Which paths to achieve fairness in algorithmic decisions?

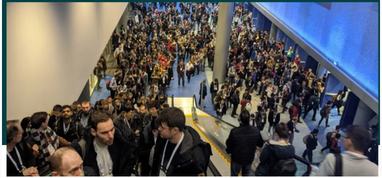
Online International Workshop, Université Paris-Dauphine

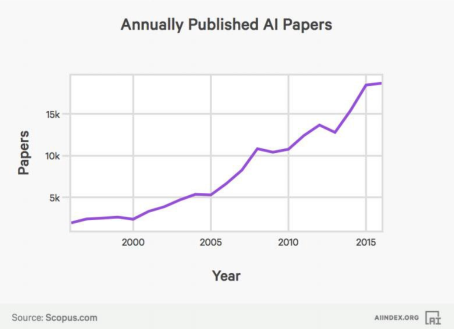
Aurélien Garivier

December 9th, 2021

Who is speaking ?

L'IA du Quotidien peut elle être Éthique ?
Loyauté des Algorithmes d'Apprentissage Automatique
P. Besse, C. Castets-Renard, A. Garivier, J-M. Loubes
Statistique et Société vol. 6 (3) Dec. 2018 pp.9-31


AI Weekly: NeurIPS proves machine learning at scale is hard



Outline

1. Success, Questions and Responsibility
2. On Biases
3. Formalizing Fairness
4. A Simple Example Expanded

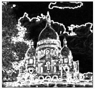
Success, Questions and Responsibility

Solving a Problem with a computer

Computer = machine able to combine arbitrarily a *small* set of elementary operations on some *data*

Problem : Input $\longrightarrow$  $\longrightarrow$ Output

Examples :

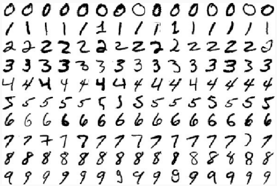
[3,2,5,1,4]	$\longrightarrow$	[1,2,3,4,5]
le petit chat	$\longrightarrow$	the little cat
	$\longrightarrow$	5
	$\longrightarrow$	

Solving a Problem with a computer

Classical way : write the **program** = sequence of elementary operations that leads from the input to the output

```
for i=1:n
    ib = i
    m = x[ib]
    for j=(i+1):n
        if x[j]>x[i]:
            ib = j
            m = x[j]
    c = x[i]
    x[i] = x[ib]
    x[ib] = c
```

Artificial intelligence : let the computer find the program itself !
→ meta-programming
Machine Learning : find the sequence using *examples* = data



Spectacular Success Stories

- Image recognition
Natural Language Processing
... and combination

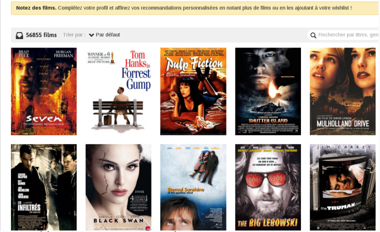
https://link.springer.com/article/10.1007

- Game solving (strategy)

AlphaGo

- Autonomous Vehicles

- Massive Recommender Systems : press, movies, ads, etc.



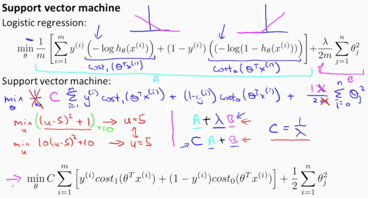
http://www.vodkaster.com/

How are these successes obtained ?

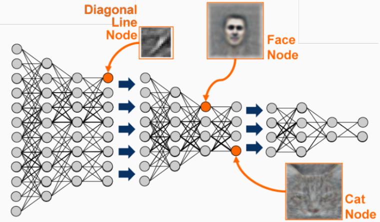
Abstraction : learn a mapping $\mathcal{X} \to \mathcal{Y}$ (mostly with vector-valued $\mathcal{X}$)

General abstract problem solved by several computationally intensive methods, including :

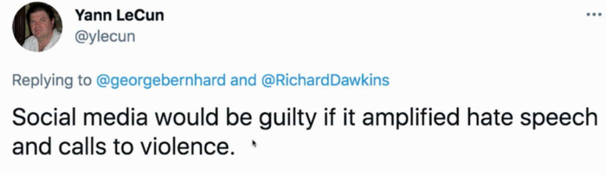
Statistical Learning



Neural networks



Difficulty : who is responsible ?



- **opacity** : not the mere formalization of an explicit process
- **dilution** : several actors involved (data / learning algorithms / choice of AI algorithm...)
- **liquid** : difficult to audit / inspect
- **impenetrability** : difficult to explain or even to interpret the results

→ more complicated than general algorithmic decision making

→ exciting on prototypes, frightening in real life

Growth crisis

Very powerful tools that are not under control

do we really want it ?

Facebook, Citing Societal Concerns, Plans to Shut Down Facial Recognition System

Saying it wants "to find the right balance" with the technology, the social network will delete the face scan data of more than one billion users.

162

https://www.nytimes.com/2021/11/02/technology/facebook-facial-recognition.html

Some causes of the crisis

- Bias in the data :
    - collected "as well as possible"
    - sometimes betraying participants' personal information
    - then considered as ground truth
- Bias in the scientific process :
    - abstraction = volunteer distance to applications, irresponsibility of the *model* (abstract world)
    - consensual technical goal = maximize average perf (dominant view, not robustness, reliability, etc.)
    - gamification (challenges with simple rules) but no certification
    - no consideration of the consequences (may augment inequalities, cf example with adult)

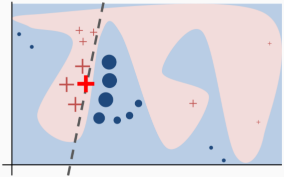
But the scientific ‘community looks for solutions !

Some causes of the crisis

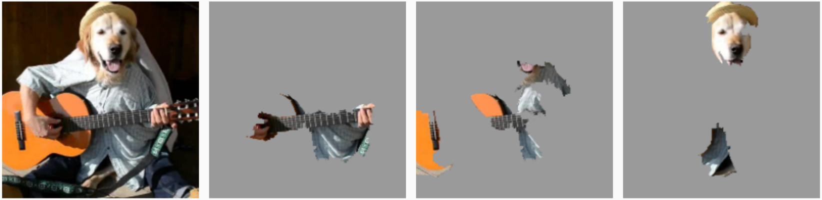
- Bias in the data :
  - collected "as well as possible" → define, detect, avoid/repair biases
  - sometimes betraying participants' personal information → differential privacy
  - then considered as ground truth → transfer learning
- Bias in the scientific process :
  - abstraction = volunteer distance to applications, irresponsibility of the *model* (abstract world) remains !
  - consensual technical goal = maximize average perf (dominant view, not robustness, reliability, etc.) → other risk measures (marginal)
  - gamification (challenges with simple rules) but no certification → XAI, research on mathematical control of the methods
  - no consideration of the consequences (may augment inequalities, cf example with adult) → fair learning

But the scientific ‘community looks for *technical* solutions !

Local Interpretable Model-Agnostic Explanations : LIME



Linear model with feature selection on local subset of data



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

Src : "Why Should I Trust You ?" Explaining the Predictions of Any Classifier, by Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin.

Local Interpretable Model-Agnostic Explanations : LIME

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Src : "Why Should I Trust You ?" Explaining the Predictions of Any Classifier, by Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin.

On Biases

Bias in the data

Src : An Introduction to Image Datasets, Malevé '19

Consequences

Underrepresentation of darker skin tones

Facial analysis datasets

LFW	77.5% male 83.5% white
IJB-A	79.6% lighter-skinned
Adience	86.2% lighter-skinned

Error Rate$_{(1\text{-}PPV)}$ By Female x Skin Type

	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
	1.7%	1.1%	3.3%	0%	23.2%	25.0%
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%

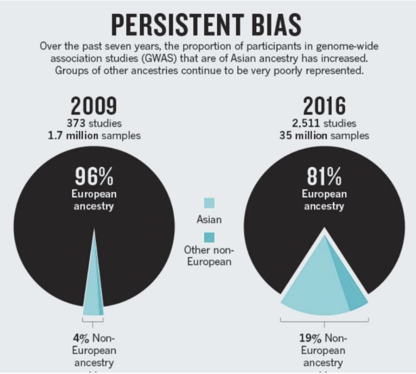
Buolamwini & Gebru FAT* 2018, Slides from Joy Buolamwini

Buolamwini & Gebru (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

This is not only about face recognition

- ...but also insurance, employment, credit risk assessment...

- ... personalized medicine : most study of pangenomic association were conducted on white/European population.
  $\Longrightarrow$ The estimated risk factors will possibly be different for patients with African or Asian origins !



Popejoy A., Fullerton S. (2016).
Genomics is failing on diversity, Nature 538

Formalizing Fairness

Detecting a bias

Detecting an individual discrimination : **Testing**

- Idea : modify just one protected feature of the individual and check if decision in changed
- Recognized by justice
- Discrimination for house rental, employment, entry in shops, insurance, etc.

Detecting a group discrimination : Discrimination Impact Assessment. Three measures :

- Disparate Impact (Civil Right Act 1971) : $DI = \dfrac{\mathbb{P}(\hat{h}_n(X) = 1|S = 0)}{\mathbb{P}(\hat{h}_n(X) = 1|S = 1)}$
- Cond. Error Rates : $\mathbb{P}(\hat{h}_n(X) \neq Y|S = 1) = \mathbb{P}(\hat{h}_n(X) \neq Y|S = 0)$
- Equality of odds : $\mathbb{P}(\hat{h}_n(X) = 1|S = 1)$ vs $\mathbb{P}(\hat{h}_n(X) = 1|S = 0)$

(Technical) Solution to the Fairness problem

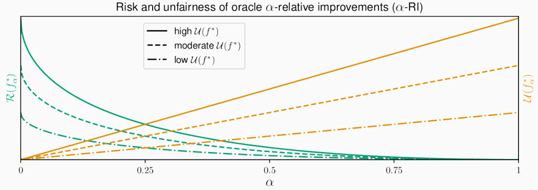
Projection to Fairness in Statistical Learning, *Le Gouic, Loubes & Rigollet* '20

A study of some trade-offs in statistical learning : online learning, generative models and fairness, *Schreuder* '21

Best fair predictor
= very inefficient

vs

Best unconstrained predictor
= very unfair



Modified Objective

Find the best predictor *among those with a Disparate Impact at most α% better than the best unconstrained predictor*

$\rightarrow$ Thanks to the theory of *optimal transport*, one shows that it takes (in some cases) an explicit form as a *interpolant between best unconstrained predictor and best perfectly fair predictor*

A Simple Example Expanded

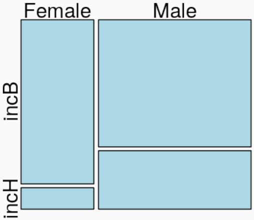
An Example in more Detail

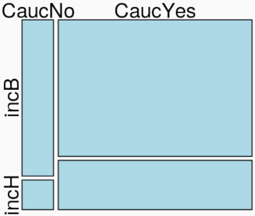
The following example is based on a Jupyter Notebook by **Philippe Besse** (INSA Toulouse) freely available (in R and python) on `https://github.com/wikistat`

Adult Census Dataset of UCI

- 48842 US citizens (1994)
- 14 features :
  - $Y =$ income threshold ($50k)
  - **age** : continuous.
  - **workclass** : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
  - **fnlwgt** : continuous.
  - **education** : Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
  - **education-num** : continuous.
  - **marital-status** : Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
  - **occupation** : Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
  - **relationship** : Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

Obvious Social Bias






```
Confidence interval for the DI
(by delta method)

round(dispImp(datBas[,"sex"],
datBas[,"income"]),3)

0.349 0.367 0.384
```

```
Confidence interval for the DI
(delta method)

round(dispImp(datBas$origEthn,
datBas$income),3)

0.566 0.601 0.637
```

Logistic Regression augments the bias !

```
log.lm=glm(income~.,data=datApp,family=binomial)

# significativity of the parameters
anova(log.lm,test="Chisq")

Df       Deviance          Resid. Df        Resid. Dev       Pr(>Chi)
NULL     NA       NA       35771    40371,72         NA
age      1        1927,29010        35770    38444,43         0,000000e+00
educNum  1        4289,41877        35769    34155,01         0,000000e+00
mariStat          3        6318,12804        35766    27836,88         0,000000e+00
occup    6        812,50516         35760    27024,38         3,058070e-172
origEthn          1        17,04639          35759    27007,33         3,647759e-05
sex      1        50,49872          35758    26956,83         1,192428e-12
hoursWeek         1        402,82271         35757    26554,01         1,338050e-89
LcapitalGain      1        1252,69526        35756    25301,31         2,154522e-274
LcapitalLoss      1        310,38258         35755    24990,93         1,802529e-69
child    1        87,72437          35754    24903,21         7,524154e-21

# Prevision
pred.log=predict(log.lm,newdata=daTest,type="response")
# Confusion matrix
confMat=table(pred.log>0.5,daTest$income)

incB     incH
FALSE    6190     899
TRUE     556      1298

tauxErr(confMat): 16,27

round(dispImp(daTest[,"sex"],Yhat),3) : 0.212 0.248 0.283

# Overall Accuracy Equality?
apply(table(pred.log<0.5,daTest$income,daTest$sex),3,tauxErr)

Female 91.81     Male 79.7
```

What about Random Forest ?

Random Forest improves significantly the predicition quality...

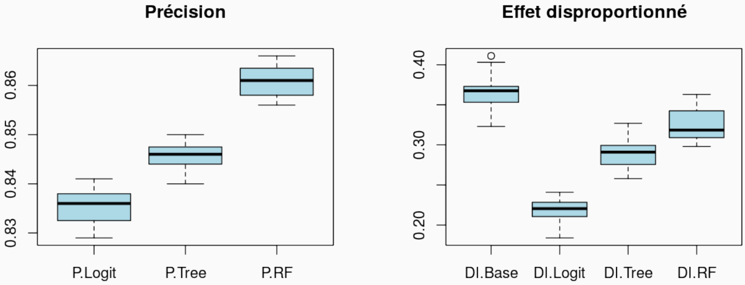
```
rf.mod=randomForest(income~., data=datApp)
pred.rf=predict(rf.mod, newdata=daTest, type="response")
confMat=table(pred.rf, daTest$income)
confMat
tauxErr(confMat)
```

```
pred.rf  incB   incH
incB     6301    795
incH     445    1402

13,87
```

```
round(dispImp(daTest[,"sex"], pred.rf),3)
0.329  0.375  0.42
```

... without augmenting the bias (here).

Summary of the results by algorithm



$\Longrightarrow$ Random Forest is here both more performant and less discriminative (BUT not interpretable)

$\Longrightarrow$ This is not a general rule ! It depends on the dataset

$\Longrightarrow$ A serious learning should consider the different algorithms, and include a discussion on the discriminative effects

Individual Biases : Testing

Are the predictions changed if the value of variable "sex" is switched ?

```
daTest2=daTest
# Changement de genre
daTest2$sex=as.factor(ifelse(daTest$sex=="Male","Female","Male"))
# Prevision du "nouvel" echantillon test
pred2.log=predict(log.lm,daTest2,type="response")
table(pred.log<0.5,pred2.log<0.5,daTest$sex)
```

Female

	FALSE	TRUE
FALSE	195	0
TRUE	23	2679

Male

	FALSE	TRUE
FALSE	1489	155
TRUE	0	4402

➔ 178 have a different prediction, in the expected direction.

Avoid Issues with Testing

Easy : use maximal prediction of all modalilities of the protected variable

```
fairPredictGenre=ifelse(pred.log<pred2.log, pred2.log, pred.log)
confMat=table(fairPredictGenre >0.5,daTest$income)
confMat;tauxErr(confMat)

incB     incH
FALSE    6145     936
TRUE     535      1327

16.45


round(dispImp(daTest$sex,as.factor(fairPredictGenre >0.5)),3)
0.24  0.277  0.314

# recall:
round(dispImp(daTest$sex,as.factor(pred.log >0.5)),3)
0.212  0.248  0.283
```

➔ No influence on the prediction quality

➔ Small bias reduction, but does not remove group over-discrimination !

Naive approach : suppress the protected variable

```
# estimation without the variable "sex"
log_g.lm=glm(income~.,data=datApp[,-6],family=binomial)

# Prevision
pred_g.log=predict(log_g.lm,newdata=daTest[,-8],type="response")
# Confusion Matrix
confMat=table(pred_g.log>0.5,daTest$income)
confMat


      incB  incH
FALSE 6157   953
TRUE   523  1310

tauxErr(confMat)

16.5



Yhat_g=as.factor(pred_g.log>0.5)
round(dispImp(daTest[,"sex"],Yhat_g),3)

0.232 0.269 0.305
```

$\Longrightarrow$ the quality of prediction is not deteriorated, but the bias augmentation remains the same !

Adapting the threshold to each class

```
Yhat_cs=as.factor(ifelse(daTest$sex=="Female",pred.log>0.4,pred.log>0.5))
round(dispImp(daTest[,"sex"],Yhat_cs),3)
tauxErr(table(Yhat_cs,daTest$income))

0.293 0.334 0.375

16.55

# Stronger correction forcing  the DI to be at least 0.8:

Yhat_cs=as.factor(ifelse(daTest$sex=="Female",pred.log>0.15,pred.log>0.5))
round(dispImp(daTest[,"sex"],Yhat_cs),3)
tauxErr(table(Yhat_cs,daTest$income))

0.796 0.863 0.93

18.57
```

$\Longrightarrow$ the prediction performance is significantly deteriorated

$\Longrightarrow$ this kind of affirmative action is a questionable choice

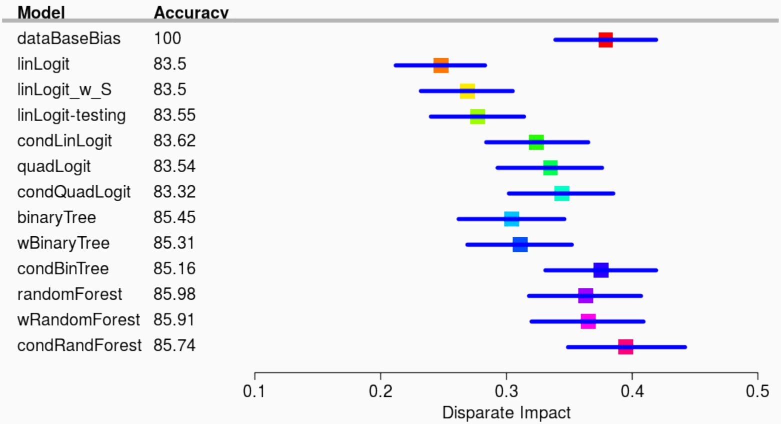
Building one classifier per class

Logistic regression $\rightarrow$ consider the interactions of the protected variable with the others

```
yHat=predict(reg.log, newdata=daTest, type="response")
yHatF=predict(reg.logF, newdata=daTestF, type="response")
yHatM=predict(reg.logM, newdata=daTestM, type="response")

yHatFM=c(yHatF, yHatM);  daTestFM=rbind(daTestF, daTestM)

# Cumulated errors
table(yHatFM>0.5, daTestFM$income)
incB    incH
FALSE   6150    935
TRUE    530     1328

table(yHat>0.5, daTest$income)
incB    incH
FALSE   6154    950
TRUE    526     1313

tauxErr(table(yHatFM>0.5, daTestFM$income))
16.38

tauxErr(table(yHat>0.5, daTest$income))
16.5

# Bias with an without class separation
round(dispImp(daTestFM[,"sex"], as.factor(yHatFM>0.5)),3)
0.284 0.324 0.365

round(dispImp(daTest[,"sex"], as.factor(yHat>0.5)),3)
0.212 0.248 0.283
```

$\Longrightarrow$ it reduces the bias

Comparison of several classifiers



Summary

- Automatic classification can *augment* the social bias
- All algorithms are not equivalent
- Linear classifiers should be particularly watched
- Random Forest can (at least sometimes) be less discriminative
- The bias augmentation diminishes with the consideration of variable interactions
- Removing the protected variable from the analysis is not sufficient
- Fitting different models on the different classes is in general a quick and simple way to avoid bias augmentation...
- ... if the protected variable is observed !