Machine Learning and Data Mining

Zürcher Hochschule
für Angewandte Wissenschaften

L02: Data Processing
Part 1

Lecturer: **Mark Cieliebak**

Based on material by Andreas Weiler (wele), Mark Cieliebak (ciel), and Thilo Stadelmann (stdm)

Lecture Plan

1. Introduction
2. **Data Processing**
3. Evaluation
4. Recommender Systems
5. Association Rules
6. Linear/Logistic Regression
7. Support Vector Machines
8. Decision Trees + Naive Bayes
9. Clustering
10. Deep Learning
11. Guest Lectures
12. Exam Preparation / Lecture Wrapup

Repetition

Previous Lecture

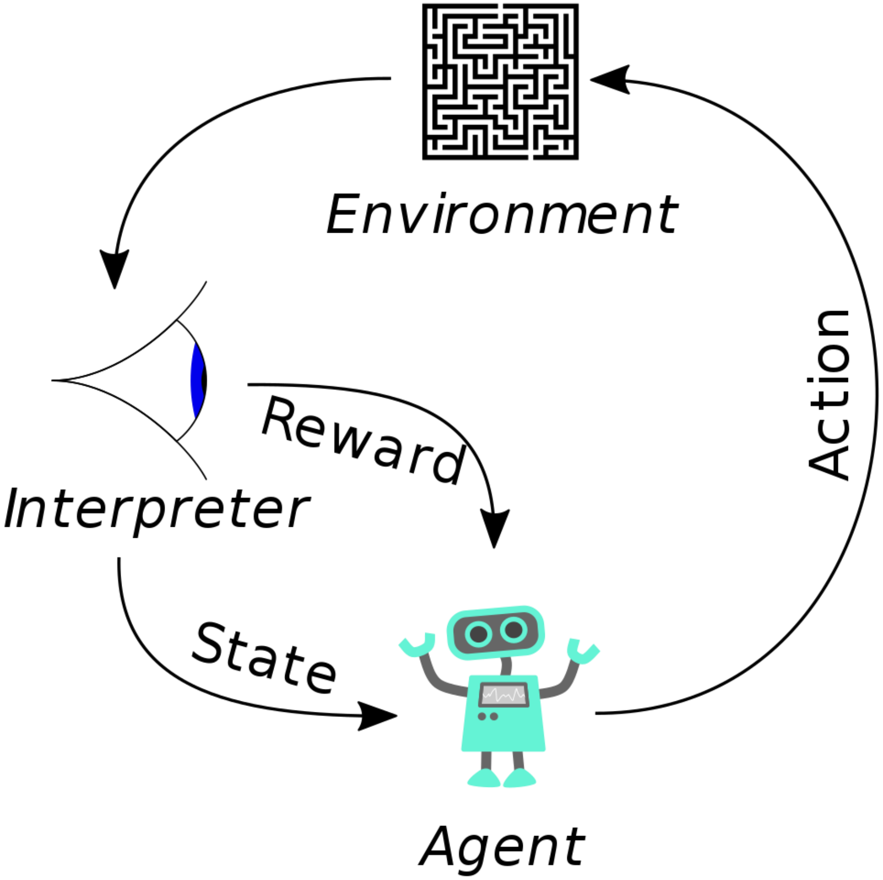
- Current rise of AI is based on Data, Computation Power, Expertise and Algorithms

- Data Mining discovers patterns in large amounts of data

- Machine Learning uses patterns in data to perform a specific task

- Machine Learning Types:
  - Unsupervised
  - Supervised
  - Reinforcement Learning

Reinforcement Learning Model

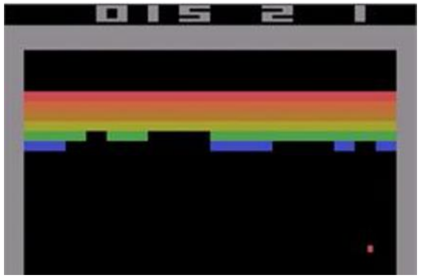


- At each step t the agent:
  - Executes action A_t
  - Receives observation O_t
  - Receives scalar reward R_t
- The environment:
  - Receives action A_t
  - Emits observation O_{t+1}
  - Emits scalar reward R_{t+1}
- t increments at env. step

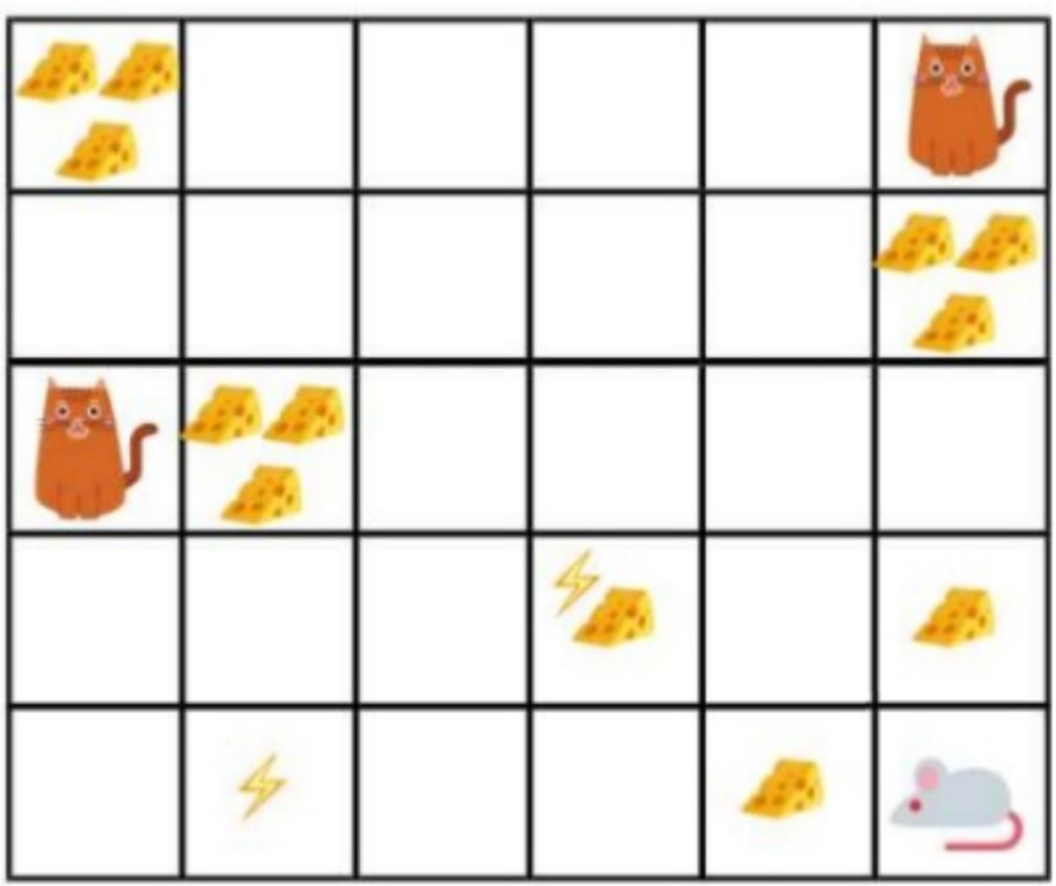
https://en.wikipedia.org/wiki/Reinforcement_learning#/media/File:Reinforcement_learning_diagram.svg

Examples of Reinforcement Learning

- Fly stunt manoeuvres in a helicopter
- Defeat the world champion at Backgammon
- Manage an investment portfolio
- Control a power station
- Make a humanoid robot walk
- Play Atari games better than humans
- Autonomous cars



Sample World: Mice and Cheese



https://medium.com/@julsimon/talk-an-introduction-to-reinforcement-learning-e26177338787

Agent is the mouse

Actions: go up, down, left, right

Reward:

- zero reward for blank cell
- positive for finding cheese
- negative for electric shock
- very negative for meeting cat

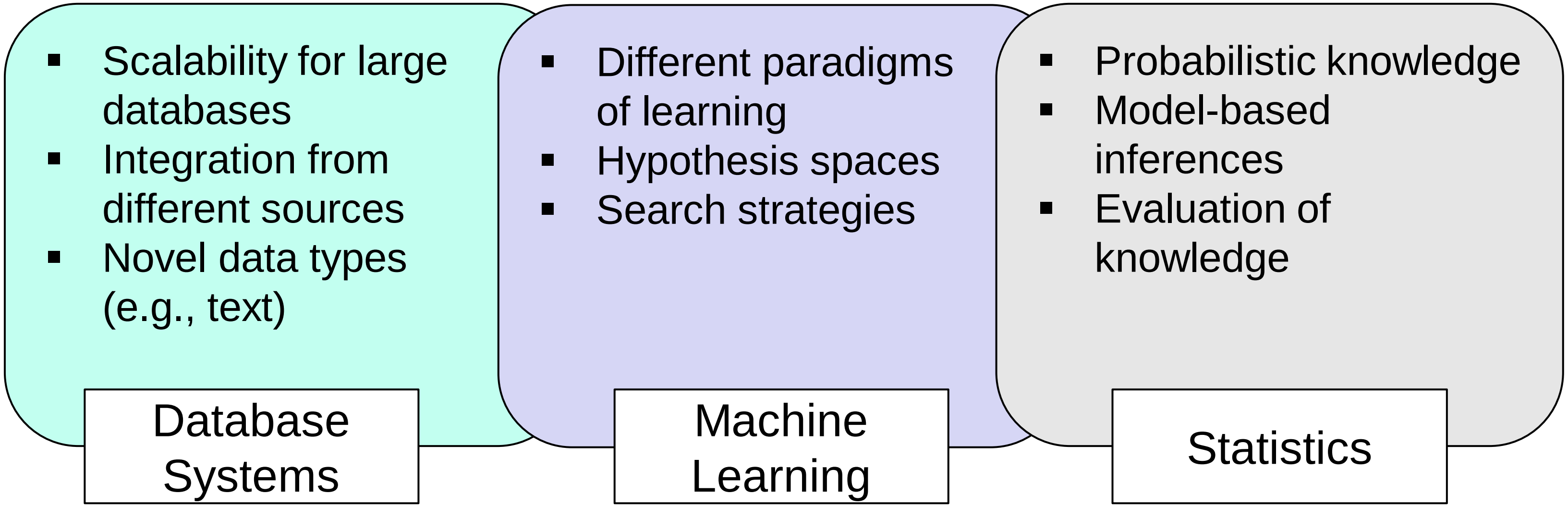
Episode: until mouse dies or max 50 steps

Learning Objectives

- Understand the process of "Knowledge Discovery in Databases" (KDD)
- Distinguish between different data types and data classes
- Understand data preprocessing, cleaning, and exploration

Knowledge Discovery in Databases

- KDD is the process of (semi-) automatic extraction of knowledge from databases which is
  - valid
  - previously unknown
  - and potentially useful
- Interdisciplinary field



- Scalability for large databases
- Integration from different sources
- Novel data types (e.g., text)

Database Systems

- Different paradigms of learning
- Hypothesis spaces
- Search strategies

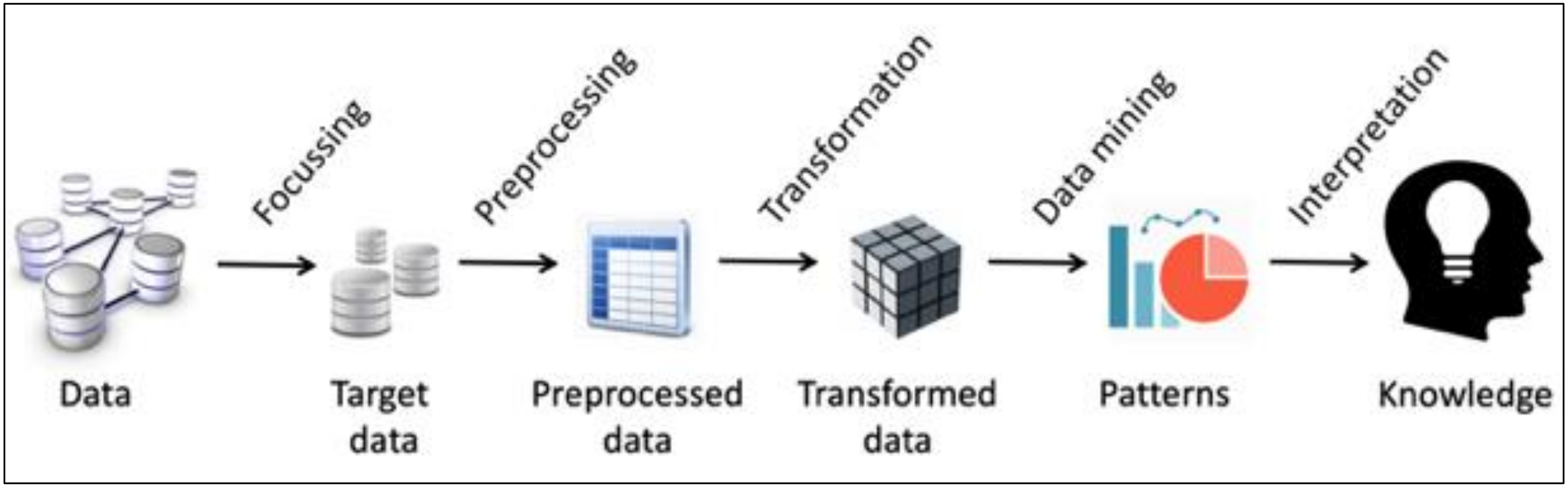
Machine Learning

- Probabilistic knowledge
- Model-based inferences
- Evaluation of knowledge

Statistics

Knowledge Discovery in Databases

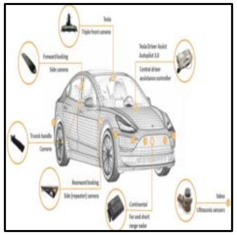
- Interactive and iterative process
- Continuous optimization of the different tasks

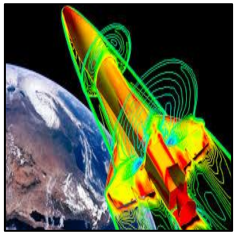


Data

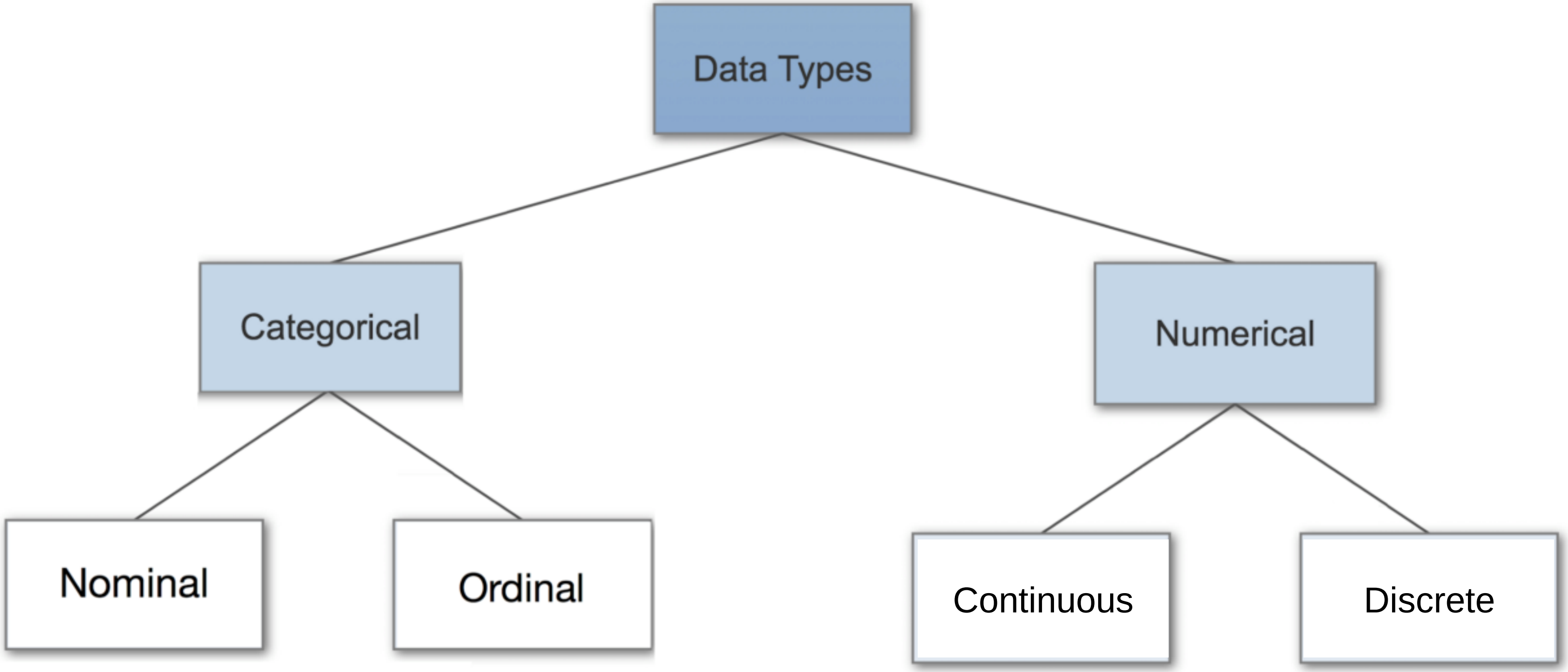
Data

- Has many sources, e.g.
  - sensor data
  - survey data
  - simulation data
  - social media data
  - textual data
  - financial data
  - multimedia data
  - ERP systems data
- Independent of the **data source**, each data point has a **data type**





Data Types


Data Types
Categorical
Numerical
Nominal
Ordinal
Continuous
Discrete

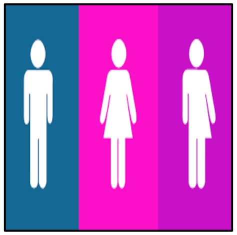
Data Types

Nominal Data

- Nominal scales are used for labeling variables, without any quantitative value
- No numerical significance
- Nominal data has no order
- Scales could simply be called “labels”

Examples:

- gender
- hair color
- race
- marital status



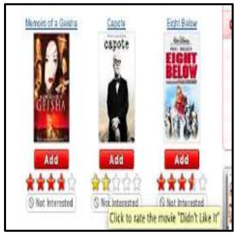
Data Types

Ordinal Data

- Represent discrete and ordered units
- Nearly the same as nominal data, except that it's ordering matters
- No distance between the different categories

Examples:

- military rank
- movie rating by number of stars
- educational background
- difficulty of cooking recipe

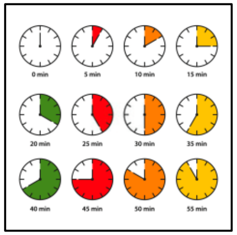


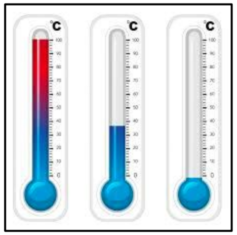
Click to rate the movie "Didn't Like It"

Numeric Data Types

Discrete Numeric Data

- Represent items that can be counted
- Values may go from 0, 1, 2, on to infinity (making it countably infinite)





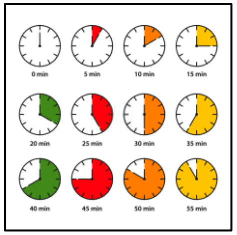
Examples:

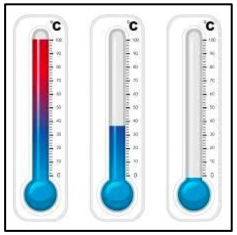
- Number of persons in a room
- Number of "heads" in 100 coin flips

Numeric Data Types

Continuous Numeric Data

- Also known as "Interval Data"
- Often measurements
- Possible values cannot be counted and can only be described using intervals on the real number line.



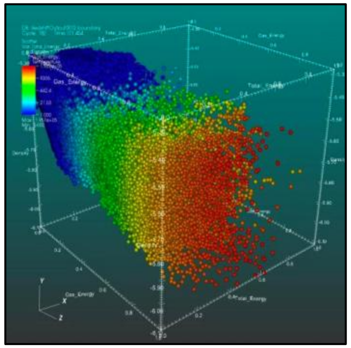


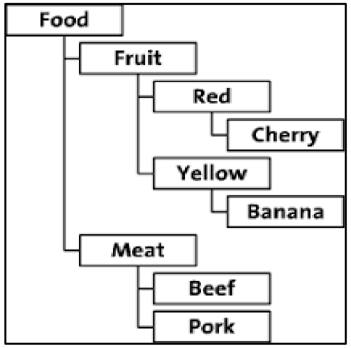
Examples:

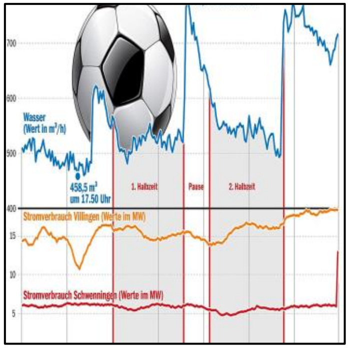
- Exact amount of gas purchased at the pump for cars with 20-gallon (represented by [0, 20]
- time e lapsed in a 100m run
- lifetime of a battery (0 hours to an infinite number of hours)

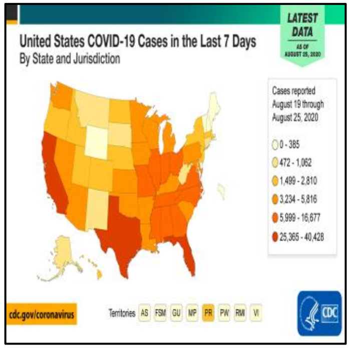
Typical Data Classes

- One-dimensional data
- Multi-dimensional data
- Network data
- Hierarchical data
- Time-series
- Geographic data









Question

Give examples for data types / classes that you could use to describe the movement of a flock of lions.

- Nominal?
- Ordinal?
- Numeric?
- Network Data?
- Hierarchical?
- Time-Series?
- Geographic?

Data Cleaning

Data Cleaning

- Process of improving the data quality
- Low-quality data will lead to low-quality mining results
- Removing or modifying incorrect or improperly formatted data

Example dataset “Breaking Bad”:

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
sunset	25.10.2011	9.3	6	3	47
grilled	20.12.2009	9.3	2	2	46

Task (2 min): Data Issues

Which data issues can you find in this data?

id	title	date	rating	episode	season	length
1	live free or die	01.11.2012	9.3	1	5	43
2	madrigal	06.12.2013	89	2	5	48
3	sunset	25.10.2011	9.3	6	3	47
4	grilled	20.12.2009	9.3	2	2	46
5	down	42.12.2009	8.3	4	2	333
6	cancer man	19.10.2008	8.3	4	1	48
7	live fre or die	01.11.2012	9.3	1	5	43

firstname	lastname	bdate	age	gender	phone
bryan	cranston	03-07-1956	65	1	999-9999
aaron	paul	27-08-1979	44	m	777-53474
anna, gunn	gunn	08-11-1968	52	0	040-15627

Data Cleaning:
(Near) Duplicates

Data Cleaning: (Near) Duplicates

Example dataset “Breaking Bad”:

id	title	date	rating	episode	season	length
1	live free or die	01.11.2012	9.3	1	5	43
2	madrigal	06.12.2013	89	2	5	48
3	sunset	25.10.2011	9.3	6	3	47
4	grilled	20.12.2009	9.3	2	2	46
5	down	42.12.2009	8.3	4	2	333
6	cancer man	19.10.2008	8.3	4	1	48
7	**live fre or die**	01.11.2012	9.3	1	5	43

Data Cleaning: (Near) Duplicates

- Compare the content of the attributes
  - For numeric values: cosine similarity of feature vectors
  - For text: pairwise Levensthein distance between the texts

	1	2	3	4	5	6	7
live free or die	0	15	15	14	15	13	1
madrigal	15	0	8	7	7	7	14
sunset	15	8	0	6	6	8	14
grilled	14	7	6	0	7	10	13
down	15	7	6	7	0	9	14
cancer man	13	7	8	10	9	0	12
live fre or die	1	14	14	13	14	12	0

Data Cleaning:
Missing Values

QUESTION: Missing Values

How could you deal with missing values?

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
sunset	25.10.2011	🟥	6	3	47
grilled	20.12.2009	9.3	2	2	46
down	22.12.2009	8.3	4	2	🟥
cancer man	19.10.2008	8.3	4	1	48
shotgun	09.11.2012	8.7	5	4	47

Data Cleaning: Missing Values

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
~~sunset~~	~~25.10.2011~~	🟥	~~6~~	~~3~~	~~47~~
grilled	20.12.2009	9.3	2	2	46
~~down~~	~~22.12.2009~~	~~8.3~~	~~4~~	~~2~~	🟥
cancer man	19.10.2008	8.3	4	1	48
shotgun	09.11.2012	8.7	5	4	47

ignore the tuples

Data Cleaning: Missing Values

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
sunset	25.10.2011	9.3	6	3	47
grilled	20.12.2009	9.3	2	2	46
down	22.12.2009	8.3	4	2	47
cancer man	19.10.2008	8.3	4	1	48
shotgun	09.11.2012	8.7	5	4	47

fill in the missing value manually

Data Cleaning: Missing Values

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
sunset	25.10.2011	-1	6	3	47
grilled	20.12.2009	9.3	2	2	46
down	22.12.2009	8.3	4	2	-1
cancer man	19.10.2008	8.3	4	1	48
shotgun	09.11.2012	8.7	5	4	47

use a global constant

Data Cleaning: Missing Values

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
sunset	25.10.2011	8.8	6	3	47
grilled	20.12.2009	9.3	2	2	46
down	22.12.2009	8.3	4	2	46.5
cancer man	19.10.2008	8.3	4	1	48
shotgun	09.11.2012	8.7	5	4	47

use the attribute mean

Data Cleaning: Missing Values

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
sunset	25.10.2011	9.3	6	3	47
grilled	20.12.2009	9.3	2	2	46
down	22.12.2009	8.3	4	2	47
cancer man	19.10.2008	8.3	4	1	48
shotgun	09.11.2012	8.7	5	4	47

use the most probable value

Data Cleaning: Missing Values

ignore the tuple

+ can be easily done
+ no computational effort

- loss of information
- unnecessary if the attribute is not needed

enter value manually

+ for small datasets effective
+ “real” value

- not feasible for large datasets
- time consuming
- error-prone

use attribute mean

+ simple to implement

- not the most accurate approximation of the value

use a global constant

+ can be easily done
+ missing values are marked

- values can not be used in algorithms

use most probable value

+ most accurate approximation of the value

- most computational effort

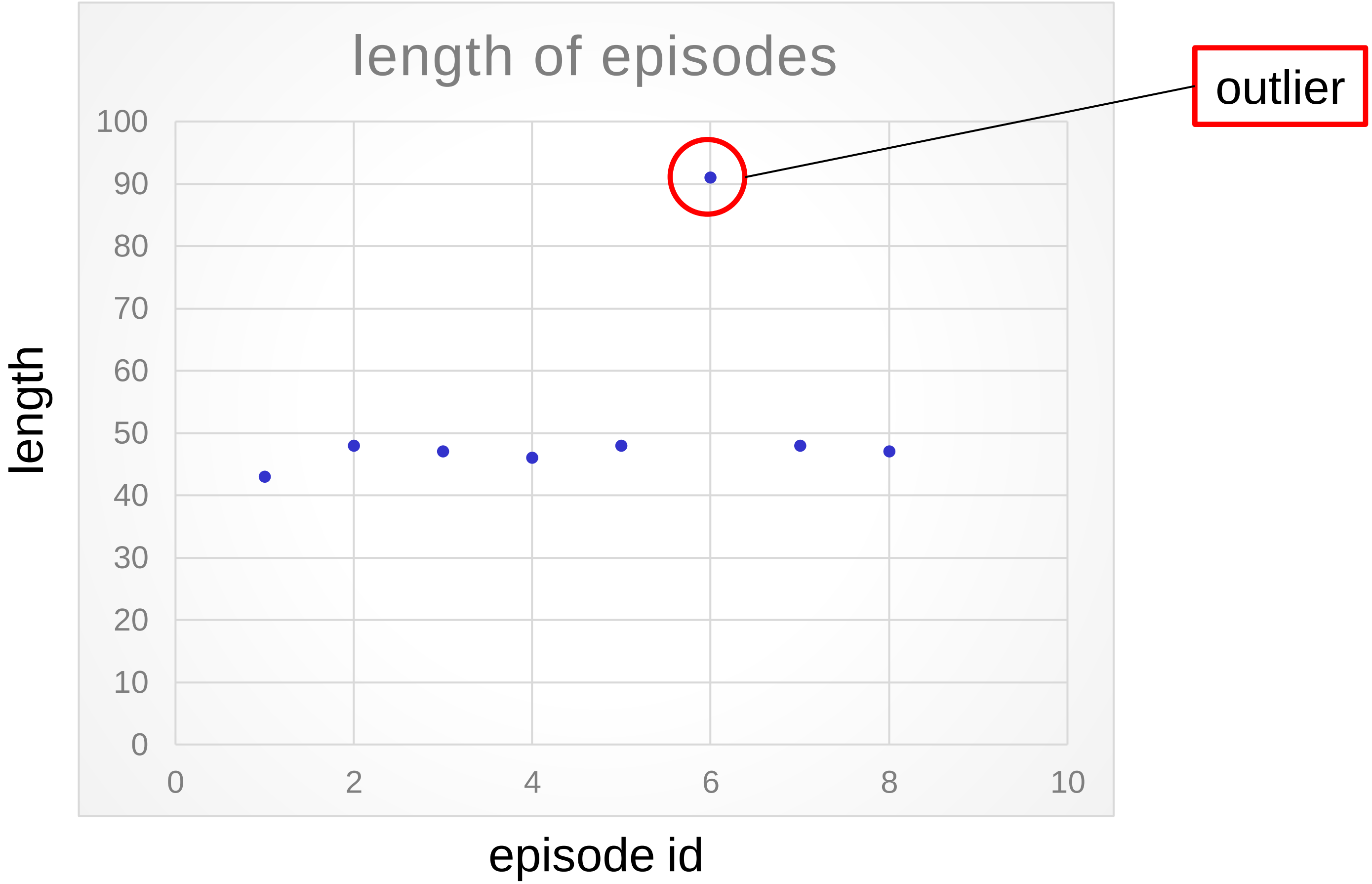
Data Cleaning:
Noisy Data

Data Cleaning: Noisy Data

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
sunset	25.10.2011	9.3	6	3	47
grilled	20.12.2009	9.3	2	2	46
down	22.12.2009	8.3	4	2	**91**
cancer man	19.10.2008	8.3	4	1	48
shotgun	09.11.2012	8.7	5	4	47

Data Cleaning: Noisy Data

Example dataset “Breaking Bad”:



Data Cleaning: Smoothing

Data Cleaning: Smoothing

Binning

- Sort data and partition into (equi-depth) bins and then smooth by bin means, bin median, bin boundaries, etc.

Regression

- Smooth by fitting a regression function

Clustering

- Cluster data and remove outliers
- Fully automatic or via manual exploration

Data Cleaning: Smoothing

Equal-width Binning

- divides the range into N intervals of equal size
- width of intervals: width = (max - min) / N
- simple method
- outliers may dominate result

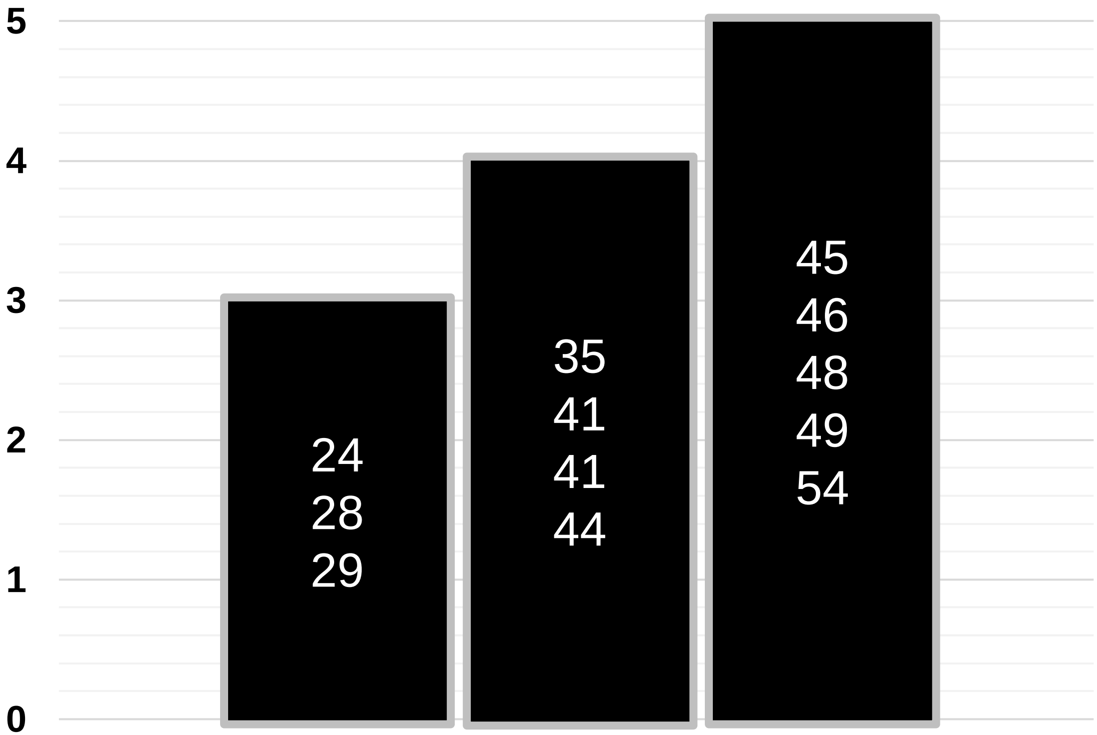
Equal-depth Binning

- divides the range into N intervals
- each interval contains approximately the same number of records
- skewed data is also handled well

Data Cleaning: Smoothing

Equal-width Binning

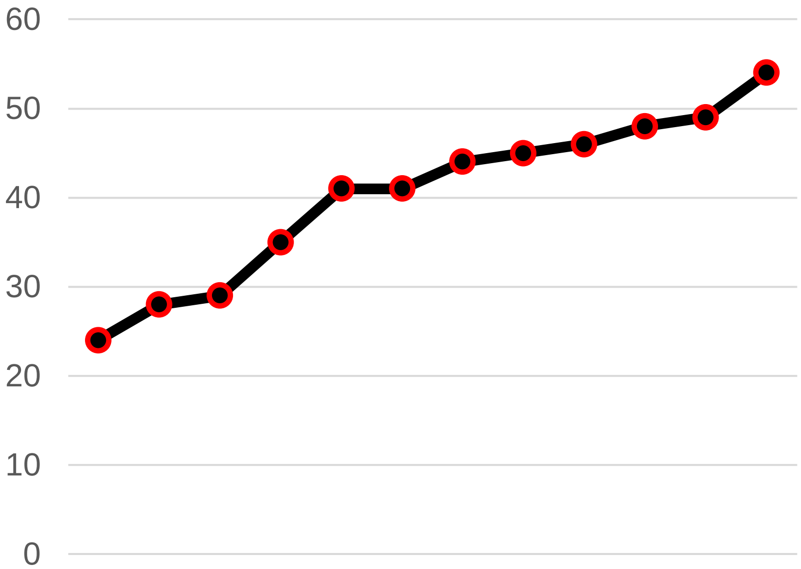
- Sorted price values: 24, 28, 29, 35, 41, 41, 44, 45, 46, 48, 49, 54
- N = 3 (user-defined)
- Width = 54 – 24 / 3 = 10



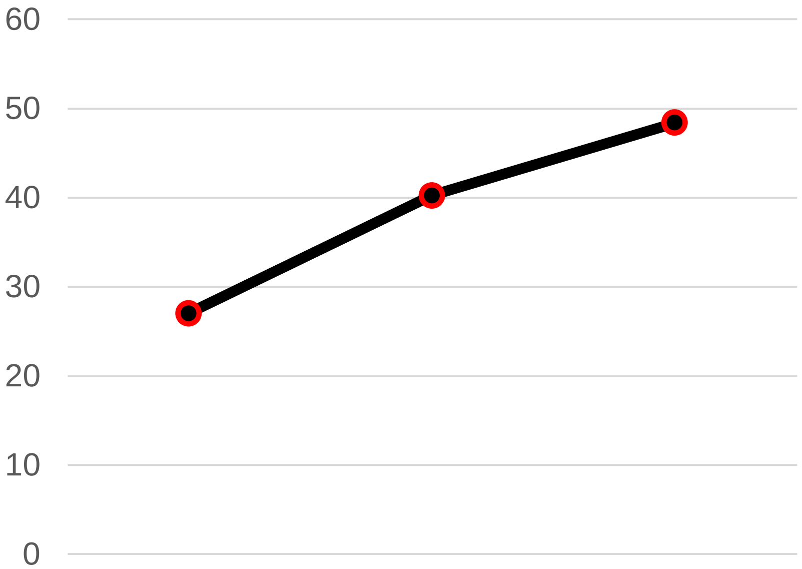
Data Cleaning: Smoothing

Equal-width Binning

- sorted price values: 24, 28, 29, 35, 41, 41, 44, 45, 46, 48, 49, 54
- N = 3 (user-defined)
- width = 54 – 24 / 3 = 10



original

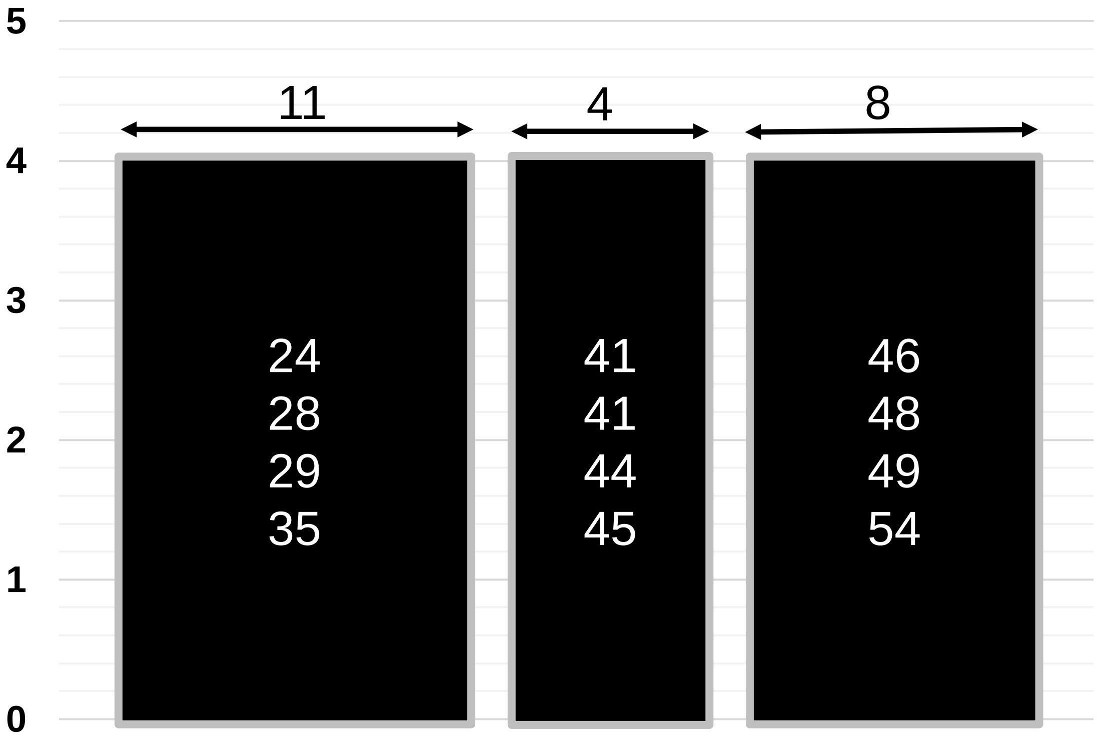


3 bin averages

Data Cleaning: Smoothing

Equal-depth Binning

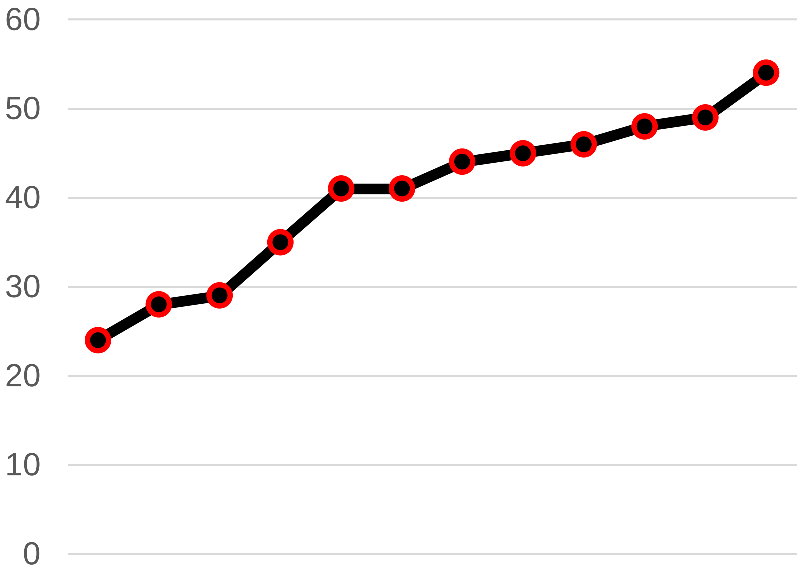
- Sorted price values: 24, 28, 29, 35, 41, 41, 44, 45, 46, 48, 49, 54
- N = 3



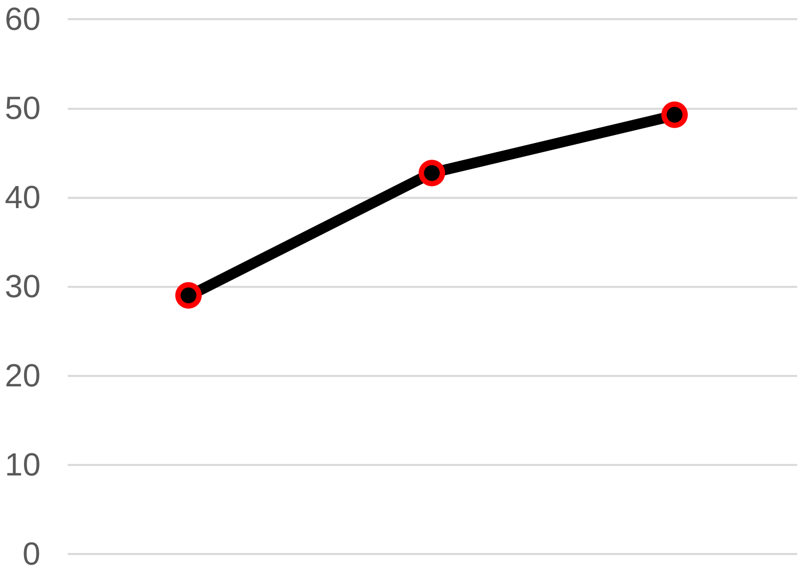
Data Cleaning: Smoothing

Equal-depth Binning

- Sorted price values: 24, 28, 29, 35, 41, 41, 44, 45, 46, 48, 49, 54
- N = 3



original



3 bin averages

Data Cleaning: Smoothing

Equal-depth Binning

- Sorted price values: 24, 28, 29, 35, 41, 41, 44, 45, 46, 48, 49, 54
- N = 3
- Partition into 3 equal-depth bins:
  - [24, 28, 29, 35], [41, 41, 44, 45], [46, 48, 49, 54]

- **Smoothing by bin means:**
  - replace each value by the mean value of the bin)
  - [29, 29, 29, 29], [43, 43, 43, 43], [49, 49, 49, 49]

- **Smoothing by bin boundaries:**
  - replace each value by closest boundary value
  - [24, 24, 24, 35], [41, 41, 45, 45], [46, 46, 46, 54]

Questions and Answers

Zürcher Hochschule
für Angewandte Wissenschaften

Data Normalization

Data Normalization

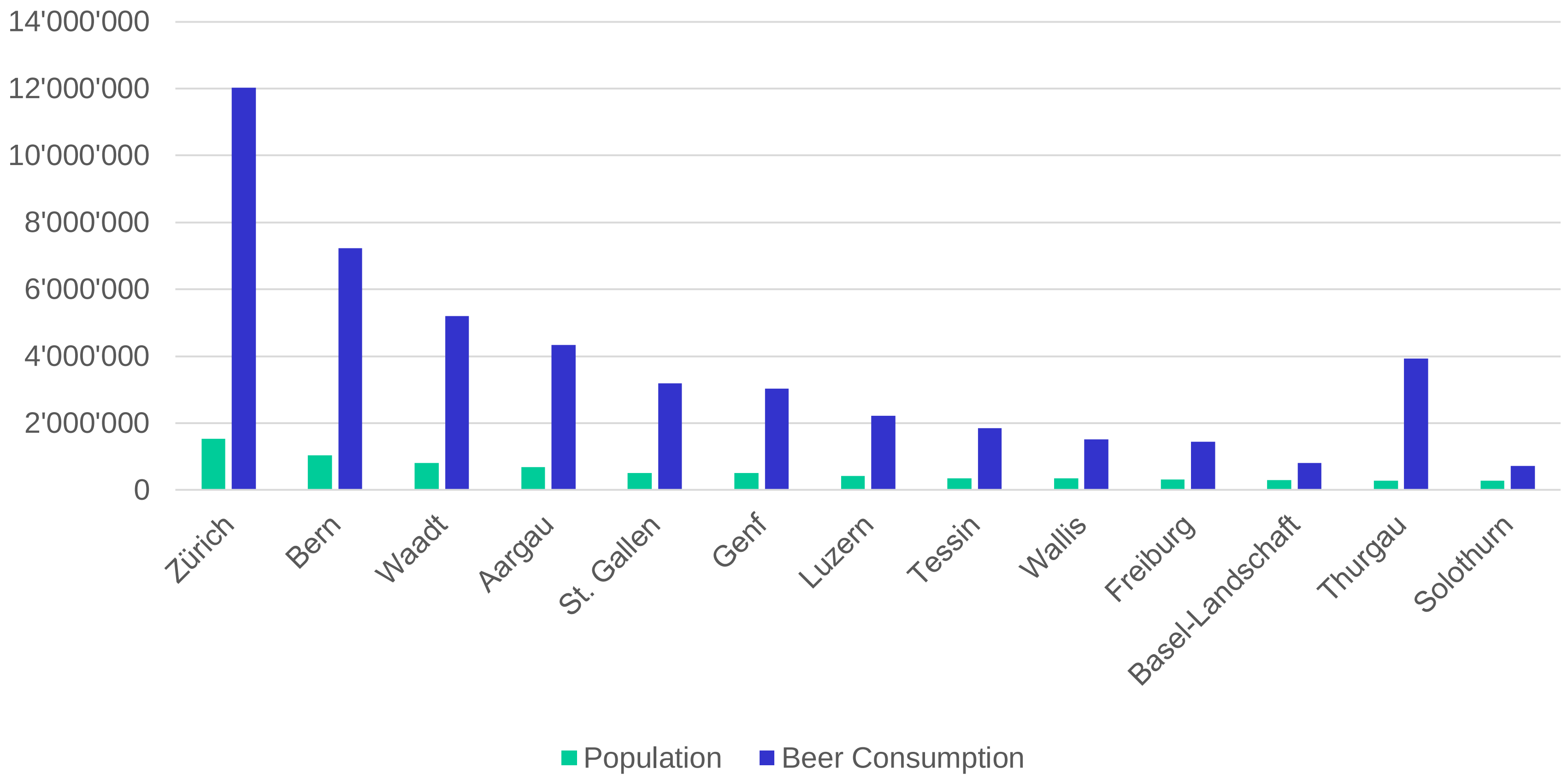
Goal: Change the values of numeric columns to a common scale, without distorting differences in the ranges of values

Canton	Population
Zürich	1539275
Bern	1039474
Waadt	805098
Aargau	685845
St. Gallen	510734
Genf	504128
Luzern	413120
Tessin	351491
Wallis	345525
Freiburg	321783
Basel-Landschaft	289468
Thurgau	279547
Solothurn	275247

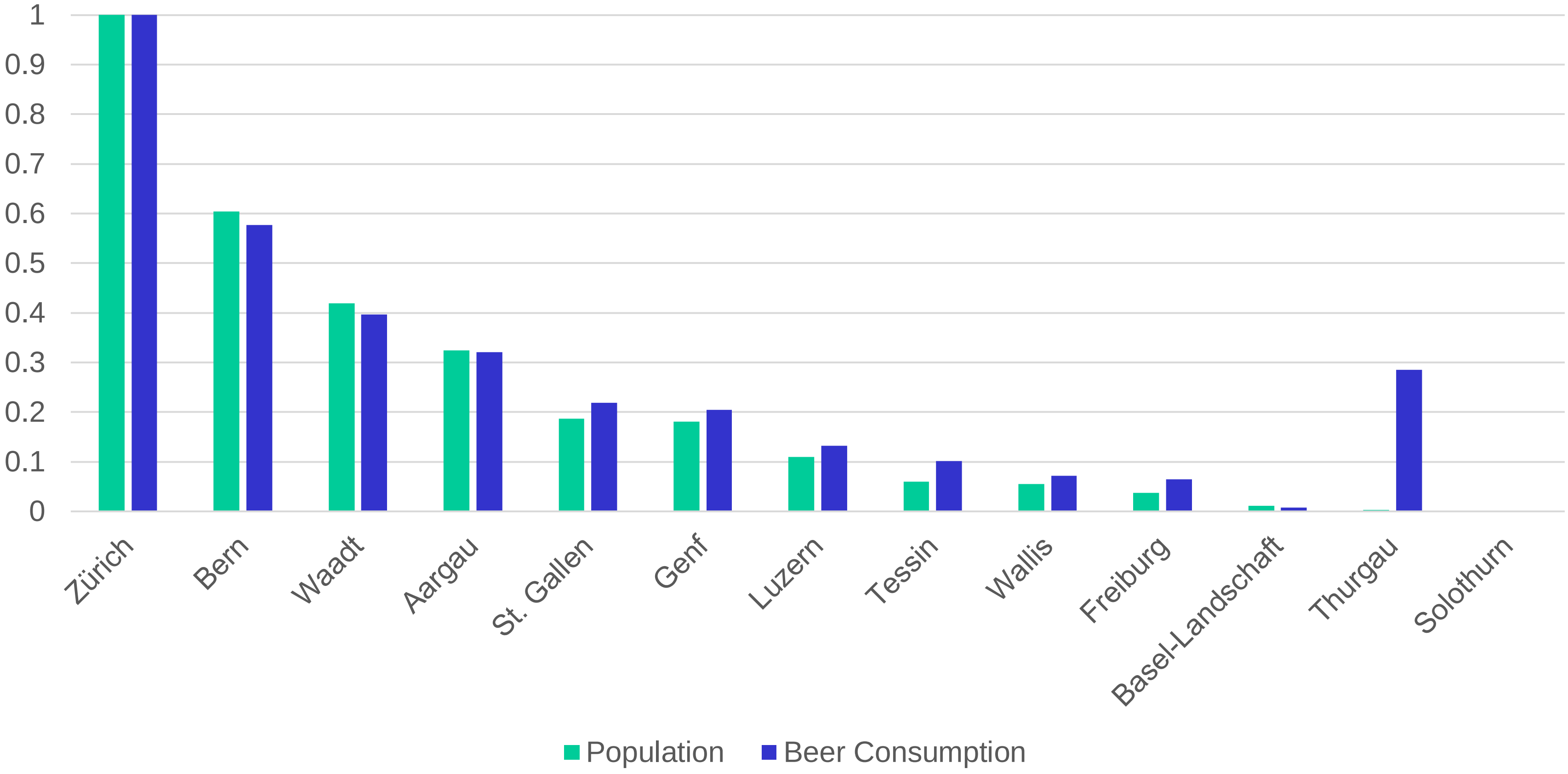
Canton	Beer Consumption
Zürich	12021124
Bern	7232844
Waadt	5195496
Aargau	4342480
St. Gallen	3195642
Genf	3031348
Luzern	2212584
Tessin	1855445
Wallis	1522087
Freiburg	1440590
Basel-Landschaft	807930
Thurgau	3936703
Solothurn	716128

Data Normalization

Goal: change the values of numeric columns to a common scale, without distorting differences in the ranges of values



Data Normalization



Data Normalization

- Linear normalization

$$f_{lin}(v) = \frac{v - min}{\max - min}$$

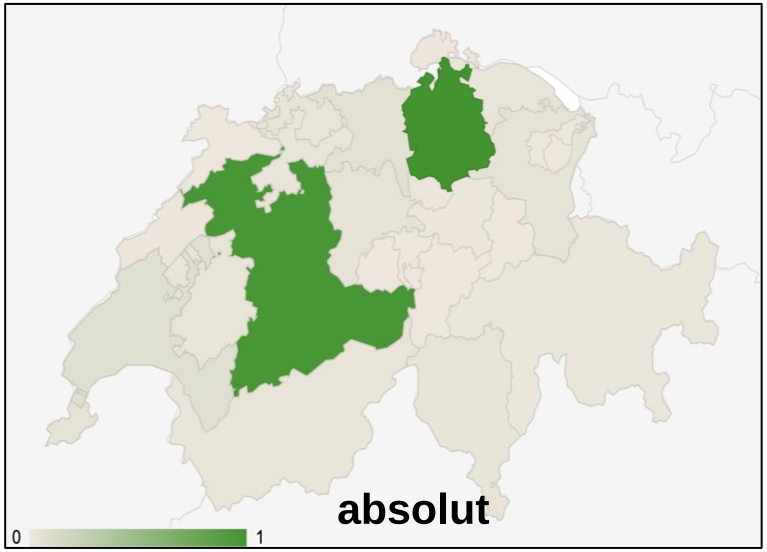
- Square root normalization

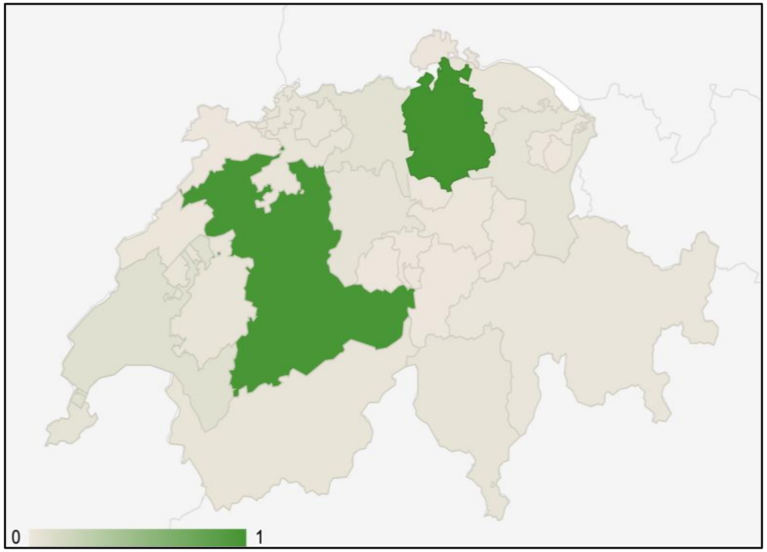
$$f_{sq}(v) = \frac{\sqrt{v} - \sqrt{min}}{\sqrt{max} - \sqrt{min}}$$

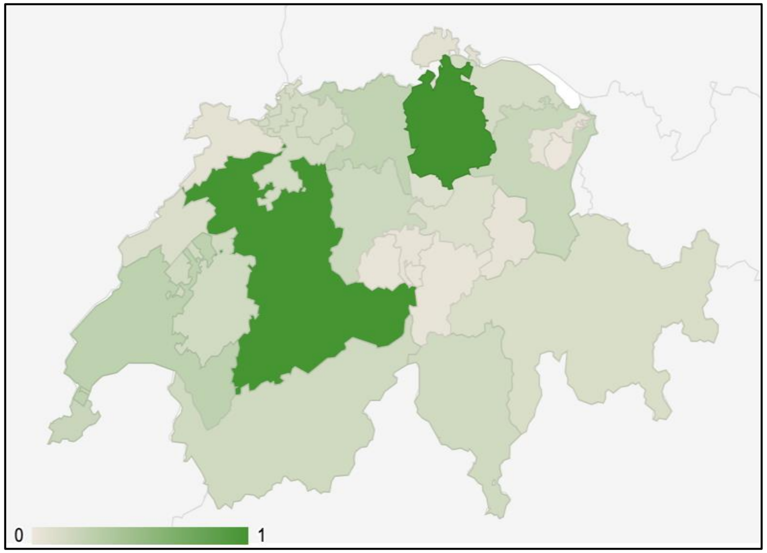
- Logarithmic normalization

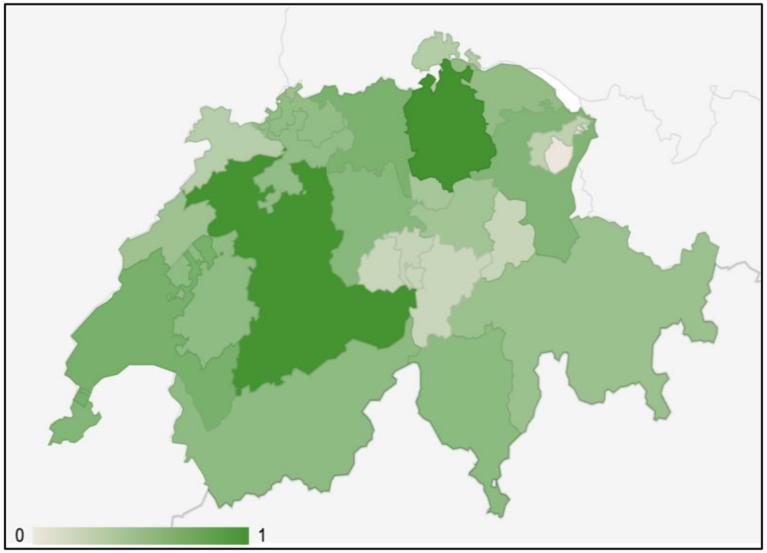
$$f_{ln}(v) = \frac{\ln(v) - \ln(\min)}{\ln(\max) - \ln(\min)}$$

TASK (2 min): The 3 unlabeled maps correspond to linear, square and logarithmic normalization. Which is which?









Lessons Learned

- KDD is the process of (semi-) automatic extraction of knowledge from databases
- KDD handles categorical and numerical data
- Data needs to be cleaned before analysis
- Data cleaning includes filling missing values, handling noisy data and normalization

Questions and Answers

Zürcher Hochschule
für Angewandte Wissenschaften

Solutions

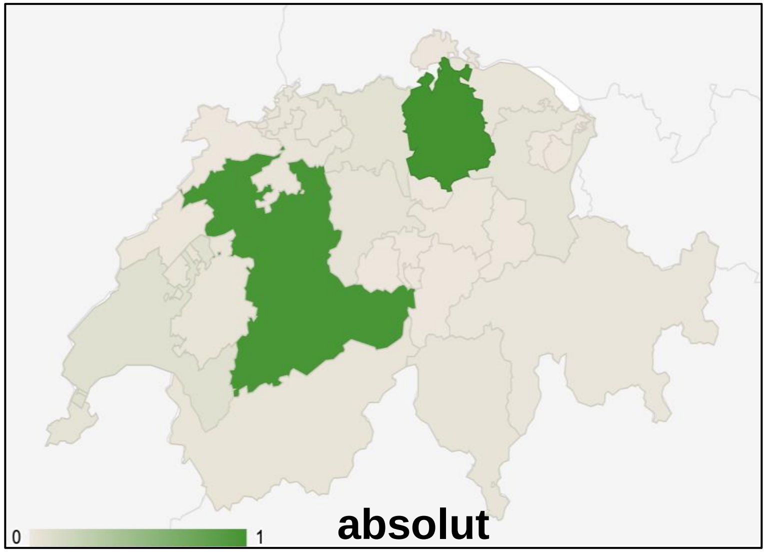
SOLUTION: Data Issues

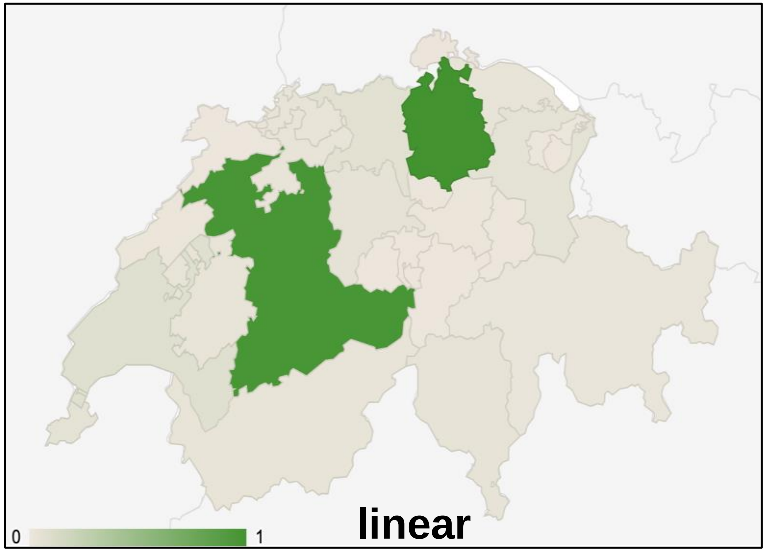
Which data issues can you find in this data?

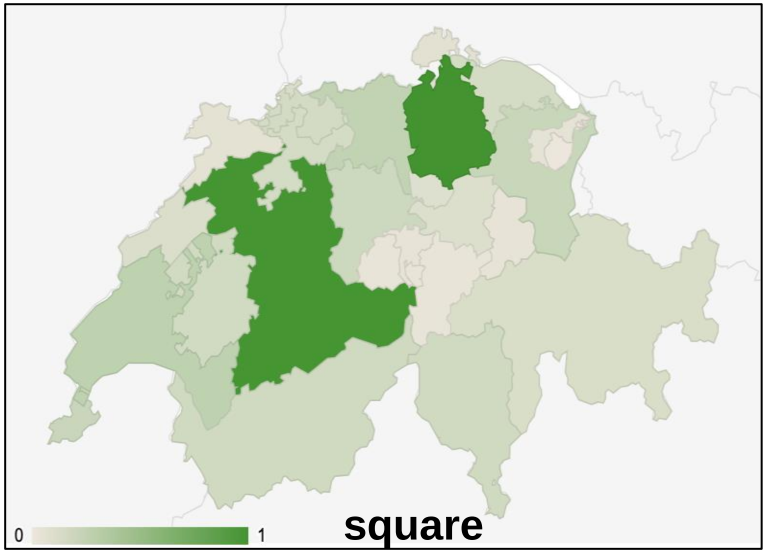
id	title	date	rating	episode	season	length
1	live free or die	01.11.2012	9.3	1	5	43
2	madrigal	06.12.2013	**89**	2	5	48
3	sunset	25.10.2011	9.3	6	3	47
4	grilled	20.12.2009	9.3	2	2	46
5	down	**42.12.2009**	8.3	4	2	**333**
6	cancer man	19.10.2008	8.3	4	1	48
7	**live fre or die**	01.11.2012	9.3	1	5	43

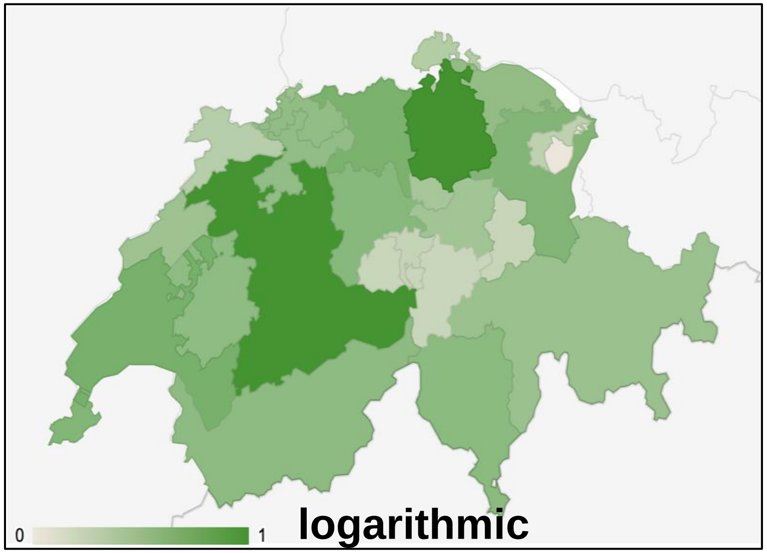
firstname	lastname	bdate	age	gender	phone
bryan	cranston	03-07-1956	65	male	**999-99999**
aaron	paul	**27-08-1979**	**24**	**0**	777-53474
anna, gunn	gunn	08-11-1968	52	female	040-15627
betsy	brandt	03-14-1973	47	**femalle**	333-49858

SOLUTION: The 3 unlabeled maps correspond to linear, square and logarithmic normalization. Which is which?








0
1
logarithmic