SPONSORED BY THE
Federal Ministry of Education and Research

MODAL FORSCHUNGSCAMPUS

FORSCHUNGS CAMPUS
öffentlich-private Partnerschaft für Innovationen

Improving quality of OR data sets: Detecting and removing data errors that involve superhuman complexity

Inci Yüksel-Ergün, Thorsten Koch, Janina Zittel

21.04.2023

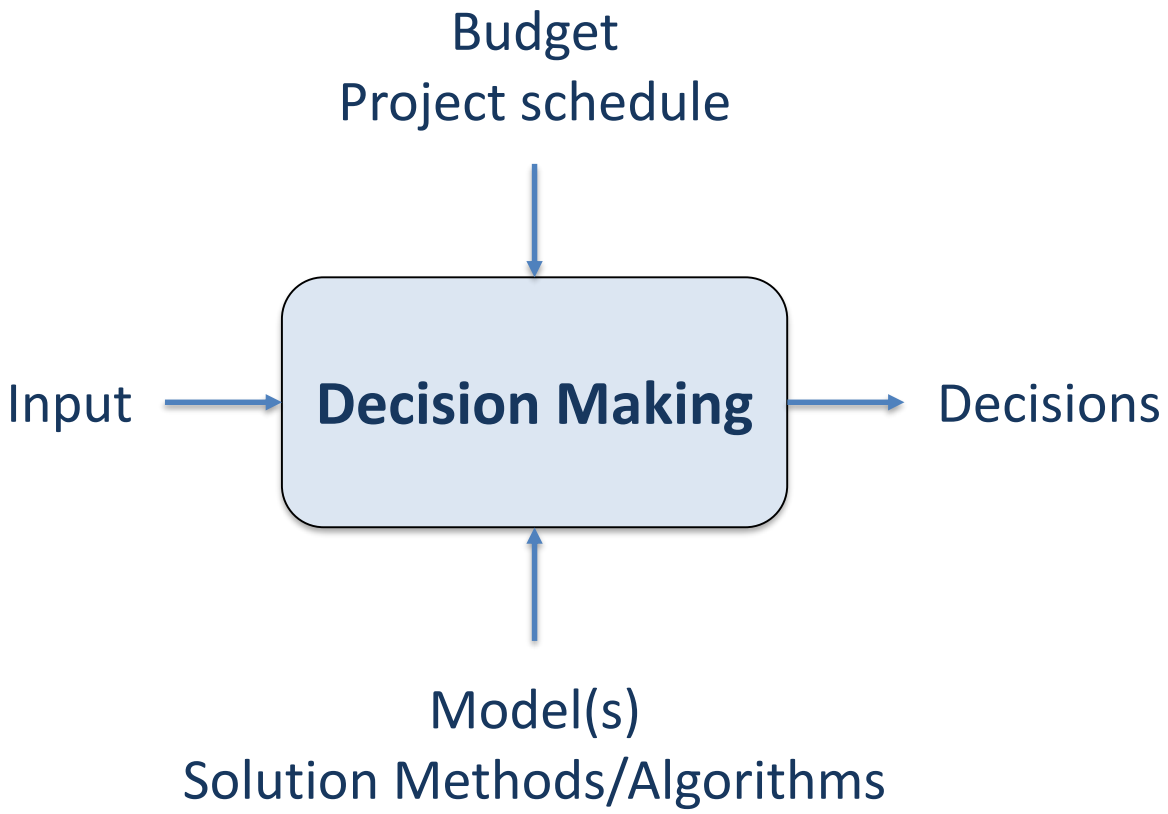
EURO Practitioners' Forum
4th Annual Conference

Chair of
Software and Algorithms for
Discrete Optimization

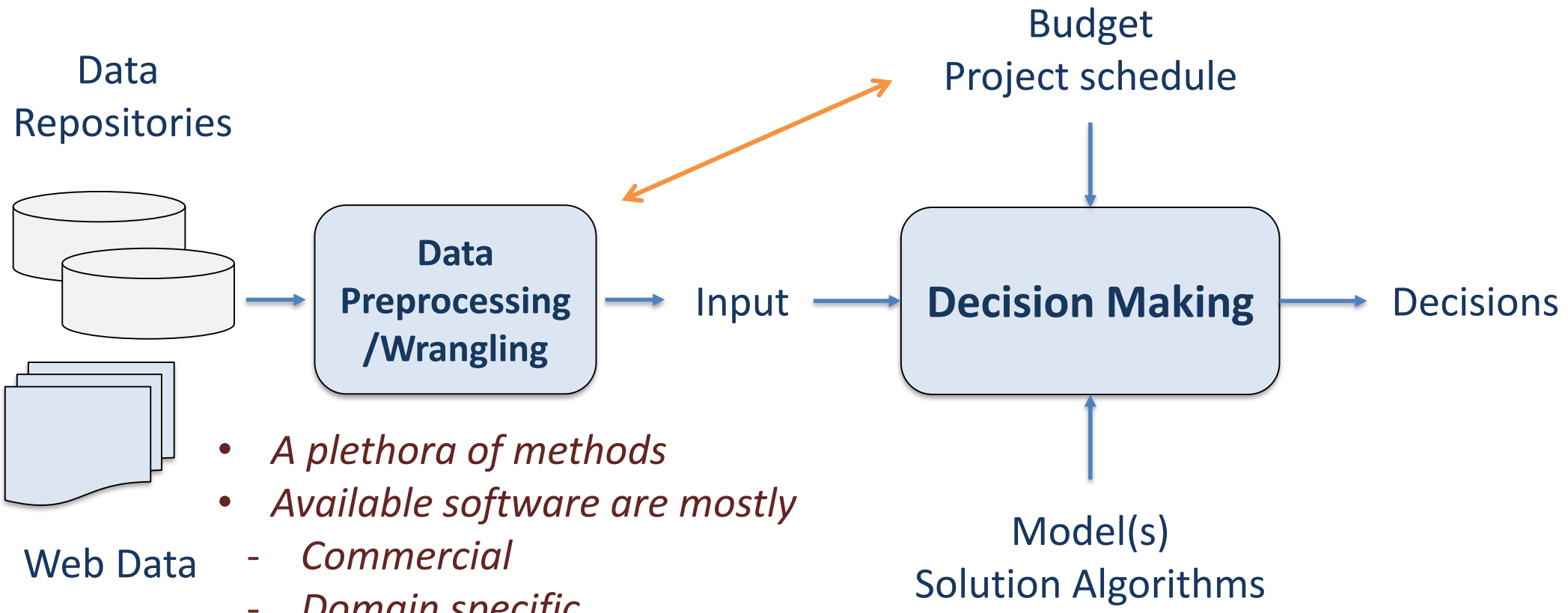
Department:
Applied Algorithmic
Intelligence Methods

ZIB
ZUSE INSTITUTE BERLIN

Data in decision making

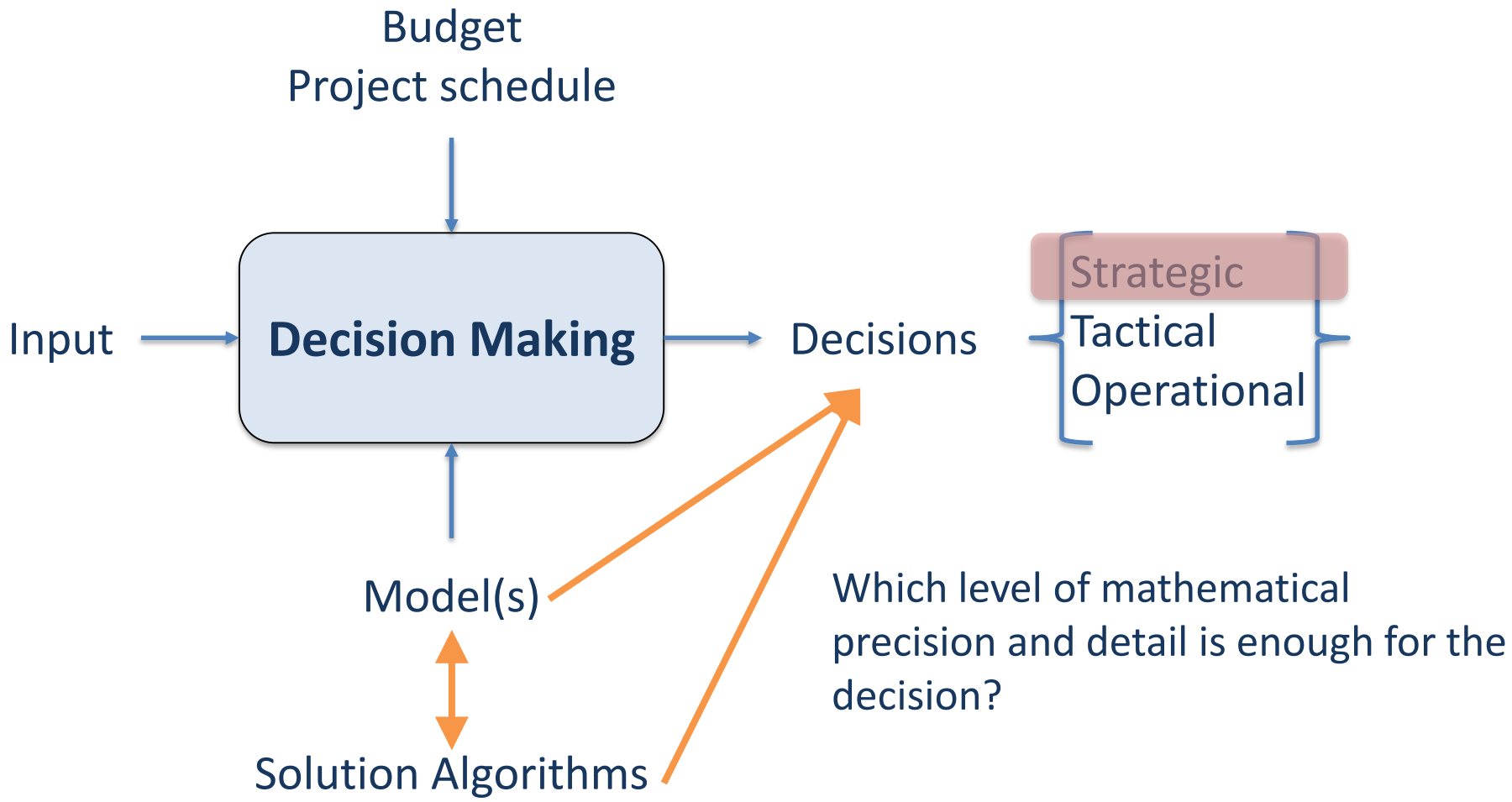


Data Preprocessing



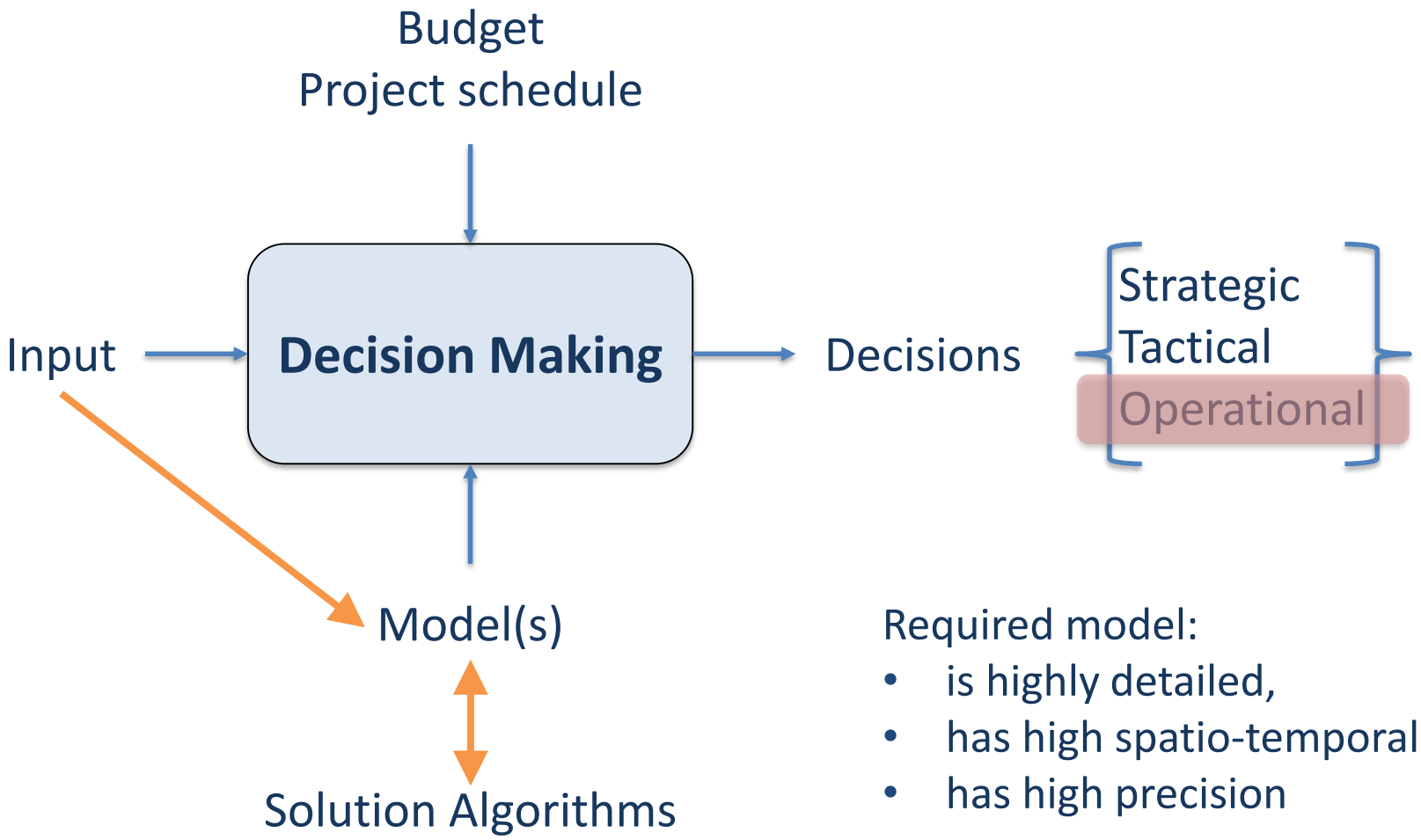
- *A plethora of methods*
- *Available software are mostly*
  - *Commercial*
  - *Domain specific*

Modeling Complex Decisions



Data Quality for Modeling Complex Decisions

Is the input data set of the sufficient quality?



Required model:

- is highly detailed,
- has high spatio-temporal resolution
- has high precision

Dimensions of Data Quality

Accuracy is the ability to reflect reality, i.e., to be able to correctly describe real world objects

Completeness is existence of all required data in the data set, comprehensiveness

Validity is conformance to predetermined type, range and format

Consistency means conflict-free data, i.e., attributes of data entities addressing the same physical attribute should be consistent

Uniqueness means representing each real world object in the data set exactly once

Timeliness shows whether data is up-to-date data depending on the analysis context

[1] Askham et al. The six primary dimensions for data quality assessment. Technical Report, 2013, DAMA UK.

Supply Infrastructures

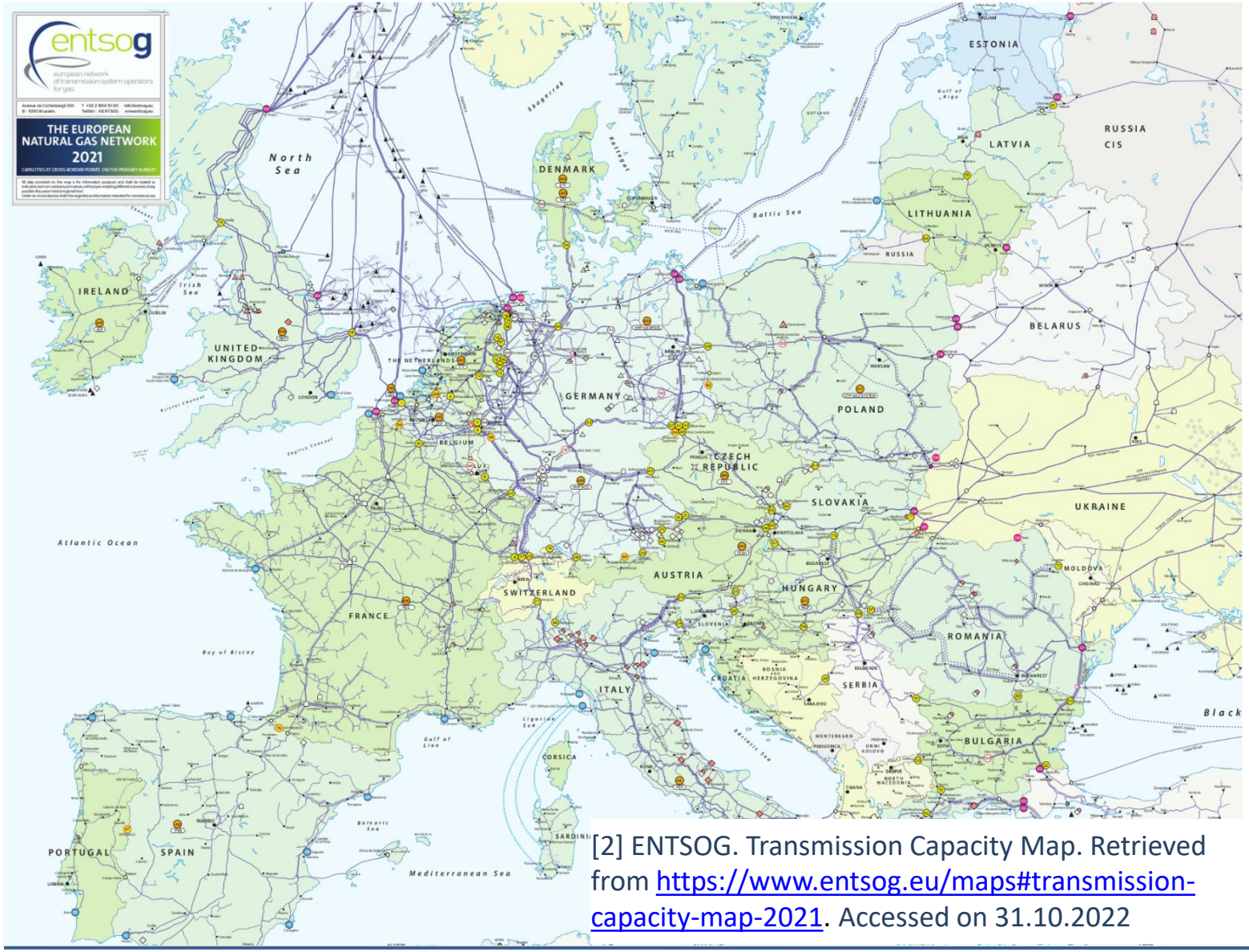
- Transport power, gas, water, data, goods from suppliers to customers
- Consist of complex networks
- Digitization of planning and operation of such networks is essential for vital problems, i.e.,
  - Security of supply of energy, supply chain management, energy transition, decarbonization, etc.

High quality data → Reliable analysis results

High quality data is incredibly costly to obtain both in commercial and public applications, since supply infrastructures

- have complex network structures
- were mostly built before digitization age
- consist of layered and connected structures that may be operated by different parties
- include complex facilities with intricate structures that can be handled by detailed mathematical models

Example: The European Gas Transport Network

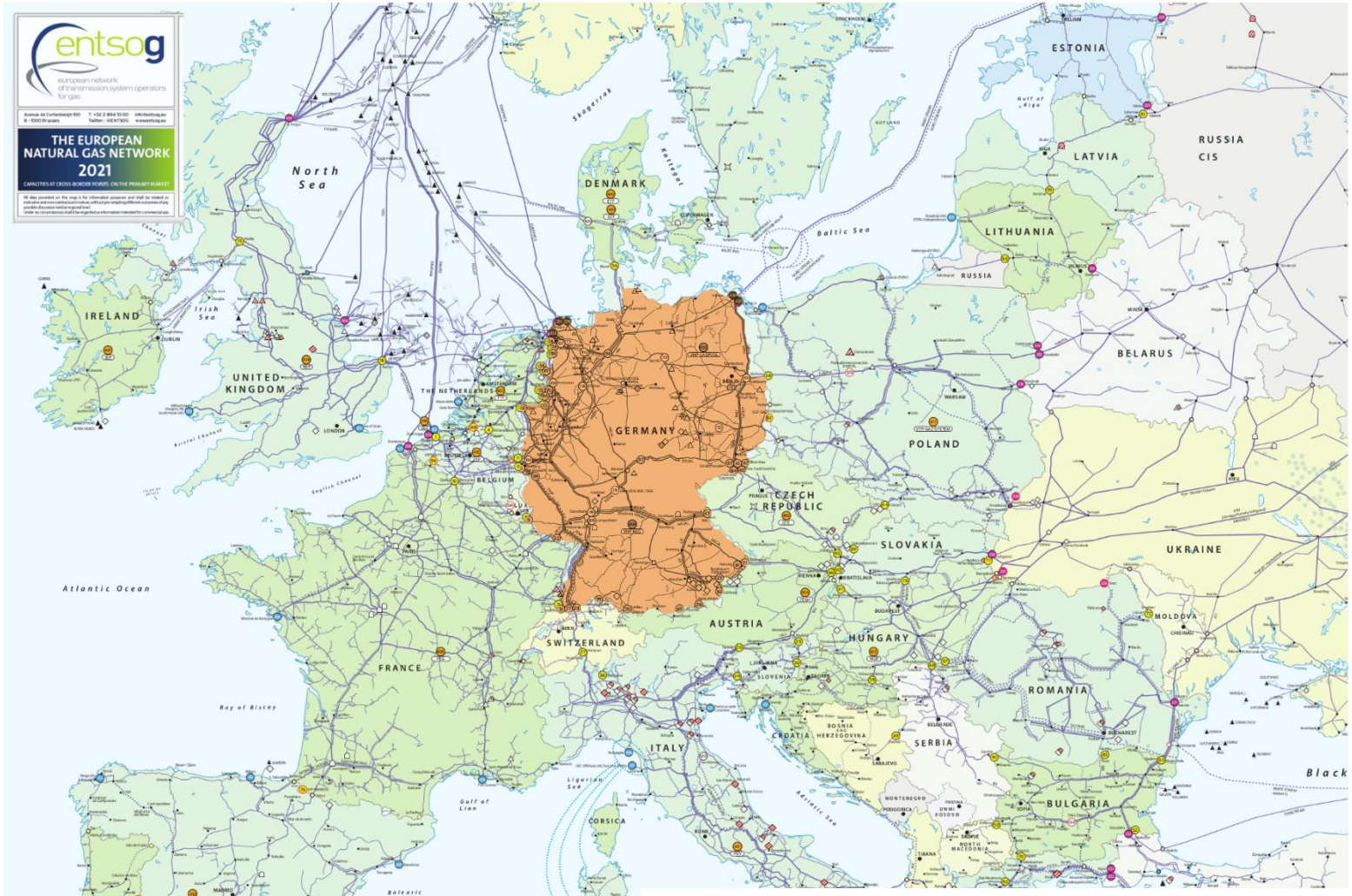


[2] ENTSOG. Transmission Capacity Map. Retrieved from https://www.entsog.eu/maps#transmission-capacity-map-2021. Accessed on 31.10.2022

European gas transport network in numbers:

- 42 member, 10 associated partner, 2 observer TSOs
- > 200 interconnection points, > 170 storages
- ≈ 200,000 km transmission pipelines (EU+UK)

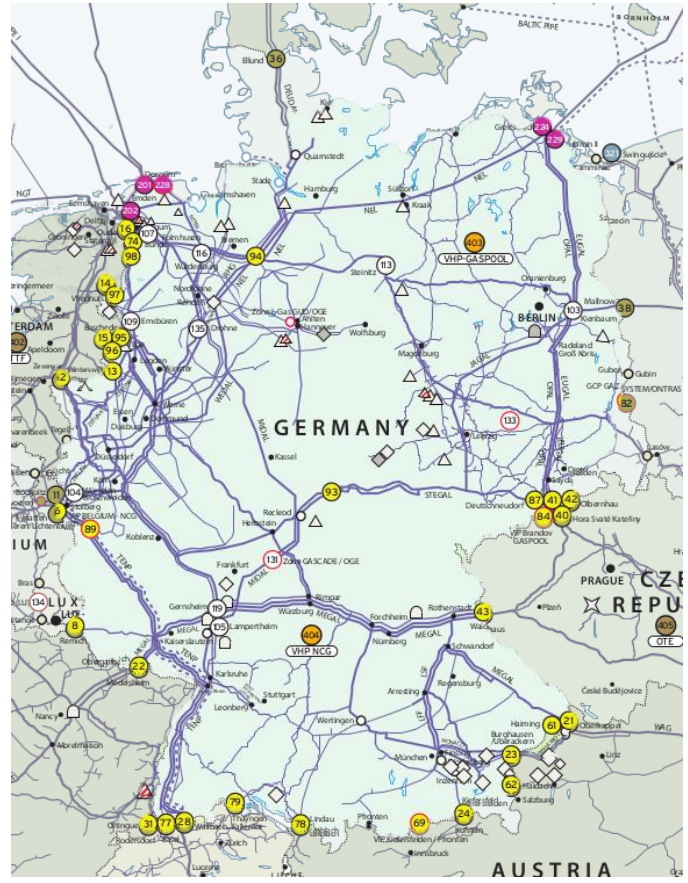
Example: The European and Germany Gas Transport Network



[2] ENTSOG. Transmission Capacity Map. Retrieved from https://www.entsog.eu/maps#transmission-capacity-map-2021. Accessed on 31.10.2022

German gas transport network in numbers:

- 16TSOs
- 44interconnection points, 60storages
- ≈ 31K km transmission pipelines (in total >500K km)



Supply Infrastructure Data - Challenges

Infrastructure network data consists of

- Tabular data: technical bounds, physical properties, etc.
- Spatial data: network topology
- Complex components: often induce non-linear relations of data elements

It is challenging to detect and correct data errors:

- Highly connected
- Requires subject matter expertise to explain errors
- Optimization is actively exploiting errors

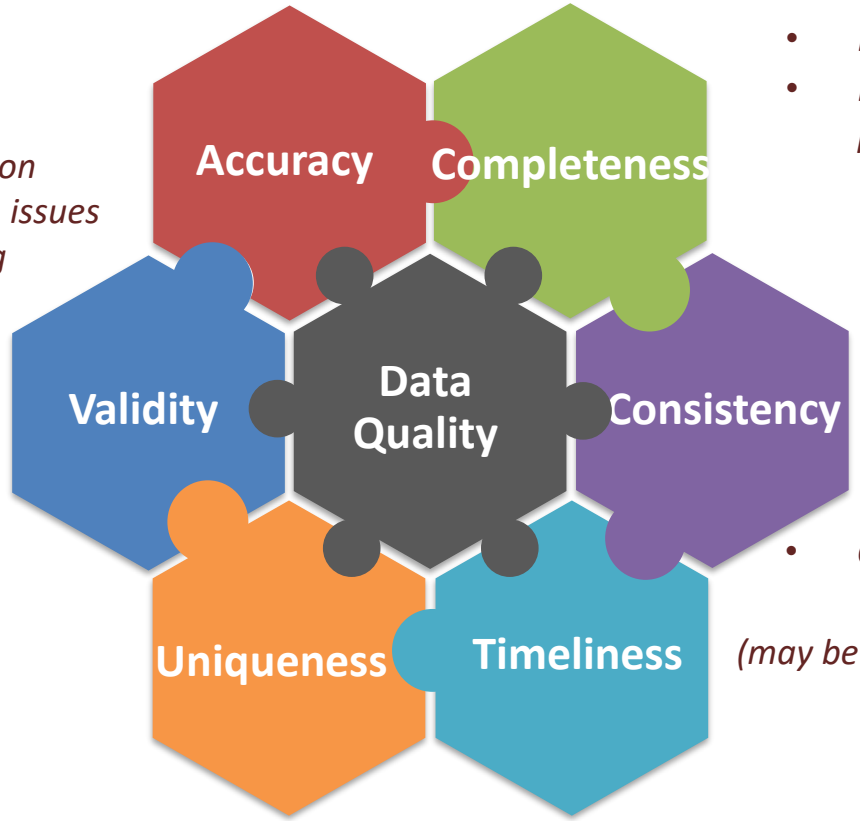
Data Errors: Non-conformances to data quality

Accuracy is the ability to reflect reality, i.e., to be able to correctly describe real world objects

- *Insufficient detail*
  → cumulative capacities
- *Wrong modeling assumptions*
  → flow capacity computation
- *Data preprocessing/integration issues*
  → erroneous node coupling

Completeness is existence of all required data in the data set, comprehensiveness

- *Missing attributes → missing pipes, valves, etc.*
- *Missing entities → missing pipe maximum pressure values, node height values, etc.*



Validity is conformance to predetermined type, range and format

- *Typos in names/non-conformance to bounds (mostly data entry related issues)*

Consistency means conflict-free data, i.e., attributes of data entities addressing the same physical attribute should be consistent

- *Conflicts in the data sets*
  → node pressure vs. pipe pressure bounds

(may be induced by data set integration or data preprocessing assumptions)

Uniqueness means representing each real world object in the data set exactly once

- *Repeated use of object names*

Timeliness shows whether data is up-to-date data depending on the analysis context

- *Untimely data → wrongly forecasted infrastructure related data for future extensions due to limited information*

[1] Askham et al. The six primary dimensions for data quality assessment. Technical Report, 2013, DAMA UK.

Eliminating Data Errors

Data Improvement Method	Related Data Quality Dimension	Error
Schema validation	Uniqueness Validity	Format mismatch Nonconformance to bounds
Data augmentation	Completeness Accuracy	Insufficient detail Missing entries/attributes
Data generation	Completeness Accuracy	Insufficient detail Missing entries/attributes
Consistency check heuristics	Consistency Accuracy	Conflicts in the data set Wrong modeling assumptions Data preprocessing errors

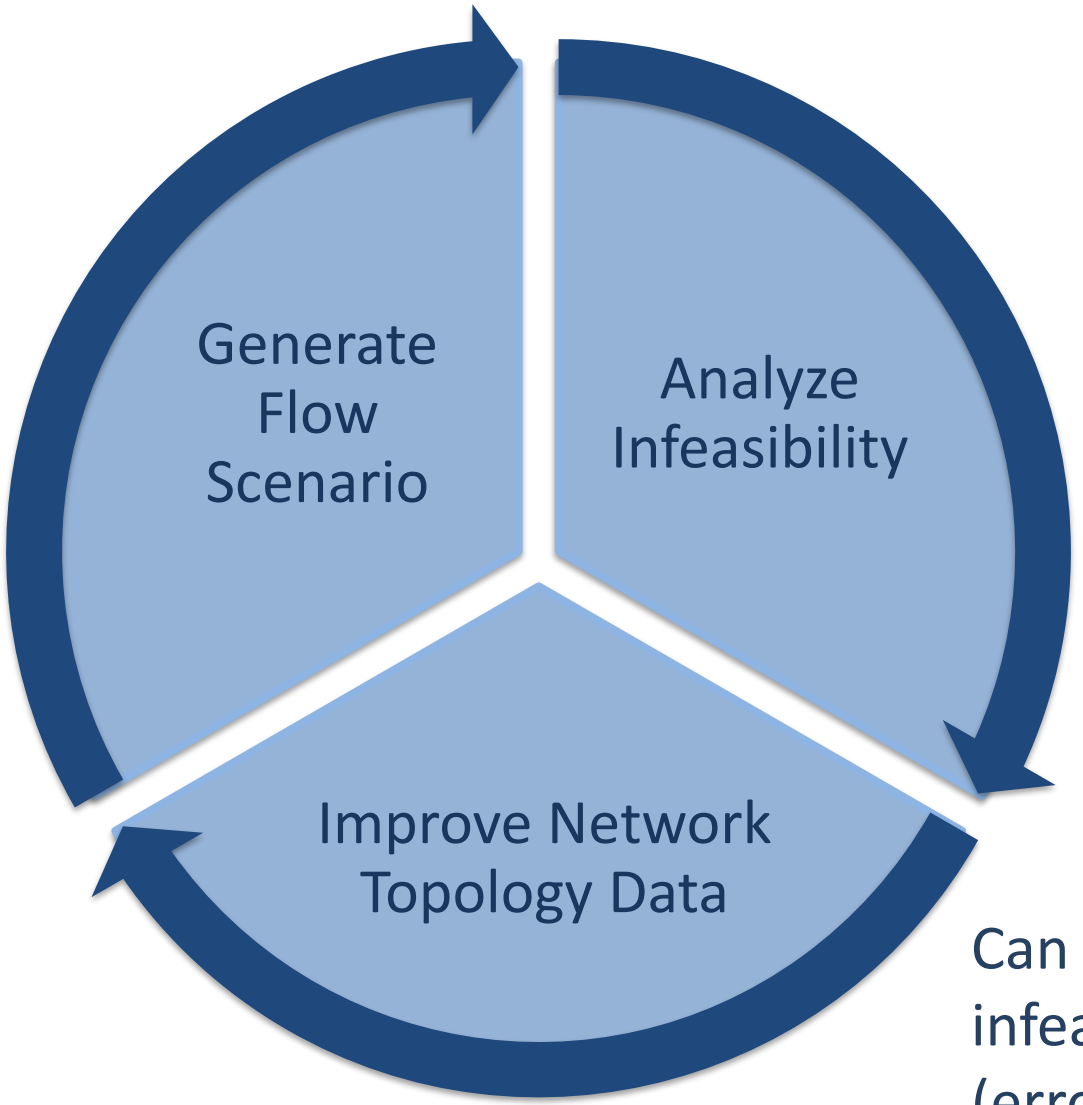
Examples for consistency check heuristics: [3] Inci Yueksel-Erguen, J. Zittel, Y. Wang, F. Hennings, T. Koch. Lessons learned from gas network data preprocessing. Technical Report 20-13. Zuse Institute Berlin, Takustr. 7, 14195 Berlin: ZIB, 2020.

Problem definition

- Highly-connected data in industrial applications
  - provided by industrial partners
  - consolidated from public data sources
  - generated using mathematical models
- Data errors too complex for humans to understand detected by analysis tools, i.e.,
  - irreducible infeasible subsystems (IIS) of large mixed-integer programs
  - bottlenecks in the pressure-coupled pipeline network that is non-linear
- Detection and correction of such errors are extremely difficult

Extensive Scenario Analysis

Generate practically valid scenarios using historical flow data

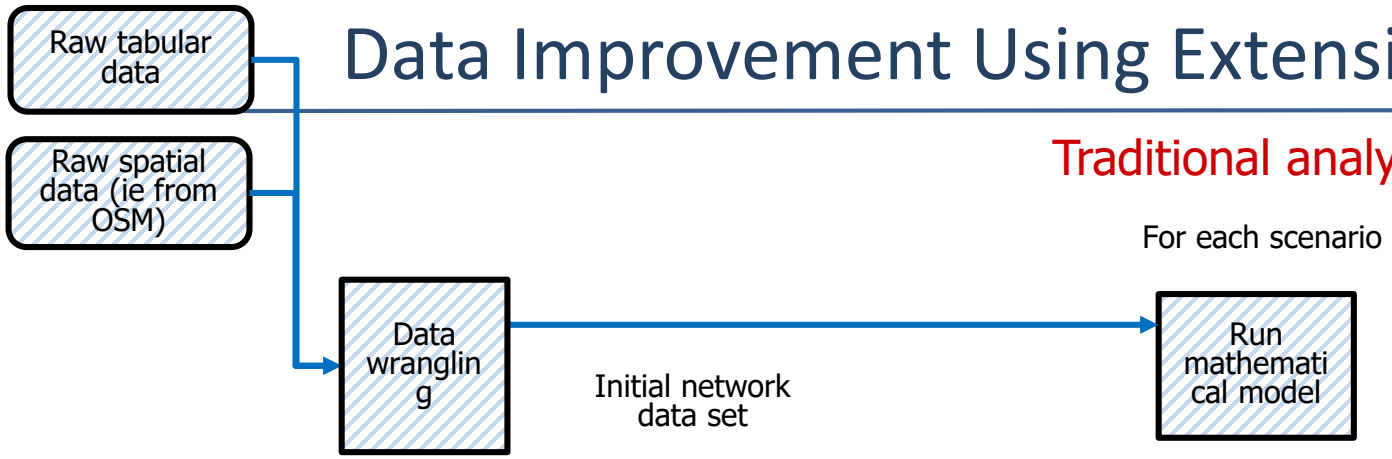


Utilize mathematical modeling methods to detect infeasibility

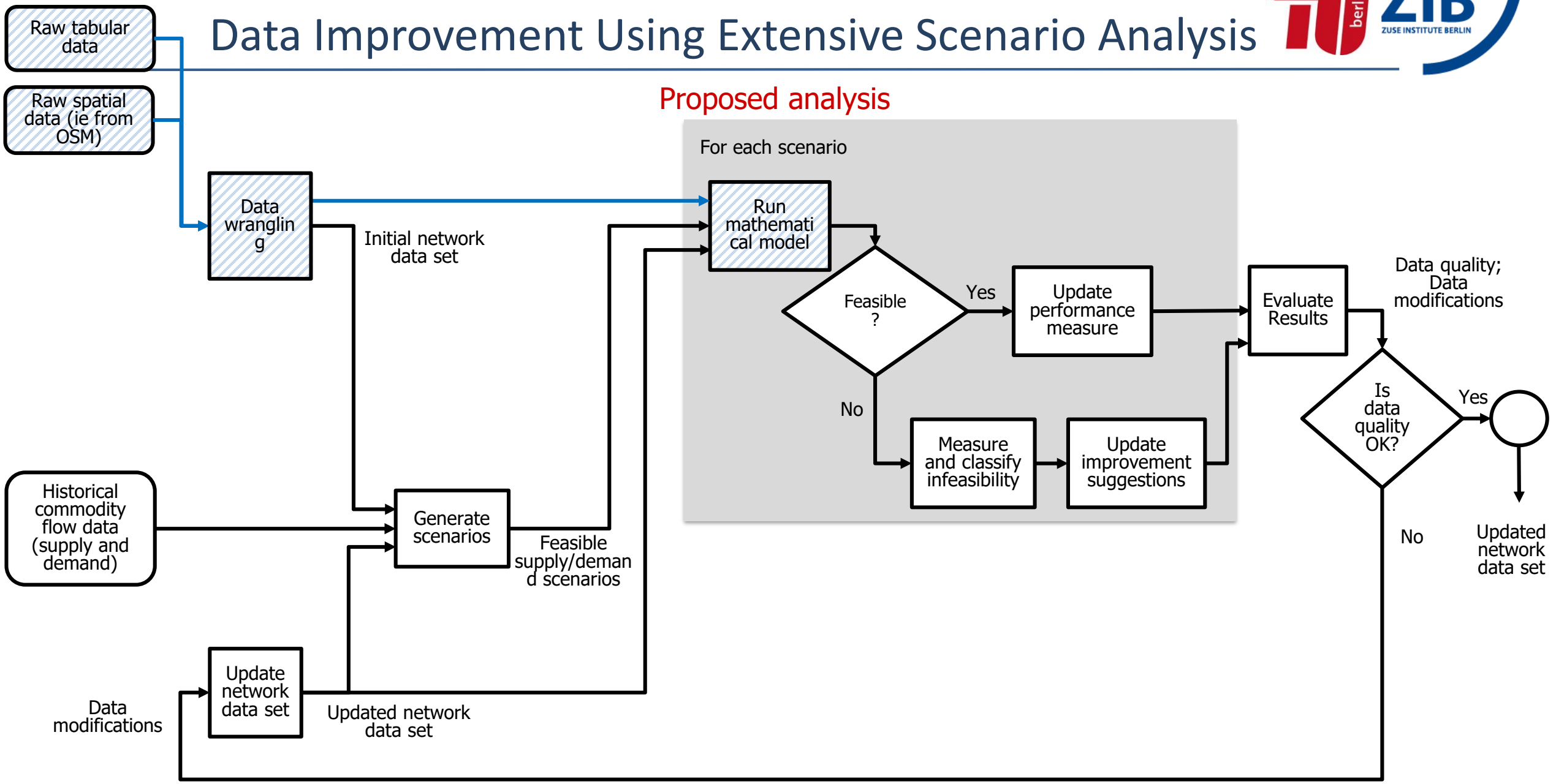
Can we relate infeasibility to data (errors)?

Data Improvement Using Extensive Scenario Analysis

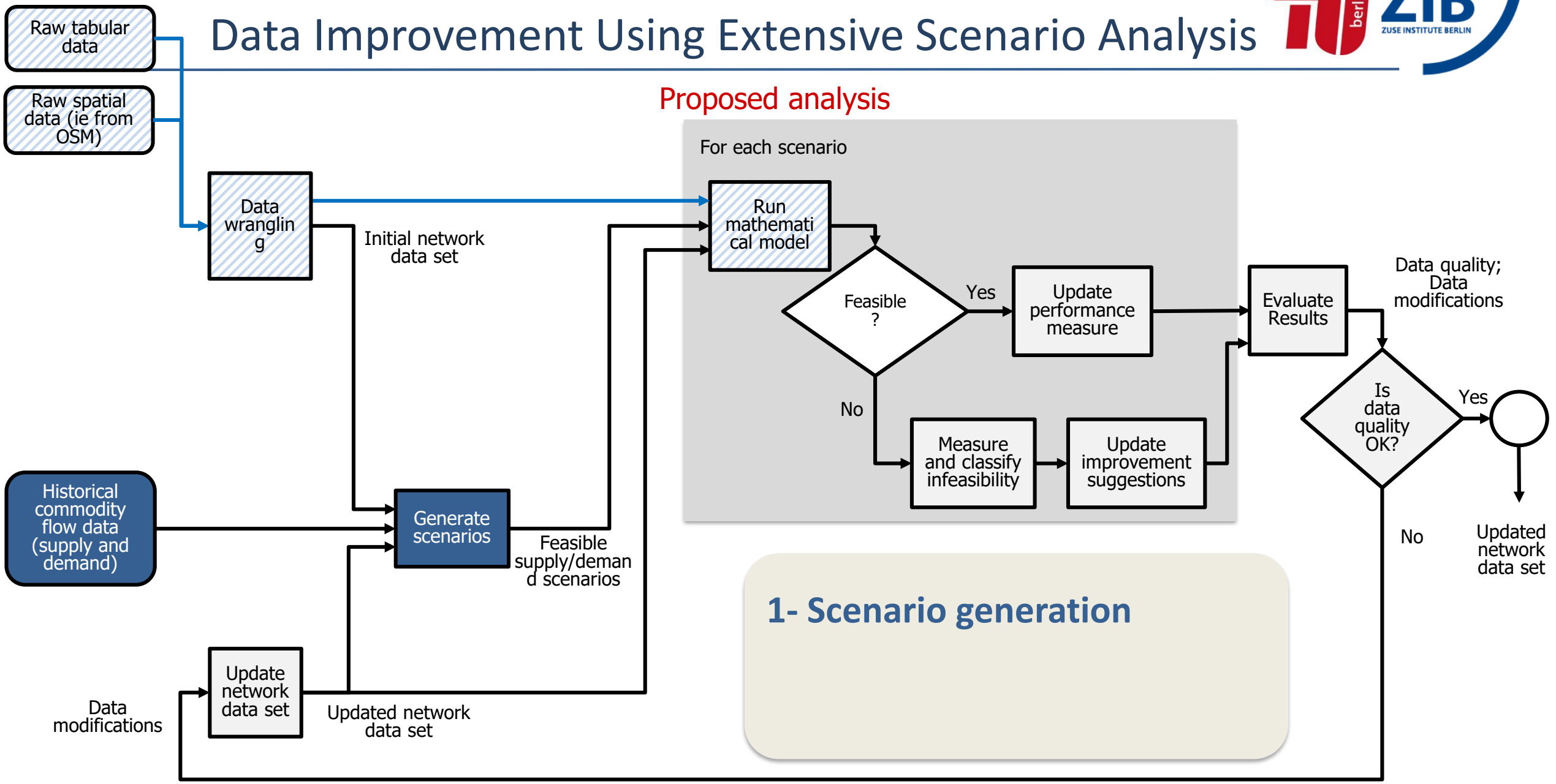
Traditional analysis



Data Improvement Using Extensive Scenario Analysis



Data Improvement Using Extensive Scenario Analysis



An Example Scenario Generator– Gas Networks

Related Data Sets

- ENTSOG IP data set
- GIE Storage data set

M1: High-level Pan-European Gas Network Model

Cumulative gas in-/out- flow of Germany based on IPs

Historical flow data

Source:
ENTSOG Security of Supply Report Scenarios

Related Data Sets

- DE network topology data set
- Node pressure data from TSOs

M2: German Gas Transport Network Model (Linear)

Gas in-/out- flow of Germany based on physical entry/exit nodes

- Estimated pipeline capacities
- ENTSOG IP- DE network associations

Resulting data

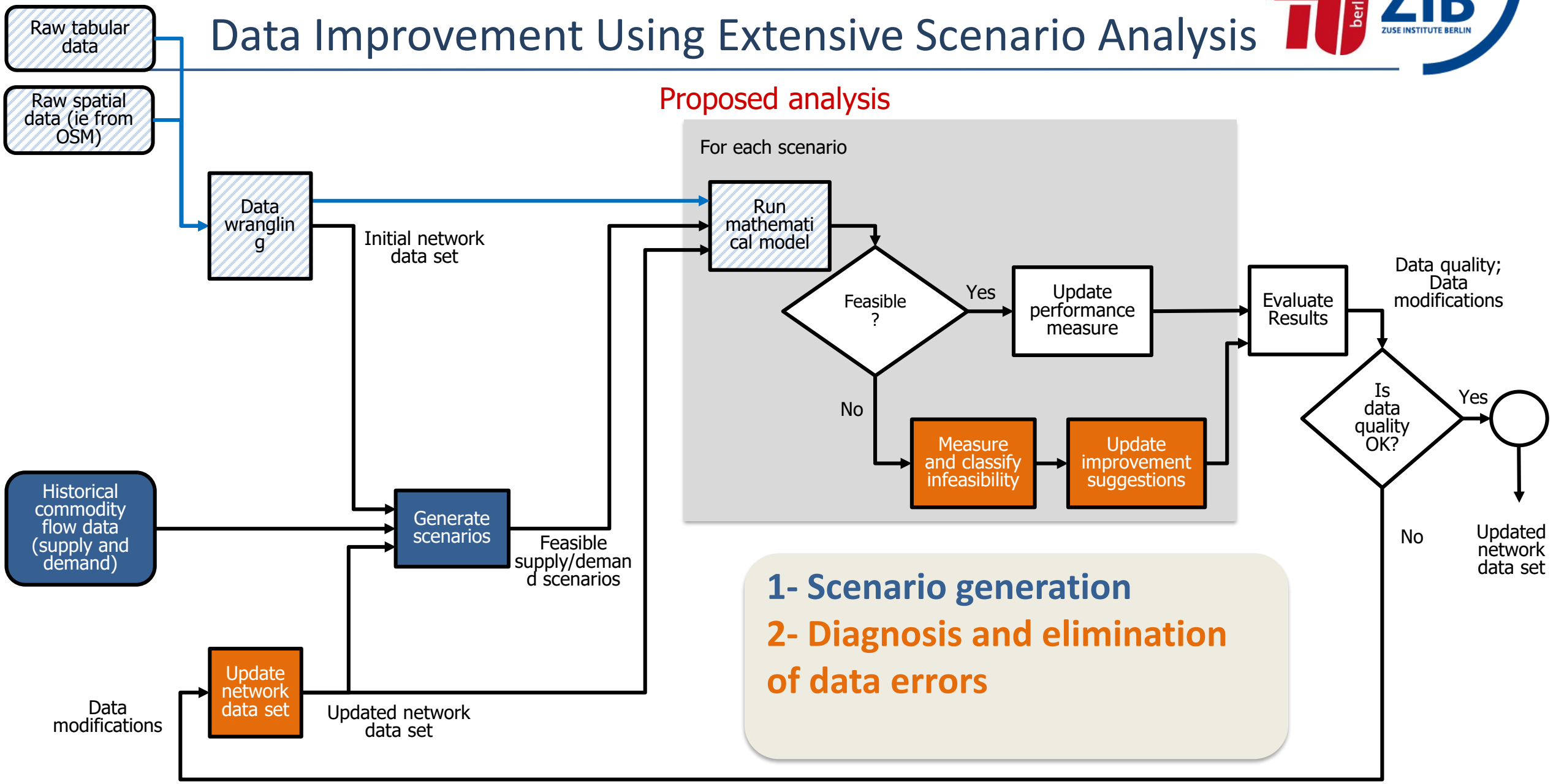
- DE gas network
- CS data

M3: German Gas Transport Network Model (Nonlinear)

- Parameters related to gas properties

Details for the models: [4] I. Yueksel-Erguen, D. Most, L. Wyrwoll, C. Schmitt, J. Zittel. Modeling the transition of the multimodal pan-European energy system including an integrated analysis of electricity and gas transport. Technical Report 22-17. Zuse Institute Berlin, Takustr. 7, 14195 Berlin, 2022.

Data Improvement Using Extensive Scenario Analysis



Mathematical Modeling Methods for Infeasibility Analysis

Slack Formulations

- Different aspects of the formulations can be relaxed with slack variables
- The objective is to minimize the deviation from the original model → zero objective function
- The smallest distance from the feasibility

(Minimum) Irreducible Infeasible Subsystems (IIS)

- Isolates the infeasibility by variables and constraints
- Not an explicit reason why a IIS is infeasible
- Long computation times for large-scale MIPs
- A trivial IIS is not informative

ns1158817 from MIPLIB 2010*	Constraints	Variables	NZ
Problem	68,455	1,804,022	2,842,044
IIS	2,003	6,002	12,002

**[5] Koch et al. MIPLIB 2010. Math. Prog. Comp. 3, 103 (2011). https://doi.org/10.1007/s12532-011-0025-9*

Semi-Automated Infeasibility Analysis Method

S1. Initiate scenario analysis:

relax all non-linear constraints in the slack formulation

if there is at least one scenario in the scenario set S, select a scenario s from the scenario set and go to S2, else go to S5

S2. Solve the mathematical model

If feasible save the solution, delete the scenario s from the scenario set S and go to S1; else go to S3

S3. Solve the slack formulation

if infeasible go to S4

if feasible with non-zero slack: correct the scenario and turn to S3

if feasible with a zero objective function value, tighten one set of the nonlinear constraints in the slack formulation if available and turn to S3, else save the solution and scenario correction, delete the scenario s from the scenario set S and go to S1

S4. Find the minIIS using the LP minIIS model

if feasible, save the minIIS solution for the scenario, rescale the scenario by 0.95, and go to S4

if infeasible go to S1

S5. Analyze the saved minIISs

to detect the frequency of existence of network components and constraint types in the minIIS and correct data set, and go to S1

Automated Infeasibility Analysis Method

S0. Initiate data set:

restore the scenario set: S:=S0; reset the data rating DR:=0

S1. Initiate scenario analysis:

relax all non-linear constraints in the slack formulation

if there is at least one scenario in the scenario set S, select a scenario s from the scenario set and go to S2; else go to S5

S2. Solve the mathematical model

If feasible save the solution and the scenario neighborhood scale, delete the scenario s from the scenario set S and go to S1; else go to S3

S3. Solve the slack formulation

if infeasible go to S4

if feasible with non-zero slack: correct the scenario and turn to S3

if feasible with a zero objective function value, tighten one set of the nonlinear constraints in the slack formulation if available and turn to S3, else update the scenario neighborhood scale and go to S2

S4. Find the minIIS using the LP minIIS model

if feasible, save the minIIS solution for the scenario, rescale the scenario by 0.95, and go to S4

if infeasible go to S1.

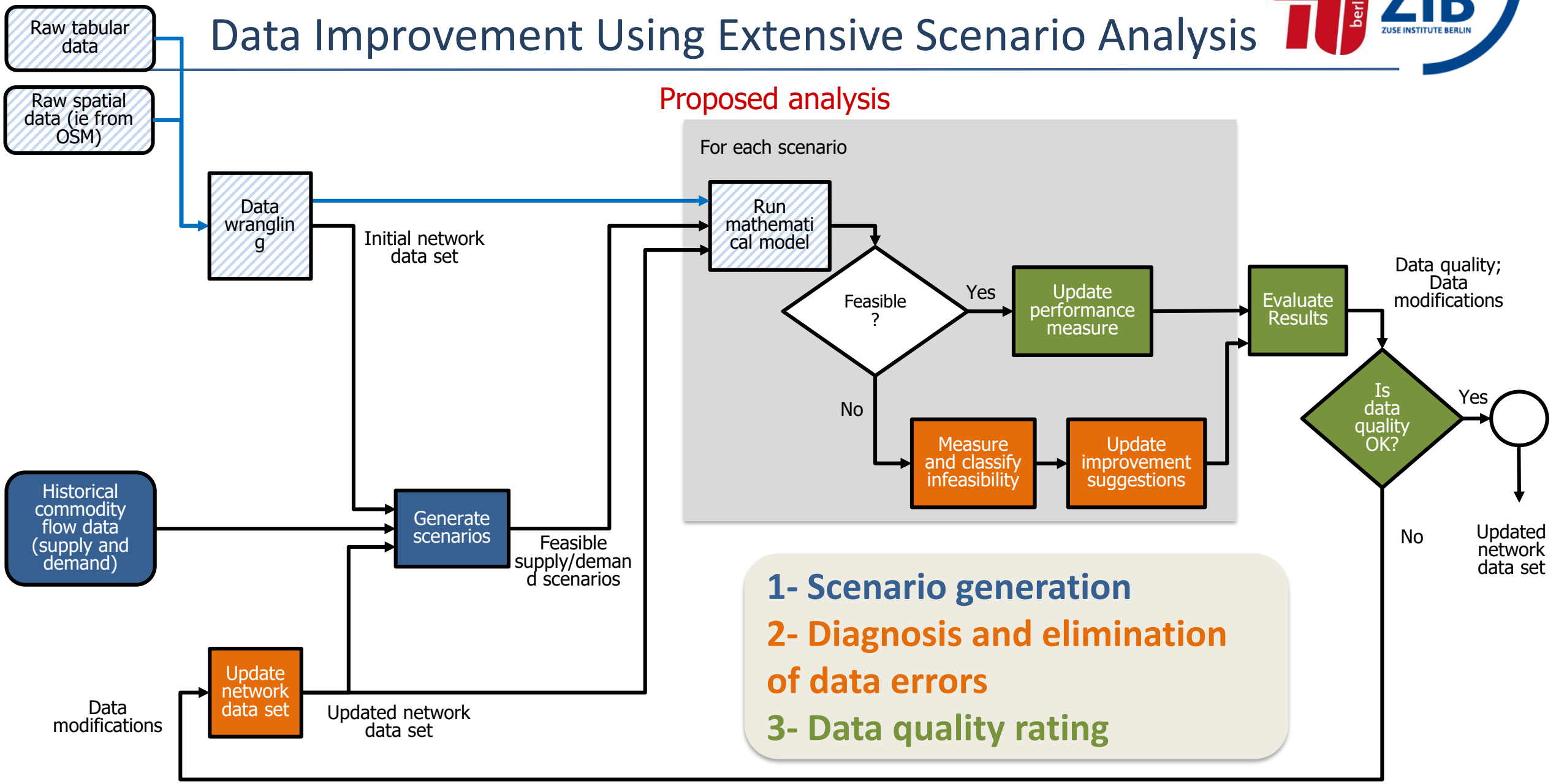
S5. Analyze the saved minIISs

to detect the frequency of existence of network components and constraint types in the minIIS, correct data set,

S6. Measure the data quality

If data quality is not sufficient, got to S1, else terminate.

Data Improvement Using Extensive Scenario Analysis



Data Quality Rating

A data rating measure to **facilitate the automated improvement** by enabling us to

- understand whether the improved data set is of **sufficient quality level for the aimed analysis**
- **compare** the performance of alternative improvements

.. and also lead us/the search to **the potential errors and error sources…**

Example: Number of scenarios generated from the historical data that the data set finds a feasible routing

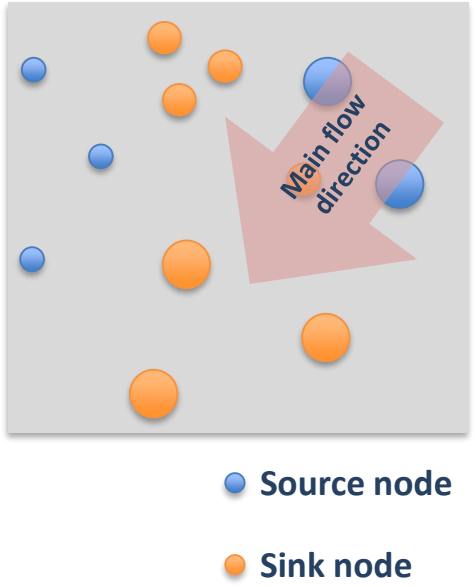
- Not informative enough
- Especially in the very beginning of the improvement process – we may get 0 for all scenario sets

Data Quality Rating of Infrastructure Network Data

An infrastructure network is an engineered system that is operational:

- Designed to meet **certain quality of service requirements given operational requirements**
- **Robust** against **uncontrollable** factors in the working environment

If a scenario gives a feasible result with the data set,
then a set of scenarios in its **neighborhood** is expected to be feasible.

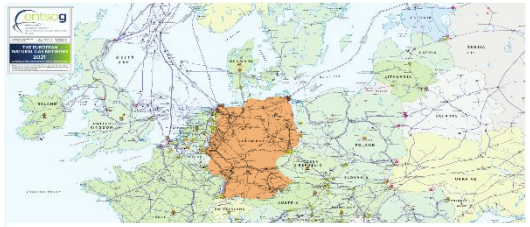


Historical flow scenarios of an infrastructure network have two main measurable characteristics:

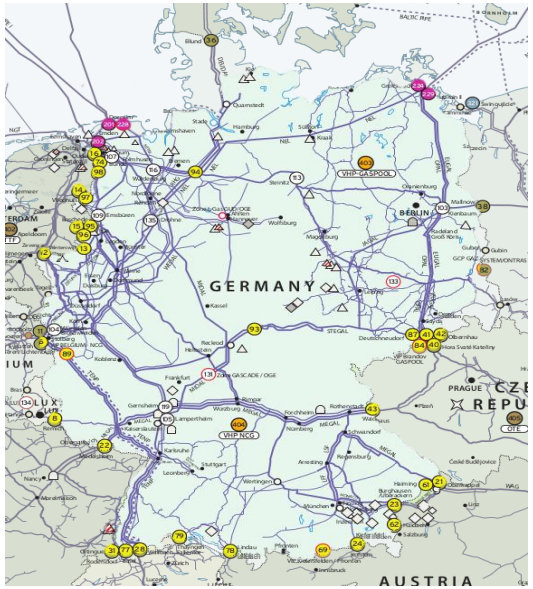
- The **amount** of commodity that can be routed
- The **distribution** of the commodity in the network

Example: Gas Transport Network Data

European Gas Transport Network: Interconnections



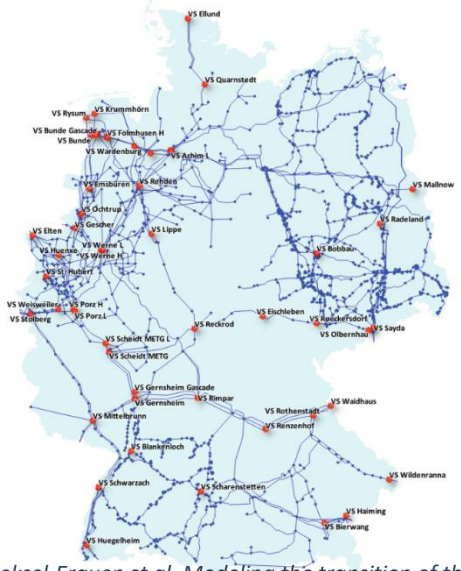
[2] ENTSOG. Transmission Capacity Map. Retrieved from https://www.entsog.eu/maps#transmission-capacity-map-2021. Accessed on 31.10.2022



[2] ENTSOG. Transmission Capacity Map. Retrieved from https://www.entsog.eu/maps#transmission-capacity-map-2021. Accessed on 31.10.2022

Germany Gas Transport Network: Topology & Complex facilities

- ≈70 entry & ≈900 exit points
- ≈1650 inner nodes
- ≈1770 pipes
- ≈95 control valves
- 58 compressor stations
  - 129 compressors & drivers
  - 200 valves



[4] Yueksel-Erguen et al. Modeling the transition of the multimodal pan-European energy system including an integrated analysis of electricity and gas transport. Technical Report 22-17. Zuse Institute Berlin, Takustr. 7, 14195 Berlin, 2022.
[6] Kunz et al. Reference Data Set: Electricity, Heat, and Gas Sector Data for Modeling the German System (Version 1.0.0), 2017. https://doi.org/10.5281/zenodo.1044463.

Gas Data Files: Technical properties

Network topology data: .net file (>34K lines; >28K data attributes)

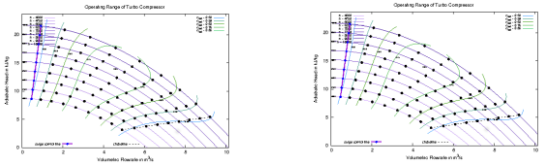
```
<source id="N1" x="-100" y="500">
  <height unit="m" value="200.0"/>
  <pressureMin unit="bar" value="30.0"/>
  <pressureMax unit="bar" value="70.0"/>
  <flowMin unit="1000m_cube_per_hour" value="50.0"/>
  <flowMax unit="1000m_cube_per_hour" value="750.0"/
  <gasTemperature unit="Celsius" value="10.0"/>
  <calorificValue value="37" unit="MJ_per_m_cube"/>
  <normDensity value="0.785" unit="kg_per_m_cube"/>
  <coefficient-A-heatCapacity value="32.23"/>
  <coefficient-B-heatCapacity value="-0.01"/>
  <coefficient-C-heatCapacity value="0"/>
  <molarMass value="19.5" unit="kg_per_kmol"/>
  <pseudocriticalPressure value="44.5" unit="bar"/>
  <pseudocriticalTemperature value="190" unit="K"/>
</source>
```

```
<pipe from="N1" id="P12" to="N2">
  <flowMin unit="1000m_cube_per_hour" value="0.0"/>
  <flowMax unit="1000m_cube_per_hour" value="73500"/>
  <length unit="km" value="50"/>
  <diameter unit="m" value="1.1"/>
  <roughness unit="mm" value="0.01"/>
  <heatTransferCoefficient unit="W_per_m_square_per_K" value="2.0"/>
</pipe>
```

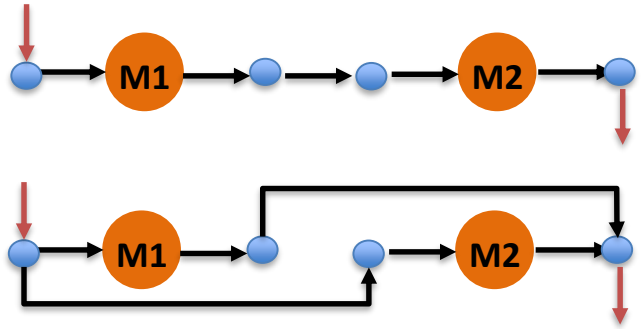
Compressor station data: .cs file (>54K lines)

```
<compressorStation id="CS3">
  <compressors>
    <turboCompressor drive="P_CS3_M1" id="T_CS3_M1">
    <turboCompressor drive="P_CS3_M3" id="T_CS3_M3">
  </compressors>
  <drives>
    <gasTurbine id="P_CS3_M1">
    <gasTurbine id="P_CS3_M3">
  </drives>
  <configurations>

    <configuration confId="1" nrOfSerialStages="2">
      <stage stageNr="1" nrOfParallelUnits="1">
      <stage stageNr="2" nrOfParallelUnits="1">
    </configuration>

    <configuration confId="3" nrOfSerialStages="1">
      <stage stageNr="1" nrOfParallelUnits="1">
    </configuration>

    <configuration confId="2" nrOfSerialStages="1">
      <stage stageNr="1" nrOfParallelUnits="2">
    </configuration>

    <configuration confId="4" nrOfSerialStages="1">
      <stage stageNr="1" nrOfParallelUnits="1">
    </configuration>

  </configurations>
</compressorStation>
```

Compressor station characteristic diagrams



Alternative compressor machine configurations



Example: Gas Transport Network Quality Rating

Representative flow scenarios generated using open data published by ENTSOG

[7] ENTSOG. Union-wide simulation of gas supply and infrastructure disruption scenarios (SoS simulation), 2017. Retrieved from https://www.entsog.eu/sites/default/files/entsog-migration/publications/sos/ENTSOG%20Union%20wide%20SoS%20simulation%20report_INV0262-171121.pdf. Accessed on 03.05.2023

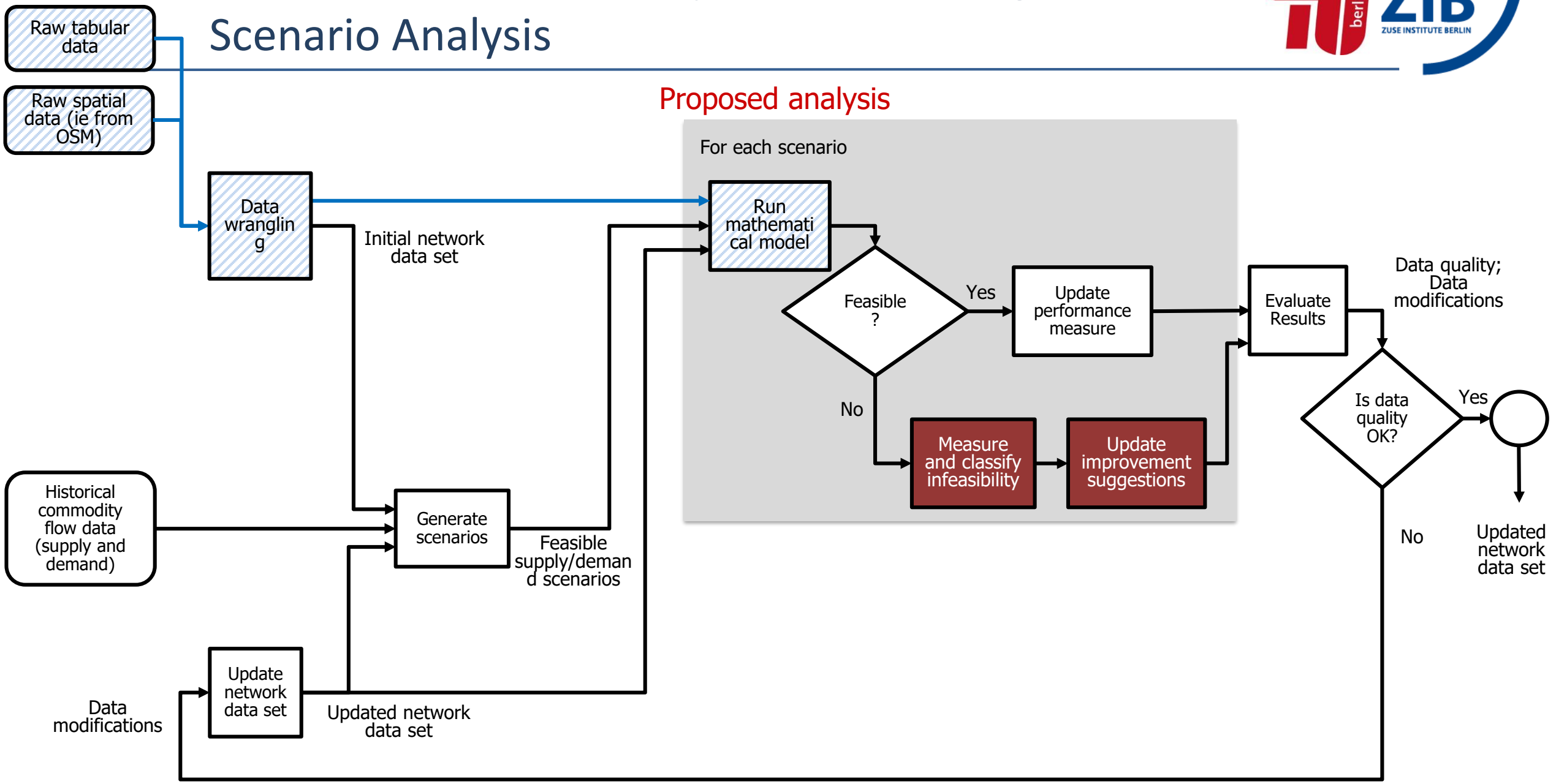
- Schema validation
- Data augmentation
- Data generation
- Consistency check heuristics

Less than %40 of representative historical flow scenarios at 90% commodity level

- Extensive scenario analysis

All representative historical flow scenarios at 95% commodity level

Future Outlook: Data Improvement Using Extensive Scenario Analysis



Thank you for listening!

yueksel-erguen@zib.de

References

[1] Nicola Askham, Denise Cook, Martin Doyle, Helen Fereday, Mike Gibson, Ulrich Landbeck, Rob Lee, Chris Maynard, Gary Palmer, Julian Schwarzenbach. The six primary dimensions for data quality assessment. Technical Report, 2013, DAMA United Kingdom.

[2] ENTSOG. Transmission Capacity Map. Retrieved from https://www.entsog.eu/maps#transmission-capacity-map-2021. Accessed on 31.10.2022

[3] Inci Yueksel-Erguen, Janina Zittel, Ying Wang, Felix Hennings, Thorsten Koch. Lessons learned from gas network data preprocessing. Technical Report 20-13. Zuse Institute Berlin, Takustr. 7, 14195 Berlin, 2020. Available online https://opus4.kobv.de/opus4-zib/frontdoor/index/index/docId/7826

[4] Inci Yueksel-Erguen, Dieter Most, Lothar Wyrwoll, Carlo Schmitt, Janina Zittel. Modeling the transition of the multimodal pan-European energy system including an integrated analysis of electricity and gas transport. Technical Report 22-17. Zuse Institute Berlin, Takustr. 7, 14195 Berlin, 2022. Available online < https://opus4.kobv.de/opus4-zib/frontdoor/index/index/docId/8777>

[5] Thorsten Koch, Tobias Achterberg, Erling Andersen, Oliver Bastert, Timo Berthold, Robert E. Bixby, Emilie Danna, Gerald Gamrath, Ambros M. Gleixner, Stefan Heinz, Andrea Lodi, Hans Mittelmann, Ted Ralphs, Domenico Salvagnin, Daniel E. Steffy, Kati Wolter MIPLIB 2010. Math. Prog. Comp. 3, 103 (2011). https://doi.org/10.1007/s12532-011-0025-9

[6] Friedrich Kunz, Jens Weibezahn, Philip Hauser, Sina Heidari, Wolf-Peter Schill, Björn Felten, Mario Kendziorski, Matthias Zech, Jan Zepter, Christian von Hirschhausen, Dominik Möst, Christoph Weber. Reference Data Set: Electricity, Heat, and Gas Sector Data for Modeling the German System (Version 1.0.0), 2017. https://doi.org/10.5281/zenodo.1044463.

[7] ENTSOG. Union-wide simulation of gas supply and infrastructure disruption scenarios (SoS simulation), 2017. Retrieved from https://www.entsog.eu/sites/default/files/entsog-migration/publications/sos/ENTSOG%20Union%20wide%20SoS%20simulation%20report_INV0262-171121.pdf. Accessed on 03.05.2023