Introduction to Data Science and Machine Learning

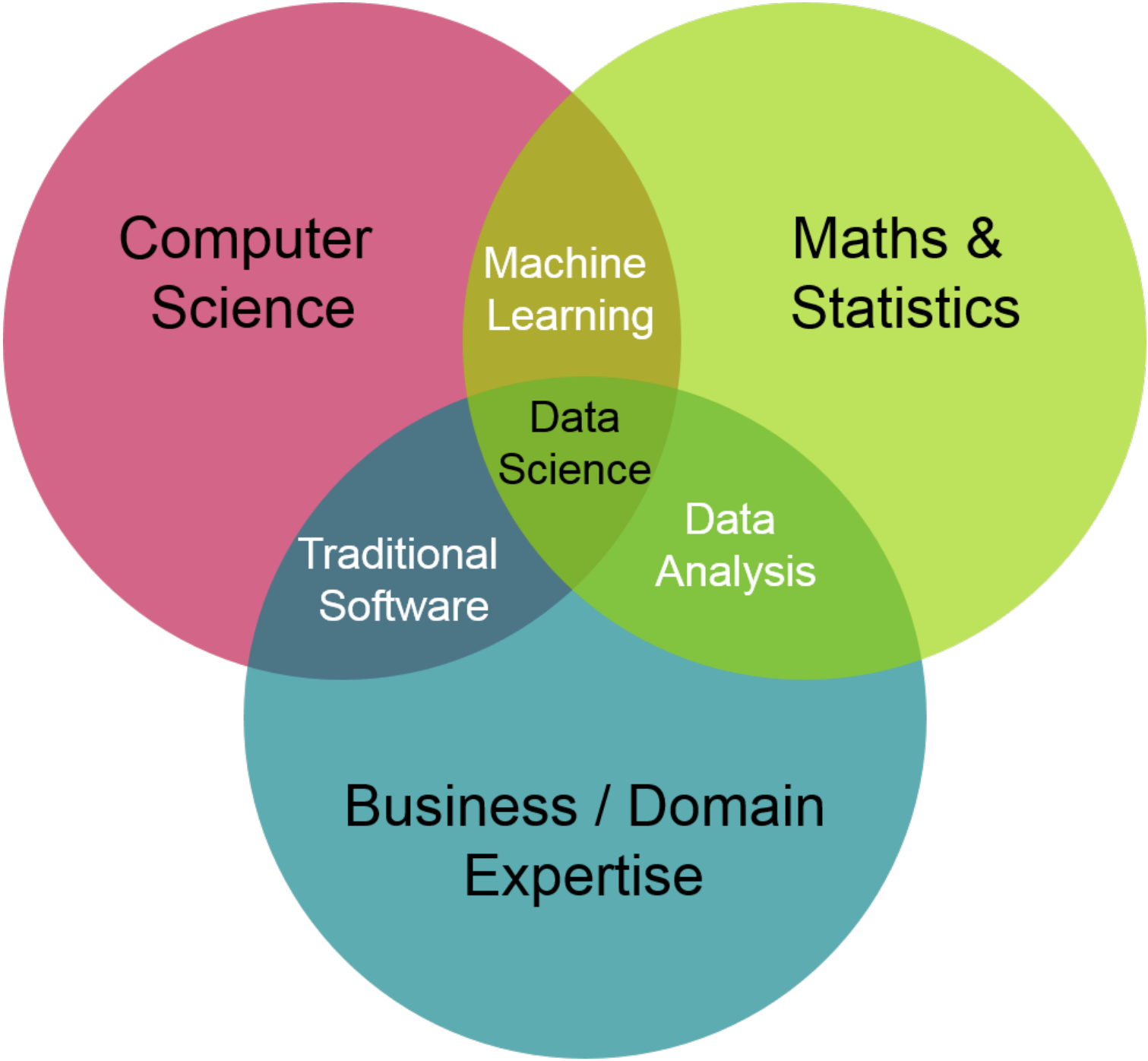
Chapter 1

Dr. Cathy Poliak, cpoliak@uh.edu

University of Houston

Statistics versus Data Science

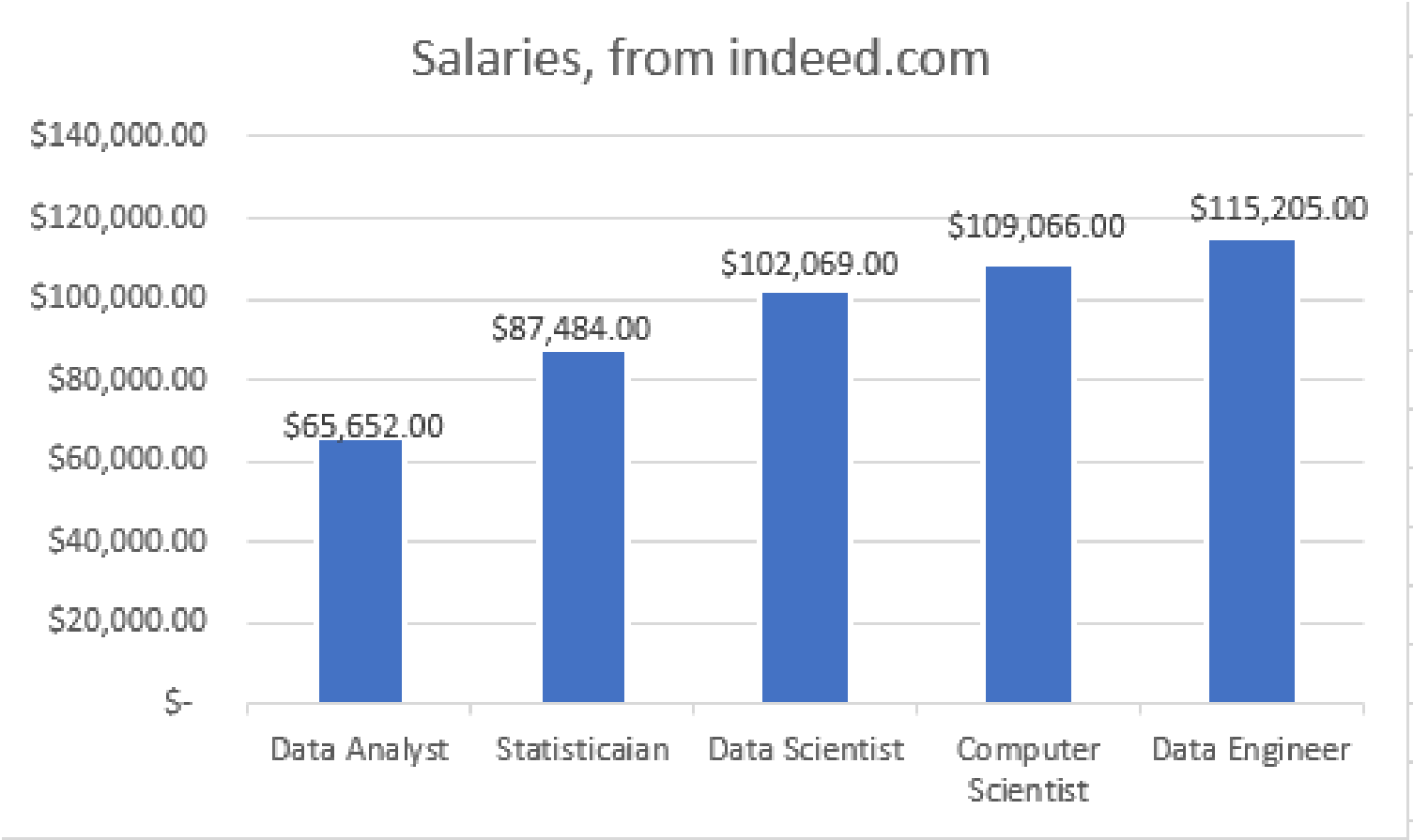
- **Statistics** is a mathematically-based field which seeks to collect and interpret quantitative data.
- **Data science** is a multidisciplinary field which uses scientific methods, processes, and systems to extract knowledge from data in a range of forms. Data scientists use methods from many disciplines,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


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

Different Job Titles

- **Data scientists** are information technology professionals who are responsible for interpreting large amounts of raw data.
- A **data engineer** creates and maintains analytics infrastructure. They build, maintain, and operate databases and large-scale processing systems. They also create data processes. While a data scientist helps ask the right questions, a data engineer helps gather and collect the answers in addition to storing and processing them.
- A **data analyst** collects data, organizes and uses it to reach meaningful conclusions. They are responsible for digesting the data and creating a report to explain the findings.
- A **statistician** uses mathematical formulas and statistical techniques to analyze and interpret data. Statisticians also design mathematical models and use statistical software for data analysis. Statistics is the science of inferring the proportions of the population based on a limited research or survey pool. It's essential for forming conclusions from studies and learning how to solve real-world problems.
- A **computer scientist** uses computer technology to create solutions for a variety of industries. This work includes designing computer systems and writing software to run on websites, computers, phones and other devices. Computer scientists often work with engineers and other specialists in offices or laboratories to develop new technologies or new approaches to using existing technology.

https://www.indeed.com/career/data-engineer


Salaries, from indeed.com
$140,000.00
$120,000.00
$100,000.00
$80,000.00
$60,000.00
$40,000.00
$20,000.00
$-
$65,652.00
$87,484.00
$102,069.00
$109,066.00
$115,205.00
Data Analyst
Statisticaian
Data Scientist
Computer Scientist
Data Engineer

CHARACTERISTICS OF A DATA WHISPERER

Data scientists aren't born—they're made. IT pros from all backgrounds are working to gain the types of skills companies need as the demand for data scientists outpaces the supply of qualified candidates. These are some common personality traits and skills of a data scientist.

Personality traits:
Intellectual curiosity combined with skepticism and good intuition. A tireless problem-solver driven to find a needle in a haystack. Creativity to guide further investigation with the goal of uncovering new information.

Interpersonal skills:
A storyteller who knows how to present data insights to drive business value and who can communicate with people at all levels of an organization.

Business skills:
Data scientists need knowledge far beyond data analysis and statistics. They need the business savvy to discover patterns that can be used to identify risks and opportunities and the leadership skills to influence business leaders to make data-driven decisions.

Education:
Bachelor's degree in statistics, data science, computer science or mathematics.

Specialized skills:
Data mining, machine learning and distributed computing. Ability to integrate structured and unstructured data. Experience with statistical research techniques, including modeling, data mining, clustering and segmentation.

Tools of the trade:
Familiarity with Hadoop, Pig, Hive, Spark and MapReduce. Comfortable with SQL, Python, Perl or other scripting languages, as well as statistical computing languages such as R.

CONTENT: BRIDGET BOTELHO, EDITORIAL DIRECTOR, BUSINESS APPS & INFO MANAGEMENT GROUP; DESIGN: LINDA KOURY; PHOTO: 4X6/GETTY IMAGES

Syllabus

Fall 2024 Syllabus

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

Statistical Learning versus Machine Learning

- Machine Learning is a sub-field fo Artificial Intelligence.
- Statistical Learning is a sub-field of Statistics.
- Both focus on supervised and unsupervised problems
  - Machine learning has an emphasis on large scale applications and prediction accuracy.
  - Statistical learning emphasizes models and their interpretability, precision, and uncertainty.

Adapted from [1]

[1] https://hastie.su.domains/ISLR2/Slides/Ch1_Inroduction.pdf

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 Usupervised 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 Algorthims

- 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

What is Statistical Inference?

- A **statistical inference** aims at learning characteristics of the population from a sample
- The population characteristics are *parameters*
- The sample characteristics are *statistics*

The Characteristics

In order to use the correct inference we need to know what type of characteristics we have from the data.

- Are these characteristics quantitative or categorical?
- Do we have a response variable (output or dependent variable) and factors (input, independent variable or predictor)?

Example `mtcars`

- Response variable - *mpg*
- Variable 1 - *cyl*
- Variable 2 - *disp*
- Variable 3 - *hp*
- Variable 4 - *wt*
- Variable 5 - *am*

Regression

Example 2

Response = Delayed or not

Classification

Data file: ontime.csv

- variable 1 - DAY.OF.MONTH
- variable 2 - DAY.OF.WEEK
- variable 3 - CARRIER
- variable 4 - ORIGIN airport where the flight left
- variable 5 - DEST, airport where the flight landed
- variable 6 - DEP.DELAY, number of minutes that the departure was delayed, if negative left early
- variable 7 - ARR.DELAY, number of minutes that the landing was delayed, if negative came in early
- variable 8 - DISTANCE

1500 flights randomly sampled from the month of May 2021.

General Approach for Supervised Learning

- Let Y be the response (dependent variable).
- Let $X = (X_1, X_2, \ldots, X_p)$ be p different predictors (independent) variables.
- We assume there is some sort of relationship between X and Y, which can be written in the general form

$$Y = f(X) + \epsilon$$

- Statistical leaning refers to a set of approaches for estimating f.

Reasons for Estimating f

- Prediction: we want to predict Y, using $\hat{Y} = \hat{f}(X)$.
  - $\hat{f}$ is often treated as a **black box**.
  - The black box means that we are not typically concerned about the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.
- Inference: we want to know how Y is affected as X changes.
  - In this situation we wish to estimate f.
  - Thus $\hat{f}$ cannot be considered as a black box because we do want to know the exact form of $\hat{f}$.

Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**
  - $\hat{f}$ is not a perfect estimate for f, and this inaccuracy will introduce some error.
  - We can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique to estimate f.
- **Irreducible error**
  - However, even if it were possible to form a perfect estimate for f, so that our estimated response took the form $\hat{Y} = f(X)$, our prediction would still have some error in it!
  - This is because Y is also a function of ϵ, which, by definition, cannot be predicted using X.
  - Therefore, variability associated with ϵ also affects the accuracy of our predictions.

Inference Questions

If we are interested in inference we are asking we are interested in answering the the following questions:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Inference and Prediction Examples

- Flight delays - our response variable (output) is if a flight arrival is delayed.
- Prediction - using our model to determine if any future fights will be delayed.
- Inference - determining what variables (inputs) affect the delay of the flight.
- Some models are only used for prediction, some for inference, some can do both. It all depends on what question you want to know.

Common Characteristics to Estimate f

- We will have n different data points **observations**.
- These observations can be split into **training data** and **test data**
  - Training data - observations used to train or teach the method how to estimate f.
  - Test data - observations used to determine the accuracy of estimating f.

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*