# CSE4334/5334 Data Mining

# 2 Data and Data Preprocessing

Deokgun Park

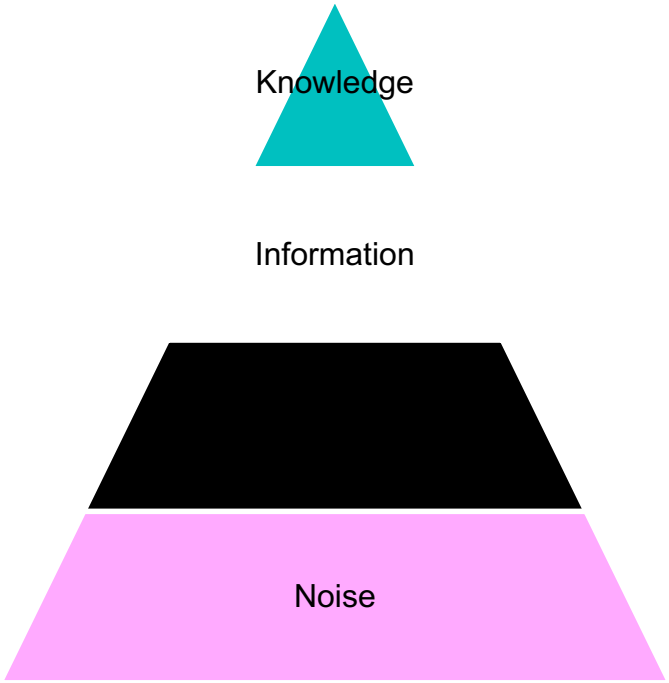
University of Texas at Arlington

(Slides courtesy of Pang-Ning Tan, Michael Steinbach, Vipin Kumar, and Chengkai Li)

# Key concepts

- Noise / Data / Information / Knowledge
- 4Vs of Big Data
- Structured vs Unstructured data / semi-structured data
- Types of Data
- Types of Attributes
- Noise and Outlier
- Similarity and Distance
- Data Preprocessing

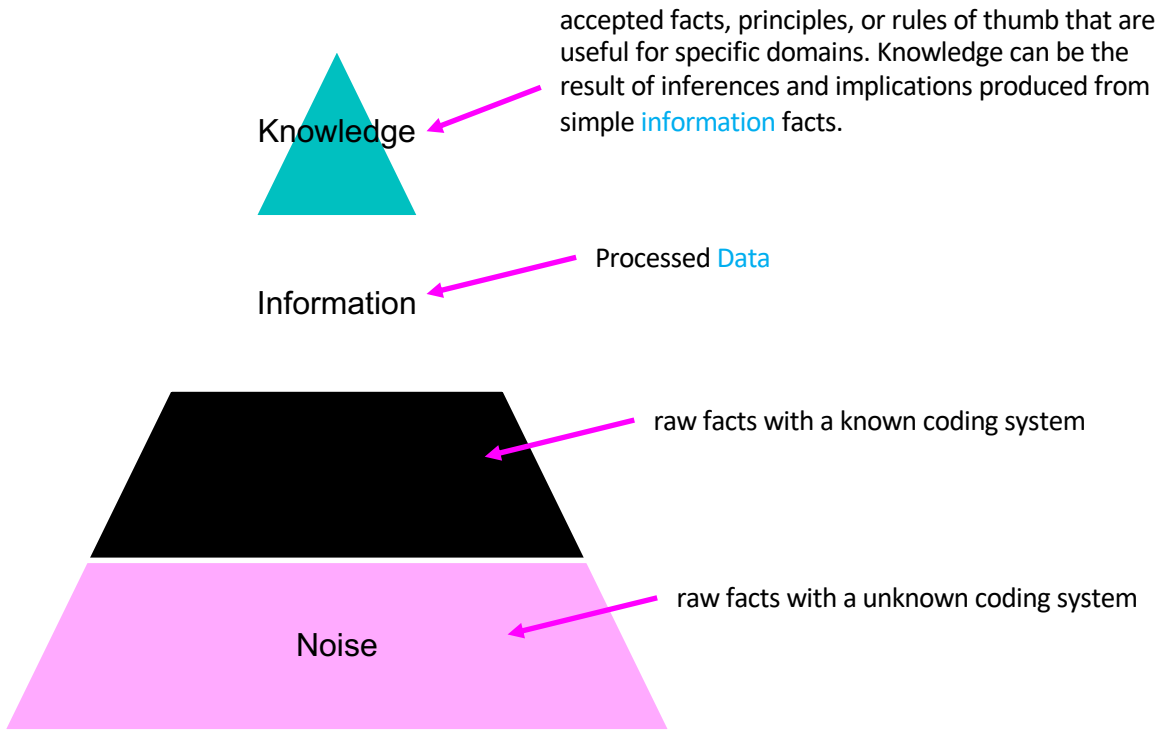
# Raggad's classification




Bel G. Raggad, Information Security Management: Concepts and Practice: CRC Press, 2010.



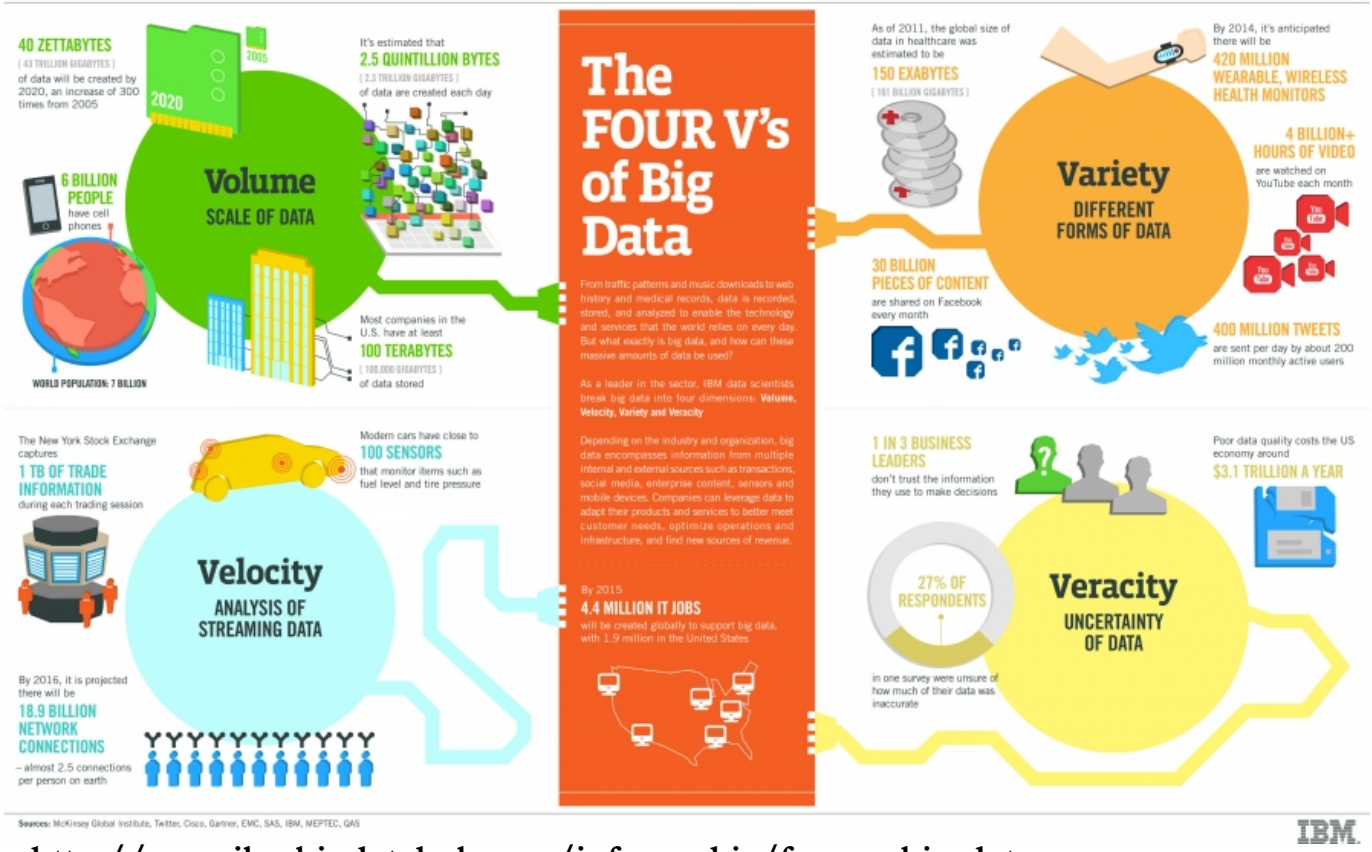
# Raggad's classification




Bel G. Raggad, Information Security Management: Concepts and Practice: CRC Press, 2010.



# Big Data




http://www.ibmbigdatahub.com/infographic/four-vs-big-data





# Big Data

## The 4 Vs

- Volume
- Variety
- Velocity
- Veracity



# Volume: How much data is out there?

## Every Day We Create 2.5 Quintillion Bytes of Data

*IBM study of 1,734 chief marketing officers from 64 countries*

This is a Press Release edited by StorageNewsletter.com on 2011.10.21

BOOKMARK

A new IBM Corp.'s study of more than 1,700 chief marketing officers from 64 countries and 19 industries reveals that the majority of the world's top marketing executives recognize a critical and permanent shift occurring in the way they engage with their customers, but question whether their marketing organizations are prepared to manage the change.


http://www.sciencedaily.com/releases/2013/05/130522085217.htm

## Big Data, for better or worse: 90% of world's data generated over last two years

*Date:* May 22, 2013

*Source:* SINTEF

*Summary:* A full 90 percent of all the data in the world has been generated over the last two years. Internet-based companies are awash with data that can be grouped and utilized. Is this a good thing?

**Share This**

- Email to a friend
- Facebook
- Twitter
- LinkedIn
- Google+
- Print this page

http://www.storagenewsletter.com/rubriques/market-reportsresearch/ibm-cmo-study/



# Variety: Types of Data

Structured data
- (relational) database tables
- CSV/TSV files

Semi-structured data
- XML, JSON, RDF

Unstructured data
- text data (documents, Web pages, short texts, e.g., social media)

Multimedia data
- images, videos, audios

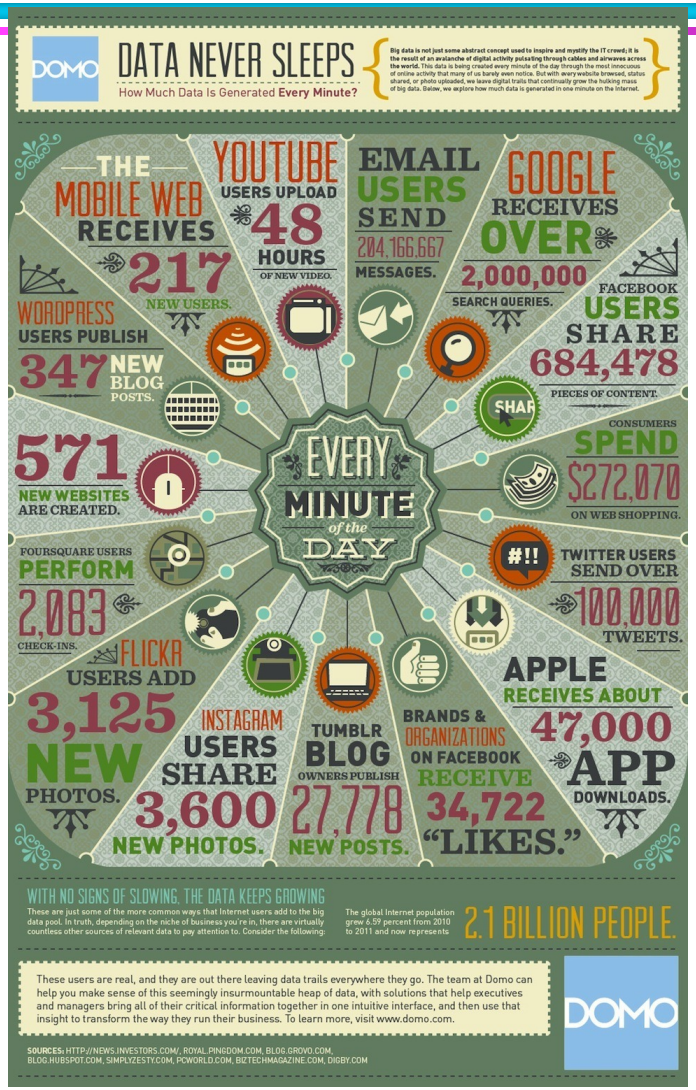
Other types of data
- matrices, graphs, sequences, time-series, spatio-temporal

8



# Velocity: Streaming Data

- Stock trades
- Highway sensors
- Weather data
- Social media
- Telephone calls
- Video streaming




http://mashable.com/2012/06/22/data-created-every-minute/

9



# Veracity: uncertain and imprecise data







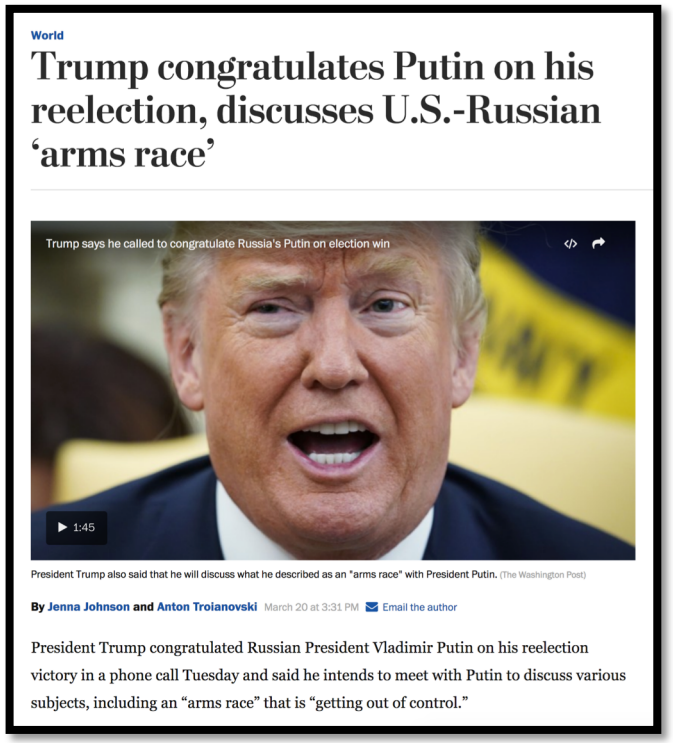
# Defending Democracy against organized attacks



"A Carnegie Mellon scholar, Dov H. Levin, ... 81 by the United States and 36 by the Soviet Union or Russia between 1946 and 2000"



# Best Comment Attack Pattern



World

Trump congratulates Putin on his reelection, discusses U.S.-Russian 'arms race'

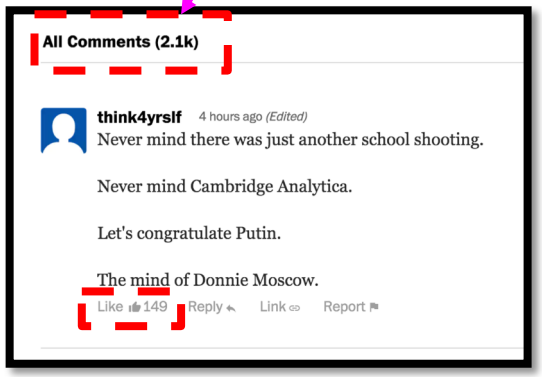


President Trump also said that he will discuss what he described as an "arms race" with President Putin. (The Washington Post)

By Jenna Johnson and Anton Troianovski March 20 at 3:31 PM Email the author

President Trump congratulated Russian President Vladimir Putin on his reelection victory in a phone call Tuesday and said he intends to meet with Putin to discuss various subjects, including an "arms race" that is "getting out of control."

The battle ground for Best comments




Herd behavior



# What is Data?

- Collection of ***data objects*** and their ***attributes***
- An ***attribute*** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an ***object***
  - Object is also known as record, point, case, sample, entity, or instance

**Attributes**

**Objects**

| *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 |

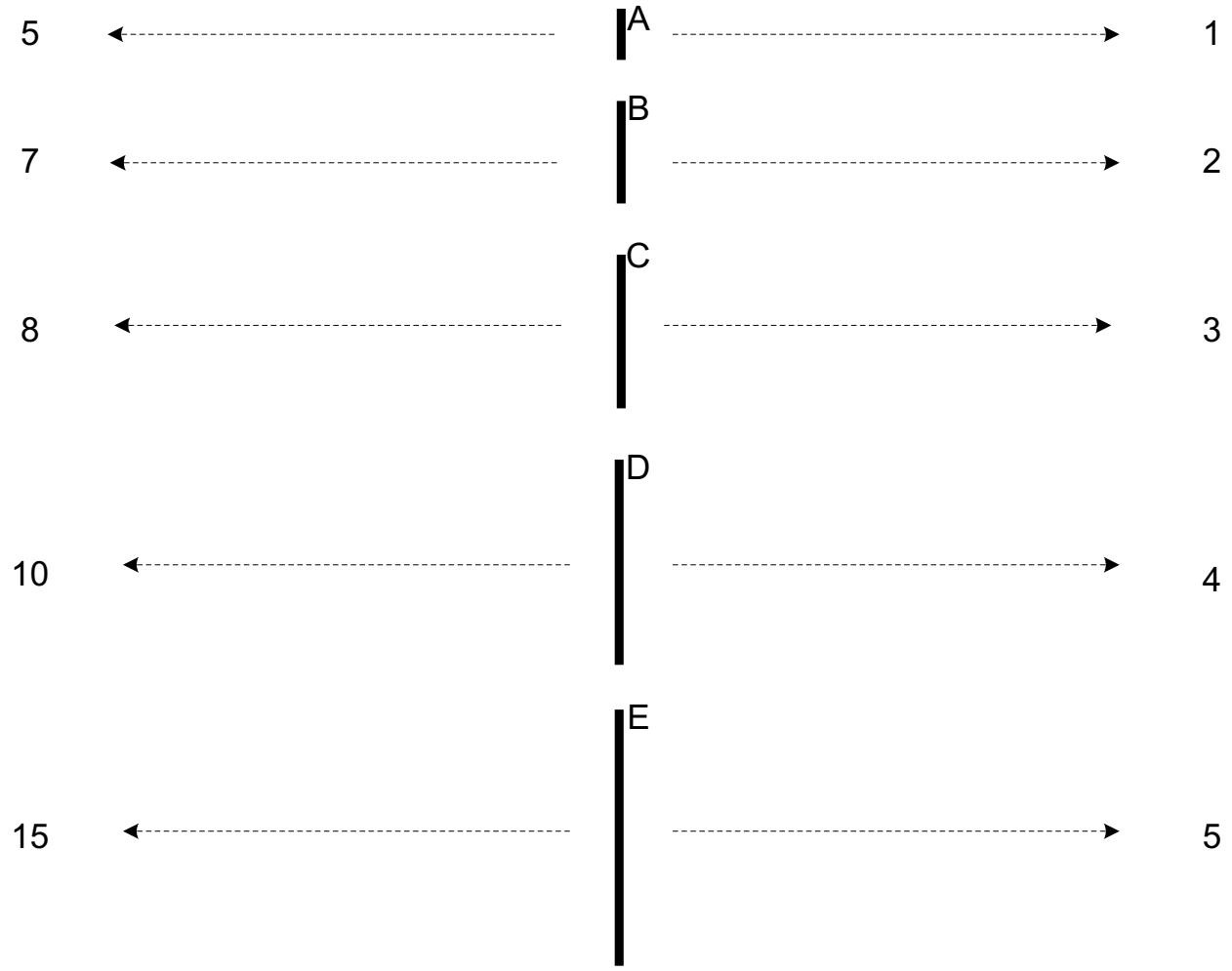
# Attribute Values

- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object

- Distinction between attributes and attribute values
  - 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 integers
    - But properties of attribute values can be different

# Measurement of Length

- The way you measure an attribute may not match the attributes properties.




**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 from 1-10), grades, height {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/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful : + -
  - Ratios are meaningful * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & meaningful differences
  - Ratio attribute: all 4 properties/operations

# 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 only distinguish. (=, ≠) | 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.,<br>*new_value* = *f(old_value)*<br>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*<br>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.
  - Note: 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 attribute value) is regarded as 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

# 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 only on the distribution
  - Many times what is meaningful is measured by statistical significance
  - But in the end, what is meaningful is measured by the domain

# Types of data sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data



# Important Characteristics of Data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data



# 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* by *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 |



# Document Data

- Each document becomes 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 |



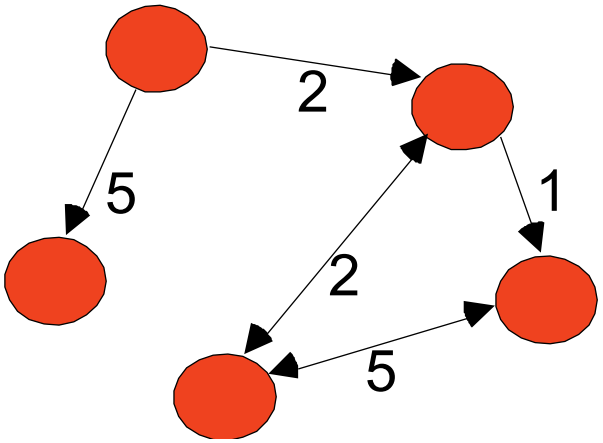
# Transaction Data

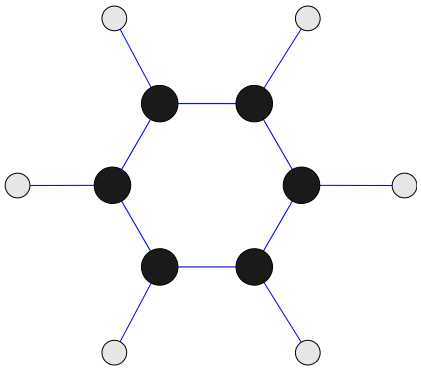
- A special type of record data, where
  - Each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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

# Graph Data

- Examples: Generic graph, a molecule, and webpages



Benzene Molecule: C6H6

**Useful Links:**

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

**Knowledge Discovery and Data Mining Bibliography**

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

**Book References in Data Mining and Knowledge Discovery**

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

**General Data Mining**

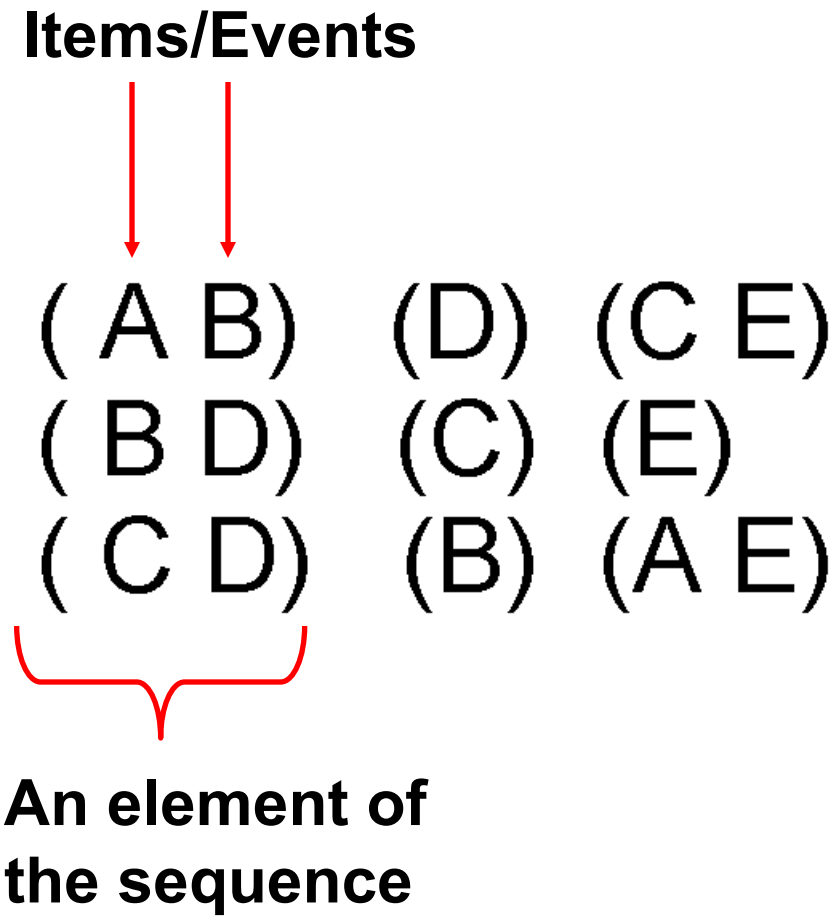
Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.



# Ordered Data

- Sequences of transactions






# Ordered Data

- Genomic sequence data

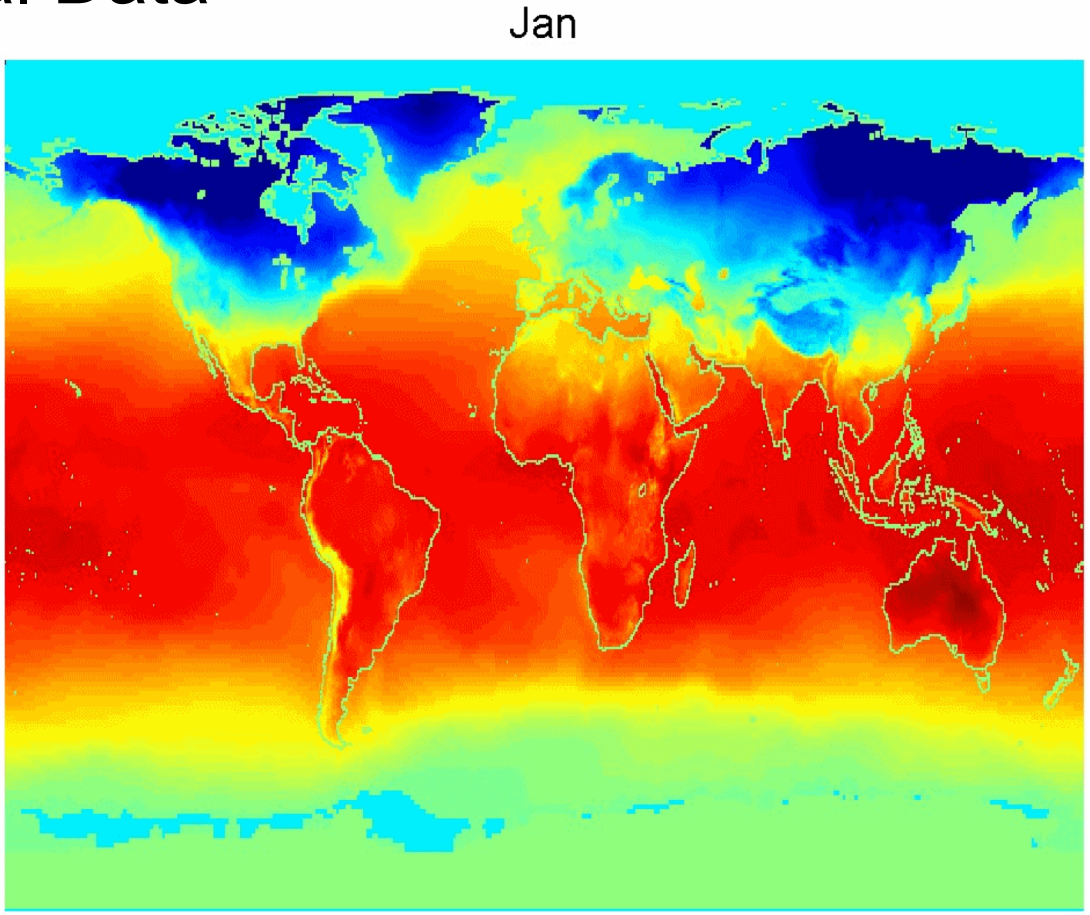
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```



# 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

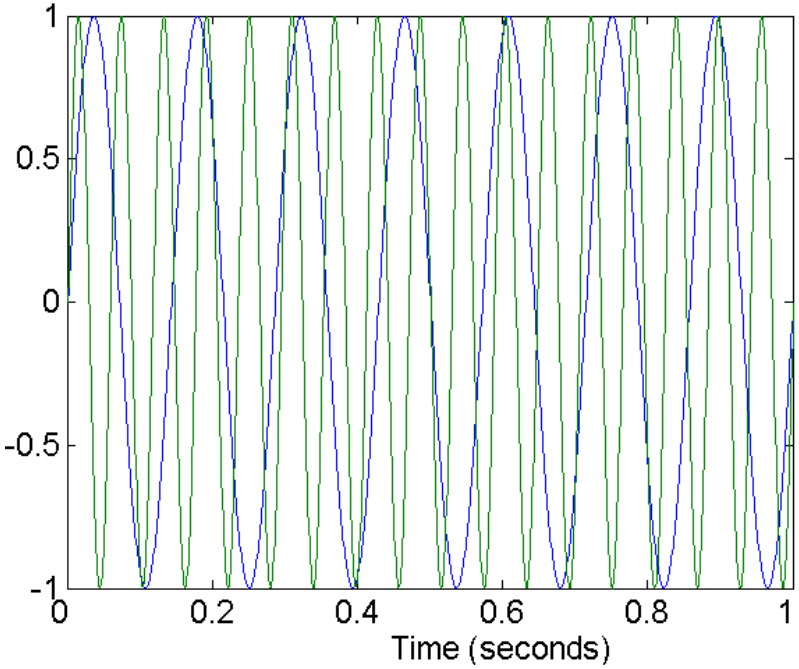
# Data Quality …

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

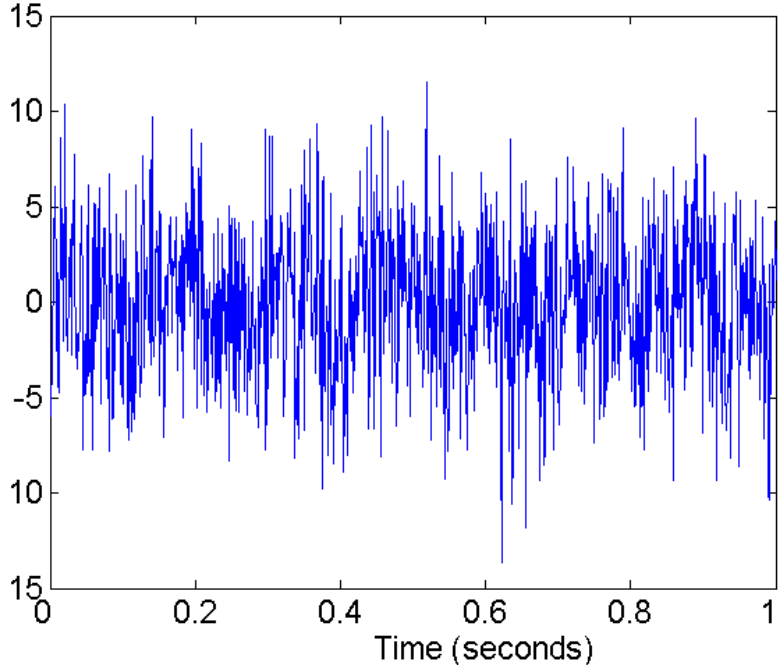
- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data

# Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



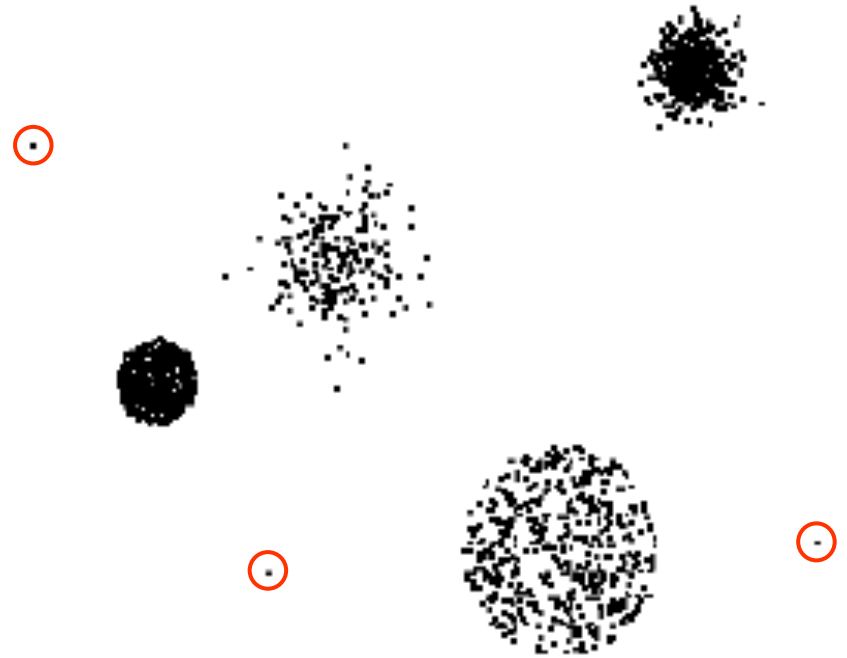

**Two Sine Waves**




**Two Sine Waves + Noise**

# 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



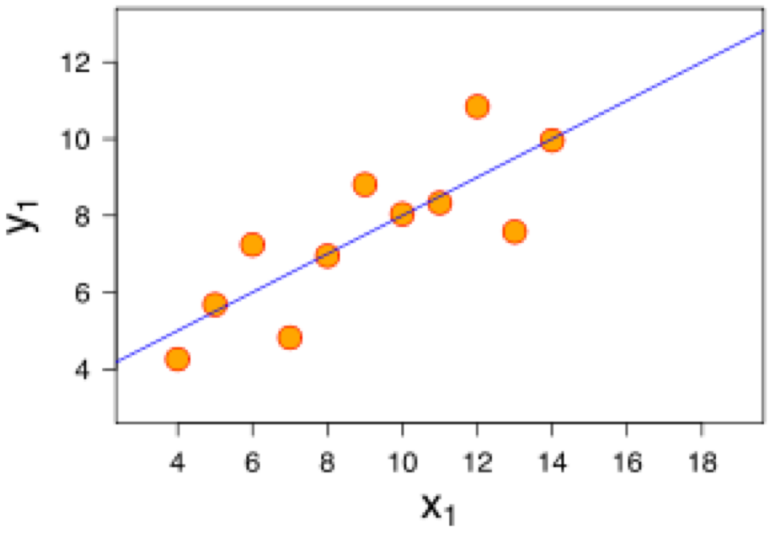
- Causes?

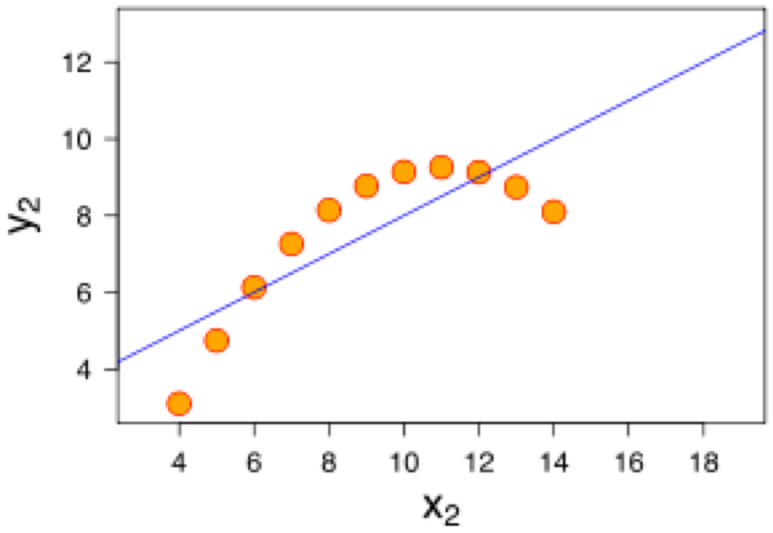


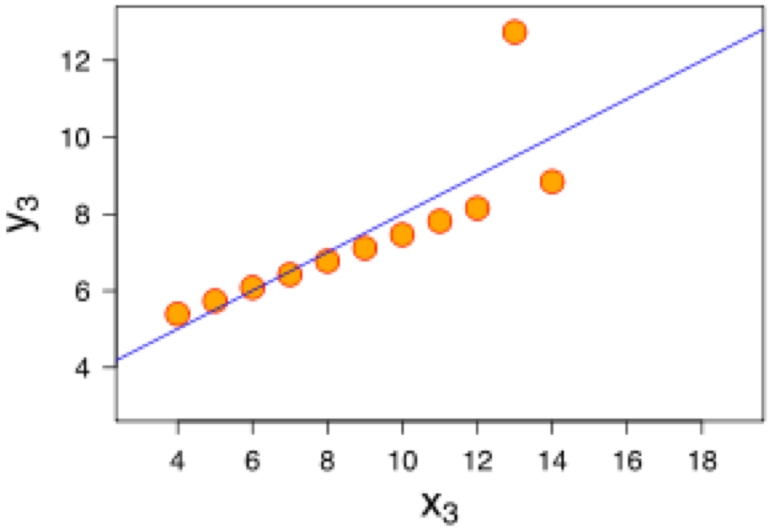
# Anscombe's Quartet

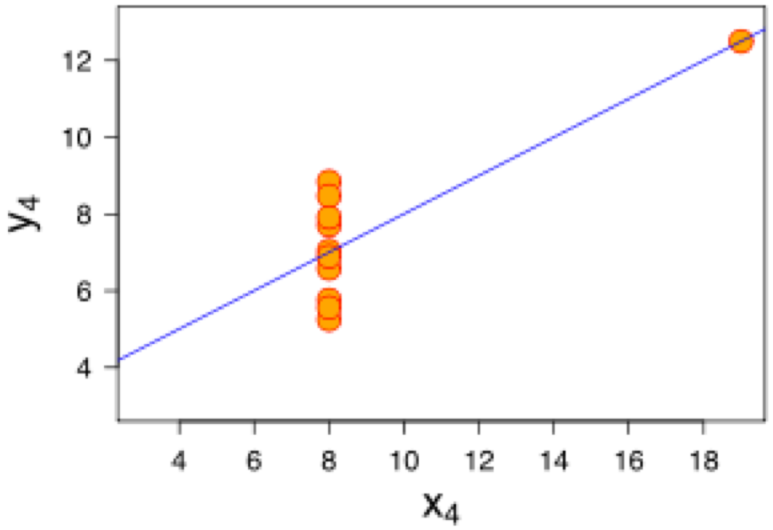
| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

| Property | Value |
|---|---|
| Mean of $x$ in each case | 9 (exact) |
| Sample variance of $x$ in each case | 11 (exact) |
| Mean of $y$ in each case | 7.50 (to 2 decimal places) |
| Sample variance of $y$ | 4.122 or 4.127 (to 3 decimal places) |
| Correlation between $x$ and $y$ | 0.816 (to 3 decimal places) |
| Linear regression | $y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively) |

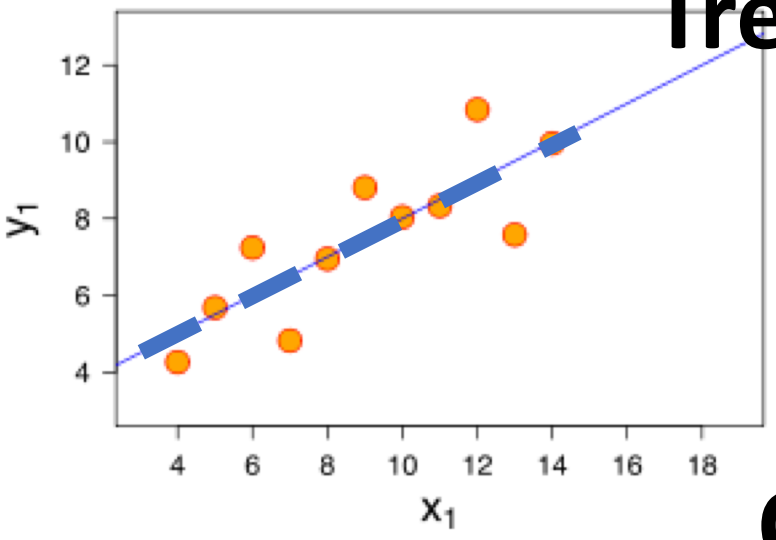

y1
x1
4
6
8
10
12
14
16
18
4
6
8
10
12

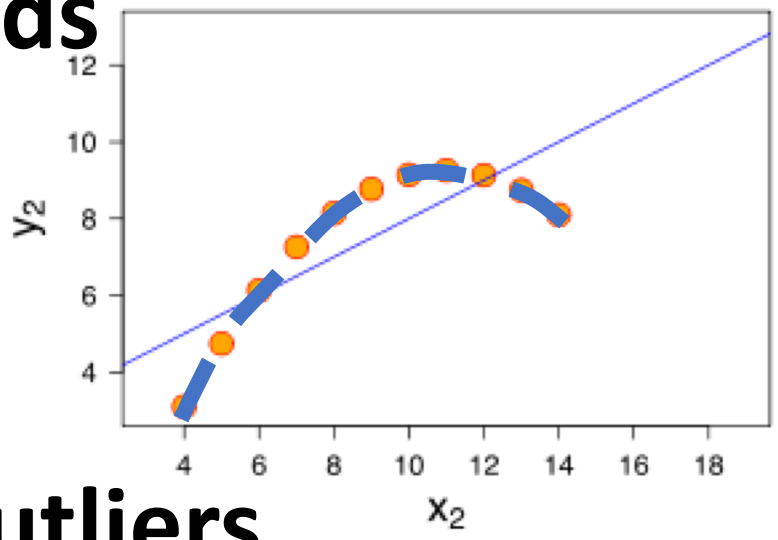

y2
x2
4
6
8
10
12
14
16
18
4
6
8
10
12


y3
x3
4
6
8
10
12
14
16
18
4
6
8
10
12

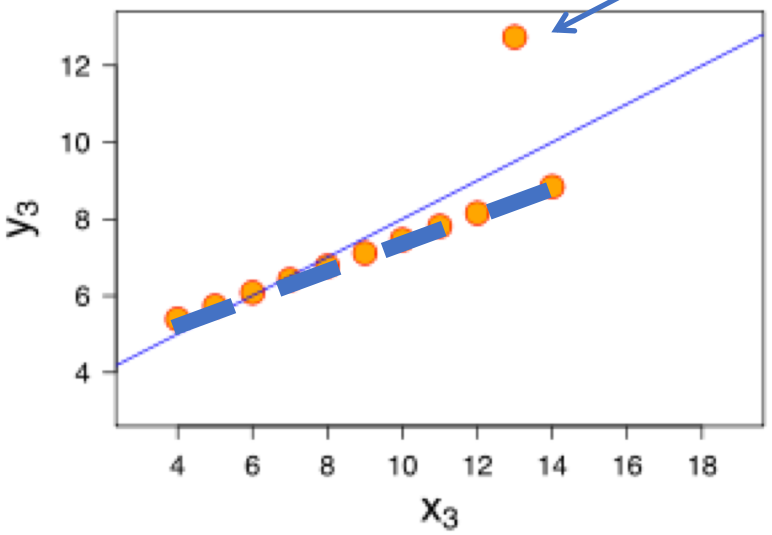

y4
x4
4
6
8
10
12
14
16
18
4
6
8
10
12

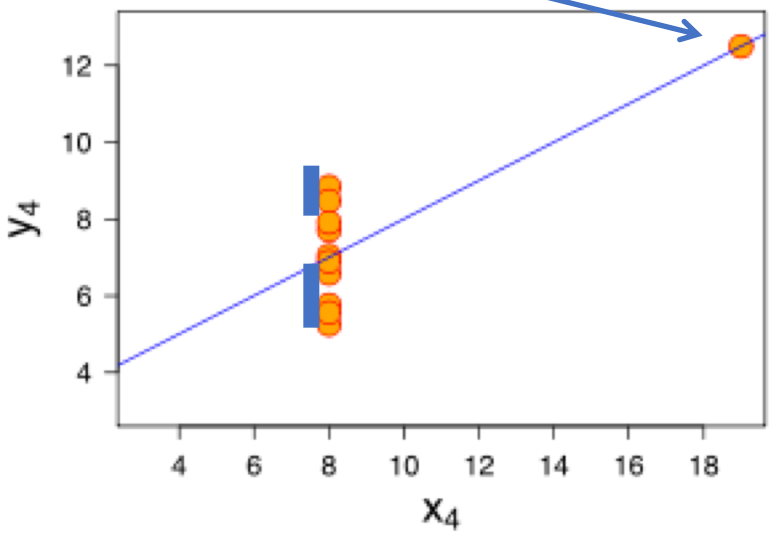
# Trends




# Outliers




**Get answers to the questions you didn't ask yet.**

# 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 variables
  - Estimate missing values
    - Example: time series of temperature
    - Example: census results
  - Ignore the missing value during analysis

# Duplicate Data

- Data set 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
  - Process of dealing with duplicate data issues
- When should duplicate data not be removed?

# Similarity and Dissimilarity Measures

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity measure
  - Numerical measure of how different two data objects are
  - 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

The following table shows the similarity and dissimilarity between two objects, $x$ and $y$, with respect to a single, simple attribute.

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



# Euclidean Distance

- Euclidean Distance

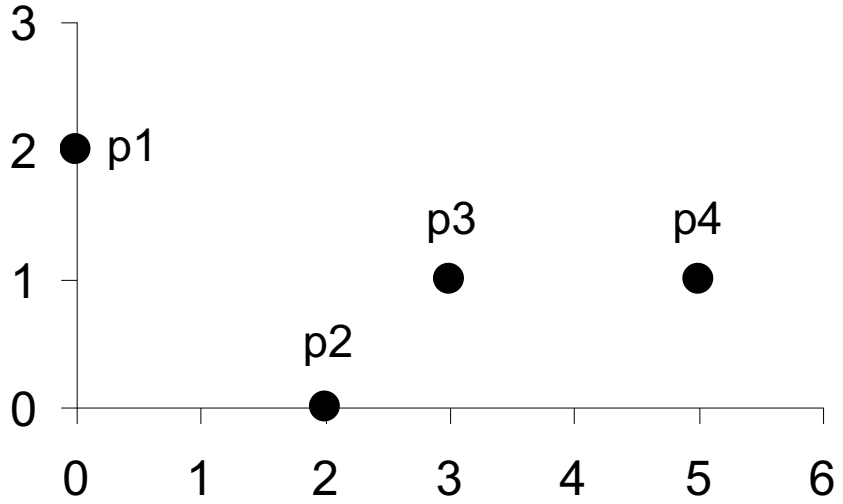
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where *n* is the number of dimensions (attributes) and $x_k$ and $y_k$ are, respectively, the $k^{th}$ attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

- 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

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where $r$ is a parameter, $n$ is the number of dimensions (attributes) and $x_k$ and $y_k$ are, respectively, the $k^{\text{th}}$ attributes (components) or data objects $\boldsymbol{x}$ and $\boldsymbol{y}$.

# Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, $L_1$ 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
- $r \rightarrow \infty$. "supremum" ($L_{max}$ norm, $L_\infty$ 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.

# 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 |

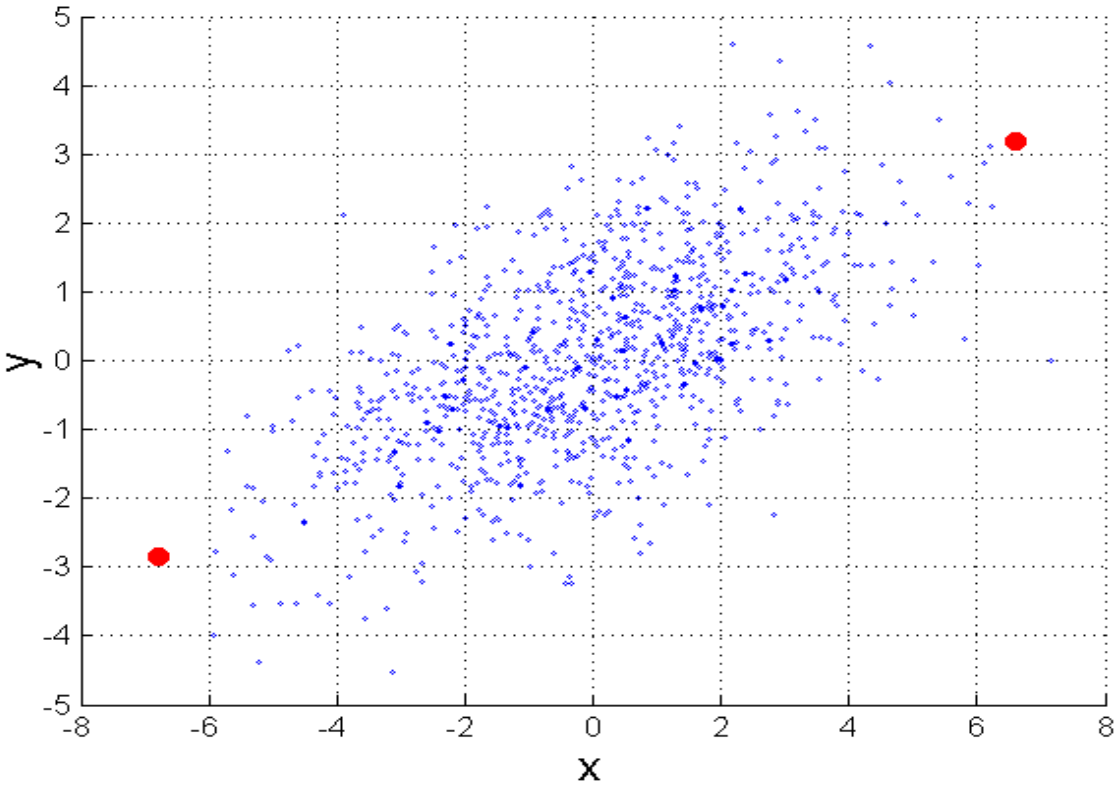
| L2 | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| p1 | 0 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0 | 2 |
| p4 | 5.099 | 3.162 | 2 | 0 |

| $L_\infty$ | 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

$$\mathbf{mahalanobis}(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})$$

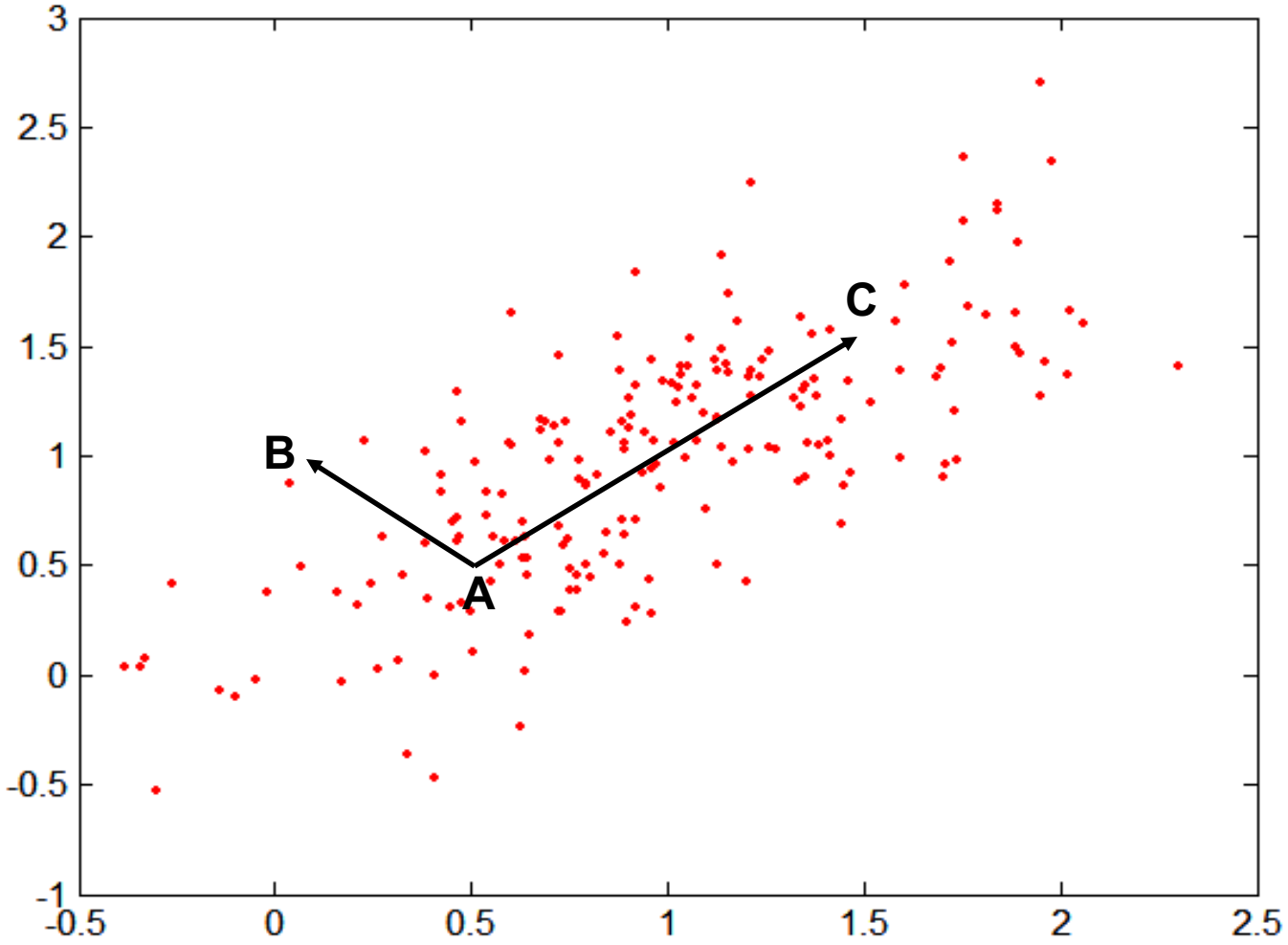



$\Sigma$ is the covariance matrix

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



# Mahalanobis Distance




**Covariance Matrix:**

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

**A: (0.5, 0.5)**

**B: (0, 1)**

**C: (1.5, 1.5)**

**Mahal(A,B) = 5**

**Mahal(A,C) = 4**



# Common Properties of a Distance

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

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $x$ and $y$ and $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$. (Positive definiteness)
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)
3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (Triangle Inequality)

where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

- A distance that satisfies these properties is a metric

# Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$.

2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

where $s(\mathbf{x}, \mathbf{y})$ is the similarity between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.



# Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes

- Compute similarities using the following quantities

  $f_{01}$ = the number of attributes where *p* was 0 and *q* was 1

  $f_{10}$ = the number of attributes where *p* was 1 and *q* was 0

  $f_{00}$ = the number of attributes where *p* was 0 and *q* was 0

  $f_{11}$ = the number of attributes where *p* was 1 and *q* was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  = $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

  J = number of 11 matches / number of non-zero attributes

  = $(f_{11}) / (f_{01} + f_{10} + f_{11})$



# SMC versus Jaccard: Example

**x** = 1 0 0 0 0 0 0 0 0 0

**y** = 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$ (the number of attributes where $p$ was 0 and $q$ was 1)

$f_{10} = 1$ (the number of attributes where $p$ was 1 and $q$ was 0)

$f_{00} = 7$ (the number of attributes where $p$ was 0 and $q$ was 0)

$f_{11} = 0$ (the number of attributes where $p$ was 1 and $q$ was 1)

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

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



# Cosine Similarity

- If $\mathbf{d}_1$ and $\mathbf{d}_2$ are two document vectors, then

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\| ,$$

where $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, $\mathbf{d}_1$ and $\mathbf{d}_2$, and $\| \mathbf{d} \|$ is the length of vector $\mathbf{d}$.

- Example:

$$\mathbf{d}_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$

$$\mathbf{d}_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

$\langle \mathbf{d}_1, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\| \mathbf{d}_1 \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\| \mathbf{d}_2 \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$



# Extended Jaccard Coefficient (Tanimoto)

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

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

# Correlation measures the linear relationship between objects

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12$$

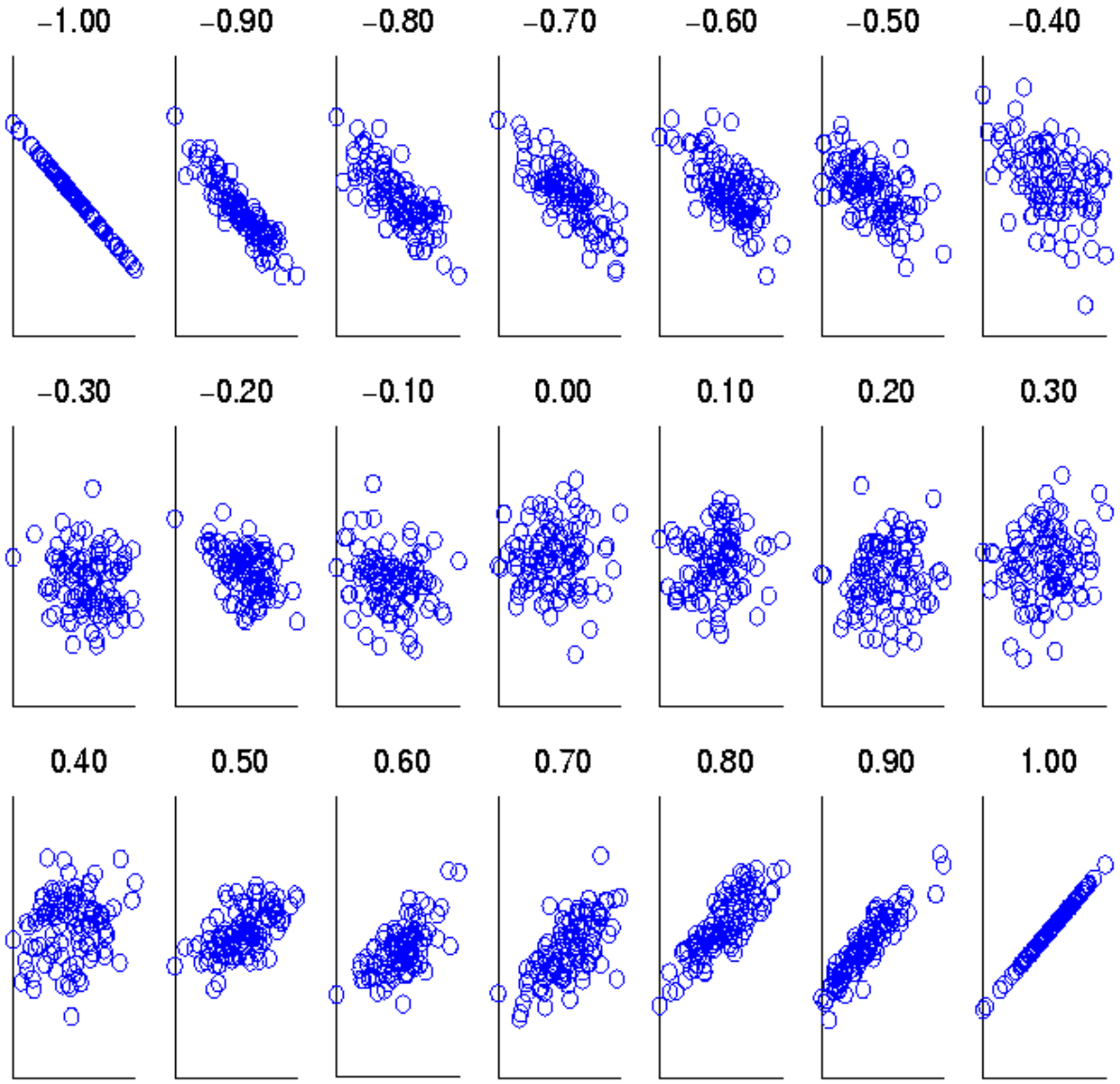
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

# Visually Evaluating Correlation




**Scatter plots showing the similarity from –1 to 1.**



# Drawback of Correlation

- **x** = (-3, -2, -1, 0, 1, 2, 3)
- **y** = (9, 4, 1, 0, 1, 4, 9)

$y_i = x_i^2$

- mean(**x**) = 0, mean(**y**) = 4
- std(**x**) = 2.16, std(**y**) = 3.74

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



# 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
- However, 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
  - And the number of observation in the $i^{\text{th}}$ 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.

1: For the $k^{\text{th}}$ attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range [0, 1].

2: Define an indicator variable, $\delta_k$, for the $k^{\text{th}}$ attribute as follows:

$\delta_k$ = 0 if the $k^{\text{th}}$ attribute is an asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing value for the kth attribute

$\delta_k$ = 1 otherwise

3. Compute $\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$

# Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use non-negative weights $\omega_k$
  - $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \omega_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \omega_k \delta_k}$

- Can also define a weighted form of distance

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

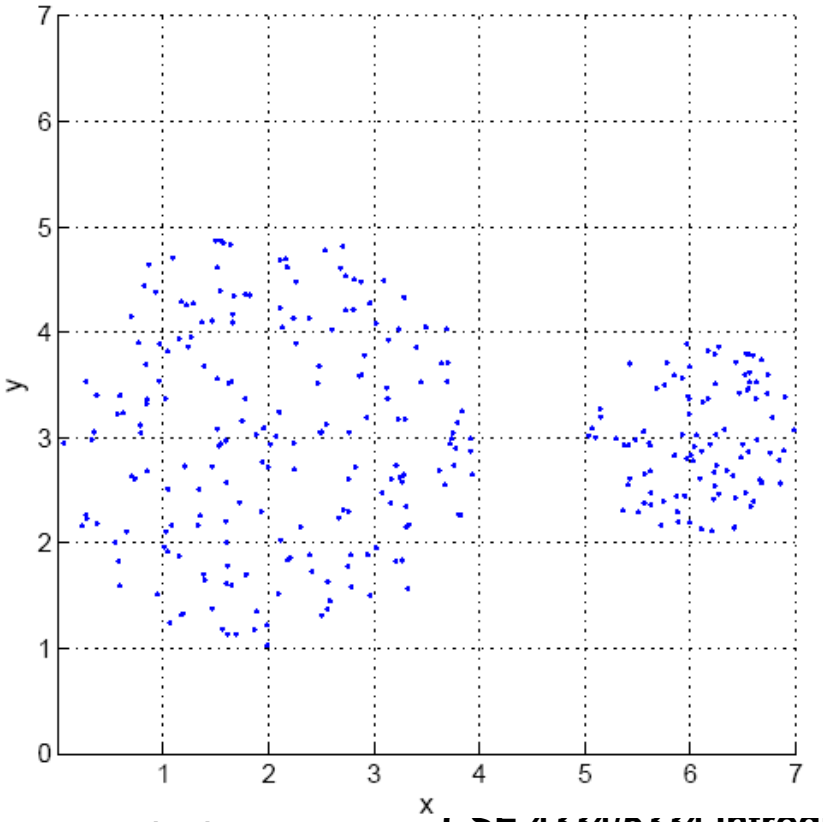


# Density

- Measures the degree to which data objects are close to each other in a specified area
- The notion of density is closely related to that of proximity
- Concept of density is typically used for clustering and anomaly detection
- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
    - Estimate what the distribution of the data looks like
  - Graph-based density
    - Connectivity

# Euclidean Density: Grid-based Approach

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains




| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 17 | 18 | 6 | 0 | 0 | 0 |
| 14 | 14 | 13 | 13 | 0 | 18 | 27 |
| 11 | 18 | 10 | 21 | 0 | 24 | 31 |
| 3 | 20 | 14 | 4 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# Euclidean Density: Center-Based

- Euclidean density is the number of points within a specified radius of the point

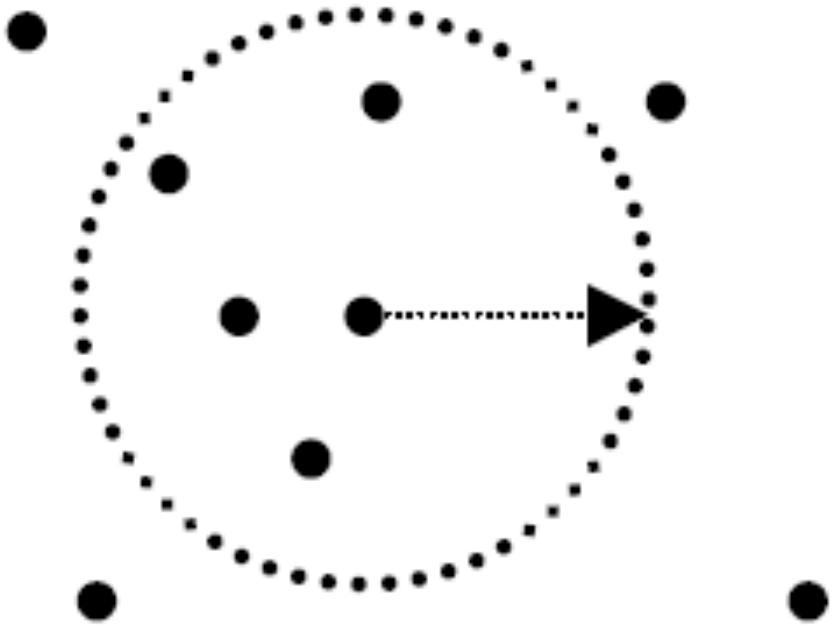


**Illustration of center-based density.**



# 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.
    - Days aggregated into weeks, months, or years
  - More “stable” data
    - Aggregated data tends to have less variability

# Example: Precipitation in Australia

- This example is based on precipitation in Australia from the period 1982 to 1993.

  The next slide shows

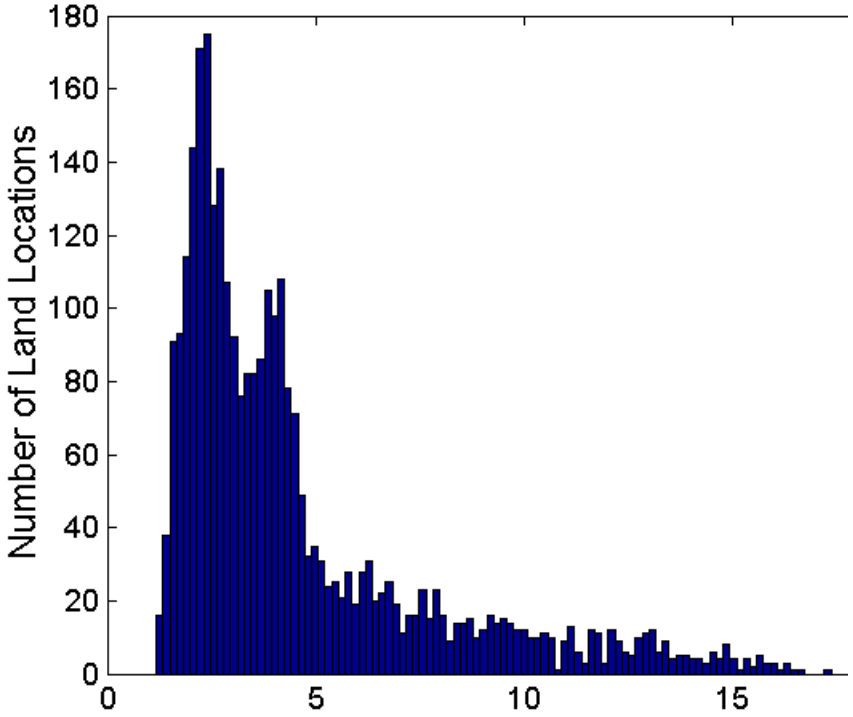
  - A histogram for the standard deviation of average monthly precipitation for 3,030 0.5◦ by 0.5◦ grid cells in Australia, and
  - A histogram for the standard deviation of the average yearly precipitation for the same locations.

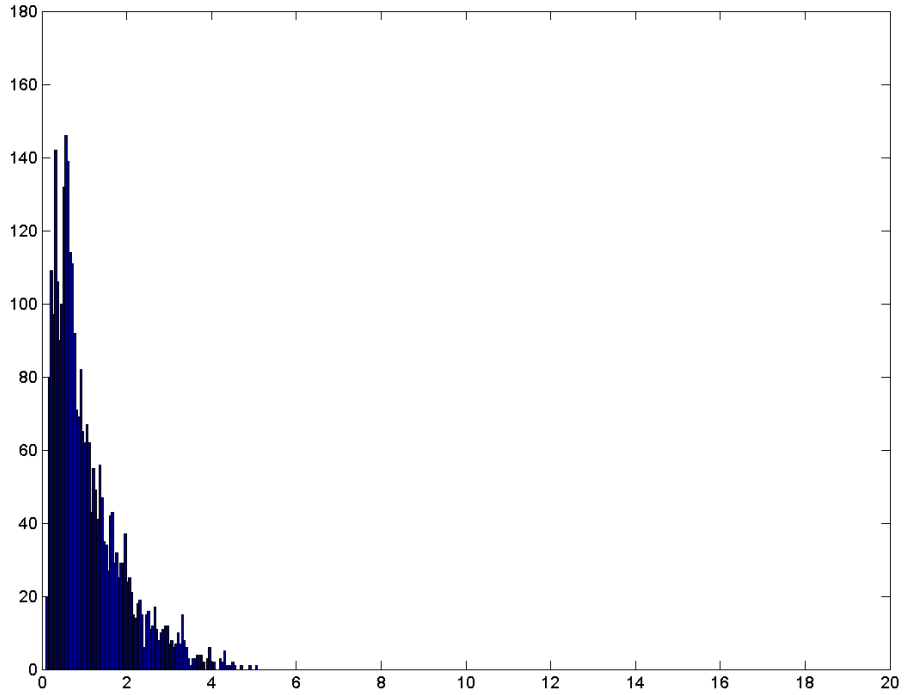
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.



# Example: Precipitation in Australia …

## Variation of Precipitation in Australia




**Standard Deviation of Average Monthly Precipitation**

**Standard Deviation of Average Yearly Precipitation**



# Sampling

- Sampling is the main technique employed for data reduction.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.

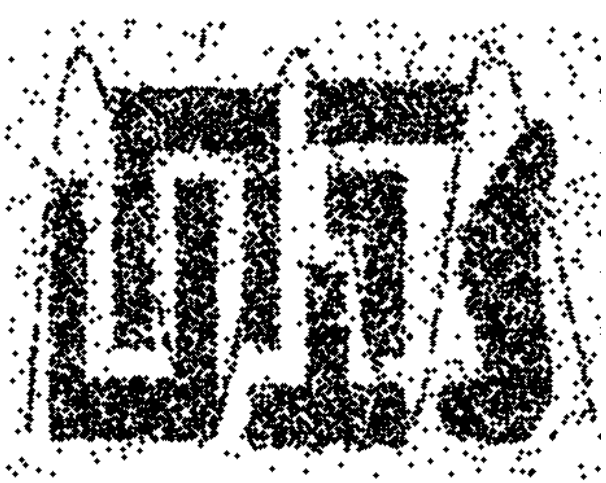
- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

# Sampling …

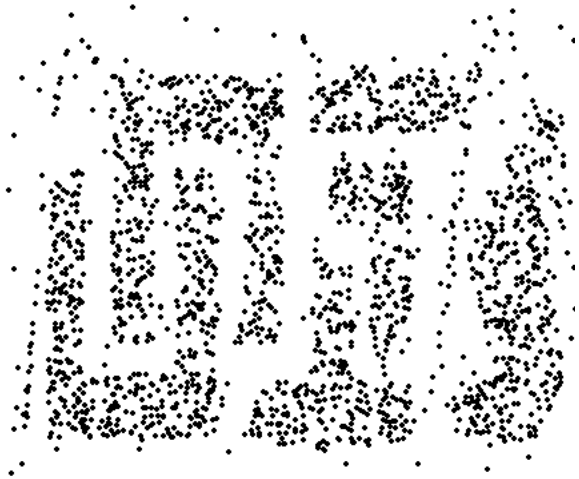
- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, if the sample is representative
  - A sample is representative if it has approximately the same properties (of interest) as the original set of data



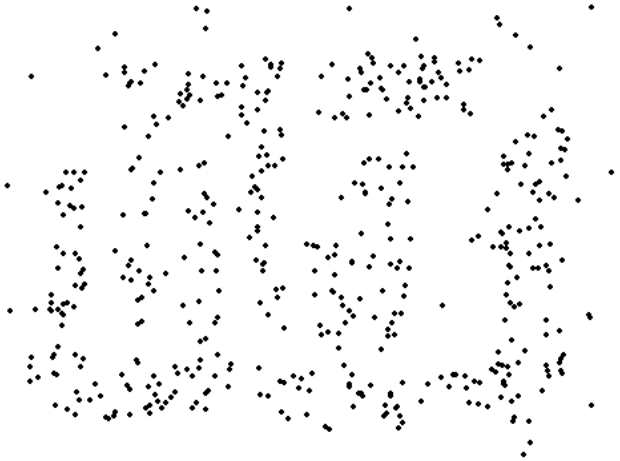
# Sample Size



8000 points



2000 Points



500 Points

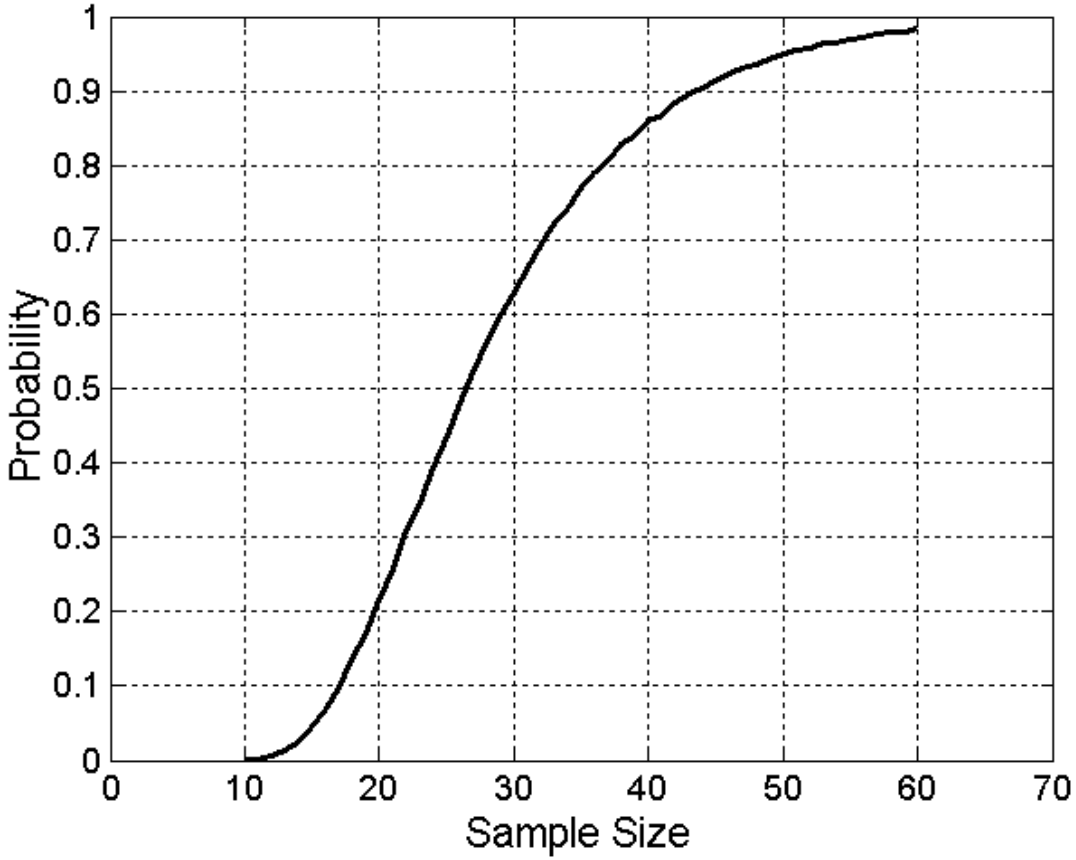


# Types of Sampling

- 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.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

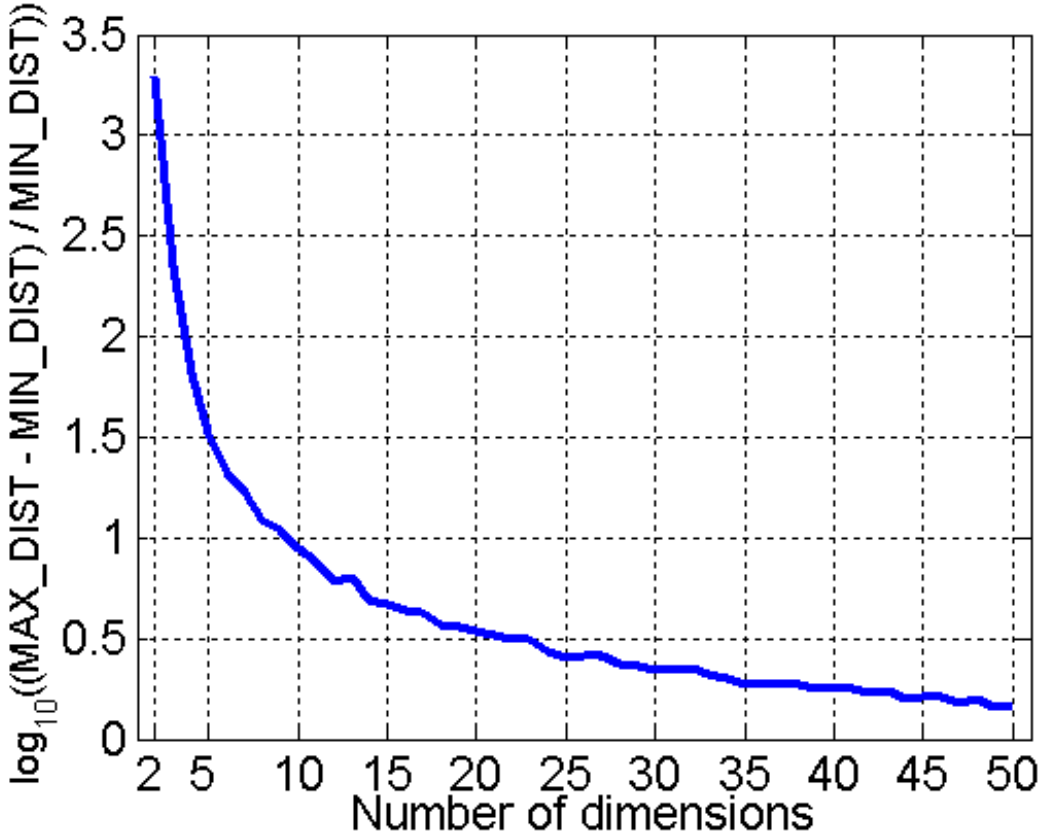
# Sample Size

- What sample size is necessary to get at least one object from each of 10 equal-sized 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 are 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**

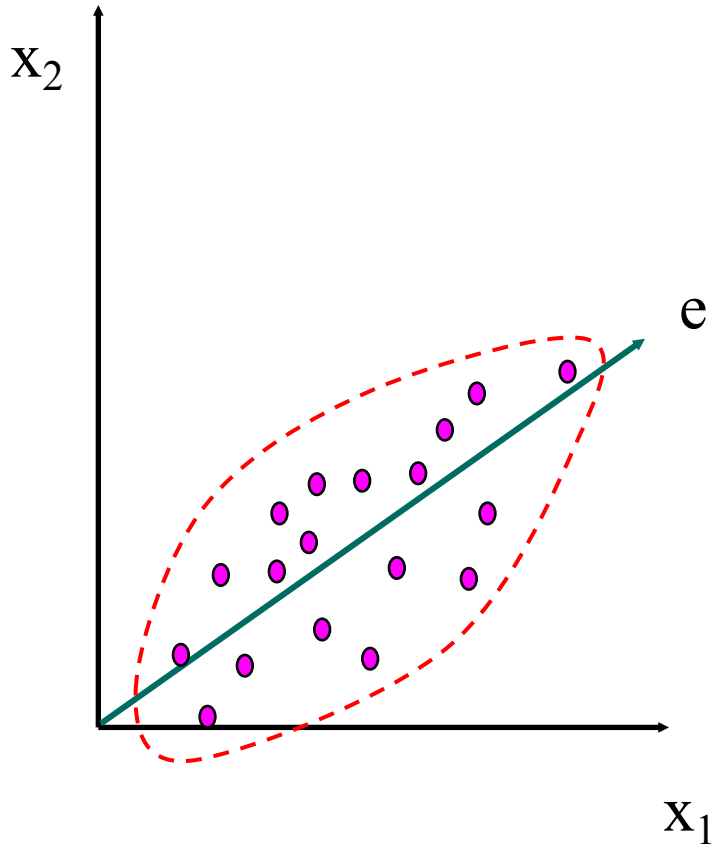


# Dimensionality Reduction

- 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
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

# Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data






# Dimensionality Reduction: PCA




# Feature Subset Selection

- Another way 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' GPA
- Many techniques developed, especially for classification



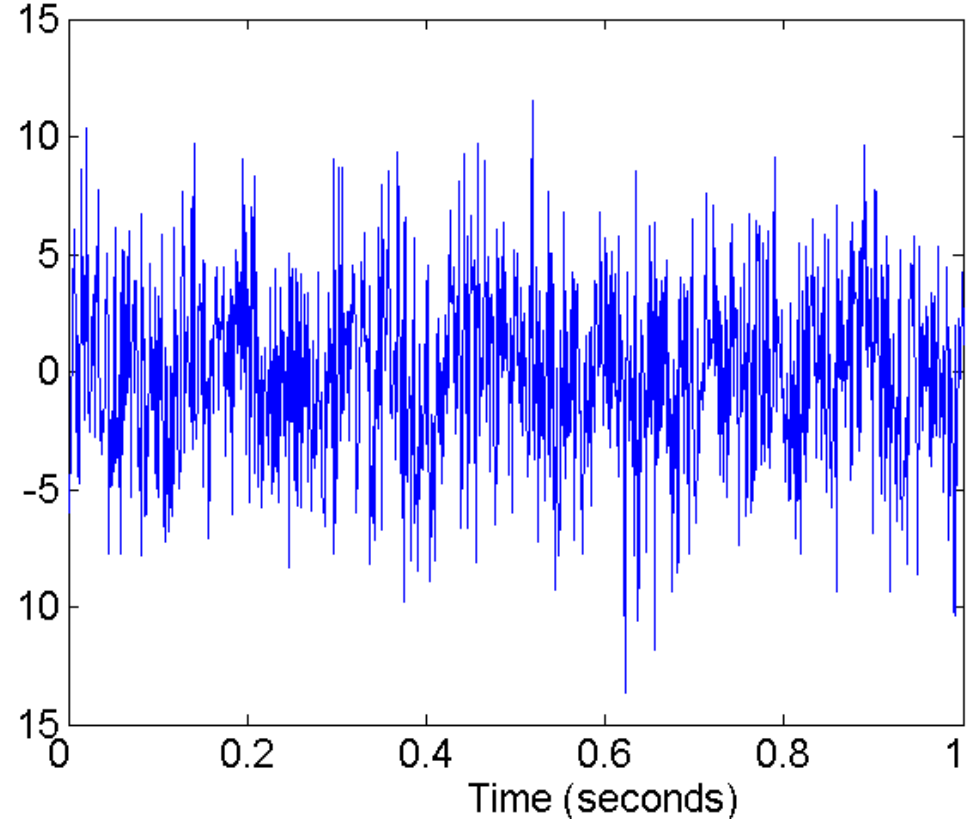
# Feature Creation

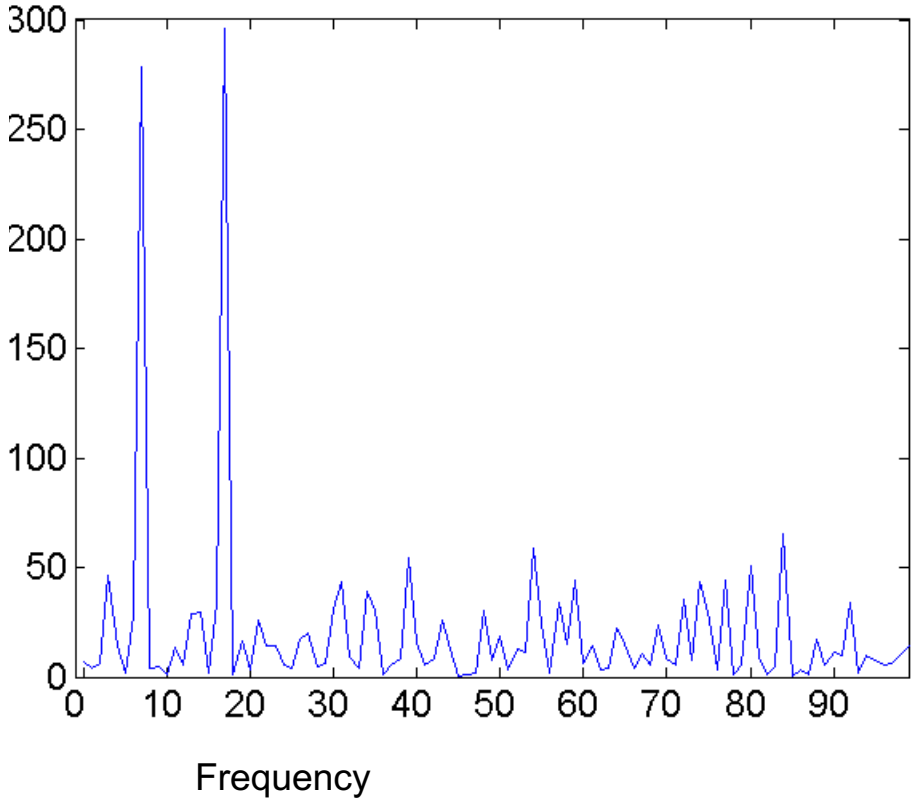
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

# Mapping Data to a New Space

- Fourier and wavelet transform




**Two Sine Waves + Noise** **Frequency**



# Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent and dependent variables have only a few values
  - We give an illustration of the usefulness of discretization using the Iris data set

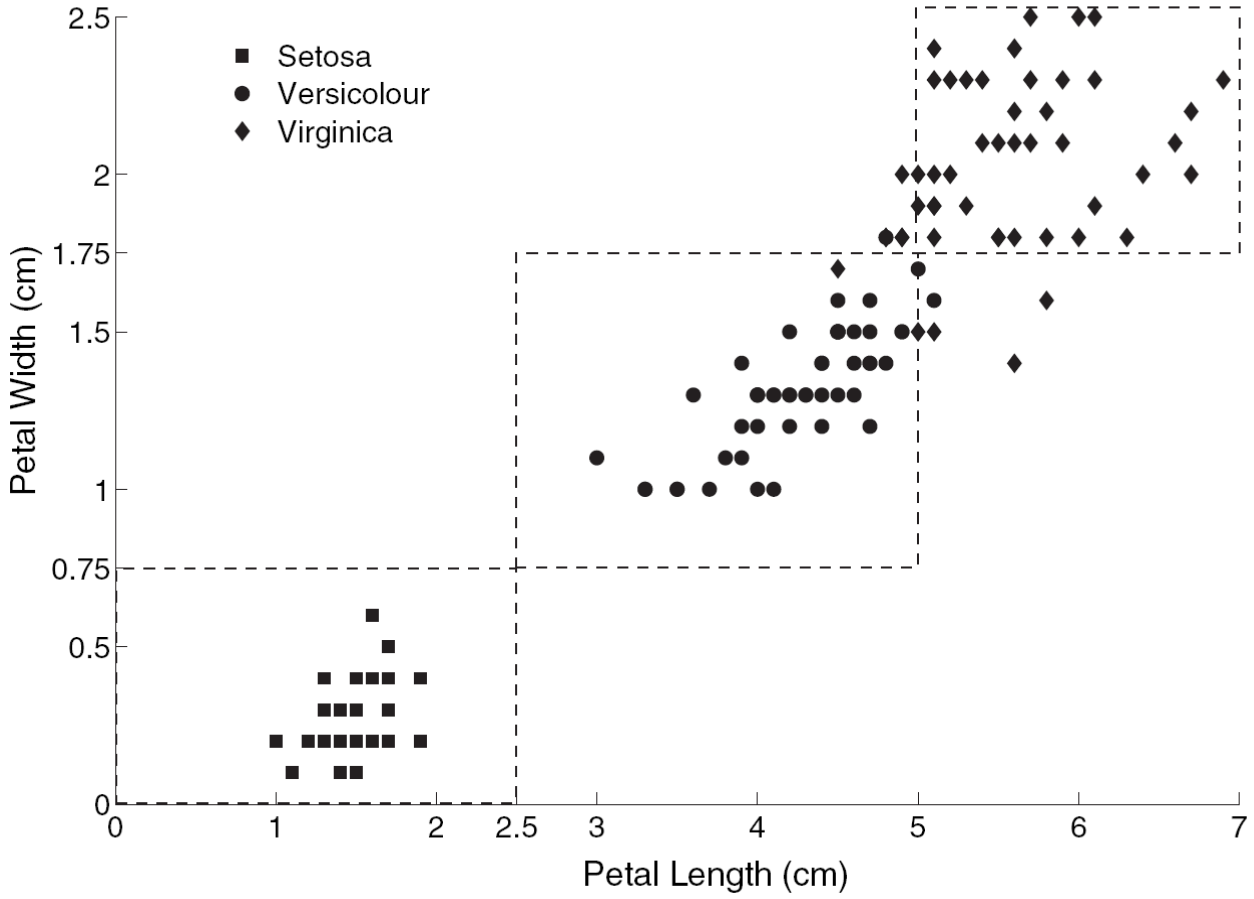


# Iris Sample Data Set

- Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Versicolour
    - Virginica
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

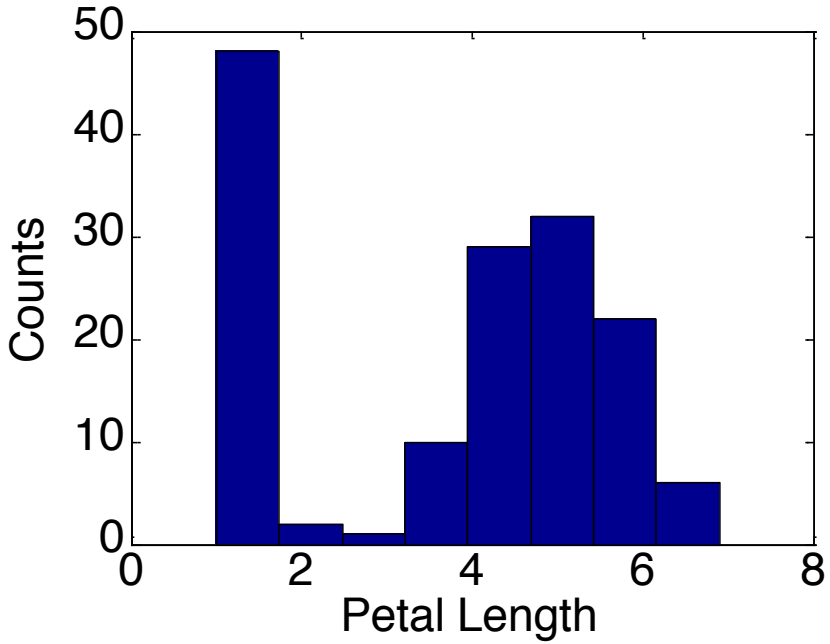
# Discretization: Iris Example




Petal width low or petal length low implies Setosa.
Petal width medium or petal length medium implies Versicolour.
Petal width high or petal length high implies Virginica.

# Discretization: Iris Example …

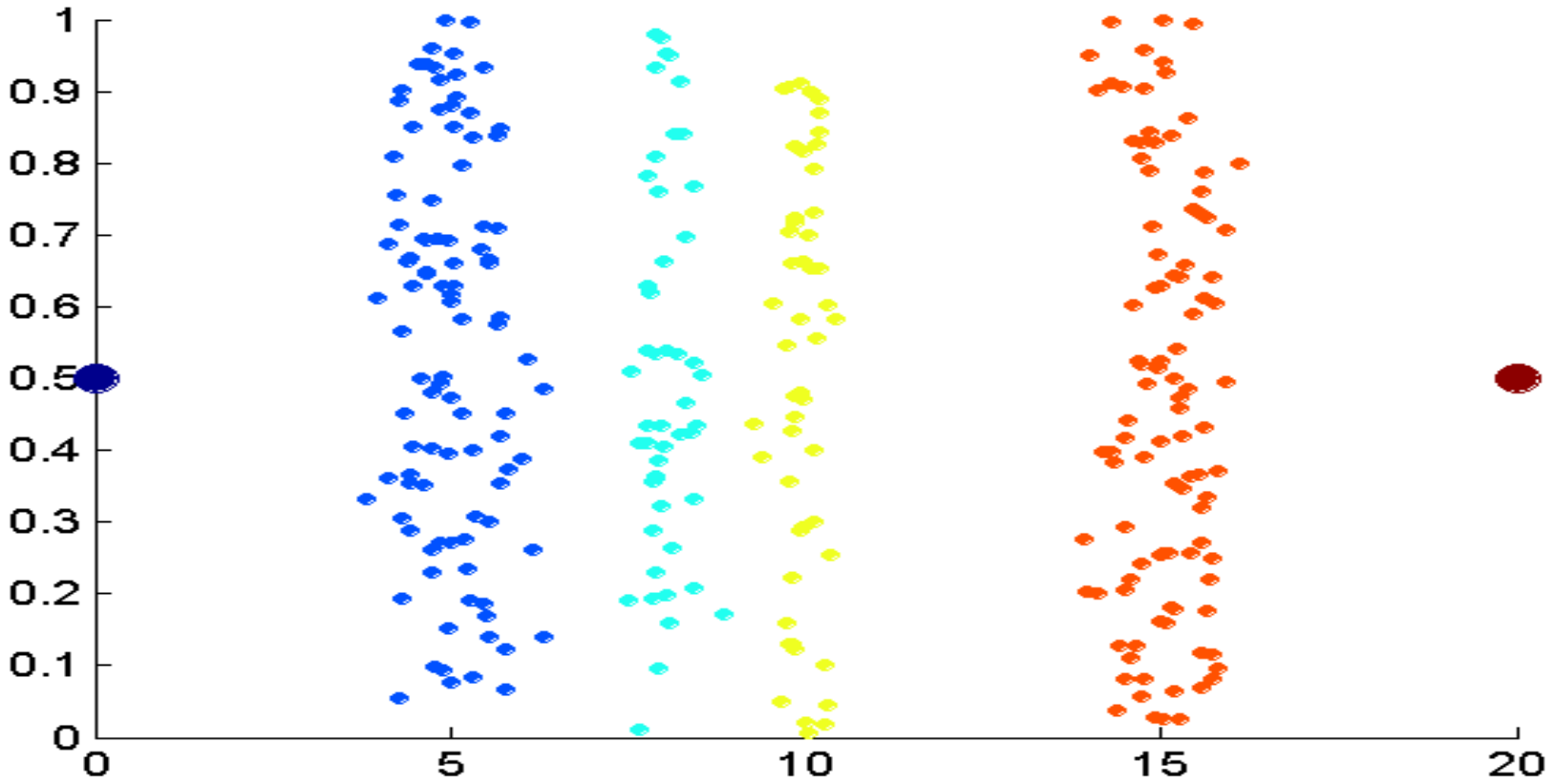
- How can we tell what the best discretization is?
  - Unsupervised discretization: find breaks in the data values
    - Example: Petal Length




  - Supervised discretization: Use class labels to find breaks



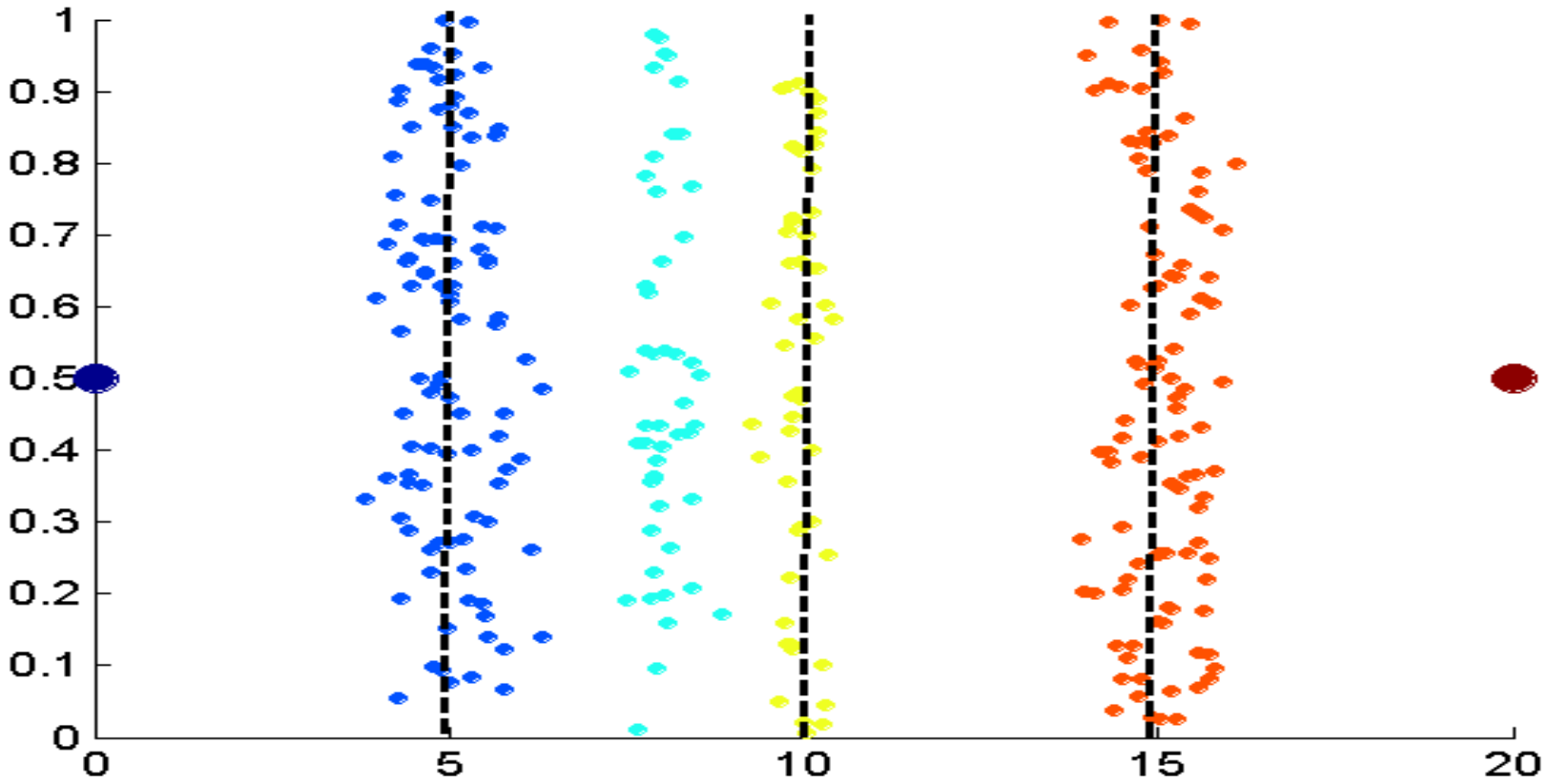
# Discretization Without Using Class Labels




Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.



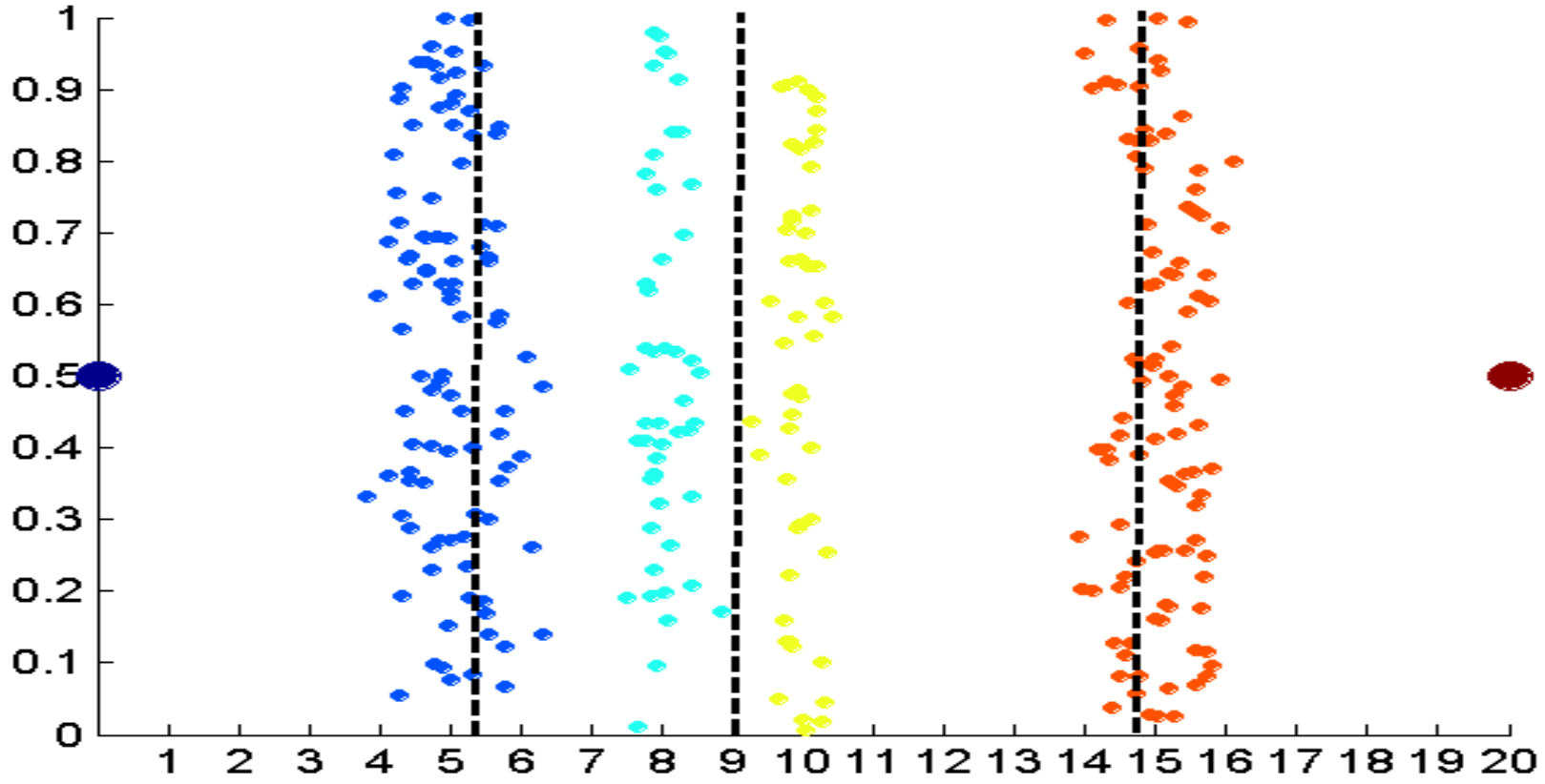
# Discretization Without Using Class Labels




**Equal interval width approach used to obtain 4 values.**



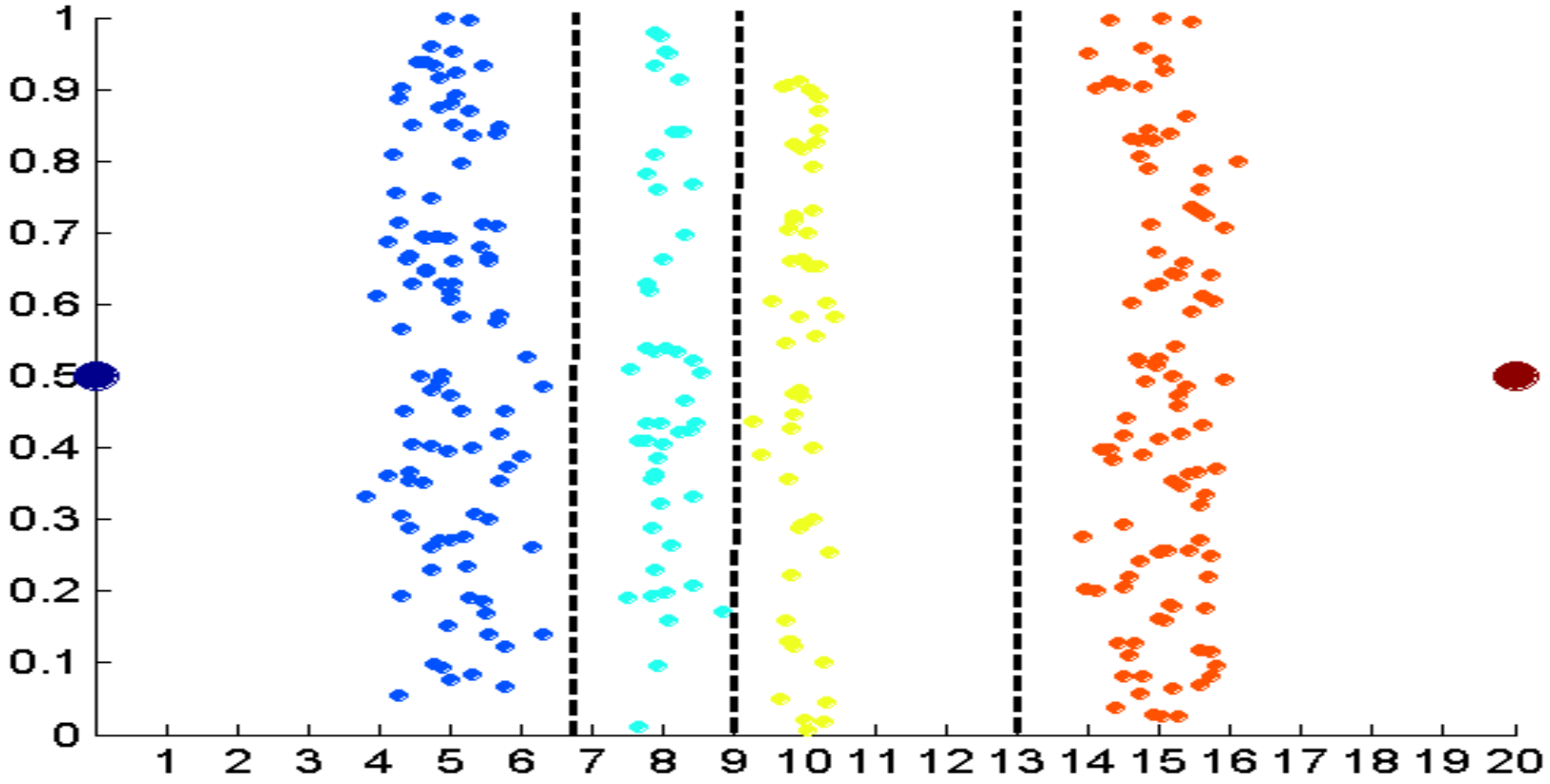
# Discretization Without Using Class Labels




**Equal frequency approach used to obtain 4 values.**



# Discretization Without Using Class Labels




**K-means approach to obtain 4 values.**



# Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables
- Typically used for association analysis
- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

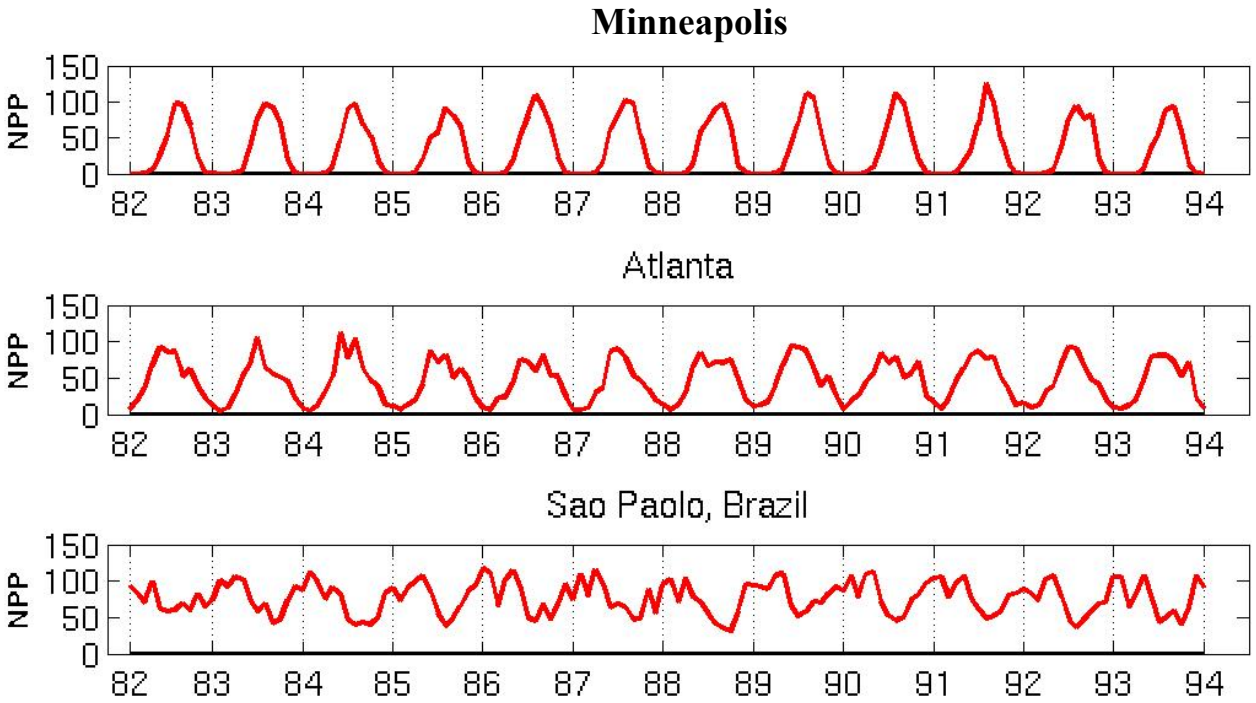


# Attribute Transformation

- An attribute transform is 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|
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation



# Example: Sample Time Series of Plant Growth



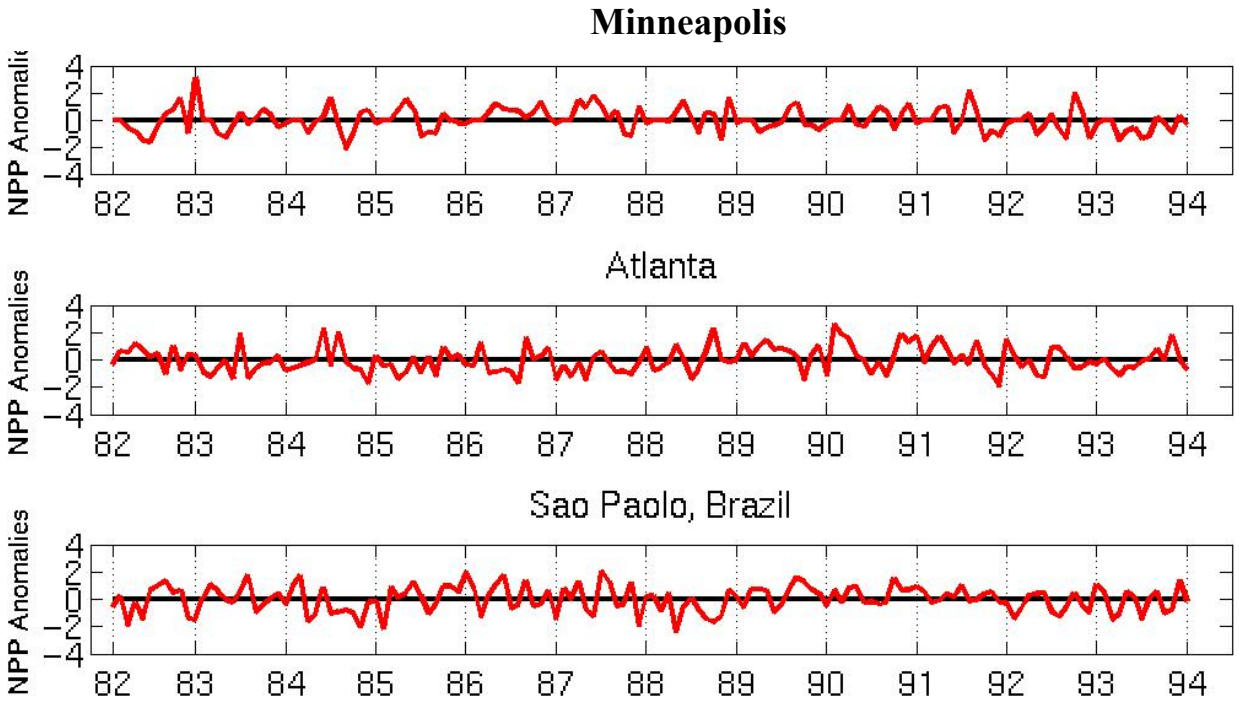

**Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.**

## Correlations between time series

| | Minneapolis | Atlanta | Sao Paolo |
|---|---|---|---|
| Minneapolis | 1.0000 | 0.7591 | -0.7581 |
| Atlanta | 0.7591 | 1.0000 | -0.5739 |
| Sao Paolo | -0.7581 | -0.5739 | 1.0000 |



# Seasonality Accounts for Much Correlation




Normalized using monthly Z Score:

Subtract off monthly mean and divide by monthly standard deviation

**Correlations between time series**

| | Minneapolis | Atlanta | Sao Paolo |
|---|---|---|---|
| Minneapolis | 1.0000 | 0.0492 | 0.0906 |
| Atlanta | 0.0492 | 1.0000 | -0.0154 |
| Sao Paolo | 0.0906 | -0.0154 | 1.0000 |