INTRODUCTION TO MACHINE LEARNING IN R

Amit Sharma

Sr Director

Architecture and Data Science (ADP)

About me

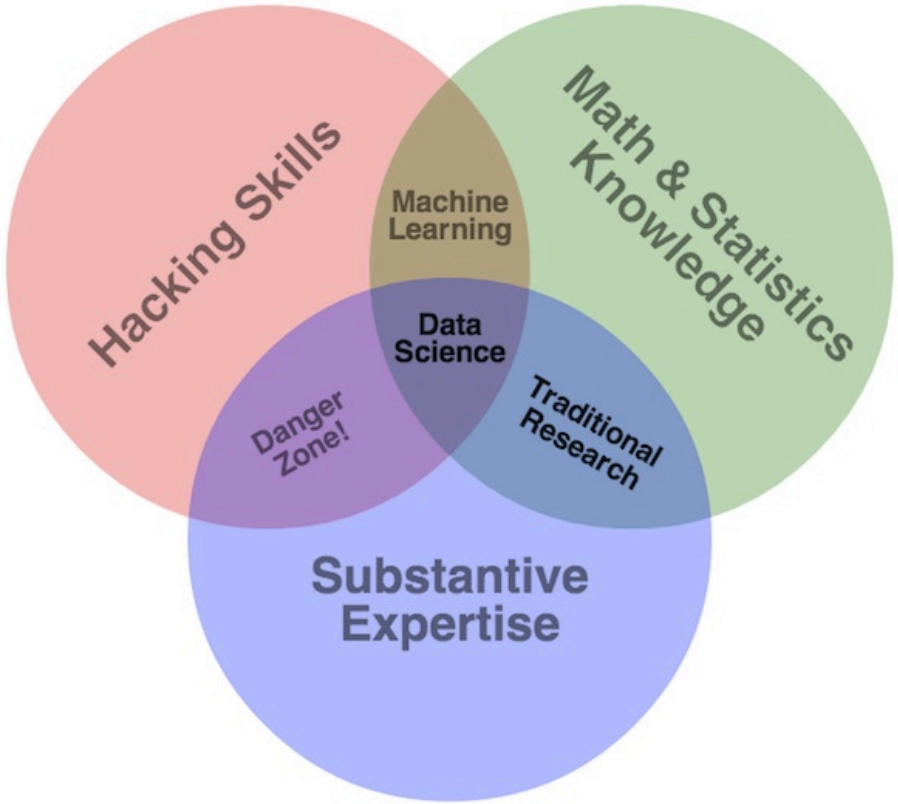
- BTech from NIT, Rourkela
  - Electronics and Instrumentation Engineering
- 17+ years of experience in Product Development
- Java/JEE Design and Architecture
- Data Science
- Speaker in many forums
  - NASSCOM
  - Institute of Product Leadership
  - Zinnov

amit_sharma_hyd 

Amit Sharama (ADP) 

Amit.Sharma.Hyd@gmail.com 

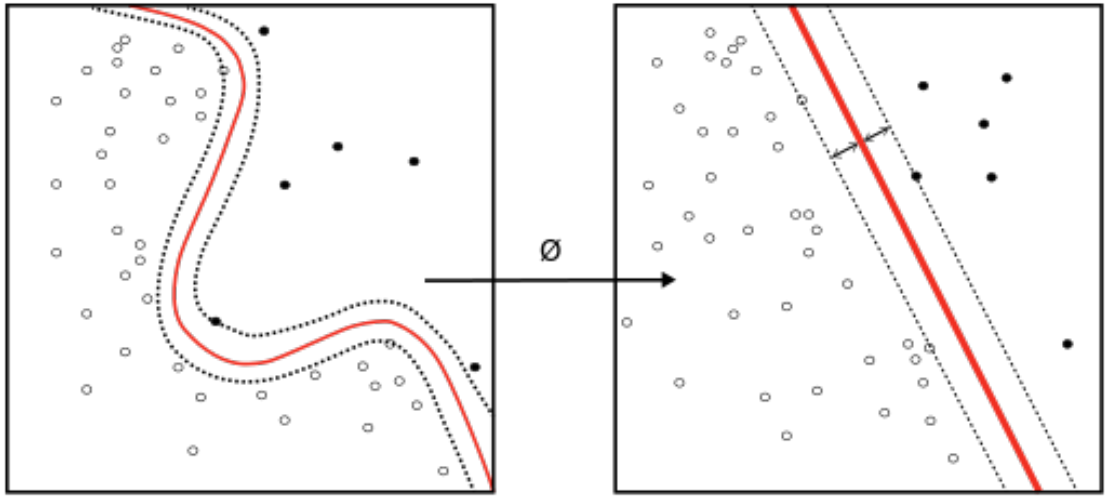
Data Scientist



http://b-i.forbesimg.com/gilpress/files/2013/05/Data_Science_VD.png

What is Machine Learning

- Field of Artificial Intelligence
- Evolved from Pattern Recognition
- Overlaps with Computational Statistics



https://en.wikipedia.org/wiki/Machine_learning#/media/File:Kernel_Machine.svg

Types of Machine Learning

- Classification
- Regression
- Optimization

Supervised

- Clustering
- Association
- Dimensionality reduction

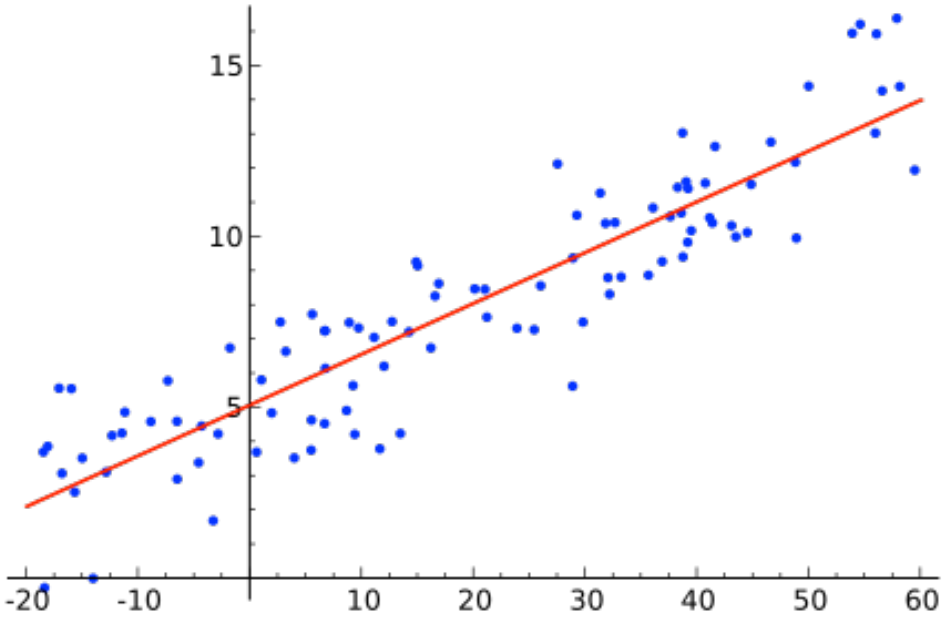
Unsupervised

- Recommenders
- Collaborative Filtering

Reinforcement

Linear Regression

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

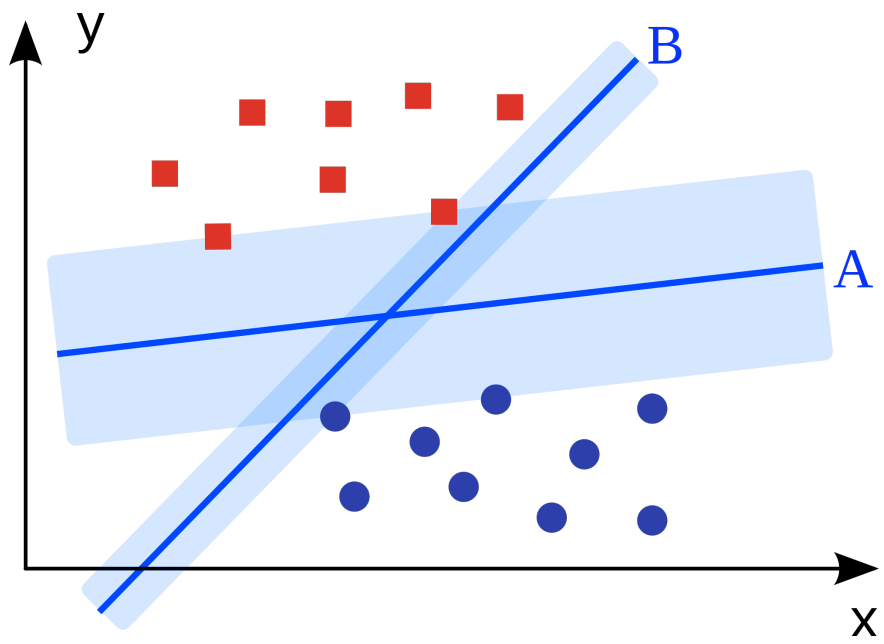


https://upload.wikimedia.org/wikipedia/commons/3/3a/Linear_regression.svg

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
```

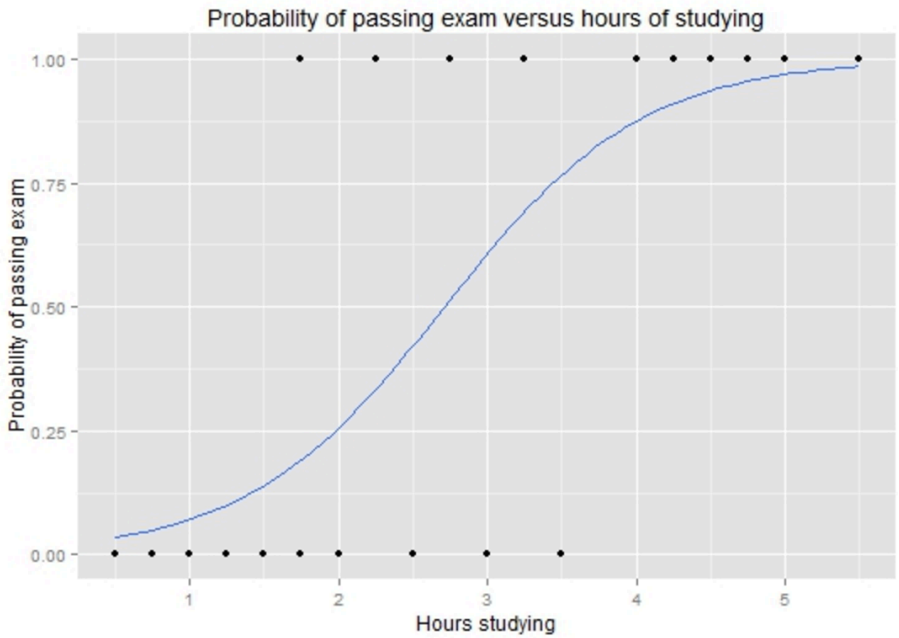
Classification - Linear

- Which is a better classifier – A or B



Logistic Regression

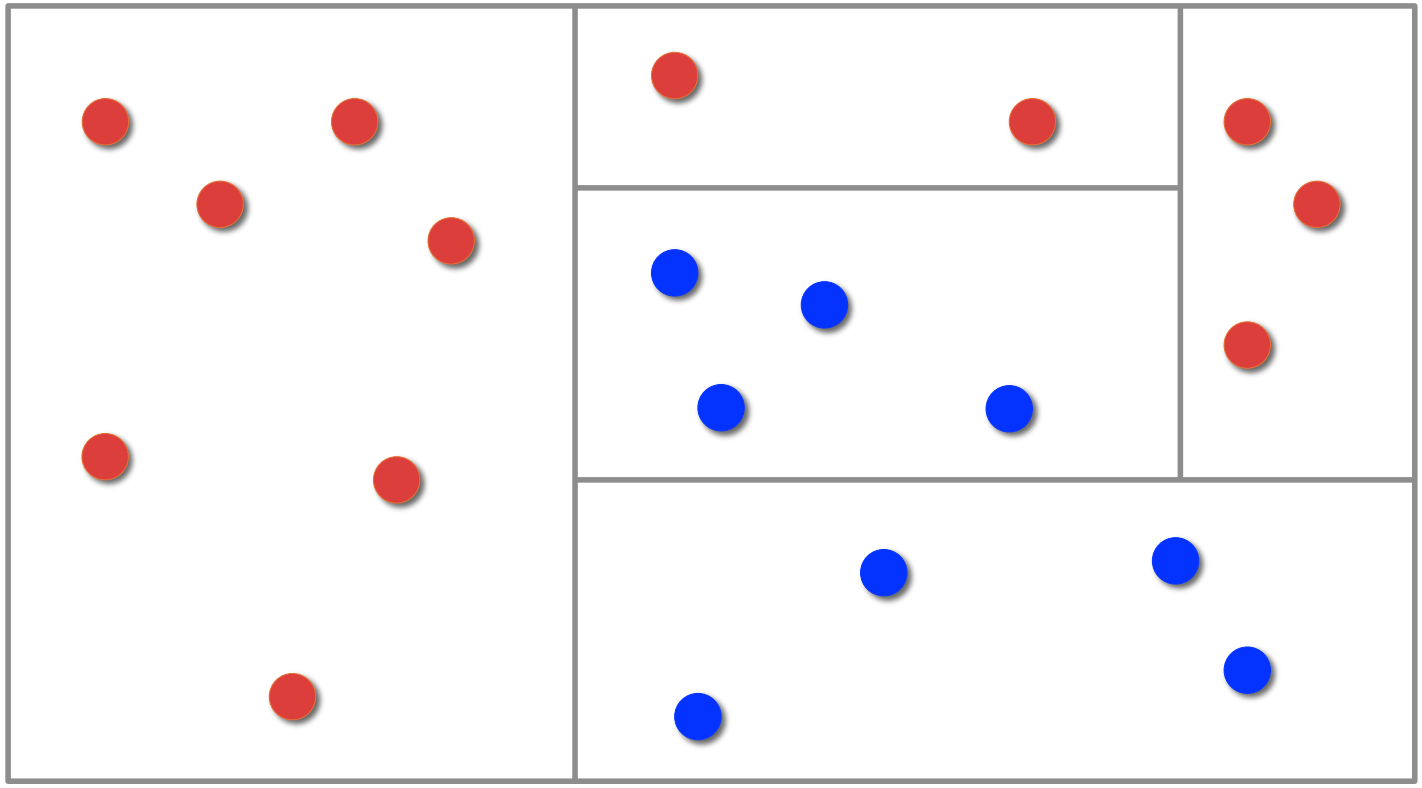
$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



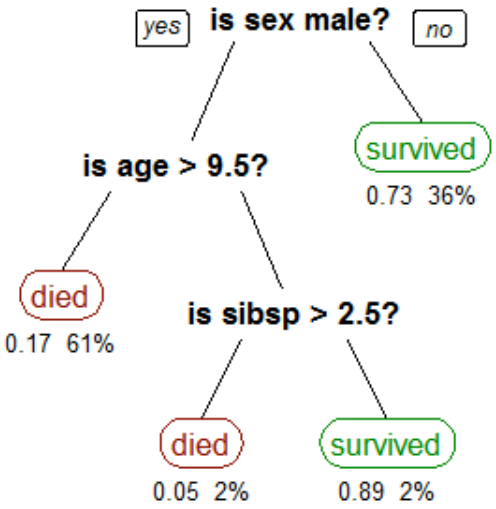

```
model <- glm(Survived ~.,family=binomial(link='logit'),data=train)
```

Classification – Non-Linear

- How would you separate the red from blue dots



Decision Tree

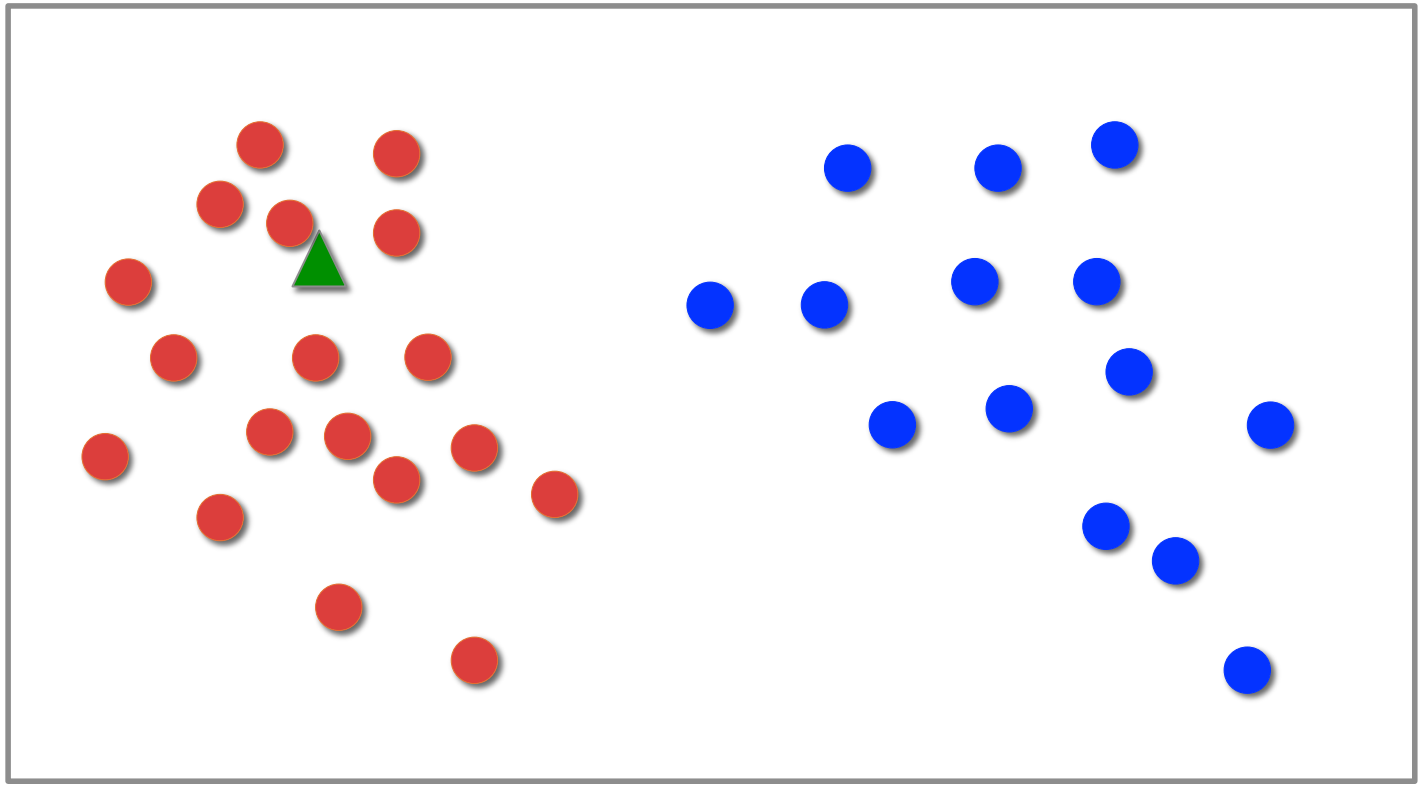


https://upload.wikimedia.org/wikipedia/commons/f/f3/CART_tree_titanic_survivors.png

```
fit <- rpart(Mileage~Price + Country + Reliability + Type,
   method="anova", data=cu.summary)
```

Similarity based

- K-nearest neighbors



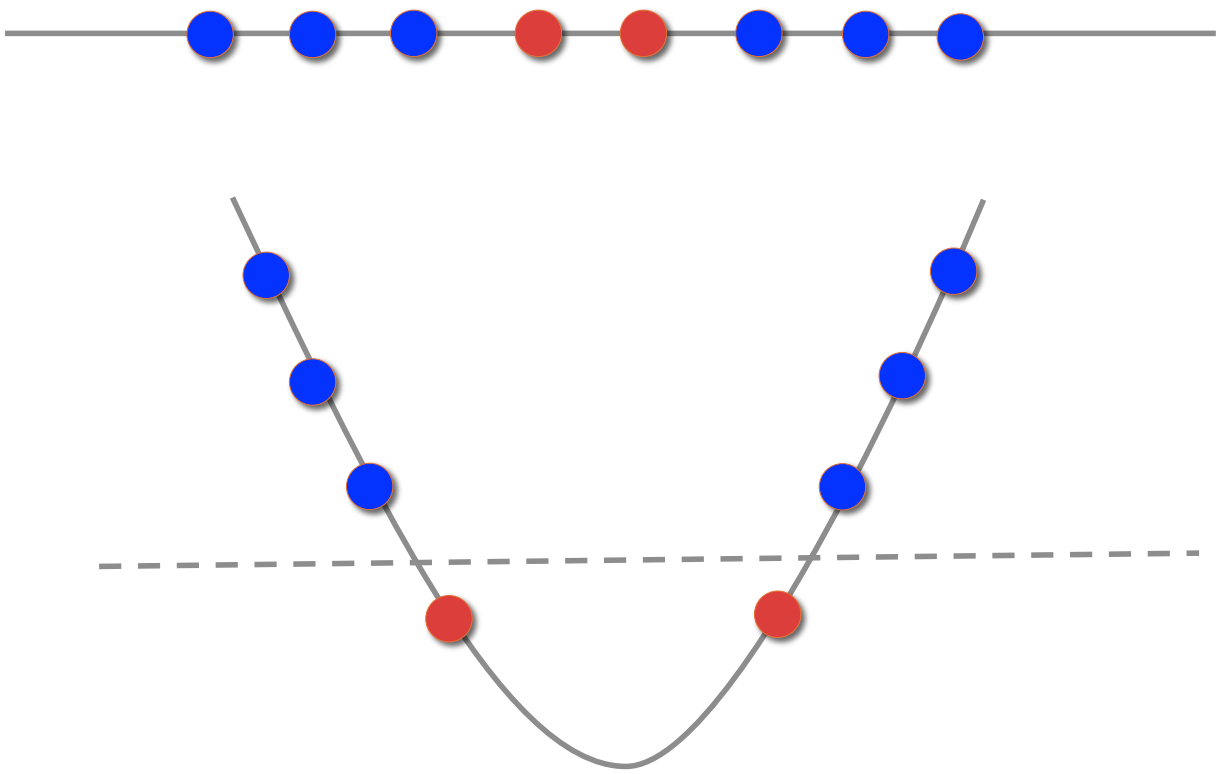
Bayesian

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

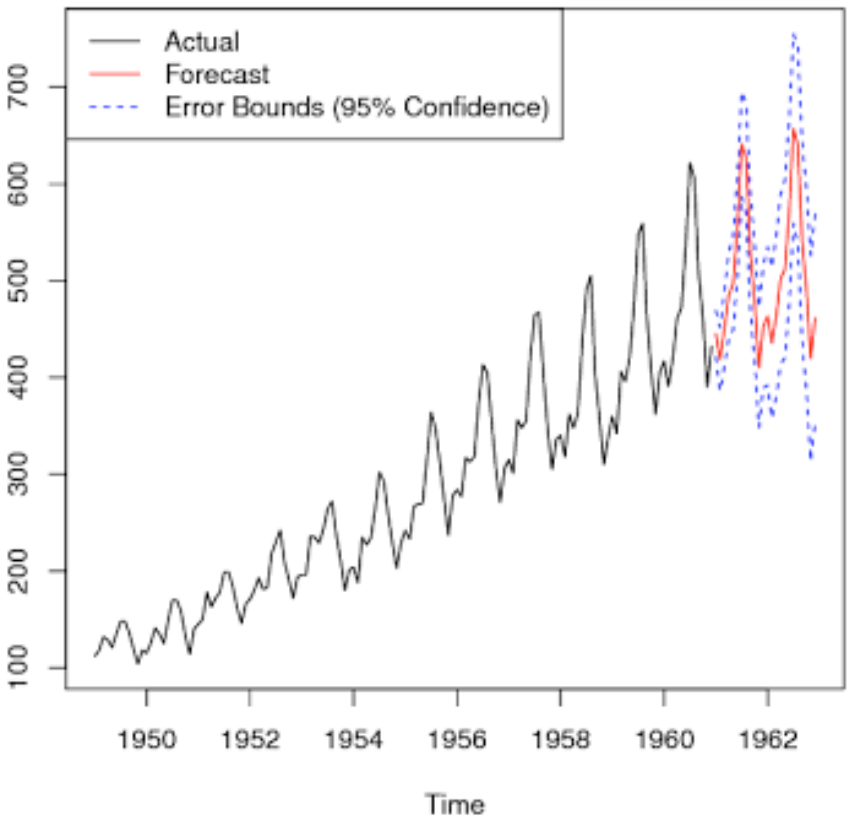
$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

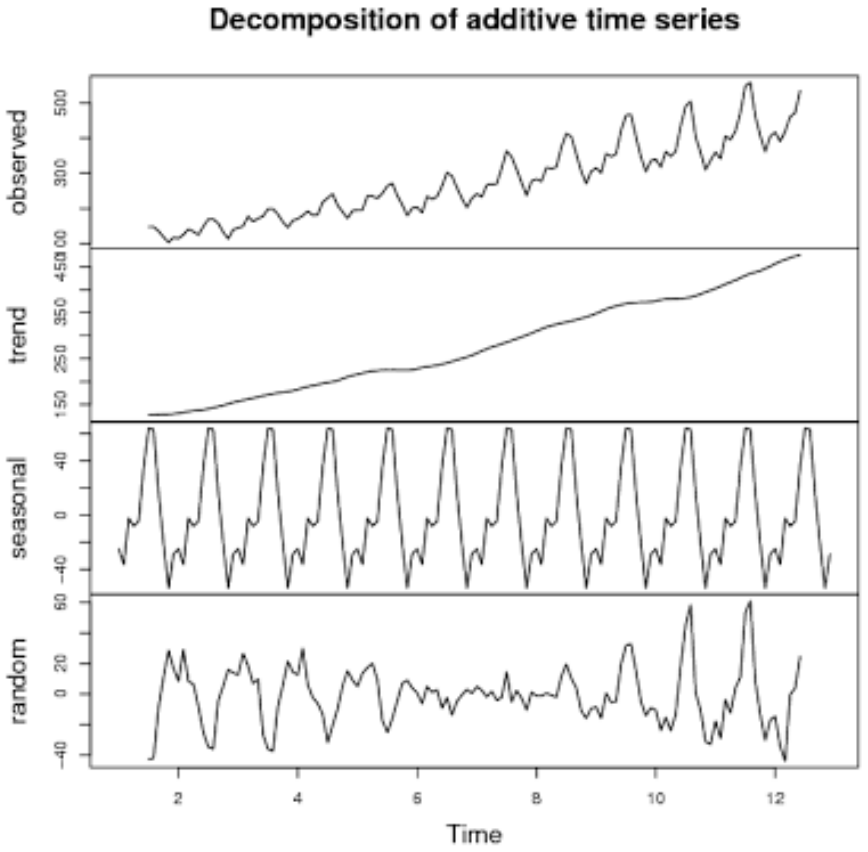
Support Vector Machines

- Mapping data to a higher dimension for linear classification



Time Series Analysis

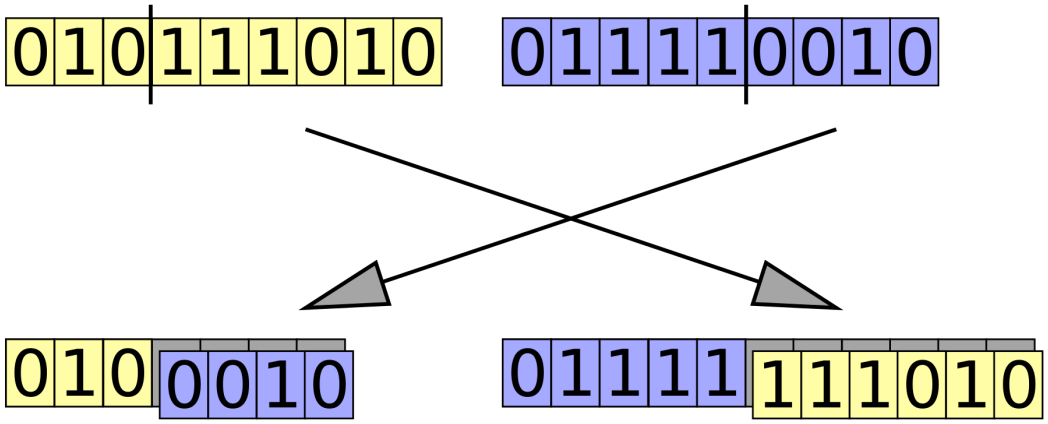




http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html

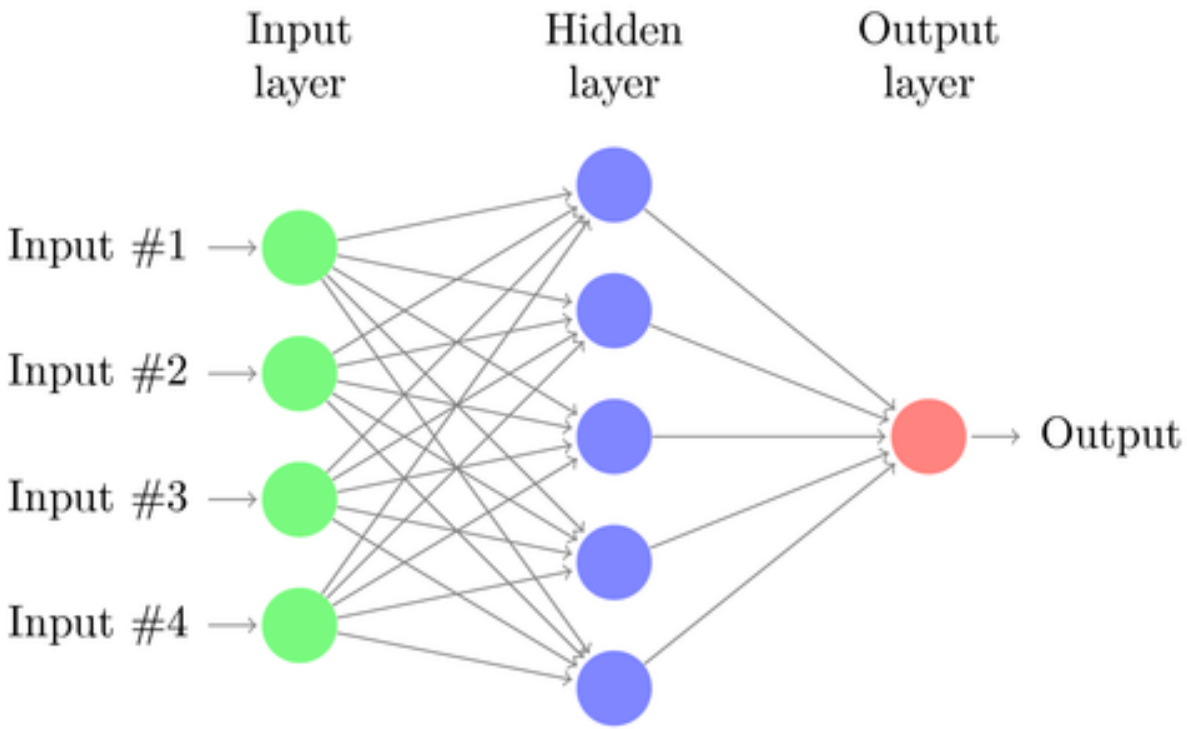
Genetic Algorithm

- Goodness measure
- Cross
- Mutate



http://creationwiki.org/Genetic_algorithm

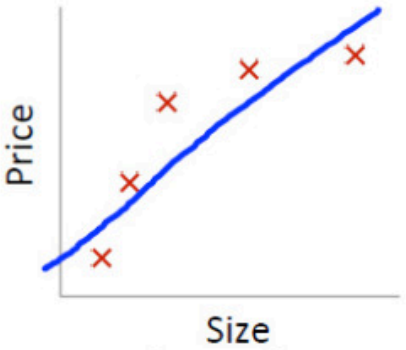
Neural Networks



http://www.texample.net/media/tikz/examples/PNG/neural-network.png

```
library(neuralnet)
```

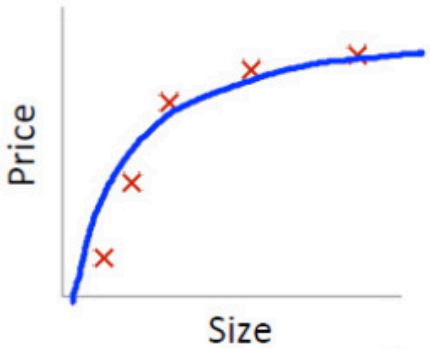
Bias and Variance



$\theta_0 + \theta_1 x$

High bias
(underfit)

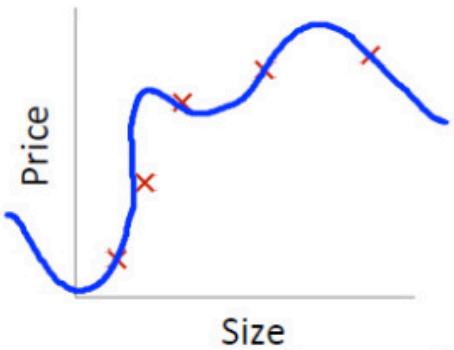
$d = 1$



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

$d = 2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

$d = 4$

http://bigdataexaminer.com/data-science/understanding-bias-and-variance-using-python/

Validating models

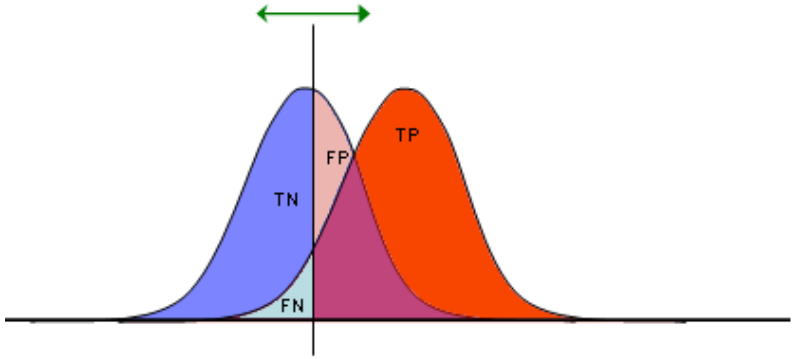
- Linear
  - R-sq
  - P-value
  - Confidence Interval
- Classification
  - Confusion Matrix
  - Sensitivity/Recall
  - Specificity

Confusion Matrix

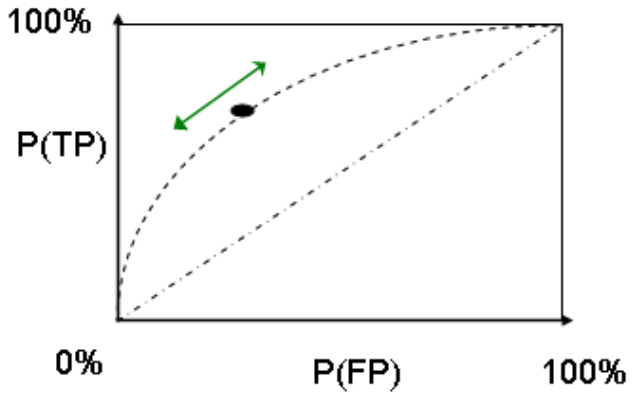
		CONDITION determined by "Gold Standard"	
	TOTAL POPULATION	CONDITION POS	CONDITION NEG
TEST OUT-COME	TEST POS	True Pos TP	*Type I Error* False Pos FP
	TEST NEG	*Type II Error* False Neg FN	True Neg TN

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

ROC Curve



TP	FP
FN	TN
1	1



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

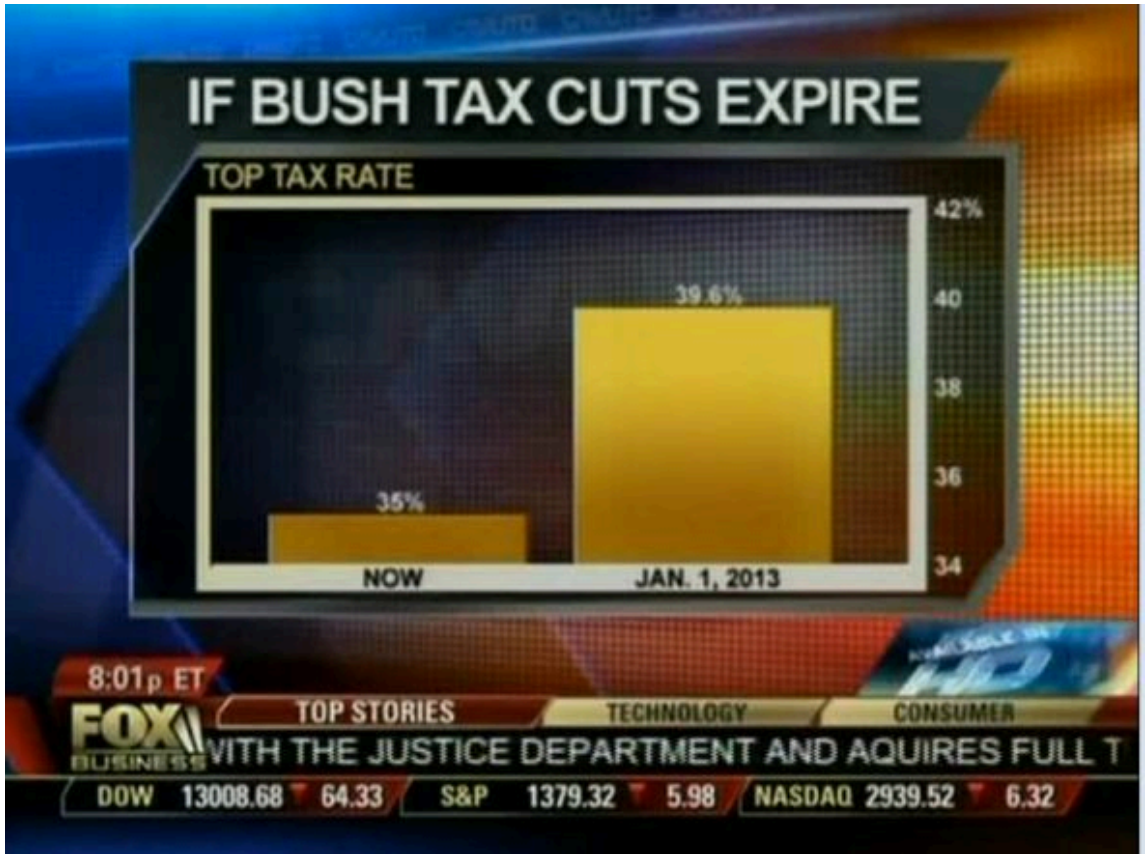
Model Goodness

- Simpler one wins
- Explicability
- Actionability

Visualization

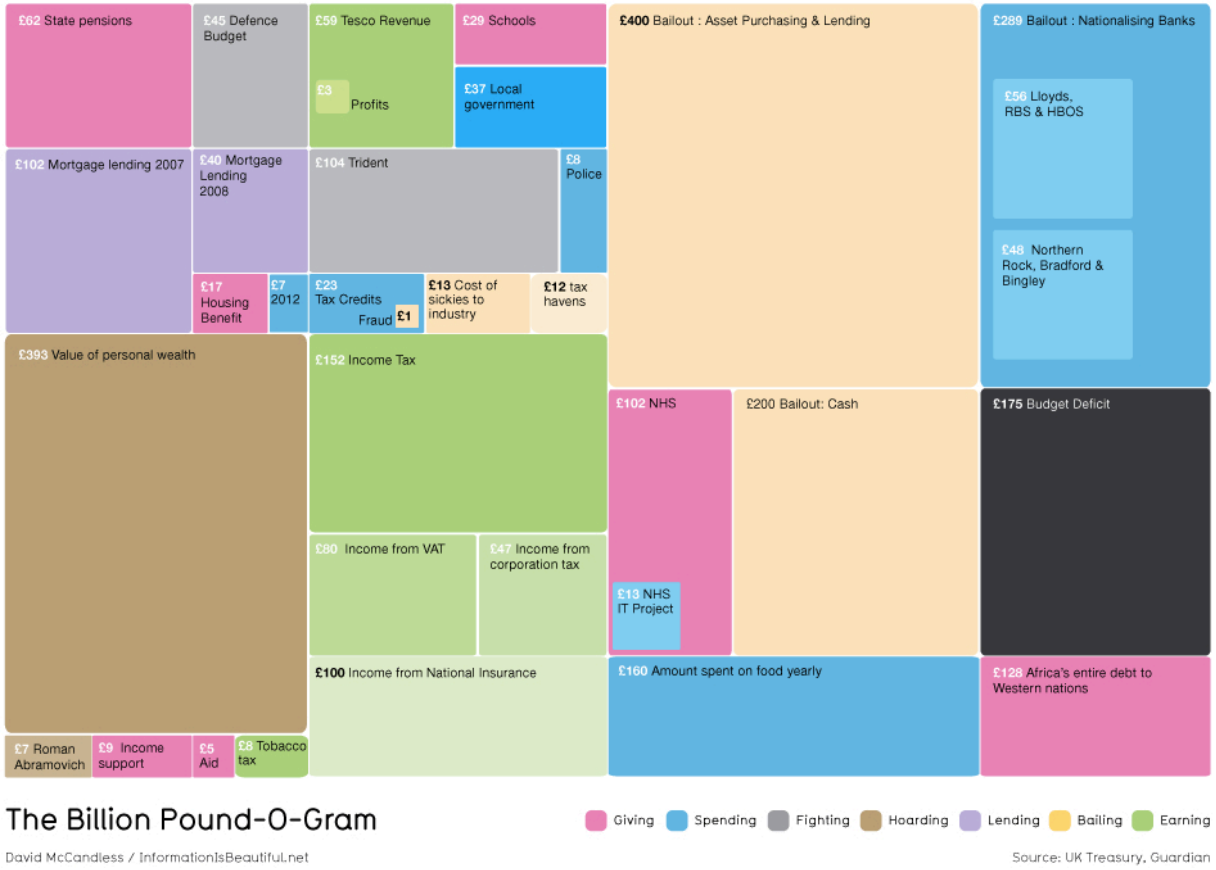
- Clarity of message/story
- Simplicity
- Truthful/Integrity (Not Misleading)

Visualization Example 1



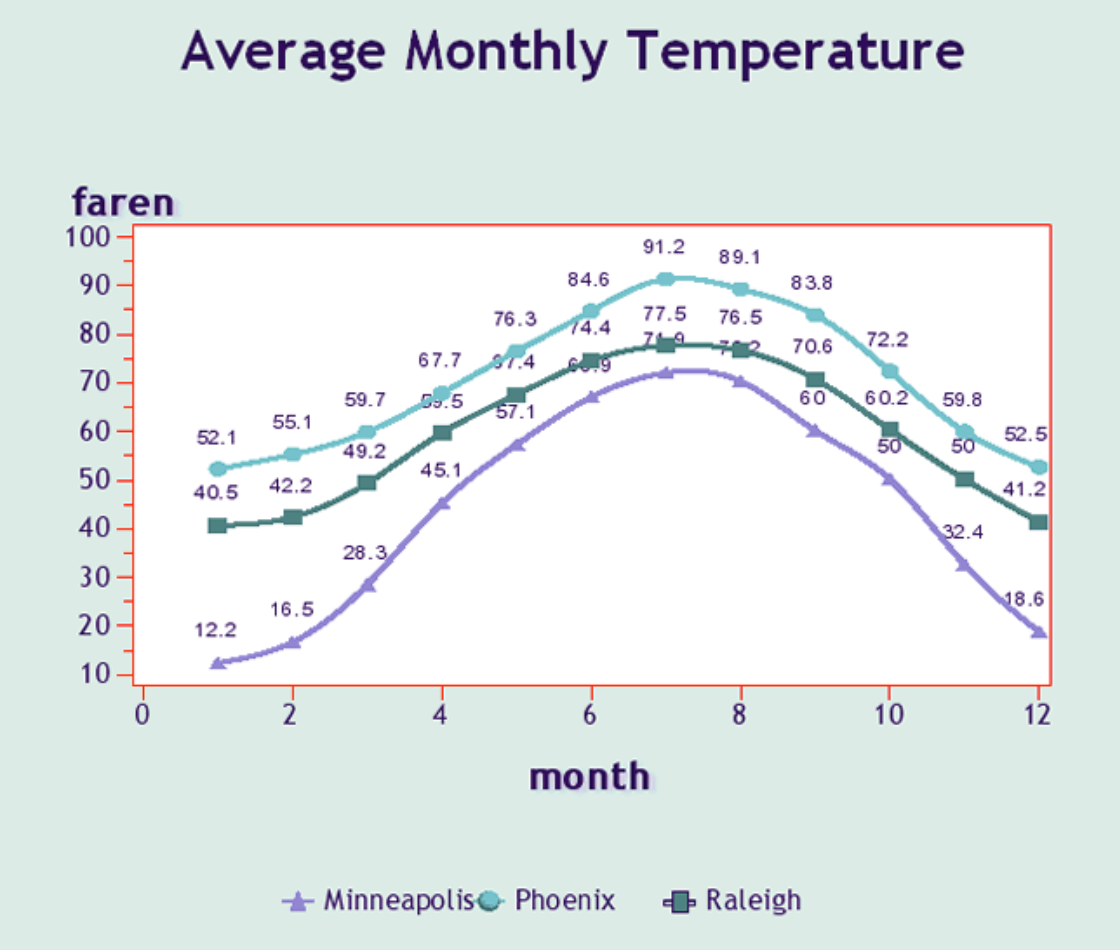
http://mediamatters.org/research/2012/10/01/a-history-of-dishonest-fox-charts/190225

Visualization Example 3



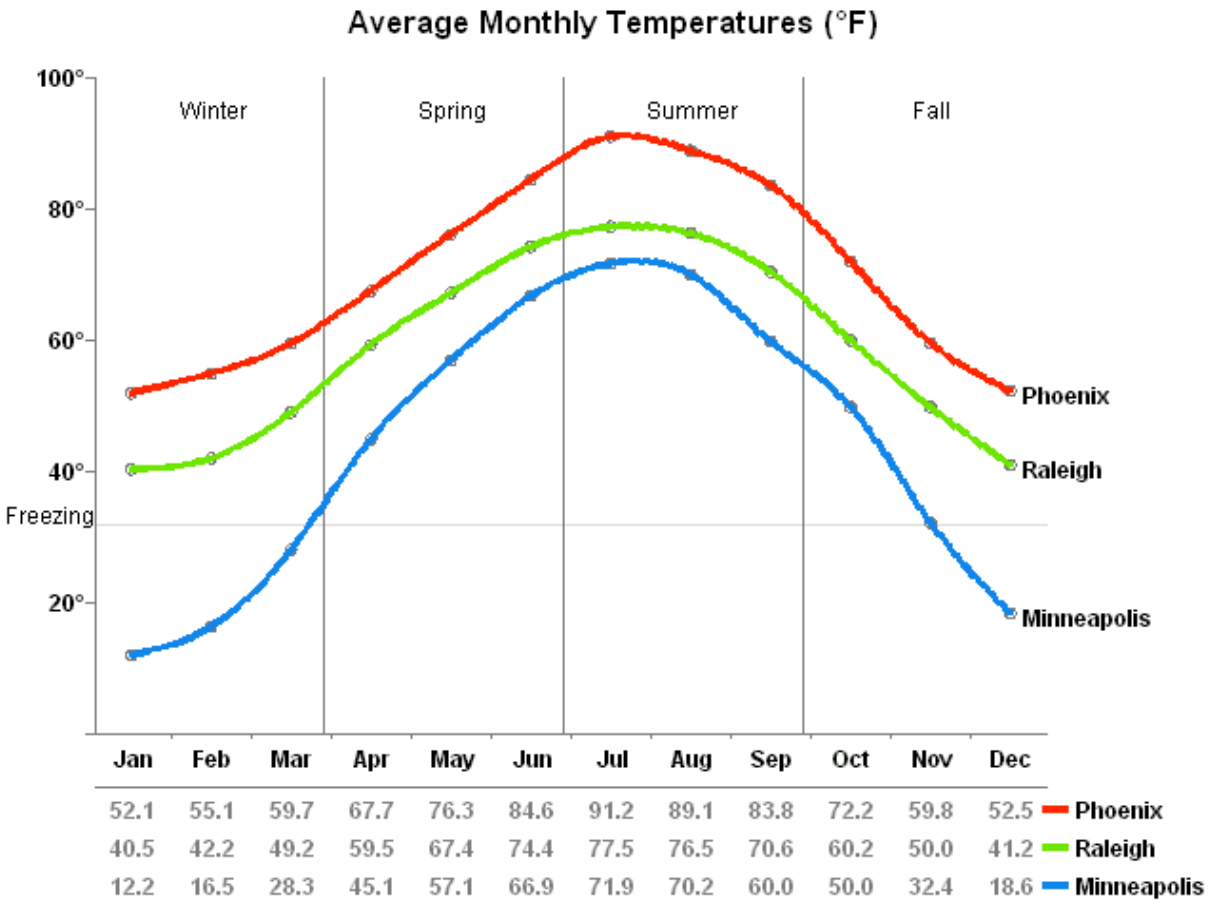
https://www.perceptualedge.com/example19.php

Visualization Example 4



https://www.perceptualedge.com/example19.php

Visualization Example 5



https://www.perceptualedge.com/example19.php