**Introduction**

Data science is multidisciplinary and encompasses both mathematical and non-mathematical skills. It requires a wide range of skills like data collection, preparation, analysis, visualization, along with non-mathematical skills like understanding the big picture impact of the story the data is trying to tell.

**1.1 Many Skills**

The skills required through the life cycle of a data science problem are:

1. Learning application domain: Understanding the context of the data of the problem.
2. Communication: Work loses value if the technical analysis of the data is not conveyed properly to the end-users.
3. Big picture: Understanding how the problem we are trying to solve works in the larger scheme of things.
4. Knowing data representation: Understanding how the data is stored and what the metadata signifies is crucial to proceed.
5. Data transformation and analysis: Changing the raw data to be meaningful and then conducting the analysis of the data is one of the most key steps in the process.
6. Visualization and presentation: Explaining the data analysis in a more intuitive and palatable way to the end-user.
7. Attention to quality: Quality of both the data and the technical analysis determines the end result.

**1.2 Steps in Doing Data Science**

Has 4 A’s during the lifecycle of a data science problem, namely: data architecture, data acquisition, data analysis and data archiving.

The 4 A’s works in the following ways:

1. Architecture: How the data is stored, shared, who has access to which part of data within the organization.
2. Acquisition: How the data is obtained
3. Analysis: It is the heart of data science. Identifies what insights we want to derive and which technique is suitable and then conducts the analysis.
4. Archiving: Making the decision, where and how to store the results of the analysis.

**Chapter 1: About Data(Data Science Overview Video)**

The data transforms in the following way:

1. Raw Data: Data coming from the source, won’t have any insight.
2. Information: Starts to provide value. Has some context.
3. Knowledge: A trend analysis can be done with knowledge.
4. Intelligence: gives insights about the big picture.
5. Wisdom: new business strategies are determined with wisdom.

**Chapter 2: Identifying Data Problems(Data Science Overview Video)**

* For a data science scope, boundaries, focus area, and context are important.
* Anomalies in data or outliers are a result of an error or exceptional circumstances. It is crucial to deal with them to avoid producing erroneous results from the analysis.
* Domain identification is essential to zero in on a problem.
* Subject Matter Experts(SMEs) provide the domain knowledge to data scientists by explaining what the data at hand implies and the value that can be derived out of the analysis of the data.

**Chapter 3: Getting Started With R**

* R is an open-source software program and is a popular choice amongst data scientists.
* In the console, R is ready when the ‘>’ pops in a new line. The R command in the console goes after the symbol.
* In the console, basic mathematical computations can be conducted.

Some R commands with syntax are as follows:

1. variableName <- ()  
   Declares a variable ‘variableName’ and assigns a value to it. To assign a string value, the string must be within double quotes (“”).
2. +, -, \*, /  
   These operators are addition, subtraction, multiplication and division. Writing the words of the operators will give an error.
3. sum()  
   Gives the sum of the values in the brackets
4. mean()  
   Gives the average of the values in the brackets
5. range()  
   Gives the range of the values in the brackets
6. c(11, 12, 45, 90)  
   Creates a vector with 4 elements 11, 12, 45, 90. ‘c’ stands for concatenate. The indices in the vector begin from 1.

**Questions from the video:**

* What is the difference between structured and unstructured data? GIve some examples of each.
* Structured data is highly organized and has clearly defined metadata labels, whereas, unstructured data is a mashup of various types of data and does not fit into a box. Parts of unstructured data may be structured, but overall, it has little to no structure,  
  Data in tabular form or comma separated files, with column entries, is an example of structured data.  
  Blog entries, webpages, user data stored by sites like Facebook, Netflix are all examples of unstructured data.

Unstructured data could have a lot more information than structured data, however, the analysis for unstructured data is a lot more challenging.

* Where might there be data analysis in the process? Provide some values.
* As a store owner, all the customers’ receipts can be analysed to see whether there are any items that are bought frequently with cookies and rearrange the store’s layout to nudge the customers into buying other things on the way between the two items.
* Another analysis could be which brands and products of cookies are being sold more and arrange the shelves such that the popular ones are at the eye-level whereas, the less popular ones are on the higher or lower shelves.

**Question to the professor:**

With other options of programming languages, how do we decide between the programming languages? Do they have limitations when it comes to a particular type of data science problem?