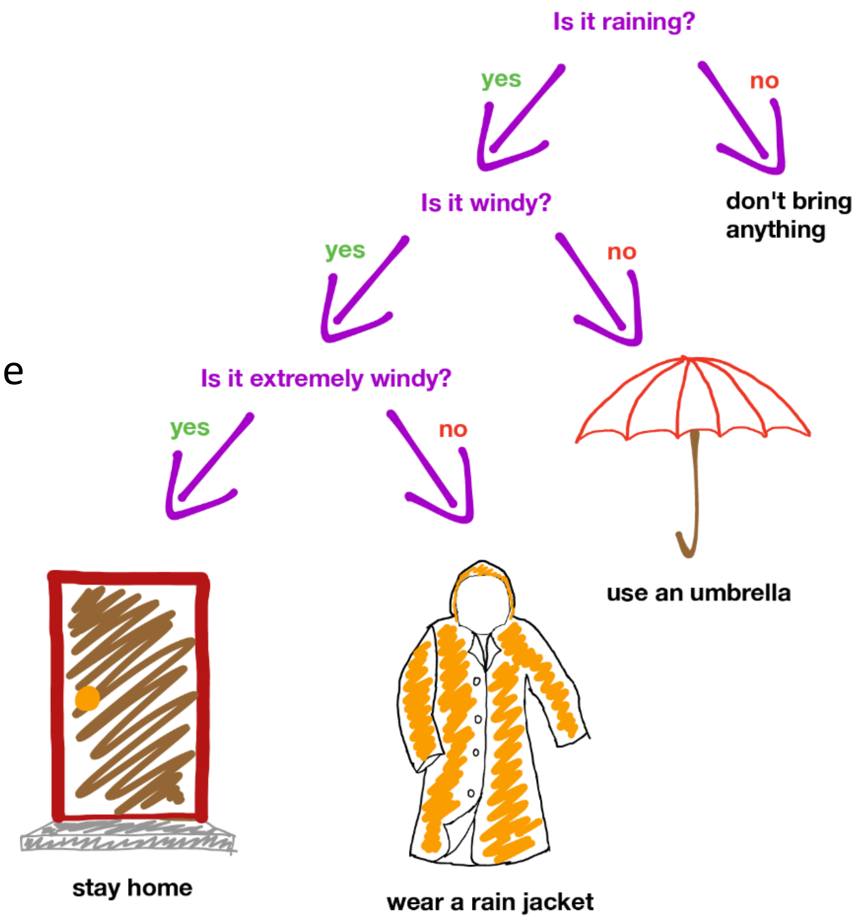


Decision Rules: Tree Interpretation

- Decision tree can be 'linearized' into *decision rules*.
  - One rule per path from root to leaf.
  - Rule outcome = Leaf node

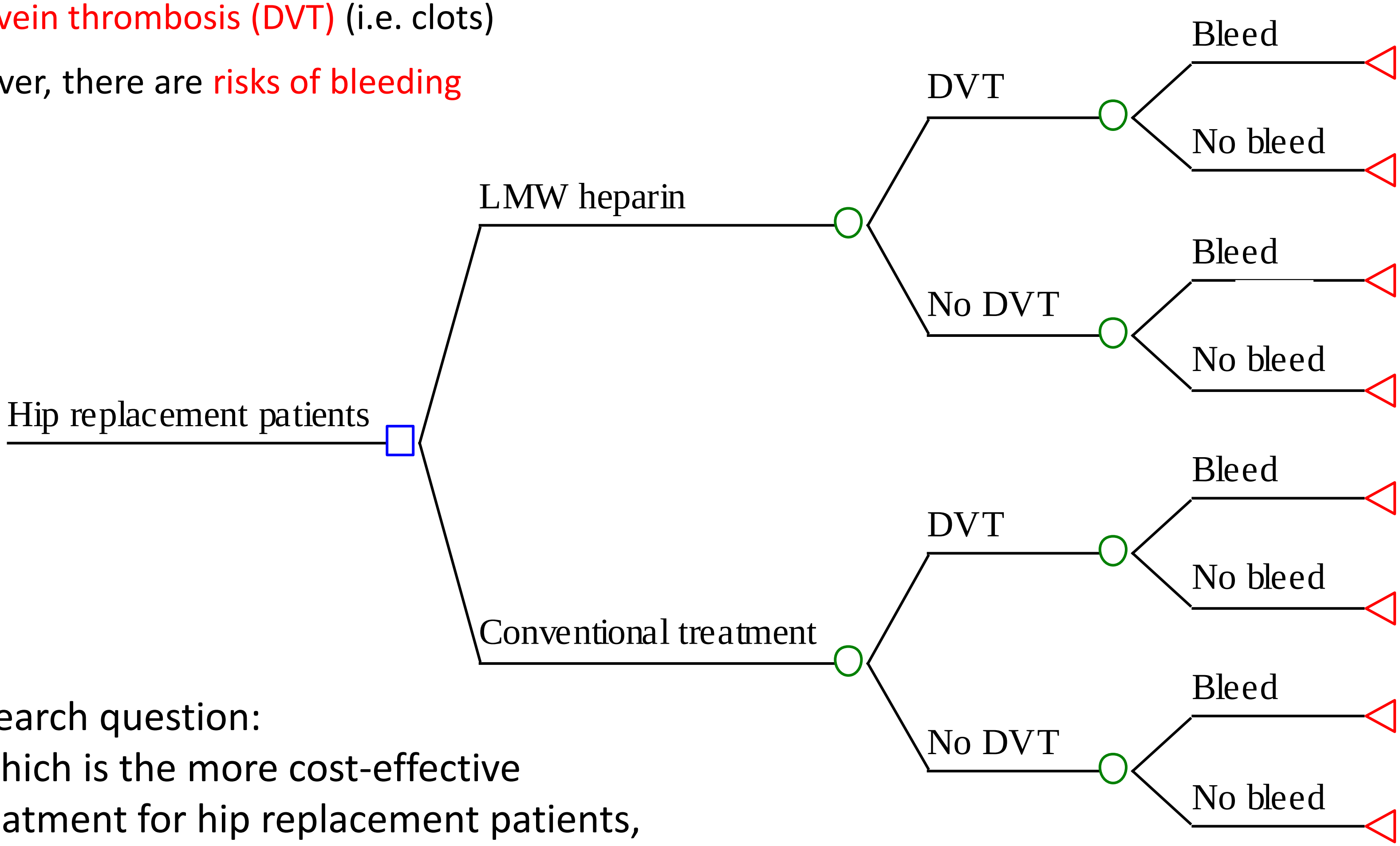
- Rule:
  - *<u>If</u>* [condition1] *and* [condition2] *<u>Then:</u>* outcome

- Examples:
  - *<u>If</u>* [not raining] *<u>Then:</u>* Don't bring anything
  - *<u>If</u>* [is raining] *and* [not windy]
    <u>Then:</u> use an umbrella



Decision Tree for Heparin

- Heparin (anticoagulant) injection for the prevention of deep vein thrombosis (DVT) (i.e. clots)
- However, there are risks of bleeding



The research question:

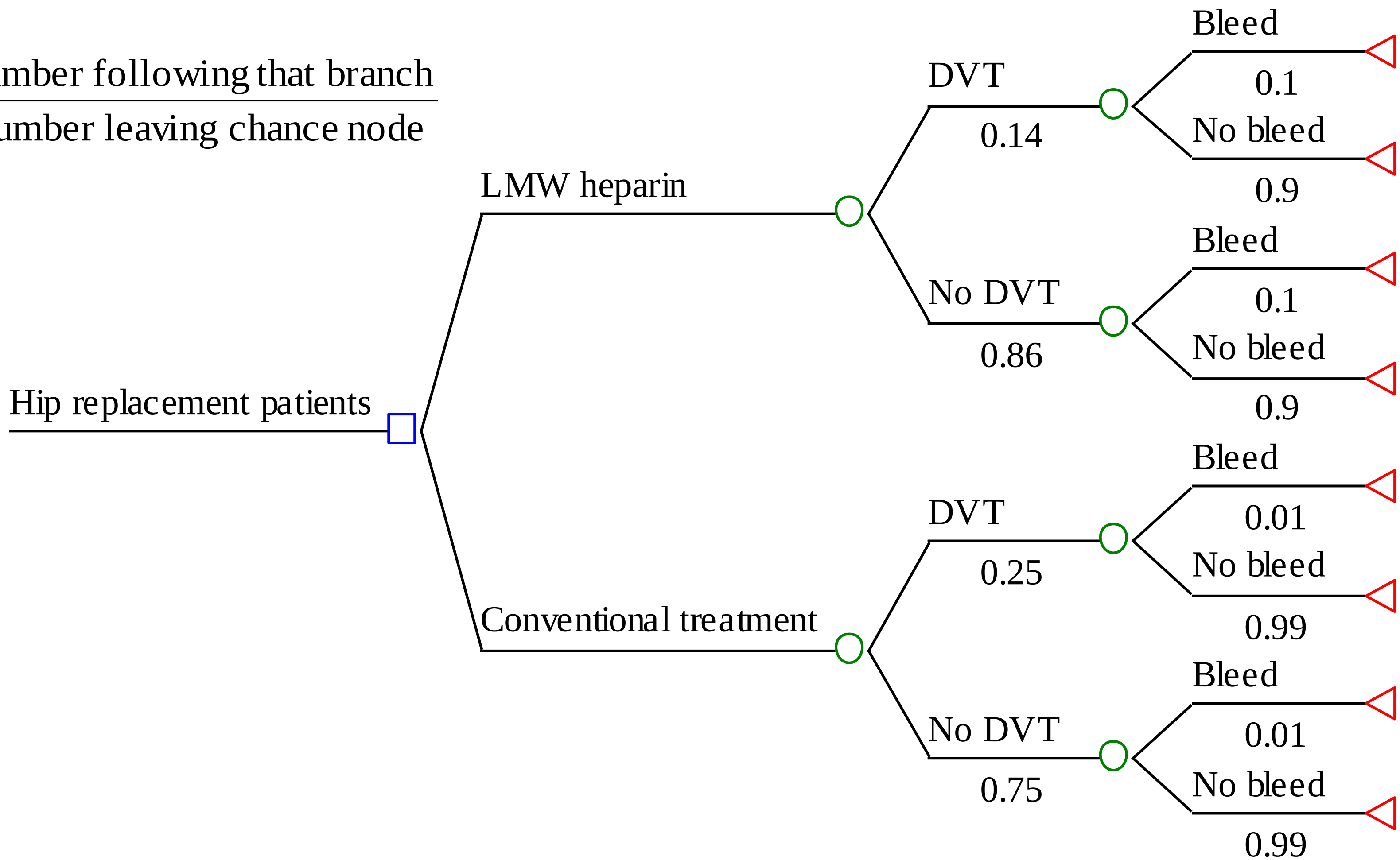
'Which is the more cost-effective treatment for hip replacement patients, heparin or conventional treatment?'

https://www.economicsnetwork.ac.uk/sites/default/files/Raymond%20Oppong/Introduction%20to%20modelling.pptx

Decision Tree for Heparin

- Entering probabilities

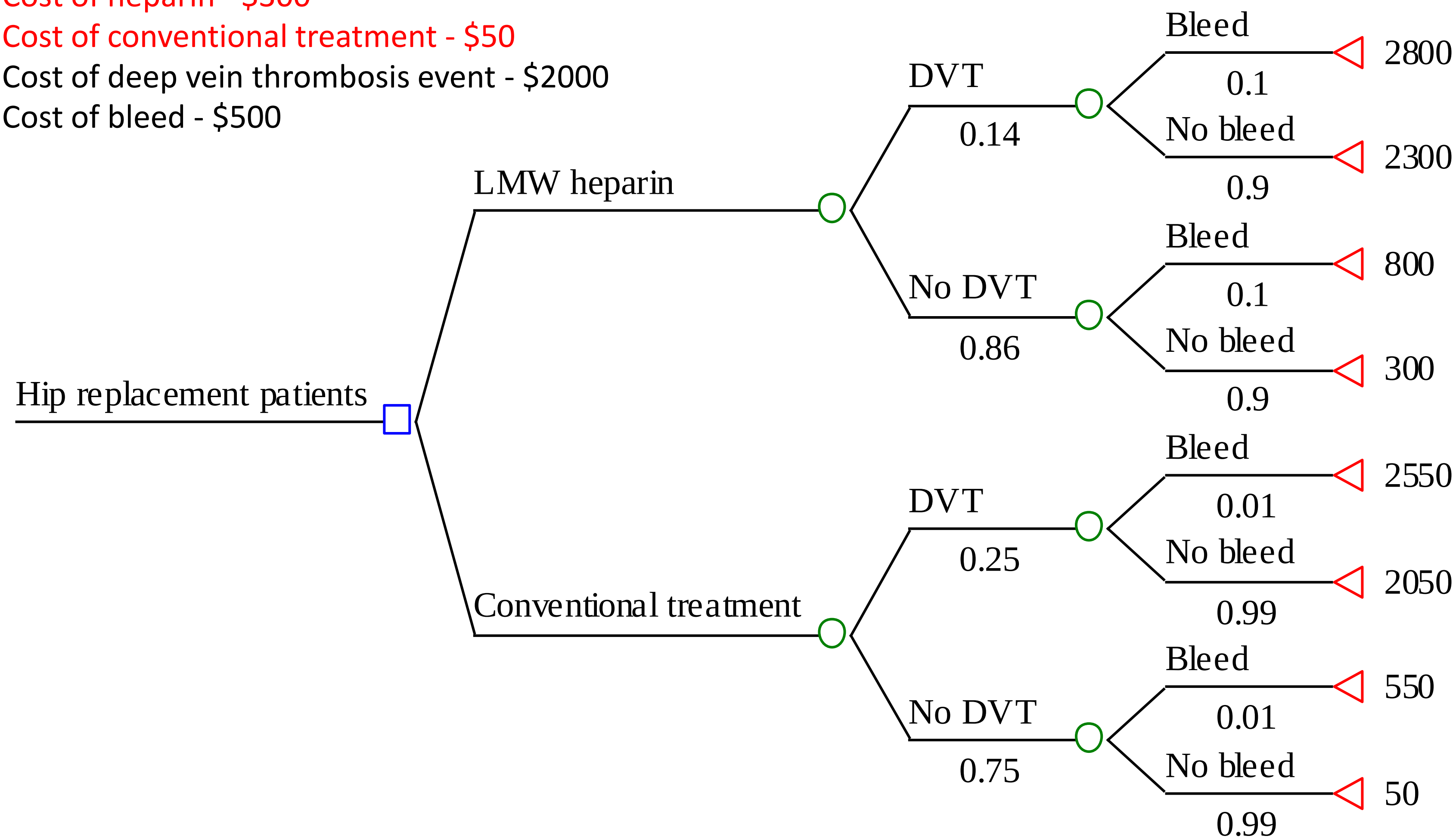
$$P = \frac{\text{Number following that branch}}{\text{Number leaving chance node}}$$



https://www.economicsnetwork.ac.uk/sites/default/files/Raymond%20Oppong/Introduction%20to%20modelling.pptx

Evaluating Outcome Costs

- Costs assumed
  - Cost of heparin - $300
  - Cost of conventional treatment - $50
  - Cost of deep vein thrombosis event - $2000
  - Cost of bleed - $500

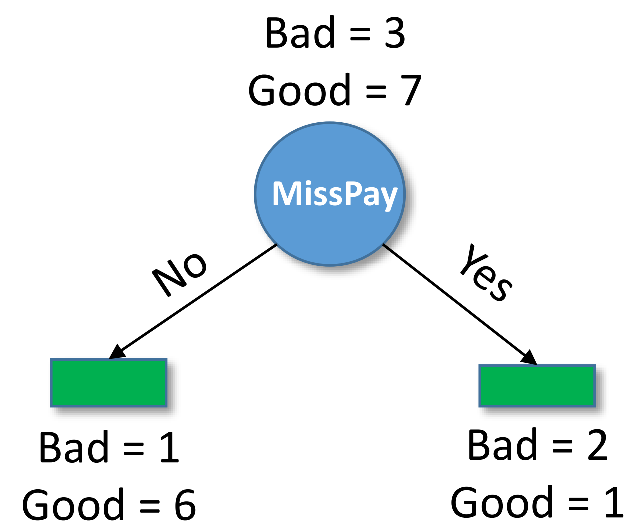


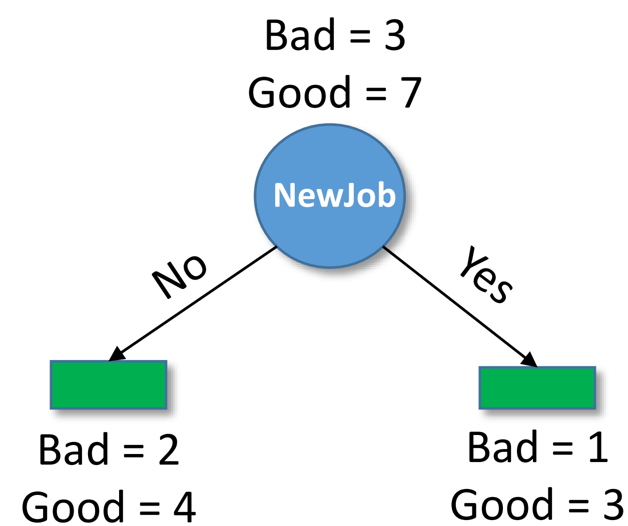
https://www.economicsnetwork.ac.uk/sites/default/files/Raymond%20Oppong/Introduction%20to%20modelling.pptx

Decision Tree: Choosing a Split

- E.g. Predicting Credit Risk
- What feature to split on?
- Want correct classification in fewest number of tests/branches.

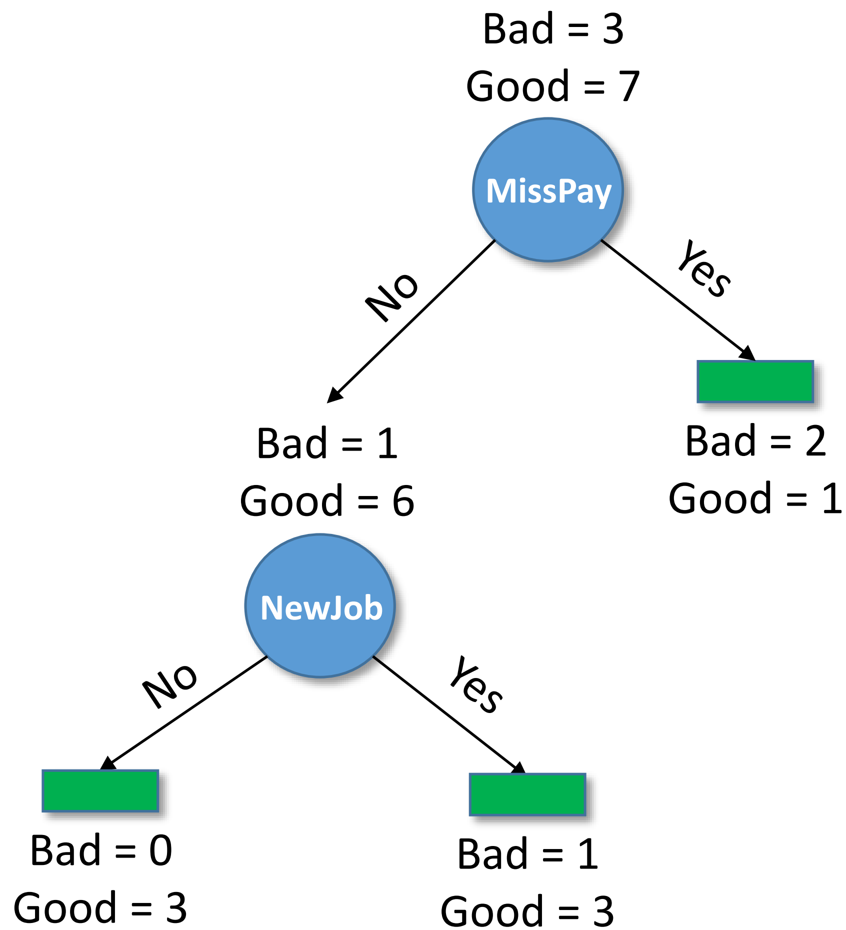
	< 2 years at current job	Missed payments?	Credit
S1	N	N	Good
S2	Y	N	Bad
S3	N	N	Good
S4	N	N	Good
S5	N	Y	Bad
S6	Y	N	Good
S7	N	Y	Good
S8	N	Y	Bad
S9	Y	N	Good
S10	Y	N	Good



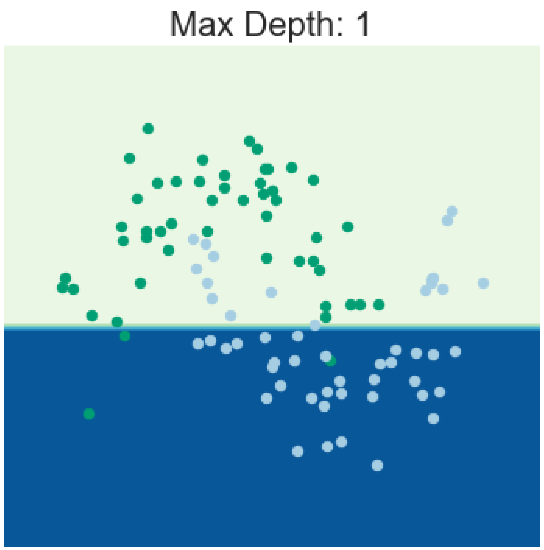


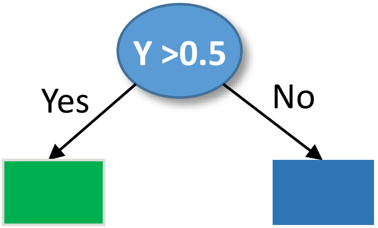
Decision Tree: Choosing a Split

	< 2 years at current job	Missed payments?	Credit
S1	N	N	Good
S2	Y	N	Bad
S3	N	N	Good
S4	N	N	Good
S5	N	Y	Bad
S6	Y	N	Good
S7	N	Y	Good
S8	N	Y	Bad
S9	Y	N	Good
S10	Y	N	Good

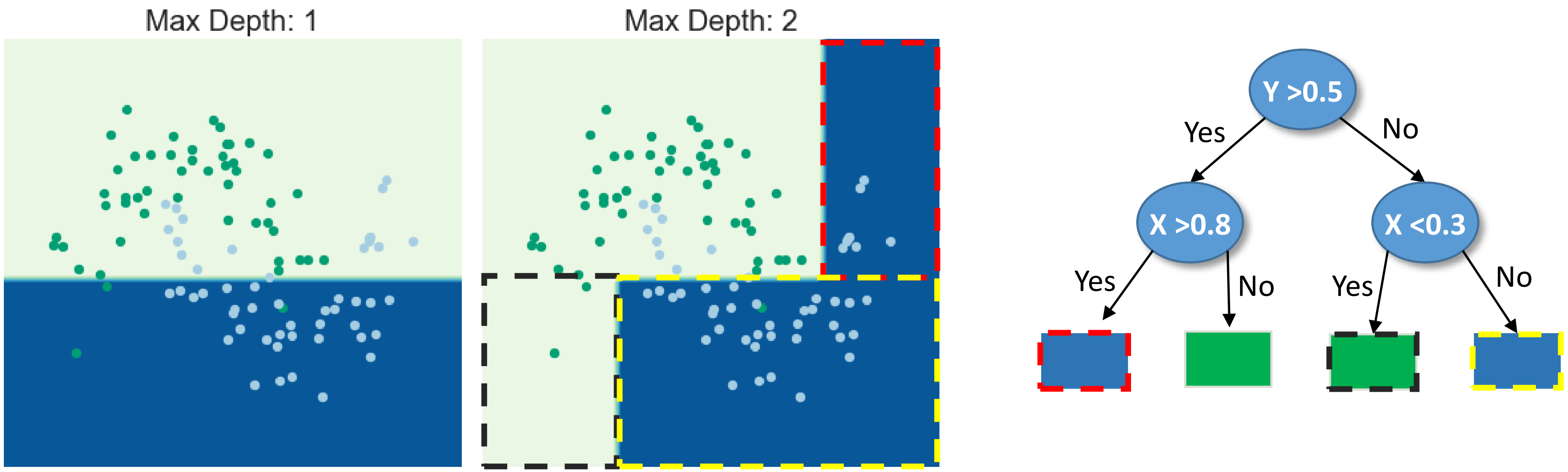


Decision Tree: Fitting with Splits

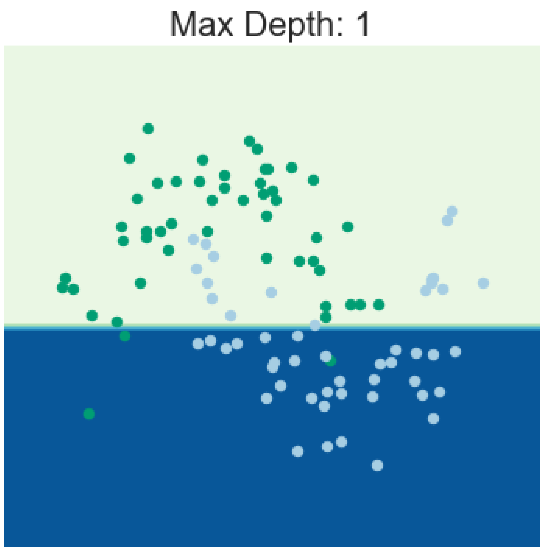


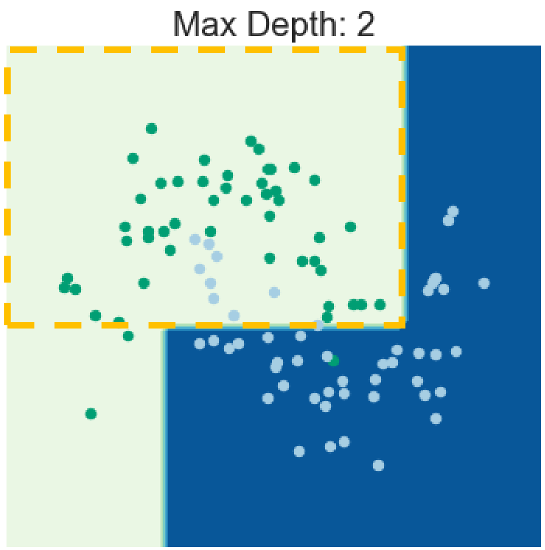


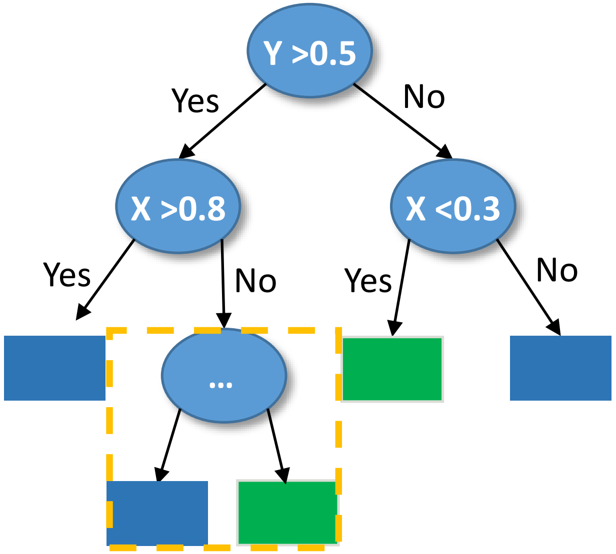
Decision Tree: Fitting with Splits



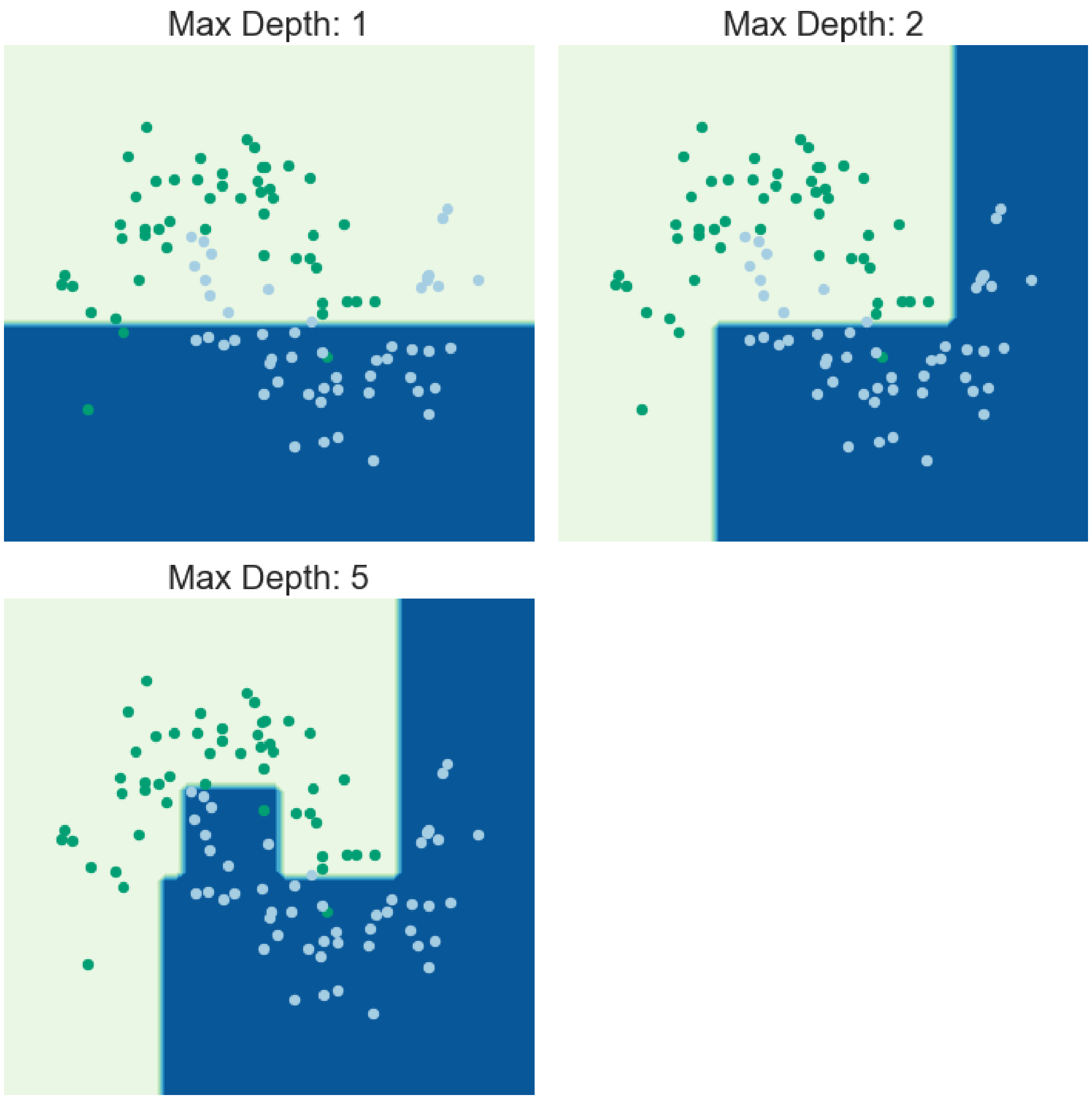
Decision Tree: Fitting with Splits

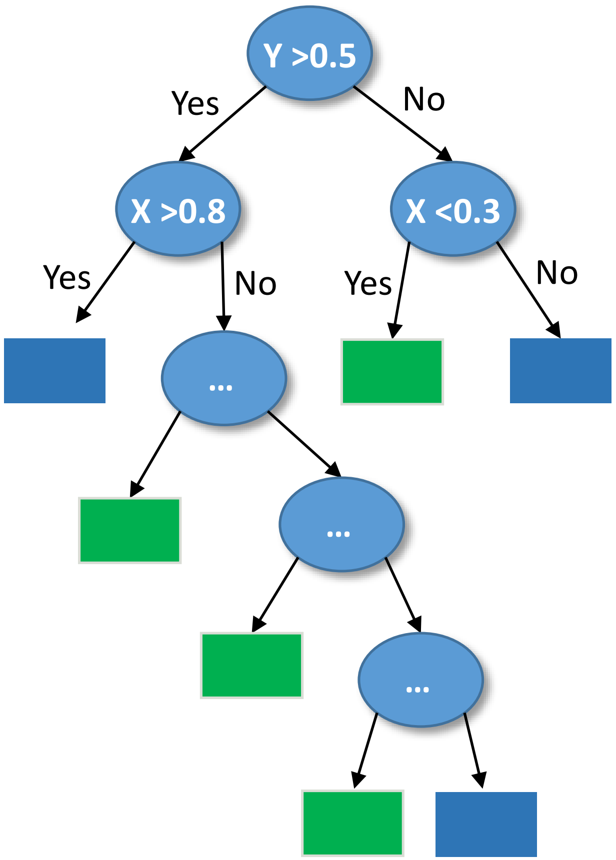




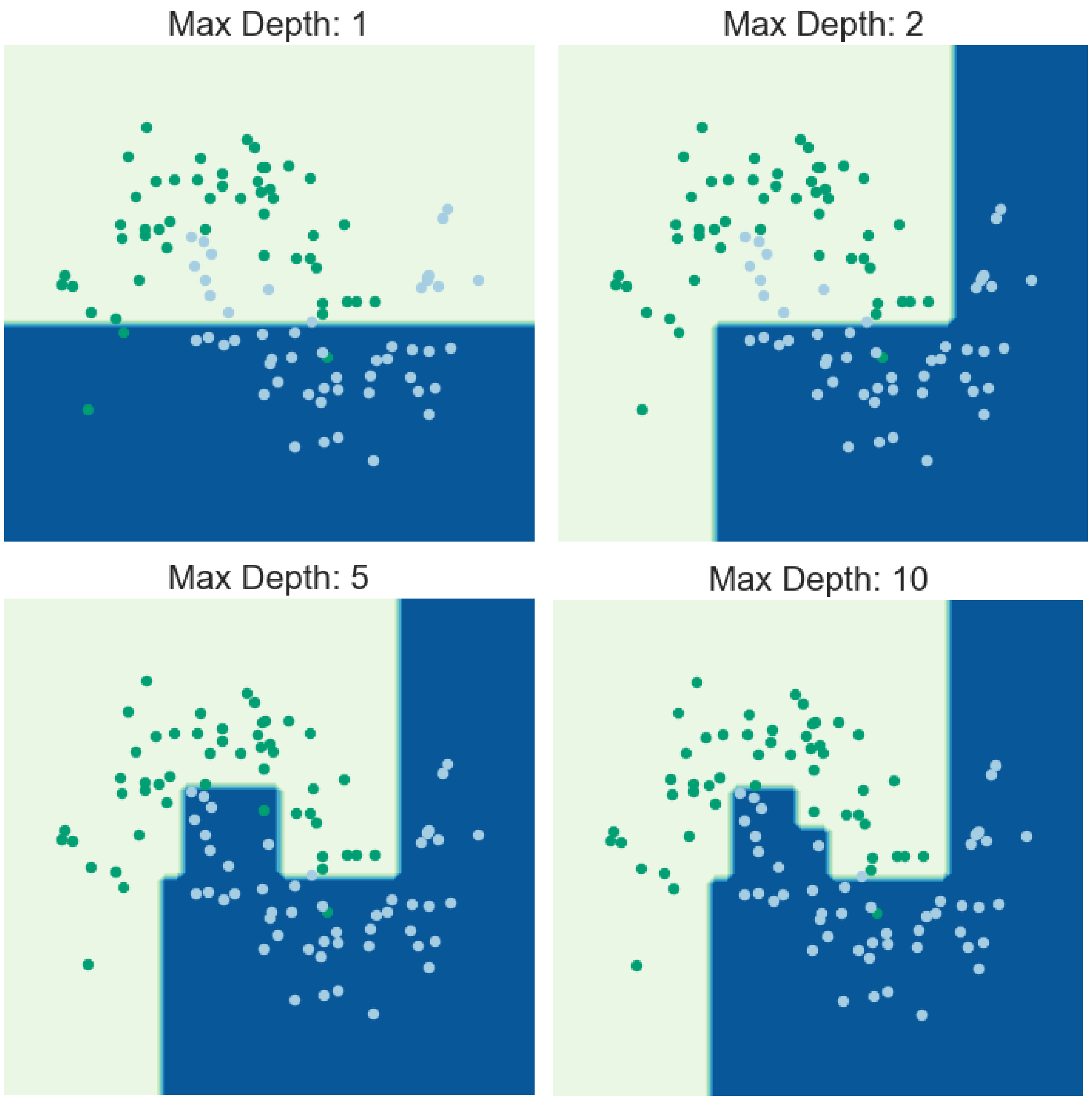


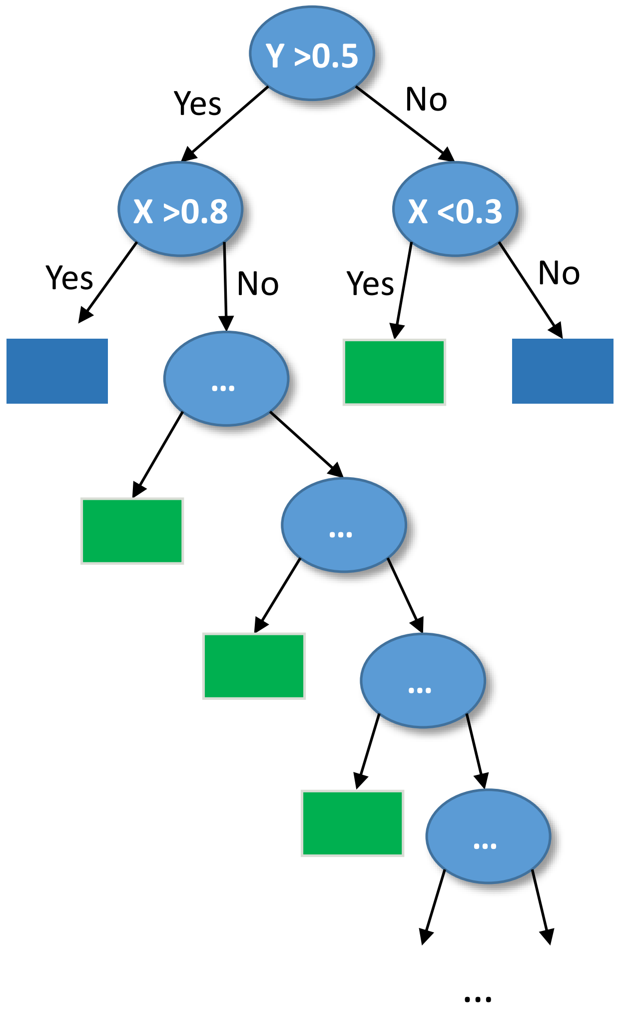
Decision Tree: Fitting with Splits





Decision Tree: Fitting with Splits





Decision Tree Challenges

- How do we decide best feature or value to split on?
- When should we stop splitting?
- What do we do if we can't achieve perfect classification?
- What if the tree is too large? Can we approximate a smaller one?

Where to start in selecting a method?

- If there is a strong, simple relationship among variables, most methods will find it.

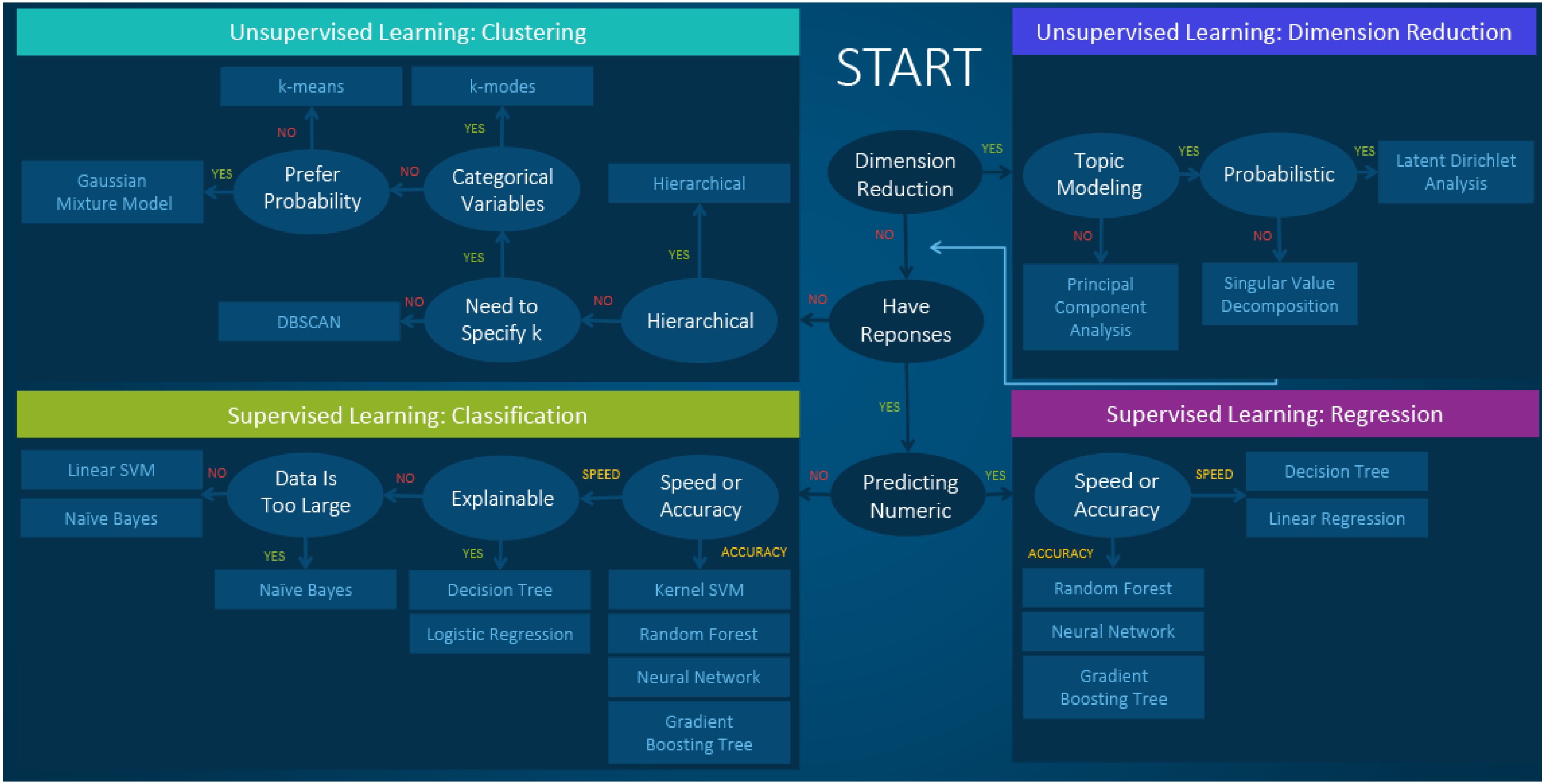
- Generally start with simpler methods if you know nothing about the problem.

- When possible, limit the search space with knowledge/assumptions about the problem.
  - E.g. If we want to know if there are linear patterns, use linear regression.

- Incorrect assumptions will limit or invalidate what can be found.

Considerations When Choosing an ML Algorithm

- Data – Labeled?, Endpoint?
- Training Time / Run Speed
- Number and Importance of Parameters
- Data Size – Features, Instances
- Interpretability
- Assumptions



ML Performance Evaluation

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

$$TruePositiveRate = \frac{TruePositive}{FalseNegative + TruePositive}$$

$$FalsePositiveRate = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

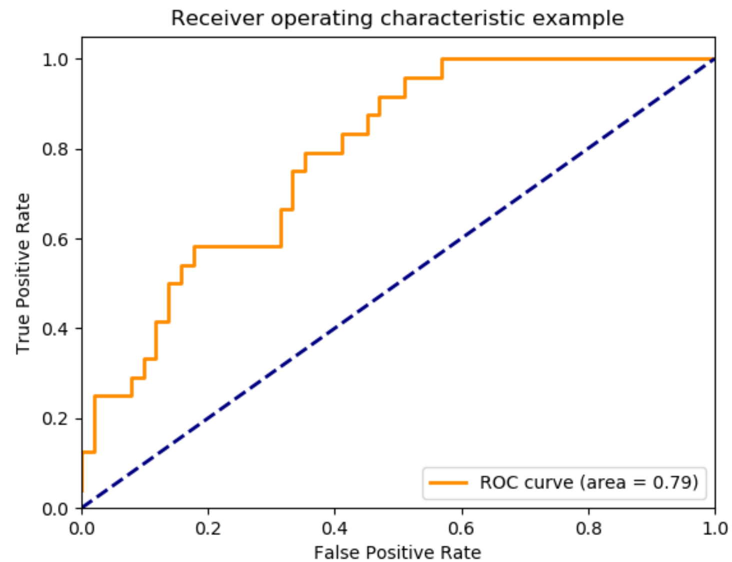
Confusion Matrix

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

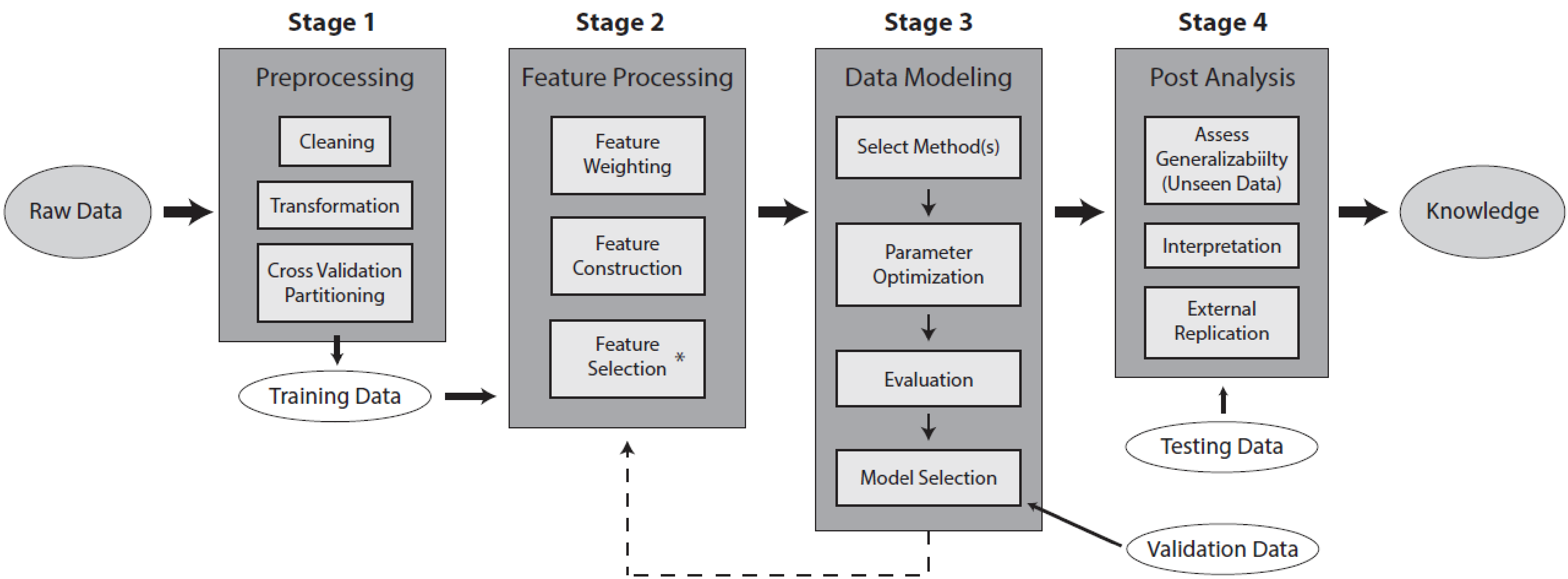


$$MeanAbsoluteError = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|$$

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Data Mining Pipeline



Common Machine Learning Pitfalls

- **Working with bad data**
- Data leakage
- Not understanding the target problem
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables

Dirty

Noisy

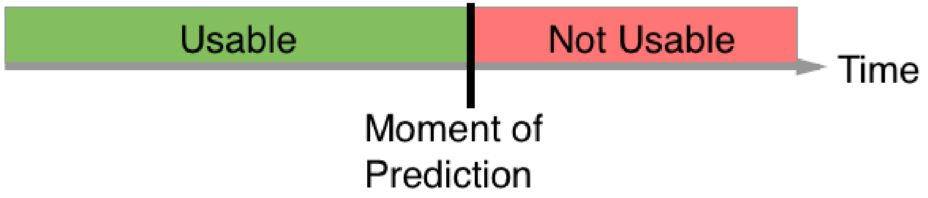
Duplicate

Biased

Sparse

Common Machine Learning Pitfalls

- Working with bad data
- **Data leakage**
- Not understanding the target problem
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables



Common Machine Learning Pitfalls

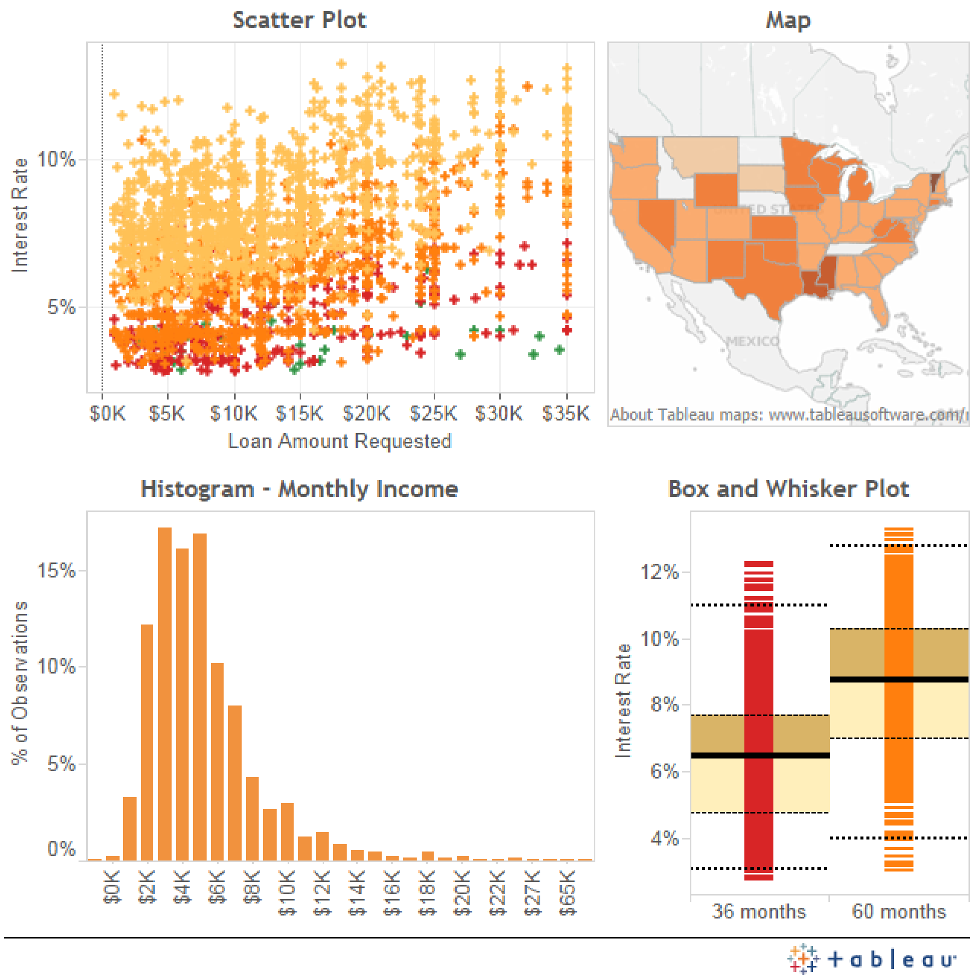
- Working with bad data
- Data leakage
- **Not defining the target problem/goals**
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables

„A problem well stated is a problem half solved."

Charles Kettering (1876-1958)

Common Machine Learning Pitfalls

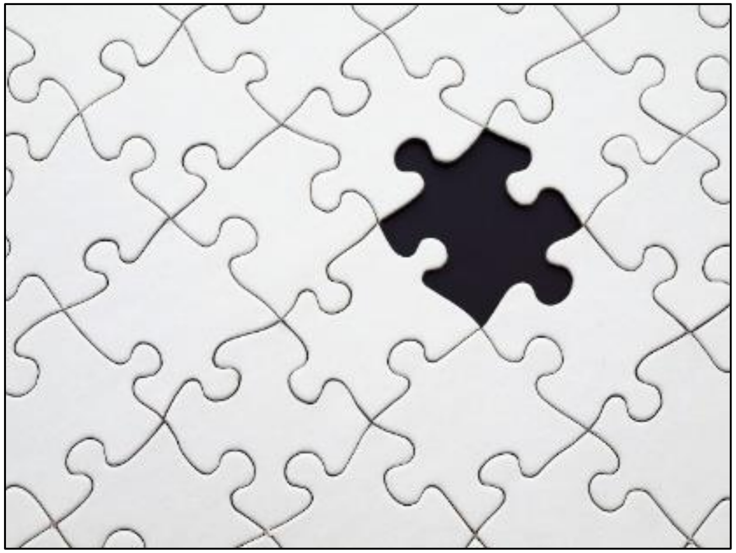
- Working with bad data
- Data leakage
- Not defining the target problem/goals
- **Ignoring exploratory analysis**
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables



Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- **Handling missing data**
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables

- Different types of 'missingness'



- Handling:
  - Removal
  - Imputation
  - Encoding as Features

Common Machine Learning Pitfalls

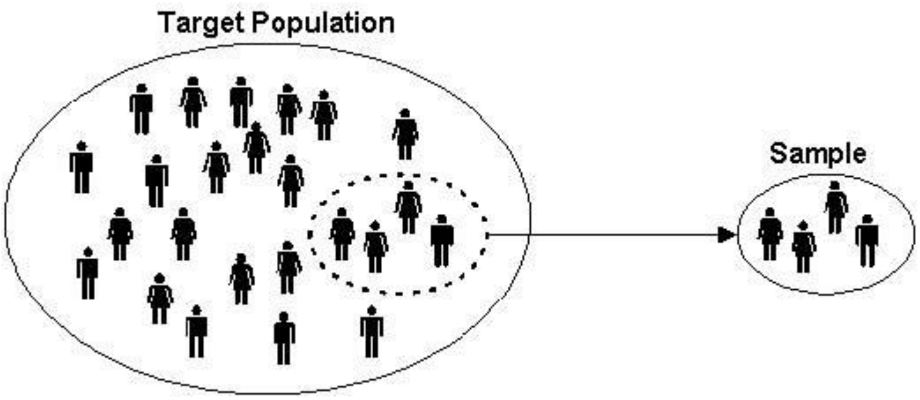
- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- **Ignoring assumptions**
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables

Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- **Representable does not imply learnable**
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables

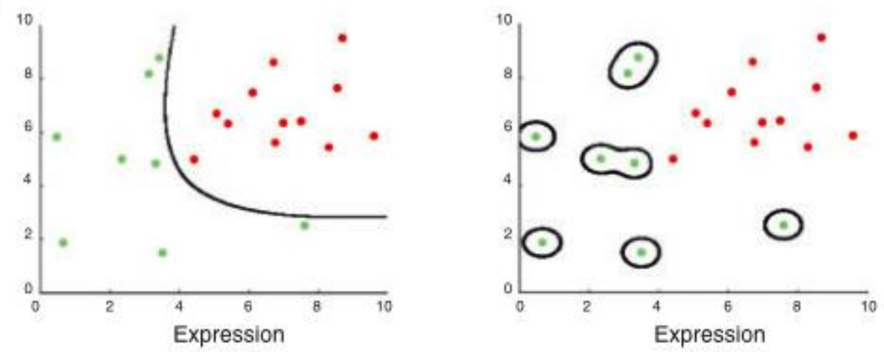
Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- **Sampling bias**
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables



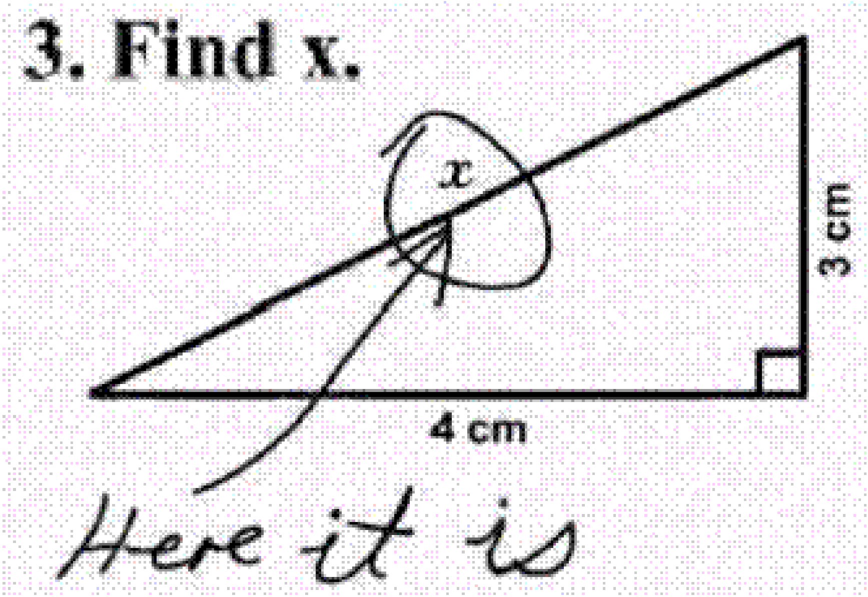
Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- **Overfitting**
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables



Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- **Simplicity does not imply better generalizability**
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables



Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- **Using the default parameters**
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables

Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- **Failing to use an appropriate evaluation metric**
- Data dredging
- Mistaking correlation for causation
- Failing to consider confounding variables

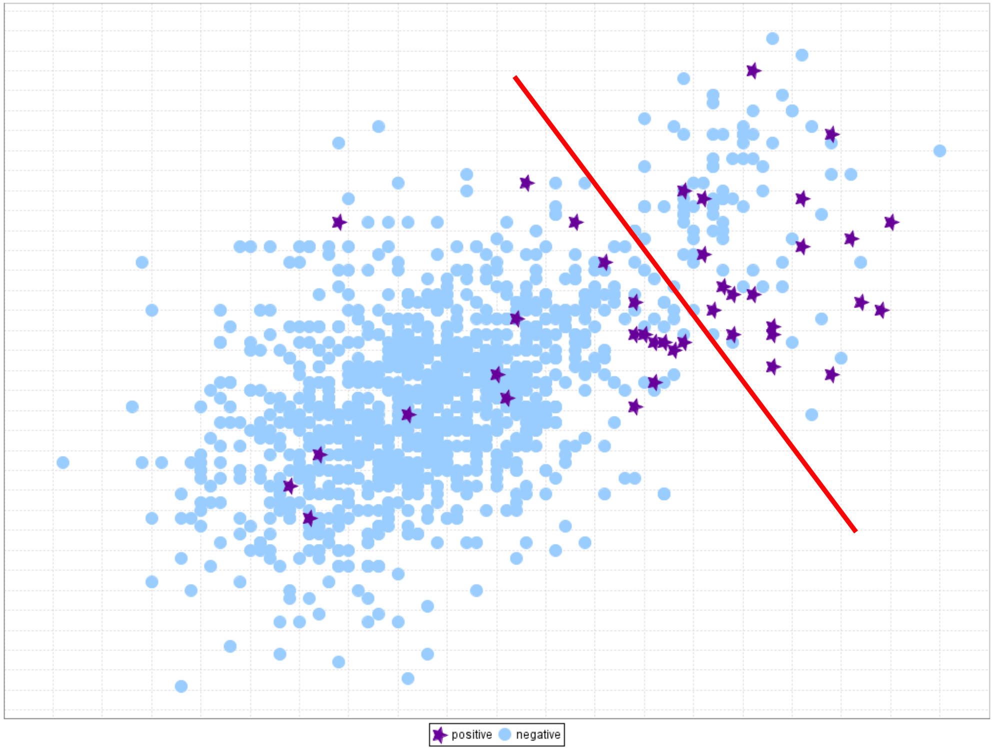
Accuracy

F1 Score

Recall

AUC

Precision



Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- **Data dredging**
- Mistaking correlation for causation
- Failing to consider confounding variables

> If you torture the data long enough, it will confess.
>
> — *Ronald Coase* —

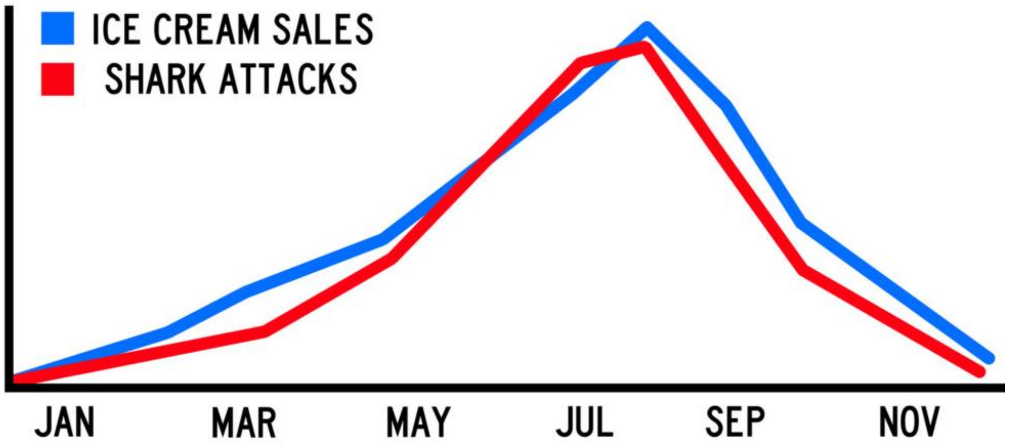
Data Fishing

Data Snooping

P-hacking

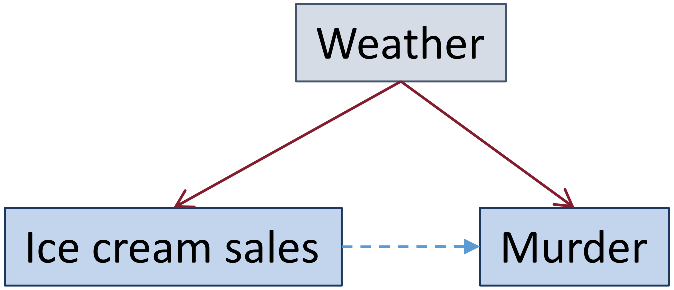
Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- **Mistaking correlation for causation**
- Failing to consider confounding variables



Common Machine Learning Pitfalls

- Working with bad data
- Data leakage
- Not defining the target problem/goals
- Ignoring exploratory analysis
- Handling missing data
- Ignoring assumptions
- Representable does not imply learnable
- Sampling bias
- Overfitting
- Simplicity does not imply better generalizability
- Using the default parameters
- Failing to use an appropriate evaluation metric
- Data dredging
- Mistaking correlation for causation
- **Failing to consider confounding variables**

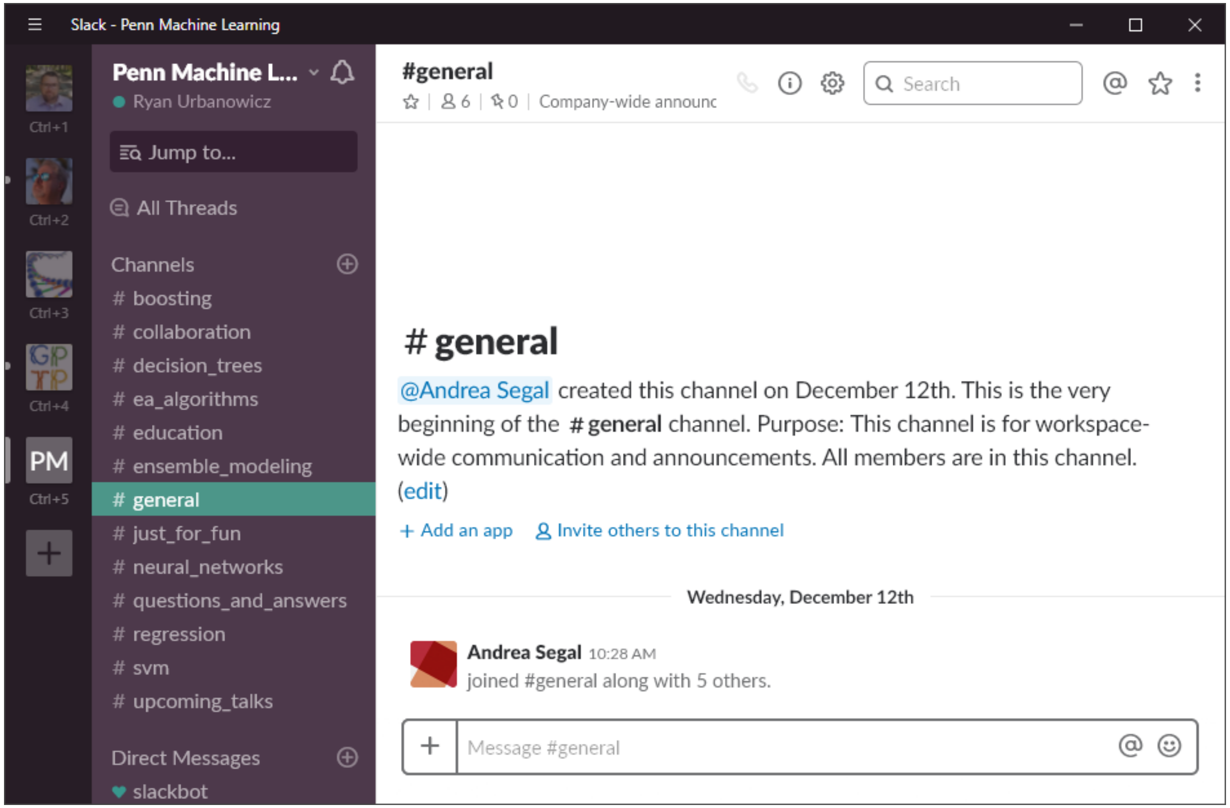


Where do we go from here?

- Data preparation
- How do different ML methods work?
- Feature selection
- Selecting run parameters
- Software/code to run ML
- Evaluation and statistical analysis
- Ensemble learning
- Model interpretation

UPenn ML slack

- Penn Machine Learning – Slack Workspace
- pennmachinelearning.slack.com



Acknowledgements and Funding

- Pennsylvania Commonwealth Universal Research Enhancement Program (CURE)

Penn LDI
Leonard Davis Institute of Health Economics

Department of Biostatistics Epidemiology & Informatics

Center for Health Incentives & Behavioral Economics

Institute for Biomedical Informatics