Making algorithmic decision-making justifiable and contestable; some technical, legal and institutional possibilities

Reuben Binns
Department of Computer Science, University of Oxford

Supporting Algorithm Accountability using Provenance — Opportunities and Challenges. Provenance Week 2018 at King's College London, 12 July, 2018.

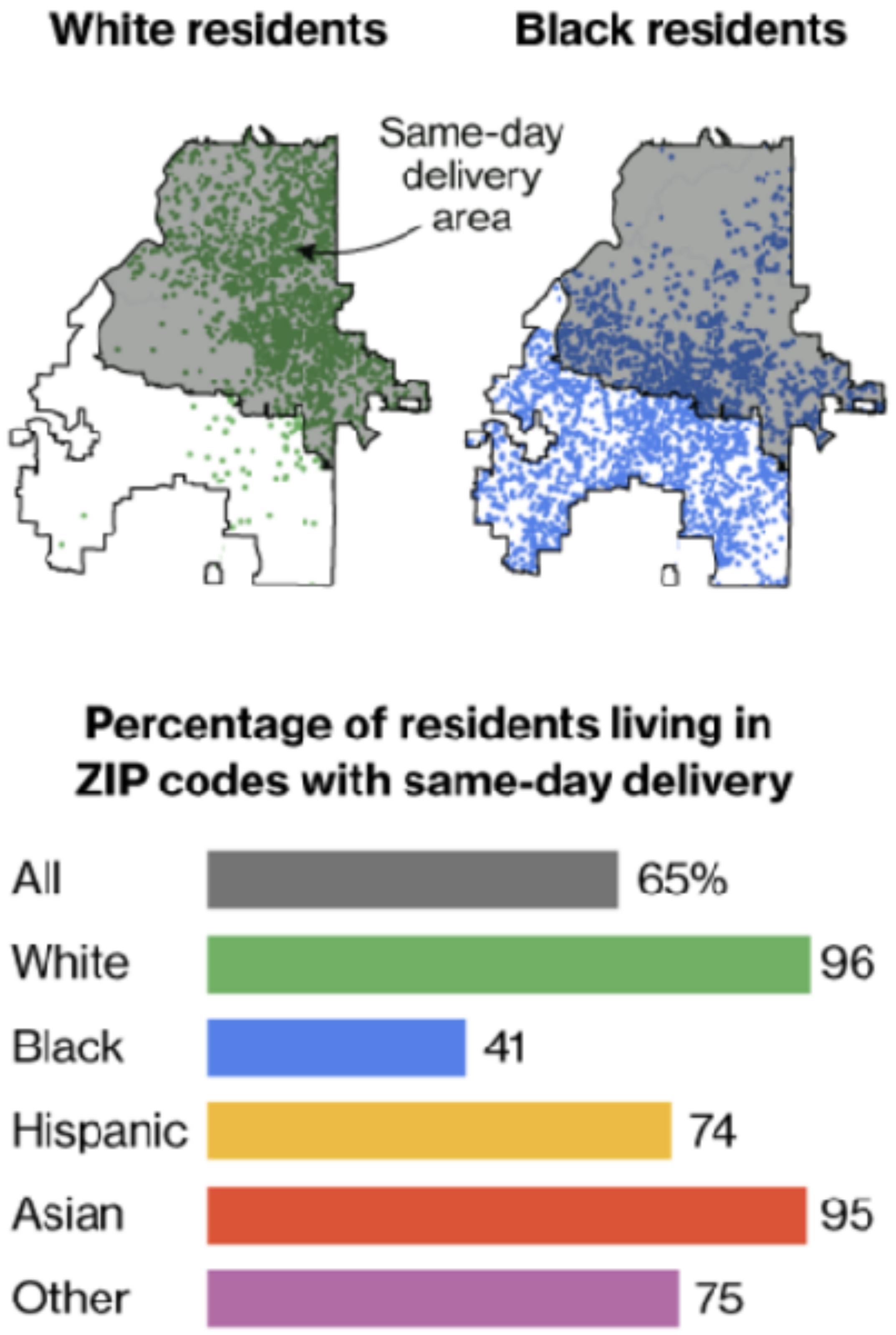
Outline

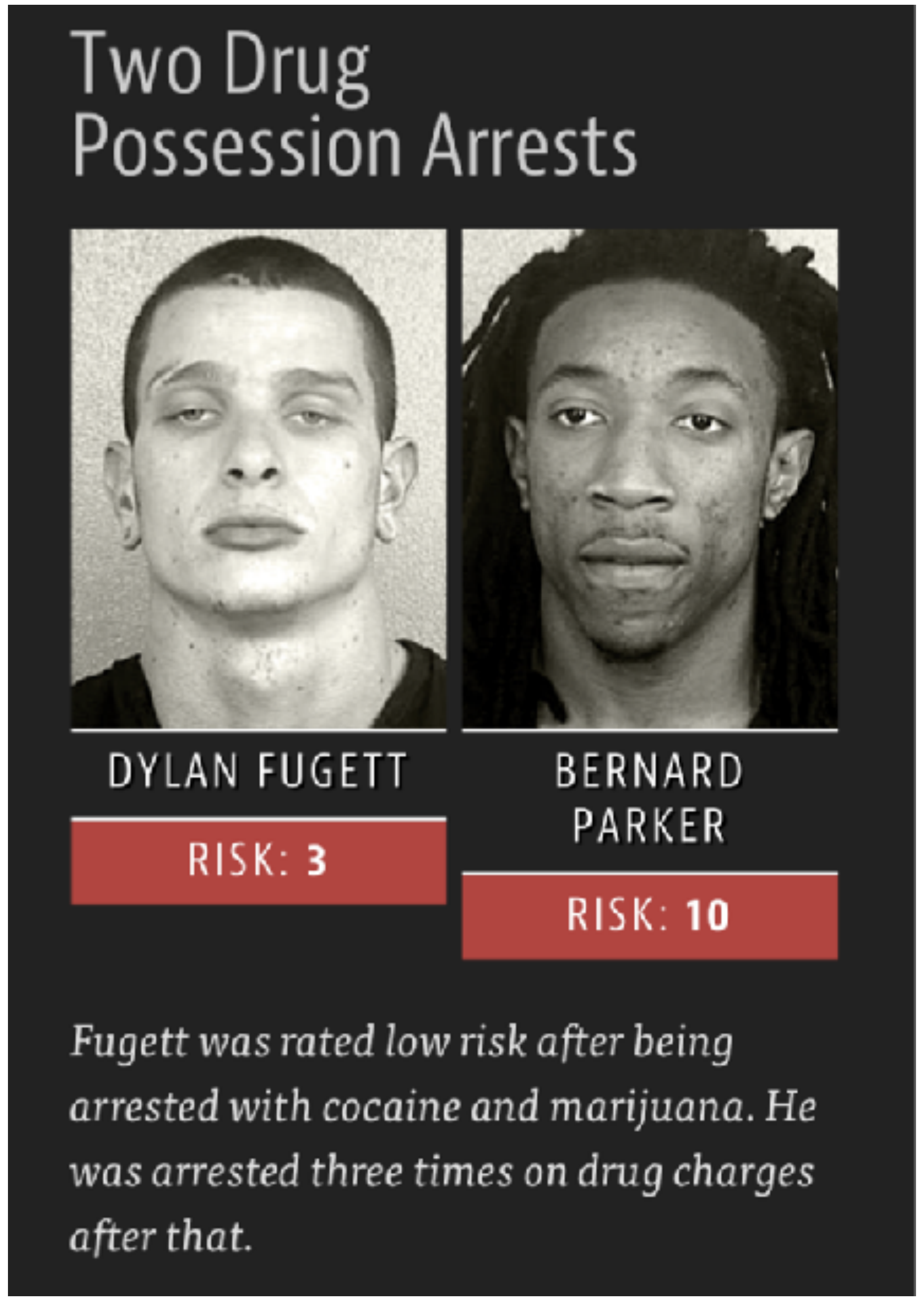
- Why might people want to hold algorithms accountable?
- What does ‘accountability’ mean? In general, and in data protection law.
- What technologies have been proposed in aid of accountability?

why might people want to hold algorithms accountable?

“Algorithmic decision-making”

- Decisions that are primarily based on the outputs of a machine learning model.
- Important, high-stakes decisions about *people,* e.g.
  - Who gets a loan? Who gets hired?


White residents
Black residents
Same-day delivery area
Percentage of residents living in ZIP codes with same-day delivery
All 65%
White 96
Black 41
Hispanic 74
Asian 95
Other 75


Two Drug Possession Arrests
DYLAN FUGETT
RISK: 3
BERNARD PARKER
RISK: 10
Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Fairness, Accountability, and Transparency in Machine Learning

Sponsors

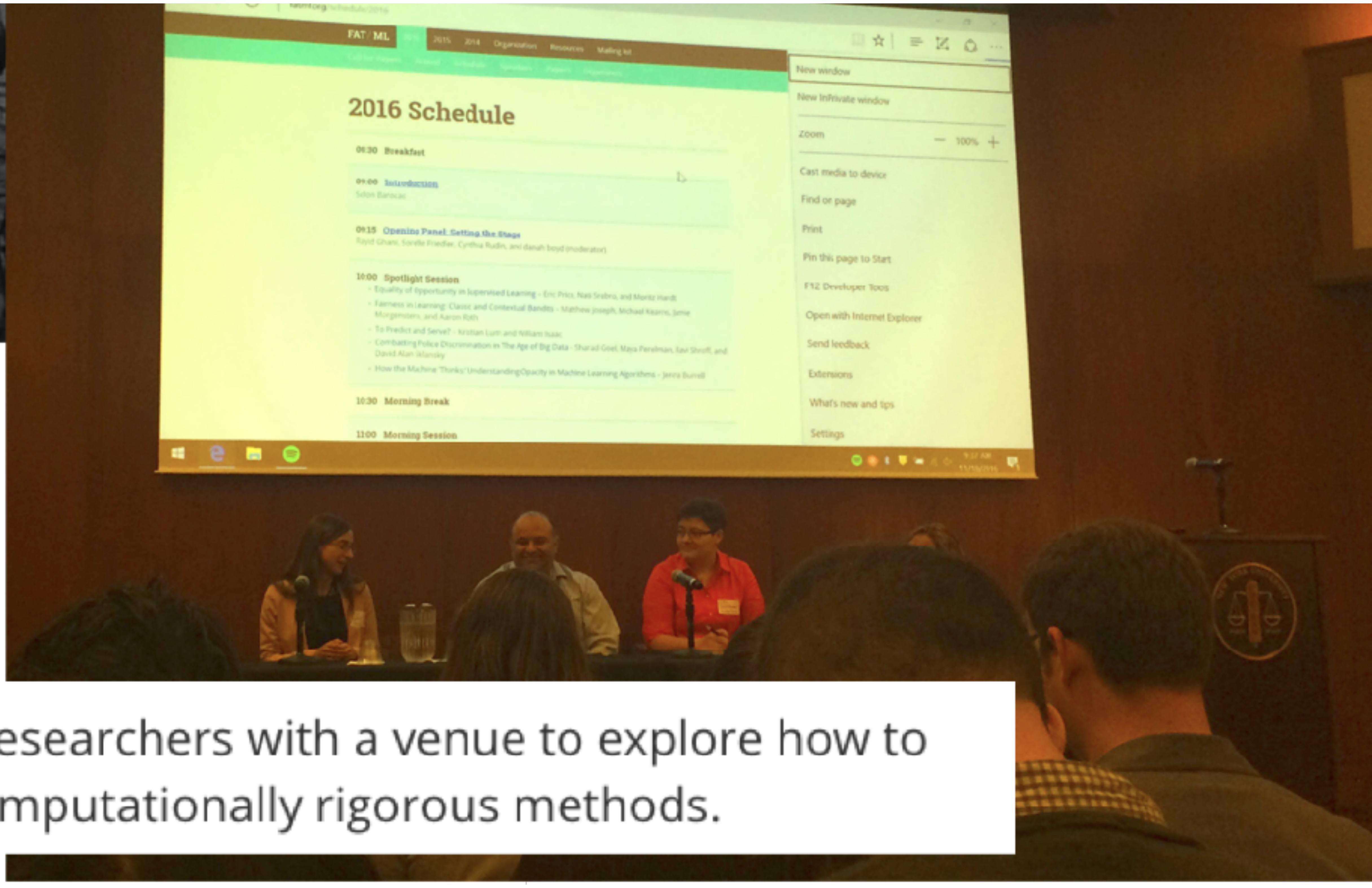
Google

Microsoft

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

At the same time, there is increasing alarm that the complexity of machine learning may



The goal of our 2016 workshop is to provide researchers with a venue to explore how to characterize and address these issues with computationally rigorous methods.

what does accountability mean?

Accountability?!?!?

- "Accountability is one of those golden concepts that no one can be against", a "hurrah-word" (Bovens 2015)
- Origins in William 1st, 1085; property holders provide a count of their possessions
- Now about powerful entities providing an account (a count) of their actions, decisions, procedures (Milgan 2000)

Accountability?!?!?

- 1970+: private sector management into public sector Schedler (1999)
- 2000’s: public sector governance modality now imposed on private sector through regulation (De Hert and Stefanatou 2015)

Accountability?!?!?

- Drawing from Bovens (2015):
- An **account-giving relationship**, between the **accountor** and **accountee**.
- Accountor has an obligation to **explain and justify** conduct
- Not just information, but debate, judgement, and possible sanctions or rewards

Accountability?!?!?

- **Distinct from fairness**: could be fair in an unaccountable way (“just trust us!”)

- **Distinct from transparency**: it’s not just about revealing what you’re doing, but explaining, justifying, and possibly facing judgement and sanctions

Accountability in data protection law

- OECD guidelines (1980): “A data controller should be accountable for complying with measures which give effect to the principles stated above”

- Two elements: **responsibility**, and **demonstrating compliance**

- “(…) accountability means more than ‘responsibility’. One can always act ‘responsibly’ without reference to anyone else. Accountability is always directed towards an external agent; responsibility is not” (Bennett)

Accountability in data protection law

- GDPR: Article 5(2): “The controller shall be **responsible** for, and be able to **demonstrate compliance** with, paragraph 1 (‘accountability’).”

- Both substantive compliance, and procedural demonstration (Urquhart et al 2017)

  - 1: Comply with the principles

  - 2: Demonstrate how

Accountability in data protection law

- Measures intended to ‘make controller responsible’ include:
  - Appointing a DPO
  - Documentation of interactions (e.g. keep a record of consent)
  - Conduct a DPIA

Accountability in data protection law

- Take any specific act of processing of personal data, and obtain a record of all the compliance-related activity that preceded it:
  - What was the controller’s purpose for processing
  - How was it decided on?
  - Who was involved in the decision? Who is or was the data protection officer?
  - How were decisions made about the balancing of rights (both between DP and other rights, and within DP), and other interests?

technologies for accountability

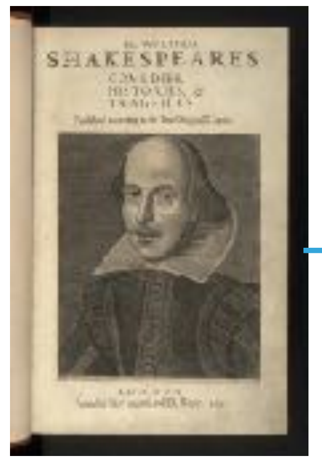
proving things about data / processes

hash functions

Inputs

Outputs

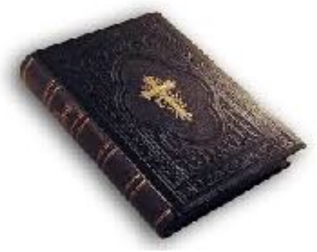
Cj7FHH7MwS3A



Ot20tf9R9d1s

I9ZQI6DVVCAT

Hash (   )



Peter Todd @peterktodd · 13 Aug 2017

Can you give an example?

3 2

Peter Todd @peterktodd · 13 Aug 2017

Actually, better yet, give me a hash commitment to an example... Let's not do @VitalikButerin's homework for him.

2 2 25

Alphonse Pace @alpacasw · 13 Aug 2017

Wow that is sad he doesn't know, the answer us super easy.

1

Peter Todd
@peterktodd

Follow

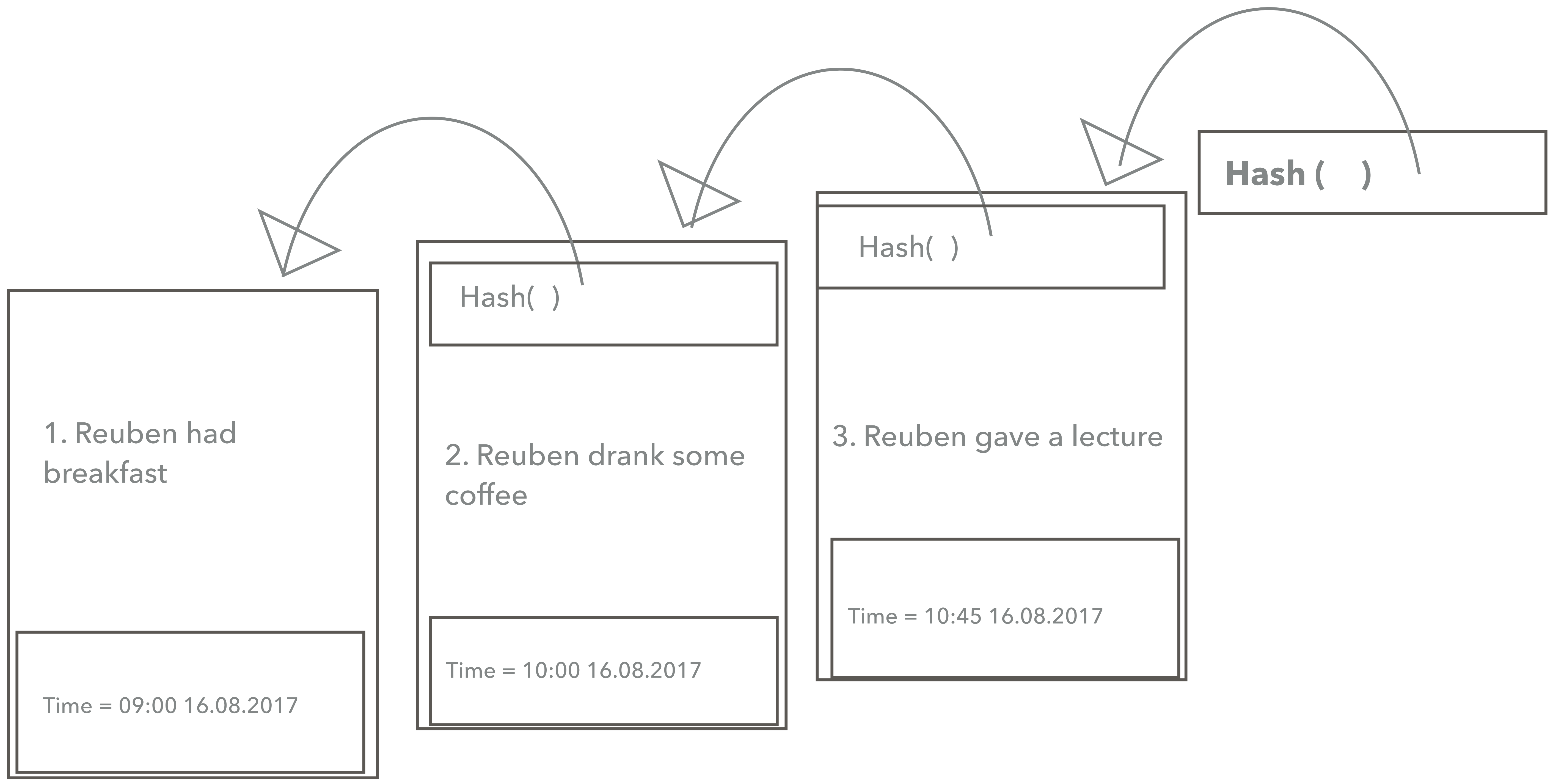
Replying to @alpacasw @fluffypony @VitalikButerin

I'd suggest you write up an answer and post a hash commitment to it, like I did a few hours ago. :)

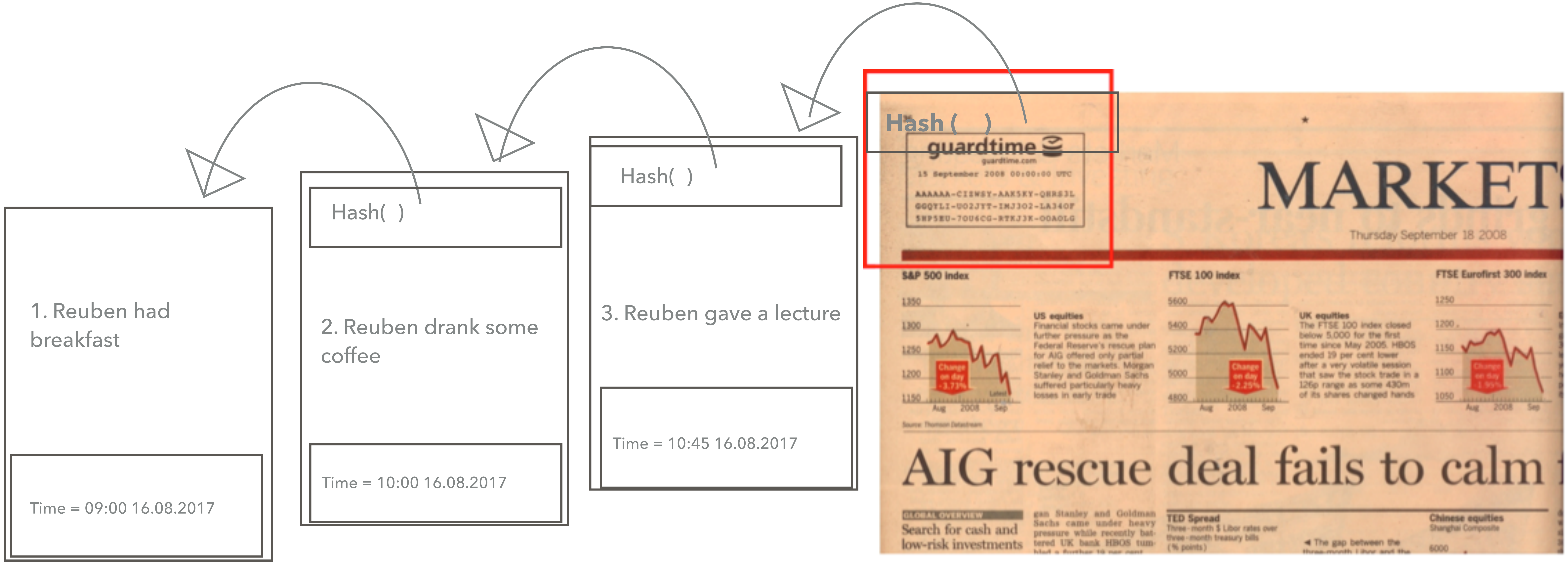
5:15 AM - 13 Aug 2017

1 Like

Secure time-stamping



Secure time-stamping



Verifiable logs for auditing data use


HAL
archives-ouvertes.fr

Log Analysis for Data Protection Accountability

Denis Butin, Daniel Le Métayer

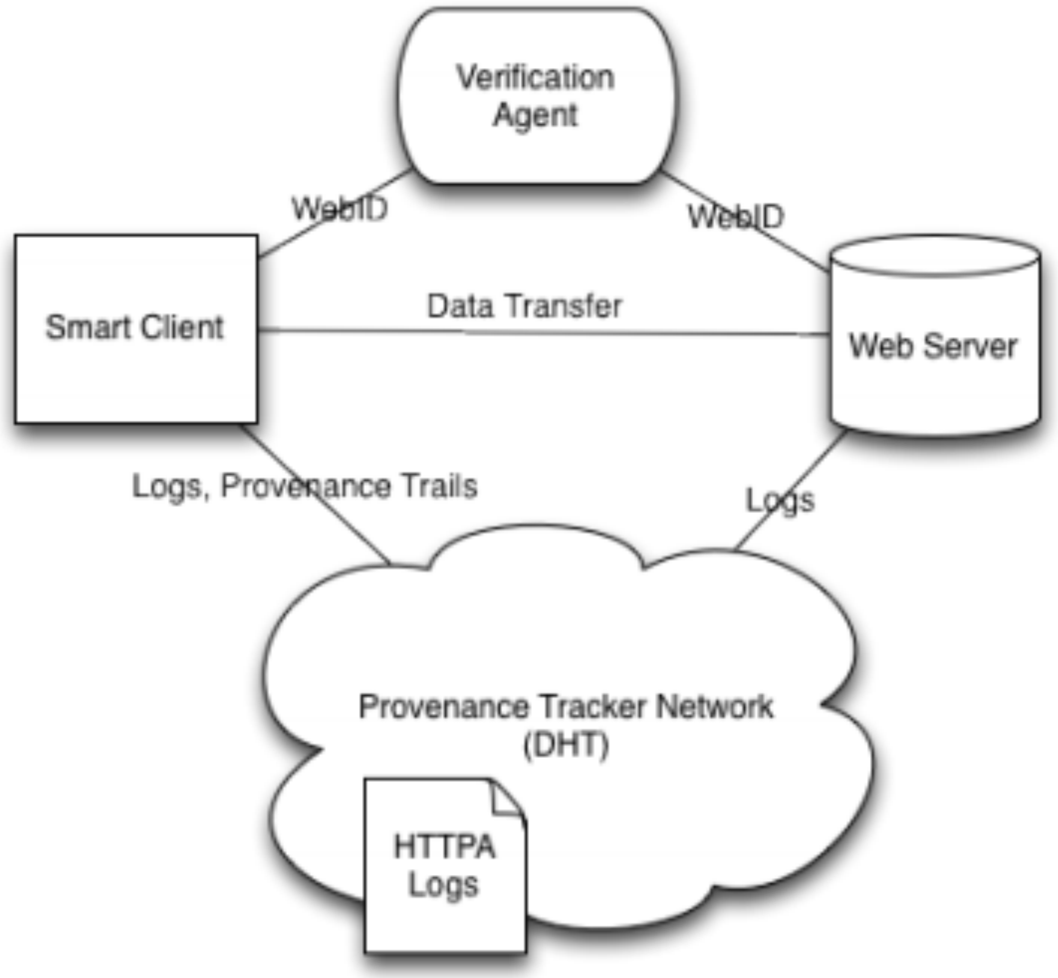
▶ To cite this version:

Denis Butin, Daniel Le Métayer. Log Analysis for Data Protection Accountability. FM2014 - 19th International Symposium on Formal Methods, May 2014, National University of Singapore (NUS), Singapore. Springer, pp.163-178, 2014, Lecture Notes in Computer Science. <http://www.comp.nus.edu.sg/ pat/FM2014/>. <10.1007/978-3-319-06410-9_12>. <hal-00984308>

HAL Id: hal-00984308

https://hal.inria.fr/hal-00984308

Submitted on 28 Apr 2014



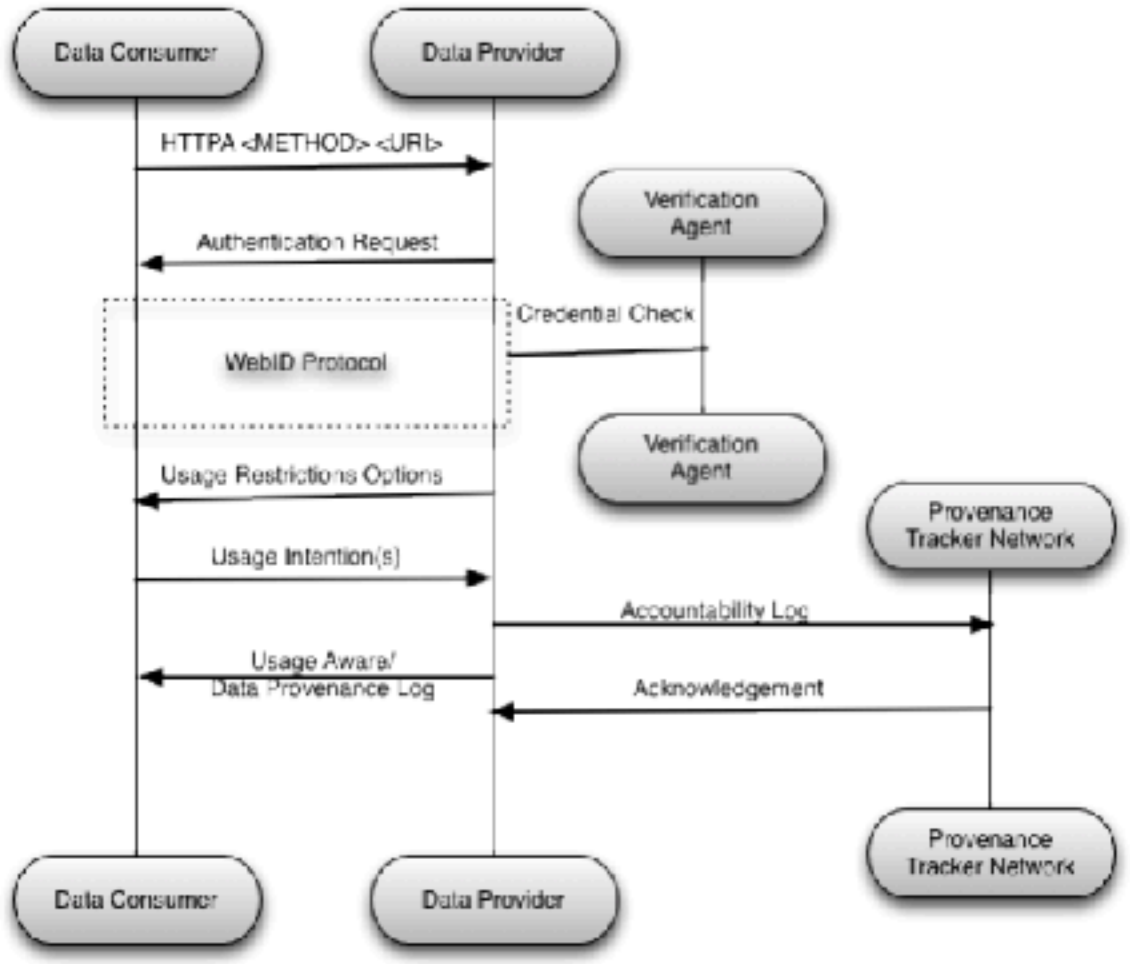


Figure 2: Data Creation HTTPA Method

Butin, Denis, and Daniel Le Métayer. "Log analysis for data protection accountability." *International Symposium on Formal Methods*. Springer, Cham, 2014.

Seneviratne, Oshani, and Lalana Kagal. "HTTPa: accountable HTTP." *IAB/w3C internet privacy workshop*. Vol. 42. 2010.

Verifiable logs for auditing data use

- Software is constantly publishing logs of events during runtime
- logs are immutable, encrypted, propagation restricted to allowed purposes
- If misuse of data is discovered, in theory the perpetrator can be found through the chain of users who have shared the data

Verifiable logs for algorithm accountability?

- An accountor can make public commitments such that they cannot deny them later, including:
  - training datasets, modelling processes, data storage, parameterisation, tuning and tweaking, thresholds, etc.
- Later, accountee (e.g. data subject, regulator) can ask to check this model is the same as that model which has been verified as meeting certain constraints

Prove this model is the same as that one

Prove this output came from this model

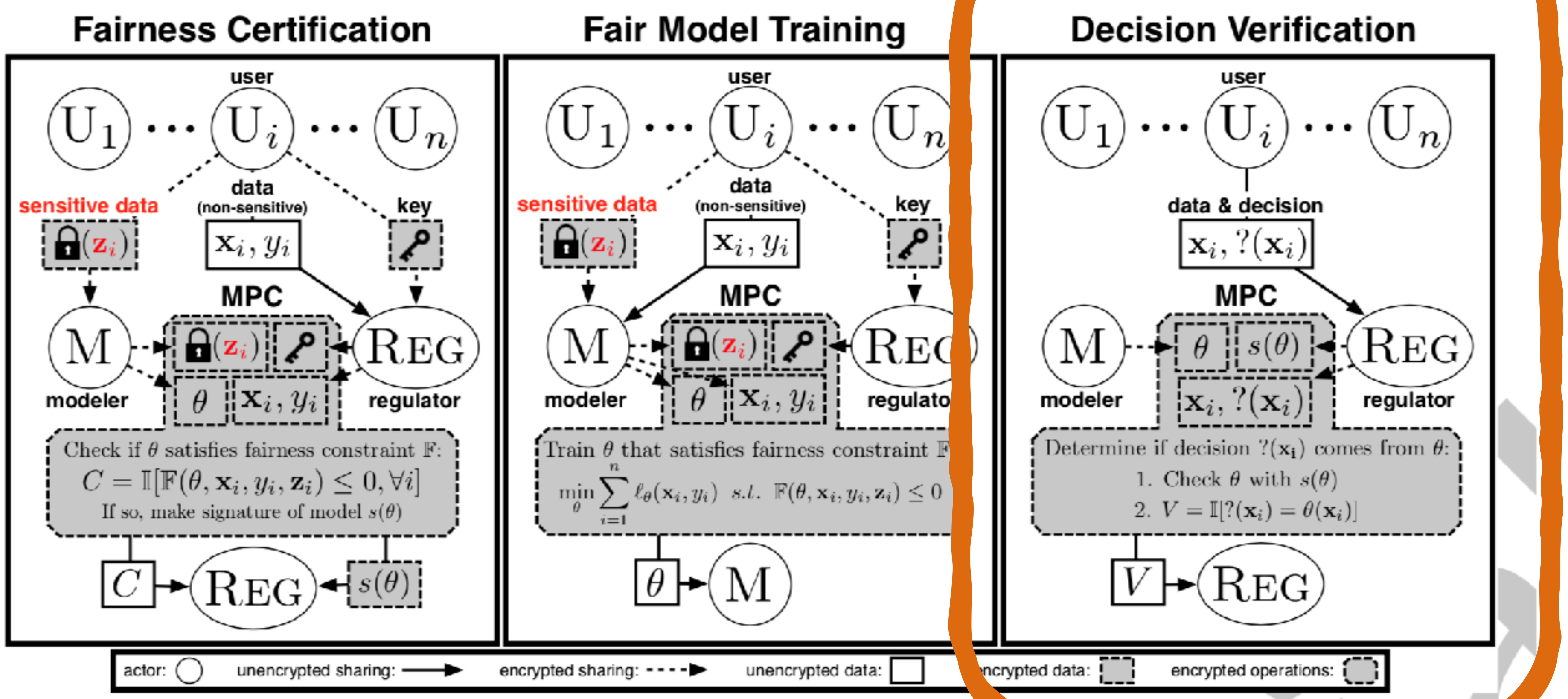
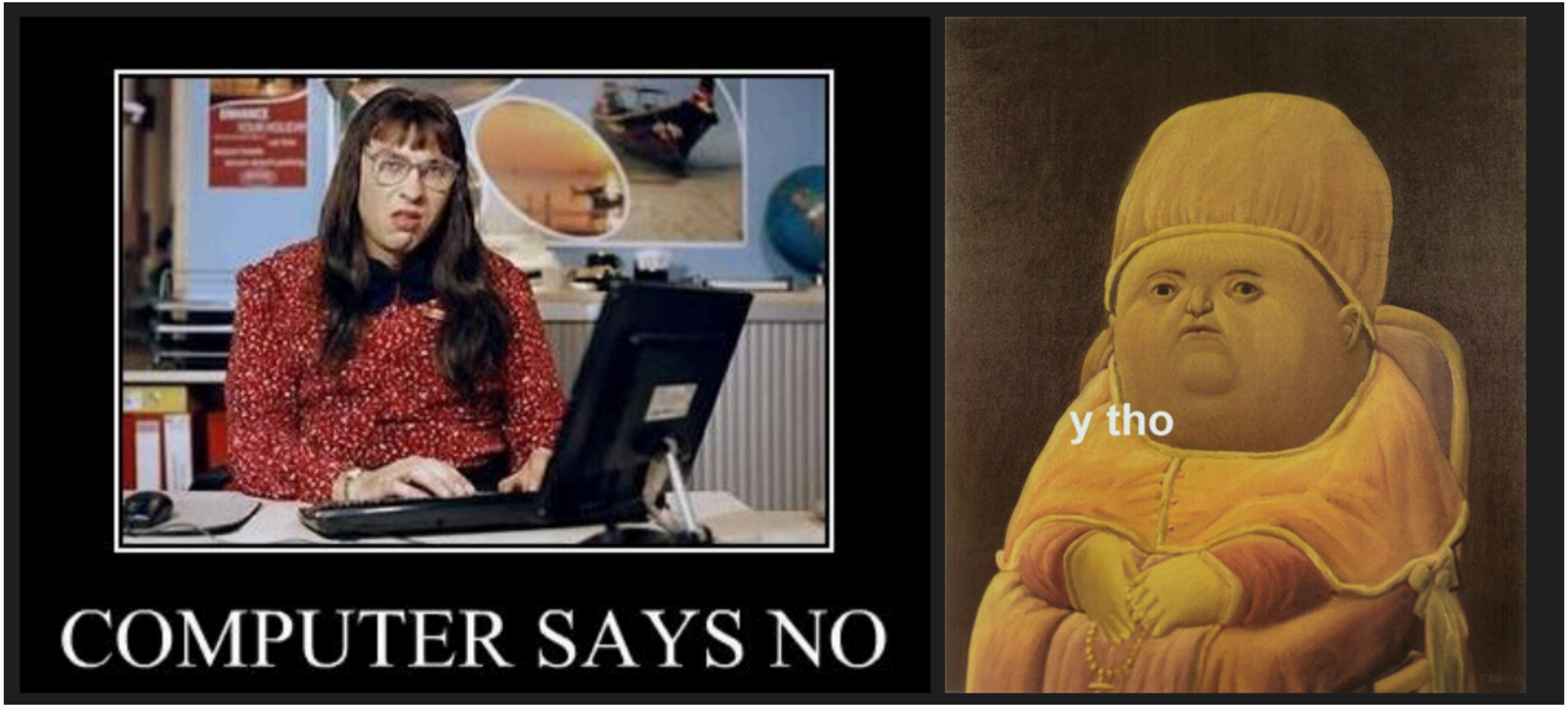


Figure 1. Our setup for *Fairness certification* (*Left*), *Fair model training* (*Center*), and ...

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P., & Weller, A. (2018). Blind Justice: Fairness with Encrypted Sensitive Attributes. *arXiv preprint arXiv: 1806.03281*.

explaining algorithm outputs



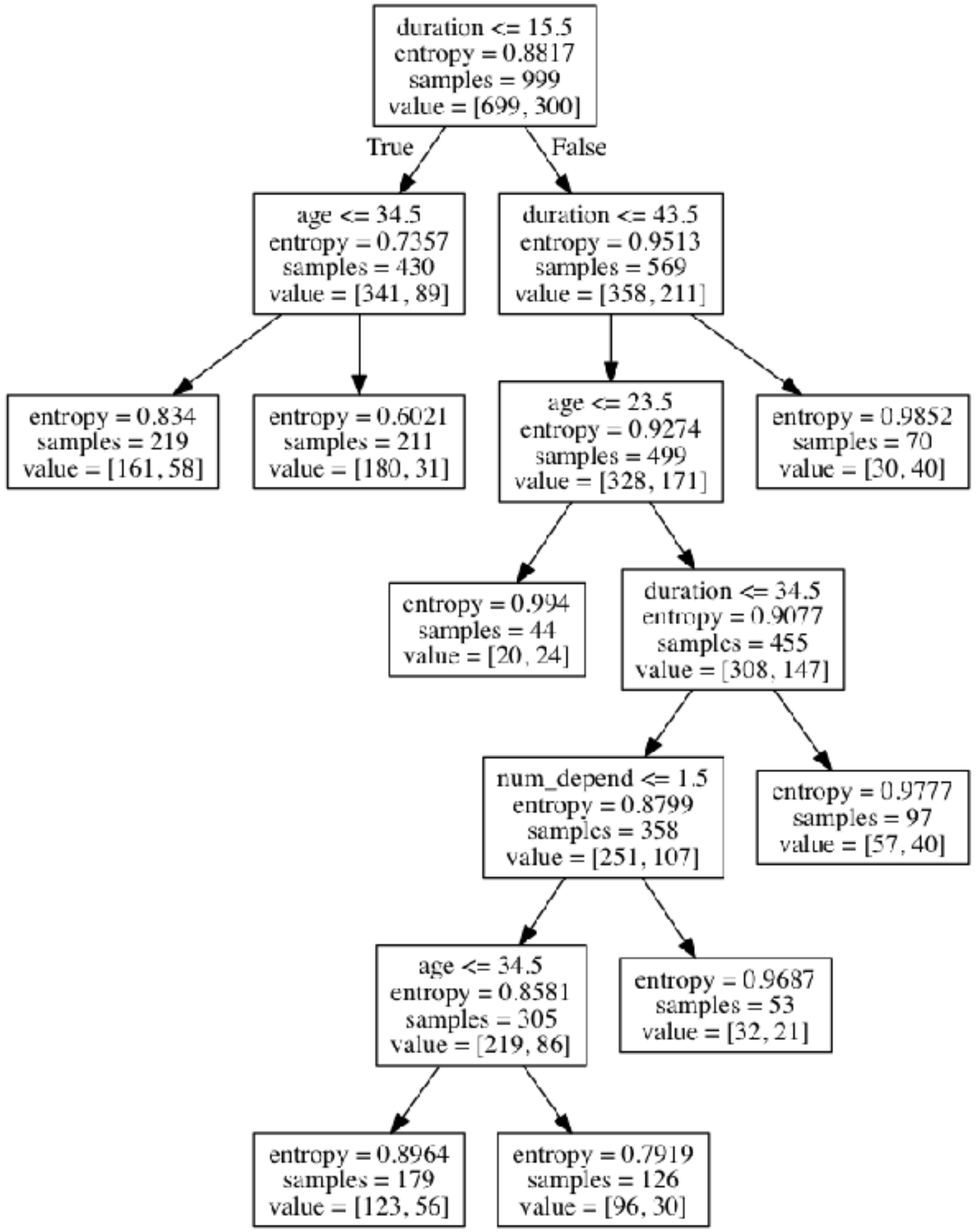
Who might want explanations, and why?

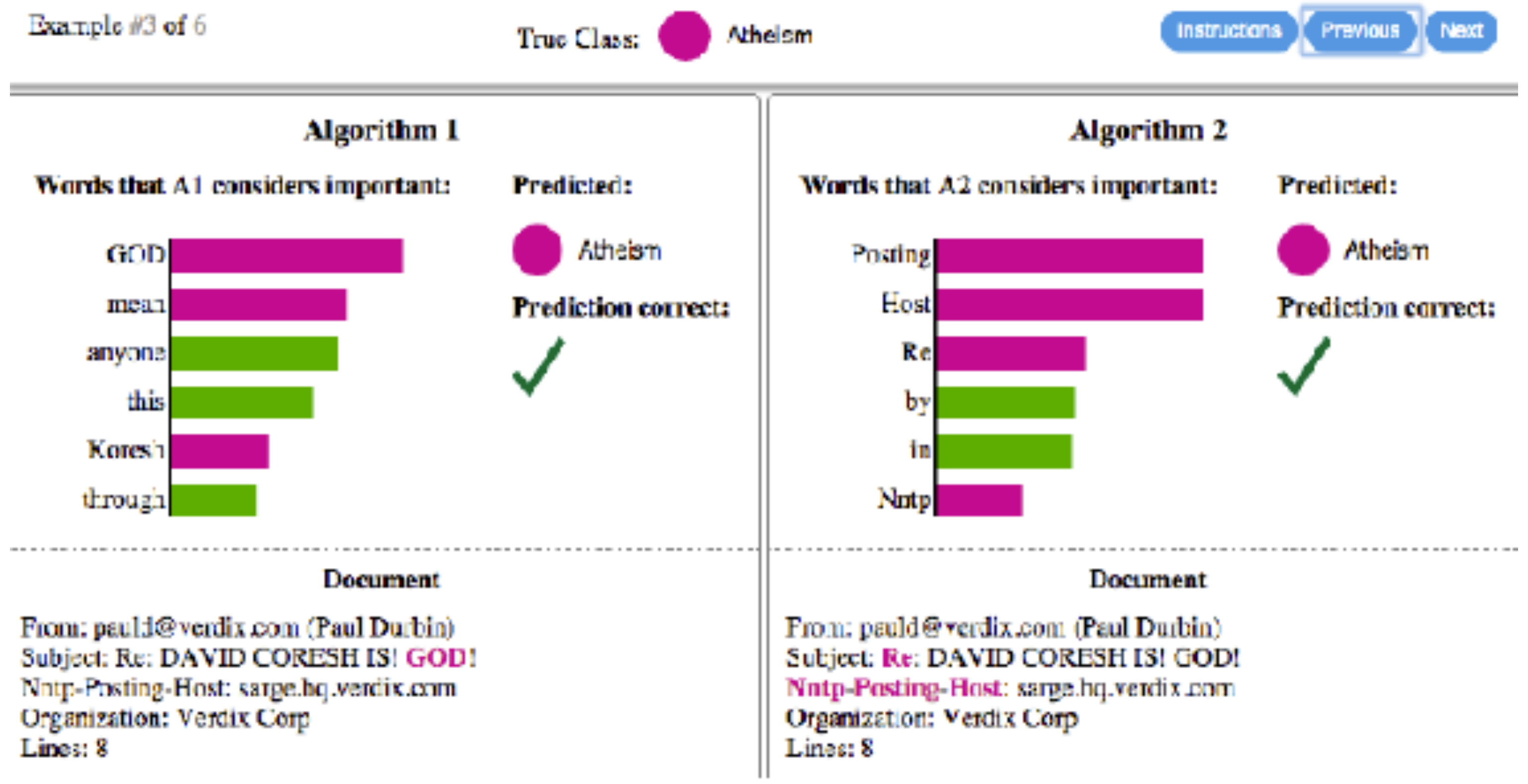
	action\question	**Is it fair?** *Using legal/socially acceptable logics*	**Does it work?** *Does it fail unevenly, or over time?*	**Do I get it?** *Can I profile the profilers?*
Decision subject	**Mount a legal or regulatory challenge**			
	Opt for a human review (art. 22)			
	Avoid product or service			
	Name-and-shame			
	Act to change your data representation			
Decision maker	**Lower business risk**			
	Regulatory compliance			
	User trust			
	Mitigate automation bias			

Explanation approaches

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)





How do ML explanations affect perceptions of procedural justice?

- Tested people’s perceptions of justice in response to various hypothetical cases

- Perceptions of justice in decision-making: *informational, procedural, distributive* (Colquitt 2015)

- Binns, Reuben, et al. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018.

- Different *contexts*: loans, employment, insurance, travel, fraud

    - e.g. ‘Sarah has been evaluated at work by a computer system…’

Page 8: Promotion at work

The human resources department at a large company is selecting current employees for a promotion to a new role of senior salesperson. Their system for assessing applications is based on a computer model, which predicts how well the applicant is likely to perform in the role of senior salesperson. The computer model makes its predictions based on data collected about thousands of previous recruits and how well they performed after promotion to the role.

Each applicant is given a prediction based on the data held about them by the human resources department. Applicants who are predicted to perform to a high standard will be automatically considered for promotion.

Ali is applying for the promotion to senior salesperson.

- He has been working in sales full-time for 3 years.
- He makes an average of 126 sales per month
- He has an average customer satisfaction of 7/10
- He has arrived late for a shift 13 times in the last year
- He scored 98% on his skills assessment test

Based on this information, the computer model has decided not to select Ali for promotion to senior salesperson.

The HR department provides Ali with the following information about the computer's decision:

- Same decision, different *explanation styles*:

  - e.g. ‘If you had 2 years more experience, and better sales numbers, you’d be promoted’

a new role of senior salesperson. Their system for assessing applications is based on a computer model, which predicts how well the applicant is likely to perform in the role of senior salesperson. The computer model makes its predictions based on data collected about thousands of previous recruits and how well they performed after promotion to the role.

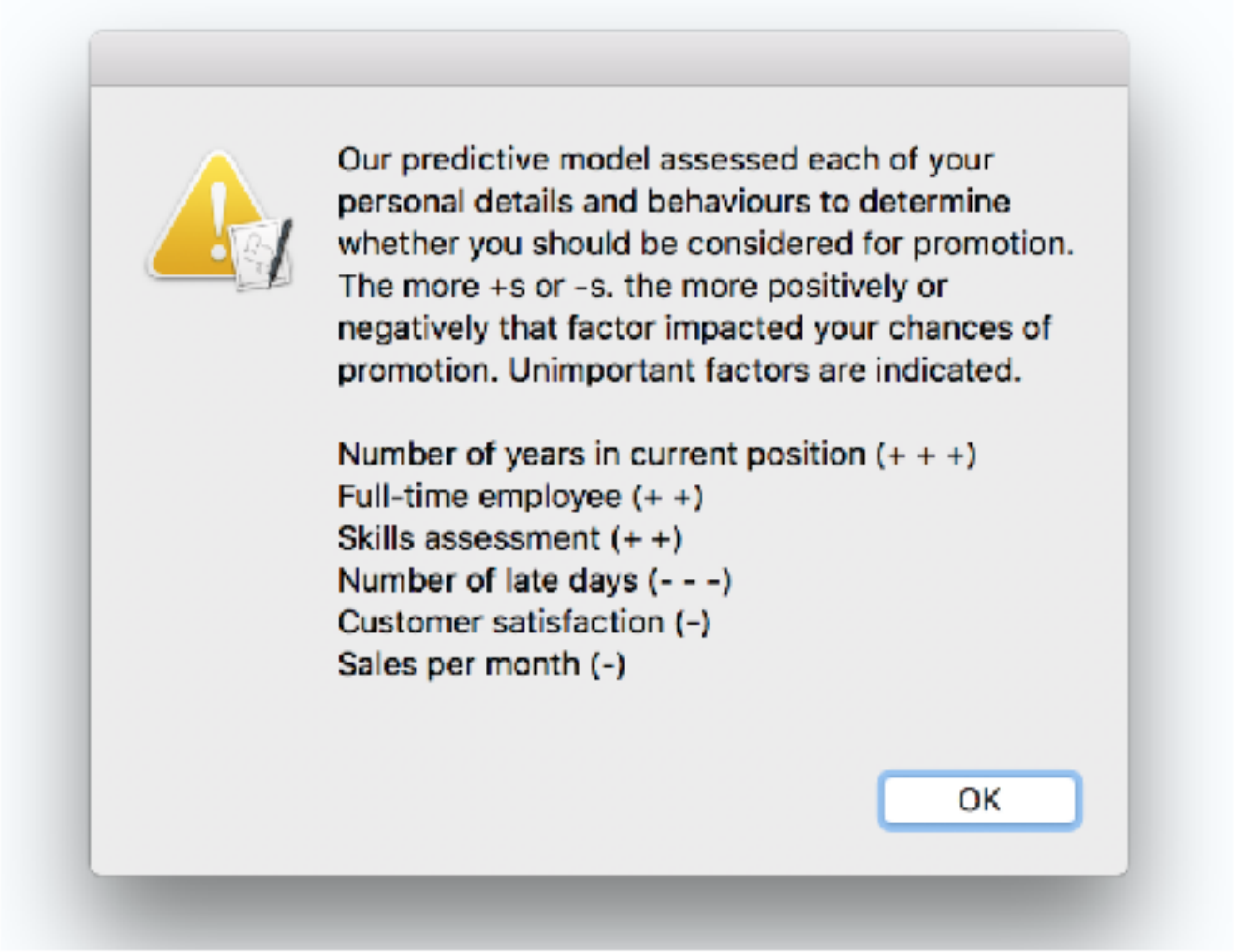
Each applicant is given a prediction based on the data held about them by the human resources department. Applicants who are predicted to perform to a high standard will be automatically considered for promotion.

Ali is applying for the promotion to senior salesperson.

- He has been working in sales full-time for 3 years.
- He makes an average of 126 sales per month
- He has an average customer satisfaction of 7/10
- He has arrived late for a shift 13 times in the last year
- He scored 98% on his skills assessment test

Based on this information, the computer model has decided not to select Ali for promotion to senior salesperson.

The HR department provides Ali with the following information about the computer's decision:



Please rate your agreement with the following statements

I agree with the decision * *Required*

Explanation styles

- Case-based
- Sensitivity
- Input influence
- Demographic

Case–based explanation

This decision was based on thousands of similar cases from the past. For example, a similar case to yours is a previous customer, Claire. She was 38 years old with 18 years of driving experience, drove 850 miles per month, occasionally exceeded the speed limit, and 25% of her trips took place at night. Claire was involved in one accident in the following year.

ok!

Sensitivity–based explanation

> If 10% or less of your driving took place at night, you would have qualified for the cheapest tier.

> If your average miles per month were 700 or less, you would have qualified for the cheapest tier.

ok!

Input influence–based explanation

Our predictive model assessed your personal information and driving behaviour in order to predict your chances of having an accident. The more +s or -s, the more positively or negatively that factor impacted your predicted chance of accidents. Unimportant factors are indicated.

\> Your age (---)

\> Driving experience (---)

\> Level of adherence to speed limit (-)

\> Number of trips taken at night (++)

\> Miles per month (+)

ok!

Demographic–based explanation

\> 29% of female drivers qualified for the cheapest tier.

\> 31% of drivers in your age group [30–39] qualified for the cheapest tier.

\> 35% of drivers with 17 years of experience qualified for the cheapest tier.

\> 15% of drivers who have been on one accident which was not their fault qualified for the cheapest tier.

\> 26% of drivers who regularly travel at night qualified for the cheapest tier.

\> 21% of drivers who exceed the speed limit once ever two months qualified for the cheapest tier

ok!

Questions about the system design

'Oh that's so mean! [...] I can't do the maths, but why is it so specific? Hmmm. I don't understand. I don't know why the cut-off is like that.'

Questions about training data (sample size)

'I don't know how many previous customers they're basing it on…'

'I'm gonna assume that it looked at more than just John!'

Explanation is not enough (reasons)

'Perhaps it's unfair to make the decision by just comparing him to other people and then looking at the statistics, he isn't the same person.' [...] 'They [...] seem like [...] just random stats, not reasons for why you'd make a decision'

Explanation is not enough (interaction)

‘there’s no sense of negotiation’

‘no opportunity for ‘human interaction’

Explanation $\subset$ Accountability

- Algorithm explanations may be a necessary part of accountability, but probably insufficient
- What we want to challenge is not necessarily just the algorithm, but the entire system, values, governance processes…
- Most of this will not be stored as structured data!

Justification and contestation

- These technologies are focused on ***proving*** properties of algorithms or ***explaining*** their outputs
- But accountability is fundamentally about **justification, contestation,** and potentially **sanctions**. What role might provenance technologies play in supporting these broader goals?
- How could provenance be combined with emerging HCI work on algorithmic accountability and GDPR-compliant ML? (Veale et al 2018)

- Veale, Michael, Reuben Binns, and Max Van Kleek. "Some HCI Priorities for GDPR-Compliant Machine Learning." *arXiv preprint arXiv:1803.06174* (2018).