STATISTICAL METHODS IN DATA MINING

Dr. Alper VAHAPLAR

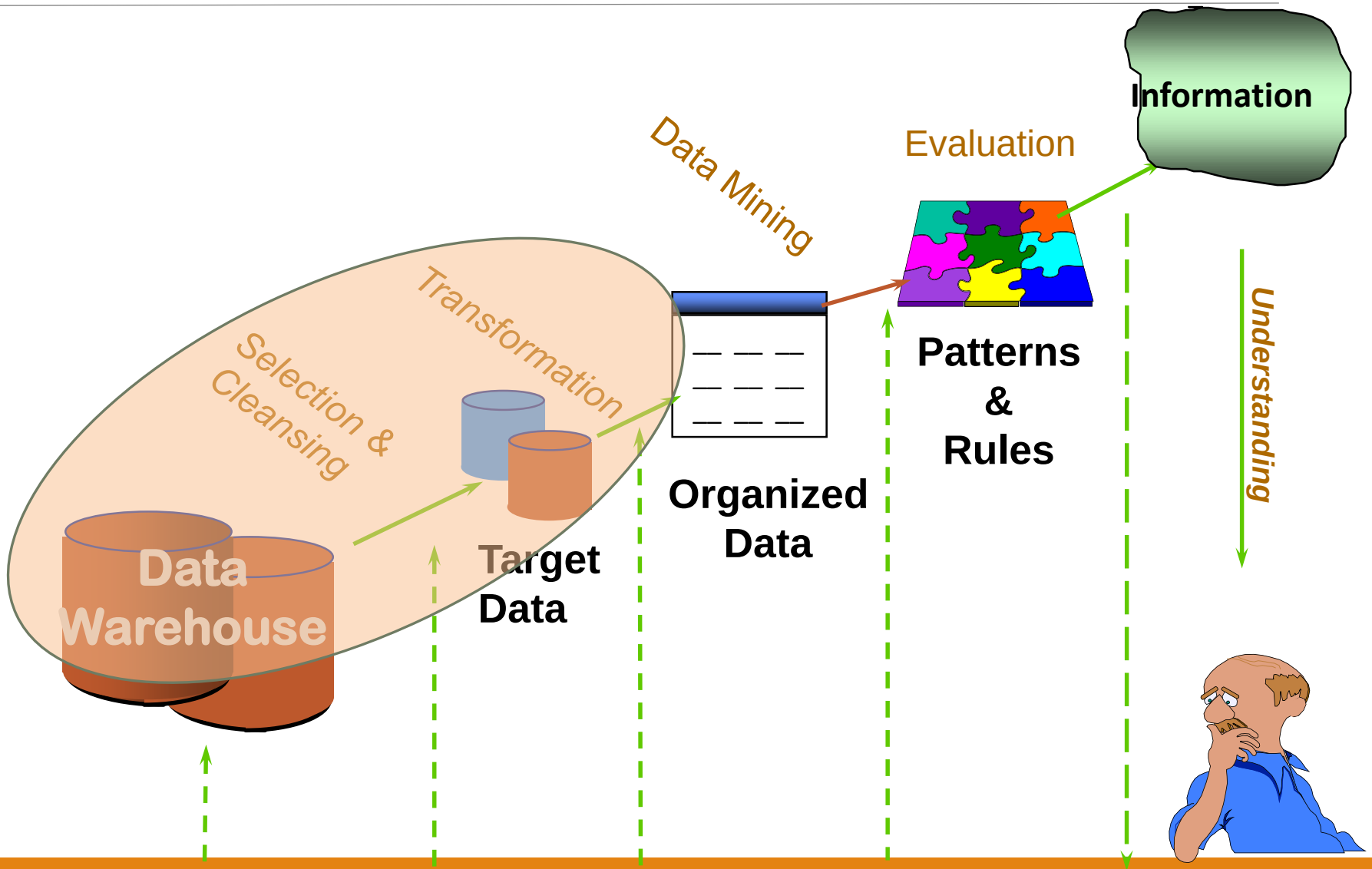
Data Understanding

- ✓Data Preparation,
- ✓Data Understanding,
- ✓Data Visualization,
- ✓IBM SPSS Modeler

References:

- ✓*Han, J. , Kamber, M., Pei, J., (2011). Data Mining: Concepts and Techniques.*
- ✓*Larose, Daniel T. (2005). Discovering Knowledge In Data – An Introduction to Data Mining.*
- ✓*Tan, P., Steinbach, M., Kumar, v. (2006) Introduction to Data Mining.*
- ✓*Bramer, M., (2007) Principles of Data Mining.*
- ✓*Birant, D. Lecture Notes (2012).*

Knowledge Discovery in Databases



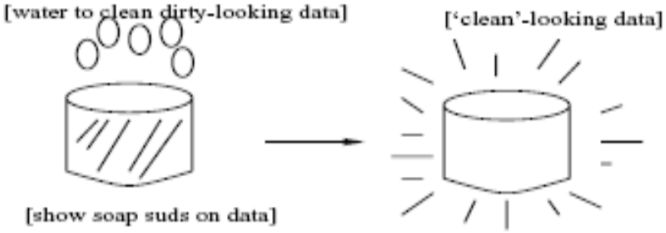
Data Preprocessing

- ✓ Why to pre-process data?
  - Data may be in different location/media,
  - Fields that are obsolete or redundant
  - Missing values
  - Outliers
  - Data in a form not suitable for data mining models
  - Values not consistent with policy or common sense
- ✓ The aim is to *minimize GIGO*
  - *Garbage In Garbage Out*

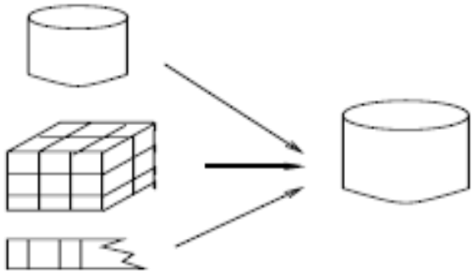
Data Preprocessing

- ✓Data Integration, Selection
- ✓Data Cleaning,
- ✓Data Transformation,
- ✓Data Reduction

Data Cleaning



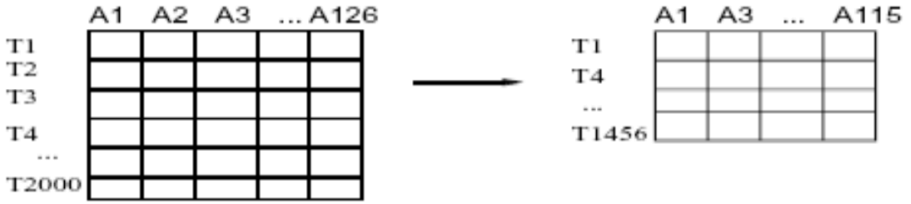
Data Integration



Data Transformation

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

Data Reduction



Data Integration

✓Data Integration

- Obtain and collect data from various sources
  - Ex: Outside or inside of the company
- Combine data from multiple sources into a coherent store
  - Printed Forms, (bills, reports, etc.)
  - Tables
    - Excel Tables, SPSS, Minitab, MATLAB, etc…
    - Database Tables (Access, SQL Server, Oracle ...)
  - Files (unstructured or structured files)
    - XML Files
  - Web Pages
  - Data Cubes, Data Marts

Data Integration

- ✓ Possible Problems:
  - Same person, different spellings
    - e.g. Hasan Hüseyin, H. Hüseyin, H. H., Hasan H., Haso, Hüso, ...
    - e.g.Dokuz Eylül Üniversitesi, 9 Eylül Üniversitesi, DEÜ, DEU ..
  - Same person, different addresses
    - Kaynaklar Kampüsü, Tınaztepe Yerleşkesi
  - Homonyms (same name for different entities)
    - e.g. (Öğrenci.)no, (SGK.)no
    - e.g. Unit incompatibilities (US Dollar, Canada Dollar)
  - Synonyms (different names for the same entities)
    - e.g. İçel, Mersin

Data Integration

- ✓ Possible Problems:
  - Different metrics
    - e.g., cm^3, m^3
  - Schema errors
    - e.g., A.cust-id ≡ B.cust-no
    - e.g. TCK, KimlikNo, TCKimlik
  - Redundancy
    - e.g. Year of birth, age
    - e.g. Annual income – salary,
    - Can be detected by correlation analysis

$$r_{A,B} = \frac{\Sigma(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B}$$

$$\overline{A} = \frac{\Sigma A}{n}$$

$$\sigma_A = \sqrt{\frac{\Sigma(A-\overline{A})^2}{(n-1)}}$$

Data Selection

✓Data Selection

- Database may store terabytes of data
  - Complex data analysis / mining
  - May take a very long time to run on the complete data set
- Selecting a target dataset
- Obtains a reduced representation of the data set

  (smaller in volume but yet produces the same (or almost the same) results)

✓How to select?

- Feature Subset Selection
- Sampling

Data Selection

- ✓ Feature Subset Selection
  - Selecting related attributes
  - Eliminating redundant fields
- ✓ Sampling
  - 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.
  - 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

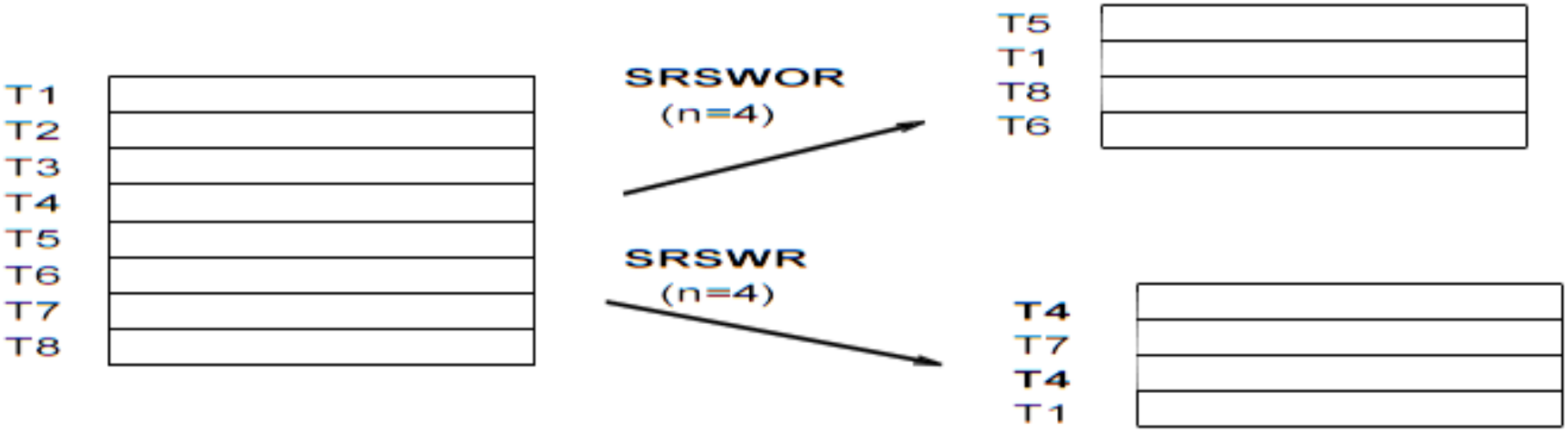
Data Selection

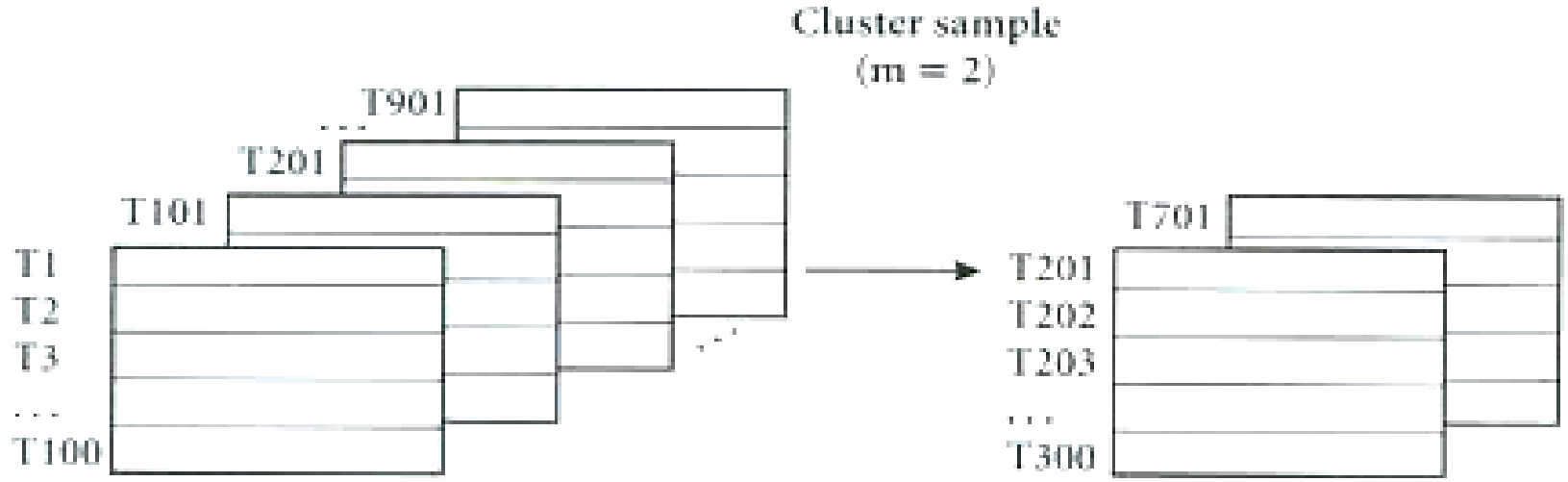
✓Sampling

- ✓ 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 same object can be picked up more than once population as they are selected for the sample,
- ✓ Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition
- ✓ Cluster Sampling
  - Heterogenous clusters are constructed, one or more clusters are chosen as the sample

Data Selection

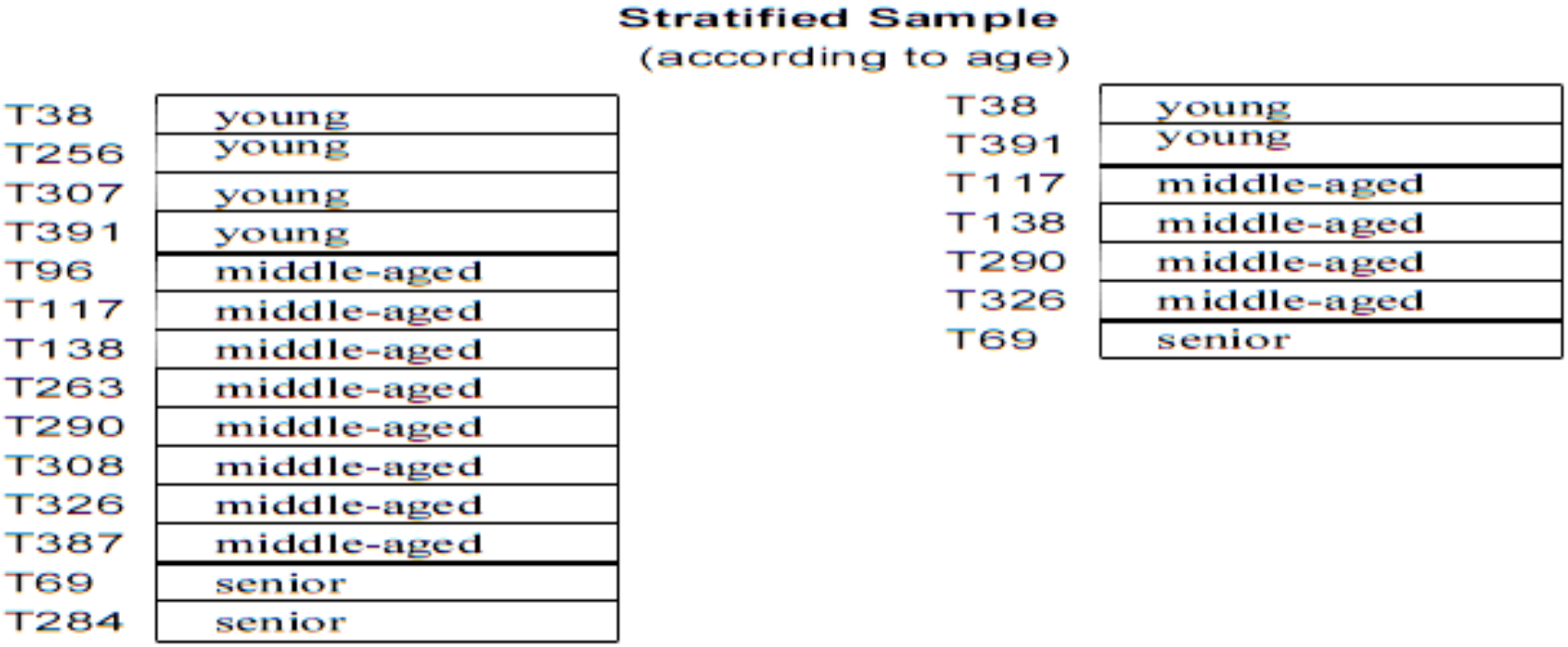
✓Sampling





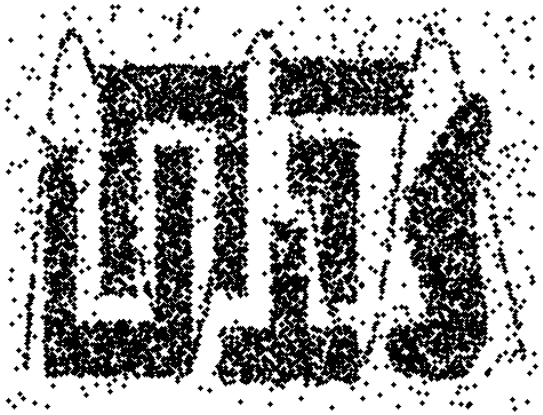
Data Selection

- ✓Sampling

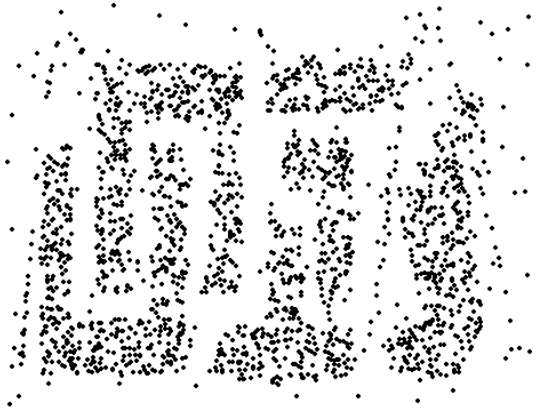


Data Selection

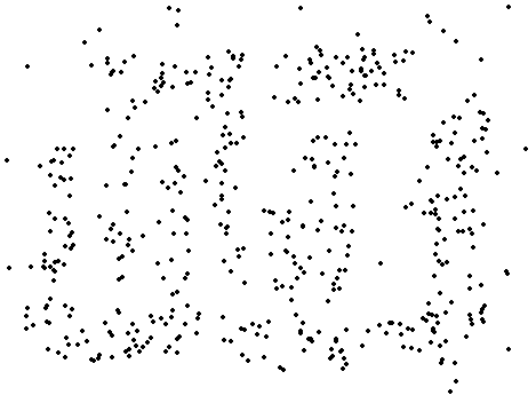
✓Sampling



8000 points



2000 Points



500 Points

Data Cleaning

- ✓ Real-world data tend to be ***incomplete***, ***noisy***, and ***inconsistent***.
- ✓ Data mining focuses on detection and correction of data quality errors, and using algorithms that can tolerate poor data quality.
- ✓ Tasks in data cleaning
  - Detect errors,
  - Fill in missing values,
  - Smooth noise,
  - Identify outliers,
  - Correct inconsistencies,

Data Cleaning

- ✓ Detecting errors – types of errors:
  - Typographic errors, user entry problems,
  - Data type errors (integer – floating point)
  - Measurement and data collection errors,
  - Missing values,
  - Noise and artifacts,
  - Outliers,
  - Duplicate data,
  - Inconsistent values.

Data Cleaning

- ✓ 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)
- ✓ Methods for handling with missing values:
  1. Ignore the tuple,
  2. Fill the missing value manually,
  3. Use a global constant to fill in the missing value ("*Unknown*", "∞"),
  4. Use the attribute mean/mod/median to fill in the missing value,
  5. Use the attribute mean for all samples belonging to the same class as the given tuple,
  6. Use the most probable value to fill in the missing value.

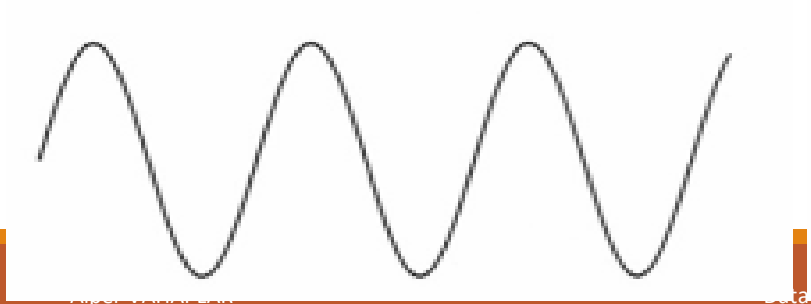
Data Cleaning

- ✓Find the appropriate missing salary value for ID = 1028.

ID	Gender	Age	Marital Satatus	Education	Region	Salary	Cluster
1021	F	41	Married	Master	Izmir	1200	**C1**
1022	M	27	Married	Bach.	Ankara	1000	**C1**
1023	M	20	NeverM	High School	Izmir	1000	C2
1024	F	34	Married	Bach.	İstanbul	1000	C3
1025	M	74	Married	Middle	Ankara	500	C2
1026	M	32	Married	PhD	İstanbul	2000	C2
1027	M	18	NeverM	High School	Ankara	800	C3
1028	F	43	Married	Master	Izmir	**???**	**C1**

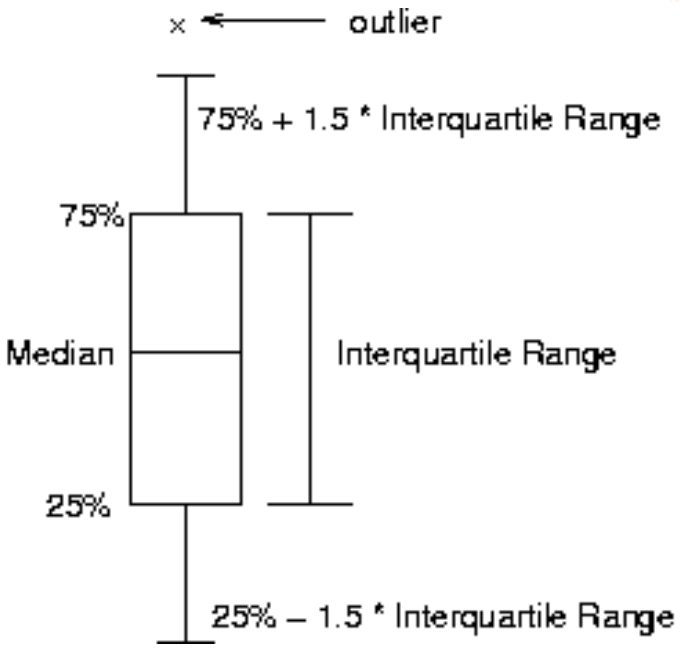
Data Cleaning

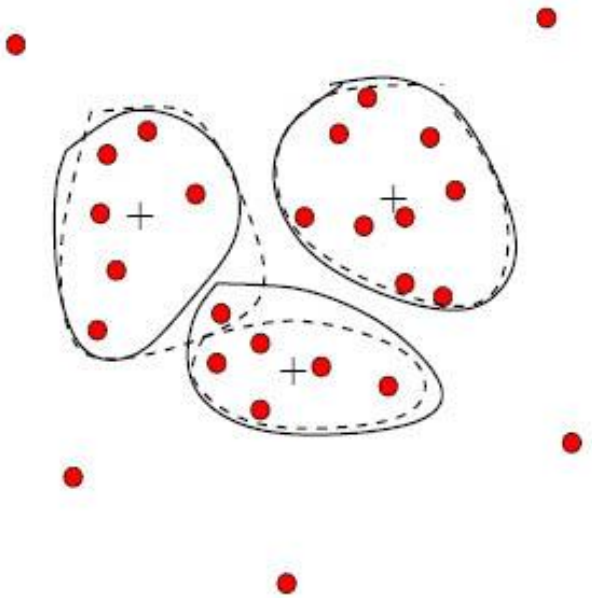
- ✓ Noise and Artifacts
- ✓ Noise is the random component of a measurement error.
- ✓ Noise is a random error or variance in a measured variable.
- ✓ Precision and Bias
  - Precision = mean – real value, Bias = standart deviation
  - E.g. Assess a weight scale with a mass of 1 kg.
  - (1.015, 0.990, 1.013, 1.001, 0.986)
  - Bias = 0.001, Precision = 0.013



Data Cleaning

- ✓ Outliers are
  - Data objects that have characteristics different from most of the other data,
- ✓ *Techniques* for *Smoothing Noise*
  - Histogram
  - Binning
  - Clustering
  - Regression





Data Cleaning

- ✓ Binning
- ✓ First data are sorted, then partition data into bins.
- ✓ **Equal-width partitioning**
  - Divides the range into N intervals of equal size
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
- ✓ **Equal-depth partitioning**
  - Divides the range into N intervals, each containing approximately same number of samples

Equal width	B1	B1	B2	B2	B2	B2	B2	B2	B2	B3	B3	B3
Price in €	4	6	14	16	18	19	21	22	23	25	27	33
Equal depth	B1	B1	B1	B1	B2	B2	B2	B2	B3	B3	B3	B3

Data Cleaning

- ✓Binning
  - (a) Equal Depth and Smoothing by Bin Means
  - (b) Equal Depth and Smoothing by Bin Boundaries
  - (c) Equal Width and Smoothing by Bin Means
  - (d) Equal Width and Smoothing by Bin Boundaries

Equal width	B1	B1	B2	B2	B2	B2	B2	B2	B2	B3	B3	B3
Price in €	4	6	14	16	18	19	21	22	23	25	27	33
Equal depth	B1	B1	B1	B1	B2	B2	B2	B2	B3	B3	B3	B3

Equal-Width Partitioning

(33-4) / 3 ~ 9
Bin1 (4-13) : 4 6
Bin2 (14-23) : 14 16 18 19 21 22 23
Bin3 (24-33) : 25 27 33

Equal-Depth Partitioning

Bin1 : 4 6 14 16
Bin2 : 18 19 21 22
Bin3: 23 25 27 33

Data Cleaning

- ✓ Binning – Smoothing
- ✓ Replace all values in a BIN (smoothing values)

Price in €	4	6	14	16	18	19	21	22	23	25	27	33
Equal depth	B1	B1	B1	B1	B2	B2	B2	B2	B3	B3	B3	B3
Smoothing by bin means	10	10	10	10	20	20	20	20	27	27	27	27
Smoothing by bin boundaries	4	4	16	16	18	18	22	22	23	23	23	33

Price in €	4	6	14	16	18	19	21	22	23	25	27	33
Equal width	B1	B1	B2	B2	B2	B2	B2	B2	B2	B3	B3	B3
Smoothing by bin means	5	5	19	19	19	19	19	19	19	28	28	28
Smoothing by bin boundaries	4	6	14	14	14	23	23	23	23	25	25	33

Binning Example

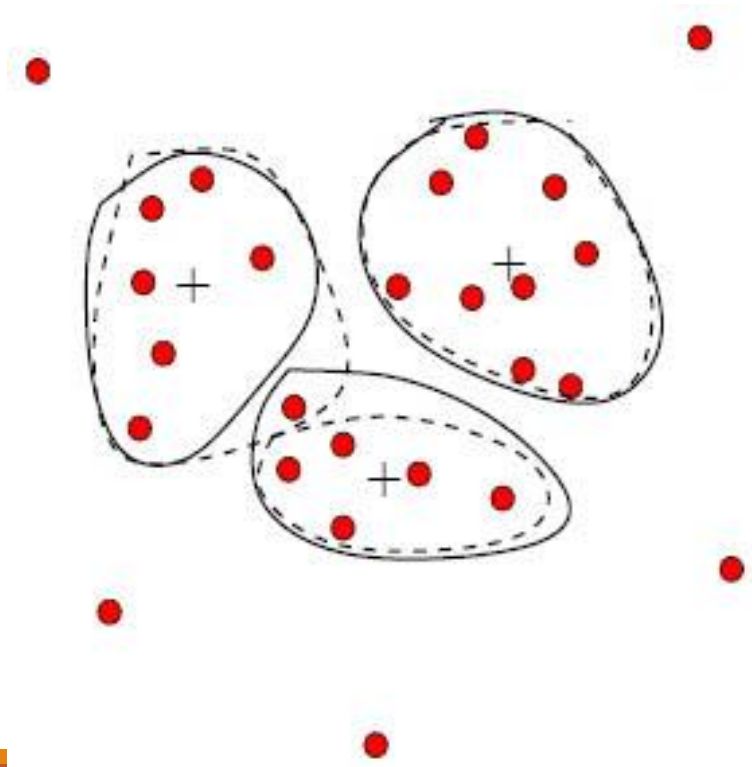
- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into Equal-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

* Partition into Equal-Width bins:
  - **Bin 1:** 4, 8, 9
  - **Bin 2:** 15, 21, 21, 24
  - **Bin 3:** 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, 9, 9
  - **Bin 2:** 15, 24, 24, 24
  - **Bin 3:** 25, 25, 25, 25, 34

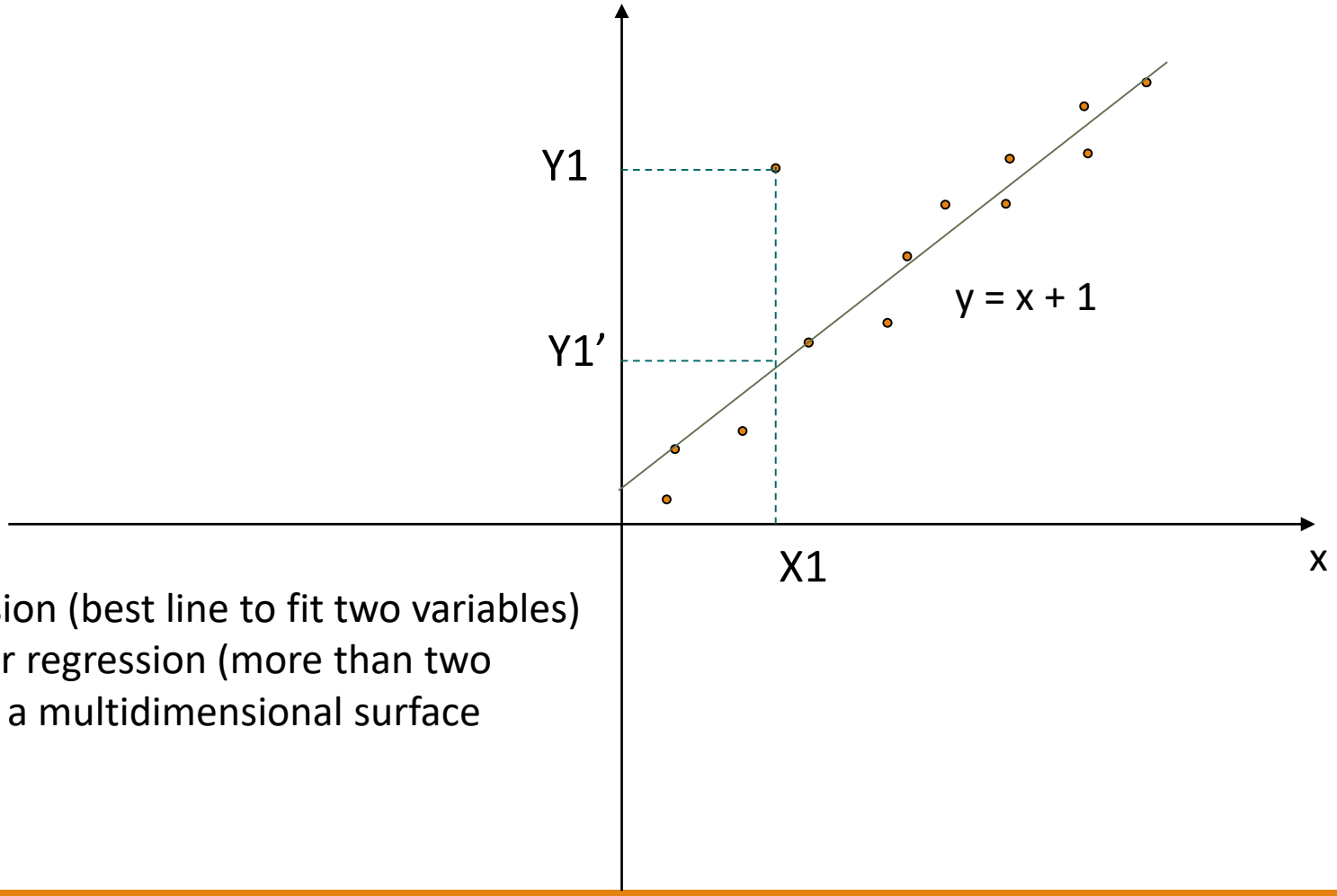
Data Cleaning

- ✓Clustering
  - Process of creating groups with similar objects.
  - Homogenity in cluster, heterogenity between clusters.



Data Cleaning

✓Regression



- Linear regression (best line to fit two variables)
- Multiple linear regression (more than two variables, fit to a multidimensional surface

Data Transformation

- Smoothing,
- Generalization,
- Normalization,
- Reduction,
- Feature construction

Data Transformation

- ✓Smoothing
  - Binning, Clustering, Regression
- ✓Generalization:
  - where low level data are replaced by higher level concepts through the use of concept hierarchies.
  - Ex: street attribute can be generalized like city;
  - age →child, young,middle aged, senior

Data Transformation

- ✓ Normalization: scaled to fall within a small, specified range
  - Min-max normalization
  - Z-score normalization
  - Normalization by decimal scaling

Normalization

- Min-max normalization: to [0, 1]

$$v' = \frac{v - min_A}{max_A - min_A}$$

- Min-max normalization: to [new_min_A, new_max_A]

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

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

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

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\max(|v'|) < 1$

Normalization Example

- ✓ Min-max normalization: $v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$
  - Ex. Let income range \$12,000 to \$98,000 normalized to [0, 1]. Then \$73,600 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1 - 0) + 0 = 0.716$$

- ✓ Z-score normalization
  - Ex. Let μ = 54,000, σ = 16,000. Then

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

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

- ✓ Normalization by decimal scaling $\frac{73{,}600}{100{,}000} = 0.736$ $v' = \frac{v}{10^j}$

Normalization Example

Price in €	4	6	14	16	18	19	21	22	23	24	27	34
Min-max [0,1]	0	.06	.33	.4	.46	.5	.56	.6	.63	.66	.76	1
Z-score	-1.8	-1.6	-0.6	-0.3	-0.1	0	0.2	0.4	0.5	0.6	1	1.8
Decimal Scaling	.04	.06	.14	.16	.18	.19	.21	.22	.23	.24	.27	.34

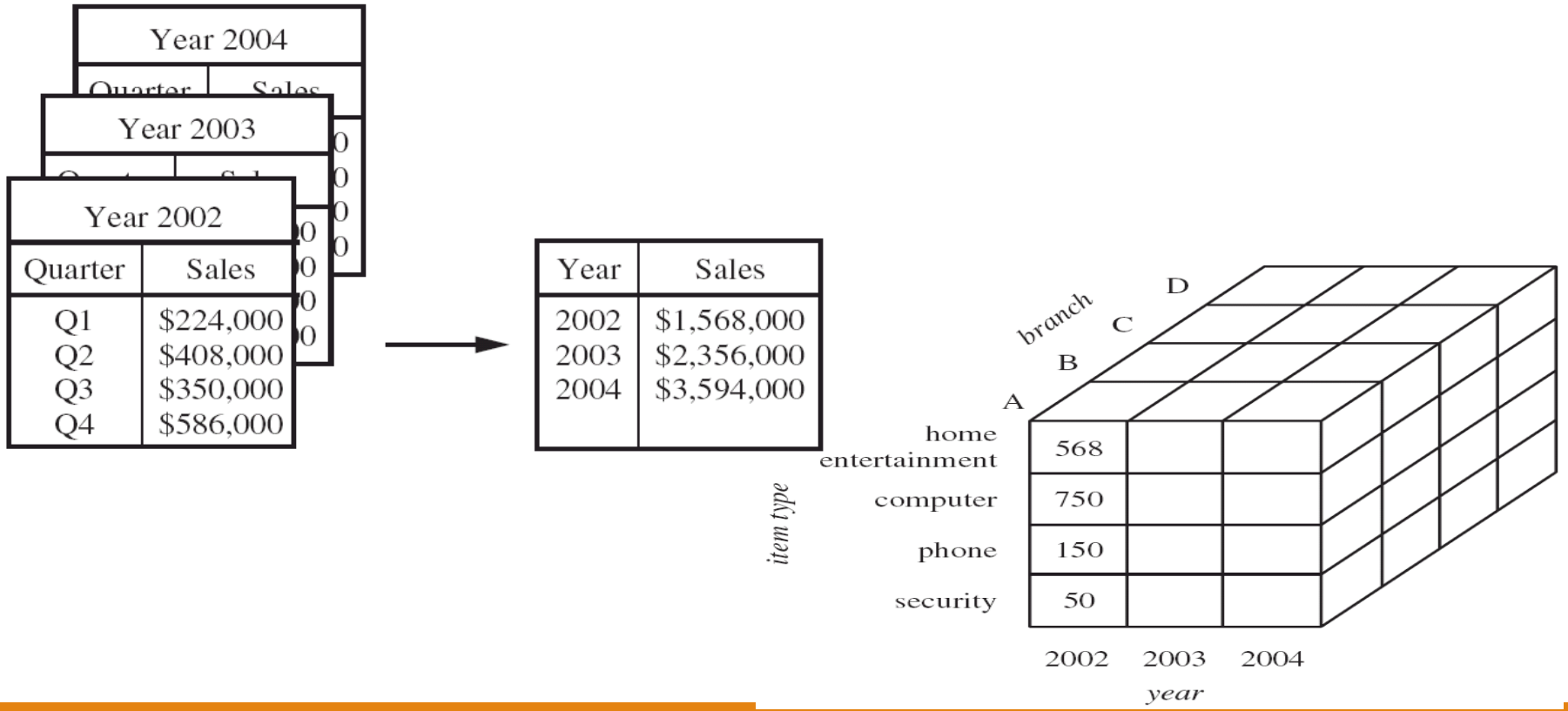
$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \qquad v' = \frac{v - \mu_A}{\sigma_A} \qquad v' = \frac{v}{10^j}$$

Data Reduction

- ✓Sampling,
- ✓Data Cube Aggregation,
- ✓Dimension Reduction,
- ✓Data Compression,
- ✓Numerosity Reduction,
- ✓Discretization and Concept Hierarchy Generation

Data Reduction

- ✓ **Aggregation:** Combining two or more attributes (or objects) into a single attribute (or object)



Data Reduction

- ✓ **Dimensionality Reduction:** reducing the data set size by
  - removing some attributes, or
  - producing new attributes (or models) to represent the older ones.
- ✓ Some methods
  - Stepwise Forward Selection,
  - Stepwise Backward Elimination,
  - Combination of both,
  - Principal Component Analysis,
  - Singular Value Decomposition,
  - Decision Tree Induction,

Data Reduction

Forward Selection

Initial attribute set:
{A1, A2, A3, A4, A5, A6}

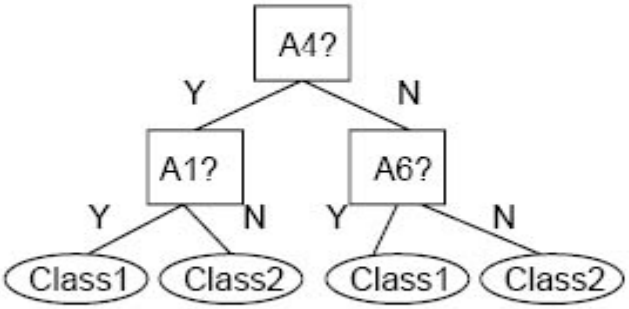
Initial reduced set:
{}
-> {A1}
--> {A1, A4}
---> Reduced attribute set:
{A1, A4, A6}

Backward Elimination

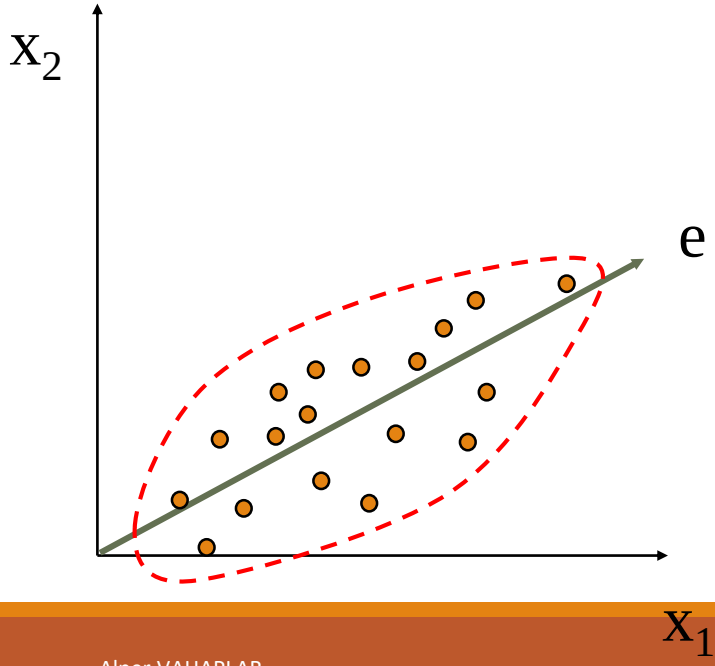
Initial attribute set:
{A1, A2, A3, A4, A5, A6}
-> {A1, A3, A4, A5, A6}
--> {A1, A4, A5, A6}
---> Reduced attribute set:
{A1, A4, A6}

Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



---> Reduced attribute set:
{A1, A4, A6}



Data Reduction

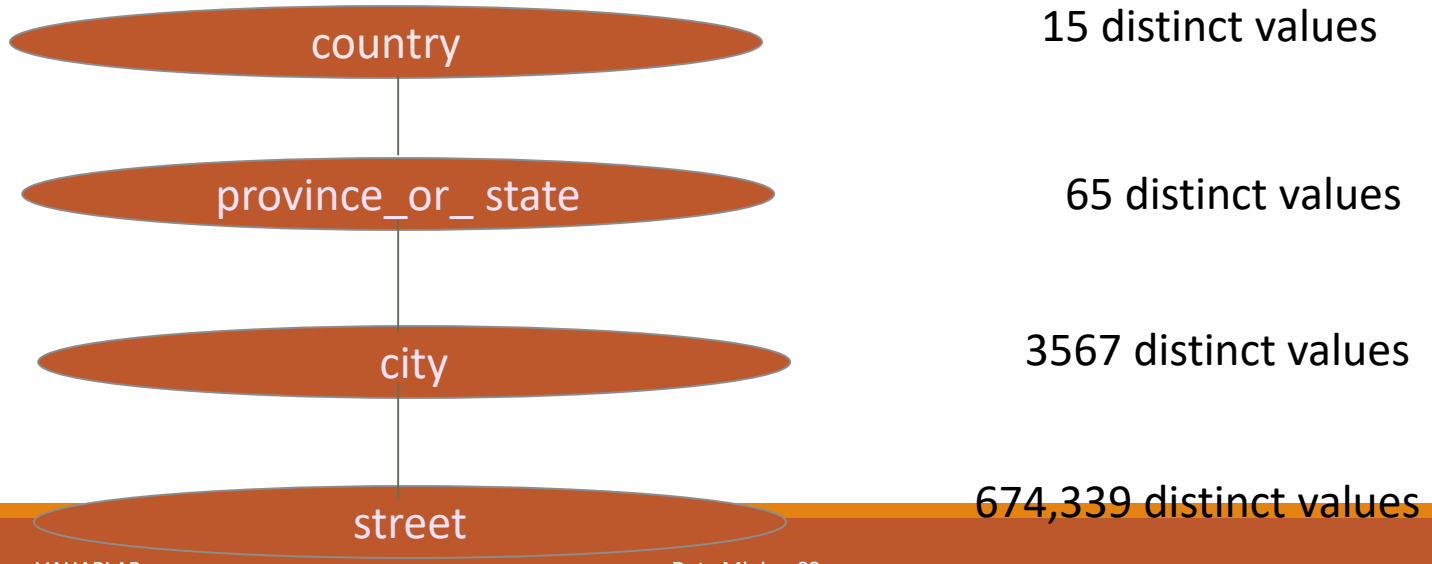
- ✓ Compression: encoding data to obtain a reduced representation.
- ✓ Losless – Lossy
- ✓ Methods for Compression
  - Standardization,
  - Normalization,
  - Logarithmic tranformation (log(x))
  - Fourier Transformation,
  - Wavelet Transformation.

Data Reduction

- ✓ Numerosity Reduction
- ✓ Parametric Methods: a model is used to estimate the data.
  - Linear Regression ($Y = \alpha + \beta X$),
  - Multiple Regression ($Y = b0 + b1\ X1 + b2\ X2$),
  - Logistic Regression, $(P = 1/(1+\exp(-(B0 + B1*X1 + B2*X2 + \ldots + Bk*Xk)))$
  - Poisson Regression,
  - Log Linear Models.
- ✓ Non Parametric Methods
  - Histogram Analysis,
  - Clustering,
  - Sampling.

Data Reduction

- ✓ Discretization and Concept Hierarchy Generation
- ✓ Discretization: reducing the number of values for a given continuous attribute by dividing the range of the attribute into intervals.
- ✓ Concept Hierarchy: reduce the data by collecting and replacing low level concepts by higher level concepts.
  - Ex: age → {young, middle-aged, old, fossil)



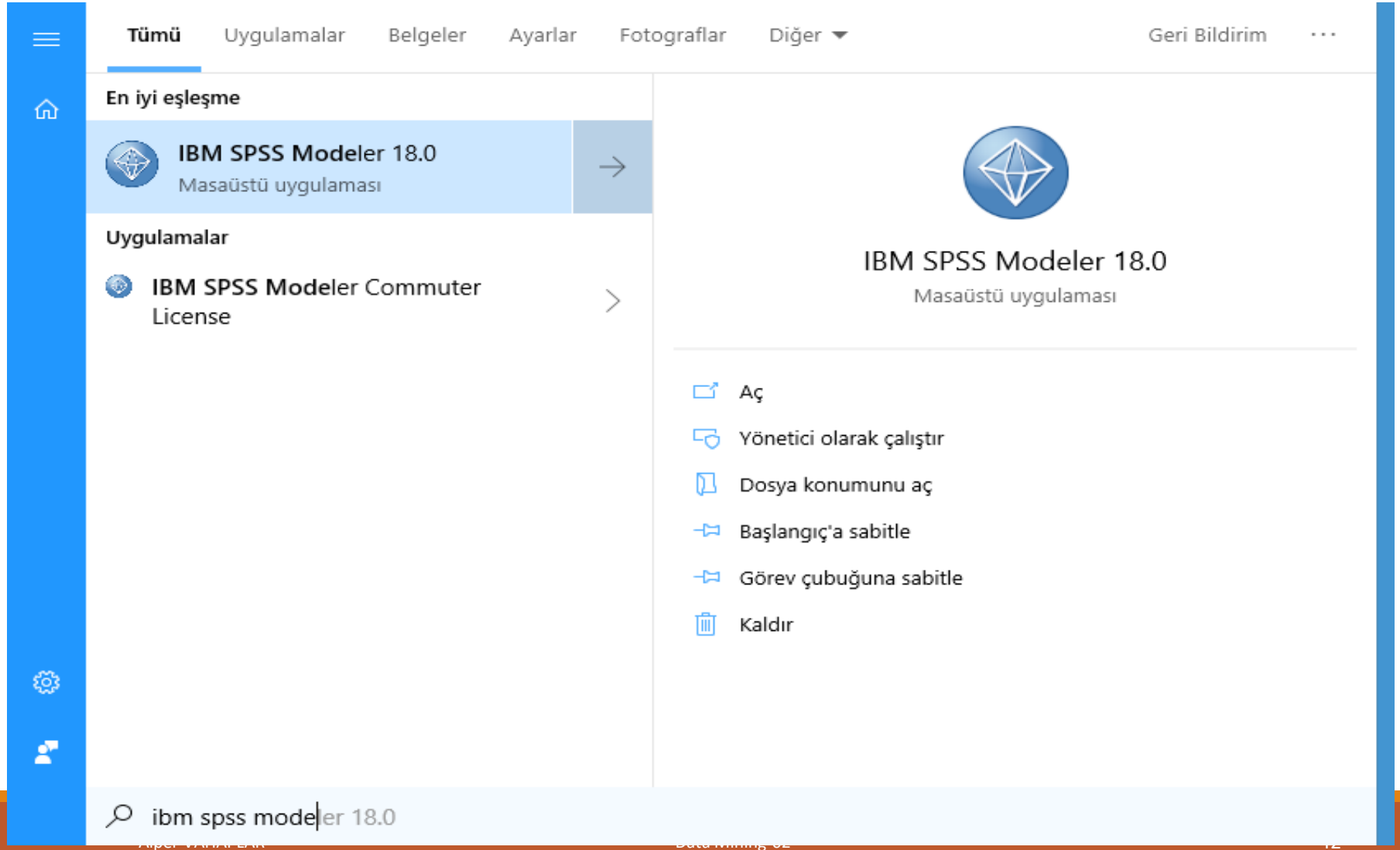
Data Reduction

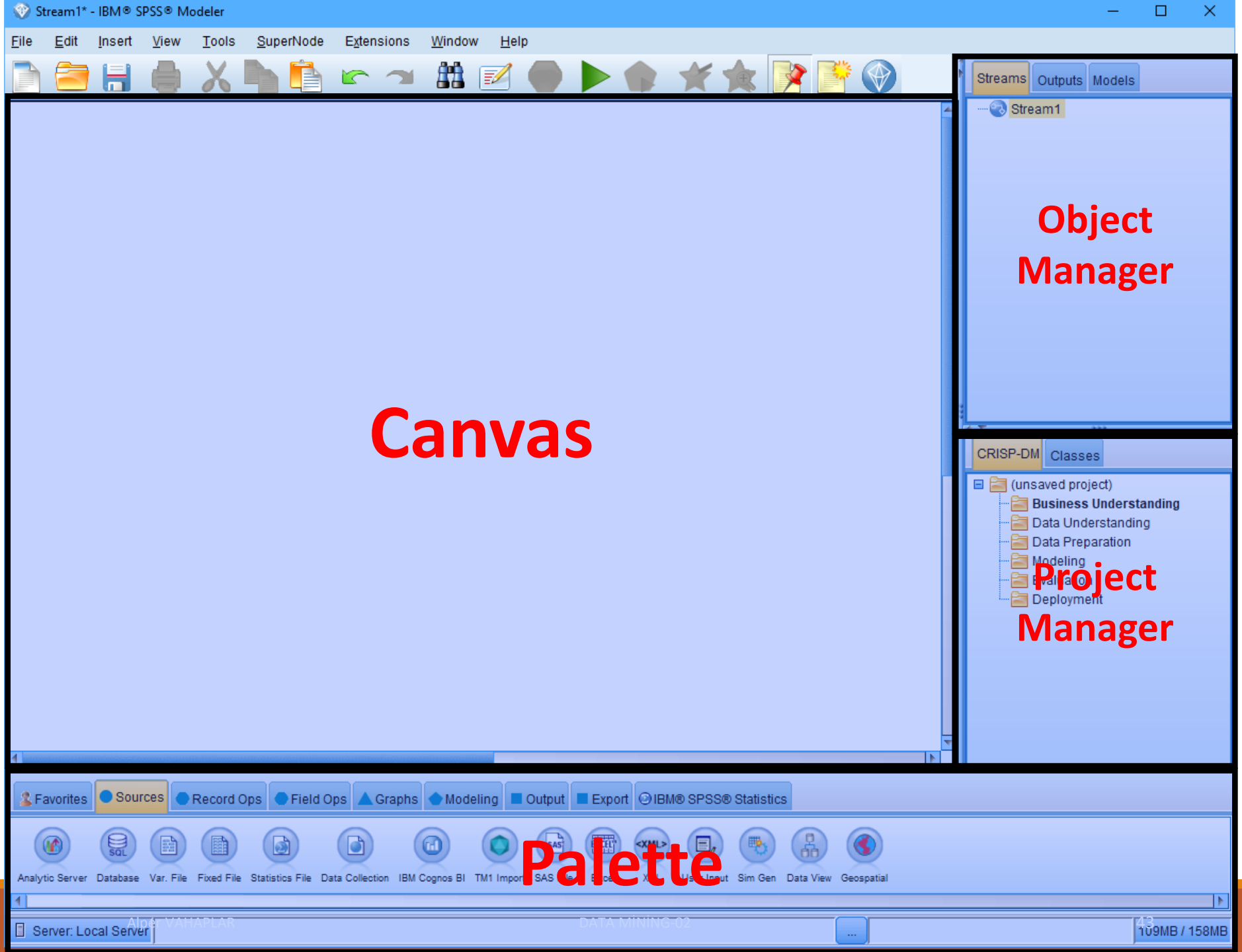
- ✓ Discretization and Concept Hierarchy Generation for Numeric Data
  - Binning
  - Histogram Analysis
  - Cluster Analysis
  - Entropy Based Discretization
  - Segmentation by Natural Partitioning.

IBM SPSS Modeler

An Introduction to Modeler 18.0

IBM SPSS Modeler 18.0




Stream1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
Streams Outputs Models
Stream1
Object Manager
Canvas
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Deployment
Project Manager
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics
Analytic Server Database Var. File Fixed File Statistics File Data Collection IBM Cognos BI Sim Gen Data View Geospatial
Palette
Server: Local Server
109MB / 158MB