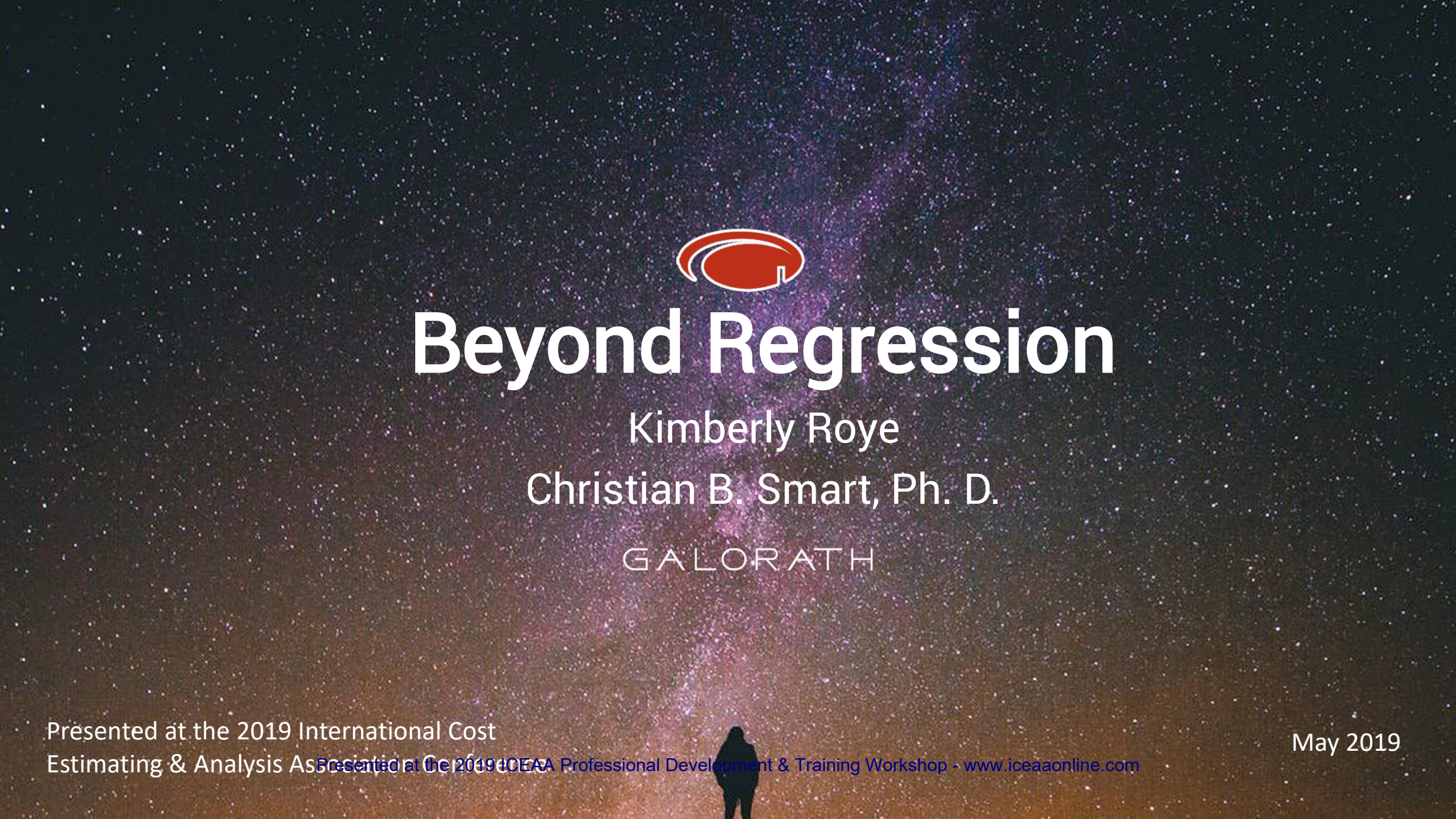


# Beyond Regression

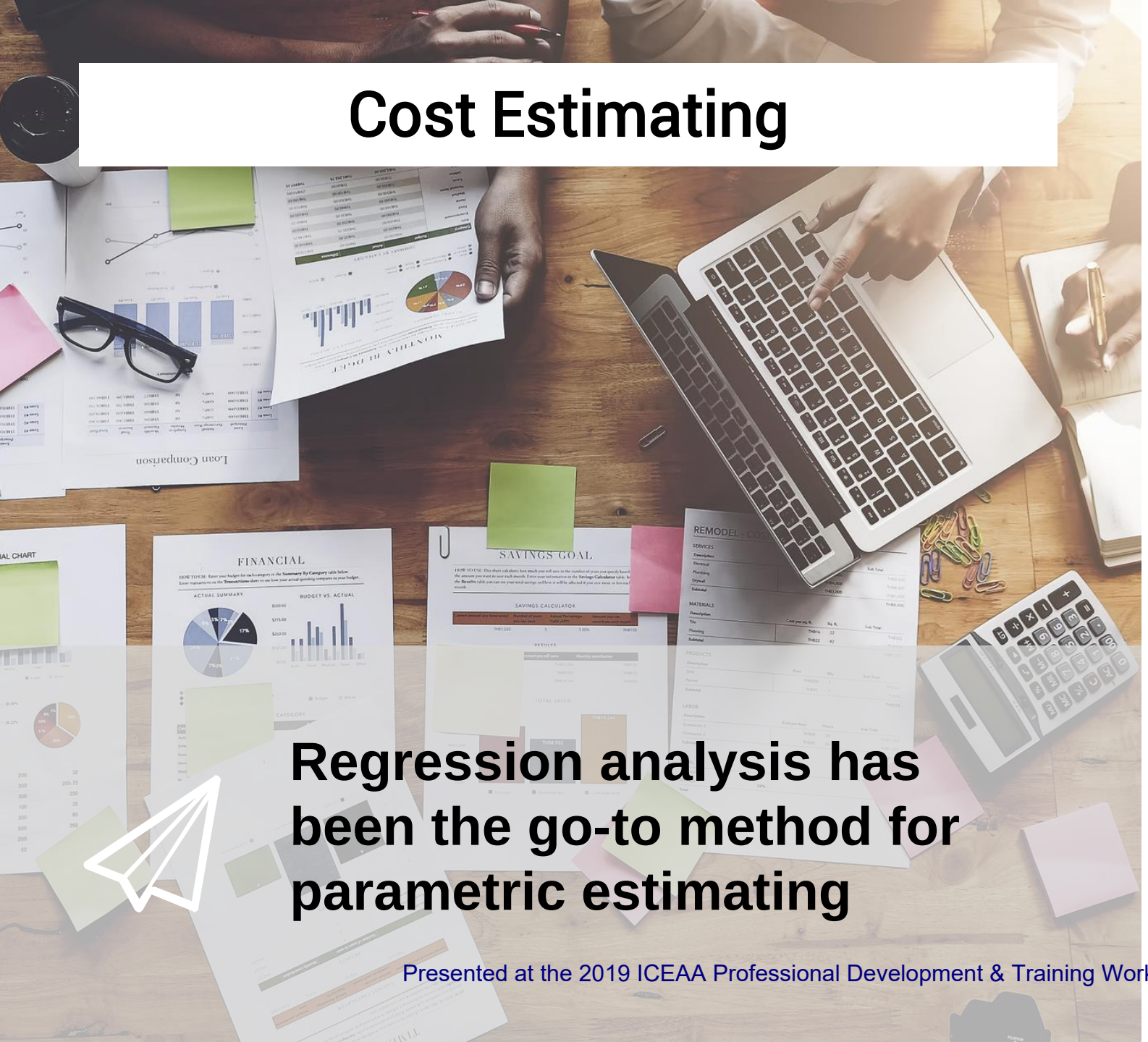
Kimberly Roye
Christian B. Smart, Ph. D.
GALORATH

Presented at the 2019 International Cost Estimating & Analysis Association Conference



May 2019

# Cost Estimating



**Regression analysis has been the go-to method for parametric estimating**

## 1 Traditional Method

Regression analysis is one tool of many in data science

## 2 Alternative Methods

A variety of supervised methods can be explored for predictive cost analysis

## 3 Review of Methods

A cross-sectional example was used to illustrate similarities and difference of the techniques

## 4 Pros and Cons

The advantages and disadvantages of each method is discussed

# Least Squares and Regression Analysis

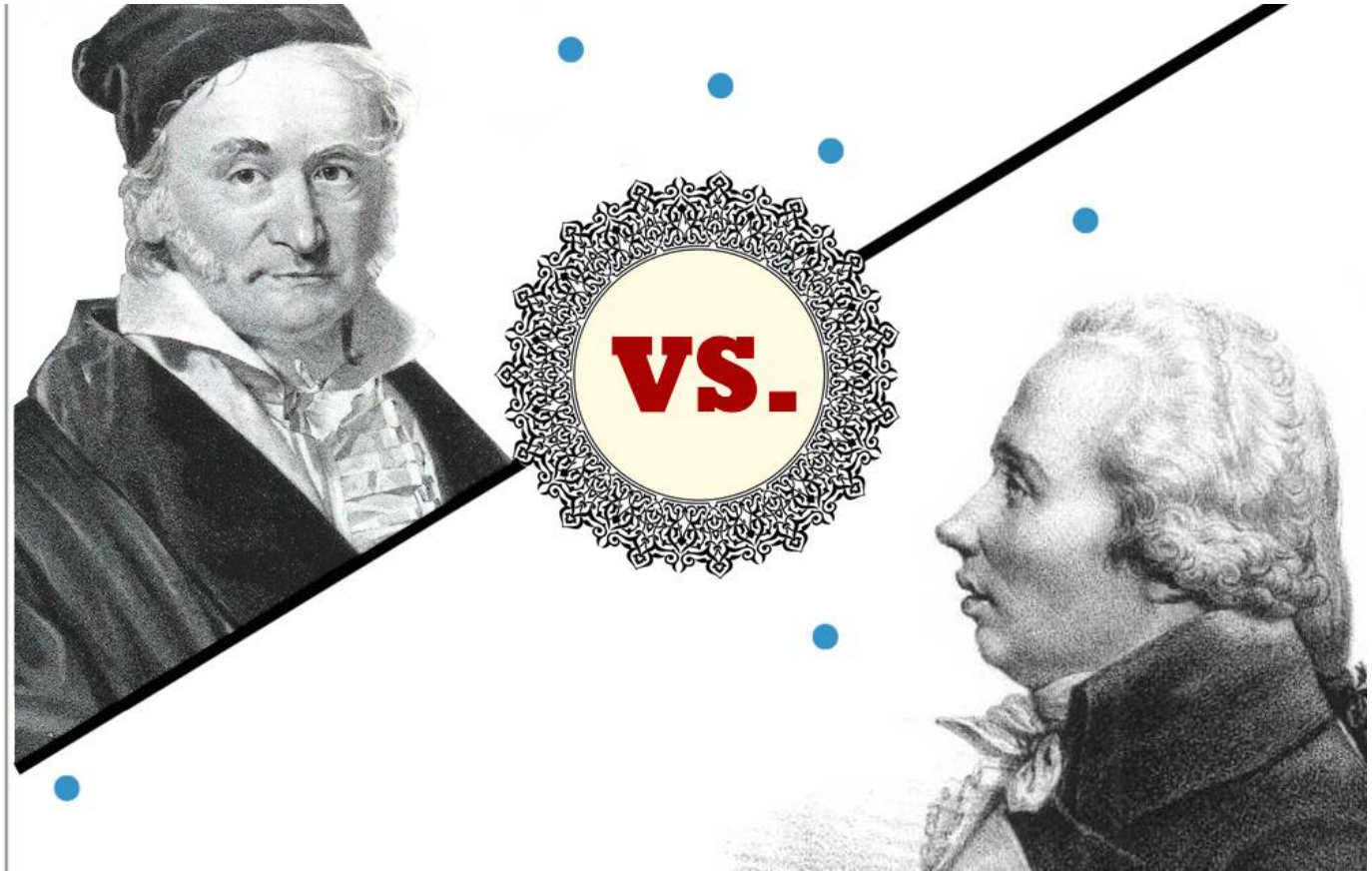
## History

The method of least squares was originally used to predict the orbits of heavenly bodies using observed data. Francis Galton applied the technique to find linear predictive relationships between various phenomena, such as relationships between the heights of fathers and sons.

Given the linear equation of the form $Y = a + bX$ and a set of data $(X_1, Y_1)$, $(X_2, Y_2)$,..., $(X_n, Y_n)$, the residuals are defined as $\varepsilon_i = Y_i - (a + bX_i) = Actual - Estimated$

The estimated cost linear regression finds the "best fit" by finding the parameters, a and b, that minimize the sum of the squares of the residuals

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2 = \sum_{i=1}^{n} (Actual_i - Estimated_i)^2$$




**Least Squares method was first developed by mathematicians Legendre and Gauss in the early 19th century**

## Nonlinear Regression

**In the spacecraft** and defense industry, nonlinear relationships between cost and cost drivers are most common

## Newer Techniques

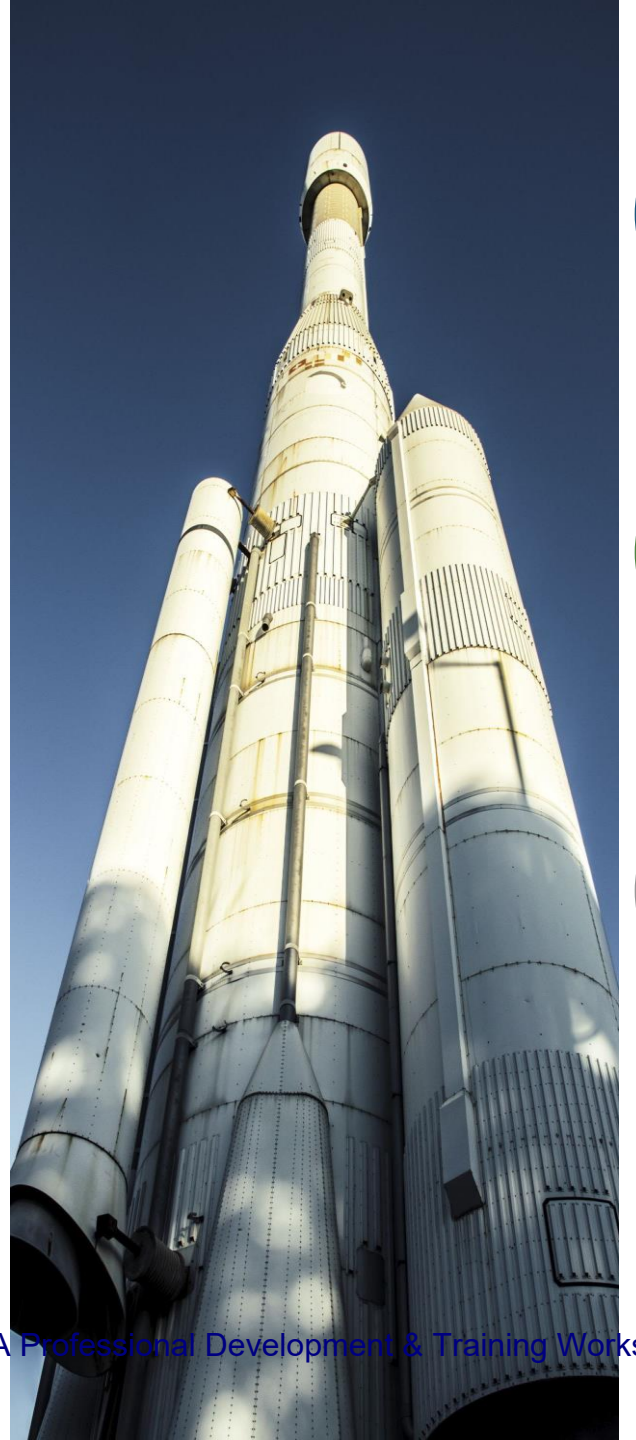
**"Newer"** techniques exist for predictive analysis

## Data Scientist

**Harvard Business Review** declared data scientist as the "sexiest job of the 21st century"

## New Terms

**Data science** and machine learning are new terms for the application of mathematics and statistics to solve real-world problems



## Application

**The key factor** that differentiates data science and machine learning from statistics is the application of computer science

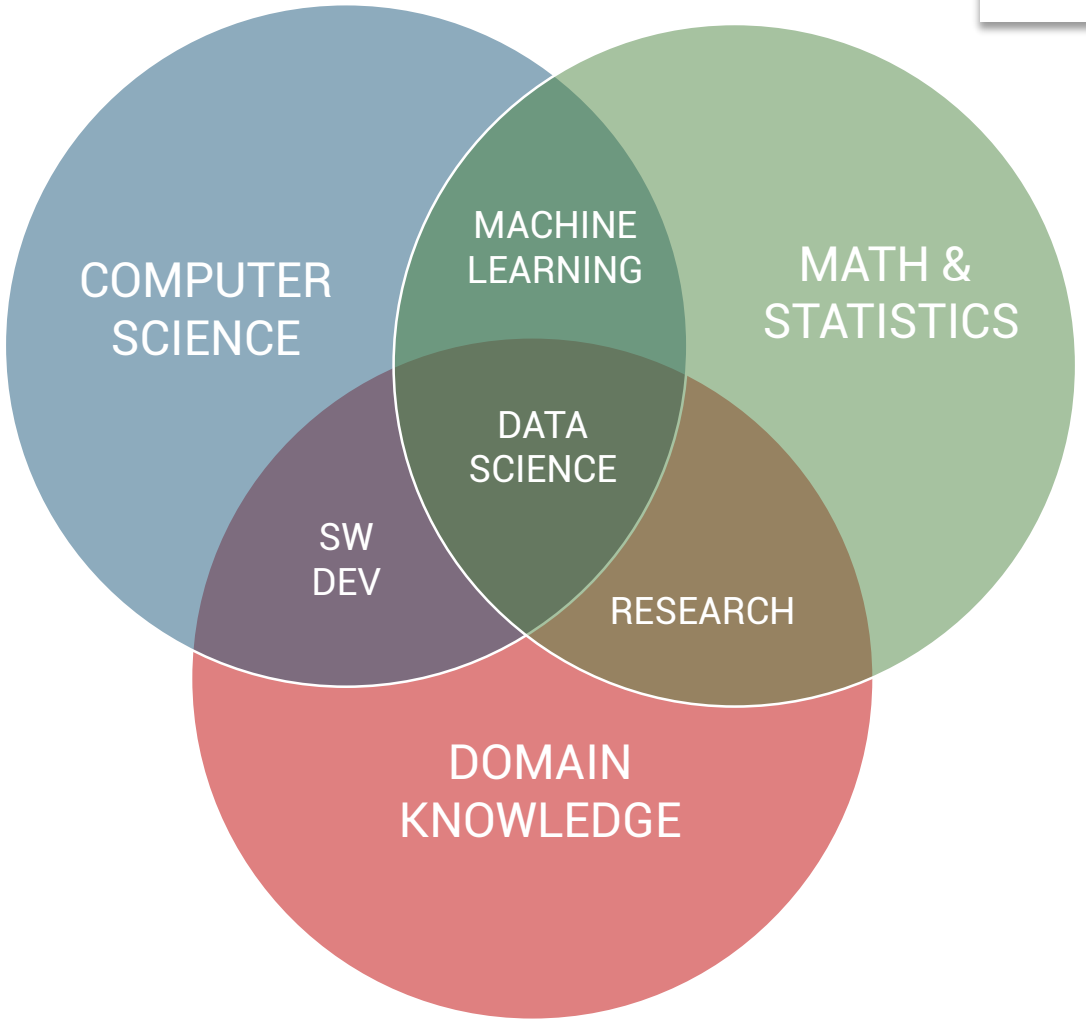
## Computational Complexity

**Computer scientists** have made significant contributions to the field of statistics

## Machine Learning

**Modern statistical methods** are taught in computer science programs under the rubric "machine learning"

# Relationships Between Topics




## Data Science

**Data science** is where statistics, computer science and domain knowledge meet

## Machine Learning

**Machine learning** is the culmination of statistics and computer science

# Big Data

The Hype and the Reality

01

## "Hype"

There is lots of "hype" surrounding the potential for machine learning techniques to supplant traditional methods

02

## "The End of Theory"

In 2008, Wired Magazine claimed big data makes the scientific method obsolete

03

## Similar Claims

In the 1990s, similar claims were made about neural networks

04

## Hype ≠ Reality

Big data, machine learning, and artificial intelligence have much to offer but these methods aren't replacing traditional methods, especially within the government

05

## Transparency

"Curve fitting without benefit of a model is notoriously a black art" (Newell and Rosenbloom 1981)

06

## Small 'n'

In cost estimating, we need to get more data before we have to worry about "big data"

# Categories of Machine Learning

## Supervised Learning

**Inputs and outputs** are labeled

Example: Regression Analysis

## Unsupervised Learning

**Only outputs** are labeled

Example: Cluster Analysis

## Reinforcement Learning

**Focus** is on actions that maximize a reward function

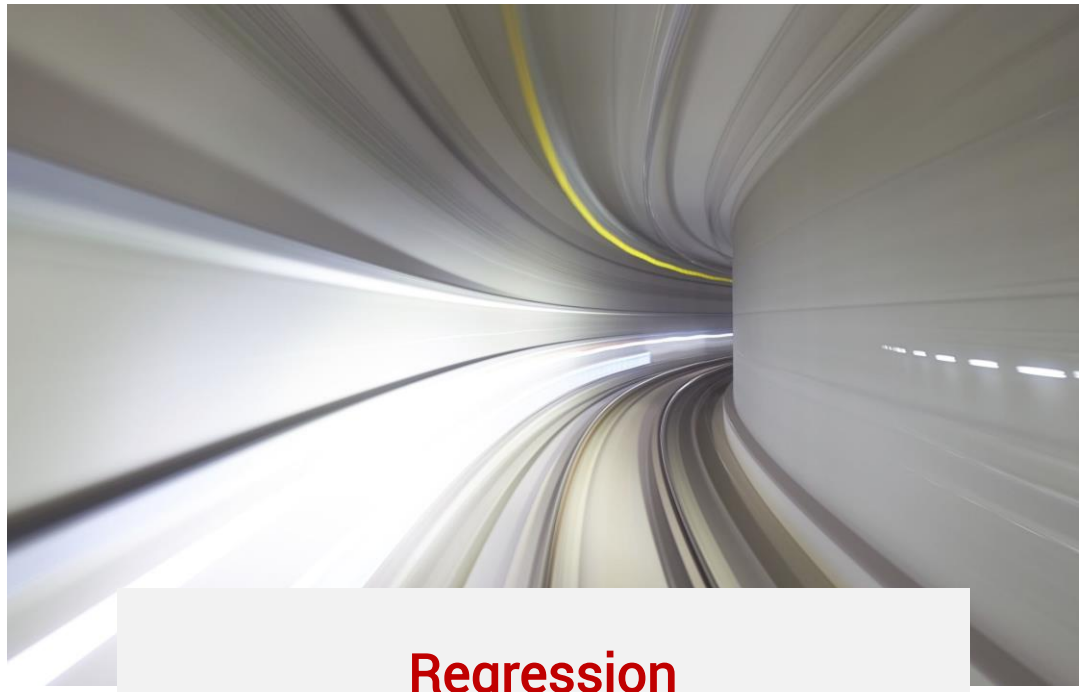
Example: Dynamic Programming

## Our Focus

**In this presentation,** we will focus on supervised learning techniques

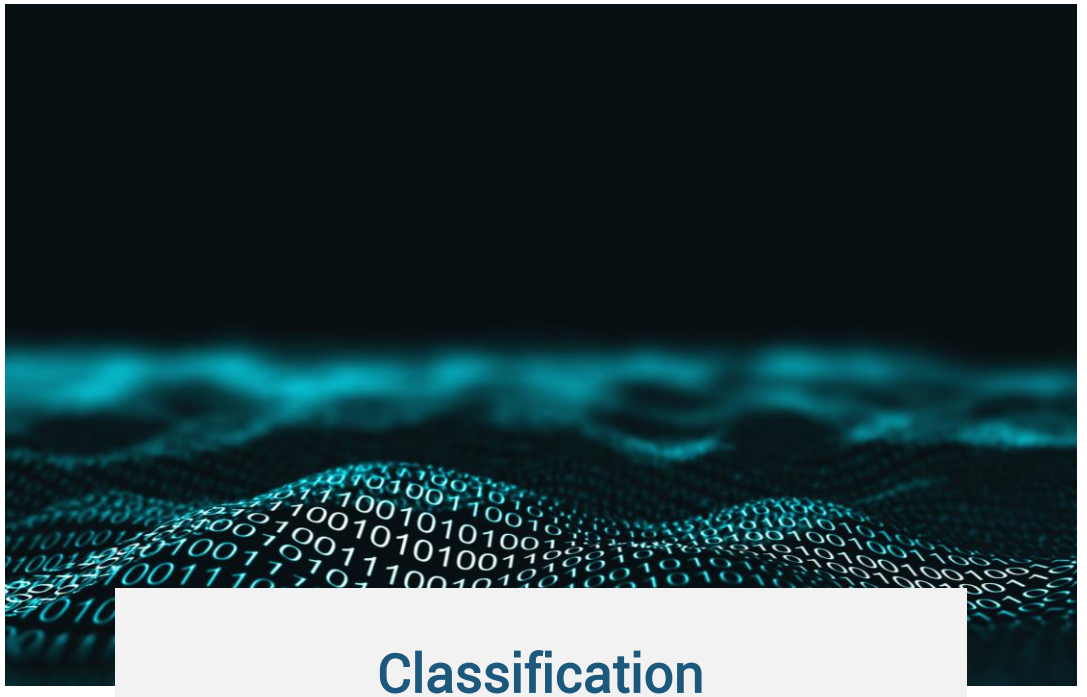
# Supervised Learning Techniques

Using input variables and an output variable along with an algorithm to learn the mapping function from the input to the output



## Regression

**When the output variable** is a real value, such as dollars or weight, we use methods appropriate for continuous variables

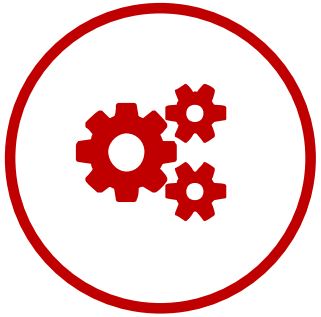


## Classification

**When the output variable** is a category, such as "disease" and "no disease", we focus on classification methods

"I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail."

- Abraham Maslow, the Psychology of Science



## K-Nearest Neighbor

**Non-parametric** method using weighted analogies to predict a numerical target based on a similarity measure

## Neural Networks

**Predictive set of algorithms** that can be trained to recognize patterns that are numerical, contained in vectors, or included in real-world data

## Regression Trees

**Method** that focuses on partitioning data into smaller regions where interactions are manageable

## Random Forest

**Ensemble approach** using decision trees to create a random forest. This method combines the estimates of multiple trees to produce an average

"I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail."

- Abraham Maslow, the Psychology of Science

## Support Vector Machines

**Originating in the 1990s,** this method seeks to maximize the margin between classes, which yield maximally robust classification

## Text Analytics

**Method** that explores large quantities of textural data and identifies patterns

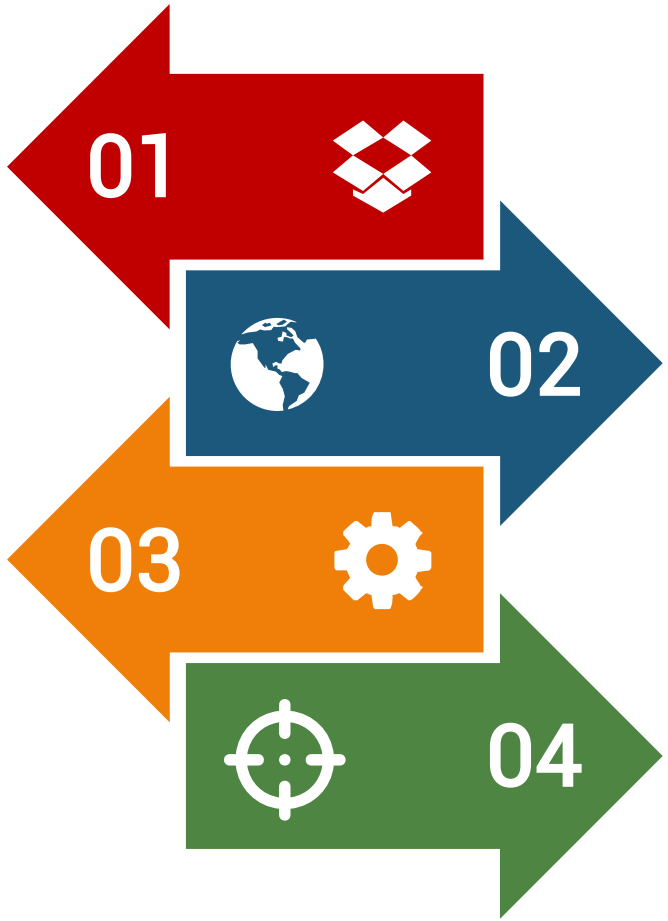
# K-Nearest Neighbors (KNN)

Simple method using analogies most commonly for classification of data, but also for predicting a numerical target

## STEP ONE

**Calculate** the Euclidean distance for the continuous variables

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$




## STEP TWO

**Rank** the data points from lowest to greatest distance

## STEP THREE

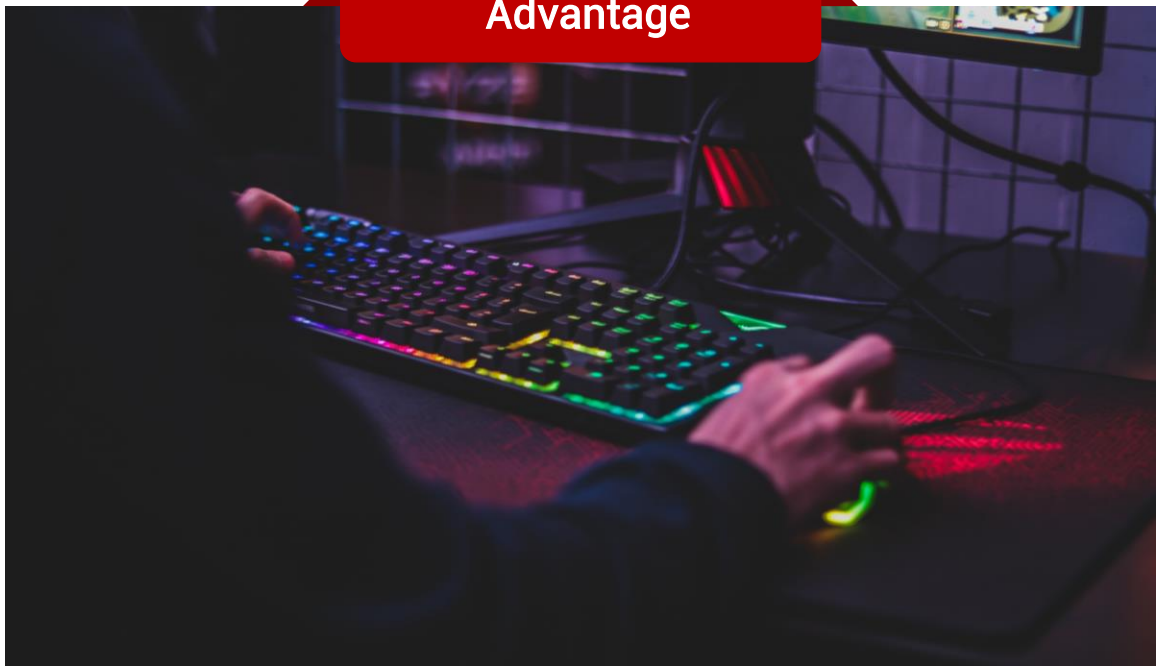
**Choose value** for 'k' and select the 'k' nearest data points to inputs being used to estimate the new dependent variable

## STEP FOUR

**Calculate** the average of the outputs for the 'k' nearest data points to determine the value for the new data point

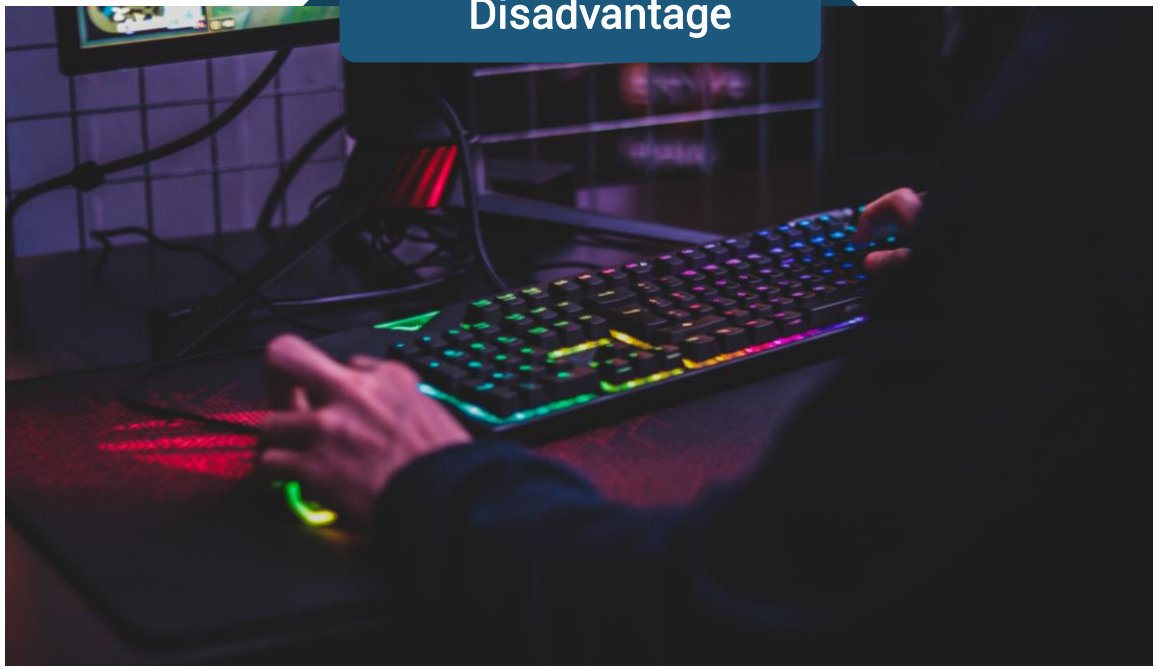
# KNN Advantages and Disadvantages

**Advantage**



**One advantage** of KNN technique is the method's simplicity. This method also provides predictions when conventional regression methods are unable to produce significant estimating relationships

**Disadvantage**



**A disadvantage** is the frequency of overfitting data by choosing an improper value of k

# Neural Networks

Simplified models of brain networks introduced in 1944 by Warren McCollough and Walter Pitts

## Neural Net

**The net** is composed of processing nodes.

## Flow

**As input enters** the network, weights score the input and map it to a set of guesses the network makes at the end.

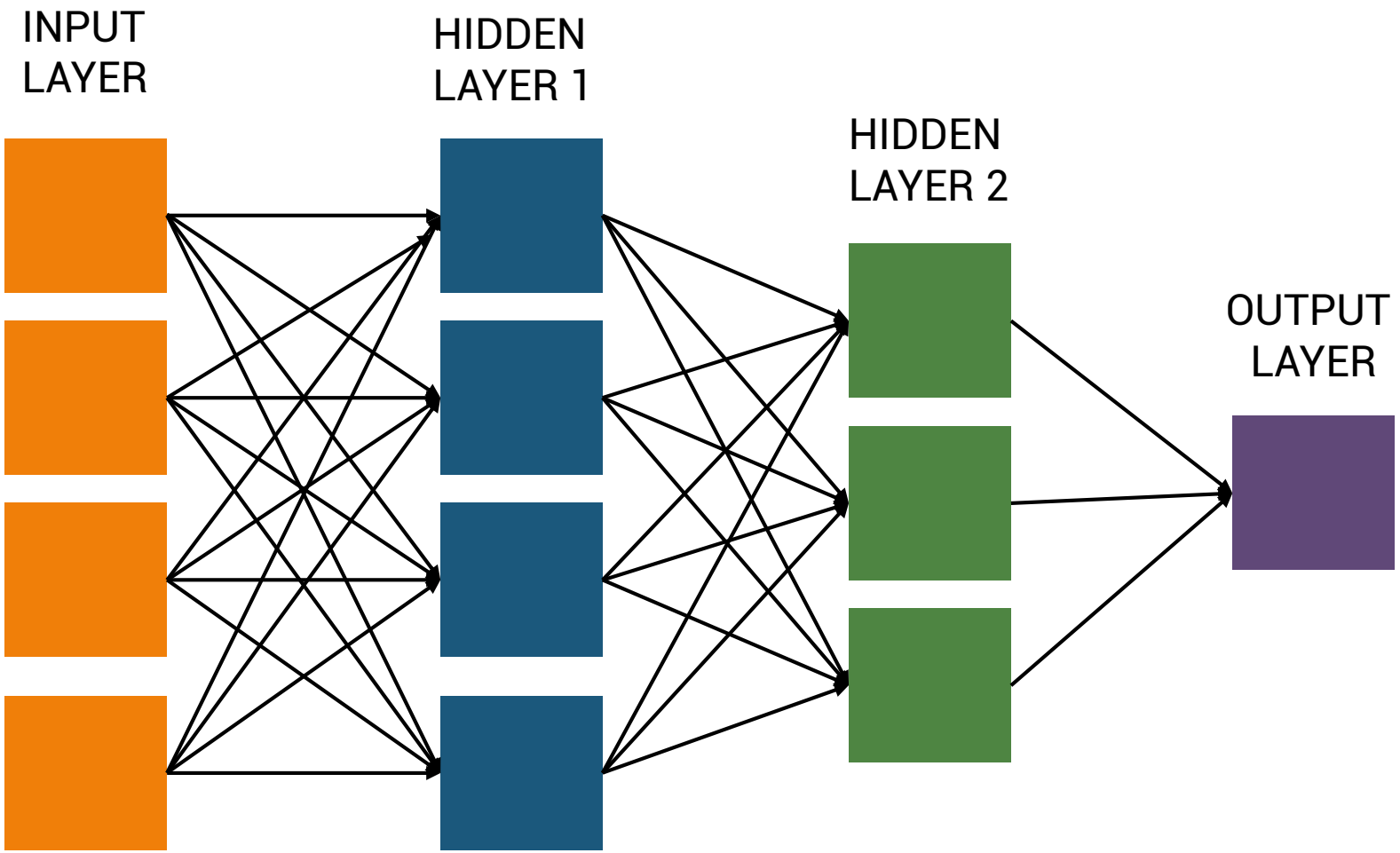
## Weighting

**The neural then takes a guess** and compares the result to the data. The network measures error and adjust weights to the extent they contributed to error. These steps are repeated until the error is minimized.

## Estimating Function

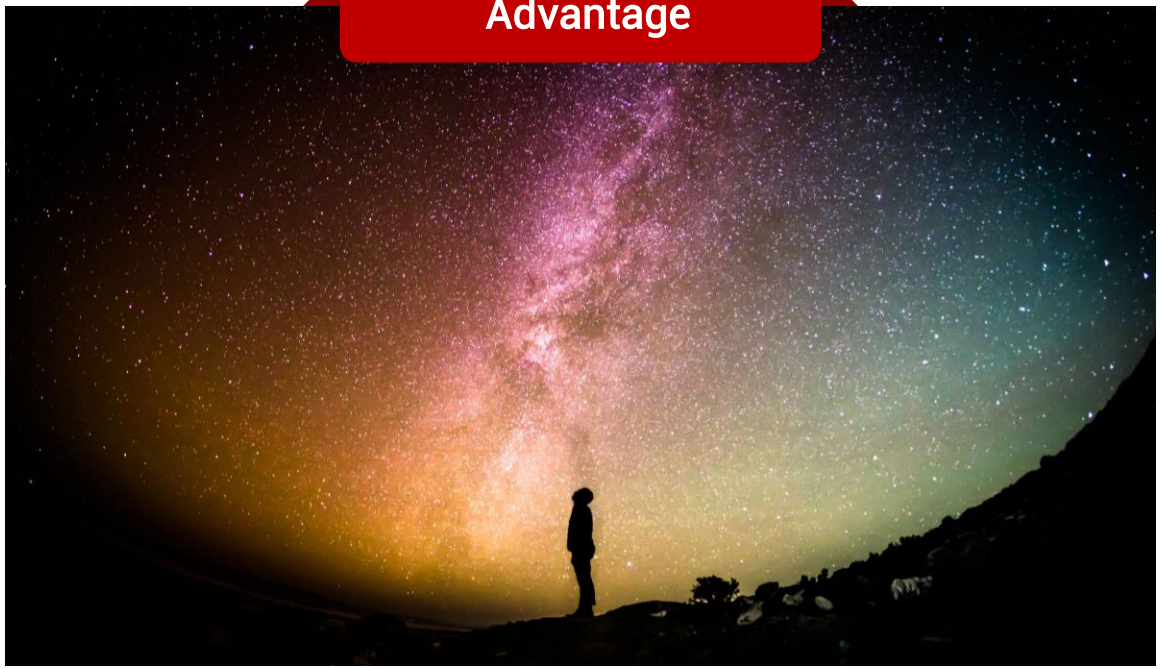
**A function f(x)=y** between an input, x, and output, y, is determined based on the neural network learning how x and y are related by understanding training datasets. Neural networks group unstructured data according to similarities within the inputs and classify data based on the labeled datasets used to train the predictive engine.

# Neural Networks



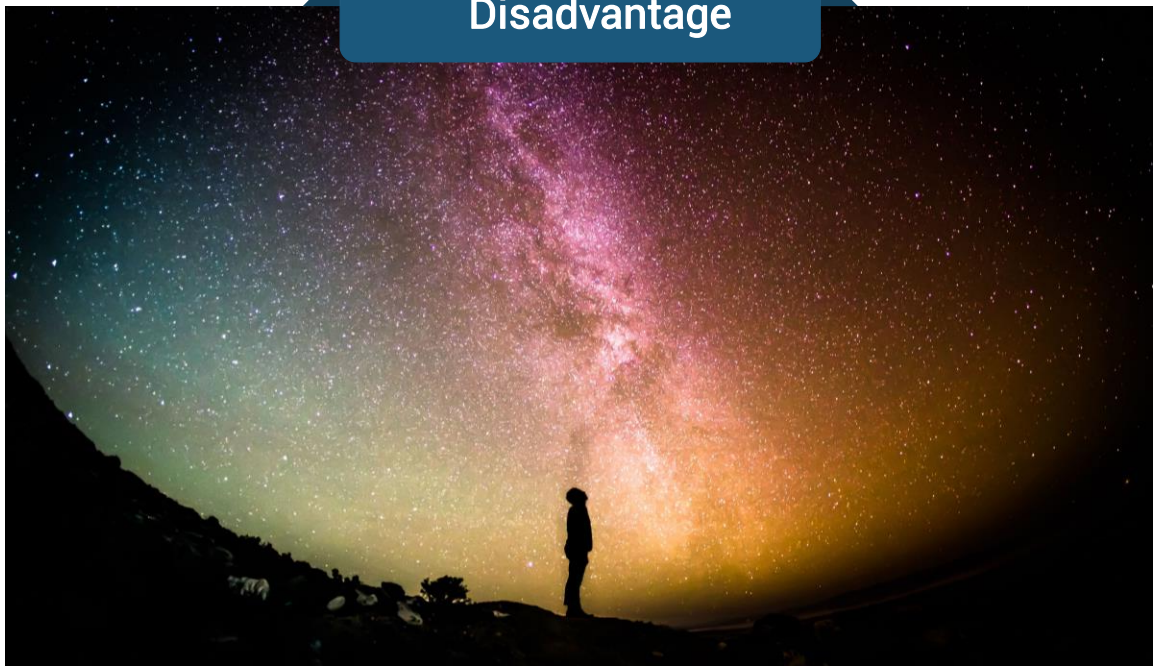
# Neural Networks Advantages and Disadvantages

**Advantage**



**One advantage** of neural networks is the ability to learn and model complex, non-linear relationships. Also, there are no restrictions on the input variables or assumptions on their distributions that have to be taken into consideration before the algorithm can be applied.

**Disadvantage**



**A disadvantage** to consider is that neural networks do not explain how they converge on a solution

# Regression Trees

Visual way to split data into homogenous, smaller regions for data exploration

## Tree Layout

**The decision trees** recursively split a dataset into partitions based on a criterion

## Terminal Node

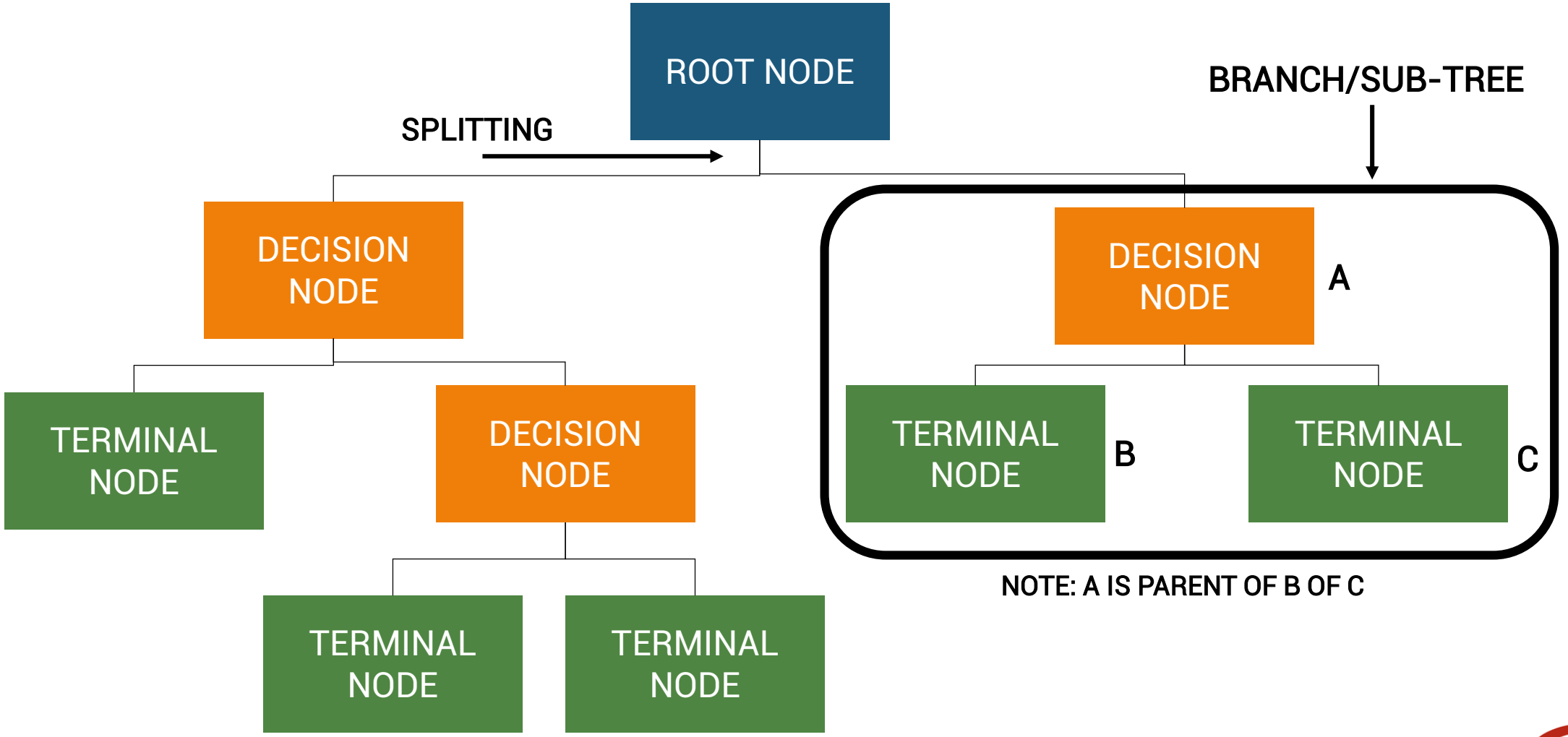
**The branch** ends at the terminal node when the data cannot be split further and the error is minimized

## Output

**The data points** at the terminal node are averaged to provide an estimate of the variables of interest based on the closest data points

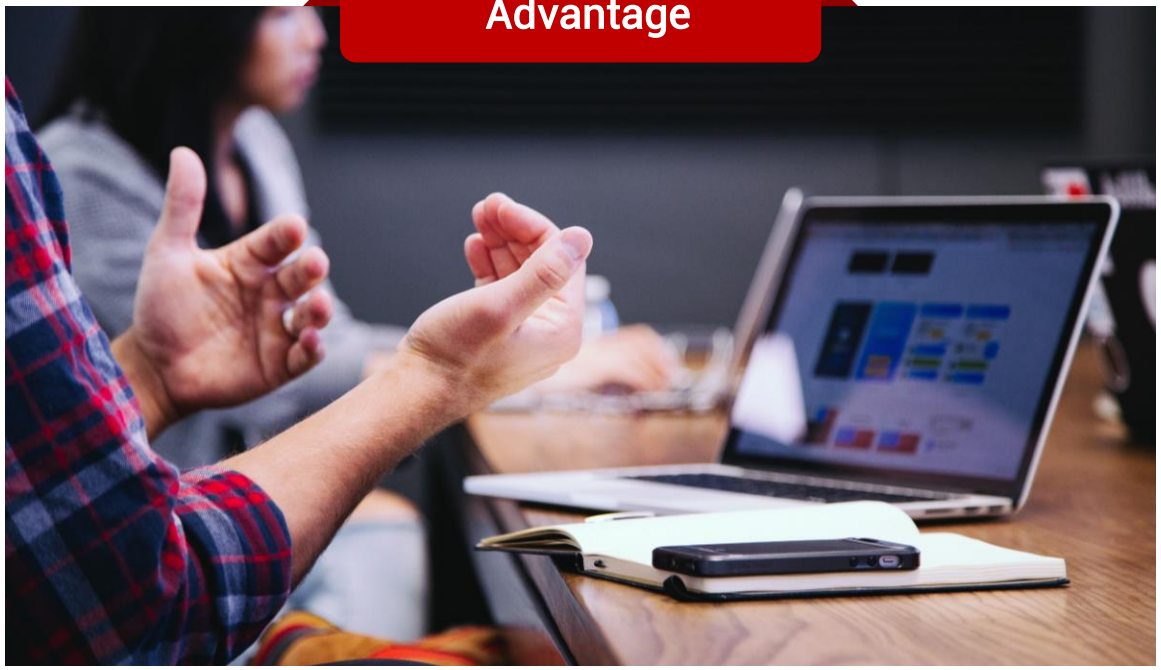
# Regression Trees

Diagram of a Regression Tree



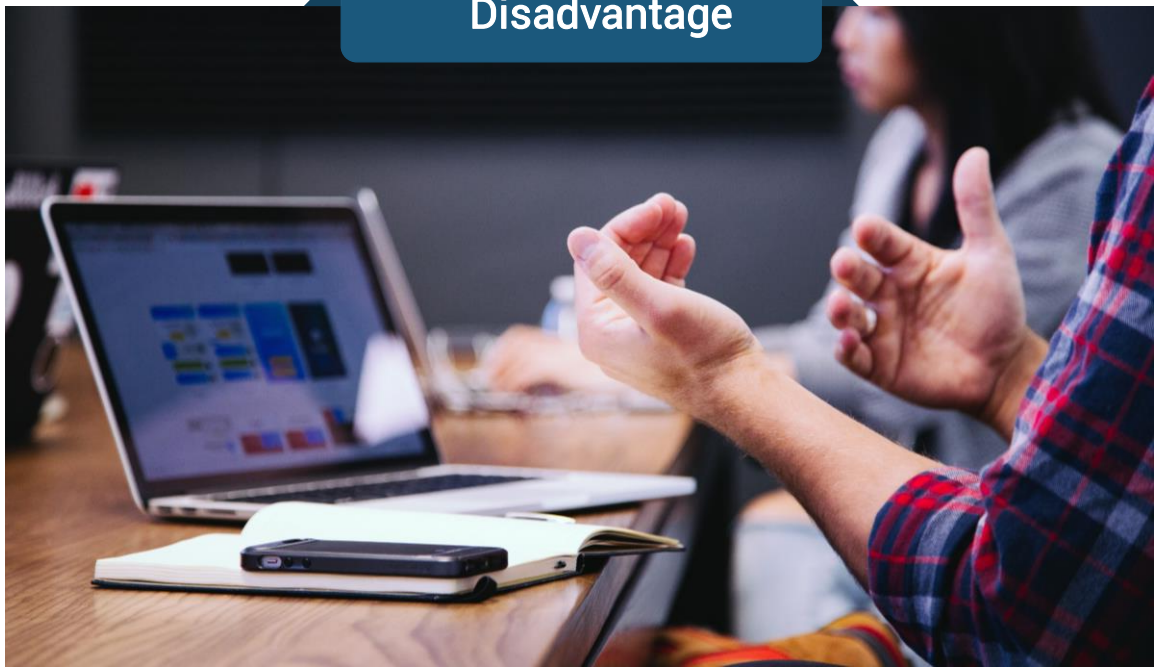
# Regression Trees Advantages and Disadvantages

## Advantage



**Visually,** they are very easy to understand. Regression trees can be used in preliminary data exploration to understand the most significant variables within a dataset. Also, since this method is nonparametric, it does not rely on data belonging to a particular type of distribution.

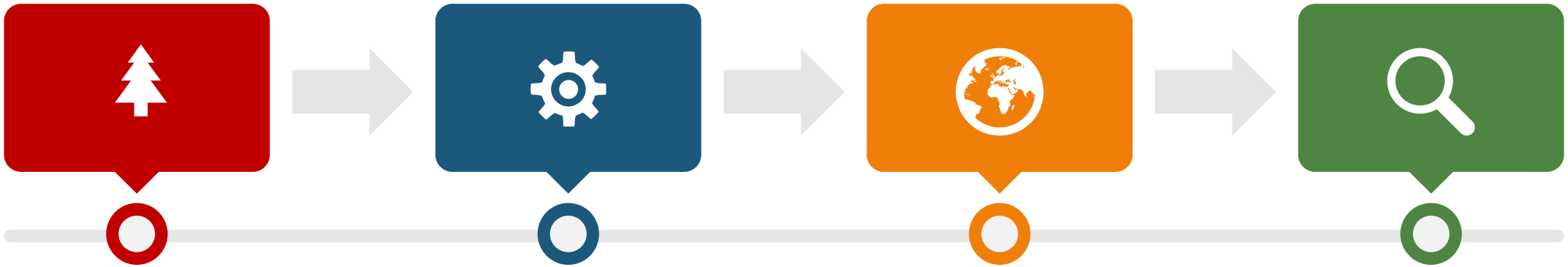
## Disadvantage



**As with KNN**, overfitting is also a common issue. Another disadvantage is despite the methods ability to handle both numerical and categorical variables, decision trees can lose information when it categorizes continuous variables in different categories

# Random Forests

An approach using multiple decision trees to produce an average for prediction



## Decision Trees

**As with Regression Trees,** decision trees are constructed by seeking the most important factors in minimizing error

## Crowd Approach

**Random Forests** have been proven to provide better prediction since they are more stable

## Random Generation

**The method** uses random sampling to generate training subsets of data and create trees for each subset

## Final Prediction

**The average** of 100-500 individual trees will provide the result of the random forest analysis

# Random Forests Advantages and Disadvantages

**Advantage**

**The advantage** of random forests is that they provide better prediction over one single tree. Random forests are more robust to small amounts of noise

**Disadvantage**

**An important disadvantage** to consider is that since the predictions are rather complex, there is no single equation or Cost Estimating Relationship. This may make communicating the model to management challenging

# Support Vectors

Goal is to maximize the margin between classes, which yields maximally robust classification


### Ideal Equation

**Find an equation** that is as "flat" as possible and has emphasis on sparseness

### Goals

**The goal for the equation** is to minimize the residuals that are outside a specified range of the estimate

### Optimization

**Must choose** optimal separating hyperplanes that best classify the data

### Calculation

**Excel solver (or R)** can be used to find the best equation that minimizes residuals

### Results

**A cost estimating relationship** will be obtained using specified independent variables

# Support Vectors Advantages and Disadvantages

**Advantage**

**The method** that is the closest in form to traditional regression, support vectors bring in concepts that make regression more robust and less prone to error. It is easy to implement in a spreadsheet using Excel Solver

**Disadvantage**

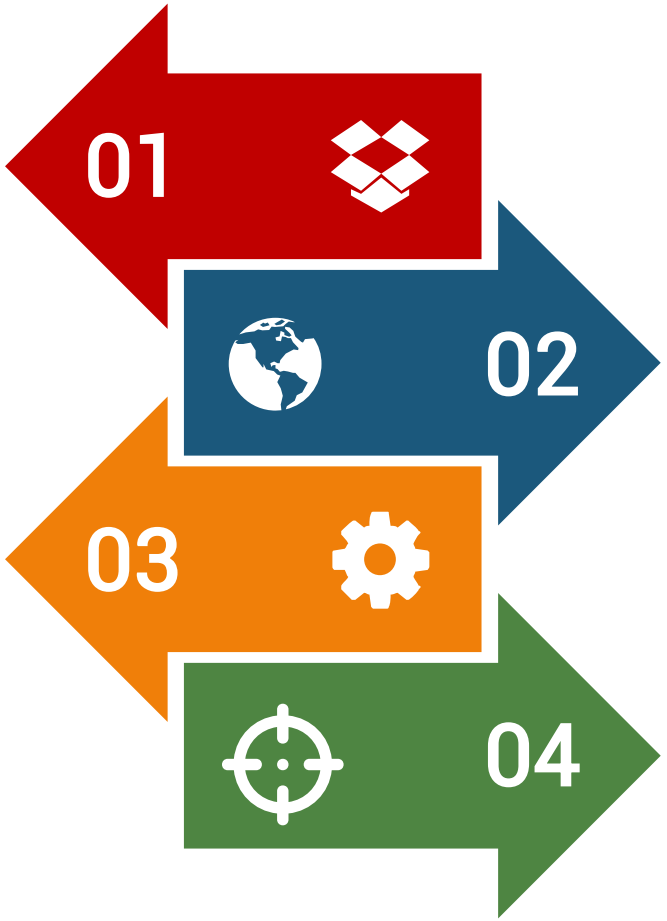
**A disadvantage** to support vector machines is speed and size, both in training and testing. Selection of the kernel is also a big limitation

# Text Analytics

Method that explores large quantities of textual data to find patterns

### STEP ONE

**Algorithm scans** text and large amounts of structured or unstructured data from different mediums and outputs data related to a specific topic




### STEP TWO

**Method seeks** to find correlations between multiple documents, groups of works and single words

### STEP THREE

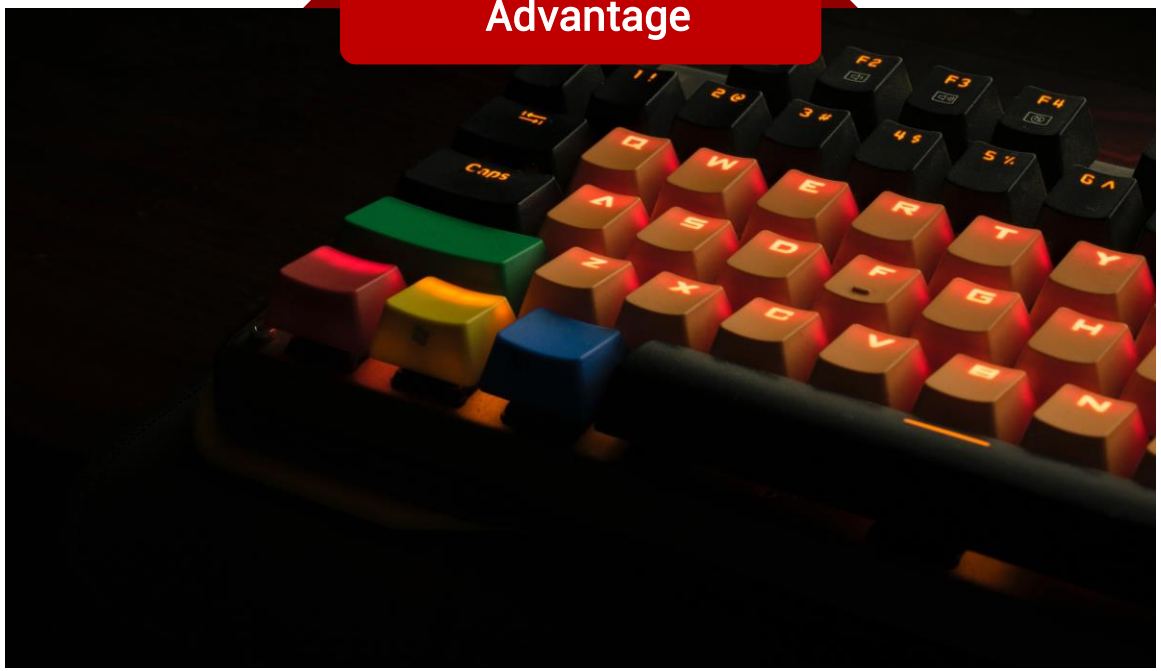
**The main goal** is to break down large quantities of data into smaller, more manageable chunks to facilitate analysis

### STEP FOUR

**Data** is then used to answer complex questions

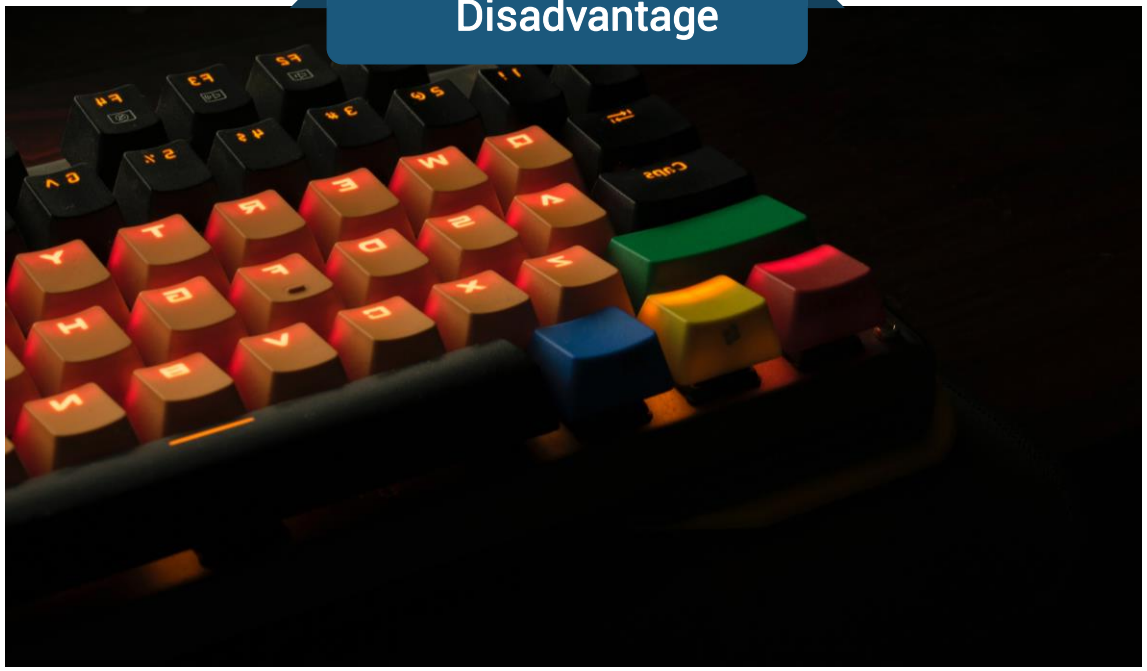
# Text Analytics Advantages and Disadvantages

**Advantage**



**Having the ability** to scan and dissect large amounts of data quickly is invaluable to analysts. This method not only can find information on a specified topic but can also understand nuances in language

**Disadvantage**



**A disadvantage** is that emotion cannot be captured during the review of text with this method. Without understanding tone and intent, sometimes text can be misinterpreted and information misused

# Results Comparison Slide

Space Dataset Results

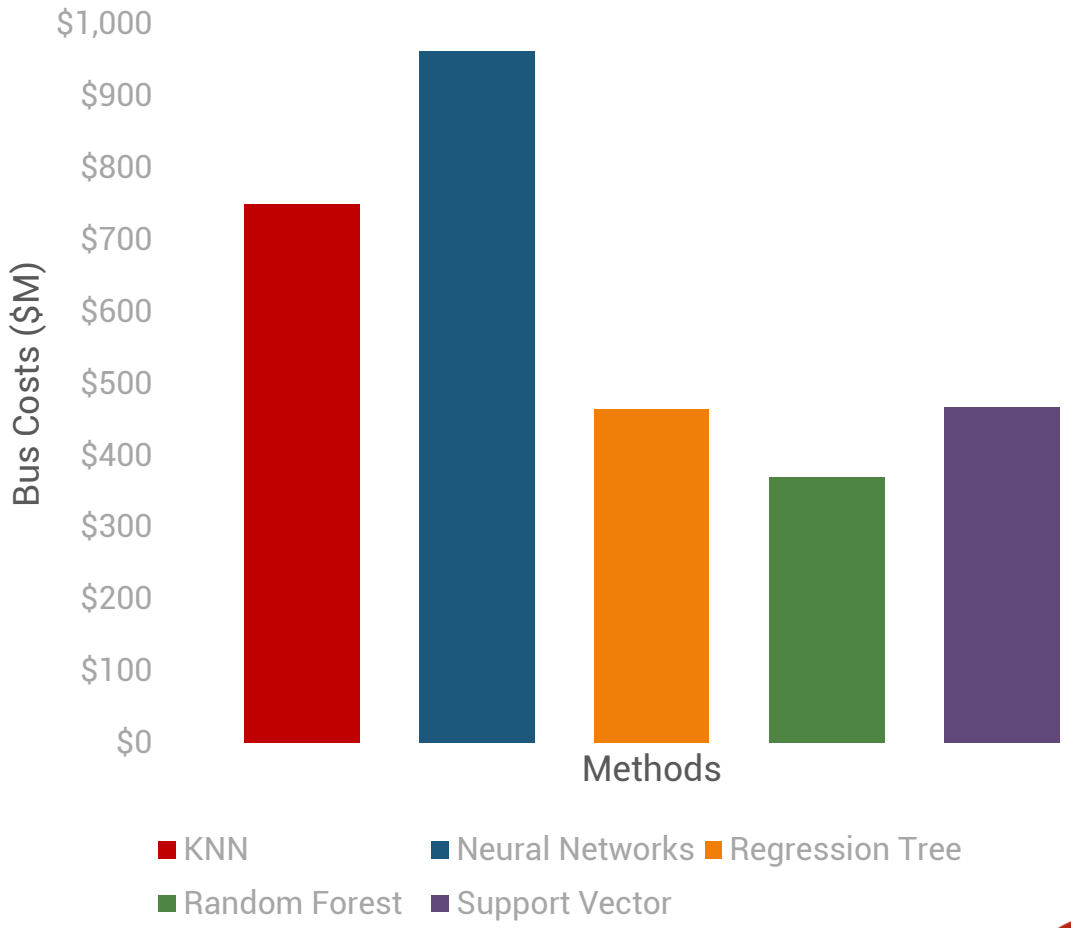
## Dataset Description

**A dataset of 47 data points** of space bus cost data was analyzed using the aforementioned methods. The goal was to compare estimated the estimated bus cost given the same inputs for four variables

Variables: Bus Weight, Schedule (in log-transformed months), Design Life (in log-transformed months), and Testing (0- Testing did not occur; 1 – Testing occurred)

With the same inputs, the methods predicted the following Bus Costs: KNN - $750M; Neural Networks - $963M; Regression Tree - $464M; Random Forests - $369M; and Support Vector Machines - $467M

Further exploration of why Neural Networks produced a much higher bus cost is required



# Tools for Machine Learning

What to use to implement these methods

### R

Free scripting language with powerful statistic packages available

### Python

Free programming language with numerous tools for machine learning


### Programming

Both R and Python require programming/script writing. Many resources online to help find the code to produce the required results.

### Limitation

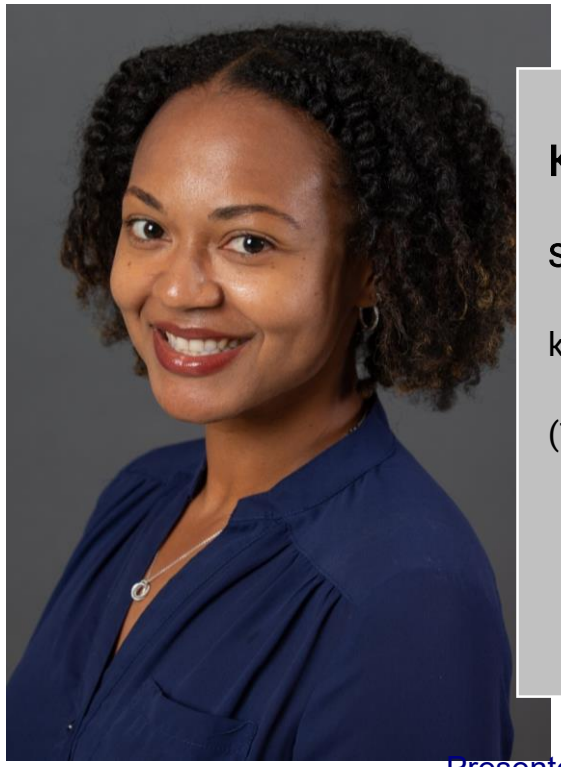
There may be issues installing these on government computers

# References


- Anderson, C., “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," Wired, June 2008.
- Davenport, D.H., and D.J. Patil, “Data Scientist: The Sexiest Job of the 21st Century," Harvard Business Review, October 2012.
- Dean,E., “Neural Network Cost Estimating Relationships," Proceedings of the 2010 ISPA-SCEA Conference, San Diego.
- Hutchings, C., “An Approach Towards Determining Value Through the Application of Machine Learning," Proceedings of the 2018 ICEAA Conference.
- Kaluzny, B.L, “An Application of Data Mining Algorithms for Shipbuilding Cost Estimation," Proceedings of the 2011 ISPA-SCEA Conference.
- Mourikas, K., J. King, and D. Nelson, “Machine Learning Approach to Cost Analysis," Proceedings of the 2017 ICEAA Conference.
- Newell, A. and P.S. Rosembloom, “Mechanisms of Skill Acquisition and the Law of Practice," in R Anderson(Ed.), Cognitive Skills and Their Acquisition, Hillsdale, NJ, Erlbaum.
- Pincus, J., and O. Akbik, “Social Media and Submarines: How Machine Learning and Unconventional Methods Can Change Cost Estimating," Proceedings of the 2018 ICEAA Conference.
- Rao, Venky, “Introduction to Classification & Regression Trees", January 2013.
- Singh,Aishwarya, “A Practical Introduction to K-Nearest Neighbors Algorithm for Regression", August 2018.

# Presenters

Galorath Federal




**Kimberly Roye**

**Senior Consultant**

kroye@galorath.com

(703) 966-3192



**Dr. Christian Smart**

**Chief Scientist**

csmart@galorath.com

(256) 457-3354