Data Processing
Data wrangling aka data munging

1. Objectives

- Data wrangling software is a very critical step in the data processing
- Data wrangling involves getting the data into structured form
- Data extraction, cleaning, and organization are the most time-consuming process and they take about 50-80% of the total data science project time.
- It is the process of removing errors and combining complex data sets to make them more accessible and easier to analyze.
- It also consists of reorganizing, transforming, and mapping data from one "raw" form into more usable and formatted data for analysis
- Incomplete:
    - Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data:
        - e.g., occupation=“”
- Noisy:
    - Containing errors or outliers
        - e.g., Salary=“-10”
- Inconsistent:
    - Containing discrepancies in codes or names
        - e.g., Age=“42” Birthday=“03/07/1997”
        - e.g., Was rating “1,2,3”, now rating “A, B, C”
        - e.g., discrepancy between duplicate records

2. Why Is Data Dirty?

- Incomplete data comes from
    - n/a data value when collected

    - Different consideration between the time when the data was collected and when it is analyzed.
    - Human/hardware/software problems

- Noisy data comes from the process of data
    - Collection
    - Entry
    - Transmission

- Inconsistent data comes from
    - Different data sources
    - Functional dependency violation

3. Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
    - Quality decisions must be based on quality data
        - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
        - Data warehouse needs consistent integration of quality data

- Data extraction, cleaning, and transformation comprise the majority of the work of building a data warehouse. —Bill Inmon.

4. Data Wrangling Tasks

- Data cleaning
    - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

- Data integration
    - Integration of multiple databases, data cubes, or files:
        - CSV, PDF, API/Json, and HTML web scraping

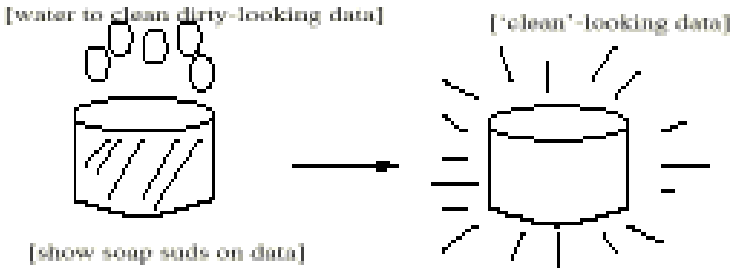
- Data transformation
    - Normalization and aggregation

- Data reduction
    - Obtains reduced representation in volume but produces the same or similar analytical results

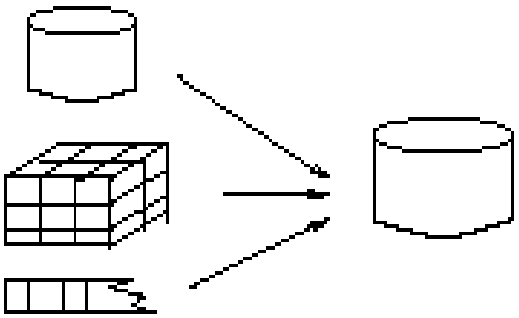
- Data discretization
    - Part of data reduction but with particular importance, especially for numerical data

5. Forms of Data Processing:

Data Cleaning



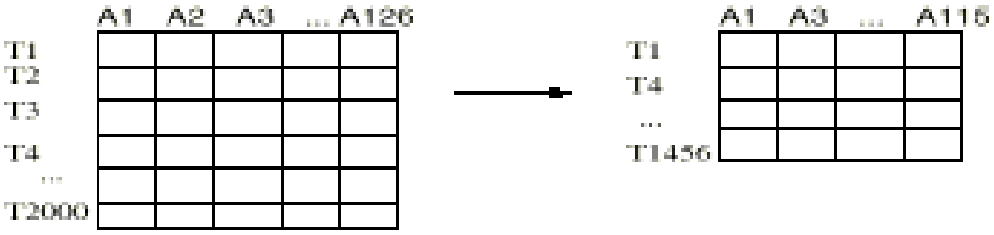
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



6. Data Cleaning

- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

7. Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Not register history or changes of the data
- Missing data may need to be inferred.

- How to Handle Missing Data?

  - Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not

effective when the percentage of missing values per attribute varies considerably.
    - Fill in the missing value manually: tedious + infeasible?
    - Fill in it automatically with
        - A global constant: e.g., “unknown”, a new class?!
        - the attribute mean
        - the attribute mean for all samples belonging to the same class: smarter
        - the most probable value: inference-based such as Bayesian formula or decision tree

8. Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
    - faulty data collection instruments
    - data entry problems
    - data transmission problems
    - technology limitation
    - inconsistency in naming convention
- Other data problems which requires data cleaning
    - duplicate records
    - incomplete data
    - inconsistent data

- How to Handle Noisy Data?

    - Binning method:
        - first sort data and partition into (equi-depth) bins

        - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
    - Clustering
        - detect and remove outliers
    - Combined computer and human inspection
        - detect suspicious values and check by human (e.g., deal with possible outliers)
    - Regression
        - smooth by fitting the data into regression functions

9. Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
    - Divides the range into *N* intervals of equal size: uniform grid
    - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
    - The most straightforward, but outliers may dominate presentation
    - Skewed data is not handled well.

- Equal-depth (frequency) partitioning:
    - Divides the range into *N* intervals, each containing approximately same number of samples
    - Good data scaling
    - Managing categorical attributes can be tricky.

- Binning methods

    - They smooth a sorted data value by consulting its "neighborhood", that is the values around it.

- The sorted values are partitioned into a number of buckets or bins.

- **Smoothing by bin means**: Each value in the bin is replaced by the mean value of the bin.

- **Smoothing by bin medians**: Each value in the bin is replaced by the bin median.

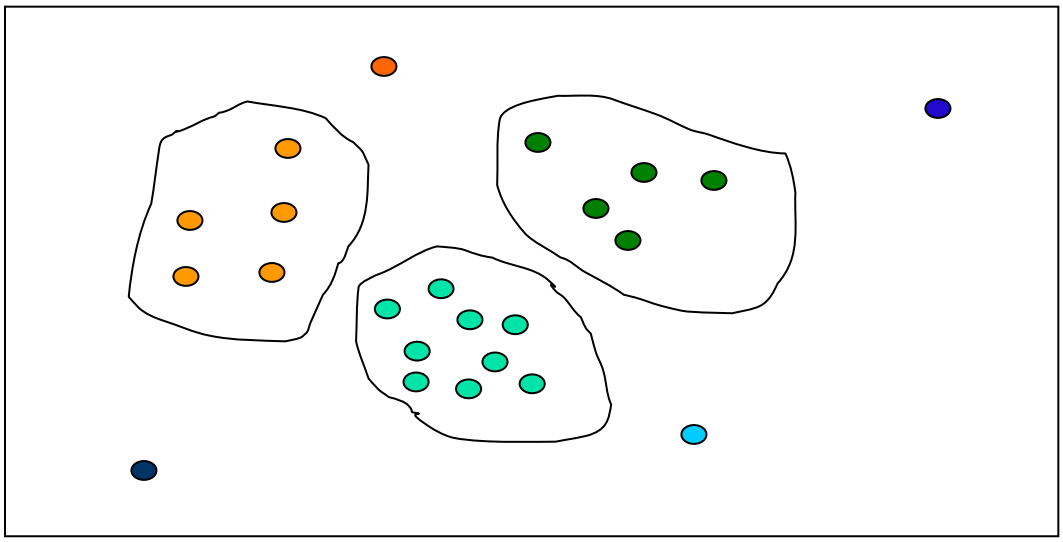
- **Smoothing by boundaries**: The min and max values of a bin are identified as the bin boundaries.

- Each bin value is replaced by the closest boundary value.

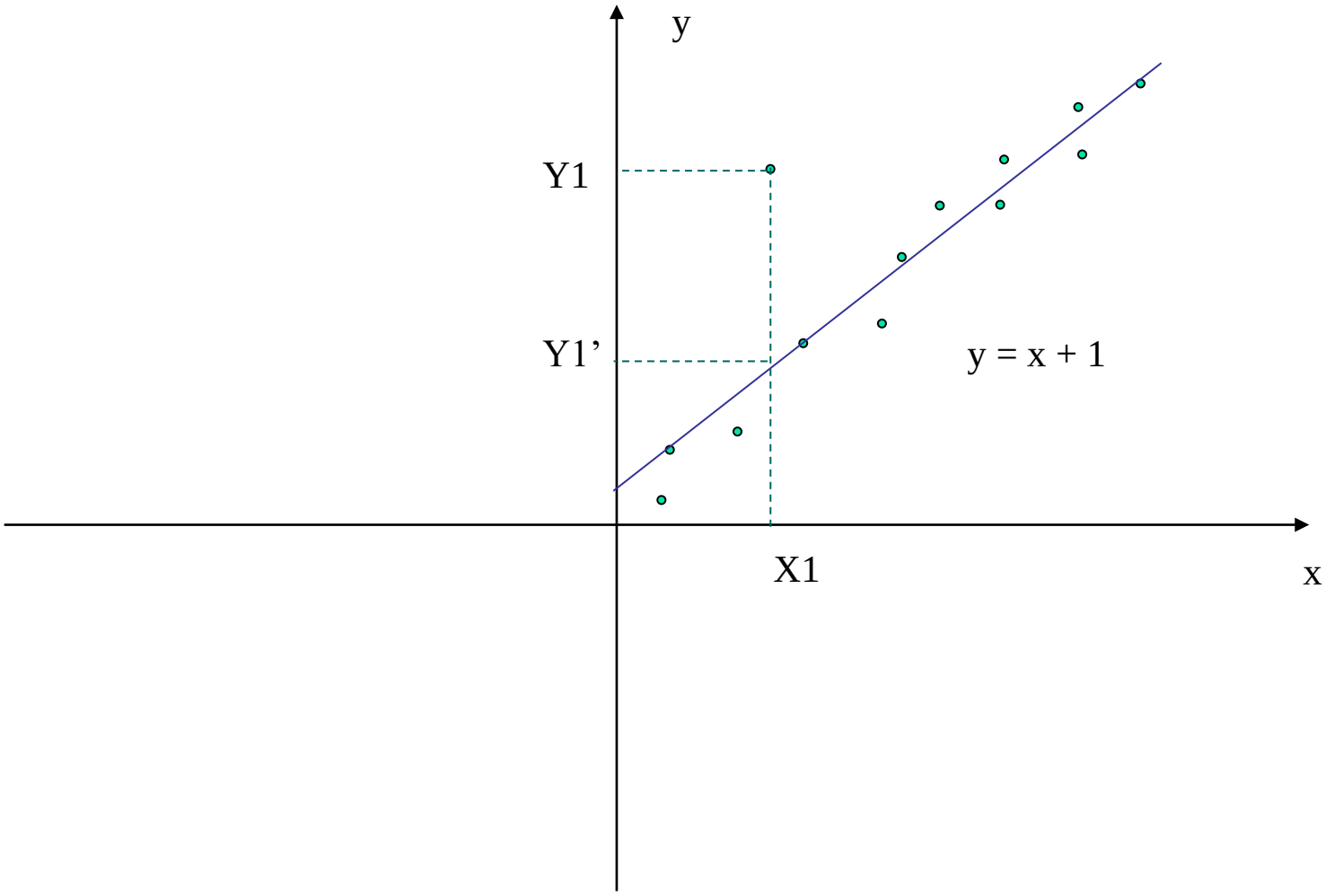
- Example: Binning Methods for Data Smoothing

  - Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
  - Partition into (equi-depth) bins:
    - **Bin 1**: 4, 8, 9, 15
    - **Bin 2**: 21, 21, 24, 25
    - **Bin 3**: 26, 28, 29, 34
  - Smoothing by bin means:
    - **Bin 1**: 9, 9, 9, 9
    - **Bin 2**: 23, 23, 23, 23
    - **Bin 3**: 29, 29, 29, 29
  - Smoothing by bin boundaries:
    - **Bin 1**: 4, 4, 4, 15
    - **Bin 2**: 21, 21, 25, 25
    - **Bin 3**: 26, 26, 26, 34

10. Cluster Analysis



11. Regression



12. Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration
  - Integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id º B.cust-#
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

- Handling Redundancy in Data Integration

  - Redundant data occur often when integration of multiple databases
    - The same attribute may have different names in different databases
    - One attribute may be a “derived” attribute in another table, e.g., annual revenue
  - Redundant data may be able to be detected by correlational analysis
  - Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

13. Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization:

$$v' = \frac{v - minA}{maxA - minA}(new\_maxA - new\_minA) + new\_minA$$

  - z-score normalization:

$$v' = \frac{v - meanA}{stand\_devA}$$

  - normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|)<1

- Attribute/feature construction
  - New attributes constructed from the given ones

14. Data reduction Strategies

- A data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation
  - Dimensionality reduction—remove unimportant attributes
  - Data Compression
  - Numerosity reduction—fit data into models
  - Discretization and concept hierarchy generation

15. Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

15.1. Similarity/Dissimilarity for Simple Attributes

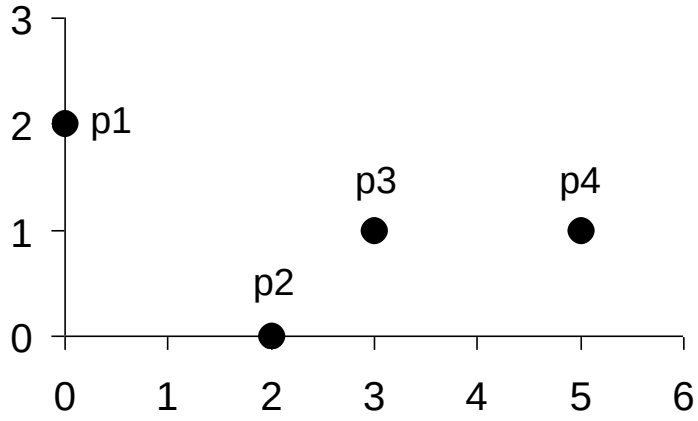
- p and q are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

15.2. Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

15.3. Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance:

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

- **Minkowski Distance: Examples**

  - $r = 1$. City block (Manhattan, taxicab, L1 norm) distance.
    - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

  - $r = 2$. Euclidean distance

  - $r \to \infty$. "supremum" (Lmax norm, L∞ norm) distance
    - This is the maximum difference between any component of the vectors

  - Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

15.4. Mahalanobis Distance

$$mahalanobis(p,q) = (p-q)\sum^{-1}(p-q)^T$$

Where

Σ is the covariance matrix of the input data *X*

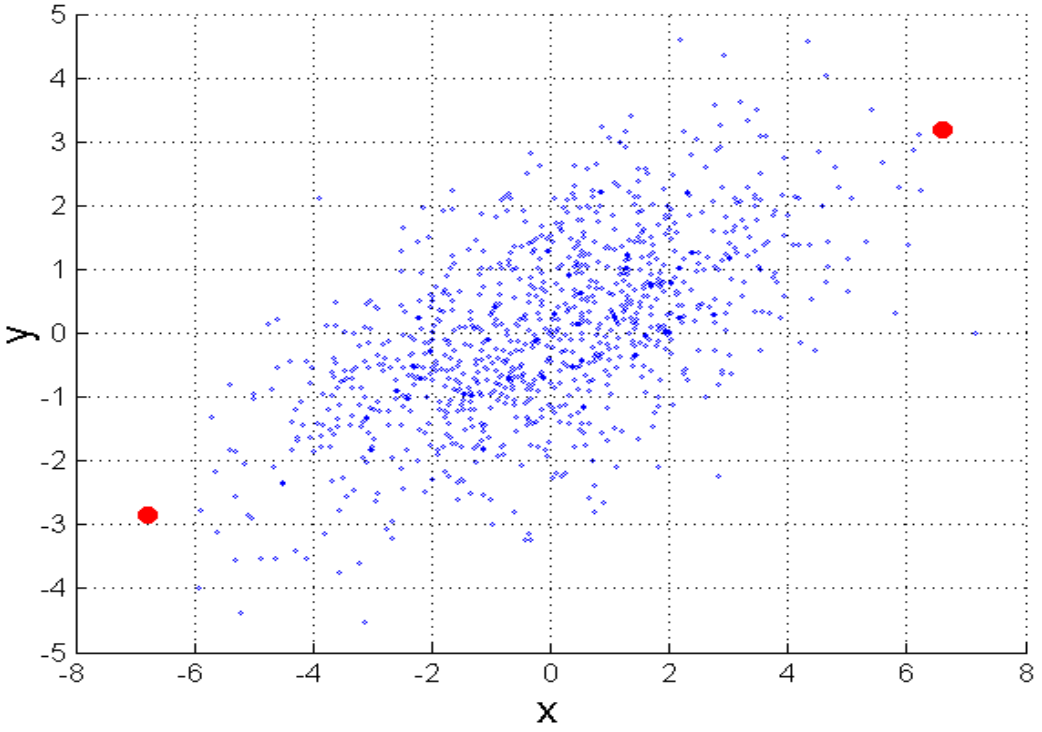
If X is a column vector with n scalar random variable components, and μk is the expected value of the kth element of X, i.e., μk = E(Xk), then the covariance matrix is defined as:

∑ = E[(X-E[X]) (X-E[X])T] =

$$\sum = E[(X - E[X])\,(X - E[X])^T]$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \dots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \dots & \\ \dots & \dots & \dots & \dots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & \dots & \dots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

The (i,j) element is the covariance between X_i and X_j.



For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

- If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called the normalized Euclidean distance:

15.5. Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

    1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
    2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
    3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

  where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a metric

15.6. Common Properties of a Similarity

- Similarities, also have some well known properties.

    1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.
    2. $s(p, q) = s(q, p)$ for all p and q. (Symmetry)

    where $s(p, q)$ is the similarity between points (data objects), p and q.

15.7. Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities
  M01 = the number of attributes where p was 0 and q was 1
  M10 = the number of attributes where p was 1 and q was 0
  M00 = the number of attributes where p was 0 and q was 0
  M11 = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes
  = (M11 + M00) / (M01 + M10 + M11 + M00)

  J = number of 11 matches / number of not-both-zero attributes values
  = (M11) / (M01 + M10 + M11)

- **SMC versus Jaccard: Example**

  p = 1 0 0 0 0 0 0 0 0 0
  q = 0 0 0 0 0 0 1 0 0 1

M01 = 2 (the number of attributes where p was 0 and q was 1)
M10 = 1 (the number of attributes where p was 1 and q was 0)
M00 = 7 (the number of attributes where p was 0 and q was 0)
M11 = 0 (the number of attributes where p was 1 and q was 1)

SMC = (M11 + M00)/(M01 + M10 + M11 + M00) = (0+7) / (2+1+0+7) = 0.7

J = (M11) / (M01 + M10 + M11) = 0 / (2 + 1 + 0) = 0

15.8. Cosine Similarity

- If *d1* and *d2* are two document vectors, then
  cos(*d1, d2*) = (*d1* • *d2*) / ||*d1*|| ||*d2*|| ,

  Where • indicates vector dot product and || *d* || is the length of vector *d*.

- Example:

  d1 = **3 2 0 5 0 0 0 2 0 0**
  d2 = **1 0 0 0 0 0 0 1 0 2**

d1 • d2= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

||d1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)**0.5** = (42) **0.5** = 6.481

||d2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) **0.5** = (6) **0.5** = 2.245

cos(d1, d2) = .3150

15.9. Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

15.10. Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$
$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

16. Data Wrangling Tools:

- Parsehub: https://www.parsehub.com/
- Scrapy:https://scrapy.org/
- Talend: https://www.talend.com/products/data-preparation/
- Alteryx APA Platform: https://www.alteryx.com/products/apa-platform
- Altair Monarch: https://www.altair.com/monarch/
- Microsoft Power Query: https://powerquery.microsoft.com/en-us/
- Tableau Desktop: https://www.tableau.com/products/desktop