Data Mining: Data

Tom Heskes

Class Project

- Hands-on experience with data mining
- Two options:
  - “Problem”
  - “Algorithm”
- Two phases:
  - Proposal: November 15 (soft deadline)
  - Final report: January 24 (hard deadline)

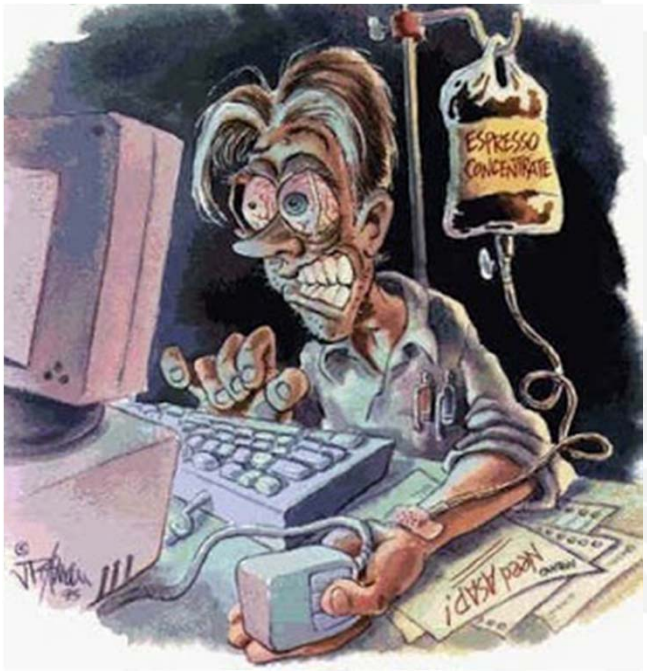
“Problem”

- Find a data set and corresponding problem
- Find a work bench of data mining / machine learning algorithms
- Apply at least two different algorithms to your problem
- Write a report

“Algorithm”

- Choose one or at most two challenging data mining algorithms that you'd like to implement yourself
- Implement them and apply them to a suitable data set
- Write a report

Practicalities



- Groups of 2 or by yourself
- Data: see Blackboard
- Software: see Blackboard
- If you’re stuck: ask!
- Do not underestimate!!!!!!

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

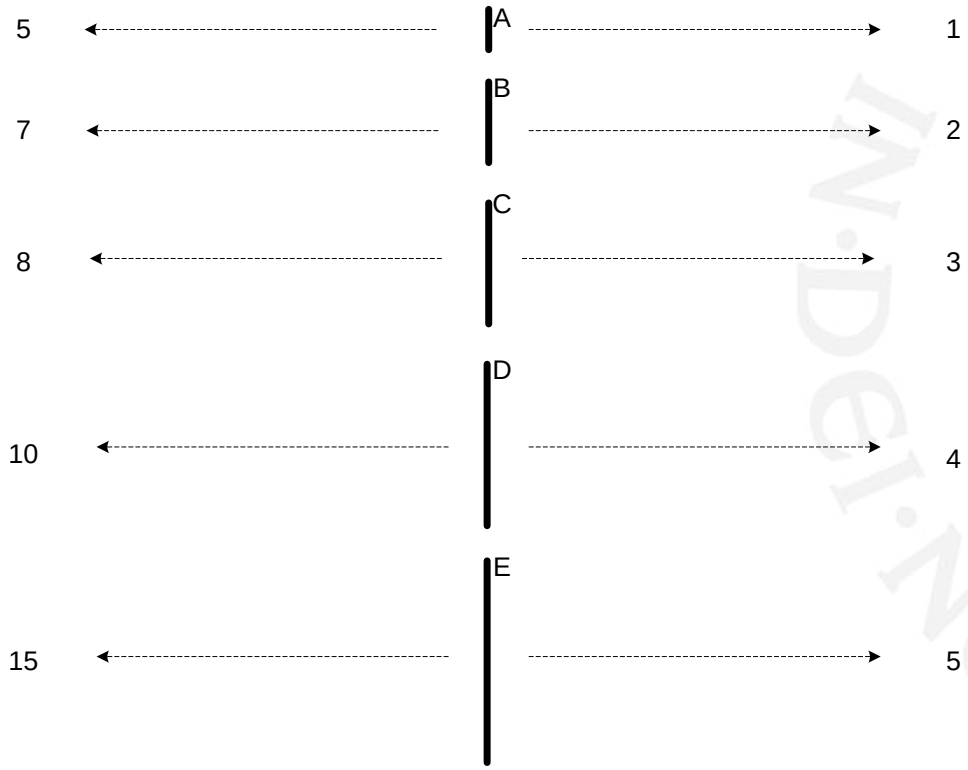
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- 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

Measurement of Length

- The way you measure an attribute is somewhat arbitrary.
- Rely on common sense.



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

Properties of Attribute Values

- Mathematical properties / operations:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /
- The type of an attribute depends on which of these apply:
  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

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.

Examples

For each of the following attributes, say whether it's binary, discrete, or continuous and whether it's nominal, ordinal, interval, or ratio.

- Age in years
- Time in terms of AM or PM
- Brightness as measured by a light meter
- Brightness as measured by people's judgments
- Bronze, silver, and gold medals as awarded at the Olympics
- Height above sea level
- Number of patients in a hospital
- ISBN numbers for books
- Military rank
- Distance from the center of campus
- Temperature in degrees Kelvin
- Temperature in degrees Celsius
- Coat check number

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

Record Data

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

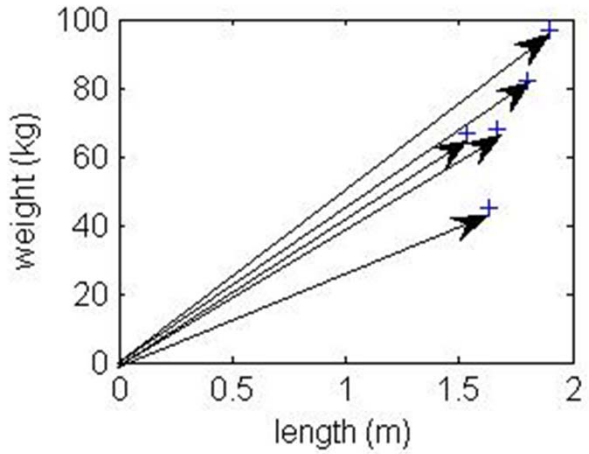
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

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

Tid	Length	Weight
1	1.80	82
2	1.53	67
3	1.67	68
4	1.90	97
5	1.63	45



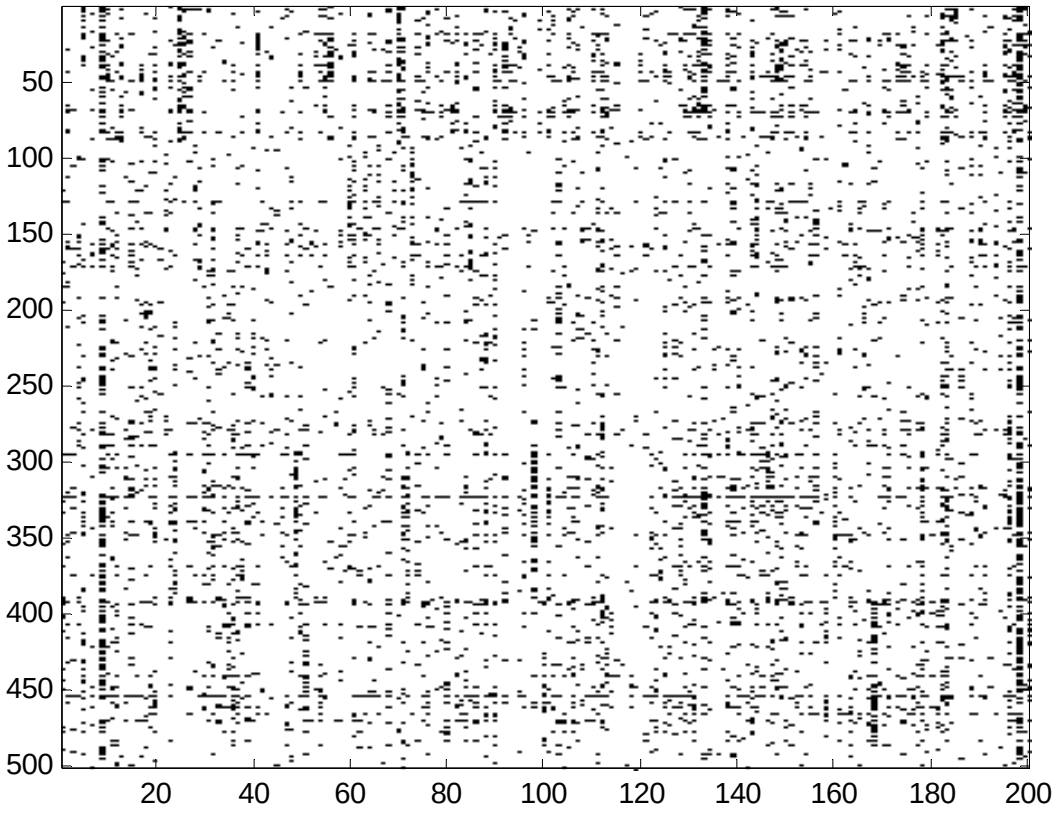
Document Data

- 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

Document	team	coach	play	ball	score	game	win	lost	timeout	season
1	3	0	5	0	2	6	0	2	0	2
2	0	7	0	2	1	0	0	3	0	0
3	0	1	0	0	1	2	2	0	3	0
4	1	4	0	2	3	0	1	6	2	1
5	2	3	3	1	6	1	3	0	0	4

Sparse Document Matrix

Text documents



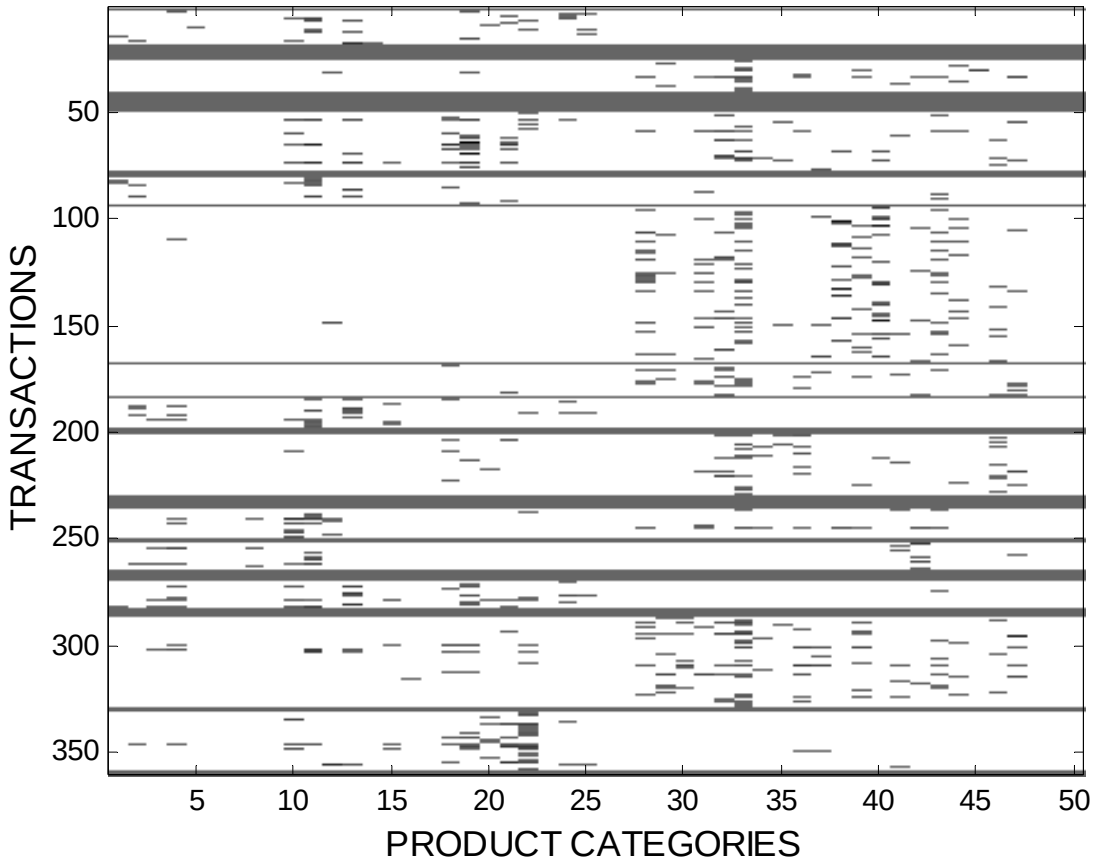
Word IDs

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 constitutes a transaction, while the individual products that were purchased are the items.

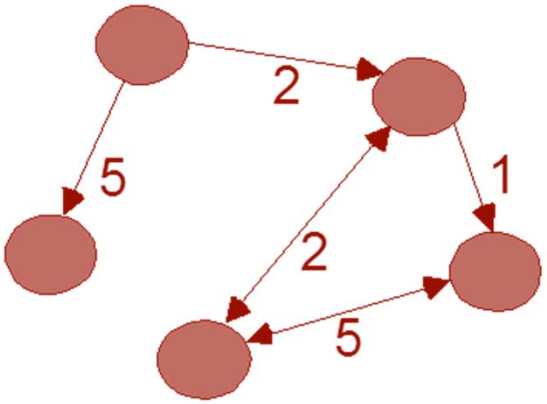
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Market Basket Data



Graph Data

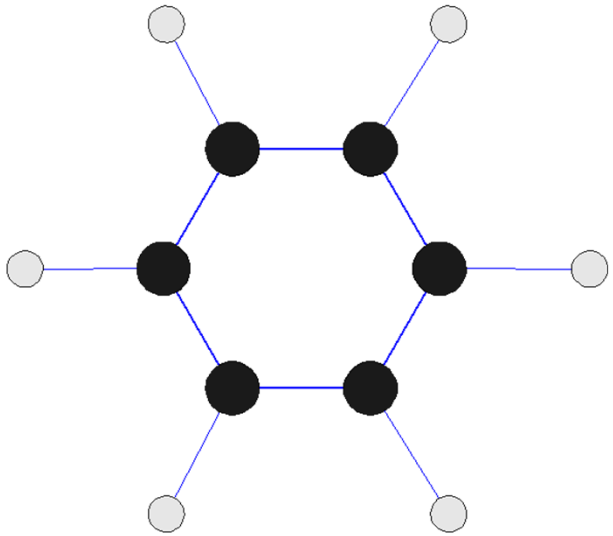
- Examples: Generic graph and HTML Links




```
<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



Sequence (Web) Data

128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,

User	*Visits*
1	2 3 2 2 3 3 3 1 1 1 3 1 3 3 3 3
2	3 3 3 1 1 1
3	7 7 7 7 7 7 7 7
4	1 5 1 1 1 1 5 1 5 1 1 1 1 1
5	5 1 1 5

Genomic Sequence Data

ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADA
BDBBDABABBCDDDCDDABDCBBDBDBCBBABBBCBBABCBB
ACBBDBAACCADDADBDBBCBBCCBBBDCABDDBBADDBBBB
CCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBACDCA
DCBACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCC
ACACACCDABDDBCADADBCBDDADABCCABDAACABCABAC
BDDDCBADCBDADDDDCDDCADCCBBADABBAAADAAABCCB
CABDBAADCBCDACBCABABCCBACBDABDDDADAABADCDC
CDBBCDBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCD
CCBCCCDCCADAADACABDABAABBDDBCADDDDBCDDBCCB
BCCDADADACCCDABAABBCBDBDBADBBBBCDADABABBDA
CDCDDDBBCDBBCBBCCDABCADDADBACBBBCCDBAAADDD
BDDCABACBCADCDCBAAADCADDADAABBACCBB

Genomic Sequence Data

ADACABDABAABBDDBCADDDDBCDDBC**CBBCC**DADADAADA
BDBBDABABBCDDDCDDABDCBBDBDBCBBABBBCBBABCBB
ACBBDBAACCADDADBDBB**CBBCC**BBBDCABDDBBADDBBBB
CCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBACDCA
DCBACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCC
ACACACCDABDDBCADADBCBDDADABCCABDAACABCABAC
BDDDCBADCBDADDDDCDDCADCCBBADABBAAADAAABCCB
CABDBAADCBCDACBCABABCCBACBDABDDDADAABADCDC
CDBBCDBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCD
CCBCCCDCCADAADACABDABAABBDDBCADDDDBCDDBC**CB**
BCCDADADACCCDABAABBCBDBDBADBBBBCDADABABBDA
CDCDDDBBCDBB**CBBCC**DABCADDADBACBBBCCDBAAADDD
BDDCABACBCADCDCBAAADCADDADAABBACCBB

Spatio-Temporal Data

Average monthly temperature of land and ocean

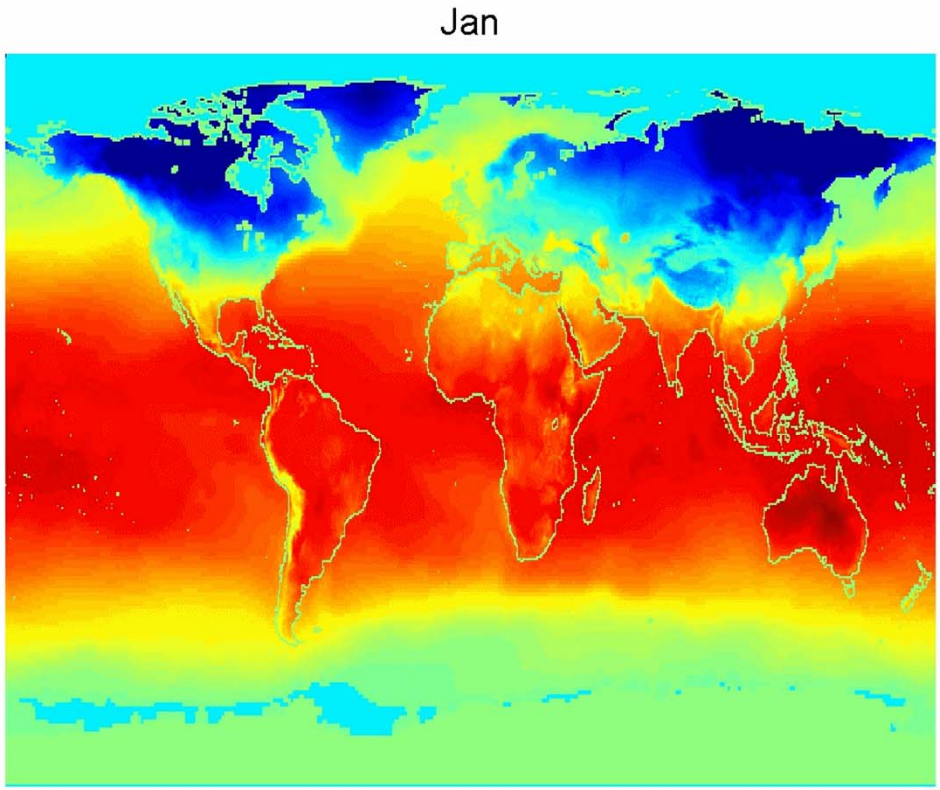
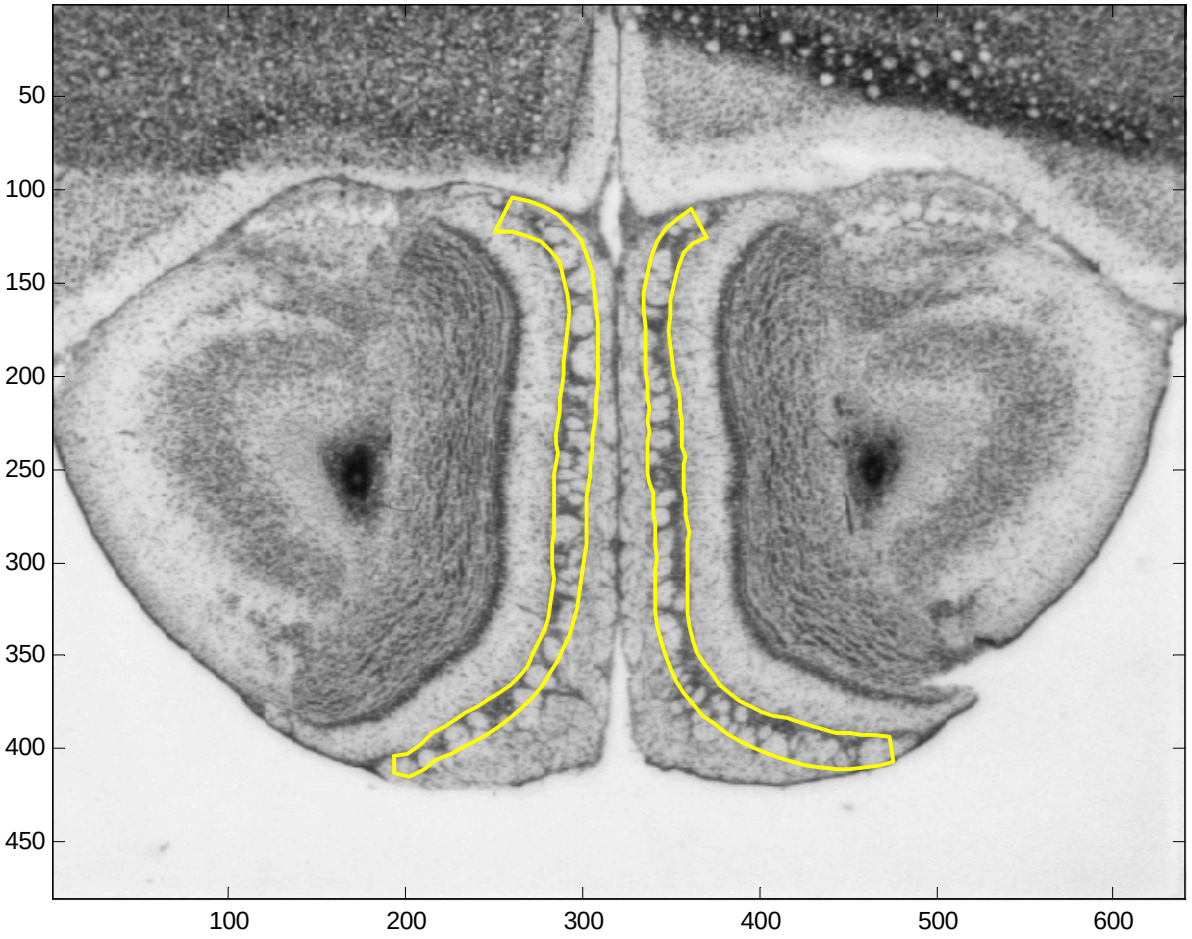
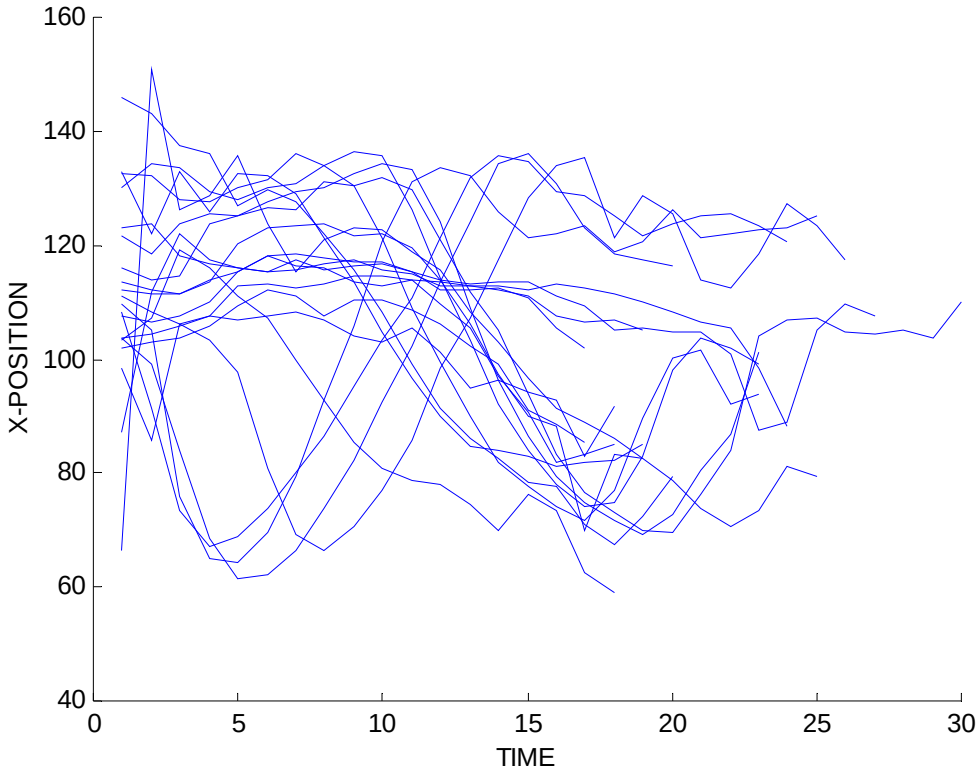


Image Data

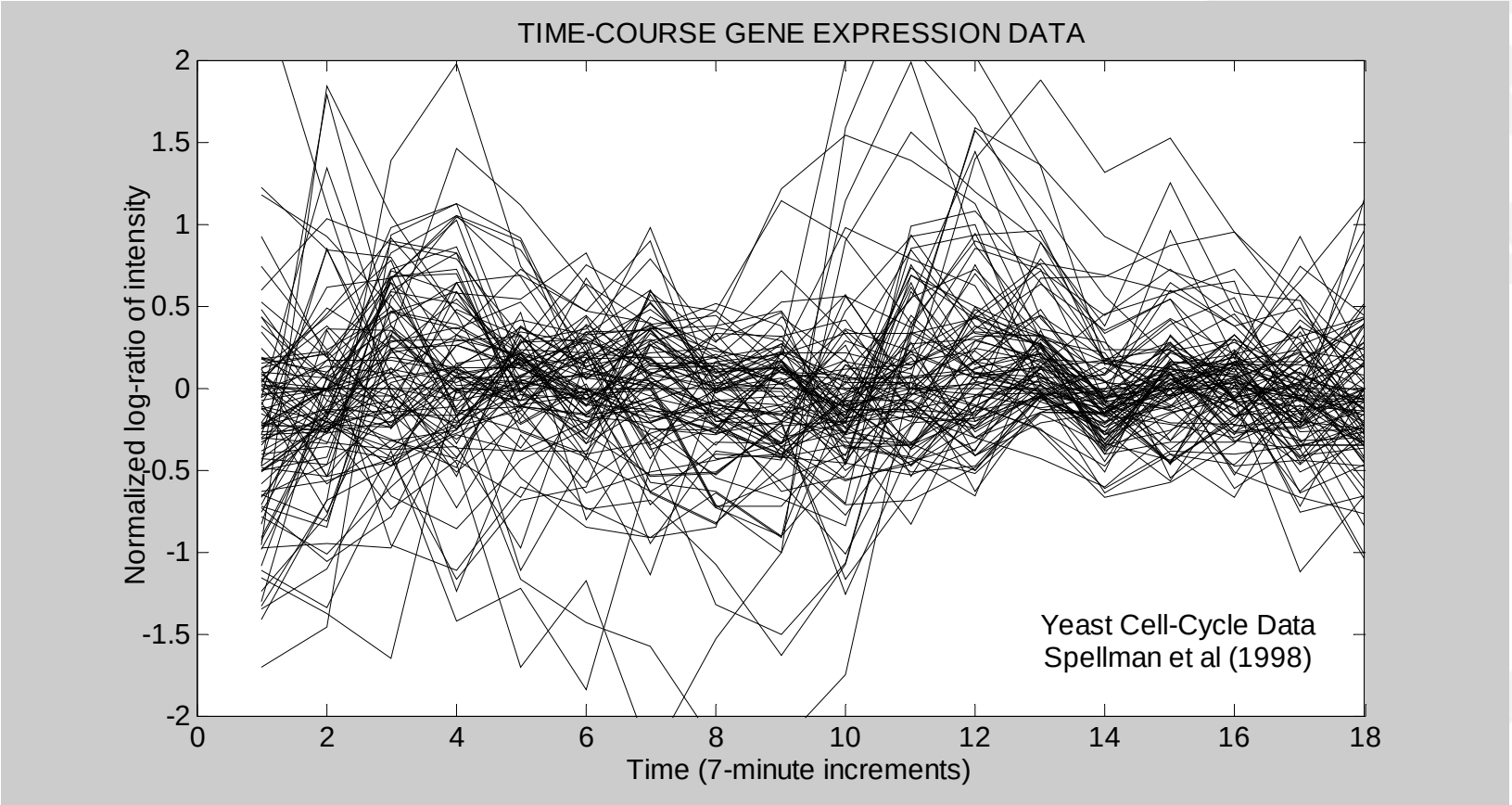


Time Series Data

Trajectories of centroids of moving hand in video streams

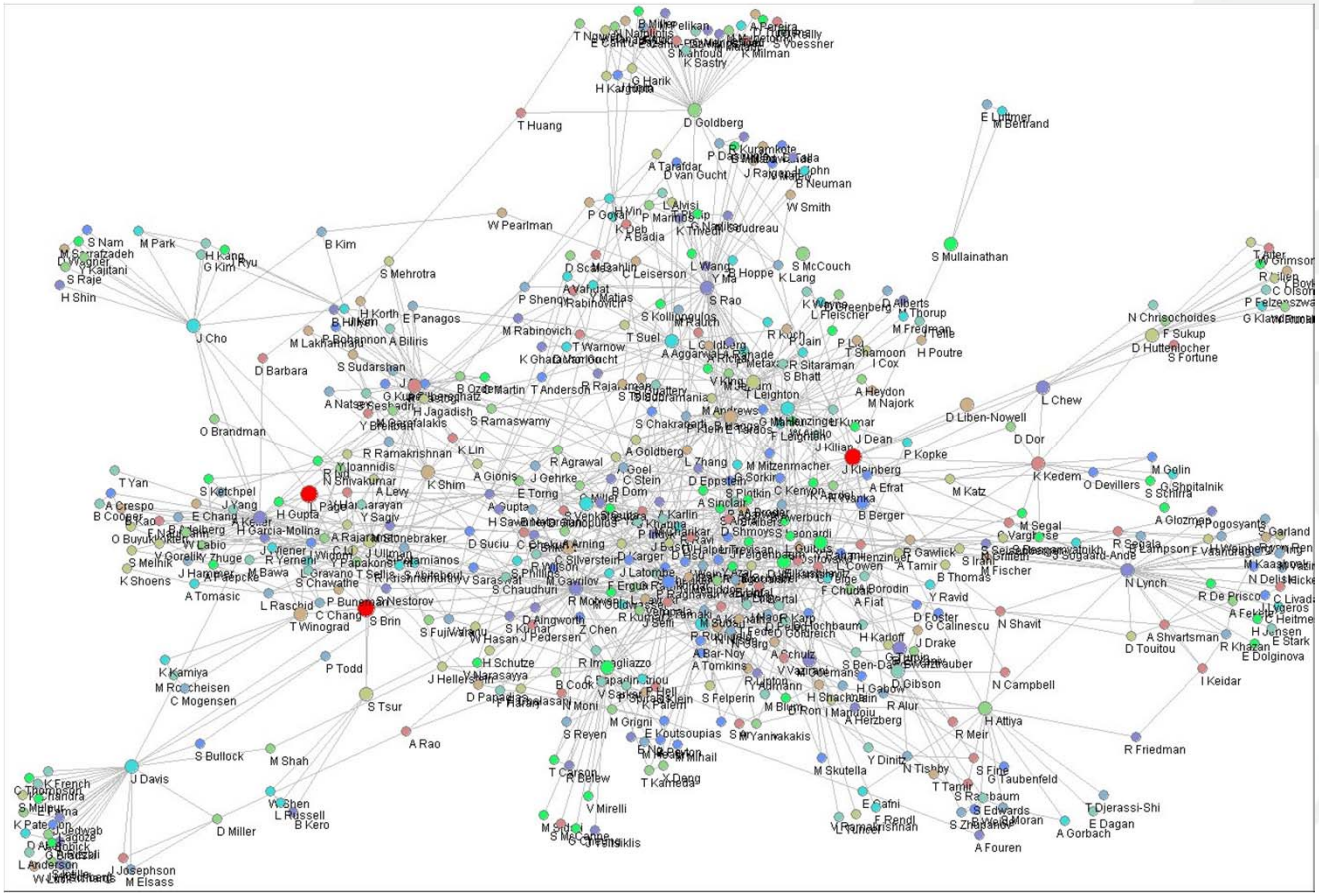


Biological Time Series

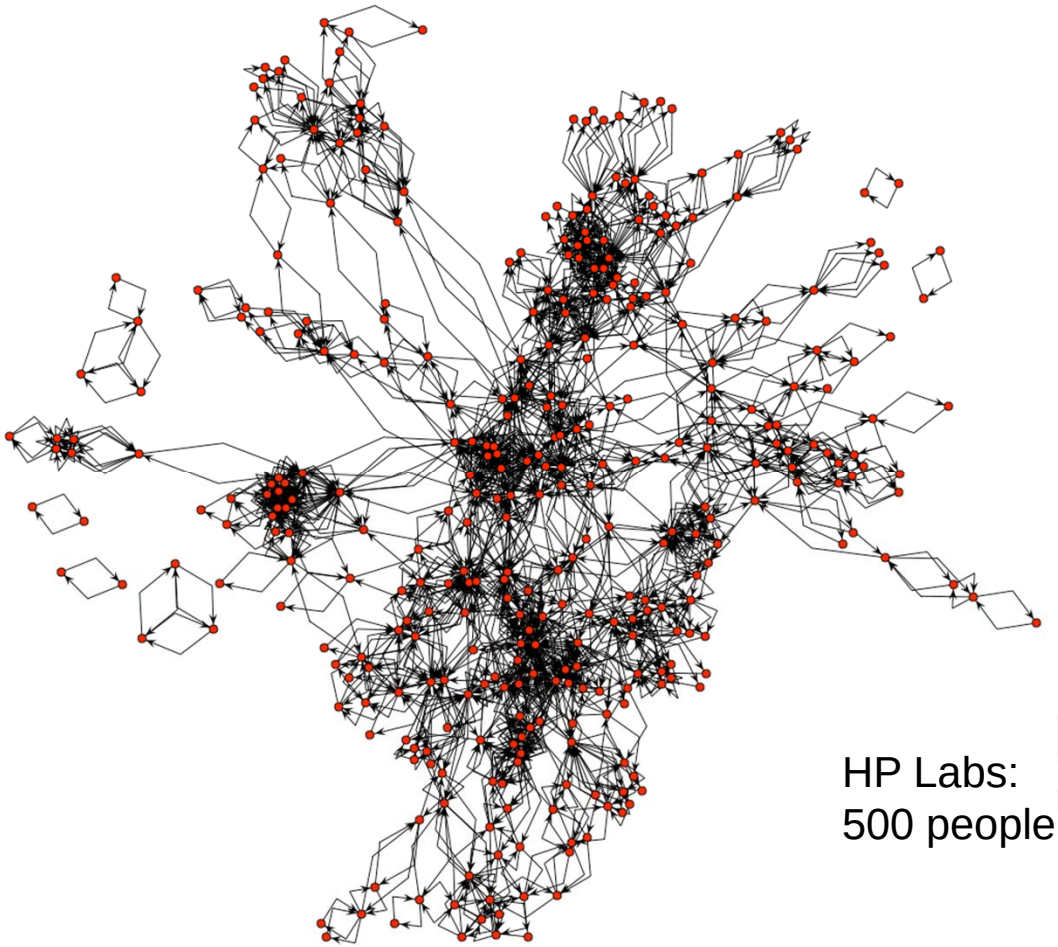


Relational Data

Co-author network



Email Network



HP Labs:
500 people, 20k relationships

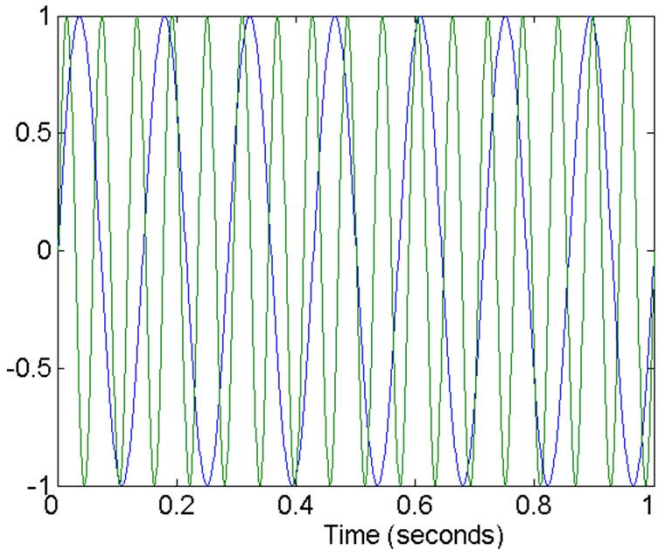
Data Quality

- Data is of high quality if they
  - Are fit for their intended use
  - Correctly represent the phenomena they correspond to

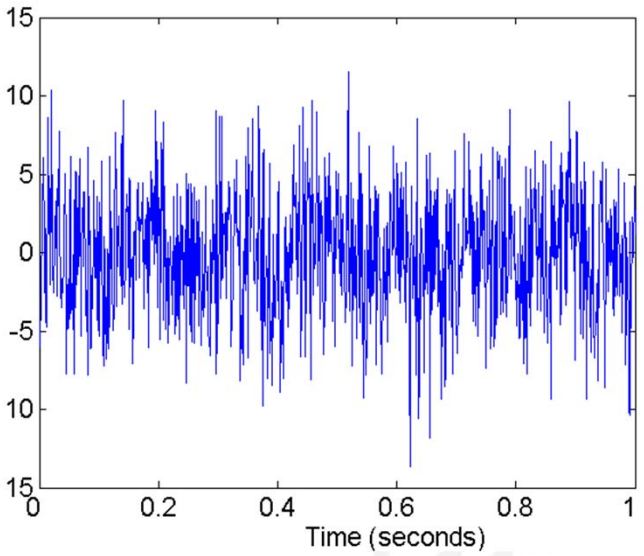
- Examples of data quality problems:
  - noise and outliers
  - missing values
  - duplicate data

Noise

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



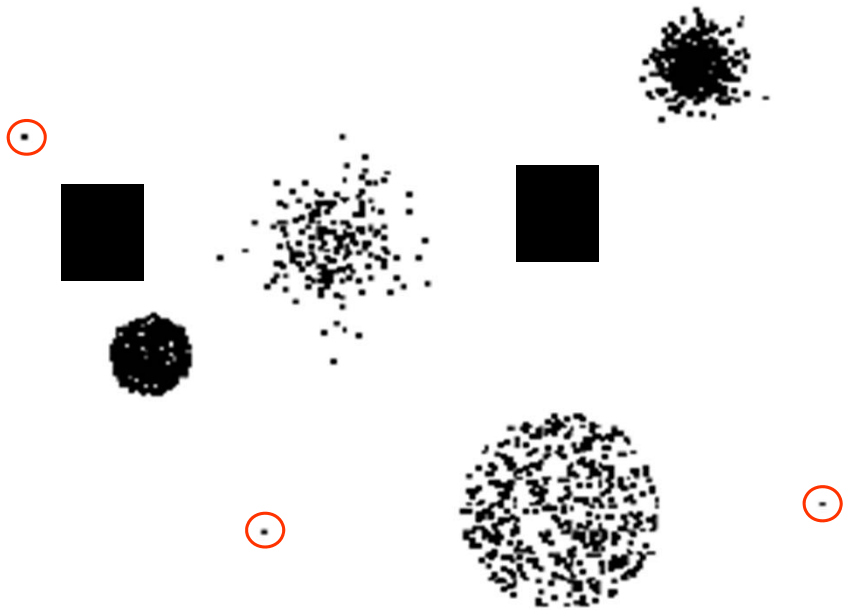
Two sine waves



Two sine waves + noise

Outliers

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



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
  - 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 heterogeneous sources
- Example: same person with multiple email addresses
- Data cleaning

Data Preprocessing

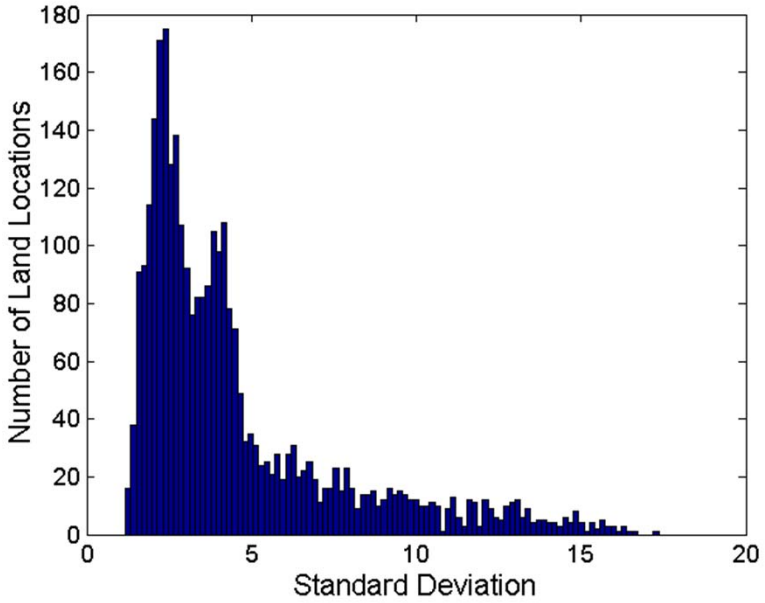
- Aggregation
- Sampling
- Dimensionality reduction
- Feature subset selection
- Feature creation
- Discretization
- Attribute transformation

Aggregation

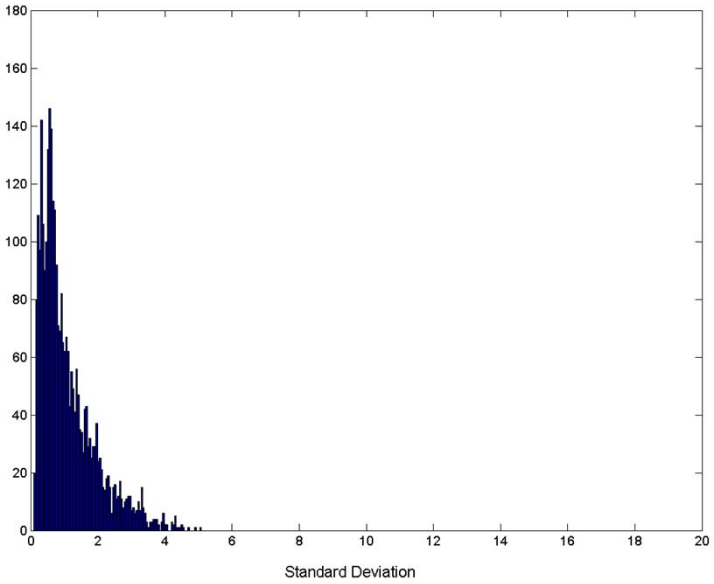
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction: reduce the number of attributes or objects
  - Change of scale: cities aggregated into regions, states, countries, etc
  - More “stable” data: aggregated data tends to have less variability

Aggregation

Variation of Precipitation in Australia



Standard deviation of average monthly precipitation



Standard deviation of average yearly precipitation

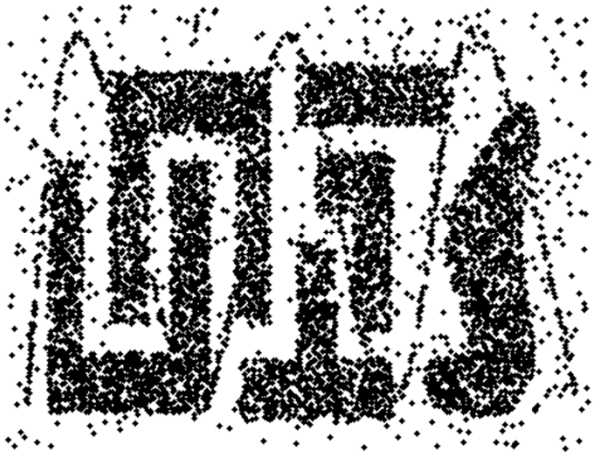
Sampling

- Sampling is the main technique employed for data selection.
  - Used for both the preliminary investigation of the data and the final 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.
- Key principle for effective sampling:
  - Using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

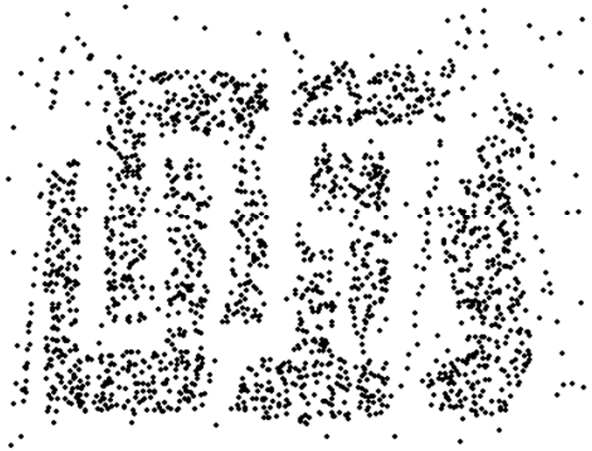
Types of Sampling

- 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: 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

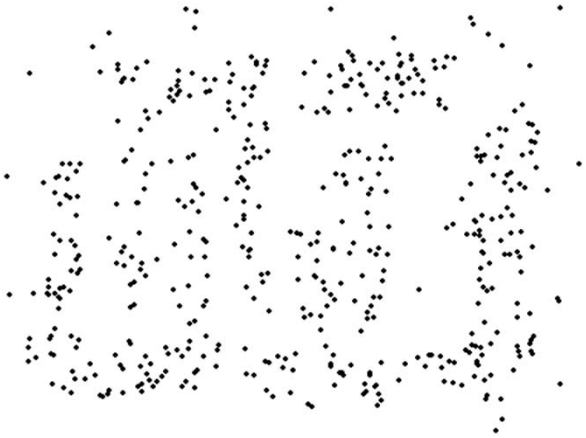
Sample Size



8000 points



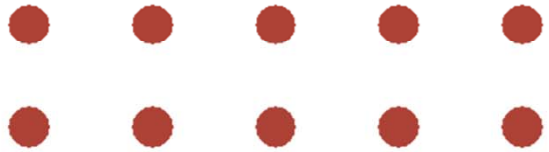
2000 Points

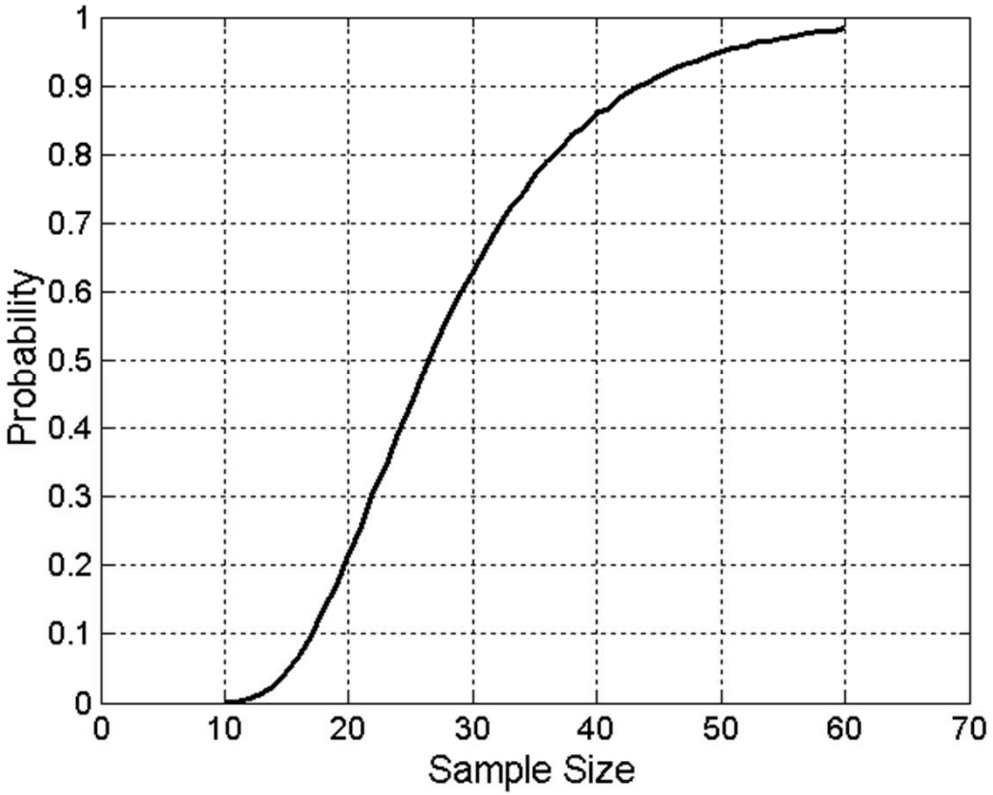


500 Points

Sample Size

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





MATLAB Code Sample Size (1)

```
nclusters = 10;
npoints = 40;
nsamples = 10000;

% Uniform probability over the clusters

probability = ones(1,nclusters)/nclusters;

% Draw nsamples times from a multinomial probability

x = zeros(nsamples,nclusters);
for i=1:nsamples,
    x(i,:) = mnrnd(npoints,probability);
end
                % e.g., x(7,:) = [2,5,3,0,4,8,4,7,2,5]
```

MATLAB Code Sample Size (1)

```
nclusters = 10;
npoints = 40;
nsamples = 10000;

% Uniform probability over the clusters

probability = ones(1,nclusters)/nclusters;

% Draw nsamples times from a multinomial probability

x = mnrnd(npoints,probability,nsamples);


                % e.g., x(7,:) = [2,5,3,0,4,8,4,7,2,5]
```

MATLAB Code Sample Size (2)

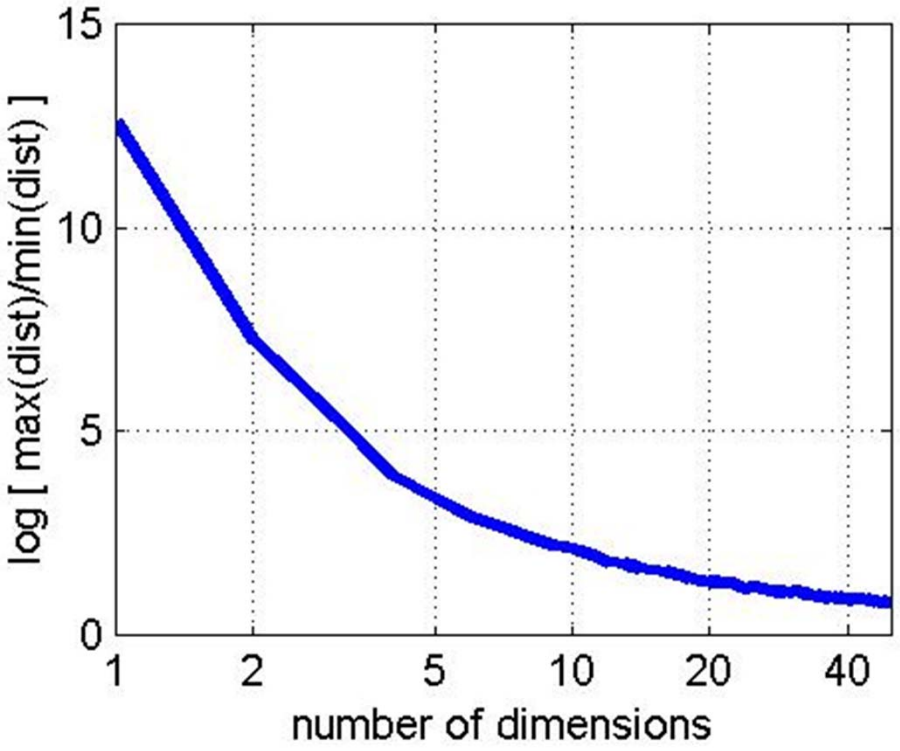
```
% Count samples without zeros

teller = 0;
for i=1:nsamples,
    if all(x(i,:) > 0),
        teller = teller+1;
    end
end

% Estimate probability

ppp = teller/nsamples
```

MATLAB Code Sample Size (2)

```
% Count samples without zeros

teller = sum(all(x,2));



% Estimate probability

ppp = teller/nsamples
```

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



- Randomly generate 500 points
- Compute (log) ratio of max and min distance between any pair of points

MATLAB Code Curse of Dimensionality

```
dimension = 40;
ndatapoints = 500;

x = rand(dimension,ndatapoints);  % uniform from [0,1]

% plot data points

switch dimension
    case 1
        hist(x,20)
    case 2
        plot(x(1,:),x(2,:),'.')
    case 3
        plot3(x(1,:),x(2,:),x(3,:),'.')
end
```

MATLAB Code Curse of Dimensionality

```
% compute distance

distance = dist(x);   % Euclidean distance between rows

% show histogram

upperdiag = triu(distance);   % take upper diagonal
hist(upperdiag(upperdiag > 0),20);

% compute min and max

mindist = min(distance(distance > 0))
maxdist = max(distance(distance > 0))
```

Dimensionality Reduction

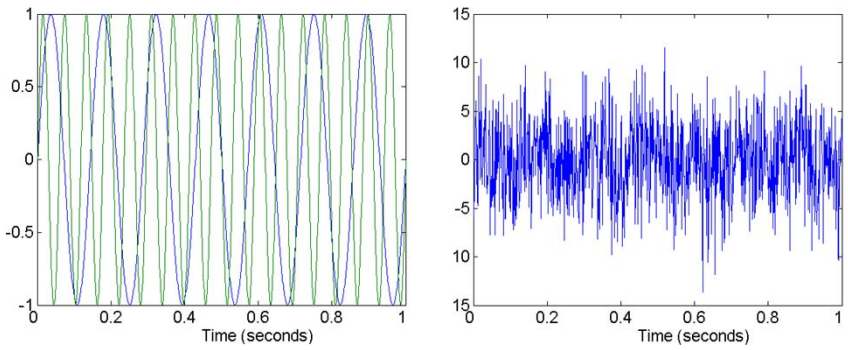
- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Techniques
  - Principal Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

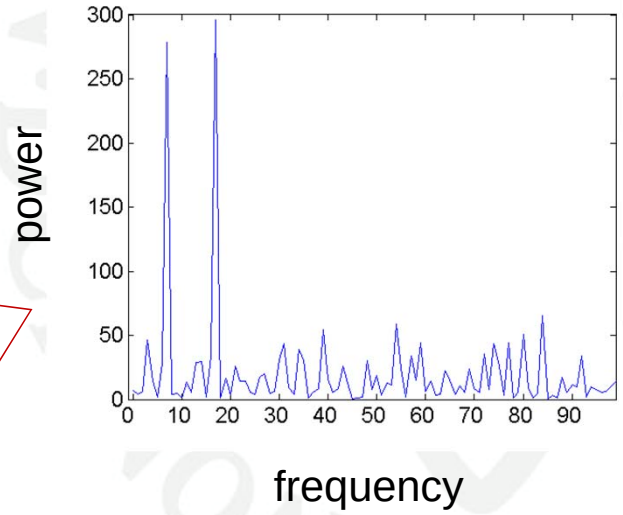
Feature Subset Selection

- Techniques:
  - Brute-force approach: Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches: Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches: Features are selected before data mining algorithm is run
  - Wrapper approaches: Use the data mining algorithm as a black box to find best subset of attributes

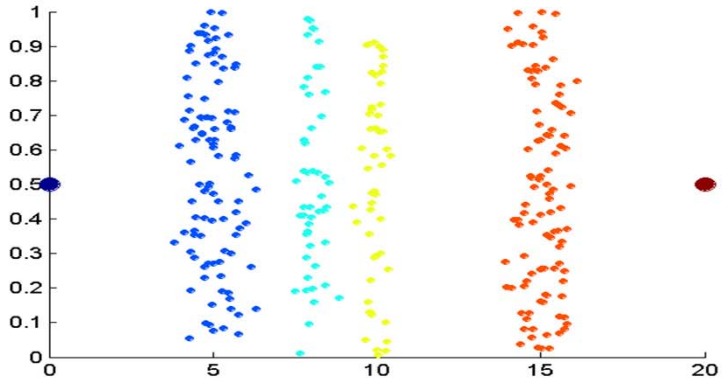
Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Combining features
  - For example, BMI instead of length and weight separately
  - Particularly relevant for restricted (e.g., linear) models
- Mapping data to a new space
  - For example, Fourier transform

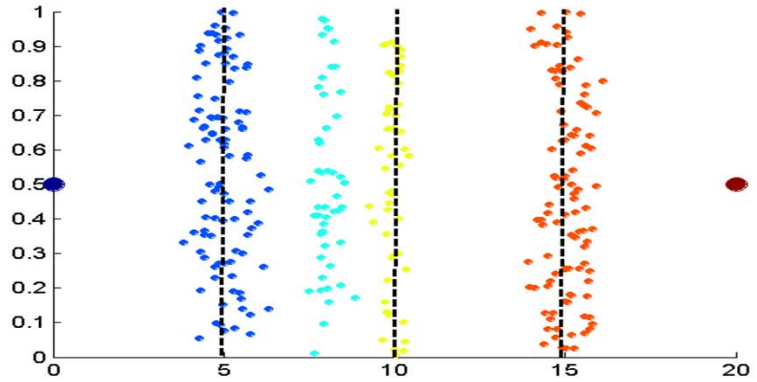




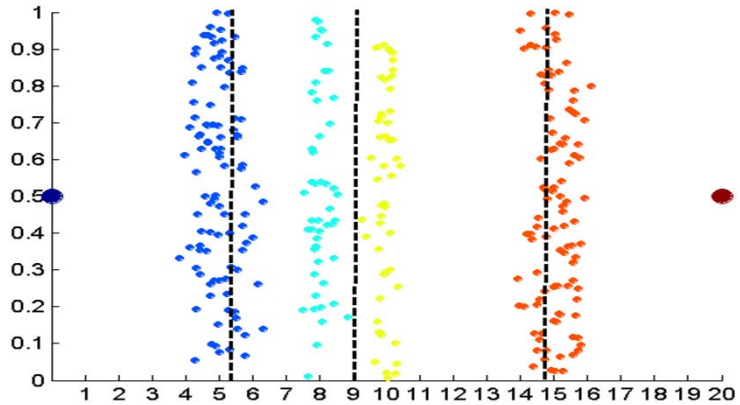
Discretization Without Using Class Labels



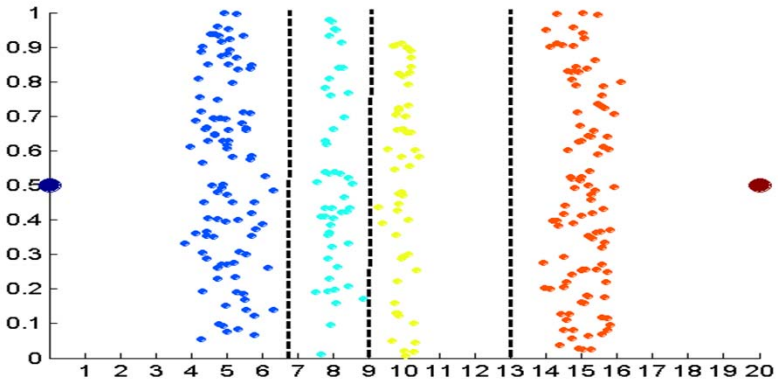
Data



Equal interval width



Equal frequency



K-means

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of values such that each old value can be identified with one of the new values
  - Simple functions: $x^k, \log x, e^x, |x|$
  - Standardization and normalization
- For many data mining algorithms, continuous features are preferentially more or less normally distributed
- Log transformation often useful for positive features such as income, height, etc.

Quiz Question

You're given a set of newspaper articles on different topics.

- Describe a classification problem based on this data.
- Describe a clustering problem based on this data.

"You can't keep adjusting the data to prove that you would be the best Valentine's date for Scarlett Johansson."