U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Artificial Intelligence and Society

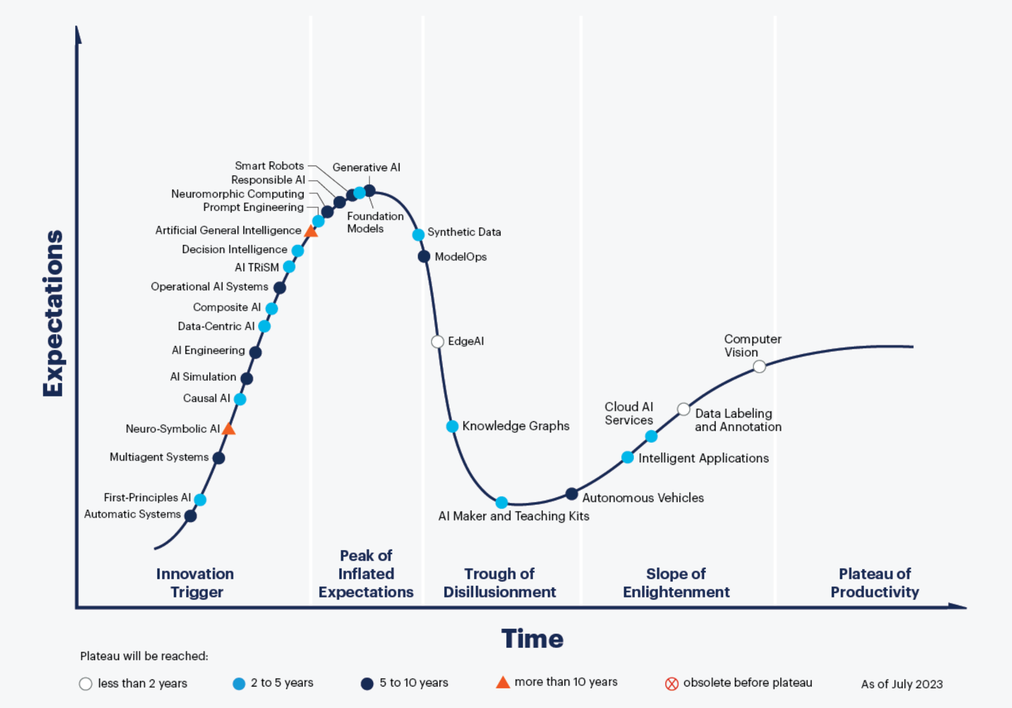
Module 01: Data-Centric AI & Data Profiling

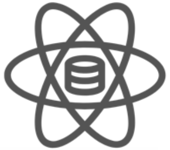
Miriam Seoane Santos
LIAAD, INESC TEC, FCUP, University of Porto
miriam.santos@fc.up.pt

Previously…

Hype Cycle for Artificial Intelligence

Innovation & Impact for Business and Academia





Data-Centric AI

AI TRiSM

Responsible AI

Synthetic Data

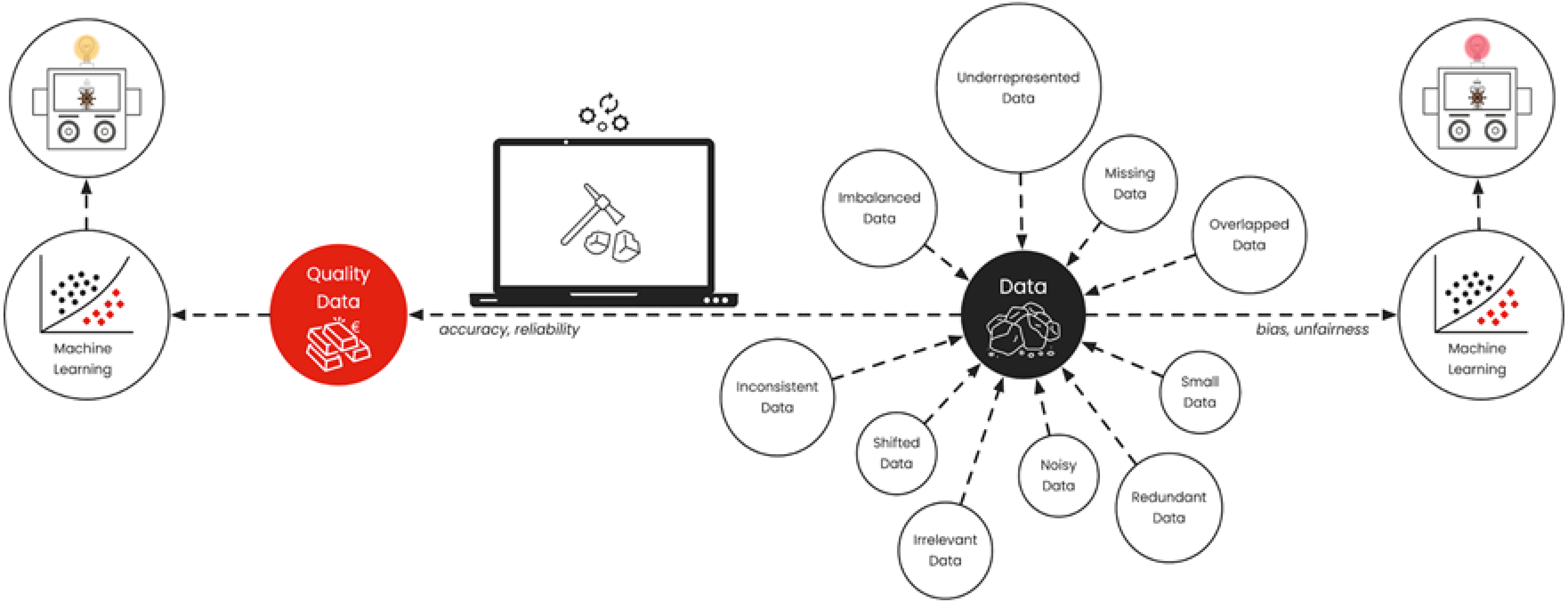
**Gartner, What's New in Artificial Intelligence from the 2023 Gartner Hype Cycle*

Data-Centric AI

An innovation trigger for Machine Learning Research

Data is our most valuable asset

Yet, data quality is (still) the problem

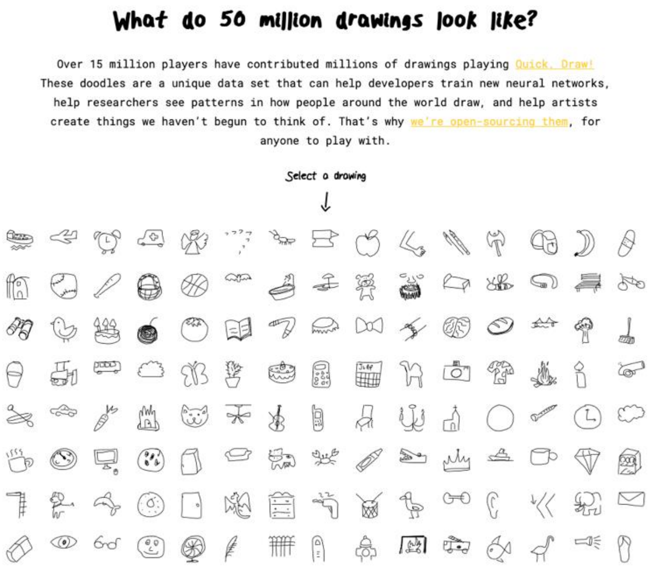


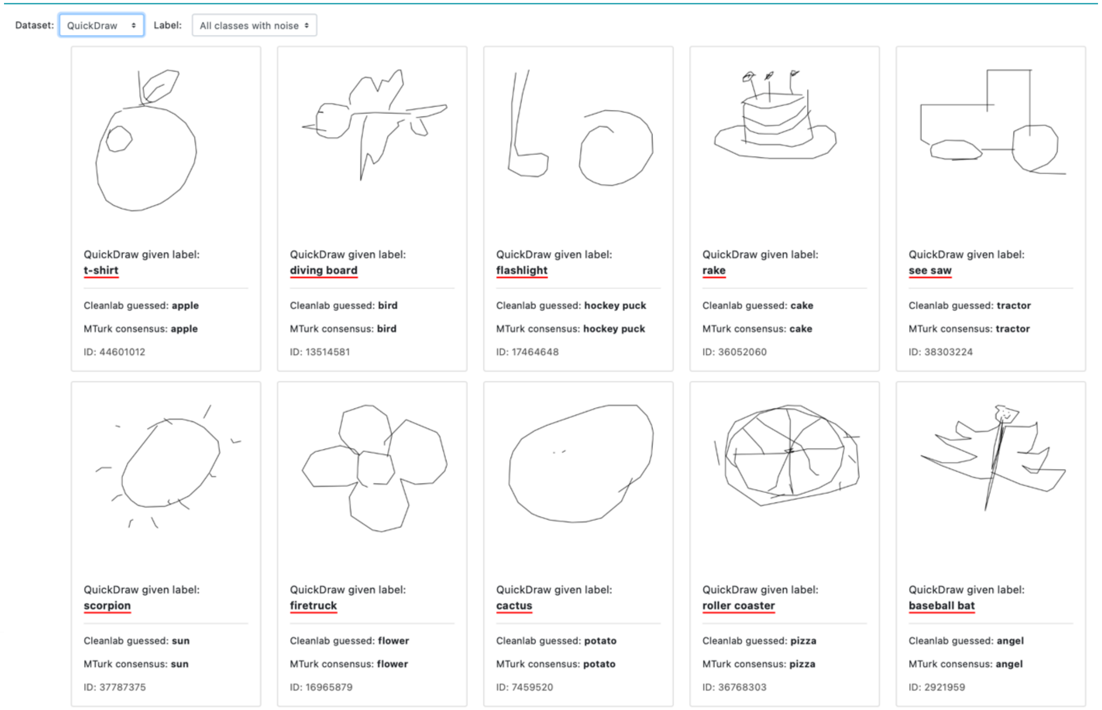
**Forbes, A Data Gap continues to inhibit Artificial Intelligence, 2023*

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Errors are bound to happen in real-world domains

Data quality issues plague almost every industry





***Label Errors in ML Test Sets** *(https://labelerrors.com)*

Imperfect Data versus Smart Data

- Do we need **Big Data** to uncover valuable *(business, research)* insights?
- The 5 V property of the **Big Data Problem**: *Volume, Velocity, Veracity, Variety, Value*

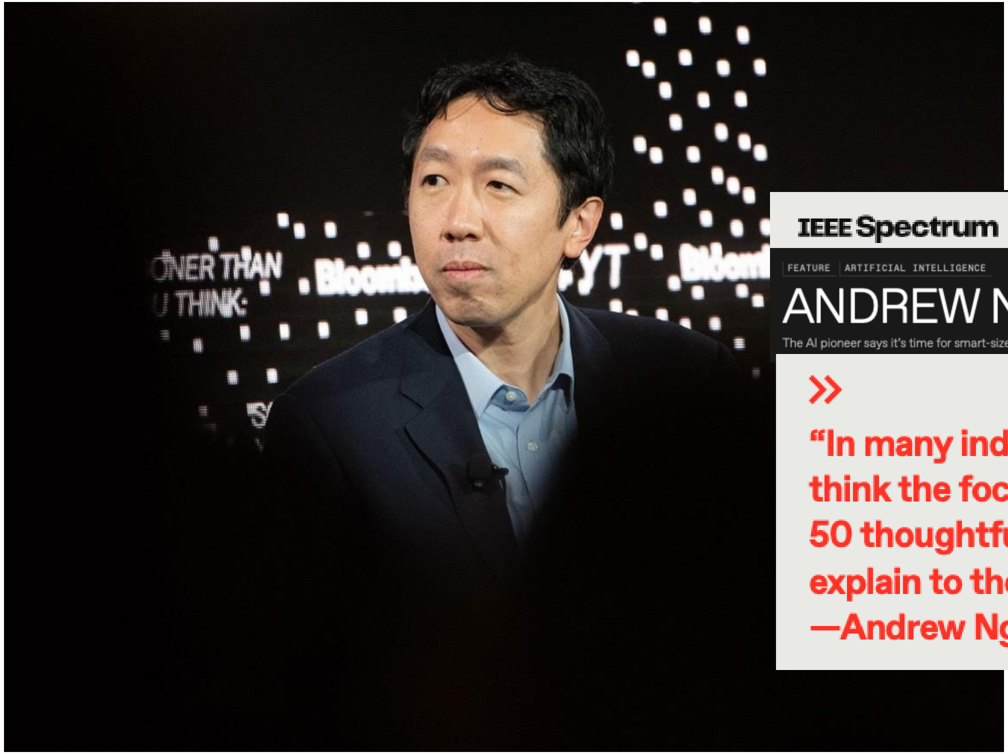
Table 1: Explanation on 5vs of Big Data

No	5V's	Description
1.	Volume	The quantity of data relative to the ability to store and manage it
2.	Velocity	The speed of calculation needed to query the data relative to the rate of change of the data[2]
3.	Variety	A measure of the number of different formats the data exist in (e.g. text, audio, video, logs etc.)
4.	Veracity	Refers to the messiness or the trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable (posts with hashtags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of the content) but big data and analytics technology now allows us to work with these type of data. The volumes often make up for the lack of quality or accuracy.
5.	Value	There is another v to take into account when looking at Big Data: Value! Having access to big data is no good unless we can turn it into value. Companies are starting to generate amazing value from their big data.

"Do we really need to keep stored big amounts of raw data that may be innacurate just for the sake of it? Storing data does not come for free and a way of finding sustainable storage is becoming imperative."

Triguero, Isaac, et al. "Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9.2 (2019): e1289.

Imperfect Data versus Smart Data



IEEE Spectrum FOR THE TECHNOLOGY INSIDER

FEATURE | ARTIFICIAL INTELLIGENCE

ANDREW NG: UNBIGGEN AI

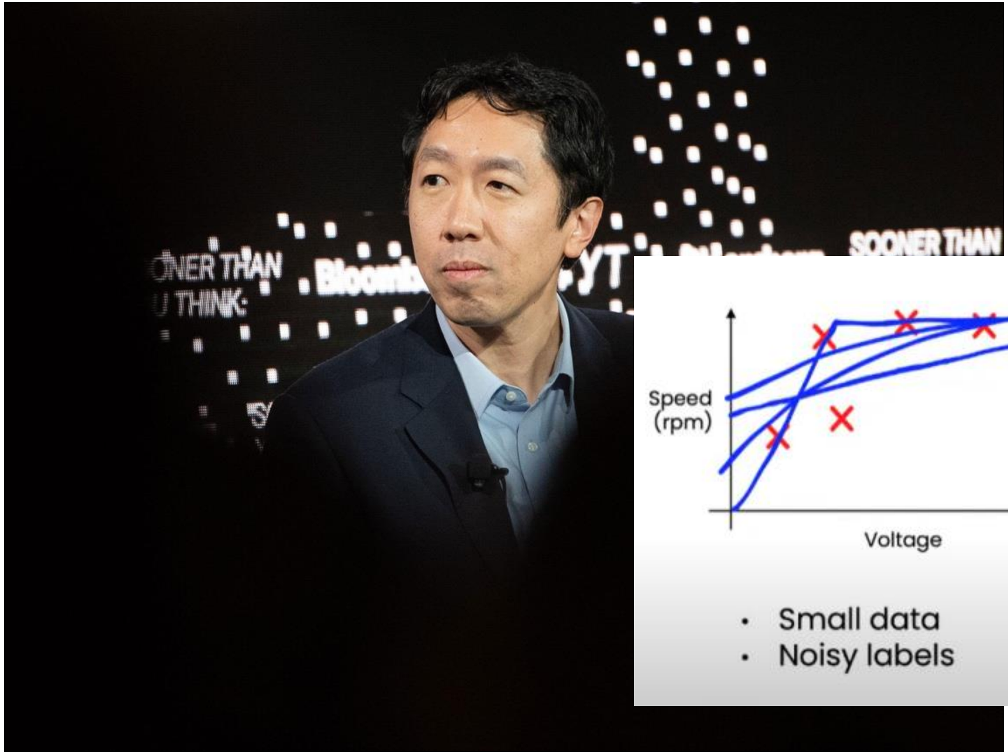
The AI pioneer says it's time for smart-sized, "data-centric" solutions to big issues

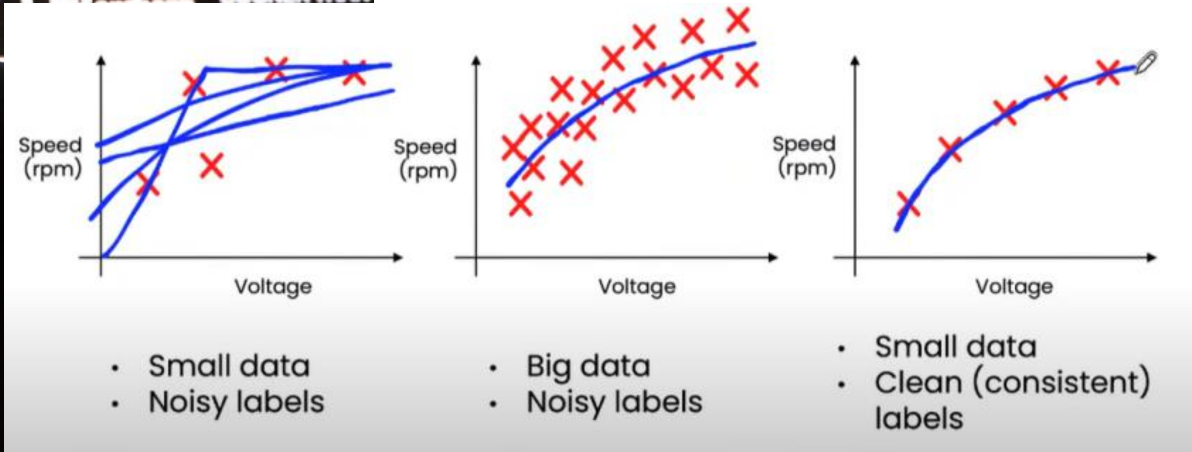
»

"In many industries where giant data sets simply don't exist, I think the focus has to shift from big data to good data. Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn."
—Andrew Ng, CEO & Founder, Landing AI

Andrew Ng: Unbiggen AI, IEEE Spectrum

Imperfect Data versus Smart Data

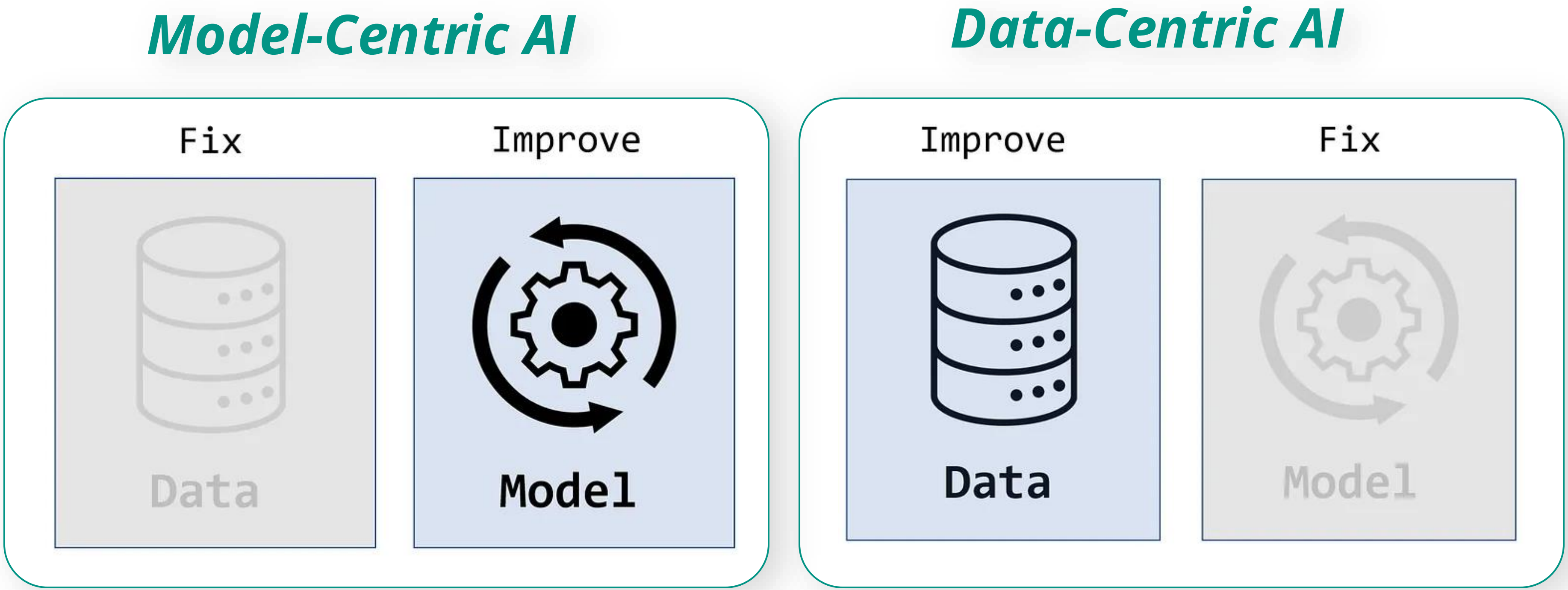




Andrew Ng: Unbiggen AI, IEEE Spectrum

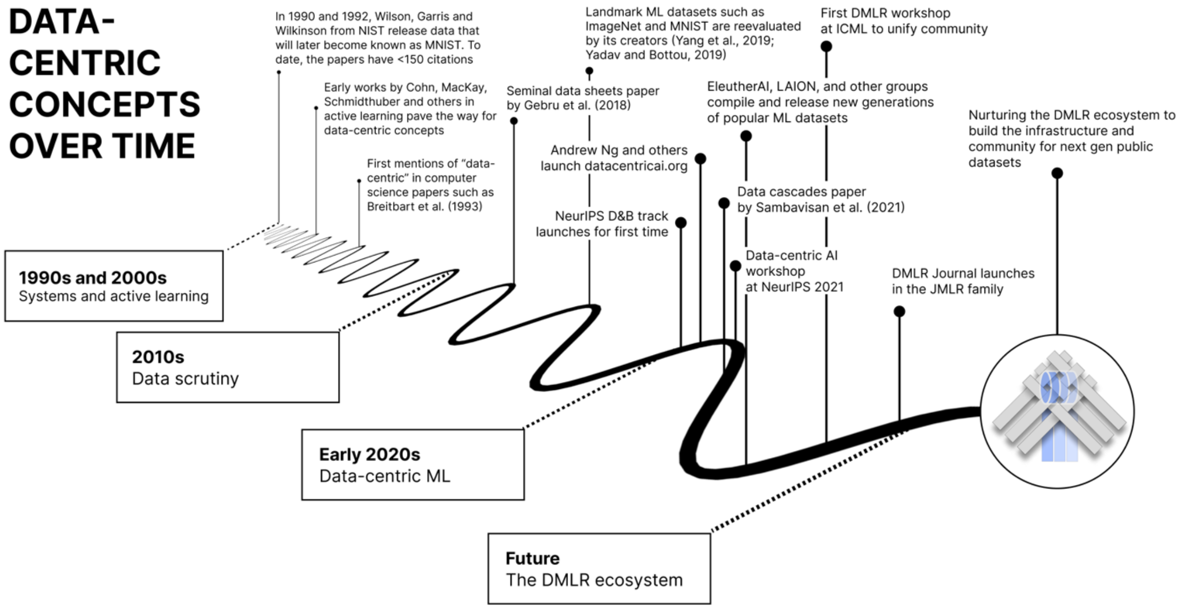
Data-Centric AI Artificial Intelligence

- Model-Centric AI has reached a **point of saturation**. In terms of improvement potential, there is now more gain in shifting our attention towards **improving data.**



Data-Centric AI Artificial Intelligence

- This process of moving from imperfect to intelligent data in a ***systematic and continuous*** manner is referred to as **Data-Centric AI.** Acknowledging the importance of data quality is far from novel ("*garbage-in, garbage out*" mantra). **So, what is the difference between standard Data Preprocessing and Data-Centric AI?**



[DMLR: Data-Centric Machine Learning Research]()

Imperfect Data versus Smart Data

- What ultimately determines the success of machine learning applications is therefore their ability to **transform *dirty, raw, imperfect data* into *high-quality, intelligent, actionable smart data*** i.e., data of sufficient quality to allow classifiers to draw accurate and reliable inferences on the domain (*machine learning perspective*).

Imperfect Data

Flawed, Inconsistent, Redundant, Erroneous, Ambiguous, Imbalanced, Missing
(...)

Smart Data

Well-Structured, Unbiased, Representative, Balanced, Complete
(...)

Data-Centric AI requires a shift in ML culture

There is a much **higher need to focus on**:

- Systematic, methodical, scalable approaches to improve data
- Data quality guidelines and standards
- Data literacy and training
- Automation (approaches and tools)
- Data provenance, governance, auditing, and monitoring
- Data security and privacy
- Metadata and documentation
- Data Labelling
- High-quality, explainable examples
- Mitigating bias and unfairness
- (...)

Data-Centric AI requires a shift in ML culture

- Data-Centric AI fosters a mindset of **systematically improving data**, which comprises two main components:

Developing Tools

Develop tools to automatically detect and clean data inconsistencies. These systems should operate at scale and be wide in their scope. Ideally they should also be explainable/interpretable.

Leverage Domain Knowledge

Human-in-the-loop approaches are required to validate and improve the approaches. This involves interpreting the information collected and scrutinizing datasets for consistency and real-world value.

Data-Centric AI in machine learning modeling

- **DC-Check**: an actionable checklist-style framework to elicit data-centric considerations at different stages of the ML pipeline: Data, Training, Testing, and Deployment

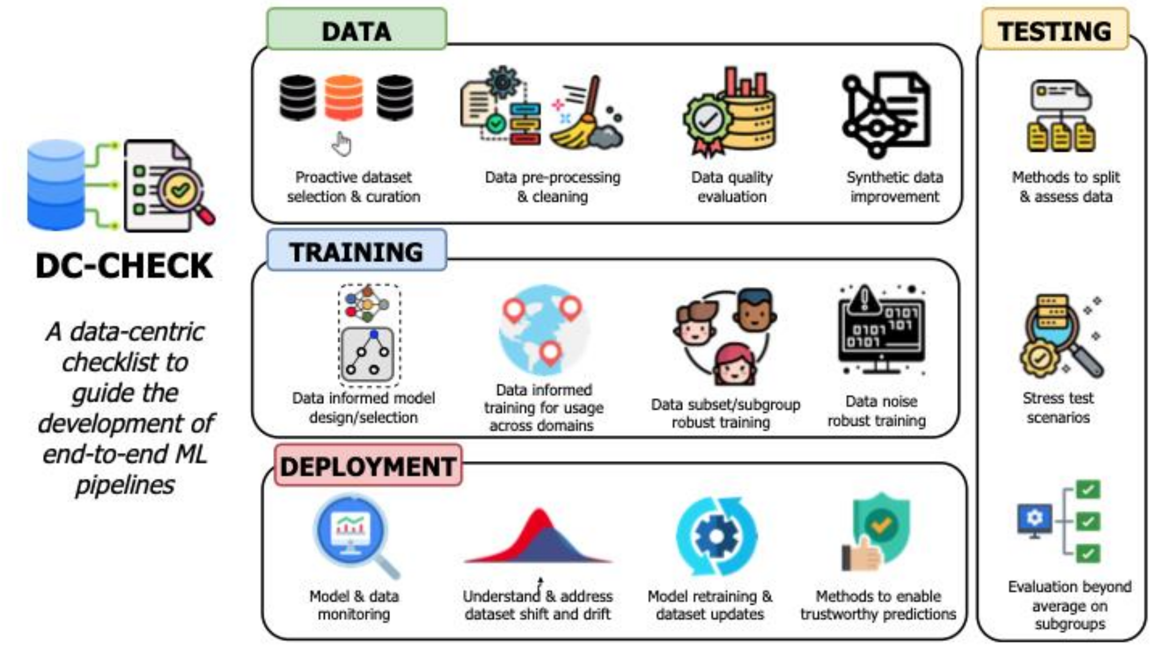
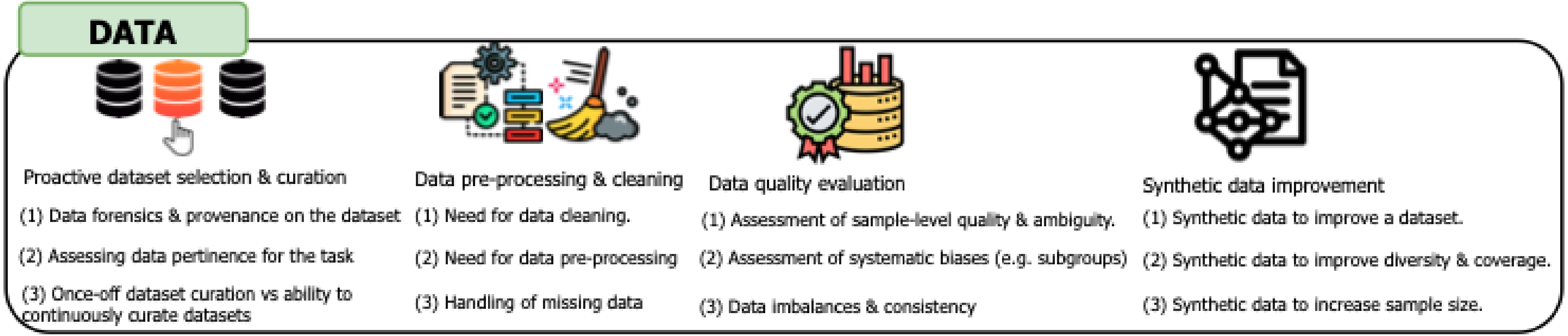


Figure 1: DC-Check: Detailed components with a data-centric lens considered across the pipeline.

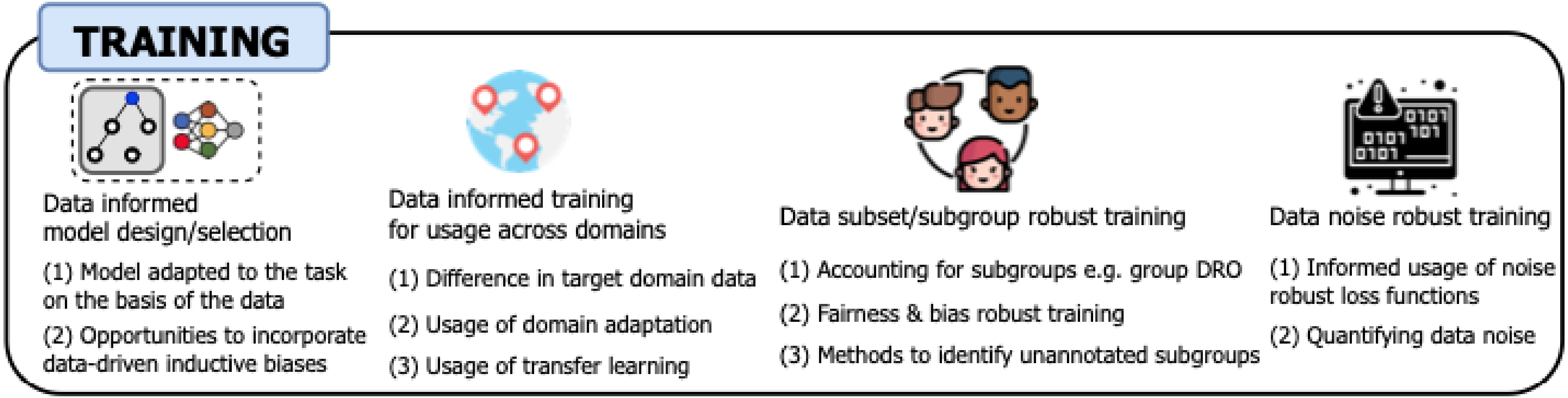
Data-Centric AI on the data component

- How did you select, collect, or curate your datasets?
- What data cleaning and/or preprocessing, if any, has been performed?
- Has data quality been assessed?
- Have you considered synthetic data?



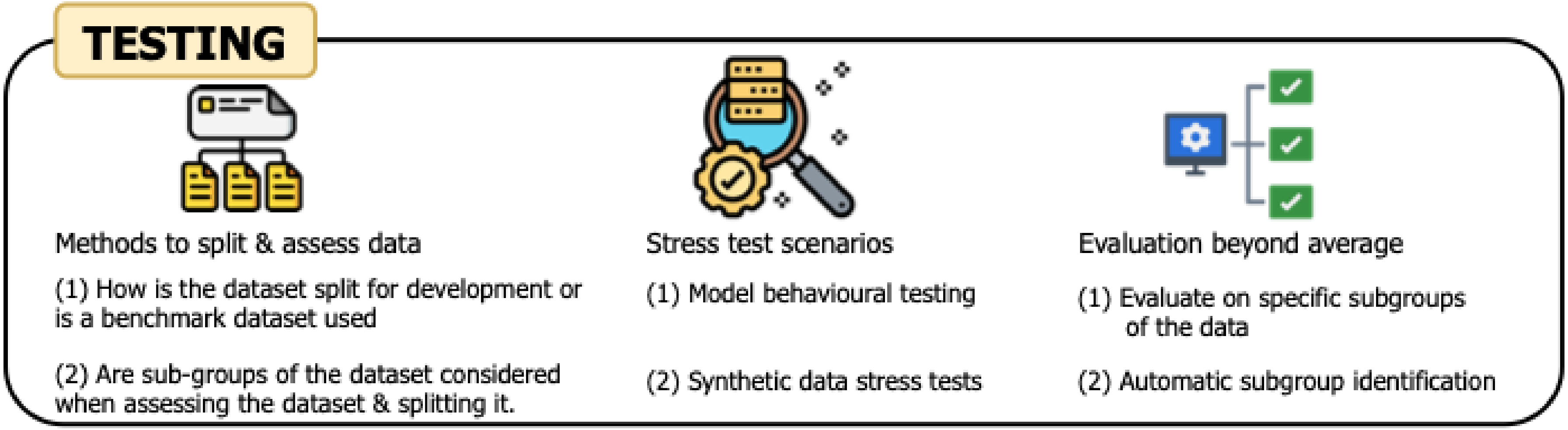
Data-Centric AI on the training component

- Have you conducted a model architecture and hyperparameter search?
- Does the training data match the anticipated use?
- Are there different subsets of groups of interest?
- Is the data noisy, either in features or labels?



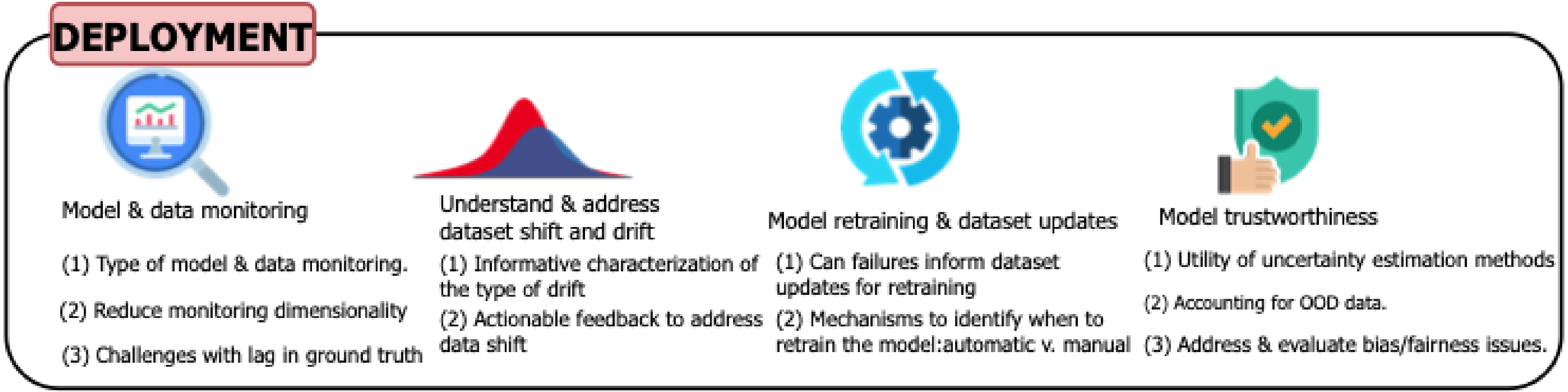
Data-Centric AI on the testing component

- How has the dataset been split for model training and validation?
- How has the model been evaluated (e.g., metrics & stress tests)?



Data-Centric AI on the deployment component

- Are you monitoring your model?
- Do you have mechanisms in place to address data shifts?
- Have you incorportated tools to engender model trust?



Data-Centric AI versus Current Approaches

Current	DC-Check
DATA	
Benchmark/Highly curated datasets	Proactive selection/curation
Fixed datasets	Continuous dataset curation
Manual data forensics	Automated data forensics
Ad hoc data pre-processing	Systematic data cleaning tools (AutoML/RL agents)
Manual dataset improvement	Synthetic data beyond privacy preservation
TRAINING	
Performance based model architecture search	Data informed architecture selection
Heuristic/manual robust learning	Data informed robust learning
Domain adaptation and transfer learning	Improving these methods for limited data
Fairness and group robust methods	Methods to balance fairness/robustness with performance
Learning robust to noisy data	Data-centric informed usage of such methods
DEPLOYMENT	
Limited or Low-dimensional monitoring	New methods for high-dimensional moonitoring
Naive data shift remedies	Actionable and understandble shift remedies
Naive model retraining (batch)	Continual learning (streaming)
Naive dataset updates	Selt-tuning datasets
Overconfident models	Uncertainty estimation & OOD detection
TESTING	
Fixed data evaluation	Synthetic stress test based evaluation
Average/population-level evaluation	Subset/subgroup evalution evaluation

Data-Centric AI landscape in the industry

- **DataPrepOps: MLOps for Data-Centric AI:** *What makes "good" data "good"? Can a dataset be "good" for one application and "bad" for another? How are data characteristics related to the choice of a suitable classifier? How is the downstream task (e.g., classification, regression, clustering) impacted by the quality of the data? How can data quality be validated? What features are relevant for this use case? How much data do we need for this application?*

MLOps

Automate and streamline the end-to-end machine learning lifecycle, from development to production.

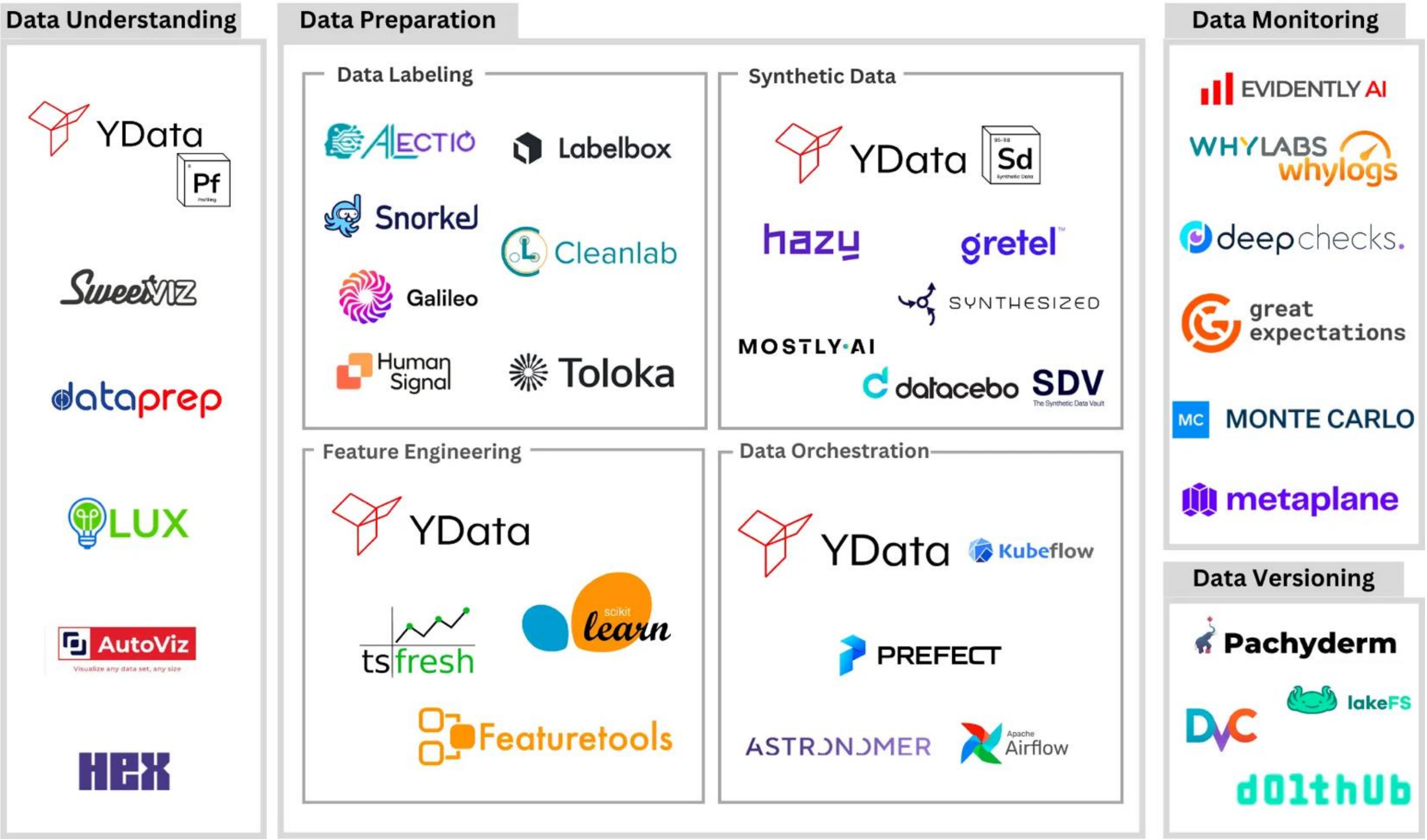
DataOps

Efficient and reliable management of data analytics processes, from organization, storage, versioning, and security. **Improves data at a structural level.**

DataPrepOps

Orchestration and automation of data-centric tasks, including profiling, cleaning, transformation, improvement, filtering, synthetic data, bias mitigation, annotation and (re)labeling. **Improves data value.**

Data-Centric AI landscape in the industry



Exploring Data Quality

Data Quality Issues vs. Data Intrinsic Characteristics

Data Quality

- **Data Quality is a rather broad term** that encompasses several definitions, for which there is no widely established standard. DAMS-NL report comprises about 65 data quality dimensions!

- Common dimensions include: ***Timeliness, Uniqueness, Validity, Consistency, Accuracy, and Completeness***

Quality Aspects	Dimensions	Definition
Reliability	Accuracy	The degree to which data is reliable and describes real-world values
	Uniqueness	Ensures that there are no duplicated records
	Validity	Assures that data conform a specific format and complies with the defined business rules
Availability	Accessibility	The Extent to which data is available and easily accessible
	Security	Ensures that access to information is appropriately restricted
Usability	Ease of Manipulation	The degree to which data could be used and manipulated for its intended use
	Completeness	Assures that there are no missing values, and all the expected attributes have values
	Readability	Refers to the ease of understanding of information [19]
Relevancy	Freshness	Refers to how recent and up-to-date the data is
	Consistency	The extent to which data are coherent and does not contain contradictions
	Credibility	Refers to how much data is credible and can be trusted

Data Quality

Data Quality Team	Functions
Chief Quality Officer	A business executive who oversees the organization's data stewardship, data administration, and data quality programs.
Data Steward	A business person who is accountable for the quality of data in a given subject area.
Subject Matter Expert	A business analyst whose knowledge of the business and systems is critical to understand data, define rules, identify errors, and set thresholds for acceptable levels of data quality.
Data Quality Leader	Oversees a data quality program that involves building awareness, developing assessments, establishing service level agreements, cleaning and monitoring data, and training technical staff.
Data Quality Analyst	Responsible for auditing, monitoring, and measuring data quality on a daily basis, and recommending actions for correcting and preventing errors and defects.

Data Quality Specialist

Siemens Energy

Lisboa, Lisbon, Portugal (On-site)

1 connection works here

Viewed · Promoted · 5 applicants

Data Quality Specialist

Siemens Energy · Lisboa, Lisbon, Portugal (On-site)

Apply Save

What You Bring / Skills, Capabilities

- Technical skills in all areas of Data Quality Management, such as data profiling through low-code industry tools, SQL, informatica, IDQ, SAP and dashboard reporting.
- Profound experience with data profiling tools, definition and execution of technical and business quality rules
- Min. 3 years recent experience in engineering solutions that enable DQ Management.
- Experience of developing data profiling routines and data quality scorecards
- Ability to support the technical rollout of a data quality tool.
- Knowledge of the elements of the ED&AA data governance capabilities and operating model, and how they will come together in the business to deliver value. In particular, data quality, data asset management (metadata etc), enterprise data modelling, data compliance
- Understanding of the Siemens Energy corporate strategy and the key business problems/opportunities of the organisation - either globally or in critical process areas
- Good admin and organizing skills for the creation of a well-run community
- Ability to work well in diverse teams and actively contribute to an inclusive team culture Flexibility and growth mindset, to change and grow as the team matures
- You are passionate about data and analytics, and driving data value across all parts of an organization and ecosystems
- You are C2 level in both spoken and written English, knowledge of German or any other language is a plus.

Data Flaws in machine learning modeling

- Data Quality Issues: Arise due to errors in data acquisition, transmission, collection, storage, manipulation processes. **Issues mostly related to structure and format**.

- Data Intrinsic Characteristics, Data Irregularities, Data Complexity Factors: Result from the intrinsic nature of the domains. **Issues related to the nature of data and problem.**

Data Quality Issues

Erroneous Formats, Inconsistent Data, Duplicate Records, Invalid Values, Incorrect Value Formats (...)

Data Characteristics

Imbalanced Data, Missing Data, Biased Data, Class Overlap, Outliers, Small Disjuncts, Lack of Data, Data Shift (...)

Data Intrinsic Characteristics

- While data quality issues are often a concern of data engineers, **we're interested in assessing data quality from a machine learning perspective**, i.e., analysing what are the data characteristics that impact classifiers and how to mitigate those issues.

  These often comprise:

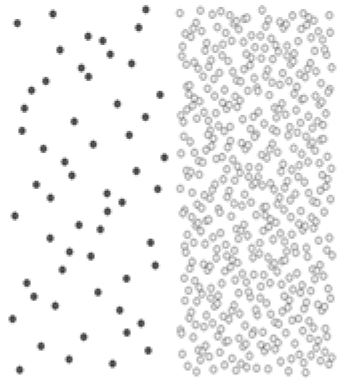
  - Class Imbalance
  - Small Disjuncts
  - Missing Data
  - Class Overlap
  - Noisy Data
  - Lack of Data
  - Dataset Shift
  - ***Data Complexity?***

Data flaws are not restricted to structure and format

Some data characteristics need to be considered

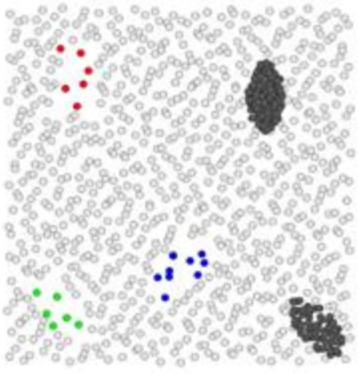
Imbalanced Data

Disproportion between concepts of interest. Worsens with concept rarity.



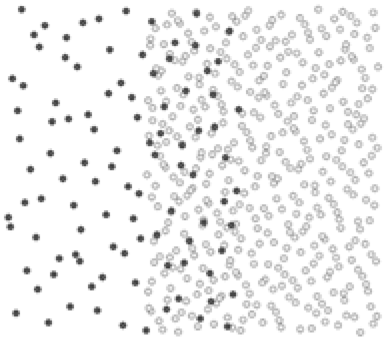
Underrepresented Data

Concept subgroups with the same outcome, despite having different characteristics.



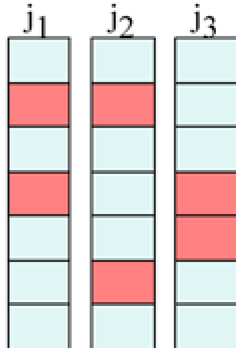
Overlapped Data

Concepts with similar characteristics but distinct outcomes.

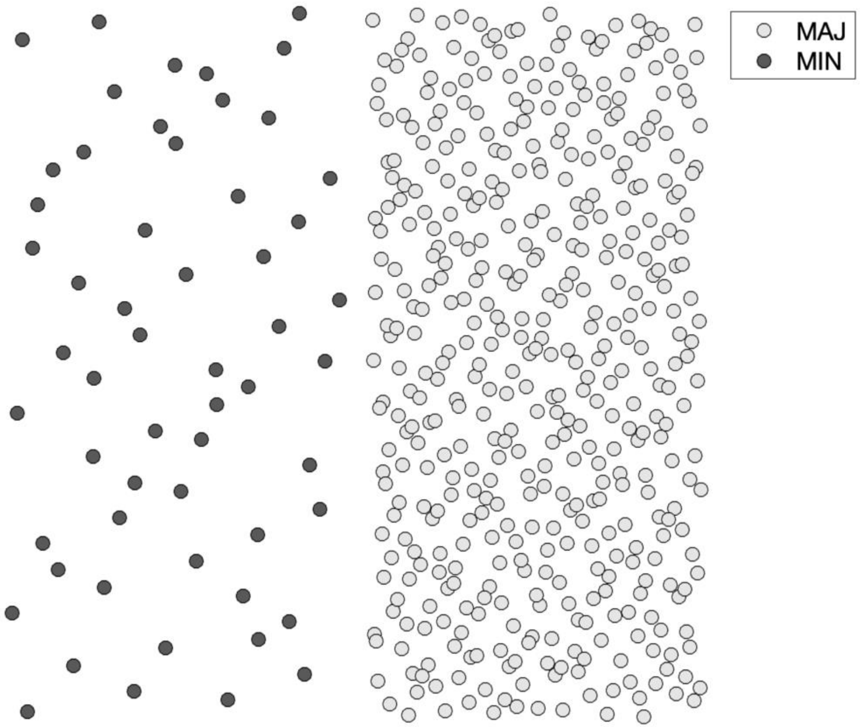


Missing Data

Missing information due to several reasons, e.g., non-disclosure and transmission/collection errors.

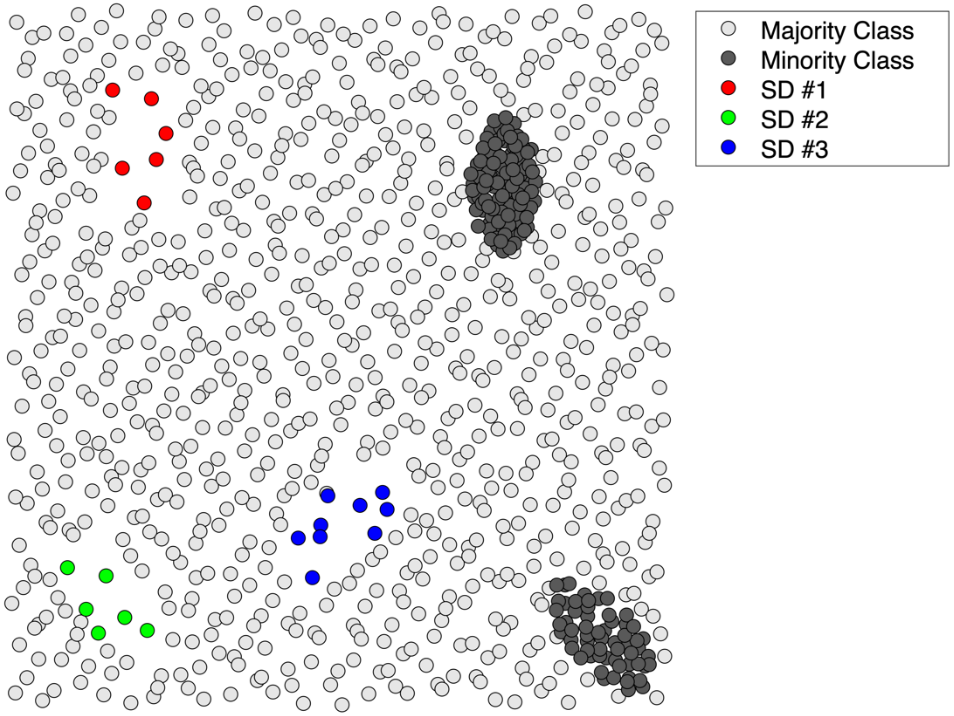


Imbalanced Data: between-class imbalance



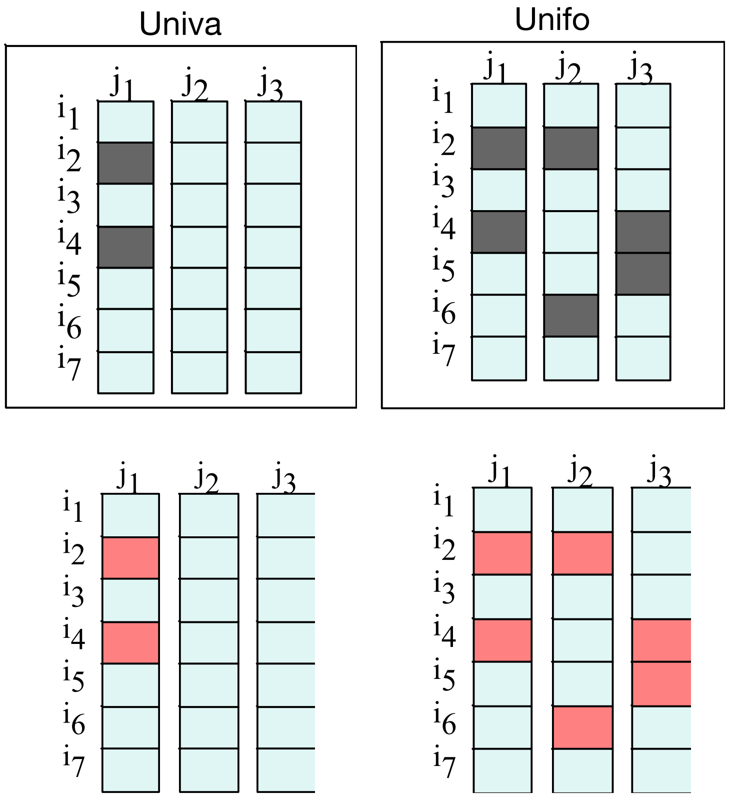
- **Definition:** Disproportion between the number of examples of each class.
- **Problem:** Standard classifiers are traditionally biased towards more well-represented concepts.
- **Example:** Diseased vs. healthy patients in a database.
- **Ongoing Research:** Combination of class imbalance with other difficulty factors.
- **Implications to Society?**

Small Disjuncts: within-class imbalance



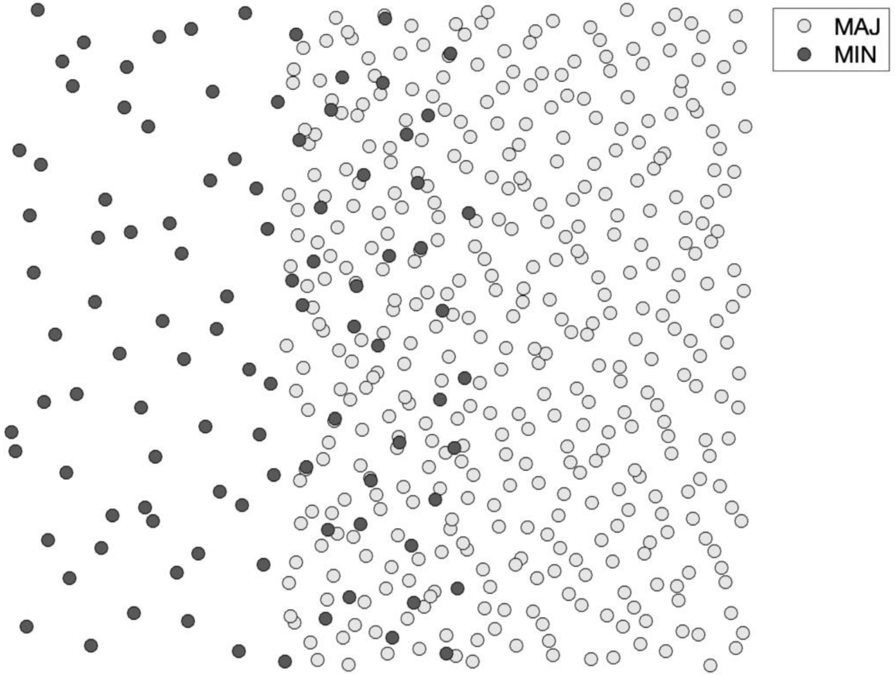
- **Definition:** Underrepresented sub-concepts associated with within-class imbalance (*small disjuncts*).
- **Problem:** Classifiers learn by generating rules for larger disjuncts , overfitting smaller disjuncts.
- **Example:** Clusters of patients with the same outcome but distinct characteristics.
- **Ongoing Research:** Distinguishing between rare cases, core concepts, and noise.
- **Implications to Society?**

Missing Data



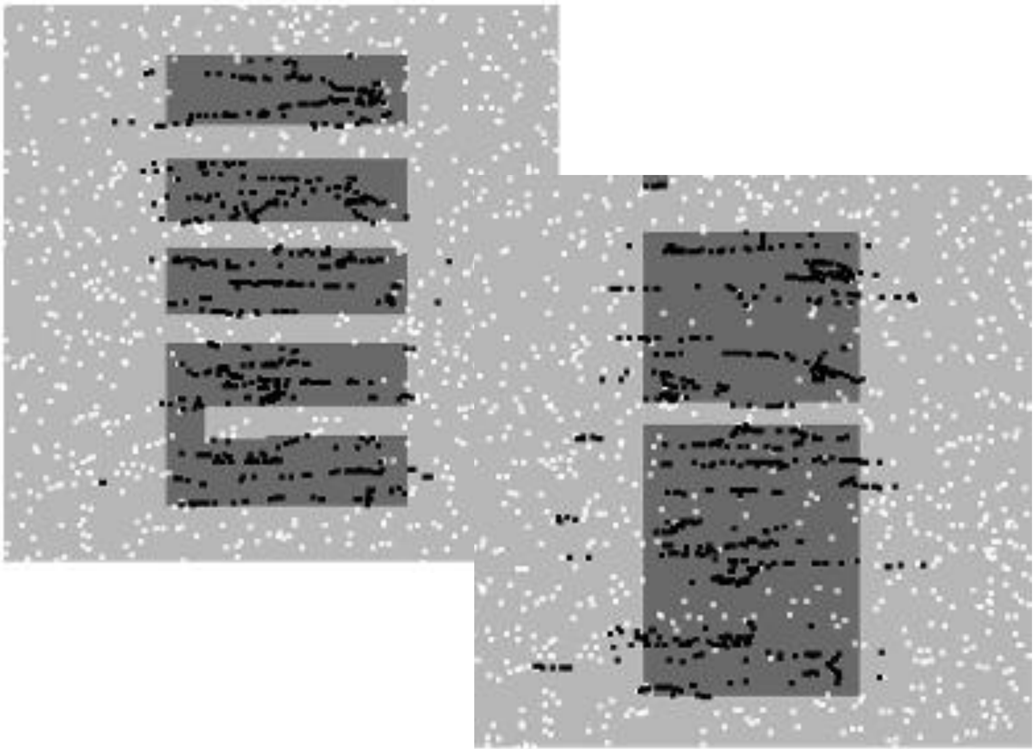
- **Definition:** Absent observations from data, due to 3 possible missing mechanisms (MCAR, MAR, MNAR).
- **Problem:** Standard classifiers expect the input data to be complete.
- **Example:** A patient that misses a survey question (*MCAR*). Values of "weight" are missing for older women (*MAR*). A sensor shuts down for high values of blood pressure (*MNAR*).
- **Ongoing Research:** Diagnosing missing mechanisms, imputation with distributed data, optimizing perfoemance vs. fidelity.
- **Implications to Society?**

Class Overlap



- **Definition:** Instances from different classes coexist in the same region of the input space.
- **Problem:** Finding a suitable decision boundary to discriminate between concepts.
- **Example:** Early stage in disease vs. healthy patients.
- **Ongoing Research:** Mapping overlap as an heterogeneous concept comprising several sources of complexity.
- **Implications to Society?**

Noisy Data

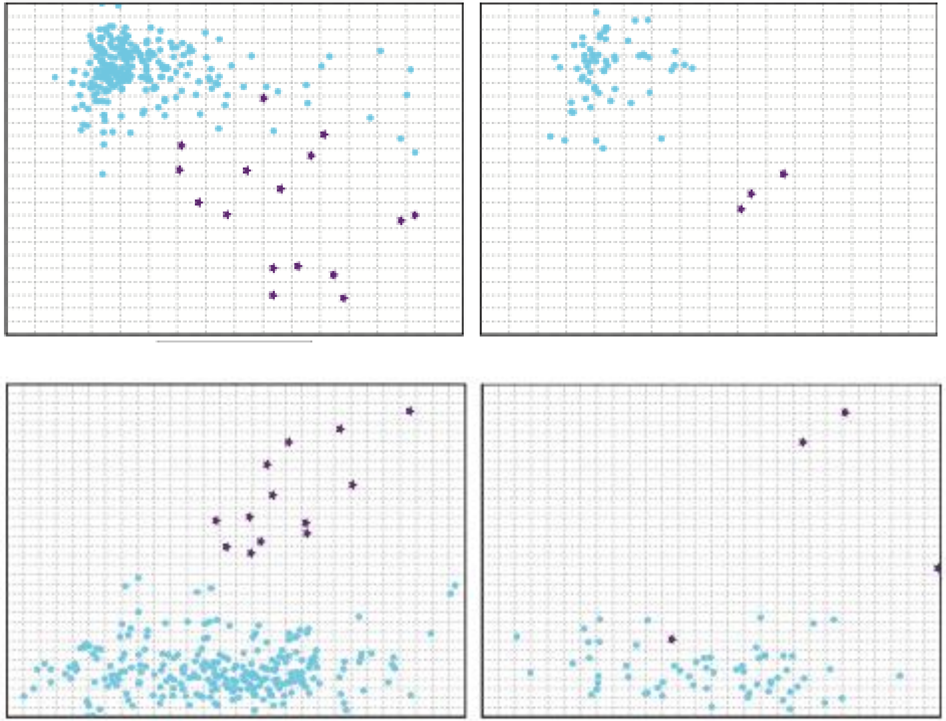


- **Definition**: Feature-level or class-level inconsistencies that affect learning performance (e.g., Gaussian noise, mislabeled examples).
- **Problem:** Standard classifiers expect consistent and correctly labelled instances.
- **Example:** Faulty device outputs erroneous values for blood pressure (feature noise). Human errors in data transcription (class noise).
- **Ongoing Research:** Development of specialized identification and cleaning algorithms. Distinguishing between *noisy* and *valid* instances.
- **Implications to Society?**

Lack of Data or Lack of Density

- **Definition:** Insufficient number of training examples to define the decision boundary.
- **Problem:** Classifiers do not have enough information to generalize for unseen cases.
- **Example:** Patient data collected from a single regional center.
- **Ongoing Research:** Specialized sampling techniques, synthetic data generation.
- **Implications to Society?**

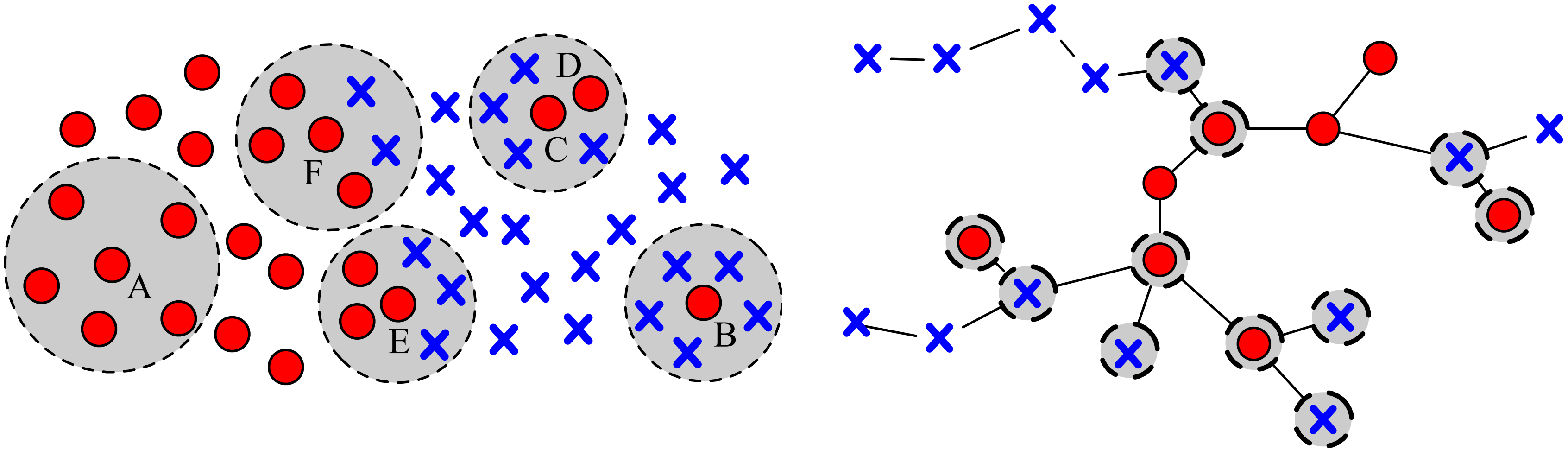
Dataset Shift



- **Definition:** The conditions differ between training and test states. The training data might not be representative of the domain.
- **Problem:** Standard classifiers expect some consistency between training and test settings.
- **Example:** A “no-smoke” policy changes patients’ smoking habits, which leads to a shift in ”# cigaretts/day”.
- **Ongoing Research:** New design of validation strategies, machine learning monitoring, specialized evaluation metrics.
- **Implications to Society?**

Other Data Intrinsic Characteristics

- Studies along this line discuss the estimation of the inherent complexity of the dataset, namely through the quantification of **borderline examples** and **instance hardness** measures. (*We will discuss meta-features and data complexity in the following module).*



Interplay between Data Intrinsic Characteristics

- In real-world domains, data characteristics **arise simultaneously**. However, we still lack a profound **understanding** of their interplay and methods to fully **define** and **quantify** them.
- There are current **open challenges in the intersection** between: *imbalance and overlap*, *imbalance and missing data*, *imbalance and privacy*, *privacy and fairness*, *imbalance and fairness*, …

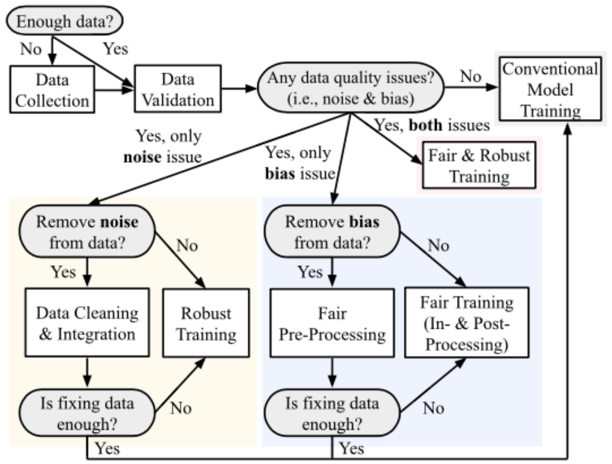
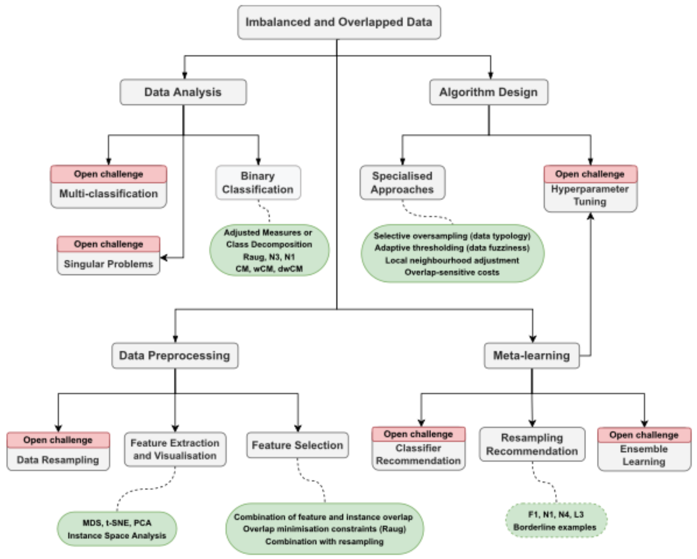
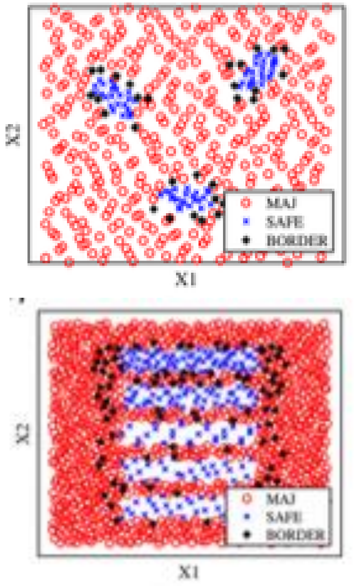


Fig. 2 Decision tree on how data-centric AI techniques connect with each other in one workflow





Other Types of Data

- Tabular
- Time-series
- Image / Video
- Text
- ***Disclaimer: We will be working mostly with tabular, binary-classification problems.***

Data Profiling

Validating and Understanding Data

Data Profiling: Validating and Understanding Data

- Data Profiling involves ***iteratively*** examining the **structure**, **characteristics**, and **quality** of a dataset. This comprehends:
    - **Metadata Analysis:** Structure of data, including types, formats, constraints. Data should match the expected formats.
    - **Statistical Properties:** Basic statistical descriptors of data and feature distribution.
    - **Data Quality Assessment:** Checking for anomalies (e.g., inconsistencies, duplicates) or complicating factors (e.g., missing data, noisy data).
    - **Relationship and Interaction Analysis:** Identifying relationships in data and deriving possible insights to investigate further (e.g., dependencies, contraints).

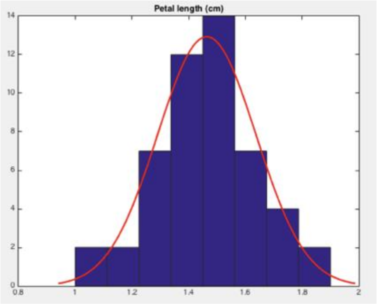
Data Profiling: Real-World Applications

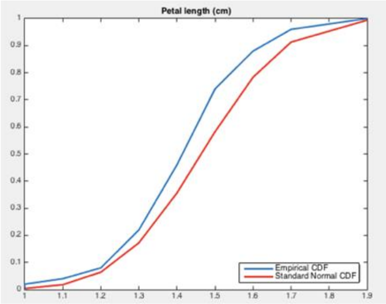
- **Data Profiling is highly relevant in several real-world applications:**
  - **Ensures Data Quality:** Detecting errors, inconsistencies, etc. that may impact the downstream analysis.
  - **Improves Data Understanding:** Which leads to better decision-making.
  - **Supports Data Integration:** Merging multiple datasets and migrating to other systems.
  - **Guides Data Preparation:** Provides important insights for data transformation, cleaning, and enrichment.
  - **Continuous Assessment:** Comparing multiple versions of data, as development occurs.
  - **Handling Sensitive Data:** Or at least hinting at it.
- ***Use cases across healthcare, finance, retail/e-commerce, telecomm, manufacturing, energy…***

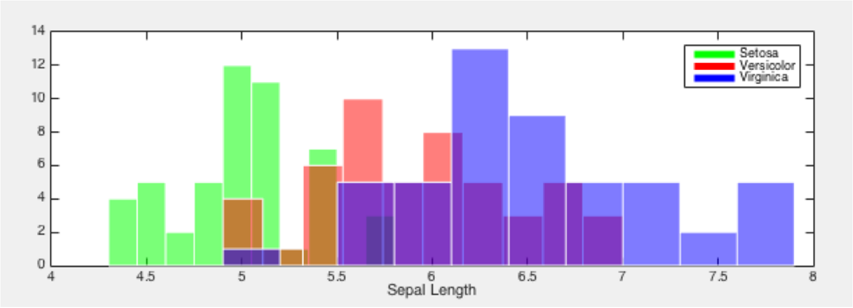
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

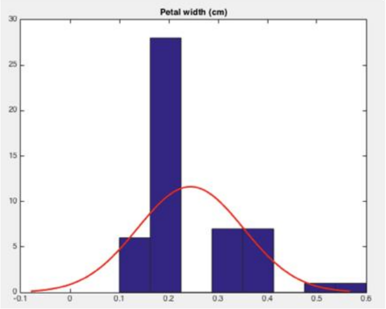
Data Profiling: Data Visualization

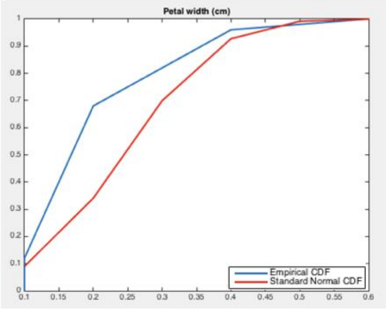
- Visualization goes hand in hand with data profiling, since it is crucial for feature assessment (feature distribution, outliers, symmetry, discriminative power...)

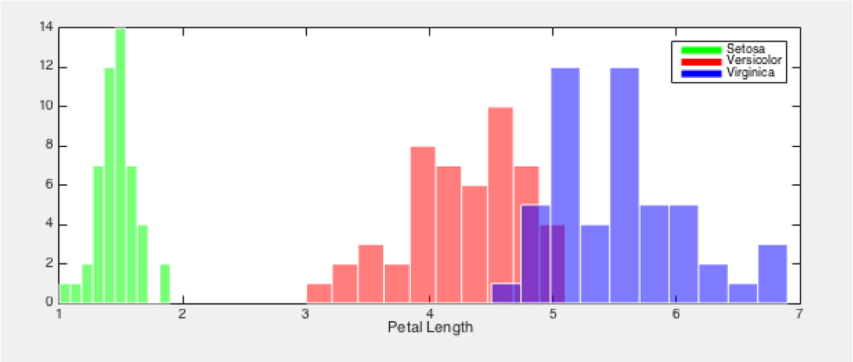








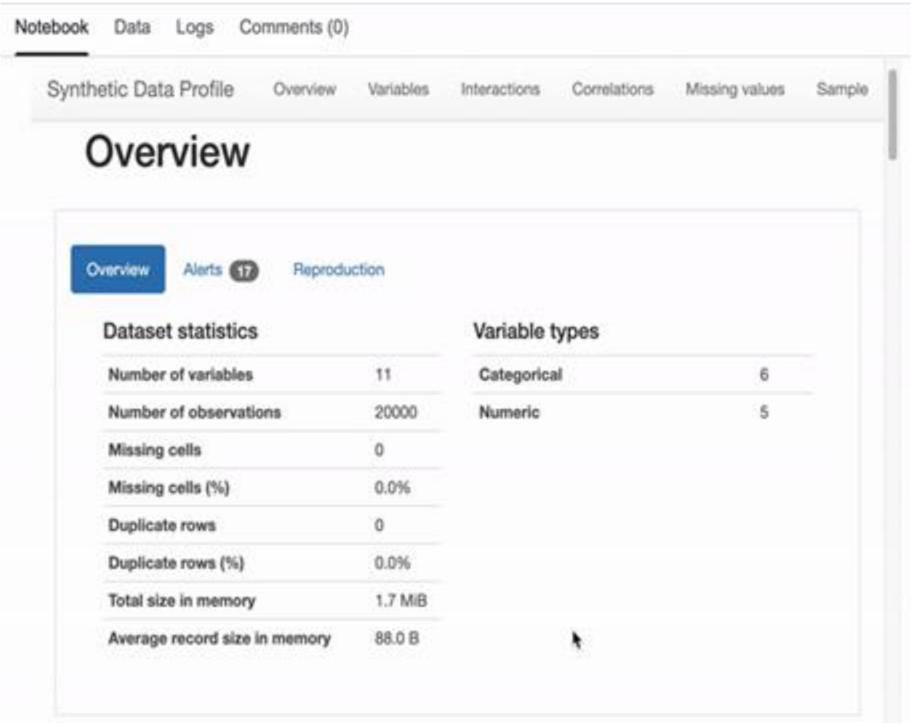




Data Profiling

Practice with Python

Data Profiling OSS: **YData-Profiling** (previously Pandas-Profiling)



- Automatic Generation of Data Quality Alerts
- Supports Tabular and Time-Series Data
- Comparison Report

Unique | Missing | Uniform | Zeros

High cardinality | High correlation | Constant

https://docs.profiling.ydata.ai

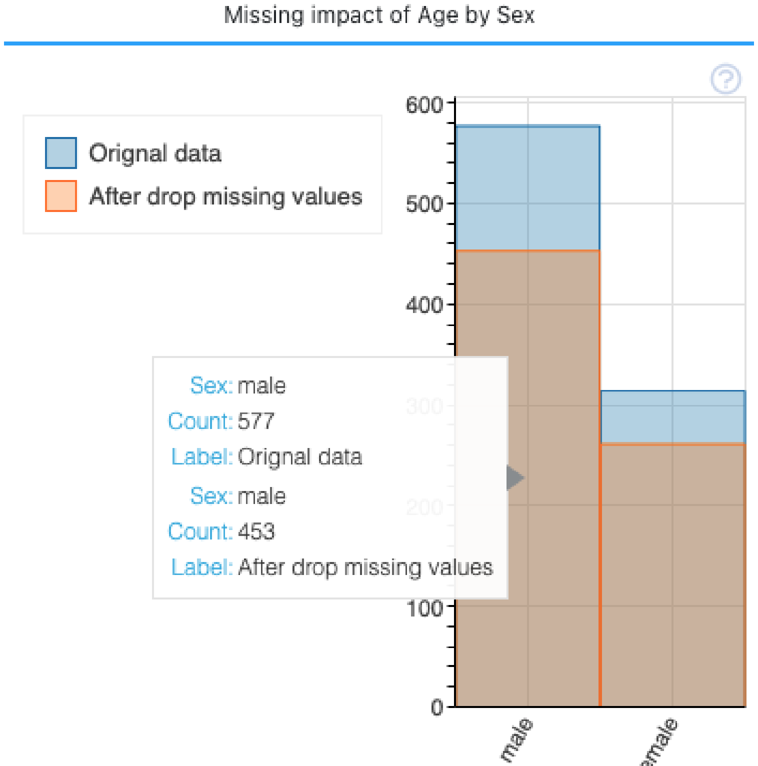
12.2K stars

1.6K forks

`pip install ydata-profiling`

Clemente et al. (2023). ydata-profiling: Accelerating data-centric AI with high-quality data. Neurocomputing

Data Profiling OSS: **Dataprep**



- Exploratory Data Analysis
- Clean and standardize data

```
from dataprep.datasets import load_dataset
from dataprep.eda import create_report
df = load_dataset("titanic")
create_report(df).show_browser()
```

https://docs.dataprep.ai/index.html

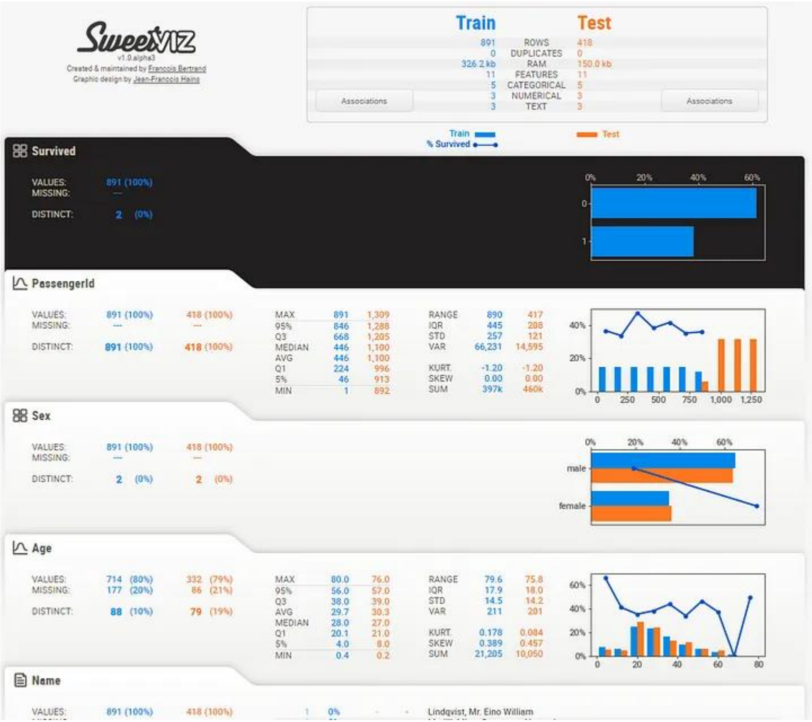
2K stars

201 forks

`pip install dataprep`

[Peng et al. (2023). DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python]()

Data Profiling OSS: **SweetViz**



- In-depth Exploratory Data Analysis (target analysis, comparison, feature analysis, correlation).

```
import sweetviz as sv

my_report = sv.analyze(my_dataframe)
my_report.show_html() # Default arguments will generate to "SWEETVIZ_REPORT.html"
```

https://github.com/fbdesignpro/sweetviz

2.9K stars

269 forks

`pip install sweetviz`

Data Profiling OSS: **AutoViz**

```
##################################################################################
Classifying variables in data set...
Data cleaning improvement suggestions. Complete them before proceeding to ML modeling.
```

	Nuniques	dtype	Nulls	Nullpercent	NuniquePercent	Value counts Min	Data cleaning improvement suggestions
Hallmark	165	object	0	0.000000	100.000000	1	combine rare categories, possible ID column: drop
MCV	130	float64	0	0.000000	78.787879	0	
Ferritin	84	float64	80	48.484848	50.909091	0	fill missing, skewed: cap or drop outliers
Hemoglobin	71	float64	3	1.818182	43.030303	0	fill missing
Total_Bil	62	float64	5	3.030303	37.575758	0	fill missing, skewed: cap or drop outliers
Age	51	int64	0	0.000000	30.909091	0	
Dir_Bil	41	float64	44	26.666667	24.848485	0	fill missing, skewed: cap or drop outliers
PS	5	object	0	0.000000	3.030303	5	
Encephalopathy	3	object	1	0.606061	1.818182	4	fill missing, fix mixed data types
Gender	2	object	0	0.000000	1.212121	32	
Alcohol	2	object	0	0.000000	1.212121	43	
Outcome	2	object	0	0.000000	1.212121	63	
HBeAg	1	object	39	23.636364	0.606061	126	fill missing, invariant values: drop, fix mixed data types
O2	1	int64	0	0.000000	0.606061	0	invariant values: drop

```
    14 Predictors classified...
        3 variables removed since they were ID or low-information variables
        List of variables removed: ['Hallmark', 'HBeAg', 'O2']
Number of All Scatter Plots = 15
```

- Build automatic vizualizations and data cleaning recommendations.

https://github.com/AutoViML/AutoViz

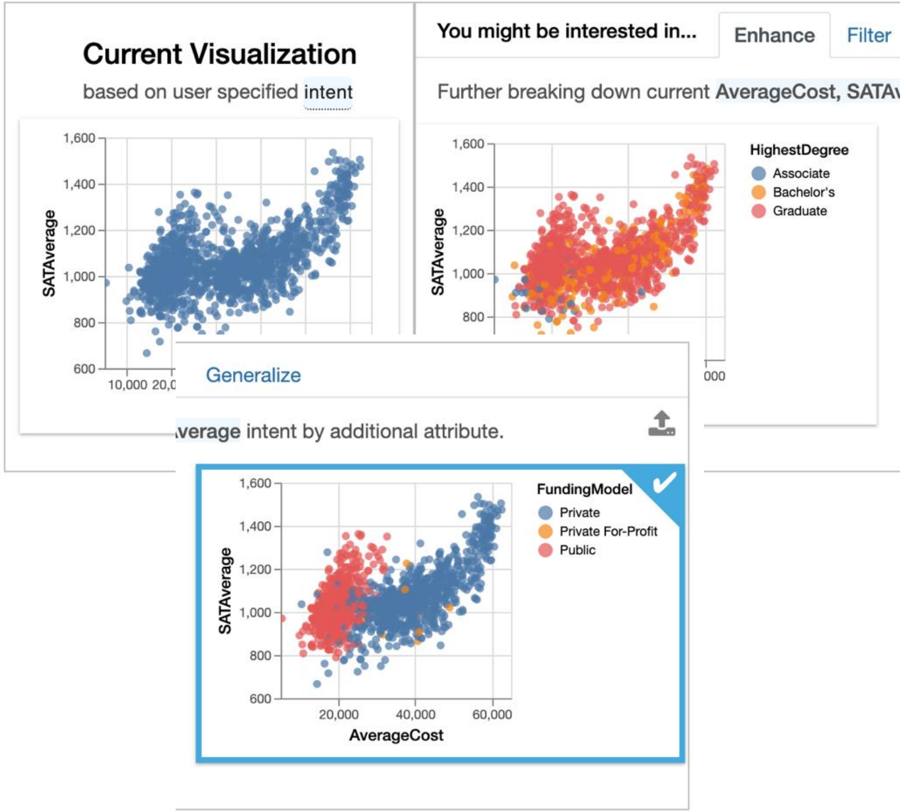
1.7K stars

196 forks

```
pip install autoviz
```

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Data Profiling OSS: **Lux**



- Build high-quality datasets and computer vision models

```
import lux
import pandas as pd
```

Lux can be used without modifying any existing Pandas code. Here, we use Pandas's read_csv command to load in a dataset of colleges and their properties.

```
df = pd.read_csv("https://raw.githubusercontent.com/lux-org/lux-datasets/master/data/college
df
```

https://github.com/lux-org/lux

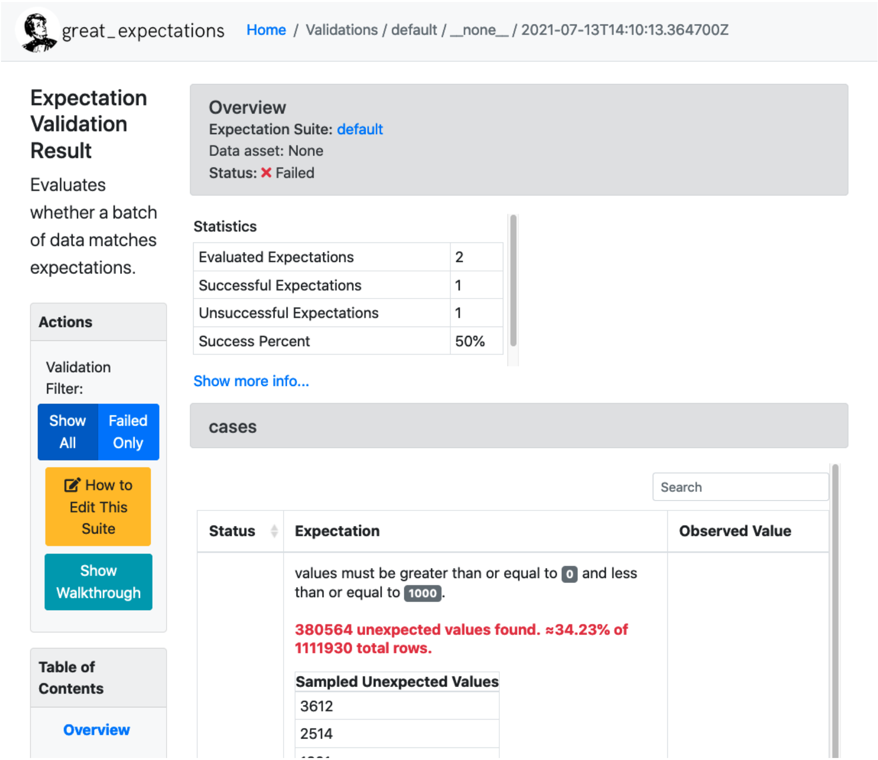
5.1K stars

362 forks

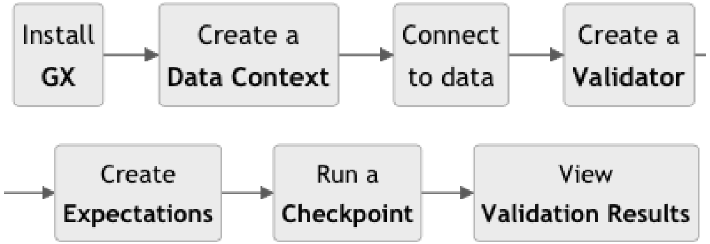
`pip install lux`

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Data Profiling OSS: **GreatExpectations**



- Create and Validate Expectations



https://docs.greatexpectations.io/docs/oss/

9.6K stars

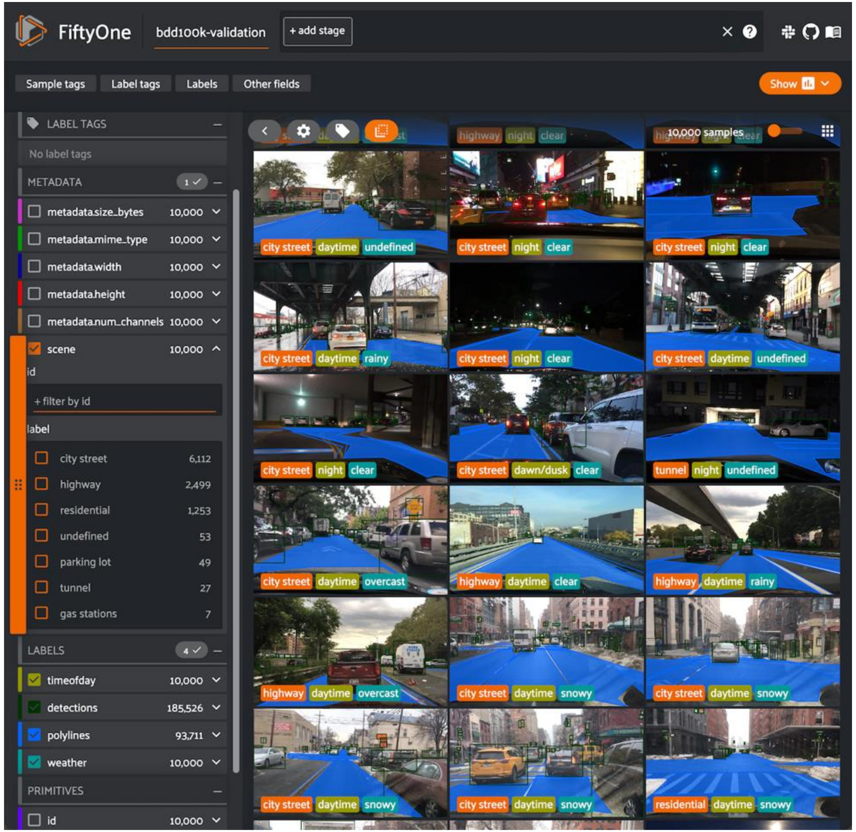
1.5K forks

`pip install great_expectations`

[Use Case Example: Validating Synthetic Data with Great Expectations]()

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Data Profiling OSS: **FiftyOne**



- Build high-quality datasets and computer vision models

```python
import fiftyone as fo
import fiftyone.zoo as foz

dataset = foz.load_zoo_dataset("quickstart")
session = fo.launch_app(dataset)
```

https://docs.voxel51.com

7.8K stars

516 forks

`pip install fiftyone`

References and Further Reading

- Das, S. Datta, B. Chaudhuri, Handling data irregularities in classification: Foundations, trends, and future challenges (2018), Pattern Recognition 81, 674–693.

- A. Fernández, S. García, M. Galar, M., R. Prati, B. Krawczyk, F. Herrera, Data Intrinsic Characteristics (2018), Springer International Publishing. pp. 253–277.

- I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, F. Herrera, Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data (2019), Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, e1289.

- Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M. J., Benítez, J. M., & Herrera, F. (2014). Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(5), 380-409.

- Seedat, N., Imrie, F., & van der Schaar, M. (2022). Dc-check: A data-centric ai checklist to guide the development of reliable machine learning systems. *arXiv preprint arXiv:2211.05764*.

- Jakubik, J., Vössing, M., Kühl, N., Walk, J., & Satzger, G. (2024). Data-centric artificial intelligence. *Business & Information Systems Engineering*, 1-9.

- Whang, S. E., Roh, Y., Song, H., & Lee, J. G. (2023). Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, *32*(4), 791-813.

- Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*.

Tutorial

T01: Data-Centric AI and Data Profiling

Artificial Intelligence and Society

Module 01: Data-Centric AI & Data Profiling

Miriam Seoane Santos
LIAAD, INESC TEC, FCUP, University of Porto
miriam.santos@fc.up.pt