AEQUITAS
unbias AI

WORKSHOP – HANDS on THE SYSTEM

The AEQUITAS System: work so far

3 June 2024

Funded by
the European Union

Agenda and goals

14.00-14.10	Welcome Agenda Review	
14.10-15.00	DEMO AEQUITAS SYSTEM	Giovanni Ciatto (UNIBO)
15.00-15.30	USE CASES: where are we?	Andrea Borghesi/ Joseph Giovanelli (UNIBO) Paul Lemmens (PRE)
15.30-15.45	METHODOLOGY: where are we?	Andrea Borghesi(UNIBO) Catelijne Muller (ALLAI)
15.45-16.15	Feedback and open discussion	
16.15-16.30	Closing	

- Status & Feedback
- Plan: 20 June 24

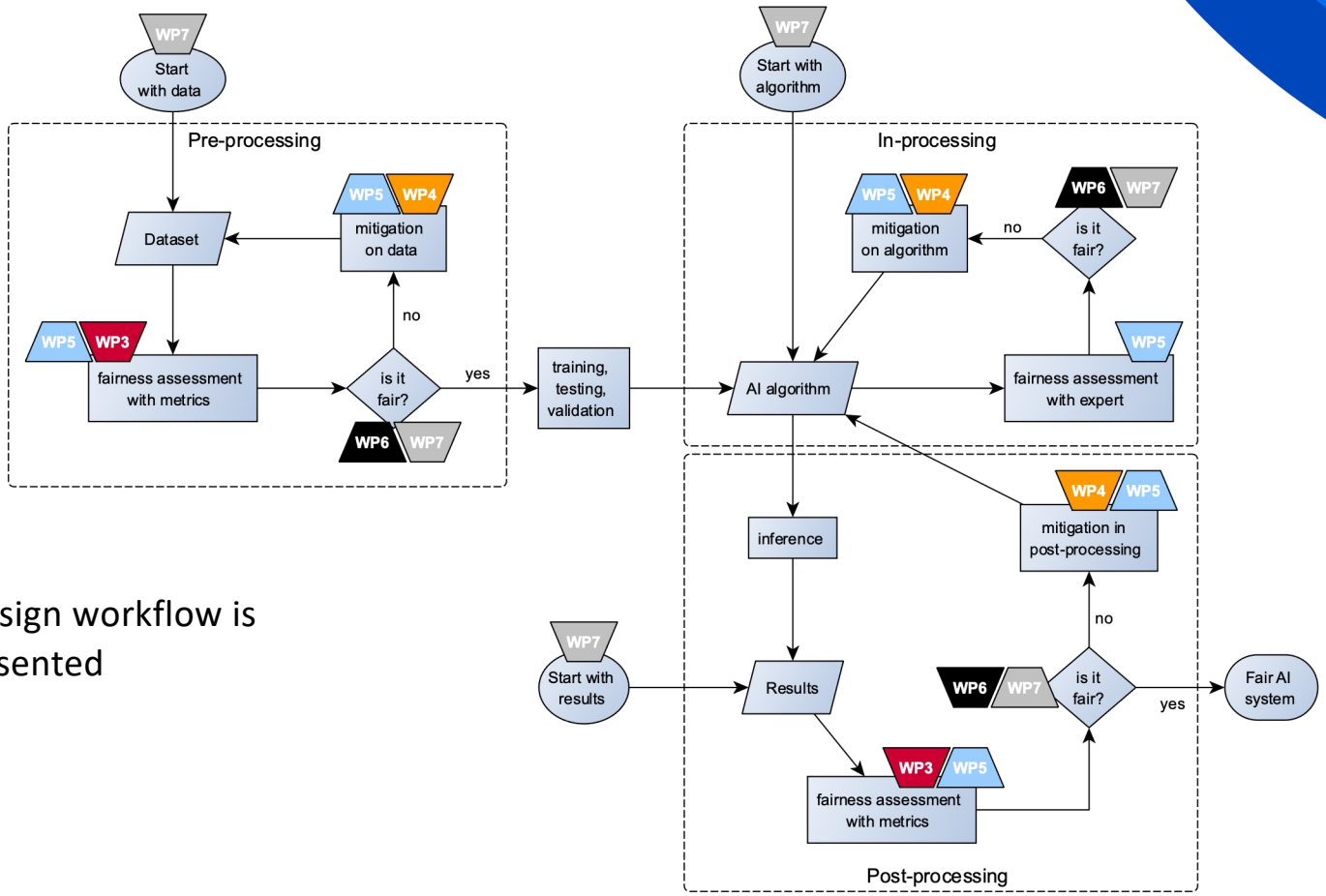
Funded by
the European Union

AEQUITAS Detection-Mitigation Workflow

A workflow for Fair AI

Funded by
the European Union

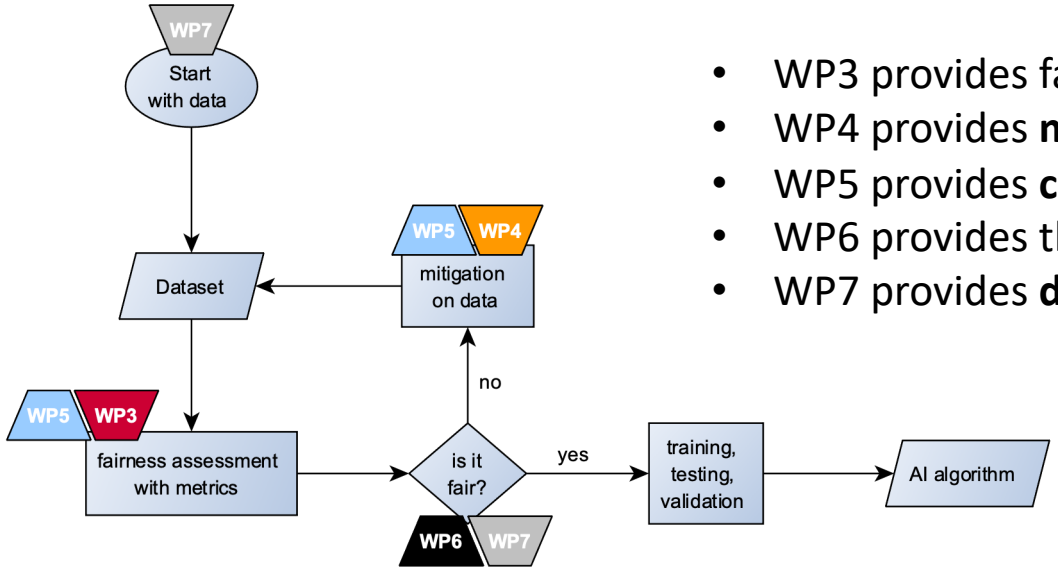
AEQUITAS Detection-Mitigation Workflow Overview



- fair-by-design workflow is not represented

Funded by the European Union

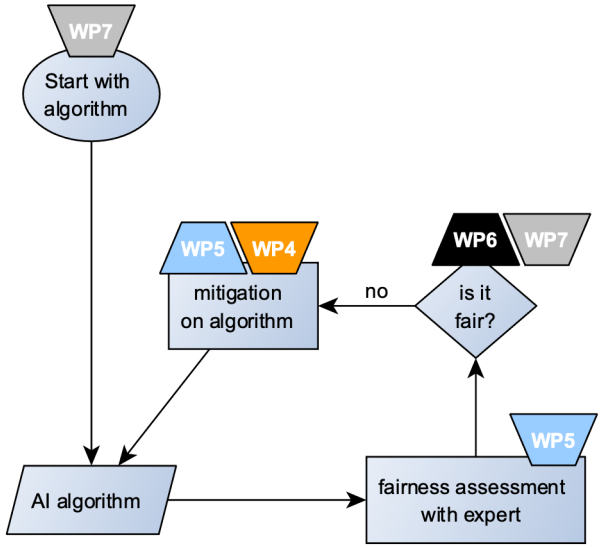
AEQUITAS Workflow
Details on **pre**-processing



- WP3 provides fairness **metrics** for **data**
- WP4 provides **mitigation algorithms** for **data**
- WP5 provides **criteria** to **select** metrics / algorithms w.r.t. domains
- WP6 provides the **socio-legal** criteria / **constraints**, fairness notions
- WP7 provides **data** (real or synthetic) and **use-cases**

Funded by the European Union

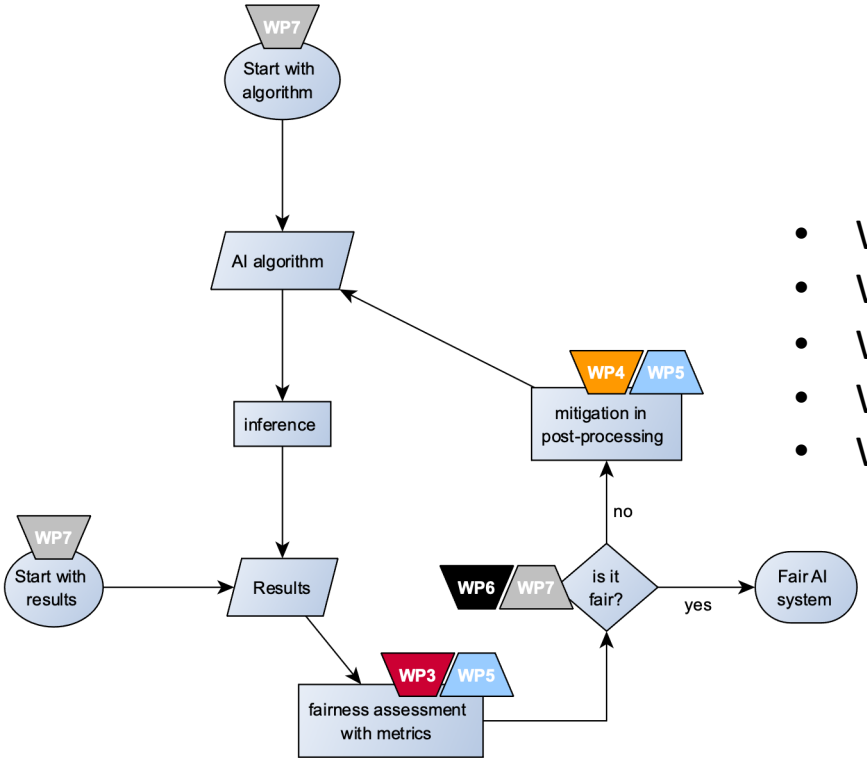
AEQUITAS Workflow
Details on **in**-processing



- WP5 provides **fairness-assessment** criteria for **AI algorithms**
- WP4 provides **mitigation algorithms** for **AI algorithms**
- WP5 provides **criteria** to **select** mitigation algorithms
- WP6 provides the **socio-legal** criteria / **constraints**, fairness notion
- WP7 provides **data** (real or synthetic) and **use-cases**

Funded by the European Union

AEQUITAS Workflow
Details on **post**-processing



- WP3 provides fairness **metrics** for **data**
- WP4 provides **mitigation algorithms** for **post-processing**
- WP5 provides **criteria** to **select** metrics / algorithms w.r.t. domains
- WP6 provides the **socio-legal** criteria / **constraints**, fairness notion
- WP7 provides **data** (real or synthetic) and **use-cases**

Funded by the European Union

AEQUITAS Software

A software system supporting the workflow

Funded by
the European Union

AEQUITAS Technology Desiderata

1. The software **should help** users/**organizations** in
   - **creating** fair AI (*datasets, algorithms, result interpretation)*
   - **making AI** algorithms, datasets, or predictions **fair**
   - be(come) **compliant** w.r.t. **regulations** and social **conventions**

2. The software should **not**:
   - **replace** humans in decision-making
   - Awareness for legal or ethical compliance

Funded by
the European Union

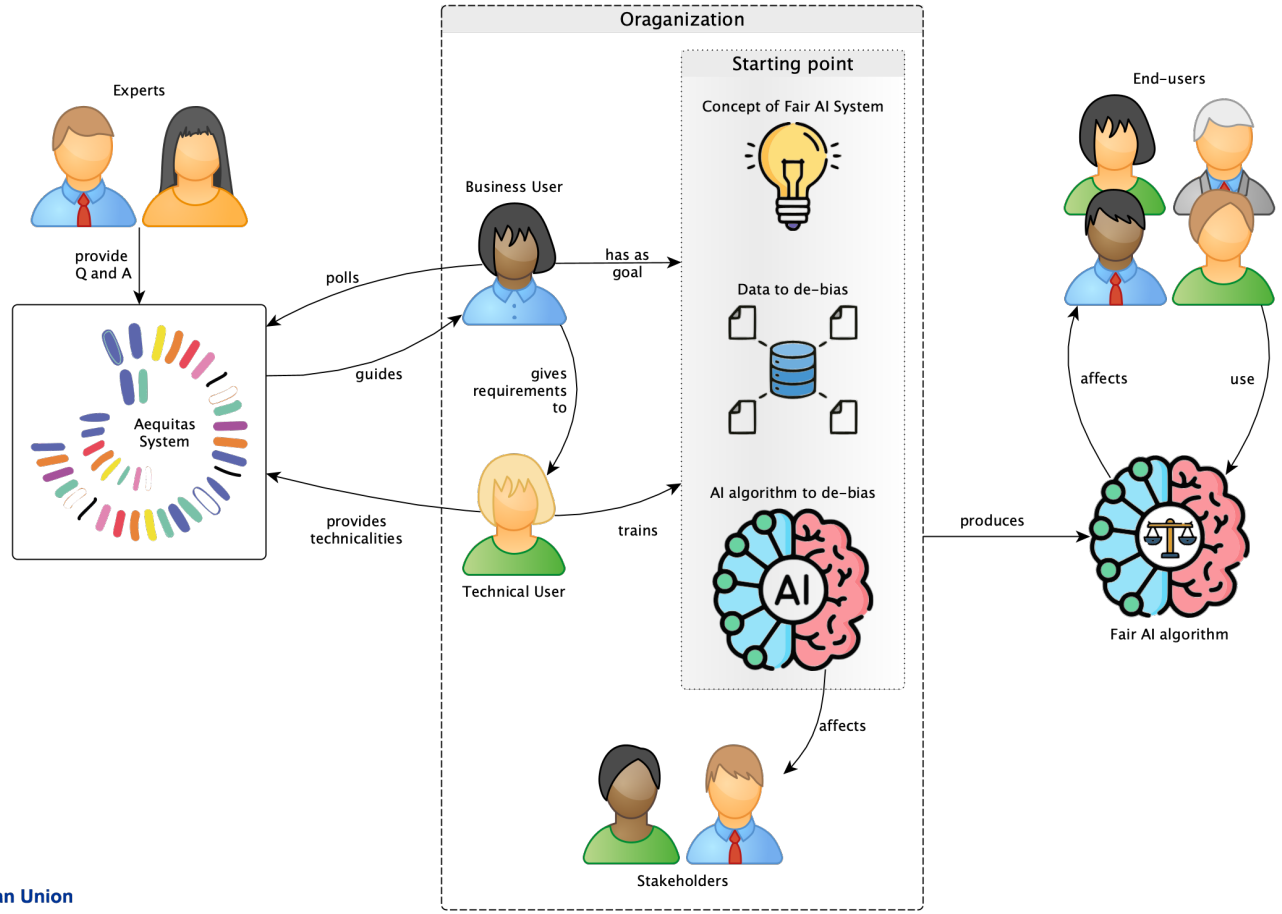
AEQUITAS Technology
Key insights

1. Two sorts of **personas**, modelling the **end-user**:
   a. **Business User (BU):** responsible for decision-making, socio-technical background
   b. **Data Scientist (DS):** technical background in fair AI, coding, data manipulation/interpretation

2. Need for **2 user interfaces**, supporting the **interaction** among BU and DS:
   a. AEQUITAS **supports BU** in **eliciting** fairness **requirements/constraints**
      - via Q/A
   b. AEQUITAS **mediates cooperation** among BU and DS
      - via backend
   c. DS **provides** data / **information** to enable BU's **decision-making**
      - computed via ad-hoc software
   d. AEQUITAS **steers BU**'s decision-making
      - via Q/A

Funded by
the European Union

AEQUITAS Technology Concept



Funded by the European Union

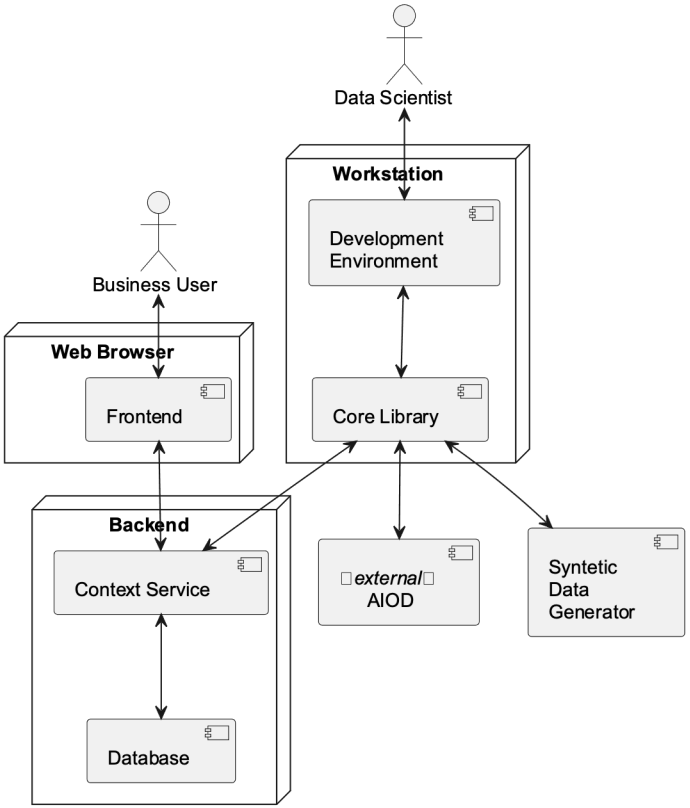
AEQUITAS Technology
Technical hints

1. A **Web-app**, supporting decision-makers...

2. … backed by several, individually-**reusable tools**
   - a **software library** of detection/mitigation algorithms
   - a synthetic **data generator**

Funded by
the European Union

AEQUITAS Technology Architecture



1. BU uses friendly, in-browser **frontend**
   - mostly, Q/A and reports
2. DS uses whatever **development env** they prefer
   - e.g. Jupyter Notebooks, IDE
3. **Context service** mediates interaction among the 2
   - of course, has a database
4. **Core-library** contains algorithmic facilities
   - for mitigation, detection, fair-by-design
   - ideally, reusable à la SciKit-Learn
5. Synthetic **Data Generator** supports many mitigation algorithms

Funded by the European Union

AEQUITAS Technology
Core Library – List of features

https://github.com/aequitas-aod

Goal	Category	Subcategory	Name
detection	metric	group	disparate_impact_ratio
detection	metric	group	equal_opportunity_difference
detection	metric	group	average_odds_difference
detection	metric	group	average_odds_error
detection	metric	group	class_imbalance
detection	metric	group	kl_divergence
detection	metric	group	conditional_demographic_disparity
detection	metric	group	smoothed_edf
detection	metric	group	df_bias_amplification
detection	metric	group	between_group_generalized_entropy_error
detection	metric	group	mdss_bias_scan
detection	metric	group	mdss_bias_score
detection	metric	individual	generalized_entropy_index
detection	metric	individual	generalized_entropy_error
detection	metric	individual	theil_index
detection	metric	individual	coefficient_of_variation
detection	metric	individual	consistency_score
detection	detector	-	bias_scan
mitigation	pre-processing	-	DisparateImpactRemover
mitigation	pre-processing	-	LFR
mitigation	pre-processing	-	OptimPreproc
mitigation	pre-processing	-	Reweighing
mitigation	in-processing	-	AdversarialDebiasing
mitigation	in-processing	-	ARTClassifier
mitigation	in-processing	-	GerryFairClassifier
mitigation	in-processing	-	MetaFairClassifier
mitigation	in-processing	-	PrejudiceRemover
mitigation	in-processing	-	ExponentiatedGradientReduction
mitigation	in-processing	-	GridSearchReduction
mitigation	post-processing	-	CalibratedEqOddsPostprocessing
mitigation	post-processing	-	EqOddsPostprocessing
mitigation	post-processing	-	RejectOptionClassification

Funded by
the European Union

Exemplifying the software

Assess and repair AI-assisted **recruiting** tool to mitigate bias based on:

- e.g. gender
- e.g. race
- e.g. age
- […]

Funded by
the European Union

Example Scenario

1. **ADECCO** (use case HR1) has a dataset of **job applicants**
   - and wants to **automate** their **selection** via AI, in a fair way

2. A BU from ADECCO **creates a new project** on the AEQUITAS **frontend**

3. The BU must **answer** several **questions**, in order to:
   - **focalise** fairness **requirements** & **constraints**
   - **select sensitive features**, protected groups, etc.
   - choose fairness notion, metrics, mitigation algorithms

4. The BU leaves the stage to the DS, which *actually* performs measurements/mitigation
   - via the **Core Library**

5. Whenever data distributions need to be controlled, **synthetic data generator** is exploited

Funded by
the European Union

About Q/A

GOALS:

- Gathering **contextual information** through a guided questionnaire
- to suggest to the user the most suitable fairness metrics in their context
- and more effective mitigation techniques

Remark:

- questions are designed to help users check for AI fairness from a legal perspective:
  - e.g. in **AI Act, art. 10.2.g**, focuses on identification of data gaps that could result in unfairness.

Hence AEQUITAS will propose:

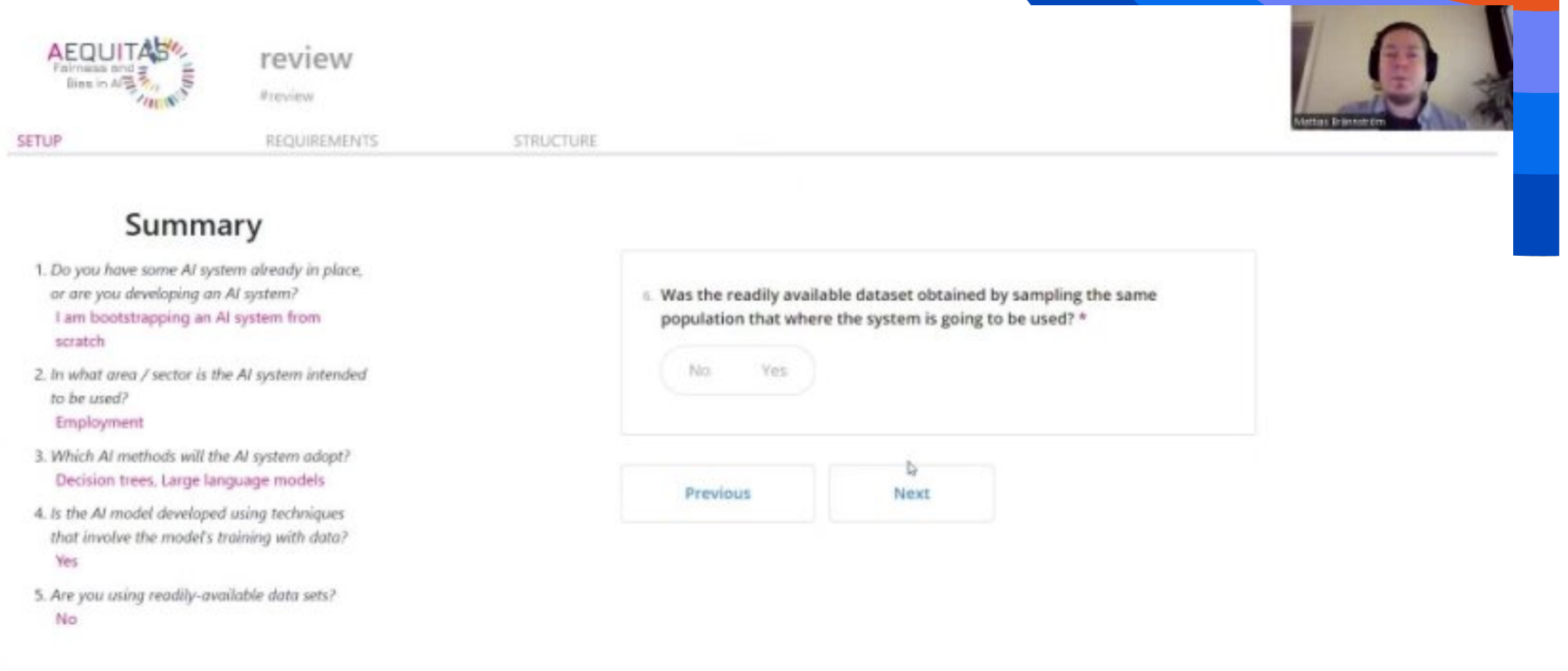
- **metrics** which can score gaps such as under-represented groups
  - (e.g., disparate impact, intersectional fairness)
- **mitigation techniques** for rebalancing/augmenting the data
  - (e.g. resampling, data generation, etc.)

Funded by the European Union

Videos

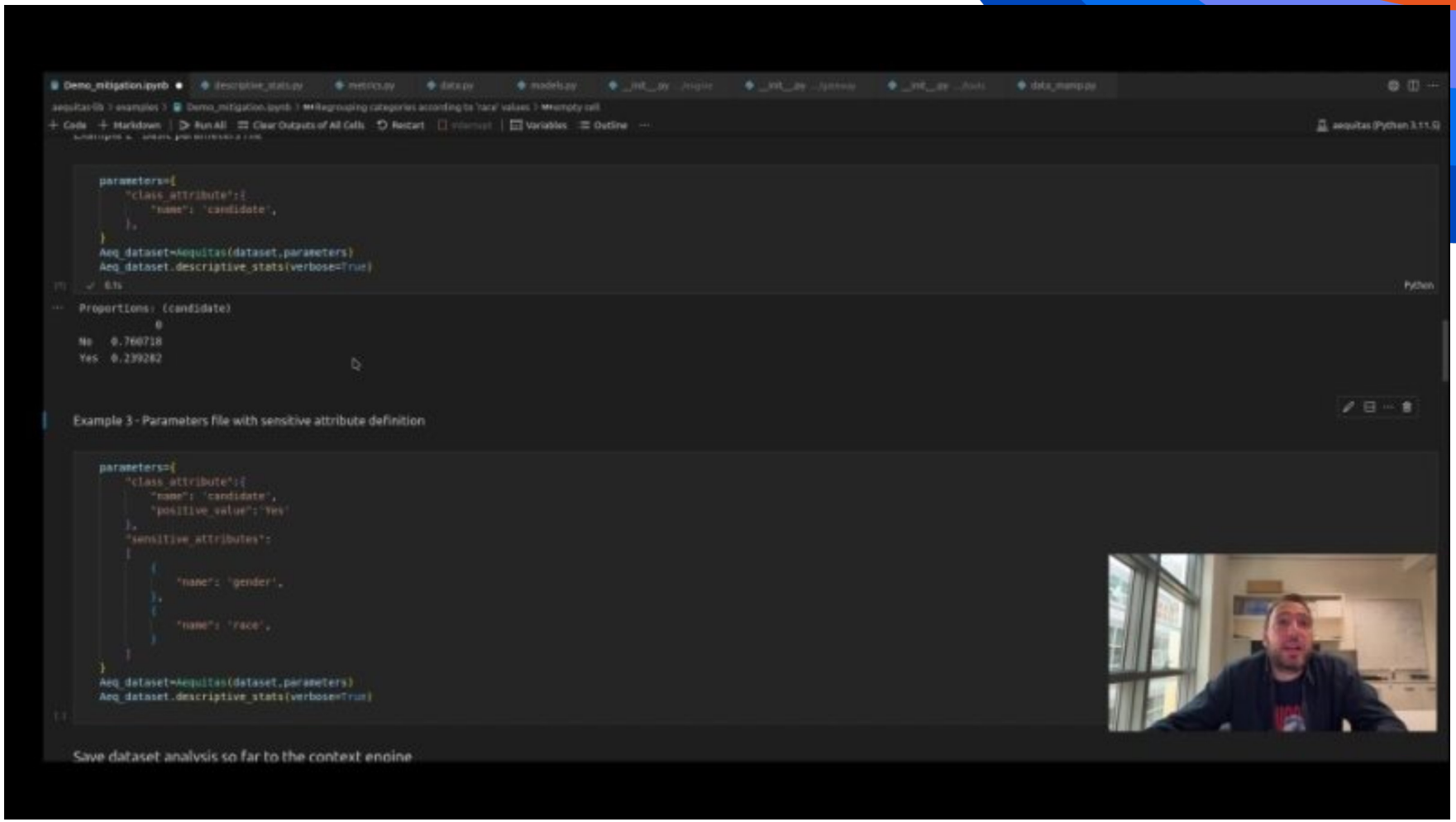
Funded by
the European Union

BU interface (full video here)



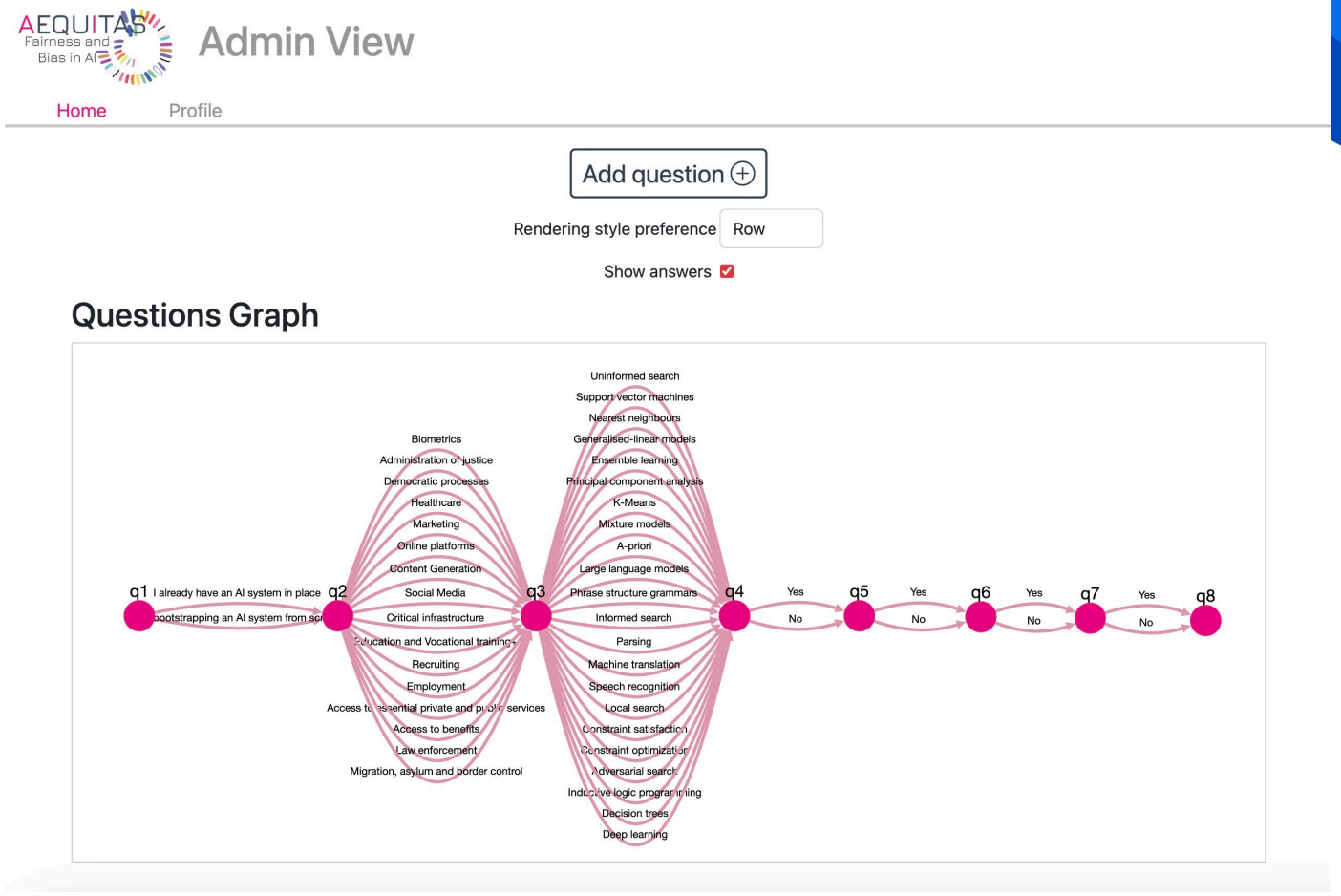
Funded by
the European Union

TU interface (full video here)



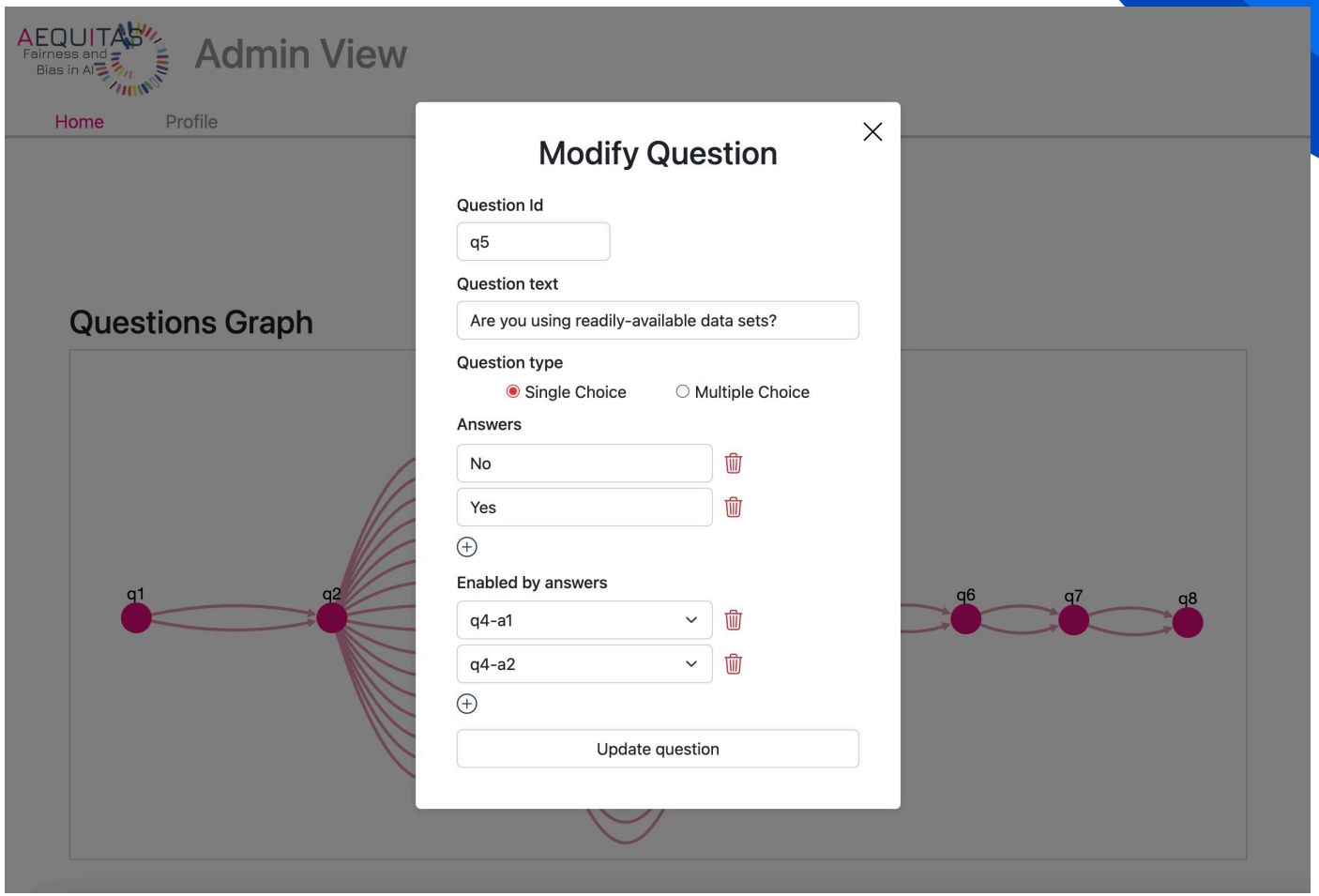
Funded by
the European Union

Admin interface pt. 1
(to let experts add new Q/A)



Funded by the European Union

Admin interface pt. 2
(to let experts add new Q/A)

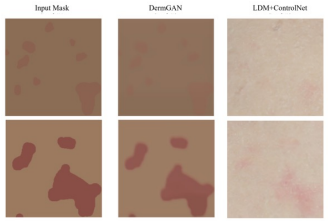
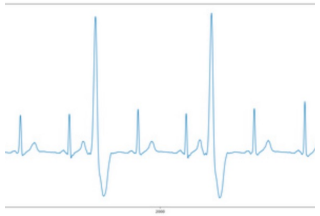
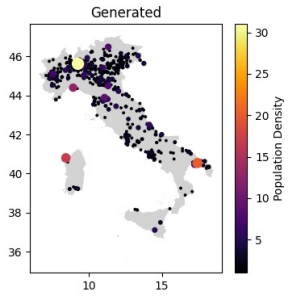
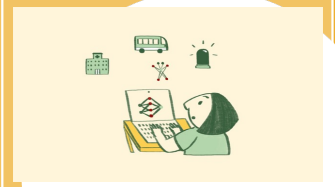


Funded by the European Union

Use cases: status

Funded by
the European Union

Use cases

Use cases					
HC1: dermatology images	HC2: synthetic ECG signals	HR1: job matchmaking dataset	HR2: candidate selection match-making algorithm	S1: child-abuse and neglect	S2: disadvantaged students
					
Lead: UNIBO	Lead: PRE	Lead: PRE	Lead: UNIBO	Lead: UNIBO	Lead: PRE

Funded by the European Union

HC1: dermatology images

- Data cleaned, anonymized, cropped
- Published

https://doi.org/10.6092/unibo/amsacta/7714

S.No	Class	Count
1	maculo-papular rash	1025
2	scabies	1041
3	chickenpox	1026
4	viral exanthema	1056
5	merbilliform rash	1032
6	pediculosis	1026
7	exanthem-polymorphic-like	1044
8	urticaria	1044
9	iatrogenic-drug-induced rash	1026

Funded by
the European Union

HC1: dermatology images: synthetizer

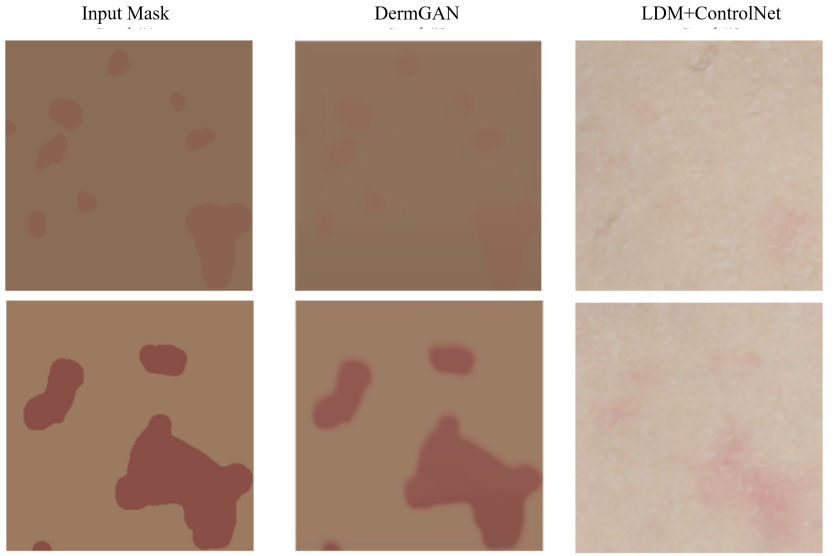
- The initial dataset comprises a collection of **2495 images**, belonging to **187 different medical cases**, captured by doctors using cameras
  - Images taken at the Italian hospital "IRCCS Azienda Ospedaliero Universitaria di Bologna"
  - Photos capture a specific body part related to the patient's condition

- **LDM+ControlNet** to generate synthetic images and expand the training set

Diffusion models + Control (constraints)

Model	FID
DermGAN	311.8
Unconditional LDM	120.7
Controlnet + LDM	74.1

Fréchet Inception Distance -> the lower the better



Synthetic Images, conditioned over the input masks

Funded by the European Union

Borghesi, A., Calegari, R: Generation of Clinical Skin Images with Pathology with Scarce Data (2024), W3PHIA24

HC1: dermatology images: polarized synthesizer (bias generation)

- We started experimenting with ***text-to-image diffusion models***
- The goal was to directly attack the bias in the training data
  - ➢ Generate synthetic skin disease images ***conditioned on the color of the skin***
- We started by improving the image segmentation to obtain better masks:

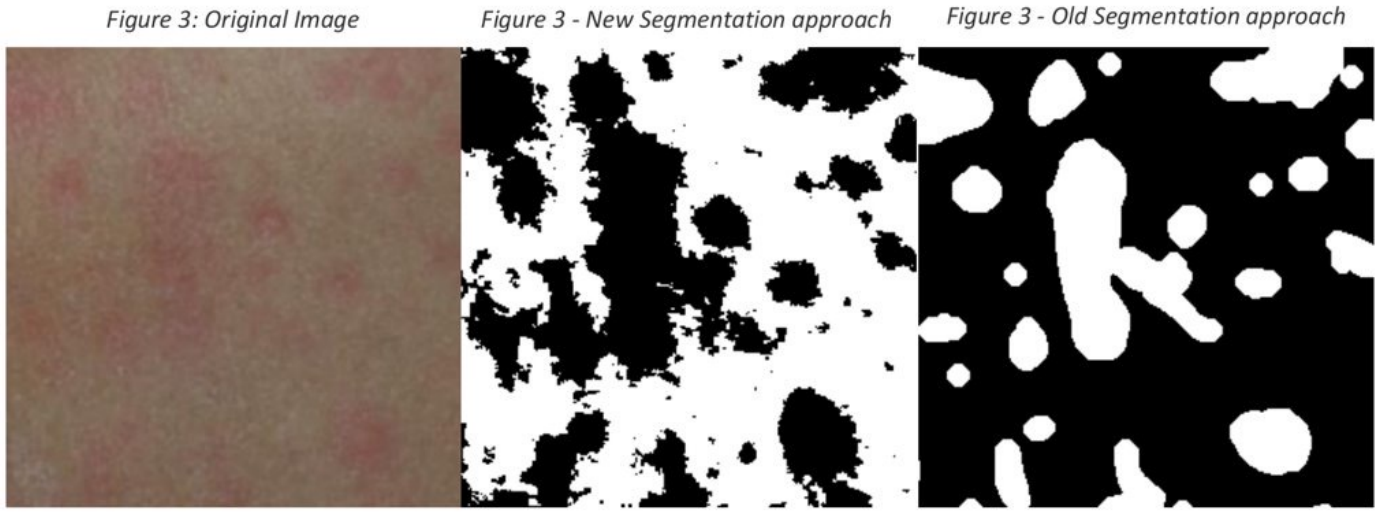


Figure 3: Original Image *Figure 3 - New Segmentation approach* *Figure 3 - Old Segmentation approach*

Funded by
the European Union

HC1: dermatology images: prediction

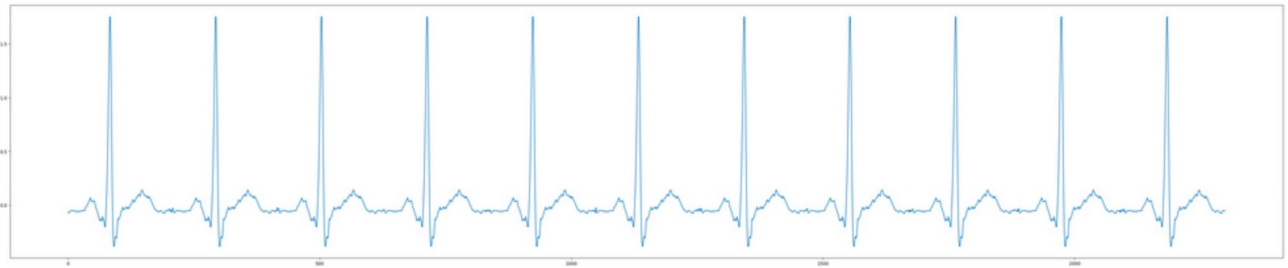
- Follow-up study: **identify the skin-diseases** in an effective and reliable way using DL particularly classification and detection for all the skin tones
  - We employ ResNet as classification model
- We used a modified version of the original dataset to illustrate how a classifier can be unfair for underrepresented classes
  - We use Computer Vision algorithms to change the skin tones of the fair-skinned images

Model	Testing Accuracy (On fair skin-tone)	Testing Accuracy (On dark skin-tone)
ResNet - 50 (Trained only on fair skin dataset)	91.03%	22.72%
ResNet - 50 (Trained on both fair and dark skin dataset)	90.29%	84.22%

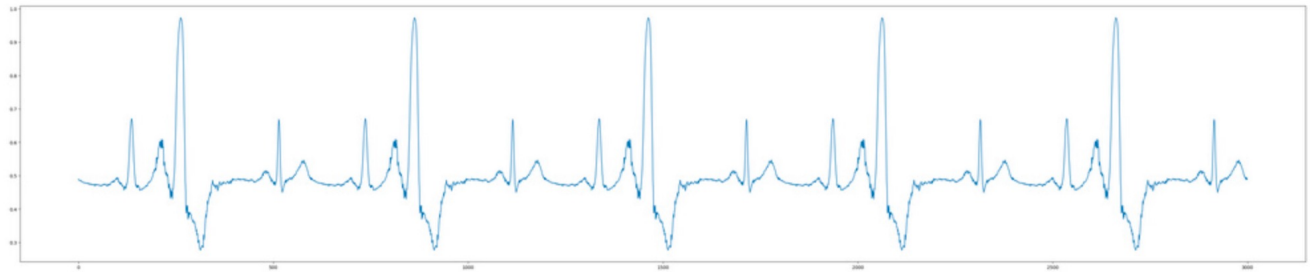
Funded by the European Union

HC2: cVAE-Based approach for beat-by-beat signal generation

- Conditional Variational autoencoder (cVAE) model trained on data from Incart data base St Petersburg INCART 12-lead Arrhythmia Database v1.0.0 (physionet.org)
- Normal ECG signal single beats stitched together



- ECG signal with abnormal beats, here PVC beats are stitched together in a N-N-PVC-N-N-PVC Sequence. The beats are generated using a conditional VAE which is conditioned on the type of beat
  - Each bit is assigned a label based on its type and the types of the previous and next beats



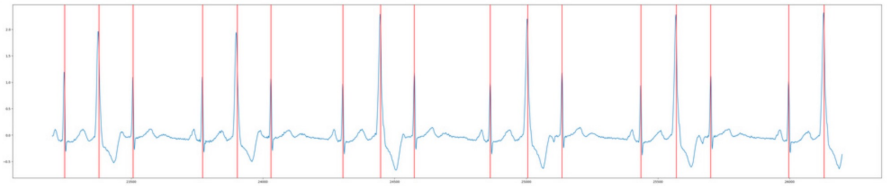
Pipeline adjustments

- Scripted input that controls synthetic data generation, e.g. number of cycles and pattern for ectopic beats, randomness can also be included, e.g. random number of normal beats, from given range, before, between and after episodes
- Addition of noise to synthetic signal
- Annotation of type of beat/episode

HC2: VAE-based approach for beat-by beat generation

Data format, annotation, and processing

Training was performed on INCART dataset. The dataset consists of 75 half-hour 12 leads recordings with per-beat annotation based on the Physionet scheme. Annotations are placed roughly into the middle of QRS complexes (Red lines):



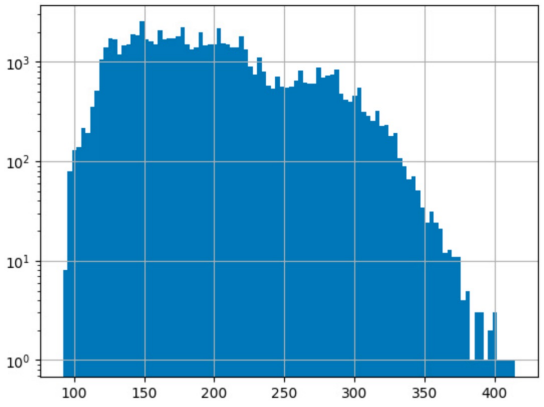
For the first attempt we took healthy ecgs without additional comorbidities (for instance without MI/history of MI/coronary syndrome etc) and only normal and PVC (premature ventricular contraction) beats.

Data was split into beats in the following way:

$$B_n = ((R_n + R_{n-1}) / 2 : (R_n + R_{n+1}) / 2]$$

Where R_n is an index of R peak

This gives the following beat size distribution (y-axis is on log scale):



For training beats are padded on both sides with real values to have length of 500 and annotated in the following scheme:

Each beats has annotation based on its type (N or V) and types of previous and next beats:

[NNN, NNV, NVN, VNN, VNV]

Funded by the European Union

HC2: synthetic ECG signals

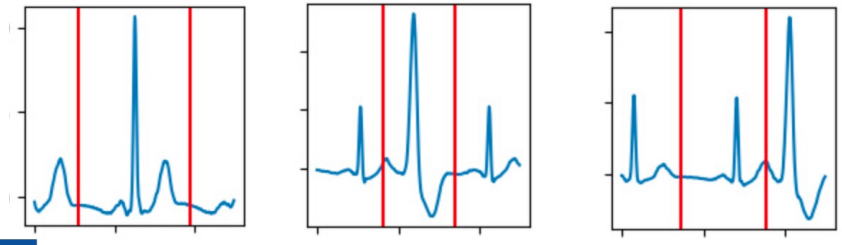
VAE approach for ECG beat-by beat generation

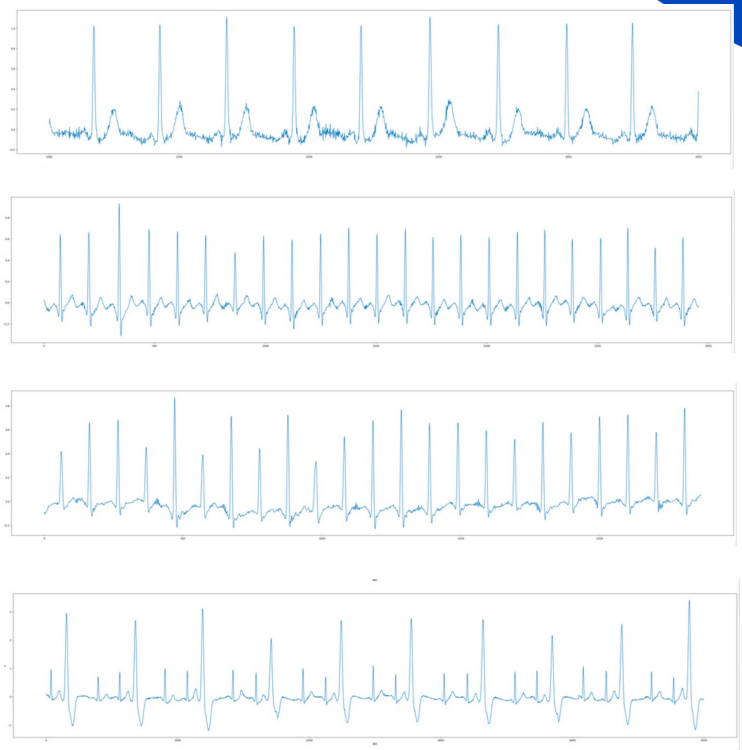
Signal generation process

- Specify heart rate and its variability (mean and std)
- Specify sequence of beats, for instance:

  N,N,V,N,N,V,N,N,V,N,N,V …
- For each normal beat sample RR from the specified distribution
- For each ectopic beat predict its RR with the tree based model
- Predict beat size in pixels based on its type and RR with another tree based model
- Generate beats conditioned on their types and RRs
- Cut the beats from both sides so that they have predicted length plus small padding (10 pixels from each side)
- Stitch beats together with linear fusion on tails

Figures show generated beats and predicted borders





Funded by the European Union

HR1: job matchmaking dataset

- Data cleaned, anonymized
- Published

https://doi.org/10.6092/unibo/amsacta/7715
https://doi.org/10.6092/unibo/amsacta/7716

The dataset contains 331,822 records in total, 39 columns;
Each record is pairing of a candidate with a job
There are 33,338 unique AssociateId (candidates) and 5,066 unique job_id (jobs).

index	max unique vals per AssociateId	max unique vals per job_id
job_id	10	
AssociateId		10633
Gender	1	2
Age_bucket	1	5
PAST_synonyms	1	71
PAST_InailName	1	73
PAST_ProfessionalCategoryName	1	77
PAST_j_EnSkillName	1	127
PAST_j_ItSkillName	1	127
PAST_PostDefinition	1	131
c_lat	1	516
c_long	1	549
c_ZIPCode	1	623
Zipcode	1	631
c_EnSkillName	1	4338
c_ItSkillName	1	4340
associate_jobtitles_extracted_normalized	1	5491
associate_jobtitles_extracted	1	6986
associate_skills_extracted_normalized	1	9394
associate_skills_extracted	1	9425

Candidates

index	max unique vals per AssociateId	max unique vals per job_id
OfferNumber	5	1
BranchId	10	1
WorkOrderNumber	10	1
workorder_jobtitles_extracted	10	1
workorder_skills_extracted	10	1
workorder_jobtitles_extracted_normalized	10	1
workorder_skills_extracted_normalized	10	1
j_ZIPCode	10	1
j_lat	10	1
j_long	10	1
PostDefinition	10	1
ProfessionalCategoryName	10	1
InailName	10	1
synonyms	10	1
j_EnSkillName	10	1
j_ItSkillName	10	1
distance_km	10	549
match_score	10	10336
rev_match_rank	5	5

Jobs

Funded by
the European Union

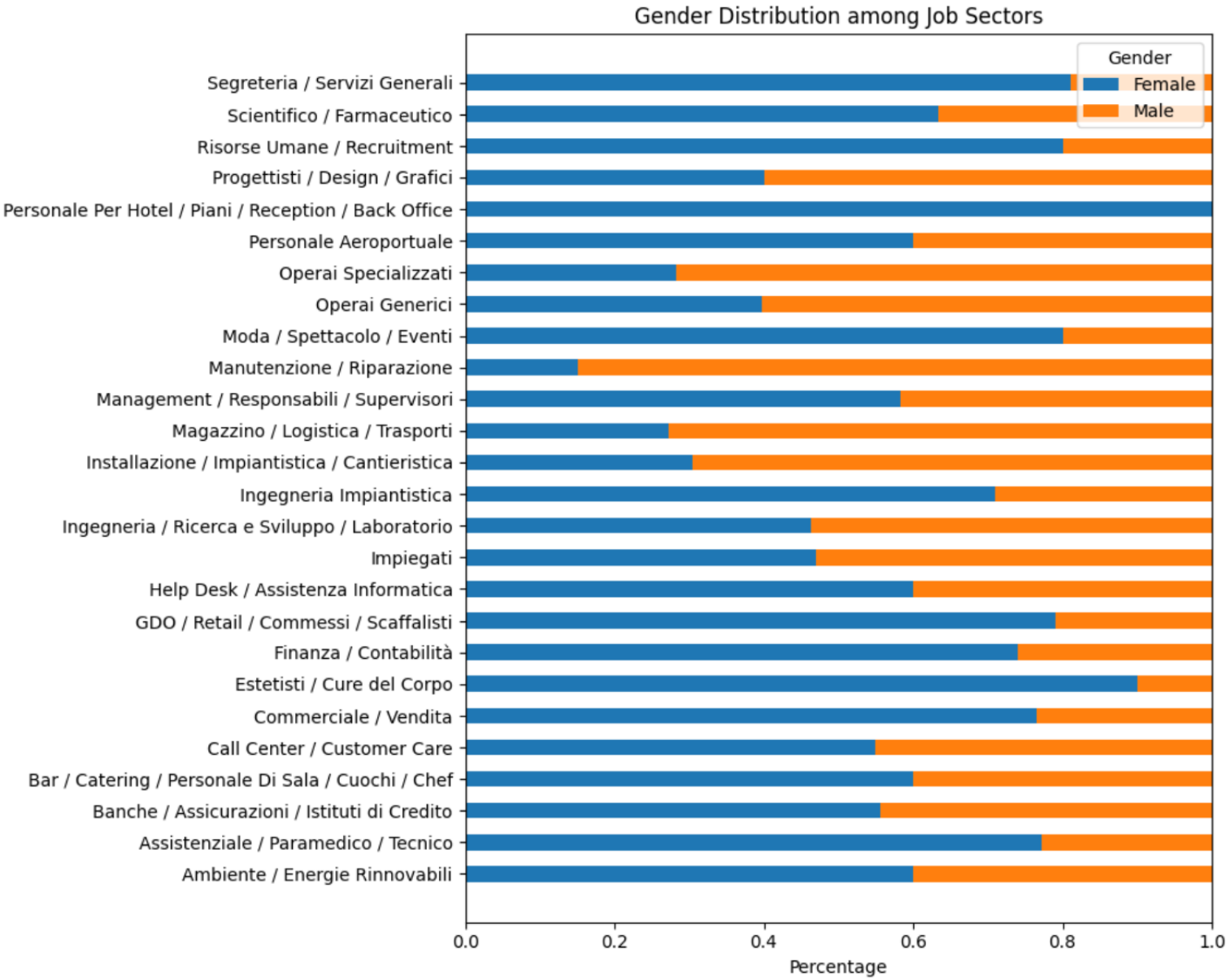
HR1: job matchmaking dataset

- We started analysing the data produced as output by the match-making algorithm
  - Goal #1: identify bias
  - Goal #2: mitigate bias (if present)
- We considered matching algorithms:
  - Direct – find the best 10 candidates for a job position
  - Reverse – find the best 10 job positions for a given candidate
- The dataset described in HR1 is the results of the application of the match-making algorithm
- Big issue: missing data due to fields not necessarily filled by candidates
- We consider gender, location, and age group as potential sensitive attributes

Funded by
the European Union

HR1: candidate

- Some type of bias are implicit in the data itself
- Probably, a reflection of actual bias in the society
  - E.g., most of mechanics are male or all (!?) hotel employees and receptionists are female



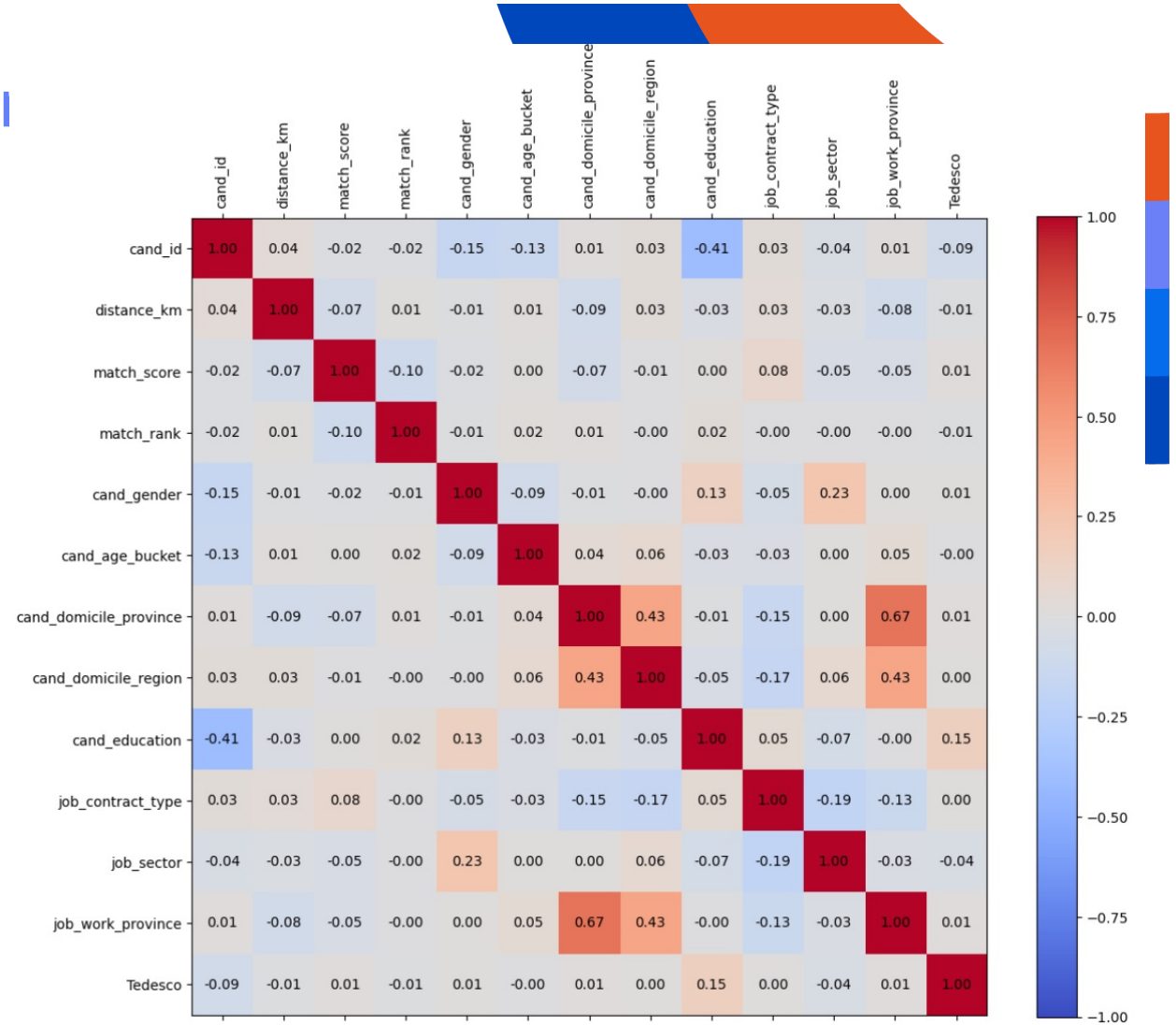
Funded by
the European Union

HR1: candidate selection

Bias Detection

```
f.get_sector_metric(df_processed, "job_sector", 22, "cand_gender")
```

	Sector	Job	Disparate_Impact	Statistical_Parity_Difference
0	22	OFF_1011_1427	0.760404	-0.000397
1	22	OFF_14991_0059	0.760404	-0.000397
2	22	OFF_15007_0059	0.126734	-0.002172
3	22	OFF_15205_0059	0.760404	-0.000397
4	22	OFF_1960_0319	0.488831	-0.000989
5	22	OFF_1963_0100	1.140605	0.000194
6	22	OFF_2276_0192	1.710908	0.000786
7	22	OFF_2498_0165	1.710908	0.000786
8	22	OFF_2516_0130	0.126734	-0.002172
9	22	OFF_2645_0457	4.562421	0.001969
10	22	OFF_2814_0009	0.488831	-0.000989
11	22	OFF_3194_1698	1.140605	0.000194
12	22	OFF_3305_0533	0.760404	-0.000397
13	22	OFF_3805_0251	0.488831	-0.000989
14	22	OFF_5512_0180	0.488831	-0.000989
15	22	OFF_6831_0144	1.710908	0.000786
16	22	ORD_11579_0687	4.562421	0.001969
17	22	ORD_16296_0110	4.562421	0.001969
18	22	ORD_16440_0110	0.000000	-0.002764
19	22	ORD_19665_0190	0.760404	-0.000397
20	22	ORD_22748_0354	0.000000	-0.002764
21	22	ORD_3024_1427	0.285151	-0.001581



Correlation Matrix

HR1: Multinom generator

- Sample categorical data from **multinomial distribution with predefined category probabilities**
  - inferred from the data or set manually
- Useful for generating independent categorical variables
  - Note, if you need to generate dependent variables, you may consider methods like CTGAN

```
{
      "name": "Gender",
      "type": "multinom",
      "labels": ["Male", "Female"],
      "prob": [0.2, 0.8]
    },
    {
      "name": "Age_bucket",
      "type": "multinom",
      "labels": ["25-34", "35-44", "45-54", "15-24", "55-74"],
      "prob": [0.24,  0.12,  0.06,  0.08,  0.5]
    },
    {
      "name": "Zipcode",
      "type": "multinom",
      "data_file": "data/Zipcode_biased.csv"
}
```

Configuration on the left will generate three data columns:
Gender, Age_bucket, Zipcode

"labels" and "probs" fields contains the names of categories and their probabilities

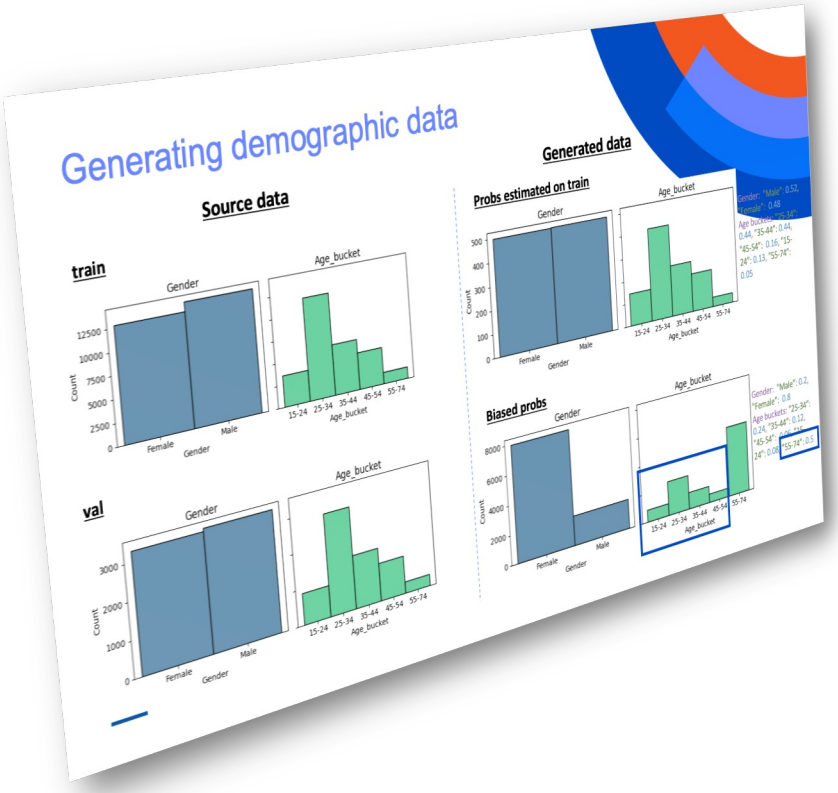
You can also specify the categories via "data_file"

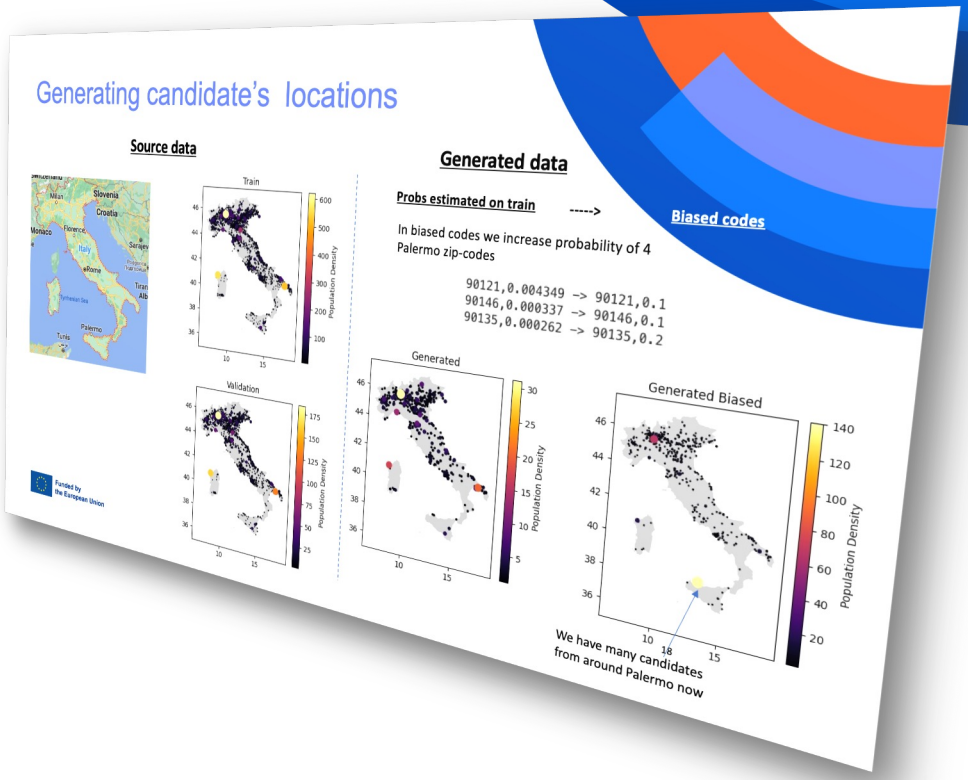
Please refer to the README.md for more details on configuring the multinom generator

Funded by
the European Union

HR1: Synthetic data for HR use cases

Demo





Funded by the European Union

HR1: Language models (LM) for skill generation

Idea: treat skills as tokens (words) in natural language. Train a model to generate skills.

Limitation: differently from words in language, the skill order does not matter

Converting skill columns into input data:

- We have four skill-related columns: associate_jobtitles_extracted_normalized, associate_skills_extracted_normalized, c_EnSkillName, PAST_j_EnSkillName
- We assume that skills within the same columns and between different columns are correlated
- To model all skills together we prepend the skills coming from each column with id of this column:

Input skills

associate_jobtitles_...	associate_skills_extracted_normalized	c_EnSkillName	PAST_j_EnSkillName
"podofo", "manage staff"	"Italian", "coordinate security", "Pizza", "use cooking techniques", "Xcas", "French", "teach survival skills", "total quality control", "English", "geographic areas", "body language", "adapt sets", "NeL", "Svelte", "first aid", "PROSE", "read books"	"(Main) Course Preparation Supervision", "(Main) Course Preparation", "Alimentary Techniques", "Microsoft Office", "International Cuisine", "Italian Regional Cuisine", "Cold Dish Preparation", "Fish Cleaning", "Meat Cleaning", "Course Preparation", "Food decoration", "Hotel experience", "Restaurant experience"	"Fruit and Vegetable Preparation", "Shelf Stocking", "Customer Assistance", "Delicatessen Counter Management", "Retail Sales", "Goods Packaging"

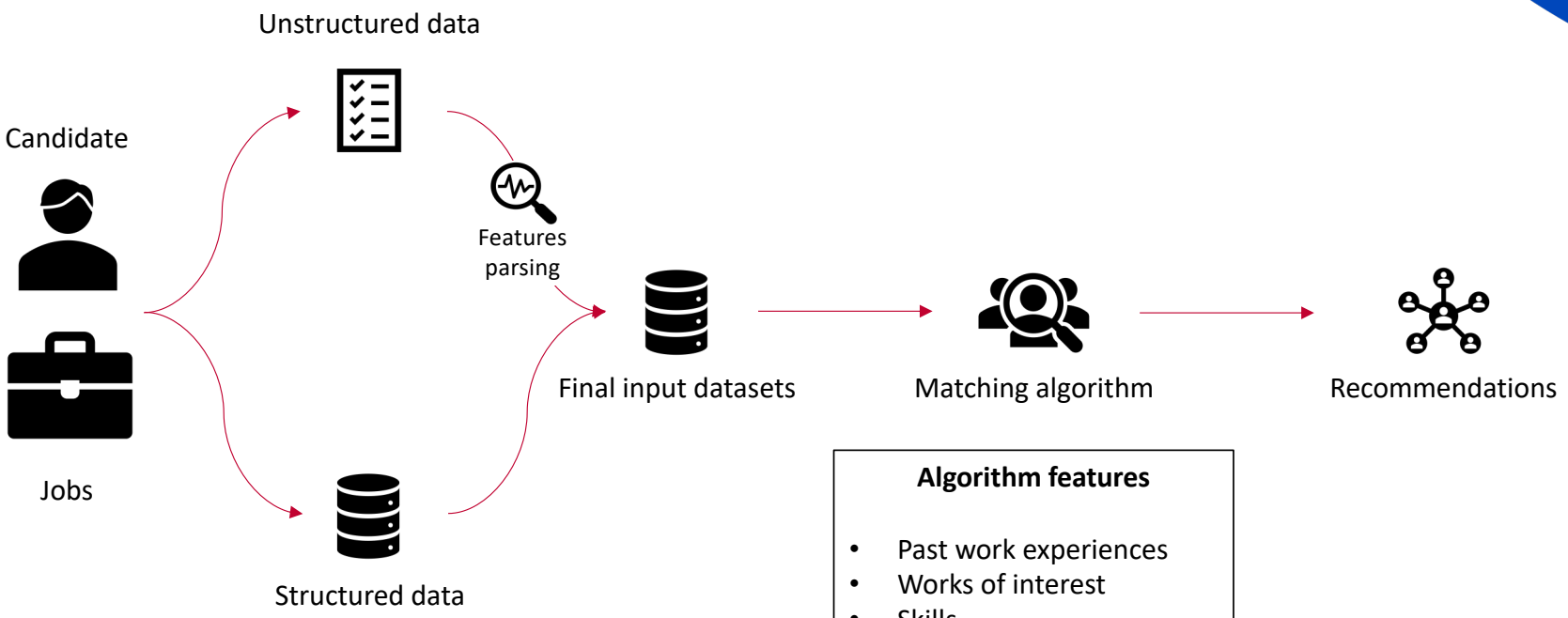
"Skill" language sequence to be modeled by LMs

0_podofo, 0_manage staff, 1_Italian, 1_coordinate security, 1_Pizza, 1_use cooking techniques, 1_Xcas, 1_French, 1_teach survival skills, 1_total quality control, 1_English, 1_geographic areas, 1_body language, 1_adapt sets, 1_NeL, 1_Svelte, 1_first aid, 1_PROSE, 1_read books, 2_(Main) Course Preparation Supervision, 2_(Main) Course Preparation, 2_Alimentary Techniques, 2_Microsoft Office, 2_International Cuisine, 2_Italian Regional Cuisine, 2_Cold Dish Preparation, 2_Fish Cleaning, 2_Meat Cleaning, 2_Course Preparation, 2_Food decoration, 2_Hotel experience, 2_Restaurant experience, 3_Fruit and Vegetable Preparation, 3_Shelf Stocking, 3_Customer Assistance, 3_Delicatessen Counter Management, 3_Retail Sales, 3_Goods Packaging

Funded by the European Union

HR2: candidate selection match-making algorithm

The new candidate matching algorithm is now in testing phase.
Below the new data flow and algorithm:



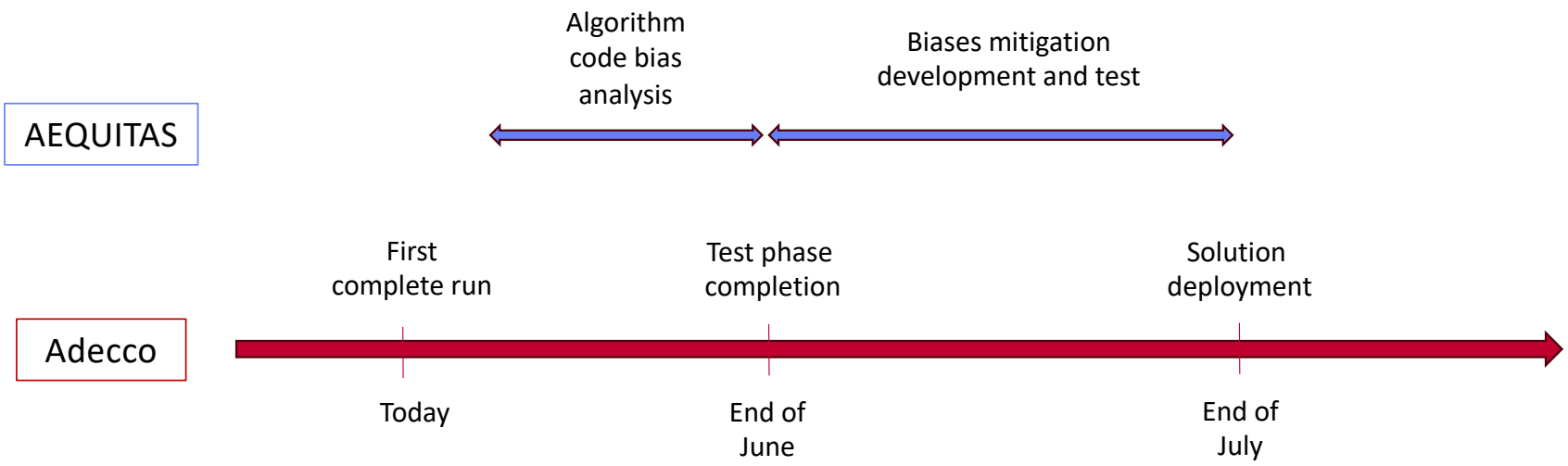
Algorithm features

- Past work experiences
- Works of interest
- Skills
- Education
- Languages
- Distance domicile -> work

Funded by the European Union

HR2: candidate selection match-making algorithm

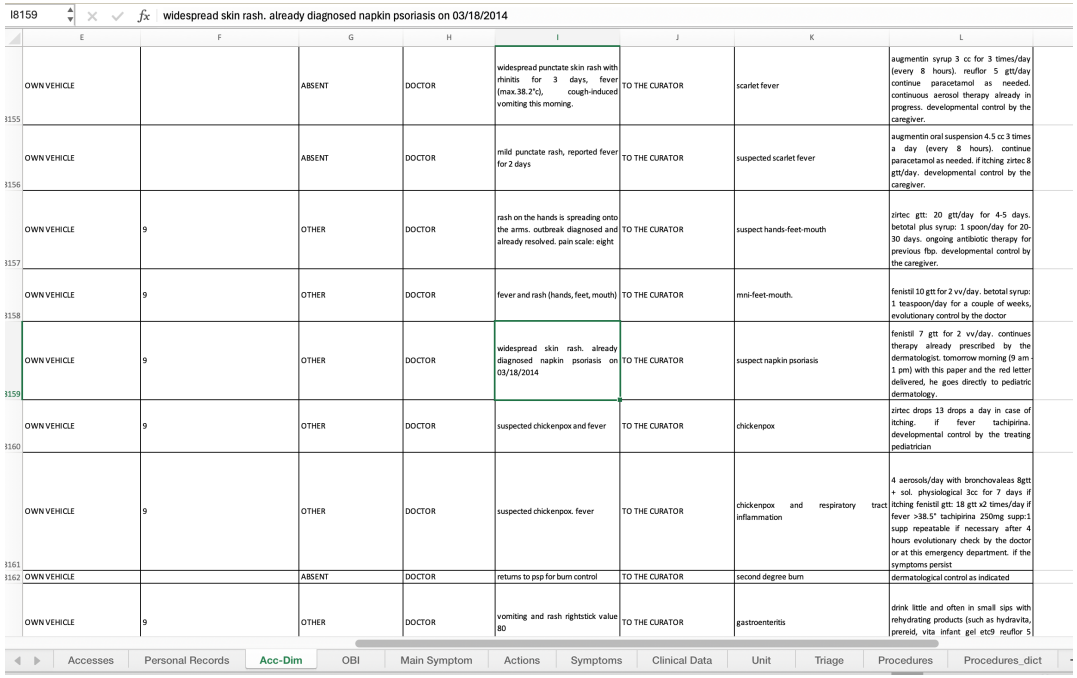
ROADMAP



Funded by the European Union

S1: child-abuse and neglect

- Data collected and anonymized
- Started the analysis



I8159 | widespread skin rash. already diagnosed napkin psoriasis on 03/18/2014

	E	F	G	H	I	J	K	L
3155	OWN VEHICLE		ABSENT	DOCTOR	widespread punctate skin rash with rhinitis for 3 days, fever (max.38.2°c), cough-induced vomiting this morning.	TO THE CURATOR	scarlet fever	augmentin syrup 3 cc for 3 times/day (every 8 hours). reuflor 5 gtt/day continue paracetamol as needed. continuous aerosol therapy already in progress. developmental control by the caregiver.
3156	OWN VEHICLE		ABSENT	DOCTOR	mild punctate rash, reported fever for 2 days	TO THE CURATOR	suspected scarlet fever	augmentin oral suspension 4.5 cc 3 times a day (every 8 hours). continue paracetamol as needed. if itching zirtec 8 gtt/day. developmental control by the caregiver.
3157	OWN VEHICLE	9	OTHER	DOCTOR	rash on the hands is spreading onto the arms. outbreak diagnosed and already resolved. pain scale: eight	TO THE CURATOR	suspect hands-feet-mouth	zirtec gtt: 20 gtt/day for 4-5 days. betotal plus syrup: 1 spoon/day for 20-30 days. ongoing antibiotic therapy for previous fbp. developmental control by the caregiver.
3158	OWN VEHICLE	9	OTHER	DOCTOR	fever and rash (hands, feet, mouth)	TO THE CURATOR	mni-feet-mouth.	fenistil 10 gtt for 2 vv/day. betotal syrup: 1 teaspoon/day for a couple of weeks, evolutionary control by the doctor
3159	OWN VEHICLE	9	OTHER	DOCTOR	widespread skin rash. already diagnosed napkin psoriasis on 03/18/2014	TO THE CURATOR	suspect napkin psoriasis	fenistil 7 gtt for 2 vv/day. continues therapy already prescribed by the dermatologist. tomorrow morning (9 am - 1 pm) with this paper and the red letter delivered, he goes directly to pediatric dermatology.
3160	OWN VEHICLE	9	OTHER	DOCTOR	suspected chickenpox and fever	TO THE CURATOR	chickenpox	zirtec drops 13 drops a day in case of itching. if fever tachipirina. developmental control by the treating pediatrician
3161	OWN VEHICLE	9	OTHER	DOCTOR	suspected chickenpox. fever	TO THE CURATOR	chickenpox and respiratory tract inflammation	4 aerosols/day with bronchovaleas 8gtt + sol. physiological 3cc for 7 days if itching fenistil gtt: 18 gtt x2 times/day if fever >38.5° tachipirina 250mg supp:1 supp repeatable if necessary after 4 hours evolutionary check by the doctor or at this emergency department. if the symptoms persist
3162	OWN VEHICLE		ABSENT	DOCTOR	returns to psp for burn control	TO THE CURATOR	second degree burn	dermatological control as indicated
	OWN VEHICLE	9	OTHER	DOCTOR	vomiting and rash rightstick value 80	TO THE CURATOR	gastroenteritis	drink little and often in small sips with rehydrating products (such as hydravita, prereid, vita infant gel etc9 reuflor 5

Accesses | Personal Records | Acc-Dim | OBI | Main Symptom | Actions | Symptoms | Clinical Data | Unit | Triage | Procedures | Procedures_dict

Analysis started

Funded by
the European Union

https://doi.org/10.5281/zenodo.11171863

S2: disadvantaged students

- Novel benchmark for AI fairness in **education**.
- Focus on improving student performance and reducing dropout rates.
- Data from students, families, principals, and teachers in the Canary Islands, Spain; across four academic year, reaching 83 thousands students.

Along with **data & meta-data**, we propose

- the **key goals**,
- the underlying **AI tasks**,
- the **interventions** to guarantee **fairness**.

G1	Fair selection of promising students
T1	Predict academic performance
T2	Rank students by predicted performance
I1	Rebalance under- and over-sampled schools

G2	Explainable detection of underperforming students
T3	Detect performance threshold
T2	Predict academic performance
I2	Assure model in T2 is explainable
I3	Handle and augment missing values

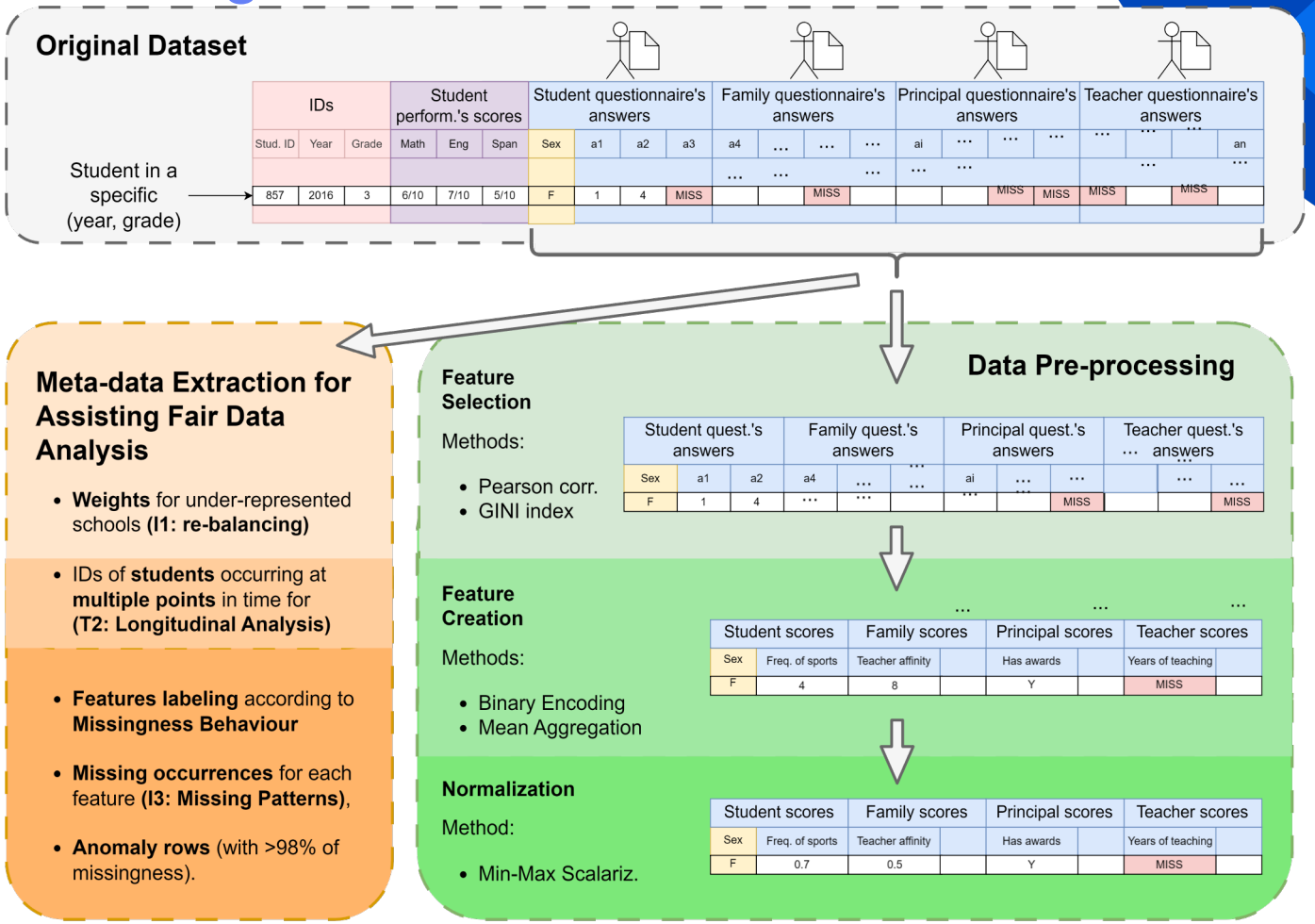
G3	Early drop-off prediction
T4	Identify potential drop-offs
I4	Address sampling bias and infer drop-offs

G4	Impact of socio-economic factors on performance
T2	Train performance classifier/regressor
T5	Extract global explanations
T6	Learn socio-economic interdependencies
I4	Ensure no discrimination against any groups

G5	Most-likely explanations of performance
T6	Use graphical models for interdependencies
I5	Apply "most probable explanation" algorithm

Funded by the European Union

S2: disadvantaged students



Meta-data Extraction provides all meta-data to apply **fairness interventions** in the proposed **goals** and **tasks**.

Data Pre-processing aggregates questionnaire answers to define indicators.

- **degree of agreement** to a statement
- **frequency of** a certain activity,
- **holding** (or not) a characteristic.

Funded by the European Union

Synthetizer

The Aequitas synthetic data generator (also known as data synthesizer) comprises two elements that both are **Python modules**:

- a synthesizer for various types of **structured data** with a component that can handle less structured data, and
- a synthesizer for **images**.

➢ Generation of ***polarized*** data

https://github.com/aequitas-aod

Funded by
the European Union

Methodology: status

Funded by
the European Union

Building Blocks

- Current visual for the fair-by-design building blocks
- All building blocks are interconnected.
- Stakeholder engagement is the connective tissue that unifies all project methodologies. Our Stakeholder Engagement Methodology will detail this critical process.
- The placement of the building blocks might alter as the writing progresses.
  - Ex: 'Social, Legal and Ethical Fairness Constraints Identification' could be expanded to the 'scoping' phase.

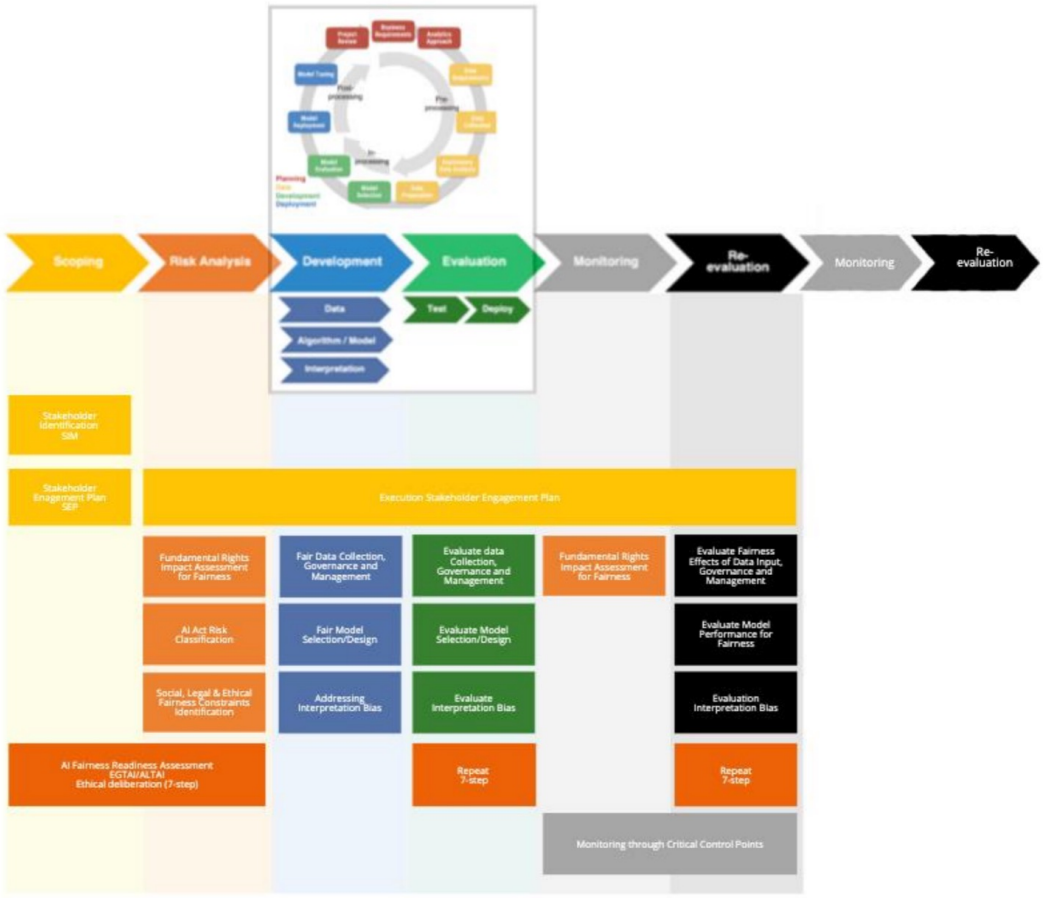


Figure 4: AEQUITAS AI Lifecycle – Fair by Design methodology building blocks

Funded by the European Union

Guidelines for the Building Blocks

- **Stakeholder Identification Methodology – Completed (minor adaptations)**
- **Stakeholder Engagement Methdology – In Progress (10%)**
- **Fundamental Rights Impact Assessment for Fairness – In Progress (50%)**
- **AI Act Risk Classification – In Progress (10-15%)**
- **AI Fairness Readiness Assessment/ 7 Step Process – Completed**
- Social, Legal and Ethical Fairness Constraints Identification – Pending Start
- **Fair Data Collection, Governance, and Management Methodology – Completed (95%)**
- Fair Model Selection / Design – Pending Start
- **Addressing Interpretation Bias – In Progress (15%) / Research completed under WP4**
- Evaluate Data Collection, Governance, and Management Methodology - Pending Start
- Evaluate Fair Model Selection / Design – Pending Start
- Evaluate Interpretation Bias – Pending Start
- Evaluate Fairness Effects of Data Input, Governance, and Management – Pending Start
- Evaluation Performance for Fairness – Pending Start
- Evaluation Interpretation Bias – Pending Start

Funded by
the European Union

Q/A

	ID	From	Article text	Engine	Conc	Question	Detection & Diagnosis	Mitigation & Repair	Fair by design	Ren
163	Q80.1	CoFR.21.1.ND			Algorithm	optimize for?				
164	Q81	CoFR.21.1.ND			Algorithm	Can the variables used result in discriminatory outcomes?				
165	Q82	CoFR.21.1.ND			Algorithm	Does the algorithm categorize, profile, score individuals or groups of individuals?				
166	Q83	CoFR.21.1.ND			Data	sensitive data lead to discriminatory outcomes based on the sensitive attributes?				
167	Q84	CoFR.21.1.ND			Data	Can the combination of multiple sensitive data points increase the discrimination potential?				
168	Q85	CoFR.21.1.ND			Algorithm	human intervention (e.g., decision approval)?				
169	Q86	CoFR.21.1.ND			Interpretati	require human interpretation of the output?				
170	Q87	CoFR.21.1.ND			Interpretati	human agent? Can such display induce bias in their interpretation?				
171	Q88	CoFR.21.1.ND			Interpretati	delivers the interpretation challenge or rectify the system's output and input?				
172	Q89	CoFR.21.2.ND	Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited		Data	Does the data (training, testing, validation) hold information on nationality?				
173	Q90	CoFR.21.2.ND			Data	Does the data (training, testing, validation) hold information that can be a proxy for nationality?				
174	Q91	CoFR.21.2.ND			Algorithm	application have an inherent/ longstanding history of discrimination and bias based on nationality?				
175	Q92	CoFR.21.2.ND			Algorithm	parameter within the system's process?				
176	Q93	CoFR.21.2.ND			Algorithm	Does the system incorporate the possibility for human intervention in the final output?				
177		Questions from the Charter								
178										

Art. 6 | Questions | Answers | Ordering | Articles | Lifecycle Phases | Stakeholders

Statistiche cartella di lavoro | Filtri applicati | 100%

Funded by the European Union

Techniques & Metrics: status

Funded by
the European Union

Assessment (Publications)

TITLE	RANK	TOPIC
Impact based fairness framework for socio-technical decision making		
Generalized Disparate Impact for Configurable Fairness Solutions in ML	A*	DI metrics extended for continuous protected attributes

In the pipeline
Enhancing the Applicability of Fair Learning with Continuous Attributes
A new metric for AI fairness: integrating social, legal and technical perspectives
Beyond explicit bias: Counterfactual explanations as unfairness indicators

Funded by
the European Union

Mitigation (Publications)

TITLE	RANK	TOPIC
FAiRDAS: Fairness-Aware Ranking as Dynamic Abstract System		Fairness constraints in ML optimization
Ensuring Fairness Stability for Disentangling Social Inequality in Access to Education: the FAiRDAS General Method	A*	Fairness constraints in ML optimization
Unveiling Opaque Predictors via Explainable Clustering: The CReEPy Algorithm		XAI for fairness
Achieving Complete Coverage with Hypercube-Based Symbolic Knowledge-Extraction Techniques		XAI for fairness
ExACT Explainable Clustering: Unravelling the Intricacies of Cluster Formation		XAI for fairness
Unlocking Insights and Trust: The Value of Explainable Clustering Algorithms for Cognitive Agents		

In the pipeline
Mitigating Intersectional Fairness: a Practical Approach with FaUCI
AutoML and argumentation for fairness

Funded by the European Union

Methodology / Data / Benchmark

TITLE	RANK	TOPIC
Impact based fairness framework for socio-technical decision making		Socio-technical Framework for fairness
Assessing and Enforcing Fairness in the AI	A*	Tech methodology for AI fairness: link to the AI system lifecycle
A geometric framework for fairness		
Generation of Clinical Skin Images with Pathology with Scarce Data		Synthetic data generation

In the pipeline
Unfair Inequality in Education: A Benchmark for AI-Fairness Research
AI-fairness: the AEQUITAS approach to practically bridge the gap between socio-legal and technical perspectives

Funded by
the European Union

AEQUITAS
unbias AI

Consortium

Funded by the European Union. Views and opinions expressed are however those of the authors) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by
the European Union

www.aequitas-project.eu
info@aequitas-project.eu