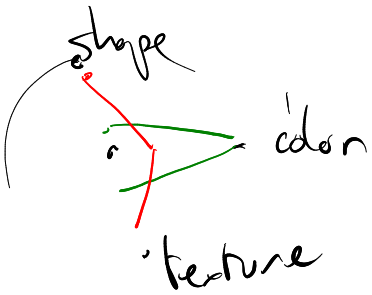
Interpretability

- ex: medical diagnosis



- Biases
  - Binary classif° task:
  - classificat° of diseases

→ Landscape
↘ town

A, B

d

Hospital 1
scans → A

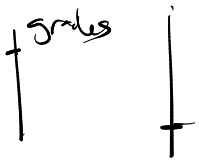
Hospital 2
scans → B

ex: MRI scan
CT

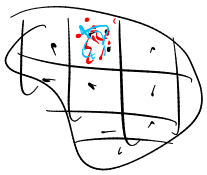
config°: - contrast
- noise
- frequencies

Societal impact: "Weapons of Maths Destruction" by Cathy O'Neil 2016?

- hiring (for avatar jobs)
- Firing (in a US school)



- job offers : CV
- COMPAS : jails : recidivism → biased against black
- police patrol optimo



→ Crucial: Feedback (from people involved), explainable, right to contest/appeal

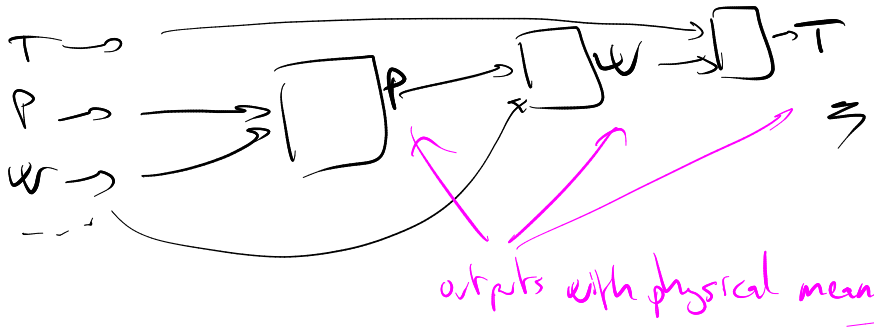
→ think twice about the impact of your algorithms before deploying them

Be responsible & careful

AI ethics: Montreal declaration for responsible AI

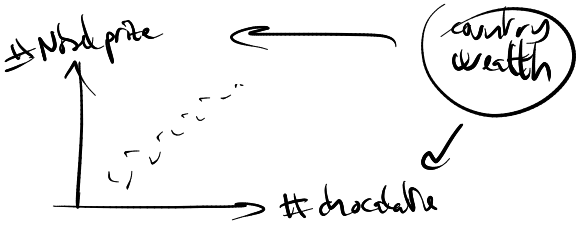
↳ FAT-ML: Fairness, Accountability, Transparency in ML

Interpretability by design : xAI

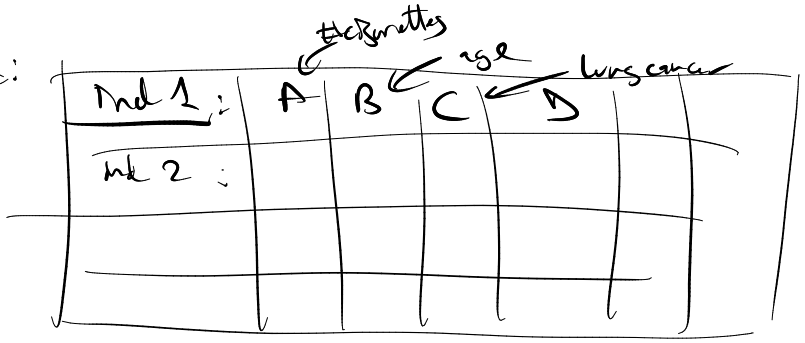


Causality

Correlat° ≠ causality



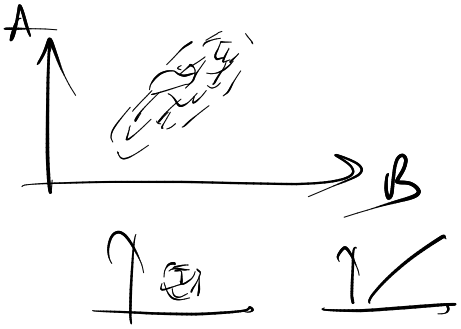
Data:



Fixed dataset

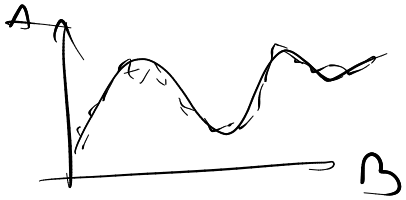
correlat°

causality ?



A = f(B) ?

B = f(A) ?



B → A

Dataset poisoning

- Forge a dataset:

- #samples to compromise an LLM : 250

Attacking an ML system

- adversarial ex: adversarial attacks to fool a system
- by prompting:
  - remove safe guards
  - hidden prompts: white prompt on white background

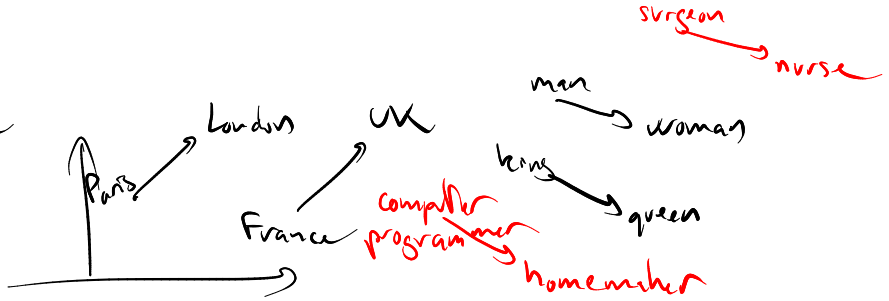
[AgenticAI

Dataset contamination

- train on training set
  test on unseen data — LLM
  ↑ how to get never-publicly-released data?
- internet is full of AI-generated content
  → data quality?!

Fairness

intro : might be subtle
word2vec



↳ upgrade training set
→ build better representation
more fair

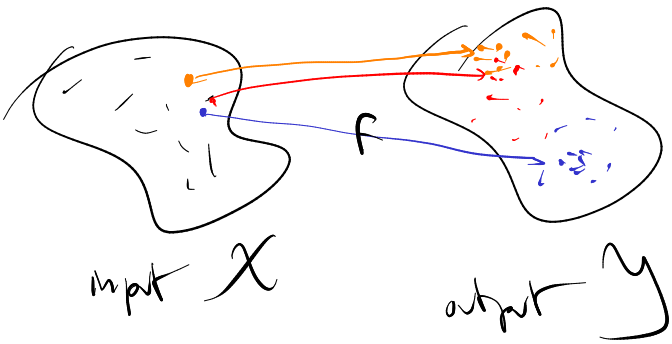
Def° 1 : fairness by unawareness

- case of matching CV with job offers
→ spot sensitive features : gender, ethnicity ...
↳ remove them from data
→ not sufficient : can use build proxies
ex : H/F : Stas gender ↔ dance

Def° 2 : fairness by awareness

Cynthia Dwork 2012

stochastic predictions



1) which metrics ?
$d_{\mathcal{D}(Y)}$, $d_{\mathcal{X}}$

x, x'

$$d_{\mathcal{Y}}\left(\mathcal{D}_p(x), \mathcal{D}_p(x')\right) \leq d_{\mathcal{X}}(x, x')$$ Fair

Def° 3 Equal opportunity / ε-fairness — group-based

- inputs (X, A) ← "sensitive" attribute
- Binary outcome variable: Y $Y=1$ = success (hired)
- predict: $\hat{Y}$

Equal opportunity:

$$\forall a, a', \quad p(\hat{Y}=1 \mid A=a, Y=1) = p(\hat{Y}=1 \mid A=a', Y=1)$$

ε-fairness: approximately: $|p(\cdot=\cdot) - p(\cdot=\cdot)| < \varepsilon$

↳ 3 similar def° but not exactly identical ⇒ not compatible

Def° 4: causality (counterfactual fairness)

$$(X, A) \rightarrow \hat{Y}$$
$$(X, A') \rightarrow ?$$

Algorithms: 3 types

- before training: pre-process data to remove all sensitive inform°
  ↳ Inform° bottleneck concepts:

$\max_{\phi} I(X,Z) - I(A,Z)$

$z = \phi(x)$

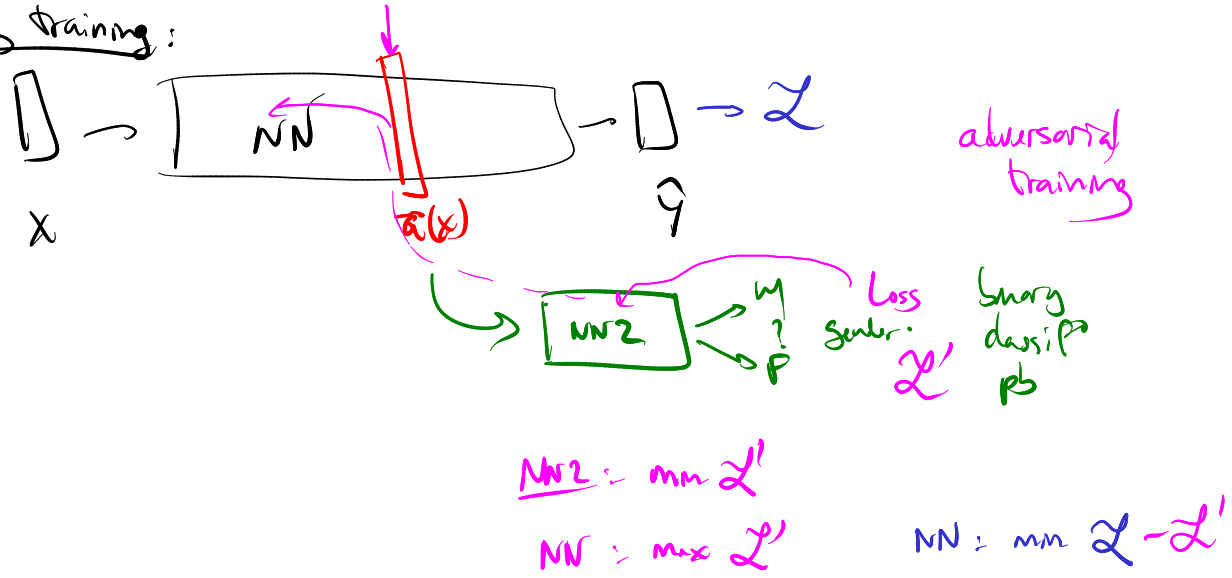
$I(X,Z)$: as high as possible: keep as much inform° as possible from X

$I(A,Z)$: as low as α (>0)

⚠ mutual information I: in practice, difficult to estimate

↳ solutions: HSIC — Hilbert-Schmidt Independence Criterion

- during training:



NN2 : $\min \mathcal{L}'$

NN : $\max \mathcal{L}'$ NN : $\min \mathcal{L} - \mathcal{L}'$

- after training:

- NN is biased : $\hat{y}(x) > 0 \rightarrow$ hire
  $< 0 \rightarrow$ don't

$g(x) > \tau_M$ $g(x) > \tau_F$ positive discr.

- ⚠ decision is taken as a f° of the sensitive attribute

Differential privacy

Why care?

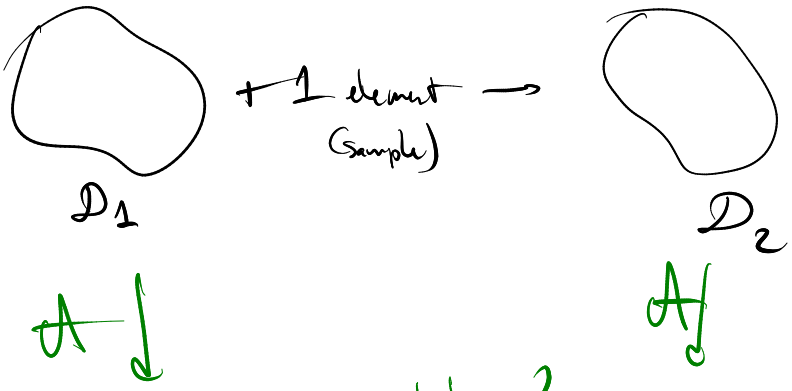
- anonymize?
  Netflix prize 2007 : recommandt° systems
  ↳ IMDB : cross-ref
- 87% of US citizens : can be identified from - birth date
  - gender
  - zipcode
  ↳ 1997 : group insurance medical data
  ↳ combine with the voter roll database
  ↳ re-identify
  ↳ president record release of that commission
- DNA

What if no dataset sharing?

Queries on databases

- statistical queries: → average salary in the company? + noise <> # queries
  → +1 + noise
  → exact knowledge of newcomer's salary

ε-differential privacy : Cynthia Dwork 2006 → Gödel prize 2017



same predictions?

Stochastic setting

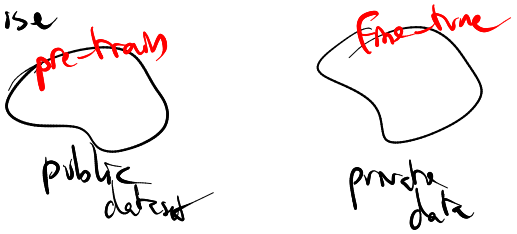
$$\forall S, \text{subsets of } \mathrm{Im}(\mathcal{A}) \text{ (of outputs)}: \quad p\left(\mathcal{A}(\mathcal{D}_1) \in S\right) \leq e^{\varepsilon}\, p\left(\mathcal{A}(\mathcal{D}_2) \in S\right) + \delta$$

e.g. outputs $\in \mathbb{R}$ $\quad S = [a, b]$

$\Leftrightarrow$ $\mathcal{A}$ is said to be (ε, δ)-differential private

Ensuring privacy

→ add noise
→



Limit # calls to private data + noise

DP-SGD :

→ $\nabla_\theta \mathcal{L}(x_i)$: clip the $\|\nabla\|$ to mitigate the influence of that sample

minibatch x_1 x_2 !

→ ∇ on minibatch $\left(\sum_{x_i} \text{clipped } \nabla\right)$ + noise $\mathcal{N}(0, \sigma)$ ↑ guarantees theory $\nabla \hookrightarrow \varepsilon$

Types of privacy attacks:

- membership inference: x_0 you know $\in ?$ $\mathcal{D}_{\text{Training}}$
- attribute " : know some features of some sample
  ↳ infer other features
- training data reconstruction:
  NN → prior knowledge over $\mathcal{D}$ → rebuild part of $\mathcal{D}$