Data Mining: Data

Chapter 2

Introduction to Data Mining

by

Tan, Steinbach, Kumar

(slides modified by Predrag Radivojac, 2024)

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Student**	**Assignment**	**Assignment**	**Assignment**	**Assignment**	**Midterm Ex**	**Final Exam**	**Mini Projec**	**Assignment**	**Activity**	**Final**	**Grade**
2	Perfectus Discipulus	160	100	50	180	57	57	50	490	10	100	
3	Student #1	160	99	50	178	50	49	50	487	10	94.55317	A
4	Student #2	155	99	46	178	49	48	49	478	10	92.80039	
5	Student #3	155	100	47	162	52	51	49	464	5	91.54851	
6	Student #4	159	99	50	175	49	46	50	483	5	90.40476	
7	Student #5	151	95	46	175	44	45	50	467	10	89.81991	A
8	Student #6	158	86	47	161	52	50	46	452	5	88.96294	A-
9	Student #7	156	100	47	178	39	42	50	481	10	87.87003	
10	Student #8	159	95	47	168	44	39	50	469	10	87.83709	
11	Student #9	158	100	47	178	39	39	50	483	10	86.93985	
12	Student #10	137.7	98	49	167	44	47	49	451.7	5	86.58493	A-
13	Student #11	160	95	49	154	40	48	49	458	5	85.91801	B+
14	Student #12	155	97	0	170	51	46	47	422	5	85.87182	
15	Student #13	96	80.1	47	153	45	48	49	376.1	10	85.15811	
16	Student #14	129	92	43	153	46	50	45	417	5	84.21482	
17	Student #15	158	88	44	176	41	45	45	466	5	83.70605	
18	Student #16	152	82	44	148	43	48	45	426	5	83.01146	B+
19	Student #17	143	97	44	160	42	42	45	444	5	81.65736	B
20	Student #18	160	95	47	169	41	39	42	471	5	80.40691	
21	Student #19	159	93	47	166	38	43	42	465	5	80.39044	
22	Student #20	152	94	40	158	44	33	47	444	5	80.20122	B
23	Student #21	142	90	47	165	28	37	42	444	5	73.49069	C+

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

Class

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

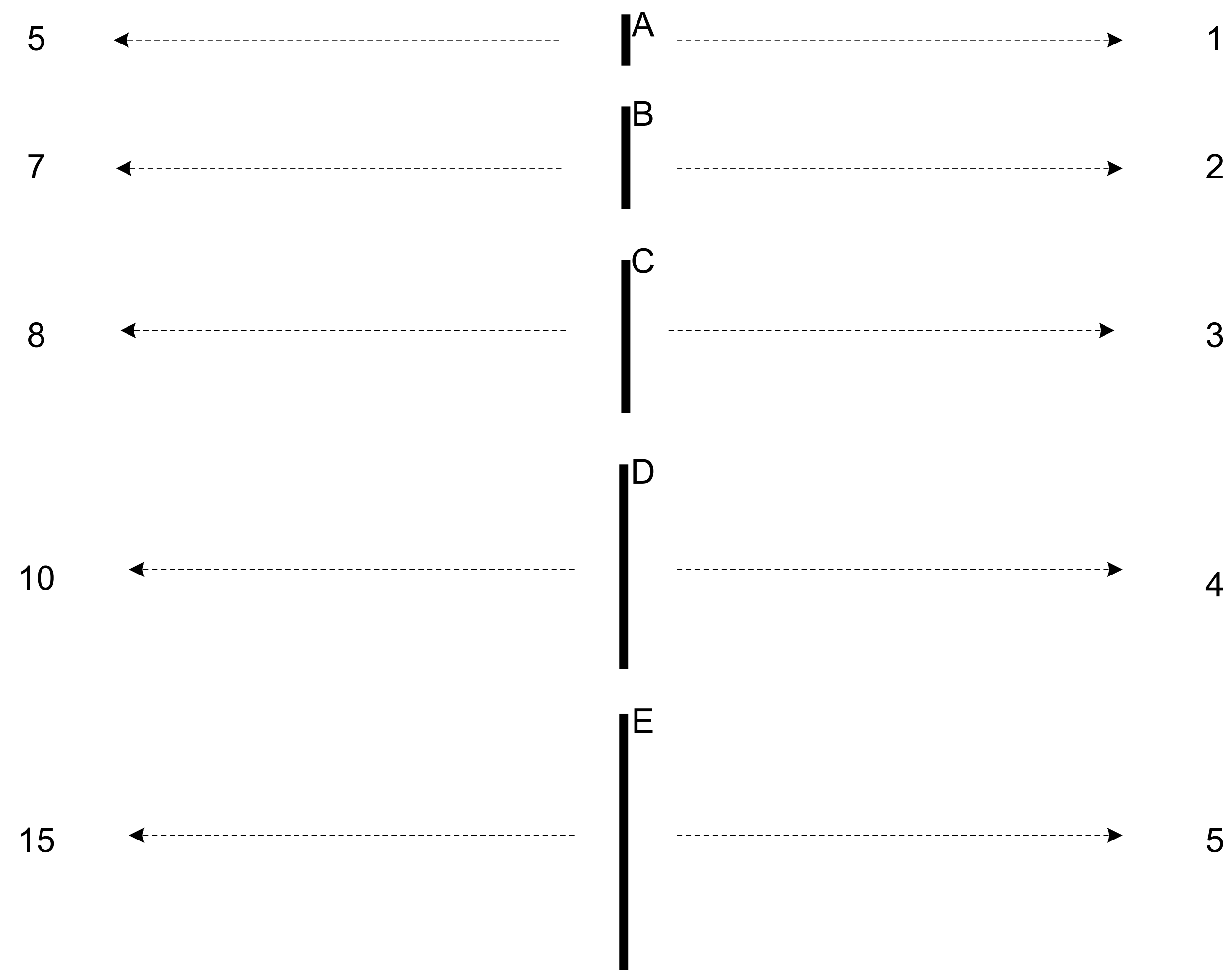
- Attribute values are numbers or symbols assigned to an attribute

- 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
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Measurement of Length

- The way you measure an attribute is something that may not match the attributes properties



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

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $= \quad \neq$
  - Order: $< \quad >$
  - Addition: $+ \quad -$
  - Multiplication: $* \quad /$

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ($=$, $\neq$)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. ($<$, $>$)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. ($+$, $-$)	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, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	**Transformation**	**Comments**
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}.
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.

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.

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 Structured Data

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

$$
\begin{bmatrix}
 & & & & \\
 & & & 1 & \\
1 & & & & \\
 & 1 & & & \\
 & & & & \\
 & & & & \\
 & & & & \\
 & & 1 & & \\
 & & 1 & & \\
 & & & & 1
\end{bmatrix}
$$

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

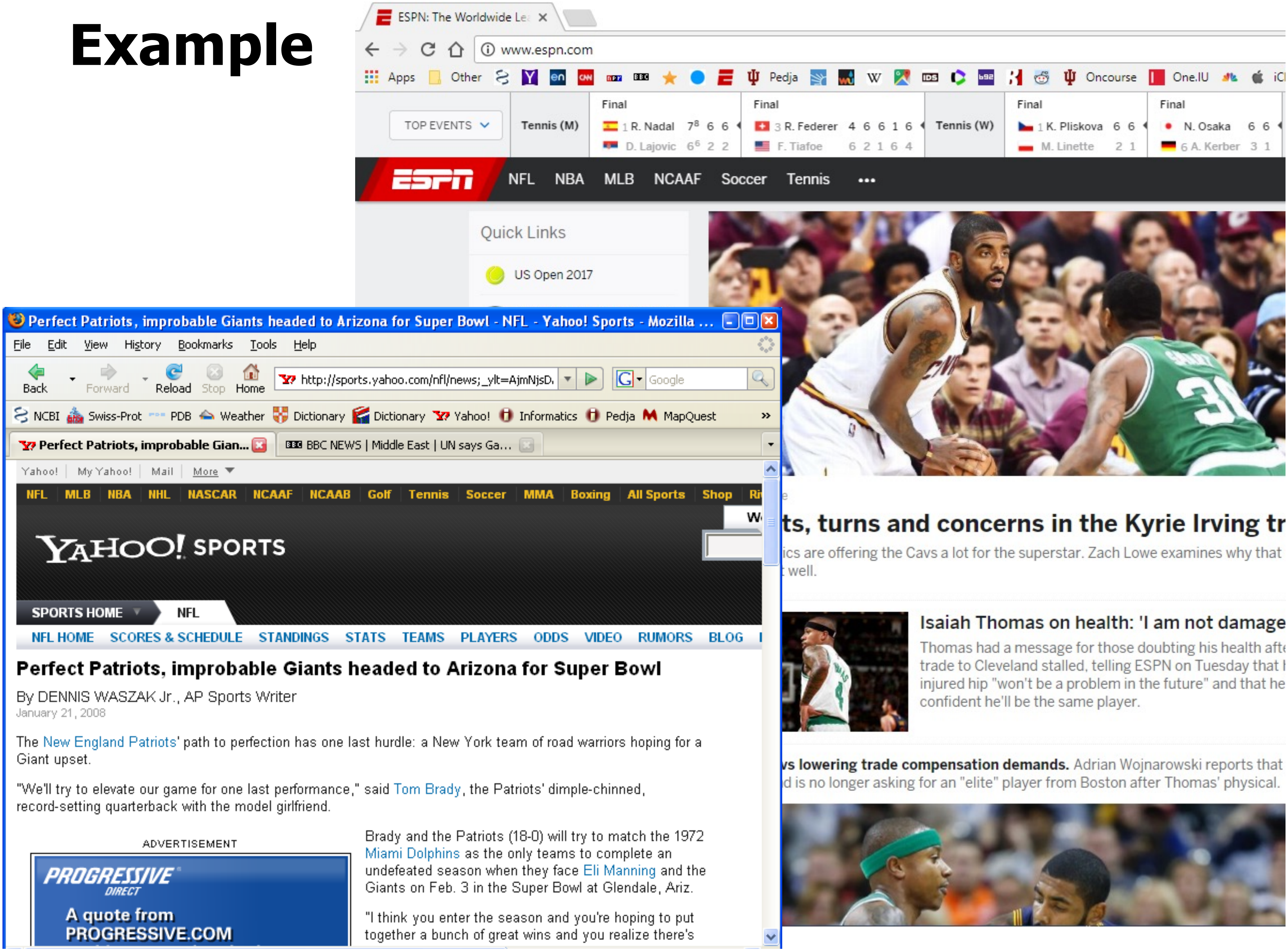
Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data (bag of words)

- 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	pla y	ball	score	game	wi n	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

Example


ESPN: The Worldwide Le
www.espn.com
Apps
Other
Pedja
Oncourse
One.IU
TOP EVENTS
Tennis (M)
Final
1 R. Nadal 7 6 6
D. Lajovic 6 2 2
Final
3 R. Federer 4 6 6 1 6
F. Tiafoe 6 2 1 6 4
Tennis (W)
Final
1 K. Pliskova 6 6
M. Linette 2 1
Final
N. Osaka 6 6
6 A. Kerber 3 1
NFL NBA MLB NCAAF Soccer Tennis
Quick Links
US Open 2017
ts, turns and concerns in the Kyrie Irving tr
ics are offering the Cavs a lot for the superstar. Zach Lowe examines why that
well.
Isaiah Thomas on health: 'I am not damage
Thomas had a message for those doubting his health afte
trade to Cleveland stalled, telling ESPN on Tuesday that h
injured hip "won't be a problem in the future" and that he
confident he'll be the same player.
s lowering trade compensation demands. Adrian Wojnarowski reports that
d is no longer asking for an "elite" player from Boston after Thomas' physical.
Perfect Patriots, improbable Giants headed to Arizona for Super Bowl - NFL - Yahoo! Sports - Mozilla ...
File Edit View History Bookmarks Tools Help
Back Forward Reload Stop Home
http://sports.yahoo.com/nfl/news;_ylt=AjmNjsD
Google
NCBI Swiss-Prot PDB Weather Dictionary Dictionary Yahoo! Informatics Pedja MapQuest
Perfect Patriots, improbable Gian...
BBC NEWS | Middle East | UN says Ga...
Yahoo! My Yahoo! Mail More
NFL MLB NBA NHL NASCAR NCAAF NCAAB Golf Tennis Soccer MMA Boxing All Sports Shop
YAHOO! SPORTS
SPORTS HOME NFL
NFL HOME SCORES & SCHEDULE STANDINGS STATS TEAMS PLAYERS ODDS VIDEO RUMORS BLOG
Perfect Patriots, improbable Giants headed to Arizona for Super Bowl
By DENNIS WASZAK Jr., AP Sports Writer
January 21, 2008
The New England Patriots' path to perfection has one last hurdle: a New York team of road warriors hoping for a Giant upset.
"We'll try to elevate our game for one last performance," said Tom Brady, the Patriots' dimple-chinned, record-setting quarterback with the model girlfriend.
ADVERTISEMENT
PROGRESSIVE DIRECT
A quote from PROGRESSIVE.COM
Brady and the Patriots (18-0) will try to match the 1972 Miami Dolphins as the only teams to complete an undefeated season when they face Eli Manning and the Giants on Feb. 3 in the Super Bowl at Glendale, Ariz.
"I think you enter the season and you're hoping to put together a bunch of great wins and you realize there's

Recommendations Data

- Sparse matrix
  - each row is a person
  - each column is a movie (book, disease, …)
  - each number is a rating

Movies

Persons

	Spiderman	Ocean's 11	Matrix	Titanic	JFK	Star wars	Creed	Rocky
Person 1		3		4				
Person 2								5
Person 3							4	5
Person 4	1		3				2	

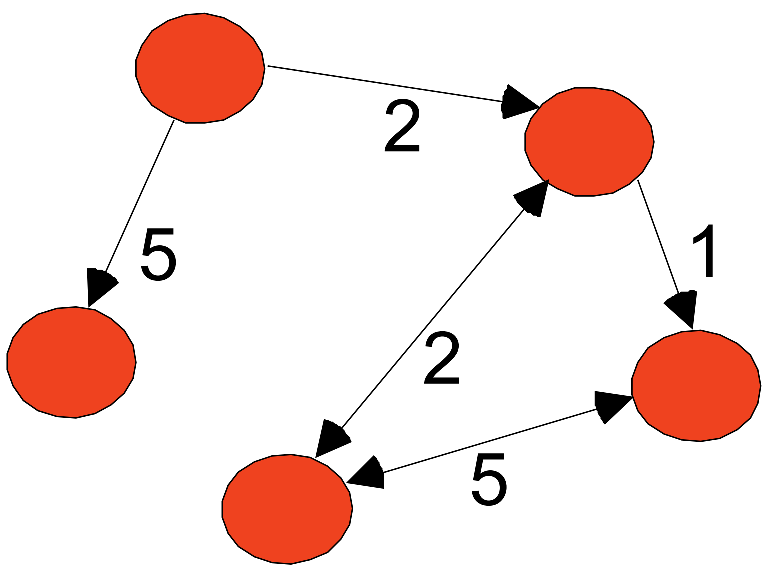
Transaction Data

- 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



Data Mining

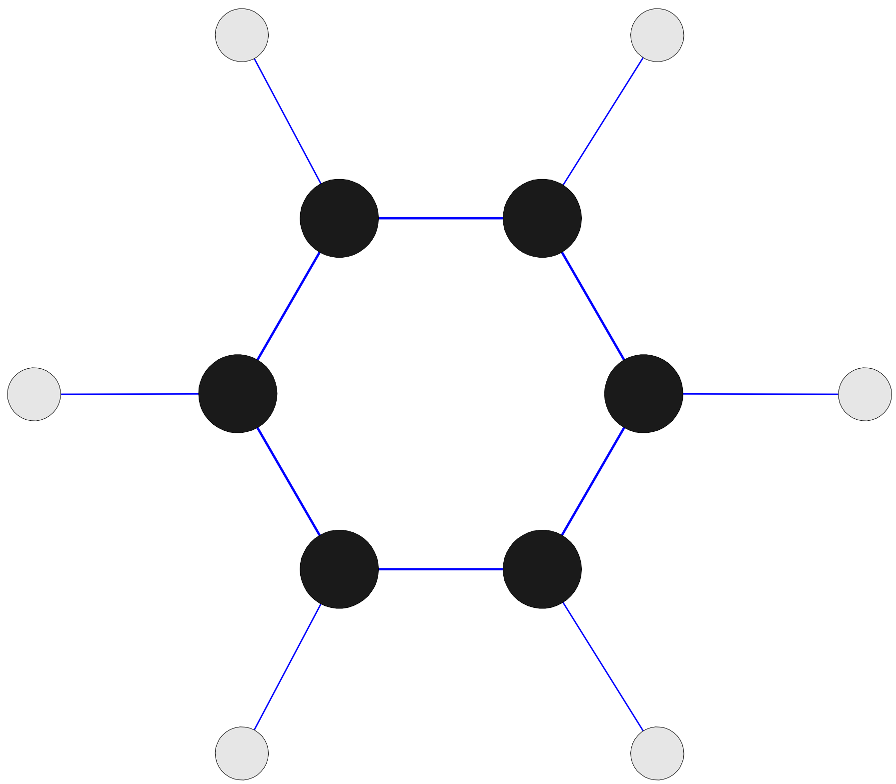
Graph Partitioning

Parallel Solution of Sparse Linear System of Equations

<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

Items/Events

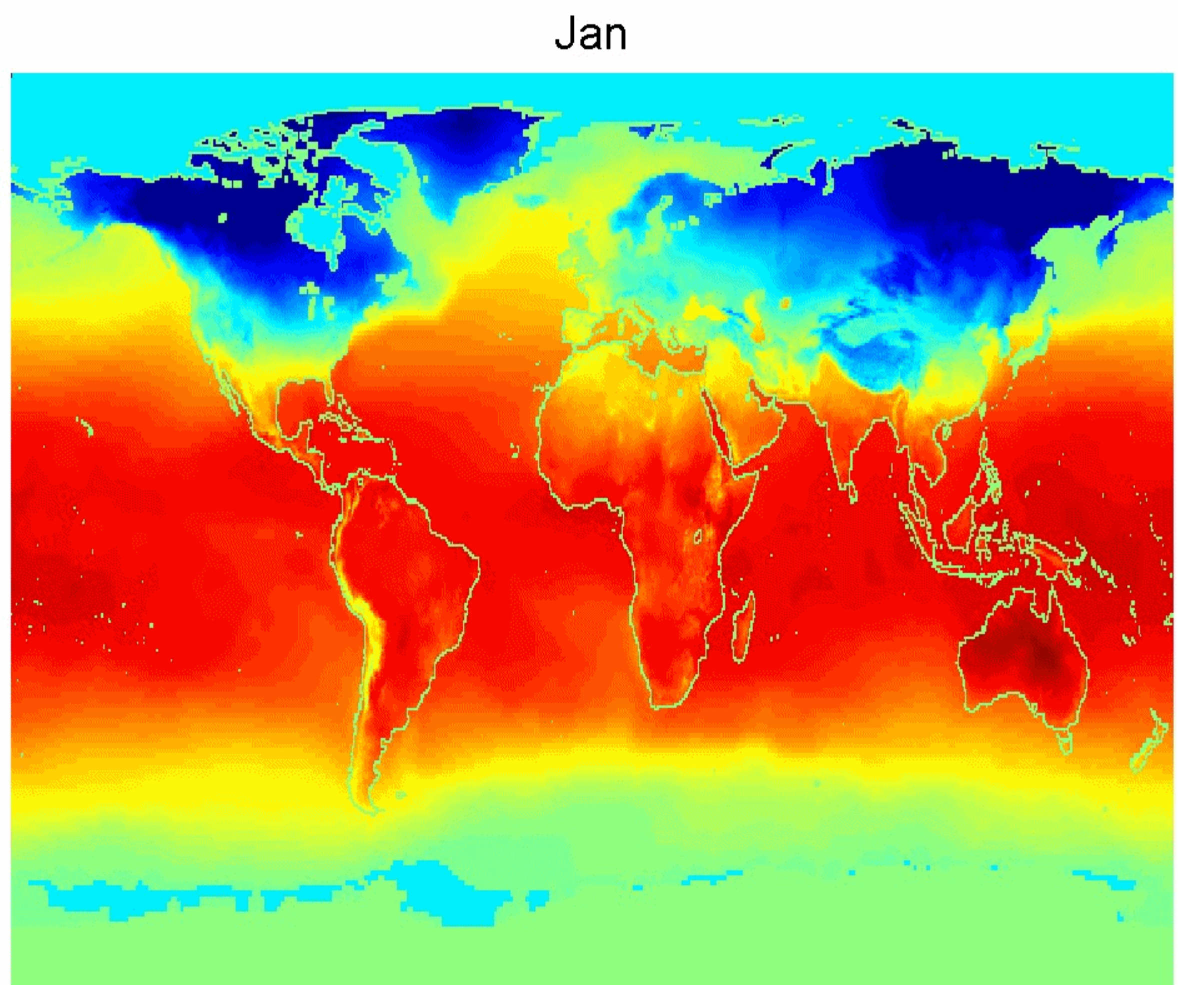
(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

Ordered Data

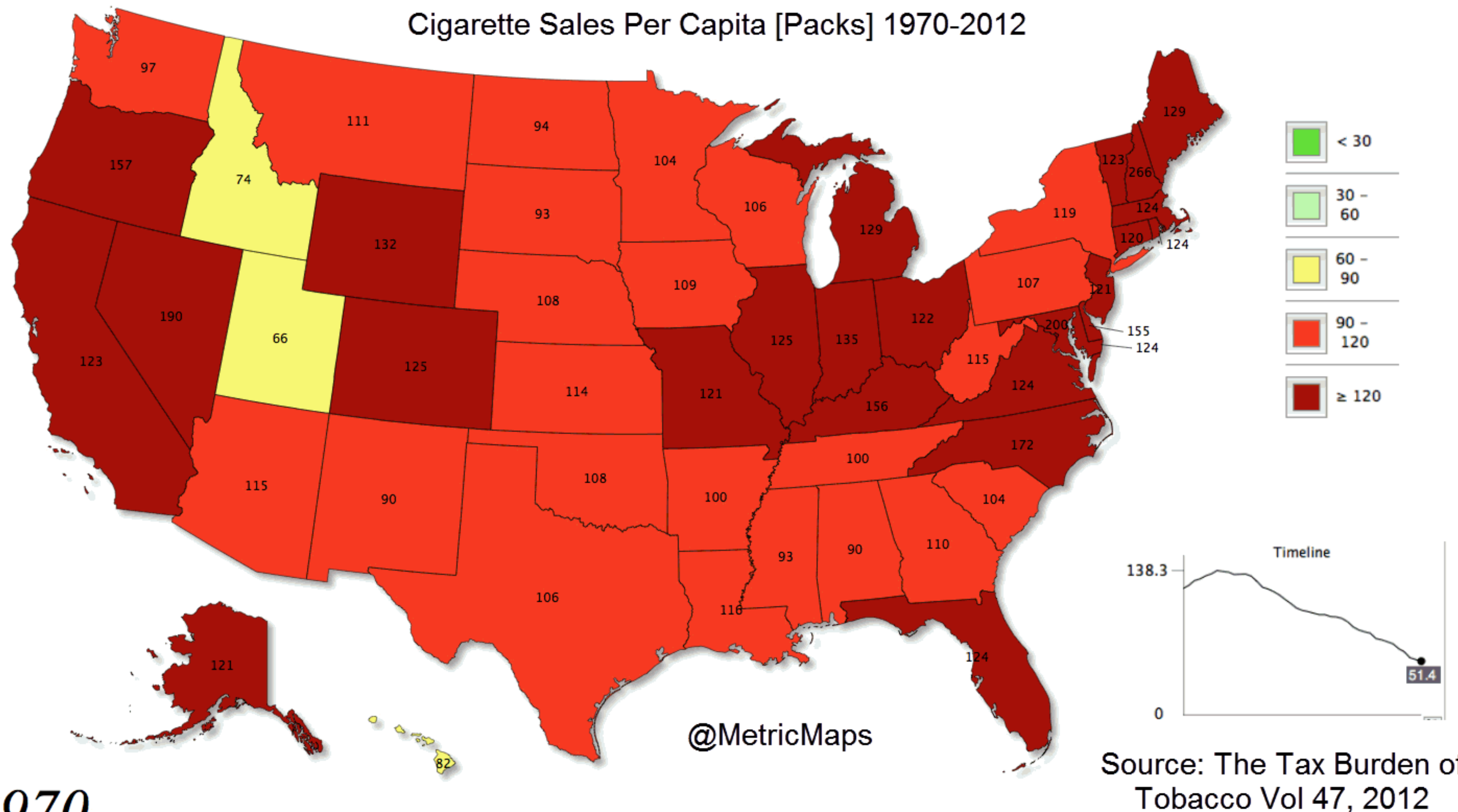
- Spatio-Temporal Data

Average monthly temperature of land and ocean



Ordered Data

- Spatio-Temporal Data

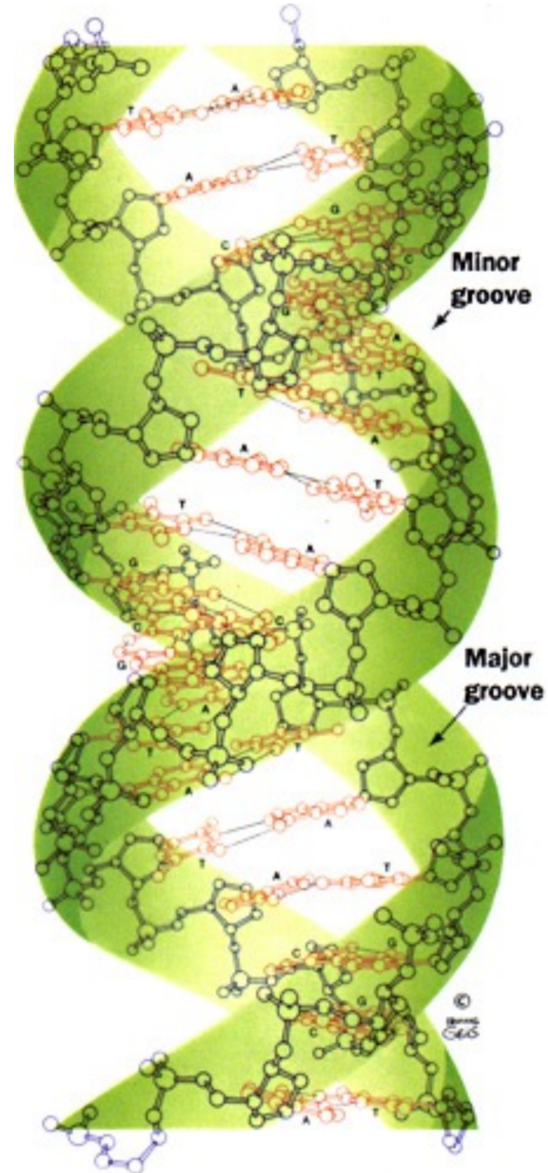


1970

Ordered/Sequence Data

- Genomic sequence data

```
1:   ...GGTTCCGCCTTCAGCCCCCCGCC...   0
2:   ...GGTTCCGCGTTCAGCCCCGCGCC...   1
3:   ...GGTTCCGCCTTCAGCCCCCCGCC...   0
4:   ...GGTTCCGCCTTCAGCCCCGCGCC...   0
5:   ...GGTTCCGCCTTCAGCCCCTCGCC...   0
6:   ...GGTTCCGCCTTCAGCCCCGCGCC...   0
7:   ...GGTTCCGCCTTCAGCCCCTCGCC...   0
8:   ...GGTTCCGCATTCAGCCCCCCGCC...   1
9:   ...GGTTCCGCCTTCAGCCCCGCGCC...   0
```



Data Repositories

- UCI Machine Learning Repository
  - contains a number of user deposited ML problems
  - https://archive.ics.uci.edu/ml/index.php

- Stanford Large Network Dataset Collection
  - Graphs of different kinds and sizes
  - https://snap.stanford.edu/data/

- Kaggle
  - Challenges for machine learning
  - https://www.kaggle.com

- Google it

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

Why Is Data Dirty?

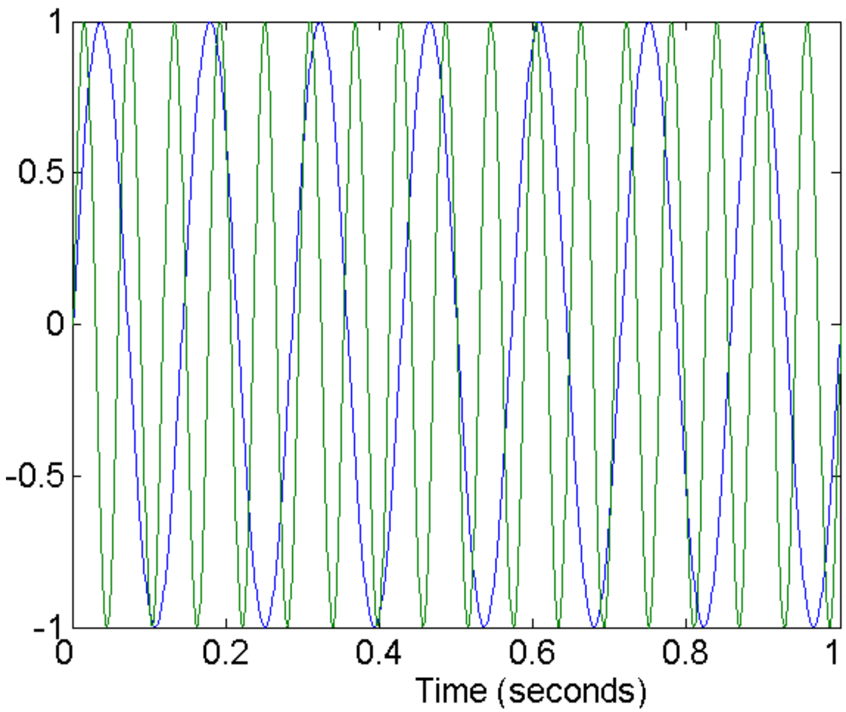
- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records may also need to be cleaned

Why Is Data Preprocessing Important?

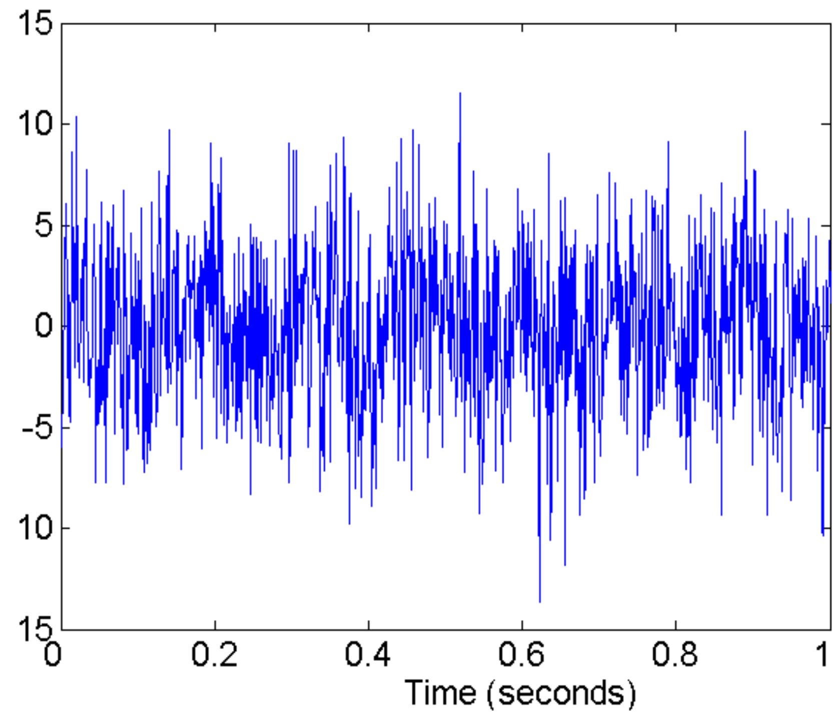
- **No quality data, no quality mining results!**
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

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



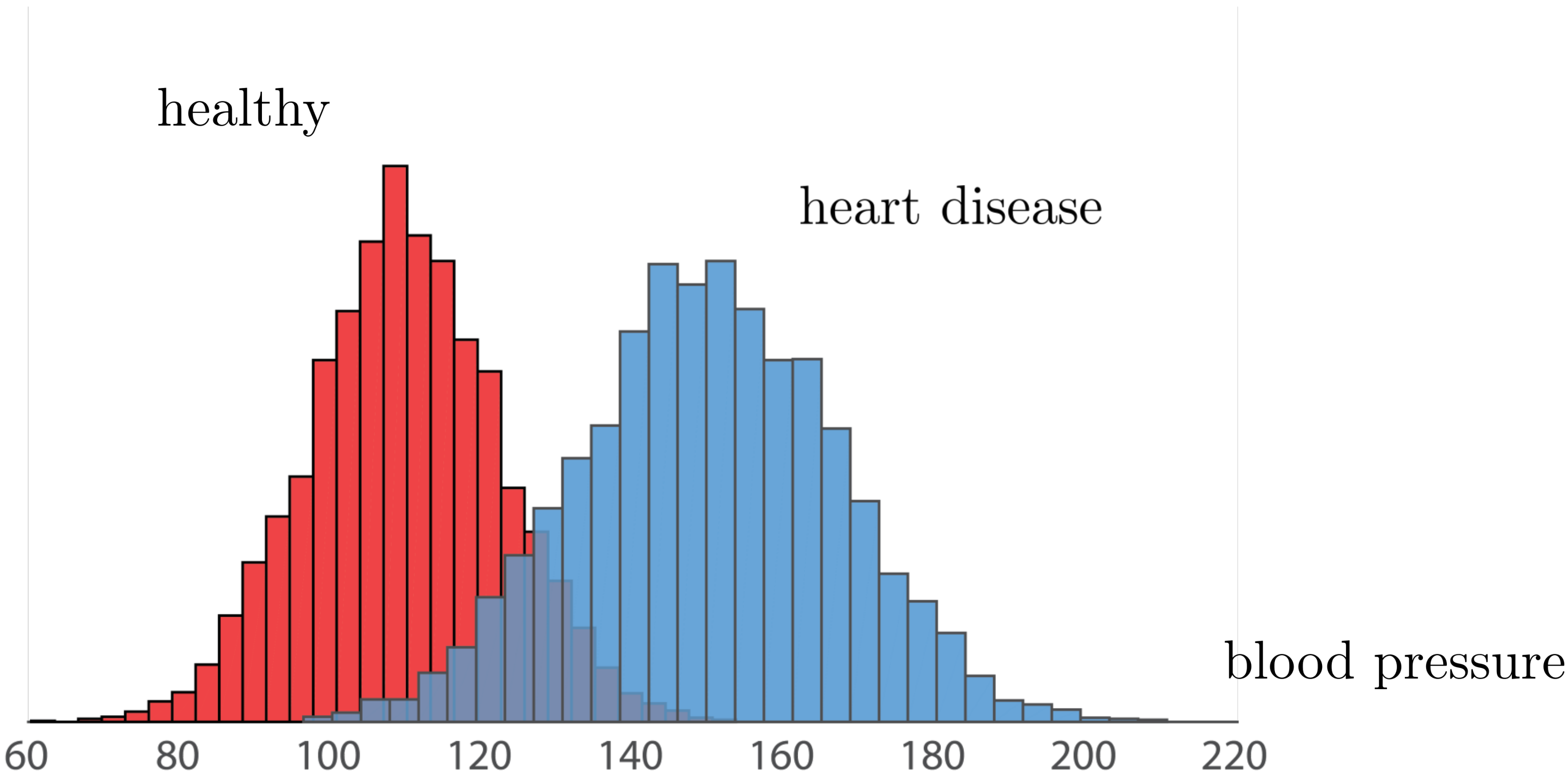
Two Sine Waves



Two Sine Waves + Noise

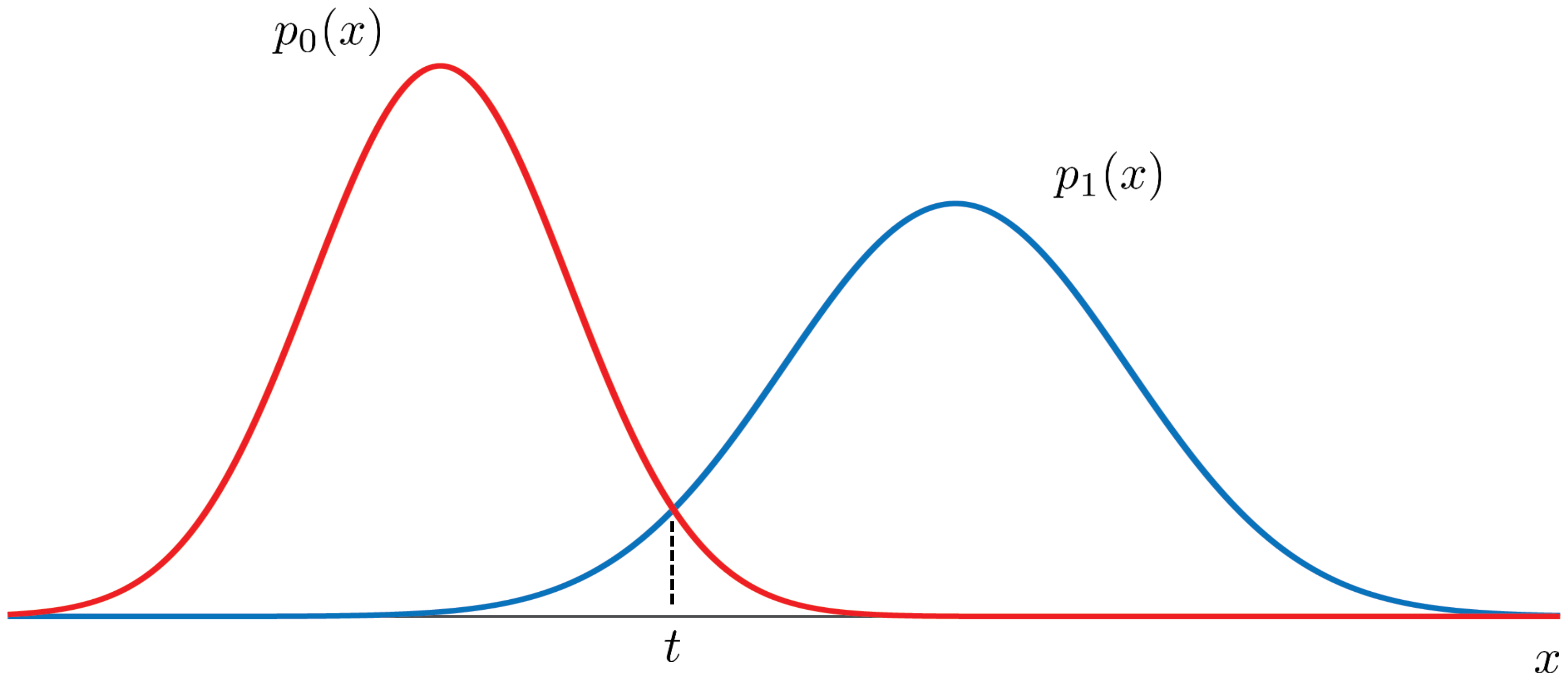
Noise vs. uncertainty

- Distribution overlap is commonly confused with noise
  - noise implies the true value is modified



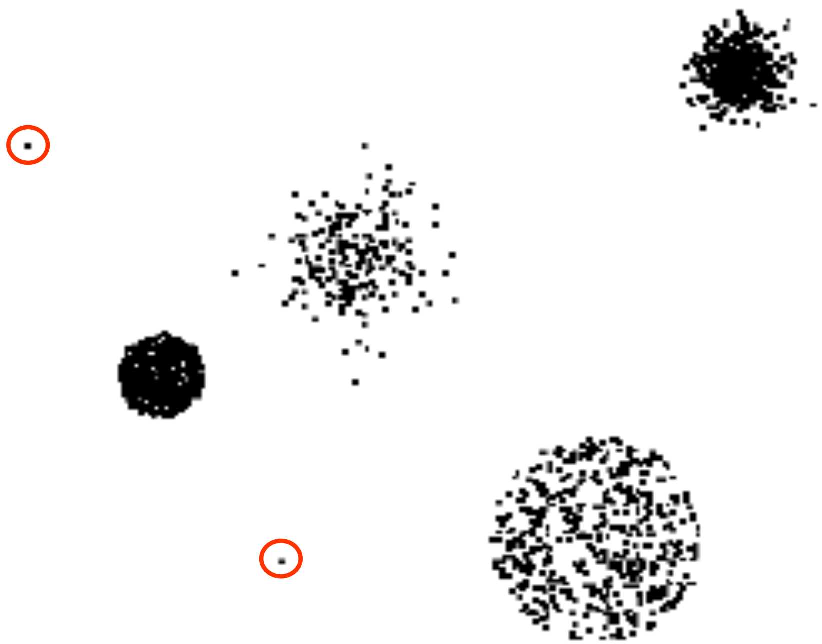
Noise vs. uncertainty

- Distribution overlap is commonly confused with noise
  - noise implies the true value is modified



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)

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)

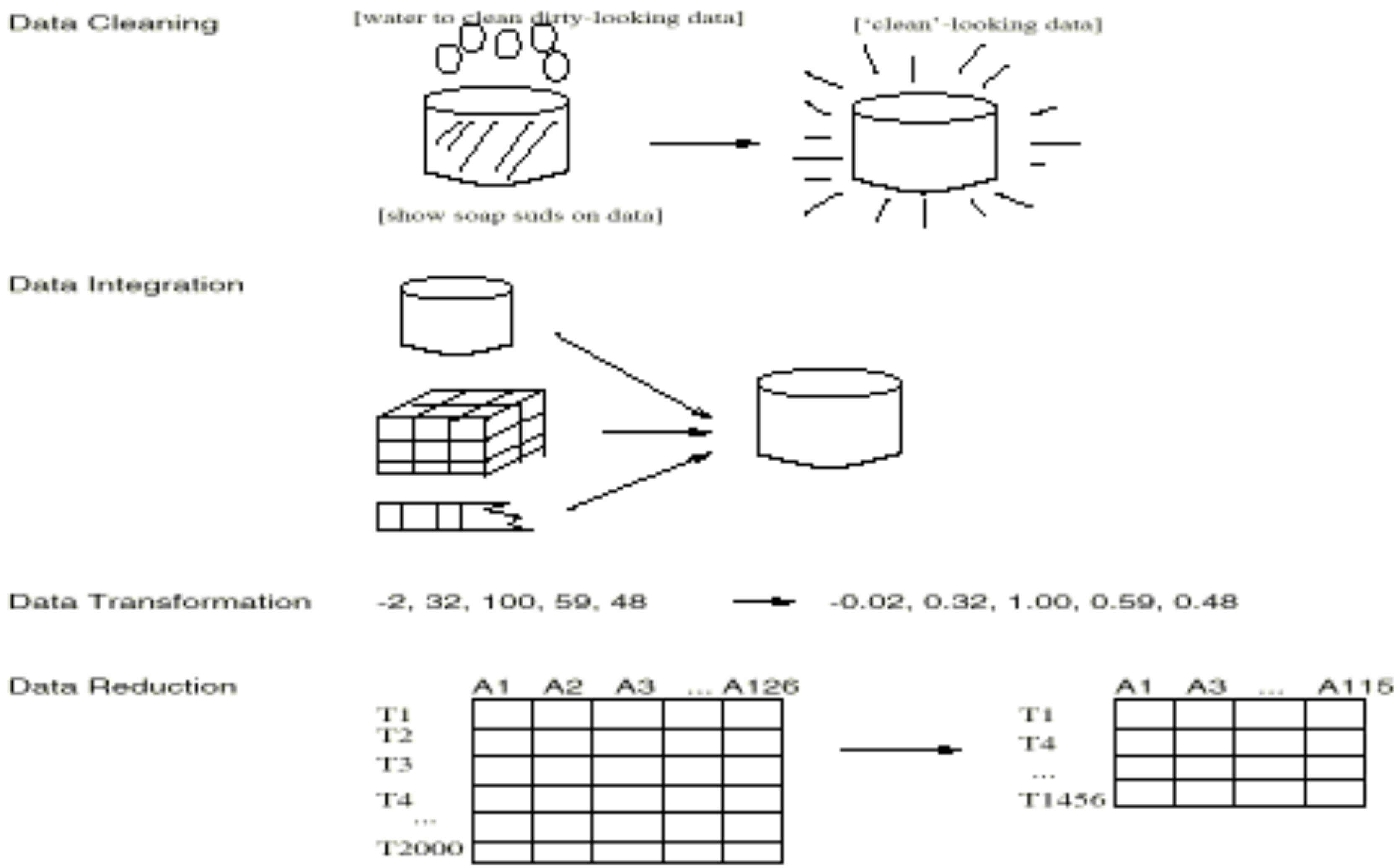
Duplicate Data

- Data set may include data objects that are duplicates, or near-duplicates of one another
  - Major issue when merging data from heterogeous sources

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Duplicates are generally removed

Forms of Data Preprocessing



Data Preprocessing

- Integration
- Data cleaning
- Aggregation
- Sampling
- Dimensionality reduction
- Feature subset selection
- Feature creation
- Discretization and binarization
- Data transformation

Data Cleaning

- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing" — Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"— (Data Catalyst Institute (DCI) survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

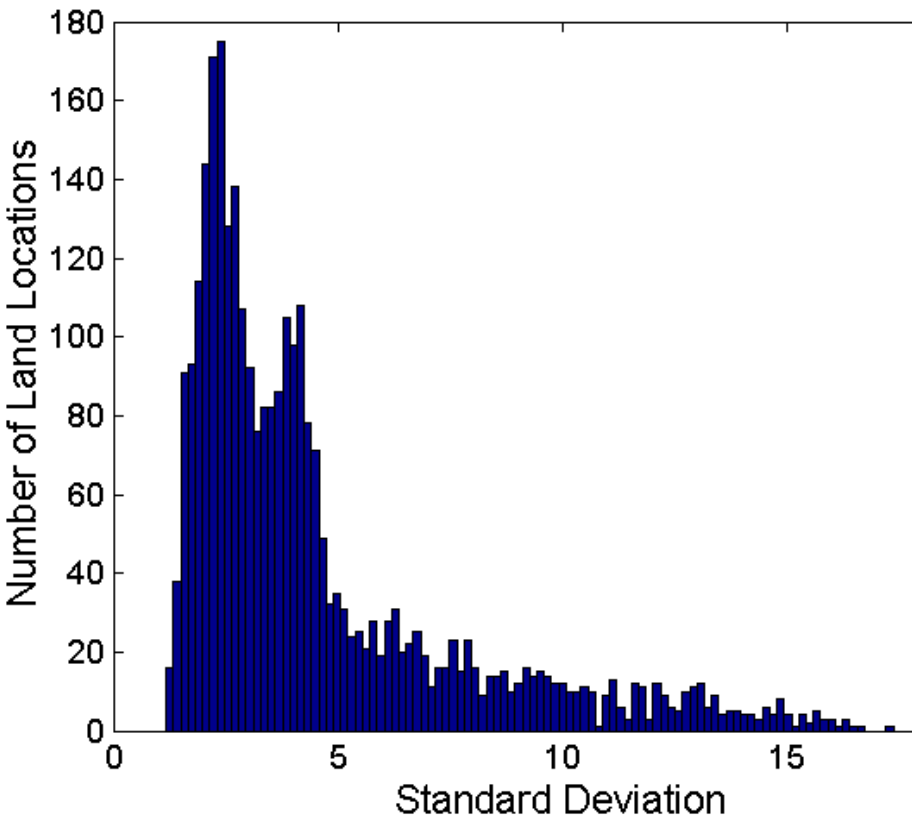
Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

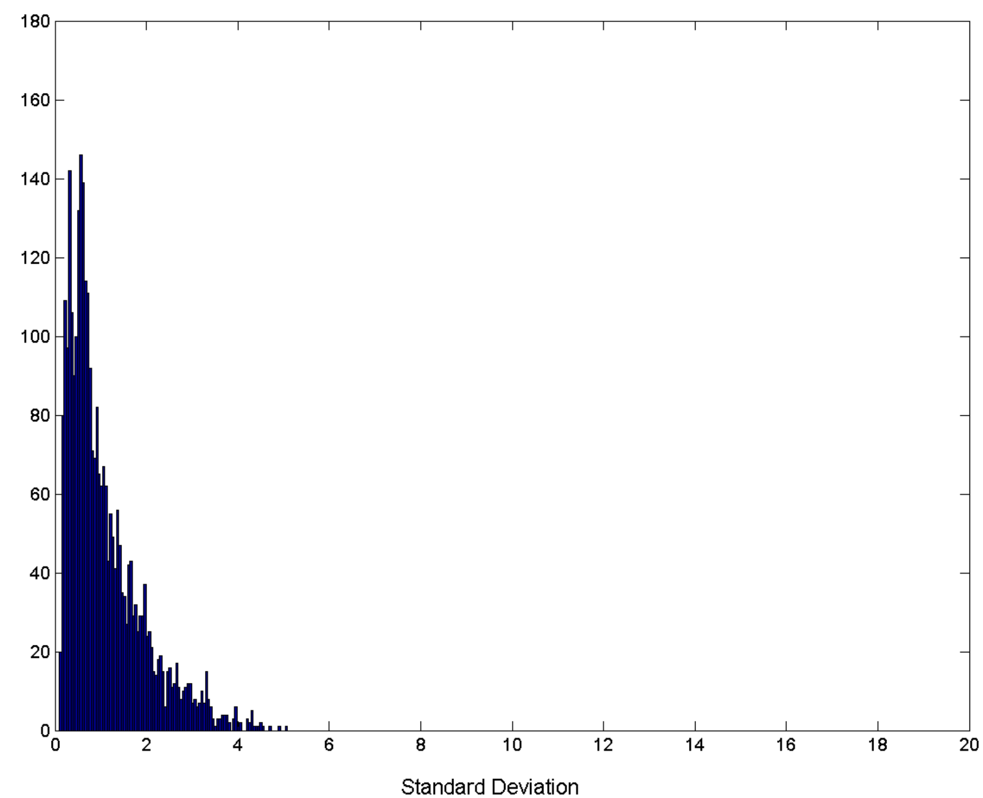
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average
Monthly Precipitation



Standard Deviation of Average
Yearly Precipitation

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id $\equiv$ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Sampling

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

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

- Sampling is 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 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

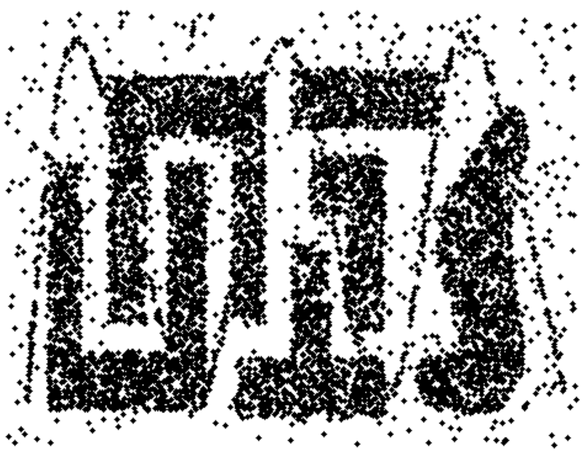
Types of Sampling

- Simple random sampling
  - There is an equal probability of selecting any particular item
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

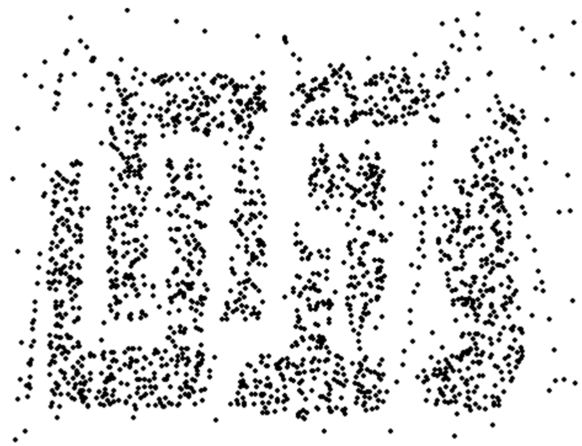
---

- 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.

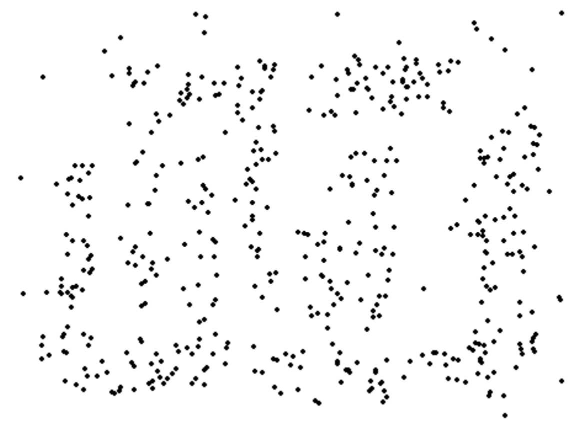
Sample Size



8000 points



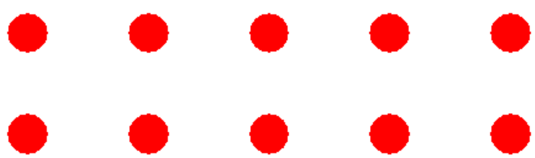
2000 Points

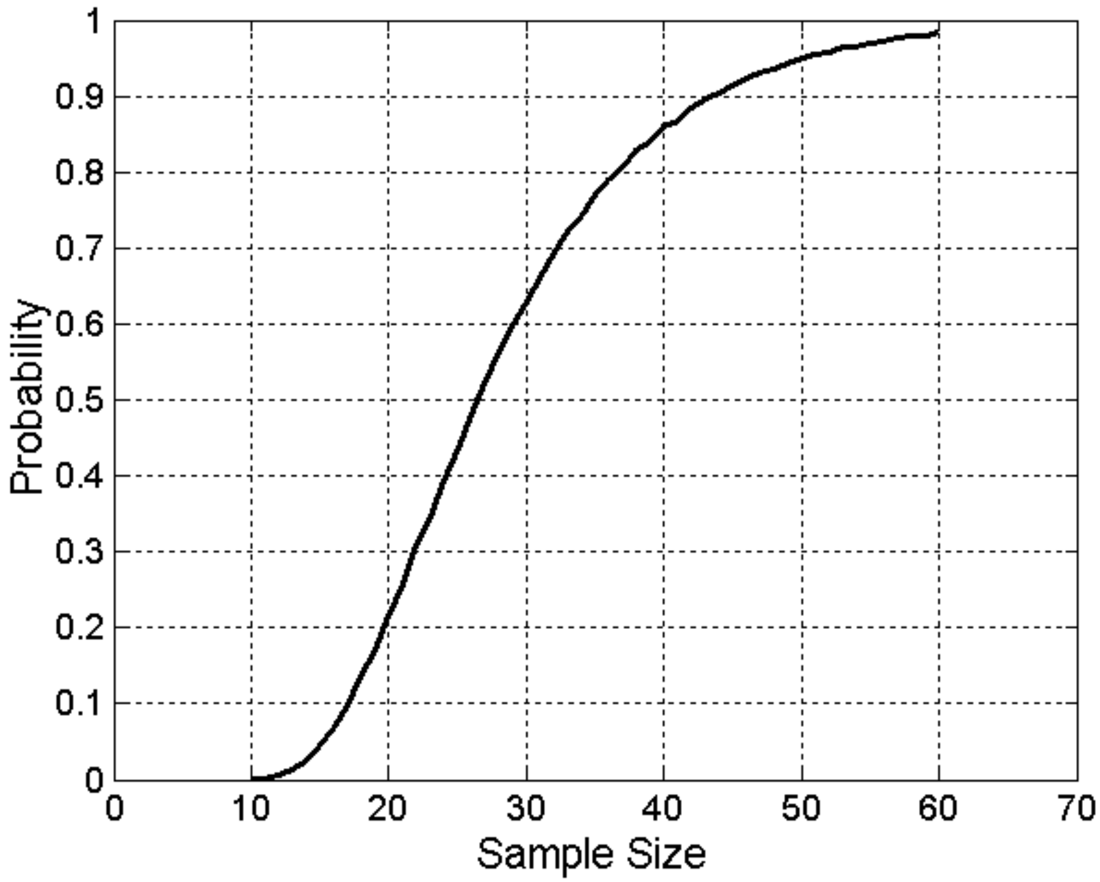


500 Points

Sample Size

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



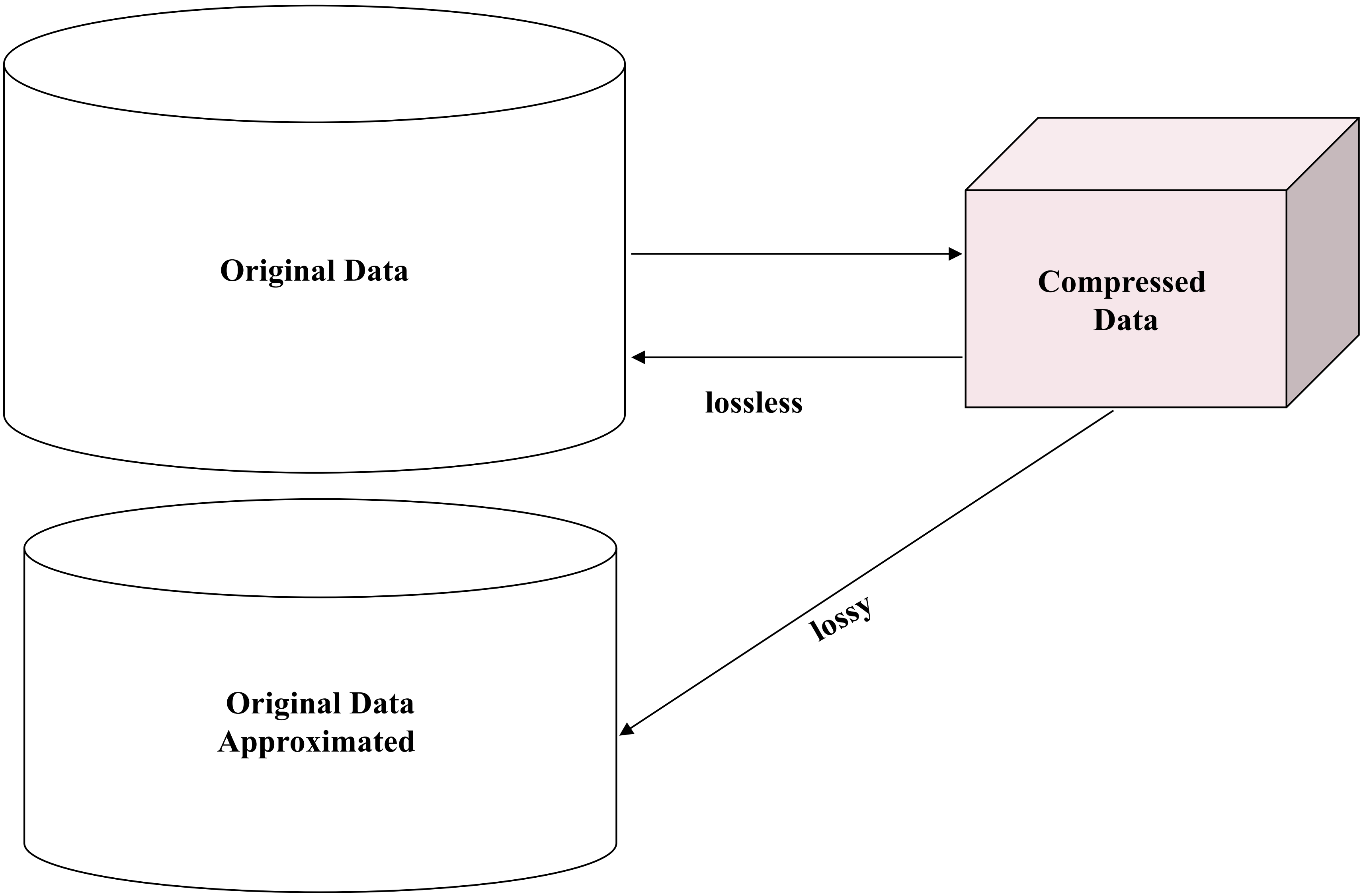


Dimensionality Reduction

- 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
  - may help to avoid stability problems

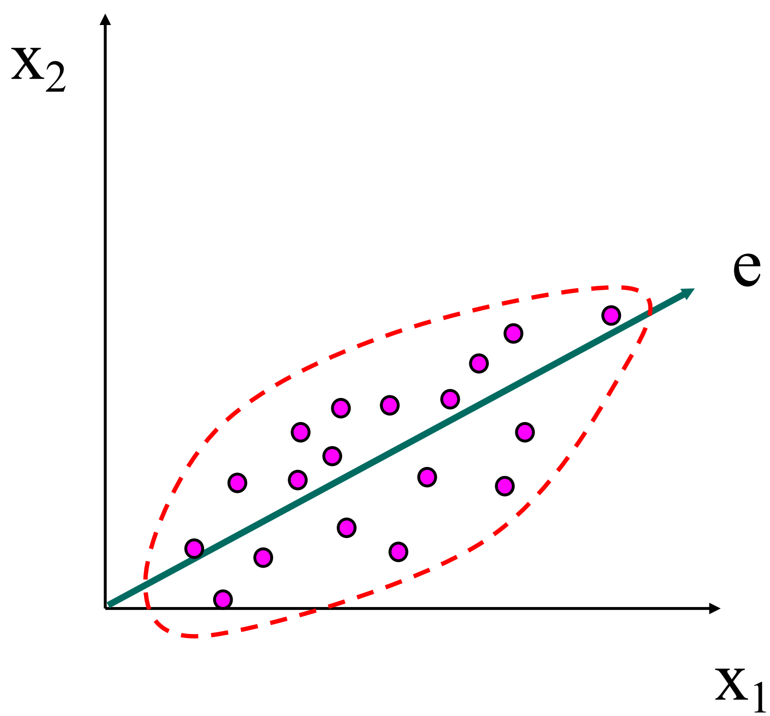
- Techniques
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Others: supervised and non-linear techniques

Relationship to Data Compression



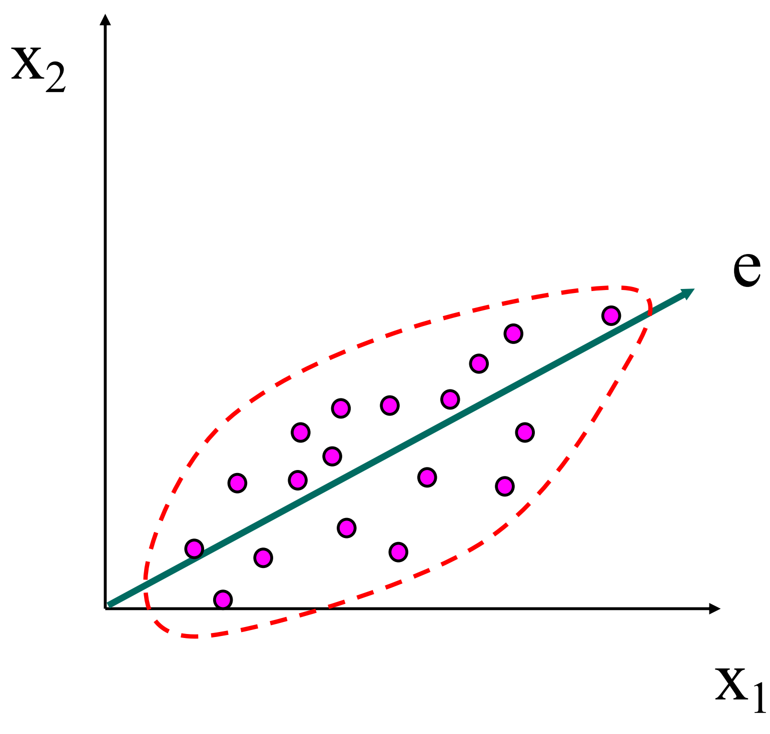
Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



Dimensionality Reduction: PCA

- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space

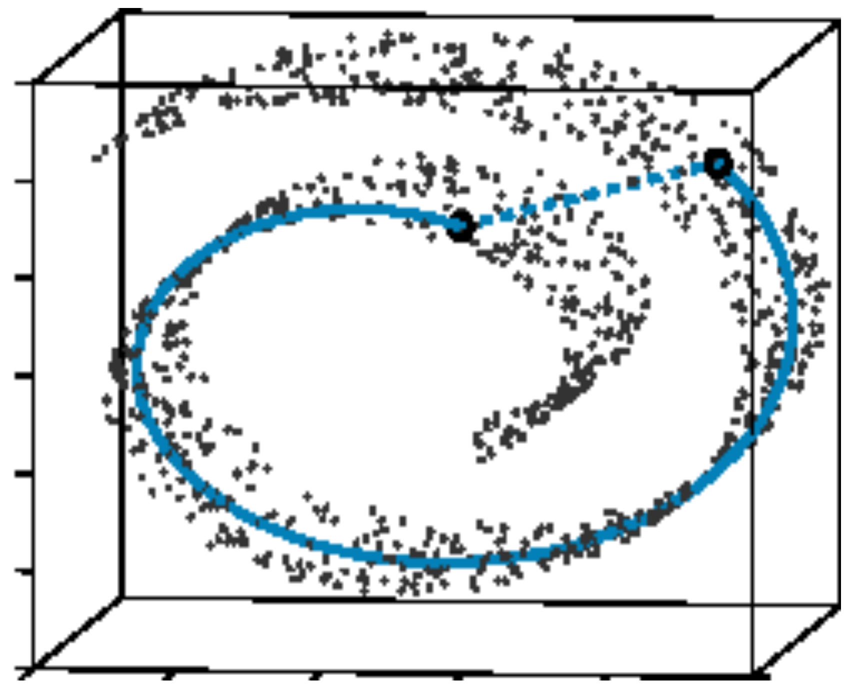


Dimensionality Reduction: PCA

Dimensions = 206

Dimensionality Reduction: ISOMAP

By: Tenenbaum, de Silva, Langford (2000)



- Construct a neighbourhood graph
- For each pair of points in the graph, compute the shortest path distances – geodesic distances

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

Feature Subset Selection

- **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 (usually one pass through data):**
  - Features are selected before data mining algorithm is run
- **Wrapper approaches (usually many passes through data):**
  - Use the data mining algorithm as a black box to find best subset of attributes

	The	Game	Play	Football	Baseball	Brady	Deflate	Gate
Document 1	12	2	3	14		4	4	6
Document 2	18	5	5		3			5
Document 3	24						4	5
Document 4	56	15					24	

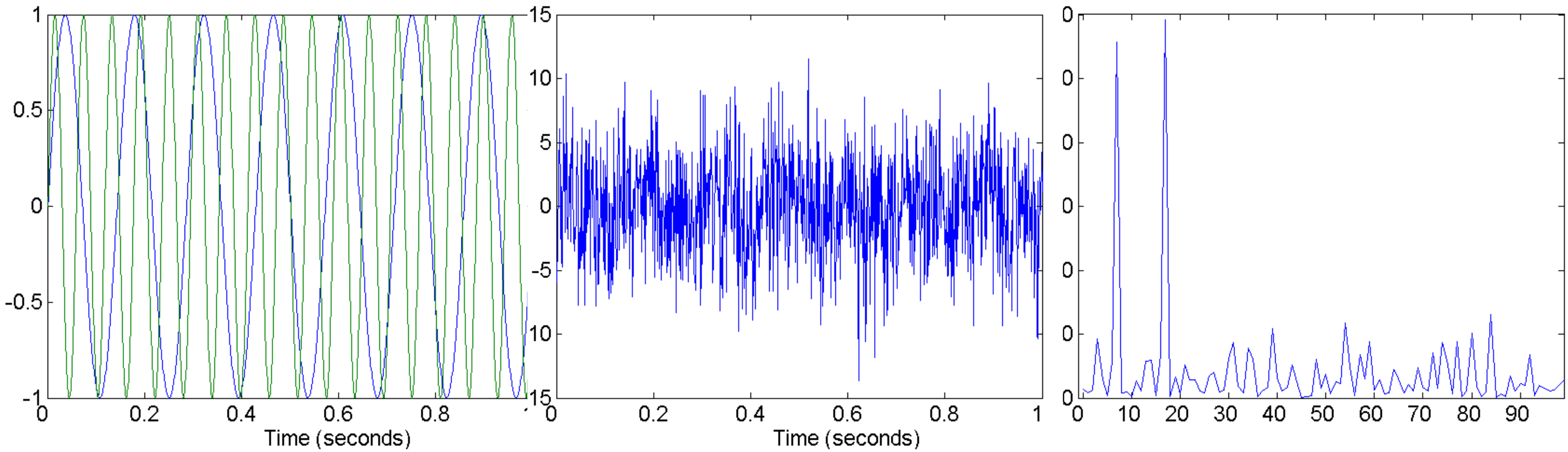
Feature Creation

- 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
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

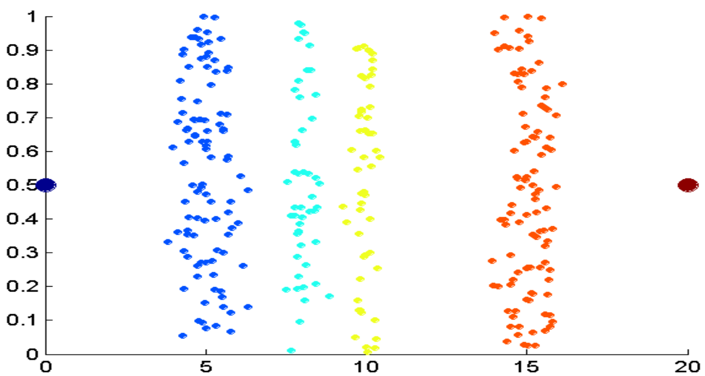
Mapping Data to a New Space

- Fourier transform
- Wavelet transform

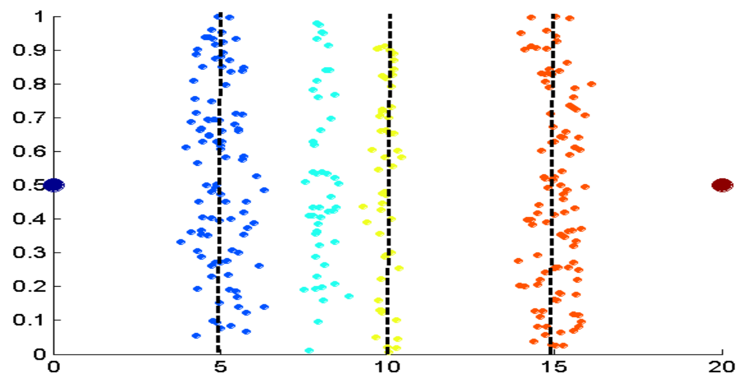


Two Sine Waves **Two Sine Waves + Noise** **Frequency**

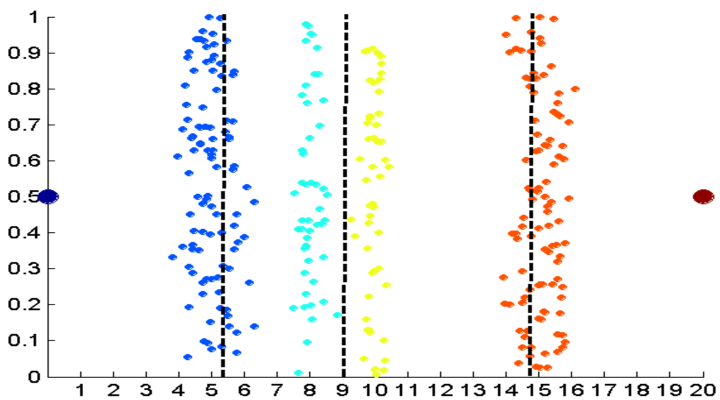
Discretization Without Using Class Labels



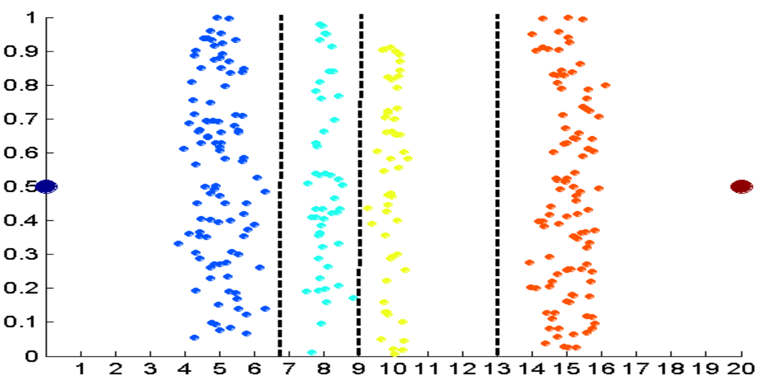
Data



Equal interval width



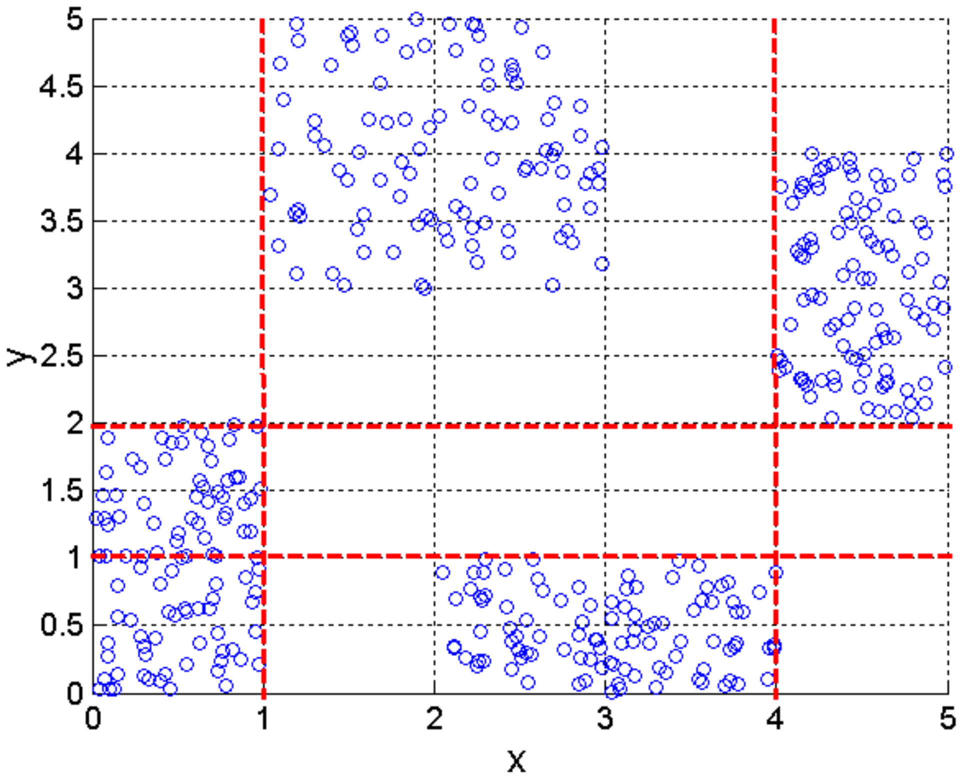
Equal frequency



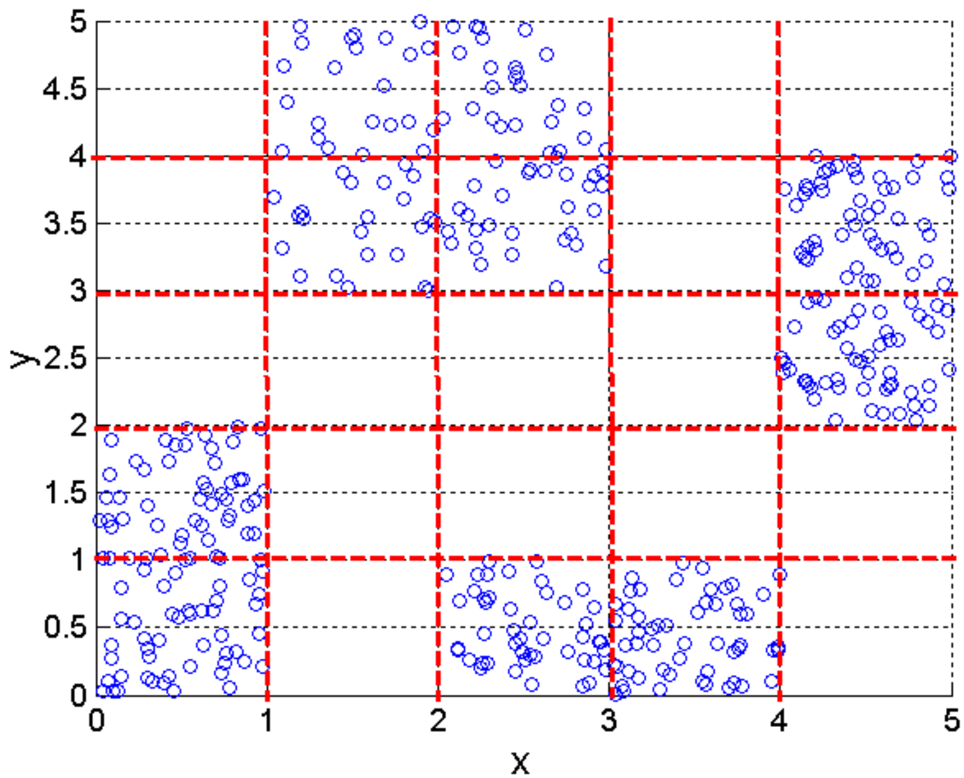
K-means

Discretization Using Class Labels

- Entropy based approach



3 categories for both x and y



5 categories for both x and y

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:

  - **Bin 1**: 4, 8, 9, 15
  - **Bin 2**: 21, 21, 24, 25
  - **Bin 3**: 26, 28, 29, 34

* Smoothing by bin means:

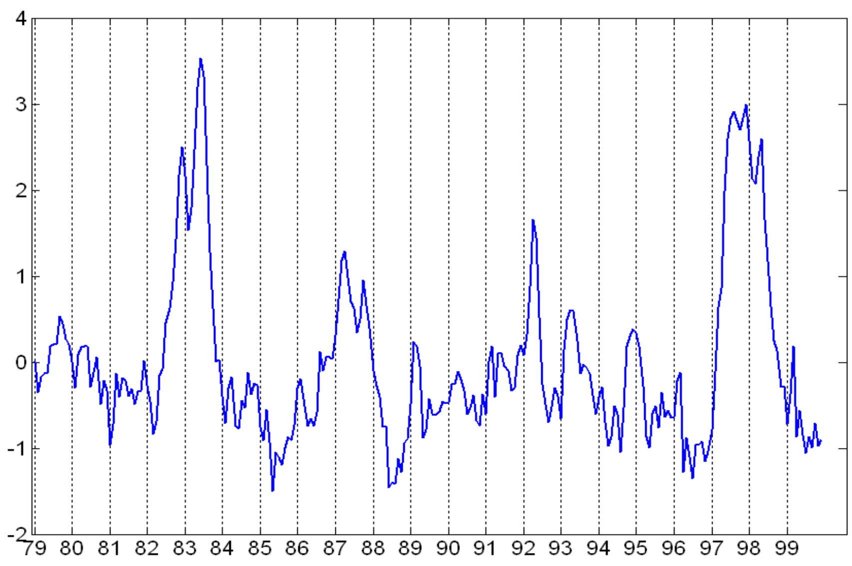
  - **Bin 1**: 9, 9, 9, 9
  - **Bin 2**: 23, 23, 23, 23
  - **Bin 3**: 29, 29, 29, 29

* Smoothing by bin boundaries:

  - **Bin 1**: 4, 4, 4, 15
  - **Bin 2**: 21, 21, 25, 25
  - **Bin 3**: 26, 26, 26, 34

Attribute Transformation

- 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, $|x|$
  - Standardization and normalization



Data Transformation: Normalization

- Min-max normalization: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  – Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- Z-score normalization (μ: mean, σ: standard deviation):

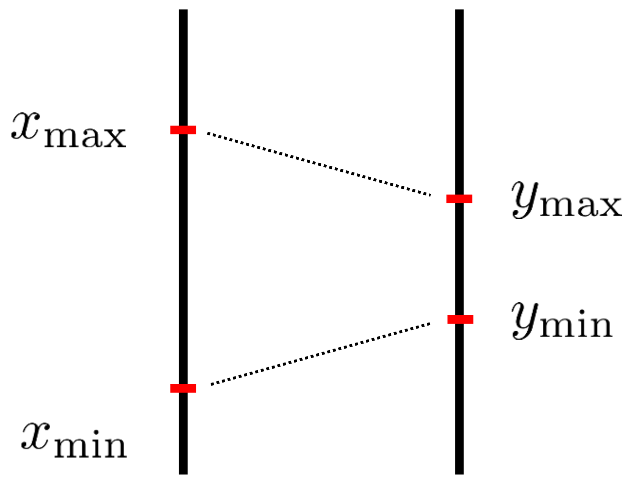
$$v' = \frac{v - \mu_A}{\sigma_A}$$

  – Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|ν'|) < 1

Deriving the Min-Max Normalization



Find linear transform:

$$y = ax + b$$

$$x_{\min} \to y_{\min}$$

$$x_{\max} \to y_{\max}$$

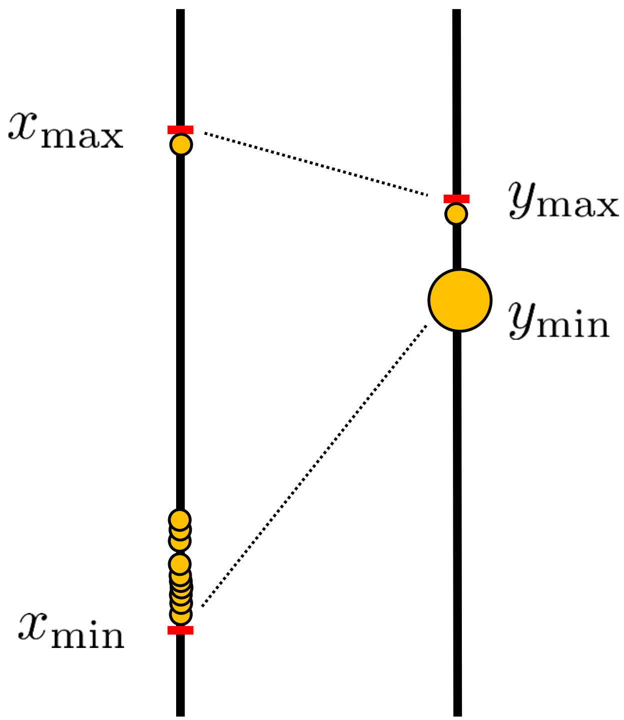
Equation of a line given two points;

i.e., (x_1, y_1) and (x_2, y_2)

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

You take it from here...

Min-max Normalization: Problems?

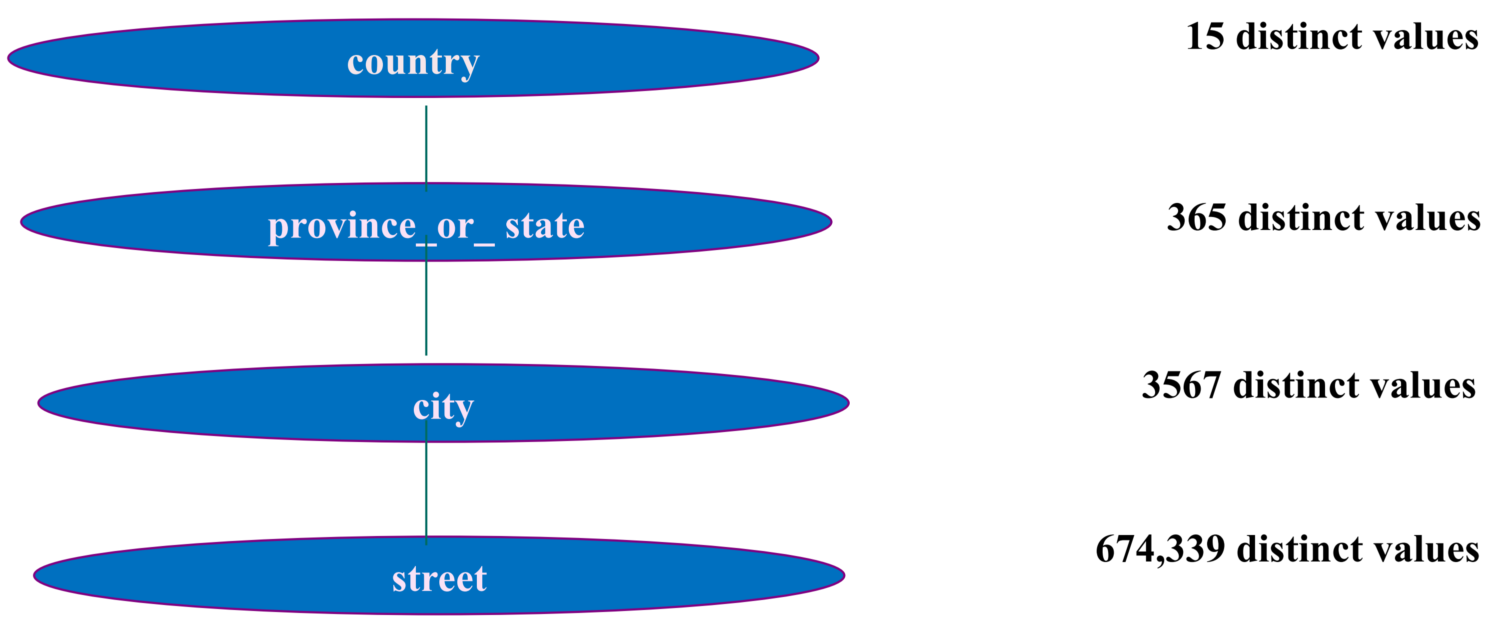


Concept Hierarchy Generation for Categorical Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Boston, Newton, Brookline} < Massachusetts
- Specification of only a partial set of attributes
  - E.g., only street < city, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {street, city, state, country}

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but data preprocessing still an active area of research