22-12-2020, ONLINE

Data Science Challenges

COMMON PITFALLS & ETHICS

About me

Short Bio

10+ years working in Telecom in Marketing & Customer Value Management. Now leading Data Science team enabling NOS to make better decisions using the latest analytical artillery. Nuno holds one MSC in Electrotechnical Engineering (University of Coimbra), a Msc of Data Analytics (University of Porto) and is a PHD candidate of the MAP-i Doctoral Program in Computer Science **studying Fairness in ML.**

"Trustworthy AI" - PhD Motivation

New cycle of learning on a broad AI/ML scope;

Tech skills & tools refresh;

Business relevancy;

Social mission;

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Why are we here today?

At the end of session you should have new knowledge to start applying on your Data Science Projects

1. Improve Data Science Projects execution

Support your development team in planning DS activities making sure that critical path includes tasks to mitigate common pitfall risks;

2. Move Ethics out of the lecture theahtre and put it into practice

Apply ethical framework in your Data Science projects learning the concepts and identifying case studies patterns;

3. Promote Fairness discussions at Model development

Ask the right questions together with all the stakeholders involved in a Data Science Product development / maintenance by measuring and discussing the reason of biases and fairness on your models.

Managing expectations

1. Online interaction is challenging

A few work group sessions are included on the agenda and I'm counting on active participation through the session:)

2. Communication

To be inclusive concepts intuition will be provided preferably in English (and less in math notation);

3. Not Comprehensive lecture on tech DS

More focus given around predictive and prescriptive analytics – there are many more technical AI disciplines that can be used to create value in a DS Project;

4. Ethics in a Data Science Context

Ethics is a multi-disciplinary topic however in this presentation is discussed mostly in an applied algorithmic policy context.

Mini-breakout sessions

1	2	3	4
XXXX	XXXX	XXXX	XXXX
XXXX	XXXX	XXXX	XXXX
XXXX	XXXX	XXXX	XXXX

- We'll have a mini breakout session for each module to experiment with lectured concepts;
- Groups are supposedly mixing different teams to maximize diversity! (any improvements please let me know)
- We'll use Zoom breakout feature and I'll set a counter – there'll be a briefing before to make sure we're on same page;
- We'll do the discussions together using jamboard – there'll be a link provided for each challenge.
- Presentation (and related materials for challenges) in a share for your convenience:

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

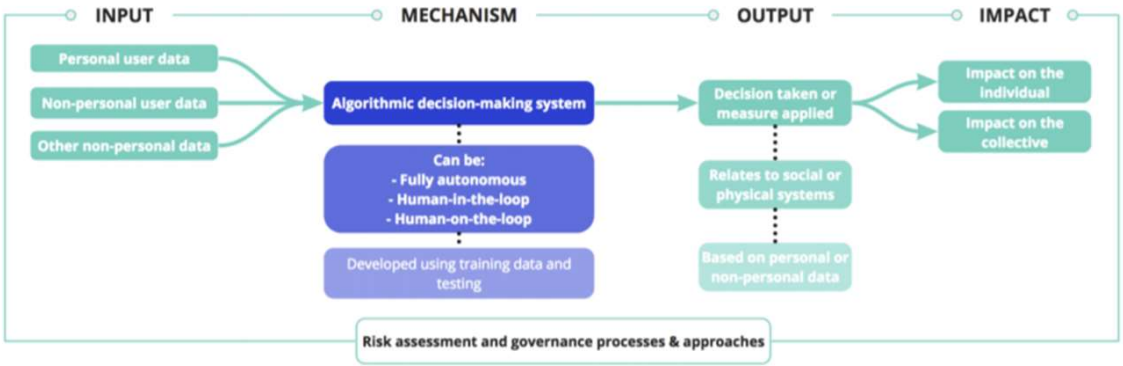
Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

[1] **algo:aware Raising awareness on algorithms** - https://actuary.eu/wp-content/uploads/2019/02/AlgoAware-State-of-the-Art-Report.pdf

Algorithmic Decision Making

Also known as predictive analytics, can be seen as statistical risk assessment, i.e., we aim to predict the probability of an outcome (a given event occurs in the future).

Working Definition



A software system – including its testing, training and input data, as well as associated governance processes – that, autonomously or with human involvement, takes decisions or applies measures relating to social or physical systems on the basis of personal or non-personal data, with impacts either at the individual or collective level.[1]

Applications

Scoring applications - e.g. credit, fraud, insurance, hiring, college admission, bail determination, preventive healthcare, etc

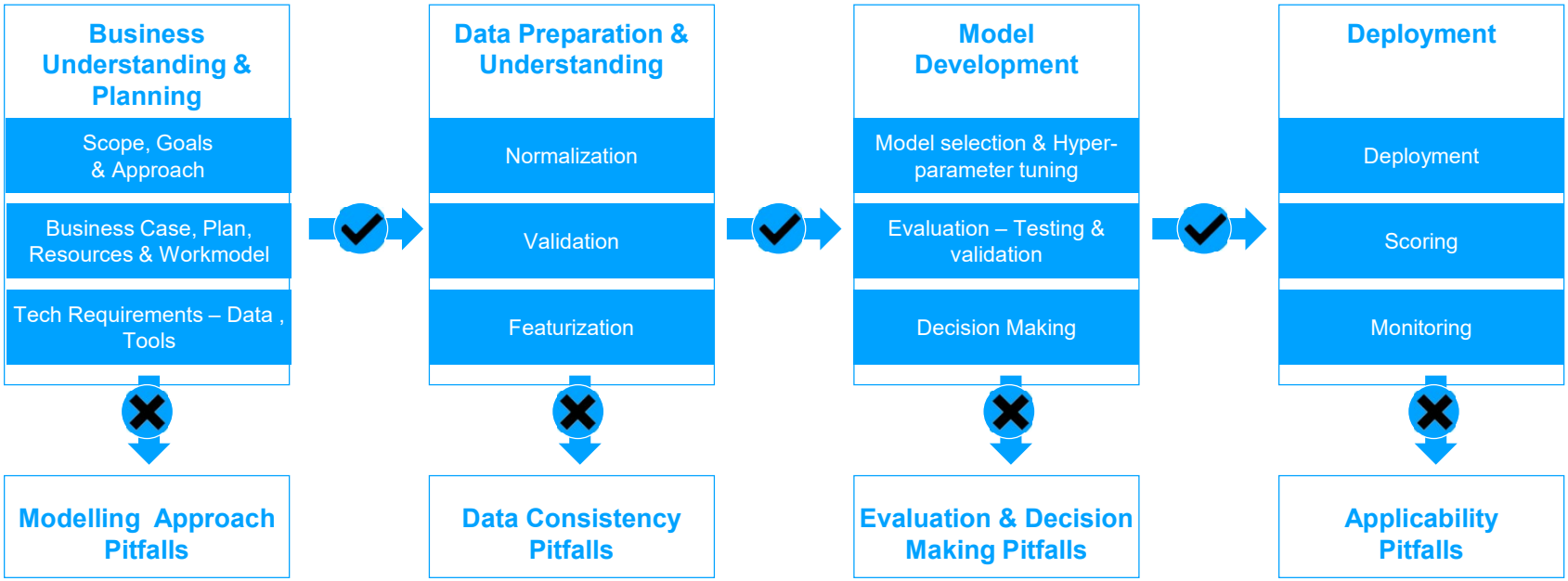
- Different types of **search engines**, including general, semantic, and meta search engines;
- **Aggregation applications**, such as news aggregators, which collect, categorise and regroup information from multiple sources into one single point of access;
- **Forecasting, profiling and recommendation applications**, including targeted advertisements, selection of recommended products or content, **personalised pricing and predictive policing**;
- **Content production applications** (e.g. algorithmic journalism)
- **Filtering and observation applications**, such as spam filters, malware filters, and filters for detecting illegal content in online environments and platforms.
- Other 'sense-making' applications, crunching data and drawing insights.

Team Data Science Process - https://docs.microsoft.com/pt-pt/azure/machine-learning/team-data-science-process/lifecycle-data

Common Pitfalls in Predictive Analytics

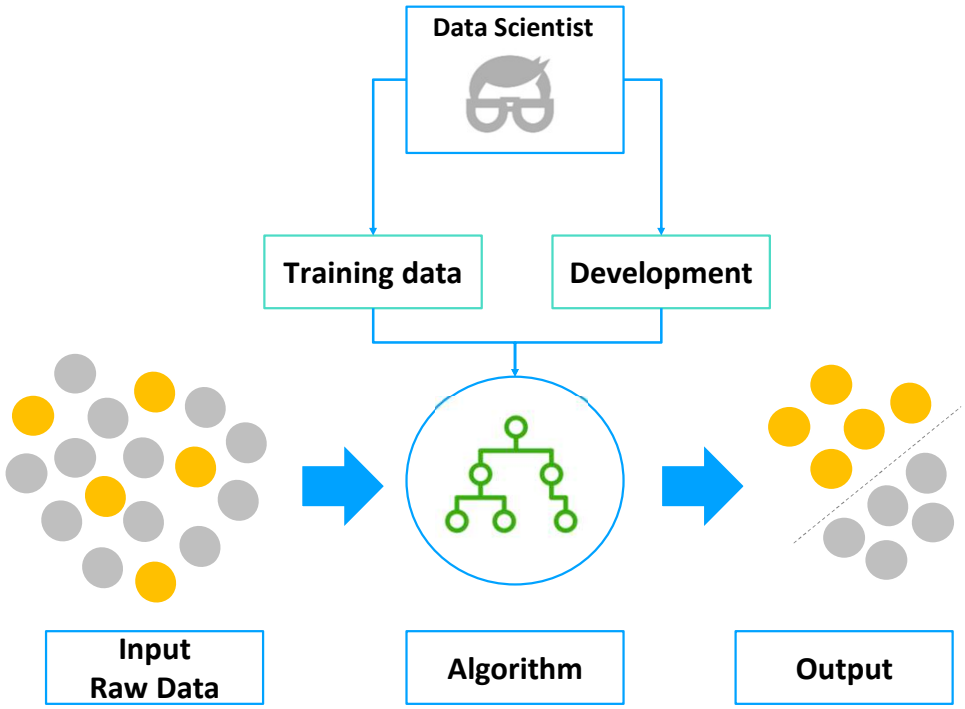
Data Science is a multidisciplinary field that may fail in unexpected ways during development or operational phase

Illustrative

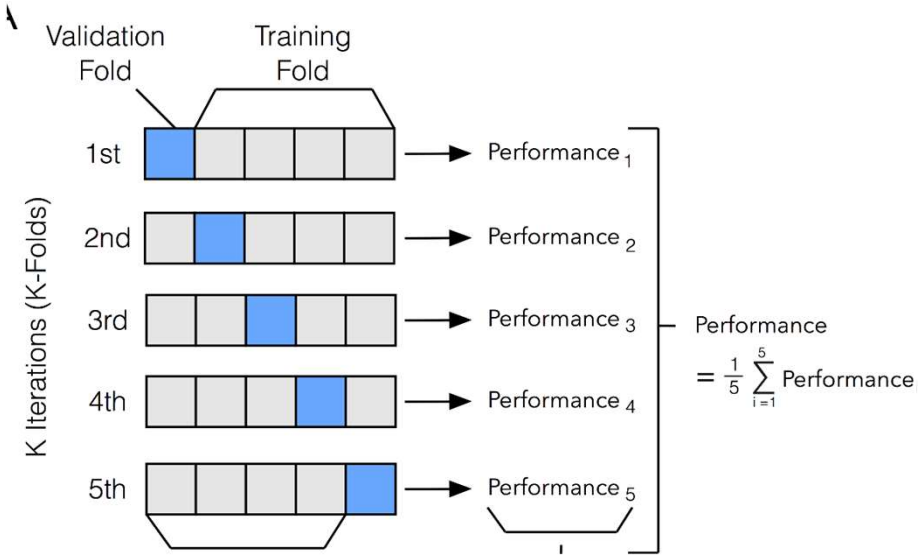


Classification Refresher

Global Process

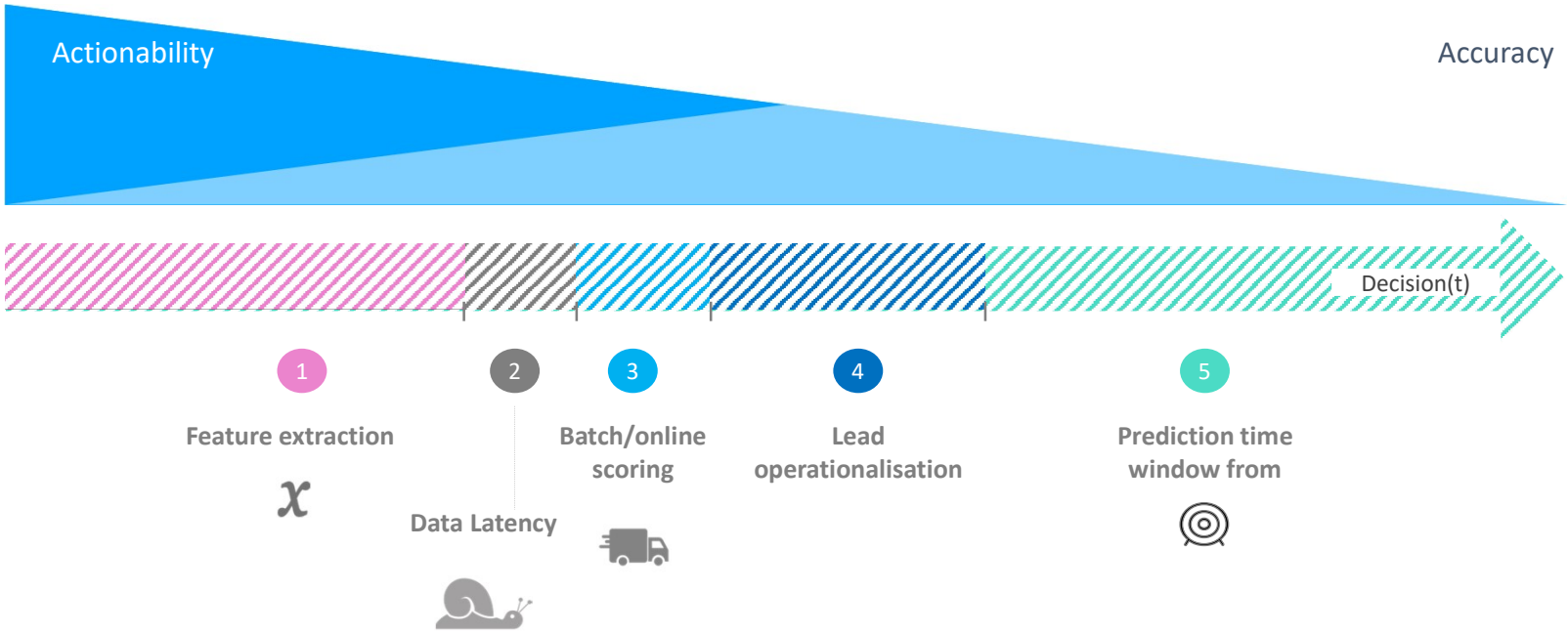


Learning (Training)



Classification Refresher

There are a few degrees of freedom that need to be discussed to set-up the modelling approach



Classification Refresher

Scenario Evaluation

Scenario 1

Scenario 2

Scenario N

Key definitions

Objective	Historic	Data Volume
Population	Latency	Quality
Target	Deployment type	Benefits
Features	Time2 Market	Costs

Algorithm Selection

+Prediction Accuracy — **+Explainability**

Neural Network

Hidden Markov Models

Gradient Boosted Trees

Support Vector Machines

Random Forest

Naive Bayes

Decision Tree

Logistic Regression

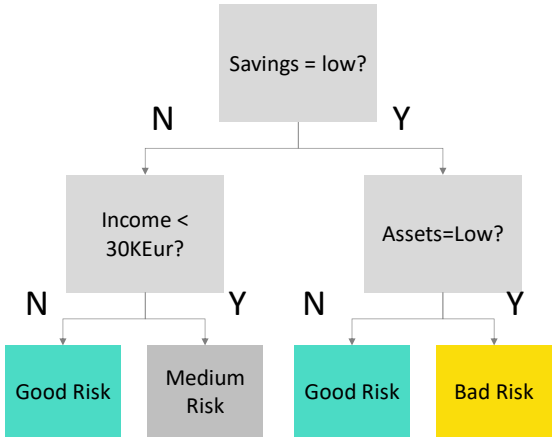
Nearest Neighbours

+ more than 200 Models available in programatic frameworks

To be detailed

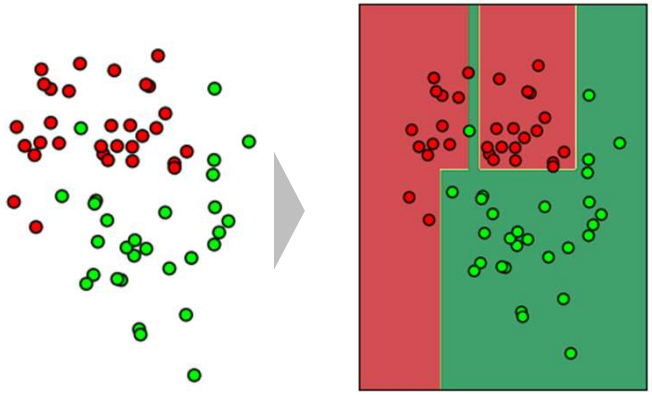
Classification Refresher: Decision Tree

A. Intuition



- Decision trees are built creating decision nodes that select the best variable at each step that best separates each class in the training dataset.

B. Decision boundary



- Space is divided in “rectangles” and each one is labelled with the more dominant k-class.
- Non-Linear relationships can be found on data.

C. Configuration & Strenghts/Weaknesses

Parametrization	Strenghts / Weaknesses
▪ Depth of tree ▪ Minimum samples Split ▪ Minimum samples leaf ▪ Maximum no features	(+) Highly Interpretable (+) Transparent; (+) little effort in data preparation (eg: can deal both continuous and discrete data); (-) less accurate while compared to other available predictors;

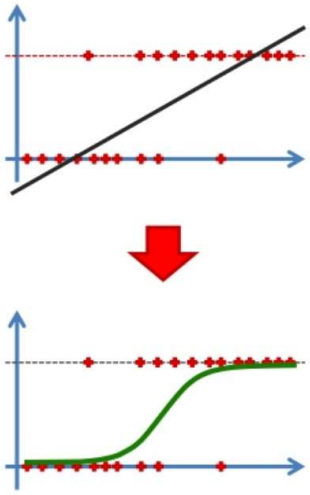
Classification Refresher: Logistic Regression

A. Intuition

y=bo + b1*x

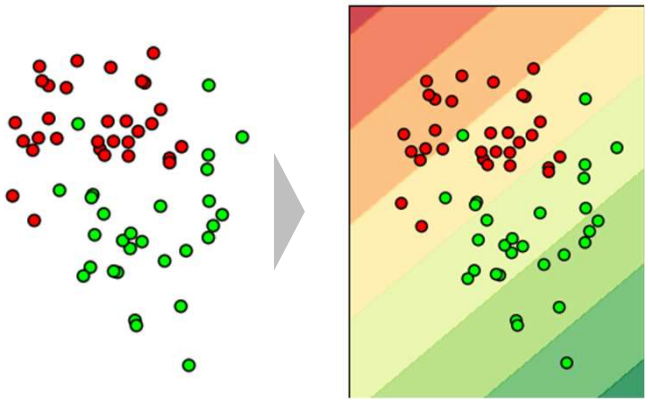
p=1/(1+e^-y)

ln(p/1-p)=
bo + b1*x



- We can think of Logistic Regression as na extension of the linear regression model but where the output is fitted through a sigmoid into a value (prob) ranging from 0 to 1.

B. Decision boundary



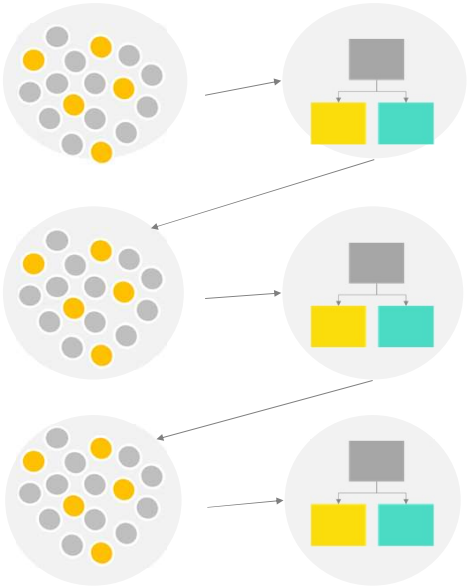
- The space is cut by a linear "hyperplane", which sets the decision criteria to classify the instances.

C. Configuration & Strenghts/Weaknesses

Parametrization	Strenghts / Weaknesses
▪ Regularization parameters to control trade-off between overfit and simplicity (underfit)	(+) simple and efficient; (+) low variance; (+) it provides probability score for observations; (-) doesn't handle large number of features; (-) requires feature processing (eg: non-linear features, categorical, multicollinearity, etc)

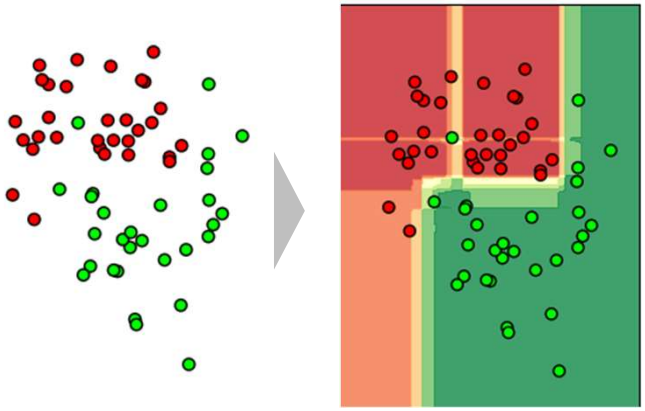
Classification Refresher: Gradient Boosted Trees

A. Intuition



- Data is fitted first in a simple tree model, then subsquent models focus on resolving classification errors;
- In the end all the predictors are combined through weights in a single model.

B. Decision boundary



- Decision boundaries are more complex since multiple trees are fitted on areas with prediction error is greater;

C. Configuration & Strenghts/Weaknesses

Parametrization	Strenghts / Weaknesses
▪ Number of trees; ▪ Learning Rate ▪ Maximum depth of tree; ▪ Min number samples to split tree node; ▪ Min number samples for a node; ▪ Max number features to consider for a split;	(+) accuracy – extensions of the boosting idea have a track record of winning Kaggle competitions; (+) similar with DT - little effort in data preparation (-) interpretability (-) takes more time to train since it's sequential (-) more prone to overfit

Classification Refresher: Evaluation Metrics

Evaluation metric is a number that measures the performance that your machine learning model when it comes to assigning observations to certain classes.

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Most Common Dervied Ratios - https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

Online Confusion Matrix - https://onlineconfusionmatrix.com/

Classification Refresher: Evaluation Metrics

Rank	Score	Label	Predict
1	0.997	1	1
2	0.993	1	1
3	0.986	1	1
4	0.982	1	1
5	0.971	0	0
6	0.965	1	0
7	0.964	0	0
8	0.961	0	0
9	0.953	0	0
10	0.932	1	0
11	0.918	0	0
12	0.873	0	0
13	0.854	0	0
14	0.839	0	0
15	0.777	0	0
16	0.723	0	0
17	0.634	0	0
18	0.512	0	0
19	0.487	0	0
20	0.473	0	0

Predicted Positive: 4

Predicted Negative: 16

Population

- Total Label Positives: 6
- Total Label Negatives: 14
- Prevalence 6/20=0.3

For threshold > 0.980 or top k=4:

		True Condition	
	Total Population	Positive	Negative
Predicted Condition	Positive	TP 4	FP 0
	Negative	FN 2	TN 14

True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$
False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$

- True Positive Rate (Recall) = 4/6 = 0.66
- False Positive Rate: 0/14=0

- False Negative Rate = 2/6=0.33
- True Negative Rate = 14/14=1.0
- Precision = 4/4 =1.0

Adapted from Pedro Saleiros's "Tutorial: Fairness in decision-making with AI: a practical guide & hands-on tutorial using Aequitas https://www.youtube.com/watch?v=yOR71zBm3Uc

Online Confusion Matrix - https://onlineconfusionmatrix.com/

Classification Refresher: Evaluation Metrics

Rank	Score	Label	Predict
1	0.997	1	1
2	0.993	1	1
3	0.986	1	1
4	0.982	1	1
5	0.971	0	1
6	0.965	1	1
7	0.964	0	1
8	0.961	0	1
9	0.953	0	1
10	0.932	1	1
11	0.918	0	0
12	0.873	0	0
13	0.854	0	0
14	0.839	0	0
15	0.777	0	0
16	0.723	0	0
17	0.634	0	0
18	0.512	0	0
19	0.487	0	0
20	0.473	0	0

Predicted Positive: 10

Predicted Negative: 10

Population

- Total Label Positives: 6
- Total Label Negatives: 14
- Prevalence 6/20=0.3

For threshold > 0.920 or top k=10:

		True Condition	
	Total Population	Positive	Negative
Predicted Condition	Positive	TP 6	FP 4
	Negative	FN 0	TN 10

True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = Σ True positive / Σ Condition positive	False positive rate (FPR), Fall-out, probability of false alarm = Σ False positive / Σ Condition negative
False negative rate (FNR), Miss rate = Σ False negative / Σ Condition positive	Specificity (SPC), Selectivity, True negative rate (TNR) = Σ True negative / Σ Condition negative

- True Positive Rate (Recall) = 6/6 = 1.0
- False Positive Rate: 4/14=0.29

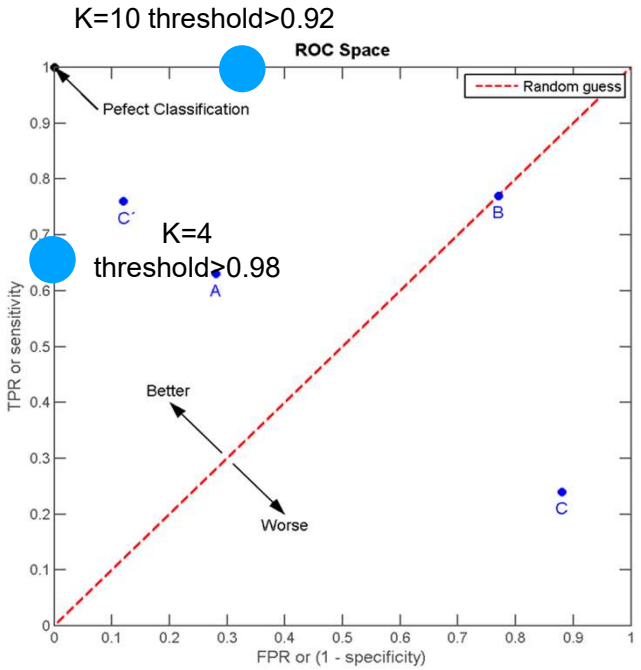
- False Negative Rate = 0/6=0
- True Negative Rate = 10/14=0.71
- Precision = 6/10 =0.6

Adapted from Pedro Saleiros's "Tutorial: Fairness in decision-making with AI: a practical guide & hands-on tutorial using Aequitas https://www.youtube.com/watch?v=yOR71zBm3Uc

Receiver Operating Characteristic Curves Demystified (in Python) - https://www.kdnuggets.com/2018/07/receiver-operating-characteristic-curves-demystified-python.html

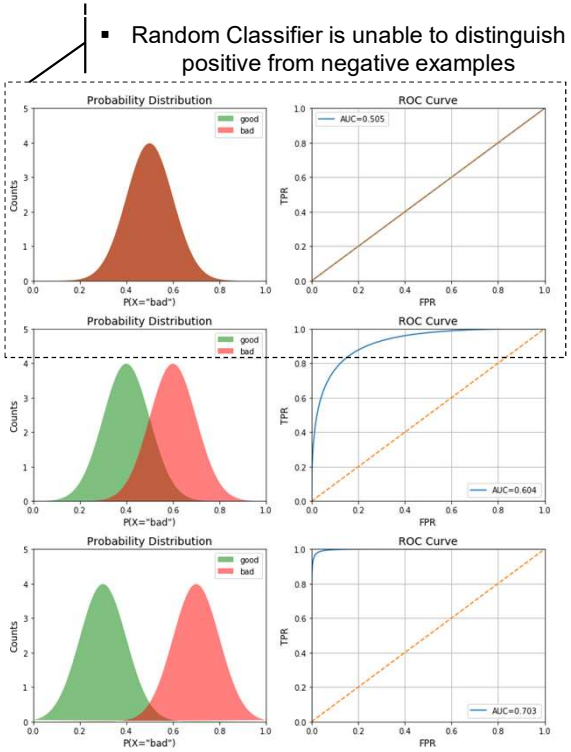
Classification Refresher: Evaluation Metrics

A. ROC COORDINATES



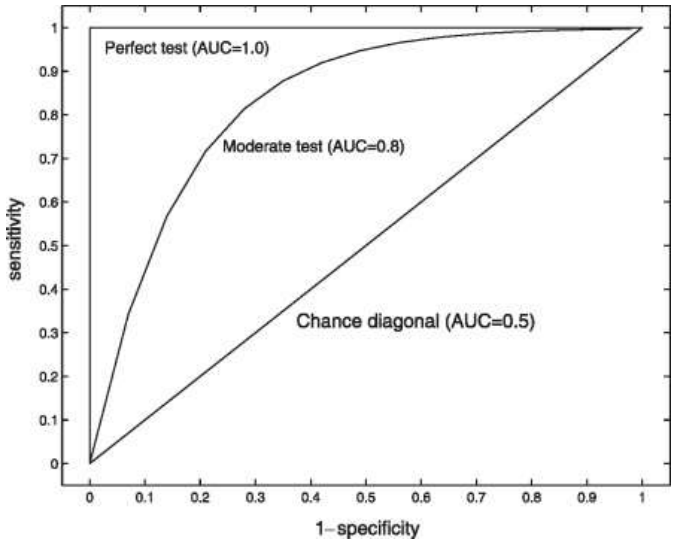
- To get ROC coordinates we define a threshold for 1/0 predictions and then we calculate the False Positive rate and Recall values for that threshold.

B. ROC PROFILES



- ROC curve is an analytical tool to assess model performance (if your dataset is not highly unbalanced)

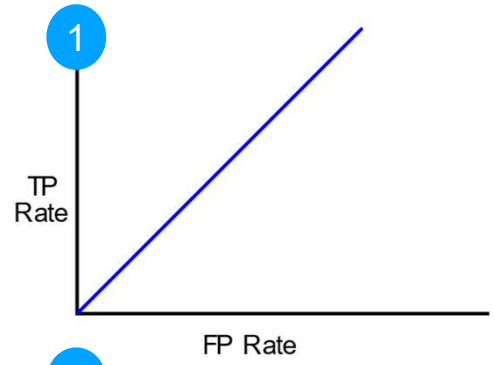
C. AUC

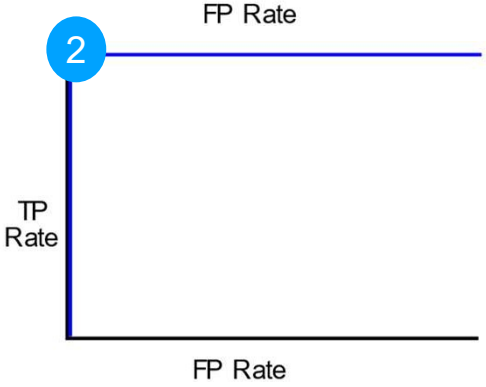


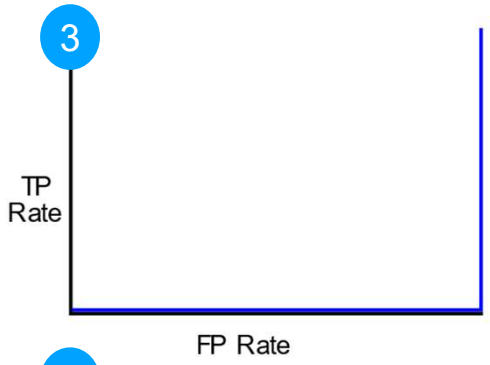
- AUC is the area under the ROC expresses how much a model can distinguish two classes along the score distribution in a given sample.

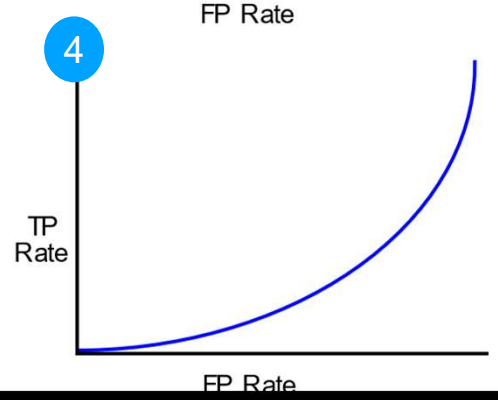
Classification Refresher: Evaluation Metrics

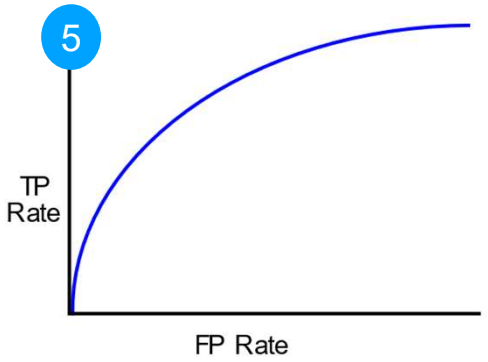
Which of the following ROC curves produce AUC values greater than 0.5?









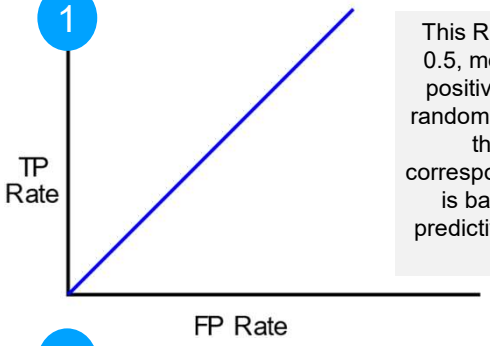


Classification: Check Your Understanding (ROC and AUC)
https://developers.google.com/machine-learning/crash-course/classification/check-your-understanding-roc-and-auc

Classification Refresher: Evaluation Metrics

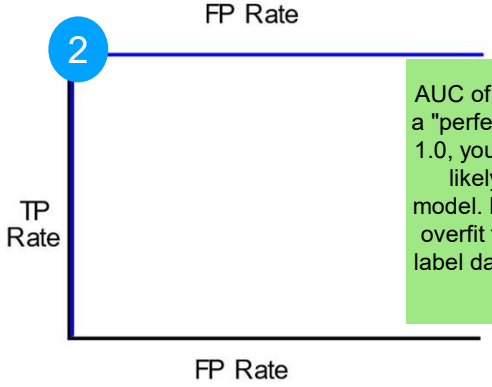
Which of the following ROC curves produce AUC values greater than 0.5?

1



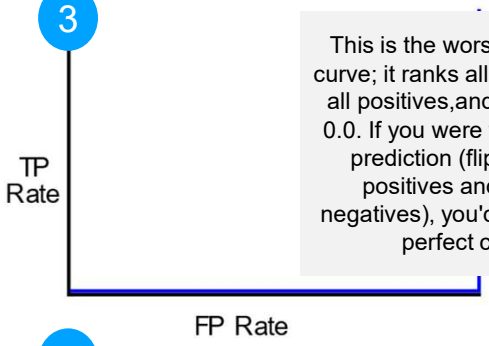
This ROC curve has an AUC of 0.5, meaning it ranks a random positive example higher than a random negative example 50% of the time. As such, the corresponding classification model is basically worthless, as its predictive ability is no better than random guessing.

2



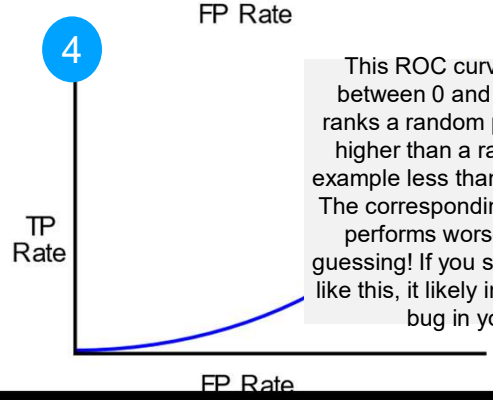
AUC of 1.0, In practice, if you have a "perfect" classifier with an AUC of 1.0, you should be suspicious, as it likely indicates a bug in your model. For example, you may have overfit to your training data, or the label data may be replicated in one of your features.

3



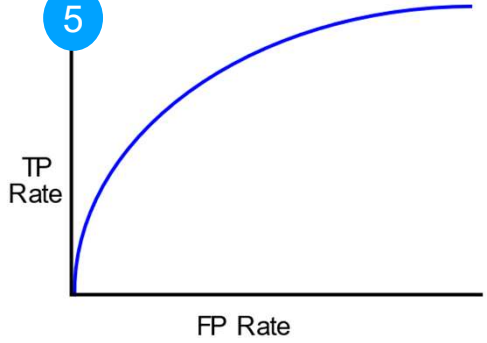
This is the worst possible ROC curve; it ranks all negatives above all positives,and has an AUC of 0.0. If you were to reverse every prediction (flip negatives to positives and postives to negatives), you'd actually have a perfect classifier!

4



This ROC curve has an AUC between 0 and 0.5, meaning it ranks a random positive example higher than a random negative example less than 50% of the time. The corresponding model actually performs worse than random guessing! If you see an ROC curve like this, it likely indicates there's a bug in your data.

5



This ROC curve has an AUC between 0.5 and 1.0, meaning it ranks a random positive example higher than a random negative example more than 50% of the time. Real-world binary classification AUC values generally fall into this range.

Binary Classification Measures Cheat Sheet -
https://www.templateroller.com/download/284921/binary-classification-regression-performances-measure-cheat-sheet.html?type=pdf&vers=main

Classification Refresher: Evaluation Metrics

Cheat Sheet

Binary classification performances measure cheat sheet

Damien François – v1.0 - 2009 (damien.francois@uclouvain.be)

Confusion matrix for two possible outcomes *p* (positive) and *n* (negative)

		Actual: p	Actual: n	Total
Predicted	p'	**true positive**	**false postive**	P
	n'	**false negative**	**true negative**	N
	total	P'	N'	

Classification accuracy
(TP + TN) / (TP + TN + FP + FN)
Error rate
(FP + FN) / (TP + TN + FP + FN)

Paired criteria

Precision: (or Positive predictive value) proportion of predicted positives which are actual positive
TP / (TP + FP)

Recall: proportion of actual positives which are predicted positive
TP / (TP + FN)

Sensitivity: proportion of actual positives which are predicted positive
TP / (TP + FN)

Specificity: proportion of actual negative which are predicted negative
TN / (TN + FP)

True positive rate: proportion of actual positives which are predicted positive
TP / (TP + FN)

True negative rate: proportion of actual negative which are predicted negative
TN / (TN + FP)

Positive likelihood: likelihood that a predicted positive is an actual positive
sensitivity / (1 - specificity)

Negative likelihood: likelihood that a predicted negative is an actual negative
specificity / (1 - sensitivity)

Combined criteria

BCR: Balanced Classification Rate
½ (TP / (TP + FN) + TN / (TN + FP))
BER: Balanced Error Rate, or **HTER**: Half Total Error Rate: 1 - BCR

F-measure harmonic mean between precision and recall
2 (precision . recall) / (precision + recall)
F_β-measure weighted harmonic mean between precision and recall
$(1+\beta)^2$ TP / $((1+\beta)^2$ TP + β^2 FN + FP)

The harmonic mean between specificity and sensitivity is also often used and sometimes referred to as F-measure.

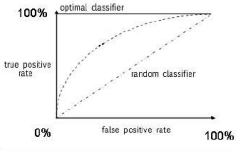
Youden's index: arithmetic mean between sensitivity and specificity
sensitivity - (1 - specificity)

Matthews correlation correlation between the actual and predicted
(TP . TN – FP . FN) /
$((TP+FP)\ (TP+FN)\ (TP+FP)\ (TN+FN))^{1/2}$
comprised between -1 and 1

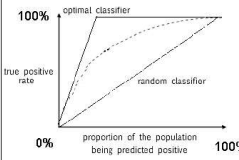
Discriminant power normalised likelihood index
sqrt(3) / π .
(log (sensitivity / (1 – specificity)) + log (specificity / (1 - sensitivity)))
<1 = poor, >3 = good, fair otherwise

Graphical tools

ROC curve receiver operating characteristic curve : 2-D curve parametrized by one parameter of the classification algorithm, e.g. some threshold in the « true postivie rate / false positive rate » space
AUC The area under the ROC is between 0 and 1



(Cumulative) Lift chart plot of the true positive rate as a function of the proportion of the population being predicted positive, controlled by some classifier parameter (e.g. a threshold)

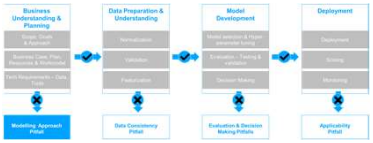


Relationships

sensitivity = recall = true positive rate
specificity = true negative rate
BCR = ½ . (sensitivity + specificity)
BCR = 2 . Youden's index - 1
F-measure = F_1measure
Accuracy = 1 – error rate

References

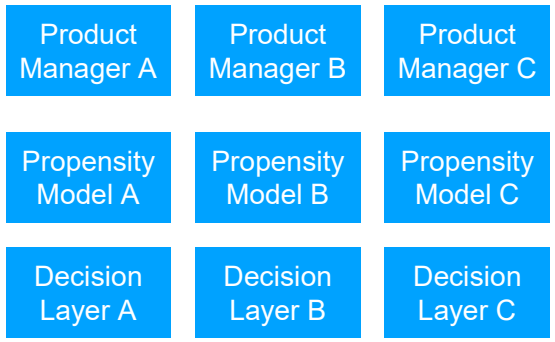
Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45, 4 (Jul. 2009), 427-437.
Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30



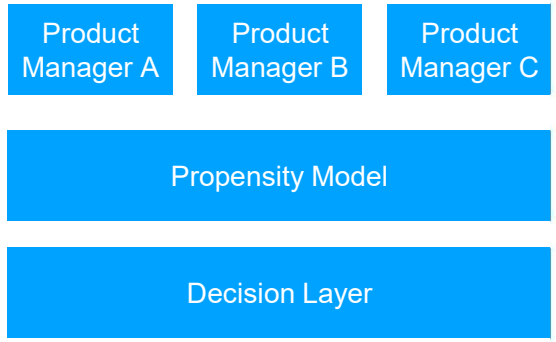
Pitfall #1: Lack of breadth and depth in modelling approach

Can result in proliferation of similar pattern use cases that will take resources for new opportunities (or the inclusion of more analytical features)

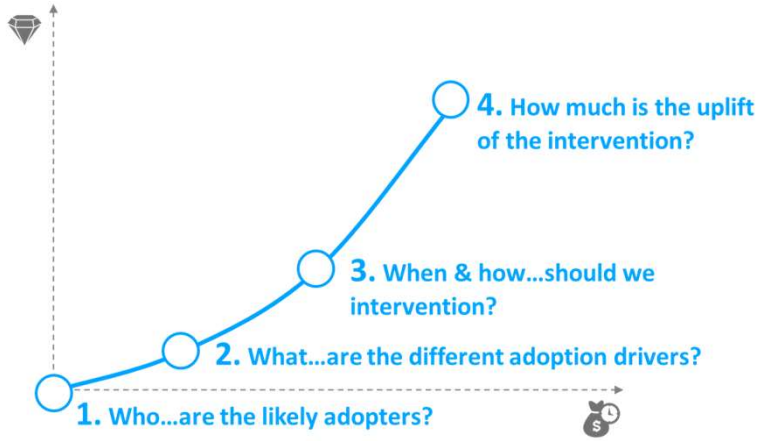
Pitfall: Data Science products can mirror your organization functions spending development/maintenance resources on a similar problem



Example: Next-Best-Offer related projects can have various stakeholders but the pattern is always the same: Customer has product X and we want to offer product Z,Y



Solution: (usually) go for a global product that addresses a fundamental pattern and invest in layers of analytical sophistication/experimentation as a better strategy to deliver more value.



Pitfall #2:Not looking at data before modelling

Each model can have different set of assumptions that have to be met, looking at data helps us to understand its characteristics and form an hypothesis of which model is best to capture those characteristics.

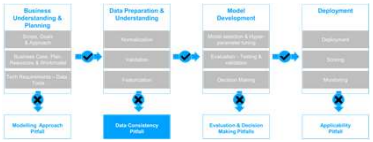
Pitfall: Not checking model assumptions can lead to poor results – "garbage in, garbage out"

Example: depending on the model the a not met assumption can affect training behavior

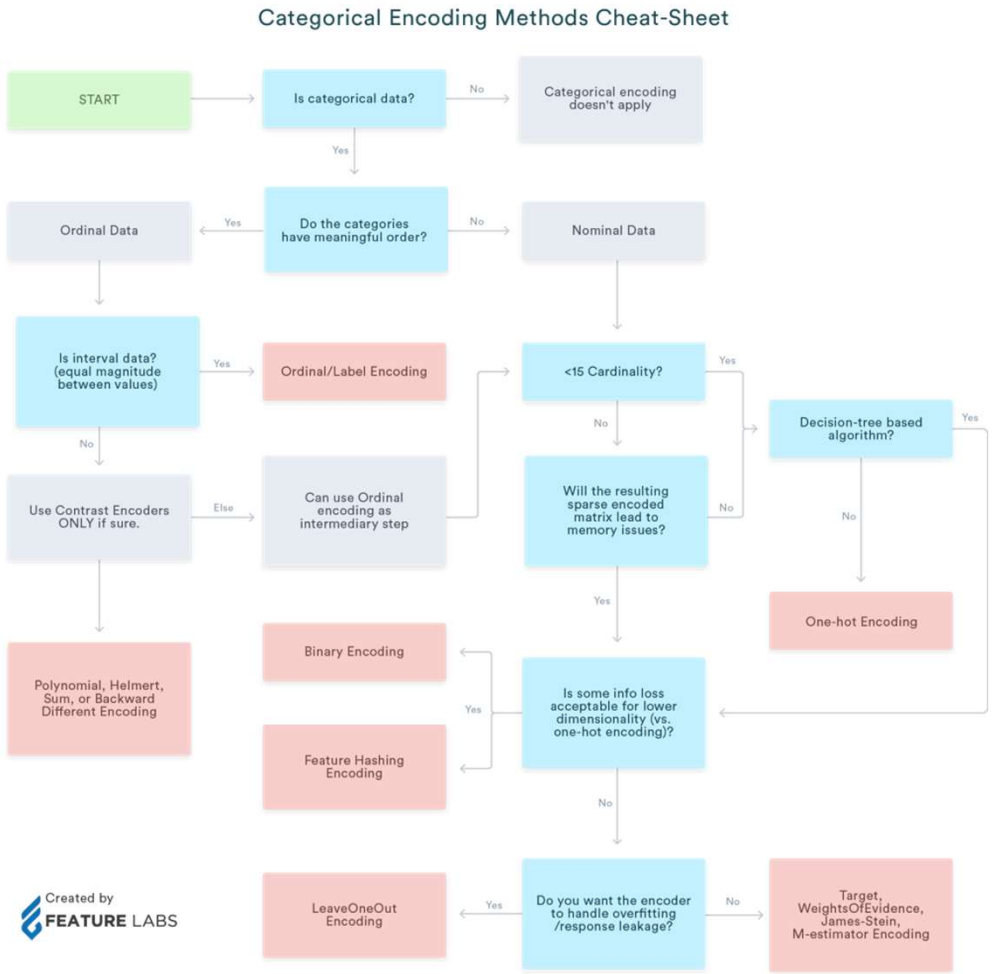
Solution: conscious development safeguards to address these matters (usually a baseline model more robust)

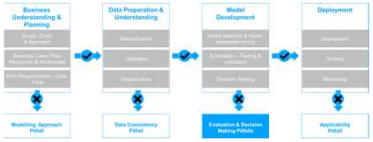
Model(s)	Assumption	Impact	Solution
Regression	Independence between features	Model will not calculate correctly coefficients and will not be able to identify which x variables have most statistical influence on the y var.	Plotting correlations between variables to diagnose and apply feature selection if issue is detected.
Distance (or Gradienc Descent) based models	Features are in same scale	▪ In distance based models features with higher scale will dominate solution; ▪ In models with gradient descent otimization will may end up taking long time to learn.	Scale features using a normalization process such as Z (better for outliers!) or Min-Max normalization.
SVMs, NNets, GBTs	Input features are numerical.	▪ Training process will left out categorical variables if they're not encoded properly.	Experiment with different encoding methods depending on categorical/ordinal data, cardinality and overfitting handling.

Encode Smarter: How to Easily Integrate Categorical Encoding into Your Machine Learning Pipeline
https://innovation.alteryx.com/encode-smarter/



Pitfall #2:Not looking at data before modelling

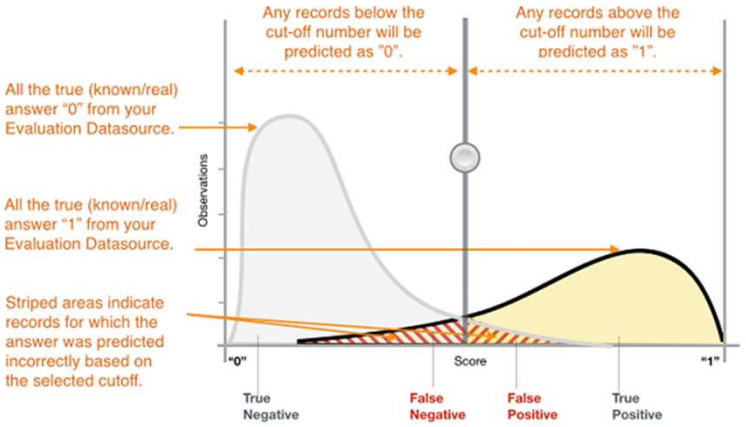




Pitfall #3: Out-of-the-Box Binary Classification

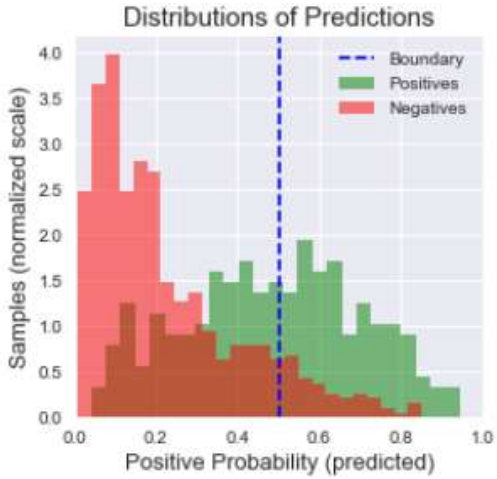
Decision-Making is different than just classification (e.g. images or text). Consequential decision-making implies an action informed by the predictions and this actions are constrained by a budget or capacity.

Pitfall: Using out-of-the-box binary predictions assume that if the model score is > 0.5 then it is 1, else it is 0.



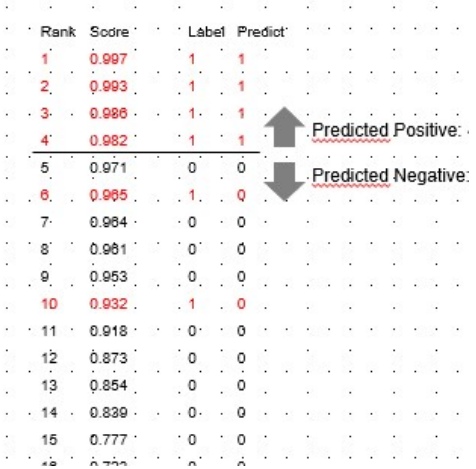
AWS Dev Guide/Binary Classification – https://docs.aws.amazon.com/machine-learning/latest/dg/binary-classification.html

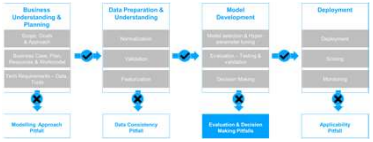
Example: Predicting Bank Loan Default – Positive Class = Payment Default



German Credit Risk Dataset – Jupyter Notebook

Solution: score all instances and select those above a defined threshold either for batch or online predictions.





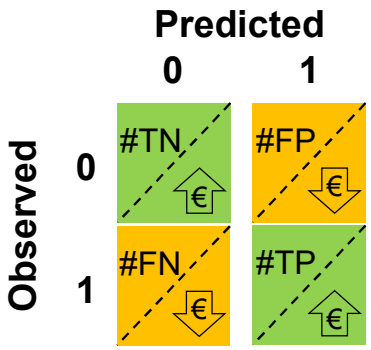
Pitfall #4: Missing link betw. model threshold and value

xxxxxxxx

Pitfall: Value created by the model is not optimized accordingly with the model score threshold

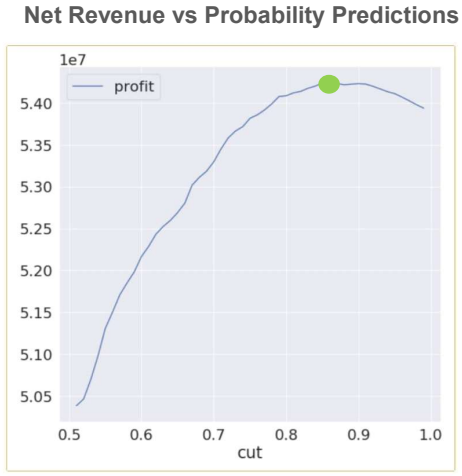
Example: Loan application illustrates which costs are critical in threshold (or model) selection – desirably with a low number of False negatives.

Solution: Plotting the Net Revenue vs Threshold reveals the best cut-off to maximize profit.



TN Revenue	FP Cost
▪ Average annual income for a payer **12.000Eur**	▪ Opportunity cost of denying a loan to a payer **-12.000Eur**
FN Cost	**TP Cost**
▪ Amount lost due to sanction a defaulter's loan **-10.00.000**	▪ Forfeiting application processing fee **-3.000Eur**

Net Revenue=
*#TN*Revenue*
*-#FP*Cost*
*-#FN*Cost*
*-#TP*Cost*



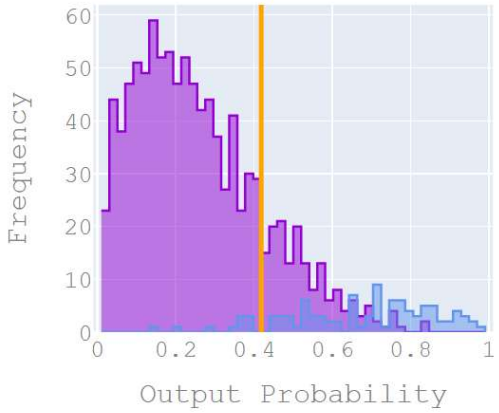
Pitfall #5: Accuracy (and ROC) can be misleading

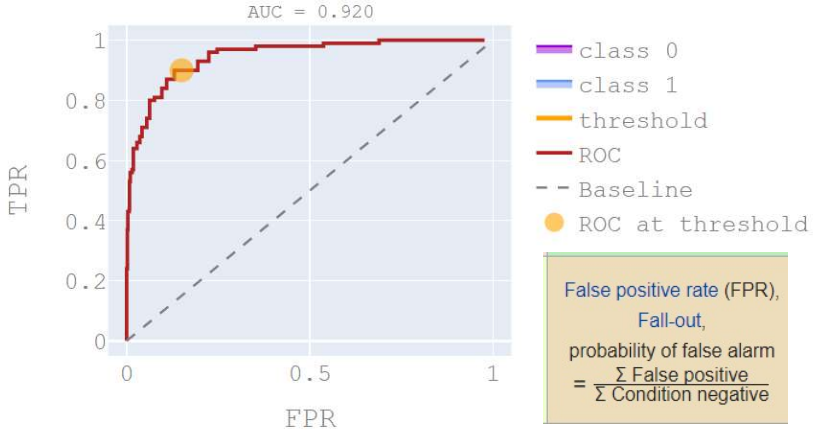
xxxxxxxx

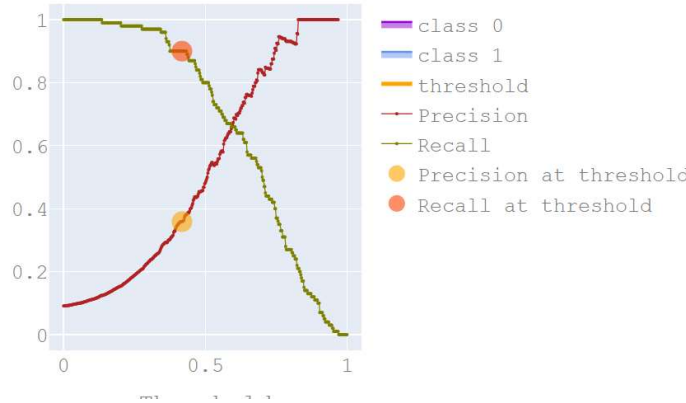
Pitfall: for imbalanced data, the right metrics for measuring become critical.

Example: For example, a fraud use case with prevalence of 1.2 percent, a model would produce **98.8 percent accuracy**, predicting no default in all cases.

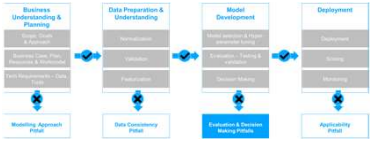
Solution: for highly unbalanced data FPR is not sensitive since Total Real Negatives is "huge". Use **precision instead.**







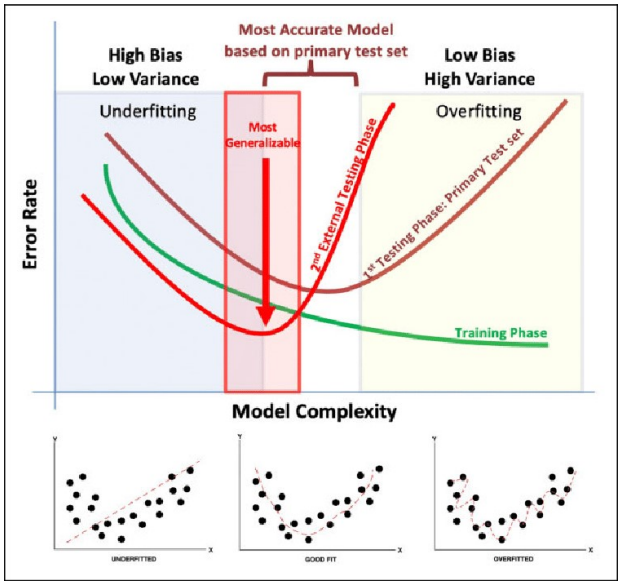
Data Science Bits - https://github.com/felipepenha/data-science-bits/blob/master/performance_metrics/



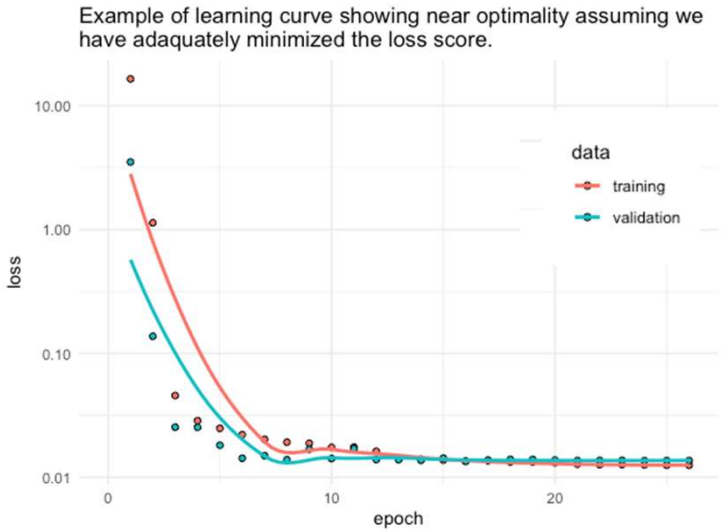
Pitfall #6: Diagnose bias and variance trade-off in learning

The bias-variance dilemma is a widely known problem in the field of machine learning, looking at the profile of learning curves a proper diagnostic can be found and solutions to fix the issue.

Pitfall: Not assessing properly model behavior to improve the learning task



Example: "Optimal Fit" is the goal of the learning algorithm



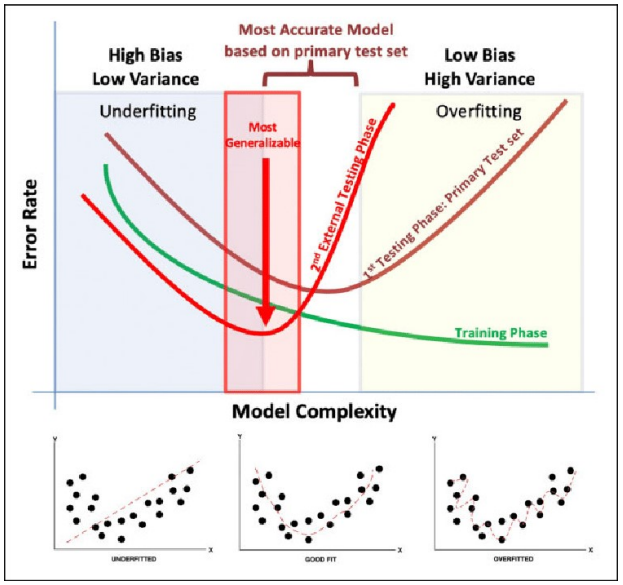
Solution: In this scenario there's no need to continue training (or it'll lead to overfitting).

- The plot of training loss decreases to a point of stability.
- The plot of validation loss decreases to a point of stability.
- The generalization gap is minimal (nearly zero in an ideal situation).

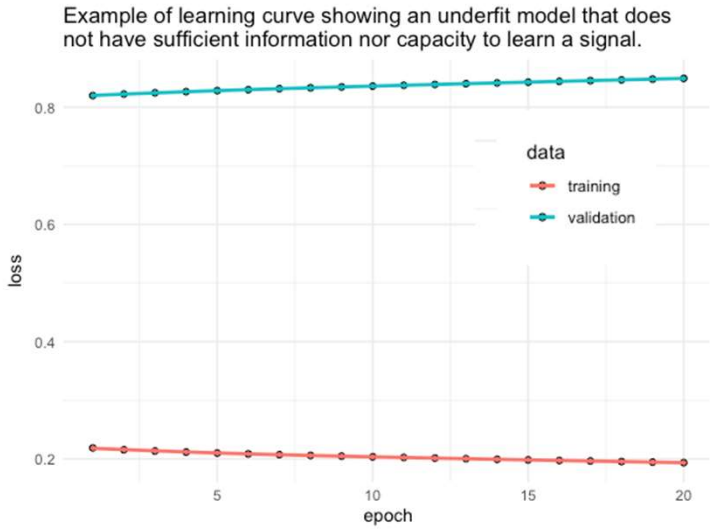
Pitfall #6: Diagnose bias and variance trade-off in learning

The bias-variance dilemma is a widely known problem in the field of machine learning, looking at the profile of learning curves a proper diagnostic can be made and solutions to fix the issue.

Pitfall: Not assessing properly model behavior to improve the learning task

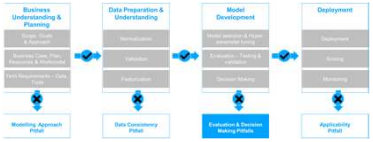


Example: "Underfit"
is a model that hasn't learned the training dataset to get a low error value – flat curve.



Solution: In this scenario model isn't capable of learning.

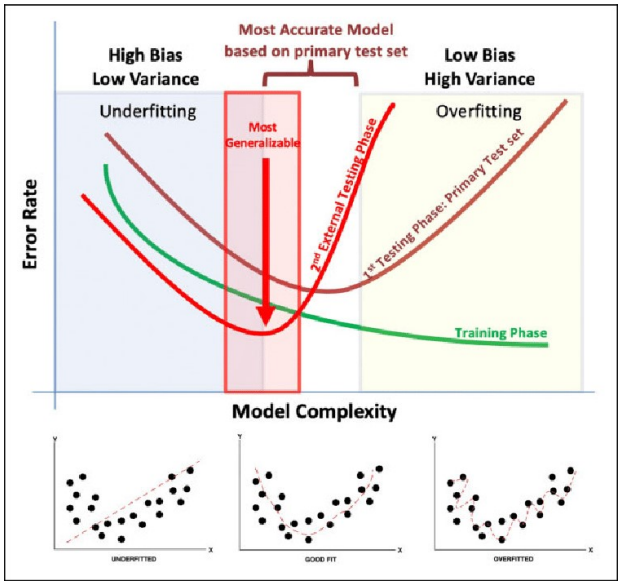
- Add more observations. You may not have enough data for the existing patterns to become strong signals.
- Add more features
- If you have explicit regularization parameters specified (i.e. dropout, weight regularization), remove or reduce these parameters;
- Model capacity may not be large enough to capture and learn existing signals – hyperparameter otimization



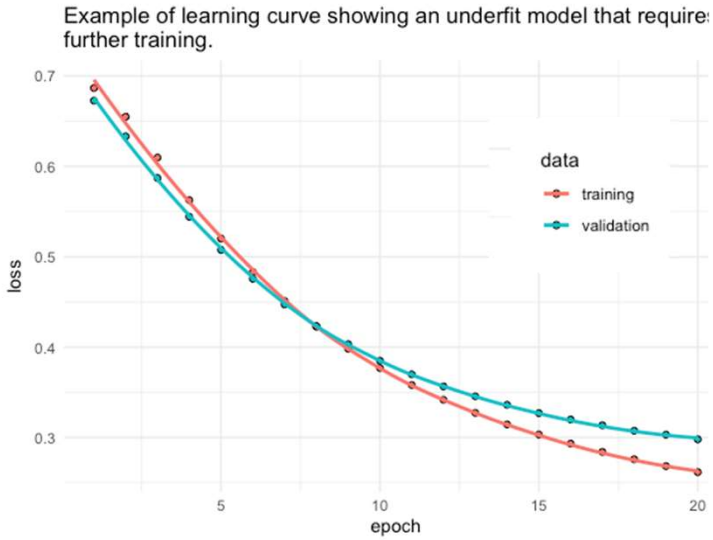
Pitfall #6: Diagnose bias and variance trade-off in learning

The bias-variance dilemma is a widely known problem in the field of machine learning, looking at the profile of learning curves a proper diagnostic can be made and solutions to fix the issue.

Pitfall: Not assessing properly model behavior to improve the learning task

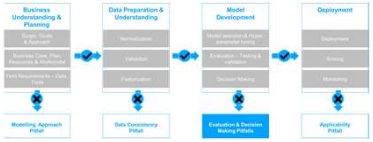


Example: "Underfit" learning curves still on decrease trend.



Solution: In this scenario learning process was stopped prematurely

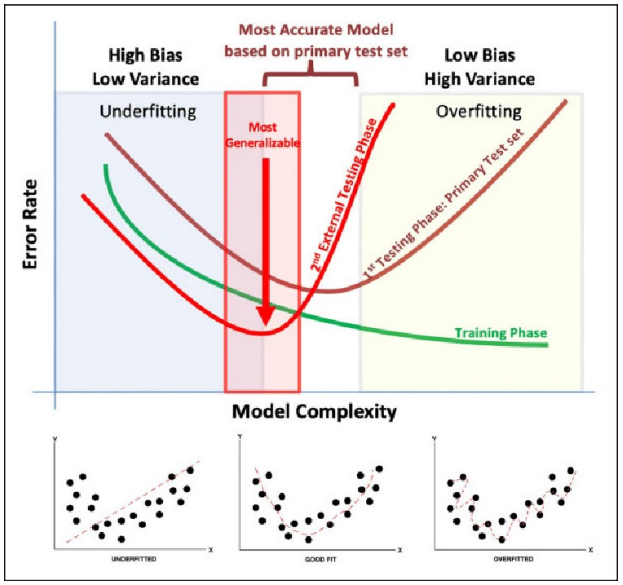
- Increase the number of epochs until the validation curve has stopped improving.
- If it is taking a long time to reach a minimum for the validation curve, increase the learning rate;



Pitfall #6: Diagnose bias and variance trade-off in learning

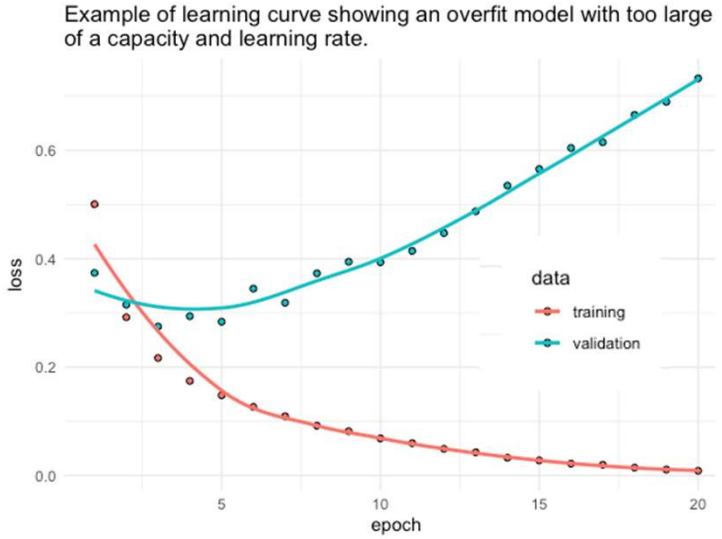
The bias-variance dilemma is a widely known problem in the field of machine learning, looking at the profile of learning curves a proper diagnostic can be made and solutions to fix the issue.

Pitfall: Not assessing properly model behavior to improve the learning task



Example: "Overfit"

Training loss continues to decrease while validation loss has a strong U-shape curve



Solution: overfitting is not necessarily a bad thing signals model extracted all what had to be learned up to a certain point

- Apply regularization techniques depending on the algorithm used – eg: if it is NN reducing the number and/or size of hidden layers;
- Earlystopping the model once the validation curve has stopped improving.- using the model parameters for the best training metric;

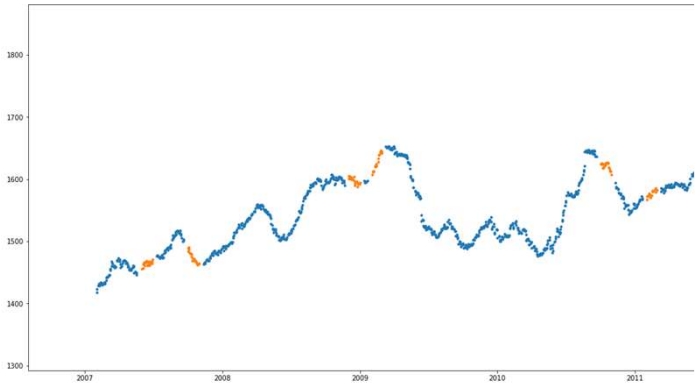
Pitfall #7: Time series validation

Standard cross-validation methods do reflect sequential discovery of time series, creating a risk of "future leakage"

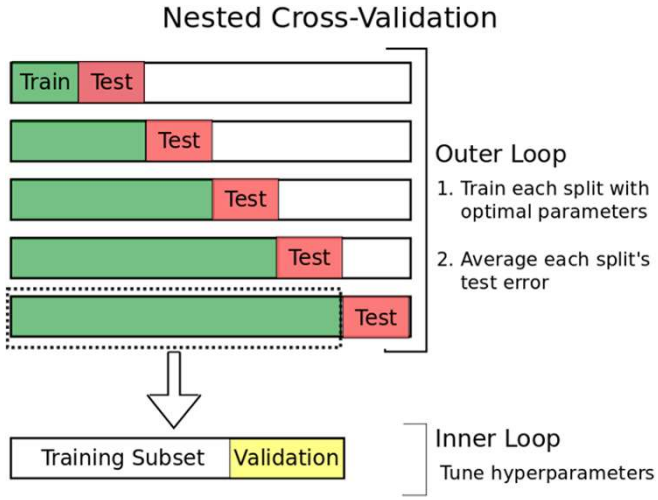
Pitfall: data with a sequential structure must be accounted for in model validation otherwise results in production can be quite different from development

Example: estimating Customer usage/churn or revenue/lifetime value using time-series data is a typical use case

Solution: ensure that cross-validation folds are created respecting the time sequence to get model validation results as close as production setting.



Customer	Day	Usage
1	1	321
1	2	313
1	3	233
1	4	244
1	5	140
1	6	80
1	7	30
1	8	2
1	9	0



Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Interview Practice with Precision and Recall - https://kiwidamien.github.io/interview-practice-with-precision-and-recall.html

Group Challenge – Practice Metric Evaluation Selection

Goal: Discuss what evaluation metric(s) should be used in a classification problem translating the technical metric in the context of each challenge.

Challenge Briefing

- Get Together with your Group in the **dedicated Zoom link** to discuss the questions related with your Use Case **[10min]**;
- **Come back to the main zoom meeting** and:

1) present your solution for your use case in <insert jamboard link here>**[5min]**

2) when you're not presenting, **you're invited to comment the other's group use cases solutions.**

#1 Breathalyzer Tests

A breathalyzer registers someone's blood alcohol content to tell if they are "over the limit" or "under the influence" of alcohol.
It is typically used at roadside police stops to determine if someone is legally able to drive.

#3 Detect malicious programs

When running a program for the first time, we are running some information about the program (such as where it was downloaded, size of the executable, etc) through a classifier. If the program is deemed safe, it will run. If it is deemed unsafe, the user will be prompted to confirm that the program is safe before running.

1. What is the positive class?
2. What does 85% recall mean?
3. What does 75% precision mean?
4. What is the most important metric?

#2 Should we unlock a phone?

We are building a facial recognition algorithm to allow people to unlock their phone. If the phone recognizes the person as the authorized user, it will unlock the phone. If it doesn't recognize the user, it will prompt them to try again or try an alternative method (such as a passphrase).

#4 De-duplicate records

You are writing a deduplication algorithm. Its goal is to flag entries in your database that are duplicates of existing records. It is more complicated than checking if two records are identical (Which is an easy problem), but instead tries to assess if differences are meaningful.Records that our algorithm detects as duplicates will be reviewed, and if we are sure they are duplicates, they are removed.

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

“the most important thing is not life, but the good life.”

Socrates, 399 BC

Technology shapes how we’re seeking “good life” and it’s not ethically neutral.

Aren’t we all Ethical already?

Strong culture organisations already have Ethics as a value however they’re not specifically catered to social and ethical processes needed to develop AI systems

<Insert your company mission & values ☺ >

…**Data Ethics** is an extension that **evaluates moral problems related with Data, algorithms and development practices** in order to provide **morally good solutions aligned with the organisation values**.

It should go from high-level abstraction concepts down to to **practical steps that can support creation & maintenance of AI systems that help us make decisions leading to fair and equitable outcomes.**

Understanding artificial intelligence ethics and safety - A guide for the responsible design and implementation of I systems in the public sector
https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

Why Data Ethics?

There are different potential harms caused by AI systems that can be managed by a solid Data Ethics practice

1

Bias and Discrimination

Systems that gain insights from existing structures and societal dynamics can reproduce and reinforce patterns of inequality.

2

Denial of Individual Autonomy, Recourse, and Rights

AI systems can potentially automate decision making processes and this can complicate accountability since it can be distributed through the system,

3

Non-transparent, Unexplainable, or Unjustifiable Outcomes

Depending on the ML model there may be challenging to understand the resulting algorithmic rules which can hide traces of discrimination, bias, inequity or fairness.

4

Invasions of privacy

AI systems are developed frequently using personal data which can be captured, extracted or handled in a way without knowledge or proper consent of the data subject.

5

Isolation and Disintegration of Social Connection

Societies are built on relationships of trust, empathy and mutual understanding, as AI gets more pervasive human to human interaction is reduced and everything that comes along.

6

Unreliable, Unsafe, or Poor-Quality Outcomes

Negligent design, development or deployment of AI systems can lead to outcomes that hurt individuals and undermine public trust in the responsible use of societally beneficial AI tech.

[1] Goodman, B. F. (2016). European Union regulations on algorithmic decision-making and a "right to explanation." ICML Workshop on Human Interpretability in Machine Learning
[2] In https://www.accenture.com/us-en/insights/technology/testing-ai
[3] https://www.bcg.com/publications/2020/europe-can-catch-up-in-ai-but-must-act-today?utm_medium=Email&utm_source=esp&utm_campaign=most_innovative_companies&utm_description=ealert&utm_topic=covid&utm_geo=global&utm_content=202007&utm_usertoken=CRM_76ee82224b32d3ff8ede5bdf349924b2f40a5cd2

Why does Data Ethics matter to organizations?

Data Ethics Practice can be a new competitive advantage

- "…the data controllers should assure implementation of measures that prevents, inter alia, discriminatory effects…also the "right-to-explanation" is deeply embedded on GDPR in Articles 13-15, specifying that individuals (or "data subjects") have the right to know which data is being collected and how it is being used and what for." [1]

- There's a nation race for AI adoption (where Portugal is lagging) taking an ethical approach will be a key enabler for that adoption in our economy as a growth driver [3]

- "Business executives today want to win customers' trust as growth comes from customers trusting the business. And, aligned to that customers have to trust you. You need to ensure that your systems are supporting that trust imperative. And that's exactly why businesses need to care… customers trust businesses which have verifiable, explainable, trustworthy systems" [2]

- Going through the process of auditing a Model for an explanation to better understand "Fairness" dimensions leads us to an insight journey, building internal trust on the DS practice. Also, DS practice will be challenged to another level of sophistication since explanations and auditing/mitigation bias are cutting edge technical matters that can be leveraged to create competitive value.

Are there any Data Ethics frameworks?

Organizations are introducing policy frameworks to enhance and regulate AI based services or products

Table 1: Trustworthy technology classes related to FAT* frameworks. *✗– no mention, ✓– mentioned, ✓✓– emphasised*

Framework	Year	Document Owner	Entities	Country	Fairness	Explainability	Safety	Auditability
Top 10 principles of ethical AI	2017	UNI Global Union	Ind	Switzerland	✓	✓	✓	✓
Toronto Declaration	2018	Amnesty International	Gov, Ind	Canada	✓✓	✓	✗	✓
Future of work and Education For the Digital Age	2018	T20: Think 20	Gov	Argentina	✓✓	✓	✓	✓
Universal Guidelines for AI	2018	The public voice coalition	Ind	Belgium	✓✓	✓	✓	✓
Human Rights in the Age of AI	2018	Access Now	Gov, Ind	United States	✓✓	✓	✓✓	✓
Preparing for the Future of AI	2016	US national Science, and Technology Council	Gov, Ind, Acad	United States	✓	✓	✓✓	✓
Draft AI R&D Guidelines	2017	Japan Government	Gov	Japan	✗	✓	✓✓	✓
White Paper on AI Standardization	2018	Standards Administration of China	Gov, Ind	China	✓	✗	✓✓	✓✓
Statements on AI, Robotics and 'Autonomous' Systems	2018	European Group on Ethics in Science and New Technologies	Gov, Ind, Acad	Belgium	✓	✓	✓	✓✓
For a Meaningful Artificial Intelligence	2018	Mission assigned by the French Prime Minister	Gov, Ind	France	✓	✓	✗	✓✓
AI at the Service of Citizens	2018	Agency for Digital Italy	Gov, Ind	Italy	✓	✓	✓	✓
AI for Europe	2018	European Commission	Gov, Ind	Belgium	✓	✓	✓	✓
AI in the UK	2018	UK House of Lords	Gov, Ind	United Kingdom	✓✓	✓	✓	✓
AI in Mexico	2018	British Embassy in Mexico City	Gov	Mexico	✓	✗	✓	✓
Artificial Intelligence Strategy	2018	German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs	Gov, Ind	Germany	✓	✓	✓	✓✓
Draft Ethics Guidelines for Trustworthy AI	2018	European High Level Expert Group on AI	Gov, Ind, Civ		✓✓	✓	✓✓	✓
AI Principles and Ethics	2019	Smart Dubai	Ind	UAE	✓✓	✓	✓✓	✓
Principles to Promote FEAT AI in the Financial Sector	2019	Monetary Authority of Singapore	Gov, Ind	Singapore	✓	✓	✗	✓
Tenets	2016	Partnership on AI	Gov, Ind, Acad	United States	✓	✓	✓	✓
Asilomar AI Principles	2017	Future of Life Institute	Ind	United States	✓	✓	✓	✓
The GNI Principles	2017	Global Network Initiative	Gov, Ind	United States	✗	✓	✓	✓
Montreal Declaration	2018	University of Montreal	Gov, Ind, Civ	Canada	✓✓	✓✓	✓✓	✓
Ethically Aligned Design	2019	IEEE	Ind	United States	✓	✓	✓	✓✓
Seeking Ground Rules for AI	2019	New York Times	Ind, GeP	United States	✓	✓	✓	✓
European Ethical Charter on the Use of AI in Judicial Systems	2018	Council of Europe: CEPEJ	Gov	France	✓	✓	✓	✓
AI Policy Principles	2017	ITI	Gov, Ind	United States	✓	✓	✓✓	✓
The Ethics of Code	2017	Sage	Ind	United States	✓	✗	✗	✓
Microsoft AI Principles	2018	Microsoft	Ind	United States	✓	✓	✓✓	✓
AI at Google: Our Principles	2018	Google	Ind	United States	✓	✓	✓✓	✓
AI Principles of Telefónica	2018	Telefónica	Ind	Spain	✓	✓	✓	✗
Guiding Principles on Trusted AI Ethics	2019	Telia Company	Ind	Sweden	✓	✓	✓✓	✓
Declaration of the Ethical Principles for AI	2019	IA Latam	Ind	Chile	✓	✓	✓✓	✓

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2019). The relationship between trust in AI and trustworthy machine learning technologies.

Framework	Year	Document Owner	Entities	Country	Fairness	Explainability	Safety	Auditability
Microsoft AI Principles	2018	Microsoft	Ind	United States	✓	✓	✓✓	✓
AI at Google: Our Principles	2018	Google	Ind	United States	✓	✓	✓✓	✓
AI Principles of Telefónica	2018	Telefónica	Ind	Spain	✓	✓	✓	✗
Guiding Principles on Trusted AI Ethics	2019	Telia Company	Ind	Sweden	✓	✓	✓✓	✓
Declaration of the Ethical Principles for AI	2019	IA Latam	Ind	Chile	✓	✓	✓✓	✓

The principles within each theme are:

Privacy:
Privacy
Control over Use of Data
Consent
Privacy by Design
Recommendation for Data Protection Laws
Ability to Restrict Processing
Right to Rectification
Right to Erasure

Accountability:
Accountability
Recommendation for New Regulations
Impact Assessment
Evaluation and Auditing Requirement
Verifiability and Replicability
Liability and Legal Responsibility
Ability to Appeal
Environmental Responsibility
Creation of a Monitoring Body
Remedy for Automated Decision

Safety and Security:
Security
Safety and Reliability
Predictability
Security by Design

Transparency and Explainability:
Explainability
Transparency
Open Source Data and Algorithms
Notification when Interacting with an AI
Notification when AI Makes a Decision about an Individual
Regular Reporting Requirement
Right to Information
Open Procurement (for Government)

Fairness and Non-discrimination:
Non-discrimination and the Prevention of Bias
Fairness
Inclusiveness in Design
Inclusiveness in Impact
Representative and High Quality Data
Equality

Human Control of Technology:
Human Control of Technology
Human Review of Automated Decision
Ability to Opt out of Automated Decision

Professional Responsibility:
Multistakeholder Collaboration
Responsible Design
Consideration of Long Term Effects
Accuracy
Scientific Integrity

Promotion of Human Values:
Leveraged to Benefit Society
Human Values and Human Flourishing
Access to Technology

Further inform methodology Artificial Intell in Ethical and *(Berkman Kle cyber.harvar*

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M., Dushkin, R., Tolmatov, A., & Onatsik, K. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rightsbased Approaches to Principles for AI.

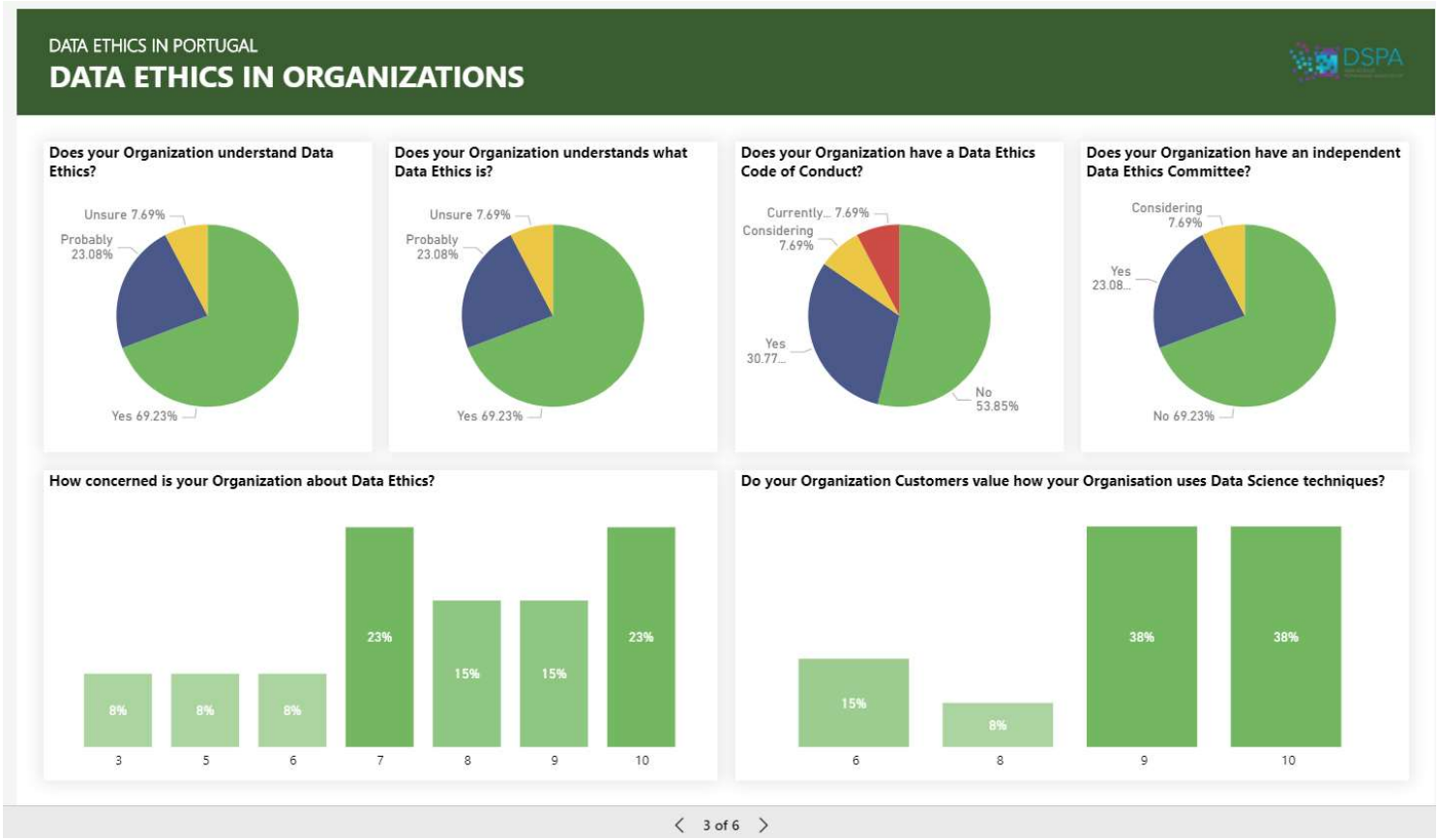
Also other organisations are starting more focused and tactical:

https://axa-rev-research.github.io/static/AXA_Booklet_Bias.pdf

DSPA Survey Data Ethics in Portugal – https://dspa.pt/observatorio-dataethics/?lang=en

And in practice? Are we doing it?

The abundance of Data Ethics frameworks doesn't translate to many PT organizations in practice



Understanding artificial intelligence ethics and safety
https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

So, where to start from?

A structured framework will organize the main principles. The Alan Turing Institute has a detailed guide with the FAST principles:

Fairness

- AI systems that process personal data must be designed to meet a mininum threshold of discriminatory non-harm. This means that all the tech end-to-end pipeline is designed and implemented from data up to the outcomes that causes to individuals in an unbiased way.

Accountability

- Answerability and auditability has to be included by design, requiring accountable humans in the loop across the entire design and implementation chain, with defined processes that allow end-to-end oversight and review.

Sustainability

- Social impact has to be determined through bringing to light unseen risks that can affect individuals or groups.
- Also technical sustainability has to be ensured for the AI system that needs to be safe, accurate, reliable, secure and robust.

Transparency

- AI system needs to be transparent in terms of being able to explain the design/implementation process but also the in terms of producto, ie: the contente and justificationof the outcome.

Integrating Ethics within Machine Learning Courses –
https://dl.acm.org/doi/10.1145/3341164

So, where to start from?

A structured framework will organize the main principles, the Alan Turing Institute has a detailed guide with the FAST principles:

Fairness

- How have you identified and minimized any bias in the data or model?
- How was any potential bias coming from the development team was identified and then mitigated?

Accountability

- Which laws and regulations might be applicable to this project?
- How is ethical accountability being achieved?
- Can we trace back the causal origins of its behavior?
- Who's accountable for actions and policies encoded on the system?

Sustainability

- How might the legal rights of organizations and individuals be impinged by our use of data?
- How might an individuals' privacy and anonimity be impinged by via aggregation and linking of data?

Transparency

- How transparent does the model needs to be and how is that achieved?
- What are likely misinterpretations of results and what can be done to prevent those misinterpretations?
- How transparency helps to debug/improve system?
- How can help tailor interventions?

What happens when something goes wrong?

A few Data Ethics cautionary tales that made "success" on social media…

Hiring [1]

Amazon scraps secret AI recruiting tool that showed bias against women

Pricing [2]

Staples website displays different prices to people depending on distance from a rival brick-and-mortar store

Justice [3]

An algorithm used by courts to assist in bail decisions was biased against non-white defendants

[1] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
[2] https://www.wsj.com/articles/SB10001424127887323777204578189391813881534
[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

What happens when something goes wrong?

A few Data Ethics cautionary tales that made "success" on social media…

Self Driving Cars [1]

The uncertainty around assigning liability in AV highlights the desirability of a coordinated effort by policymakers and manufacturers to devise organisational structures and regulatory measures which support a fair criteria for culpability attribution

Social Welfare Tax [2]

*The **robodebt automated welfare recovery scheme** matched annualised pay information from the Australian Taxation Office and income data to claw back overpaid welfare payments…*
*But **one-in-five debt letters sent were based on false** information. The Robodebt class action bought by Gordon Legal has been settled at a **cost to the government of around $1.2 billion**.*

Health Care [3]

*AI system as **only advisory and expect that accountability is thereby secure, because human clinicians still make the final decision**. …. **dilemma with two equally undesirable choices**. Either clinicians spend the time to develop their own opinions as to the best course of action, meaning that the artificial intelligence system adds little value; or clinicians must accept the advice blindly, further weakening both the control and epistemic conditions of moral accountability.*

[1] Legal issues in automated vehicles: critically considering the potential role of consent and interactive digital interfaces https://www.nature.com/articles/s41599-020-00644-2
[2] https://7news.com.au/business/finance/robodebt-class-action-everything-you-need-to-know-about-the-12-billion-lawsuit-brought-by-gordon-legal-c-1595625
[3] https://www.who.int/bulletin/volumes/98/4/19-237487/en/

What happens when something goes wrong?

A few Data Ethics cautionary tales that made "success" on social media…

Data Brokers [1]

The type of location data in question is collected from millions of phones, with most people unaware that their movements are being tracked this way and unable to find out who has access to that information. There are few laws regulating location data companies, and government agencies have used this to their advantage, spending millions to gain access to this information.

Facebook [2]

Facebook breached data protection laws by failing to keep users' personal information secure, *allowing Cambridge Analytica to* ***harvest the data of up to 87 million people without their consent worldwide****. The now-defunct firm worked for the Trump presidential campaign and used the data to influence several elections around the world.*

Google [3]

Google has agreed to pay a record ***$170 million penalty*** *to settle accusations that YouTube broke the law when it knowingly tracked and sold ads targeted to children …The settlement involves the largest-ever penalty under the Children's Online Privacy Protection Act, which* ***YouTube allegedly violated by collecting user information from kids to fuel its behavioral advertising business.***

[1] https://www.vox.com/recode/22038383/dhs-cbp-investigation-cellphone-data-brokers-venntel
[2] https://www.npr.org/2019/10/30/774749376/facebook-pays-643-000-fine-for-role-in-cambridge-analytica-scandal?t=1607184810121
[3] https://edition.cnn.com/2019/09/04/tech/google-youtube-ftc-settlement/index.html

What happens when something goes wrong?

A few Data Ethics cautionary tales that made "success" on social media…

Uber & Lyft [1]

Uber and Lyft blamed for San Francisco's congested streets… Lyft and Uber caused a 51 per cent increase in vehicle hours on the road – far outweighing the impact of other causes like changes in population and employment.

Predatory Ads [2]

…company's artificial intelligence technology "enables us to efficiently serve prospective students as well as customize our outreach and engagement with current students to help them stay and succeed in school… settlements worth over $500 million last year to resolve major actions pursued by the Federal Trade Commission and 49 state attorneys general alleging deceptive practices.

Apple [3]

[1] https://www.dezeen.com/2018/10/18/uber-lyft-transportation-network-companies-blamed-san-francisco-traffic-congestion/
[2] https://www.republicreport.org/2020/biden-must-cut-off-taxpayer-billions-to-predatory-colleges-that-ruin-students-lives/
[3] https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

[1] **AI Ethics into Practice** - https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust?ab=at_articlepage_whattoreadnext

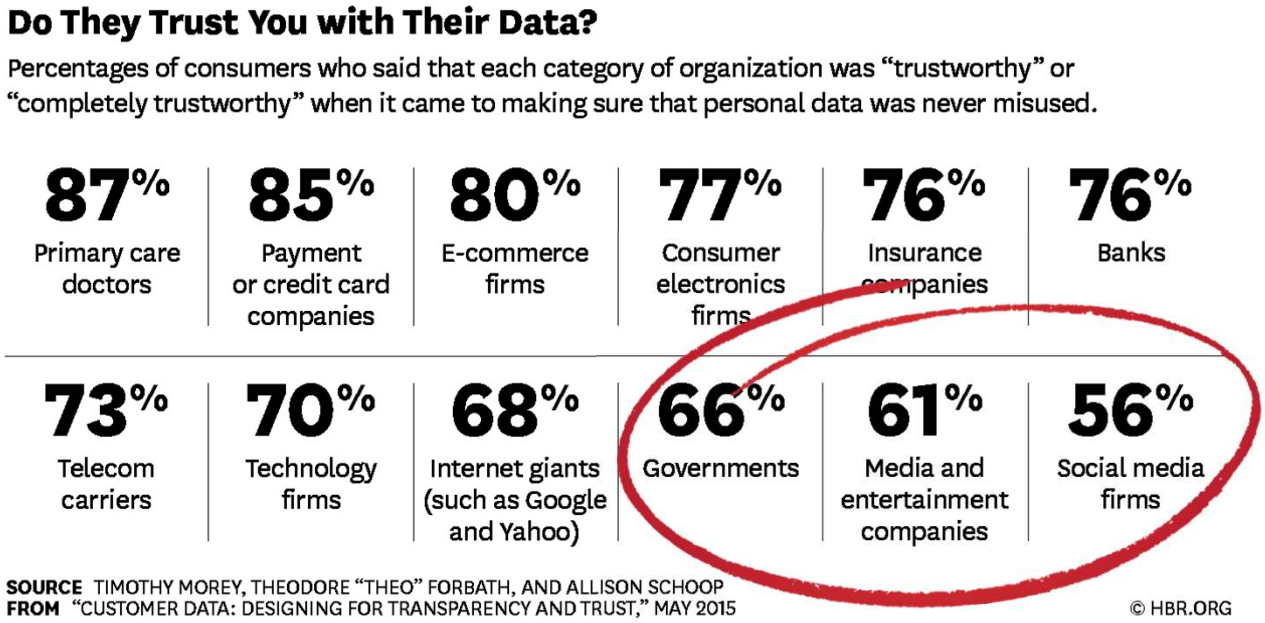
Trust as a competitive differentiator in an AI economy

Transparency reinforces trust and Customer willingness to share data, that by its turn allows AI to flourish new use cases, as opposing to hinder innovation and growth

Survey Methodology

46 companies surveyed representing seven categories of business around the world. asked Consumers to rate the firms on the following scale:

- **completely trustworthy** - would freely share sensitive personal data with a firm because they trust the firm not to misuse it;
- **trustworthy** - would "not mind" exchanging sensitive data for a desired service;
- **untrustworthy -** would provide sensitive data only if required to do so in exchange for an essential service;
- **completely untrustworthy -** would never share sensitive data with the firm.



[1] **AI Ethics into Practice** - https://cbicdnend.azureedge.net/media/3172/ai-ethics-report.pdf

How to go from principles to ground practice?

Depending on your DS Ethics maturity you may complete current practice or focus in specific "pain" areas

Not Exhaustive

Embed: Put Ethics at the core of your approach to AI

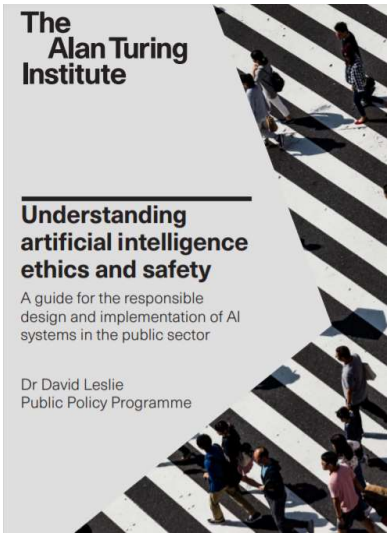
Engage: Enpower your workforce when it comes to AI

Explain: Engage Customers and the public on AI

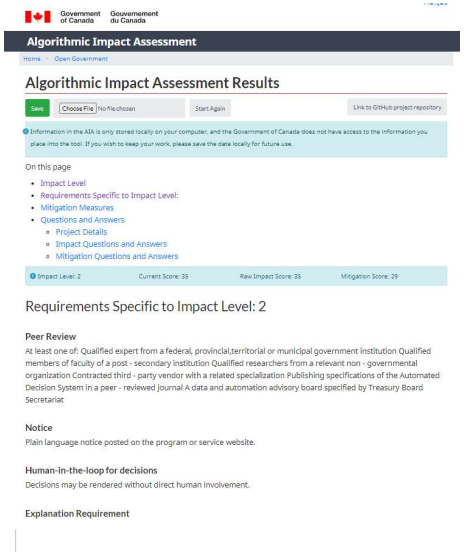
	Update/create governance processes	Engage Worforce	Improve data and diversity to eliminate unfair bias	Increase data security and privacy to boost consumer trust	Put trust in the hands of consumers by explaining AI
Challenge	• AI is a general purpose technology with a potential adoption across the organization leading to (re)definition of responsabilities on building accountability into how AI is used.	• AI has the potential to automate certain types of tasks, increasing the need to continuous learning & upskilling of the workforce.	• Data & Teams must reflect diversity mitigating risk of harming particular groups while spoting biased data or assumptions that are ethical blindspots.	• AI systems can bring new risks with new kinds of complexity or simply by making the existing risks more challeging to control.	▪ With explainability, we can check for errors, problems, or unfair bias in AI decisions and re commendations. Is worth investing in accurately tracking an algorithm's performance that could lead to better outcomes for business.
Lines of Action	• Diagnose DS Ethics maturity in organization; • Map how AI fits existing frameworks & principles; • Consider creating new governance frameworks & for AI; • Set up a AI ethics Board;	• Prepare employees for working in an environment that is rapidly changing; • Encourage employees to explore how to introduce AI in their roles; • Retain your AI talent making sure that they feel the "right thing"	**Data:** • Test & monitor technically data & models; • Make data more inclusive; • Collect new data with AI in mind; **Team:** Review existing HR targets for diversity on teams.	▪ Introduce/review good data practices: robust & audited data access protocols in place, who can access data & when. Deidentify data where possible; • Increase trust by making your Data privacy policy accessible to consumers	• Map stakeholders and put yourself in their shoes to decide how to communicate on AI – explanations should be tailored for context & audience; • Inform/highlight people when interaction/decisions are being taken by algorithms.

[1] Understanding artificial intelligence ethics and safety - https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf
[2] Algorithmic Impact Assessment - https://canada-ca.github.io/aia-eia-js/
[3] https://www.gov.uk/government/publications/data-ethics-framework
[4] https://www.tech-transformed.com/product-development/

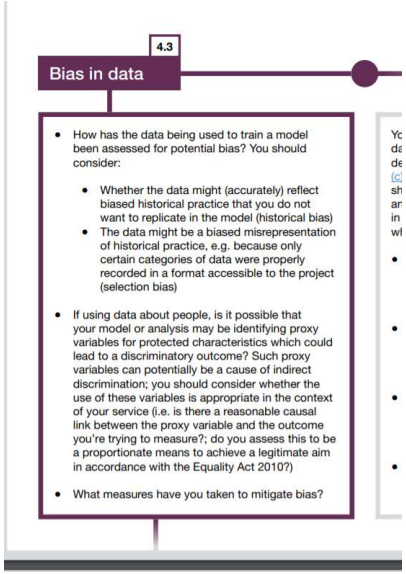
There's plenty of (free!) tools to structure consequential bias & fairness policy discussions on your projects!



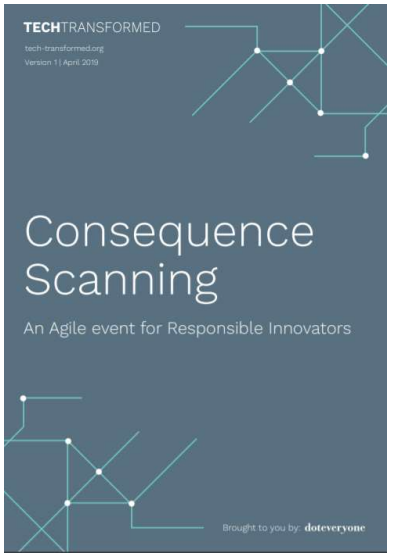
- Thorough guide on values, principles and Process Based Framework.with human centered implementation processes. [1]



- Self assessment tool with impact score plus a series of recommendations of mitigation requirements and measures.



- Other self-assessment tool to Help you cook the project prepared by the UK governement [3]



- There's also a free kit prepared by TechTransformed to upgrade your agile practice to consider impact of products/services on people & society.

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Group Challenge – Practice Data Ethical Issues Identification

Goal: Identify ethical issues on an AI system and discuss potential harms for each involved stakeholder.

Challenge Briefing

- Spend some time to read the article;
- Get Together with your Group in the **dedicated Zoom link** to discuss the questions **[20min]**;

- **Come back to the main zoom meeting** and:

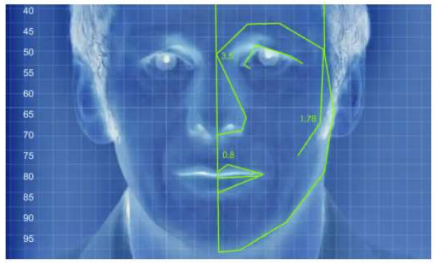
1) present your view to the audience – write it on the jamboard <insert jamboard link> **[5min/group]**

2) when you're not presenting, **you're invited to comment the other's group ethical issues.**

https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions

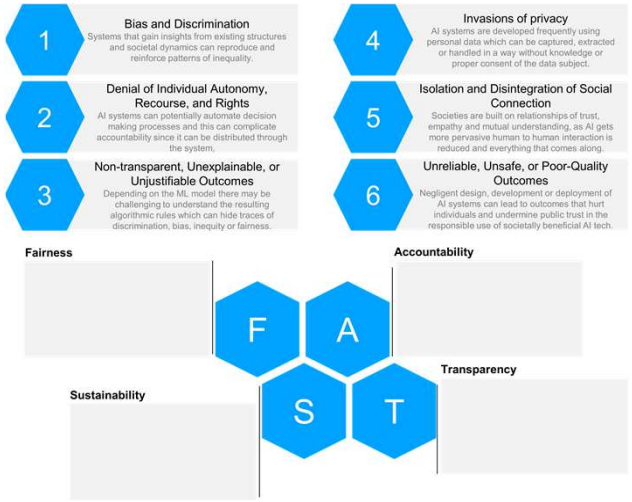


An illustrated depiction of facial analysis technology similar to that used in the experiment. Illustration: Alamy

Artificial intelligence can accurately guess whether people are gay or straight based on photos of their faces, according to new research that suggests machines can have significantly better "gaydar" than humans.

The study from Stanford University – which found that a computer algorithm could correctly distinguish between gay and straight men 81% of the time, and 74% for women – has raised questions about the biological origins of sexual orientation, the ethics of facial-detection technology, and the potential for this kind of software to violate people's privacy or be abused for anti-LGBT purposes.

The machine intelligence tested in the research, which was published in the Journal of Personality and Social Psychology and first reported in the Economist, was based on a sample of more than 35,000 facial images that men and women publicly posted on a US dating website. The researchers, Michal Kosinski and Yilun Wang, extracted features from the images using "deep neural networks", meaning a sophisticated mathematical system that learns to analyze visuals based on a large dataset.



1) Which **potential harms** and **related ethical issues** do you identify on the AI system presented on the article?
2) Can you think of an **ethical application** of this system?

Group Challenge – Practice Data Ethical Issues Identification

Helper – it could help you to structure your brainstorm output in this format

	Article Quote	Potential Harm/Implication		Article Quote	Potential Harm/Implication
F	"AI can accurately guess whether people are gay or straight based on photos of their faces	**Bias & Discrimination** Acknowledging sexual orientation as a binary classification in scientific research can reinforce discrimination on groups which are not included.	S		
F			S		
T			A		

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

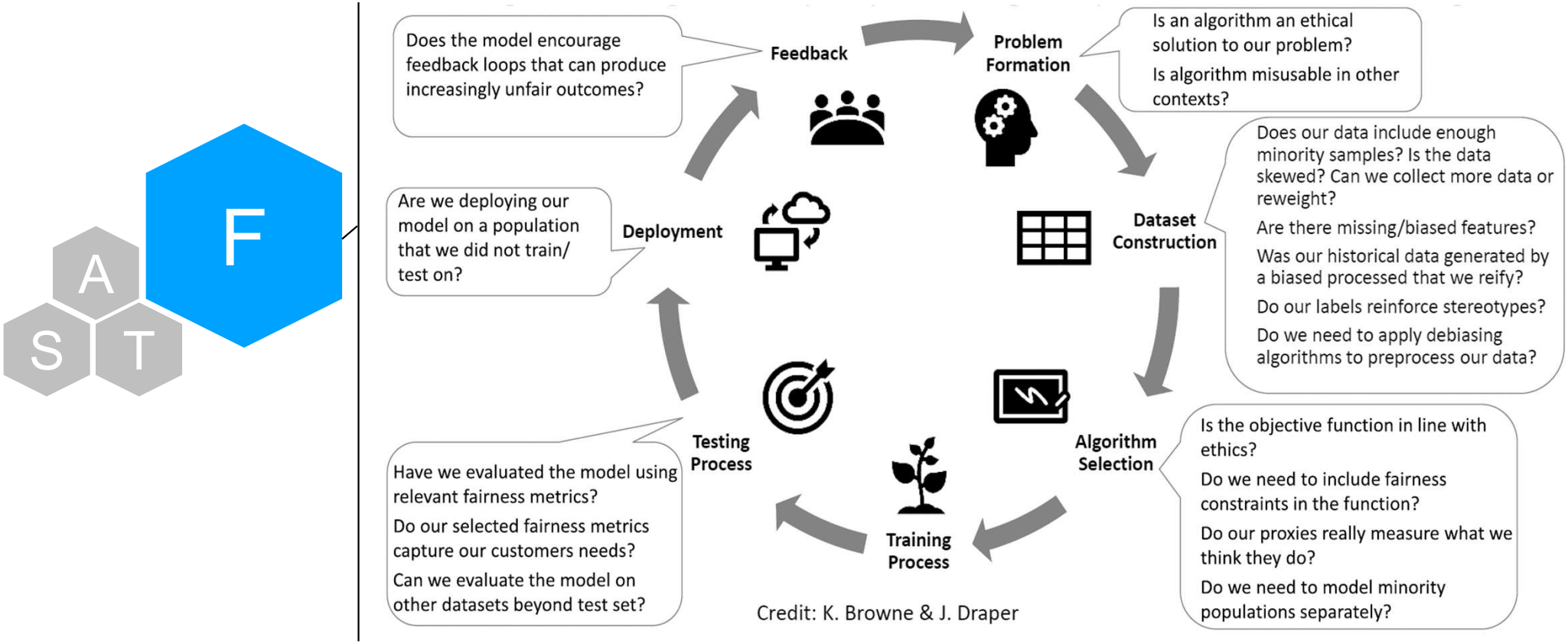
Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Team Data Science Process - https://docs.microsoft.com/pt-pt/azure/machine-learning/team-data-science-process/lifecycle-data

How do we develop AI systems that help make decisions leading to fair and equitable systems?

Each step on the pipeline may introduce potential bias to the downstream tasks



Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. CoRR, abs/1901.10002. http://arxiv.org/abs/1901.10002

Data is a social mirror of an unfair world?

Type of Bias in Data

Bias Type	Example
Historical Bias *Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection*	An example of this type of bias can be found in a 2018 image search result where **searching for women CEOs ultimately resulted in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were woman**—which would cause the search results to be biased towards male CEOs
Representation Bias *Representation bias happens from the way we define and sample from a population*	Lacking **geographical diversity in datasets like ImageNet** is an example for this type of bias, where **Western countries such as US and Great Britain are over-represented**
Measurement Bias *Measurement bias happens from the way we choose, utilize, and measure a particular feature*	In the recidivism risk prediction tool COMPAS, where **prior arrests and friend/family arrests were used as proxy variables** to measure level of "riskiness" or "crime"—-which on its own can be viewed as **mismeasured proxies**. This is due to the fact that **minority communities are controlled and policed more frequently, so they have higher arrest rates.**
Population Bias *Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platfor from the original target population*	**Different user demographics on different social platforms**, such as women being more likely to use Pinterest, Facebook, Instagram, while men being more active in online forums like Reddit or Twitter.

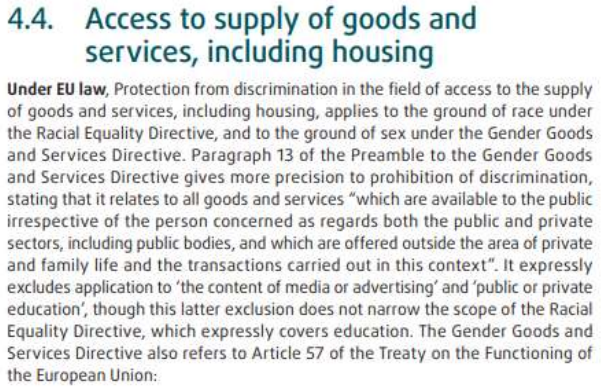
Other Sources of Bias

Evaluation	Aggregation	Simpson's Paradox	Longitudinal Data Fallacy
Sampling	Behavioral	Content Production	Linking
Temporal	Popularity	Algorithmic	User Interaction
Social	Emergent	Self Selection	Omitted Variable
Cause Effect Bias	Observer Bias	Funding Bias	Sample Size

There’s no unique mathematical definition of fairness!

Fairness isn’t a general concept, it can be concerned “**with important opportunities that affect people’s life chances**” (Solon Barocas · Moritz Hardt, 2017) - **domain specific** - or concerned with “**social salient qualities have served as the basis for unjustified and systematically adverse treatment in the past” - feature specific**.

Domain Specific



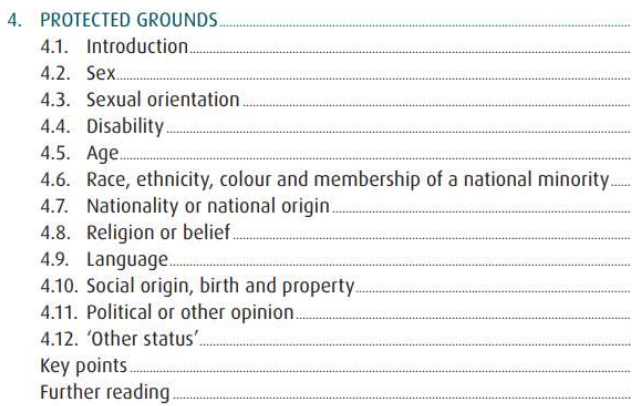
4.4. Access to supply of goods and services, including housing

Under EU law, Protection from discrimination in the field of access to the supply of goods and services, including housing, applies to the ground of race under the Racial Equality Directive, and to the ground of sex under the Gender Goods and Services Directive. Paragraph 13 of the Preamble to the Gender Goods and Services Directive gives more precision to prohibition of discrimination, stating that it relates to all goods and services “which are available to the public irrespective of the person concerned as regards both the public and private sectors, including public bodies, and which are offered outside the area of private and family life and the transactions carried out in this context”. It expressly excludes application to ‘the content of media or advertising’ and ‘public or private education’, though this latter exclusion does not narrow the scope of the Racial Equality Directive, which expressly covers education. The Gender Goods and Services Directive also refers to Article 57 of the Treaty on the Functioning of the European Union:

from Handbook on European non- discrimination law

- “Under EU law, Protection from discrimination in the field of access to the supply of **goods and services**, including housing, applies to the ground of race under the Racial Equality Directive, and to the ground of sex under the Gender Goods”

Feature Specific



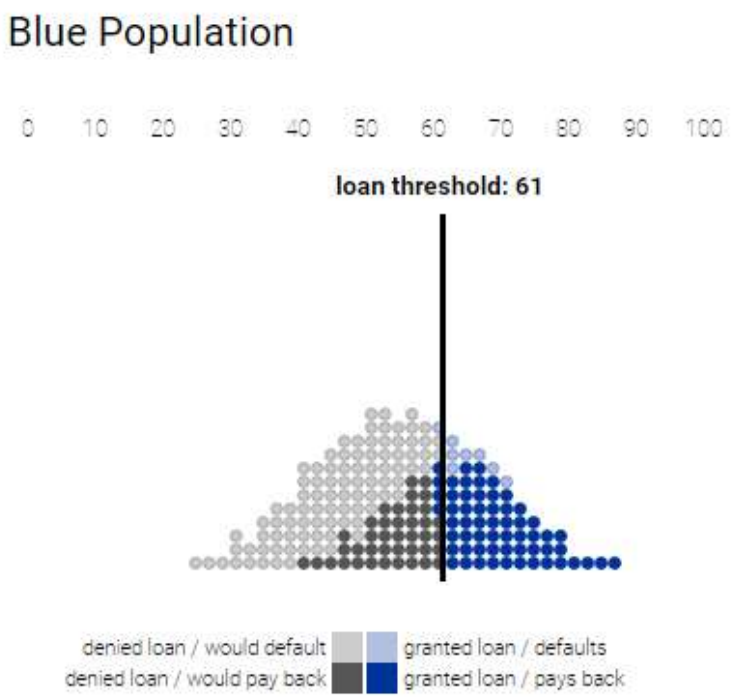
4. PROTECTED GROUNDS
   - 4.1. Introduction
   - 4.2. Sex
   - 4.3. Sexual orientation
   - 4.4. Disability
   - 4.5. Age
   - 4.6. Race, ethnicity, colour and membership of a national minority
   - 4.7. Nationality or national origin
   - 4.8. Religion or belief
   - 4.9. Language
   - 4.10. Social origin, birth and property
   - 4.11. Political or other opinion
   - 4.12. ‘Other status’
   - Key points
   - Further reading

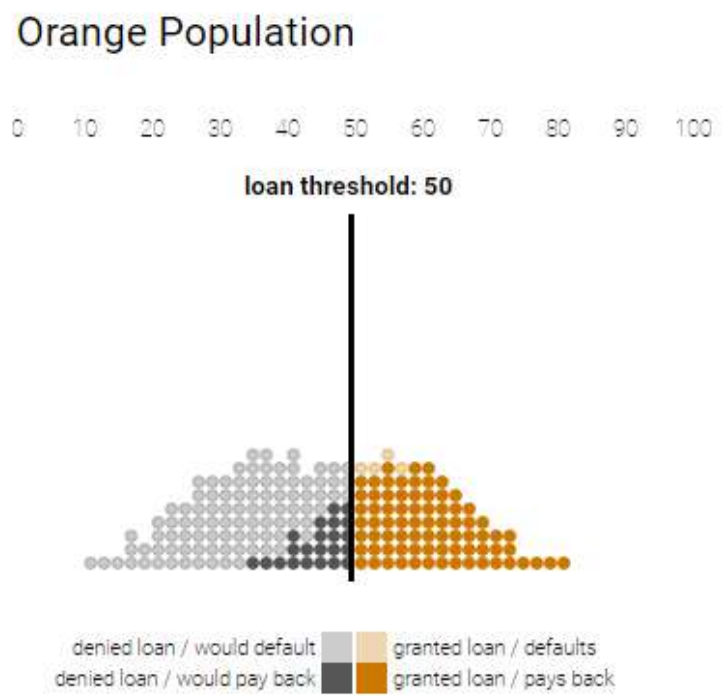
from Handbook on European non- discrimination law

- “The European non-discrimination directives prohibit differential treatment that is based on **certain ‘protected grounds’**, containing a fixed and limited list of protected grounds”

“Attacking discrimination with smarter machine learning”
https://research.google.com/bigpicture/attacking-discrimination-in-ml/

To understand Fairness implications, we can simulate a loan application use case for two distinct populations with different default probabilities…





Different threshold criteria can lead to different Fairness definition and outcome impact.

Loan Strategy Maximize profit with:	**Loan Treshold** cut-off for loan granting		**True Positive Rate** percentage of paying applications getting loans		**Positive Rate** percentage of all applications getting loans		**Incorrect** loans denied to paying applicants and granted to defaulters		**Profit** sum of the costs and profits of granted paying and default loans		
	Blue	Orange	Blue	Orange	Blue	Orange	Blue	Orange	Blue	Orange	Total
Max Profit - The most profitable, since there are no constraints. But the two groups have different thresholds, meaning they are held to different standards.	61	50	60%	78%	34%	41%	24%	13%	$ 12 100	$ 20 300	$ 32 400
Group unaware - Both groups have the same threshold, but the orange group has been given fewer loans overall. Among people who would pay back a loan, the orange group is also at a disadvantage.	55	55	81%	60%	52%	30%	21%	21%	$ 8 600	$ 17 000	$ 25 600
Demographic parity - The number of loans given to each group is the same, but among people who would pay back a loan, the blue group is at a disadvantage	60	52	64%	71%	37%	37%	23%	16%	$ 11 900	$ 18 900	$ 30 800
Equal Opportunity - Among people who would pay back a loan, blue and orange groups do equally well. This choice is almost as profitable as demographic parity, and about as many people get loans overall.	59	53	68%	68%	35%	40%	22%	17%	$ 11 700	$ 18 700	$ 30 400

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Group Challenge - Which could be the sources of bias a loan use case?

Goal: Go through the process of brainstorming sources of bias in a specifc use case

Challenge

- Get Together with your Group in the to read article + discuss the questions related with Use Case **[15min]**;
- **Come back to the main zoom meeting** and:

1) add & present your solution for your use case in the <insert jamboard link> **[5min]**

2) when you're not presenting, **you're invited to comment the other's group use cases solutions.**

Case Study - Bank Loan

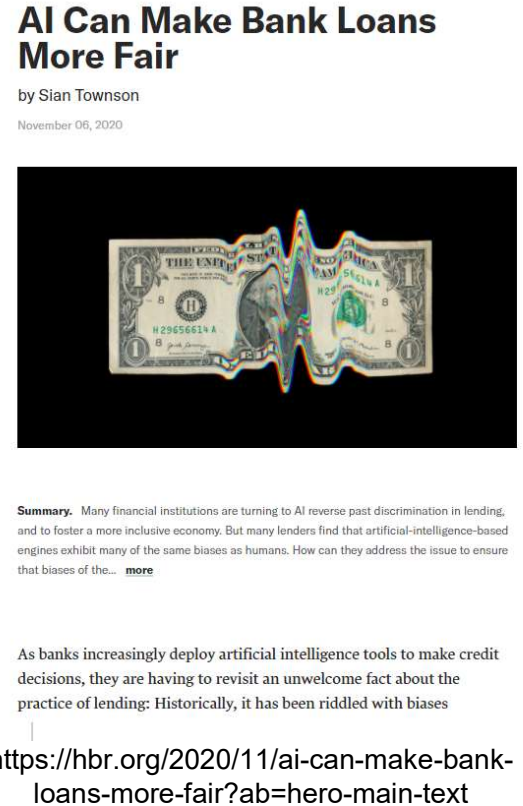
Description: Acme Bank is developing a system to decide which loan applications to deny based on predicted risk of lender not paying back their loan in time.

- **Goal:** Increase repayment rates for bank loans
- **Data:** Historical loans and payments, credit reporting data, background checks, <other that you may find relevant!>
- **Analysis:** Build model to predict risk of not repaying on time
- **Actions:** Deny loan or increase interest rate/penalties

Question: What are some potential sources of bias in the underlying data?

From KDD workshop

Suggested reading



AI Can Make Bank Loans More Fair

by Sian Townson

November 06, 2020

Summary. Many financial institutions are turning to AI reverse past discrimination in lending, and to foster a more inclusive economy. But many lenders find that artificial-intelligence-based engines exhibit many of the same biases as humans. How can they address the issue to ensure that biases of the... more

As banks increasingly deploy artificial intelligence tools to make credit decisions, they are having to revisit an unwelcome fact about the practice of lending: Historically, it has been riddled with biases

https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair?ab=hero-main-text

Playing with AI Fairness
https://pair-code.github.io/what-if-tool/ai-fairness.html

If blue = women and orange = man, how would you decide fairly?

Expert #1. Profit Max

Expert #1 defends that what matters is maximizing profit without restrictions and thus otimize the cut-offs having in account the distributions minimizing loan granting to the womens population.

Expert #2: Group unaware:

What's fair, says Expert #2, is to totally disregard the gender mix of the applicants who are given loans. If no women at all make it into the pool, that's fair so long as people were chosen purely on their merits–what's sometimes called “group unaware” fairness.
Otherwise, you'd have to kick out a qualified man to make room for a less qualified woman, because there are a limited number of approvals possible.
company should go to great lengths to exclude gender and gender-proxy information from the data set, and then go for the most accurate predictions of who will repay a loan.

#3. Demographic Parity

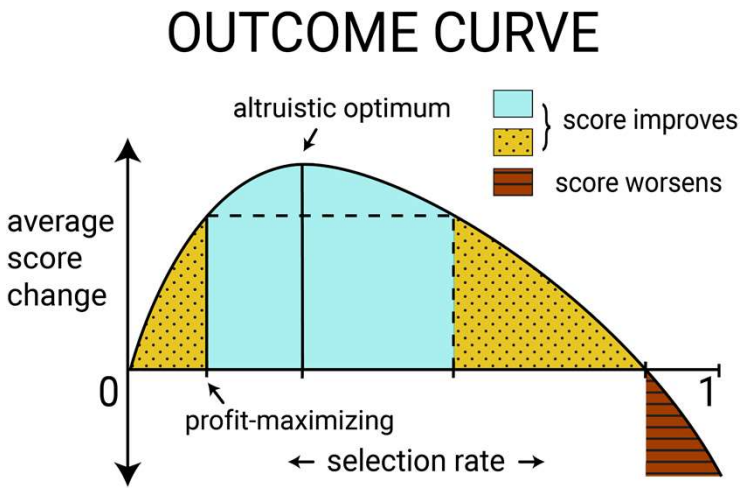
No, says Expert #3. The composition of the set of people who are granted loans should reflect the percentage of applicants: if 30 percent of the applicants are women, then 30 percent of the pool of approved applicants ought to be women.

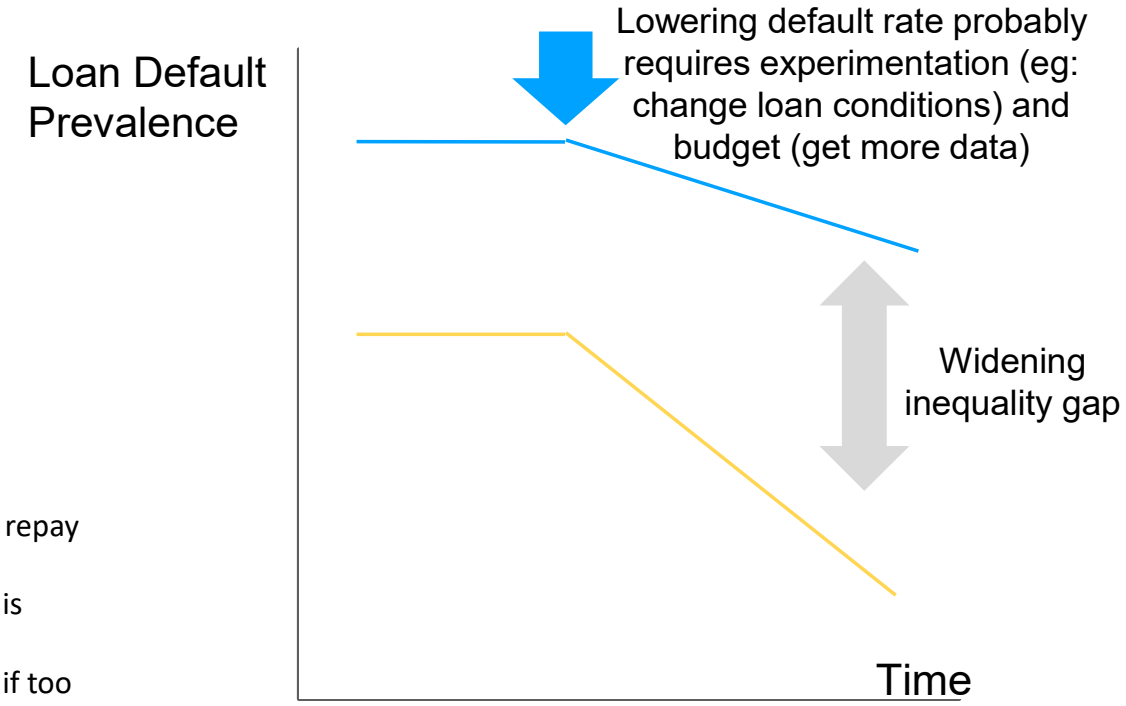
#4. Equal Opportunity

Expert #4 disagrees because if you just want equal proportions of men and women to get loans, you could randomly offer loans to women whether or not the individuals are good risks. Instead, it will be “Equal Opportunity” fairness if the same percentage of men and women who are likely to succeed at loans are given loans. What you don't want is for 90 percent of the male good risks to make it into the acceptance pile, but only 40 percent of the female good risks to do so.

Delayed Impact of Fair Machine Learning
https://bair.berkeley.edu/blog/2018/05/17/delayed-impact/

Did you consider the impact of your fairness policy over time?





- When enough individuals in a group are granted loans and successfully repay them, the **average credit score in that group is likely to increase**;
- Increasing selection rate up to a point where the average score change is lower than under **unconstrained profit-maximization but still positive**;
- Selection rates in dotted **yellow regions are causing relative harm** and if too many individuals are **unable to repay their loans, credit score for the group will decrease**, as is the case in the red striped region.

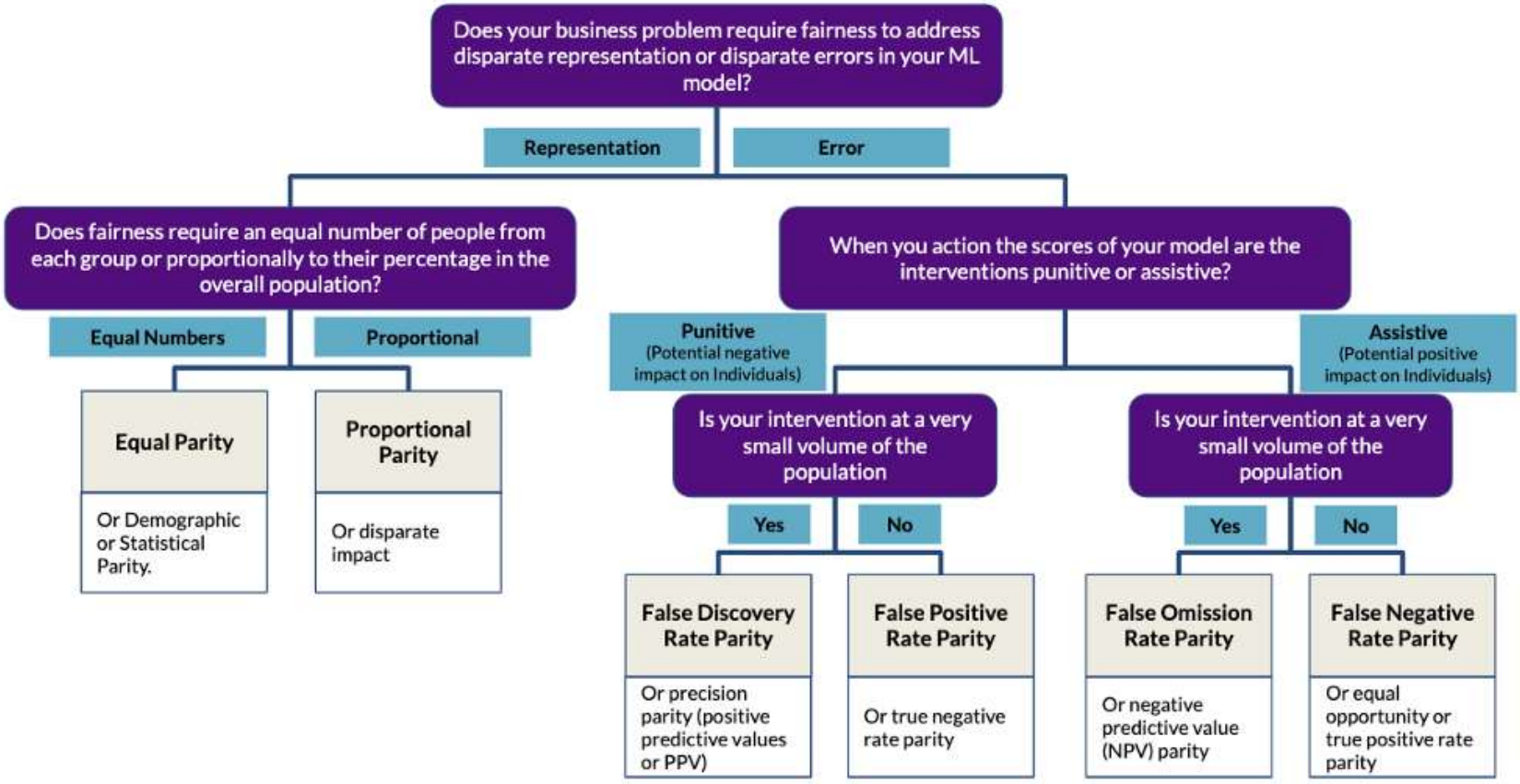
Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Which metrics do we care about for our problem?



The Fairness Tree, Adapted from http://www.datasciencepublicpolicy.org/projects/aequitas/

Use the fairness tree as a tool to support discussions and evaluate scenarios

Punitive Scenario – **Jail Determination**

Metric	Probabilistic Term	Metric Translation to Use Case Domain	Caveat
Equal Count of False Positives	P(wrongly jailed group i)=C ∀ I	*Number of people wrongly jailed from each group is the same*,	If groups are not the same size it can be arguable as a fair criteria.
Group Size-Adjusted False Positives	P(wrongly jailed | group i)=C ∀ i	*Just by virtue of the fact that an individual is a member of a given group, what are the chances they'll be wrongly convicted?*	% innocents is lower on group B Group | Pop. Size | No Bail | FP | FP/Popul. | %Innocent A | 10 000 | 100 | 10 | 0,1% | 90% B | 30 000 | 100 | 30 | 0,1% | 70%
False Discovery Rate (FDR)	P(wrongly jailed | jailed, group i)=C ∀ i	*Focused on the group of people which are interventioned - 100/each group jailed on example above – "discovers" the the ratio of FP on the interventioned sub-groups.*	Considering two distinct groups A & B with a criminal process with equal FDR and group sized-adjusted false positives: Group | Pop. Size | Guilty | Innocent | Decision = Jail A | 1 000 | 900 | 100 | 100 B | 3 000 | 300 | 2700 | 100 $\frac{FP_B}{n_B} = \frac{30}{3000} = 1.0\%$ $\frac{FP_A}{n_A} = \frac{10}{1000} = 1.0\%$ $FDR_B = \frac{30}{300} = 10.0\%$ $FDR_A = \frac{10}{100} = 10.0\%$ $FPR_B = \frac{30}{2730} = 1.1\%$ $FPR_A = \frac{10}{10} = 100.0\%$
False Positive Rate (FPR)	P(wrongly jailed | innocent, group i)=C ∀ i	*For an innocent person, what are the chances they will be wrongly convicted by virtue of the fact that they're a member of a given group?*	

- There'll be cases where it's needed to balance trade-offs while selecting evaluation metrics

Online Confusion Matrix - https://onlineconfusionmatrix.com/

Classification Refresher: Evaluation Metrics

Rank	Score	Label	Predict	Skin Color
1	0.997	1	1	non-white
2	0.993	1	1	white
3	0.986	1	1	white
4	0.982	1	1	non-white
5	0.971	0	1	white
6	0.965	1	1	white
7	0.964	0	1	white
8	0.961	0	1	non-white
9	0.953	0	1	non-white
10	0.932	1	1	non-white
11	0.918	0	0	non-white
12	0.873	0	0	white
13	0.854	0	0	white
14	0.839	0	0	white
15	0.777	0	0	white
16	0.723	0	0	non-white
17	0.634	0	0	white
18	0.512	0	0	non-white
19	0.487	0	0	white
20	0.473	0	0	non-white

Predicted Positive: 10

Predicted Negative: 10

Population

- Total Label Positives: 6
- Total Label Negatives: 14
- Prevalence 6/20=0.3

For threshold > 0.920 or top k=10:

		True Condition	
	Total Population	Positive	Negative
Predicted Condition	Positive	TP 6	FP 4
	Negative	FN 0	TN 10

True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$

False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$

False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$

Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$

- True Positive Rate (Recall) = 6/6 = 1.0
- False Positive Rate: 4/14=0.29

- False Negative Rate = 0/6=0
- True Negative Rate = 10/14=0.71
- Precision = 6/10 =0.6

Adapted from Pedro Saleiros's "Tutorial: Fairness in decision-making with AI: a practical guide & hands-on tutorial using Aequitas https://www.youtube.com/watch?v=yOR71zBm3Uc

Online Confusion Matrix - https://onlineconfusionmatrix.com/

Classification Refresher: Evaluation Metrics

Rank	Score	Label	Predict	Skin Color
1	0.997	1	1	non-white
2	0.993	1	1	white
3	0.986	1	1	white
4	0.982	1	1	non-white
5	0.971	0	1	white
6	0.965	1	1	white
7	0.964	0	1	white
8	0.961	0	1	non-white
9	0.953	0	1	non-white
10	0.932	1	1	non-white
11	0.918	0	0	non-white
12	0.873	0	0	white
13	0.854	0	0	white
14	0.839	0	0	white
15	0.777	0	0	white
16	0.723	0	0	non-white
17	0.634	0	0	white
18	0.512	0	0	non-white
19	0.487	0	0	white
20	0.473	0	0	non-white

Predicted Positive: 10

Predicted Negative: 10

Population

- Total Label Positives: 6
- Total Label Negatives: 14
- Prevalence 6/20=0.3

True positive rate (TPR), Recall, Sensitivity, probability of detection, Power
$= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$

False positive rate (FPR), Fall-out, probability of false alarm
$= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$

False negative rate (FNR), Miss rate
$= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$

Specificity (SPC), Selectivity, True negative rate (TNR)
$= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$

For threshold > 0.920 or top k=10:

Non-White		True Condition	
	Population	Positive	Negative
Predicted Condition	Positive	3	2
	Negative	0	4

White		True Condition	
	Population	Positive	Negative
Predicted Condition	Positive	3	0
	Negative	2	6

Metric	Non-White	White
FPR	2/6=0.33	2/6=0.25
Recall	3/3=1	3/3=1
Precision	3/5=0.6	3/5=0.6
FNR	0	0

Adapted from Pedro Saleiros's "Tutorial: Fairness in decision-making with AI: a practical guide & hands-on tutorial using Aequitas https://www.youtube.com/watch?v=yOR71zBm3Uc

[1] **Adverse Impact Analysis / Four-Fifths Rule**
https://www.prevuehr.com/resources/insights/adverse-impact-analysis-four-fifths-rule/

How to assess level of fairness disparity between groups?

1. Compare a given metric with a reference group (e.g. historically favoured group)

$$FPR_g\ disp = \frac{FPR_{a_i}}{FPR_{a_r}} = \frac{\Pr(\widehat{Y}=1|Y=0,A=a_i)}{\Pr(\widehat{Y}=1|Y=0,A=a_r)}$$

2. Define a fairness range for the disparity measure across groups

$$\tau \leq DisparityMeasure_{group_i} \leq \frac{1}{\tau}$$

This can be regulated depending on domain/geography

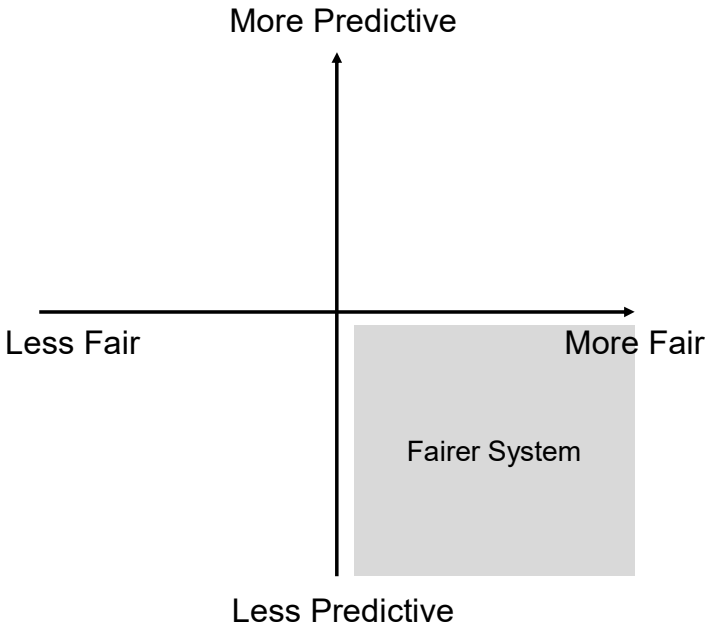
"The four-fifths or 80% rule is described by the guidelines as "a selection rate for any race, sex, or ethnic group which is less than four-fifths (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact…" [1]

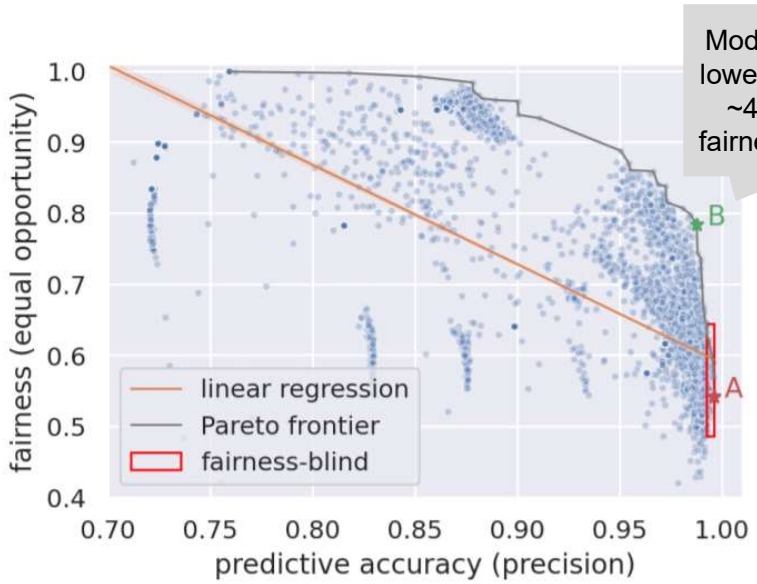
3. Calculate & evaluate disparate impact

Models	Accuracy	Disparate impact
TF without mitigation	0.73	0.16
TF with adversarial debiasing	0.71	0.42

[1] A Bandit-Based Algorithm for Fairness-Aware Hyperparameter Optimization
https://arxiv.org/abs/2010.03665

Once defined one (or a set) of fairness definition(s) we can do a better judgment of which model best suits us





Fairness-accuracy trade-off of thousands of models [1]

A Survey on Bias and Fairness in Machine Learning from (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019)

Fairness is a cross-sectional topic in AI with different domains already benefiting of growing research activity

In a recent survey different domains & subdomains of ML/AI are examined to structure research around state-of-art methods to mitigate bias

Domain	Sub-Domain	Domain	Sub-Domain	Domain	Sub-Domain
ML	Binary Classification	NLP	Word Embedding	DNN	Variational Auto-Encoders
	Regression		Coreference Resolution		Adversarial Learning
	PCA		Language Model		
	Community Detection		Sentence embedding		
	Clustering		Machine translation		
	Graph embedding		Semantic Role Labeling		
	Causal Inference		Classification		

https://www.linkedin.com/pulse/overview-some-available-fairness-frameworks-packages-murat-durmus/
https://modeloriented.github.io/fairmodels/articles/Advanced_tutorial.html#tradeoff-between-bias-and-accuracy-1

There is a proliferation of open source frameworks to support the development of Fair models

Not Exhaustive

Tool	Latest Commit	Nbr of Commits	Nbr Branches	Nbr Releases	Nbr of Contributors	Github Reference
Tensorflow – what if tool *	30/08/2019	2.776	21	25	176	tensorflow/tensorboard/tree/master/tensorboard/plugins/interactive_inference
SHAP	31/08/2019	964	3	33	64	slundberg/shap
Aequitas	11/07/2019	832	11	35	8	dssg
Fairness Comparison	05/06/2019	503	10	2	10	algofairness/fairness-comparison
Fairtest	29/05/2017	474	5	0	7	columbia/fairtest
LIME	23/07/2019	433	5	17	34	marcotcr/lime
Fairness Measures	29/12/2017	260	3	0	3	megantosh/fairness_measures_code
AIF360	30/08/2019	149	26	4	16	IBM/AIF360
Fairml	23/03/2017	104	3	0	2	adebayoj/fairml
ThemisML	31/07/2018	65	1	3	1	cosmicBboy/themis-ml
FAT-Forensics	03/05/2019	59	1	0	4	liamdalg/FAT-Forensics
Microsoft Fair Learn	09/08/2019	37	10	2	6	github.com/microsoft/fairlearn
Audit-AI	25/06/2019	35	2	2	6	pymetrics/audit-ai

https://www.linkedin.com/pulse/overview-some-available-fairness-frameworks-packages-murat-durmus/
https://modeloriented.github.io/fairmodels/articles/Advanced_tutorial.html#tradeoff-between-bias-and-accuracy-1

There is a proliferation of open source frameworks to support the development of Fair models

Not Exhaustive

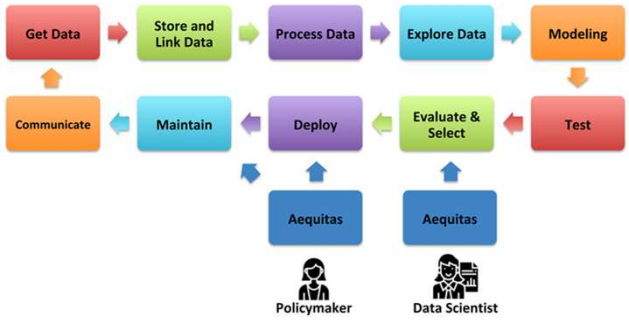
Tool	Latest Commit	Nbr of Commits	Nbr Branches	Nbr Releases	Nbr of Contributors	Github Reference
Tensorflow – what if tool *	30/08/2019	2.776	21	25	176	tensorflow/tensorboard/tree/master/tensorboard/plugins/interactive_inference
SHAP	31/08/2019	964	3	33	64	slundberg/shap
Aequitas	11/07/2019	832	11	35	8	dssg
Fairness Comparison	05/06/2019	503	10	2	10	algofairness/fairness-comparison
Fairtest	29/05/2017	474	5	0	7	columbia/fairtest
LIME	23/07/2019	433	5	17	34	marcotcr/lime
Fairness Measures	29/12/2017	260	3	0	3	megantosh/fairness_measures_code
AIF360	30/08/2019	149	26	4	16	IBM/AIF360
Fairml	23/03/2017	104	3	0	2	adebayoj/fairml
ThemisML	31/07/2018	65	1	3	1	cosmicBboy/themis-ml
FAT-Forensics	03/05/2019	59	1	0	4	liamdalg/FAT-Forensics
Microsoft Fair Learn	09/08/2019	37	10	2	6	github.com/microsoft/fairlearn
Audit-AI	25/06/2019	35	2	2	6	pymetrics/audit-ai

Bias and Fairness Audit Toolkit
http://aequitas.dssg.io/
http://www.datasciencepublicpolicy.org/projects/aequitas/

Aequitas – Bias & Fairness Audit

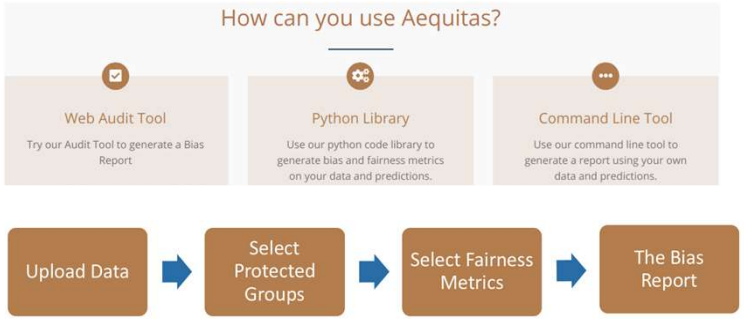
Aequitas is an open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive tools.

What is the Target Audience?



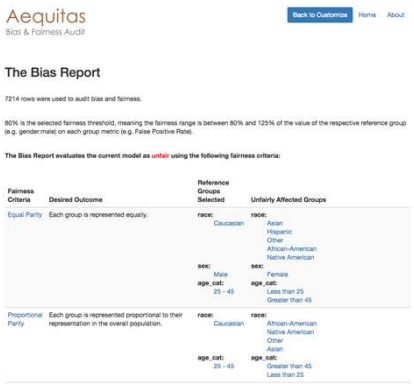
- Aequitas was designed having in mind technical Data profiles but also policymakers creating a common ground for discussion.

How to use it?



- For non-tech profiles a bias report is quite straightforward to get through web app – also there are available python api and batch loading of data.

What do we get?



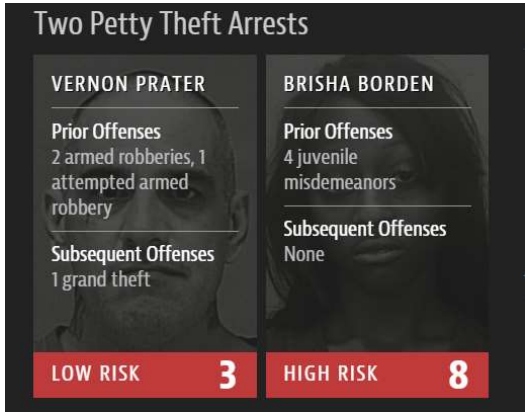
- Aequitas provides a detailed audit to your models calculating detailed fairness and bias statistics, mitigation can be done through an informed model selection.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study – Compas Recidivism Risk

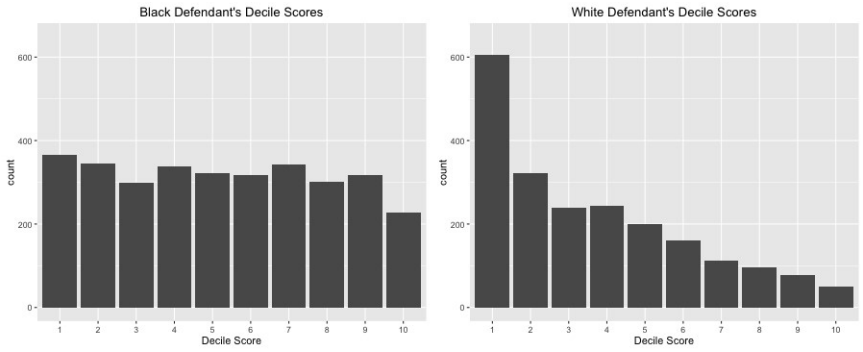
Risk models are used in criminal justice to inform high-stakes decisions in people's life such as bail, sentencing and early release. In May 2016 journalists at Propublica explored the bias in such models, choosing a more widely solution in use – Correctional Offender Management Profiling for Alternative Sanctions (Compas)

COMPAS Scores & Profiles



- Scores of individuals with inconsistent criminal record were found to be unreliable by Propublica journalists.

COMPAS scores for "Risk of Recidivism"



- Propublica showed that scores for white defendants were skewed toward lower-risk categories, while black defendants were evenly distributed across scores.

COMPAS Fairness Metrics

Fairness Metric	Caucasian	African American
False Positive Rate (FDR)	23%	45%
False Negative Rate (FNR)	48%	28%
False Discovery Rate (FDR)	41%	37%

- Propublica and Northpoint (Compas vendor) were arguing over different Fairness Metrics

http://aequitas.dssg.io/audit/27py3o00/compas_for_aequitas/report-1.html#false-positive-rate-parity-span-red-initfailedspan-red-end

Case Study – Compas Recidivism Risk

Checking audit results on the Aequitas Bias Report example for Compas use case we can find same results as reported by Propublica

Audit Results: Group Metrics Values

race

Attribute Value	Group Size Ratio	Predicted Positive Rate	Predicted Positive Group Rate	False Discovery Rate	False Positive Rate	False Omission Rate	False Negative Rate
African-American	0.51	0.66	0.59	0.37	0.45	0.35	0.28
Asian	0	0.0	0.25	0.25	0.09	0.12	0.33
Caucasian	0.34	0.26	0.35	0.41	0.23	0.29	0.48
Hispanic	0.09	0.06	0.3	0.46	0.21	0.29	0.56
Native American	0	0.0	0.67	0.25	0.38	0.17	0.1
Other	0.05	0.02	0.21	0.46	0.15	0.3	0.68

Go to Previous

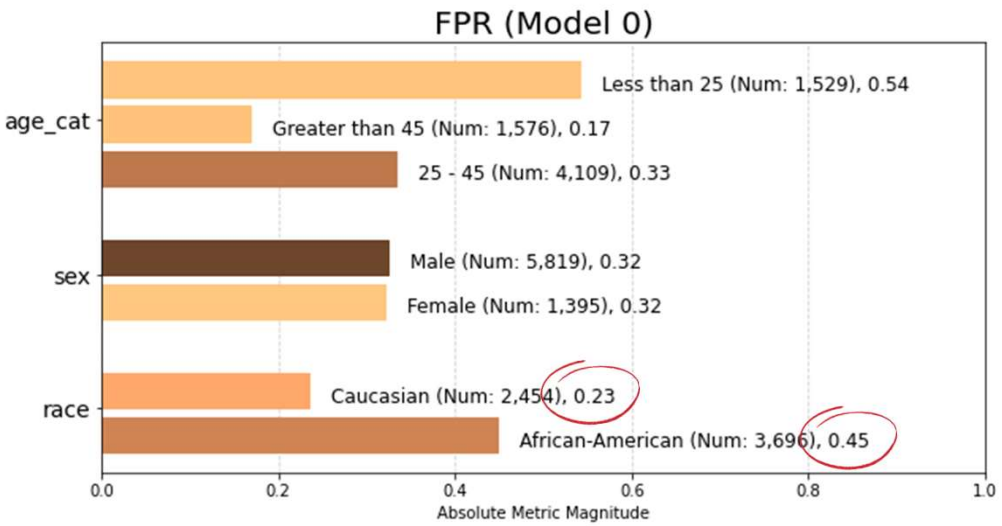
Go to Top

https://github.com/dssg/aequitas/blob/master/docs/source/examples/compas_demo.ipynb

Case Study – Compas Recidivism Risk

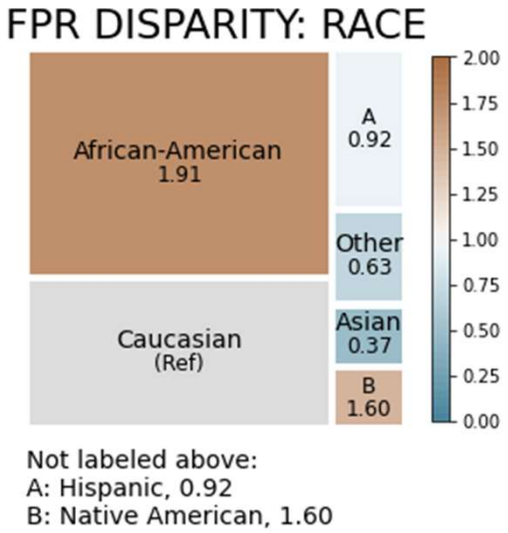
If we need more flexibility, we can easily use aequitas python api to audit our Compas use case

Fairness Metrics for Group above volume threshold




```
fnr = aqp.plot_group_metric(xtab, 'fpr', min_group_size=0.05)
```

Visualizing disparities between groups using Treemap




```
aqp.plot_disparity(bdf, group_metric='fpr_disparity',
attribute_name='race', significance_alpha=0.05)
```

https://www.linkedin.com/pulse/overview-some-available-fairness-frameworks-packages-murat-durmus/
https://modeloriented.github.io/fairmodels/articles/Advanced_tutorial.html#tradeoff-between-bias-and-accuracy-1

There is a proliferation of open source frameworks to support the development of Fair models

Not Exhaustive

Tool	Latest Commit	Nbr of Commits	Nbr Branches	Nbr Releases	Nbr of Contributors	Github Reference
Tensorflow – what if tool *	30/08/2019	2.776	21	25	176	tensorflow/tensorboard/tree/master/tensorboard/plugins/interactive_inference
SHAP	31/08/2019	964	3	33	64	slundberg/shap
Aequitas	11/07/2019	832	11	35	8	dssg
Fairness Comparison	05/06/2019	503	10	2	10	algofairness/fairness-comparison
Fairtest	29/05/2017	474	5	0	7	columbia/fairtest
LIME	23/07/2019	433	5	17	34	marcotcr/lime
Fairness Measures	29/12/2017	260	3	0	3	megantosh/fairness_measures_code
AIF360	30/08/2019	149	26	4	16	IBM/AIF360
Fairml	23/03/2017	104	3	0	2	adebayoj/fairml
ThemisML	31/07/2018	65	1	3	1	cosmicBboy/themis-ml
FAT-Forensics	03/05/2019	59	1	0	4	liamdalg/FAT-Forensics
Microsoft Fair Learn	09/08/2019	37	10	2	6	github.com/microsoft/fairlearn
Audit-AI	25/06/2019	35	2	2	6	pymetrics/audit-ai

https://github.com/Trusted-AI/AIF360/
https://www.slideshare.net/AnimeshSingh/aif360-trusted-and-fair-ai

AIF360 – Bias & Fairness Audit

AI Fairness 360
https://github.com/IBM/AIF360

AIF360 toolkit is an open-source library to help detect and remove bias in machine learning models.

The AI Fairness 360 Python package includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models.

Toolbox
Fairness metrics (70+)
Fairness metric explanations
Bias mitigation algorithms (10)

IBM CODE

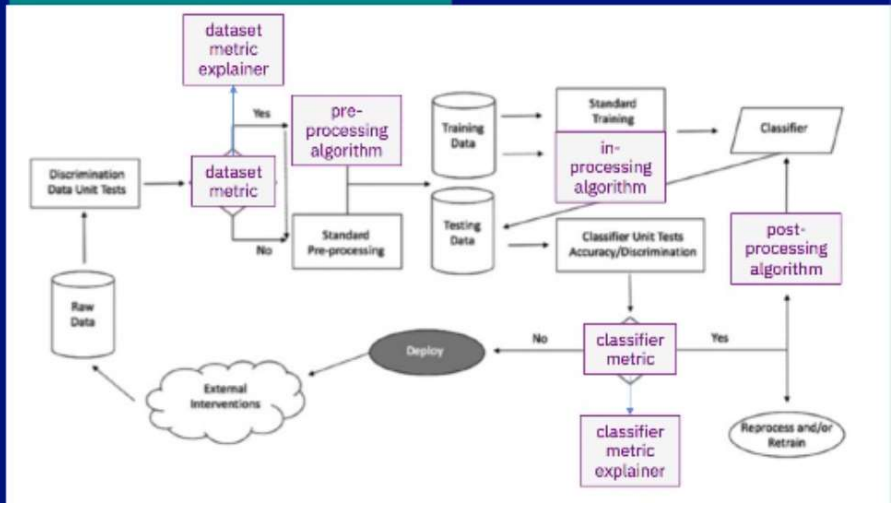
AIF360

Supported bias mitigation algorithms

Optimized Preprocessing (Calmon et al., 20
Disparate Impact Remover (Feldman et al., 2015)
Equalized Odds Postprocessing (Hardt et al., 2016)
Reweighing (Kamiran and Calders, 2012)
Reject Option Classification (Kamiran et al., 2012)
Prejudice Remover Regularizer (Kamishima et al., 2012)
Calibrated Equalized Odds Postprocessing (Pleiss et al., 2017)
Learning Fair Representations (Zemel et al., 2013)
Adversarial Debiasing (Zhang et al., 2018)

Supported fairness metrics

Comprehensive set of group fairness metrics derived from selection rates and error rates
Comprehensive set of sample distortion metrics
Generalized Entropy Index (Speicher et al., 2018)



We can group the existing unfairness mitigation methods in 4 different classes

Not Exhaustive

1 Data Collection

- Better investment in data collection processes as a effective way to migitate bias and/or better data stewardship processes.

- Datasheets for Datasets (Gebru, et al. 2018)
- Bias Source Analysis (I. Chen, Johansson, & Sontag, 2018)

2 Pre-Processing

- Mitigate bias on the dataset tipically by learning a data transformation or oversampling minority classes;

- Optimized Preprocessing (Calmon et al., 2017)
- Manipulating labels (Kamiran & Calders, 2012);
- Reweighing (Kamiran & Calders, 2012);

3 In-Processing

- Modifications of existing learning algorithms to minimize discrimination during the training stage can be grouped as in-processing techniques

- Prejudice removal by regularization (T Kamishima, Akaho, & Sakuma, 2011);
- Adversarial de-biasing: (Beutel, Chen, Zhao, & Chi, 2017)

4 Post-Processing

- Performed after model training, it relies typically on the prediction, protected attribute and ground truth label.

- Equalized Odds Postprocessing (Hardt et al., 2016)

We can group the existing unfairness mitigation methods in 4 different classes

Not Exhaustive

1 Data Collection	2 Pre-Processing	3 In-Processing	4 Post-Processing
▪ Better investment in data collection processes as a effective way to migitate bias and/or better data stewardship processes.	▪ Mitigate bias on the dataset tipically by learning a data transformation or oversampling minority classes;	▪ Modifications of existing learning algorithms to minimize discrimination during the training stage can be grouped as in-processing techniques	▪ Performed after model training, it relies typically on the prediction, protected attribute and ground truth label.
▪ Datasheets for Datasets (Gebru, et al. 2018) ▪ Bias Source Analysis (I. Chen, Johansson, & Sontag, 2018)	▪ Optimized Preprocessing (Calmon et al., 2017) ▪ Manipulating labels (Kamiran & Calders, 2012); ▪ Reweighing (Kamiran & Calders, 2012);	▪ Prejudice removal by regularization (T Kamishima, Akaho, & Sakuma, 2011); ▪ Adversarial de-biasing: (Beutel, Chen, Zhao, & Chi, 2017)	▪ Equalized Odds Postprocessing (Hardt et al., 2016)

https://towardsdatascience.com/reducing-bias-from-models-built-on-the-adult-dataset-using-adversarial-debiasing-330f2ef3a3b4

Adversarial debiasing is one of the most popular in-processing techniques

Intuition

- Train a model that **predicts Y from X, but encodes information from the protected atribute Z in a way that is not recoverable** by an adversarial model.

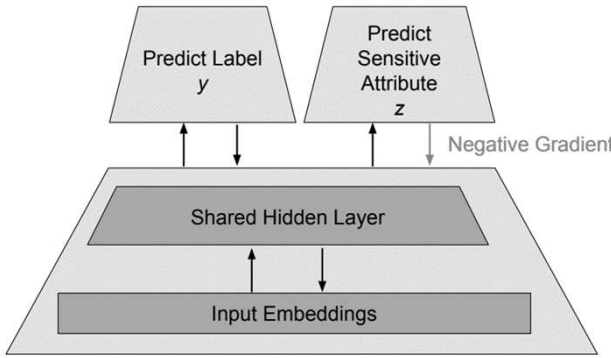
Goal

- Learn a representation g(X) that makes it **easy for f to predict Y, but difficult for a to predict Z**. In optimization terms, we want to **minimize our prediction loss L_y(f(g(X)), Y) and maximize the adversarial loss L_z(a(g(X)), Z).**

Model

We let *f* be our prediction function where Y=*f*(*g*(X))

Let a be our adversary where Z=a(g(X))



g(X) to be the shared learned embedding of our input.

Results

Models	Accuracy	Disparate impact
TF without mitigation	0.73	0.16
TF with adversarial debiasing	0.71	0.42

For a loss of 2pp accuracy we have improved 2.7x the DI fairness metric

Method is also:

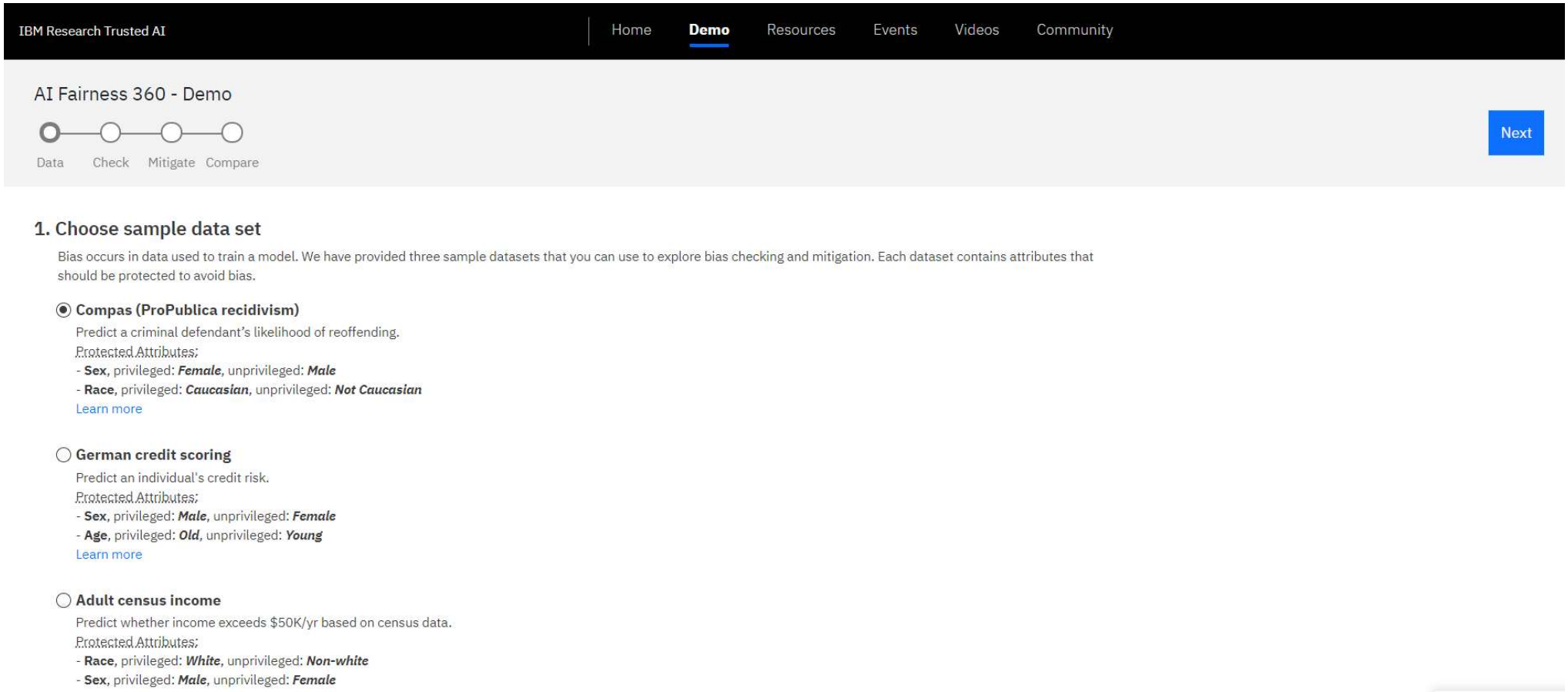
- Flexible, it can be tuned for the desired trade-off between "fairness" and accuracy (or other metric);
- Applicable to different fairness definitions;
- Can support regression and classification.

$$DI = \frac{P_{positive\ unprivileged}}{P_{positive\ privileged}}$$

https://github.com/Trusted-AI/AIF360/
https://www.slideshare.net/AnimeshSingh/aif360-trusted-and-fair-ai
http://aif360.mybluemix.net/

AIF360 – Bias & Fairness Audit

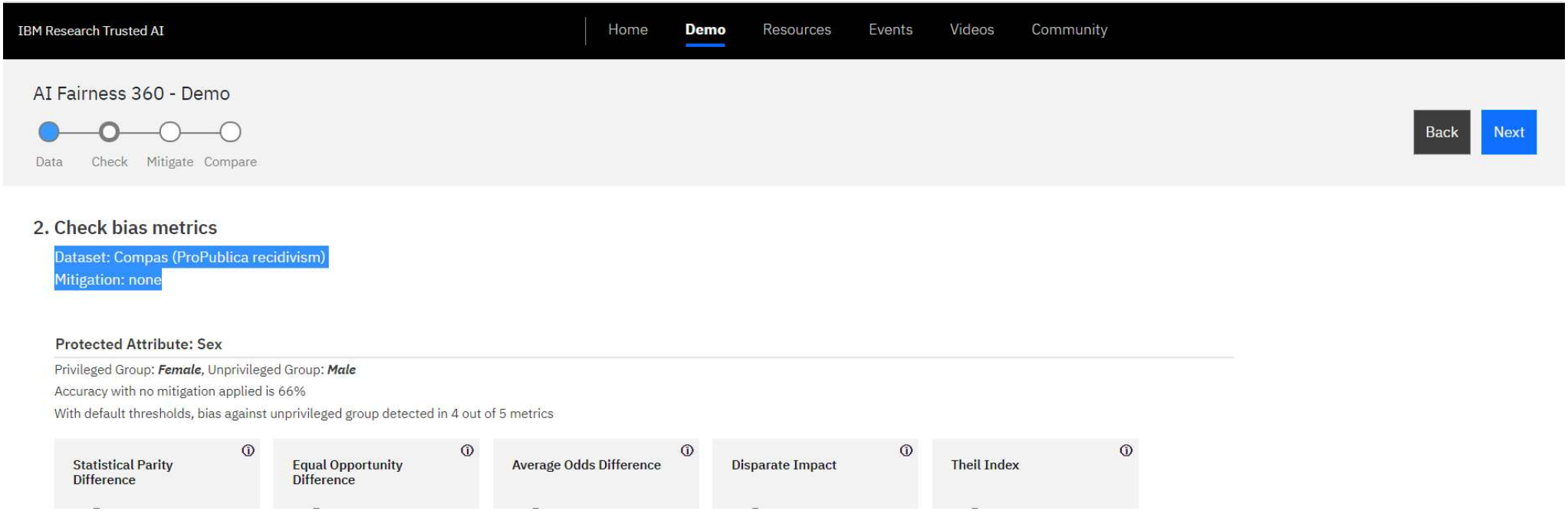
We can test the same dataset but using a mitigation method



https://github.com/Trusted-AI/AIF360/
https://www.slideshare.net/AnimeshSingh/aif360-trusted-and-fair-ai
http://aif360.mybluemix.net/

AIF360 – Bias & Fairness Audit

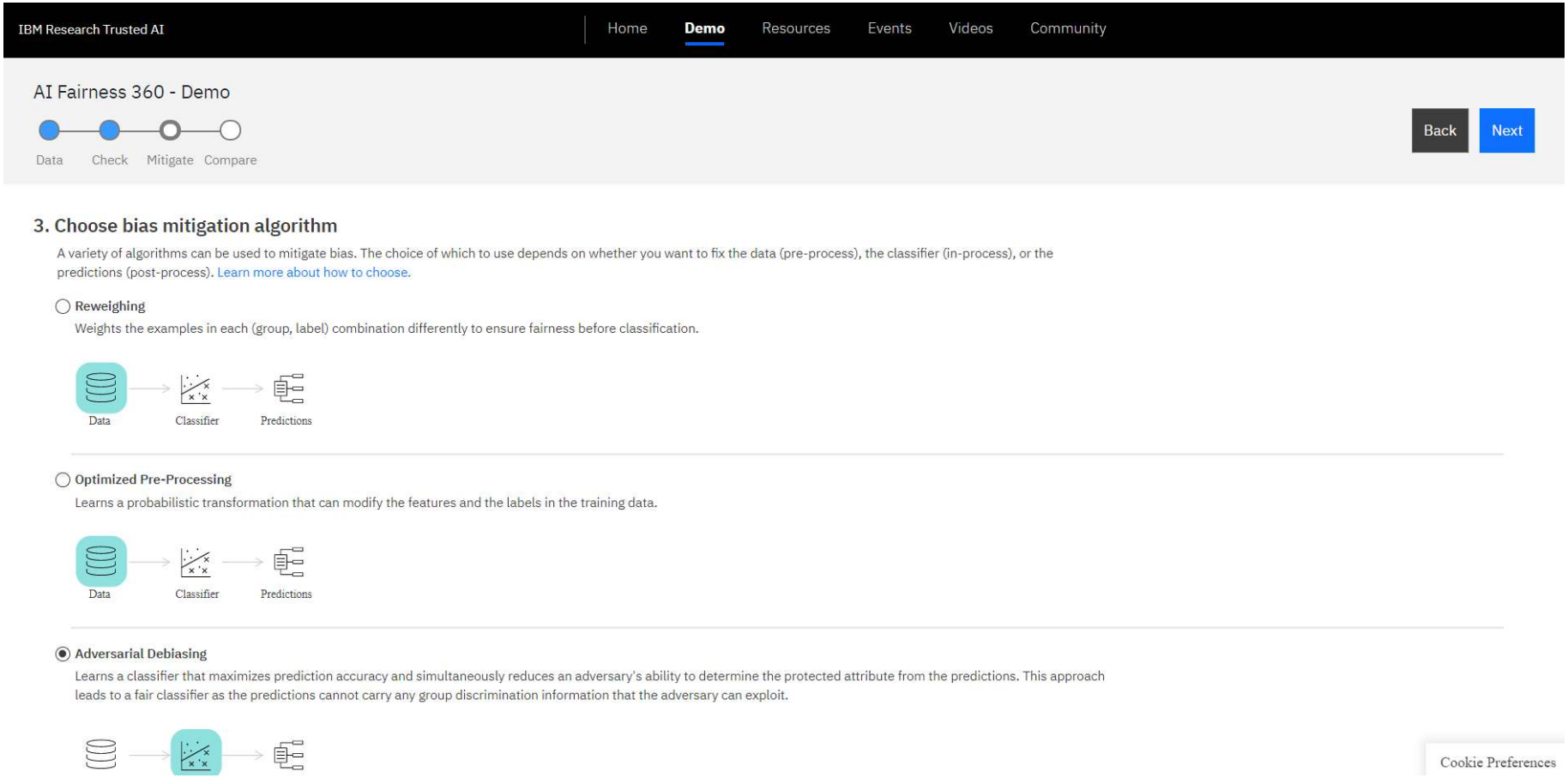
We can test the same dataset but using a mitigation method



https://github.com/Trusted-AI/AIF360/
https://www.slideshare.net/AnimeshSingh/aif360-trusted-and-fair-ai
http://aif360.mybluemix.net/

AIF360 – Bias & Fairness Audit

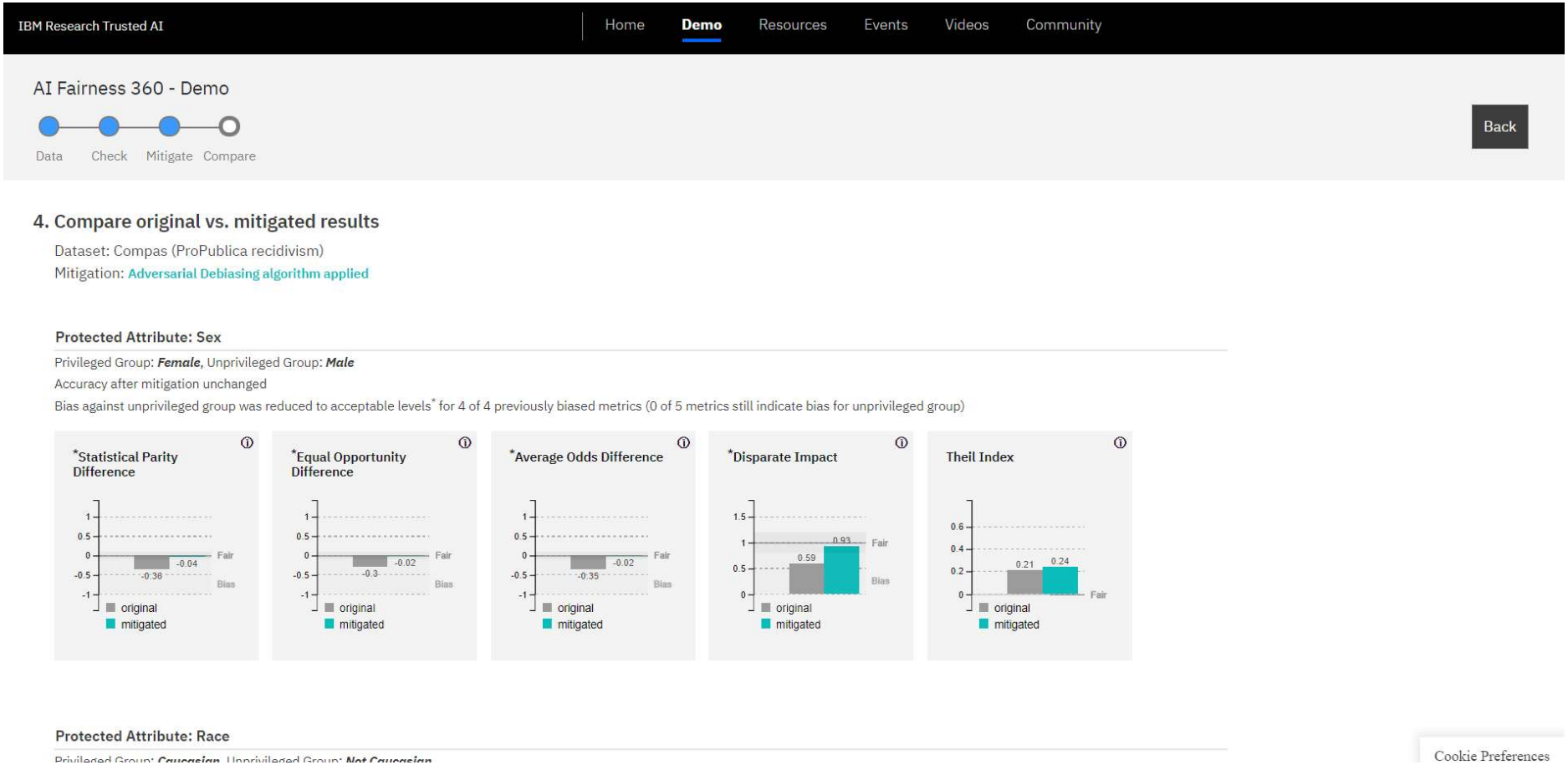
We can test the same dataset but using a mitigation method



https://github.com/Trusted-AI/AIF360/
https://www.slideshare.net/AnimeshSingh/aif360-trusted-and-fair-ai
http://aif360.mybluemix.net/

AIF360 – Bias & Fairness Audit

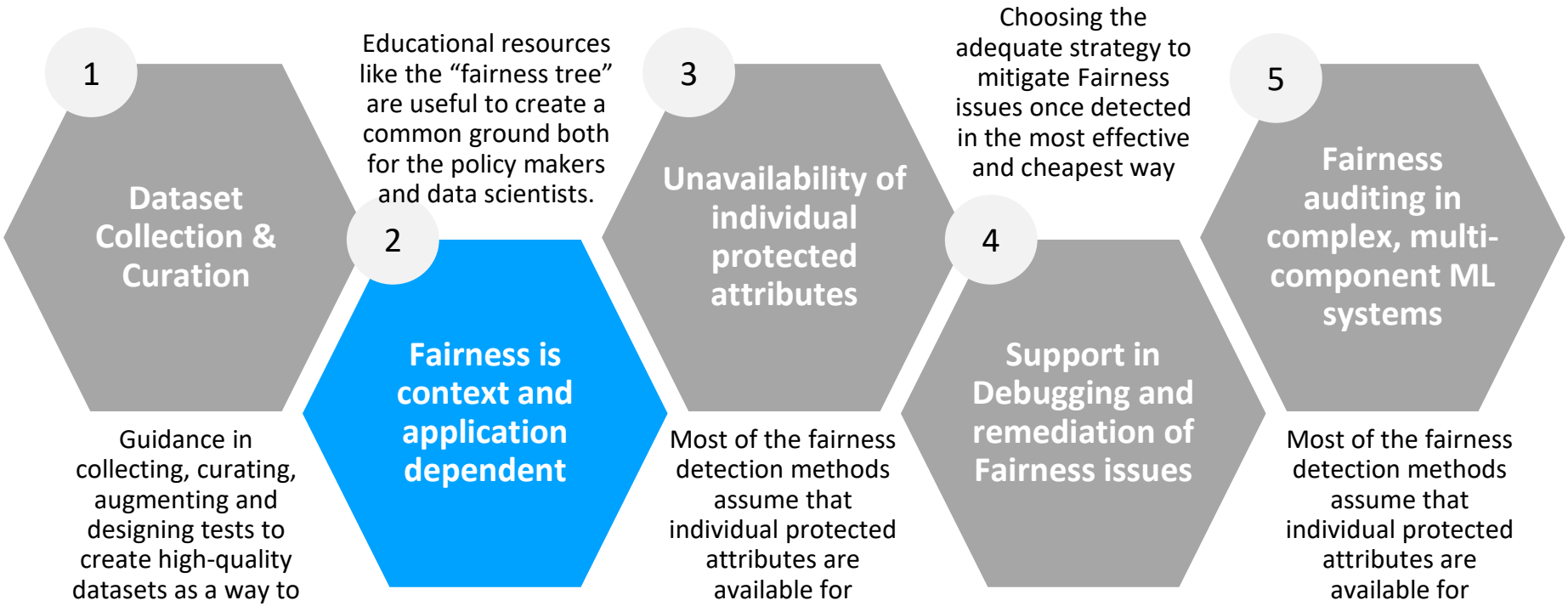
We can test the same dataset but using a mitigation method



*Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? 2019 ACM CHI Conference on Human Factors in Computing Systems.

Even for the existing (global) practice there are open challenges

A recent study[1] identifies challenges that are not so much covered by existing research topics on Fairness



We need to address and share more industry use-cases patterns

Illustrative

Use Case	Fairness Discussion	Existing Research
Pricing / Next-Best-Offer	▪ In acquisition are we covering all the socio-economical segments? ▪ What pricing/offer is being provided to different groups of Customers and how fair is the current practice? ▪ Which are the possible definitions of a Fair Price?	▪ Seele, P., Dierksmeier, C., Hofstetter, R., & Schultz, M. D. (2019). Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing. Journal of Business Ethics. https://doi.org/10.1007/s10551-019-04371-w ▪ Maestre, R., Duque, J., Rubio, A., & Arévalo, J. (2018). Reinforcement Learning for Fair Dynamic Pricing. ArXiv E-Prints, arXiv:1803.09967.
Content Recommendation Systems	▪ Is cultural or popularity bias reinforcing recommendations? ▪ Which metrics such diversity could be a more "fair" metric? Is it a trade-off, or can we find items dissimilar but relevant for user's preferences? ▪ If a product is provided for a household and users are obfuscated how can we define fairness as well? ▪ Should the suppliers have a fair opportunity?	▪ Porcaro, L., Castillo, C., & Gómez, E. (2019). Music Recommendation Diversity: A Tentative Framework and Preliminary Results. ▪ Yao, S., & Huang, B. (2017). Beyond Parity: Fairness Objectives for Collaborative Filtering. CoRR, abs/1705.08804. http://arxiv.org/abs/1705.08804 ▪ Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (2018). Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. 2243–2251. https://doi.org/10.1145/3269206.3272027
Customer Service & Operations	▪ What defines a fair level of service either in a Customer Service context? For instance how do I fairly route calls to agent or set up priotization in tasks? ▪ What other more complex applications such as speech to text and associated nlp applications have Fairness opportunities? ▪ What defines a fair Quality of Service?	▪ Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117; ▪ Timan, T., & Grommé, F. (2020). *A framework for social fairness; Insights from two algorithmic decision-making controversies in The Netherlands*.

Agenda

Today: encounter new ideas, engage through discussions with your peers and reflect on your learning.

Time	Topic
9:00 am	Welcome, logistics & agenda
9:15 am	**Common Pitfalls in Data Science**
10:00 am	Group Challenge #1
10:30 am	Break: Grab your coffee!
10:40 am	**DS Ethics**
11:10 am	Group Challenge #2

Time	Topic
11:50 am	Break
12:00 am	**Fairness in Machine Learning**
12:10 am	Group Challenge #3
12:30 am	Audit & Mitigation
12:55 am	Wrap up & Farewell

Key takeaways

- Include fairness and equity as part of AI Systems development process;
- Make sure that fairness metrics are translated to the ones that make sense to your context;
- Audit bias & fairness;
- Explore bias reduction strategies;

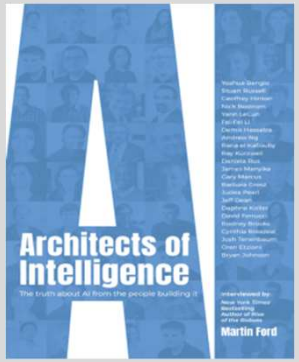
Call2action

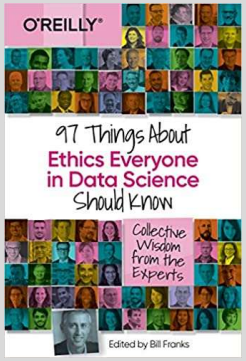
What could be ideas to kick-start or accelerate these practices starting tomorrow?

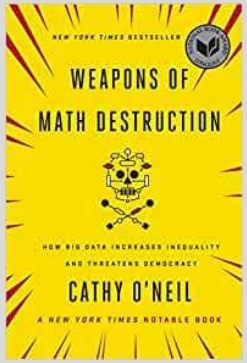
- High-level list your internal DS initiatives – current & about to start;
- Prepare a strutured workshop to go through bias & fairness potential harm impacts;
- Schedule a workshop, discuss and fill list potential harms;
- Structure a brief presentation to management;
- Discuss next steps;

Resources

Inspire







Acquire

- Berkeley CS 294: Fairness in ML
- Cornell INFO 4270: Ethics and policy DS
- Princeton COS 597E: Fairness in ML
- CMU 10718/94889: ML for Public Policy Lab
- KDD 2020 Hands-on Tutorial

Engage

Data Science for Social Good
PORTUGAL