Introduction to Data Science and Machine Learning

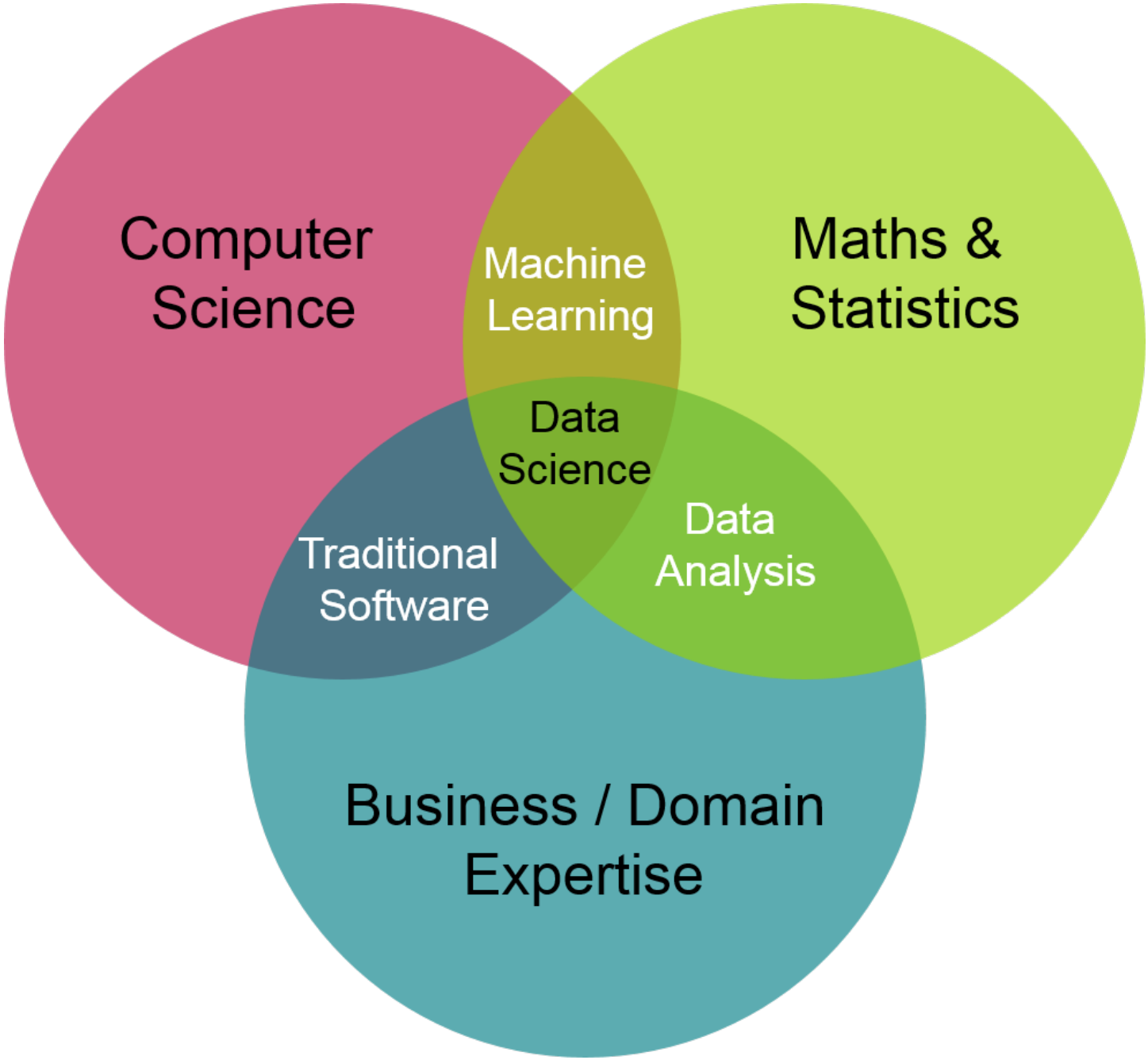
Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

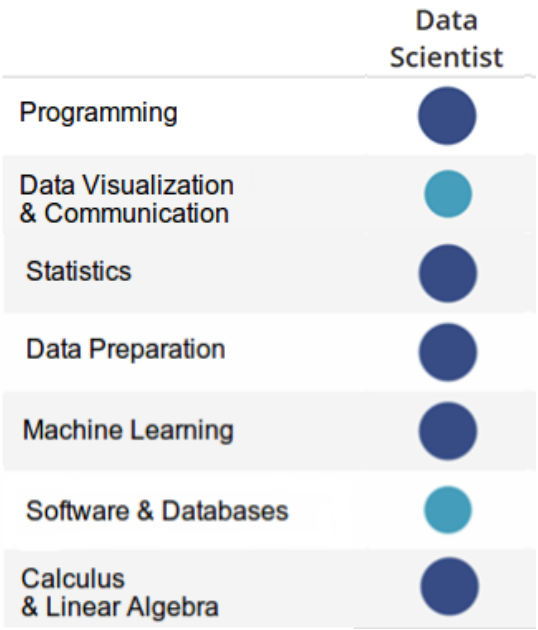
Outline

Statistics vs. Data Science

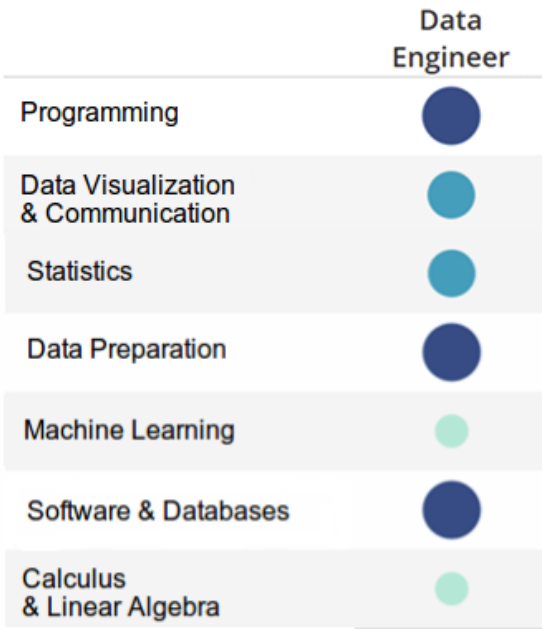
- **Statistics** is a mathematically-based field which seeks to collect and interpret quantitative data.
- **Data science** is a multidisciplinary field which uses scientific methods, processes, and systems to extract knowledge from data in a range of forms. Data scientists use methods from many disciplines, including statistics.
- However, the fields differ in their processes, the types of problems studied, and several other factors.
- Reference: `https://www.displayr.com`


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise

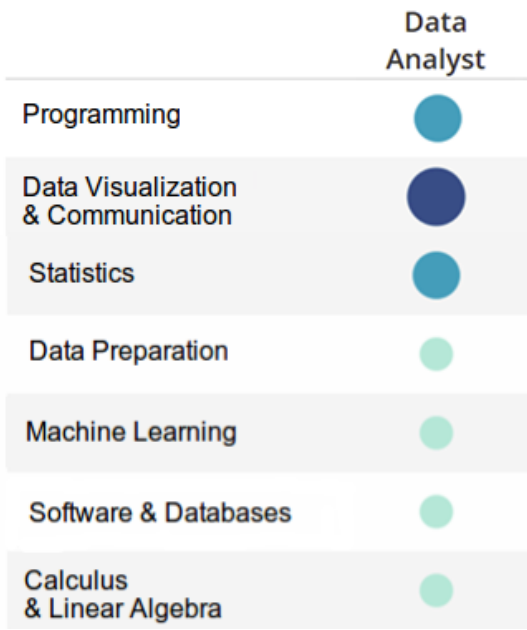
Skills Needed



Salary: $115,503



Salary: $11,061



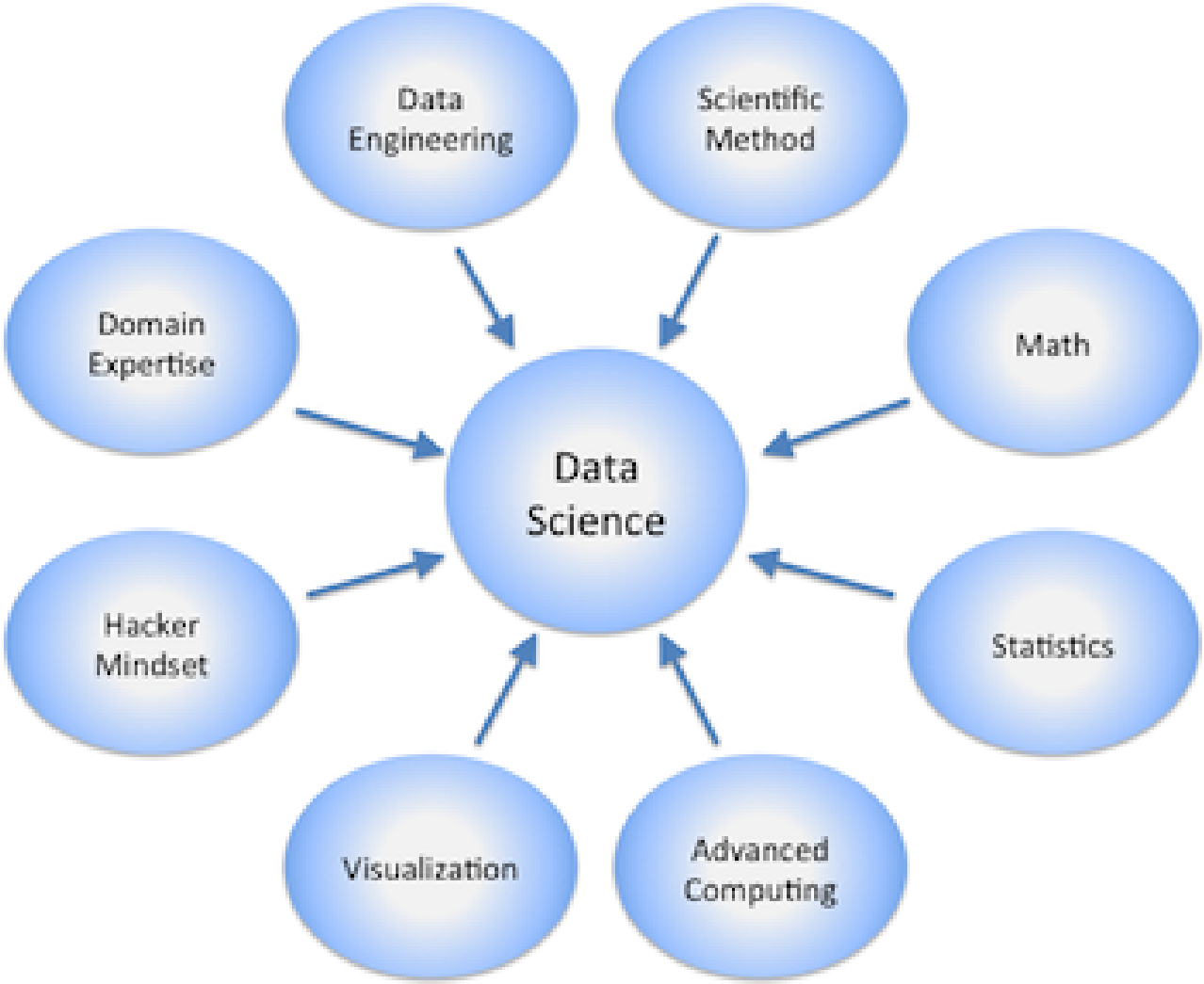
Salary: $82,704 [1]

[1] Source: Glass Door August 13, 2021, average salary.

Data Scientists Versus Data Analyst

Data Analyst Skills	**Data Scientist Skills**
Math & Statistics	Math & Statistics
Programming languages like Python, R , SQL, HTML, JavaScript	Programming languages like Python, R, SAS, Matlab, SQL, Pig, Hive, and Scala.
Spreadsheet Tools (Excel)	Business Acumen
Data Visualization Tools like Tableau	Story-telling and Data Visualization.
	Distributed Computing frameworks like Hadoop.
	Machine Learning Skills

Source: `https://www.dezyre.com/article/difference-between-data-analyst-and-data-scientist/`


Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

Syllabus

Syllabus Fall 2021

Goal of Statistical Learning

The main goal of statistical learning theory is to provide a framework for studying the problem of inference, that is of gaining knowledge, making predictions, making decisions or constructing models from a set of data. This is studied in a statistical framework, that is there are assumptions of statistical nature about the underlying phenomena (in the way the data is generated). [2]

[2] *Introduction to Statistical Learning Theory, O. Bousquet, S. Boucheron, and G. Lugosi*

What is Statistical Learning?

- **Statistical learning** refers to a vast set of tools for understanding data.
- These tools can be classified as *supervised* or *unsupervised*.
  - **Supervised statistical learning** involves building a statistical model for predicting, or estimating, an output based on one or more inputs.
  - **Unsupervised statistical learning** involves inputs but no supervising output.

Source: "An Introduction to Statistical Learning with Applications in R", page 1

Supervised Learning Examples

1. House prices
   - Inputs: square footage, number of rooms, features, whether a house has a garden or not, etc.
   - Outputs: the prices of these houses
2. Will a customer default on their credit card?
   - Inputs: Income, outstanding loans, ect.
   - Output: Default or not default.

Regression versus Classification Problems

- A **regression** problem involves predicting a *continuous* or *quantitative* output value. The *house prices* is an example of a regression problem.
- A **classification** problem involves predicting a non-numerical value—that is, a *categorical* or *qualitative* output value. The *default* problem is an example of the classification problem.

Examples of Unsupervised Learning

1. Clustering is an unsupervised technique where the goal is to find natural groups or clusters in a feature space and interpret the input data.
   - Commonly used for determining customer segments in marketing data.
   - Different segments of customers helps marketing teams approach these customer segments in unique ways. (Think of features like gender, location, age, education, income bracket, and so on.)

2. Dimensionality reduction is a commonly used unsupervised learning technique where the goal is to reduce the number of random variables under consideration.

Supervised Learning Algorithms

- Linear regression
- Logistic regression
- Linear discriminant analysis
- Decision trees
- K-nearest neighbor algorithm
- Neural Networks (Multilayer perceptron)
- Support Vector Machines

Unsupervised Learning Algorithms

- Clustering
  - Hierarchical clustering
  - k-means
  - mixture models
- Neural Networks
- Approaches for learning latent variable models such as
  - Expectation–maximization algorithm (EM)
  - Method of moments
  - Principal component analysis
  - Singular value decomposition