How to use international standard to be compliant to regulation in the era of AI

prof. Alessandro Simonetta

Maria Cristina Paoletti

Rome, the 28th of November 2025

INTRODUCTION

FAIRNESS METRICS

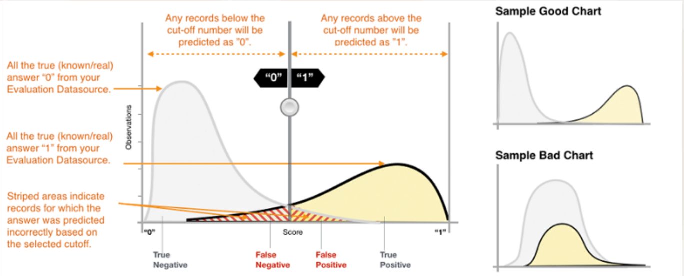
CASE STUDY

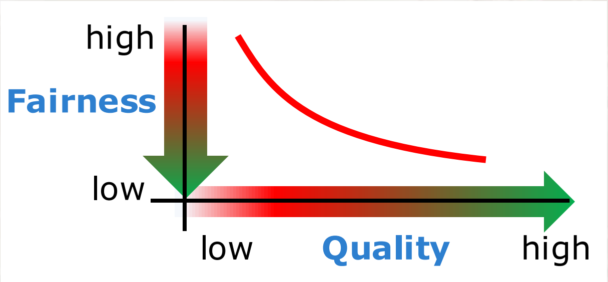
CONCLUSION & FUTURE

Agenda

- The blueprint of my careers
- The EU's Digital Decade 2030 Strategy and Digital Compass
- Introduction to AI Act
- International and harmonized standards
- Artificial intelligence systems
- The blueprint of my careers
- Fairness measurement in the ML classification
- Case study: the COMPAS dataset
- The notion of data completeness
- The relationship between poor data quality and unfair outcome
- Mitigating the harmful effects of bias
- Conclusions
- References

European Commission | 2030 DIGITAL DECADE





The blueprint of my career

INTRODUCTION
FAIRNESS METRICS
CASE STUDY
CONCLUSION & FUTURE

Public Sector Employee — 29 years

20/02/00 – today — ~26 years

INAIL

- Actuary Professional
- Sector Coordinator – Rate And Services Actuarial Statistical Consulting

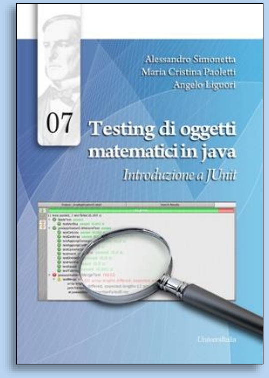
20/05/96 – 31/01/00

Ministry of justice, department of penitentiary administration

- Statistical Assistant - Level VI — 3 years

Contract Professor

- Programming Fundamentals — A.A. 2012/2013



09/01/13 – today

UNI UNINFO

UN MONDO FATTO BENE

UNI/CT 504 Software engineering

UNI/CT 533 Artificial Intelligence

ISO/IEC JTC1 SC7 Software and systems engineering

ISO/IEC JTC1 SC 42

CEN/CLC JTC 21

ISO IEC cen CENELEC

12 years

issa

GOOD PRACTICE AWARD EUROPE COMPETITON

INAIL

2019 GDPR and data transparency: a quality approace

2016 Measuring worker well-being is necessary for good governance

Our research background

INTRODUCTION

FAIRNESS METRICS

CASE STUDY

CONCLUSION & FUTURE

ISO/IEC standards and design of an Artificial Intelligence system
IWESQ 2024, 3rd December 2024, https://ceur-ws.org/Vol-3916/ pp.39-43
Simonetta A., Paoletti M.C.

The SQuaRE series as a guarantee of ethics in the results of AI systems
IWESQ 2023, 4th December 2023 https://ceur-ws.org/Vol-3612/ pp.17-21
Simonetta A., Paoletti M.C., Nakaijma T.

Ethics in Artificial Intelligence Systems
INAIL seminar, December 4-6, 2023, Sapienza University of Rome,
Paoletti M.C., Simonetta A., Natale D.

Application of AI for social and labor reintegration in the operation of complex machinery
INAIL seminar, December 4-6, 2023 Sapienza University of Rome
Muratore M., Paoletti M.C., Simonetta A., Colafemmina.G.

Fairness Metrics and Maximum Completeness for the prediction of discrimination (*)
IWESQ 2022, Tokyo, 6th December 2022, https://ceur-ws.org/Vol-3356/
Simonetta A., Nakaijma T., Paoletti M.C., Venticinque A.

Bias
AI
SQuaRE

The use of Maximum Completeness to estimate bias in AI based recommendation systems
SYSTEM 2022, Brunico July 23-31, https://ceur-ws.org/Vol-3360/
Simonetta A., Paoletti M.C., Venticinque A.

Using the SQuaRE series as a guarantee for GDPR compliance (*)
IWESQ 2021, 8th December 2021, http://ceur-ws.org/Vol-3114/paper-05.pdf
Simonetta A., Paoletti M.C., Venticinque A.

Integrating SQuaRE data quality model with ISO 31000 risk management to measure and mitigate software bias (*)
IWESQ 2021, 8th December, 2021, http://ceur-ws.org/Vol-3114/paper-04.pdf
Simonetta A., Vetrò A., Paoletti M.C., Torchiano M.

Metrics for identifying bias in datasets (*)
ICYRIME 2021, https://ceur-ws.org/Vol-3118/p02.pdf
Simonetta A., Trenta A., Paoletti M.C., Vetrò A.

(*) Research articles referred by CEN/CLC/TR **18115:2024** "Data governance and quality for AI within the European context"

Security / Software

Code protection techniques when distributed in source format: an Adobe Connect Pod written in JavaScript
SYSTEM 2021, July 27-29, 2021, https://ceur-ws.org/Vol-3092/p06.pdf
Simonetta A., Rinaldi F.

A forensic methodology for the identification of illicit data leakage
SYSTEM 2021, July 27-29, 2021, https://ceur-ws.org/Vol-3092/p01.pdf
Simonetta A., Fazio L., Paoletti M.C.

A simple method for extracting real dependencies between data and software applications
INAIL seminar, 23-25 October 2018, Simonetta A.

New Computing Architectures

Multi-Valued Logic digital circuits for realizing a complete computer architecture, ICYRIME 2022, August 26-29, 2022, https://ceur-ws.org/Vol-3398, Simonetta A., Paoletti M.C., Venticinque A.

A New Approach for Designing of Computer Architectures Using Multi-Value Logic, IJASEIT, Vol. 11 (2021) No. 4, pages: 1440-1446, DOI:10.18517/ijaseit.11.4.15778, Simonetta A., Paoletti M.C., Muratore M.

Designing Digital Circuits in Multi-Valued Logic, IJASEIT, Vol. 8 (2018) No. 4, DOI:10.18517/ijaseit.8.4.5966, Simonetta A. & Paoletti M.C.

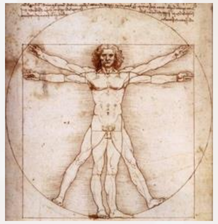
European Commission

The main objective

2030 DIGITAL DECADE

INTRODUCTION
FAIRNESS METRICS
CASE STUDY
CONCLUSION & FUTURE

The EU is pursuing a human-centric, sustainable vision for **digital society** throughout the digital decade to empower citizens and businesses.




https://digital-strategy.ec.europa.eu/en/policies/europes-digital-decade

Digital society and digital technologies bring with them new ways to learn, entertain, work, explore, and fulfil ambitions.

They also bring new **freedoms** and **rights** and give EU citizens the opportunity to reach out beyond physical communities, geographical locations, and social positions.


2030 DIGITAL DECADE

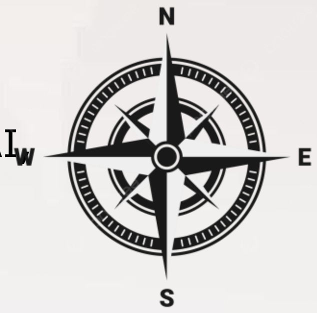
Strategic action plan: 2030 Digital Compass

https://digital-strategy.ec.europa.eu/en/policies/digital-decade-policy-programme

DIGITALLY SKILLED CITIZENS & HIGHLY SKILLED DIGITAL PROFESSIONALS

- 20 million ICT specialists and gender balance
- 80% of the population with digital skills

cardinal points



DIGITAL TRANSFORMATION OF BUSINESSES

- 75% of EU enterprises using Cloud, AI or Big Data
- To double the number of unicorn startups in the EU ($1 billion without being listed)
- 90% of SMEs using new technologies

DIGITALISATION OF PUBLIC SERVICES

- 100% of essential public services online
- 100% of citizens with eID digital identity and access to digital health records

SECURE, PERFORMANT AND SUSTAINABLE DIGITAL INFRASTRUCTURES

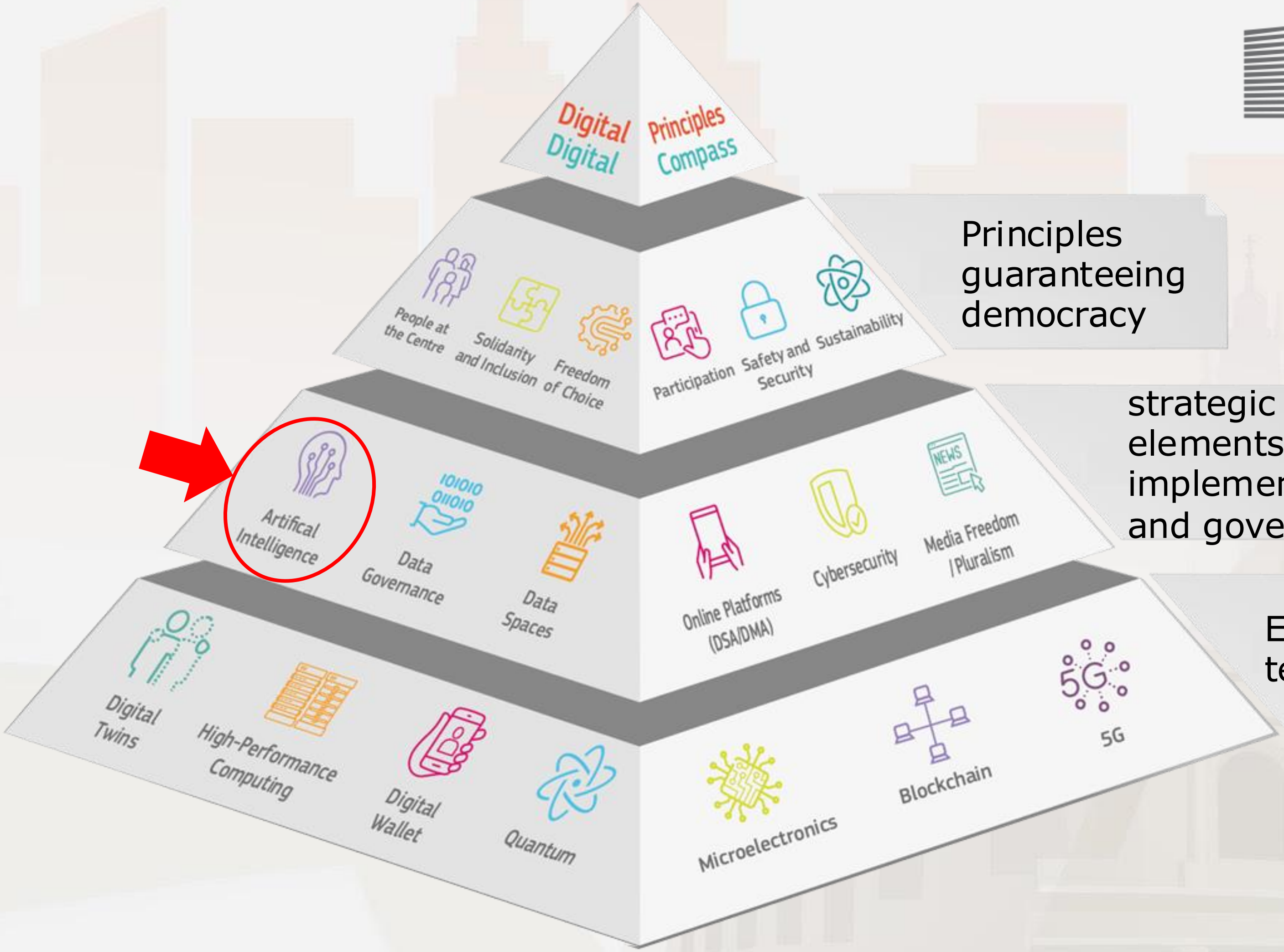
- Gigabit connectivity for everyone
- High-speed mobile coverage (at least 5G) everywhere
- EU reaches 20% of global semiconductor production
- 10,000 zero-impact edge cloud nodes
- First European Quantum Computer in 2025

2030 DIGITAL DECADE

INTRODUCTION
FAIRNESS METRICS
CASE STUDY
CONCLUSION & FUTURE

Key elements of the European digital strategy

European Commission



Principles guaranteeing democracy

strategic elements for implementation and governance

Enabling digital technologies

European Commission

AI as a driver of digital transformation

The goal is to create a European ecosystem of public and private actors that develop and deploy AI systems in line with the **Union's values** and **unlock the potential** of the digital transformation across all Union regions (ref. Digital Compass 2030).

The problem of **digital illiteracy** is one of the main obstacles to the widespread adoption of AI systems.

Collaboration among all stakeholders is necessary to overcome the lack of knowledge, as well as the Commission's need to adapt the regulation

AI ACT

Official Journal of the European Union

EN L series

2024/1689

12.7.2024

REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 13 June 2024

laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

Whereas (8), Art. 4

Expectations for AI Systems

- The selection is **objective**, without influences resulting from social prejudices (Bias).
- The result is fair and does not unfairly discriminate against groups of people based on ethnicity, gender, age, or similar characteristics.
- The response is **immediate** or near real-time.
- The evaluation considers a **comprehensive universe of information**, therefore the best possible outcome.
- The decision-making process can be automated, either replacing human decisions or providing decision support.

These expectations could be met if, and only if, the data used to build the learning models is of **high quality**.

AI Act's goals

(1) (2) (8), Art.1

To improve the functioning of the internal market by establishing a uniform legal framework for AI systems

To promote the development and adoption of human-centric and trustworthy Artificial Intelligence (AI)

To protect citizens and society against the possible harmful effects of AI systems

To promote innovation by avoiding limitations on the development, sale, and use of AI systems

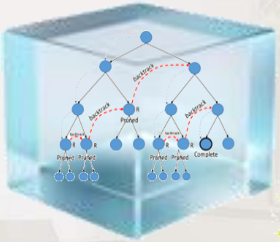
Establishing the EU as a leader in the adoption of trustworthy AI

Ethical principles

In 2019, the AI **High-level expert group on artificial intelligence** (HLEG) group elaborated seven non-binding ethical principles for AI to ensure the trustworthiness and ethics of an AI system.

Principles that should be **by-default** and **by-design**:

1. Human monitoring and intervention 
2. Technical robustness and safety (cybersecurity and resilience) 
3. Privacy and data governance (data quality) 


4. Transparency (traceability and explainability)
5. Diversity, non-discrimination, and fairness (bias or prejudice) 
6. Societal and environmental well-being 
7. Accountability 

Risk-Based Approach

(27), Art.3

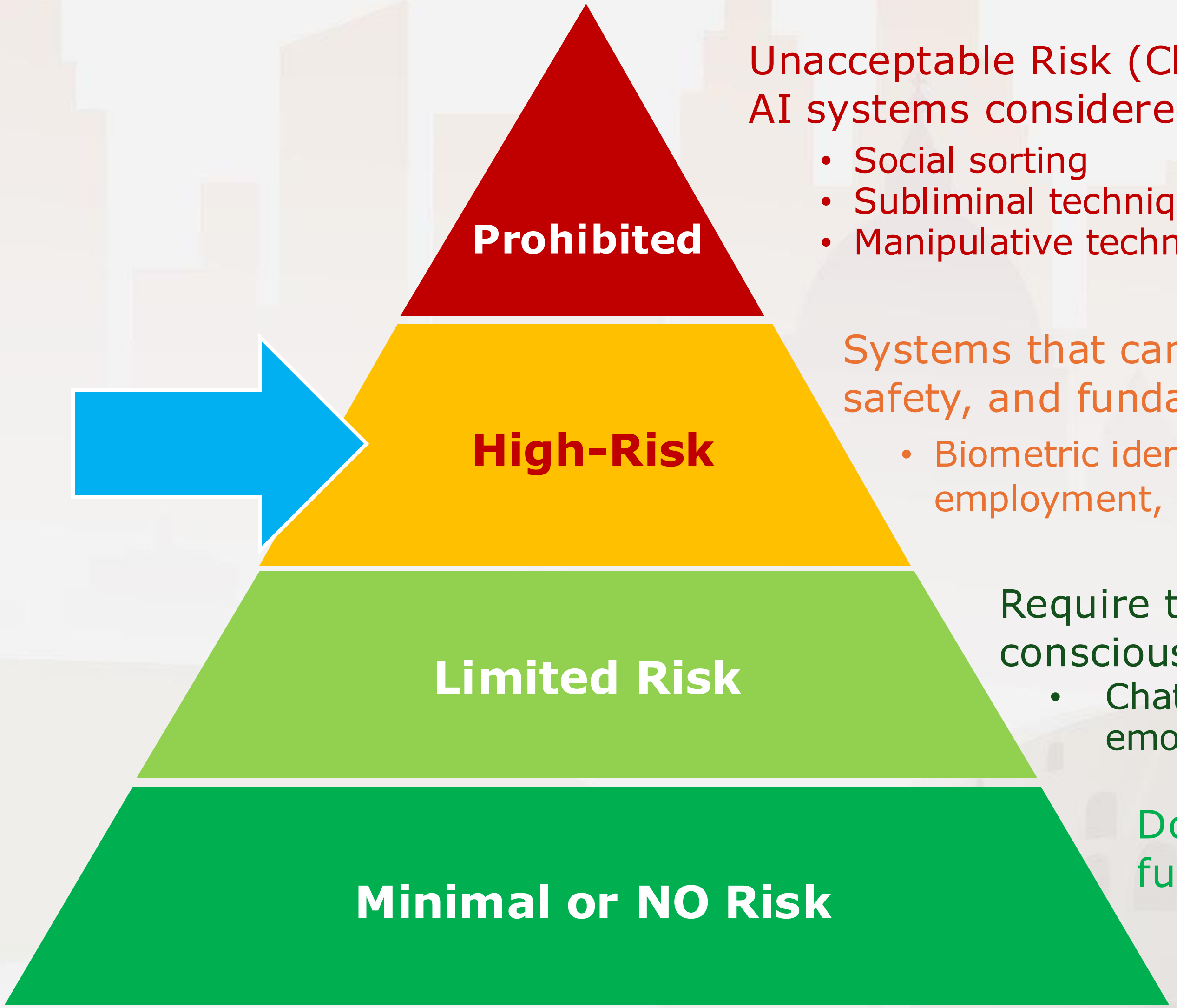
Similar to the GDPR, the approach is based on **risk assessment**.

The risk is the combination of the probability of harm occurring and the severity of the harm itself.

The regulation establishes sanctions for non-compliance, with fines that can reach the greater of **€ 35 million** and 7% of the company's global turnover.

It classifies artificial intelligence systems based on four categories of risk.

Risk-Based Classification of AI Systems



Unacceptable Risk (Chapter II)
AI systems considered a threat to fundamental rights.

- Social sorting
- Subliminal techniques
- Manipulative techniques to induce unwanted behavior

Systems that can have a significant impact on health, safety, and fundamental rights (Chapter III)

- Biometric identification, critical infrastructure, education, employment, essential services, migration, justice

Require transparency obligations to ensure conscious use by users (Chapter IV):

- Chatbots, Biometric categorization systems, emotion recognition, deep fake.

Does not present significant risks to safety or fundamental rights (Chapter V):

- videogames, Spam filters,...

High-Risk AI Systems

Art. 6

High-Risk

An AI system is considered high-risk if:

1) It is part of or falls under systems subject to safety assessment as specified in Annex I (Machinery Directive, toys, boats, lifts, medical devices, etc.)

High-Risk AI Systems

Annex III, Art. 6 **High-Risk**

2) An AI system is considered high-risk if it is used in a context provided for in **Annex III**:

- Biometrics/emotions (excluding authentication)
- Critical infrastructure (digital, traffic, electricity, gas,...)
- Education and vocational training
- Employment, worker management
- Essential private and public services (local authorities, healthcare,...)
- Law enforcement activities
- Migration, asylum, and border control management
- Administration of justice and democratic processes (elections/referendums)

Exemption for high-risk AI systems

(53), Art. 6 **High-Risk**

An AI system listed in Annex III is not considered high-risk if a prior risk assessment has documented that:

1. it does **not pose a significant risk of harm** to the health, safety, or fundamental rights of natural persons,
2. it does **not materially influence the outcome** of the decision-making process

This exemption applies when at least one of the following conditions is met:

a) The AI system has **limited involvement** in the process.

b) The AI system's purpose is **to improve the outcome** of a human activity already completed (i.e., the AI supports the human activity).

c) The AI system is **not intended to replace** or **influence** human assessment.

d) The AI system is only intended to perform a **preparatory task**.

Risk management system

Art. 9

High-Risk

In relation to high-risk AI systems, a risk management system is (1) established (a risk management system is actually put in place), (2) implemented, (3) documented (accountability), and (4) maintained (constant upkeep).

The **risk management system** is a continuous, iterative process planned and executed throughout the entire life cycle of a high-risk AI system, requiring constant and systematic review and updating. The phases are:

a) the identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose;

b) the estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse;

c) the evaluation of other risks possibly arising, based on the analysis of data gathered from the post-market monitoring system referred to in Article 72;

d) the adoption of appropriate and targeted risk management measures designed to address the risks identified pursuant to point (a)

Risk management system

(67)

High-Risk

High-quality data and access to high-quality data plays a vital role in providing structure and in ensuring the performance of many AI systems, especially when techniques involving the training of models are used, with a view to ensure that the high-risk AI system performs as intended and safely and it does not become a source of discrimination prohibited by Union law.

Biases can for example be inherent in underlying data sets, especially when historical data is being used, or generated when the systems are implemented in real world settings. Results provided by AI systems could be influenced by such inherent biases that are inclined to gradually increase and thereby perpetuate and amplify existing discrimination, in particular for persons belonging to certain vulnerable groups, including racial or ethnic groups.

Data and data governance

Art. 10 **High-Risk**

Article 10: «Training, validation and testing data sets shall be relevant, **sufficiently representative**, and to the best extent possible, free of errors and **complete** in view of the intended purpose. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used….»

How do you measure data quality and the risk that data bias can perpetuate forecasts?

Are there International Standards that can help us?

Who develops standards?

International Standardization Organizations

International Organization for Standardization

International Electrotechnical Commission

International Telecommunication Union

ISO JTC1 IEC
INFORMATION TECHNOLOGY STANDARDS

UNI ITALIAN STANDARDIZATION BODY

UNINFO

The Joint Technical Committee 1 (https://jtc1info.org/) is responsible for the standardization aspects of Information Technology (IT) for ISO and IEC

European Standardization Organizations

European Committee for Electrotechnical Standardization

COMITATO ELETTROTECNICO ITALIANO

European Committee for Standardization

European Telecommunications Standards Institute

Overview of the international AI standard

References

- Fairness Metrics and Maximum Completeness for the prediction of discrimination, IWESQ 2022
- Using the SQuaRE series as a guarantee for GDPR compliance IWESQ 2021
- Integrating SQuaRE data quality model with ISO 31000 risk management to measure and mitigate software bias IWESQ 2021
- Metrics for identifying bias in datasets ICYRIME 2021

INTRODUCTION
FAIRNESS METRICS
CASE STUDY
CONCLUSION & FUTURE

Life Cycle
- ISO/IEC/IEEE 12207:18 — Sw life cycle process

Fundation
- ISO/IEC 2382:15 — IT Vocabulary
- ISO/IEC 31000:18 — Risk Mgmt Guidelines
- ISO 9001:15 — Quality Mgmt Systems — Requir.

Software
- ISO/IEC **25000**:14 — System & SW engineering — **SQuaRE**
- ISO/IEC 29119:24 — SW testing (11 parts))

Life Cycle
- ISO/IEC TR 5338:23 — AI Life Cycle Process

Fundation
- ISO/IEC 22989:22 — IA concepts & term.
- ISO/IEC 23053:22 — Framework AI Systems Using ML

Yellow boxes represents the standards used in this research

AI Governance
- ISO/IEC 38507:22 — AI Governance implic.
- CEN/CLC TR **18115**:24 — AI Data governance & quality

Risk Mgmt
- ISO/IEC 23894:23 — AI Risk Mgmt

Functional Safety
- ISO/IEC TR 5469:24 — AI Functional Safety

Quality
- ISO/IEC CD TR 42106:xx — Overview benchmark vs AI quality characteristic
- ISO/IEC 25059:23 — AI Quality Model
- ISO/IEC 25058:24 — Guidance AI Quality evaluation

Data Quality (DQ) for Analytics & ML
- ISO/IEC 24668:22 — Process Mgmt Framework for Big Data Analytics
- ISO/IEC 5259-1:24 — DQ for Analytics and ML – Overview & examples
- ISO/IEC **5259-2**:24 — DQ for Analytics and ML – Model & measures
- ISO/IEC 5259-3:24 — DQ for Analytics and ML – Mmgt Req. & Guideline
- ISO/IEC 5259-4:24 — DQ for Analytics and ML – Process Framework
- ISO/IEC 5259-5:25 — DQ for Analytics and ML – Governance Framework
- ISO/IEC CD TR 5259-6:xx — DQ for Analytics and ML – Visualization Framework

Trustworthiness
- ISO/IEC TR 24028:20 — AI Trustworthiness
- ISO/IEC TR **24027**:21 — AI Bias
- ISO/IES DTS 6254:xx — Objectives for AI explainability and interpretability
- ISO/IEC TS 8200:24 — AI controllability
- ISO/IEC TR 24368:22 — AI Overview of ethical concerns

Testing
- ISO/IEC TR 29119-11:24 — Guidelines AI Testing
- ISO/IEC 42119:xx — AI Testing

AI Mgmt System
- ISO/IEC 42001:23 — AI – Mgmt System
- ISO/IEC 42005:xx — AI impact assessment
- ISO/IEC 42006:xx — Req. for audit & cert.

Assessment
- ISO/IEC TR 24029:21 — Robustness of Neural Network
- ISO/IEC TS **4213**:22 — ML performance

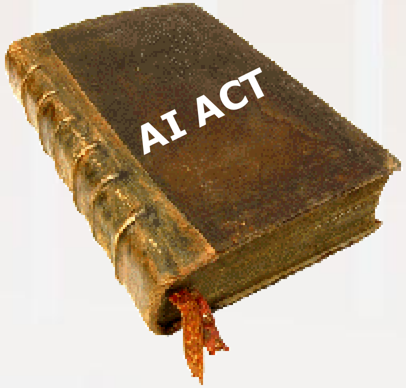
Application
- ISO/IEC TR 24030:24 — AI use cases
- ISO/IEC 5339:24 — AI Application Guidance
- ISO/IEC TR 24372:21 — Overview of computational approaches
- ISO/IEC 5392:24 — Ref. Architecture of knowledge eng.

Big Data
- ISO/IEC 20546:19 — Overview & Vocabulary
- ISO/IEC 20547:20 — Series (5 parts)

How to use international standard to be compliant to regulation in the era of AI

From the law versus harmonized standards

European Commission



STANDARDIZATION REQUEST

European Standardization Organizations

CEN

CENELEC

ETSI

HARMONIZED STANDARD

A harmonized standard (hEN) is a European Standard (EN) developed in response to a formal Standardization Request (SR) of the EC

INTRODUCTION
FAIRNESS METRICS
CASE STUDY
CONCLUSION & FUTURE

AI Act standardisation request

23 June 2025

List of new harmonised standards and European standardisation deliverables to be drafted on:

1. **risk management** systems for AI systems
2. **governance and quality** of datasets used to build AI systems
3. **record** keeping through **logging** capabilities by AI systems
4. **transparency** and information provisions for users of AI systems
5. **human oversight** of AI systems
6. **accuracy** specifications for AI systems
7. **robustness** specifications for AI systems
8. **cybersecurity** specifications for AI systems
9. **quality management systems** for providers of AI systems, including post-market monitoring processes
10. **conformity assessment** for AI systems

https://ec.europa.eu/transparency/documents-register/detail?ref=C(2025)3871&lang=en

The blueprint of my career

INTRODUCTION

FAIRNESS METRICS

CASE STUDY

CONCLUSION & FUTURE

Public Sector Employee — 26 years

01/09/22 – today — **3 years**

- Inspector General's Staff
- Director of IT, Logistics, and Digital Transition **13/10/22 – 14/07/24**
- Head of Digital Transition (Art. 17, Italian Law Code of Digital Administration)

02/03/99 –31/08/22 — **23 years**

- IT Professional
- Sector Coordinator - Technology Innovation Consulting

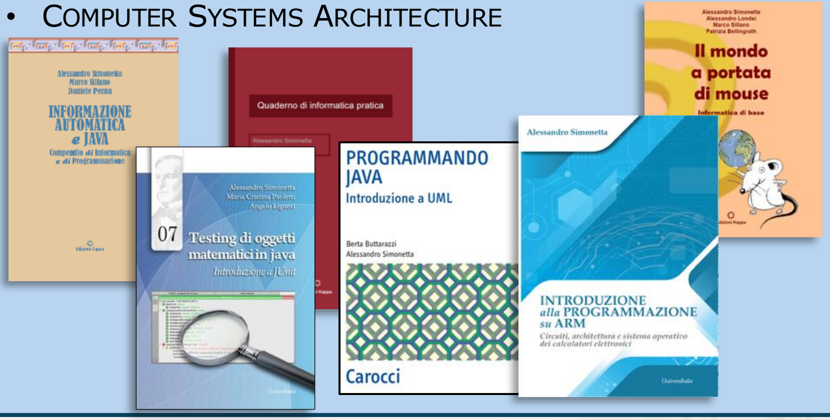
Private Sector Employee — 4 years

01/02/95 - 30/06/95

01/07/95 - 01/03/99

Contract Professor — 24 years

- Computer Literacy **A.A. 2001/2002**
- Information Theory and Digital Processing Techniques
- IT Techniques
- Programming Fundamentals
- Computers and Operating Systems
- Computer Systems Architecture **A.A. 2025/2026**



13/05/09 – today

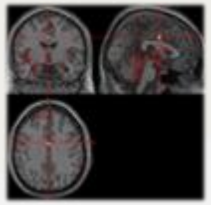
UNI/CT 504 Software engineering
UNI/CT 510 Secutity
UNI/CT 533 Artificial Intelligence

ISO/IEC JTC1 SC7 Software and systems engineering
ISO/IEC JTC1 SC27 Information security, cybersecurity and privacy
ISO/IEC JTC1 SC 42 CEN/CLC JTC 21

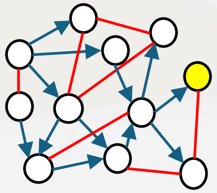
16 years

1994

Prof. Daniele Nardi



fMRI



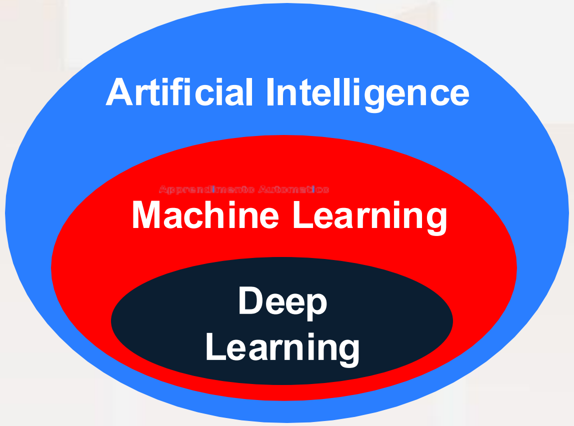
BCI

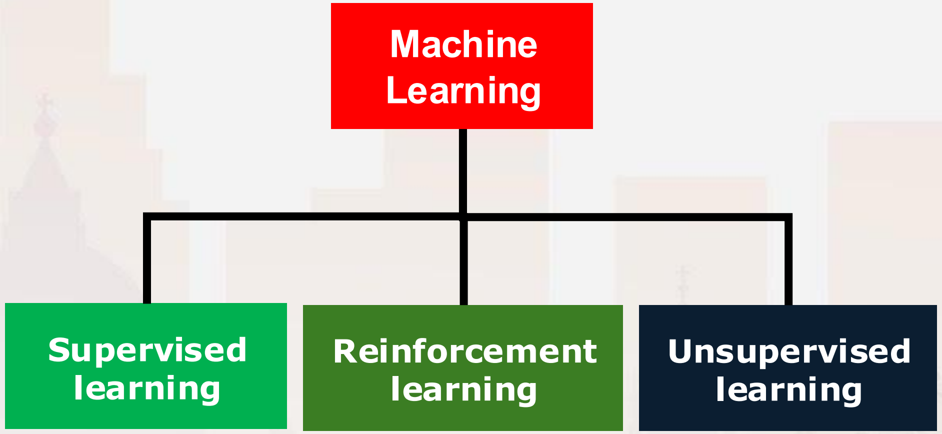
JOIV INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

INTERNATIONAL JOURNAL ON ADVANCED SCIENCE ENGINEERING INFORMATION TECHNOLOGY

Editorial Team

Artificial intelligence systems





Supervised learning: A model is built from already labeled training data, allowing predictions to be made.

Reinforcement learning: A system is built that improves its performance based on interactions with the environment, through a reward signal. This is a special case of supervised learning.

Unsupervised learning: The system is able to extract useful information from the data without the guidance of a result label or a reward function. It is capable of identifying clusters in the data.

INTRODUCTION

FAIRNESS METRICS

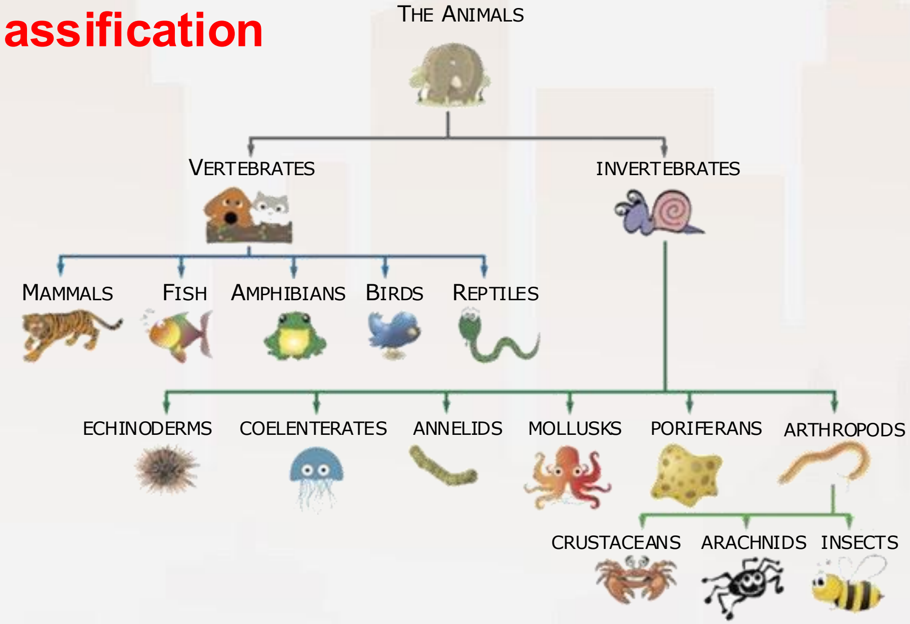
CASE STUDY

CONCLUSION & FUTURE

Supervised learning

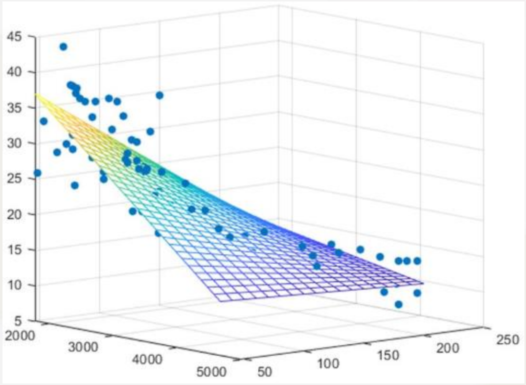
Supervised learning is typically divided into two categories depending on the domain of the target variables' values: discrete or continuous.

Classification



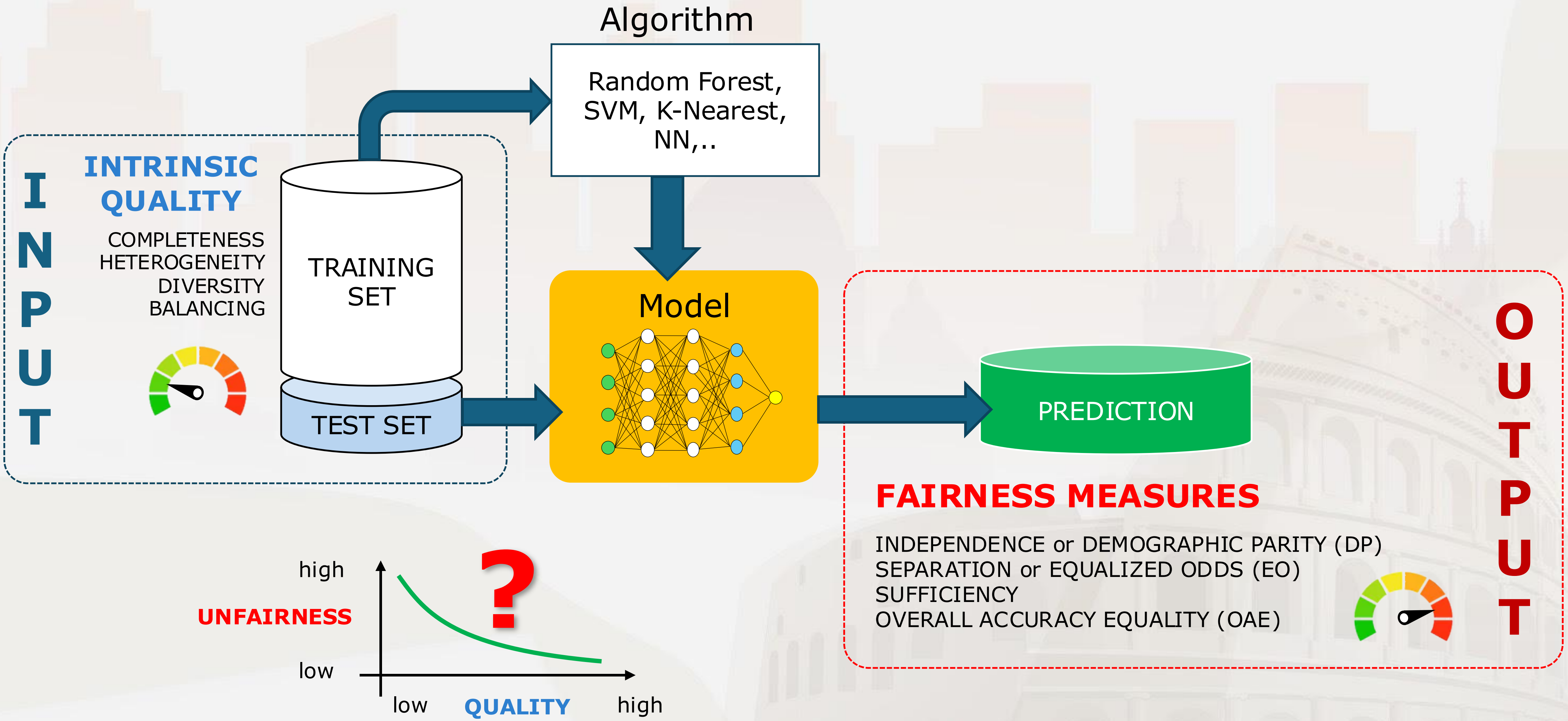
It deals with categorical target variables, which represent **discrete** classes or labels. For example, classifying emails as spam or non-spam, or predicting the necessity of replacing a mechanical part based on wear. Classification algorithms are able to match the input features to a discrete value within the scope of predefined classes.

Regression



It deals with **continuous** numerical target variables. For example, predicting an insurance premium based on the risk type, or the probability of contracting a disease based on analysis results. Regression algorithms are able to match the input features to a real numerical value

Where and what to measure?



Formalization of classification

https://fairmlbook.org/classification.html

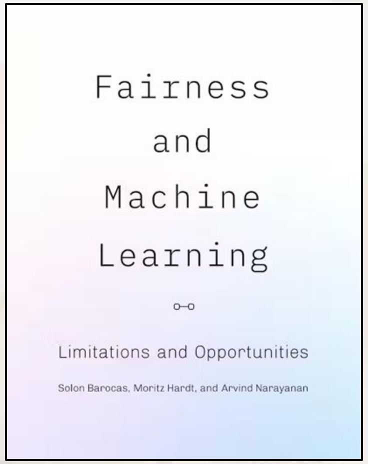
The objective of classification is to determine a plausible value for an unknown **target variable** Y based on the observed **covariates** X:

$$\hat{Y} = f(X)$$

The function f is called the *classifier* or *predictor*.

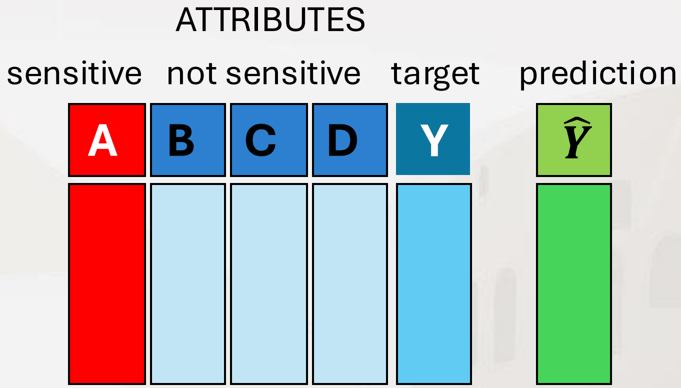
The output of the classifier is called the *label* or *prediction*

The covariates X and the target Y are **random variables** jointly distributed. This means that there is a probability distribution over pairs of values (x,y) that the random variables (X,Y) can assume. This probability distribution models a population of instances of the classification problem



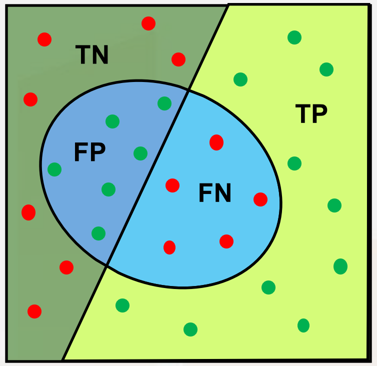
Solon Barocas, Moritz Hardt, Arvind Narayanan

$$\hat{Y} = \begin{cases} 1 & R \geq t \\ 0 & R < t \end{cases}$$



INTRODUCTION
FAIRNESS METRICS
CASE STUDY
CONCLUSION & FUTURE

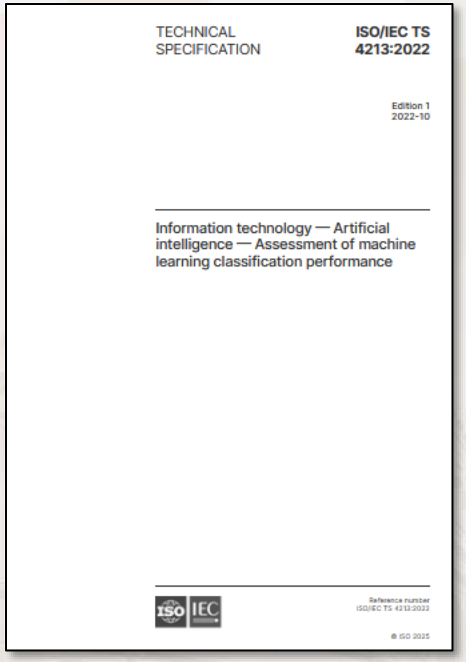
Core elements for metric computation



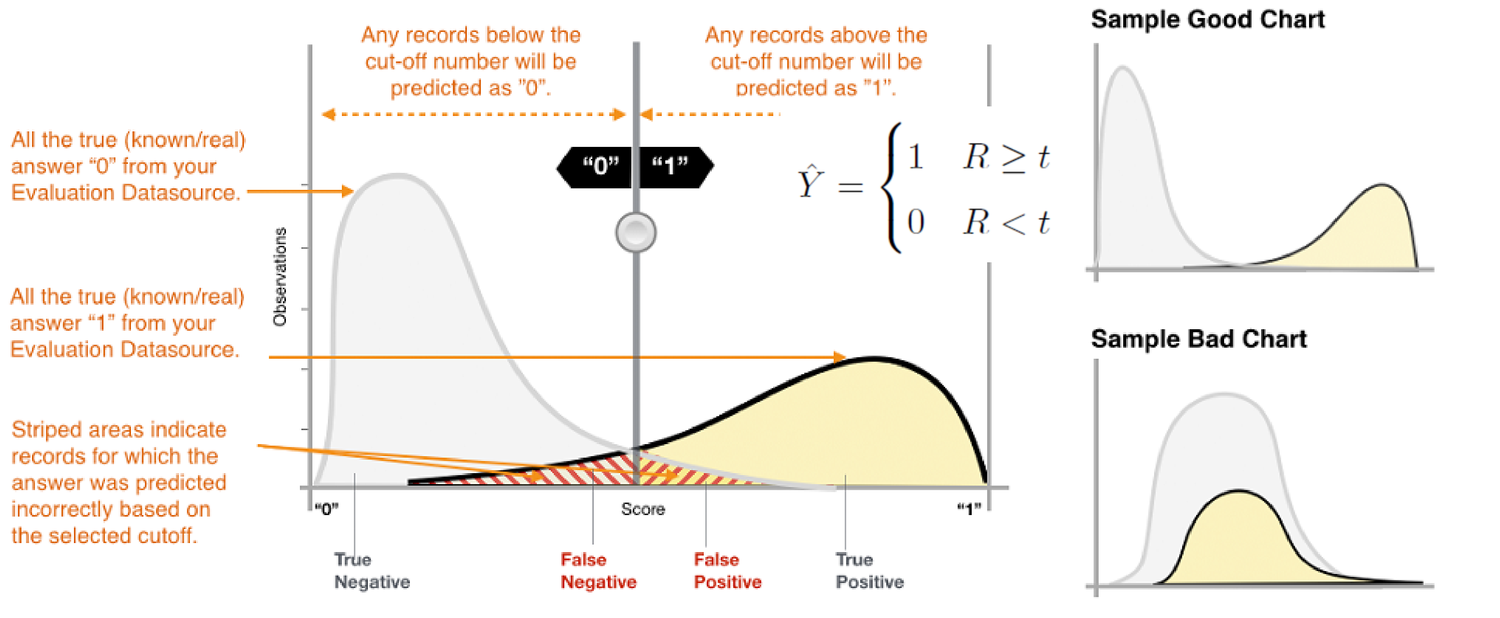
Common classification criteria

Event	Condition	Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$)	
$\hat{Y} = 1$	$Y = 1$	True positive rate, recall	sensitivity
$\hat{Y} = 0$	$Y = 1$	False negative rate	miss rate
$\hat{Y} = 1$	$Y = 0$	False positive rate	fall-out
$\hat{Y} = 0$	$Y = 0$	True negative rate	specificity

ISO/IEC TS 4213:22
Information technology — Artificial intelligence — Assessment of machine learning classification performance



This publication was last reviewed and confirmed in 2025



$$\hat{Y} = \begin{cases} 1 & R \geq t \\ 0 & R < t \end{cases}$$

https://developers.google.com/machine-learning/crash-course/classification?hl=it

Core elements for metric computation

Accuracy and **precision** are two of the data quality characteristics recognized in the **ISO/IEC 25012** and **5259-2** standards (Data quality measures), and the calculation method is described in the **ISO/IEC TS 4213** standard.

Confusion matrix

	$\hat{Y}=1$	$\hat{Y}=0$
Y=1	*TP*	*FN*
Y=0	*FP*	*TN*

Equalized Odd

$TPR = \frac{TP}{TP+FN} = 1 - FNR$ *True Posivitve Rate or **Recall** or **Sensitivity***	$FNR = \frac{FN}{TP+FN} = 1 - TPR$ *False Negative Rate or False Rejection Rate (FRR) (*)*
$FPR = \frac{FP}{FP+TN} = 1 - TNR$ *False Positive Rate or True Acceptance Rate (TAR) (*)*	$TNR = \frac{TN}{FP+TN} = 1 - FPR$ *True Negative Rate or **Specificity***

$PPV = \frac{TP}{TP+FP} = 1 - FDR$ *Positive Predictive Value or **Precision***	$FOR = \frac{FN}{FN+TN} = 1 - NPV$ *False Omission Rate*
$FDR = \frac{FP}{TP+FP} = 1 - PPV$ *False Discovery Rate*	$NPV = \frac{TN}{FN+TN} = 1 - FOR$ *Negative Predictive Value*

Sufficiency

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Prevalence = \frac{TP+FN}{TP+FP+TN+FN}$$

(*) Advances in Computer Vision and Pattern Recognition
Handbook of BiometricAnti-Spoofing
Trusted Biometrics under Spoofing Attacks
Sébastien Marcel, Mark S. Nixon, Stan Z. Li

Core fairness measurement

$\mathbb{P}\{\hat{Y} = 1 \mid A = a_i\} = \mathbb{P}\{\hat{Y} = 1 \mid A = a_j\}\ \forall_{i,j}$	IND	Independence Demographic Parity (DP)
$\mathbb{P}\{\hat{Y} = Y \mid A = a_i\} = \mathbb{P}\{\hat{Y} = Y \mid A = a_j\}\ \forall_{i,j}$	OAE	Overall Accuracy Equality
$\mathbb{P}\{\hat{Y} = 1 \mid Y = 1, A = a_i\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 1, A = a_j\}\ \forall_{i,j}$ $\mathbb{P}\{\hat{Y} = 1 \mid Y = 0, A = a_i\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 0, A = a_j\}\ \forall_{i,j}$	TPR FPR	Equalized Odds (EO) Separation
$\mathbb{P}\{Y = 1 \mid \hat{Y} = 1, A = a_i\} = \mathbb{P}\{Y = 1 \mid \hat{Y} = 1, A = a_j\}\ \forall_{i,j}$ $\mathbb{P}\{Y = 1 \mid \hat{Y} = 0, A = a_i\} = \mathbb{P}\{Y = 1 \mid \hat{Y} = 0, A = a_j\}\ \forall_{i,j}$	PPV NPR	Sufficiency

Fairness impossibility theorem

If the prevalence of a positive outcome is different among groups, AI system developers must make an ethical choice about which form of fairness to implement, accepting that the system **will be intrinsically "unfair"** (or inequitable) according to the other definitions.

The choice of an index is a decision on where the residual bias will fall:

- If **Demographic Parity** (DP) is chosen, it is accepted that the model may be more or less accurate for one group compared to another (violating Equalized Odds).
- If **Equalized Odds** (EO) are chosen, it is accepted that one group may have fewer positive outcomes overall (violating Demographic Parity).

Demographic Parity (DP) ⟷ Equalized Odds (EO)

Fairness impossibility theorem

If a classifier satisfies both DP and EO, the following occurs:

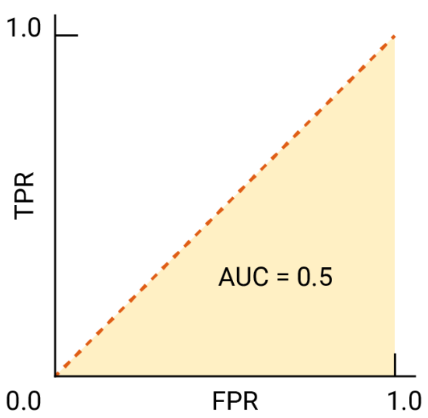
$$\Delta PR \cdot \Delta \mathbb{P} = 0$$

where:

$$\Delta PR = TPR - FPR$$
$$\Delta \mathbb{P} = [\mathbb{P}(Y = 1|A = a_i) - \mathbb{P}(Y = 1|A = a_j)]$$

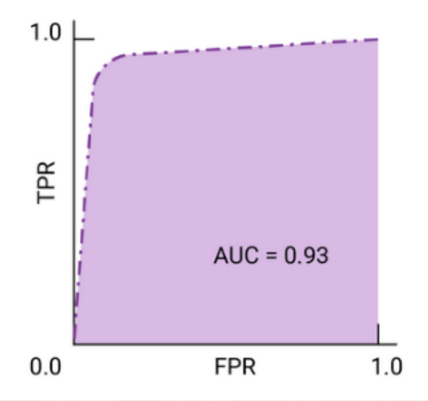
condition 1 $TPR = FPR$

This condition is not a goal, **but a failure**. This means that the algorithm does not discriminate between real positive and negative outcomes.



condition 2 $\Delta \mathbb{P} = 0$

The condition of equal prevalence means that $\mathbb{P}(Y = 1|A = 0) = \mathbb{P}(Y = 1|A = 1)$, meaning **the prevalence of the real positive outcome is exactly the same** in both protected groups. This eliminates the conflict between DP and EO and allows the classifier to be both useful $TPR \gg FPR$) and fair.



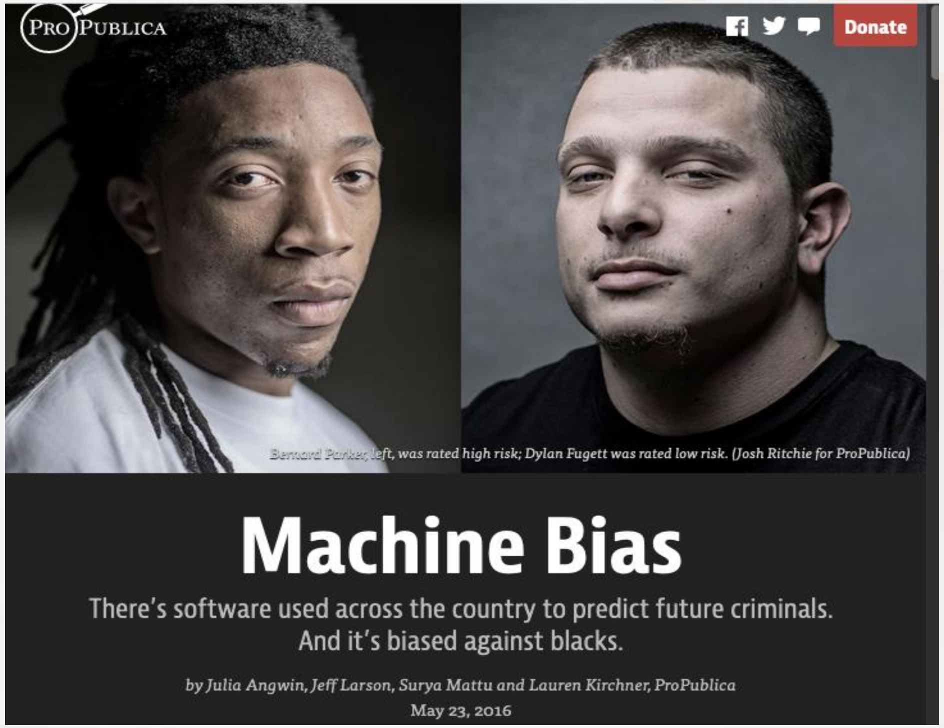
The measure of independence (demographic parity)

Without loss of generality, to explain our work, we will consider the measure of independence (or demographic parity), which is the most intuitive. Independence is defined between two groups of sensitive attributes as the distance between the two probabilities:

$$Ind(a_i\,,a_j) = \left|\mathbb{P}\{\hat{Y} = 1|A = a_i\} - \mathbb{P}\{\hat{Y} = 1|A = a_j\}\right|$$

1. to find a way to synthesize the notion of independence into a single dimensionality, since it is defined between pairs of values of a sensitive attribute (measure on the OUTPUT)
2. to identify an index that highlights poor quality in the training sets (measure on the INPUT)
3. to study the relationship that exists between poor quality in the training sets and an unfair outcome, and, if possible, to anticipate the probability that a bias in the data may perpetuate itself into the predictions (INPUT vs OUTPUT).
4. to nullify the harmful effects of bias in the data when no further learning data is available

Case study: COMPAS dataset



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

In this famous case, the system incorrectly predicted a higher degree of recidivism for African-American defendants, while in reality, it was the white defendants who had a greater propensity to reiterate crimes.

Calculation of independence among different ethnicities

Recalling the notion of independence between two groups of sensitive attributes:

$$Ind(a_i, a_j) = \left|\mathbb{P}\{\hat{Y} = 1|A = a_i\} - \mathbb{P}\{\hat{Y} = 1|A = a_j\}\right|$$

$A = a_i$	$P(\hat{Y}=1\|A = a_i)$
Caucasian	0.33
Hispanic	0.28
Other	0.20
Asian	0.23
African-American	0.58
Native-American	0.73

	Asian	Caucasian	Hispanic	Native American	Other
African-American	35,03%	24,51%	29,90%	15,12%	37,20%
Asian		10,51%	5,12%	50,15%	2,17%
Caucasian			5,39%	39,63%	12,69%
Hispanic				45,03%	7,29%
Native American					52,32%

1 Identifying a single indicator (Output measure)

The ideal situation of perfect fairness for a sensitive attribute A occurs when all probabilities $\mathbb{P}\{\hat{Y} = 1 | A = a_i\}$ are equal.

A single independence index representative of the set of independence measures between pairs of sensitive attribute values can be calculated by:

- **A** Considering the mean of the independence values between all pairs of sensitive attribute values
- **B** In relation to the distance between groups of probabilities that are similar in treatment
- **C** Based on the maximum disparity between the probability values $\mathbb{P}\{\hat{Y} = 1 | A = a_i\}$
- **D** Utilizing the notion of mutual information and entropy

A Use of the average of the distances

Some authors have proposed the average of the independence measures:

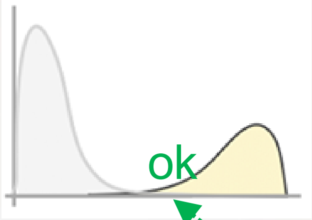
$$\mathfrak{U}(a_1, \ldots, a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} |P(R=1 \mid A=a_i) - P(R=1 \mid A=a_j)|$$

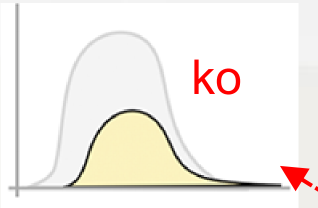
$$Ind(A) = \frac{2}{m\cdot(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} Ind(a_i, a_j) = 24{,}8\ \%$$

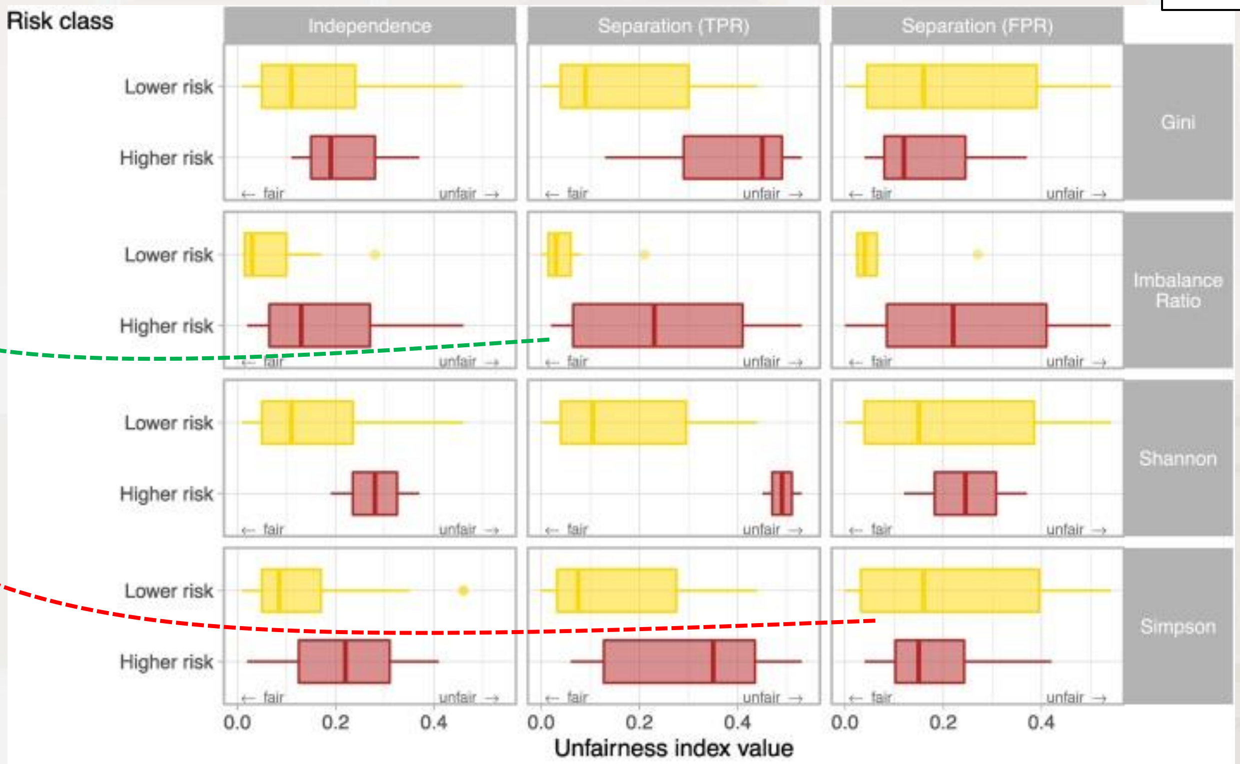
Government Information Quarterly
Volume 38, Issue 4, October 2021, 101619

A data quality approach to the identification of discrimination risk in automated decision making systems

Antonio Vetrò, Marco Torchiano, Mariachiara Mecati







Evidence of different treatment groups

The confirmation of the existence of multiple treatment groups is proven by considering the concept of fairness as a multi-dimensional vector whose components are the individual fairness measures (independence, TPR, FPR, PPV, NPV, and OAE).The relationships among the vectors, calculated using Pearson's index, allow for the grouping of ethnicities based on analogies of treatment

Race	Fairness Index					
	Ind.	*SepTPR*	*SepFPR*	*SufPPV*	*SufNPV*	*OAE*
Caucasian	33,10%	50,36%	22,01%	59,48%	29,00%	67,19%
Hispanic	27,70%	41,80%	19,38%	56,03%	29,89%	66,21%
Other	20,41%	33,87%	12,79%	60,00%	30,04%	67,93%
Asian	22,58%	62,50%	8,70%	71,43%	12,50%	83,87%
African American.	57,61%	71,52%	42,34%	64,95%	35,14%	64,91%
Native American.	72,73%	100%	50%	62,50%	0,00%	72,73%

Race	Race					
	African-A.	*Native A.*	*Caucasian*	*Hispanic*	*Other*	*Asian*
African-American	1	0,901	0,801	0,680	0,562	0,848
Native American	0,901	1	0,515	0,364	0,210	0,596
Caucasian	0,801	0,515	1	0,983	0,941	0,991
Hispanic	0,680	0,364	0,983	1	0,984	0,956
Other	0,562	0,210	0,941	0,984	1	0,900
Hispanic	0,848	0,596	0,991	0,956	0,900	1

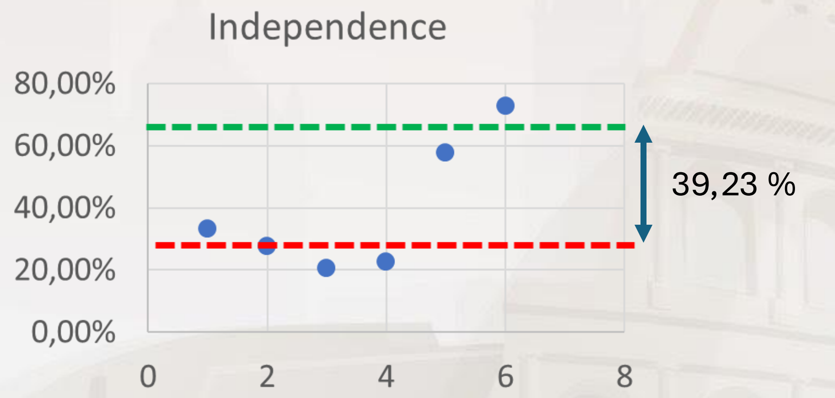
Groups	Races			
	1	*2*	*3*	*4*
G0	African-American	Native American		
G1	Caucasian	Hispanic	Other	Asian

B The notion of independence between groups

By utilizing clustering techniques (such as the DBSCAN algorithm), it is possible to group data into homogeneous treatment areas, which will also have similar probability values $\mathbb{P}\{\hat{Y} = 1|A = a_i\}$.

The cluster centers (or centroids) can then be chosen as representatives of the groups, and the average distance can be applied. In the COMPAS dataset, the conditional probabilities cluster around two centroids representing the two treatment groups.

$A = a_i$	$P\ (\hat{Y}=1\|A = a_i)$	Centroid
Caucasian	0.33	
Hispanic	0.28	0.26
Other	0.20	
Asian	0.23	
African-American	0.58	0.65
Native-American	0.73	

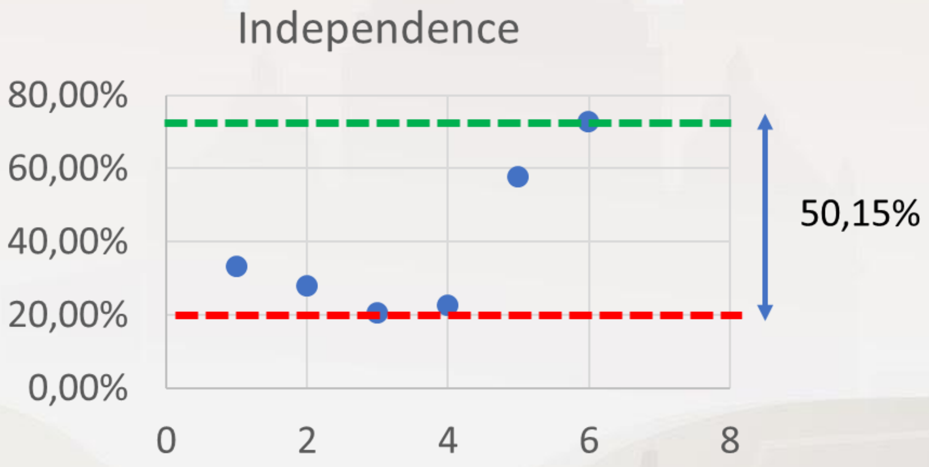


This signifies an independence fairness within the Group, and therefore an unfairness between groups

C

Maximum treatment disparity (MaxMin algorithm)

This method is based on the idea that a measure of the extent of unfairness can be calculated by considering the worst-case scenario, which is the highest independence value.



Mutual information and entropy

Given two random variables, they are **statistically independent** if their mutual information is zero:

$$I(\hat{Y}, A) = 0$$

We can calculate the mutual information using the notion of entropy:

$$I(\hat{Y}, A) = H(\hat{Y}) + \mathrm{H(A)} - \mathrm{H}(\hat{Y}, \mathrm{A})$$

Specifically, from which:

$$\sum_{i=1}^{n} P(r_i)log(P(r_i)) + \sum_{i=1}^{n} P(a_i)log(P(a_i) - \sum_{i=1,j=1}^{n,m} P(r_i \cap a_j)log(P(r_i \cap a_j))$$

By utilizing this approach, it is possible to calculate all the fairness indices analyzed previously (independence, sufficiency, separation, and overall accuracy parity)

2 Metrics for assessing training data quality (Input measure)

The main statistical indices used in literature are:

- **Gini Index** (or **Gini Coefficient**): Measures the inequality in the distribution (heterogeneity) of a variable in a dataset.
- **Shannon Index** (or **Shannon-Weaver Index**): Measures the species diversity in an ecosystem.
- **Imbalance Ratio**: Represents the degree to which one data class is more frequent than another class.
- **Simpson Index**: Measures the diversity of a group of individuals, specifically the probability that two individuals chosen at random belong to the same species.

Gini Index (Heterogeneity)	$\mathcal{G}_n = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^{m} f_i^2\right)$	**Shannon Index (Diversity)**	$H' = -\frac{1}{\ln m} \sum_{i=1}^{m} f_i \ln f_i$
Imbalance Ratio	$I_n = \frac{min\, f_i}{max\, f_i}$	**Simpson Index (Diversity)**	$D_n = \frac{1}{m-1}\left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$

2 The notion of data completeness

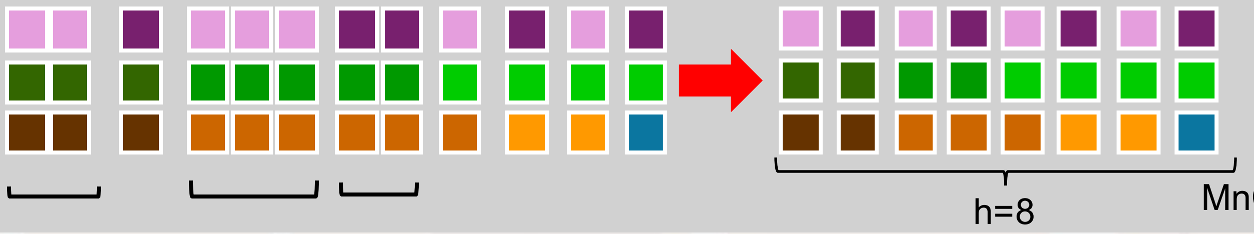
The notion of data completeness is present in ISO/IEC 25012 and has also been included in ISO/IEC 5259-2 (Artificial intelligence — Data quality for analytics and machine learning (ML) - Part 2: Data quality measures).

Minimal data Completeness	Measures the ratio between the distinct number of tuples with sensitive attributes and the ideal case where the dataset contains all possible distinct combinations of sensitive attributes
Maximal data Completeness	Measures the ratio between the number of tuples with sensitive attributes and the ideal case where the dataset contains the number of distinct combinations of sensitive attributes repeated exactly the number of times of the predominant combination

Calculation of minimal data completeness

Example dataset to analyze

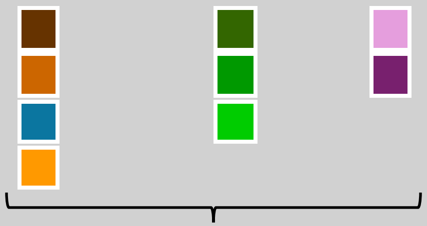
Sex
Age Group
Ethnicity



MnCI = 33.34%

Sensitive Attributes

Ethnicity CS0	Age Group CS1	Sex CS2

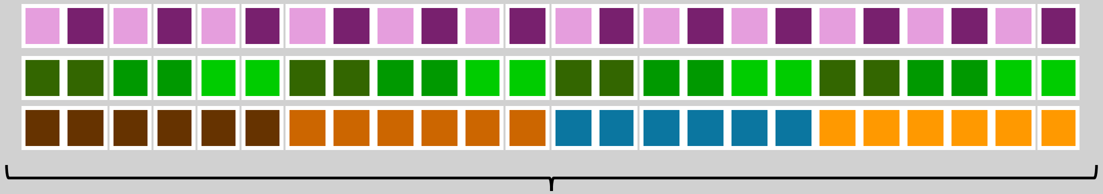



```
k=(   df['CS0'].unique().size *
      df['CS1'].unique().size *
      df['CS2'].unique().size )
h=len (df.drop_duplicates())
```

Minimal data completeness

MnCI = h / k

Ideal dataset

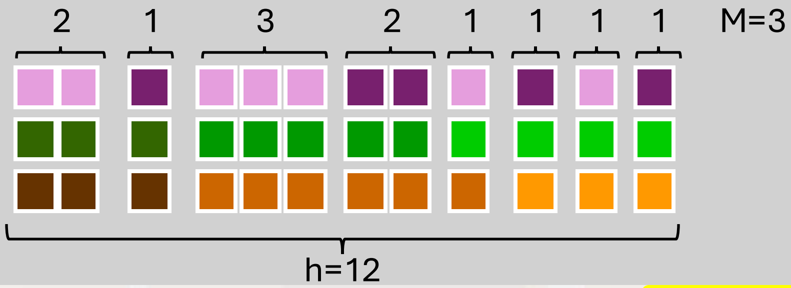


100%

2

Calculation of maximal data completeness

Example dataset to analyze



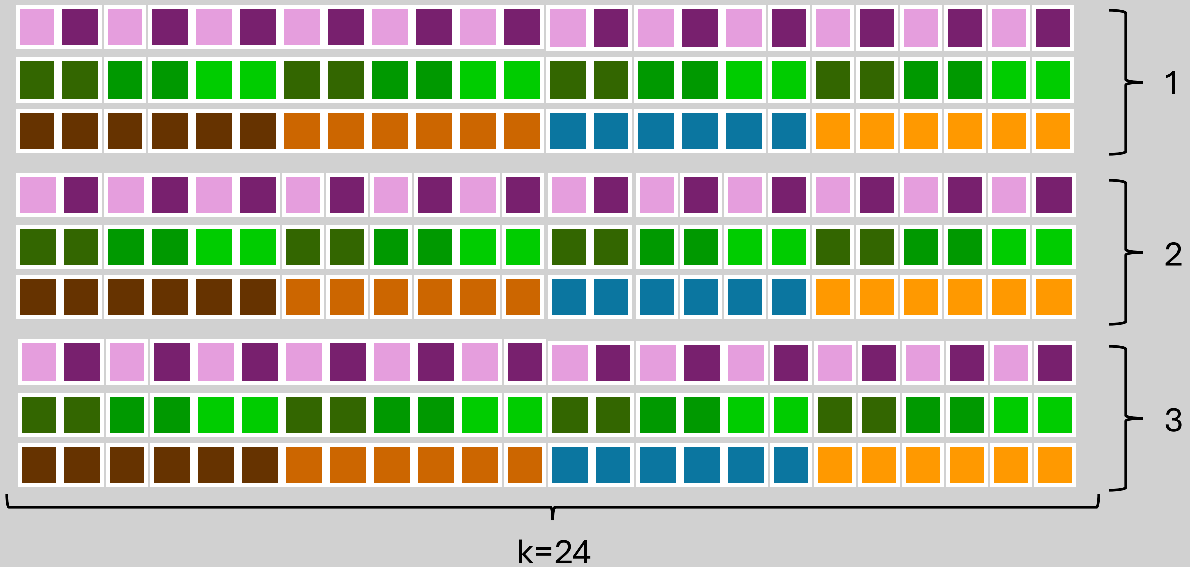
MxCI = 16.67%

```
M=df.groupby(['CS0','CS1','CS2']).size()
.reset_index(name='counts').counts.max()
h=len (df)
```

Maximal data completeness

MxCI = h /(M*k)

Ideal dataset



100%

3 The relationship between poor data quality and unfair outcome

The challenge lies in the ability to anticipate treatment disparities in the results of an AI system by evaluating the quality of the data in the training set.

The idea is to find **predictive markers** to confine the risk that a defect in the data might propagate within the learning system, perpetuating, or even amplifying, societal prejudices concerning ethnic minorities, gender, etc.

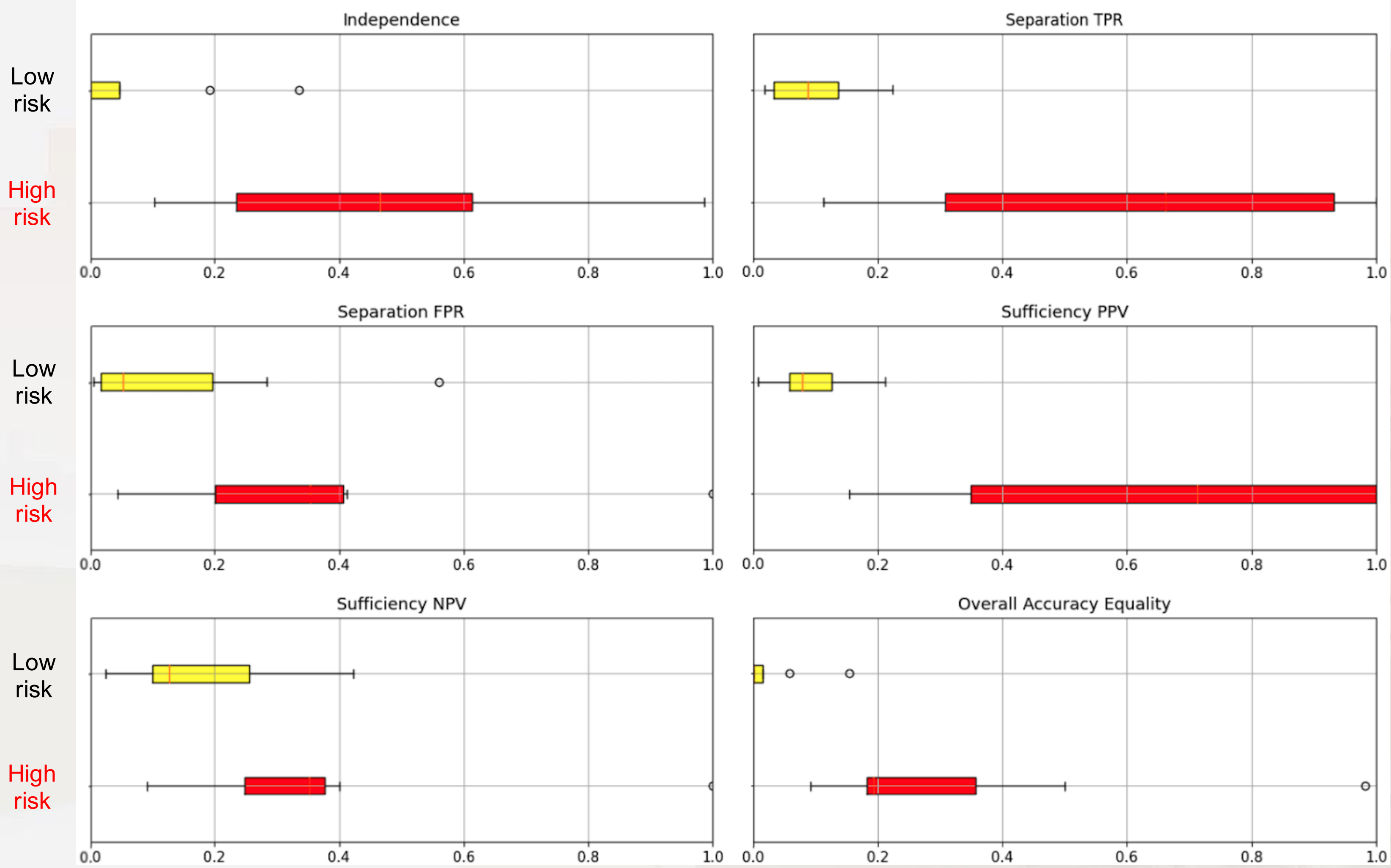
This is the goal of the present research…

Datasets used in this research

- COMPAS Recidivism Dataset
  https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
- Recidivism in juvenile justice
  https://www.ojjdp.gov/ojstatbb/compendium/
- UCI Statelog German Credit
  https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data
- default of credit card clients Data Set
  https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
- Adult Data Set
  https://archive.ics.uci.edu/dataset/2/adult
- Student Performance Data Set
  https://archive.ics.uci.edu/dataset/320/student+performance

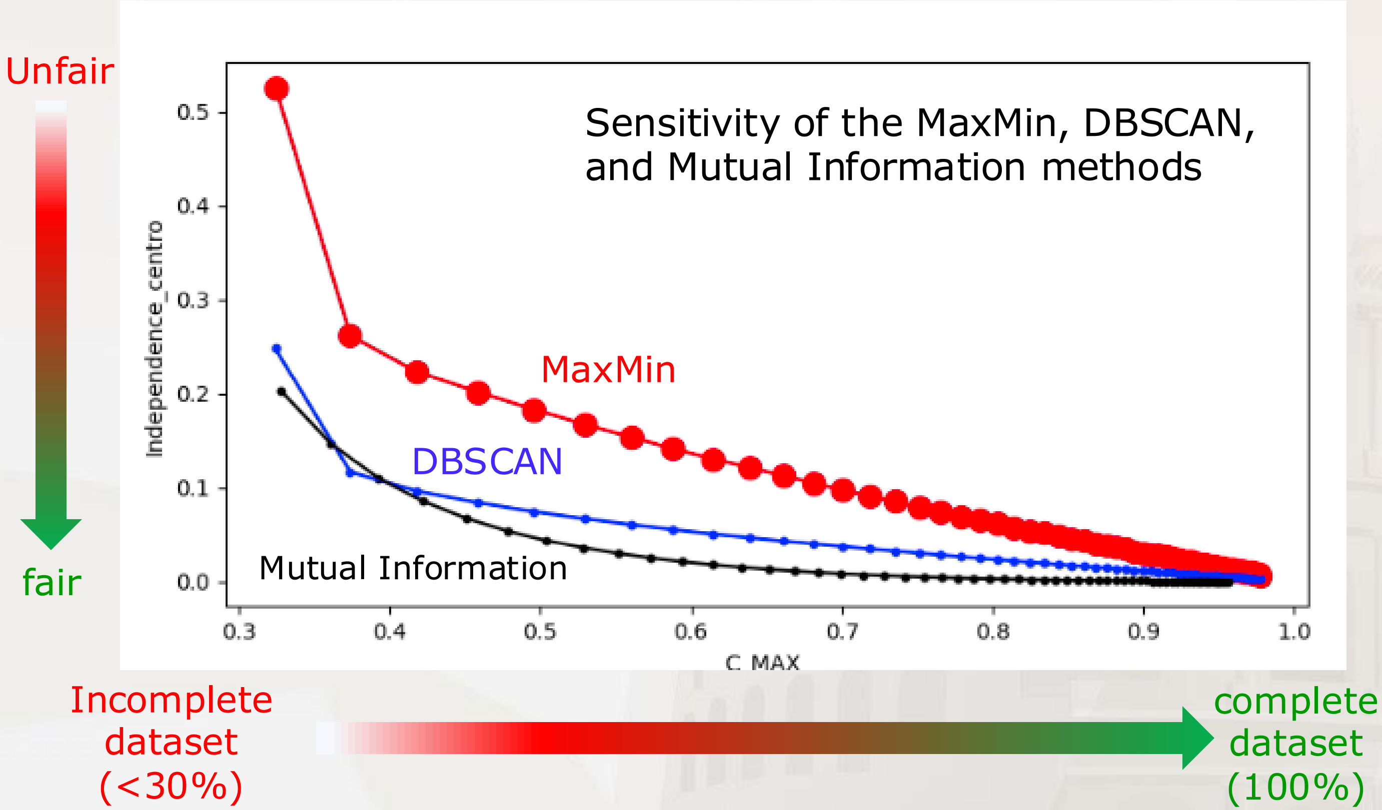
Etica nei sistemi AI

3 Relationship between Cmax and fairness measures calculated with MaxMin



4 Mitigating the harmful effects of bias

When it's not possible to enrich the dataset with new elements, statistical data augmentation techniques can be used. Among these, an example is bootstrapping, which utilizes a resampling method with replacement to generate new datasets from an original sample



Conclusions and future work

The use of AI systems in decision-making processes presents the risk of perpetuating, or even amplifying, the prejudices present in the data (Whereas (67)).

The presence of incomplete or imbalanced data can lead to biased results.

While awaiting the production of harmonized standards, we can use existing international standards for informed deployment and to comply with regulation.

However, statistical metrics fail to distinguish between discrimination and pre-existing inequality; therefore, causal models can be employed to analyze bias through the notion of Counterfactual Fairness. This will be a new line of study to address.

Bibliography

1) "*ISO/IEC standards and design of an Artificial Intelligence system*", https://ceur-ws.org/Vol-3916/ PP.39-43, Simonetta A., Paoletti M.C.
2) "*ISO/IEC quality standards for AI engineering*", Oviedo J., Rodriguez M., Trenta A., Cannas D., Natale D., Piattini M., Computer Science Review, Volume 54, November 2024
3) "*The SQuaRE series as a guarantee of ethics in the results of AI systems*",https://ceur-ws.org/Vol-3612, Simonetta A., Paoletti M.C., Nakaijma T.
4) "*Etica nei sistemi di Intelligenza Artificiale*", https://www.inail.it/cs/internet/comunicazione/pubblicazioni/catalogo-generale/pubbl-atti-sem-aggior-prof-ctss-csa-cit-salute-sicurezza.html, Paoletti M.C., Simonetta A., Natale D.
5) "*Fairness Metrics and Maximum Completeness for the prediction of discrimination*", https://ceur-ws.org/Vol-3356/ PP.13-20, Simonetta A., Nakaijma T., Paoletti M.C., Venticinque A.
6) "*The use of Maximum Completeness to estimate bias in AI based recommendation systems*", https://ceur-ws.org/Vol-3360, Simonetta A., Paoletti M.C., Venticinque A.
7) "*Using the SQuaRE series as a guarantee for GDPR compliance*", http://ceur-ws.org/Vol-3114, Simonetta A., Paoletti M.C., Venticinque A.
8) "*Integrating SQuaRE data quality model with ISO 31000 risk management to measure and mitigate software bias*", http://ceur-ws.org/Vol-3114, Simonetta A., Vetrò A., Paoletti M.C., Torchiano M.
9) "*Metrics for identifying bias in datasets*", https://ceur-ws.org/Vol-3118/, Simonetta A., Trenta A., Paoletti M.C., Vetrò A
10) "*A data quality approach to the identification of discrimination risk in automated decision making systems*", Government Information Quarterly, Volume 38, Issue 4 October 2021, Vetrò A., Torchiano M., Mecati M.
11) "*Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data*", Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021, Mecati M., Vetrò A., Torchiano M.,

Kiitos Kasih Diolch Sheun Shnorhakalutiun Teşekkür Hvala Todah Shokrun Mamnoon Spaas Dekuju/Dekujeme Shokriya or Faleminderit Dhanyavaad Dank Gamsahapnida Ngiyabonga Ačiū Grazie Takk umesc Cam on Xie Xie Dankie Dakujem Waad Dhanyvaalu Salamat Kop Grazie Merci Kruthagnathalu Dhanyavad Dziękuję Daw Thank you kun Mul Arigatou krap Or Shukriya Dhonnobaad Tack Nandree Gracias Blagodariya Gomapsupnida dank Euxaristo Khopjai ederim Grazzi Fyrir Terima Danke Kun Asante raibh Enkosi daa Hain Dhan

prof. Alessandro Simonetta
alessandro.simonetta@gmail.com

INAIL
Maria Cristina Paoletti
m.paoletti@inail.it