LOYOLA
UNIVERSITY MARYLAND

MAA
MATHEMATICAL ASSOCIATION OF AMERICA

Data Science:

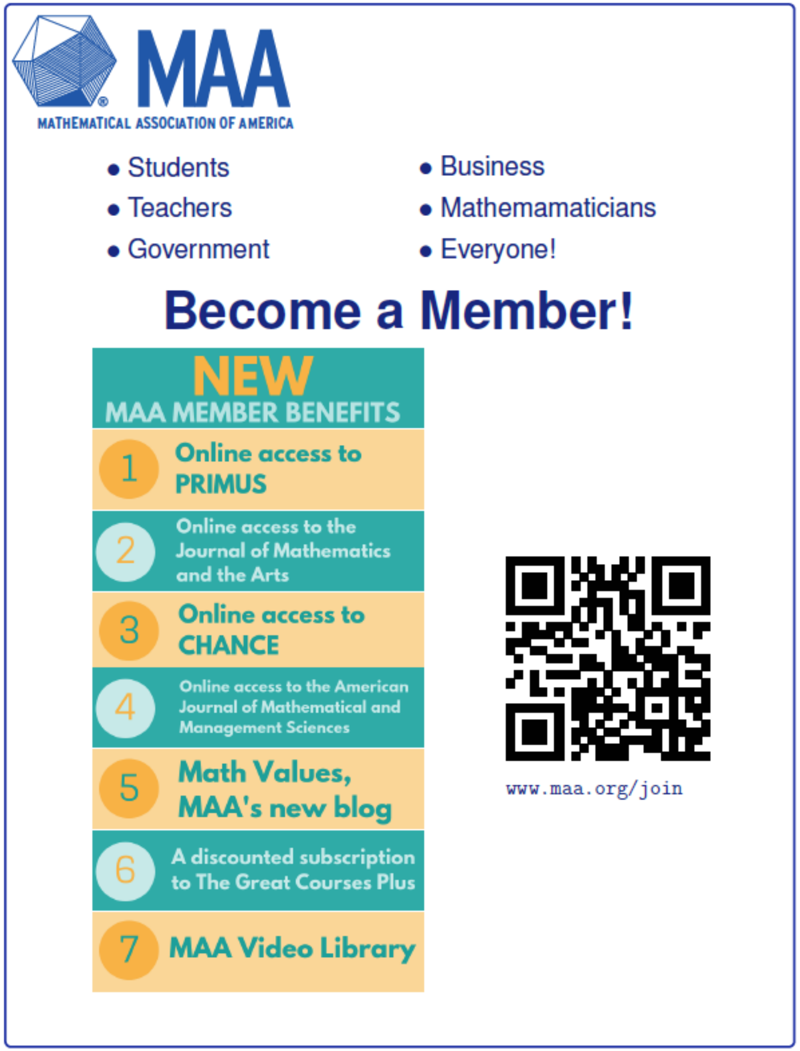
The Discipline, Methods, and Preparation

Christopher Morrell
Professor of Statistics
Director of Data Science Programs
Mathematics and Statistics Department
Loyola University Maryland
chm@loyola.edu

WASHINGTON
STATISTICAL
SOCIETY

MAA Webinar Recording Policy

- The MAA is recording the webinar.
- The MAA retains the right to show it again and to distribute it.
- By participating, you are agreeing that your contributions become part of the recording.



MAA

MATHEMATICAL ASSOCIATION OF AMERICA

- Students
- Teachers
- Government
- Business
- Mathematicians
- Everyone!

Become a Member!

NEW

MAA MEMBER BENEFITS

1. Online access to PRIMUS
2. Online access to the Journal of Mathematics and the Arts
3. Online access to CHANCE
4. Online access to the American Journal of Mathematical and Management Sciences
5. Math Values, MAA's new blog
6. A discounted subscription to The Great Courses Plus
7. MAA Video Library

www.maa.org/join

ASA and WSS Membership

- To join WSS: Go to http://washstat.org/joinus.html . If you are already an ASA member, follow the instructions for “Full Membership”. If you are not an ASA member, follow the instructions for “Associate Membership”
- To join ASA: Go to https://www.amstat.org/ and click on Join.
- To join BOTH ASA and WSS: Join the ASA as above. When you get to the 2nd page, click on “Select Chapters”, scroll down to “District of Columbia” and click on Washington Statistical Society.
- **Contact Carol Blumberg at cblumberg@gmail.com if you have any questions**

Outline

- Data Science
  - What is it?
  - Demand
  - Ethics
- Data Science process
  - Getting the data
  - Analysis methods
- Preparation

What is Data Science?

- Extracts knowledge and insights from data
- Evolving interdisciplinary field:
  - computer science, statistics
  - domain knowledge
- *“data science is the art of turning data into actions”*
  - from *The Field Guide to Data Science*, 2nd Edition, BoozAllen, 2015.

Need for Data Scientists

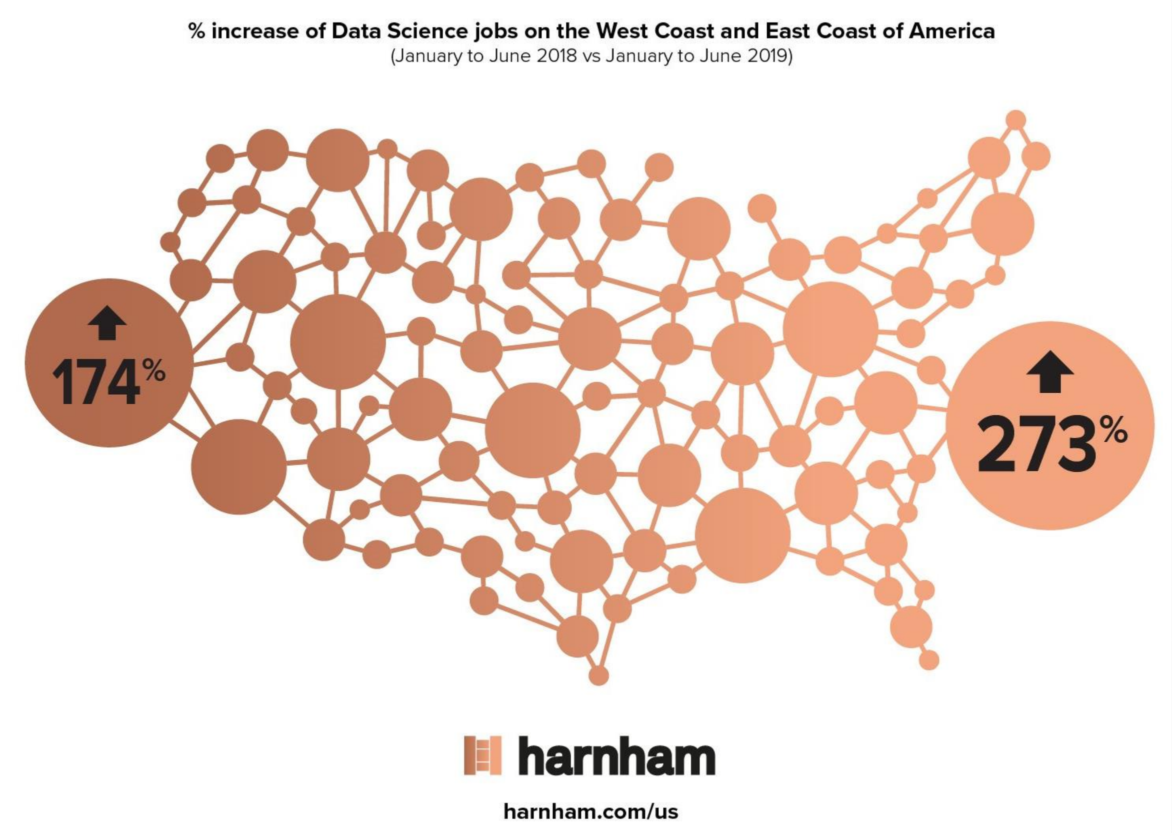
- McKinsey Global Institute report (2011) concludes, "a **shortage of** the analytical and managerial **talent** necessary to make the most of Big Data is a significant and pressing challenge (for the U.S.)."
  - Large numbers of positions will only be filled through training or retraining.
  - Project a need more managers and analysts with deep analytical and technical skills “who can ask the right questions and consume the results of analysis of big data effectively”
- IBM Predicts **Demand For Data Scientists Will Soar 28% by 2020** (from a Forbes article)

McKinsey Global Institute. "Big data: The next frontier for innovation, competition, and productivity", May 2011

LinkedIn Workforce Report | United States | August 2018

- **Skills Gaps | Demand for data scientists is off the charts**
- Data science skills shortages are present in almost every large U.S. city. Nationally, we have a shortage of 151,717 people with data science skills. As more industries rely on big data to make decisions, data science has become increasingly important across all industries, not just tech and finance. In that sense, it’s a good proxy for how today’s cutting-edge skills like AI & machine learning may spread to other industries and geographies in the future.

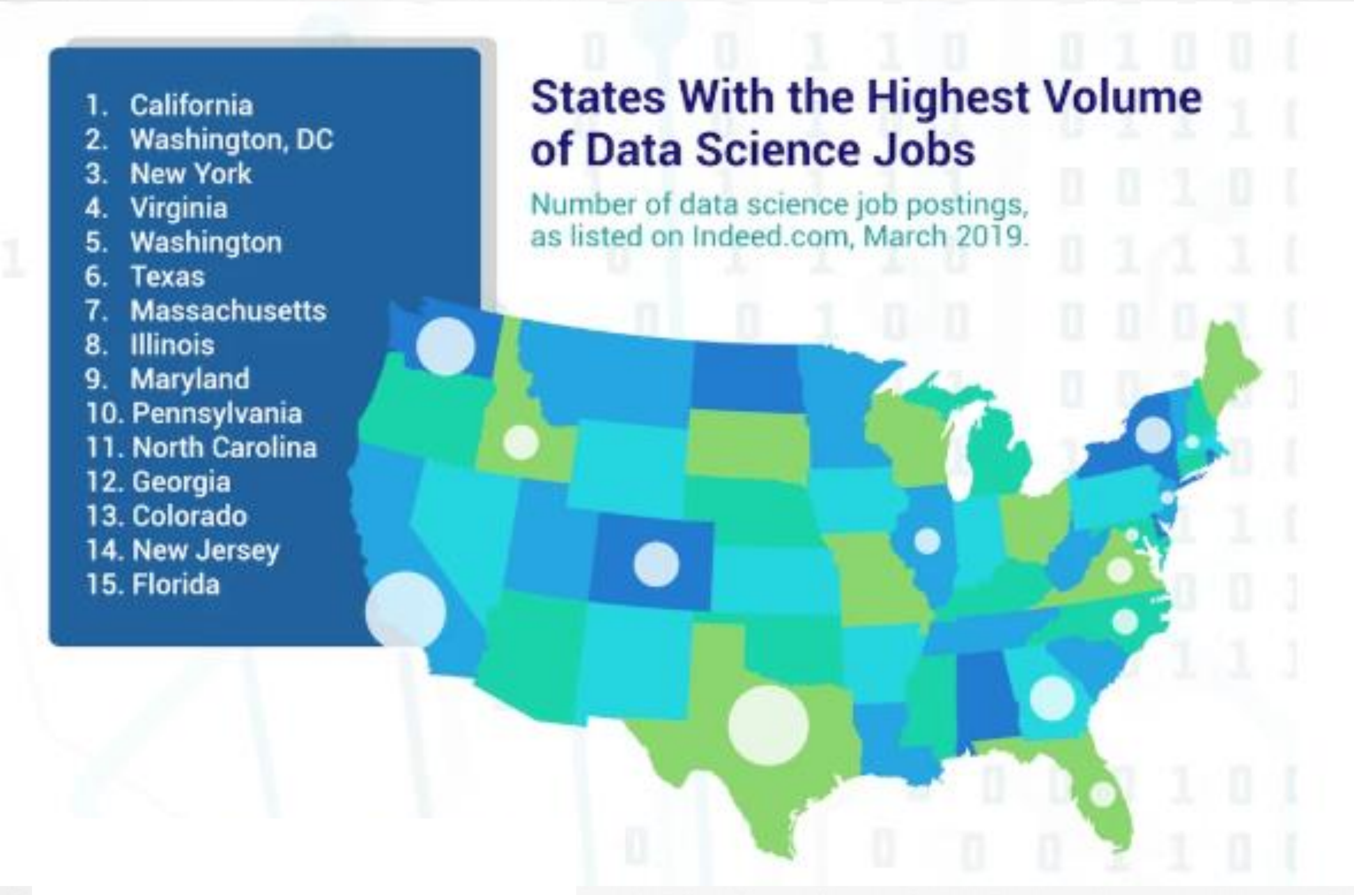
What is the Need?



National need:

- Number of Job Openings: 32,665 positions in US (3,208 within 50 miles of Baltimore) on glassdoor.com (3/12/2020)

https://www.pcmag.com/news/370627/looking-to-switch-careers-data-science-is-booming

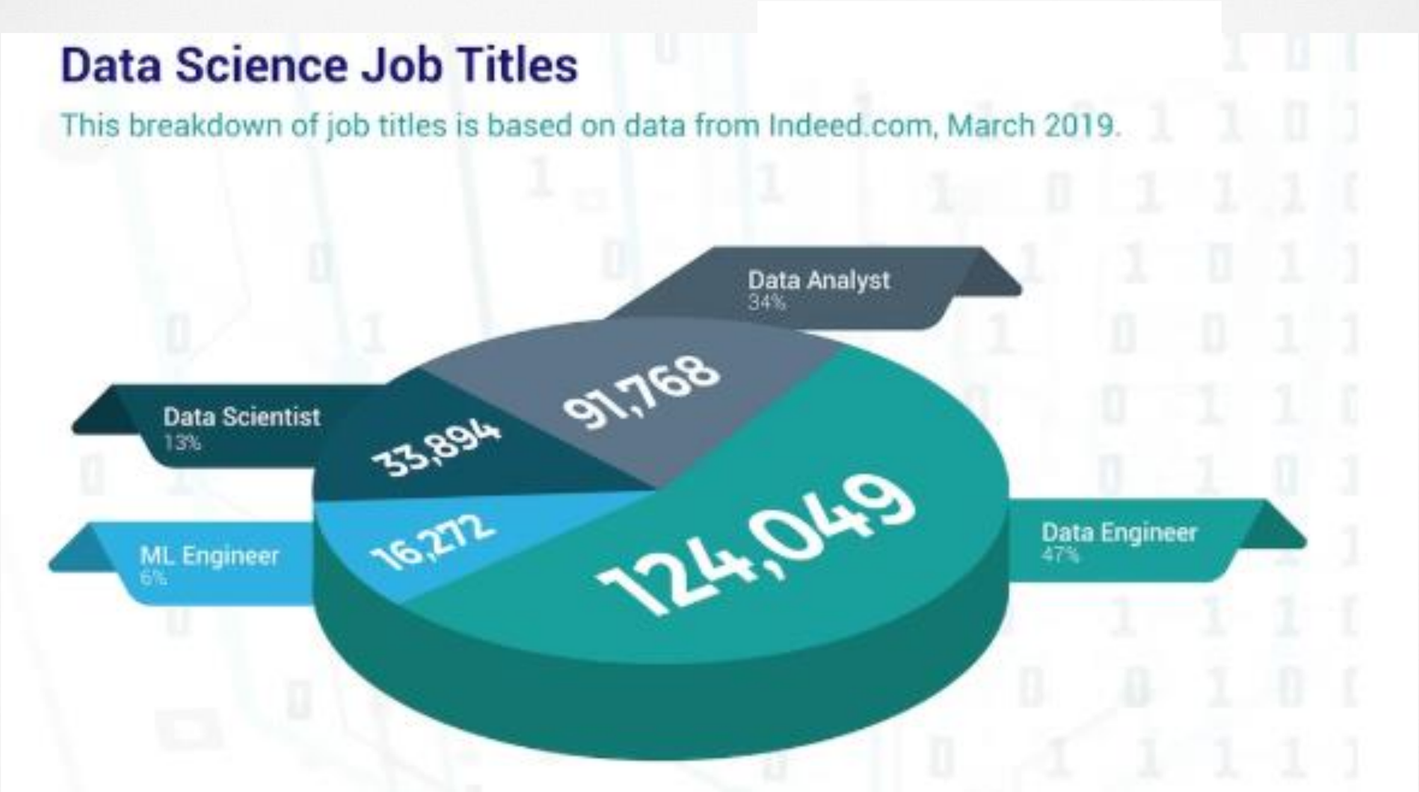


States With the Highest Volume of Data Science Jobs

Number of data science job postings, as listed on Indeed.com, March 2019.

1. California
2. Washington, DC
3. New York
4. Virginia
5. Washington
6. Texas
7. Massachusetts
8. Illinois
9. Maryland
10. Pennsylvania
11. North Carolina
12. Georgia
13. Colorado
14. New Jersey
15. Florida

https://www.springboard.com/blog/data-science-salaries/



Data Science Job Titles

This breakdown of job titles is based on data from Indeed.com, March 2019.

https://www.springboard.com/blog/data-science-salaries/

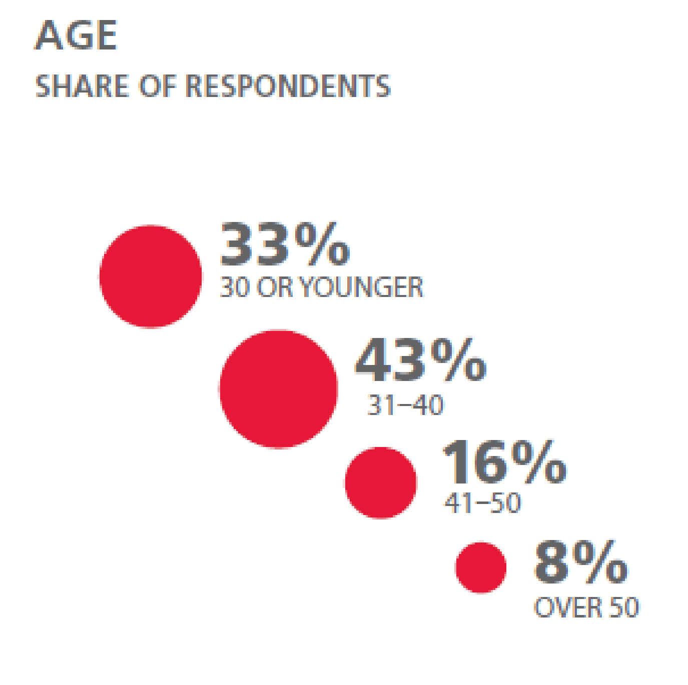
Work Features

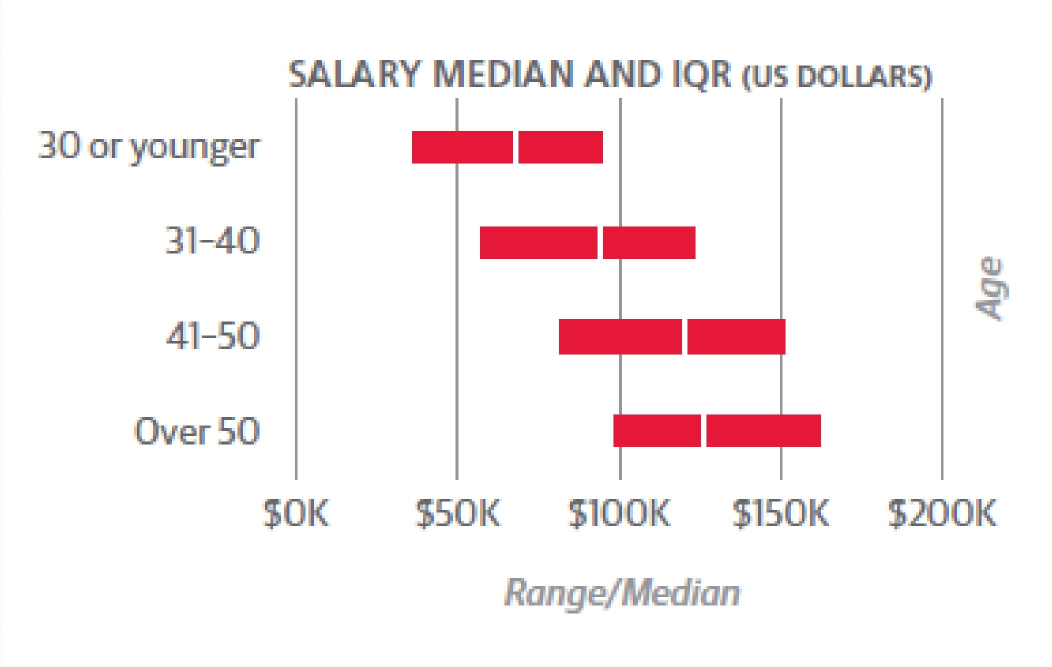
- Number 1 job for work-life balance[1]
- Salary:
  - National average: $113,309[2]

1. https://www.today.com/money/best-jobs-work-life-balance-data-scientists-seo-managers-more-t51326

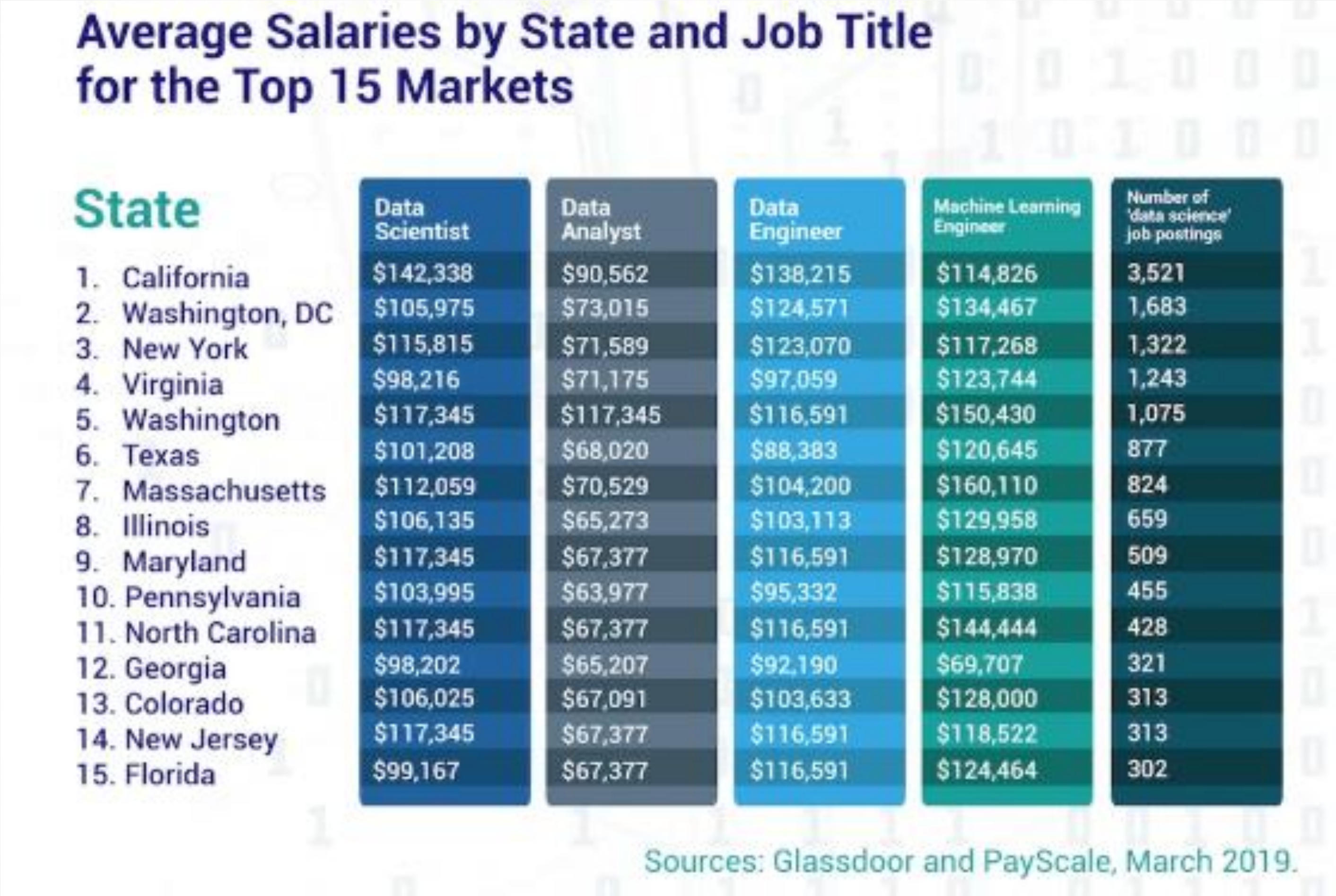
2. https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm (3/12/2020)

Salary





Source *- O’Reilly 2017 Data Science Salary Survey*



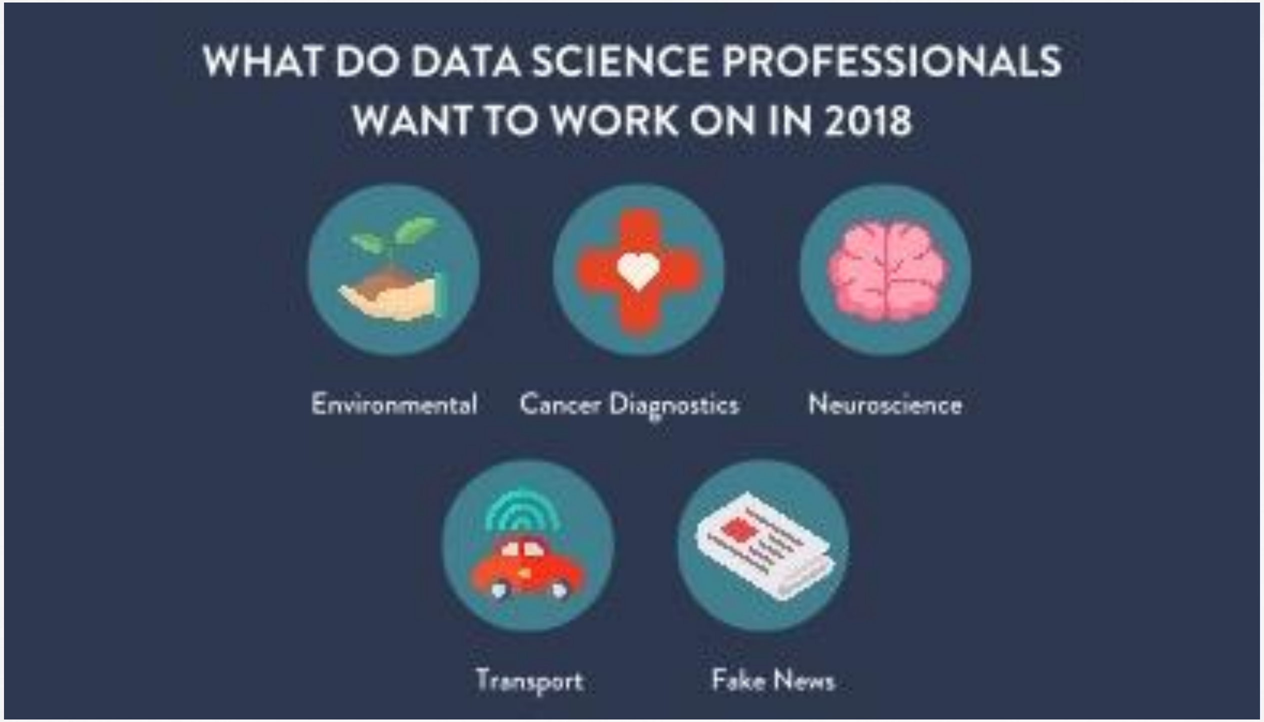
Average Salaries by State and Job Title for the Top 15 Markets

State	Data Scientist	Data Analyst	Data Engineer	Machine Learning Engineer	Number of 'data science' job postings
1. California	$142,338	$90,562	$138,215	$114,826	3,521
2. Washington, DC	$105,975	$73,015	$124,571	$134,467	1,683
3. New York	$115,815	$71,589	$123,070	$117,268	1,322
4. Virginia	$98,216	$71,175	$97,059	$123,744	1,243
5. Washington	$117,345	$117,345	$116,591	$150,430	1,075
6. Texas	$101,208	$68,020	$88,383	$120,645	877
7. Massachusetts	$112,059	$70,529	$104,200	$160,110	824
8. Illinois	$106,135	$65,273	$103,113	$129,958	659
9. Maryland	$117,345	$67,377	$116,591	$128,970	509
10. Pennsylvania	$103,995	$63,977	$95,332	$115,838	455
11. North Carolina	$117,345	$67,377	$116,591	$144,444	428
12. Georgia	$98,202	$65,207	$92,190	$69,707	321
13. Colorado	$106,025	$67,091	$103,633	$128,000	313
14. New Jersey	$117,345	$67,377	$116,591	$118,522	313
15. Florida	$99,167	$67,377	$116,591	$124,464	302

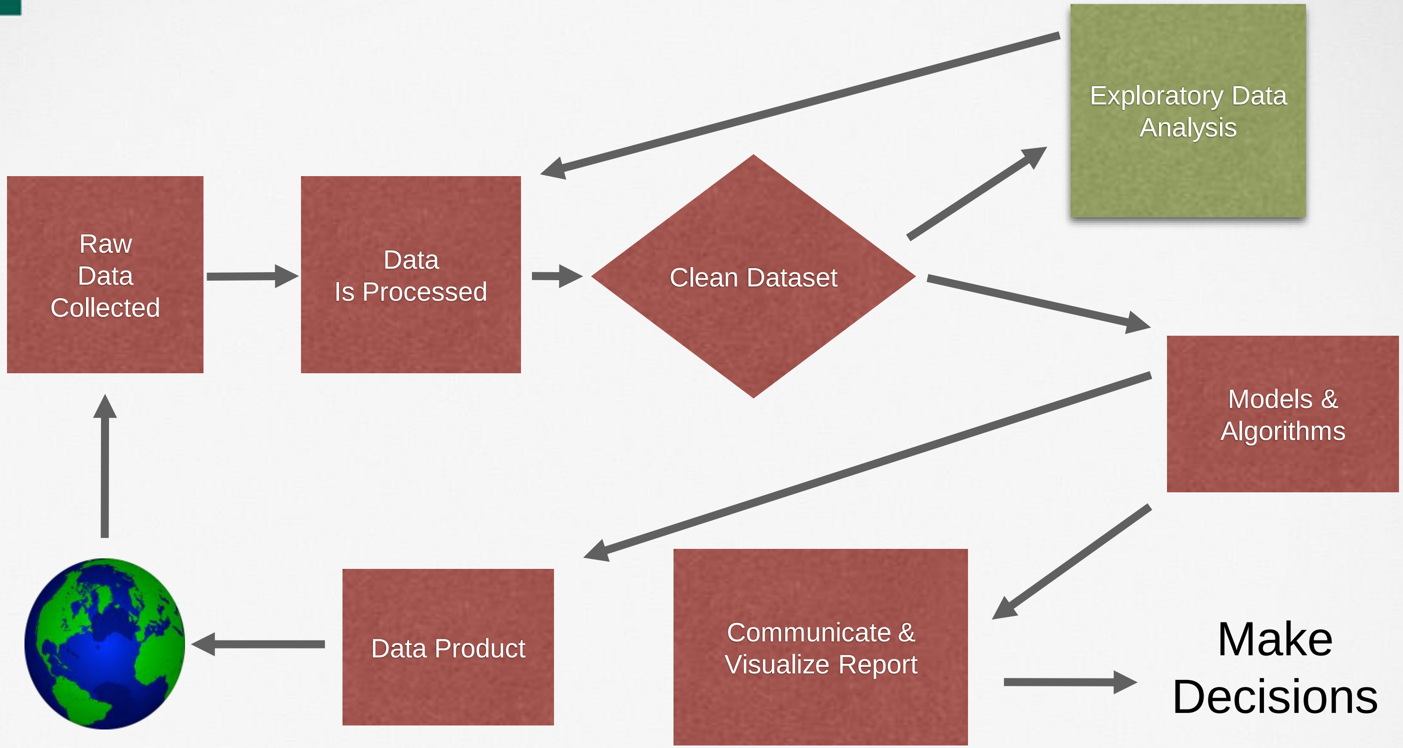
Sources: Glassdoor and PayScale, March 2019.

https://www.springboard.com/blog/data-science-salaries/

Application Areas



https://www.kdnuggets.com/2018/01/2018-data-science-salary-report.html



Data Science Process

Brings together disparate data sources to recognize new opportunities and improve current practices

Obtaining Data

- From various sources and types
- Structured
  - Can be displayed in rows and columns
  - Number, dates, strings
- Unstructured
  - Images, audio, video, text files, documents, spreadsheets, …

Feature/variable engineering

- For unstructured data, what aspects/features of the item are to be recorded and used in the analyses.
  - Character string: length, # vowels, …
  - Text: word count, N-grams (phrases), Term Frequency-Inverse Document Frequency (TF-IDF)
  - Image: Grayscale Pixel Values, Mean Pixel Value of Channels, Extracting Edge Features
  - Audio: Zero Cross Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Roll off, Mel-Frequency Cepstrum Coefficients (MFCC), Chroma Vector, Chroma Deviation

Feature/variable selection

- Automatically or manually select those features/variables which contribute most to your model or prediction.
- Feature selection techniques are used for several reasons:
  - simplification of models to make them easier to interpret,
  - shorter training times,
  - to avoid the curse of dimensionality,
  - enhanced generalization by reducing overfitting.

“Machine Learning”

- **Common Machine Learning Algorithms**

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. Support Vector Machines
5. Naive Bayes
6. K Nearest Neighbors
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

“Machine Learning” – traditional statistics approaches

- **Common Machine Learning Algorithms**

1. Linear Regression
2. Logistic Regression
3. Decision Tree (proposed by statistician, Breiman, in 1984)
4. Support Vector Machines
5. Naive Bayes
6. K Nearest Neighbors
7. K-Means (Cluster Analysis)
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms (developed in 1990s by Breiman and Friedman)

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

More Machine Learning

- Neural Networks
- Deep Learning
- Ensemble Algorithms

Approaches to the analysis of data

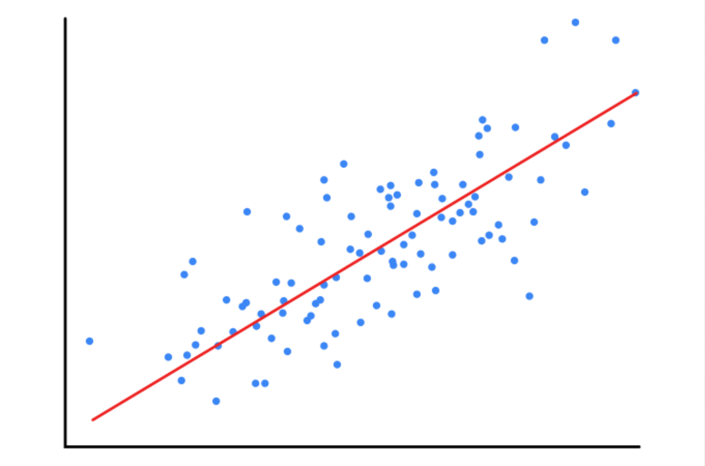
- **Supervised Learning:** Supervised Learning are methods in which a model is "trained" using a set of pre-existing labeled data.
- **Unsupervised Learning:** A class of methods in which a model is built without the use of labeled data.

Problem Types

- **Classification**: Assigning or predicting a observation's membership in discrete class
- **Regression:** Predicting a continuous value based on the observations’ values
- **Clustering:** Identifying groupings within a dataset
- **Dimensionality Reduction:** Reducing the number of variables in a feature set

Linear Regression

- Regression is concerned with modeling the relationship between a numerical response variable and a number of explanatory variables.
- Example: Want to predict the demand for a product based on a number of factors.
- $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

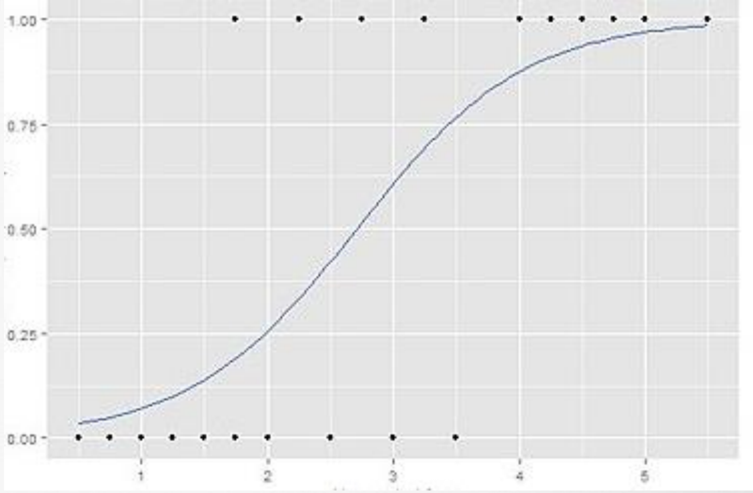


https://backlog.com/blog/gradient-descent-linear-regression-using-golang/

Logistic Regression

- Logistic regression is concerned with modeling the relationship between a binary or categorical response variable and a number of explanatory variables.
- Example: Will a customer purchase a product (yes or no) based on the customer’s characteristics and previous purchase behavior.
- $p(x) = \frac{e^{X\beta}}{1+e^{X\beta}}$

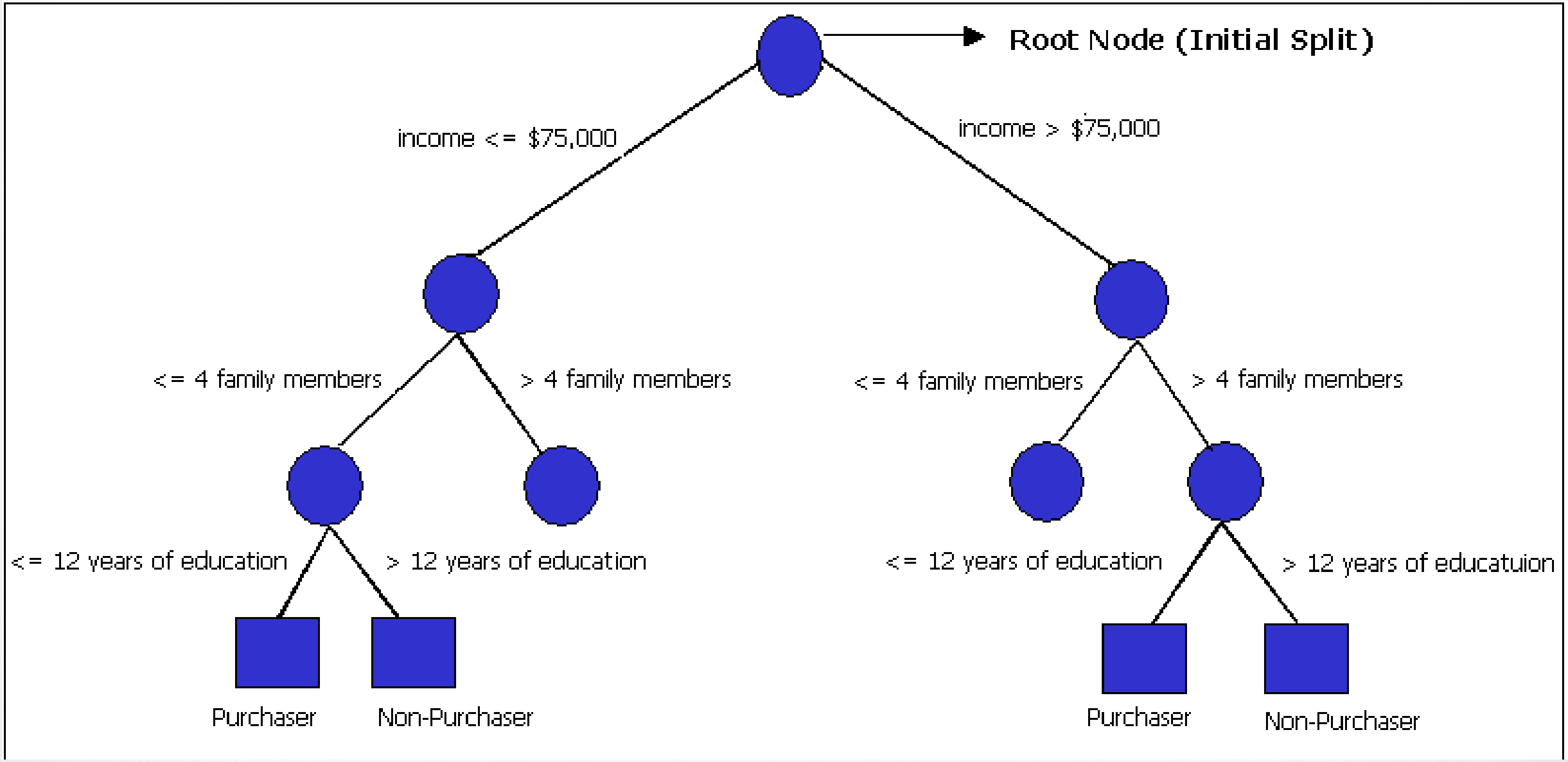
https://en.wikipedia.org/wiki/Logistic_regression



Decision Tree

- A type of supervised learning algorithm that is mostly used for classification problems.
- Works for both categorical and continuous dependent variables (classification trees and regression trees).

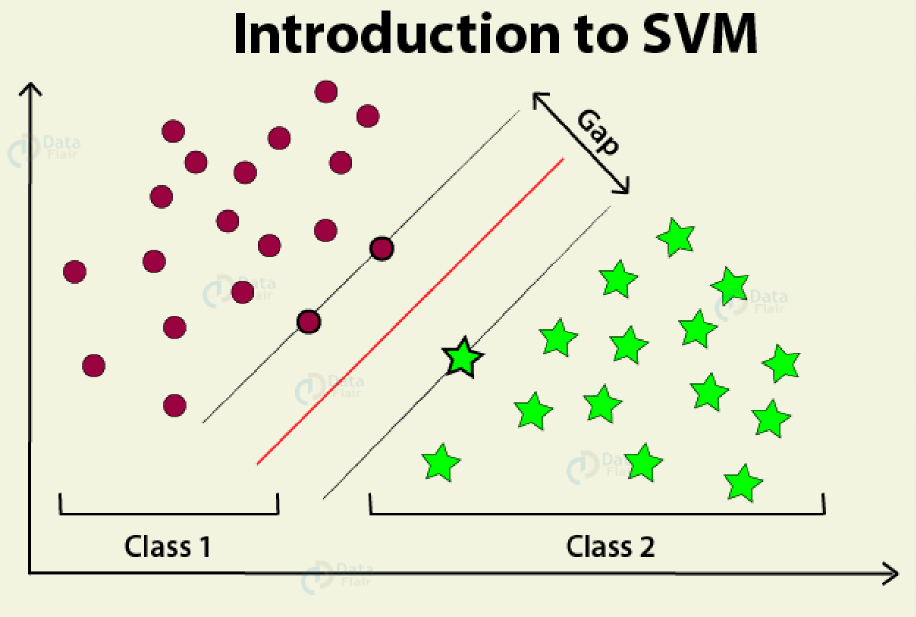
Decision Tree - example



https://www.solver.com/classification-tree

Support Vector Machine

- A supervised machine learning model that uses classification algorithms for two-group classification problems.
- Plot data in the n-dimensional space.
- Find the ideal hyperplane that differentiates between the two classes.



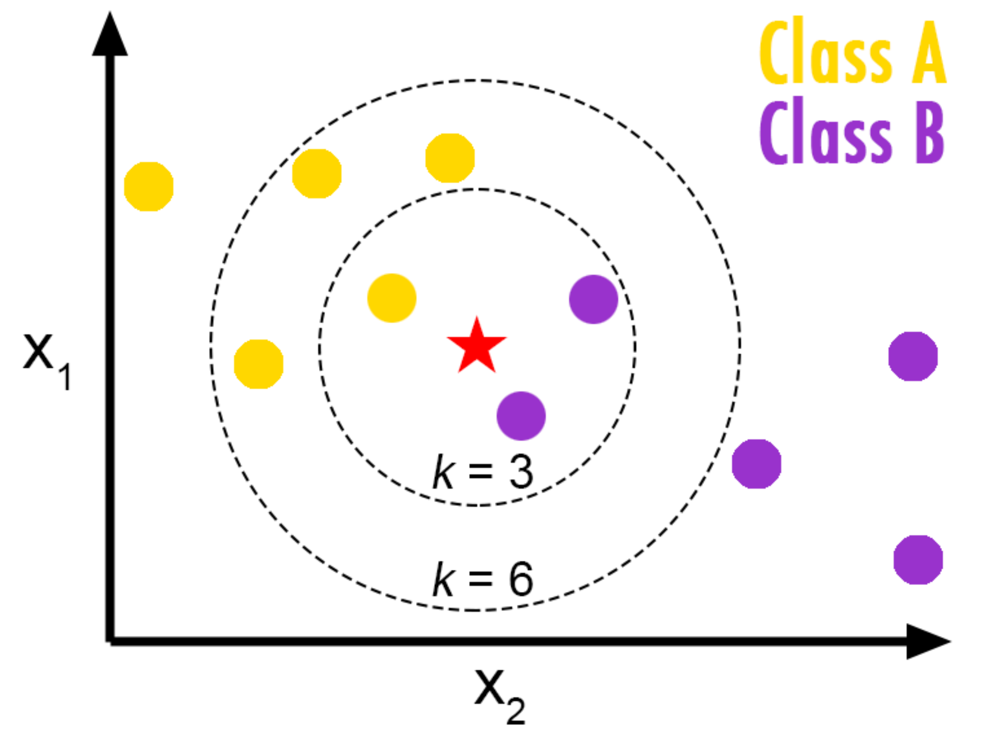
https://data-flair.training/blogs/svm-support-vector-machine-tutorial/

Naïve Bayes

- Naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.
- Predict the probability of different class based on various attributes.
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\{P(B_1|A)\times\cdots\times P(B_n|A)\}P(A)}{P(B)}$

K Nearest Neighbors

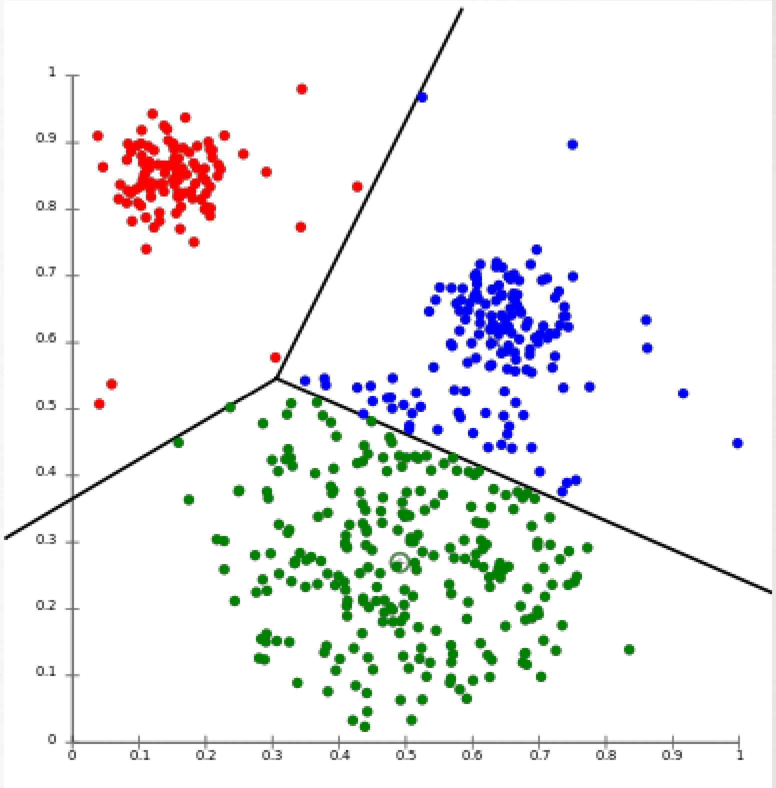
- Used in classification problems.
- Classifies new cases by a majority vote of its k neighbors.



https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d

K-Means

- Unsupervised clustering algorithm that attempts to group observations into different clusters.
- The goal of the algorithm is to minimize the difference *within* clusters and maximize the difference *between* clusters.

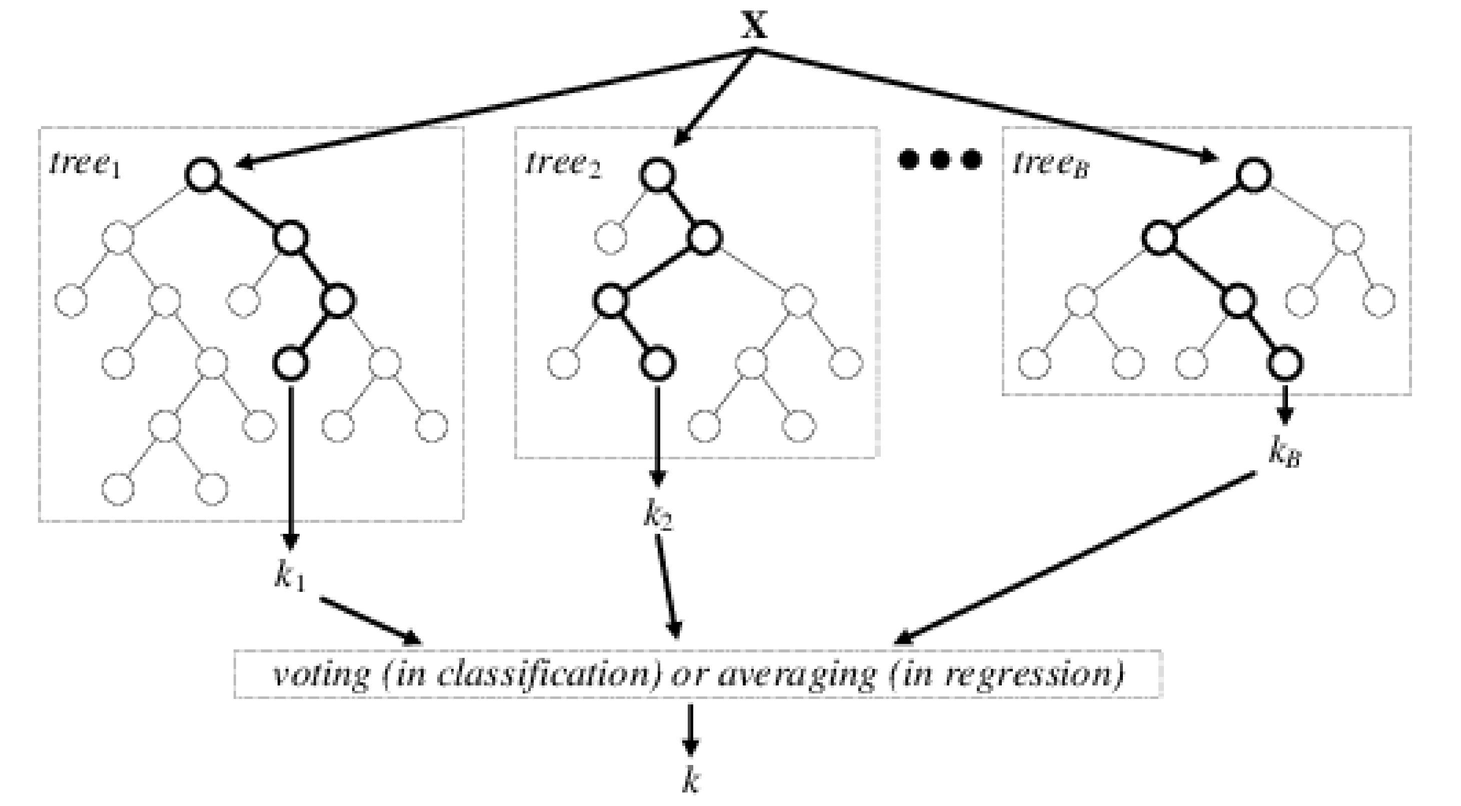


https://aws.amazon.com/blogs/machine-learning/k-means-clustering-with-amazon-sagemaker/

Random Forest

- An ensemble of decision trees.
- To classify/predict a new object based, each tree gives a classification/prediction.
- The forest chooses the classification having the most votes (over all the trees in the forest).
- In a regression setting, the forest averages the predictions.

Random Forest



https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643

Dimensionality Reduction Algorithms

- Have many variables.
- The variables may be correlated (so all do not provide independent information).
- Want to reduce the number of variables.
- One method: Principal Component Analysis (PCA): Variables are transformed into a new set of variables which are linear combination of original variables.
- Choose the number of PCs that account for a suitable amount of the variation in the data.

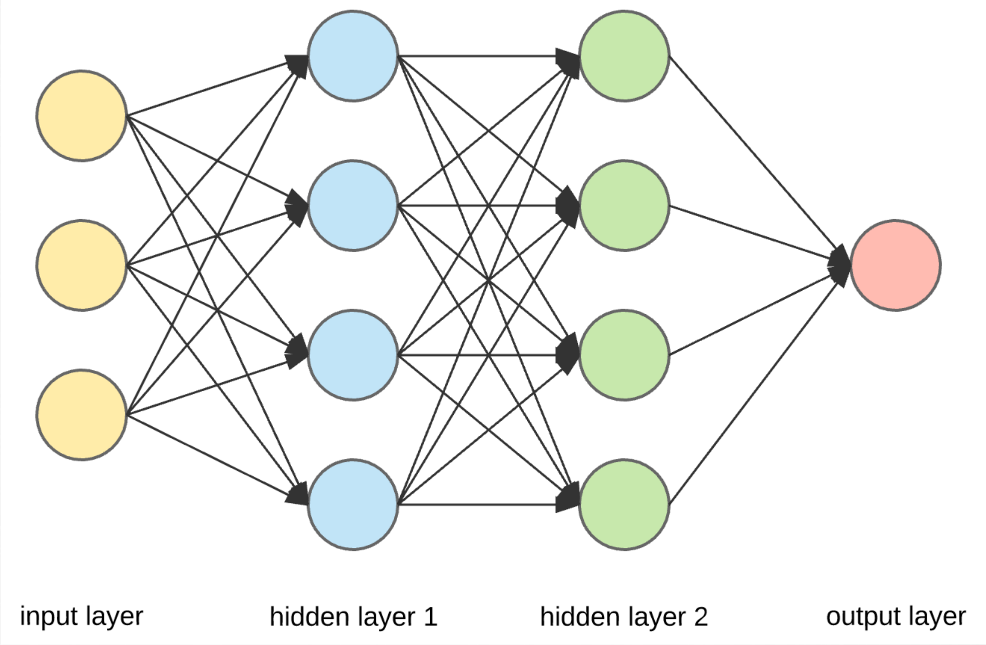
Gradient Boosting algorithms

- For regression and classification problems.
- Produces a prediction model from an ensemble of weak prediction models, typically decision trees.
- Builds the model in a stage-wise fashion.

Neural Networks

- Inspired by the biological neural networks that constitute animal brains.

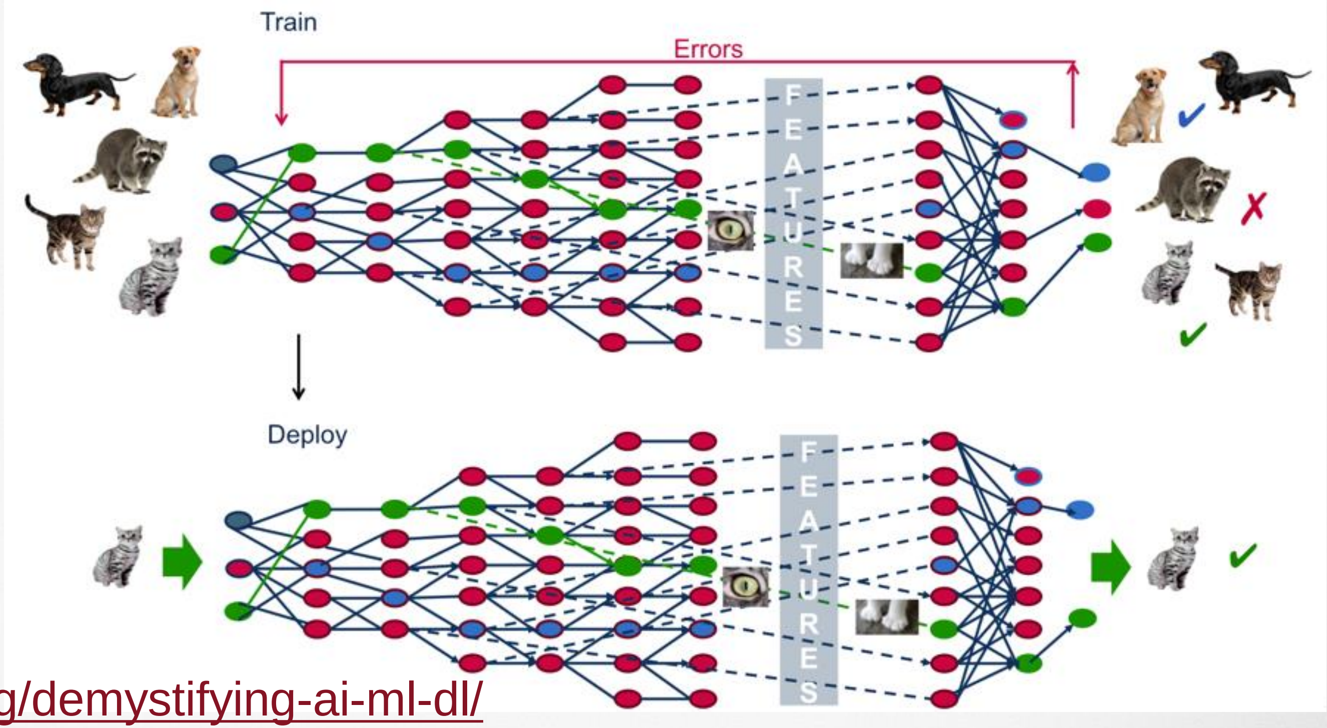
- Systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules.



https://towardsdatascience.com/converting-a-simple-deep-learning-model-from-pytorch-to-tensorflow-b6b353351f5d

Deep Learning

- Neural networks for BIG data
- Have many (thousands?) hidden layers.
- Need GPUs.



https://mapr.com/blog/demystifying-ai-ml-dl/

Ensemble Algorithms

- Combine approaches to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.
- Goal: decrease variance (bagging), bias (boosting), or improve predictions (stacking).

Mathematics & Statistics Preparation

- As much mathematics as possible (though many programs require little mathematics):
  - calculus
  - discrete methods
  - linear algebra
  - …
- Statistics
  - Experience analyzing data

Computing Preparation

- Introduction to Computer Science/Programming
- One should to be able to
  - solve problems using Python
  - use control structures including functions, if-statements, and loops
  - utilize lists
  - have some experience reading and writing files

- Domain knowledge

Questions?

- Christopher Morrell
- chm@loyola.edu

MAA

MATHEMATICAL ASSOCIATION OF AMERICA

WASHINGTON
STATISTICAL
SOCIETY