Rashika Singh
IST 687

# Lecture Notes-Module 1

**Data Science:**
Data Science is emerging area of work concerned with collection, preparation, analysis, visualization, management and preservation of large collections of information. It has a number of skills associated with it and is different from mathematics and statistics, you need to communicate, identify problems, analyze and find solutions. Data Scientists play the most active role in the four A's of data" data architecture, data acquisition, data analysis, and data archiving. Data Science is multidisciplinary field consisting of visualizations, statistics, ML among others.

**Life cycle of Data Science:**
1. Learning the application domain- The data scientist must learn how data will be used.
2. Communicating with data users- They must possess strong skills for learning preference of users.
3. Seeing the big picture- They must imagine how data will be used for different systems.
4. Knowing how data can be generated- Must have a clear understanding of how data is stored, used and information about metadata.
5. Data transformation and analysis- Data scientists must know how to transform, summarize, and make inferences from the data.
6. Visualization and presentation- Visualizations must be used to effectively communicate insights about data.
7. Attention to quality- They must know limitations of data and know how to work with accuracy.

**What is the difference between structured and unstructured data? Provide examples of each.**
Structured data is a data which is clearly defined and is easily readable and do analysis on it. There are well defined rows and columns which makes it easy to fit in databases and is organized. Examples of it include- data in csv files, SQL databases, tables. Unstructured data is everything else which is not easily searchable and interpretable. Examples of unstructured data include text, audio, websites, social media data

**Where might there be data analysis in this process? (cookie example)**
Suppose we consider a case of cookie being sold in the supermarket. There is data analysis involved in most of the steps of selling a cookie. When a customer wants to purchase something, he writes down the list of things which is data. The supermarket uses inventory to manage the stock of items in the market. If the cookies on the shelf are less, then they replace the stock. Even when a customer hands over a coupon and the cashier scans it, he offers them a discount. Data analysis is also involved in the process of determining if a particular marketing campaign is effective in increasing sales. The sales market could use visualizations to determine which types of cookies were sold and how many which could be the basis for inventory management. Raw data does not have value and data should be turned into information.

**Getting Value from Data:**
Data->Information->Knowledge->Intelligence->Wisdom

Rashika Singh
IST 687

**Data Science process:**

1. Domain analysis, identification, understanding- to focus on problem or opportunity
2. Identify subject matter expert- engage people essential in the process
3. Question/interview/observation process
    a. Stories-Identify what people do, how they do, information produced, process and information touch points, decisions
    b. Anomalies- define typical events, processes, activities and identify the exceptions
    c. Risks and uncertainty

**Getting started with R:**

R is open source data analysis which is powerful, flexible, extensible and command line oriented.
Strings- sequence of characters
Example- myText<- "this is a piece of text" #Sample string
Integer list is all same type/mode. R refers to list as vector. Vector is basic form of storage in R and it consists of list of items of same mode.

**Use of c ()**- The letter c in front of opening parenthesis stands for concatenate which is join things together. Numbers and text can be collected in lists (Vector)
A vector can be stored in names location using the location (<-)
You can get data object named location by typing the name.

Example-
myFamilyAges<- c (43,42,12,8,5) #vector of numbers
sum(myFamilyAges) #sum of all numbers
myRange<-range(myFamilyAges) #gets smallest and largest numbers

**Error example:**

Test <- this is a test
#gives an error message due to strings not specified in "".
Test<- "this is a test"

**Code for condition-**

myAge <- 40 #define age as 40
if(myAge<50)"not so old" #if age is less than 50, it prints "not so old"

**Question- Is R better for structured or unstructured data?**