Unlocking Data Insights - Introduction to Data-Centric AI

Introduction

Funded by the European Union NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Future Artificial Intelligence Research
UniBa Università degli Studi di Bari Aldo Moro
KDDE

Before going to the deeper ...

How many of you have ever heard of Data-Centric AI?

Data-Centric AI: transforming raw data into smart data

Executive summary

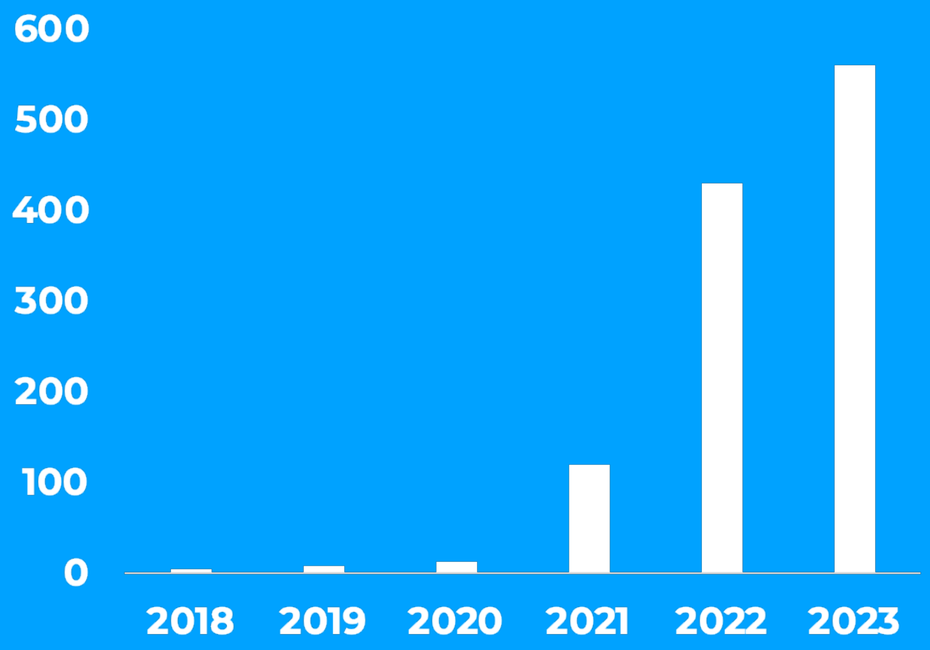
- From Model-Centric AI to Data-Centric AI
- Data-Centric AI pipeline
- Data-Centric AI Open Challenges + Project Ideas

Source: dalle2.gallery

Data-Centric AI: SOTA Trend

The statistics are collected by querying Google Scholar with exactly matched phrase “Data-Centric AI”

Tendency of “Data-Centric AI” topic over the past years



DATA-CENTRIC AI in Scientific Events

NeurIPS Data-Centric AI Workshop 2021

35th International Conference on Advanced Information Systems Engineering

NeurIPS 2023 Datasets and Benchmarks Track

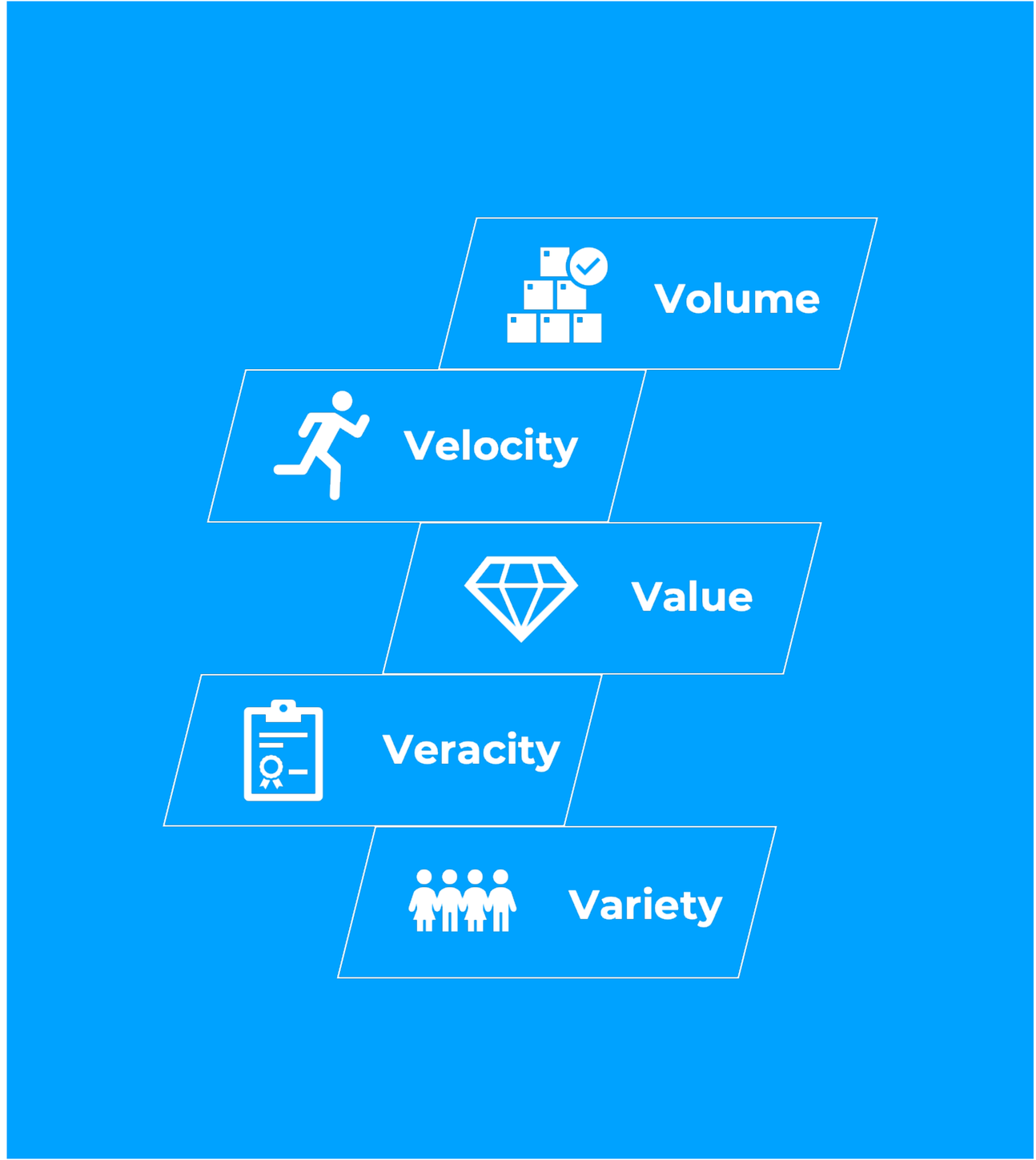
From Model-Centric to Data-Centric

+ Encourages model advancements

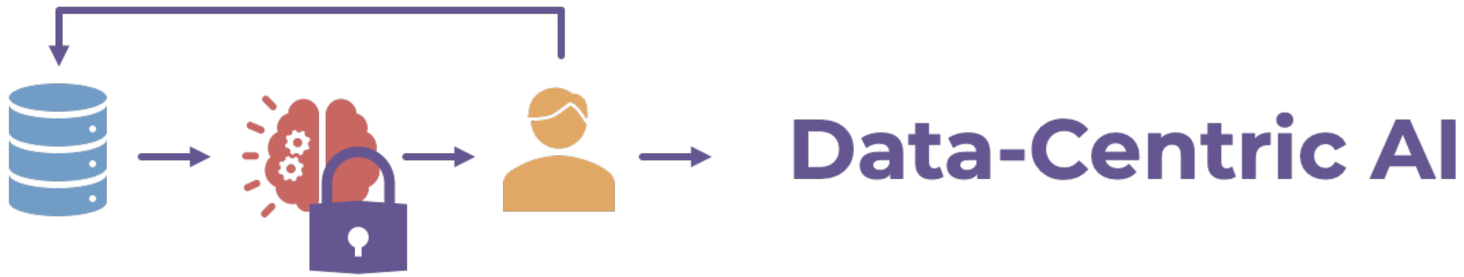
- Requiring high trust in data

From Model-Centric to Data-Centric

Due to the increasing availability of big data in multiple scenarios (e.g. satellite data, process logs)



From Model-Centric to Data-Centric



+ Start to pay more attention to the data used to build powerful AI systems

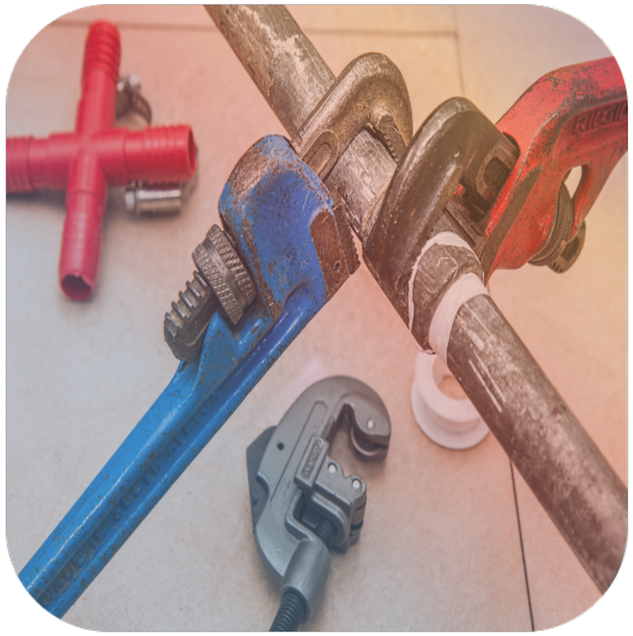
Data-centric AI is the discipline of systematically engineering the data used to build an AI system.

Source: Data-Centric AI Resource Hub

Data-Centric AI - Big Picture

Training Data Development

Inference Data Development



Data Maintenance

Training Data Development

1. Data Collection

- Generate new data from scratch
- Dataset discovery
- Data integration

2. Data Labeling

- Crowsourcing
- Semi-supervised labeling
- Active learning

3. Data Preparation (extracting smart data from raw data)

- Data cleaning
- Feature extraction
- Data transformation

4. Data Reduction/Augmentation

- Feature selection
- Summary extraction
- Data augmentation

Inference Data Development

In-distribution Evaluation

- Data slicing
- Algorithmic resource

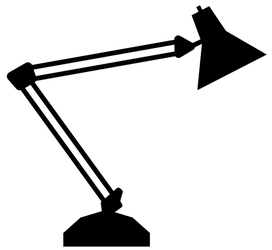
Out-distribution Evaluation

- Adversarial perturbation
- Distribution shift

Prompt Engineering

- Manual prompting
- Automated prompting

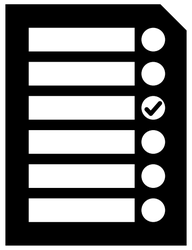
Data Maintenance



Data understanding

•Data visualization to represent data in a more intuitive form

•Data valuation to understand what type of data is most valuable



Data quality assurance

•Quality assessment to determine the value of data

•Quality improvement in different stages of a data pipeline (domain experts, collective intelligence ...)

Data acceleration

•Resource allocation to balance resources and minimize throughput and latency

•Query acceleration to achieve rapid data retrieval by minimizing the number of disk accesses

Data-Centric AI Keyword Cloud

Data-Centric AI Example: ChatGPT

Where is the novelty?

Training Data Development

The quantity and quality of the data used for training GPT models have seen a significant increase through better data collection, data labeling, and data preparation strategies

Data-Centric AI Example: ChatGPT

GPT history
GPT-1

Dataset: BooksCorpus
This dataset contains 4629.00 MB of raw text, covering books from a range of genres such as Adventure, Fantasy, and Romance

Data-centric AI strategies: None
Result: Pertaining GPT-1 on this dataset can increase performances on downstream tasks with fine-tuning

Data-Centric AI Example: ChatGPT

GPT history
GPT-2

Dataset: WebText
This is an internal dataset in OpenAI created by scraping outbound links from Reddit.

Data-centric AI strategies:
1. Curate/filter data by only using the outbound links from Reddit, which received at least 3 karma.
2. Use tools Dragnet and Newspaper to extract clean contents.
3. Adopt de-duplication and some other heuristic-based cleaning (details not mentioned in the paper)

Result: 40 GB of text is obtained after filtering. GPT-2 achieves strong zero-shot results without fine-tuning

Data-Centric AI Example: ChatGPT

GPT history
GPT-3

Dataset: Common Crawl
Common Crawl is a nonprofit 501(c)(3) organization that crawls the web and freely provides its archives and datasets to the public

Data-centric AI strategies:
1. Train a classifier to filter out low-quality documents based on the similarity of each document to WebText, a proxy for high-quality documents
2. Use Spark's MinHashLSH to fuzzily deduplicate documents
3. Augment the data with WebText, books corpora, and Wikipedia.

Result: 570GB of text is obtained after filtering from 45TB of plaintext (only 1.27% of data is selected in this quality filtering). GPT-3 significantly outperforms GPT-2 in the zero-shot setting

Data-Centric AI Example: ChatGPT

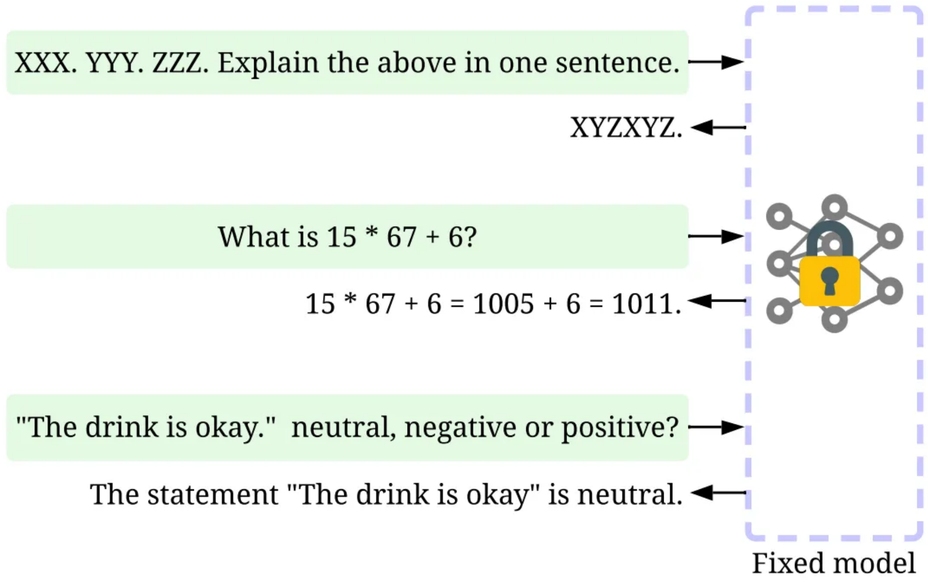
GPT history
InstructGPT

Let humans evaluate the answer to tune GPT-3 so that it can better align with human expectations. They have designed tests for annotators, and only those who can pass the tests are eligible to annotate. They have even designed a survey to ensure that the annotators enjoy the annotating process

Data-centric AI strategies:
1. Use human-provided answers to prompts to tune the model with supervised training.
2. Collect comparison data to train a reward model and then use this reward model to tune GPT-3 with reinforcement learning from human feedback (RLHF)

Results:
InstructGPT shows better truthfulness and less bias, i.e., better alignment

Data-Centric AI Example: ChatGPT



Source: S. Salehi and A. Schmeink, "Data-Centric Green Artificial Intelligence: A Survey" in IEEE Transactions on Artificial Intelligence, vol. 1, no. 01, pp. 1-18, 5555.

Where is the novelty?

Inference data development

As recent GPT models are already sufficiently powerful, we can achieve various goals by tuning prompts (or tuning inference data) with the model fixed. For example, we can conduct text summarization by offering the text to be summarized alongside an instruction like “summarize it”

Data-Centric AI Example: ChatGPT

Where is the novelty?

Data maintenance

ChatGPT/GPT-4, as a commercial product, is not only trained once but rather is updated continuously and maintained.

Data-centric AI strategies:

1. Continuous data collection: When we use ChatGPT/GPT-4, our prompts/ feedback could be, in turn, used by OpenAI to further advance their models. Quality metrics and assurance strategies may have been designed and implemented to collect high-quality data in this process

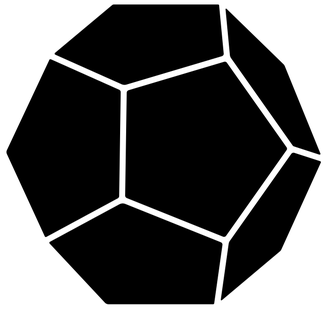
2. Data understanding tools: Various tools could have been developed to visualize and comprehend user data, facilitating a better understanding of users' requirements and guiding the direction of future improvements

3. Efficient data processing: As the number of users of ChatGPT/GPT-4 grows rapidly, an efficient data administration system is required to enable fast data acquisition

Data-Centric AI Open Challenges*

1. **Inference Data & Data Maintenance** - E.g. handling with concept drift, adversarial samples
2. **Cross-task Techniques** E.g. transforming raw data into smart data within the model learning and coupled with the model explanation
3. **Data-model Co-Design** E.g. Explainability as enabler of data model co-design
4. **Data Bias** E.g. Multi-View Data, Imbalanced data, Feature Robustness vs Accuracy
5. **Benchmarks** E.g. Multi-Objective Data, Data Quality, Explainability

*Zha, Daochen, et al. "Data-centric ai: Perspectives and challenges." Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics, 2023.



Smart Data

Variety, Velocity, Value, Veracity Volume are some of the major challenges of big data.

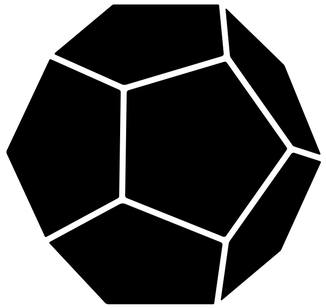
We need to handle these challenges to produce smart representations of big raw data:

- **Context-aware embeddings** (e.g Vord2Vec, BERT)
- **Attention, self-attention** (e.g. ViTs)

...

Difficulty: ★★★☆☆

Creativity: ★★★★★



Smart Data

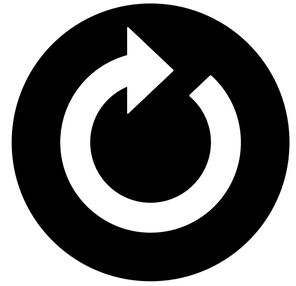
Project idea

New smart representations

Recent stream learning literature has explored different approaches to handle concept drifts paving the way for handling the challenge in the Data-Centric AI paradigm

To handle concept drifts:

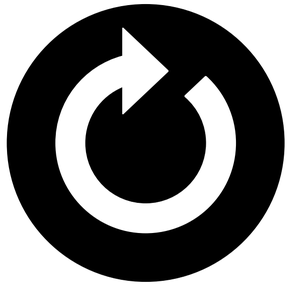
- **Adaptive algorithms** (e.g. Hoeffding Adaptive Tree, Adaptive Random Forest, ...)
- **Periodic updating of data**
- **Selection of more stable data**



Concept Drift

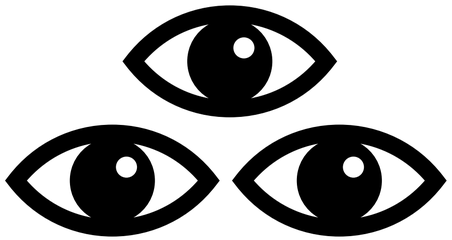
Difficulty: ★★★★★

Project idea



Concept Drift

New deep-learning solutions in stream environment

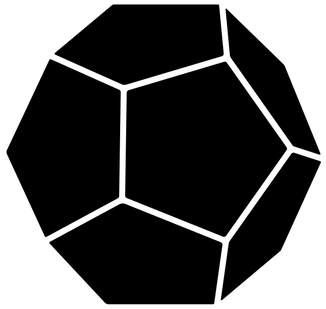


Multi-view & Multi-Objective data

Big amounts of data from several perspectives with different objectives collected for the same phenomenon

- **Health dataset:** exams (tabular data), x-ray (image data), etc
- **Event log:** activity perspective, resource perspective, etc
- **Sensor data:** data from different sensors/devices
- **Multi-Objective data:** Object Centric Event Logs

Difficulty: ★★★★☆



Multi-view &
Multi-Objective data

Project idea

New multi-view approaches

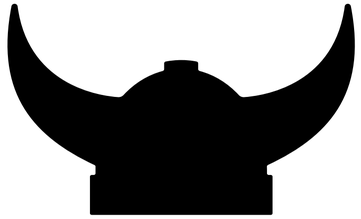
Tips: Contrastive learning?

The Model-Centric paradigm ensures data quality by:

- **Handling different representations of the raw data** (e.g. sequence, image)
- **Removing outliers from data** (e.g., filtering techniques)

Adversarial Samples can compromise the robustness of AI models:

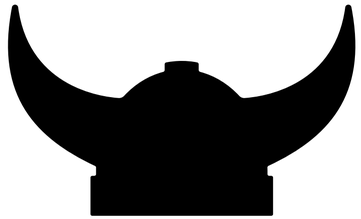
- **Offensive AI**
- **Defensive AI**



Adversarial

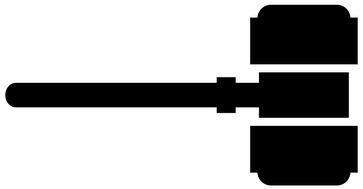
Difficulty: ★★★☆☆

Project idea



Adversarial

Try Adversarial learning on your data



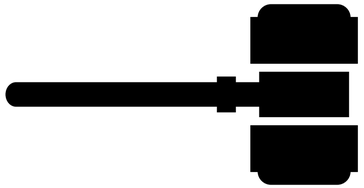
Explainability

Accurate models are not enough

According to the General Data Protection Regulation (GDPR) of the European Union the individual data subject (i.e. the person who was rejected for the loan) **has the right to ask** the business company to motivate the decision

Difficulty: ★★★☆☆

Project idea



Explainability

Novel XAI methods to identify possible data issues in the learning stage

Rare data have the same value of frequent data

Handling Imbalanced Data:

- **Example Selection and/or Generation**
- **Deep metric learning**

Imbalanced

Difficulty: ★★★☆☆

Project idea

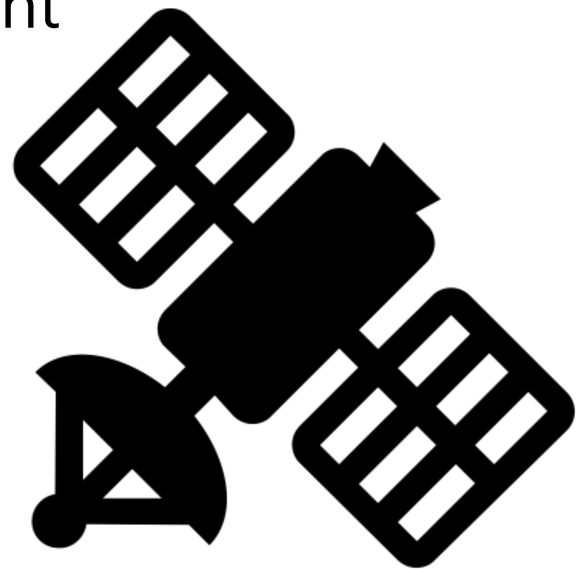
Imbalanced

Model-Centric VS Data-Centric approches

Data-Centric AI - use cases

Data-centric AI in spatial data analysis

[1] has recently described the main principles of the DCAI paradigm in both **remote sensing and geospatial data applications**. This study shows that geospatial **data acquisition and curation** should receive as much attention as data engineering and model development and evaluation

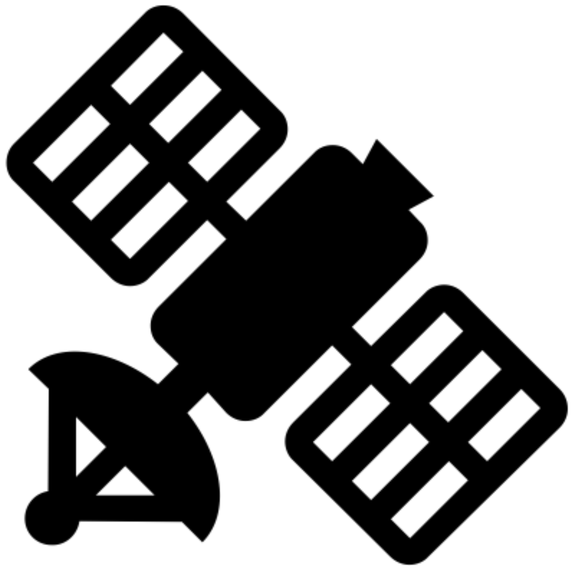


[1] Roscher, R., Rußwurm, M., Gevaert, C., et al., Data-Centric Machine Learning for Geospatial Remote Sensing Data, arXiv, 2023.

Data-centric AI in spatial data analysis

Contrastive learning has been recently used to handle lack of labels in land cover classification by resorting to semi-supervised learning [2]

In addition, [3] has recently explored **data-driven approaches** for deep feature extraction in Sentinel-2 data



[3] Ienco, D., Gaetano, R., and Interdonato, R., A constrastive semi-supervised deep learning framework for land cover classification of satellite time series with limited labels, Neurocomputing, 2024.

[4] Phillips, J., Zhang, C., Williams, B., et al., "Data-Driven Sentinel-2 Based Deep Feature Extraction to Improve Insect Species Distribution Models," EGU General Assembly, 2022

Data-centric AI in Healthcare

Given its myriad capabilities and potential benefits, DCAI is increasingly embraced and integrated into **precision healthcare**, reshaping traditional healthcare into digitized, **patient-centred** healthcare [4]

[5] has recently illustrated a systematic review of emerging information technologies used for data modeling and analytics to achieve **Data-Centric Health-Care (DCHC)** for sustainable healthcare

[4] Oberste, L., and Heinzl, A., “User-centric explainability in healthcare: A knowledge-level perspective of informed machine learning,” IEEE Trans. Artif. Intell., vol. 4, no. 4, pp. 840-857, Aug. 2023

[5] Zahid, A., Poulsen, J., Sharma, R., et al. A systematic review of emerging information technologies for sustainable data-centric health-care. International Journal Of Medical Informatics. 149 pp. 104420 (2021)

Data-centric AI in Industry

Data-Centric AI principles have decisively influenced not only academia but **industrial research and development**.

Small-Medium Enterprises often encounter obstacles such as limited data, lack of labels, data drift and insufficient knowledge in ML and DL techniques which hinder their data science implementation efforts

[7] Luley, P., Deriu, J., Yan, P., et al. From concept to implementation: The data-centric development process for AI in industry. 2023 10th IEEE Swiss Conference On Data Science (SDS). pp. 73-76 (2023)

Data-centric AI in Industry

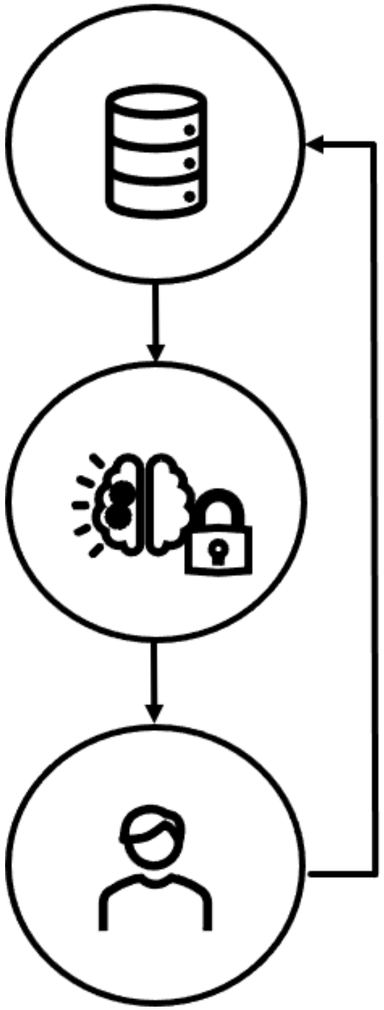
The Data-Centric AI paradigm, however, prioritizes the systematic engineering of data used in constructing an AI system. In [6], the authors describe **a tangible, adaptable implementation of a Data-Centric AI development process** tailored for industrial applications, particularly in machining and manufacturing sectors

[6] Luley, P., Deriu, J., Yan, P., et al. From concept to implementation: The data-centric development process for AI in industry. 2023 10th IEEE Swiss Conference On Data Science (SDS). pp. 73-76 (2023)

Course Agenda

1. Introduction
2. Data-Centric XAI
3. New life to your data
4. Learning from data streams: A gentle introduction
5. Self-supervised learning

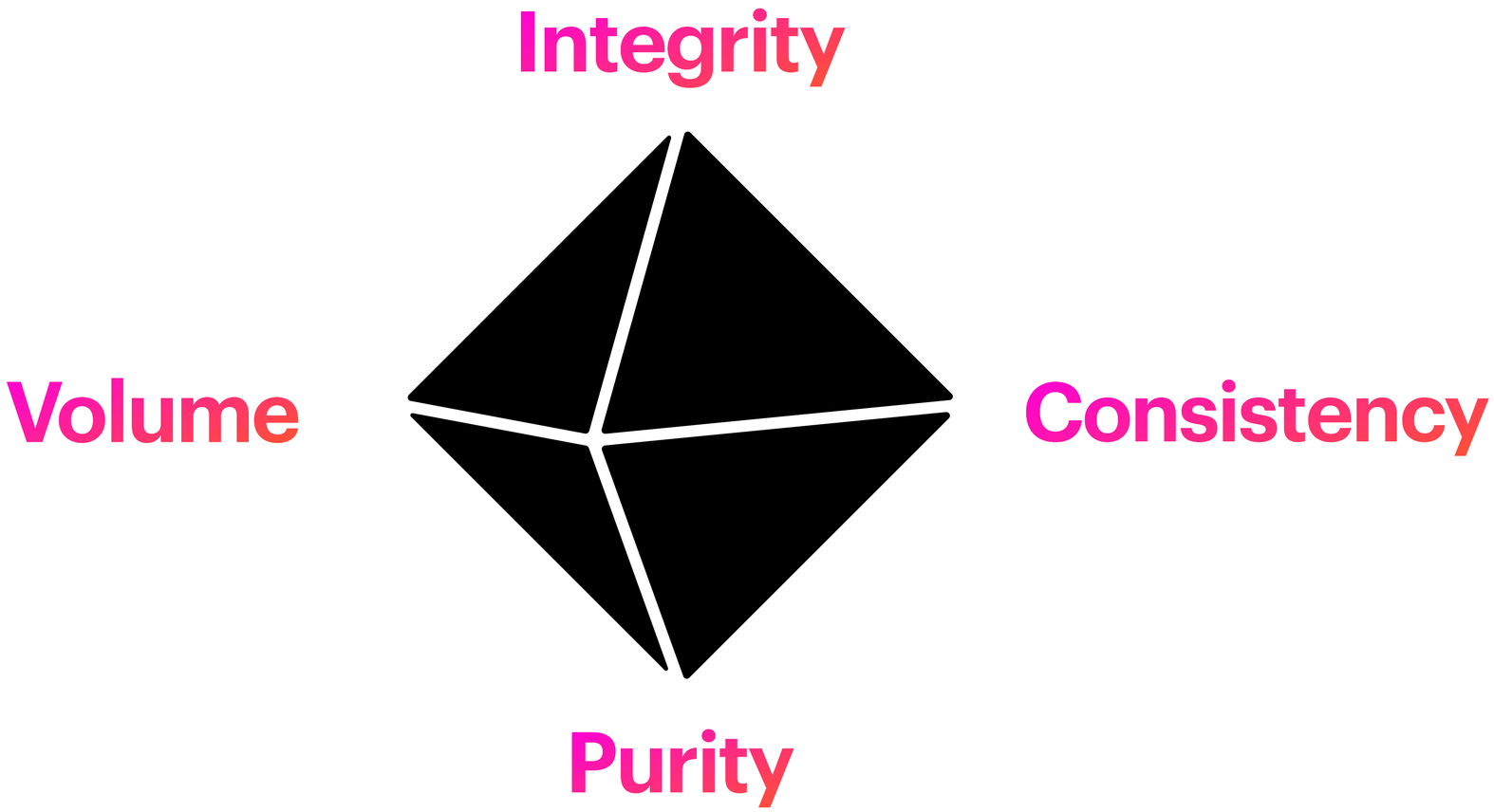
1 Introduction



from Model-Centric to Data-Centric

1 Introduction

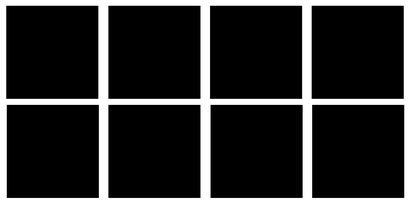
2 **Data-Centric XAI**



1 Introduction

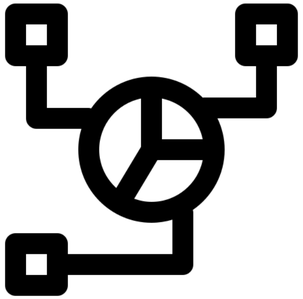
2 Data-Centric XAI

3 **New life to your data**



Contextual Emedding

Image



Hybrid approach

1 Introduction

2 Data-Centric XAI

3 New life to your data

4 **Learning from data streams: A gentle introduction**

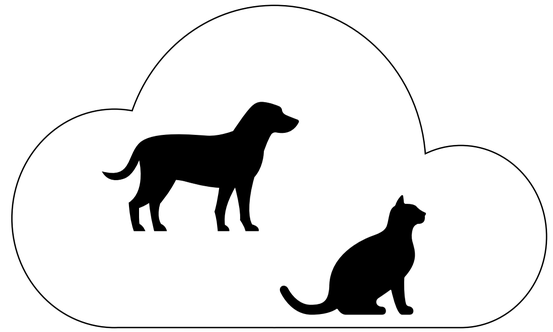
1 Introduction

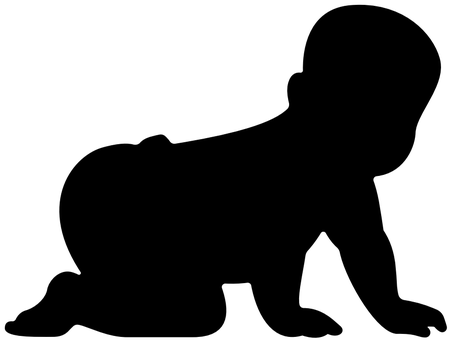
2 Data-Centric XAI

3 New life to your data

4 The influence of the synthetic data on Data-Centric paradigm

5 **Self-supervised learning**





Thanks for attention

Funded by the European Union NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
FAIR Future Artificial Intelligence Research
UniBa UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
KDDE