Interpretability – part 2

Societal impact & approaches

ex: medical diagnosis

"Weapons of maths destruction" by Cathy O'Neil

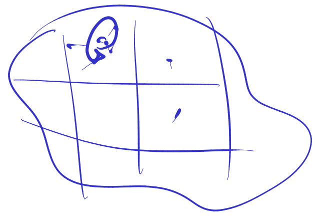
→ black box software (at large scale) for important matters

- ↳ hiring — jobs – CV
- ↳ firing — (bis school)
- ↳ loans — (bank)

→ behaving arbitrarily / stochastic
→ no feedback / questioning possible
→ large scale ⇒ arbitrariness nightmare
→ illegal criteria / proxy

ex: COMPAS: recidivism prediction
↳ high bias

· self-reinforcement (self-fulfilling prophecy): police patrols



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

Be responsible & careful

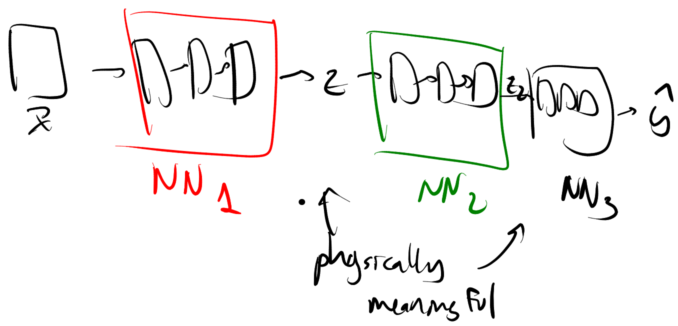
⇒ AI ethics → Montreal declaration for responsible AI

"FAT"-ML: Fairness, Accountability & Transparency

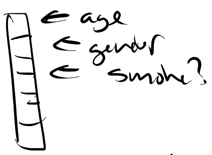
Interpretability by design

"X-AI"
Explainable

cf website

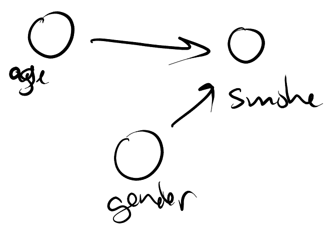


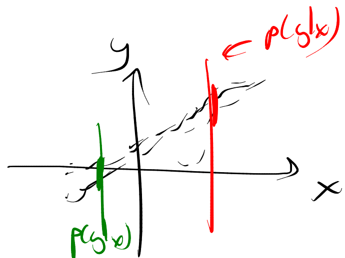
Causality



$x = 1$ file

~~Correlations~~ ? Causation

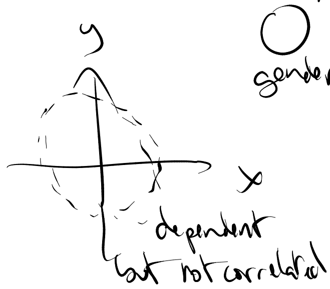




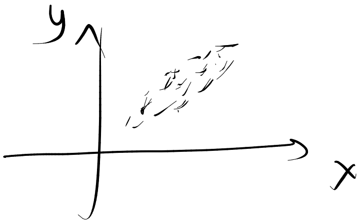
$E[(x-\bar{x})(y-\bar{y})]$

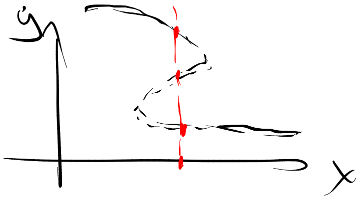
$x^2+y^2=1$

$y = \pm\sqrt{1-x^2}$



independent ⇒ correlated
$p(y|x) = p(y)$
$y \perp x$





$x = F(y)$ simpler

$y = g(x)$

Occam's razor

$(Y, X) \rightarrow$ $X \perp Y$ X, Y, Z

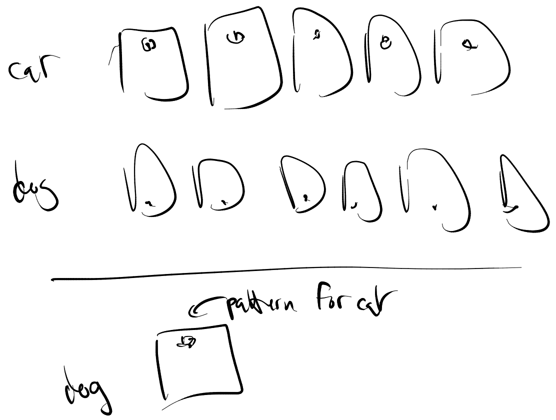
X — Y

$X \rightarrow Y$

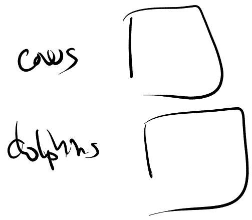
$X \leftarrow Y$

Issues related to datasets

Dataset poisoning



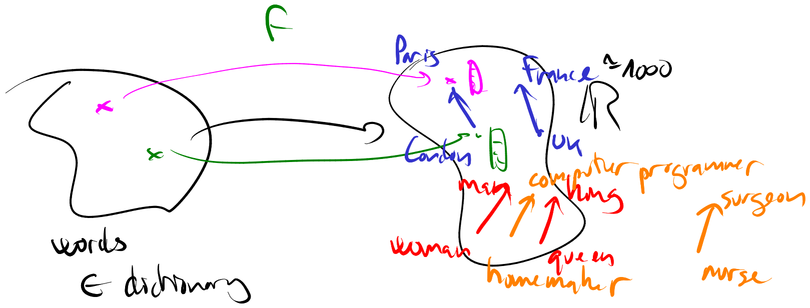
easy to detect for a ML pipeline



Fairness

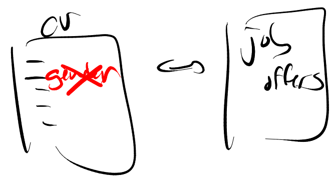
Introduct°

word2vec :

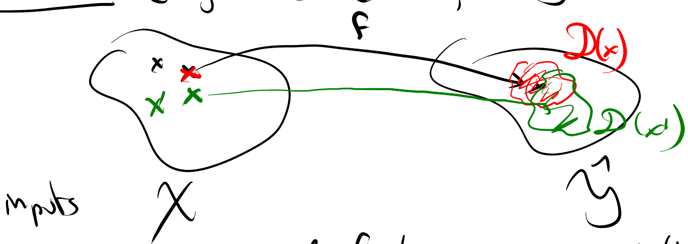


Definition 1 :

- simplistic : unawareness
  → do not include sensitive features



- awareness [Cynthia Dwork & al, 2012]



F: stochastic

$d_D(D(x), D(x')) \leq d_X(x, x')$

⚠ need relevant metric

eg for d_D : - Kullback-Leibler (KL)
- optimal transport
- MMD

Def° 2: Equal opportunity / ε-fairness

- input: (X, A) ⟶ output: Y : desired, $\hat{Y}$: predicted

1: hired
0: not hired

↑ sensitive attribute (ex: ethnicity)

ex: being hired

Equal opportunity:

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

⚠ group-based def°

⚠ statistics

Relax: ε-fairness:

$$\left| p(\hat{Y}=1 \mid A=a, Y=1) - p(\hat{Y}=1 \mid A=a', Y=1) \right| < \varepsilon$$

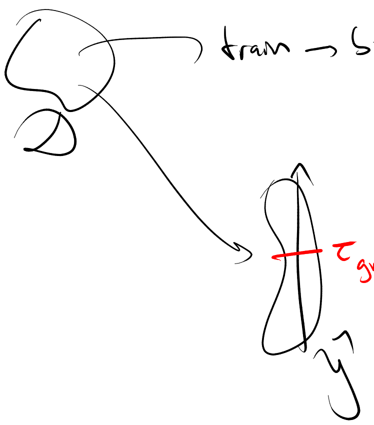
Def° 3: group-based

same distrib° of outputs / errors

Algorithms

fairness / accuracy trade-off

- [after training]



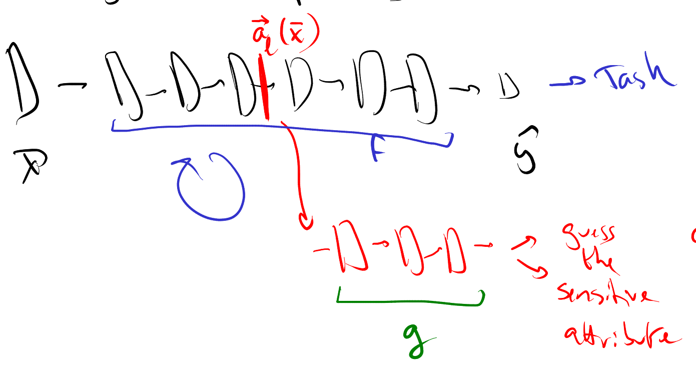
train → bias → debias at post-processing

(easy for group-based fairness)

⇒ one threshold per group

⚠ need to know the group
i.e. the sensitive attribute

- [during training] while optimizing



Adversarial training

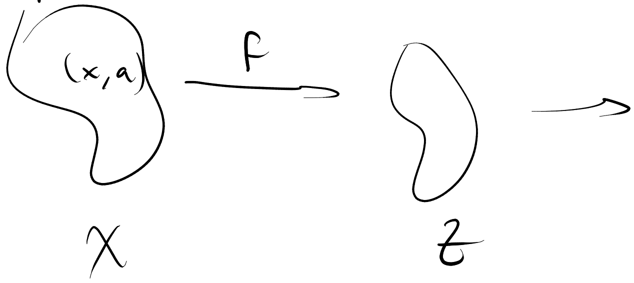
classificat° task

g :

f: make g fail

↳ criterion

[pre-processing] Information Bottleneck

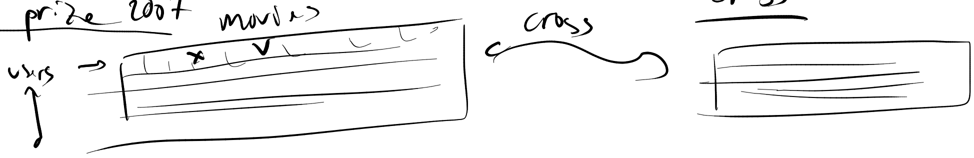


$\max I(X, Z)$ mutual information
while
$\min I(A, Z)$

Differential privacy

Why care about privacy?

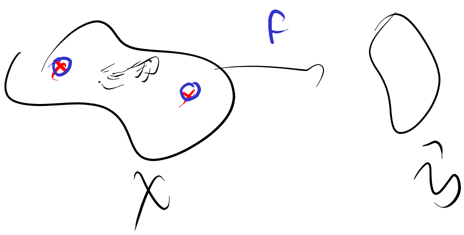
- Netflix prize 2007



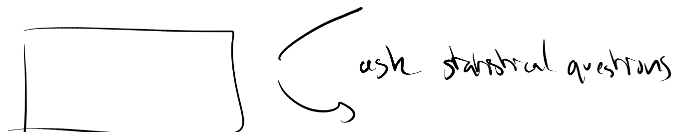
- Electricity consumption
- 87% of US citizens : identifiable from
  - birth date
  - gender
  - zip code

Group Insurance ⟷ voter roll database

If no data sharing



Queries on a database



private database

eg: wages in a company

1 person join: N persons ⟶ average s
$N+1$ persons ⟶ s'
} retrieve the salary of new employee

$(s' - s) \times N$

Principle:
1) add noise
2) bound the # of requests

ε-differentiable privacy [Dwork, 2006]

algorithm : A

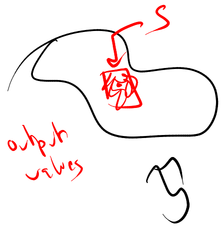
dataset : D_1

" : D_1 + one element : D_2

A has (ε, δ)-privacy if:

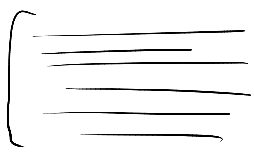
$\forall$ subsets S of $\mathrm{Im}(A)$, $\forall$ datasets D_1 & D_2 differing only by 1 element,

$$p(A(D_1) \in S) \leq \underbrace{e^{\varepsilon}}_{\simeq 1 \; (\varepsilon \sim 0)} p(A(D_2) \in S) + \underbrace{\delta}_{(\delta \sim 0)}$$



Gödel prize 2017

eg: DNA

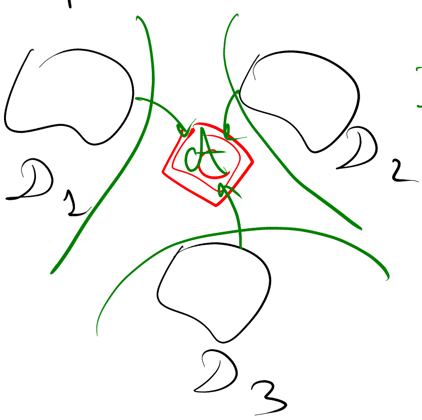


train a generative model

public dataset

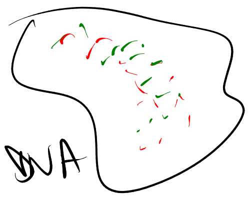
Federated learning

eg: hospitals & medical task



- no data transfer
- transfer parameters (encrypted)

AATS



$$\min_{x \in D_{real}} d(x', x)$$

generated