Data Science Cognizance

Prithvi Macherla, Dr. Atif M. Farid, Saith Kumar Gundu, Sujitha Gali, Sushma Venkatesh Reddy and Vidhya Sagar Budagam

***Abstract*—"Without data you are just another person with an opinion"-Andreas Schleicher. This quote clearly depicts the importance of data in the real world. Unless we extract knowledge from data we will not know its valued implications and this extraction of knowledge involves various domains like computer science, math and statistics, and subject matter expertise which are framed within the field named Data Science. Searching for real data scientists are considered as rare as unicorns due to the number of domains an individual has to expertise. As the need for data scientists erupt fake data scientists have appeared and it is getting tough to find their validity. Here are some of the most important questions which a real data scientist should know and we have tried to make justice to them.**

Recommendation Engine

A **recommendation engine** is a feature (not a product) that filters items by predicting how a user might rate them. It solves the problem of connecting your existing users with the right items in your massive inventory (i.e. tens of thousands to millions) of products or content. Though the research is done in this particular field from many years, the interest remains the same as it constitutes a problem rich area and abundance of practical applications. It help users to deal with information overload and provide personalized recommendations, content and services to them. Examples of such applications include recommending books, CDs and other products at Amazon.com, movies by Netflix, etc. Moreover, some of the vendors have incorporated recommendation capabilities into their commerce servers.

Need: As the definition conveys, it is a solution to large amounts of good data. Helps in reducing cognitive load on users (browsing in interesting manner where they get what interests them) and to improve quality control (to define business rules like amazon can suggest products 10% based on season and 90% based on previous browsing history).

Overview of recommendation system

Candidate generation – automatically identify items of interest to users (focus of talk)
Filtering – Filters: near duplicate, already seen, dismissed
Rank – Order recommendation based on temporal (seasonal or based on time), diversity (keep relevant but do not show same type of products to the user – not all Harry Potter books also include lord of rings), personalized
User Feedback – Track user feedback such as their dislikes, clicks etc.

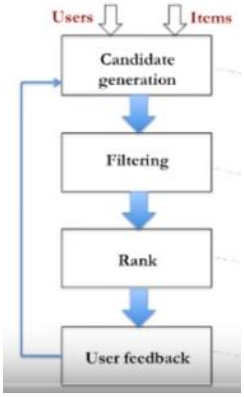

Fig 1. Overview of recommendation system

Candidate generation is the heart of the recommendation system so recommendation stands on it. Some of the approaches that can be utilized are discussed below.

Naive Recommendation System:
In this type of system, rating of each item is aggregated and then item with maximum ratings is recommended. But here the problem is not everyone like the same products so historical information of each user is very important.
Recommender systems typically produce a list of recommendations in one of two ways - through collaborative or content-based filtering. **Collaborative** filtering approaches build a model from a user's past behavior (items previously purchased or selected and/or numerical ratings given to those items) as well as similar decisions made by other users. This model is then used to predict items (or ratings for items) that the user may have an interest in. **Content**-based filtering approaches utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties. These approaches are often combined and they are called **Hybrid** Recommender Systems.

Collaborative filtering can be done in two ways, one is by comparing likelihood based on product and other is based on user. When recommendation is based on product, it compares the closeness among a group of products and then suggest which product to offer to a given user. In user-based recommendation, comparison is done between different users to find similarities and then based on similar users, recommendation of a product is made which he/she might buy.
Two forms of Collaborative filtering (CF) are discussed below.

Item - based: Prediction of user's rating for an item i based on his ratings for other item/items. Given a user u with I(u) preferred items. One of the approach to recommend based on scores of each item is as shown below.

$$\text{score(i, u)} = \sum_{j \exists l(u)}^{1} rating(u, j) * sim(i, j)$$

For example, here we try to predict which movie to recommend for a user.

Given user ratings for Harry Potter and The Matrix movies,

	Harry Potter	The Matrix
Rating	0.8	0.3

Table 1. User rating

Similarity ratings for The Chronicles of Narnia (N), Star Wars (S), Harry Potter and The Matrix movies are as given below.

Item	Harry Potter	The Matrix
The Chronicles of Narnia (N)	1.0	0.3
Star Wars (S)	0.2	0.8

Table 2. Similarity rating

score(u,N) = 1.0*0.8+0.3*0.3=0.89
score(u,S) = 0.2*0.8+0.3*0.8=0.4

From the score we can see that 'The Chronicles of Narnia' (N) is recommended compared to that of 'Star Wars' (S).

One more apprach to recommend items is based on cosine similarity between items.

Cosine similarity between items: Following are some of the assumptions made by this approach.

- Items are represented as u-dimensional vectors over user space
- Similarity is cosine of the angle between two vectors
- Score ranges between 1 (perfect) and -1 (opposite)

	U1	U2
A	0.8	0.45
B	0.4	0.8
C	0.3	0.3

Table 3. Cosine between A, B and C

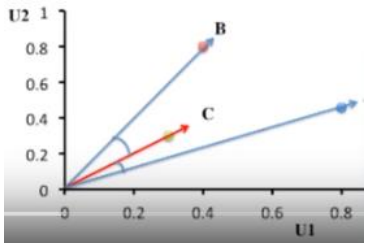

Fig 2. Cosine angle representation

Cosine similarity:

$$sim(i, j) = \frac{I * J}{||I|| * ||J||}$$

Cosine doesn't take care of magnitude like for example if user likes it very much is equal to that of user who just likes it. If there is a denser vector then value is going to be biased towards it. Hence the problem is not solved on its entirety.
Example: U(j) is the set of users who has seen harry potter and then when this is compared to other movie then the intersection is going to be non-zero or nonempty set but when taken a movie that is less popular then we get smaller set so this is where the bias is created.

Magnitude-aware measure:

$$sim(i, j) = \frac{U(i) \cap U(j)}{\sqrt{|U(i)| * |U(j)|}}$$

To check based on click then magnitude-aware measure works better.
User-based: One of the best approaches in user-based recommendation is K-Nearest Neighbors (KNN).
In this approach, users are clustered after representing them as feature vectors. To cluster it will use the similarity measure and it can also use cosine similarity measure. Users above some similarity are placed in one group. This is hard cluster, which means that the users cannot belong to more than one group.

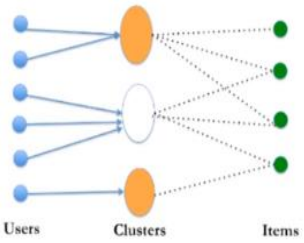

Fig 3. User-based clustering

Data Collection methods:
As recommendation system learns from data, it is very important to collect data and it can be done in two different ways:

Explicit feedback: Explicitly collecting the feedback from users. Ex: ratings, dismiss
Implicit feedback: Collecting the information implicitly based on user behavior. Ex: number of views, purchases

Evaluation: One of the measures to find the accuracy of the system is using RMSE.
Root Mean Squared Error (RMSE): Differences between value (sample and population values) predicted by a model or an estimator and the values actually observed. The square root of the mean divided by average of the square of all of the error.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Consider the prediction by the system is as below.

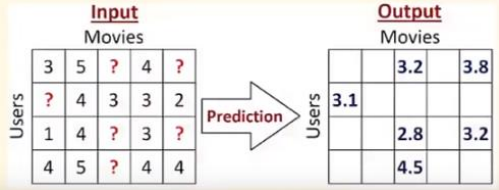

Fig 4. Movies prediction

Now to make sure if the recommendation is working fine, we should hide some of the values known in our dataset and then predict the values for those. Then we should compare the predict values with that of the correct values. If the RMSE value is less it implies that the recommendation system works well.

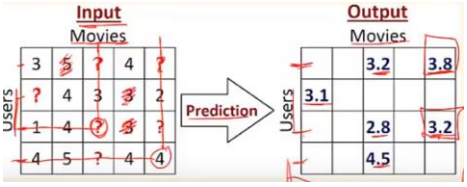

Fig 5. Root mean square error validation

Challenges:

There are many challenges involved with respect to recommendation system. Few of the important ones are noted below.

Data Sparsity: It is caused when the users rarely purchase, rate, or click.

The more you see the less you know: Increasing users or items increases the dimensions we need to learn.

Cold Start Problem: One of the biggest challenges can be that there is not sufficient historical data at the start. For example consider FOUNDD, a young Berlin-based startup for movie recommendations. It did not have a long purchase history like Netflix NFLX -6.00%, thus the algorithm will not be able to recommend anything useful in the beginning. Fully aware of that issue, the founder Lasse Clausen created a "hot or flop" page in the beginning. Each customer has to rate 10 movies before the system begins to recommend anything. However content based filtering can be utilized when we do not have historic data. Where user needs to input information on their personal tastes, though not on the same scale so cold start problem doesn't affect this.

Ratings or other metrics may contain biases: Some of them are: Harry Potter 1 is more popular than 2 because 1 already has more information collected. Ads or documents often suffer from this problem. First sight, fatigue, seasonality are some of the other issues.

Correlation between nearest neighbors: Harry potter sequels like if you see part 1 then no guarantee that you see part 2 so both are independent.

Scalability and recommendation accuracy not production - friendly

No Surprises: In case there were sufficient data, then the second problem from recommendation engines – if executed badly – is that there might be no surprises. An advice to read the book Harry Potter 3 after you looked at Harry Potter 6 might not be all to insightful. It just states the obvious. Recommendation engines work best, therefore, in the long tail of the data – because here are the unexpected results.

Solution to some of the challenges:

Dimensionality Reduction

Say every user who likes "Harry Potter" also likes "The chronicles of Narnia"

Generalize movies into generic latent semantic characteristics – {ex: fantasy, novel-based movies, etc.}

- Reduces dimensions to track and improves scalability.
- Reduces data sparsity and improves prediction accuracy.

Singular value decomposition (SVD): It is the most commonly used dimensionality reduction method.

- Represent datasets from multiple users and items into a matrix
- Apply SVD and pick k-dimensions to reduce our datasets
- Map new users into this low k-dimensional space
- Compute similarity between users in this space
- Provide recommendations based on similarity users

It can be computed easily using R. However basic structure of its working is as given below.

Takes an m*n matrix (Ex: m users and n movies) and produces three matrices:

S: a m*n diagonal matrix with non-negative numbers

U: a m*m matrix

V: a n*n matrix

$$M = U * S * V^T$$

SVD collapses the matrix to a smaller matrix retaining important features. Pick k dimensions and chop off the matrixes. S stores the information about decomposition.

A user X comes in with some ratings in the original feature space, map it to k-dimensional vector.

$$B = B^T * U * S^{-1}$$

Computing SVD:

$$A = \begin{bmatrix} 8 & 0 \\ -2 & 1 \end{bmatrix} \quad A^T = \begin{bmatrix} 8 & -2 \\ 0 & 1 \end{bmatrix} \quad A^T.A = \begin{bmatrix} 64 & 0 \\ 0 & 5 \end{bmatrix}$$

$$A^T.A - cI = \begin{bmatrix} 64-c & 0 \\ 0 & 5-c \end{bmatrix} \quad |A^T.A - cI| = (64-c) * (5-c) - 0 = 0$$

Fig 6. Singular Value Decomposition

Suppose C1=8, C2=1 then

$$S = \begin{bmatrix} \sqrt{8} & 0 \\ 0 & \sqrt{1} \end{bmatrix}$$

Explore – Exploit Strategy: This is in order to solve the **bias** problem. Recommend items be sent at random to a small randomly chosen users. Ensures each user's information on each item. Full randomization may not be possible.
Recommender systems need to choose between:

- Exploiting a model to improve quality
- Exploring a new item to reduce uncertainty.

False positive and false negative:

Retrieved items that are not relevant are called **false positives**. Mistakenly rejecting a null hypothesis. Also called **Type 1** error.

Relevant items that are not retrieved are called **false negatives**. Failing to reject a null hypothesis or accepting a null hypothesis without support. Also called **Type 2** error. Statistical power increases when type 2 error decreases.

Both types of errors are problems for individuals, corporations, and data analysis.

	Patient develops cancer within 1 yr	Patient does NOT develop cancer within 1 yr
Cancer suspected based on mammogram	True Positive	False Positive
Normal mammogram	False Negative	True Negative

Fig 7. False positive and false negative

Here while testing for a cancer condition, a false positive (a condition being detected when none exists) causes unnecessary worry or treatment, while a false negative (a condition going undetected when it is present) gives the patient the dangerous illusion of good health and the patient might not get an available treatment.

Based on the real-life consequences of an error, one type may be more serious than the other. In many applications there is a trade-off between these errors, particularly when classifying based on a threshold: a lower threshold for positive results yields more false positives but fewer false negatives.

For example, in high-cost or life-and-death situations, like space exploration or military equipment, the cost of defects is very high (a mission fails or someone dies), and thus one has very strict tolerances. Thus NASA engineers would prefer to waste some money and throw out an electronic circuit that is really fine (false positive) than to throw out less but use one on a spacecraft that is actually broken (false negative). In this situation false positives use more money but increase mission safety, but a false negative would save some money but would risk the entire mission.

On the other hand, in many legal traditions there is a presumption of innocence, as stated in Blackstone's formulation that:

"It is better that ten guilty persons escape than that one innocent suffer",

That is, the false negatives (a guilty person is acquitted and escapes) are far preferable to false positive (an innocent person is convicted and suffers). This is not universal, however, and some systems prefer to jail many innocent, rather than let a single guilty escape – the tradeoff varies between legal traditions.

R, SAS and Tableau

Tableau: It produces a family of interactive data visualization products focused on business intelligence. It helps people to see and understand data. Tableau allows quickly connecting, visualizing, and sharing data with a seamless experience from the PC to the iPod. Allows creating and publishing dashboards and sharing them without requiring any programming skills. It is very easy to use and it is one among the top business intelligence tools. Some of the important features of Tableau are Metadata management, Mobile-ready dashboards, Security permissions at any level, Tableau public for data sharing, automatic updates etc.

R: It is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R is freeware so it is a security concern for large companies. R has good books by example, online user documentation, R mailing list, and R meet-ups. Users rely on what others put out there about the software. There is disconnect in the world-wide user group because the developers are so spread out. Packages are not written by the R Development Core-Team therefore they are not well polished and some could have questionable validity. It is also difficult to direct an issue to a particular person or support system.

SAS (Statistical Analysis System): It is software suite developed by SAS Institute for advanced analytics, multivariate analysis, business intelligence, data management and predictive analytics. SAS is a software suite that can mine, alter, manage and retrieve data from a variety of sources and perform statistical analysis on it. SAS has extensive online documentation, expert technical support, professional training courses, and many excellent books in press, and a tight knit user group and web based community. Problems can be addressed to SAS directly via tech support that replies very quickly and will work with the user to solve the problem.

SAS and R dominate the programming choices in the statistics field. Industry is heavily dominated by SAS whereas R is used widely in academia due to being free and open source software structured around users being able to write and share their own functions.

Some of the pros and cons of these are addressed below.
We need to understand that there is no clear winner. Both packages have their strengths and weaknesses. They need to co-exist. Users need to keep their technology toolkit up to date. SAS and R have some great websites for learning about new technology advances. The winning solution would be to use both technologies to leverage data manipulation and analysis.

Data Scientists

Larry Page: He is ranked as the top data scientist by Forbes and he is also a powerful person as according to it. Page built Google with Sergey Brin while studying for a Masters in Computer Engineering at Stanford University. He's the inventor of PageRank, an algorithm used by Google that ranks the popularity of webpages in their search engine results, and is a pioneer of data analytics.

Jeff Hammerbacher: Hammerbacher with DJ Patil were the first individuals to form Data teams at Facebook and LinkedIn. They even coined the term "Data Scientist". His early support for the open-source database project, Hadoop, has allowed the company to make huge profits in targeted advertising. Now he is a chief scientist at Cloudera and has commented on the importance of data in industries like oil & gas, retail, and life sciences.

Dhanurjay "DJ" Patil: Currently Chief Data Scientist at Greylock partners. He has held a number of high profile jobs - ranging from Chief Security Officer for LinkedIn to advisory work for the U.S. Department of Defense. He employed social network analysis to help anticipate emerging threats. He used NOAA to improve numerical weather forecasting. He famously described his role at LinkedIn as "making big data small" and was involved in creating the "people you may know" feature on the website. Has won numerous awards, including 2014 Young Global Leader by the economic Forum, Forbes – The World's 7 Most Powerful Data Scientist and CNN - 36 of tech's most powerful disruptors.

Todd Park: CTO, Department of Health and Human Services. Park is leading the charge to transform American healthcare into a data driven business. From medical diagnostics to insurance reimbursement to community health statistics, he is finding ways to use data to make healthcare more effective and affordable.

Startups:

Confluent:

Founded: September 2014. CEO: Jay Kreps. Capital raised: $30.9 million.
One of the biggest challenges in big data is working with high volumes of real-time streaming data. One technology that's catching on for tackling the problem of streaming data is Apache Kafka, an open-source, highly scalable messaging system that can be used in conjunction with other technologies to provide real time analysis and rendering of streaming big data.
Confluent was launched to provide technology and services that help businesses adopt and use the Kafka system. Based in Mountain View, Calif., Confluent was co-founded by Jay Kreps, Neha Narkhede and Jun Rao, who created Kafka while working at LinkedIn and then contributed to the Apache Software Foundation and spun out as a separate company.
As more businesses implement Internet-of-Things systems to collect and analyze huge volumes of streaming data, Kafka could prove to be a critical technology. And Confluent could play a major role in its adoption.

AtScale:

Founded: 2013. CEO: Dave Mariani. Capital Raised: $9 Million. It was created to solve the problem of using familiar business intelligence tools and interfaces, such as SQL and Tableau, with big data storage technologies such as Hadoop. The goal is to be able to perform analysis with the data in place, rather than moving it to a specialized analysis tool. Customers include Aetna, Comcast, and Cloudera.
While more corporate data is being collected and stored in Hadoop, there are few straightforward ways to access and analyze that data with the reporting and business analytics tools many information workers use today. And that's proving to be a stumbling block for many big data projects.
AtScale, aims to bridge that disconnect. The San Mateo, Calif.-based company exited stealth mode in April and debuted its AtScale Intelligence Platform software that allows commonly used business intelligence tools to access data stored in Hadoop clusters. The technology creates a semantic layer between Hadoop and business analytics tools, turning Hadoop into an OLAP server.

H2O.ai:

Founded: 2011. CEO: SriSatish Ambati. Capital raised: $33.6 million. The Company, rebranded as H2O.ai from 0xdata in Nov 2014, offers an open source machine learning platform that works with Hadoop and Spark. It can be used through a Web UI or programming environments such as R, Java, Scala, Python, and JSON. It supports common database and file types, including Microsoft Excel, R Studio, and Tableau. Customers include AT&T, Comcast, Kaiser Permanente, McKesson, Walgreens, Capital One, and Progressive.

Long Format: In the long format each row speaks about a particular subject at a particular time point. If a particular subject's responses are taken at different time points it will be appearing in multiple rows. Any of the variables that doesn't change with time remain the same in all the rows.

Student ID	Student Name	Semester #	GPA
800898006	Prithvi Macherla	1	4
800898006	Prithvi Macherla	2	3
800898006	Prithvi Macherla	3	3.5
880898006	A Srimukhi	1	4
880898006	A Srimukhi	2	3.75
880898006	A Srimukhi	3	3.75

Table 3. Long Format Example

In the above example, the subject 'Prithvi Macherla' is measured at different time points that is i different semesters and therefore appears in multiple rows. The attributes namely 'StudentID' and 'StudentName' doesn't change with time. Therefore they remain the same in all of the subject's rows.

Wide Format: On the other hand in wide format, the subject's repeated responses will be in a single row and each response is given a separate column.

Student ID	Student Name	Sem 1 GPA	Sem 2 GPA	Sem 3 GPA
800898006	Prithvi Macherla	4	3	3.5
880898006	A Srimukhi	4	3.75	3.75

Table 4. Wide Format Example

In the above example, all of the subject's responses (GPA values in different semesters) are shown in a single row giving each semester (time point) a separate column.

These formats dominate one another in different scenarios. To see this let us consider different scenarios.

1. Missing data: Deletion is the default approach used to handle the missing data. That is when one of the attributes go missing that particular tuple is dropped to handle the missing values. In the long format, if a child is missing at one point, only that time point is dropped from the analysis and the subject's measures in other time points exist. But in the case of wide format, the subject itself will be dropped from the entire analysis.

Student ID	Student Name	Sem #	GPA
80089006	Prithvi Macherla	1	4
800898006	Prithvi Macherla	3	3.5
880898006	A Srimukhi	2	3.75
880898006	A Srimukhi	3	3.75

Table 5. Missing data (Long Format)

In the above example, though the subjects at certain time points are dropped but they are still part of the entire analysis.

Student ID	Student Name	Sem 1 GPA	Sem 2 GPA	Sem 3 GPA
880898006	A Srimukhi	4	3.75	3.75

Table 6. Missing data (Wide Format)

In table 4, we notice that when a row related to the subject is dropped, that subject will not be present in the entire analysis.

2. Flexibility: In the real world, the time is likely to be treated as continuous rather than considering it as 2 or more categories. This can be implemented with ease in the long format and is not possible in the case of wide format. In table 5 given below, time at which the fee payment is done is noted and in this case time has to dealt as continuous.

Student ID	Student Name	Amount Paid	Payment Date
800898886	Ross Geller	100$	3/19/2016
800898886	Ross Geller	1200$	4/22/2016

Table 7. Depicting time in long format

3. Building larger models: Suppose in the above example, the list of courses taken in each of the semesters should also be shown. This is can be incorporated in a simple way in long format. On the other hand, in the case of wide format building larger models from smaller ones is not simple. Long format is flexible when compared to the wide.

Student ID	Student Name	Sem #	GPA`	Course List
800898006	Prithvi Macherla	1	4	SPL,SSDI, DB
800898006	Prithvi Macherla	2	3	MAD, CN, IS, AC
800898006	Prithvi Macherla	3	3.5	KDD, DW

Table 8. Larger model

In other cases like Table 5, wide format is better than long format for analysis purposes.

```
subject.id tx measure.1 measure.2 measure.3
         1  A  27.61686  19.57415  24.10536
         2  A  41.46307  25.26103  32.82880
         3  A  33.97125  26.57679  42.66070
         4  A  31.03210  13.75620  23.50204
         5  A  28.27894  19.57316  27.29456
         6  A  37.97978  24.92336  30.52496
```

Table 9. Wide format for analysis

Edward Tufte's concept of "chart junk"

According to Edward Tufte, chart junk is unnecessary information that doesn't add any real value or that distracts the user from the actual information. Markings and visual elements can be called as chart junk if they don't use minimal set of visuals that are necessary for communication of the information. The chart junk is easy to produce when compared to fetching proper results. The background color or figure camouflaging the chart itself can be an example of chart junk as it distracts the user and doesn't add any real value to the chart.

According to Edward Tufte, irrespective of the reason of decoration, the extra ink used doesn't add any real value and is considered to be chart junk. His work also states there are better ways to portray the essence than to tangle it with the statistics. There are various types of chart junk among which unintentional optical art is widespread.

- **Unintentional optical art:** These arts produce a vibration or a movement effect, which distracts the human eye. Using such optical arts is considered to be chart junk as it is harmful and deviates the reader.
    - Such optical arts appeared in different publications ranging from 2% to more than 50% of the sample used for experimentation.
    - Examples of unintentional optical art can be an unnecessary usage of graphics in simple graph where some of the particles of the graph itself camouflage other particles.
- **The Grid problem:** This is another area where chart junk comes into the picture. Dark grid lines can deviate from the actual plotting and therefore falls under the concept of chart junk.

Method to determine whether the statistics published in an article (e.g. newspaper) are either wrong or presented to support the author's point of view, rather than correct, comprehensive factual information on a specific subject

Statistics in an article or newspaper can be evaluated based on the sampling set used. If there is no information given about the sampling set that was taken into consideration then there are brighter chances of the sample set is biased. For example, the overall doctors that subscribe a particular drug might be 70% but around 90% can fill the sample being tested. So, if there is no proper whereabouts about the sample then there are high chances that the survey conducted is a false one.
On the other hand, if proper details are given hypothesis testing can be carried out to prove the correctness of the survey. Hypothesis testing is the testing often used by the statisticians on the sample data to see whether a hypothesis holds or not. Here, one of the hypotheses can be the showcased result.

Screening outliers

Outliers are observations that vary largely from rest of the major observations. In other words, outliers are the objects that are many standard deviations away from the mean.
Screening outliers can be very useful in a variety of ways. Screening outliers and removing them can better the results of various important data mining algorithms like K-means etc. On the other hand screening outliers and studying them can be a whole other scenario.
The reasons for the existence of outliers can be incorrect observations or the object being different from others, which is a topic of scientific importance.

The outliers have to be detected in a manner and this should be done carefully. There are separate methods to detect single outliers and multiple outliers. A method for detecting single outliers shouldn't be applied sequentially on a data set to detect multiple outliers. One of the most commonly used techniques to screen outliers is known as box plots. The data set is plotted on a graph and all the points that fall below the 10th percentile and above the 90th percentile (based on the value on which the plotting is done)

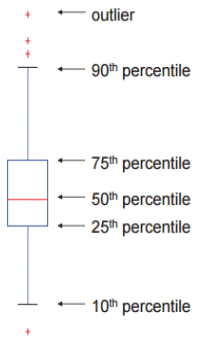

Fig 8. Box plots

PR and ROC curves and relationship between them
Basically PR (Precision-Recall) curves and ROC (Receiver Operating Characteristic) curves are used to understand the performance of an algorithm or system and its results.
Precision is the fraction of retrieved results that are relevant. It is also called as **Positive Predictive Value (PPV)** and measures how useful the retrieved results are.
Precision= no of relevant results retrieved/ total no of retrieved results
Recall is the fraction of relevant results that are retrieved. It is also called as **Sensitivity** and measures how complete the results are.
Recall= no of relevant results retrieved/ total no of relevant results
Let us consider an example to better understand about precision and recall, a search engine returns 30 pages of which only 20 are relevant and 30 are irrelevant and fails to return 40 additional relevant pages, then its precision= 20/30= 2/3 and recall= 20/60= 1/3. Precision measures exactness of quality while recall measures completeness or quantity.
Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of

reducing the other. Brain surgery example helps us to better understand this. Consider a brain surgeon tasked with removing a cancerous tumor from a patient's brain. The surgeon needs to remove all of the tumor cells since any remaining cancer cells will regenerate the tumor. Conversely, the surgeon must not remove healthy brain cells since that would leave the patient with impaired brain function. The surgeon may be more liberal in the area of the brain she removes to ensure she has extracted all the cancer cells. This decision increases recall but reduces precision.
On the other hand, the surgeon may be more conservative in the brain she removes to ensure she extracts only cancer cells. This decision increases precision but reduces recall. That is to say, greater recall increases the chances of removing healthy cells (negative outcome) and increases the chances of removing all cancer cells (positive outcome). Greater precision decreases the chances of removing healthy cells (positive outcome) but also decreases the chances of removing all cancer cells (negative outcome).
Usually, precision and recall scores are not discussed in isolation. Instead, either values for one measure are compared for a fixed level at the other measure (e.g. precision at a recall level of 0.75) or both are combined into a single measure. Examples for measures that are a combination of precision and recall are the F-measure (the weighted harmonic mean of precision and recall), or the Matthews correlation coefficient, which is a geometric mean of the chance-corrected variants: the regression coefficients.

		Predicted condition	
	Total population	Predicted Condition positive	Predicted Condition negative
True condition	condition positive	True positive	False Negative (Type II error)
	condition negative	False Positive (Type I error)	True negative

Figure 9: Confusion matrix

true positive (TP)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error

sensitivity or **true positive rate (TPR)**
eqv. with hit rate, recall

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

specificity (SPC) or **true negative rate (TNR)**

$$SPC = \frac{TN}{N} = \frac{TN}{FP + TN}$$

precision or **positive predictive value (PPV)**

$$PPV = \frac{TP}{TP + FP}$$

negative predictive value (NPV)

$$NPV = \frac{TN}{TN + FN}$$

fall-out or **false positive rate (FPR)**

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - SPC$$

false discovery rate (FDR)

$$FDR = \frac{FP}{FP + TP} = 1 - PPV$$

Probabilistic Interpretations:
It is possible to interpret precision and recall not as ratios but as probabilities:

- Precision is the probability that a (randomly selected) retrieved document is relevant.
- Recall is the probability that a (randomly selected) relevant document is retrieved in a search.

The random selection should be such that all documents in the set are equally likely to be selected.
Another interpretation for precision and recall is as follows. Precision is the average probability of relevant retrieval. Recall is the average probability of complete retrieval. Here we average over multiple retrieval queries.

Receiver Operating Characteristic (ROC):
This is basically a technique for visualizing, organizing and selecting classifiers based on their performance. ROC curve is created by plotting **True Positive Rate (TPR) vs False Positive Rate (FPR)**. TPR is also called sensitivity or recall while FPR is also called fall-out. To understand about fall-out, it's first important to explain about specificity. Specificity is the fraction of negatives that are correctly identified as such. Specificity is also called as Total Negative Rate (TNR).
TNR/ Specificity= true negative/ total negative
Fall-out= 1- specificity

The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.
A ROC space is defined by FPR and TPR as *x* and *y* axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent to sensitivity and FPR is equal to 1 − specificity, the ROC graph is sometimes called the sensitivity vs (1 − specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space.
Relation between PR and ROC curves:
Both curves are used to analyze the performance of an algorithm or system but ROC are best used to represent results for binary decision problems while PR curves give more informative picture of an algorithm's performance when dealing with highly skewed data sets.
There is a very important difference between what a ROC curve represents vs that of a PR curve. Remember, a ROC curve represents a relation between sensitivity (recall) and specificity (not precision). Sensitivity is the other name for recall but specificity is not precision. Recall/Sensitivity is the measure of the probability that your estimate is 1 given all the samples whose true class label is 1. It is a measure of how many of the positive samples have been identified as being positive. Specificity is the measure of the probability that your estimate is 0 given all the samples whose true class label is 0. It is a measure of how many of the negative samples have been identified as being negative. Precision on the other hand is different. It is a measure of the probability that a sample is a true positive class given that your classifier said it is positive. It is a measure of how many of the samples predicted by the

classifier as positive is indeed positive. Note here that this changes when the base probability or prior probability of the positive class changes. Which means precision depends on how rare is the positive class. In other words, it is used when positive class is more interesting than the negative class.

So, if your problem involves kind of searching a needle in the haystack when for ex: the positive class samples are very rare compared to the negative classes, use a precision recall curve. Otherwise use a ROC curve because a ROC curve remains the same regardless of the baseline prior probability of your positive class (the important rare class).

In other words, in principle, ROC and PR are equally suited to compare results. But for the example case of a result of 20 hits and 1980 misses they show that the differences can be rather drastic, as shown in below figure.

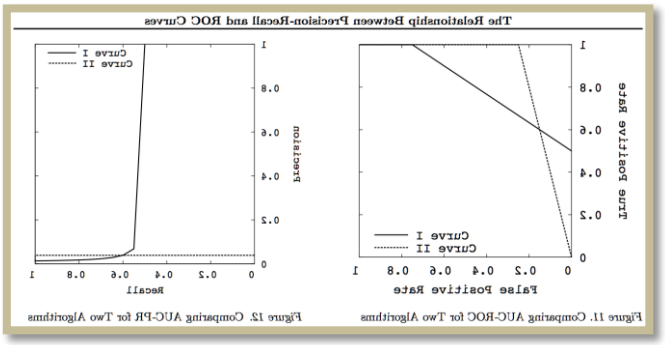

Fig 10. Terminologies and derivations from matrix

Result/curve (I) describes a result where 10 of the 20 hits are in the top ten ranks and the remaining 10 hits are then evenly spread out over the first 1500 ranks. Result (II) describes a result where the 20 hits are evenly spread over the first 500 (out of 2000) ranks. So in cases where a result "shape" like (I) is preferable, this preference is clearly distinguishable in PR-space, while the AUC ROC of the two results are nearly equal.

Validation of a model created to generate a predictive model of a quantitative outcome variable using multiple regression

Multiple regression:
Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the **dependent variable**. The variables we are using to predict the value of the dependent variable are called the **independent variables**.
For example, you could use multiple regression to understand whether exam performance can be predicted based on revision time, test anxiety, lecture attendance and gender. Alternately, you could use multiple regression to understand whether daily cigarette consumption can be predicted based on smoking duration, age when smoking, smoker type, income and gender started.
Multiple regression also allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained. For example, you might want to know how much of the variation in exam performance can be explained by revision time, test anxiety, lecture attendance and gender "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

Regression validation:
Now in order to validate the model created using multiple regression we use various validation techniques as validation using R^2, analysis of residuals and cross-validation. The validation process can involve analyzing the goodness of fit of the regression, analyzing whether the regression residuals are random, and checking whether the model's predictive performance deteriorates substantially when applied to data that were not used in model estimation.

1) Validation using R^2: coefficient of determination
The coefficient of determination is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.
R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data. Values of R^2 outside the range 0 to 1 can occur where it is used to measure the agreement between observed and modeled values and where the "modeled" values are not obtained by linear regression and depending on which formulation of R^2 is used. One problem with the R^2 as a measure of model validity is that it can always be increased by adding more variables into the model, except in the unlikely event that the additional variables are exactly uncorrelated with the dependent variable in the data sample being used.

2) Analysis of residuals
The residuals from a fitted model are the differences between the responses observed at each combination values of the explanatory variables and the corresponding prediction of the response computed using the regression function. Mathematically, the definition of the residual for the i^{th} observation in the data set is written

$$e_i = y_i - f(x_i; \hat{\beta}),$$

with y_i denoting the i^{th} response in the data set and x_i the vector of explanatory variables, each set at the corresponding values found in the i^{th} observation in the data set.
If the model fit to the data were correct, the residuals would approximate the random errors that make the relationship between the explanatory variables and the response variable a statistical relationship. Therefore, if the residuals appear to behave randomly, it suggests that the model fits the data well. On the other hand, if non-random structure is evident in the residuals, it is a clear sign that the model fits the data poorly.

3) Cross- validation

Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting, give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem), etc. One round involves partitioning a sample of data into complementary sub sets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

Regularization

Regularization is a technique used in an attempt to solve the overfitting problem in statistical models. Let's first understand about overfitting in order to gain good understanding about regularization. Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been over fit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

The possibility of overfitting exists because the criterion used for training the model is not the same as the criterion used to judge the efficacy of a model. In particular, a model is typically trained by maximizing its performance on some set of training data. However, its efficacy is determined not by its performance on the training data but by its ability to perform well on unseen data. Overfitting occurs when a model begins to "memorize" training data rather than "learning" to generalize from trend. As an extreme example, if the number of parameters is the same as or greater than the number of observations, a simple model or learning process can perfectly predict the training data simply by memorizing the training data in its entirety, but such a model will typically fail drastically when making predictions about new or unseen data, since the simple model has not learned to generalize at all.

Let's take a look at the simple curve fitting problem to understand regularization and over-fitting. Given a set of points $(x_1,y_1),(x_2,y_2)...(x_n,y_n)$. Our goal is to find a model, which is a function y=f(x) that fits the given data. To do this, we can use the method least-square error. For simplicity, suppose f(x) is just a first order linear function, f(x)=Wx+bf. Our job is to figure out what W and b are. We set up an error function that looks like:

$$\sum_{i=1}^{N} (y_i - (Wx_i + b))^2$$

To figure out what W and b are, we need to minimize the error function above. However, in minimizing the error function, we get into a problem called over-fitting, when the model we found fits very well with the training data but fails miserably if we apply new data (that is, get another set of data points).

To do this, we introduce a new terms into the error function, which implies that the coefficient W are also derived from a random process. The error function now looks like:

$$\sum_{i=1}^{N} (y_i - (Wx_i + b))^2 + \lambda(W^2)$$

The added lambda parameter is called the regularized term.

To illustrate, suppose f(x) can be any order. To generate testing data, we first have a function y=g(x). Next, from points belonging to y=g(x), we added some noise and make those points our training data. Our goal is to derive a function y=f(x) from those noisy points, that is as close to the original function y=g(x) as possible. The plot below shows over-fitting, where the derived function (the blue line) fits well with the training data but does not resemble the original function.

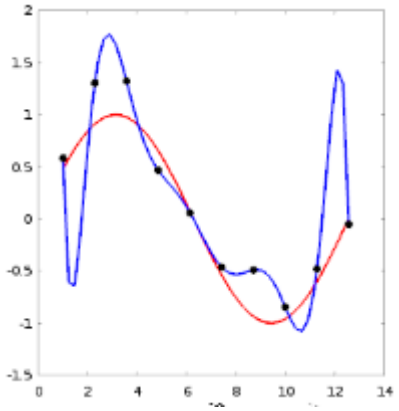

Fig 11. Before using regularization

After using regularization, the derived function looks much closer to the original function, as shown below:

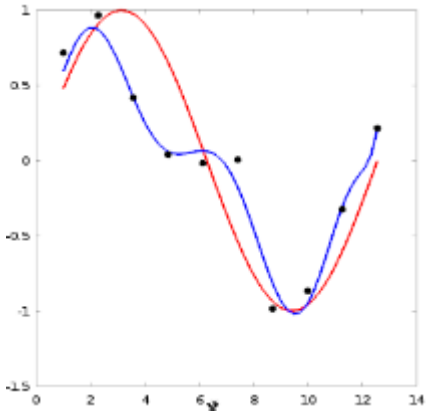

Fig 12. After using regularization

Representation of five dimensions in a chart

Dimensions in visualizations are not necessary orthogonal spatial dimensions. With visual attributes such as color, size and shape one can easily add some more dimensions to a visualization. There can be several ways to represent five dimensions in a chart of which some would be represented here.

For example let us consider the well-known Iris data set with four numerical dimensions (sepal length, sepal width, petal length, petal width) and one nominal attribute – the specific class of the referring iris plant. A small sample of the data set looks as below:

ID	sepal length	sepal width	petal length	petal width	class
Row9	4.4	2.9	1.4	0.2	Iris-setosa
Row29	5.2	3.4	1.4	0.2	Iris-setosa
Row48	4.6	3.2	1.4	0.2	Iris-setosa
Row52	6.4	3.2	4.5	1.5	Iris-versicolor
Row60	5.2	2.7	3.9	1.4	Iris-versicolor
Row81	5.5	2.4	3.8	1.1	Iris-versicolor
Row126	7.2	3.2	6.0	1.8	Iris-virginica
Row132	7.9	3.8	6.4	2.0	Iris-virginica
Row140	6.9	3.1	5.4	2.1	Iris-virginica

Fig 13. Table with five dimensions

The visual variables color, shape and size are assigned in the following way:

- Size: sepal length
- Color: sepal width
- Shape: class
- X-column: petal length
- Y-column: petal width

R tool can be used to plot the graph. All the visual variables described above can be assigned to a scatter plot using a script. The only way to get different shapes will be to add the points of the different classes successively. A resulting graph would be something like below:

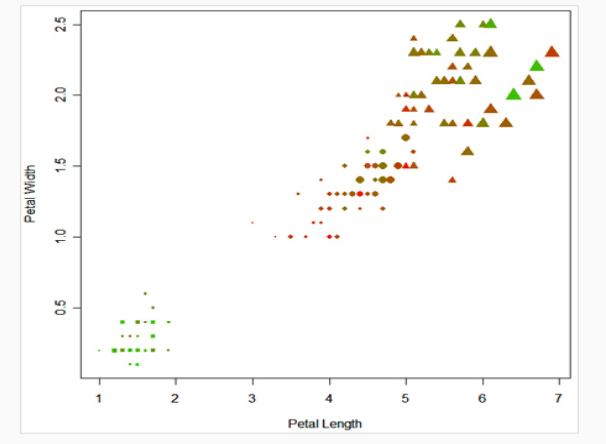

Fig 14. Five dimension representation in R

KNIME is one more tool that can be used. KNIME is an open source data analytics, reporting and integration platform. A graphical interface allows assembly of nodes for data preprocessing (extraction, transform and load) for modelling and data analysis and visualization. Simple the visual variables are attached to the data table pushed through the pipeline and the graph would be created. The screenshot below shows the KNIME flow and one example dialog to assign the color to the attribute "sepal width".

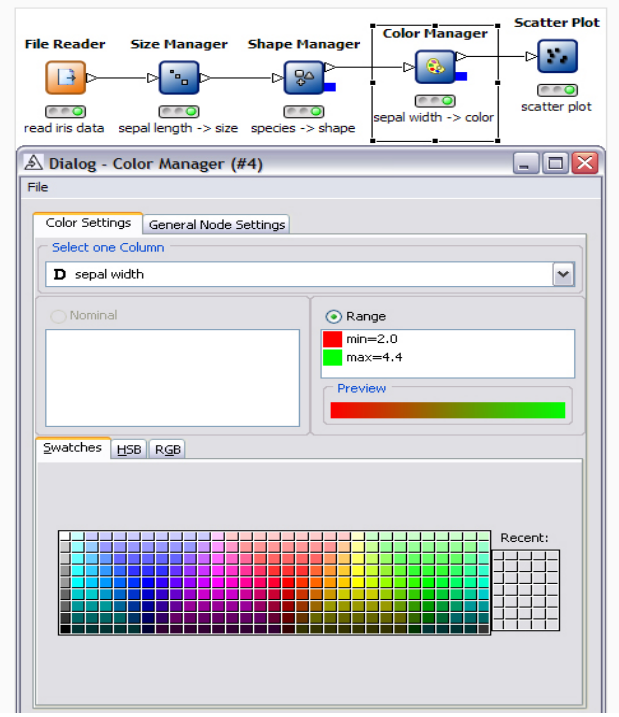

Fig 15. Tools for 5 Dimension representation

This results in a plot like below and this is a 2D scatter plot displaying all 5 dimensions of the data set.

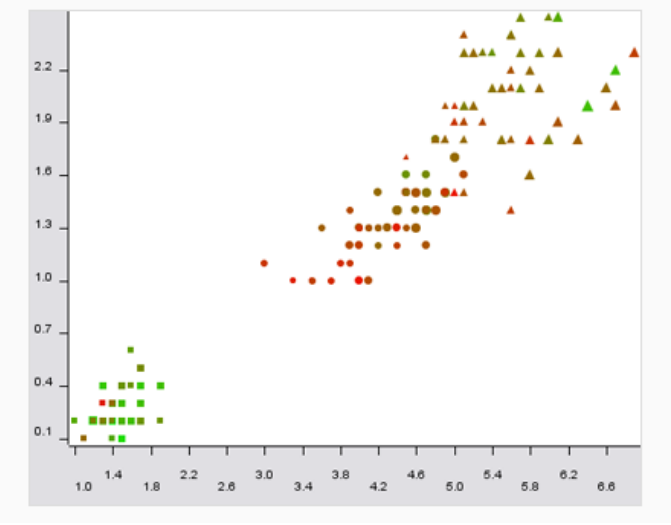

Fig 16. Five dimension representation using KNIME

Excel is also one of the tool that allows to create a variety of charts. Bubble chart is a type of chart that displays three dimensions of data. Each entity with its triplet of associated data is plotted as a disk that expresses two of the v_i values through the disk's xy location and the third through its size. Creating a bubble chart in excel along with color and motion

as two other attributes can be used to represent the five dimensions as shown below:

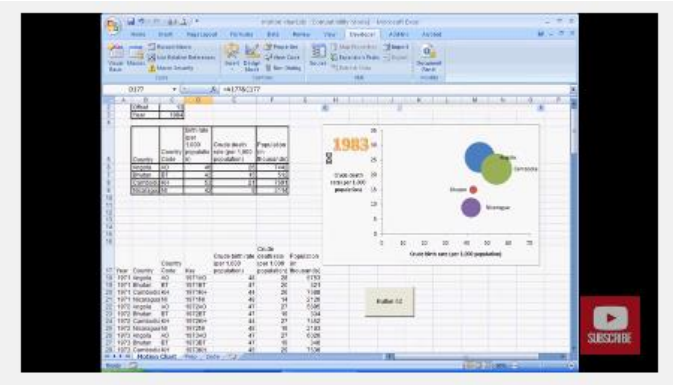

Fig 17. Five dimension representation in Excel

Also there are several web based charts and online flash-based API's available that can be used to create five dimensional charts. It's very difficult to decide which would be the best method out of all the options and completely depends on the individual and the data set used, to decide on the most efficient approach.

To prove that one improvement you've brought to an algorithm is really an improvement over not doing anything:

Often it is observed that in the pursuit of rapid innovation, the principles of scientific methodology are violated leading to misleading innovations, i.e. appealing insights that are confirmed without rigorous validation. One such scenario is the case that given the task of improving an algorithm to yield better results, you might come with several ideas with potential for improvement.

Frequently it is watched that in the quest for fast advancement, the standards of investigative technique are violated prompting deluding innovations, i.e. engaging experiences that are affirmed without thorough acceptance. One such situation is the situation when you fail to test that one improvement you've brought to an algorithm is really an improvement over not doing anything

The main reason behind this is, the urge to showcase these ideas really quick and get them implemented. This can be eliminated by following some simply checklist. To have supporting data, having results in such a way that are very likely to be not impacted by selection bias (known or unknown) or a misleading global minima (due to lack of appropriate variety in test data).

While the exact approach to prove that one improvement you've brought to an algorithm is really an improvement over not doing anything would depend on the actual case at hand. However there are a few common guidelines following which can make sure there is improvement in algorithm. They are:

- Test data used used for performance comparison should not be selection bias.
- Ensure that the test data has sufficient variety in order to be symbolic of real-life data
- Ensure that "controlled experiment" principles are followed i.e. while comparing performance, the test environment (hardware, etc.) must be exactly the same while running original algorithm and new algorithm
- Ensure that the results are repeatable with near similar results
- Examine whether the results reflect local maxima/minima or global maxima/minima

One common way to achieve the above guidelines is through A/B testing, where both the versions of algorithm are kept running on similar environment for a considerably long time and real-life input data is randomly split between the two. This approach is particularly common in Web Analytics.

While doing A/B testing, you can always check the model performance after adding or removing a features, if the performance of model is dropping or improving you can see if the inclusion of that variable makes sense or not. Apart from that, you tweak different inbuilt model parameters like you increase number of trees to grow or number of iterations to do in random forest, you add a regularization term in linear regression, you change threshold parameters in logistic regression, you assign weights to several algorithms, if you compare the accuracies and other statistics before and after making such change to model, you can understand if these result into any improvement or not.

A glance about a how typical A/B testing results look like

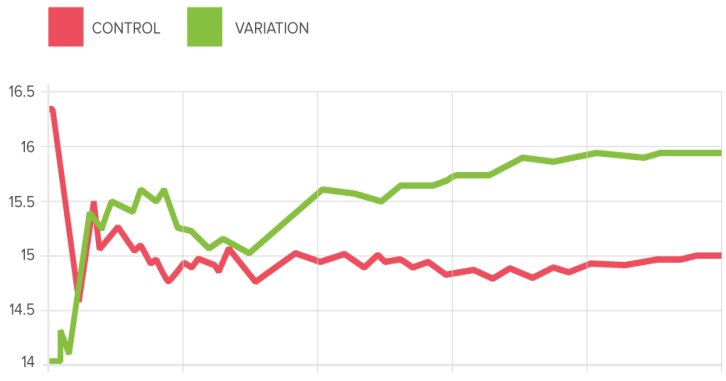

Fig 18. A/B Testing

Consider the following things while conducting testing your algorithm:

1. Focus: Get clarity on what your algorithm is targeting and focus on testing that part thoroughly.
2. Cohorts: Try to identify and segregate the user base into cohorts and perform testing based on the cohort scenarios. This helps you to cover multiple use cases.

3. Usage: Never ignore to check the improvements based on the usage of the algorithm.
4. New users: For better results its always advisable to bring in new users each time you test your algorithm. This helps to identify the unknown errors caused if any due to improvements you made.

Statistical Power

The **power of a statistical test** of a null hypothesis (H_0) is the probability that the H_0 will be rejected when it is false, that is, the probability of obtaining a statistically significant result. In plain English, statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected

Need: Any report formally testing a theory will incorporate a related p value and confidence interval, and another statistical idea that is in some ways more critical is the power of a study. Unlike p value and confidence interval, the issue of power ought to be considered even before embarking on a clinical study. Measurements are useful in analyzing most collections of information. This is similarly valid for hypothesis testing which can legitimize conclusions notwithstanding when no logical hypothesis exists.

Real world applications of hypothesis testing include:

- Establishing authorship of documents
- Evaluating the effect of the full moon on behavior
- Selecting the best means to stop smoking
- Checking whether bumper stickers reflect car owner behavior
- Testing the claims of handwriting analysts

Factors: Statistical power is affected chiefly by the size of the effect and the size of the sample used to detect it. Bigger effects are easier to detect than smaller effects, while large samples offer greater test sensitivity than small samples.

Type I error: Rejecting the null hypothesis when it is true is called a Type I error.
Type II error: that is failing to reject the null hypothesis when it is, in fact, not true.

Factors: Statistical power is affected chiefly by the size of the effect and the size of the sample used to detect it. Bigger effects are easier to detect than smaller effects, while large samples offer greater test sensitivity than small samples.

The statistical power of an experiment is determined by the following:

(a) The level of significance to be used
(b) The variability of the data (as measured, for example, by its standard deviation)
(c) The size of the difference in the population it is required to detect.
(d) The size of the samples

By setting the power to 80% and with any three of these four values, the remaining one can calculated. However, since we usually use a 5% level of significance, we need to set three out of (b), (c), (d) and the power to determine the other. The variability of the data (b) needs to approximately assessed, usually from previous studies or from the literature, then the sample sizes (d) can be determined for a given difference (c) or, alternatively, for a specific sample size the difference likely to be detected can be calculated.

Many online calculators are available to calculate the statistical power. Below is one such online calculator for statistical power: https://www.dssresearch.com/KnowledgeCenter/toolkitcalculators/statisticalpowercalculators.aspx

Root cause analysis

Root cause analysis is a problem solving method. It is used to identify the root causes or faults of the problem. A root cause is an initiating cause of a simple condition or a casual chain that leads to an outcome. A root cause analysis can be used either when something goes wrong or something goes well. A root cause is casually identified as one, main cause. Focusing on single cause limits the solution set. There is a fair chance of solutions being missed.
In some fields like medicine, it is easy to understand the difference between treating, symptoms and curing. For example, pain killers will just reduce the pain instantly but will not heal the problem. Similarly, when you have a problem at work, you don't jump to the treatments. You look deeper into the problems, figure out what's causing the problem and fix it.

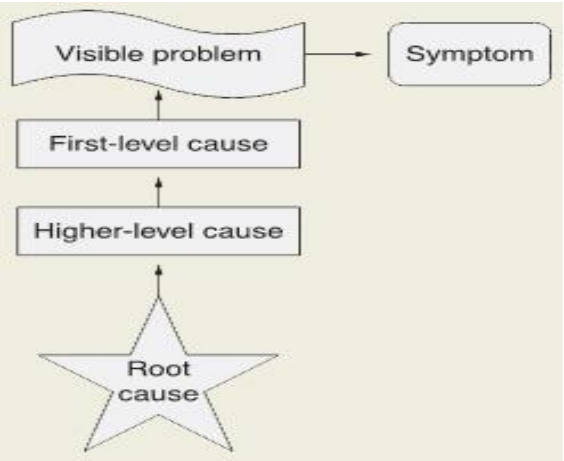

Fig 19. Root Cause Analysis

It is fundamentally connected by three basic questions:
What's the problem? Why did it happen? and What will be done to prevent it?

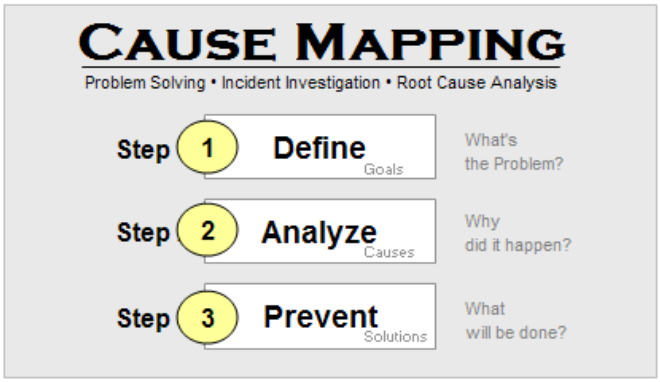

Fig 20. Cause Mapping

A Cause Map is a visual explanation of occurrence of the incident. It can be either very basic or can be extremely detailed depending on the issue.

There are three basic steps to the Cause Mapping method:

1. Define the issue by its impact to overall goals
2. Analyze the causes in a visual map
3. Prevent or mitigate any negative impact to the goals by selecting the most effective solutions.

Usually there are three types of causes:

Physical causes – Material items failed in some way, like a machine breakdown.
Human causes – People doing something wrong or not doing it right or doing something not needed. Human causes typically lead to physical causes. Like not plugging off the phone after full charging leads to battery failure.
Organizational causes – A system, process, or policy that people use to make decisions or do their work is faulty.
To make it clear. A car's breaks not working is a physical cause, which is because no one filled the brake fluid which is a physical cause. Everyone assuming someone had filled the brake fluid and they are not responsible for the damage is the organizational cause.
The root cause analysis have been defined in five steps:

1) Defining the problem.
2) Collecting the data.
3) Identifying the possible casual factors.
4) Identifying the root causes.
5) Recommending and Implementing the solutions

The 5 why's technique:

One of the famous technique for the root cause of a problem which shows the relationship of causes by repeatedly asking the question, "Why?", until you find the root of the problem. This technique is commonly called "5 Whys", although is can be involve more or less than 5 questions.

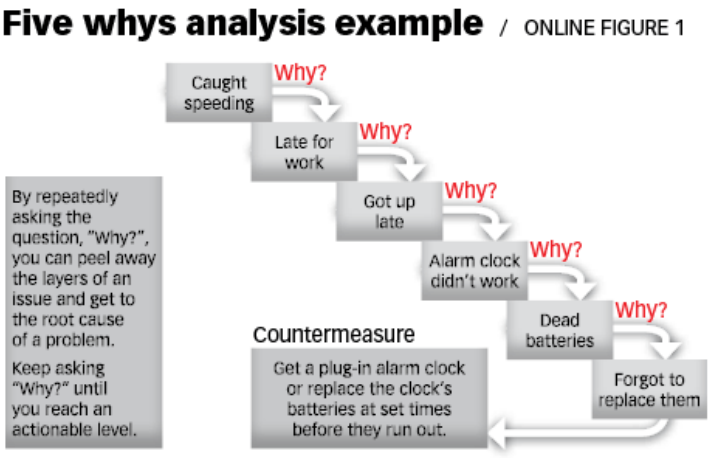

Fig 21. Five whys

Price elasticity demand:

Price elasticity demand can be defined as a measure of the responsiveness
of quantity demanded to a price change. It might cause a price change just to
have a small or large impact on quantity demanded. Goods like gasoline which
are inelastic or classified as inelastic probably are still purchased at the same quantity even when there is raise in price as they are mandatory goods that are required. While in other case elastic goods such as restaurant meals, movies usually follow the pattern or trend of the law of demand and will drop its trend based on price change or demand of that particular commodity on a particular season.

So as we see there is no change in demand in case of an inelastic commodity or a good here as mentioned above it's a mandatory or an essential good where the demand never experiences a drop or fall though there is change in price.

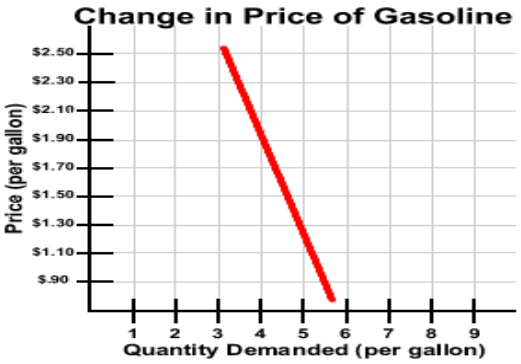

Fig 22. Change in Price of Gasoline

Now when it comes to the case of an non-essential or a luxury commodity the drop of the demand with respect to the change in price is quite high as when we take a look at the pizza graph we notice that people tend to request only one slice when a slice costs 3 $. So this graph depicts or serves as a good example to explain the very fact that on luxurious

commodities like hotel reservations, restaurant meals the demand drops by a considerable margin when compared to or unlike essential or elastic goods.

Fig 23. Change in Price of Pizza

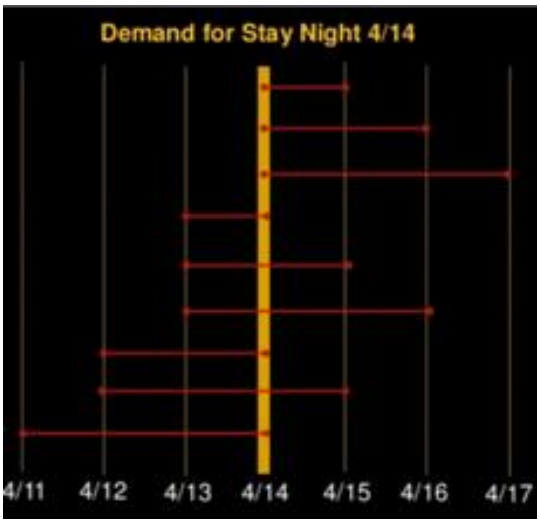

Fig 24. Demand for Hotel

Price Optimization:

Price optimization can be better explained as the amount of usage of mathematical analysis by a company to detect how customers respond to price changes for a particular good in the market. In other words, to be short it uses and is based on big data to test the potential buyers. They observe, figure out the pattern and as part of their mathematical analysis and this determines will best meet its objectives and main functions such as maximizing an operating profit. The data that is used in price optimization indulges operating costs, inventories and history of the prices and sales. Price optimization is a healthy process and is implemented in various fields such as banking, retail, airlines, hotels , casinos etc as this analysis is useful in and serves as template for their business and improves the standard and quality and sales for a particular business in respected fields.

Example:

Hotel revenue management and dynamic pricing:-

Optimizing on what prices to charge on a customer would help a Hotel business and would make them aware on the type of reservations to accept as hotels tend to accept and charge maximum on room reservation and generate maximum revenue from them. To give an insight a hotel management would not likely tend to accept a reservation on a Friday night at $300 as there would be high chances and possibilities if renting the same room by a different customer on Friday as well Saturday as being on a weekends people tend to stay more so accepting the reservation on both saturday and friday nights would fetch a hotel 250$*2=500$ which in that case any management would tend to earn and take $500 rather than 300$ single day reservation. So this is the way price optimization and its mathematical analysis would work in case of a hotel business. The picture would depict the tendency of weekend reservations that was mentioned earlier.

Inventory Management:

Inventory management can be bestowed as the overseeing and the controlling of the ordering storage or alignment and use of the components that a particular firm or company will face in the production or manufacturing of a particular god or item. It will sell as well as the overseeing and controlling of the fine and finished final products for sale. The major asset here is the business inventory represents an investment that is tied up till a particular item is released in the market and sold or used in the production of an item that is sold. There is a lot of expenditure that needs to be spent to store, track and insure an inventory. Inventories, which are not managed properly or in other terms mismanaged can create a significant financial problem for a business, whether a particular mismanagement results in an inventory glut or an inventory shortage.

Successful inventory management always involves creating a particular plan and a dedicated prompt that will ensure that their particular goods or commodities manufactured by that particular firm are always available when they are needed most in the market or in other words should always match up with the demand. The other way around to deal with this that companies tend to maintain sales forecasts of a significant period and then manufacture and release their products into the market accordingly with the time stamp.

Example:

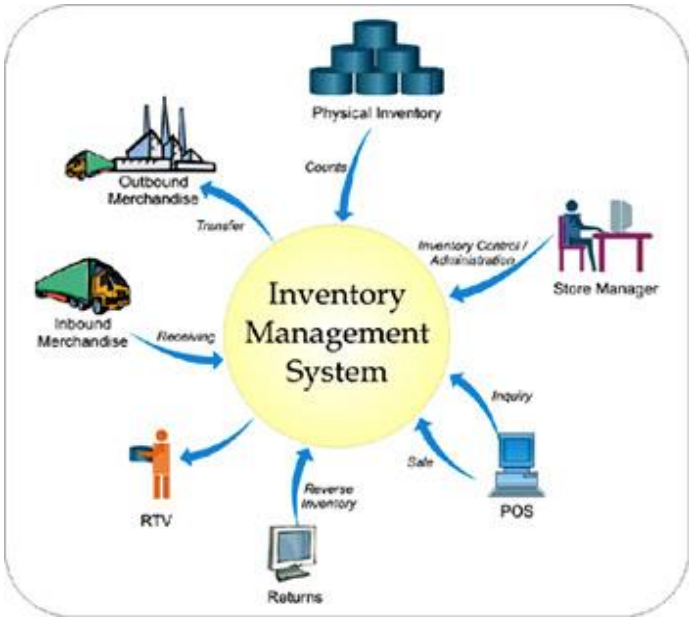

Fig 25. Inventory Management System

Competitive Intelligence:

Competitive intelligence is the action of defining, gathering, analyzing, and distributing intelligence about products, customers, competitors, and any aspect of the environment needed to support executives and managers making strategic decisions for an organization.

Examples:
stock traders who analyze the data on prices and price movements to determine the best investments, Japanese automobile industry's analysis of the U.S.-automobile market in the 1970s, AT&T's database of in-company experts, a final example is how Wal-Mart stores studied problems Sears had with distribution, and built a state-of-the art distribution system so that Wal-Mart customers were not frustrated by out-of-stock items, as were Sears's customers.

Resampling Methods:
Resampling methods refers to the use of already available data or data generating mechanisms to generate new hypothetical data samples that represent the underlying population, the results of which can be analyzed. But the definition of resampling in Wikipedia is defined as:
In statistics, resampling is any of a variety of methods for doing one of the following:

Estimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping)
Exchanging labels on data points when performing significance tests (permutation tests, also called exact tests, randomization tests, or re-randomization tests)
Validating models by using random subsets (bootstrapping, cross validation)

Common resampling techniques include bootstrapping, jackknifing and permutation tests.

Types of Resampling Methods:
There at least four major types of resampling methods.

A. **Boot Strapping:** The definition of bootstrapping according to Wikipedia is "Bootstrapping is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample, most often with the purpose of deriving robust estimates of standard errors and confidence intervals of a population parameter like a mean, median, proportion, odds ratio, correlation coefficient or regression coefficient."

   This technique was invented by Bradley Efron (1979, 1981, and 1982) and further developed by Efron and Tibshirani (1993). In bootstrapping, we try to derive the estimates of standard error and confidence intervals, by generating samples with replacement from the original set. This approach is used in hypothesis testing and is used to calculate the approximate sample size for experimental design.

B. **Jackknife:** It is also known as the Quenouille-Tukey Jackknife, this tool was invented by Maurice Quenouille in 1949 and later developed by John W. Tukey in 1958. As the father of EDA, John Tukey attempted to use Jackknife to explore how a model is influenced by subsets of observations when outliers are present.

   Jack knifing is used in statistical inference to estimate the bias and standard error (variance) of a statistic, when a random sample of observations is used to calculate it. This method provides a systematic method of resampling with a mild amount of calculations. It offers "improved" estimate of the sample parameter to create less sampling bias. The basic idea behind the jackknife estimator lies in systematically re-computing the statistic estimate leaving out one observation at a time from the sample set. From this new "improved" sample statistic can be used to estimate the bias can be variance of the statistic.

C. **Randomization Exact Test:** Randomization tests are also known as re-randomization tests, Exact Tests, Permutation Tests. They were developed by R.A. Fisher in 1939, but subsequently lost interest in it because of lack of proper technology to further develop it. As described earlier a permutation test (also called a randomization test, re-randomization test, or an exact test) is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic

under rearrangements of the labels on the observed data points.

So, it is a type of statistical significance test, in which a reference distribution is obtained by calculating all possible values of test statistic under rearrangement of the labels on the observed data points.

D. **Cross- Validation:** Cross-validation is a statistical method for validating a predictive model. Subsets of the data are held out for use as validating sets; a model is fit to the remaining data (a training set) and used to predict for the validation set. Averaging the quality of the predictions across the validation sets yields an overall measure of prediction accuracy. Cross-Validation is employed repeatedly in building decision trees.
Simple cross-validation was proposed by Kurtz (1948) as a remedy for the Rorschach test. Based on Kurtz's simple cross-validation, Mosier (1951) developed double cross-validation, which was later extended to multicross-validation by Krus and Fuller (1982).

Advantages: Permutation tests exist for any test statistic, regardless of whether or not its distribution is known. Thus one is always free to choose the statistic which best discriminates between hypothesis and alternative and which minimizes losses.

Limitations: An important assumption behind a permutation test is that the observations are exchangeable under the null hypothesis. An important consequence of this assumption is that tests of difference in location (like a permutation t-test) require equal variance.
Other limitations with resampling are:

1. **Assumption:** Stephen E. Fienberg mocked resampling by saying, "You're trying to get something for nothing. You use the same numbers over and over again until you get an answer that you can't get any other way. In order to do that, you have to assume something, and you may live to regret that hidden assumption later on" (cited in Peterson, 1991, p. 57).

   Every theory and procedure is built on certain assumptions and requires a leap of faith to some degree. Indeed, the classical statistics framework requires more assumptions than resampling does.

2. **Generalization:** Some critics argued that resampling is based on one sample and therefore the generalization cannot go beyond that particular sample. One critic even went further to say, "I do wonder, though, why one would call this (resampling) inference?" (cited in Ludbrook & Dudley, 1998) Nevertheless, Fan and Wang (1996) stated that assessing the stability of test results is descriptive, not inferential, in nature.

3. **Bias and Bad data:** Bosch (2002) asserted that confidence intervals obtained by simple bootstrapping are always biased though the bias decreases with sample size. If the sample comes from a normal population, the bias in the size of the confidence interval is at least n/(n-1), where n is the sample size. Nevertheless, one can reduce the bias by more complex bootstrap procedures. Some critics challenged that when the collected data are biased, resampling would just repeat and magnify the same mistake.

4. **Accuracy:** The other limitation of resampling is accuracy. The researcher has to perform enough experimental trials in order to produce accurate results. If the researcher doesn't conduct enough experimental trials, resampling may be less accurate than conventional parametric methods. However, this doesn't seem to be a convincing argument because today's high-power computers are capable of running billions of simulations.

Selection bias:

Selection bias occurs when proper randomization of sample is not achieved when any sample is selected. It means that when any sample is selected for any study or analysis, the sample selected or generated is not completely random or doesn't represent the complete set of population of data.

Researchers are always encountered with the problem of selection bias. What researchers basically mean when they selection bias is "Selection bias is a common type of error where the decision about who to include in a study can throw findings into doubt". Bias is a type of error that systematically skews the results or favors the results in one direction. So, selection bias usually occurs when the researcher decide who should be included in a particular study.

For example: we want to study the health performance of employees in a software company between people who work at 9-5 in the morning and people who work night shifts. We collect the data from all the employees working. We compare the rate at which health problems are reported between both the types of employees.
In this type of study there is a problem of occurrence of selection bias. The main trouble is the study group from the morning and night shifts might be completely different. Another trouble is that, the eating habits of morning and night shift employees might be completely different.
So when the results are derived from the available data, the results might favor the night shift employees, stating that there is a higher occurrence if health issues in night shift employees.

Selection bias also occurs when there is under coverage. That is the sample data might not be representative of the entire population when there is under coverage of some members from the population.

The other example for selection bias is voluntary response bias. When any study is conducted the members included in the sample might be self-selected volunteers. For example, we try to induce a program where we try to improve the health habits of the employees. Now the people will participate in the program are self-selected people who are more health conscious than others. Now the program cannot be deemed as effective because there might be other factors involved in the improved health of people in the study than the program itself.

Non response bias is also another example of selection bias where the people who are selected for the study are unwilling to participate in the survey. This problem usually occurs in email surveys where people are uninterested in the survey.

There are different types of bias:
Sampling bias
Time Interval
Data
Studies
Observer selection
These are few different types of selection bias that can occur.

Sampling bias occurs when there is ambiguity in the sample selected for any study. Sampling bias is systematic error due to a non-random sample of a population, causing some members of the population to be less likely to be included than others, resulting in a biased sample, defined as a statistical sample of a population (or non-human factors) in which all participants are not equally balanced or objectively represented. It is mostly classified as a subtype of selection bias.

Time Interval Bias occurs when the trial that is being conducted is terminated early when the results are about to support the desired conclusion.
A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.

Data bias is partitioning data based on the knowledge of the contents in the data and then applying tests and analyzing the data on blindly chosen partitions. Post hoc alteration of data inclusion based on arbitrary or subjective reasons, including:
Cherry picking, when specific subsets of data are chosen to support a conclusion (e.g. citing examples of injuries in matches as evidence of sports being unsafe, while ignoring the far more common example of matches where no injuries occurred)
Rejection of "bad" data on arbitrary grounds, instead of according to previously stated or generally agreed criteria or discarding "outliers" on statistical grounds that fail to take into account important information that could be derived from "wild" observations

Observer selection
Data is filtered not only by study design and measurement, but by the necessary precondition that there has to be someone doing a study. In situations where the existence of the observer or the study is correlated with the data observation selection effects occur, and anthropic reasoning is required.

An example is the past impact event record of Earth: if large impacts cause mass extinctions and ecological disruptions precluding the evolution of intelligent observers for long periods, no one will observe any evidence of large impacts in the recent past (since they would have prevented intelligent observers from evolving). Hence there is a potential bias in the impact record of Earth. Astronomical existential risks might similarly be underestimated due to selection bias, and an anthropic correction has to be introduced.

Prevention of Selection bias:
All efforts must be made to avoid selection bias when selecting any participants for any study. Selection bias can be avoided by

1) Study population that is being considered for the study has to be clearly identified.
2) The choice of right reference/comparison group for the study is crucial
    a) For example, in the example of in an occupational cohort study, rather than comparing workers with the general population (which includes people who are too ill to work),
        i) Select an external comparison group ie., a group outside of the work force and where the workers group selected has some degree of exposure to the job.
        ii) Select an internal comparison group ie., a group of people within the workforce of which few had exposure and others do not.
3) In the cohort study,
    a) The exposed and unexposed group selected for the study must be identical.
    b) Also, the selection of exposed and unexposed group for the study must be done without any prior knowledge of the outcome.
4) In case controlled study,
    a) The control group should represent the exposure of population which gave rise to the case.
    b) Precise case definition and exposure definition must be used by the investigators.
5) In the intervention study,
    a) Select the participants through randomization so that all of them have an equal chance of intervention.
    b) Whether the randomization has been successful or not can be checked by comparing baseline factors between the intervention and control group afterwards, and seeing if the groups are similar in all the other aspects apart from receiving the intervention.

Experimental design to answer a question about user behavior:
Experimental design is a planned interference in the natural order of events by the researcher. A researcher does something more than just recording the observations. The emphasis is mostly on the results that these experiments produce, thereby

reflecting the importance of these experiments. This is applicable to everyday life since most of the knowledge or data that is gained in all the fields is through experiments conducted. In experiment, there more than just observing the results of an event. A condition or change is introduced in the experiment being conducted and the resulting events are carefully examined to understand the effects of the changed induced in the experiment. These observations or measurements illuminate the effect of the changes in the conditions.

Experimental research tests a hypothesis and establishes causation by using independent and dependent variables in a controlled environment.

Dependent Variable is a subject of the experiment that is influenced by the manipulated aspect.

Independent Variable is the manipulated aspect of the experiment.

Experiments are generally the most precise studies. They derive the most approximate conclusion. These experiments are particularly effective in supporting various hypotheses about cause and effect relationships. However, since the conditions in an experiment are assumptions, they may not apply to everyday situations.

A well-designed experiment has control of random variables to make sure that the effect measured is caused by the independent variable being manipulated. The features included in these experiments have control over the random variables. These features include random assignment, use of a control group, and use of a single or double-blind design.

An experimenter decides how to manipulate the independent variable while measuring only the dependent variable. In a good experiment, only the independent variable will affect the dependent variable.

For example we are trying to determine the user behavior while trying to access a web page on the internet. In order to analyze the user behavior in this particular situation we try to implement the experimental design by following the below steps,

1: Formulate the Research Question: What are the effects of page load times on user satisfaction ratings?

2: Identify variables: We identify the cause & effect. Independent variable -page load time, Dependent variable- user satisfaction rating.

3: Generate Hypothesis: Lower page access time will have more effect on the user satisfaction rating for a web page. Here the factor we try to analyze is page load time.

4: Determine Experimental Design: We consider experimental complexity i.e. vary one factor at a time or multiple factors at one time in which case we use factorial design (2^k design). A design is also selected based on the type of objective (Comparative, Screening, Response surface) & number of factors. Here we also identify within-participants, between-participants, and mixed model. For e.g.: There are two versions of a page, one with Buy button (call to action) on left and the other version has this button on the right.

Within-participants design - both user groups see both versions. Between-participants design - one group of users see version A & the other user group version B.

5: Develop experimental task & procedure: Detailed description of steps involved in the experiment are defined. Tools used to measure user behavior, goals and success metrics are also explained. Collect qualitative data about user engagement to allow statistical analysis.

6: Determine Manipulation & Measurements:

Manipulation: One level of factor will be controlled and the other will be manipulated. We also identify the behavioral measures:

i. Latency- time between a prompt and occurrence of behavior (how long it takes for a user to click buy after being presented with products).
ii. Frequency- number of times a behavior occurs (number of times the user clicks on a given page within a time)
iii. Duration-length of time a specific behavior lasts(time taken to add all products)
iv. Intensity-force with which a behavior occurs (how quickly the user purchased a product)

7: Analyze results: Identify user behavior data and support the hypothesis or contradict according to the observations made for e.g. how majority of users satisfaction ratings compared with page load times.

Experimental design is done to test answers to certain questions during surveys, product preference in marketing etc. Let's take a example. Certain company wants to launch a new Shampoo and wants to understand their customer's preferences and reviews before launch so that it can advertise the product accordingly. There are many ways in which we can derive customer feedback through this experiment. One method could be to select 100 candidates from similar profile for whom the shampoo is intended. These candidates are advised to use the product for few days and then their review is recorded. Another method could be where only half of these 100 candidates are asked to apply the product and then the comparison is done between these two groups of candidates to understand the impact of shampoo. (Example given here is for analyzing product behavior, but we can apply the same to understand user behavior.)

Conclusion

Data Scientists will be capable of analyzing any data, in-turn experts in computer science, math and statistics, and subject matter. People experts in one particular discipline and insist that their discipline is the one and only true data science are called fake data scientists. Since the topics covered in this paper varies from general to expertise levels covering machine learning, big data, statistics and other mathematics models, it is very hard for anyone to fake it. However, if the answers presented here are accepted by large population, they offer lessons for all fake data scientists in order to evolve towards the true data scientists.

REFERENCES

[1] http://www.datacommunitydc.org/blog/2013/05/recommendation-engines-why-you-shouldnt-build-one
[2] http://www.forbes.com/sites/lutzfinger/2014/09/02/recommendation-engines-the-reason-why-we-love-big-data/#4fadeed8218e
[3] http://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html?_r=1&partner=permalink&exprod=permalink&_r=0
[4] https://www.youtube.com/watch?v=TSv6eLAOt78
[5] https://www.youtube.com/watch?v=NSscbT7JwxY
[6] http://www.aaai.org/ojs/index.php/aimagazine/article/view/2361/2231
[7] http://ids.csom.umn.edu/faculty/gedas/papers/recommender-systems-survey-2005.pdf
[8] https://en.wikipedia.org/wiki/Tableau_Software
[9] https://www.linkedin.com/company/tableau-software
[10] https://en.wikipedia.org/wiki/SAS_(software)
[11] https://en.wikipedia.org/wiki/R_(programming_language)
[12] http://support.sas.com/resources/papers/proceedings13/348-2013.pdf
[13] http://dataconomy.com/six-data-scientists-you-should-know/
[14] http://www.forbes.com/pictures/lmm45emkh/
[15] http://www.informationweek.com/big-data/9-hot-big-data-and-analytics-startups-to-watch/d/d-id/1324664
[16] http://www.crn.com/slide-shows/data-center/300077457/the-10-coolest-big-data-startups-of-2015-so-far.htm
[17] http://exceluser.com/blog/1133/good-examples-of-bad-charts-chart-junk-from-a-surprising-source.html
[18] http://www.theanalysisfactor.com/wide-and-long-data/
[19] wikipedia.org, 2005
[20] http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00040Z
[21] http://www.referenceforbusiness.com/management/Bun-Comp/Competitive-Intelligence.html
[22] http://www.econedlink.org/teacher-lesson/551/Price-Elasticity-Tires-Toothpicks
[23] https://en.wikipedia.org/wiki/Price_optimization
[24] https://www.economicshelp.org/blog/7019/economics/examples-of-elasticity/
[25] http://perfectprice.io/price-optimization/
[26] http://www.investopedia.com/terms/i/inventory-management.asp
[27] http://www.investinganswers.com/financial-dictionary/financial-statement-analysis/inventory-management-5999
[28] http://searchmanufacturingerp.techtarget.com/definition/Inventory-management
[29] http://asq.org/learn-about-quality/root-cause-analysis/overview/overview.html
[30] http://asq.org/quality-progress/2015/02/back-to-basics/the-art-of-root-cause-analysis.html
[31] https://en.wikipedia.org/wiki/Root_cause_analysis
[32] http://www.thinkreliability.com/Root-Cause-Analysis-CM-Basics.aspx
[33] https://www.mindtools.com/pages/article/newTMC_80.html
[34] http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1428
[35] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018227/
[36] https://www.youtube.com/watch?v=VFMcGdWp0MQ
[37] https://www.youtube.com/watch?v=nkmzsyLg0tY
[38] https://en.wikipedia.org/wiki/Statistical_power
[39] https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Use_and_importance
[40] Cohen, Jacob. "Statistical Power Analysis". *Current Directions in Psychological Science* 1.3 (1992): 98–101, Published by: Sage Publications, Inc. on behalf of Association for Psychological Science
[41] www.kdnuggets.com
[42] https://declara.com/content/61XmXxMg
[43] https://en.wikipedia.org/wiki/A/B_testing
[44] http://www.slideshare.net/vonreventlow/developing-and-testing-search-engine-algorithms
[45] https://www.linkedin.com/pulse/answer-20-questions-detect-fake-data-scientists-anuj-kumar
[46] https://www.boundless.com/psychology/textbooks/boundless-psychology-textbook/researching-psychology-2/types-of-research-studies-27/experimental-research-126-12661/
[47] http://www.kdnuggets.com/2016/02/21-data-science-interview-questions-answers-part2.html
[48] http://liutaiomott2ola.com/myth/expdesig.html
[49] https://en.wikipedia.org/wiki/Selection_bias/
[50] http://stattrek.com/statistics/dictionary.aspx?definition=selection%20bias
[51] http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_Bias/EP713_Bias_print.html
[52] http://www.businessdictionary.com/definition/selection-bias.html
[53] http://www.investopedia.com/terms/s/sample_selection_basis.asp
[54] http://www.iwh.on.ca/wrmb/selection-bias
[55] https://wiki.ecdc.europa.eu/fem/w/wiki/preventing-bias
[56] http://pareonline.net/getvn.asp?v=8&n=19
[57] http://userwww.sfsu.edu/efc/classes/biol710/boots/rs-boots.htm
[58] https://en.wikipedia.org/wiki/Resampling_(statistics)
[59] http://onlinelibrary.wiley.com/doi/10.1002/1097-0193(200102)12:2%3C61::AID-HBM1004%3E3.0.CO;2-W/full