# CHAPTER 4

## DOMAIN III - DATA

### WHAT WILL YOU LEARN IN THIS CHAPTER?

Analytics is defined by INFORMS as the scientific process of transforming data into insight for making better decisions. In this section we will see how data collection, manipulation and analysis support the analytic framework from problem identification to model building and management. Data transformation starts with data element definition and potential source identification. Once sources are identified, collection of new data and extraction and transformation of existing data can begin. Often data will need to be cleaned to address incorrect and/or missing data points. Finally, the data must be properly formatted for use in the common database and loaded into it.

### Learning Objectives

By the end of this chapter, you should be able to

1. Identify and prioritize data needs and resources
2. Identify means of data collection and acquisition
3. Determine how and why to harmonize, rescale, clean and share data
4. Identify ways of discovering relationships in the data
5. Determine the documentation and reporting of findings
6. Use data analysis results to refine business and analytics problem statements

### Key Concepts/Fundamentals

### OBJECTIVE 1. IDENTIFY & PRIORITIZE DATA NEEDS & RESOURCES

Data reduces our uncertainty about the values assigned to variables of interest in the analysis.

Analysis typically uses `hard data', i.e., data that is obtained by scientific observation and measurement (e.g., experimentation). But much of our information is frequently soft, e.g., gleaned from interviews and reflective opinions and preferences. Hence it will be important to convert this soft information into scientific data. The traditional way in which soft data is converted into hard data is to hypothesize an artificial individual whose preferences and beliefs can be completely described with hard data. (In economics, this artificial individual is called the `economic man' and is viewed as totally rational.) We then determine what hard data would be required so that this artificial individual's behavior coincides with that of the actual individual with soft data. We then solve the analytical problem as if our actual individual could be described by this artificial individual.

Probably the most successful example of this approach is conjoint measurement or analysis which posits that the behavior of the actual individual can be described by an artificial individual whose preferences are described by a utility function. The utility function for various outcomes is first specified as a parametric function of observable attributes of that item. If this utility function were known, then it is, in theory, straightforward to specify which of several items an individual would choose or how an individual would rank several items. To determine the parameters of this utility function, individuals are then asked to either specify which, of several hypothetical alternatives, they prefer or how they would rank different items in order of preferability. The parameters are then calibrated so as to minimize the disparity between what the individual actual prefers and what the model predicts the individual should prefer.

Other methods analogous to conjoint have been developed when uncertainty is involved. For example, it may be necessary to develop a `subjective probability' as a summary measure of an individual's beliefs about whether an event occurs. To assess an individual's subjective probability, consider a random mechanism (e.g., a roulette wheel in a casino or a table of random numbers). For any frequency *f* between zero and one, this random mechanism can be used to define an uncertain event A(*f*) which occurs with probability *f*. An individual's belief in whether an event E occurs can be specified by asking an individual to choose between betting on event E occurring or A(*f*) occurring. For some frequencies *f*, the individual will prefer E to E(*f*) and for others, they will prefer A(*f*) to E. The point at which the individual is indifferent is called the individual's `subjective probability' for event E.

While conjoint is focused on assessing utility functions for known outcomes, the decisions which will be informed by analysis typically are gambles which do not have guaranteed outcomes. As a result, it becomes important to extend the concept of utility to gambles with uncertain outcomes. To construct these utilities, define an experiment where M is some best possible outcome and m is a worst possible outcome. Consider a gamble which leads to M with frequency *f* and m otherwise. Again consider a carefully designed laboratory environment where the individual must decide between the consequence and the gamble. Then there will be some maximum value of *f* for which the individual still prefers the consequence to the gamble. This maximum value measures the individual's preference in the consequence.

In gathering data, it is usually important to have some measure of the confidence which is placed on each of the various data points. To translate this notion of confidence into something tangible, consider two individuals, both whose measure of belief in event E is described by the subjective probability p. Consider a carefully designed laboratory experiment in which each individual observes one success in one trial. Each individual's new belief in the event is then measured. Suppose the resulting value for both individuals is U. A parallel experiment is then run in which the individual's belief in event E is measured after the individual, instead of observing success in the one trial, observes a failure. Let L be the measure of belief which both individuals now have in the event. Now suppose that one of the individual's original assessment of p is based only on observing n trials. *(More precisely, we assume that the individual had a non-informative prior over p and then updated that based on the information in n trials.)* Then it can be shown that n=1/(U-L). Suppose that the other individual's beliefs are based on soft data. Then for analytical purposes, it still is legitimate to use 1/(U-L) as a measure of confidence in p.

These examples assumed a carefully designed laboratory experiment. Just as a physical experiment presumes that the physical environment has been prepared to eliminate contaminating influences, so these laboratory experiments must be designed to eliminate contaminating influences like ambiguity, reference point effects, etc.

## OBJECTIVE 2. IDENTIFY MEANS OF DATA COLLECTION & ACQUISITION

The focus of this stage is on identifying which kinds of data collection will have the most favorable impact on the quality of the actions and recommendations supported by the analysis. An especially useful tool for doing this analysis is the decision tree. (While the decision tree as applied to uncertainty was formalized in the mid-twentieth century, it can be argued that the Pythagorean Y might have been the first decision tree.) Consider the following very simple decision tree where there are two choices : continue the present course or make a specific change. If a change is made, the outcome of the change could be favorable or unfavorable. We can write this decision tree in outline form as

1. Continue present course
    a. Get an average outcome
    b. Implement a change
2. Get a good outcome
    a. Get a poor outcome

There are two possible outcomes of making the change. If the chance of getting a good outcome is high enough, then it will be better to implement the change. Otherwise implementing the change will be unwise. For example, suppose that we attach a probability p to getting a good outcome if we make a change. Suppose we believe that U is the value (utility) of making a change with the good outcome, L is the value (utility) of making a change if the poor outcome occurs and u is the value (utility) of continuing the present source. Then we will only make a change if

$$p\,U + (1 - p)\,L \;\; > u.$$

Suppose we find that the best decision (i.e., the one with highest utility to the customer) is to continue the present course. Then we will get utility score u.

But instead of simply making a decision, we could have chosen to gather data and then make our decision based on the results of the data gathering exercise. If we chose to gather data, then our decision tree becomes

1. Gather data and get favorable information
    a. Continue present course
        i. Get an average outcome
    b. Implement a change
        i. Get a good outcome
        ii. Get a poor outcome
2. Gather data and get unfavorable information
    a. Continue present course
        i. Get an average outcome
    b. Implement a change
        i. Get a good outcome
        ii. Get a poor outcome

Now suppose we gather data and get favorable information. This increases the probability of getting a good outcome given we implement a change. Suppose the change in probability is not enough to justify implementing the change. So our conditional decision is *if we get favorable information, we continue with the present course.* Now suppose that instead of getting favorable information, our data gathering led us to collect unfavorable information. This lowers the probability of getting a good outcome given we implement a change. As a result, our other conditional decision is *if we get unfavorable information, continue with the present course.* Thus our two conditional decisions are *if we get favorable information, we continue with the present course; if we get unfavorable information, continue with the present course*. Hence regardless of the outcome of the information, we continue our present course. This simple example demonstrates an important principle: *Before collecting the information, think about everything you might discovery from collecting the information. If none of these discoveries would lead you to change your decision, then do not collect the information.* Of course, sometimes people collect information—even though they know what decision they will make—in order to defend themselves against criticisms from others. And sometimes people collect information to postpone making the decision.

When would information be valuable? Suppose that the favorable information led to a substantial change in the probability of getting a good outcome. Suppose that this change in probability was enough to justify implementing the change. Then our two conditional decisions would be *if we get favorable information, implement a change; if we get unfavorable information, continue with the present course.* We can assign a value (or utility) to these two conditional decisions. Let u* be the utility of implementing a change, given that we get favorable information. Let u be the utility of continuing the present course, given that we can unfavorable information. Let q be the probability of our getting favorable information if we collect data. Then the utility if we decide to gather data will be

$$q\, u^* + (1 - q)\, u.$$

Since the utility if we did not gather data was u, this tells us that our overall utility has increase from u to q u* + (1 - q) u. Since u* >u, collecting the information can only improve our utility. This demonstrates a well-established principle: *the value of information is non-negative,* i.e., it can never make you worse off if you behave rationally.

But in reality, there is a cost to collecting this information. Suppose that paying this cost would reduce our utility by some factor d. Thus our utility if we collect information is

$$d\, (q\, u^* + (1 - q\,)u)\ ,$$

while it is u if we do not collect information. So if we knew q and d, the decision on whether to buy information would depend on (u* - u)/u.

What determines u*? Before making a decision, the chance of getting a good outcome after making a decision was p. Suppose that if we get favorable information, this probability changes to p* while if we get unfavorable information, it changes to p**. Then if q is the chance of getting favorable information, the rules of probability require that p=qp*+(1-q)p**. Thus while the utility of making the change was originally

$$p\ U + (1 - p)L,$$

the utility now changes—given a favorable outcomes—to u* where

$$u^*=p^*\ U + (1 - p^*)\ L \quad = p^*\ (U - L) + L.$$

So the critical value u* depends upon p* and, in particular, on how much p* differs from p.

The degree to which the new information can change the value of p depends upon the confidence in the original value of p as well as in the impact of the data. One key question is *if the new information tells us something unexpected (i.e. , the favorable outcome), how much will our initial beliefs change?* But given that they do change, we need to know what the potential payoff might be. In this example, H was the maximum payoff if we knew for certain that there would be a good outcome. If the potential payoff, H, were small, then gathering more information would also be pointless.

The final consideration is cost. Since analysts often collect information from the client's subject matter experts, it is important to treat the time of these subject matter experts as precious. If they feel their time is being wasted, then they will complain to the client who will eventually begin to wonder about the value of doing your analysis. There are many cases in which an organization chooses a flawed heuristic over a more sophisticated procedure just because the flawed heuristic seems to require less painful information collection.

There are also privacy issues. Invasion of privacy can lead to a loss of customer good will and, in some cases, legal repercussions. And if we are gathering information that is potentially proprietary intellectual property issues become paramount. The fact that information technology has made it easier to collect information does not mean that information collection is costless.

Once you identify the variables on which you should collect data, the next step is collecting that data. Data collection is analogous to asking certain subjects certain close-ended questions under certain circumstances. Hence there are five steps involved in data collection:

1. Determining how to identify subjects (the sample design)
2. Determining how many subjects to identify (the sampling plan)

3. Determining the questions to be asked
4. Determining the possible answers to the question (the granularity of the experiment)
5. Determining a control group

## SAMPLE DESIGN

The population of subjects that could be recruited should be identified. It is common to require random sampling, i.e., to conduct sampling so as to give each subject an equal chance of being part of the sample. This reflects the fact that convenience sampling, e.g., asking those subjects that happen to be easy to identify, has been shown to lead to significant biases. But if the event of interest is highly unlikely, it may be advantageous to bias the sampling toward sampling those individuals most likely to have experienced the event of interest. The analysis will, however, have to take into account this systematic deviation, called stratified random sampling, from conventional random sampling.

Typically each subject has different characteristics (or covariates). To determine how these covariates affect the results of the experimenter, it is tempting—but inefficient—to change one factor at a time and record the change in response from the factor as the impact of that factor. Design of experiments has been shown to be a much more efficient way of assessing the impact of changing factors. This typically involves changing several factors simultaneously. If a full factorial design is used, it is possible to identify the impact of each factor as well as the impact of all possible two-way, three-way, etc. interactions between factors. When it is not necessary to know these higher-order interactions, the less time-consuming fractional factorial designs are used.

It is common to use response surface modeling (and especially regression) to specify the value of interest as a function of the covariates. The independent variable reflects the covariates and are commonly represented using dummy variables for categorical and interval data. When the variable is ratio scale, Box-Cox Transformations are often used to achieve normality. When the dependent variable is categorical, the regression model is typically logistic. When the dependent variable is ordinal, the regression model is typically ordered logit. When the dependent variable is ratio, standard regression is often used. If Y is the dependent variable and X1,…Xn represent the independent (or explanatory variables), then the typical regression model has the form y=E[Y] + e where e is a normally distributed error term and E[Y], the expected value of Y is some parameterized function of (X1,…Xn). In the interests of making this function linear, it is common to write

$$E[Y] = g(a1\ X1' + \ldots\ an\ Xn')$$

where g is a `link' function and X1',…Xn' are monotonic transformations of X1,… Xn. This becomes a generalized linear model if we generalized the error term to be a member of the exponential family of distributions (which includes the normal distribution, the exponential distribution and a remarkably large number of other distributions.)

Because time is often an important dimension, there are a separate body of time-series methods when observations are collected over time. Time-series analysis typically corrects for seasonal patterns (e.g., unusually high sales during holiday seasons) and provides a natural way of identifying trends.

## SAMPLING PLAN

How many individuals should be sampled? This is typically  determined by the existing amount of uncertainty in the quantity of interest, how much that uncertainty needs to be reduced to facilitate the making of a decision and the degree to which an individual's responses is contaminated with random error. A simple rule of thumb is that quadrupling the number of individuals sampled reduces the uncertainty by half. While uncertainty is commonly measured by the standard deviation, there are situations in which the standard deviation does not exist and the difference between the third and first fractile of the uncertainty distribution is more appropriate.

If we are willing to describe our uncertainty using the previously mentioned exponential family of distributions, the rule for updating uncertainties based on sample information has a very simple form. If our uncertainty is described by an exponential family distribution, it will have two parameters.  (In some cases, the parameters may be vectors.) The data is described by the number of observations and the sum of a score for each individual observation. This score for each individual observation will depend upon the exponential family distribution being assumed (If the data consists of coin flips, the score might be one for successes and zero otherwise.) Based on this observed data, the original distribution of the uncertainty is updated. The updated distribution will have the same form as the original distribution with two changes.  The first parameter is increased by the summed score while the second parameter is increased by the number of observations. In effect, the original uncertainty about the variable—which reflects soft data—can be treated as if it were generated by a hypothetical set of observations. Pooling this hypothetical data with the actual data then generates a new set of hypothetical data.

## DETERMINING THE QUESTIONS TO BE ASKED

A key issue in designing the experiment is determined the nature of the variable being assessed. Is the variable categorical (e.g., values of the variable are blue, red, white) where there is no natural ordering between the values of the variable? When we have categorical scales, the data can be summarized by the proportion of observations which assumed each of the possible values of the categorical variables (e.g., the proportion of blue responses, red responses, etc.)

One can ask YES/NO questions or multiple-choice questions for nominal scales. One extension (likert-type questions) asks subjects to indicate whether they fully agree, partially agree, are neutral, partially disagree or fully disagree with the statement.)

Alternatively the variable might be ordinal (e.g., short, medium, tall) where there is a natural ordering between the values of the variables. When we have ordinal sales, it is possible to define the normalized quantity for each response x by the fraction of responses less than or equal to x (e.g., the fraction of people who are either short or medium.)

A second approach, semantic differential, has the form: `what is your experience navigating our web-site' with answers like `very hard, somewhat hard, OKAY, somewhat easy, very easy' where the two ends of the scale represented opposites. In this case, the response is ordinal.

In both Likert and semantic differential scales, the response scales may be improved by providing concrete examples of what would have to be true for a `fully agree' or a `fully disagree' response to be true.

A third approach, rank-order, asks individuals to rate various factors in order of importance.

Alternatively the variable might be interval (e.g., thirty degrees centrigrade, forty degrees centrigrade, fifty degrees centigrade) where the differences between values (e.g., forty degrees minus thirty degrees) are meaningful. Note that when we have interval scales, it is possible to define a normalized quantity for each response x by subtracting the lowest possible value from x and dividing the result by the difference between the highest and lowest value.

A fourth approach is the simple multiple-choice question.

It is important to remember that subjects will often answer a question even when they have no idea about what question is being asked or about what their answer means. (For example, individuals will generally answer the question which is more important diamonds or water even though the answer clearly depends upon whether the individual feels that the choice is between having no water at all for a week (and dying of dehydration) or simply go without an added glass of water for an hour or two.) Questions must be designed with care.

## DETERMINING A CONTROL GROUP

Measurements are typically only meaningful if there is reference to some kind of underlying standard. Thus in extensive measurement, there is some base unit of measure. The score of an item is the number of multiples this basic unit required to create an object that is comparable to the item of interest.  By for many non-physical cases, there is no meaningful unit of measurement. In these cases, one creates a benchmark group of units, some smaller than the item of interest and some larger. The score of an item is the proportion of items in a benchmark group which the item outranks. When the item is an uncertain quantity, the score of an item is the probability of the item outranking a randomly chosen item from the benchmark group. This benchmark group is commonly referred to as a control with the item's score being called its effect size.

While some data needs to be created and collected, some data already exists. The purpose of extraction is to collect all this data from the many sources in which it appears so that it can eventually be loaded into a common database. In extracting this data, it is critical to know the data source from which each data element was taken, *i.e., the data must be traceable to its source*. If the results of an analysis depend critically on the data element, then understanding the validity of this data element becomes critical. In addition, if there is some change in the clients for the analysis, it will be important to transition the database to reflect the data sources which these new clients consider important. This requirement is called traceability and typically requires careful documentation.

Another increasingly important issue in using existing data sources is privacy. It is now fairly easy to get personal information on customers by buying such information from vendors. But the customers who provided this information often had an expectation that the information was to be used for a specific purpose, e.g., for enabling them to buy a product over the internet. When these customers discover that their information is being used for another purpose, some customers feel that their privacy is being violated. On top of the privacy issue are intellectual property issues. Even though it may be easy to access the information, there may be copyright or other issues which limit your ability to use it without permission or without compensating the owner of the information.

Data cleaning, while often the least glamorous phase of analysis, is often the most necessary. This is especially the case with pre-existing databases. Because pre-existing databases were collected for other purposes, the quality of the data will be driven by what was important in the original use of this data and hence need not satisfy the quality requirements for the analysis at hand. For example, vendors often have to fill in various forms in order to get reimbursed for their services. Sometimes third parties successfully get their own questions added to these surveys. But both vendor and buyer are primarily interested in the fields which determine how much the vendor gets compensated for their services. As a result, these decision-relevant fields get scrutinized carefully and the rest do not.

There are many other reasons why survey quality may be deficient:

1. Individuals asked to fill out a lengthy survey will get fatigued and simply put in default values so that they can finish the survey. If there are five possible answers to a survey, they may also check a neutral response. Or in a survey of satisfaction, they may either indicate that they are satisfied with everything or satisfied with nothing.

2. Individuals may also be offended by questions about their age, income, marital status, ethnicity and—if the survey forces them to fill in an answer—will often deliberately fill in a false answer. This is especially true given increasing concerns about privacy

3. Biases can often arise because most people, when asked to fill out a survey, simply refuse. Those who did fill out the survey are often people with more leisure time or with more emotional commitment to the organization asking that the survey be filled out.

As a result, data cleaning includes:

1. Identifying the range of valid responses for each question and labeling the data field

2. Identifying invalid data responses (e.g., where letters are used where numbers are required)

3. Identifying inconsistent data encodings (e.g., different abbreviations might be used for state)

4. Identifying suspicious data responses (e.g., when physically questionable numbers are put in for a response) Are there outliers that don't seem to make sense?

5. Identify suspicious distribution of values (e.g., when one finds that 99% of the respondents in a survey of poor neighborhoods have incomes of more than a million dollars.) Descriptive statistics can be very helpful in identifying suspicious distributions. For example, histograms specify the frequency with which various data response are used. Box and Whisker charts as well as stem and leaf plots provide compact descriptions of the variation in the data within a field and help identify outliers. Scatterplots show how the value of one set of variables depends on another. Summary statistics like the mean, median, upper and lower fractiles can also be useful

6. Identify suspicious interrelationships between fields. We first identify whether there is any correlation between data fields—possibly using factor analysis or principle component analysis. The creator of the database may have created a new variable—by combining existing fields—which was useful for their analysis but is no longer useful for your analysis

So a key part of data cleaning is determining whether the data makes sense. It also involves handling null or missing values. There are several possible solutions:

1. **Deletion:** Dropping the observation containing the missing value

2. **Deletion when necessary:** Not using the observation in analyses requiring a valid response for the missing item. This approach means that one might have a sample of 1000 people for one kind of analysis and a sample of 950 people for a second kind of analysis.

3. **Imputing a value:** In other words, we use regression to attempt to predict what the answer to this question would have been—based on the answers the subject gave to other questions.

4. **Randomly imputing a value:** The problem with imputing a value is that it pretends that we do know the value that the subject filled it for this question. Thus understates our uncertainty about the value (and thus overstates our sample size) which can lead to biases in the analysis. Random imputation in theory reruns the analysis for all possible responses the subject might have given to the question, weighted by the regression-based probability of the subjective giving that response. Efficient algorithms have been developed for doing multiple imputation.

It is important to determine whether important observations (e.g., observations from a specific group of sub-users) is missing.

A field should be created with the data of each observation (a date stamp.) A field should also be created identifying the data source from which this information is collection. This field will be important in the next step where information from different data sources is combined into a single database

While the individual responses come from different data sources, they need to be placed into a common database (which typically is organized into rows representing observations and columns representing observed characteristics of that observation). This requires that all of the data be summarized at a common level of granularity.

For example, we might have 1000 observations of one product, 5000 observations of a product and its location and 3 observations of a product, its option content and its location. If details about a product's location are not relevant for the analysis, then we can sum up our observations so that all data is at this less granular level. In other cases, we need to go to this more granular level. If we simply dropped all the observations that did not have this information, there could be insufficient sample size to support a meaningful analysis. Alternatively we may rewrite all of our 9000 records at this more granular level with fields for the product, its location and its option content. We now must treat many of our records as if they had missing values for location and option content.

Sometimes data is aggregated in different ways. Thus some information on vehicles is stored with certain vehicles being called two-door Chevrolets. Other information is stored with certain vehicles called Chevy Cruzes. Still other information is stored as General Motors compacts. In this case, there is no a single categorization that is more granular than any other categorization. As a result, we may simply need to define a record which has enough fields to contain the information from each of these observations. Thus there might be a field indicating whether the vehicle was two-door or not, a field for the vehicle's model name (Cruz), a field indicating the vehicle's body-type (compact) and a field reflecting the vehicle's division and manufacturer.

In some cases, the model may require information on a variable which is not in the data-base but can be computed from items in the data-base. This may require the creation of a new field in the database for this derived variable.

In some cases, a single observation may reflect the responses of 10,000 people while another observation may reflect the responses of 100 people. As opposed to creating a database with 10,100 rows for these two observations, it may be useful to introduce a weighting field that identifies the number of respondents associated with the observation.

Because different datasets are typically generated with different data architectures and different programming languages, these languages may use different standards for encoding information. Thus missing values can be represented by spaces, the words NA, the words Not/Available, etc.

Some decisions may be required in how to handle textual fields. This could be handled by creating numeric columns describing the textual field and—without deleting the textual field—using the columns to classify the field.  For example, the textual field might contain verbatim user expressions of satisfaction. A column might be created which expresses the encoder's interpretation of that field as expressing satisfaction, dissatisfaction or neutrality.

Before loading the database, it is useful to assess whether certain fields have the same value across all datasets. If this is the case, then it may be worth deleting those fields.

The data is then loaded into the common database. Information is typically normalized so that any given item of information only occurs in the database exactly once. This is the place to do some final checks on the quality of the data:

1. Completeness: Are all the fields of the data complete?

2. Correctness: Is the data accurate?

3. Consistency: Is the data provided under a given field and for a given concept consistent with the definition of that field and concept?

4. Currency: Is the data obsolete?

5. Collaborative: Is the data based on one opinion or on a consensus of experts in the relative area?

6. Confidential: Is the data secure from unauthorized use by individuals other than the decision maker?

7. Clarity: Is the data legible and comprehensible?

8. Common Format: Is the data in a format easily used in the application for which it is intended?

9. Convenient: Can the data be conveniently and quickly accessed by the intended user in a time-frame that allows for it to be effectively used?

10. Cost-effective: Is the cost of collecting and using the data commensurate with its value?

The term data warehouse is generally used to describe:

1. A staging area, i.e, the operational data sets from which the information is extracted
2. Data integration which is the centralized source where the data is conveniently stored
3. Access layers, i.e., multiple OLAP (on-line analytical processing) data marts which store the data in a form which will be easy for the analysis to retrieve

The data mart is organized along a single point of view (e.g., time, product type, geography) for efficient data retrieval. It allows analysts to

1. slice data, i.e., filtering data by picking a specific subset of the data-cube and choosing a single value for one of its dimensions;
2. dice data, i.e., grouping data by picking specific values for multiple dimensions;
3. drill-down/up, i.e., allow the user to navigate from the most summarized (high-level) to the most detailed (drill-down);
4. roll-up, i.e., summarize the data along a dimension (e.g., computing totals or using some other formula);
5. pivot, i.e., interchange rows and columns (`rotate the cube’).

Fact tables are used to record measurements or metrics for specific events at a fairly granular level of detail. Transaction fact details record facts about specific events (like sales events), snapshot fact tables record facts at a given point in time (like account details at month end) and accumulating snapshot tables record aggregate facts at a given point in time. Dimension tables have a smaller number of records compared to fact tables although each record may have a very large number of attributes. Dimension table includes time dimension tables, geography dimension table, product dimension table, employee dimension table, and range dimension tables.

Each dimension is typically ranged into hierarchies, e.g., the geography dimension might be arranged in stores, cities, states and countries. These hierarchies are often dynamic, e.g., a firm may redraw its organizational boundaries. In the star schema, there is often a single fact table with many dimensional tables surrounding it.

The leads to a data mart which will service the analysts in an efficient matter. However the data warehouse and data marts are not finished until they are documented in a way that makes them usable by external parties. While it is tempting to assume that the modeler will know what the variables mean, the reality is that there will often be requests to revisit the data months or years after the analysis is done. These requests may come from the client or they may come from peer reviewers interested in replicating your work. In this case, failure to document your data fields as well as the sources of the data can be very costly.

## OBJECTIVE 4. IDENTIFY WAYS OF DISCOVERING RELATIONSHIPS IN THE DATA

The many ways of understanding data can be organized into nine steps. The following list from Booz-Allen-Hamilton' s The Field Guide to Data Science describes some of the techniques which can be useful in implementing each of these steps:

1. Filtering
    a. Filtering can involve using relational algebra projection and selection to add or remove data based on its value.
    b. Filtering usually involves outlier removal, exponential smoothing and the use of either Gaussian or median filters.
2. Filling in missing data with imputation:
    a. If other observations in the dataset can be used, then values for missing data can be generated using random sampling or Monte Carlo Markov Chain methods.
    b. To avoid using other observations, imputation can be done using the mean, regression models or statistical distributions based on existing observations.
3. Reducing the number of dimensions in the data:
    a. Principle component analysis or factor analysis can help determine whether there is correlation across different dimensions in the data.
    b. For unstructured text data, term frequency-inverse document frequency identifies the importance of a word in some document in a collection by comparing the frequency with which the word appears in the document to the frequency with which the word appears in the collection to which the document belongs.
    c. When data has a variable number of features, feature hashing is an efficient method for creating a fixed number of features which form the indices of an array.

d. Sensitivity analysis and wrapper methods are typically essential when you don't know which features of your data are important. Wrapper methods, unlike sensitivity analysis, typically involving identifying a set of features on a small sample and then testing that set on a holdout sample.

e. Finally self-organizing maps and Bayes nets are helpful in understanding the probability distribution of the data.

4. Extracting features

a. Duplicate data elements must be corrected with de-duplication methods.

b. Normalization is required to ensure your data stays within common ranges. This prevents the scales in which data was collected from obscuring the interpretation and analysis of that data.

c. Format conversion is typically required when data is in binary format.

d. Fast Fourier Transforms and Discrete wavelet transforms are used for frequency data.

e. Coordinate transformations are used for geometric data defined over Euclidean.

5. Collecting and summarizing data

a. Basic statistics (raw counts, means, medians, standard deviations, ranges) are helpful in summarizing data.

b. Box plots, scatter plots, box and whisker plots provide compact representations of how data is distributed. But when the data can be reasonably described by parametric distributions, distribution fitting are even more efficient ways of summarizing data.

c. `Baseball card' aggregation is an effective way of summarizing all the information available on an entity.

6. Adding new information to the data

a. Annotation is recommended for tracking source information and other user-defined parameters.

b. Relational algebra rename and feature addition (e.g., geography, technology, weather) can be helpful in processing certain data fields together or in using one field to compute the value of another.

7. Segmenting the data to find natural groupings

    a. Connectivity-Based methods: Hierarchical clustering generates an ordered set of clusters with variable precision.

    b. Centroid–Based methods: When the number of clusters is known, k-means is a popular technique. When the number is unknown, x-means is a useful extension of k-means that both creates clusters and searches for the optimal number of clusters. Canopy clustering is an alternate way of enhancing k-means when the number of clusters is unknown.

    c. Distribution-based methods: Gaussian mixture models, which typically used the expectation-maximization (EM) algorithm, are appropriate if you want any data element's membership in a segment to be `soft.'

    d. Density-based methods: For non-elliptical clusters, fractal and DB scan are useful.

    e. Graph-Based methods: Such methods, often based on constructing cliques and semi-cliques, are useful when you only have knowledge of how one item is connected to another.

    f. For text data, topic modeling allows for segmentation of the data

8. Determining which variables are important

    a. When the structure of the data is unknown, tree-based methods are helpful.

    b. If statistical measures of importance are needed, generalized linear models are appropriate. But if statistical measures of importance are not needed, regression with shrinkage (e.g., LASSO, elastic net) and stepwise regression may be preferable.

9. Classifying data into existing groups

    a. If you are unsure of feature importance, neutral nets and random forests are helpful. But if you require a highly transparent model, decision trees (e.g., CART, CHAID) can be preferable.

    b. When the number of data dimensions is less than twenty, k nearest neighbor methods often work. But if you have a large dataset with an unknown classification signal, naïve Bayes may be preferable.

c. Hidden Markov models are useful in estimating an unobservable state based on observable values.

### OBJECTIVE 5. DETERMINE THE DOCUMENTATION & REPORTING OF FINDINGS

Learning Objectives 5 and 6 go together. Raw data and relationships, while interesting to analysts, will not hold the attention of your business stakeholders for long. You will need to tie your findings to the analytics problem and from there to the business problem. Going back to the example in the business problem domain of the manufacturing plant, a key relationship unearthed by your data analysis might be an inverse proportionality of WIP to queue and wait time. This would need to be communicated clearly to your stakeholders along with a recommendation to reduce the rate at which material is being released to the floor to enable faster delivery. Another key relationship might be a workcenter's scrap rate being higher than its normal or "should fail" rate.

### OBJECTIVE 6. USE DATA ANALYSIS RESULTS TO REFINE BUSINESS & ANALYTICS PROBLEM STATEMENTS

Having solid data and relationships allows the first true refinement of your analytics and business problem, as you now have the ability to go beyond anecdotes of the situation and describe the situation with some level of mathematical rigor. You may find at this point that the true constraint of the system isn't what you thought it was, and that therefore the analytics problem needs to be reframed around that newly surfaced constraint. Or you may find that the business problem itself missed a key facet (interrelationships between customers and purchases, a time-series effect in the data, or anything else) that needs to be included prior to continuing. Once in a while, you actually do get the business problem and the analytics problem right the first time and you can proceed to selecting your methodology and creating your model.

### SUMMARY

It is no accident that the CAP exam weights data the most heavily of the seven domains. Without proper data gathering, cleaning, transformation, and loading, all you have are nice anecdotes. With reliable data sorted usefully, you can actually solve your problem in a meaningful way.

## FURTHER READING

Booz-Allen-Hamilton, 2013, The Field Guide to Data Science, http://www.boozallen.com/media/file/The-Field-Guide-to-Data-Science.pdf.

Hubbard DW (2010) *How to Measure Anything: Finding the Value of "Intangibles" in Business*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Hillier F, Hillier M (2010) *Introduction to Management Science: A Modeling and Case Study Approach,* 4th ed. (McGraw-Hill Higher Education, New York).

Vose D (2008) *Risk Analysis: A Quantitative Guide,* 3rd ed. (John Wiley & Sons, Chichester, UK).