PENNSTATE

Chapter 2: Data

Presentation extended from the slides of the textbook, Introduction to Data Mining by Tan et al. and supplementary material

Overview

- What is data?
  - Data Objects and Attributes
  - Data Types
  - Data Quality
- Data Preprocessing
- Data Similarity and Dissimilarity

What is Data?

- Data captures things, phenomena, etc, in forms of collection of *data objects* and their *attributes*
- An *attribute* is a property or characteristic of an object
  - E.g., eye color of a person, temperature, etc.
  - Attribute is also known as variable, feature, field or characteristic
- An *object* is described by a set of attributes
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

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

Objects

Attribute Values

- Attribute values are *numbers* or *symbols* assigned to an attribute
- Distinction between attributes and attribute values
  - *Attribute is the semantic notation while the attribute value is the numeric measure or symbolic representation.*
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and Age are both integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

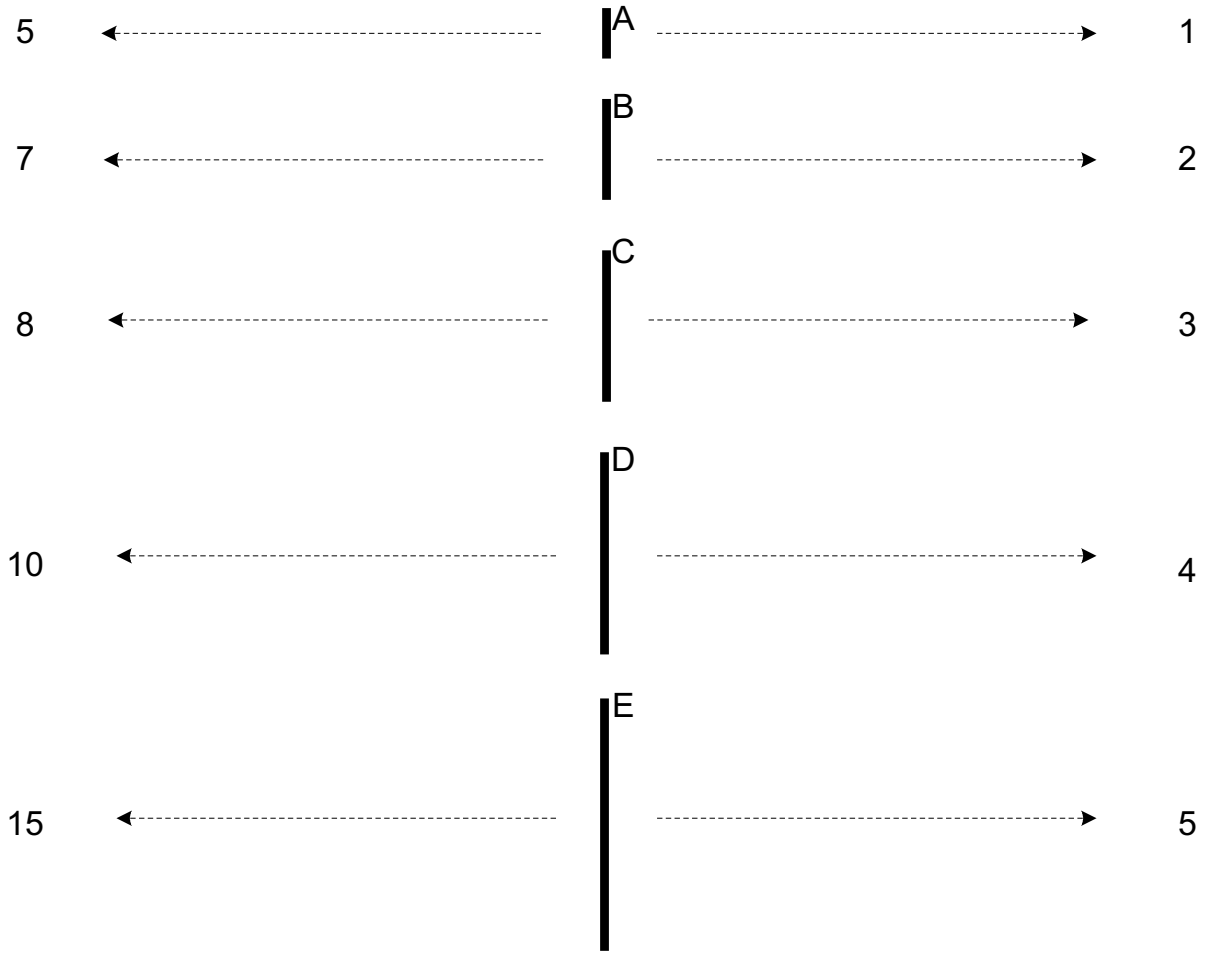
Process of Measurement

- A measurement scale is a rule (function) that associate a numerical or symbolic value with an attribute of an object.
- The process of measurement is the application of a measurement scale to associate a value with a particular attribute of an object.
- The properties of an attribute may not be the same as the *intended* properties of the values used to measure the attribute
  - *Choose a measure carefully!*
  - Integers can be used to represent Employee attributes such as Age and ID Number, but not all integer operations can be meaningfully applied to them.

Measurement of Length

Different measurements can be used to capture the desired properties of attributes, e.g., length, based on application requirements.

This scale preserves only the ordering property of length.



This scale preserves the ordering and additvity properties of length.

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale of 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + - (differences are meaningful)
  - Multiplication: * / (ratios are meaningful)

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10 ° is twice that of 5° on
  - the Celsius scale?
  - the Fahrenheit scale?
  - the Kelvin scale?

- Consider measuring the height above average
  - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?
  - Is this situation analogous to that of temperature?

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values are only to distinguish objects. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative	Interval	*new_value = a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

This categorization of attributes is due to S. S. Stevens

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as *integer* variables.
  - *Binary* attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as *floating-point* variables.

Asymmetric Attributes

- Only presence (a non-zero value) is important
  - Words present in documents
  - Items present in customer transactions
- If we met a friend in the grocery store, would we ever say the following?

  "I see our purchases are very similar since we didn't buy most of the same things."
- We need two asymmetric binary attributes to represent one ordinary binary attribute
  - Association analysis uses asymmetric attributes
- Asymmetric attributes typically arise from objects that are sets

Some Extensions and Critiques

- Velleman, Paul F., and Leland Wilkinson. "Nominal, ordinal, interval, and ratio typologies are misleading." The American Statistician 47, no. 1 (1993): 65-72.
- Mosteller, Frederick, and John W. Tukey. "Data analysis and regression. A second course in statistics." Addison-Wesley Series in Behavioral Science: Quantitative Methods, Reading, Mass.: Addison-Wesley, 1977.
- Chrisman, Nicholas R. "Rethinking levels of measurement for cartography."Cartography and Geographic Information Systems 25, no. 4 (1998): 231-242.

Critiques

- Incomplete
  - Asymmetric binary
  - Cyclical
  - Multivariate
  - Partially ordered
  - Partial membership
  - Relationships between the data
- Real data is approximate and noisy
  - This can complicate recognition of the proper attribute type
  - Treating one attribute type as another may be approximately correct

Critiques …

- Not a good guide for statistical analysis
  - May unnecessarily restrict operations and results
    - Statistical analysis is often approximate
    - E.g., using interval analysis for ordinal values may be justified
  - Transformations are common but don't preserve scales
    - Can transform data to a new scale with better statistical properties
    - Many statistical analyses depend only on the distribution

More Complicated Examples

- ID numbers
  - Nominal, ordinal, or interval?
- Number of cylinders in an automobile engine
  - Nominal, ordinal, or ratio?
- Biased Scale
  - Interval or Ratio

Key Messages for Attribute Types

- The types of operations you choose should be "meaningful" for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are only four properties of data
  - The data type you see (often numbers or strings) may not capture all the properties or may suggest properties that are not there
  - Analysis may depend on these other properties of the data (many statistical analyses depend on distribution)
  - Many times what is meaningful is measured by statistical significance
  - But in the end, what is meaningful is measured by the domain

Data Sets

- Many types of data sets
  - Record Data, e.g., transaction, document, etc.
  - Graph Data, e.g., World Wide Web, molecular structures, social networks, etc.
  - Ordered Data, e.g., temporal data, sequential data, genetic sequence, spatial data, etc.
- Important characteristics of data Sets
  - Dimensionality: curse of dimensionality
  - Distribution: statistical analysis usually assume certain distribution (many DM algorithms do not make assumption but distribution does have strong impact)
    - Sparsity: only presence (with non-null values) counts. This could be an advantage in terms of computing and storage.
  - Resolution: patterns depend on the scale.

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

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

Data Matrix

- If data objects have the same fixed set of *numeric* attributes, then the data objects can be thought of as *points in a multi-dimensional space*, where each dimension represents a distinct attribute

- Such data set can be represented by an $m \times n$ matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Transaction Data

- Special type of record data & sparse data matrix each record (transaction) involves a set of items.
  - Consider a grocery store. The set of products (items) purchased by a customer during one shopping trip constitute a transaction.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

  - Transaction dataset are usually represented as the above instead of sparse data matrix.

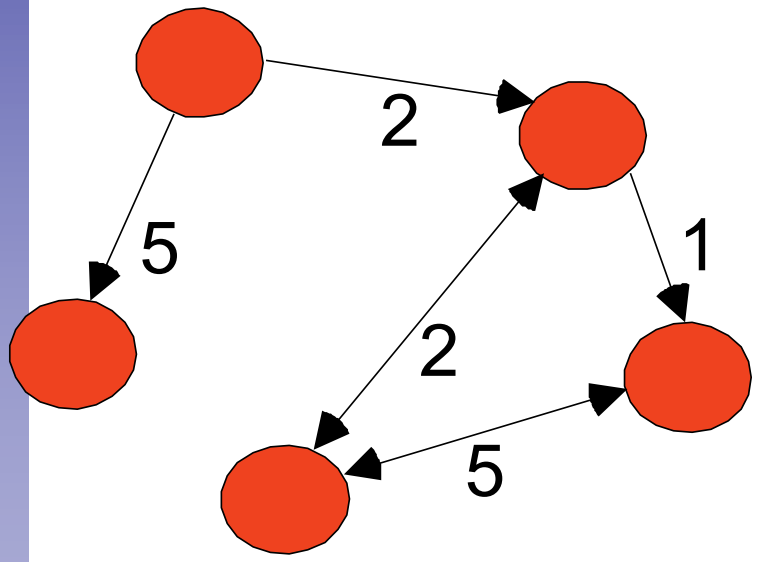
Document Data

- Each document is a *term vector*,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Graph Data

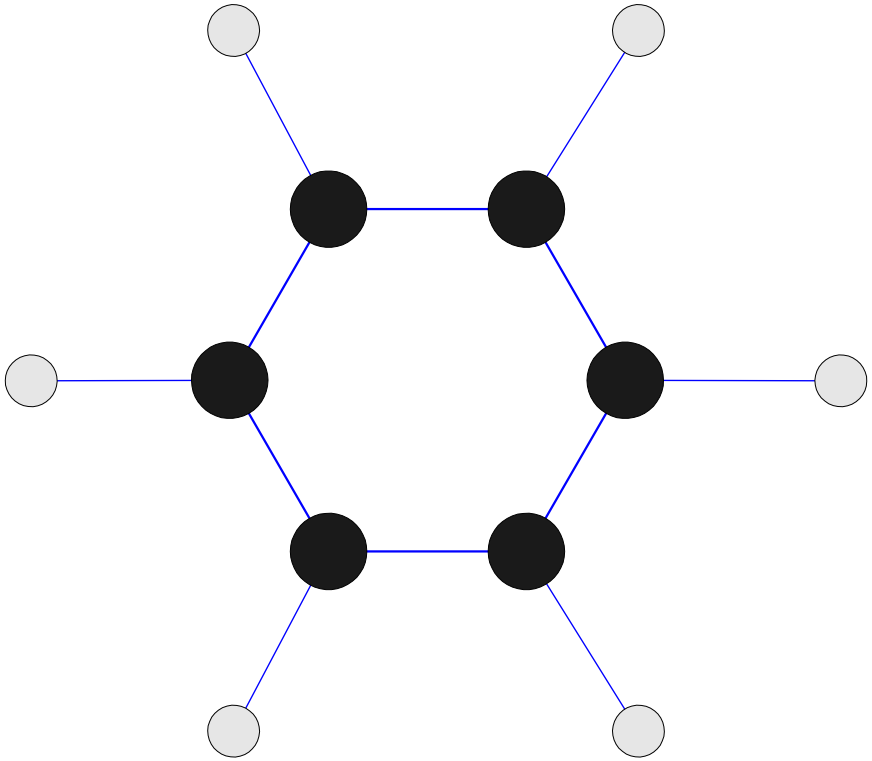
- A graph is a powerful representation of data
  - captures *relationship* among objects
  - captures *complex structure* of objects
- Examples: Generic graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

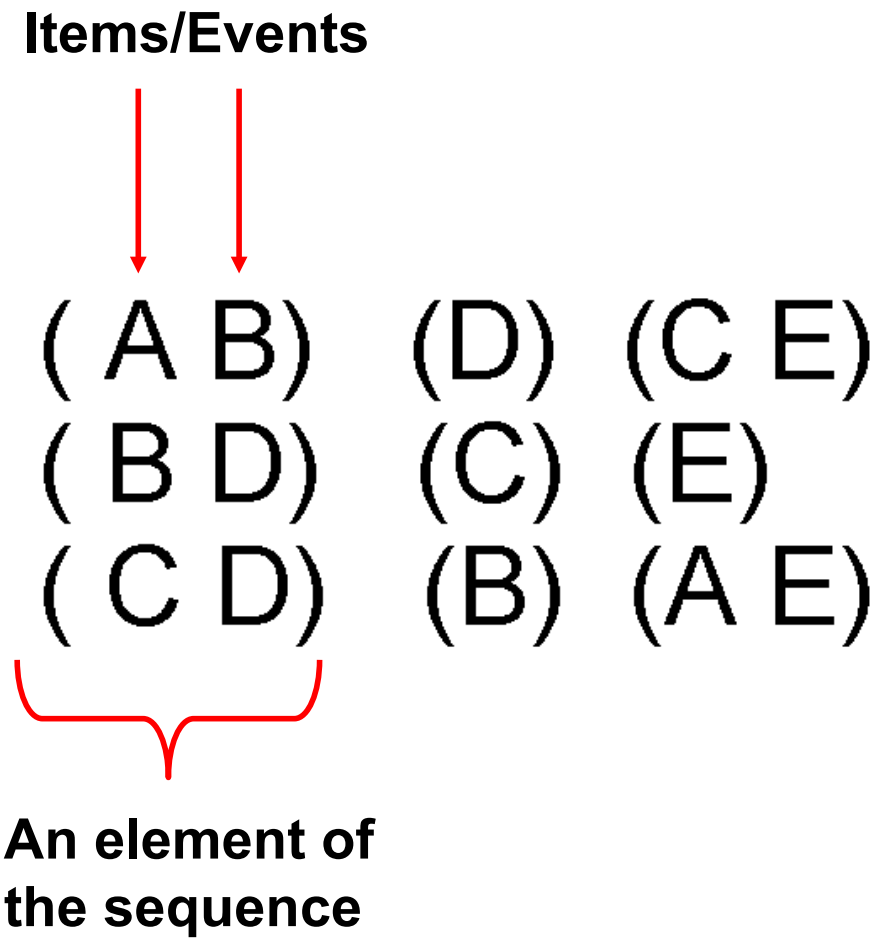
Chemical Data

- Benzene Molecule: C_6H_6



Ordered Data

- For some data sets, the attributes involve *order* in time and space.
- E.g., sequences of transactions.



Ordered Data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

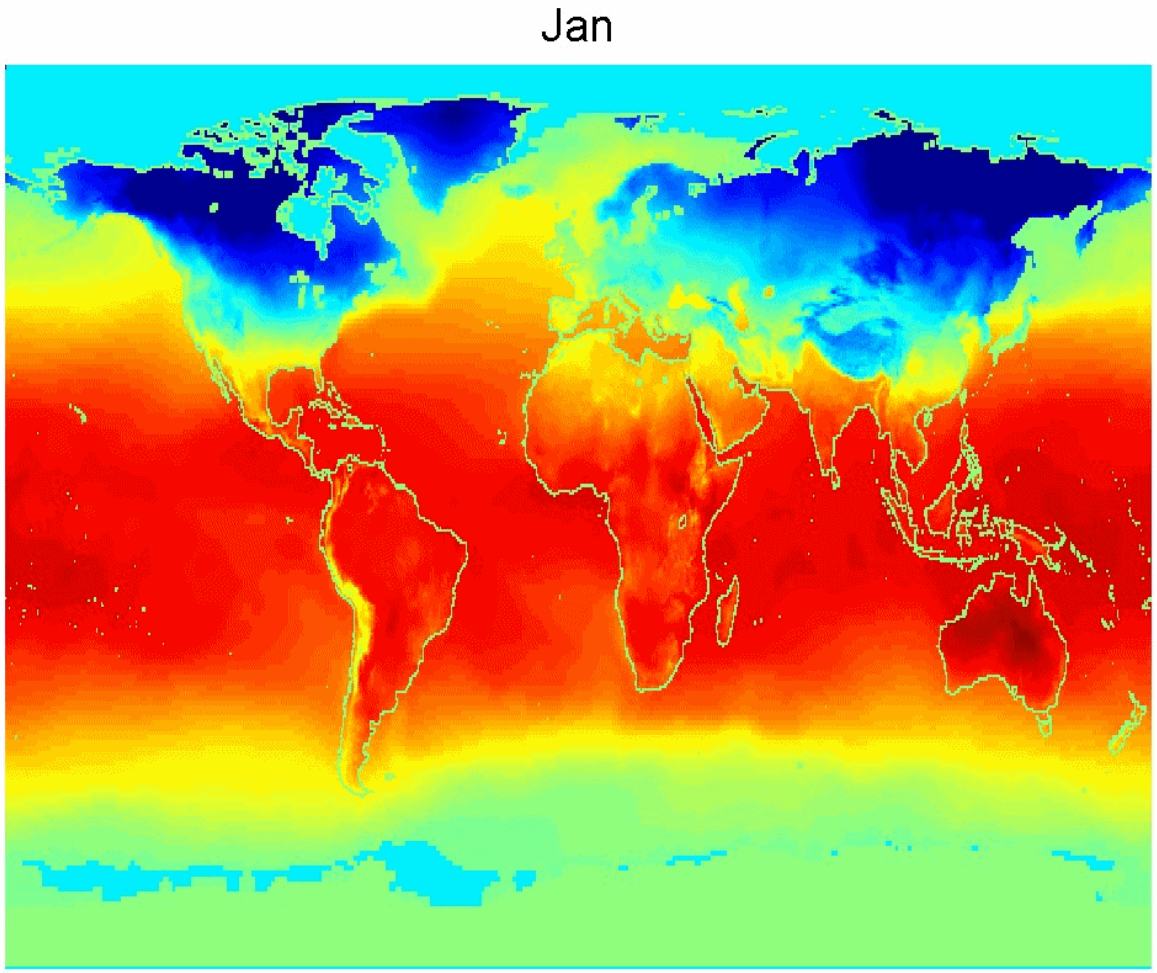
Ordered Data

- Time series data

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Data Quality

- Poor data quality negatively affects many data processing efforts

  "The most important point is that poor data quality is an unfolding disaster. Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate."

  Thomas C. Redman, DM Review, August 2004

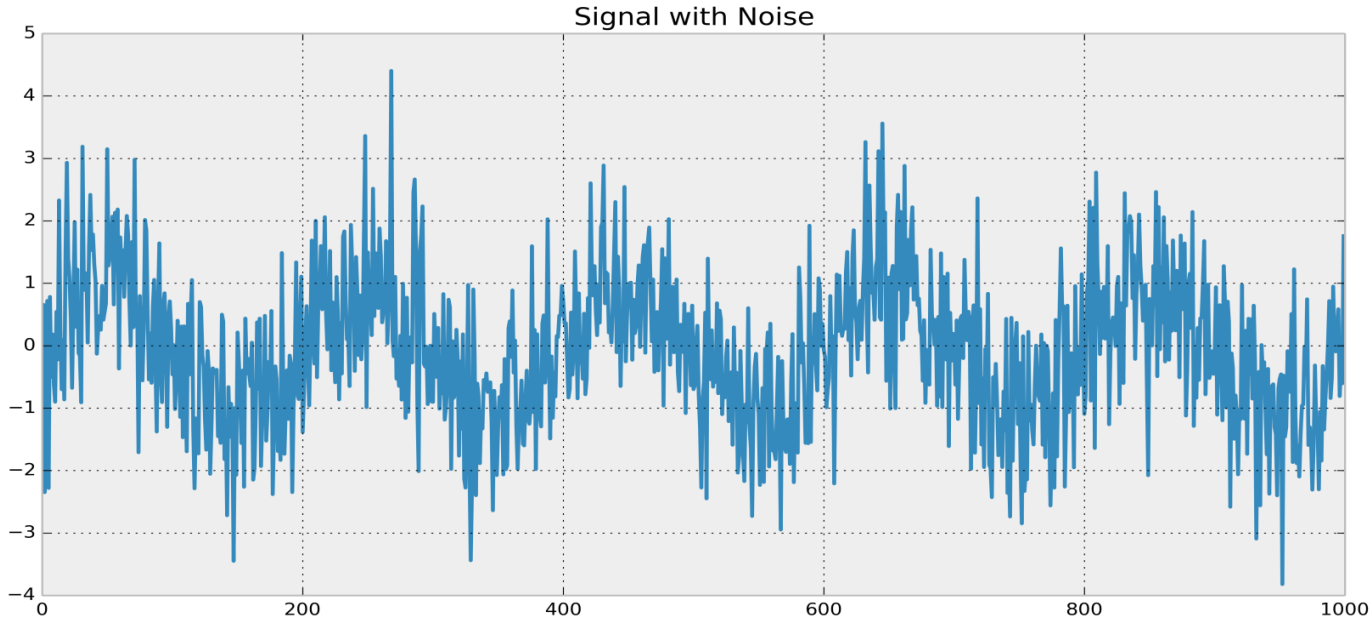
- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

Data Quality

- Data mining applications often use data collected for other (or future) applications, and thus facing serious data quality issues.
  - What kinds of data quality problems?
  - How can we detect problems with the data?
  - What can we do about these problems?
- Many data quality issues are related to the process of *measurement* and *data collection*. For examples:
  - Noise and outliers
  - missing values, inconsistent values
  - duplicate data

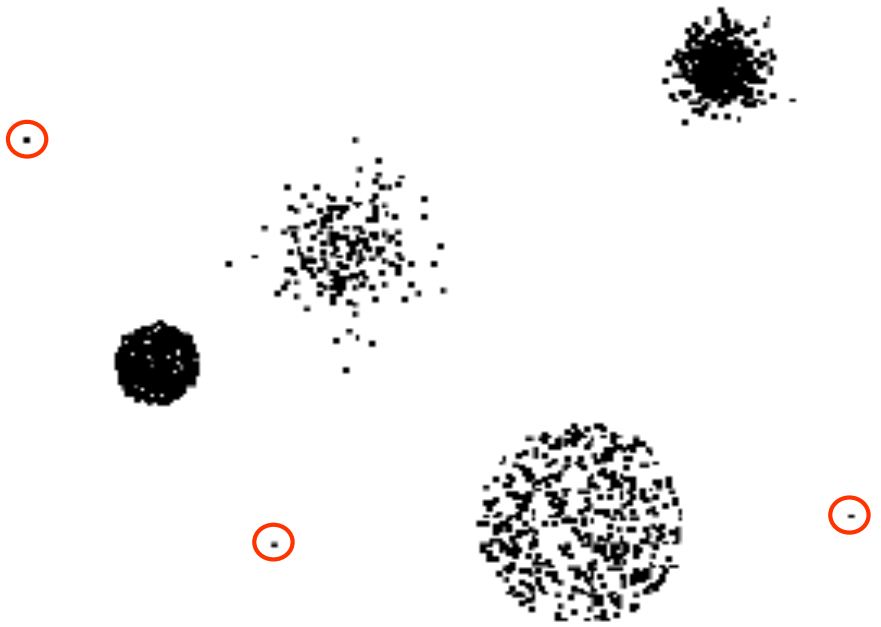
Noises

- Noise refers to deviation from the original values
  - E.g., distortion of a person's voice when talking on a poor phone and "snow" on television screen
  - It's the random component of a measurement error
  - Eliminating noises is usually difficult
- Data mining focuses on devising robust algorithms to produce acceptable result in presence of noises.



Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - Case 1: Outliers are noise that interferes with data analysis
  - Case 2: Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection
- Outliers may also refer to unusual attribute values

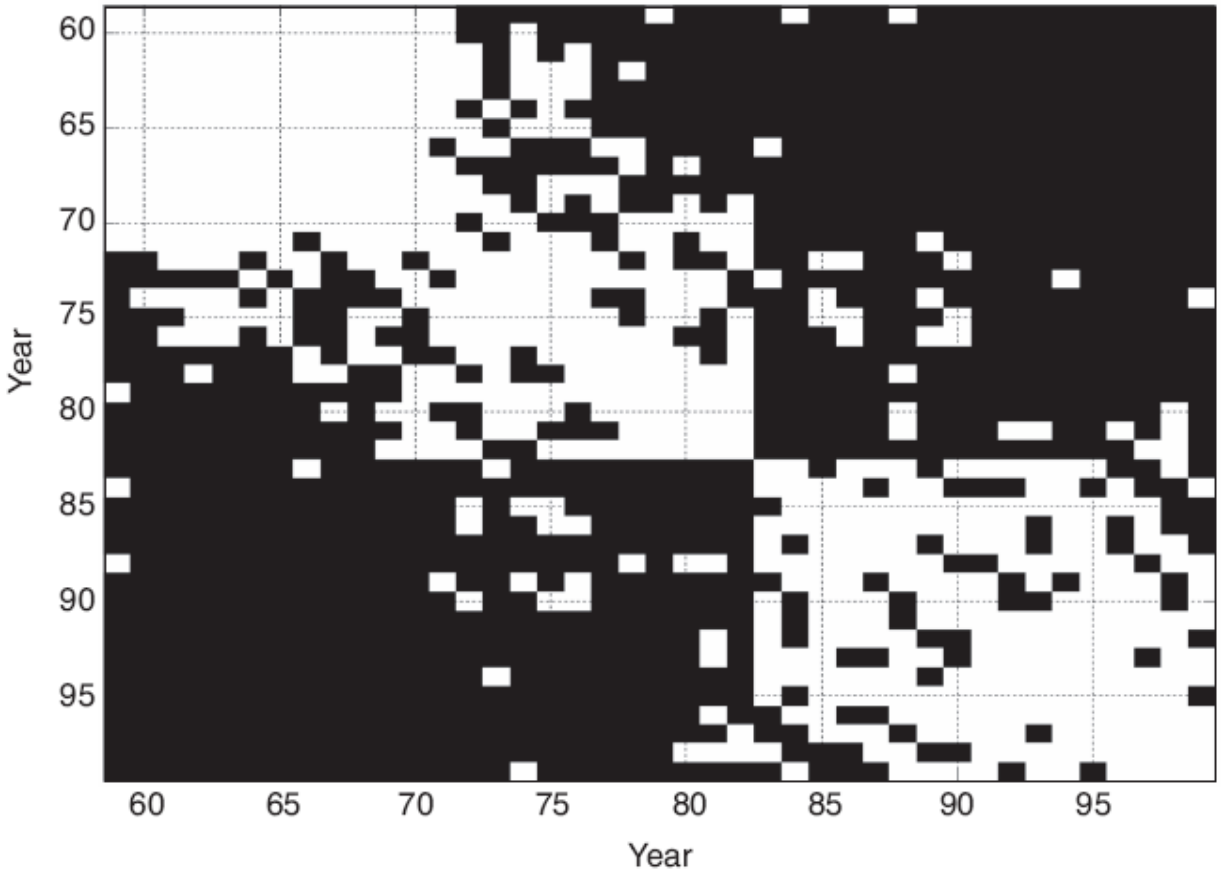


Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate Data Objects (or attributes w/ caution)
  - Estimate Missing Values
  - Ignore the Missing Value during Analysis
  - Replace with all possible values (weighted by their probabilities)

Inconsistent Data

- Inconsistent data needs to be detected/corrected, if all possible.
- The figure plots correlation of sea surface temperature from two different sources



Duplicate Data

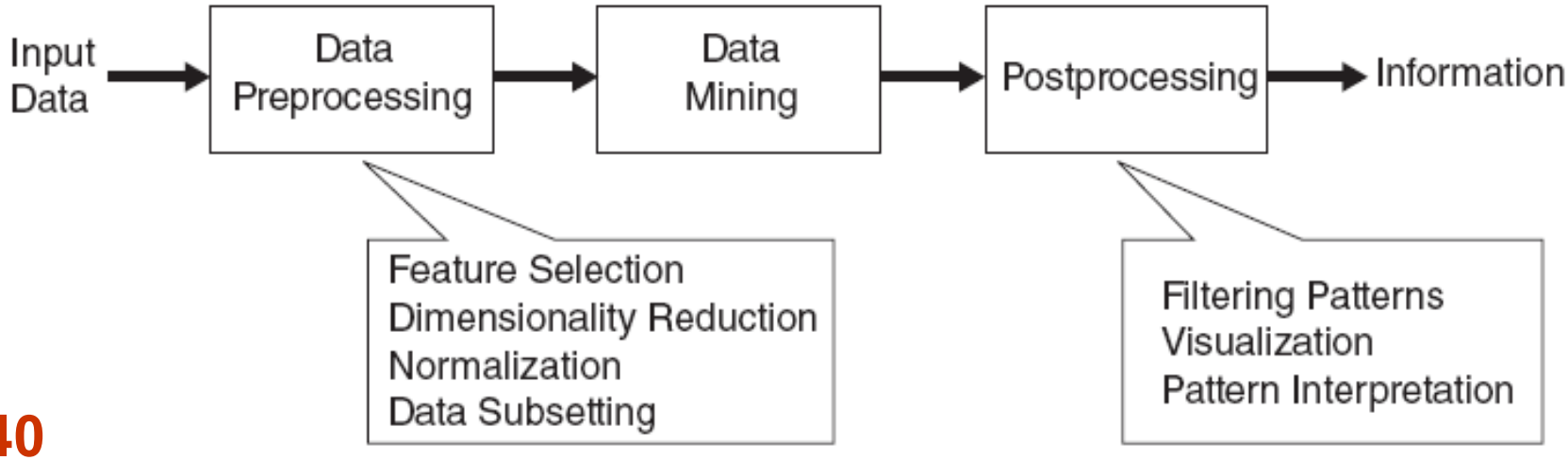
- Data sets may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning (Deduplication)
  - Process of dealing with duplicate data issues

Application-related Quality Issues

- Data is high quality if it's suitable for intended use
- General issues related to applications
  - Timeliness: Some data starts to age as soon as it's collected, e.g., purchasing behavior of customers
  - Relevance: Data must contain the information necessary for the application. Sampling bias occurs when a sample does not contain different types of objects in proportion to their actual occurrence in the population.
  - Knowledge about the data: The quality of documentation that describes the data may either aid or hinder the analysis/mining of the data.

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation



Aggregation

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

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation: Sales Data

- Change the scope/scale of data to provide a higher-level view.

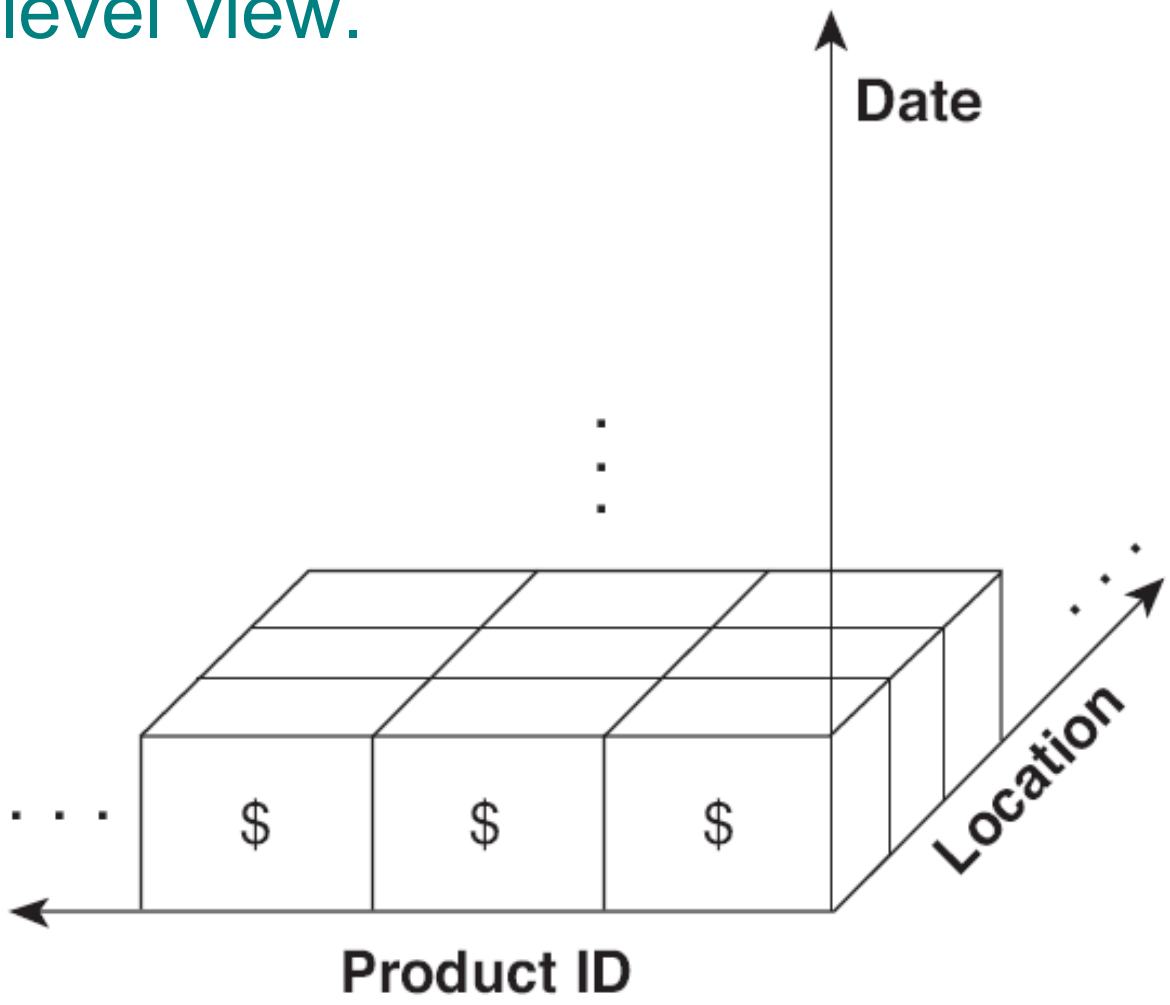
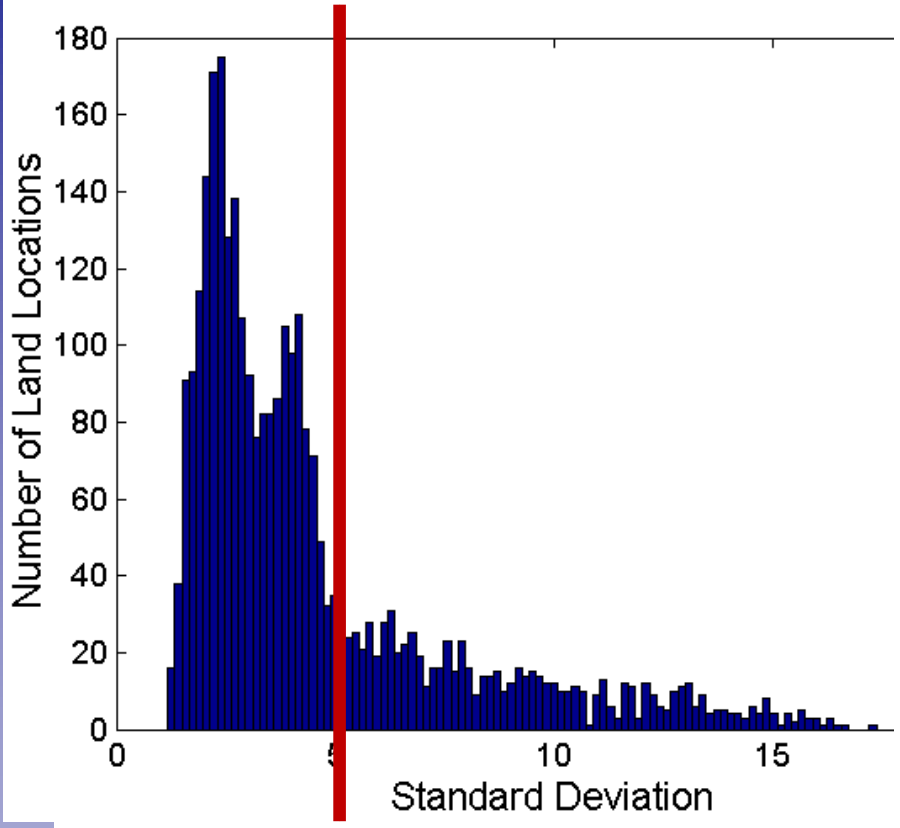


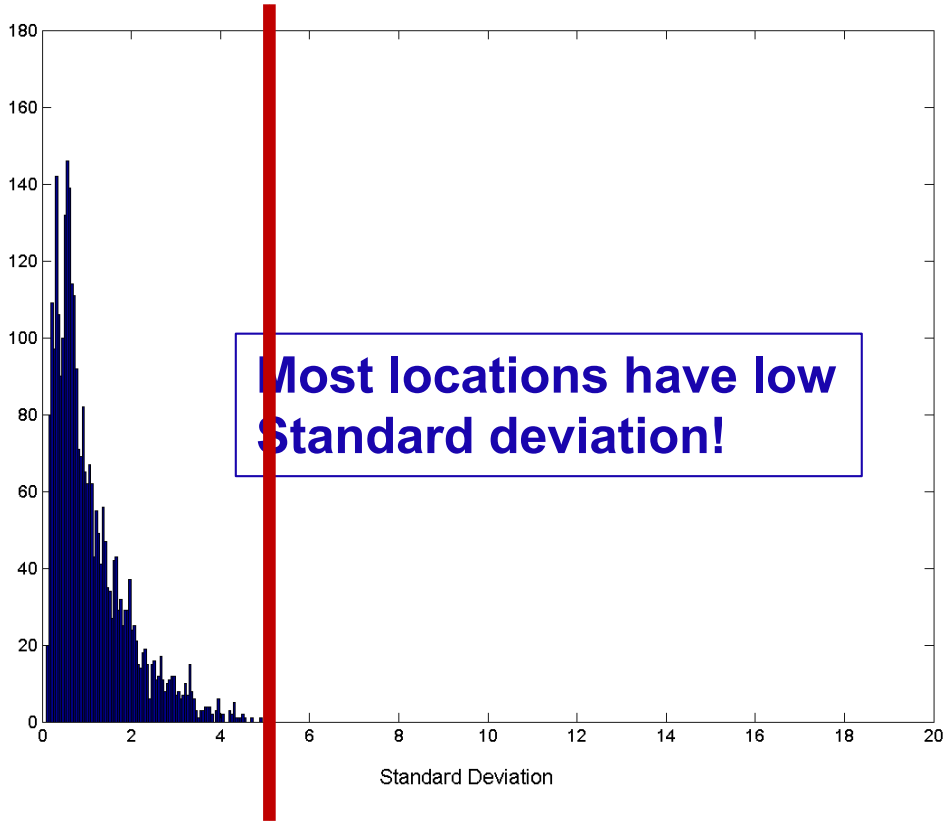
Figure 3.31. Multidimensional data representation for sales data.

Aggregation – Precipitation Data

The behavior of group is usually more stable than individuals.



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because *obtaining* the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining due to scalability issue because *processing* the entire set of data of interest is too expensive or time consuming.
  - Big data computing frameworks aim to address the processing needs. Recent advances help but sampling is still needed sometimes.

Sampling Principle

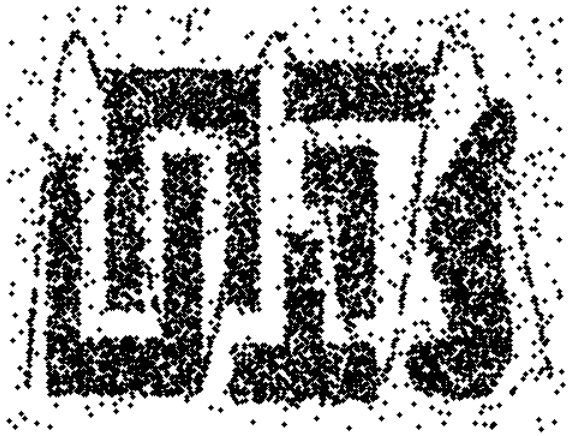
- The key principle for effective sampling is the following:
  - *Using a sample will work almost as well as using the entire data sets, if the sample is representative*
  - *A sample is representative if it has approximately the same property (of interest) as the original set of data*

Sampling Approaches

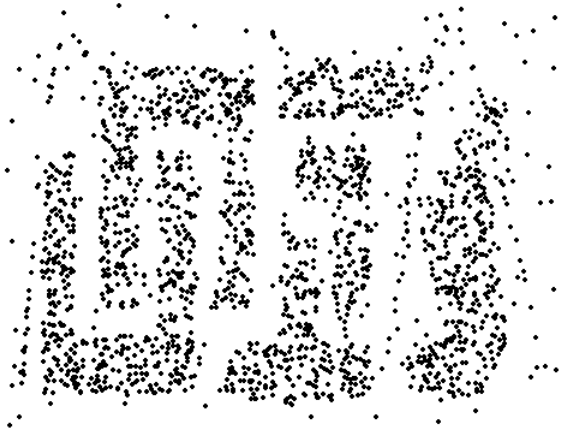
- Simple Random Sampling
  - There is an equal probability of selecting any item
- Sampling without replacement
  - An item is removed from the population after being selected
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - the same object can be picked up more than once
    - Easy to analyze as probability not affected by sampling.
- Stratified sampling
  - Split the data into pre-specified groups; then draw random samples from each group

Sample Size

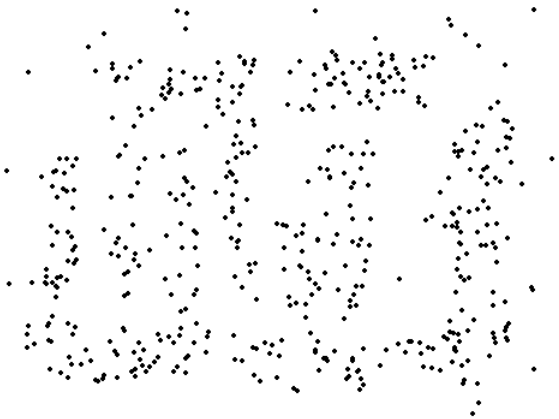
- Large sample is representative but loose the advantage of sampling.
- Small sample may result in missing or erroneous patterns



8000 points



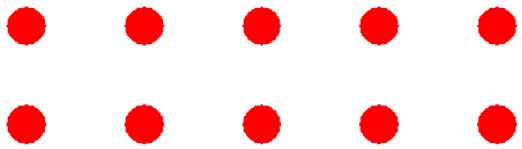
2000 Points

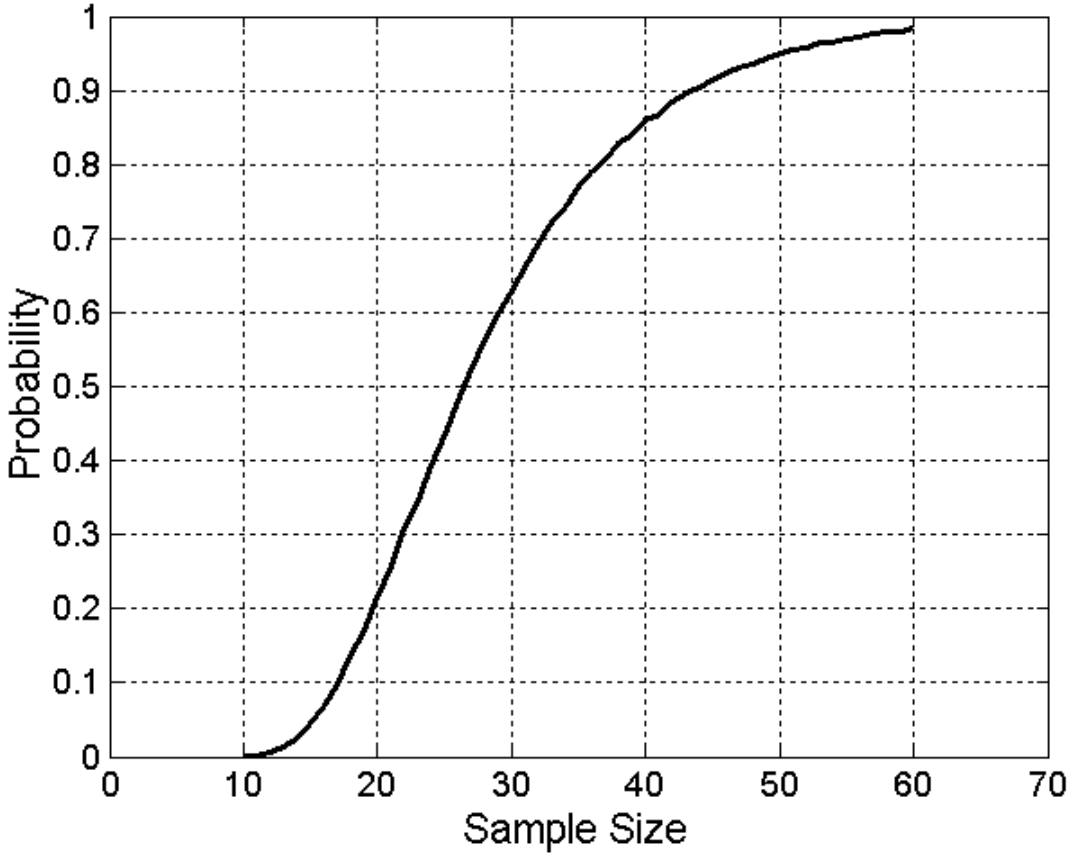


500 Points

Determine Sample Size

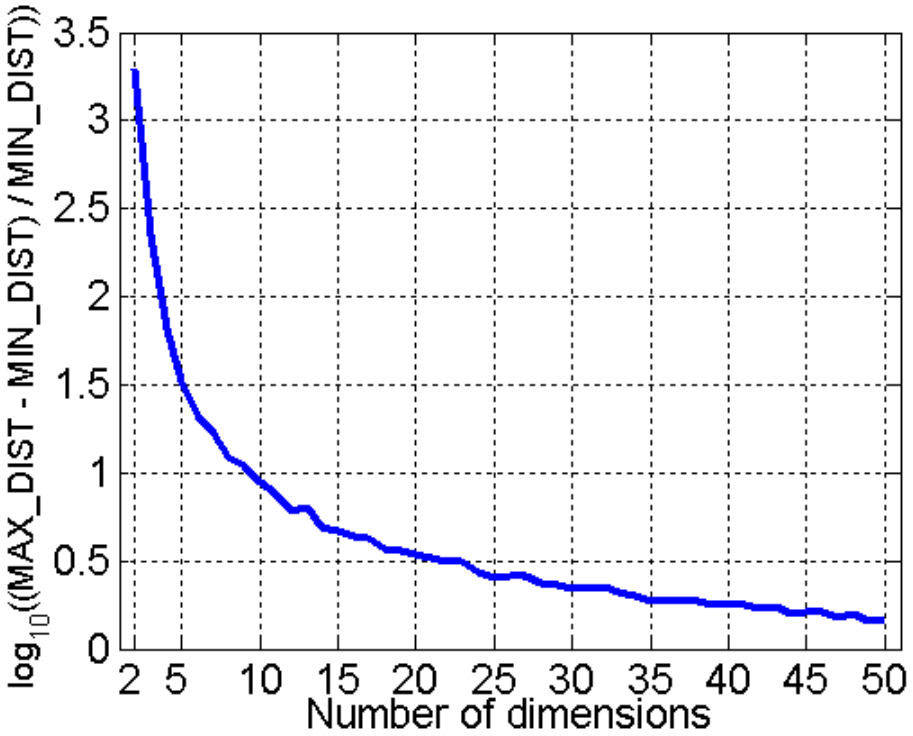
- What sample size is necessary to get at least one object from each of 10 groups?





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



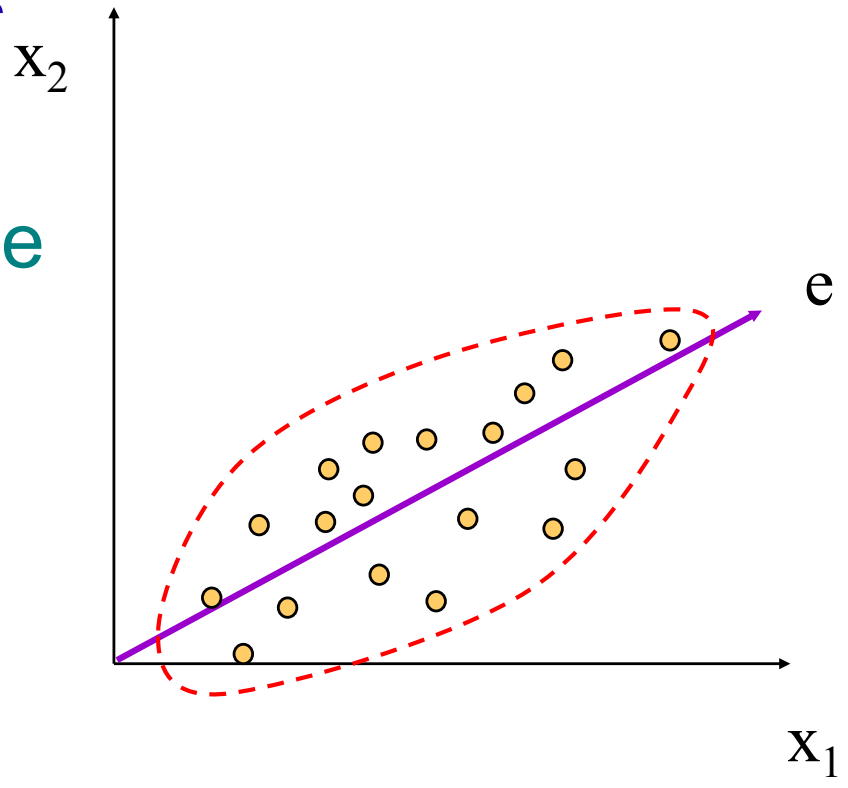
Randomly generate 500 points. Compute difference between max and min distance between any pair of points.

Dimensionality Reduction

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

Principle Component Analysis

- Find orthogonal projections (i.e., linear combinations of original dimensions) that captures the largest amount of variation in data
- In Linear Algebra, this is to find the *eigenvectors* of the covariance matrix
- The eigenvectors (principle components) defines the new dimension-reduced space



Feature (Subset) Selection

- Select some features, eliminate others.
  - *Will this cause information loss?*
- 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' GPA

Feature Selection

- Brute-force (ideal) approach:
  - Try *all possible feature subsets* as input to data mining algorithm
- Embedded approaches:
  - Feature selection occurs naturally as part of the data mining algorithm
  - Decision tree operates in this manner
- Filter approaches:
  - Features are selected *before* data mining is performed
  - E.g., select attributes with low pair-wise correlation
- Wrapper approaches:
  - Use the data mining algorithm as a black box to find the best subset of attributes

Steps for Feature Subset Selection

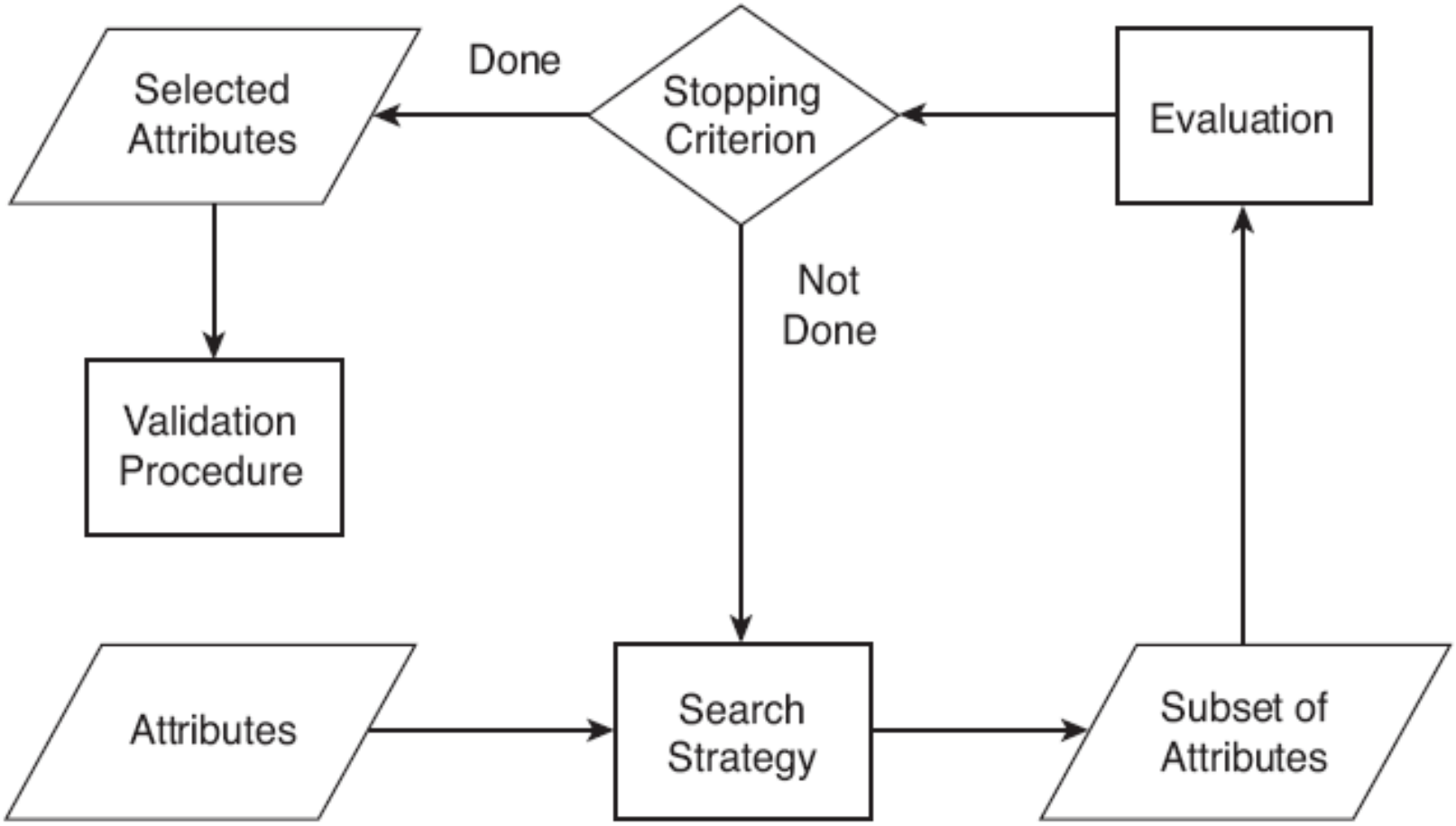


Figure 2.11. Flowchart of a feature subset selection process.

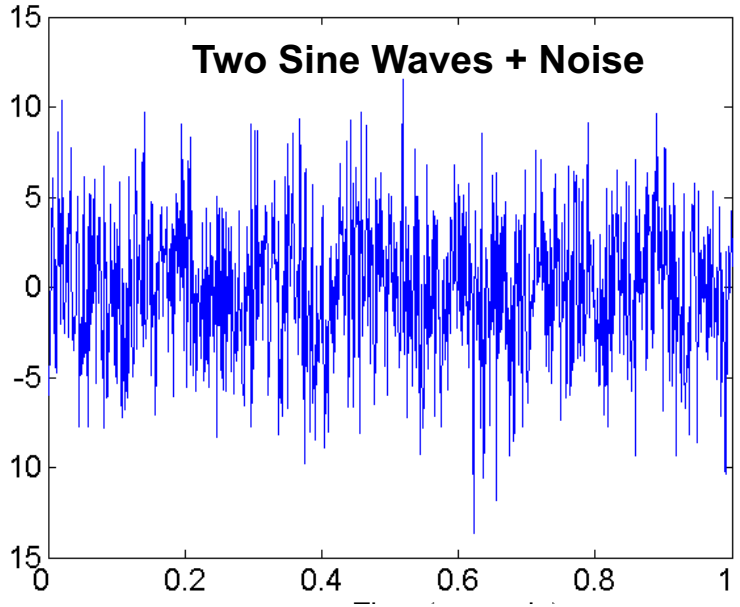
Feature Creation

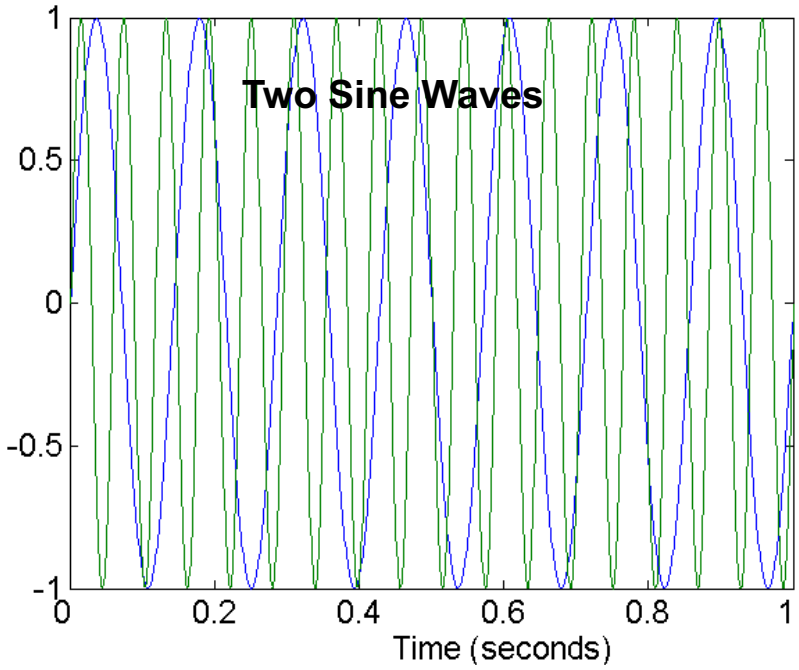
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

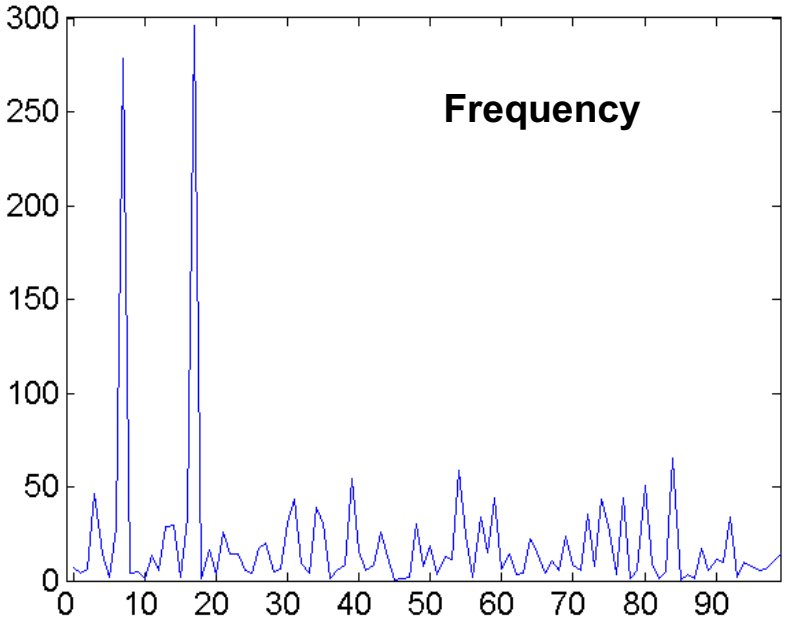
- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Feature Transformation
    - mapping data to new space
  - Feature Construction
    - combining features
    - E.g., derive speed from distance and interval from a vehicle trajectory dataset

Feature Transformation: Mapping Data to a New Space

- Fourier transform
- Wavelet transform







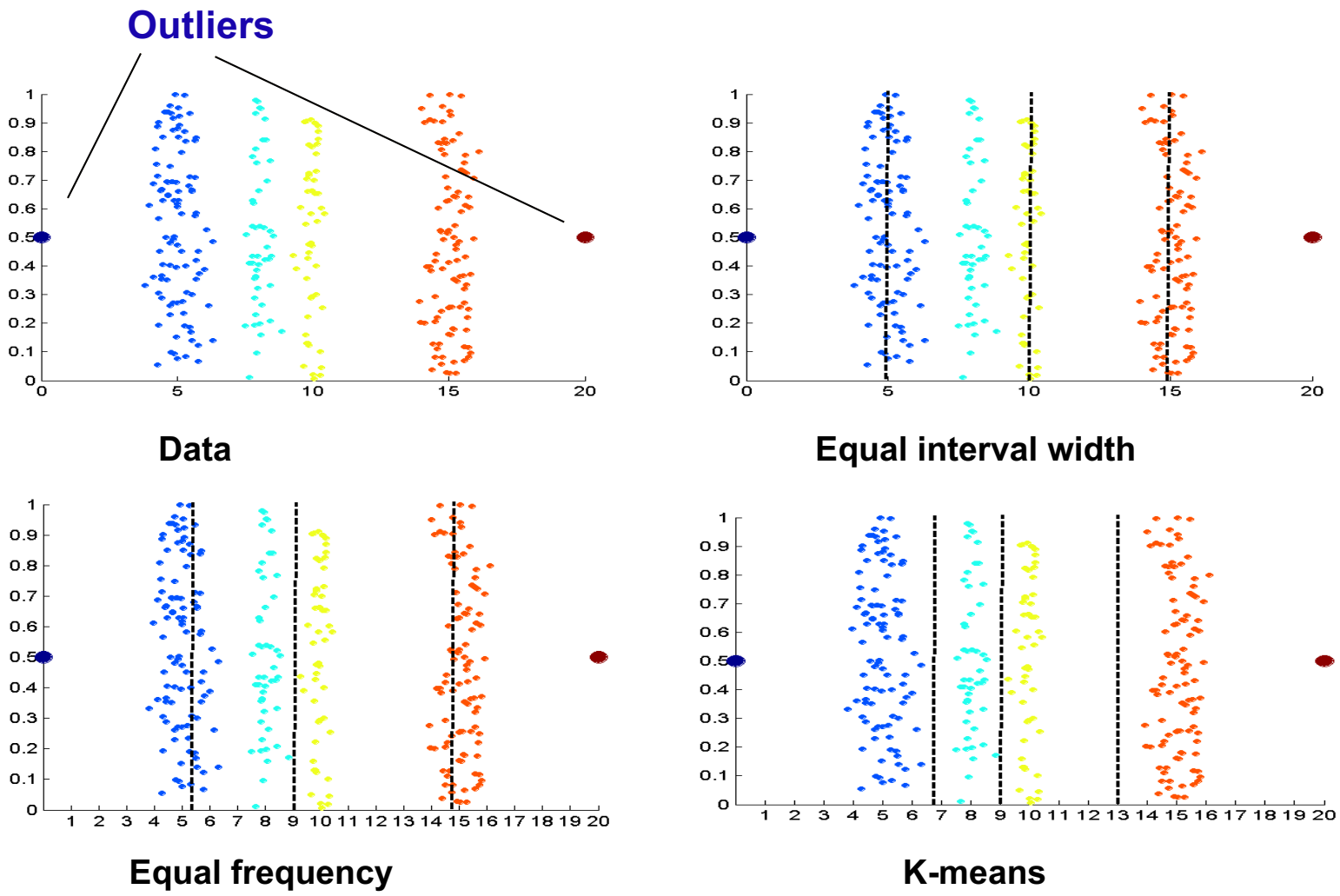
Discretization and Binarization

- *Discretization:* transform a continuous attribute into a categorical one.
  - Some data mining algorithms, e.g., classification, require categorical attributes.
- *Binarization*: transform continuous and categorical attributes into binary one
  - Association pattern discovery may require binary attributes.

Discretization of Continuous Attributes

- Sort the attribute values and divide them into *n* intervals (by specifying *n-1* split points).
  - The key issues are how many split points to choose and where to put them.
- Depending on class information (i.e., labels) are used or not, discretization methods can be classified as follows:
  - Unsupervised discretization
    - Equal Width, Equal Depth/Frequency, Clustering-based
  - Supervised discretization
    - Class information is useful, as unsupervised methods usually result in intervals of objects with mixed labels.

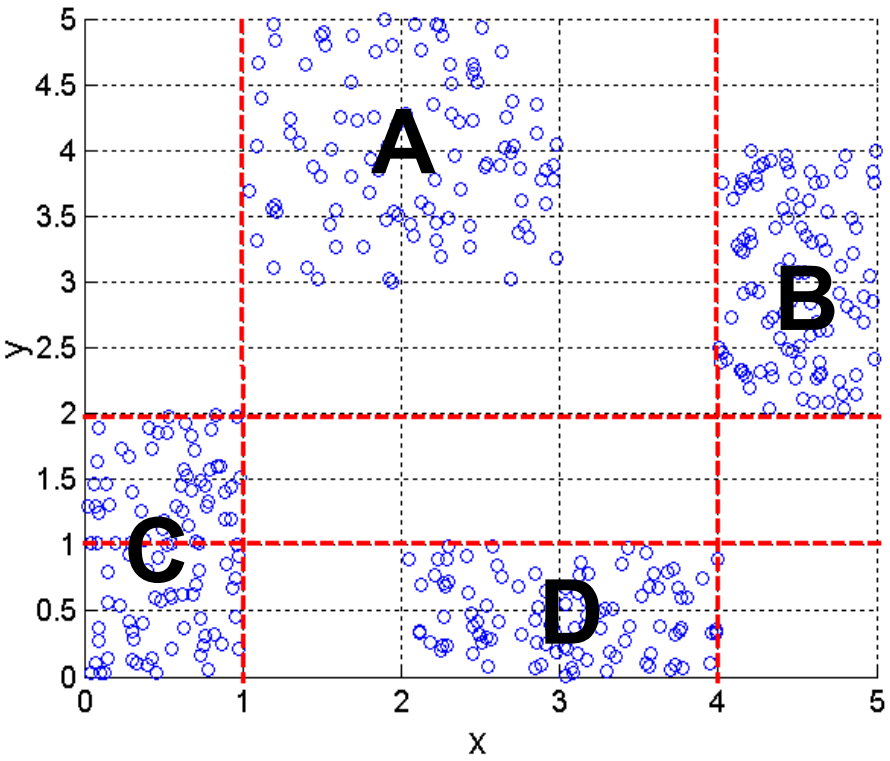
Discretization Without Using Class Labels (Unsupervised)



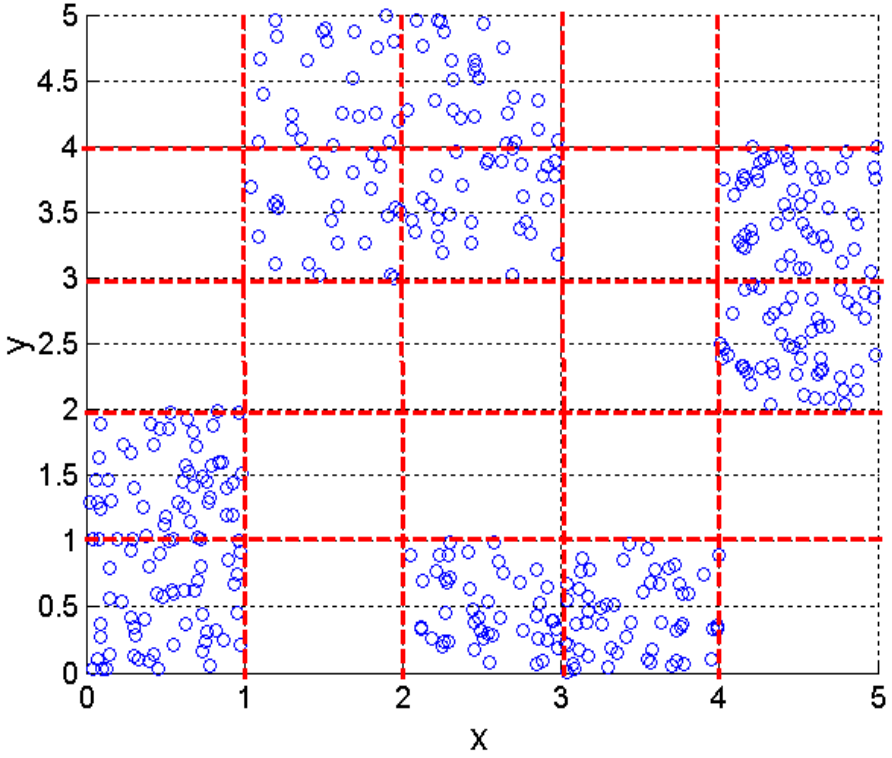
Split the points into 4 Intervals

Discretization Using Class Labels

- Entropy based approach: use entropy as a *purity* measure to divide the objects.



3 categories for both x and y



5 categories for both x and y

- In left figure, the separation in one dimension is not as good as two dimensions. In right figure, it's ok.

Binarization of Categorical Attributes

- Simple technique: *m* categories expressed in *log m* bits.
  - Sometimes create unintended relationship between bits
- One hot representation: *m* categories expressed in *m* bit. Only one bit is set to 1 and the rest bits are 0.
  - Bits are independent of each other

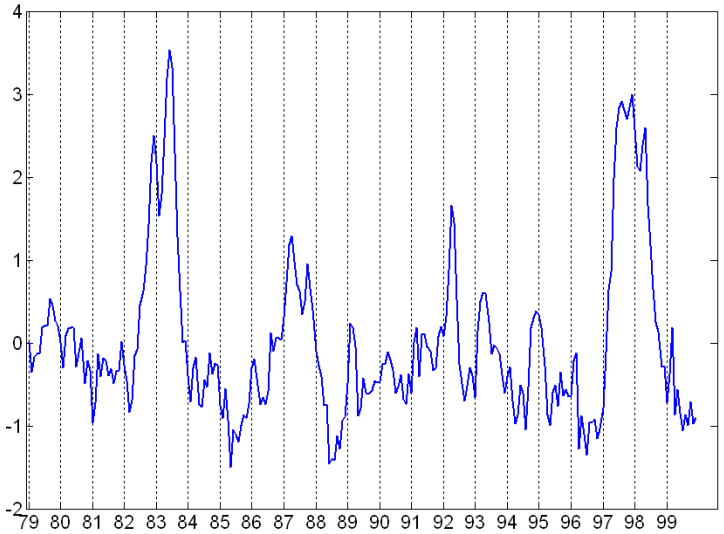
Categorical Attributes with Too Many Values

- *How to we handle it?*

- For ordinal attributes, *discretization* techniques could be used
- Categorical attributes with a lot of values can be combined, based on some relationship or taxonomy, to reduce the number of values
  - E.g., EE, CSE, IE all belong to College of Engineering.

Attribute Transformation

- Sometimes we need to transform attribute values into different form to amplify/smooth their effect in data mining algorithms.
  - E.g., salary and age are considered together by weighted sum.
- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, $\log(x)$, e^x, $|x|$
  - *Standardization* and *Normalization*



Similarity and Dissimilarity

- *Similarity* and *dissimilarity* are important for many data mining techniques, e.g., clustering.
- 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 objects*
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

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 $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Euclidean Distance

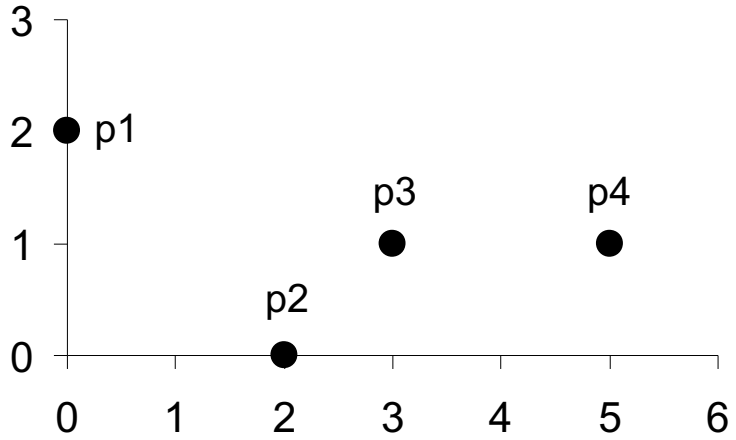
- Euclidean Distance

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

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.

- Standardization is necessary, if scales differ.

Euclidean Distance



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

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

Minkowski Distance: Examples

- r = 1. Manhattan distance (L_1 norm, City block, taxicab).
  - 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 (L_2 norm)
- $r \rightarrow \infty$. Supremum distance (L_{max} norm, L_∞ norm)
  - This is the maximum difference between any component/attribute of the vectors
- Do not confuse r (order) with *n (dimension)*, i.e., all distances are defined for all numbers of dimensions.

Minkowski Distance

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

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.*

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.

Similarity Between Binary Vectors (or Two Sets)

- A common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities with 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* and *Jaccard Coefficients*

SMC = number of matches / number of attributes
= $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

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

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

$M_{01} = 2$ (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**

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\|d_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5}$

$= 6.481$

$\|d_2\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5}$

$= 2.245$

$\cos(d_1, d_2) = .3150$

Extended Jaccard Coefficient (Tanimoto)

- Jaccard Coefficient is mainly for measuring binary attributes
- Extended Jaccard (Tanimoto) Coefficient is a variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

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

Correlation

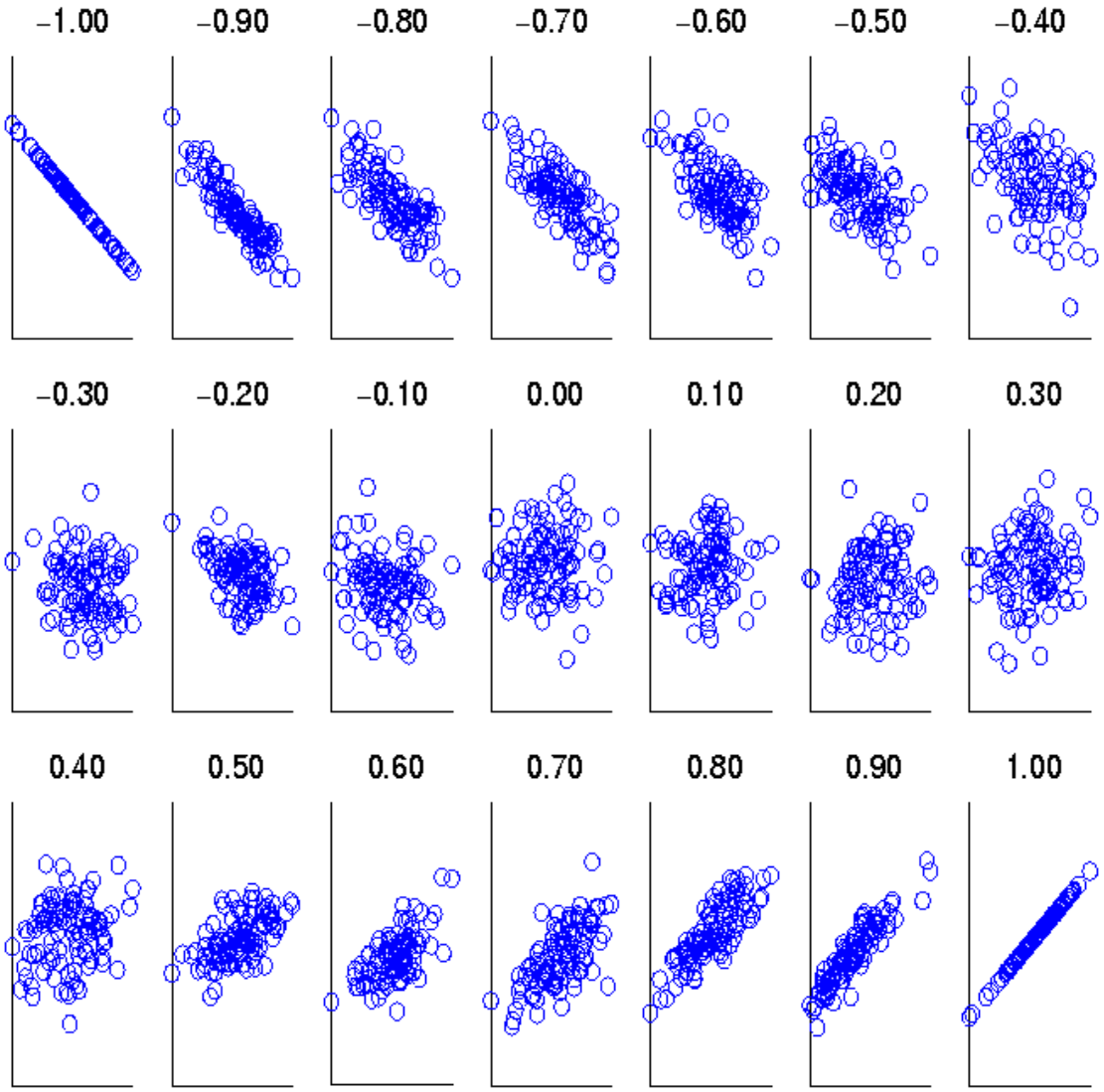
- Correlation measures the *linear relationship* between objects
- To compute correlation, we standardize data objects, *p* and *q*, and then take their dot product

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

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

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

Visually Evaluating Correlation



Scatter plots showing the correlations from –1 to 1.

Drawback of Correlation

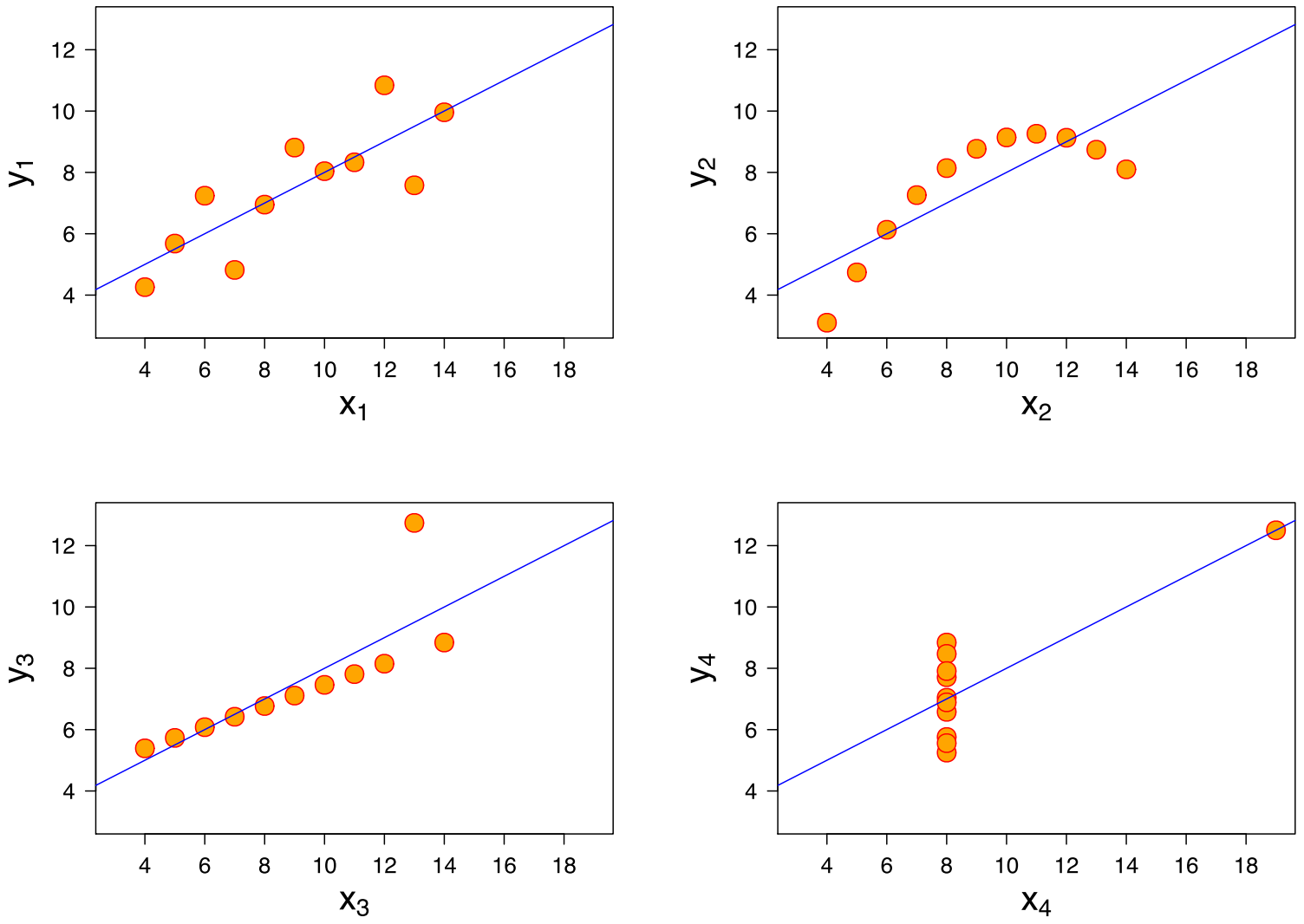
- $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$
- $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$

$y_i = x_i^2$

- $\text{mean}(\mathbf{x}) = 0$, $\text{mean}(\mathbf{y}) = 4$
- $\text{std}(\mathbf{x}) = 2.16$, $\text{std}(\mathbf{y}) = 3.74$

- corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)
  = 0

Anscombe's Quartet



Comparison of Proximity Measures

- Domain of application
  - Similarity measures tend to be specific to the type of attribute and data
  - Record data, images, graphs, sequences, 3D-protein structure, etc. tend to have different measures
- One can talk about various properties that you would like a proximity measure to have
  - Symmetry is a common one
  - Tolerance to noise and outliers is another
  - Ability to find more types of patterns?
  - Many others possible
- The measure must be applicable to the data and produce results that agree with domain knowledge

Information Based Measures

- Information theory is a well-developed and fundamental disciple with broad applications
- Some similarity measures are based on information theory
  - Mutual information in various versions
  - Maximal Information Coefficient (MIC) and related measures
  - General and can handle non-linear relationships
  - Can be complicated and time intensive to compute

Information and Probability

- Information relates to possible outcomes of an event
  - Transmission of a message, flip of a coin, or measurement of a piece of data

- The more certain an outcome, the less information that it contains and vice-versa
  - For example, if a coin has two heads, then an outcome of heads provides no information
  - More quantitatively, the information is related the probability of an outcome
    - The smaller the probability of an outcome, the more information it provides and vice-versa
  - Entropy is the commonly used measure

Entropy

- For
  - a variable (event), X,
  - with n possible values (outcomes), $x_1, x_2 \ldots, x_n$
  - each outcome having probability, $p_1, p_2 \ldots, p_n$
  - the entropy of X , $H(X)$, is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Entropy is between 0 and $\log_2 n$ and is measured in bits
  - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = (-p \log_2 p) + (-q \log_2 q)$$

  - For $p = 0.5$, $q = 0.5$ (fair coin) $H = 1$
  - For $p = 1$ or $q = 1$, $H = 0$

- What is the entropy of a fair four-sided die?

Entropy for Sample Data: Example

Hair Color	Count	p	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Maximum entropy is $\log_2 5 = 2.3219$

Entropy for Sample Data

- Suppose we have
  - a number of observations (m) of some attribute, X, e.g., the hair color of students in the class,

    where there are n different possible values
  - the number of observation in the i^{th} category is m_i
  - Then, for this sample

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- For continuous data, the calculation is harder

Mutual Information

- Information one variable provides about another
- Formally, $I(X,Y) = H(X) + H(Y) - H(X,Y)$,

  where $H(X,Y)$ is the joint entropy of X and Y,

$$H(X,Y) = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

  where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together
- For discrete variables, this is easy to compute
- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y)$, where n_X (n_Y) is the number of values of X (Y)

Mutual Information Example

Student Status	Count	p	$-p\log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p\log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p\log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

Mutual information of Student Status and Grade = 0.9928 + 1.4406 - 2.2710 = 0.1624

Maximal Information Coefficient

Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. "Detecting novel associations in large data sets." *science* 334, no. 6062 (2011): 1518-1524.

- Applies mutual information to two continuous variables
- Consider the possible binnings of the variables into discrete categories
  - $n_X \times n_Y \leq N^{0.6}$ where
    - n_X is the number of values of X
    - n_Y is the number of values of Y
    - N is the number of samples (observations, data objects)
- Compute the mutual information
  - Normalized by $\log_2(\min(n_X, n_Y))$
- Take the highest value

General Approach for Combining Similarities

- Sometimes attributes are of *many different types*, but an overall similarity is needed.
- A simple strategy is to take *average* of effective attributes.

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.
2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$