ULB

ECOLE POLYTECHNIQUE DE BRUXELLES

UNIVERSITÉ LIBRE DE BRUXELLES

TRAIL TRUSTED AI LABS BY DIGITALWALLONIA / SPW-RECHERCHE

Biases in Machine Learning

Christine Decaestecker

LISA (Laboratory of Image Synthesis and Analysis)

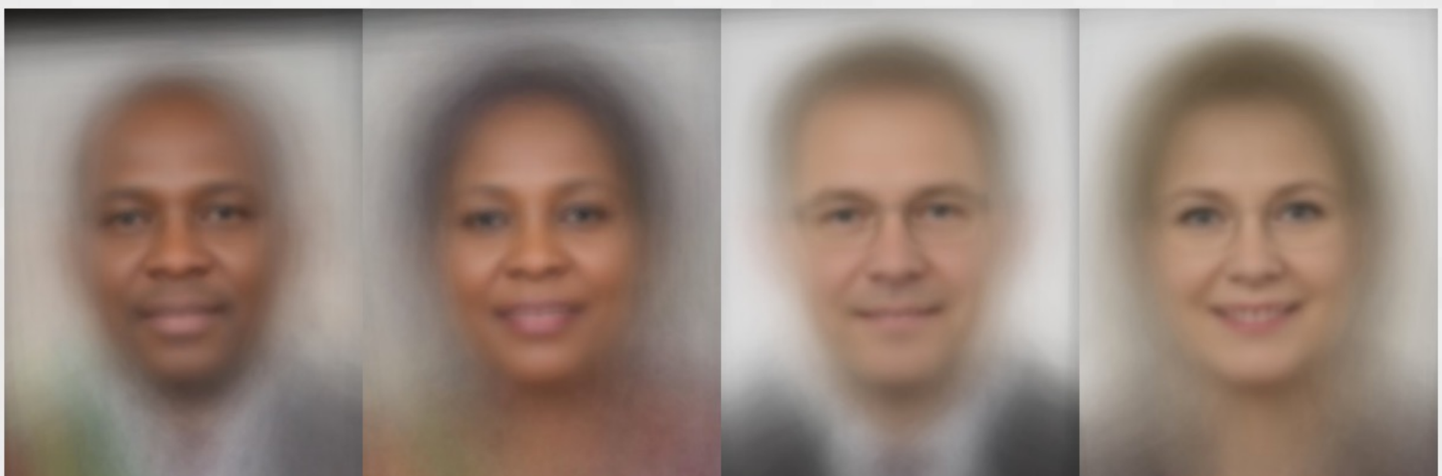
ULB - FNRS

Some overwhelming findings …

- **In male vs. female image recognition:** > 30 % of the dark-skinned female images are marked as male.

Michelle Obama, Oprah Winfrey and Serena Williams, were misidentified as male by Amazon and Microsoft.

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



http://gendershades.org/overview.html

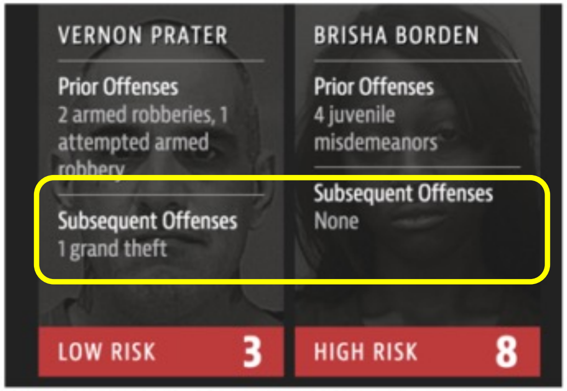
ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Some overwhelming findings …

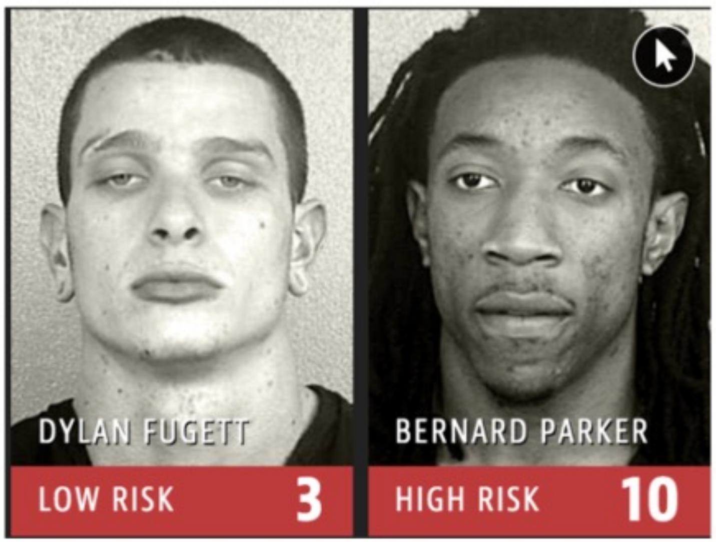
- **In male vs. female image recognition:** 34.7 % of the dark-skinned female images are marked as male.
- **Amazon's recognition system** wrongly identified 28 of members of the U.S. Congress as criminals

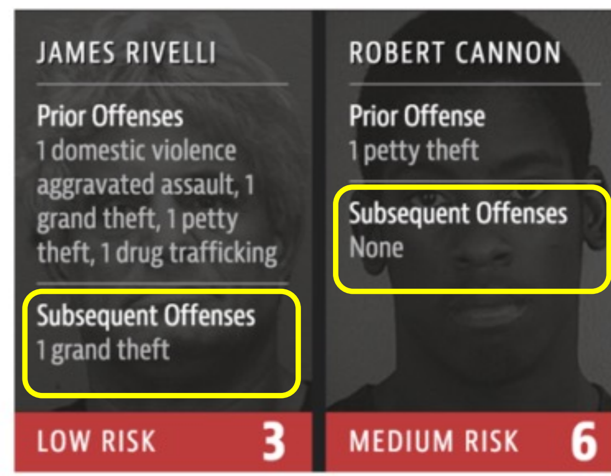
Prediction of criminal recidivism risk



reality

prediction





https://www.backprop.fr/les-biais/

Some overwhelming findings …

- **In male vs. female image recognition:** 34.7 % of the dark-skinned female images are marked as male.
- **Amazon’s recognition system** wrongly identified 28 of members of the U.S. Congress as criminals
- **Apple Pay Card algorithm** granted a higher credit limit to men than women, despite equivalent incomes

Some overwhelming findings …

- **In male vs. female image recognition:** 34.7 % of the dark-skinned female images are marked as male
- **Amazon's recognition system** wrongly identified 28 of members of the U.S. Congress as criminals
- **Apple Pay Card algorithm** granted a higher credit limit to men than women, despite equivalent incomes

BIASES

Recent changes:

- **In 2018,** Amazon abandoned an AI system for IT staff recruitment because of a bias against women.
- **In 2020,** letter to US Congress: "IBM no longer offers general purpose IBM facial recognition or analysis software. "

https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/

What happened with AI ?

- AI and Machine Learning have come **out of the research labs** massively.

What happened with AI?

- AI and Machine Learning have come **out of the research labs** massively.
- Were society, companies and users **ready**?
  The press and media are telling us:

NO or, at least, **NOT YET**

What happened with AI?

- AI and Machine Learning have come **out of the research labs** massively.
- Were society, companies and users **ready**?

NO or, at least, **NOT YET**

- **What to do** to make AI more responsive to the needs of society?

ULB

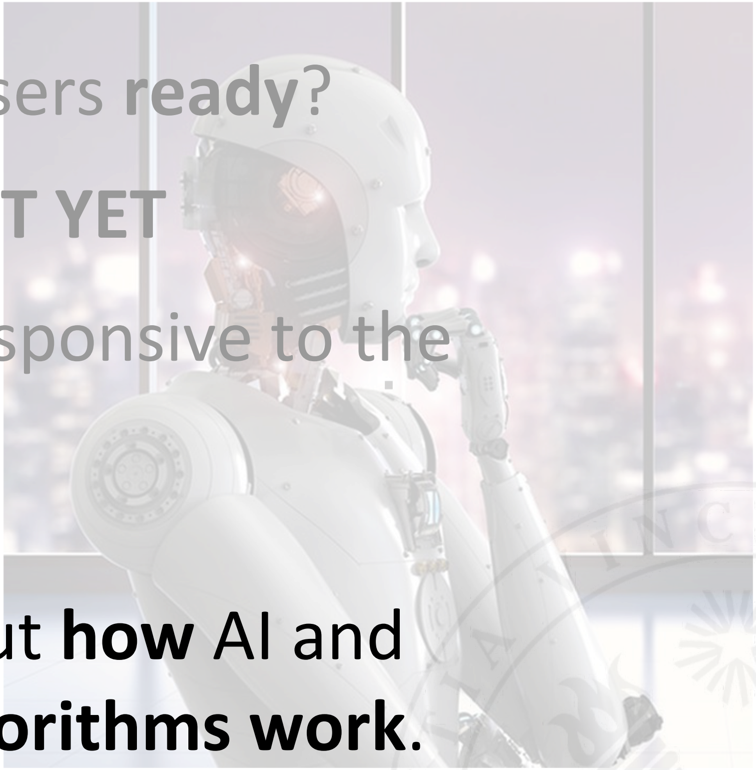
UNIVERSITÉ LIBRE DE BRUXELLES

What happened with AI?

- AI and Machine Learning have come **out of the research labs** massively.
- Were society, companies and users **ready**?

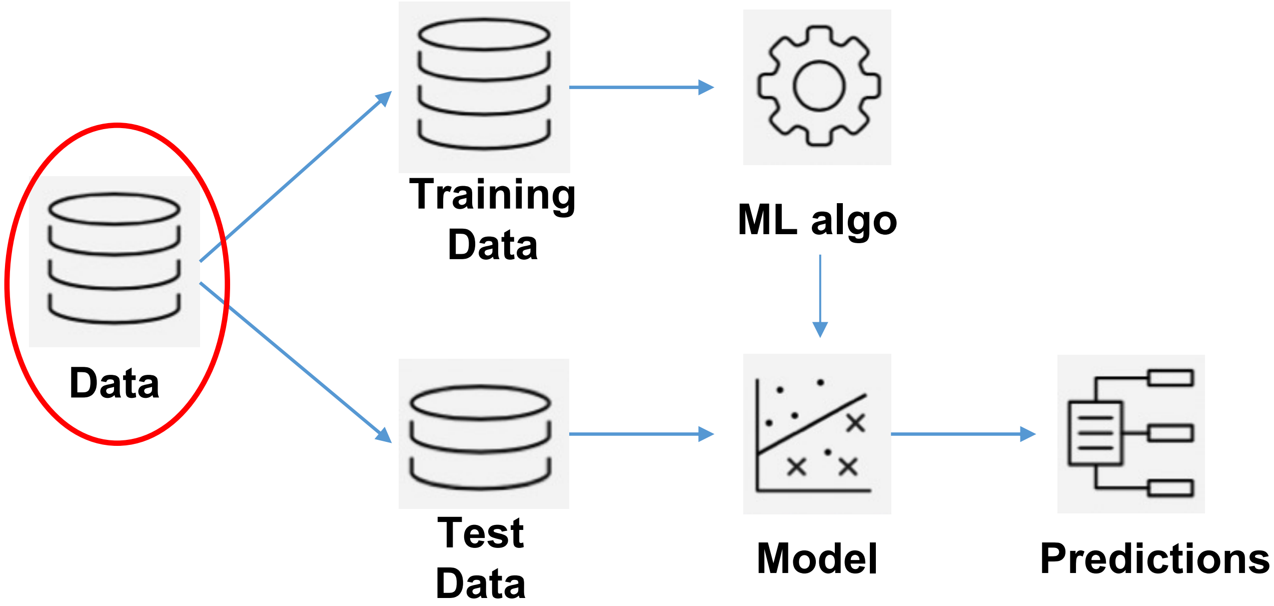
NO or, at least, **NOT YET**

- **What to do** to make AI more responsive to the needs of society?
- One essential key:
  Having enough **knowledge** about **how** AI and especially **machine learning algorithms work**.



Machine learning: basic principles

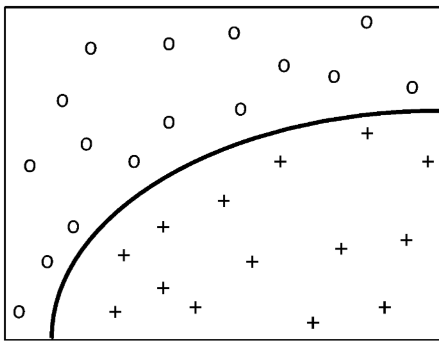
- **Training a model and validating it:**



- **And after that deploying/applying it in the real world: production phase**

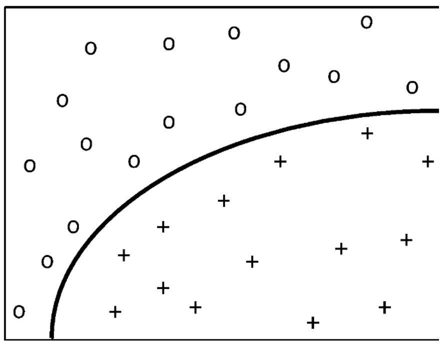
Machine learning: basic principles

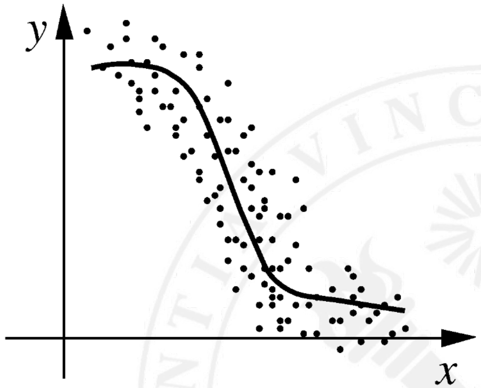
- **Machine learning = data-centric methodology**
  Using training data to extract **statistical characteristics** and **relationships** able to
  - classify data into labeled groups
    (e.g. fraud detection) = classification task



Machine learning: basic principles

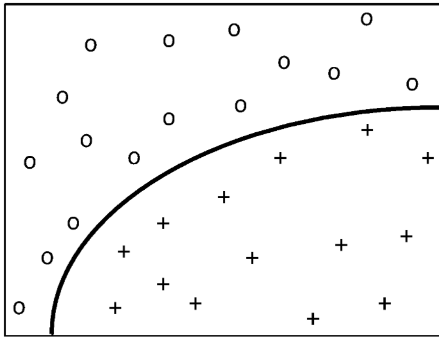
- **Machine learning = data-centric methodology**
  Using training data to extract **statistical characteristics** and **relationships** able to
  - classify data into labeled groups
    (e.g. fraud detection) = classification task
  - predict a quantitative feature
    (e.g., insurance pricing) = regression task
  - etc

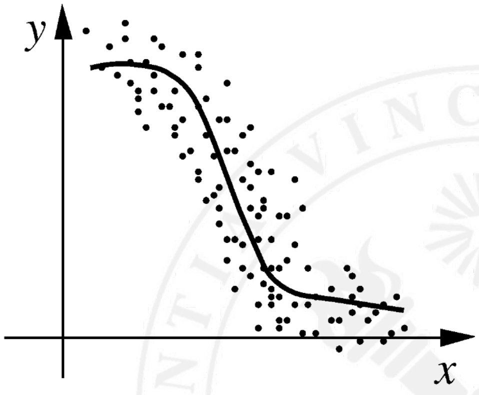




Machine learning: basic principles

- **Machine learning = data-centric methodology**
  Using training data to extract **statistical characteristics** and **relationships** able to
  - classify data into labeled groups
    (e.g. fraud detection) = classification task
  - predict a quantitative feature
    (e.g., insurance pricing) = regression task
  - etc

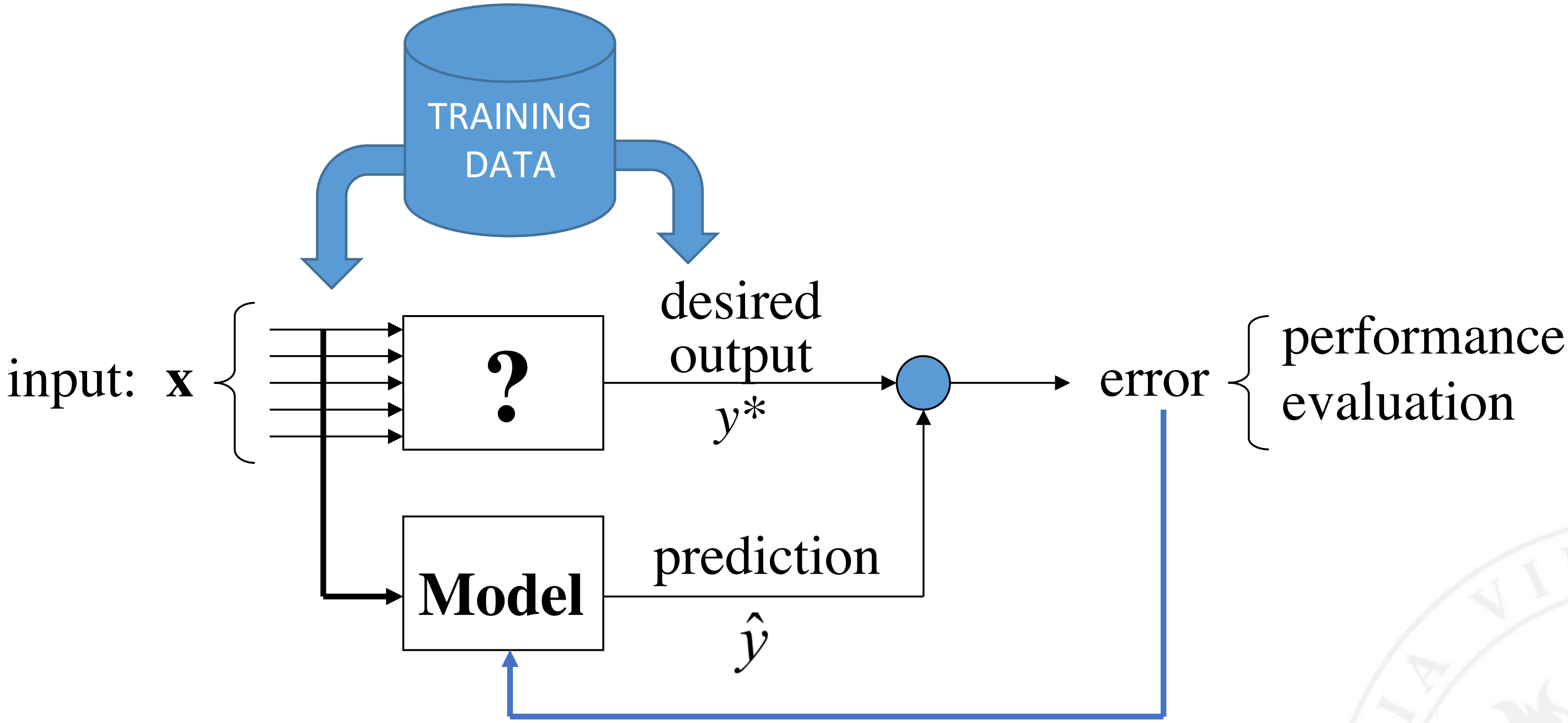




⇨ **ML algorithms rely on training data as ground truth, i.e. as representative of the real world and the job to do !**

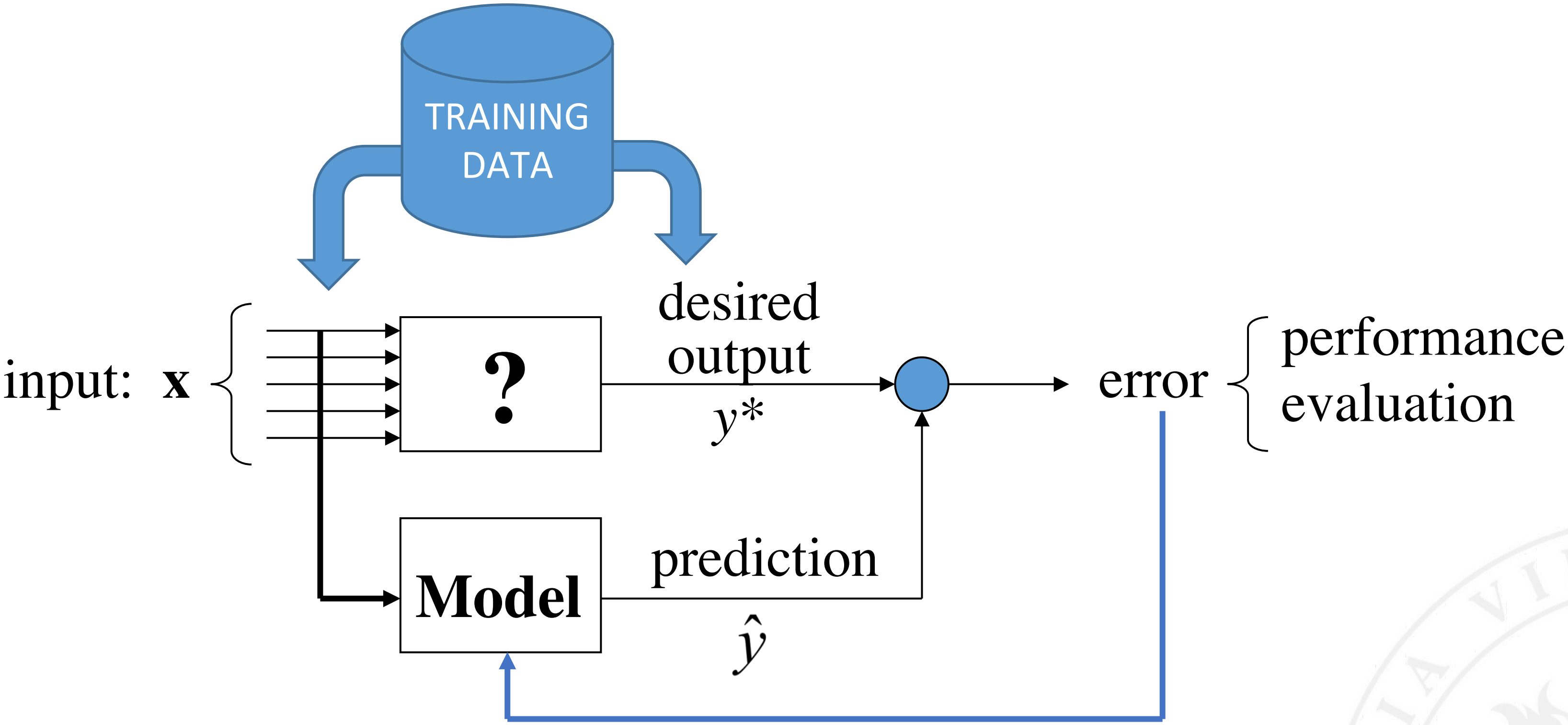
Machine learning: basic principles

- **Supervised training from data:**



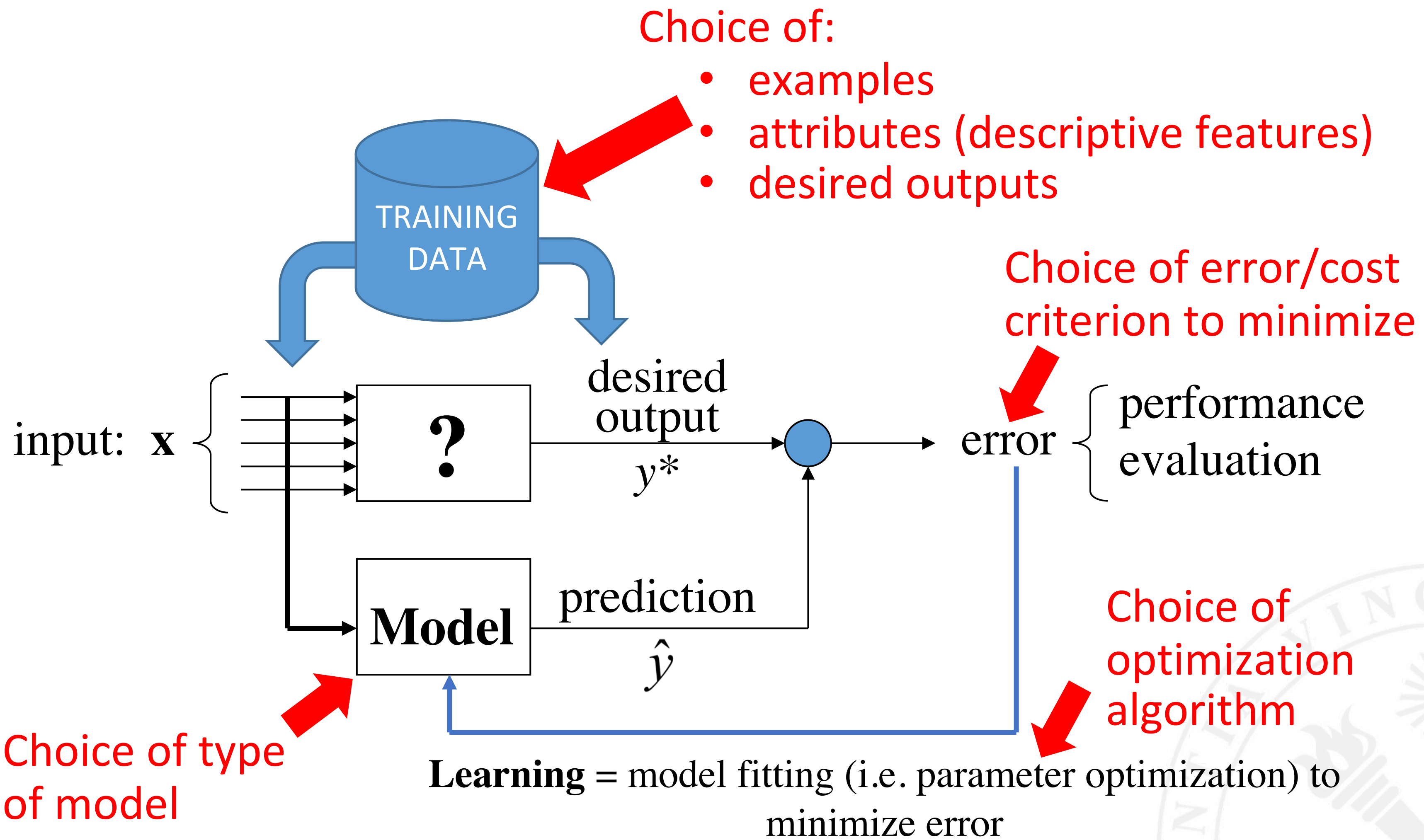
Learning = model fitting (i.e. parameter optimization) to minimize error

Machine learning: **bias sources**

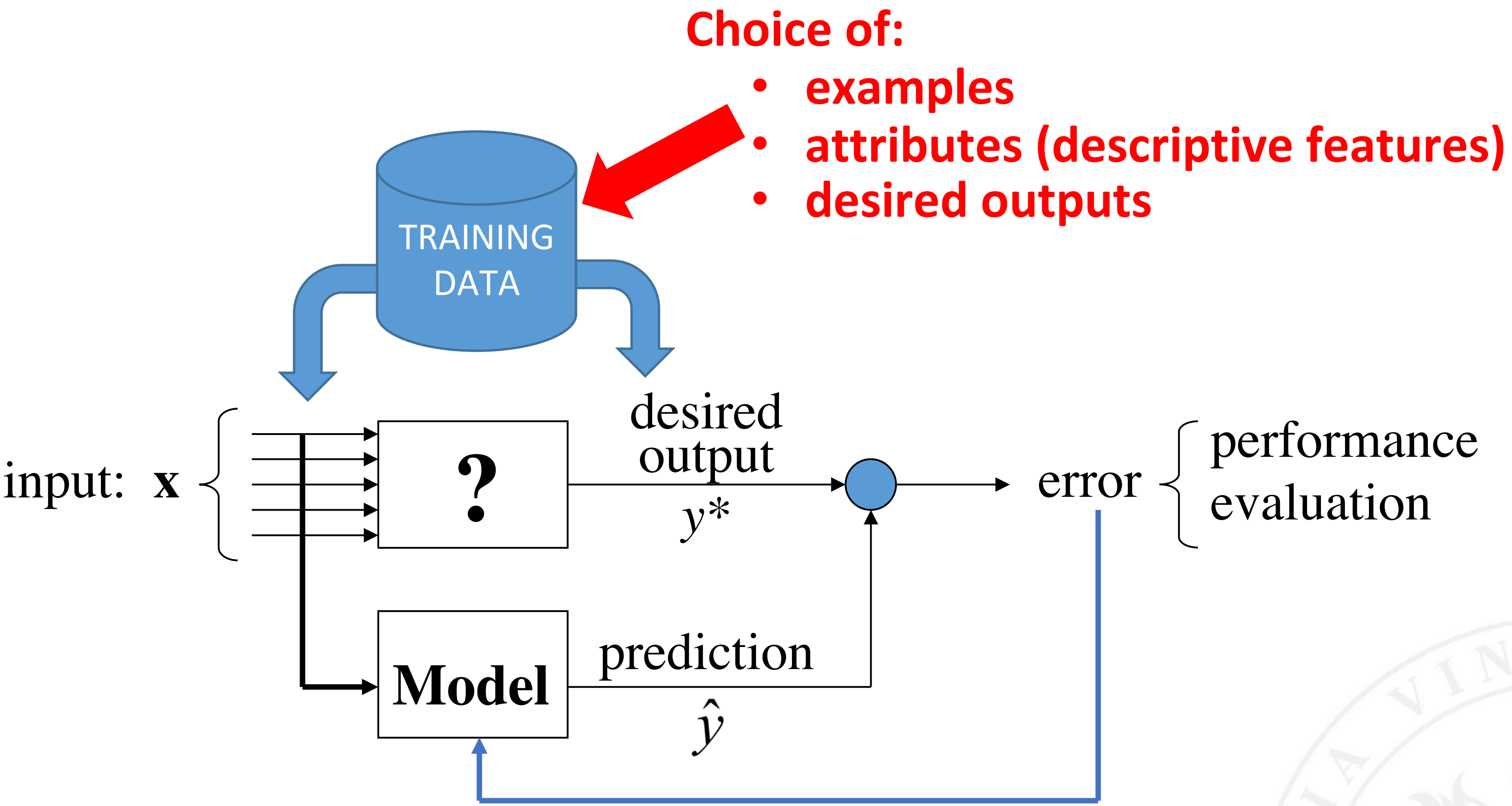


Learning = model fitting (i.e. parameter optimization) to minimize error

Machine learning: **bias sources**



Machine learning: bias sources



Learning = model fitting (i.e. parameter optimization) to minimize error

Bias source: data imperfections

- **Representativeness:**
  - **lack of representativeness** of certain sub-groups or minorities (may result from societal and/or historical prejudices)
  - **too old:** not adapted to the future application context and to changes in society (e.g. CV database)

Bias source: data imperfections

- **Representativeness:**
  - **lack of representativeness** of certain sub-groups or minorities (may result from societal, historical prejudices)
  - **too old:** not adapted to the future application context and to changes in society
- **Attribute quality:**
  - **errors** in attribute values
  - **not informative enough** to be able to solve the problem
  - include **potential discrimination** sources (e.g. gender, race, age, nationality …) in databases

Bias source: data imperfections

ML models can only be as good as the data on which they are trained:

- inherit the historical prejudices (from prior decision makers) and/or the widespread biases that persist in society

Bias source: data imperfections

ML models can only be as good as the data on which they are trained:

- inherit the historical prejudices (from prior decision makers) and/or the widespread biases that persist in society

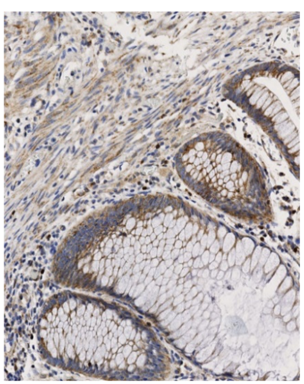
Even if potentially discriminatory attributes are omitted:

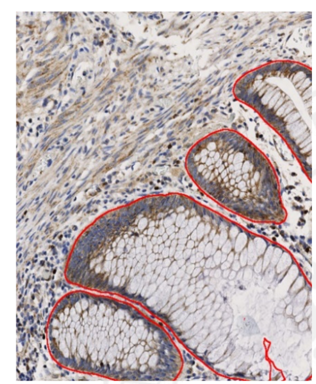
- may extract and then use **(hidden) data regularities** that are preexisting patterns of exclusion and inequality (e.g. hidden link between gender and hobbies in CV)

Bias sources: raw data

Standard ML approach

Input (raw data) → Extraction of expert features → Machine learning algorithm → Output





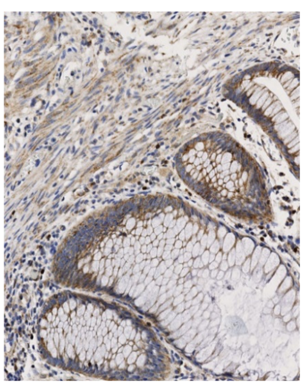
Bias sources: raw data

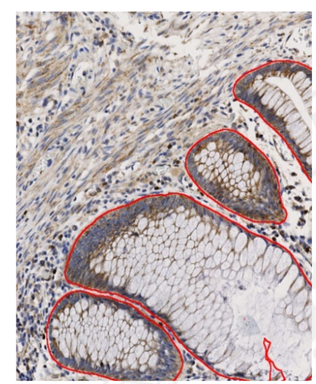
Deep learning on raw data

(text/signal/image/video processing)

Input (raw data) → Extraction of expert features → Machine learning algorithm → Output

Input (raw data) → **Deep learning** algorithm → Output





Automatic extraction of features

No selection bias but possible other biases more difficult to identify (e.g. related to the data source and acquisition process)

Bias source (data): desired outputs

- The formalization of the desired outputs (target variable to be predicted) can be not obvious:
  - often subjective translation of a decision problem into a question about the value of a target variable

Bias source (data): desired outputs

- The formalization of the desired outputs (target variable to be predicted) can be not obvious:
  - often subjective translation of a decision problem into a question about the value of a target variable
- Values are either provided by humans/experts:
  - possible errors in difficult tasks
  - not always well defined (e.g. what is good, what is bad?)

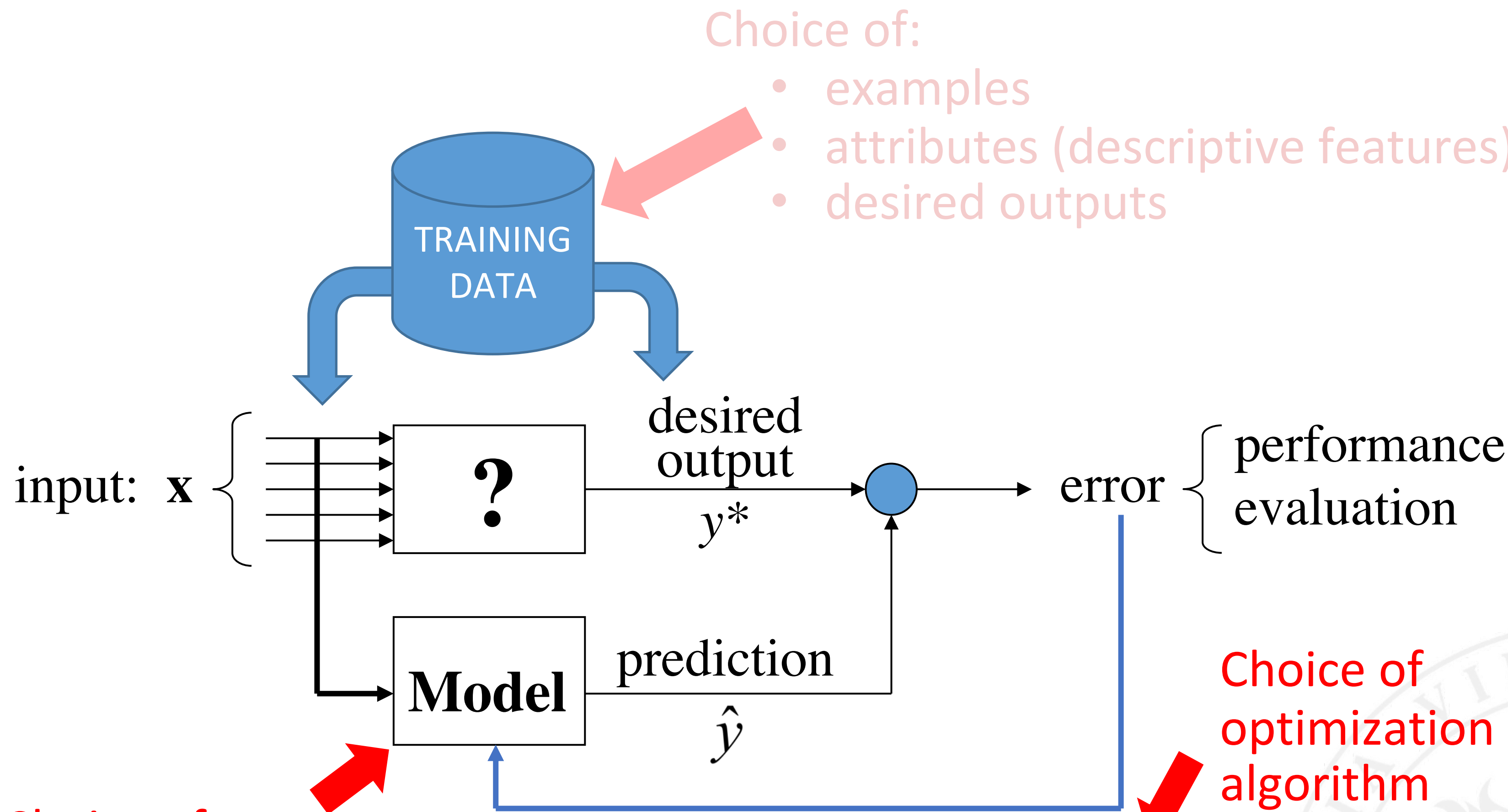
  => “supervisor” dependency

Bias source (data): desired outputs

- The formalization of the desired outputs (target variable to be predicted) can be not obvious:
  - often subjective translation of a decision problem into a question about the value of a target variable
- Values are either provided by humans/experts:
  - possible errors in difficult tasks
  - not always well defined (e.g. what is good, what is bad?)

  => “supervisor” dependency
- Or by a rule, calculation, simulation, or resulting from a costly / time-consuming process
  - may be also biased or erroneous

Machine learning: bias sources

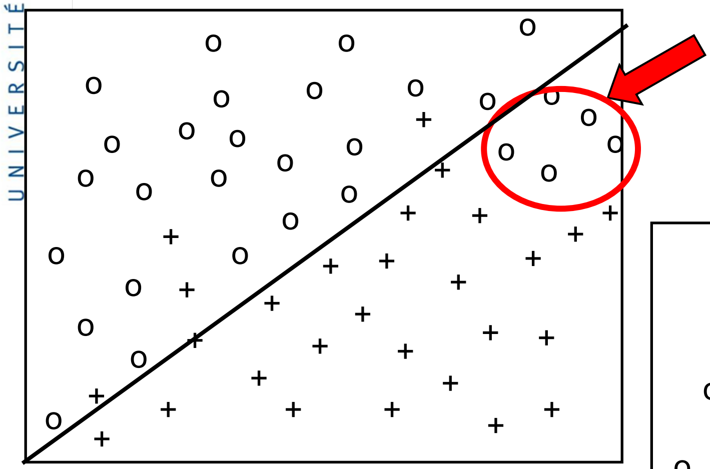


Learning = model fitting (i.e. parameter optimization) to minimize error

Bias source: model & optimization algorithm

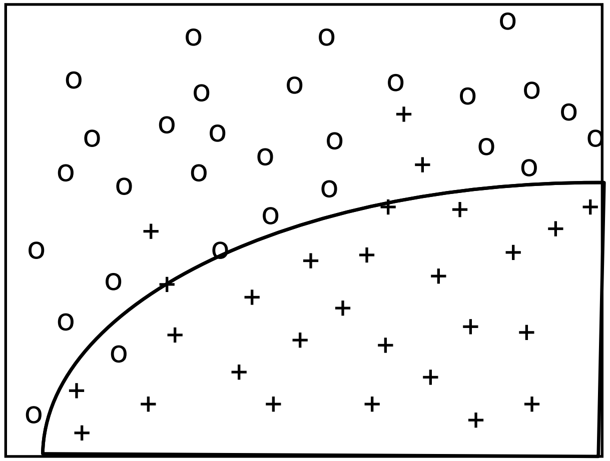
- Model flexibility/complexity adapted or not to the task?

Well known problem!

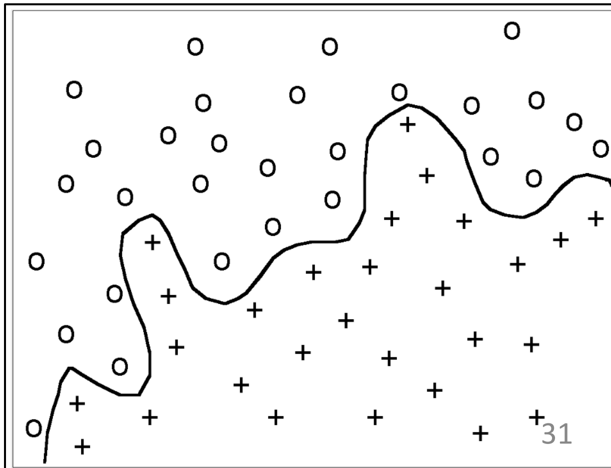


Underfitting

Best solution

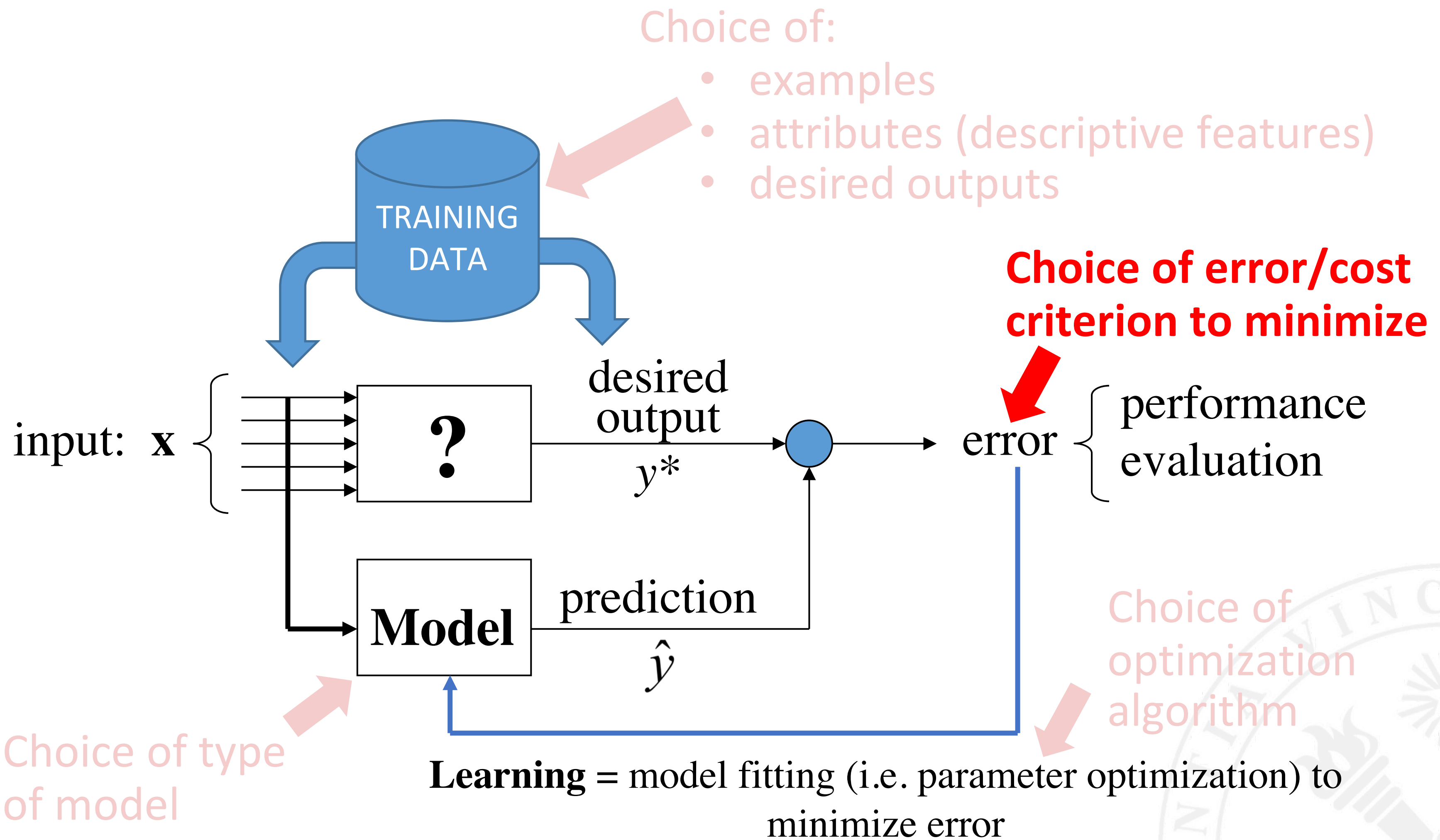


Overfitting



Standard strategies exist to solve it

Machine learning: bias sources



Learning = model fitting (i.e. parameter optimization) to minimize error

Bias source: error criterion

- What is the error/optimization criterion used for training?

Bias source: error criterion

- What is the error/optimization criterion used for training?
- For classification: balancing the false positive and false negative rates

Bias source: error criterion

- What is the error/optimization criterion used for training?
- For classification: balancing the false positive and false negative rates
- Many different mathematical definitions of such a balance

Bias source: error criterion

- What is the error/optimization criterion used for training?
- For classification: balancing the false positive and false negative rates
- Many different mathematical definitions of such a balance
- Should be adapted to the application: a biased balance can be more appropriate for some tasks!
    - In disease screening: avoid false negatives even if an increase of false positives (which will be identified by subsequent examinations)

Combination of bias sources

- **Effects of unbalanced class priors on classification error rates:**

	Prediction		
True class	**A**	**B**	**C**
A (n = 50)	50	0	0
B (n = 15)	0	10	5
C (n = 35)	0	15	20

- **Global** error rate: (5+15)/100 = **20%**
- **Mean** error rate **per class:** (0 + 33.6 + 42.9)/3 = **25.4%**

Combination of bias sources

- **Effects of unbalanced class priors on classification error rates:**

	Prediction		
True class	**A**	**B**	**C**
A (n = 50)	50	0	0
B (n = 15)	0	10	5
C (n = 35)	0	15	20

  - **Global** error rate: (5+15)/100 = **20%**
  - **Mean** error rate **per class:** (0 + 33.6 + 42.9)/3 = **25.4%**

- Standard error criteria are based on **sum of errors:** bias the model to **perform better for the most frequent class(es) in the training data,** possibly to the detriment of the other classes.

Combination of bias sources

- **Effects of unbalanced class priors on classification error rates**:

	Prediction		
True class	**A**	**B**	**C**
A (n = 50)	50	0	0
B (n = 15)	0	10	5
C (n = 35)	0	15	20

  - **Global** error rate: (5+15)/100 = **20%**
  - **Mean** error rate **per class:** (0 + 33.6 + 42.9)/3 = **25.4%**
- Standard error criteria are based on **sum of errors**
- Have a look on **detailed error distribution** to detect possible biases and use **normalized and "disentangled" metrics**

Detect and mitigate biases

Numerous studies in AI/machine learning:

- E. Celis, et al., "**Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees**", FAT* '19: Conference on Fairness, Accountability, and Transparency, **2019**
- T. Speicher, et al., "**A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices**", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **2018**.
- B. Hu Zhang, et al., "**Mitigating Unwanted Biases with Adversarial Learning**", AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, **2018**
- F. P. Calmon, et al., "**Optimized Pre-Processing for Discrimination Prevention**", Conf. on NIPS, **2017**
- G. Pleiss, et al., "**On Fairness and Calibration**", Conference on NIPS, **2017**.
- M. Hardt, et al., "**Equality of Opportunity in Supervised Learning**", Conference on NIPS, **2016**.
- M. Feldman, et al., "**Certifying and Removing Disparate Impact**", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **2015**.
- R. Zemel, et al., "**Learning Fair Representations**", Int. Conf. on Machine Learning, **2013**.
- F. Kamiran, T. Calders, "**Data Preprocessing Techniques for Classification without Discrimination**", Knowledge and Information Systems, **2012**.
- F. Kamiran, et al., "**Decision Theory for Discrimination-Aware Classification**", IEEE International Conference on Data Mining, **2012**.
- T. Kamishima, et al., "**Fairness-Aware Classifier with Prejudice Remover Regularizer**", Joint European Conference on Machine Learning and Knowledge Discovery in Databases, **2012**.

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Some (very) general guidelines: training models and after!

- **Control data** used for training, validating and accuracy evaluation of the algorithm
  - Balance the **representativeness of each (sub)group of interest**: collect more data, weight their impact in the error criteria, use data augmentation techniques, …

Data augmentation: to balance training data and avoid biases

- Generating new <u>realistic</u> samples to enrich minority subgroups:
  use of Generative Adversarial Networks (GAN, deep learning) to avoid racial bias in face recognition

Original image

Generated images

Some (very) general guidelines: training models and after!

- **Control data** used for training, validating and accuracy evaluation of the algorithm
  - Balance the **representativeness of each (sub)group of interest**: collect more data, weight their impact in the error criteria, use data augmentation techniques, …
- **Control algorithm behavior in real situations, but also extreme cases, <u>before</u>** going into production

ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Some (very) general guidelines: training models and after!

- **Control data** used for training, validating and accuracy evaluation of the algorithm
  - Balance the **representativeness of each (sub)group of interest**: collect more data, weight their impact in the error criteria, use data augmentation techniques, …
- **Control algorithm behavior in real situations, but also extreme cases, before** going into production
- **During production: regularly check** that the context of the application has not changed
  - requires model retraining or refining or output post-processing

Technical resource

- **IBM AI Fairness 360:** open source Python toolkit
  - to examine, report, and mitigate discrimination and bias in (data and) machine learning models **throughout the AI application lifecycle**
  - comprehensive set of fairness metrics for datasets and models
  - explanations for algorithms to mitigate bias in datasets and models

  http://aif360.mybluemix.net/

Technical resource

- **IBM AI Fairness 360:** open source toolkit
  - to examine, report, and mitigate discrimination and bias in (data and) machine learning models **throughout the AI application lifecycle**
  - comprehensive set of fairness metrics for datasets and models
  - explanations for algorithms to mitigate bias in datasets and models

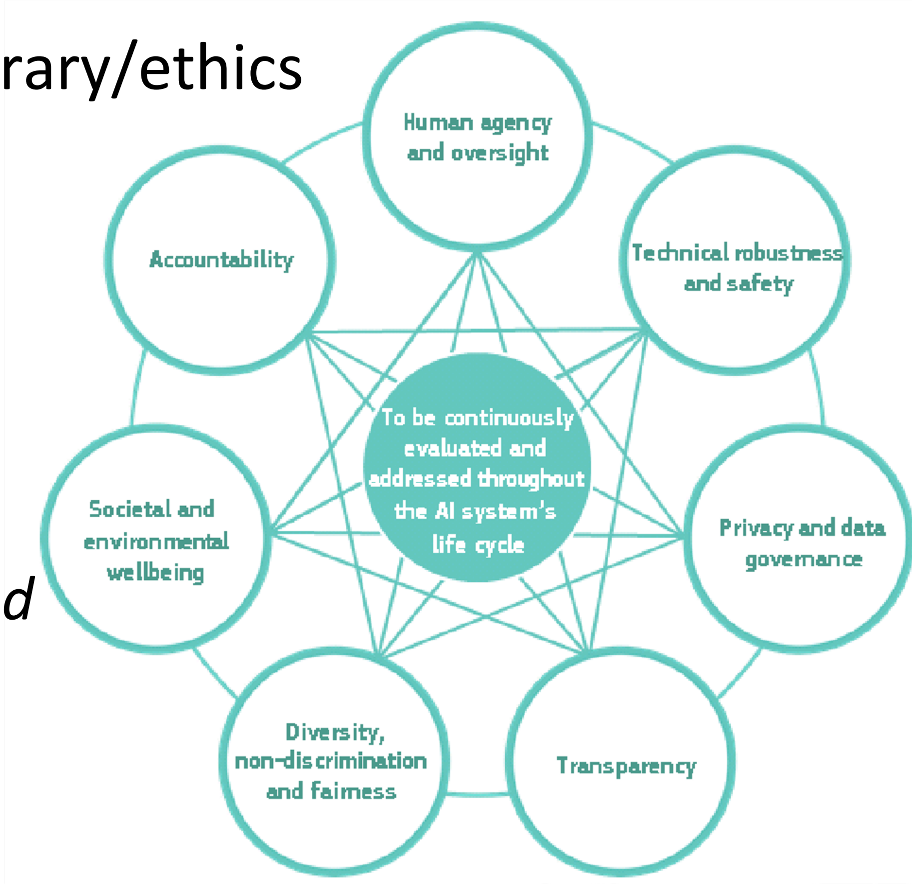
  http://aif360.mybluemix.net/

- Numerous interesting blogs: https://towardsdatascience.com/**understanding-and-reducing-bias-in-machine-learning**-6565e23900ac

Ethics resource from EU

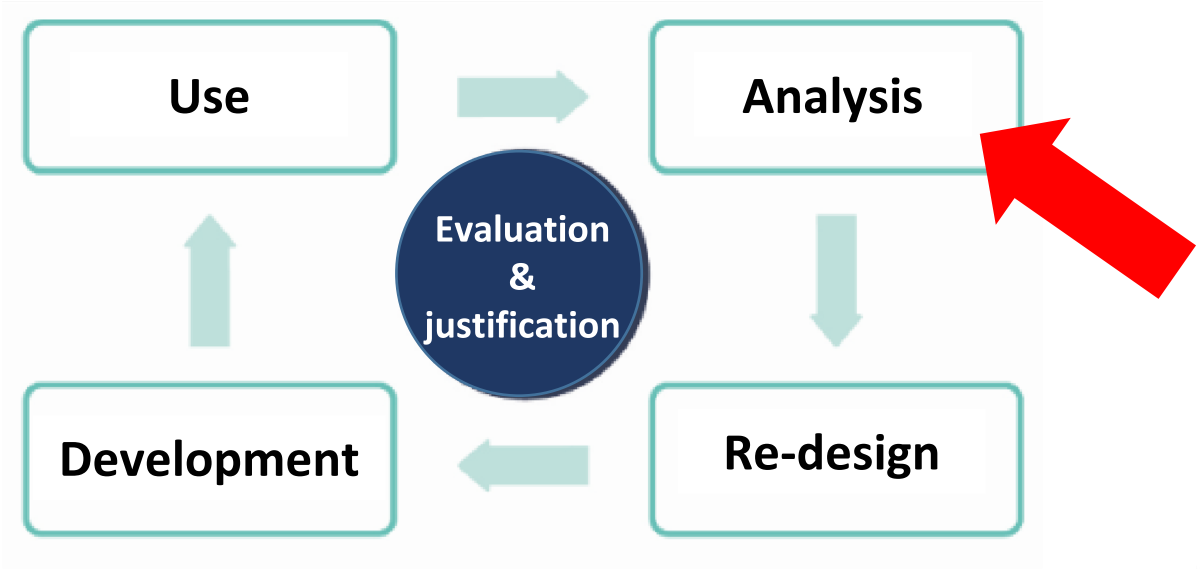
- **Ethics Guidelines for Trustworthy AI** https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle



Conclusion

- Ideally, AI systems are continuously evolving and acting in a dynamic environment



- **A lot of exciting things to do at different levels, both in and out of research labs!**

Thank you for your attention