

Automatic Decision by Machine Learning and Fairness

CFIES 2019

Aurélien Garivier

Strasbourg



Table of contents

1. Automatic Decision Systems based on Machine Learning
2. On Fairness and Machine Learning
3. Understanding the Algorithms' Predictions?

Automatic Decision Systems based on Machine Learning

Automatic Decision Systems based on Machine Learning

Data and Learning Algorithms

Classification Framework

On Fairness and Machine Learning

What is fair?

How to fix the problem?

Understanding the Algorithms' Predictions?

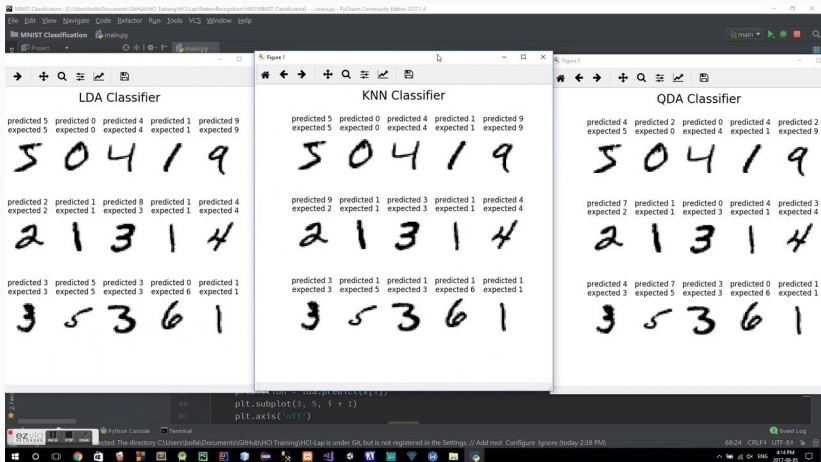
What is Machine Learning?

- ▶ Algorithms operate by building a model from **example** inputs in order to make data-driven **predictions or decisions**...
- ▶ ...rather than following strictly static program instructions: useful when designing and programming explicit algorithms is unfeasible or poorly efficient.

Within Artificial Intelligence

- ▶ evolved from the study of pattern recognition and computational learning theory in artificial intelligence.
- ▶ AI: emulate cognitive capabilities of humans (big data: humans learn from abundant and diverse sources of data).
- ▶ a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".

Example: MNIST dataset



Definition

Arthur Samuel (1959)

Field of study that gives computers the ability to learn without being explicitly programmed

Tom M. Mitchell (1997)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P , improves with experience E .

Machine Learning: Typical Problems

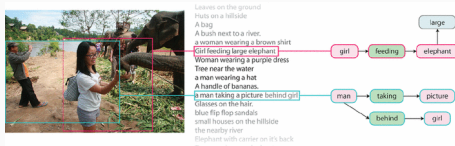
- ▶ spam filtering, text classification
- ▶ optical character recognition (OCR)
- ▶ search engines
- ▶ recommendation platforms
- ▶ speech recognition software
- ▶ computer vision
- ▶ bio-informatics, DNA analysis, medicine
- ▶ ...

For each of this task, it is possible but very inefficient to write an explicit program reaching the prescribed goal.

It proves much more succesful to have a machine infer what the good decision rules are.

Spectacular Success Stories

- Image recognition
- Natural Language Processing
- ... and combination

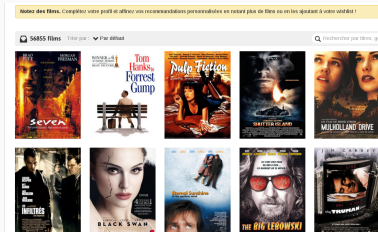


<https://link.springer.com/article/10.1007>

- Game solving (strategy)
- Autonomous Vehicles



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



<http://www.vodkaster.com/>

Automatic Decision Systems based on Machine Learning

Data and Learning Algorithms

Classification Framework

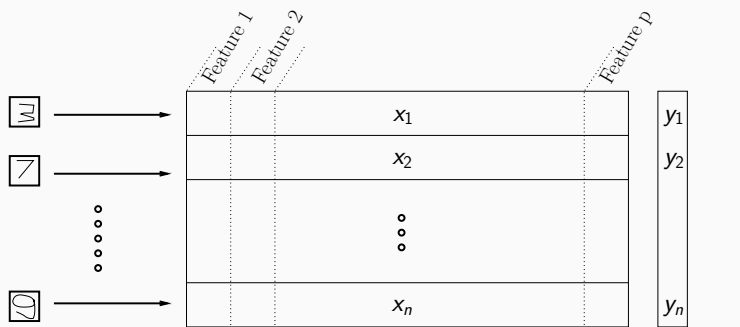
On Fairness and Machine Learning

What is fair?

How to fix the problem?

Understanding the Algorithms' Predictions?

What is a classifier?



$$X \in \mathcal{M}_{n,p}(\mathbb{R})$$

$$Y \in \mathcal{Y}^n$$

Data: n -by- p matrix X

- ▶ n examples = points of observations
- ▶ p features = characteristics measured for each example

Classifier \mathcal{A}_n

$$\begin{array}{c} \downarrow \\ h_n : \mathcal{X} \rightarrow \mathcal{Y} \\ \boxed{6} \mapsto 6 \end{array}$$

- ▶ Inside R: package datasets
- ▶ Inside python/scikitlearn: package `sklearn.datasets`
- ▶ UCI Machine Learning Repository



- ▶ Challenges: Kaggle, etc.

Statistical Learning Hypothesis

Assumption

- ▶ The examples $(X_i, Y_i)_{1 \leq i \leq n}$ are iid samples of an unknown joint distribution \mathcal{D} ;
- ▶ The points to classify later are also independent draws of the *same* distribution \mathcal{D} .

Hence, for every *decision rule* $h : \mathcal{X} \rightarrow \mathcal{Y}$ we can define the *risk*

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(X,Y) \sim \mathcal{D}}(h(X) \neq Y) = \mathcal{D}\left(\{(x,y) : h(x) \neq y\}\right).$$

The goal of the learning algorithm is to *minimize the expected risk*:

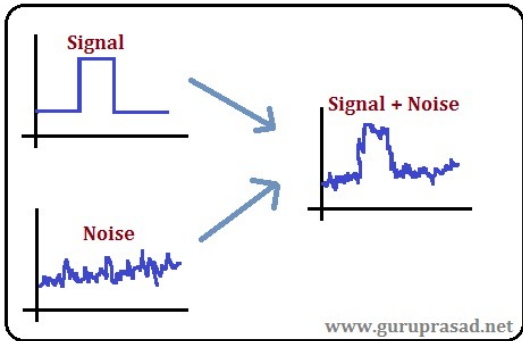
$$R_n(\mathcal{A}_n) = \mathbb{E}_{\mathcal{D}^{\otimes n}} \left[L_{\mathcal{D}} \left(\underbrace{\mathcal{A}_n((X_1, Y_1), \dots, (X_n, Y_n))}_{\hat{h}_n} \right) \right]$$

for every distribution \mathcal{D} , using only the examples.

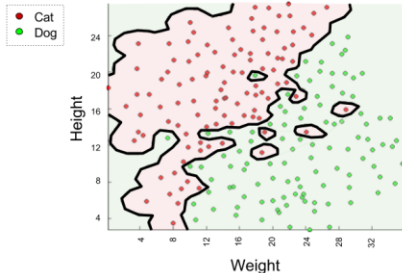
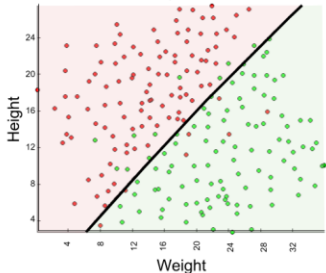
Signal and Noise

new york times bestseller
noise and the noi
the signal and th
and the noise and
the noise and th
why so many noi
predictions fail—
but some don't th
and the noise and
nate silver the nc

"Could turn out to be one of the more momentous books of the decade." —*The New York Times Book Review*



www.guruprasad.net



Example: Character Recognition

Domain set \mathcal{X}	28 × 28 images
Label set \mathcal{Y}	{0, 1, ..., 9}
Joint distribution \mathcal{D}	?
Prediction function $h \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$	
Risk $R(h) = P_{\mathcal{D}}(h(X) \neq Y)$	
Sample $S_n = \{(X_i, Y_i)\}_{i=1..n}$	MNIST dataset
Empirical risk $L_{S_n}(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{h(X_i) \neq Y_i\}$	
Learning algorithm $\mathcal{A} = (\mathcal{A}_n)_n, \mathcal{A}_n : (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{H}$	neural nets, boosting...
Expected risk $R_n(\mathcal{A}) = \mathbb{E}_n [L_{\mathcal{D}}(\mathcal{A}_n(S_n))]$	

Performance Limit: Bayes Classifier

Consider binary classification $\mathcal{Y} = \{0, 1\}$, $\eta(x) := \mathcal{D}(Y = 1|X = x)$.

Theorem

The Bayes classifier is defined by

$$h^*(x) = \mathbb{1}\{\eta(x) \geq 1/2\} = \mathbb{1}\{\eta(x) \geq 1 - \eta(x)\} = \mathbb{1}\{2\eta(x) - 1 \geq 0\}.$$

For every classifier $h : \mathcal{X} \rightarrow \mathcal{Y} = \{0, 1\}$,

$$L_{\mathcal{D}}(h) \geq L_{\mathcal{D}}(h^*) = \mathbb{E}\left[\min(\eta(X), 1 - \eta(X))\right].$$

The Bayes risk $L_{\mathcal{D}}^ = L_{\mathcal{D}}(h^*)$ is called the **noise** of the problem.*

More precisely,

$$L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h^*) = \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}\{h(X) \neq h^*(X)\}\right].$$

Extends to $|\mathcal{Y}| > 2$.

Benchmarks, Challenges: Competition to reach the Bayes Risk

The screenshot shows the Kaggle website's 'Competitions' section. At the top, there are navigation links for 'Search kaggle', 'Competitions', 'Datasets', 'Kernels', 'Discussion', 'Jobs', and a 'Sign In' button. Below the navigation is a 'Competitions' header with 'Learn more' and 'InClass' buttons. The main content area is titled '17 Active Competitions' and lists several challenges with their respective prizes and team counts:

Competition Name	Prize	Teams
2018 Data Science Bowl	\$100,000	2,992 teams
Google Cloud & NCAA® ML Competition 2018-Men's	\$50,000	934 teams
Google Cloud & NCAA® ML Competition 2018-Women's	\$50,000	505 teams
TalkingData AdTracking Fraud Detection Challenge	\$25,000	1,532 teams
iMaterialist Challenge (Furniture) at FGVC5	\$2,500	94 teams
Google Landmark Retrieval Challenge	\$2,500	91 teams

The screenshot shows the UCI Machine Learning Repository website. The header includes the UCI logo, a search bar, and navigation links for 'About', 'Classes/Policy', 'Dataset & Data Set', and 'Contact'. Below the header, there is a 'Welcome to the UC Irvine Machine Learning Repository!' message and a paragraph of text explaining the repository's mission and providing links for more information.

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 429 data sets as a service to the machine learning community. You may view all data sets through our searchable interface. Our data sets are still available, for those who prefer the old format. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to contact the [Repository Manager](#). We have also set up a [wiki page](#) for the Repository.

Supported by: in Collaboration With:

The screenshot shows the 'Latest News' and 'Newest Data Sets' sections of the UCI Machine Learning Repository website. The 'Latest News' section lists recent updates, including the addition of new data sets and the release of new versions of existing data sets. The 'Newest Data Sets' section lists the most recent data sets added to the repository, including 'English-Canadian Mailbox Messages', 'UCI-MM GPU kernel performance', 'chess', 'Newspaper Popularity in Multiple Social Media Platforms', and 'Residential Building Data Set'. The 'Most Popular Data Sets (hits since 2007)' section lists the most popular data sets, including '150', 'Adult', 'Wave', 'Car Evaluation', and 'Breast Cancer Wisconsin (Diagnostic)'.

- ▶ Logistic Regression
- ▶ Support Vector Machines (with kernels)
- ▶ Classification and Regression Trees
- ▶ Random Forests
- ▶ Boosting
- ▶ Neural Networks and Deep Learning...

But do these classifiers really what we want?

- ▶ most effort tends to minimize the *average risk* for *entirely autonomous* systems
- ▶ is it always what we want?
- ▶ do we have guarantees on the *reliability?* on the *robustness?* are automatic decisions *certifiable?*
- ▶ is it *legal* to use them, what are the limitations?
- ▶ are the automatic decisions *explainable?* *interpretable?*
- ▶ are they *acceptable* by society?

On Fairness and Machine Learning

Automatic Decision Systems based on Machine Learning

Data and Learning Algorithms

Classification Framework

On Fairness and Machine Learning

What is fair?

How to fix the problem?

Understanding the Algorithms' Predictions?

CE : Ethic guidelines for trustworthy AI (2019)

The principle of fairness

The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: *ensuring equal and just distribution of both benefits and costs*, and **ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.**

If unfair biases can be avoided, AI systems could even increase societal fairness.[. . .] Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models.

Diversity, non-discrimination and fairness. . . Identifiable and discriminatory bias should be removed in the collection phase where possible. [. . .]

Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system ?

Article 225-1 du code pénal

- Constitue une discrimination toute distinction opérée entre les personnes physiques sur le fondement de leur origine, de leur sexe, de leur situation de famille, de leur grossesse, de leur apparence physique, de la particulière vulnérabilité résultant de leur situation économique, apparente ou connue de son auteur, de leur patronyme, de leur lieu de résidence, de leur état de santé, de leur perte d'autonomie, de leur handicap, de leurs caractéristiques génétiques, de leurs mœurs, de leur orientation sexuelle, de leur identité de genre, de leur âge, de leurs opinions politiques, de leurs activités syndicales, de leur capacité à s'exprimer dans une langue autre que le français, de leur appartenance ou de leur non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée.
- Constitue une discrimination indirecte une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés.

2 sources of biases:

- ▶ Social Bias (in the data)
- ▶ Algorithmic unfairness (ML creates or reinforces bias)

+ risk of "self-fulfilling prophecy"

Ex: higher credit insurance cost augment the risk of default

⇒ need concepts and tools to

- ▶ **identify** and (if possible)
- ▶ **correct** the biases.

Une étude démontre les biais de la reconnaissance faciale, plus efficace sur les hommes blancs

Lorsqu'il s'agit de reconnaître le genre d'un homme blanc, des logiciels affichent un taux de réussite de 99 %. La tâche se complique lorsque la peau d'une personne est plus foncée, ou s'il s'agit d'une femme.

<http://www.lemonde.fr/pixels/>

Joy Buolamwini (MIT) has studied three face recognition software (by IBM, MICROSOFT and FACE++) on 1 270 official portraits of politicians from Rwanda, Senegal, South Africa, Finland and Sweden, asking to **predict their gender**.

Buolamwini Study

Average results are good: 93,7% success rate for MICROSOFT, 90% for FACE++, and 87,9% pour IBM.

BUT

- ▶ Less successful for women than for men: for example, FACE++ classifies correctly 99,3% of the men but only 78,7% of the women.
- ▶ Less successful for dark skins than for pale skins: for the IBM softwares, success rates are 77;6% versus 95%.
- ▶ 93,6% of the mistakes of the Microsoft software were on dark skins, and 95,9% of the mistakes of Face ++ were on women!

Why? Bias in the data!

"Men with white skin are over-represented, and in fact white skins in general are."

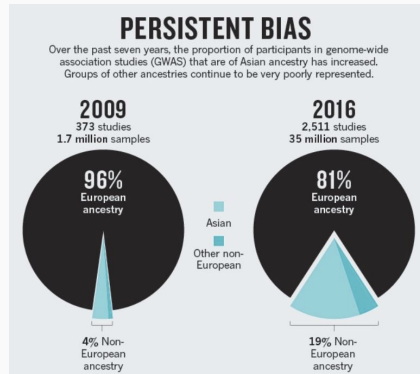
http://www.lemonde.fr/pixels/article/2018/02/12/une-etude-demontre-les-biais-de-la-reconnaissance-faciale-plus-efficace-sur-les-hommes-blancs_5255663_4408996.html#EZuQd0CJvJ3kYTiL.99

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.

⇒ 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

Detecting a bias

Detecting an individual discrimination: **Testing**

- ▶ Idea: modify just one protected feature of the individual and check if decision is 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): $DI = \frac{\mathbb{P}(\hat{h}_n(X) = 1|S = 0)}{\mathbb{P}(\hat{h}_n(X) = 1|S = 1)}$
- ▶ 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 $\mathbb{P}(\hat{h}_n(X) = 1|S = 0)$

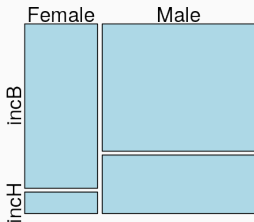
An Example in more Detail

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.
 - > **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
 - > **sex**: Female, Male.
 - > **capital-gain**: continuous.
 - > **capital-loss**: continuous. hours-per-week: continuous.
 - > **native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, ...
- ▶ We ignore **fnlwgt** and **nativ-country** (pretty redundant with race), **relationship** → child, **race** → CaucYes / CaucNo

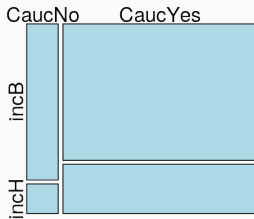
Obvious Social Bias



Confidence interval for the DI
(by delta method)

```
round(displmp(datBas["sex"],  
             datBas["income"]),3)
```

0.349 0.367 0.384



Confidence interval for the DI
(delta method)

```
round(displmp(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	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(displmp(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	Male
1	91.81	79.7

What about Random Forest?

Random Forest improves significantly the prediction 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)
```

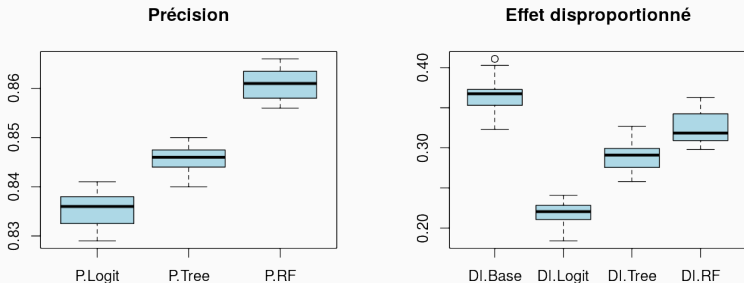
```
pred.rf  incB    incH
incB     6301    795
incH     445    1402
```

```
13,87
```

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

... without augmenting the bias (here).

Summary of the results by algorithm



⇒ Random Forest is here both more performant and less discriminative (BUT not interpretable)

⇒ This is not a general rule! It depends on the dataset

⇒ 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"))
# Pr vision du "nouvel" chantillon test
pred2.log=predict(log.lm, daTest2, type="response")
table(pred.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.

Automatic Decision Systems based on Machine Learning

Data and Learning Algorithms

Classification Framework

On Fairness and Machine Learning

What is fair?

How to fix the problem?

Understanding the Algorithms' Predictions?

Avoid Issues with Testing

Easy: use maximal prediction of all modalities 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(displmp(daTest$sex , as.factor(fairPredictGenre >0.5)),3)  
0.24 0.277 0.314
```

```
# recall :  
round(displmp(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(disImp(daTest[, "sex"], Yhat_g), 3)

0.232 0.269 0.305
```

⇒ 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$sex=="Female" , pred.log > 0.4, pred.log > 0.5))  
round(displmp(daTest[, "sex"], Yhat_cs), 3)  
tauxErr(table(Yhat_cs, daTest$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$sex=="Female" , pred.log > 0.15, pred.log > 0.5))  
round(displmp(daTest[, "sex"], Yhat_cs), 3)  
tauxErr(table(Yhat_cs, daTest$income))
```

```
0.796 0.863 0.93
```

```
18.57
```

⇒ the prediction performance is significantly deteriorated

⇒ this kind of affirmative action is a questionable choice

Building one classifier per class

Logistic regression → 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(displmp(daTestFM[, "sex"], as.factor(yHatFM>0.5)),3)
```

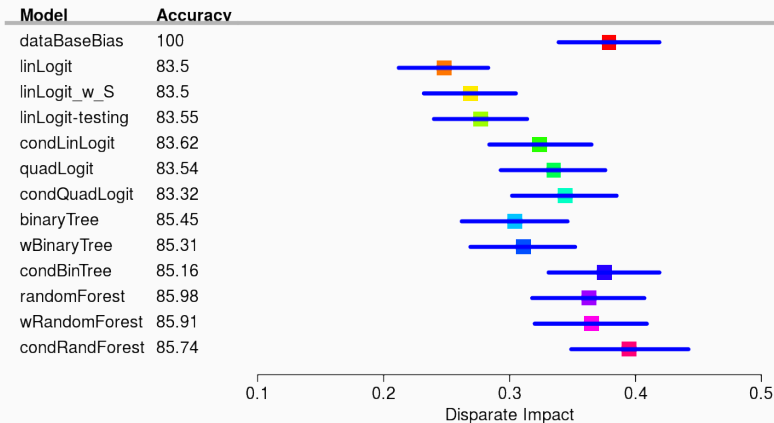
```
0.284 0.324 0.365
```

```
round(displmp(daTest[, "sex"], as.factor(yHat>0.5)),3)
```

```
0.212 0.248 0.283
```

⇒ 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!

Understanding the Algorithms' Predictions?

Article 22 (RGPD) : Décision individuelle automatisée, y compris le profilage

1. La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire.
2. Le paragraphe 1 ne s'applique pas lorsque la décision :
 - a. est nécessaire à la conclusion ou à l'exécution d'un contrat entre la personne concernée et un responsable du traitement ;
 - b. est autorisée par le droit de l'Union ou le droit de l'état membre auquel le responsable du traitement est soumis et qui prévoit également des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ; ou
 - ▶ c.] est fondée sur le consentement explicite de la personne concernée.
3. Dans les cas visés au paragraphe 2, points a) et c), le responsable du traitement met en œuvre des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée, au moins du droit de la personne concernée d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision.
4. Les décisions visées au paragraphe 2 ne peuvent être fondées sur les catégories particulières de données à caractère personnel (cf. article 9 : biométriques, génétiques, de santé, ethniques ; orientation politique, syndicale, sexuelle, religieuse, philosophique) sous réserve d'un intérêt public substantiel et que des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ne soient en place.

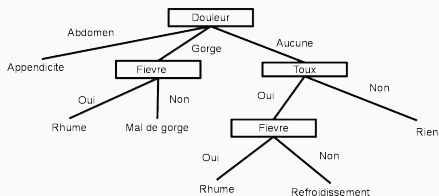
Explainability vs Interpretability

Two distinct notions (but the vocabulary is misleading: we follow here

<https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf>).

A decision rule is said to be:

interpretable if we understand how a prediction is associated to an observation; typical example: decision tree



<http://www.up2.fr/>

explainable if we understand what feature values led to the prediction, possibly by a counterfactual analysis; for example: "if variable X_3 had taken that other value, then the prediction would have been different".

Explainability vs Interpretability

Two distinct notions (but the vocabulary is misleading: we follow here

<https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf>).

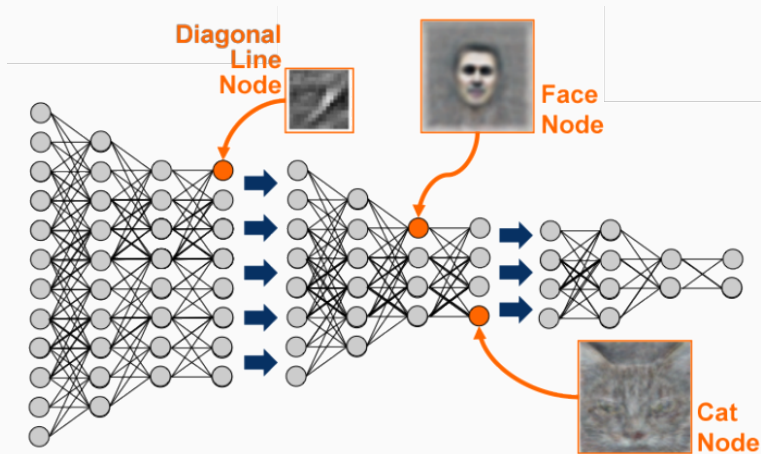
A decision rule is said to be:

interpretable if we understand how a prediction is associated to an observation; typical example: decision tree

explainable if we understand what feature values led to the prediction, possibly by a counterfactual analysis; for example: "if variable X_3 had taken that other value, then the prediction would have been different".

Explainability relates to the statistical notions of *causal inference* and *sensitivity analysis*

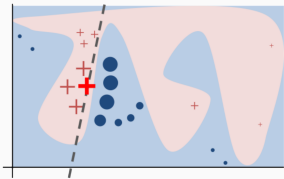
Interpreting a deep Neural Work : the Founding Dream



<http://aiehive.com>

An audacious scientific bet...

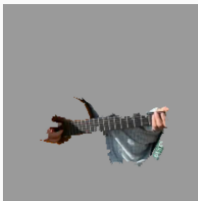
Local Interpretable Model-Agnostic Explanations: LIME



Linear model with feature selection on local subset of data



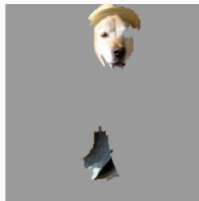
(a) Original Image



(b) Explaining *Electric guitar*



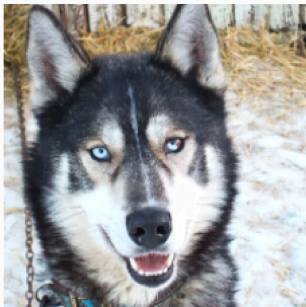
(c) Explaining *Acoustic guitar*



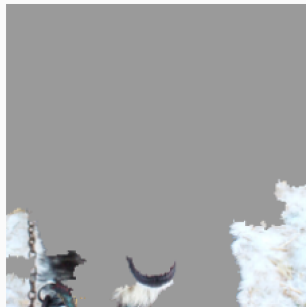
(d) Explaining *Labrador*

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.

- ▶ Huge need for more research and good practice
- ▶ Not only average performance matters
- ▶ Fairness should be included in data analysis with human impact
- ▶ Important issues that everyone should be aware of
- ▶ Interesting experiments to run at every level