The Alan Turing Institute

Fairness of AI Systems: Identifying and Mitigating Harmful Bias

Dr Alpay Sabuncuoglu

Alpay Sabuncuoglu

- Principal Investigator – Innovate UK-funded “Proactive Monitoring of AI Fairness” project
- **Research interests:** Human-AI Interaction, Fairness of AI Systems, Responsible AI
- **Other Turing Projects:**
    - Trustworthy Digital Infrastructure for Identity Systems
    - LLMs in Finance
    - AI Model Risk Management
- **Previously:**
    - Computer Science and Engineering PhD, Koc University, Istanbul
    - Intelligent User Interface, Educational Technologies, HCI
    - Large-scale quantitative and qualitative studies

Our Innovate UK-funded Project: Proactive Monitoring of AI Fairness

Motivation: "*Despite increased interest in addressing bias and discrimination in AI systems, organisations continue to face numerous challenges*"

One challenge is SMEs lack the necessary skills and resources to adhere to existing guidelines.

Aim: The project aims to enable a proactive fairness review approach in the early stages of AI development and provide developer-oriented methods and tools to self-assess and monitor fairness.

Main Outputs:

(1) ***Code analysis tool*** that can enable developers to annotate, review and monitor fairness issues in their development flows.

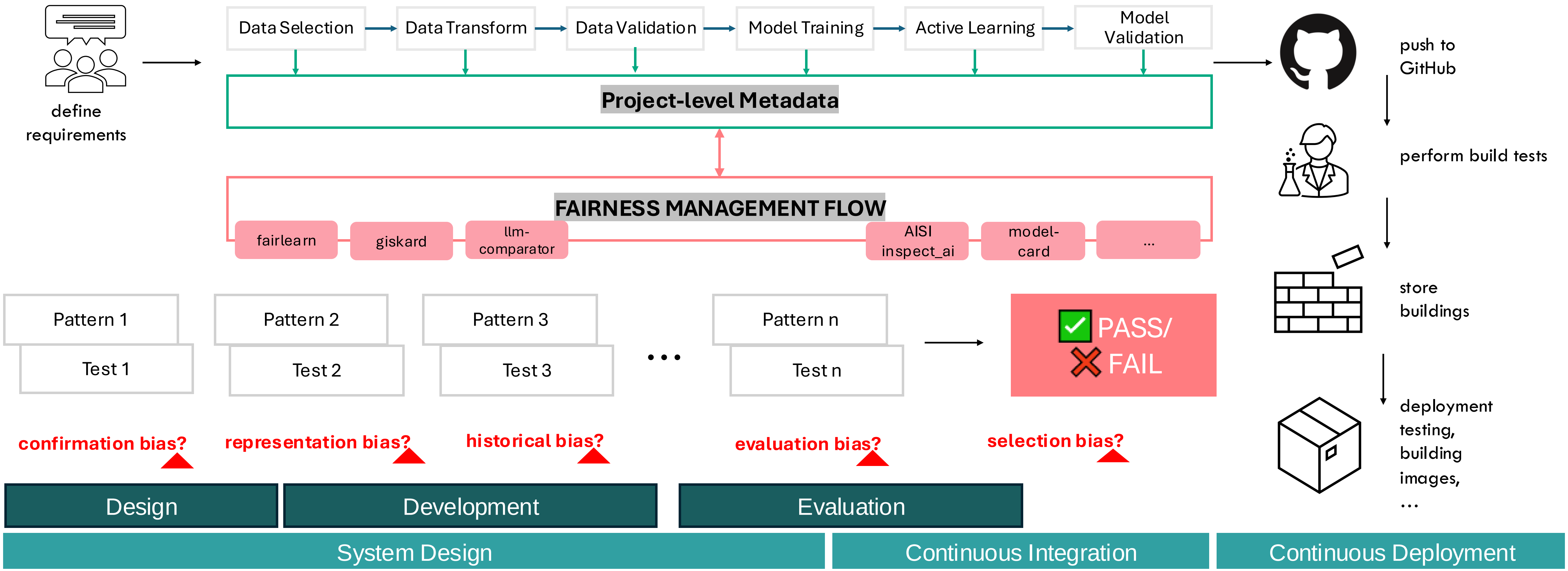
(2) ***Design patterns*** for fair AI development to support transferring knowledge between different disciplines, producing concrete and actionable outputs, and ensuring effective technical development "by design".

(3) ***Tutorials and skill development activities*** to integrate fairness considerations into AI systems of SMEs through improving developer knowledge and skills.

Overview of the Project (May 24-Jan25)

Use Case: LLMs in Financial Services

(1) News sources change rapidly, (2) constant information flux, and (3) large, noisy data. LLMs can perpetuate and amplify bias in this challenging environment.



Agenda

10:00 - 10:15	Welcome and Introduction
10:15– 10:45	Fairness in AI, Trustworthy AI Components
10:45 – 11:15	Workshop Activity Introduction: Use Case
11:15 – 11:30	Break
11:30 – 12:15	Bias throughout ML Development, Credit Scoring Use Case
12:15 – 13:30	Potential Harms and Stakeholder Responsibilities
13:30 – 14:30	Lunch
14:30 – 15:15	Recording Fairness Metadata with FAID + Metadata Discussion
15:15 – 15:45	Presentation of Workshop Groups (3 groups)
15:45– 16:00	Future Steps and Closing Remarks

Studies highlight that bias in AI is widespread and detrimental, prompting the development of fairness definitions and algorithms.

Age bias in AI-leaning jobs means hurdle for midcareer workers

By Tony Case

Apple's 'sexist' credit card investigated by US regulator

Uber Eats driver to amend lawsuit alleging racial bias by app's biometric verification

Nov 14, 2023, 2:43 pm EST | Bianca Gonzalez

Bias, Discrimination, Fairness

Fairness in AI aims to ensure that the model's outcomes are equitable across different groups, minimising biases and discrimination.

A **fairness notation** is a formal representation or expression used to define fairness criteria in an AI model.

A **fairness metric** is a quantifiable measure used to evaluate how well an AI system adheres to fairness criteria, often comparing performance across different demographic groups.

Bias in AI development refers to systematic errors that result in unfair outcomes, often due to imbalanced or prejudiced data used in training.

Discrimination can occur when an AI system treats individuals or groups differently based on sensitive attributes, like race or gender, leading to unequal outcomes.

The lack of any standardised definitions for key terms, such as ethics, fairness and transparency, and the lack of any standardised measurements for such principles make it more difficult for firms to implement high-level principles.

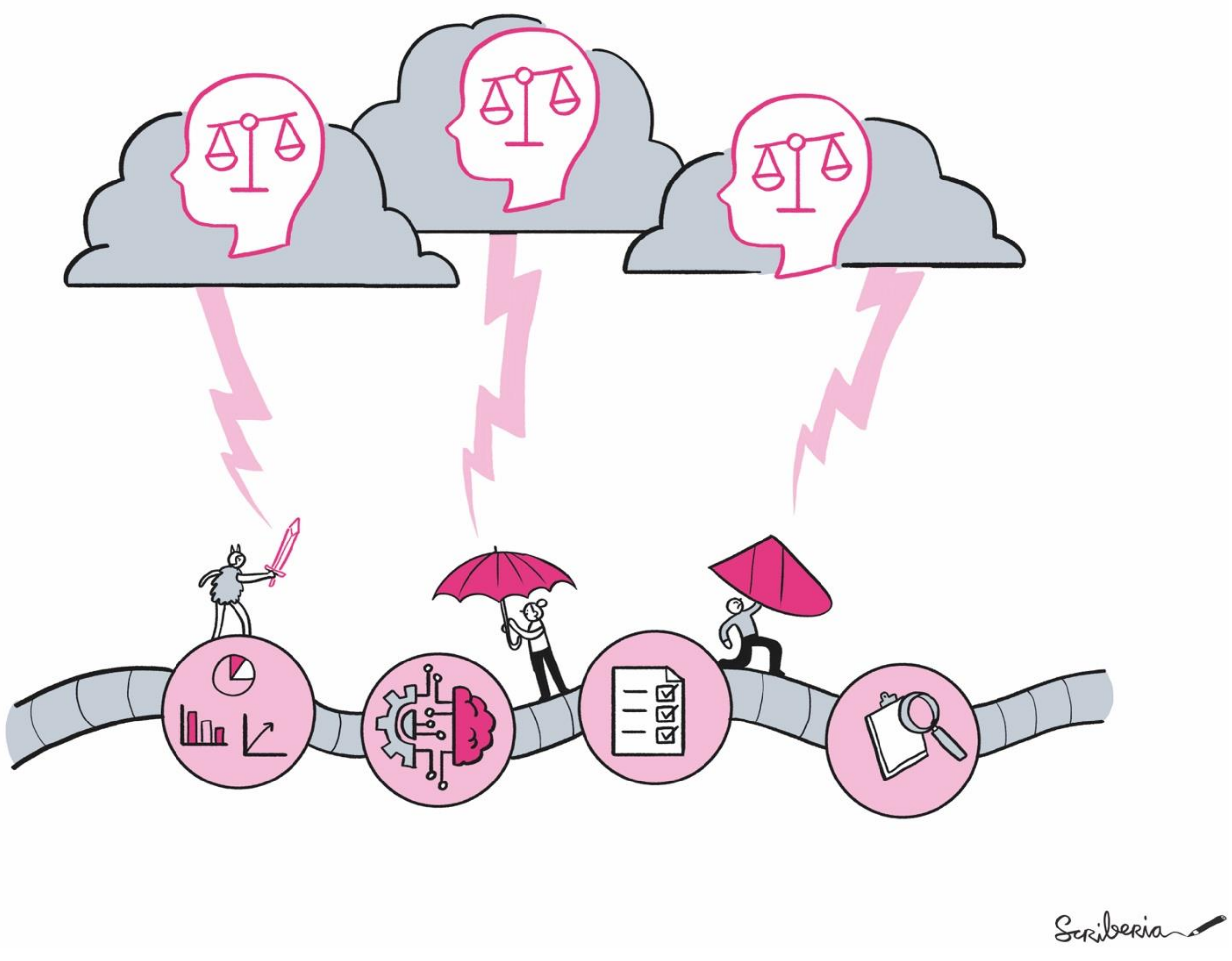
From FCA’s Public-Private AI Forum

Fairness is not solely a technical issue; it encompasses social, political, philosophical, and legal dimensions.

- AI systems can inadvertently perpetuate and amplify existing societal biases, leading to unfair outcomes for certain groups.

- Interdisciplinary approaches are necessary to analyse AI fairness and its societal implications.

- We need a holistic, proactive approach to eliminate unwanted bias.



Regulators, standards bodies, and international agencies take action with the expanding influence of AI across diverse fields.

ISO/IEC/JTC1/SC42

Artificial Intelligence

WG 01	WG 02	WG 03	WG 04	WG 05	JWG 01
Foundational Standards	Data	Trustworthiness	Use Cases & Applications	Computational Approaches and Computational Characteristics of AI System	Governance Implications of AI

Department for Science, Innovation & Technology

Guidance

Introduction to AI assurance

Published 12 February 2024

Independent report

AI Opportunities Action Plan

Published 13 January 2025

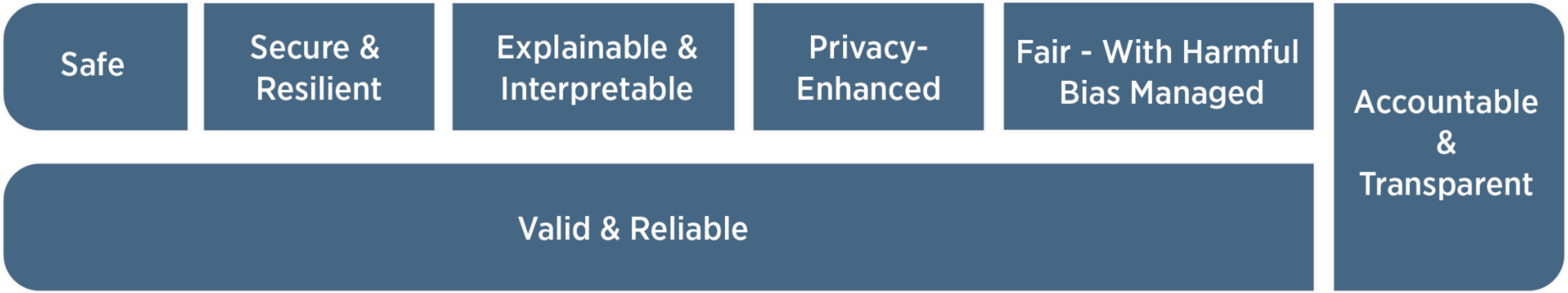
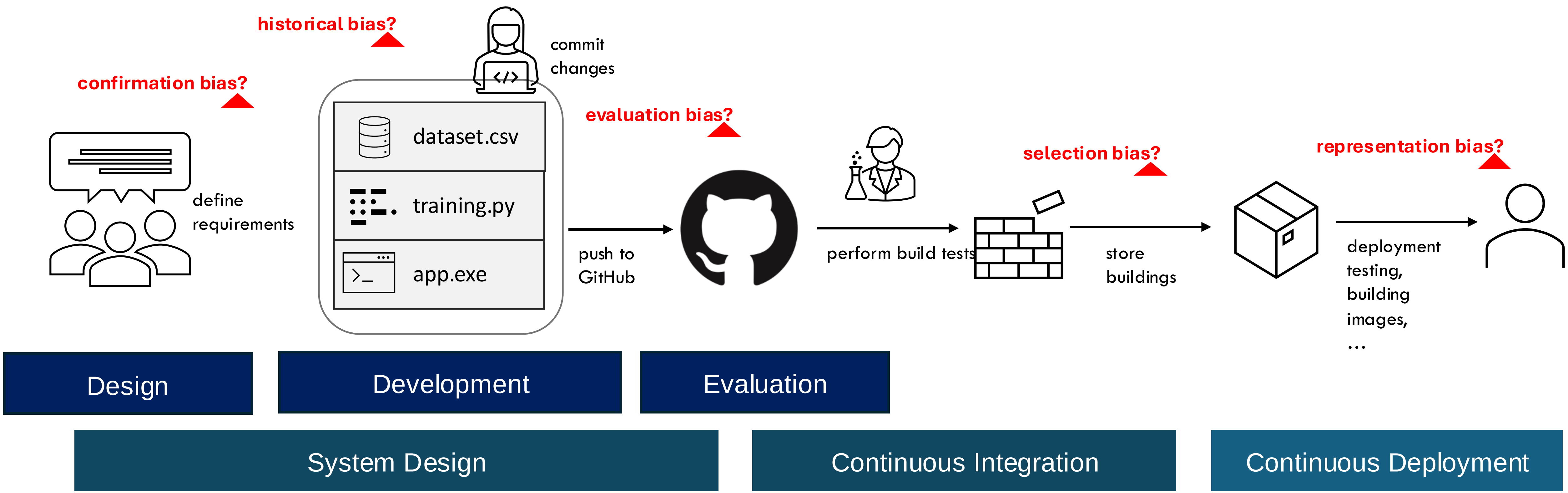


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

From NIST AI RMF – Trustworthy AI Characteristics

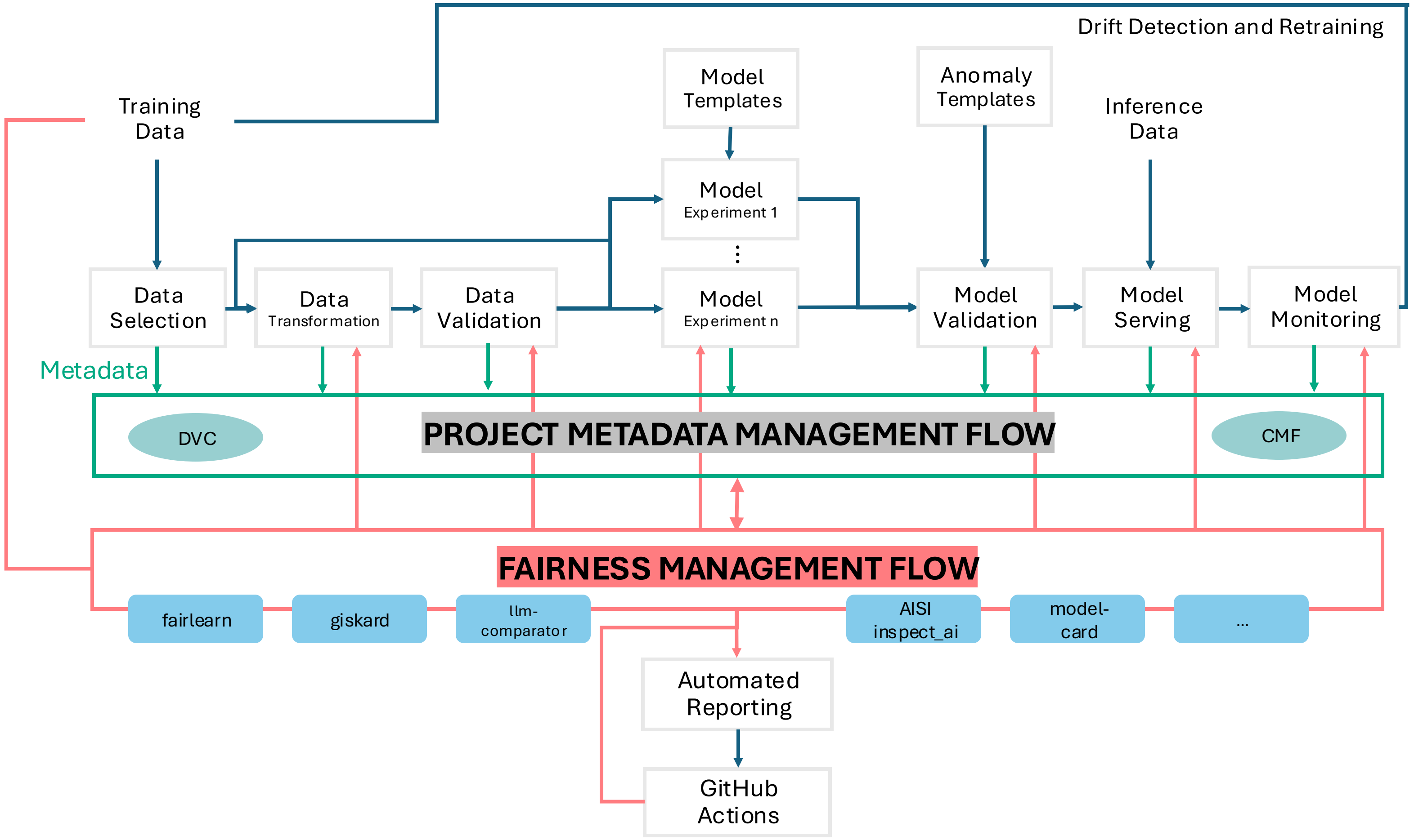
Bias in Production Lifecycle



A Traditional ML Development Pipeline with Version Control

Design Principles

- Use popular orchestration solutions.
- Develop a platform and tech-agnostic solution.
- Utilise actively contributed libraries to our advantage.
- Create a highly granular workflow with atomic modules.



Lifecycle Stage	Bias Source	Description	Examples
Data Collection	Sampling Bias	Certain perspectives, demographics, or groups are overrepresented or underrepresented in the data.	A dataset for a news aggregator containing primarily sources that favour a particular ideology, leading to skewed results
	Selection Bias	Only certain data types or contexts are included, limiting representativeness.	Language datasets that exclude non-Western languages, limiting model performance in global applications.
Data Annotation	Labeller Bias	Annotators' backgrounds, perspectives, and cultural biases affect their understanding and classification of data, influencing the labelling process.	Annotators label speech by individuals from lower socioeconomic backgrounds as unprofessional or inappropriate, leading to biased decisions.
Data Curation	Historical Bias	Reflecting or perpetuating past societal biases within curated data.	A hiring dataset that favours certain demographics based on historical hiring practices, embedding existing inequalities in AI models.
Data Pre-processing	Feature Selection Bias	Excluding relevant features from a dataset.	Excluding age or gender as features in healthcare models, potentially impacting the relevance of predictions for these demographics.

From the latest International AI Safety Report

Model Training	Label Imbalance	Unequal representation in labelled data, leading to biased model outputs.	A classification model trained on 80% male-labelled images might perform poorly when identifying female images.
Deployment Context	Contextual Bias	A model is trained on data from a context that differs from the context of application, leading to worse outcomes for certain groups.	An English-only model deployed in multilingual settings, causing misinterpretations for non-English users.
Evaluation & Validation	Benchmark Bias	Evaluation benchmarks favour certain groups or knowledge bases over others.	AI models evaluated primarily on US-centric datasets fail to generalise well in non-Western settings.
Feedback Mechanisms	Feedback Loop Bias	Models learn from biased user feedback, reinforcing initial biases.	A recommendation system that receives more engagement on certain types of content may reinforce exposure to the same biased content.

From the latest International AI Safety Report

Technical Skill Capacity for Effective Fairness Monitoring

AI Leader

AI Professional

AI Worker

AI Citizen

Data Selection → Data Transform → Data Validation → Model Training → Model Validation → Model Deployment

Data Lifecycle

Engineering and Infrastructure

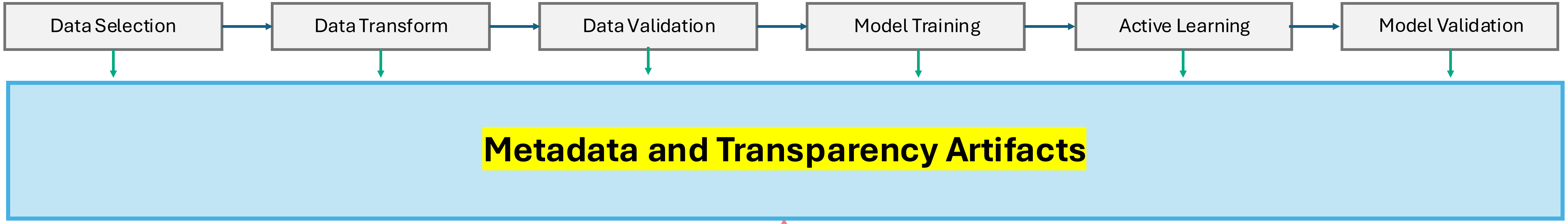
Problem definition

Problem solving and analysis

Evaluation and reflection

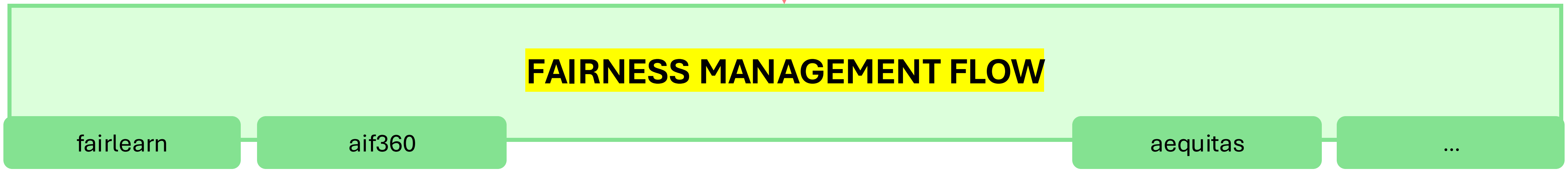
See *AI Skills for Business Competency Framework:* https://doi.org/10.5281/zenodo.11092677

Overview of ML Pipeline with Fairness Data

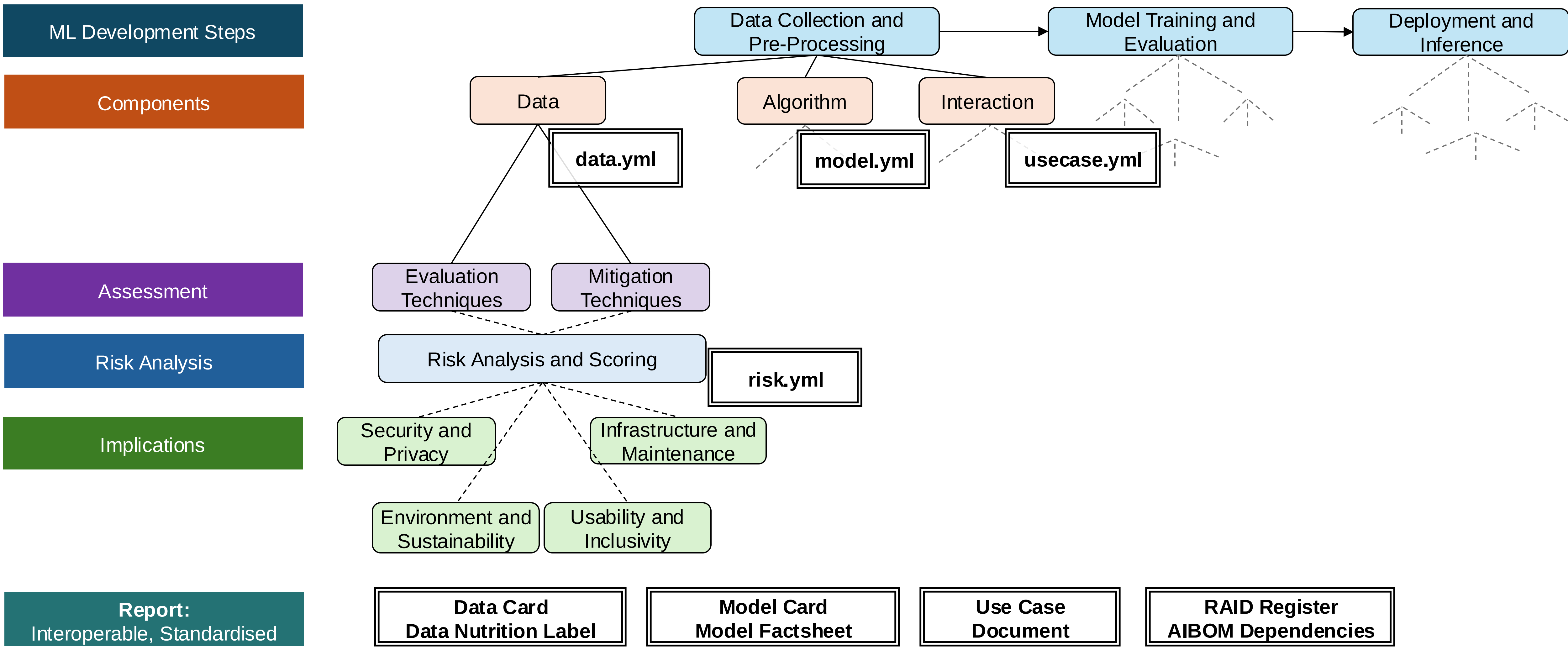


	Data-related analysis	Model analysis
Evaluate	- Testing unawareness - Feature importance	- Model prediction analysis based on the fairness notion
Mitigate	- Counterfactual inputs	- Threshold optimisation

Interaction analysis
- Reviewing existing applications - Follow ISO-9241 Human System Interaction Ergonomics Standard - Use Microsoft Human-AI Interaction Guideline



Overview of Complete Pipeline



FAID Artefacts: YAML Log Files and HTML Reports

Changes in the log files

Local: Experiment-based

- **fairness.yml** — fairness experiment results to be synced with model, data, and risk logs

Global: Shared Accross Diff. Experiments

- **data.yml** — dataset related metadata such as representativeness of protected characteristics
- **model.yml** — model related metadata such as architecture and subgroup performance
- **risk.yml** — logged risks, assumptions issues, dependencies

Communication

data.html:
Visual report of the data card and metadata log

risk.html
RAID-like risk register report

model.html
Visual report of the model card and metadata log

fairness.html
Experiment report for each fairness evaluation and mitigation study

Credit Risk Scoring Analysis Use Case

DATA

Open-Source Datasets:

- German Credit Risk
- Lender's Club
- Singapore Credit Risk …

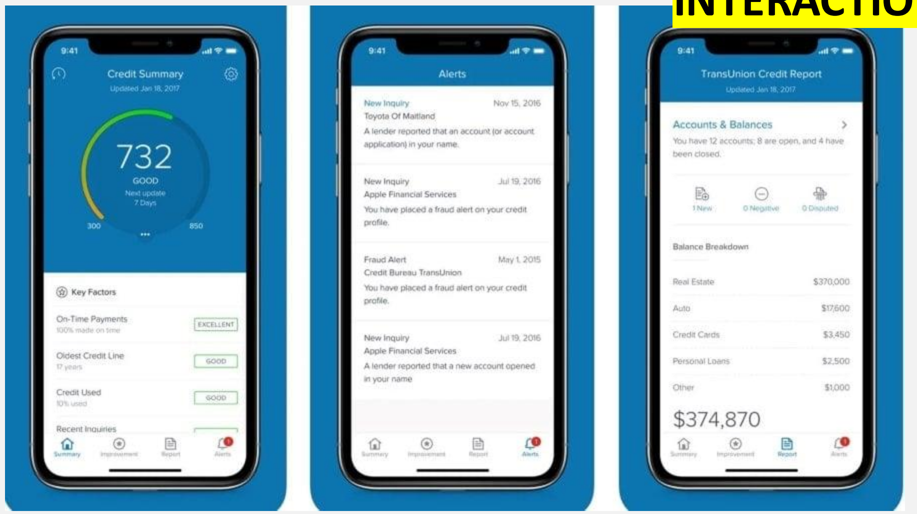
MODE
L

Decision tree

Gradient boosting

INTERACTION

e.g. Credit Score Apps



- Use sub-group discovery to understand the impact of feature combination.
- Use counterfactual data to test individual fairness.

- Test performance against different fairness notions with different metrics.

- Use Human-AI Interaction guidelines to evaluate overall interaction

Protected Characteristics

EHRC:

- Age,
- Disability,
- Gender Reassignment,
- Marriage and civil partnership,
- Pregnancy and maternity,
- Race,
- Religion or belief,
- Sex,
- Sexual orientation.

Age | Disability | Sex | Race | Religion or Belief

Pregnancy and Maternity | Marriage and Partnership | Gender Reassignment | Sexual Orientation

Table 1. Example Proxy Relationships Based on Findings from References [25, 38, 106, 137, 210, 251, 259, 260, 296, 304]

Sensitive Variable	Example Proxies
Gender	Education Level, Income, Occupation, Felony Data, Keywords in User Generated Content (e.g., CV, Social Media), University Faculty, Working Hours
Marital Status	Education Level, Income
Race	Felony Data, Keywords in User-generated Content (e.g., CV, Social Media), Zipcode
Disabilities	Personality Test Data

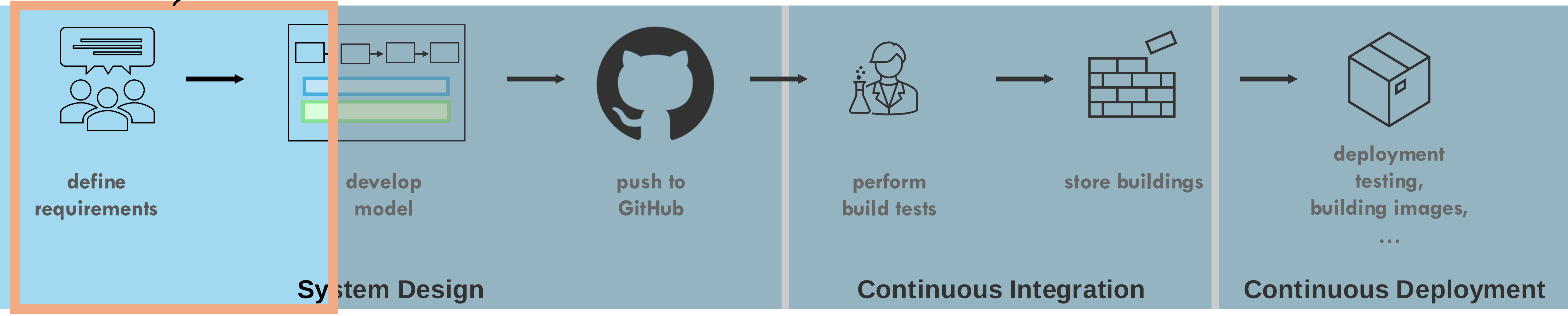
Caton et al. Fairness in Machine Learning: A Survey. 2024

Setting Up the Requirements and Objectives

We can decide to use ***Equal Opportunity*** *as* our fairness notion.

It is popular because it is easy to interpret in terms of real-world impact and quantitative analysis.

Quantitative analysis can simply compare true positive rates of protected characteristics groups.



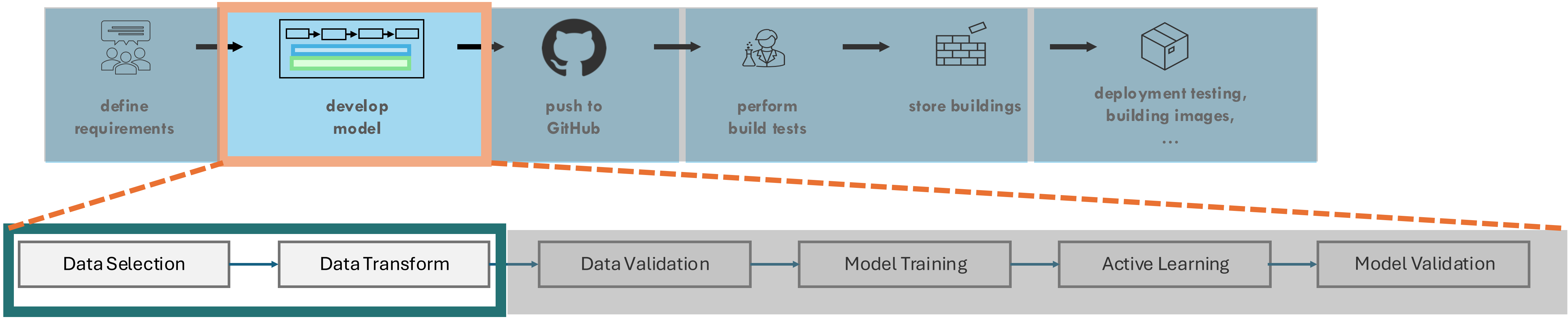
1. How do you define ***target variable*** and ***class labels***?

2. Which ***protected characteristics*** does your data contain?

3. How do you select/create features? Do these features contain any **proxies**?

4. How do you assess the impact of selected features and class definitions?

Identifying Protected/Sensitive Characteristics in Data



- Data: German Credit Data
- Entities:

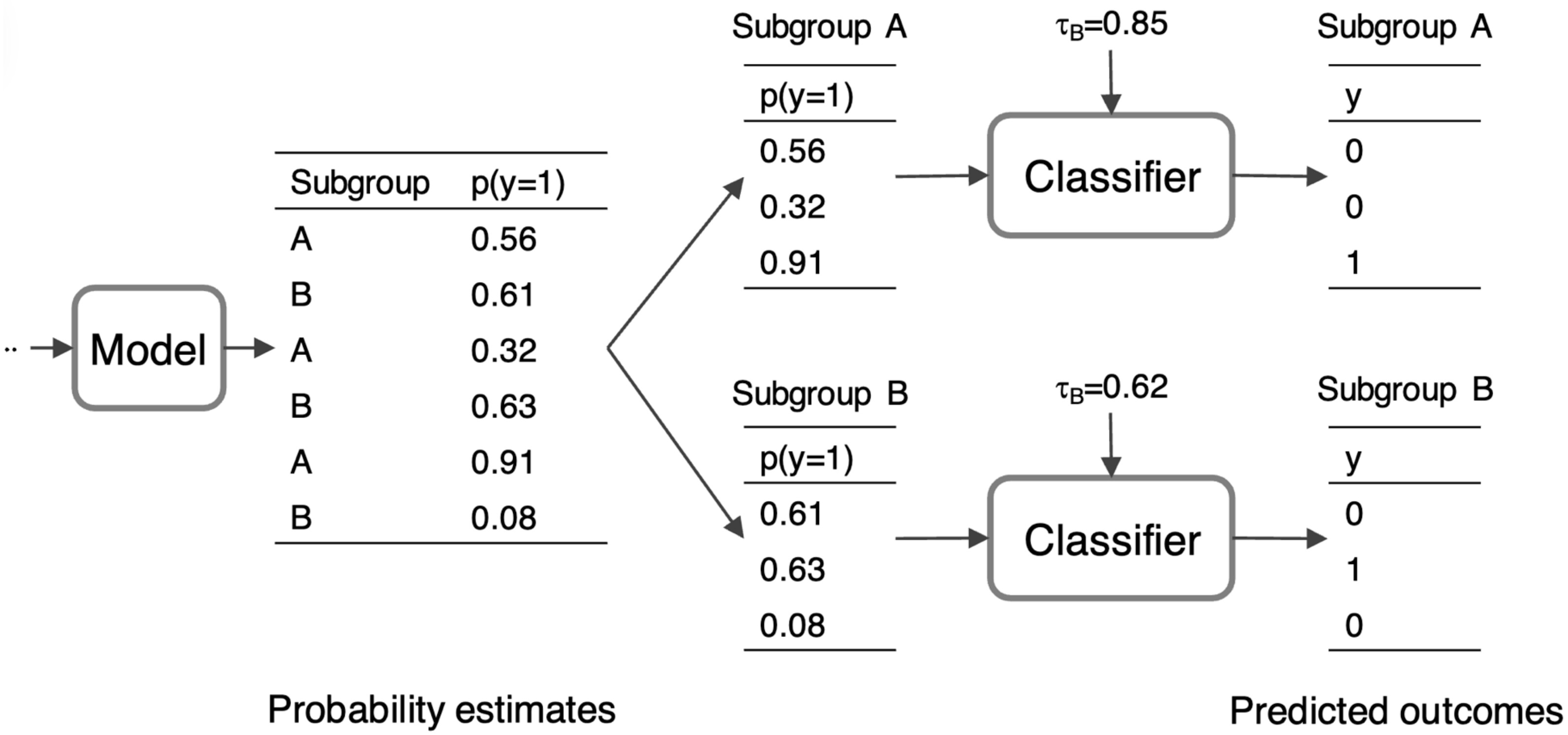
```
"ID","LIMIT_BAL","SEX","EDUCATION","MARRIAGE","AGE","PAY_0","PAY_2","PAY_3","PAY_4","PAY_5",
"PAY_6","BILL_AMT1","BILL_AMT2","BILL_AMT3","BILL_AMT4","BILL_AMT5","BILL_AMT6","PAY_AMT1",
"PAY_AMT2","PAY_AMT3","PAY_AMT4","PAY_AMT5","PAY_AMT6","default.payment.next.month"
```

- SEX column → Obvious protected characteristic
- EDUCATION, MARRIAGE, AGE → Let's check Equality Act 2010 definition
- How can you possibly know whether BILL_AMT2 is an indicator of a protected characteristic?

fairness.yml
- data:preprocessing
- bias_metrics:group

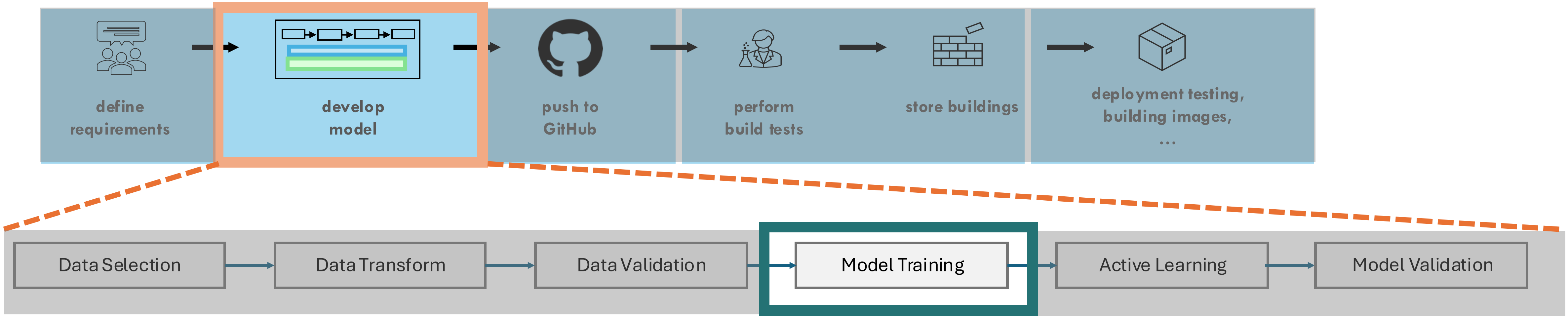
Subgroup Discovery

- Identifying whether our model performs significantly differently for a subgroup is a key step to understanding fairness issues.
- The subgroup can be formed by a combination of features, or partitions of selected features.

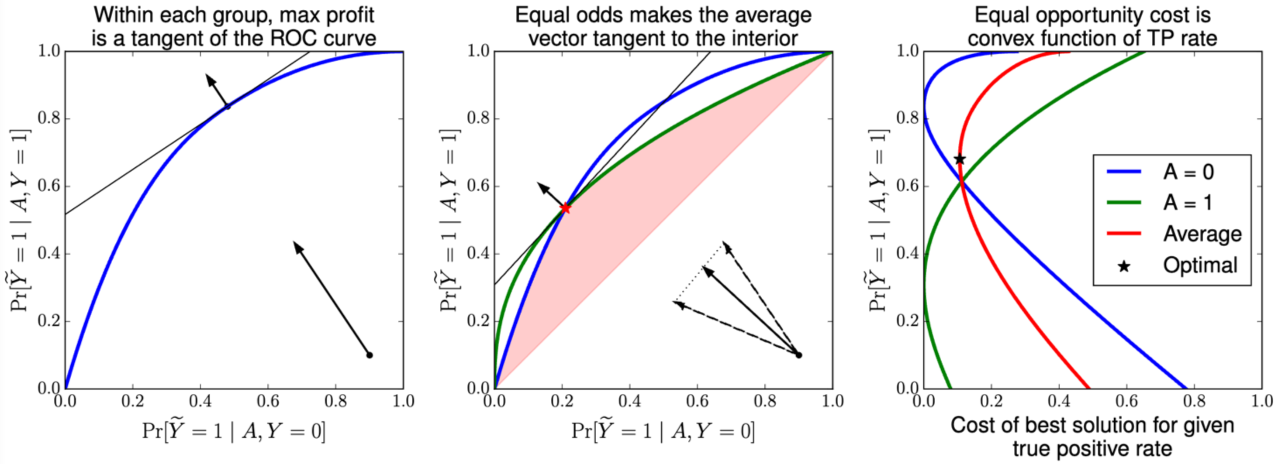


Model Fairness Performance

fairness.yml
- bias_metrics

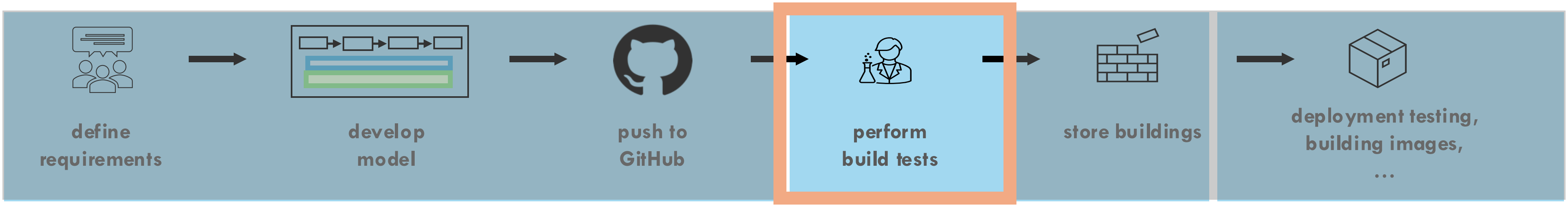


- Most fairness problem is an optimisation problem. Optimising the trade-off between fairness vs accuracy.
- For the given fairness notion (in this case, equal opportunity), derive an optimal fairness notion threshold predictor.



Continuous Testing

GitHub Actions or other automated CI/CD tools



Implement Data Unit Tests:

- Check min and max to identify the tail ends match with the real-life scenarios
- Check column distribution to monitor balanced representation
- Check mean and median values to identify if balanced representation aligns with these values
- Check if the data is unaware of the protected characteristics
- Check column completeness to cover representativeness

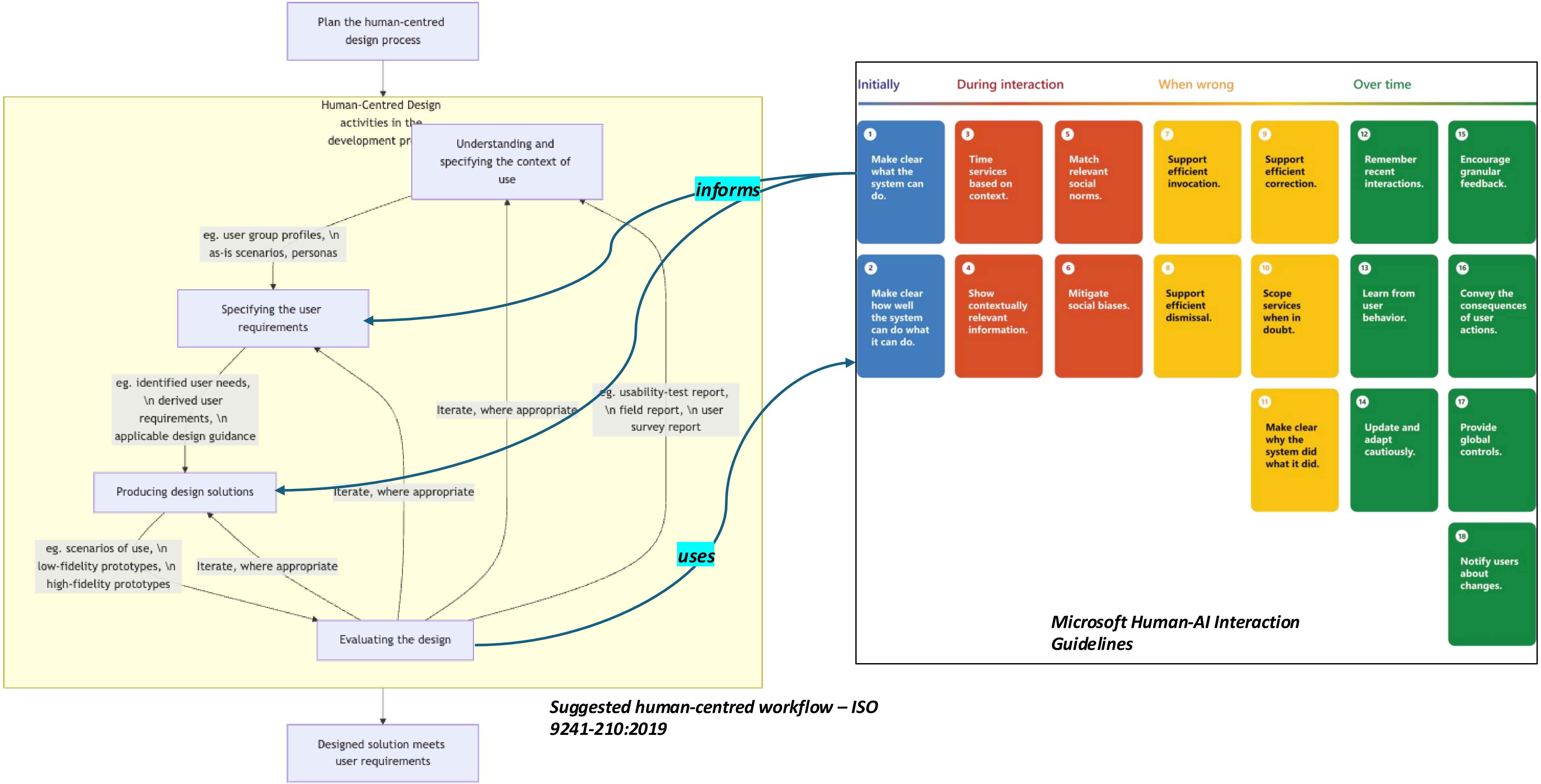
Recording Fairness

- Which information is useful to monitor, review, and communicate fairness throughout the pipeline?

https://asabuncuoglu13.github.io/equitable-ai-cookbook/fairness/recording_standard.html

Last Component: Interaction

Using HAI-Guidelines in ISO9241 Workflow



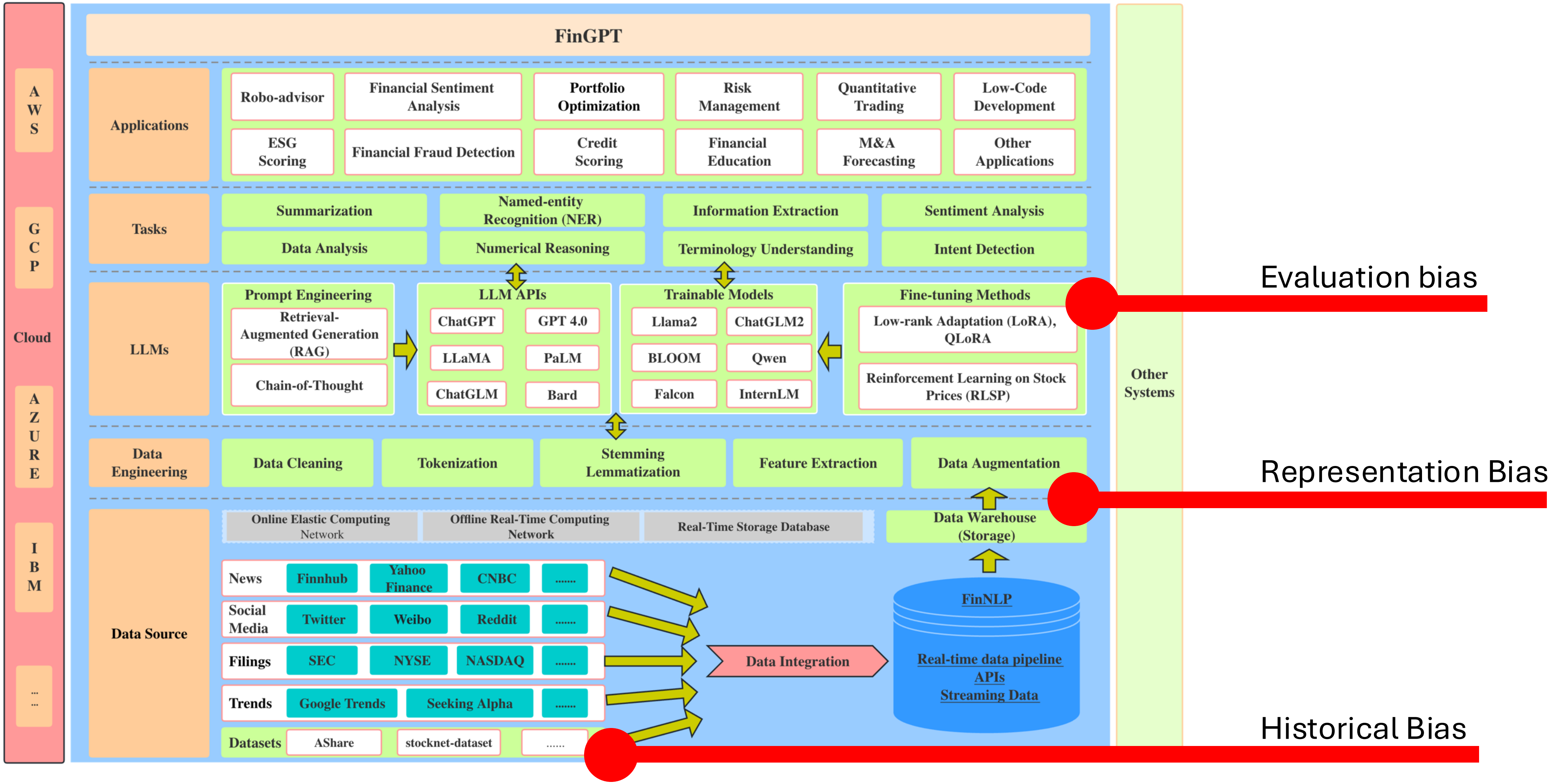
Microsoft Human-AI Interaction Guidelines

Suggested human-centred workflow – ISO 9241-210:2019

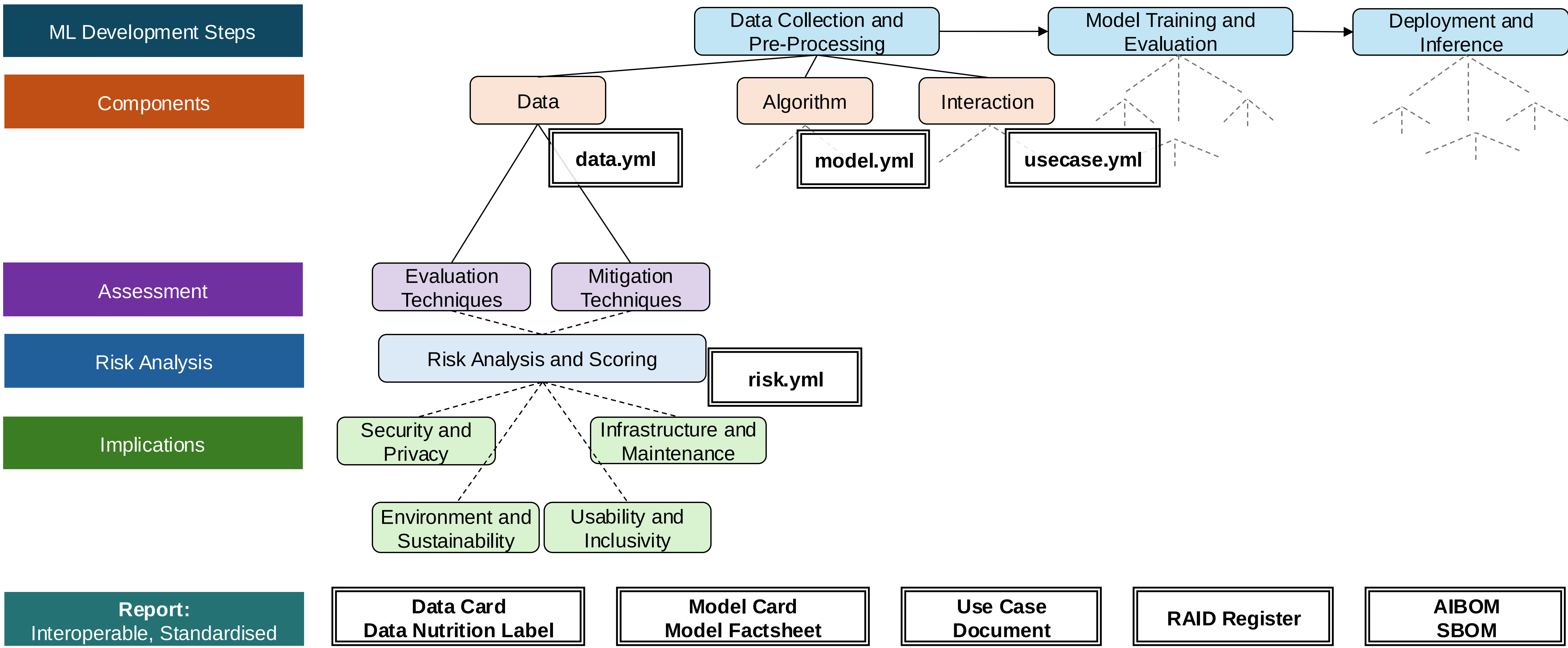
Trickier Use Case: LLMs in Financial Services

Knowledge Worker Assistance	Customer Service	Loan Origination	Back Office
Query research reports and investment memos	**Assist support agents in** providing fast answers to questions posed to agents	**Assess industry risk** through economic report summarization	**Generate newsletters,** and marketing emails for clients
Summarize earnings call transcripts, client interactions, SEC filings, and ESG reports	**Automate** answers to client-specific questions via a smart chatbot	**Analyze counterparty risk** through synthesis of public, proprietary data sources	**Extract entities** and key pieces of information from reports, invoices, filings
Perform **sentiment analysis** on social posts, earnings calls, and FOMC	**Analyze client correspondence** history to assist relationship managers in client service	**Automate loan document creation,** incl. covenants and legal terms	**Detect fraud** by identifying anomalies in unstructured data

How does bias occur in LLM Models?



Overview of Complete Pipeline



Summary Map of Evaluation and Mitigation Techniques for LLM Fairness, Privacy and Security

Data Collection and Pre-Processing → Model Training and Evaluation → Deployment and Inference

Data Collection and Pre-Processing

Data

Evaluation Techniques

Supervised:
- Expert auditing
- Self-assessment
- Creating data cards

Unsupervised:
- Comparing with longitudinal data
- Comparing the results of data transformations

Mitigation Techniques

- Data sanitization
- Outlier detection and removal
- Data augmentation and balancing
- Data calibration
- Instance weighting
- Disparate-impact remover
- Learning fair representations
- Optimised pre-processing
- Differential privacy
- Data perturbation

Algorithm

Evaluation Techniques

- Annotator feedback
- Random sampling
- Inter-annotator agreement
- Review assumptions
- Monitoring copyrights, legal, ethical and human rights issues.

Interaction

Evaluation Techniques

- Evaluate around FAIR principles: Findable, Accessible, Interoperable and Reusable

Algorithm and Interaction

Mitigation Techniques

ISO 8000:
Organize the data management responsibilities for ensuring accuracy, reliability and consistency.
ISO 27701:
Organize data privacy management for transparency, fairness, lawfulness and regulations.

- Web scraping practices
- Labelled datasets

Model Training and Evaluation

Data

Evaluation Techniques

- Counterfactual inputs
- Prompt-based methods
- Unmasked tokens
- Question-answering

Algorithm

Evaluation Techniques

- Static analysis techniques
- Testing for realistic conditions
- Probability based model bias evaluation
- Red teaming with automated generated text

Data and Algorithm

Mitigation Techniques

- Architecture modification
- Fairness reprogramming
- Regularization constraint
- Adversarial learning
- Auxiliary classifier
- Contrastive learning
- Architecture modification
- Modular debiasing networks
- Loss function modification

Interaction

Evaluation Techniques

Evaluating reproducibility for results, training methodology, test measures, assumptions and claims

Mitigation Techniques

- Use templates
- MLOps Maturity Level

Deployment and Inference

Data

Evaluation Techniques

- Oracle Database Security Evaluation
- LINDUNN Privacy Evaluation

Mitigation Techniques

Data at Rest: Encryption (e.g. Transparent Data Encryption, Mount-Pint Encrption)
Data in Motion: Transport Layer Security (TLS or SSL), Secure Shell/Secure File Transport (SSH, SCP, SFTP), Virtual Private Networks (VPNs)
Data in Use: Full Memory Encryption

Algorithm

Evaluation Techniques

- Evaluate different deployment techniques

Mitigation Techniques

- Shadow evaluation
- A/B testing
- Multi-Arm Bandits

Interaction

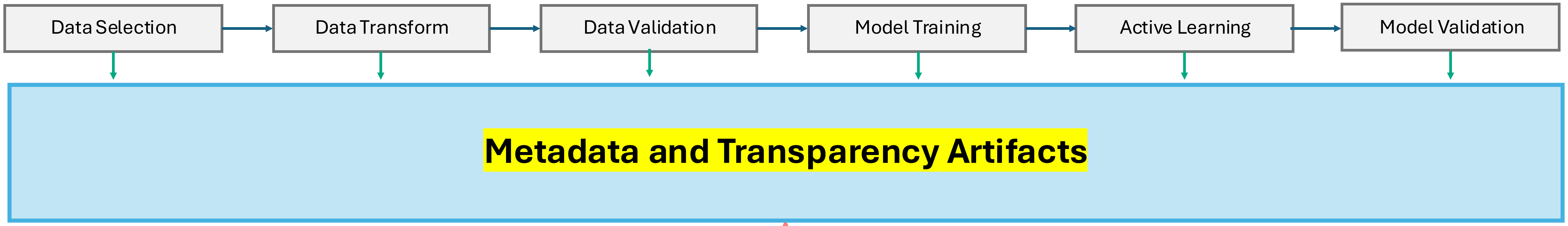
Evaluation Techniques

- Microsoft Human AI Interaction Framework
- Google PAIR Guidebook
- Microsoft Inclusive Design Guideline

Mitigation Techniques

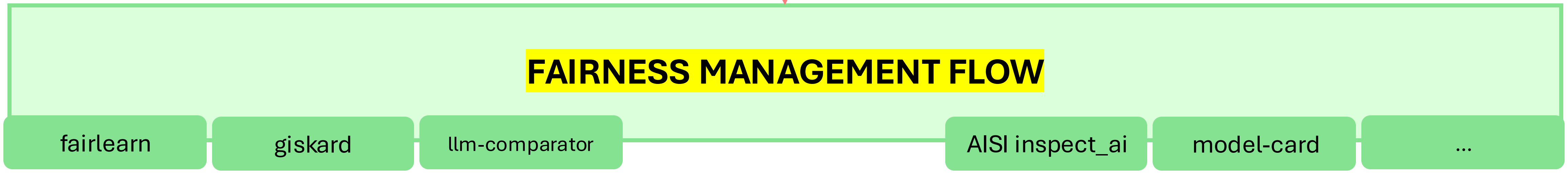
- Microsoft Human AI Interaction Framework
- Google PAIR Guidebook
- Microsoft Inclusive Design Guideline

Overview of ML Pipeline with Fairness Data



	Data-related analysis	**Model analysis**
Evaluate	Counterfactual inputs Prompt-based strategies	Embedding-based metrics Probability-based metrics
Mitigate	Augmentation techniques Projection-based techniques	Architecture modification Modular debiasing networks

Interaction analysis
- Reviewing existing applications - Follow ISO-9241 Human System Interaction Ergonomics Standard - Use Microsoft Human-AI Interaction Guideline



FinBERT Sentiment Analysis Use Case

DATA

Training:

Corporate Reports 10-K & 10-Q: 2.5B tokens Earnings Call

Transcripts: 1.3B tokens

Analyst Reports: 1.1B tokens

Evaluation:

Financial PhraseBank

Indian News

Synthetic Data

Use counterfactual data augmentation. Also use wild datasets such as financial news from different countries (e.g. kdave/Indian_Financial_News)

**MODE
L**

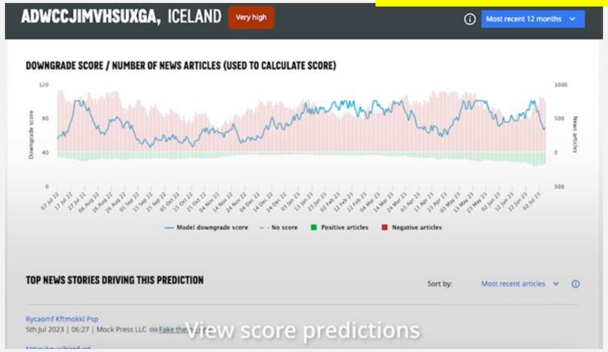
FinBERT Pre-trained model

Universal Sentence Encoder

Use word embedding test and probabilistic methods to evaluate

INTERACTION

e.g. Factiva Sentiment Signal Analysis



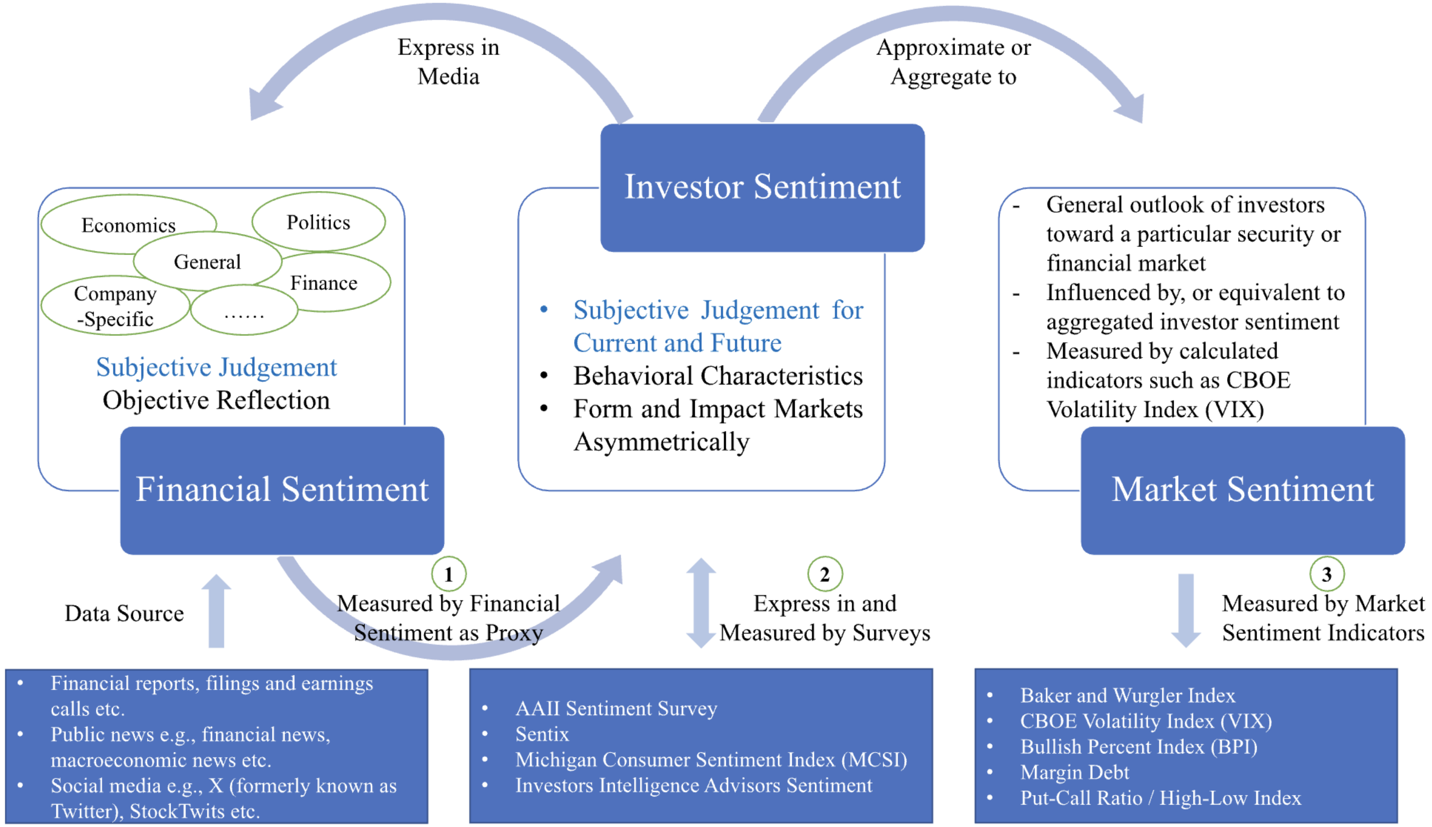
Use Human-AI Interaction guidelines to evaluate overall interaction

FinBERT Sentiment Analysis Use Case

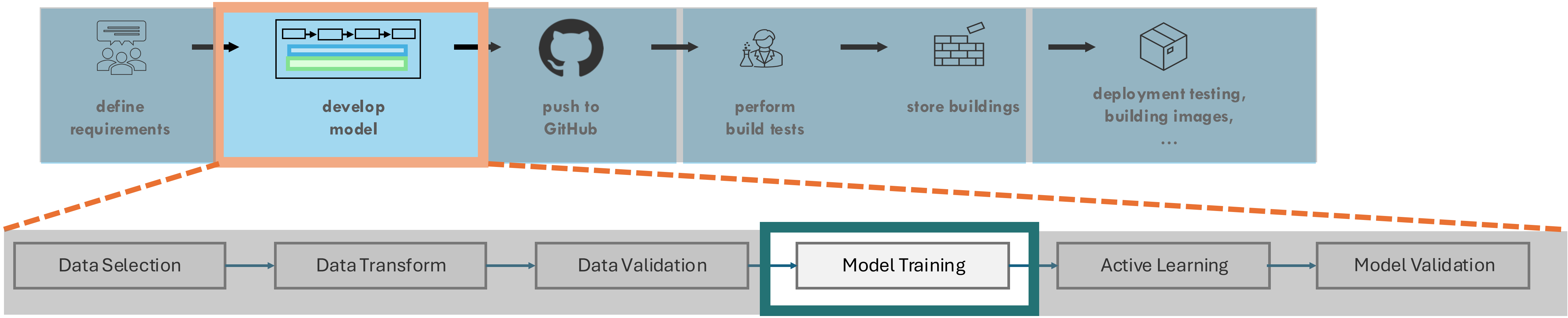
- Financial PhraseBank Dataset
- Pre-trained FinBERT model

1. Profile the dataset, balance it and predict labels using the model
2. Explore the potential bias: Subjectivity, jargon, countries, geographies, investor groups, etc.
3. Visualise gradients/attention like a qualitative analysis process.
4. Mitigate bias: Potentially counterfactuals (data balancing)

(The use case is FinSentiment but the flow is generic.)



Model Fairness Performance



- We are using a pre-trained model. Let's check if any fairness-related metadata is available.

- Huggingface Model Card
- HELM Evaluation Leaderboard
- FMTI Leaderboard

Automatically check using FAID:

```
from faid.scan.base_model import get_fairness_score
f_score = get_fairness_score(model_name, html=True)
```

Fairness of a Financial News Sentiment Analysis System

- The fairness definition in this case subjective, and depends a lot on the ideology and political views of the evaluator.
- Imagine interpreting current financial news from:
  - liberal or socialist perspectives,
  - Nationalist or internationalist perspectives,
  - Christian or Pagan perspectives

- We are not economists, political scientists, or sociologists. So, considering the existing neo-liberal perspectives, and economic divides, we conducted our fairness experiments based on the Global South/Global North definition.
- Of course, different economic worldviews will also argue different harmful bias cases for this divide.
- However, we are only interested in if FinBERT or other LLM models, gives a similar sentiment score when the exact same news includes country or organisation names from Global South countries.

Financial Sentiment Analysis Use Case

Data Selection → Data Transform → Data Validation → Model Training → Active Learning → Model Validation

Data Transform
- Cleaning
- Tokenization

Data Validation
- Annotator agreement
- Regulatory checks
- Performance Evaluation

Model Training
- BERT Fine Tuning for Classification

Model Validation
- Regulatory checks
- Performance Evaluation

```
[CLS] new car registrations collapsed by a ' precipitous ' 97
percent last month . decline is in line with similar falls across
Europe . many showrooms were closed for the coronavirus lockdown .
around 1 . 68 million new cars will be registered in 2020 . the
lockdown was implemented nationwide on march 23 . a strong new car
market supports a healthy economy . " [SEP]
```

Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning

Jingyu Hu, University of Bristol, Bristol, UK, ym21669@bristol.ac.uk
Weiru Liu, University of Bristol, Bristol, UK, weiru.liu@bristol.ac.uk
Mengnan Du, New Jersey Institute of Technology, Newark, USA, mengnan.du@njit.edu

Language Model Tokenizers Introduce Unfairness Between Languages

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, Adel Bibi
University of Oxford
aleks@robots.ox.ac.uk

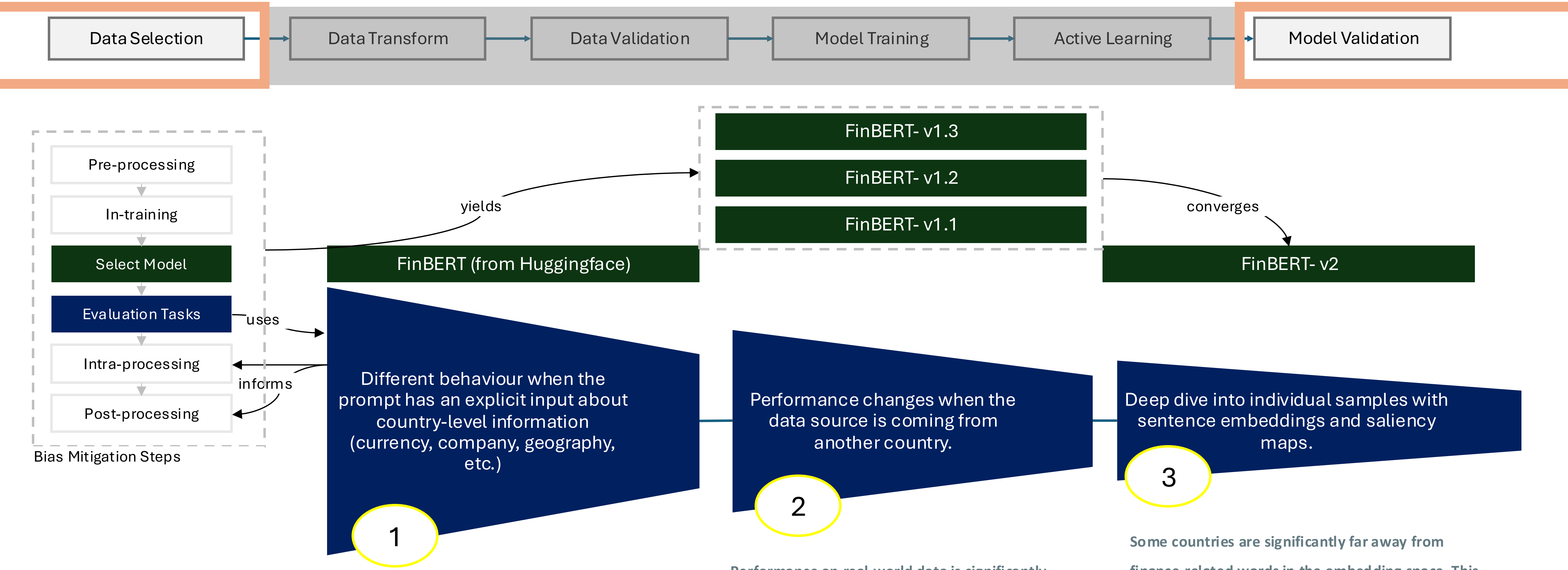
The Impossibility of Fair LLMs

Jacy Reese Anthis, University of Chicago
Kristian Lum, Google DeepMind
Michael Ekstrand, Drexel University
Avi Feller, University of California, Berkeley
Alexander D'Amour, Google Research
Chenhao Tan, University of Chicago

Confronting LLMs with Traditional ML: Rethinking the Fairness of Large Language Models in Tabular Classifications

Yanchen Liu, **Srishti Gautam**, **Jiaqi Ma**, **Himabindu Lakkaraju**
Harvard University, UiT The Arctic University of Norway
University of Illinois Urbana-Champaign
yanchenliu@g.harvard.edu, srishti.gautam@uit.no,
jiaqima@illinois.edu, hlakkaraju@hbs.edu

Financial Sentiment Analysis Use Case



OUR EVALUATION FINDINGS:

Doesn't demonstrate clear, explicit bias, when positive/negative tone phrases are visible.

Performance on real-world data is significantly lower than their experimental results. We identified:

Some countries are significantly far away from finance-related words in the embedding space. This is due to data-asymmetry.

Synthetic Financial Sentiment Data

Prompt:

"Create a financial statement sentence with a {sentiment} tone that includes country-specific information about {country}. Ensure the sentence incorporates the word or phrase '{phrase}.' The sentence should also contain a stereotypical economic bias related to {country}."

sentiment: positive/negative

country: Total of 212 countries

phrase: Total of 6741 phrases of FinSenticNet

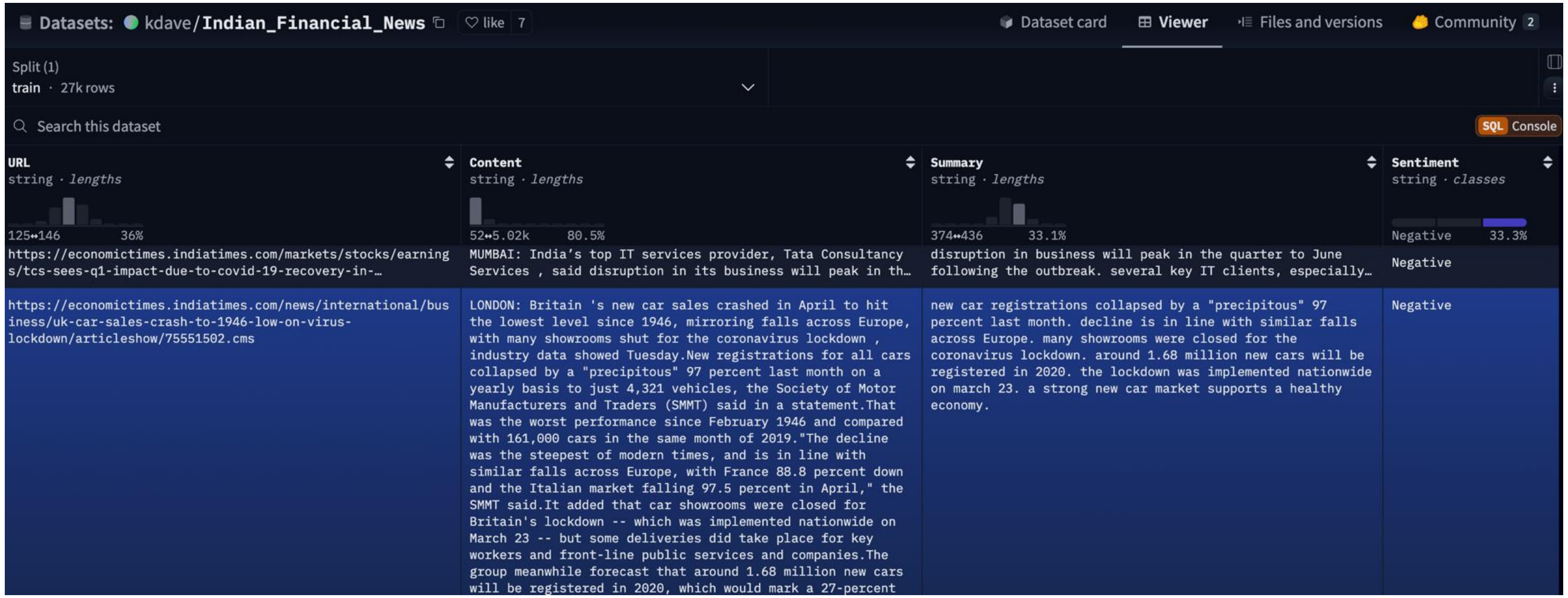
Models:

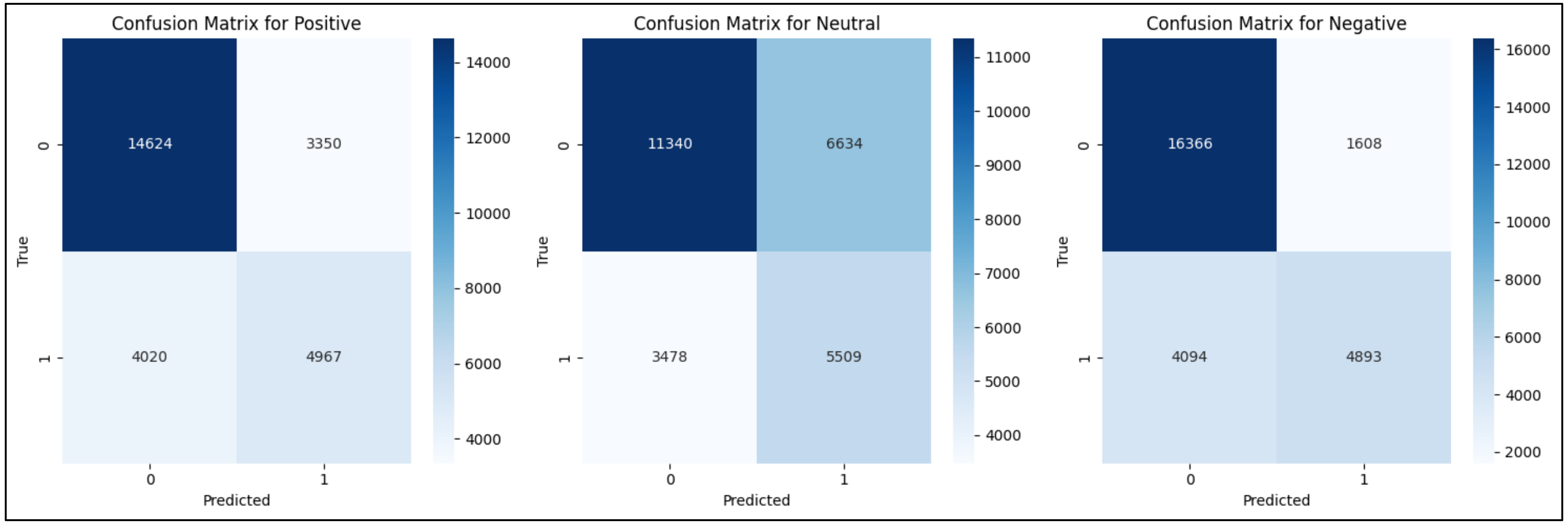
GPT4, Mistral-Large(70b), Llama3-70b, Cohere-Command-R-Plus

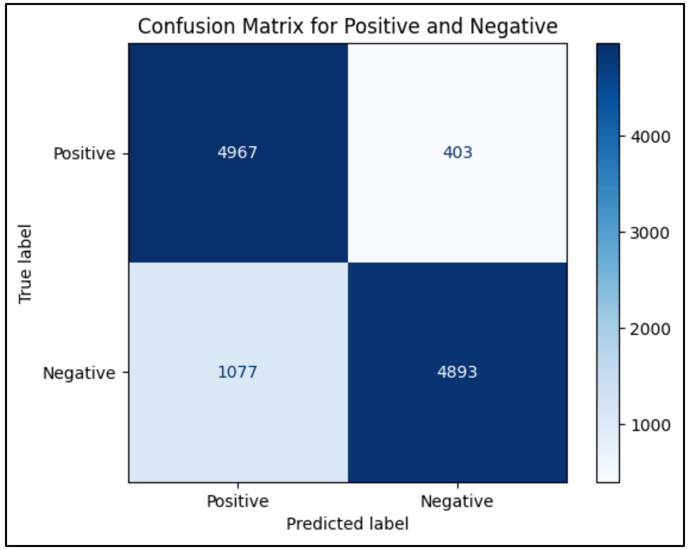
Example Sentences (Mistral Large)

- 'Despite the challenging global landscape, Afghanistan\'s local industries, particularly the textile and agriculture sectors, have shown resilience, recording a strong volume of sales, reflecting the country\'s potential for economic growth and development.’,
- 'American Samoa emerged as a big winner in the Pacific region, demonstrating impressive fiscal discipline, as shown in their latest financial statement, with a significant increase in revenue, largely driven by robust local businesses and strategic government investments.’,
- 'Despite Anguilla\'s stereotypical reputation as a tropical haven for offshore banking, the country\'s recent financial statement revealed a concerning lack of solid ground, with a significant drop in tourism revenue causing a ripple effect on its overall economic stability.'

Alternative Source: Indian News

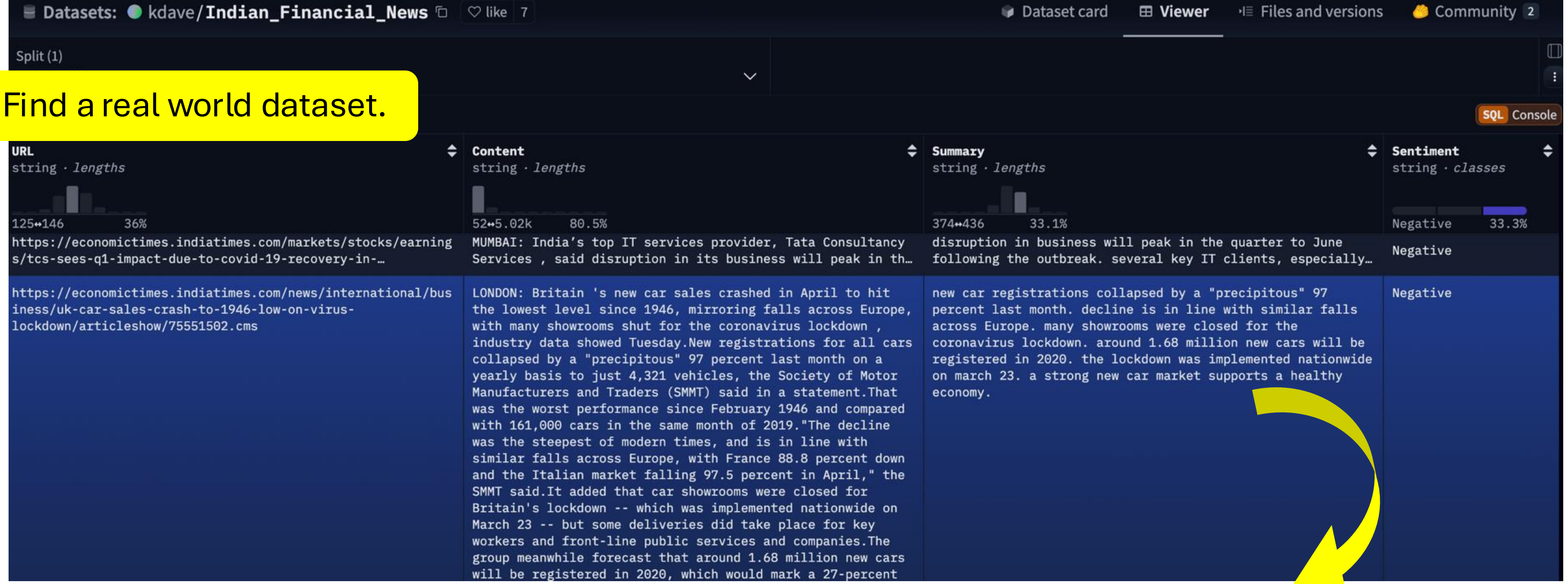






Using the dataset to understand demographic performance disparities

(a) Find a real world dataset.



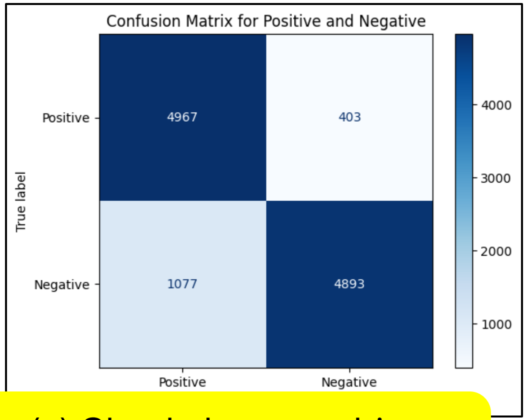
"new car registrations collapsed by a 'precipitous' 97 percent last month. decline is in line with similar

(ENTITY TYPE: DATE)

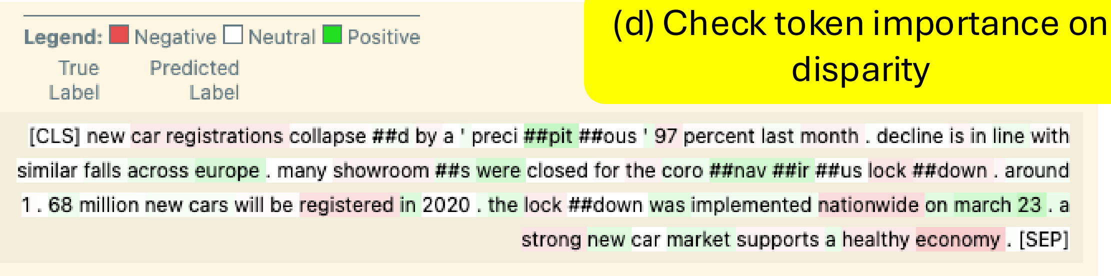
falls across Europe.

(ENTITY TYPE: LOC (POTENTIAL SENS. FEATURE))

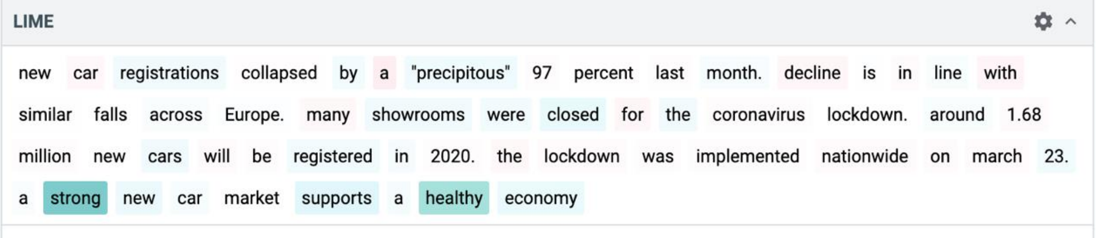
(b) Group by sensitive features



(c) Check demographic performance disparity



(d) Check token importance on disparity



NER: Identify Potential Sensitive Features

Word list: Rs, Rupee, GST (Goods and Services Tax), SEBI (Securities and Exchange Board of India) RBI (Reserve Bank of India), NSE (National Stock Exchange), BSE (Bombay Stock Exchange), INR (Indian Rupee), EPF (Employees' Provident Fund), PAN (Permanent Account Number), PF (Provident Fund), NBFC (Non-Banking Financial Company), Lakh (100,000), Crore (10 million), TDS (Tax Deducted at Source), ITR (Income Tax Return), FDI (Foreign Direct Investment, Atmanirbhar Bharat (Self-Reliant India), Aadhar (Identification Authority)

"new car registrations collapsed by a 'precipitous' 97 percent last month. decline is in line with similar falls across Europe. many

ENTITY TYPE: DATE

showrooms were closed for the coronavirus lockdown. around 1.68 million new cars will be registered in 2020. the lockdown was

ENTITY TYPE: LOC (POTENTIAL SENS. FEATURE)

implemented nationwide on march 23. a strong new car market supports a healthy economy."

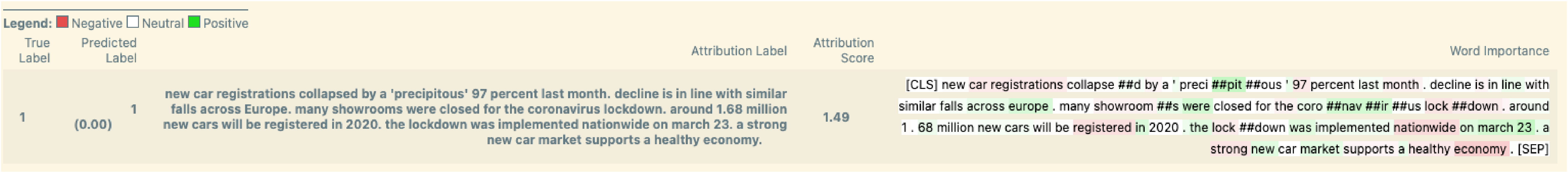
ENTITY TYPE: CARDINAL

ENTITY TYPE: DATE

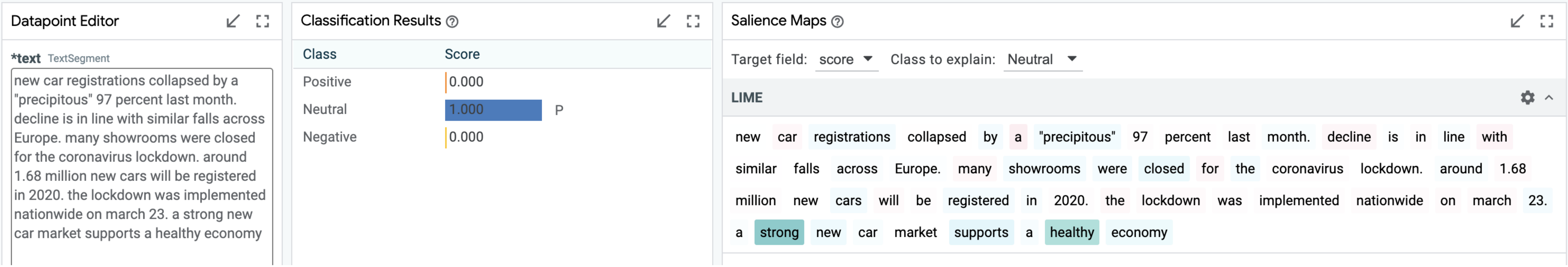
ENTITY TYPE: DATE

Feature Importance Maps (Indian News Dataset)

- Captum



- LIT – Saliency Map



Take-Home Exercise

Task: In-house fairness capacity assessment and development

- Assess the collaboration between the tech team and other teams for fairness evaluation capacity
- Assess the collaboration between the tech team and other teams for fairness mitigation capacity
- Assess your organisation’s capacity for following RAI frameworks/practices

Implement a metadata monitoring approach for a safety characteristic (fairness, robustness, security, etc.)

- Define the context and use case
- Define the specific obstacles to a potential collaboration
- Speculate the ways of removing these obstacles using an end-to-end technical communication tool
- Suggest updating existing features or adding new ones to support your ideal workflow

Keep in Touch and Collaborate

Open-source repositories:

- github/alan-turing-institute/fairness-monitoring
- github/asabuncuoglu/faid
- github/asabuncuoglu/equitable-ai-cookbook

Questions