

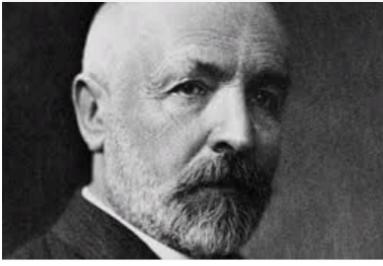
# Bias in, Bias out? Building Fair Models from Imbalanced Data

# Example: Hiring Algorithm

Google mathematician

Alle Bilder Maps News Videos Mehr Einstellungen Tools

william jones albert einstein indian atiyah math mathematics michael atiyah srinivasa ramanujan indian mathematicians

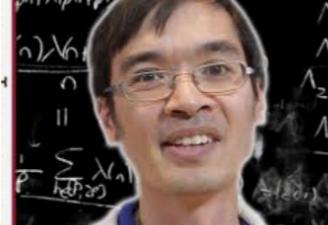
The 10 best mathematicians | Culture | The G...  
the guardian.com

John Wilson (mathema...  
en.wikipedia.org

John Casey (mathem...  
en.wikipedia.org

Biography of Joseph Louis Lagrange, ...  
thoughtco.com

That's Maths: Where are the successors to...  
irishtimes.com

Charles Babbage - Mathematician Bi...  
famousmathematicians.net

The 10 best mathematicians | Culture | ...  
theguardian.com

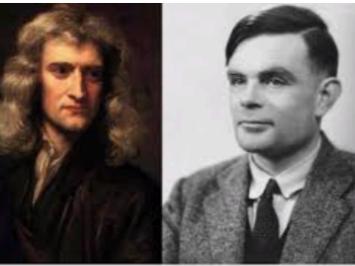
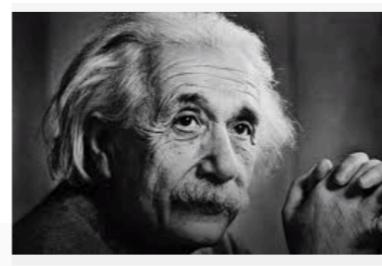
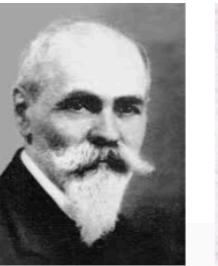
5 brilliant mathematici...  
mnn.com

The World's Best Mathematician (\*) - Nu...  
youtube.com

Mathematician – WeUseMath.org  
weusemath.org

Florence Nightingale as a Mathematici...  
worldhistory.us

Explainer: Has Michael Atiyah conque...  
irishtimes.com

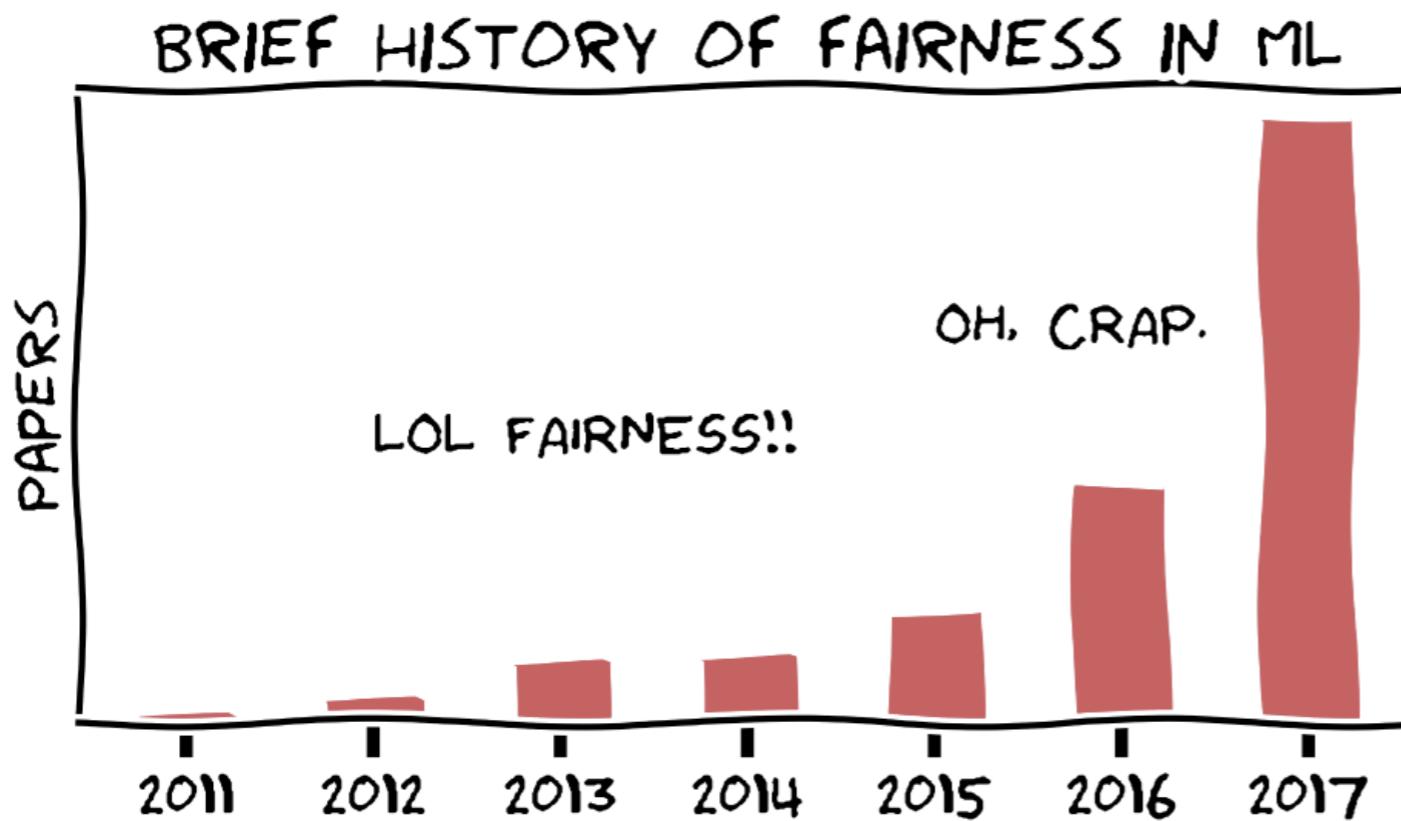
      

Gottfried Wilhelm Leibniz: Who wa...  
5 brilliant mathematicians and their impact o...  
The 10 best mathematicians | Culture | T...  
Albert Einstein - Mathematician Biography, ...  
Robert Osserman, not...  
Stanisław Zaremba (...  
Srinivasa Ramanujan | Famous Mathemati...

 Age: 50  
Gender: Male  
Birth place: Europe

[1] M. C. Tschantz and A. Datta. Automated experiments on ad privacy settings.  
In *Proceedings on Privacy Enhancing Technologies*, 2015.

# Fair ML Research



[2] CS 294: Fairness in Machine Learning  
at Berkeley, by Moritz Hardt  
(<https://fairmlclass.github.io>)

One Goal:  
Define and formalise "fair"  
→ Fairness metrics

# I) Individual Fairness

"Similar individuals should have similar outcomes"

A model M is **fair** if it satisfies the following:

**Definition** (Lipschitz mapping). A mapping  $M: V \rightarrow \Delta(A)$  satisfies the  $(D, d)$ -Lipschitz property if for every  $x, y \in V$ , we have

$$D(Mx, My) \leq d(x, y). \quad (1)$$

When  $D$  and  $d$  are clear from the context we will refer to this simply as the *Lipschitz* property.

V: set of individuals  
M: "model", maps individuals to outcomes  
d, D: metrics in input/output space

[3]Dwork, Cynthia, et al. "Fairness through awareness."

In *Proceedings of the 3rd innovations in theoretical computer science conference. ACM, 2012.*

# II) Group Fairness

"Different groups should have similar outcomes"

Groups defined via a **protected attribute A** (e.g. gender, age or race).

Feature vector now becomes  $X = (x_1, \dots, x_n, a)$

## A) Same distribution of outcomes per group ("statistical parity")

A model M is **fair** iff

$$P\{M(X) = 1 | A = a\} = P\{M(X) = 1 | A = b\}.$$

## B) Same error rates per group ("equalized odds")

A model M is **fair** iff

$$P\{M(X) = 1 | A = 0, Y=y\} = P\{M(X) = 1 | A = 1, Y=y\} \text{ for } y \in \{0,1\}.$$

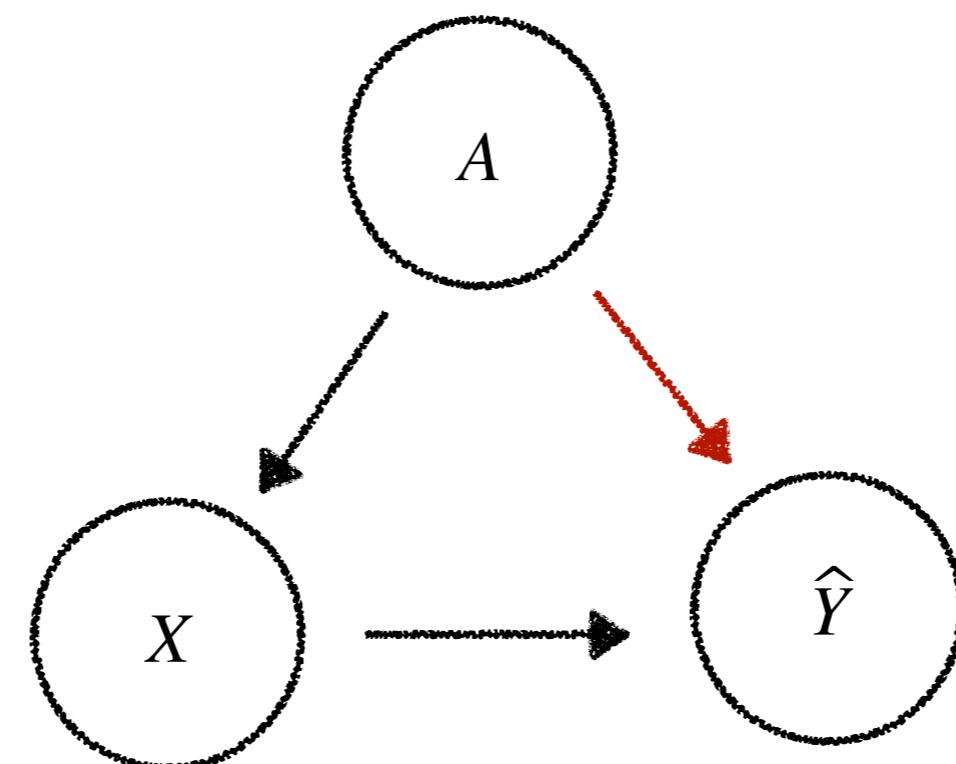
[4] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323).

[5] Barocas, S., Hardt, M. and Narayanan, A.. Fairness and Machine Learning , [www.fairmlbook.org](http://www.fairmlbook.org), 2019.

# III) Causal Fairness Criteria

A model is fair if it doesn't display any **unresolved discrimination**:

**Definition** (Unresolved discrimination). A variable  $V$  in a causal graph exhibits *unresolved discrimination* if there exists a directed path from  $A$  to  $V$  that is not blocked by a resolving variable and  $V$  itself is non-resolving.



no unique definition

contradicting  
definitions

# Some critical notes...

context

tools for analysis rather than  
solutions



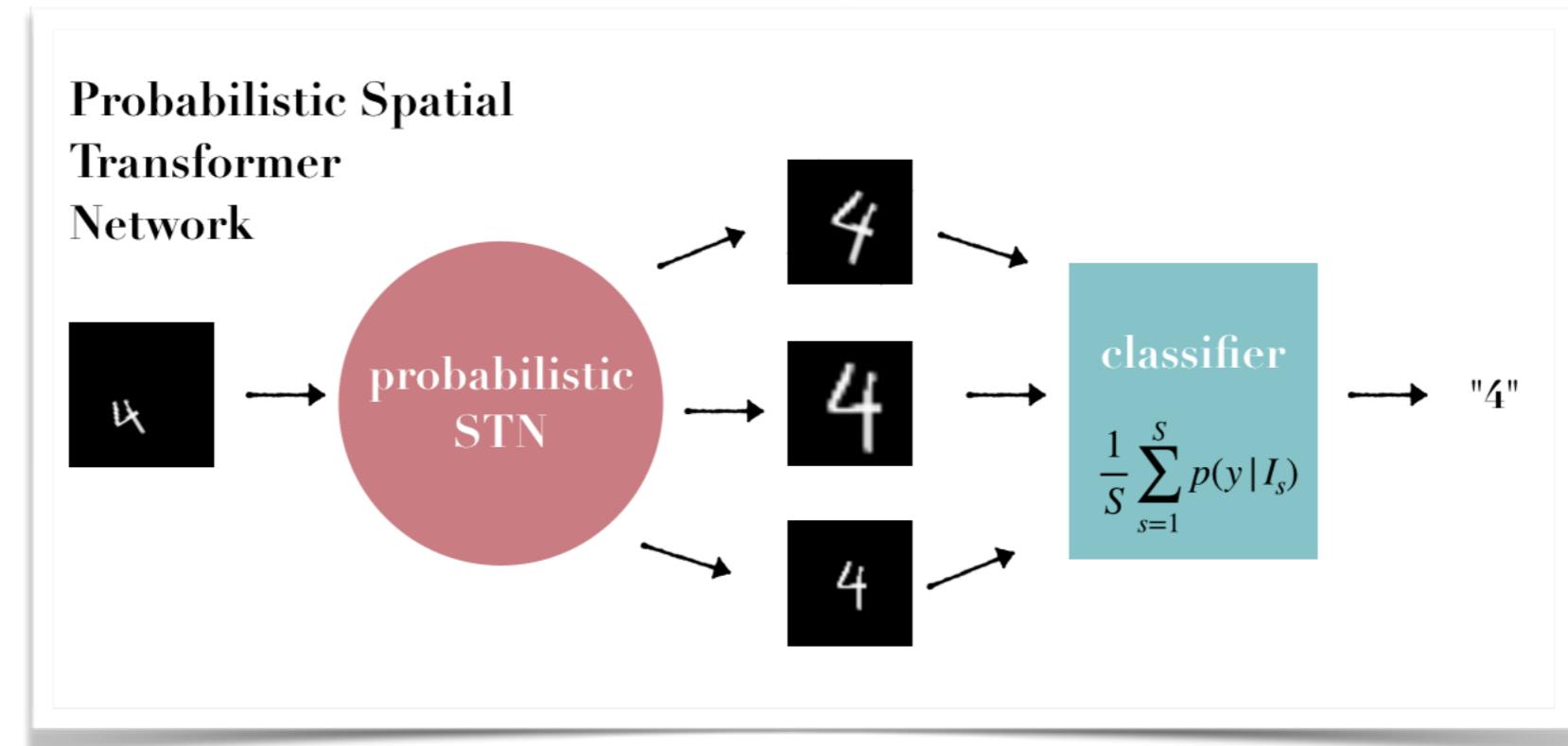
Goal: Operationalise ethics concepts and translate them into formulas and code, thereby making them accessible for the technical community to work with.

# My Research: Data Augmentation

Data augmentation: Artificially extend datasets that are too small.  
Usually done via ad hoc assumptions.



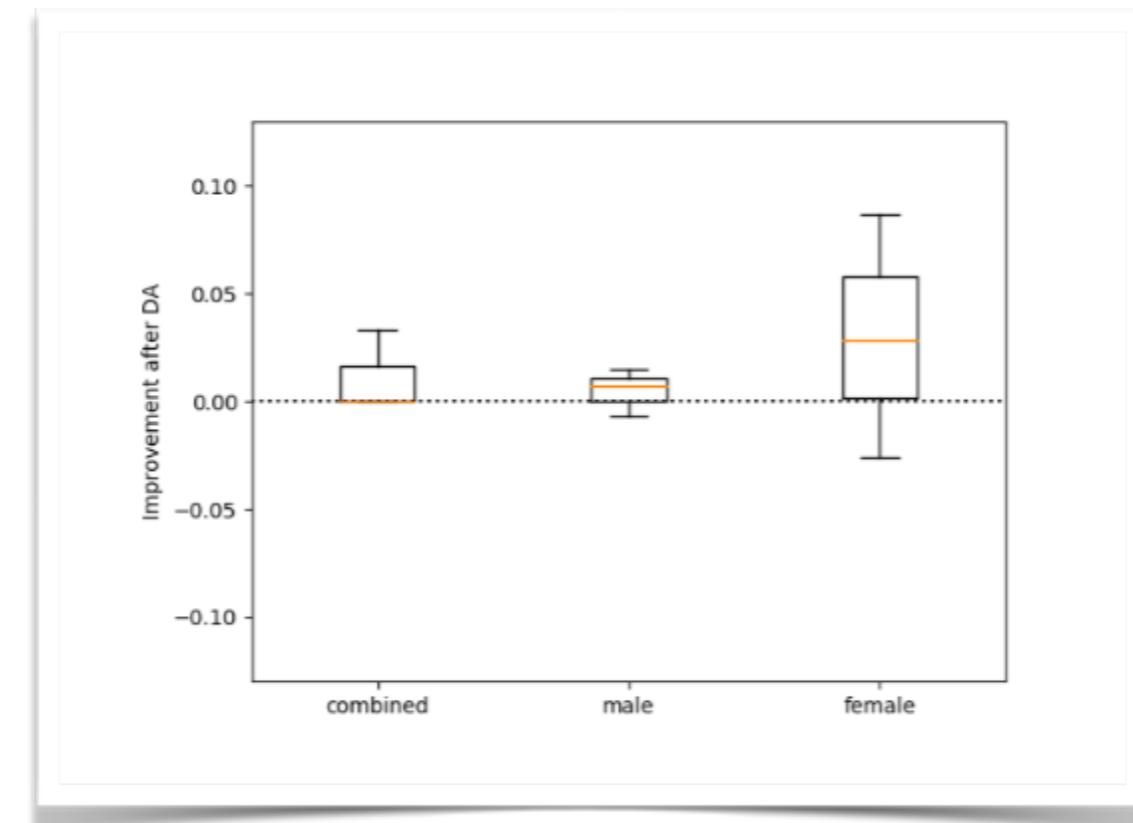
Our method: Estimate good data augmentation scheme from data.



# Data Augmentation for Bias-Correction

Now: Only augment underrepresented group.

- upsampling
- more balanced dataset
- bias-correction!



Some first results building on

[7] Piotr Sapiezynski, Valentin Kassarnig, and Christo Wilson.  
Academic performance prediction in a gender-imbalanced environment. 2017.

on data from

[8] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone,  
Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann.  
Measuring large- scale social networks with high resolution. PloS one, 2014.

# Thanks!

Sounds interesting?  
02456 project

[posc@dtu.dk](mailto:posc@dtu.dk)