ALGORITHMS FOR SOCIAL GOOD: FAIRNESS AND BIAS IN DATA-DRIVEN DECISION-MAKING SYSTEMS

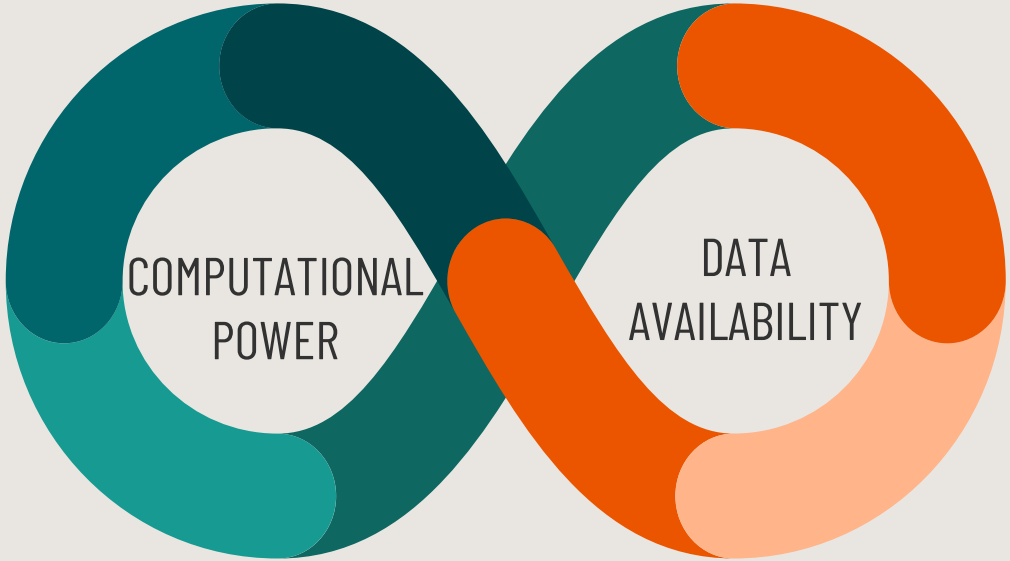
Elena Beretta — PhD candidate (XXXIII cycle) – Thesis Defense
Nexa Center for Internet & Society, Politecnico di Torino, Italy
Fondazione Bruno Kessler, Trento, Italy

Supervisors — Prof. Juan Carlos De Martin, Politecnico di Torino
Bruno Lepri, FBK

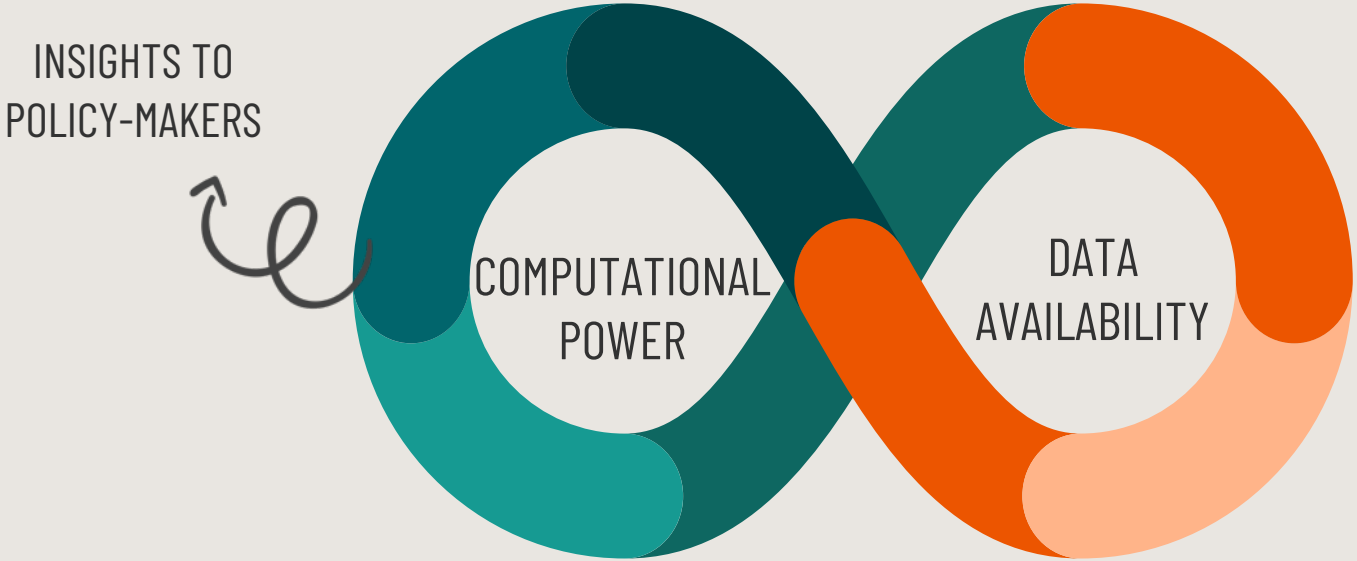
Advisor — Antonio Vetrò, Politecnico di Torino

TABLE OF CONTENTS

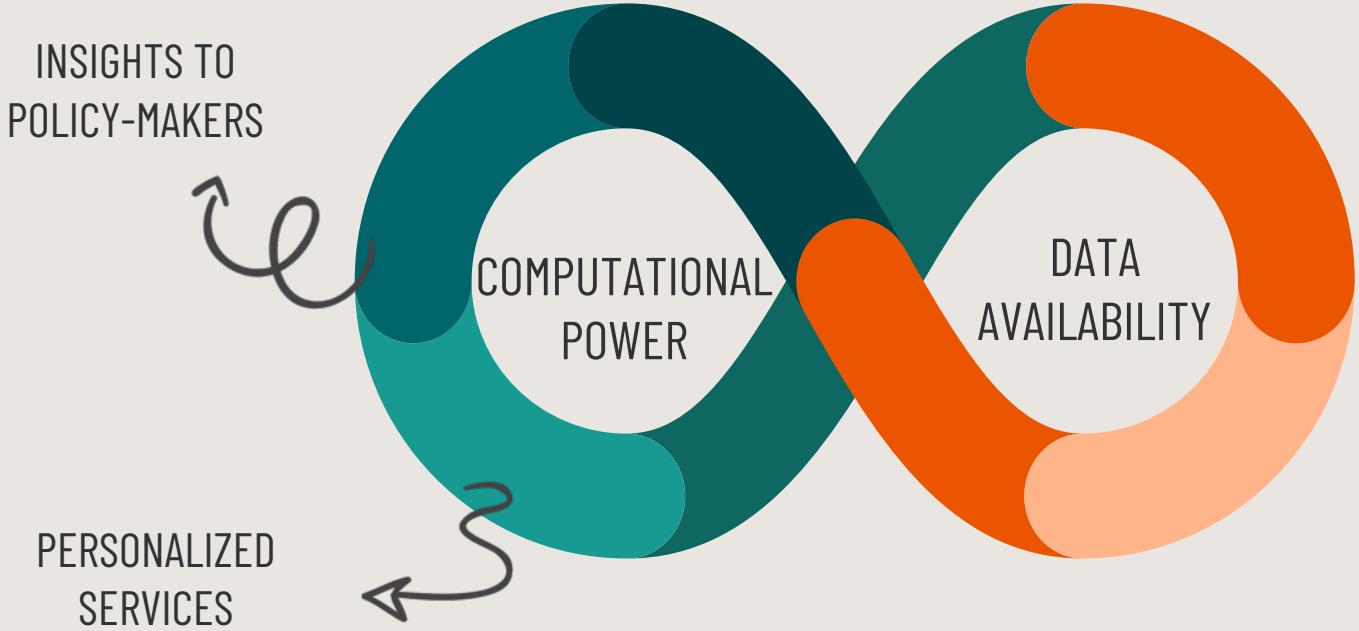
RESEARCH CONTEXT



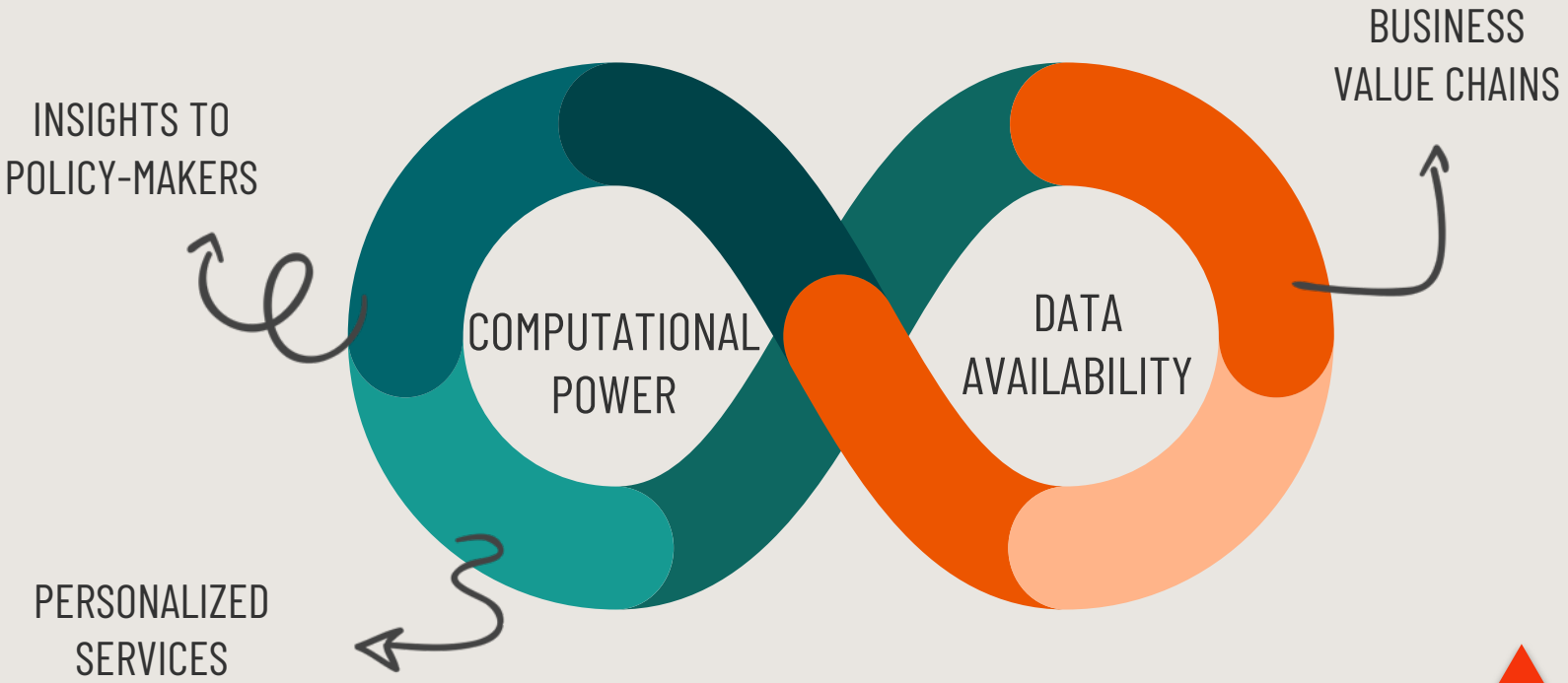
RESEARCH CONTEXT



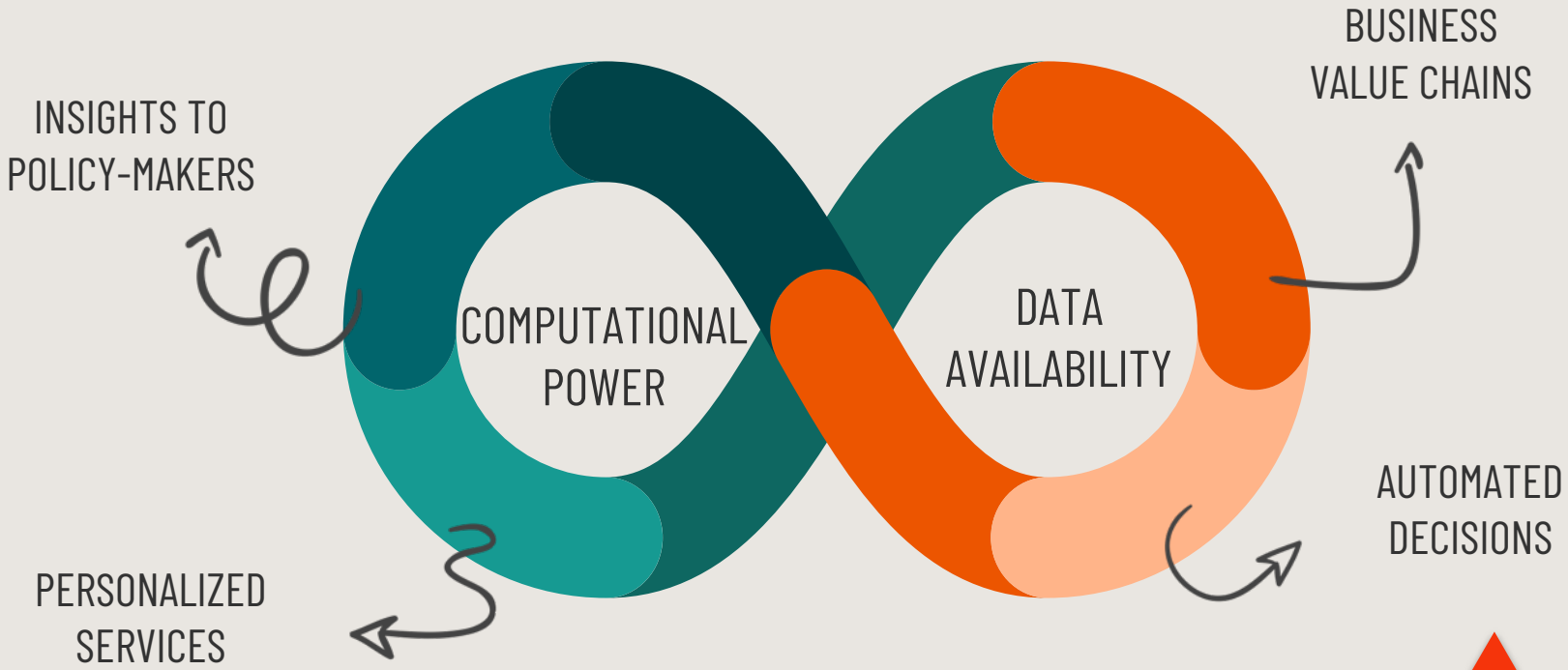
RESEARCH CONTEXT



RESEARCH CONTEXT



RESEARCH CONTEXT



RESEARCH CONTEXT

RESEARCH CONTEXT

SEARCH ENGINE

RESEARCH CONTEXT

SEARCH ENGINE

PRODUCTS

RESEARCH CONTEXT

SEARCH ENGINE

PRODUCTS

CULTURE

RESEARCH CONTEXT

SEARCH ENGINE

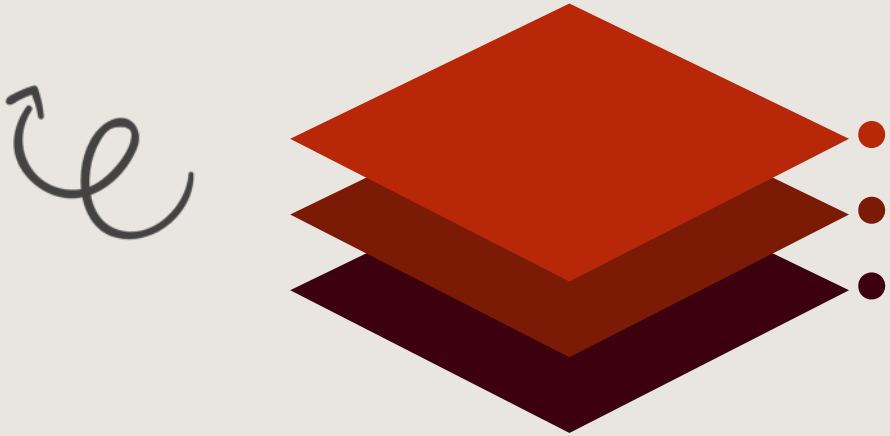
PRODUCTS

CULTURE

PEOPLE

AUTOMATED DATA-DRIVEN DECISION-MAKING SYSTEMS

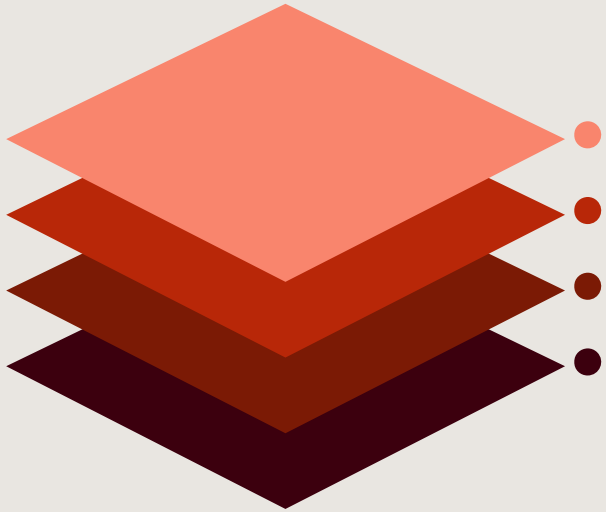
AUTOMATED DECISION SYSTEM



- COMPUTER PROGRAMS
- ALGORITHMS
- DATA

AUTOMATED DATA-DRIVEN DECISION-MAKING SYSTEMS

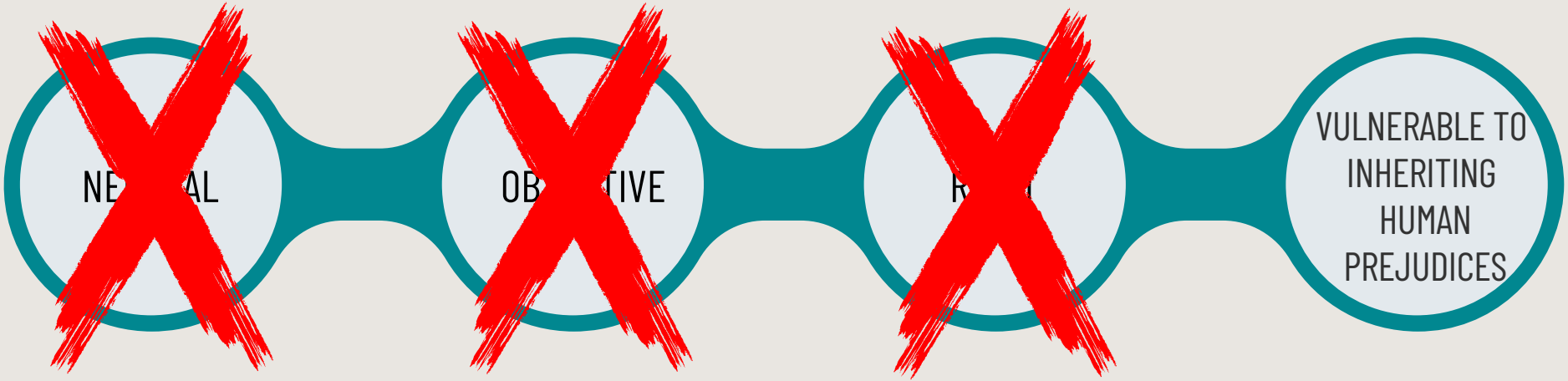
AUTOMATED DATA-DRIVEN DECISION-MAKING SYSTEM (ADMs)



- MACHINE LEARNING SYSTEMS
- COMPUTER PROGRAMS
- ALGORITHMS
- DATA

DISCRIMINATION IN ADMs

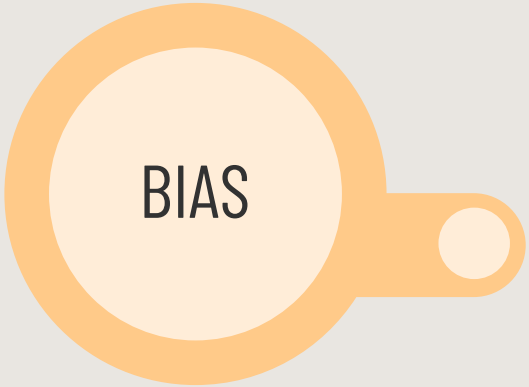
DISCRIMINATION IN ADMs



WHAT ABOUT THE TERM *BIAS*?

BIAS

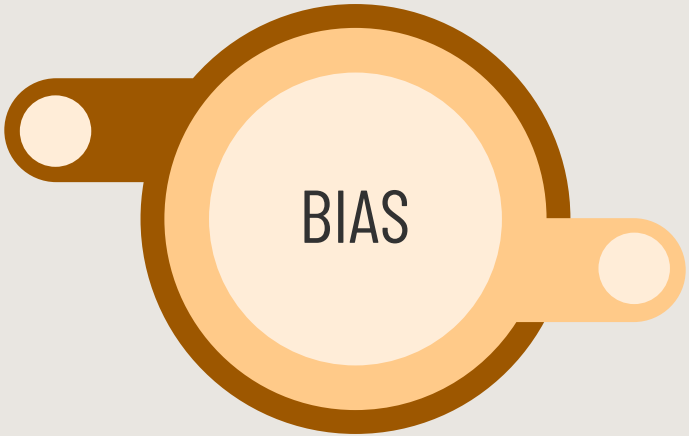
WHAT ABOUT THE TERM *BIAS*?



BIASED DATA COLLECTION

WHAT ABOUT THE TERM *BIAS*?

BIASED ESTIMATOR

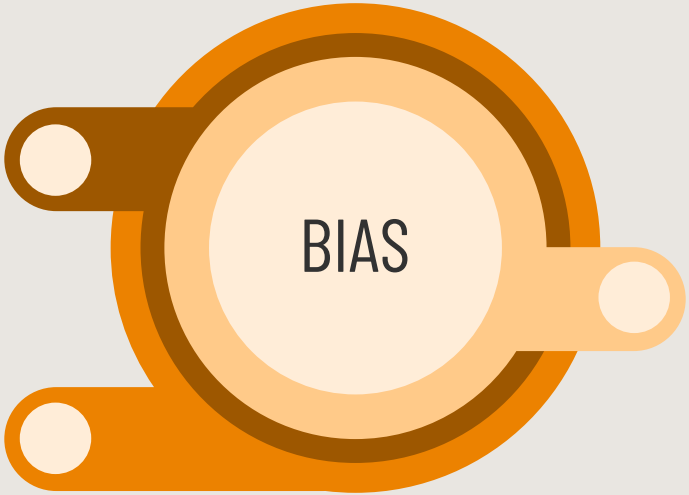


BIASED DATA COLLECTION

WHAT ABOUT THE TERM *BIAS*?

BIASED ESTIMATOR

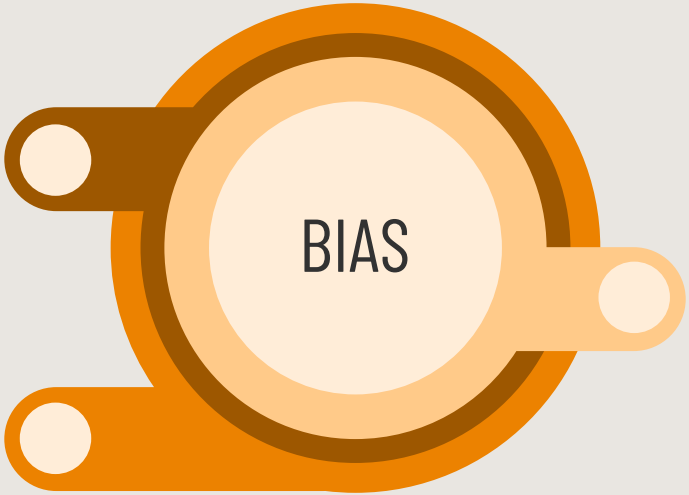
UNJUSTIFIED DISCRIMINATION



BIASED DATA COLLECTION

WHAT ABOUT THE TERM *BIAS*?

BIASED ESTIMATOR

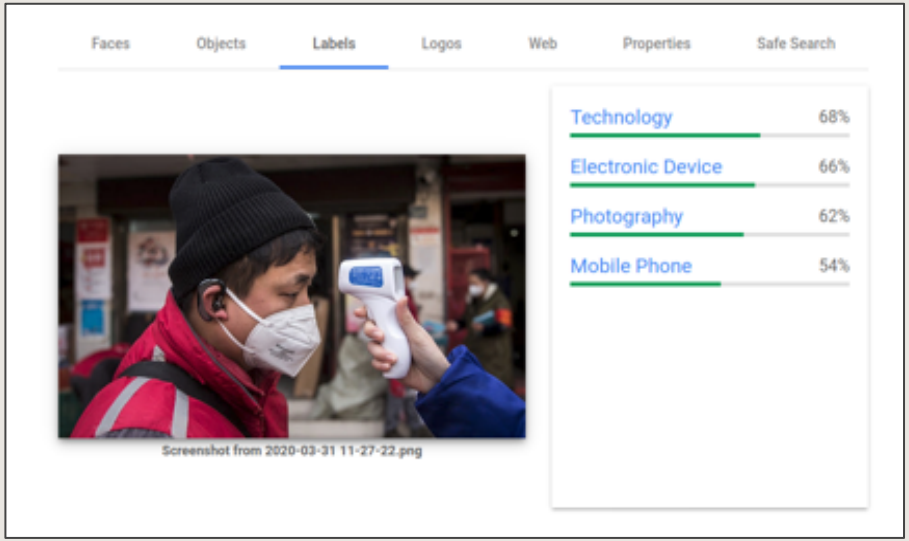


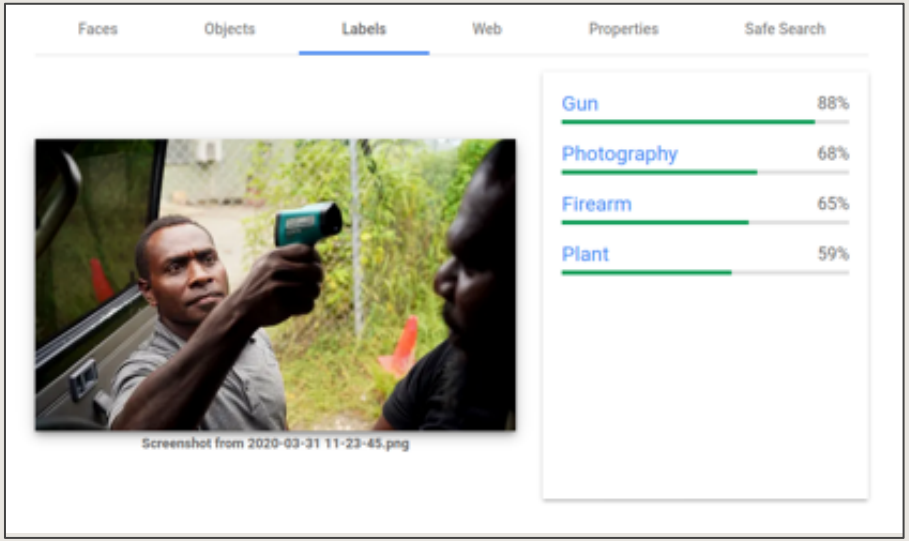
BIASED DATA COLLECTION

UNJUSTIFIED DISCRIMINATION

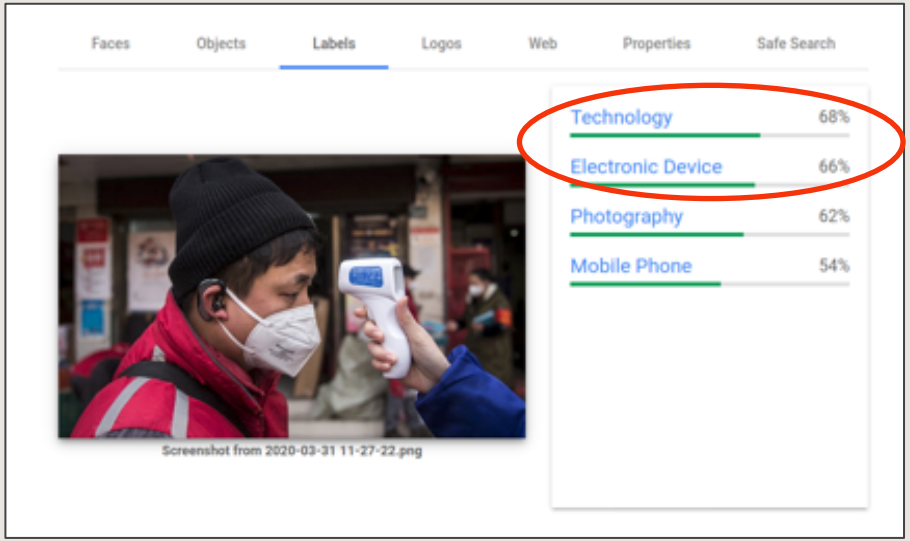
GENDER, ETHNICITY, RELIGION...ETC

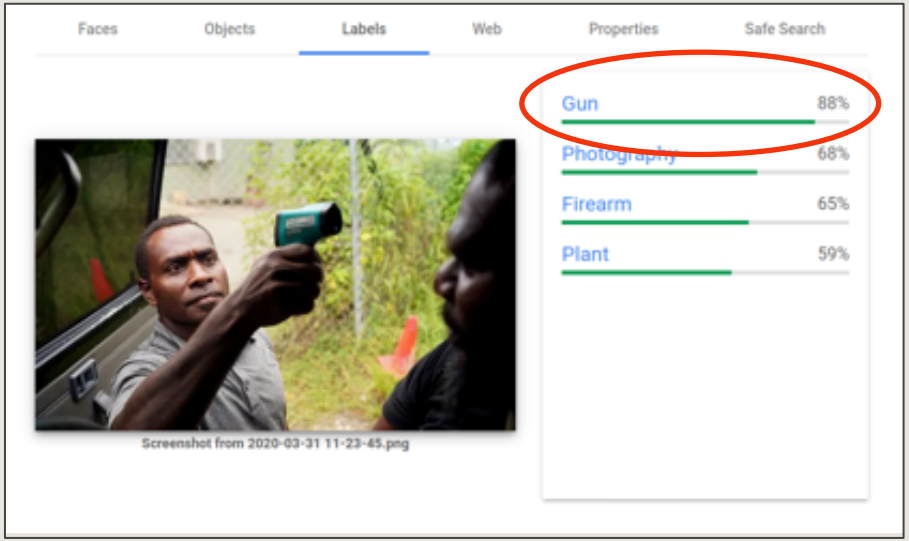
UNJUSTIFIED DISCRIMINATION: GOOGLE VISION AI





UNJUSTIFIED DISCRIMINATION: GOOGLE VISION AI





OVERALL GOALS

ADDRESSING PROBLEMS OF UNJUSTIFIED DISCRIMINATION IN ADMs

LEVERAGING MULTIPLE DISCIPLINES TO MITIGATE THESE UNDESIRABLE EFFECTS AND TO PROVIDE CROSS-DISCIPLINARY PERSPECTIVES OF ANALYSIS

THESIS STRUCTURE

THESIS STRUCTURE

INTERSECTION OF SCIENCE, TECHNOLOGY AND SOCIETY

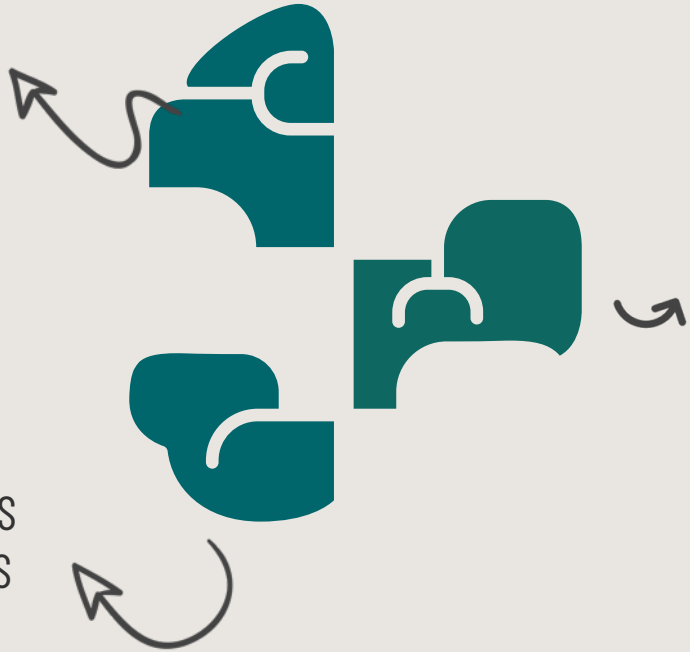
THESIS STRUCTURE

INTERSECTION OF SCIENCE, TECHNOLOGY AND SOCIETY

NOT PURELY TECHNICAL

THESIS STRUCTURE

INTERSECTION OF SCIENCE, TECHNOLOGY AND SOCIETY



NOT PURELY TECHNICAL

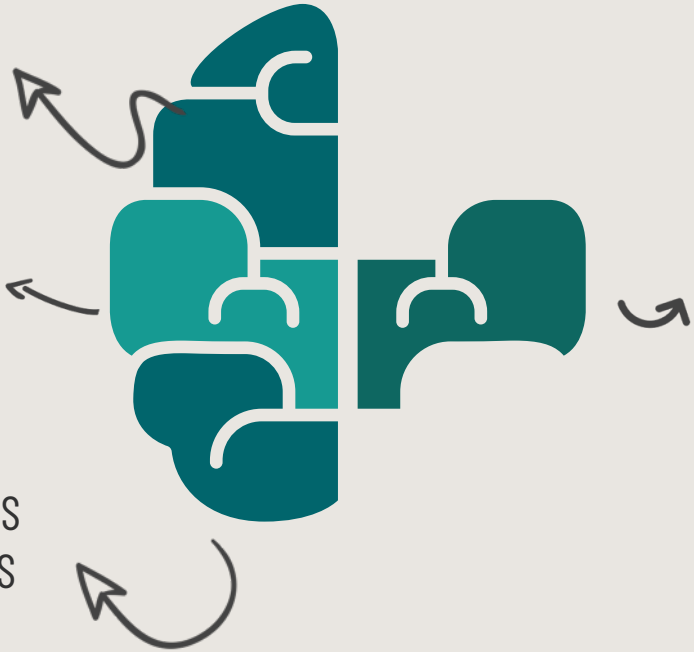
SOCIAL SCIENCES AND HUMANITIES

THESIS STRUCTURE

INTERSECTION OF SCIENCE, TECHNOLOGY AND SOCIETY

INTERDISCIPLINARY PERSPECTIVE

SOCIAL SCIENCES AND HUMANITIES

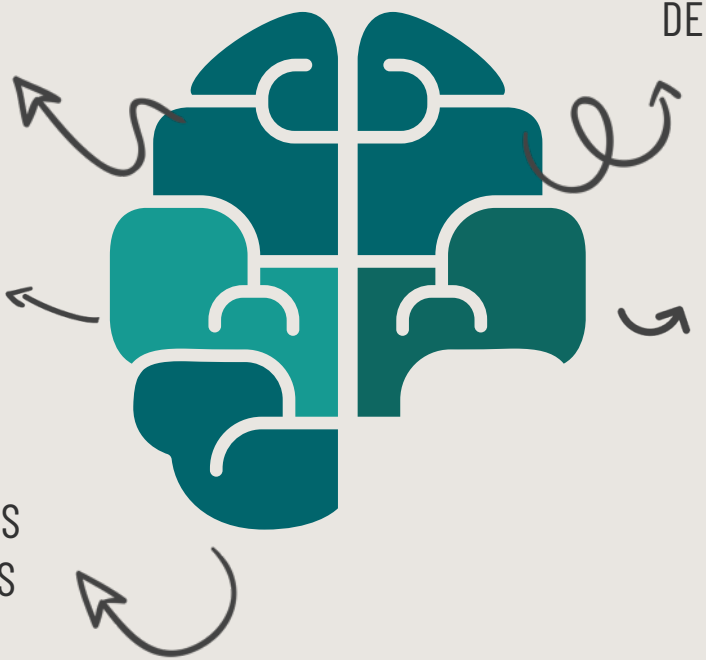


NOT PURELY TECHNICAL

THESIS STRUCTURE

FAIRNESS AND BIAS IN AUTOMATED DATA-DRIVEN DECISION-MAKING SYSTEMS

INTERSECTION OF SCIENCE, TECHNOLOGY AND SOCIETY



INTERDISCIPLINARY PERSPECTIVE

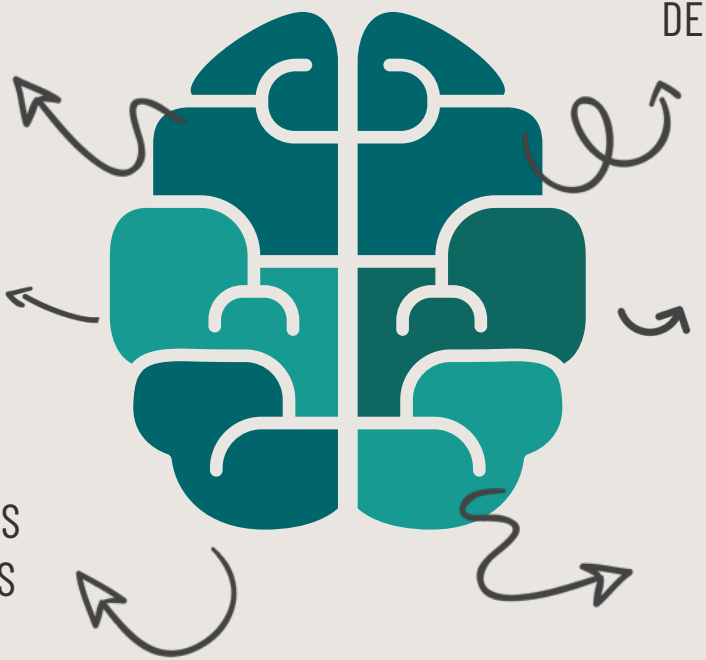
NOT PURELY TECHNICAL

SOCIAL SCIENCES AND HUMANITIES

THESIS STRUCTURE

FAIRNESS AND BIAS IN AUTOMATED DATA-DRIVEN DECISION-MAKING SYSTEMS

INTERSECTION OF SCIENCE, TECHNOLOGY AND SOCIETY



INTERDISCIPLINARY PERSPECTIVE

NOT PURELY TECHNICAL

SOCIAL SCIENCES AND HUMANITIES

THEME CHAPTERS

THESIS CONTRIBUTION

DATA ANNOTATION SYSTEM

PREDICTING FUTURE DISCRIMINATORY RISK BASED ON BIASED DATA

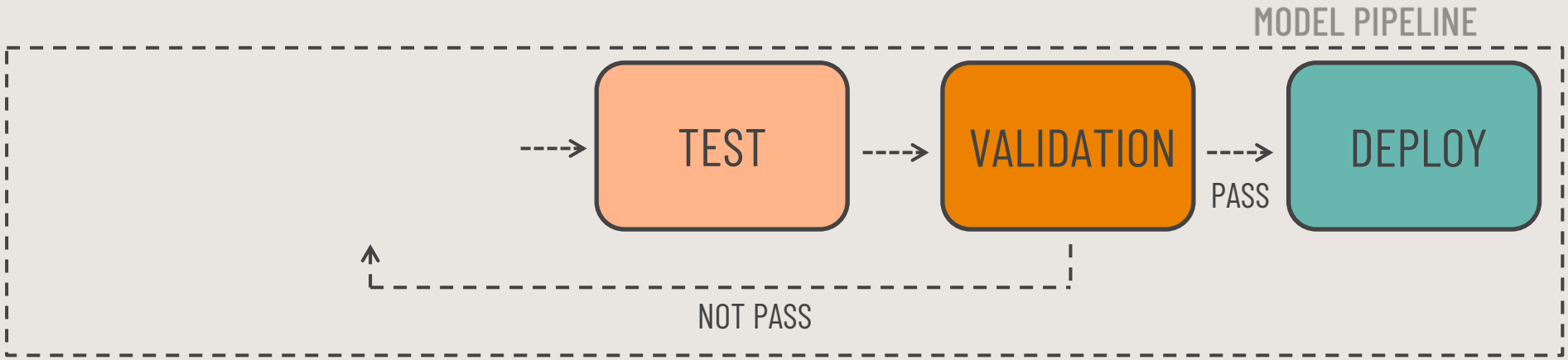
RANKING SYSTEM

PROPOSING A FAIR-DISTRIBUTIVE RANKING SYSTEM

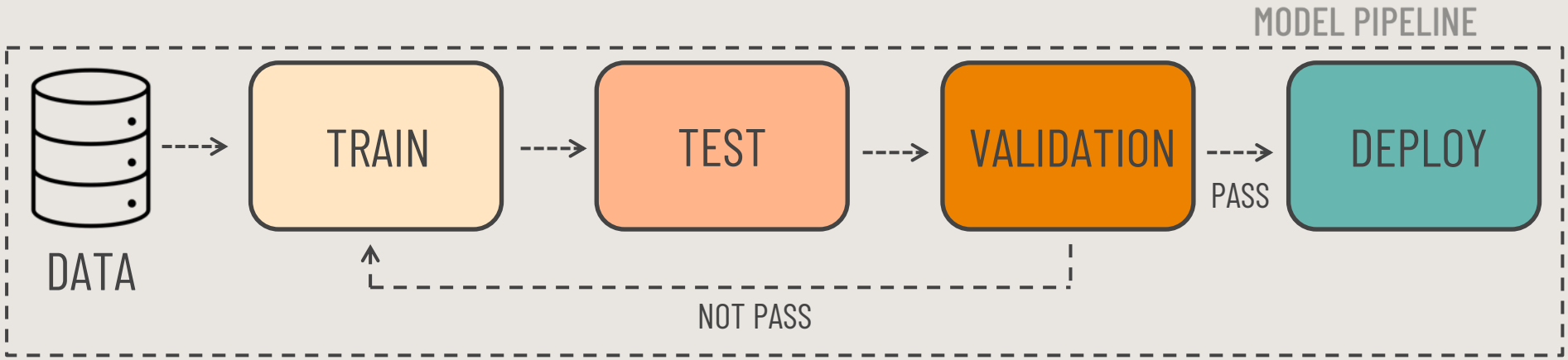
DECISION SUPPORT SYSTEM

MODELING INDIVIDUAL DYNAMICS FOR LONG-TERM FAIRNESS

DATA ANNOTATION SYSTEM



DATA ANNOTATION SYSTEM



RESEARCH QUESTIONS

- IS IT POSSIBLE TO ESTABLISH THE A PRIORI PROBABILITY OF THE TRAINING DATA DISTRIBUTION FROM THE AVAILABLE DATASET?
- HOW TRAINING DATA COULD INFORM ABOUT THE RISK OF FUTURE DISCRIMINATION?

THESIS CONTRIBUTION

DATA ANNOTATION SYSTEM

PREDICTING FUTURE DISCRIMINATORY RISK BASED ON DATA BIAS

RANKING SYSTEM

PROPOSING A FAIR-DISTRIBUTIVE RANKING SYSTEM

DECISION SUPPORT SYSTEM

MODELING INDIVIDUAL DYNAMICS FOR LONG-TERM FAIRNESS

RANKING SYSTEM

RANKING SYSTEM

FAIRNESS IN RANKING SYSTEMS IS LESS EXPLORED THAN MACHINE LEARNING

RANKING SYSTEM

THE MAJORITY OF THE STUDIES PROVIDES A DEFINITION OF EQUITY RATHER THAN GIVING A SOLUTION TO INEQUALITY

FAIRNESS IN RANKING SYSTEMS IS LESS EXPLORED THAN MACHINE LEARNING

RESEARCH QUESTIONS

- ARE RANKING SYSTEMS BASED ON A DISTRIBUTIVE FAIRNESS CONSTRAINT ABLE TO PRESERVE THE ACCURACY OF THE RANKING AND THE MODEL'S OVERALL UTILITY BY PROVIDING A RANKING OF THE BEST CANDIDATES?
- WHAT ARE THE FACTORS AFFECTING THE FAIRNESS UTILITY TRADE-OFF IN A FAIRNESS CONSTRAINED RANKING SYSTEM?

THESIS CONTRIBUTION

DATA ANNOTATION SYSTEM

PREDICTING FUTURE DISCRIMINATORY RISK BASED ON DATA BIAS

RANKING SYSTEM

PROPOSING A FAIR-DISTRIBUTIVE RANKING SYSTEM

DECISION SUPPORT SYSTEM

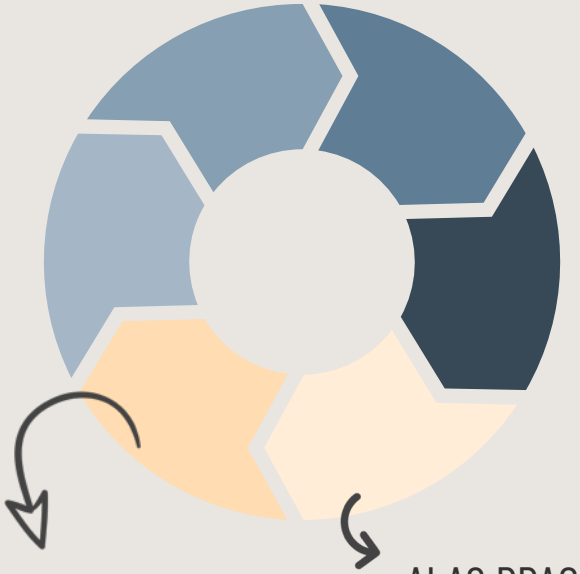
MODELING INDIVIDUAL DYNAMICS FOR LONG-TERM FAIRNESS

DECISION SUPPORT SYSTEM

DECISION SUPPORT SYSTEM

AI AS PRACTICAL CHALLENGE

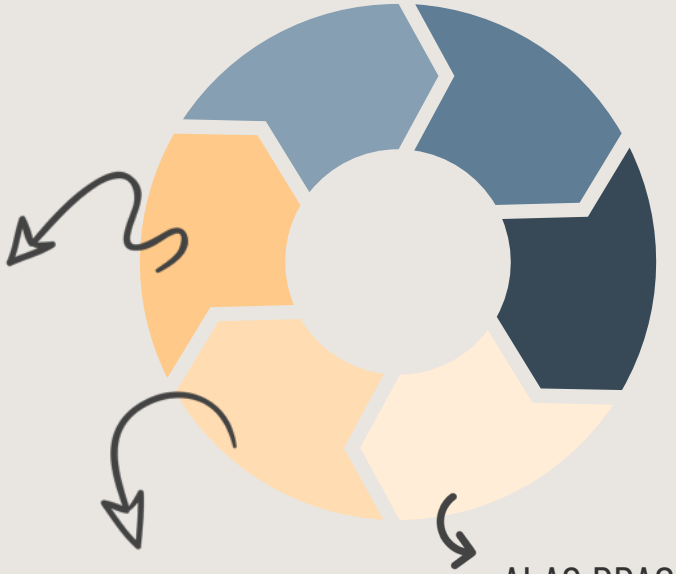
DECISION SUPPORT SYSTEM



IMPACT ON SOCIETY

AI AS PRACTICAL CHALLENGE

DECISION SUPPORT SYSTEM



ML STATIC OBJECTIVES

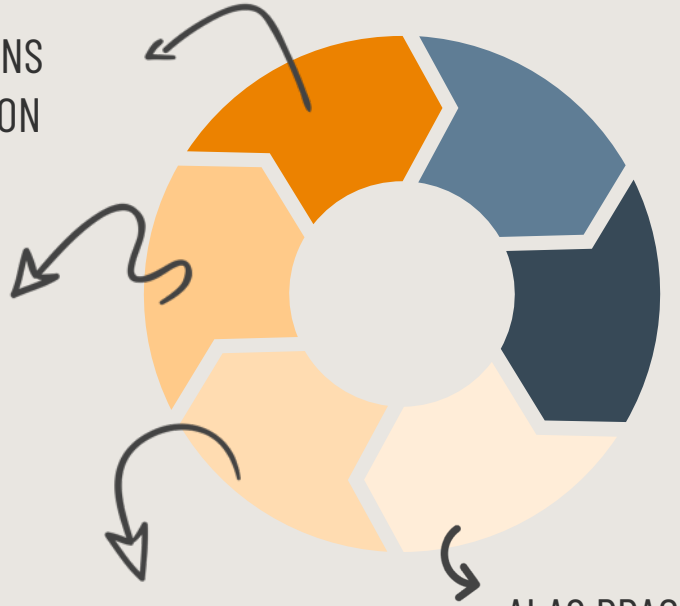
IMPACT ON SOCIETY

AI AS PRACTICAL CHALLENGE

RESEARCH QUESTIONS

DECISION SUPPORT SYSTEM

CONSEQUENTIAL DECISIONS RESHAPE THE POPULATION



ML STATIC OBJECTIVES

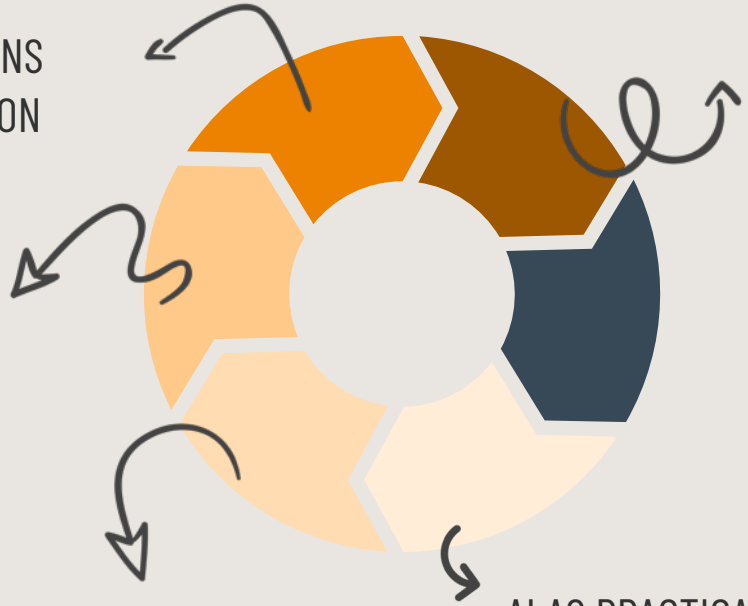
AI AS PRACTICAL CHALLENGE

IMPACT ON SOCIETY

DECISION SUPPORT SYSTEM

CONSEQUENTIAL DECISIONS RESHAPE THE POPULATION

LONG-TERM FAIRNESS STILL UNEXPLORED



ML STATIC OBJECTIVES

AI AS PRACTICAL CHALLENGE

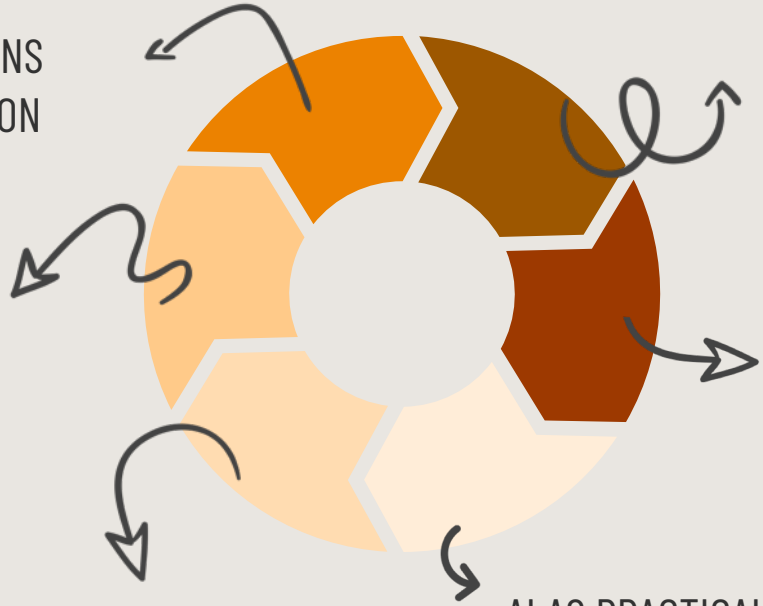
IMPACT ON SOCIETY

RESEARCH QUESTIONS

DECISION SUPPORT SYSTEM

CONSEQUENTIAL DECISIONS RESHAPE THE POPULATION

LONG-TERM FAIRNESS STILL UNEXPLORED



ML STATIC OBJECTIVES

INDIVIDUAL DECISION-MAKING DYNAMICS CAN AFFECT THE EFFECTIVENESS OF A POLICY

IMPACT ON SOCIETY

AI AS PRACTICAL CHALLENGE

RESEARCH QUESTIONS

- HOW TO CHOOSE THE BEST POLICY TO ENSURE LONG-TERM FAIRNESS?
- DO THE FAIRNESS CONSTRAINTS KEEP THEIR VALIDITY FOR AS LONG AS THEY ACT?

DETECTING DISCRIMINATORY RISK THROUGH DATA ANNOTATION BASED ON BAYESIAN INFERENCES

PRINCIPAL TECHNIQUES/CONCEPTS

- BAYESIAN INFERENCE
- BIAS AND FAIRNESS IN DATA AND ML

CONTENT

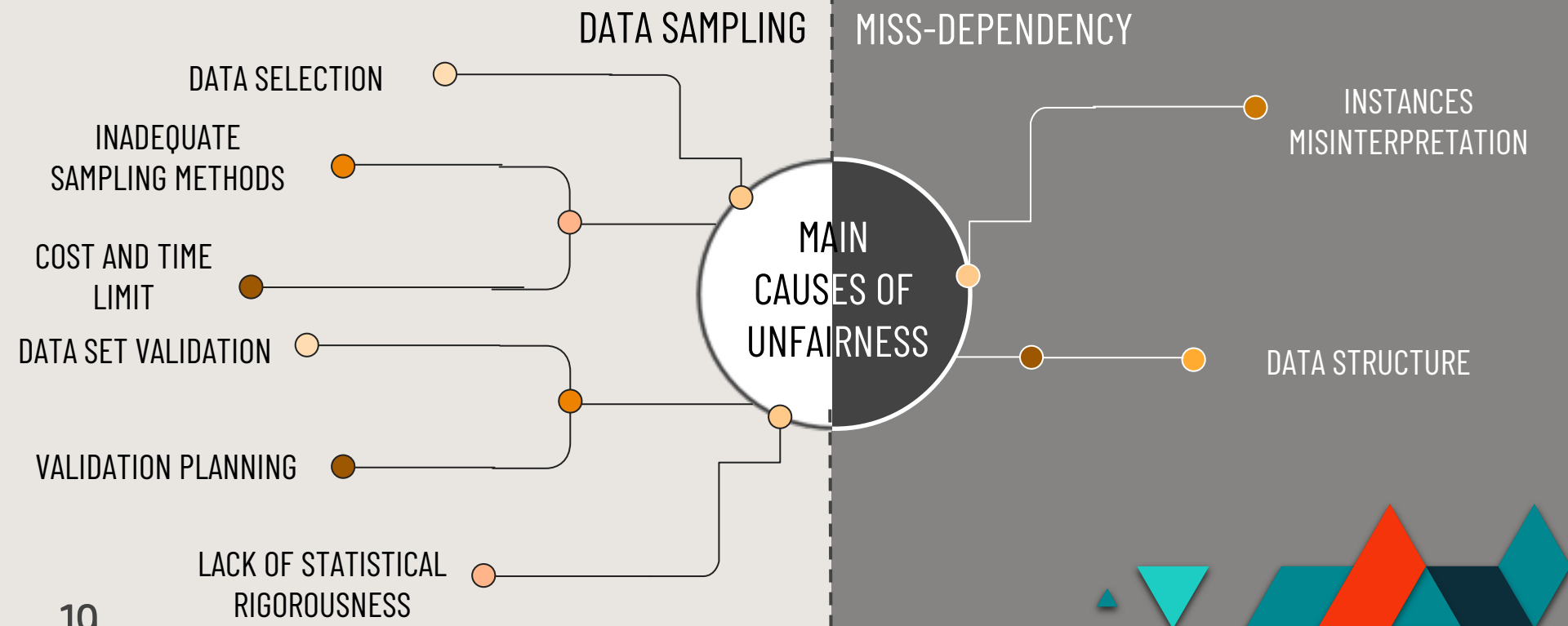
A METHOD OF DATA ANNOTATION BASED ON BAYESIAN STATISTICAL INFERENCE THAT AIMS TO WARN ABOUT THE RISK OF DISCRIMINATORY RESULTS OF A GIVEN DATA SET

RESEARCH QUESTIONS

IS IT POSSIBLE TO ESTABLISH THE A PRIORI PROBABILITY OF THE TRAINING DATA DISTRIBUTION FROM THE AVAILABLE DATASET?

HOW TRAINING DATA COULD INFORM ABOUT THE RISK OF FUTURE DISCRIMINATION?

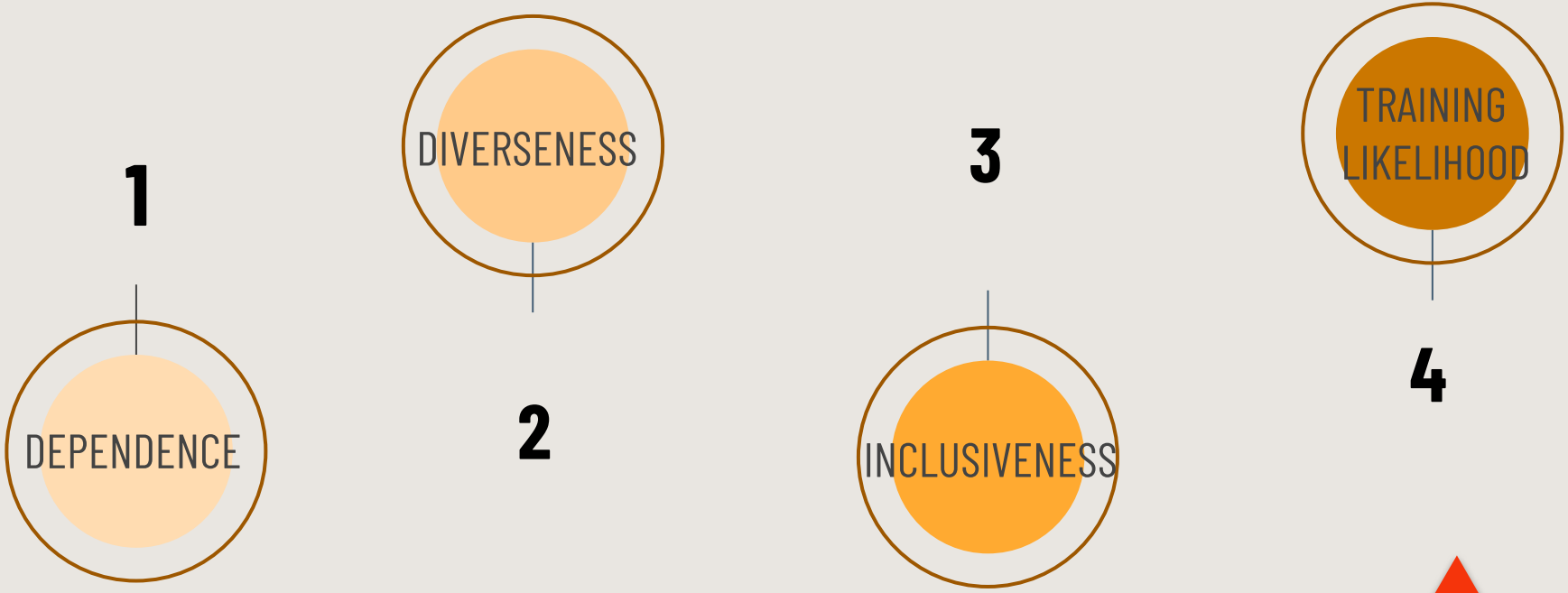
BACKGROUND


DATA SAMPLING
MISS-DEPENDENCY
DATA SELECTION
INADEQUATE SAMPLING METHODS
COST AND TIME LIMIT
DATA SET VALIDATION
VALIDATION PLANNING
LACK OF STATISTICAL RIGOROUSNESS
MAIN CAUSES OF UNFAIRNESS
INSTANCES MISINTERPRETATION
DATA STRUCTURE

RESEARCH GOAL

A DIAGNOSTIC FRAMEWORK TO WARN ABOUT THE RISK OF DISCRIMINATORY RESULTS

THEORETICAL MODEL



METHDOLOGY

DEGREE OF CONNECTION

METHDOLOGY

SIMPLIFIED INTERPRETATION OF THE DEPENDENCY

DEGREE OF CONNECTION

METHDOLOGY

SIMPLIFIED INTERPRETATION OF THE DEPENDENCY

DEGREE OF CONNECTION

EFFECT SIZE INDEX w

MAGNITUDE	VALUE
SMALL	w = 0.1
MEDIUM	w = 0.3
LARGE	w = 0.5

METHDOLOGY

1

DEPENDENCE

DIVERSENESS

2

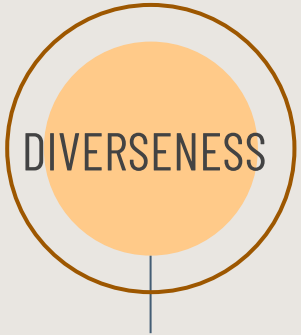
3

INCLUSIVENESS

TRAINING LIKELIHOOD

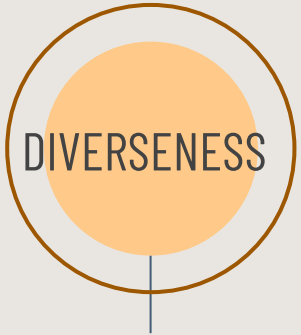
4

METHDOLOGY



TRAINING DIVERSIFICATION PROBABILITY

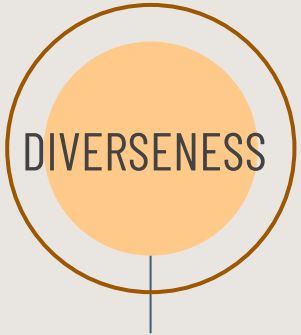
METHDOLOGY



HOW LIKELY A PROPERTY WILL OCCUR

TRAINING DIVERSIFICATION PROBABILITY

METHDOLOGY



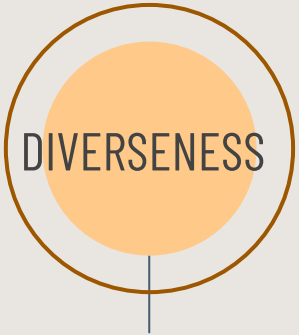
TRAINING DIVERSIFICATION PROBABILITY

HOW LIKELY A PROPERTY WILL OCCUR

FORMULA	PROBABILITY
$P(Y = 0)$ $P(Y = 1)$	$P = 0.48$ $P = 0.52$
$P(A = \text{white})$ $P(A = \text{black})$ $P(A = \text{Asian})$	$P = 0.6$ $P = 0.35$ $P = 0.15$

METHDOLOGY

DEGREE OF CONNECTION



PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET

TRAINING DIVERSIFICATION PROBABILITY

OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS BEFORE THE TRAINING SET IS SAMPLED

METHDOLOGY

PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET

METHDOLOGY

PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET

THE PROBABILITY THAT THE TRAINING SET SIMULTANEOUSLY SHOWS THE PROPERTY
Y = y AND A = a

METHDOLOGY

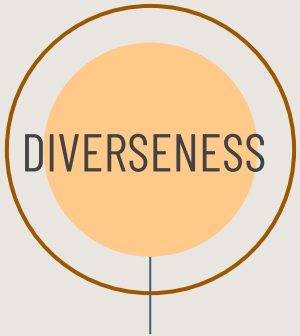
PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET

THE PROBABILITY THAT THE TRAINING SET SIMULTANEOUSLY SHOWS THE PROPERTY
Y = y AND A = a

FORMULA	PROBABILITY
$P(Y=0 \cap A=white)$	P=0.42
$P(Y=0 \cap A=black)$	P=0.07
$P(Y=0 \cap A=Asian)$	P=0.09
$P(Y=1 \cap A=white)$	P=0.18
$P(Y=1 \cap A=black)$	P=0.28
$P(Y=1 \cap A=Asian)$	P=0.06

METHDOLOGY

DEGREE OF CONNECTION



TRAINING DIVERSIFICATION PROBABILITY

PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET

OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS BEFORE THE TRAINING SET IS SAMPLED

METHDOLOGY

OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS BEFORE THE TRAINING SET IS SAMPLED

METHDOLOGY

OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS BEFORE THE TRAINING SET IS SAMPLED

i) WHAT IS THE PROBABILITY OF BELONGING TO AN ETHNIC GROUP WITH RESPECT TO THE OUTCOME VARIABLE?

ii) WHAT IS THE PROBABILITY OF OBTAINING A CERTAIN OUTCOME WITH RESPECT TO THE ETHNIC GROUP?

METHDOLOGY

OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS BEFORE THE TRAINING SET IS SAMPLED

i) WHAT IS THE PROBABILITY OF BELONGING TO AN ETHNIC GROUP WITH RESPECT TO THE OUTCOME VARIABLE?

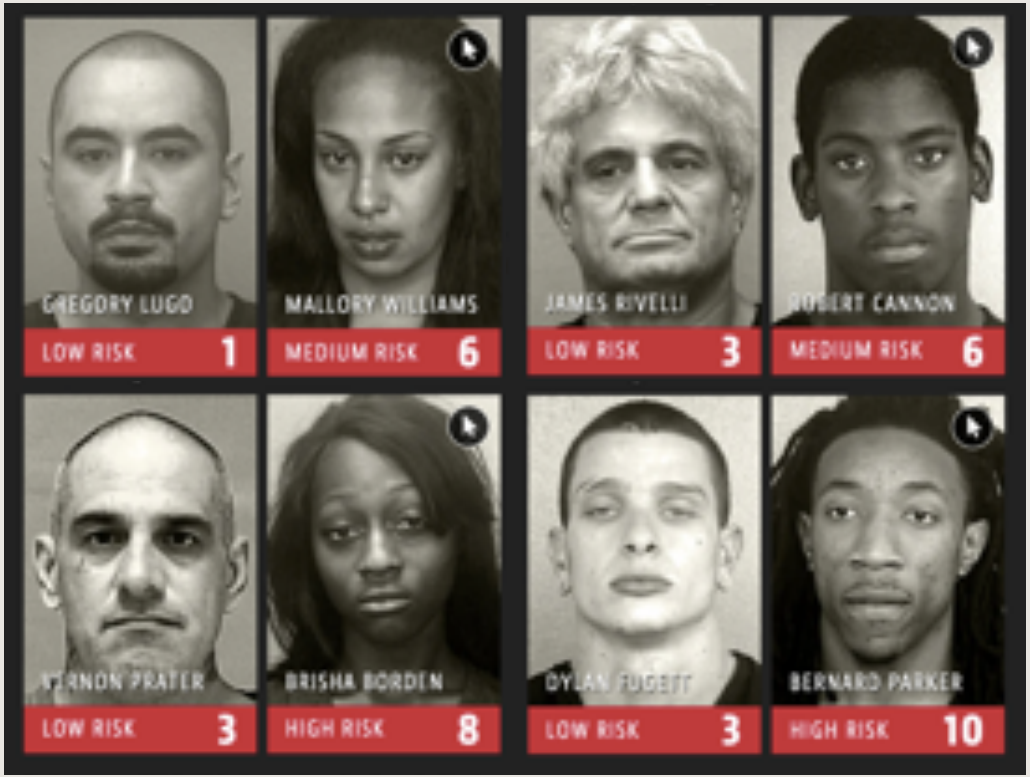
ii) WHAT IS THE PROBABILITY OF OBTAINING A CERTAIN OUTCOME WITH RESPECT TO THE ETHNIC GROUP?

FORMULA	PROBABILITY
$P(Y=0\|A=white)$	P=0.7
$P(Y=0\|A=black)$	P=0.2
$P(Y=0\|A=Asian)$	P=0.6
$P(Y=1\|A=white)$	P=0.3
$P\ (Y=1\|A=black)$	P=0.8
$P\ (Y=1\|A=Asian)$	P=0.4
$P(A=white\|Y=1)$	P=0.34
$P(A=white\|Y=0)$	P=0.87
$P(A=black\|Y=1)$	P=0.53
$P(A=black\|Y=0)$	P=0.15
$P(A=Asian\|Y=1)$	P=0.11
$P(A=Asian\|Y=0)$	P=0.18

DATASETS PROMINENT PROPERTIES

	COMPAS	DRUG CONSUMPTION	ADULT CENSUS DATASET
SIZE	6172x9	1885x31	48842x15
TARGET VARIABLE	0 ⟶ NO	0 ⟶ NON USER	0 ⟶ > 50K
	1 ⟶ YES	1 ⟶ USER	1 ⟶ ≤ 50K
LEVELS OF ETHNICITY ATTRIBUTE	ASIAN	ASIAN BLACK	AMERICAN-INDIAN/ESKIMO
	BLACK	BLACK/ASIAN	ASIAN-PAC-ISLANDER
	CAUCASIAN	CAUCASIAN	BLACK
	HISPANIC	WHITE/ASIAN	CAUCASIAN
	NATIVE AMERICAN	WHITE/BLACK	OTHER
	OTHER	OTHER	

VALIDATION



COMPAS

CORRECTIONAL OFFENDER MANAGEMENT PROFLING FOR ALTERNATIVE SANCTIONS

1 = RECIDIVE
0 = NOT RECIDIVE

BLACK PEOPLE = MORE LIKELY TO BE LABELED AS RECIDIVISTS

WHITE PEOPLE = UNDERESTIMATED RISK OF RECIDIVISM

VALIDATION

COMPAS

CORRECTIONAL OFFENDER MANAGEMENT PROFLING FOR ALTERNATIVE SANCTIONS

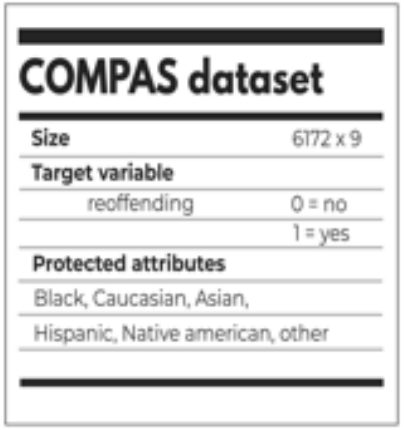
1 = RECIDIVE
0 = NOT RECIDIVE

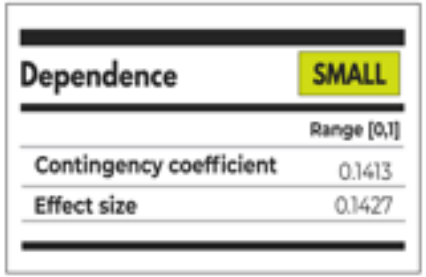
BLACK PEOPLE = MORE LIKELY TO BE LABELED AS RECIDIVISTS

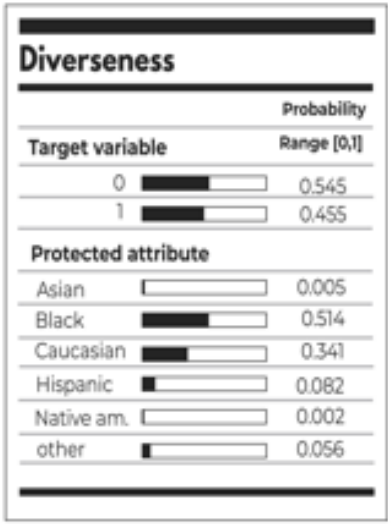
WHITE PEOPLE = UNDERESTIMATED RISK OF RECIDIVISM

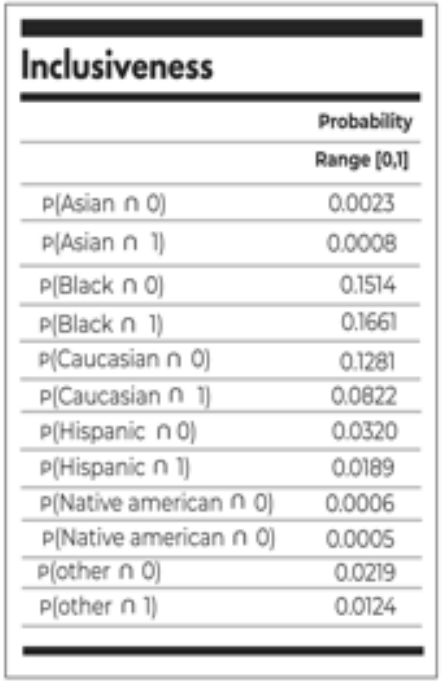
RESULTS

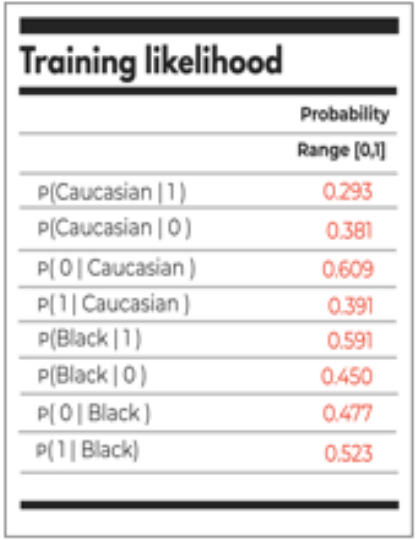
GRAPHICAL VISUALIZATION OF THE DATA ANNOTATION SYSTEM



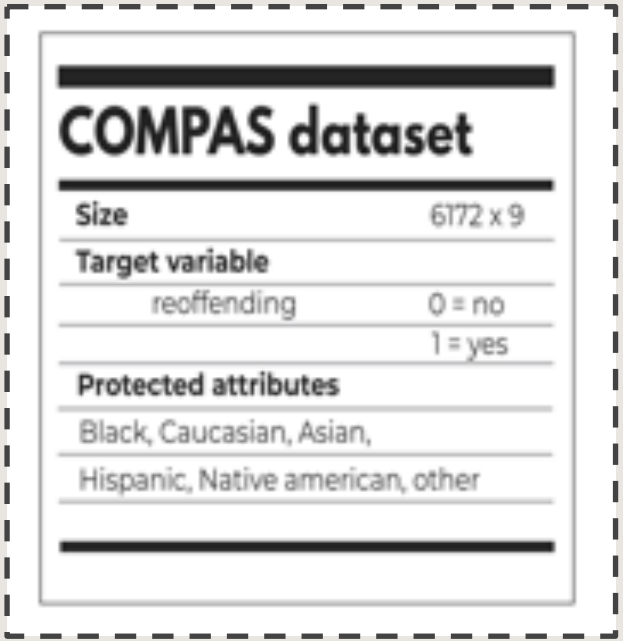








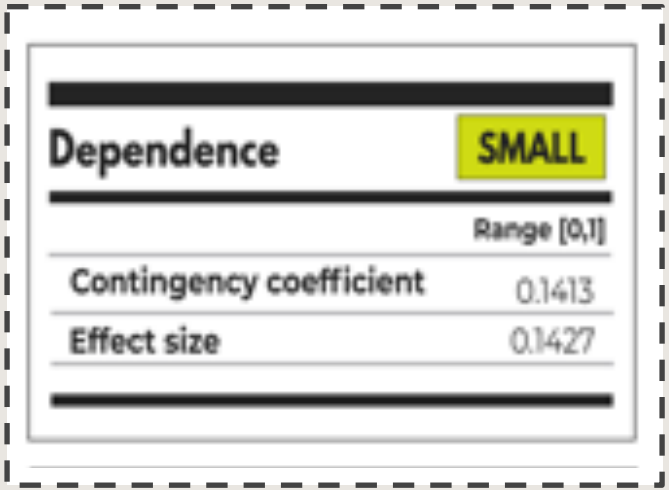
RESULTS



COMPAS dataset	
Size	6172 x 9
Target variable	
reoffending	0 = no
	1 = yes
Protected attributes	
Black, Caucasian, Asian,	
Hispanic, Native american, other	

DATASET PROMINENT PROPERTIES

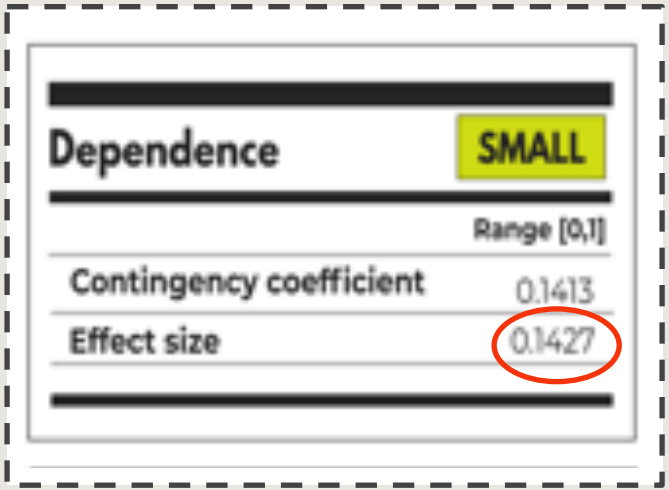
RESULTS



Dependence	SMALL
	Range [0,1]
Contingency coefficient	0.1413
Effect size	0.1427

DEPENDENCY RELATIONSHIPS BETWEEN RECIDIVISM AND ETHNIC PROPERTY

RESULTS

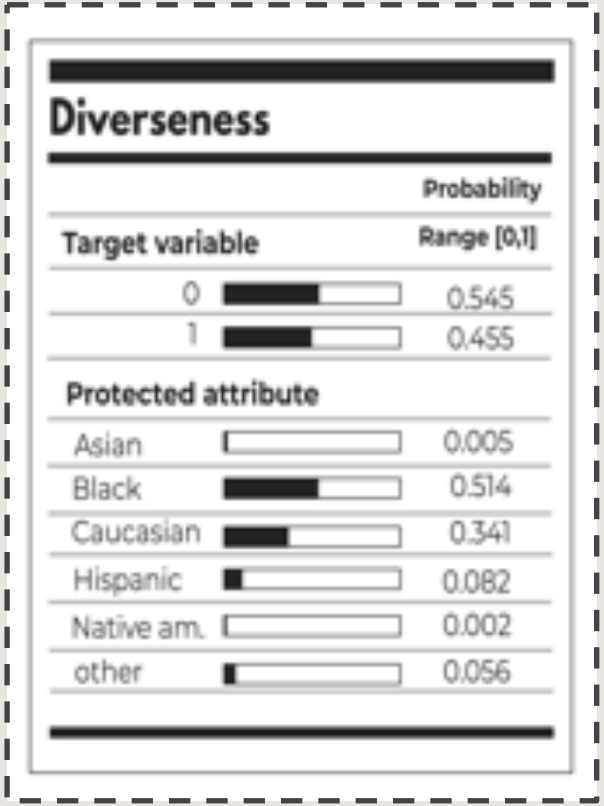


Dependence	SMALL
	Range [0,1]
Contingency coefficient	0.1413
Effect size	0.1427

DEPENDENCY RELATIONSHIPS BETWEEN RECIDIVISM AND ETHNIC PROPERTY

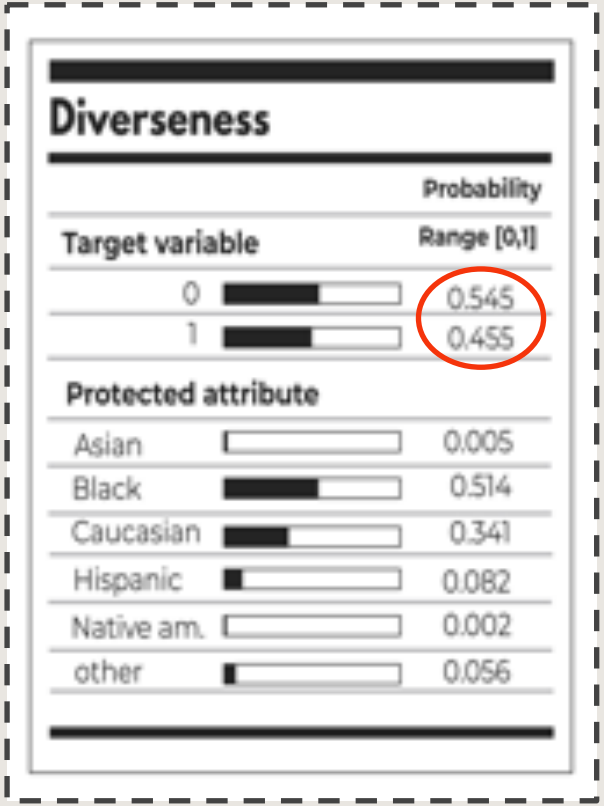
SMALL DEPENDENCE

RESULTS



THE PROBABILITY THAT THE TRAINING SET WILL BE EQUALLY COMPOSED BY ETHNIC MINORITIES AND ETHNIC MAJORITIES

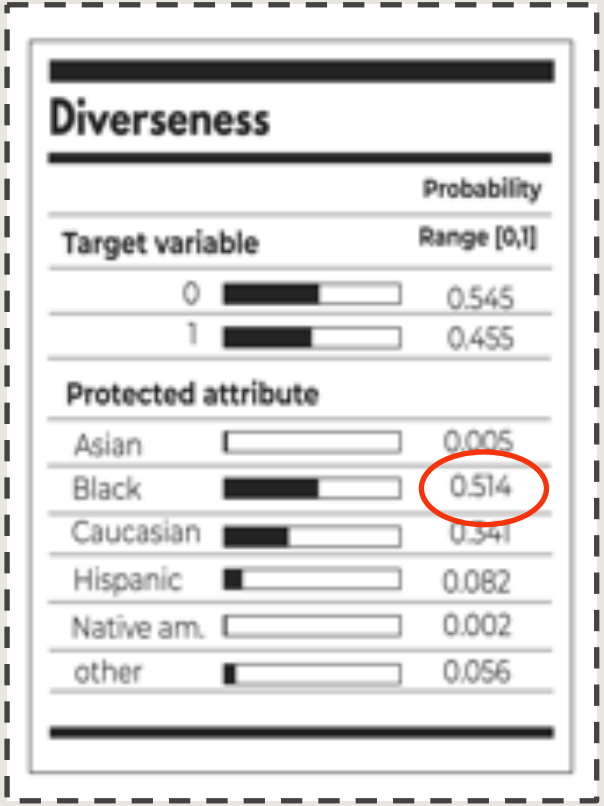
RESULTS



THE PROBABILITY THAT THE TRAINING SET WILL BE EQUALLY COMPOSED BY ETHNIC MINORITIES AND ETHNIC MAJORITIES

TARGET VARIABLE EQUALLY DISTRIBUTED

RESULTS



THE PROBABILITY THAT THE TRAINING SET WILL BE EQUALLY COMPOSED BY ETHNIC MINORITIES AND ETHNIC MAJORITIES

TARGET VARIABLE EQUALLY DISTRIBUTED

PREVALENCE OF BLACK PEOPLE

RESULTS

Inclusiveness

	Probability Range [0,1]
P(Asian ∩ 0)	0.0023
P(Asian ∩ 1)	0.0008
P(Black ∩ 0)	0.1514
P(Black ∩ 1)	0.1661
P(Caucasian ∩ 0)	0.1281
P(Caucasian ∩ 1)	0.0822
P(Hispanic ∩ 0)	0.0320
P(Hispanic ∩ 1)	0.0189
P(Native american ∩ 0)	0.0006
P(Native american ∩ 0)	0.0005
P(other ∩ 0)	0.0219
P(other ∩ 1)	0.0124

THE PROBABILITY THAT THE TRAINING SET WILL BE EQUALLY COMPOSED BY ETHNIC MINORITIES AND ETHNIC MAJORITIES SHOWING SIMILAR TARGET LEVELS

HIGHER LIKELIHOOD THAT CAUCASIANS WILL NOT RECIDIVATE THAN THAT THEY WILL RECIDIVATE

RESULTS

Training likelihood

	Probability Range [0,1]
P(Caucasian | 1)	0.293
P(Caucasian | 0)	0.381
P(0 | Caucasian)	0.609
P(1 | Caucasian)	0.391
P(Black | 1)	0.591
P(Black | 0)	0.450
P(0 | Black)	0.477
P(1 | Black)	0.523

THE PROBABILITY THAT THE OCCURRENCE OF REOFFENDING IS GIVEN BY THE PROPERTIES OF THE PROTECTED ATTRIBUTE ETHNICITY

RESULTS

Training likelihood

	Probability Range [0,1]
P(Caucasian | 1)	0.293
P(Caucasian | 0)	0.381
P(0 | Caucasian)	0.609
P(1 | Caucasian)	0.391
P(Black | 1)	0.591
P(Black | 0)	0.450
P(0 | Black)	0.477
P(1 | Black)	0.523

THE PROBABILITY THAT THE OCCURRENCE OF REOFFENDING IS GIVEN BY THE PROPERTIES OF THE PROTECTED ATTRIBUTE ETHNICITY

GIVEN AS VERIFIED THE RECIDIVIST PROPERTY, THE PROBABILITY THAT THE INDIVIDUAL IS BLACK IS MUCH HIGHER THAN THE PROBABILITY THAT THE INDIVIDUAL IS WHITE

RESULTS

Training likelihood

	Probability Range [0,1]
P(Caucasian | 1)	0.293
P(Caucasian | 0)	0.381
P(0 | Caucasian)	0.609
P(1 | Caucasian)	0.391
P(Black | 1)	0.591
P(Black | 0)	0.450
P(0 | Black)	0.477
P(1 | Black)	0.523

THE PROBABILITY THAT THE OCCURRENCE OF REOFFENDING IS GIVEN BY THE PROPERTIES OF THE PROTECTED ATTRIBUTE ETHNICITY

GIVEN AS VERIFIED THE RECIDIVIST PROPERTY, THE PROBABILITY THAT THE INDIVIDUAL IS BLACK IS MUCH HIGHER THAN THE PROBABILITY THAT THE INDIVIDUAL IS WHITE

GIVEN AS VERIFIED THE NOT RECIDIVIST PROPERTY, THE PROBABILITY THAT THE INDIVIDUAL IS BLACK IS LOWER THAN THE PROBABILITY THAT THE INDIVIDUAL IS WHITE

RESULTS

- CLASSICAL SAMPLING VS MACHINE LEARNING PRACTISES
- REAL POPULATION VS AVAILABLE DATA
- THE STRUCTURE OF THE DATA AFFECTS THE PROBABILITY OF PROPERTIES DISTRIBUTION

AFteRS: AN AUTOMATED FAIR-DISTRIBUTIVE RANKING SYSTEM FOR SOCIAL JUSTICE IN AI

PRINCIPAL TECHNIQUES/CONCEPTS

- RANKING SYSTEMS
- EQUALITY OF OPPORTUNITY
- DISITRIBUTIVE JUSTICE

CONTENT

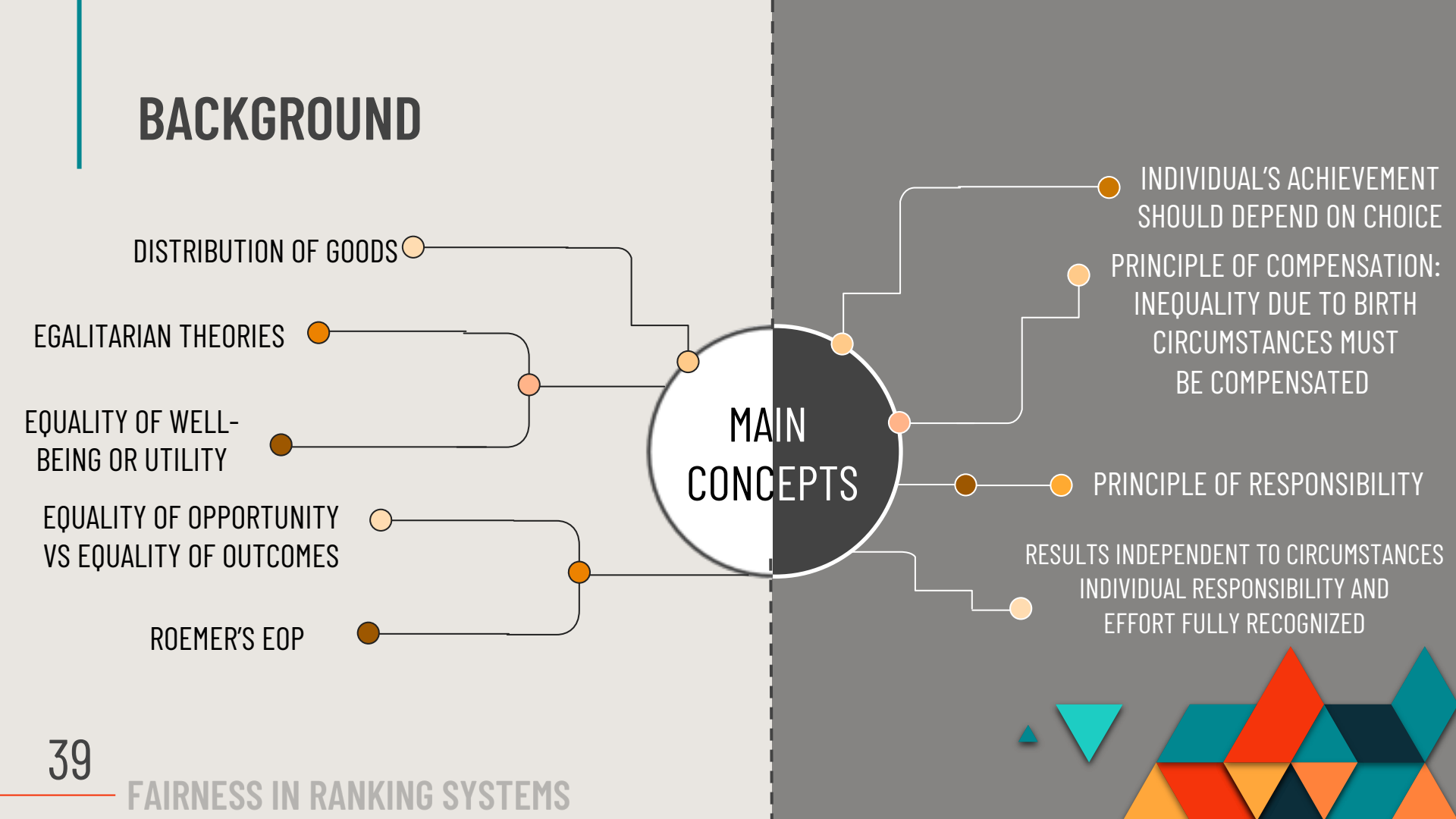
AN AUTOMATED FAIR-DISTRIBUTIVE RANKING SYSTEM

RESEARCH QUESTIONS

ARE RANKING SYSTEMS BASED ON A DISTRIBUTIVE FAIRNESS CONSTRAINT ABLE TO PRESERVE THE ACCURACY OF THE RANKING AND THE MODEL'S OVERALL UTILITY BY PROVIDING A RANKING OF THE BEST CANDIDATES?

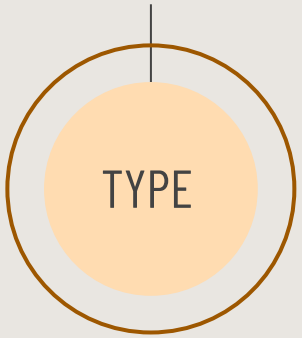
WHAT ARE THE FACTORS AFFECTING THE FAIRNESS UTILITY TRADE-OFF IN A FAIRNESS CONSTRAINED RANKING SYSTEM?

BACKGROUND


DISTRIBUTION OF GOODS
EGALITARIAN THEORIES
EQUALITY OF WELL-BEING OR UTILITY
EQUALITY OF OPPORTUNITY VS EQUALITY OF OUTCOMES
ROEMER'S EOP
MAIN CONCEPTS
INDIVIDUAL'S ACHIEVEMENT SHOULD DEPEND ON CHOICE
PRINCIPLE OF COMPENSATION: INEQUALITY DUE TO BIRTH CIRCUMSTANCES MUST BE COMPENSATED
PRINCIPLE OF RESPONSIBILITY
RESULTS INDEPENDENT TO CIRCUMSTANCES INDIVIDUAL RESPONSIBILITY AND EFFORT FULLY RECOGNIZED

THEORETICAL MODEL

GROUP SAHRING
SAME CIRCUMSTANCES



SAME CHOICES
DIFFERENT CIRCUMSTANCES
DIFFERENT OTUCOME

NEUTRAL TOWARDS INEQUALITIES
WITHIN TYPES

RESEARCH GOAL

AN AUTOMATED FAIR-DISTRIBUTIVE RANKING SYSTEM

1. THE BEST TOP-N-RANKING IN A SET OF CANDIDATES
2. MAXIMIZING UTILITY AND SATISFYING FAIRNESS CONSTRAINTS

METHODOLOGY

MORAL THEORY

EQUALITY OF OPPORTUNITY AND DISTRIBUTIVE JUSTICE

CIRCUMSTANCES

VARIABLES DESCRIBING INDIVIDUALS

FAIRNESS CONSTRAINTS

A SET OF POLICY: EQUITY, EQUALITY, NEED

TYPE

POPULATION PARTITIONED IN A SERIES OF NON-OVERLAPING SETS

EFFORT

DEGREE OF EFFORT PEOPLE EXERT TO ACCOMPLISH A TASK

METHODOLOGY

01

TYPES ESTIMATE

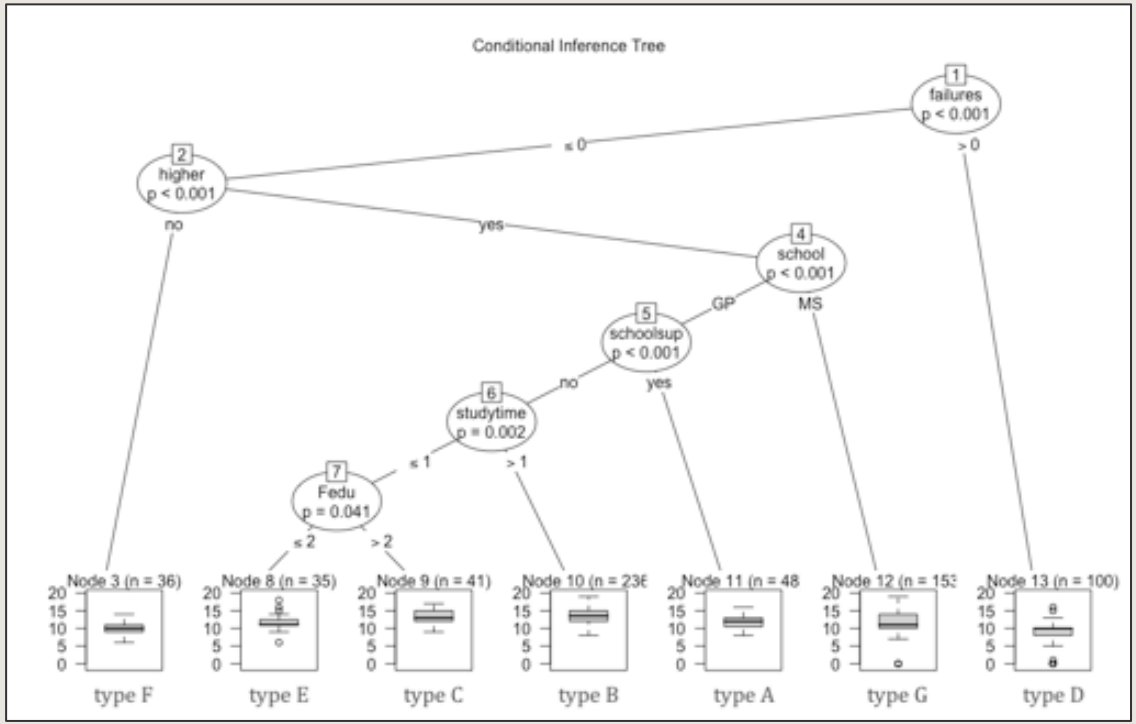
$$T_k = \begin{cases} S_i^X & if\ H_0^i: P(Y|X_i) = P(Y) \\ recursion\ stops & otherwise \end{cases}$$

S_i^X = set of all x^i possible realizations

METHODOLOGY

01

TYPES ESTIMATE



METHODOLOGY

01

TYPES ESTIMATE

$$T_k = \begin{cases} S_i^X & if\ H_0^i: P(Y|X_i) = P(Y) \\ recursion\ stops & otherwise \end{cases}$$

02

EFFORT ESTIMATE

$$CDF_{type}(\lambda)$$

(λ)= quantile of the Cumulative Distribution Function

METHODOLOGY

01

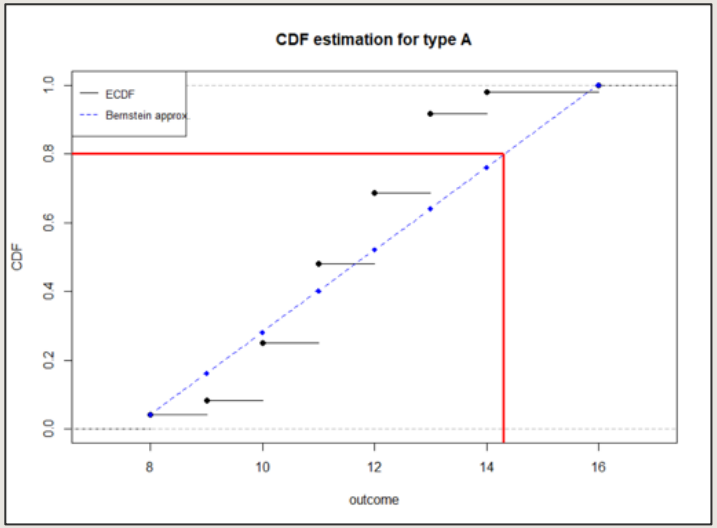
TYPES ESTIMATE

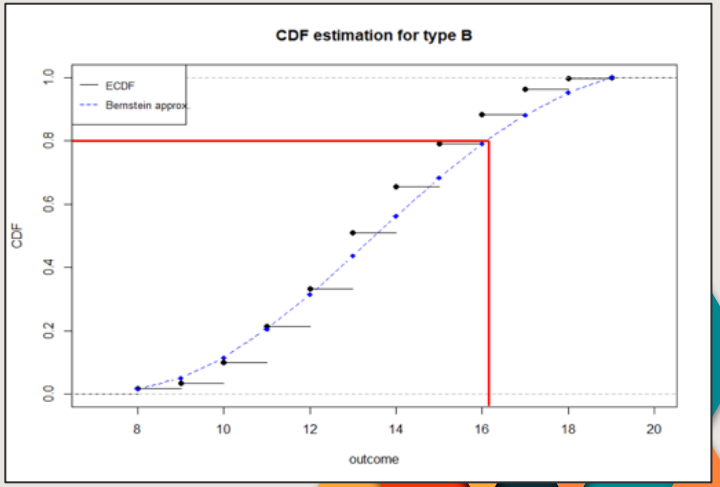
$$T_k = \begin{cases} S_i^X & if\ H_0^i: P(Y|X_i) = P(Y) \\ recursion\ stops & otherwise \end{cases}$$

02

EFFORT ESTIMATE

$$CDF_{type}(\lambda)$$





METHODOLOGY

01

TYPES ESTIMATE

$$T_k = \begin{cases} S_i^X & if\ H_0^i: P(Y|X_i) = P(Y) \\ recursion\ stops & otherwise \end{cases}$$

02

EFFORT ESTIMATE

$CDF_{type}(\lambda)$

03

UNFAIRNESS DEGREE ESTIMATE

UNFAIRNESS= $Gini\ Index\ \sum CDF_{type}(\lambda)$

(λ)= quantile of the Cumulative Distribution Function

METHODOLOGY

01

TYPES ESTIMATE

$$T_k = \begin{cases} S_i^X & if\ H_0^i: P(Y|X_i) = P(Y) \\ recursion\ stops & otherwise \end{cases}$$

02

EFFORT ESTIMATE

$CDF_{type}(\lambda)$

03

UNFAIRNESS DEGREE ESTIMATE

UNFAIRNESS= $Gini\ Index \sum CDF_{type}(\lambda)$

$$y_{\rightarrow} = f(CDF_{type}(\lambda), \sum Gini\ Index(CDF_{type}(\lambda)$$

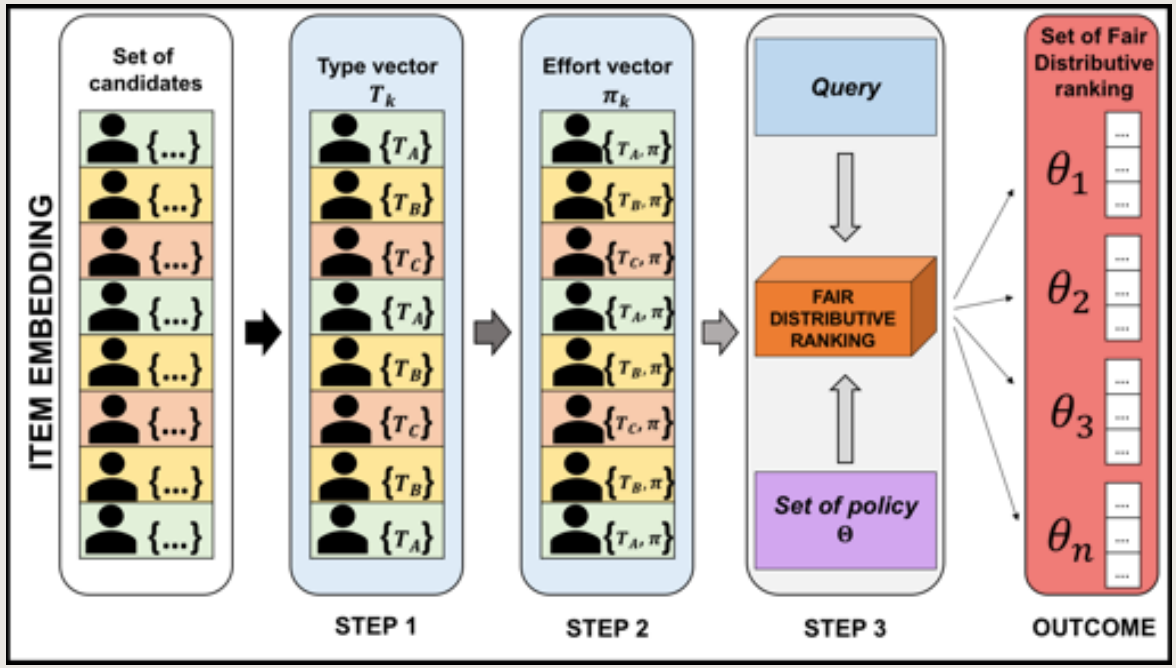
04

OUTCOME

METHODOLOGY

	EQUITY (ROEMER EOp)	EQUALITY	NEED
01 TYPES ESTIMATE	✅	✅	❌
02 EFFORT ESTIMATE	✅	❌	❌
03 UNFAIRNESS DEGREE ESTIMATE	✅	❌	❌
04 OUTCOME	THE OUTCOME INDIVIDUALS WOULD HAVE GOTTEN IF THEY HAD NOT BELONGED TO THEIR TYPE	DEMOGRAPHIC PARITY ON TYPES	DEMOGRAPHIC PARITY ON PROTECTED ATTRIBUTES

APPLICATION SETTING



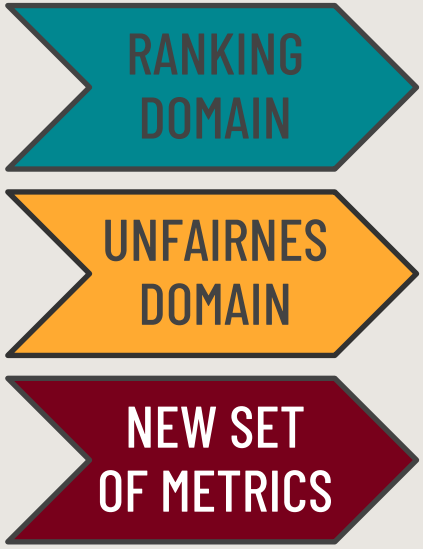
DATA: STUDENT PERFORMANCE DATASET

SCENARIO: HYPOTHETICAL SCENARIO OF A UNIVERSITY SELECTION PROCESS IN WHICH THE DECISION-MAKER DETERMINES WHICH STUDENTS ARE SUITABLE ON THE BASIS OF THEIR PERSONAL QUALIFICATIONS AND ACHIEVEMENTS, SO AS TO MAXIMIZE THE INSTITUTION UTILITY

SYSTEM'S GOAL:

$$\Gamma = maxmin(utiliy, unfairness)$$

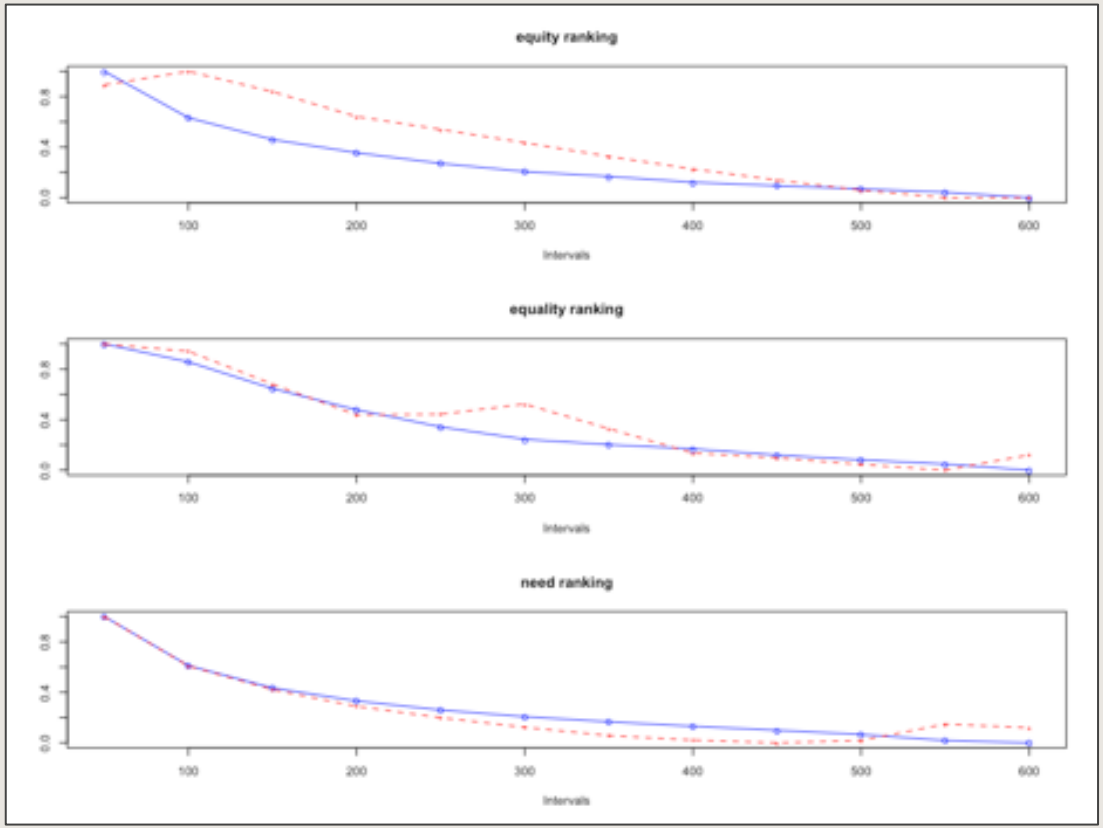
VALIDATION



VALIDATION


RANKING DOMAIN
EXPOSURE
RELEVANCE
EXPECTED RANKING
UNFAIRNES DOMAIN
GINI
DIVERSITY
NEW SET OF METRICS
OPPORTUNITY TYPES RATE
OPPORTUNITY LOSS/GAIN RATE
ETC

RESULTS



UNFAIRNESS-UTILITY TRADE-OFF FOR RANKINGS UNDER FAIRNESS CONSTRAINTS

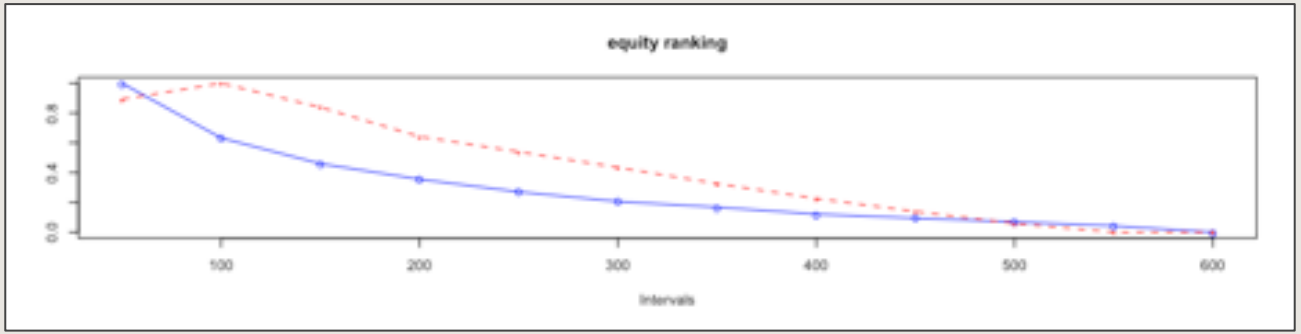
RED LINE: UNFAIRNESS

BLUE LINE: UTILITY

Y-AXIS: UTILITY AND INEQUALITY VALUES RANGING FROM 0 TO 1

X-AXIS: NUMEROSITY OF THE RANKINGS

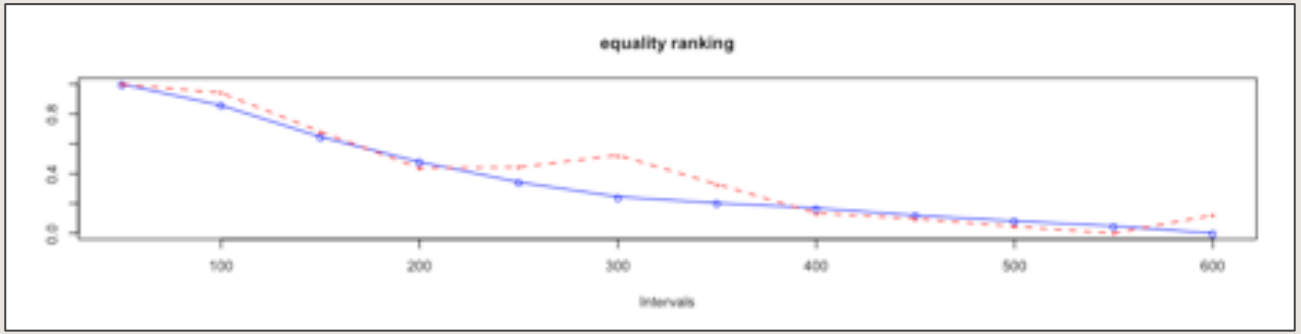
EQUITY POLICY: UNFAIRNESS-UTILITY TRADE-OFF



GOOD GENERAL UTILITY IN THE FIRST TOP-N-RANKING

HIGH LEVELS OF INEQUALITY

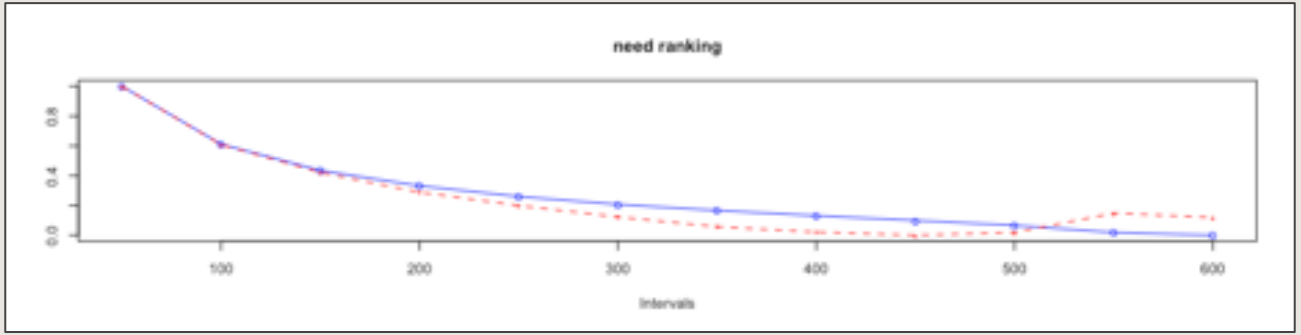
EQUALITY POLICY: UNFAIRNESS-UTILITY TRADE-OFF



LOW LEVELS OF INEQUALITY

LOW GENERAL UTILITY IN THE FIRST TOP-N-RANKING

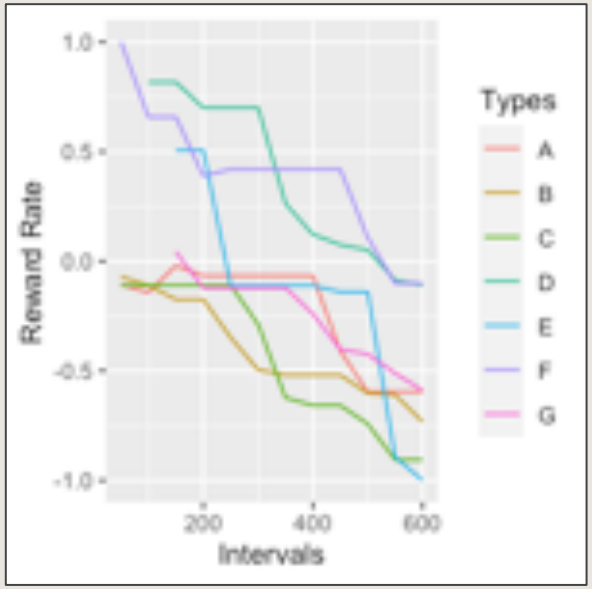
NEED POLICY: UNFAIRNESS-UTILITY TRADE-OFF



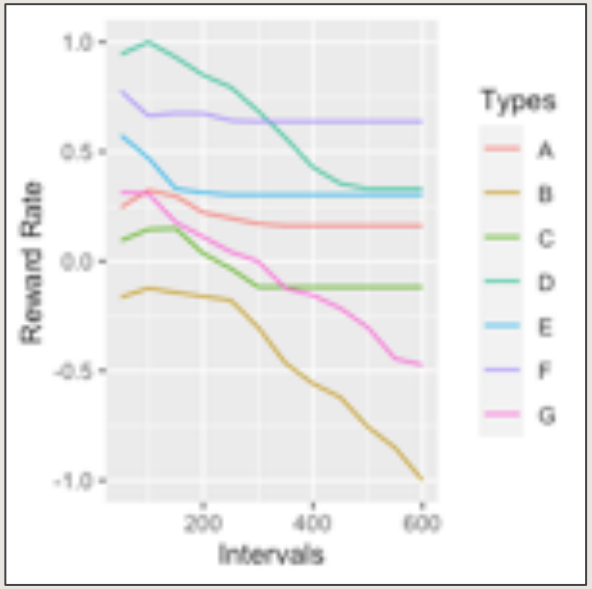
UNIFORM LEVELES OF INEQUALITY AND UTILITY

OVERALL WORST PERFORMANCES

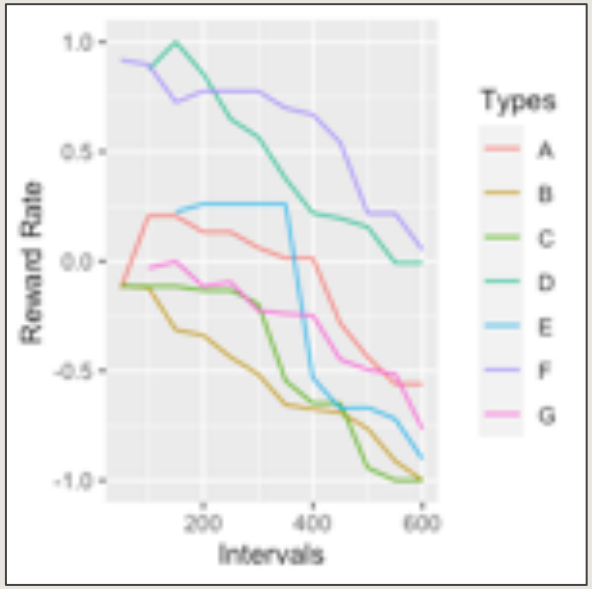
REWARD RATE



EQUITY POLICY



EQUALITY POLICY



NEED POLICY

RESULTS

- GROUPS VS INDIVIDUAL
- EQUITY AND UTILITY RESULTS ARE POLICY DEPENDENT
- EQUITY POLICY: BEST CHOICE FOR NUMEROUS RANKING
- EQUALITY POLICY: BEST CHOICE FOR LESS DENSE RANKING
- THE MORAL GROUND IS CONTEXT DEPENDENT

A DECISION SUPPORT SYSTEM FOR LONG-TERM FAIRNESS

PRINCIPAL TECHNIQUES/CONCEPTS

- DECISION THEORY
- LONG-TERM FAIRNESS FOR CLASSIFICATION SYSTEMS

CONTENT

A DECISION SUPPORT SYSTEM TO ENSURE LONG-TERM FAIRNESS IN MACHINE LEARNING SYSTEMS

RESEARCH QUESTIONS

HOW TO CHOOSE THE BEST POLICY TO ENSURE LONG-TERM FAIRNESS?

DO THE FAIRNESS CONSTRAINTS KEEP THEIR VALIDITY FOR AS LONG AS THEY ACT?

RESEARCH GOAL

A DECISION SUPPORT SYSTEM TO ENSURE LONG-TERM FAIRNESS

1. DECISION THEORY APPLIED TO ALGORTHMIC DECISION-MAKING
2. INDIVIDUAL DYNAMICS INTEGRATED IN THE MODEL

THEORETICAL MODEL

THEORETICAL MODEL

INSTITUTION'S UTILITY

$$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$$

g_A, g_B = fractions of the population

γ_1 = positive behavior

THEORETICAL MODEL

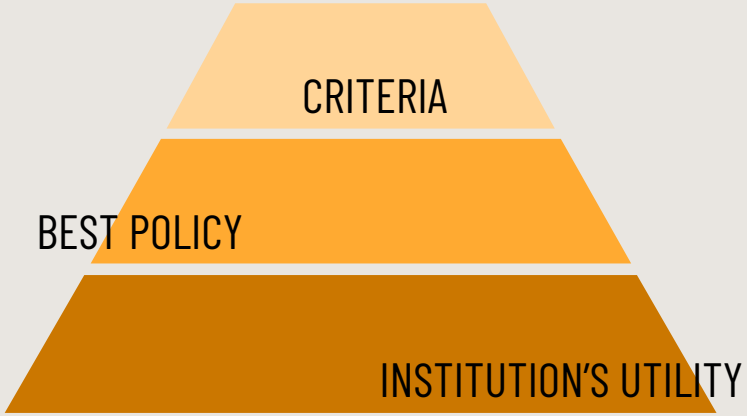
POPULATION IS INDUCED TO PERFORM A POSITIVE BEHAVIOR

$$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$$

g_A, g_B = fractions of the population

γ_1 = positive behavior

THEORETICAL MODEL



$$max \int_0^{\infty} \gamma_1(g_A) + \gamma_1(g_B)$$

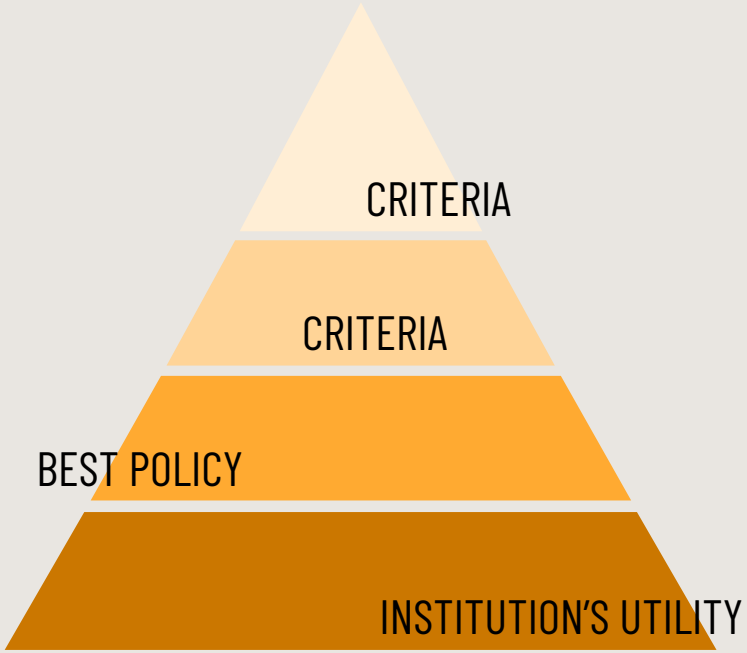
POPULATION IS INDUCED TO PERFORM A POSITIVE BEHAVIOR

$$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$$

g_A, g_B = fractions of the population

γ_1 = positive behavior

THEORETICAL MODEL



$$min \left| \int_0^{\infty} \gamma_1(g_A) - \gamma_1(g_B) \right|$$

$$max \int_0^{\infty} \gamma_1(g_A) + \gamma_1(g_B)$$

POPULATION IS INDUCED TO PERFORM A POSITIVE BEHAVIOR

$$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$$

g_A, g_B = fractions of the population

γ_1 = positive behavior

THEORETICAL MODEL

SOCIETY

g_A, g_B MINORITY AND MAJORITY GROUPS

γ_1 POSITIVE BEHAVIOR

$\theta \in \Theta$ SET OF INDIVIDUALS' ATTRIBUTES

$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$ INSTITUTION'S UTILITY

$max \int_0^\infty \gamma_1(g_A) + \gamma_1(g_B)$ MAXIMIZING DOMINANCE

$min \left| \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \right|$ MINIMIZING DOMINANCE AMONG GROUPS

THEORETICAL MODEL

SOCIETY

INDIVIDUAL DYNAMICS

g_A, g_B MINORITY AND MAJORITY GROUPS

ALTERNATIVES X

γ_1 POSITIVE BEHAVIOR

$\theta \in \Theta$ SET OF INDIVIDUALS' ATTRIBUTES

$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$ INSTITUTION'S UTILITY

$max \int_0^\infty \gamma_1(g_A) + \gamma_1(g_B)$ MAXIMIZING DOMINANCE

$min \left| \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \right|$ MINIMIZING DOMINANCE AMONG GROUPS

THEORETICAL MODEL

SOCIETY		INDIVIDUAL DYNAMICS	
g_A, g_B	MINORITY AND MAJORITY GROUPS	ALTERNATIVES	X
γ_1	POSITIVE BEHAVIOR	SCENARIOS	Ω
$\theta \in \Theta$	SET OF INDIVIDUALS' ATTRIBUTES		
$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$	INSTITUTION'S UTILITY		
$max \int_0^\infty \gamma_1(g_A) + \gamma_1(g_B)$	MAXIMIZING DOMINANCE		
$min \left\| \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \right\|$	MINIMIZING DOMINANCE AMONG GROUPS		

THEORETICAL MODEL

	SOCIETY	INDIVIDUAL DYNAMICS	
g_A, g_B	MINORITY AND MAJORITY GROUPS	ALTERNATIVES	X
γ_1	POSITIVE BEHAVIOR	SCENARIOS	Ω
$\theta \in \Theta$	SET OF INDIVIDUALS' ATTRIBUTES	IMPACTS	F
$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$	INSTITUTION'S UTILITY		
$max \int_0^\infty \gamma_1(g_A) + \gamma_1(g_B)$	MAXIMIZING DOMINANCE		
$min \left\| \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \right\|$	MINIMIZING DOMINANCE AMONG GROUPS		

THEORETICAL MODEL

	SOCIETY	INDIVIDUAL DYNAMICS	
g_A, g_B	MINORITY AND MAJORITY GROUPS	ALTERNATIVES	X
γ_1	POSITIVE BEHAVIOR	SCENARIOS	Ω
$\theta \in \Theta$	SET OF INDIVIDUALS' ATTRIBUTES	IMPACTS	F
$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$	INSTITUTION'S UTILITY	UTILITY FUNCTION	f
$max \int_0^\infty \gamma_1(g_A) + \gamma_1(g_B)$	MAXIMIZING DOMINANCE		
$min \left\| \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \right\|$	MINIMIZING DOMINANCE AMONG GROUPS		

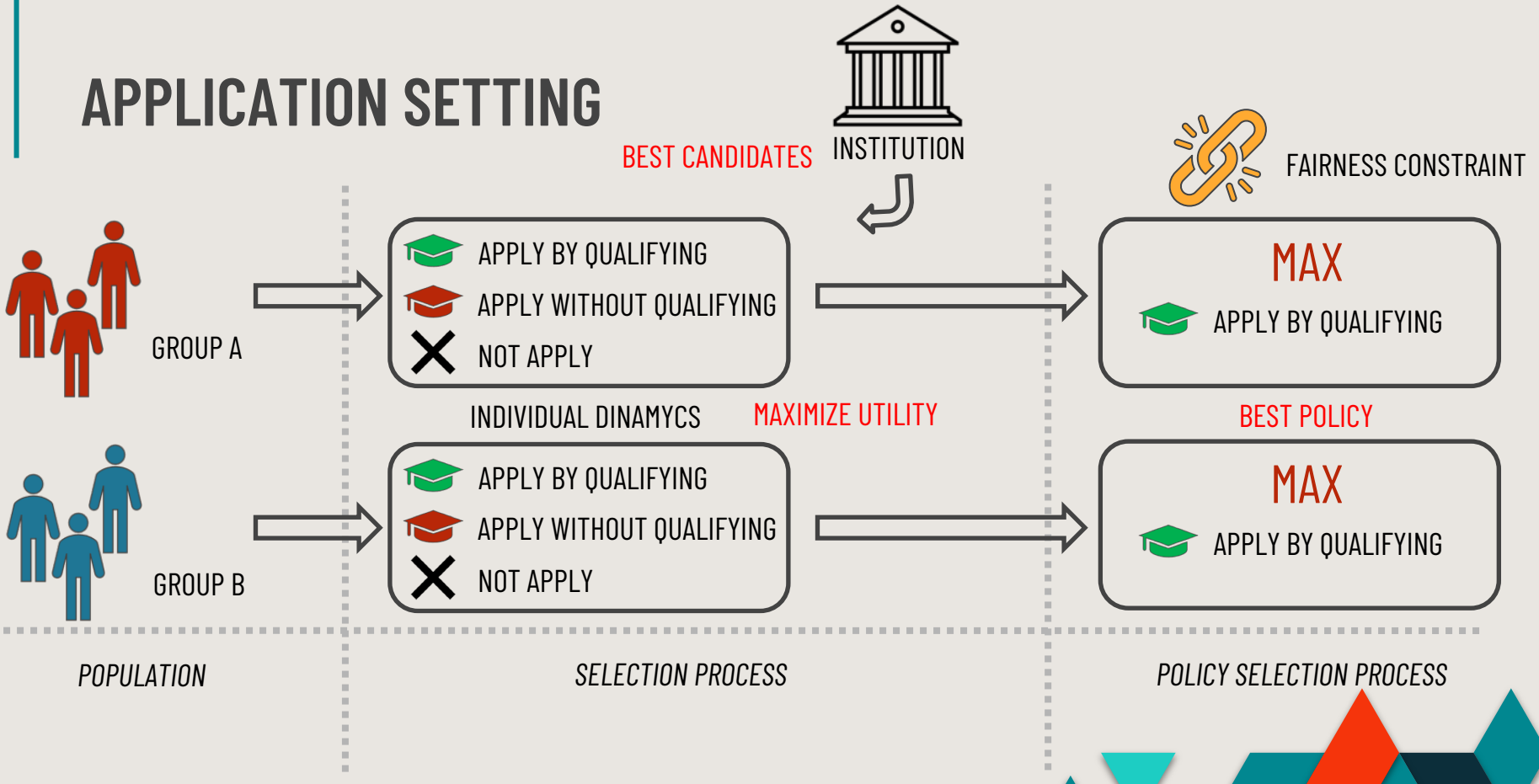
THEORETICAL MODEL

	SOCIETY	INDIVIDUAL DYNAMICS	
g_A, g_B	MINORITY AND MAJORITY GROUPS	ALTERNATIVES	X
γ_1	POSITIVE BEHAVIOR	SCENARIOS	Ω
$\theta \in \Theta$	SET OF INDIVIDUALS' ATTRIBUTES	IMPACTS	F
$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$	INSTITUTION'S UTILITY	UTILITY FUNCTION	f
$max \int_0^\infty \gamma_1(g_A) + \gamma_1(g_B)$	MAXIMIZING DOMINANCE	DECIDERS	D
$min \left\| \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \right\|$	MINIMIZING DOMINANCE AMONG GROUPS		

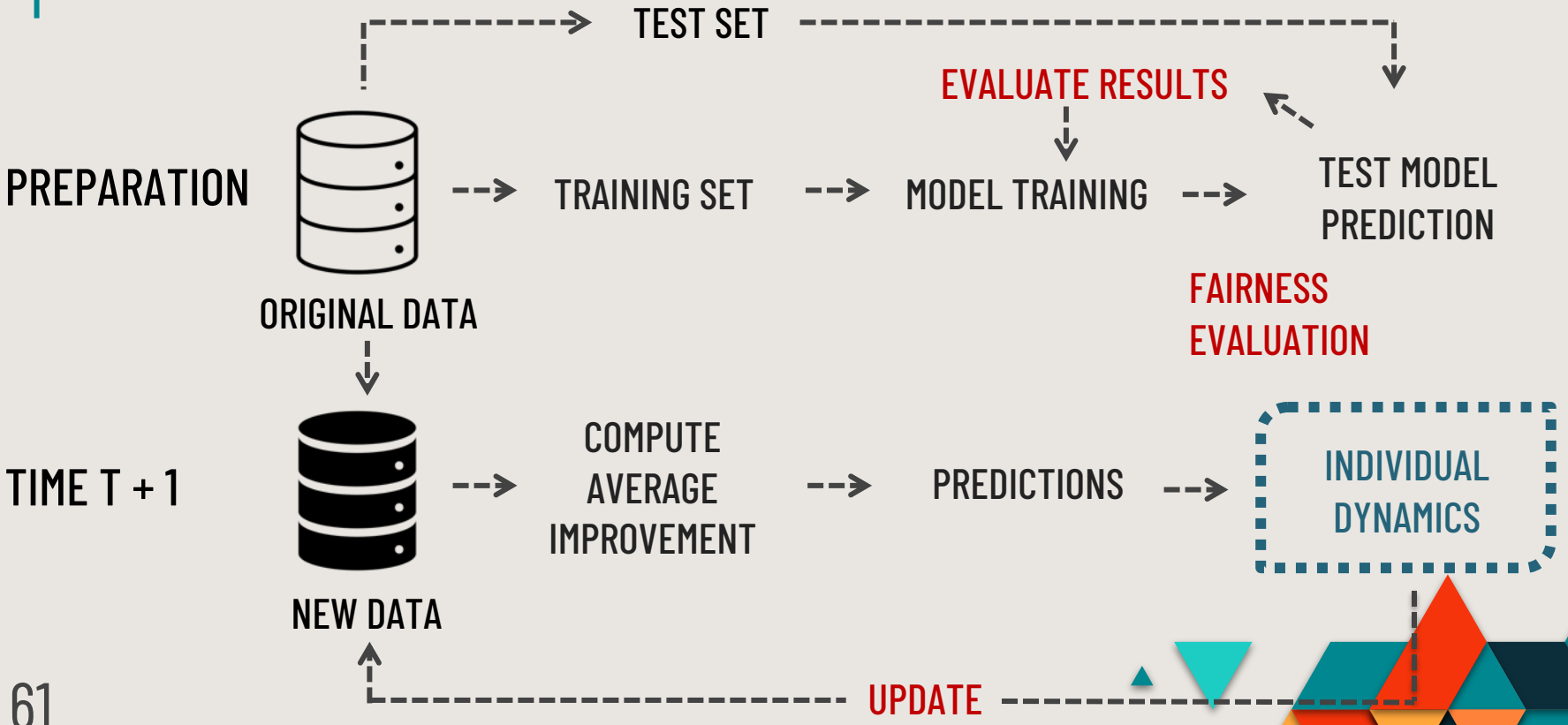
THEORETICAL MODEL

	SOCIETY	INDIVIDUAL DYNAMICS	
g_A, g_B	MINORITY AND MAJORITY GROUPS	ALTERNATIVES	X
γ_1	POSITIVE BEHAVIOR	SCENARIOS	Ω
$\theta \in \Theta$	SET OF INDIVIDUALS' ATTRIBUTES	IMPACTS	F
$U_I^* = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta)$	INSTITUTION'S UTILITY	UTILITY FUNCTION	f
$max \int_0^\infty \gamma_1(g_A) + \gamma_1(g_B)$	MAXIMIZING DOMINANCE	DECIDERS	D
$min \left\| \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \right\|$	MINIMIZING DOMINANCE AMONG GROUPS	PREFERENCES' FUNCTION	Π

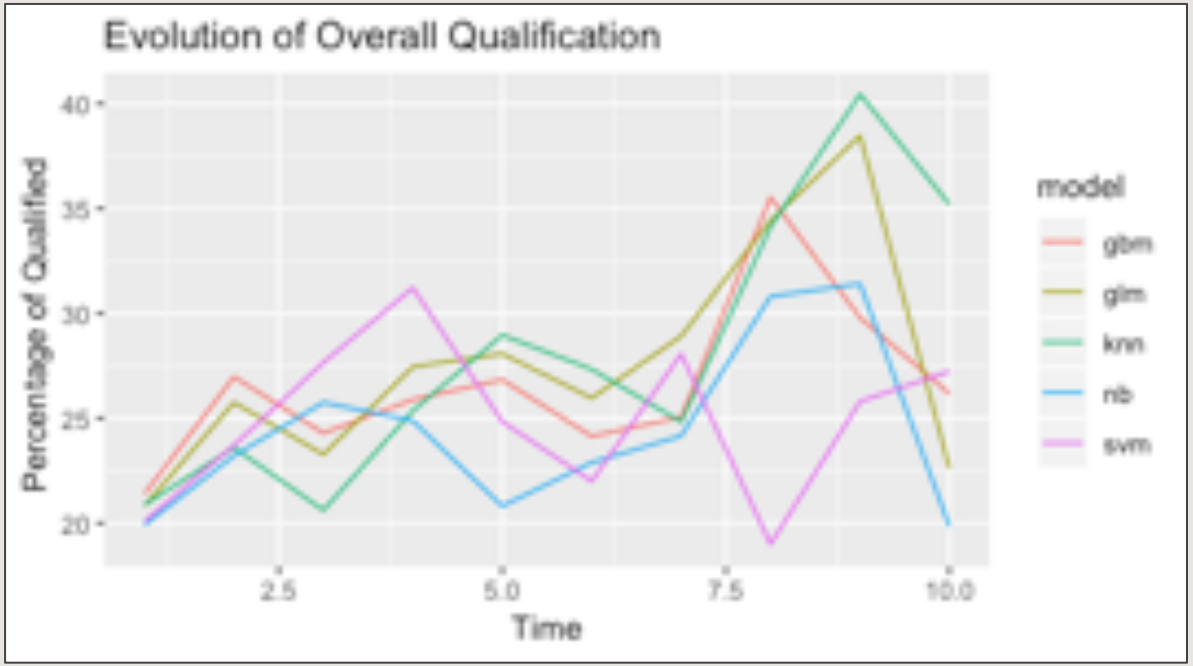
APPLICATION SETTING



APPLICATION SETTING

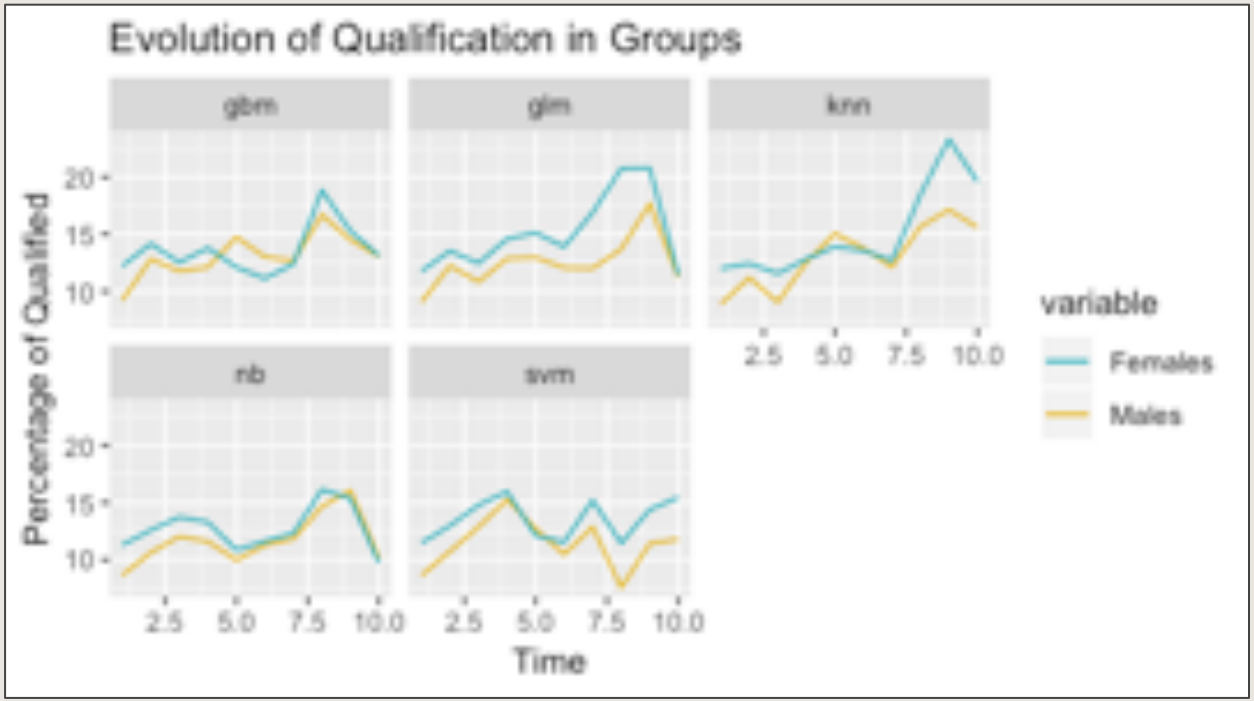


RESULTS



INCREASE IN OVERALL QUALIFICATION PROFILES

RESULTS



POSITIVE TREND IN BOTH GROUPS

RESULTS

	MAXIMIZING DOMINANCE $\int_0^{10} \gamma_1(g_A) + \gamma_1(g_B)$	MINIMIZING DOMINANCE AMONG GROUPS $\left\| \int_0^{\infty} \gamma_1(g_A) - \gamma_1(g_B) \right\|$
GRADIENT BOOSTING MACHINE	242.31	3.48
GENERALIZED LINEAR MODEL	253.33	15.83
K-NEAREST NEIGHBOUR	223.82	9.02
NAIVE BAYES CLASSIFIER	225.95	17.83
SUPPORT VECTOR MACHINE	254.06	25.19

RESULTS

	MAXIMIZING DOMINANCE $\int_0^{10} \gamma_1(g_A) + \gamma_1(g_B)$	MINIMIZING DOMINANCE AMONG GROUPS $\left\|\int_0^{\infty} \gamma_1(g_A) - \gamma_1(g_B)\right\|$
GRADIENT BOOSTING MACHINE	242.31	3.48
GENERALIZED LINEAR MODEL	253.33	15.83
K-NEAREST NEIGHBOUR	223.82	9.02
NAIVE BAYES CLASSIFIER	225.95	17.83
SUPPORT VECTOR MACHINE	254.06	25.19

RESULTS

	MAXIMIZING DOMINANCE $\int_0^{10} \gamma_1(g_A) + \gamma_1(g_B)$	MINIMIZING DOMINANCE AMONG GROUPS $\left\lvert \int_0^{\infty} \gamma_1(g_A) - \gamma_1(g_B) \right\rvert$
GRADIENT BOOSTING MACHINE	242.31	3.48
GENERALIZED LINEAR MODEL	253.33	15.83
K-NEAREST NEIGHBOUR	223.82	9.02
NAIVE BAYES CLASSIFIER	225.95	17.83
SUPPORT VECTOR MACHINE	254.06	25.19

RESULTS

- OUR SYSTEM IS EFFICIENT IN ANALYZING THE LONG-TERM EFFECTS OF POLICIES
- POLICIES HAVE DIFFERENT INFLUENCES ON GROUPS IN A NON-ONE-STEP MODEL
- FAIRNESS IS NOT CONSISTENT OVER TIME
- FAIRNESS CONSTRAINTS DO NOT NECESSARILY KEEP THEIR VALIDITY FOR AS LONG AS THEY ACT
- INDIVIDUAL DYNAMICS AFFECT SYSTEM OUTCOMES

THESIS CONTRIBUTION

DATA ANNOTATION SYSTEM

PREDICTING FUTURE DISCRIMINATORY RISK BASED ON BIASED DATA

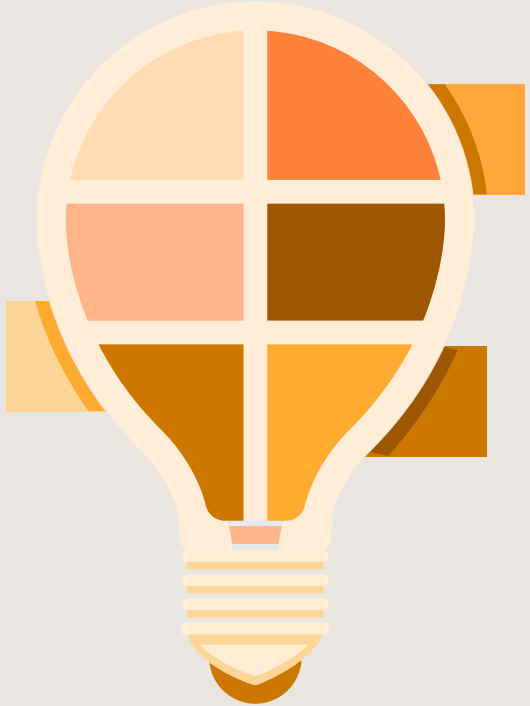
RANKING SYSTEM

PROPOSING A FAIR-DISTRIBUTIVE RANKING SYSTEM

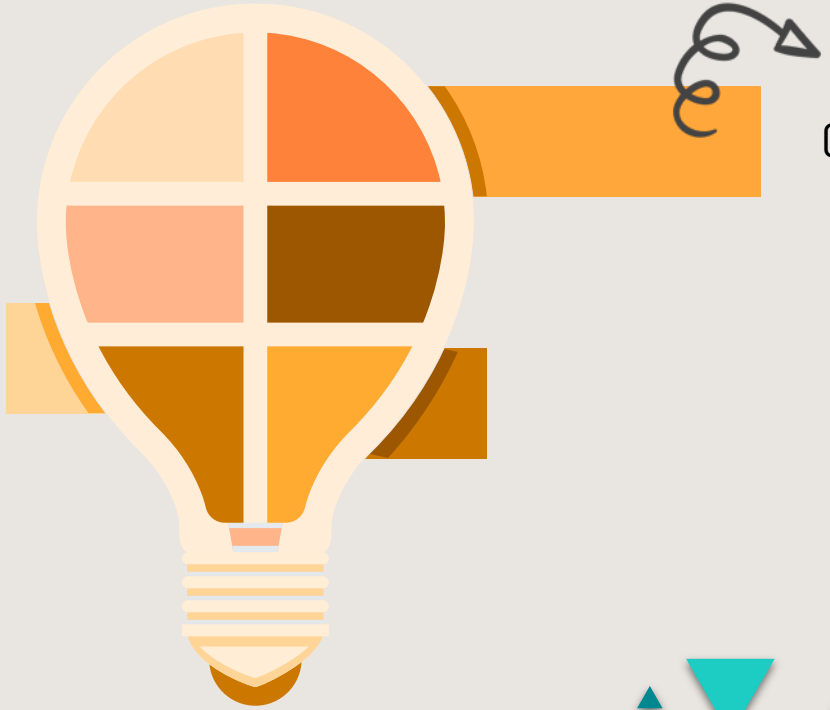
DECISION SUPPORT SYSTEM

MODELING INDIVIDUAL DYNAMICS FOR LONG-TERM FAIRNESS

CHALLENGES AND NEW TRAJECTORIES

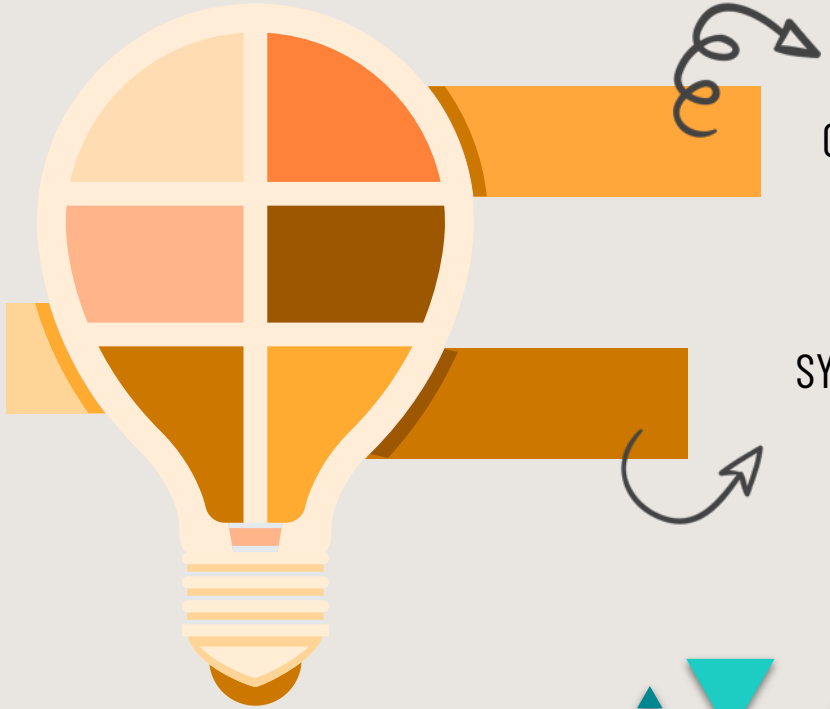


CHALLENGES AND NEW TRAJECTORIES



CROSS-DISCIPLINARY VALIDATION

CHALLENGES AND NEW TRAJECTORIES



CROSS-DISCIPLINARY VALIDATION

SYSTEMATIC GROUND ENCODING

CHALLENGES AND NEW TRAJECTORIES

MULTI-HIGH-INTERPRETABILITY

CROSS-DISCIPLINARY VALIDATION

SYSTEMATIC GROUND ENCODING

PUBLICATIONS

Published

Elena Beretta, Antonio Vetrò, Bruno Lepri, Juan Carlos De Martin. (2021) **Detecting discriminatory risk through data annotation based on Bayesian inferences**. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 794–804. DOI:https://doi.org/10.1145/3442188.3445940

Antonio Vetrò, Antonio Santangelo, Elena Beretta, Juan Carlos De Martin. (2019) **AI: from rational agents to socially responsible agents**, *Digital Policy, Regulation and Governance*, https://doi.org/10.1108/DPRG-08-2018-0049

Elena Beretta, Antonio Santangelo, Antonio Vetrò, Bruno Lepri, Juan Carlos De Martin. (2019) **The Invisible Power of Fairness. How Machine Learning Shapes Democracy**. In: Meurs MJ., Rudzicz F. (eds) *Advances in Artificial Intelligence. Canadian AI 2019*. Lecture Notes in Computer Science, vol 11489. Springer, Cham. https://doi.org/10.1007/978-3-030-18305-9_19

Elena Beretta, Antonio Vetrò, Bruno Lepri, Juan Carlos De Martin. (2018) **Ethical and Socially-Aware Data Labels**, Information Management and Big Data. SIMBig 2018. *Communications in Computer and Information Science*, vol 898, pp. 320-327, Springer, Cham.

Accepted for publication

Elena Beretta, Antonio Vetrò, Bruno Lepri, Juan Carlos De Martin. (2021) **Equality of opportunity in ranking: a Fair-Distributive Model**, Second International Workshop on Algorithmic Bias in Search and Recommendation (April 2021)

THANK YOU

Elena Beretta PhD candidate (XXXIII cycle) – Thesis Defense
Nexa Center for Internet & Society, Politecnico di Torino, Italy
Fondazione Bruno Kessler, Trento, Italy

Supervisors Prof. Juan Carlos De Martin, Politecnico di Torino
Bruno Lepri, Fondazione Bruno Kessler

Advisor Antonio Vetrò, Politecnico di Torino

APPENDIX

APPENDIX

DATA BIAS AWARENESS

DATA BIAS AND CONDITIONAL PROBABILITIES

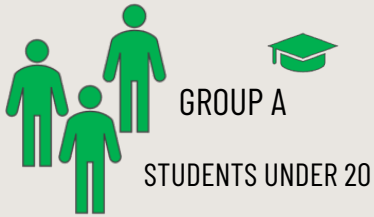
QUANTIFYING DEPENDENCE

ESTIMATING DIVERSENESS

ESTIMATING INCLUSIVENESS

ESTIMATING TRAINING LIKELIHOOD

DATA BIAS AND CONDITIONAL PROBABILITIES



GROUP A

STUDENTS UNDER 20

GROUP B

STUDENTS OVER 20

CASE 1

THE ACTUAL PROPERTIES DIFFER ACROSS GROUPS

OBSERVED SPACE = CONSTRUCT SPACE

DECISION SPACE ⟶ CORRECT MAPPING

CASE 2

THE ACTUAL PROPERTIES ARE DIFFERENT FROM THOS OBSERVED

OBSERVED SPACE ≠ CONSTRUCT SPACE

DECISION SPACE ⟶ ERRONEOUS MAPPING

POPULATION

CASES

SPACES

QUANTIFYING DEPENDENCE

DEPENDENCE

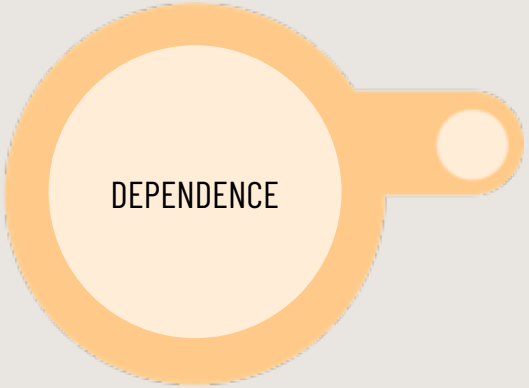
DEPENDENCE
ASSESSES THE DEGREE OF CONNECTION AMONG THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE

QUANTIFYING DEPENDENCE

DEPENDENCE

ASSESSES THE DEGREE OF CONNECTION AMONG THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE

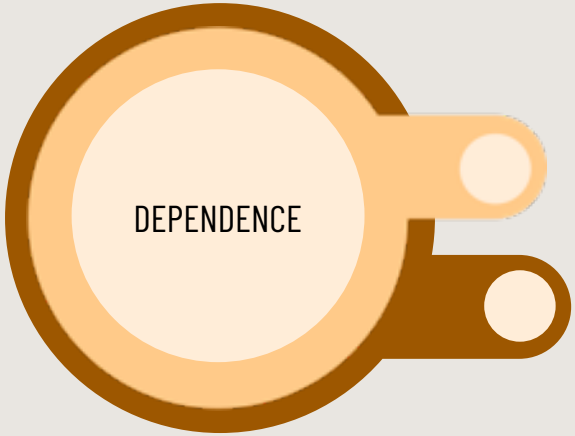
$$C(x_i;y_j) = f(x_i,y_j) - f'(x_i,y_j)$$



QUANTIFYING DEPENDENCE

DEPENDENCE

ASSESSES THE DEGREE OF CONNECTION AMONG THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE



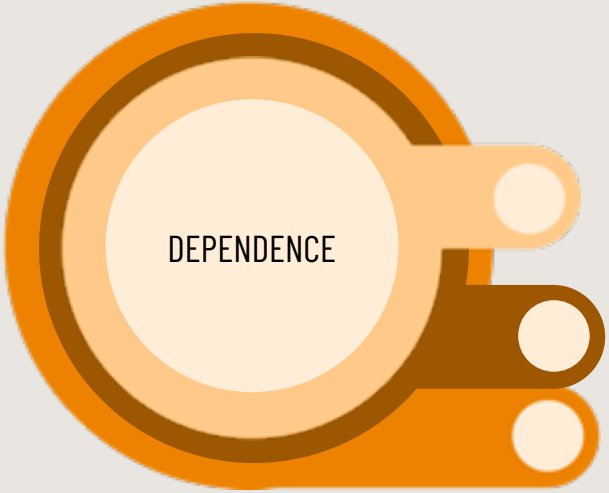
$$C(x_i;y_j) = f(x_i,y_j) - f'(x_i,y_j)$$

$$\chi^2 = \sum_{i,j} \frac{C^2(x_i;y_i)}{n_{i,j}} = n\left(\sum_{i,j} \frac{n_{i,j}^2}{n_{i,0}n_{0,j}}\right)$$

QUANTIFYING DEPENDENCE

DEPENDENCE

ASSESSES THE DEGREE OF CONNECTION AMONG THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE



$$C(x_i;y_j) = f(x_i,y_j) - f'(x_i,y_j)$$

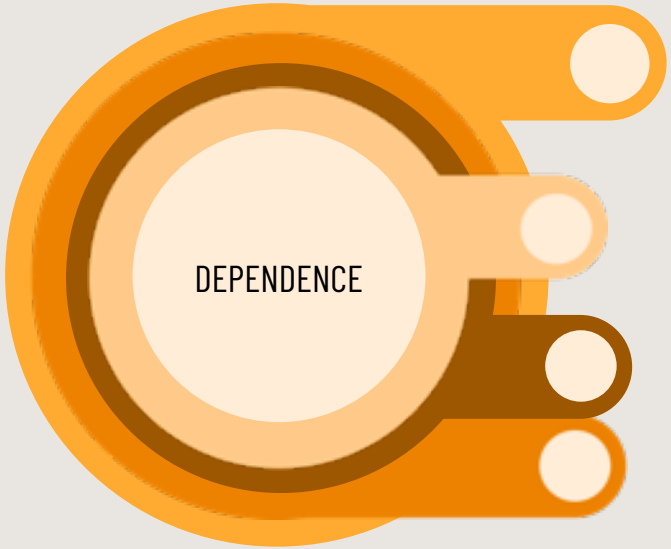
$$\chi^2 = \sum_{i,j} \frac{C^2(x_i;y_i)}{n_{i,j}} = n\left(\sum_{i,j} \frac{n_{i,j}^2}{n_{i,0}n_{0,j}}\right)$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

QUANTIFYING DEPENDENCE

DEPENDENCE

ASSESSES THE DEGREE OF CONNECTION AMONG THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE



$$C(x_i;y_j) = f(x_i,y_j) - f'(x_i,y_j)$$

$$\chi^2 = \sum_{i,j} \frac{C^2(x_i;y_i)}{n_{i,j}} = n\left(\sum_{i,j} \frac{n_{i,j}^2}{n_{i,0}n_{0,j}}\right)$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

MAGNITUDE	VALUE
SMALL	w = 0.1
MEDIUM	w = 0.3
LARGE	w = 0.5

$$w = \sqrt{\sum_{i=0} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$$

EFFECT SIZE INDEX w

$$C = \sqrt{\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{w^2}{w^2+1}} \qquad w = \sqrt{\frac{C^2}{1-C^2}}$$

ESTIMATING DIVERSENESS

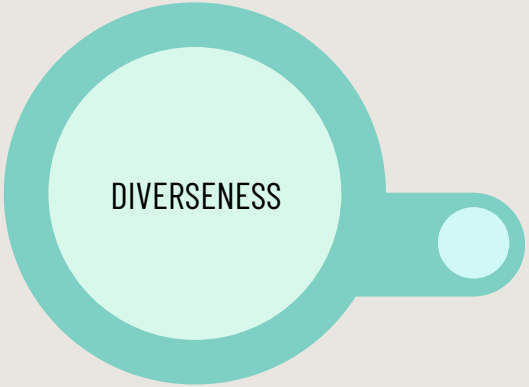
DIVERSENESS PROVIDES THE TRAINING DIVERSIFICATION PROBABILITY IN RESPECT TO EACH LEVEL OF THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE

DIVERSENESS

ESTIMATING DIVERSENESS

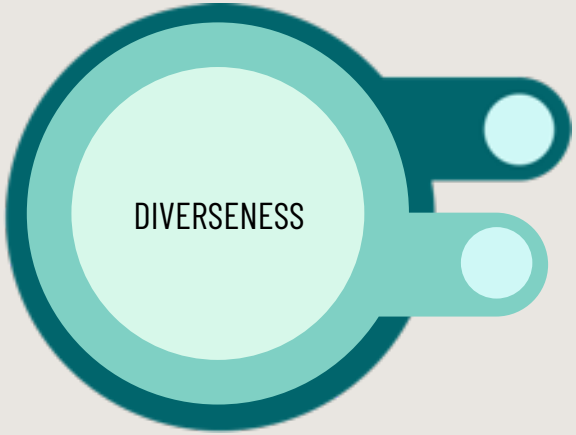
DIVERSENESS PROVIDES THE TRAINING DIVERSIFICATION PROBABILITY IN RESPECT TO EACH LEVEL OF THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE

$$P = \frac{number\ of\ favorable\ cases}{number\ of\ possibles\ cases}$$



ESTIMATING DIVERSENESS

DIVERSENESS PROVIDES THE TRAINING DIVERSIFICATION PROBABILITY IN RESPECT TO EACH LEVEL OF THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE

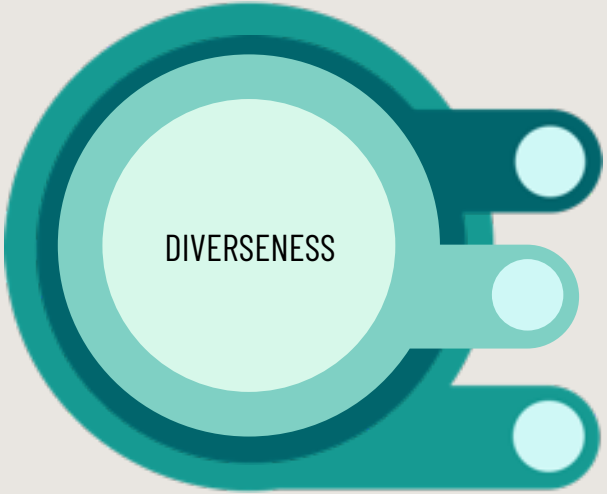


$$P = \frac{number\ of\ favorable\ cases}{number\ of\ possibles\ cases}$$

$$P = \frac{number\ of\ favorable\ properties}{number\ of\ possibles\ properties}$$

ESTIMATING DIVERSENESS

DIVERSENESS PROVIDES THE TRAINING DIVERSIFICATION PROBABILITY IN RESPECT TO EACH LEVEL OF THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE



$$P = \frac{number\ of\ favorable\ cases}{number\ of\ possibles\ cases}$$

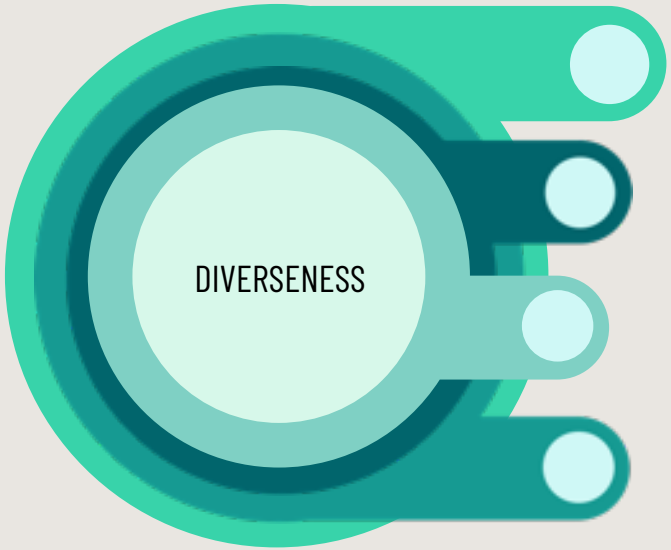
$$P = \frac{number\ of\ favorable\ properties}{number\ of\ possibles\ properties}$$

$$P = (Y = y)$$
$$P = (A = a)$$

PRIOR PROBABILITIES

ESTIMATING DIVERSENESS

DIVERSENESS PROVIDES THE TRAINING DIVERSIFICATION PROBABILITY IN RESPECT TO EACH LEVEL OF THE PROTECTED ATTRIBUTE AND THE TARGET VARIABLE



$$P = \frac{number\ of\ favorable\ cases}{number\ of\ possibles\ cases}$$

$$P = \frac{number\ of\ favorable\ properties}{number\ of\ possibles\ properties}$$

$P = (Y = y)$
$P = (A = a)$

PRIOR PROBABILITIES

FORMULA	PROBABILITY
$P(Y = 0)$ $P(Y = 1)$	$P = 0.48$ $P = 0.52$
$P(A = \text{white})$ $P(A = \text{black})$ $P(A = \text{Asian})$	$P = 0.6$ $P = 0.35$ $P = 0.15$

ESTIMATING INCLUSIVENESS

INCLUSIVENESS

PROVIDES THE PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET

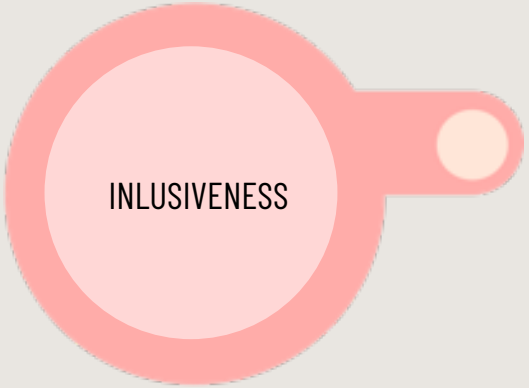
INLUSIVENESS

ESTIMATING INCLUSIVENESS

INCLUSIVENESS

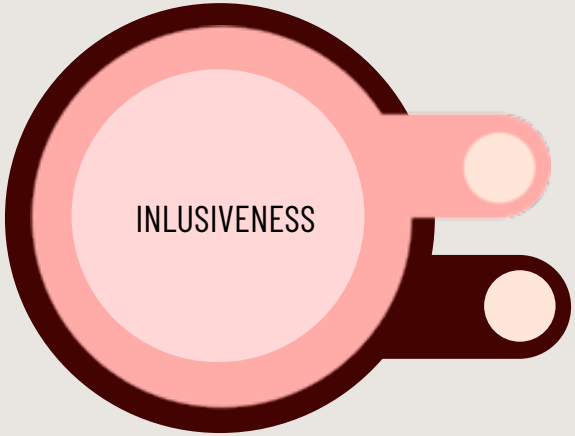
PROVIDES THE PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET

$P(A|B)$ POSTERIOR PROBABILITIES



ESTIMATING INCLUSIVENESS

INCLUSIVENESS
PROVIDES THE PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET

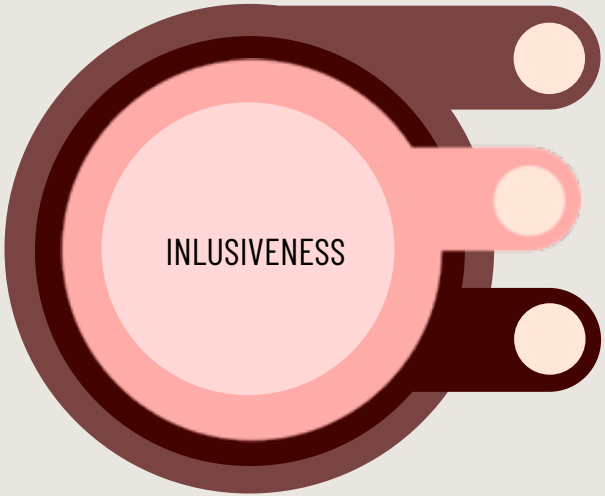


$P(A|B)$ POSTERIOR PROBABILITIES

$$P(A=a \cap Y=y)=P(A=a)P(Y=y|A=a)$$
$$P(Y=y \cap A=a)=P(Y=y)P(A=a|Y=y)$$

ESTIMATING INCLUSIVENESS

INCLUSIVENESS
PROVIDES THE PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET



$P(A|B)$ POSTERIOR PROBABILITIES

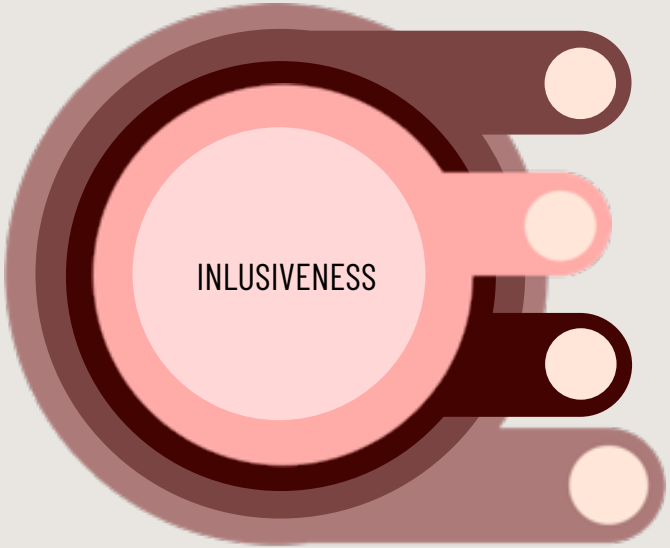
$$P(A=a \cap Y=y)=P(A=a)P(Y=y|A=a)$$
$$P(Y=y \cap A=a)=P(Y=y)P(A=a|Y=y)$$

$$P\,(A=a \cap Y=y) \;=\; P\,(Y=y \cap A=a)$$

COMPOUND PROBABILITY THEOREM

ESTIMATING INCLUSIVENESS

INCLUSIVENESS
PROVIDES THE PROBABILITY THAT TWO PROPERTIES ARE SIMULTANEOUSLY INCLUDED IN THE TRAINING SET



$P(A|B)$ POSTERIOR PROBABILITIES

$$P(A=a \cap Y=y)=P(A=a)P(Y=y|A=a)$$
$$P(Y=y \cap A=a)=P(Y=y)P(A=a|Y=y)$$

$$P\,(A=a \cap Y=y) \;=\; P\,(Y=y \cap A=a)$$

COMPOUND PROBABILITY THEOREM

FORMULA	PROBABILITY
$P(Y=0 \cap A=white)$	P=0.42
$P(Y=0 \cap A=black)$	P=0.07
$P(Y=0 \cap A=Asian)$	P=0.09
$P(Y=1 \cap A=white)$	P=0.18
$P(Y=1 \cap A=black)$	P=0.28
$P(Y=1 \cap A=Asian)$	P=0.06

ESTIMATING TRAINING LIKELIHOOD

TRAINING LIKELIHOOD PROVIDES THE OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS - AND VICE VERSA - BEFORE THE TRAINING SET IS SAMPLED

TRAINING LIKELIHOOD

ESTIMATING TRAINING LIKELIHOOD

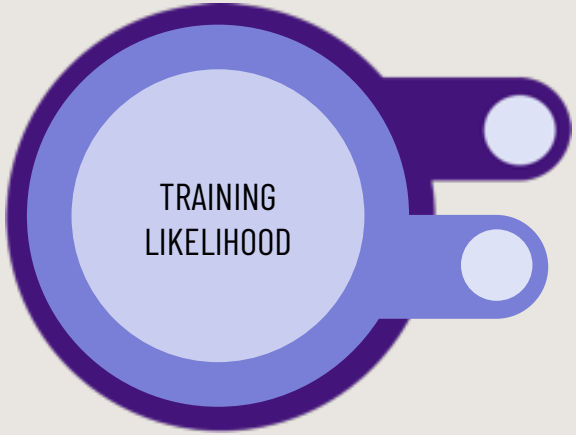
TRAINING LIKELIHOOD PROVIDES THE OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS - AND VICE VERSA - BEFORE THE TRAINING SET IS SAMPLED

$$P(A = a|Y = y) = \frac{P(A = a)P(Y = y|A = a)}{P(Y = y)}$$

$$P(Y = y|A = a) = \frac{P(Y = y)P(A = a|Y = y)}{P(A = a)}$$

ESTIMATING TRAINING LIKELIHOOD

TRAINING LIKELIHOOD PROVIDES THE OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS - AND VICE VERSA - BEFORE THE TRAINING SET IS SAMPLED



$$P(A = a|Y = y) = \frac{P(A = a)P(Y = y|A = a)}{P(Y = y)}$$

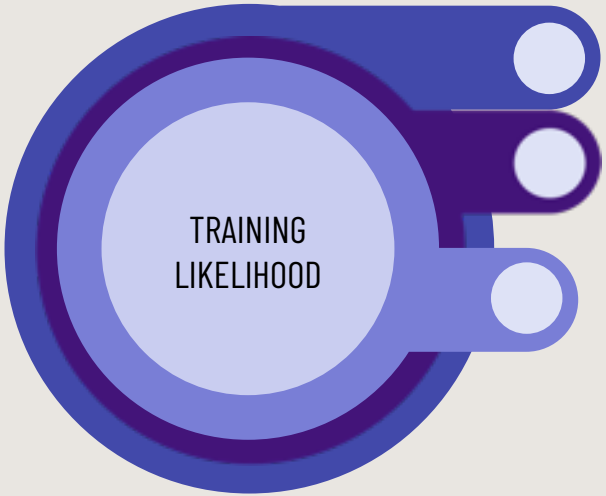
$$P(Y = y|A = a) = \frac{P(Y = y)P(A = a|Y = y)}{P(A = a)}$$

$$\Omega{:}\cup_{i=1}^{N} Y_i = \Omega,\ hence \sum_{i=1}^{N} P(Y_i) = P\left(\cup_{i=1}^{N} Y_i\right)$$

$$\Omega{:}\cup_{i=1}^{N} A_i = \Omega,\ hence \sum_{i=1}^{N} P(A_i) = P\left(\cup_{i=1}^{N} A_i\right)$$

ESTIMATING TRAINING LIKELIHOOD

TRAINING LIKELIHOOD PROVIDES THE OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS - AND VICE VERSA - BEFORE THE TRAINING SET IS SAMPLED



$$P(A = a|Y = y) = \frac{P(A = a)P(Y = y|A = a)}{P(Y = y)}$$

$$P(Y = y|A = a) = \frac{P(Y = y)P(A = a|Y = y)}{P(A = a)}$$

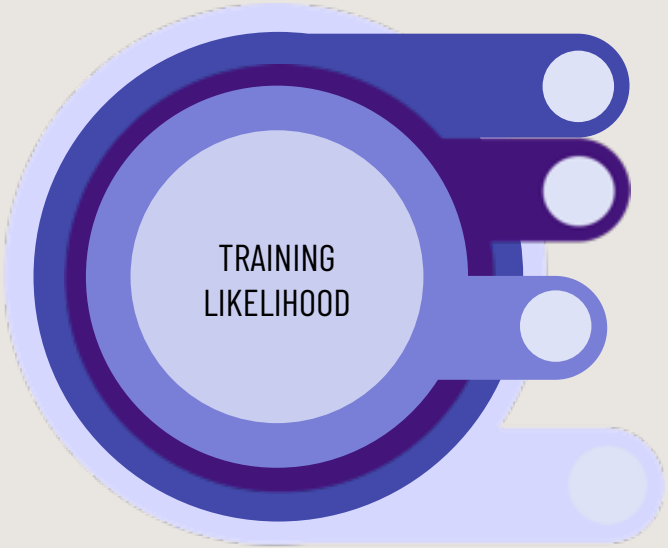
$$\Omega{:}\cup_{i=1}^{N} Y_i = \Omega,\ hence \sum_{i=1}^{N} P(Y_i) = P\left(\cup_{i=1}^{N} Y_i\right)$$

$$\Omega{:}\cup_{i=1}^{N} A_i = \Omega,\ hence \sum_{i=1}^{N} P(A_i) = P\left(\cup_{i=1}^{N} A_i\right)$$

$$P(Y = y|A) = \frac{P(Y = y)P(A|Y = y)}{P(A)} = \frac{P(Y = y)P(A|Y = y)}{\sum_{i=1}^{N} P(A|Y_i)P(Y_i)}$$

ESTIMATING TRAINING LIKELIHOOD

TRAINING LIKELIHOOD PROVIDES THE OCCURRENCE LIKELIHOOD OF THE PROTECTED ATTRIBUTE LEVELS GIVEN THE TARGET VARIABLE LEVELS - AND VICE VERSA - BEFORE THE TRAINING SET IS SAMPLED



FORMULA	PROBABILITY
$P(Y=0\|A=white)$ $P(Y=0\|A=black)$ $P(Y=0\|A=Asian)$	P=0.7 P=0.2 P=0.6
$P(Y=1\|A=white)$ $P\ (Y=1\|A=black)$ $P\ (Y=1\|A=Asian)$	P=0.3 P=0.8 P=0.4
$P(A=white\|Y=1)$ $P(A=white\|Y=0)$	P=0.34 P=0.87
$P(A=black\|Y=1)$ $P(A=black\|Y=0)$	P=0.53 P=0.15
$P(A=Asian\|Y=1)$ $P(A=Asian\|Y=0)$	P=0.11 P=0.18

FAIRNESS IN RANKING SYSTEMS

APPENDIX

MORAL GROUND

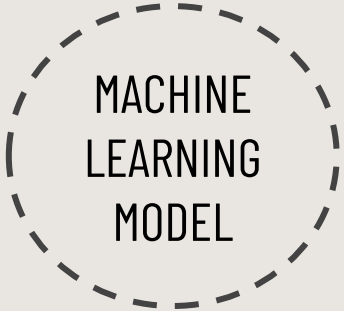
AFteRS: ALGORITHM 1

AFteRS: ALGORITHM 2

TABLE OF METRICS

MORAL GROUND

MORAL GROUND

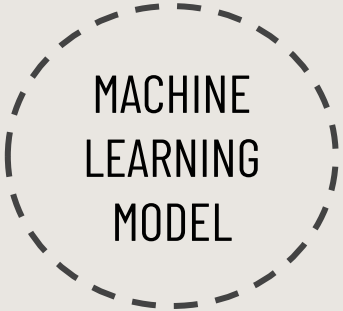


MORAL GROUND

MORAL GROUND

CASE 1

MORAL GROUND



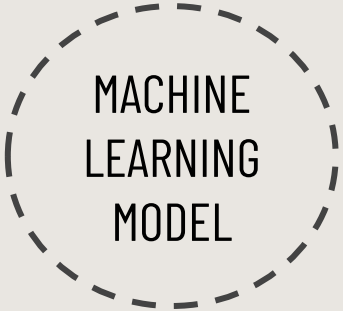
GROUND TRUTH

TRAINING
DATA

TARGET: PAST PERFORMANCE

CASE 1

MORAL GROUND



GROUND TRUTH

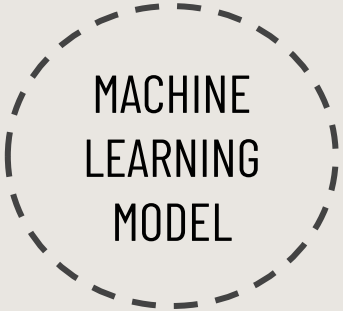
TRAINING
DATA

TARGET: PAST PERFORMANCE

REPLICATE HUMAN
DECISIONS

CASE 1

MORAL GROUND



GROUND TRUTH

TRAINING
DATA

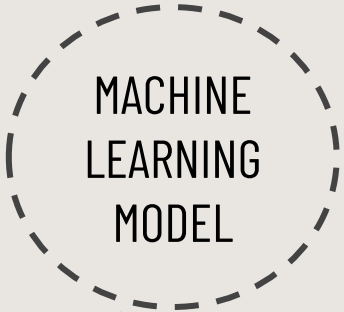
TARGET: PAST PERFORMANCE

REPLICATE HUMAN
DECISIONS

CASE 1

CASE 2

MORAL GROUND



GROUND TRUTH

TRAINING DATA

TARGET: PAST PERFORMANCE

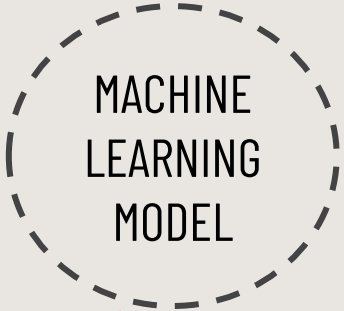
TARGET: ANNUAL REVIEWS

REPLICATE HUMAN DECISIONS

CASE 1

CASE 2

MORAL GROUND



GROUND TRUTH

TRAINING DATA

TARGET: PAST PERFORMANCE

TARGET: ANNUAL REVIEWS

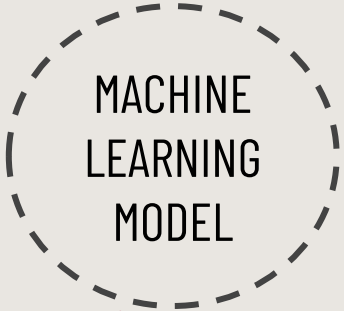
REPLICATE HUMAN DECISIONS

REPLICATE HUMAN MANAGER DECISIONS

CASE 1

CASE 2

MORAL GROUND



GROUND TRUTH

TRAINING DATA

TARGET: PAST PERFORMANCE

TARGET: ANNUAL REVIEWS

REPLICATE HUMAN DECISIONS

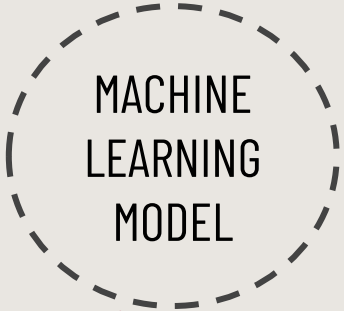
REPLICATE HUMAN MANAGER DECISIONS

CASE 1

CASE 2

CASE 3

MORAL GROUND



GROUND TRUTH

TRAINING DATA

TARGET: PAST PERFORMANCE

REPLICATE HUMAN DECISIONS

CASE 1

TARGET: ANNUAL REVIEWS

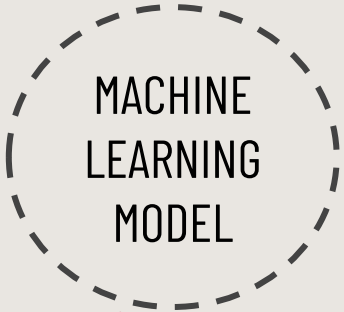
REPLICATE HUMAN MANAGER DECISIONS

CASE 2

TARGET: SALES FOR THE YEAR

CASE 3

MORAL GROUND



GROUND TRUTH

TRAINING
DATA

TARGET: PAST PERFORMANCE

REPLICATE HUMAN
DECISIONS

CASE 1

TARGET: ANNUAL REVIEWS

REPLICATE HUMAN
MANAGER DECISIONS

CASE 2

TARGET: SALES FOR THE YEAR

REPLICATE
CIRCUMSTANCES

CASE 3

AFteRS: Algorithm 1

Algorithm 1 Automated Fair Distributive Ranking – Step 1-2 (Figure 3)
The algorithm partitions the population in n types (Section 3.2.1), derives effort (Section 3.2.2), and computes the Standardized Distribution (Equation 5)

Step 1
input: dataset D
output: non-overlapping subsets of D $\implies$ population partitioned in T_k types

for all $X_i \in D$ **do**
  Test the null hypothesis of independence between Y and all X_i
  if $H^0_{partial} : P(Y|X) = P(Y)$ couldn't be rejected **then**
    Stop
  else
    1. select X_i with the strongest association to Y (smallest *adj p-value)*
    2. find the splitting point $C*$ for X_i such that
      $S_n^{x_i} \subset \chi_i$ are all the possible disjoint sets of the sample space χ_i
  end if
end for
return T_k vectors $\subset D$

Step 2
input: T_k vectors $\subset$ D
output: Standardized Outcome $\tilde{y}_i^t(\lambda)$

partition each T_k in 10 sets Ψ_n, such that $\Psi_{k,n} \subset T_k$
training set $1^{st} - 9^{th}$ T_k sets
test set 10^{th} T_k set
for all $\Psi_{k,n} \subset T_k$ **do**
  1. perform the Bernstein polynomials log-likelihood on the training set to estimate the best type-distribution approximation $LL_B(p_m = \sum_{i=1}^{n} \log f_B(x_j, p_m))$
  2. predict the CDF of T_k on the test set
end for
estimate the *Standardize Distribution* $= y_i^t(\lambda)\frac{\mu}{\mu^\lambda}$
return $\tilde{y}_i^t(\lambda)$

AFteRS: Algorithm 2

Algorithm 2 Automated Fair Distributive Ranking – Step 3 (Figure 3)
The algorithm computes the Γ ranking based on policies Equity, Equality and Need

Step 3
input: Standardized Outcome $\tilde{y}_i^t(\lambda)$
output: ranking Γ constrained by a policy $\theta \in \Theta$

if $\theta = equity$ **then**
for all $\tilde{Y}_{t,\lambda} \in D$ **do**
    compute the counterfactual outcome from stnd. outcome and decomposed Gini
    $\Gamma \leftarrow$ ranking ordered by counterfactual outcome
end for

if $\theta = equality$ **then**
for all $T_k \in D$ **do**
    $sorted_{T_k} \leftarrow$ type-ranking ordered by decreasing stnd. outcome
end for
for all $(j) \in sorted_{T_k}$ **do**
    $row_n \leftarrow j$ element of $sorted_{T_k}$
    array[j] $\leftarrow row_n$ ordered by decreasing std. outcome
end for
$\Gamma \leftarrow$ *merge* all j array

if $\theta = need$ **then**
$G_k \leftarrow n$ subsets $\in D$ grouped by protected attribute A
for all $(z) \in sorted_{G_k}$ **do**
    $row_n \leftarrow z$ element of $sorted_{G_k}$
    array[z] $\leftarrow row_n$ ordered by decreasing std. outcome
end for
$\Gamma \leftarrow$ *merge* all z array

return Ranking Γ

TABLE OF METRICS

Metric	Formula	Input
Expected ranking	$r = argmaxU(ranking_n \mid q)$	Score distribution
Exposure	$\frac{1}{\log(1+j)}$	Original Distribution
Relevance	$\beta(Rel(item_n \mid user_n, q))$	ScDistr., adj ScoreDistr
Expected ranking-policy	$\Gamma = argmax_{\theta \in \Theta} u^t(q \mid e_i(\lambda), \theta)$	Adj ScoreDistr
Exposure-policy	$\max_{\theta \in \Theta} \int_0^1 min_t exp^t(\lambda, \theta)\, d\lambda$	Adj ScoreDistr
Gini Index	$1 - \frac{1}{\mu} \int_0^\infty (1 - F(y))^2 dy$	All distributions
Decomposed Gini	$Gini_\lambda^t$	Stand. score distribution
Richness	n^t	Types diversity
Margalef	$\frac{T-1}{lnN}$	Types diversity
Shannon-Wiener Index	$H = \sum_{i=1}^R p_i \ln p_i$	Types diversity
Simpson	$1 - \sum \frac{n^t(n^t-1)}{N(N-1)}$	Types diversity
Theil Index	$\frac{1}{N} \sum_{i=1}^N ln(\frac{\mu}{y_i})$	All Distributions
Opportunity-Types Profile	$min/max(y^t - \mu(y))$	Score distribution
Opportunity-Types Rate	$y^t - \mu(y)$	ScDistr
Opportunity-L/G Profile	$min/max(y_\lambda^t - \mu(y_\lambda))$	Stnd. distribution
Opportunity-L/G Rate	$y_\lambda^t - \mu(y_\lambda)$	StndDistr
Unexplained Inequality Rate	$\frac{1}{N} \sum y_i - \tilde{y}_i$	ScDistr, stndDistr
Reward Profile	$min/max(j(y_\lambda^t) - j(adj(\tilde{y}_\lambda^t)))$	ScDistr, adj score distr.
Reward Rate	$j(y_\lambda^t) - j(adj(\tilde{y}_\lambda^t))$	ScDistr, adj score distr.

Table 3: Summary of metrics employed. *Notation*: F(y)= cumulative distribution function of the score, μ = mean score; R = number of types, p_i = frequency of types; y_λ^t = score distribution aggregated by type and quantile; $\tilde{y}_i$= standardized score; $adj(\tilde{y}_\lambda^t)$= adjusted mean-type score at each effort degree (after policy); j = ranking position

LONG-TERM FAIRNESS

APPENDIX

APPLICATION SETTING

INDIVIDUAL DYNAMICS

APPLICATION SETTING

SCENARIO	UNIVERSITY SELECTION PROCESS	MAXIMIZE LONG-TERM SELECTION	BEST CANDIDATES SELECTION	FAIRNESS AS POSITIVE BEHAVIOR	QUALIFICATION AT TIME t + 1
DATA	SYNTHETHIC DATA	RETAKING SAT STATISTICS	GPA, SAT AND GRE SCORE, AGE, SEX	ONLY IMPROVE SAT SCORE	ESTABLISH MAXIMUM SCORE
POLICY	GRADIENT BOOSTING MACHINE	GENERALIZED LINEAR MODEL	K-NEAREST NEIGHBOUR	NAIVE BAYES CLASSIFIER	SUPPORT VECTOR MACHINE
TIME	NON EVOLVING PREFERENCES	DATA AT TIME t + 1	KNOWLEDGE OF THE STATE OF THE NATURE	NOT A DETERMINISTIC MODEL	POLICY SELECTION 10 YEARS

INDIVIDUAL DYNAMICS

ALTERNATIVES → SCENARIOS → IMPACTS AND UTILITY FUNCTION → PREFERENCES

x_1: APPLYING WITH QUALIFICATION

x_2: APPLYING WITHOUT QUALIFICATION

x_3: NOT APPLYING

x_1: Q such that X = [Q, NQ, Q]
x_2: NQ
x_3: N

INDIVIDUAL DYNAMICS

ALTERNATIVES → SCENARIOS → IMPACTS AND UTILITY FUNCTION → PREFERENCES

$\pi(\omega\|x)$	Q	NQ	N
OPTIMISTIC	0.75 | 0.86	0.75 | 0.76	0.75 | 0.76
PESSIMISTIC	0.85 | 0.86	0.85 | 0.76	0.85 | 0.76
AGNOSTIC	0.8 | 0.86	0.8 | 0.76	0.8 | 0.76

INDIVIDUAL DYNAMICS

ALTERNATIVES → SCENARIOS → IMPACTS AND UTILITY FUNCTION → PREFERENCES

$f(\omega\|x)$	Q	NQ	N
OPTIMISTIC	0.9	1	-1
PESSIMISTIC	0.9	0	0
AGNOSTIC	0.9	0	0

INDIVIDUAL DYNAMICS

ALTERNATIVES → SCENARIOS → IMPACTS AND UTILITY FUNCTION → PREFERENCES

$f(\omega \mid x)$	Q	NQ	N
OPTIMISTIC	0.9	1	-1
PESSIMISTIC	0.9	0	0
AGNOSTIC	0.9	0	0

$$max_{x \in X} Laplace(x) = max_{x \in X} \frac{\sum_{\omega \in \Omega} f(x, \omega)}{|\Omega|}$$

INDIVIDUAL DYNAMICS

ALTERNATIVES → SCENARIOS → IMPACTS AND UTILITY FUNCTION → PREFERENCES

$f(\omega \mid x)$	Q	NQ	N
OPTIMISTIC	0.9	1	-1
PESSIMISTIC	0.9	0	0
AGNOSTIC	0.9	0	0
Laplace(x)	0.9	0.33	- 0.33

$$max_{x \in X} Laplace(x) = max_{x \in X} \frac{\sum_{\omega \in \Omega} f(x, \omega)}{|\Omega|}$$