Responsible AI

Lecture 1

Who am I?

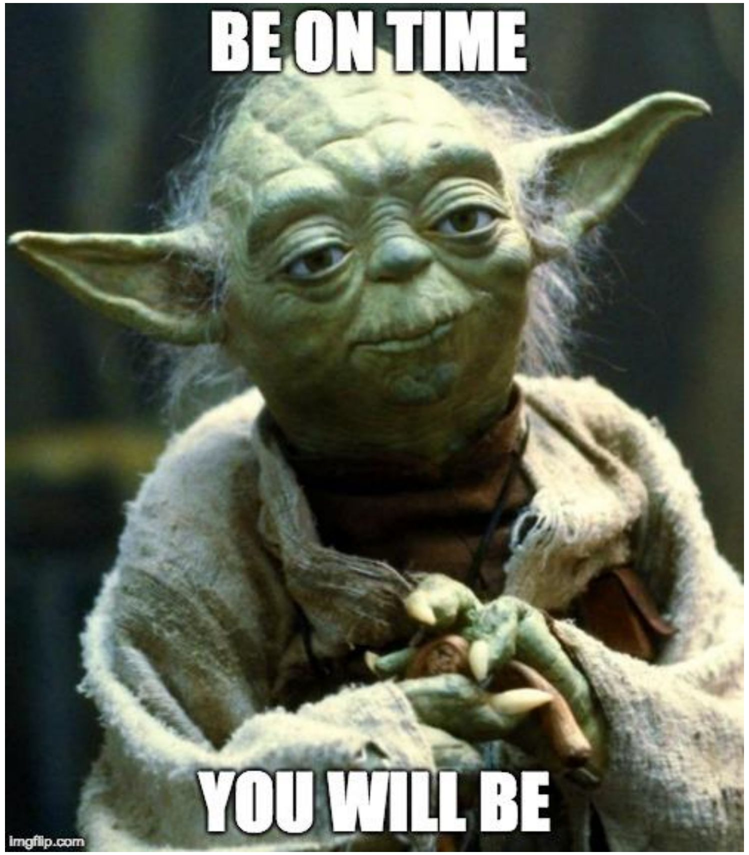
TL;DR

Lifeline 1/2

1. Responsible AI chapter 1-2, policy documents
2. Research paper, Ch 2, 9
3. Research paper, Ch 2, 9
   - Use case on Financial Service
   - Projects : Fairness and Proxy - Cosine similarity, Distance method, Mutual Information
   - Assessment Item – Class Participation
4. Chapters from book, research paper on fairness metrics
   - Lab - Python hands-on
   - Assessment Item - Use case on Financial Service
5. Ch 8 of RAI book, What-if by tensor flow
   - Lab
     - Python libraries and other XAI modules
     - Feature explanation - PDP, ALE, Sensitivity analysis
     - Model explanation - Global & local explanation, Morris sensitivity
   1. Assessment Item - Use case on Financial Service

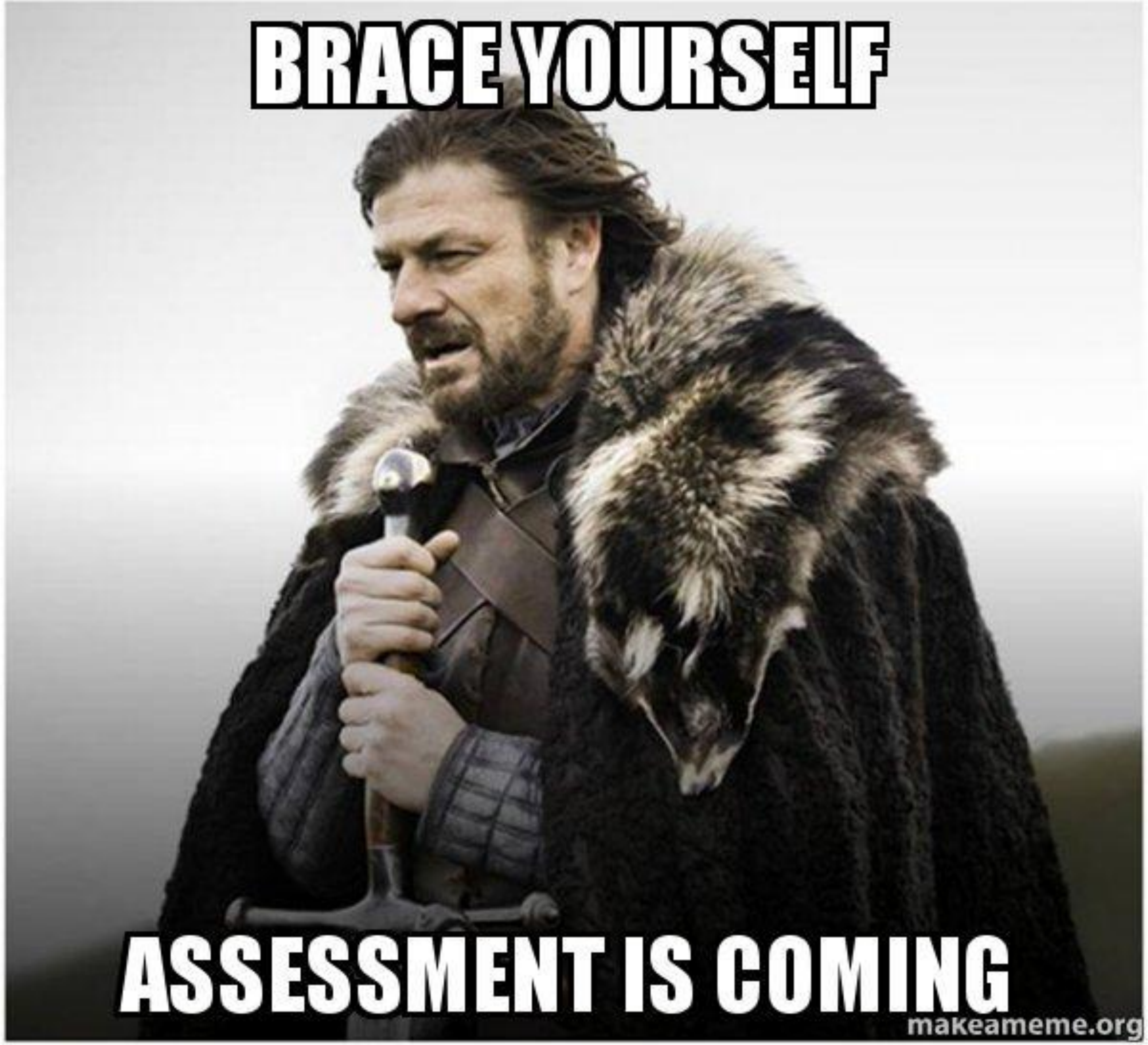
Lifeline 2/2

6. Chapter 5 to 7, Research paper on reweighting, exploring AIF 360 by IBM
    - Lab - Implementation using python
    - Assessment Item - Use case mid-term submission
7. Research paper on ACF
    - Lab - Python implementation
    - Assessment Item - Use case on Financial Service
8. Responsible AI course on Coursera
9. Paper on Reject option, AIF 360 GitHub
    - Lab - Python code on multiple use case
    - Assessment Item - Implementing Two other methods of ROC
10. Chapter 8, research paper on Monitoring, regulator papers
    - Lab
        - Python implementation
        - Data drift - Jensen–Shannon distance, Wasserstein distance
        - Concept drift - Brier score, Hierarchical Linear Four Rate (HLFR)
    - Assessment Item - FS use case, implementation of monitoring metric on Tableau

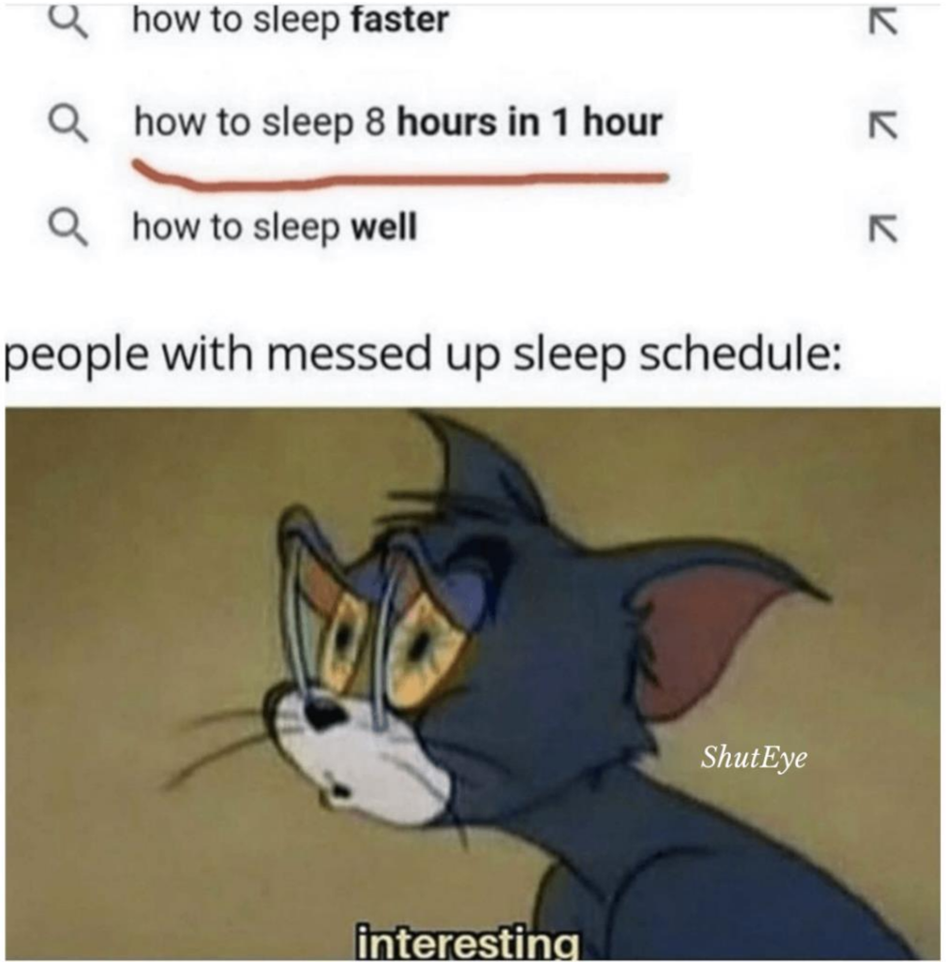

BE ON TIME
YOU WILL BE
imgflip.com


NO PHONES IN CLASS.
CALM DOWN, YOU'RE NOT THAT POPULAR.
memegenerator.net


IF YOU COULD JUST STOP IGNORING ME
THAT'D BE GREAT


BRACE YOURSELF
ASSESSMENT IS COMING
makeameme.org

What you need to take care



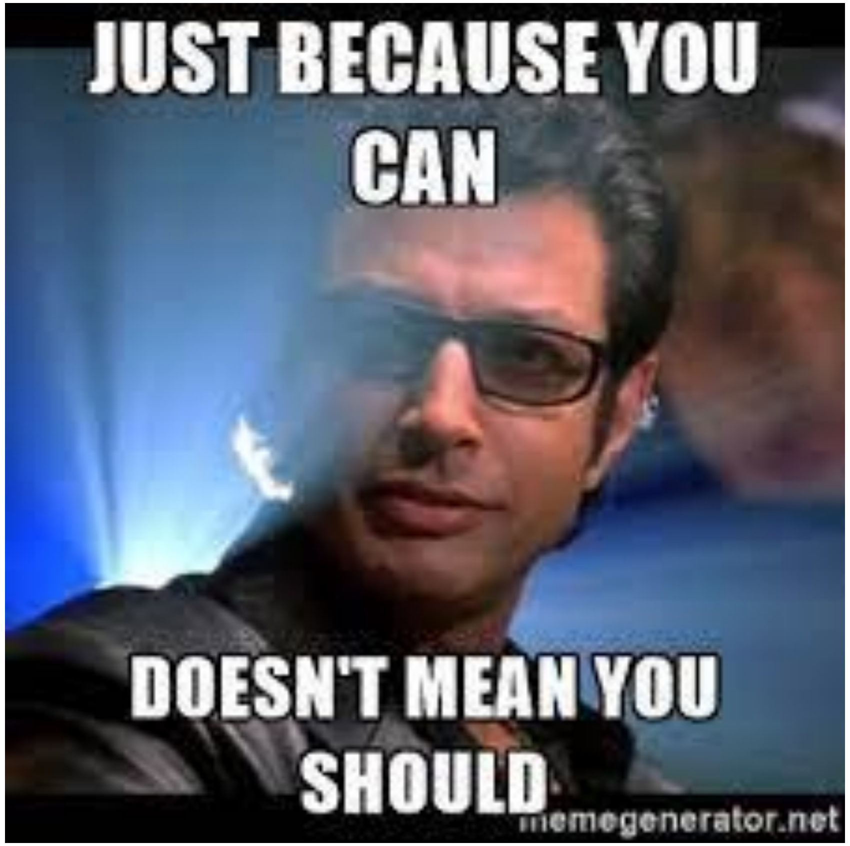
- Class Participation – 5%
- Class attendance – 10%
- 24 hour Assignments – 10%
- > 24 hours Assignments – 15%

Project:
- 1st submission – 10%
- 2nd Submission – 20%
- 3rd Submission + Presentation – 30%

Today

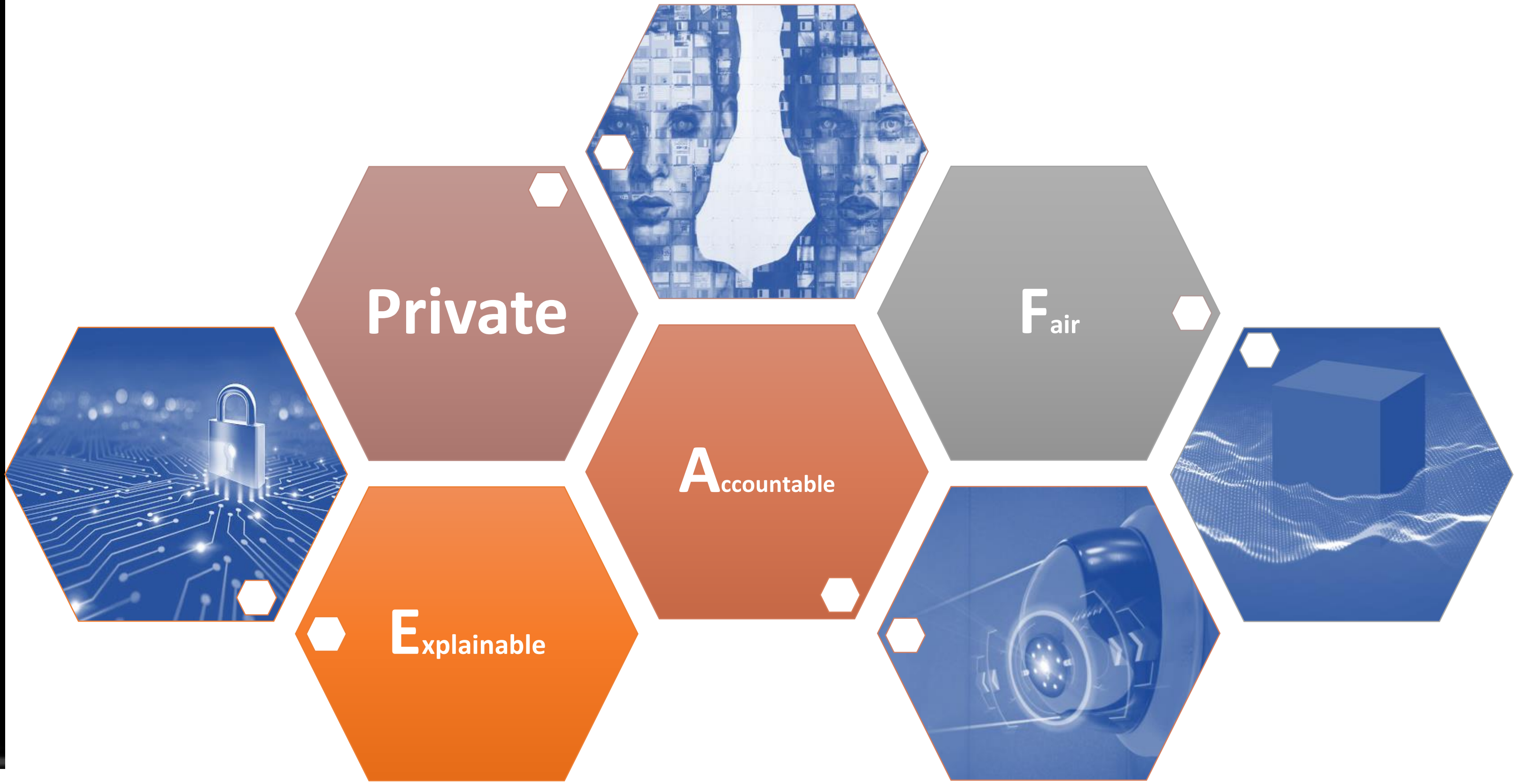
1. What is Responsible AI
   1. Facets of Responsible AI
   2. Fair AI
   3. Explainable AI
   4. Accountable AI
   5. Privacy

What?

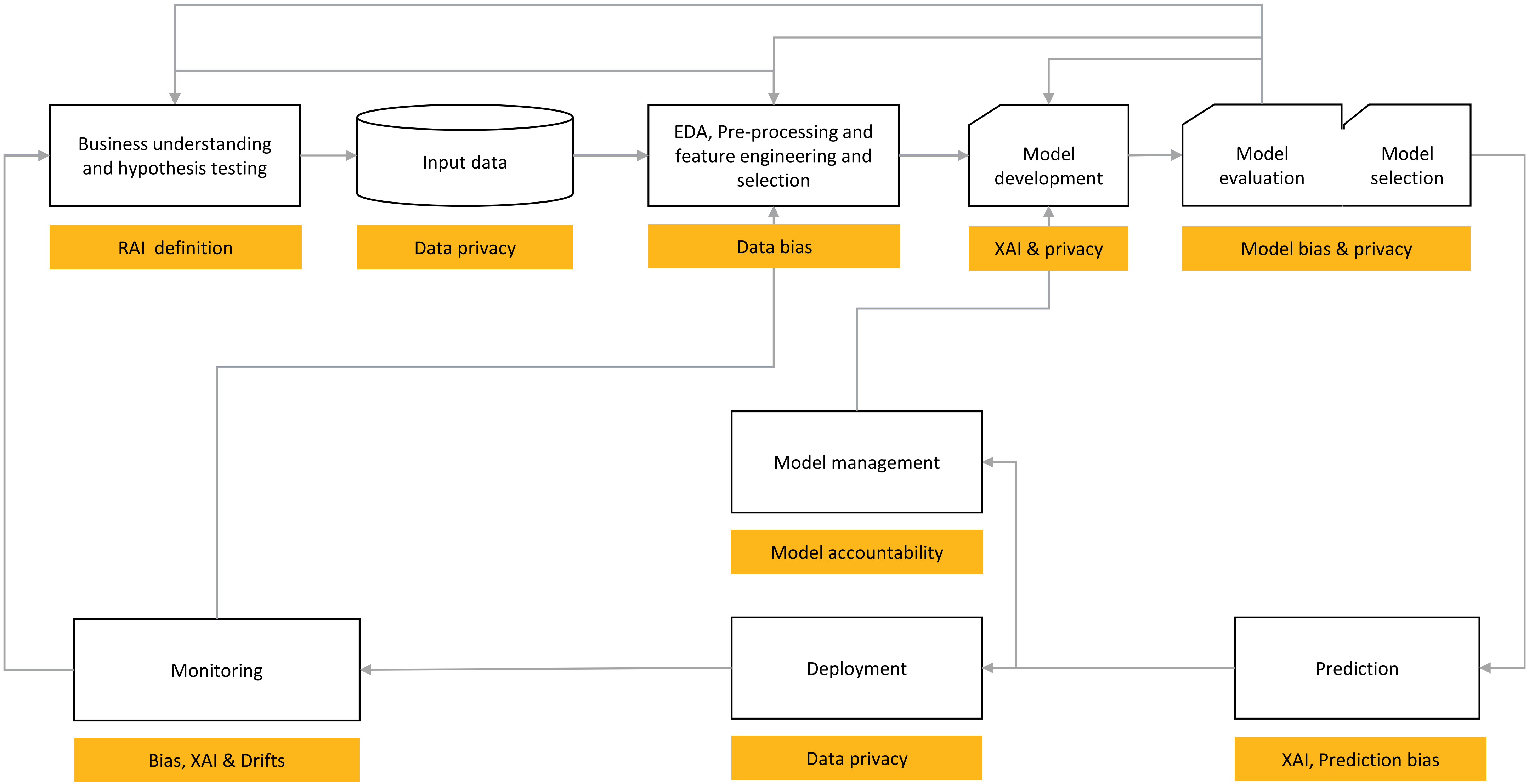

JUST BECAUSE YOU CAN
DOESN'T MEAN YOU SHOULD
memegenerator.net


WHAT IF I TOLD YOU
Without Ethical AI,
the Future of Humanity looks bleak.


DO YOU CARE TO SEE..
HOW DEEP THE
RABBIT HOLE GOES?


Private
Fair
Accountable
Explainable

Data science lifecycle with RAI



Why RAI

In Group of three tell three reason why RAI is:

- Very important
- Next big thing
- Why now

Responsible AI 2.0

[When a professional salesperson on LinkedIn doesn't exist]

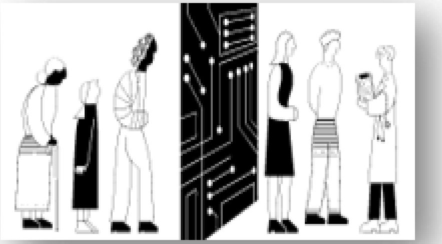
[When AI is both a threat and a boon to creatives]

[When a drug-developing AI invents 40,000 potentially lethal molecules in a few hours]

[Deepfake Democracy: When South Korean presidential candidate's avatar is a huge hit]

[Can algorithms predict a teenage pregnancy]

[What happens when an AI doctor misdiagnoses you?]



[Can we trust AI to be fair and inclusive?]

Should you reveal underlying AI systems?

Legal and regulatory constraints are still unclear – we are being challenged

Critical questions:

- Are these images inspired by existing artists
- Should the art be attributed to the original authors
- Is the consumer and the original creator informed about it's use?

Copyright Ambiguity: With AI now generating vast amounts of new data, copyright issues are emerging as content inspired from human authors are being recreated in creative ways

Source: *The Dawn of Generative AI*

When a drug-developing AI invents 40,000 potentially lethal molecules in a few hours

- While the capacity for **AI-assisted drug discovery** to have a positive impact is evident, the experiment proved that the opposite is also true
- Model put into a **"bad actor" mode** to seek out, rather than weed out toxicity
- We see **humans as key actors** in the process of an AI workflow with a firm moral and ethical 'don't-go-there' voice to intervene"
- Responsibly account for who's using the resources

The Verge Publication

Responsible AI Framework

Principles

NEW

Social well-being & planet-inclusive
Cost-benefit analyses for net positive on humans, society and planet

Privacy & Safety
Respect and protection from mis-use

Fairness & Equity
Ensure equitable outcomes

NEW

Robustness & Stability
Reliable, repeatable outcomes

Accountability
Humans responsible/governance

Transparency
All deserve white-box decision choices

Behaviors

Contestability
Newspaper test

Human-Centricity
Evaluate biases

Adaptability
Augment signals and update

NEW

Upskilling - Future of Work
Unlearning, relearning positively

Explainability
Decision factors traceable

NEW

Attribution
Copyright infringement and monetization

Enablers

Codebase
Plug-ins/Cloud agnostic APIs

Chief RAI
RAI diagnostic & Committees

Education
RAI Training Courses

Adoption
Nudge toolkit, RAI Certification

NEW

ESG diagnostic
Circular Design

NEW

Synthetic Data
Privacy by design

Fair AI

The impartial and just treatment or behaviour without favouritism or discrimination

… this happed a many months ago

Dave Edwards @dedwards93 · Nov 9, 2019
Replying to @dhh @AppleCard
As a former senior Apple employee, this @applecard issue is very disappointing to me. I feel betrayed. Apple is positioned as the good team in tech. And I believe they are. But this is an issue that they have to fix. They have to think different and be better.

Steve Wozniak
@stevewoz

I'm a current Apple employee and founder of the company and the same thing happened to us (10x) despite not having any separate assets or accounts. Some say the blame is on Goldman Sachs but the way Apple is attached, they should share responsibility.

3,101 7:06 AM - Nov 10, 2019

605 people are talking about this

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Steve Wozniak
@stevewoz

Replying to @dhh
The same thing happened to us. We have no separate bank accounts or credit cards or assets of any kind. We both have the same high limits on our cards, including our AmEx Centurion card. But 10x on the Apple Card.

184 6:58 AM - Nov 10, 2019

99 people are talking about this

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even though she has the higher credit score

DHH @dhh · Nov 8, 2019
The @AppleCard is such a [redacted] sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

Steve Wozniak
@stevewoz

The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

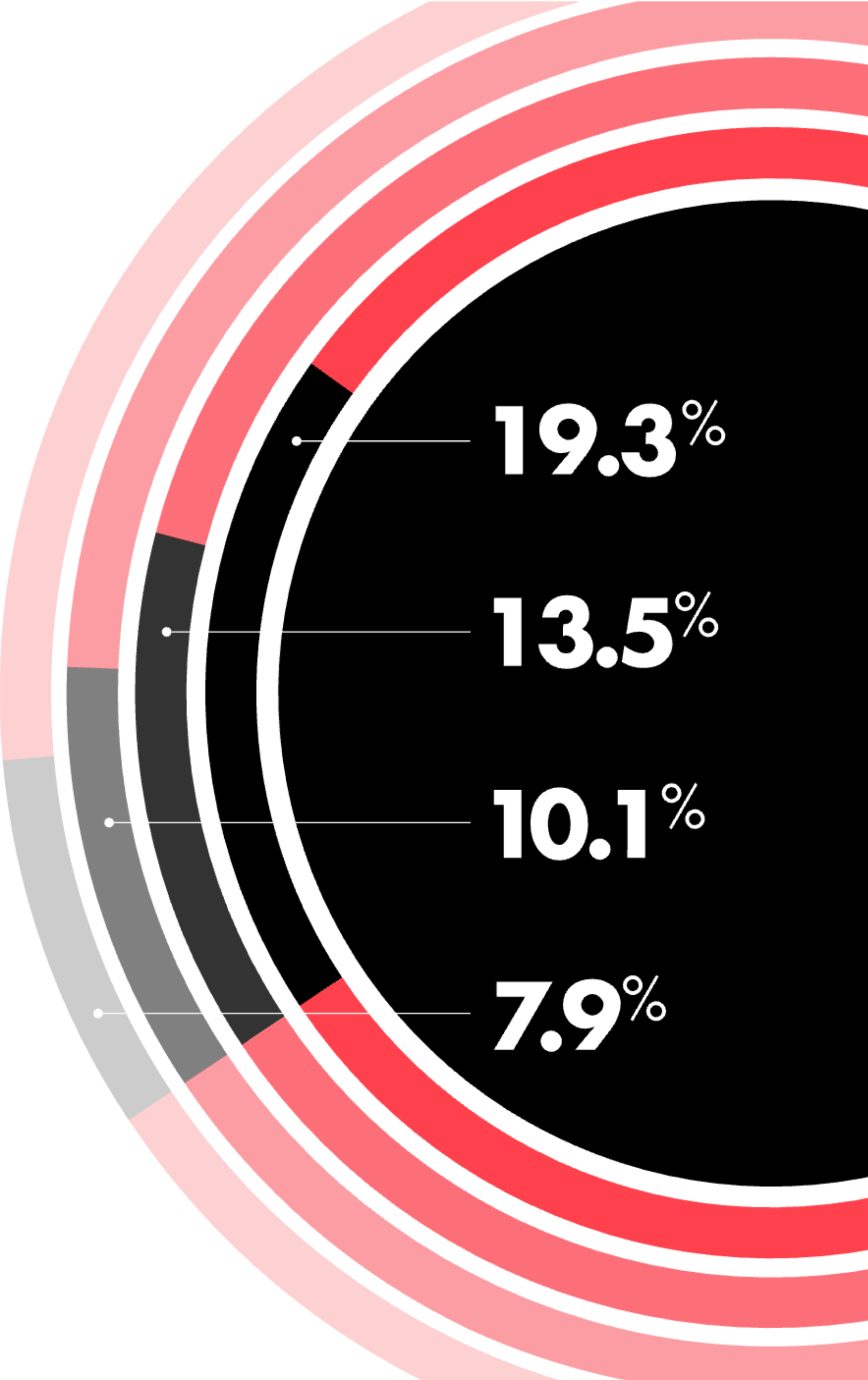
Share

Fair ML will underpin ethical practices

- Amazon’s AI algorithm **discriminatory job** selection biased towards men
- Google’s photo recognition AI led to **colored people being misidentified** as primates (Simonite, 2018)
- Google displayed adverts for a **higher paying jobs 1,852 times to the male** group and only 318 times to the female group (Post, 2015)
- Court intervened and asked a company to **stop using proxies for race**, to make hiring decisions (Ajunwa, 2015)
- Names and place of birth was used **to identify race or nationality** (Schwartz, 2019)
- Bank of America's Countrywide Financial business has agreed to pay a record fine of $335m (£214m) to settle discrimination charges when around 200,000 qualified African-American and Hispanic borrowers were charged with **higher rates solely because of their race or national origin**

Fair ML will underpin ethical practices while ensuring that regulations covering biases and interpretability are met

- 19.3% of African-American and 13.5% of Hispanic borrowers in the US were turned down for a conventional loan
- 10.1% of Asian applicants in the US were denied a conventional loan. By comparison, just 7.9% of white applicants were denied
- Bank of America's Countrywide Financial business has agreed to pay a record fine of $335m (£214m) to settle discrimination charges when around 200,000 qualified African-American and Hispanic borrowers were charged with higher rates solely because of their race or national origin



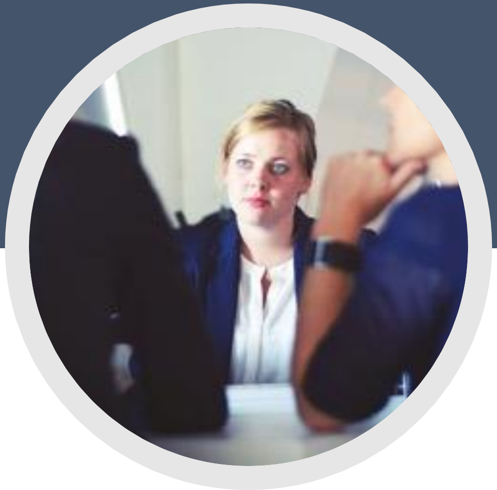
Bias is everywhere

1.

ML system used by a bank in India assigned higher weight to features related to income, resulting in systematic categorisation of women as less suitable for a mortgage loan because historically they earned less, while ignoring the fact that they have a better payment history

2.

ML system used for hiring decisions has favored men because women represented in the underlying data have historically been promoted less as compared to men



3.

Google displayed adverts for a higher paying jobs 1,852 times to the male group and only 318 times to the female group

Defining ‘Fairness’?

Fairness

- noun [U]
- **UK** /ˈfeə.nəs/

- The ability of an algorithm to treat various each group in the data without intrinsic bias:
- The model performance does not favor or discriminate against any group in the data.

Synonyms of fairness

- Equality of odds, disparate impact, parity ratio, equal opportunity, odds ratio

Words related to fairness

- Algorithmic fairness, data fairness, model accuracy, predictive parity, treatment equality

Antonyms of fairness

- Bias, discrimination, favor

Also:

- Trade off, threshold, fairness cost, fairness utility, predictive equality, fair calibration, equal error rate

Parity and disparity in ML algorithms

Color | Gender | Race | Age | Marital Status | Nationality

Bias in facial recognition

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%

Facial recognition based on gender and color shows varied accuracies

	Type I	Type II	Type III	Type IV	Type V	Type VI
Microsoft	1.7%	1.1%	3.3%	0%	23.2%	25.0%
FACE++	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%
IBM	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%

Facial recognition based on age and color (within a gender group) shows varied accuracies

Policy makers and regulators have laid down guidelines for the financial services industry

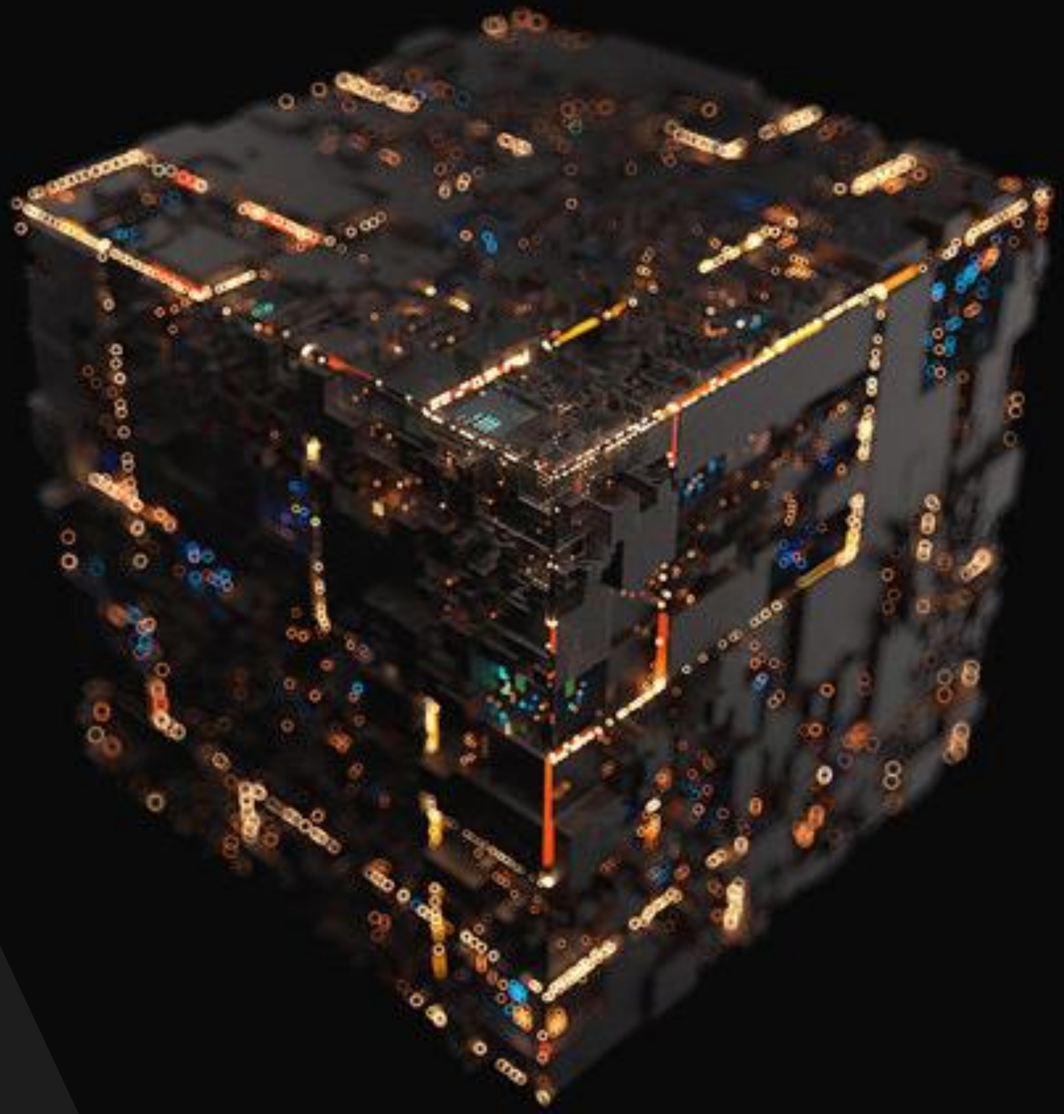
- **Communication from EU parliament** states, “The way in which AI systems are developed (e.g. algorithms’ programming) may also suffer from unfair bias. This could be counteracted by putting in place oversight processes...” and adds, “the way in which AI systems are developed may also suffer from bias”
- **Ethics guidelines for trustworthy AI by European Commission** stresses on, “Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness? Did you establish mechanisms to ensure fairness in your AI systems? Did you consider other potential mechanisms?”
- **A paper by EU commission** highlights how, “if algorithmic systems and/or their outcomes are biased, this may block equality of opportunity and/or outcome and systematically disadvantage certain social groups” and “that bias in algorithms can be discriminatory, where it disadvantages demographic groups with protected characteristics.”

The UK government also has come out with multiple papers highlighting the need for ethics in ML and AI

- **Centre for Data Ethics and Innovation (CDEI)** states that “decision-making processes that are driven by algorithms can share some of the same vulnerabilities as a human decision-making process”… “Another potential problem is that the complexities of algorithmic decision making can throw up unintended results. For example, while an employer might remove details of ethnicity before conducting recruitment sifting by algorithm, the system may use other data as a proxy for those characteristics – for example, postcodes that correlate closely with race.”
- In the interim report review into bias in algorithmic decision making by UK govt cites, “…a credit-scoring algorithm may rate consumers who routinely buy clothes in certain kinds of shop less favorably because the algorithm indicates that this is a good predictor that they are less likely to pay back loans. However, if these shops are largely selling women’s clothes, the algorithm will recommend fewer loans to women”.

Frameworks to help you develop interpretable and inclusive ML models and deploy them with confidence

Explainable AI



Explainability

In Group of three discuss what is XAI:

- What you mean
- What needs to be explained
- Why XAI

Accountable AI

Requires both the function of guiding action and the function of explanation by placing decisions in a broader context

Accountability

- To ensure models don’t get stale and decay
- To ensure your model is retrained on time on right data distribution
- To ensure you still using significant features
- To ensure your results today are similar to that during model development
- To ensure you considering right and topical data points (e.g., Pre vs Post Covid)
- To keep a check on error
- To alert in case a model goes rogue

Private AI

- As AI evolves, it magnifies the ability to use personal information in ways that can intrude on privacy interests

“In 2006, Netflix released a dataset containing ~100M movie ratings by ~500K users (1/8 of the Netflix user base)”

- Narayanan and Shmatikov, 2008

As of January 2022, the Netherlands come in hot at ***second place of top data breaches in Europe with 92,657 reports***

Dutch municipality Assen, an employee had sent a file containing ***530 persons' personal data to the wrong email***

An unsecured server resulted in the ***exposure of 3 terabyte of data*** *including airport employee records*

In 2014, hackers gained access to databases full of sensitive data via credentials of 3 employees of ebay

What is Privacy in Data?

Privacy Enhancing Techniques by technology

- Differential Privacy
- Federated Analysis

Whether or how data is shared with third parties.

Privacy by design

- Data minimization
- Purpose Limitation
- Security(technical organizational)

Misuse of Personal data

- sell the information on the dark web
- to carry out financial or identify fraud
- Selling of information to Marketing firms
- charge a ransom to safely return it to the victim.

How data is legally collected or stored.

Threats to Privacy:

- Malware
- Non-secure third-party services
- Data Breaches

Regulatory restrictions such as GDPR, HIPAA, GLBA, or CCPA.

More...

Sustainability and RAI

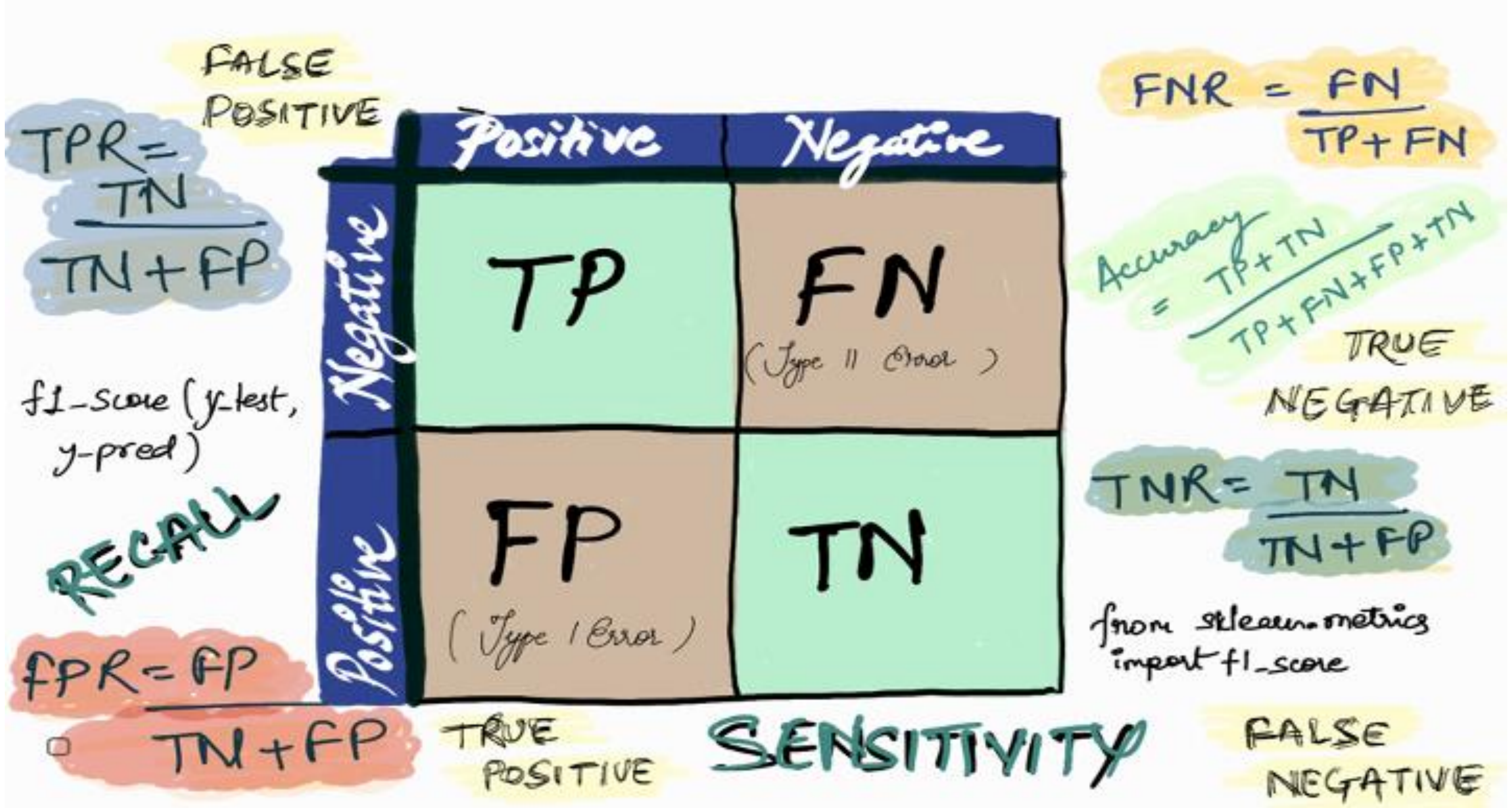
Metaverse and RAI

Blockchain and RAI

Gen AI and RAI

Recap and Review

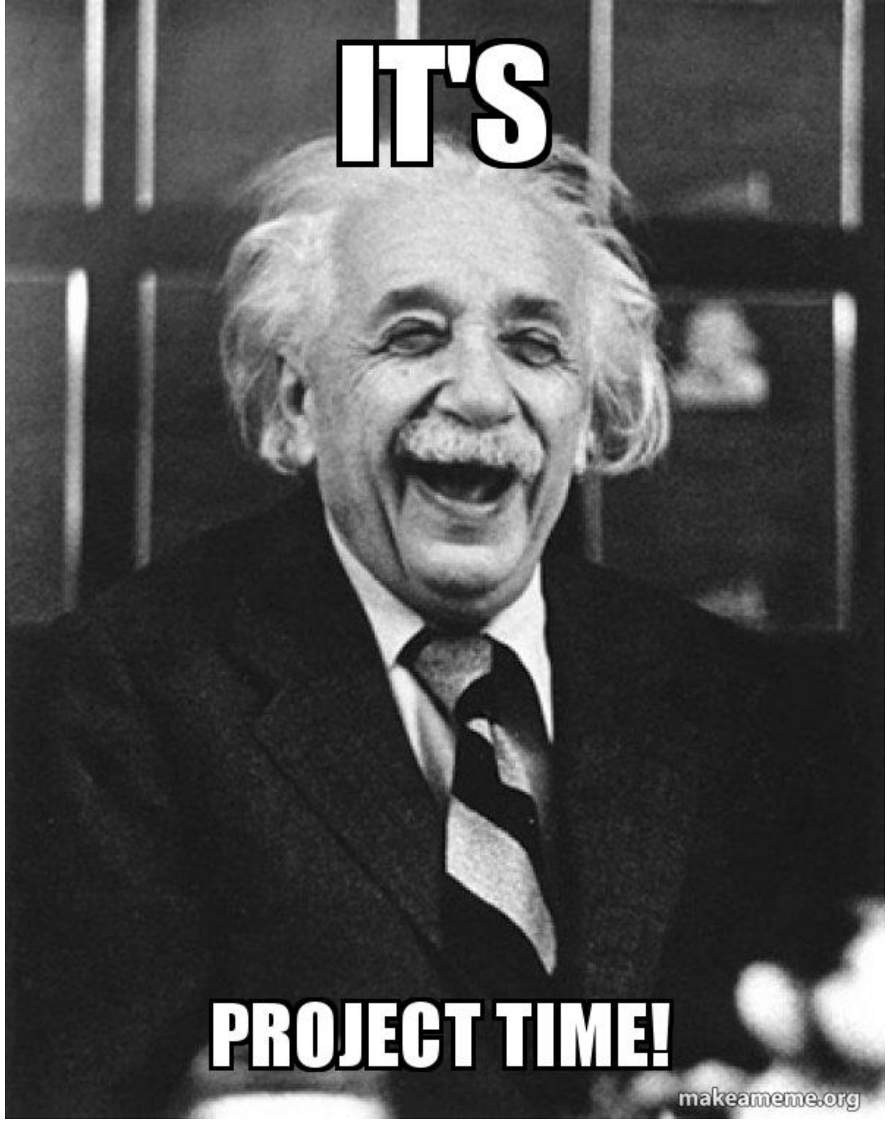
METRIC	FORMULA	ESSENCE
Recall, Sensitivity, TPR	TP / [FN +TP]	TP / P
FPR	FP / [TN + FP]	FP / N
Specificity, TNR	TN / [TN + FP]	TN / N OR, 1-FPR
Precision	TP / [TP + FP]	
FNR	FN / [FN + TP]	FN / P
Accuracy	TP + TN / [P + N]	



In group of total class / 10:

<u>One page</u> submission on '**Top 3 policies by government on Ethical and Responsible AI**'

Font: Times New Roman
Spacing: 1
Font Size: 10



Group Size: Total class / 9

Project Scope:

1. Choose a data set (any industry) that must contain PII data – **Get it approved by 3rd lecture**

2. Implement 2 Privacy algorithms – One based on class discussion, one based on your research

3. Investigate and report any proxy features

4. Submission **due before 5th lecture**

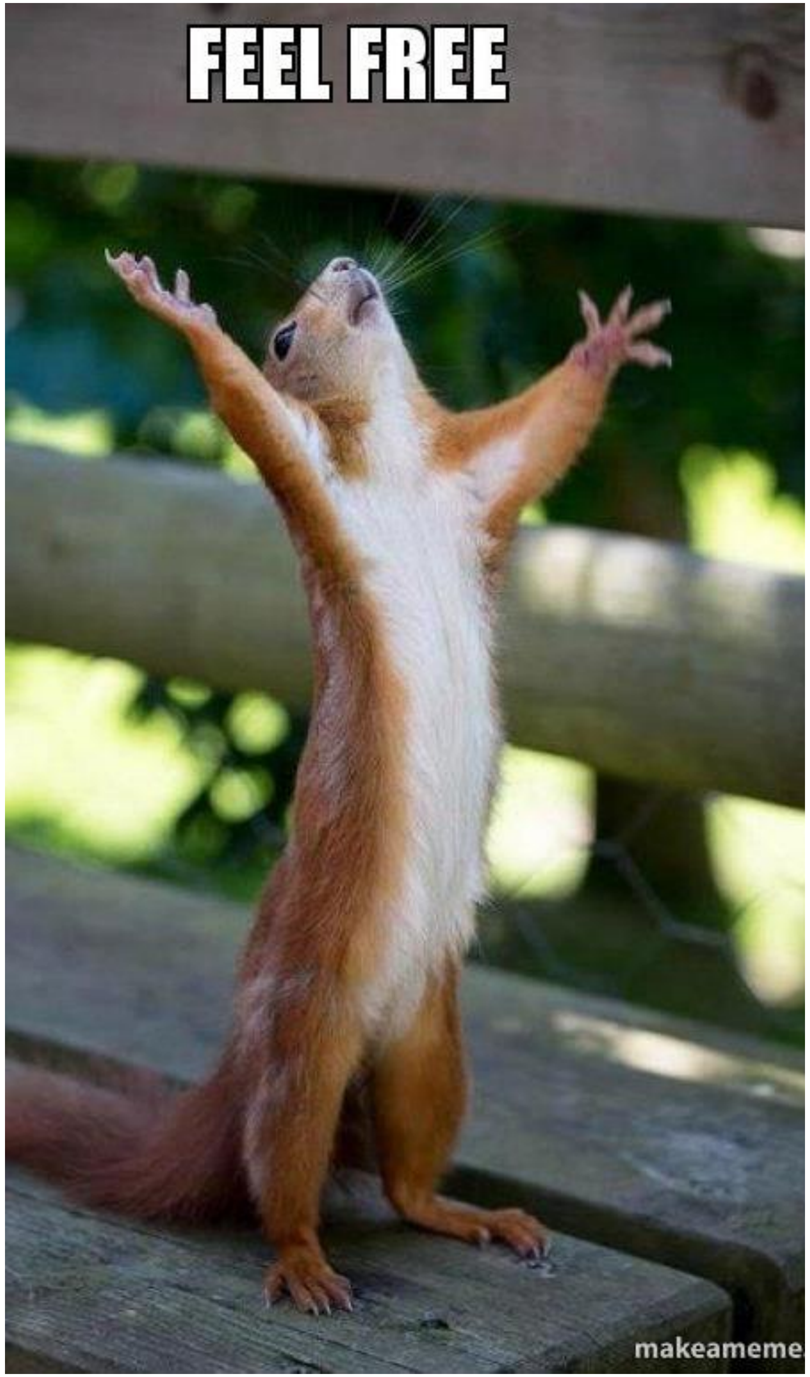
This is 1st of 3 parts of final project

Next

1. Introduction to privacy
2. Differential privacy
3. Differentially private ML algorithms
4. Introduction to discrimination in ML
5. Key parameters
6. Common accuracy metrics

Read chapter 2 and 9 from RAI book

Complete your Coursera Course


FEEL FREE
makeameme


YOU CAN CALL ME


DID YOU REALLY JUST SEND THAT EMAIL!


KNOCK, KNOCK!