POLITECNICO
MILANO 1863

Reconciling Statistical and Causal Metrics of Fairness in Machine Learning on Data-Driven Systems

Chiara Criscuolo, Davide Martinenghi and Jing Huang

The 4th Italian Conference on Big Data and Data Science
9-11 September 2025, Turin, Italy

Introduction

Fairness, Statistical Fairness Metrics, Causal Fairness Metrics

“Fairness is the absence of prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics.” [1]

Focus: **distinction and integration between statistical and causal fairness metrics**

[1] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.

Why fairness?

Racial Bias in Mortgage Lending Chatbots (2024)[1]

Researchers at Lehigh University found that AI chatbots used in mortgage applications exhibited racial bias. In a study analyzing 6,000 simulated loan applications, the chatbots recommended denials for Black applicants more frequently than for identical white counterparts.

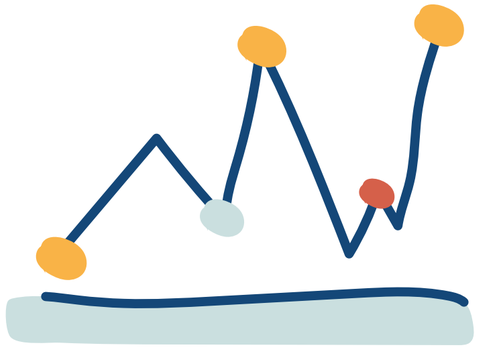
Bias in UK Welfare Fraud Detection AI (2024)[2]

An internal analysis revealed that the UK government's AI system for detecting welfare fraud exhibited bias against individuals based on age, disability, marital status, and nationality.

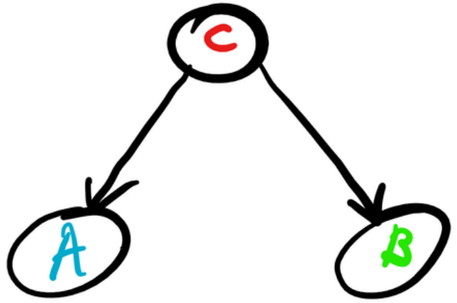
[1] https://news.lehigh.edu/ai-exhibits-racial-bias-in-mortgage-underwriting-decisions
[2] https://www.theguardian.com/society/2024/dec/06/revealed-bias-found-in-ai-system-used-to-detect-uk-benefits

Terminology

- **Statistical-based Fairness Metrics**: assess fairness based on the observed statistical outcomes of different demographic groups, which focus on ensuring similar outcomes across groups.



- **Causal-based Fairness Metrics**: consider the causal relationships between sensitive attributes and outcomes, ensuring fairness by intervening/simulating different scenarios to understand biases caused by underlying causal mechanisms.



Research Questions

Q1 What are the main **concepts** and **differences** between statistical-based and causal-based Fairness?

Q2 Which **datasets** are most commonly used in Fairness research, and are there **differences** between those used in causal-based and statistical-based studies?

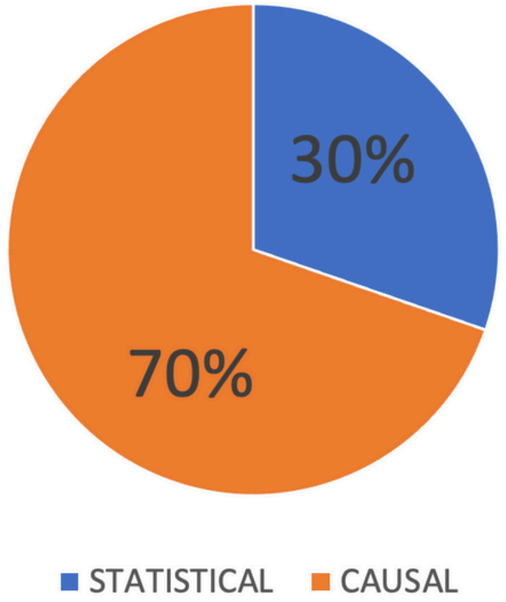
Q3 Is it possible to have a **common vision** between causal and statistical Fairness?

Q4 How to choose the **most suitable metric** considering both perspectives?

Results

We identified and selected **29 relevant papers** for detailed analysis:

- **30%** of the papers focus on **statistical-based fairness metrics**.
- **70%** of the papers concentrate on **causal-based fairness metrics**.



Research Results

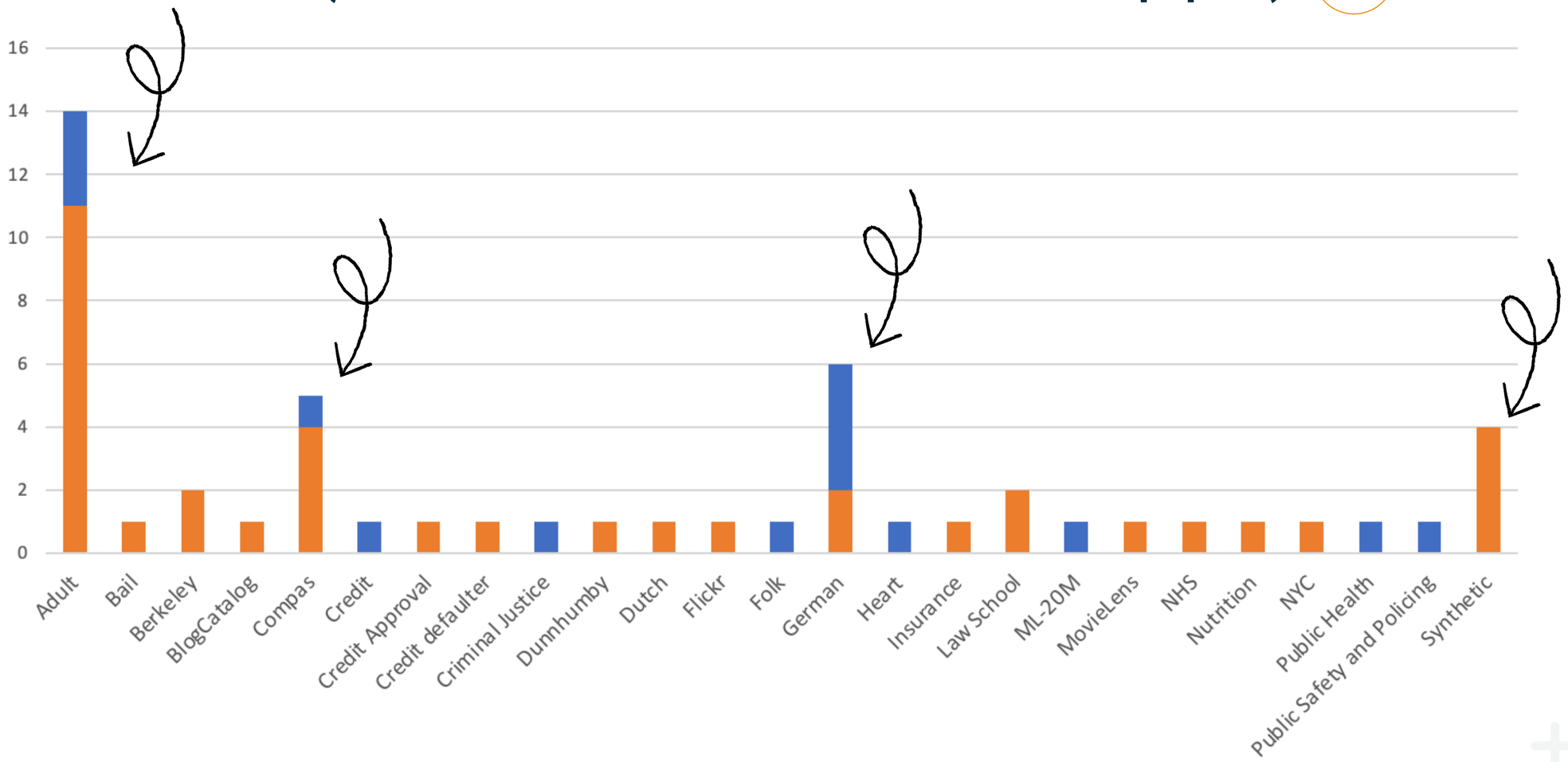
Research Results (a little part of selected papers)

Q1 Q2

PAPER	CAUSAL/NO CAUS/	FAIRNESS and METRICS	ANALYSIS TYPE (	CONTENT	VENUE	DATE	CONTEXT	EXPERIMENTAL DATASETS	METHODS (pre / in / post process) (/ se non	AUTHORS
Aequitas: A Bias and Fairness Audit Toolkit	NO CAUSAL	Equal Selection Parity, Demographic Parity, FP Parity, FDR Parity, FP/GS Parity, FPR Parity, FN Parity, FOR Parity, FN/GS Parity,	Solution proposal: Tool	Aequitas toolkit: audit machine learning models for discrimination and bias	arXiv:1811.05577	2018	machine learning	Criminal Justice, Public Health, Public Safety and Policing	In-process	Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby
Avoiding Discrimination through Causal Reasoning	CAUSAL	Demographic parity, equal opportunity, equal odds, individual fairness, counterfactual fairness	Solution proposal: new perspective	assay a general causal framework for reasoning about discrimination	NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems	2017	machine learning	/	/	Niki Kilbertus, Mateo Rojas-Carulla, Giambattista
Automated Data Cleaning Can Hurt Fairness in Machine Learning-based Decision Making	NO CAUSAL	Predictive parity, Equal opportunity	Evaluation: data cleaning impact	automated data cleaning impact	2023 IEEE 39th International Conference on Data Engineering (ICDE), Anaheim, CA, USA, pp.	2023	machine learning	Adult, Folk, Credit, Heart	/	Shubha Guha, Falaah Arif Khan, Julia Stoyanovich, Sebastian Schelter
BeFair: Addressing Fairness in the Banking Sector	BOTH	group fairness, individual fairness, causal fairness	Solution proposal: Tool	general roadmap for fairness in ML and the implementation of a toolkit called BeFair: a collection of tools	2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, pp. 3652-3661	2020	machine learning	/	Pre-Process, In-Process, Post-Process	Alessandro Castelnovo, Riccardo Crupi, Giulia Del Gamba, Greta
CAPUCHIN: CAUSAL DATABASE REPAIR FOR ALGORITHMIC FAIRNESS	CAUSAL	Causal fairness (Interventional Fairness)	Solution proposal: New algorithm	Capuchin system: repairing the training data	SIGMOD '19: Proceedings of the 2019 International Conference on Management of Data	2019	machine learning	Adult, Compas	Pre-Process, Post-Process	Babak Salimi, Luke Rodriguez, Bill Howe, Dan Suciu
Causal Modeling-Based Discrimination Discovery and Removal: Criteria, Bounds, and Algorithms	CAUSAL	direct/indirect discrimination	Solution proposal: New algorithm	discovering both direct/indirect discrimination from historical data, and removing them before	IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 11, pp. 2035-2050	2019	databases	Adult, Dutch	Pre-Process	Lu Zhang, Yongkai Wu, Xintao Wu
Causal Reasoning for Algorithmic Fairness	CAUSAL	Equalised Odds, Test fairness, Demographic Parity, Individual Fairness, Causal Fairness	Solution proposal: new perspective	argue that it is important to understand where sources of bias come from in order to rectify them,	arXiv:1805.05859	2018	databases	Berkeley admissions	/	Joshua R. Loftus, Chris Russell, Matt J. Kusner, Ricardo Silva
Counterfactual Fairness	CAUSAL	Causal fairness (Counterfactual Fairness)	Definition	classification and definition of counterfactual fairness	Advances in Neural Information Processing Systems	2017	databases	Law School	/	Matt J. Kusner, Joshua Loftus, Chris Russell, Ricardo Silva
Controlling Fairness and Bias in Dynamic Learning-to-Rank	NO CAUSAL	Group Fairness, Individual Fairness	Solution proposal: New algorithm	Fairness in dynamic learning to rank	SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in	2020	Ranking	ML-20M	In-process	Marco Morik, Ashudeep Singh, Jessica Hong, Thorsten Joachims
Database Repair Meets Algorithmic Fairness	BOTH	Statistical Fairness (Demographic Parity, Conditional Statistical parity, Equalized Odds, Predictive Parity), Causal Fairness	Classification: causal approach for fair ML	statistical fairness classification, causal fairness classification, Capuchin system	ACM SIGMOD Record, Volume 49, Issue 1, pp 34–41	2020	machine learning	Adult, Compas	Pre-Process, Post-Process	Babak Salimi, Bill Howe, Dan Suciu
DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks	CAUSAL	Causal fairness	Solution proposal: New algorithm	DEbiasing CAusal Fairness (DECAF), a generative adversarial network (GAN) that leverages causal	Advances in Neural Information Processing Systems, Volume 27, pp. 22221 - 22233, 35th Conference on	2021	machine learning	Adult, Credit Approval	Pre-Process	Boris van Breugel, Trent Kyono, Jeroen Berrevoets, Mihaela van
Explaining Algorithmic Fairness Through Fairness-Aware Causal Path Decomposition	CAUSAL	Causal fairness	Solution proposal: New algorithm	propose to a novel framework to explain algorithmic fairness with the causal graph	KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data	2021	Supervised learning	Synthetic, Adult, COMPAS, Nutrition	Pre-Process	Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, Fei Wang
Fair Inference on Outcomes	CAUSAL	Fair Inference	Definition	presence of discrimination = presence of a certain path-specific effect	AAAI'18/IAAI'18/EAAI'18: Proceedings of the Thirty-Second AAAI Conference on Artificial	2018	machine learning	Adult, Compas	Pre-Process	Razieh Nabi, Ilya Shpitser
Fairness Definitions Explained	BOTH	most prominent definitions of fairness (Statistical parity, Conditional statistical parity, Predictive parity, Predictive	Classification: collection of most prominent definitions	fairness classification	FairWare '18: Proceedings of the International Workshop on Software Fairness	2018	databases	German	/	Sahil Verma, Julia Rubin
Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality	CAUSAL	group fairness (fair on average causal effect, fair on average causal effect on the treated)	Definition	fair on average causal effect (FACE) and fair on aveage causal effect on the treated (FACT) definitions	WWW '19: The World Wide Web Conference	2019	machine learning	Synthetic, Adult, NYC Stop and Frisk	Pre-Process	Aria Khademi, Sanghack Lee, David Foley, Vasant Honavar

Research Result (number of datasets used across selected papers)

Q2



Fairness Decision Tree

Q3 Q4

A fairness decision tree is a tool that helps guide the selection of the fairness metrics

ORIGINAL DECISION TREE

statistical-based fairness metrics

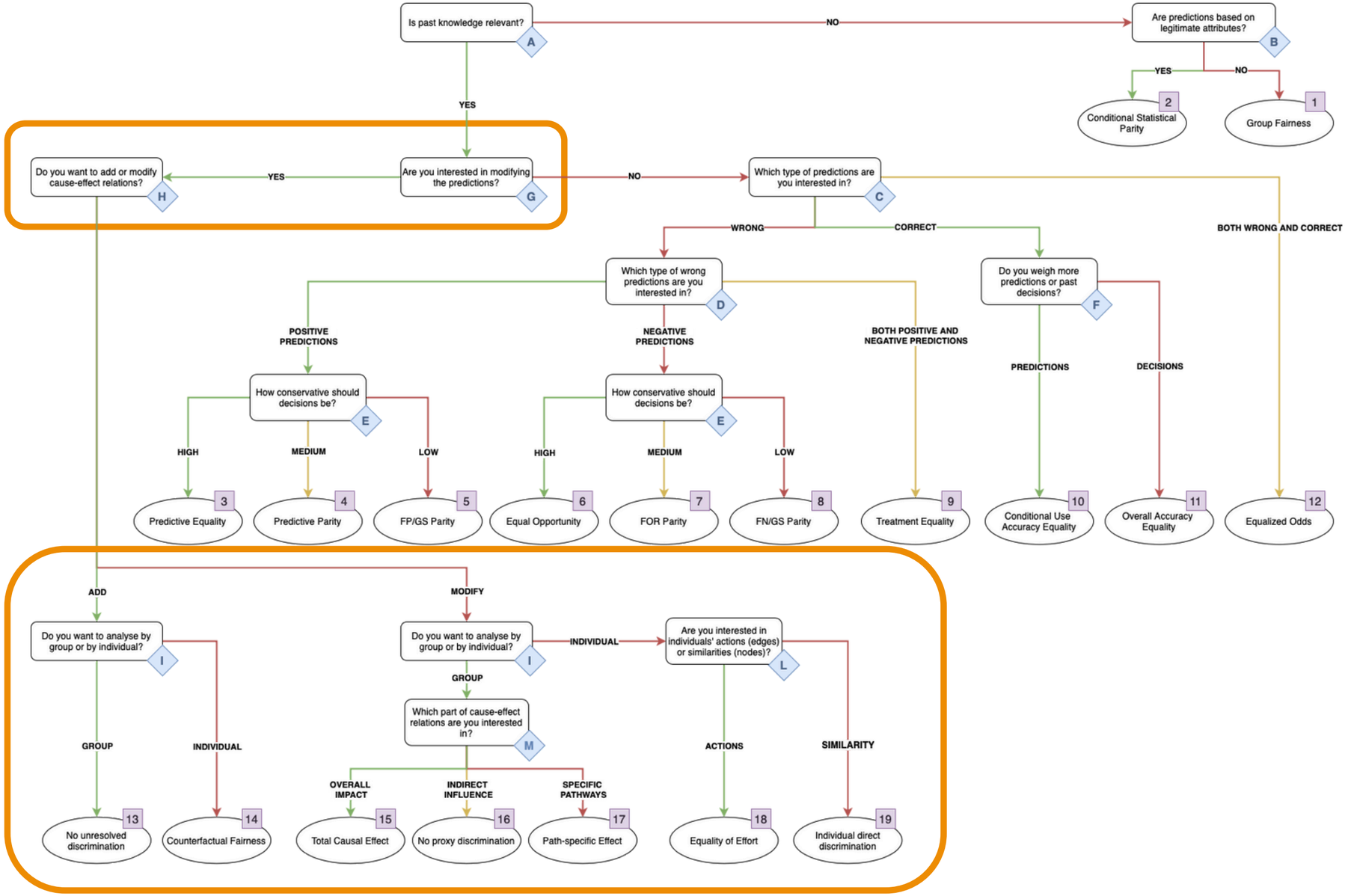
NEW DECISION TREE

statistical-based and **causal** fairness metrics

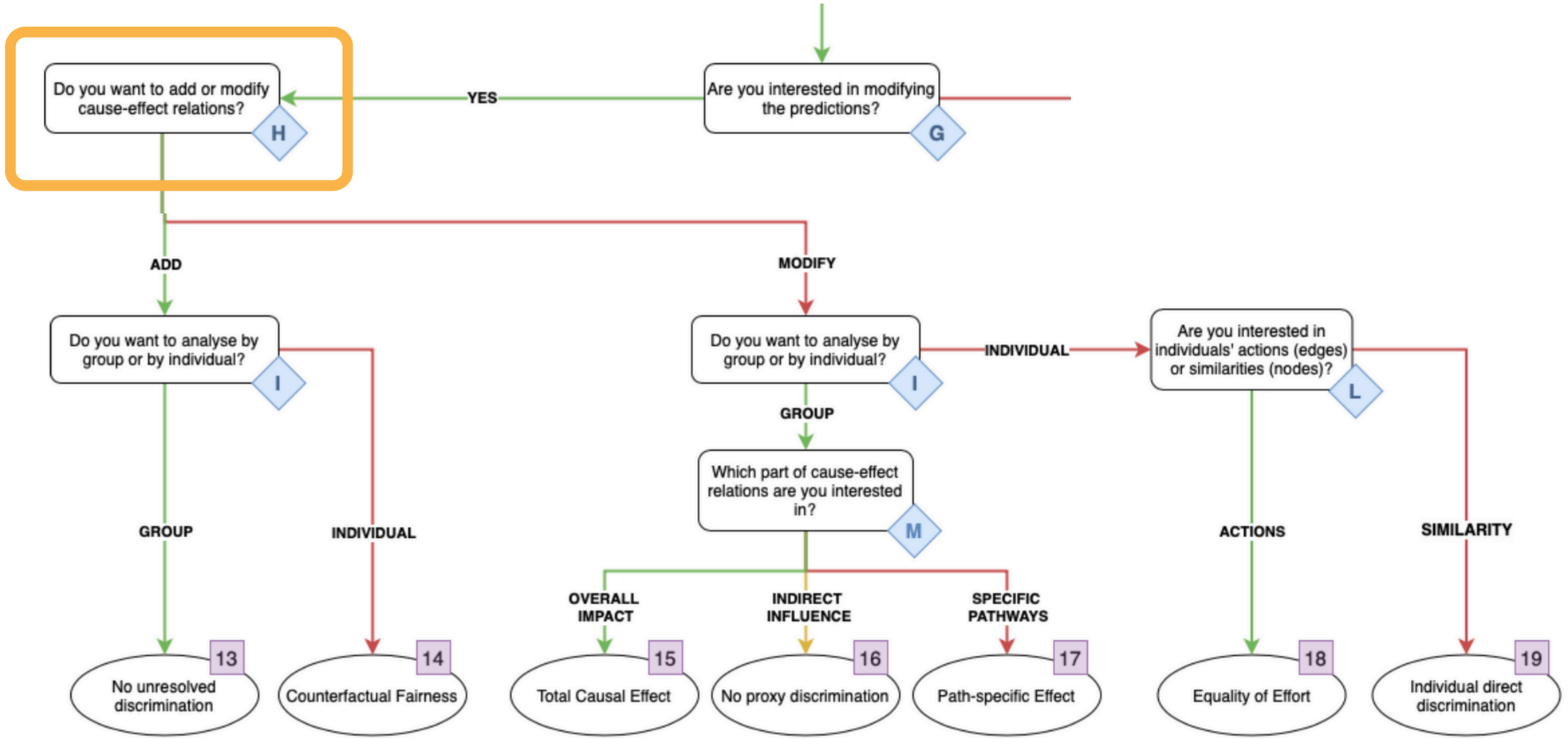
New Fairness Decision Tree

The new causal part of the decision tree includes the following nodes and metrics:

- questions **G, H, I, L, M**
- metrics **13 to 19**

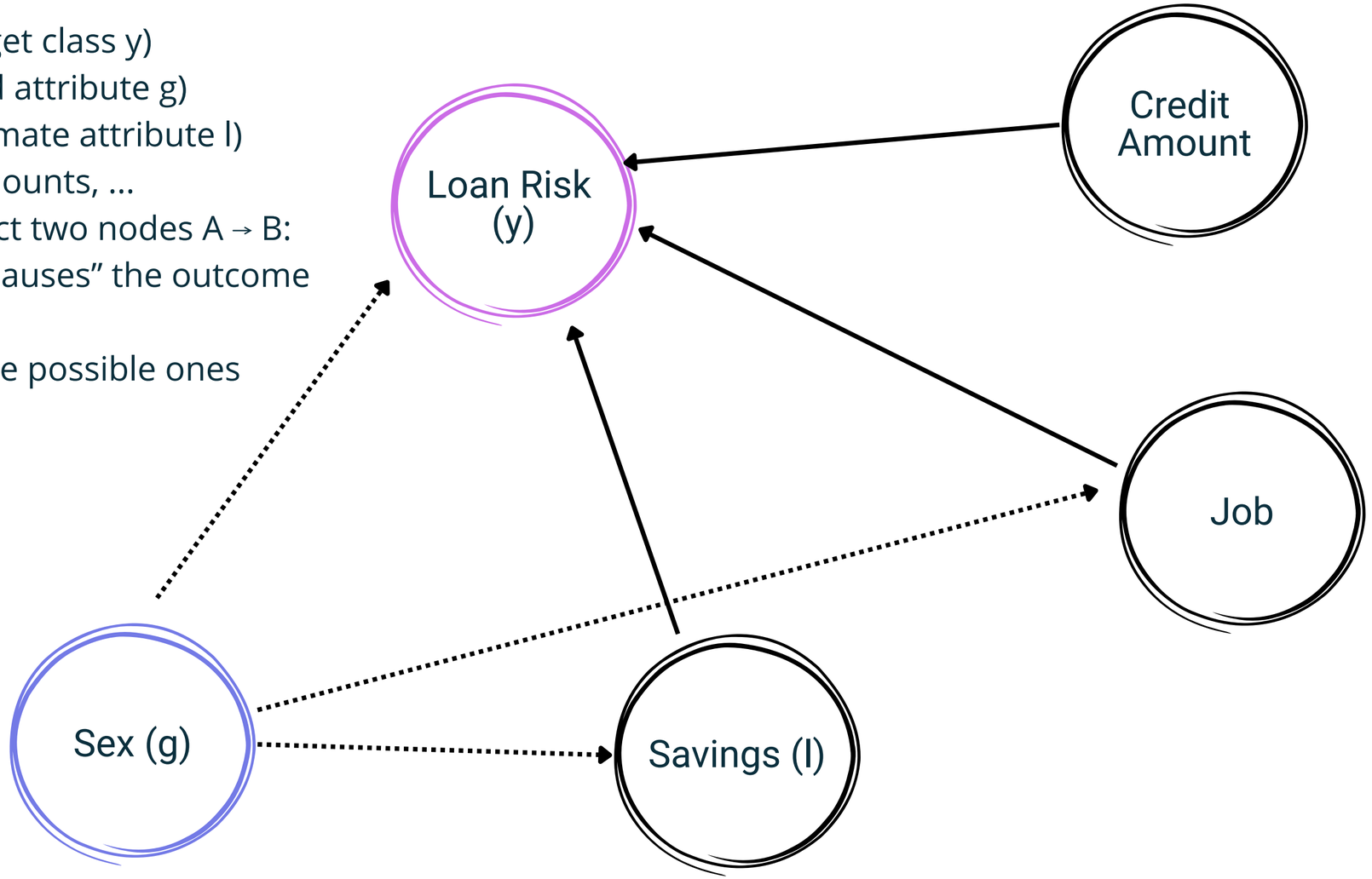


New Fairness Decision Tree



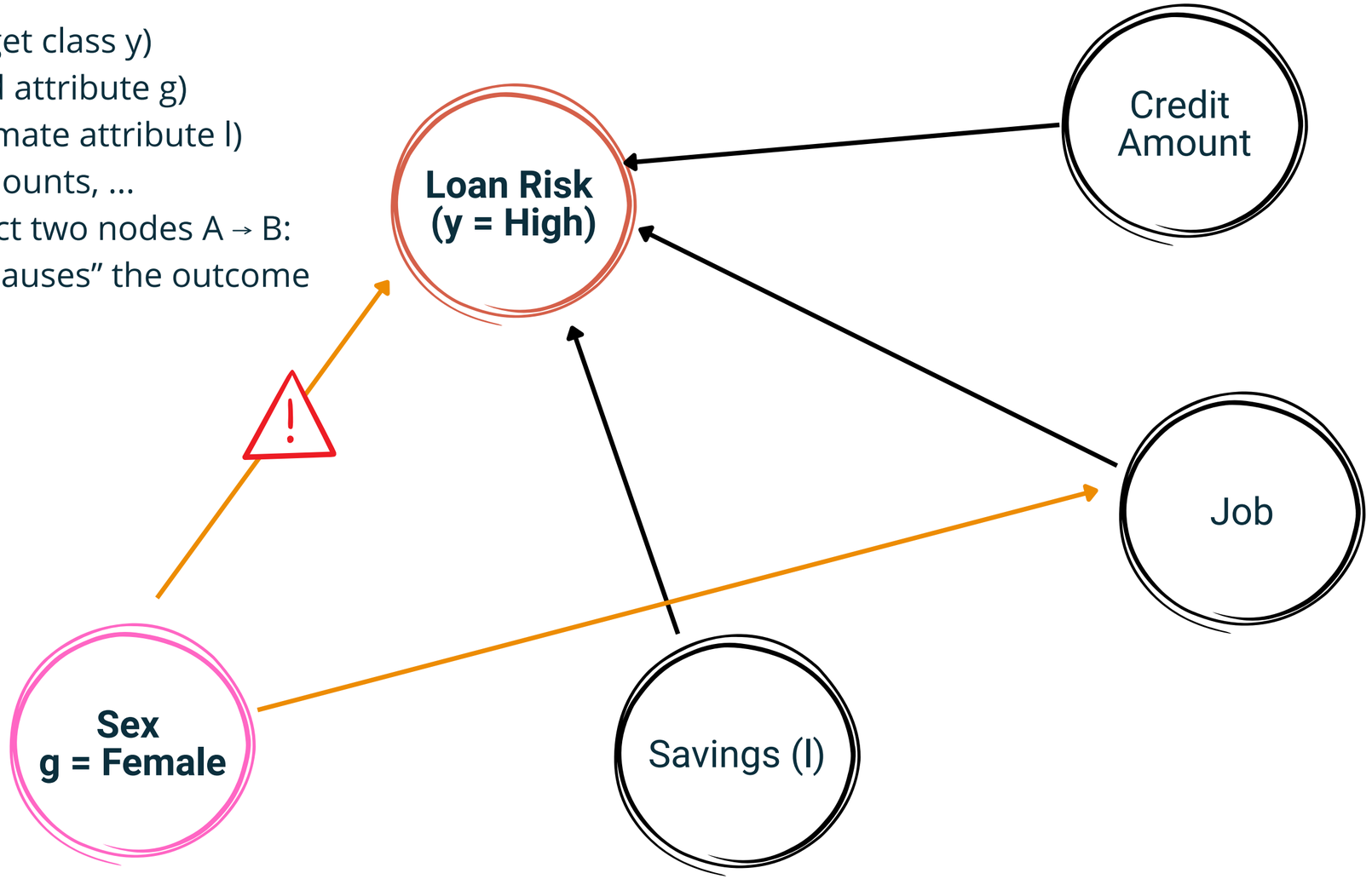
German Credit Example

- Nodes:
  - Loan risk (target class y)
  - Sex (protected attribute g)
  - Savings (legitimate attribute l)
  - Job, Credit Amounts, …
- Arrow that connect two nodes A → B: if the value of A “causes” the outcome of B
- Dashed arrows are possible ones

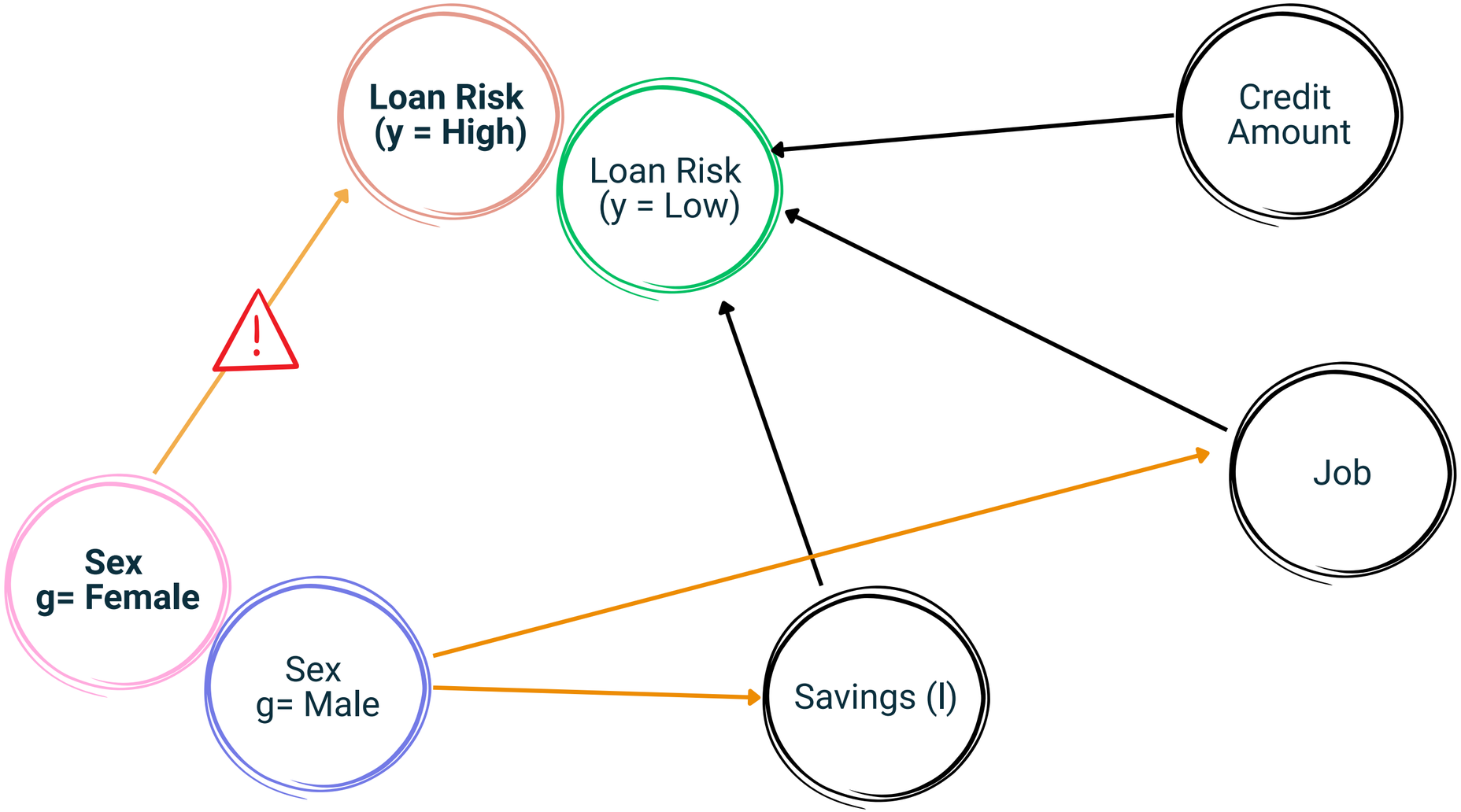


Interventional (Modify)

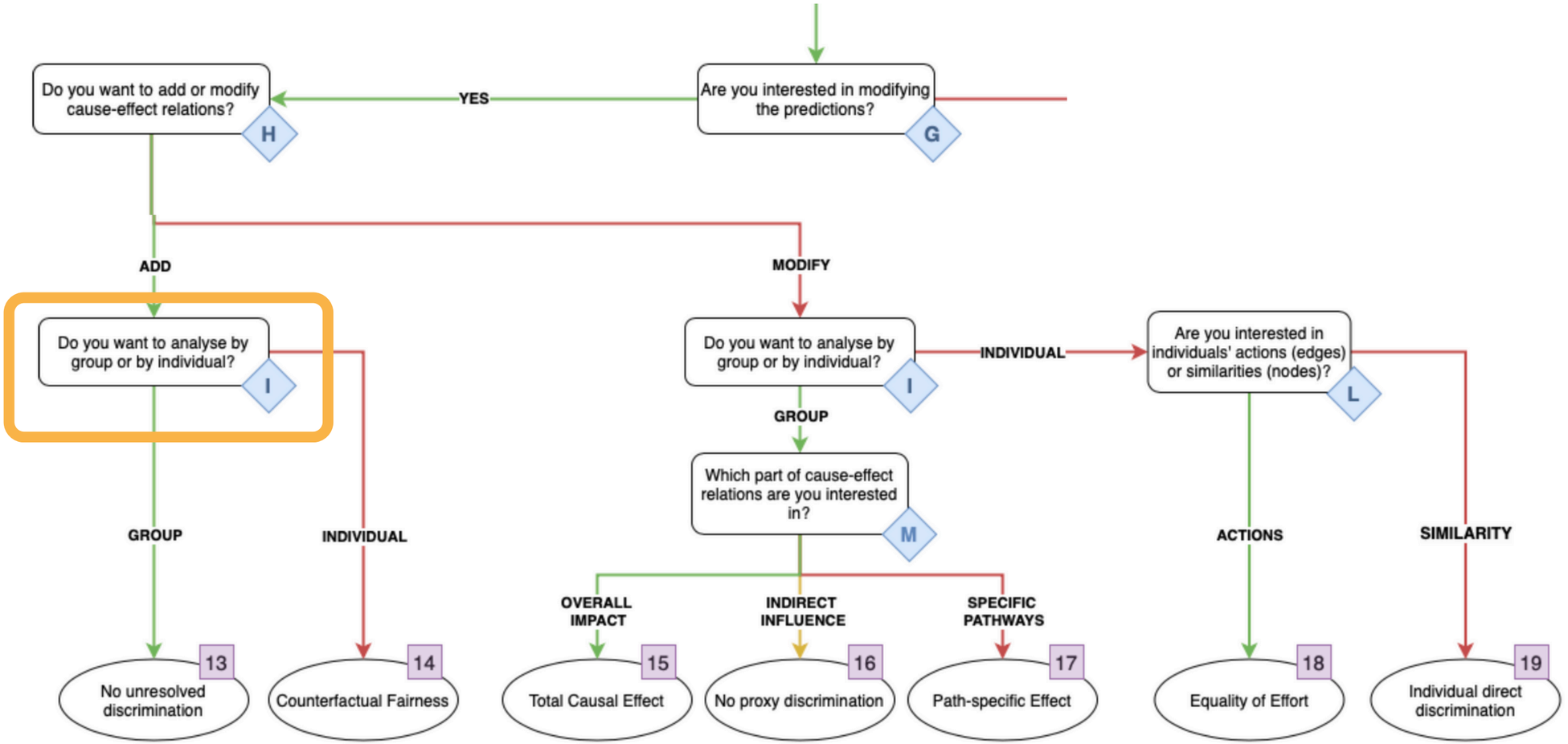
- Nodes:
  - Loan risk (target class y)
  - Sex (protected attribute g)
  - Savings (legitimate attribute l)
  - Job, Credit Amounts, …
- Arrow that connect two nodes A → B: if the value of A “causes” the outcome of B



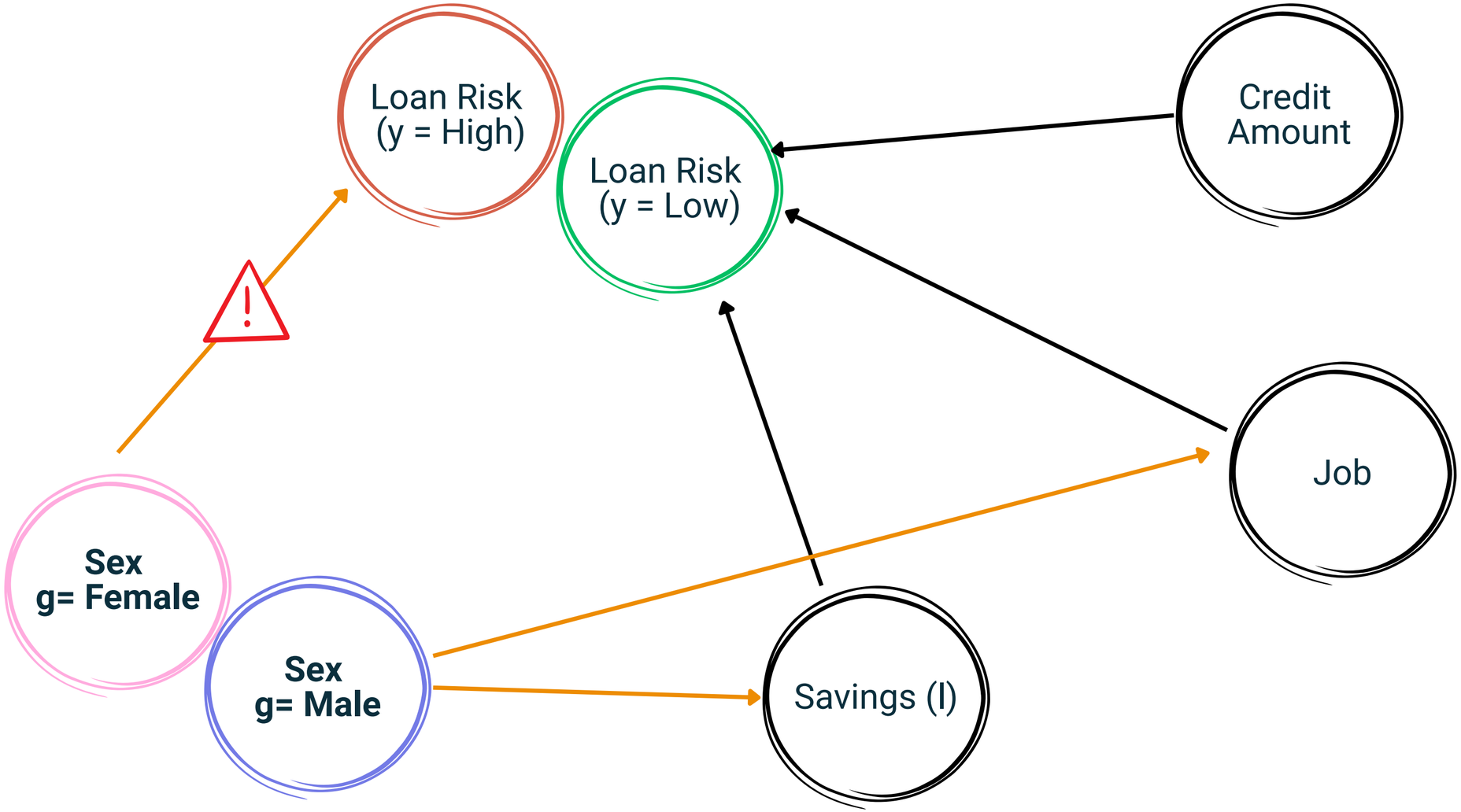
Counterfactual (Add)



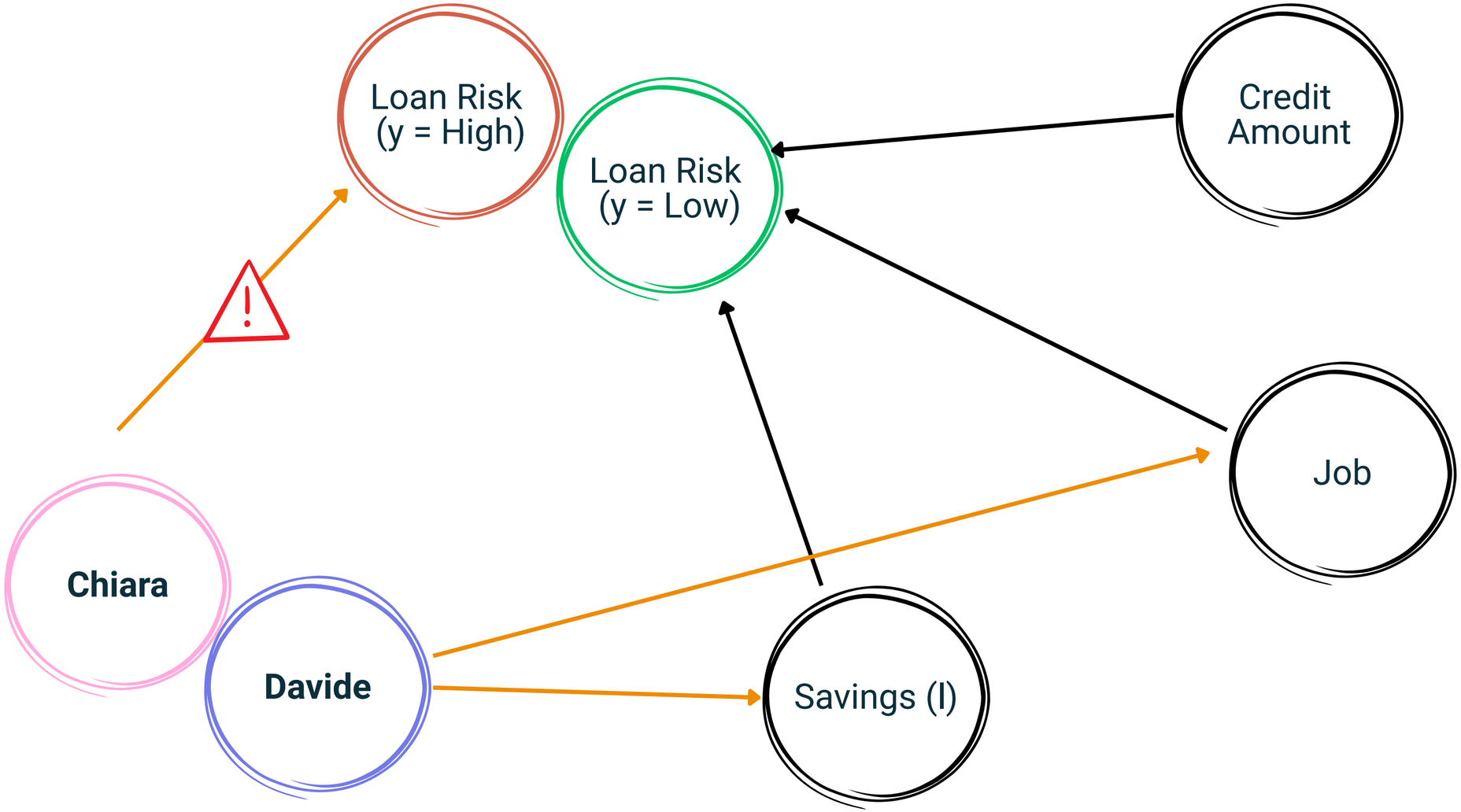
New Fairness Decision Tree



No Unresolved Discrimination (Group)



Counterfactual Fairness (Individual)



Conclusion

Contributions

- **Results of Literature Analysis**: Insights on the distinction between statistical and causal fairness.
- **Decision Tree**: A novel structured process to select the most suitable fairness metrics based on the specific context of their application.

Future Work

- **Expanding the Decision Tree**: Extend the decision tree to include new emerging fairness metrics.
- **Validation Through Practical Applications**: Apply the decision tree to real-world scenarios.

List of analyzed papers

- Saleiro, Pedro, et al. "Aequitas: A bias and fairness audit toolkit." *arXiv preprint arXiv:1811.05577* (2018).
- Kilbertus, Niki, et al. "Avoiding discrimination through causal reasoning." *Advances in Neural Information Processing Systems* 30 (2017).
- Guha, Shweta, et al. "Automated data cleaning can hurt fairness in machine learning-based decision making." *IEEE Transactions on Knowledge and Data Engineering* (2024).
- Castelnovo, Alessandro, et al. "Befair: Addressing fairness in the banking sector." *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 3652-3661.
- Salimi, Babak, et al. "Capuchin: Causal database repair for algorithmic fairness." *arXiv preprint arXiv:1902.08283*(2019).
- Zhang, Lu, Yizhou Wu, and Xindong Wu. "Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms." *IEEE Transactions on Knowledge and Data Engineering* 31.11 (2018): 2035-2050.
- Loftus, J. R., et al. "Causal reasoning for algorithmic fairness." *arXiv preprint arXiv:1805.05859* (2018).
- Kusner, Matt J., et al. "Counterfactual fairness." *Advances in Neural Information Processing Systems* 30 (2017).
- Morik, Matthias, et al. "Controlling fairness and bias in dynamic learning-to-rank." *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020): 429-438.
- Salimi, Babak, Bill Howe, and Dan Suciu. "Database repair meets algorithmic fairness." *ACM SIGMOD Record* 49.1 (2020): 34-41.
- Van Breugel, Bart, et al. "Decaf: Generating fair synthetic data using causally-aware generative networks." *Advances in Neural Information Processing Systems* 34 (2021): 22221-22233.
- Pan, Wenyuan, et al. "Explaining algorithmic fairness through fairness-aware causal path decomposition." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021): 1287-1297.
- Nabi, Razieh, and Ilya Shpitser. "Fair inference on outcomes." *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (2018).
- Verma, Sahil, and Julia Rubin. "Fairness definitions explained." *Proceedings of the International Workshop on Software Fairness* (2018): 1-7.
- Khademi, Amirhossein, et al. "Fairness in algorithmic decision making: An excursion through the lens of causality." *The World Wide Web Conference* (2019): 2907-2914.
- Du, Minhao, et al. "Fairness in deep learning: A computational perspective." *IEEE Intelligent Systems* 36.4 (2020): 25-34.

List of analyzed papers

- Galhotra, Siddhant, Yuriy Brun, and Alexandra Meliou. "Fairness testing: Testing software for discrimination." *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering* (2017): 498-510.
- Huan, Wei, et al. "Fairness through equality of effort." *Companion Proceedings of the Web Conference 2020* (2020): 743-751.
- Ma, Junteng, et al. "Learning fair node representations with graph counterfactual fairness." *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (2022): 695-703.
- Guo, Ruitong, et al. "Learning individual causal effects from networked observational data." *Proceedings of the 13th International Conference on Web Search and Data Mining* (2020): 232-240.
- Binns, Reuben. "On the apparent conflict between individual and group fairness." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020)
- Wu, Yizhou, Lu Zhang, and Xindong Wu. "PC-Fairness: A unified framework for measuring causality-based fairness." *Advances in Neural Information Processing Systems* 32 (2019).
- Yan, Jingyu N., et al. "Silva: Interactively assessing machine learning fairness using causality." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020): 1-13.
- Carey, Andrew N., and Xindong Wu. "The causal fairness field guide: Perspectives from social and formal sciences." *Frontiers in Big Data* 5 (2022): 892837.
- Kilbertus, Niki, et al. "The sensitivity of counterfactual fairness to unmeasured confounding." *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (2020): 616-626.
- Kasirzadeh, Atefeh, and Arran Smart. "The use and misuse of counterfactuals in ethical machine learning." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021): 228-236.
- Li, Yuxuan, et al. "Towards personalized fairness based on causal notion." *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021): 1054-1063.
- Sato, Mai, et al. "Unbiased learning for the causal effect of recommendation." *Proceedings of the 14th ACM Conference on Recommender Systems* (2020): 378-387.
- Baresi, Luciano, Claudia Criscuolo, and Carlo Ghezzi. "Understanding fairness requirements for ML-based software." *2023 IEEE 31st International Requirements Engineering Conference (RE)*. IEEE, 2023: 341-346.

POLITECNICO
MILANO 1863

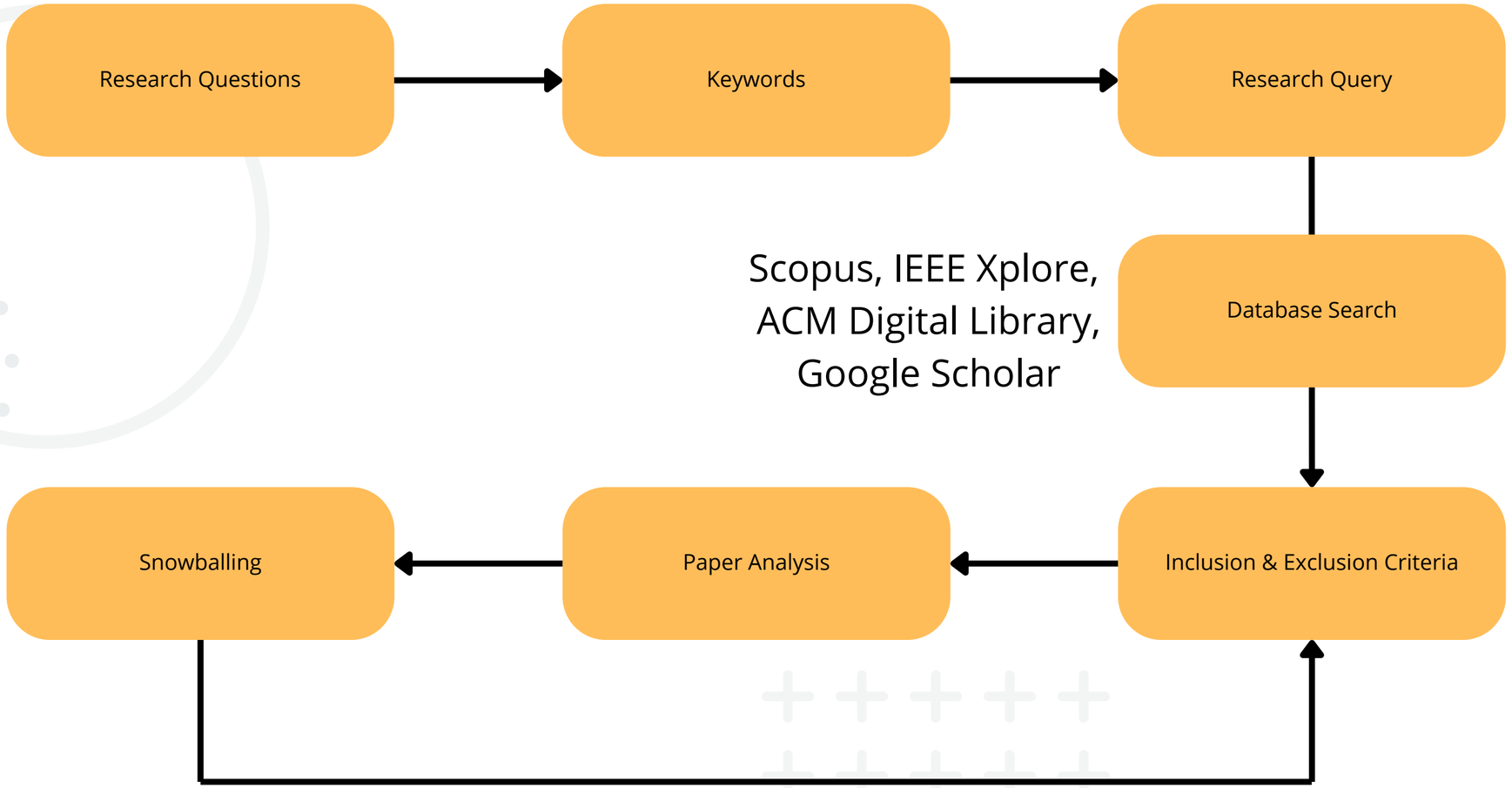
THANKS

FOR YOUR ATTENTION

ITA DATA 2025

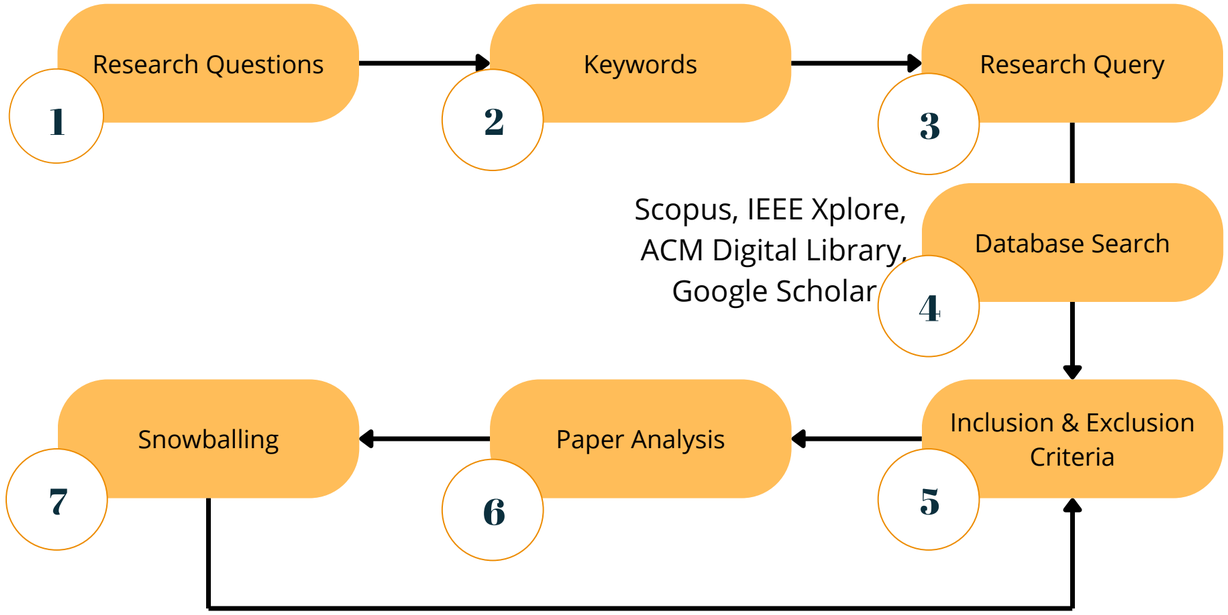
Chiara Criscuolo, Davide Martinenghi and Jing Huang

Methodology



Methodology

1. **Research Questions:** Develop specific research questions to guide the investigation.
2. **Keywords:** Develop a set of keywords that reflect the core topics of the research.
3. **Research Queries:** Formulate precise search queries using the identified keywords.
4. **Database Search:** Identify the most relevant search engines to collect research materials.
5. **Inclusion and Exclusion Criteria:** Define criteria for selecting relevant papers.
6. **Analyze Papers**: Collect all the information to address the research questions.
7. **Snowballing:** Use the reference of selected papers to discover additional studies.



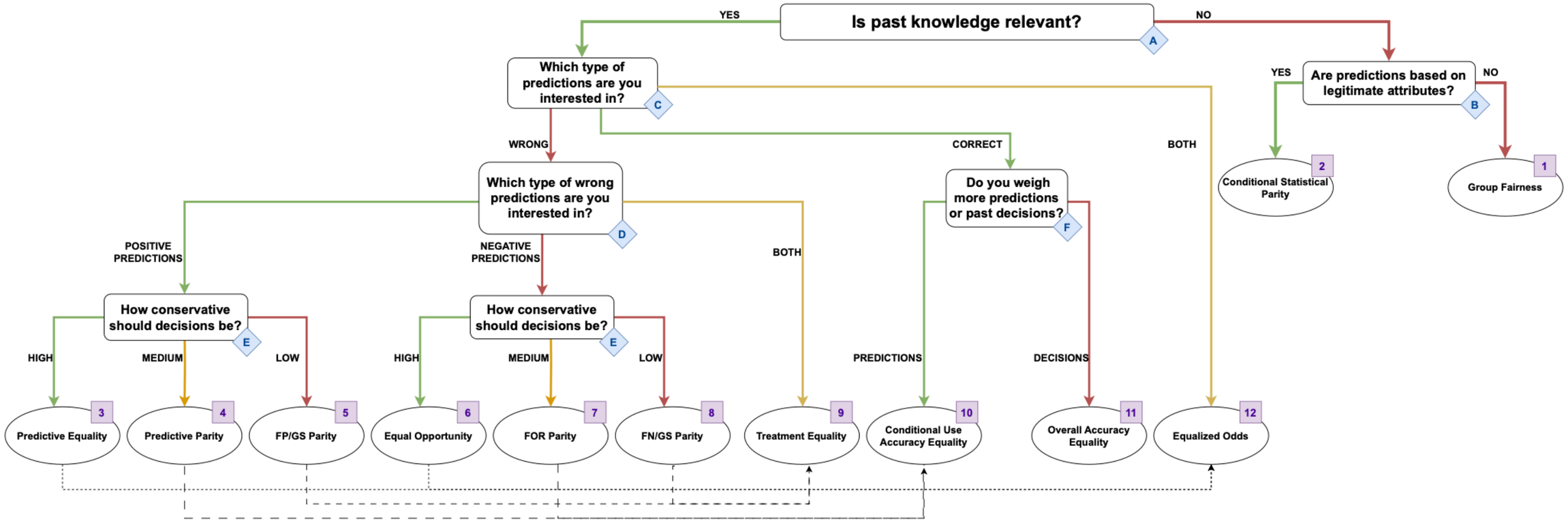
2,3 Research Strategy

Keywords	Machine Learning, ML, Data, Ethics, Metric, Definition, Solution, Mitigation, Measure, Causal, Statistical, Group, Individual, Counterfactual, Interventional
Query	(fair" OR discriminat" OR unfair" OR bias") **AND** (causal" OR statistic" OR individual" OR group" OR counterfact" OR intervention" OR parity") **AND** (metric" OR measur" OR defin" OR solut" OR mitigat") **AND** ((machine learning) OR data" OR ethic" OR (artificial intelligence))

5 Research Strategy

Inclusion Criteria	• Article or Papers with more than 5 citations; • A paper that addresses about Fairness; • A paper that proposes a mitigation; • A paper that proposes a tool; • A paper that proposes a new perspective; • A paper that proposes a new Fairness metric
Exclusion Criteria	• Theses or reports; • Surveys; • A paper that do not discusses Fairness; • A paper that is not written in English

Q4 Original Decision Tree



Causal Fairness Metrics

Individual Direct Discrimination	It aims to discover direct discrimination at the individual level by comparing the individual with similar individuals from both protected and unprotected groups.
Equality of Effort	It detects discrimination by comparing the effort required to reach the same level of the actual outcome for individuals from protected and unprotected groups.
Path-specific Effect	Given a causal path set, the path-specific effect is defined as the value change of the sensitive attribute on the decision along specific causal paths. It is considered fair if the difference is within the fair threshold.
No proxy Discrimination	Requires that no path exists from the protected attribute to the predicted outcome that is blocked by a proxy variable.
Counterfactual Fairness	Requires that the predicted outcome in the model does not depend on a descendant of the protected attribute.
Total Causal Effect	It is defined as the effect of changing the sensitive attribute on the decision along all causal paths. It is considered fair if the difference between the conditional distributions is within the fair threshold.
No unresolved discrimination	Requires that no path exists from the protected attribute to the predicted outcome.