Best Practices in Data Governance, Preparation and Analytics (CA6003)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Data Profiling and Cleaning

Smitha K G
Senior Lecturer
College of Computing and Data Science

Course Learning Outcomes

ILO 1	Apply data governance principles across the AI lifecycle—including data collection, consent management and documentation—by addressing legal and ethical considerations (e.g., GDPR, PDPA, HIPAA) and identifying strategies to mitigate risks related to data privacy, ownership, and regulatory non-compliance in real-world AI projects
ILO 2	Demonstrate the ability to prepare raw data for machine learning by performing profiling, cleaning, validation, transforming, and enrichment, using appropriate tools to transform and preprocess data—including handling missing values, encoding, feature engineering, and standardization.
ILO 3	Analyze datasets to detect and address bias, outliers, and integrity issues by evaluating their impact on model performance and applying mitigation strategies
ILO 4	Design robust EDA by interpreting summary statistics and visualizations, identifying and avoiding analytical fallacies and applying strategies for accurate and meaningful data interpretation.
ILO 5	Design ethical and transparent AI data pipelines by ensuring traceability, explainability, and reproducibility

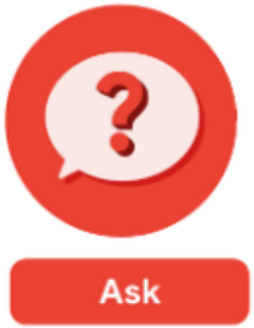
Data Science

Data science is the process of extracting knowledge and insights from complex, often varied data (multi-modal and heterogeneous) through analysis and exploration. Its goal is to enable better decision-making by systematically collecting, preparing, managing, and explaining data and its results.

- **Focus:** Application-specific extraction of insights and knowledge from data
- **Methods:** data cleaning, data profiling, statistics, machine learning, visualization
- **Applications:** Business intelligence, predictive analytics, scientific discovery, machine learning models, data-based decision-making

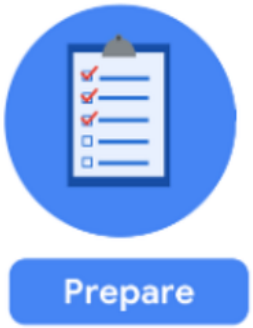
M. Tamer Özsu: Data Science - A Systematic Treatment. Commun. ACM 66 (7), (2023), 106-116. https://doi.org/10.1145/3582491

Data science thinking process



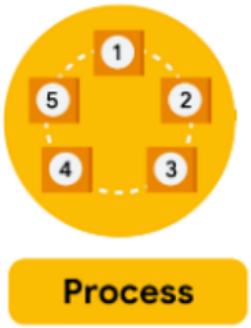
Ask

Ask questions and define the problem.



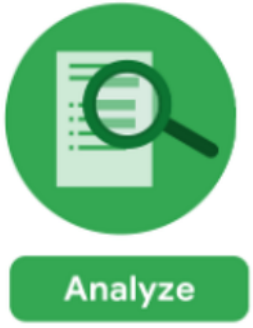
Prepare

Prepare data by collecting and storing the information.



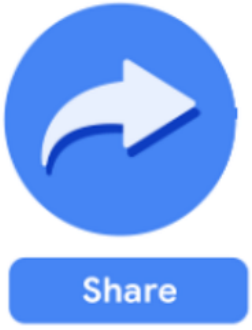
Process

Process data by cleaning and checking the information.



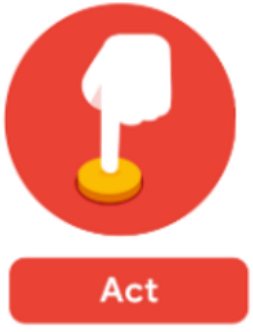
Analyze

Analyze data to find patterns, relationships, and trends.



Share

Share data with your audience.



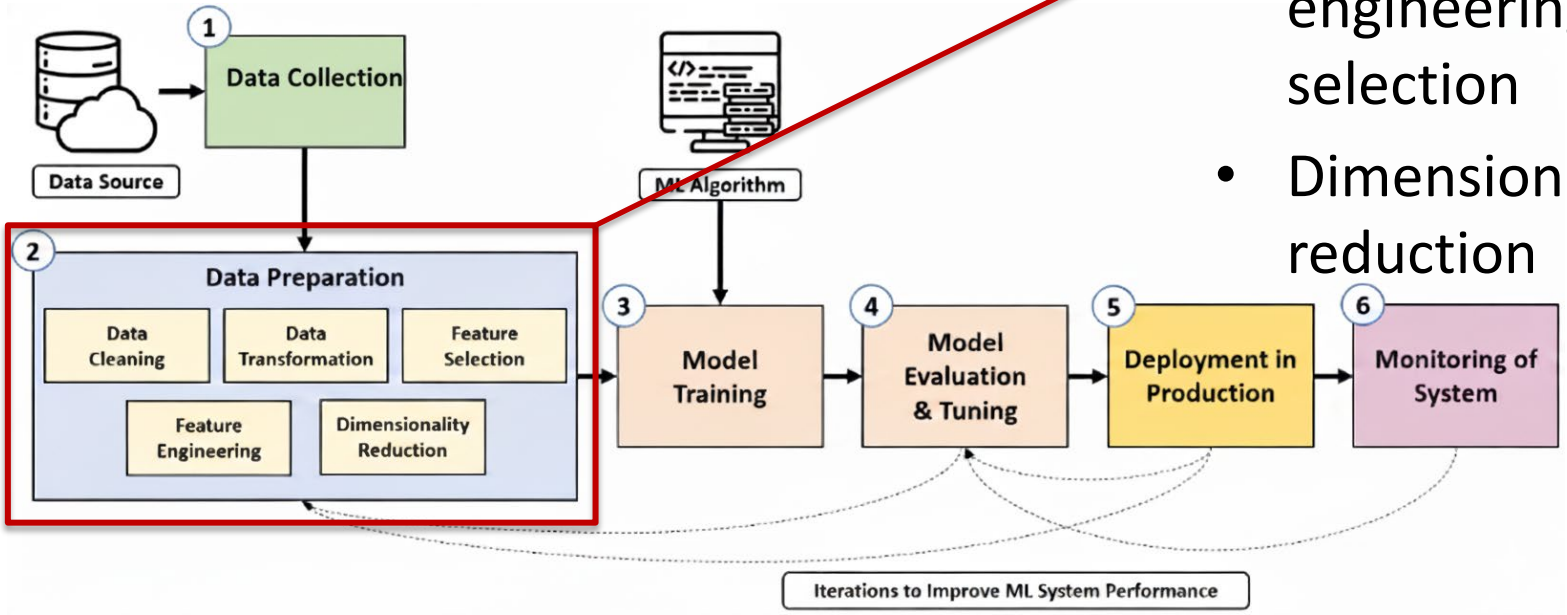
Act

Act on the data and use the analysis results.

Comparing Data Science with Statistics and AI

Aspect	Data Science	Statistics	Artificial Intelligence (AI)
Primary Focus	Extracting insights and patterns from large volumes of structured and unstructured data	Understanding variability, uncertainty, and drawing conclusions from data	Building systems that simulate human intelligence and decision-making
Core Objective	Transform data into actionable knowledge through analytical and computational methods	Develop mathematical models for inference, prediction, and hypothesis testing	Enable machines to learn, reason, and act autonomously
Key Methods / Techniques	Data cleaning, preprocessing, visualization, machine learning, predictive modeling, big data analytics	Descriptive and inferential statistics, regression, probability modeling, sampling, hypothesis testing	Machine learning, deep learning, natural language processing, computer vision, reinforcement learning
Data Handling	Collects, organizes, processes, and interprets data to discover insights	Focuses on data collection, organization, and inference accuracy	Uses data to train models that learn patterns and make intelligent predictions
Outcome	Data-driven insights and decision-making	Statistically valid conclusions and model-based predictions	Intelligent and adaptive systems capable of learning and reasoning

Data science pipeline



- Data profiling
- Data cleaning
- Feature engineering/ selection
- Dimensionality reduction

Assumptions

- Data collection including practical motivation, Sample collection is already been done
- Key points to note
  - Data integrity and completeness (ensures the data you collected is sound and reliable for analysis)
  - Representation and distribution assumption(ensures your sample accurately reflects the population you intend to model)
  - Ethical and Legal compliance (ensures the data can be legally and ethically used for the intended purpose)

- Key assumption: We believe this raw data is the complete, unbiased, and compliant representation of the real world we need to solve the problem

Assumptions- Data collected

#	Data collection Assumption	Details	Critical Risk If Violated
1	Integrity & completeness	We have enough data that is verifiably real, complete, and comes directly from the intended source.	Model Learns Nonsense: If the volume is too small or the data is fake/corrupted, the model will not generalize to the real world.
2	Representativeness & Relevance	The collected data is unbiased, accurately reflecting the diversity of the target population, and its patterns are still relevant today.	Model is Unfair & Outdated: If the data is biased (e.g., missing a demographic), the model will discriminate. If it's too old, the predictions are irrelevant.
3	Ethical & Legal Compliance	All necessary user consent, anonymity, and privacy measures (e.g., GDPR/PDPA) are secured, and we have the legal right to use the data for ML.	Project Termination & Fines: Using data without proper legal or ethical clearance can result in major lawsuits and irreparable reputational damage.

Data Profiling, Cleaning

1. What is Data Preparation and Profiling?
2. Data Cleaning: Handling missing data(deletion, mean/median/mode imputation)
3. Removing duplicates, noise, and irrelevant features

Data Preparation

- What is the need for data preparation
- Data preparation is the process of gathering, combining and structuring and organizing data so it can be used in Analytics and Data visualization
- The components are Data profiling, Cleaning, Feature engineering and Transformation.

Common Data Types

Two Primary Data Types (Structured and Unstructured)

Structured Data

Highly Organized, Easy to Analyze Numeric/Factor, Time Series, Network

- Data that is **highly organized** and resides in a fixed field within a record or file (often tabular). Making a query or filtering is fast
- Data types are **Numeric** (e.g., age, price), **Categorical/Factor** (e.g., gender, country code), **Time Series** (e.g., stock prices).
- Example : **Bank Transactions:** Date, account number, amount, transaction type. 2. **Inventory Logs:** Product ID, quantity, warehouse location. 3. **Sensor Readings:** Timestamp, temperature value, sensor ID.
- **Tools: Pandas** (Python), **Spreadsheets** (Excel/Google Sheets), **SQL** (PostgreSQL, MySQL, Oracle).

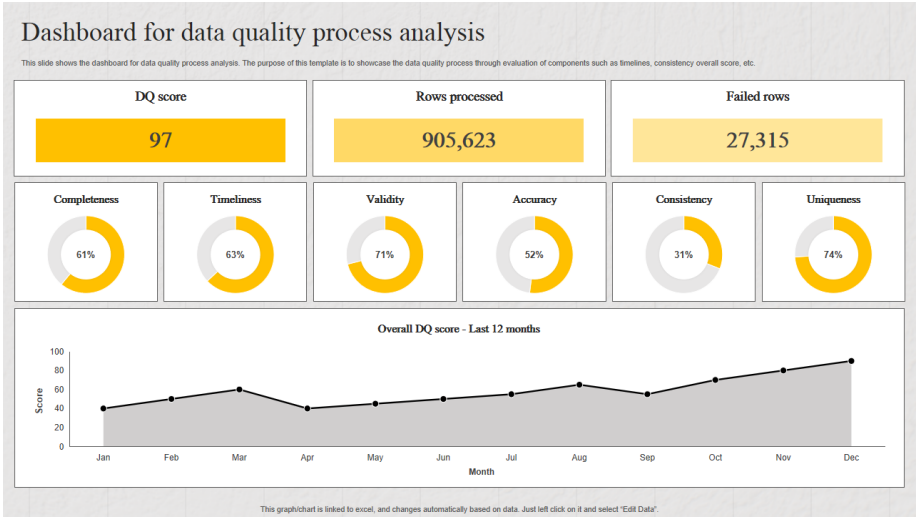
Common Data Types

Unstructured Data

Highly Unorganized and Contextual Text, Image, Voice, Videos

- Data that lacks a pre-defined format or schema. It is **highly contextual** and much harder for machines to interpret directly.
- Requires advanced AI/ML models (Deep Learning, NLP, Computer Vision) to find structure within the raw data.
- Data types are **Text** (emails, tweets, articles), **Image**, **Audio** (voice recordings, music), **Video** (surveillance, films).
- Examples: 1. **Customer Feedback:** Text of an email complaint or support ticket. 2. **Medical Imaging:** X-ray or MRI scans. 3. **Video Content:** Frames and audio from security cameras.
- Tools**: Python** (TensorFlow, PyTorch), **Spark**, and **OpenCV** (Image/Video).

Data profiling



- Data profiling is like doing a health check on your data. **Data profiling** is the process of examining, analyzing, and summarizing data from an existing source (like a database or dataset) to understand its structure, content and interrelationships.
  - Detect data quality issues (e.g., missing, invalid, or duplicate data)
  - Understand data patterns and distributions
  - Prepare for data cleaning and transformation
  - Ensure data integrity before analysis

Data Profiling

Step / Discovery Type	Goal	Process	Typical Statistical / Analytical Tools	Typical Outputs
Structure Discovery	Verify schema, data types, formats	Validate structure, datatypes, patterns	Pattern analysis, data type inference, format checks, regex validation	Schema summary, type mismatches
Content Discovery	Assess data quality and value distributions	Examine actual values (Identifies missing values, anomalies and inconsistencies in data)	Descriptive statistics, frequency analysis, outlier detection, missing value analysis	Missing value $\%$, mean, std. dev., outliers
Relationship Discovery	Identify dependencies and correlations between attributes	Identify column/table relationships (Analyses connection between different data elements and tables)	Correlation coefficients, chi-square tests, covariance, association rule mining	Correlation matrix, key integrity

Very essential for ensuring data readiness for analysis and informed decision making

Types of Data Profiling

- **Single Column Profiling:** Examines individual columns for cardinalities, patterns and data types and Value distributions
  - Cardinality:-Number of rows, Minimum, Maximum, Null values, Distinct values
  - Patterns and Data types: -Generic data type, size , decimals, Histogram of patterns
  - Value Distributions:-Frequency histograms , quartiles
- **Cross/ multi-Column Profiling:** Looks at relationships between columns.
  - Example: Correlations, Clusters, Outliers and Summaries

Data cleaning

- Data cleaning (also called data scrubbing) is the process of detecting and correcting errors, inconsistencies, and inaccuracies in a dataset to ensure that it is accurate, consistent, and ready for analysis or machine learning. It usually follows data profiling and involves handling issues like:
  - Missing or incomplete values (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data, e.g., *Occupation*=“ ” (missing data))
  - Duplicated rows
  - Outliers or invalid values
  - Inconsistent data formats (Noisy Data)(containing discrepancies in codes or names, e.g., *Age*=“42”, *Birthday*=“03/07/2010”, Was rating “1, 2, 3”, now rating “A, B, C”
  - Typographical or case errors, e.g., *Salary*=“−10” (an error)

Handling Missing values

Type of Missingness	Meaning	Example	Handling Suggestion
MCAR — Missing Completely at Random	Missing has no relationship with other data	Randomly unrecorded age	Safe to impute (mean/median)
MAR — Missing at Random	Related to other observed features	Missing income depends on education level	Impute based on correlated columns
MNAR — Missing Not at Random	Missingness depends on the value itself	People with low income don't report income	Avoid simple imputation; model separately or use flag

- ✔ Check **percentage of missing values per column**
- ✔ Check if missingness follows a **pattern** (e.g., all missing in a group)
- ✔ Never fill before **splitting into train/test** — prevents *data leakage*
- ✔ Create a **missing flag column** if missingness itself carries information
- ✔ Choose the right method based on **data type and importance**

Handling Missing values- how to choose right method?

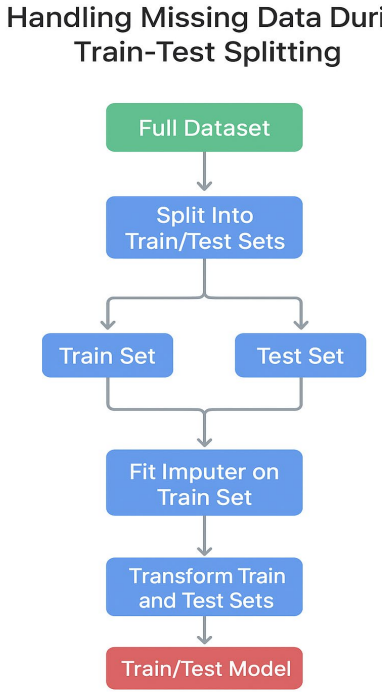
Type	Method	Description	Example	Impact
Deletion	Drop missing rows or columns	If only a few missing	Drop 2 rows missing Age	Simple, but loses data
Constant Value Imputation	Replace with fixed value	Useful for categorical	Replace missing country with "Unknown"	Keeps data but may bias counts
Mean / Median / Mode Imputation	Replace with central tendency	For numerical data	Replace missing Age with median	Maintains scale but reduces variance
Forward/Backward Fill	Copy neighboring values	For time series data	Fill missing temperature with previous value	Preserves continuity, may hide jumps
KNN Imputation	Replace with average of nearest neighbors	Based on similar rows	Use KNNImputer from sklearn	More accurate, computationally heavy
Multiple Imputation	Create multiple imputed datasets, average results	Statistical method for uncertainty	MICE (Multiple Imputation by Chained Equations)	Best for robust analysis, complex

Handling Missing values

Method	Description	Best For	Pros	Cons
Deletion	Removing rows/columns with missing values	When missingness <5%	Simple	Can lose important data
Constant Value Imputation	Replace with fixed value (0, "Unknown")	Categorical features	Easy, fast	Adds bias
Mean/Median/Mode Imputation	Statistical replacement	Numerical or categorical	Simple, baseline	Ignores relationships
Forward/Backward Fill	Fill with previous or next value	Time-series	Good for trends	Not suitable for random missingness
KNN Imputation	Uses nearest neighbors	Mixed datasets	Captures relationships	Slow for large data
Binning-Based Imputation	Assign data to bins and impute	Numerical features	Robust, simple	Information loss
Clustering-Based Imputation	Uses cluster statistics to fill gaps	Natural group data	Subgroup-aware	Needs tuning

Golden rule in machine learning

- Never compute imputation values (mean, median, etc.) using the full dataset before splitting!
  - Because if you use all data to fill missing values, you are allowing information from the test set to influence the training process — this is data leakage- you may get optimistic results
  - Split the data into training and test sets
  - Fit the imputer only on the training set
  - Transform both training and test sets using that fitted imputer



Benefits of data profiling and cleaning

- Fix errors quickly:- helps to catch errors before processing
- Produce top quality data- cleaning and reformatting datasets ensures that all data used in analysis will be high quality
- Makes better decisions: can be processed and analyzed more quickly ad efficiently leading to timely , efficient and high quality decisions