- ❖ **WEEK 1**

What is Data Science?

- Data science is the field of exploring, manipulating, and analyzing data, and using data to answer questions or make recommendations.

Fundamentals of Data Science

- It’s essential to clarify the question that the organization wants answered
- "What data do we need to solve the problem, and where will that data come from?"

The Many Paths to Data Science

- As data science is not a discipline traditionally taught at universities, contemporary data scientists come from diverse backgrounds such as engineering, statistics, and physics.

Advice for New Data Scientists

- Curiosity
- Judgemental
- Good story teller (Argumentative)

SUMMARY

- Data science is the study of large quantities of data, which can reveal insights that help organizations make strategic choices.
- There are many paths to a career in data science; most, but not all, involve a little math, a little science, and a lot of curiosity about data.
- New data scientists need to be curious, judgemental and argumentative.
- Why data science is considered the sexiest job in the 21st century, paying high salaries for skilled workers.

Old problems, new problems, Data Science solutions

- Transport: Uber, understand traffic performance (peak hours, congested routes)
- Environment: Solve bacteria problems (explore bacterias across the coast), algorithmic models to assess the findings, better predictions
- Identify the problem and establish a clear understanding of it, gather the data for analysis, identify the right tools to use and develop a data strategy.
- Once the data is extracted, you can develop a machine learning model

Data Science Topics and Algorithms

- Regression: regression allows you to compute that constant that you didn't know. That it was $2.50, and it would compute the relationship between the fare and the distance and the fare and the time.
- Data visualization (R language)
- Artificial neural networks
- Nearest neighbor: Using complicated machine learning algorithms does not always guarantee achieving a better performance. Occasionally, a simple algorithm such as k-nearest neighbor can yield a satisfactory performance comparable to the one achieved using a complicated algorithm. It all depends on the data.
    - In any field, and data science is no different, a simple solution is always preferred over a complicated one, especially if the performance is comparable.
- Structured data

Cloud for Data Science

- Allows multiple entities to work with same data at the same time
- Enhances productivity

<u>LESSON SUMMARY</u>

- The typical work day for a Data Scientist varies depending on what type of project they are working on.
- Many algorithms are used to bring out insights from data.
- Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily.

- ❖ WEEK 2

Foundations of Big Data

- Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.
- V's of Big Data
    - Velocity: the speed at which data accumulates. Ex: hours of footage are uploaded to YouTube
    - Volume: scale of the data, or the increase in the amount of data stored. Ex: from seven billion people, approximately 2.5 quintillion bytes are stored everyday.
    - Variety: diversity of the data. Data comes from different sources, machines, people, and processes, both internal and external organizations. Ex: text, pictures, film, sound, health data from wearable devices
    - Veracity: quality and origin of data, and its conformity to facts and accuracy. Consistency, completeness, integrity, and ambiguity
    - Value: ability and need to turn data into value

What is Hadoop?

- Mapper process and the second process is called a reduce process
- These big data clusters scale linearly
- Yahoo hired someone named Doug Cutting who had been working on a clone or a copy of the Google big data architecture and now that's called Hadoop.
- How does data science differ from traditional subjects like statistics?
    - Machine learning
    - We can take really large data sets and look for patterns
    - the combination of traditional [technique] areas computer science probability, statistics, mathematics all coming together in this thing that we call Decision Sciences

Data Science Skills & Big Data

- Python
- Statistics

Data Mining

1. The first step in data mining requires you to **set up goals** for the exercise
2. **Preprocessing data** is an important step in data mining. Often raw data is messy, containing erroneous or irrelevant data. Data should be subject to checks to ensure integrity. Lastly, you must develop a formal method of

dealing with missing data and determine whether the data are missing randomly or systematically.

3. After the relevant attributes of data have been retained, the next step is to **determine the appropriate format in which data must be stored**. An important consideration in data mining is to reduce the number of attributes needed to explain the phenomena.
    a. This may require transforming data Data reduction algorithms, such as *Principal Component Analysis*
4. The transformed data must be **stored** in a format that makes it conducive for data mining. The data must be stored in a format that gives unrestricted and immediate read/write privileges to the data scientist.
5. After data is appropriately processed, transformed, and stored, it is subject to **data mining**. This step covers data analysis methods, including parametric and non-parametric methods, and machine-learning algorithms. A good starting point for data mining is data visualization.
6. After results have been extracted from data mining, you do a **formal evaluation of the results**. Formal evaluation could include testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data. This is known as an “in-sample forecast”. In addition, the results are shared with the key stakeholders for feedback, which is then incorporated in the later iterations of data mining to improve the process.

LESSON SUMMARY

- How Big Data is defined by the Vs: Velocity, Volume, Variety, Veracity, and Value.
- How Hadoop and other tools, combined with distributed computing power, are used to handle the demands of Big Data.
- What skills are required to analyze Big Data?
- About the process of Data Mining, and how it produces results.

What’s the difference?

- The term big data refers to data sets that are so massive, so quickly built, and so varied that they defy traditional analysis methods such as you might perform with a relational database.
    - Big data is often described in terms of five V's; velocity, volume, variety, veracity, and value.
- Data mining is the process of automatically searching and analyzing data, discovering previously unrevealed patterns.
    - Once this is done, insights and patterns are mined and extracted using various tools and techniques ranging from simple data visualization tools to machine learning and statistical models.

- Machine learning is what enables machines to solve problems on their own and make accurate predictions using the provided data.
    - Deep learning is a specialized subset of machine learning that uses layered neural networks to simulate human decision-making
        - Deep learning algorithms can label and categorize information and identify patterns. It is what enables AI systems to continuously learn on the job and improve the quality and accuracy of results by determining whether decisions were correct.
    - A neural network in AI is a collection of small computing units called neurons that take incoming data and learn to make decisions over time. Neural networks are often layer-deep and are the reason deep learning algorithms become more efficient as the data sets increase in volume, as opposed to other machine learning algorithms that may plateau as data increases.
- Data Science is the process and method for extracting knowledge and insights from large volumes of disparate data.
    - It's an interdisciplinary field involving mathematics, statistical analysis, data visualization, machine learning, and more.
    - Data Science can use many of the AI techniques to derive insight from data.

Neural Networks and Deep Learning

- How does a neural network work?
    - So a neural network is trying to use computers, a computer program that will mimic neurons, how our brains use neurons to process things, neurons and synapses and building these complex networks that can be trained. So this neural network starts out with some inputs and some outputs, and you keep feeding these inputs in to try to see what kinds of transformations will get to these outputs. And you keep doing this over, and over, and over again in a way that this network should converge. So these inputs, the transformations will eventually get these outputs.
- Speech recognition is an example of deep learning

Applications of Machine Learning

- Predictive analytics
- Decision trees, Bayesian Analysis, naive Baayes, lots of different techniques
- Recommendations (like in Netflix or Instagram)
- Fraud detection (banking and finance)
    - You have to learn from all of the transactions that have happened previously and build a model, and when the charge comes through you have to compute all this stuff and say, "Yeah we think that's ok," or

"hmm, that's not so good. Let's route it to, you know, our fraud people to check."

Regression

- Why regress? A whole host of questions could be put to regression analysis. Some examples of questions that regression (hedonic) models could address include:
    - How much more can a house sell for an additional bedroom?
    - What is the impact of lot size on housing price?
    - Do homes with brick exteriors sell for less than homes with stone exteriors?
    - How much does a finished basement contribute to the price of a housing unit?
    - Do houses located near high-voltage power lines sell for more or less than the rest?

LESSON SUMMARY

- The differences between some common Data Science terms, including Deep Learning and Machine Learning.
- Deep Learning is a type of Machine Learning that simulates human decision-making using neural networks.
- Machine Learning has many applications, from recommender systems that provide relevant choices for customers on commercial websites, to detailed analysis of financial markets.
- How to use regression to analyze data.

❖ WEEK 3

How Should Companies Get Started in Data Science?

1. Start gathering, capturing, archiving, measuring data
2. Apply algorithms and analytics

Applications of Data Science

- Recommendations (Amazon, Netflix, restaurants, places)
- Personal assistants (Siri, Alexa)
- Route guidance systems (GPS, UPS)

LESSON SUMMARY

- Data Science helps physicians provide the best treatment for their patients, and helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
- That companies can start on their data science journey by capturing data. Once they have data, they can begin analyzing it.

- Some ways that data is generated by consumers.
- How businesses like Netflix, Amazon, UPS, Google, and Apple use the data generated by their consumers and employees.
- The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.

Recruiting for Data Science

- Curiosity
- Sense of humor
- Storytelling
- Technical skills

LESSON SUMMARY

- Data Scientists need programming, mathematics, and database skills, many of which can be gained through self-learning.
- Companies recruiting for a Data Science team need to understand the variety of different roles Data Scientists can play, and look for soft skills like storytelling and relationship building as well as technical skills.
- High school students considering a career in Data Science should learn programming, math, databases, and, most importantly, practice their skills.
- The length and content of the final report will vary depending on the needs of the project.
- The structure of the final report for a Data Science project should include a cover page, table of contents, executive summary, detailed contents, acknowledgements, references and appendices.
- The report should present a thorough analysis of the data and communicate the project findings.