INTRODUCTION TO DATA ANALYTICS

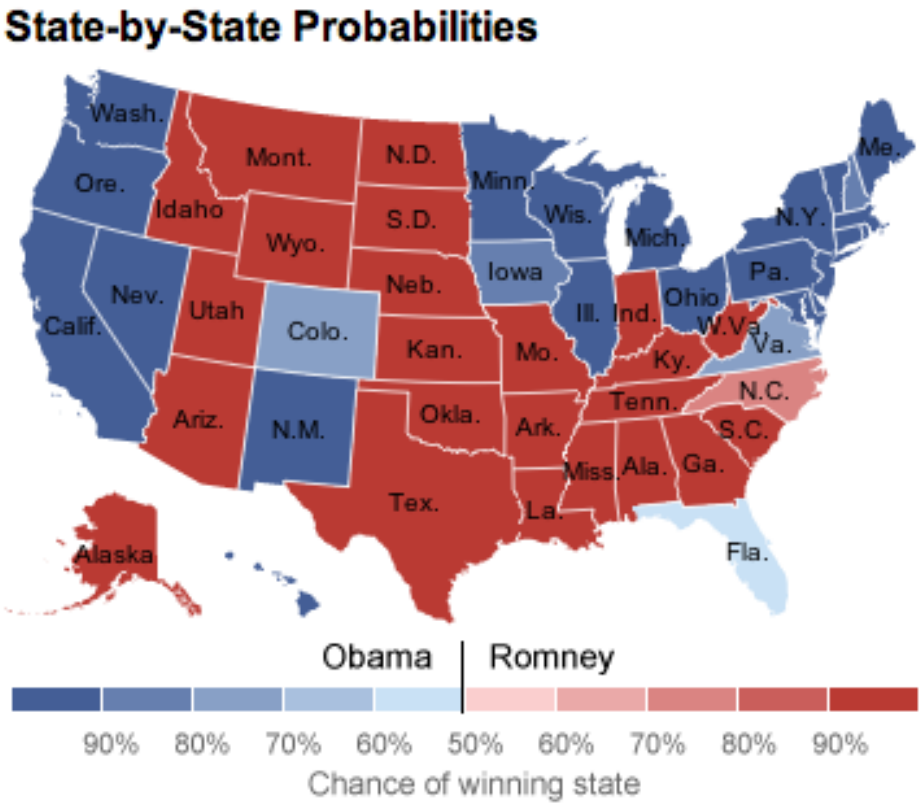
What is Data Analytics?

- Process that turns data into knowledge, insights and actions that improves decision making
- In other words it is deriving intelligence out of data
- Involves data collection, transformation, analysis and reporting

Data Analytics helps us:

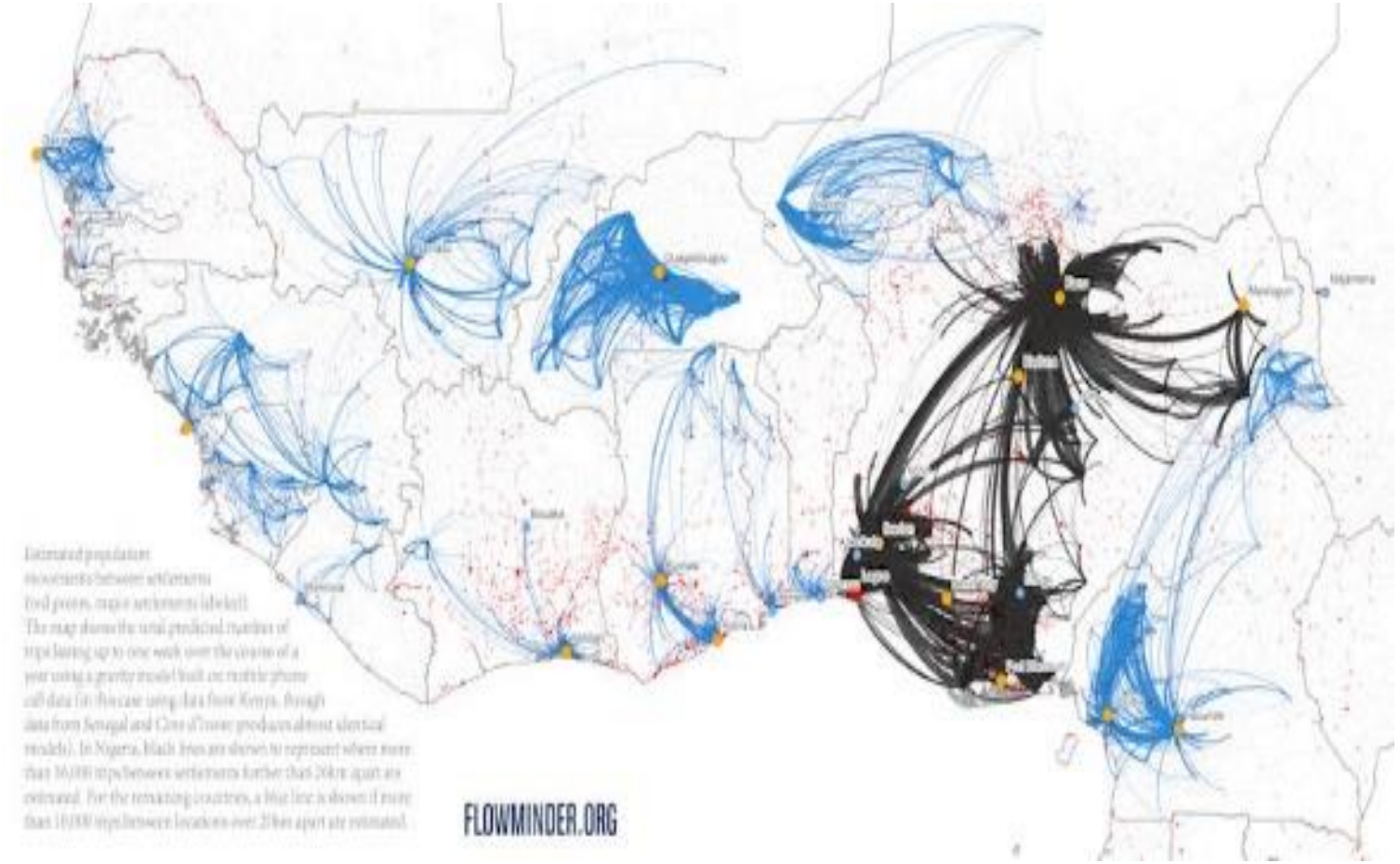
- To understand our world better
- To make better decisions
- To optimize processes

Predicting who is going to be the next US President



http://fivethirtyeight.com/

Predicting where Ebola is going to strike next



http://www.worldpop.org.uk/ebola/

Some practical cases

- Targeted customer marketing
- Optimizing stock management in manufacturing by predicting demand
- Product affinity analysis in retailing
- Attrition management
- Improving customer retention
- Text mining to gather feedback
- Finding potential candidate molecules that can be developed into drugs
- Prediction of revenue for a planned new store

Types of Analysis

Descriptive	What happened?
Diagnostic	Why did it happen?
Predictive	What will happen?
Prescriptive	What should I do?

Steps in Data Analysis Process

- Start with an interesting question
- Collect data
- Clean and Explore the data
- Create data models
- Interpret and Communicate the results

Job Titles

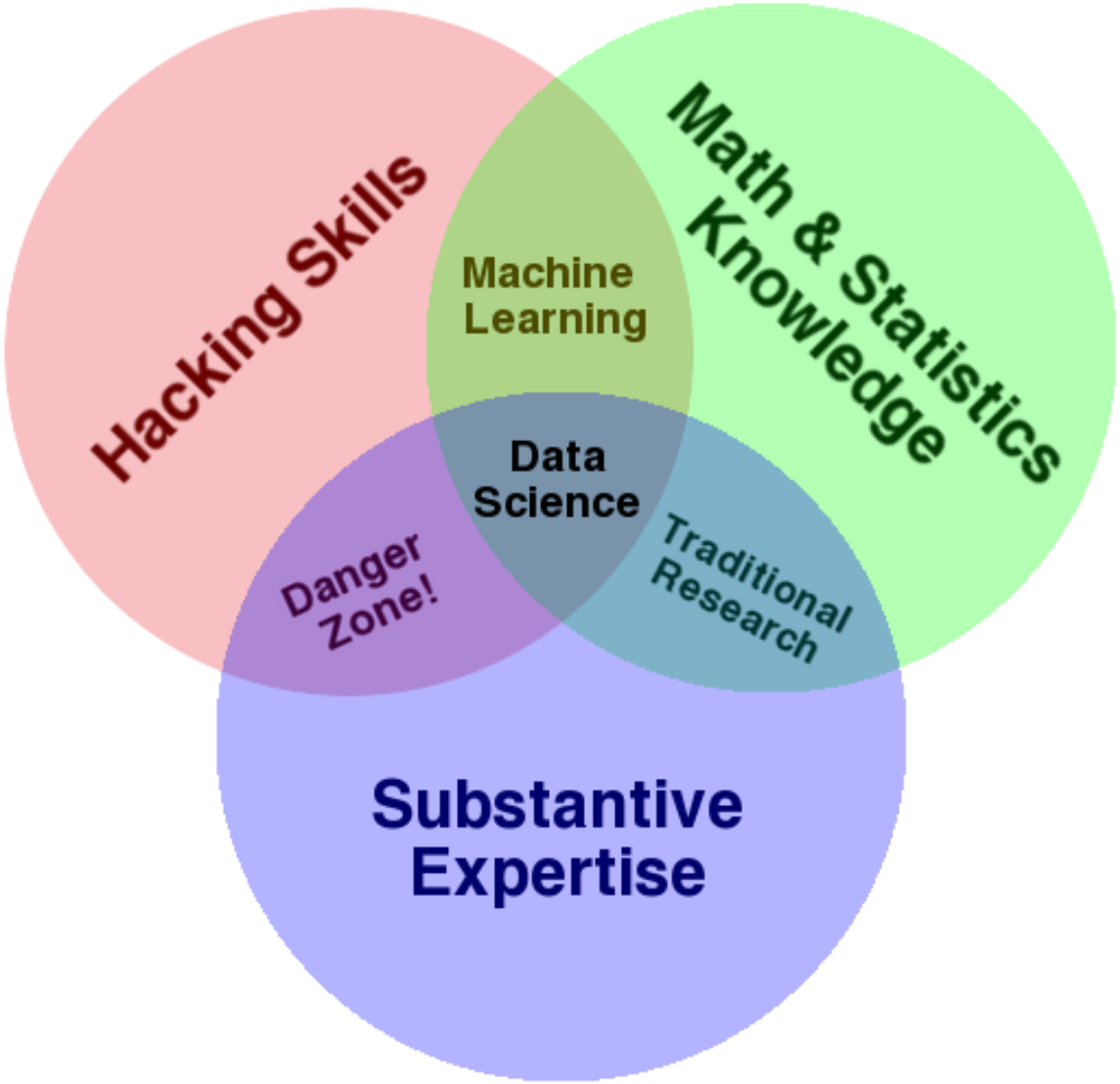
- Analyst
- Big Data Analyst
- Business Analyst
- Consultant
- Data Analyst
- Data Modeler
- Data Scientist
- Decision Scientist
- Machine Learning Specialist

Who is a data scientist?

Someone who knows more statistics than a computer scientist and more computer science than a statistician

- Joshua Blumenstock

Skills required



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Skills required

- Know how to analyse data and create data visualizations
- Experience with any statistical programming language (R, Python etc.)
- Experience in database querying languages (like SQL or MySQL)
- A good understanding of machine learning tools and techniques
- A good understanding of statistics (Hypothesis Testing, Summary Statistics etc.)
- Understanding of data wrangling / data munging
- Familiarity with big data tools (Hive, Pig etc.)

NICHE areas in Data Analytics

- Social Media Analytics
- Marketing Analytics
- Customer Analytics
- Supply Chain Analytics
- Demand Forecasting
- HR Analytics
- Risk Analytics
- Web Analytics

Tools used in Data Analytics

- R
- SAS
- Python
- MATLAB
- STATA
- SPSS
- Julia
- ...

Why R?

- Most comprehensive statistical analysis package
- Free and open source software
- Most popular data analysis tool
- More then 5000 packages (libraries) available
- Cross-platform capability
- Active user groups

Disadvantages

- Memory Limitation

Supervised Learning

- Variables under study can be split into 2 groups: explanatory variables and dependent variables
- Target is to specify a relationship between these two variables
- The relationship is obtained by analysing the training data
- Example: Prediction of revenue for a planned new store
- Explanatory variables: Store size, Location etc.
- Dependent variable: Revenue
- Training Data: Data from existing stores

Unsupervised Learning

- No distinct dependent variable
- All variables are treated the same way
- Target is to find patterns in data
- Example: Segmenting customers into distinct groups

Exercise

Classify whether the following problems as super vised or unsupervised:

- Predicting future stock market prices
- Identifying major topics people are tweeting about
- Classifying a tumour as either malignant or non-malignant
- Detecting spam in email
- Identifying groups of houses according to their house type, value and location