Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Project Study (M25)
Automatize Machine Learning Processes

PROF. DR. FRANK FUCHS-KITTOWSKI

PAUL SCHULZE, M. SC.

Organizational matters

- Timetable: 01.04. – 31.09.2023
- scheduled meetings (every 2 weeks)
- Tools?

GitLab

DS

Zotero

What do we exect!

- Procedure according to CRISP-DM
- Using scientific methods!

(1) Scientific Question
(2) Hypothesis
(3) Experimental design
(4) Prepare the experiment
(5) Conducting the experiment
(6) Evaluation of the experiment
(7) Conclusion

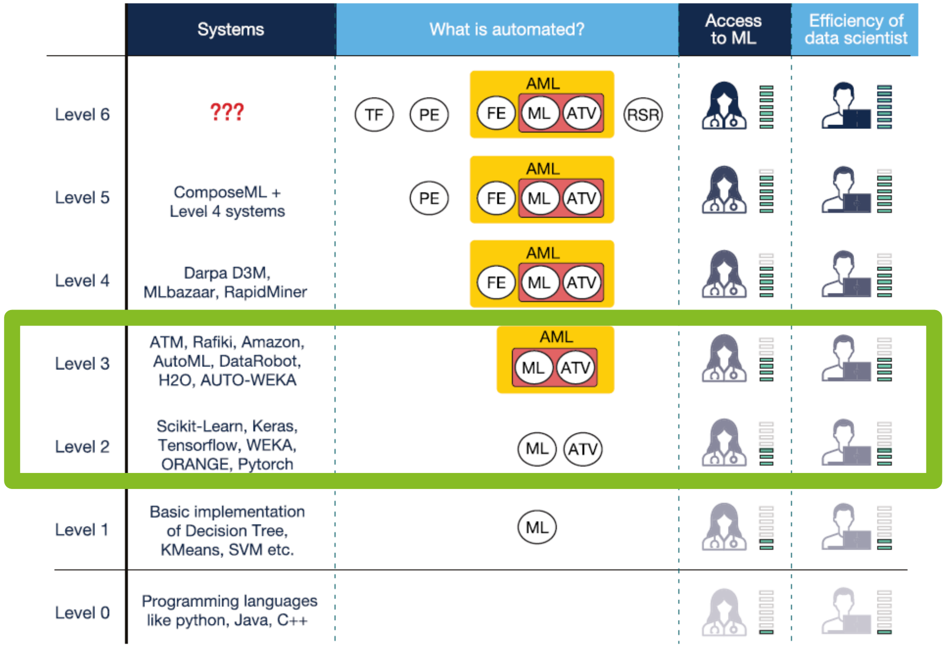
If there is a problem: Tell us!

It’s aloud to fail in the experiment, if you can tell us why and how to avoid it the next time

- **<u>Prepare a short summary to every meeting with results und questions!</u>**
- A scientific report of 20 – 50 pages (with the points above)

What do you expect?

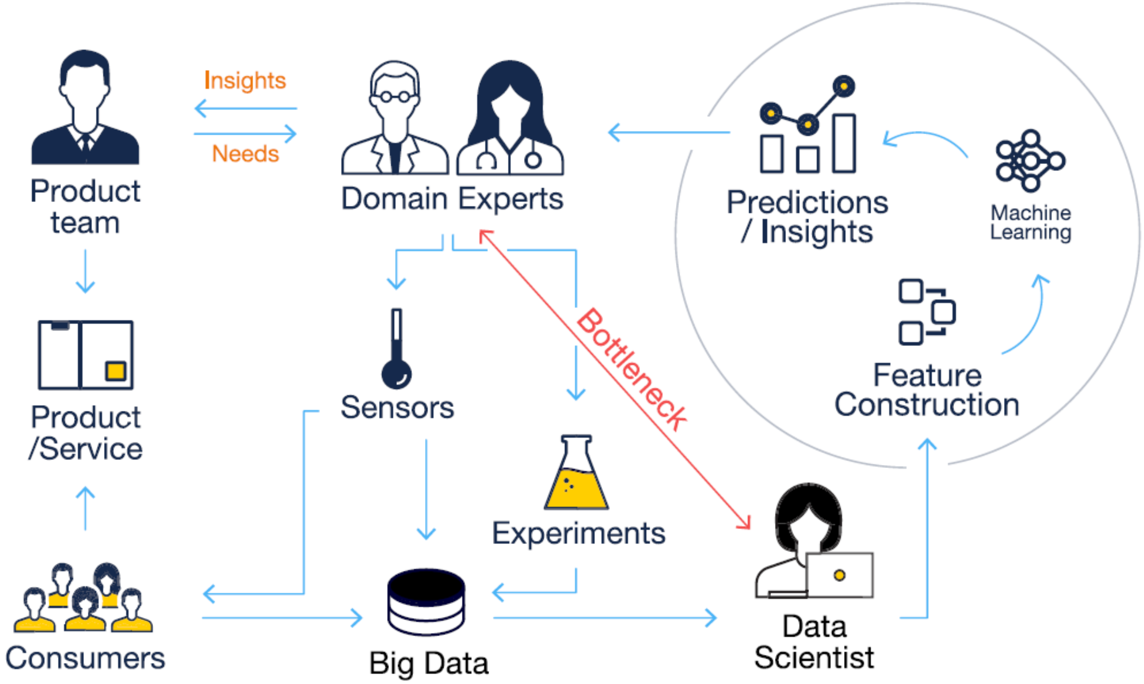
- „To become an expert in AutoML“
- What do you want to learn?



	Systems	What is automated?	Access to ML	Efficiency of data scientist
Level 6	???	TF PE AML FE ML ATV RSR		
Level 5	ComposeML + Level 4 systems	PE AML FE ML ATV		
Level 4	Darpa D3M, MLbazaar, RapidMiner	AML FE ML ATV		
Level 3	ATM, Rafiki, Amazon, AutoML, DataRobot, H2O, AUTO-WEKA	AML ML ATV		
Level 2	Scikit-Learn, Keras, Tensorflow, WEKA, ORANGE, Pytorch	ML ATV		
Level 1	Basic implementation of Decision Tree, KMeans, SVM etc.	ML		
Level 0	Programming languages like python, Java, C++			

What can you expect from us

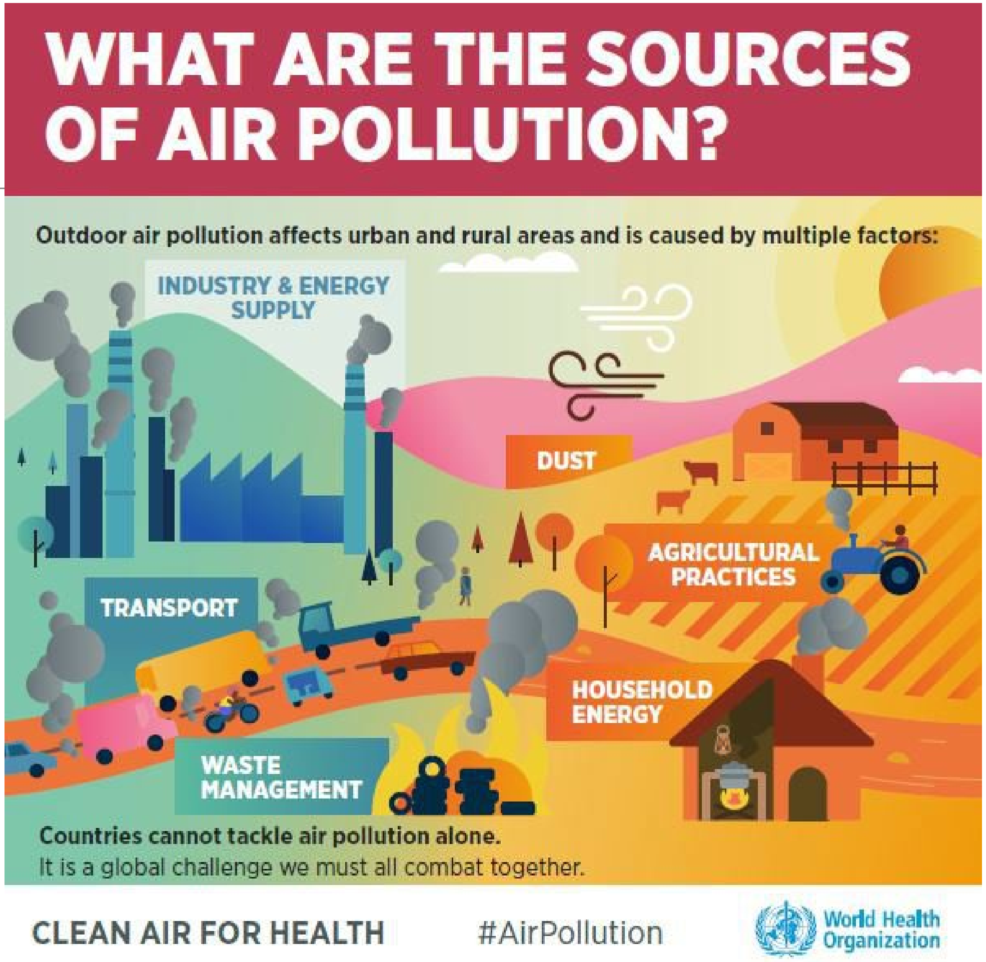
- We are the experts for air pollution
- We hopefully provide you with data
- If you have a problem, talk to us!



Air pollution

- About 4.2 Mio. person died prematurely in 2016 (WHO)
- cause respiratory diseases such as asthma, COPD, lung cancer, …
- many possible pollutants!

➔ Air pollution is a major environmental health threat



WHAT ARE THE SOURCES OF AIR POLLUTION?

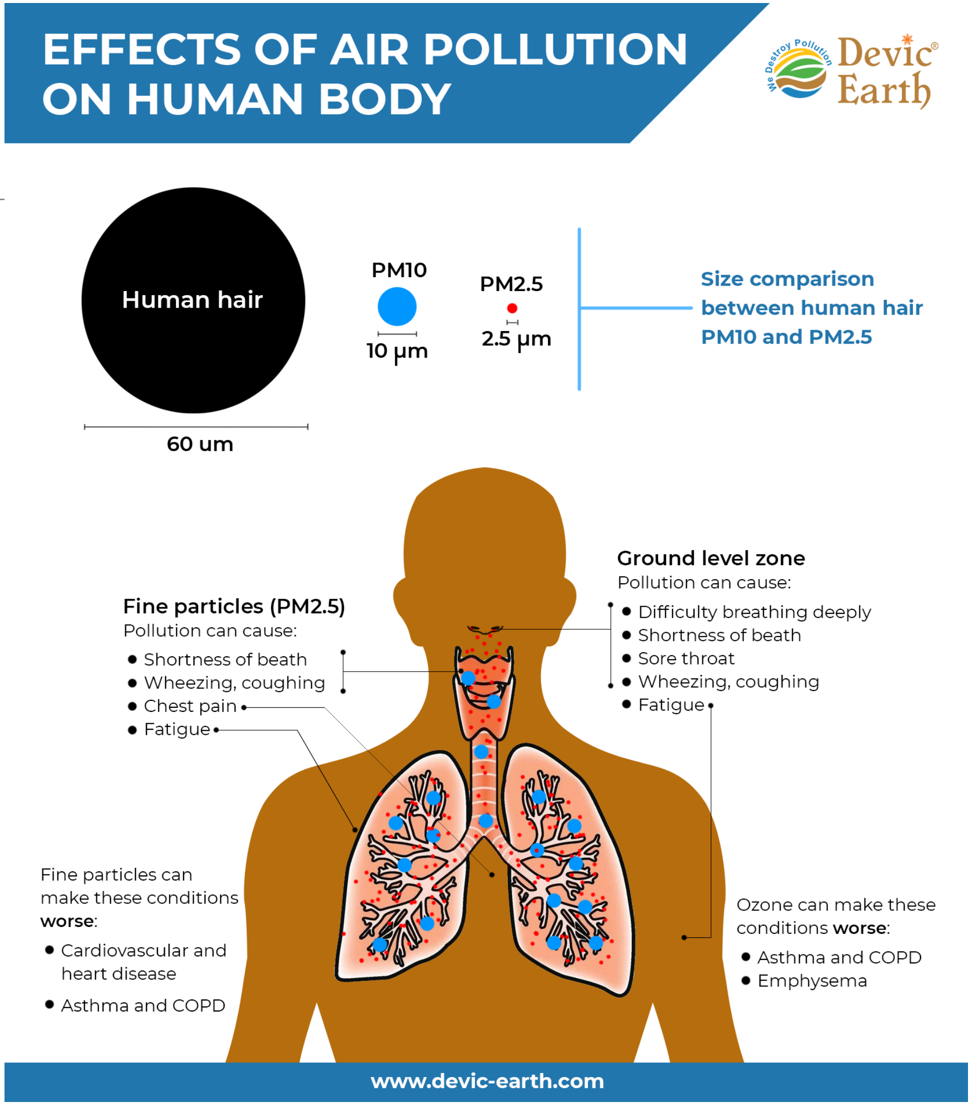
Outdoor air pollution affects urban and rural areas and is caused by multiple factors:

Countries cannot tackle air pollution alone.
It is a global challenge we must all combat together.

CLEAN AIR FOR HEALTH #AirPollution World Health Organization

PM10 (particulate matter)

- Particles with diameter **≤ 10 µm**
- no consideration of the chemical substance
- could deeply penetrate the lungs
- Sources:
  - dust from unsealed roads
  - smoke from fires
  - sea salt
  - car and truck exhausts
  - industry



EFFECTS OF AIR POLLUTION ON HUMAN BODY

Fine particles (PM2.5)
Pollution can cause:

- Shortness of beath
- Wheezing, coughing
- Chest pain
- Fatigue

Ground level zone
Pollution can cause:

- Difficulty breathing deeply
- Shortness of beath
- Sore throat
- Wheezing, coughing
- Fatigue

Fine particles can make these conditions **worse**:

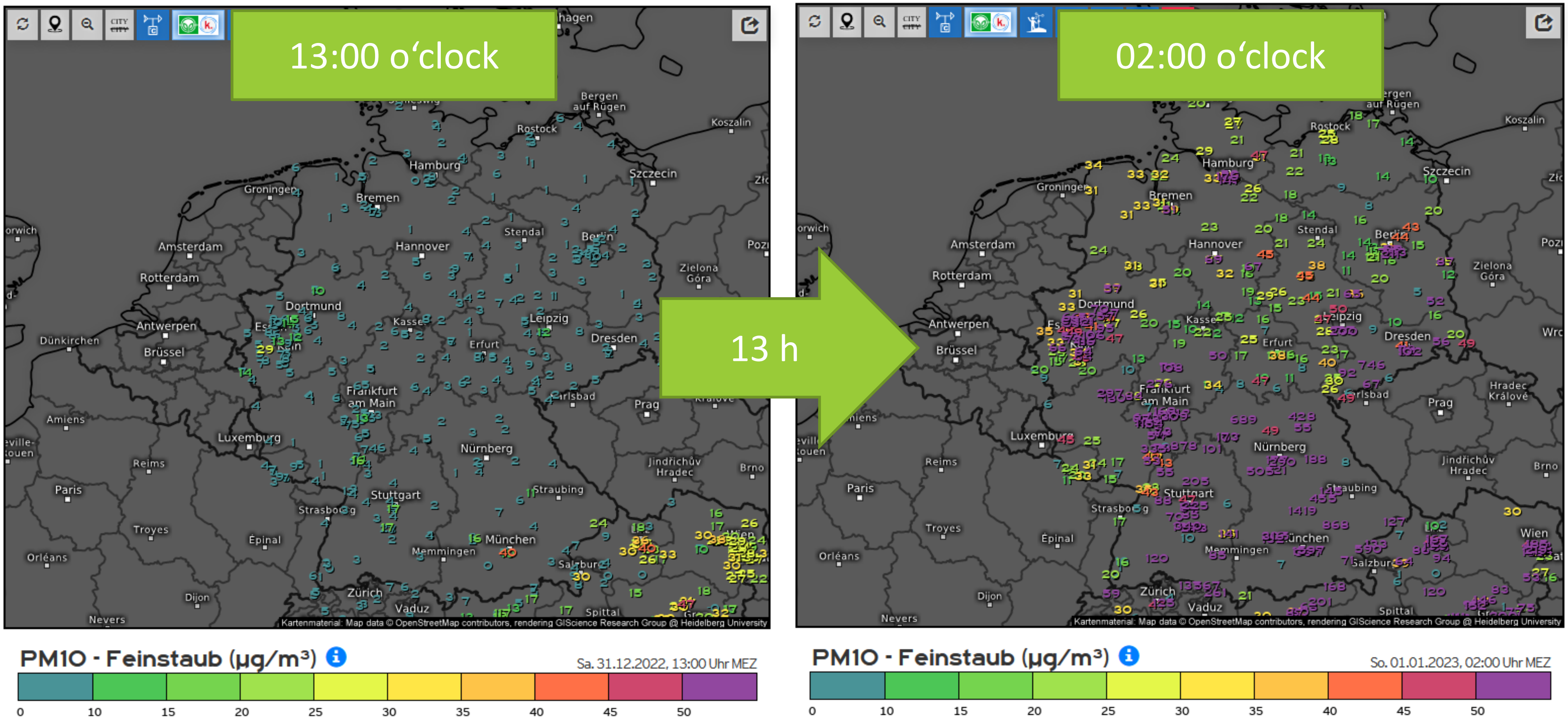
- Cardiovascular and heart disease
- Asthma and COPD

Ozone can make these conditions **worse**:

- Asthma and COPD
- Emphysema

www.devic-earth.com

Silvester: Fireworks!



13:00 o‘clock

13 h

02:00 o‘clock

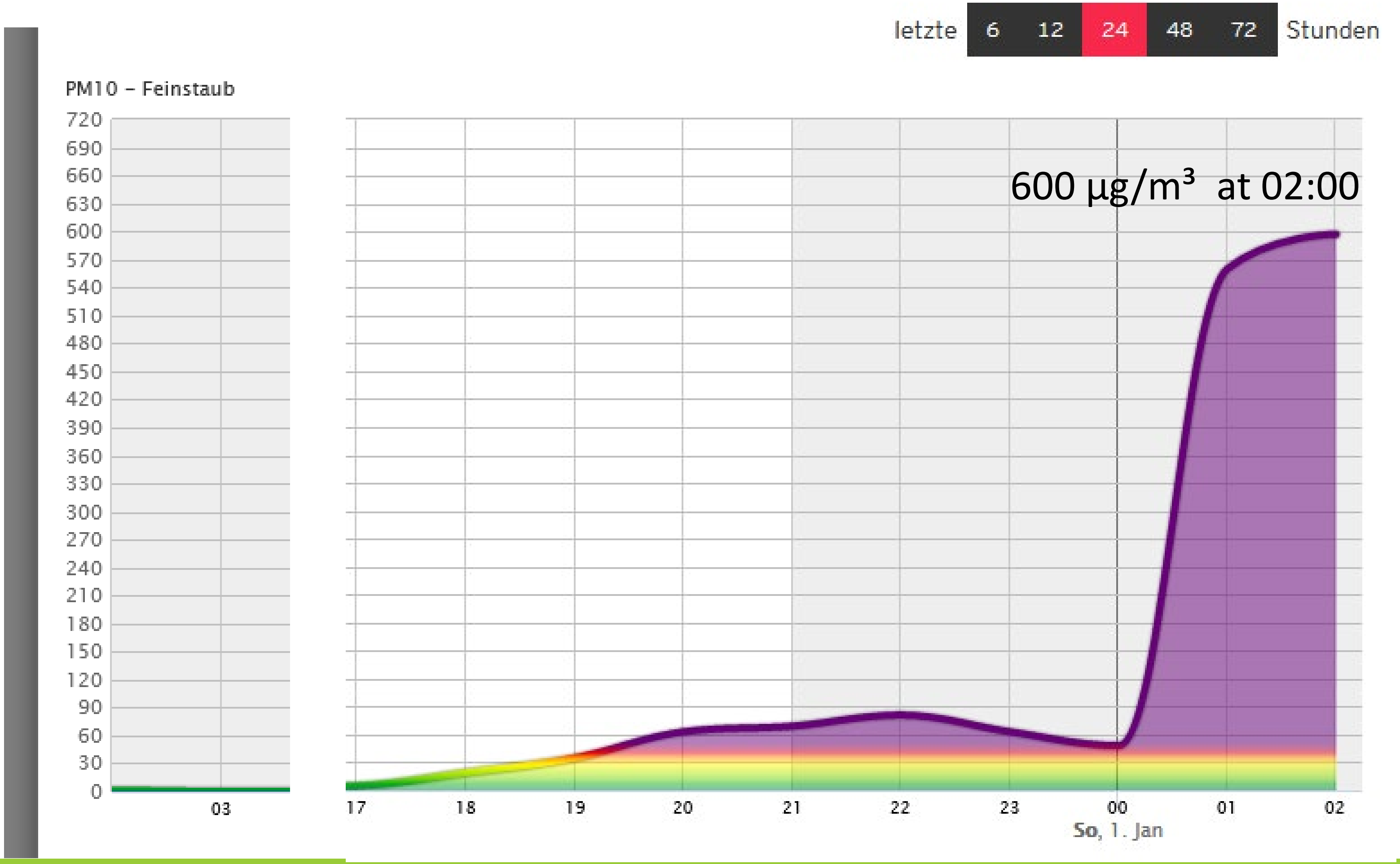
PM10 - Feinstaub (µg/m³)

Sa. 31.12.2022, 13:00 Uhr MEZ

So. 01.01.2023, 02:00 Uhr MEZ

Kartenmaterial: Map data © OpenStreetMap contributors, rendering GIScience Research Group @ Heidelberg University

München / Johanneskirchen



600 µg/m³ at 02:00

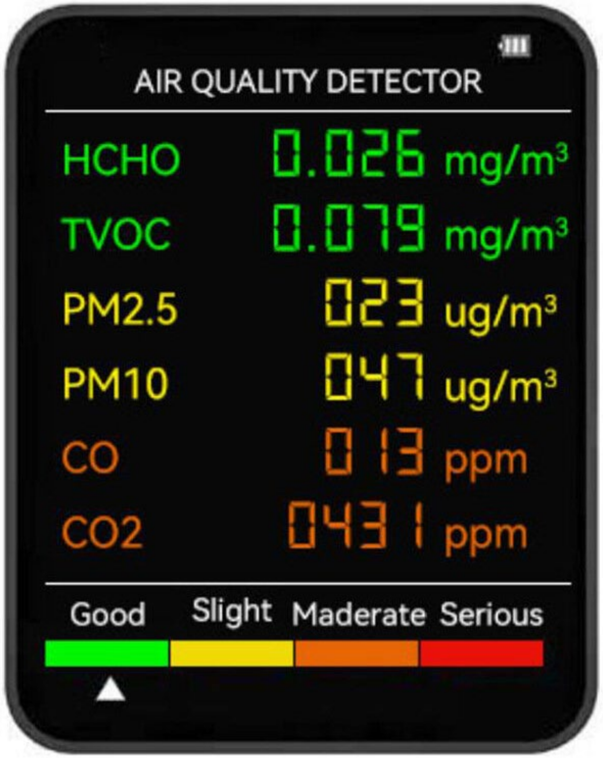
What can you do on a personal level when PM10 levels are high?

You can‘t see them!

➔ Avoid the contract!

- wear a mask
- avoid being outside
- close your windows and door

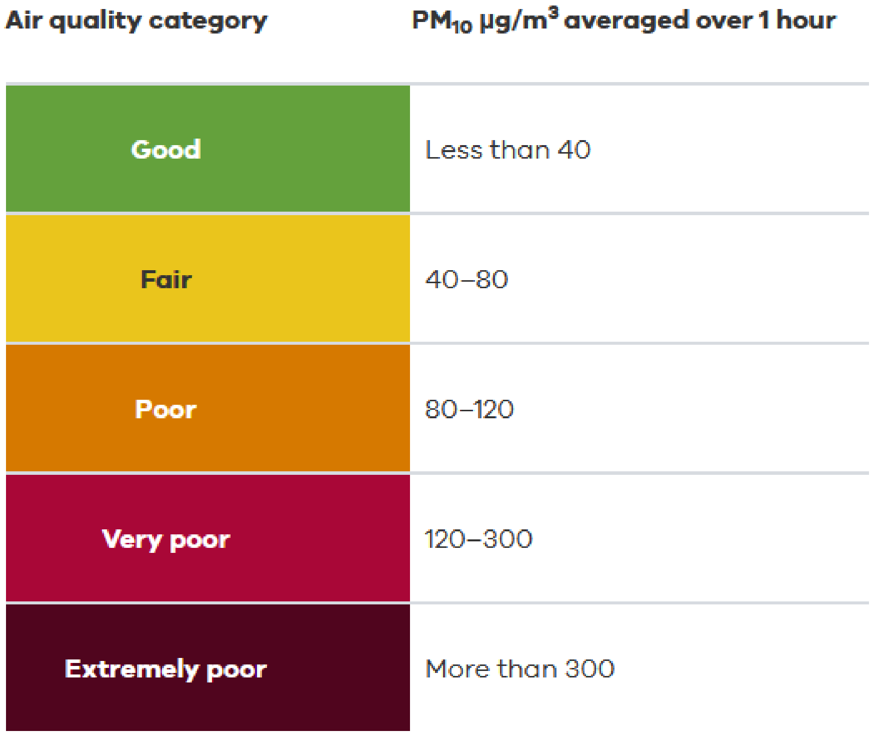
- But:
  How can you tell that the air quality is bad?



Possible scientific questions?

- Can we **forecast** the **concentration** of PM10?
- Or can we **forecast**, when the concentration in the next hour is **exceeding threshold**?
- Or can we **forecast**, when the daily **threshold** is exceeding at the **next day**?
- …

The question leads to the methods used for analysis!



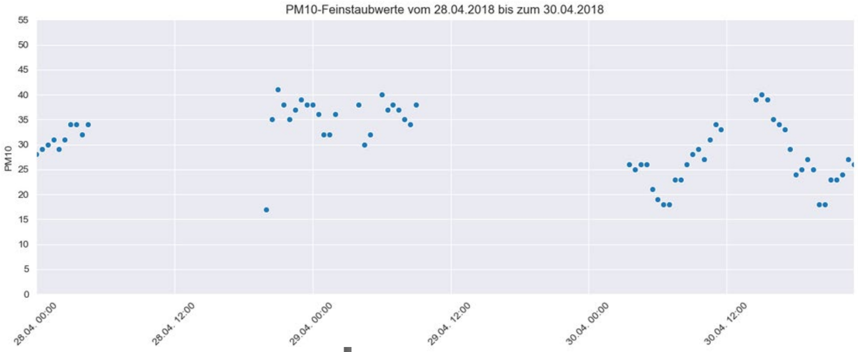
Air quality category	PM_{10} $\mu g/m^3$ averaged over 1 hour
Good	Less than 40
Fair	40–80
Poor	80–120
Very poor	120–300
Extremely poor	More than 300

Threshold for Germany:
50 µg/m³ within one day
(24 Values per hour where averages to one Value per day)

First step: Data exploration!

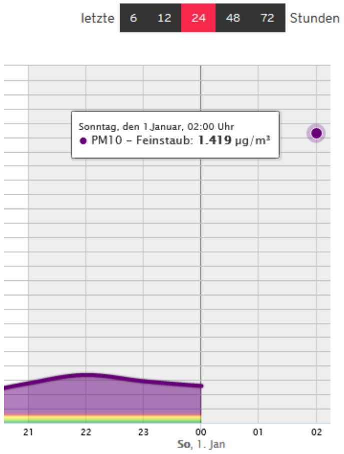
Do we have the data to answer the question?

Is there missing data?



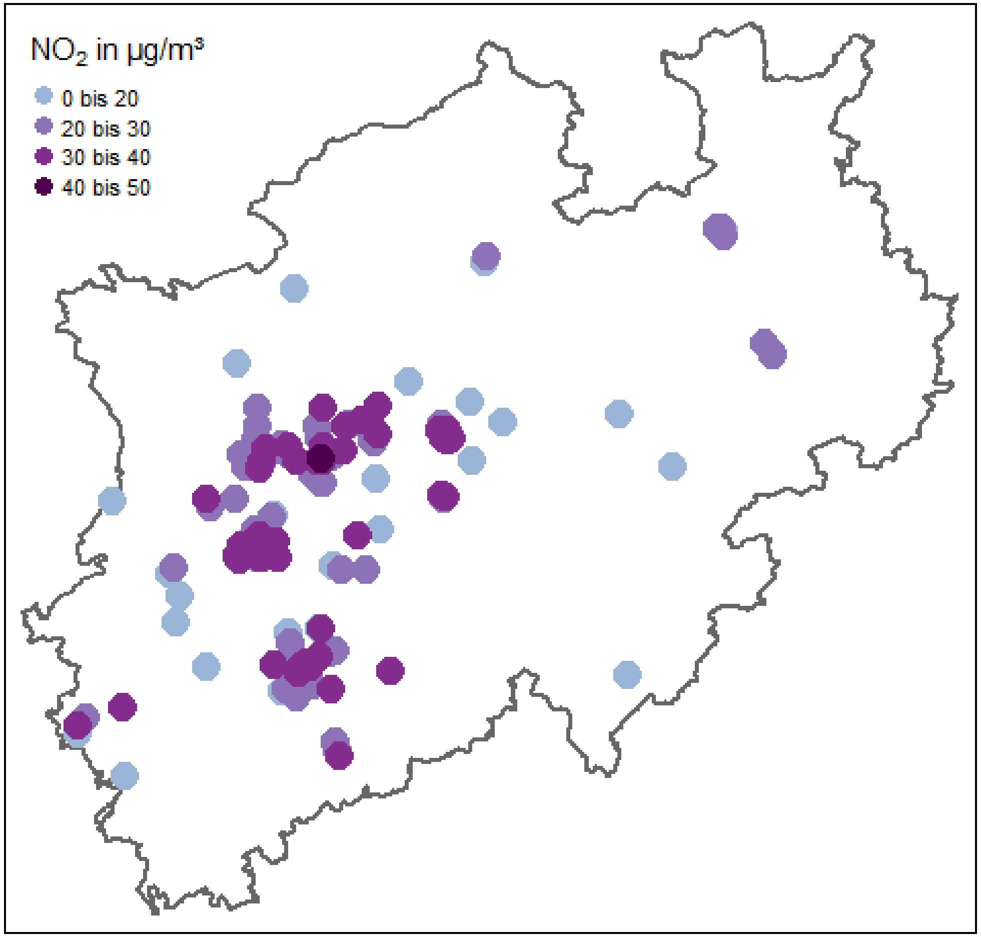
Error in the data?

1419 µg/m³ at 02:00 on 01.01.2023

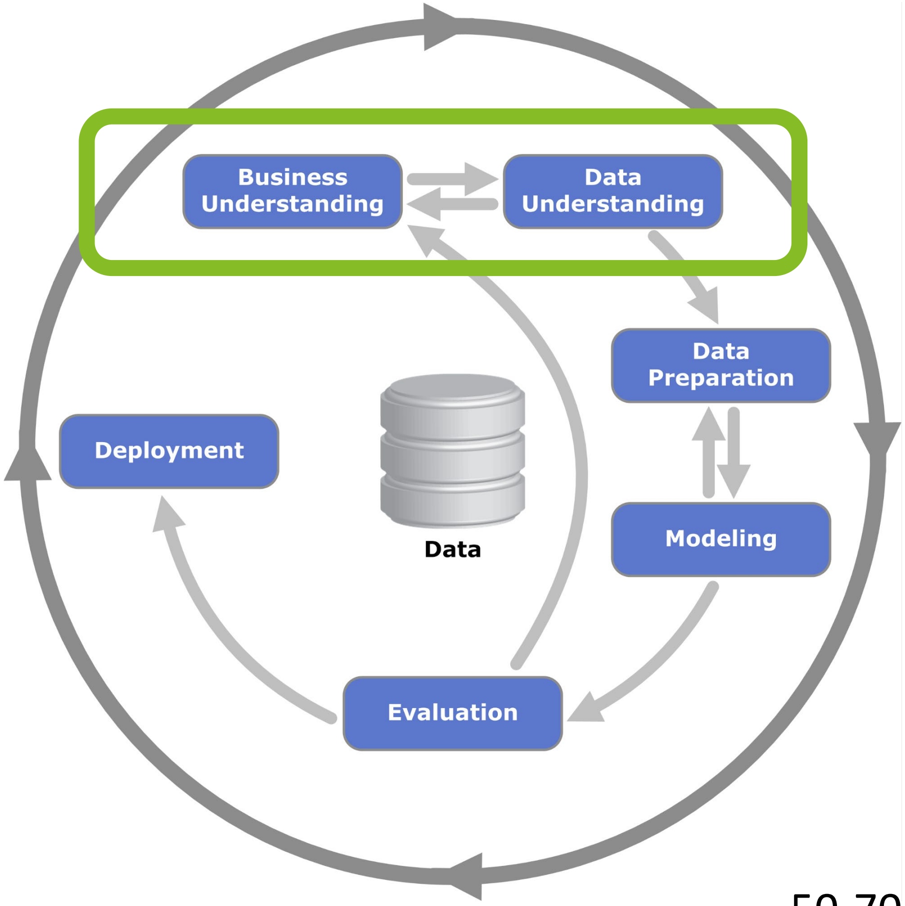


letzte
6
12
24
48
72
Stunden
Sonntag, den 1.Januar, 02:00 Uhr
● PM10 – Feinstaub: 1.419 µg/m³
21
22
23
00
01
02
So, 1. Jan

Monitoring sites in NRW (Nordrhein-Westfalen)



Where are we?



Phase 1: Business Understanding

- Determination of the business problem
- Situation assessment
- Determination of analytical goals
- Preparation of the project plan

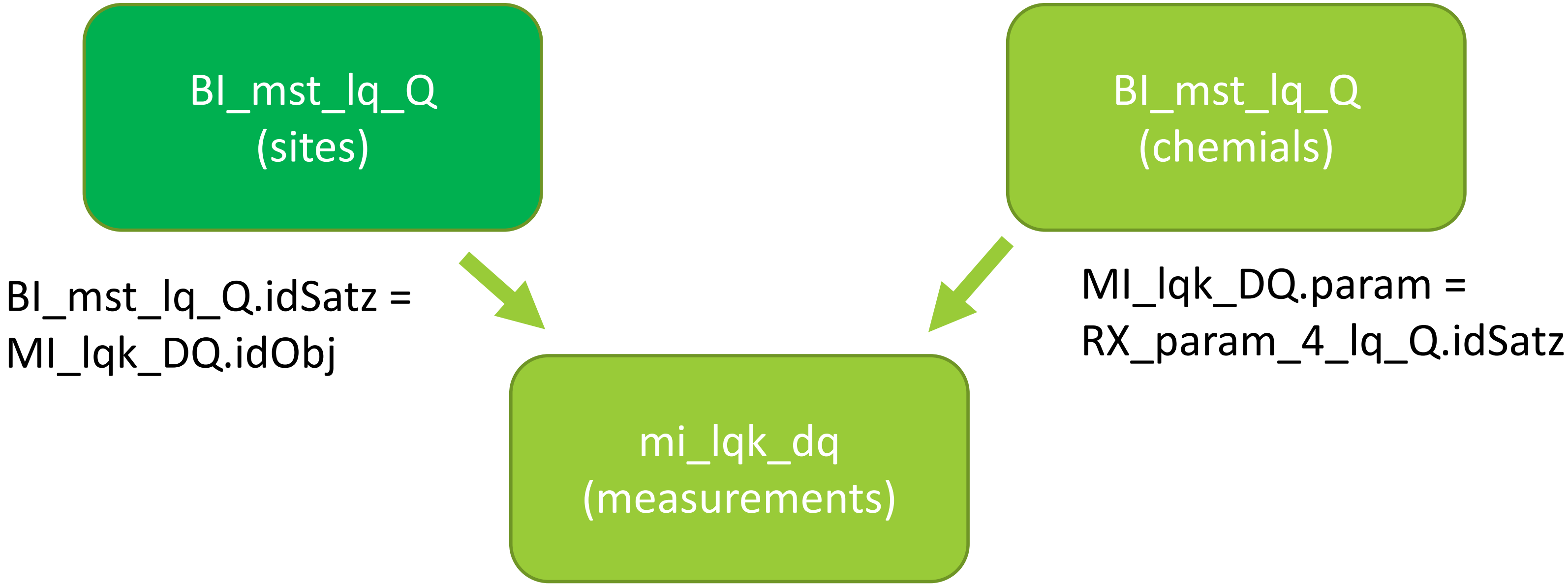
Phase 2: Data Understanding

- collect data
- decribe data
- analyse data
- check data

50-70 % of the time needed for data preparation

Dataset

- data from all monitoring station (173) from NRW
- data is hourly
- **~ 500 Million datasets**



BI_mst_lq_Q
(sites)

BI_mst_lq_Q
(chemials)

BI_mst_lq_Q.idSatz =
MI_lqk_DQ.idObj

MI_lqk_DQ.param =
RX_param_4_lq_Q.idSatz

mi_lqk_dq
(measurements)

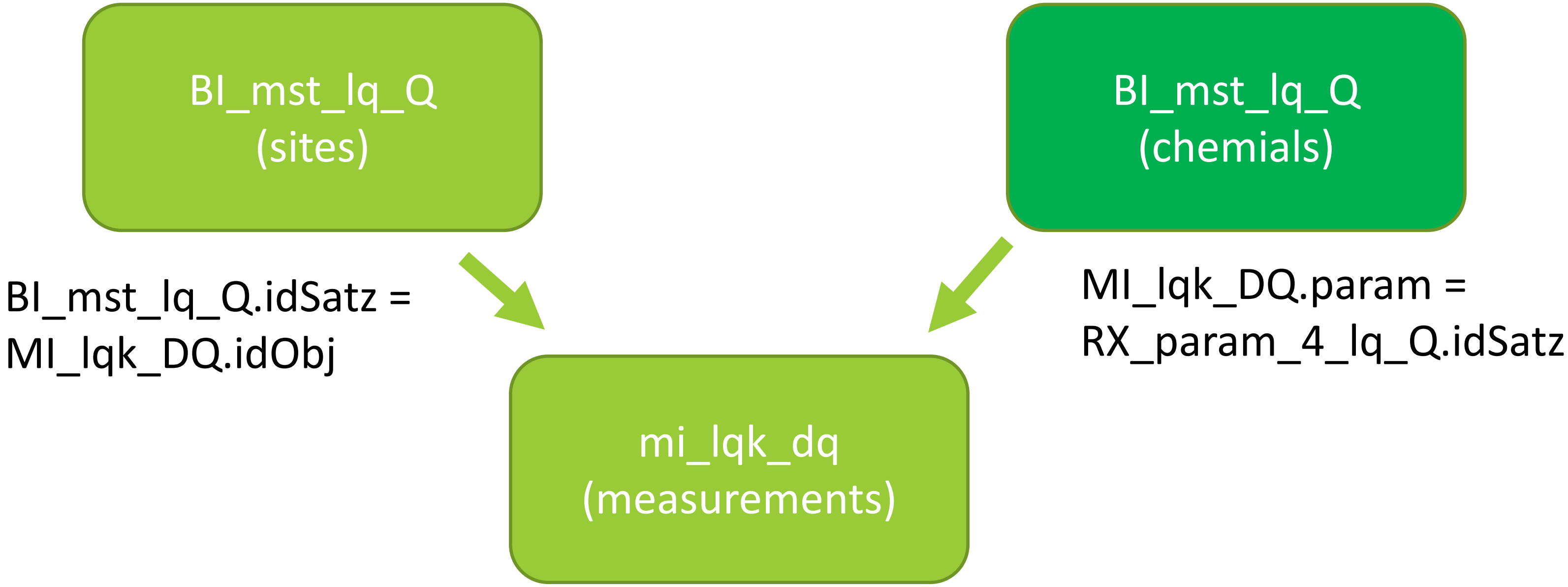
Possible Approach: How to predict PM10 concentration?

- Data is hourly!
- We assume that the concentration of PM10 in the last 6 hours has an influence on the concentration in one hour.
- We assume that we can calculate the future concentration of PM10 from the data of the last 6 hours.
- The scientific question has an influence on the necessary data preparation!

„BI_mst_lq_Q“ description of measurement sites

German	English	Description	Example
idsatz	**IdDataSet**	**identifier**	**2319, 2319, …**
idsrcobj	**idsrcobj**	**site as Identifier** **(Bad Berleburg ➔ BBER)**	**ACBU, ACMI, …**
idsrccls	idsrccls	objtyp as id numner	303211000
namowner	namowner	governmental organizational unit	FB 42 / FB43
idprz		table identifier	
objkey	objkey	same as idsrcobj	ACBU, ACMI, …
objnam	objnam	objectname (adress of measurement site)	„Aachen – Burtscheid“
objbez	objdecs	object decsrition	{ empty }
objtyp	objtyp	type of measurement site (in all cases: air quality measurement sites)	Luftqualitätsmessstelle
objbem	objcom	object comment: Description of the area and surounding	

Datasets



BI_mst_lq_Q
(sites)

BI_mst_lq_Q
(chemials)

BI_mst_lq_Q.idSatz =
MI_lqk_DQ.idObj

MI_lqk_DQ.param =
RX_param_4_lq_Q.idSatz

mi_lqk_dq
(measurements)

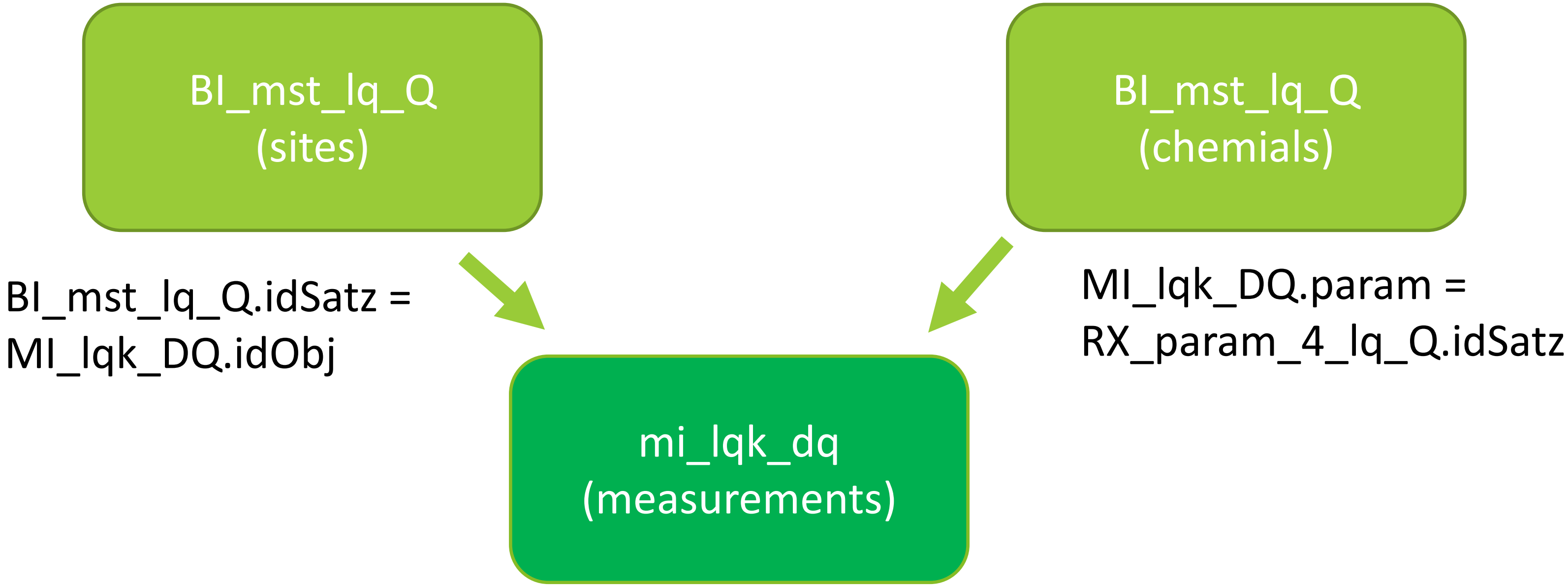
“rx_param_4_lq_q” chemials / reagents

German	English	Description	Example
idsatz	IdDataSet	Identifier	73636, 73627
idcls		table identifier	
idsrcobj	**idsrcobj**	**Multiple ids for measurement parameter** **e.g. 3024 = (chemial, reagent) Pentachlorbiphenyle** **1 = type of measurement** **13 = time interval of measurement** **6 = place of measurement**	**3024_1_13_6**
objkey	objkey	Chemical/reagent as key	PentaCB
objnam	objnam	Chemical/reagent descrition (long form)	Pentachlorbiphenyle
objbez	objdec	Place of measurement	deposition
objtyp	objtyp	chemial group	PCB, BTEX, PAK, …
objbem	objcom	Object comment	{ empty }
einh	unit	measurement unit	µg/m^3*h

“rx_param_4_lq_q” chemials / reagents

German	English	Description	Example
meth_erf	meas_plc	place of measurement	Staubniederschlag (dust collection)
meth_erf	meas_des	place of measurement description	{ description of meth_erf }
code		unknown identifier	{ empty }
t_bez	t_des	time interval description	
p_bez		unknown identifier	

Datasets



BI_mst_lq_Q
(sites)

BI_mst_lq_Q
(chemials)

BI_mst_lq_Q.idSatz =
MI_lqk_DQ.idObj

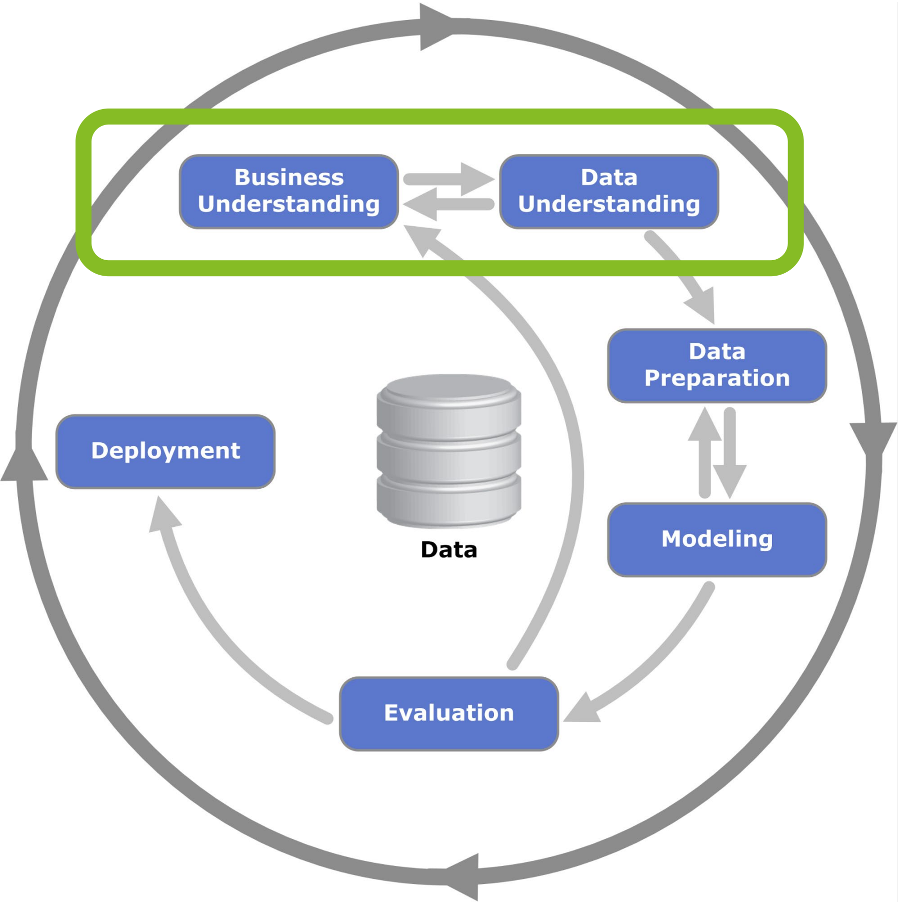
MI_lqk_DQ.param =
RX_param_4_lq_Q.idSatz

mi_lqk_dq
(measurements)

“mi_lqk_dq” measurement values

German	English	Description	Example
idsatz	iddataset	identifier	1598, 1599
idcls		unknown identifier	
idobj	**idobj**	**Measurement site from file „BI_mst_lq_Q“ column „idsatz“**	**372** **(Aachen-Burtscheid)**
param	**param**	**Measured parameter from file „rx_param_4_lq_q” column “idsatz”**	**2053** **(4_1_1_1, Stickstoffmonooxid)**
dtbeg		unknown identifier	
tmbeg	**Date**	**Measurement date**	**2020-01-01**
tmend	**time**		**00:00:00**
prefix	**Prefix**	**Identifiert if value is below detection limit**	**{ < }**
Wert	**Values**	**Measurement value**	**7**
rohwert	Raw values	Raw value from measurement device	1.002154
Nwg	Limit	detection limit	7
Status	State	Is value confimed?	
Freigabe	access	Is values free for access?	
idprz		table identifier	

Next steps



<u>Phase 1: Business Understanding</u>

- **Determination of the business problem (Scientific question?)**
- **Situation assessment (Do you have all the tools needed?)**
- **Determination of analytical goals (what do you want to aquire?)**
- **Preparation of the project plan (your time table)**

<u>Phase 2: Data Understanding</u>

- **describe data (get to know your data set)**
- analyse data (frist statistical analysis)
- check data (is amount of data is sufficient and usable for the analysis?)

ToDos for the next meeting?

Determination of the business problem

- Describe possible Scientific questions or elaborate on the given questions.
- Search for literature

Situation assessment

- Do you have all the tools needed?

Determination of analytical goals

- what do you want to aquire?

Preparation of the project plan

- Create a time table for your project with milestones

Describe data

- get to know your data set
- Write import script for data set and prepare data filtering (cities, pollutants, …)