1. Crisp DM and its phases.
   Crisp DM namely stands for Cross-Industry Standard Process for Data Mining. To define the above in a single sentence "**This is an written index of a book even before a single chapter is written**". This is the rule of hand that is followed in most data science projects and governs the direction of any project which has **Data at its epi-center.**

   For people to understand and work simultaneously on a project it is divided into six phases. (Just like forking and pipelining in a regular coding project). However it has to be pointed out that this model works best with the Agile projects as there are many correlated phases as we will see later.

   For the reasons discussed above the Model is defined in 6 phases.

   a. **Business Understanding**
      This is the ground level of the project where the right people whose business acumen allies with the project are assembled. They discuss the entire project time, resources, tools and data fields required for the project success.

   b. **Data Understanding**
      Once the project flow and people are decided the next step is to explore the data. Here all the relevant data sources are discussed which could have a causality to the project. Post this the quality of the data is determined and accessed if it can be used for the project.

   c. **Data Preparation**
      As mentioned in the slides as well as chapter 2, this part of the project should take up to 80% of the project time for the entire project if done correctly. This is the most critical part of the project where data is selected, cleaned and made more efficient for the project use. All the data related activities such as mining, construction, cleaning and formatting takes place here.

   d. **Modeling**
      This is another critical step of DM, where we actually model the data. Just like writing the code the algorithm is the unknown Hero the same goes for Modeling. We need to pick and select the best model for the project and the data on hand for the same. **More than one models are used in this step so that we can cherry pick the one's that fits the business best.**

   e. **Evaluation**
      This is the next step where we have seen the impact of various models and data sets and we make our conclusions. We generally select more than one models but pursue the one that fits the business best. Evaluate all the results and move to deployment.

f. **Deployment**

This is the final step where we deliver to the customer and show our results of mining as well as if they fall with in the domain predictions. This part generally needs constant monitoring as we sometimes need to tweak our model to counter business changes.

2. **Making Honey Sweeter**

Our project is based on the application Honey where we give it our spin and predict the best to buy a certain product.

Overview – As we know that prices of various products keeps on changing depending upon many factors such as inventory on hand, holidays, customer moods, targeted sales etc.

- **Business Understanding** – We have assigned work to different individuals depending upon skills to crawl through websites like target, Walmart and Amazon and prepare the data. Secondly we are trying to find resources that could help us with the on hand inventory of the products which manipulates the cost of these products.

- **Data Understanding –** Once we have all the data we will try to categories the data with costs, inventory, holidays and public moods for new launches. Also we will see the trends that impact the old prices when new products are launched.

- **Data Preparation –** We will be merging close to 5 data sources in one so once that is completed there will be a lot of scope for data cleaning and duplication keys.

- **Modeling –** We will be trying 3 models with 2 similar keys of prices and dates and we will see the impact of those on the entire data sets.

- **Evaluation –** Once the above is completed we will get the final model ready for our final presentation and try to run infinite time simulations and see whether the prediction lands in the domain.

- **Deployment –** We are planning to merge an UI so that we can manage the entire project through it as well as we will show in our final presentation if the predictions match our original domain.

3. The key differences are mentioned below.
   a. Supervised Learning
   - This method predicts the output.
   - Needs supervision to train the data.
   - Provides more accurate results.

   b. Un-supervised Learning
   - This method helps in finding the hidden patterns.
   - Does not need supervision to train the data.

- Results are still accurate but not as accurate as supervised learning.

4. Explanation and differences
   a. **Keyword based Search.**
      This is probably the most common method used.
      In this method we will try to find something verbatim and if those terms are found within the dataset we will find our target and use as input.

   b. **Semantic based Search**
      In this method we go a step further than above. Our goal is not only to find the keyword but also the intent behind that word.
      We need to add another layer to classify the results based on the searches that hit and classify the intent behind those matches.

   c. **Contextual based search**
      These results have extra level with the once mentioned below. This is specific for every targeted user depending upon various data points such as interest, date, place, time etc.
      Example – Restaurant near me

5. Big Data – Data that meets the threshold values of variety, velocity and volume can be determined as big data. Once the combination of variety, velocity and volume becomes incomprehensible for normal computers the data can be characterized as big data.

   a. Structured – Names, address, phone numbers where the data is categorized.
   b. Unstructured – Document collections. Invoices, records, emails, productivity applications.
   c. Semi structured data – Webpages, mailboxes and Spams.

6. **Data Clustering** : Data Clustering is a technique, which groups the unlabeled dataset. We can say a way of grouping the data points into different clusters, consisting of similar data points.

   Data Classification : Data classification is the process of organizing data by relevant categories, to make it easy to find, store, and analyze.

   Building Recommendation System : Recommendation systems are typically used to rank or rate products and consumers. A recommender system is a program that predicts how a user will rate a certain item. After then, the predictions will be ranked and returned to the user. For example, in a movie recommender system, clustering is sometimes used to present the user with recommendations based on his preferences. Its main goal is to suggest an item to a user who has a good possibility of enjoying or needing it based on his previous purchases/choices.

   Mining Association Rules : The use of machine learning models to evaluate data in a

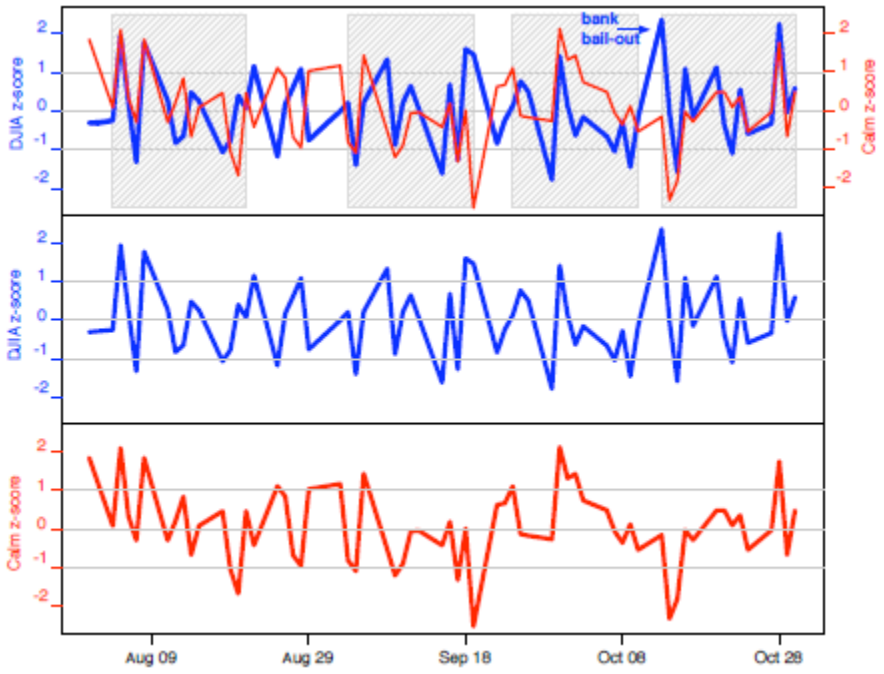
database for patterns, or co-occurrences, is known as association rule mining. It finds common if-then relationships, which are the association rules themselves.

5. Earthquake features used

    1. Keyword in a tweet
    2. Position of query word in tweet
    3. Context of target-event words
    4. The number of words

7. Twitter predicts mood.
   Summary -
   This was a Paper that used A Granger causality analysis where they were able to accurately show the graph of user mood was matching a lagging graph of the stock market which made the mood a Granger Causality of the stock market.
   Stock market = St
   Mood = Mt

   St+1 is dependent on bot St as well as Mt which proves that St is a granger Casualty of Mt and the results can be seen in the graph below.



8. **User based collaborative filtering** – The system creates a group of projects with similar tastes and then uses the ML model to do recommendations based on the group. This generally requires much more data and much more compute power.

Item Based Collaborative In this type of system the recommendation engine looks at what products other users have generally looked at after the current item. It will always look for current items and match with historical items. For ex – similar examples.

9. Alternative data is a data field that is collected from non-traditional data which provides an indication of future performance of a company excluding the common datasets used.
    a. In the world of Finance
       This has been an absolute game changer especially for hedge funds. Here they try to trac the web and app traffic to predict the mood after investment or sale of any particular commodity.

    b. Epidemiology – Here they simply used web site tracking for targeted users. Examples – Where they were searching for symptoms but collecting the non conventional data as for how long they were reading or surfing through medicines and related data in mobile apps as well as other related websites.

10. The early success was defined as Google was able to predict an epidemic 2 weeks prior to the CDC. But as soon as the 2 weeks went by, it was branded as a failure as it showed close to twice the number of results than the actual.
    This dramatic error was made in 2013 where after following social media news, people were searching for flues such as influenza without any symptoms.
    Later this error was fixed by combining the data of google and CDC.
11. Ss
12. They used Granger Causality to overcome this hurdle. They investigated the relation between DJIA values and all Twitter mood dimensions measured by GPOMS and OpinionFinder.
    Once they were able to prove Yt was a Granger Causality of Xt.

    Once they did this, they were no longer proving the causality but the actual causation but whether one time series has predictive information about the other or not.