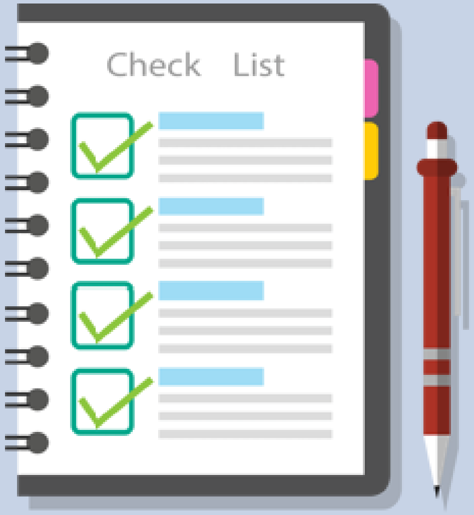
Introduction to Data Science

Digital Lync

Index



Data Science

Data

- Qualitative data
- Quantitative data

Data (Storage)

- Structured
- Semi structured
- Unstructured

Data Analytics Process

Catalog of ML methods

Utilities & Basic Setup

Data Science

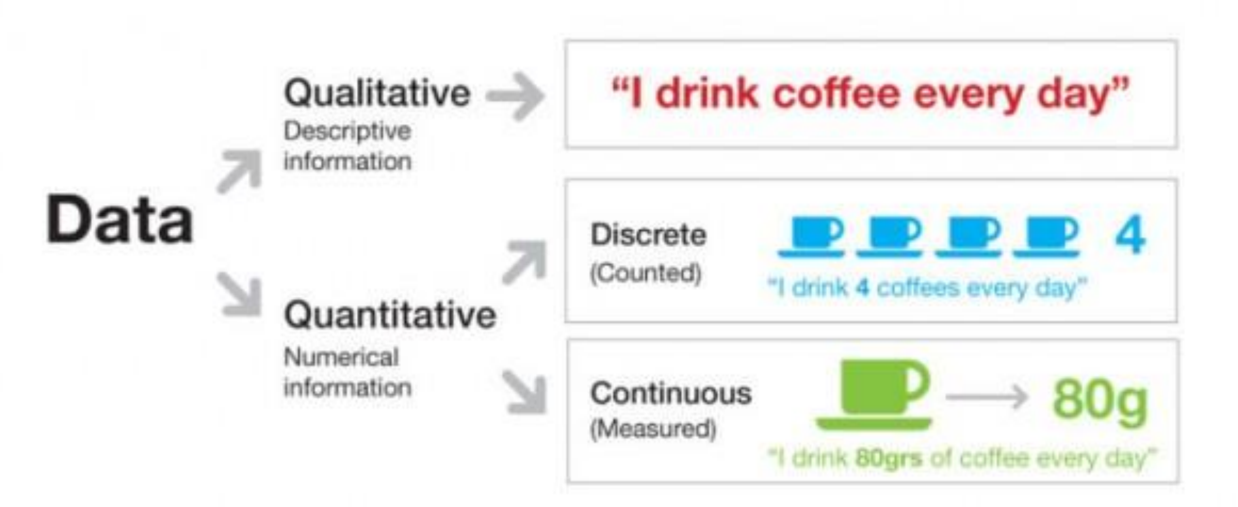
Goal of data science is to extract the meaningful insights form the data & effectively tell a story that can be easily understood by non-professionals.

Data Science Jargons

Jargons	Year	Description
Machine Learning	1980’s	Focus was on algorithms & the amount of data was limited
Predictive Analytics / Data mining	1990’s	Used Algorithms that are developed & applied on Large amount of data
Big data Analytics	2000’s	Focus was on computing on big volume of data in distributed fashion
Data Science	2010’s	Filed where complex Algorithm works on large volume of data to solve business problem Lot of emphasis on visualization & story telling

Data

Data is a collection of facts, such as numbers, words, measurements, observations or even discretion of things.



Data

Qualitative
Descriptive information

"I drink coffee every day"

Quantitative
Numerical information

Discrete
(Counted)

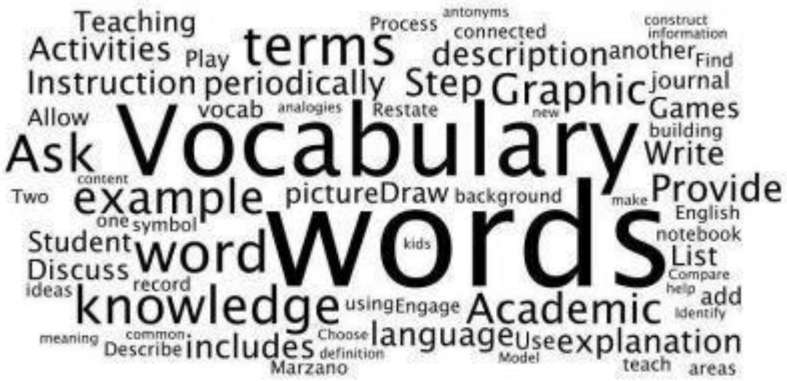
4

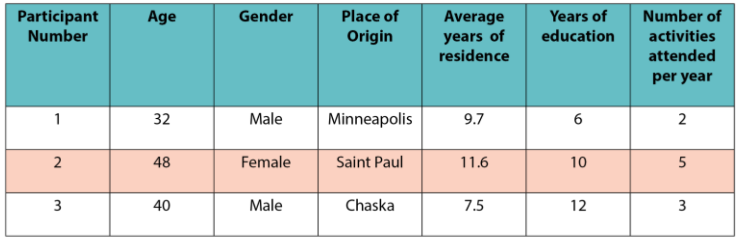
"I drink 4 coffees every day"

Continuous
(Measured)

80g

"I drink 80grs of coffee every day"





Participant Number	Age	Gender	Place of Origin	Average years of residence	Years of education	Number of activities attended per year
1	32	Male	Minneapolis	9.7	6	2
2	48	Female	Saint Paul	11.6	10	5
3	40	Male	Chaska	7.5	12	3

How can I use this data – Analytics

Qualitative data: Descriptive information

- "Your friends house is pretty good"
- "Amazon Echo is the best AI Assistant"
- "Computer Vision is the new area of research"

Usage: Text mining, NLP, Sentiment analysis.

Sentiment analysis

Quantitative data: Data represents some quantity (numerical value).

It is of two types

- Discreate data: can take certain value (whole number)
- Continues data: can take any value (range of values)

Usage: Predictive analytics, Classification, Regression

Sales Predictions

Data, Data Analytics

Various types of Data (Storage):

1.	*Structured data:*	DB, ERP systems, CRM
2.	*Semi- structured data:*	Log files, XMLs
3.	*Unstructured data:*	Facebook, twitter

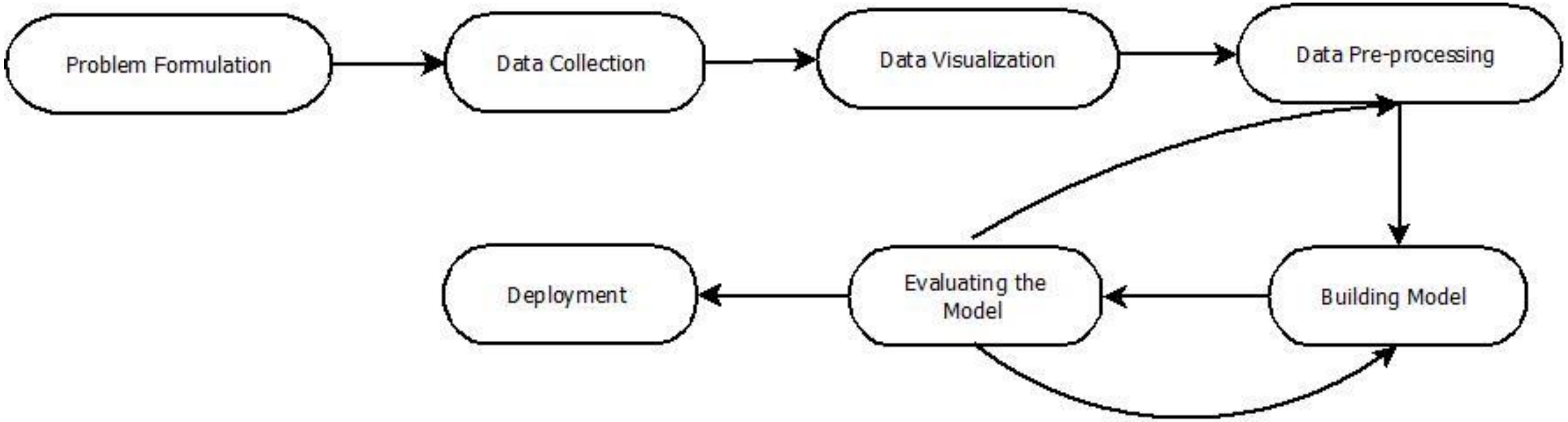


Fig: Data Analytics Process

General Task of Data Scientist

1. Get Domain Knowledge
2. Define the problem statement clearly
3. Pre-process the data to fix data issues
4. Visualize data for better understanding & to see basic patterns
5. Identify what kind of problem it is
6. Identify appropriate modeling technique & build models
7. Analyze the results & iterate if needed
8. Visualize outputs & story telling

Catalog of ML methods

Descriptive statistics

- Central tendency
- Correlations
- Sampling & distribution
- Hypothesis testing

Predictive Methods (Supervised)

- Simple Linear Regression
- Multiple Linear Regression
- Supported Vector Machines (SVM)
- Neural Networks
- Gradient Boosting

Optimization Methods

- Operational research
- Linear Programming
- Genetic Algorithm

Classification(unsupervised)

- Clustering
  - K-means
  - Hierarchical
- Association Rules
- Market basket Analysis

Classification (Supervised)

- Logistic Regression
- Decision Trees
- Bayesian Analysis or classification
- Random forest

Utilities & basic Setup

Python installation

- Download Anaconda 3: https://www.anaconda.com/download/
- IDE: Jupyter Notebook or Spyder