

SoK: What Makes Private Learning Unfair?

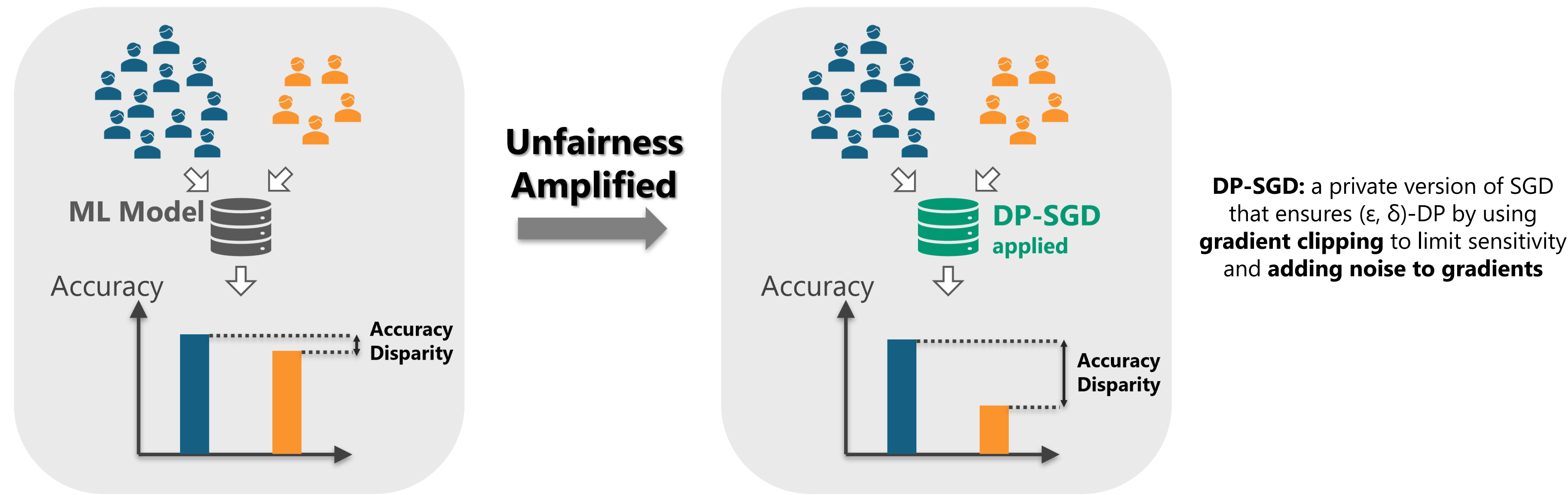
3rd IEEE Conference on Secure and Trustworthy Machine Learning, University of Copenhagen, Denmark April 9-11, 2025

Kai Yao (kai.yao@ed.ac.uk) and Marc Juarez (marc.juarez@ed.ac.uk)

Laboratory for Foundations of Computer Science, School of Informatics, University of Edinburgh

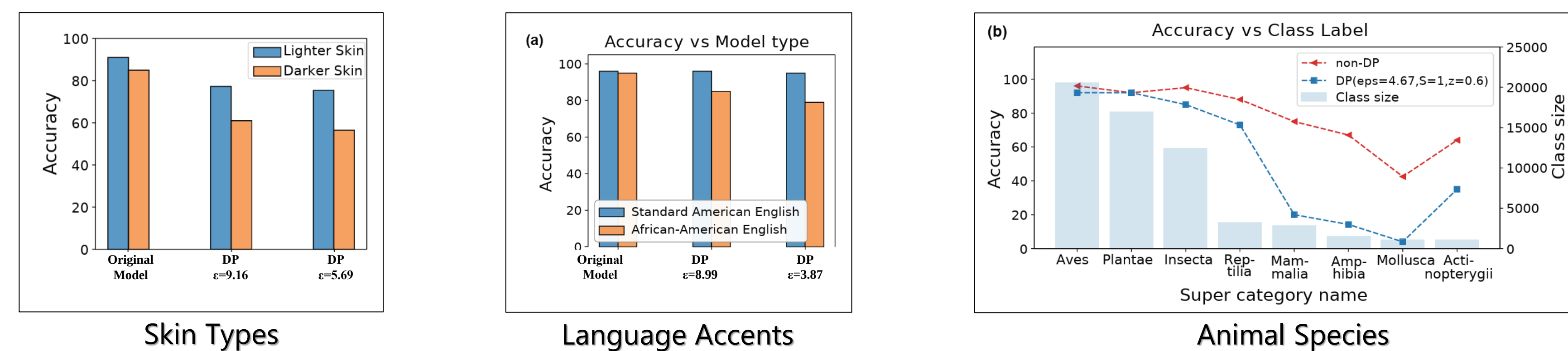
Problem Statement & Research Motivation

Differential Privacy (DP) can protect data privacy in ML, but can **amplify unfairness**

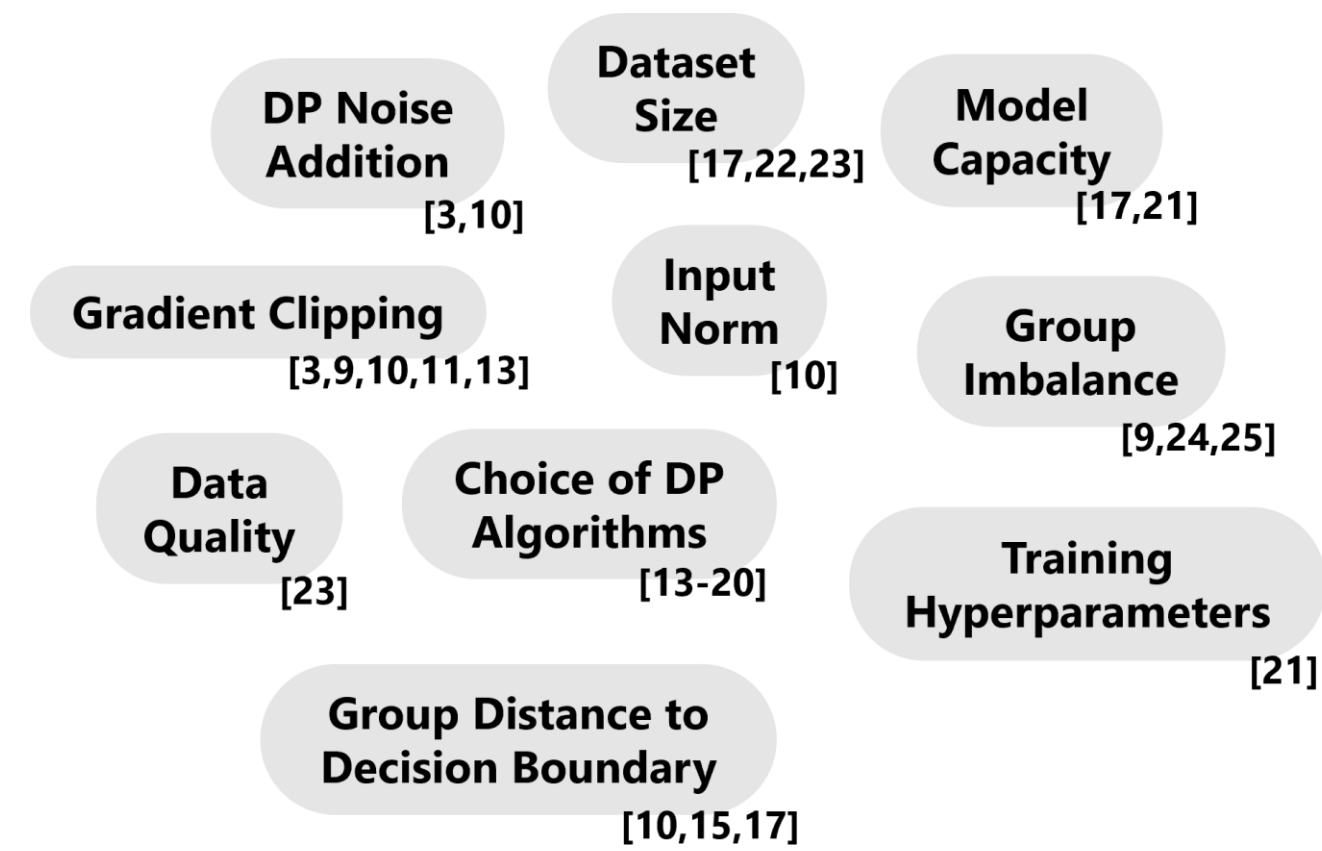


DP-SGD: a private version of SGD that ensures (ϵ, δ) -DP by using **gradient clipping** to limit sensitivity and **adding noise to gradients**

An early study found that **DP-SGD amplifies accuracy disparities** in age classification across different skin types. Moreover, this phenomenon seems **universal** across different kinds of groupings



Studies have identified several **Contributing Factors**, however ...

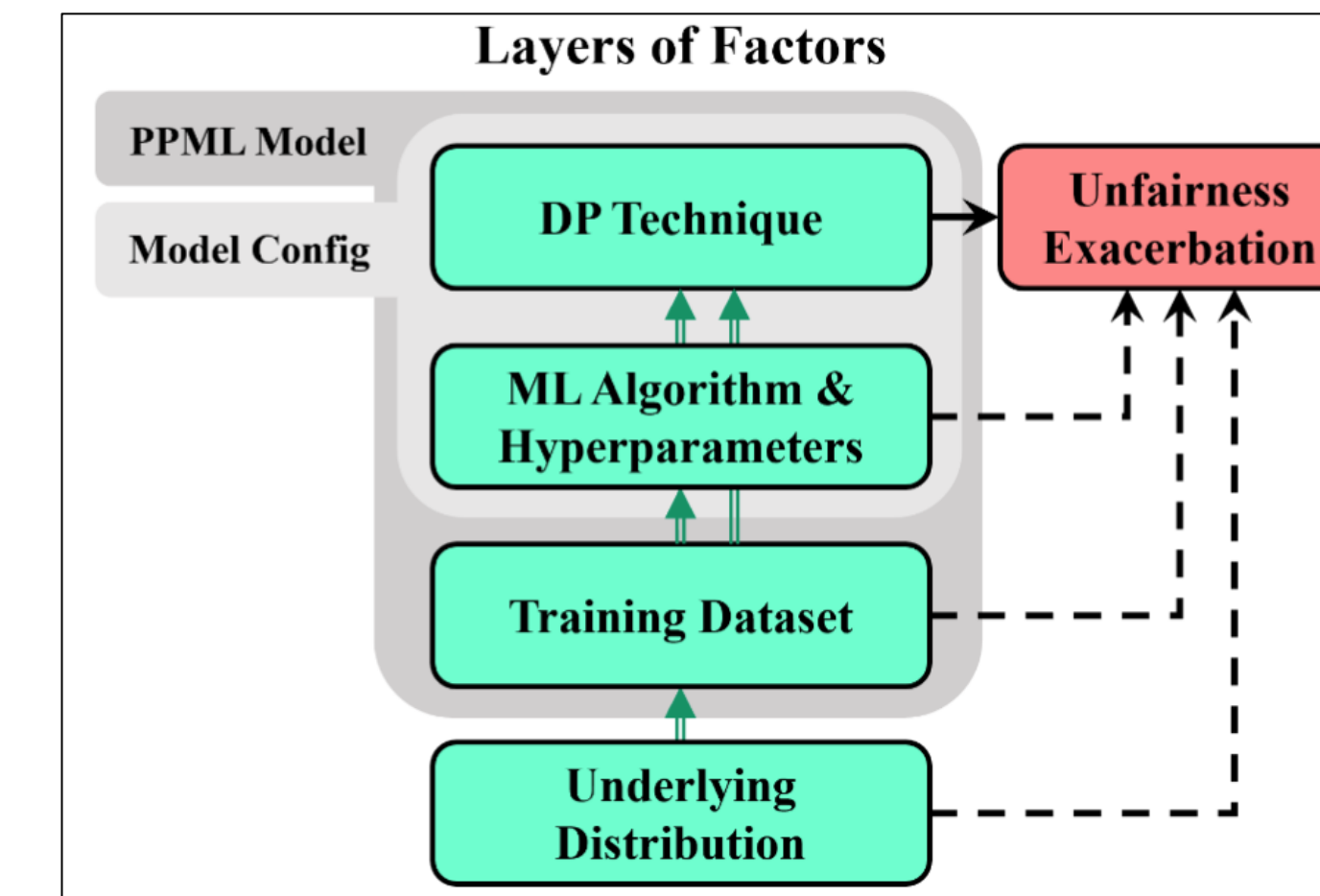


- Mechanism?
- Relationships?
- Significance?
- Necessity and sufficiency?

There is **lack of comprehensive review and analysis** of all the identified factors

Contributing Factors & Causal Analysis

Taxonomy of Contributing Factors

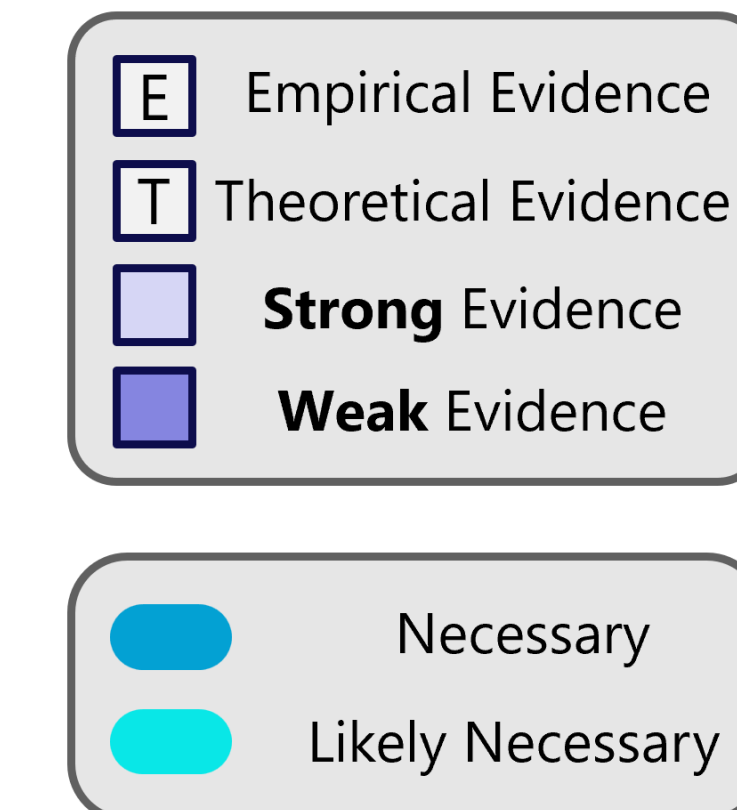


- Our taxonomy is guided by **where each factor is positioned within the ML pipeline**
- Each layer is **interconnected**, supporting **causal analysis** and revealing underexplored research areas
- Useful for both researchers to **study unfairness** and practitioners to **apply targeted mitigations**

Review and Analysis on Contributing Factors

DP Technique 12 works

- DP Noise Addition** [3,10]: Inherent for achieving DP
- Gradient Clipping** [3,9,10,11,13]: Larger gradient norm, more clipped
- Choice of DP Algorithms** [13-20]: Certain DP algorithms may mitigate unfairness



Training Dataset 4 works

- Dataset Size** [17,22,23]: Larger dataset, less noise needed
- Data Quality** [23]: Greater data quality, less utility and fairness loss
- Input Norm** [10]: Larger input norm, larger gradient norm

- No single factor is sufficient
- DP layer receives most attention
- Inter-factor relationship is unclear
- No conflicts between fairness notions

ML Algorithms 2 works

- Model Capacity** [17,21]: Larger model amplifies fairness
- Training Hyperparameters** [21]: Tuning can mitigate disparity

Underlying Distribution 6 works

- Group Imbalance** [9,24,25]: Extreme imbalance amplifies disparity; Unfair occurs even in balanced dataset
- Group Distance to Decision Boundary** [10,15,17]: Larger distance, less affected by noise

Methodology

Scope of literature

- Apply **DP** to **supervised ML**, during training or other stages in ML pipeline
- Study change in **performance disparity** across subgroups, **before and after DP** is applied

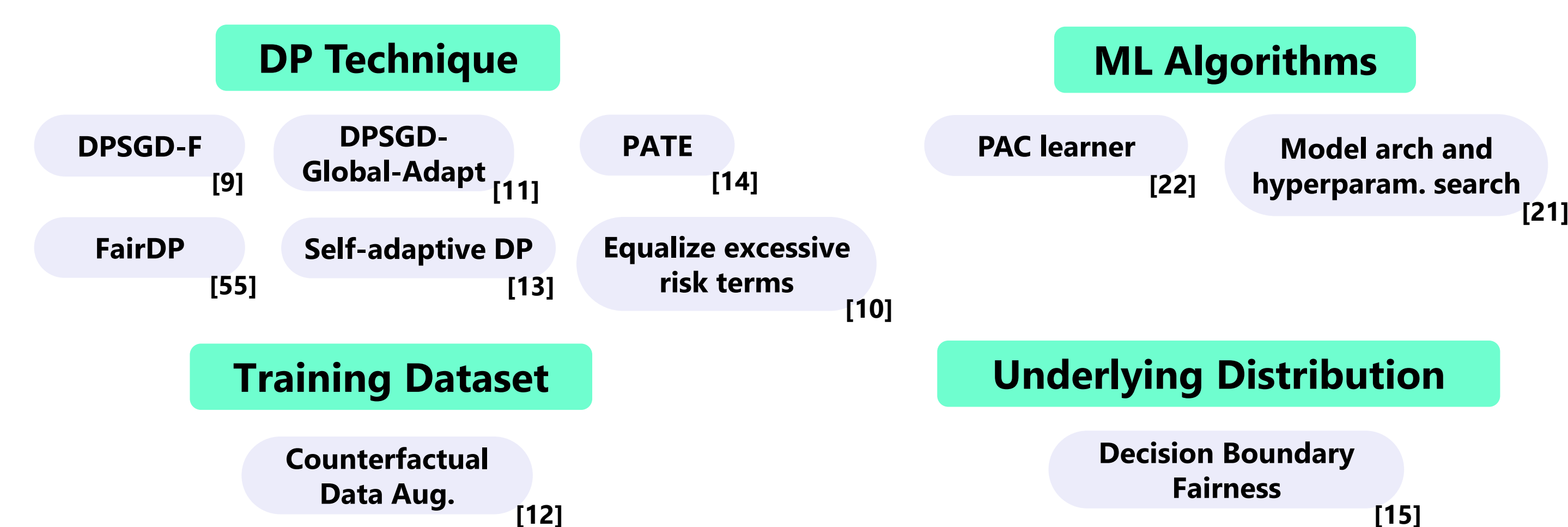
Methodology

- **Google Scholar** query populated **~200** articles, we selected **20** studies that are relevant

Literature Review

- Organize studies by key elements:
 - ✦ Fairness notions; DP techniques; ML algorithms; Tasks and data modalities
 - ✦ Identified contributing factors
- Compare all studies to analyze the significance and causal role of each factor

Mitigation Solutions



Mitigation efforts often focus on **DP techniques**, as lower layers are harder to adjust due to embedded **societal biases**

Challenges and Directions

Current Causal Analysis Faces Challenges

- ✉ New field with **limited studies** (~20); few explore **factor importance or interactions**
- ⚖️ Inconsistent **fairness definitions** without clear justification hinder comparison

Dataset & Distribution Need More Attention

- ✚ Focus is mostly on DP algorithms, **overlooking dataset/distribution** factors
- ❓ Open questions: What distributional properties matter most? How does decision boundary alter with DP?

Mitigation Methods Come with Limitations

- 🔒 **Revealing group labels** may breach data privacy regulations
- 🔄 Alternatives to DP-SGD can weaken both privacy and attack robustness



THE UNIVERSITY
of EDINBURGH

SECURITY, PRIVACY
& TRUST GROUP



Academic Centre of Excellence
In Cyber Security Research



in association with
National Cyber
Security Centre