Understanding data in relation to social justice

Seminar at UCL Information Security Research Group

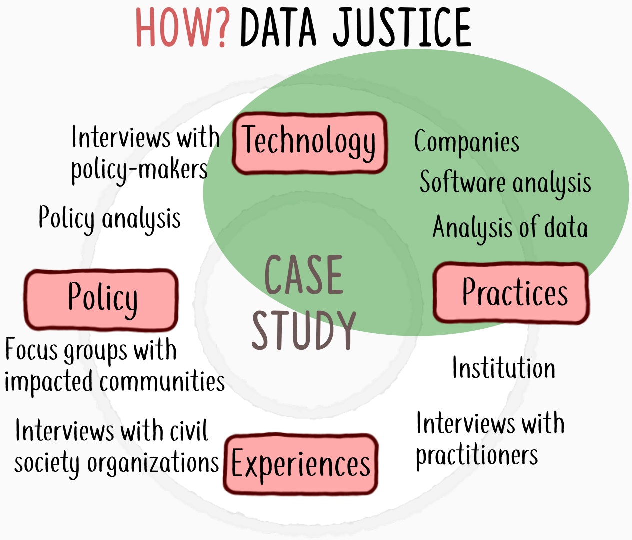
Javier Sánchez-Monedero
sanchez-monederoj at cardiff.ac.uk

August 2, 2018

Cardiff University, UK

Data Justice Lab

datajusticelab.org


HOW? DATA JUSTICE
Interviews with
policy-makers
Technology
Companies
Software analysis
Policy analysis
Analysis of data
CASE
STUDY
Policy
Practices
Focus groups with
impacted communities
Institution
Interviews with
practitioners
Interviews with civil
society organizations
Experiences

Quick summary of ML

Traditional programming

Explicit rules:

```
if email contains Viagra
  then mark is-spam;
if email contains ...;
if email contains ...;
```

Example from Jason's Machine Learning 101

Machine learning programs

Learn from examples:

```
try to classify some
emails;
change self to reduce
errors;
repeat;
```

...then use the model to label

Quick summary of ML

Traditional programming

Explicit rules:

```
if email contains Viagra
  then mark is-spam;
if email contains ...;
if email contains ...;
```

Example from Jason's Machine Learning 101

Machine learning programs

Learn from examples:

```
try to classify some
emails;
change self to reduce
errors;
repeat;
```

...then use the model to label

Since nobody is explicitly programming it, it is often assumed to be fair, non-discriminative, avoid human biases, etc.

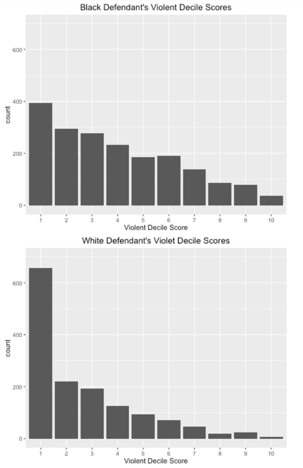
Machine learning tasks

- Prediction (classification/regression)
- Clustering, a.k.a. unsupervised machine learning
- Natural language processing
- Association rule learning
- Recommendation and search engines
- Ranking, sorting, etc.
- Some data visualization methods

How machines learn to discriminate

Some sources of discrimination (based on[Bar16]):

- Skewed sample



Source [JL16]

How machines learn to discriminate

Some sources of discrimination (based on[Bar16]):

- Skewed sample
- Tainted examples

Learn to predict hiring/loans/… decisions

How machines learn to discriminate

Some sources of discrimination (based on[Bar16]):

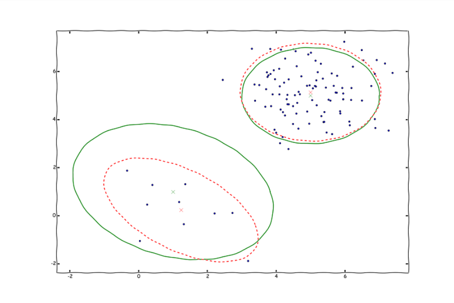
- Skewed sample
- Tainted examples
- Limited features

Are the features (equally) reliably collected for all the groups?

How machines learn to discriminate

Some sources of discrimination (based on[Bar16]):

- Skewed sample
- Tainted examples
- Limited features
- Sample size disparity



Source How big data is unfair

How machines learn to discriminate

Some sources of discrimination (based on[Bar16]):

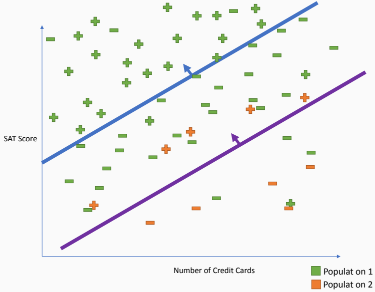
- Skewed sample
- Tainted examples
- Limited features
- Sample size disparity
- Proxy variables

{Postal code, salary} correlates to race

How machines learn to discriminate

Some sources of discrimination (based on[Bar16]):

- Skewed sample
- Tainted examples
- Limited features
- Sample size disparity
- Proxy variables
- Different features behaviour for each (sub)group



Source [Rot18]

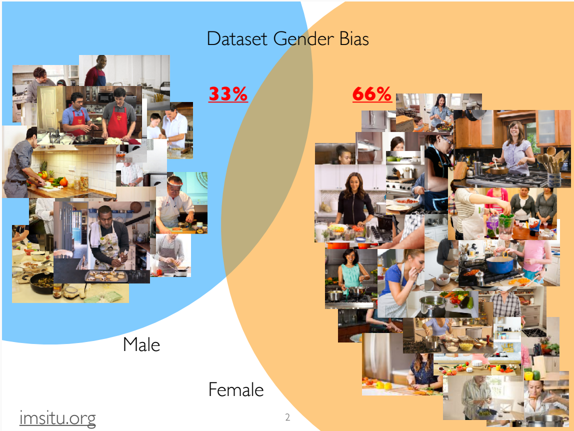
Bias reproduction and amplification

Source [LI16]

Feedback loops can reproduce and amplify discrimination [BH17, EFN$^+$17], example PredPol:

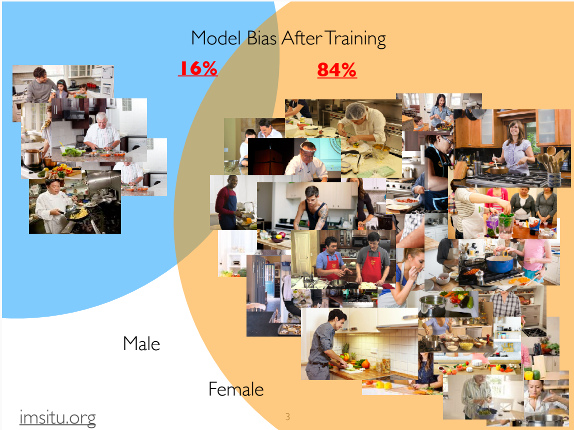
- Crime prediction in an area will send police resources to that area
- Discovered events will be added to the database
- It is less likely to observe events that contradicts predictions

Bias amplification i



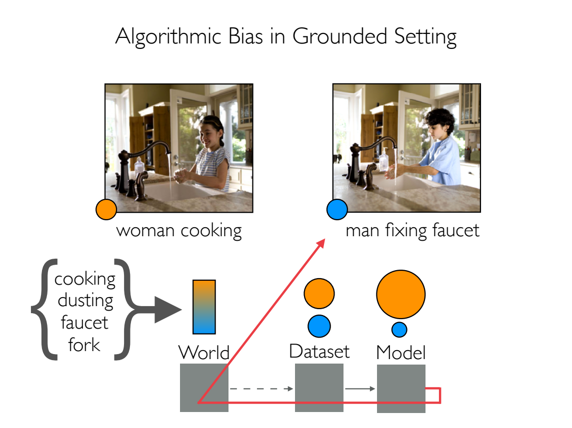
Source [ZWY+17]

Bias amplification ii



Source [ZWY+17]

Bias amplification iii



Source [ZWY+17]

How to measure discrimination?

How to evaluate *fairness*:

- Model/algorithm interpretability (what we mean with model interpretability? [Lip17])

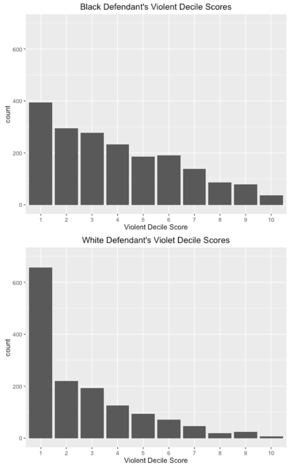
Risk of Violent Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	-0.729*** (0.127)
Age: Greater than 45	-1.742*** (0.184)
Age: Less than 25	3.146*** (0.115)
Black	0.659*** (0.108)
Asian	-0.985 (0.705)
Hispanic	-0.064 (0.191)
Native American	0.448 (1.035)
Other	-0.205 (0.225)
Number of Priors	0.138*** (0.012)
Misdemeanor	-0.164* (0.098)
Two Year Recidivism	0.934*** (0.115)
Constant	-2.243*** (0.113)
Observations	4,020
Akaike Inf. Crit.	3,022.779
*Note: *p<0.1; **p<0.05; ***p<0.01*	

How to measure discrimination?

How to evaluate *fairness*:

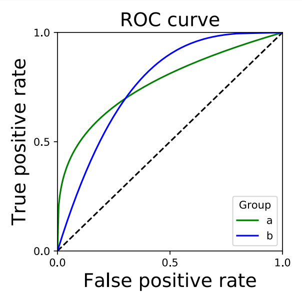
- Model/algorithm interpretability (what we mean with model interpretability? [Lip17])
- Dataset analysis



How to measure discrimination?

How to evaluate *fairness*:

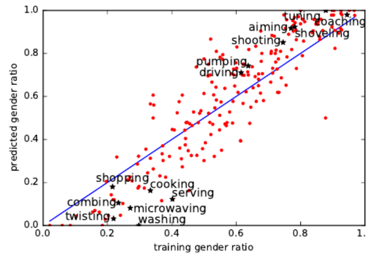
- Model/algorithm interpretability (what we mean with model interpretability? [Lip17])
- Dataset analysis
- Model performance w.r.t. subgroups and subgroups discovery ([ZN16])



How to measure discrimination?

How to evaluate *fairness*:

- Model/algorithm interpretability (what we mean with model interpretability? [Lip17])
- Dataset analysis
- Model performance w.r.t. subgroups and subgroups discovery ([ZN16])
- Model behaviour analysis



How to measure discrimination?

How to evaluate *fairness*:

- Model/algorithm interpretability (what we mean with model interpretability? [Lip17])
- Dataset analysis
- Model performance w.r.t. subgroups and subgroups discovery ([ZN16])
- Model behaviour analysis

but... we need a criteria
(Aaron Roth: "Weakly Meritocratic Fairness")

Discrimination is not a general concept

From the tutorial at NIPS [BH17], discrimination:

- It is domain specific and depends on potential impact on (marginalized) communities.
- It is feature(s) specific, with “socially salient qualities that have served as the basis for unjustified and systematically adverse treatment in the past”.

Formal setup in the community

Random variables in the same probability space ([BH17]):

- X features describing an individual
- A sensitive attribute (gender, race...)
- Y target variable
- $C = f(X, A)$ predictor estimating Y

Likelihood w.r.t. X and protected attribute A:

$$P(Y|X, = x, A = a).$$

Many FATML/FAT*ML works deal with C independence of A so that, for all groups in A (statistical parity):

$$P(C = c|X, = x, A = a) \approx P(C = c|X, = x, A = b)$$

For more conditions and definitions on fairness see [BH17] and [Rot18].

Formal setup in the community

Random variables in the same probability space ([BH17]):

- X features describing an individual
- A sensitive attribute (gender, race...) subgroup discovery!
- Y target variable
- $C = f(X, A)$ predictor estimating Y

Likelihood w.r.t. X and protected attribute A:

$$P(Y|X, = x, A = a).$$

Many FATML/FAT*ML works deal with C independence of A so that, for all groups in A (statistical parity):

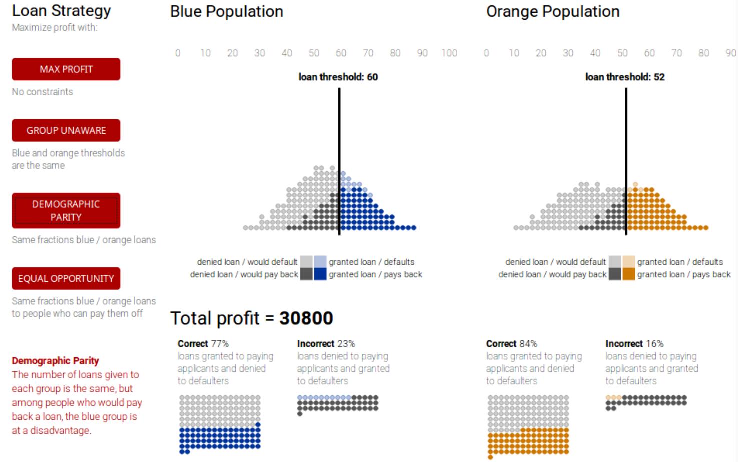
$$P(C = c|X, = x, A = a) \approx P(C = c|X, = x, A = b)$$

For more conditions and definitions on fairness see [BH17] and [Rot18].

Some fixings on classifiers

Pre-processing. E.g. feature adjustment
Post-processing. E.g. threshold calibration
Training algorithm. E.g. regularization term
Many more...

Threshold calibration



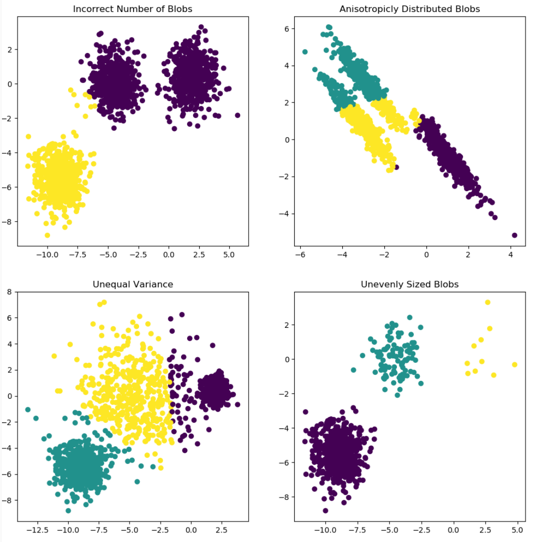
Source `http://research.google.com/bigpicture/attacking-discrimination-in-ml/`

Assumptions of methods

We should be aware of:

- Error function: What are we really optimising?
- Linearity assumption, e.g., Generalised Linear Models, K-means
- Independence of variables and variables interaction.
- ...

K-means assumptions



Source Documentation of scikit-learn

Further Questions

Everyone-is-right/wrong situations

Statistical learning will always tend to be conservative by definition

Is disparate treatment essential?

Fair facial recognition?

Non-binary group membership

...

Questions?

References i

- Barocas, Solon; Selbst, Andrew D, *Big Data's Disparate Impact*, California Law Review (2016) (en).
- Solon Barocas and Moritz Hardt, *Fairness in Machine Learning. NIPS 2017 Tutorial*, 2017.
- Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian, *Runaway Feedback Loops in Predictive Policing*, arXiv:1706.09847 [cs, stat] (2017) (en), arXiv: 1706.09847.
- Julia Angwin Jeff Larson, *How We Analyzed the COMPAS Recidivism Algorithm*, May 2016.
- Kristian Lum and William Isaac, *To predict and serve?*, Significance **13** (2016), no. 5, 14–19 (en).

- Zachary C. Lipton, *The Doctor Just Won't Accept That!*, arXiv:1711.08037 [stat] (2017) (en), arXiv: 1711.08037.
- Aaron Roth, *Course in (un)fairness in machine learning*, 2018.
- Zhe Zhang and Daniel B. Neill, *Identifying Significant Predictive Bias in Classifiers*, arXiv:1611.08292 [cs, stat] (2016) (en), arXiv: 1611.08292.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang, *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017), 2941–2951 (en), arXiv: 1707.09457.