Presented by Sophie X.

DATA READINESS FOR AI

DATA PREPROCESSING
1. Completeness
2. Feature Relevance
3. Correctness

DATA PREPARATION
4. Data Relevance
5. Timeliness
6. Outliers
7. Label Purity
8. Class Purity
9. Bias Index
10. Split Ratio

POST-MODELING
1. Assess Data Adequacy
2. Performance Drift

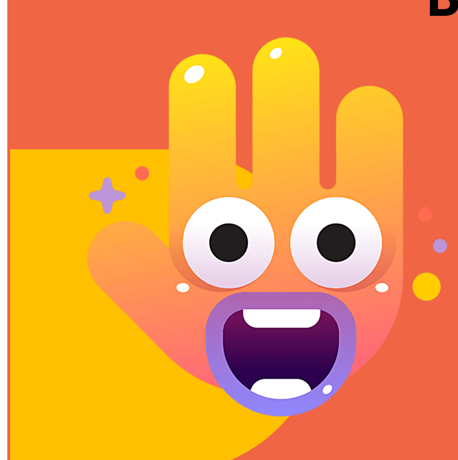
DOCUMENTATION
1. Data Lifecycle
2. Metadata

WHY DO WE CARE ABOUT DATA READINESS FOR AI PROJECTS? WHAT *IS* DATA READINESS?

Data readiness involves **ethically governing and curating high-quality data** throughout an AI system's lifetime, preventing

"GARBAGE IN, GARBAGE OUT".

BUT HOW DO WE KNOW OUR DATA IS READY?



The level of data readiness can be divided into bands:

ACCESSIBILITY: Ensure availability of the data
CORRECTNESS: Understand limitations of the data
TRANSFORMATION: Prepare the data for specific model
IN CONTEXT: Remove risk and reduce bias in data
READY: Fits the use case

Follow the stages listed to enhance data readiness for your AI project!

LIFECYCLE

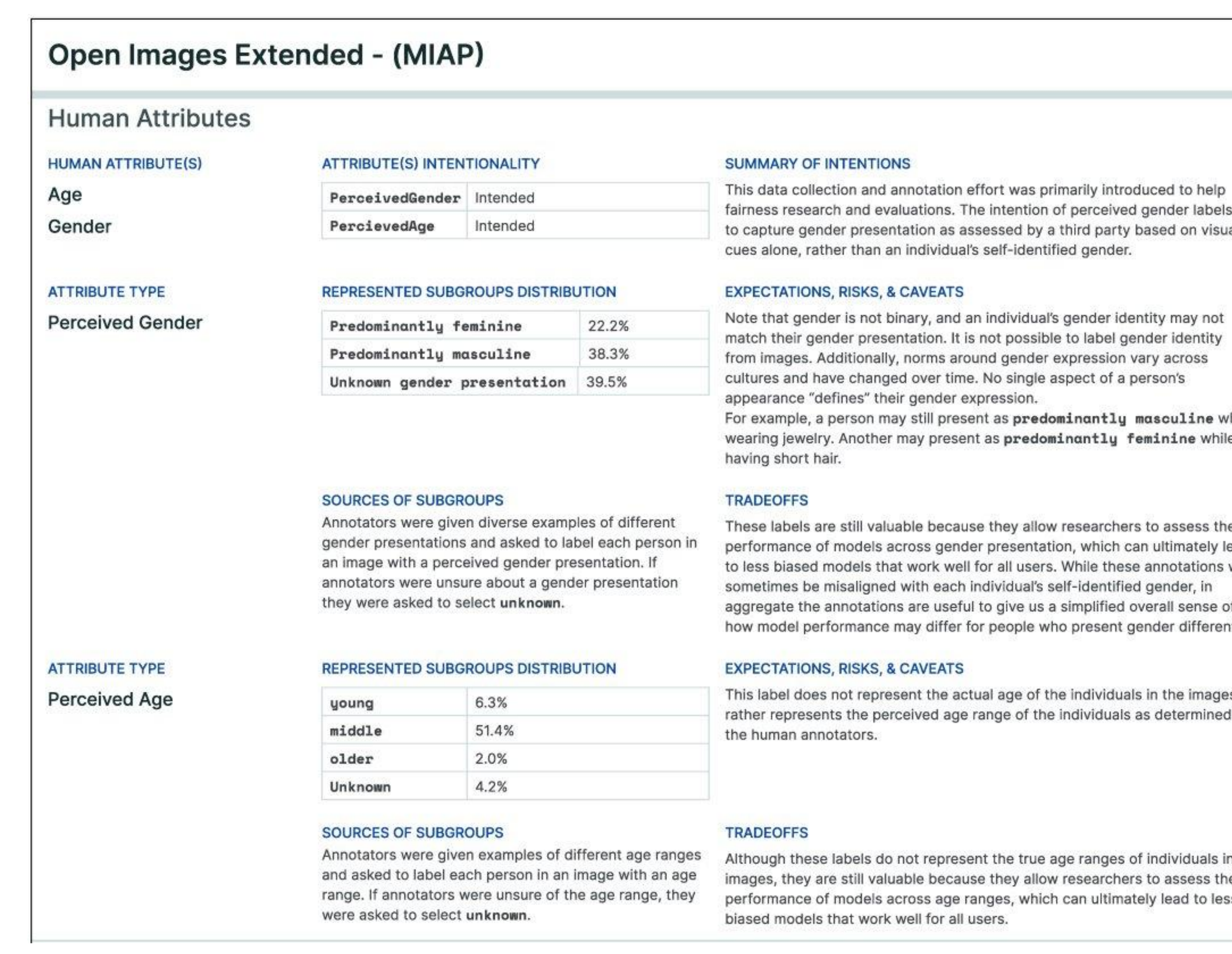
The **Data Readiness Report** documents data quality and preprocessing steps, and boosts transparency and standardization in AI documentation.

Baseline Quality and Readiness Assessment

Quality Metric	Description	Explanation	Recommendations
Feature Relevance (15%, 0.85, 85%)	*Identifies and ranks the feature based on Relevance. A score of 1 indicates that all features are relevant.*	*Less relevant features found in the dataset are 2 / 14 giving a quality score of 0.85*	*Drop less relevant features from the data.*
Data Completeness (1, 100%)	*Identifies missing values in the given data. A score of 1 indicates no missing values found in the data.*	*Missing values detected in 0 / 488415 entries giving a quality score of 1.0.*	*No action required*
Outlier Detection (11%, 0.89, 89%)	*Identifies outlier samples in the data. A score of 1 indicates no outliers found in the data.*	*Outliers detected in 3262 / 32561 rows giving a quality score of 0.89.*	*Remove Outliers*

METADATA

The **Data Card** goes beyond metadata to include explanations, rationales, and instructions related to dataset provenance, representation, usage, and fairness evaluations for ML models.



Open Images Extended - (MIAP)

Human Attributes

HUMAN ATTRIBUTE(S)
Age
Gender

ATTRIBUTE(S) INTENTIONALITY

PerceivedGender	Intended
PercievedAge	Intended

SUMMARY OF INTENTIONS
This data collection and annotation effort was primarily introduced to help fairness research and evaluations. The intention of perceived gender labels is to capture gender presentation as assessed by a third party based on visual cues alone, rather than an individual's self-identified gender.

ATTRIBUTE TYPE
Perceived Gender

REPRESENTED SUBGROUPS DISTRIBUTION

Predominantly feminine	22.2%
Predominantly masculine	38.3%
Unknown gender presentation	39.5%

EXPECTATIONS, RISKS, & CAVEATS
Note that gender is not binary, and an individual's gender identity may not match their gender presentation. It is not possible to label gender identity from images. Additionally, norms around gender expression vary across cultures and have changed over time. No single aspect of a person's appearance "defines" their gender expression.
For example, a person may still present as **predominantly masculine** while wearing jewelry. Another may present as **predominantly feminine** while having short hair.

SOURCES OF SUBGROUPS
Annotators were given diverse examples of different gender presentations and asked to label each person in an image with a perceived gender presentation. If annotators were unsure about a gender presentation they were asked to select **unknown**.

TRADEOFFS
These labels are still valuable because they allow researchers to assess the performance of models across gender presentation, which can ultimately lead to less biased models that work well for all users. While these annotations will sometimes be misaligned with each individual's self-identified gender, in aggregate the annotations are useful to give us a simplified overall sense of how model performance may differ for people who present gender differently.

ATTRIBUTE TYPE
Perceived Age

REPRESENTED SUBGROUPS DISTRIBUTION

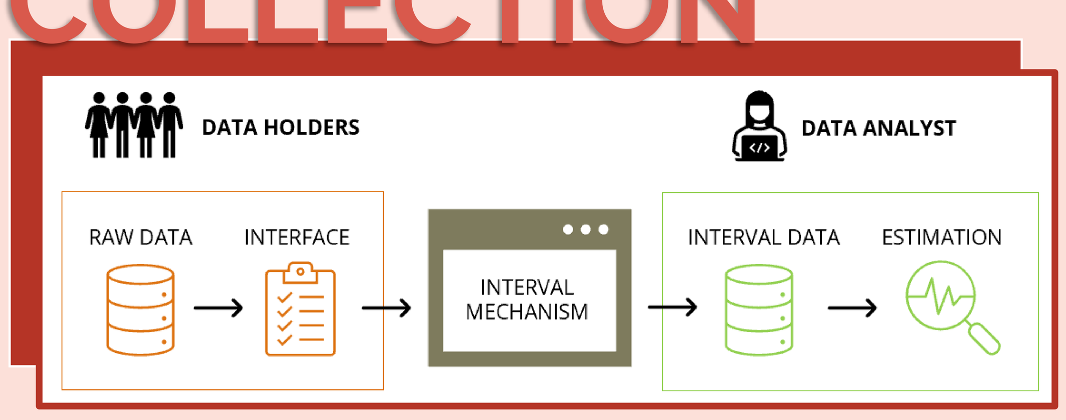
young	6.3%
middle	51.4%
older	2.0%
Unknown	4.2%

EXPECTATIONS, RISKS, & CAVEATS
This label does not represent the actual age of the individuals in the images. It rather represents the perceived age range of the individuals as determined by the human annotators.

SOURCES OF SUBGROUPS
Annotators were given examples of different age ranges and asked to label each person in an image with an age range. If annotators were unsure of the age range, they were asked to select **unknown**.

TRADEOFFS
Although these labels do not represent the true age ranges of individuals in images, they are still valuable because they allow researchers to assess the performance of models across age ranges, which can ultimately lead to less biased models that work well for all users.

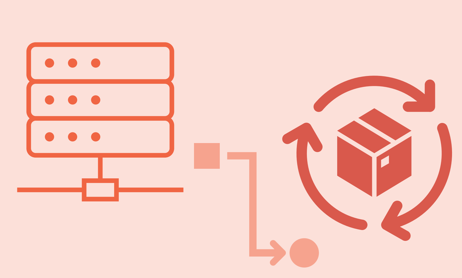
COLLECTION



Problem: Trade-off between data privacy and data completeness.

Interval privacy represents raw data as intervals, preserving privacy through ambiguity, enhancing transparency, flexibility, and fidelity.

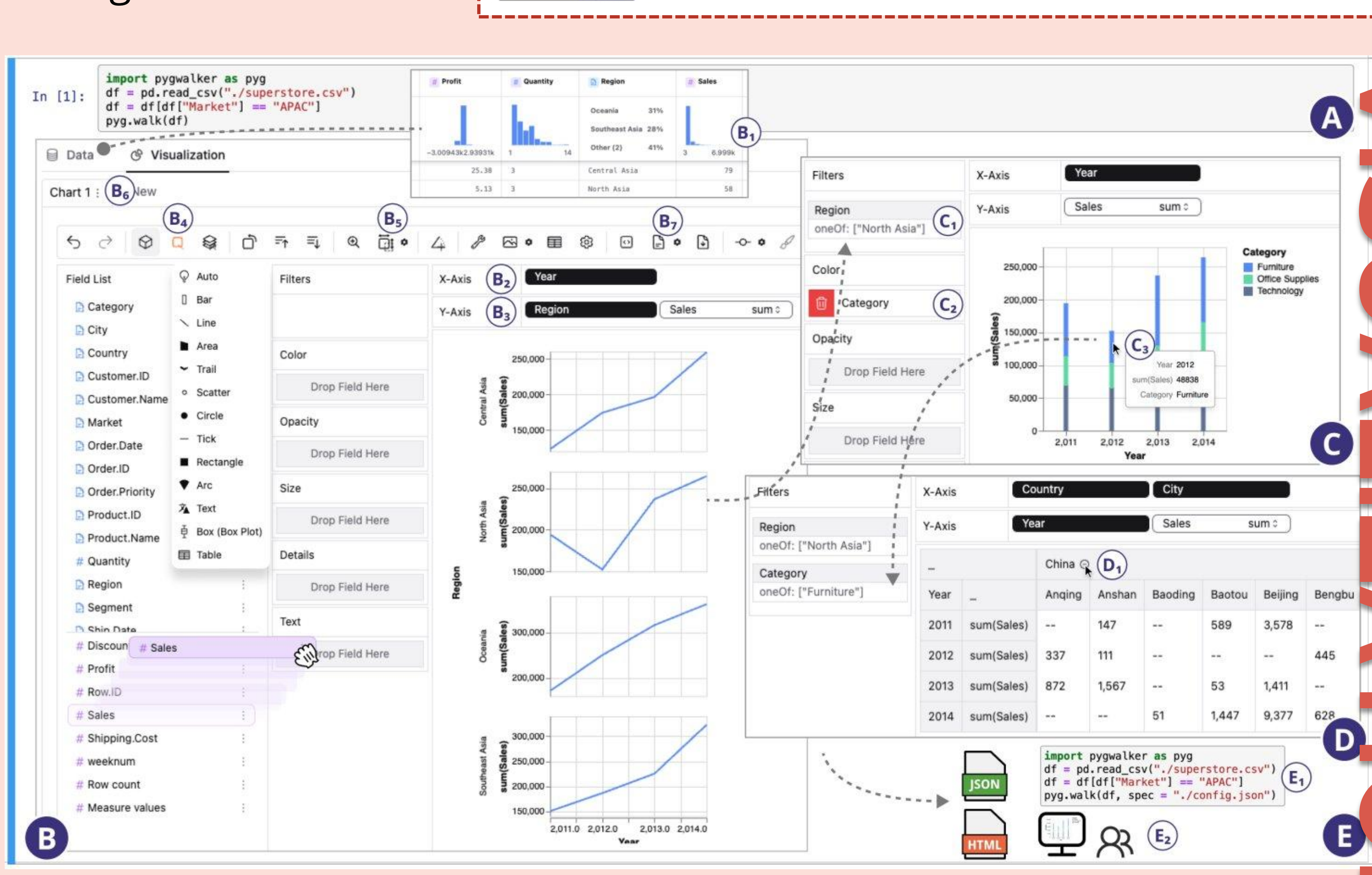
PREPROCESSING



DiffPrep automatically searches for a data preprocessing pipeline that maximizes the performance of the ML model given a dataset.

Identify Missingness	Imputation Methods	Encoding Techniques	Discretization	Outliers
MAR	Mean & Median	One Hot Encoding	Equal Width	Hypothesis Testing
	End of Distribution	Frequency Encoding	Equal Frequency	Z-Score
MNAR	Arbitrary Value	Label Encoding	K-Means	Robust Z-Score
	Frequent Value	Ordinal Encoding	Decision Tree	IQR Method
MCAR	Missing Category	Mean Encoding	Custom	DBSCAN Clustering
	K-Near Neighbors			Isolation Forest
				Data Visualization

VISUALIZATION



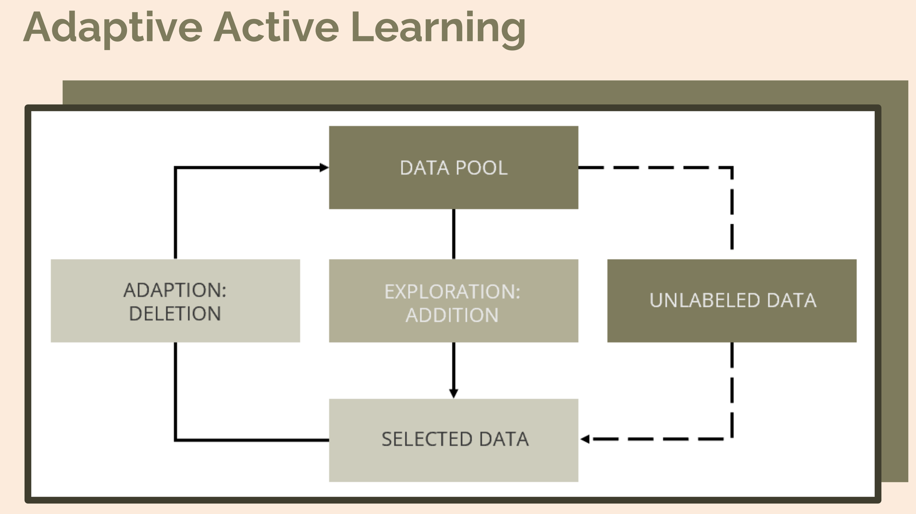
PyGWalker offers on-the-fly assistance for exploratory data analysis.

PURITY

Problem: Increasing demand for high-quality labeled datasets. Label bias occurs due to differing interpretations of ambiguous terms.

Crowdsourced annotations provide scalability, demographic bias mitigation, and disagreement information utilization.

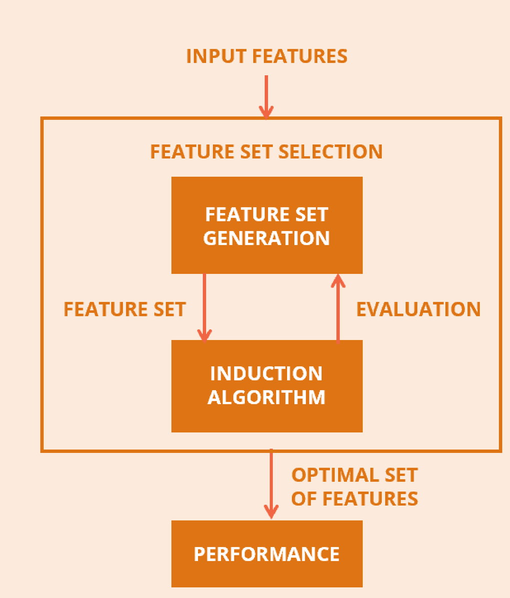
Adaptive Active Learning



Problem: Training on imbalanced dataset causes small-sample bias. This leads to poor performance in detecting rare events.

Metric	Defined as
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
Negative Predictive Value	$\frac{TN}{TN+FN}$
Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$
F_1 score	$2\frac{PRC \cdot SNS}{PRC+SNS}$
Geometric Mean	$\sqrt{SNS \cdot SPC}$
Matthews Correlation Coefficient	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Bookmaker Informedness	$SNS + SPC - 1$
Markedness	$PPV + NPV - 1$

RELEVANCE



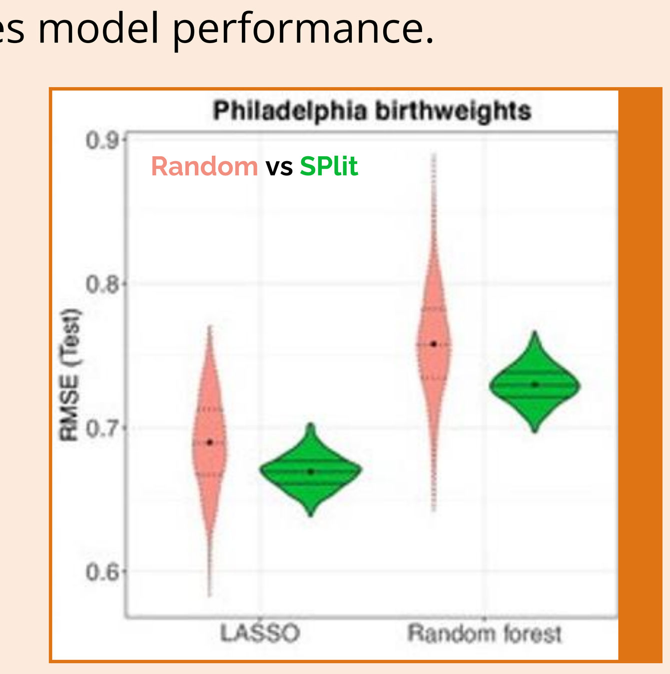
Problem: Feature Selection (Filter, Wrapper, Embedded Model) is essential for reducing data dimensionality and raising accuracy.

Data Relevance: the suitability of data for analysis. Clear project objectives and data assessment are crucial.

Timeliness: "data is made available as fast as necessary to preserve the value of the data".

SPLIT RATIO

Problem: Split ratio is crucial for data with high autocorrelation to avoid data leakage among training and test sets which exaggerates model performance.



SPlit adapts support points and uses sequential nearest neighbor for subsampling.

Rule: Small 70:20:10 Large 98:1:1 Tune Ratio

POINT IMPACT

Scikit Learn (Python) includes tools like isolation forest and local outlier factor.

ELKI (Java) is an open-source benchmarking and fairness assessment test for algorithms.

Types: Implicit bias, Selection bias, Measurement bias, Confounding bias, Algorithmic bias, Temporal bias.

Unlabeled Data framework consists of 1) pseudo labeling, 2) re- sampling and 3) fair ensemble learning.

BIAS INDEX

POST-MODELING

Problem: Decline in model performance caused by Data Drift (changes in input data) and Concept Drift (changes in real-world that make learned rules obsolete).

NannyML is an open-source python library for detecting and correcting data and concept drift.