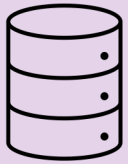
Pre-modeling Explainability

Explainable and Trustworthy AI

Eliana Pastor

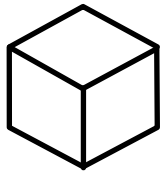
Stages of Explainability

Explainability involves the entire AI development pipeline



Pre-modelling explainability

Before building the model
- Data exploration
- Data selection
- Feature engineering



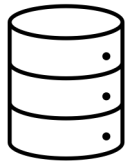
Explainable modeling

Build inherently interpretable models
- Manage the accuracy and interpretability trade-off

Post-modelling explainability

After model development
- Explaining predictions and behavior of trained models

Stages of Explainability – Pre-modelling explainability



Before the actual modeling process

The goal is to

- Gain more useful **insights from data** and use them for model development
- Preprocess them for the following model development preserving their understandability

It includes:

- **Exploratory data analysis**: extract a summary of the main characteristics of the data
- **Data description and summarization**: standardize documentation, ensure proper communication between data providers and users of datasets
- **Interpretable Feature engineering**: selection and preprocessing of features preserving their interpretability

Exploratory data analysis (EDA)

Statistical techniques and **visualizations** to gain more insights into a dataset

- Extract a **summary of the data**
- **Visualizing the dataset**

- **Compute and analyze the statistical** properties of the data
  - mean, standard deviation, percentage of missing sample, feature dimensionality, presence of outliers

- Knowing the data enable
  - to then better understand the model will be trained on such data
  - exposing biases that might exist within the data

Exploratory data analysis (EDA)

- Use common libraries
  - E.g., Numpy, Pandas, Sklearn

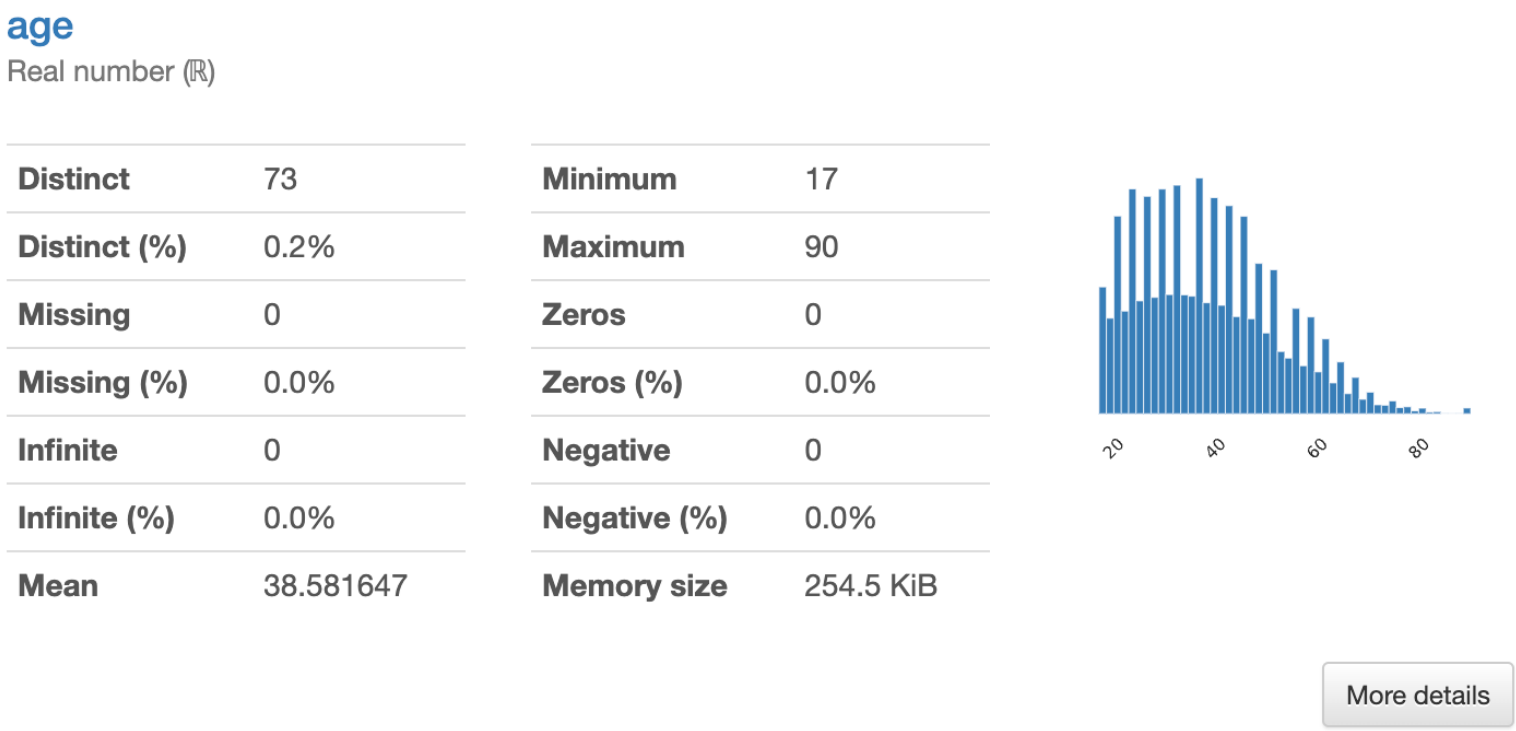
- Ad-hoc libraries and Tools
  - e.g., ydata-profiling ,FACETS, Tableu, KNIME

Exploratory data analysis – Example of tools

ydata-profiling

Analysis of a DataFrame, e.g.,

- Univariate analysis: descriptive statistics (mean, median, mode, etc) and visualizations
- Multivariate analysis: correlations, missing data, pairwise interaction
- Compare datasets: fast and complete report on the comparison of datasets



age

Real number (ℝ)

Distinct	73	**Minimum**	17
Distinct (%)	0.2%	**Maximum**	90
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	38.581647	**Memory size**	254.5 KiB

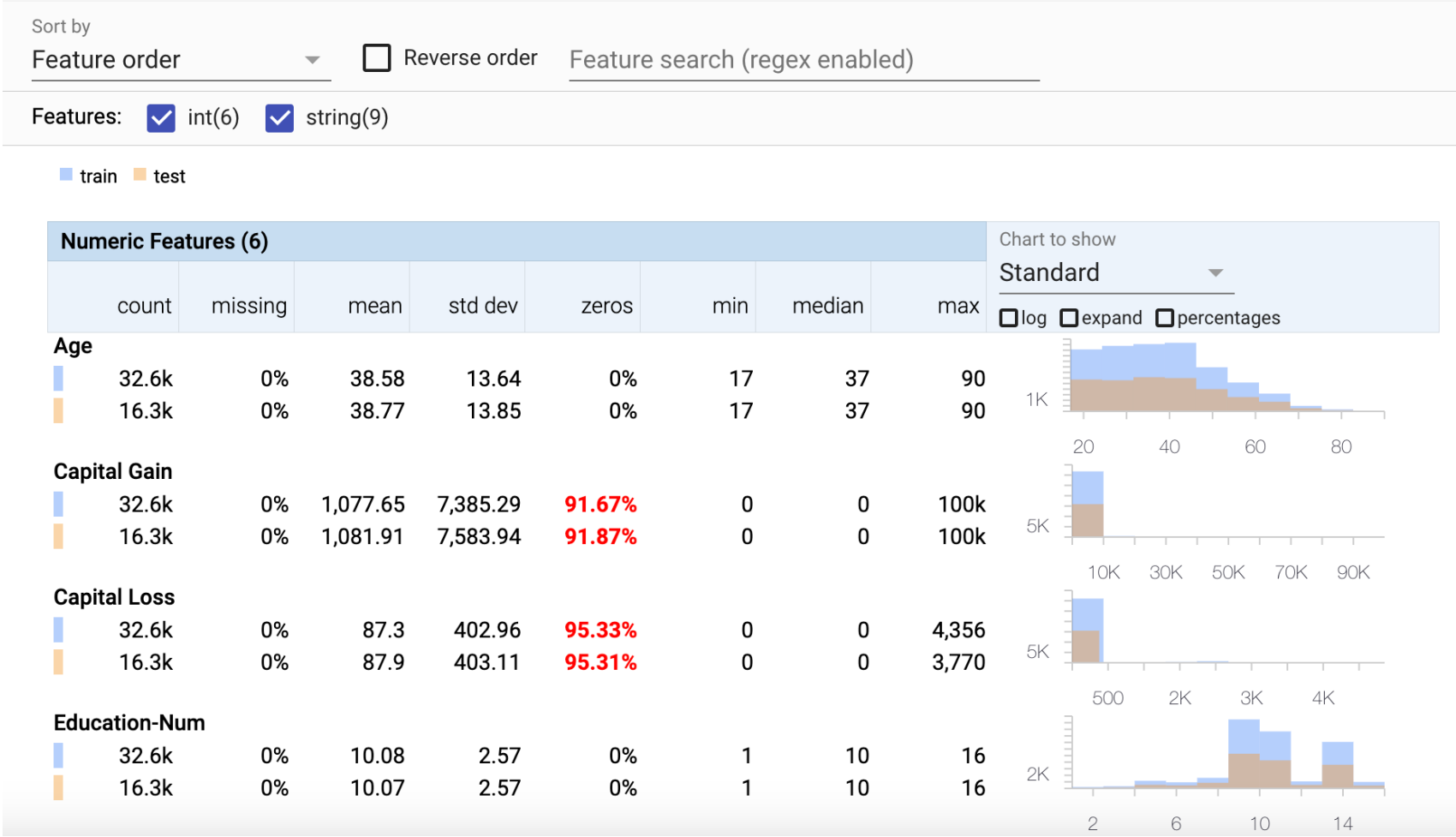
More details

https://github.com/ydataai/ydata-profiling

Exploratory data analysis – Example of tools

FACETS

- Visual feature-by-feature statistical analysis
- Data distribution, focus on common issues such as missing feature values.
- Exploration of the relationship between data points across the different features



https://pair-code.github.io/facets/

Dataset description

Goals

- **Communication**. Ensure proper between data creators and users
- **Transparency**. Clear data origin, characteristics, and potential biases. Users can understand how the data was collected, processed, and analyzed
- **Avoid misuse** of the data
- **Ethical Considerations**: Help in addressing systemic bias in models will be developed
- **Reproducibility**. Enable to reproduce the results or analyses conducted with the data

Dataset description

Goals - II

- **Data governance**. Provide guidelines and standards for data management practices
- **Collaboration and sharing**. Documented data can be easily shared and allows different users to understand and use the data effectively
- **Long-Term Preservation**. Ensures that data remains accessible and usable over time by maintaining information about its structure, format, and metadata.
- **Risk Management**. Documentation helps identify potential risks associated with the data, such as privacy concerns, security vulnerabilities, or data quality issues

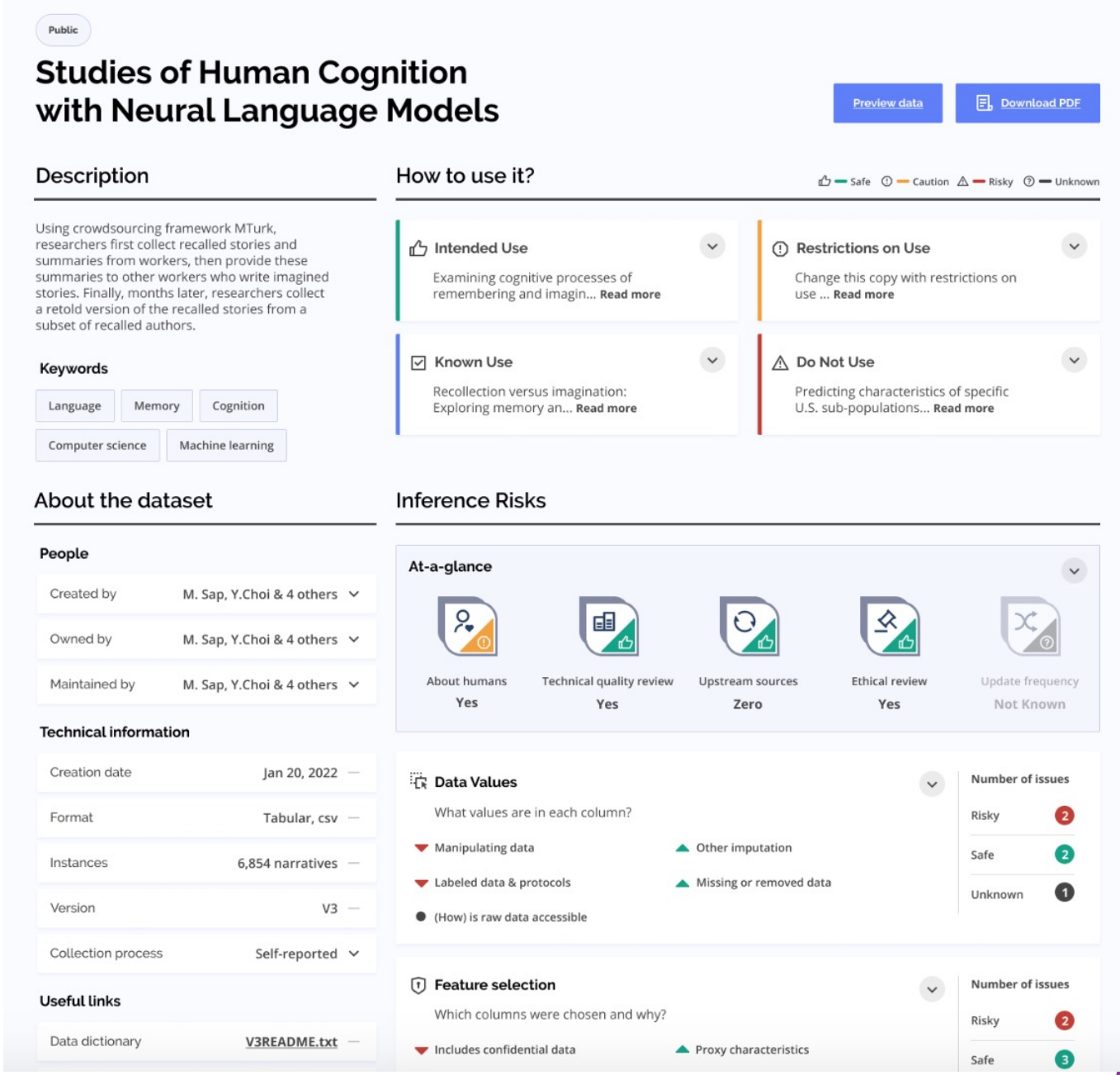
Dataset description – Points to address

Points to address

- **Motivation**: reasons for creating the dataset and info on who created or funded it
- **Composition**: info on what the dataset provides, presence of errors, noise or redundancies. Users can therefore decide if data suitable for their purposes
- **Collection process**: how the data were acquired, who involved in the collection and review
- **Preprocessing, cleansing or tagging**: info on preprocessing or software to perform it. Users can determine if data are processed in ways that are compatible with their intended uses
- **Uses**: info on the tasks for which the data may or may not be used
- **Distribution**: info on how the dataset will be disseminated, restrictions and licences
- **Maintenance**: planned maintenance and updates, support and communication to the users

Dataset description

- Multiple recommendations for standardizing dataset descriptions
  - datasheets for datasets
  - data statements
  - dataset nutrition labels



Example of dataset nutrition label

Gebru, T., et al. "Datasheets for datasets." Communications of the ACM 64.12 2021

Bender, E. and Friedman B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. TACL. 2018

Holland, Sarah, et al. "The dataset nutrition label." Data Protection and Privacy 12.12 (2020): 1.

Interpretable feature engineering

Selection and preprocessing of features preserving their interpretability

- **Selection of features**
  - Using a lower number of feature reduces the complexity and makes the process and the model easier to interpret
    - e.g., via recursive feature elimination
  - Prioritize also interpretable selection processes, e.g.,
    - driven by domain experts: they select the most important features for the process
    - correlation-based: compute correlations among features, keep only a/few representative(s) of correlated ones

Interpretable feature engineering

Selection and preprocessing of features preserving their interpretability

- **Interpretable feature engineering**
  - Creating or transforming features in a way that makes them understandable by humans
    - Discretization (from age to <30, 30-60, >60)
    - Semantic binning (from age to young, adult, senior)
    - Statistics over windows: e.g., from time series to mean, percentiles, standard deviation of windows
    - Domain Knowledge Integration: create domain-drive features that are meaningful for the problem and interpretable

References

- https://github.com/ydataai/ydata-profiling
- https://pair-code.github.io/facets/
- Gebru, T., et al. "Datasheets for datasets." Communications of the ACM 64.12 2021
- Bender, E. and Friedman B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. TACL. 2018
- Holland, Sarah, et al. "The dataset nutrition label." Data Protection and Privacy 12.12 (2020): 1.
- Bahador Khaleghi. The How of Explainable AI: Pre-modelling Explainability