THE ULTIMATE GUIDE ON BIG DATA ANALYTICS & DATA SCIENCE

upx

Move up in life

Table of Contents

1

What are Big Data Analytics & Data Science?

A. What is Big Data?

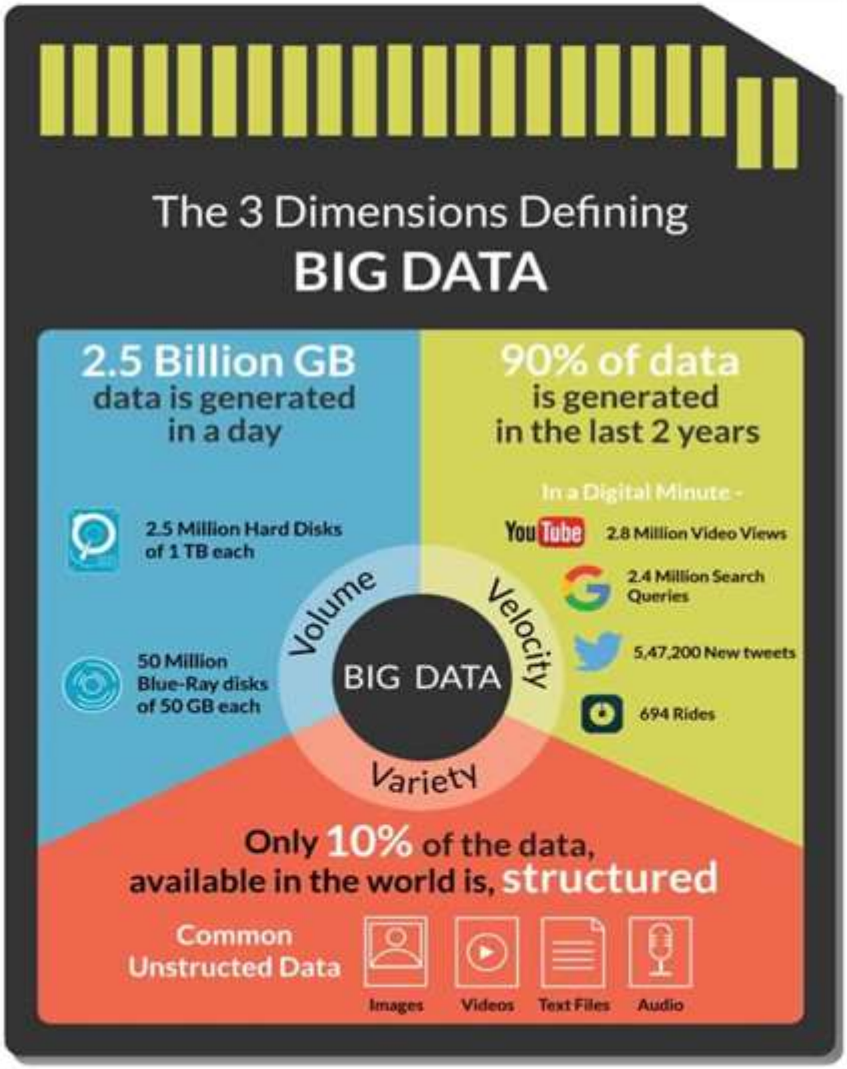
You must have heard the term "Big Data" a lot. It is indeed gaining a lot of importance these days. Some are of the opinion that it is the modern age oil! So then, what is Big Data all about?

Big data refers to the large amount of data generated by web-logs, text, videos, content & images — mainly created by online activity that demands modern and sophisticated systems for storage.

Here are 4 Big Data Essentials that you probably didn't know!

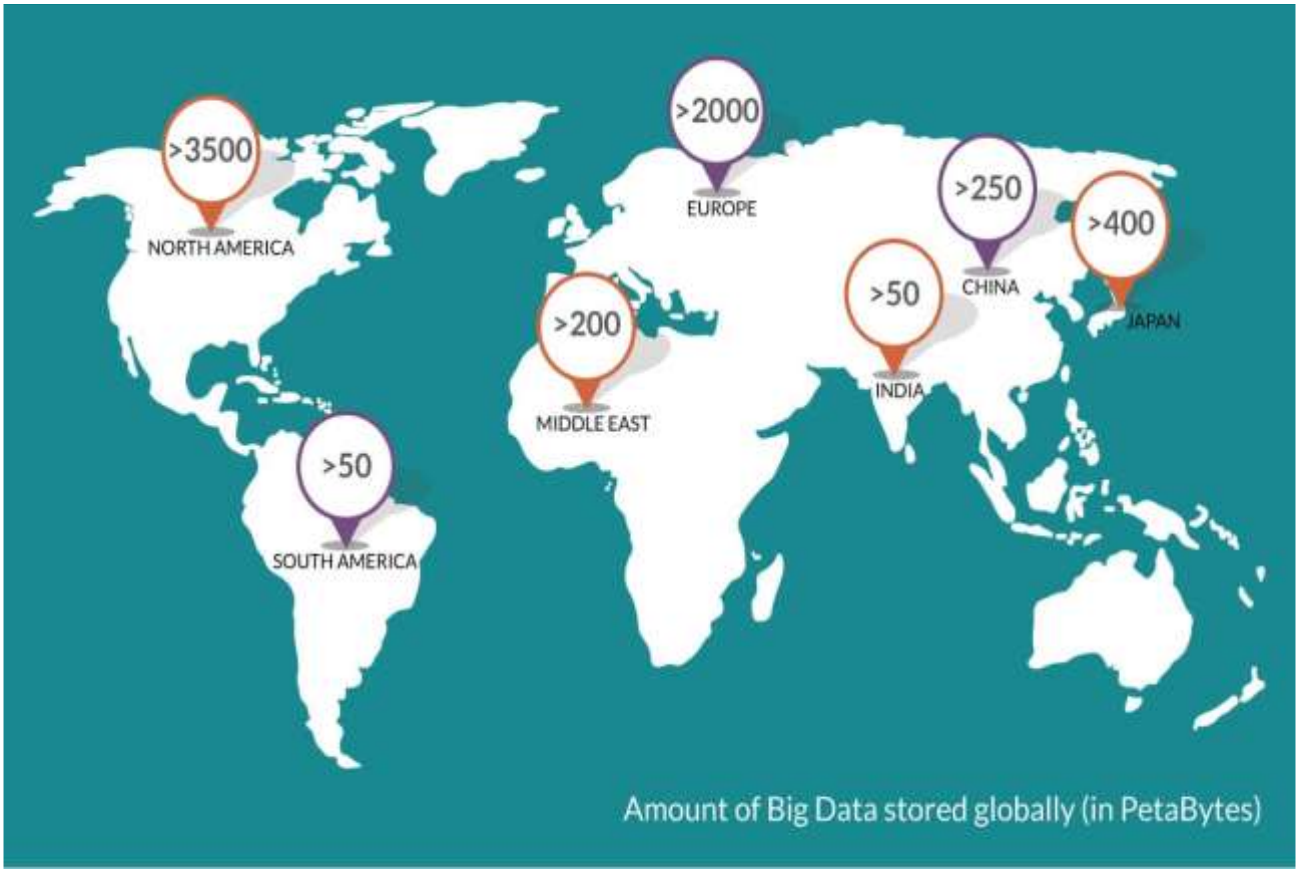
Big Data Essentials #1- Characteristics

When we talk about Big Data we don't necessarily mean the size of the data. Dough Laney defines Big Data on the basis of 3Vs, viz., **Volume**, **Variety**, and **Velocity**.



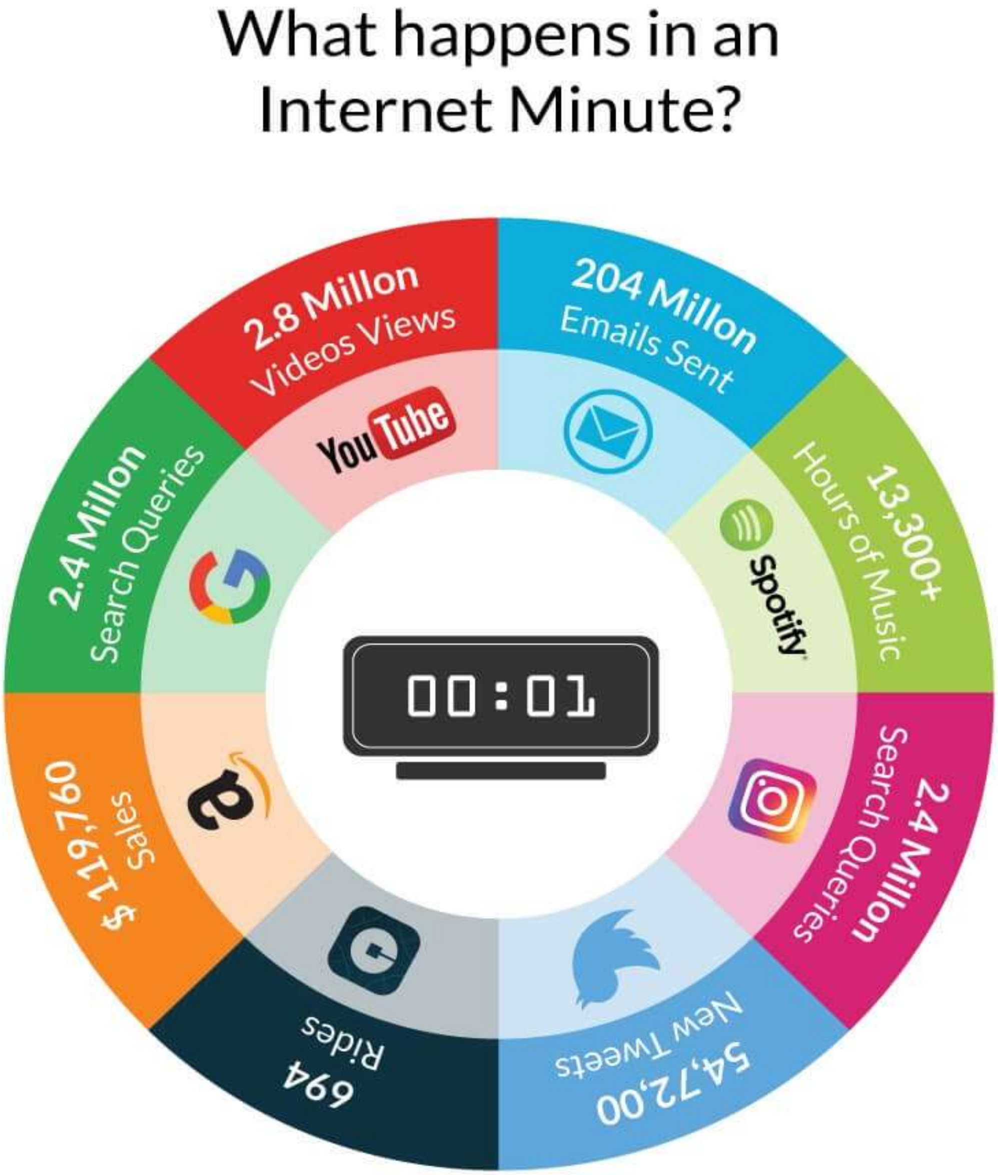
i. Volume:

The volume of data is increasing daily. As per IBM, 2.5 Exabyte of data is generated every day. By 2020, the total data will add up to 40,000 Exabyte! A single storage server cannot store such vast amounts of data. Hence, a need for network of storage devices called SANs (i.e. Storage Area Networks) arises. Companies find it harder to afford the cost of these storage servers.



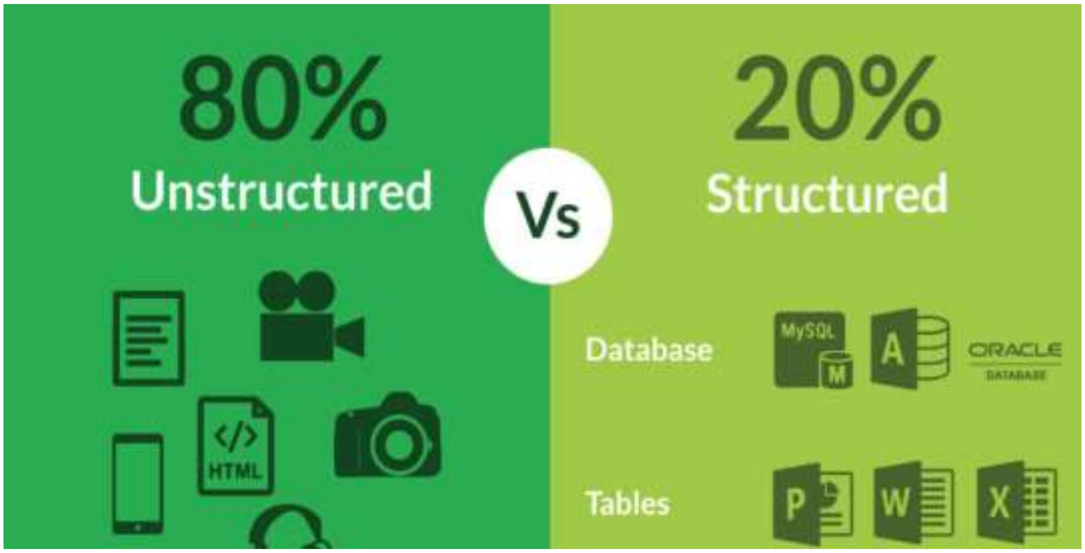
ii. Velocity:

Velocity is the speed at which data is generated and the promptness at which it needs to be processed. Some researchers believe that 90% of the world's data was generated in the last two years alone. Big Data poses a huge challenge for social networking sites. For example, Facebook needs to store petabytes of data generated by its 1.65 billion active monthly users. Such streaming data needs to be stored and queries need to be processed in real-time.



iii. Variety:

Traditional data types only include structured data, which perfectly fits in the case of an RDBMS (i.e. Relational Database Management). But most of the data we generate is unstructured. The digital world has opened up its doors to unstructured data making RDBMS no longer viable. In fact, Facebook alone generates 30+ petabytes of unstructured data in the form of web logs, pictures & messages. Almost 80% of the data today is unstructured and cannot be classified into tables. With the aid of Big Data technologies, it is now possible to consolidate this data and make sense of it.

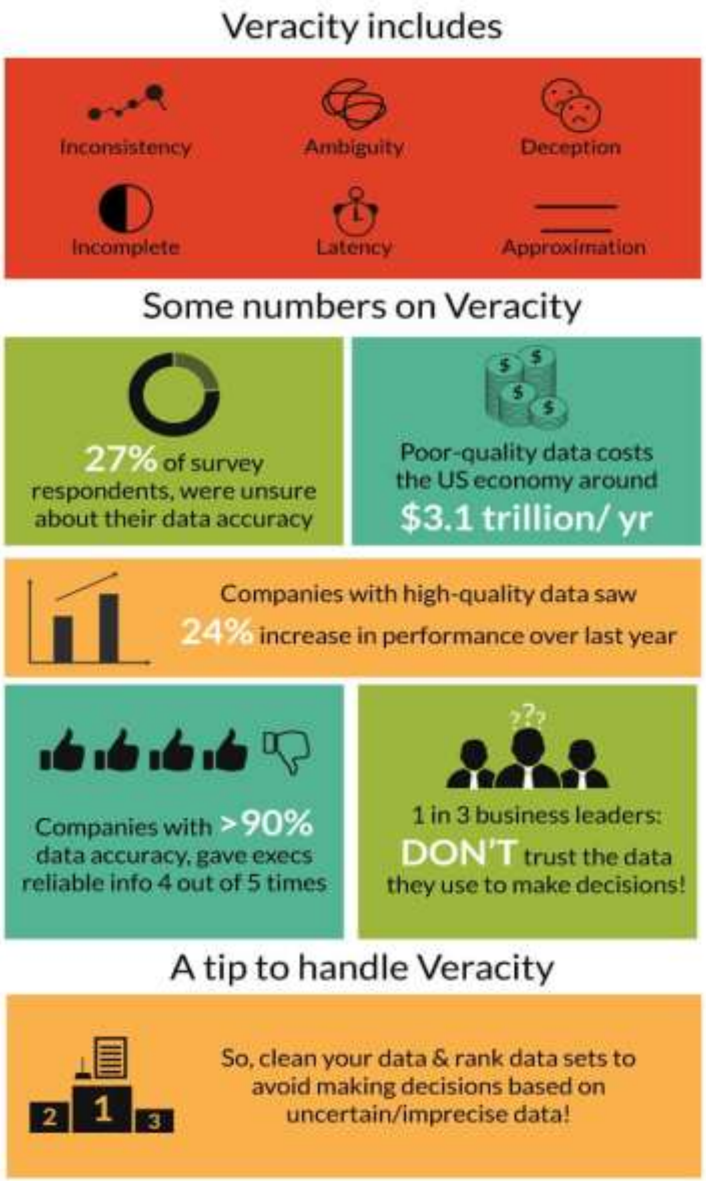


However, Big Data has been further classified to include two more Vs i.e. Veracity and Value.

iv. Veracity:

Veracity means the biases, noises and abnormality in data. In other words, how dependable, reliable or certain the data is. Uncertainty may exist due to incompleteness and inconsistency, ambiguity, latency, deception, model approximation, etc. With the amount of data surging, the accuracy of the data surely takes a toll. This then leads to the big question, “Can I trust this data and the insights it provides? A lot of “dirty data” may exist in the system. Therefore, it is recommended to use clean data while formulating the Big Data Strategy.

Tip- Assign a Data Veracity score by ranking specific data sets. Do this to avoid making decision based on analysis of uncertain and imprecise data.



Veracity includes

Some numbers on Veracity

27% of survey respondents, were unsure about their data accuracy

Poor-quality data costs the US economy around $3.1 trillion/ yr

Companies with high-quality data saw 24% increase in performance over last year

Companies with >90% data accuracy, gave execs reliable info 4 out of 5 times

1 in 3 business leaders: DON'T trust the data they use to make decisions!

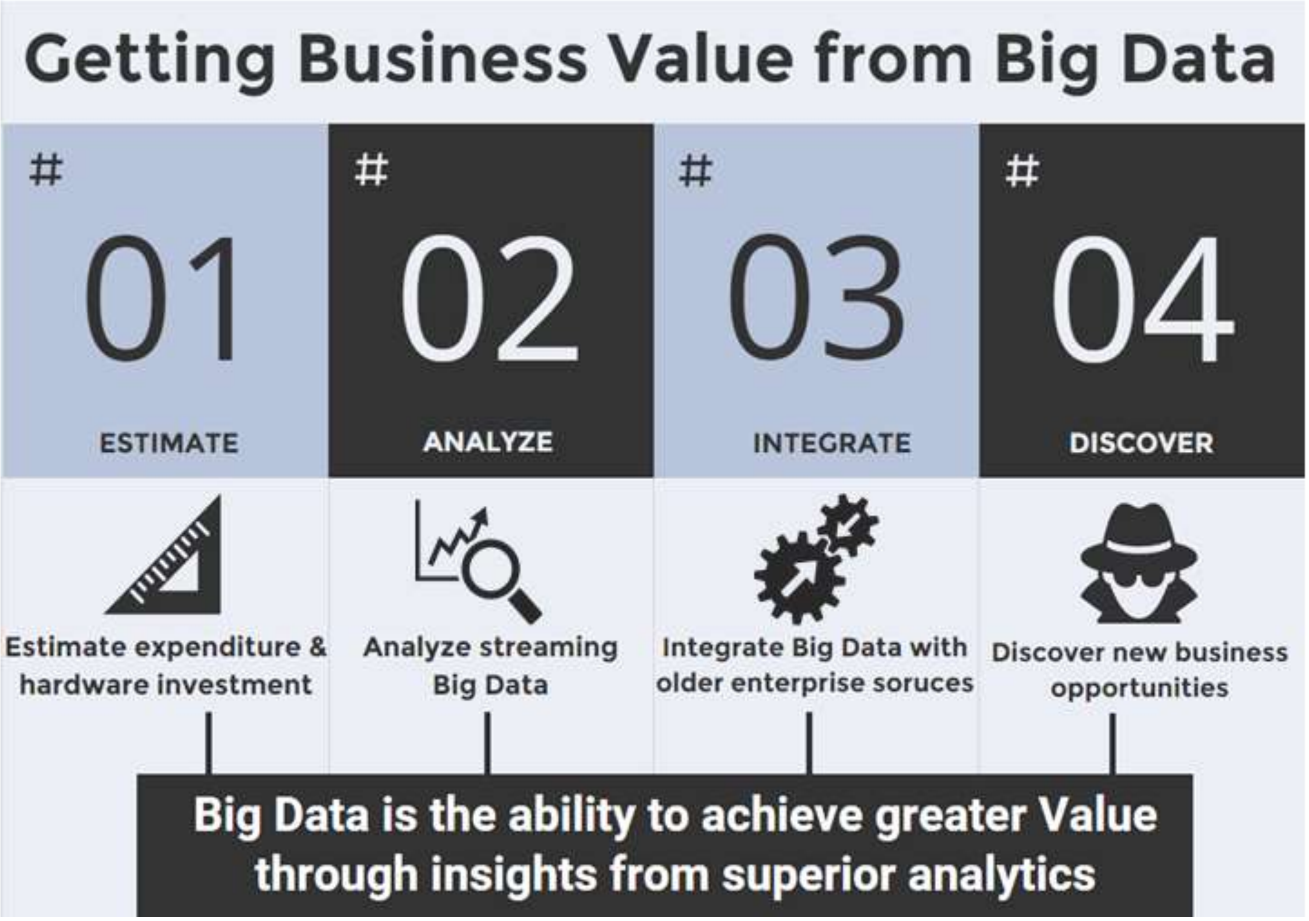
A tip to handle Veracity

So, clean your data & rank data sets to avoid making decisions based on uncertain/imprecise data!

v. Value:

From the business perspective, this V is the most vital element in Big Data. The main aim of delving into an ocean of data is to bring some value out of it.

This requires financial and hardware investments in infrastructure & resources to handle and analyze Big Data. As a result, it’s crucial to do a cost-benefit analysis before investing in Big Data projects.



Getting Business Value from Big Data

01 ESTIMATE

Estimate expenditure & hardware investment

02 ANALYZE

Analyze streaming Big Data

03 INTEGRATE

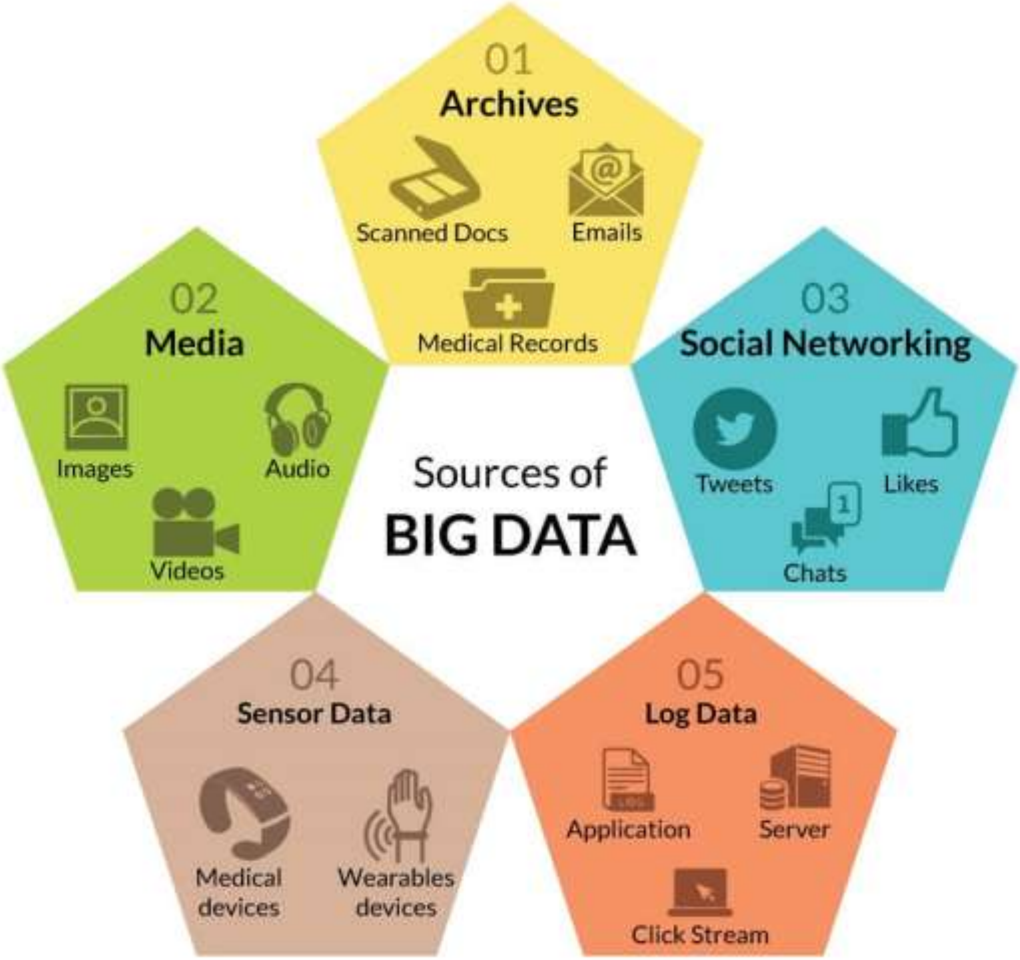
Integrate Big Data with older enterprise soruces

04 DISCOVER

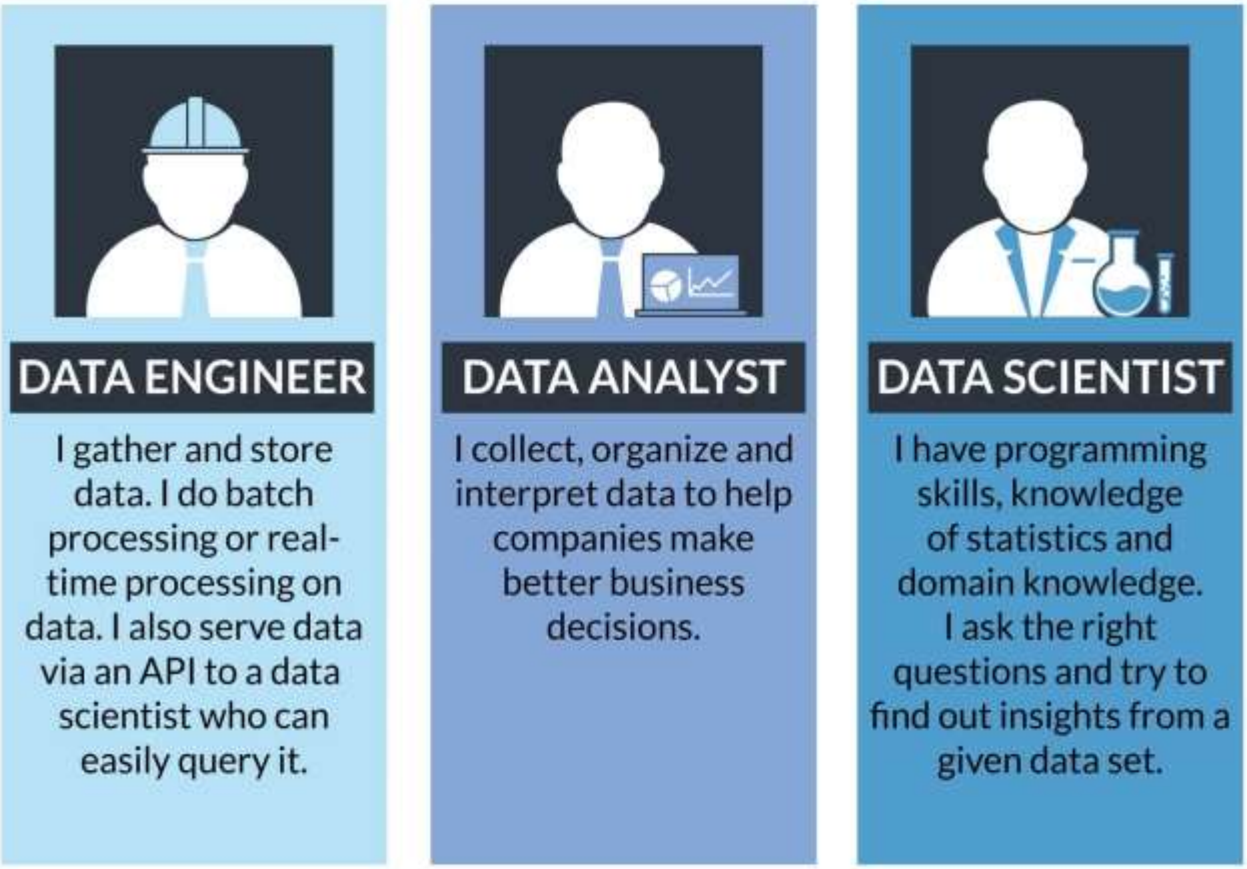
Discover new business opportunities

Big Data is the ability to achieve greater Value through insights from superior analytics

Big Data Essentials #2- Sources



Big Data Essentials #3- Careers



DATA ENGINEER

I gather and store data. I do batch processing or real-time processing on data. I also serve data via an API to a data scientist who can easily query it.

DATA ANALYST

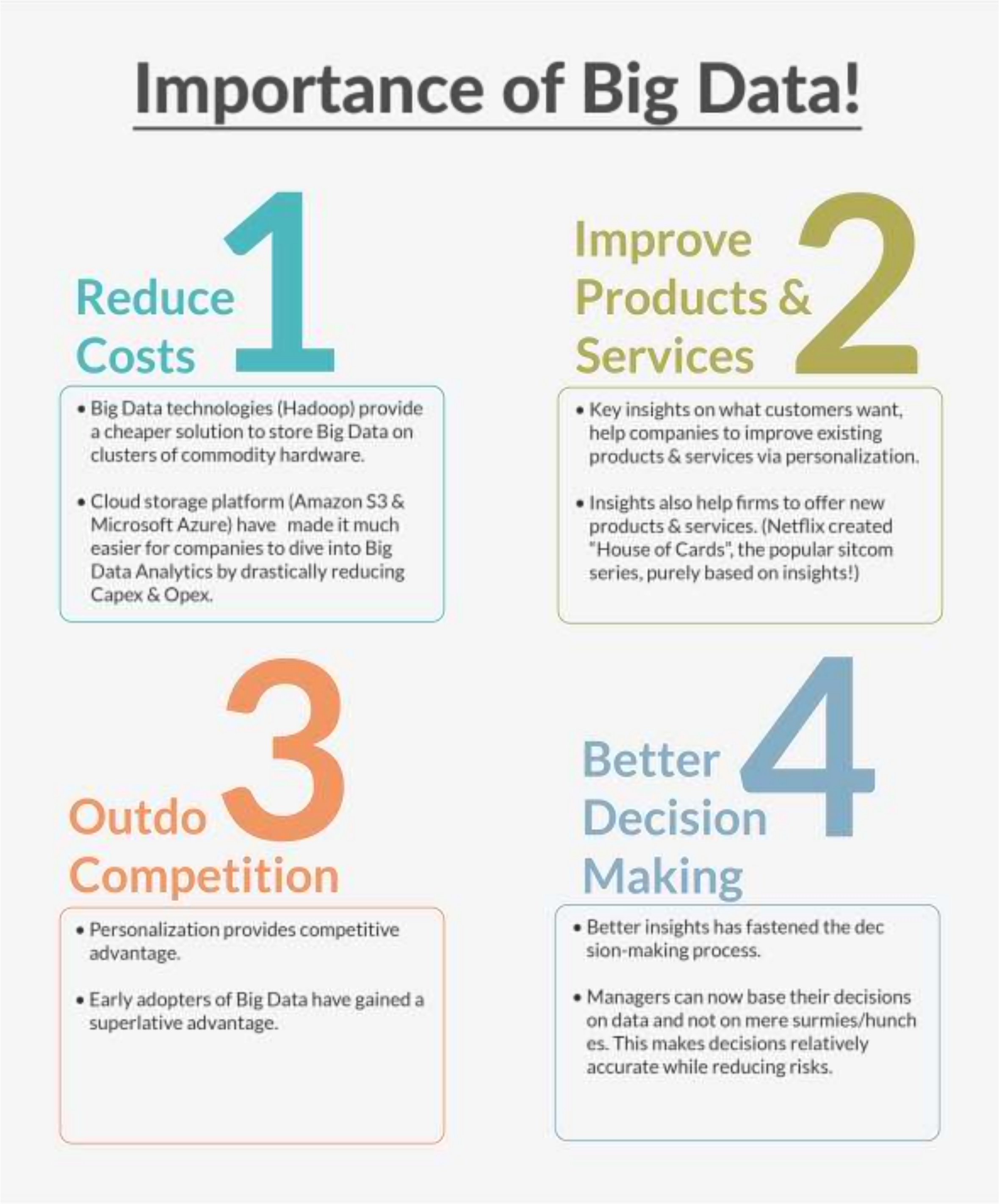
I collect, organize and interpret data to help companies make better business decisions.

DATA SCIENTIST

I have programming skills, knowledge of statistics and domain knowledge. I ask the right questions and try to find out insights from a given data set.

Big Data Essentials #4- Importance

So, here's why Big Data is a significant addition in every sector-



Importance of Big Data!

1 Reduce Costs

- Big Data technologies (Hadoop) provide a cheaper solution to store Big Data on clusters of commodity hardware.
- Cloud storage platform (Amazon S3 & Microsoft Azure) have made it much easier for companies to dive into Big Data Analytics by drastically reducing Capex & Opex.

2 Improve Products & Services

- Key insights on what customers want, help companies to improve existing products & services via personalization.
- Insights also help firms to offer new products & services. (Netflix created "House of Cards", the popular sitcom series, purely based on insights!)

3 Outdo Competition

- Personalization provides competitive advantage.
- Early adopters of Big Data have gained a superlative advantage.

4 Better Decision Making

- Better insights has fastened the dec sion-making process.
- Managers can now base their decisions on data and not on mere surmies/hunches. This makes decisions relatively accurate while reducing risks.

Some of the companies using Big Data

Now you finally know about the 4 Big Data Essentials!

Want to know how Adobe, Twitter, Google, Netflix, Facebook, etc. make smarter decisions every day based on Big Data? Then read our article on Retail, Sports & Media. Big Data is everywhere! and learn about the different use cases of Big Data in modern industries.

Conclusion:

Surely, the amount of data generated and stored every day across the global level is incredible. More so, this phenomenon is only expected to continue. At the same time, organizations having futuristic thinking are quickly evolving to include Big Data. They are thus hiring skilled professionals to interpret data and aggressively build their organization's Big Data capabilities.

B. Big Data Analytics & Data Science – Are they the same?

"Big Data Analytics, Data Science and Big Data classes soon! Sign up to learn the next big technologies! "Does this ad seem to be familiar to you? Whether you're a techie or not, don't you come across these terms quite often? On the surface, Big Data Analytics and Data Science seem to be one and the same as they are used interchangeably quite often. A quick Google result probably didn't help, it may have just confused you further.

This blog covers what these technologies are, how different they are, and how much they overlap.

What is Analytics?

Big Data cannot be converted into an asset unless it is analyzed and insights are mined from it. This is where Big Data Analytics comes into the picture.

The process of mining useful information (i.e. relevant and useful insights from raw data) from the plethora of data being generated to make smart business decisions, is Big Data Analytics. (This is how the word "information" differs from the word "data"- other pair of words that are used interchangeably.)

Analytics is a process of discovery, interpretation, and communicating meaningful patterns in data. It denotes a persons' skill to gather and use data to generate insights that lead to fact-based decision making.

Data-driven analytics provides us with unparalleled opportunities that will help to transform the vast areas concerning business, healthcare, government, etc. The application of data-driven analytics is especially valuable in areas rich with recorded information. Analytics banks on the simultaneous application of statistics, computer programming, and operation research to measure performance.

It is observed that analytics most likely favours data visualization while communicating insight.

Analytics also supports the organizations to use the generated business data. It helps the organizations to describe, predict, and enhance their business performance.

Analytics way before the Computers' age

The analysis of data led to knowledge discovery for hundreds of years now. So, starting from the data collection project by the Swedish government in 1749 to Florence Nightingale recording. And analyzing mortality data in the 1850s, to British scholar Richard Doll's tobacco and lung cancer study in the 1950s.

Each of these systems has empowered the field by responding to questions, the answers to which were unknown to us. The Swedes sought answers to the geographical distribution of their population to learn the most efficient way to sustain an appropriate military force.

Nightingale wanted to know the role that hygiene and nursing care played in mortality rates. Whereas Doll wished to know if people who smoked had more chances of suffering from lung cancer.

What does Analytics answer?

Analysis of data can reveal correlations and patterns. With the data analytics in the picture, there becomes a less need for the people to rely on hunches and intuitions. Also, the analysis of data can help us answering the following questions:

- Descriptive: What has happened?
- Diagnostic: Why did it happen?
- Predictive: What is likely to happen?
- Prescriptive: Is there anything I can do about it?

Analysis and Analytics – the same concept?

Data Analysis is a broad spectrum that includes Analysis of all kinds, on data sets of all sizes. At a basic level, working with functions and formatting data in Microsoft Excel is an example of Analysis.

Excel was the tool largely used by businesses for a long time. But as the volume of data grew, Excel couldn't be relied upon as a does it all tool. Analysis tools had to be scaled to fit **"bigger"** data as well. Therefore, new tools had to be developed to deal with Big Data. This led to the birth of Hadoop.
Analysis largely deals with analysing past data and understanding the data. **Analytics** deals with using these insights to make smart business decisions in the future.

What is Data Science?

Data Science is the science that uses smart mathematical and statistical models to mine information from data. It is a multidisciplinary field that involves Statistics, Programming and Domain Knowledge. Data Science uses a host of smart Machine Learning algorithms to make smart and informed decisions about data.

Michael E. Driscoll, the CEO of Metamarkets, said "**Data scientists: better statisticians than most programmers & better programmers than most statisticians**".

This basically sums the entire field up!

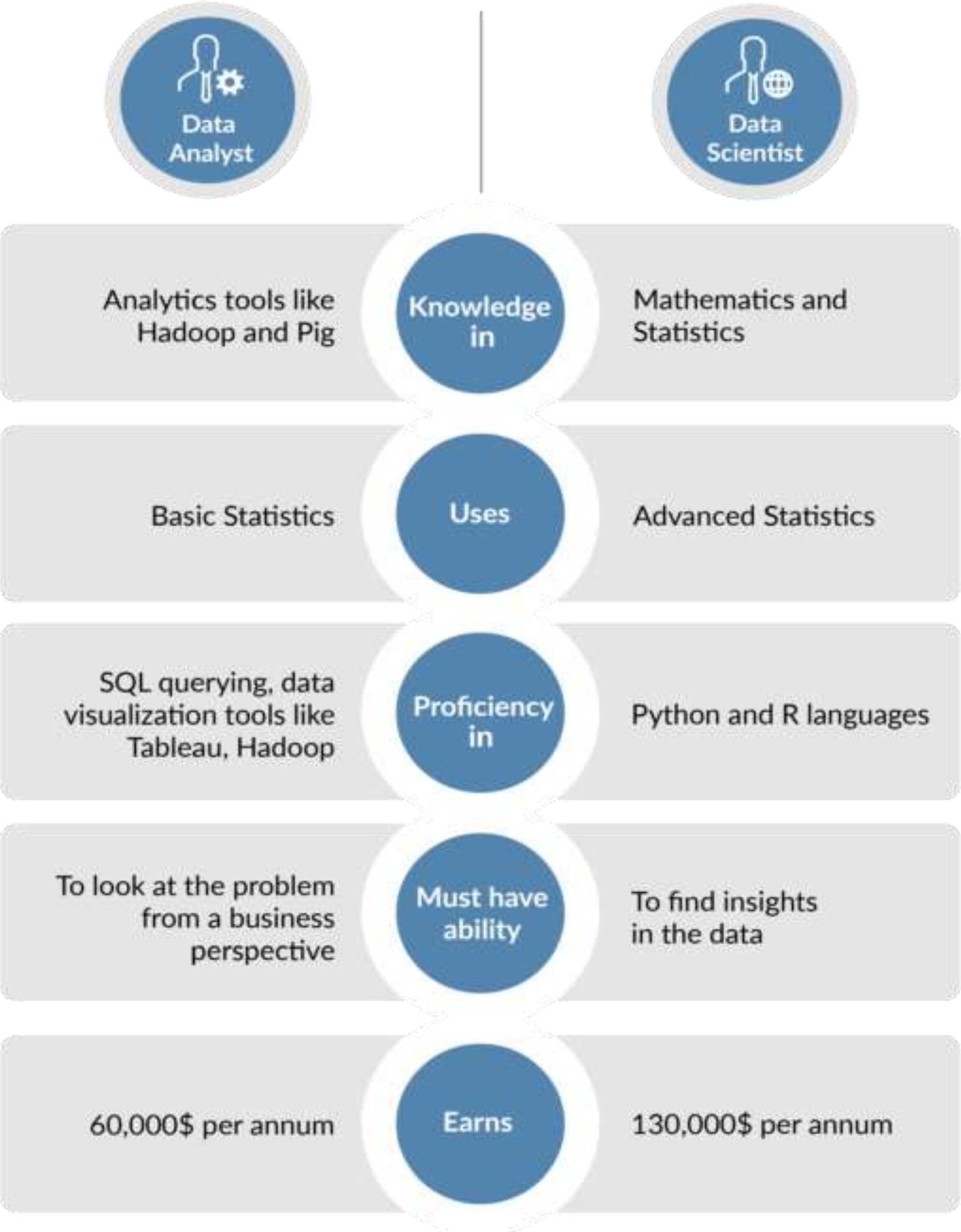
Big Data Analytics & Data Science – Is there a winner?

Broadly speaking, Big Data Analytics can be called Data Science, but Data Science cannot be called Big Data Analytics. On the surface, they perform the same operation – i.e. mining useful information from data. So, the two fields overlap significantly and often work hand in hand. Big Data Analytics involves mining useful information from raw data. Data Science uses Machine Learning to train the computer to learn without being explicitly programmed, to make future predictions. Machine Learning is what makes Data Science different from Analytics. The Machine Learning Algorithms used are – Decision Tree Learning, Artificial Neural Networks, Deep Learning, Clustering, Random Forest Classifiers, Naïve Bayes, Regression and more. The use of Machine Learning makes computers even smarter – which is one reason why Data Science is so sought after these days.

Do Big Data Analysts and Data Scientists differ?

Data Science has entered the realm of Big Data recently. Big Data Analytics has been around for a little over twenty years. Experts are still trying to develop a clear definition about the differences between Big Data Analytics and Data Science. However, as we delve deeper, we notice discernible differences. Here are some of them-

How do Big Data Analysts and Data Scientists differ?



Data Analyst		Data Scientist
Analytics tools like Hadoop and Pig	Knowledge in	Mathematics and Statistics
Basic Statistics	Uses	Advanced Statistics
SQL querying, data visualization tools like Tableau, Hadoop	Proficiency in	Python and R languages
To look at the problem from a business perspective	Must have ability	To find insights in the data
60,000$ per annum	Earns	130,000$ per annum

Are there any stark similarities between the two fields?

There seem to be some significant differences between the two fields. Does this mean that they don't coincide at all? No. There are quite a few similarities –

- Develop useful insights from raw Data
- Work on Big Data
- Attempt to use the insights achieved to make smart business decisions
- Applied in similar fields, like healthcare, finance, social media and sports.

Did you know? The Harvard Business Review named Data Scientist as the sexiest job of the 21st century!

Can you become a Data scientist or an Analyst?

Are you fed up with your current job and want to shift to the hottest new profession on the block? Here are some of the requirements the fields demand –

Data Scientist:

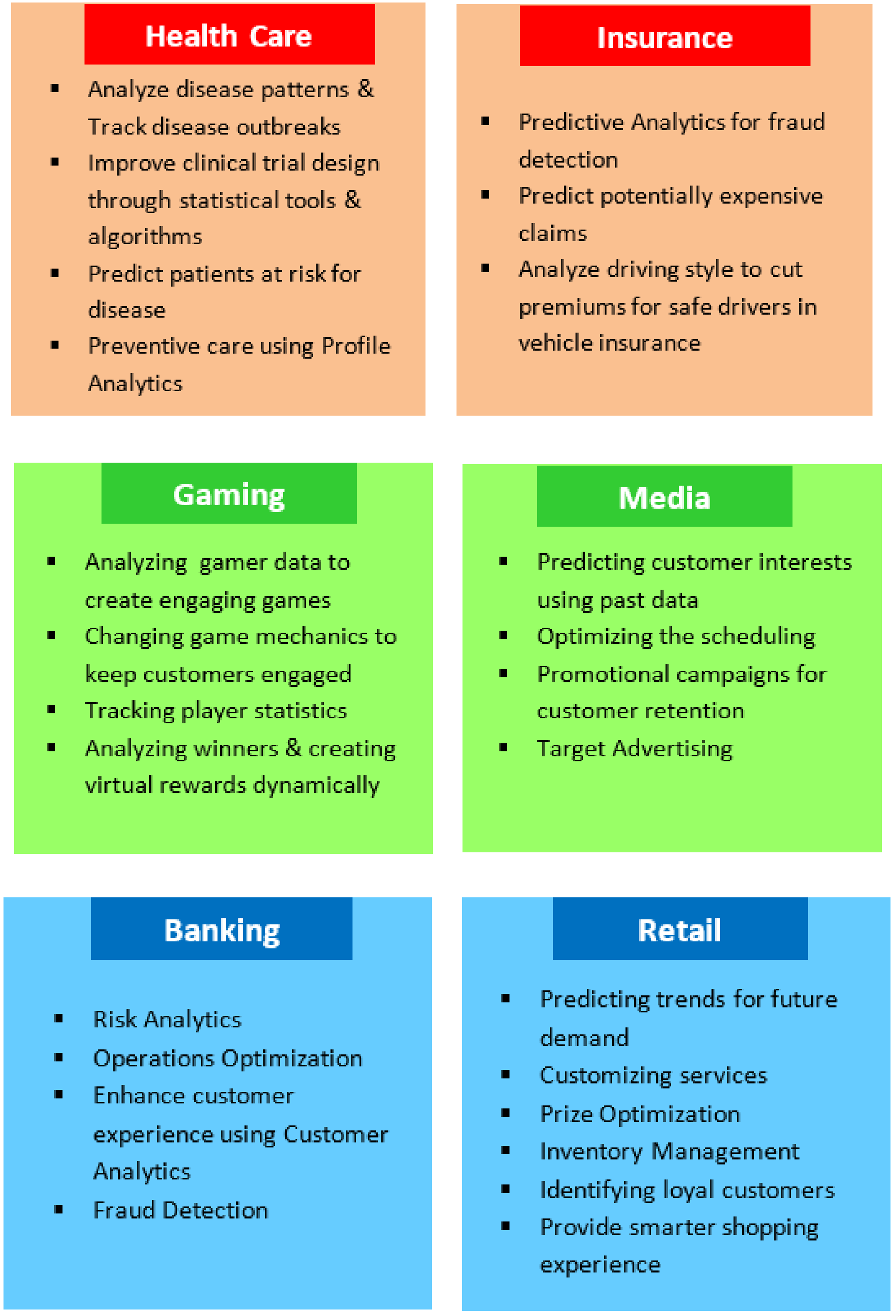
- Good Statistical and Mathematical skills
- Good programming skills (Python, R, Java)
- The ability to ask the right questions (given a data set)
- Knowledge in Machine Learning
- A fast learner

Data Analyst:

- Good Analytical skills
- A keen business mind
- The ability to analyse the results after Analytics tools are applied to data
- Knowledge in dealing with Analytics tools like Hadoop and Hive
- A fast learner

If you think that, you have (or can acquire) these abilities, then these fields are for you

Where are Big Data Analytics and Data Science used?



Health Care

- Analyze disease patterns & Track disease outbreaks
- Improve clinical trial design through statistical tools & algorithms
- Predict patients at risk for disease
- Preventive care using Profile Analytics

Insurance

- Predictive Analytics for fraud detection
- Predict potentially expensive claims
- Analyze driving style to cut premiums for safe drivers in vehicle insurance

Gaming

- Analyzing gamer data to create engaging games
- Changing game mechanics to keep customers engaged
- Tracking player statistics
- Analyzing winners & creating virtual rewards dynamically

Media

- Predicting customer interests using past data
- Optimizing the scheduling
- Promotional campaigns for customer retention
- Target Advertising

Banking

- Risk Analytics
- Operations Optimization
- Enhance customer experience using Customer Analytics
- Fraud Detection

Retail

- Predicting trends for future demand
- Customizing services
- Prize Optimization
- Inventory Management
- Identifying loyal customers
- Provide smarter shopping experience

Conclusion

While there is a major overlap between Data Science and Data Analytics and essentially, as essentially they perform the same operation, the two fields have some concrete differences.

Either way, they are making our lives easier every day, usually without us even realizing it! But, if data amazes you and you're smitten with this Big Data revolution, then hop on to the bandwagon and join us! Read more articles on Big Data Analytics, Data Science, Machine Learning and much more!

Analytics or Data Science? We'd love to know your thoughts on the two and whether you've ever used these words interchangeably.

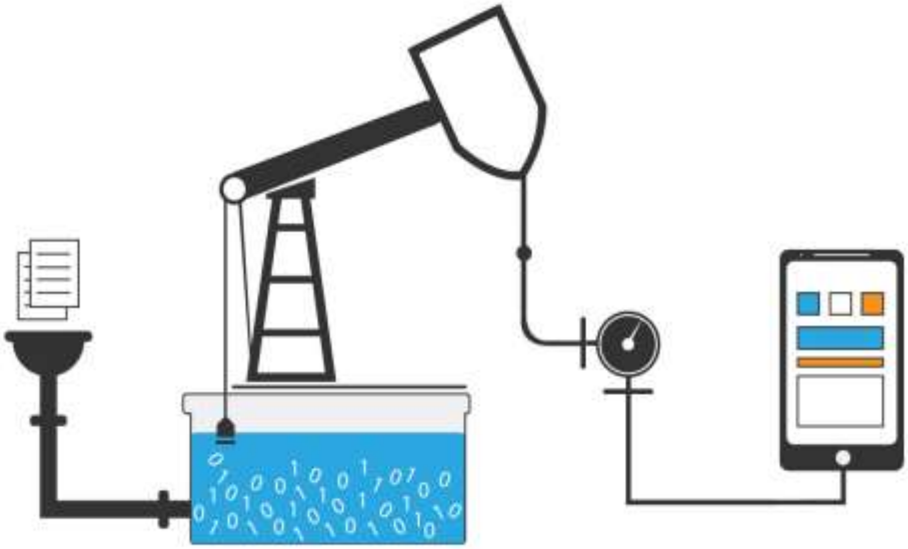
C. Introduction to Data Science

Today, all sectors have integrated technology into their day-to-day business. Technology brings along with it growing data sets. Hence, processing and analyzing such data sets is becoming more and more vital. But there is a key question that needs answering. Who would synthesize and give meaning to such scattered data existing in various forms? Enter data science.

Harvard Business Review classifies Data Scientist to be “The Sexiest Job of the 21st Century”. A McKinsey report forecasts the demand for data scientists to multiply so rapidly such that it would outgrow the supply by 50%. Surely, in the upcoming years, this would be one of the sought-after fields wherein job aspirants would opt to build their career in.

(**Did you know?** *Big companies such as* ***Amazon***, ***L’Oreal***, ***Viacom 18***, ***British American Tobacco****, and many others are hiring Data Scientists!)*

Data- The Modern Age Oil

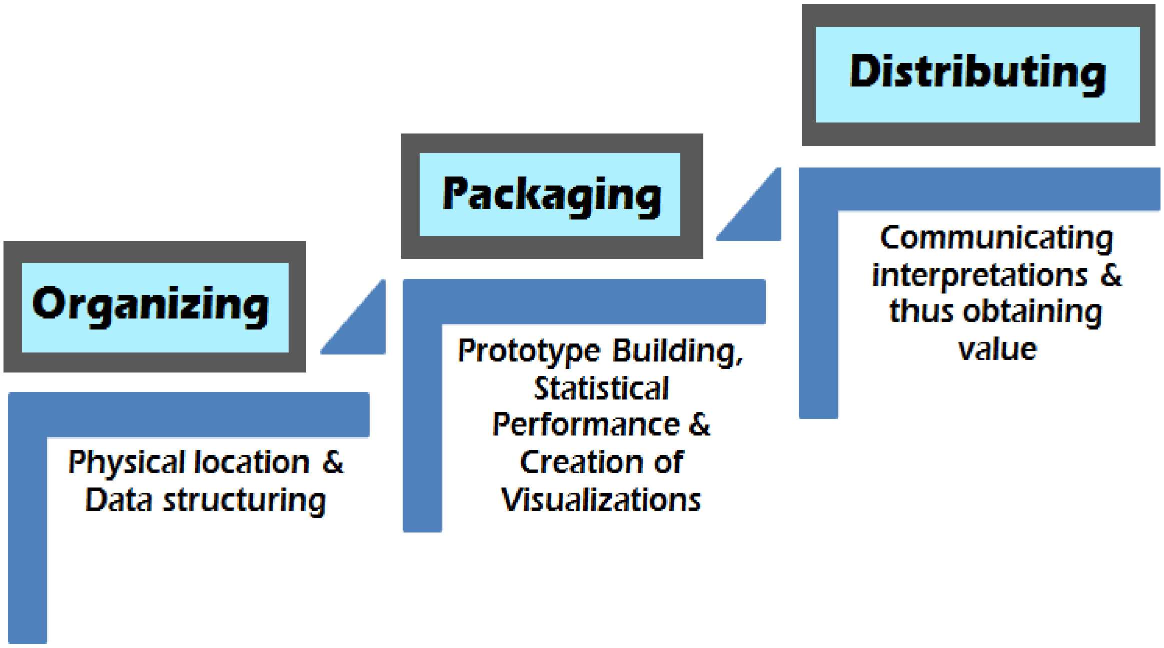


With its dramatic ups and lows, crude oil prices may witness fluctuations; thus, taking a toll on the oil producing nations. But, data remains to be priceless as it will form the crux of decision–making in future.

At the same time, an analogy can be drawn between data and oil. Certainly, data and crude oil are precious resources. But, we cannot derive value from both unless they are processed and refined. As miners extract crude oil, in the same way, Data engineers extract data and Data Scientists refine data. However, the supply chain of data is not as complex and cumbersome as that of oil. After the oil is extracted, it needs to be then transported to tankers. Which then is routed through pipelines to be finally stored in storehouses? With the tremendous rise of Cloud Computing services, transportation and storage of data have never been easier!

What is Data Science?

Data Science is nothing but a science of making sense of data. To explain, it involves the usage of automated methods to analyze data and extract information or insights to find the unknown. It creates data products which help in decision making. This further aids in driving business value and building confidence.



Organizing

Physical location & Data structuring

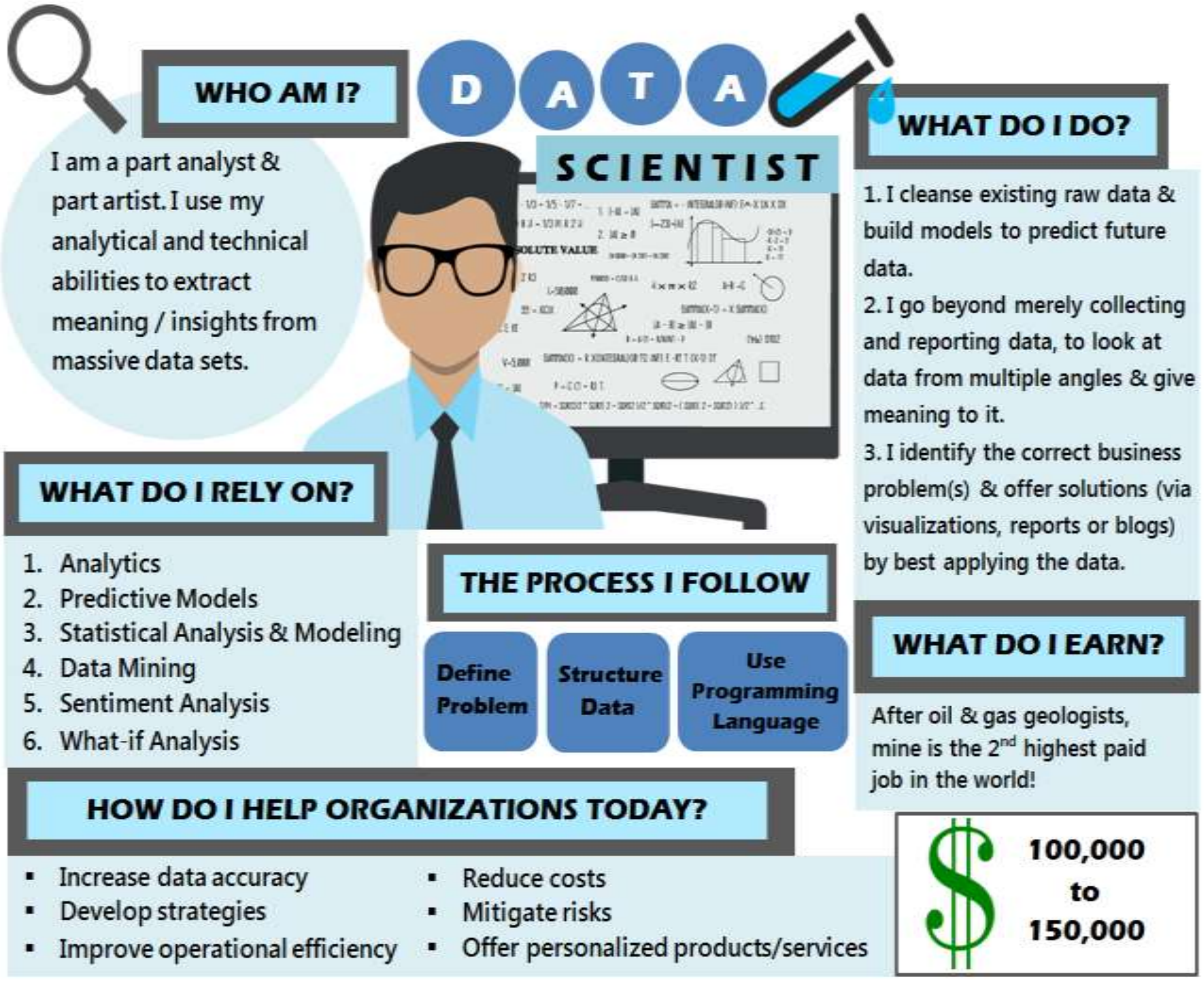
Packaging

Prototype Building, Statistical Performance & Creation of Visualizations

Distributing

Communicating interpretations & thus obtaining value

Data Scientist- At a glance



DATA SCIENTIST

WHO AM I?

I am a part analyst & part artist. I use my analytical and technical abilities to extract meaning / insights from massive data sets.

WHAT DO I DO?

1. I cleanse existing raw data & build models to predict future data.
2. I go beyond merely collecting and reporting data, to look at data from multiple angles & give meaning to it.
3. I identify the correct business problem(s) & offer solutions (via visualizations, reports or blogs) by best applying the data.

WHAT DO I RELY ON?

1. Analytics
2. Predictive Models
3. Statistical Analysis & Modeling
4. Data Mining
5. Sentiment Analysis
6. What-if Analysis

THE PROCESS I FOLLOW

Define Problem | Structure Data | Use Programming Language

WHAT DO I EARN?

After oil & gas geologists, mine is the 2nd highest paid job in the world!

$ 100,000 to 150,000

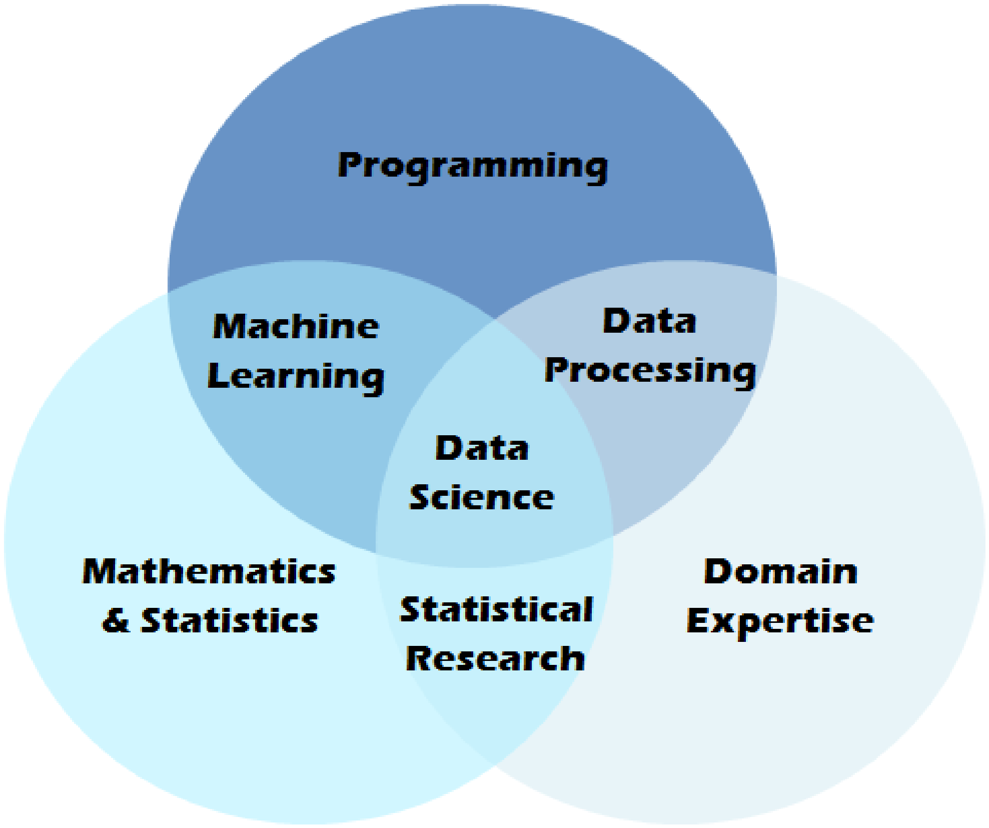
HOW DO I HELP ORGANIZATIONS TODAY?

- Increase data accuracy
- Develop strategies
- Improve operational efficiency
- Reduce costs
- Mitigate risks
- Offer personalized products/services

What skills are needed to be a Data Scientist?

Drew Conway's Venn diagram clearly illustrates the three skill sets, viz., Mathematics & Statistics, Programming & Database, and Domain Expertise, that are required to be a data scientist.

You think you don't possess extensive knowledge or expertise in all the areas? No problem. Data Science is said to be a team sport. Thus, a data scientist need not necessarily be strong in all the mentioned fields. So, you can still build a career in data science if you either have strong analytical or programming skills. This would suffice to make you a valuable player in the data science team!



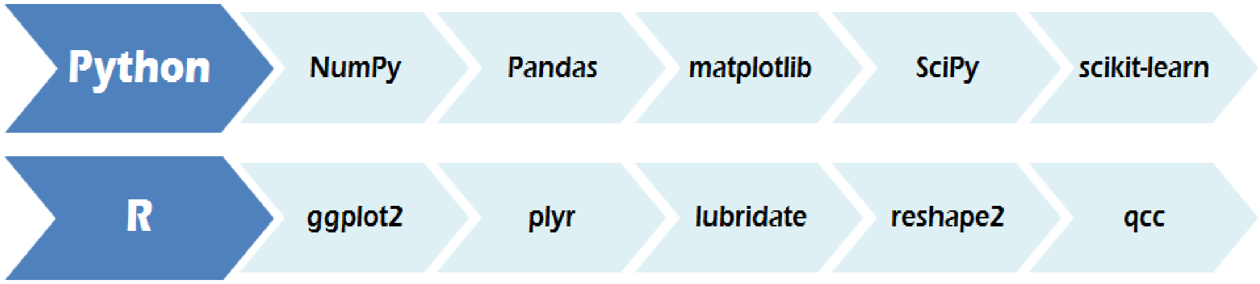
I. Mathematics & Statistics

Understanding of basic mathematics and statistics is crucial for a data scientist. Statistical knowledge helps to interpret and to analyze the data that is collected. Data Science becomes magical as applying brilliant mathematical concepts to the data yields unexpected insights! The basic concepts are Descriptive and Inferential Statistics, Linear Algebra, Graphing, etc.

II. Programming & Databases

This is the area that separates one from being a statistician or an analyst. For any given data, one needs to write programs to query and retrieve data from databases or apply machine learning algorithms. One should have a good grasp on data science libraries and modules.

Python and R provide some prebuilt libraries. Both can be used by simply importing them into programs. Thus, this makes them good programming languages to start with. Although Microsoft Excel is a great tool for processing data, it is only suitable when working with small or medium data sets. But when it comes it to Big Data, Python and R are much better. They also provide greater flexibility and control to the user.



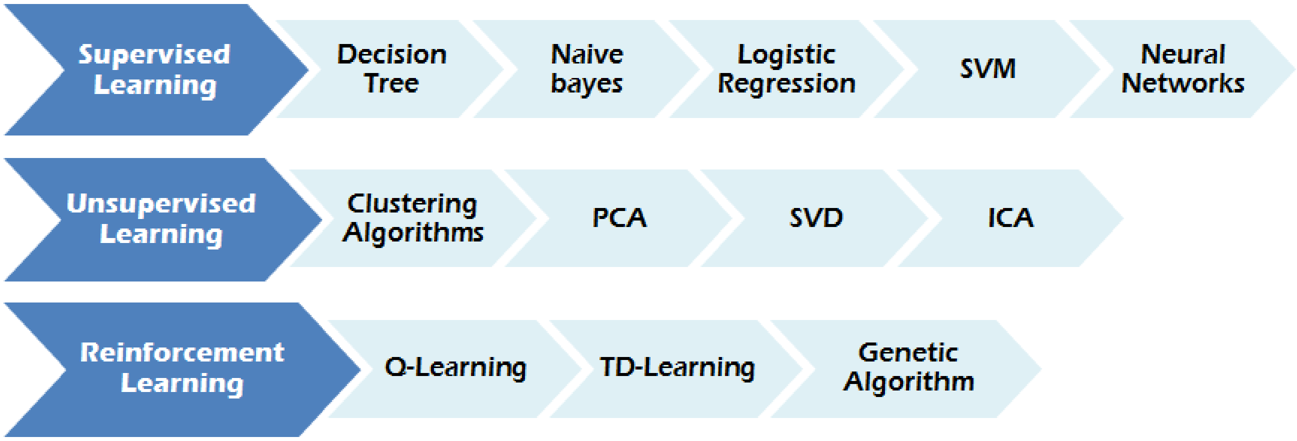
Database Systems act as a central hub to store information. These can be SQL-based or NoSQL-based. Relational Database includes PostgreSQL, MySQL, Oracle, etc. with Hadoop, Spark, and Mongo DB being among the others.

III. Domain Expertise

This involves asking the right questions by filtering the relevant data from the entire data universe. Data Scientists need to interpret the data by understanding its structure. They also need to know the problem(s) they are solving. For instance, if they are solving an online advertising problem, they should understand the type of customers visiting their website, their interaction with the website and the meaning of such data.

IV. Machine Learning

This is an integral part of Data Science, used to create predictive models. Machine Learning algorithms are very powerful, thereby eliminating the need to create a new algorithm. Having said this, one should know the common ones such as dimensionality reduction, supervised and unsupervised algorithms.



Types of Machine Learning algorithms

You've used machine learning dozens of times in a day, without even realising that you are doing so! (E.g. searching on Google requires machine learning) The reason for Google Search to work so well is because the machine learning software knows how to rank pages. The feature of auto-recognition and tagging of friends on Facebook is also due to machine learning.

How can we solve real world problems with Data Science?

All the tools and resources will help to resolve problems. But, the main aim should be to identify the right problem and to find the right tool or model which would be used while solving the problems. Data Science is now impacting myriad areas.

1. NETFLIX

Based on the movies viewed previously, **Netflix** uses collaborative filtering algorithms to recommend movies to its users.

2. SOCIAL MEDIA

Many social media sites are also using data science. Whether it is recommending you connections on **LinkedIn** or new products on **Amazon.** Whether it is personalising your **Facebook** feeds or suggesting to you people to follow on **Twitter.** Data Science does it all!

3. OTHER ONLINE SERVICES

Many online apps such as **eHarmony** (a dating Site), **Uber**, **Spotify**, **Stitch Fix**, **Hulu**, **Bombfell**, **Pandora**, etc. are collecting and using real-time data to customize and improve their users' experience. Personalization not only increases customer engagement and retention but also improves conversion rates; thus, positively impacting the organizational bottom-line.

All of the above may be tech domains, but Data Scientists also work in other domains such as-

4. BIOINFORMATICS

Scientists are working on analyzing genome sequence.

5. ASTROPHYSICS

Physicists use data science concepts while building and analyzing 100 TBs of astronomical data. (You should check this out- Interview: Kirk Borne, Data Scientist, GMU on Big Data in Astrophysics)

6. SPORTS

Toronto Raptors, an NBA team, is installing cameras on basketball courts. They collect huge amounts of data on a player's movement and playing styles. The team then analyses game trends (based on the data collected) thereby improving coaching decisions and team performance.

Want to be a Data Scientist? A piece of advice.

- Do you love data?
- You consider yourself an eye for identifying trends and patterns?
- Do you have strong foundations of statistics, code writing, and programming?
- Are you comfortable with handling unknown facts?
- Do you think you can deliver to impatient stakeholders?
- Do you think you can convince them of your findings?

If you've answered a yes to these questions, then Data Science is the right career for you!

So all you need to do is the master at one least one language or tool of your choice. Also, get your hands dirty by doing some projects or competitions! Now there are high chances that you would make a talented data scientist!

(Note- Some data science job descriptions are titled "Consumer Insights Manager". Hence, you can search for jobs with this title too, and increase your chances of being a Data Scientist!)

Data Science Learning Resources and Communities

These learning resources will make you better placed in bagging that dream data science job!

R Programming Language: Click here
ggplot2, Click here; ggpairs, Click here; reshape2, Click here

Python Programming Language: Learn Python the hard way
NumPy, Click here; Pandas, Click here; matplotlib, Click here; Scipy, Click here; scikit-learn, Click here

Github Repository: The Open Source Data Science Masters, Learn Data Science

Communities and Blogs: DataTau, Stack Exchange, GitHub

Datasets: Kaggle Competition, 6 data set lists created by data scientists, List of public data sources

2

Where are Big Data and Data Science used?

I. Banking

Banking industries are rich with data. Used or unused, there is an excess amount of data in these sectors. Most banks are, nowadays, bearing the pressure to stay profitable and simultaneously understand the needs, wants and preferences of the customers. Lately, many financial institutions have adopted some new models that will help them to sustain in the market field.

Then the banks will need to go beyond their standard business reporting and sales forecasting to be able to find out a set of crucial factors about success.

The application of data mining and predictive analytics to extract actionable insights and quantifiable predictions can help the banks to gain insights that comprise of all types of customer behavior, including channel transactions, account opening and closing, default, fraud, and customer departure.

Insights into these banking behaviors can be discovered through multivariate descriptive analytics, and predictive analytics, such as the assignment of credit score.

Banking Analytics, or applications of data mining in banking, enhances the performance of the banks by improving how banks segment, target, acquire, and retain customers. Furthermore, improvements in risk management, customer understanding, and fraud empower banks to maintain and grow a profitable customer base.

The Role of Analytics in Banking

- It helps banks become efficient by managing the myriad challenges they face.
- With the steady increase in the growing demand for the analytics, which has successfully managed to produce more sophisticated and accurate results, many more banks are deploying a range of analytics today.
- While basic reporting continues to be a relevant factor in the banks, advanced predictive and prescriptive analytics are now starting to generate potent insights.

Prominent use cases for banking analytics

1. Fraud Analysis:

- Fraud detection is a crucial activity that needs to be present, so that any fraudulent activities, either by the employers or customers, are curbed beforehand.
- Since banking is an extremely regulated industry, there is also a vast range of external compliance requirements that bank circumspect to combat against fraudulent and criminal activity.

2. Risk Analytics:

- Of all the other industries, the banking industry is predicted to be facing the greatest risk analytics investments, with 73 percent of banking respondents anticipating more than a 10 percent rise in expenditure.
- Its spending is expected to increase the most specific capabilities; in areas such as data quality and sourcing, system integration and modelling.

3. Loan amount prediction/classification:

- Based on customer details, banks can automate the loan eligibility process (real time).
- These details can be obtained from online applications as Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others.
- Building high-accuracy predictive model will assist in automating this process those are eligible for loan amount so that they can specifically target these customers.

4. Customer Analytics

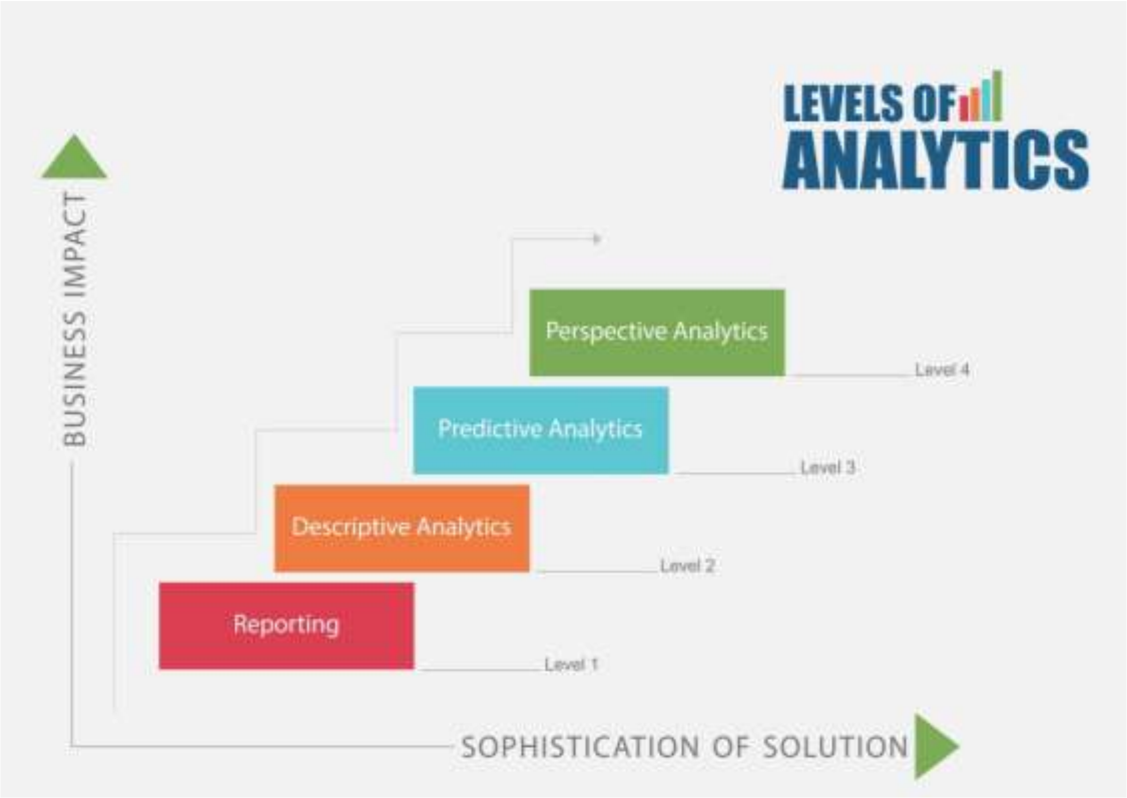
- It is said that it's profitable to retain a customer than to get a new customer in the present scenario. Banks are constantly at risk of losing customers or members, and to stem the flow, they may offer their best customers better rates, waive annual fees, and priorities treatments.
- However, such retention strategies have associated costs, and the banks cannot afford to make such offers to every single customer. The success and feasibility of such strategies are dependent on identifying the right action for the right customer.

5. Customer Insights

- Based on a customer spending patterns and transactions, the banks can create personalized experiences for individual customers using analytics and cognitive computing.

Did you know? From January 1, 2008, to April 15, 2011, the FDICclosed356 banks that failed to manage the risks building up in their residential and commercial mortgage exposures.
Interesting Fact: In the year 2012, fraud losses on cards in the United States reached US$5.33billion, up 14.5%
"By 2020, the world's computers are expected to hold 35 zettabytes (1021) of data."-IBM Corp.

Levels of Analytics



1. Reporting

- The Basic version of analytics solution that focuses on building data repositories and reporting the current situation using simple and uni/bivariate data.
- Examples in banking include; suspicious activity reporting and account validation against watch lists.

2. Descriptive analytics

- Generating actionable insights on the current situation using complex and multi-variety data.
- Examples in banking include; customer segmentation and profitability, campaign analytics, and parametric Value at Risk (VaR) calculations.

3. Predictive analytics

- Predicting the likely future outcome of events often leveraging structured and unstructured data from a variety of sources.
- Typical examples in banking include pattern recognition and machine learning to predict fraud, generating risk alerts at customer/product/ geography level, designing personalized and next-best offers, and trigger- based cross-sell campaigns.

4. Prescriptive analytics

- Prescribing action items required to deal with predicted future events using big data from a variety of sources often associated with simulations in various business scenarios.
- Typical examples in banking include behavioral PD1, LGD2, and EAD3 modelling, channel mix modelling, real-time offer models, next-best-offer models, and stress testing for mandated and custom scenarios

Areas in banking where Analytics has the maximum impact

1) Consumer and marketing analytics

2) Risk, Fraud, and AML/KYC Analytics, and

3) Product and portfolio optimization modelling.

Conclusion

Fortunately, customers leave a large data trail from an opening of an account to their day to day transaction with either debit/credit cards; this data can be leveraged by banking sector to solve their problems efficiently through building Data-driven Models and extracting valuable insights from results which in turn add a significant business value.

II. Classification of emails – Spam/Ham

-Spam Alert! 'Machine Learning' is filtering your emails

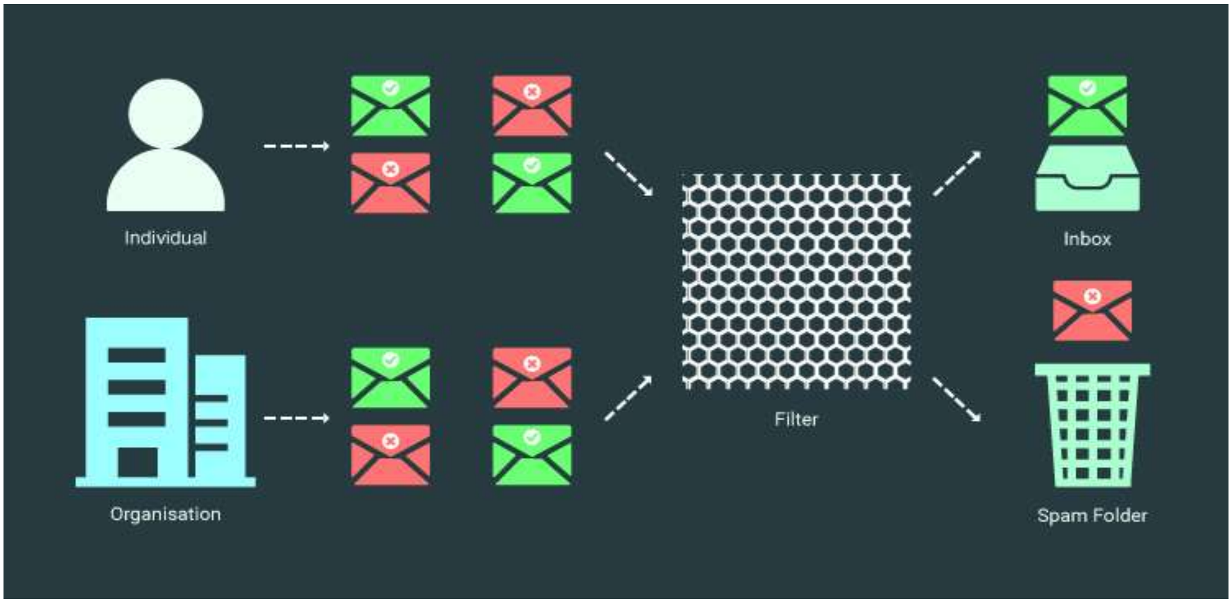
"Congratulations, you have won a lottery worth $10 million!! Reply with your credit card details to avail this."
No, I'm not offering any lottery, this is just an example of a spam e-mail. Nowadays we don't bother much about these emails because they automatically land in our spam folders. But 20 years back, this was not the case. So let's take a seat on a time machine and ride back to the past.

The 1990s – Just send it!

The 1990s marked the beginning of internet with Hotmail being the first web-based email provider. Standards such as sender authentication, whitelisting, etc. were unknown. Marketers exploited this opportunity. The strategy was 'just send it'. As a result, our inboxes get flooded with junk emails every day.

Can you imagine wasting 6 hours every day just to clean your inbox?
It was after this era when Spam Filters came to our rescue! Email providers started filtering emails based on sender's identity, analyzing message texts, etc.

What does a spam filter do?



- Email comes from various senders and organizations.
- Spam filter moves legitimate emails to inbox and rest of them to the spam folder.

How does it work?

Spam filter uses Machine Learning techniques to filter an email. It looks for several features in an email, based on which it decides whether an email is a spam or a ham (term used for legit emails).

1. Where the message came from?

Spam filter analyses the sender's address. Based on this, fraud emails can be detected. An anti-spam filter SpamAssasin tracks the network of the message source, maintains a list and looks it up in several other lists. Each time the network appears in a list, the spam score of the message is increased a little.

2. Which software sent the message?

Most legit mail comes from big email providers such as Gmail, Outlook (Hotmail), Yahoo, etc. Spam, however, is distributed by software that is designed to send out millions of messages as quickly as possible. The spammer does not want the messages to be easily traced back to their source. SpamAssassin looks for clues in the message headers that indicate that the message was sent by a spam engine rather than a real mailer.

3. What the message seems like?

Spam Filter also looks at the subject and the body of the message just as a person understands that a message "looks like spam". It searches for strings like *"lottery", "buy now", "lowest prices", "click here",* etc. It also looks for flashy HTML such as large fonts, blinking text, bright colours, etc. Many spam filters compare the amount of suspicious text to the total amount of text so that an entire 12-page paper doesn't get blocked based on a few suspicious words.

Challenges to spam filters

Spam filters although a very helpful tool poses a few challenges. Wondering how?

False Positives-

Let's say someone sent you an important email, and it went unnoticed because it landed in your spam folder. Maybe you could miss some important opportunity, invitation or anything just because of a fault in the probabilistic methods. In machine learning terminology, this case is called False Positive (i.e. a test result which wrongly indicates that a particular condition or attribute is present).

Tweaks by Spammers-

Another challenge is to face spammers who try to trick the filter by modifying emails in such a way that it could not be detected. For example, if a spam filter is made to look for the word 'buy now' then spammers modify the text like 'B-U-Y N-O-W', 'buy now', etc. to surpass the filter.

How has Google been addressing these challenges?

In February 2012, Microsoft boasted that its spam filters were removing all but 3% of the junk emails from Hotmail. Google responded by claiming that Gmail removes all but 1% of spams, adding that its false positive rate is also about 1%.

The relative success of both these companies showed that machine learning technologies were working. But 1% spam is still pretty annoying. And 1% false positive rate is, well, even more annoying.

Naturally, these companies keep improving their spam filtering techniques. Now, Google has come up with a new set of machine learning tools based on neural networks and deep learning. It has decreased their spam rate down to 0.1 percent, and false positive rate down to 0.05%.
"One of the great things about machine learning is that it adapts to changing situations." says John Rae-Grant, a senior product manager for Gmail.

End Notes

The spam filter is no doubt an important tool in the web mailing world. And, companies like Google, Microsoft make sure that they don't just filter junk mails based on pre-existing tools but also create their methods using advanced machine learning techniques.

III. E-commerce

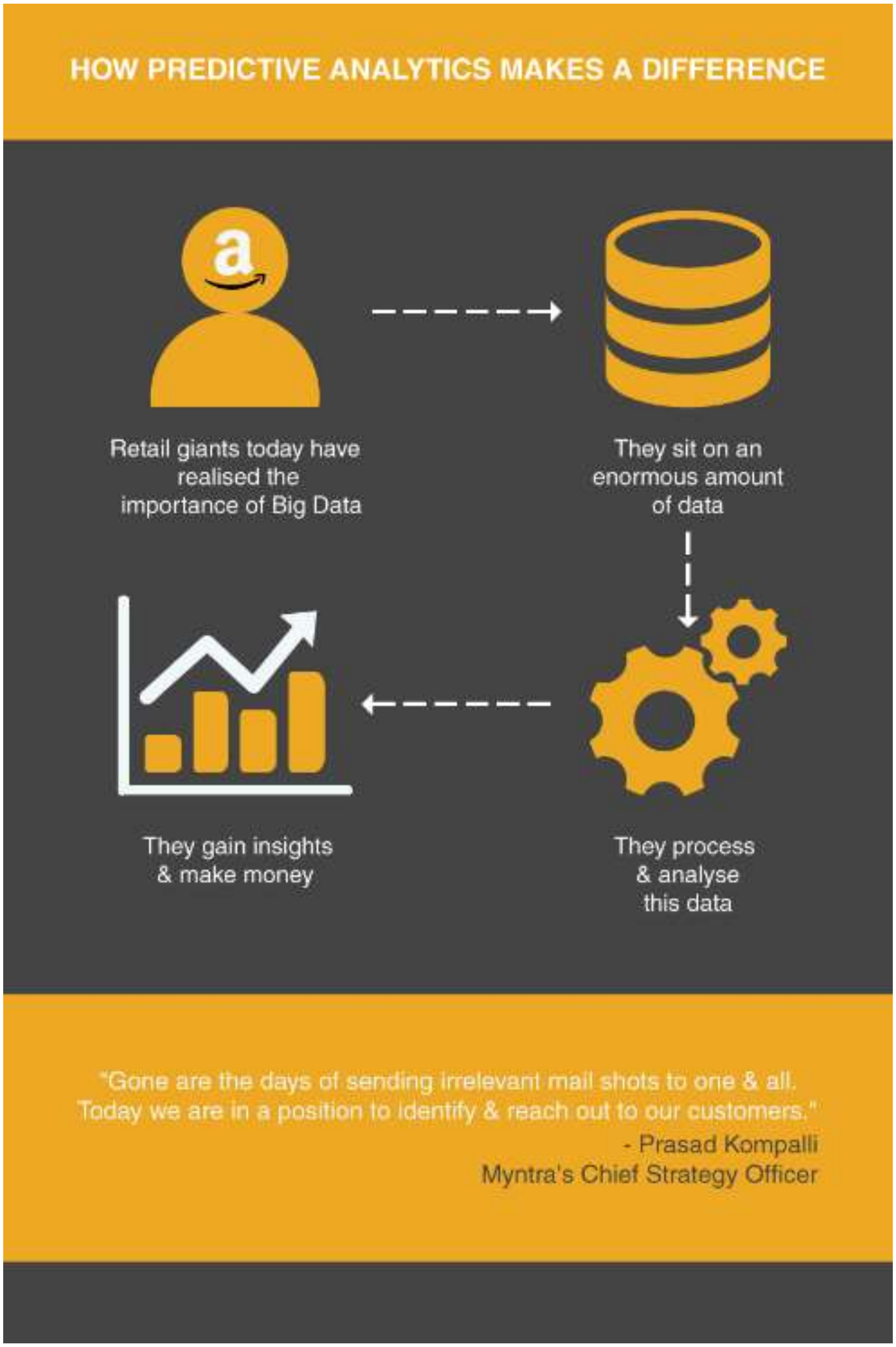
From the way e-tailers offer products to the way shoppers buy or transact online, Predictive Analytics has been reshaping and transforming the e-commerce industry.

E-commerce has taken the global market by a storm and is expected to grow exponentially in the upcoming years. A market research firm, **eMarketer**, projects e-commerce sales will eclipse $3.5 trillion within the next five years. **Walmart** (i.e. the world's largest company by revenue) has planned to close hundreds of stores to now refocus its efforts on its e-commerce site.
So, a big reason for this boom is the rapid growth in the internet and mobile users globally. E-commerce provides convenience to the users by giving better payment and advanced shipping option too. It also provides a new level of customization and personalization by using some advanced analytics tools.

Did you know? *E-tailers track and analyze your online behavior to gain insights that allows them to make billions of money?*
Have you ever wondered how Amazon knows if you are a movie buff, a gadget freak or an avid reader? Or how Google knows that you were looking for a new laptop on some website? Or how do you get the recommendation for an ideal party wear that you were looking for a long time, directly to your inbox?

In this infographic, we intend to answer these questions and know how the industry uses analytics

E-Commerce uses Predictive Analytics to deliver a great shopping experience!



HOW PREDICTIVE ANALYTICS MAKES A DIFFERENCE

Retail giants today have realised the importance of Big Data

They sit on an enormous amount of data

They process & analyse this data

They gain insights & make money

"Gone are the days of sending irrelevant mail shots to one & all. Today we are in a position to identify & reach out to our customers."

- Prasad Kompalli
Myntra's Chief Strategy Officer

3 WAYS IN WHICH ANALYTICS CAN ENHANCE SALES

CONVERTING BROWSER TO BUYER

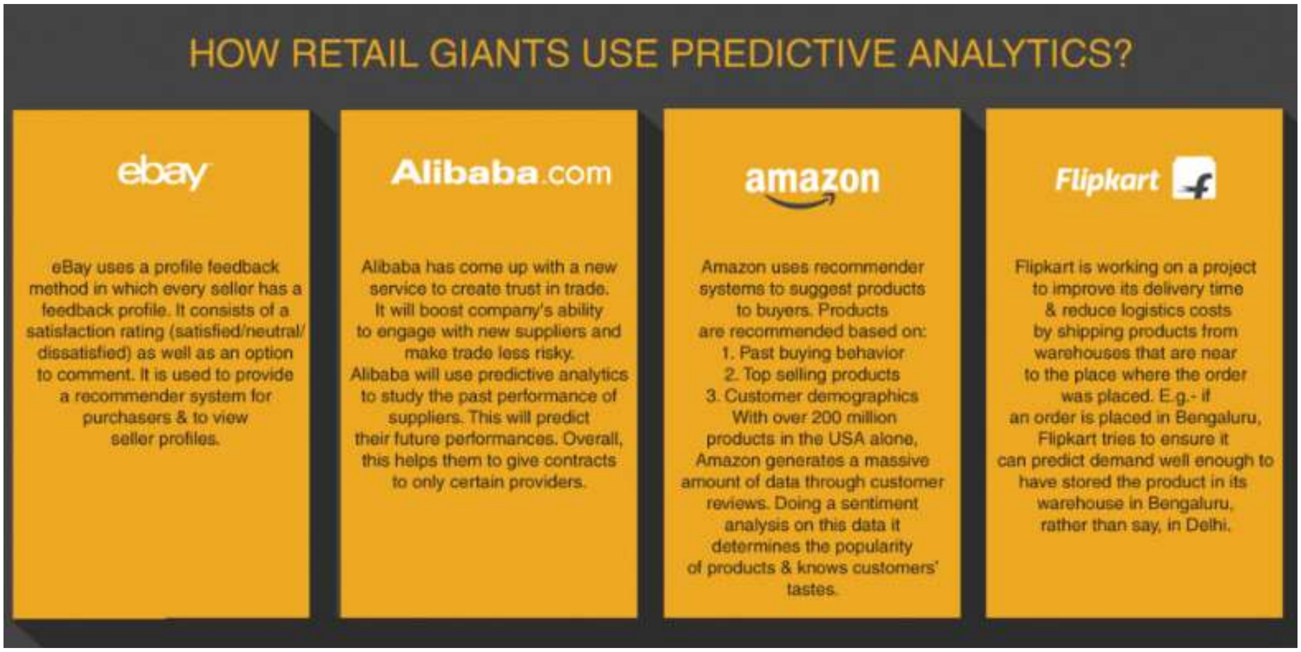
Visitors to a website often scan & browse through web sites without ever purchasing anything. Predictive Analytics helps customers offer targeted products & hence see only those products that interests them.

CROSS-SELLING

By suggesting additional products, it increases the average order size of the customer.

INCREASING LOYALTY

Seeing products as per their taste, adds an element of personalization hence increasing



HOW RETAIL GIANTS USE PREDICTIVE ANALYTICS?

ebay

eBay uses a profile feedback method in which every seller has a feedback profile. It consists of a satisfaction rating (satisfied/neutral/dissatisfied) as well as an option to comment. It is used to provide a recommender system for purchasers & to view seller profiles.

Alibaba.com

Alibaba has come up with a new service to create trust in trade. It will boost company's ability to engage with new suppliers and make trade less risky.
Alibaba will use predictive analytics to study the past performance of suppliers. This will predict their future performances. Overall, this helps them to give contracts to only certain providers.

amazon

Amazon uses recommender systems to suggest products to buyers. Products are recommended based on:
1. Past buying behavior
2. Top selling products
3. Customer demographics
With over 200 million products in the USA alone, Amazon generates a massive amount of data through customer reviews. Doing a sentiment analysis on this data it determines the popularity of products & knows customers' tastes.

Flipkart

Flipkart is working on a project to improve its delivery time & reduce logistics costs by shipping products from warehouses that are near to the place where the order was placed. E.g.- if an order is placed in Bengaluru, Flipkart tries to ensure it can predict demand well enough to have stored the product in its warehouse in Bengaluru, rather than say, in Delhi.

How Predictive Analytics helps Buyers?

- Provides products that are relevant to shoppers' needs.
- Reduces the time taken to shop.
- Reduces the hassle to choose from a variety of products.

Hence resulting in a great buying experience!

How Predictive Analytics helps Sellers?

Offering targeted and personalized products not only increases customer loyalty and retention but also improves customer satisfaction. It also aids in cross-selling and up-selling. These hence result in increased revenues and profits for the firms.

How the giants use Predictive Analytics?

- Amazon has the most sophisticated recommendation algorithms allowing it to offer targeted products.
- Alibaba- Chooses the best vendors to sell products.
- EBay: The American marketplace uses past data to rate sellers, thus helping buyers to make better decisions.
- Flipkart- Optimizes logistics and stock.

These were just some of the examples of analytics in e-commerce. Hopefully, you've now got a brief idea of how analytics is reshaping and transforming the online shopping experience.

IV. Retail chains - Starbucks

- Starbucks, Roasting Data and Brewing Analytics!

Have you ever wondered how your favourite coffee shop uses data analytics and business intelligence (i.e. BI) to deliver to you the unique 'Starbucks Experience'? This largest and the most recognizable coffee brand is one of the places where business and data analytics solutions meet in the real world. So read on to know how intelligent business strategy lies behind your freshly brewed cup of coffee!

1. Deciding a new Starbucks store location

Don't get surprised if you find Starbucks stores clustered near each other in most of the cities across the world. However, this seems contrary to common-sense expectation as it can negatively impact the profitability of these densely located stores. But, Starbucks is a clever market player in determining an optimal store location using data analytics.

So, you will find this coffee shop in some of the most prime and strategic locations all across the world. Hence, premium, high-traffic, high visibility locations near downtown, suburban retailers, work-spaces and university campuses are the major targets of Starbucks.

How BI helps a Starbucks store?

Starbucks acquires data from Atlas that is a mapping and BI platform developed by Esri. Esri (i.e Environmental Systems Research Institute) is a geographic information system (GIS) software company. This data helps to discover a potential new store location.

Step 1- Data to Insights

Starbucks carefully analyses data on parameters such as-

- Consumer demographics
- Population density
- Average income levels
- Traffic patterns
- Public transport hubs
- Types of businesses in the location under consideration.

So it connects this data with R and builds models based on pedestrian traffic and average customer spend of the location.

RESIDENTIAL
HIGH INCOME LEVEL
HIGH POPULATION
COFFEE LOVERS
STARBUCKS COFFEE

Step 2- Insights to Location

This coffee chain has an amazing BI team. Using the obtained insights, it determines the economic feasibility of opening a store in that spot. Thus, presenting a splendid example of transforming data into knowledge and knowledge into business strategy. The largest specialty coffee retailer in the world is undoubtedly becoming "The Third Place" between work and home due to the convenience of their locations.

2. Deciding Starbucks menu offerings

Are you a Starbucks fan? Your beloved coffee shop has been taking notes of your tastes! Not only does Atlas help it to bag an optimal store location but even assists in customizing menu offerings. Analyzing consumer data, Starbucks drafts its new line of products to supplement the habits captured from its own stores.

Coffee and Alcohol!

Moreover, Atlas also provided data to determine areas with the highest alcohol consumption. Based on this information, Starbucks smartly picked up its stores to serve alcohol as a part of a special menu called "Starbucks Evenings". Launched in 2010 in Seattle, it has since been expanded to other stores in different cities.

Summertide? Grab a Frappuccino!

In another example, this caffeine purveyor used Atlas to predict the arrival of heat waves in the city of Memphis (United States) and launched a local Frappuccino promo to beat the heat.

These data-driven menu enhancements enable Starbucks to reach out a large customer pool. Eventually, achieving the market dominance!

3. Starbucks loyalty program

Starbucks has one of the most sophisticated loyalty programs in the world making it a marketing success.

Do you happen to be a part of "My Starbucks Rewards"? If yes, Starbucks knows who you are and how you differ from others.

As a matter of fact, it has over 10 million loyalty program members in US and around 24,000 stores worldwide. To sum up, the largest coffee brand certainly roasts a lot of data!

How Starbucks uses this data

Can you imagine the takeaways this international brew chain derives from your data? Below is a list of the few important ones. Starbucks-

- Identifies you by linking 'What', 'Where' and 'When' of the products you buy.
- Tracks down your product purchasing behavior.
- Delivers targeted advertising and planned discounts directly to your mobile devices.

Using analytics and Business Intelligence (BI), Starbucks has successfully transformed 300 pages long reports into 11 KPIs for each store across the globe.

CLEANLINESS
PRODUCTIVITY
CUSTOMER SATISFACTION RATING
TRAFFIC
TICKET
INVENTORY
SALES
OFFERS & DISCOUNTS
CUSTOMER PROFILING
ROYALITY PROGRAM
REVENUE

Did you know?
Starbucks was the second most valuable fast food brand worldwide in the year 2015 (Data analytics and BI played a major role)!

Conclusion

So, from a single storefront in Seattle to a global coffee phenomenon, the data-driven business strategy is the soul of this coffee company. Turning disloyal customers into loyal ones highlights Starbucks' strong business intelligence. Thus, we can say that Starbucks brews every cup of coffee with strong data science and analytics beans!

V. Predicting Election Results

How data analytics powers 2016 US election

The best way to predict the future is to study past behavior. This is the underlying idea behind Big Data Analytics. The 2008 Obama election campaign was one of the first to take advantage of data-driven methods in the race to an elected office. The Obama campaign had a data analytics team of 100 people. This shows how deeply data analytics impacts the world. From recommending products to customers on e-commerce websites (i.e. using predictive analytics) to electing the most powerful official of the free world. Big Data Analytics is indeed everywhere. Data analytics has evolved itself to become the brain of every election campaign since the Obama campaign. Data analytics helps the election campaign to understand the voters better and hence adapt to their sentiments. Now let's find out how data analytics affects the elections and how election campaigns use it.

How Data Analytics affects the election

The 2016 race to the White House had data at its center and made itself an unstoppable force. The question here is how it affects the outcome of the election? Positively or Negatively? In other words, does data analytics have the ability to turn election results? Social and polling data can affect the voters.

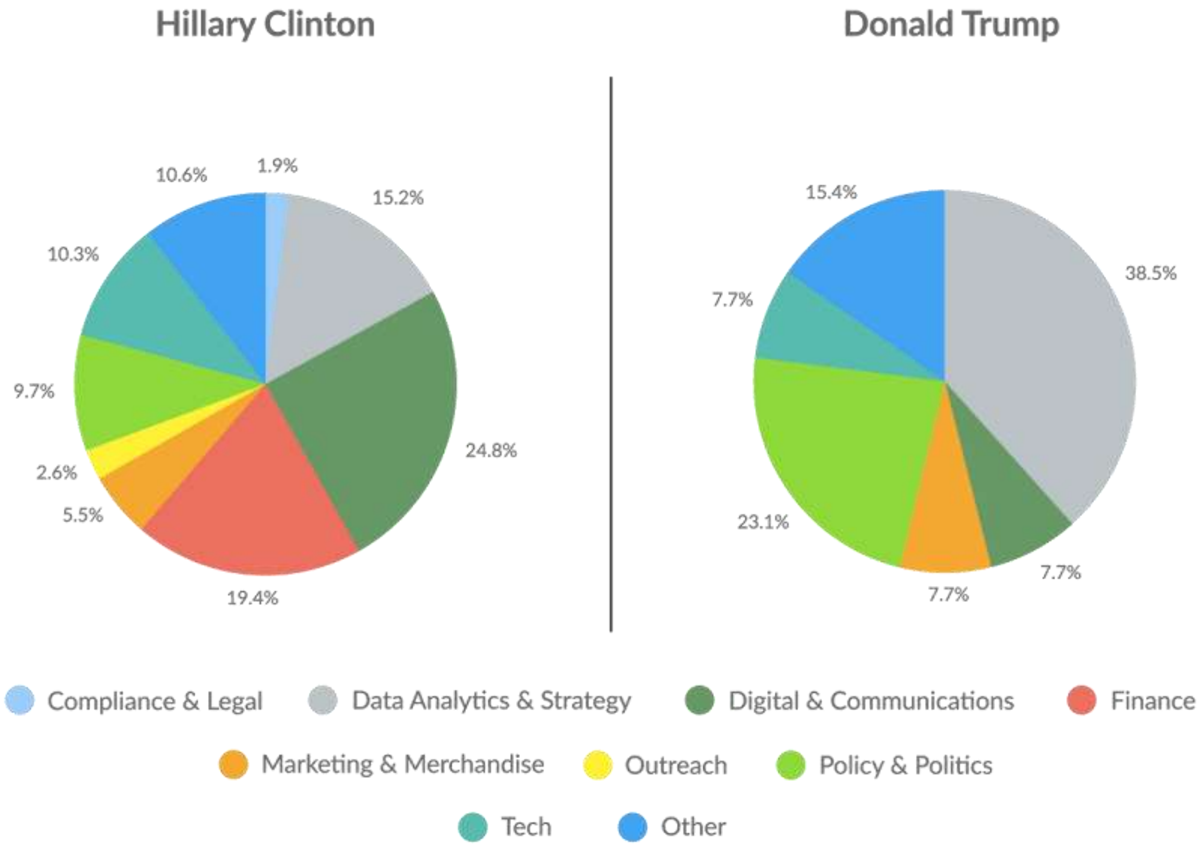
Social websites such as Facebook and Twitter optimize their feeds to the target audience to promote voting. Conversely, you see Hillary Clinton leading by 73% chance to take over the White House in the polls (released by some analytics firm) to Trump. In reality, would you agree that a good number of people would feel that the election result is obvious now?

As a result, I feel that there will be a negative impact on the voter turnout in such situations. Websites like Five Thirty Eight and Real Clear Politics use social, polling data to predict the election results. To emphasize, if they tweak those results for a single candidate, then it can give an altogether different perspective to their millions of followers, who now after knowing the probable outcome may not turn out at the booths to support their candidate. Hence, it is crucial to realize the downside as well.

How election campaigns use Data Analytics

There are two subdivisions of extracting data for an election campaign. Firstly, social data and polling data and secondly public data which becomes a part of Big Data. It helps the candidate to understand the voters better and design the campaign accordingly. Moreover, this brings more clarity to the election campaign. Both Clinton's and Trump's campaigns are relying on technology for reaching out to the voters in the 2016 race for The White House.

The Campaign job distribution for both Hillary and Trump Campaign obtained from ValuePengiun is shown below. We can observe from the graph that Data Analytics and the resulting Strategic Operations takes up a huge chunk of the workforce of both the presidential campaigns.



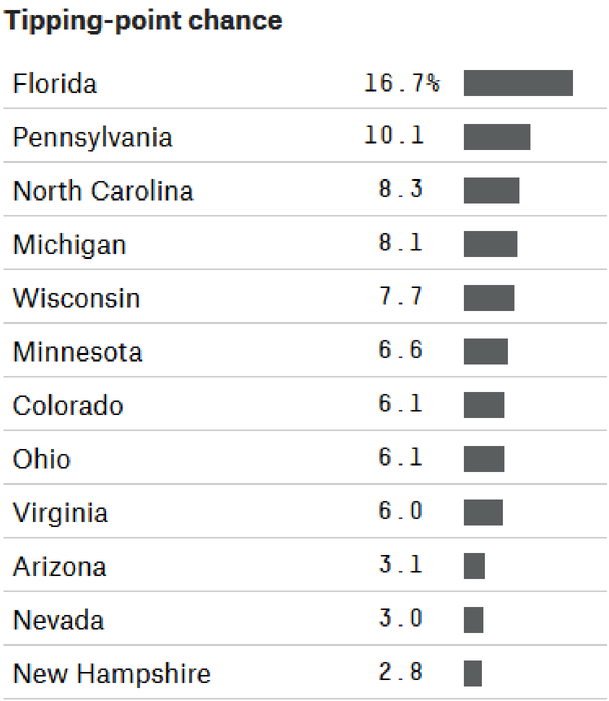
Identifying the Swing States

A swing state is a state where the two major political parties have similar levels of support among voters, viewed as important in determining the overall result of a presidential election. Swing states are one of the most important factors in the US elections.

Red states are ones that are dominated by the Republicans (i.e. Trump's party) whereas the blue ones signify the dominance of the Democrats (i.e. Clinton's party). Hence, swing states are also known as purple states as both parties have similar electoral support in these areas.

Large amounts of public data, polling data, sentimental analysis of Twitter and Facebook feeds are used to determine the swing states. In particular, winning the swing states can make a big difference in the electoral votes. These are the best opportunity for a party to gain electoral votes. So, political parties majorly focus on these states while strategizing their election campaign.

In 2016, US Presidential elections the 12 swing states are – Wisconsin, Minnesota, Nevada, Pennsylvania, New Hampshire, Colorado, Ohio, Iowa, Virginia, Florida, Michigan, and North Carolina. "Tipping-point chance" as described by FiveThirtyEight is the probability that a state will provide the decisive vote in the Electoral College. This is a good indicator of the Swing states.



Tipping-point chance

Florida	16.7%
Pennsylvania	10.1
North Carolina	8.3
Michigan	8.1
Wisconsin	7.7
Minnesota	6.6
Colorado	6.1
Ohio	6.1
Virginia	6.0
Arizona	3.1
Nevada	3.0
New Hampshire	2.8

Online and offline marketing

Using big data analytics, the election campaign analyzes the demographics of the states where they fall behind their opposition. Offline marketing like billboards and television ads is deployed strategically to target the audience using data analytics.

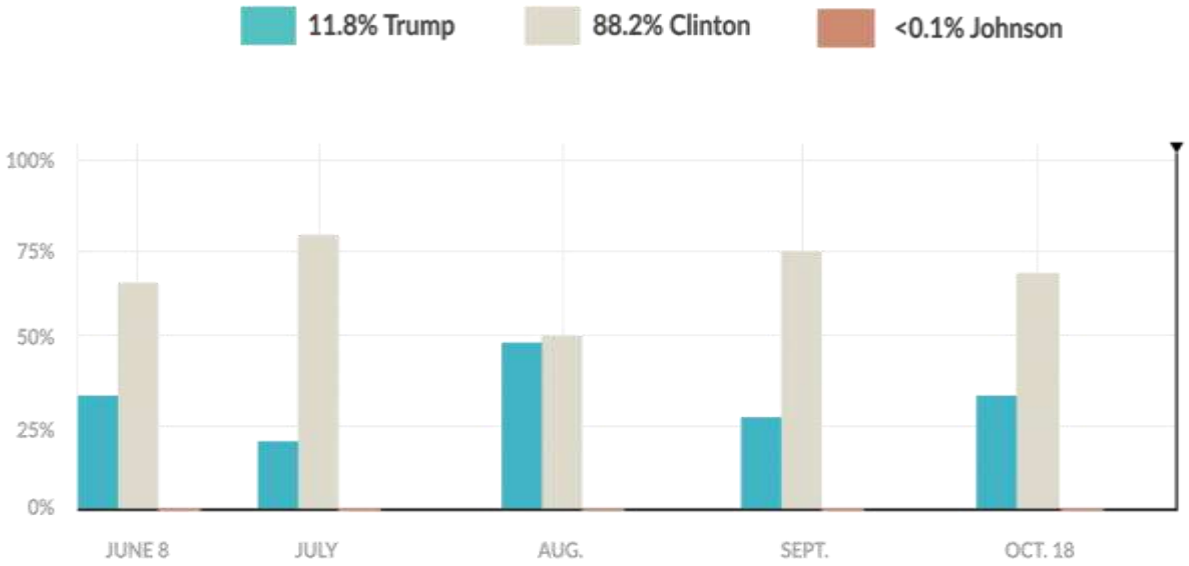
It helps them to understand the states where the campaign needs to improve on the marketing and hence turn the voter sentiments around.

Big Players in the Election Forecast

Now, let's look at some of the notable players who use Data analytics on polling, social and big data for the forecast.

Five Thirty Eight

In 2007, Nate Silver launched Five Thirty Eight. Silver made data analytics super cool with his famous 2008 US Presidential election predictions. Five Thirty Eight's 2008 presidential election forecast had 98.08% accuracy in predicting the winners in each of the states. Notably, they correctly predicted the winner of 49 of the 50 states including the District of Columbia. Overall, Indiana is the only state in which they missed out. Five Thirty Eight's prediction on "chance of winning" for the 2016 election cycle is shown below.



Real Clear Politics

John McIntyre and Tom Bevan founded RCP in 2000. They are one of the leading websites which collects a lot of polling data and generates a predictive analytical model for the forecast.

Conclusion

We are in the midst of yet another US Presidential election which is due to take place on 8th November 2016. To sum up, we have seen how highly data analytics is used by election campaigns and how it affects elections as a whole. Additionally, this also opens a whole world of possibilities on how someone can be a part of such a technological field with great impact.

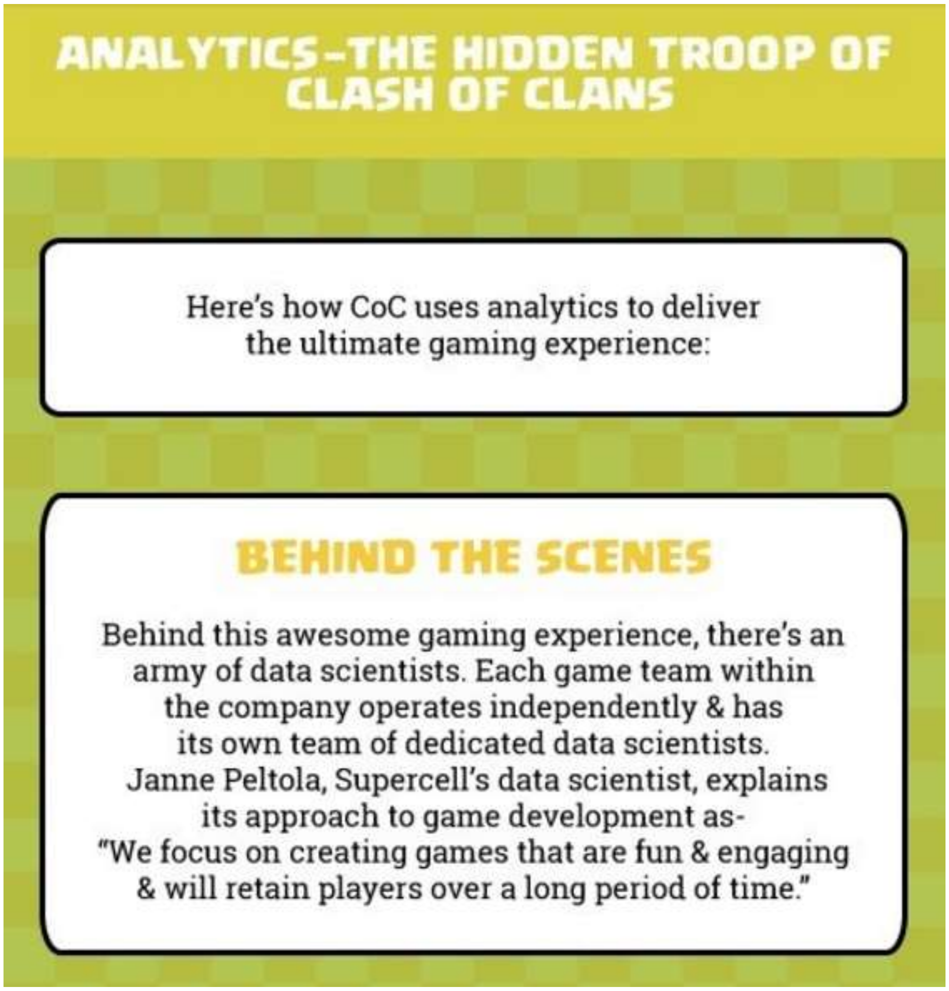
VI. Gaming

- The Hidden Troop of Clash of Clans!

Analytics has found uses in a variety of domains such as Healthcare, Insurance, Banking & Finance, etc.

Now it has forayed into the gaming sector and is finding uses here too.

Read on to know how Clash of Clans, a popular game, uses analytics to enhance its gaming experience.



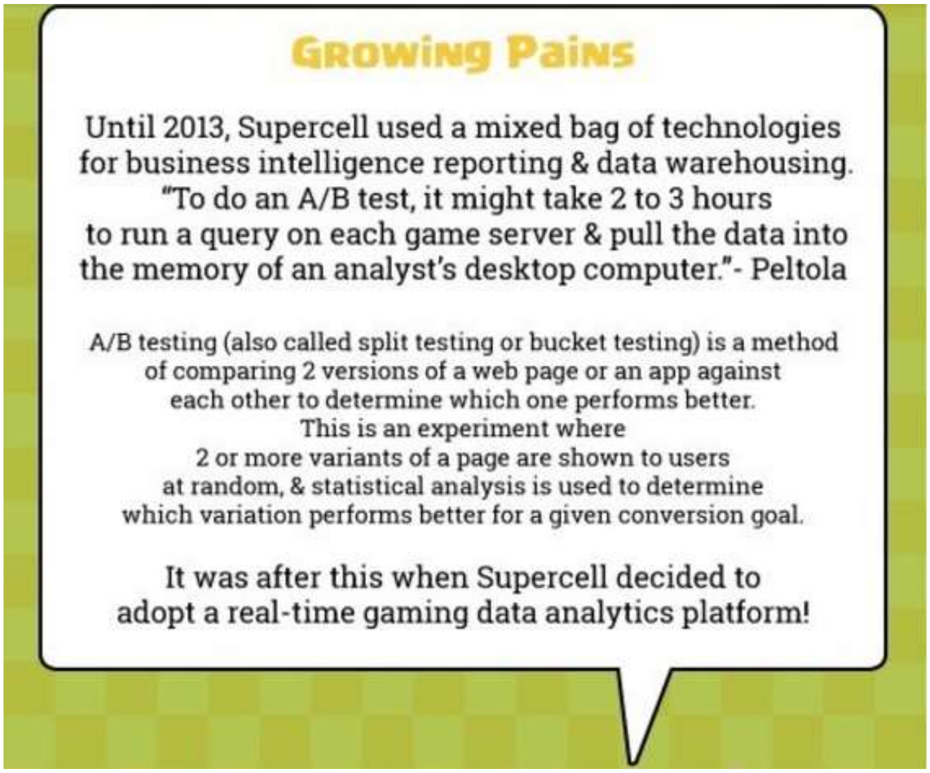
ANALYTICS–THE HIDDEN TROOP OF CLASH OF CLANS

Here's how CoC uses analytics to deliver the ultimate gaming experience:

BEHIND THE SCENES

Behind this awesome gaming experience, there's an army of data scientists. Each game team within the company operates independently & has its own team of dedicated data scientists. Janne Peltola, Supercell's data scientist, explains its approach to game development as- "We focus on creating games that are fun & engaging & will retain players over a long period of time."

Did you know? Supercell (CoC's creator) is a Finnish mobile game development company. Founded in June 2010, Supercell is currectly valued at a gigantic $10.2 billion! It has developed 3 games 'Hay Day', 'Clash of Clans' & has recently launched 'Clash Royale'.
CoC being the most popular of them all with $460,922 in daily revenue & 25,374 daily installs estimates.

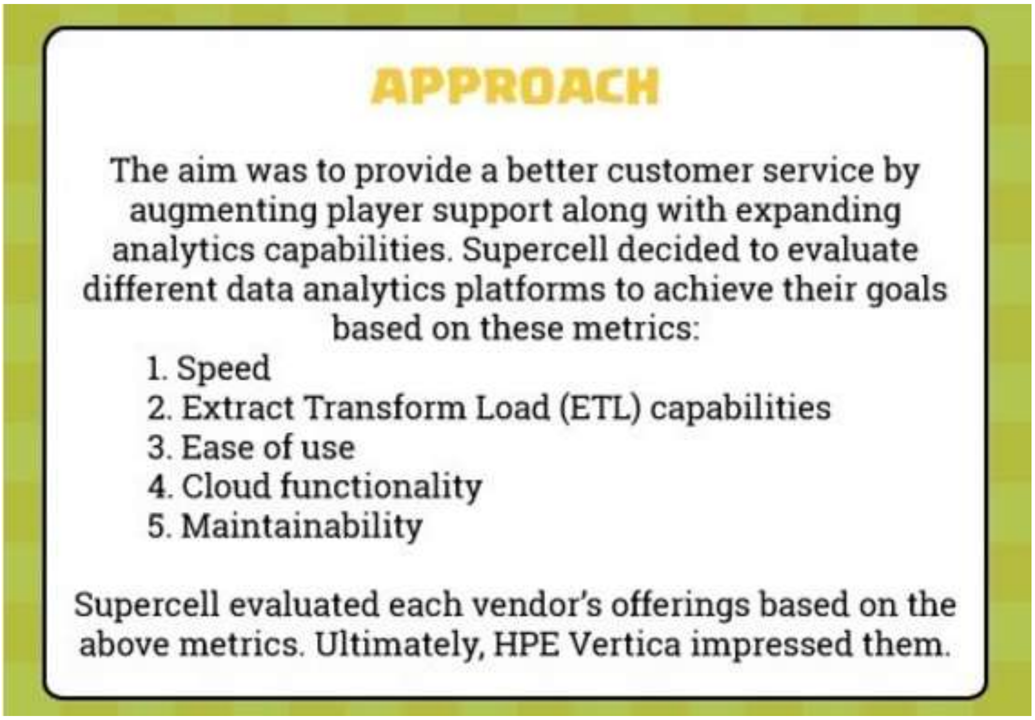


Growing Pains

Until 2013, Supercell used a mixed bag of technologies for business intelligence reporting & data warehousing. "To do an A/B test, it might take 2 to 3 hours to run a query on each game server & pull the data into the memory of an analyst's desktop computer."- Peltola

A/B testing (also called split testing or bucket testing) is a method of comparing 2 versions of a web page or an app against each other to determine which one performs better. This is an experiment where 2 or more variants of a page are shown to users at random, & statistical analysis is used to determine which variation performs better for a given conversion goal.

It was after this when Supercell decided to adopt a real-time gaming data analytics platform!



APPROACH

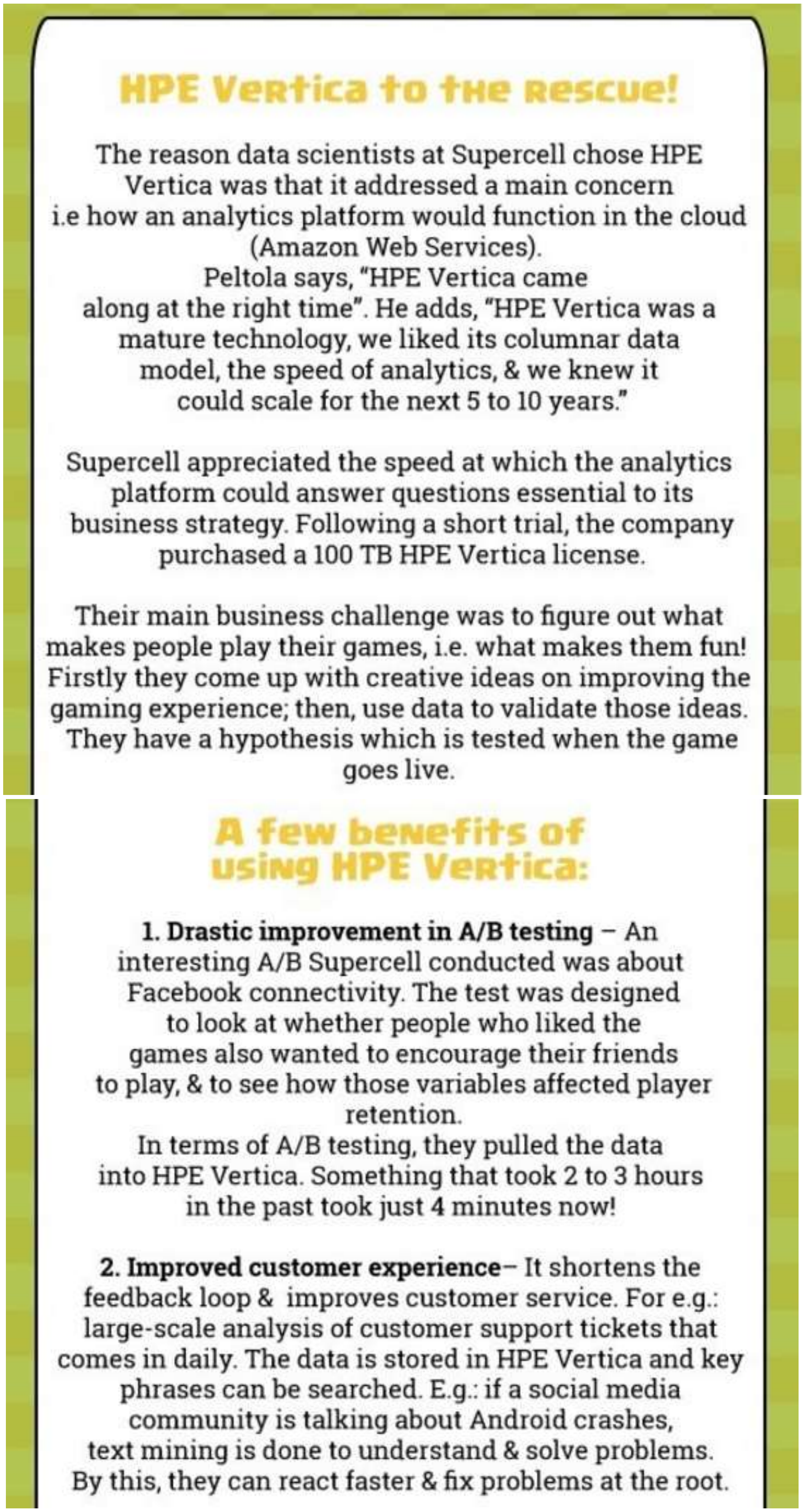
The aim was to provide a better customer service by augmenting player support along with expanding analytics capabilities. Supercell decided to evaluate different data analytics platforms to achieve their goals based on these metrics:

1. Speed
2. Extract Transform Load (ETL) capabilities
3. Ease of use
4. Cloud functionality
5. Maintainability

Supercell evaluated each vendor's offerings based on the above metrics. Ultimately, HPE Vertica impressed them.

Did you know?

Vertica Systems was an analytic database management software company. Hewlett-Packard acquired it in March 2011.



HPE Vertica to the rescue!

The reason data scientists at Supercell chose HPE Vertica was that it addressed a main concern i.e how an analytics platform would function in the cloud (Amazon Web Services).
Peltola says, "HPE Vertica came along at the right time". He adds, "HPE Vertica was a mature technology, we liked its columnar data model, the speed of analytics, & we knew it could scale for the next 5 to 10 years."

Supercell appreciated the speed at which the analytics platform could answer questions essential to its business strategy. Following a short trial, the company purchased a 100 TB HPE Vertica license.

Their main business challenge was to figure out what makes people play their games, i.e. what makes them fun! Firstly they come up with creative ideas on improving the gaming experience; then, use data to validate those ideas. They have a hypothesis which is tested when the game goes live.

A few benefits of using HPE Vertica:

1. Drastic improvement in A/B testing – An interesting A/B Supercell conducted was about Facebook connectivity. The test was designed to look at whether people who liked the games also wanted to encourage their friends to play, & to see how those variables affected player retention.
In terms of A/B testing, they pulled the data into HPE Vertica. Something that took 2 to 3 hours in the past took just 4 minutes now!

2. Improved customer experience– It shortens the feedback loop & improves customer service. For e.g.: large-scale analysis of customer support tickets that comes in daily. The data is stored in HPE Vertica and key phrases can be searched. E.g.: if a social media community is talking about Android crashes, text mining is done to understand & solve problems. By this, they can react faster & fix problems at the root.

3.Ask questions to Data – This platform can not only answer a lot of questions, but it can also go deeper. It enables data scientists to converse with the data. For e.g.: Now they can get an answer to "How are Japanese iPhone players doing?" & to a follow-up question like "Are Japanese users using both iPads and iPhones?". A conversation with the data would take a week to come in the past. Now, they can deepen the data & come to a conclusion almost on the fly, with agility and speed.

Conclusion

Be it mobile games or PC games, these can bring joy to anyone from 5-50 years of age!
Use of Big data & Analytics in the gaming industry has enabled game developers to take data-driven decisions. This case study was one such application of game analytics in today's world. Hopefully, now you've got some understanding about the role analytics plays in this industry.

Now, your barracks are full ! Analytics is your troop !

Go for the attack !

VII. Word Cloud

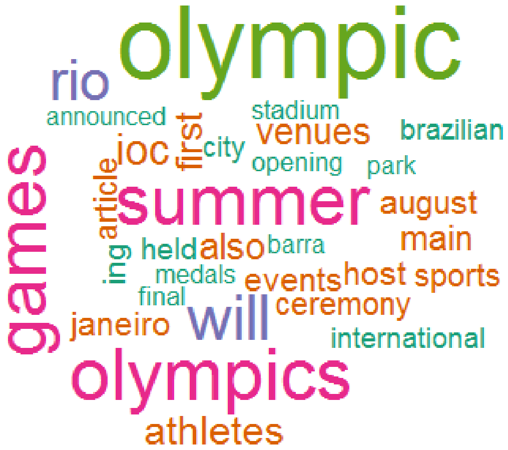
- A big thing in a small time!

Reading large textual data can be tedious sometimes. Be it a long speech, blog post or a big database, you must have encountered such monotonous situation! What will you do when you want to convey crucial information from a 10,00,00,000 (still counting the zeros!) words of text document in a very limited time? Nothing to worry. Data analytics and visualization to the rescue! Using a word cloud can make even a dull data beautifully insightful. Read on to know more about this graphical representation and even build one using R.

What is a word cloud?

Word cloud is a simple yet powerful visualization technique. It is primarily used to highlight significant textual data points. Word cloud is also known as text cloud or tag cloud. For instance, below is the word cloud created using Wiki data on Rio Olympics, 2016.

Altogether, the more frequently a specific word appears in the textual data, the greater prominence is given to that word in the cloud. Hence, the bigness and boldness of a word depend on the frequency of that particular word in the document.



Create your own word cloud!

Want to make your own tag cloud? Let's do it! Here, we will be using R to convert a source text document into a vibrant word cloud (in fact, with only few lines of code).

Pre-requisites:

1. **A working RStudio** (open source integrated development environment (IDE) for R)
2. **Source text data**

Step 1-

Install the following packages in RStudio:

- **tm** (a framework for text mining applications within R)
- **wordcloud** (package for creating pretty word clouds)
- **RColorBrewer** (provides color schemes for graphics)

Note: To install a package in RStudio, use the following command

```
> install.packages("package_name")
```

Step 2-

After successful installation of the above packages, load these packages into your R script using `library(package_name)`

```
library(tm)
library(wordcloud)
library(RColorBrewer)
```

Now, we are all set to use the functionality provided by these packages in our script.

Step 3-

To create a word cloud, you need a text document as an input to your R script.

Note: The source text data used in this blog can be downloaded from here. Make it sure that your R script and text file are in the same working directory.

```
mydata <- readLines(“Rio olympics.txt”)
#This function will read the entire text document line by line into the user defined variable (i.e. mydata)
```

Step 4-

After loading the text data, we need to preprocess it and finally convert it into a plain text. Corpus means list of documents.

We used this class from 'tm' to create a corpus from character vectors using 'VectorSource' method and passed it to user defined 'mycorpus'.

```
mycorpus <- Corpus(VectorSource(mydata))
```

Step 5-

Once our corpus is ready, we need to preprocess or modify the documents in it, e.g stopword removal, stemming, removing punctuations, numbers, symbols etc.

These transformations are done using **tm_map()** function which maps a given function to all the elements of corpus.
# write comments for each line

```
mycorpus <- tm_map(mycorpus,tolower)
mycorpus <- tm_map(mycorpus,removeWords, stopwords("english"))
mycorpus <- tm_map(mycorpus,removeNumbers)
mycorpus <- tm_map(mycorpus,removePunctuation)
mycorpus <- tm_map(mycorpus,stripWhitespace)
mycorpus <- tm_map(mycorpus, PlainTextDocument)
```

Step 6-

Done filtering your text document? Great! You are just one step away to create your own word cloud.

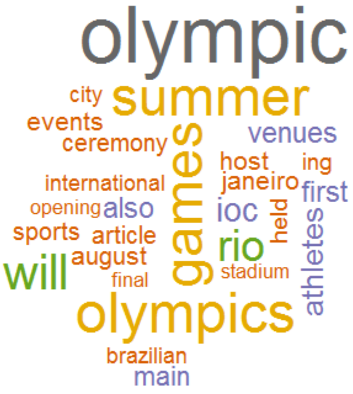
The final instance of 'mycorpus' is a plain text document without punctuatuions, white space, numbers and stopwords. Perfect condition to create a word cloud!

```
wordcloud(mycorpus, min.freq=9, colors=brewer.pal(5, "Dark2"))
```

'wordcloud()' method from 'wordcloud' package takes the plain text document (i.e mycorpus) and creates a word cloud based on the minimum threshold frequency of words (here it is 9).

Alternatively, you can create a colorful wordcloud using colors attribute in 'wordcloud()' method.

Your word cloud is ready!



Time to derive some insights

- Firstly, this text document reveals about summer Olympics.
- Secondly, Rio seems to be the hosting city.
- Thirdly, we can say that the document talks about medals, venues and stadiums.
- Since, Brazil is the hosting nation, the document also speaks something about Brazilian Olympic community.

Conclusion

Word cloud reveals the essential by popping brand names, key words etc. to the surface. Moreover, it is engaging and invokes the interest among the audience. In fact, just observing a word cloud gives you overall sense of the text. Lastly, we can say that word cloud is a handy tool for quick visualization.

Machine Learning

Introduction to Machine Learning

Machine Learning is a field that is at the forefront of computing today. In fact, it is omnipresent in the computing world! Even if you've never heard of Machine Learning before, you use it many times (in a day!) without probably even realizing it! You make a **Google search**. There's Machine Learning running in the background which ranks the results (so that you see the most relevant results first). You log in to your **email account.** It is Machine Learning that determines which emails should appear in your inbox and which ones in your spam folder.

Did you know? Machine Learning, or ML for short, is being used today to make driver-less cars, predict emergency room waiting times, identify whales in oceans based on audio recordings (so that ships can avoid hitting them), and make intelligent recommendations on which movie one should watch next on services such as Netflix.

What is Machine Learning?

Put simply, Machine Learning is, making computers learn from past experience. The concept of ML is, in essence, very much like we humans learn. Think of how we learn to walk. As infants, we found it difficult to walk, but as we grew, we learnt from our experiences of falling and taking bad steps. But eventually, after some time, our brains learn to walk comfortably, and thus we become the faultless walkers that we are now.

Similarly in ML, such experiences serve as data. (For example- this data could be information from the sensors of a car as it learns to drive.) Thus, we use past data to make the computer learn. Based on this "learning", it makes predictions about future events.

How does a computer learn?

There are different types of learning, but we will discuss those in the next section. For now, let's know how a computer learns, in a general sense, with the help of an example.

Let's say you want to predict whether a person will be hired for a certain job position or not. Then how would you go about making such a predictive model? Well, we can break the complete process down to five main steps.

First step

Getting the past data of people, who were hired or rejected for the same job position, on factors such as educational qualifications, previous work experience, etc.

Second step

Cleaning, preparing or manipulating the data collected. This step involves converting the data into a form that the computer can operate on. So, in our example, we would convert everything to numerical data. This would include converting the yeses and noes (for work experience) to 1's and 0's (i.e. binary form) and dividing educational qualifications into groups (say, Masters, Bachelors, Higher Education, etc.) and assigning a number to each group.

Third step

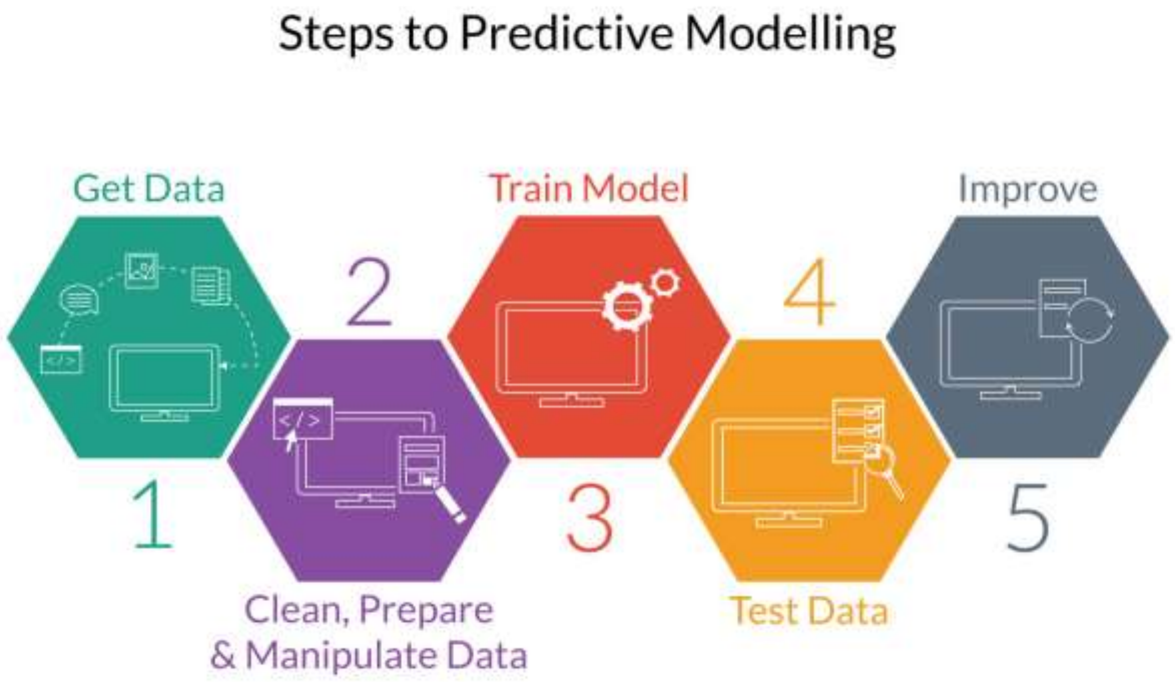
Building a mathematical model of the data. In our example, we have (what we call in ML lingo) a *classification* problem, i.e., we need to classify a person into one of two categories; whether they are going to be hired or not. A mathematical model allows us to succinctly define our data. The different parameters of this model can then be tweaked as per what we learn from our data. This would help us make the best possible predictions. This step is called *training*.

Fourth step

Using this model to make predictions. We can validate our model by running it on the same data that we used in training and then test the accuracy of our model by running it on the test data, i.e., data that the model has not been introduced to, before. This is called *testing*.

Fifth step

Making further improvements to decrease the error in our model and thus increase the accuracy.


Steps to Predictive Modelling
Get Data
1
2
Clean, Prepare & Manipulate Data
Train Model
3
4
Test Data
Improve
5

Types of Machine Learning

There are different types of learning. The type of learning you use depends on the specific problem(s) you are trying to solve.

1. Supervised Learning

This is the most common type of learning. Here, you provide the computer with sample data where the output is clearly defined. Recall our example in the last section where we had previous data that we used to "supervise" the learning procedure. Here we tell the computer what the "right" answer is, thus acting as its teacher. Two problems that you'd want to use this kind of learning for are *classification* and *regression* problems

2. Unsupervised Learning

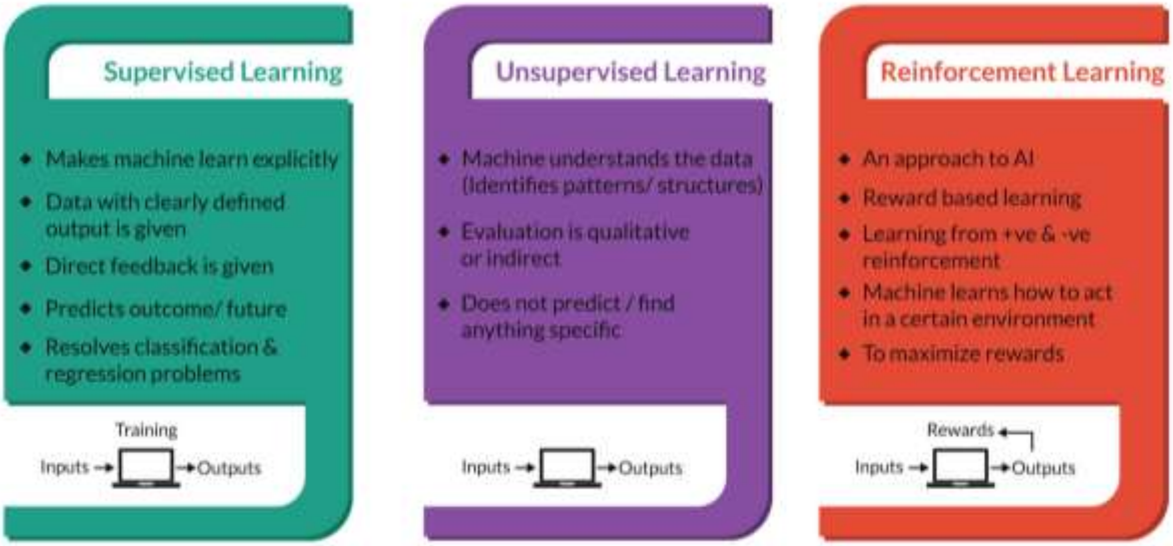
Here, we don't make the computer learn explicitly. Instead, we give it data and allow it to find underlying structures or patterns in it. *Clustering* is a problem for which unsupervised learning can be used.

Unsupervised learning is often used in *aggregation*. For e.g., Google news aggregates news articles (from various sources) on different topics. It then clubs the related topics to form a cluster. Thus enabling the user to find a host of articles on a certain topic. Another use of this learning is in *market segmentation or social network analysis*. Here one is interested in the patterns that can be found.

3. Reinforcement Learning

This is best described as a reward based system of learning. When you train your dog to do a particular task, you reward it with something so that it becomes used to performing that particular behavior. Therefore, a learner or an *agent* in ML vocabulary must learn over time what actions to take so as to maximize its reward.

This sort of learning often takes place while playing games. So, if we train a computer to play chess, we'd make it play a number games where the computer would figure out which step to take in response to the one taken by the opponent, so as to maximize its final reward in a future game.



Supervised Learning
Makes machine learn explicitly
Data with clearly defined output is given
Direct feedback is given
Predicts outcome/ future
Resolves classification & regression problems
Training
Inputs
Outputs
Unsupervised Learning
Machine understands the data (Identifies patterns/ structures)
Evaluation is qualitative or indirect
Does not predict / find anything specific
Inputs
Outputs
Reinforcement Learning
An approach to AI
Reward based learning
Learning from +ve & -ve reinforcement
Machine learns how to act in a certain environment
To maximize rewards
Rewards
Inputs
Outputs

Final notes

In this guide, we've talked about what Machine Learning is, how it works and the types of learning algorithms. Hopefully, now you have mastered the basics of this exciting field.

The 10 most popular ML algorithms

Machine Learning is a branch of Artificial Intelligence that "learns" without being explicitly instructed and programmed. Although it has been around for quite some time, it is gaining new momentum nowadays. This is the era of Big Data. Mining patterns from Big Data and using this information to make smart business decisions is a herculean and crucial task today. Using complex computational statistics in Machine Learning to do just this, is yielding brilliant results every day. Everyone has heard the term Machine Learning. It is omnipresent these days.

Regarding research, Machine Learning is probably one of the most important areas of computing today. Therefore, several algorithms have surfaced in Machine Learning, to help the computer learn as best as it can. This blog focuses on the most popular Machine Learning Algorithms used today, where you should use them, on what kind of data, and when.

1. Linear Regression

Linear Regression is the most popular Machine Learning Algorithm, and the most used one today. It works on continuous variables to make predictions. Linear Regression attempts to form a relationship between independent and dependent variables and to form a regression line, i.e., a "best fit" line, used to make future predictions. The purpose of this regression line is to minimize the distance between the data points and the regression line to make an equation in the form of

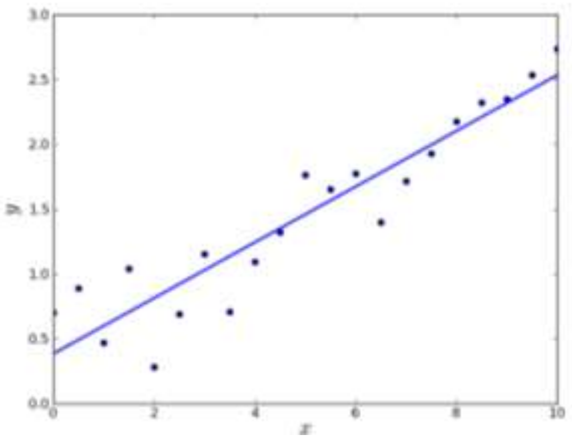
$$Y = a*X + b$$

I.e., the equation of a straight line where, "Y" is the dependent variable, "X" is the independent variable, "a" is the slope, and "b" is the intercept. The model is made based on the least squares estimation approach, among other lesser used approaches.

There are two types of Linear Regressions

– **Simple Linear Regression:** Where there is only one independent variable.

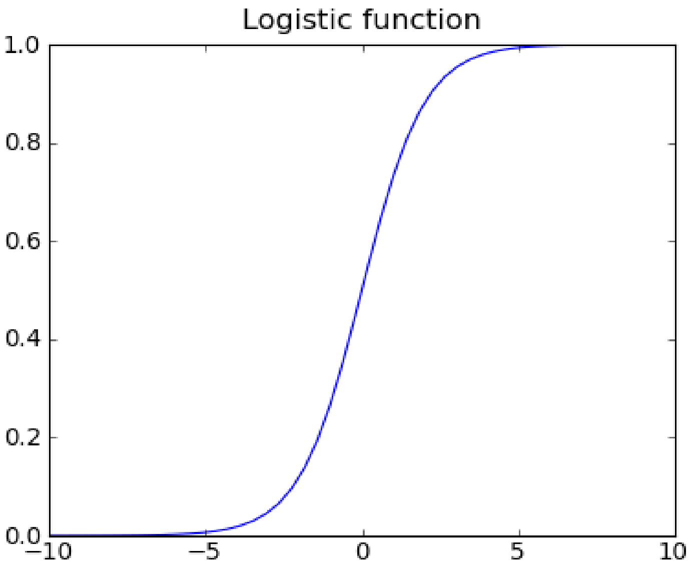
– **Multiple Linear Regression:** Where there is more than one independent variable.



3. Logistic Regression

Logistic Regression is a Machine Learning algorithm where the dependent variable is categorical. It estimates only discrete values, like 0 or 1, yes or no. The relationship between the dependent variable and the independent variables is determined by estimating probabilities using a Logistic Function. The curve plotted between the variables is an S-shaped curve as opposed to linear regression where it is a straight line. Logistic Regression is used when the outcome to be predicted binary – i.e., 0 or 1. Otherwise, other methods like Linear Regression are chosen. A logit function is used to predict the outcome variable Y when the outcome is to be categorical. The log of the probability that Y equals the independent variable. The equation becomes

$$\ln(p/(1-p)) = B0 + B1X1 + B2X2 + .. + BkXk$$



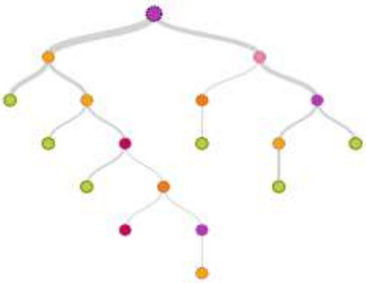
3. Decision Tree

Decision Tree is one of the most popular supervised Machine Learning algorithms. It is also the easiest to understand. Decision trees mainly perform classification, but sometimes, it forms regression trees as well. They work for categorical and continuous input and output variables. A decision tree splits the data into groups that are homogeneous but heterogeneous to each other.

So, how does the tree know when to use the methods?

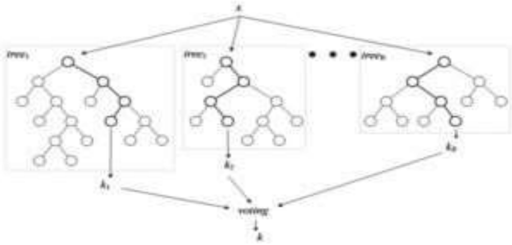
The tree makes use of a variety of algorithms to make this decision. But the goal of the algorithm is this – Splits are performed on all criteria, and the split that results in the most homogeneous subgroup is selected. The broad algorithms used are Gini Index and Information Gain.

E.g. If the credit worthiness of a customer is to be determined using Decision trees, many parameters are considered, as the income of the customer, his household size, and age. Now, a split is done based on each of these parameters. To determine which split provides the most homogeneous data, Gini Index or Information Gain is used.



4. Random Forest

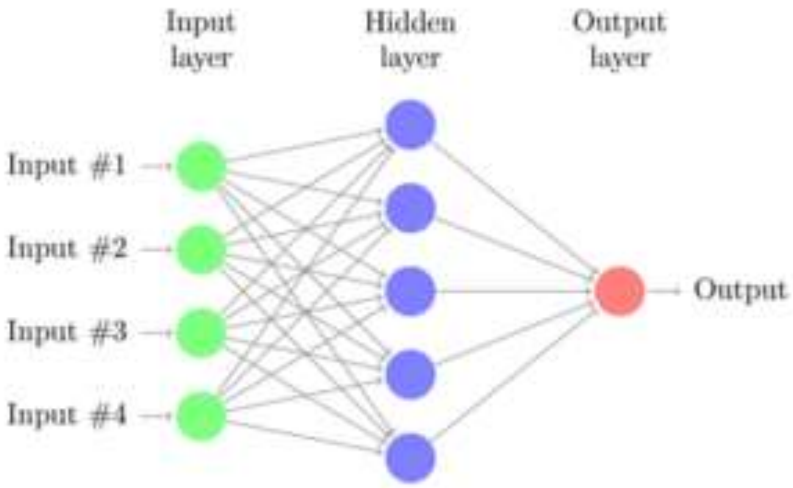
Random Forest is an ensemble learning approach that uses Decision Trees. It is an immensely popular Machine Learning algorithm that does one better than Decision Trees. Instead of a single tree, multiple trees are constructed. To classify an object based on a parameter, each of the trees gives a classification. Finally, the classification with the maximum votes is chosen. Random Forest works excellently for classification but not as well for regression. However, it is usually the go to the algorithm for most Data Scientists when dealing with a dataset that they're unsure how to classify.



5. Artificial Neural Network

Artificial Neural Network is a supervised Machine Learning algorithm that is making giant leaps in the field of Artificial Intelligence. ANN attempts to replicate the working of the brain to help the machine to “learn’ based on past data, much like our brain learns from past experiences. ANN requires a huge training data, and it takes the time to train itself. However, it performs very accurate predictions. It makes use of a concept of “hidden layers” with neurons at each layer and connections between all the neurons.

The exact working of ANN is quite a mystery, though, much like the human brain. It is something like a “black box”. ANN makes use of only two hidden layers. However, a sub-branch of ANN called Deep Learning has been making the rounds. It makes use of many more hidden layers to make bear perfect predictions. The use of more than two hidden layers is possible today because of the advancements in computation power.



6. Support Vector Machine

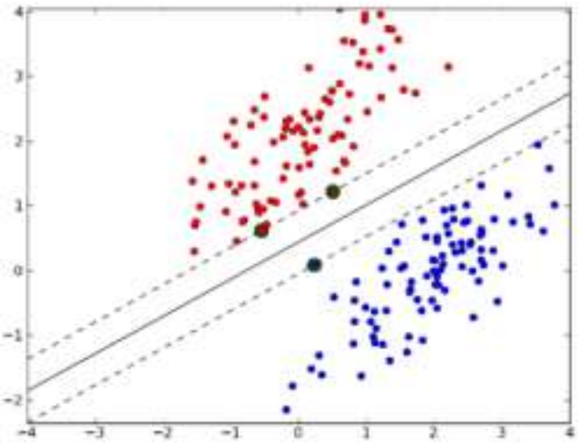
Support Vector Machine is a supervised Machine Learning algorithm that is used for classification and regression problems, but usually more for classification. It is a very effective tool that is used mainly when classifying an item into one of two categories.

What is a Support Vector?

Every individual data point's coordinate in the dataset is a Support Vector. These Support Vectors are plotted, and a classification is formed.

What is a Support Vector Machine?

Support Vector Machine is the hyper plane that divides or "classifies" the data points the best. For e.g., if you wanted to classify whether a given person was an Indian or an American, the parameters you'd choose are maybe the colour of their skin and their height. Based on the training data, we know that Americans tend to have fairer skin and are taller. Therefore, we now need to plot a line or find the best hyper plane that classifies the points in such a way that maximum distance exists between the closest data point from both the categories.



7. K-Means Clustering

K-means clustering is an unsupervised Machine Learning algorithm that deals with clustering of data. Using training data, the model finds the best structures and forms clusters.

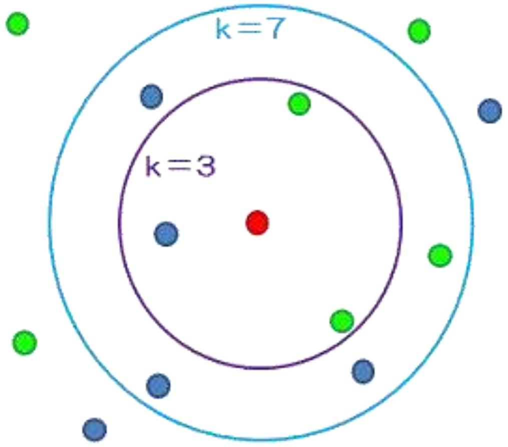
How does the algorithm work?

The algorithm takes two inputs. The number of clusters to be formed, and the training data X. Since there are no labels as it is unsupervised learning, there is no Y. Initially, the location of each of the K-clusters is the centroid. A new data point from the dataset is put into the cluster whose centroid it is the closest too. To determine the "nearness", the cluster with mean with the least sum of squares is used. As new elements are added to the cluster, the cluster centroid keeps changing. The new centroid becomes the average of the locations of all the data points currently in the cluster. These two steps of assigning a data point to the cluster and then updating the cluster centroid are done iteratively.

Towards the end, you will notice that the cluster centroid doesn't change since we've accurately created clusters that are homogeneous and heterogeneous with other clusters.

8. K-Nearest Neighbour

K-Nearest Neighbour is a simple Machine Learning algorithm that works on classification and regression problems, but more commonly on classification problems. A new element is classified into one of the classes based on a vote from its "neighbours". A parameter "k" which signifies the number of neighbours used. The selection of k is critical, as with a small value of k, the result is not accurate and results in a lot of noise. But selecting an enormous value of k is computationally infeasible, and defeats the purpose of this algorithm. There is a loose rule followed while choosing k, which is to assign it the square root of the number of samples in the dataset.



9. Naive Bayes classifier

Naive Bayes classifier is a widely used Machine Learning algorithm which uses the Bayes Theorem with the base assumption that every feature is independent i.e., no feature depends on any other feature. Bayes Theorem states that the probability of an event can be calculated based on conditions that might affect the event.

Bayes Theorem states that –

$P(c|x) = P(x|c) . P(c) / P(x)$

Where,

$P(c|x)$ is the probability that event of class c, given the predictor x

$P(x|c)$ is the likelihood of x, given c

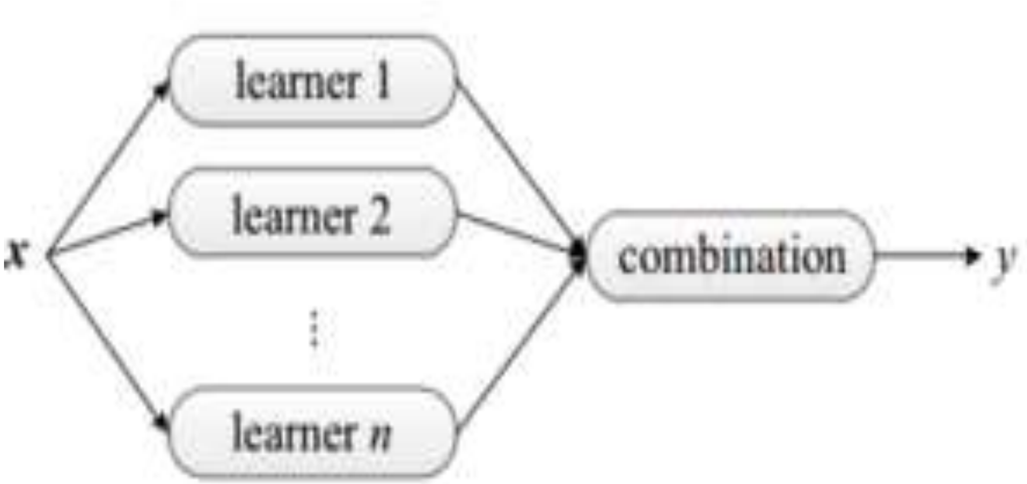
$P(c)$ is the probability of the class

$P(x)$ is the probability of the predictor

The posterior probability for each class is calculated, and classification is done based on the result of the calculation. The observation is classified into the class with the highest probability.

10. Ensemble Learning

Ensemble Learning is a machine learning which uses not one but many models to make a prediction. The underlying idea for this is that collective opinion of many is more likely to be accurate than that of one. The outcome of each of the models is combined, and a prediction is made. The outcome can either be combined using average or the outcome occurring the most, or weighted averages. Ensemble Learning attempts to find a trade-off between variance and bias. The three most common methods of Ensemble Learning are Bagging, Boosting and Stacking.



Although there are many other Machine Learning algorithms, these are the most popular ones. If you're a newbie to Machine Learning, these would be a good starting point to learn.

Which are the top Machine Learning algorithms do you think every Data Scientist should be having in their toolbox? We would love to know which are your favourite *ones.*

Machine Learning in R

Introduction

R is a programming language used in statistical computing and visualisation, which runs on several platforms. It's an open source language with ready-to-run "binaries" available for different platforms. Its source code can be downloaded and compiled for other platforms. R can be downloaded from CRAN (Comprehensive R Archive Network) or http://www.r-project.org/

R environment

It has an easy and quickly adaptable environment. So, it allows users to visualise and manipulate data, run statistical tests, calculate, and apply machine learning algorithms. Apart from this, R has a few other benefits:

- Effective data handling and storage facility
- Gives you access to rich statistical learning packages developed by top researchers.
- Most importantly, it's free.
- Graphical support for data analysis and visualisation

Getting Started

You can directly open R and start working on it, but it is advisable to first download the R-Studio which is a free and open-source integrated development environment (i.e. IDE) for R.

Now, figure out the current directory *getwd()* and change it using *setwd()*.

```
> getwd()

> setwd("C:\...")
```

Installing and loading packages

There are not many machine learning algorithms incorporated in R and so most of the times you need to download the package and then call them into your workspace.

```
>install.packages("<package name>")

> library("<package name>")
```

Need Help?

Well R has a great documentation. If you get stuck on some command or function, just type “?<function name>” in R console and there will be an output window showing results for the details of that function.

Basic Functions in R

In this section, we will cover how to read data, perform basic operations and visualize data.

I. Get your data and read it!

To load the data, you can either search it on the web, or look into your local disk, or use built-in datasets.

a. Reading data using URL

```
>url <- “<enter a URL address with CSV data>”
> data <- read.csv(url, header = TRUE)
```

b. Reading from your local drive

```
> data <- read.csv(“<location of your file>”)
```

Using built-in dataset. Here you can directly call or look into the dataset by simply typing the name of the dataset into R-console:

```
> iris
```

For reading data tables from text files, we use *read.table()*. Want to see its working details? Go ahead and call for help! *“?read.table”* .

II. Know your data

You can’t simply jump into building the model by just reading the data. It is very important to analyze the data first.

One of the first steps in data exploration is inspecting your data. However, there are myriad ways to do that but I’ll mention here only a few of the profound ways to understand your data.

```
> str(data)

> summary(data)

> head(data)
```

There are other statistical commands too for computing the mean *mean()*, variance *var()*, standard *sd()* deviation etc. to help you evaluate the dataset.

III. Visualization

One of the best things about R is that it has great graphical properties with lots of rich libraries to let us show off our visualization skills. Let's find out how:

We will be taking iris dataset for our example and try out these commands to see the awesome visualization.

```
> plot(iris$Petal.Length, iris$Petal.Width, main="Iris Data")
> plot(iris$Petal.Length, iris$Petal.Width, pch=21,

bg=c("green","red","blue" [unclass(iris$Species)],main="Iris Data")
```

Now, let's play around with ggplot:

```
>ggplot(iris, aes(x = Sepal.Length, y = Petal.Length)) +
geom_point(aes(color = Species))
```

You have to explore this *ggplot* to understand the underlying notion behind the grammar of graphics. Also, try *ggvis,* which is again an awesome visualisation package to work with.

Machine Learning

The CRAN repository has more than 10,000 active packages. Let's discuss a few of the important machine learning packages and learn how to implement them.

1. Linear regression

It is one of the very basic statistical learning approaches. *lm* is the function we use to generate our model. Since this is a built-in function you don't need to install any package for it.

Let's name the dependent variable as 'Y' and independent variables as x1 and x2. We want to find the coefficients of a linear regression and generate a summary for it.

```
>lm_model <-lm(y ~ x1 + x2, data=as.data.frame(cbind(y,x1,x2)))
> summary(lm_model)
```

2. Logistic Regression

Its syntax is very similar to that of linear regression, as here also you don't need to install any package.

```
> glm_mod <-glm(y ~ x1+x2,
family=binomial(link=”logit”),

 data=as.data.frame(cbind(y,x1,x2)))
```

3. K- Nearest Neighbour Classification

K-nearest neighbours is a simple algorithm that stores all the available cases and classifies the new ones based on stored and labelled instances i.e., by a similarity measure (e.g., distance functions).

In order to build the classifier for *knn*, you need to divide the dataset into testing and training and then add some arguments.

```
>library(class)
> knn_model <-knn(train=X_train, test=X_test,

cl=as.factor(labels), k=3)
```

4. Decision Trees

It is one of the most simple and frequently used classification algorithms. Let's see the syntax using the same formula we used in linear regression.

```
>library(rpart)
> tree_model <-rpart(y ~ x1 + x2,

data=as.data.frame(cbind(y,x1,x2)), method=”class”)
```

5. K-means Clustering

K-means is one of the simplest unsupervised learning algorithms that solves the well-known clustering problem.

If you have a matrix “*Z*,” and “*n*” is the no. of clusters then the syntax goes like this:

```
> kmeans_model <-kmeans(x=Z, centers=n)
```

6. Naïve Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with an assumption that features are independent.

```
>library(e1071)
> nB_model <-naiveBayes(y ~ x1 + x2,

data=as.data.frame(cbind(y,x1,x2)))
```

7. SVM (Support Vector Machines)

SVM is a supervised machine learning algorithm which can be used for both classification and regression problems.

Let “Z” be the matrix of features, and labels be a vector of 0-1 class labels. Let the regularization parameter be “C”.

```
>library(e1071)
> svm_model <-svm(x=X, y=as.factor(labels),

kernel =”radial”, cost=C)

> summary(svm_model)
```

There are many parameters like kernel type, the value of “C” parameter, etc. for that you have to dig deep into the algorithm and check the documentation regarding SVM.

8. Apriori Algorithm

Apriori algorithm is an algorithm for association rules to analyse the technique to uncover how items are associated with each other.

Here our dataset must be binary incidence matrix and also we have to define the support and confidence parameter for this algorithm.

```
>library(arules)
> assoc_rules <-apriori(as.matrix(dataset),

parameter = list(supp = 0.8, conf = 0.9))

> inspect(assoc_rules)
```

9. Random Forest

Random Forest is one of the most powerful ensemble techniques in machine learning algorithms for both classification and regression problems. The syntax for this algorithm is very similar to that of the linear regression model. Let's say you have a data with seven variables where *Y* is the response variable and *X1, X2, X3, X4, X5 and X6* are the independent variables.

```
>library(randomForest)
> fit <- randomForest((Y) ~ X1+X2+X3+X4+X5+X6,

data=train,

importance=TRUE,

ntree=500)
```

I hope this post was simple to understand. The sole purpose of this article is to get you acquainted with R as a tool to get you started with the basic machine learning algorithms. This article alone won't do justice to these awesome algorithms. So, you have to dig deeper into each of these algorithms. In addition, you should learn how to predict and evaluate your result, and most importantly practice a lot by taking up different kind of problems.

How to choose the best ML technique for your dataset?

Python vs R. Which one should you use?

-Comparing Python and R for Data Science:

Today, we will discuss the two most popular languages used in the field of Data Science- Python and R. When one makes the decision of learning this exciting field, they often have to choose between one of these languages. Both Python and R have their pros and cons. Hence this decision depends on various factors such as how much programming experience one has, whether they'll be working in an academic or an industry setting, etc.

Python is a general-purpose programming language that is easy to write as well as understand. It reads more like a regular human comprehensible language such as English. Over the past few years, its popularity has been increasing. It now finds uses in a wide variety of fields including web development and scientific computing. R, on the other hand, is a language that is built almost exclusively for statistical computing. Researchers in statistics widely use R.

In this post, we compare the two on eight factors. This should serve as a good resource for making the decision on which one is better for you!

#1. The Start

It is a lot easier to build your first model on R than it is on Python. You take a data set, import it, and use one of the built in libraries to run an algorithm and generate an output. In Python, however, it is harder as you have a number of options (such as the data structure to be used). That being said, after the initial obstacle, Python is easier to learn. This brings us to our next point.

#2. Learning Curve

R has a steeper learning curve than Python. This comes down mostly to syntactical details, on which Python wins by a mile.

It has an extremely expressive syntax that is in some ways very similar to regular English. Most universities teach their introductory programming courses in Python. It's much easier to pick up and get started with Python than with other languages such as C or Java which are relatively verbose in nature.

#4. Community

Python being a general purpose programming language, has a much bigger community of people using it. R has also been catching up though. Its popularity has grown over time with the internet now having many data science examples that are written in R.

#5. Speed

In the beginning, R was fairly slower than Python. But as time has progressed, a majority of R has been rewritten in C, thus making it pretty fast. Python, too, has libraries such as numpy (written in C), fastening operations!

#6. IDEs

RStudio is way popular than any other IDE for R. RStudio makes it easier to manage R libraries. In fact, it is one of the reasons for the rise in the popularity of R in recent times. Python's counterpart to RStudio is Spyder (Though PyCharm is another amazing IDE.)

#7. Data Visualization

As we've already mentioned, R's ggplot2 is a fantastic library for data visualization. It feels natural; it is easy to learn; and is flexible enough to produce any visualization of your choice!
Python has matplotlib as its standard visualization package and a lot of other Python packages also use it.

#8. Industry

Many companies employ a Python-based application stack. So, they prefer to use Python for their Data Science needs as well, since it makes integrating everything together easier, rather than using a completely different language such as R.

Conclusion

Python and R are indeed popular Data Science languages. Choosing from the two can indeed be very tough!

The final verdict- So, if you're new to programming and are looking to learn more in a short period of time, go with Python. If you're looking to work as a Data Scientist in a company, Python is a better option than too! But if you are looking to build statistical models in an academic setting, R may be the one to go with!

4

Data Science Careers- Can you become one?

Data Science Roles

Role #1: Data Analyst

What do they do?

Data Analyst will be called upon for deriving insights from numbers. They are required to process large sets of data and connect that with data actions that can drive business impact.

How much do they earn?

Min: Rs.2, 81,000

Average: Rs.6, 00,000

Max: Rs.12, 10,000

What skills do they have?

- Collect data using scripting languages.
- Explore and analyze datasets with tools like Excel
- Strong analytical, quantitative and problem-solving abilities
- Data modeling, Reporting

What tools do they use?

- Excel, Tableau, Matlab, QlikView, etc.

Role #2: Business Analyst

What do they do?

Business analyst role would require

- Analyzing and solving business problems with focus on understanding root cause
- Designing new metrics and enhance existing metrics to support future state of business.
- Enabling effective decision making by retrieving data from multiple sources.

How much do they earn?

Min: Rs.3, 72,000

Average: Rs.6, 44,857

Max: Rs.12, 50,000

What skills do they have?

- Strong operational business understanding, Project Management
- Advance working knowledge of data mining tools like Excel.
- Understanding of visualization tools such as Tableau, Datazen.

What tools do they use?

- Tableau, Excel, QlikView, etc.

Role #3: Data scientist/ Statistician

What do they do?

A Data Scientist will develop analytics and machine learning models on problems of moderate to high complexity, using statistical analysis software tools such as R, Python, and SAS.

How much do they earn?

Min: Rs.3,78,000

Average: Rs.6,50,000

Max: Rs.16,00,000

Career Path:

- Junior Data Scientist
- Data Scientist
- Senior Data Scientist
- Chief Data Scientist

What skills do they have?

- Machine learning
- Knowledge of algorithms, statistics, mathematics
- Broad knowledge of programming languages such as Python, R

What tools do they use?

- R, Excel, Matlab, Python, Wolfram alpha, etc.

Note:

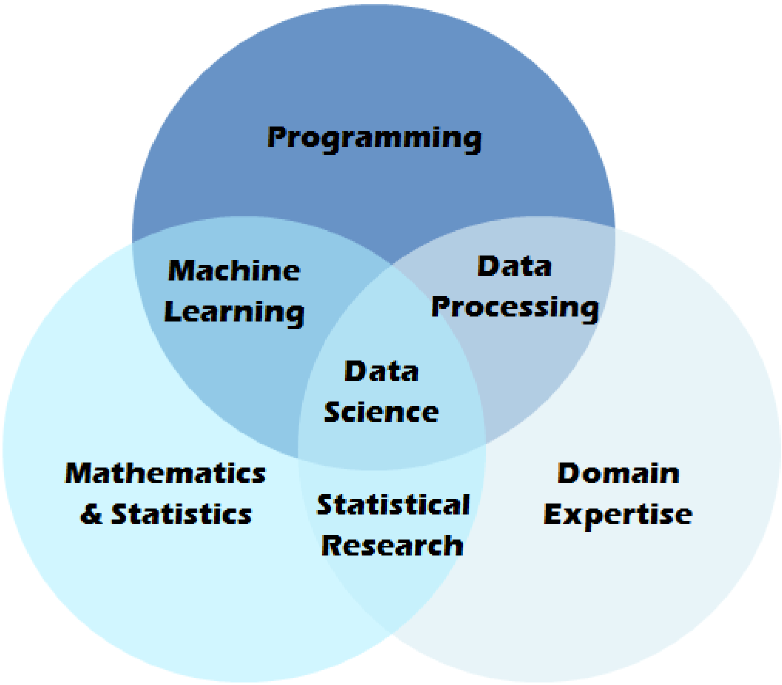
Salaries and job descriptions stated are based on reviews received from several sites such as Glassdoor, Payscale, etc. The job description and salaries are subject to change with companies requirements. Data is being collected from Glassdoor, LinkedIn, Springboard, etc.

What skills are required to be a Data Scientist?

What does being a Data Scientist entail?

The simple answer to this is – Statistics, Programming, and some domain knowledge. These are one of the steps to become a Data Scientist. Seems easy enough on the surface, doesn't it? It is, to an extent. Statistics knowledge isn't what we learnt in school. It isn't studying simple mean, median, and mode formulae (Although, smooth implementation of even these simple concepts can go a long way and provide great insights through visualization!), but studying the subject in detail.

Concepts like Normal Distribution, Random Variables, z-score, correlation and covariance, Linear Regression, Probability and Bayes Theorem are just some of the terms you will come across on a daily basis as a Data Scientist, that concern Statistics.



Python, R and SAS are the programming languages that a Data Scientist uses every day. No, a Data Scientist does not need to know each of these languages extensively. As is often said, don't be the jack of all trades, be the master of one. Because all of these languages offer the same functionality.

Being an expert in one is much better than being intermediate in all. There is a constant debate in the Data Science community about which language is better – Python or R. In fact, a quick Google search on this will show you the millions of posts on the controversial topic! There is no simple answer to this. It depends on which one a programmer is more comfortable in. Personally, I'd recommend Python, as I believe that R has a much steeper learning curve than Python.

Below is the Python code to add two numbers –

```
n1 = int(input("Enter a number"))

n2 = int(input("Enter the second number"))

sum = n1 + n2

print("Sum is" +str(sum))
```

Isn't this simple? Doesn't it feel like regular English? Python also provides excellent easily understandable scientific packages for Machine Learning and dealing with large data sets like Scikit-Learn and statsmodel. The same code in R is quite cluttered and not very easily understandable to someone that is new to programming. If you're new to the field, you should play around with both languages and decide which you find easier.

Having Domain Knowledge involves using the insights derived from applying the previous two steps to make smart business decisions and optimize operations. This is perhaps, the most crucial step, even though it might not seem like much. Unless this step is carried out effectively, the insights cannot accurately be converted to smart business decisions.

In addition to this, a Data Scientist must have the ability to tell a story with the data. To represent the insights visually in such a way that a layman can understand and process them. Essentially, a Data Scientist should be a good storyteller. "*People hear statistics, but they feel stories*" It's as simple as this; the insights gained seems worthless unless they can translate to ways that increase profit margins, i.e., make smart business decisions. Data Visualization tools often used by Data Scientists are Tableau, the scientific Python packages (matplotlib), Excel (Excel is a very powerful for data visualization. It cannot be used on "Big Data", but is an excellent solution for small data.), etc. To put it whimsically, "*if you torture the data long enough, it will confess.*", as said by Economist Ronald Coase.

Can anyone become a Data Scientist?

There is quite a lot of confusion around who can become a Data Scientist. Therefore, let me give you a few interesting examples

- Karthik Rajagopalan of AT&T has degrees in Mechanical, Industrial and Electrical Engineering, and a PhD in Solid State Physics. He is now a Data Scientist!
- Shankar Iyer, a Data Scientist at Quora says "Our data science team at Quora has people with diverse backgrounds, including physics, economics, and chemical engineering. There are indeed some team members with degrees in statistics and machine learning, but it's not a requirement."
- Luis Tandalla, a student at the University of New Orleans took a couple of free Data Science courses on Coursera and a few months later, he scored his first victory in a Kaggle competition hosted by the Hewlett Foundation where he had to devise a model for accurately grading short-answer questions on exams. This, from a student who had no idea what Machine Learning was before he signed up for the courses online!

What background do you need to have to be a Data Scientist?

What do all these examples tell you? That absolutely anybody can become a Data Scientist? While it is inspiring to think so, and the statement is true to an extent, it has a little bit of a glass ceiling. A Data Scientist is someone who has some background in Mathematics and Statistics, Programming, a creative mind to ask the right questions and to use insights for business decisions, and, absolute love for data. Someone who can weave the insights into a story using all the tools at his disposal.

Programming and Statistics can be learnt, but it depends on an individual's background and commitment, how steep the learning curve would be. The love for data, however, is something that cannot be acquired. It's something that is inbuilt in you. If this revolution of mounds and mounds of data changing the world astounds, fascinates, drives and motivates you, then you'll probably be the best Data Scientist out there soon!

Are Statistics and Programming definite prerequisites?

There are several cases in the Data Science community where by being proficient in only Excel or Tableau or any of the visualization tools or having strong business acumen, one has become a successful Data Scientist. While I'm sure this fact brought a smile to many of your faces, I must warn you that it isn't an absolute fact. Everything considered someone with programming and statistics knowledge is much more likely to be a successful Data Scientist than someone proficient in Excel. So, my take on the subject is this. If you're planning to get into the field, start getting familiar with programming and statistics. Programming might be a term scaring a lot of you, but rest assured that in Data Science, in-depth knowledge in complex languages like Java is unnecessary. Start out with R or Python, very easy languages to learn.

As far as Statistics is concerned, again, you do not need to become an expert in the field. Start taking basic stats courses particular to Data Science that will teach you just enough.

What you should do right away to become a Data Scientist!

I believe that the best way to learn anything is by throwing yourself headfirst into it and playing with it. Learning Data Science concepts will be for nothing unless you start implementing your new found knowledge in practical applications. Start by taking any of the plethora of courses offered in Machine Learning online. But, do not wait to finish the entire course, keep revising their examples, familiarizing yourself with the theory. Go to Kaggle or any of the sources that are giving away free data sets, and get working!

A Day in a Life of a Data Scientist

Why has being a Data Scientist become the hottest profession?

Intuitively, it seems like, with the advantage and insight that Data Science provides, Data Scientists should have been around for much longer, helping us make decisions, isn't it? This question can be answered in one word. **Data.** Data is the bread and butter of a Data Scientist. A Data Scientist sleeps, breathes and eats data. The data that is being produced today is quite mind boggling. We have produced more data in the last two years alone than we did since the dawn of time! And this data only increases every day.

To just let the ever increasing data get produced and stay stagnant did not seem useful. Data was mounds of useful information, a tool, a Pandora's Box, full of untapped potential, waiting to be explored. This is around the time Data Science was born. Data Science is so new a field which not even actual Data Scientists can tell you what exactly the job entails!

Data Analytics has been performed on data for years. It has been around for over 20 years. So, why not just use basic Analytics tools? Why introduce an entirely new profession called Data Science? Because, with the evolution and advancement of Machine Learning, a branch of Artificial Intelligence, the kind of insights that were being mined from data were revolutionary! Statistical and Machine Learning models as simple as Linear Regression were providing immensely helpful insights that were vastly improving business decisions.

For example–

- A company may find a pattern with a section of consumers behaving differently. After digging deeper and further analysis, they discover that this subsection of consumers has a similar trait. Now, they can work on ideas to modify the consumer's behaviour or understand what can be done differently to cater this audience.

- Let's take an airline company. The company might want to know things affecting costs, things affecting revenue, and things which will help them attract and retain customers. Fuel is a major cost for any airline, so you might do some data analysis to project the future expenditure and buy more fuel when the price is low. Regarding growing the revenue, consider that customers are usually more sensitive to price when buying tickets for personal travel compared to business, as the fare comes out of their own pockets. Here, you might explore the opportunity to attract new customers by giving low-cost fares to popular vacation destinations and offering higher-cost fares with offers that suit the business travelers going.

- A website might want to know about a metric such as, how long do people spend on their site and then find certain features that are correlated with it. Then emphasize those features to boost metrics. They might also ask different questions to incorporate few changes. Like, if they can build the findings into the product features, so that if a person is looking for a review of a restaurant, they can be prompted to review it from the homepage itself. Both how they decide which restaurant to show you, because you may have looked at several places and user interface elements have a huge impact on whether you write a review.

A Data Scientist's Day

Data Science being a new profession, even Data Scientists wouldn't be able to tell you what their job entails. If you asked Data Scientists at Facebook, Yahoo, GE, etc. what exactly they do throughout the day, their answers could probably differ a little.

However, we have tried to show you how a possible Data Scientist's day could look like.

The Mornings

A Data Scientist would start off his day with a steaming cup of coffee. Nothing works better than to get those grey cells activated than a rush of caffeine! Creativity is a huge part of a Data Scientist's job. It isn't just about crunching numbers. A Data Scientist is often compared to an artist. He weaves a story out of the insights provided in such a way that, even a layman should be able to understand i©2016 UpX Academ

What's the Breakfast Menu?

At breakfast, the popular choice is eggs and bread with orange juice and fruit. Yes, it is pretty routine. A Data Scientist is human after all!

Newspapers

A Data Scientist would then go on to read the newspaper to stay informed about the world's goings on. His job involves staying on top of everything occurring in the world, since Data Science can be applied to practically every industry, be in banking, sales, finance, education, etc. from, to obtain insights. After this, he'd probably move on to read some mainstream Data Science magazines and journals. Keeping himself updated on the new research happening in the field every day (And mind you, that is a LOT! A new technology or tool emerges practically every day.) is vital. It is a field so volatile, (in the positive sense) that keeping up is crucial. Are you an aspiring Data Scientist? Then these journals are for you!

- IEEE Transactions on Knowledge and Data Engineering
- Journal of Data Science
- EPJ Data Science Journal
- Information Visualization
- Journal of Machine Learning Research
- Predictive Modelling News
- Transactions on Machine Learning and Data Mining

At work

When a Data Scientist gets to work, the first thing he would be is to communicate with members of the team about which stage the product is in the right now. Based on this, he'd put on his thinking hat and search his brain for WHAT questions should be asked at that particular phase. Because remember, not asking questions is better than asking the wrong question! The questions should maximize the advantage that Data Science would offer at that point to the highest possible extent.

Now, that the right questions have been asked, it's time for him to get his hands dirty and dive into the first step of predictive modelling – Cleaning and organizing the data. Research has shown that this is the least preferred part for a Data Scientist. Compared to his usual work, it is rather dull. Don't get me wrong, though; it is one of the most important parts of data analytics. For example, while performing Linear Regression, if two data points are outliers and they are not "managed" or "cleaned", they could jeopardize the entire model!

The next step a Data Scientist would take is a crucial one. He now has to find the right model. This depends on a variety of things. The "kind" of data, what questions are being asked, and what predictions are being sought. Not all models work for all data. Some models are fit for specific data.

How do they work?

Once the appropriate model is selected, a Data Scientist applies the model. And then, (hopefully), the model gives us the appropriate results. One of the most important jobs that a Data Scientist performs, and one of the most crucial parts of a Data Scientist's day, is understanding and interpreting the results that are obtained after running the model. The results are often a garbled bunch of numbers with random letters in between. Probably only a Data Scientist will truly be able to understand what these results mean, and interpret them to mean something that a company can use to improve operations.

Once provided with the results, insights can be given by a variety of people. Yes, a Data Scientist can usually provide the best insights, but often, a useful insight is caught by someone's eye on the management team, since they deal with the business decisions after all. However, interpreting the results of a model as it is probably beyond the scope of the management.

This is where the next part of a Data Scientist's day comes in. It is the job of a Data Scientist to weave the results into a story (See why they are called artists?), probably through visualization tools that can be understood by anyone that sets their eyes on it (At least, the management!).

This process isn't standalone. It works in a constant feedback loop. Once these insights have been applied, it is the job of a Data Scientist to ask, has used the insights provided the expected results? Are they not as satisfactory? What other parameters can and should be considered while revamping the model? In this way, constant optimization is exercised to make the system as efficient as possible.

Back home

Now that a Data Scientist has performed his official duties, he returns home, (maybe hits the gym?), and since he is so passionate about the budding field, reads up on it from many informative books. As they say, when you love your work, work stops feeling like work! Some of the books that are commonplace in a Data Scientist's library are-

- Data Science for Business by Foster Provost and Tom Fawcett
- The Elements of Statistical Learning by Trevor Hastie, Jerome H. Friedman, and Robert Tibshirani
- Python for Data Analysis by Wes McKinney
- Data Science in the Cloud by Stephen F.Elston
- Statistical Inference for Data Science by Brian Caffo
- An Introduction to Statistical Learning by Robert Tibshirani
- Machine Learning for Hackers by Drew Conway & John Myles White
- Agile data science by Russell Jurney
- Natural Language Processing with Python by Steven Bird et al.

A Data Scientist then probably just kicks back with a glass of wine and some healthy dinner (No, this isn't a prerequisite to becoming a Data Scientist!), and relaxes. Quite a neat little day, isn't it? The best part is there is no monotony in the job. Every day brings new challenges, every day a new creative thinking hat needs to be put on, and insights need to be mined. The future of the company probably depends on it. But, no pressure!

At the movies!

A Data Scientist is someone that deals with Mathematics and logic throughout his day. His mind is fine-tuned and conditioned to movies that involve these areas. Apparently, Data Scientists have no concept of “Don’t take your work home with you”! Some of the movies that are on every Data Scientist’s must watch list are –

- A Beautiful Mind – A true life story where a schizophrenic Maths professor called John Nash is the protagonist. Represents how powerful an impact Mathematics has on his life. He goes on to win the Nobel Prize!
- The Imitation Game – Alan Turing, the protagonist, builds the first computer to crack the German Enigma Machine to eventually execute a series of smart decisions and win WWII. Interestingly enough, it is said that this was around the time discussions about Neural Networks and using data originated among these brilliant British mathematicians!
- Moneyball – Based on a true story, Moneyball deals with using Analytics to find the most optimal sports team by assessing each player’s value.
- 21 – A group of MIT students travel to Vegas every weekend to play Blackjack. They counted cards, and won thousands of dollars, eventually getting caught. At the foundation, the film focuses on how powerful mathematics and probability are.

Conclusion

It might seem like all work and no play, but most Data Scientists have quite a relaxed life. If you’re thinking about getting into the field, read the mentioned magazines and start on the books, get familiar with various Machine Learning algorithms, and most importantly, get your hands dirty by downloading any of the various data sets from the plethora of sources online, and working on them. You’re on the path to becoming a successful Data Scientist!

How to become a Freelancing Data Scientist?

Want to become a Data Scientist? But you also don't want to work only less than one company? Would you explore yourself in the different job roles of Data Scientist?

If your answers are yes, then you are on a right way to be a freelance Data Scientist. Don't know how to find a job in this field as a freelancer?

In this blog, we will answer you how to become a freelancer Data Scientist.

Freelancing is in every field of journalism to technology. Though freelancing in the technology field is at an early stage, soon it will boom the market. To begin with, what is freelancing and how it is relevant to Data Scientist job? In simple terms, a freelancer is a person who can contribute to a company without being a part of it. A *Freelance Data Scientist* is a Data Scientist who works as an independent employee in different companies on a particular project.

Why be a Freelance Data Scientist?

Freelancer has a few advantages over others –

- Flexible working hour
- High pay
- Choose area of your strength
- Work as a free bird
- No place restriction

How to make your portfolio noticeable?

Step 1: Create your online profile on *Upwork* & *Toptal*

***Upwork*–** It is an online professional business platform where you can make your profile to connect and collaborate with business employers. It is the world's largest freelance marketplace which has recently collaborated with popular freelance platform *Elance-oDesk*. So by making the online profile, you can list yourself among the freelance Data scientists where your profile becomes noticeable by the employees.

How does it work?

1. Register yourself on Upwork. Click here.
2. Create an outstanding profile including your professional skills, experience, education, projects, etc.
3. Upload your friendly looking profile picture.
   After profile! Now it is time to get hired by the clients. Get started by sending the clients a compelling application. If your profile is highly impressive, then chances are more to be called for an interview. So, make sure your profile is-

- Professionally well-written
- Appropriate for the work you're seeking
- Presents your skills and strength

Toptal– It is an online platform where companies hire the software engineers and designers in need of freelance talent. It includes a series of process to get hired by the companies including screening and interviews. However, only 3% of applicants are included in the Toptal directory, if you are in the directory, you will get the world-class opening as a freelancer.

How does it work?

1. Register yourself on Toptal. Click here
2. Prepare your profile including all your relevant skills, experience, education, project, etc., and submit your application.
3. Update your profile picture.
4. For the acceptance of your application at Toptal you need to go through a live screening for language, technical skills, and an interview by the Toptal engineers.
5. Some test projects will be assigned to you which gets over in one – three weeks.
   Note: – The completion of application process takes approximately 1 month.

Step 2: A few projects on Github or your blogs

Putting a few projects on Github will increase the application approval chances. It will be easier for the hiring managers to find these projects and to evaluate your skills. Now, let's see what kinds of projects can be helpful to be a freelance Data Scientist.

1. **Data cleaning project** – It is the huge part of any Data Scientist job, where in this project you take a messy data and clean it up, showing the analysis. This project displays the skills that you can take disparate datasets and make sense out of it. Doing projects like this demonstrates the skills of taking data from many sources and combining into a single dataset.

2. **Data storytelling project** – This project can be an important piece of your portfolio and seek attention to the hiring managers. It has a significant impact on business value due to the ability to extract insights from data and persuade others. However, it demonstrates the skills of taking a set of data and telling an exact narrative of it.
3. **An end-to-end project** – Broadly, speaking it is a fusion of Data cleaning and storytelling. Whereas to build an end-to-end project the main components are:-
   - To understand the context
   - To explore the data and figure out the nuances
   - Write a high-performance code so that runs quickly and uses minimal resources
   - Document the installation of the code well so that it can use by others.
4. **A Data Science competition projects** – Already, got your hands dirty on a few Data Science Competition projects. Then you need to upload the projects to get your application accepted by the freelance hiring platforms like Upwork and Toptal.

Some Data Science competition platforms

1. **Kaggle** – It is a leading platform for Data Science competitors to utilise their skills and potential to crack some datasets. Here, the companies and the researchers post their datasets; the statisticians mined data to compete with others to produce the best models. Taking part in such data mining competitions can be proved helpful for you to get your application noticed among the crowd.

2. **DrivenData** – It is the world's biggest social challenge platform where they bring edge-cutting practices in data science and crowdsourcing. They host online competition which lasts for 2-3 months. Here, a global community of data scientist comes to compete over others by giving the best statistical models for difficult datasets that make a difference.

Step 3: Growing a network/ build connections

Since it is said, the internet is the best way to grow your network. There are some communities' where newbie of Data Scientists comes to take help on solving some tough problems. Therefore, answering that problem and the continuous practice will grow your network in these communities. Once you start answering and your peers find you helpful in solving their queries. They will reach out to you with offers to collaborate with them on a freelance work.

Some communities to grow your network

1. **GITTER**– It is a chatting and networking platform which helps to manage, connect, and grow communities' network of developers. It creates a single private chat room for messaging, content, and discovery. Most of the conversations are technical, focusing on debugging and sharing code bases.

2. **Data Science Central** – It's an online resource for Big Data practitioners. It provides a community experience in Analytics, Data Integration, and Visualization. Here, the forum dedicated to helpful resources, Tutorials, webinars, tools, latest technology, and jobs opportunities.

3. **Data Science Stack Exchange** – It is a site where you exchange question & answer on Data Science, Machine Learning, R, Neural networks, and Python. The experts can be a Data Scientist or a Machine Learning Specialist. A Freelance Data Scientist can reap the benefits by becoming an expert.

4. **Meetup** – It is an online social networking portal designed to facilitate offline groups meetings in a way to manage it in various localities around the world. It helps to find a group of peers to engage in a data science network. The user enters the postal code, and the website helps them to locate a place for a meetup. In this way, you can connect to the professionals or the individual groups themselves.

Conclusion

Since Freelancer has expanded the market over the past decades, the preference of independent contracts over employees has increased. It has also increased the level of competition for finding work. Once you become a freelancer, you need to be proactive and utilize your potential & resources in all possible ways.