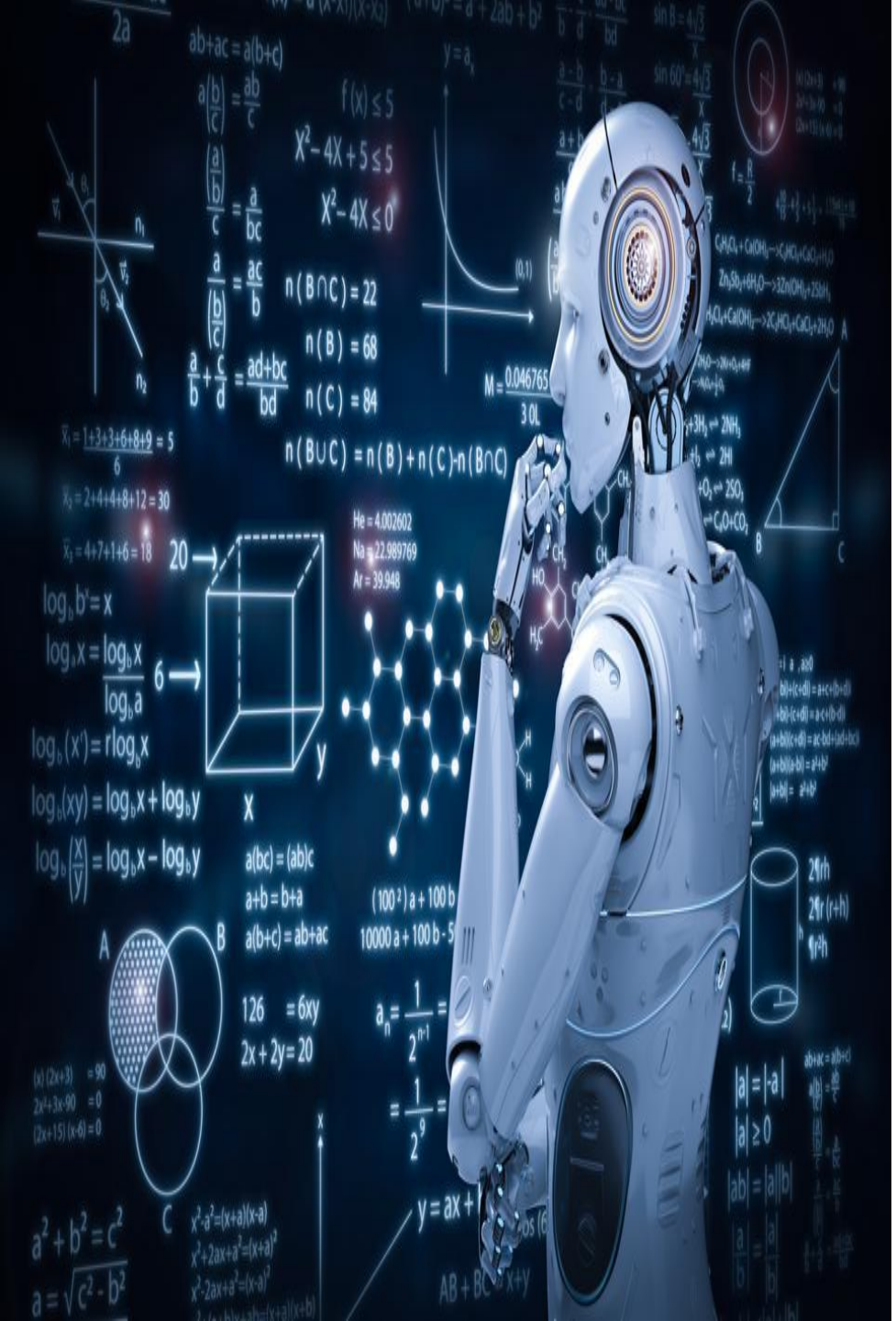


Practical Machine Learning

Day 3: Mar23 DBDA

Kiran Waghmare

Agenda

- Data
- Types of Attributes
- Preprocessing
- Transformations
- Measures
- Visualization

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, 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

Types of Data

- Categorical features come from an unordered set:
  - Binary: job?
  - Nominal: city.

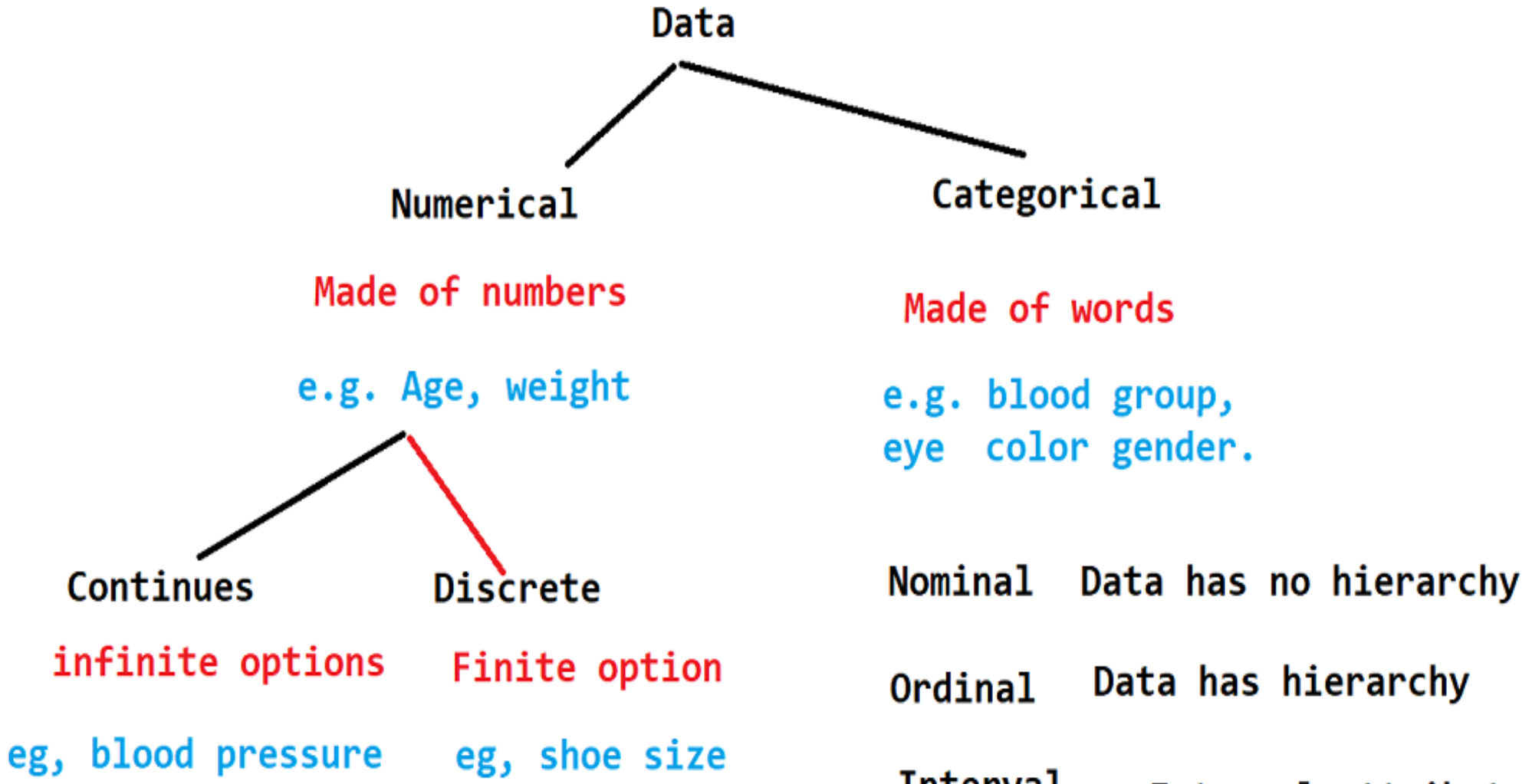
- Numerical features come from ordered sets:
  - Discrete counts: age.
  - Ordinal: rating.
  - Continuous/real-valued: height.

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 in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Types of data sets

- Record
  - Data matrix
  - Document data
  - Transaction data
- Graph
  - World Wide Web
  - Molecular structures
- Ordered
  - Spatial data
  - Temporal (time series) data
  - Sequential data
  - Genetic sequence data


Data
Numerical
Categorical
Made of numbers
e.g. Age, weight
Made of words
e.g. blood group,
eye color gender.
Continues
Discrete
infinite options
Finite option
eg, blood pressure
eg, shoe size
Nominal Data has no hierarchy
Ordinal Data has hierarchy
Interval Interval attribute
Ratio ratio variables

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* x *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

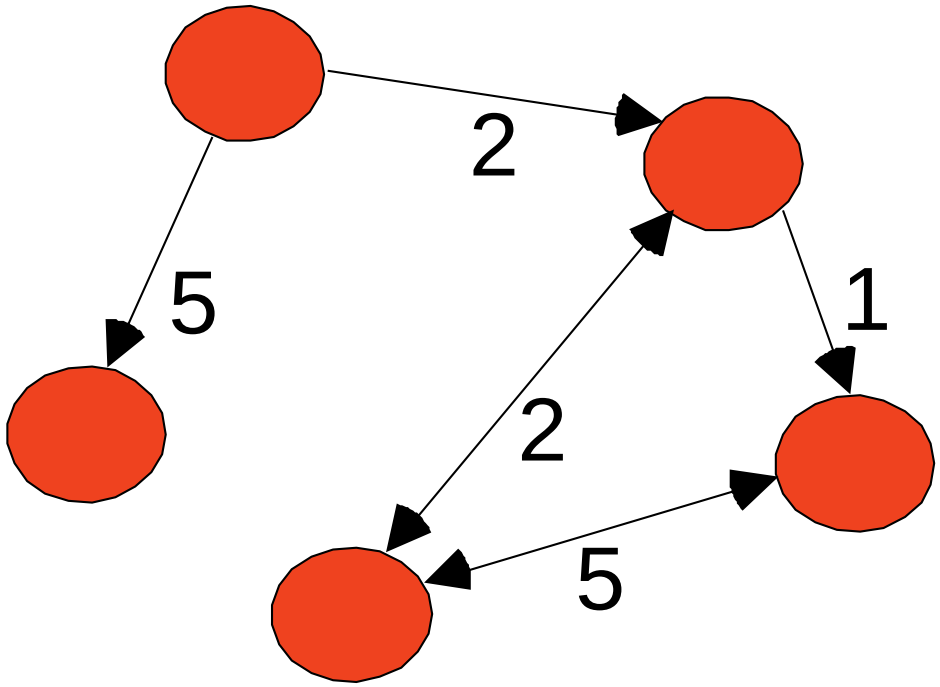
Transaction data

- 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

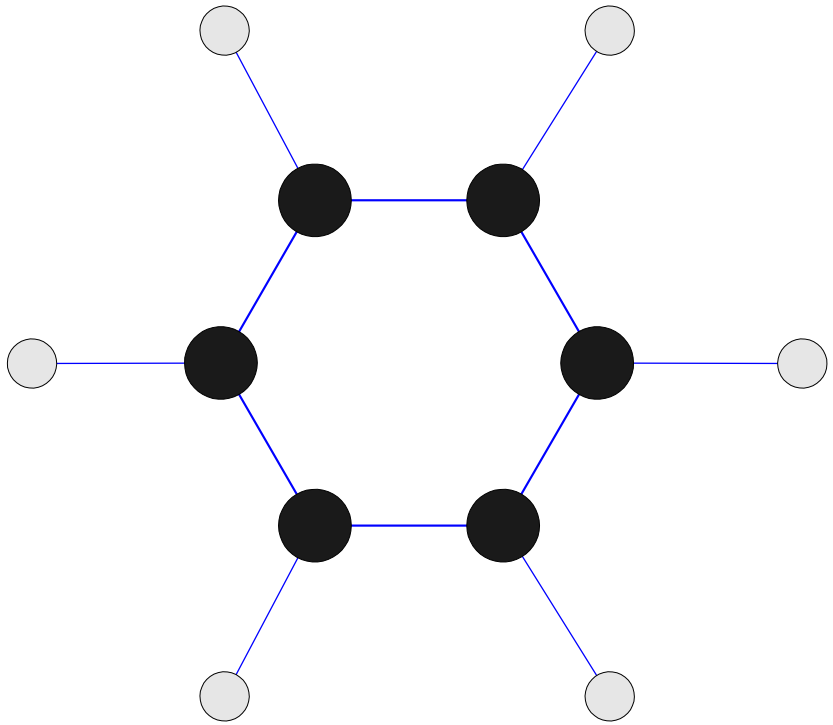
- Examples: Generic graph and HTML Links




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

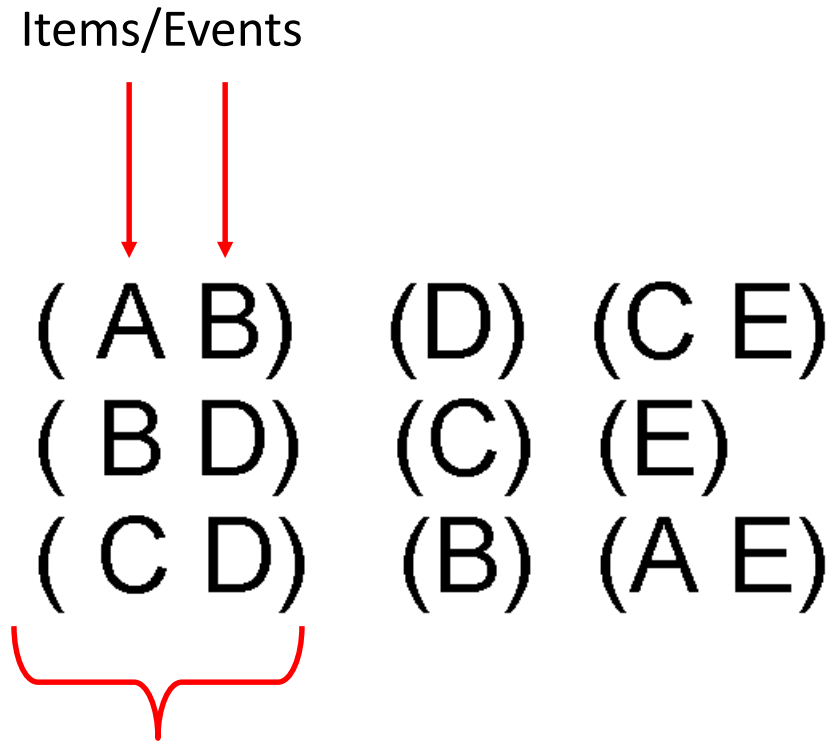
Chemical data

- Benzene molecule: C_6H_6



Ordered data

- Sequences of transactions



Ordered data

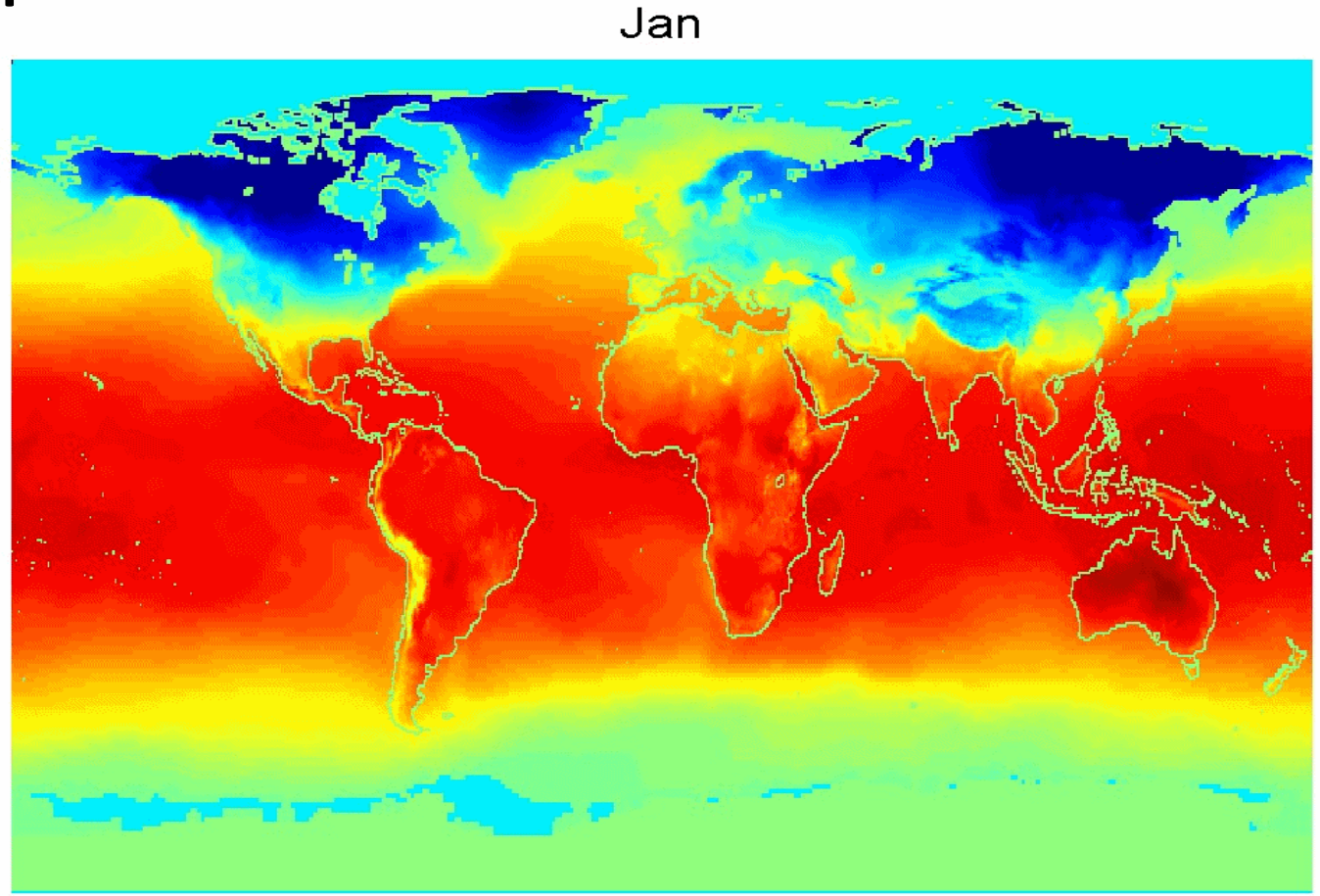
- Genomic sequence data

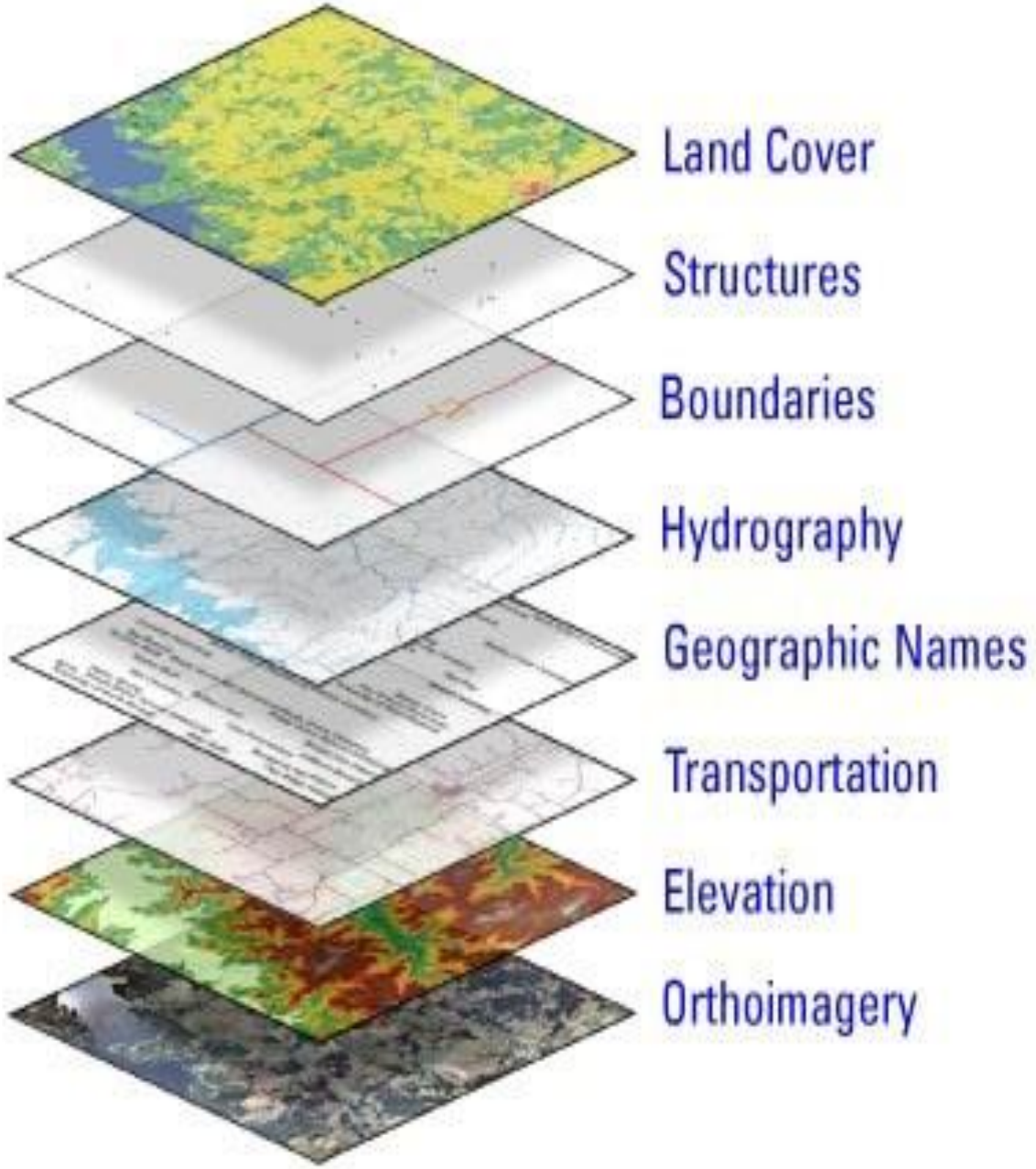
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

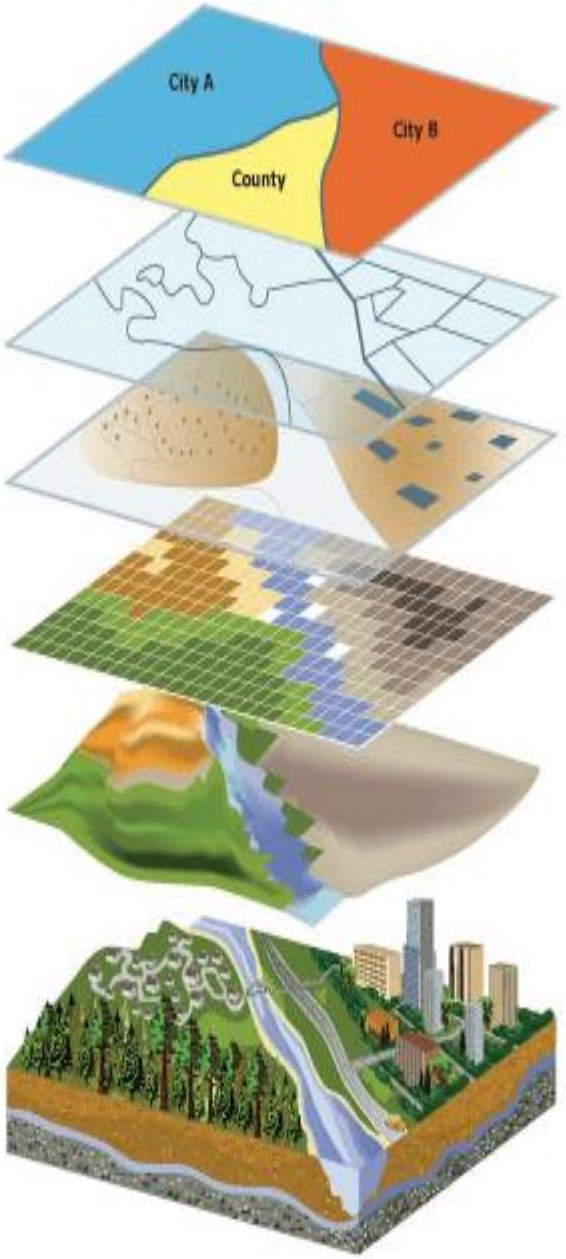
Ordered data

- Spatio-temporal data

Average monthly temperature of land and ocean




Land Cover
Structures
Boundaries
Hydrography
Geographic Names
Transportation
Elevation
Orthoimagery


City A
City B
County

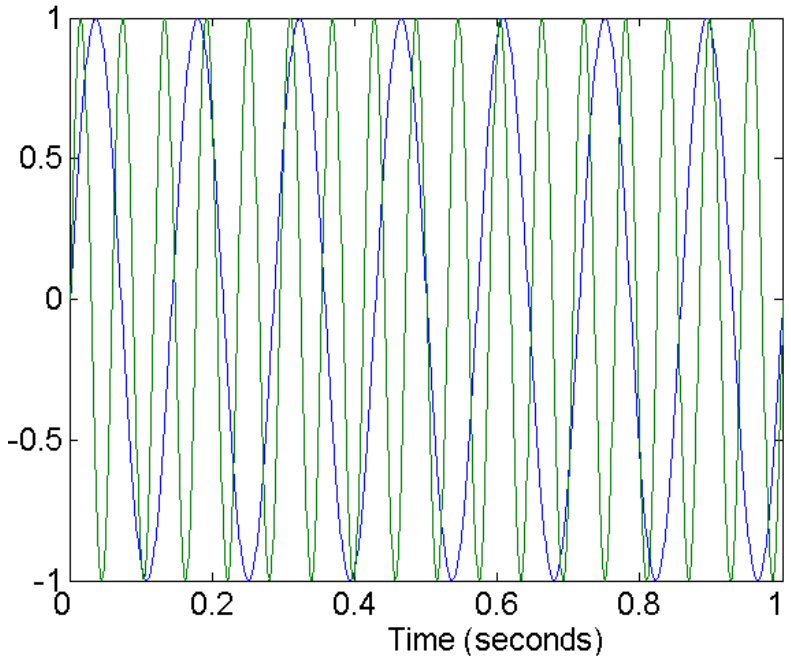
Data quality

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

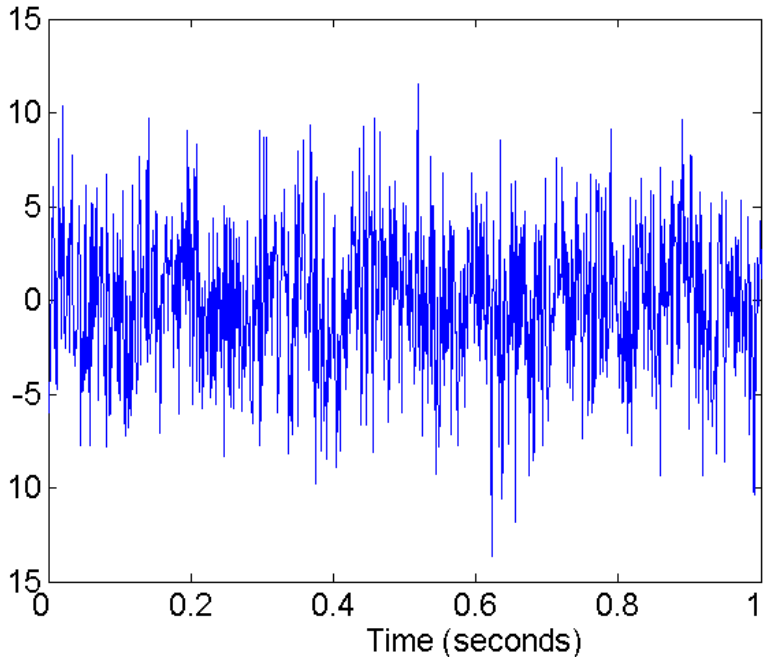
- Examples of data quality problems:
  - noise and outliers
  - missing values
  - duplicate data

Noise

- Noise refers to random modification of original values
- Examples:
  - distortion of a person's voice when talking on a poor phone
  - "snow" on television screen



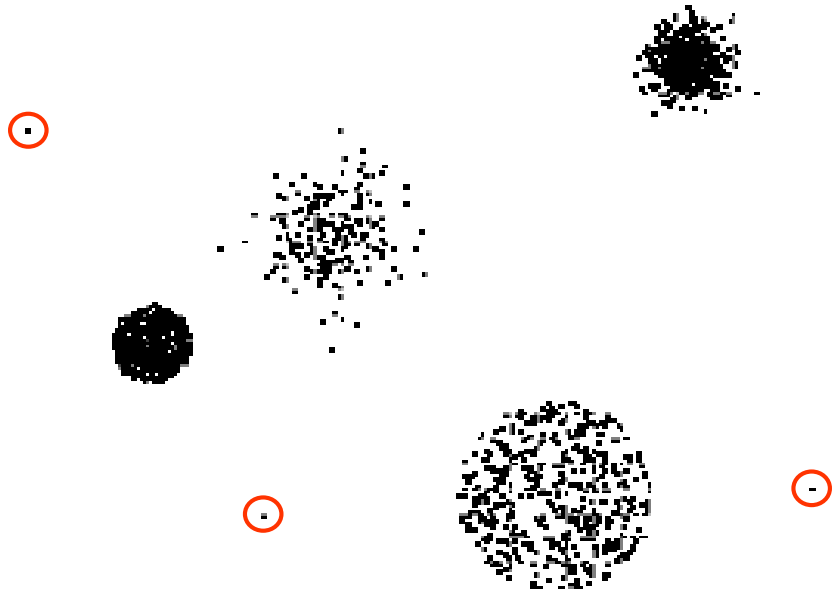
Two sine waves



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



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 (imputation)
  - Ignore the missing value during analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeous sources
- Example:
  - Same person with multiple email addresses
- Data cleaning
  - Includes process of dealing with duplicate data issues

The Question I Hate the Most…

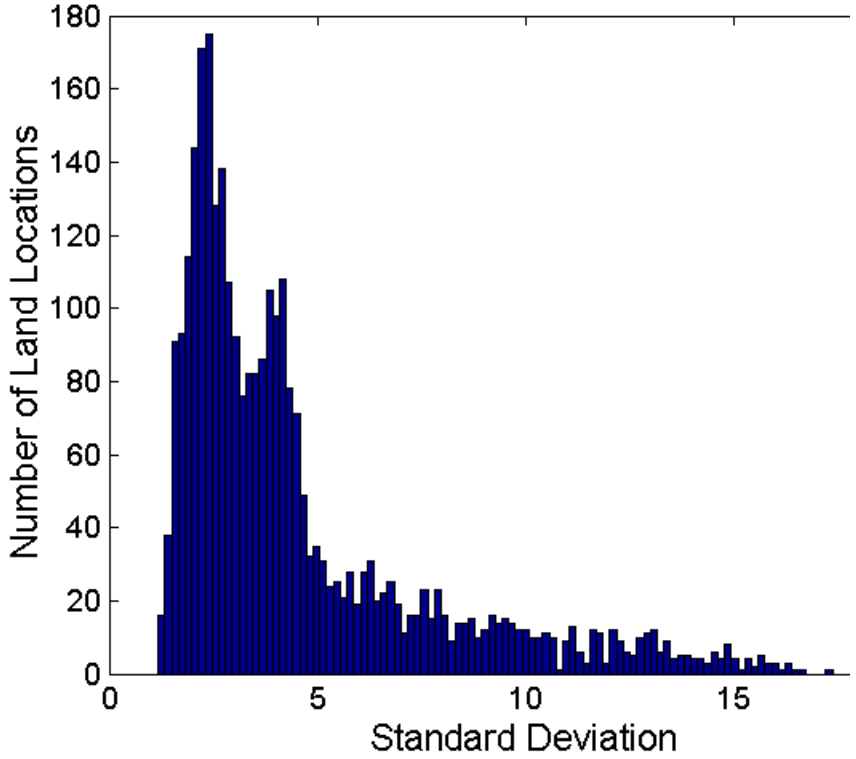
- How much data do we need?
- A difficult if not impossible question to answer.
- My usual answer: “more is better”.
  - With the warning: “as long as the quality doesn’t suffer”.
- Another popular answer: “ten times the number of features”.

Data preprocessing

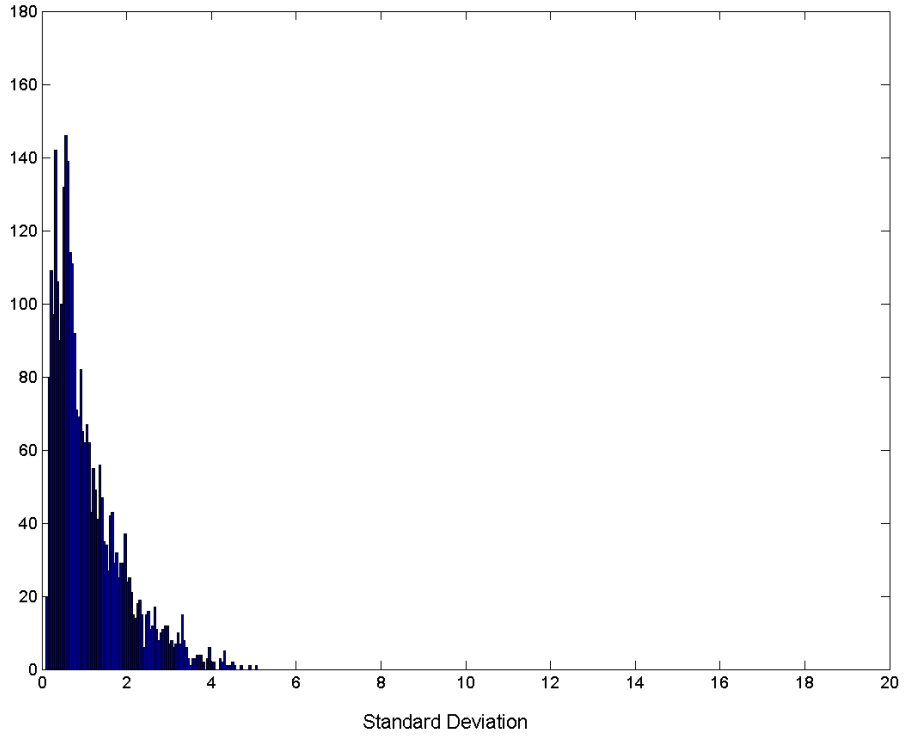
- Aggregation
- Sampling
- Discretization and binarization
- Attribute transformation
- Feature creation
- Feature selection
  - Choose subset of existing features

Aggregation

Variation of precipitation in Australia



Standard deviation of average monthly precipitation



Standard deviation of average yearly precipitation

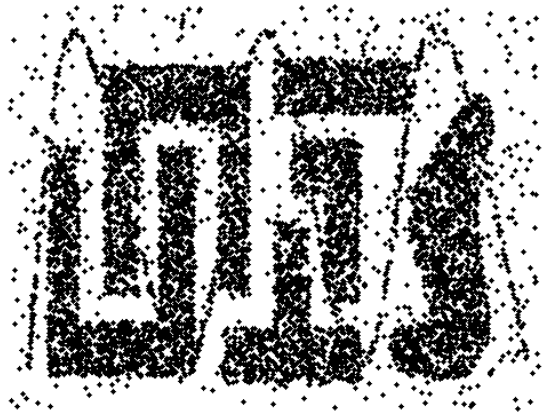
Sampling

- Sampling is the main technique employed for data selection.
  - Often used for both preliminary investigation of the data and the final data analysis.

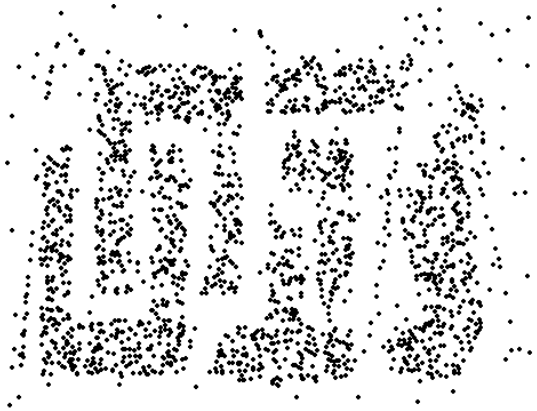
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

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

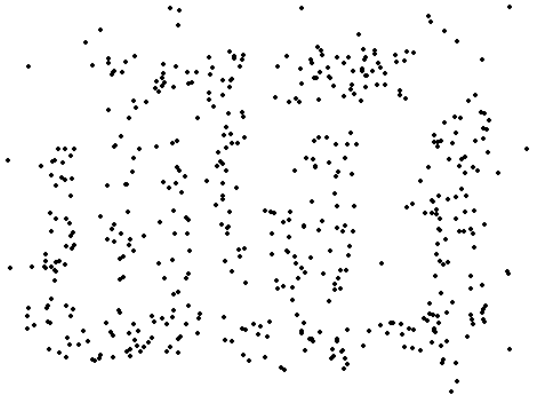
Sample size



8000 points



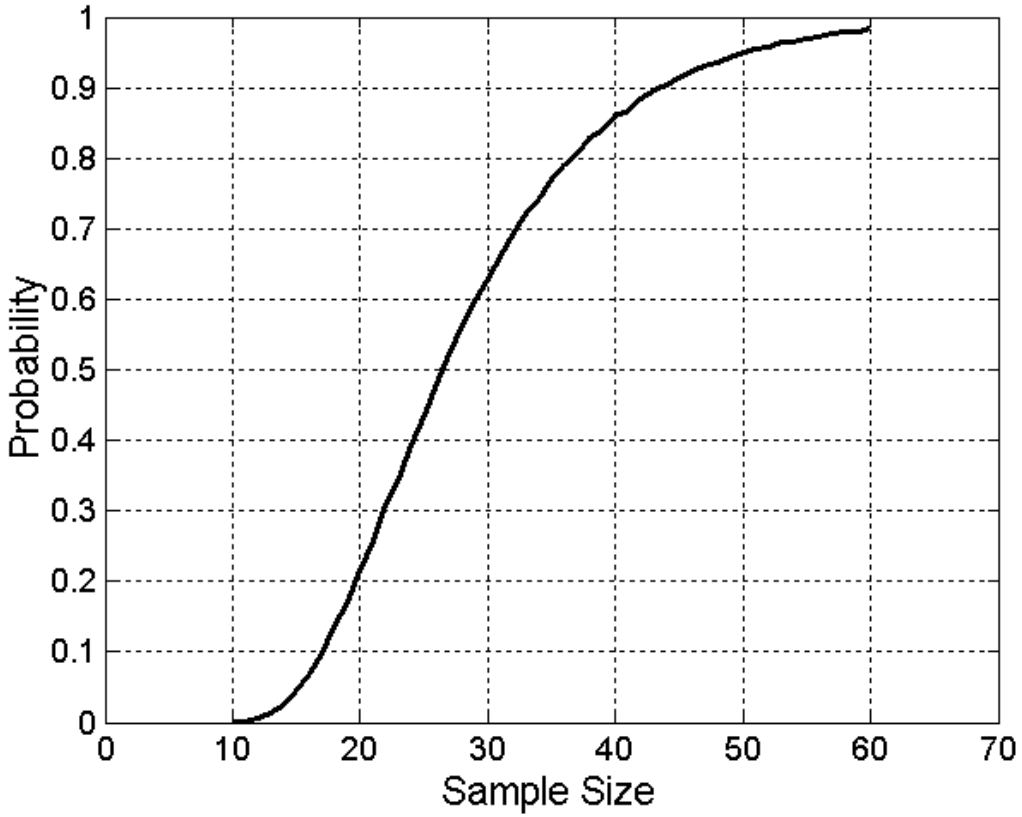
2000 Points



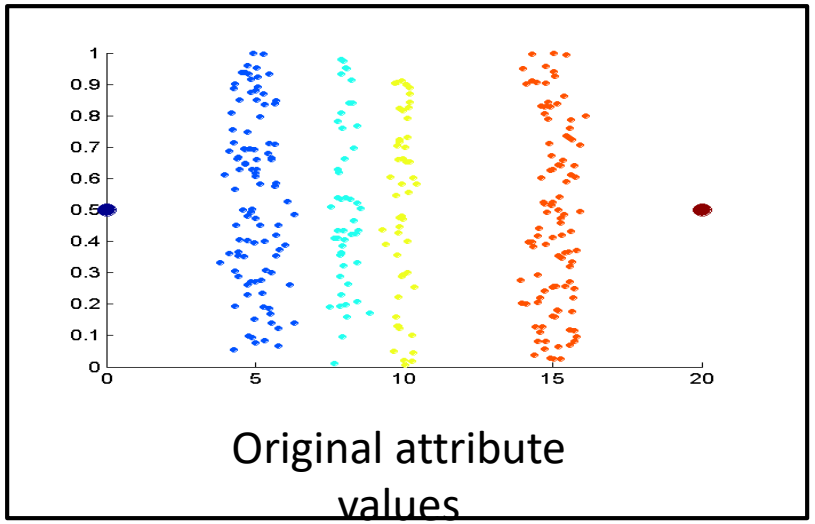
500 Points

Sample size

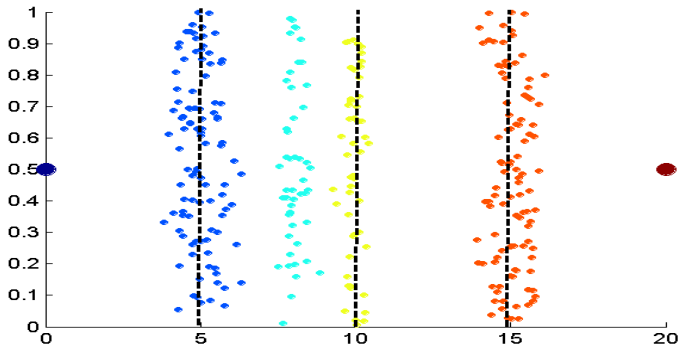
- What sample size is necessary to get at least one object from each of 10 equal-sized groups?



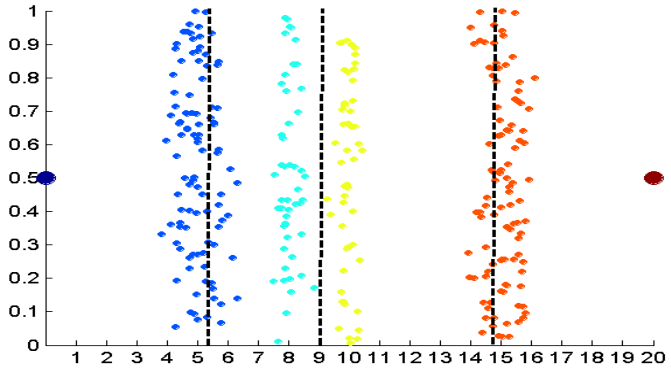
Approaches to discretization



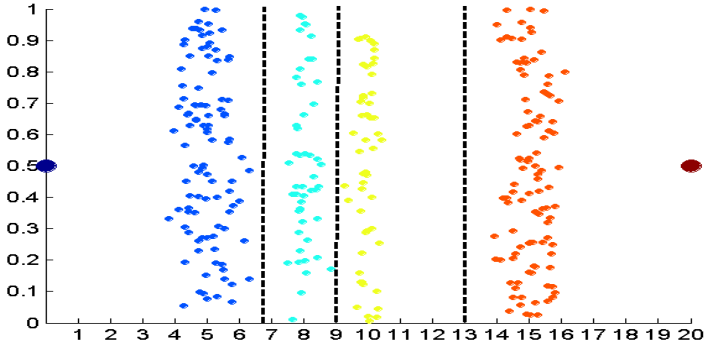
Original attribute values



Equal interval width



Equal frequency



k-means

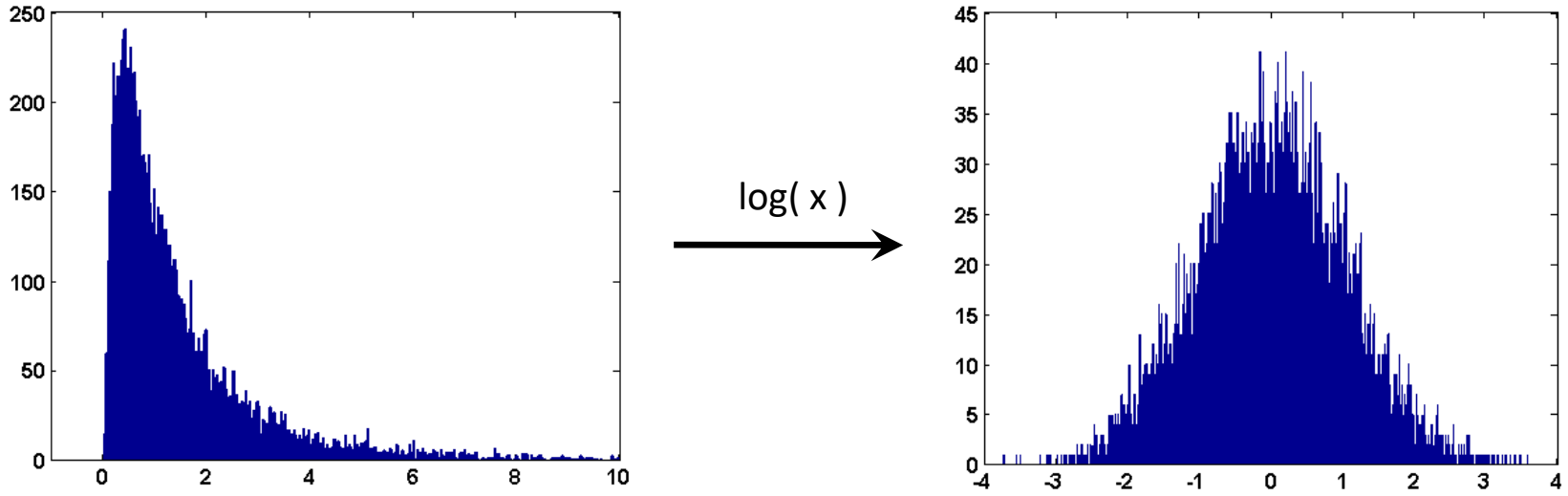
Attribute transformation

Definition:

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.

Attribute transformation

- Simple functions
  - Examples of transform functions:
    x^k $\log(x)$ e^x $|x|$
  - Often used to make the data more like some standard distribution, to better satisfy assumptions of a particular algorithm.
    - Example: discriminant analysis explicitly models each class distribution as a multivariate Gaussian



Attribute transformation

- **Standardization or normalization**
  - Usually involves making attribute:

    mean = 0

    standard deviation = 1

  - Important when working in **Euclidean space and attributes** have very different numeric scales.

  - Also necessary to satisfy assumptions of certain algorithms.
    - Example: **principal component analysis (PCA)** requires each attribute to be mean-centered (i.e. have mean subtracted from each value)

Approximating Text with Numerical Features

- Bag of words replaces document by word counts:

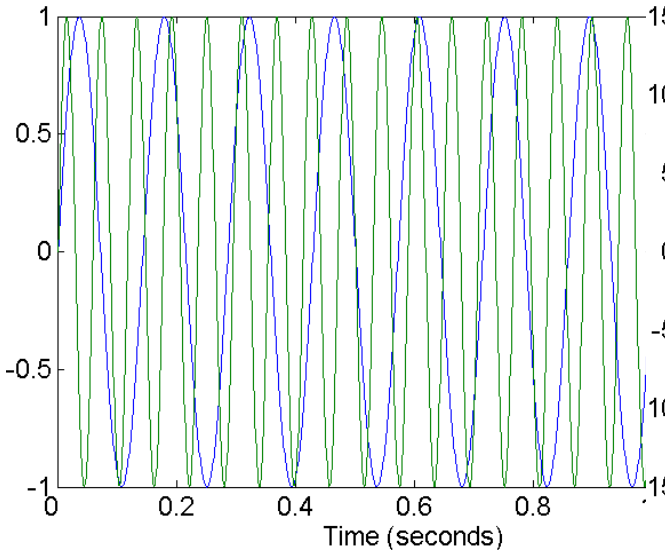
The **International Conference on Machine Learning** (ICML) is the leading international academic conference in machine learning

ICML	International	Conference	Machine	Learning	Leading	Academic
1	2	2	2	2	1	1

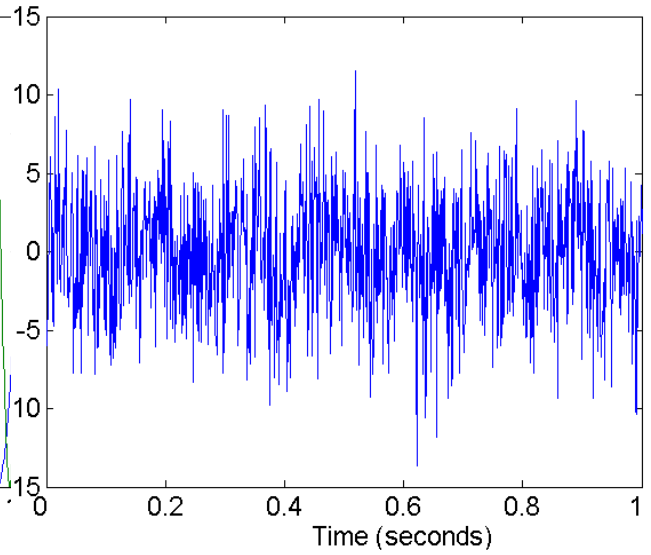
- Ignores order, but often captures general theme.
- You can compute a "distance" between documents.

Transform data to a new space

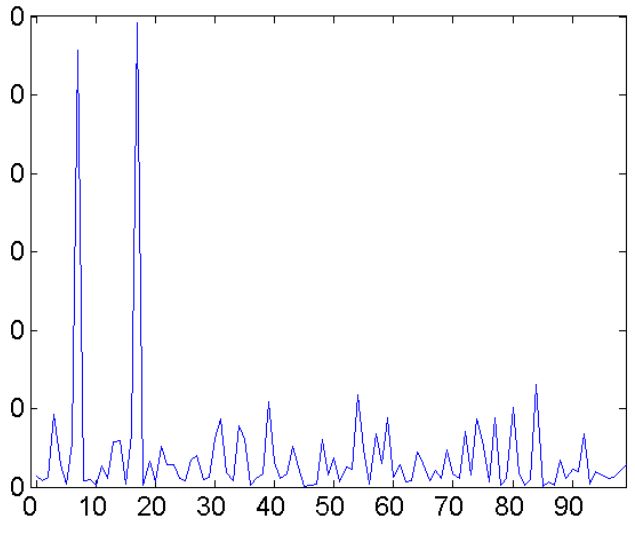
- Fourier transform
  - Eliminates noise present in time domain



Two sine waves



Two sine waves + noise



Frequency

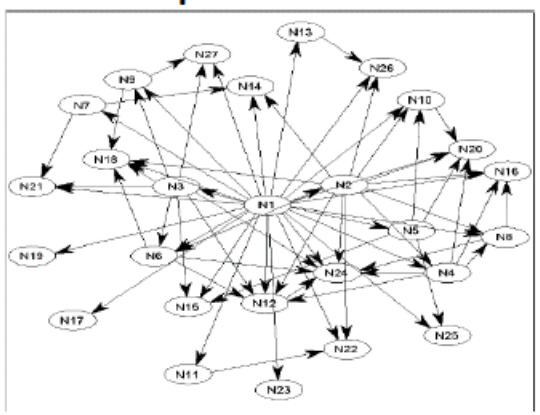
Approximating Images and Graphs

- We can think of other data types in this way:
  - Images:

graycale intensity

(1,1)	(2,1)	(3,1)	…	(m,1)	…	(m,n)
45	44	43	…	12	…	35

  - Graphs:



adjacency matrix

N1	N2	N3	N4	N5	N6	N7
0	1	1	1	1	1	1
0	0	0	1	0	1	0
0	0	0	0	0	1	0
0	0	0	0	0	0	0

Converting to Numerical Features

- Often want a real-valued example representation:

Age	City	Income
23	Van	22,000.00
23	Bur	21,000.00
22	Van	0.00
25	Sur	57,000.00
19	Bur	13,500.00
22	Van	20,000.00

Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00

- This is called a "1 of k" encoding.
- We can now interpret examples as points in space:
  - E.g., first example is at (23,1,0,0,22000).

Feature Aggregation

- Feature aggregation:
  - Combine features to form new features:

Van	Bur	Sur	Edm	Cal
1	0	0	0	0
0	1	0	0	0
1	0	0	0	0
0	0	0	1	0
0	0	0	0	1
0	0	1	0	0

BC	AB
1	0
1	0
1	0
0	1
0	1
1	0

- Fewer province “coupons” to collect than city “coupons”.

Feature Selection

- Feature Selection:
  - Remove features that are not relevant to the task.

SID:	Age	Job?	City	Rating	Income
3457	23	Yes	Van	A	22,000.00
1247	23	Yes	Bur	BBB	21,000.00
6421	22	No	Van	CC	0.00
1235	25	Yes	Sur	AAA	57,000.00
8976	19	No	Bur	BB	13,500.00
2345	22	Yes	Van	A	20,000.00

  - Student ID is probably not relevant.

Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.

Age
23
23
22
25
19
22

→

< 20	>= 20, < 25	>= 25
0	1	0
0	1	0
0	1	0
0	0	1
1	0	0
0	1	0

- Only need consider 3 values.

Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.
  - Square, exponentiation, logarithm, and so on.

Interview Questions