UMD

UNIVERSITY OF MINNESOTA DULUTH

Driven to Discover℠

Lecture Notes for Topic 2: Data

Introduction to Data Mining , 2nd Edition

by

Tan, Steinbach, Karpatne, Kumar

Modified for CS 4232/5232

Outline

- Attributes and Objects
- Types of Data
- Data Quality
- Similarity and Distance
- Data Preprocessing

Outline

- **Attributes and Objects**
- Types of Data
- Data Quality
- Similarity and Distance
- Data Preprocessing

About Data

- We have a 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.
  - Attributes are also known as variables, fields, characteristics, dimensions, **predictors** or features
- A collection of attributes describes an ***object***
  - Objects are also known as records, points, cases, samples, entities, or instances

Attributes

Objects

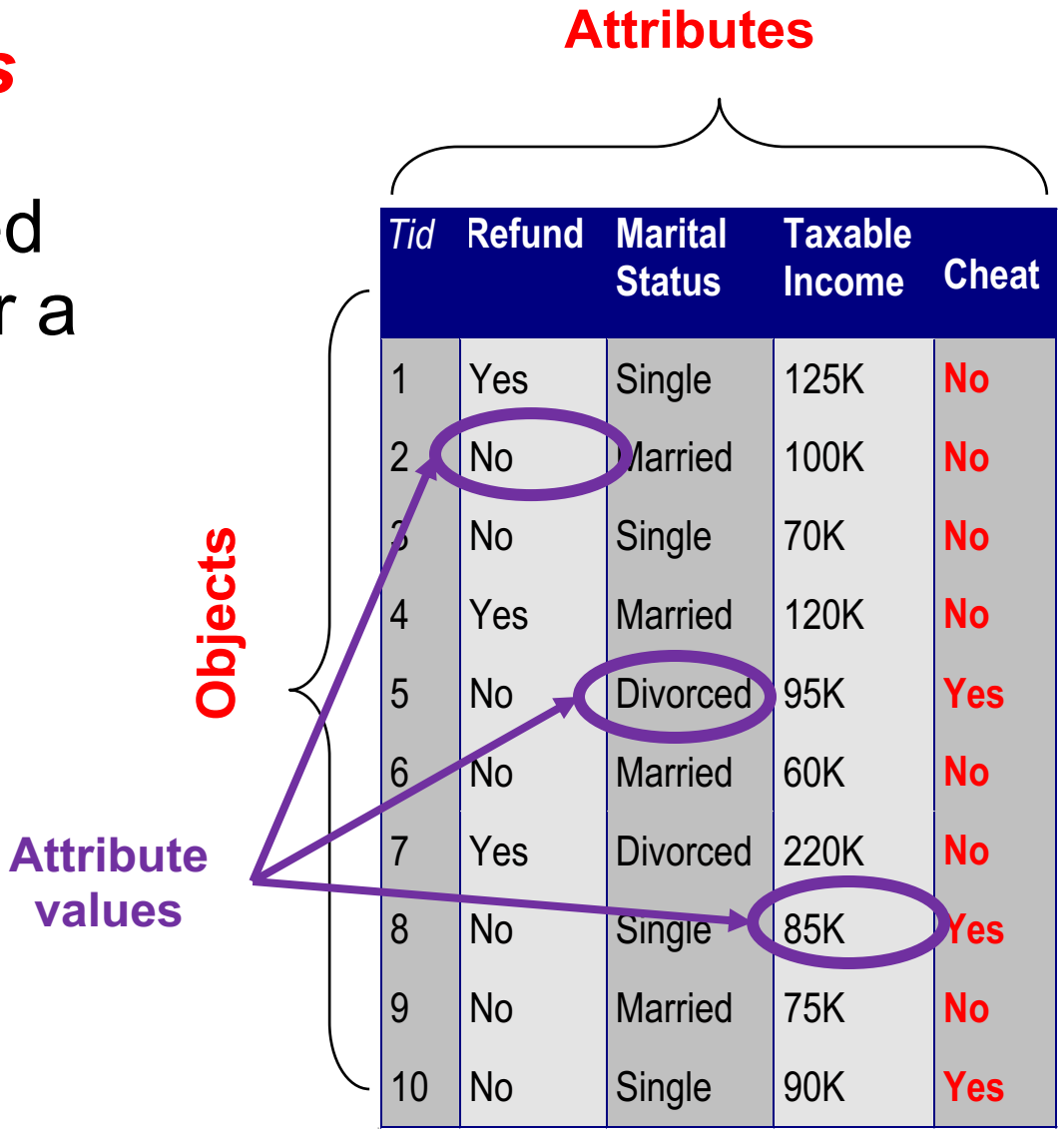
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

A More Complete View of Data

- Data may have parts
- The different parts of the data may have relationships
- More generally, data may have structure
- Data can be incomplete
- We will discuss this in more detail later

Attribute Values

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



Attribute Values (Cont'd)

- Distinction between **attributes** and **attribute values**
  - The 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 attribute 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:
  - Categorical (Qualitative)
    - 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}
  - Numeric (Quantitative)
    - 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

Properties of Attribute Values (Cont'd)

- **Nominal Attributes**: distinctness (=, ≠)
  - Does it make sense to say?:
    - Single = Married
    - Single ≠ Married

    Yes!
  - Does it make sense say?:
    - Single < Married

    No!

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

Nominal Attribute

Properties of Attribute Values (Cont'd)

- **Ordinal Attributes**: distinctness, **order**
  - Does it make sense to say?:
    - Freshman = Freshman
    - Junior ≠ Senior

    Yes!
  - Does it make sense to say?:
    - Sophomore < Junior

    Yes!

Tid	Name	Classification	Pass
1	Cody	Freshman	Yes
2	Zach	Freshman	Yes
3	Ahmed	Sophomore	Yes
4	Yu	Sophomore	No
5	Pedro	Freshman	Yes
6	Jackson	Junior	No
7	Ashley	Freshman	No
8	Bre	Senior	Yes
9	Emily	Junior	No
10	Andrew	Sophomore	Yes

Properties of Attribute Values (Cont'd)

- **Interval Attributes**: distinctness, order, **differences**
  - Does it make sense to say?:
    - Single = Married
    - Single ≠ Married
  - Does it make sense to say?:
    - Single < Married

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

Nominal Attribute

Properties of Attribute Values (Cont'd)

- **Ratio Attributes**: distinctness, order, difference, **ratio**
  - Does it make sense to say?:
    - 120K < 220K

      Yes!
  - Does it make sense say?
    - 60K is half of 120K

      Yes!

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

Ratio Attribute

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?

  } **No!**
  - the Kelvin scale?

  } **Yes!**

- 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

Operations According to Attribute Type

- **Nominal Attributes**
  - Mode:
    - Most common value(s)
      - E.g.: The mode of marital status is either ‘Single’ or ’Divorced’ (4 each)
  - Can you compute the median or the mean of marital status?

No!

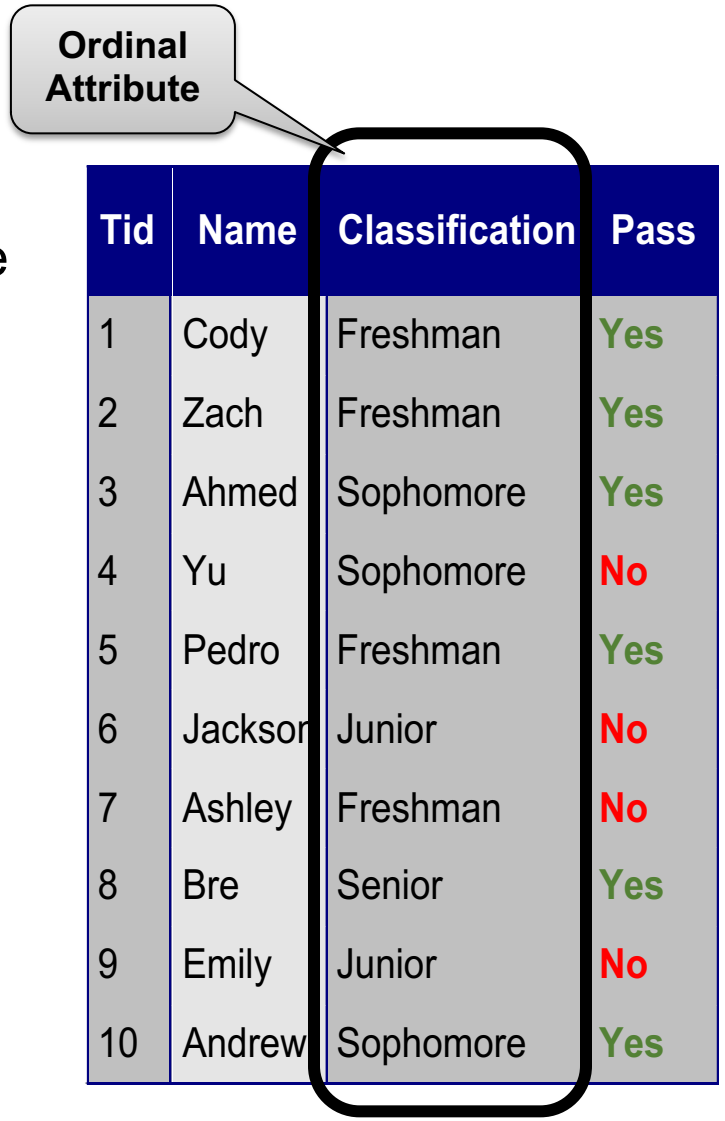
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Nominal Attribute

Operations According to Attribute Type

- **Ordinal Attributes**
  - Median:
    - The attribute value right in the middle (if all values are sorted)
      - E.g.: To compute the median of classification: {F,F,So,So,F,J,F,Se,J,So}, sort them: {F,F,F,F,**So,So**,So,J,J,Se} and find the mid value.
      - Can you compute the mean of Classification?

No!



Tid	Name	Classification	Pass
1	Cody	Freshman	Yes
2	Zach	Freshman	Yes
3	Ahmed	Sophomore	Yes
4	Yu	Sophomore	No
5	Pedro	Freshman	Yes
6	Jackson	Junior	No
7	Ashley	Freshman	No
8	Bre	Senior	Yes
9	Emily	Junior	No
10	Andrew	Sophomore	Yes

	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 Attributes
  - Have 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 Attributes
  - Have real numbers as attribute values
  - Examples: temperature, height, or weight
  - In practice, 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

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?
- Number of police stops at intersections
- Passenger class on an airplane
- Average GPA
- Age of a person

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

Example with R

- Study the file “Data Types – CS4232.Rmd” posted on Canvas. This file contains examples of the features of the different data types.

Outline

- Attributes and Objects
- **Types of Data**
- Data Quality
- Similarity and Distance
- Data Preprocessing

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 the size of the 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

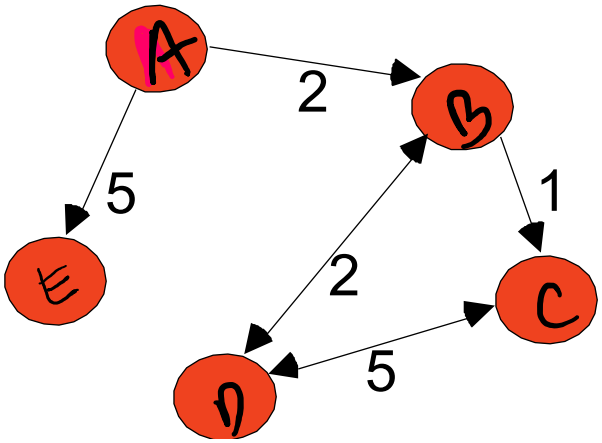
Transaction Data

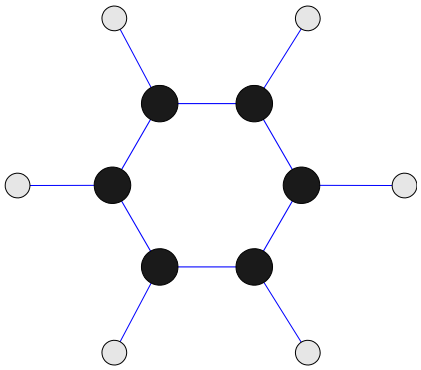
- A special type of record data, where
  - Each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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.

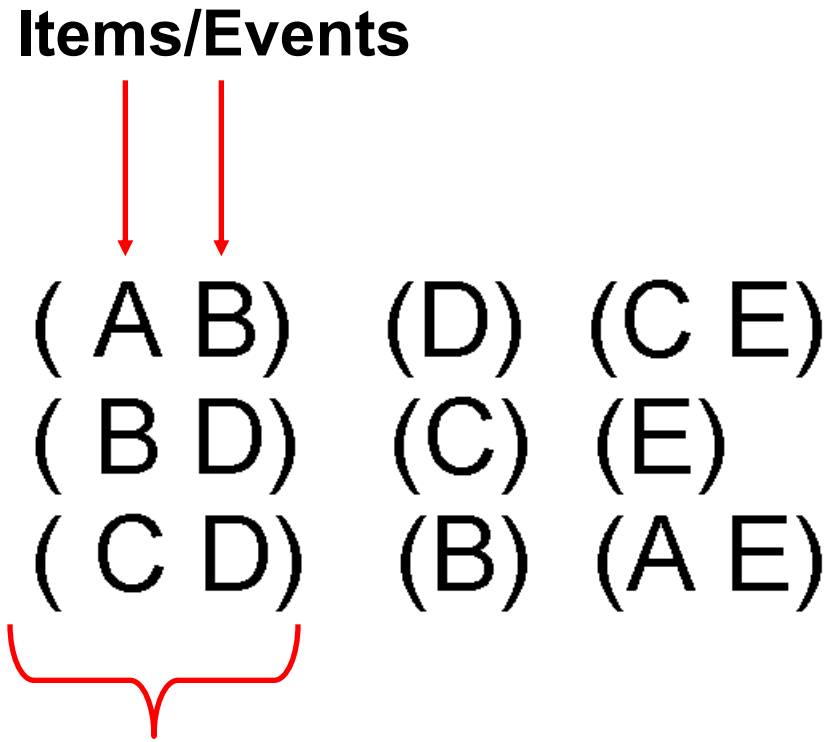
General Data Mining

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



The first of these sequences means:

- **I buy bread (B) and Milk (A) together first**
- **Then I buy pizza (D)**
- **Then I buy (together) diapers (E) and Reese's peanut butter cups (C)**

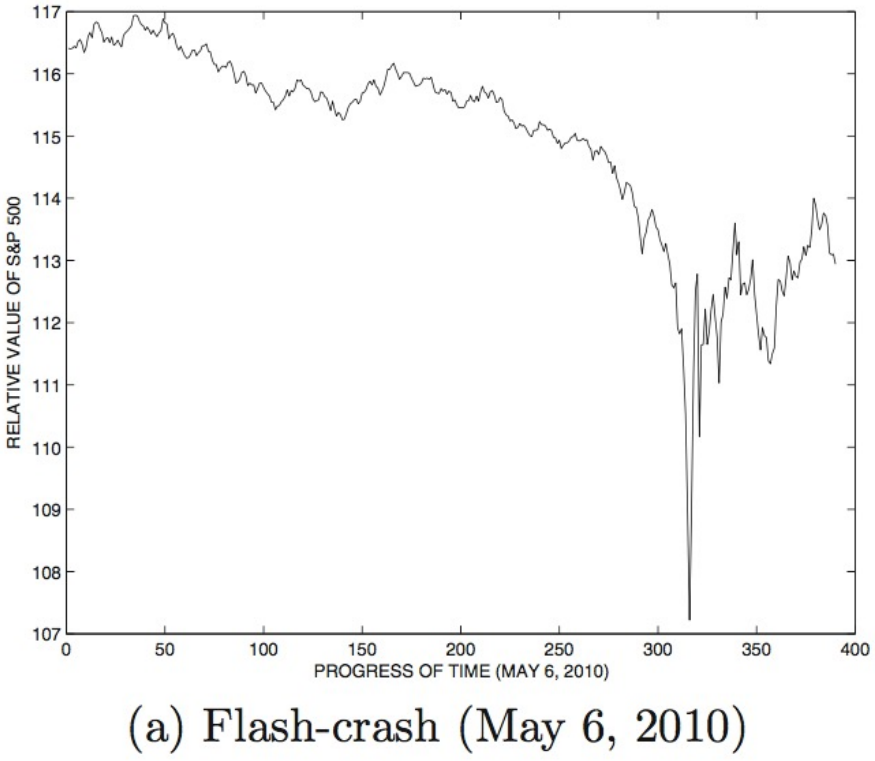
Ordered Data

- Genomic sequence data

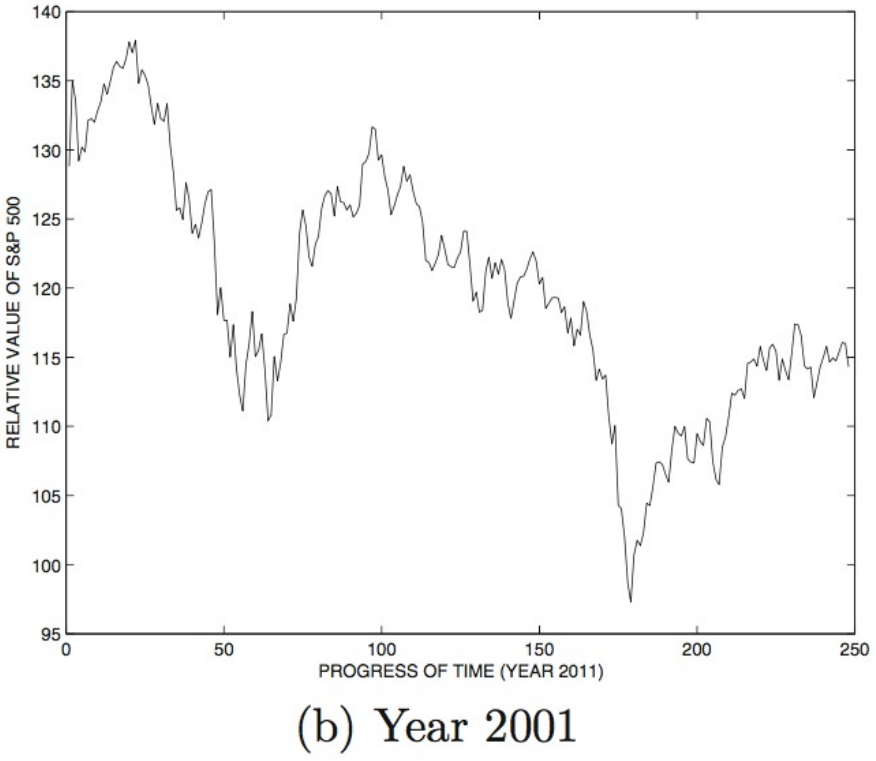
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Time Series Data

S&P500 Index on the day of the flash-crash (May 6, 2010) and during 2001



(a) Flash-crash (May 6, 2010)



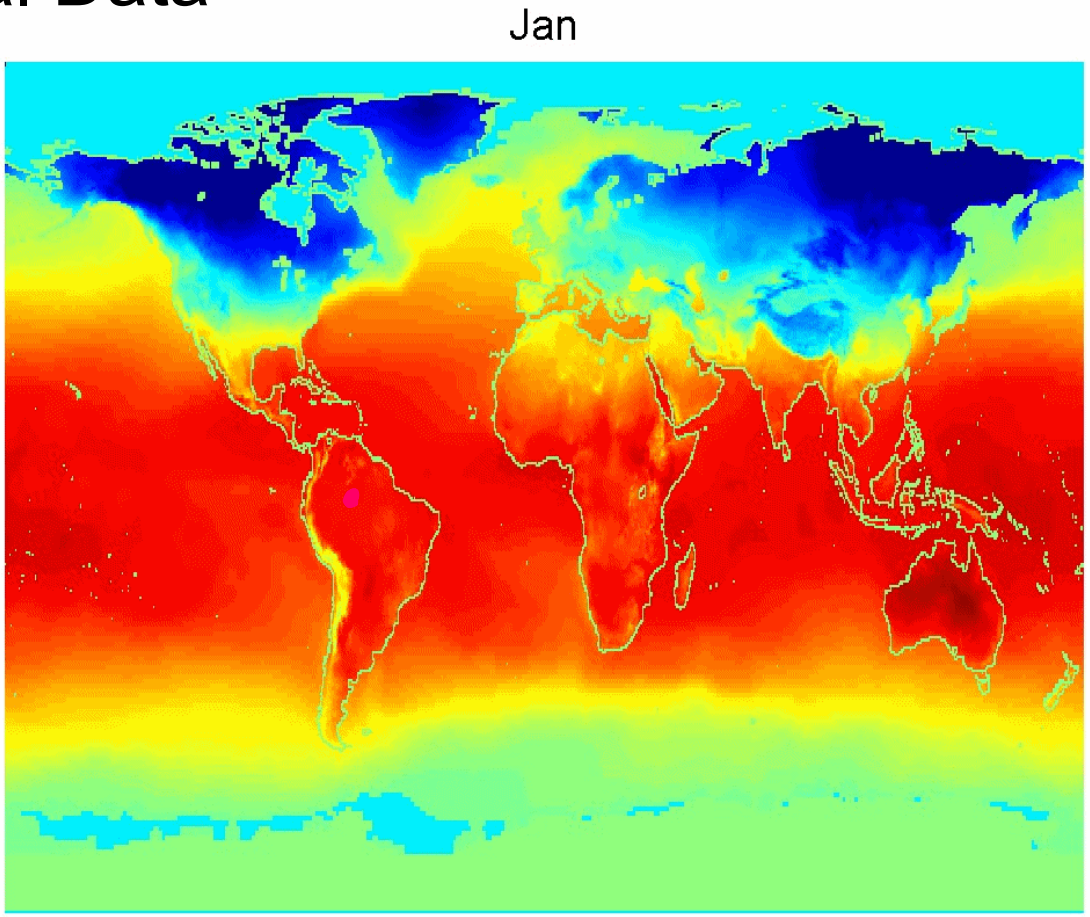
(b) Year 2001

Figure from Aggarwal, Charu. Outlier Analysis, 2nd edition, Springer 2018

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Outline

- Attributes and Objects
- Types of Data
- **Data Quality**
- Similarity and Distance
- Data Preprocessing

Data Quality


Harvard Business Review

Data | Bad Data Costs the U.S. $3 Trillion Per Year

DATA

Bad Data Costs the U.S. $3 Trillion Per Year

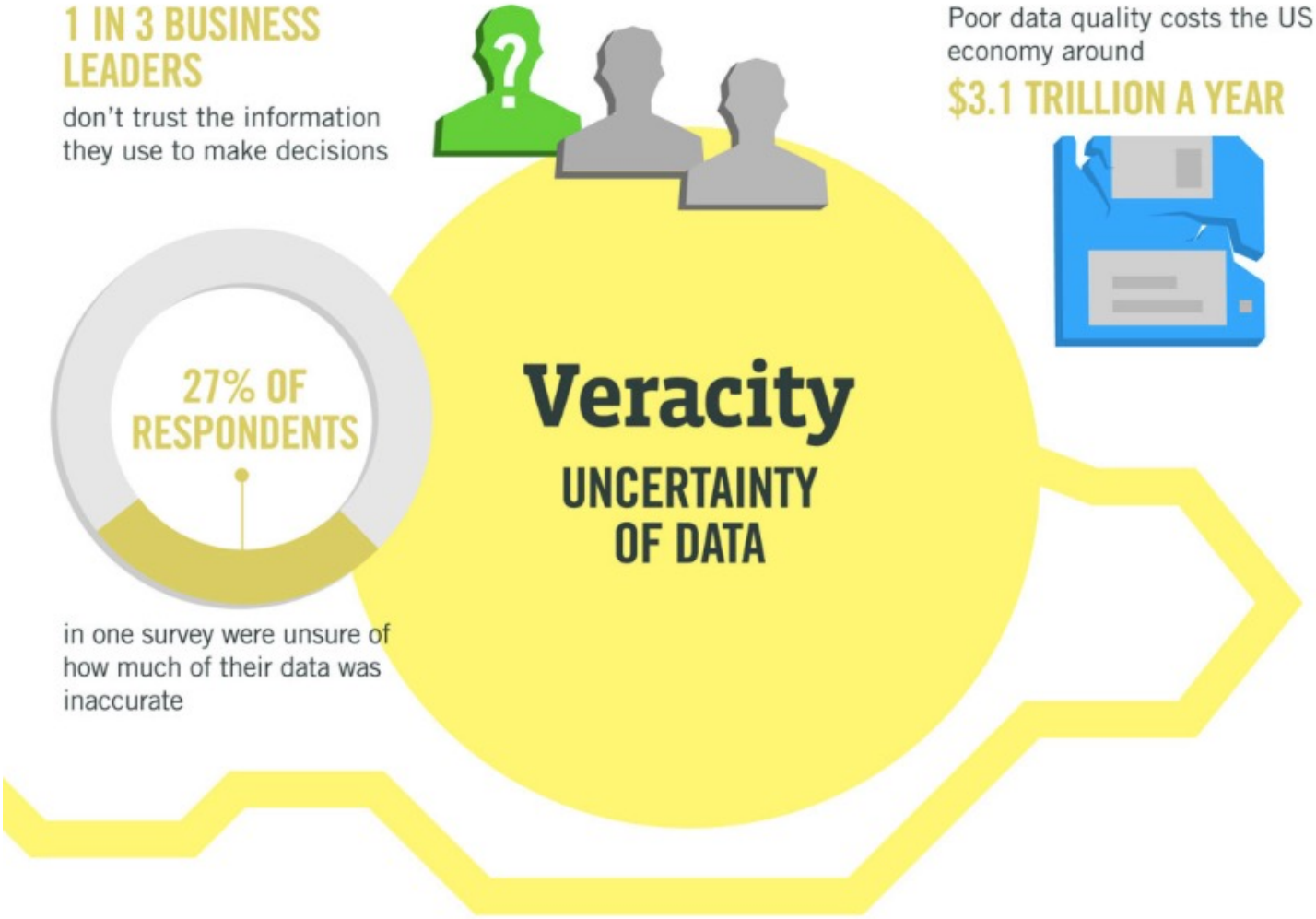
by Thomas C. Redman

September 22, 2016

Save Share 10 Comment Print $8.95 Buy Copies

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

Data Quality



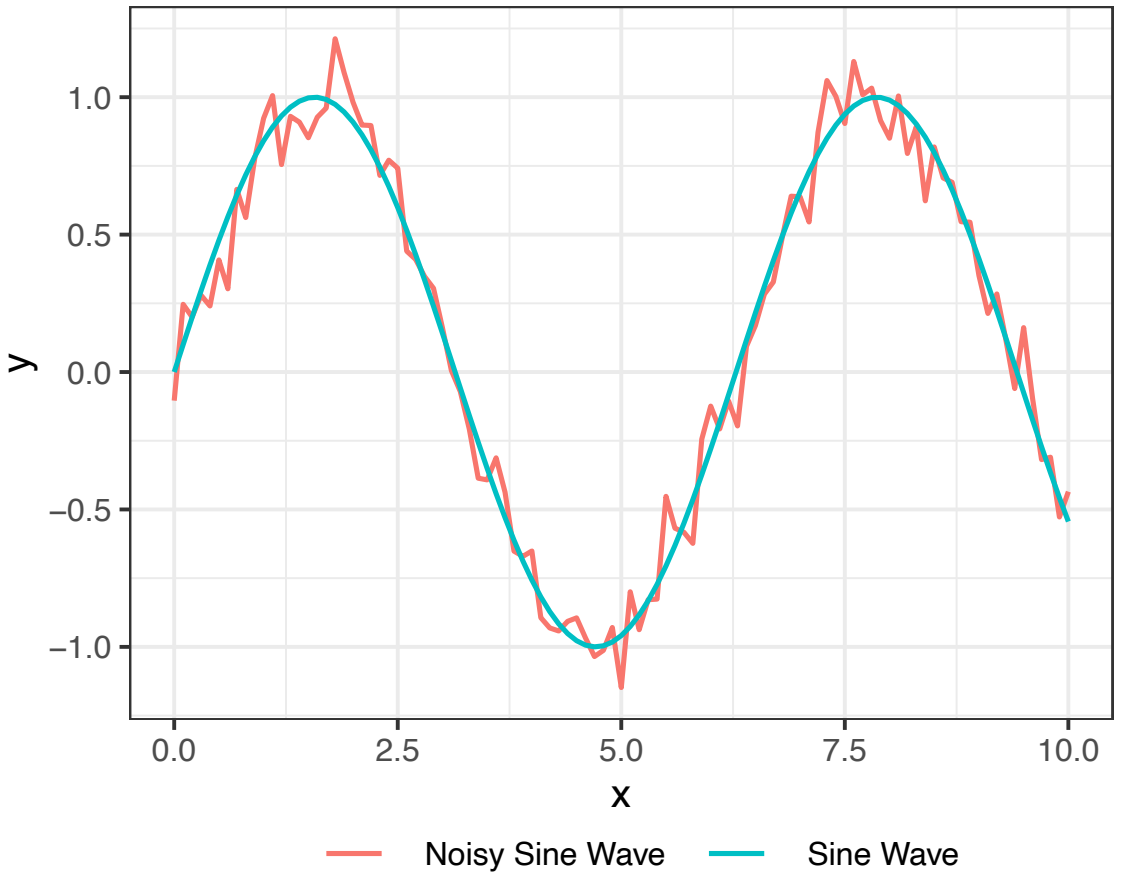
Data Quality (Cont'd)

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

Noise

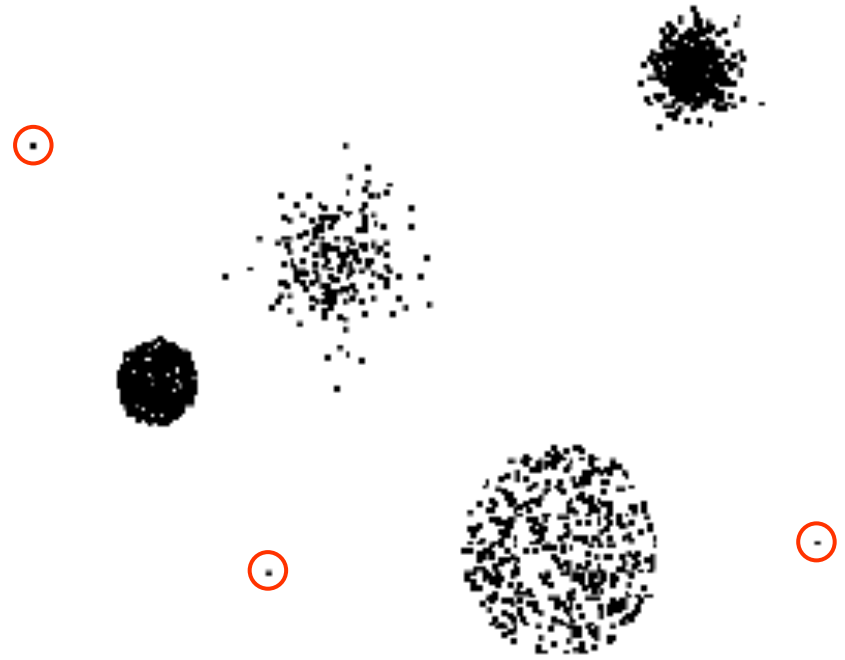
- For objects, **noise** is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on a television screen

Sine wave with superimposed gaussian noise



Outliers

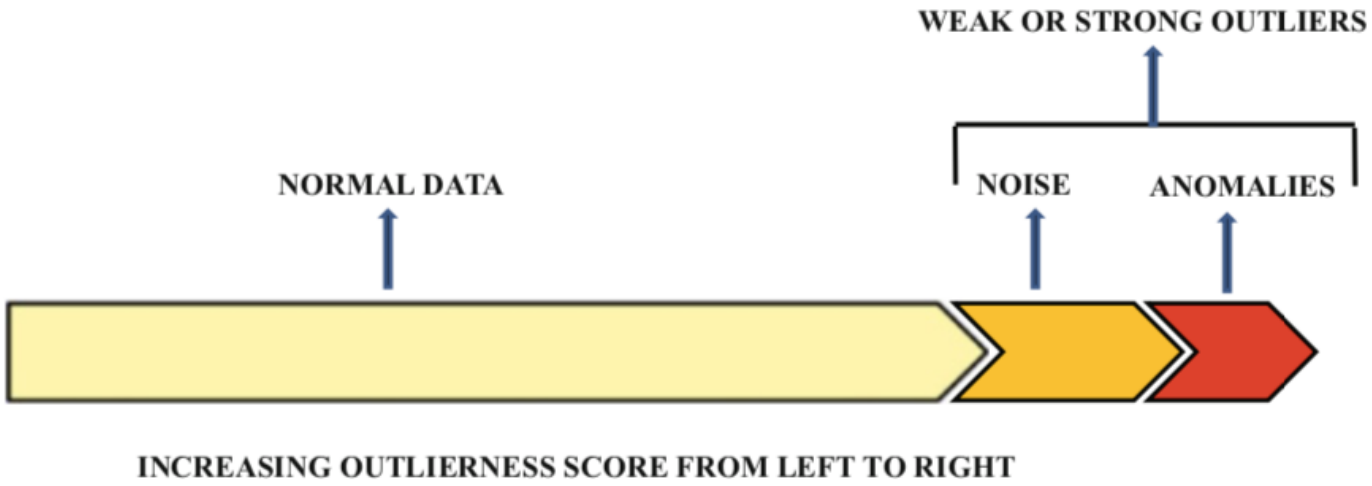
- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection



- Causes?

Outliers vs. Noise

- There is a “spectrum of anomalies”



- Outliers are anomalies that are interesting to the application, while noise is not.

Figure from C. C. Aggarwal. Outlier Analysis. Springer 2018

Missing Values

- Reasons for missing values
  - Information was 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)

age <dbl>	workclass <chr>	fnlwgt <dbl>	education <chr>	education-num <dbl>	marital-status <chr>	occupation <chr>
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty
54	Private	302146	HS-grad	9	Separated	Other-service
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving
59	Private	109015	HS-grad	9	Divorced	Tech-support
56	Local-gov	216851	Bachelors	13	Married-civ-spouse	Tech-support
19	Private	168294	HS-grad	9	Never-married	Craft-repair
54	*NA*	180211	Some-college	10	Married-civ-spouse	*NA*
39	Private	367260	HS-grad	9	Divorced	Exec-managerial
49	Private	193366	HS-grad	9	Married-civ-spouse	Craft-repair

21-30 of 32,561 rows | 1-7 of 15 columns

Previous 1 2 3 4 5 6 ... 100 Next

Missing Values (Cont'd)

- Handling missing values
  - Eliminate data objects or variables with missing values
  - Estimate missing values
    - Using the mean/median of values of the column (quantitative attributes)
    - Using the mode of the values column (quantitative attributes)
    - Finding the k-most similar observations (rows), and then computing the mean/median/mode for the values of that attribute that has the missing value.
    - Many other strategies
  - 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 Data
- Data Quality
- **Similarity and Distance**
- Data Preprocessing

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

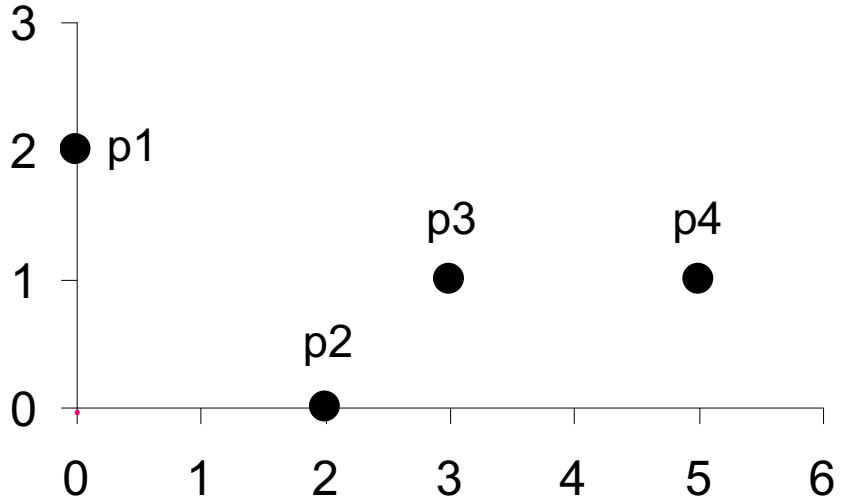
- Euclidean Distance

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

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

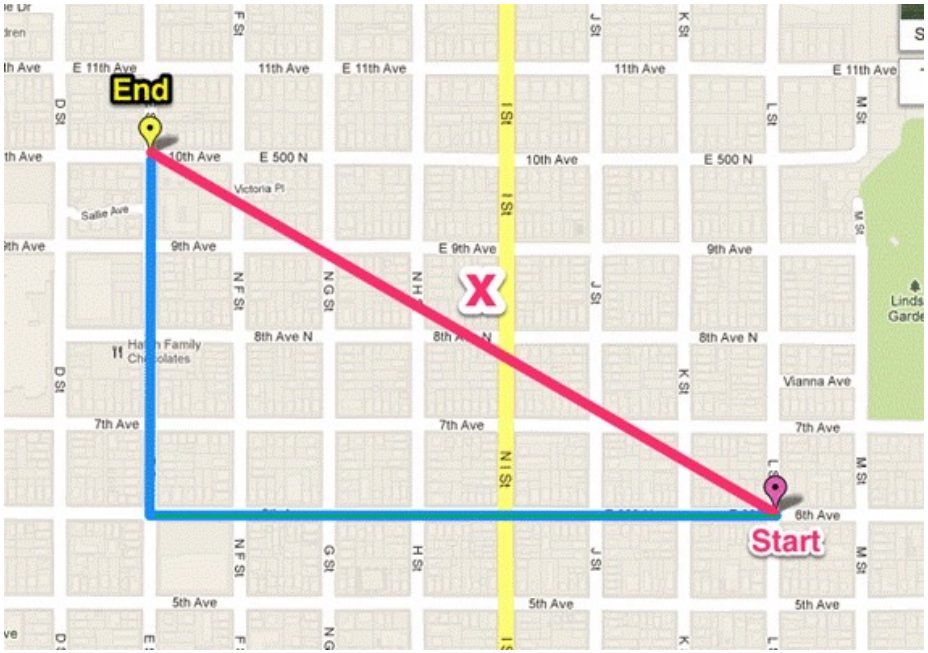
- Minkowski Distance is a generalization of Euclidean Distance

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

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{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



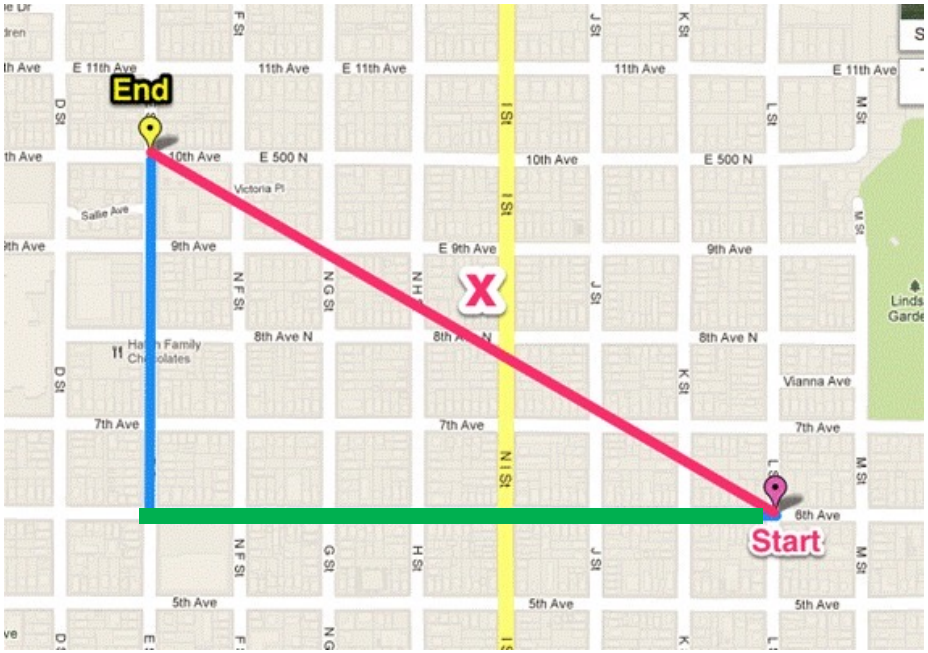
The red line is the Euclidean distance between Start and End. The blue line is the taxicab distance

Figure from https://geoidlife.com/2018/01/16/is-your-earth-flat/

Minkowski Distance: Examples

- $r \rightarrow \infty$. "supremum" or infinity (L_{max} norm, L_{∞} 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.



The red line is the Euclidean distance between Start and End. The green line is the infinity norm

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

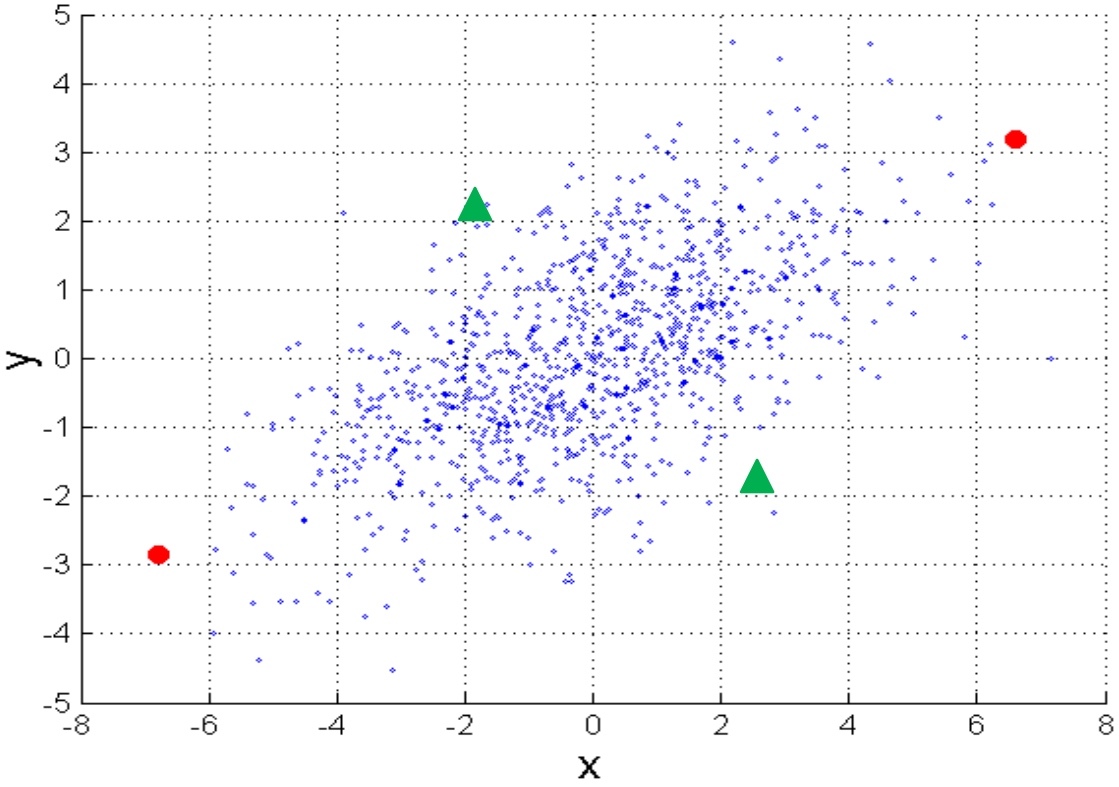
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Mahalanobis Distance

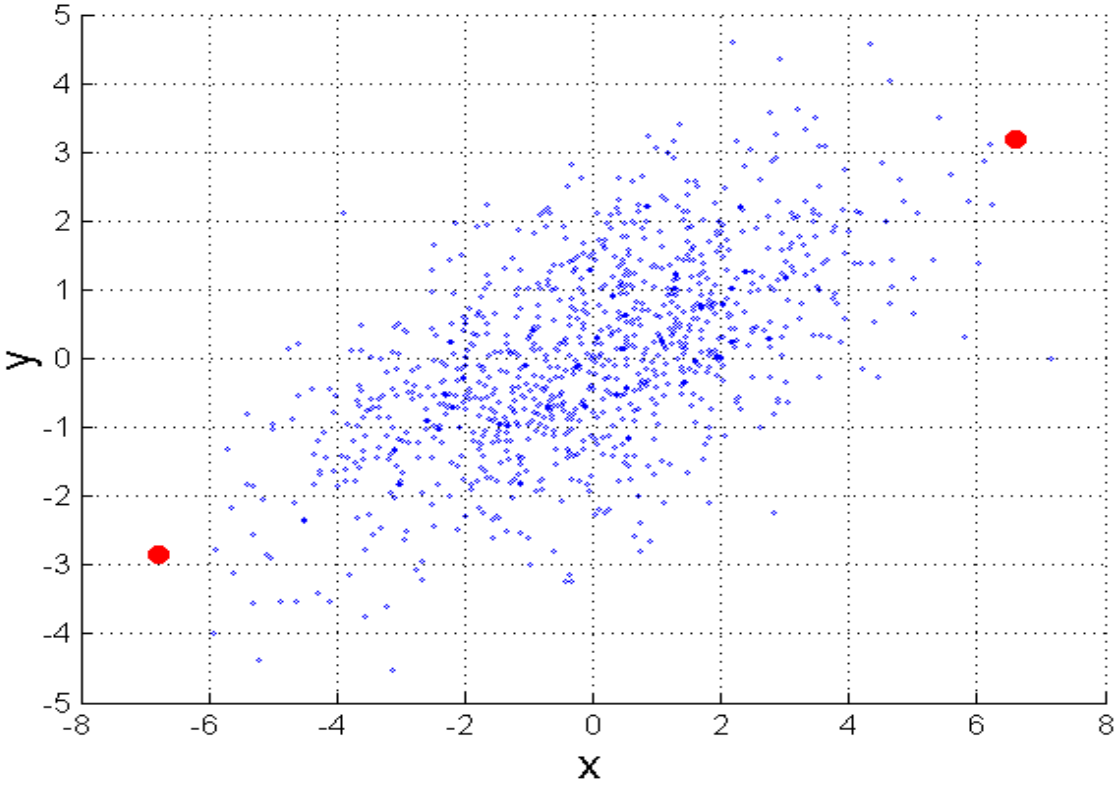
Which pair of points is farther apart from each other? The red pair or the green pair?



- The standard deviation along the line connecting the red dots is much larger than along the line connecting the green points.
- Therefore, we can argue that the green points are further apart than the red ones

Mahalanobis Distance

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



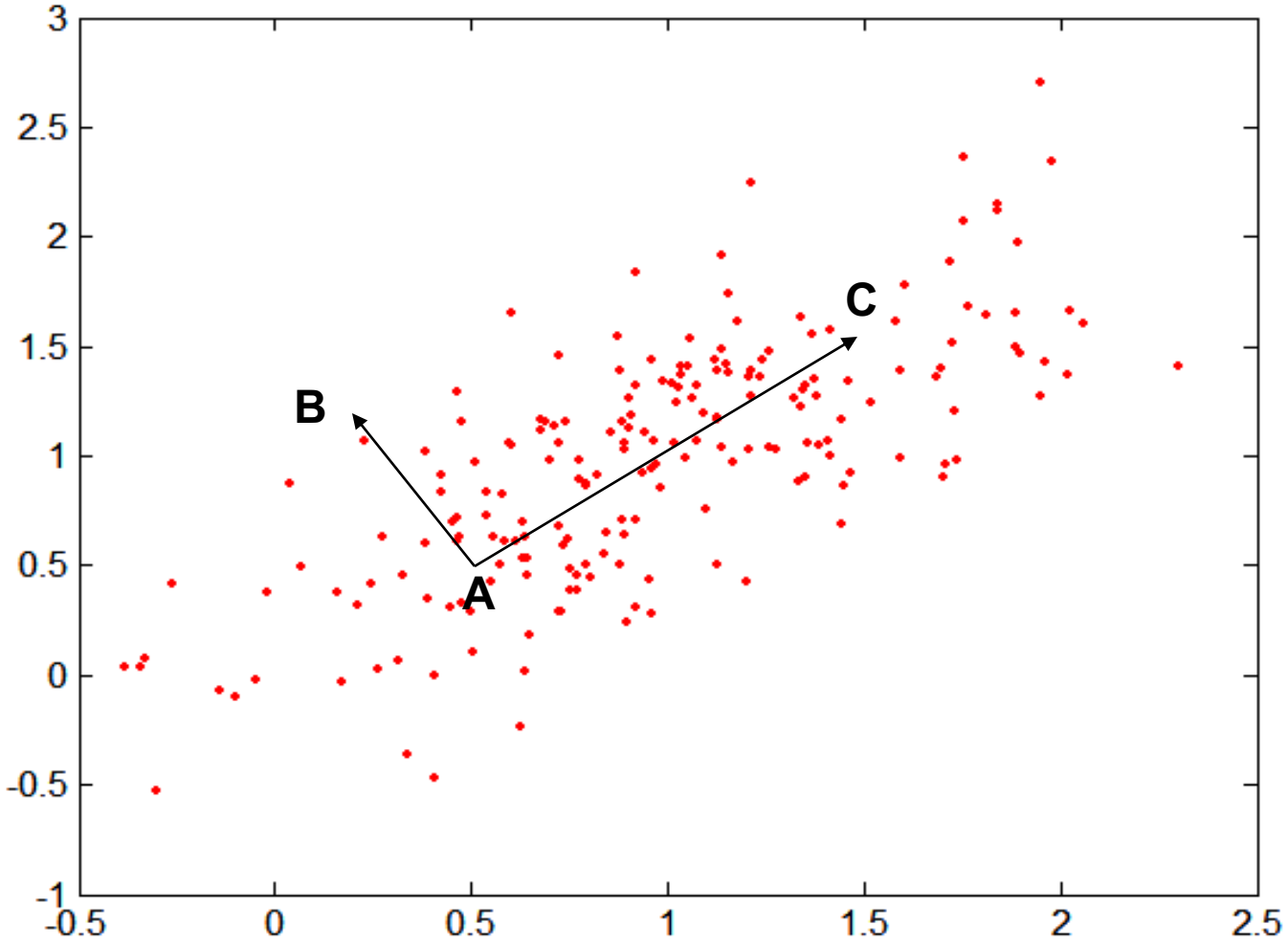
Σ **is the covariance matrix**

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y \\ \sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Σ^{-1} **is the precision matrix**

For the red points, the Euclidean distance is 14.7, while the 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)

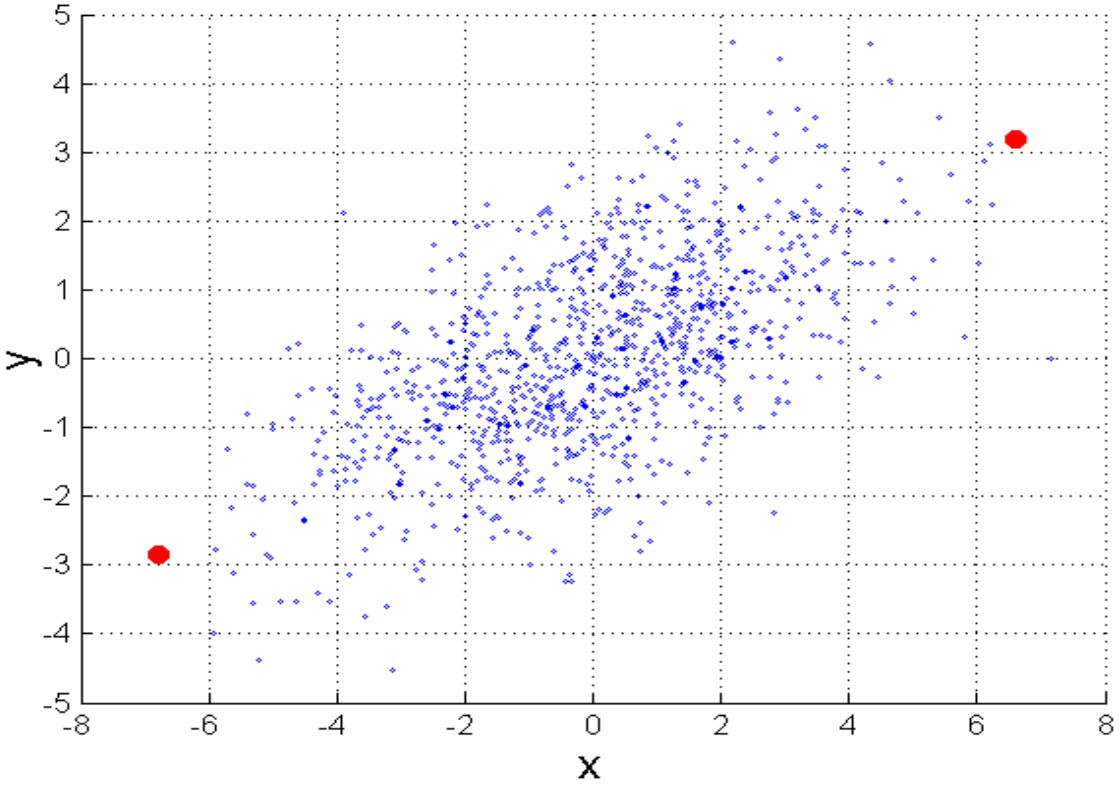
C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Mahalanobis Distance

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



Σ is the covariance matrix

Σ^{-1} is the precision matrix

What is the relationship between the Mahalanobis distance and the Euclidean distance?

Take Σ to be the identity matrix, and you get the Euclidean distance

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

- A 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

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

Jaccard ignores the f_{00} matches!

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)

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

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

Jaccard is for scenarios where the attributes are likely to have many zeros. E.g., if the vector x[i] = 1 represents "purchased item i" and x[i] = 0 means "did not purchase item i." Most likely, a person would only purchase a small portion of a store's catalog

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 norm of vector $\mathbf{d}$.

- Example:

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

$$\mathbf{d}_2 = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$$

$\langle \mathbf{d}_1, \mathbf{d}_2 \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

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

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

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

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

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

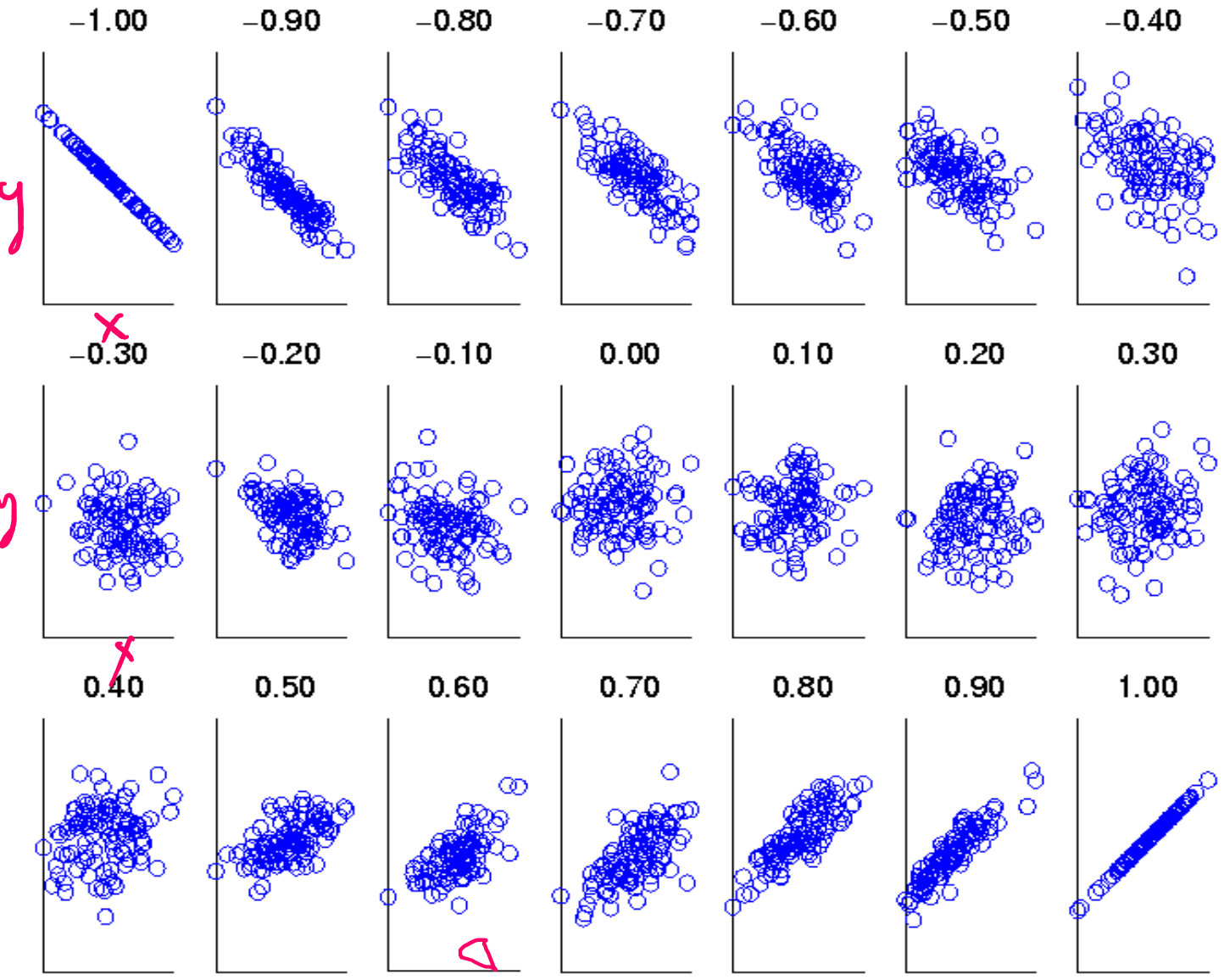
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

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

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

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

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1

Drawback of Correlation

- Suppose we have two vectors:
  - **x** = (-3, -2, -1, 0, 1, 2, 3)
  - **y** = (9, 4, 1, 0, 1, 4, 9)

satisfying $y_i = x_i^2$

- Notice that:
  - mean(**x**) = 0, mean(**y**) = 4
  - std(**x**) = 2.16, std(**y**) = 3.74
- Therefore:
  - corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74) **= 0**

The correlation is 0 despite that the two vectors are functionally related!

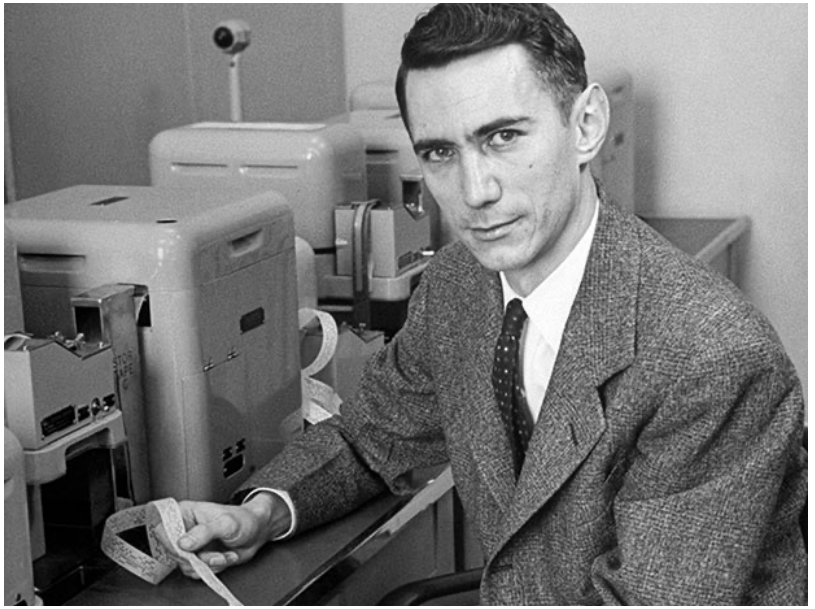
Comparison of Proximity Measures

- 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



Claude Shannon, the father of Information Theory

Information and Probability

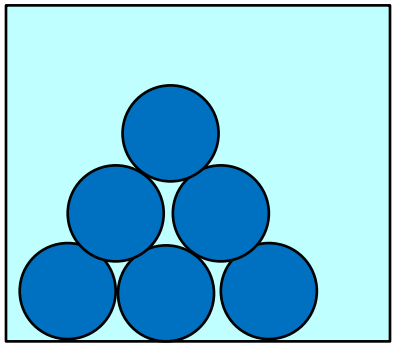
- Information relates to possible outcomes of an event
  - transmission of a message, flip of a coin, or measurement of a piece of data

- The more certain an outcome, the less information that it contains and vice-versa
  - For example, an outcome of heads for fair coin 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

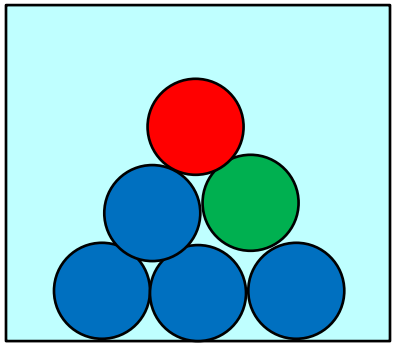
Which of the following boxes has the most entropy? Which one the least entropy?

Box 1

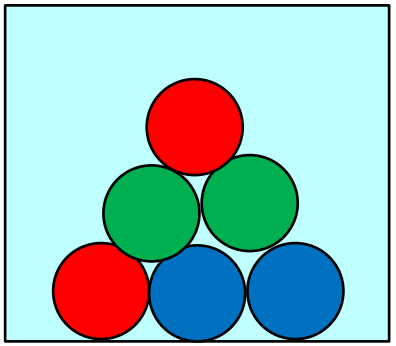


The least entropy (least impure)

Box 2



Box 3



The most entropy (most impure)

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 possible value of 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^{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

General Approach for Combining Similarities

- Suppose we are given the dataset

age <dbl>	workclass <chr>	fnlwgt <dbl>	education <chr>	education-num <dbl>	marital-status <chr>	occupation <chr>
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty
54	Private	302146	HS-grad	9	Separated	Other-service
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving
59	Private	109015	HS-grad	9	Divorced	Tech-support
56	Local-gov	216851	Bachelors	13	Married-civ-spouse	Tech-support
19	Private	168294	HS-grad	9	Never-married	Craft-repair
54	*NA*	180211	Some-college	10	Married-civ-spouse	*NA*
39	Private	367260	HS-grad	9	Divorced	Exec-managerial
49	Private	193366	HS-grad	9	Married-civ-spouse	Craft-repair

21-30 of 32,561 rows | 1-7 of 15 columns

Previous 1 2 3 4 5 6 ... 100 Next

- How do we measure the similarity between rows if they have multiple attributes, each with a different type?

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

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

2) Define an indicator variable, δ_k, for the k^{th} attribute as follows:

   δ_k = 0 if the k^{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

   δ_k = 1 otherwise

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

Using Weights to Combine Similarities

- It may be the case that we do not want to treat all attributes the same
  - Use non-negative weights ω_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

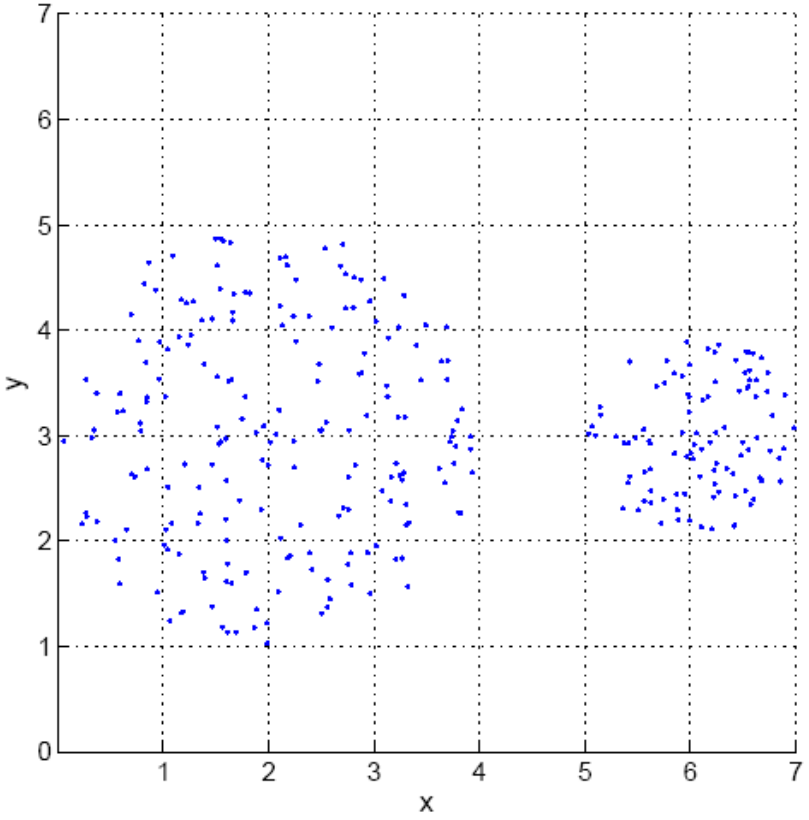
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

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

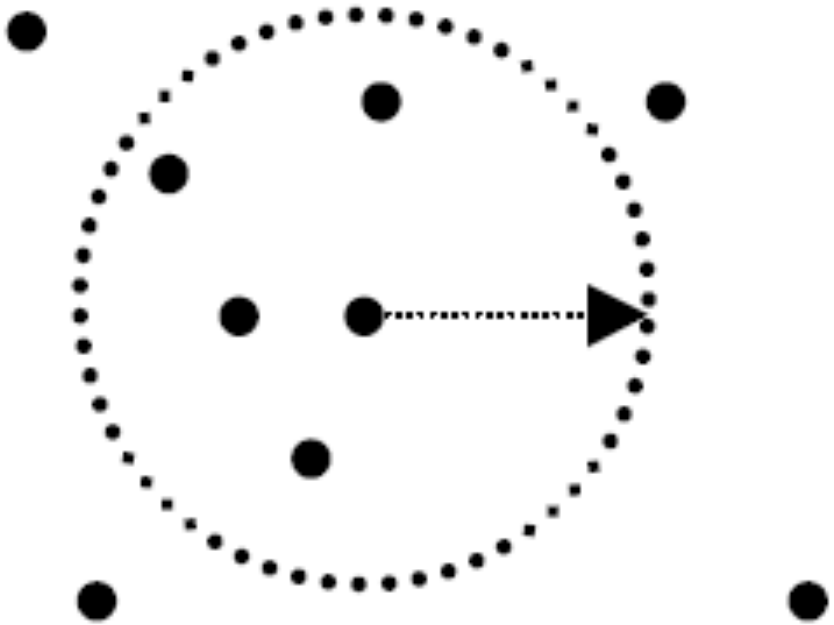


Illustration of center-based density

Outline

- Attributes and Objects
- Types of Data
- Data Quality
- Similarity and Distance
- **Data Preprocessing**

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

Aggregation Example

age <dbl>	workclass <chr>	fnlwgt <dbl>	education <chr>
39	State-gov	77516	Bachelors
50	Self-emp-not-inc	83311	Bachelors
38	Private	215646	HS-grad
53	Private	234721	11th
28	Private	338409	Bachelors
37	Private	284582	Masters
49	Private	160187	9th
52	Self-emp-not-inc	209642	HS-grad
31	Private	45781	Masters
42	Private	159449	Bachelors

- We can do **aggregation** manually with dplyr by using `mutate` and `case_when` to create a new variable

```r
```{r}
adult <- adult %>%
  mutate(`Age Category` = case_when(age <=45 ~ "Young Adult",
                                    45 < age & age <=65 ~ 'Old Age',
                                    TRUE ~ "Middle Age"),
         `Age Category` = factor(`Age Category`,
                                 levels = c('Young Adult', 'Middle Age', 'Old Age'),
                                 ordered = TRUE)) %>%
  relocate( `Age Category`, .after = age)
adult
```

age <dbl>	Age Category <ord>	workclass <fctr>	fnlwgt <dbl>	education <chr>
39	Young Adult	State-gov	77516	Bachelors
50	Old Age	Self-emp-not-inc	83311	Bachelors
38	Young Adult	Private	215646	HS-grad
53	Old Age	Private	234721	11th
28	Young Adult	Private	338409	Bachelors
37	Young Adult	Private	284582	Masters
49	Old Age	Private	160187	9th
52	Old Age	Self-emp-not-inc	209642	HS-grad
31	Young Adult	Private	45781	Masters
```

Example: Precipitation in Australia

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

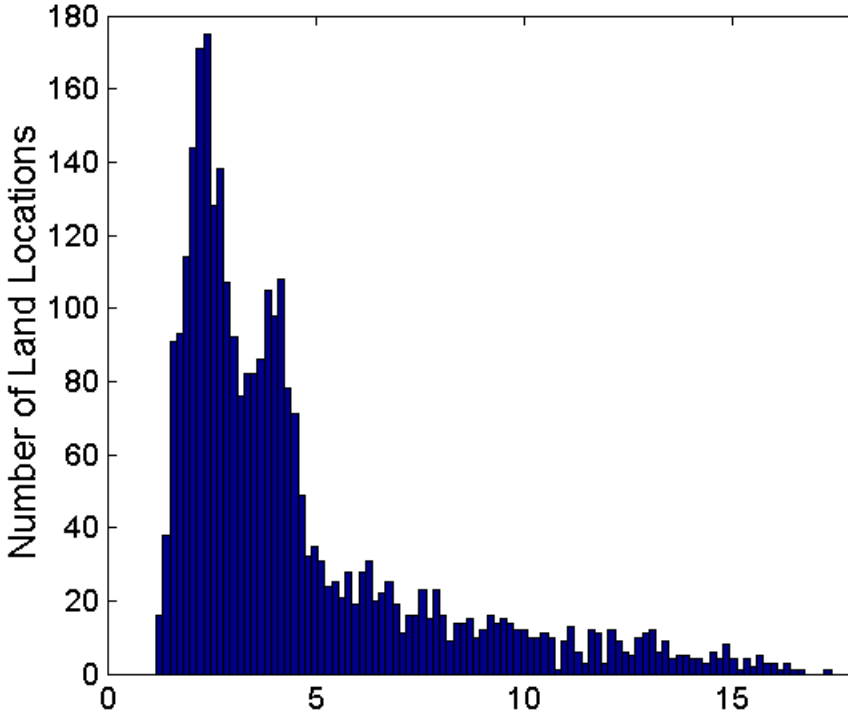
  The next slide shows

  - 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.

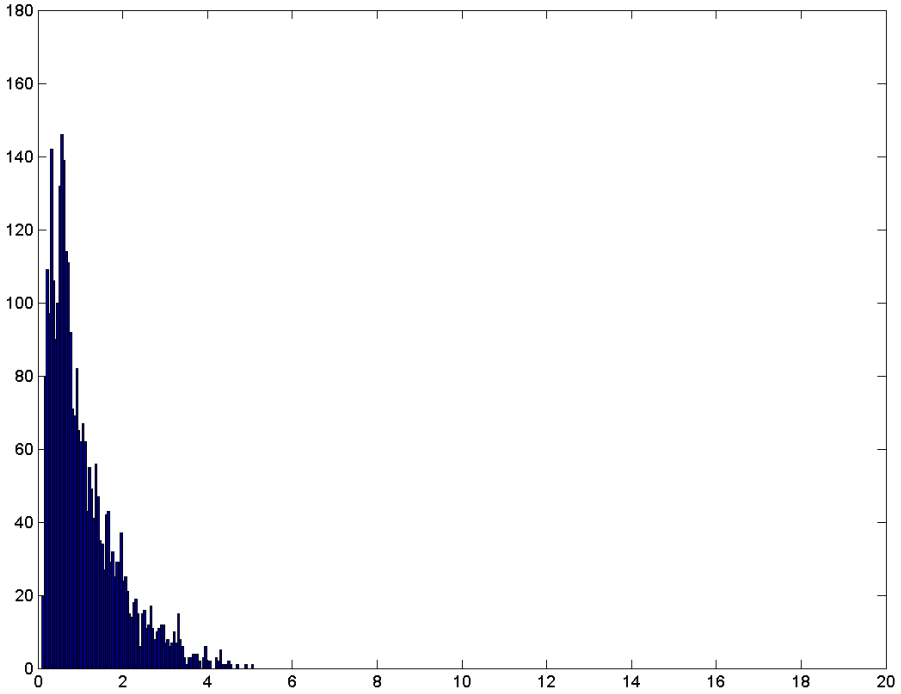
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.

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.

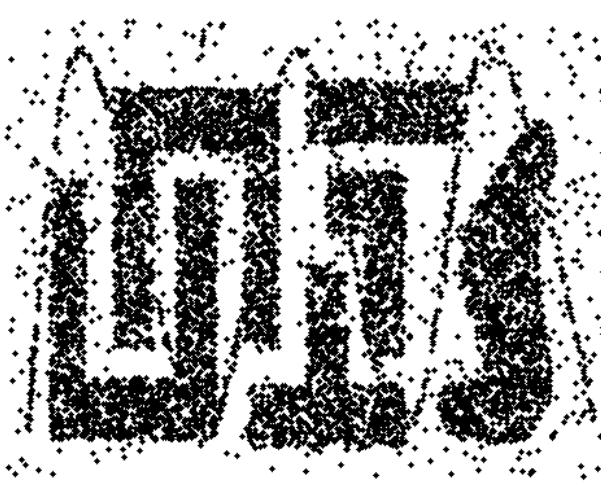
- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.

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

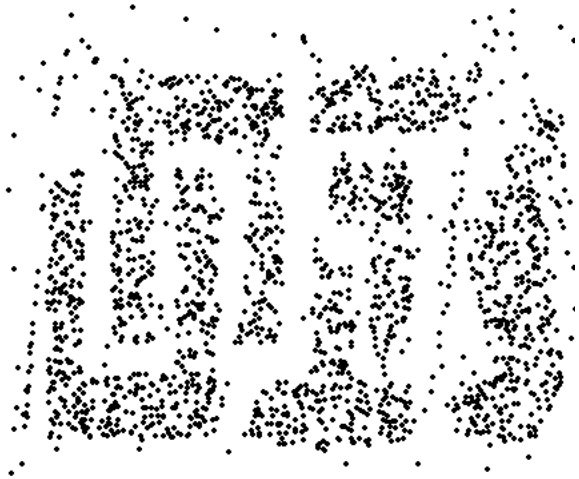
Sampling

- 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

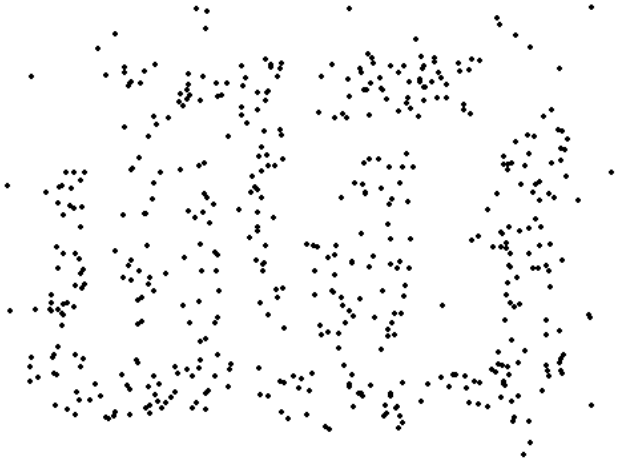
Sample Size



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

Simple Random Sampling from Data Frames in R

- We can do **simple random sampling** in R using the command sample_n of dplyr

```r
set.seed(0)
starwars %>% sample_n(size = 5, replace = FALSE)
```

Without replacement

A tibble: 5 × 14

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Han Solo	180	80	brown	fair	brown
Dexter Jettster	198	102	none	brown	yellow
Sebulba	112	40	none	grey, red	orange
Luke Skywalker	172	77	blond	fair	blue
Jar Jar Binks	196	66	none	orange	orange

5 rows | 1–10 of 14 columns

- Without dplyr, we need to use the sample command

```r
set.seed(1143)
starwars[sample(nrow(starwars), size = 5, replace = FALSE), ]
```

A tibble: 5 × 14

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Rey	*NA*	*NA*	brown	light	hazel
Chewbacca	228	112	brown	unknown	blue
Cordé	157	*NA*	brown	light	brown
Luke Skywalker	172	77	blond	fair	blue
IG-88	200	140	none	metal	red

5 rows | 1–10 of 14 columns

Stratified Random Sampling from Data Frames in R

- We can do **stratified random sampling** in R using the command **`slice_sample`** function of dplyr.
- The weights in weight_by are normalized to add up to one.

```r
starwars %>%
  group_by(sex) %>%
  mutate(count = n()) %>%
  slice_sample(prop = 0.4, replace = FALSE, weight_by = count)
```

Without replacement

A tibble: 82 × 15 | Groups: sex [3]

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr>
Luke Skywalker	172	77.0	blond	fair	blue	19.0	male
C-3PO	167	75.0	*NA*	gold	yellow	112.0	none
R2-D2	96	32.0	*NA*	white, blue	red	33.0	none
Darth Vader	202	136.0	none	white	yellow	41.9	male
Leia Organa	150	49.0	brown	light	brown	19.0	female
Owen Lars	178	120.0	brown, grey	light	blue	52.0	male
Beru Whitesun lars	165	75.0	brown	light	blue	47.0	female
R5-D4	97	32.0	*NA*	white, red	red	*NA*	none
Biggs Darklighter	183	84.0	black	light	brown	24.0	male
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	57.0	male

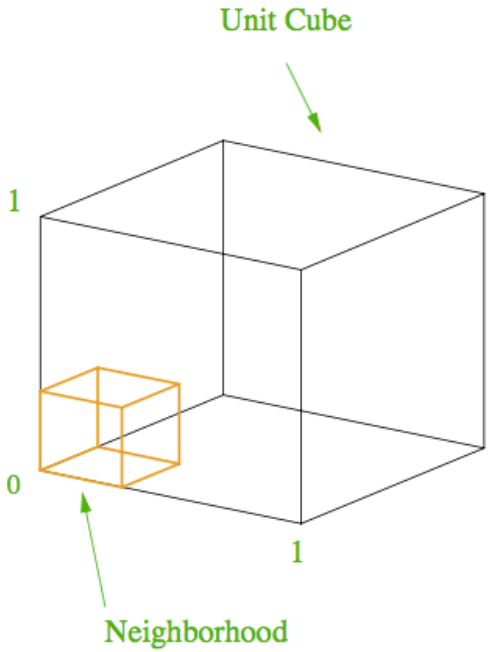
1-10 of 82 rows | 1-9 of 15 columns

Previous 1 2 3 4 5

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

In the figure on the right *p* is the dimension of the space where the cube lives.
The plot says how large each side of the yellow cube needs to be in order to obtain a fraction x of the volume of the unit cube



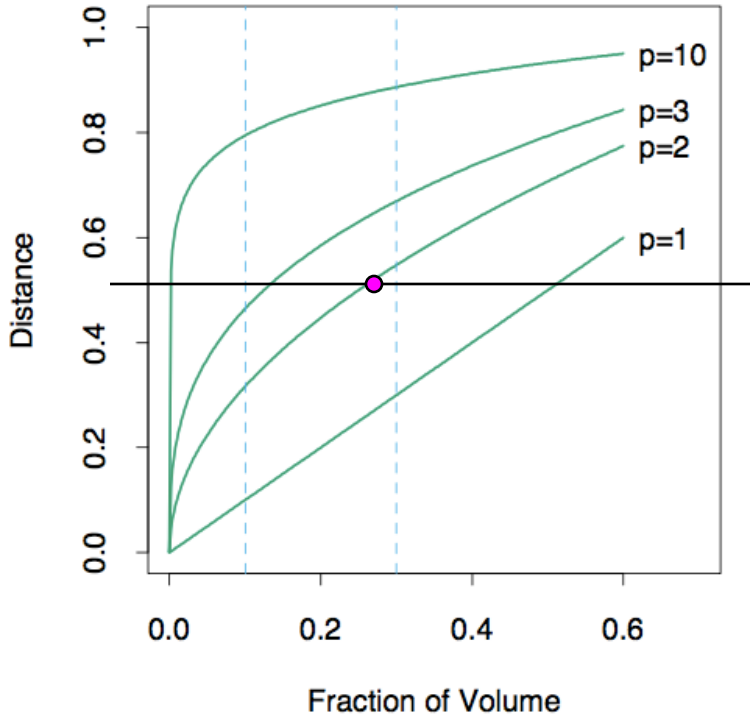


Figure from Hastie & Tibshirani, Elements of Statistical Learning, 2nd Edition.

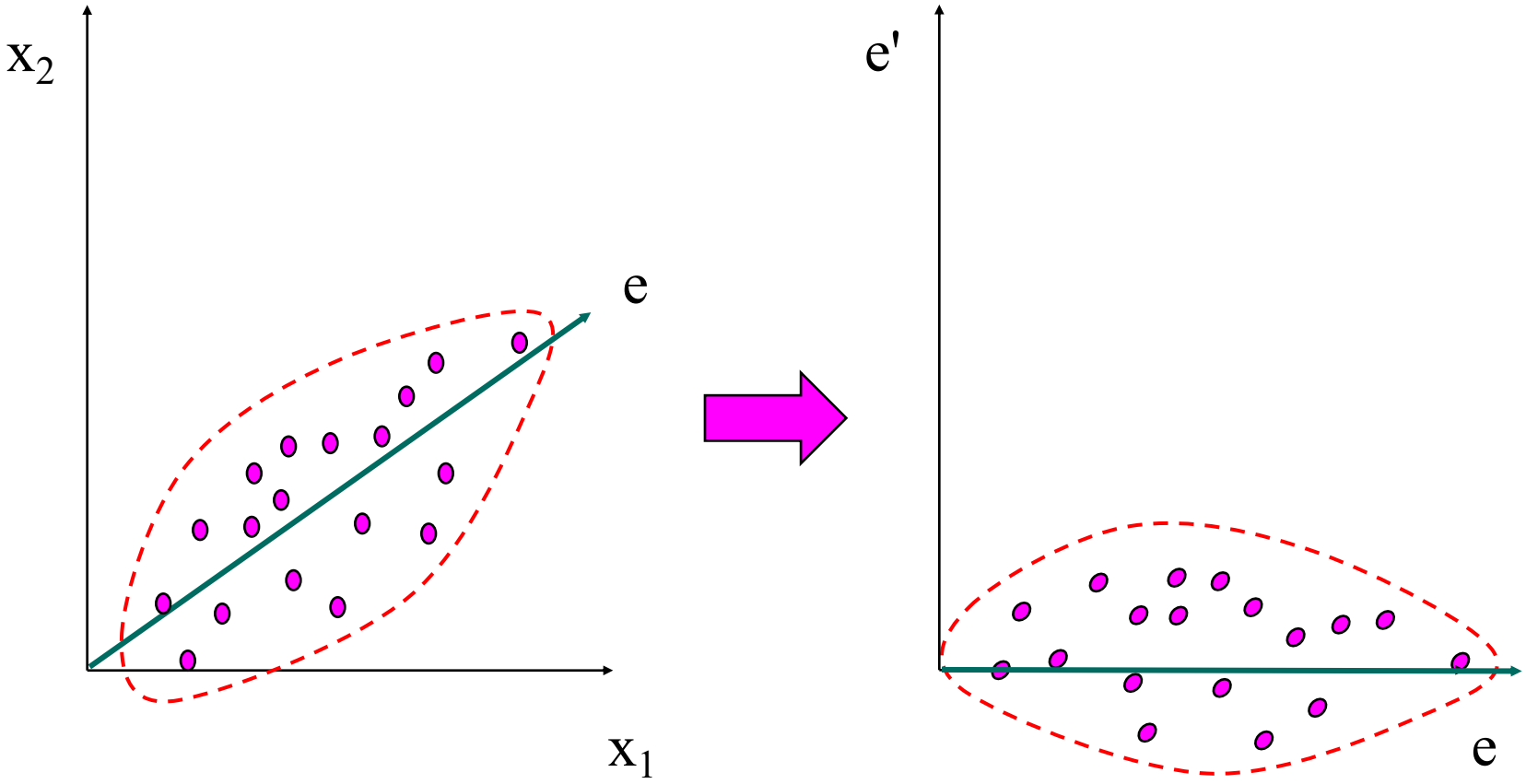
Dimensionality Reduction

- 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

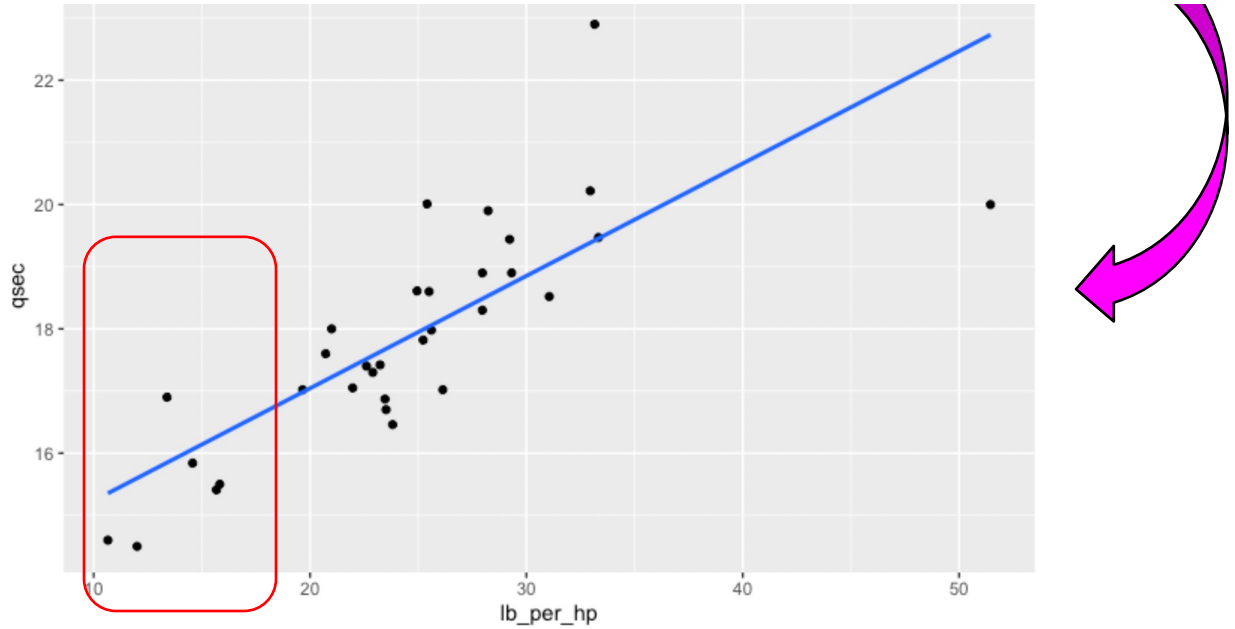
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- 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

Feature Creation Example

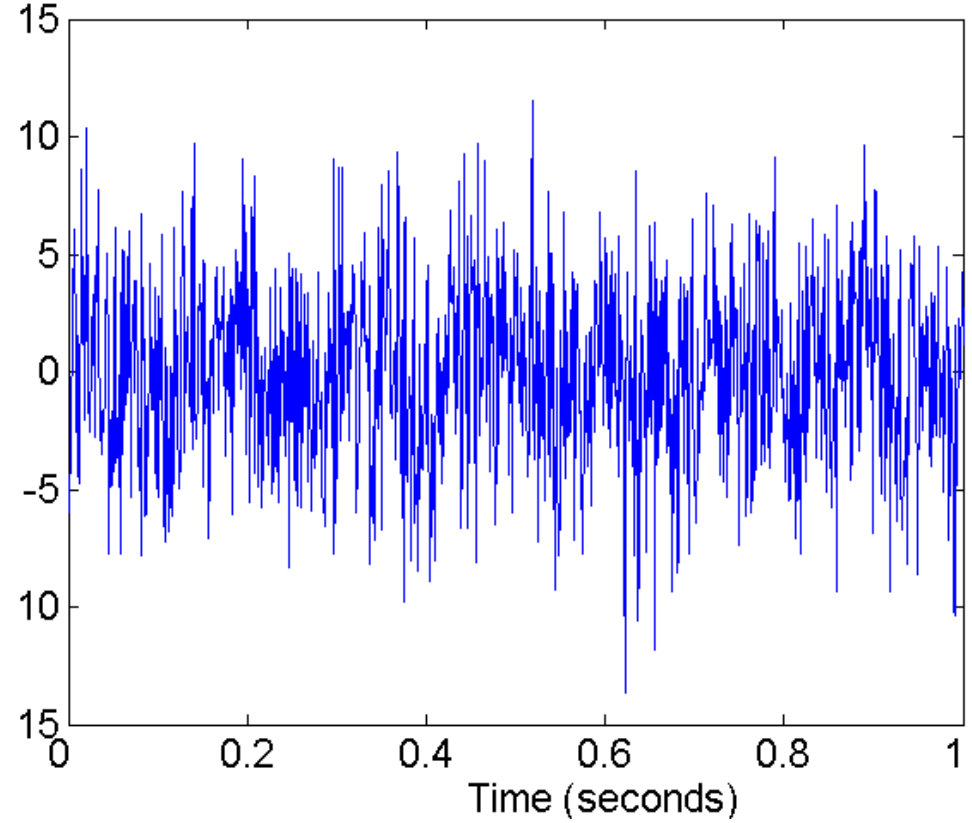
	mpg <dbl>	cyl <fctr>	disp <dbl>	hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02
Datsun 710	22.8	4	108	93	3.85	2.320	18.61
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02

```r
data %>% mutate(lb_per_hp = wt * 1000 / hp) %>%
         relocate(lb_per_hp, .after = hp) %>%
         ggplot(aes(x = lb_per_hp, y= qsec)) + geom_point() + geom_smooth(method = 'lm', se = FALSE)
```

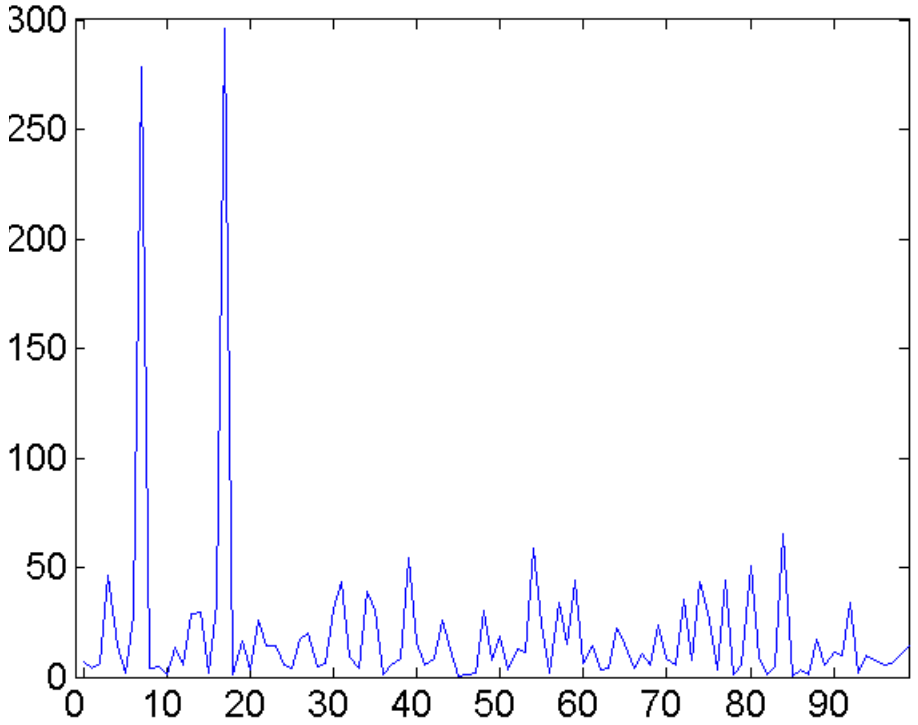


Mapping Data to a New Space

- Fourier and wavelet transform



Two Sine Waves + Noise



Frequency

Discretization

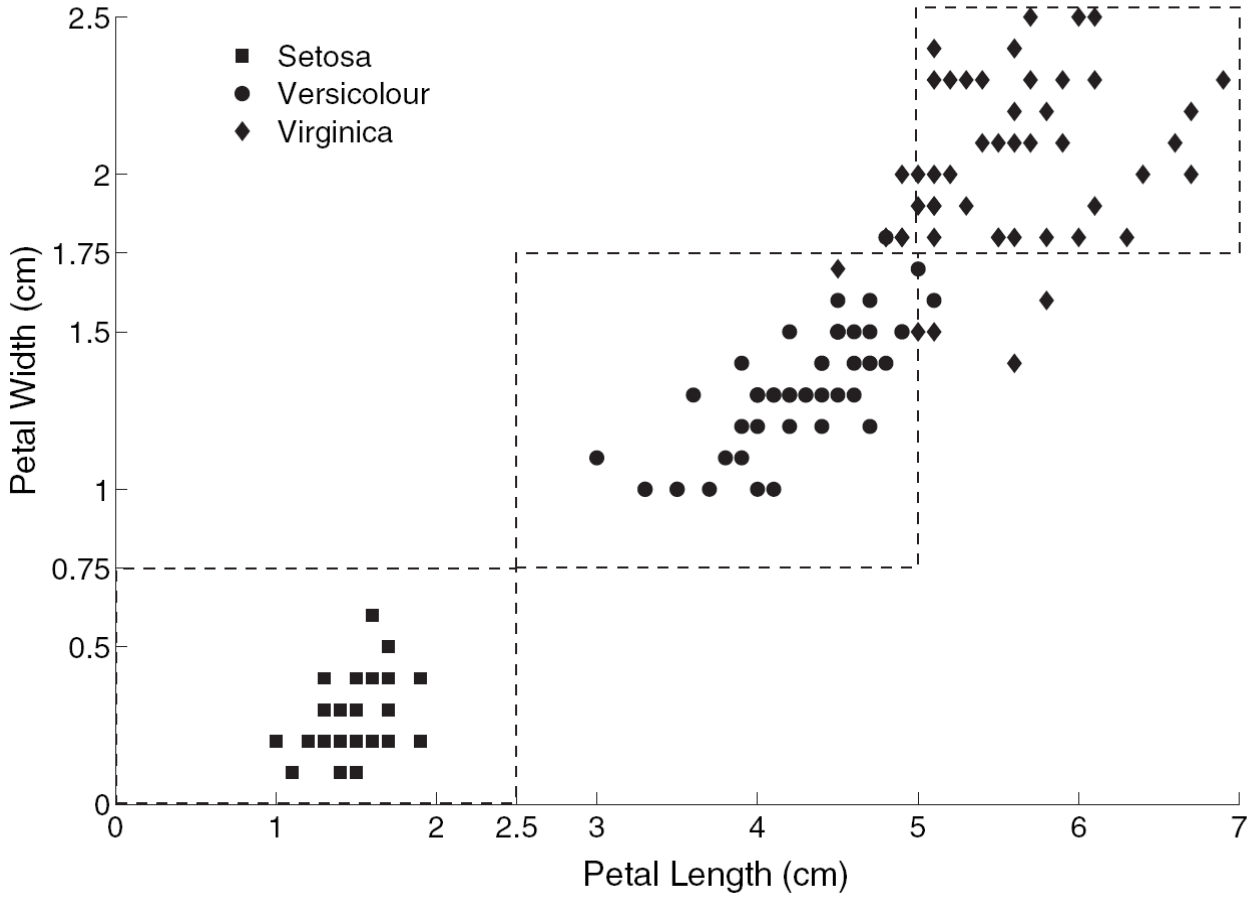
- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent 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.

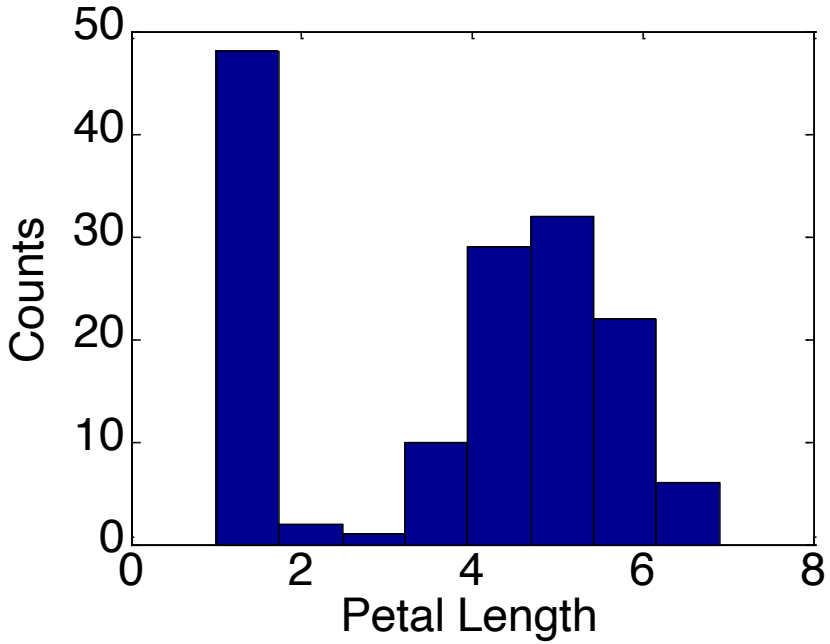
Discretization: Iris Example



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.

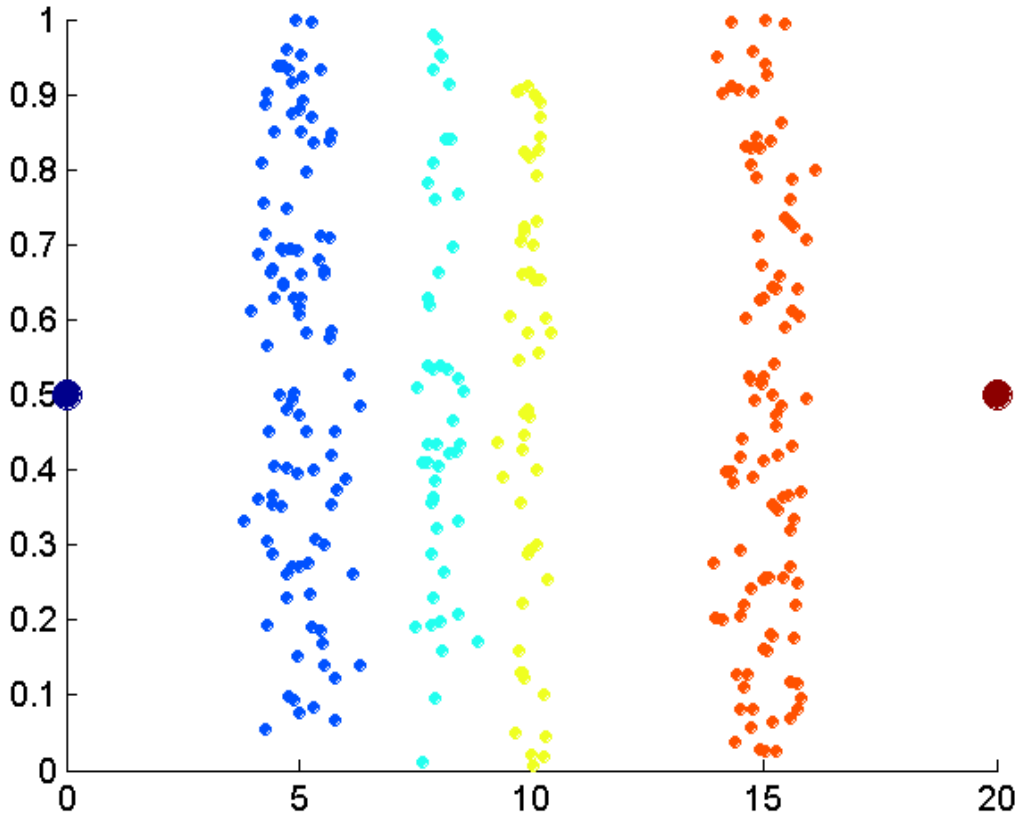
Discretization: Iris Example …

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



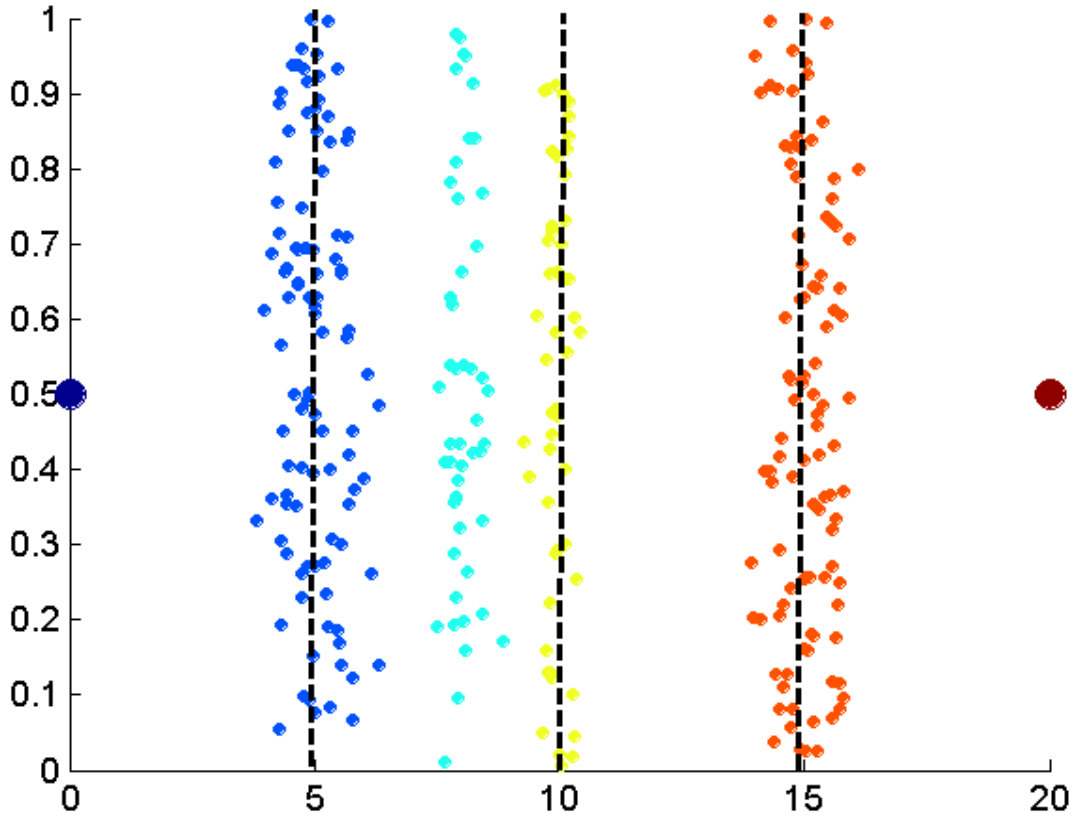
  - Supervised discretization: Use class labels to find breaks

Discretization Without Using Class Labels



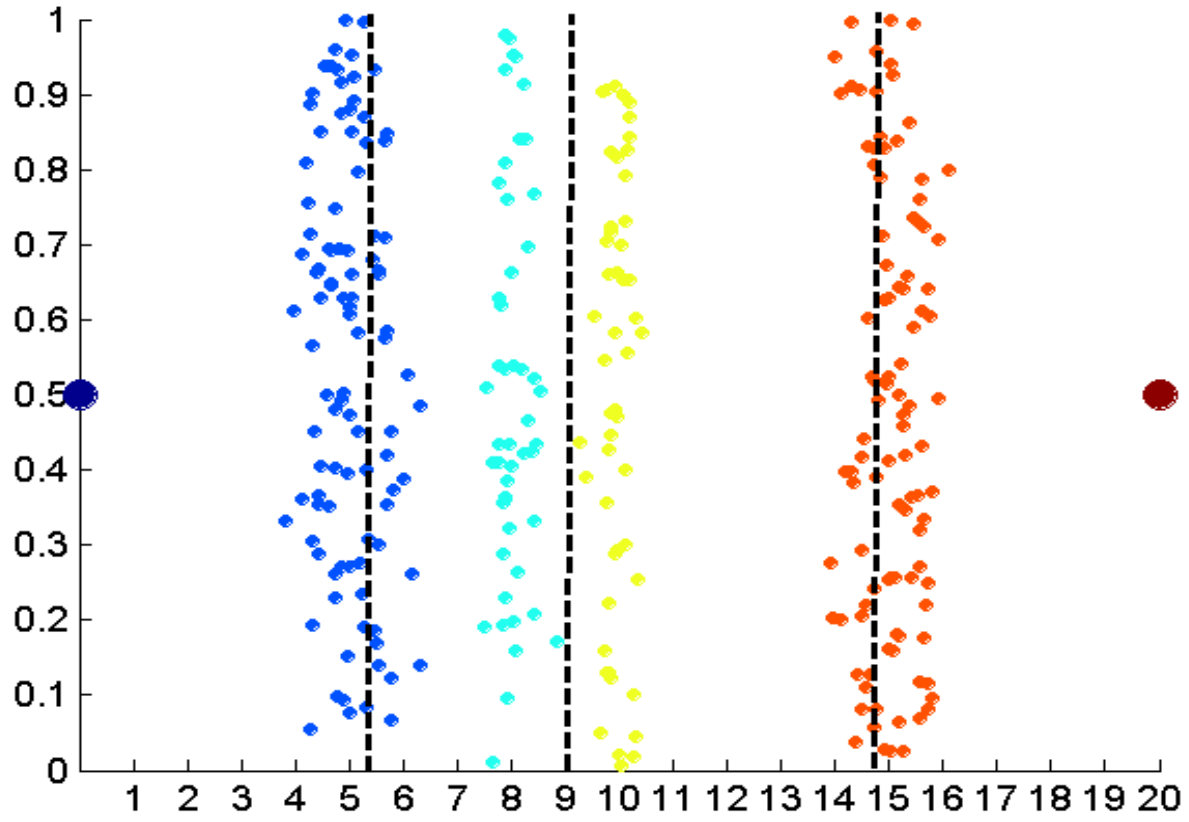
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Discretization Without Using Class Labels



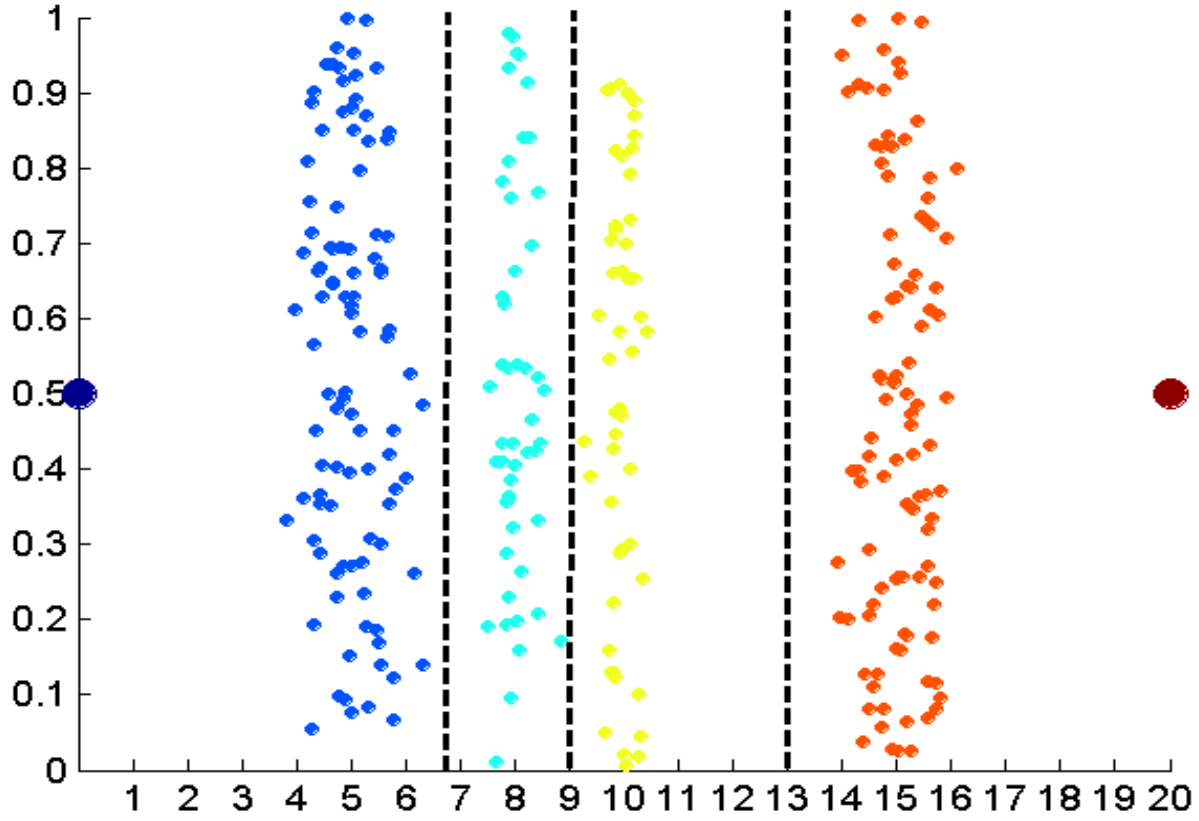
Equal interval width approach used to obtain 4 values

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 (we'll talk more about this later in the course)

Discretization in R

- There are many ways to discretize a continuous attribute in R
- The following is an example that uses the `cut` command from the `arules` package

```r
mtcars$hp
```

```
  [1] 110 110  93 110 175 105 245  62  95 123 123 180 180 180 205 215 230  66  52  65  97 150 150 245 175  66  91 113 264 175 335 109
```

```r
cut(mtcars$hp, breaks = c(0, 100, 150, 200, 250, 600), labels = c('Very Low Hp', 'Low Hp', 'Med Hp', 'High Hp', 'Super High Hp'))
```

```
 [1] Low Hp        Low Hp        Very Low Hp   Low Hp        Med Hp        Low Hp        High Hp       Very Low Hp   Very Low Hp
[10] Low Hp        Low Hp        Med Hp        Med Hp        Med Hp        High Hp       High Hp       High Hp       Very Low Hp
[19] Very Low Hp   Very Low Hp   Very Low Hp   Low Hp        Low Hp        High Hp       Med Hp        Very Low Hp   Very Low Hp
[28] Low Hp        Super High Hp Med Hp        Super High Hp Low Hp
Levels: Very Low Hp Low Hp Med Hp High Hp Super High Hp
```

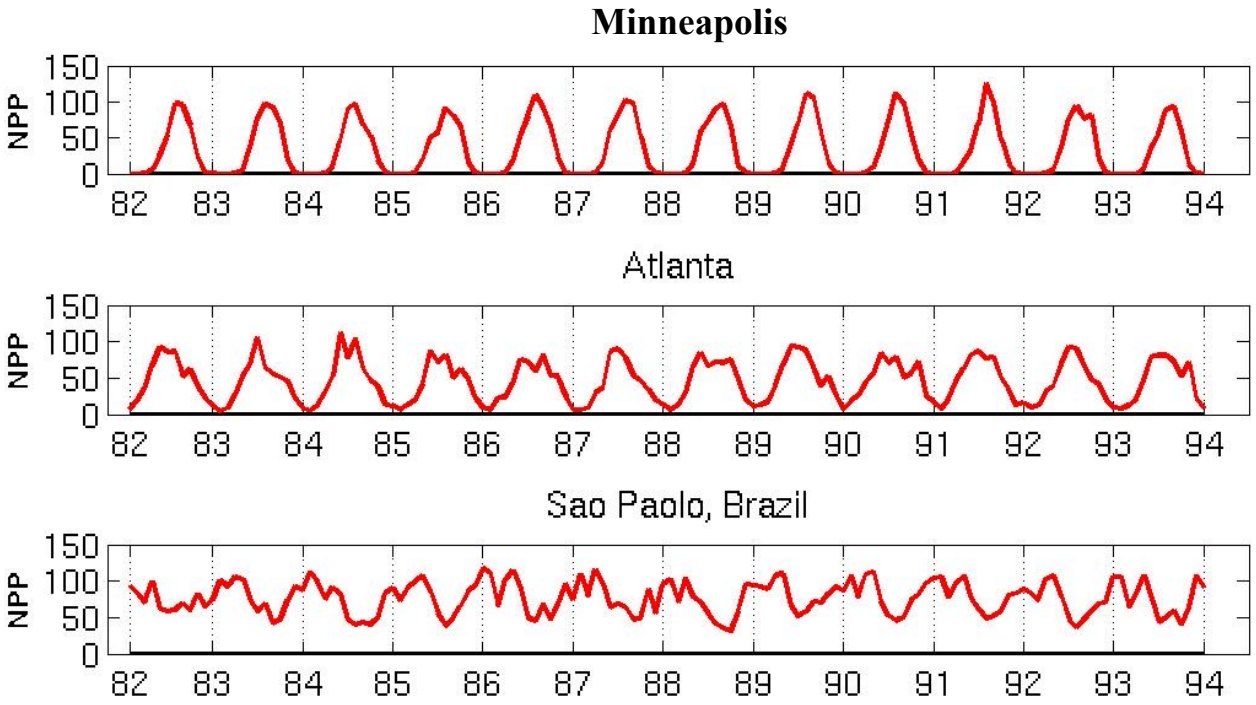
Binarization

- 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

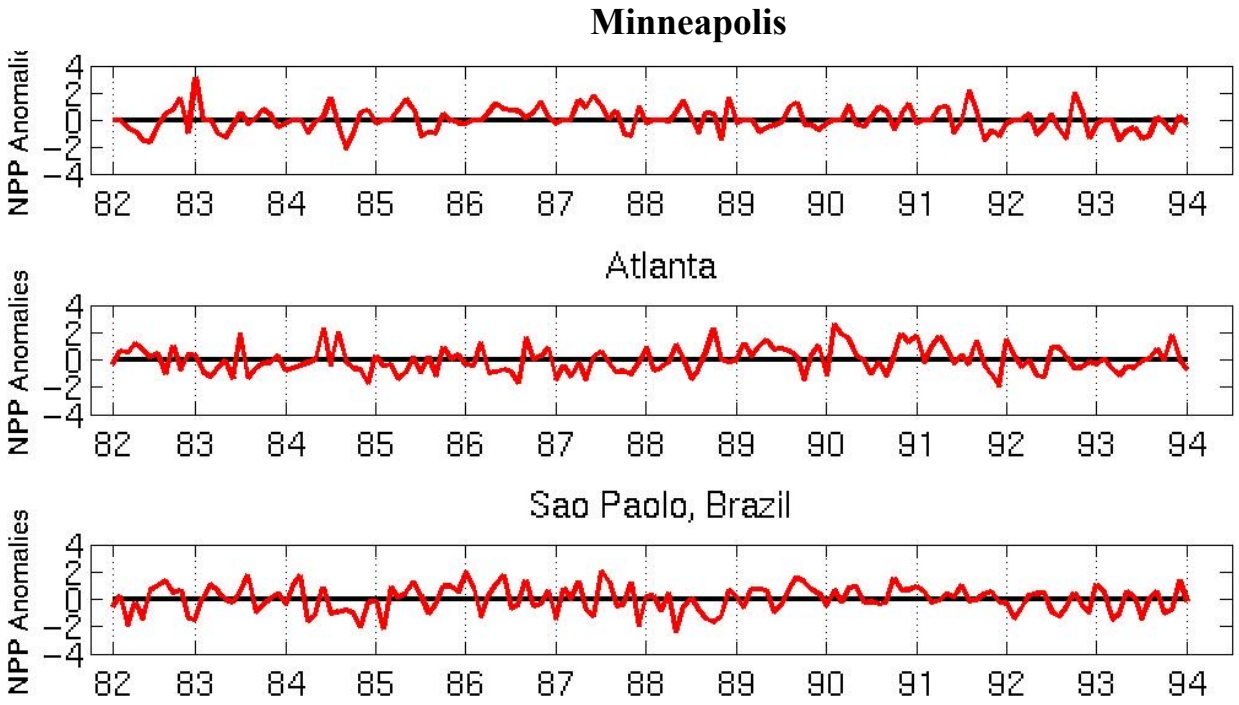


Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

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