INTRODUCTION TO DATA ANALYSIS

Partha Padmanabhan

UCSC Silicon Valley Extension
YOUR UC IN SILICON VALLEY

Class Structure

Class 1
- About the Course
- Introduction to Data Analysis

Class 2 -9
- R Studio
- Data Analysis with Programming in R
- Homework (Class 2,4,6 & 8)

Class 10
- Final Project Demo

Grading
- Homework – 30%
- Class Participation – 20%
- Project Completion – 50%

Today's Topics

- About:
  - *About Me*
  - *About You*
  - *Contents of this course*
- Landscape:
  - *Data Analysis*
  - *Analysis Samples & Examples*
- Aspects:
  - *Framework*
  - *Best Practices*
  - *Real Life Use Cases*

About Me

- My name is Partha Padmanabhan
- Working @ Cisco for the last 10+ years as Data Architect
- Teaching Data Modeling, Data Visualization & Data Analysis at UCSC Extn
- Passion: Data, Data Modeling, Data Visualization, Data Analysis & Architecture

Approach

- Learn & Share
- Interactive Class Approach
- Will accept my mistakes
- Will try to avoid draining the slides
- Ask me questions

Final Project

- Aim: turning data analysis techniques you learn in class to become your strength in dealing real world problems
- The project involves analysis of the data, implementation, preparation of a report and presentation of the results during the last week of the class. The project will be done in groups of 2~3 students. If you already working on a research project in the area of interest you are encouraged to use dataset/topic from your research provided you make some extra effort for the class.
- Detailed instructions for the project will be posted later
- Due to the current scenario of online classes, if you are not able to perform your project in groups, you can complete it individually and present it.

Enough about me.. Your turn

- Tell me about yourself
- Your familiarity to any Analytical tools
- Your expectation and outcome from this course

Goals & Objectives

•Data science is huge field, and there's no way you can master it by reading a single book or taking a single course. The goal of this course is to give a solid foundation in the most important tools.

• This course is concerned with the nuts and bolts of manipulating, cleaning, and crunching data in R. It is also a practical, modern introduction to scientific computing in R, tailored for data-intensive applications.

• This course is also about the parts of R language and libraries you'll need to effectively solve a broad set of data analysis problems.

•The overall goal of this class is to introduce you to the discipline of data analysis/mining, a science of understanding and analyzing data. This class is designed to provide you with the tools you need for solving real world problems using statistics and a better understanding of data analysis techniques.
• We plan to achieve these goals by introducing you to the relevant statistical knowledge, how to use statistical software R to perform data analysis and engage in solving problems, analysis through homework, discussion, class participation and project.

Why is big data analytics important?
Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. In his report *Big Data in Big Companies*, IIA Director of Research Tom Davenport interviewed more than 50 businesses to understand how they used big data. He found they got value in the following ways:

1.Cost reduction. Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.

2.Faster, better decision making. With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.

3.New products and services. With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

What will you learn?

First you must define your requirements or objectives. Once you have defined your requirements, next step is to think about the data. What is required and how to obtain it.

Next you need to import your data into R. This typically means that you take data stored in a file, database, or web API, and load it into a data frame in R.

Once you have imported your data, it is a good idea to tidy it. Tidying your data means storing it in a consistent from that matches the semantics of the dataset with the way it is stored. In brief, when your data is tidy, each column is a variable, and each row is an observation.

Next step is to transform the data. Transformation includes narrowing in on observations of interest (like all houses in one zip code, all data from the last year, creating new variables that are functions of existing variables (like computing velocity form speed and time know as derived variable, and calculating a set of summary statistics (like means or counts.

Tidying and transformation are called wrangling, because getting your data in a form that's natural to work with oden feels like a fight.

Once you have prepared your data, with variables you need, there are two main engines of knowledge generation***: visualization*** and ***modeling***.

Visualization is a very important. A good visualization will show you things that you did not expect, or raise new questions about data.
A good visualization might also hint that you're asking the wrong question, or you need to collect different data.

Models are complementary tools to visualization. Once you have made your questions sufficiently precise, you can use to answer them. Models are a fundamentally mathematical or computational tool.

The last step of data science **communication**, an absolutely critical part of any data analysis project. It does not matter how well your models and visualization have led you to understand the data unless you can also communicate your results to others. The format will depend upon the audience.

Above all these tools is **programming**. Programing is a cross-cutting tool that you use in every part of the project. You don't need to be an expert programmer to be a data scientist, but learning more about programming pays off because becoming a better programmer allows you to automate common tasks, and solve new problems with greater ease.

At the end of the course:
– Be able to understand the concepts, strategies, and methodologies related to the design and construction of data mining/data analysis

– You will be able to perform independent analysis of data

– Be able to comprehend several data preprocessing methods

– Be Trained how to present your project and results to a diverse audience

– You will have the tools to tackle a wide variety of data science challenges, using the best parts of R.

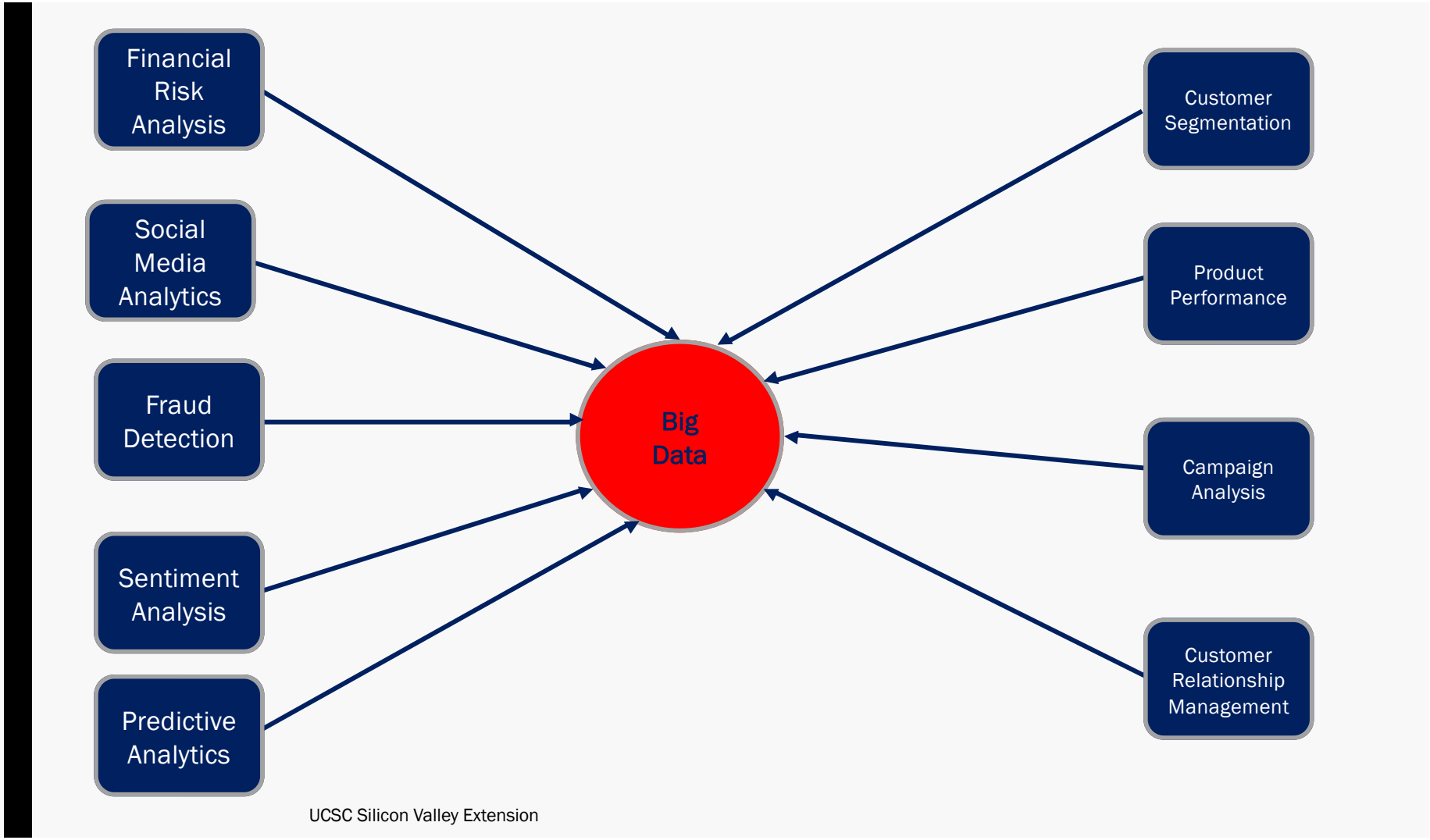
Who is this course for?

Anyone who is interested in:

- Helping companies make decisions aided by data
- Refreshing some theory learned in school, but with a practical focus
- Getting up to speed with new Open Source tools and libraries
- Curious about the new technology

Data analysis is all around us ...

Typical areas where Data Analysis can be performed –
Agriculture
– Computing
– Crime
– Ecology
– Health
– Hydrology
– Meteorology
– Social networking


Financial Risk Analysis
Social Media Analytics
Fraud Detection
Sentiment Analysis
Predictive Analytics
Big Data
Customer Segmentation
Product Performance
Campaign Analysis
Customer Relationship Management

Data Mining

- Misnomer??
- Gold mining vs. Sand (Rock) Mining
- Knowledge Discovery from Data (KDD)
- Knowledge extraction
- Data/pattern analysis
- Data archaeology
- Data dredging

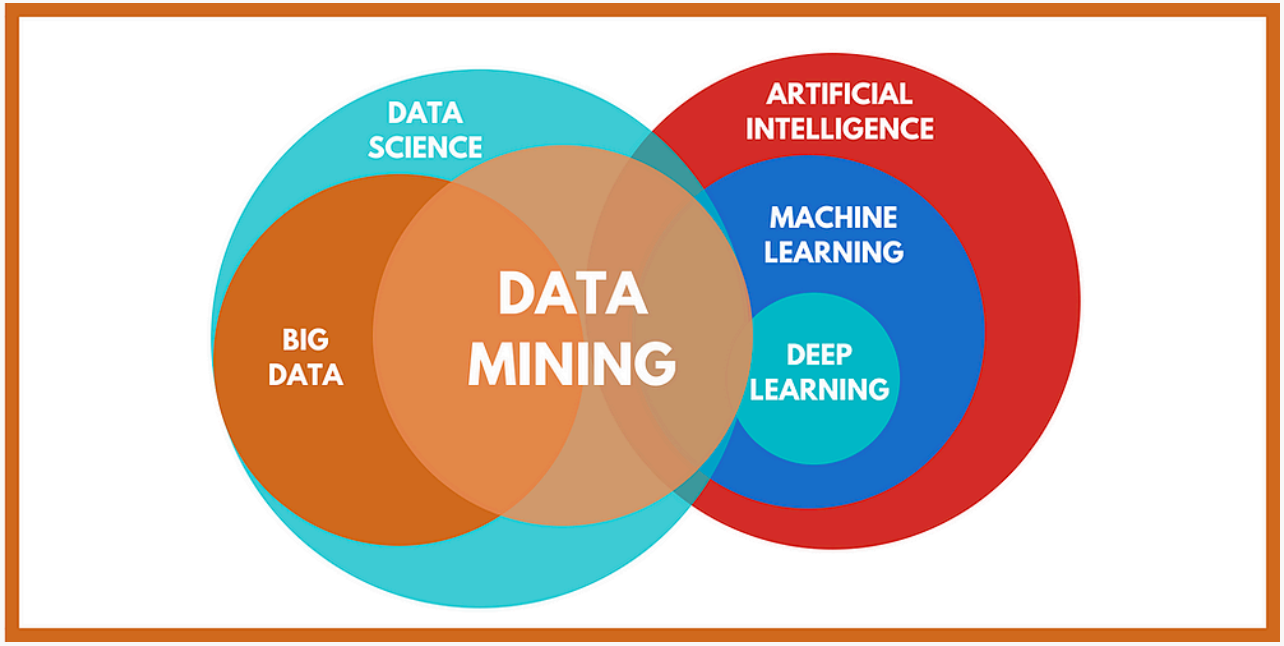
Data Mining

The main objective of data mining is knowledge discovery. What is data mining?

- Automatically examine through data to find hidden patterns, discover new insights and make predictions

- The process of extracting patterns from data. Humans have been manually extracting information from raw data for years.
http://en.wikipedia.org/wiki/Data_mining
- Exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover useful patterns.
"*Introduction to Data Mining", by Tan, Steinbach, and Kumar, 2005*.

Data mining is a confluence of many disciplines and combines the ideas of many fields such as Statistics, Artificial Intelligence, Machine learning, databases etc.



About Data Science:

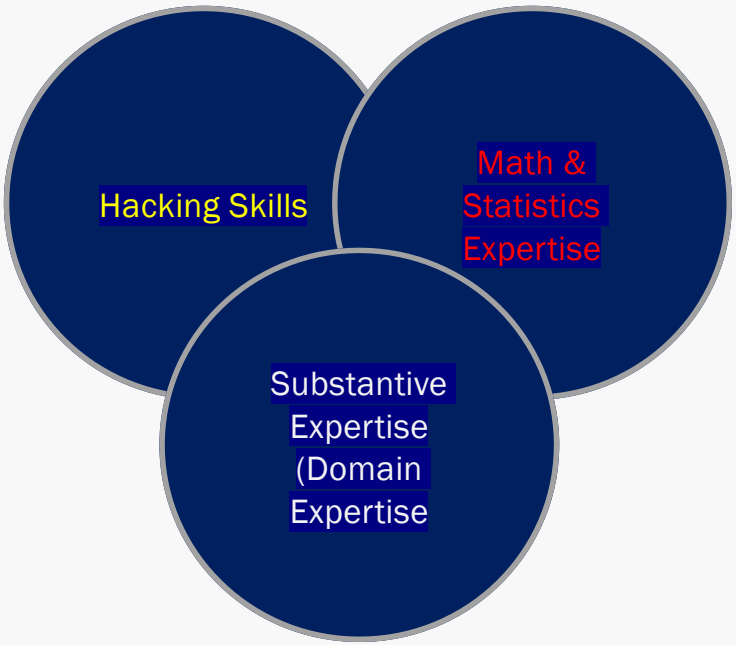
Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.
Data science is related to data mining and big data.

Data Science provides:
- Hidden insights
- Competitive Advantage

Data Scientist:
- Sexiest Job of the 21st Century
- D J Patil - Served as the Chief Data Scientist in the United States Office of Science and Technology Policy from 2015 to 2017

Rare Qualities of a Data Scientist:
- Find the Order, meaning & value in unstructured Data
- Predict Outcomes
- Automate Processes

Data Science in a Venn Diagram



= Data Science

Data Science Explores:

- Novel Sources
- Challenging Formats
- Streaming Data (Fast)
- Python & R for Data Manipulation & Modeling
- Tensorflow for Deep Learning

Forms of Mathematics used:
- Probability
- Linear Regression
- Calculus

Data Science - How & When to use:

- Choose Procedures
- Diagnose Problems
- Domain Expertise
    - What constitutes value
    - How to implement insights

It is fundamentally an interdisciplinary subject. Data science comprises three distinct and overlapping areas:

The skills of a ***statistician*** who knows how to model and summarize datasets (which are growing even larger;

the skills of a ***computer scientist*** who can design and use algorithms to effectively store, process, and visualize this data;

and the ***domain expertise*** – what we might think of as classical training in a subject – necessary both to formulate the right questions and to put their answers in context.

Sampling, estimation, and hypothesis testing from statistics.

- Searching algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning.

- Database systems for storage, indexing, and query processing.

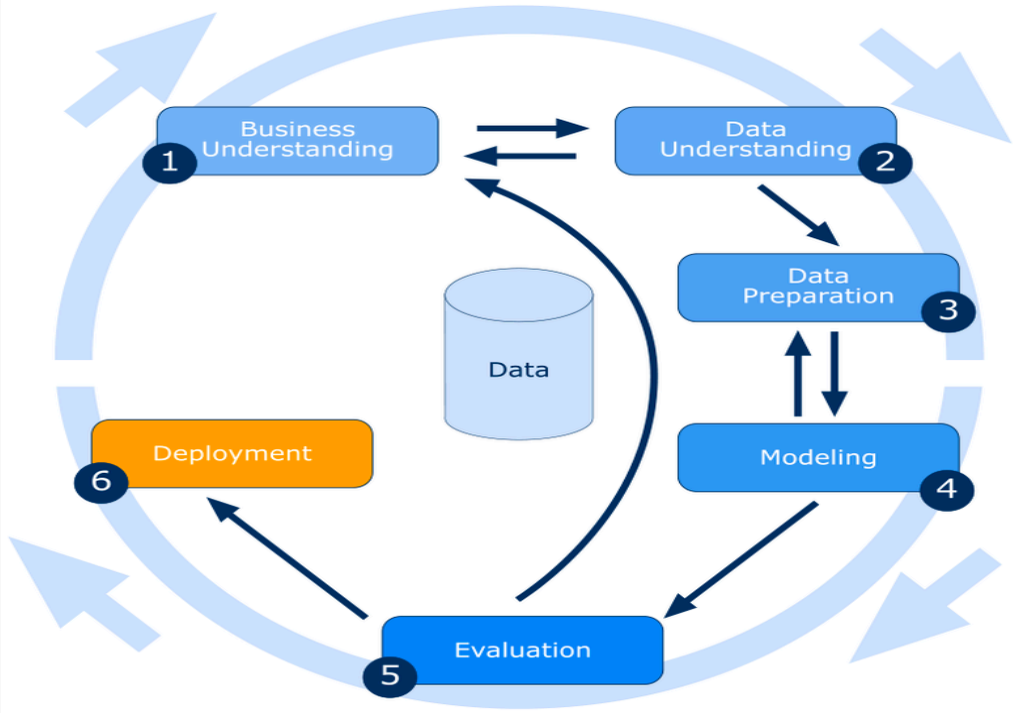
- High performance, parallel computing for massive datasets.
  Distributed computing to address size and location of data.

- Optimization, information theory, and visualization.

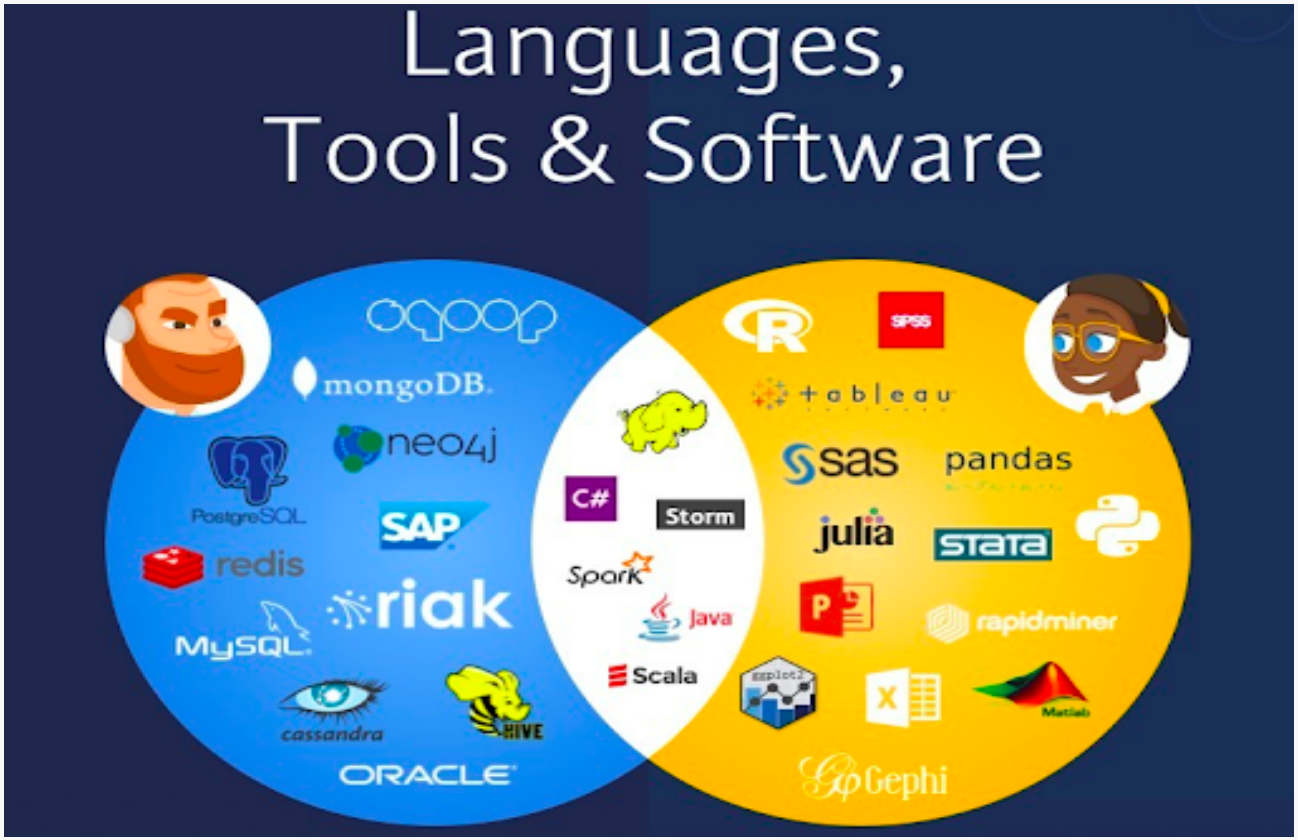
With this in mind, data science should not be considered as a new domain of knowledge to learn, but a new set of skills that you can apply within your current area of expertise.

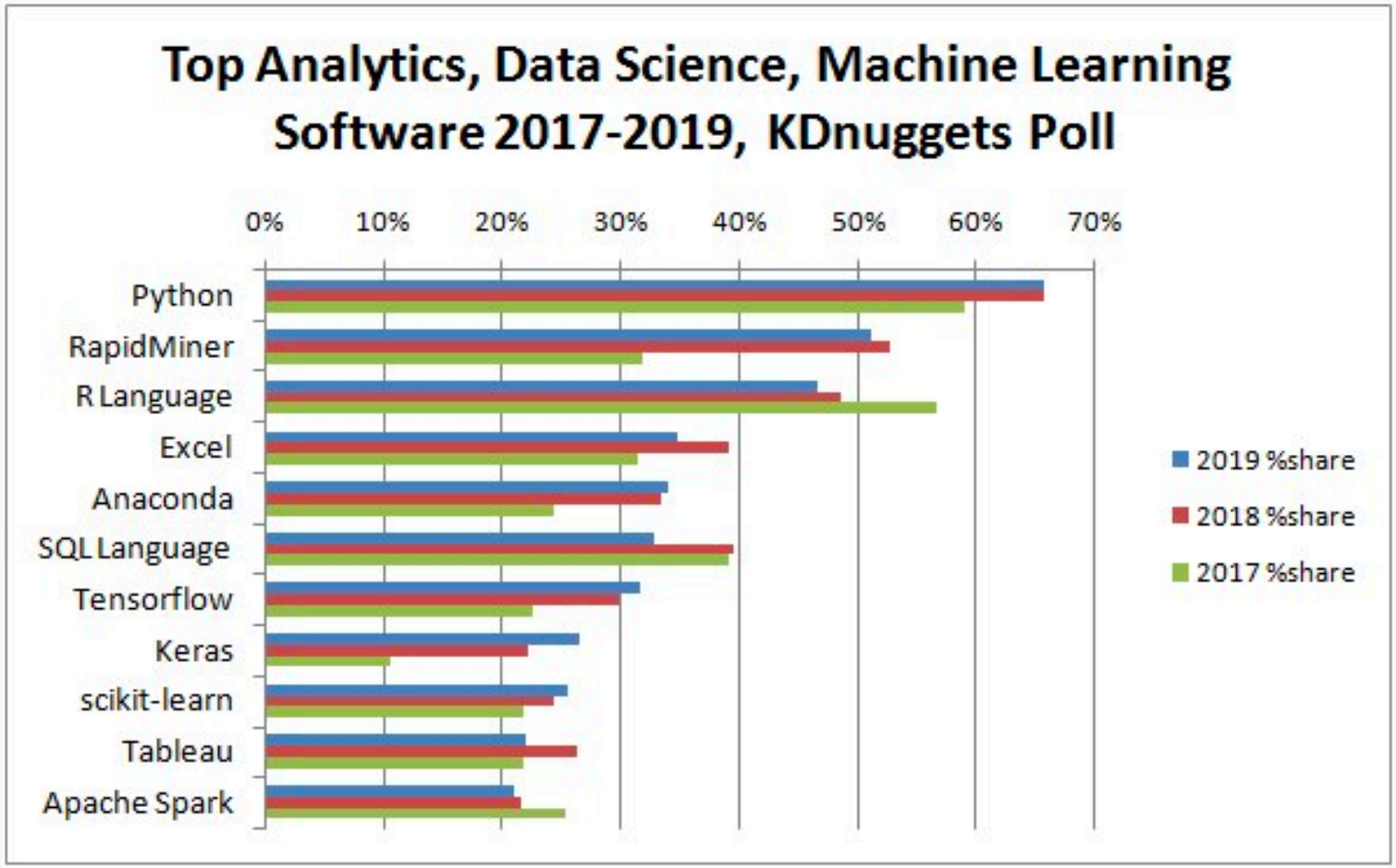
Whether you are reporting election results, forecasting stock returns, forecasting weather, or working with data in any other field, the goal of this course is to give you the ability to ask and answer new questions about your chosen subject area.

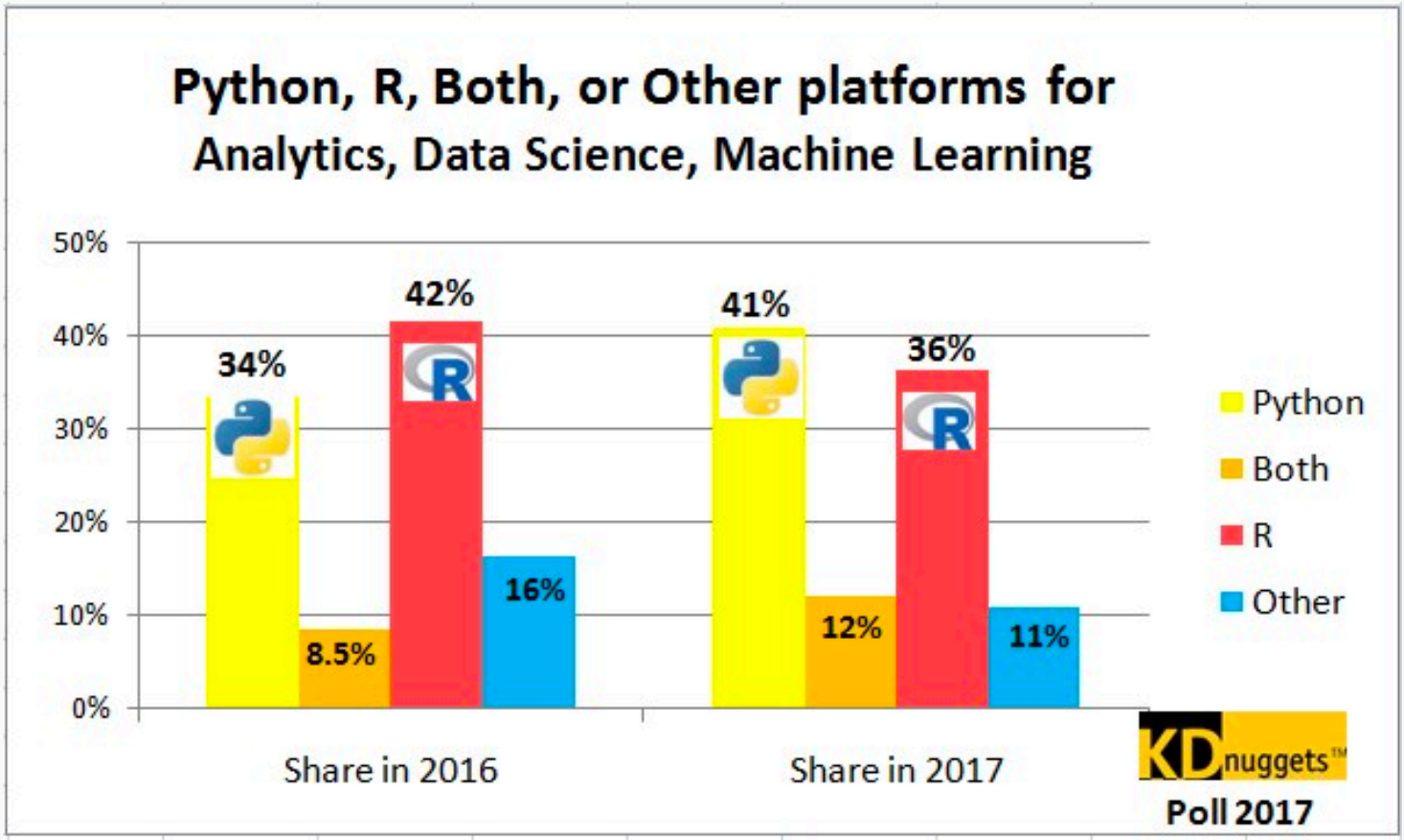
CRISP-DM (Cross Industry Standard Process for Data Mining provides a process model for data mining that consists of six major phases



Data Scientist's Toolbox




Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll
0%
10%
20%
30%
40%
50%
60%
70%
Python
RapidMiner
R Language
Excel
Anaconda
SQL Language
Tensorflow
Keras
scikit-learn
Tableau
Apache Spark
2019 %share
2018 %share
2017 %share


Python, R, Both, or Other platforms for
Analytics, Data Science, Machine Learning
50%
40%
30%
20%
10%
0%
34%
8.5%
42%
16%
41%
12%
36%
11%
Python
Both
R
Other
Share in 2016
Share in 2017
KDnuggets™
Poll 2017

Focus on the results, not the tools

- It is okay to use multiple tools
- Whatever you know to get the job done
- Specialized Tools: The best tool for the right job

What is R

R is a computer programming language which is particularly well suited to handling and sorting the large datasets associated with Big Data projects

- R is an implementation of the object-oriented mathematical programming language S. It is developed by statistician around the world and is free software.

- Syntactically and functionally it is very similar (if not identical) to S+, the popular statistics package.

What is ‘R’

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

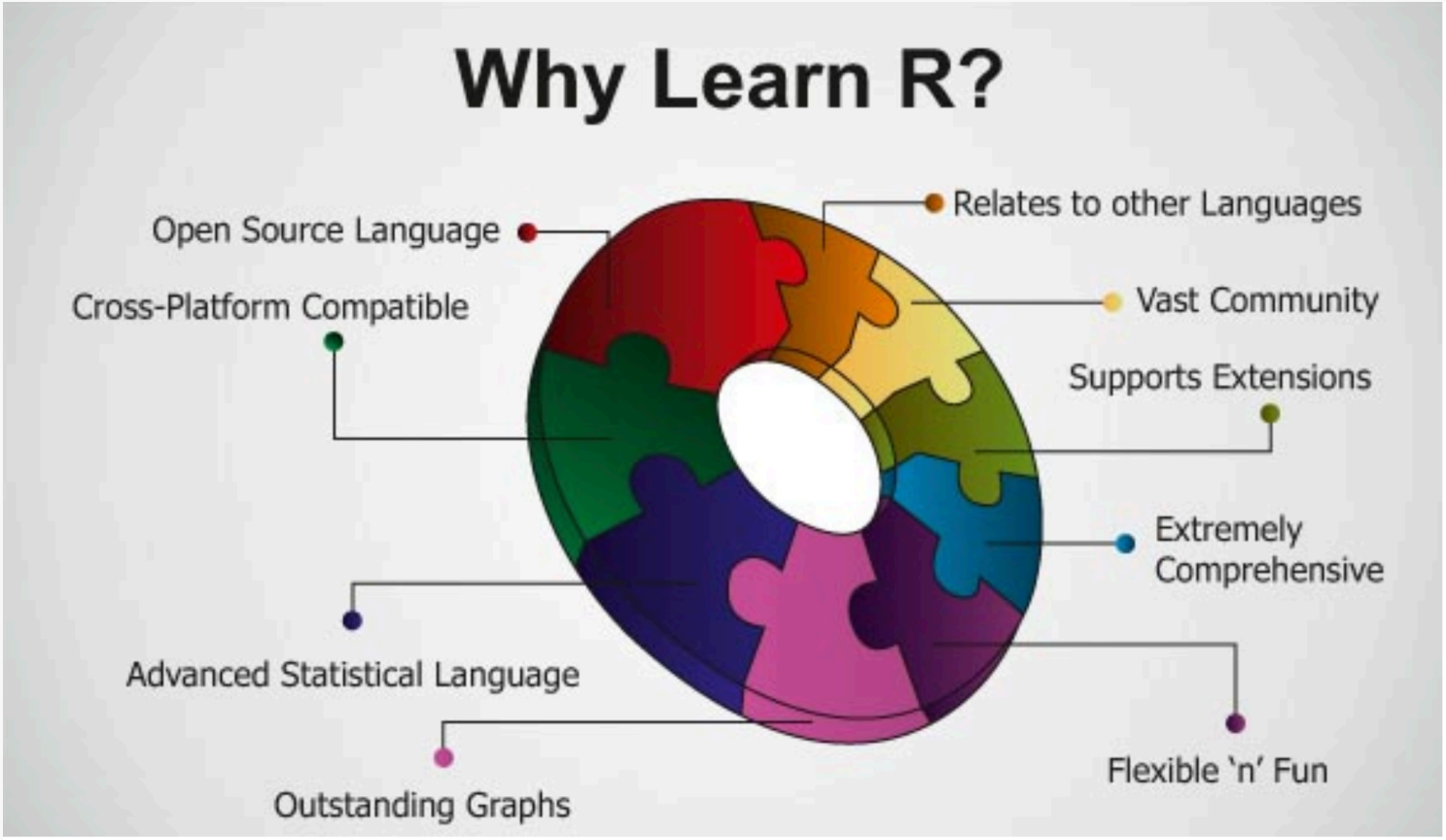
Why to use 

The style of coding is easy

- The R is free and open source. No need to pay any subscription charges.
- R is effectively platform independent
- It is a full programing language
- R is on the cutting edge, and expanding rapidly, 10000+ packages are available

R has unrivaled help resources. The community support is overwhelming. There are numerous forums to help you out.

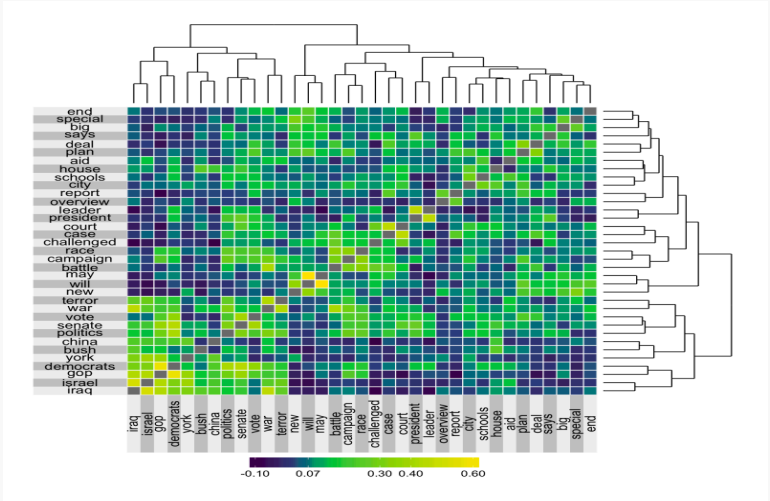
- R makes the best graphics
- One of the highly sought skill by analytics and data science companies
- comes standard with some of the most flexible and powerful graphics routines available anywhere

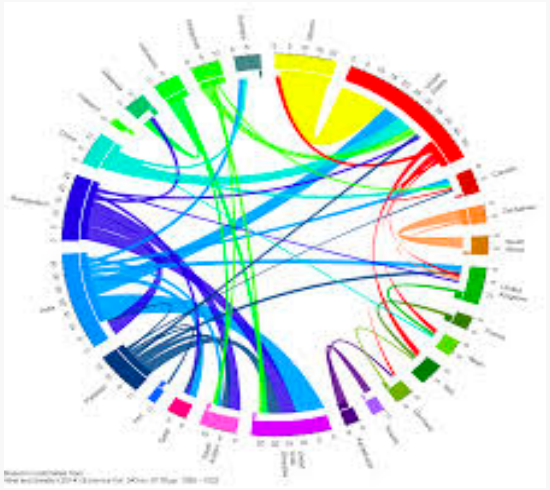

Why Learn R?
Open Source Language
Cross-Platform Compatible
Advanced Statistical Language
Outstanding Graphs
Relates to other Languages
Vast Community
Supports Extensions
Extremely Comprehensive
Flexible 'n' Fun

vs

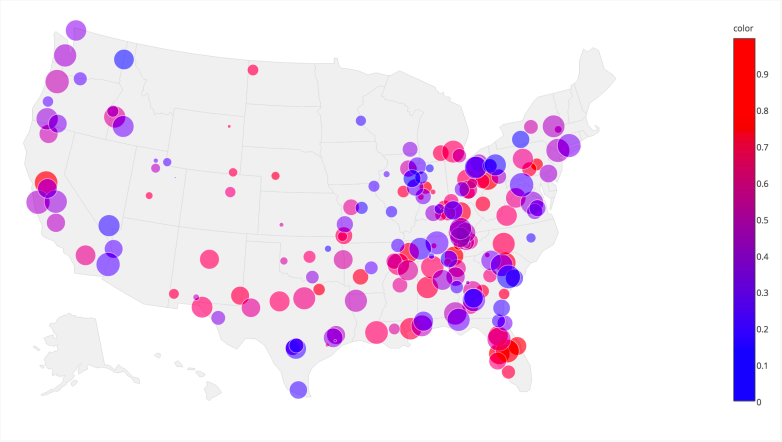
Both are good options for Analysis

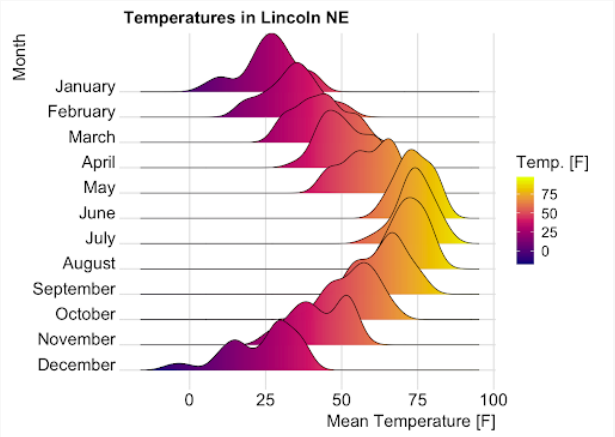
- R or Python + NumPy + SciPy + SciKit +Pandas
- R is stronger for stats. Python for General Programming
- ggplot2 > Matplotlib
- Easier to get started with R
- Python is easier to get “ into production”





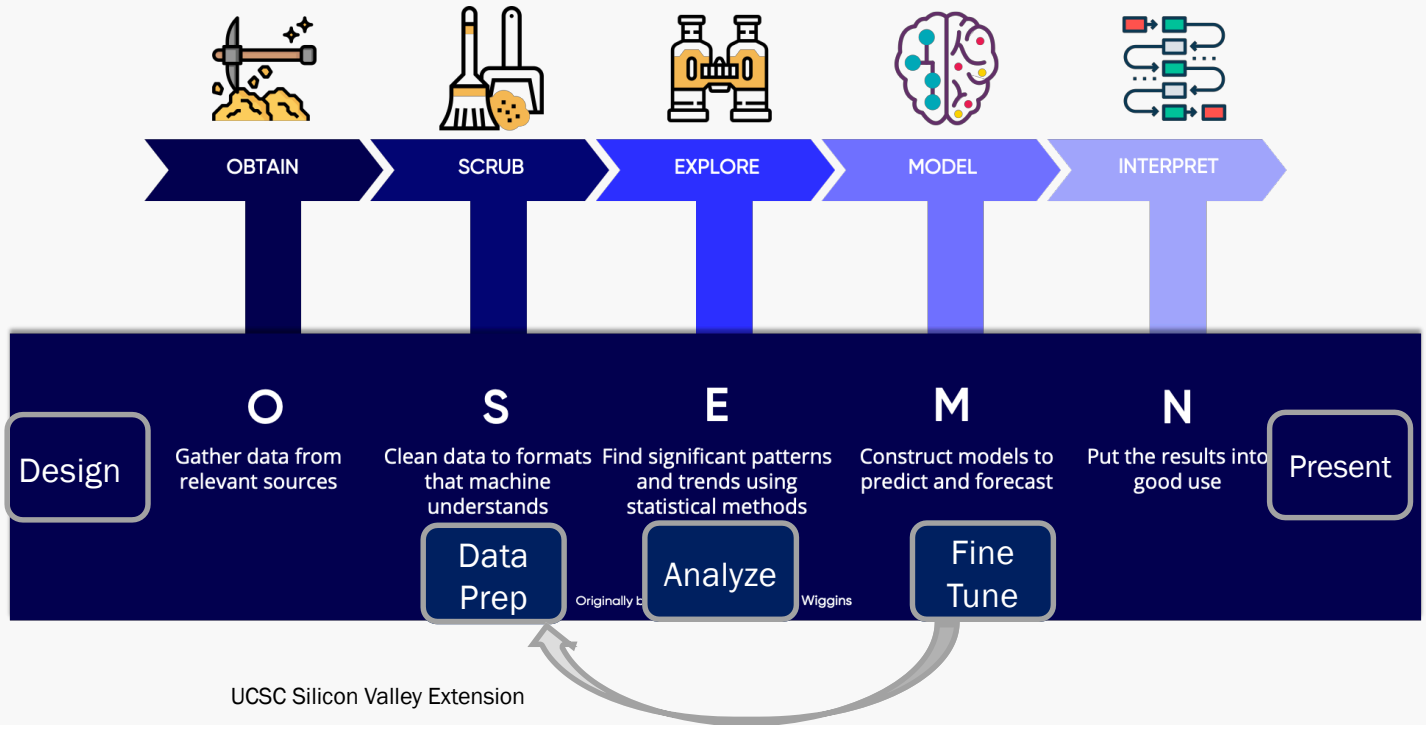
Election Analysis



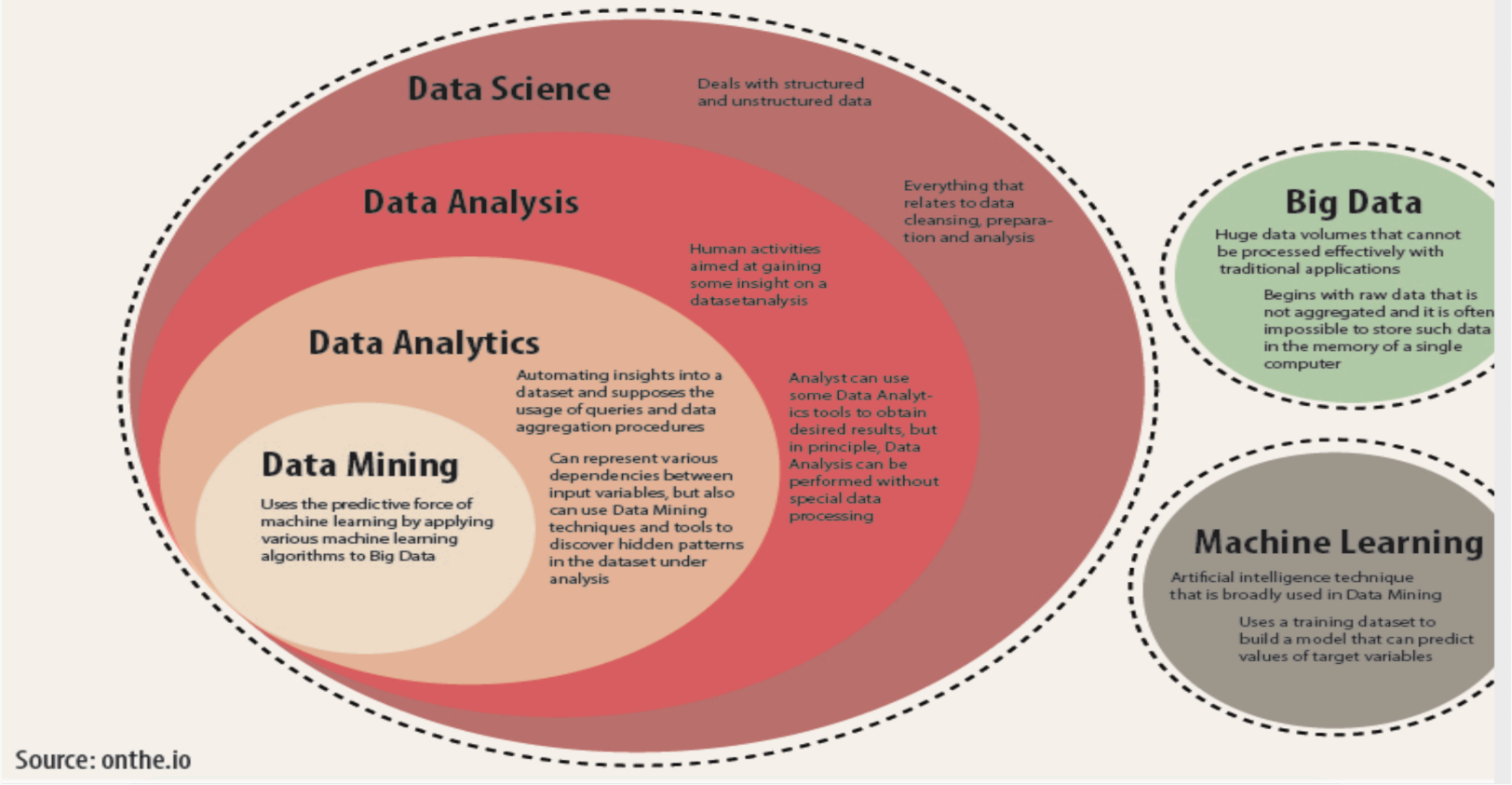


Data Science Process

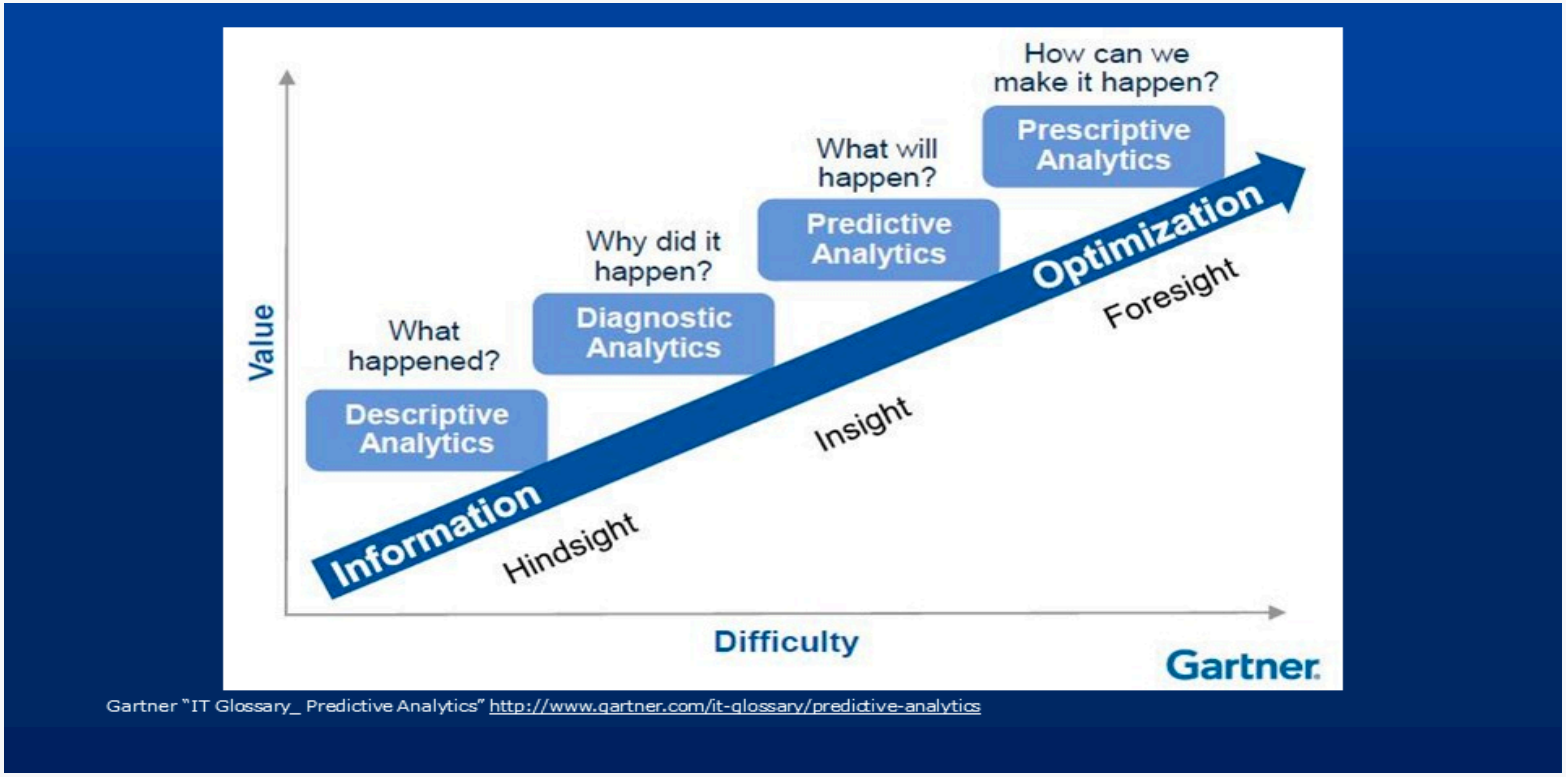
LEAD



Difference between Data Science, Data Analysis, Big Data, Data Analytics, Data Mining, Machine Learning

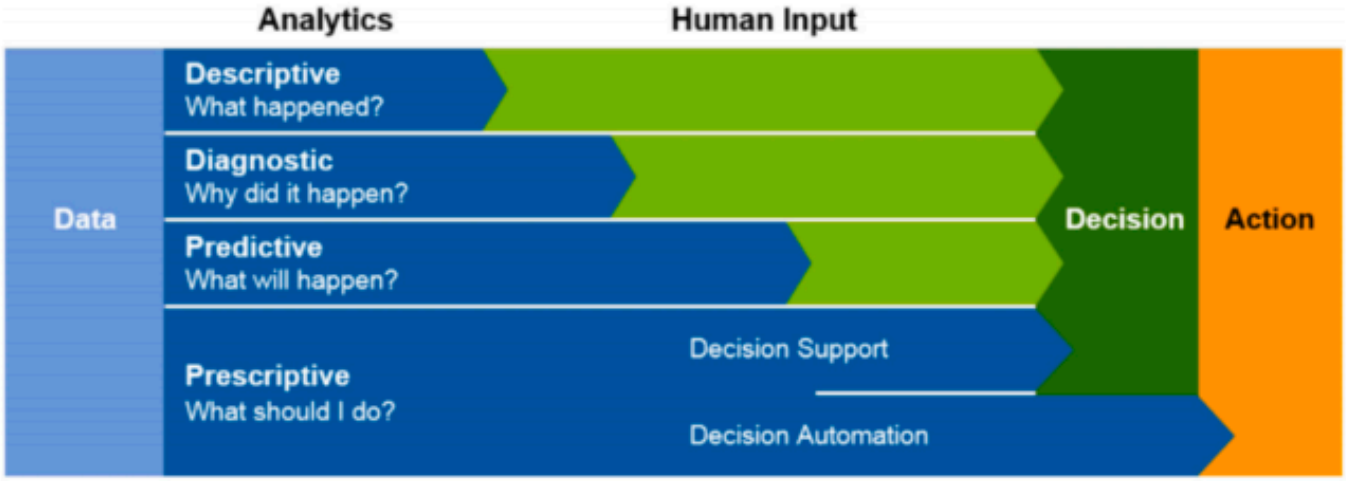


Analytics Capabilities Framework



Gartner "IT Glossary_ Predictive Analytics" http://www.gartner.com/it-glossary/predictive-analytics

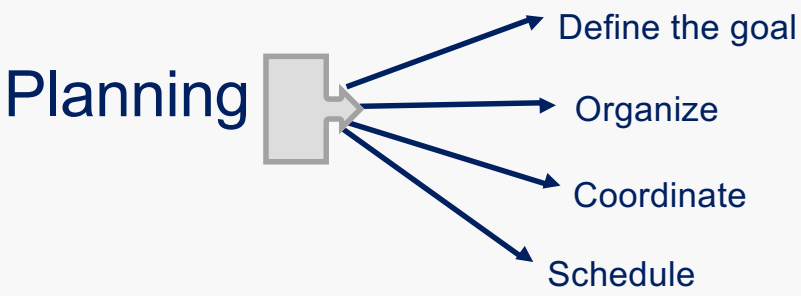
Figure 1. Analytics Capabilities Framework



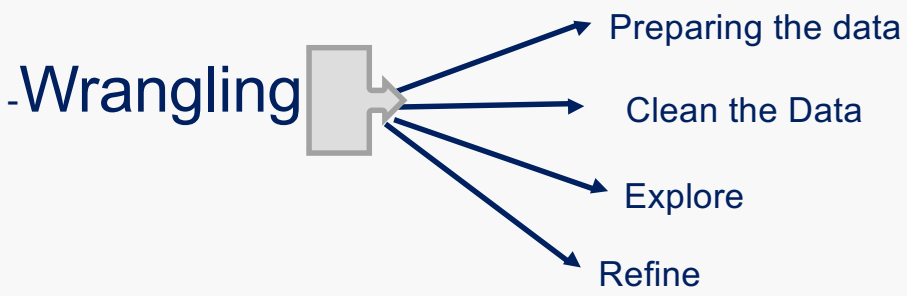
Source: Gartner (May 2015)

Gartner, How to Get Started With Prescriptive Analytics, Lisa Kart, May 5, 2015

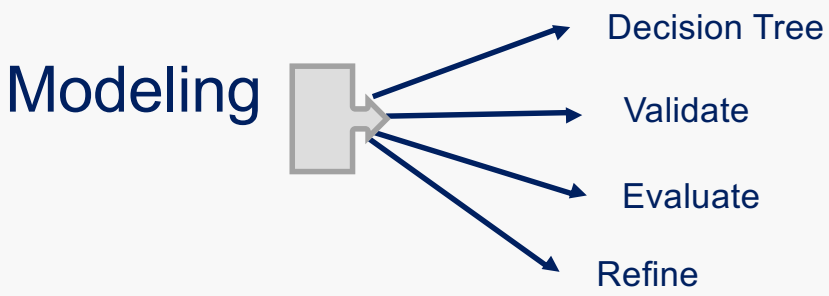
Data Science Pathway:



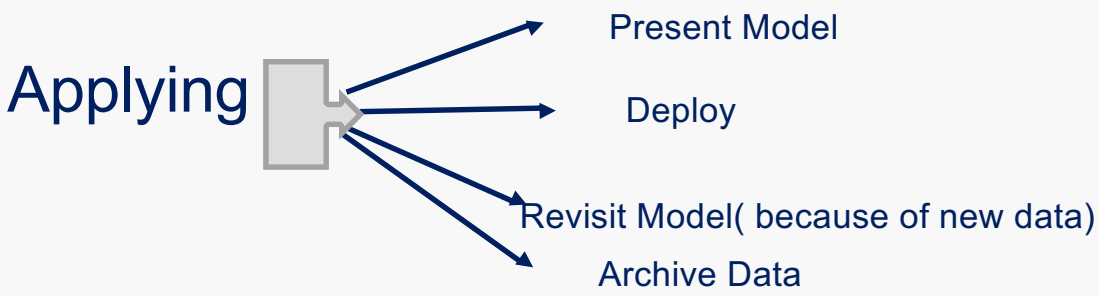
Data Science Pathway:



Data Science Pathway:



Data Science Pathway:



Data Science Roles & Terms:

- Data engineer
- ML Specialist
  - Computer Science
  - Mathematics
  - Deep Learning
  - Artificial Intelligence
- Researcher
  - Domain Specific Research
- Analyst
  - Decision Making

Data Science Roles & Terms:

- Manager
  - Data Science Projects
  - Create Solutions
  - Speak Data
- Entrepreneur
  - Data based Startups
- THE UNICORN
  - Rockstar, Ninja
  - Very Rare
  - Full Stack Data Scientist

Analytics Philosophy

- You have to get your hands dirty
- Keep trying out things
- Download data, or some code, and try to run
- Make small tweaks
- Analysis is both a science and art.
- Understand how the analysis has been put together
- There is no way to know everything. Learning is the answer
  - You learn by observing and practicing

Ask yourself this before you start the analysis

- What do I want to present?
- Which graphs will I create? and how many?
- What data will I need
- Where I can the data, i.e., source of the data
- Try a “mock plot” with dummy data
- Does it look like what I want

Getting comfortable with Big Data

- Recommend that you work with at least one data set that has >100K rows
- Over the course, you must download and use at least two OpenGov type datasets (To get into the habit: data.gov)
- For either your final project, or any open homework/assignment problem