Data Analytics with Python

Trainer: Michael Li

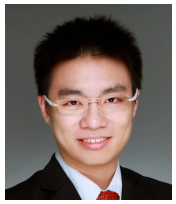
Venue: NTUC Learning Hub

Date Conducted: 23 Oct 2018

Outline

- What is Data Analytics
- What is Python
- Analytics in Industry
- About Data
- Data Tutorial
- Descriptive, Diagnostic, Predictive
- Analytics Tutorial

About Your Trainer



Michael Li

Biomedical Engineering Degree, BioInformatics Specialist

Expansive Experience

- Biomedical Devices and Diagnostics
- **Systems Engineering** and Server Administration
- **Analytics Consultancy**

Data Scientist

- **Innumerous strategies** to solve **Data-Fusion** problems
- **Solution Architect**
- **Recommendation Engines**
- **Semantic Technologies**
- **Fraud Analytics**
- **Text Mining**

www.linkedin.com/in/michaelqali

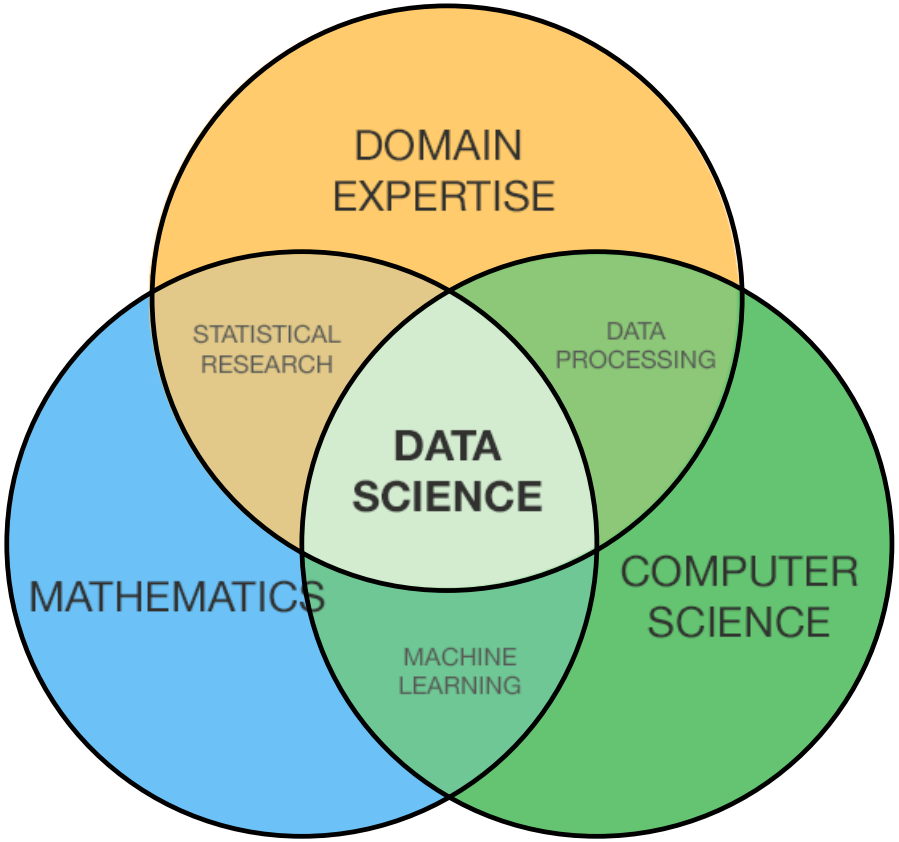
DATA ANALYTICS WITH PYTHON

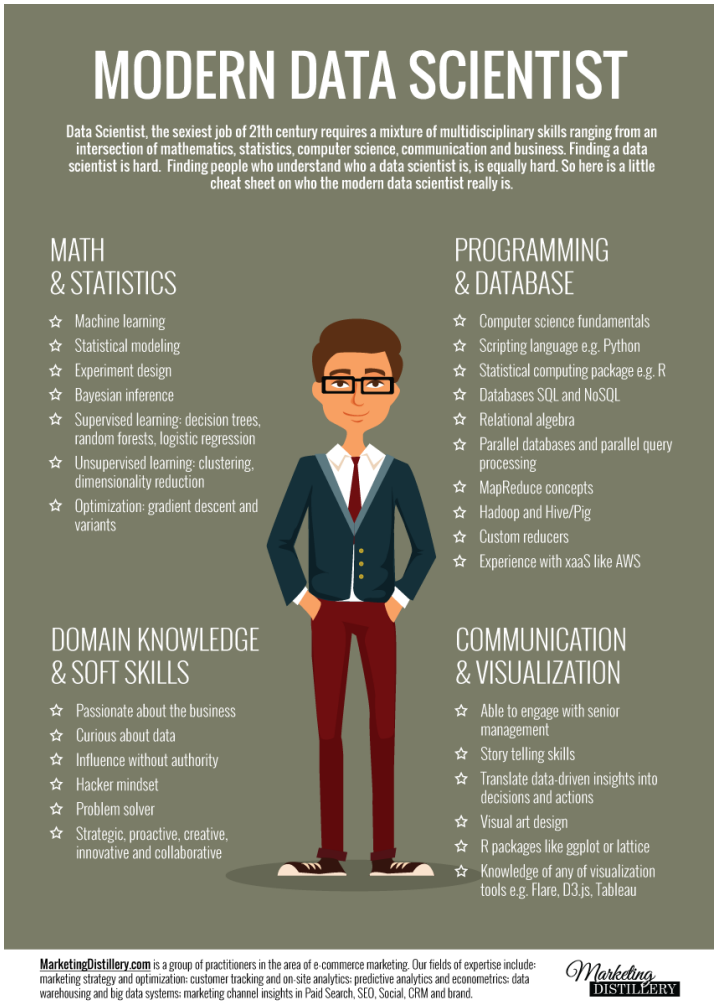
What is Data Analytics

What is Data Analytics

- The science of drawing insights from raw information sources
- Also commonly known as Data Mining, Data Science,
- Has multiple branches:
  - Big Data
  - Text Mining
  - Business Analytics
  - Machine Learning
  - Quantitative Analytics

What is Data Analytics





Data Analytics Skills

Main Skills Category	Why	Examples
Business and Communication	Understand and extract Business Goals, Key Performance Indicators, Problem Statements. Define, formulate and deliver actionable targeted analyses.	CRISP-DM
Visualization and User Interface	Create easily digestible, usable dashboards and interactive charts.	Business Intelligence (Tableau, QlikSense) Visualization Libraries (D3.js, bokeh)
Statistics	Pick out significant factors from data. Find outliers in data points. Understand the quality of the data provided. Quantify the accuracy or applicability of tests done.	Chi-square, Standard Deviation, Pearson Correlation
Software Engineering	Data Management e.g. Extract-Transform-Load. Automation of Data Processing and Preparation.	Programming skils (Java, Python, R, C/C++/C#, Hadoop)
Machine Learning	Assess feasibility of running Predictive Modelling on data. Choosing the right model for the dataset. Tuning parameters used in the model.	Clustering (Meanshift, K-means, etc.), Classification (RandomForest, Naïve bayes, etc.), Active Learning, Deep Learning
Science & Engineering	Formulate an analysis approach to a problem. Identify and apply measurable performance indicators on method used. Apply best practices and standard procedures for different domains.	AB-testing, Scientific Method, Standards/Body-Of-Knowledge
Social Skills	Keep in touch with current technology trends and techniques. Maintain network of contacts from various domains.	Knowledge Sharing between other Professionals, Confidence and Humility

Methodology

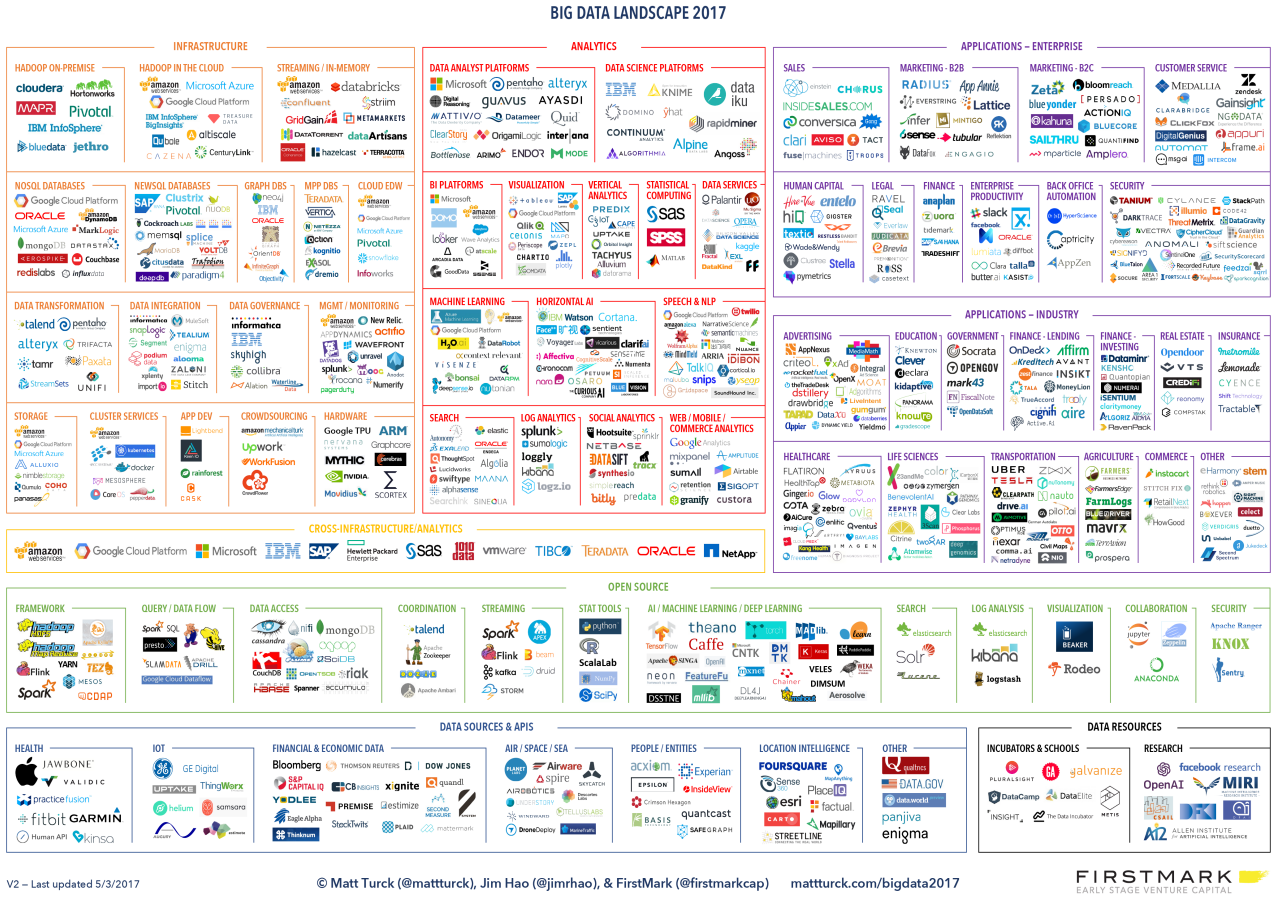
- Cross-Industry Standard Process for Data mining (CRISP-DM)
- CRISP-DM is a very general approach and widely used in analytics projects.

- Due to the highly varied nature of data from different entities, the length of an Analytics project may range from 1 week to 1 year (excluding deployment).

- Thus, it is now a recommended practice to conduct a small-scale feasibility study first.

6 Phases of CRISP-DM

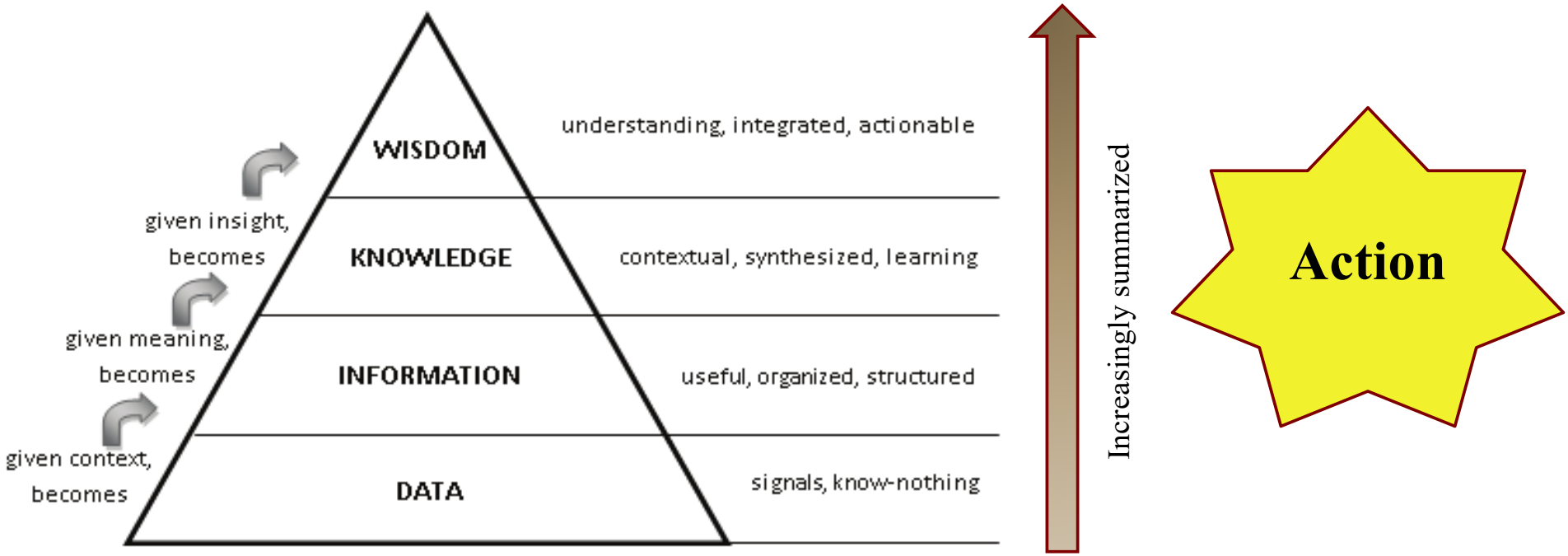
Enterprise Tools



With increasing awareness of the value of data and analytics; Many Enterprise Analytics Solutions have been appearing in the market.

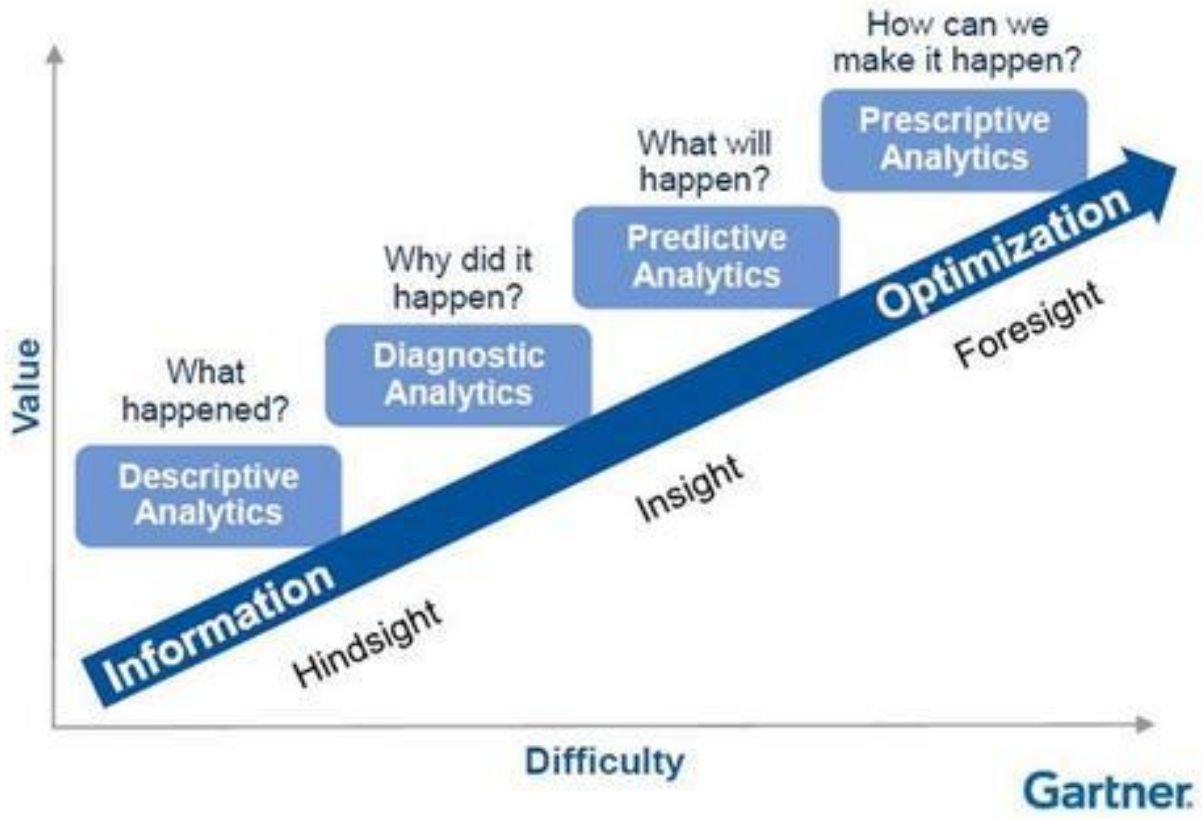
However, no solution can be fool-proof. Continual education and knowledge sharing amongst peers is essential to improving one's ability to extract insight.

From Data to Insights



DIKW Pyramid

Outcomes from Analytics



Modeling

Statistics and Correlation

Business Intelligence

DATA ANALYTICS WITH PYTHON

What is Python?

What is Python?

Official Definition:

Python is an **interpreted**, **object-oriented, high-level programming language** with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for **Rapid Application Development**, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are **available in source or binary form without charge for all major platforms**, and can be freely distributed

Why/Why Not Python?

Why:

- Community Support is very good
- Documentation of Libraries are mostly good
- Many libraries (especially the data science ones) are computationally efficient!
- Syntax is easier to read and use than most other languages
- It can do many things!

Why not:

- Some people think its still a simple scripting language and not an Enterprise ready language
- Library not built yet
- It can do many things!

Data Science Libraries

Library	Why	Remarks
IPython Notebook	Provide a data science notebook to run scripts and and in line documentation.	Essential.
NumPy	Computationally efficient numerical calculation/storage package. Included here just for kudos to the developer.	Learn-as-you-go
SciPy	Computationally efficient matrix calculation/storage package. Included here just for kudos to the developer.	Learn-as-you-go
Pandas	Data Frames in R was really cool. Python now has it too!	Essential. Very good project to follow, as each new version includes commonly used functions by scientists!
Scikit-learn	Machine Learning Library	Well documented with examples. Default parameters are already the at recommended settings.
Sklearn_pandas	Bridge between pandas and sklearn	Allows for some convenient mapping of table columns to feature matrices used by the sklearn package.
statsmodels	Another Machine Learning Library	Not as commonly used as sklearn but some examples online still uses this.
nltk	Natural Language Toolkit for Computational Linguistics	For text analysis that requires parsing the structure of the sentences

The Pandas Library

The pandas library is a Data Frame library (inspired from R) with a collection of data manipulation packages.

It is deemed essential in any Data Scientist's toolbox due to the ease of running descriptive analyses (like R Data Frames) and data manipulation.

The manipulation operations are computationally efficient as the "batteries" behind it are NumPy and SciPy.

The SKLEARN Library (SciKit-Learn)

The Scikit-learn library contains a wide collection of machine learning algorithms with extensive documentation on each family of algorithms and examples of the usage of each algorithm.

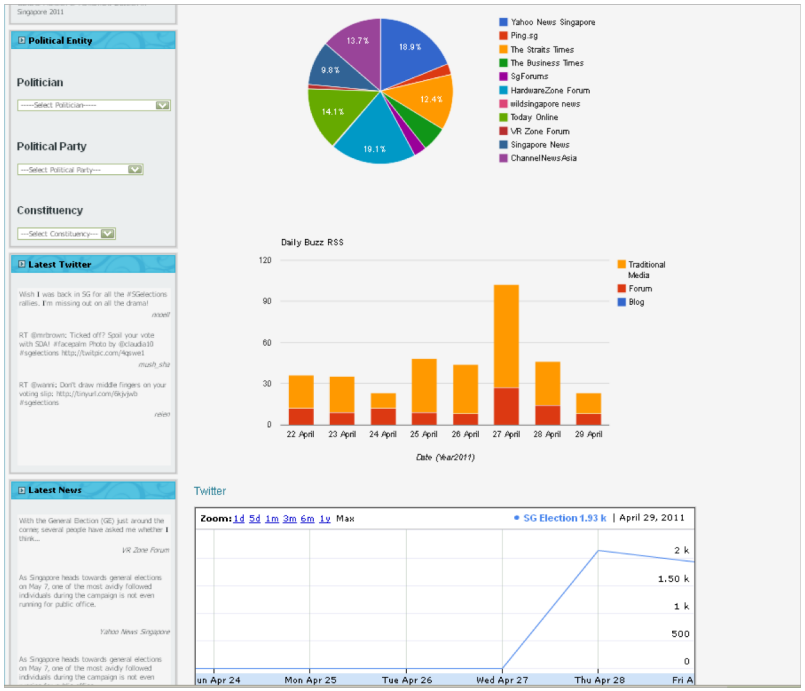
See the user guide http://scikit-learn.org/stable/user_guide.html

- Please check that you have a working copy of Python on your machine
- https://www.anaconda.com/download for those that have not yet installed Python on their system.
- There are slight differences between Python 2 and Python 3
- For our exercises; we will be using Python 3.
  - Those using Python 2 may want to set up an environment with Python 3 using https://conda.io/docs/user-guide/tasks/manage-python.html
  - Otherwise, there will be a few lines of code that need to be changed e.g. print statements

DATA ANALYTICS IN PYTHON

Data Analytics in Industry

Social Media Analytics



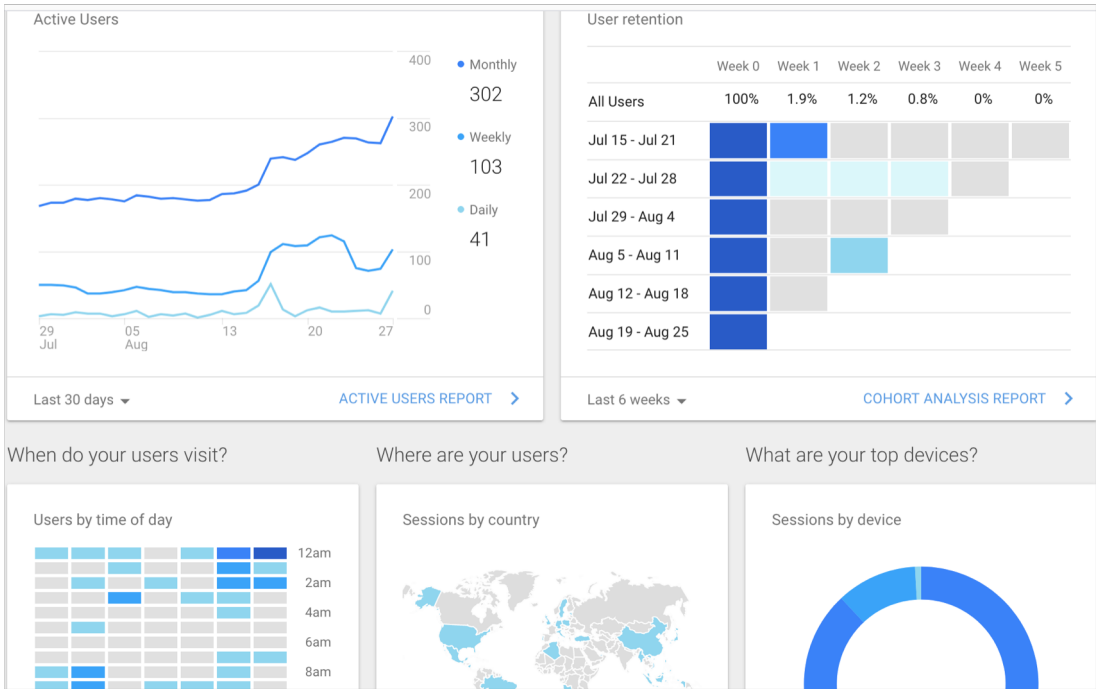
Social Media Sensing

Social Media Analytics is about tapping into different social media channels, and summarizing them into a coherent picture for the user.

Involves:

- Campaign Tracking
- Brand Tracking
- Sentiment Analysis

Web Analytics



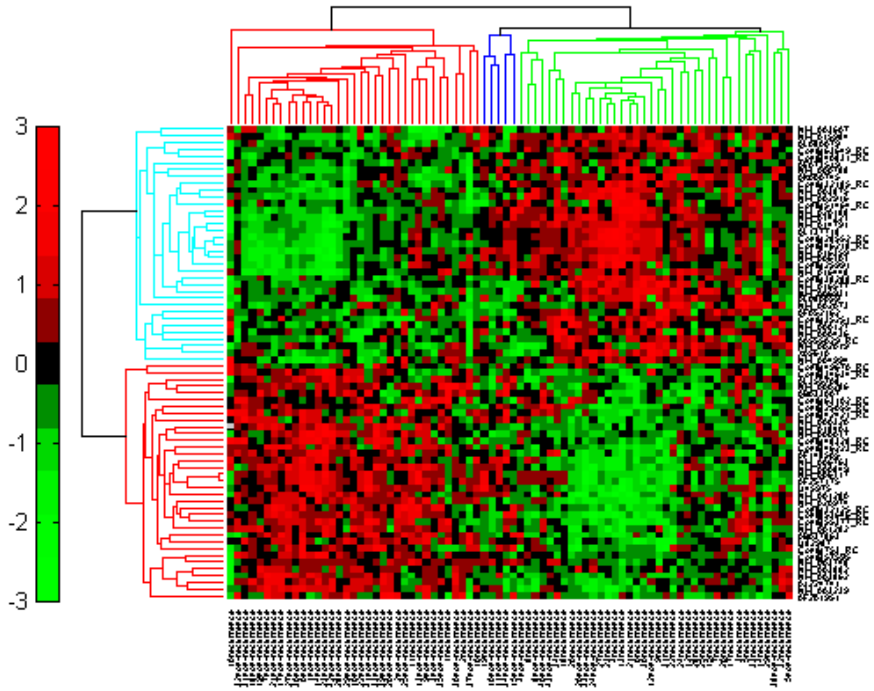
Google Analytics (Web Tracking)

Web Analytics is the collection and summarization of web data to your website. The insights are mainly used to track traffic and can be used in conjunction with search engine marketing (SEM)

Involves:

- Site Traffic Monitoring
- User Profiling
- Click Analytics

Medical Analytics



Microarray Heatmap

Medical Analytics is the application of computer science and statistical analyses to biomedical data. Medical data spurred the rise of Big Data strategies through optimization of algorithms.

Involves:

- BioInformatics
- Clinical Informatics
- Data Fusion

Text Analytics

I was really happy this morning walking into your store until when I needed to ask a question and the staff were incredibly rude. Normally you have super fast service. I wanted an item I saw online and tried to call the store but the phone was always engaged. Over experience was not very nice, the staff didn't care. Usually they are really helpful, I'm not sure what happened today.

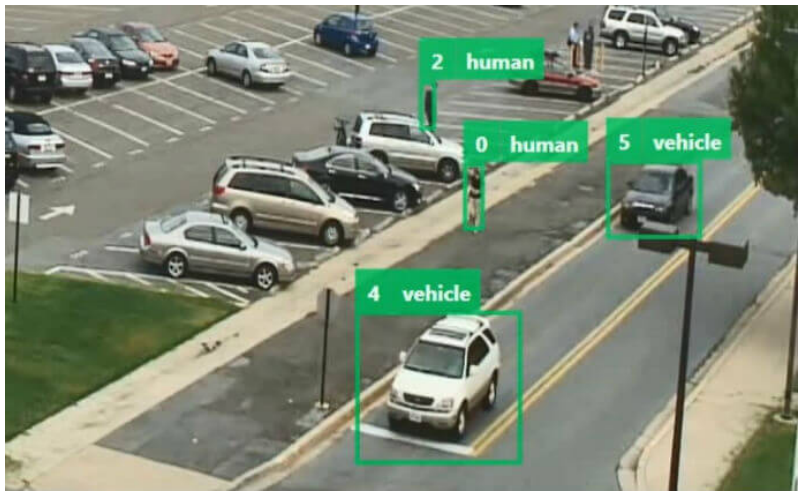
Text Extraction

Text Analytics can be domain by itself due to the complexity involved in working with text data. The majority of effort is in the information extraction from documents.

Involves:

- Document Clustering
- Sentiment Analyses
- Natural Language Processing (Linguistics)
- Named Entity Recognition

Video Analytics

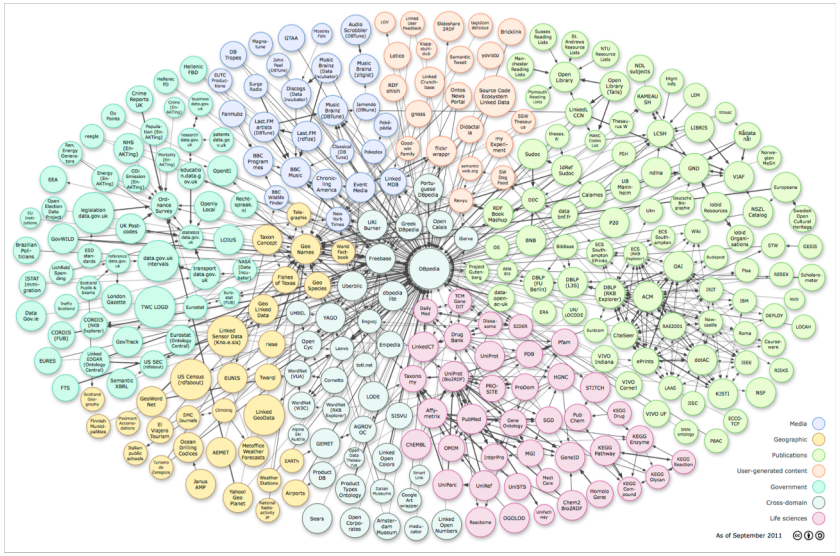


Video and Image Analytics mostly revolve around the extraction of features from the image or video data. It can be a highly complex area of analytics especially if it has to be in real-time (see autonomous vehicles)

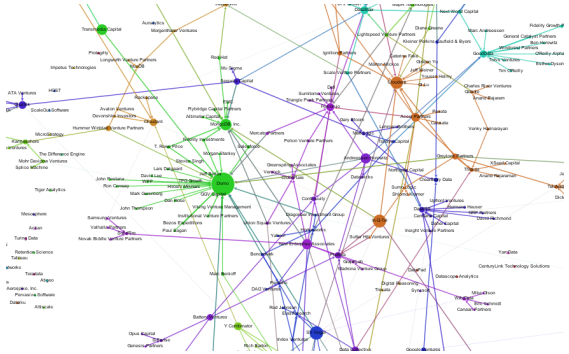
Involved in:

- OCR (Optical Character Recognition)
- Traffic Analysis
- Security Monitoring

Graph Analytics



Linked online data



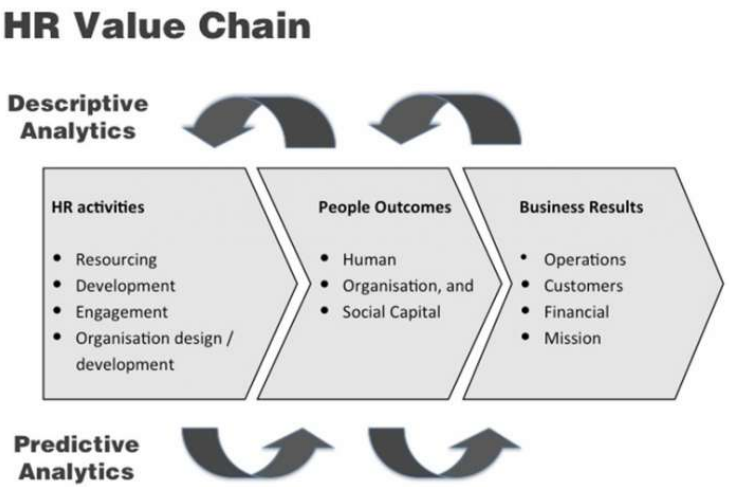
Network Analytics

The study of interactions between entities is known as graph analytics or network analytics.
From the interactions between each entity, we are able to discover insights via cliques and gateway nodes.

Involved in:

- Fraud Detection
- Clustering
- Entities-of-interest
- Semantic Technologies

Human Resource Analytics

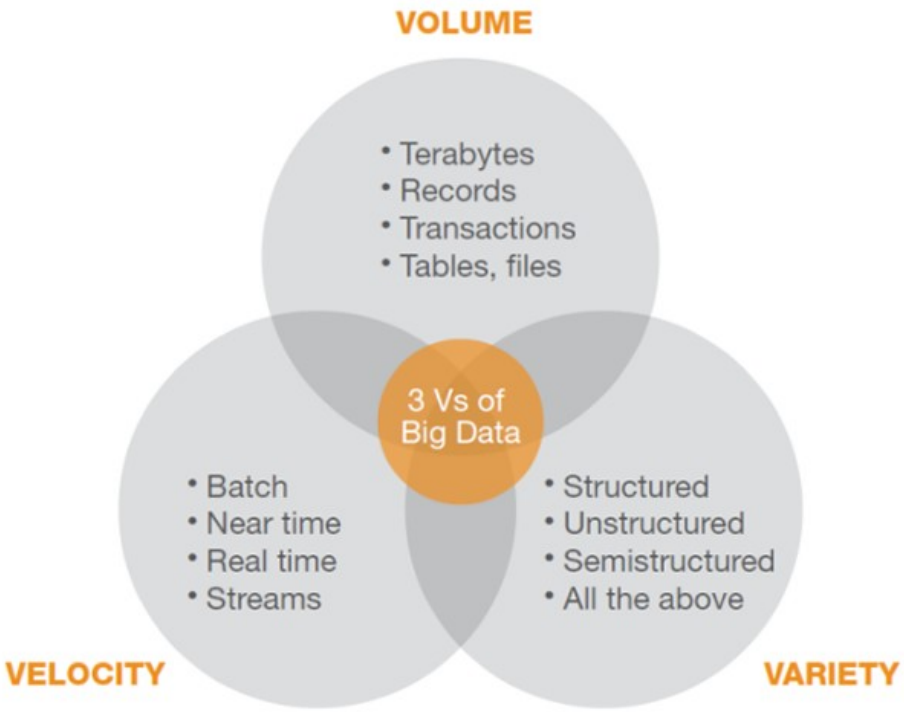


Human Resource creates a variety of data during the operation of the business. These datum can be in the form of exit surveys, employee feedback, claims, salaries, etc.

Involved in:

- Workforce/Talent Profiling
- Turnover Prediction
- Performance Tracking
- Skills Audit

Big Data



Big Data is not just about Volume. It can be any or a combination of any of the 3 Vs. Most Big Data are transactional data.

- Volume -> Size of data, quantity of data files
- Velocity -> Amount of data to process per time unit
- Variety -> How do I integrate many different sources of data

DATA ANALYTICS WITH PYTHON

About Data

What is Data

Structured vs **Unstructured** Data:

Structured data is typically well organized e.g. data tables.

Unstructured data typically refers to text data where there is no well defined form e.g. images, videos, text documents

Raw vs **Processed** vs **Summarized** Data:

Raw data or processed data is preferable over summarized data, although some information may be lost after processing data.

Different Types of Data

Identifiers:

Data fields that are used to **uniquely identify entities** of interest e.g. IC number, company registration number, purchase order number

Categorical:

Data fields that allow the data set to be grouped into **categories**. E.g. gender, occupation, department

Measures:

Data fields that hold numbers that quantify the **value** of an entity. E.g. Age, price, quantity

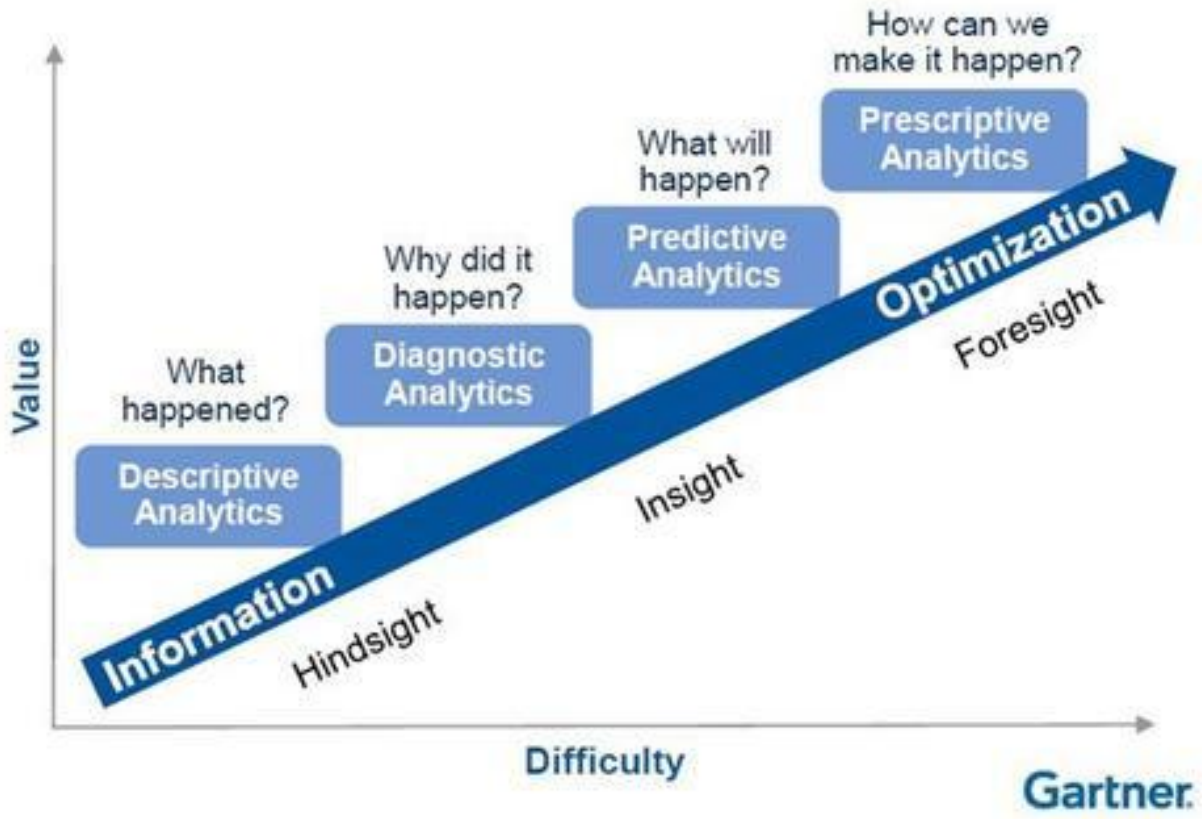
Tutorial

- IPython Refresher
- Identify different types of data

DATA ANALYTICS WITH PYTHON

Descriptive, Diagnostic, Predictive

Outcomes from Analytics



Prediction Models

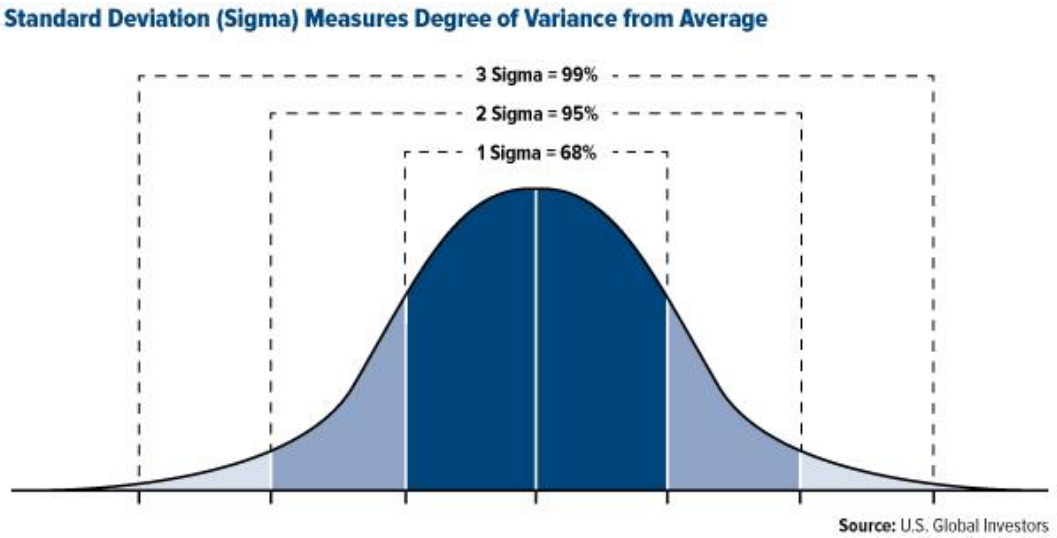
Statistics and Correlation

Business Intelligence

Descriptive Analytics

- **Descriptive Analytics** is the examination of data or content, usually manually performed, to answer the question “What happened?” (or What is happening?), characterized by traditional business intelligence (BI) and visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives.

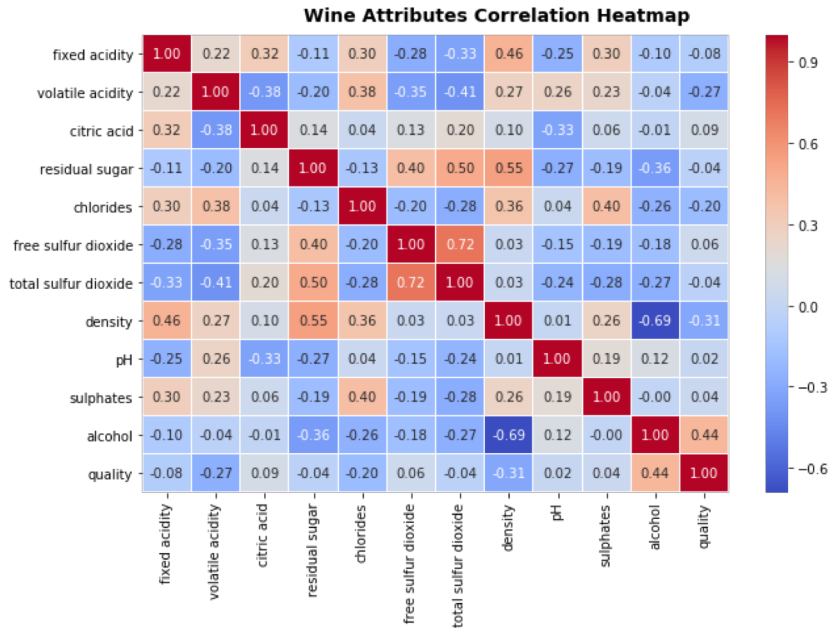
- What we do
  - Data Distribution
  - Data Visualization



Diagnostics Analytics

- **Diagnostic Analytics** examines data or content to answer the question “Why did it happen?”, and is characterized by techniques such as drill-down, data discovery, data mining and correlations.

- What we do
  - Correlation
  - Covariance
  - Cointegration



Predictive Analytics

- **Predictive Analytics** describe the use of statistics and modeling to determine future performance based on current and historical data. Predictive analytics look at patterns in data to determine if those patterns are likely to emerge again, which allows businesses and investors to adjust where they use their resources in order to take advantage of possible future events.

- What we do
  - Linear Regression
  - Classification

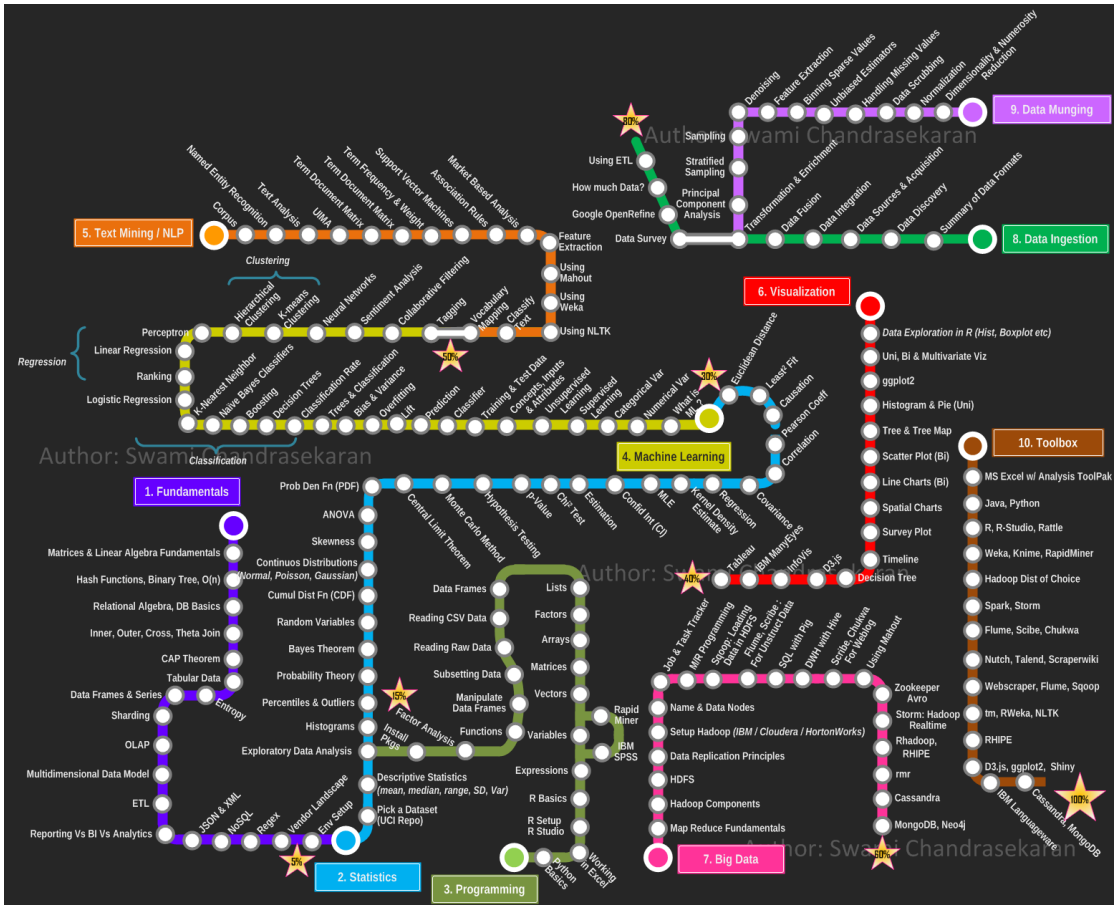
Hands-on

- Examine Data Distribution of dataset
- Anomaly detection
- Charting with pandas and matplotlib
- Correlation Study
- Modeling

DATA ANALYTICS WITH PYTHON

APPENDIX

Analytics Roadmap



http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist/

Community

Extra Resources:

http://pugs.org.sg/pages/learning.html

Local Community:

Python User Group (Singapore) http://pugs.org.sg

PyData-SG https://www.facebook.com/groups/pydatasg/

PyLadies-SG https://www.facebook.com/PyLadiesSG/

Events:

PyCon Singapore – usually held in June/July

Data Analytics with Python

Trainer: Michael Li

Venue: NTUC Learning Hub

Date Conducted: 23 Oct 2018