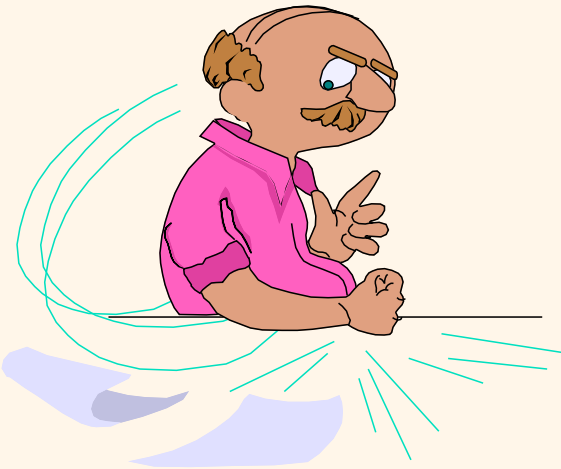


Data Warehousing/Mining

Comp 150 DW

Chapter 3: Data Preprocessing

Instructor: Dan Hebert





Chapter 3: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary



Why Data Preprocessing?

- Data in the real world is dirty
 - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - **noisy**: containing errors or outliers
 - **inconsistent**: containing discrepancies in codes or names
- No quality data, no quality mining results!
 - Quality decisions must be based on quality data
 - Data warehouse needs consistent integration of quality data



Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
 - Accuracy
 - Completeness
 - Consistency
 - Timeliness
 - Believability
 - Value added
 - Interpretability
 - Accessibility
- Broad categories:
 - intrinsic, contextual, representational, and accessibility.

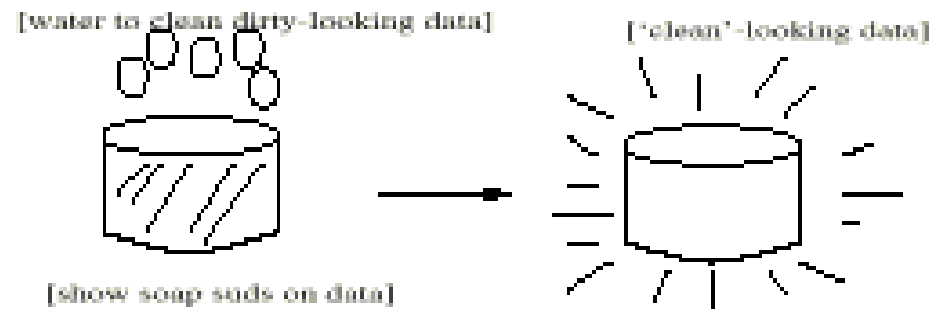


Major Tasks in Data Preprocessing

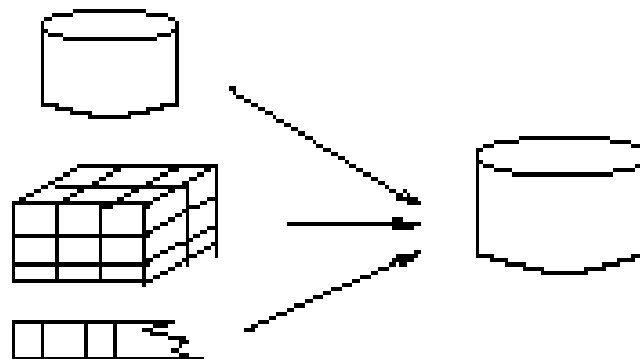
- Data cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
 - Integration of multiple databases, data cubes, or files
- Data transformation
 - Normalization and aggregation
- Data reduction
 - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
 - Part of data reduction but with particular importance, especially for numerical data

Forms of data preprocessing

Data Cleaning



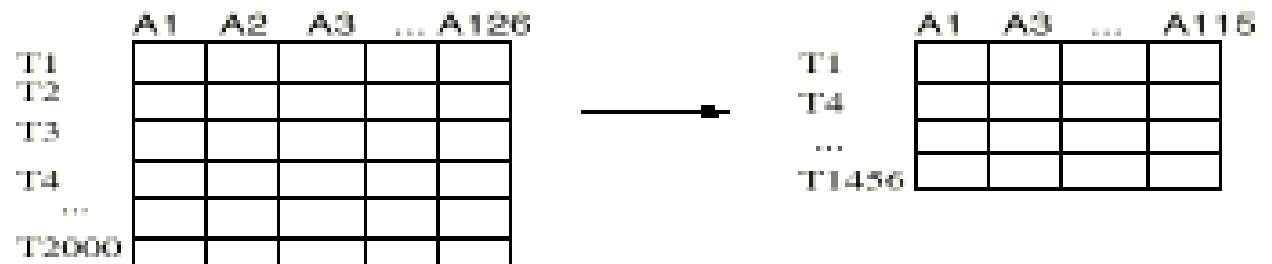
Data Integration

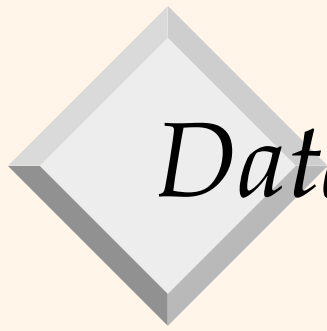


Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48


Data Reduction





Data Cleaning

- Data cleaning tasks
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data



Missing Data

- Data is not always available
 - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
 - equipment malfunction
 - inconsistent with other recorded data and thus deleted
 - data not entered due to misunderstanding
 - certain data may not be considered important at the time of entry
 - not register history or changes of the data
- Missing data may need to be inferred.



Missing Data Example

Bank Acct Totals - Historical


Name	SSN	Address	Phone #	Date	Acct Total
John Doe	111-22-3333	1 Main St Bedford, Ma	111-222-3333	2/12/1999	2200.12
John W. Doe		Bedford, Ma		7/15/2000	12000.54
John Doe	111-22-3333			8/22/2001	2000.33
James Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444	12/22/2002	15333.22
Jim Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444		12333.66
Jim Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444		

How should we handle this?



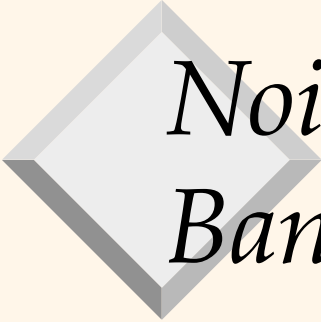
How to Handle Missing Data?

- ❑ 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?
- ❑ Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- ❑ Use the attribute mean to fill in the missing value
- ❑ Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
- ❑ Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree



Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
 - faulty data collection instruments
 - data entry problems
 - data transmission problems
 - technology limitation
 - inconsistency in naming convention
- Other data problems which requires data cleaning
 - duplicate records
 - incomplete data
 - inconsistent data



Noisy Data Example

Bank Acct Totals - Historical

Name	SSN	Address	Phone #	Date	Acct Total
John Doe	111-22-3333	1 Main St Bedford, Ma	111-222-3333	2/12/1999	2200.12
John Doe	111-22-3333	1 Main St Bedford, Ma	111-222-3333	2/12/1999	2233.67
James Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444	12/22/2002	15333.22
James Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444	12/23/2003	15333000.00

How should we handle this?



How to Handle Noisy Data?

- Binning method:
 - first sort data and partition into (equi-depth) bins
 - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
 - detect and remove outliers
- Combined computer and human inspection
 - detect suspicious values and check by human
- Regression
 - smooth by fitting the data into regression functions




Simple Discretization Methods: Binning

- **Equal-width (distance) partitioning:**
 - It 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:**
 - It 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

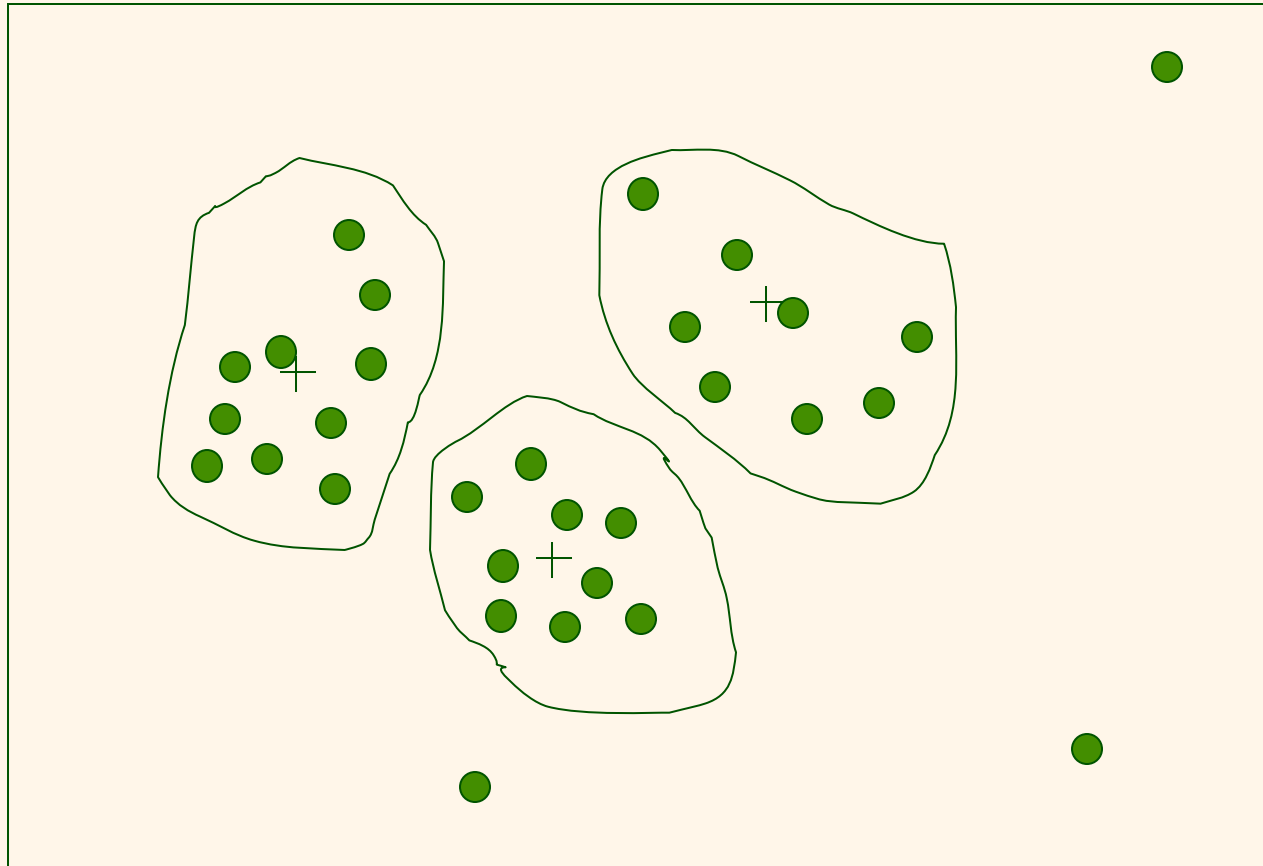
- * Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- * Partition into (equi-width) bins:
 - Bin 1 (4-14): 4, 8, 9
 - Bin 2(15-24): 15, 21, 21, 24
 - Bin 3(25-34): 25, 26, 28, 29, 34
- * Smoothing by bin means:
 - Bin 1: 7, 7, 7
 - Bin 2: 20, 20, 20, 20
 - Bin 3: 28, 28, 28, 28, 28
- * Smoothing by bin boundaries:
 - Bin 1: 4, 4, 4
 - Bin 2: 15, 24, 24, 24
 - Bin 3: 25, 25, 25, 25, 34



Binning Methods for Data Smoothing (continued)

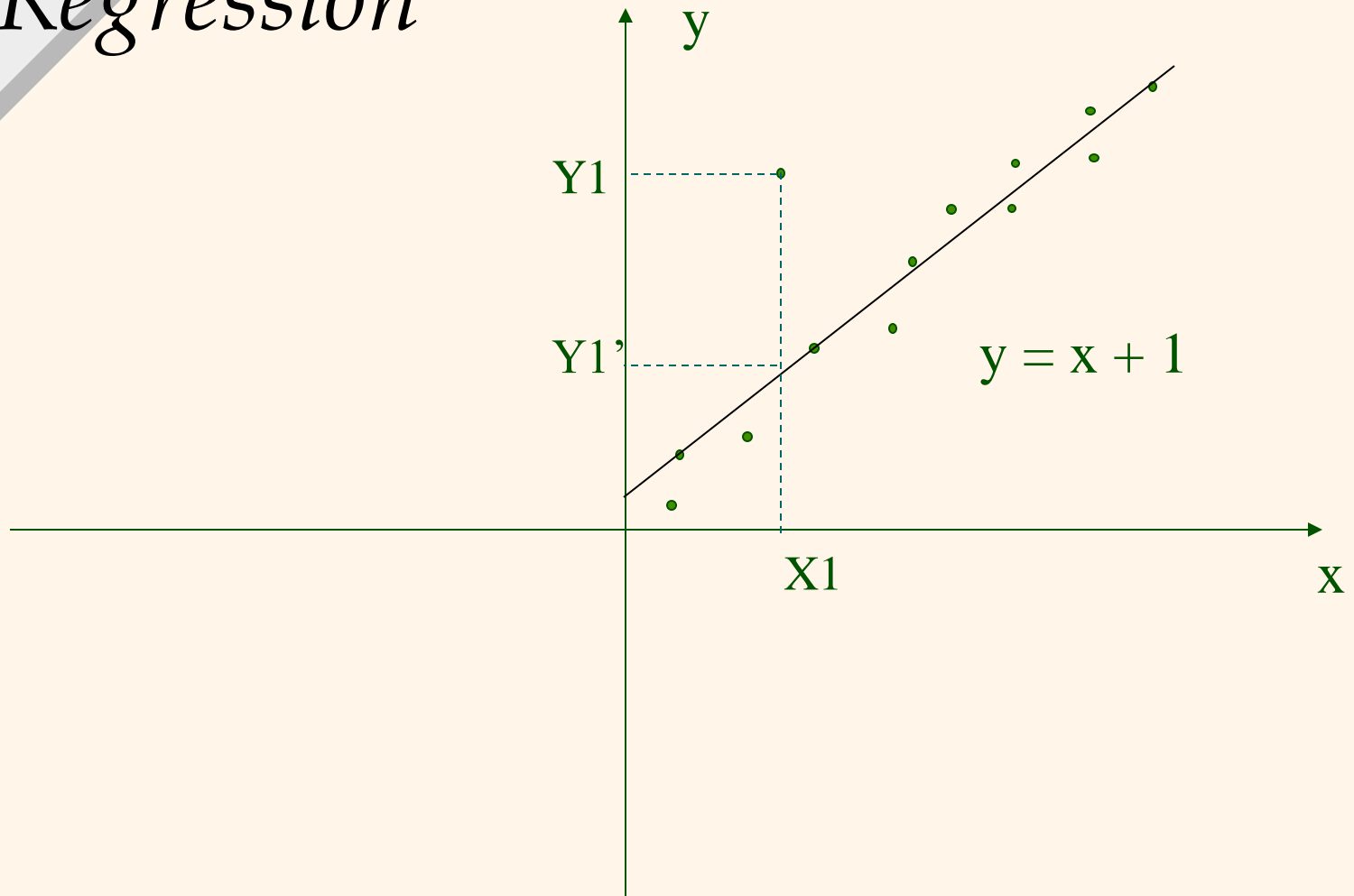
- * Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- * Partition into (equi-depth) bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- * Smoothing by bin means:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- * Smoothing by bin boundaries:
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34

Cluster Analysis

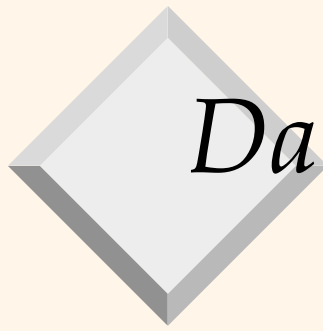


Allows detection and removal of outliers

Regression



Linear regression – find the best line to fit two variables and use regression function to smooth data



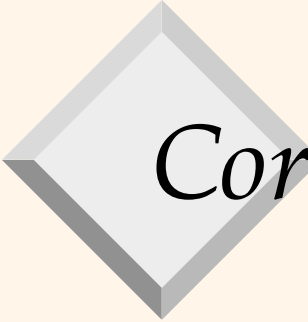
Data Integration

- Data integration:
 - combines data from multiple sources into a coherent store
- Schema integration
 - integrate metadata from different sources
 - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id \equiv B.cust-#
- 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 Redundant Data in Data Integration

- Redundant data occur often when integration of multiple databases
 - The same attribute may have different names in different databases
 - One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlational analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality



Correlational Analysis

$$\square R_{A,B} = \frac{\text{Sum } (A-A') (B-B')}{(n-1) \text{sd}A \text{sd}B}$$


Where A' = mean value of A

$$\frac{\text{sum } (A)}{n}$$

sdA = standard deviation of A

$$\text{SqRoot } \left(\frac{\text{Sum } (A-A')^2}{n-1} \right)$$

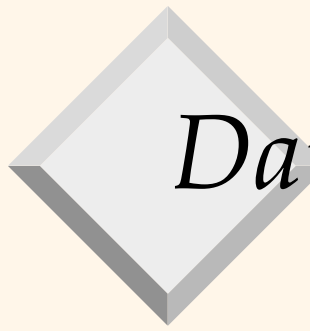
<0 negatively correlated, =0 no correlation, >0 correlated –
consider removal of A or B



Correlational Analysis Example

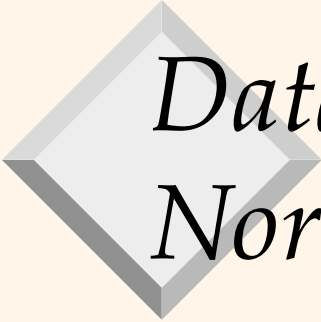
- ❖ A – 2, 5, 6, 8, 22, 33, 44, 55
- ❖ B – 6, 7, 22, 33, 44, 66, 67, 70
- ❖ $A' = 22, B' = 45$
- ❖ $\text{Sum}(A - A') = -1, \text{Sum}(B - B') = -45$
- ❖ $\underline{\text{sd}}A = .378, \underline{\text{sd}}B = 17.008$
- ❖ $R_{A,B} = 45/45.003 = .999$

$R_{A,B} > 0$ - correlated – consider removal of A or B



Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
 - min-max normalization
 - z-score (zero mean) normalization
 - normalization by decimal scaling
- Attribute/feature construction
 - New attributes constructed from the given ones to help in the data mining process



Data Transformation: Normalization

□ min-max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

□ Example – income, min \$55,000, max \$150000 – map to 0.0 – 1.0

□ \$73,600 is transformed to :
– $\frac{73600 - 55000}{150000 - 55000} (1.0 - 0) + 0 = 0.196$



Data Transformation: Normalization

- z-score normalization

$$v' = \frac{v - mean_A}{stand_dev_A}$$

- Example – income, mean \$33000, sd \$11000
- \$73600 is transformed to :
 - $\frac{73600-33000}{11000} = 3.69$




Data Transformation: Normalization

- normalization by decimal scaling

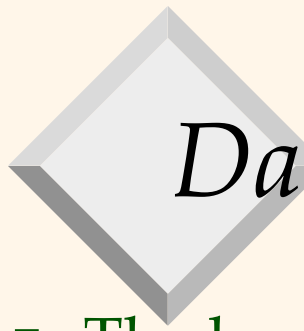
$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|) < 1$$

- Example recorded values - -722 to 821
- Divide each value by 1000
 - -28 normalizes to -.028
 - 444 normalizes to 0.444



Data Reduction Strategies

- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
 - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Data reduction strategies
 - Data cube aggregation
 - Dimensionality reduction
 - Numerosity reduction
 - Discretization and concept hierarchy generation



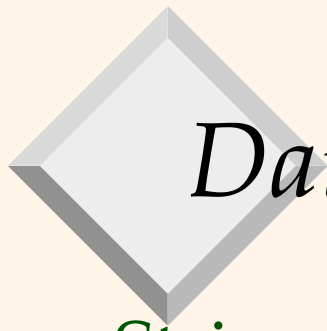
Data Cube Aggregation

- The lowest level of a data cube
 - the aggregated data for an individual entity of interest
 - e.g., a customer in a phone calling data warehouse.
- Multiple levels of aggregation in data cubes
 - Further reduce the size of data to deal with
- Reference appropriate levels
 - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible



Dimensionality Reduction

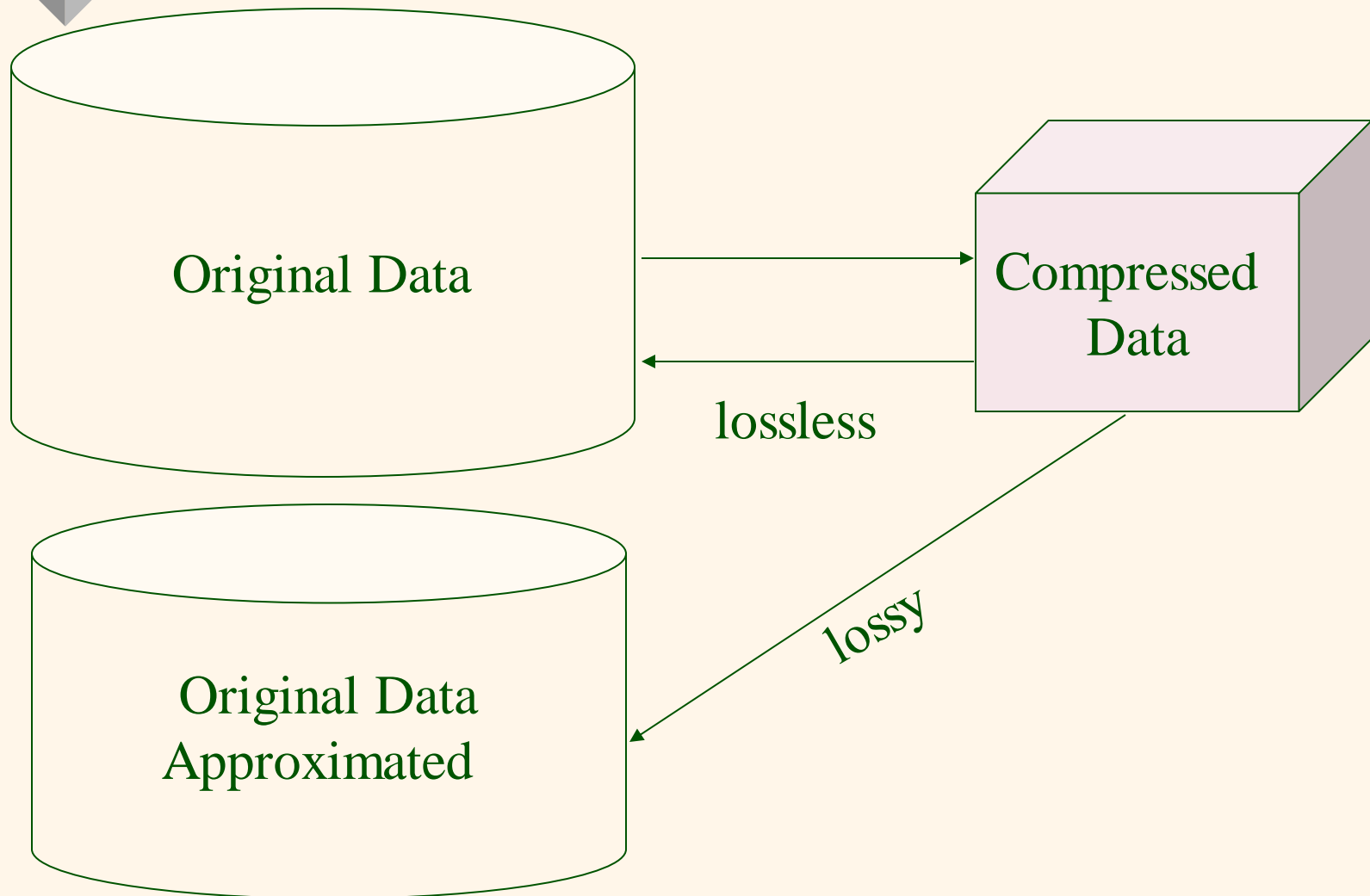
- Feature selection (i.e., attribute subset selection):
 - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
 - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (best & worst attributes determined using various methods: statistical significance, info gain, ... Chpt 5 – more detail):
 - step-wise forward selection
 - step-wise backward elimination
 - combining forward selection and backward elimination
 - decision-tree induction

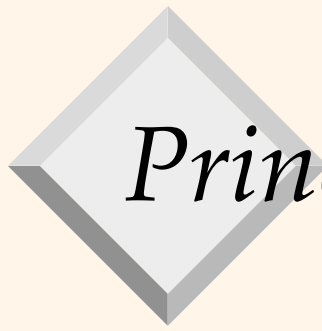


Data Compression

- String compression
 - There are extensive theories and well-tuned algorithms
 - Typically lossless
 - But only limited manipulation is possible without expansion
- Audio/video compression
 - Typically lossy compression, with progressive refinement
 - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
 - Typically short and vary slowly with time

Data Compression





Principal Component Analysis

- Given N data vectors from k -dimensions, find $c \leq k$ orthogonal vectors that can be best used to represent data
 - The original data set is reduced to one consisting of N data vectors on c principal components (reduced dimensions)
- Each data vector is a linear combination of the c principal component vectors
- Works for numeric data only
- Used when the number of dimensions is large



Numerosity Reduction

□ Parametric methods

- Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- Log-linear models: obtain value at a point in m-D space as the product on appropriate marginal subspaces

□ Non-parametric methods

- Do not assume models
- Major families: histograms, clustering, sampling

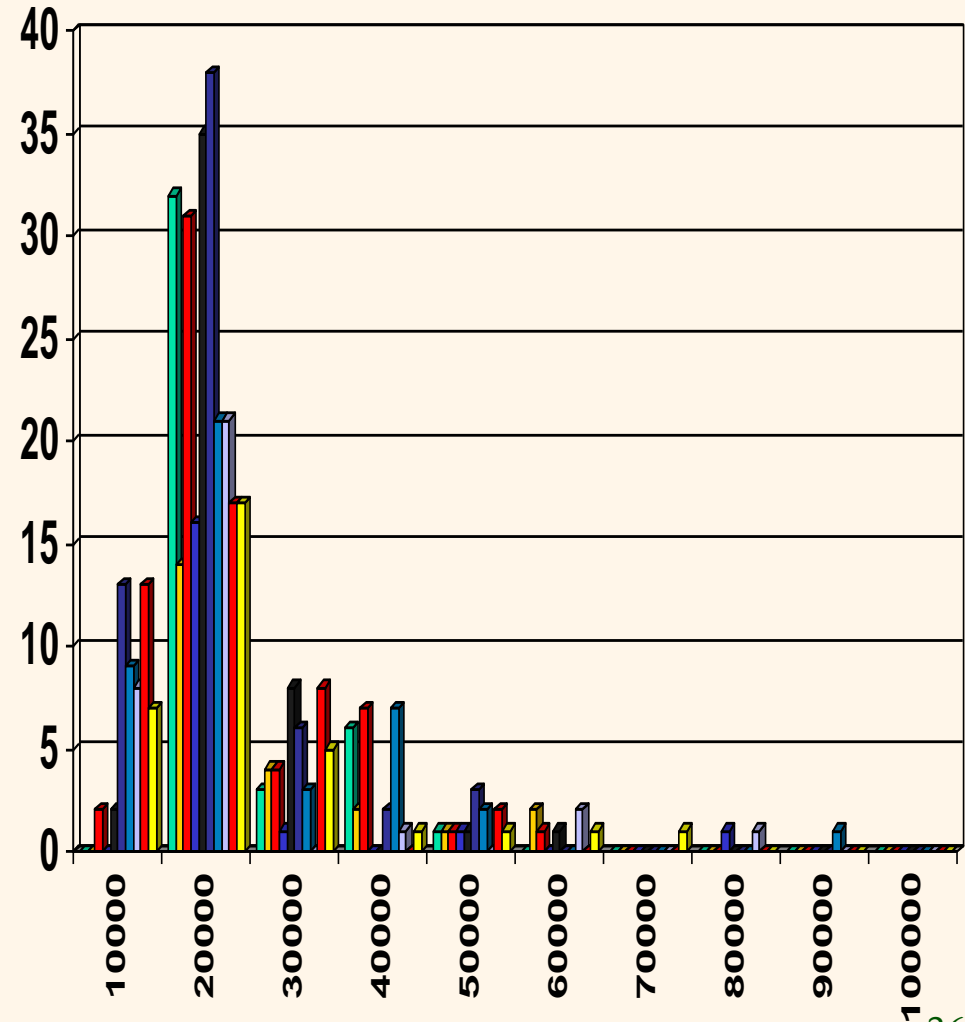


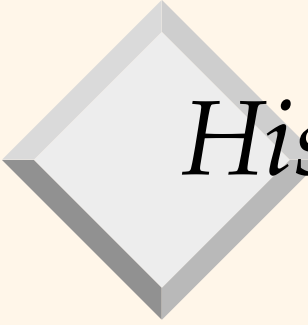
Regression and Log-Linear Models

- Linear regression: Data are modeled to fit a straight line
 - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability distributions

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.





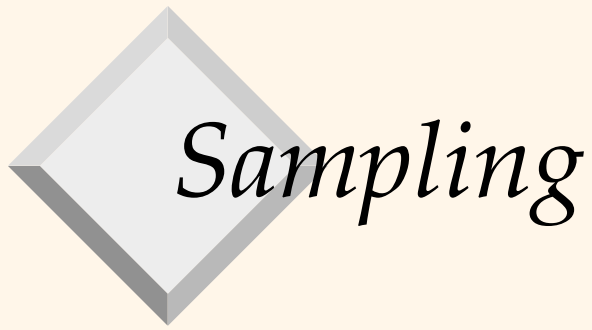
Histograms (continued)

- Several techniques for determining buckets
 - Equiwidth – width of each bucket range is uniform
 - Equidepth – each bucket contains roughly the same number of contiguous samples
 - V-Optimal – weighted sum of the original values that each bucket represents, where bucket weight = number of values in a bucket
 - MaxDiff – bucket boundary is established between each pair for pairs having the $B - 1$ largest differences, where B is user defined
- V-Optimal & MaxDiff most accurate and practical



Clustering

- Partition data set into clusters, and one can store cluster representation only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms, further detailed in Chapter 8

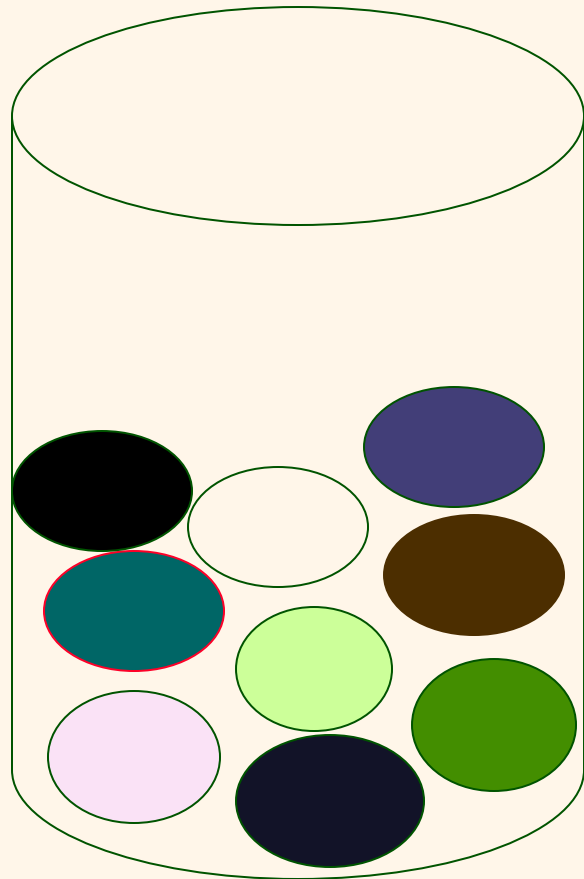


Sampling

- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a **representative** subset of the data
 - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
 - Stratified sampling:
 - Approximate the percentage of each class (or subpopulation of interest) in the overall database
 - Used in conjunction with skewed data
- Sampling may not reduce database I/Os (page at a time).

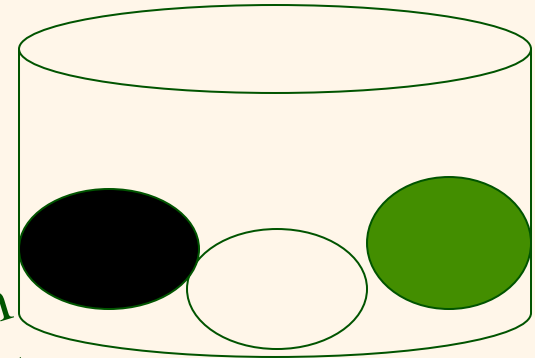
Sampling (continued)

All tuples have equal probability of selection



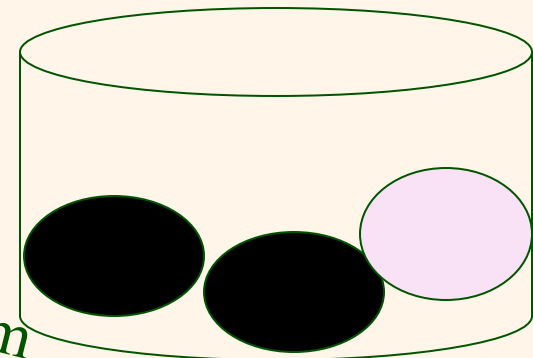
Raw Data

SRSWOR
(simple random
sample without
replacement)



Once selected, can't be
selected again

SRSWR
(simple random
sample with
replacement)

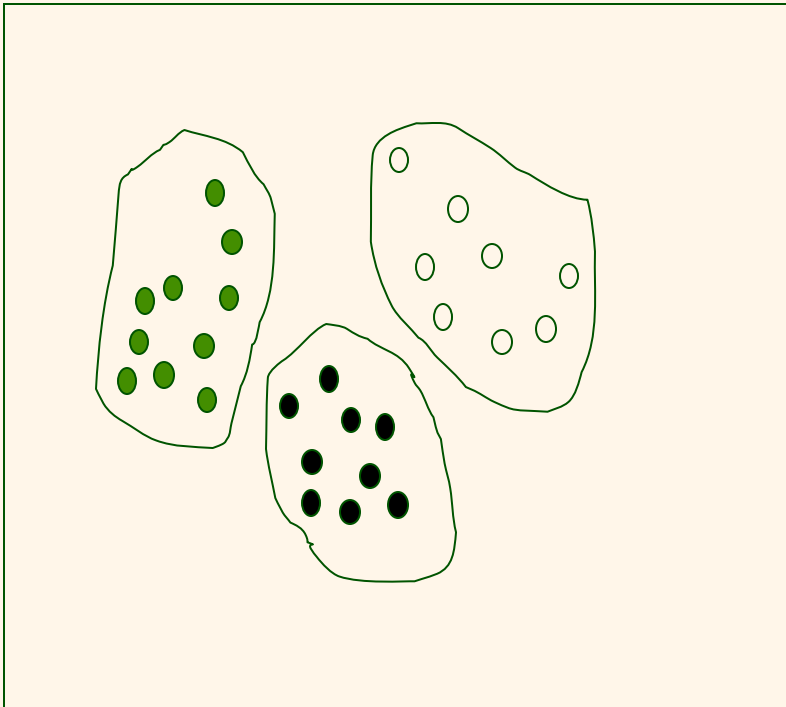


Once selected, can be
selected again

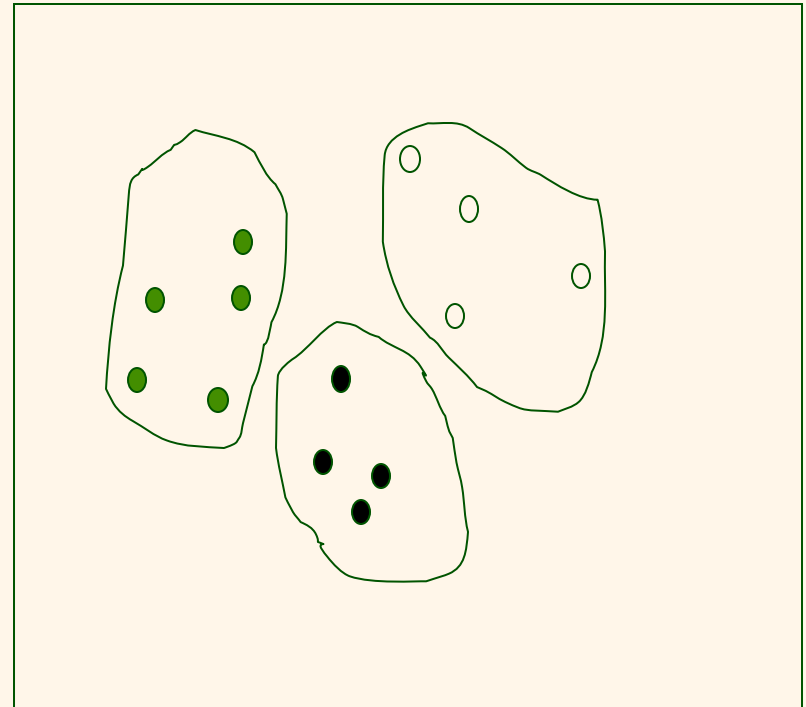
Sampling (continued)

If data is clustered or stratified, performed a simple random sample (with or without replacement) in each cluster or strata

Raw Data



Cluster/Stratified Sample





Hierarchical Reduction

- Use multi-resolution structure with different degrees of reduction
- Hierarchical clustering is often performed but tends to define partitions of data sets rather than “clusters”
- Parametric methods are usually not amenable to hierarchical representation
- Hierarchical aggregation
 - An index tree hierarchically divides a data set into partitions by value range of some attributes
 - Each partition can be considered as a bucket
 - Thus an index tree with aggregates stored at each node is a hierarchical histogram



Discretization

- Three types of attributes:
 - Nominal — values from an unordered set
 - Ordinal — values from an ordered set
 - Continuous — real numbers
- Discretization:
 - divide the range of a continuous attribute into intervals
 - Some classification algorithms only accept categorical attributes.
 - Reduce data size by discretization
 - Prepare for further analysis




Discretization and Concept hierarchy

□ Discretization

- reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.

□ Concept hierarchies

- reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).



Discretization and concept hierarchy generation for numeric data

- Binning (see sections before)
- Histogram analysis (see sections before)
- Clustering analysis (see sections before)
- Entropy-based discretization
- Segmentation by natural partitioning



Concept hierarchy generation for categorical data

- Specification of a partial ordering of attributes explicitly at the schema level by users or experts
 - Example : rel db may contain: street, city, province_or_state, country
 - Expert defines ordering of hierarchy such as street < city < province_or_state < country
- Specification of a portion of a hierarchy by explicit data grouping
 - Example : province_or_state, country : {Alberta, Saskatchewan, Manitoba} – prairies_Canada & {British Columbia, prairies_Canada} – Western Canada



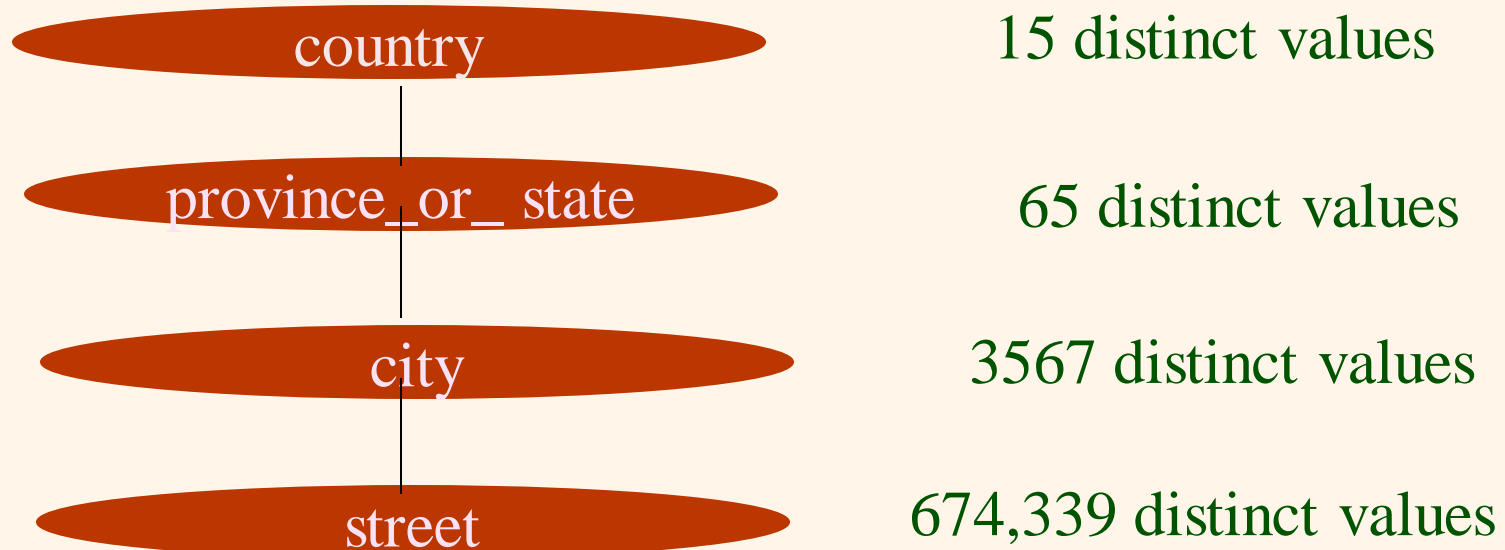
Concept hierarchy generation for categorical data

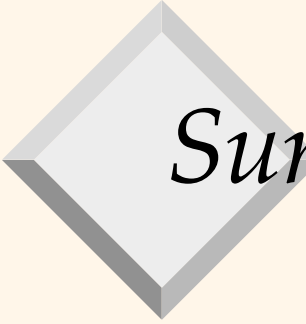
- Specification of a set of attributes, but not of their partial ordering
 - Auto generate the attribute ordering based upon observation that attribute defining a high level concept has a smaller # of distinct values than an attribute defining a lower level concept
 - Example : country (15), state_or_province (365), city (3567), street (674,339)
- Specification of only a partial set of attributes
 - Try and parse database schema to determine complete hierarchy



Specification of a set of attributes

Concept hierarchy can be automatically generated based on the number of distinct values per attribute in the given attribute set. The attribute with the most distinct values is placed at the lowest level of the hierarchy.





Summary

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