Retail
Customer relationship building is critical to the retail industry – and the best way to manage that is to manage big data. Retailers need to know the best way to market to customers, the most effective way to handle transactions, and the most strategic way to bring back lapsed business. Big data remains at the heart of all those things.

1.2. BIG DATA ANALYTICS - OVERVIEW [2]

The volume of data that one has to deal has exploded to unimaginable levels in the past decade, and at the same time, the price of data storage has systematically reduced. Private companies and research institutions capture terabytes of data about their users' interactions, business, social media, and also sensors from devices such as mobile phones and automobiles. The challenge of this era is to make sense of this sea of data. This is where big data analytics comes into picture.

Big Data Analytics largely involves collecting data from different sources, munge it in a way that it becomes available to be consumed by analysts and finally deliver data products useful to the organization business.

The process of converting large amounts of unstructured raw data, retrieved from different sources to a data product useful for organizations forms the core of Big Data Analytics.

1.3. BIG DATA ANALYTICS - DATA LIFE CYCLE [2]

1.3.1 Traditional Data Mining Life Cycle [2]

In order to provide a framework to organize the work needed by an organization and deliver clear insights from Big Data, it's useful to think of it as a cycle with different stages. It is by no means linear, meaning all the stages are related with each other. This cycle has superficial similarities with the more traditional data mining cycle as described in CRISP methodology.

CRISP-DM Methodology
The CRISP-DM methodology that stands for Cross Industry Standard Process for Data Mining, is a cycle that describes commonly used approaches that data mining experts use to tackle problems in traditional BI data mining. It is still being used in traditional BI data mining teams.
Take a look at the Fig.1.3. It shows the major stages of the cycle as described by the CRISP-DM methodology and how they are interrelated.

CRISP-DM was conceived in 1996 and the next year, it got underway as a European Union project under the ESPRIT funding initiative. The project was led by five companies: SPSS, Teradata, Daimler AG, NCR Corporation, and OHRA (an insurance company). The project was finally incorporated into SPSS. The methodology is extremely detailed oriented in how a data mining project should be specified.

Let us now learn a little more on each of the stages involved in the CRISP-DM life cycle

Business Understanding − This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition. A preliminary plan is designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.

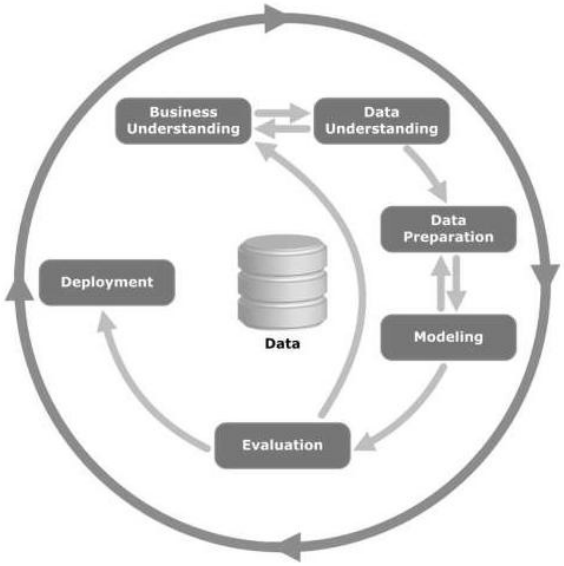

Figure 1.3 major stages of the cycle as described by the CRISP-DM methodology

Data Understanding − The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation − The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed

order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling − In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, it is often required to step back to the data preparation phase.

Evaluation − At this stage in the project, you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model thoroughly and review the steps executed to construct the model, to be certain it properly achieves the business objectives.
A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment − Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer.
Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process.

In many cases, it will be the customer, not the data analyst, who will carry out the deployment steps. Even if the analyst deploys the model, it is important for the customer to understand upfront the actions which will need to be carried out in order to actually make use of the created models.

SEMMA Methodology
SEMMA is another methodology developed by SAS for data mining modeling. It stands for Sample, Explore, Modify, Model, and Asses. Here is a brief description of its stages −

Sample − The process starts with data sampling, e.g., selecting the dataset for modeling. The dataset should be large enough to contain sufficient information to retrieve, yet small enough to be used efficiently. This phase also deals with data partitioning.

Explore − This phase covers the understanding of the data by discovering anticipated and unanticipated relationships between the variables, and also abnormalities, with the help of data visualization.

Modify − The Modify phase contains methods to select, create and transform variables in preparation for data modeling.

Model − In the Model phase, the focus is on applying various modeling (data mining) techniques on the prepared variables in order to create models that possibly provide the desired outcome.

Assess − The evaluation of the modeling results shows the reliability and usefulness of the created models.

The main difference between CRISM–DM and SEMMA is that SEMMA focuses on the modeling aspect, whereas CRISP-DM gives more importance to stages of the cycle prior to modeling such as understanding the business problem to be solved, understanding and preprocessing the data to be used as input, for example, machine learning algorithms.

1.3.2 Big Data Life Cycle [2]

In today's big data context, the previous approaches are either incomplete or suboptimal. For example, the SEMMA methodology disregards completely data collection and preprocessing of different data sources. These stages normally constitute most of the work in a successful big data project.

A big data analytics cycle can be described by the following stage

- Business Problem Definition
- Research
- Human Resources Assessment
- Data Acquisition
- Data Munging
- Data Storage
- Exploratory Data Analysis
- Data Preparation for Modeling and Assessment
- Modeling
- Implementation

Business Problem Definition

This is a point common in traditional BI and big data analytics life cycle. Normally it is a non-trivial stage of a big data project to define the problem and evaluate correctly how much potential gain it may have for an organization. It seems obvious to mention this, but it has to be evaluated what are the expected gains and costs of the project.

Research

Analyze what other companies have done in the same situation. This involves looking for solutions that are reasonable for your company, even though it involves adapting other solutions to the resources and requirements that your company has. In this stage, a methodology for the future stages should be defined.

Human Resources Assessment

Once the problem is defined, it's reasonable to continue analyzing if the current staff is able to complete the project successfully. Traditional BI teams might not be capable to deliver an optimal solution to all the stages, so it should be considered before starting the project if there is a need to outsource a part of the project or hire more people.

Data Acquisition

This section is key in a big data life cycle; it defines which type of profiles would be needed to deliver the resultant data product. Data gathering is a non-trivial step of the process; it normally involves gathering unstructured data from different sources. To give an example, it could involve writing a crawler to retrieve reviews from a website. This involves dealing with text, perhaps in different languages normally requiring a significant amount of time to be completed.

Data Munging

Once the data is retrieved, for example, from the web, it needs to be stored in an easyto-use format. To continue with the reviews examples, let's assume the data is retrieved from different sites where each has a different display of the data.

Suppose one data source gives reviews in terms of rating in stars, therefore it is possible to read this as a mapping for the response variable $y \in \{1, 2, 3, 4, 5\}$. Another data source gives reviews using two arrows system, one for up voting and the other for down voting. This would imply a response variable of the form $y \in \{\text{positive, negative}\}$.

In order to combine both the data sources, a decision has to be made in order to make these two response representations equivalent. This can involve converting the first data source response representation to the second form, considering one star as negative and five stars as positive. This process often requires a large time allocation to be delivered with good quality.

Data Storage

Once the data is processed, it sometimes needs to be stored in a database. Big data technologies offer plenty of alternatives regarding this point. The most common alternative is using the Hadoop File System for storage that provides users a limited version of SQL, known as HIVE Query Language. This allows most analytics task to be

done in similar ways as would be done in traditional BI data warehouses, from the user perspective. Other storage options to be considered are MongoDB, Redis, and SPARK.
This stage of the cycle is related to the human resources knowledge in terms of their abilities to implement different architectures. Modified versions of traditional data warehouses are still being used in large scale applications. For example, teradata and IBM offer SQL databases that can handle terabytes of data; open source solutions such as postgreSQL and MySQL are still being used for large scale applications.

Even though there are differences in how the different storages work in the background, from the client side, most solutions provide a SQL API. Hence having a good understanding of SQL is still a key skill to have for big data analytics.

This stage a priori seems to be the most important topic, in practice, this is not true. It is not even an essential stage. It is possible to implement a big data solution that would be working with real-time data, so in this case, we only need to gather data to develop the model and then implement it in real time. So there would not be a need to formally store the data at all.

Exploratory Data Analysis
Once the data has been cleaned and stored in a way that insights can be retrieved from it, the data exploration phase is mandatory. The objective of this stage is to understand the data, this is normally done with statistical techniques and also plotting the data. This is a good stage to evaluate whether the problem definition makes sense or is feasible.

Data Preparation for Modeling and Assessment
This stage involves reshaping the cleaned data retrieved previously and using statistical preprocessing for missing values imputation, outlier detection, normalization, feature extraction and feature selection.

Modelling
The prior stage should have produced several datasets for training and testing, for example, a predictive model. This stage involves trying different models and looking forward to solving the business problem at hand. In practice, it is normally desired that the model would give some insight into the business. Finally, the best model or combination of models is selected evaluating its performance on a left-out dataset.

Implementation
In this stage, the data product developed is implemented in the data pipeline of the company. This involves setting up a validation scheme while the data product is working, in order to track its performance. For example, in the case of implementing a predictive

model, this stage would involve applying the model to new data and once the response is available, evaluate the model.

1.4. BIG DATA ANALYTICS - METHODOLOGY [2]

In terms of methodology, big data analytics differs significantly from the traditional statistical approach of experimental design. Analytics starts with data. Normally we model the data in a way to explain a response. The objective of this approach is to predict the response behavior or understand how the input variables relate to a response. Normally in statistical experimental designs, an experiment is developed and data is retrieved as a result. This allows to generate data in a way that can be used by a statistical model, where certain assumptions hold such as independence, normality, and randomization.

In big data analytics, we are presented with the data. We cannot design an experiment that fulfills our favorite statistical model. In large-scale applications of analytics, a large amount of work (normally 80% of the effort) is needed just for cleaning the data, so it can be used by a machine learning model.

We don't have a unique methodology to follow in real large-scale applications. Normally once the business problem is defined, a research stage is needed to design the methodology to be used. However general guidelines are relevant to be mentioned and apply to almost all problems.

One of the most important tasks in big data analytics is statistical modeling, meaning supervised and unsupervised classification or regression problems. Once the data is cleaned and preprocessed, available for modeling, care should be taken in evaluating different models with reasonable loss metrics and then once the model is implemented, further evaluation and results should be reported. A common pitfall in predictive modeling is to just implement the model and never measure its performance.

1.5. ANALYTICS PROCESS MODEL [3]

Figure 4 gives a high-level overview of the analytics process model.[4] As a first step, a thorough definition of the business problem to be solved with analytics is needed. Next, all source data need to be identified that could be of potential interest. This is a very important step, as data is the key ingredient to any analytical exercise and the selection of data will have a deterministic impact on the analytical models that will be built in a subsequent step. All data will then be gathered in a staging area, which could be, for

example, a data mart or data warehouse. Some basic exploratory analysis can be considered here using, for example, online analytical processing (OLAP) facilities for multidimensional data analysis (e.g., roll-up, drill down, slicing and dicing). This will be followed by a data cleaning step to get rid of all inconsistencies, such as missing values, outliers, and duplicate data. Additional transformations may also be considered, such as binning, alphanumeric to numeric coding, geographical aggregation, and so forth. In the analytics step, an analytical model will be estimated on the preprocessed and transformed data. Different types of analytics can be considered here (e.g., to do churn prediction, fraud detection, customer segmentation, market basket analysis). Finally, once the model has been built, it will be interpreted and evaluated by the business experts. Usually, many trivial patterns will be detected by the model. For example, in a market basket analysis setting, one may find that spaghetti and spaghetti sauce are often purchased together. These patterns are interesting because they provide some validation of the model. But of course, the key issue here is to find the unexpected yet interesting and actionable patterns (sometimes also referred to as knowledge diamonds) that can provide added value in the business setting. Once the analytical model has been appropriately validated and approved, it can be put into production as an analytics application (e.g., decision support system, scoring engine). It is important to consider here how to represent the model output in a user-friendly way, how to integrate it with other applications (e.g., campaign management tools, risk engines), and how to make sure the analytical model can be appropriately monitored and back tested on an ongoing basis.

It is important to note that the process model outlined in Figure 1.4 is iterative in nature, in the sense that one may have to go back to previous steps during the exercise. For example, during the analytics step, the need for additional data may be identified, which may necessitate additional cleaning, transformation, and so forth. Also, the most time consuming step is the data selection and preprocessing step; this usually takes around 80% of the total efforts needed to build an analytical model.

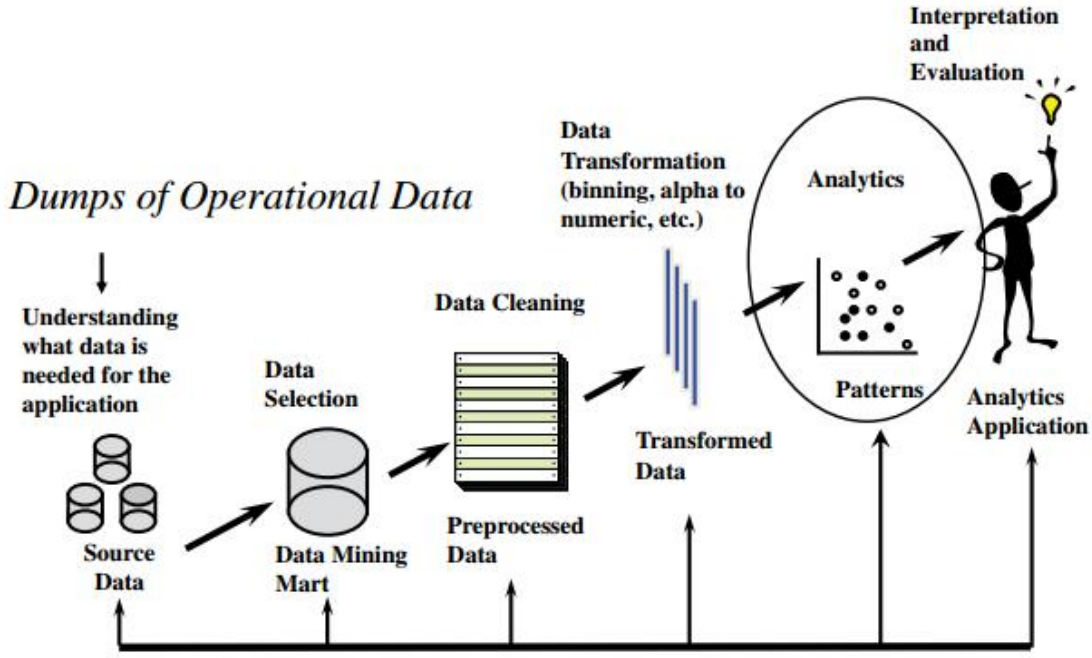

Figure 1.4 The Analytics Process Model [3]

1.6. ANALYTICAL MODEL REQUIREMENTS [3]

A good analytical model should satisfy several requirements, depending on the application area. A first critical success factor is business relevance. The analytical model should actually solve the business problem for which it was developed. It makes no sense to have a working analytical model that got sidetracked from the original problem statement. In order to achieve business relevance, it is of key importance that the business problem to be solved is appropriately defined, qualified, and agreed upon by all parties involved at the outset of the analysis.

A second criterion is statistical performance. The model should have statistical significance and predictive power. How this can be measured will depend upon the type of analytics considered. For example, in a classification setting (churn, fraud), the model should have good discrimination power. In a clustering setting, the clusters should be as homogenous as possible. In later chapters, we will extensively discuss various measures to quantify this.

Depending on the application, analytical models should also be interpretable and justifiable. Interpretability refers to understanding the patterns that the analytical model captures. This aspect has a certain degree of subjectivism, since interpretability may depend on the business user's knowledge. In many settings, however, it is considered to be a key requirement. For example, in credit risk modeling or medical diagnosis, interpretable models are absolutely needed to get good insight into the underlying data

patterns. In other settings, such as response modeling and fraud detection, having interpretable models may be less of an issue. Justifiability refers to the degree to which a model corresponds to prior business knowledge and intuition.[5] For example, a model stating that a higher debt ratio results in more creditworthy clients may be interpretable, but is not justifiable because it contradicts basic financial intuition. Note that both interpretability and justifiability often need to be balanced against statistical performance. Often one will observe that high performing analytical models are incomprehensible and black box in nature.

A popular example of this is neural networks, which are universal approximators and are high performing, but offer no insight into the underlying patterns in the data. On the contrary, linear regression models are very transparent and comprehensible, but offer only limited modeling power.

Analytical models should also be operationally efficient. This refers to the efforts needed to collect the data, preprocess it, evaluate the model, and feed its outputs to the business application (e.g., campaign management, capital calculation). Especially in a real-time online scoring environment (e.g., fraud detection) this may be a crucial characteristic. Operational efficiency also entails the efforts needed to monitor and backtest the model, and reestimate it when necessary. Another key attention point is the economic cost needed to set up the analytical model. This includes the costs to gather and preprocess the data, the costs to analyze the data, and the costs to put the resulting analytical models into production. In addition, the software costs and human and computing resources should be taken into account here. It is important to do a thorough cost–benefit analysis at the start of the project.

Finally, analytical models should also comply with both local and international regulation and legislation. For example, in a credit risk setting, the Basel II and Basel III Capital Accords have been introduced to appropriately identify the types of data that can or cannot be used to build credit risk models. In an insurance setting, the Solvency II Accord plays a similar role. Given the importance of analytics nowadays, more and more regulation is being introduced relating to the development and use of the analytical models. In addition, in the context of privacy, many new regulatory developments are taking place at various levels. A popular example here concerns the use of cookies in a web analytics context.

1.7. DATA COLLECTION, SAMPLING, AND PREPROCESSING [3]

Data are key ingredients for any analytical exercise. Hence, it is important to thoroughly consider and list all data sources that are of potential interest before starting the analysis. The rule here is the more data, the better. However, real life data can be dirty because of inconsistencies, incompleteness, duplication, and merging problems.

Throughout the analytical modeling steps, various data filtering mechanisms will be applied to clean up and reduce the data to a manageable and relevant size. Worth mentioning here is the garbage in, garbage out (GIGO) principle, which essentially states that messy data will yield messy analytical models. It is of the utmost importance that every data preprocessing step is carefully justified, carried out, validated, and documented before proceeding with further analysis. Even the slightest mistake can make the data totally unusable for further analysis. In what follows, we will elaborate on the most important data preprocessing steps that should be considered during an analytical modeling exercise.

1.7.1 Types of Data Sources [3]

As previously mentioned, more data is better to start off the analysis. Data can originate from a variety of different sources, which will be explored in what follows.

Transactions are the first important source of data. Transactional data consist of structured, low-level, detailed information capturing the key characteristics of a customer transaction (e.g., purchase, claim, cash transfer, credit card payment). This type of data is usually storedin massive online transaction processing (OLTP) relational databases.

It can also be summarized over longer time horizons by aggregating it into averages, absolute/relative trends, maximum/minimum values, and so on.

Unstructured data embedded in text documents (e.g., emails, web pages, claim forms) or multimedia content can also be interesting to analyze. However, these sources typically require extensive preprocessing before they can be successfully included in an analytical exercise.

Another important source of data is qualitative, expert-based data. An expert is a person with a substantial amount of subject matter expertise within a particular setting (e.g., credit portfolio manager, brand manager). The expertise stems from both common sense and business experience, and it is important to elicit expertise as much as possible before the analytics is run. This will steer the modeling in the right direction and allow you to interpret the analytical results from the right perspective. A popular example of applying expert-based validation is checking the univariate signs of a regression model. For example, one would expect a priori that higher debt has an adverse impact on credit risk, such that it should have a negative sign in the final scorecard. If this turns out not to be the case (e.g., due to bad data quality, multicollinearity), the expert/business user will not be tempted to use the analytical model at all, since it contradicts prior expectations.

Nowadays, data poolers are becoming more and more important in the industry. Popular examples are Dun & Bradstreet, Bureau Van Dijck, and Thomson Reuters. The core business of these companies is to gather data in a particular setting (e.g., credit risk, marketing), build models with it, and sell the output of these models (e.g., scores), possibly together with the underlying raw data, to interested customers. A popular

example of this in the United States is the FICO score, which is a credit score ranging between 300 and 850 that is provided by the three most important credit bureaus: Experian, Equifax, and Transunion. Many financial institutions use these FICO scores either as their final internal model or as a benchmark against an internally developed credit scorecard to better understand the weaknesses of the latter.

Finally, plenty of publicly available data can be included in the analytical exercise. A first important example is macroeconomic data about gross domestic product (GDP), inflation, unemployment, and so on. By including this type of data in an analytical model, it will become possible to see how the model varies with the state of the economy.

This is especially relevant in a credit risk setting, where typically all models need to be thoroughly stress tested. In addition, social media data from Facebook, Twitter, and others can be an important source of information. However, one needs to be careful here and make sure that all data gathering respects both local and international privacy regulations.

1.7.2 Sampling [3]

The aim of sampling is to take a subset of past customer data and use that to build an analytical model. A first obvious question concerns the need for sampling. With the availability of high performance computing facilities (e.g., grid/cloud computing), one could also directly analyze the full data set. However, a key requirement for a good sample is that it should be representative of the future customers on which the analytical model will be run. Hence, the timing aspect becomes important because customers of today are more similar to customers of tomorrow than customers of yesterday. Choosing the optimal time window for the sample involves a trade-off between lots of data (and hence a more robust analytical model) and recent data (which may be more representative). The sample should also be taken from an average business period to get a picture of the target population that is as accurate as possible.

It speaks for itself that sampling bias should be avoided as much as possible. However, this is not always straightforward. Let's take the example of credit scoring. Assume one wants to build an application scorecard to score mortgage applications. The future population then consists of all customers who come to the bank and apply for a mortgage—the so-called through-the-door (TTD) population. One then needs a subset of the historical TTD population to build an analytical model. However, in the past, the bank was already applying a credit policy (either expert based or based on a previous analytical model). This implies that the historical TTD population has two subsets: the customers that were accepted with the old policy and the ones that were rejected (see Figure 1.5). Obviously, for the latter, we don't know the target value since they were never granted the credit. When building a sample, one can then only make use of those that were accepted, which clearly implies a bias. Procedures for reject inference have

been suggested in the literature to deal with this sampling bias problem.[6] Unfortunately, all of these procedures make assumptions and none of them works perfectly. One of the most popular solutions is bureau-based inference, whereby a sample of past customers is given to the credit bureau to determine their target label (good or bad payer).

When thinking even closer about the target population for credit scoring, another forgotten subset is the withdrawals. These are the customers who were offered credit but decided not to take it (despite the fact that they may have been classified as good by the old scorecard). To be representative, these customers should also be included in the development sample. However, to the best of our knowledge, no procedures for withdrawal inference are typically applied in the industry.

In stratified sampling, a sample is taken according to predefined strata. Consider, for example, a churn prediction or fraud detection context in which data sets are typically very skewed (e.g., 99 percent nonchurners and 1 percent churners). When stratifying according to the target churn indicator, the sample will contain exactly the same percentages of churners and nonchurners as in the original data.

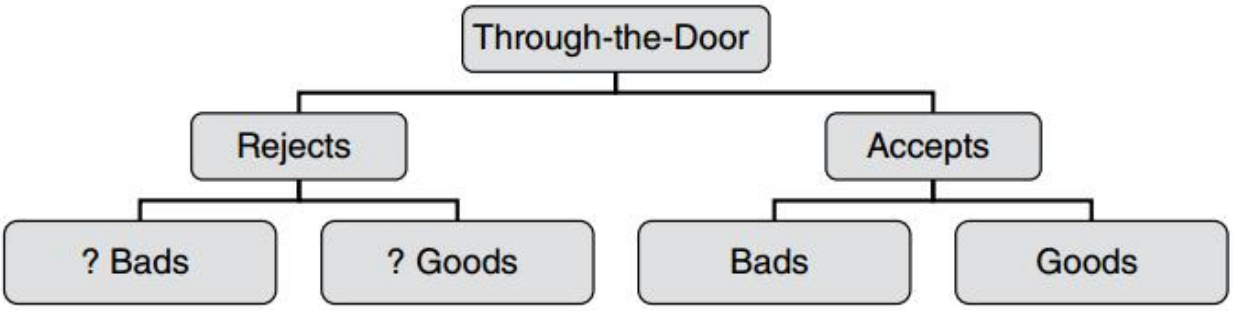

Figure 1.5 The Reject Inference Problem in Credit Scoring [3]

1.7.3 Types of data elements [3]

It is important to appropriately consider the different types of data elements at the start of the analysis. The following types of data elements can be considered:

- *Continuous*: These are data elements that are defined on an interval that can be limited or unlimited. Examples include income, sales, RFM (recency, frequency, monetary).
- *Categorical*
  - *Nominal*: These are data elements that can only take on a limited set of values with no meaningful ordering in between.
    Examples include marital status, profession, purpose of loan.
  - *Ordinal*: These are data elements that can only take on a limited set of values with a meaningful ordering in between. Examples include credit rating; age coded as young, middle aged, and old.

- *Binary*: These are data elements that can only take on two values. Examples include gender, employment status.

Appropriately distinguishing between these different data elements is of key importance to start the analysis when importing the data into an analytics tool. For example, if marital status were to be incorrectly specified as a continuous data element, then the software would calculate its mean, standard deviation, and so on, which is obviously meaningless.

1.7.4 Visual data exploration and exploratory statistical analysis [3]

Visual data exploration is a very important part of getting to know your data in an "informal" way. It allows you to get some initial insights into the data, which can then be usefully adopted throughout the modeling. Different plots/graphs can be useful here. A first popular example is pie charts. A pie chart represents a variable's distribution as a pie, whereby each section represents the portion of the total percent taken by each value of the variable. Figure 1.6 represents a pie chart for a housing variable for which one's status can be own, rent, or for free (e.g., live with parents). By doing a separate pie chart analysis for the goods and bads, respectively, one can see that more goods own their residential property than bads, which can be a very useful starting insight. Bar charts represent the frequency of each of the values (either absolute or relative) as bars. Other handy visual tools are histograms and scatter plots. A histogram provides an easy way to visualize the central tendency and to determine the variability or spread of the data. It also allows you to contrast the observed data with standard known distributions (e.g., normal distribution). Scatter plots allow you to visualize one variable against another to see whether there are any correlation patterns in the data. Also, OLAP-based multidimensional data analysis can be usefully adopted to explore patterns in the data.

A next step after visual analysis could be inspecting some basic statistical measurements, such as averages, standard deviations, minimum, maximum, percentiles, and confidence intervals. One could calculate these measures separately for each of the target classes (e.g., good versus bad customer) to see whether there are any interesting patterns present (e.g., whether bad payers usually have a lower average age than good payers).

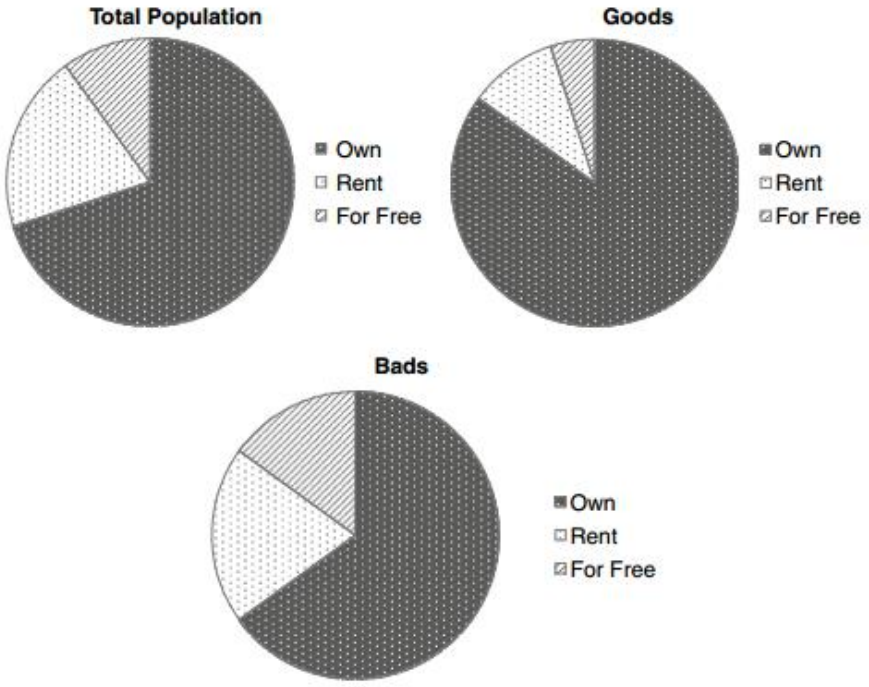

Figure 1.6 Pie Charts for Exploratory Data Analysis [3]

1.7.5 Missing values [3]

Missing values can occur because of various reasons. The information can be nonapplicable. For example, when modeling time of churn, this information is only available for the churners and not for the nonchurners because it is not applicable there. The information can also be undisclosed. For example, a customer decided not to disclose his or her income because of privacy. Missing data can also originate because of an error during merging (e.g., typos in name or ID).

Some analytical techniques (e.g., decision trees) can directly deal with missing values. Other techniques need some additional preprocessing. The following are the most popular schemes to deal with missing values:[7]

- *Replace (impute)*. This implies replacing the missing value with a known value (e.g., consider the example in Table 1.1). One could impute the missing credit bureau scores with the average or median of the known values. For marital status, the mode can then be used. One could also apply regression-based imputation whereby a regression model is estimated to model a target variable (e.g., credit bureau score) based on the other information available (e.g., age, income). The latter is more sophisticated, although the added value from an empirical viewpoint (e.g., in terms of model performance) is questionable.
- *Delete*. This is the most straightforward option and consists of deleting observations or variables with lots of missing values. This, of course, assumes

that information is missing at random and has no meaningful interpretation and/or relationship to the target.

- *Keep*. Missing values can be meaningful (e.g., a customer did not disclose his or her income because he or she is currently unemployed). Obviously, this is clearly related to the target (e.g., good/bad risk or churn) and needs to be considered as a separate category.

As a practical way of working, one can first start with statistically testing whether missing information is related to the target variable (using, for example, a chi-squared test, discussed later). If yes, then we can adopt the keep strategy and make a special category for it. If not, one can, depending on the number of observations available, decide to either delete or impute.

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1,800	?	620	Churner
2	28	1,200	Single	?	Nonchurner
3	22	1,000	Single	?	Nonchurner
4	60	2,200	Widowed	700	Churner
5	58	2,000	Married	?	Nonchurner
6	44	?	?	?	Nonchurner
7	22	1,200	Single	?	Nonchurner
8	26	1,500	Married	350	Nonchurner
9	34	?	Single	?	Churner
10	50	2,100	Divorced	?	Nonchurner

Table 1.1 Dealing with Missing Values [3]

1.7.6 Outlier detection and treatment [3]

Outliers are extreme observations that are very dissimilar to the rest of the population. Actually, two types of outliers can be considered:

1. Valid observations (e.g., salary of boss is $1 million)
2. Invalid observations (e.g., age is 300 years)

Both are univariate outliers in the sense that they are outlying on one dimension. However, outliers can be hidden in unidimensional views of the data. Multivariate outliers are observations that are outlying in multiple dimensions. Figure 1.7 gives an example of two outlying observations considering both the dimensions of income and age. Two important steps in dealing with outliers are detection and treatment. A first obvious check for outliers is to calculate the minimum and maximum values for each of the data

elements. Various graphical tools can be used to detect outliers. Histograms are a first example.

Figure 1.8 presents an example of a distribution for age whereby the circled areas clearly represent outliers.

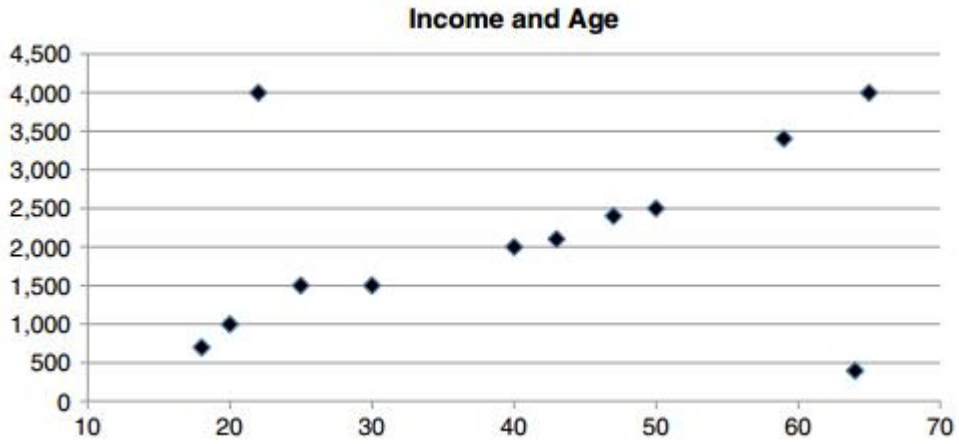

Figure 1.7 Multivariate Outliers [3]

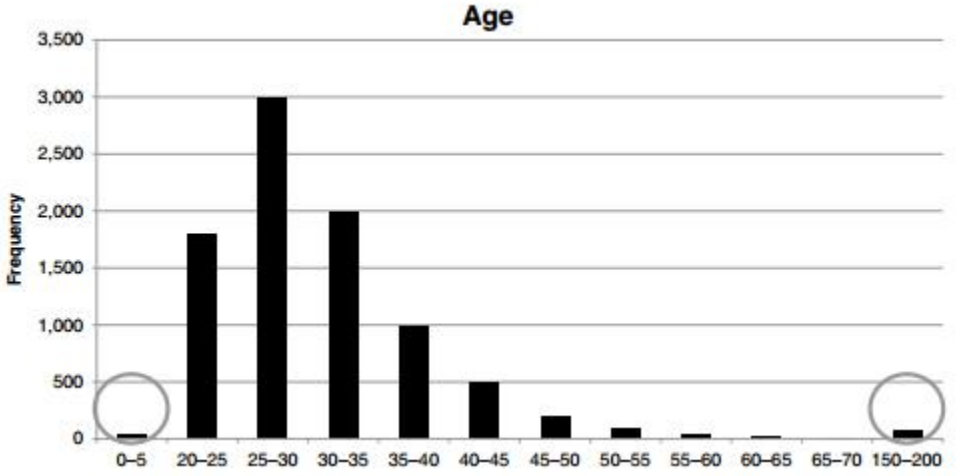

Figure 1.8 Histograms for Outlier Detection [3]

Another useful visual mechanism are box plots. A box plot represents three key quartiles of the data: the first quartile (25 percent of the observations have a lower value), the median (50 percent of the observations have a lower value), and the third quartile (75 percent of the observations have a lower value). All three quartiles are represented as a box. The minimum and maximum values are then also added unless they are too far away from the edges of the box. Too far away is then quantified as more than 1.5 * Interquartile Range (IQR = Q3 − Q1). Figure 1.9 gives an example of a box plot in which three outliers can be seen.

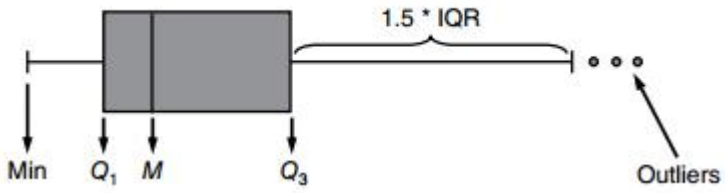

Figure 1.9 Box Plots for Outlier Detection [3]

Another way is to calculate z-scores, measuring how many standard deviations an observation lies away from the mean, as follows:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where μ represents the average of the variable and σ its standard deviation. An example is given in Table 1.2. Note that by definition, the z-scores will have 0 mean and unit standard deviation.

A practical rule of thumb then defines outliers when the absolute value of the z-score |z| is bigger than 3. Note that the z-score relies on the normal distribution.

ID	Age	Z-Score
1	30	(30 – 40)/10 = –1
2	50	(50 – 40)/10 = +1
3	10	(10 – 40)/10 = –3
4	40	(40 – 40)/10 = 0
5	60	(60 – 40)/10 = +2
6	80	(80 – 40)/10 = +4
…	…	…
	μ = 40 σ = 10	μ = 0 σ = 1

Table 1.2 Z-Scores for Outlier Detection [3]

The above methods all focus on univariate outliers. Multivariate outliers can be detected by fitting regression lines and inspecting the observations with large errors (using, for example, a residual plot).

Alternative methods are clustering or calculating the Mahalanobis distance. Note, however, that although potentially useful, multivariate outlier detection is typically not considered in many modeling exercises due to the typical marginal impact on model performance.

Some analytical techniques (e.g., decision trees, neural networks, Support Vector Machines (SVMs)) are fairly robust with respect to outliers. Others (e.g., linear/logistic regression) are more sensitive to them. Various schemes exist to deal with outliers. It highly depends on whether the outlier represents a valid or invalid observation. For invalid observations (e.g., age is 300 years), one could treat the outlier as a missing value using any of the schemes discussed in the previous section. For valid observations (e.g., income is $1 million), other schemes are needed. A popular scheme is truncation/capping/winsorizing. One hereby imposes both a lower and upper limit on a variable and any values below/above are brought back to these limits. The limits can be calculated using the z-scores (see Figure 1.10), or the IQR (which is more robust than the z-scores), as follows:

$$\text{Upper/lower limit} = M \pm 3s, \text{ with } M = \text{median and } s = IQR/(2 \times 0.6745).^{3}$$

A sigmoid transformation ranging between 0 and 1 can also be used for capping, as follows:

$$f(x) = \frac{1}{1+e^{-x}}$$

In addition, expert-based limits based on business knowledge and/or experience can be imposed.

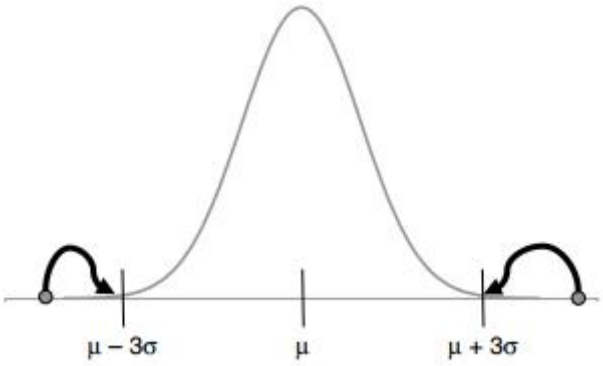

Figure 1.10 Using the Z-Scores for Truncation [3]

1.7.7 Categorization [3]

Categorization (also known as coarse classification, classing, grouping, binning, etc.) can be done for various reasons. For categorical variables, it is needed to reduce the number of categories. Consider, for example, the variable "purpose of loan" having 50 different values.

When this variable would be put into a regression model, one would need 49 dummy variables (50 − 1 because of the collinearity), which would necessitate the estimation of

49 parameters for only one variable. With categorization, one would create categories of values such that fewer parameters will have to be estimated and a more robust model is obtained.

For continuous variables, categorization may also be very beneficial. Consider, for example, the age variable and its risk as depicted in Figure 1.11. Clearly, there is a nonmonotonous relation between risk and age. If a nonlinear model (e.g., neural network, support vector machine) were to be used, then the nonlinearity can be perfectly modeled. However, if a regression model were to be used (which is typically more common because of its interpretability), then since it can only fit a line, it will miss out on the nonmonotonicity. By categorizing the variable into ranges, part of the nonmonotonicity can be taken into account in the regression. Hence, categorization of continuous variables can be useful to model nonlinear effects into linear models.

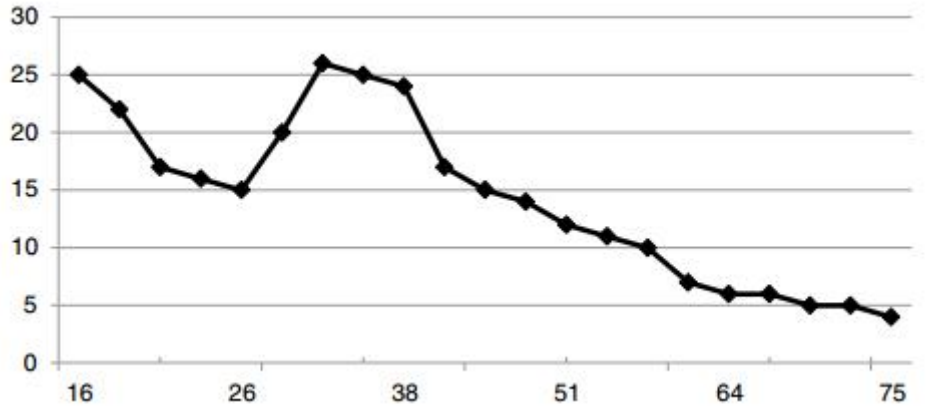

Figure 1.11 Default Risk versus Age [3]

Various methods can be used to do categorization. Two very basic methods are equal interval binning and equal frequency binning.

Consider, for example, the income values 1,000, 1,200, 1,300, 2,000, 1,800, and 1,400. Equal interval binning would create two bins with the same range—Bin 1: 1,000, 1,500 and Bin 2: 1,500, 2,000—whereas equal frequency binning would create two bins with the same number of observations—Bin 1: 1,000, 1,200, 1,300; Bin 2: 1,400, 1,800, 2,000. However, both methods are quite basic and do not take into account a target variable (e.g., churn, fraud, credit risk).

Chi-squared analysis is a more sophisticated way to do coarse classification. Consider the example depicted in Table 1.3 for coarse classifying a residential status variable.

Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6,000	1,600	350	950	90	10	9,000
Bads	300	400	140	100	50	10	1,000
Good: bad odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	9:1

Source: L. C. Thomas, D. Edelman, and J. N. Crook, *Credit Scoring and its Applications* (Society for Industrial and Applied Mathematics, Philadelphia, Penn., 2002).

Table 1.3 Coarse Classifying the Residential Status Variable

Suppose we want three categories and consider the following options:

- Option 1: owner, renters, others
- Option 2: owner, with parents, others

Both options can now be investigated using chi-squared analysis.

The purpose is to compare the empirically observed with the independence frequencies. For option 1, the empirically observed frequencies are depicted in Table 1.4.

The independence frequencies can be calculated as follows. The number of good owners, given that the odds are the same as in the whole population, is 6,300/10,000 × 9,000/10,000 × 10,000 = 5,670. One then obtains Table 1.5.

The more the numbers in both tables differ, the less independence, hence better dependence and a better coarse classification. Formally, one can calculate the chi-squared distance as follows:

$$\chi^2 = \frac{(6000-5670)^2}{5670} + \frac{(300-630)^2}{630} + \frac{(1950-2241)^2}{2241} + \frac{(540-249)^2}{249} + \frac{(1050-1089)^2}{1089} + \frac{(160-121)^2}{121} = 583$$

Attribute	Owner	Renters	Others	Total
Goods	6,000	1,950	1,050	9,000
Bads	300	540	160	1,000
Total	6,300	2,490	1,210	10,000

Table 1.4 Empirical Frequencies Option 1 for Coarse Classifying Residential Status [3]

Attribute	Owner	Renters	Others	Total
Goods	5,670	2,241	1,089	9,000
Bads	630	249	121	1,000
Total	6,300	2,490	1,210	10,000

Table 1.5 Independence Frequencies Option 1 for Coarse Classifying Residential Status [3]

Likewise, for option 2, the calculation becomes:

$$\chi^2 = \frac{(6000-5670)^2}{5670} + \frac{(300-630)^2}{630} + \frac{(950-945)^2}{945} + \frac{(100-105)^2}{105} + \frac{(2050-2385)^2}{2385} + \frac{(600-265)^2}{265} = 662$$

So, based upon the chi-squared values, option 2 is the better categorization. Note that formally, one needs to compare the value with a chi-squared distribution with k − 1 degrees of freedom with k being the number of values of the characteristic.

Many analytics software tools have built-in facilities to do categorization using chi-squared analysis. A very handy and simple approach (available in Microsoft Excel) is pivot tables. Consider the example shown in Table 1.6. One can then construct a pivot table and calculate the odds as shown in Table 1.7.

Customer ID	Age	Purpose	…	G/B
C1	44	Car		G
C2	20	Cash		G
C3	58	Travel		B
C4	26	Car		G
C5	30	Study		B
C6	32	House		G
C7	48	Cash		B
C8	60	Car		G
…	…	…		

Table 1.6 Coarse Classifying the Purpose Variable [3]

	Car	Cash	Travel	Study	House	…
Good	1,000	2,000	3,000	100	5,000	
Bad	500	100	200	80	800	
Odds	2	20	15	1.25	6.25	

Table 1.7 Pivot Table for Coarse Classifying the Purpose Variable [3]

We can then categorize the values based on similar odds. For example, category 1 (car, study), category 2 (house), and category 3(cash, travel).

1.7.8 Variable selection [3]

Many analytical modeling exercises start with tons of variables, of which typically only a few actually contribute to the prediction of the target variable. For example, the average application/behavioral scorecard in credit scoring has somewhere between 10 and 15 variables. The key question is how to find these variables. Filters are a very handy variable selection mechanism. They work by measuring univariate correlations between each variable and the target. As such, they allow for a quick screening of which variables should be retained for further analysis. Various filter measures have been suggested in the literature. One can categorize them as depicted in Table 1.8.
The Pearson correlation ρ_P is calculated as follows:

$$\rho_P = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

It measures a linear dependency between two variables and always varies between −1 and +1. To apply it as a filter, one could select all variables for which the Pearson correlation is significantly different from 0 (according to the p-value), or, for example, the ones where $|\rho_P| > 0.50$
The Fisher score can be calculated as follows:

$$\frac{\left|\bar{X}_G - \bar{X}_B\right|}{\sqrt{s_G^2 + s_B^2}},$$

where $\overline{X_G}$ ($\overline{X_B}$) represents the average value of the variable for the Goods (Bads) and s_G^2 (s_B^2) the corresponding variances. High values of the Fisher score indicate a predictive variable. To apply it as a filter, one could, for example, keep the top 10 percent. Note that the Fisher score may generalize to a well-known analysis of variance (ANOVA) in case a variable has multiple categories.
The information value (IV) filter is based on weights of evidence and is calculated as follows:

$$IV = \sum_{i=1}^{k}(Dist\ Good_i - Dist\ Bad_i) * WOE_i$$

where *k* represents the number of categories of the variable.

The following rules of thumb apply for the information value:

- < 0.02: unpredictive
- 0.02–0.1: weak predictive
- 0.1–0.3: medium predictive
- 0.3: strong predictive

Note that the information value assumes that the variable has been categorized. It can actually also be used to adjust/steer the categorization so as to optimize the IV. Many software tools will provide interactive support to do this, whereby the modeler can adjust the categories and gauge the impact on the IV. To apply it as a filter, one can calculate the information value of all (categorical) variables and only keep those for which the IV > 0.1 or, for example, the top 10%.

Another filter measure based upon chi-squared analysis is Cramer's V. Consider the contingency table depicted in Table 1.8 for marital status versus good/bad.

Similar to the example discussed in the section on categorization, the chi-squared value for independence can then be calculated as follows:

$$\chi^2 = \frac{(500-480)^2}{480} + \frac{(100-120)^2}{120} + \frac{(300-320)^2}{320} + \frac{(100-80)^2}{80} = 10.41$$

This follows a chi-squared distribution with k-1 degrees of freedom, with k being the number of classes of the characteristic. The Cramer's V measure can then be calculated as follows:

$$Cramer's\ V = \sqrt{\frac{\chi^2}{n}} = 0.10,$$

	Good	Bad	Total
Married	500	100	600
Not Married	300	100	400
Total	800	200	1,000

Table 1.8 Contingency Table for Marital Status versus Good/Bad Customer [3]

With n being the number of observations in the data set. Cramer's V is always bounded between 0 and 1 and higher values indicate better predictive power. As a rule of thumb, a cutoff of 0.1 is commonly adopted. One can then again select all variables where Cramer's V is bigger than 0.1, or consider the top 10 percent. Note that the information value and Cramer's V typically consider the same characteristics as most important.

Filters are very handy because they allow you to reduce the number of dimensions of the data set early in the analysis in a quick way.

Their main drawback is that they work univariately and typically do not consider, for example, correlation between the dimensions individually. Hence, a follow-up input selection step during the modeling phase will be necessary to further refine the characteristics. Also worth mentioning here is that other criteria may play a role in selecting variables. For example, from a regulatory compliance viewpoint, some variables may not be used in analytical models (e.g., the U.S. Equal Credit Opportunities Act states that one cannot discriminate credit based on age, gender, marital status, ethnic

origin, religion, and so on, so these variables should be left out of the analysis as soon as possible).
Note that different regulations may apply in different geographical regions and hence should be checked. Also, operational issues could be considered (e.g., trend variables could be very predictive but may require too much time to be computed in a real-time online scoring environment).

1.7.9 Segmentation [3]

Sometimes the data is segmented before the analytical modeling starts. A first reason for this could be strategic (e.g., banks might want to adopt special strategies to specific segments of customers).
It could also be motivated from an operational viewpoint (e.g., new customers must have separate models because the characteristics in the standard model do not make sense operationally for them).
Segmentation could also be needed to take into account significant variable interactions (e.g., if one variable strongly interacts with a number of others, it might be sensible to segment according to this variable).
The segmentation can be conducted using the experience and knowledge from a business expert, or it could be based on statistical analysis using, for example, decision trees, k-means, or self-organizing maps.
Segmentation is a very useful preprocessing activity because one can now estimate different analytical models each tailored to a specific segment. However, one needs to be careful with it because by segmenting, the number of analytical models to estimate will increase, which will obviously also increase the production, monitoring, and maintenance costs.

1.8. TYPES OF DATA [8]

Big Data, data science and business analytics work with structured and unstructured data. But SMART business occurs when we combine existing data sets with unstructured or semi-structured data from both internal and external sources.

Structured data
Structured data provides most of our current business insights but is often considered ‘old hat’ and a bit dull – especially in comparison to its rock star cousin, unstructured data – it is easy to ignore structured data. But that is a mistake as many Big Data insights are generated by combining structured and unstructured data.

Data that is located in a fixed field within a defined record or file is called structured data. This includes data contained in relational databases and spreadsheets.

Examples of structured data include:

- Point of sales data
- Financial data
- Customer data.

As the name would suggest structured data refers to data or information that has a predefined data model or is organized in a predetermined way.

A data model is a model of the types of business data that your business will record and how that data will be stored, processed and accessed. Within that data model the fields of data that you intend to capture need to be defined and any conventions set around how that data will be stored. For example, if you look at a standard customer database the fields that are defined will include name, address, contact telephone numbers, email address, etc. Within those fields conventions may also be set so, for example, the telephone number field will only accept numeric information. These conventions can also include drop down menus that limit the choices of the data that can be entered into a field, thus ensuring consistency of input. For example, a 'Title' field within a name structure may only give you certain options to choose from, such as Mr, Ms, Miss, Mrs, Dr, etc.
Structured data gives names to each field in a database and defines the relationships between the fields. As a result structured data is easy to input, easy to store and easy to analyze. Up until relatively recently technology just didn't have the grunt to store, never mind analyze, anything other than structured data.
Everything that didn't fit into the databases or spreadsheets was usually either discarded or stored on paper or microfiche in filing cabinets or storage facilities.
Structured data is often managed using Structured Query Language (SQL) – a programming language originally created by IBM in the 1970s for managing and querying data in relational database management systems. SQL represented a huge leap forward over paper-based data storage and analysis, but not everything in business will fit neatly into a predefined field.

Unstructured and semi-structured data

Unstructured and semi-structured data are like the popular kids at school! Everyone is talking about them and they represent the sexy new frontier lauded by Big Data. It is estimated that 80% of business-relevant information originates in unstructured or semistructured data.
It represents all the data that can't be so easily slotted into columns, rows and fields. It is usually text heavy, but may also contain data such as dates, numbers and facts or

different types of data such as images. These inconsistencies make it difficult to analyze using traditional computer programs.

Examples of unstructured and semi-structured data include:

- Photos and graphic images
- Videos
- Websites
- Text files or documents such as email, PDF, blogs, social media posts, etc.
- PowerPoint presentations.

Semi-structured data is a cross between unstructured and structured. This is data that may have some structure that can be used for analysis but lacks the strict data model structure. In semistructured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure. For example, a Facebook post can be categorized by author, data, length and even sentiment but the content is generally unstructured. Another example is word processing software that includes metadata detailing the author's name, when it was created and amended but the content of the document is still unstructured.

Internal data

Internal data accounts for everything your business currently has or could access.
This includes private or proprietary data that is collected and owned by the business where you control access.

Examples of internal data include:

- Customer feedback
- Sales data
- Employee or customer survey data
- CCTV video data
- Transactional data
- Customer record data
- Stock control data
- HR data.

External data

External data is the infinite array of information that exists outside your business.
External data is either public or private. Public data is data that anyone can obtain – either by collecting it for free, paying a third party for it or getting a third party to collect it for you. Private data is usually something you would need to source and pay for from another business or third party data supplier.

Examples of external data include:

- Weather data
- Government data such as census data
- Twitter data
- Social media profile data
- Google Trends or Google Maps.

A lot of the Big Data hype focuses on unstructured data and the allure and promise of external data, often at the expense or dismissal of internal or structured data.
It is really important to understand that no type of data is inherently better or more valuable than any other type. The key is to start with your strategy and establish your SMART questions so that those questions guide you to the best structured, unstructured, internal or external data to answer those questions and deliver the strategy.
But before we explore how to do that let's take a moment to appreciate the new forms of data that are now at your disposal as you seek to answer those questions.

1.8.1 Datification: The new forms of data [8]

Most human and computer based activities already leave a digital trace (or data) that can be collected and analysed to provide insights on everything from health to crime to business performance. Of the few activities that don't currently leave a digital trace – they soon will.
The world is being 'datafied' and there are now many forms of useful data. Some of the data forms are new such as social media posts; others have been around for a long time. For example, we've been able to record conversations for a long time but a lack of storage capacity or a way to really analyse those recordings limited their utility. But all that is changing.

Data is now being mined from:

- Our activities
- Our conversations
- Photos and video
- Sensors
- The Internet of Things.

Activity data
More and more of the activities we engage in leave a data trail.

For example, when we go online our browser logs what we are searching for and what websites we visit. Most websites will log how many people visit the site, where those people are located (using the computer ISP), how long the person stayed on the site and how they navigated or clicked through the site. Often this information is used to assess website performance and delete areas that no-one visits while improving pages that seem to generate the most interest.

If we decide to go shopping online there is a record of what we share or like and of course what we buy, how much we paid for it, when we bought it, when it was delivered and often what we then thought of the product or service through user feedback.

If we decide to read a book chances are we will increasingly turn to a Kindle, iPad, smart phone or other e-reader. There are now millions of books available in a digital format. Some books such as technology text books which change rapidly are often never even released as a physical printed book.

It has been estimated that 130 million unique books have been published since the invention of the Gutenberg printing press in 1450.

By 2012, just seven years into the Google Book Project, Google had scanned over 20 million titles or more than 15% of the world's entire written heritage![9] Amazon also gives us extensive access to old books in digital form.

When we use an e-reader we are usually not just reading a digital image of the page – the text is datafied. That means that we can change font size, add notes, highlight text or search the book.

This datafication also means that data is gathered about what we read, how long we read for, whether we skip pages, what pages we annotate and what we choose to highlight. This information could certainly prove useful for authors and publishers. I would love to know how people use my books, which sections people skip, when readers stop reading a book. This would allow me – or indeed any author – to revise content in order to shorten or improve particular parts so that readers have a better experience. Furthermore, authors and publishers may be able to identify areas of interest from frequently highlighted passages across many books to identify new topic trends on which to commission new work.

If we listen to music using our smart phone or digital music player, data is also collected on what we are listening to, how long we are listening and what tracks we are skipping past. And artists like Lady Gaga are using this data to create playlists for live gigs and influence future song creation.

Even walking to work or going to the gym will generate data if we are wearing a smart device like the 'Up' band or are using an app on our smart phone. These apps and devices can measure how many steps we take each day, how many calories we burn, how well we sleep, log activity and exercise, deliver insights and celebrate milestones. Some devices also measure our heart rate and often our heart rate variation (HRV). HRV measures the tiny variations in the interval between each heart beat and has been proven to be a

significant metric for predicting health problems. For example, since 1965 it has been common obstetric practice to monitor a baby's HRV during labour for early signs of foetal distress.[10]

In 1997 Jacqueline Dekker, Professor of Diabetes Epidemiology at the VU University Medical Centre in Amsterdam, along with her colleagues discovered that HRV was capable of predicting death, not only in babies or heart attack victims but it also predicted 'all cause mortality'.[11] Clearly, data on HRV would be useful for us all to know and devices like smart watches will be able to collect such data.

Many of these wearable devices are now Internet-enabled so that they self-generate and share data. It is also almost inevitable that many of the current wearable devices and apps will be swallowed up by the smart watch in the same way iPods were swallowed up by iPhones.

The company that makes the 'Up' band, Jawbone, now collects sleep data from millions of people around the world. This means they have unparalleled access to years' worth of sleep data – every night! No company on the planet has ever had that sort of data or that sort of volume of data.

Jawbone is then able to analyse the data to understand more about sleep, our sleeping patterns and what disrupts those patterns. For example, Jawbone could look at the data and work out how many hours of sleep are lost, on average, when the Superbowl is broadcast in the US or how long it normally takes for travelers to get back to normal sleeping patterns if they fly between New York and San Francisco or between London and Sydney.

Conversation data

Increasingly we also leave digital records of our conversations – either through text when we write an SMS message, on social media or an audio recording of a telephone call.

Just think of the billions of emails that are sent and stored every week. In fact, twenty million emails were written in the time it took to read this sentence.[12]

We are using social media to communicate and interact with each other, which is creating unfathomable amounts of data. Check out these stats:

- More than a billion tweets are sent every 48 hours.
- One million accounts are added to Twitter every day.
- Every sixty seconds, 293,000 status updates are posted on Facebook.
- Two new members join LinkedIn every second (172,800 per day)
- 72% of online adults use social networking sites.
- 25 percent of Facebook users never bother with any kind of privacy control.
- The average Facebook user creates 90 pieces of content including links, news stories, photo albums, notes, and videos each month.
- Incredibly, people in New York City received tweets about the August 2011 earthquake in Mineral, Virginia 30 seconds before they felt it.[12]

There are also already millions of website and blogs contributing to the conversation. An estimated 571 new websites are created every minute of the day. Every minute, Tumblr owners publish approximately 27,778 new blog posts and 3 million new blogs come online every month.[12]

Plus there is the data collected from our telephone conversations. If you call a customer service department we are always told the conversation may be recorded. Often that data is being mined for content and sentiment and even analysed for stress levels in someone's voice to gauge how irritated the customers are!

Audio data is also being used to improve voice recognition and translation software. For example, Google decided to venture into translation in 2006 as part of its mission to 'organize the world's information and make it universally accessible and useful'. Most translation software utilize perfectly translated pages of text to create the algorithms but Google used the entire global Internet and more. Their system sucked in every translation – good and bad – that it could find in order to train the translation computers. As a result of the sheer volume of data that they could access and use Google translation is more accurate than any other system. By mid 2012 its dataset covered more than 60 languages and even accepts voice input in 14 languages for fluid translation.[9] It's still not perfect but as the system learns from the correct translation and the incorrect translation chances are it will be in the future.

Photo and video image data

Again the data being collected and stored is staggering. Digital cameras and smart phones are taking and sharing more photos and videos than ever before. Check out these stats:

- Each day 350 million photos are uploaded to Facebook, which equates to 4,000 photos per second.
- Flickr users upload 3.5 million photos to the site each day.
- Approximately 100 hours of video is uploaded to YouTube every minute.
- More than 45 million pictures are uploaded to Instagram every day.
- As of June 2013, Instagram users have shared more than 16 billion photos.[12]

Granted, sharing what we had for dinner or a picture of our new Labrador puppy won't change the world but this plethora of photo, video (and text data) is actually already saving lives in disaster areas.

When typhoon Hiayan hit the Philippines in 2013, for example, over 6,000 people were killed and 1.1 million homes were damaged or destroyed in hours. In the UK, a team of volunteers were creating a vital map of the damaged areas using just social media.

Because it is now very common for people to share their experiences as they happen in almost real time, photos, tweets (#Hiayan) and videos about the disaster were being posted on social media. In the aftermath of Hiayan the volunteers were receiving on average a million photos, messages, tweets, videos, etc., every day!
After filtering the millions of messages using artificial intelligence to pick out the ones that could be important the team of volunteers then made an assessment of what they saw. For example, for a photograph they would be asked, 'How much damage do you see?' and they simply needed to click the appropriate button: 'none', 'mild', or 'severe'. For text based messages such as tweets or Facebook updates the volunteer was asked to decide if the text was 'not relevant', 'request for help', 'infrastructure damage', 'population displacement', 'relevant but other', etc. Each piece of data (picture, video or message) was then assessed by between three to five different people to make sure the assessment was consistent and therefore probably accurate.
By pinpointing where the data was coming from in the Philippines (using GPS sensors in the photos or through the text) the work of the volunteers then created an online map, not just of the disaster zone but of the needs in each area.
That meant that when the disaster relief effort arrived in the Philippines they didn't need to waste days working out what was happening and where the worst hit areas were. They already knew from the map – created by people half way around the world – who needed water, who needed food, where the dead bodies were and where people had been displaced, where the most damage was and what hospitals were least damaged, and therefore more able to help the injured.[13]

How cool is that?

In addition to all the photo and video data created by individuals via their digital tech or smart phone there is also all the CCTV camera footage. In days gone by companies may video record their premises or retail store and store the recording for a week or so before recording over older recordings. Now some of the larger data savvy stores are keeping all the CCTV camera footage and analysing it to study how people walk through the shops, where they stop, what they look at and for how long so they can make alterations to offers and boost sales. Some are even using face recognition software so it probably won't be long before a combination of data sources such as CCTV camera footage, loyalty card information and face recognition software will see us being welcomed to a store on our smart phones and directed to particular special offers or promotions that may be of interest to us based on our previous buying habits!

Sensor data

There is also an increasing amount of data being generated and transmitted from sensors. There are sensors everywhere.

Have you ever wondered that makes your smart phone (or smart anything for that matter) smart? Basically what makes them smart is the inclusion of various sensors that capture data. In your smart phone for example there is a:

- GPS sensor
- Accelerometer sensor
- Gyroscope
- Proximity sensor
- Ambient sensor, and
- Near Field Communications (NFC) sensor.

The GPS (Global Positioning System) sensor lets us (and others) know where we are using the GPS satellite navigation system.
The GPS sensors in our phone can pinpoint our location within a few meters (assuming we are with our phone of course!). The accelerometer sensor is a motion sensor and measures the acceleration or how quickly the phone is moving. It's this technology that allows you to take better photos with your smart phone because it's this sensor that triggers the shutter when it detects the camera is stationary or stable. The gyroscope sensor is used to maintain orientation and is used to rotate the screen. It is this sensor that is often utilized in gaming apps where you have to tilt the screen to direct the character or steer the car. As the name would suggest the proximity sensor senses proximity and how close we are to other objects or locations. Ambient sensors are the ones that detect changes in the ambience or atmosphere so it is this sensor that adjusts the backlight on your phone or saves power when it's not being actively used. And finally the NFC sensor is one of the latest communication protocols being utilized in smart phones. It is these NFC sensors that when enabled, allow you to transfer funds just by bumping phones or waving your phone close to an appropriate payment machine.
There are also sensors in the natural environment, for example, in the oceans for measuring the health, temperature and changes of the oceans in real time. Also in Japan there are sensors in the soil to collect data on how healthy the soil is and companies are combining that data with weather data. Farmers can then subscribe to the service to get information to optimize yield, including how much and when to put fertilizer on their crops.
Increasingly more and more machines are equipped with sensors to monitor performance and provide information on when best to service or repair the machines.
For example, Rolls Royce manufactures nearly half the world's passenger jet engines including the Trent 1000, the engine that powers many of our transatlantic flights. When in operation these engines reach incredibly high temperatures – half the temperature of the surface of the sun and 200 degrees above that temperature when the metal used to

make the engine melts! The only reason it doesn't melt is because the engines are being cooled through special passageways and channels that keep the heat away from the metal.
Needless to say it's vital to know that everything is working and doing its job, as you wouldn't want the plane you are taking to visit your friends in New York to melt at 30,000 feet!
The engine is therefore full of vital components all engineered with absolute precision including an on-board computer that is the brains of the engine, controlling it and also collecting and monitoring data from sensors buried deep within the engine measuring 40 parameters 40 times per second including temperatures, pressures and turbine speeds.
All the measurements are stored in the computer and streamed via satellite back to Rolls Royce HQ in Derby, England. And that's true for the entire fleet of Rolls Royce engines, which is a lot of data when you consider that a Rolls Royce powered engine takes off or lands somewhere in the world every two and a half seconds.
Whenever those thousands of engines are in the air they are gathering data which is continuously sent back to HQ and constantly monitored using clever data analytics that are looking for anything unusual going on in the engine, or any sign that it may need to be serviced early or repaired. In Derby, computers then sift through the data to look for anomalies. If any are found they are immediately flagged and a human being will check the results and if necessary telephone the airline and work out what needs to be done – normally before the issue escalates into an actual problem.
These sensors therefore allow for dynamic maintenance based on actual engine-by-engine performance rather than some automatic rota system based on time alone. Instead of pulling an expensive piece of equipment out of service every three or six months these sensors allow the airlines to maintain their fleet much more cost effectively and, more importantly, these sensors make the planes much safer.[13]

Modern cars are also full of similar sensors that measure everything from fuel consumption to engine performance, which again allows for dynamic servicing and better long term performance.
On-board sensors also alert the driver if they get too close to another car or object and can even parallel park the car without the driver having to do anything!
In the retail industry, data has long been collected via barcode; however, the sensors known as Radio Frequency Identification (RFID) systems increasingly used by retailers and others are generating 100 to 1,000 times more data than the conventional barcode system.[14]
There are sensors everywhere.

The Internet of Things

The Internet of Things (IoT) is a result of more objects being manufactured with embedded sensors and the ability of those objects to communicate with each other.

IDC describes the IoT as:

> *'a network connecting – either wired or wireless – devices (things) that are characterized by automatic provisioning, management, and monitoring. It is innately analytical and integrated, and includes not just intelligent systems and devices, but connectivity enablement, platforms for device, network and application enablement, analytics and social business, and applications and vertical industry solutions. It is more than traditional machine-to-machine communication. Indeed, it is more than the traditional Information and Communications Technology (ICT) industry itself.'*[15]

This concept explores what is and will be possible as a result of advances in smart, sensor-based technology and massive advances in connectively between devices, systems and services that go way beyond business as usual. For example, research groups such as Gartner and ABI Research estimate that by 2020 there will be between 26 and 30 billion devices wirelessly connected to the IoT.

And the resulting information networks promise to create new business models and improve business processes and performance, while also reducing cost and potentially risk.

The day will come, not far from now when your alarm will be synced to your email account and if an early meeting is cancelled your alarm will automatically reset to a later time, which will also postpone the coffee machine to the new wake-up time. Your fridge will know what's in it and place online orders to replenish stocks without you having to do anything. You'll put on your suit, with a payment chip in the sleeve so you can swipe payment for lunch without a credit card.

Your wearable device or smart watch will monitor your health through the day, watching your calorie intake and making sure you stay active and don't sit too long at your desk. As you get in your car to drive home at night the car will automatically check the route with traffic and weather information to get you home as quickly and safely as possible. On arriving home, the temperature will be perfect and your fridge will tell you what you can make for dinner based on what you currently have in stock.

As you settle down to watch TV with your family, you may be enjoying a film rated 18 when your 5-year-old child walks in and your smart TV will suspend the film and change channel. Oh and if your elderly mother is ever house sitting while you are away your smart carpet will measure and monitor her movements and patterns – perhaps she goes to the kitchen at 10.30 a.m. every morning to make a cup of coffee or always goes to bed at 11 p.m. Should those patterns change you will be alerted to get in touch and check everything is OK.

The wired and wireless networks that connect the Internet of Things often use the same Internet Protocol (IP) that connects the Internet – hence the name. These vast networks create huge volumes of data that's then available for analysis. When objects use sensors to sense the environment and communicate with each other, they become tools for understanding complexity and responding to it quickly. The resulting physical information systems are now beginning to be deployed, and some of them operate without human intervention.

Pill-shaped micro-cameras already traverse the human digestive tract and send back thousands of images to pinpoint sources of illness. Precision farming equipment with wireless links to data collected from remote satellites and ground sensors can take into account crop conditions and adjust the way each individual part of a field is farmed. There are even billboards in Japan that monitor passers-by, assess how they fit consumer profiles, and instantly change displayed messages based on those assessments. Advances in wireless networking technology and the greater standardization of communication protocols make it possible to collect data from these sensors almost anywhere at any time. Ever-smaller silicon chips are gaining new capabilities, while costs are falling. Massive increases in storage and computing power, some of it available via cloud computing, make number crunching possible on a very large scale and at declining cost.[16]

All coming together to create Big Data.

1.8.2 The anatomy of Big Data [8]

When we consider the types and forms of data that now exists it's easy to see how people become overwhelmed and bamboozled by the possibilities of Big Data. Although, as I've said I think the term will disappear and what we consider Big Data today will just be 'data' tomorrow.

For a start, what is uncommon and exciting now will become commonplace. Plus the term may be simple and easy to remember but it's overly simplistic and places far too much emphasis on the volume of data. But volume is just one of the four V's of Big Data:

- ***Volume*** – relating to the vast amounts of data generated every second.
- ***Velocity*** – relating to the speed at which new data is generated and moves around the world. For example, credit card fraud detection tracks millions of transactions for unusual patterns in almost real time.
- ***Variety*** – relating to the increasingly different types of data that is being generated from financial data to social media feeds; from photos to sensor data; from video footage to voice recordings.
- ***Veracity*** – relating to the messiness of the data being generated – just think of Twitter posts with hash tags, abbreviations, typos, text language and colloquial speech.

Big Data backlash

As with any new frontier, the frontier of Big Data is also under attack. There are those that believe that it's a storm in a teacup and the theory of Big Data is so far removed from the reality for most businesses that it will never yield much, if any fruit for the vast majority of business.

Certainly there are some companies that already have these huge data sets; however, most businesses will never have access to the volume and variety of data that an Amazon, eBay or Facebook will have. But as I've said before that's OK because most businesses don't need access to oceans of data.

The other area of attack is around consumer data and privacy.

The reputation of Big Data has suffered with the revelations by whistleblower Edward Snowden that the US National Security Agency (NSA) has been systematically using Big Data analytics to 'spy' on everyone's communications as well as perform targeted surveillance of individuals and companies. We can all be certain that the US is not the only government agency in the world to collect and use Big Data. For example, former French foreign minister, Bernard Kouchner, stated, 'Let's be honest, we eavesdrop too.

Everyone is listening to everyone else. But we don't have the same means as the United States, which makes us jealous.'

Despite high profile Snowden-type media stories, most people are completely unaware of just how much data about them is freely available online. Even if someone takes the time to complete privacy settings on social media and is deliberately vague and cautious about over-sharing – there is still a phenomenal amount of information being collected, stored and analysed. Most of us are, for example, almost entirely oblivious to the fact that the GPS sensor in their smart phone makes it possible to identify where a picture was taken within a few meters, regardless of whether the person sharing the photo adds a tag, message or caption. They don't realize how open and freely available their social media sites are, how much of what they post is saved and analysed – even when the platform tells its users that the photo or video will self-destruct in 10 seconds! Those images may not be accessible to the user after a set time but they are saved. They have no idea that their web browser is monitoring their every move or even that people can easily hack into the camera on their laptop and watch them!

In 2013 a 19-year-old US student was charged with hacking Miss Teen USA's webcam. The FBI found that he had used malicious software to remotely operate webcams to get nude photos and videos of at least seven women as they changed clothes. Some of these women he knew personally and others he found by hacking Facebook pages. In the UK in 2014 another man received a suspended sentence for the same thing. Probably best to cover your webcam when you're not using it – just in case!

So far people have not really cottoned on to the dangers or the inherent value of their own data and are happy to freely share that data in exchange for services they want, such as Facebook.

Facebook is already a gigantic data mining paradise with unbelievable amounts of data at their disposal, all enthusiastically provided by the users of Facebook. Remember the stats from earlier – 350 million photos a day, 293,000 status updates a minute and 25% of users never bother with privacy!

Facebook knows what we look like, who our friends are, what our views are, what our interests are, when our birthday is, whether we are in a relationship or not, where we are, what we like and dislike, and much more. That is an awful lot of information (and power) in the hands of one commercial company.

People may start to get uncomfortable about the amount of data that is known and held amount them. But how much of a difference would it really make? Take Facebook again: even if we all stopped using Facebook today (which is very unlikely), the company would still have more information about people than any other private company on the planet. Google may come close but they don't have the plethora of detailed personal data that Facebook has. Of course it's not just Facebook.

The challenge is that once companies have access to the data they won't stop. And we don't have to be a loyalty card member for the companies to know about us: in addition to social media, they can also track our credit card use and use face recognition software to record what we are doing in store.

A recent study showed that it is possible to accurately predict a range of highly sensitive personal attributes simply by analyzing the 'Likes' we have clicked on Facebook. The work conducted by researchers at Cambridge University and Microsoft Research shows how the patterns of Facebook 'Likes' can very accurately predict characteristics such as your sexual orientation, satisfaction with life, intelligence, emotional stability, religion, alcohol use and drug use, relationship status, age, gender, race and political views among many others.[17]

The fact is that the data collectively held on you by banks, credit card companies, insurance companies, supermarkets and social media is astonishing and it's growing all the time.

Even if people did become uncomfortable in enough numbers to bring about changes to legislation it may be too late. It would be like shutting the barn door once the horse had bolted. It may be that legislation may push for at least some of the most sensitive data to be anonymized, i.e. markers that identify an actual person to be removed, but it will still be used and the datifaction of the world will not stop.

Whether we like it or not, or are ready for it or not, the future will involve Big Data. Our ability to harness that power with intelligence, common sense and practicality will see us turn it into meaningful SMART Data.