# Automatic Decision by Machine Learning and Fairness

Café Stat' Lyon

Philippe Besse

Aurélien Garivier

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

UNIVERSITÉ DE LYON | ENS DE LYON

UMPA ENS DE LYON

LIP

# Table of contents

# Automatic Decision Systems based on Machine Learning

# Outline

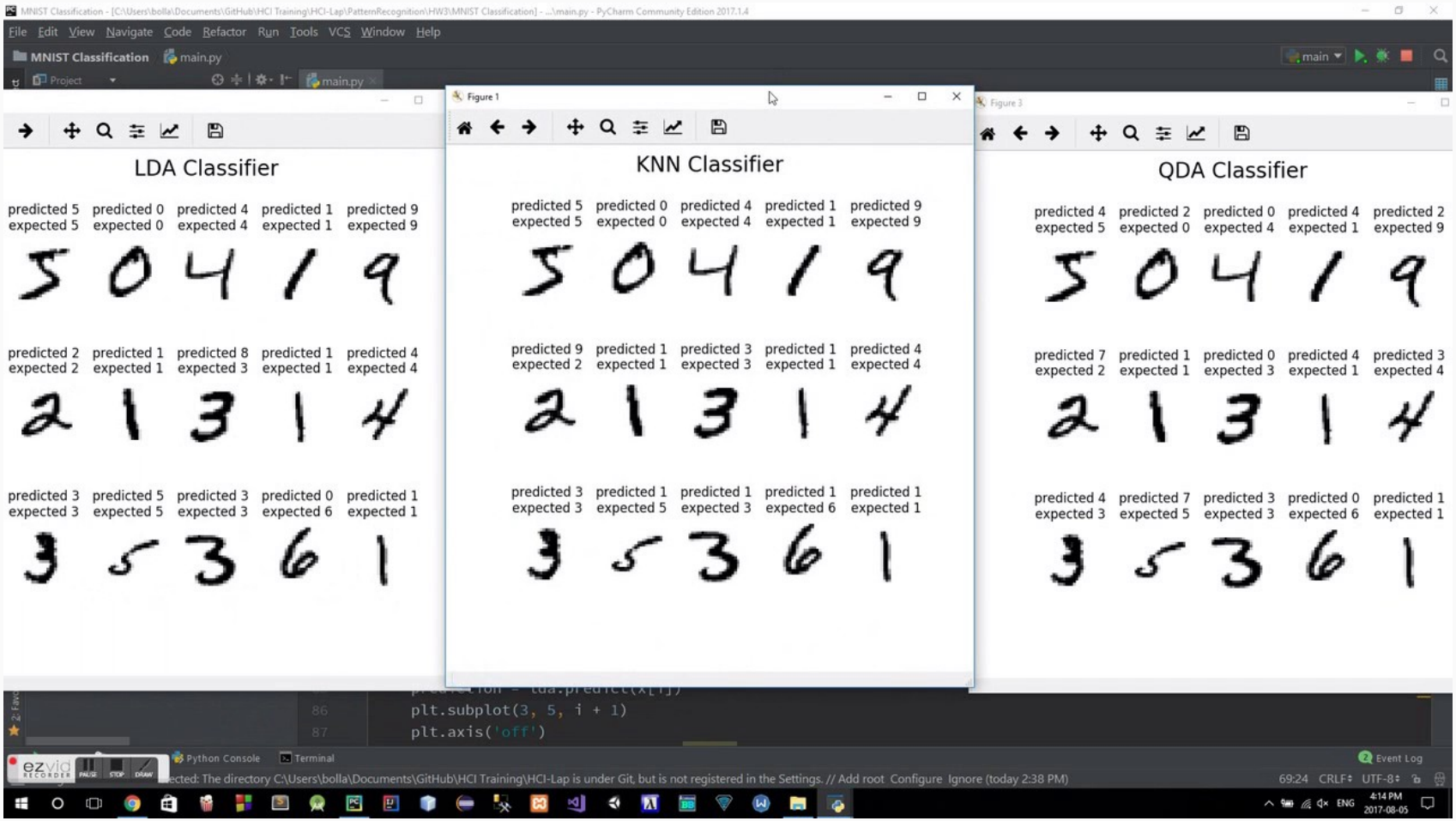
# What is Machine Learning?

- 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
- speach 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

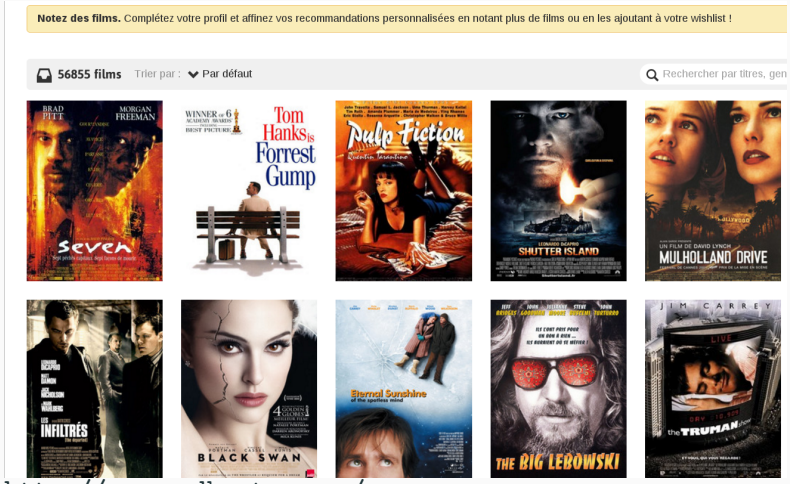

https://link.springer.com/article/10.1007

- Game solving (strategy)

AlphaGo

- Autonomous Vehicles

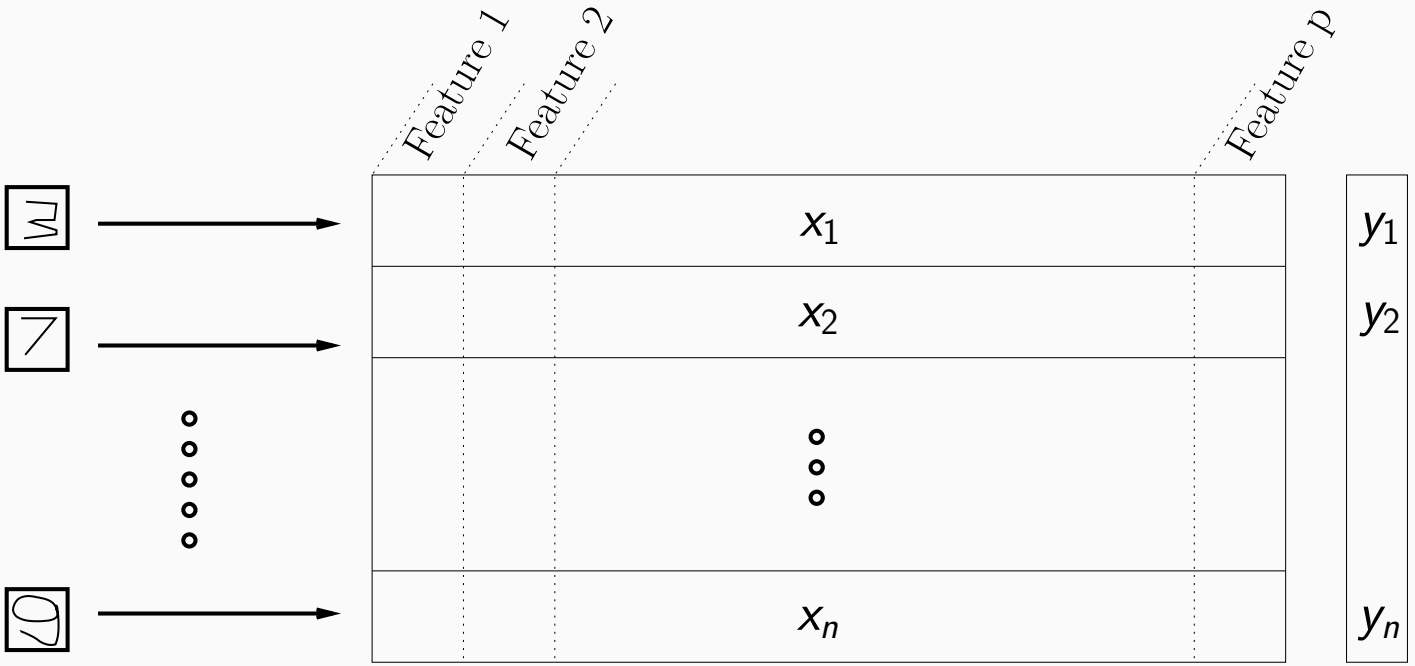
- Massive Recommender Systems: press, movies, ads, etc.




http://www.vodkaster.com/

# Outline

# What is a classifier?



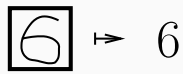

$$X \in \mathcal{M}_{n,p}(\mathbb{R}) \qquad Y \in \mathcal{Y}^n$$

Data: $n$-by-$p$ matrix $X$

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

Classifier $\mathcal{A}_n$

$$h_n : \mathcal{X} \to \mathcal{Y}$$

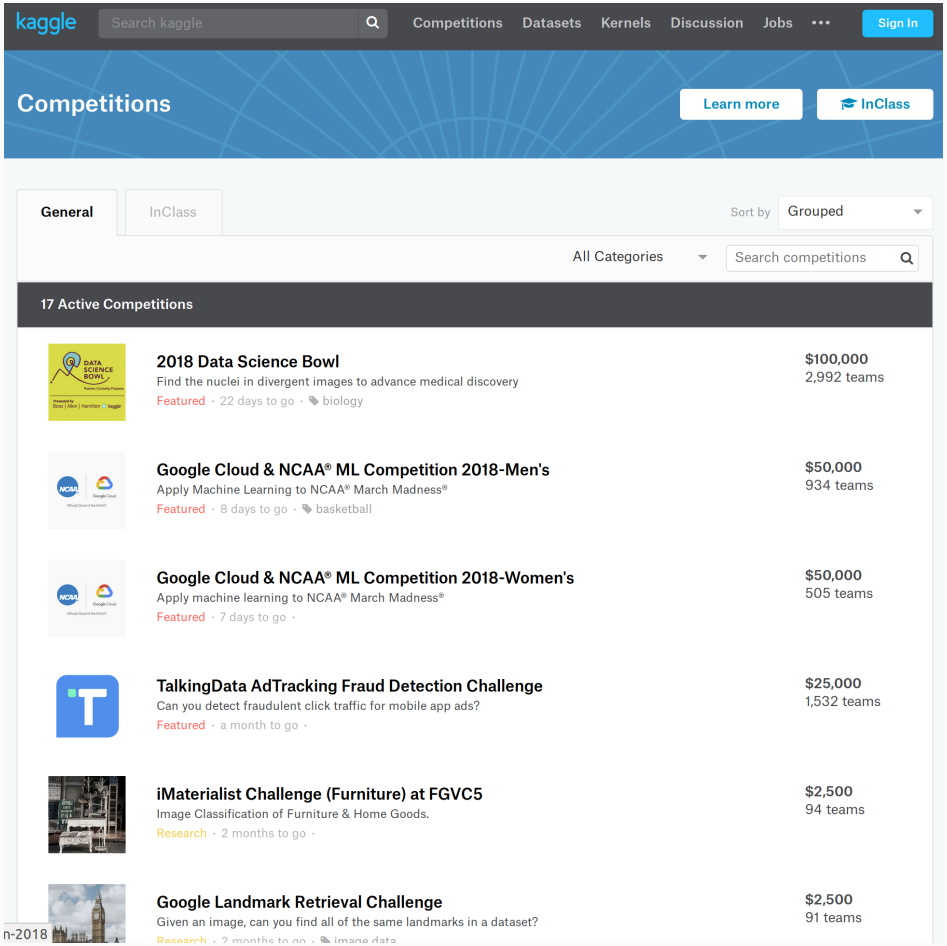
 $\mapsto 6$

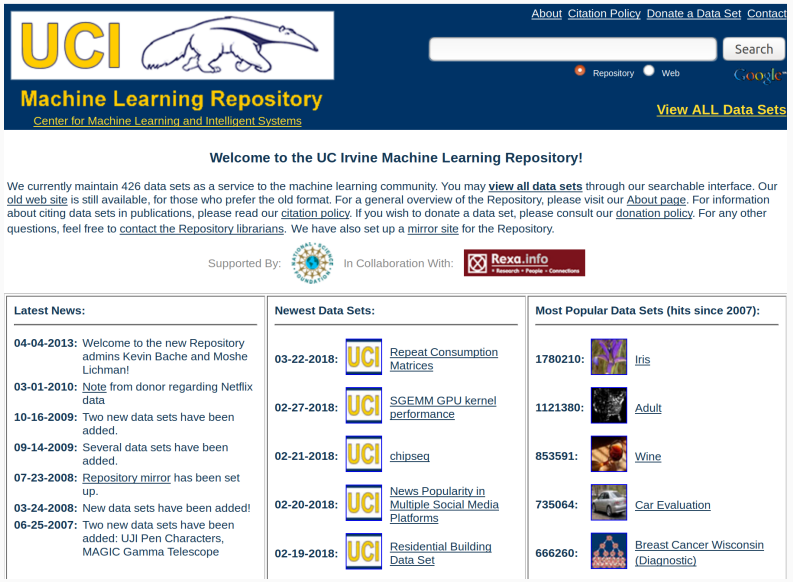
# Data repositories

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


- Challenges: Kaggle, etc.

# Benchmarks, Challenges: Competition to reach the Bayes Risk




- 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

# Outline

# 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 murs, 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.

# Fairness in automatic decision

**2 sources of biases**:

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

+ risk of "self-fullfilling prophecy"
Ex: higher credit insurance cost augment the risk of default

$\Longrightarrow$ need concepts and tools to

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

# Bias in the Data


`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 policitians 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!

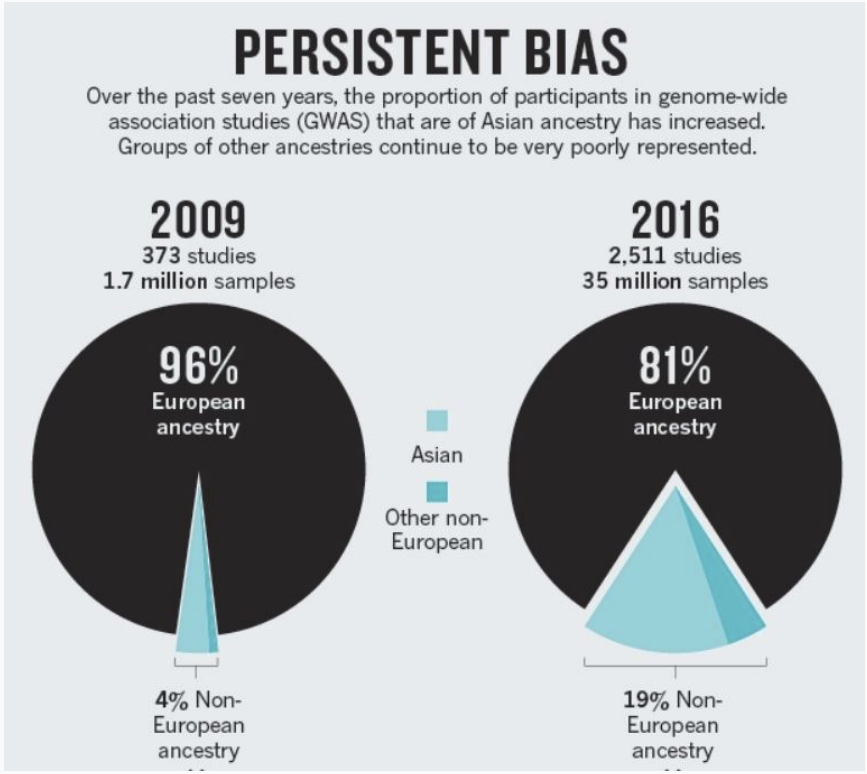
Why? Bias in the data!

"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#EZuQdOCJvJ3kYTiL.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.
  $\Longrightarrow$ 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

# 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): $DI = \frac{\mathbb{P}(\hat{h}_n(X) = 1 | S = 0)}{\mathbb{P}(\hat{h}_n(X) = 1 | S = 1)}$
- 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 $\mathbb{P}(\hat{h}_n(X) = 1 | S = 0)$

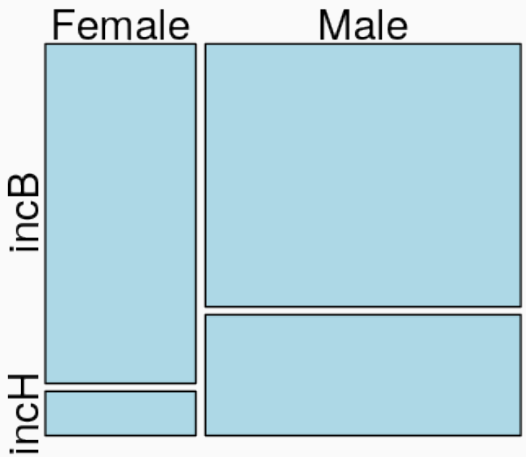
# An Example in more Detail

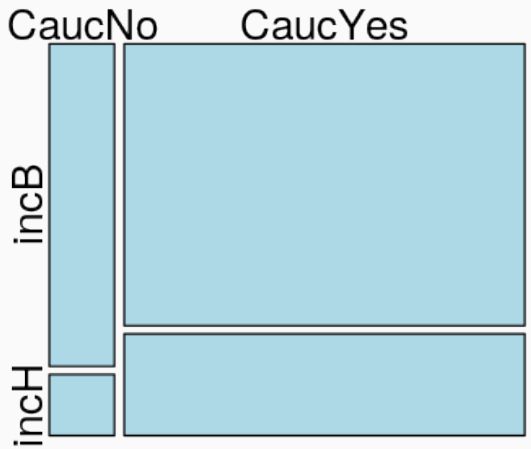
The following example is based on a Jupyter Notebook by **Philippe Besse** (INSA Toulouse) freely available (in R and python) on `https://github.com/wikistat`

# 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** $\rightarrow$ child, **race** $\rightarrow$ CaucYes / CaucNo

# 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...

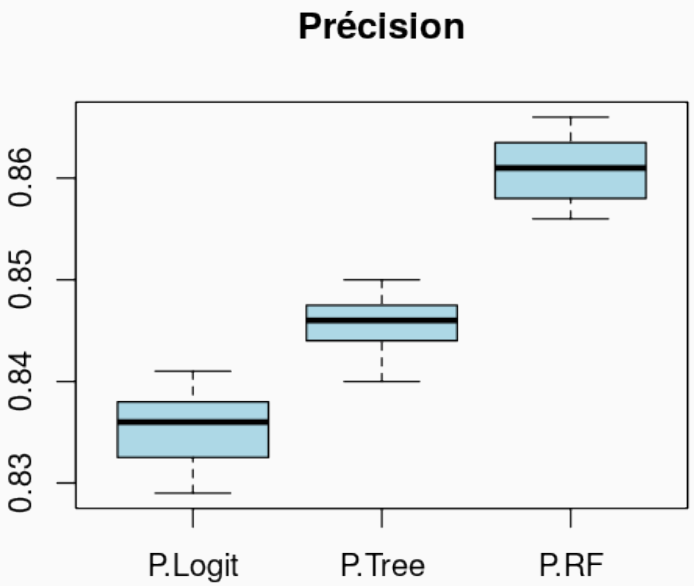
```
rf.mod=randomForest(income~.,data=datApp)
pred.rf=predict(rf.mod,newdata=daTest,type="response")
confMat=table(pred.rf,daTest$income)
confMat
tauxErr(confMat)
```

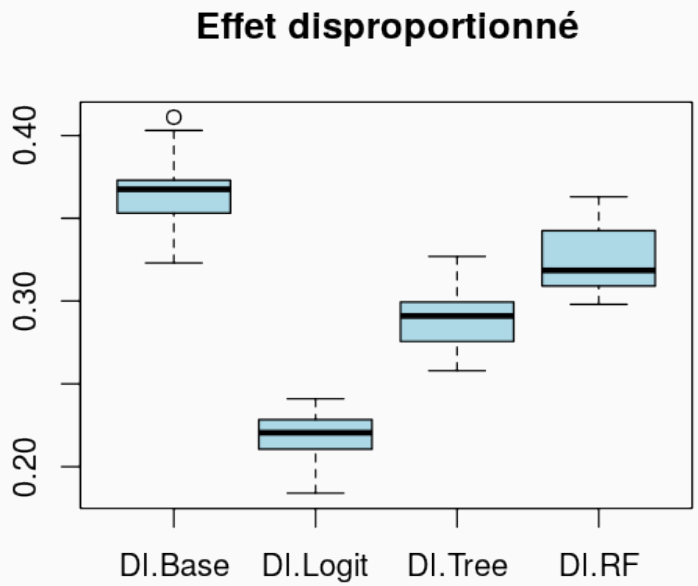
```
pred.rf  incB    incH
incB     6301     795
incH     445     1402

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"))
# Prvision du "nouvel" chantillon 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.

# Outline

# 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

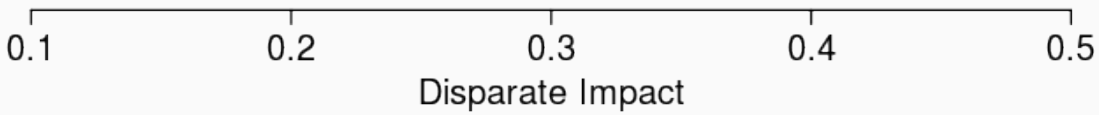
Logistic regression $\rightarrow$ consider the interactions of the protected variable with the others

```
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

| Model | Accuracy |
|---|---|
| dataBaseBias | 100 |
| linLogit | 83.5 |
| linLogit_w_S | 83.5 |
| linLogit-testing | 83.55 |
| condLinLogit | 83.62 |
| quadLogit | 83.54 |
| condQuadLogit | 83.32 |
| binaryTree | 85.45 |
| wBinaryTree | 85.31 |
| condBinTree | 85.16 |
| randomForest | 85.98 |
| wRandomForest | 85.91 |
| condRandForest | 85.74 |



# 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 maniére 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.

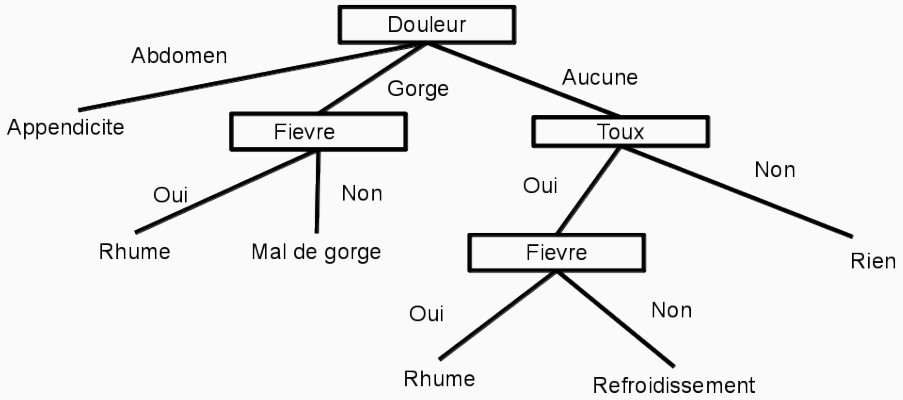
**4.** Les décisions visées au paragraphe 2 ne peuvent être fondées sur les catégories particuliéres de données à caractére personnel (cf. article 9 : biométriques, génétiques, de santé, ethniques ; orientation politique, syndicale, sexuelle, religieuse, philosophique) sous réserve d'un intérêt public 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 flllow 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 flllow here https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf ).

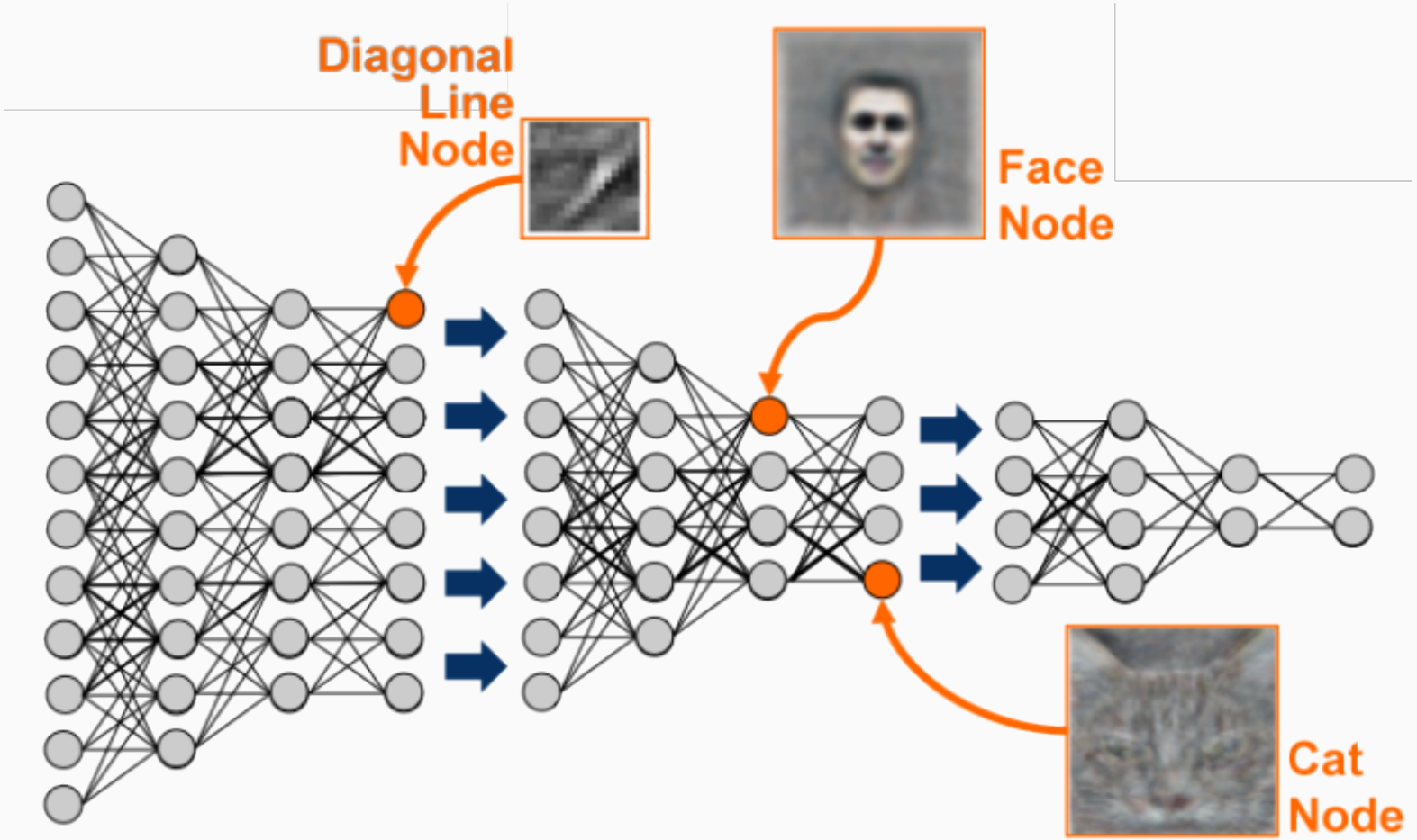
A decision rule is said to be:

**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".

Expainability relates to the statistical notions of *causal inference* and *sensibility analysis*

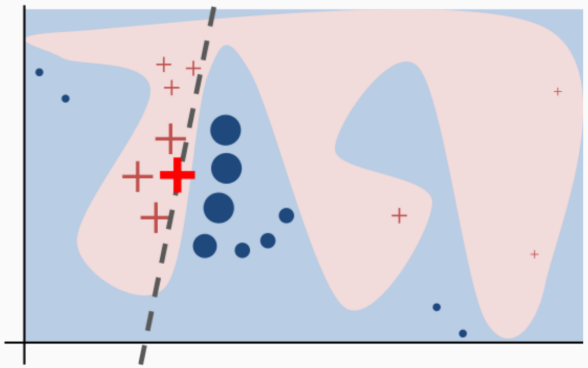
# Interpreting a deep Neural Work : the Founding Dream




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.

# Conclusion

- 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