Business Intelligence

TICS-423
Universidad Adolfo Ibáñez

Week 02: 08-13 August, 2016

Claudio Diaz
Sebastián Moreno
Gonzalo Ruz

Data and Measurement

What is data?

- Collection of entities and their attributes
- Attribute: property or characteristic of an entity (e.g., eye color, temperature)
- Entity: collection of attributes
  Aka: record, point, case, sample, object, or instance
- The values of the attributes are numbers or signs assigned to the attribute

Attributes

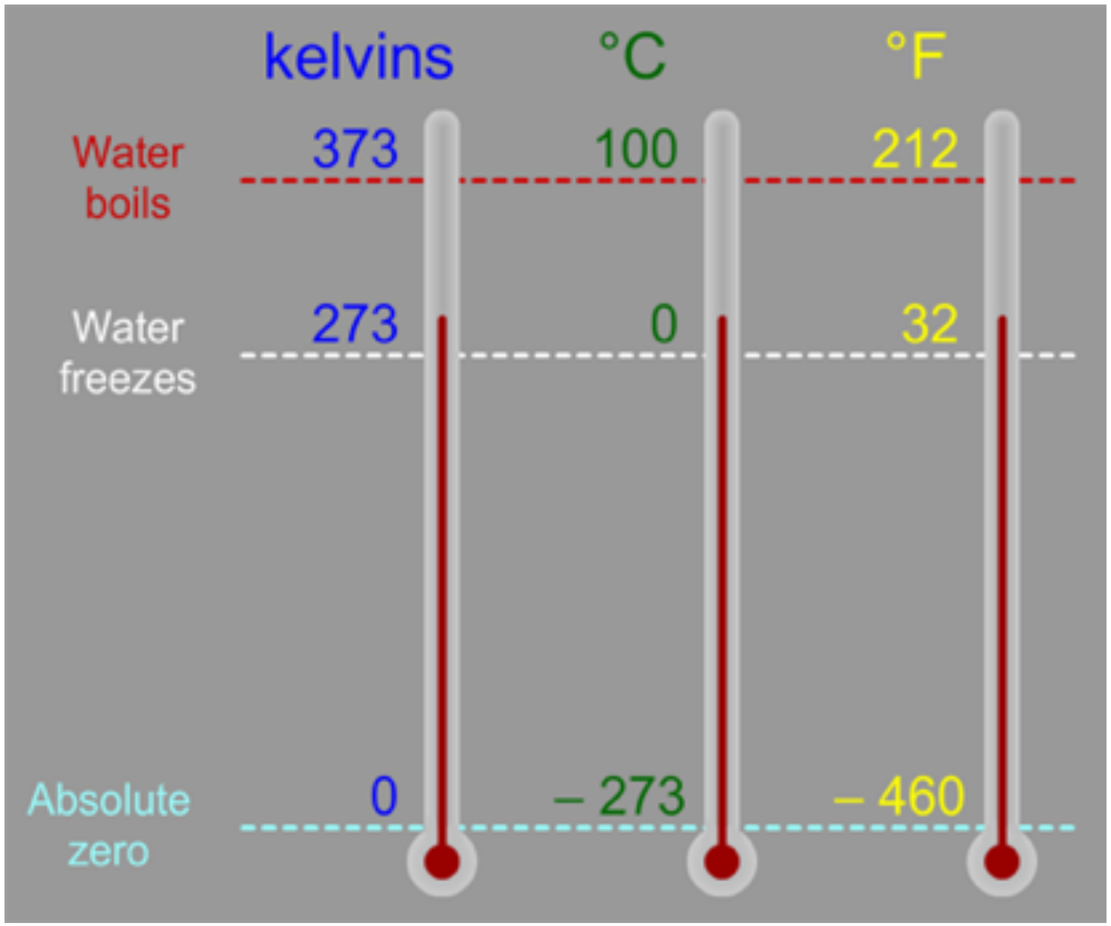
Entity

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

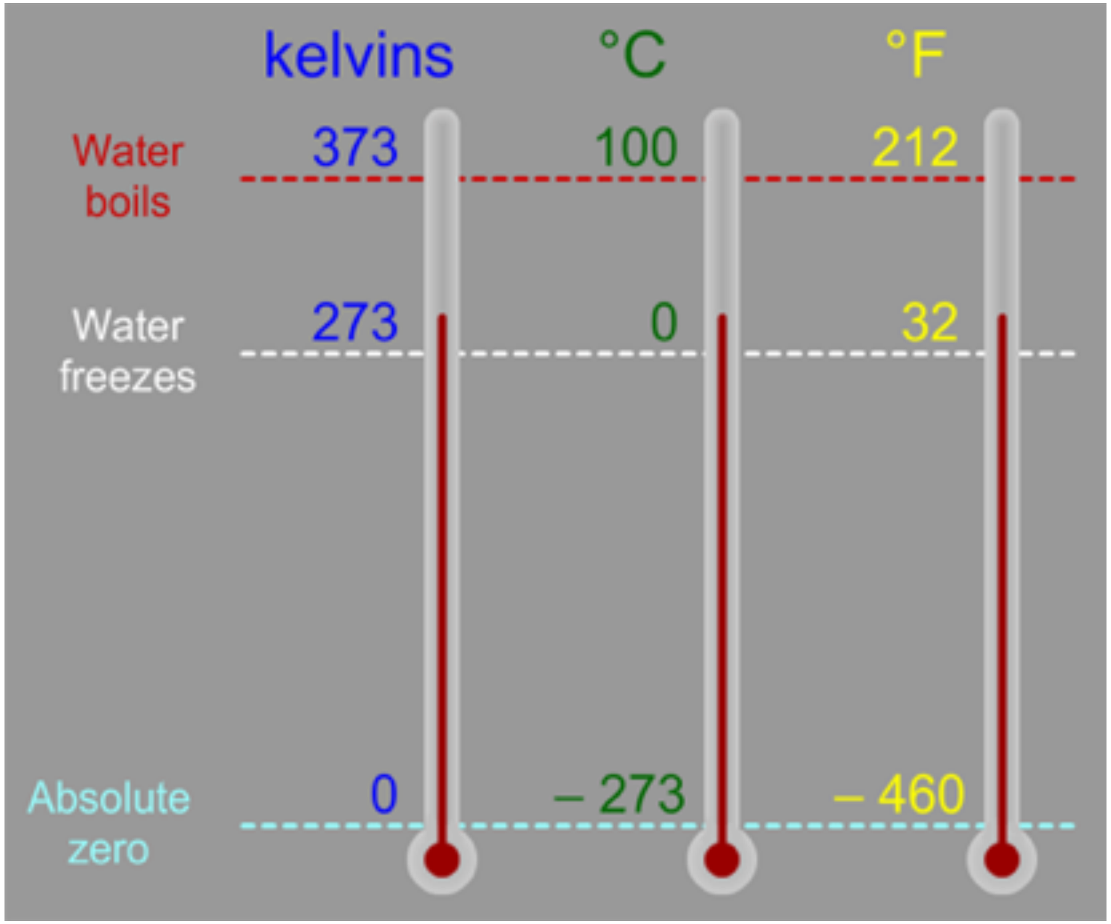
Type of measurements

- **Nominal:** Categorial values, without any order
- **Ordinal:** Ordered values, without meaningful distance between points.
- **Interval:** Ordered values, with meaningful distance between points.
- **Ratio:** Ordered values, with meaningful distance between points, and a clear definition of zero.

Type of measurements, example



Type of measurements, example



Ratio Intervals

Ordinal

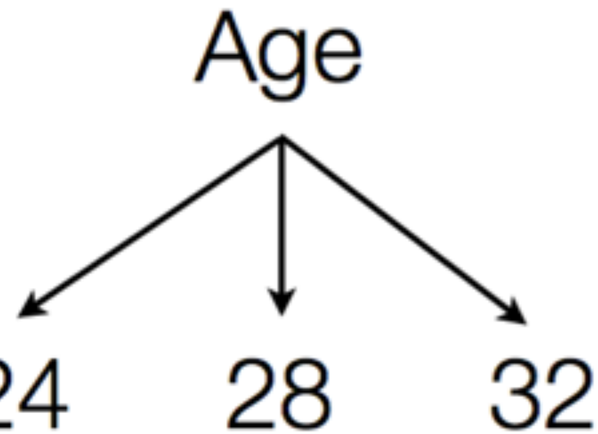
Nominal

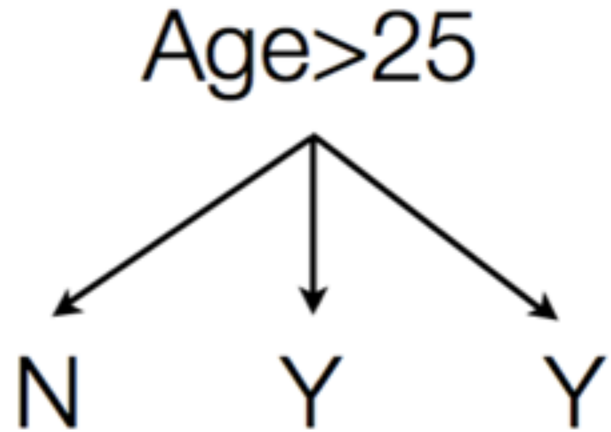
Type of measurements, attributes

- **Discrete:**
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, set of words in a collection of documents
  - Often represented as integer variables
- **Continuous**
  - Has real numbers as attribute values
  - Examples: temperature, height
  - Continuous attributes are typically represented as floating-point variables

Type of measurements, naming convention

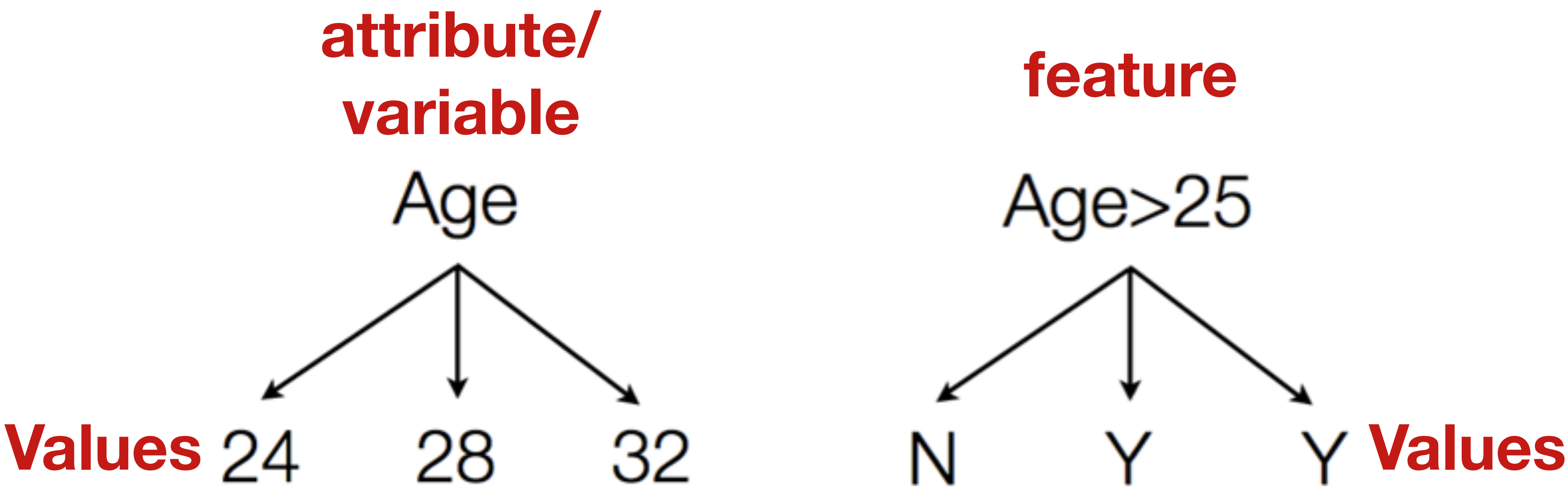
feature, attribute, or variable?

Values 



Values

Type of measurements, naming convention



Types of data

Types of data: tabular data

- Collection of records, each of which consists of a fixed set of attributes.

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Types of data: document data

- Each document is represented as a term vector, where each attribute records the number of times the term occurs in the document

	Docs																			
Terms	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
data	1	1	0	0	2	0	0	0	0	0	1	2	1	1	1	0	1	0	0	0
examples	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
introduction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
mining	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
network	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1
package	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Types of data: transaction data

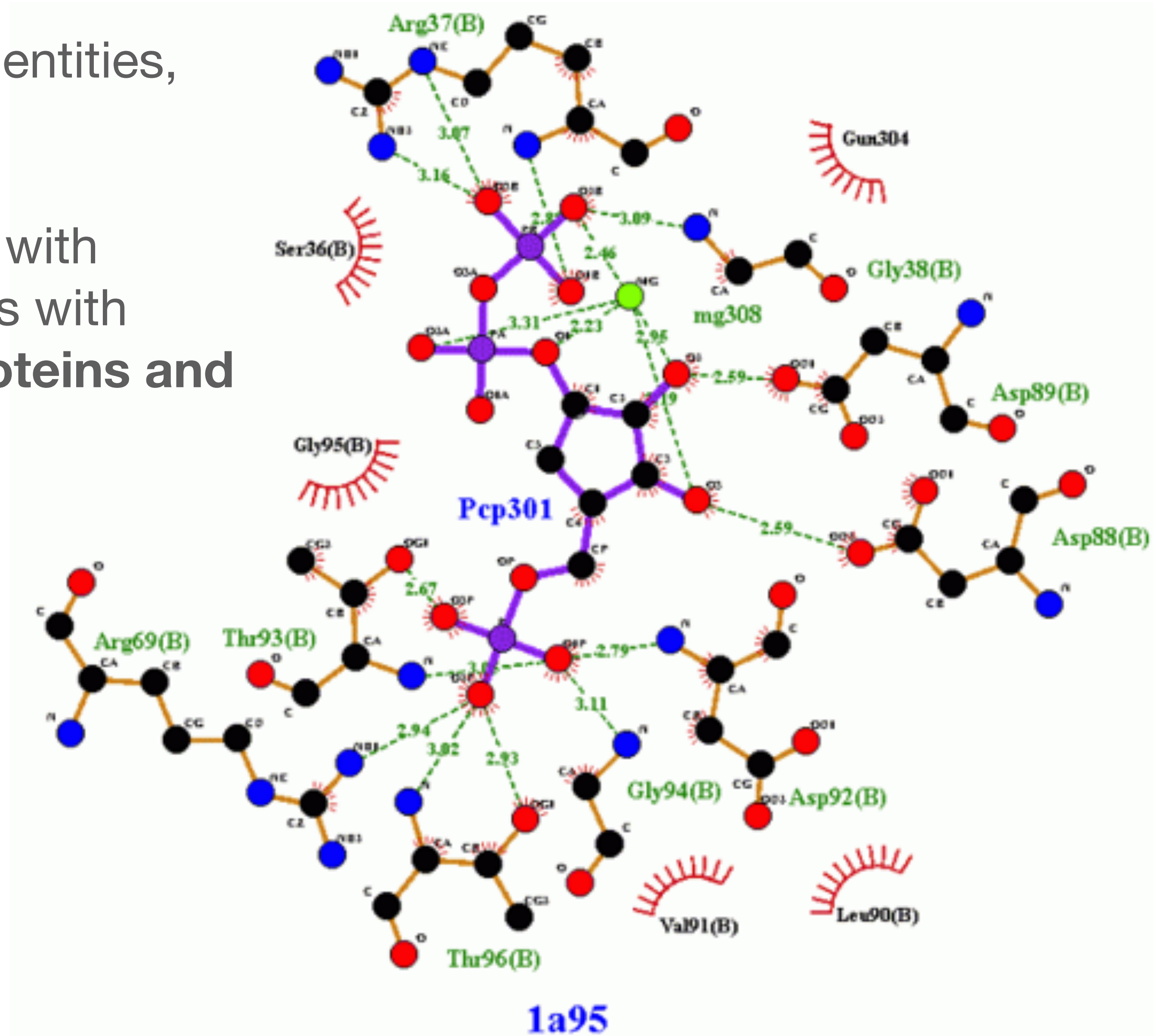
- Each record corresponds to a transaction that involves a set of items
  Example: In a grocery store purchase, the set of products purchased by a customer constitute a transaction, while the individual products that were purchased are the items

Table 6.22. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a,d,e\}$
1	0024	$\{a,b,c,e\}$
2	0012	$\{a,b,d,e\}$
2	0031	$\{a,c,d,e\}$
3	0015	$\{b,c,e\}$
3	0022	$\{b.d.e\}$
4	0029	$\{c,d\}$
4	0040	$\{a,b,c\}$
5	0033	$\{a,d,e\}$
5	0038	$\{a,b,e\}$

Types of data: graph data

- Nodes correspond to entities, edges correspond to relationships
  Example: Web graph with HTML links, molecules with atoms and bonds, **proteins and their interactions.**



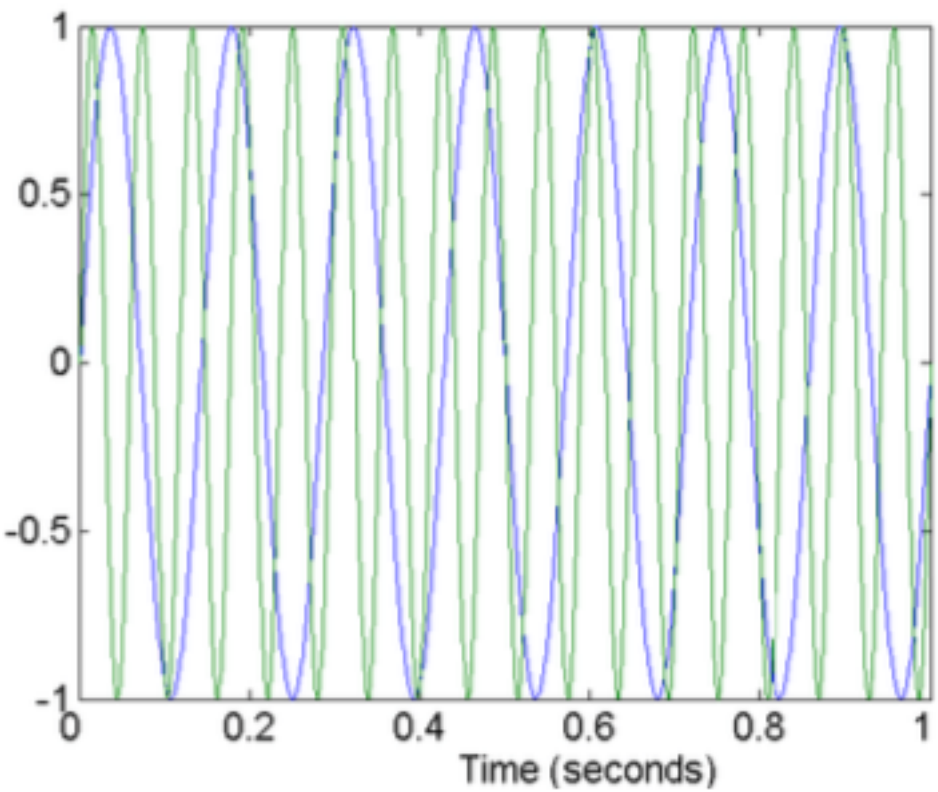
Data quality

Data quality

- Several times the collected data presents some important problems such as:
  - Noise
  - Outliers
  - Missing values
  - Duplicate data

Data quality: noise

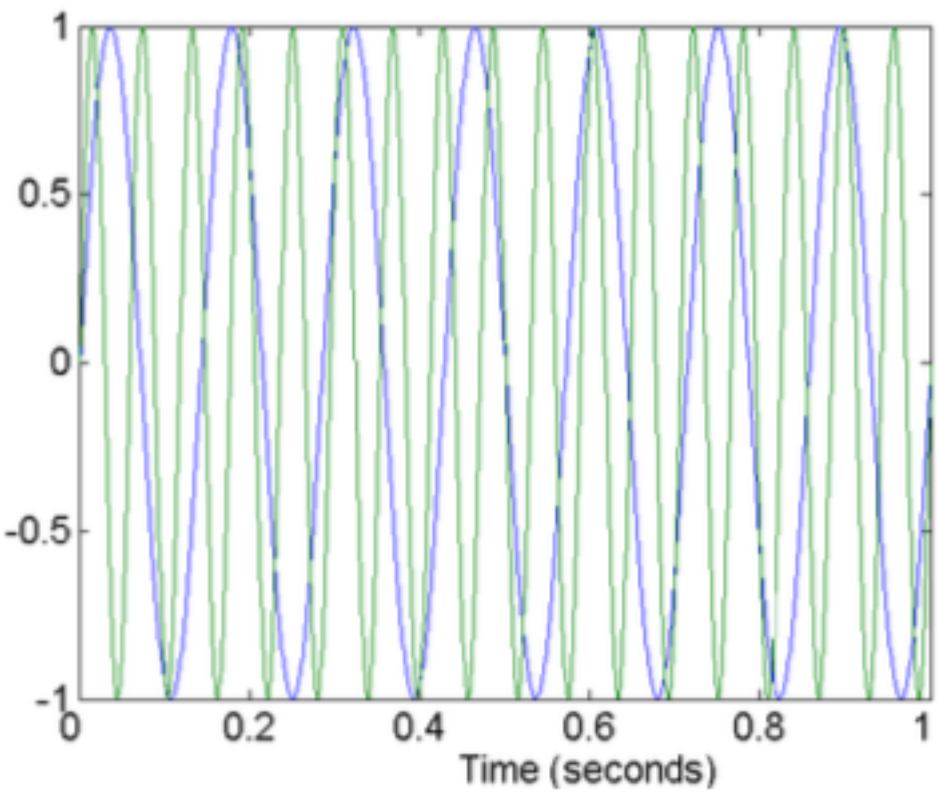
- Noise refers to measurement error in data values
  Could be random error or systematic error



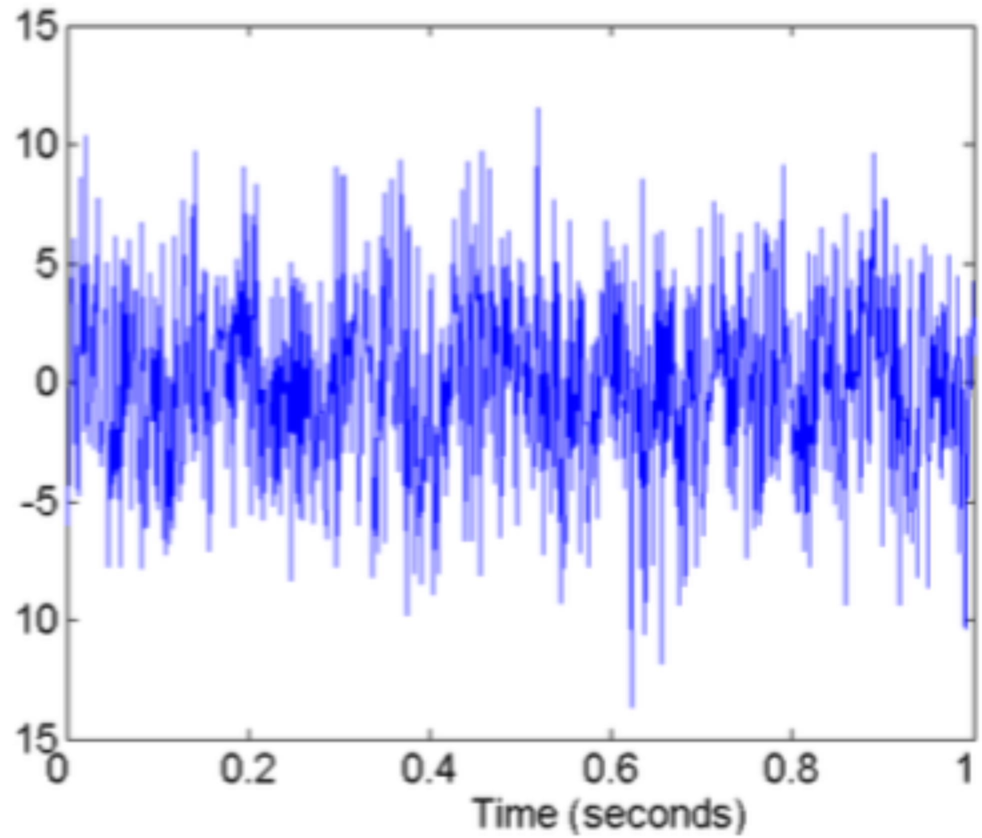
Two Sine Waves

Data quality: noise

- Noise refers to measurement error in data values
  Could be random error or systematic error



Two Sine Waves

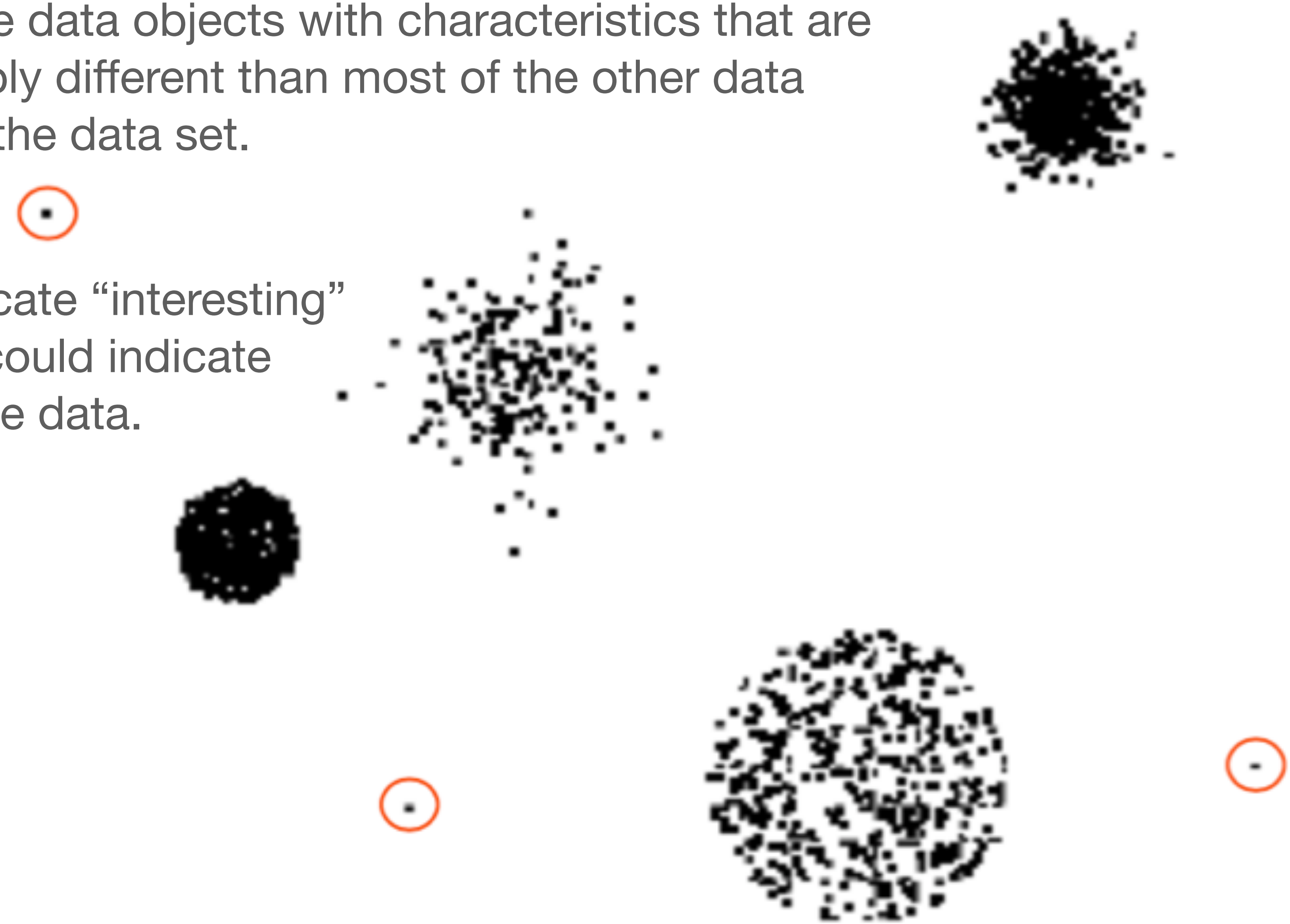


Two Sine Waves + Noise

Data quality: outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.

- Could indicate “interesting” cases, or could indicate errors in the data.



Data quality: missing values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Ways to handle missing values
  - Eliminate entities with missing values
  - Estimate attributes with missing values
  - Ignore the missing values during analysis
  - Replace with all possible values (weighted by their probabilities)
  - Impute missing values

Data quality: duplicate values

- Data set may include data entities that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
  - Example: same person with multiple email addresses
- Data cleaning
  - Finding and dealing with duplicate entities
  - Finding and correcting measurement error
  - Dealing with missing values

Data pre-processing

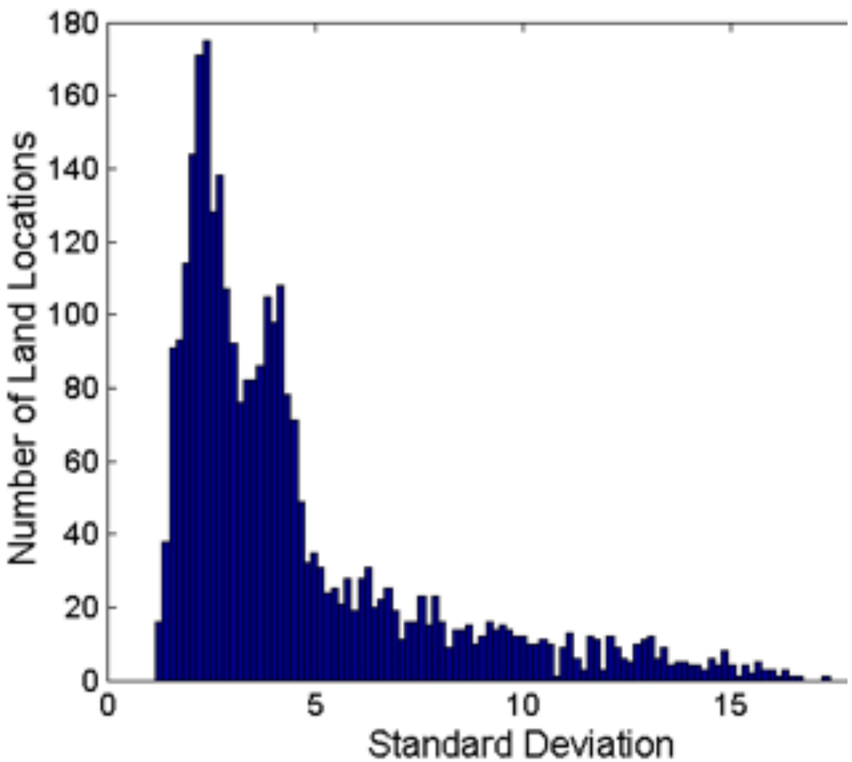
Data pre-processing

- Data pre-processing “clean” the data by eliminating corrupted, redundant, and irrelevant data.
  - Aggregation
  - Sampling
  - Dimensionality reduction
  - Feature subset selection
  - Discretization

Data pre-processing: aggregation

- Combines two or more attributes (or objects) into a single attribute (or object)

Variation of Precipitation in Australia

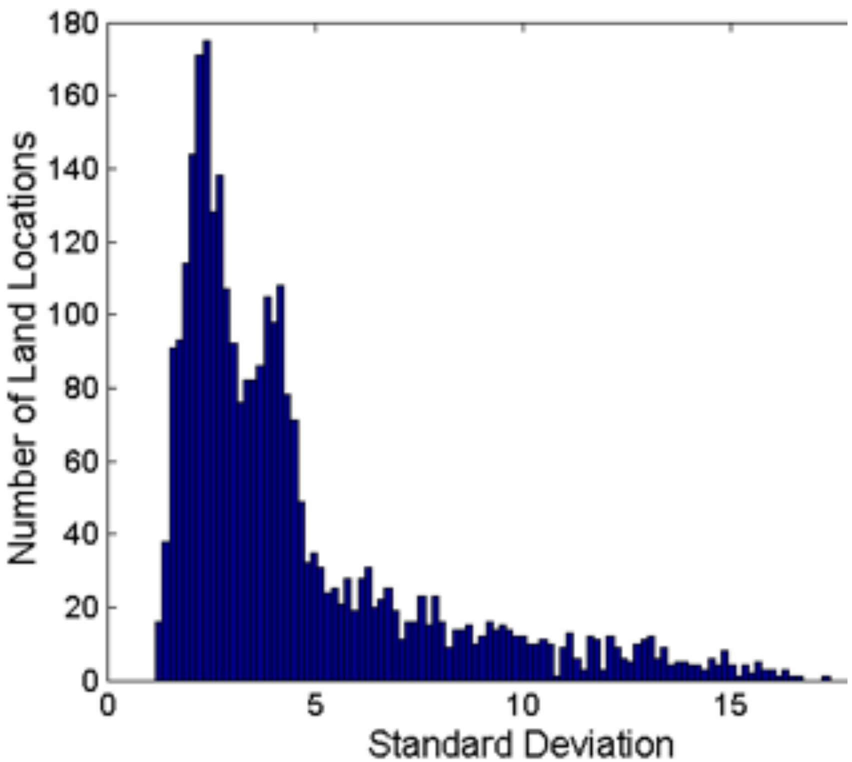


Standard Deviation of Average Monthly Precipitation

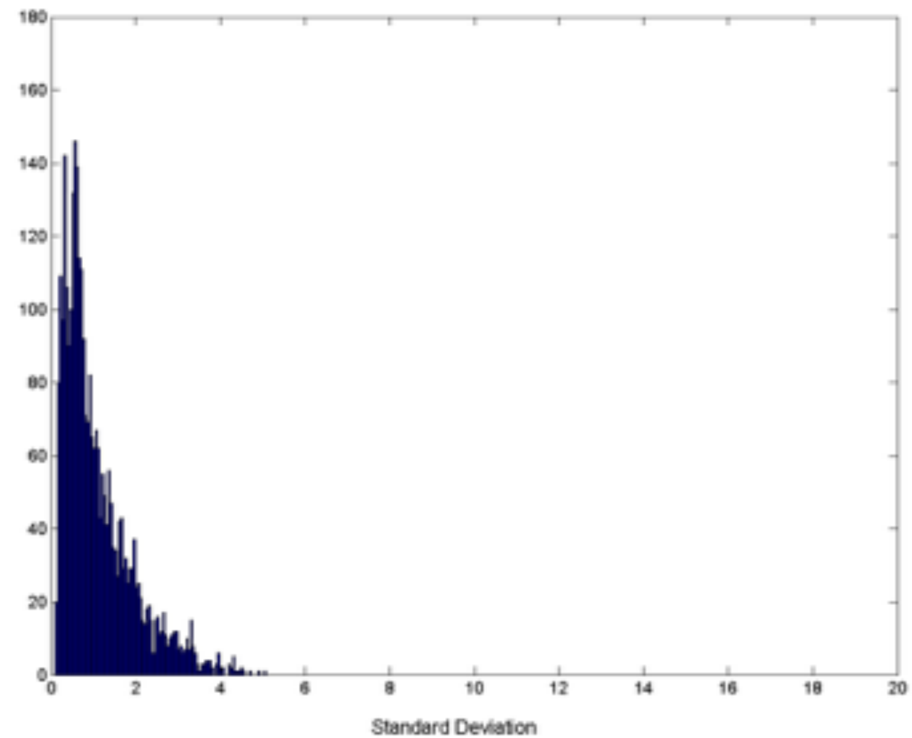
Data pre-processing: aggregation

- Combines two or more attributes (or objects) into a single attribute (or object)

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

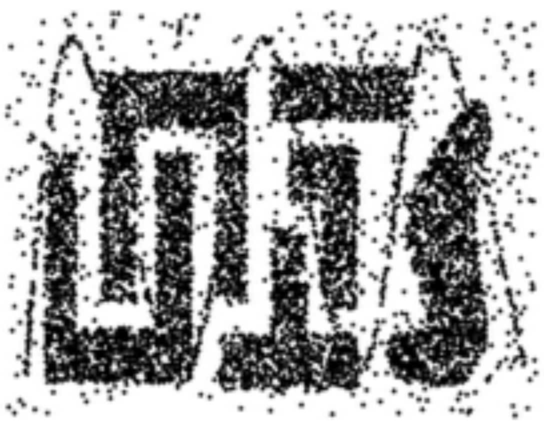
Data pre-processing: sampling

- Sampling is the main technique employed for data selection.
    - In data mining/statistics sampling is used because processing/obtaining the entire set of data of interest is too expensive or time consuming (BIG DATA).
- The sample must be representative (it has approximately the same property (of interest) as the original set of data).

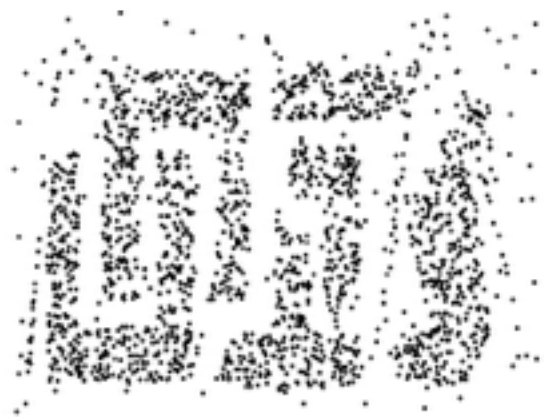
Data pre-processing: sampling

- Types of sampling:
  - Simple Random Sampling: There is an equal probability of selecting any particular item.
  - Sampling without replacement: As each item is selected, it is removed from the population
  - Sampling with replacement: Objects are not removed from the population as they are selected for the sample.
  - Stratified sampling: Split the data into several partitions; then draw random samples from each partition

Data pre-processing: sample size of the sampling



8000 points

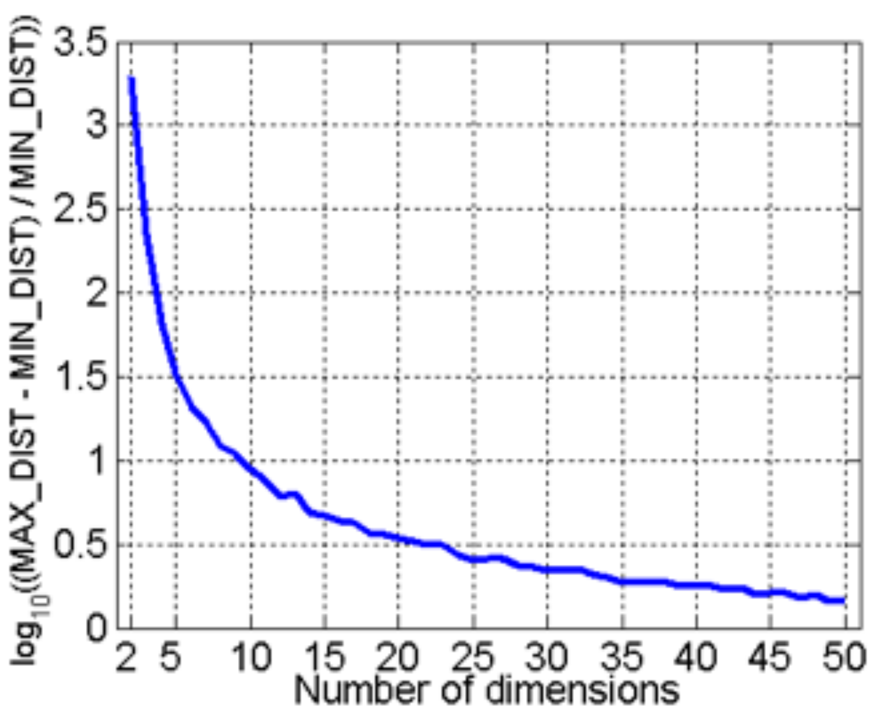


2000 Points

500 Points

Data pre-processing: curse of dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**

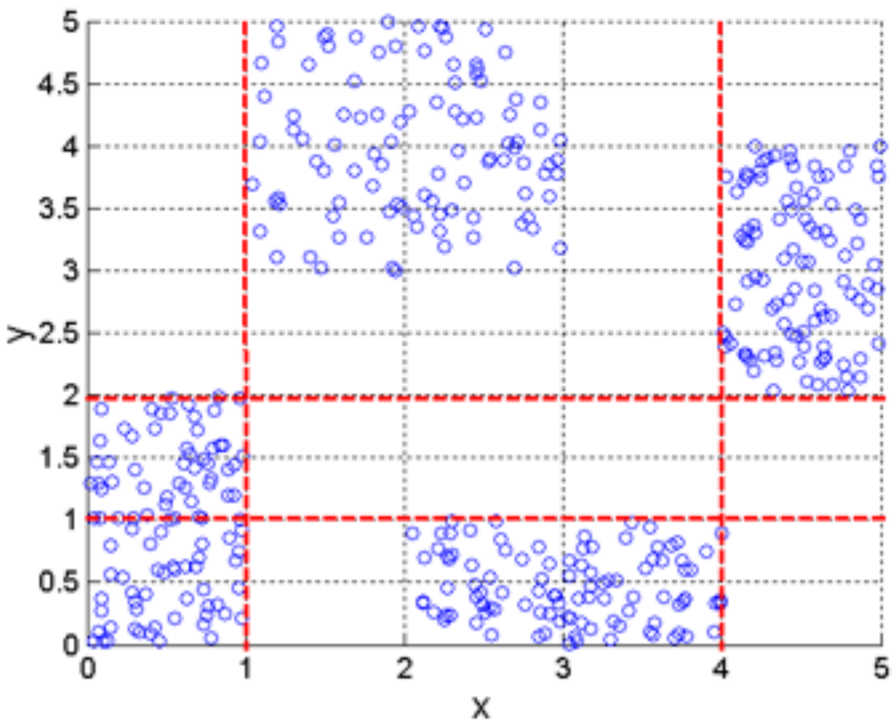
Data pre-processing: dimensionality reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualised
  - May help to eliminate irrelevant features or reduce noise
- Techniques:
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

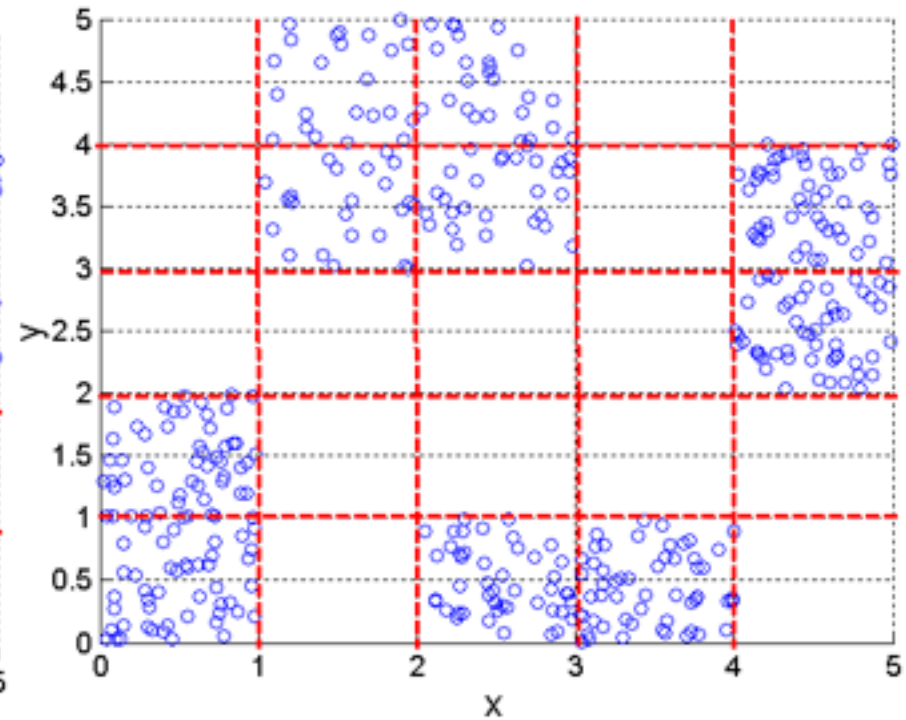
Data pre-processing: feature subset selection

- A method to reduce dimensionality of data
- Redundant features: duplicate much or all of the information contained in one or more other attributes
  Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features: Contain no information that is useful for the data mining task at hand
  Example: students ‘ID' is often irrelevant to the task of predicting students grade.

Data pre-processing: discretization



3 categories for both x and y



5 categories for both x and y

Similarity and distance

Similarity

- **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

Similarity

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

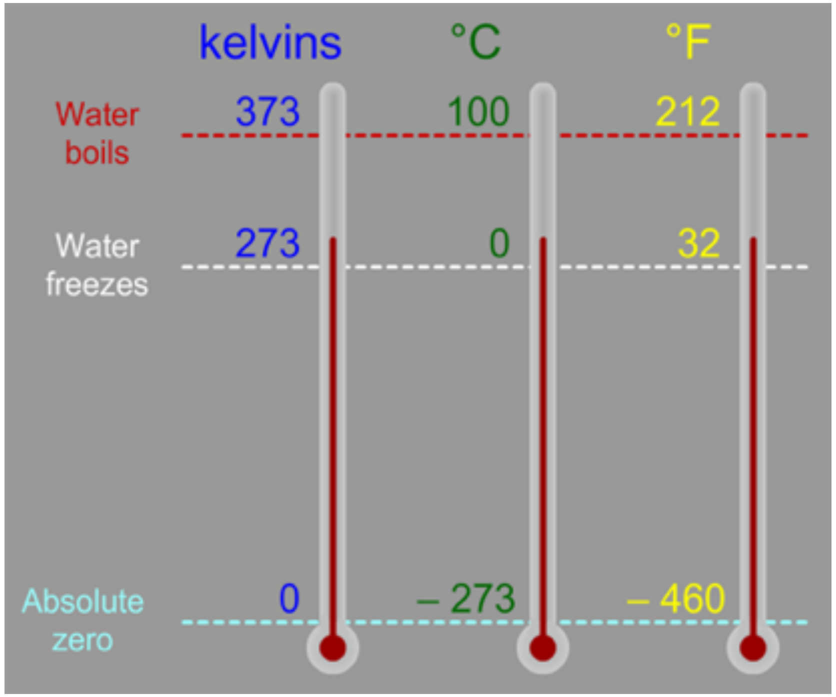
Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is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

Similarity, example

- **Nominal** S(p,q)=0

- **Ordinal** S(p,q)=1-(5-4)/(5-1)=0.75

- **Intervals** p=35 C, q=40 C
  => s(p,q) = -5
  => s(p,q) = 1/(1+5) = 0.166



Similarity between binary vectors

- Let **p** and **q** vectors with only binary attributes. To calculate the similarities between these two vectors, we use the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1

- **Simple Matching Coefficient (SMC) =** number of matches / number of attributes
  = $(M_{00}+M_{11})/(M_{00}+M_{01}+M_{10}+M_{11})$

- **Jaccard Coefficient (J) =** number of 11 matches / number of not-both-zero attributes
  = $(M_{11})/(M_{01}+M_{10}+M_{11})$

Similarity between binary vectors, example

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

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)
$M_{10} = 1$ (the number of attributes where p was 1 and q was 0)
$M_{00} = 7$ (the number of attributes where p was 0 and q was 0)
$M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

Cosine similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

where • indicates vector dot product and $||d||$ is the length of vector d.

- Example:

d_1 = 3 2 0 5 0 0 0 2 0 0
d_2 = 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 = 2.449

$\cos(d_1,d_2) = 0.3150$

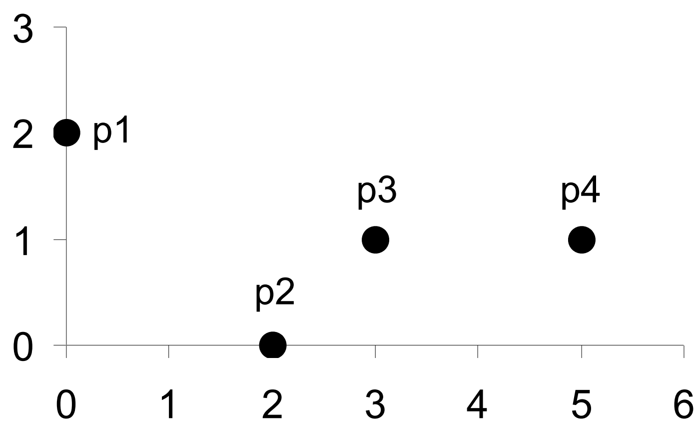
Distance

- A metric or distance function is a function that defines a distance between each pair of elements of a set.
- Given two points x and y, a metric or distance function must satisfy the following conditions
  - non negativity => $d(x,y)>0$
  - identity => $d(x,y)=0 \iff x=y$
  - symmetry => $d(x,y)=d(y,x)$
  - triangle inequality => $d(x,z)<=d(x,y)+d(y,z)$

Euclidean distance

- One of the most well known and used distance between points. Let p and q be two m dimensional vectors

$$d(p,q) = \sqrt{\sum_{k=1}^{m}(p_k - q_k)^2}$$



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

Minkowski distance

- Is a generalization of the euclidean distance.
  Let p and q be two m dimensional vectors

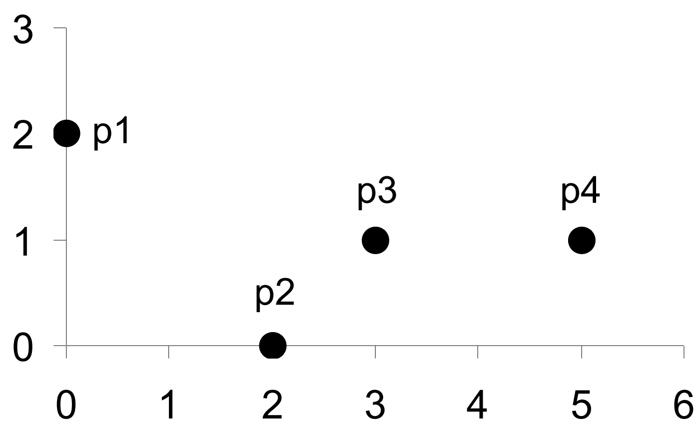
$$d(p,q) = \left(\sum_{k=1}^{m} |p_k - q_k|^r \right)^{\frac{1}{r}}$$

- For r=1 => City block (Manhattan, taxicab, L_1 norm) distance.

- For r=2 => Euclidean distance, L_2 norm.

- For $r \to \infty$ => supremum distance: the the maximum difference between any component of the vectors.

Minkowski distance, example



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

Mahalanobis distance

- It considers the variance of the data to calculate the distance.

$$d(p,q) = \sqrt{(p-q)\Sigma^{-1}(p-q)^T}$$

where ∑ is the covariance matrix of the input data.

- Example:
grade test 1: 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0
grade test 2: 1.0, 3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.4, 4.5, 7.0

  What are the Mahalanobis distances between grades 1.0 and 7.0 for each test?

Mahalanobis distance

- It considers the variance of the data to calculate the distance.

$$d(p,q) = \sqrt{(p-q)\Sigma^{-1}(p-q)^T}$$

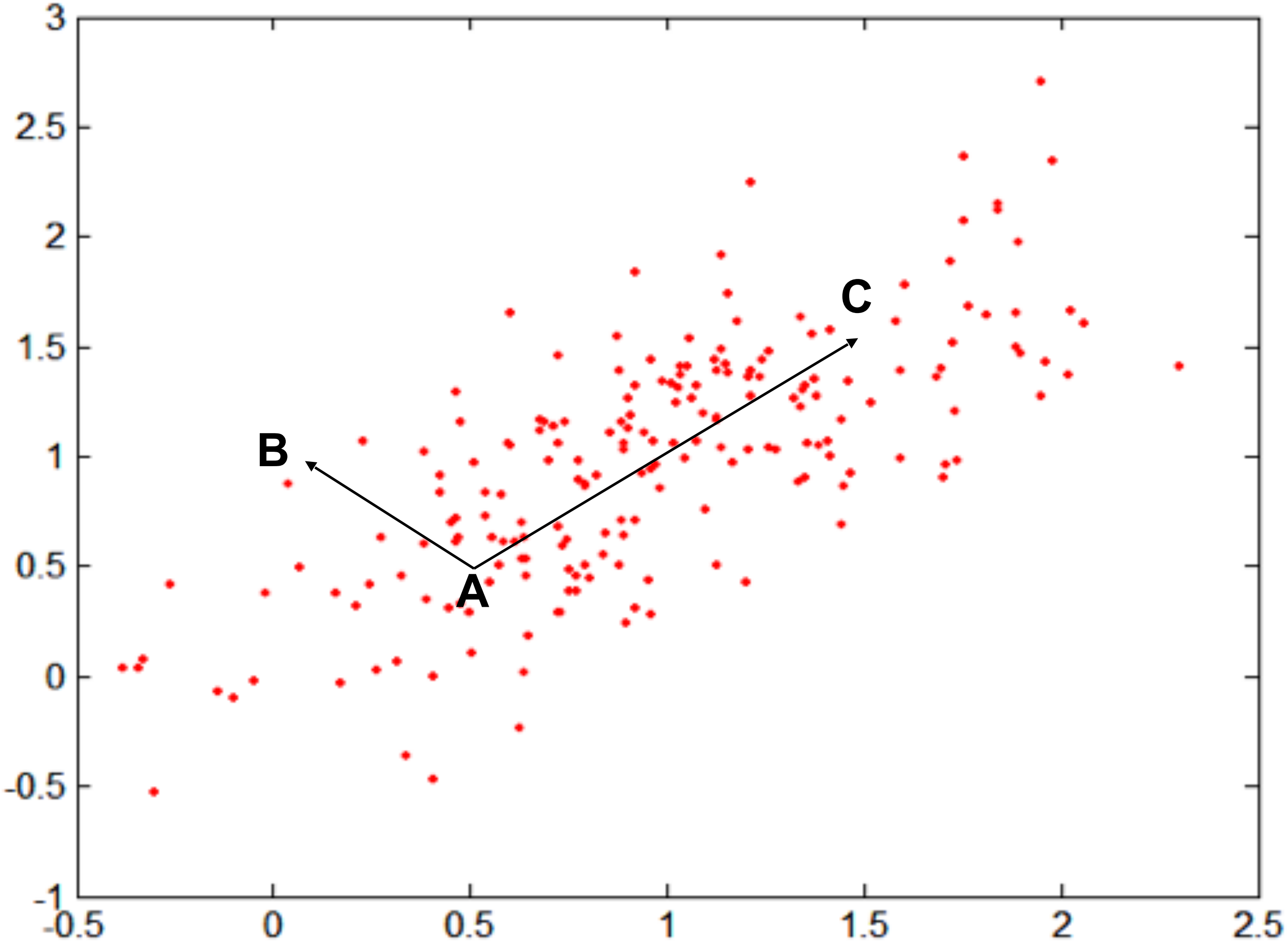
where $\sum$ is the covariance matrix of the input data.

- Example:
  grade test 1: 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0
  grade test 2: 1.0, 3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.4, 4.5, 7.0

  What are the Mahalanobis distances between grades 1.0 and 7.0 for each test?

  For test 1 => d(7.0,1.0)=3.08
  For test 2 => d(7.0,1.0)=4.76

Mahalanobis distance, example



A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Correlation

Correlation

- Correlation measures the linear relationship between attributes.

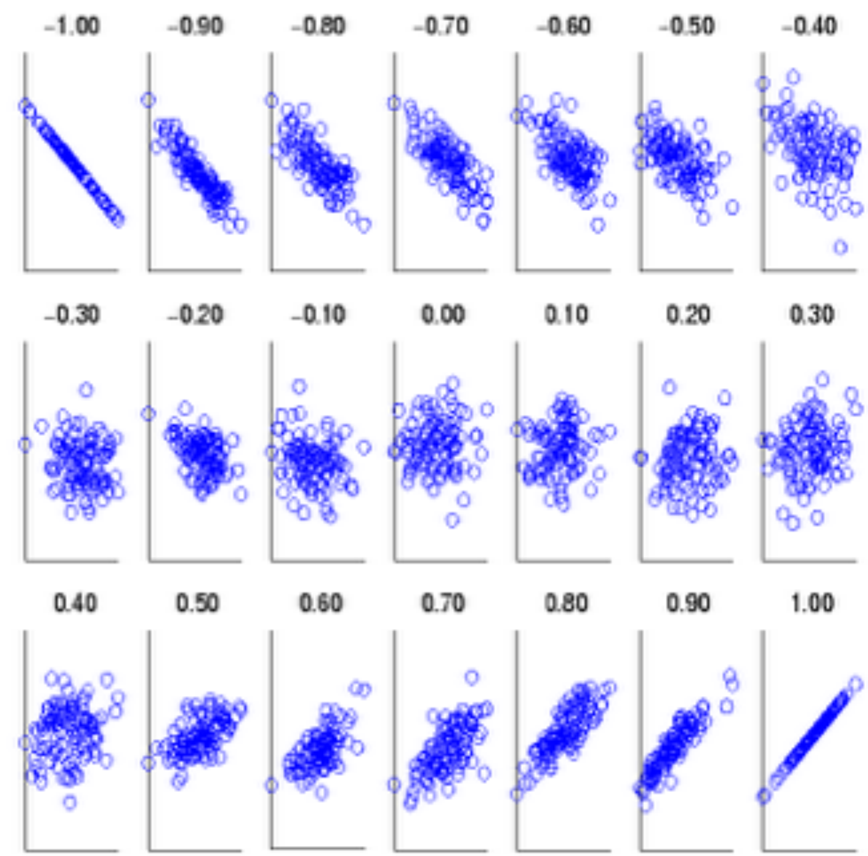
$$\rho(X,Y) = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Correlation between attributes X and Y

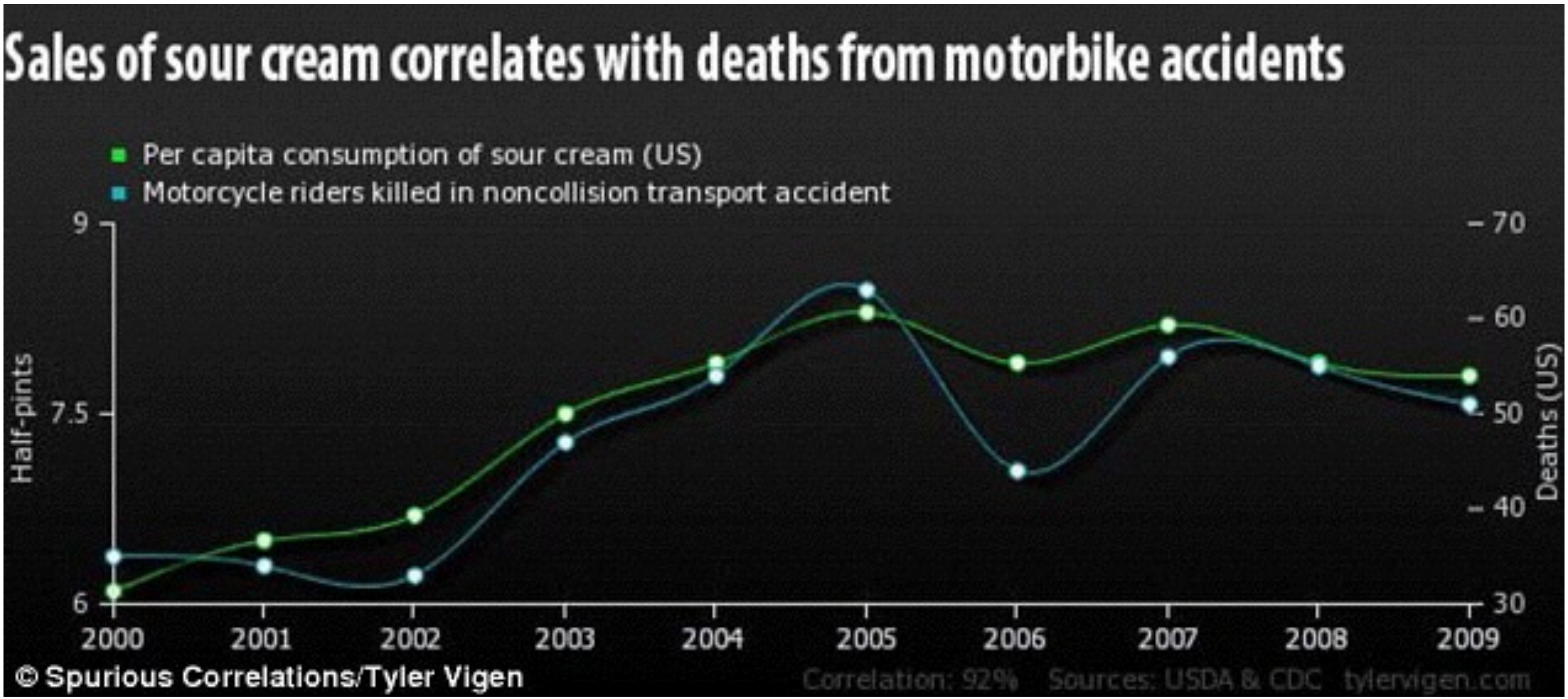
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n s_x s_y}$$

Pearson correlation using the samples of the attributes between X and Y

Data with different correlation values



Correlation



Correlation