

Data Science 101

Husein Zolkepli, Machine learning Engineer at
DigitalHill Sdn Bhd

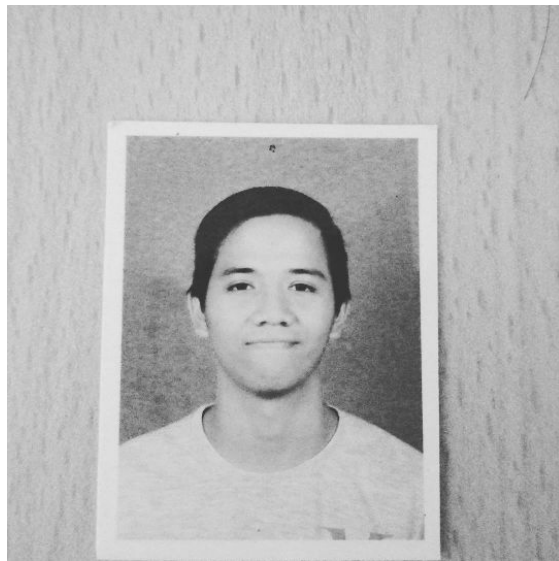


Introduction of myself

Call me Husein.

Experienced with Machine Learning about 2+ years

1. Developed OLAP cube
2. Confident Stock Engine
3. Closed sentiment analysis engine
4. Closed tagging analysis engine
5. Hyper Regression engine
6. Used bayesian to predict traffic system
7. [So much more my github is here](#)



Right now at DigitalHill,

I do local sentiment analysis.



What does your today tweet said? What it's polarity? Is it bring a positive or negative impact to society? Is it relatable with any subjects?



Join some Kaggle or other online competitions

	House Prices: Advanced Regression Techniques In progress - Top 5% · Competition Ineligible for Medals 11 entries as a solo competitor	77 th of 1691
	New York City Taxi Trip Duration In progress - Top 10% · Competition Ineligible for Medals 8 entries as a solo competitor	113 th of 1224
	Personalized Medicine: Redefining Cancer Treatment In progress - Top 12% · Competition Ineligible for Medals 15 entries as a solo competitor	135 th of 1175

 **data.gov.my**

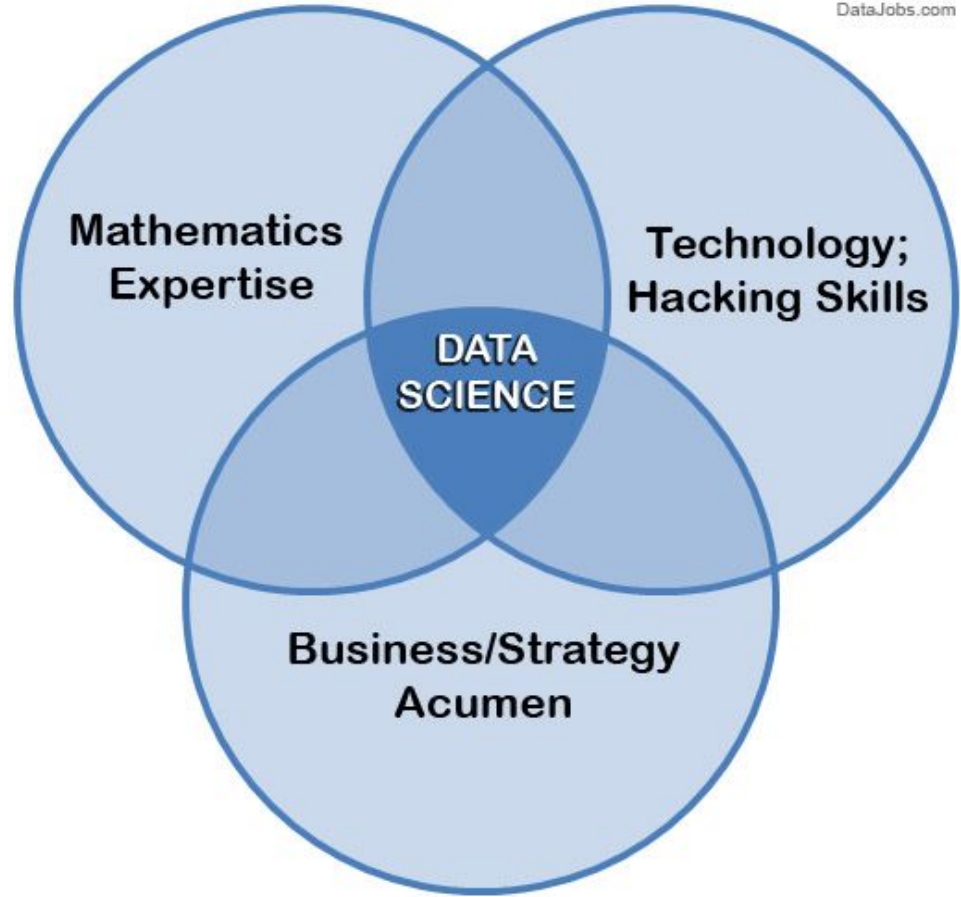
Ru	Text Normalization Challenge - Russian Language In progress - Top 47% 2 entries as a solo competitor	38 th of 81
En	Text Normalization Challenge - English Language In progress - Top 30% 5 entries as a solo competitor	57 th of 192
	Zillow Prize: Zillow's Home Value Prediction (Zestimate) In progress - Top 14% 9 entries as a solo competitor	401 st of 2866
	Carvana Image Masking Challenge In progress - Top 86% 1 entries as a solo competitor	542 nd of 631

kaggleTM

What is Data Science?

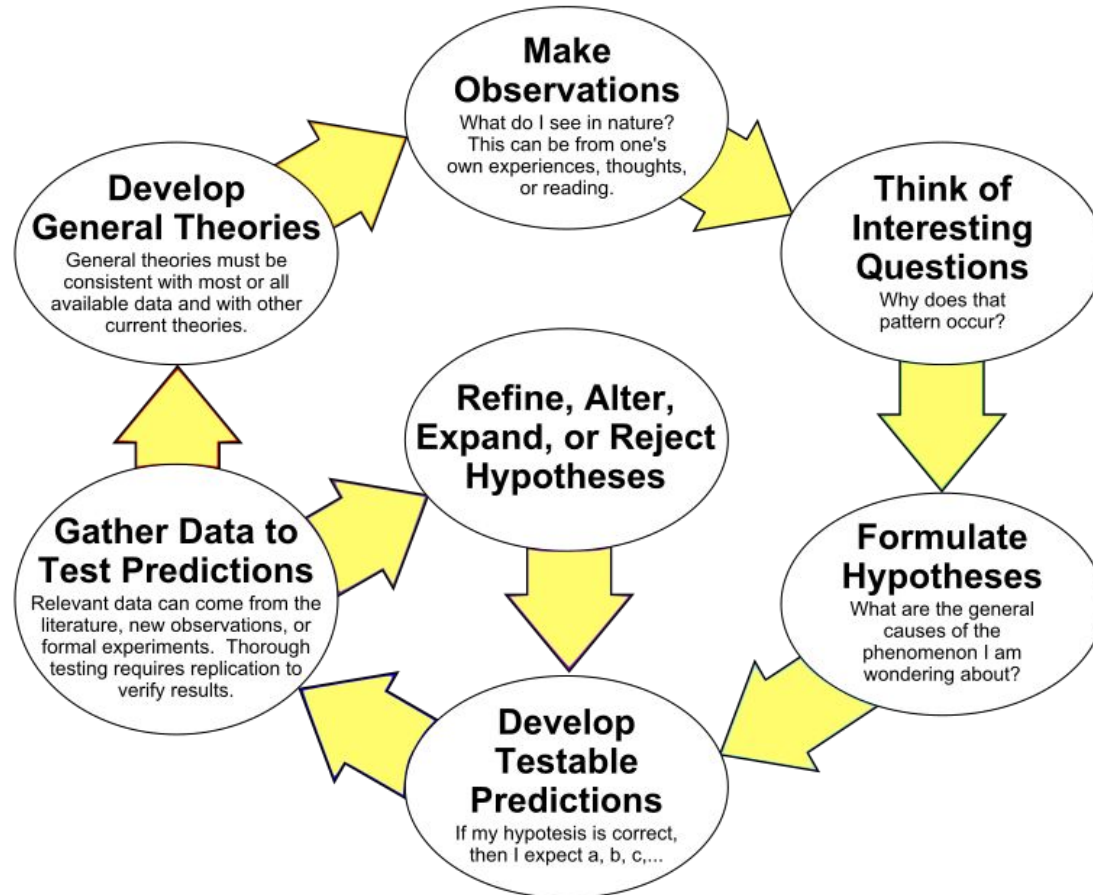
Your future decision based on data you got.

Technically said, **data-driven** science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured.



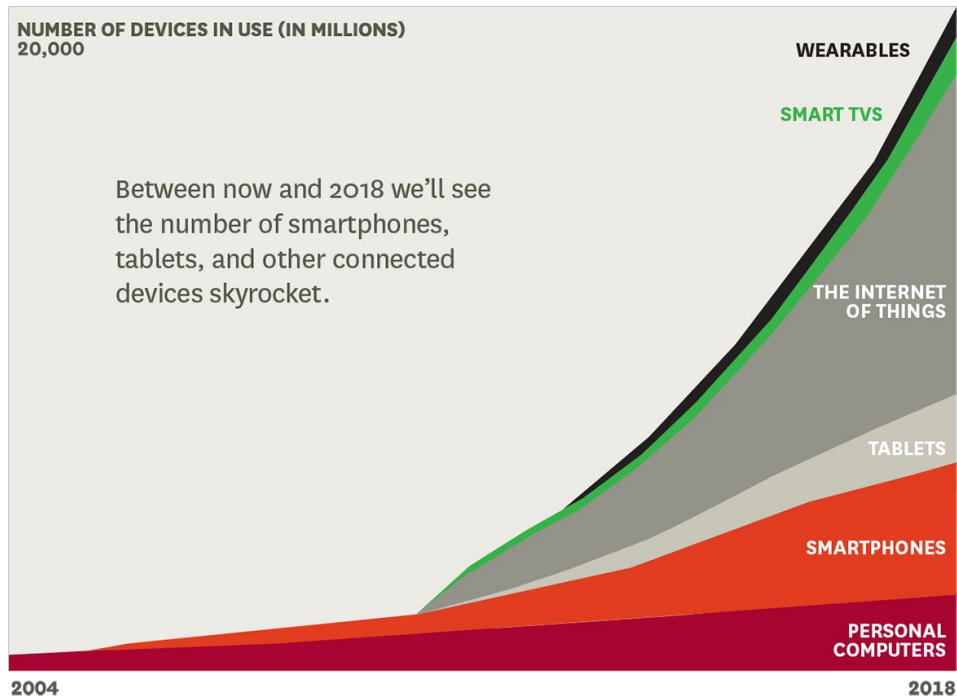
The Scientific Method as an Ongoing Process

scientific
methods



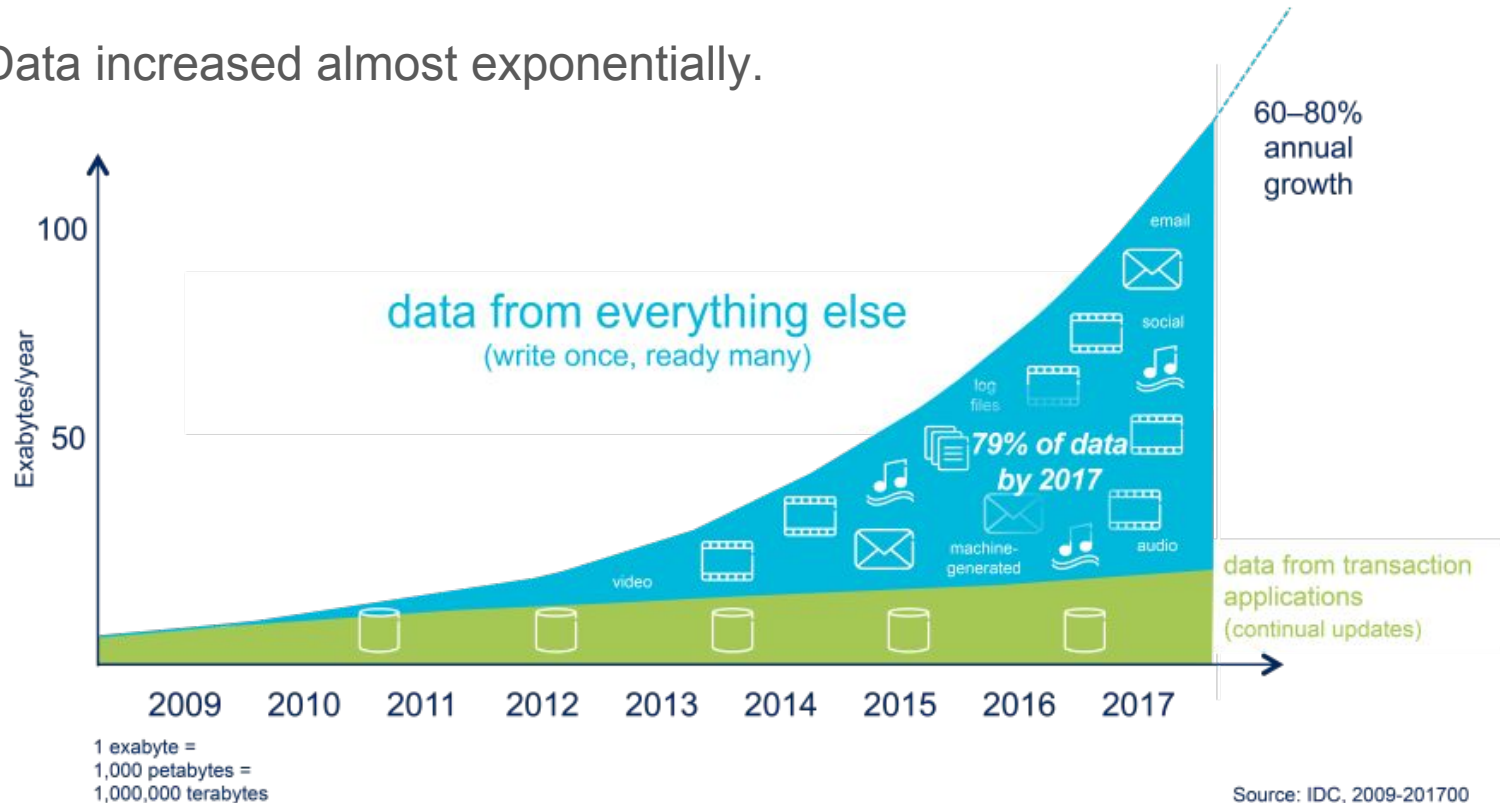
Why suddenly data science become a boom?

"Data scientist" has become a popular occupation with Harvard Business Review dubbing it "The Sexiest Job of the 21st Century" and McKinsey & Company projecting a global excess demand of 1.5 million new data scientists.



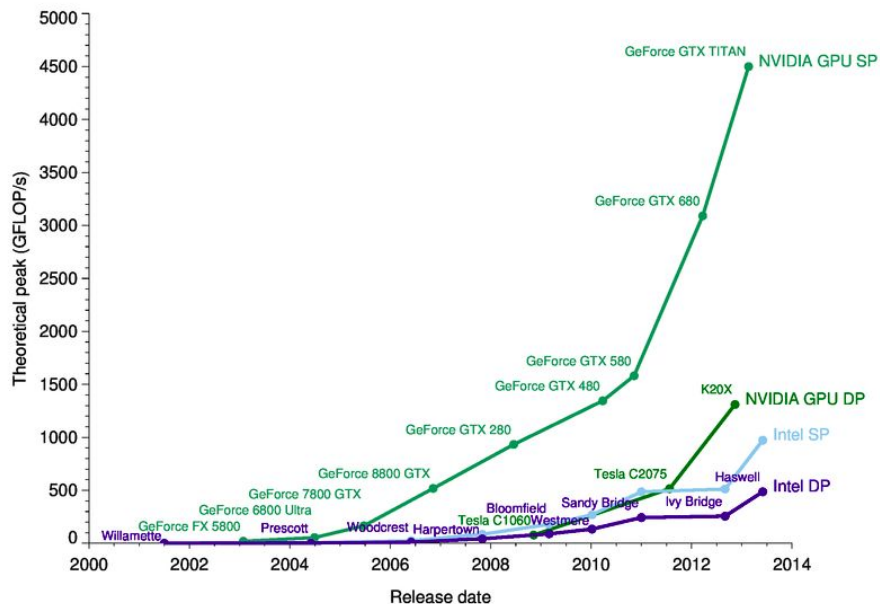
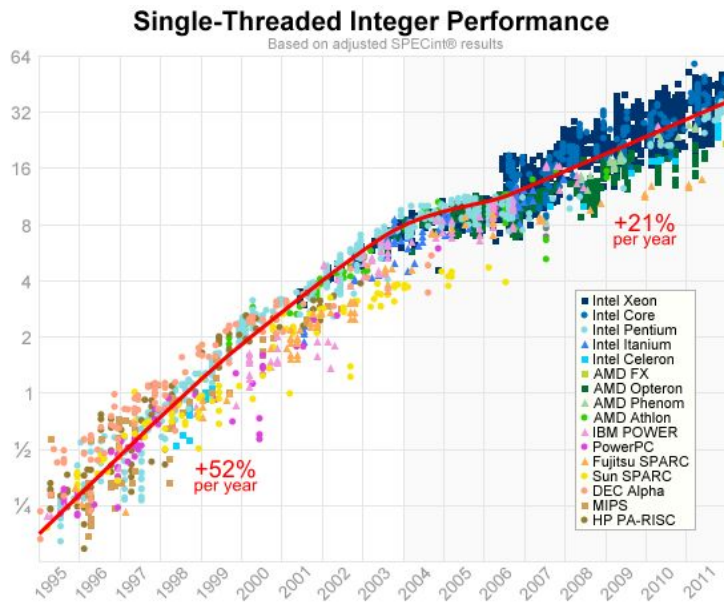
Why suddenly data science become a boom?

1- Data increased almost exponentially.

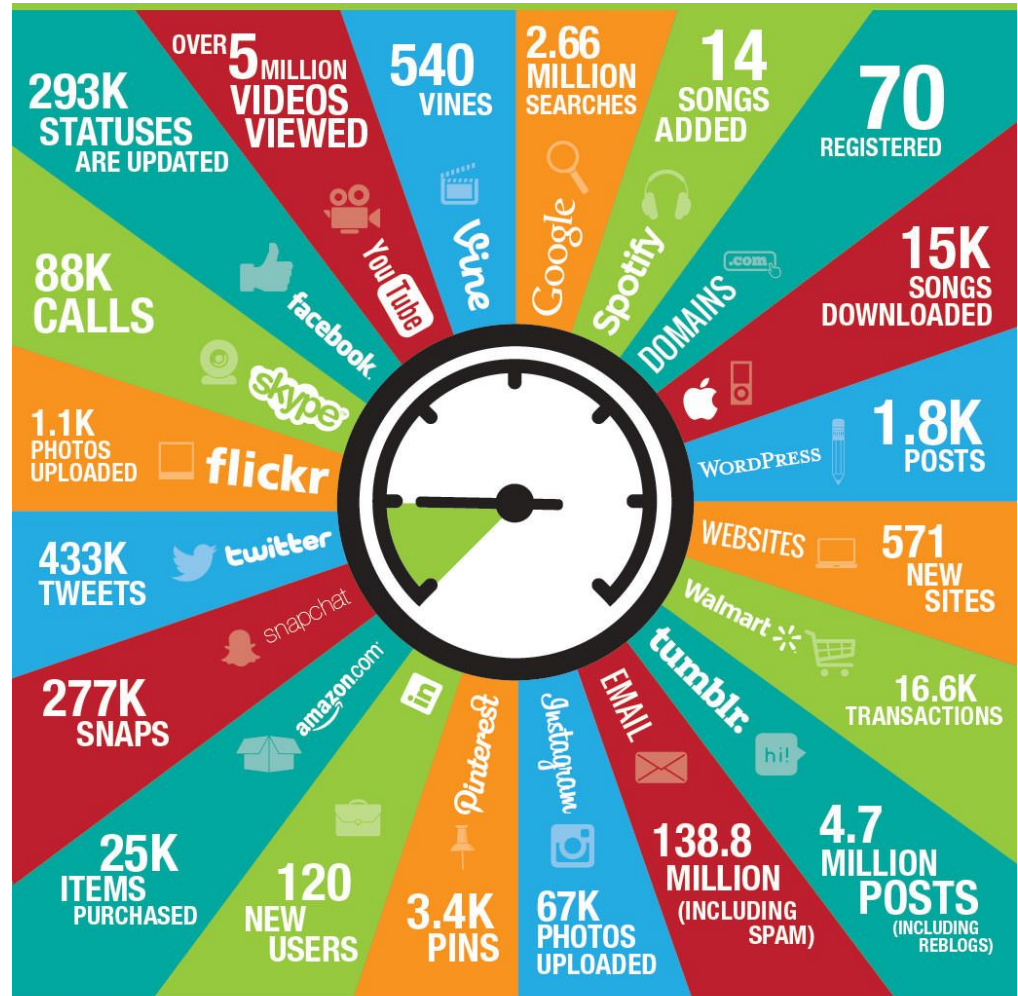


Why suddenly data science become a boom?

2- Machine capabilities increased almost exponentially.

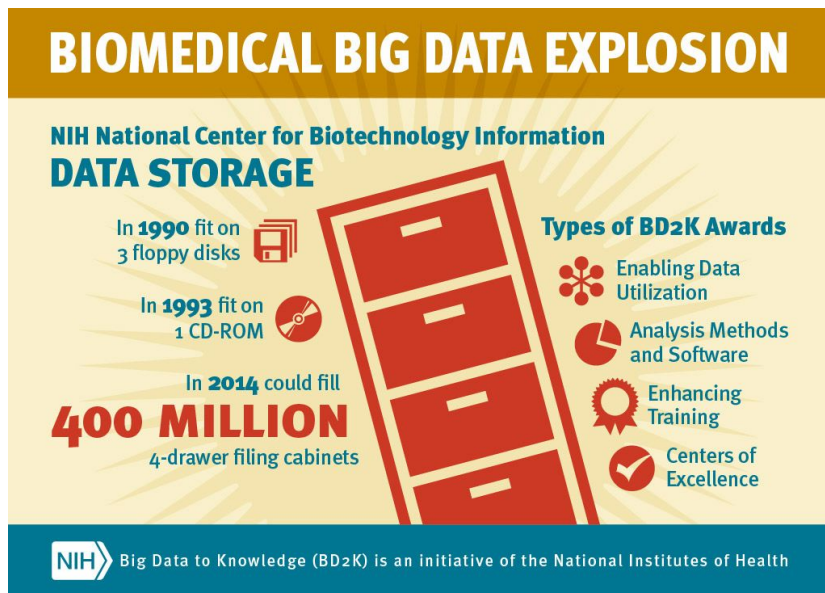


Why suddenly data science become a boom?

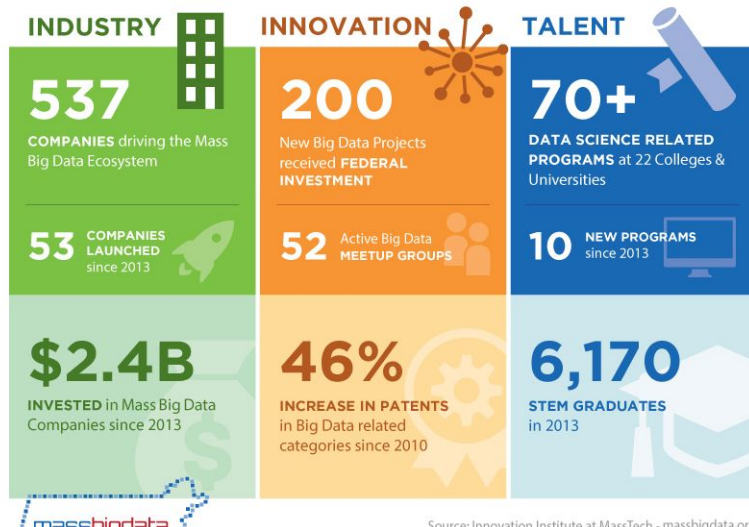


Why suddenly data science become a boom?

3- Now literally everything can be represented as digital data!



THE MASS BIG DATA ECOSYSTEM 2015



2 different job perspective for Data Scientist

1- Business Data Scientist / business analyst / data analyst

someone who analyzes an organization or business domain (real or hypothetical) and documents its business or processes or systems, assessing the business model or its integration with technology.

2- Technical Data Scientist / data scientist

Someone who analyzes an organization or business domain (real or hypothetical) and documents its technical requirement, directly data manipulation by coding or interface.

Business Analyst

Business
Data
Scientist /
business
analyst /
data
analyst

```
public class TcpClientSample
{
    public static void Main()
    {
        byte[] data = new byte[1024]; string input, stringData;
        TcpClient server;
        try
        {
            server = new TcpClient("...", port);
        } catch (SocketException)
        {
            Console.WriteLine("Unable to connect to server");
            return;
        }
        NetworkStream ns = server.GetStream();
        int recv = ns.Read(data, 0, data.Length);
        stringData = Encoding.ASCII.GetString(data, 0, recv);
        Console.WriteLine(stringData);
        while (true)
        {
            input = Console.ReadLine();
            if (input == "exit") break;
            if (input != null)
            {
                newChild.Properties["id"] = "1";
                newChild.CommitChanges();
                newChild.Close();
            }
        }
    }
}
```

What my friends think I do



What my parents think I do



What society thinks I do



What my boss thinks I do



What I think I do

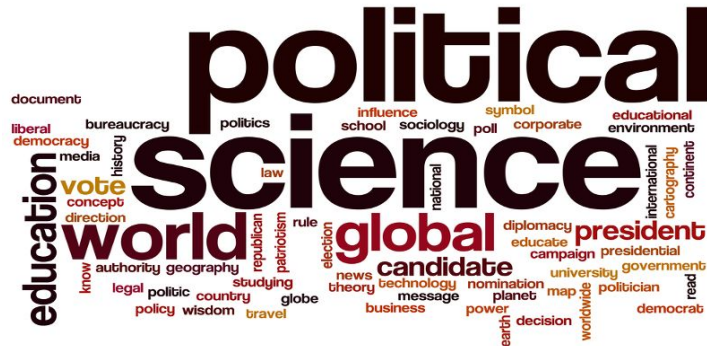


What I actually do

Business Data Scientist / business analyst / data analyst

1- can be mostly by anyone

Business administration, finance, political science, economics, anthropology, psychology also can.



Business Data Scientist / business analyst / data analyst

2- business heuristics for decision making

Focus on business insight from data analysis to apply on business strategic marketing.



Business Data Scientist / business analyst / data analyst

3- define business problem from the use case and translate into technical analysis for data scientist

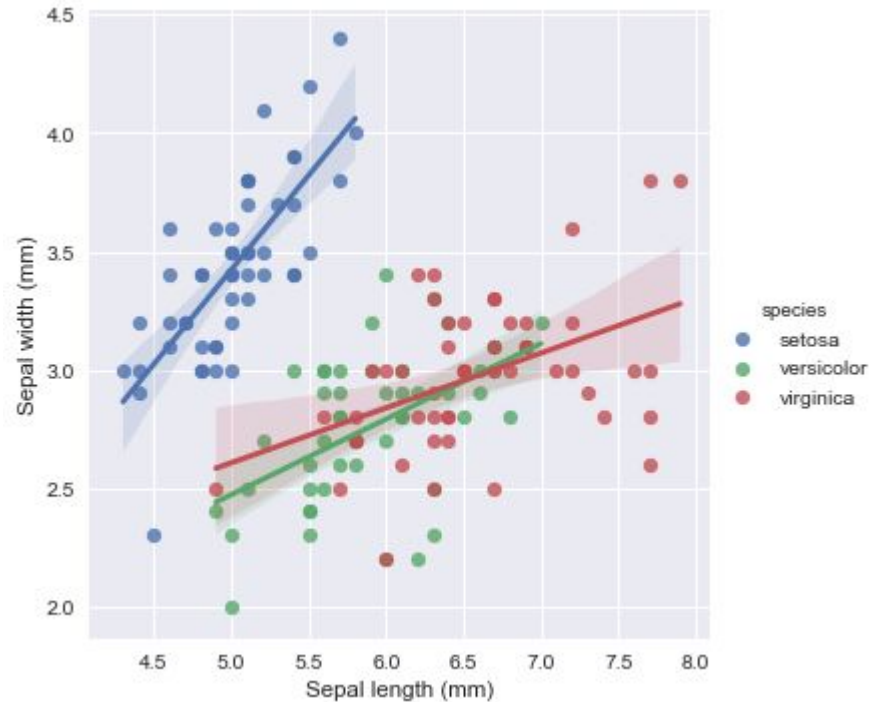
The problem with technical person, they really hard to close the gap between client and technical understanding



Business Data Scientist / business analyst / data analyst

4- study structured / unstructured data using data analysis for business users

They do data visualization using layman term.



Business Data Scientist / business analyst / data analyst

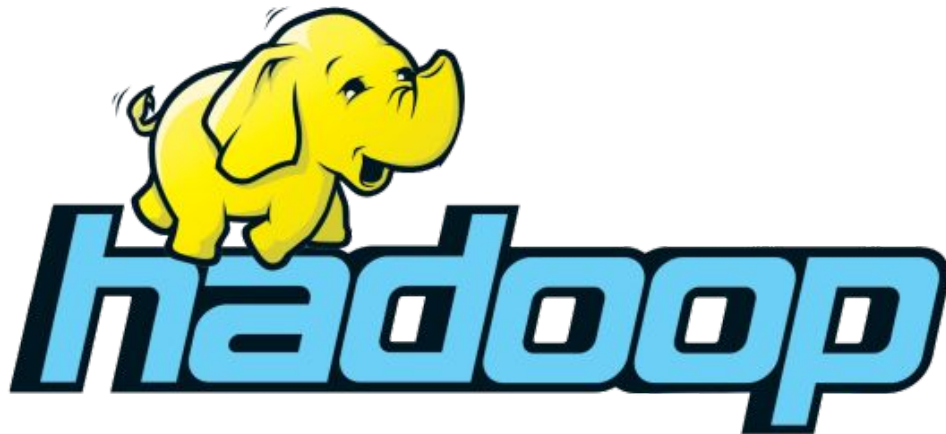
5- understand on how to integrate statistical models with business models



Business Data Scientist / business analyst / data analyst

6- define suitable architecture for both client and technical sides.

They will proposed to use Hadoop, or Spark for reducing technique.



Technical
Data
Scientist /
data
scientist

Data Scientist



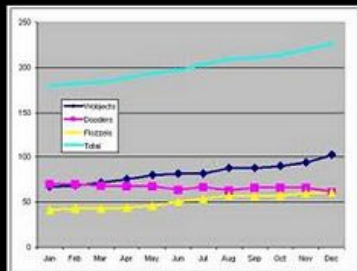
What my friends think I do



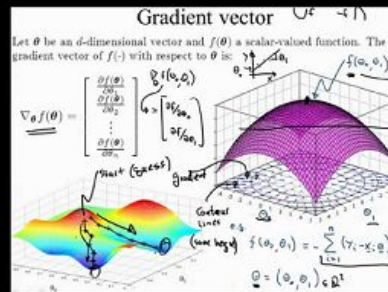
What my mom thinks I do



What society thinks I do



What my boss thinks I do



What I think I do



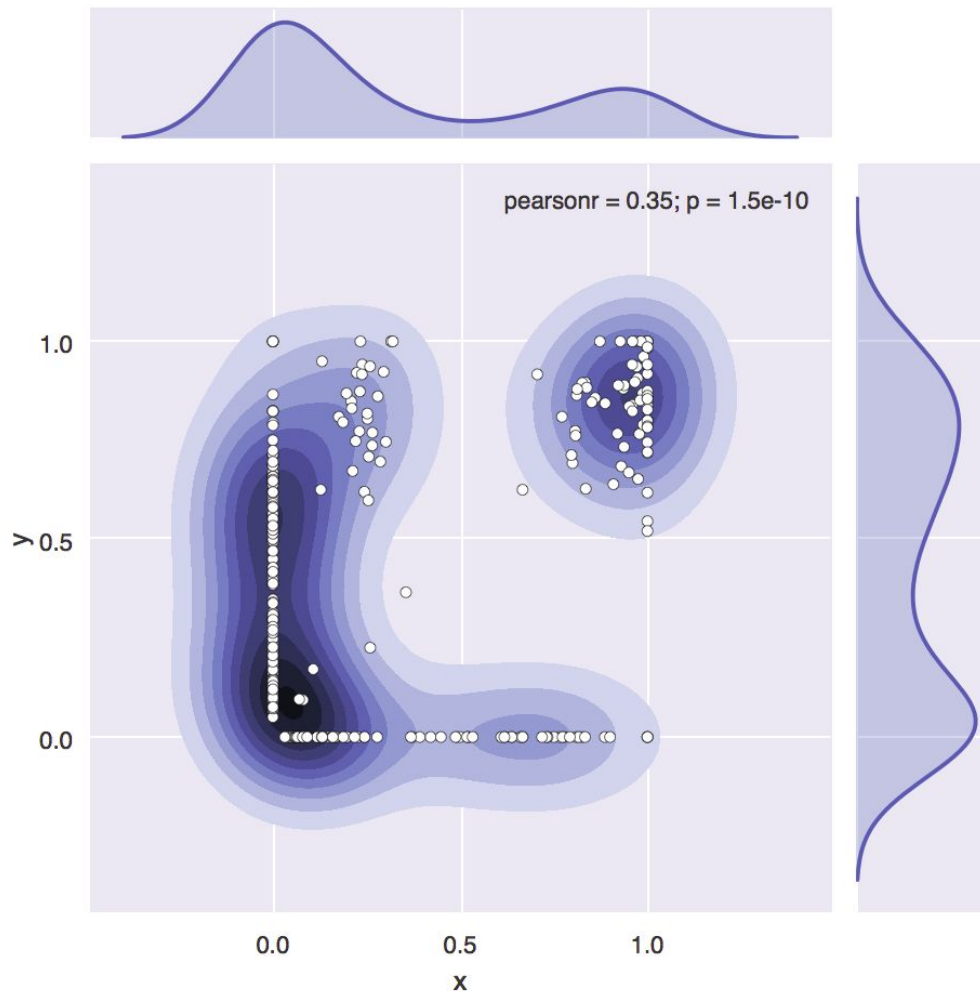
What I actually do

Technical Data Scientist / data scientist

1- Not everyone can do

For seriously, you need math and statistical understanding.

No need to go deep, but at least understand abstract on how to apply it.

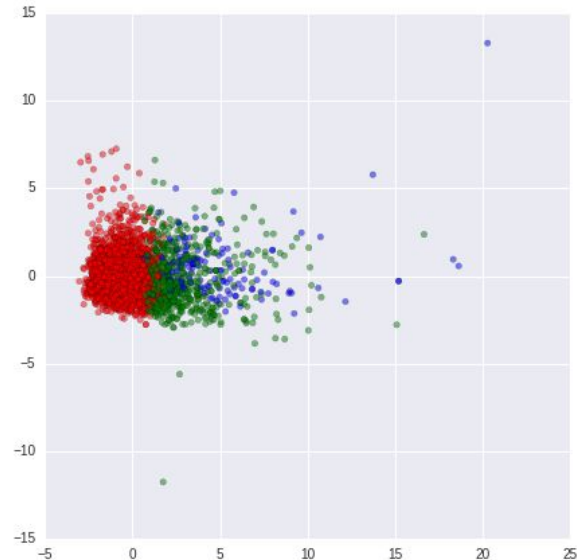


Technical Data Scientist / data scientist

2- mathematical heuristics for decision making.

What algorithm do we need to apply for that data, for this data.

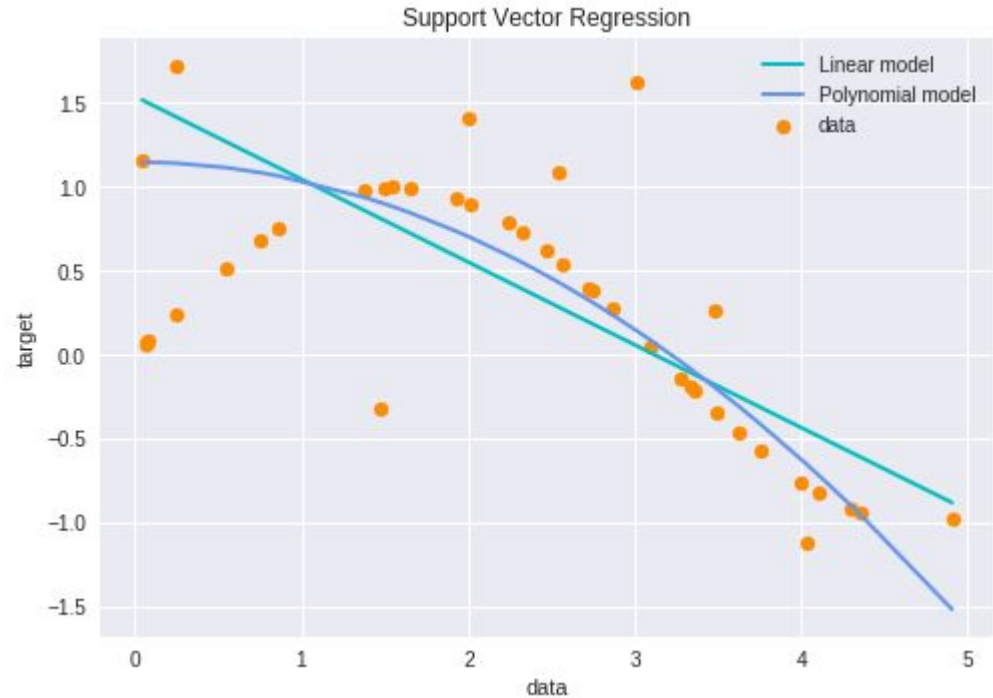
Overlapping data and clustered nearly easy to tackle, standardize the data by change it into unit variance, and any circle-like sampling technique can help for regression like naive bayes



Technical Data Scientist / data scientist

3- Define technical problem from business use case explained by our data analyst.

We cannot explain our data is scattered following polynomial pattern.



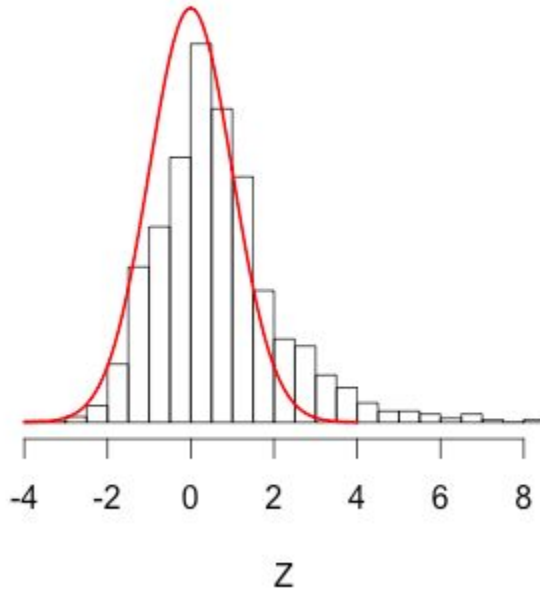
Technical Data Scientist / data scientist

4- study structured / unstructured data using data analysis for technical users.

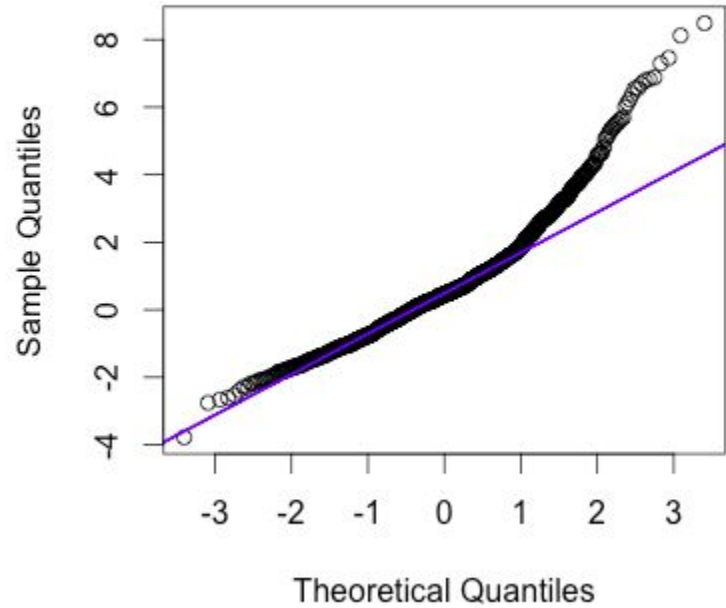
They do data visualization and explained why they use this technique to data analyst.

Why Kernel Density Function? Why plot on using Q-Q plot? Why the data skewed like this? How many outliers effect our data?

Skewed Right



Normal Q-Q Plot

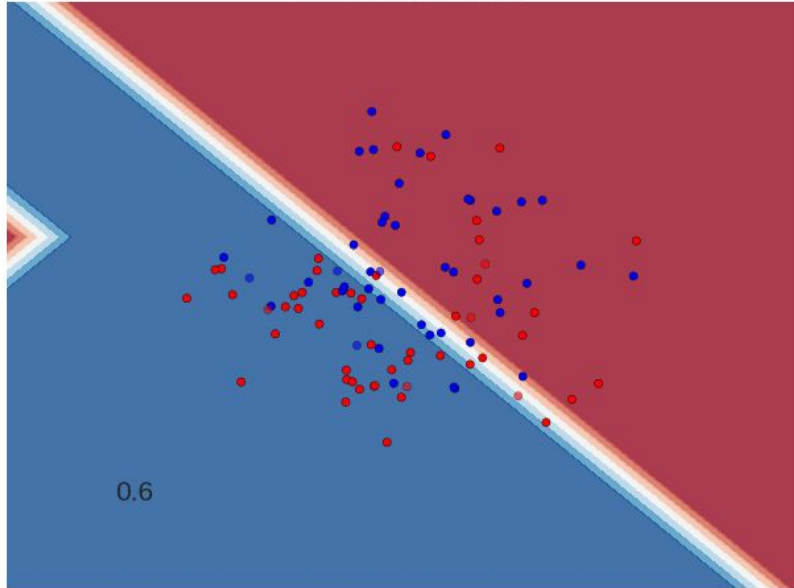


Why plot on using Q-Q plot? Why the data skewed like this? How many outliers effect our data?

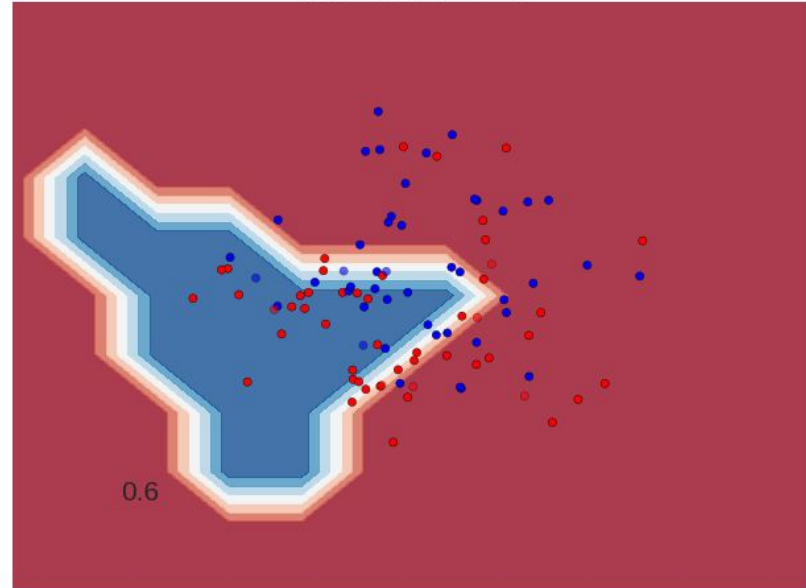
Technical Data Scientist / data scientist

5- understand on how to integrate statistical models with machine learning models

3 layers Neural Network



4 layers Neural Network



Technical Data Scientist / data scientist

6- Code for proposed architecture. SQL, noSQL, PySpark.



So which one you want to be?

Data Analyst or Data Scientist?

Business Data Scientist or Technical Data Scientist?

If business, you do not require to understand a lot of applied math, but handle macro processes.

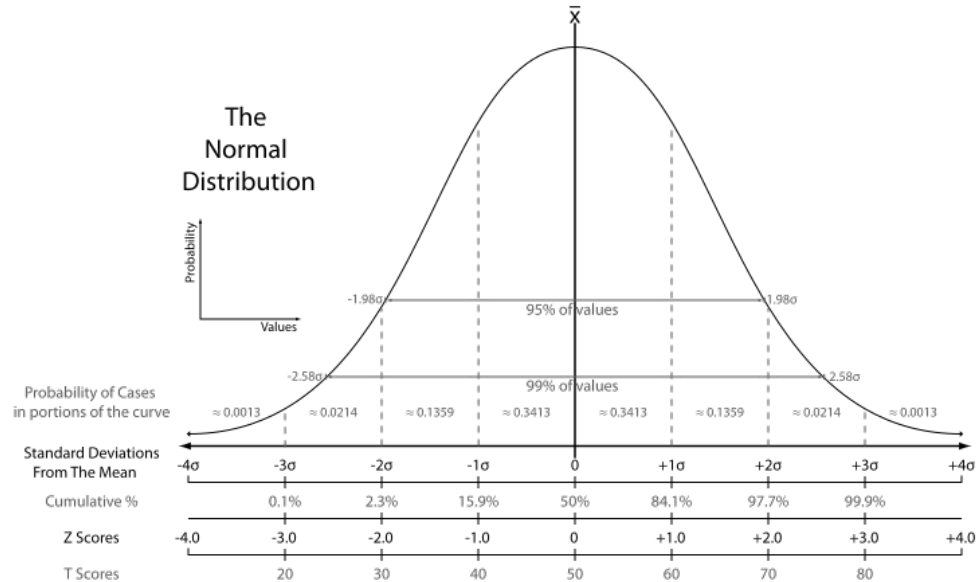
If technical, you do not require to handle macro processes with client.

This one really depends on your forte.

What type of analysis required?

1- Quantitative Analysis

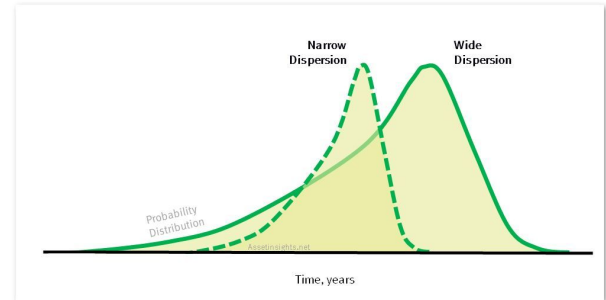
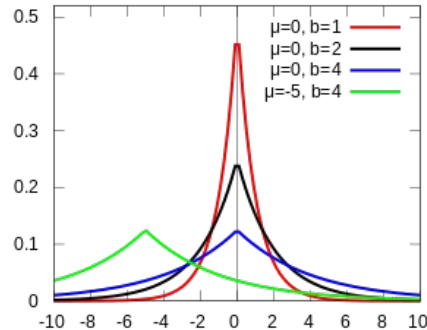
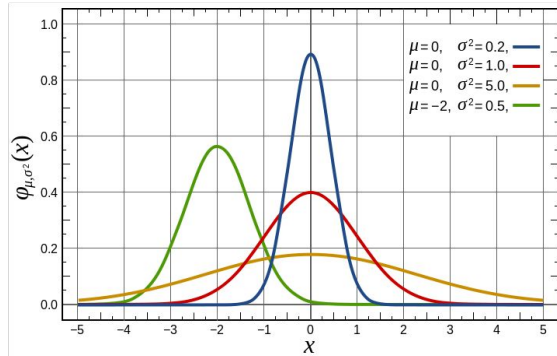
Or short said, Statistics. We study the collection, analysis, interpretation, presentation, and organization of data.



Descriptive statistics are most often concerned with two sets of properties of a distribution (sample or population)

central tendency (or location) seeks to characterize the distribution's central or typical value

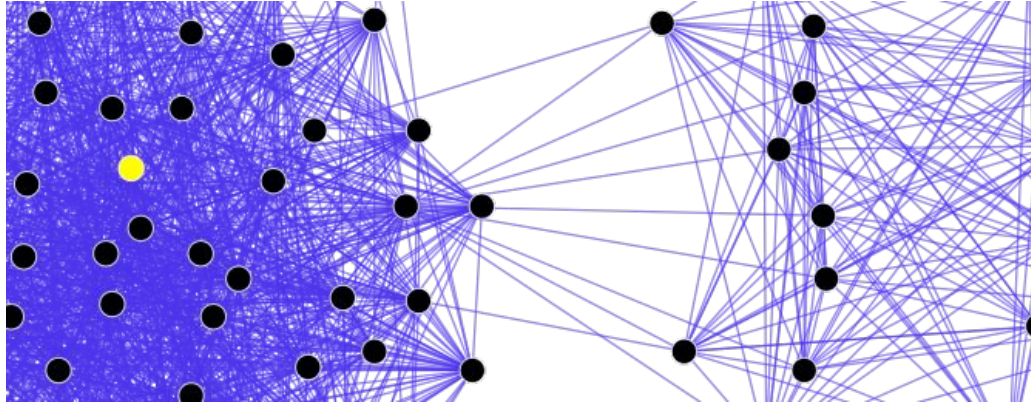
dispersion (or variability) characterizes the extent to which members of the distribution depart from its center and each other.



What type of analysis required?

2- Quantitative Analysis

Qualitative methods examine the why and how of decision making, not just what, where, when, or "who", and have a strong basis in the field of sociology to understand the quantitative analysis



Qualitative methods produce information only on the particular cases studied and any more general conclusions are considered propositions.

Quantitative methods can then be used to seek empirical support for such research hypotheses.



Data Driven vs Business Driven

Data Driven

You collect student feedbacks regarding facilities. What do they want. And the result is, most of the student want vending machine.

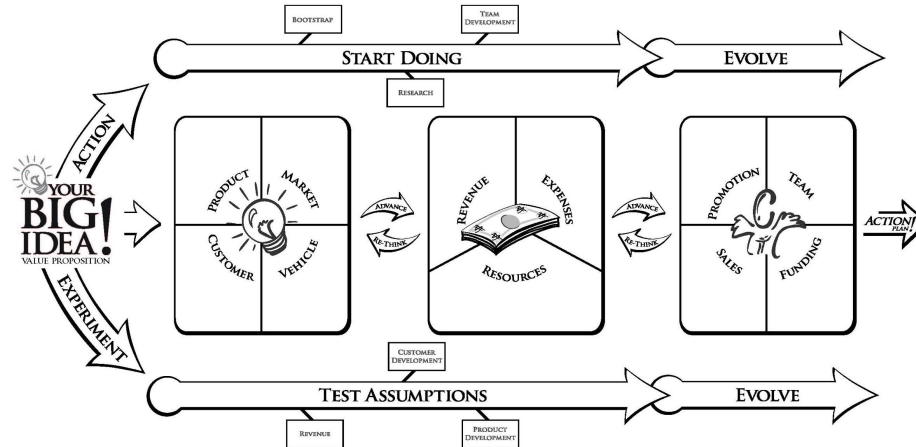
So your college bought 2 vending machine.



Business Driven

You collect student feedbacks regarding facilities. What do they want. And the result is, most of the student want vending machine.

But your upper management doubt to buy the vending machine. Instead they create a new business use case from the feedbacks.



Data Driven vs Business Driven

Data Driven drives / compelled by data after do quantitative analysis, rather than by intuition or by personal experiences.

Business driven drives / compelled by intuition or personal experiences, more measure on qualitative analysis.

Data Driven is cheap since the density of data increased exponentially, and huge samples already can generalized the whole population.

Business driven is expensive because require a long term running of business experiences, played by own judgement and heuristics.

But we cannot trust very much on Data Driven

Here is why,

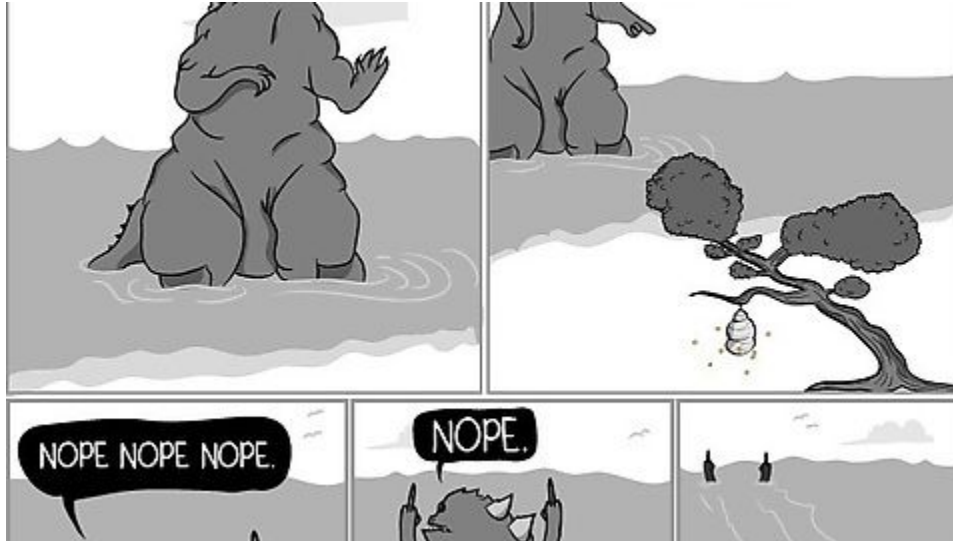
How many of us fill in any feedback truly with their heart?

Not really. We do cincai, other also do cincai. Other samples also do cincai. By combining multiple samples can bring cincai generalization report representing the population.

This is the problem with Data Driven.

But we cannot trust very much on Data Driven

Quantitative Analysis is easy, correlation study is easy, but why our correlation shows that value?



One more important thing,

Data Science is not an architecture
driven!

If you deal with a company, and they proposed what is the physical architecture before business use-case, they are totally no idea.

So how to be a Data Scientist?

If you are asking
questions and using data
to find answers,

**YOU ARE A
DATA
SCIENTIST.**



But still,

You still need to learn on how to code to manipulate directly the data using preferred language like Python or R.

So let's get our hand dirty on
some coding!