Data Mining, or Data Analytics

Data Understanding

- A key element of DATA MINING is developing a thorough understanding of the data at hand
- Hence, one must understand the attribute domains, and must understand how values are distributed within the domain
    - **Descriptive statistics**

Descriptive Statistics

- The goal of descriptive statistics is to provide values that describe a data set without someone having to examine all of the data. For example:
  - it is simpler to say that the class average on an exam was 78% than to list all of the individual scores.
- Some useful descriptive statistics tools are:
  - Functions to calculate all of these statistics (available in Oracle, MySQL, Excel and Python)
  - If your dataset is not too large, Excel can be a valuable tool for calculating and visualizing your descriptive statistics.
- More and more, Jupyter Notebook is used to write Python code that invokes machine learning and statistical libraries
- R Language

Indicators per data type: Numeric domain

Count: How many values are there?

- This is not the same as how many rows are in the table as some values may be NULL.
- The number of values, **n**, present is important for how much confidence you have in the statistic.
- If your **n** is small you may not put as much faith into the other statistics that you capture.

Range What is the range of the value? This can be found using the MAX() and MIN() functions.

Mean The arithmetic mean of the attribute values: a useful statistic, but it can sometimes be misleading if there are any outliers: values that are much bigger or smaller than most. A common operation is to “centre” values around a mean. That is to take each value and subtract the mean from it. This will make values that are less than the mean negative and values that are greater than the mean positive.

Median The median is that value that splits the data in half. Half of the values are larger and half are smaller.

- If the median is lower than the mean then there are more low values and
- if the median ishigher than the mean then there are more high values.

Indicators per data type: **Numeric domain**

Mode: The mode is the most frequently occurring value. With continuous values it may be necessary to round or otherwise "bin" the values before calculating the mode.

- For an example of binning, think of temperature data. In Calgary we have temperatures that range from about -35.0 degrees to +35.0 degrees. For some applications we could create bins that span 5 degrees. We could to the nearest 5 and end up with -35, -30, -25, -20, -15, -10, -5, 0, 5, 10, 15, 20, 25, 30, 35.

Standard Deviation: The standard deviation measures the variance in the data set. For normally distributed data 2/3 of the values are within +/- one standard deviation of the mean.

- A small standard deviation means values tend to be close to the mean.
- If you find that the mode, median and mean are all the same (or very
- close) you can assume that the data is *normally distributed*

Indicators per data type: Non-**numeric domain**

Obviously, most of the statistics above will not work for non-numeric

data, but it is possible to look at the range, the mode and at a frequency distribution for non-numeric values.

- For example the distribution of letter grades is something I often calculate.

Frequency Distributions

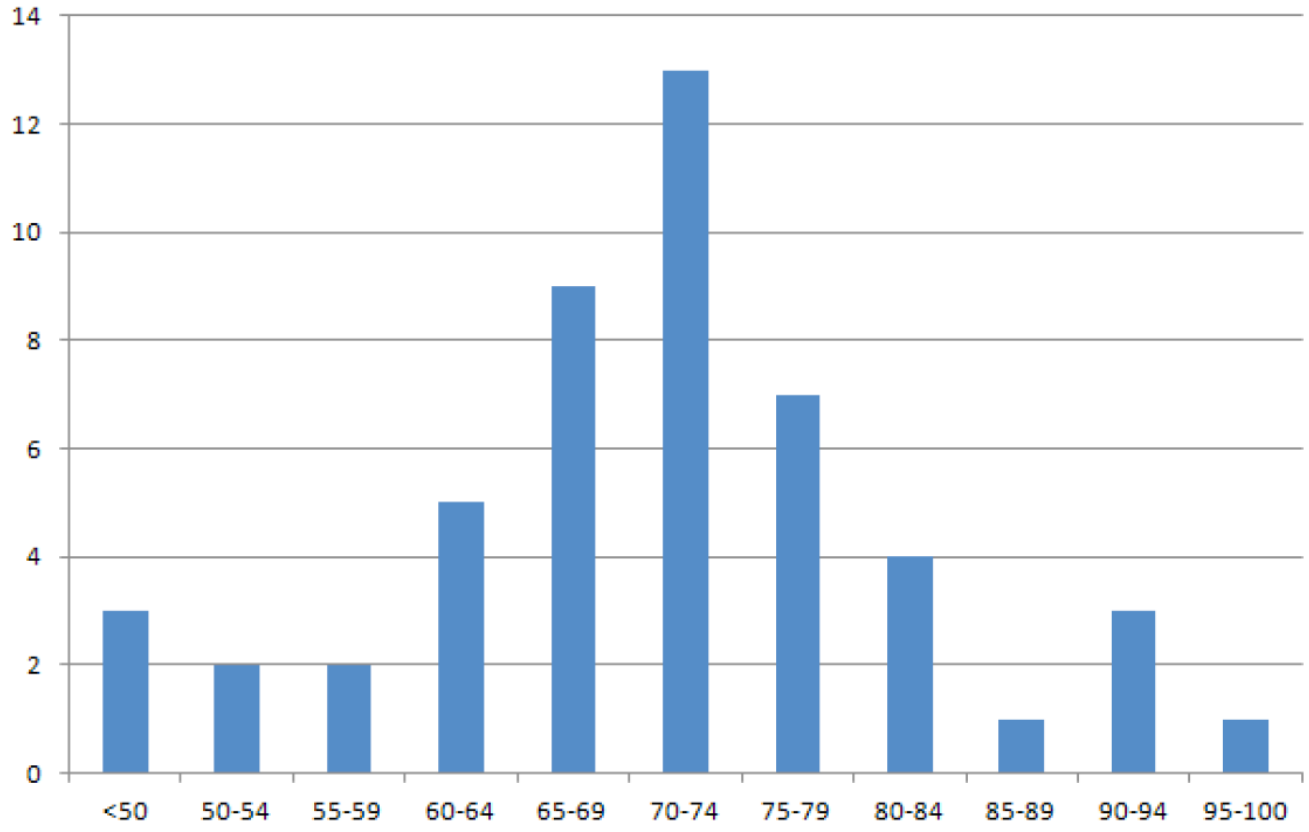
A very useful way to visualize the data is with a frequency distribution histogram. This works for continuous, discrete and even non numeric data.

- If there are a large number of possibilities, the idea here is to place the data into bins.
- For example, if looking at integer class grades one could create bins of < 50, 50-54, 55-59, 60-64, ... 95-100.
- Then count how many grades fall into each bin.
- If you have data with a smaller number of possibilities (a more restricted
- domain), like letter grades, you can just use the data as is.

Frequency Distributions

Distributions are one of the most useful and interesting things you can do when exploring data.

- Is it a normal distribution?
- Is it skewed?
- Are there outliers?

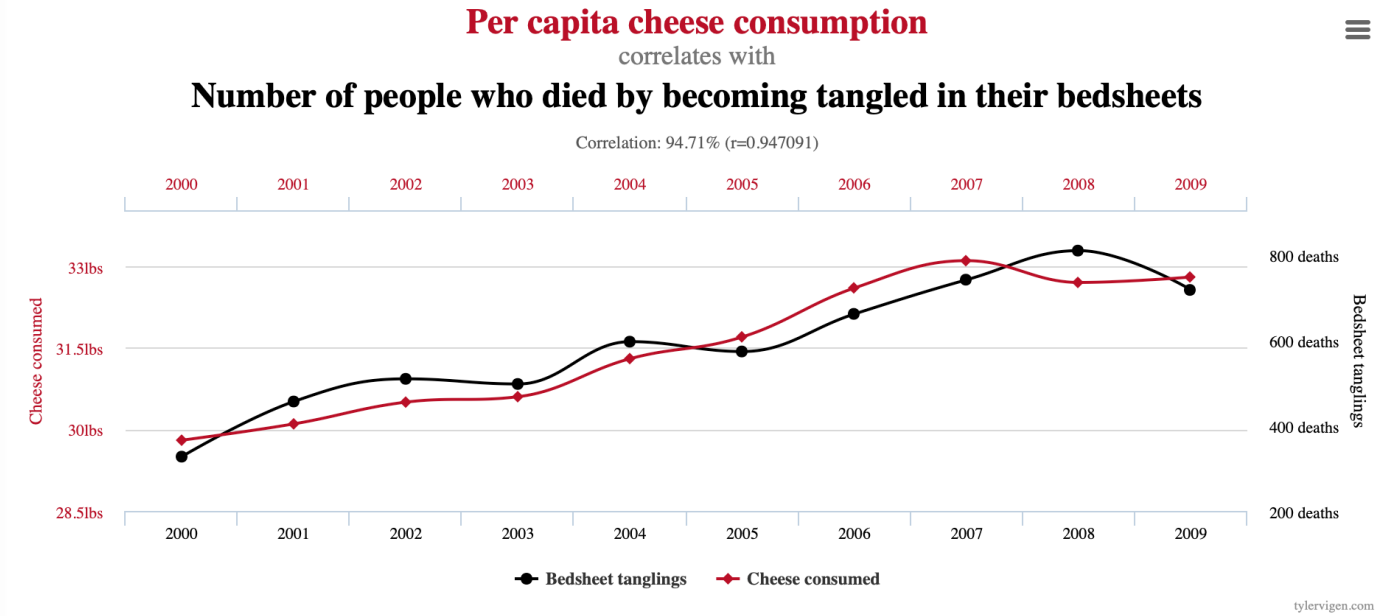


Relationships between attributes

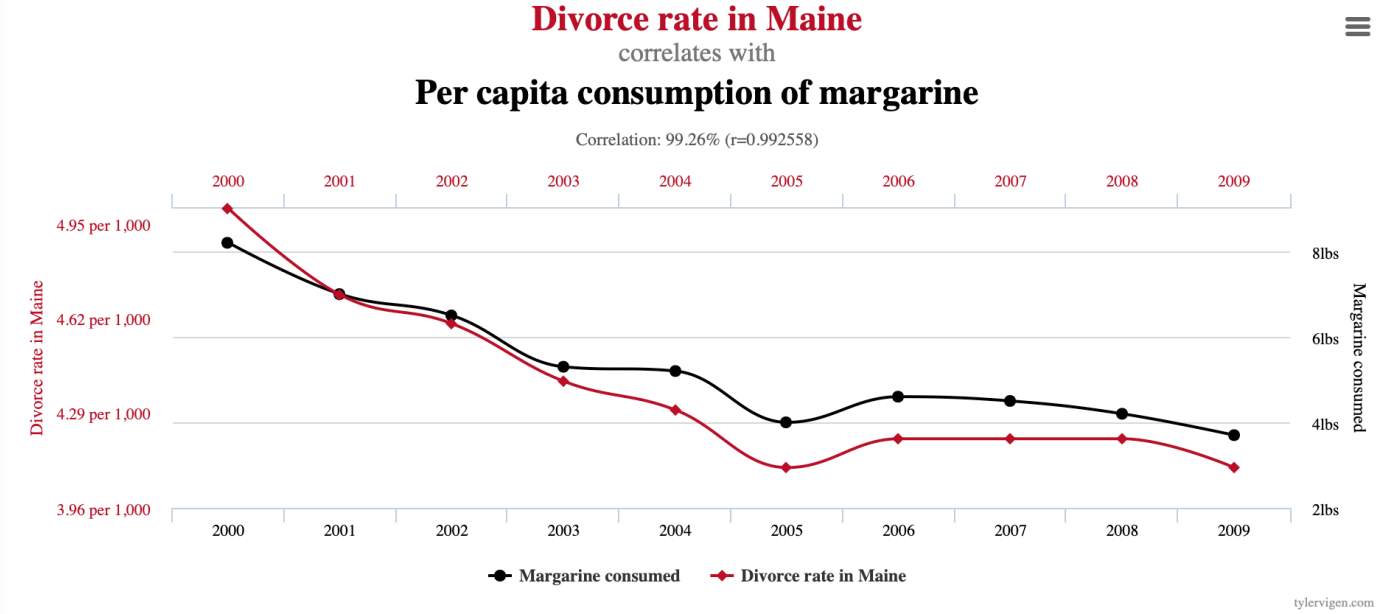
- So far we have been looking at only one attribute at a time.
- If you suspect that two attributes (A and B) may have some kind of relationship you can calculate the correlation coefficient between two data sets.
  - The correlation coefficient is a number between -1 and 1.
  - If the value is 0, then there is no relationship between the two values.
  - If the value is 1, then there is a strong positive correlation
    - that is when the value of A increases, the values of B also increases.
  - If the value is -1, then there is a strong negative correlation
    - that is when the value of A increases, the values of B decreases.
  - When the value is between 0 and 1 it is a judgement call about how strong a correlation exists.

- Remember that just because a two variables have a high correlation coefficient it does not necessarily mean there is any relationship between the two, it may simply be coincidence.
- The correlation value does not say anything about cause

Relation



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention



Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

Inferential Statistics

Descriptive statistics describe the data, as it is

- **Inferential statistics** are used to make predictions.
- You use the data that you have to create a formula that can be used to predict values outside of the data you have already collected.

Regression and Making Predictions

Linear regression is the most common type of inferential statistics

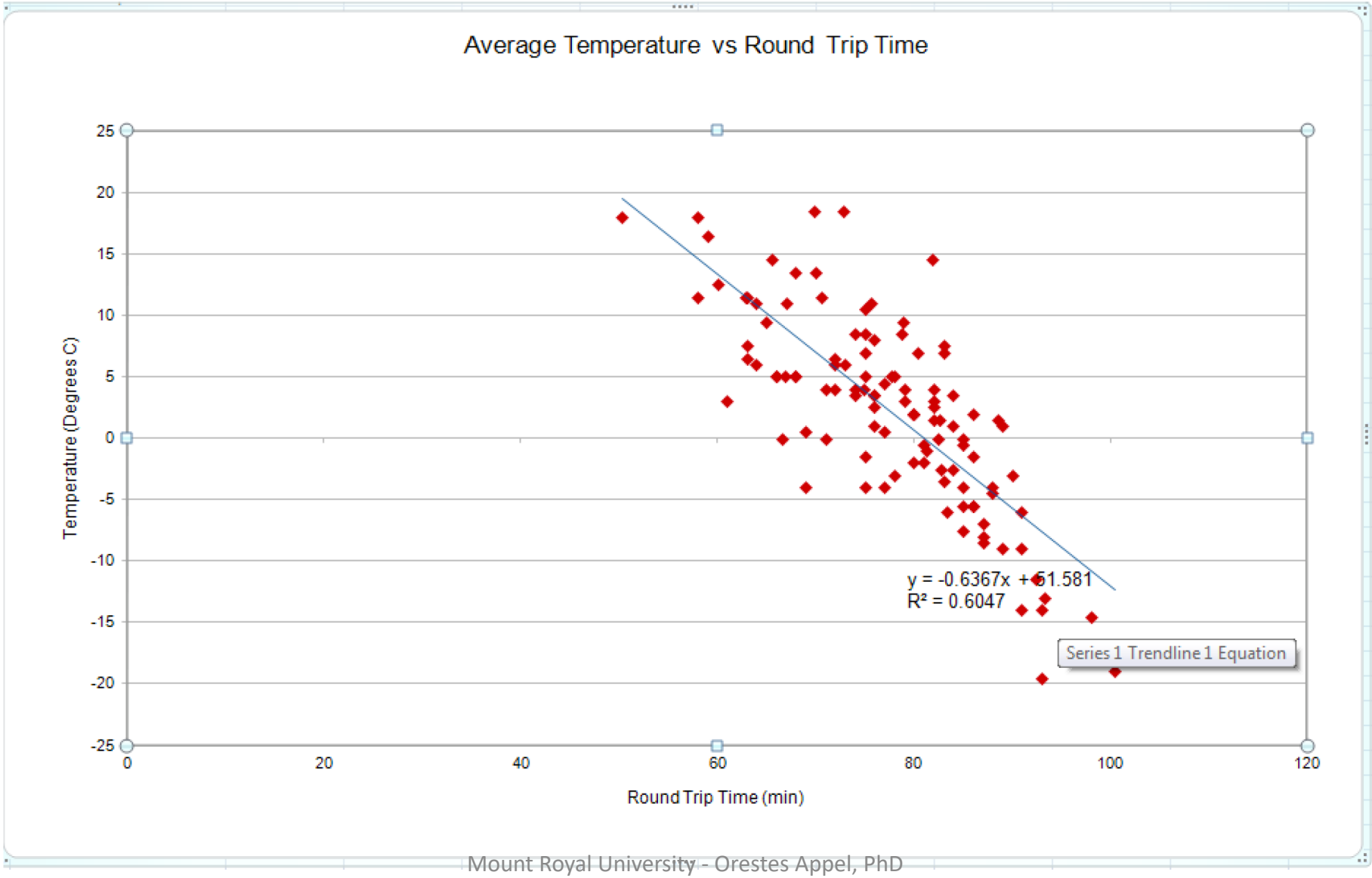
- **One tries to find a linear equation that describes how the attributes are related**.
  - $y = mx + b$

Note: It is possible to try to find polynomial, linear equations with many variables, exponentials etc. but these are beyond the scope of this course.

Regression and Making Predictions

- Usually, a scatter plot is created graphing one of the attributes on the x-axis and one on the y-axis.
  - Then, a line of best fit can be calculated.
- The Seaborn package in Python and Excel, and Oracle all have built-in functions to do these.

Scatter Plot



Example in Oracle

```
/* Using Oracle to perform linear regression */

/* Create a data table and populate it with data. */
DROP TABLE foo;
CREATE TABLE foo ( x INTEGER, y INTEGER);

INSERT INTO foo VALUES (1,3);
INSERT INTO foo VALUES (2,5);
INSERT INTO foo VALUES (3,7);
INSERT INTO foo VALUES (5,10);
INSERT INTO foo VALUES (10,22);
INSERT INTO foo VALUES (11,23);
INSERT INTO foo VALUES (12,25);

/* For a linear equation we need to find the slope, m, and the y-inter

   We also want to estimate how well our line fits the data. The
   R2 coefficient of determination R^2 gives us a number betweeen
   0 and 1. 1 is a perfect fit.
```

```
SELECT
   REGR_SLOPE(y, x) SLOPE,
   REGR_INTERCEPT(y, x) INTERCEPT,
   REGR_COUNT(y, x) COUNT,
   REGR_R2(y, x) R2
FROM foo;
```

Commercial Software

Data Mining

MonkeyLearn | No-code text mining tools

RapidMiner | Drag and drop workflows or data mining in Python

Oracle Data Mining | Predictive data mining models

IBM SPSS Modeler | A predictive analytics platform for data scientists

Weka | Open-source software for data mining

Knime | Pre-built components for data mining projects

H2O | Open-source library offering data mining in Python

Orange | Open-source data mining toolbox

Apache Mahout | Ideal for complex and large-scale data mining

SAS Enterprise Miner | Solve business problems with data mining

Data Analytics

Microsoft Power BI

SAP BusinessObjects

TIBCO Spotfire

SAS Business Intelligence

Tableau

Google Data Studio

IBM Cognos

KNIME

Jupyter Notebook

Pyhthon, Excel, R

Association Rules

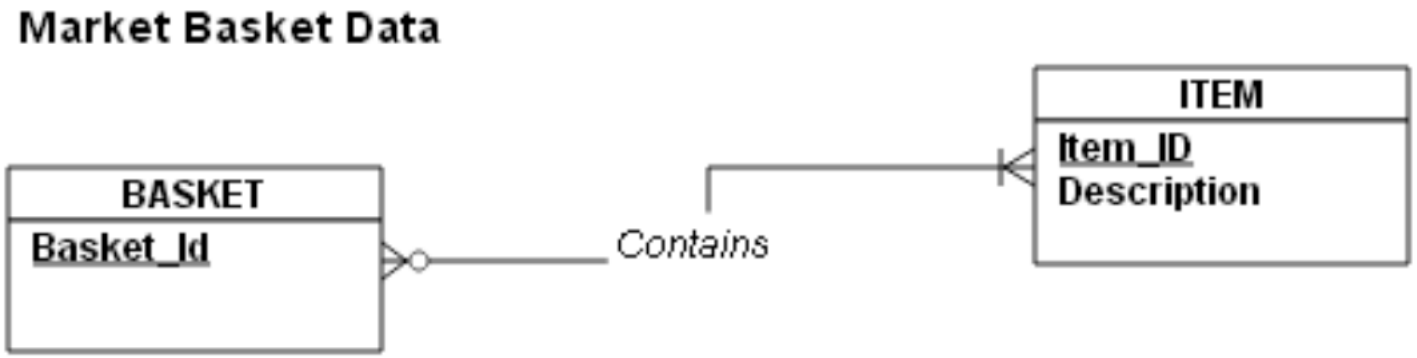
Association Rules

- For example, go to http://chapters.ca and select a book for purchase.
  - You will immediately be presented with a selection of books that others also looked at or purchased.
  - Which other books to display is determined by association rules that have been found in the purchase data at Chapters.
  - A grocery store may find, from **data mining**, that customers who buy milk often also buy bread, so they can place the bread and milk in the store in such a way that customers are enticed to buy other stuff as well (tofu, pork-rinds).

By Sander van der Wel from Netherlands [360/365] Private, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=34928584

Association Rules – Market Baskets and Rules

- Finding associations is a common goal in data mining
- Most often, data used for finding associations in organized into a Market Basket model
- Items can be objects for purchase in a store, or other things like hashtags in a tweet.
- Usually, there are a large number of baskets and some smaller number of items



Association Rules – Market Baskets and Rules

- We are interested in finding patterns in what goes into the baskets, that is, what items were bought together, or which hashtags appear together in a set of tweets.
- These patterns are usually expressed as rules.
- The formula milk ⟹ bread denotes an association rule between milk and bread. Formally this is read as milk implies bread.
  - milk is the antecedent and bread is the consequent.
- For an association to make any sense it must have a population.
  - The population is a set of instances, sometimes called baskets (like a shopping basket)
- For example, at a grocery store:
  - an instance would be the list of all products purchased during one trip to the store.
  - For the on-line book sellers an instance may be all items that the customer has ever bought (*when* they bought the book may be less important).
  - For the tweet example a basket may be a single tweet.

Properties: Support

- First property to understand is the *Support*
- Support is defined as the percentage of baskets that have a particular item or items in them.
  - Suppose we have a population P where |P| = 10,000 of grocery store baskets
  - If there are 100 baskets that contain **tofu**, then:
    - Supt(tofu) = 100 / |P| = 0.01 = 1%
  - If there are 500 that contain **pork-rinds**, then:
    - Supt(pork-rinds) = 500 / |P| = 0.05 = 5%
  - If there are 2 instances with both **tofu** and **pork-rinds** in the same basket:
    - Supt(pork-rinds, tofu) = 2 / |P| = 0.02%

Given a rule R: A == > B, the support of R is the percentage of instances in the population that have both the A and B in them.

Properties: Confidence

- Given a rule R: A == > B, the confidence of R is the likelihood that B appears in a basket given that A is also there
- If someone buys milk, what is the likelihood that they will also buy bread?
  - It can be calculated by finding the support of the rule and dividing by the support of antecedent: Conf(R:A == > B) = Supt(R) / Supt(A)
  - If we have a rule R: tofu == > pork rinds, then:

    Conf(R) = Supt(R) / Supt(A) = 0.02 / 1 = 0.02 (not much confidence in the rule)
  - Let's see another example: Supt(bread) = 30, Supt(milk) = 40, and Supt(bread, milk) = 26
  - From Rule R: bread == > milk, Supt(R) = 26 AND Conf(R) = Supt(bread,milk) / Supt(bread) = 26/30 = 87% (lots of confidence in this rule!)

Properties: Lift

- The Lift of a Rule R: A == > B is the ratio of how many times B will appear when A appears: Lift(R: A == > B) = Conf(R) / Supt(B)
- It may be defined as well as: Supt(R) / (Supt(A) x Supt(B))
- A Lift of 1 means there is no association between A and B
- A Lift greater than 1, means that there is a likelihood A and B will appear together
- A Lift less than 1, means there is no likelihood A and B will be bought together

Properties: Interest

- There is another measure, the **Interest**, that may be of *interest*.
- If we have a rule, R, A ⟹ B then the interest is defined as

  I(R) = Conf(R) – Supt(B)
- If R has no influence on B the fraction of baskets including A and B would be the same as the fraction of all baskets that contain B, and so would have an interest of 0.
- High positive interest means R influences B to appear
- High negative interest mean R influences B to not appear.

Finding frequent Itemsets

- To find rules from a dataset, one first finds frequent itemsets. These are sets of items that occur often, or formally, they have high support.
- If there are only a small number of items one could produce a list of all the possible subsets of the items and check the support of each.
- However, the number of subsets grows exponentially, and the problem quickly becomes intractable.
  - For example, if a store sells 1000 different items (a modest amount by today's big store standards) then there are 21000 possible combinations of purchases or possible instances.
  - Clearly it is not possible to check them all for support or confidence.

Many algorithms have been developed to find frequent itemsets. One is the **A-priori method**.

A-priori Method

1. Decide on a threshold of support, for example 60%.
2. Create all the possible single item itemsets and calculate the support for each one.
3. Discard all items that have less than the threshold of support.
4. With the remaining items create all possible pair itemsets and calculate the support for each of these.
5. Discard all pairs with less than the threshold of support.
6. Create itemsets with three items and continue the process.

- This works by pruning the search space and by the fact that for a set to have high support, all of its subsets must also have high support.
- Now that we have a set of frequent itemsets we need to find rules from them.

Finding Rules

If we have an itemset J with n items, we can create n rules of the form:

J – {j} ⟹ j for all j ∈ J.

- For example, if we have an itemset {A,B,C} we can generate three rules:
  - B , C ⟹ A
  - A, C ⟹ B
  - A, B ⟹ C

Which of these rules should we use? We can calculate the confidence, support and lift for each rule, which is not hard as we already calculated the support figures when we found the frequent itemsets.

We want high support, high confidence, high lift rules.

Example

An example with phrases.

Item	Text
1	Cat, and, dog, bites
2	Yahoo, news, claims, a, cat, mated, with, a, dog, and, produced, viable, offspring
3	Cat, killer, likely, is, a, big, dog
4	Professional, free, advice, on, dog, training, puppy, training
5	Cat, and, kitten, training, and, behavior
6	Dog, &, Cat, provides, dog, training, in, Eugene, Oregon
7	"Dog, and, cat", is, a, slang, term, used, by, police, officers, for, a, male-female, relationship
8	Shop, for, your, show, dog, grooming, and, pet, supplies

Data Quality

Information Cleansing or Scrubbing

A process that weeds out and fixes or discards inconsistent, incorrect, or incomplete information.

- Software tools use sophisticated algorithms to parse, standardize, correct, match and consolidate warehouse information.
- Process is done during the ETL process and once it is in the warehouse.
- Critical when data exits in several operational systems.

Information Cleansing or Scrubbing

Customer Contact Data in Operational Systems

4.4

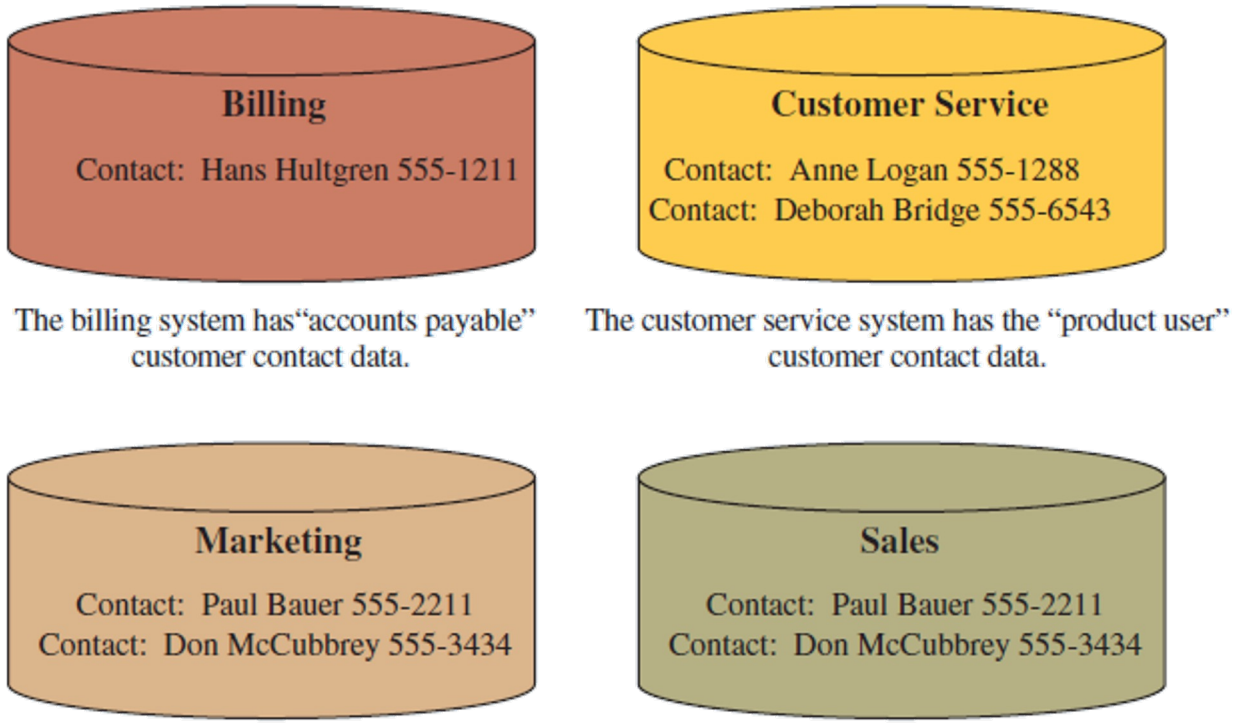


Figure 4.9

Standardizing Customer Name from Operational Systems

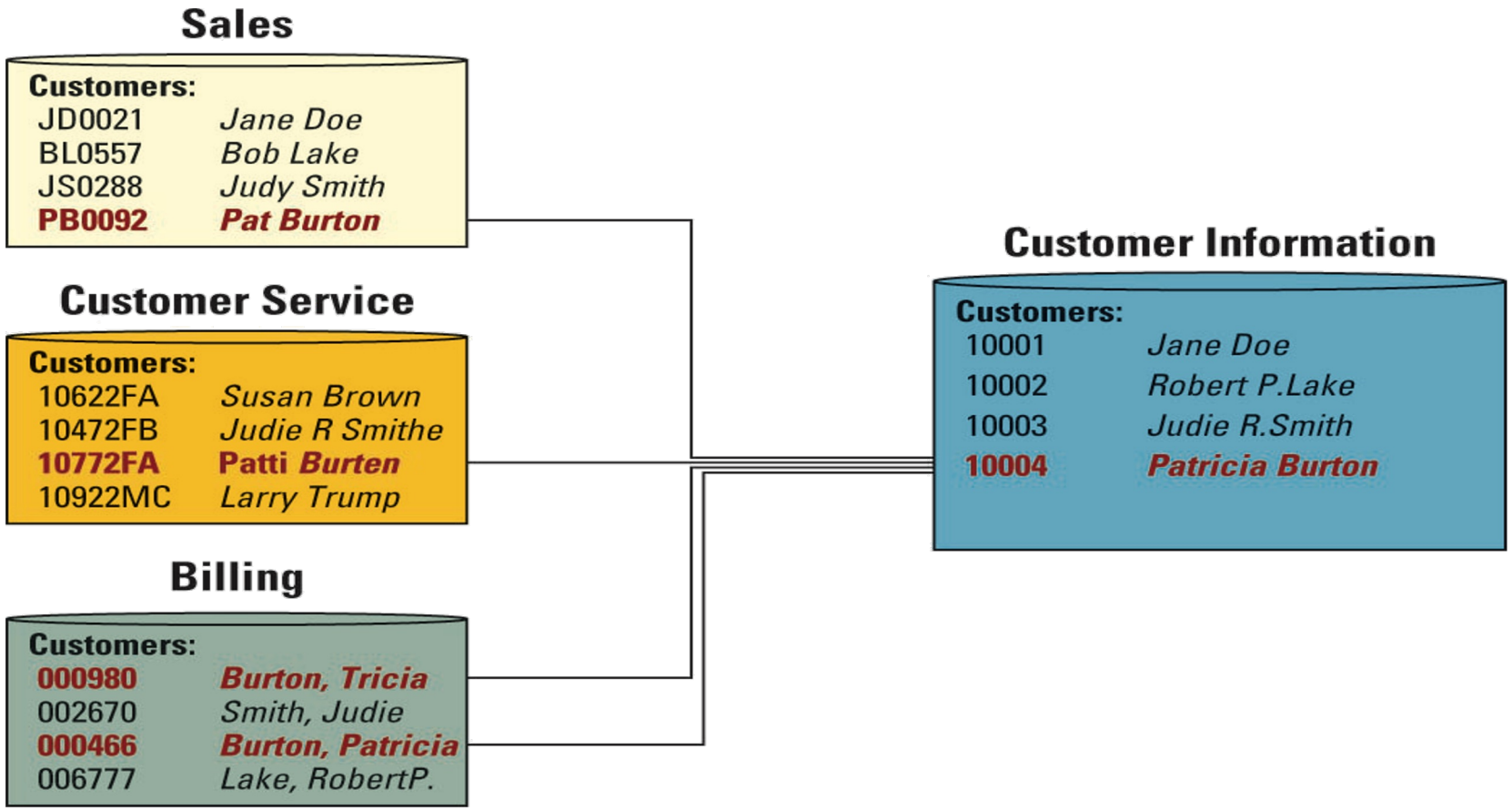


Figure 4.10

Information Cleansing or Scrubbing

4.4

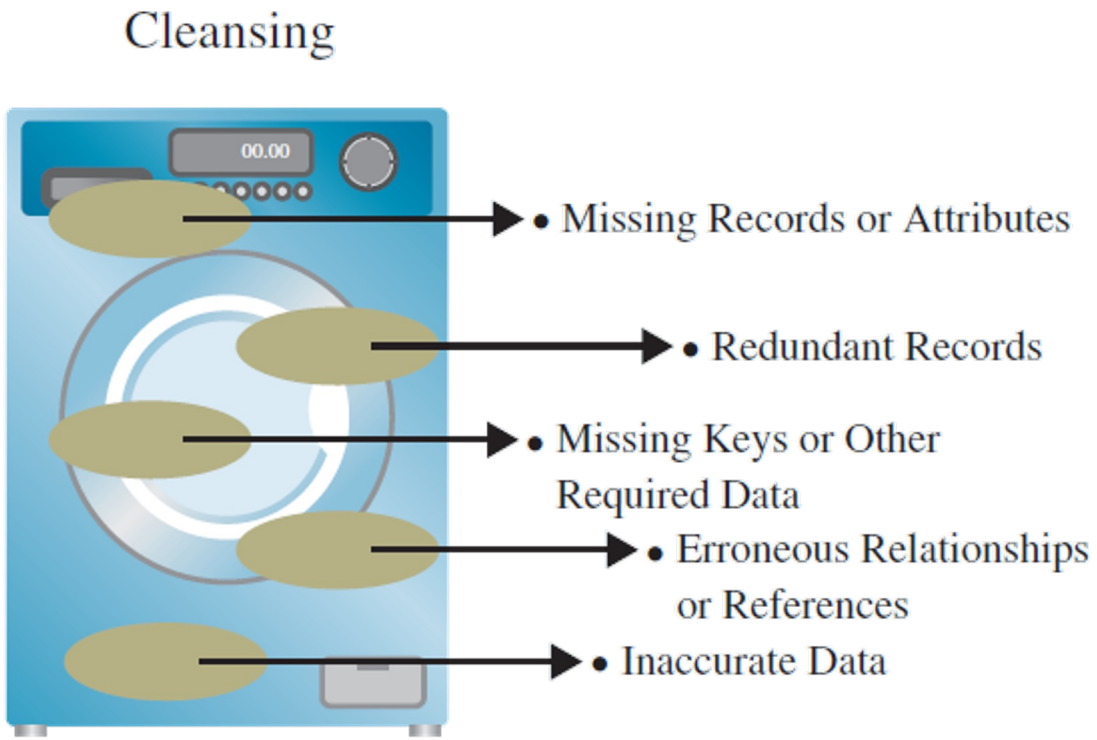


Figure 4.11

Accurate and Complete Information

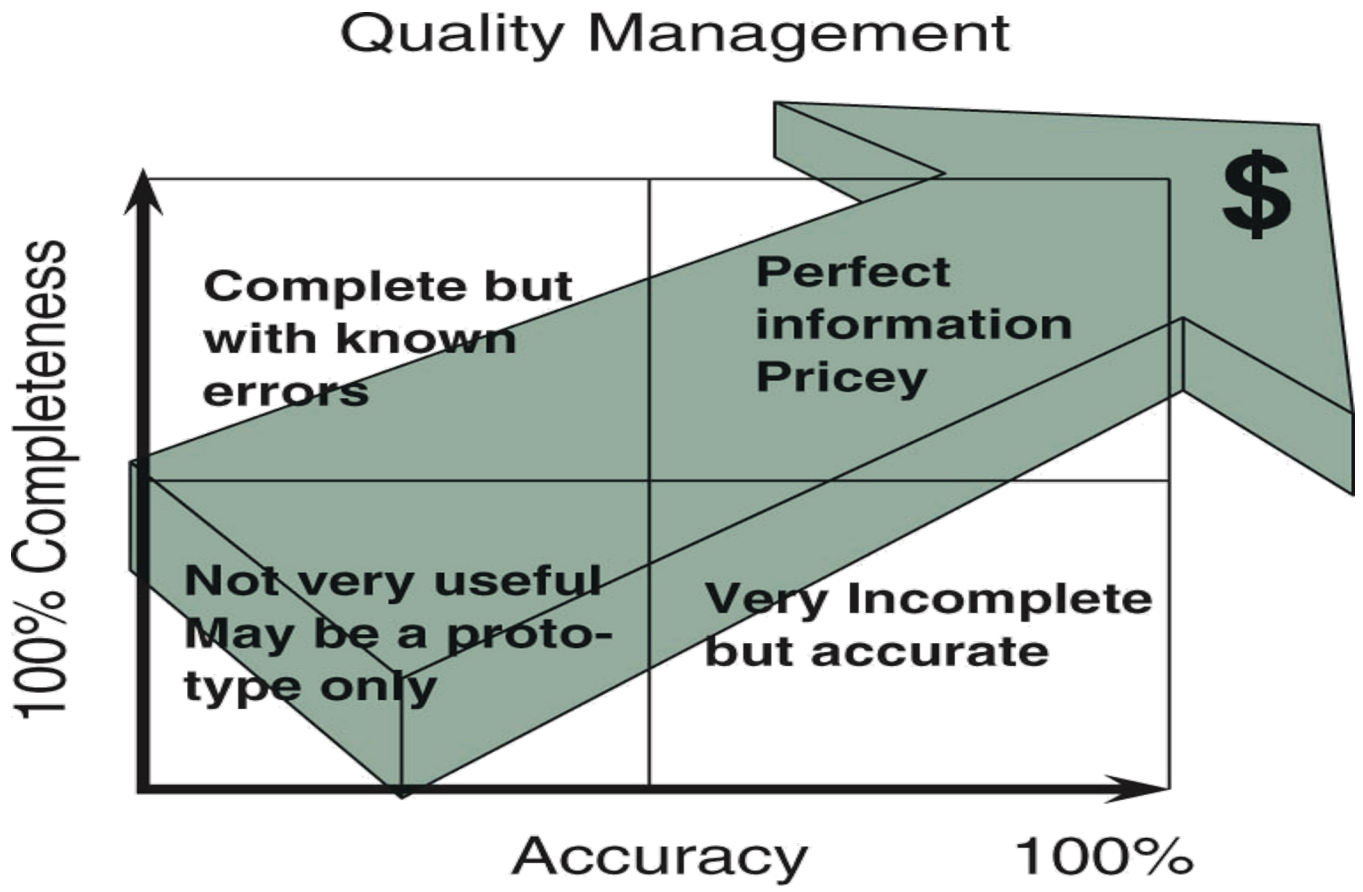


Figure 4.12

Data Quality

- Drawing conclusions from wrong data will, most likely, lead to wrong outcomes
- What can we do about it?
- Do we care about data quality as Data Miners or Data Analysts?

Data Quality Dimensions*

Accuracy

- We have accuracy when data reflects reality.
- For example, this can refer to correct names, addresses or represent factual and up to date data

Completeness

- Data is considered complete when all the data required for a particular use is present and available to be used.
- It's not about ensuring 100% of your data fields are complete. It's about determining what data is critical and what is optional.
- Consider patient records consisting of personal details and medical history. Missing information on allergies is a serious data quality problem

* https://www.gov.uk/government/news/meet-the-data-quality-dimensions

Data Quality Dimensions*

Uniqueness

- Uniqueness measures the number of duplicates.
- Data is unique if it appears only once in a data set. A record can be a duplicate even if it has some fields that are different.
  - For example, two patient records may have different addresses and contact numbers, but if they both refer to the same patient there is duplication

Consistency

- Consistency is achieved when data values do not conflict with other values within a record or across different data sets.
- For example, the first characters in a postcode should correspond to the locality of the address.
- Similarly, date of birth for the same person in two different data sets should be the same.
- Consistent data improves the ability to link data from multiple sources

* https://www.gov.uk/government/news/meet-the-data-quality-dimensions

Data Quality Dimensions*

Timeliness

- Timeliness indicates whether the data is available when expected and needed.
- Timeliness means different things for different uses.
  - In a hospital setting, timeliness is critical in ensuring the most up to date data in a bed allocation system.
  - However, it may be acceptable to use previous quarterly figures from healthcare records to forecast care needs and plan health services.
  - Data quality may diminish over time. For example, someone might provide the correct address or job title when the data is captured, but if the same individual changes their address or job these data items will become outdated.

Validity

- Validity is defined as the extent to which the data conforms to the expected format, type, and range.
- For example, an email address must have an '@' symbol
- Postcodes are valid if they appear in the Royal Mail postcode list
- Month should be between one and twelve.
- Having valid data means that it can be used with other sources.

* https://www.gov.uk/government/news/meet-the-data-quality-dimensions