David Laner
david.laner@tuwien.ac.at

Data quality assessment and uncertainty characterisation

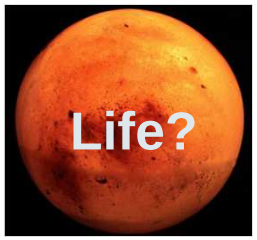
TECHNISCHE UNIVERSITÄT WIEN

Special Session
2017 Joint Conference, ISIE and ISSST
25–29 June 2017, Chicago, USA

Nature of uncertainty[1]

- **Aleatory variability**
  - Due to the randomness of processes
  - “irreducible” uncertainty

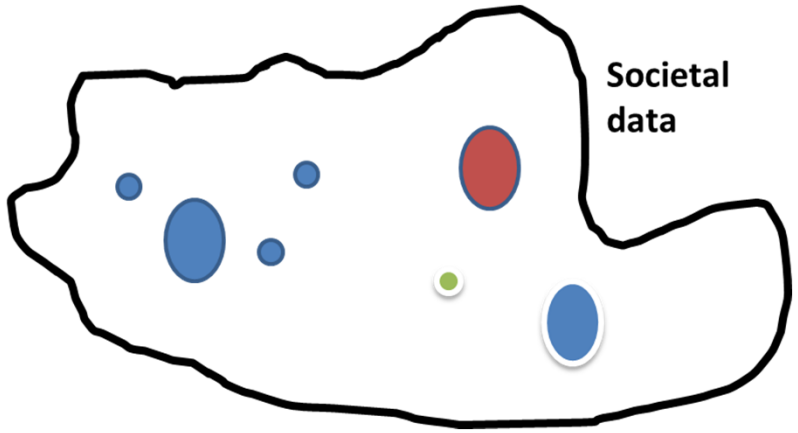
- **Epistemic uncertainty**
  - Due to a lack of knowledge
  - “Reducible” uncertainty



[1]Hoffman, Hammonds (1994): Propagation of Uncertainty in Risk Assessments: The Need to Distinguish Between Uncertainty Due to Lack of Knowledge and Uncertainty Due to Variability. Risk Analysis, 14: 707–712.

Data situation in Industrial Ecology

- Data from different sources with different quality generated in different ways
  - Characteristics of input data should reflect these differences

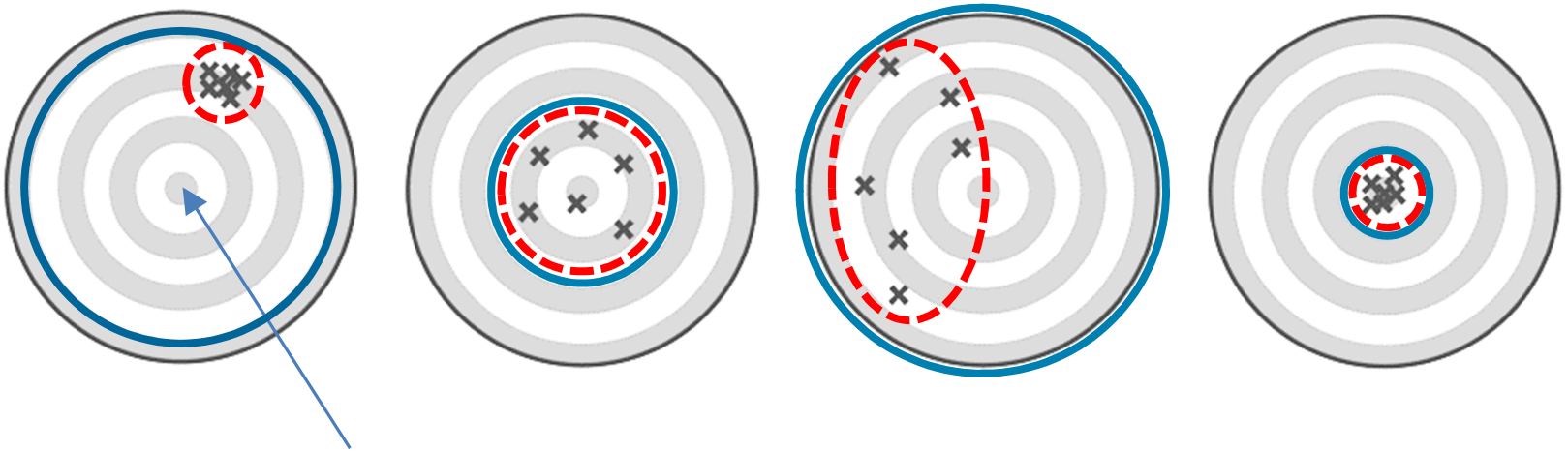


- Kinds of data
  - Sparse measurements
  - Official statistical data
    - Regional or company level
    - National level
  - Regionalized data (top down estimates)
  - Literature data
  - Expert estimates
  - …

Data Quality (1)

Measurement precision

Data quality



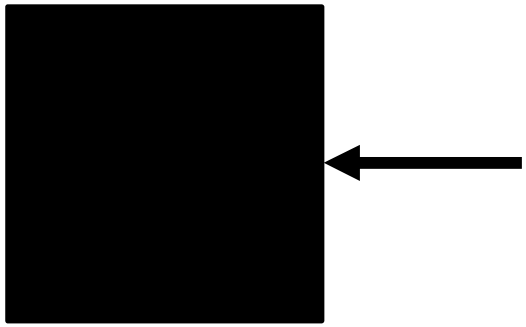
That´s what we want to know

Data quality ≠ Measurement uncertainty

Data Quality (2)

Data quality
is a measure of the reliability of data in the context of the application purpose

- „good“ vs. „bad“ data
- only evaluable on a given context
- subjective (to a certain degree)



Status Quo:
Data quality is often a „Black box“.

Data classification

- Categorisation approach

Level	Source of information	Example
1 (interval */1.1)	Official statistics on local level.	Number of households, cars, apartments, small houses.
	Information from authorities/ construction/production.	Cr content in steel for a specific application.
2 (interval */1.33)	Official statistics on (local), regional and national levels.	Percentage of leather shoes among shoes.
	Information from authorities/ construction/production.	Amount of Pb and Cu in power cables. Cr content in leather. Thickness of Ni and Cr layer on plating. Paint per area.
3 (interval */2)	Official statistics on national level downscaled to local level.	Share of Volvo cars among all cars.
	Information on request from authorities/construction/production.	Annual use of stainless steel on roofs and fronts.
4 (interval */4)	Information on request from authorities/construction/production.	Weight of catalytic converters.
5 (interval */10)		Cd content in Zn in a type of good, eg galvanised goods.

Hedbrant, Sörme (2001): Data Vagueness and Uncertainties in Urban Heavy-Metal Data Collection. Water, Air, & Soil Pollution: Focus 1(3): 43-53.

Data quality indicators

- PEDIGREE matrix

Table 1 Pedigree matrix with 5 data quality indicators

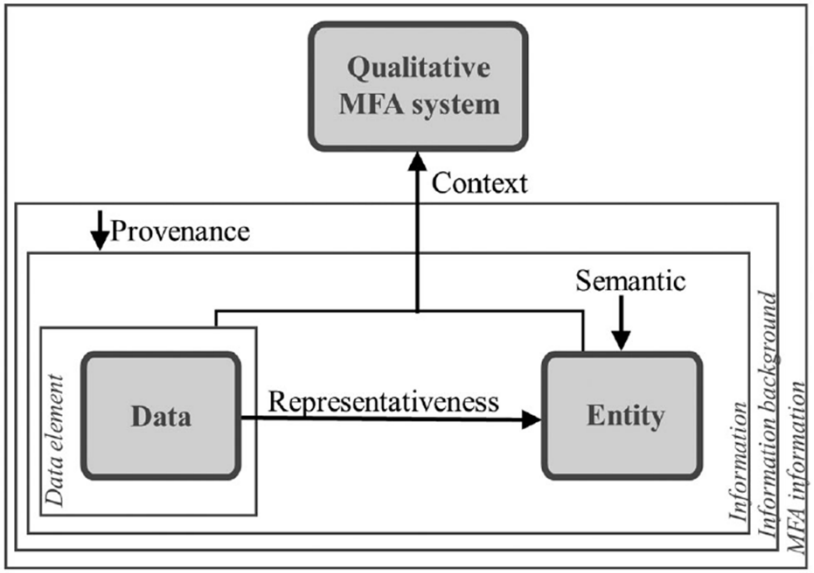
Indicator score	1	2	3	4	5
Reliability	Verified[a] data based on measurements[b]	Verified data partly based on assumptions or non-verified data based on measurements	Non-verified data partly based on assumptions	Qualified estimate (e.g. by industrial expert)	Non-qualified estimate
Completeness	Representative data from a sufficient sample of sites over an adequate period to even out normal fluctuations	Representative data from a smaller number of sites but for adequate periods	Representative data from an adequate number of sites but from shorter periods	Representative data but from a smaller number of sites and shorter periods or incomplete data from an adequate number of sites and periods	Representativeness unknown or incomplete data from a smaller number of sites and/or from shorter periods
Temporal correlation	Less than three years of difference to year of study	Less than six years difference	Less than 10 years difference	Less than 15 years difference	Age of data unknown or more than 15 years of difference
Geographical correlation	Data from area under study	Average data from larger area in which the area under study is included	Data from area with similar production conditions	Data from area with slightly similar production conditions	Data from unknown area or area with very different production conditions
Further technological correlation	Data from enterprises, processes and materials under study	Data from processes and materials under study but from different enterprises	Data from processes and materials under study but from different technology	Data on related processes or materials but same technology	Data on related processes or materials but different technology

[a]Verification may take place in several ways, e.g. by on-site checking, by recalculation, through mass balances or cross-checks with other sources.
[b]Includes calculated data (e.g. emissions calculated from inputs to a process), when the basis for calculation is measurements (e.g. measured inputs). If the calculation is based partly on assumptions, the score should be two or three.

Weidema, Wesnæs (1996): Data quality management for life cycle inventories—an example of using data quality indicators. Journal of Cleaner Production 4(3–4): 167-174.

Data quality assessment framework

MFA data quality evaluation (often isolated values instead of datasets) needs to be systematic and transparent.



Result:
Information defect for each flow of the MFA:
ID → 0 good quality (zero defect)
ID → 1 poor quality (max defect)

Representativeness	The sample size is not large enough to represent a complex population.
Semantic	The meaning of data is not clearly defined.
Provenance (Origin)	Doubts about the producer of information and the applied method.
Context	The given data deviates from the system context.

Schwab, Laner, Rechberger (2016): Quantitative evaluation of data quality in regional Material Flow Analysis. Journal of Industrial Ecology.
Schwab, Zoboli, Rechberger (2017): A Data Characterization Framework for Material Flow Analysis. Journal of Industrial Ecology, 21(1): 16-25.

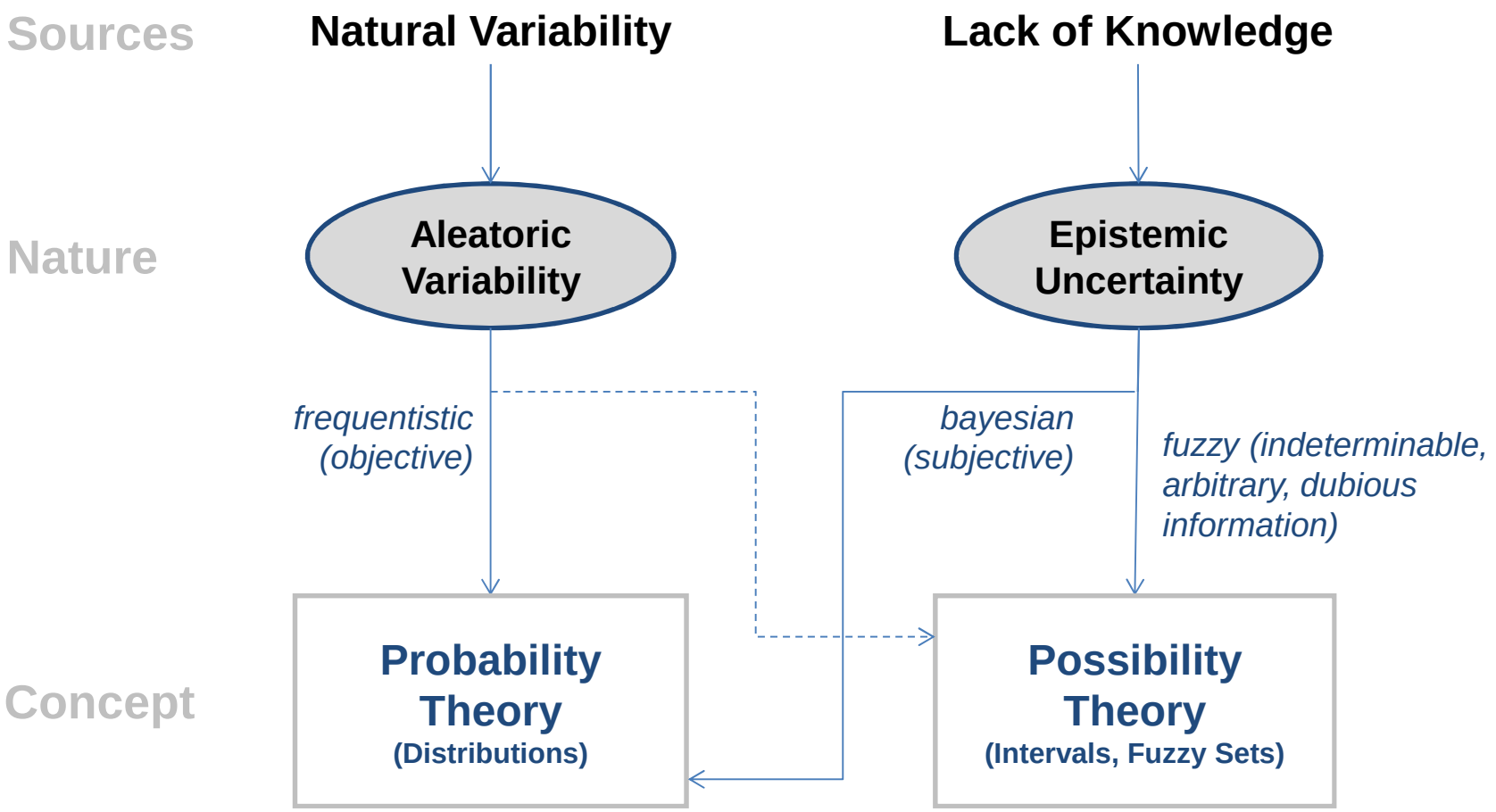
Characterization of Uncertainties

Goal:

Transformation of data quality into uncertainties based on mathematical functions

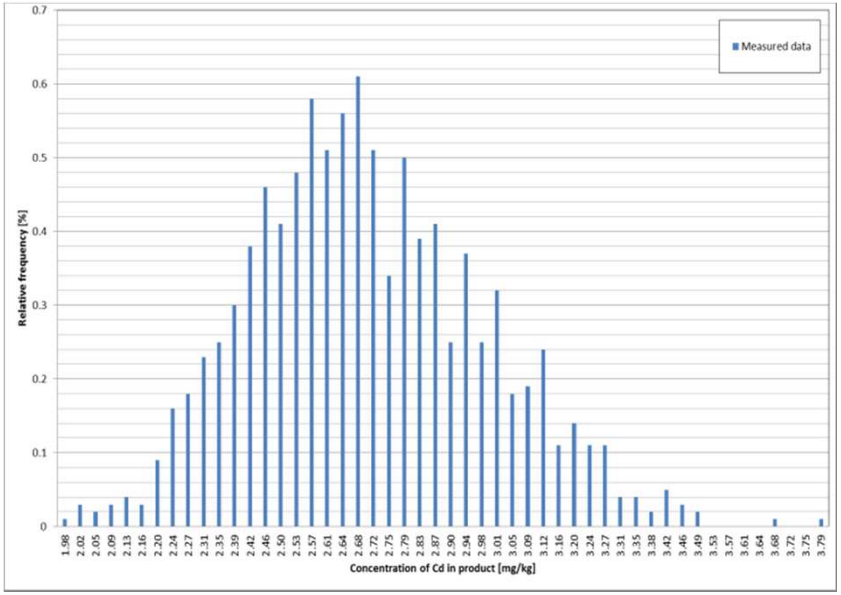
Avoid:

Mathematical description of uncertainty

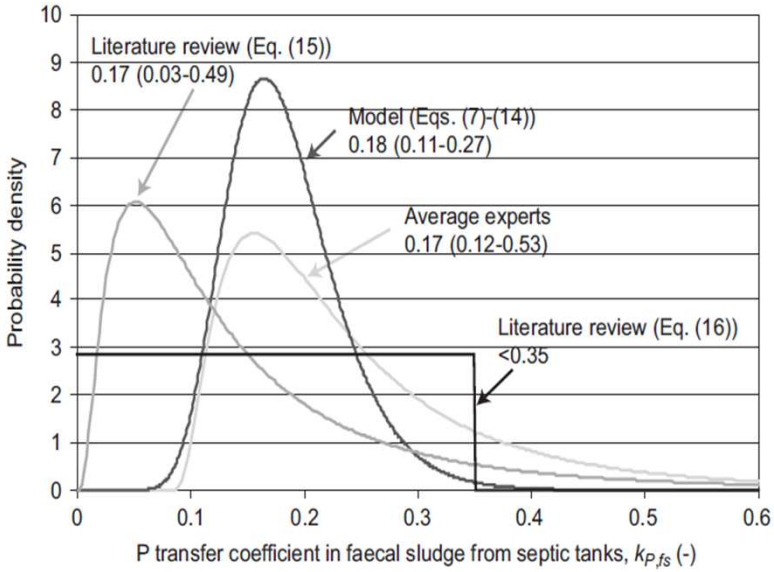


Probability density functions

Repeated Measurements



Expert Estimates

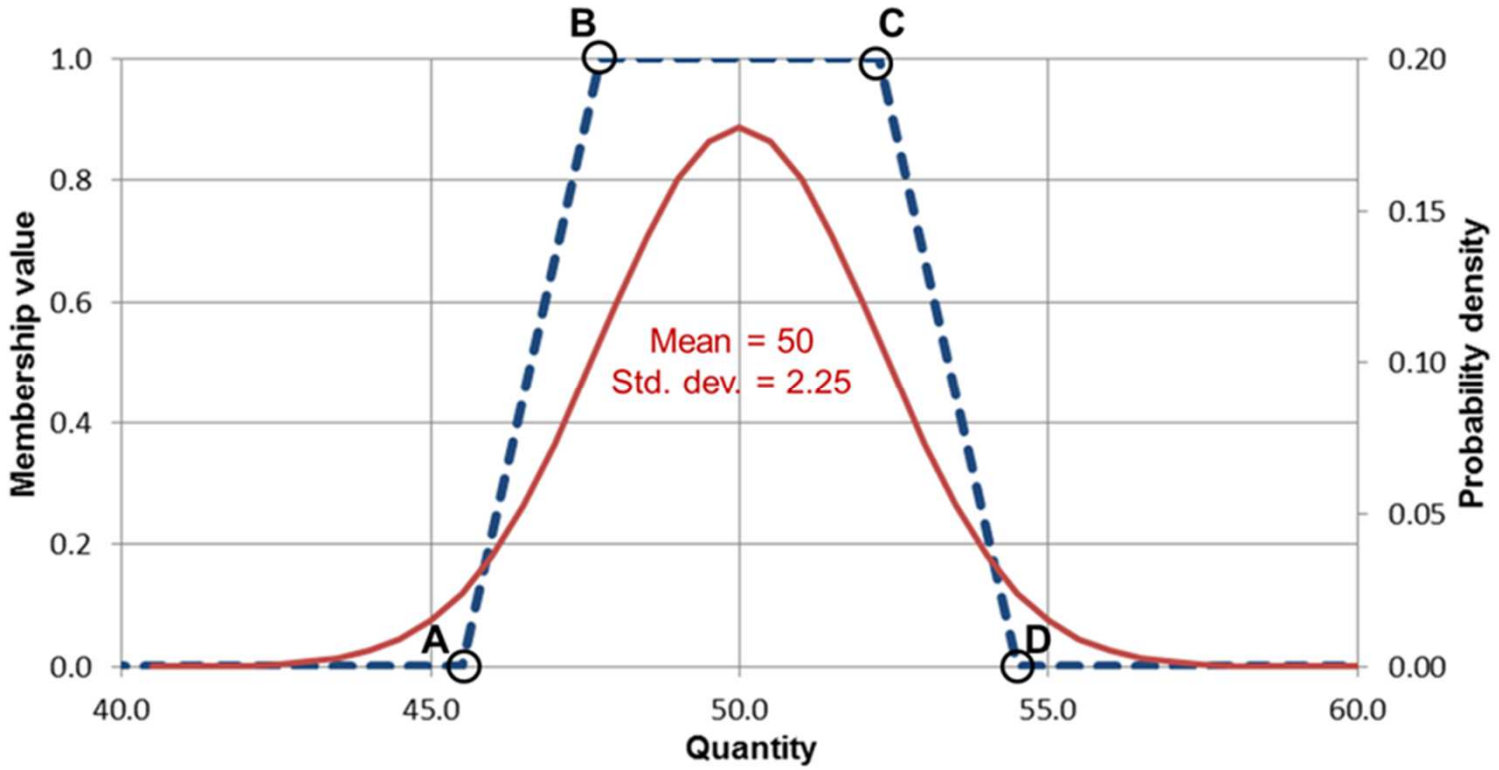


What if we lack the information to construct a PDF?

Membership functions

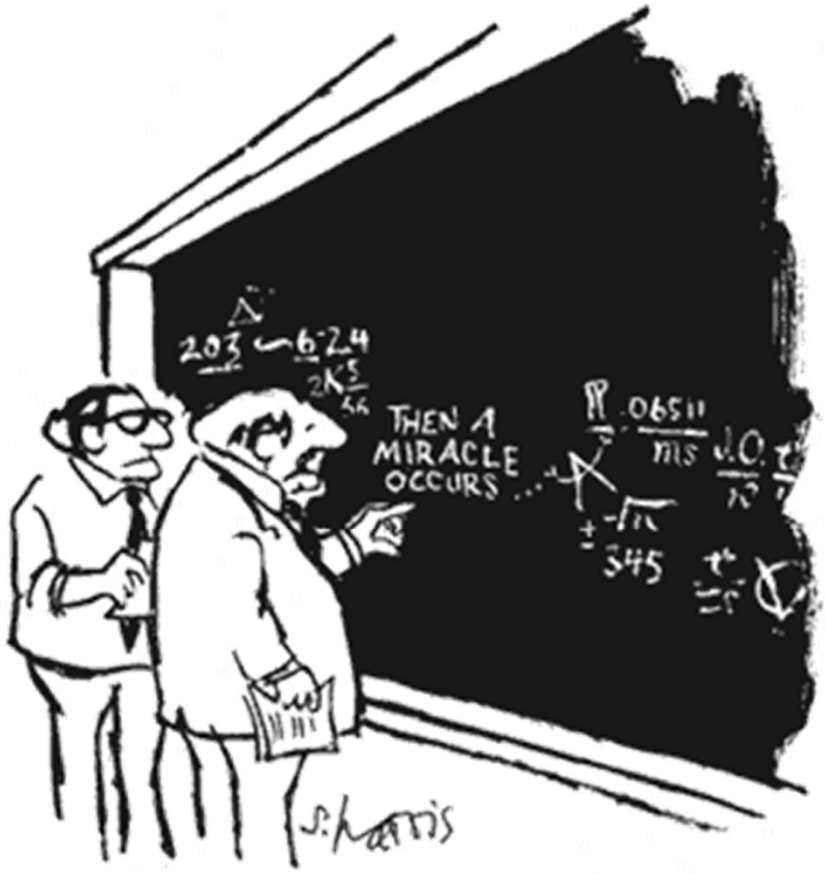
Fuzzy sets to define possible areas (poor information)

- Intervals (Min – Max)
- Membership functions can take on various forms



Laner, Rechberger, Astrup (2015): Applying Fuzzy and Probabilistic Uncertainty Concepts to the Material Flow Analysis of Palladium in Austria. Journal of Industrial Ecology, 19(6): 1055-1069.

Thank you for your attention!



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Source: Simon Harris, 2015

Discussion

- What is a reasonable effort for data characterization and quality evaluation? How can we facilitate it?

- Is uncertainty assessment in Industrial Ecology always subjective?

- Should we consider using different mathematical concepts (e.g. probabilistic vs. possibilistic) for uncertainty caharcaterisation in the same model?