

## Data Science: Key Concepts

In this chapter we will look at the five disruptions that are caused in the market place by data science. Once the context and its importance is understood it's easy to simplify and demonstrate what data science actually is. Then we will also study traditional architecture versus Data science architecture and understand the importance of Signal detection, which we shall study in chapter 3 and the machine learning techniques that help with this signal detection is studied from chapter 9 onwards, although we have covered few machine learning concepts in this chapter. This chapter shall also discuss solution architecture and the three critical components that are required for any solution.

### Five Disruptive Products

The five quick disruptive products launched in the market place will be discussed now:

1. A very simple Japanese App
2. Healthcare App
3. Coursera
4. Sensory device in Agriculture Sector
5. Autonomous Car

### The Japanese App

The first one is a very simple Japanese app, which essentially helps two people to discover each other. Essentially, what the App does is, for every individual a set of questions has to be answered. When these questions are answered it gives a characteristics score that tells if the person likes music, books, viewpoints on philosophy, religion etc. Whatever the parameters are, the questions have to be answered and each person gets a score attached to each question answered.

The other score that is attached to this device is the location. If a device is carried while walking on the street it will tell how many people with similar scores are around you within a 1 km radius. This app will enable strangers to find one another and have coffee, chat or get to know one another better. Using similarity score and location they are able to discover one another.

## 2 INTRODUCTION TO DATA SCIENCE USING 'R'

**Disruption:** An app that leveraged and capitalized on new social norms of today's casual meetups. Revolutionized the way people find others with similar taste/interests. Usage of data to find patterns and clusters from humongous set of entries and present to the users in a meaningful way, which is 'right match' in this case. Turning Data to Insights.



FIGURE 1.1 Japanese dating app

---

### The Healthcare App

The second one is in the healthcare space. In this healthcare app a heart implant is able to communicate information such as rate of heartbeat, condition of heart in real time with your mobile phone. The mobile app also communicates remotely to the doctor.

**Disruption:** Reduction in visits to the clinic, reduction in non-medical costs. Continuous monitoring of organ health vs. one time data captured during the physician visit. Presents an opportunity to track patterns and higher chance of identifying an anomaly and hence act early/on time.

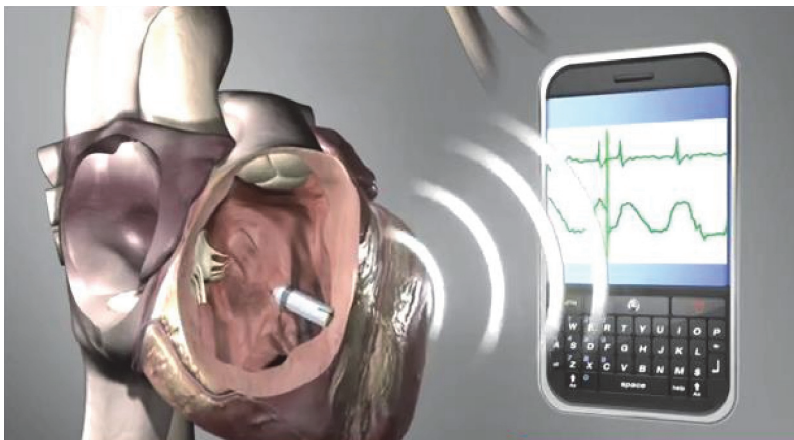


FIGURE 1.2 Heart implants

---

## Coursera

The third disruptive product is Coursera, an online educational platform where one can learn various kinds of courses for free. There are a lot of educational videos and tutorials online. When students watch these videos it is possible to pinpoint those places in the video when students pause or stop. Those jump and exit points are noted and this enables to figure out how to re-orchestrate the content, to make the content more engaging.

**Disruption:** While MOOCs have expanded the access to education to learners by overcoming lack of infrastructure/resources, COURSERA aimed to continuously improve the quality of the content delivered by collecting data on focus/topics of interest from thousands of students from across the world. By redesigning UX, and fine tuning content COURSERA disrupted the way online education was delivered by its predecessors like Khanacademy, MIT OCW, etc.

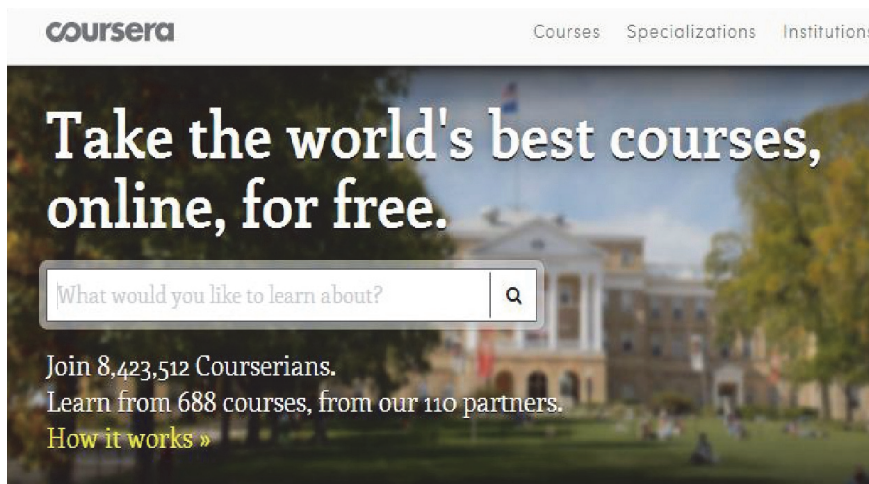


FIGURE 1.3 MOOC

## Sensory Device in Agriculture Sector

Fourth, disruptive product is in the Agriculture sector. Netherlands agriculture is a big part of their economy. They make the worlds best cheese and butter. One of the problems farmers face there is understanding the health of cows, which are carrying. Therefore now they have attached a sensory device to the cow's ears, through which farmers can remotely (communicated via a satellite), monitor their cow's health.

**Disruption:** Livestock farming techniques and the sensors help with cattle health monitoring and action can be taken immediately if the cattle are unwell. This helps within time detection of disease and helps prevention of spread of disease to the other cows through prediction.



FIGURE 1.4 Sensored Cows in Netherland

---

### Autonomous Car

Lastly, the autonomous car, an autonomous car is special in that the car moves without a driver. This device tracks and scans the surroundings of the car at high speeds. It has the intelligence to process all kinds of real-time information and communicates it back to the steering wheel.

**Disruption:** Processing data from images and supplementary sensors, self-driving cars create a virtual world through which they navigate. By reducing the reaction time by millions of folds than human level, they aim to eliminate human error driven accidents and traffic congestions. Significant improvement in time and fuel efficiency whilst saving lives.



FIGURE 1.5 Google's autonomous car

---



A look at all the five uses shows one thing that is common to all of these and that is a data product which is working behind the scenes, very silently humming. To create a data product a data science process is needed, which will unlearn patterns from that data and create a bigger product. So in the five examples that happen in our everyday like how our health gets taken care of, how we learn, how we fall in love, how we farm and how we drive, all of these are touched increasingly by data products. Data science needs to be an integral part of any organization you consider, else there is a very high probability that you will lose the market place.

One of the biggest secrets of winners is that they are able to see patterns faster. So a core team, which uses data science techniques to process all the structured, unstructured data and looks at patterns around it and acts on it in real time is what most companies are aiming at today.

## **Data Science Vs Traditional Methods**

It's similar to an iceberg floating on water. Most organizations just see the tip of the iceberg. For example they just know how much sales is happening. They fail to realize what is driving sales. If there is a change in the promotions by 5% what is the expected growth in sales? There are lots of unknown questions for which answers are required.

Most organizations have tons of data on sales, finance aspects; call centre data and reports, which are typically delivered on Business Objects, Cognos, and Microsoft Analysis Services. These reports quickly answer few important basic questions such as which call centre agent has the best all round time. What happens in Data science is inserting a process called analytical modeling process where there are specific techniques such as segmentation, scoring models, text-mining models, which will process the data and give a different lens. This will enable one to see patterns in the data.

## **Difference in Architecture**

Here is a detailed architecture of traditional companies versus the new age companies. Both of them have a Data Repository and a Dashboard but where they are different is in the four layers. There is Machine Learning Process (Text Mining, Collaborative filtering) in-between the data repository and Dashboards, which will change the game. They detect what is called a signal. A Signal is nothing but a pattern, so once the pattern is detected via an action, they keep a close watch on that action. This is a simplified view of the Data science architecture.

### What are 4 core differences between DataScience & Dashboards?

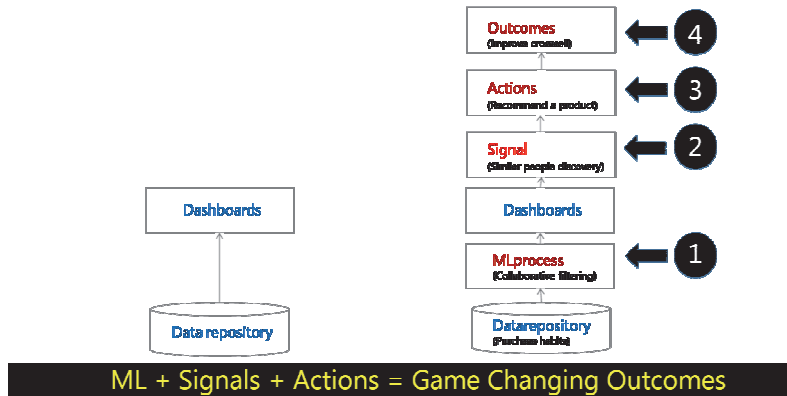


FIGURE 1.6 4 core differences between data science and dashboards

## Demystifying Machine Learning

The goal of Data scientist is to use data to discover signals that cause changes and which ultimately have an impact on the revenue of the firm. Even for a data scientist, it is humanly impossible to analyze big data. But with the aid of a computer, it can be easily done. Yet, a computer can only compute what has been programmed into it. So how do data scientists cope with this scenario, where analysis of the data will require the computer to pick up the 'trends' on its own? This is where machine learning comes in.

Machine Learning is a remarkable application of artificial intelligence that enables computing systems to perform tasks through a process of “self-learning” without their being specifically programmed for the same. As data scientists cannot pinpoint exactly what sorts of patterns, the computer should recognize, this application of “machine learning comes in extremely handy. Thus, machine learning facilitates the computer to automatically adapt to new patterns and signals in data, while “learning” or recognizing previous trends and data computations. When Google’s search bar uses “auto-complete” before you type in your query, it is an example of machine learning, as the Google server has learnt to give you ‘predictions’ of what you might want to search based on your previous search history.

We will now familiarize with five techniques


### Technique 1: Segmentation

This process involves breaking data into various chunks based on shared characteristics. The analyst then picks the clusters through an iterative process looking for uniqueness between segments. We could segment based on demographic, need based, behavior based etc. The statistical techniques that we use for segmentation are K Means, Hierarchical clustering and Discriminant analysis, as shown in figure 1.7.

Some business questions that are answered by segmentation are:

- What are the behavioral personas about customer, which lie buried in my raw customer transactions in the database? This is explained in Figure 1.8
- Which specific customer behavior discriminates a high value segment from low value segment? This is explained in Figure 1.9
- How do customer behavior segments migrate across time and what does it reveal to us? This is explained in Figure 1.10 and 1.11

## A real life customer segmentation case study



- Customer Context
  - A large owner of fleets in US
  - Each truck driver given a fuel card
  - Driver info + Mileage + Refuelling behaviour + Location
- Customer Challenge
  - Aligning Service Models to Customer Segments
  - Drive Growth & Ability to Cross-sell & Up-sell
- Data Science Technique
  - K means clustering
  - Analysed over 120,000,000 Customer Records & Profiles
  - Analysed over 110,000,000 Million Customer Service Rep Comments

FIGURE 1.7 A Real Life customer segmentation case study

### Fleet related master data

1. Fleet id
2. SIC Code
3. No of trucks in fleet
4. No of drivers/cards

### Fleet spend data

1. Avg\_Gallons\_Per\_month
2. Avg\_Spend\_on\_non\_fuel
3. Avg\_Transaction\_Per\_Month
4. Total\_Active\_Cards
5. MOM(3 months) growth(gallons)
6. Avg\_Credit\_utilization (3 months)

### Current product holdings flag

1. Has\_OPIS\_Suite\_of\_Reports\_flag
2. Has\_EFPS\_Discount
3. Has\_Smart
4. Has\_Rewards
5. Has\_Screen\_Now\_Report
6. Has\_Volume\_or\_Service\_Discount
7. Has\_Exception\_Reports

### Touch point data

1. Avg No of inbound calls per month
2. Recency of last call
3. Total no of phone calls per year

FIGURE 1.8 Behavioral components considered for fleet card segmentation

Dimensions of fleet behavior measured and segmented

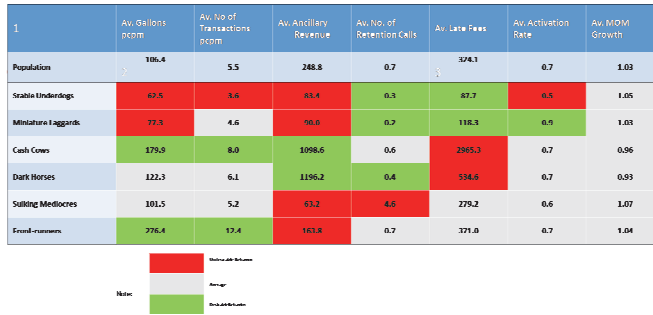


FIGURE 1.9 Dimensions of fleet behavior measured and segmented

Segment-3: Cash Cows...Segment Profile

- Definition: The large size fleets, that are mostly medium tenure customers having very high spends but also having high late fees incidences
- Constitutes 5% of total fleets and contributes 22% of total spend.

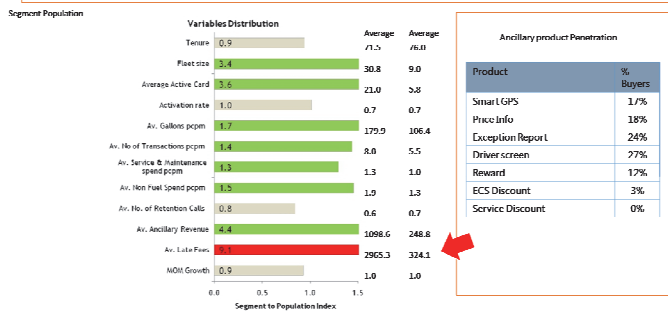


FIGURE 1.10 Cash cow - segment profile

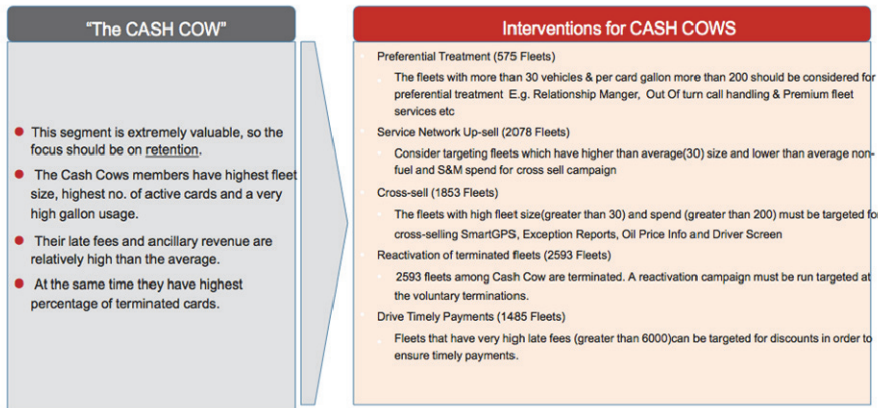


FIGURE 1.11 Cash cow – behavior portrait and target action

## Segmenting in BANKING Industry

In order to give the right offer and product to the right customer and to do it the efficient way you will need to use a segmentation method. In banking we could classify and segment the customers into 5 clusters and their line of credit, pricing and campaign intervention for each segment can be studied as seen in the graph 1.12

### Clustering

It is considered the most important unsupervised learning problem. Cluster analysis is in simple language dividing data into different clusters or groups.

## Segmentation in Banking Industry



### Key cluster observations

- Cluster Observation-1: Low balance, Lowrisk, Reached credit limit often
- Possible treatment strategy: Extend Line of credit and possibly charge fixed fee depending on # of times they reach credit limit
- Cluster observation-2: Low balance, moderate risk, reach credit limit often
- Possible treatment strategy: Possibly charge fixed fee depending on # of times they reach credit limit
- Cluster observation-3: High balance, moderate risk, Do not reach credit limit often
- Possibly run a focused outbound campaign to sell short term fixed deposit
- Cluster observation-4&5: Moderate balance, High risk, Moderate usage
- Since risk is high, interest rates and Pricing strategy

5 segments and LOC, Pricing, Campaign interventions for each customer segment

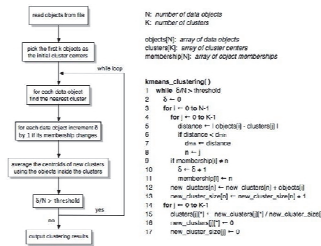
**FIGURE 1.12 Segmentation in banking industry**

The greater the similarity within a group the better is the cluster. The greater the dissimilarity between groups the cluster is more distinct. One technique of clustering is the k means technique. This technique is used to separate data into the best-suited group based on information the algorithm already has. Once data is separated one has to specify the number of cluster that will be created to be able to produce effective data mining results. Each cluster had a centre point called the centroid, which each observation is assigned to. Associating every observation with the nearest mean creates K clusters.

Then one has to calculate the centroid mean for each cluster. This becomes the new mean and the above two steps are repeated till convergence has been reached.



## The Mathematics behind Clustering



- K means algorithm
- Specify K the number of clusters to create
- Choose K points at Random as Cluster centroids
- Assign each observation to the cluster centroid it is closest to
- Calculate centroid mean for each cluster
- Use it as the new centroid of the cluster
- Iterate till cluster centre does not change

FIGURE 1.13 The mathematic behind clustering

## Technique 2: Unstructured Text Mining

The second technique is unstructured text mining. Here we use data to discover signals and process changes that create an impact. So lets take a store manager as an example. A store manager gets a lot of feedback and there is a lot of unstructured data that comes in. It's very important to process this into a structure. Text mining can process all this feedback and give a glimpse of what is called the sentiment analysis as how many people like the store and how many don't like the store.

Another example of text mining is in the health care domain. For example, a patient goes to a doctor and gets admitted to a hospital. He comes in contact with a doctor, lab technicians and the nurse. The Doctor makes a record of the state of his health condition, the Lab technician tests his blood and writes down some inferences and the nurse regularly checks his vital parameters and makes notes of the health condition of the patients. Now all this data can be run through a text mining activity and triangulate the state of the patient by mashing up all the three datasets, the doctors dataset, the lab technicians dataset and the nurse's dataset.

### Real world Unstructured text mining in health care

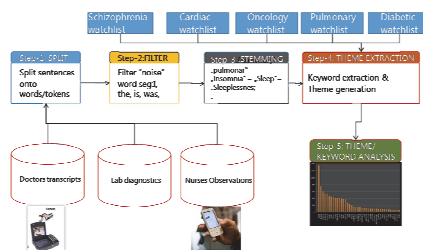


FIGURE 1.14 Real world unstructured text mining in the health care

Another example of text mining is insurance domain. Auto insurance companies deal with large number of claims every day from collision damage, fire and theft damage and accidental damage. Large amount of time and money is spent in identifying fraudulent claims. One of the datasets used is a historical claim, coverage and settlements dataset. If we have considered taking this dataset for our analysis, A text mining solution layer is added above the structured data that quickly reads and understands the claim details to highlight missing facts, inconsistencies and changed stories to identify with a probability that the claim is fraudulent.

### Technique 3: Scoring for Signal Processing

The third technique is scoring models. Scoring refers to what is the best action that a person is likely to take? For example most of us have shopped at Flip kart at least once. Flip kart tries to figure out what is the next best action that a customer is likely to perform; will one be redeeming a coupon or is the customer just browsing and adding items to the wish list? Studying the past history and trying to characterize the future behavior is a scoring model. It is a binary result, its either an action or not, in other words 1 or 0.

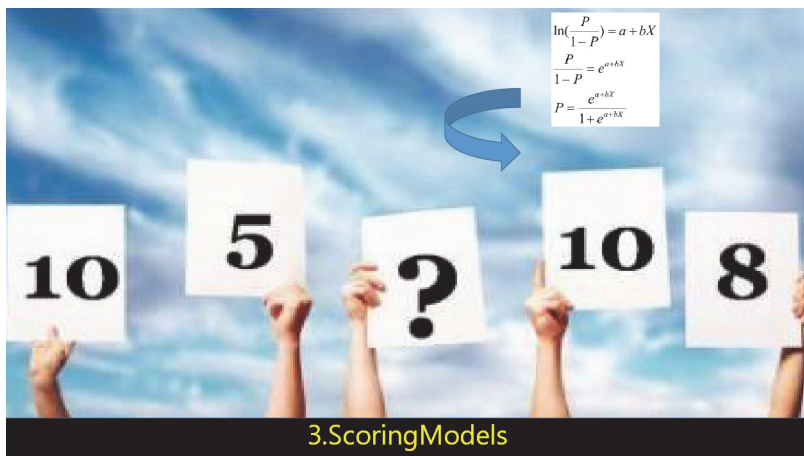


FIGURE 1.15 Scoring models

### Technique 4: Forecasting

Fourth technique is Forecasting. When trying to predict the sales for the next three months, if the outcome is a number then this is Forecasting as seen in Figure 1.16.

### Technique 5: Recommenders

The fifth technique is Recommenders. Looking at recommenders, if a person likes pizza and salad and another person likes pizza, salad and coke. Based on the actual purchase behavior one will club these two people as similar to each other. As both people have



### Snapshot of Machine Learning Techniques & Reference Architecture

As in the above topic we have just seen few machine learning technique below is a graph that displays a snapshot of six techniques and a brief idea of what falls under each of the techniques. We will study Machine learning as a whole topic in chapter 9.

## Snapshot of Machine Learning Techniques

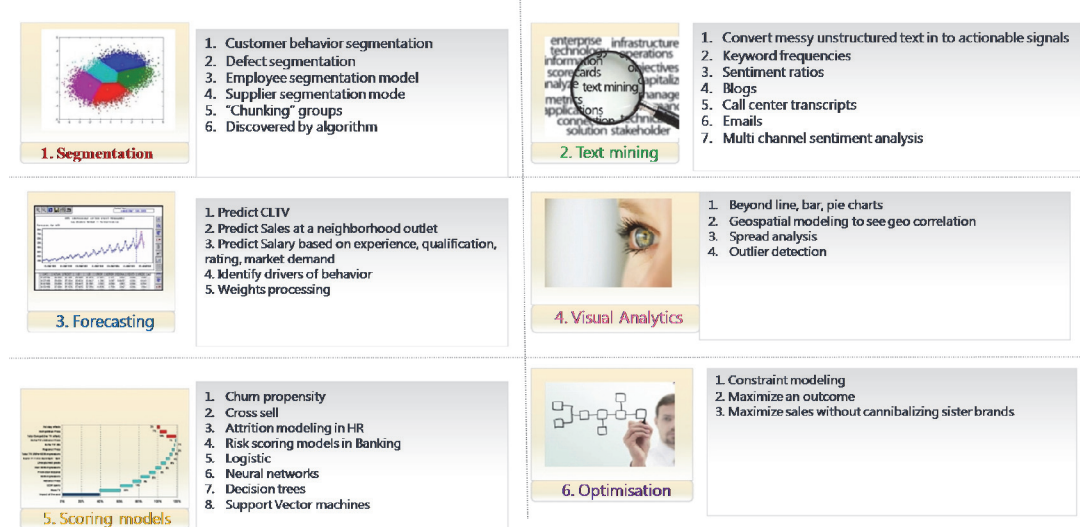


FIGURE 1.18 Machine learning Techniques

### Reference Architecture

A look at the reference architecture for any of these projects has three layers. Machine Learning Reference Architecture is an algorithm that can process raw data to provide a big picture that combines all the major and minor aspects of the data being analyzed. Real time data analysis is carried out using an optimization process,

1. The store layer that captures and stores data: where one can store the data in Hadoop, Hive, Hana or any other database. More important than storing is what is done with the data, how to extract signals from the data and that's where data science comes in.
2. The Sense Layer: in this layer a text-mining model or a scoring model is used to detect a pattern. Then it mines the collected data from historical trends and patterns that act as reference points, this pattern is then monetized.
3. The Respond Layer: compare previous trends with the latest data collected, to predict an outcome and to recommend the next decision to be taken. This is where the analysis of the data is presented in relatable terms, and the patterns are contextualized in terms of the business/company's needs. Analytic workflows are computed and insights are generated.

## Machine Learning Reference Architecture

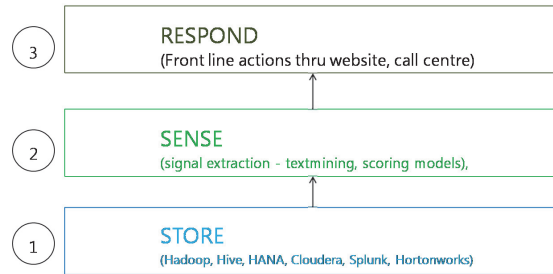


FIGURE 1.19 Machine learning reference architecture

### Hands on Segmentation

You will learn the path to download R in chapter 4, here you can have a quick view of some commands in R specific to clustering technique and its visualization.

1. To segment the dataset we use the “kmeans” command to find clusters. To view the cluster statistics we input the “fit” command.

### Hands on Segmentation... Using K means to find clusters

Execute clustering algorithm

```
> fit <- kmeans(retail_data_clustering, 5)
> show(fit)
```

Shows cluster statistics

Is it above or below population average? How does this help characterise the segment?

Shows membership of each segment

```
K-means clustering with 5 clusters of sizes 61, 35, 78, 74, 52

Cluster means:
  AvgPurchValue  AvItemPurchased  ShopFrequency  Spendtilldate
1      38170.67         6.803279      3.622951      50209.34
2      48370.06        12.771429      8.457143     140760.31
3      49059.06        11.846154      5.487179     96592.00
4       17516.01         9.756757      5.621622     81245.81
5       16695.88        12.326923      8.173077    133726.56

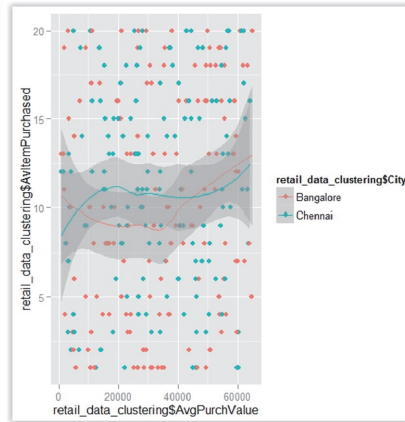
Clustering vector:
 [1] 4 4 4 2 3 2 3 3 2 3 4 3 3 5 4 4 4 3 3 1 3 2 3 5 4 4 5 3 2 5 5 2 3 3 1 2 5 2
 [38] 5 3 1 4 2 5 4 3 4 1 5 3 2 2 4 2 4 4 4 1 4 5 3 3 4 1 1 4 1 3 1 2 1 3 5 4 5
 [75] 5 3 1 1 3 4 3 4 3 2 3 5 1 3 4 5 2 5 2 1 2 1 1 1 1 1 5 1 4 2 3 4 4 3 4 4 3
 [112] 1 5 5 4 1 5 3 1 3 3 4 3 1 4 4 1 5 1 2 1 3 4 5 4 3 5 5 4 4 5 3 1 1 1 1 4 5
```



- To visualize the data we use the “ggplot” command. On the X axis we have plotted average purchase value and on the Y axis we have plotted average items purchased for two cities Bangalore and Chennai.

### ggplot() – How to draw a quick scatter plot? Visual relationship

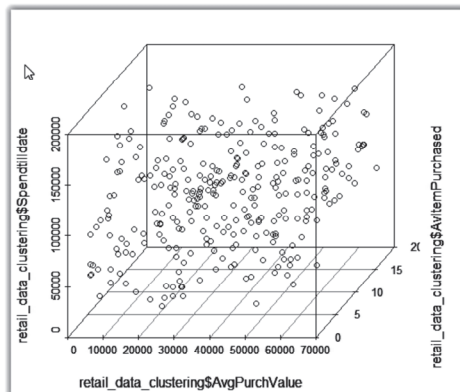
```
> ggplot( retail_data_clustering, aes(x=retail_data_clustering$AvgPurchValue,
+ y=retail_data_clustering$AvItemPurchased,color = retail_data_clustering$City ))+
+ geom_point() +
+ theme_minimal()
```



- Another way of visualizing is in 3D format as shown in the figure below using the “scatterplot3d” command.

### 3D Visualisation

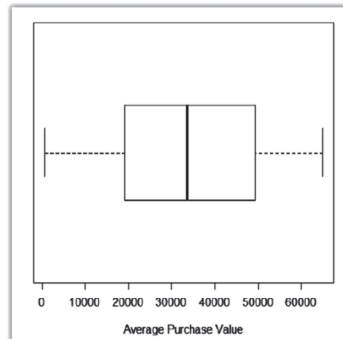
```
> scatterplot3d( retail_data_clustering$AvgPurchValue, retail_data_clustering$AvItemPurchased, retail_data_clustering$Spendtilldate)
```



4. We could draw a box plot to analyze the spread by inputting the “boxplot” command for the retail data.

boxplot()—How to draw a quick box plot to analyse spread?

```
> boxplot(retail_data_clustering$AvgPurchValue, horizontal = TRUE, xlab = "Average Purchase Value")
```



## Summary

Any organization today is drowning in data. They have tones of data with them. Its clearly evident in the job market that the role of a data scientist is scarce and a look at the job portals shows how companies such as KPMG, Pfizer etc., are all recruiting data scientists. Data is an asset that can add value to their organizations, and can help them predict future market demands. Using this data is a must to improve decision-making, and it can only be done through data analysis. Traditional methods of analyzing data can no longer be used because of the huge amount of data flowing in, making it unavoidable to use Data Science techniques, and specifically Machine Learning.

While data mining has been used earlier to discover previously hidden information and anomalies, Machine learning goes one step further by using the stored information collected previously to predict future trends. Undoubtedly, Data Science is indispensable to businesses that deal in share markets, advertising, political campaigning, voter behavior and advanced medicine or any other domain. With more and more efficient algorithms being developed, it is possible to filter the necessary information in today's data-rich world that can aid in making vital decisions. As an introduction to data science we have understood why data science is important, and the option to make this as a career as data products surround us. We have also touched five situations to understand this.