Big Data with Social & Search Networks

STATS 157 Final Report

Imran Yousuf

Big Data w/ Social $ Search Networks

By Imran Yousuf

Table of Contents Page#

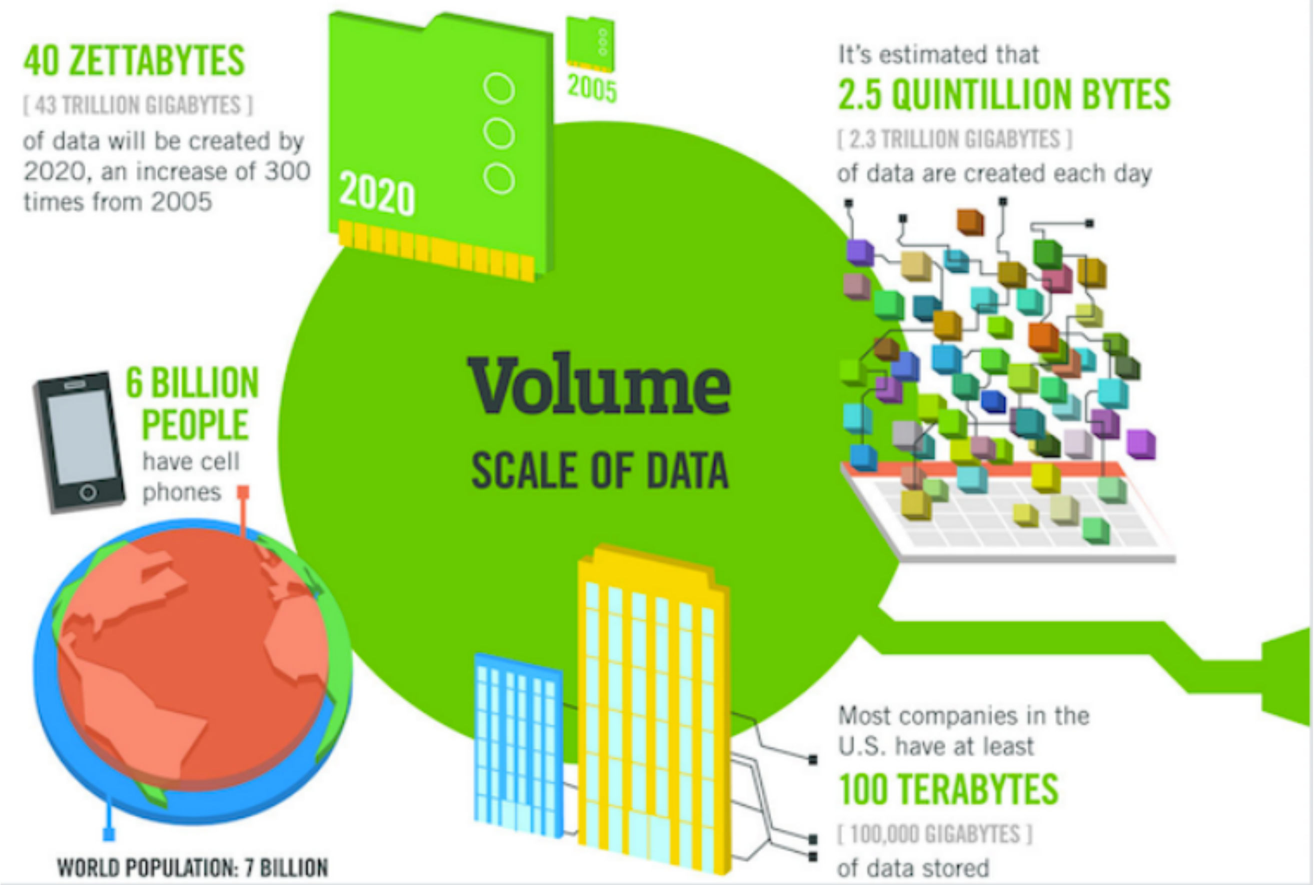
Table of Contents Page#

What is Big Data ?

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications.

The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization & privacy violations.

How Big is Big Data?


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

It is huge, estimated to that there will be 43 trillion gigabytes of data created by the year 2020.

About 2.3 Trillion Gigabytes of data is created everyday.

For more Reference:

Most companies have 100 Terabytes of Data Storage and backups in contrast to 100GB in 2000.

Six out of Seven people in the world has a cell phone.

Important Questions ?

Now the questions are:

1. What are these data ?
2. What do they consists of ?
3. Are there DATA that's not stored ?
4. Can we use these DATA ?
5. Where does statistics come into play ?
6. What is the process like ?
7. Is it worth it ?
8. Ethical Boundaries ?
9. Experience with Big Data ?

I will try to answer them in the following pages!

What are these Data ?

There are different types of data

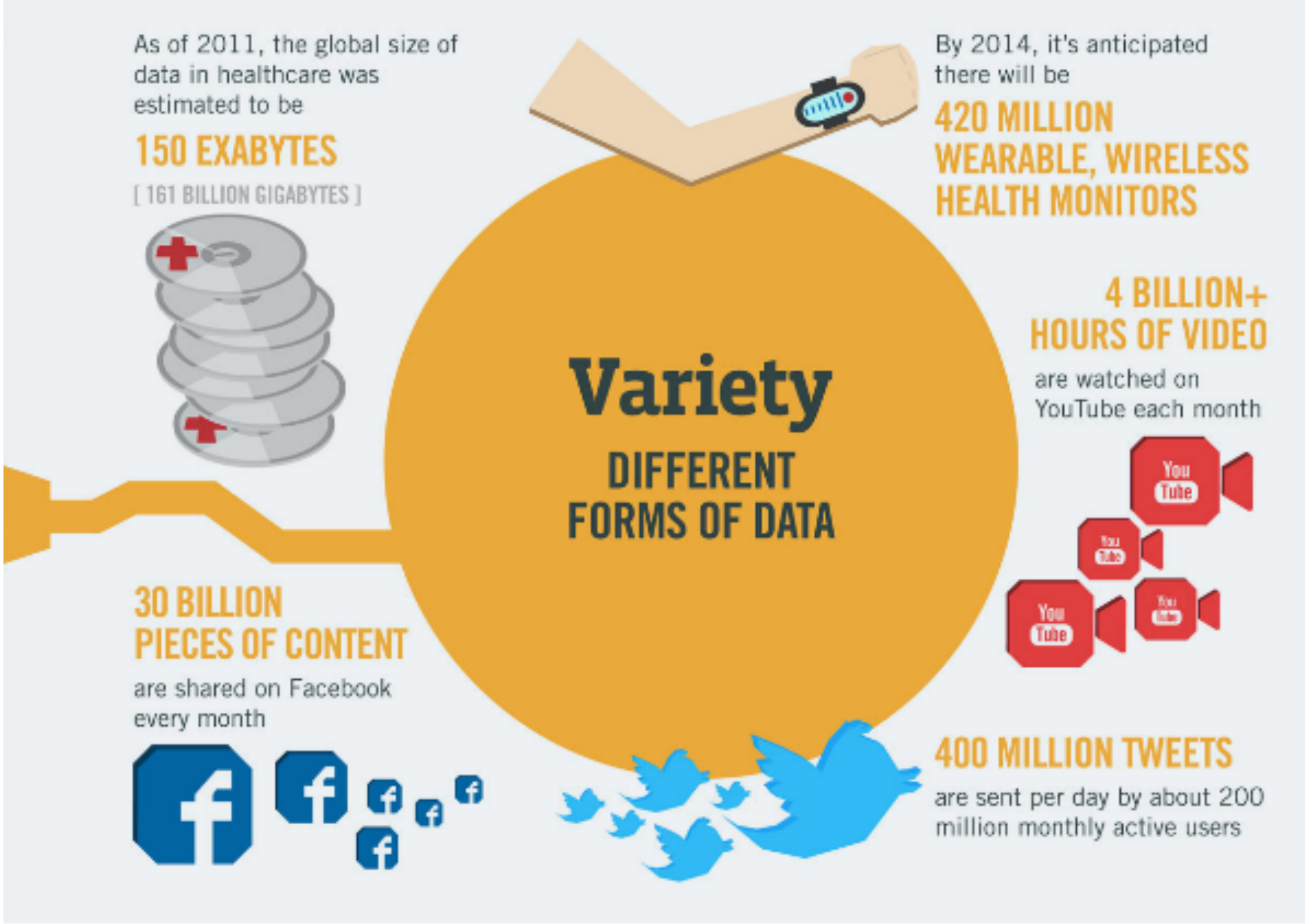
Variety of Data.

There are 30 billion pieces of content data from Facebook alone.

We watch 4+ billion hours of youtube streaming data.

There will be 420 million wearables wireless health monitors by 2015.

Obviously there are more data for instance the 161


As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
You Tube

Velocity: Analysis of the data

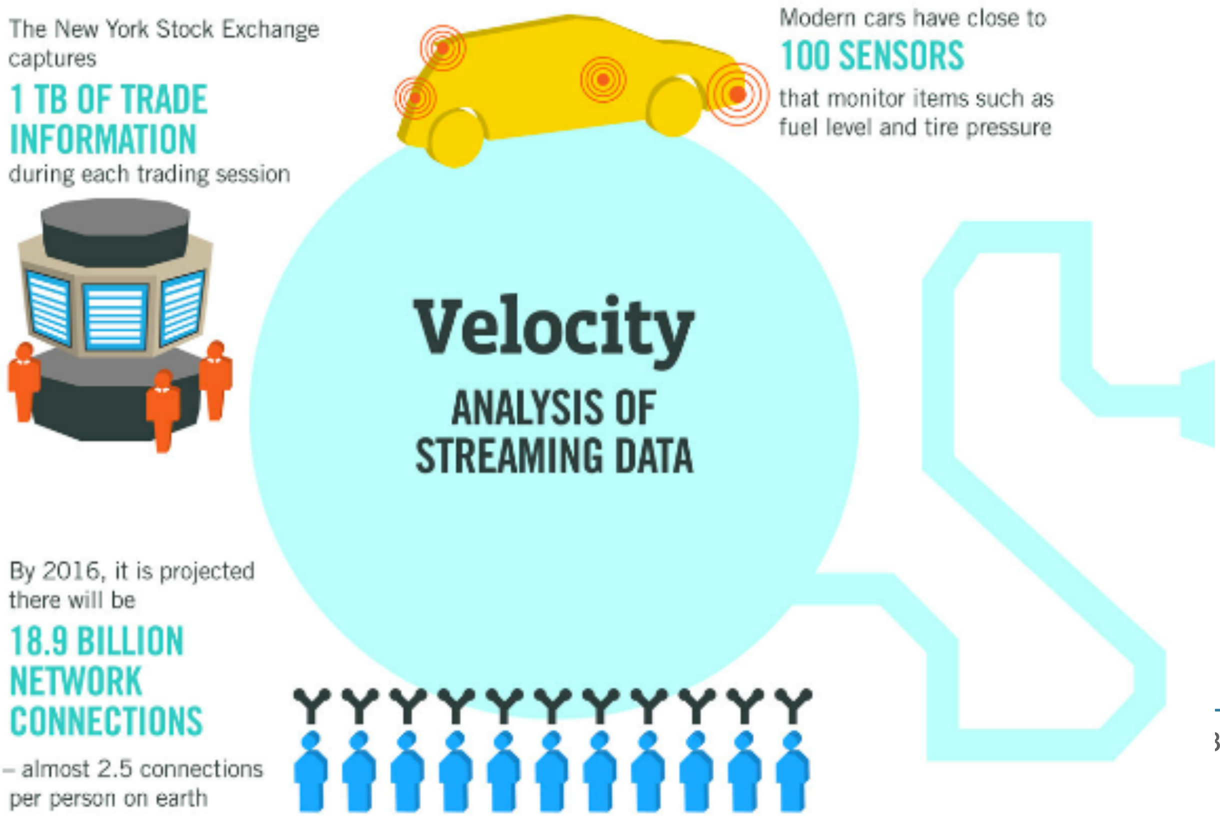
Next it is the **Velocity** of the streaming data.

In a modern car there are close to at least 100 sensors that communicates with the fuel level and much more just by streaming such data.

Or we can think about the NYSE which has a 1TB of trade data transferred everyday.

These kind of data is often not stored but transmitted and replaced.

In fact it is estimated that there will be 19 billion endpoints connection within 2016. That is the same as every human on planet earth having 2.5 phones or other devices.

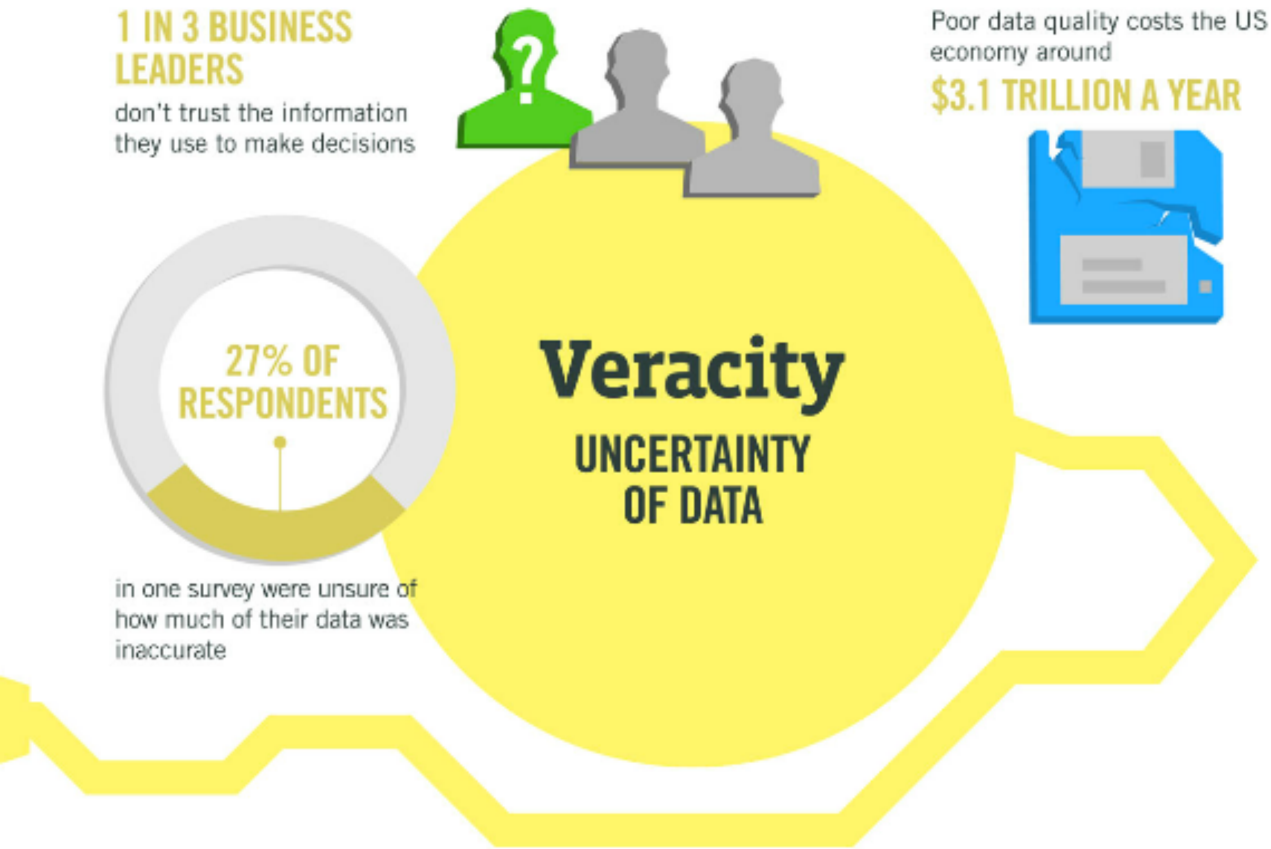

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Can we use any Data ?

The biggest is issue with data is its **relevance**. Getting data is often not the hardest part but to understanding the data takes a lot of work.

The **Veracity** i.e the uncertainty of data is such a big problem that 1 in 3 business leaders does not trust the information they use to make the decision.

In fact the bad quality of the data costs the US about 3.1 Trillion dollars a year.

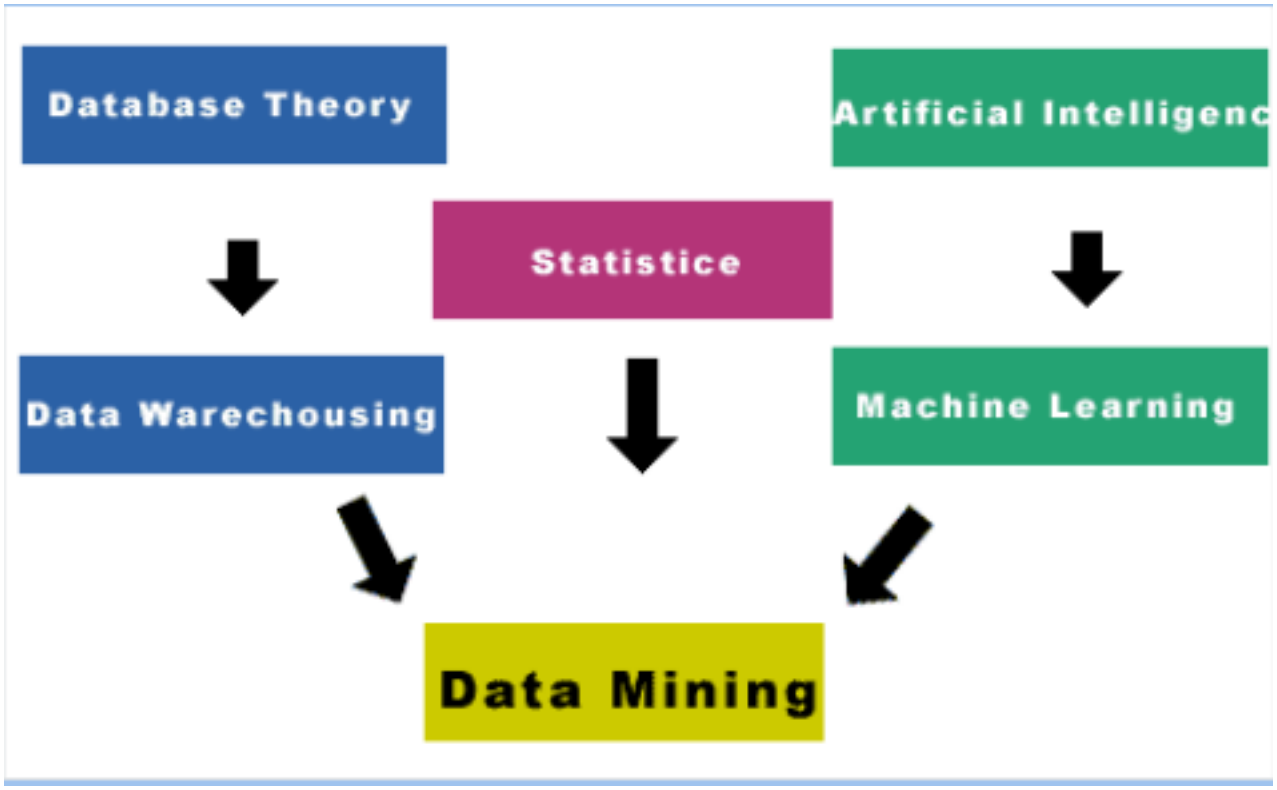

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

What Data can we use ?

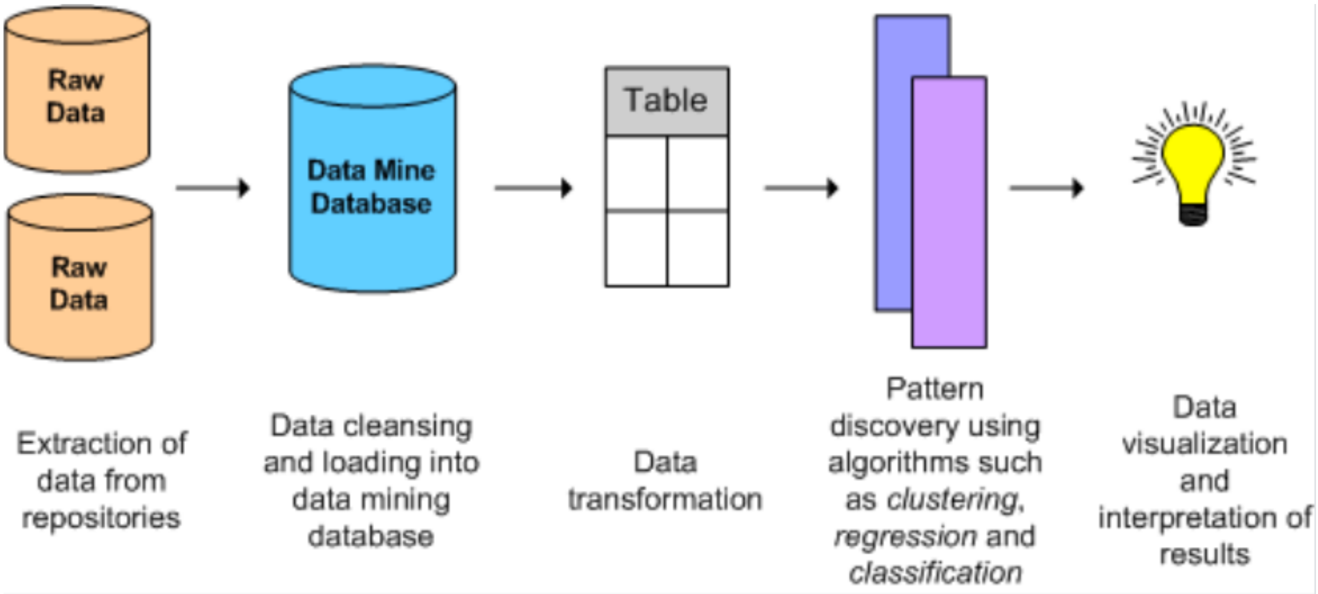
This is where **Statistics** comes into play largely.

We only use selected data from **DATA MINING**

Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. **The ultimate goal of data mining is prediction**



Process of Data Mining


Raw Data
Raw Data
Data Mine Database
Table
Extraction of data from repositories
Data cleansing and loading into data mining database
Data transformation
Pattern discovery using algorithms such as *clustering*, *regression* and *classification*
Data visualization and interpretation of results

The Process of Data Mining can be categorized into three stages:

1. Exploration

2. Model building and validation

3. Deployment

Brief Introduction about the process

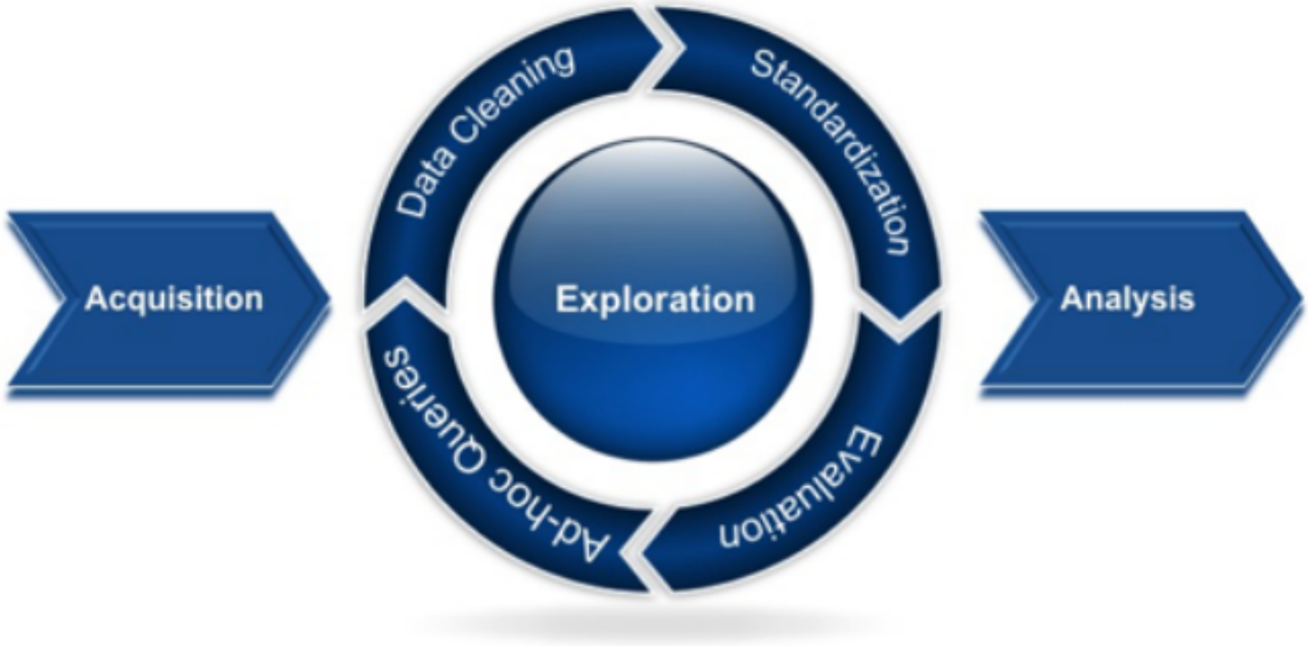
Exploration: Analyzing the huge set of data with data preparation which may involve cleaning data, data transformations, selecting subsets of records.

Model building and validation: This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples)

Deployment: That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

I will further deep into these processes deeply one by one.

Exploration

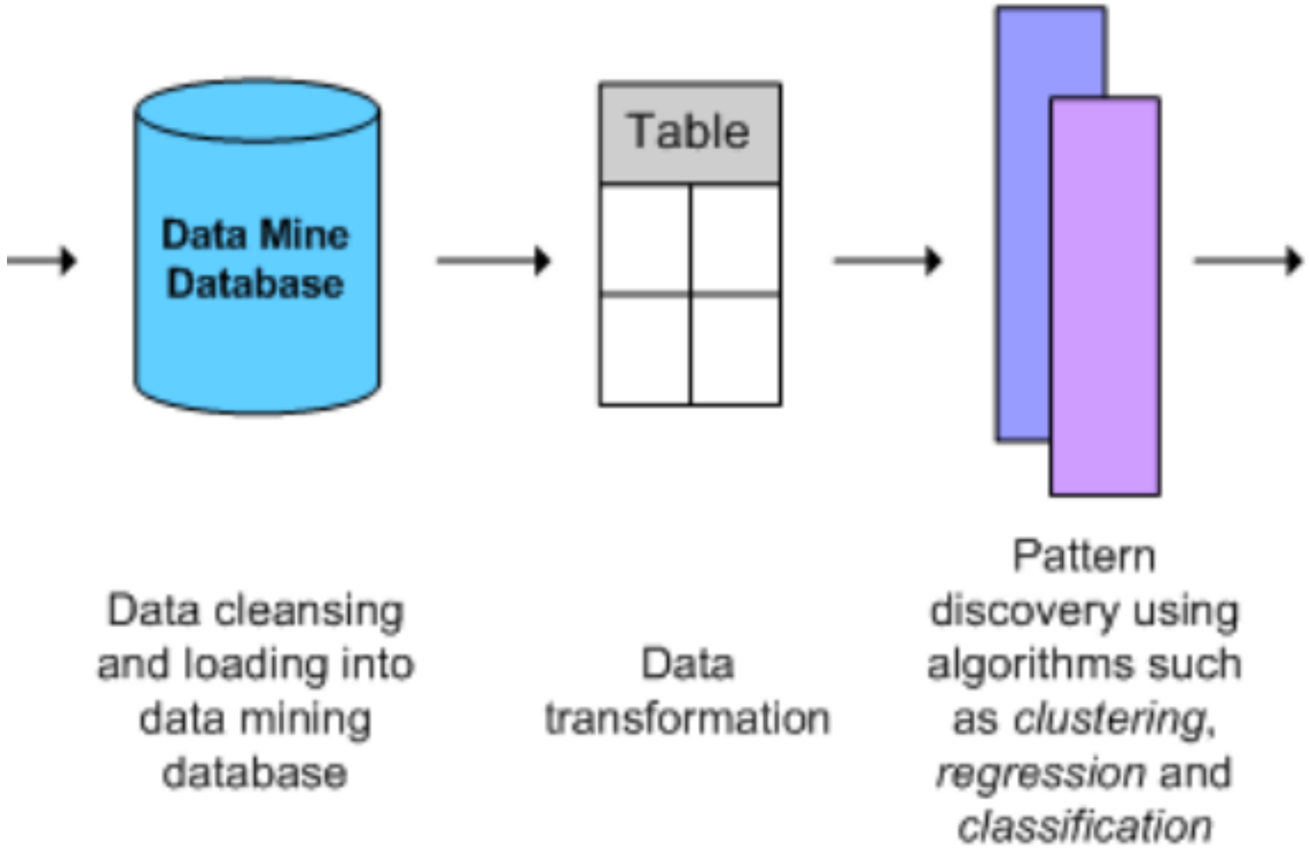


The idea here is to carefully understand what kind of data is present and what data do we **actually need**. There is huge cleansing of data at this stage.

It is estimated that during this process **88%** of data is ignored and only 12% of the data is collected to be moved to be validated onto the next stage.

Model Building and Validation

Once we have the extracted data then we get into the core of data mining which is building an appropriate model by choosing the best predictive performance algorithm to the set of data and then finally validating for deployment.


Data Mine Database
Table
Data cleansing and loading into data mining database
Data transformation
Pattern discovery using algorithms such as *clustering*, *regression* and *classification*

Deployment

After a satisfactory model or set of models has been identified (trained) for a particular application, we usually deploy those models so that predictions or predicted classifications can quickly be obtained for new data.

For example, a credit card company may want to deploy a trained model or set of models (e.g., neural networks, meta-learner) to quickly identify transactions which have a high probability of being fraudulent.

Now lets move onto some algorithms that makes these possible.

For Exploration

1. GIGO: Stands for "garbage-in-garbage-out". Very helpful when you want to clean your data knowing what you need and what you don't need. A fair amount of it is dealt with writing scripts and regular expression.

2. Drill-Down: Only get the data you need and forget about the rest of the data. For instance for web pricing you need the item number and the price the competitor is selling for.

FOR MODELING AND VERIFICATION

1. Bagging (Voting, Averaging): The concept of bagging is to combine the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data. This is basically machine learning algorithm for generating weights for weighted prediction.

2. Boosting: The concept of boosting is to generate multiple models or classifiers (for prediction or classification), and to derive weights to combine the predictions from those models into a single prediction or predicted classification.

For Deployment

1. Meta-Learning: The concept of meta-learning is to combine the predictions from multiple models. It is particularly useful when the types of models included in the project are very different.

2. Other **Machine Learning Algorithms**

Why go through such a hard process?

Data Mining becomes very **profitable**, specially when there is a **trend and connection** within each sample.

Examples:

1. Targeted Advertisement: Purchase rate tend to be 73% higher.

2. Sales forecasting: Competitive analysis of our product or our competitor.

3. Customer Loyalty: Helps build a brand.

4. Card Marketing: collects the information from usage, identify customer segments

.........and so much **more!!**

How do we make such connection ?

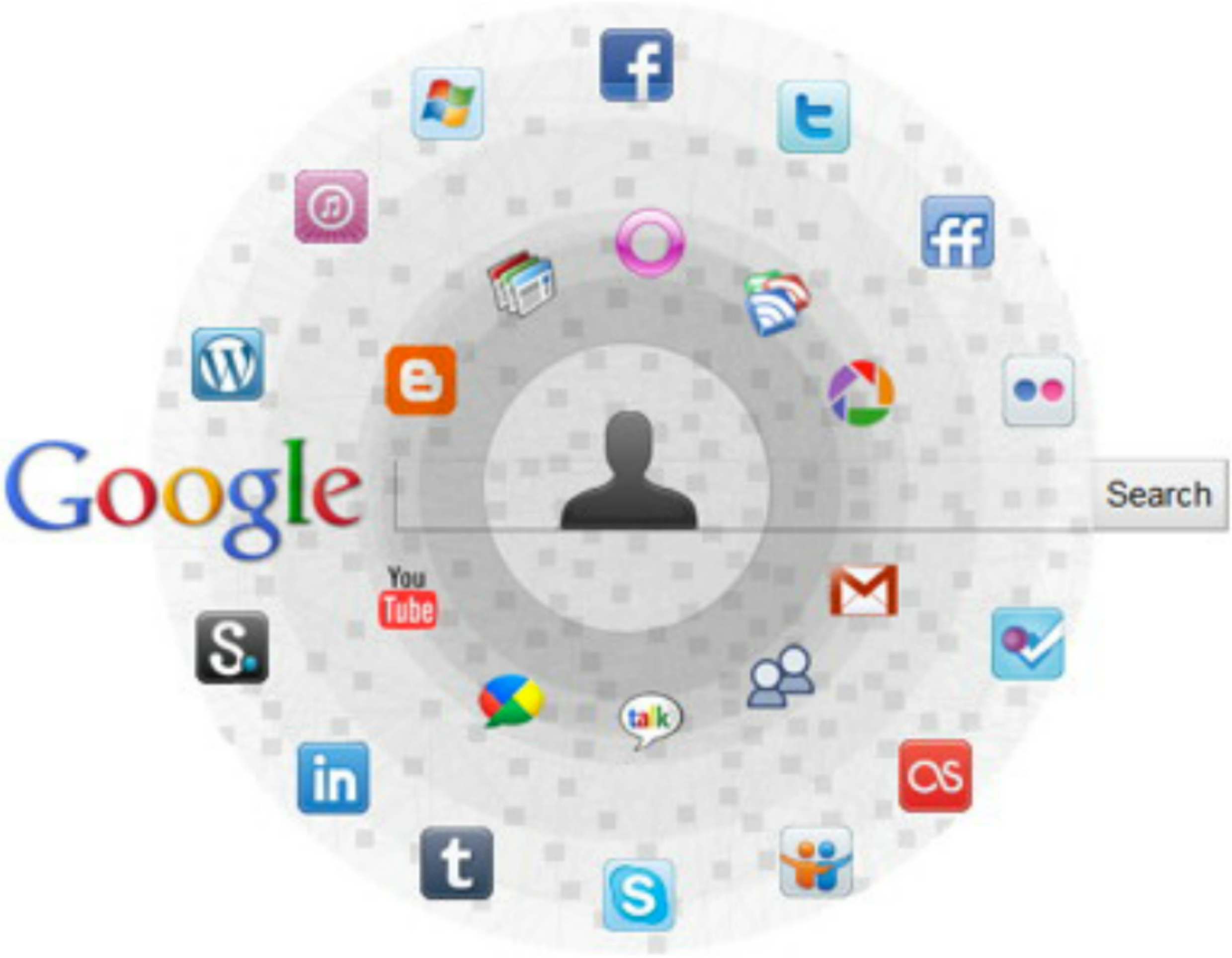
We make these Connections by:

Scrapping data and **buying data** of the web.

FACT: Almost everyone scrapes,sells and buys data

Data is not very useful if you cannot make **association** with it.

Who is Almost Everyone ?



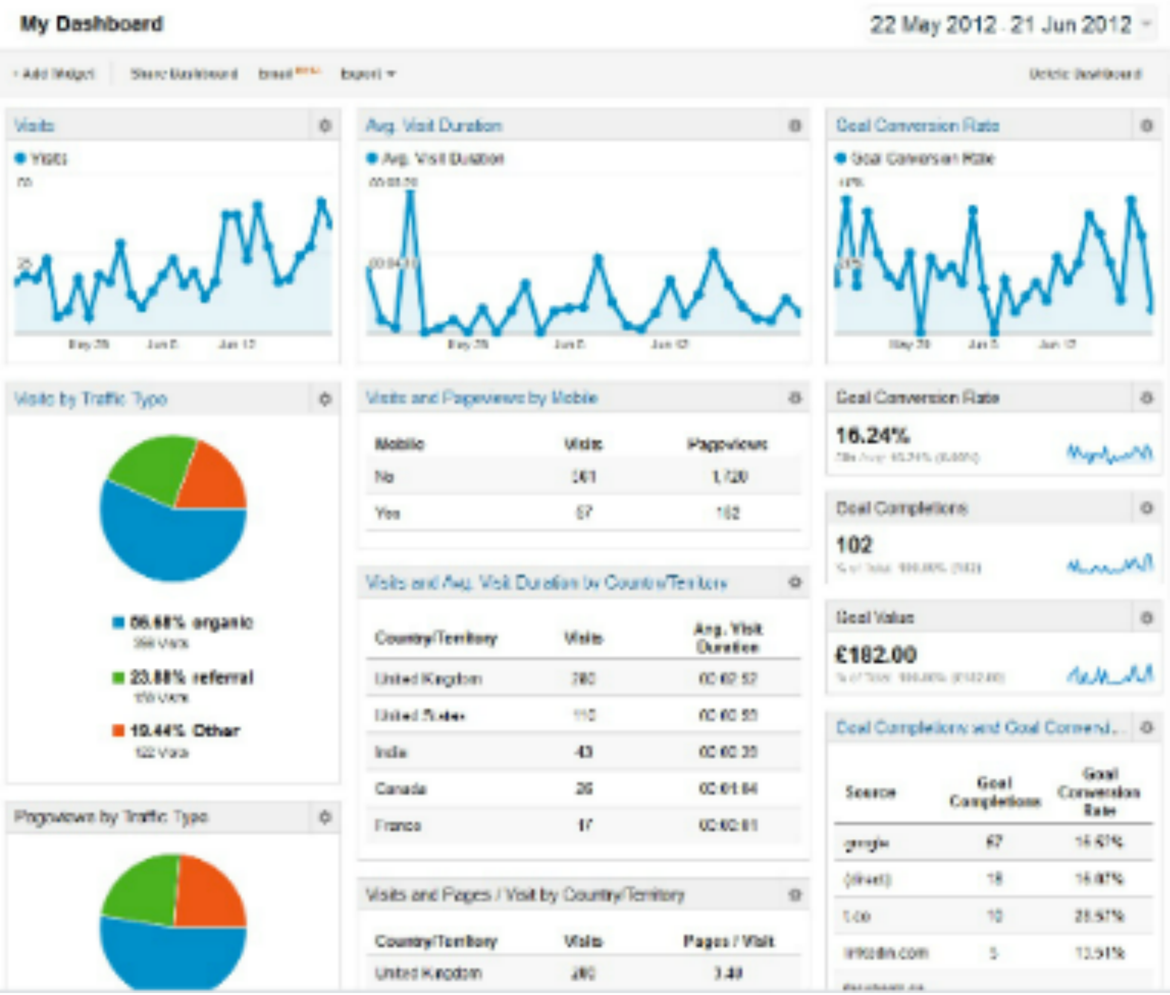
Any social and search media you can think of relies on data. Not only third party companies, but our **government** scrapes data. **The main point in Big Data is to make prediction** and there is no better way of making data instead of using their relative data and making association with it.

Search & Social

If we are able to make the connection with the search and social then there is huge social and economic interest that promotes predictions.

Prediction of products can lead into the beneficial of a company and in return plays a huge role in our current stock market.

How is it possible ?



Visits and Pageviews by Mobile

Mobile	Visits	Pageviews
No	361	1,720
Yes	57	162

Visits and Avg. Visit Duration by Country/Territory

Country/Territory	Visits	Avg. Visit Duration
United Kingdom	280	00:02:52
United States	110	00:00:50
India	43	00:00:33
Canada	26	00:01:04
France	17	00:00:01

Visits and Pages / Visit by Country/Territory

Country/Territory	Visits	Pages / Visit
United Kingdom	280	3.48

Goal Completions and Goal Conversi...

Source	Goal Completions	Goal Conversion Rate
google	67	16.67%
(direct)	18	16.07%
t.co	10	28.57%
linkedin.com	5	13.51%

All of our interactions are being **watched**. It is scary to think that not only everything we do on the web is tracked, but also a lot of our offline activities are tracked. The offline data is often called **"reports"** and we often accept it during installation of a product.

Where are the Ethical Boundaries?

Although there are Ethical Data Mining such as getting anonymous reports without knowing who the person is, but in practice very few data miners actually follow any instruction or are restricted by any Ethical boundaries.

In fact, there were an increase of **258%** (source: amstat.org) of new Data Mining companies and a lot of these companies have their servers and data **over seas** in order of risk of getting **sued.**

My Experience with Big Data

As a summer project I was given a task to get all events from a specific query from a website (as in in a given date and area, find all the events going on) and store it in a relational database.

I was able to get all the events in that area using data mining techniques and doing cross reference verifications.

Even with a small processing unit I was able to generate a **145GB XML** file.

I used **Scrappy** (a python web scrapping library).

There are other tools along with such libraries:

iPython

NodeJs Crawler

Selenium Web Driver

……………………….& much more

Thank You

I am glad that I took this class. Thank you so much Professor, David Aldous.

I honestly feel as if I understand the world better.

Thank you again.

Source:

"Big Data." *Wikipedia*. Wikimedia Foundation. Web. 18 Dec. 2014. <http://en.wikipedia.org/wiki/Big_data>.

"Data Mining: What Is Data Mining?" *Data Mining: What Is Data Mining?* Web. 18 Dec. 2014. <http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm>.

"What Is Data Mining (Predictive Analytics, Big Data)." *What Is Data Mining, Predictive Analytics, Big Data*. Web. 18 Dec. 2014. <http://www.statsoft.com/textbook/data-mining-techniques>.