

Chapter II: Data

Knowledge Discovery in Databases

Luciano Melodia M.A.

Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg

Summer semester 2021



Chapter II: Getting to know your data

This is our agenda for this lecture:

Data objects and attribute types

Basic statistical descriptions of data

Data visualization

Measuring data similarity and dissimilarity

Summary

Types of data sets

Records:

Relational records.

Data matrix, e.g. numerical matrix, crosstabs.

Document data: text documents, **term-frequency vectors**.

Transaction data.

	team	couch	play	ball	score	game
Document1	3	0	5	0	2	6
Document2	0	7	0	2	1	0
Document3	0	1	0	0	1	2

Graph and network:

World wide web.

Social of information networks.

Molecular structures.

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

Types of data sets

Ordered data:

- Video data: sequences of images.
- Temporal data: time series.
- Sequential data: transaction sequences.
- Genetic sequence data.

Spatial, image and multimedia:

- Spatial data: maps.
- Image data.
- Video data.

Important characteristics of structured data

Dimensionality:

Curse of dimensionality (sparse high-dimensional data spaces).

Sparsity:

Only presence counts.

Resolution:

Patterns depend on the scale.

Distribution:

Centrality and dispersion.

Data objects

Data sets are made up of data objects.

A data object represents an entity.

Examples:

Sales database: customers, store items, sales.

Medical database: patients, treatments.

University database: students, professors, courses.

They are also called:

Samples, examples, instances, data points, objects, tuples, ...

Data objects are described by attributes:

Database rows → data objects.

Columns → attributes.

Attributes

Attribute:

Sometimes also in other context: field, dimension, feature, variable, ...

A data field encodes the property of an entity or feature of a data object.

E.g. customer_ID, name, address.

Types:

- Nominal.

- Binary.

- Ordinal.

- Numerical:

 - Interval scaled.

 - Ratio scaled.

Thank you for your attention.
Any questions about the second chapter?

Ask them now, or again, drop me a line:
✉ luciano.melodia@fau.de.