CS5344
Lesson 1b: Data

https://canvas.nus.edu.sg/courses/38824

Anthony Tung
Department of Computer Science
anthony@comp.nus.edu.sg

Acknowledgement: Modified from Introduction to Data Mining , 2nd Edition By Tan, Steinbach, Kumar

Outline

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

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

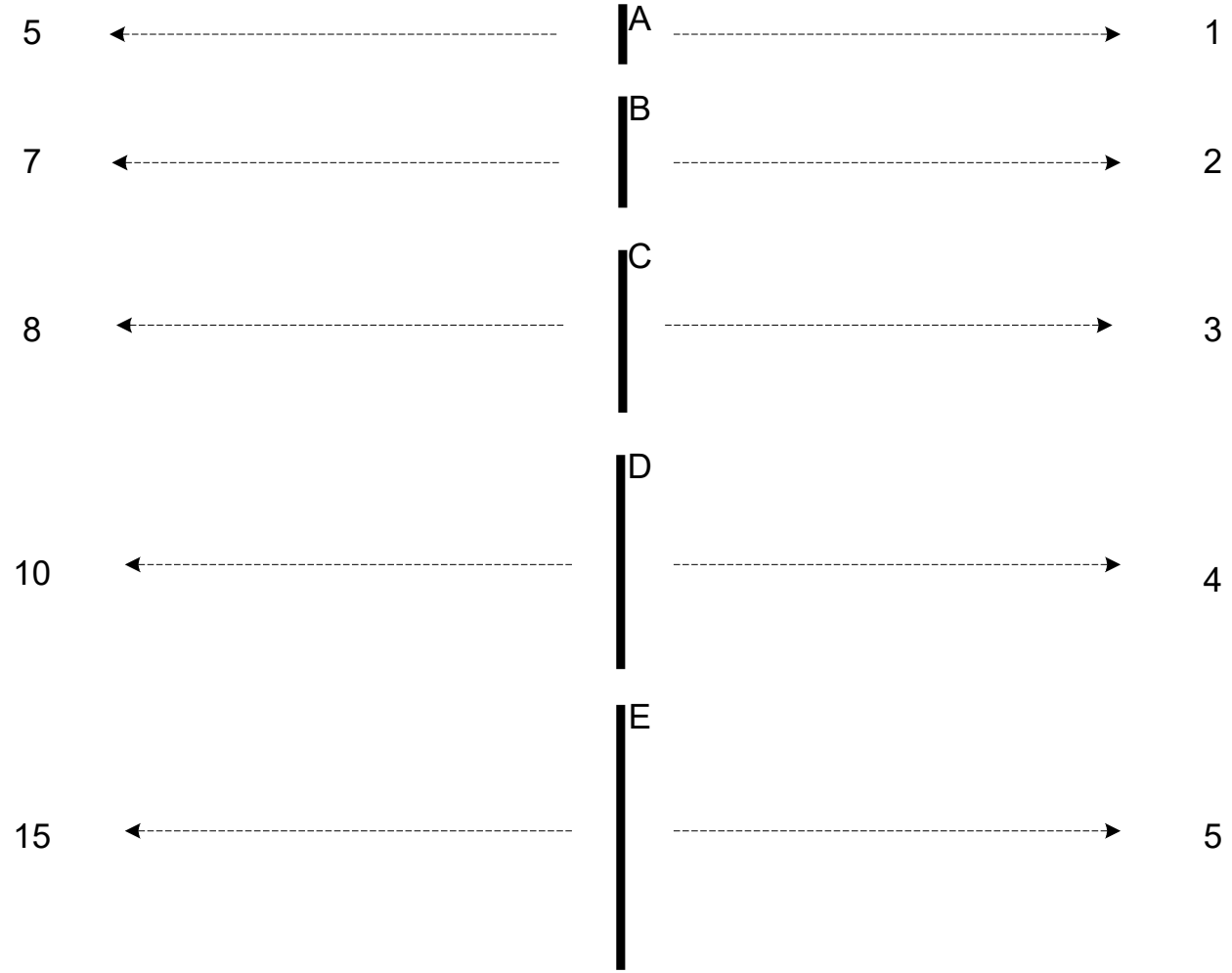
    Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values

    Example: Attribute values for ID and age are integers

- **But properties of attribute can be different from the properties of the values used to represent the attribute**

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, counts, elapsed time (e.g., time to run a race)

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., *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.
- 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."

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 (among many possible) 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 present
- Analysis may depend on these other properties of the data

  Many statistical analyses depend only on the distribution
- In the end, what is meaningful can be specific to domain

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**

Outline

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

Types of Big Data

- Relational data
- High-dimensional data
- Sequence
- Tree
- Graph
- Mixed types
  - Sequences in a graph (social network)
  - Spatial-temporal data
  - Spatial-textual data
  - High-dimensional time series

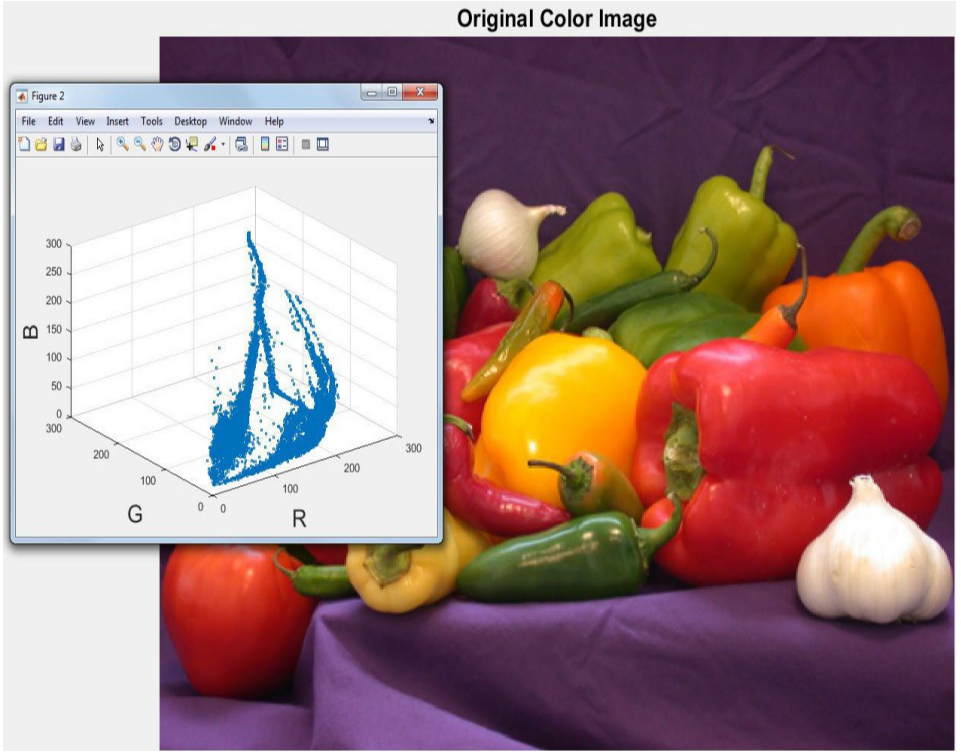
Types of Big Data(I): Relational Data

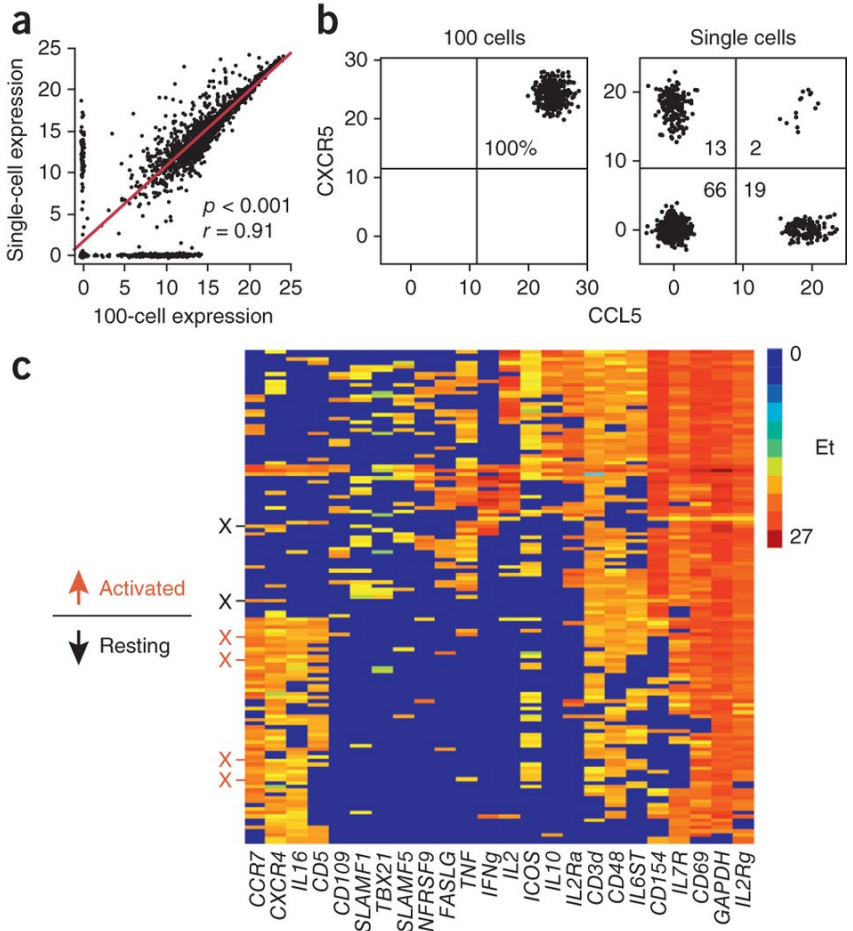
NUS National University of Singapore

Numeric(integer, float), categorical, binary, textual

age	salary	credit	sex	country	spending	Zoo	Orchard Road	Sentosa	Casino
35	30k	poor	M	USA	500	0	1	1	1
25	76k	good	F	China	10,000	1	1	1	1
40	90k	good	F	India	2,000	0	0	1	1
30	100k	poor	M	Taiwan	10,000	1	0	1	1
25	110k	good	F	Malaysia	2,000	0	1	0	1
30	50k	good	M	Malaysia	5,000	1	1	0	1
35	35k	poor	F	China	100,000	0	0	0	1
45	15k	poor	M	Indonesia	15,000	1	0	0	1

Types of Big Data(II): High-dimensional Data





Types of Big Data(III): Sequence

Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA Crime DNA

GTCGACCGGTGACCGTGCGTACACAGTGCTCCGGATAGCTGATAGCTCCGGTG

CAGCTGGCCACTGGCACGCATGTGTCACGAGGCCTATCGACTATCGAGGCCAC

Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA Suspect 1 DNA

GTCCCAGCCGGACCGTACCGGTAGATCAGCCGGTAGATTGATAGCGTGATGTG

CAGGGTCGGCCTGGCATGGCCATCTAGTCGGCCATCTAACTATCGCACTACAC

Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA Suspect 2 DNA

GTCTACGTAATCGTAGCCATCCGGACAGTGTGCACGATCGTACATGCTACGTG

CAGATGCATTAGCATCGGTAGGCCTGTCACACGTGCTAGCATGTACGATGCAC

Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA Suspect 3 DNA

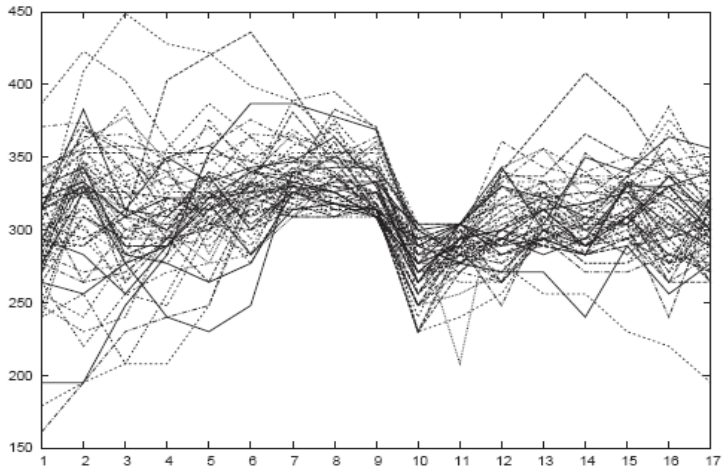
GTCGACCGGTGACCGTGCGTACACAGTGCTCCGGATAGCTGATAGCTCCGGTG

CAGCTGGCCACTGGCACGCATGTGTCACGAGGCCTATCGACTATCGAGGCCAC

Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA Suspect 4 DNA

GTCTCCATCCGGACTACCATACATCTGGTGTACCCGGTGATATCGTCCGGGTG

CAGAGGTAGGCCTGATGGTATGTAGACCACATGGGCCACTATAGCAGGCCCAC



CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

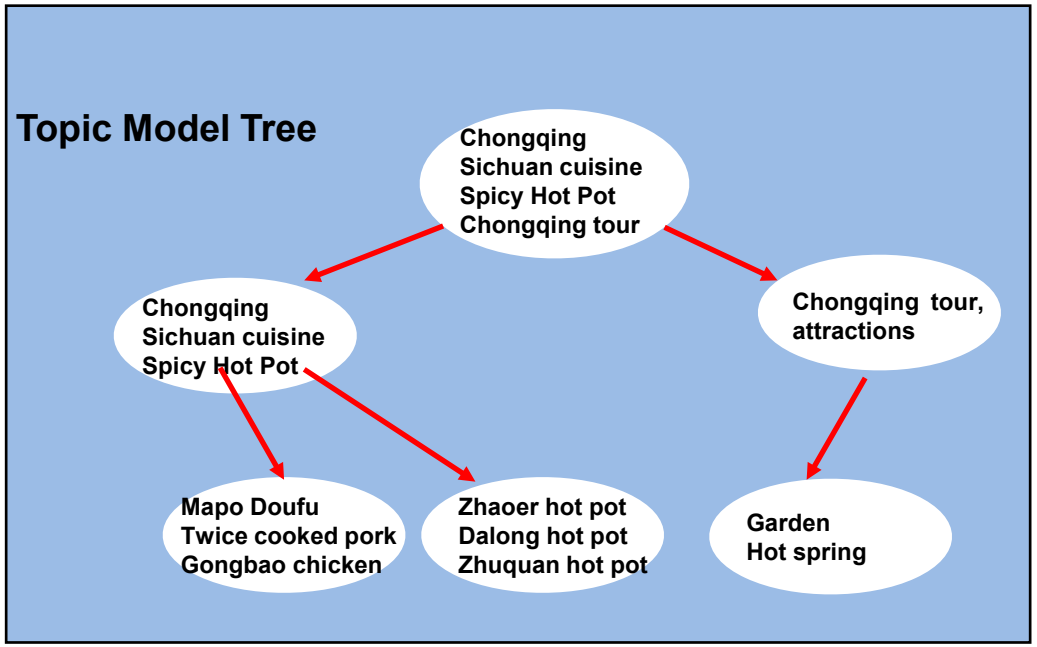
So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

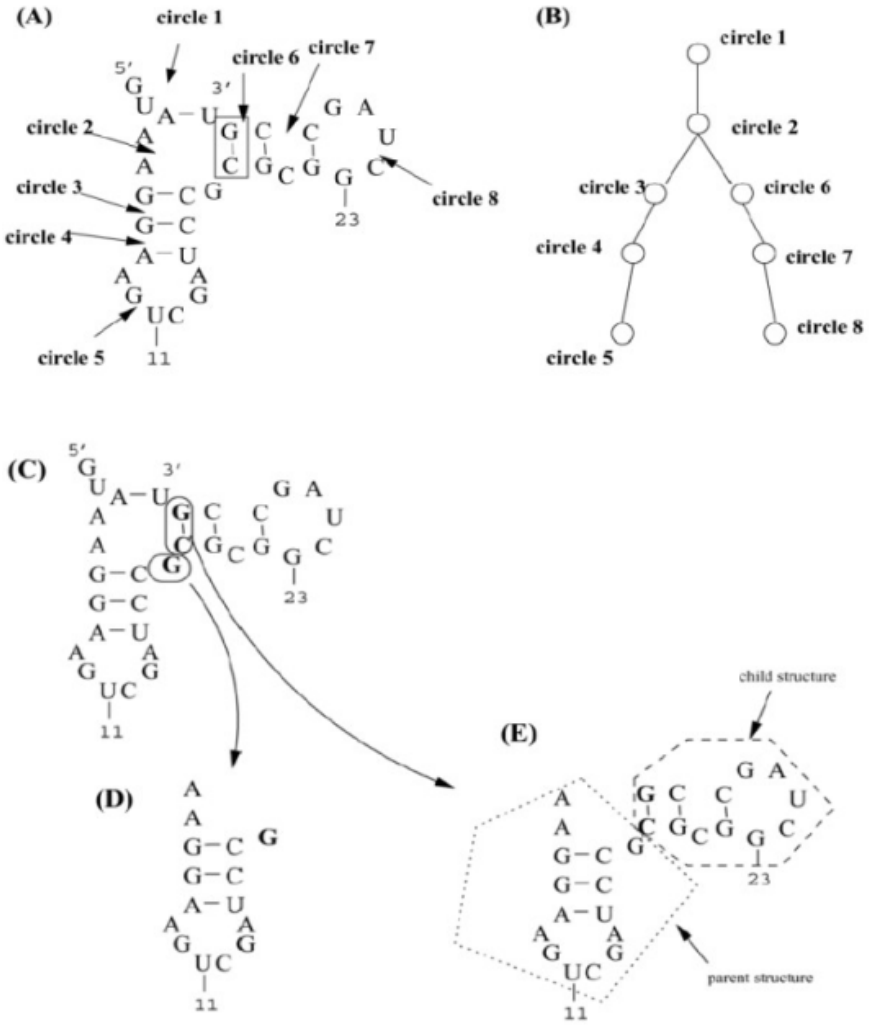
There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.

In another moment down went Alice after it, never once considering how in the world she was to get out again.

The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself falling down a very deep well.

Types of Big Data(IV): Tree





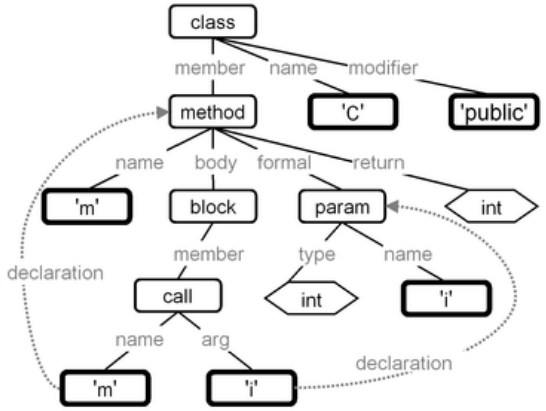
Java Code

```
public class C {

int m(int i) {

    m(i);
  }
}
```

Simplified AST

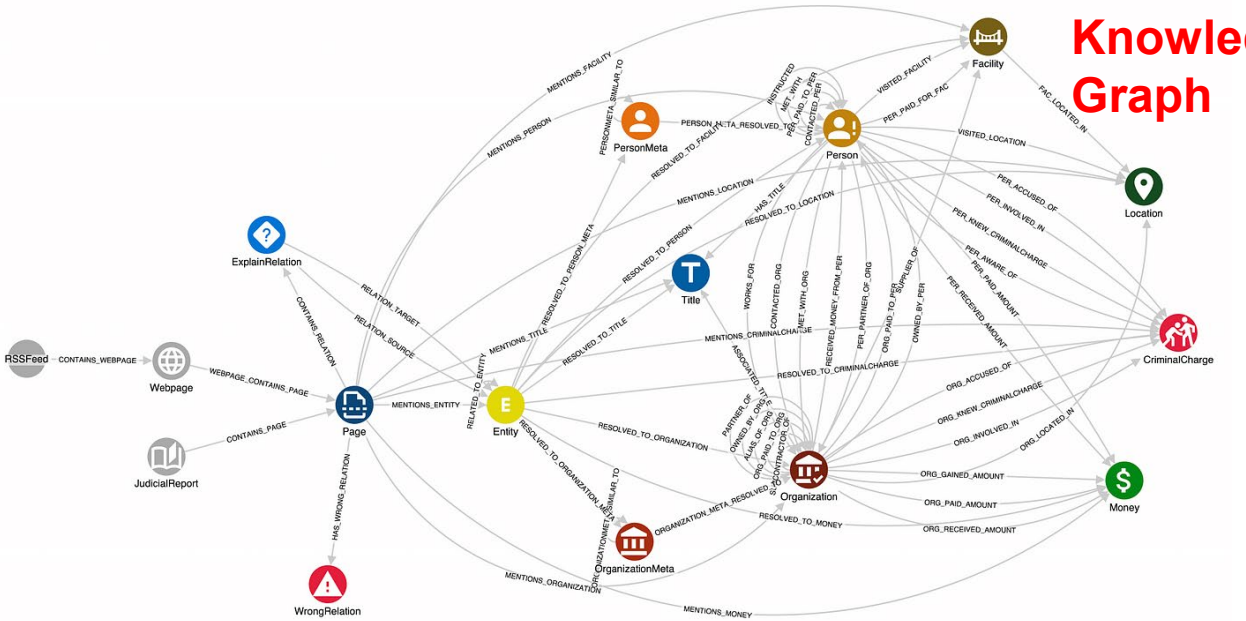


Logic Representation

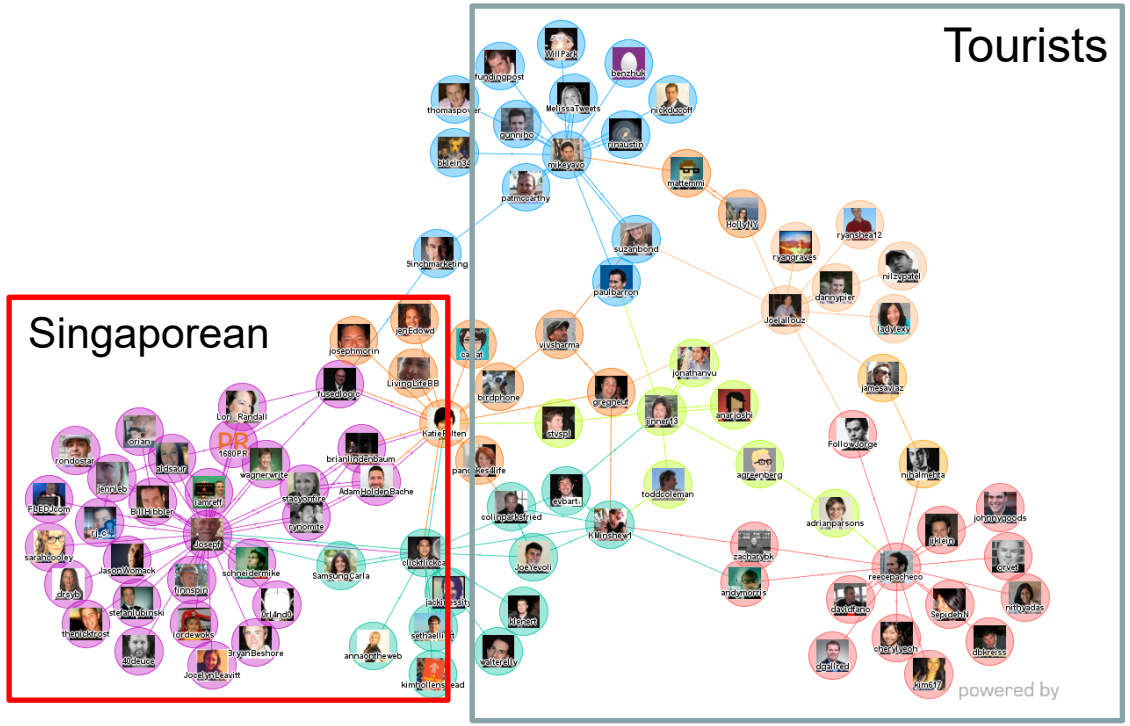
```
class(1, 'C').
  modifier(1, 'public').

method(2, 1, 'm', [9], 'void').
  param(9, 2, 'int, 'i').
  block(10, 2, [12]).
    call(12, 10, ,this', 'm', [13], 2).
      ident(13, 12, 'i', 9).
```

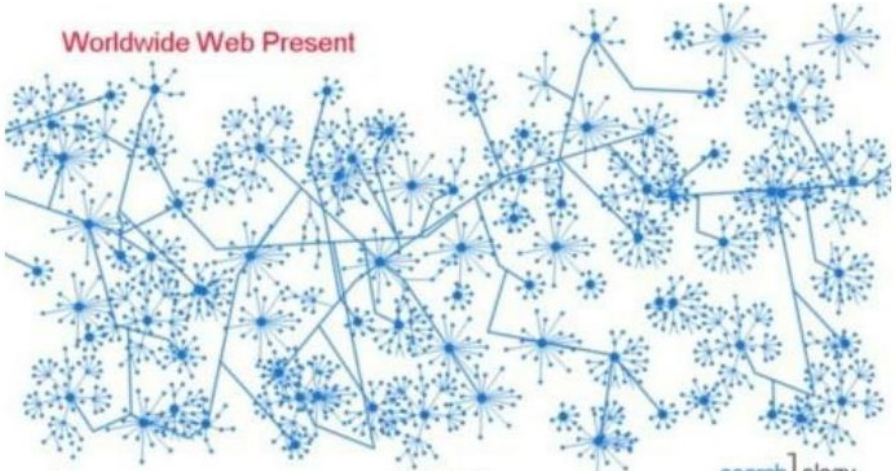
Types of Big Data(V): Graph



Knowledge Graph



Social network



Worldwide Web

Types of Big Data

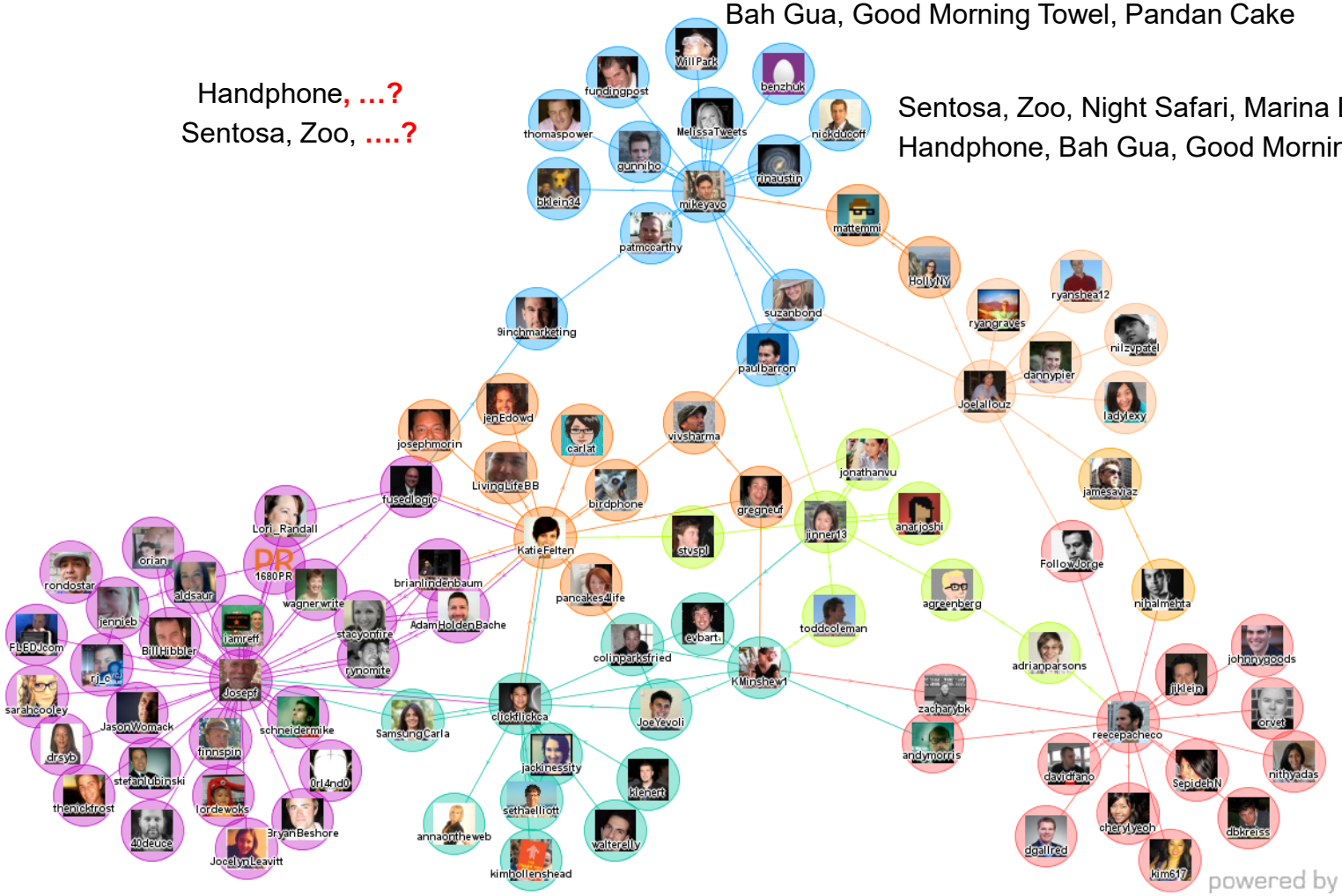
- Relational data
- High-dimensional data
- Sequence
- Tree
- Graph
- Mixed types
  - Sequences in a graph (social network)
  - Spatial-temporal data
  - Spatial-textual data
  - High-dimensional time series

Types of Big Data(IV): Sequences in a Graph

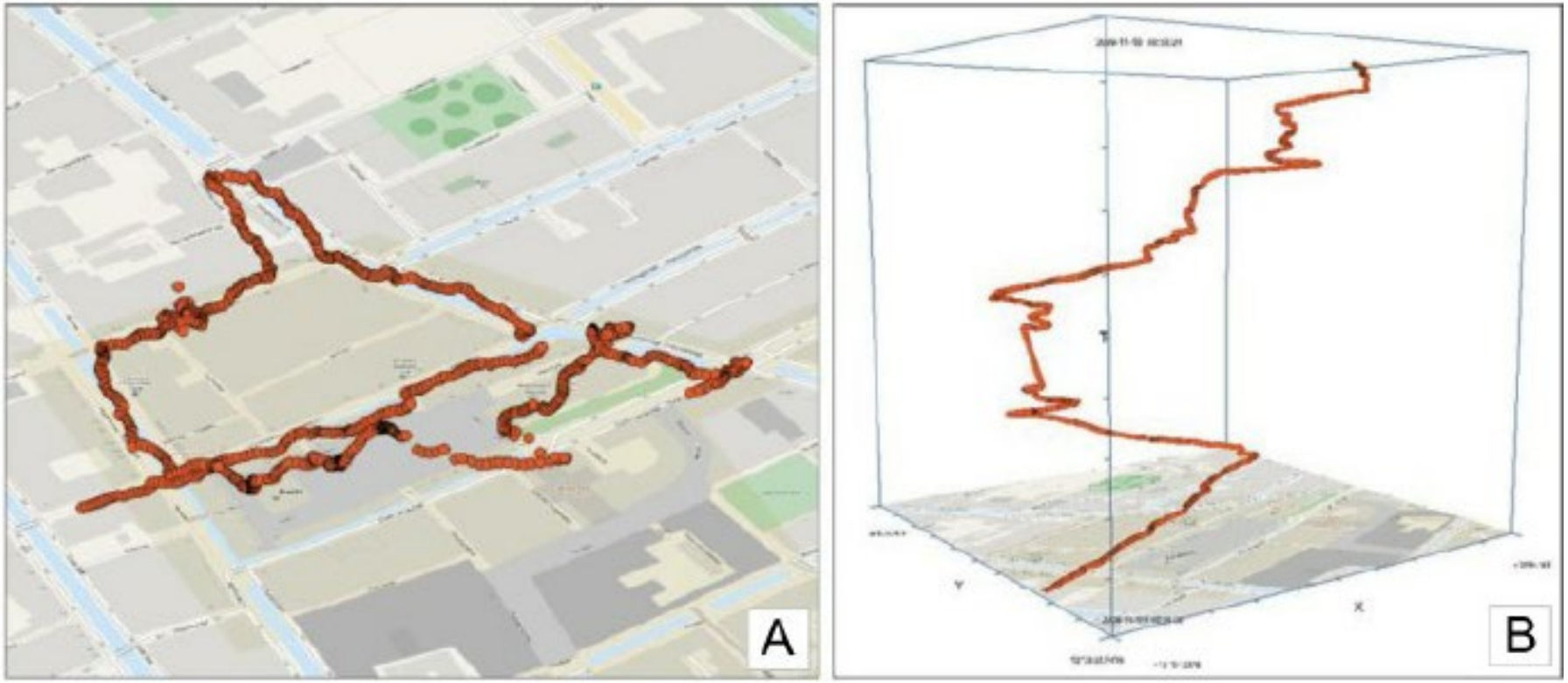
Sentosa, Night Safari, Marina Bay Sand
Bah Gua, Good Morning Towel, Pandan Cake

Handphone, ...?
Sentosa, Zoo, ...?

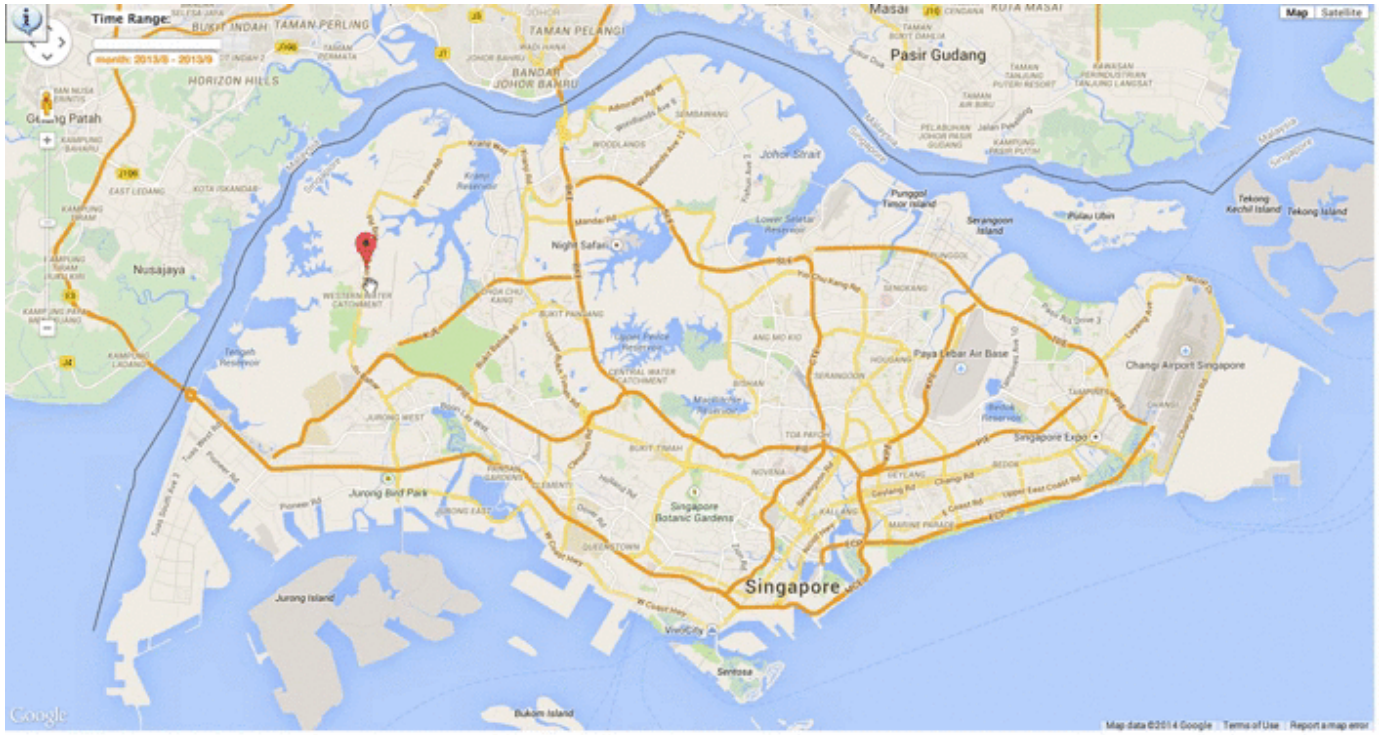
Sentosa, Zoo, Night Safari, Marina Bay Sand
Handphone, Bah Gua, Good Morning Towel

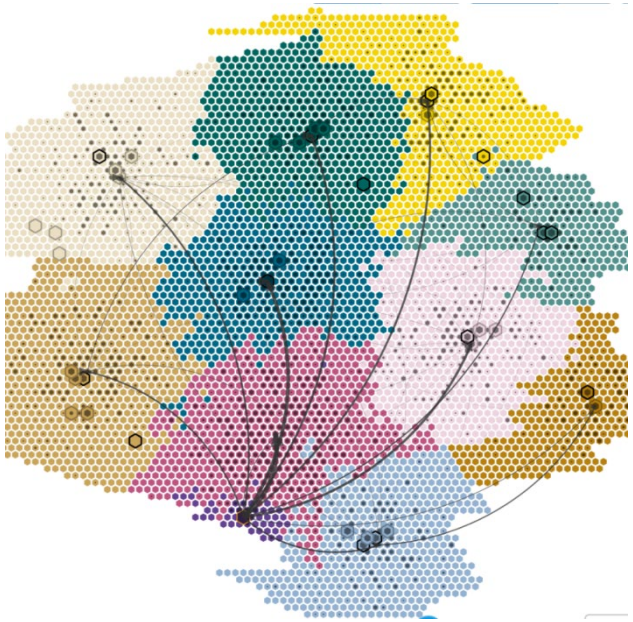


Types of Big Data(V):Spatial-temporal Data

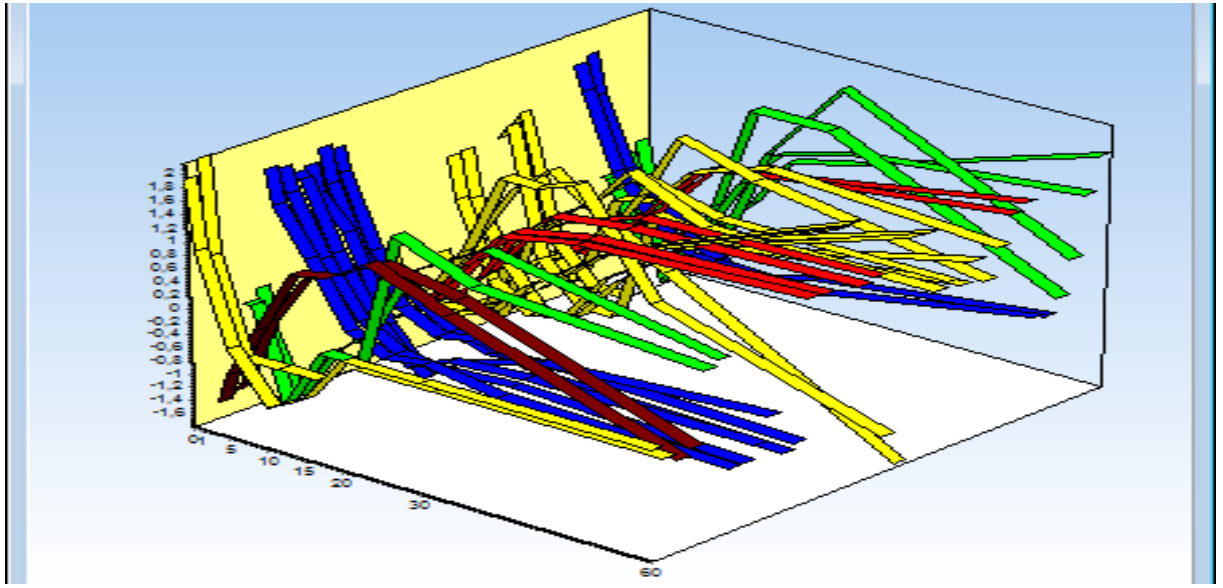


Types of Big Data(VI): Spatial-textual Data





Types of Big Data(VII): High-dimensional Time Series



Types of Big Data: Summary

- The type of big data somehow depends on how data is generated or collected.
- In many cases, the types of data depends on the problem and application themselves.
- Big data applications often need to process multiple types of data
- Distilling raw data into appropriate data format is one of the most fundamental problem.

Outline

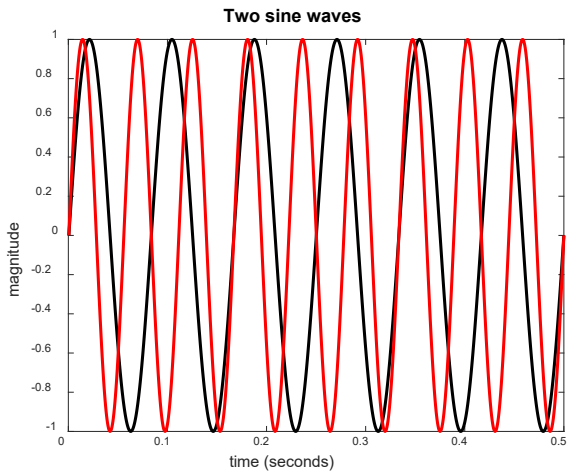
- Attributes and Objects
- Types of Big Data
- Data Quality 
- Data Preprocessing

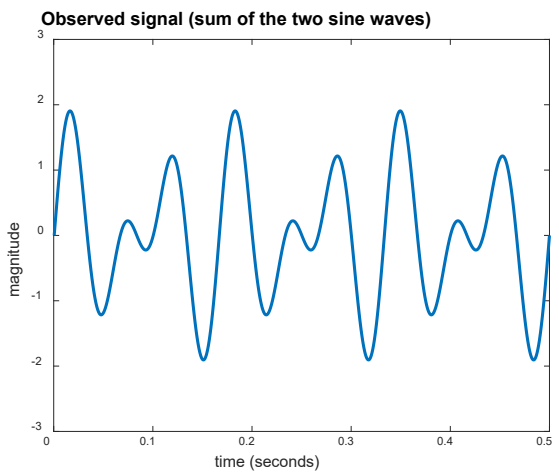
Data Quality …

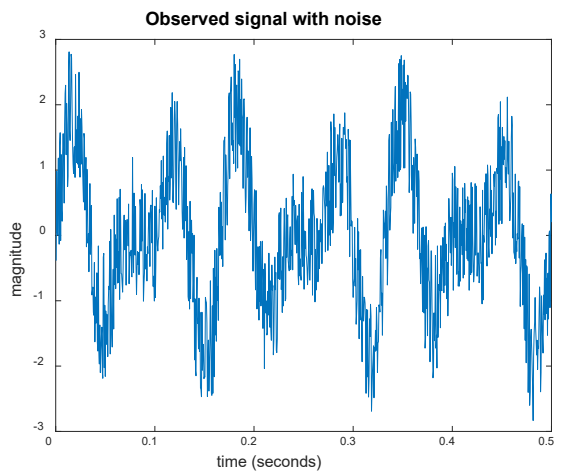
- **What kinds of data quality problems?**
- **How can we detect problems with the data?**
- **What can we do about these problems?**
- **Examples of data quality problems:**
  - Noise and outliers
  - Wrong data
  - Fake data
  - Missing values
  - Duplicate data

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
  - The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise
    - The magnitude and shape of the original signal is distorted







Outliers

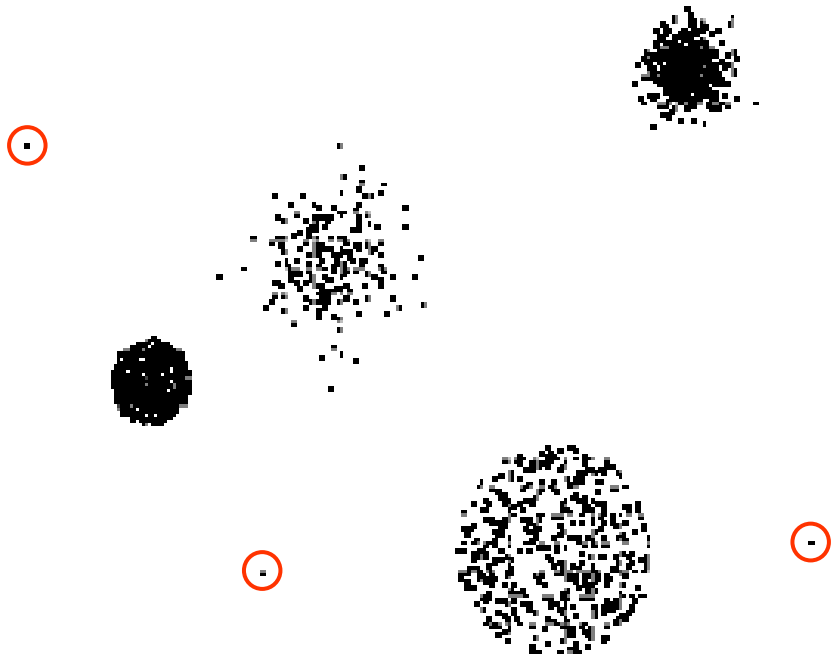
Outliers **are data objects with characteristics that are considerably different than most of the other data objects in the data set**

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?



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?**

Outline

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

Data Preprocessing

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

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

Table 2.4. Data set containing information about customer purchases.

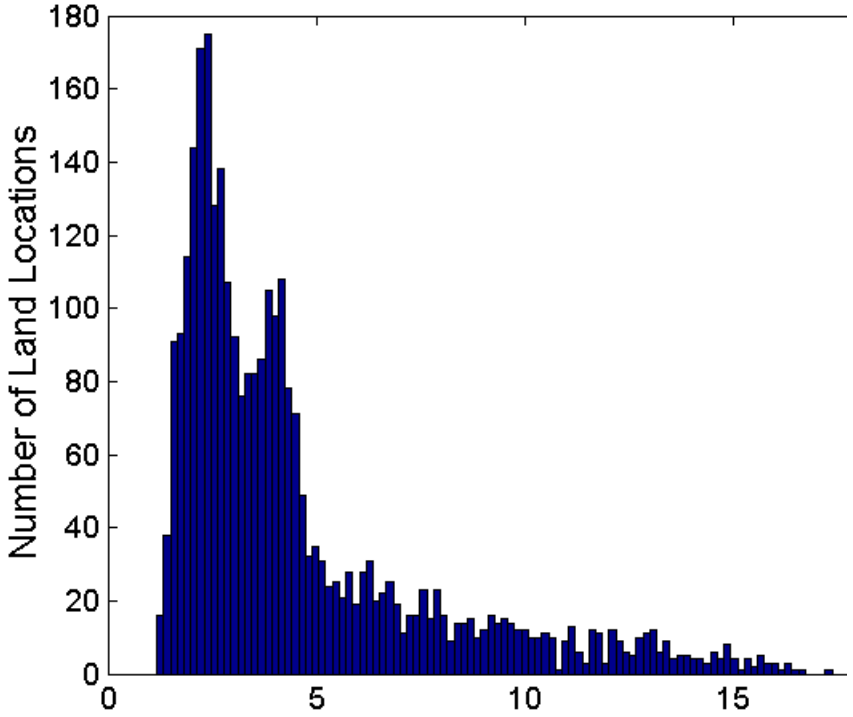
Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	

Example: Precipitation in Australia

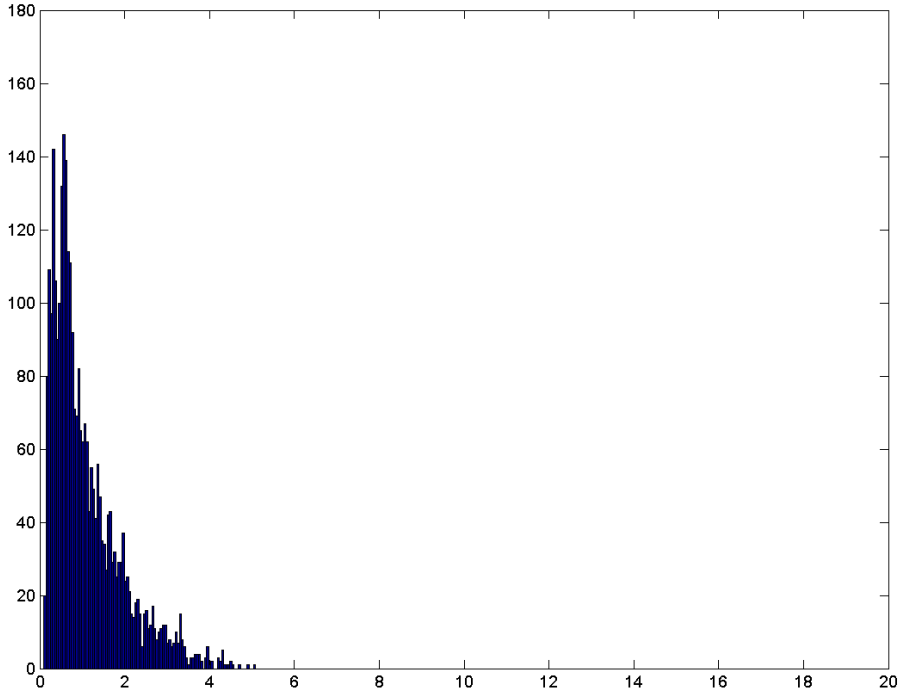
- **This example is based on precipitation in Australia from the period 1982 to 1993.**
  - The next slide shows
  - A histogram for the standard deviation of average monthly precipitation for 3,030 0.5◦ by 0.5◦ grid cells in Australia, and
  - A histogram for the standard deviation of the average yearly 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**

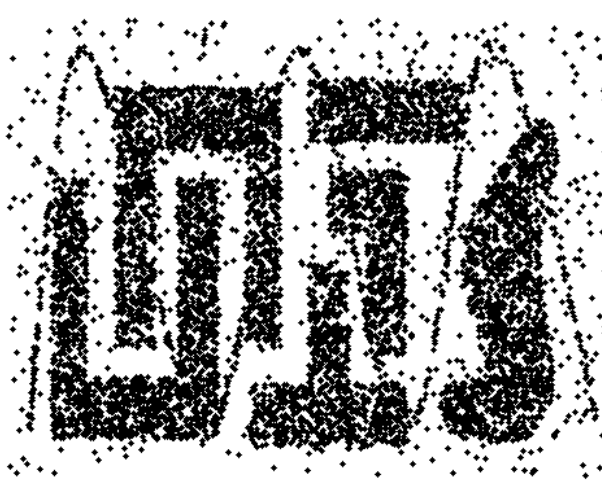
Sampling …

The key principle for effective sampling is the following:

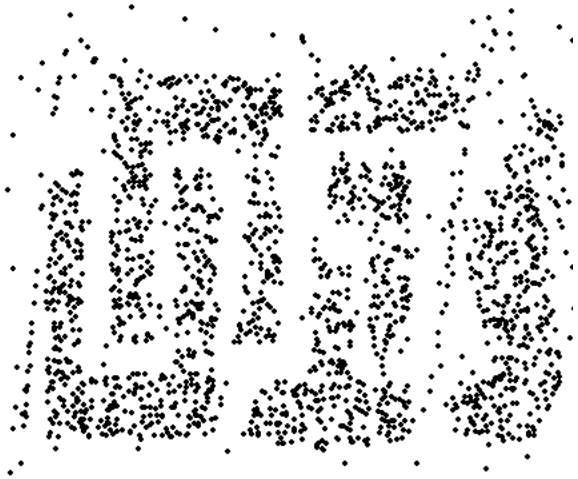
Using a sample will work almost as well as using the entire data set, if the sample is representative

A sample is representative if it has approximately the same properties (of interest) as the original set of data

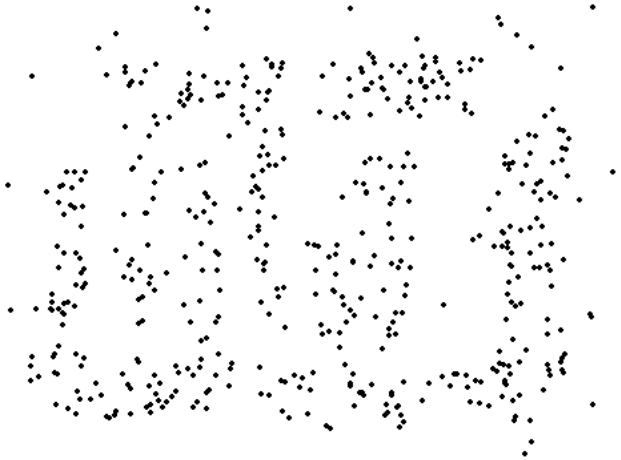
Sample Size



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.

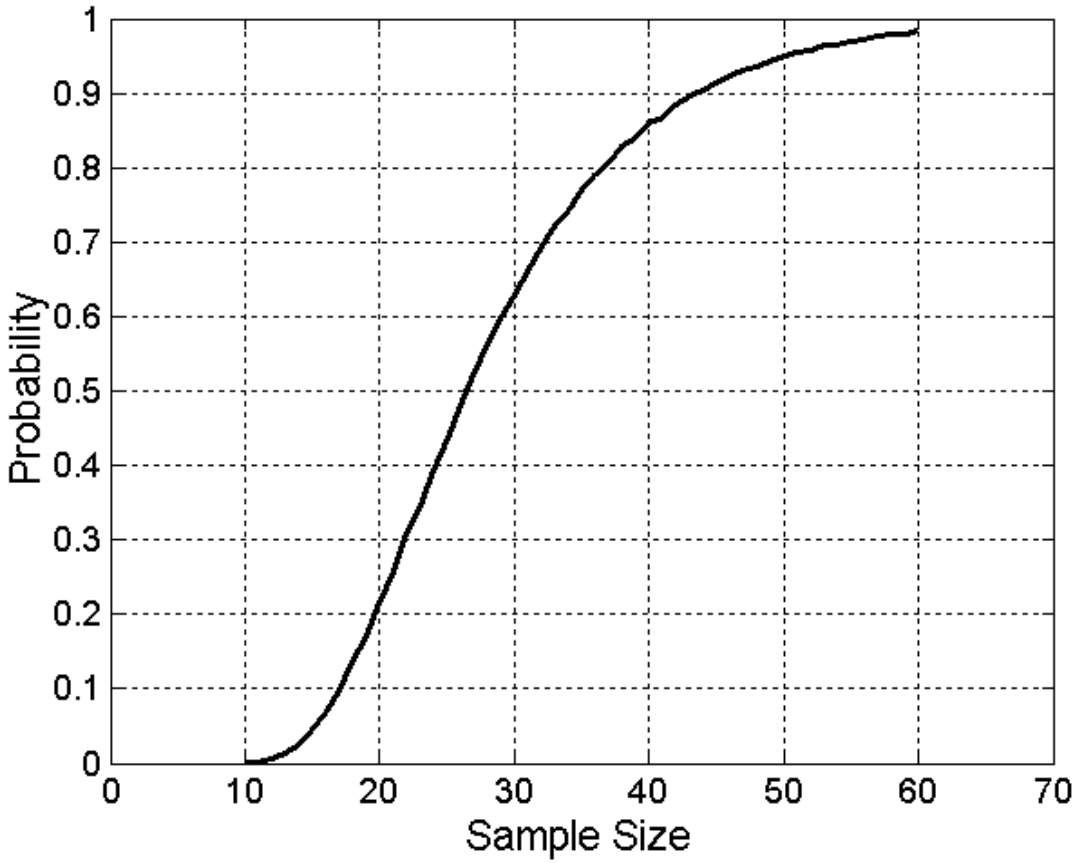
In sampling with replacement, the same object can be picked up more than once

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.

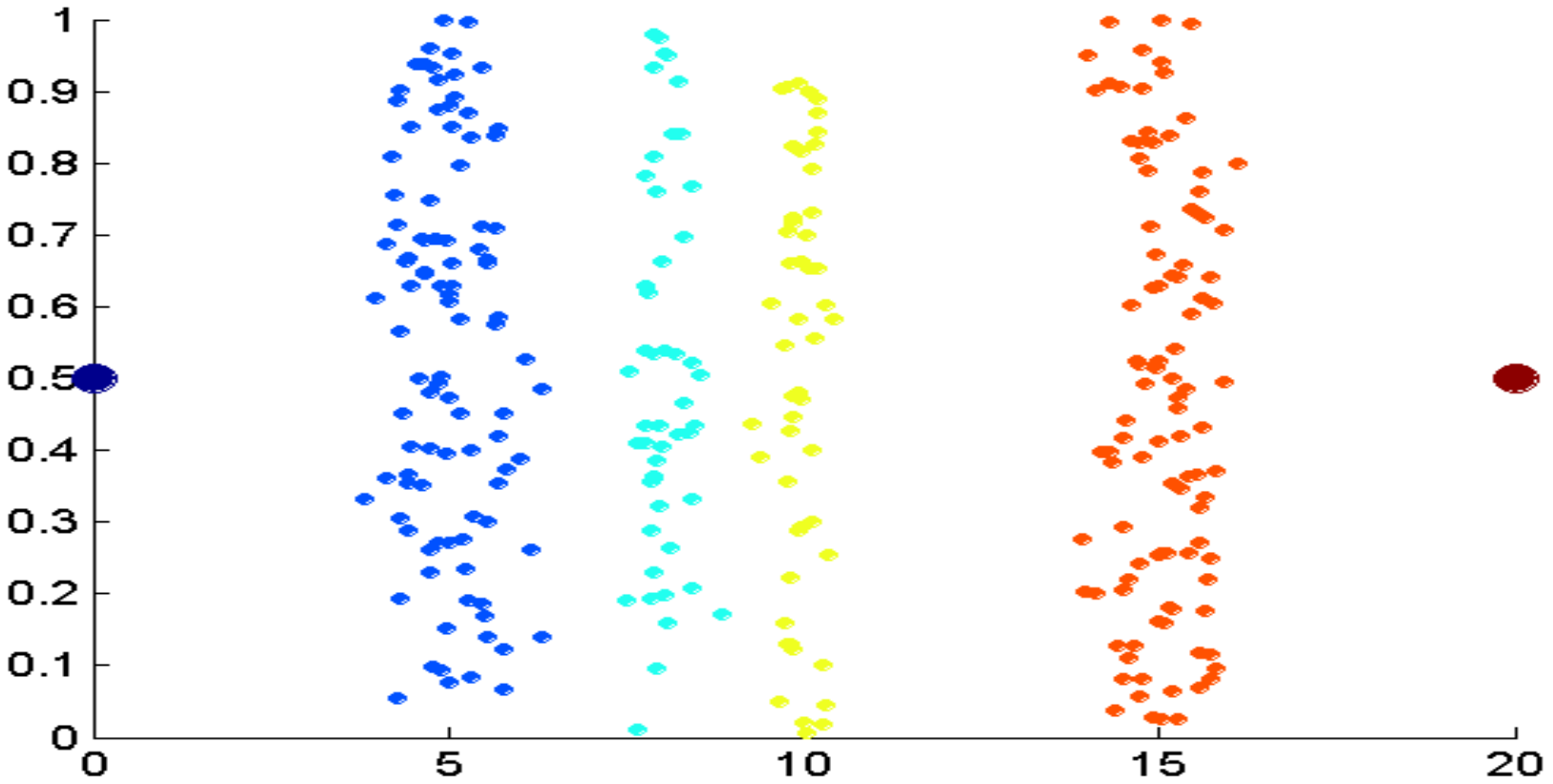


Discretization

Discretization is the process of converting a continuous attribute into an ordinal attribute

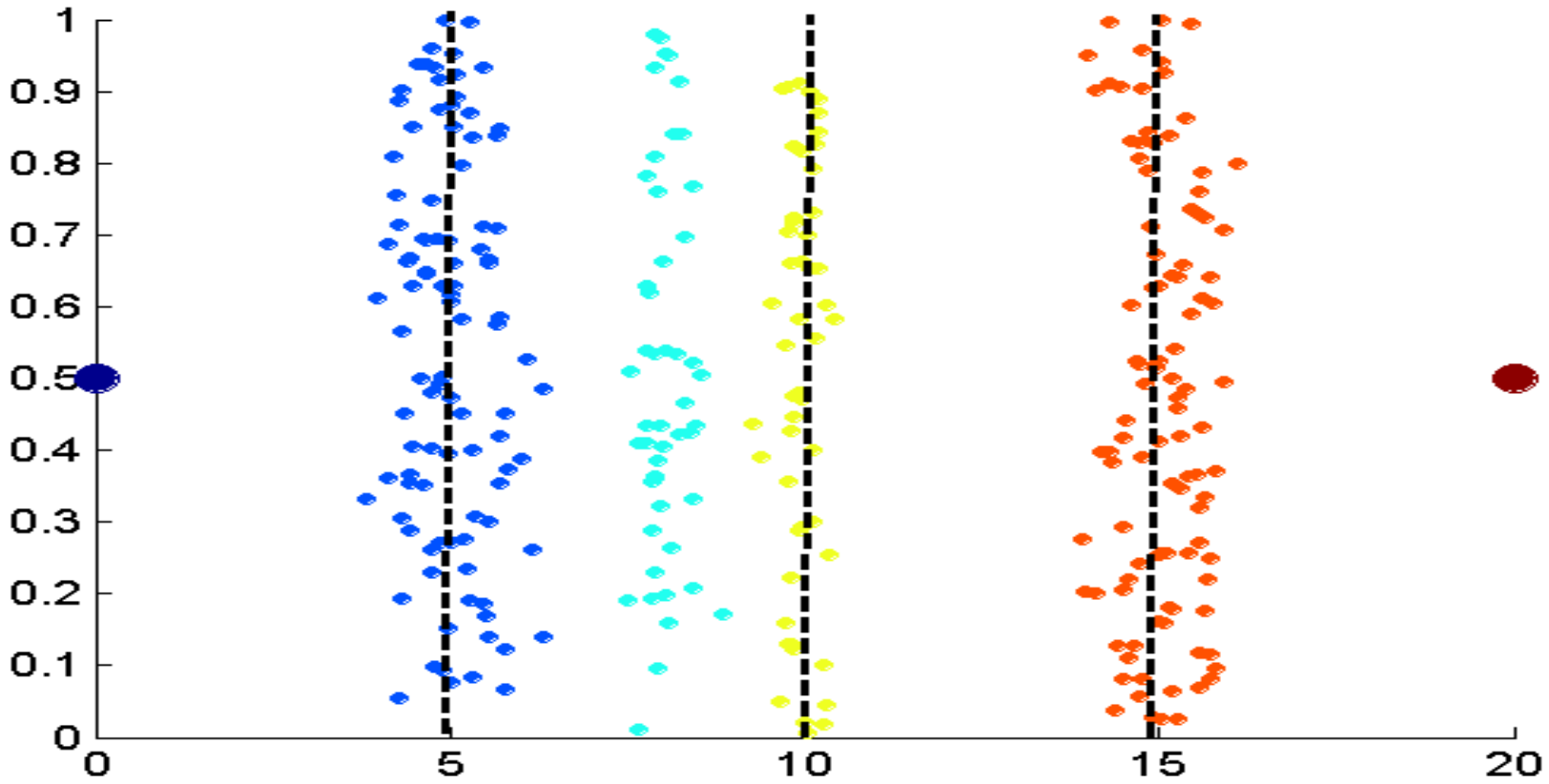
- A potentially infinite number of values are mapped into a small number of categories
- Discretization is used in both unsupervised and supervised settings

Unsupervised Discretization



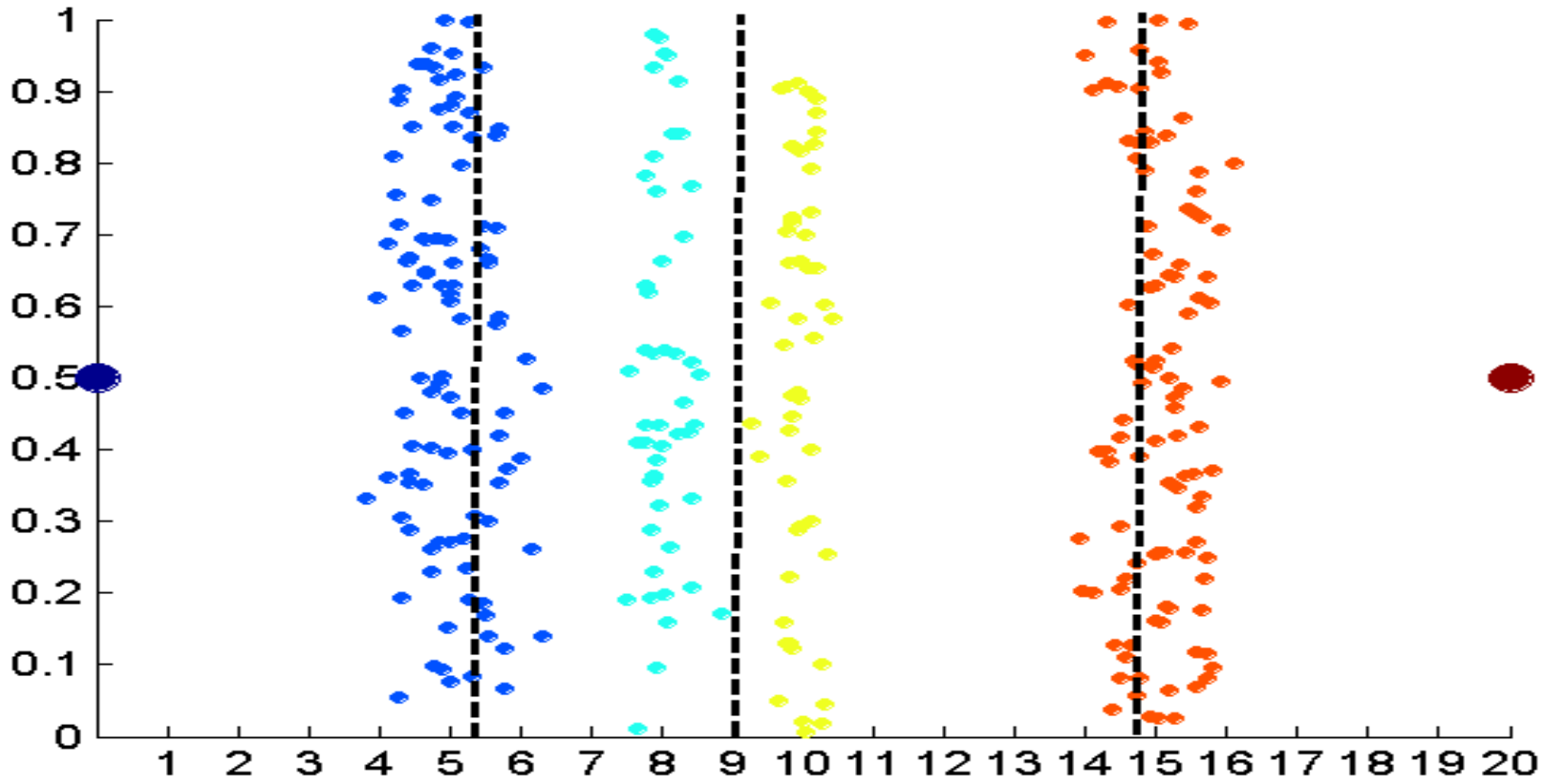
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Unsupervised Discretization



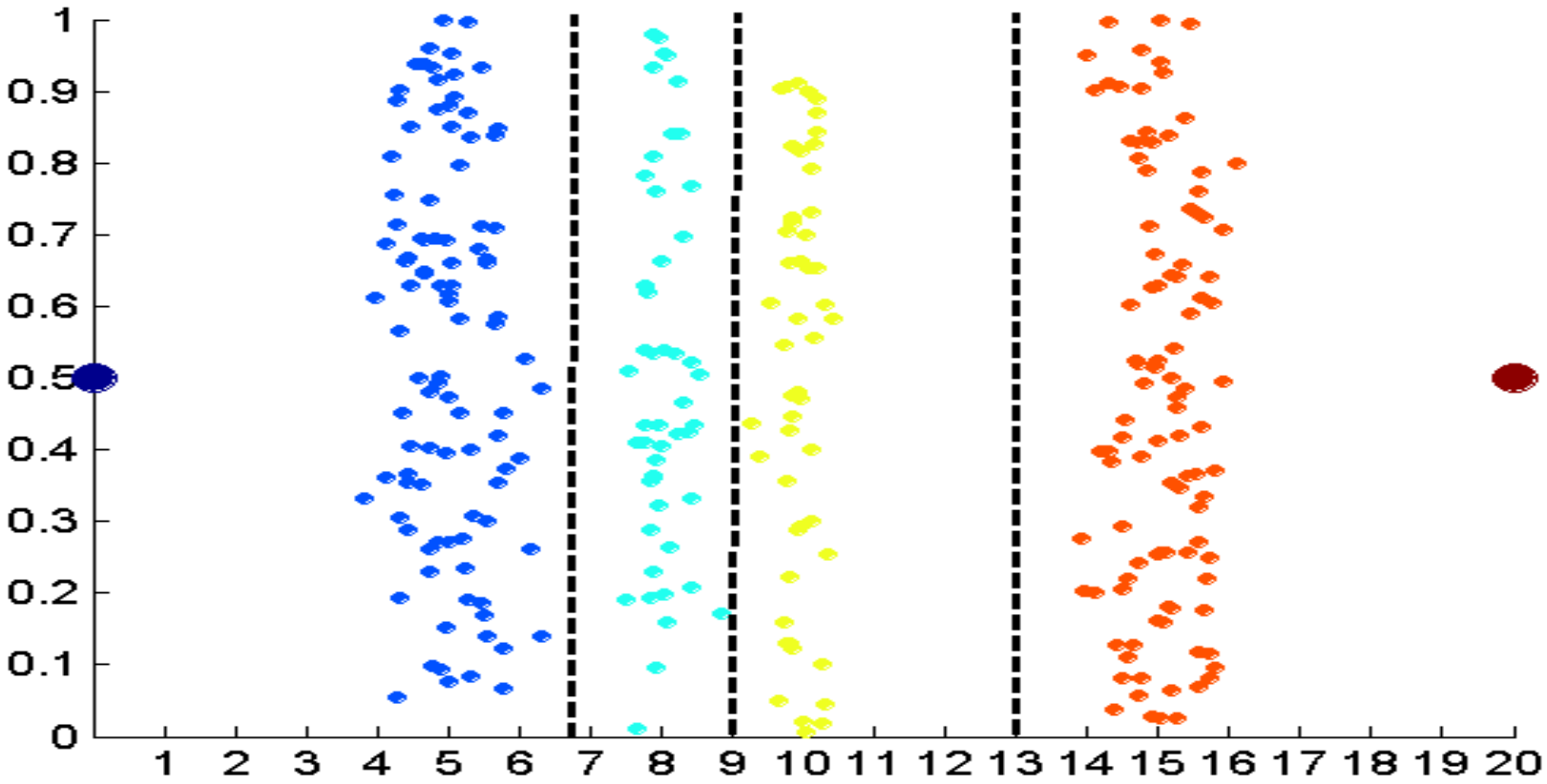
Equal interval width approach used to obtain 4 values.

Unsupervised Discretization



Equal frequency approach used to obtain 4 values.

Unsupervised Discretization

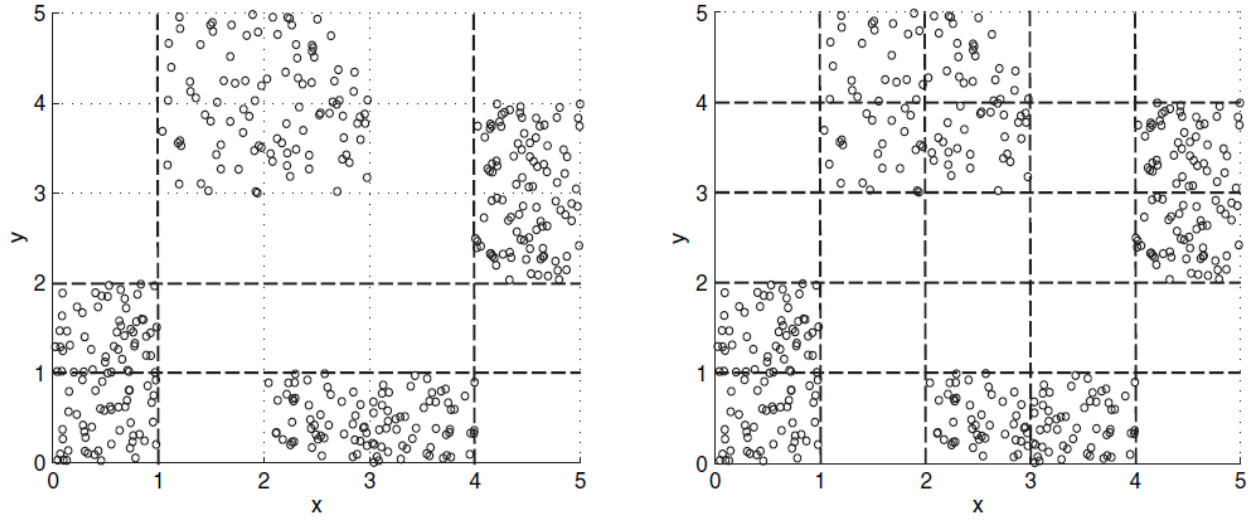


K-means approach to obtain 4 values.

Discretization in Supervised Settings

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 following example.



(a) Three intervals (b) Five intervals

Figure 2.14. Discretizing x and y attributes for four groups (classes) of points.

Binarization

Binarization maps a continuous or categorical attribute into one or more binary variables

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

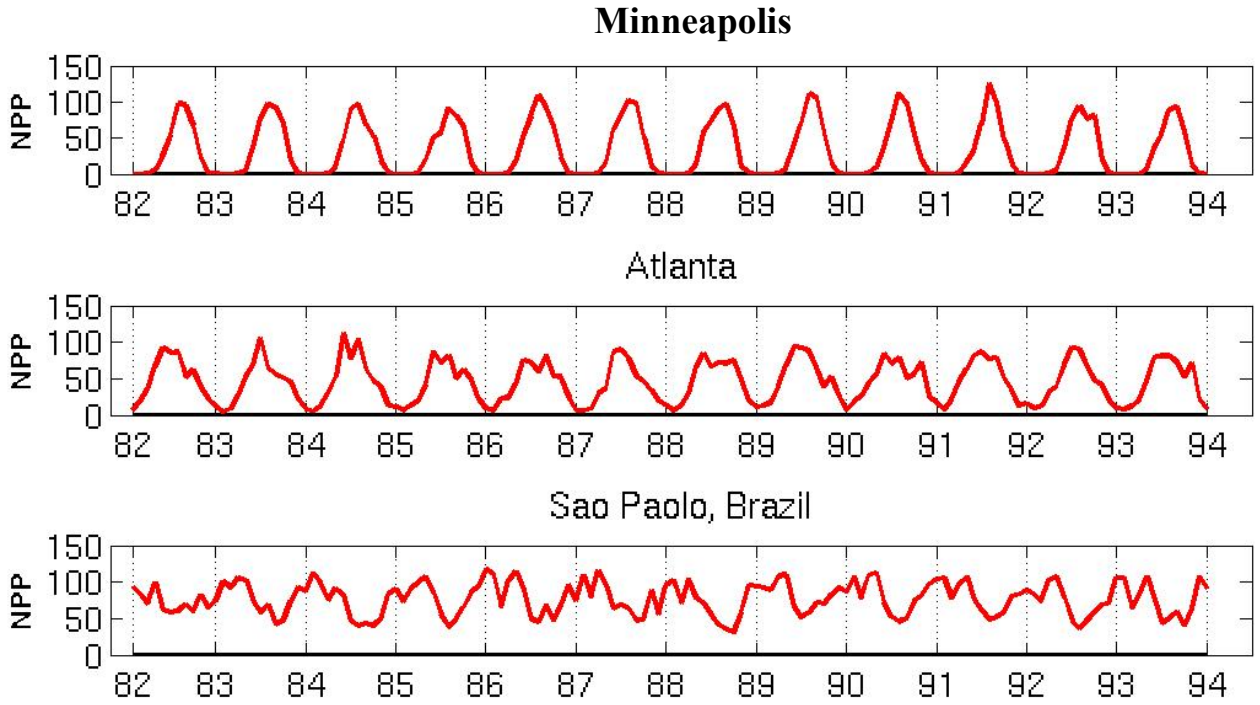
Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

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

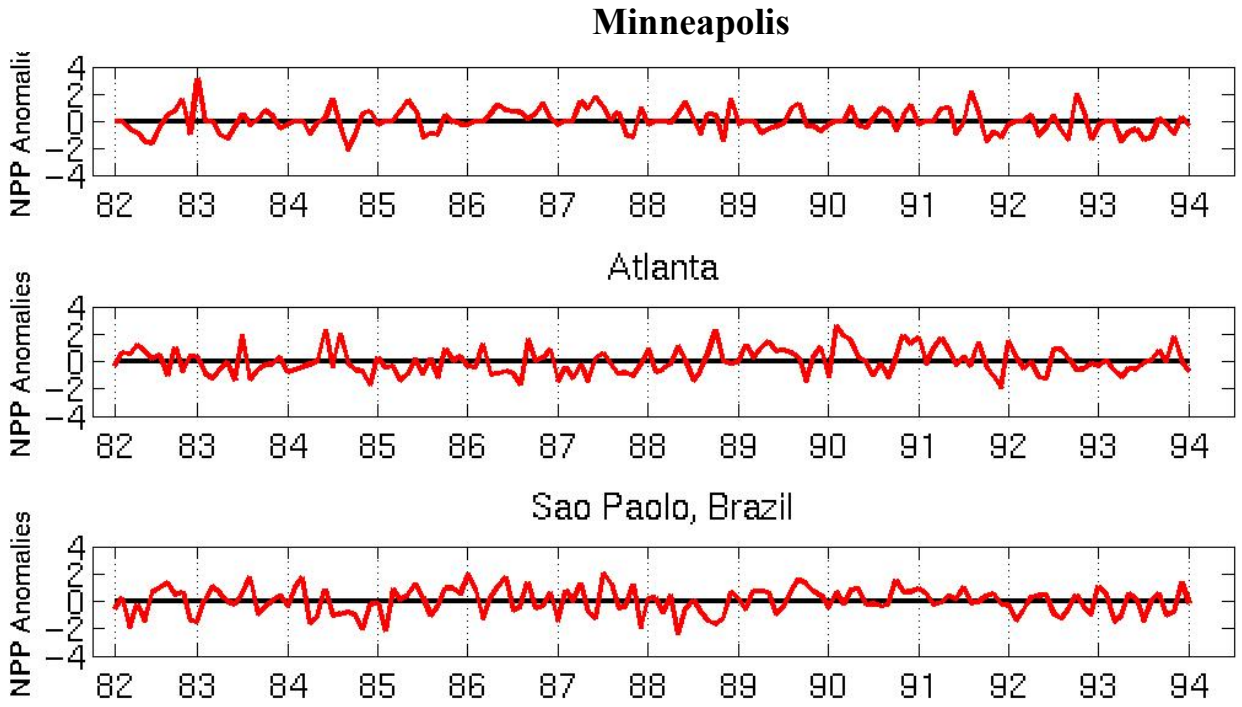


Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

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

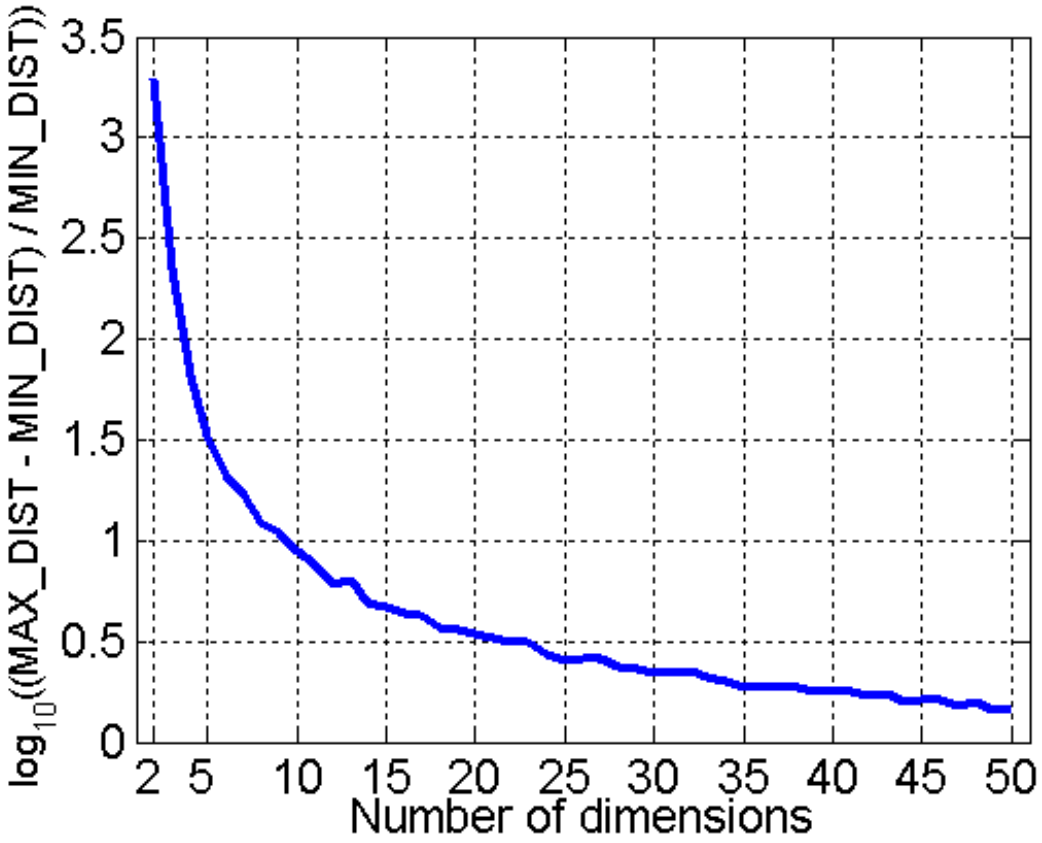
Correlations between time series

	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

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:

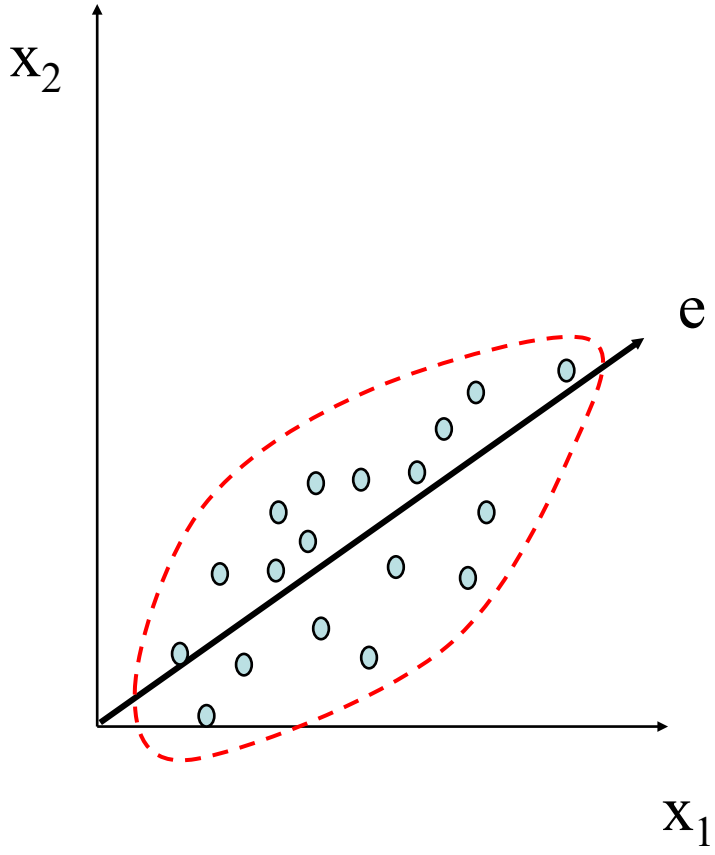
- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

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 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 Engineeering/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