1. Data driven decision making (DDD)--what does it involve?

#### **practice of basing decisions on the analysis of data, rather than purely on intuition.  (Davenport, pg 5)**

Definition: Data-driven decision making (DDDM) involves making decisions that are backed up by hard data rather than making decisions that are intuitive or based on observation alone.

(Davenport, chap 1)

Data-driven decision-making (DDD) refers to the practice of basing decisions on the analysis of data, rather than purely on intuition. For example, a marketer could select advertisements based purely on her long experience in the field and her eye for what will work. Or, she could base her selection on the analysis of data regarding how consumers react to different ads. She could also use a combination of these approaches. DDD is not an all-or-nothing practice, and different firms engage in DDD to greater or lesser degrees.

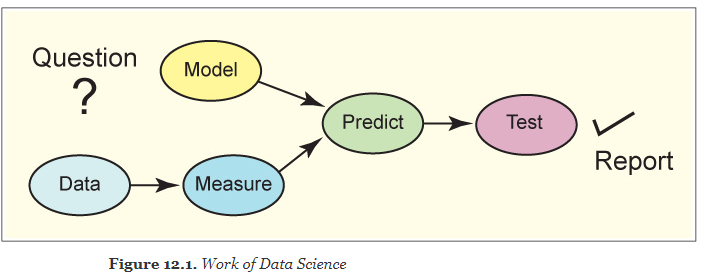
The benefits of data-driven decision-making have been demonstrated conclusively. Economist Erik Brynjolfsson and his colleagues from MIT and Penn’s Wharton School conducted a study of how DDD affects firm performance (Brynjolfsson, Hitt, & Kim, 2011). They developed a measure of DDD that rates firms as to how strongly they use data to make decisions across the company. They show that statistically, the more data-driven a firm is, the more productive it is—even controlling for a wide range of possible confounding factors. And the differences are not small. One standard deviation higher on the DDD scale is associated with a 4%–6% increase in productivity. DDD also is correlated with higher return on assets, return on equity, asset utilization, and market value, and the relationship seems to be causal.(Data Science for business, chap 1)

1. Data science capability

#### Data science capability – **the capability to extract useful knowledge from data, should be considered a strategic asset. (Davenport, pg 9)**

A **data science capability** moves an organization beyond performing pockets of **analytics** to an enterprise approach that uses analytical insights as part of the normal course of business. Building a **data science capability** in any organization isn't easy—there's a lot to learn, with roadblocks and pitfalls at every turn.

**data science capabilities** as a fundamental element in an organization's competitiveness and ability to differentiate itself.

* + Data and data science capability are (complementary) strategic assets. Under what conditions can a firm achieve competitive advantage from such an asset? First of all, the asset has to be valuable to the firm. This seems obvious, but note that the value of an asset to a firm depends on the other strategic decisions that the firm has made.(Data Science for business, chap 12)
    - 

1. Big data technologies—what are they used for?

#### Big data technologies—what are they used for? – **Examples are Hadoop, HBase and MongoDB.  They are used for data processing in support of data mining, data engineering, and occasionally for implementing data mining. (Davenport, pg 8)**

5 V's of Big Data

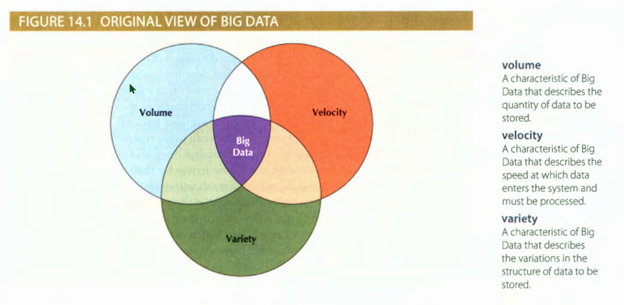
Volume, Velocity (of generation & analysis), Variety, Veracity (uncertainty of data), Value

<https://quizlet.com/392037548/big-data-technologies-flash-cards/>

<https://www.datamation.com/big-data/big-data-technologies.html>

<https://www.forbes.com/sites/gilpress/2016/03/14/top-10-hot-big-data-technologies/#60dc63bc65d7>

Big Data generally refers to a set of data that displays the characteristics of volume, velocity,and variety (the "3 Vs") to an extent that makes the data unsuitable for management by a relational database management system.



1. What is a “traditional” research model? Classical model? Bayesian? Predictive model?

##### **Predictive model? –** predictive analytics involves searching for meaningful relationships among variables and representing those relationships in models.  (Miller, pg 2)

They are **probabilistic in nature**, cannot be 100% sure the predictive is correct.

Two common predictive models:

* + - *Regression methods* help us to predict a response with meaningful magnitude, such as quantity sold, stock price, or return on investment.

*Classification methods* help us to predict a categorical response.

##### **“Traditional” research model? –** begins with a specification of a theory or model.  Classical or Bayesian methods of statistical inference are employed.  Traditional methods such as linear regression and logistic regression estimate parameters for linear predictors.  Model building involves fitting models to data and checking them with diagnostics (Miller, pg 3)

##### Classical and Bayesian statistics represent alternative approaches to inference, alternative ways of measuring uncertainty about the world. **Classical hypothesis testing** involves making null hypotheses about population parameters and then rejecting or not rejecting those hypotheses based on sample data. Typical null hypotheses (as the word null would imply) state that there is no difference between proportions or group means, or no relationship between variables. Null hypotheses may also refer to parameters in models involving many variables.(Miller, A**.** Data Science Methods, A2)

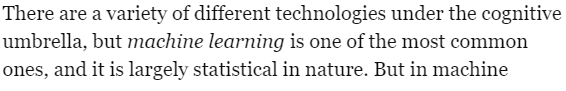
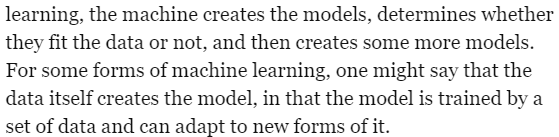
##### Bayesian**?**

While the classical approach treats parameters as fixed, unknown quantities to be estimated, the Bayesian approach treats parameters as random variables. In other words, we can think of parameters as having probability distributions representing our uncertainty about the world.

The Bayesian approach takes its name from Bayes’ theorem, a famous theorem in statistics. In addition to making assumptions about population distributions, random samples, and sampling distributions, we can make assumptions about population parameters. In taking a Bayesian approach, our job is first to express our degree of uncertainty about the world in the form of a probability distribution and then to reduce that uncertainty by collecting relevant sample data.

How do we express our uncertainty about parameters? We specify prior probability distributions for those parameters. Then we use sample data and Bayes’ theorem to derive posterior probability distributions for those same parameters. The Bayesian obtains conditional probability estimates

from posterior distributions.(Miller, A**.** Data Science Methods, A2)

1. What is model dependent research?
   * begins with specification of a model and uses that model to generate data, predictions or recommendations.  Simulations and mathematical programming methods are examples of model-dependent research
2. What is machine learning? Test set and training set? What is machine learning used for?
   * Machine learning is a collection of methods that are used to extract (predictive) models from data.  (Provost and Fawcett pg 39)
   * ****
   * ****
   * **(Davenport pg,16)**
   * machine learning, sometimes called statistical learning or data mining. Data-adaptive methods adapt to the available data, representing nonlinear relationships and interactions among variables. The data determine the model. Data-adaptive methods are data-driven. As with traditional models, we validate data-adaptive models before using them to make predictions.(Miller, chap 1)
3. What is a **data adaptive approach**?
   * When we employ a data-adaptive approach, we begin with data and search through those data to find useful predictors. We give little thought to theories or hypotheses prior to running the analysis. This is the world of machine learning, sometimes called statistical learning or data mining. Data-adaptive methods adapt to the available data, representing nonlinear relationships and interactions among variables. The data determine the model. Data-adaptive methods are data-driven. As with traditional models, we validate data-adaptive models before using them to make predictions (Miller, ch 1.1)
   * **a simulation approach?**
     + Manipulation of parameters using mathematics and/or rule bases to model how different values would generate a result. The simulated result can be used to obtain an optimal output or to predict a certain behaviour. Pharmaceutical researchers might use biosimulation to study how cells or other living entities react to chemical or other interventions (Davenport & Harris, p96)
4. Define statistical inference.
   * Statistical inference is the process of using [data analysis](https://en.wikipedia.org/wiki/Data_analysis) to deduce properties of an underlying [distribution of probability](https://en.wikipedia.org/wiki/Probability_distribution).[[1]](https://en.wikipedia.org/wiki/Statistical_inference#cite_note-Oxford-1) Inferential statistical analysis infers properties of a [population](https://en.wikipedia.org/wiki/Statistical_population), for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is [sampled](https://en.wikipedia.org/wiki/Sampling_(statistics)) from a larger population.

* + Inferential statistics can be contrasted with [descriptive statistics](https://en.wikipedia.org/wiki/Descriptive_statistics). Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.
  + From <<https://en.wikipedia.org/wiki/Statistical_inference>>

* + The traditional approach to research, statistical inference, and modeling begins with the specification of a theory or model. Classical or Bayesian methods of statistical inference are employed. (Miller, Chap 1.1)

* + We use information technology and database systems as far as they can take us, and then we do more, applying what we know about statistical inference and the modeling techniques of predictive analytics.(Miller, chap 12)
  + How shall we draw inferences from data? Formal scientific method suggests that we construct theories and test those theories with sample data. The process involves drawing statistical inferences as point estimates, interval estimates, or tests of hypotheses about the population. Whatever the form of inference, we need sample data relating to questions of interest. For valid use of statistical methods we desire a random sample from the population.(Miller, chap 12)

1. Define model dependent research
   * Model-dependent research is the third approach. It begins with the specification of a model and uses that model to generate data, predictions, or recommendations. Simulations and mathematical programming methods, primary tools of operations research, are examples of model-dependent research. When employing a model-dependent or simulation approach, models are improved by comparing generated data with real data. We ask whether simulated consumers, firms, and markets behave like real consumers, firms, and markets. The comparison with real data serves as a form of validation. (miller, chap1.1)
2. Define the following methods and when they are used:
   * scatter plots,
   * **Definition**: Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The variable or attribute which is independent is plotted on the X-axis, while the dependent variable is plotted on the Y-axis. These plots are often called scatter graphs or scatter diagrams.
   * **When is it used:** Scatter diagram has many applications and usages nowadays. Here are some of them:

* When trying to find out whether there is a relationship between 2 variables.
* When having paired numerical data.
* When working with [root cause analysis tools](http://intellspot.com/root-cause-analysis-tools/) to identify the potential for problems.
* When just want to visualize the correlation between 2 large datasets without regard to time.

  From <<http://www.intellspot.com/scatter-plot/>>

* Examples from Miller:

1. **Figure 4.3.** *Market Basket Association Rules: (Chap4)*
   * + **Figure 10.1.** *California Housing Data: Correlation Heat Map for the Training Data (Chap 10)*
   * correlation matrix,

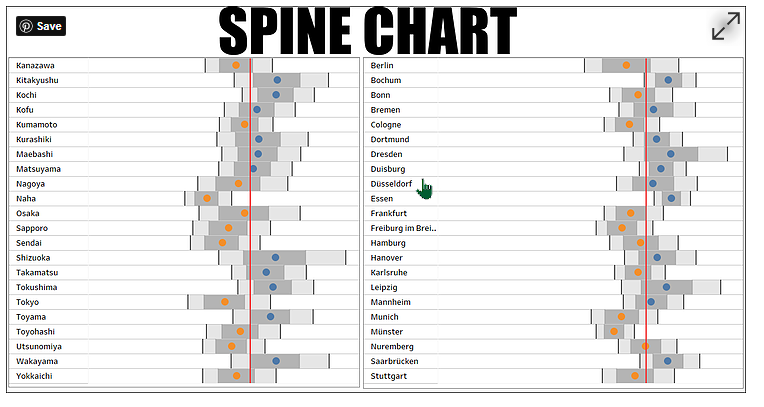
**Definition:** The prototypical validation study involves a multitrait-multimethod matrix, as described by Campbell and Fiske ([1959](https://learning.oreilly.com/library/view/modeling-techniques-in/9780133892123/bib01.html#bib_89)). This correlation matrix provides a structure for demonstrating construct validity. The matrix is partitioned by measurement methods. Within each method there are rows and columns associated with traits (attributes). Each element of the matrix represents a trait-method unit. The components of the matrix are the reliability diagonal, validity diagonals, heterotrait-monomethod triangles, and heterotrait-heteromethod triangles.( miller, A data science methods)

[Components defined (heterotrait-monomethod triangles for ex)](https://conjointly.com/kb/multitrait-multimethod-matrix/#:~:text=The%20Heterotrait%2DMonomethod%20Triangles,the%20same%20method%20of%20measurement.&text=If%20these%20correlations%20are%20high,a%20strong%20%E2%80%9Cmethods%E2%80%9D%20factor.)

* + - **When is it used:**used to investigate the dependence between multiple variables at the same time. The result is a table containing the correlation coefficients between each variable and the others.
  + horizon plot,
    - **Definition**:The horizon plot reduces the amount of space needed to display multiple time series and, through the use of color, makes it easier to detect relationships across time series. Granger causality tests and dynamic regressions notwithstanding, we may detect a leading indicator in a horizon plot long before statistical indicators reveal its existence. (ch. 5)
    - **When is it used:**We can use a horizon plot to visualize text measures in time. This is similar to what we do for horizon plots of economic time series. Here we compute standard scores for each text measure and each document year and then plot the resulting multiple time series. (figure 7.6, miller chap 7)
  + histogram,
    - **Definition:** A histogram is a graphical representation that organizes a group of data points into user-specified ranges. It is similar in appearance to a [bar graph](https://www.investopedia.com/terms/b/bar-graph.asp). The histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.
    - **When is it used:**
      * A histogram is a bar graph-like representation of data that buckeths a range of outcomes into columns along the x-axis.
      * The y-axis represents the number count or percentage of occurrences in the data for each column and can be used to visualize data distributions.
      * In trading, the MACD histogram is used by technical analysts to indicate changes in momentum.



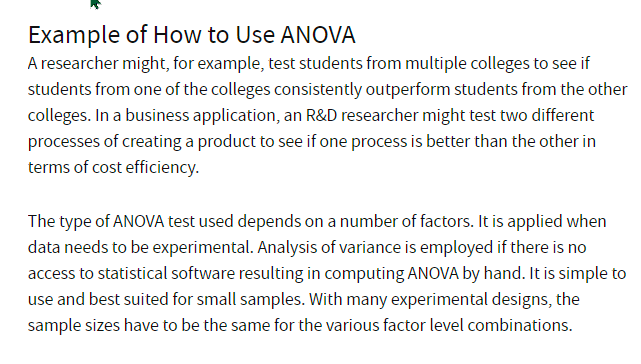
* + spine chart,
    - **Definition:** To display the results of the conjoint analysis, we use a special type of dot plot called the spine chart, shown in [figure 3.1](https://learning.oreilly.com/library/view/modeling-techniques-in/9780133892123/ch03.html#ch03fig01). In the spine chart, part-worths can be displayed on a common, standardized scale across attributes. The vertical line in the center, the spine, is anchored at zero. The part-worth of each level of each attribute is displayed as a dot with a connecting horizontal line, extending from the spine. Preferred product or service characteristics have positive part-worths and fall to the right of the spine. Less preferred product or service characteristics fall to the left of the spine.
    - **When is it used:** To display the results of the conjoint analysis

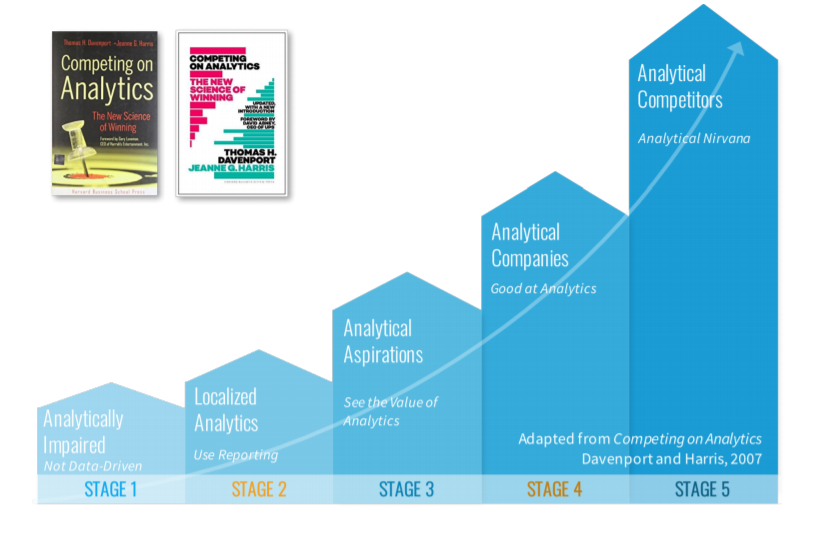


Spine chart is a visualization that shows at a glance how one area compares with others across a range of indicators. Summary statistics of each measure such as the average are aligned to form one central line or ‘spine’ on the chart. This spine is the comparator and is usually the national or regional average to which local area values are compared. Most often the vertebrae show the range of values – the minimum, maximum, 25th and 75th percentiles. Area specific values are placed on top of these summaries, allowing the viewer to understand how the area compares with the rest of areas for that measure. Most often, color is applied on the area specific values to represent significance of its value compared to the comparator.

* + regression (linear regression):
    - **Definition:** Help us to predict a response with meaningful magnitude (Miller, 2)
    - **When is it used**
  + ANOVA:
    - * **Definition:**What is Analysis of Variance (ANOVA)?

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

* + - * **When is it used:**
    - 

1. What is a “method” in data analytics
   * Qualitative and quantitative
2. What data science technologies would you use in a promotional campaign?
   * Chapter 2 of Miller (Modeling Techniques, p16-32) Advertising and Promotion. Exploratory graphics like box plot, scatter plots, strip plots, linear modeling, fitted predictive modeling. Are all used in the bobblehead example.
   * <https://towardsdatascience.com/20-practical-ways-to-implement-data-science-in-marketing-e10da4a6d0b2>
3. What is market basket analysis? What is it used for?
   * Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy. (online)
     + Also see Miller Pages 43-55
4. Be able to describe the concepts from weekly discussions.
5. What is conjoint analysis and what is it used for?
   * Typically used to **evaluate the strength and direction of customer preferences** for a combination of product or service attributes. For example, a CA might be used to determine which factors- price, quality, dealer location, and so on- are most important to customers who are purchasing a new car” (Davenport & Harris, p132)
6. What is an enterprise-level approach?
   * Part of DELTA (Data, Enterprise, Leadership, Targets, Analysts). Instead of managing their analytics resources in disconnected sios, highly analytical firms manage these resources- including data, technology, and analysts- in a coordinated fashion across the enterprise. (Davenport & Harris, p 46-47, covered in-depth on pages 51-54)
   * (essentially centralized management of all aspects of analytics)
   * 
7. Define/list Davenport & Harris’s levels of analytic competition
8. What is an analytically competitive organization? What are its basic characteristics? What are characteristics of analytical leaders?

#### **(Davenport, p 61-65)**

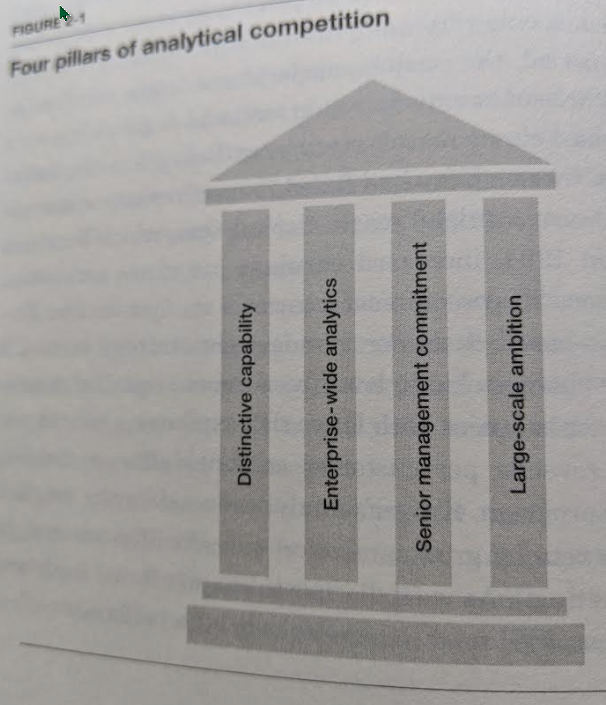
Stage 1: Analytically Impaired (pg65)*- “Have some desire to become more analytical, but thus far they lack both the will and skill”*

Stage 2: Localized Analytics (p65)- “*Exhibit the typical localized analytics approach to ‘business intelligence’ in the past… an emphasis on reporting w/ pockets of analytical activity… do analytical work but no intention of competing with it”*

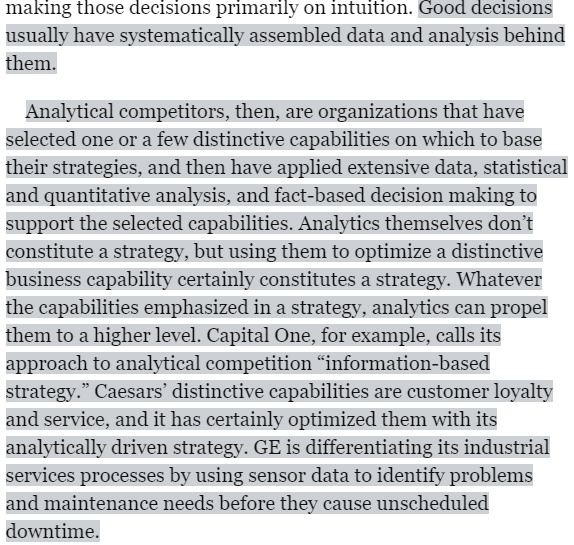
Stage 3: Analytical Aspirations (p64)- *“Grasp the value and promise of analytical competition, but they’re in the early stages of it.. Typically face major capability hurdles and are a long way from overcoming them”*

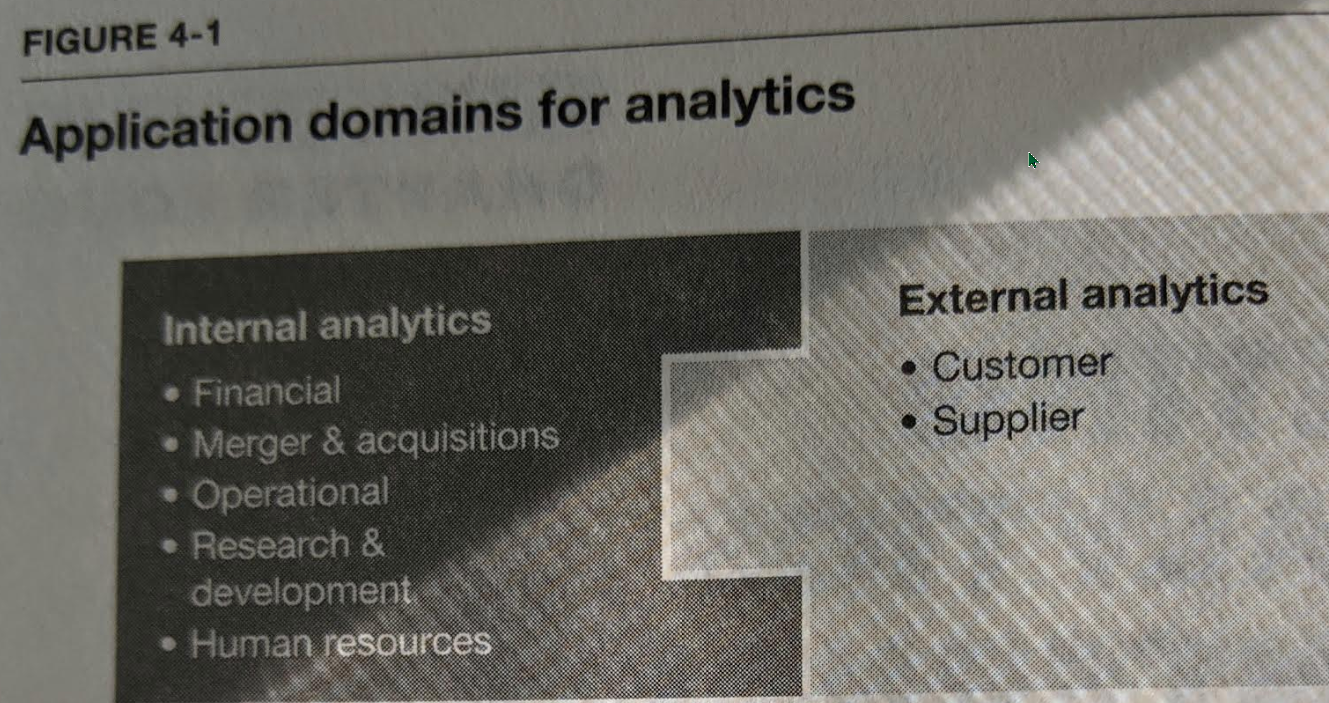
Stage 4: Analytical Companies (p62)- *“On the verge of analytical competition but still have a few minor hurdles to get there in full measure. With only a small increase in emphasis, these companies could move into analytical competitor status”*

Stage 5: Analytical Competitors (p62)- “*High degree of the four factors. Analytical activities are clearly in support of a distinctive capability, taking an enterprise-wide approach, execs are passionate and driving, and their analytical initiatives are aimed at substantial results”*

**

##### Analytically competitive organization (What is an ?) What are its basic characteristics?



1. Define internal and external processes
   * (Internal Process Applications on Davenport, p 92)
   * ****
2. What is supply chain?
   * (Davenport , p148 - Analytical Techniques for SC)

A supply chain is a network between a company and its suppliers to produce and distribute a specific product to the final buyer. This network includes different activities, people, entities, information, and resources. The supply chain also represents the steps it takes to get the product or service from its original state to the customer. (online def)

1. **What are customer related processes? Customer centered management? – maybe davenport**
2. Define a consumer choice model
   * Keep consumer preference in mind
   * **choice-based conjoint (CBC)**, full-profile, adaptive conjoint analysis (ACA), menu-based conjoint, adaptive choice based conjoint, and other approaches have different ways to manage the balance between the number of attributes that can be included and the relative complexity of the choices that need to be shown in order to get good quality data.
3. Define a peer reviewed article
   * Peer review is the evaluation of work by one or more people with similar competences as the producers of the work. It functions as a form of self-regulation by qualified members of a profession within the relevant field.
4. What is a bag of words approach?
   * In developing text measures for the movie document collection, we take what is affectionately called a *bag-of-words* approach to this analysis. We create one document for each of the forty years by combining the taglines from all movies of each year. A bag-of-words approach does not evaluate the meaning of taglines, the order of words, or word combinations. Rather, it considers each word as an individual item for analysis. A bag-of-words approach is distinct from a natural language processing (NLP) approach to text analytics. A bag-of-words approach makes sense for what we are doing here because we are aggregating taglines from hundreds to thousands of movies in the document for each year. These yearly documents are, in fact, bags of words.(Miller, chap7)
5. What is text analysis?
   * To do text analytics we must find ways to structure text so that it can be understood by computers. The two primary ways to do this are the *bag of words* approach and *natural language processing.*
6. What is NLP (Natural Language Processing)
   * Computer programs for natural language processing use linguistic rules to mimic human communication and convert natural language into structured text for further analysis. Natural language processing is a broad area of academic study itself, and an important area of computational linguistics. The location of words in sentences is a key to understanding text. More generally, we are parsing natural language text to arrive at structured text. (Miller, chap 7)
7. What is sentiment analysis? What is it used for?

#### **Sentiment analysis is text analytics with a purpose. It is the use of text measures to learn about the past and make predictions about the future. Sometimes called opinion mining, sentiment analysis draws upon positive and negative word sets (lexicons, dictionaries) that convey human emotion or feeling. These word sets are specific to the language being spoken and the context of application.**

To do sentiment analysis correctly, we need to design text measures that work.  Methods for developing text measures of sentiment, including list-based measures, item-weighted measures, and models for text classification. We employ a training-and-test regimen in evaluating the predictive performance of text measures and models. (Miller, chap8)

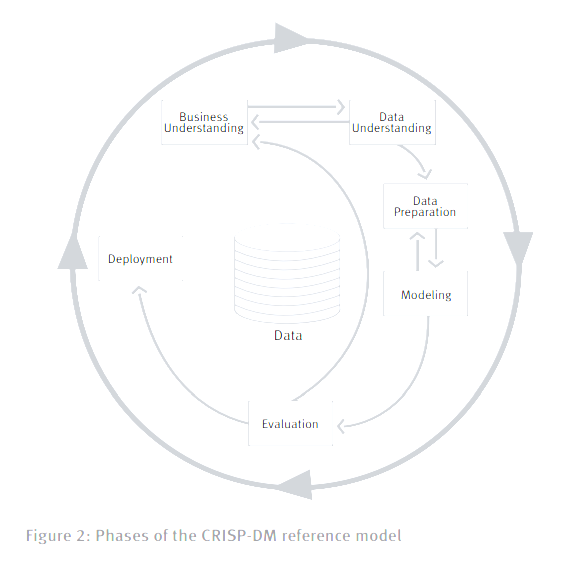
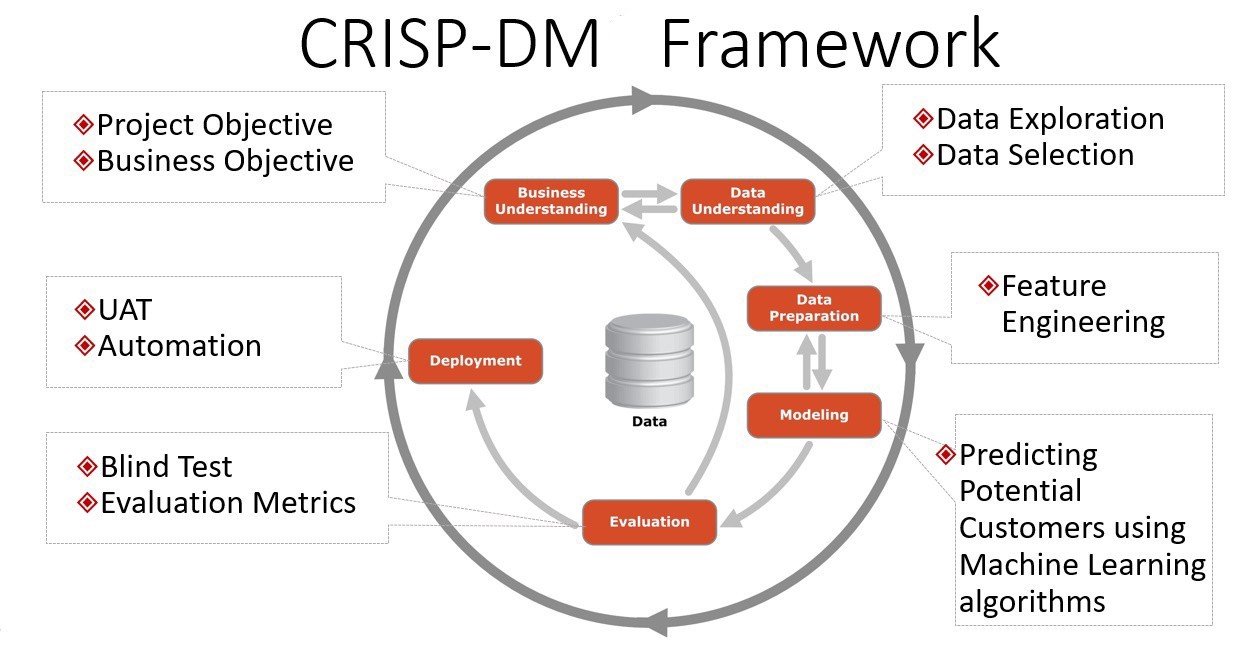
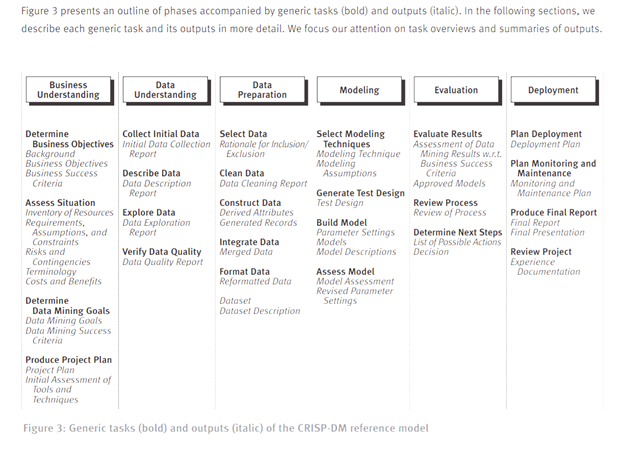
Sentiment analysis is a method of text analysis

* + that attempts to determine if a statement conveys a positive, negative, or neutral attitude about a topic. ( DB Systems, coronel, pg 10) - week 5 reading

1. What is a balanced scorecard approach?
   * A balanced scorecard is a strategic management performance metric used to identify and improve various internal business functions and their resulting external outcomes. Balanced scorecards are used to measure and provide feedback to organizations. Data collection is crucial to providing quantitative results as managers and executives gather and interpret the information and use it to make better decisions for the organization.
   * **management tool** enabling organizations to clarify vision and strategy and translate them into action
   * A method of evaluating a firm's performance using performance measures
2. What is a dashboard?
   * represents complex in a format that  
     is readily recognizable and easily understood using graphics
3. What is the CRISP-DM process?
   * **(Provost, p 27) (CRISP-DM pdf)**

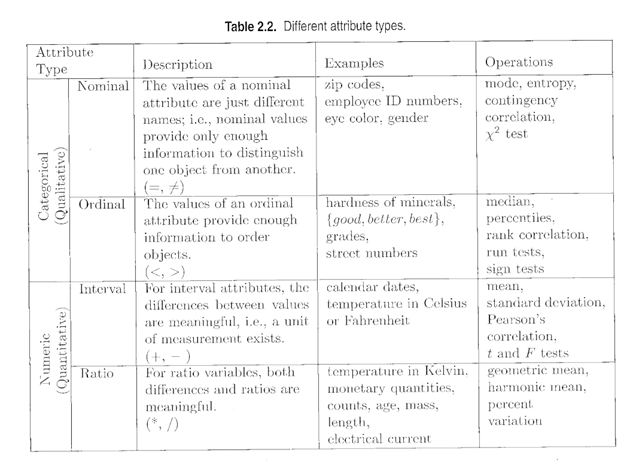
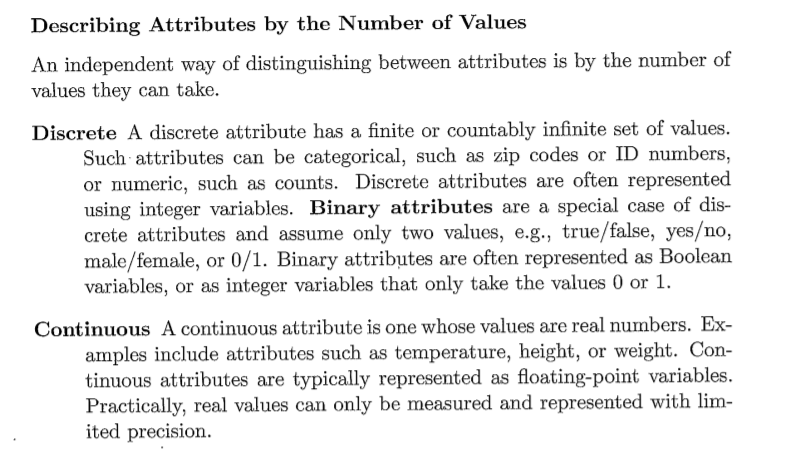
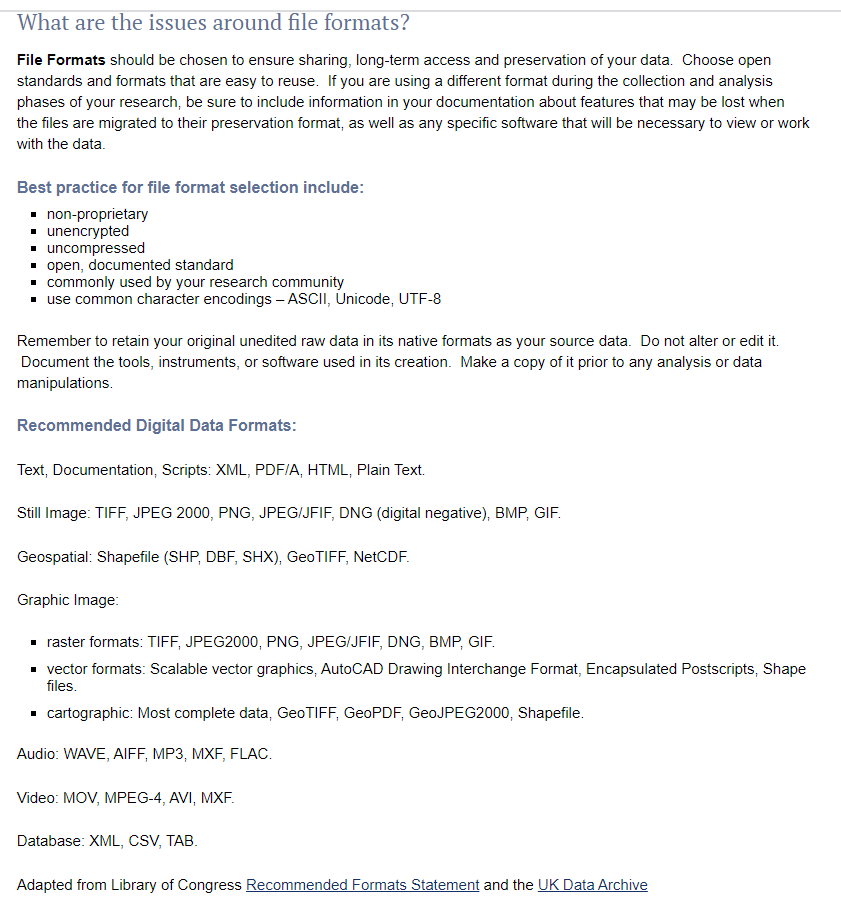
Cross Industry Standard Process for Data Mining

According to CRISP-DM, the data scientist first starts by identifying the business problem and business objectives in the business understanding phase.**!** **True**

* + 
  + 
  + 

1. Define and contrast structured versus unstructured data
   * **Structured data** is data that has been organized to fit a predefined data model. **Unstructured data** is data that is not organized to fit into a predefined data model. **Semistructured** data combines elements of both-some parts of the data fit a predefined model while other parts do not. Relational databases rely on structured data. Most of the data in the world is semistructured or unstructured. Unstructured data

includes maps, satellite images, emails, texts, tweets, videos, transcripts, and a whole host of other data forms (DB Systems, coronel, pg 10) - week 5 reading

1. List and describe the four levels of measurement (nominal, ordinal, interval, ratio). What are categorical data?
   * (Intro to Data Mining pdf pg 26) - nominal, ordinal, interval, ratio
   * ****
2. Define and contrast discrete versus continuous data
   * ****
3. Define and contrast quantitative versus qualitative data
   * Qualitative and Quantitative: Qualitative refers to text, images, video, sound recordings, observations, etc.  Quantitative refers to numerical data.
4. Kinds of file formats and how they are classified: e.g. plaintext, JSON file format
   * 

JSON- JavaScript Object Notation is an open standard file format, and data interchange format, that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and array data type

CSV- Comma Separated Values

XML-Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable

ASCII- text is stored in this format

Lossy, lossless, RAW, bitmap, vector graphics

1. Define CRM based research
   * Customer relationship management (CRM) systems go beyond sales information systems by tracking many customer interactions with a ﬁrm.
   * A key difference between survey-based consumer research and CRM-based research is the way we interact with research participants. In consumer research there is a tradition of protecting research participants from subsequent contact by sales and service personnel. Conﬁdentiality and privacy protection are part of the code of conduct of consumer researchers. Research participants opt-in when ﬁlling out a survey, and they can just as easily opt-out. They're Anonymous Consumers. The customers CRM,on the other hand,are anything but anonymous. They are well identiﬁed,with recorded sales transactions and customer service operations. With CRM,research and service activities are intertwined. (Marketing Data Science, Miller, pdf, p305-6)
2. Define and contrast SQL/NoSQL
   * **SQL** -  is a domain-specific language used in programming and designed for managing data held in a relational database management system, or for stream processing in a relational data stream management system

Data in relational database tables may be accessed/inserted/deleted using SQL.

**true**

* + **NoSQL** is the unfortunate name given to a broad array of non relational database technologies that have developed to address the challenges represented by Big Data. There are literally hundreds of products that can be considered as being under the broadly defined term NoSQL. Most of these fit roughly into one of four categories*: key-value data stores, document databases, column-oriented databases,* and *graph databases.*

1. What is a relational database?
   * (ex. SQL) Relational databases impose a structure on the data when the data is captured and stored. Relational databases often aggregate these into database blocks.  Relational databases rely on structured data (Database Systems- Design…, pdf)
2. What are the uses/purposes of data warehousing?

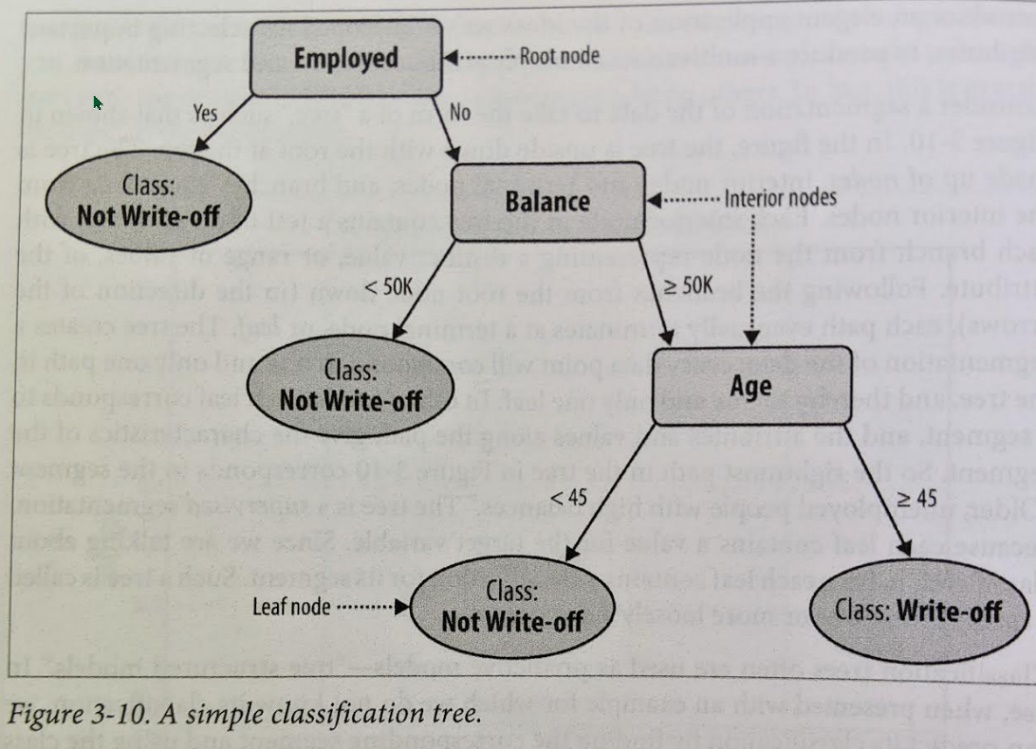
DW- A consolidated/integrated view of corporate data drawn from disparate operational data sources and a  range of end-user access tools capable of supporting simple to highly complex queries to support decision making.

Deemed the solution to meet the requirements of a system capable of supporting decision making, receiving data from multiple operational data sources (Database Systems - A Practical Approach to Design...  pdf, p 75)

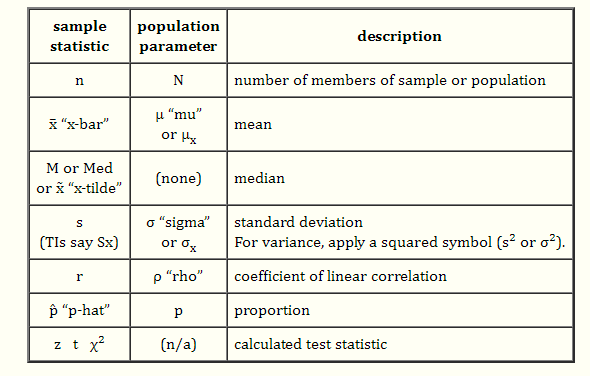
Data warehousing providers focus on providing technology and services aimed toward integrating data from multiple sources.

**true**

1. What is an algorithm?
   * A process or set of operations (Database Systems - Design, Implementation and Mgt pdf)
2. What does it mean to “normalize” data?
   * In transactional systems, normalization is used to decompose complex data into related tables to reduce redundancy and to improve the speed of rapid manipulation of small sets of data. (*Database Systems: Design, Implementation, & Management, pdf, module 5, p 666)*
   * to create normal forms  
     to make data follow strict rules e.g. a field can't hold multiple values  
     values depend on unique primary keys
3. What are accuracy, precision and recall metrics?
   * Accuracy
     + Accuracy (error rate)
       - The rate of correct (incorrect) predictions made by the model over a dataset (cf. coverage). Accuracy is usually estimated using an independent (holdout) dataset that was not used at any time during the learning process. More complex accuracy estimation techniques, such as cross-validation and the bootstrap, are commonly used, especially with datasets containing a small number of instances.
         * Accuracy is a common evaluation metric that is often used in data mining studies because it reduces classifier performance to a single number and it is very easy to measure
         * Accuracy = (Number of correct decisions made)/(Total number of decisions made)
         * Equal to 1 - Error Rate
         * (Provost, p 189)
         * The rate of correct (incorrect) predictions made by the model over a dataset (cf coverage). Accuracy is usually estimated using an independent (holdout) dataset that was not used at any time during the learning process. More complex accuracy estimation techniques, such as cross-validation and the bootstrap, are commonly used, especially with datasets containing a small number of instances (Provost, p 357, Glossary)
         * how close a measurement is to actual value
         * positional, temporal, attributue, logical consistency, data completeness
   * Precision
     + Recall that a classification model takes an instance for which we do not know the class and predicts its class. Part of the evaluation (
       - *Precision* and *Recall* are often used, especially in text classification and information retrieval. Recall is the same as true positive rate, while precision is TP/(TP + FP), which is the accuracy over the cases predicted to be positive. *F-measure* is the harmonic mean of precision and recall at a given point, and is:
         * F-measure= 2 \* ((precision \* recall)/(precision + recall))
       - You may also hear about the *positive predictive value*, which is the same as precision (Provost, p 204)
         * variance value when repeated measurements are taken
   * recall metrics – see above precision
4. Define: symmetrical simulation, clusters in a segmented dataset, unsupervised segmentation
   * symmetrical simulation
     + In statistics symmetry is when you have a normal distribution.
   * clusters in a segmented dataset
     + Classification tree is an example of supervised segmentation.
     + “Segment the population into groups that differ from each other with respect to some quantity of interest. In particular, how can we segment the population with respect to something we would like to predict or estimate?”  (Provost, p43) (Supervised Segmentation- Provost 43-69)
     + (clustering in supervised data is typically called classification)
   * unsupervised segmentation:
     + We may want to group similar items together into clusters, for example to see whether our customer base contains groups of similar customers and what these groups have in common. Previously we discussed supervised segmentation; this is unsupervised segmentation.
       - The idea of finding natural groupings in the data may be called unsupervised segmentation, or more simply, *clustering.* Clustering is another application of our fundamental notion of similarity. The basic idea is that we want to find groups of objects(consumers, businesses, whiskeys, etc.) where the objects within the groups are similar, but the objects in different groups are not so similar. (Provost, p 164)
5. What is a classification tree or a decision tree?
   * Consider a segmentation of the data to take the form of a “tree”, such as that shown in fig 3-10. In the figure, the tree is upside down with the root at the top. The tree is made up of *nodes*, interior nodes and terminal nodes, and branches emanating from the interior nodes. Each interior node in the tree contains a test of an attribute, with each branch from the node representing a distinct value, or range of values, of the attribute. Following the branches from the root node down (in the direction of the arrows), each path eventually terminates at a terminal node, or *leaf*. The tree creates a segmentation of the data: every data point will correspond to one and only one path in the tree, and thereby to one and only one leaf. In other words, each leaf corresponds to a segment, and the attributes and values along the path give the characteristics of the segment. So the rightmost path in the tree in Fig 3-10 corresponds to the segment. “Older, unemployed people with high balances”. The tree is a *supervised* segmentation, because each leaf contains a value for the target variable. Since we are talking about classification, here each leaf contains a classification for its segment. Such a tree is called a *classification tree* or more loosely a *decision tree.* (Provost, p 62-66)

**

1. What is a “nearest neighbor” and other distance related models?
   * “Now that we have a way to measure distance, we can use it for many different data-analysis tasks. Recalling examples from the beginning of the chapter, we could use this measure to find the companies most similar to our best corporate customers, or the online consumers most similar to our best retail customers. Once we have found these, we can take whatever action is appropriate in the business context. For example, IBM does this to help direct its sales force. Online advertisers do this to target ads. These most-similar instances are called *nearest-neighbors*” (Provost, p 144-155)
2. Define and state when use is (or is not) appropriate: SRS (simple random sample); stratified sample, cluster sample, opportunistic sample, nonprobability sample, purposive sample, deliberate sample.
   * **SRS  (simple random sample) –** Select units from the sample frame by a truly random process.  Use when there is a complete and accurate sampling frame and when sample bias is a critical issue
   * **stratified sample –** Randomly sample a known quota of units within defined strata of the population.  Use when there is detailed information about the characteristics of the population, cost of contacting and screening respondents is low, and sampling error must be avoided
   * **nonprobability sample –** can be generalized to the population only when making strong assumptions (should be avoided)
     + ***cluster sample –*** *Every unit in the aggregate unit selected from the sampling frame is included in the sample.  Use when interviewing costs are high, and the clusters correspond to collections of observations that are useful in research.  (All mid-level managers in 10 out of 100 firms)*
     + ***opportunistic (convenience) sample –*** *Units are selected because they are convenient.  Use with populations that are organized into groups, large organizations that are divided into departments.  (People exiting malls)*
     + ***purposive sample –*** *when targeted information about small subpopulations is needed, the number of potential observations are limited or hard to obtain, or the population is small*
     + ***deliberate sample -*** *non probability sample design in which researcher purposively selects certain units of the universe to form a sample. (purposive sampling)*
3. Sampling methods for underrepresented groups.
   * One way to compensate for underrepresented groups in a polling sample is to:  
     a. eliminate all answers from those who are  
     part of the underrepresented group.  
     b. reduce the weight given to the  
     underrepresented group so that it does not  
     count excessively.  
     c. add extra weight to correct for the  
     underrepresented group.
4. Symbols in sampling (n, N)
   * Symbol in sampling (n, N) - **n is the sample number, N is the population number**



1. Why sample?
   * reduces time/cost/processing of responses
2. What is the difference between sampling with replacement and sampling without replacement?
   * **Sampling with replacement –** two items are independent.  Have 1/n chance of choosing every item
   * **Sampling without replacement –** you do not replace the item before you choose a second
3. What are sampling error, response bias, face validity, construct validity, content validity, reliability? To what which of these does Korzybski’s statement that “The map is not the territory, and the name is not the thing named” relate?
   * Sampling Error:
     + A sampling error is a statistical error that occurs when an analyst does not select a sample that represents the entire population of data and the results found in the sample do not represent the results that would be obtained from the entire population. Sampling is an analysis performed by selecting a number of observations from a larger population, and the selection can produce both sampling errors and non-sampling errors.
       - sampling bias – some members of the sampling frame are given no or reduced chance of selection
       - sampling variance – given the design for the sample, by chance many different sets of frame elements could be drawn
   * Response bias : **Response bias** is a general term for a wide range of tendencies for participants to respond inaccurately or falsely to questions. These biases are prevalent in research involving participant [self-report](https://en.wikipedia.org/wiki/Self-report_study), such as [structured interviews](https://en.wikipedia.org/wiki/Structured_interview) or [surveys](https://en.wikipedia.org/wiki/Statistical_survey).[[1]](https://en.wikipedia.org/wiki/Response_bias#cite_note-Furnham-1) Response biases can have a large impact on the validity of [questionnaires](https://en.wikipedia.org/wiki/Questionnaire) or surveys.[[1]](https://en.wikipedia.org/wiki/Response_bias#cite_note-Furnham-1)[[2]](https://en.wikipedia.org/wiki/Response_bias#cite_note-Nederhof-2)

response deviations described above are systematic, that is,

if there is a consistent direction of the response deviations over trials, then

"response bias" might result. "Bias" is the difference between the expected value

(over all conceptual trials) and the true value being estimated. Bias is a systematic

* + - distortion of a response process.
  + face validity :
    - **Face validity** is the extent to which a test is [subjectively](https://en.wikipedia.org/wiki/Subjectivity) viewed as covering the concept it purports to measure. It refers to the transparency or relevance of a test as it appears to test participants.[[1]](https://en.wikipedia.org/wiki/Face_validity#cite_note-1)[[2]](https://en.wikipedia.org/wiki/Face_validity#cite_note-2) In other words, a test can be said to have face validity if it "looks like" it is going to measure what it is supposed to measure.[[3]](https://en.wikipedia.org/wiki/Face_validity#cite_note-3) For instance, if a test is prepared to measure whether students can perform multiplication, and the people to whom it is shown all agree that it looks like a good test of multiplication ability, this demonstrates face validity of the test.
  + Content Validity:
    - Face validity is often contrasted with [content validity](https://en.wikipedia.org/wiki/Content_validity) and [construct validity](https://en.wikipedia.org/wiki/Construct_validity). **content validity** (also known as **logical validity**) refers to the extent to which a measure represents all facets of a given construct. For example, a [depression](https://en.wikipedia.org/wiki/Depression_(mood)) scale may lack content validity if it only assesses the [affective](https://en.wikipedia.org/wiki/Affective) dimension of depression but fails to take into account the [behavioral](https://en.wikipedia.org/wiki/Behavioral) dimension. An element of subjectivity exists in relation to determining content validity, which requires a degree of agreement about what a particular [personality trait](https://en.wikipedia.org/wiki/Personality_trait) such as [extraversion](https://en.wikipedia.org/wiki/Extraversion) represents. A disagreement about a personality trait will prevent the gain of a high content validity.[[1]](https://en.wikipedia.org/wiki/Content_validity#cite_note-1)
  + Reliability:
    - low reliability is when you have response variance that leads to instability in the values of estimates over trials
  + **(errors of observation –** 
    - deviations from answers given to a survey question and the underlying attribute being measured)
  + **(errors of non-observation –** 
    - deviations of a statistic estimated on a sample from that of the full population)
  + Korzybski’s statement : .
    - in all thought or perception or communication about perception, there is a transformation, a coding, between the report and the thing reported, the Ding an sich (the thing in itself)... Naming is always classifying, and mapping is essentially the same as naming.

1. How to properly form survey questions?
   * Survey questions and answer pdf
2. Briefly, what is imputation? When is it used?
   * The process of replacing missing data with substituted value.  Used when a survey item is missing. In a data set before analysis.
3. What is the purpose of the Nuremberg Code?
   * Ensures the rights of subjects in medical research.  is a set of [research ethics](https://en.wikipedia.org/wiki/Research_ethics) principles for [human experimentation](https://en.wikipedia.org/wiki/Human_experimentation) created as a result of the [Nuremberg trials](https://en.wikipedia.org/wiki/Subsequent_Nuremberg_trials) at the end of the [Second World War](https://en.wikipedia.org/wiki/World_War_II).
   * <https://en.wikipedia.org/wiki/Nuremberg_Code>
4. What was the Declaration of Geneva?
   * <https://en.wikipedia.org/wiki/Declaration_of_Geneva>
   * It is a declaration of a [physician](https://en.wikipedia.org/wiki/Physician)'s dedication to the [humanitarian](https://en.wikipedia.org/wiki/Humanitarian) goals of [medicine](https://en.wikipedia.org/wiki/Medicine), a declaration that was especially important in view of the medical crimes which had just been committed in [German-occupied Europe](https://en.wikipedia.org/wiki/German-occupied_Europe). The Declaration of Geneva was intended as a revision[[1]](https://en.wikipedia.org/wiki/Declaration_of_Geneva#cite_note-1) of the [Hippocratic Oath](https://en.wikipedia.org/wiki/Hippocratic_Oath) to a formulation of that oath's moral truths that could be comprehended and acknowledged in a modern way.[[2]](https://en.wikipedia.org/wiki/Declaration_of_Geneva#cite_note-wma.net-2) Unlike the case of the Oath of Hippocrates, the World Medical Association calls the statement a "pledge".
5. What was the point of the article called “The New Deal on Data”?
   * That brings up interesting points around who owns the asset and how to regulate it
   * Setting up ownership rights, eligibility to opt-in and opt-out from data sharing is essential to gain the trust of the data generators.
   * raises the valid point that without regulation, a potential future disaster could occur, which would bring about a much higher level and potentially industry killing level of regulation in the future.
   * Who's going to hold the complete picture of your life? Some credit-rating service? Google?
6. What is a sample statistic (the point estimate) in a sampling equation, e.g. (p)(1-p)?
   * The p-value, itself a sample statistic, gives the probability of rejecting the null hypothesis under the assumption that it is true.
7. [Expected value](https://www.statisticshowto.com/probability-and-statistics/expected-value/#formula) of the sampling distribution of P̄:  
   E(p̄) = p.
8. [Variance](https://www.statisticshowto.com/probability-and-statistics/variance/) for the sampling distribution of P̄:  
   p(1-p) / n.
9. [Standard Error](https://www.statisticshowto.com/what-is-the-standard-error-of-a-sample/)(SE) of the Sample Proportion:  
   √ (p(1-p) / n). **Note**: as the [sample size](https://www.statisticshowto.com/probability-and-statistics/find-sample-size/) increases, the standard error decreases.
10. You can use the [normal distribution](https://www.statisticshowto.com/probability-and-statistics/normal-distributions/) if the following two formulas are true:
    1. np≥5
    2. n(1-p)≥5.
11. [Z Score](https://www.statisticshowto.com/probability-and-statistics/z-score/) for sample proportion: z = (P̄ – p) / SE
    * 1. Point Estimate (Miller, p 281)
12. Human Subject Review:
    * *to ensure research involving humans applies ethical principles and complies with federal regulatory requirements for protecting the rights and welfare of human participants*