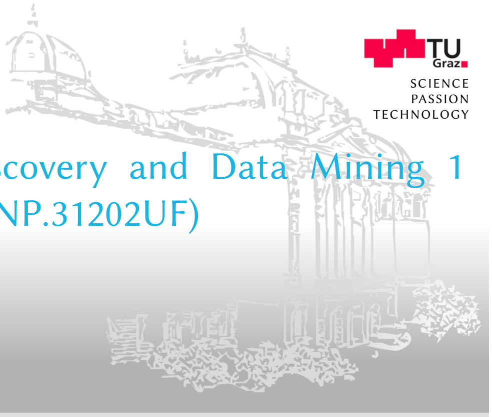


SCIENCE
PASSION
TECHNOLOGY

Knowledge Discovery and Data Mining 1 (INP.31101UF/INP.31202UF)

Roman Kern <rkern@tugraz.at>

Version 2.1.0

> **Motivation**:
Working with data is a key competence, which in today's work is of high importance.
> **Goal**:
This course aims at providing the key elements of working with data and extracting valuable information out of it.

Knowledge Discovery and Data Mining 1 (INP.31101UF/INP.31202UF)

Outline

1. Motivation & Introduction
2. Historic Example
3. Contemporary Example
4. Theoretical Part (VO)
5. Practical Part (KU)
6. Data Team

Motivation & Introduction

Why this course? Why this name? What to expect?

Motivation & Introduction

Goals of the course

The overall goal of KDDM1 and related courses is to learn how to discover patterns in data, and how to model the data. We aim to discover patterns that are:

i **Valid**: hold for new data with high probability

ii **Useful**: we can base further actions on them

iii **Unexpected**: non-obvious

iv **Understandable**: humans can interpret them

Data Science

Data science popularised by Peter Naur in the book “Concise Survey of Computer Methods” (1974) [1]

> *A basic principle of data science is this: The data representation must be chosen with due regard to the transformation to be achieved and the data processing tools available. This stresses the importance of concern for the characteristics of the data processing tools.* [2]

> Go beyond statistics.
> Skills from multiple disciplines are needed.
> Development of proper tools for data representation and processing.
> See http://www.naur.com/Conc.Surv.html.

Knowledge Discovery

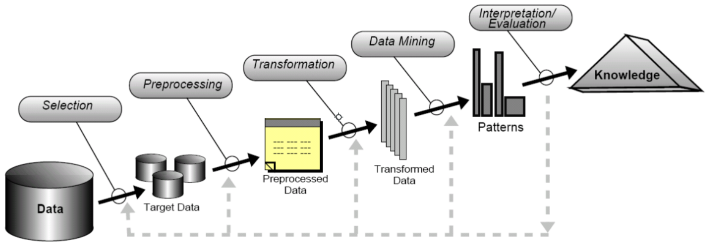
Knowledge discovery from databases (KDD) process proposed by Fayyad (1996)

> *Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.* [3]

> Knowledge discovery process represents a sequence of steps starting with the data and as its final stage providing knowledge.
> Impact of data collection and pre-processing not explicitly stated and stressed.

> Initially, **from databases** was used to highlight the size of the data (today, one would use the term “Big Data” instead).

Knowledge Discovery Process

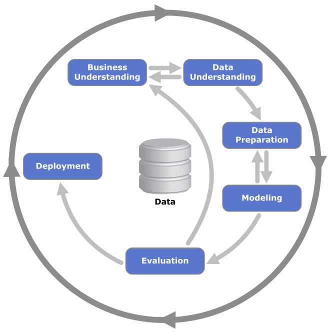


> Multiple (numbered) steps of the knowledge discovery process.
> Importantly, there are also cycles.

> Missing from this overview is the **collection of data**, which also plays an important role (and will be the focus of one lecture).

> This course is strongly aligned with the KDD process.

CRISP-DM



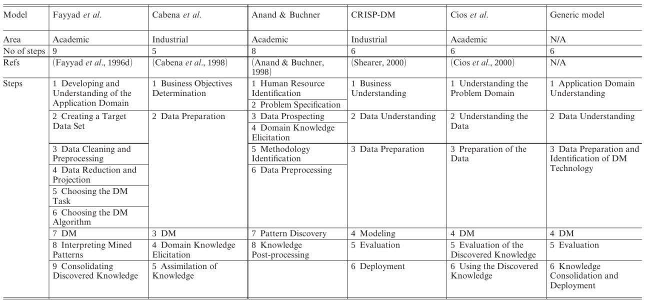
Importance of **data understanding** as highlighted by CRISP-DM (2000) [4]

> KDDM does not consider the business understanding.
> But additionally considers the data collection part.

> The KDD process is more focused on academia, and CRISP-DM has more of a business background.

Motivation & Introduction

Comparisons of Approaches



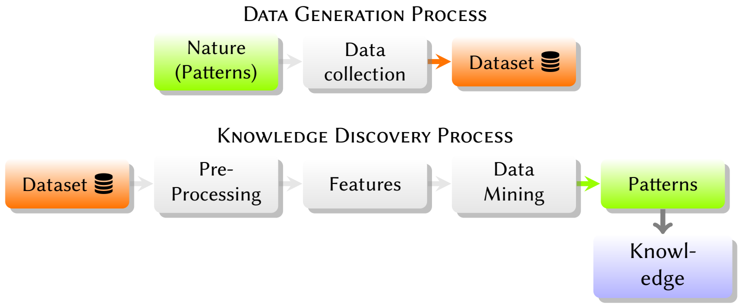
Model	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N/A
No of steps	9	5	8	6	6	6
Refs	(Fayyad *et al.*, 1996d)	(Cabena *et al.*, 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios *et al.*, 2000)	N/A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification 2 Problem Specification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting 4 Domain Knowledge Elicitation	2 Data Understanding	2 Understanding the Data	2 Data Understanding
	3 Data Cleaning and Preprocessing 4 Data Reduction and Projection 5 Choosing the DM Task 6 Choosing the DM Algorithm		5 Methodology Identification 6 Data Preprocessing	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

Comparison of approaches, each having different emphasis [5]

> The KDDM1 course aims to cover the main aspects of all methods.

Motivation & Introduction

Data Processing Pipeline



> **Data science as two-body problem**. The upper row represents an exemplary data generation process, the lower row a typical knowledge discovery / data science process. The aim is to uncover patterns in data, given just a dataset of finite observations. To achieve this, the relation between a data generation process and its resulting data need to be understood.

> One can only loose information (more on this in an upcoming lecture covering the data processing inequality).

Historic Example

Where did knowledge discovery start?

Historic Example

Historic Example

1854 Broad Street cholera outbreak

- Cholera pandemic of 1846-1860
  - Millions death world-wide
- In historical London
  - No proper sewer system in the area
  - Dump into rivers
    - Water companies take water from Thames
    - ... from different sites and varying filtering methods

> See https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Historic Example

- Cholera was not understood
- Multiple theories
  - Miasma (bad air)
    - Common theory
  - Germs
    - Theorised by **John Snow**

> Obviously, in this case the “**business understanding**” is lacking - a.k.a. subject matter expertise / domain knowledge.

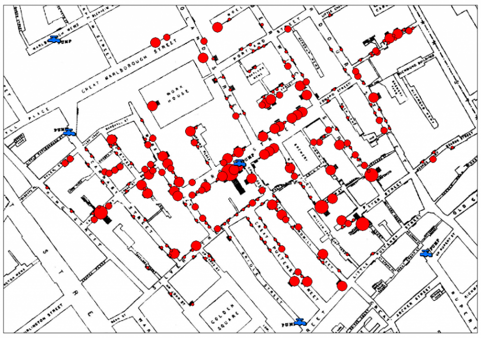
Historic Example

Data Collection

- Motivated by an strong outbreak in 1854
- John Snow started to collect data
  - Talking to people in the area
- Later plotting the data
  - Making use of a dot map
  - Marking the location of water pumps
  - ... and location of deaths

> First step is to collect the necessary data, and to select data.

Historic Example



> Location of the water pumps in blue, deaths in red, with the size encoding the number of deaths.

Historic Example

Exploratory Data Analysis

- Some visual patterns emerge
  - Cluster of death in certain areas
  - Anomalies in certain areas

> To gain the necessary data understanding, visual tools are often helpful.

Historic Example

Pre-Processing

- Mapping the deaths to the closest water pump
  - Via “walking distance”
  - Creating grouping (clustering) of the data

> This step is similar to a part of the k-means clustering algorithm.

Historic Example

Analysis on single patterns

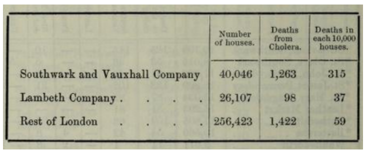
- One cluster is closest to Broad Street water pump
  - → Intervention of removing the handle
  - ... and death count dropped
- One anomaly (fewer deaths) can be mapped to a brewery
  - Workers are assumed to drink beer (heat treatment)
  - ... instead of water

> An **intervention** (optimally a randomized controlled experiment) is often conducted to confirm a hypothesis.

Historic Example

Further analysis on aggregated data

- Found patterns between
  - The water pump’s death counts
  - ... and the responsible water company
  - Other factors (age, wealth, etc.) can be ruled out



	Number of houses.	Deaths from Cholera.	Deaths in each 10,000 houses.
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company	26,107	98	37
Rest of London	256,423	1,422	59

> This can be seen as **double-blind study**.

> Image taken from https://archive.org/details/b28985266/page/86/mode/2up?view=theater

Historic Example

Recap

i Valid: as interventions shown drop in cases

ii Useful: lives saved

iii Unexpected: against mainstream ideas

iv Understandable: now we do have the domain knowledge

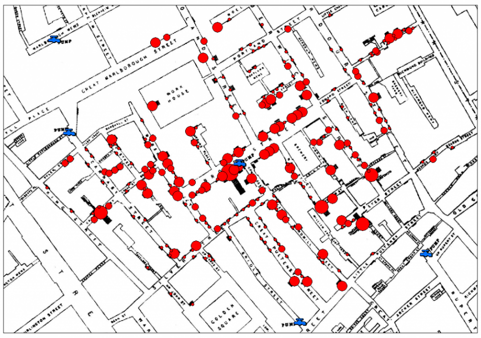
Today tools and methods are far more advanced, as well as the available data.

> ... and this course will present some of these tools and methods.

Contemporary Example

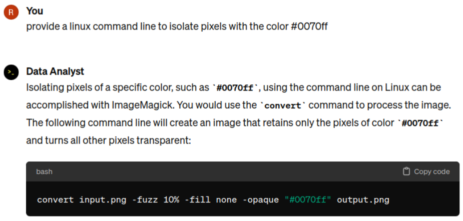
What does data science look like today?

Contemporary Example



> Familiar image, first split wells and cases, based on color.

Contemporary Example



You
provide a linux command line to isolate pixels with the color #0070ff

Data Analyst
Isolating pixels of a specific color, such as `#0070ff`, using the command line on Linux can be accomplished with ImageMagick. You would use the `convert` command to process the image. The following command line will create an image that retains only the pixels of color `#0070ff` and turns all other pixels transparent:

```bash
convert input.png -fuzz 10% -fill none -opaque "#0070ff" output.png
```

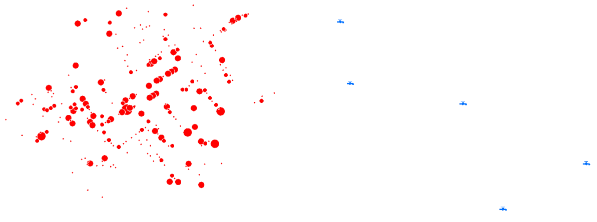
> Asked ChatGPT how to isolate coloured pixels.

“provide a linux command line to isolate pixels with the color #0070ff”

convert cholera-dot-map.png -fuzz 10% -fill white +opaque "#0070ff" faucet.png

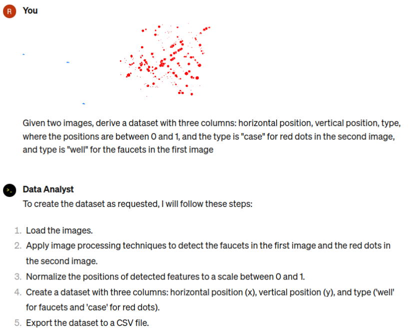
convert cholera-dot-map.png -fuzz 10% -fill white +opaque "#ff0000" cases.png

Contemporary Example



> Isolated cases (red pixel), and wells (blue pixels).

Contemporary Example



You

Given two images, derive a dataset with three columns: horizontal position, vertical position, type, where the positions are between 0 and 1, and the type is "case" for red dots in the second image, and type is "well" for the faucets in the first image

Data Analyst

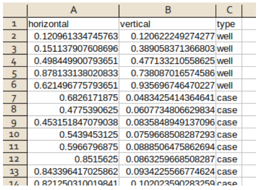
To create the dataset as requested, I will follow these steps:

1. Load the images.
2. Apply image processing techniques to detect the faucets in the first image and the red dots in the second image.
3. Normalize the positions of detected features to a scale between 0 and 1.
4. Create a dataset with three columns: horizontal position (x), vertical position (y), and type ('well' for faucets and 'case' for red dots).
5. Export the dataset to a CSV file.

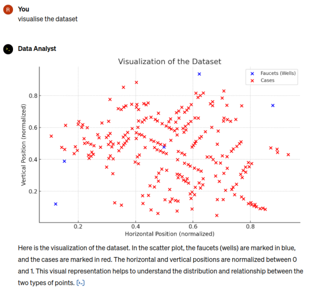
> Convert the two images to a tabular dataset.

Given two images, derive a dataset with three columns: horizontal position, vertical position, type, where the positions are between 0 and 1, and the type is "case" for red dots in the second image, and type is "well" for the faucets in the first image

Contemporary Example



	A	B	C
1	horizontal	vertical	type
2	0.120961334745763	0.120622249274277	well
3	0.151137907608696	0.389058371366803	well
4	0.498449900793651	0.477133210558625	well
5	0.878133138020833	0.738087016574586	well
6	0.621496775793651	0.935696746470227	well
7	0.6826171875	0.0483425414364641	case
8	0.4775390625	0.0607734806629834	case
9	0.453151847079038	0.0835848949137096	case
10	0.5439453125	0.0759668508287293	case
11	0.5966796875	0.0888506475862694	case
12	0.8515625	0.0863259668508287	case
13	0.843396417025862	0.0934225566774624	case
14	0.821250310019841	0.102023590283259	case



You

visualise the dataset

Data Analyst

Visualization of the Dataset
Faucets (Wells)
Cases
0.8
0.6
0.4
0.2
Vertical Position (normalized)
0.2
0.4
0.6
0.8
Horizontal Position (normalized)

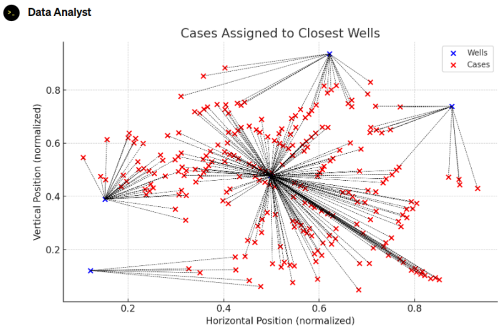
Here is the visualization of the dataset. In the scatter plot, the faucets (wells) are marked in blue, and the cases are marked in red. The horizontal and vertical positions are normalized between 0 and 1. This visual representation helps to understand the distribution and relationship between the two types of points. [>_]

> Excerpt of the data, and visualisation from GPT, with the vertical axis flipped.

Contemporary Example

- Extracted the instances from the image
- Next steps
  - Assign the cases to the wells
  - Rank the well w.r.t to the case count

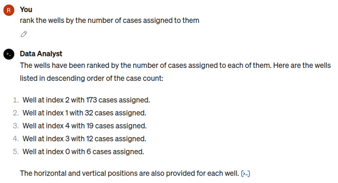
Contemporary Example



Data Analyst

> “cluster the data by assigning each case to the closest well”

Contemporary Example



You

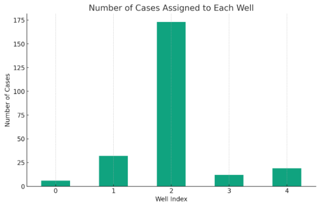
rank the wells by the number of cases assigned to them

Data Analyst

The wells have been ranked by the number of cases assigned to each of them. Here are the wells listed in descending order of the case count:

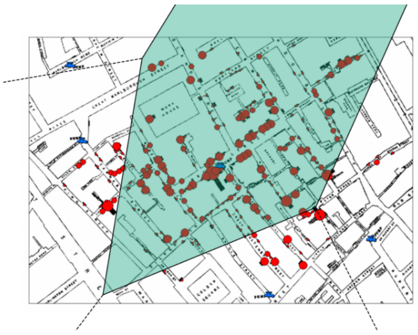
1. Well at index 2 with 173 cases assigned.
2. Well at index 1 with 32 cases assigned.
3. Well at index 4 with 19 cases assigned.
4. Well at index 3 with 12 cases assigned.
5. Well at index 0 with 6 cases assigned.

The horizontal and vertical positions are also provided for each well.



> “rank the wells by the number of cases assigned to them”

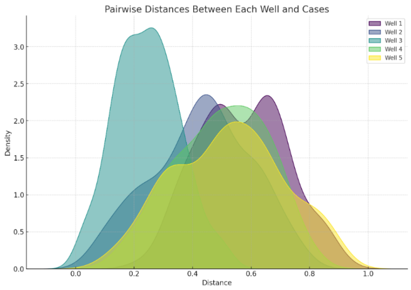
Contemporary Example



> “color the clusters in the map with a voronoi segmentation with different background colors”

Please note that the Voronoi here and the assignments (2 slides ago) do not match.

Contemporary Example



> “provide a plot with the pair wise distances between each well and case with one distribution per well, visually separated by different colours”

Theoretical Part (VO)

Course Organisation

Lecturers

- Name: **Roman Kern**
  - E-mail: rkern@tugraz.at
    - Please use `[KDDM1]` in the subject
- Name: **Denis Helic**
  - E-mail: dhelic@tugraz.at

Language

- Default language: **English**
  - i.e., all materials, homework, etc. are in English
- Communication also possible in German

Lectures

- Lectures are scheduled for
  - Monday, 12:00 - 14:00, c.t.
  - HS i12
- Special sessions
  - Introduction (04.03.2024)
  - Q&A Session for KU (27.05.2024)

https://qnode.eu/ows/hackathon/

Outline I

- Dataset collection
  - Web crawling, databases, surveys
- Visual data science
  - IQR, QQ, etc.
- Statistical data science
  - Correlation, assumptions

Theoretical Part (VO)

Outline II

- Pre-processing
  - Feature extraction
  - Feature engineering
  - Outlier detection
  - Missing value imputation
  - Dataset augmentation

Theoretical Part (VO)

Outline III

- Unsupervised
  - Dimensionality reduction
  - Clustering
- Supervised
  - Prediction (classification, regression)
  - Forecast

Theoretical Part (VO)

Outline IV

- Interactive systems
  - Pattern mining
  - Recommender systems
- Evaluation
- Special topics
  - Class imbalance
  - AutoML
  - XAI

Theoretical Part (VO)

TeachCenter

- Lectures, slides, videos, etc.
- Homework
  - Optional way of examination
    - If participated at homework, no registration for an exam is needed
  - Questions posted as PDF
    - Posted about a month before deadline
  - Answers submitted as PDF
    - Deadline Homework 1: 13.05.2024
    - Deadline Homework 2: 24.06.2024

Practical Part (KU)

Project Organisation

Practical Part (KU)

Study Assistants

- Names: Theresa Doppelhofer, Daniel Hebenstreit
- E-mails: theresa.doppelhofer@student.tugraz.at daniel.hebenstreit@student.tugraz.at
- Please put [KDDM1] in the subject of your e-mail
- Questions also via TeachCenter forum

Practical Part (KU)

Language

- Default language: **English**
  - i.e., all emails, practicals, etc. are in English
- Communication also possible in German

Practical Part (KU)

TeachCenter

- Practical
  - Groups of 4 students
    - Group registration 31.03.
    - Deadline for projects 23.06.
  - Topic: House Price Prediction
  - Each Group gets a slightly different dataset
  - Everything will be uploaded to TC until 31.03

Practical Part (KU)

Project

- For this task, imagine a company/research institution asks you as data scientists to estimate the value of houses
  - We provide details on the houses, such as
    - Location
    - Square meters
    - Previous owner name
    - ...
  - It is your task to build a pipeline to predict the price of houses

Practical Part (KU)

Project cont.

- We created a dataset containing many properties which you will also find in the real world
  - Different types of variables
  - Outliers
  - Missing values
  - ...
- Just like a real data scientist, you will have to evaluate your model yourself!

Practical Part (KU)

Practical submission

- Necessary files for evaluation
  - Group_x.pdf (presentation)
  - Group_x.zip (source code)
- Dropzone (23.06.)
  - https://cloud.tugraz.at/index.php/s/mN2smDjipbjBngb

Practical Part (KU)

Submission Interview

- Everything MUST be uploaded to Dropzone
- Slot selection for oral Q&A
  - TBA
- Presentation (max. 10 minutes)
  - point deduction if you take longer
- Discussion (max. 5 minutes)
- Submission Interview in person

Practical Part (KU)

Remark on Projects

- The goal of the projects
  - ... is not only to optimise (e.g., prediction) performance
- But, to understand
  - ... the data and
  - ... how is has been generated
- And, to understand
  - ... the implications of pre-processing/feature engineering/etc
  - on the results

Data Team

Data Team

Who are we?

- Group of students who are interested in solving Data Science & Machine Learning challenges
- Weekly meetings with discussions and soft drinks!
- If you're interested, join our discord:
  - https://discord.gg/MR69kh3m6v

Thank You!

... for your attention

References I

[1] P. Naur, **Concise survey of computer methods**. Petrocelli Books, 1974.

[2] P. Naur, **Concise survey of computer methods**, [Online; accessed 2022-04-11]. (2001).

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, **From data mining to knowledge discovery in databases**, AI magazine, vol. 17, no. 3, pp. 37–37, 1996.

[4] R. Wirth and J. Hipp, **Crisp-dm: Towards a standard process model for data mining**, Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, Manchester, vol. 1, 2000, pp. 29–40.

[5] L. A. Kurgan and P. Musilek, **A survey of knowledge discovery and data mining process models**, The Knowledge Engineering Review, vol. 21, no. 1, pp. 1–24, 2006.