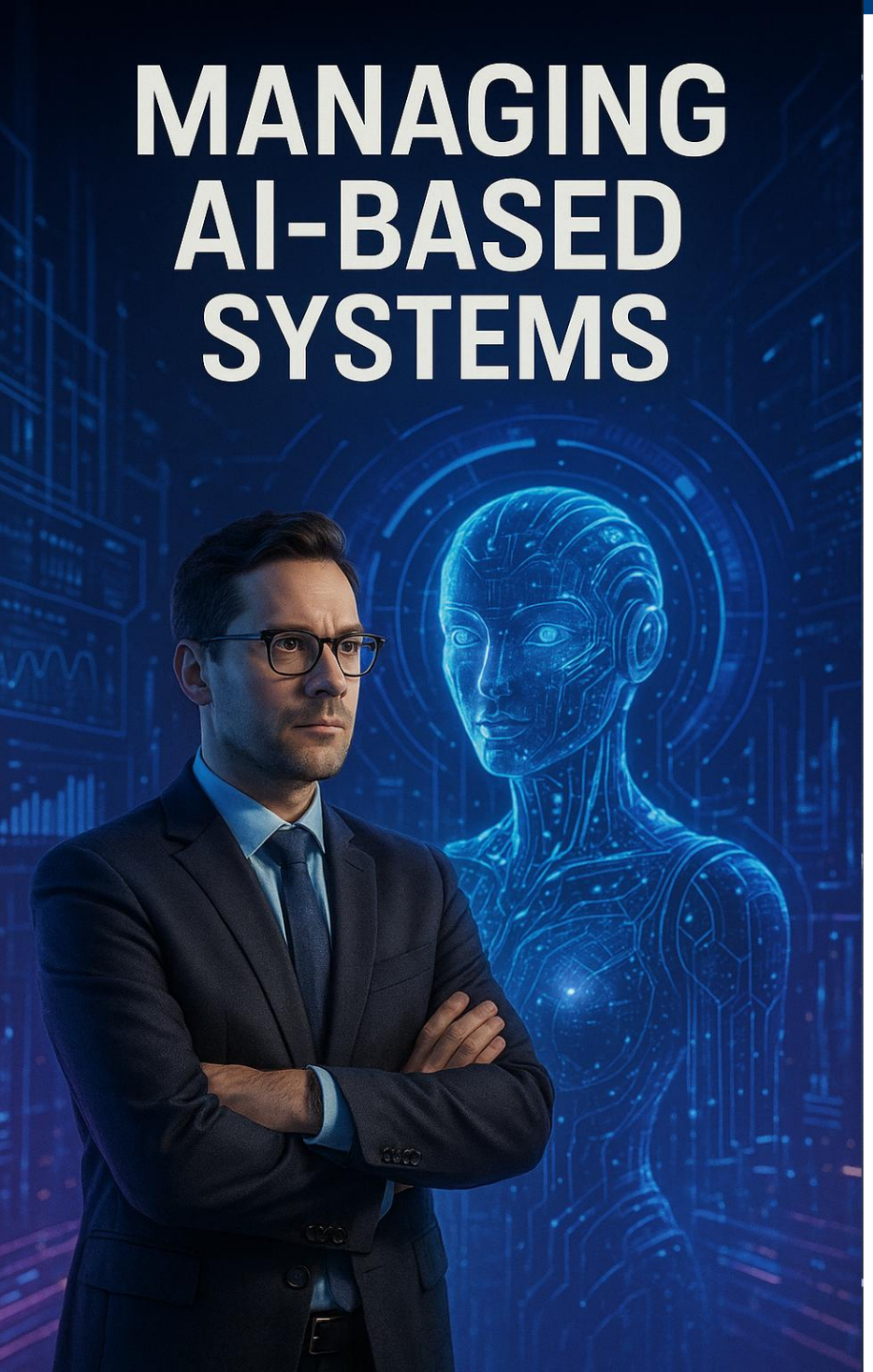



# Session 10: Data Management and Model Transparency

## Managing AI-based Systems

**Prof. Dr. Nils Urbach**

Frankfurt University of Applied Sciences,
Research Lab for Digital Innovation & Transformation

FIM Forschungsinstitut für Informationsmanagement

Fraunhofer-Institut für Angewandte Informationstechnik FIT,
Institutsteil Wirtschaftsinformatik

www.ditlab.org
www.fim-rc.de
www.wirtschaftsinformatik.fraunhofer.de

# Creative Commons Copyright

# Course navigator

2
3 Basic understanding of AI and its business potential
4

## AI Ideation

5 Application potentials of AI technologies

6 Identification, design and evaluation of AI use cases

## AI Strategizing

7 Evaluation an organization's AI readiness

8 Management and governance of AI

## AI Design & Development

9 Architectures of AI applications

10 **Data Management and Model Transparency**

11 Design of human-AI interactions

## AI Operations at Scale

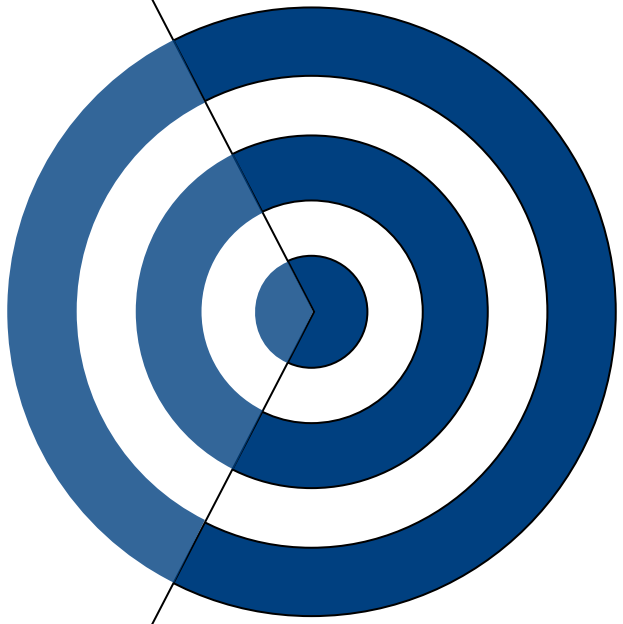
12 Monitoring and KPI-based control

13 Ethical, legal and social implications of AI

Implementation

# Objectives of today's lecture

1. Comprehend the relevance data literacy in the context of AI
2. Comprehend the necessity of AI model lineage
3. Know how to evaluate and manage data quality

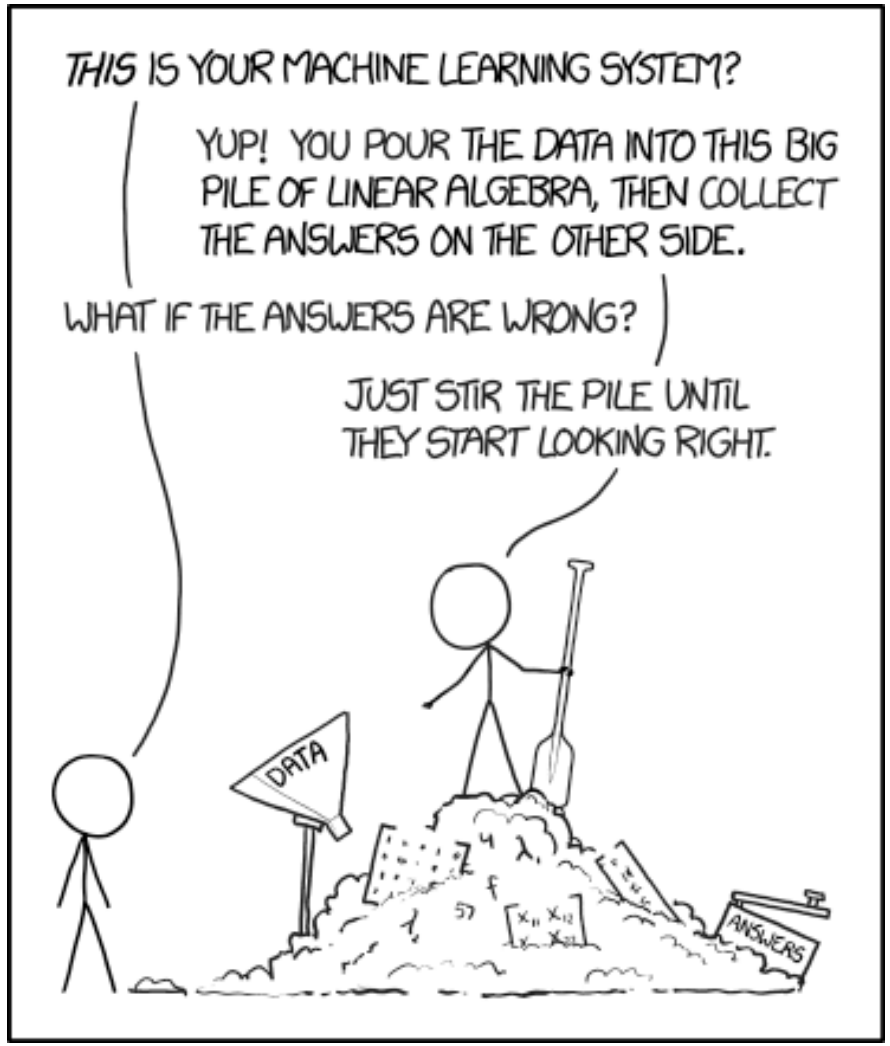


01 Data literacy

02 Model lineage

03 Data quality management

# Why we need to understand what to do with data…




*https://chartmogul.com/blog/data-literacy-deficit/* (2023)

# Data literacy

**Definition data literacy:**
The ability to understand and use data effectively to inform decisions

**Data literacy in the context of machine learning:**
- Involves understanding the intricacies of the data that fuel machine learning models
- Enables individuals to make informed decisions throughout the model development lifecycle
- Leads to more accurate, reliable, and impactful machine learning solutions

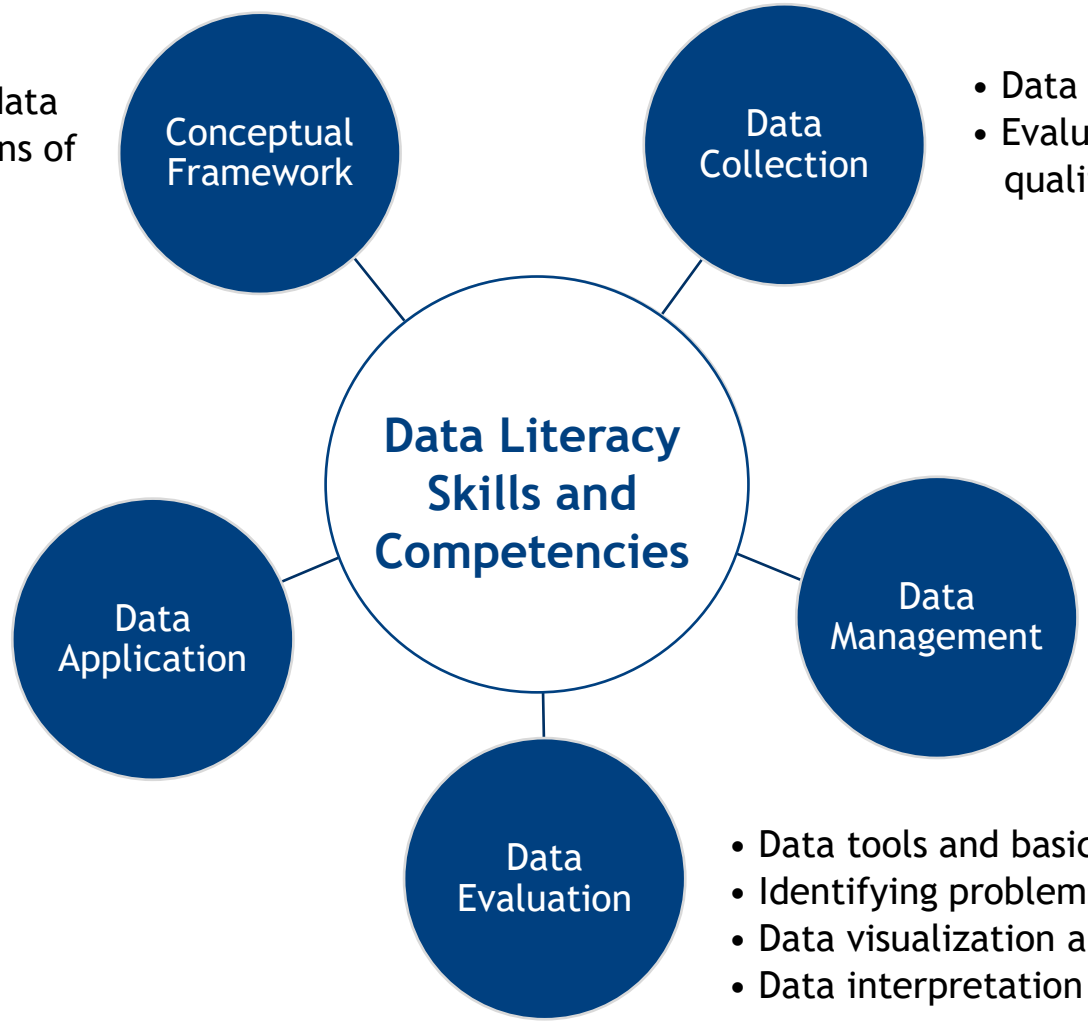
Specific skill set + Knowledge base

Transform data into information and ultimately into actionable knowledge

Monitoring AI systems requires an understanding of the quality, diversity, and potential biases present in the training data

Wolff (2016)

# Relevant skills




- Knowledge and understanding of data
- Use and applications of data

- Data discovery and collection
- Evaluating and ensuring quality of data and sources

- Data manipulation
- Data conversion
- Metadata creation and use
- Data curation, security and re-use
- Data preservation
- Data organization

- Data tools and basic data analysis
- Identifying problems using data
- Data visualization and presenting data
- Data interpretation

- Data sharing
- Data citation
- Data culture
- Data ethics
- Evaluating decisions based on data
- Critical thinking

Ridsdale et. Al. (2015)

# Relevant skills

AI FOR BUSINESS
BUSINESS FOR AI

**Conceptual Framework**

- Comprehending the **significance of data** and its societal impact
- Recognizing how **evolving definitions** of data can affect data collection methods and give rise to new debates and discussions
- **Grasping data comprehensively** involves knowledge of data types, data origins, data sources, data collection techniques, and various applications of data in knowledge and innovation
- Acknowledging that **access to data and data ownership** greatly influence power dynamics, thereby shaping biases and inequalities within society

**Data Collection**

- Involves **discovering and gathering data** for a defined purpose
- Assessing **data quality, relevance, and validity**
- Collection methods vary in complexity, but **accuracy** is always crucial
- Personal data collection involves **privacy** concerns (GDPR addresses data privacy)

Guler, Gulsen (2019); adapted from Ridsdale et. Al. (2015)

# Relevant skills

AI FOR BUSINESS
BUSINESS FOR AI

**Data Management**

- Includes wide range of activities to **properly deal with collected data**
- **Gains significance** in various fields particularly due to the growth of open access to data
- Requires **technical skills** (like data conversion, metadata creation and use)
- Securing data, especially personal and online data, is a critical component of data management (**privacy and online data security**)

**Data Evaluation**

- Involves **understanding data and its value** to be able to develop hypotheses and identify correlations
- Asking critical questions about **data sources, methods, and relevance**
- **Cleaning, analyzing, visualizing and communicating** data

Guler, Gulsen (2019); adapted from Ridsdale et. Al. (2015)

# Relevant skills

Data Application

- Translating data into **actionable practices** (closely related to data driven decision making)
- **Data ethics** (data, algorithms and practices)
- **Critical thinking** is crucial in identifying, defining, analyzing, and self-correcting biases in data, making it a fundamental skill for data literacy

**Data-driven decision making** (DDDM) as practical application of data literacy:

1. Analyzing and evaluating data
2. Setting goals
3. Determining a strategy
4. Implementing and executing the strategy

Guler, Gulsen (2019); adapted from Ridsdale et. Al. (2015)

# Tips for improving data literacy (in organizations)

**Data Literacy Program Recommendations**

1. **Distinguish data from technology:**
   - Emphasize data over complex technical tools
   - Make technology user-friendly to allocate more time for data
2. **Assess employee skills:**
   - Start with a baseline assessment of employee data skills
   - Develop a plan for upskilling based on skill levels
3. **Use common language:**
   - Establish a common data language
   - Avoid jargon and imprecise terms for clear communication
4. **Foster a learning culture:**
   - Encourage curiosity and continuous learning
   - Reward curiosity instead of punishing lack of data literacy

Brown, S. (2021)

# Tips for improving data literacy (in organizations)

**Data Literacy Program Recommendations**

5. **Recognize diverse learning styles:**
   - Customize training and enablement to cater to different learning preferences

6. **Define success metrics:**
   - Develop measurable performance indicators
   - Link data literacy to real projects with tangible results

7. **Engage leadership:**
   - Ensure top executives are actively involved
   - Model the desired data-driven behavior

8. **Data literacy is part of a larger picture:**
   - Progress in data literacy can be achieved within a year
   - Consider data maturity, data-driven leadership, and data-driven decision-making as integral components of a data-driven organization

Brown, S. (2021)

# Handling data management in organizations

- AI systems need to have access to **high-quality, reliable data for accurate insights and predictions**
- Especially **data management is a critical task** in organizations to build a foundation for informed decision-making and data comprehension in the field of data literacy

Important thematic areas to be considered: 5Cs of data management

# 5Cs of Data Management

AI FOR BUSINESS
BUSINESS FOR AI

**Establish security and governance**

- Manage data processing
- Audits and quality control
- Organizational resources

**CONTROL**

**Gather data, put infrastructure in place**

- Transactional databases
- Web mining
- User-generated content big data extraction

**COLLECT**

**Catalog, index and streamline data access**

- Relational databases
- Data warehouses
- Big data storage

**CURATE**

**User data**

- Requirements
- Data processing & visualization
- Data science applications

**CONSUME**

**Data management work**

- Requirements analysis
- Conceptual modeling
- Business intelligence

**CONCEPTUALIZE**

Chua et al.(2022)

01 Data literacy

02 Model lineage

03 Data quality management

# Model lineage in machine learning

**Definition model lineage:**
Documentation and recording of the development history and relationships between different models

**Purpose of model lineage in ML:**
- Facilitate collaboration
- Troubleshooting
- Model quality validation and maintenance

## Typical components

| | |
|---|---|
| Dependencies | Development History |
| Versioning | Metadata |

Kühl et al. (2022); Makinen et al. (2021); Makemeanalyst.com (2022)

# Model lineage in the different stages of ML

AI FOR BUSINESS
BUSINESS FOR AI

### Data preparation

- Clear documentation of the dataset (sources, preprocessing steps)
- How will data be split into training, validation, and test sets for model evaluation?

### Model development

- Code tracking by version control for changes
- Documentation of made choices like model type, hyperparameters and settings

### Training and validation

- Documentation of training details and metrics (accuracy, loss, etc.) of each iteration
- Model monitoring (over- and underfitting)
- Records of every experiment run

Kühl et al. (2022); Makemeanalyst.com (2022); Neptune.ai (2023); Cloud.google.com (2022)

# Model lineage in the different stages of ML

AI FOR BUSINESS
BUSINESS FOR AI

**Model deployment**

- Documentation of model versions (date, environment, etc.)
- Log interfaces, endpoints to access the model

**Data versioning**

- Documentation of impact on model performance

**Model evaluation and monitoring**

- Implementation of frequent performance evaluation and monitoring
- Identify input data shifts (by using tools)

Kühl et al. (2022); Makemeanalyst.com (2022); Neptune.ai (2023); Cloud.google.com (2022)

# Throughout the process

**Communication**

- Internal and external communication over lifecycle decisions for trust and transparency reasons

**Challenges**

- Complexity of larger companies and teams
- Privacy and security
- Changes in data sources/ infrastructure

Kühl et al. (2022); Makemeanalyst.com (2022); Neptune.ai (2023); Cloud.google.com (2022)

01 | Data literacy

02 | Model lineage

03 | Data quality management

# Why data quality can be crucial…


*https://datactivist.coop/opendatadays/1/img/baddata2.png*

(2023)

# Bad data quality

Nearly 60% of organizations don’t measure the annual financial cost of poor-quality data

Costly on-premises data quality tools with an average of $208,000 which prevents more pervasive adoption of tools

Poor data quality causes costs of $15 million as the average annual financial cost in 2017 (Gartner's Data Quality Market Survey)

Gartner.com (2018)

# Recap lecture 2 - Data quality

**Data quality** is one of the most important problems in **data management**, since dirty data often leads to inaccurate data analytics results and incorrect business decisions

The best learner is useless if not supplied with **balanced** and **high-quality** data in sufficient quantity

**Data quality**

- Accuracy
- Completeness
- Purity
- Recency
- Consistency

# Is your data fit for use?

| Data quality factors | Consequences of poor-quality data |
| --- | --- |
| Accuracy | Unfit data comes with incorrect insights, flawed decision-making and compromised business processes |
| Completeness | Lacking or wrong data leads to incomplete, biased picture |
| Purity | Redundant or irrelevant data leads to inefficiencies and skewed analysis results |
| Recency | Once perfect data can be outdated over time and be unreliable |
| Consistency | Uniformity in data formats, units and definition must be ensured for good data integration and analysis |

BUT achieving perfect data quality might be impractical, as the required level depends on the context and conflicting user needs

Tayi et al. (1998)

# Dealing with data quality challenges

AI FOR BUSINESS
BUSINESS FOR AI

## 4 steps to overcome data quality challenges

1. **Measure value -** annual cost of poor-quality data (missed business growth opportunities, increased risks, lower ROI)
2. **Establish the critical data quality roles -** key roles (data stewards or data quality analysts) are shifting from IT to either purely business or an IT-business hybrid combination
3. **Optimize cost of data quality tools** – find a combination of well-established and innovative sourcing methods for tools to meet required performance and needs at optimal costs
4. **Estimate a realistic time frame to deploy data quality tools** – often overestimated deployment time leading to distrust between the business and IT and barriers for data quality programs

Gartner.com (2018)

# Measuring data quality

***How do we know if our data fits the required standards?***

**Common tools for data quality assessment**

Data profiling tools

Statistical analysis tools

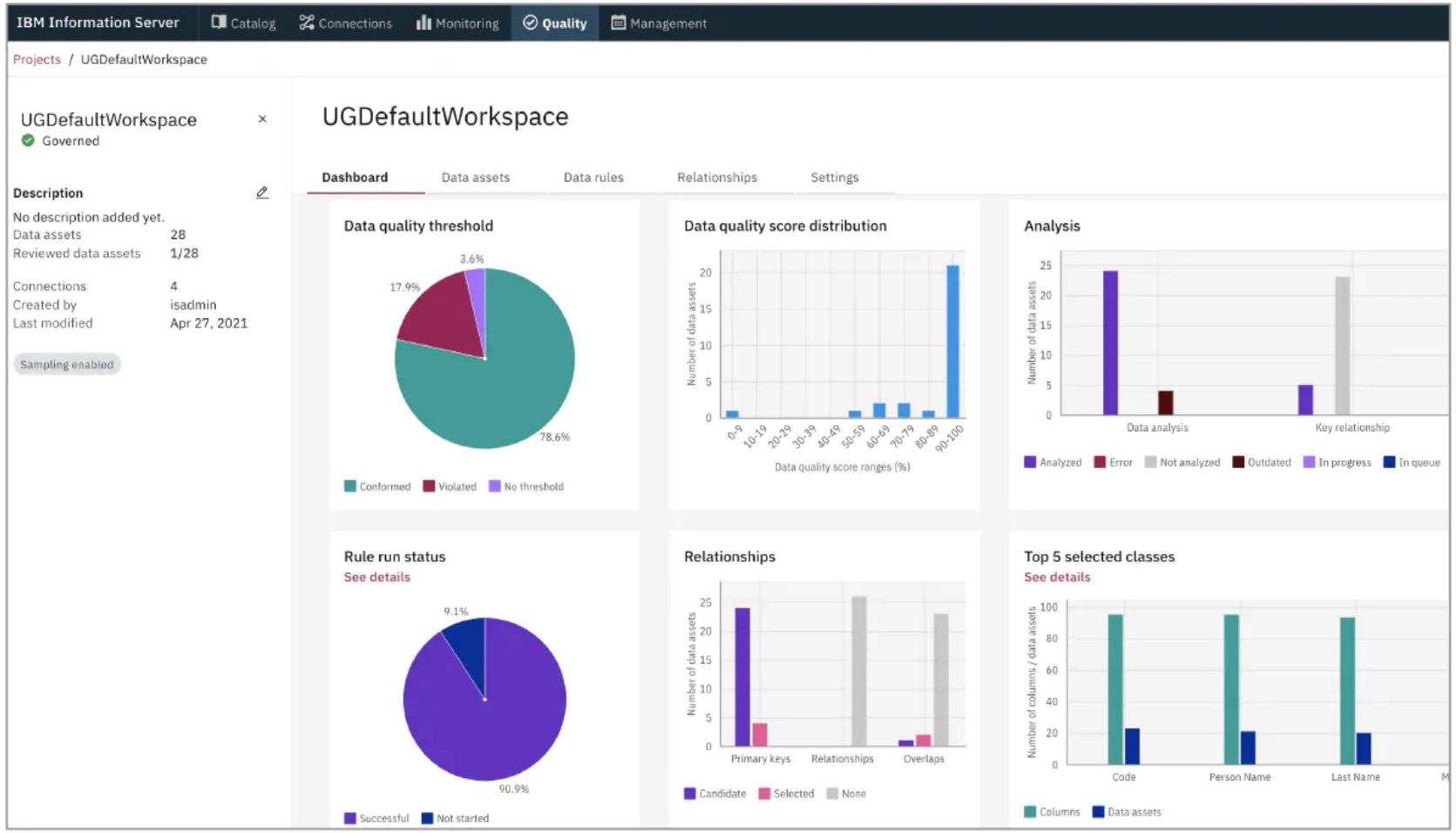
Data quality frameworks (DQAF)

Duplicate matching software

Pipino at al. (2002)

# Example: IBM InfoSphere Information Server for Data Quality

- Cleanse data and monitor data quality on an ongoing basis
- Turn your data into trusted information
- Understand your data and its relationships
- Analyze and monitor data quality continuously
- Cleanse, standardize and match data
- Maintain data lineage




IBM (2023)

# Today's lecture at a glance

- We have gained a sound understanding of data literacy and related skills in the AI context
- We have gained clarity over model lineage in different ML stages
- We know how to manage data quality

# Questions, comments, observations

# Scientific references


- Brown, Sara. How to build data literacy in your company. MIT Sloan. Retrieved July, 2021, 12. Jg., S. 2022.
- Chua et al.(2022): MISQ Research Curation on Data Management.
- D. Kreuzberger, N. Kühl and S. Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," in *IEEE Access*, vol. 11, pp. 31866-31879, 2023, doi: 10.1109/ACCESS.2023.3262138.
- Guler, Gulsen. (2019). Data literacy from theory to reality: How does it look?. 10.13140/RG.2.2.27537.35680.
- Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. Commun. ACM 45, 4 (April 2002), 211–218. https://doi.org/10.1145/505248.506010
- Reddy, S.; Allan, S.; Coghlan, S.; Cooper, P. (2020) A governance model for the application of AI in health care. Journal of the American Medical Informatics Association. 27(3)
- Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., ... & Wuetherick, B. (2015). Strategies and best practices for data literacy education: Knowledge synthesis report. Dalhousie University, Canada. S. 38.
- S. Makinen, H. Skogstrom, E. Laaksonen and T. Mikkonen, "Who needs MLOps: What data scientists seek to accomplish and how can MLOps help?", *Proc. IEEE/ACM 1st Workshop AI Eng. Softw. Eng. AI (WAIN)*, pp. 109-112, May 2021.
- Schneider, J.; Abraham, R.; Meske, C.; vom Brocke, J. (2022) AI Governance for Businesses. Information Systems Management
- Tayi, Giri Kumar, and Donald P. Ballou. "Examining data quality." *Communications of the ACM* 41.2 (1998): 54-57.
- Wirtz, B.; Weyerer, J.; Kehl, I. (2022) Governance of artificial intelligence: A risk and guideline-based integrative Framework. Government Information Quaterly. 39(4)

# Scientific references


- Wolff, Annika, et al. Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*, 2016, 12. Jg., Nr. 3.
- Wolff, Annika; GOOCH, Daniel; KORTUEM, Gerd. Data literacy to support human-centred machine learning. 2016.

# Non-scientific references


- https://www.ibm.com/analytics/common/smartpapers/ai-governance-smartpaper/
- "Canada's New Federal Directive Makes Ethical AI a National Issue." *Digital*, 8 March 2019. Accessed 15 June 2020.
- https://futurium.ec.europa.eu/en/european-ai-alliance/best-practices/implementing-ai-governance-framework-practice#:~:text=The%20AIGA%20AI%20Governance%20Framework,-The%20framework%20developed&text=It%20is%20aimed%20at%20supporting,of%20technology%2C%20and%20professional%20responsibility
- https://www.ibm.com/analytics/common/smartpapers/ai-governance-smartpaper/
- https://makemeanalyst.com/what-is-model-lineage-artifact-tracking/
- https://neptune.ai/blog/tools-for-ml-model-governance-provenance-lineage
- https://cloud.google.com/vertex-ai/docs/pipelines/lineage
- https://www.gartner.com/smarterwithgartner/how-to-stop-data-quality-undermining-your-business
- https://www.gartner.com/smarterwithgartner/how-to-stop-data-quality-undermining-your-business
- https://www.boltic.io/blog/data-profiling-tools
- https://www.ibm.com/products/infosphere-info-server-for-datamgmt
- https://www.ibm.com/information-server
- https://techcrunch.com/2023/05/10/clearview-ai-another-cnil-gspr-fine/
- "SR 11-7: Guidance on Model Risk Management." Board of Governors of the Federal Reserve System Washington, D.C., Division of Banking Supervision and Regulation, 4 April 2011. Accessed 15 June 2020.

# Pictures

- GDPR - Forbes
- Cartoon 1 - Chartmogul
- Cartoon 2 - Datactivist
- Informatica Logo - Hevo Data
- IBM Logo - Product-Information
- Statistical analysis tools - Selecthub
- Syncsort Logo - Businesswire
- Talend Logo - Talend
- Melissa Logo - Printingnews
- Experian Logo - Printingnews