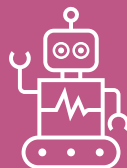# Empowering our teams with Generative AI

RECAP

Aim of the workshop

1

understanding functionalities & capabilities of genAI

understanding the dark side of genAI

understanding concepts of human evaluation

take-aways for using and consuming genAI

What is genAI?

Generative AI (GenAI) refers to **a class of artificial intelligence models** designed to **generate new content** based on input data. This can include generating **text, images, music, code, and other forms of content**. These models learn patterns, structures, and relationships from the training data and use this understanding to produce new, coherent, and contextually relevant outputs.

Key characteristics
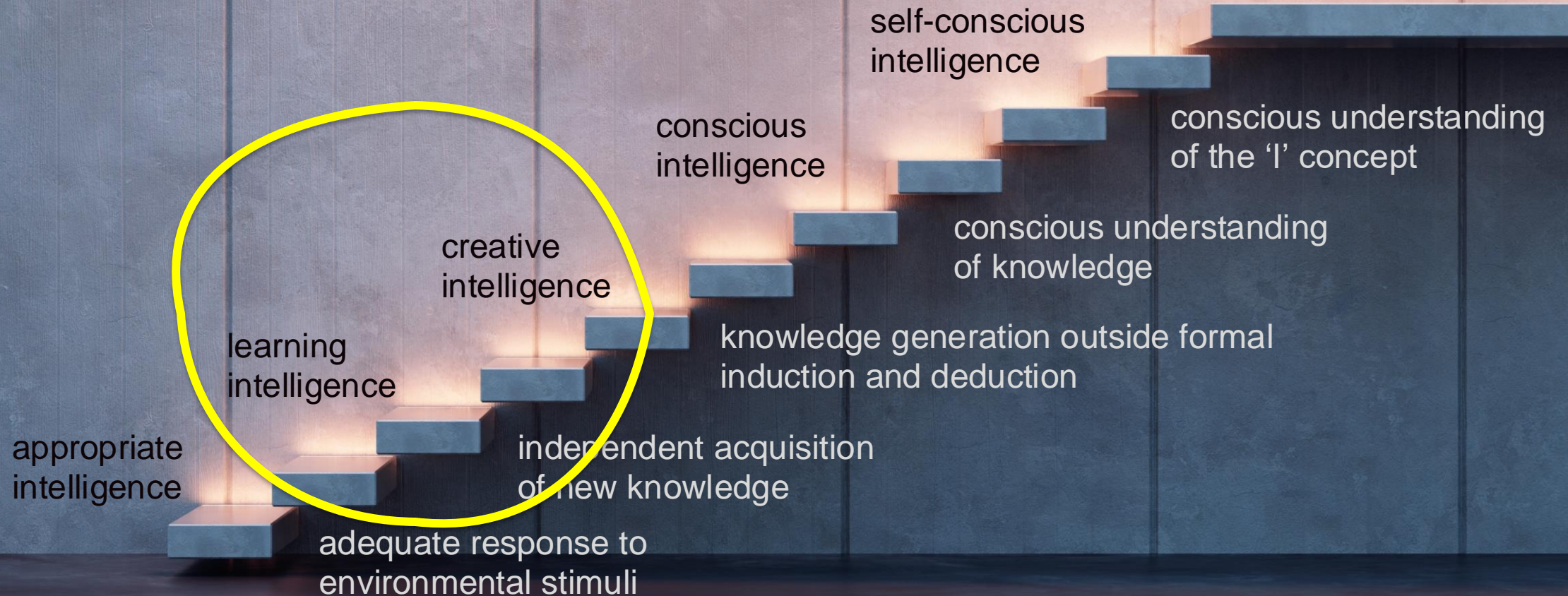
„creativity"

learning from data

versatility

interactivity

1

Capabilities of genAI

what is AI **able to**?	what is AI **not able to**?
pattern recognition	creativity
decision-making based on data and algorithms	emotional intelligence
natural language processing	moral decisions
recognition/analysis of visual data	self-awareness

2

2 Capabilities of genAI

self-conscious intelligence

conscious intelligence

conscious understanding of the 'I' concept

creative intelligence

conscious understanding of knowledge

learning intelligence

knowledge generation outside formal induction and deduction

appropriate intelligence

independent acquisition of new knowledge

adequate response to environmental stimuli

3 Dark sides of genAI

„The term deepfake originated in 2017 by a Reddit user who used deep learning technologies" (**"deep"**) to create fake videos (**"fake"**)" (Mitra et al. 2024).
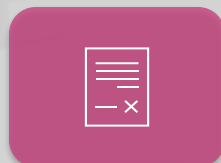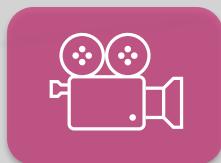
source: Adobe Stock

The European roadmap for AI – approach for excellence and trust

2018
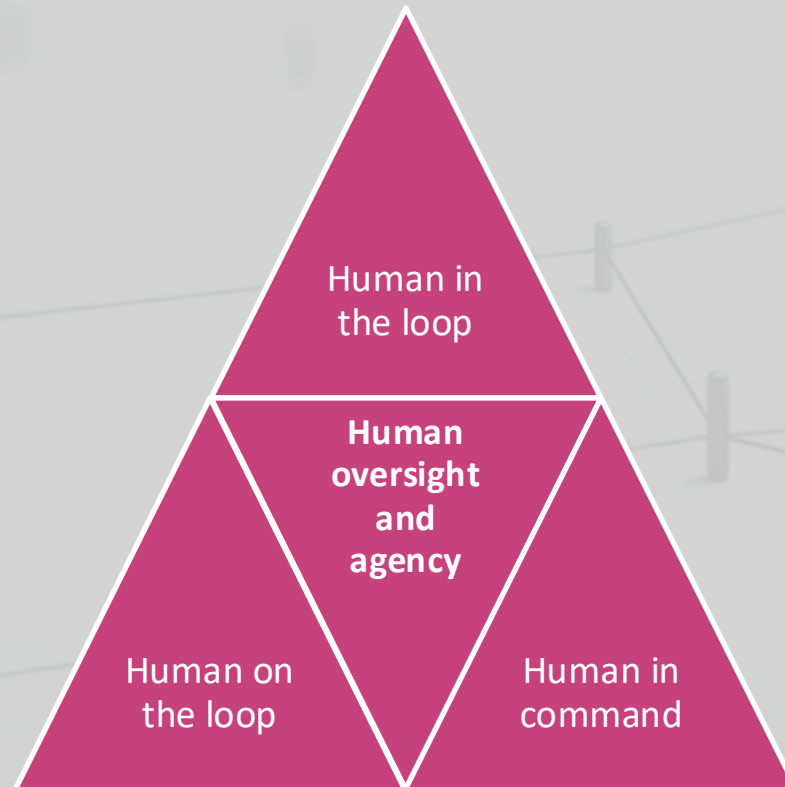- Set up of the high-level expert group for AI

2020
- European White paper on AI (2020)

2021
- Commission: updated coordination plan on AI
- European Commission: Proposal for a regulation laying down harmonised rules on AI

4 Human agency and oversight

Need for proper oversight mechanisms

- Human in the loop
- Human oversight and agency
- Human on the loop
- Human in command

ETHICS GUIDELINES FOR TRUSTWORTHY AI

.. and what does that mean now?

6

Data
- data quality
- data diversity

AI model
- choose the right model
- how was the model trained?

Input
- frame the context
- give detailed information

Output
- double check
- rate and evaluate results
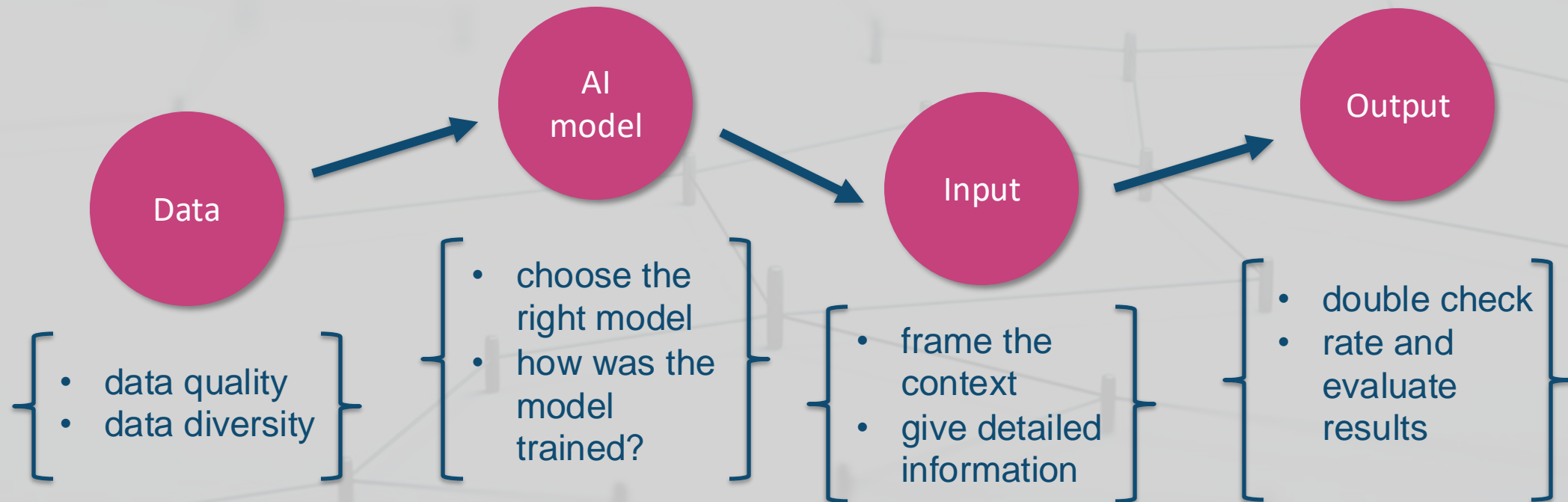
Session
Data Cleaning
Techniques

FOR Generative AI Models

Presenter:

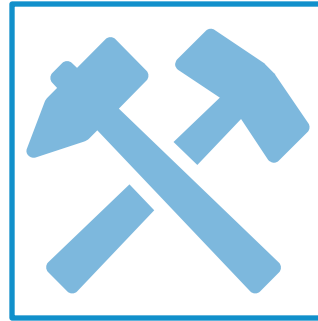**Dr. Sphesihle Ndlela**

Speaker

Introduction and research interests

Sphesihle Ndlela:

Software Developer

Researcher

(Universität Siegen)

Technology:

Data Science

Data Integration

Enterprise Integration

Business Models (AI)

SME Management :

Longevity/Survival

Transformation

Cognitive managerial capabilities(innovation, creative thinking, critical thinking, creative problem solving)

EDIH Südwestfalen   ATTRACT

Usage of Generative AI

- ❑ Generative AI (GenAI) has opened new possibilities for businesses

- ❑ Gen AI generate images, text responses, music and other files by training the huge amount of data

- ❑ Generative AI applications are built on top of large language models (LLMs) models.

- ❑ LLMs are built on a publicly available data example Chat-GPT

- ❑ But that doesn't stop organizations not to use their own data to leverage on this technology.

- ❑ Organizations can borrow existing LLM models alongside their data, or create their own LLM from scratch

LLM and Data

❑ Storage and extraction

LLM and Data

Where do Generative AI models get data to process?

Storage Type	Description	Example
Unstructured	Can't be organized in databases.	Email, PDFs, text documents, call center recordings, chat logs and video footage , etc.
Semi-structured	Contains some information in defined data fields	HTML,XML , CSV, etc.

EDIH Südwestfalen

ATTRACT

LLM and Data

Where do Gen AI models get data to process?

Storage Type	Description	Example
Structured data	Defined spreadsheet or defined database fields	Server log files, Internet of Things (IoT) sensor data, customer relationship management (CRM) databases and enterprise resources planning (ERP)

Data extraction synopsis

How is the data extracted?

Extracting :data from databases such as (SQL) establish connections

The data extracted from the structured database can be saved in **semi-structured** format, i.e CSV, JSON, XML etc.

Process and clean the data to the format like JSON(or any format expected by the model).

Pass the formatted file to the model and process.
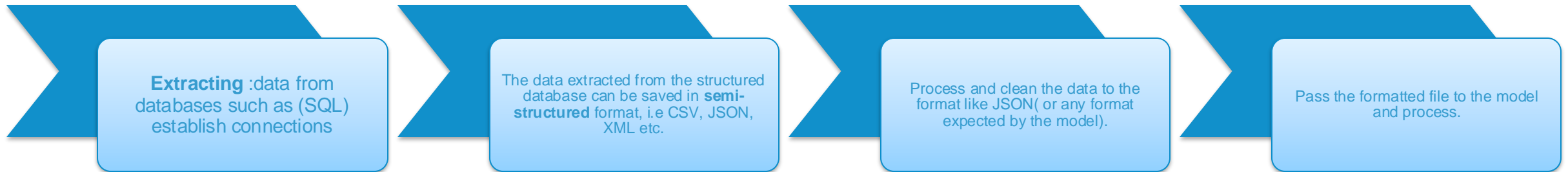
Data quality

- The quality of your data will impact the performance of Gen AI models for your business.

- The quality of data fed into Gen AI must be:

- Accurate

- Relevent

- Complete and

- Must have no baises

The Gen AI models are therefore as good as the data at which they are trained (Databricks, 2024).
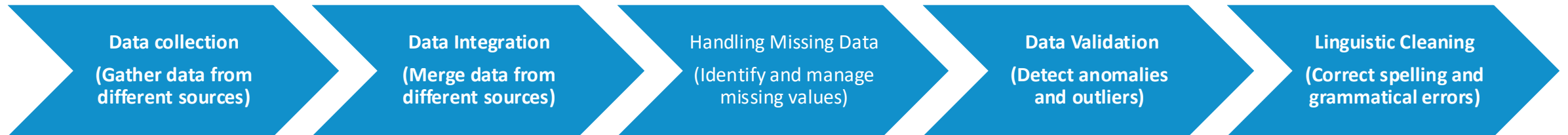
Ways to improve data quality

❑ Implement Robust Data Cleaning Processes

❑ Identifying and **handling outliers** using statistical methods, **visualization**, and **domain expertise**

❑ Addressing **missing** or **incomplete** data through imputation, deletion, or flagging

Data Cleaning Techniques for Generative AI

Synopsis

Data collection

(Gather data from different sources)

Data Integration

(Merge data from different sources)

Handling Missing Data

(Identify and manage missing values)

Data Validation

(Detect anomalies and outliers)

Linguistic Cleaning

(Correct spelling and grammatical errors)

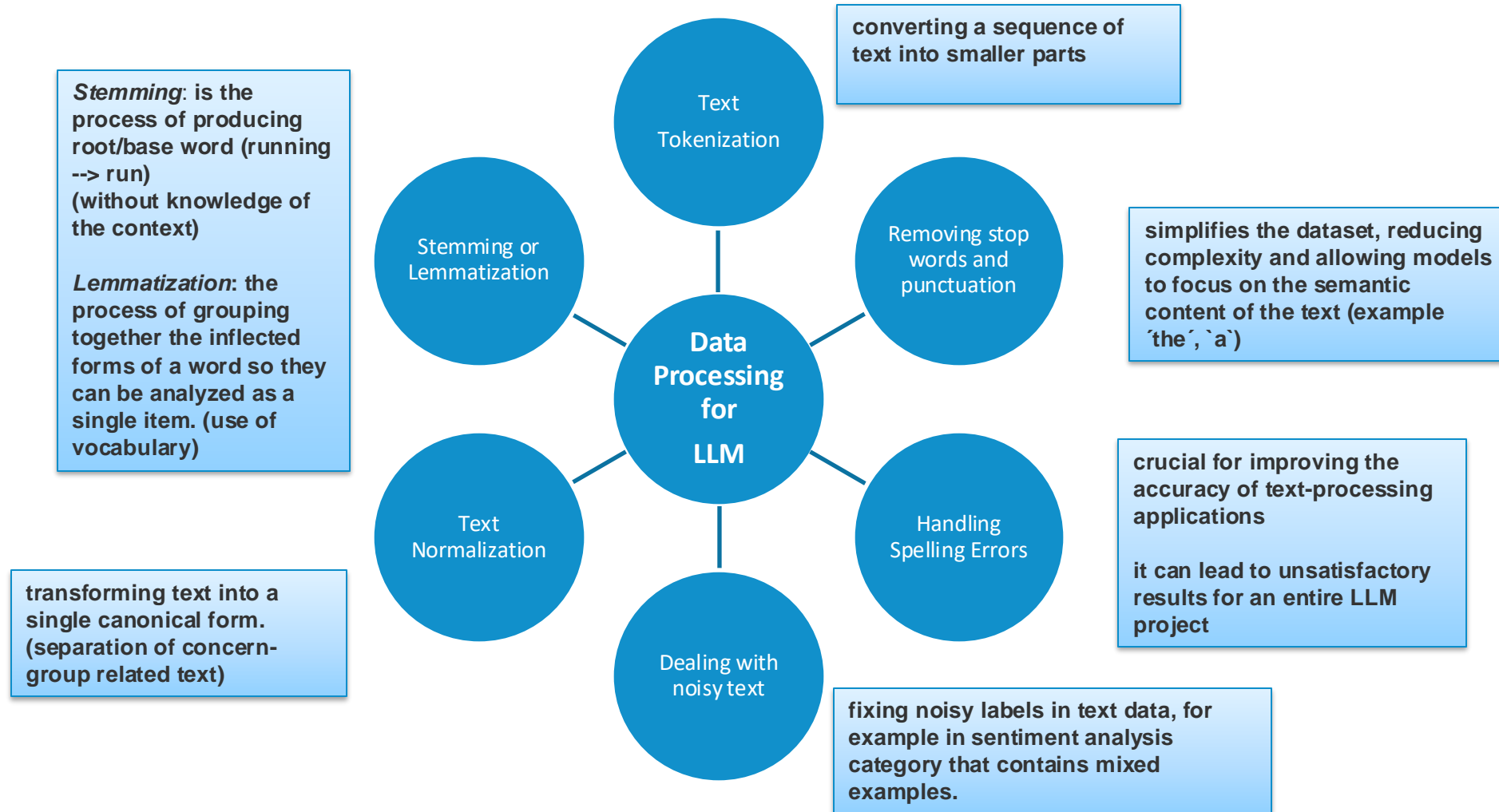EDIH Südwestfalen ATTRACT

Data Preparation and Cleaning Process

General common practises of cleaning: Tubular format
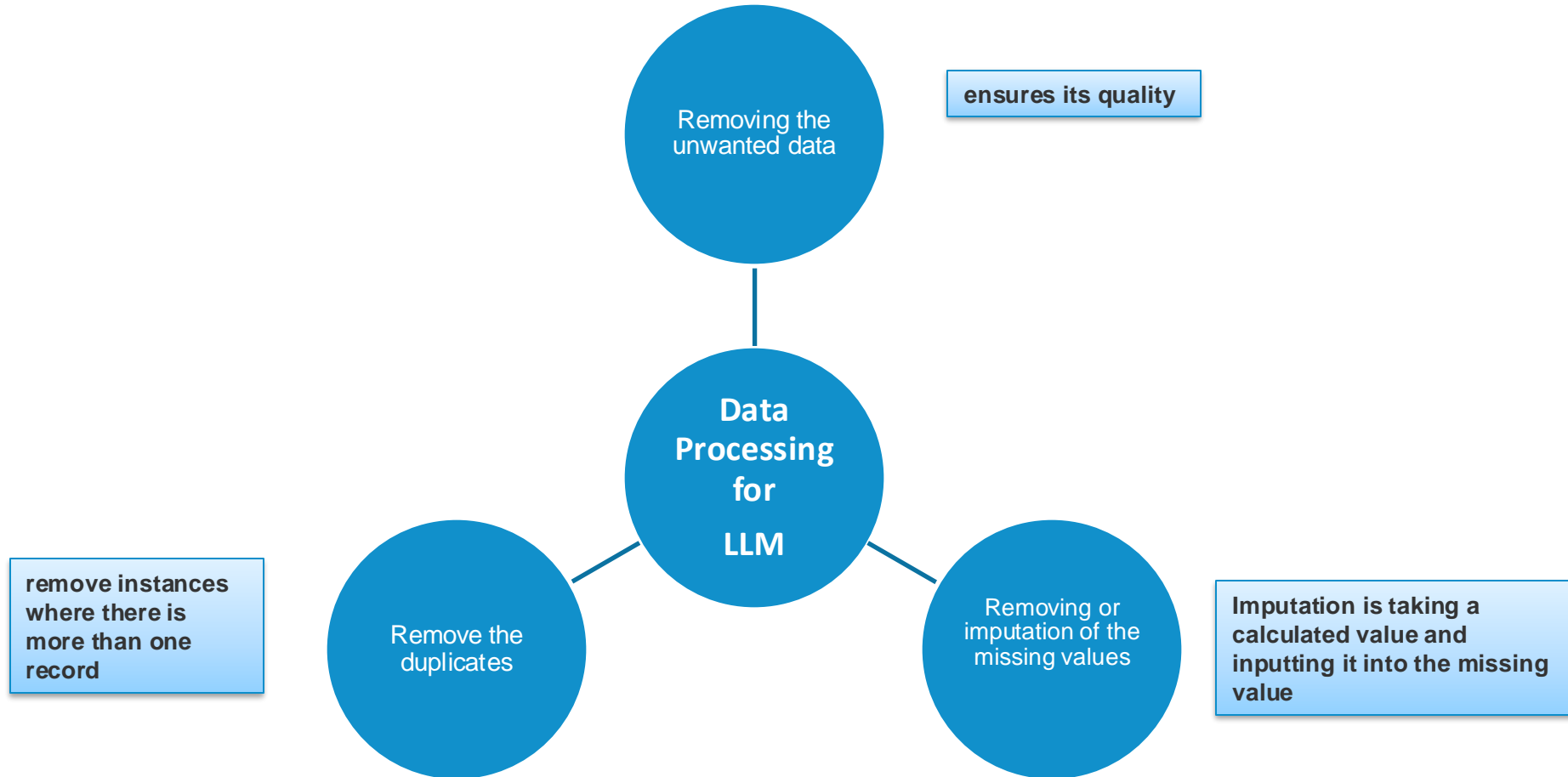
Data cleaning and noise reduction

Common noise reduction techniques in the text

Stemming: is the process of producing root/base word (running --> run) (without knowledge of the context)

Lemmatization: the process of grouping together the inflected forms of a word so they can be analyzed as a single item. (use of vocabulary)

converting a sequence of text into smaller parts

simplifies the dataset, reducing complexity and allowing models to focus on the semantic content of the text (example ´the´, `a`)

crucial for improving the accuracy of text-processing applications

it can lead to unsatisfactory results for an entire LLM project

transforming text into a single canonical form. (separation of concern-group related text)

fixing noisy labels in text data, for example in sentiment analysis category that contains mixed examples.

Text Tokenization

Removing stop words and punctuation

Stemming or Lemmatization

Data Processing for LLM

Text Normalization

Handling Spelling Errors

Dealing with noisy text

Key Benefits of Data Cleaning

❑ **Ensure Accuracy**: Eliminates mistakes and maintains consistency, reducing the risk of model confusion and hallucinations.

❑ **Improve Quality:** Ensures models operate with reliable, consistent information for accurate inferences.

❑ **Facilitate Analysis:** Simplifies data interpretation and analysis; for instance, plain text models may struggle with tabular data.

Data cleaning conclusions

- ❑ **No One-Size-Fits-All**: Data cleaning varies and lacks a universal template.

- ❑ **Critical Thinking**: Data scientists must continuously ask questions and delve deeper for meaningful models.

- ❑ **Pre-Model Checklist**: Ensure data is:

  - Clean

  - Consistent

  - Accurate

End

Thank you all for attending this session.