FAT Predictive Analytics

Facets of (un)fairness and (non)transparency

prof. dr. Mykola Pechenizkiy

http://www.win.tue.nl/~mpechen/

Responsible Data Science, VU Amsterdam

19 October 2017

Fear of AI vs. practical challenges

What are we told to fear?

- AI will take our jobs
- *General* AI is coming, artificial brain
- AI will conquer the world

What is actually going on?

- more of *specialized* AI is used everywhere
- experts using AI will be replacing experts who don't

What are our actual challenges for next 5-10 years?

- Transparency, interpretability, interactivity and ethics of predictive analytics as specialized AI
- Have new generations of data scientists ready

Predictive analytics impacts everyone

- Everyday life: services by Google/Facebook/…

- Mid-term and long-term via contribution to
  - Science
  - Industries
  - Healthcare
  - Education
- This all has little to do with AI conquering us

What is Predictive Analytics?

AI-readable big data matrix

Ground truth: known correct answers Y

$$X * w \overset{?}{=} Y$$

What we try to minimize

$$\text{Error} = Y - X*w$$

Black-box magic to learn to guess correct answers Y

Focus on accuracy and efficiency

- More complex and expressive models
  - ensembles and deep neural networks
- Support for handling 5V' of Big Data
  - more data, data types & operational settings
- More robust models
  - handling anomalies and changes in evolving data

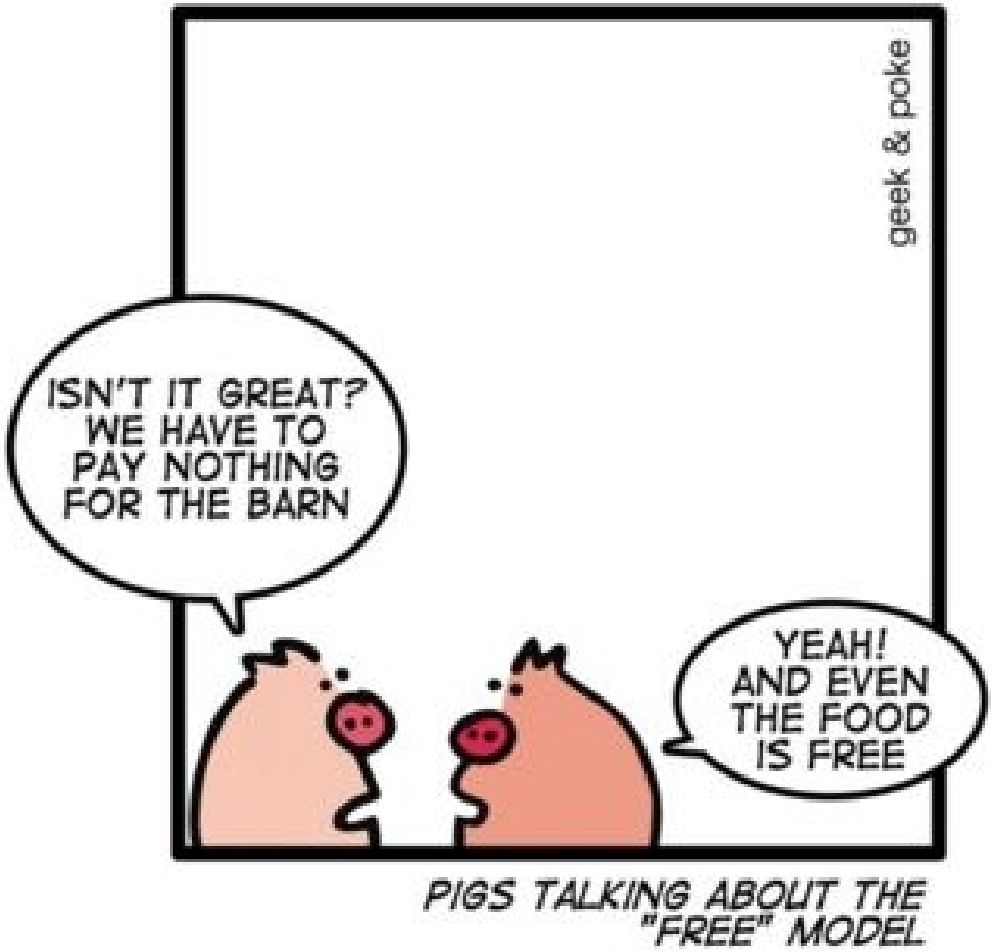
"Anything you can do, *AI* can *do better*"

Massive Automation of Decision Making

What these services are really optimizing for?
What should/could be made public about how the algorithms work?

What companies are optimizing for?

"I want everything I touch to turn to gold"



Do companies really know what they are optimizing for?

What are they really optimizing for?

- Inaccurate predictions vs. “accurate” (≠ perfect), but unethical
  - discrimination by gender, race, religion
- Is it customer centric or organization centric?
  - an average person vs. each individual?
- Based on what data?
  - Biases, noise, …
- With help of what Big Data analytics?
  - Can a chosen approach give an answer to the set optimization goal?

Facets of fairness

- Existential questions: “cars programmed to kill” vs. **quantifiable** fairness/discrimination
- Defining and measuring fairness
  - Achieving parity or satisfying preferences?
  - Focus on treatment or on impact?
  - Individual or group level
- Achieving fairness (by design)
  - Discrimination-aware classification, regression, recommendation, …

Facets of transparency

- Model transparency
  - How an induced model works (deep learning)
  - How the model was induced/fine-tuned – what was it optimized for (e.g. TP/FP tradeoff)
  - How well does the model work (and for what kinds of input; quantitative vs. qualitative)
- Model output transparency
  - Linking to evidence that **may** explain a decision
- Model performance transparency, wrt
  - considered trade-offs and defined fairness
  - Average/group level vs. individual level
- Stakeholders perspectives
  - Regulators vs. domain experts/decision makers vs. analysts vs. scientists

Decision making: Automation vs. Support

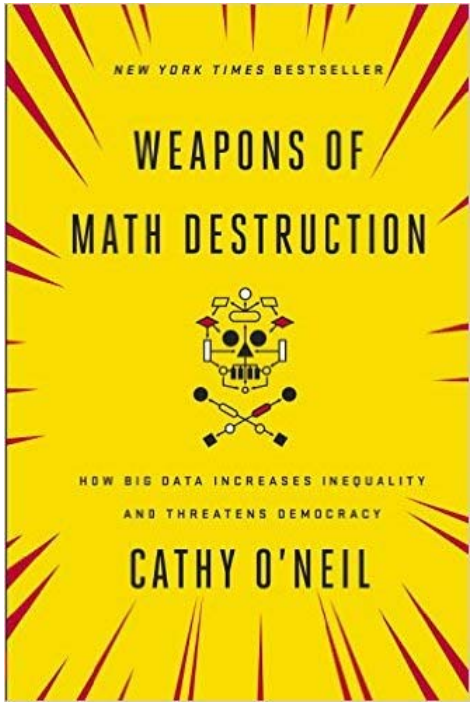
Develop foundations and techniques for next-generation Data Mining for Data Science

"Anything you can do, *AI* can *do better*"

"Anything you can do, *AI* can *help you do better*"

Facilitating more *reliable*, *transparent* and *safe* predictive analytics with *experts-in-the-loop*

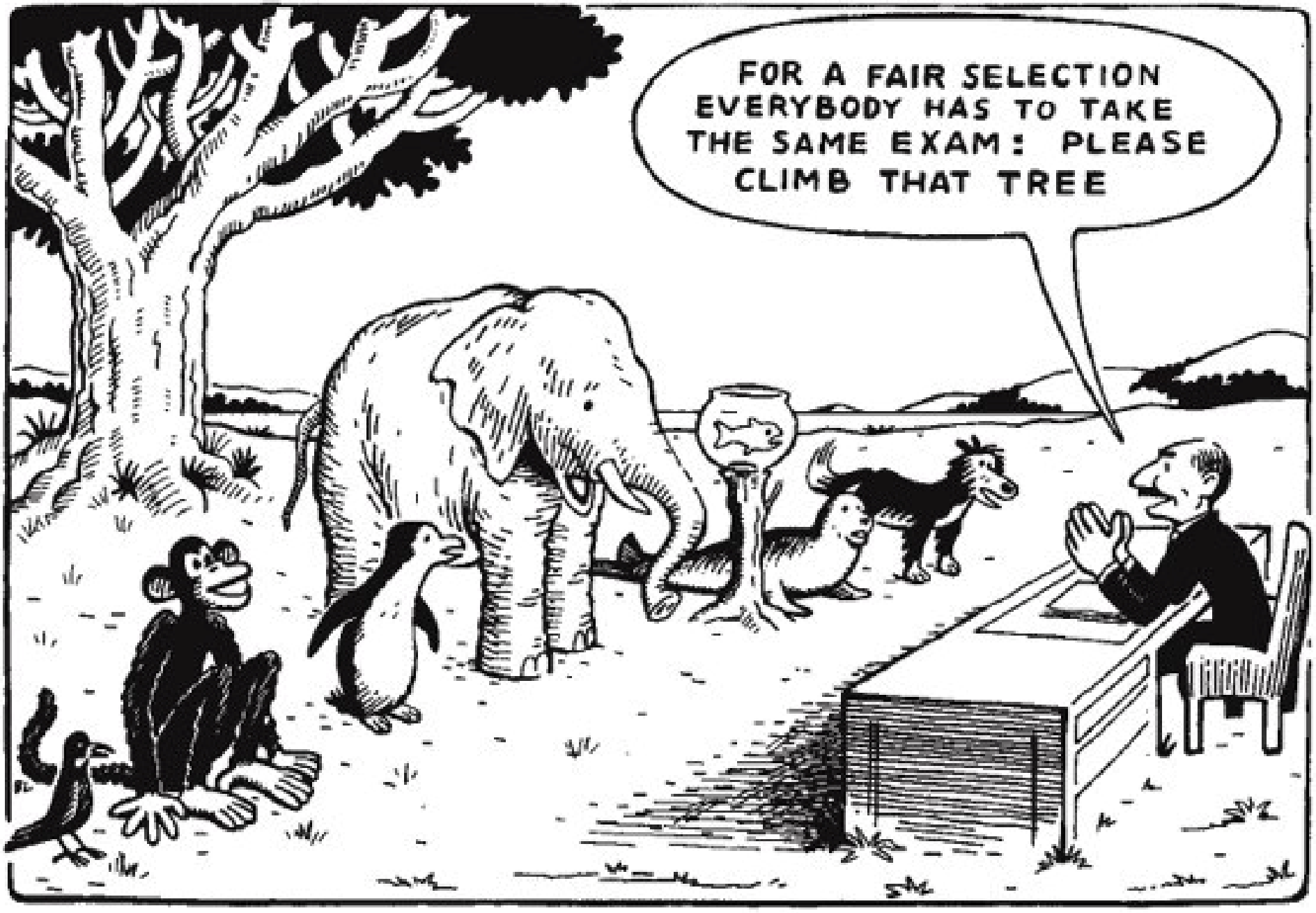
Dangers of optimizing for KPIs



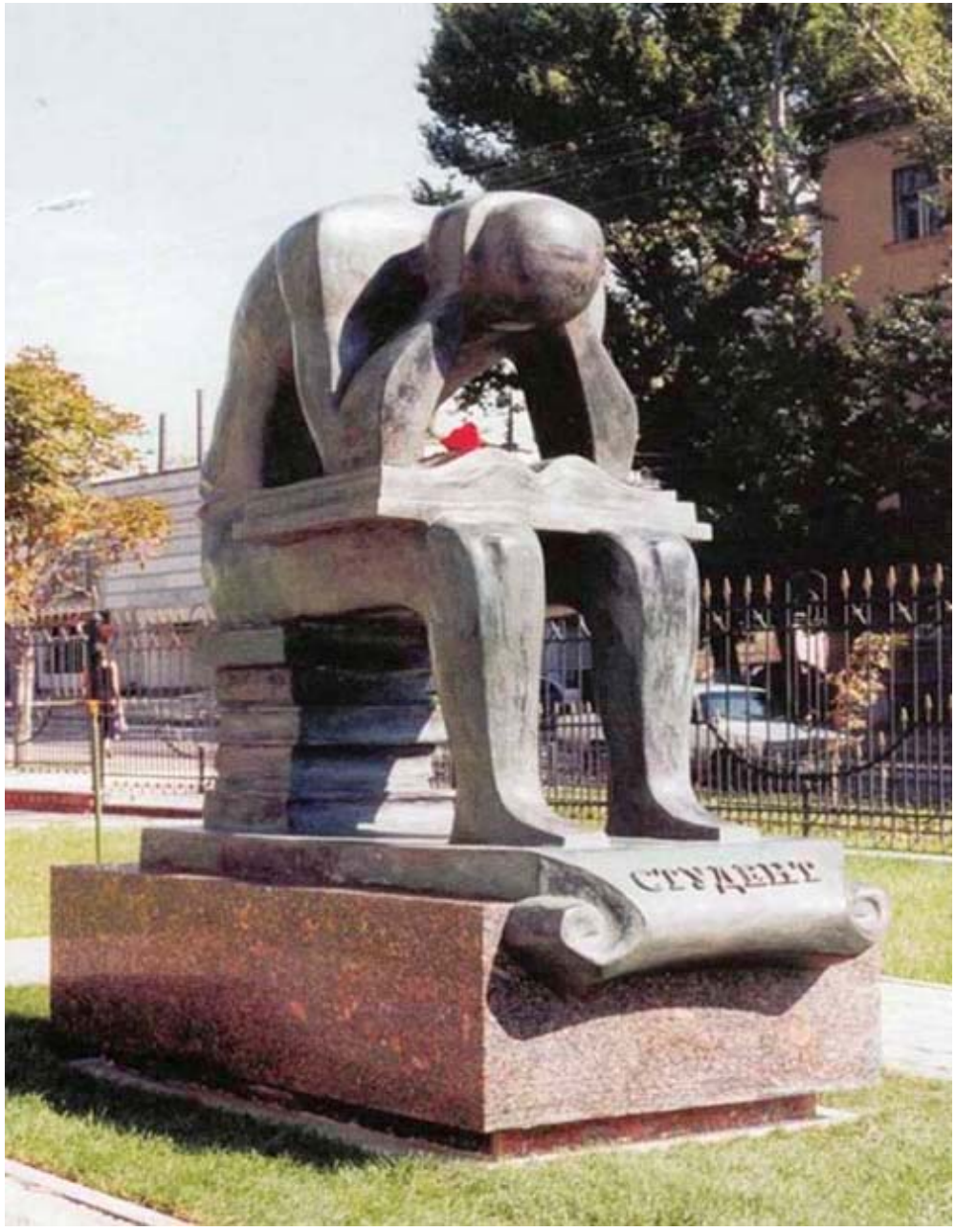
- Education ecosystem
- Academic/research ecosystem
- Police and justice

Things can go rather wrong despite of good intentions behind the set KPIs

Dangers of optimizing for KPIs

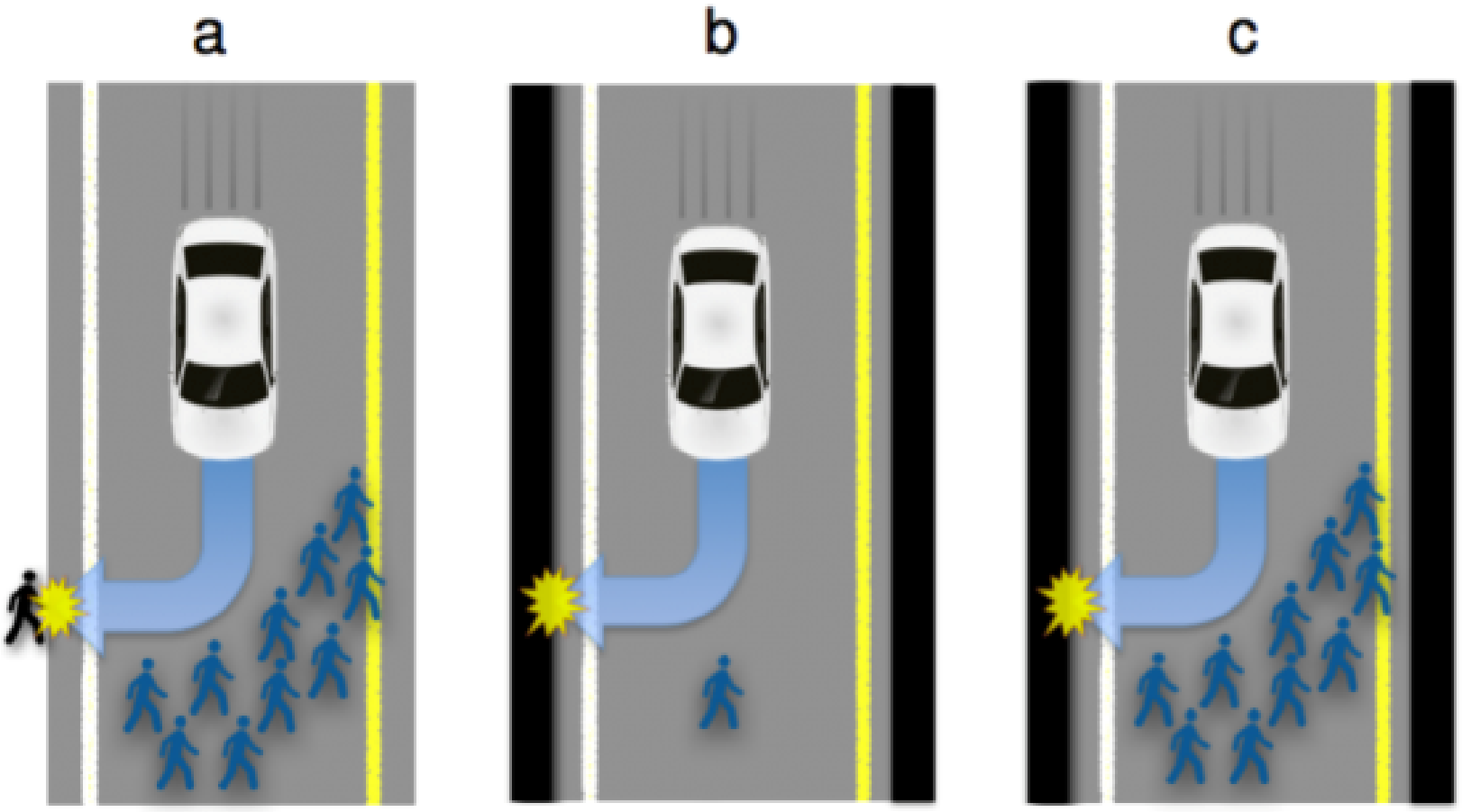


Optimizing for passing the tests



https://nl.pinterest.com/pin/567312884283385075/

Self-Driving Cars Programmed to Kill?



Alain Kornhauser, director of the transportation program at Princeton University. “It doesn’t work in all circumstances. Drivers don’t necessarily know when the car goes from tracking fine to a gray area when the car is confused, and then to a situation when the car doesn’t know where it’s going. These things aren’t well-defined.”

http://www.technologyreview.com/view/542626/why-self-driving-cars-must-be-programmed-to-kill/

Computational Journalism

"… automated decisions of algorithms deserve every bit as much scrutiny as other powerful and influential actors." http://www.nickdiakopoulos.com/

- What is the basis for a prioritization decision? Is it fair and just, or discriminatory?
- What are the limits of an algorithm and when is it known to (have high chances to) fail?
- How has the algorithm been tuned to privilege FPs/FNs? Is it fair/just?
- What are the potential biases of the training data?

Computational Journalism: Google Ads

Control your Google ads

You can control the ads that are delivered to you based on your Google Account, across devices, by editing these settings. These ads are more likely to be useful and relevant to you.

Your interests

- [x] Air Force
- [x] Bicycles & Accessories
- [x] Cleaning Agents
- [x] Dogs
- [x] Executive Branch
- [x] Flora & Fauna
- [x] Geographic Reference
- [x] Holidays & Seasonal Events
- [x] Arts & Entertainment
- [x] Books & Literature
- [x] Computers & Electronics
- [x] East Asian Music
- [x] Finance
- [x] Food & Drink
- [x] Government
- [x] Home & Garden
- [x] Autos & Vehicles
- [x] Cats
- [x] Cruises & Charters
- [x] Education
- [x] Fish & Aquaria
- [x] Games
- [x] Hair Care
- [x] Home Improvement

+ ADD NEW INTEREST VIEW 26 MORE INTERESTS WHERE DID THESE COME FROM?

Google Ads Settings

- "setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male"

  Automated Experiments on Ad Privacy Settings
  http://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf

- ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity

  http://dataprivacylab.org/projects/onlineads/1071-1.pdf

- target people who live in low-income neighborhoods with high-interest loans

Predictive analytics in Decision support

- Police, security, intelligence – screening suspects
- Judges – deciding on pre-trial period of suspects
- eCommerce – cookie-based price adjustments
- Education – giving a (negative) study advice
- Medical diagnostics, personalized medicine, …
- Mortgages, car insurances, CV screening, jobs, salaries, funding decisions, …

Why are ML algorithms so valuable?

- ability to replicate human decision-making efficiently
- ability to perform with higher accuracy and lower pre-justice than humans

Predictive analytics that works!?

"All models are wrong, but some are useful"

hard to make 100% accurate ⇔ many trade-offs:

- Well formulated and well studied:
  - precision-recall; bias-variance; robustness-sensitivity;
- Not so well formulated, and not so well studied:
  - accuracy-fairness, accuracy-confidentiality, accuracy-privacy, accuracy-transparency, …
  - accuracy-human resources: inducing, applying and improving/maintaining does not come for free

Discrimination Prevention: Why?

AI algorithms are considered to be inherently objective, or to have no bad intent, but:

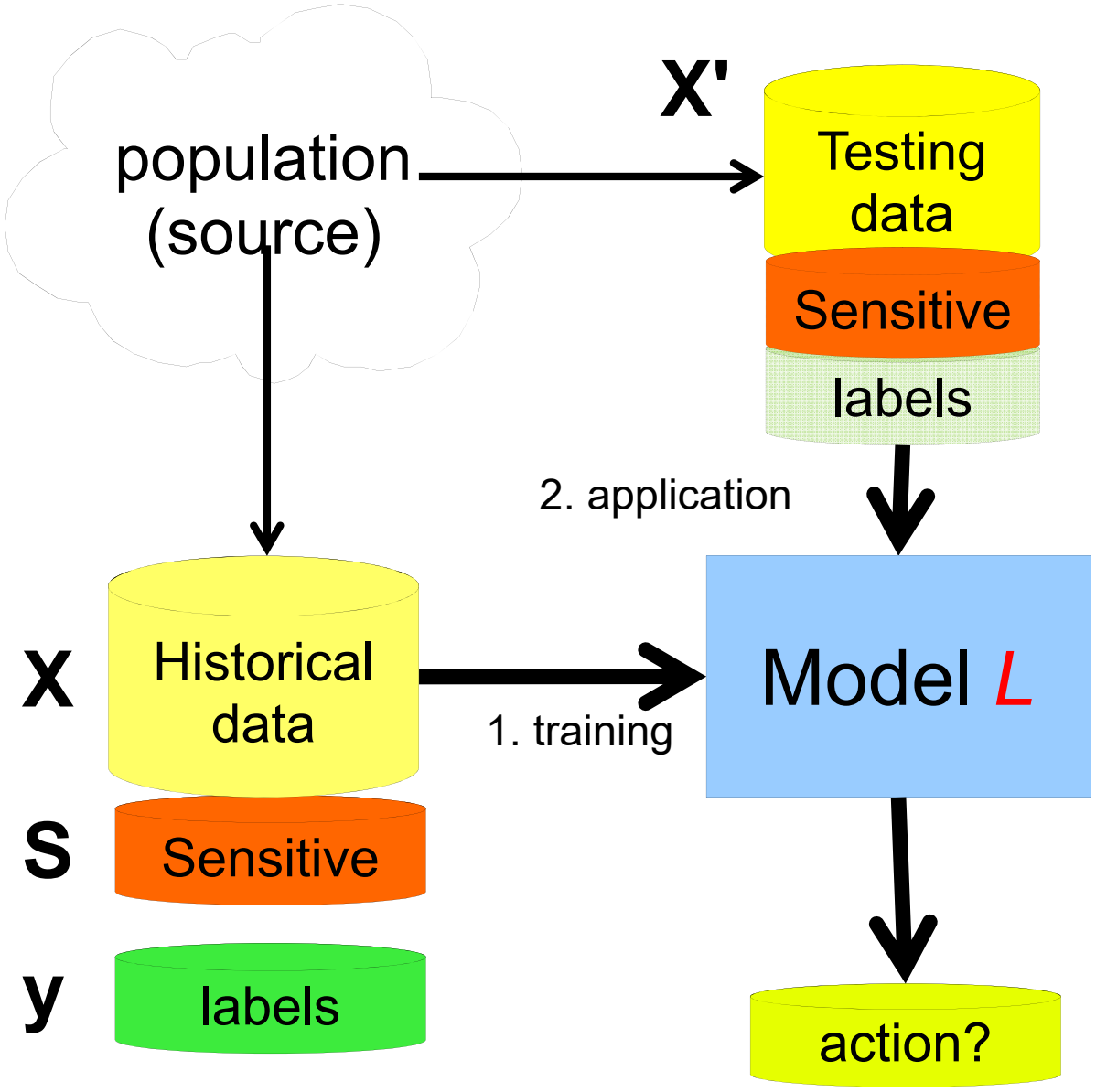
- models are as good as the underlying (biased) data they learn from,
- algorithms may reinforce human prejudices – while human decision makers may discriminate occasionally, algorithms would discriminate systematically

St George's Hospital: Student admission

- An algorithm to automate the 1st round of the Med School admissions process.
- used historical patterns in the characteristics of rejected candidates to filter out new candidates (whose profiles matched those of the least successful applicants).
- The admissions data showed bias against females and people with non-European-looking names

British Medical Journal 296(623), p.657, 1988

Predicting with Sensitive Attributes



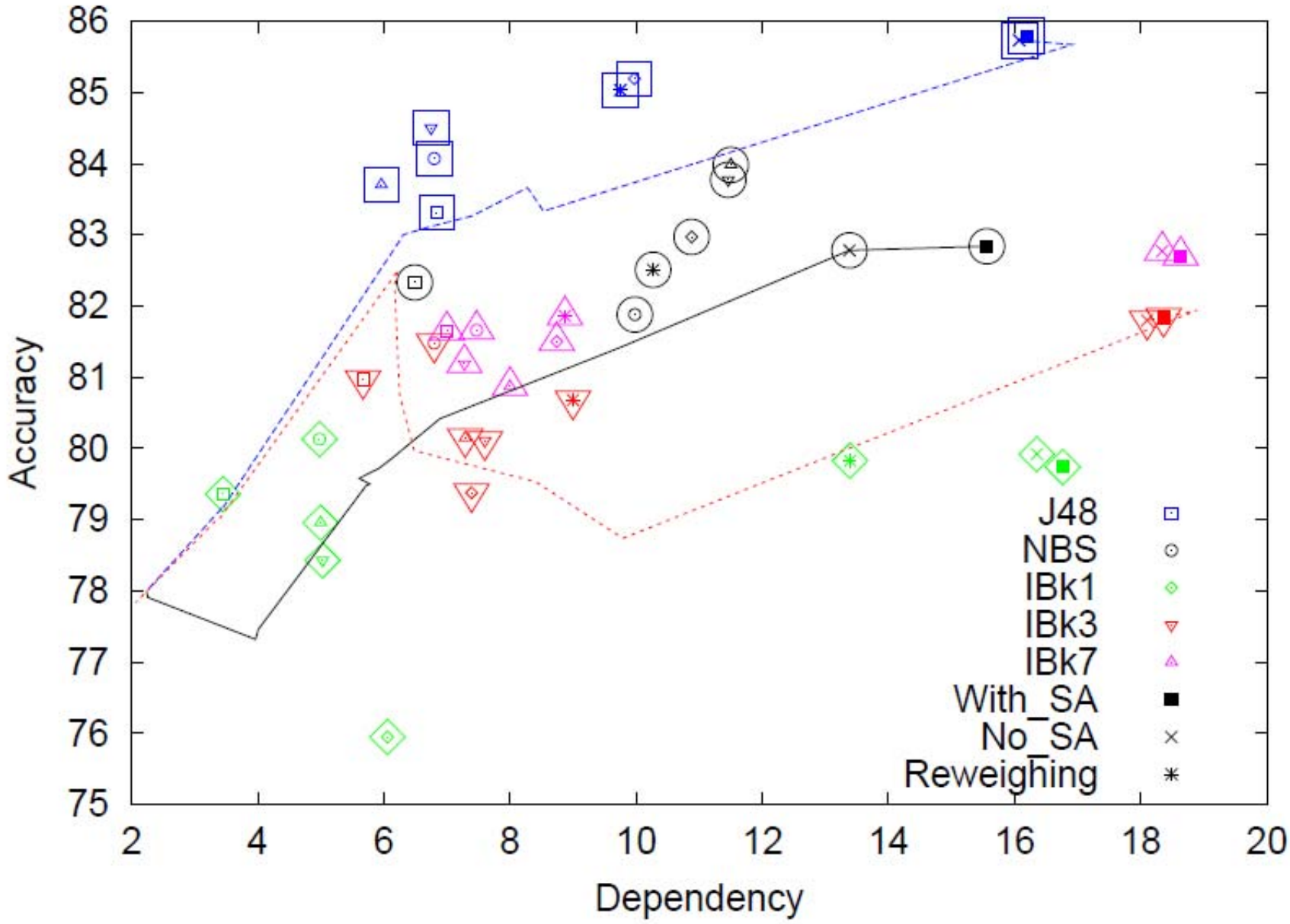
Training:

y = *L* (X, S)
minimizing prediction error & enforcing P(y=1|X,S='male') = P(y=1|X,S='female')

Application:
use *L*
for new data

y' = *L* (X',S')

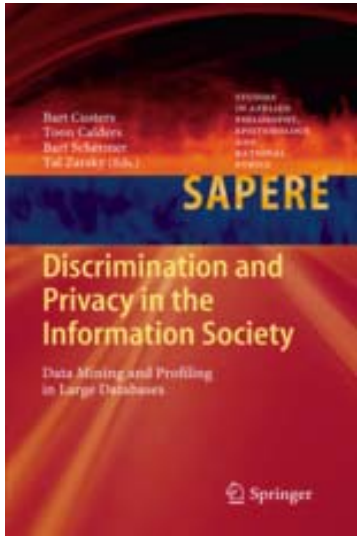
Accuracy – Discrimination Trade-off



Predicting with Sensitive Attributes

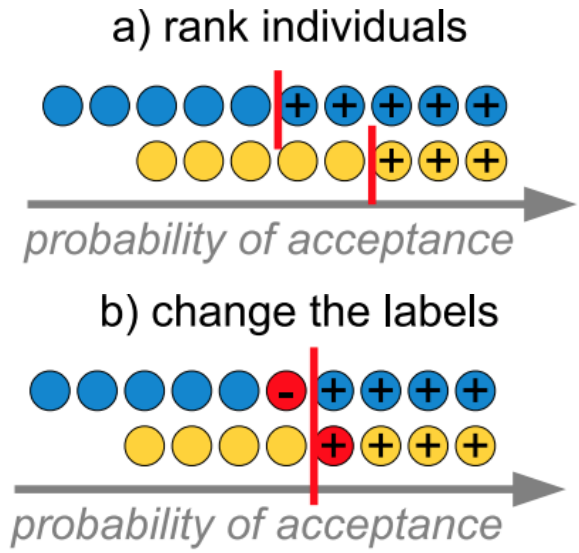
Paradox: we need to use personal data to control for unethical predictive analytics

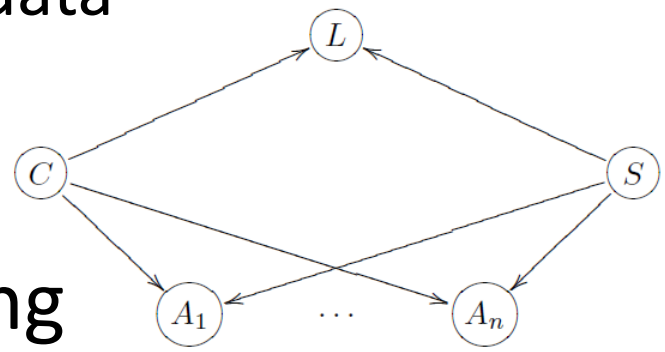
- "Fairness through awareness" Dwork et al.
- "It's Not Privacy, and it's Not Fair" Dwork & Mulligan
- "Discrimination and Privacy in the Information Society" Custers et al. (Eds)
  - Data mining for discrimination discovery
  - Explainable/conditional vs. unethical discrimination
  - Accuracy-discrimination tradeoff



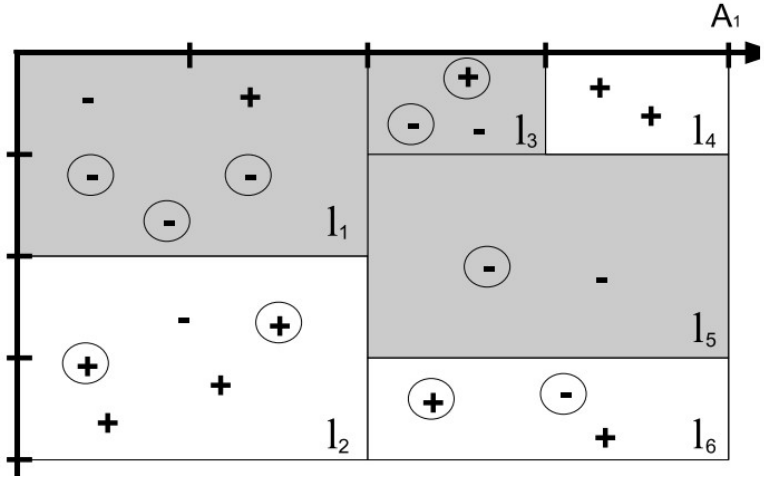
Discrimination-aware Solutions

- ~~Remove sensitive attributes?~~
- Preprocessing – "data massaging"
  - Modify input data (labels)
  - Resample input data





- Constraint learning
  - Algorithm-specific, e.g. Bayesian, decision trees, SVMs



- Postprocessing
  - Modify models
  - Modify outputs

Further reading (technical)

Bookchapter:

- Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. *Discrimination and Privacy in the Information Society*, Springer

Videotutorial and slides:

- http://videolectures.net/kdd2016_tutorial_algorithmic_bias/
- http://francescobonchi.com/KDD2016_Tutorial_Part1&2_web.pdf
- http://francescobonchi.com/KDD2016_Tutorial_Part3&4_web.pdf

Recent papers, events:

- http://www.fatml.org/
- http://www.responsibledatascience.org/
- http://www.zliobaite.com/non-discriminatory

Non-technical:

- https://mathbabe.org/author/mathbabe/

Thank you!

- Collaboration, proposals:

  m.pechenizkiy@tue.nl

- Staying connected:

  nl.linkedin.com/in/mpechen/