

# Introduction to Web Data Mining: Data

# Outline

---

- ▶ Attributes and Objects
- ▶ Types of Data
- ▶ Data Quality
- ▶ Data Preprocessing

# Quick Questions

---

- ▶ What are the most time consuming part in DM?

Data Preprocessing

- ▶ What are the most important steps for finishing a given DM task?

Data Understanding and Preprocessing

Domain Knowledge Discovery

Visualization

Algorithm

# Simple Comparison

---

Medical Care		Data Mining	
First Concern		First Concern	
Patient & symptoms	<b>Medicine</b>	Data & Applications	<b>Algorithms</b>

# What is Data

- ▶ Collection of data objects and their attributes (属性)
- ▶ 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, or feature
- ▶ A collection of attributes describe an object
  - ▶ 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

# Attributes

---

- ▶ **Attribute** (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - ▶ *E.g., customer \_ID, name, address*

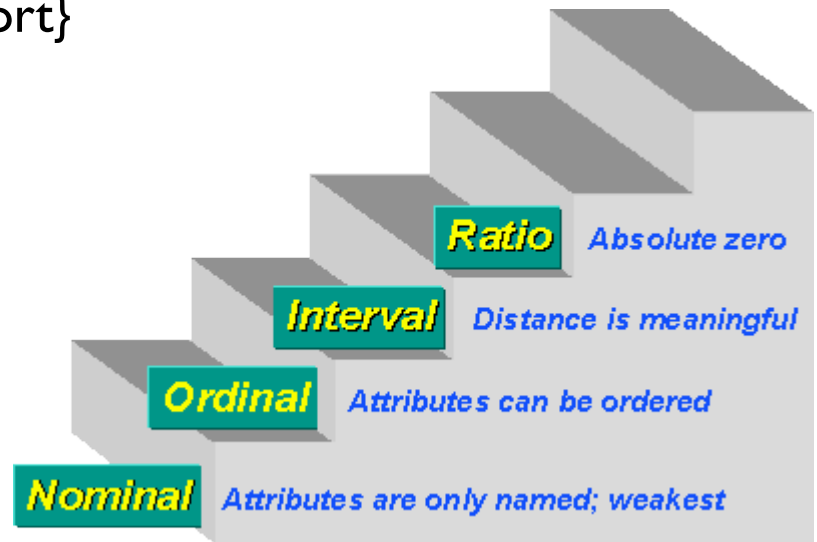
# Attribute Values

---

- ▶ 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

# Types of Attributes

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





# Discrete and Continuous Attributes

---

## ▶ Discrete Attribute

- ▶ Has only a finite or countably infinite set of values
  - ▶ E.g., 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
  - ▶ Nominal attribute with only 2 states (0 and 1)
  - ▶ Symmetric binary: both outcomes equally important
    - e.g., gender
  - ▶ Asymmetric binary: outcomes not equally important
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

## ▶ Continuous Attribute

- ▶ Has real numbers as attribute values
  - ▶ E.g., 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.

# Important Characteristics of Structured Data

---

- ▶ **Dimensionality**
  - ▶ Curse of Dimensionality
- ▶ **Sparsity**
  - ▶ Only presence counts
- ▶ **Resolution**
  - ▶ Patterns depend on the scale

# Types of data sets

---

- ▶ **Record**
  - ▶ Relational records
  - ▶ Data matrix, e.g., numerical matrix, crosstabs
  - ▶ Document data: text documents: term-frequency vector
  - ▶ Transaction data
- ▶ **Graph and network**
  - ▶ World Wide Web
  - ▶ Social or information networks
  - ▶ Molecular Structures
- ▶ **Ordered**
  - ▶ Video data: sequence of images
  - ▶ Temporal data: time-series
  - ▶ Sequential Data: transaction sequences
  - ▶ Genetic sequence data
- ▶ **Spatial, image and multimedia:**
  - ▶ Spatial data: maps
  - ▶ Image data:
  - ▶ Video data:

# Record Data

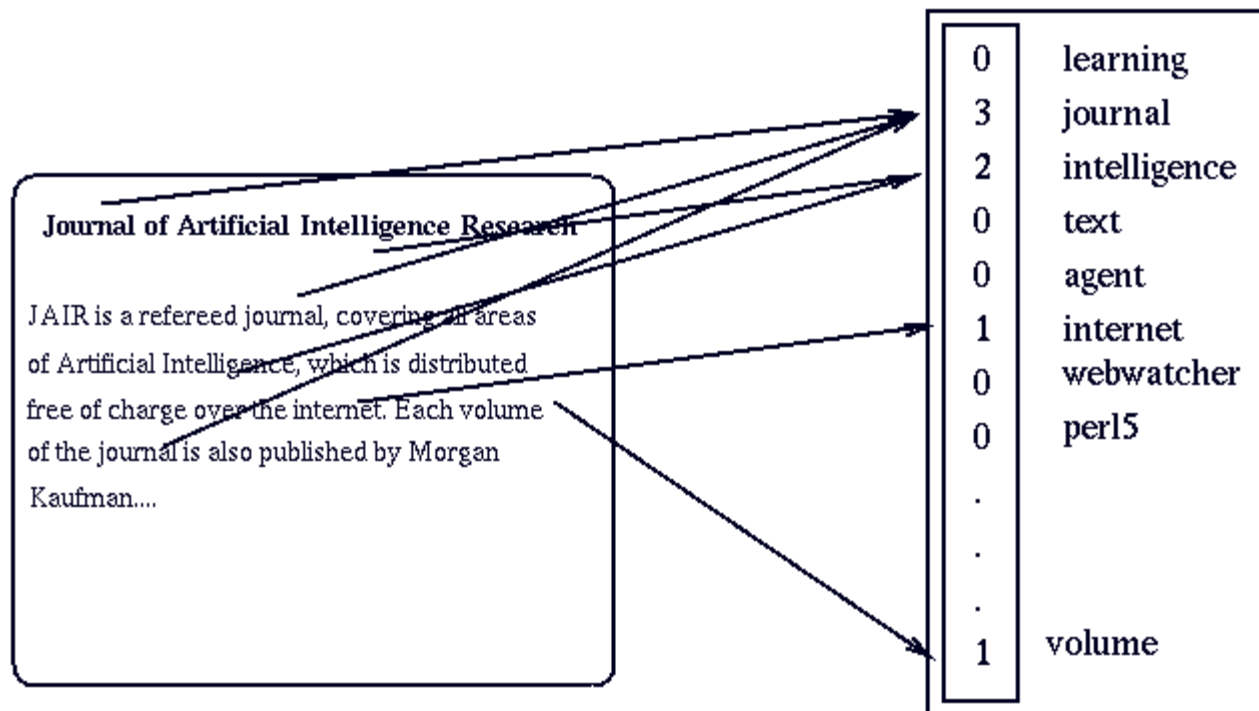
---

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

<i>Tid</i>	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

# Document Data

---



Bag-of-words

# 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

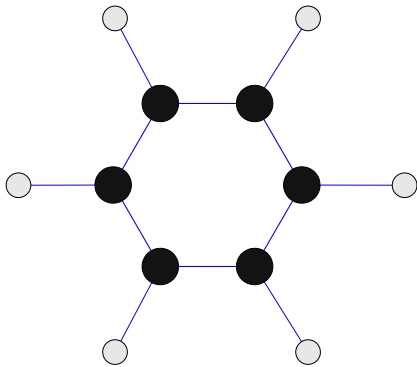
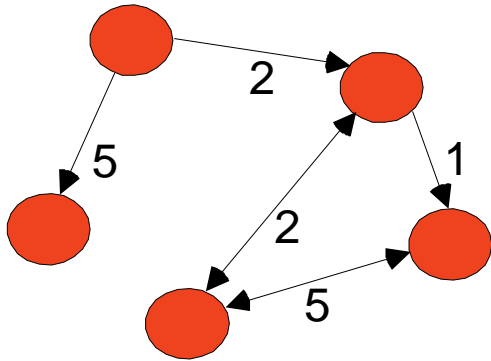
---

- ▶ 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.

<i><b>TID</b></i>	<i><b>Items</b></i>
<b>1</b>	<b>Bread, Coke, Milk</b>
<b>2</b>	<b>Beer, Bread</b>
<b>3</b>	<b>Beer, Coke, Diaper, Milk</b>
<b>4</b>	<b>Beer, Bread, Diaper, Milk</b>
<b>5</b>	<b>Coke, Diaper, Milk</b>

# Graph Data

## ► Examples: Generic graph, a Molecule, and Webpages



Benzene Molecule: C<sub>6</sub>H<sub>6</sub>

### 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.

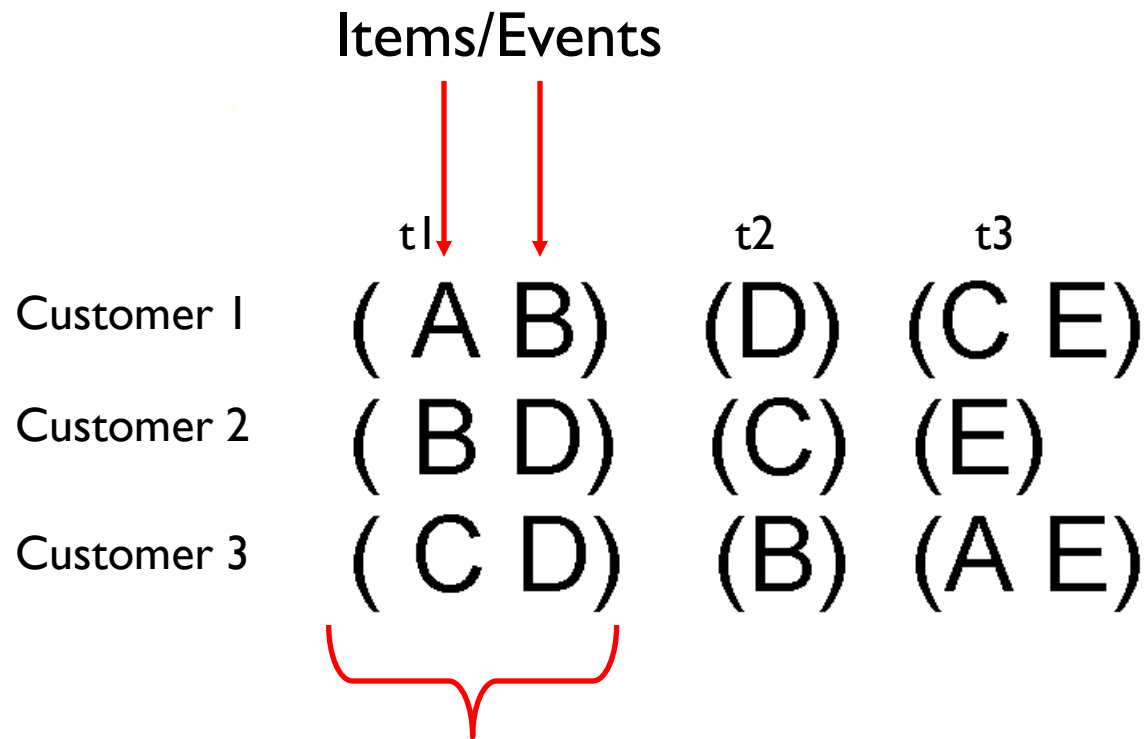
Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.



# Ordered Data

---

## ► Sequences of transactions



An element of the  
sequence

# Ordered Data

---

- ▶ Genomic sequence data

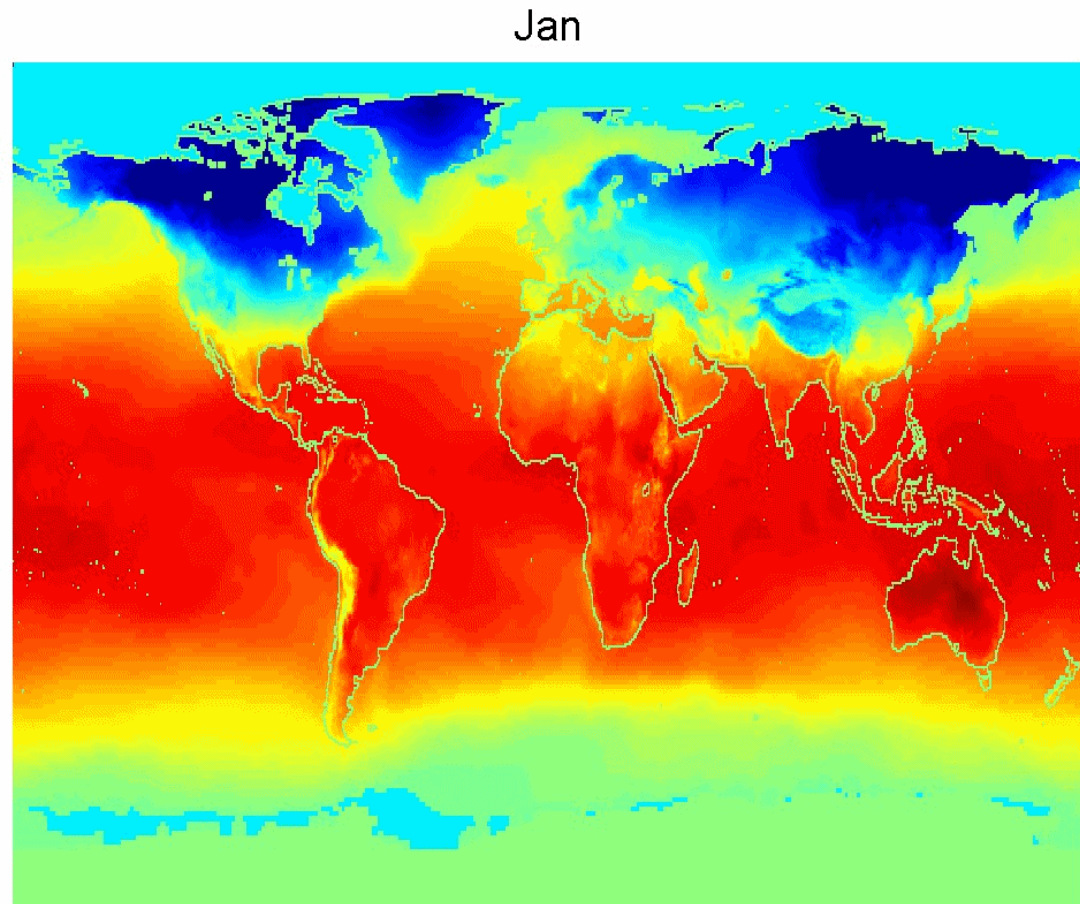
**GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG**

# Spatial 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 with loan risks is built using poor data
  - ▶ Some credit-worthy candidates are denied loans
  - ▶ More loans are given to individuals that default

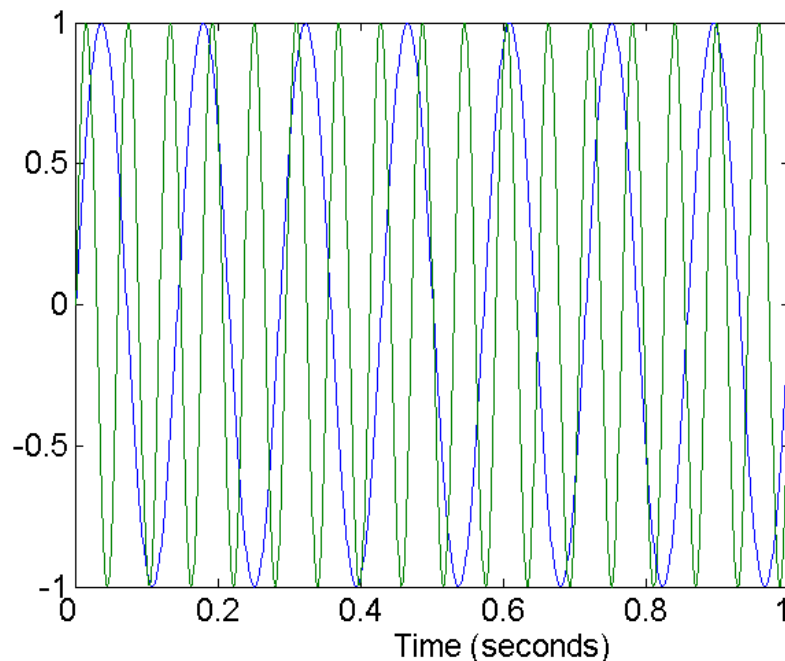
# Data Quality ...

---

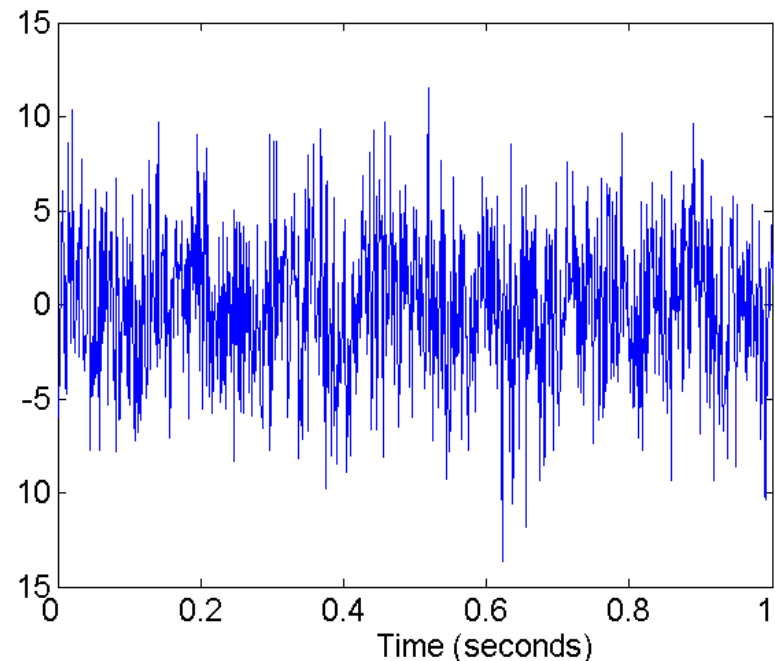
- ▶ What kinds of data quality problems?
- ▶ How can we detect problems in the data?
- ▶ What can we do about these problems?
  
- ▶ Examples of data quality problems:
  - ▶ Noise and outliers
  - ▶ Missing values
  - ▶ Duplicate data

# Noise

- ▶ 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

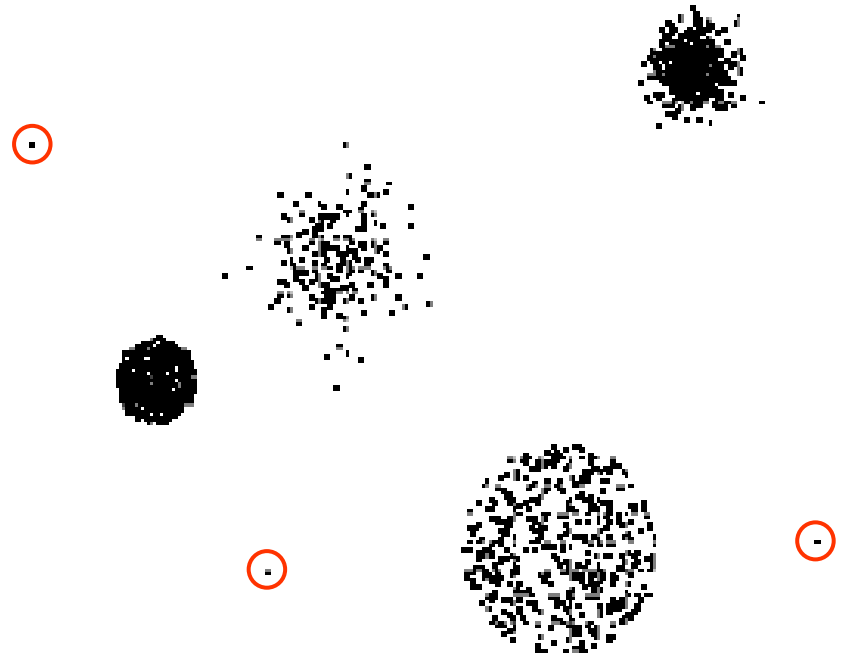


Two Sine Waves + Noise

# 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



# 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
  - ▶ 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

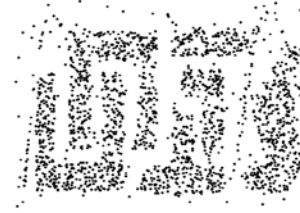
# Data Preprocessing

---

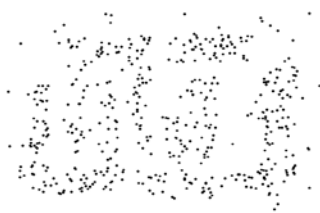
- ▶ Aggregation
- ▶ Sampling
- ▶ Dimensionality Reduction
- ▶ Feature subset selection
- ▶ Feature creation
- ▶ Discretization and Binarization
- ▶ Attribute Transformation



8000 points



2000 Points



500 Points

# 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

# 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.

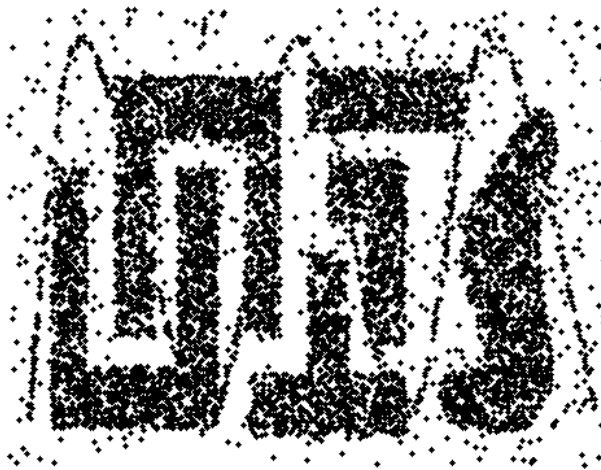
# Sampling

---

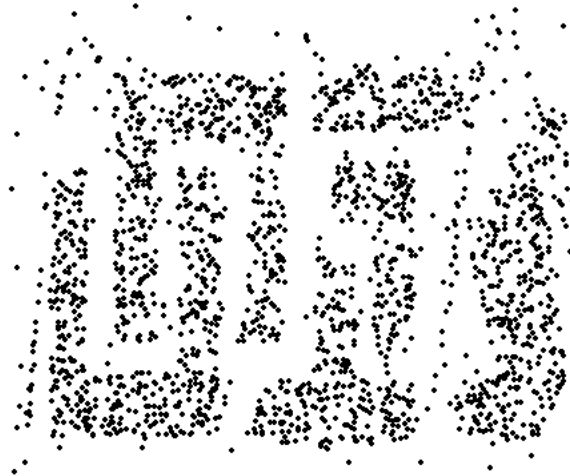
- ▶ The key principle for effective sampling is the following:
  - ▶ Using a sample will work almost as well as using the entire data sets, if the sample is representative
  - ▶ A sample is representative if it has approximately the same property (of interest) as the original set of data

# Sample Size

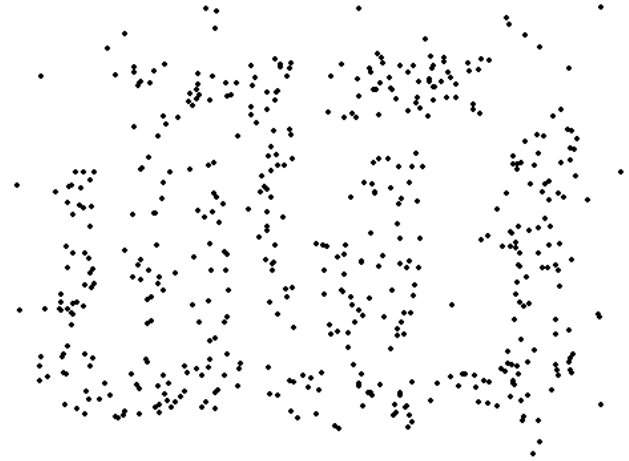
---



8000 points



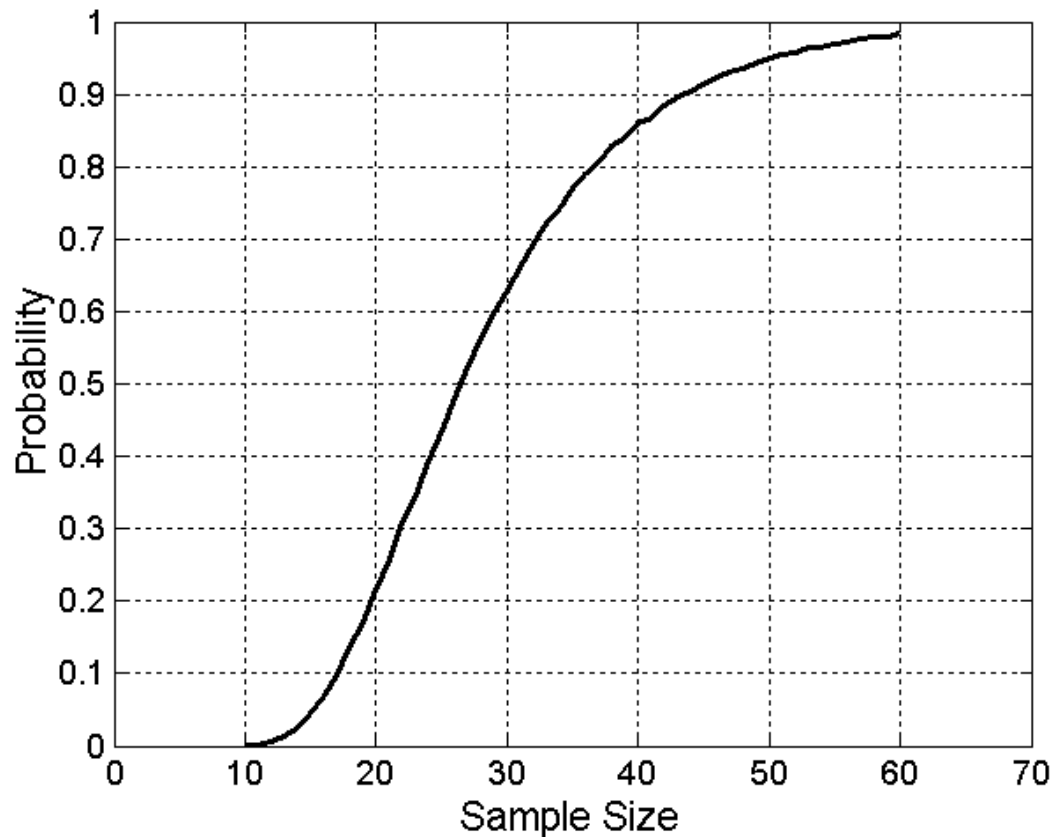
2000 Points



500 Points

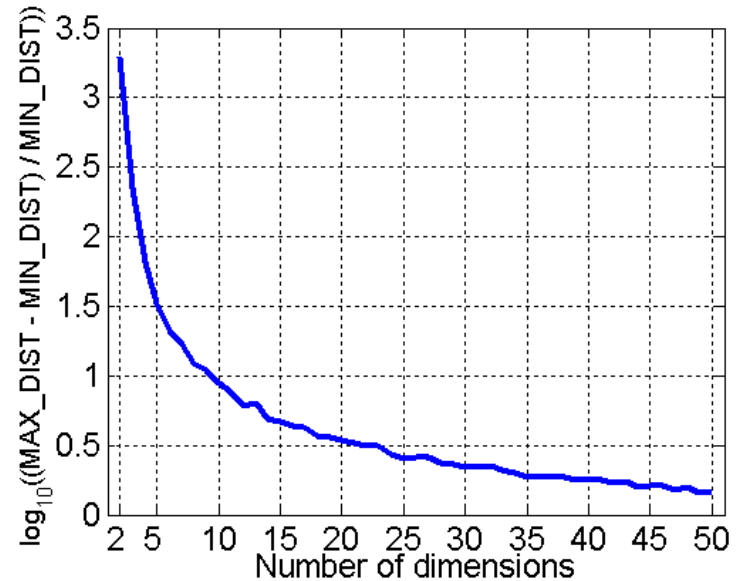
# Sample Size

- ▶ E.g., 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 is critical for clustering and outlier detection, become less meaningful
- ▶ If  $N_1 = 100$  represents a dense sample for a single input problem, then  $N_{10} = 100^{10}$  is the sample size required for the same sampling density with dimension 10.
- ▶ The proportion of a hypersphere with radius  $r$  and dimension  $d$ , to that of a hypercube with sides of length  $2r$  and dimension  $d$  converges to 0 as  $d$  goes to infinity — nearly all of the high-dimensional space is “far away” from the center



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



# Curse of Dimensionality

---

- ▶ Typical text categorization problem:
  - ▶ *TREC-AP* headlines (Cohen&Singer,2000): 319,000+ documents, 67,000+ words, 3,647,000+ word 4-grams used as features.
- ▶ *How can you learn with so many features?*
  - ▶ For speed, exploit *sparse* features.
  - ▶ Use simple classifiers (linear or loglinear)



# 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

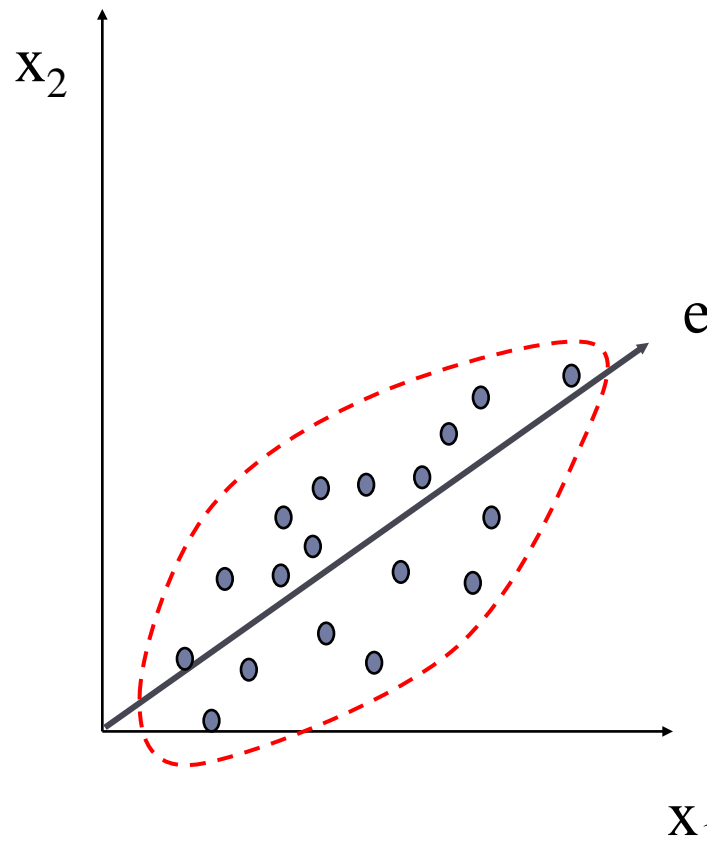
## ▶ Techniques

- ▶ 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

---

256



# 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

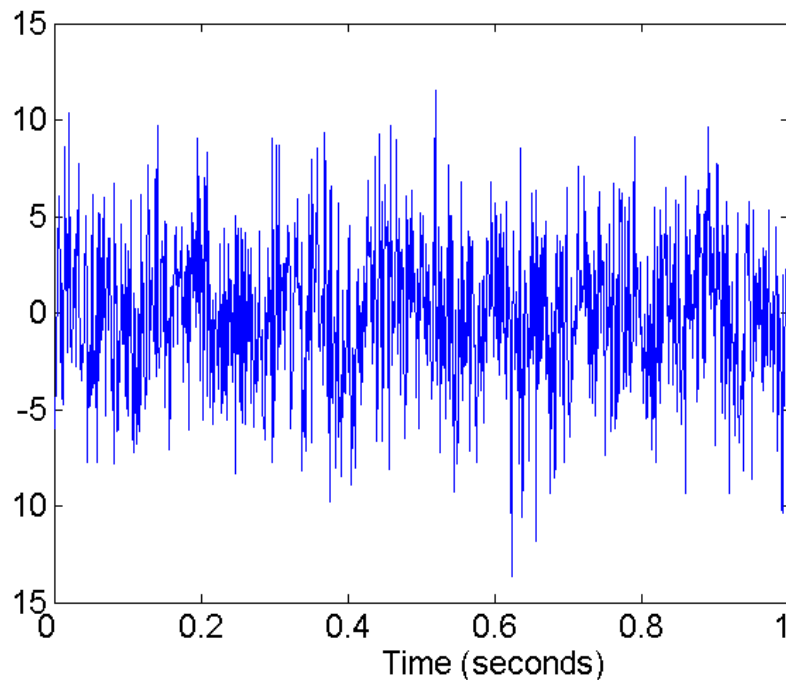
---

- ▶ 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

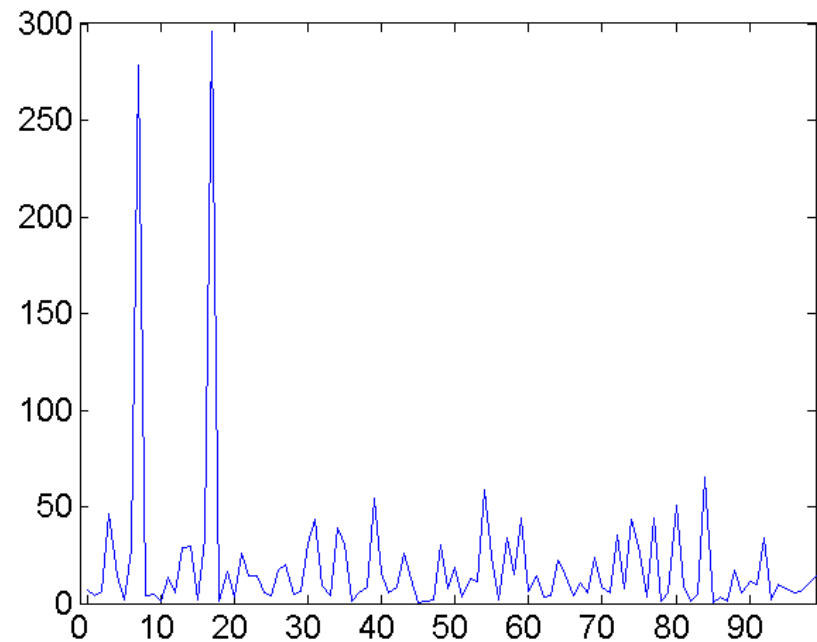
# Mapping Data to a New Space

---

## ► Fourier and wavelet transform



Two Sine Waves + Noise



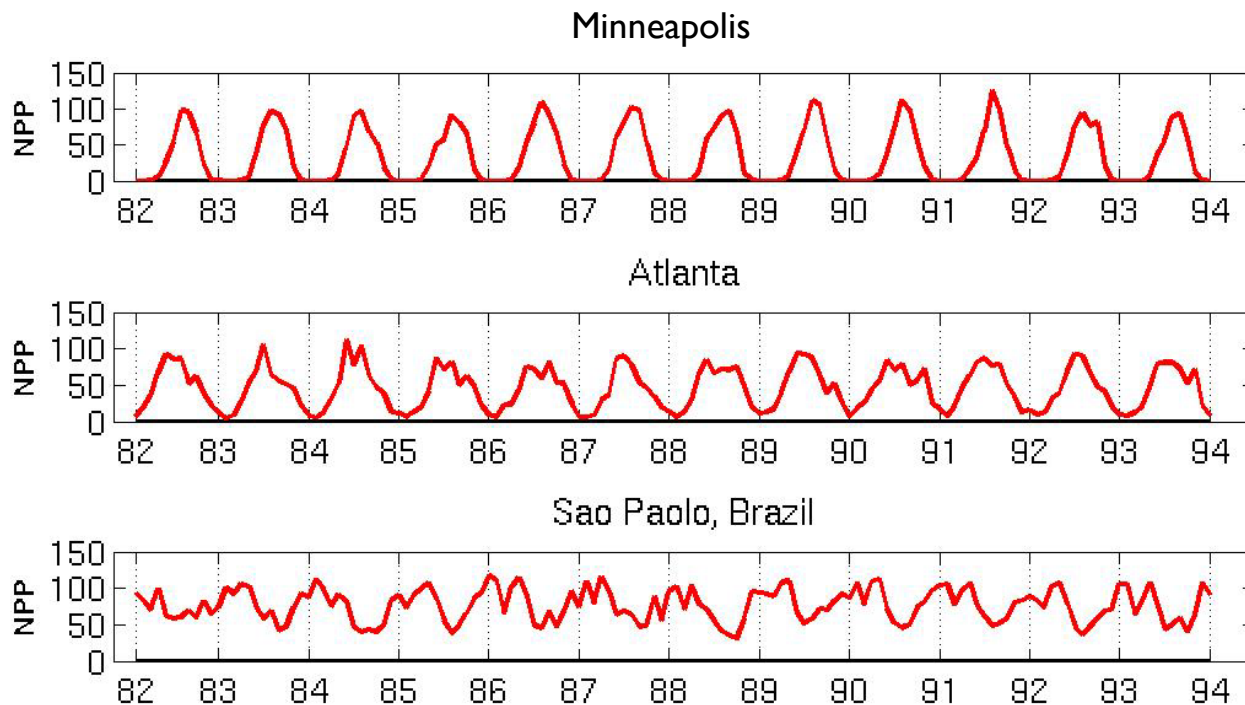
Frequency

# 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, magnitude
  - ▶ In statistics, **standardization** refers to subtracting off the means and dividing by the standard deviation

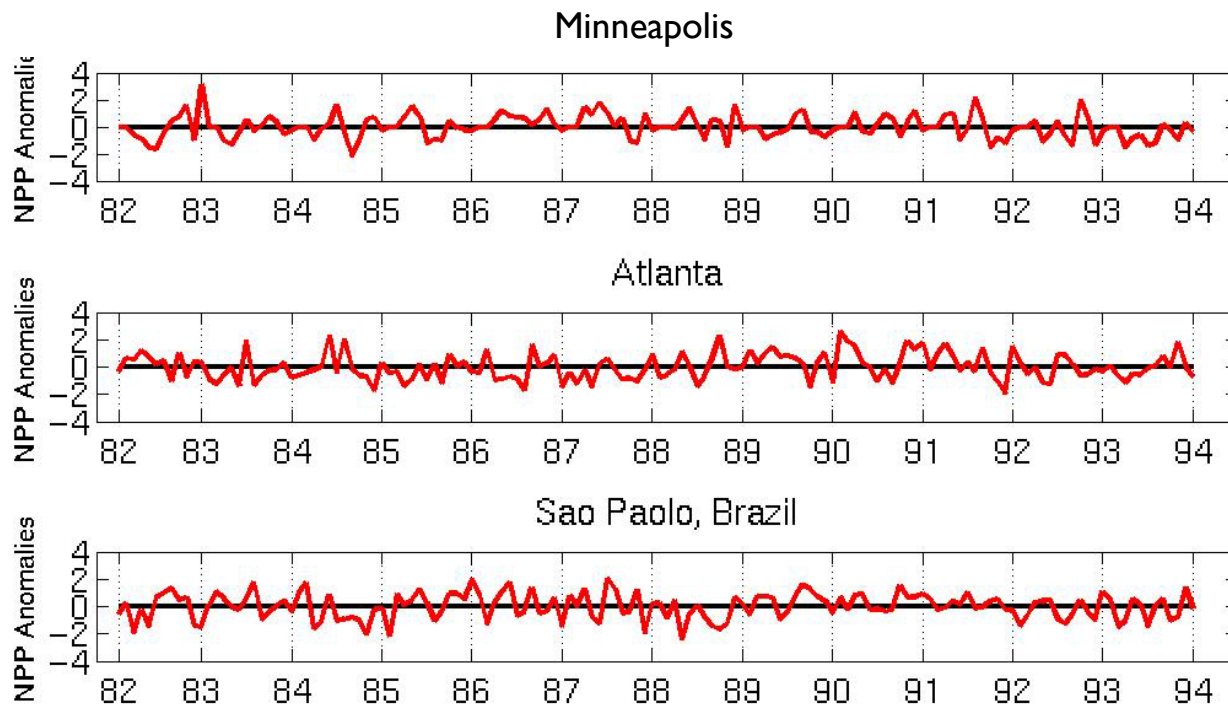




Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paulo	-0.7581	-0.5739	1.0000



Normalized using  
monthly Z Score:

Subtract off monthly  
mean and divide by  
monthly standard  
deviation

### Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paulo	0.0906	-0.0154	1.0000

# Summary

---

- ▶ **Attributes and Objects**

- ▶ Attribute types: nominal / ordinal / interval / ratio; discrete / continuous

- ▶ **Types of Data**

- ▶ Record, graph and network, ordered, spatial...

- ▶ **Data Quality**

- ▶ Noise, outliers, missing values, duplicate data

- ▶ **Data Preprocessing**

- ▶ Sampling, dimensionality reduction, feature selection...
  - ▶ **Curse of dimensionality**