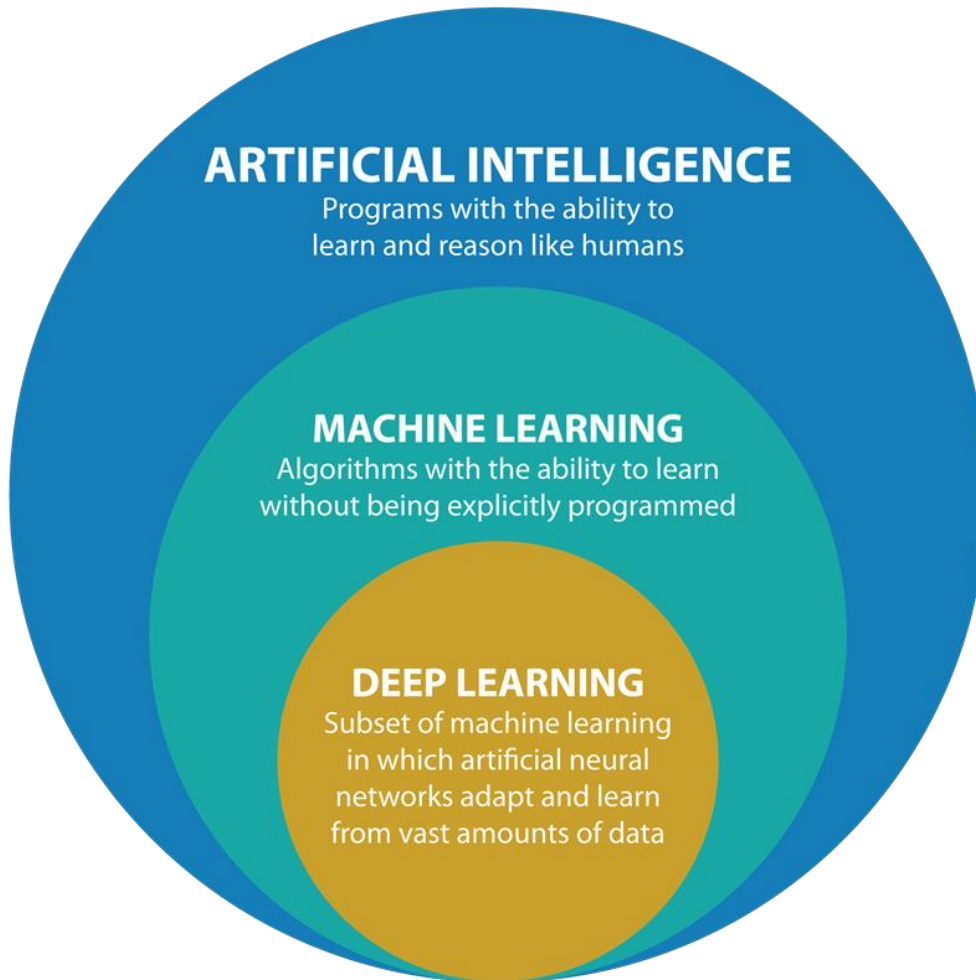


Algorithmic Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making

Adriano Koshiyama

- ❖ **Introduction to AI & Machine Learning (Algorithms)**
- ❖ **Key Components of Algorithmic Impact Assessment**
- ❖ **Algorithmic Explainability**
- ❖ **Algorithmic Fairness**
- ❖ **Algorithmic Robustness**

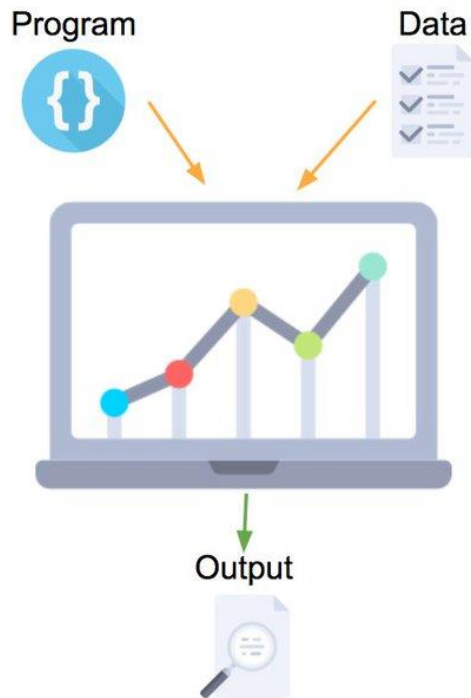
- ❖ Principles and goals
- ❖ Ways to learn in a nutshell
- ❖ Supervised Learning
- ❖ Evaluating Supervised Learning
- ❖ Typical modelling pipeline
- ❖ When learning works, and fails...
- ❖ Further reading



“Whenever I hear people saying AI is going to hurt people in the future I think, yeah, technology can generally always be used for good and bad and you need to be careful about how you build it ...”

Mark Zuckerberg

Traditional Programming



Goal: generating an Output
through Data (2,2) & Program (+)
 $(2 + 2 = ?)$

Machine Learning

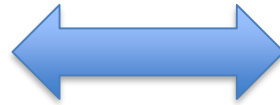


Goal: learning a Program
through Data (2,2) & Output (4)
 $(2 ? 2 = 4)$

Traditional Programming



Both are Algorithms!



Machine Learning



Often

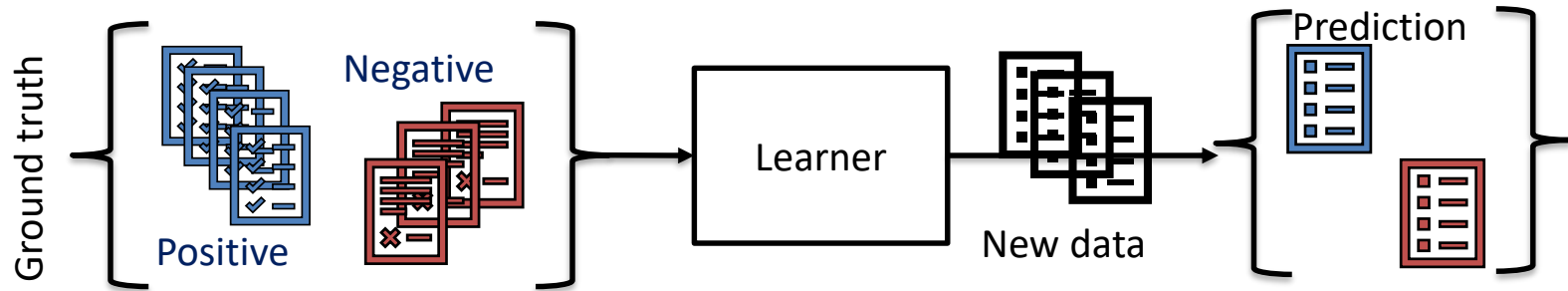
- ❖ Static
- ❖ Rule-based
- ❖ Easier to verify

Often

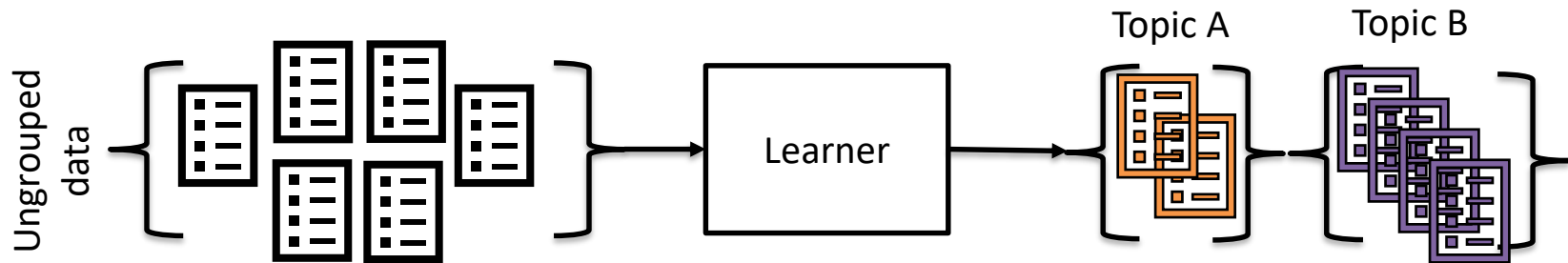
- ❖ Dynamic
- ❖ Functional
- ❖ Harder to verify

Ways to learn in a nutshell

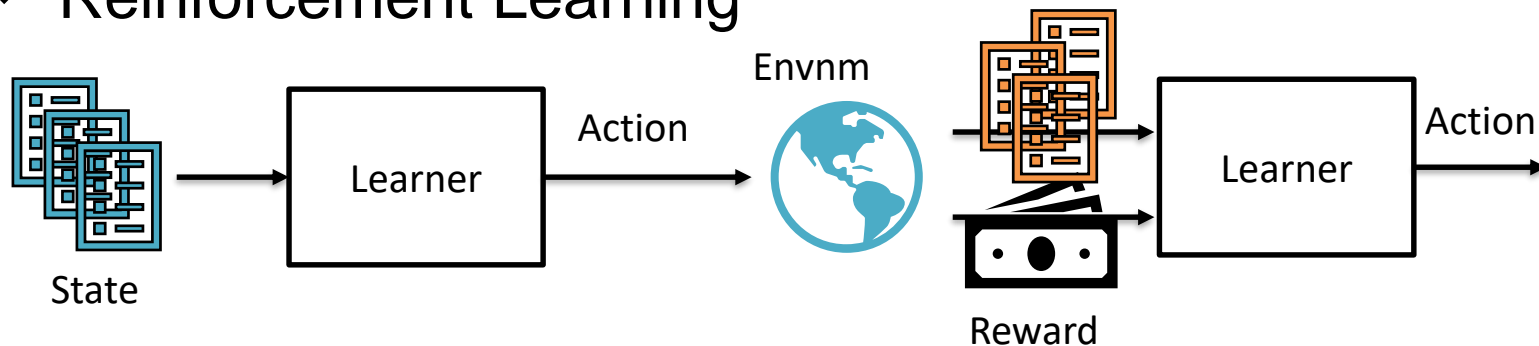
❖ Supervised Learning



❖ Unsupervised Learning

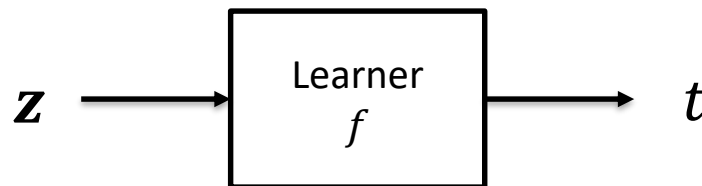


❖ Reinforcement Learning



❖ Formal definition

- ❖ given a set of pairs $D = \{(\mathbf{z}_i, v_i)\}_{i=1}^n$ where
 - ❖ Inputs/Independent/Features/RHS: $\mathbf{z}_i = (z_1, \dots, z_j, \dots, z_J)$
 - ❖ Output/Dependent/Target/LHS: t_i
 - ❖ D is the dataset with $i = 1, \dots, n$ samples
- ❖ Our goal is to uncover the functional link $f: Z \rightarrow V, f(\mathbf{z}) \approx t$



- ❖ Depending on how we label t , learning f is called
 - ❖ Classification: if t can take values in a finite set (e.g. {Yes, No})
 - ❖ Regression: if t can take values in an interval (e.g. [-10, 10])

Classification (recognition):
What?



t
 $\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$
Dog
Cat
Person
...

Encoding
(conv&pool)

Feature
map

Combining
features

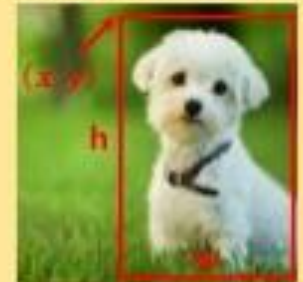
$+$ $=$

Bounding box regression (localization):
Where?



t
 $\begin{bmatrix} x \\ y \\ w \\ h \end{bmatrix}$

Objection Detection

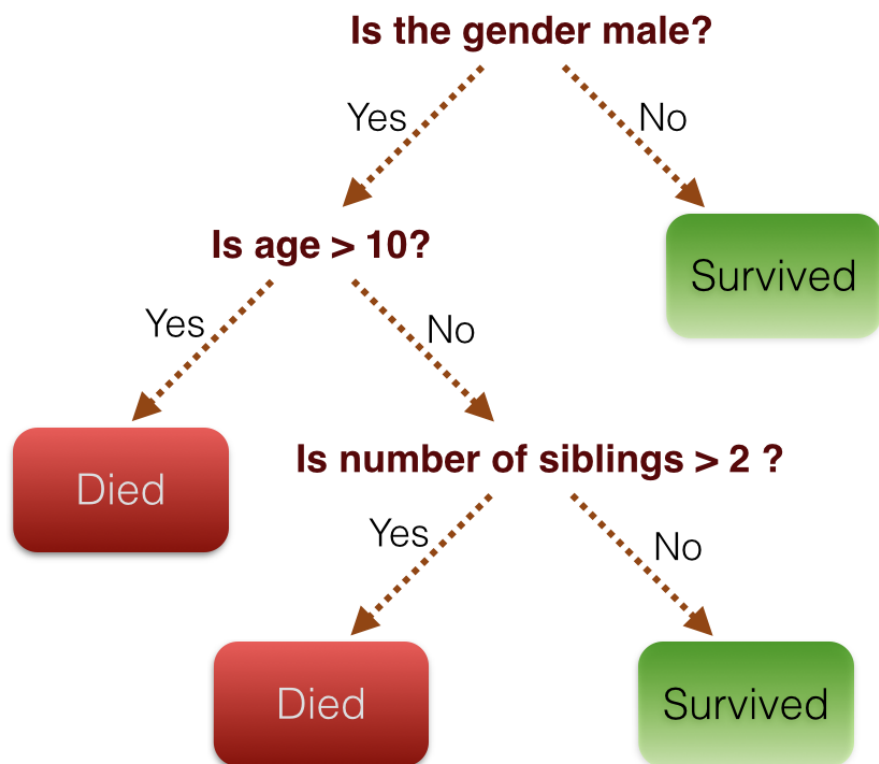


A dog at (x, y, w, h)

Bounding box information

- (x, y) : top left corner position
- w = width
- h = height

❖ Predicting who survived from the Titanic disaster



Decision Tree Confusion Matrix

Obs // Pred	Yes	No
Yes	4	1
No	1	4

Name	Sex	Age	Siblings	Survived	Prediction
Mr. William Thompson Sloper	M	28	0	Yes	No
Mrs. John Bradley (Florence Briggs Thayer) Cumings	F	38	1	Yes	Yes
Miss. Laina Heikkinen	F	26	0	Yes	Yes
Miss. Torborg Danira Palsson	F	8	3	No	Yes
Mr. William Henry Allen	M	35	0	No	No
Mr. James Moran	M	27	0	No	No
Mr. Timothy J McCarthy	M	54	0	No	No
Master. Gosta Leonard Palsson	M	2	3	No	No
Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	F	27	0	Yes	Yes
Mrs. Nicholas (Adele Achem) Nasser	F	14	1	Yes	Yes

Evaluating supervised learning

❖ Confusion matrix-based metrics

		True condition			
		Total population	Condition positive	Condition negative	
				Prevalence $= \frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$
Predicted condition	Predicted condition positive	True positive, Power 4	False positive, Type I error 1	Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error 1	True negative 4	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$
		False negative rate (FNR), Miss rate $= \frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$	
		F ₁ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$			

Decision Tree Performance

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Decision Tree	80%	80%	80%	0.8

Typical modelling pipeline

Data and Task Setup

- Preparing queries and fetching data
- Defining the learning task (Classification, Regression, etc.)



Feature pre-processing and Engineering

- Scaling and transforming some features
- Creating new features, using z-scores, lagging, embedding, etc.



Model Selection

- Defining baselines, performance metrics and model evaluation
- Hypothesis space and hyperparameters to explore



Post-processing/Reporting

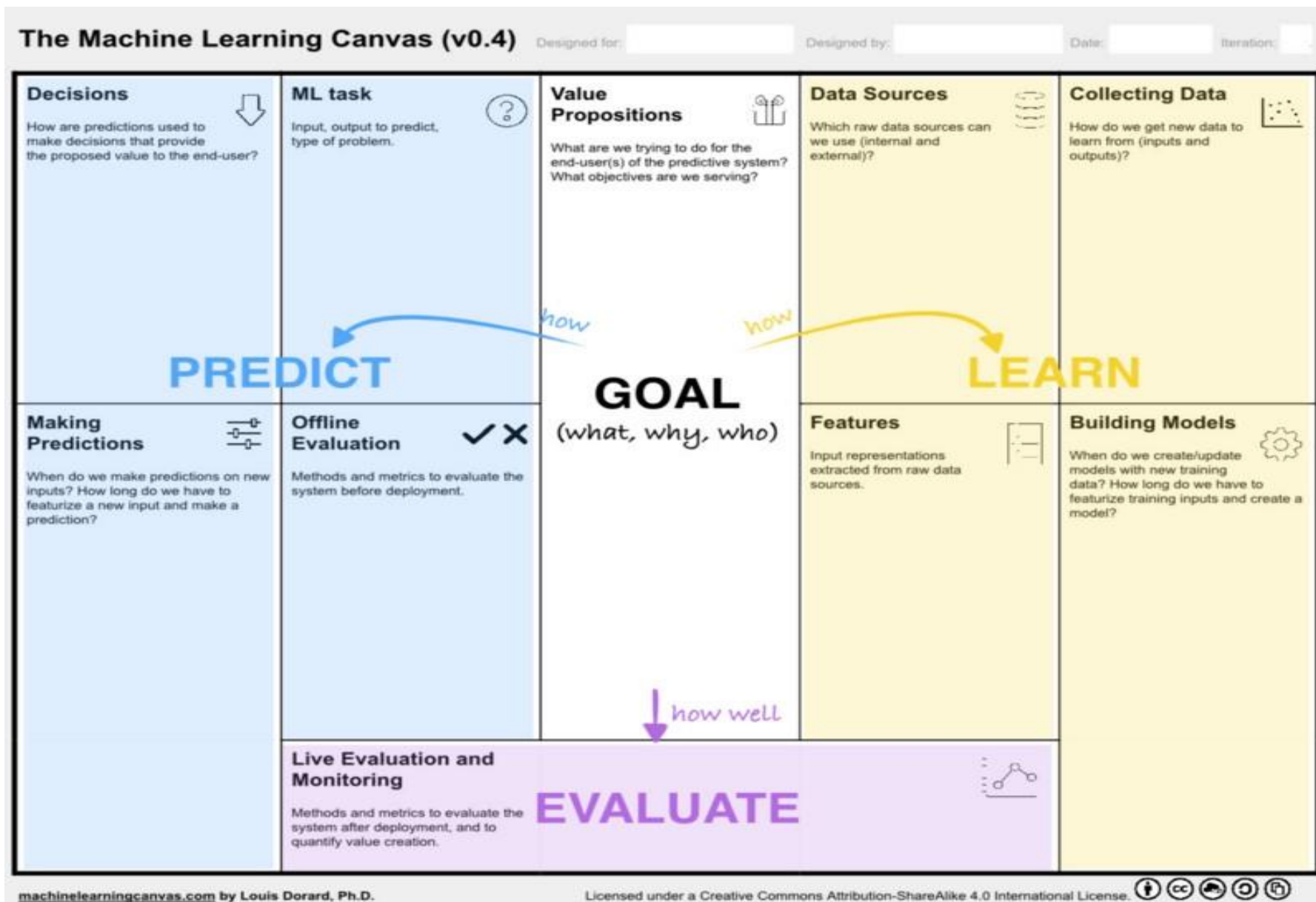
- Adding constraints to some models, feature importance, etc.
- Plotting results, compiling metrics and showing value



Productionizing and Deploying

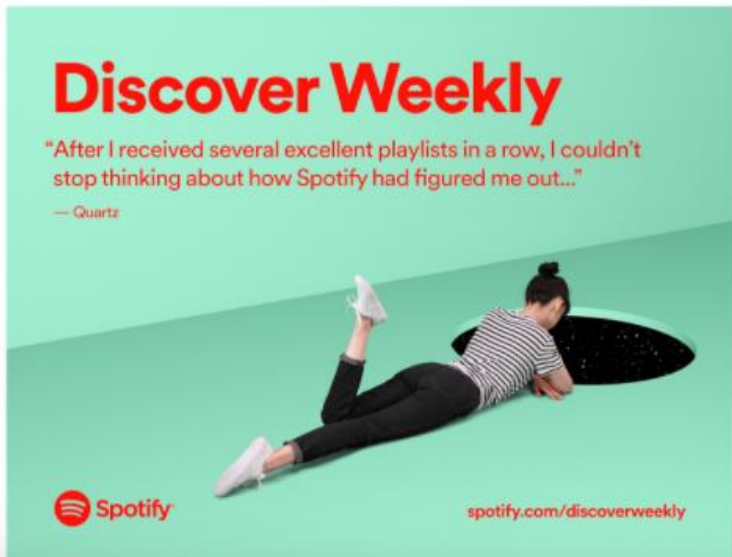
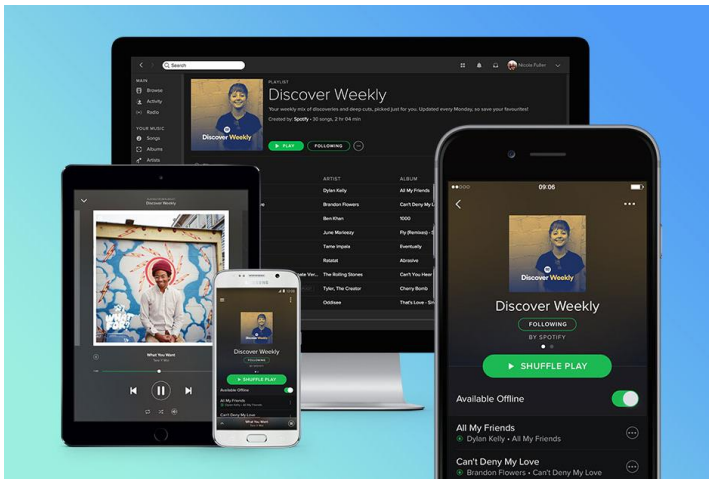
- Stress-testing models and streamlining re-training and performance
- Setting up servers/hosts, liaising with potential users, etc.

Typical modelling pipeline



When learning works, and fails...

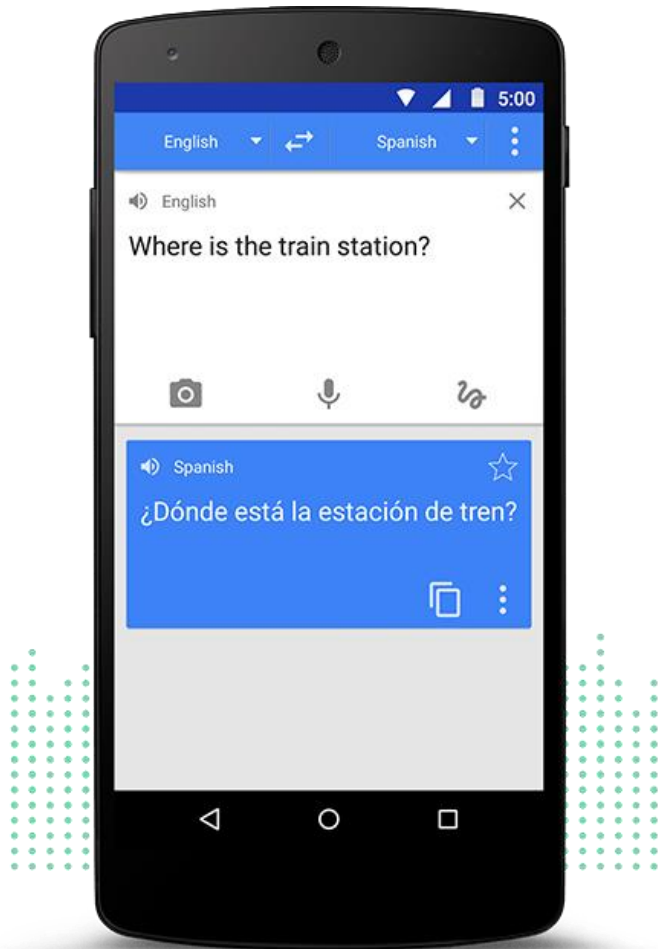
Spotify Discover



Correctional Offender Management and Profiling Alternative Section



Google Translate

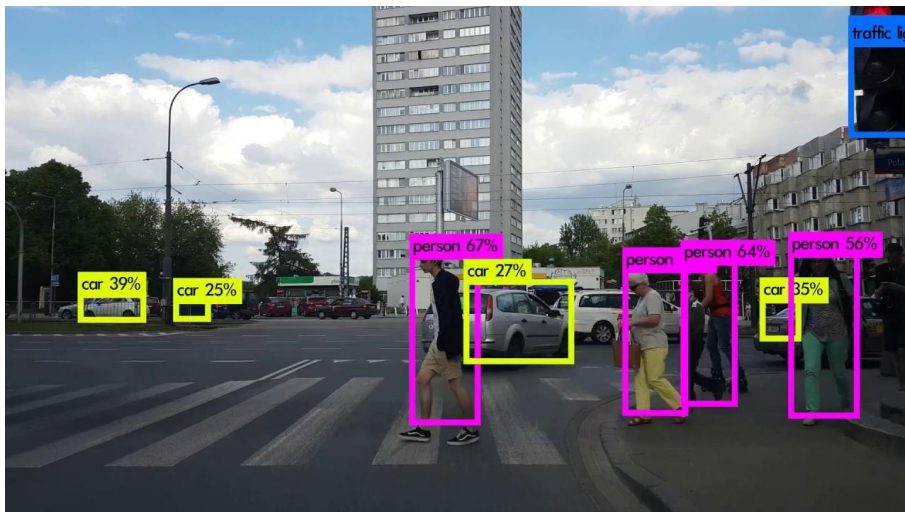
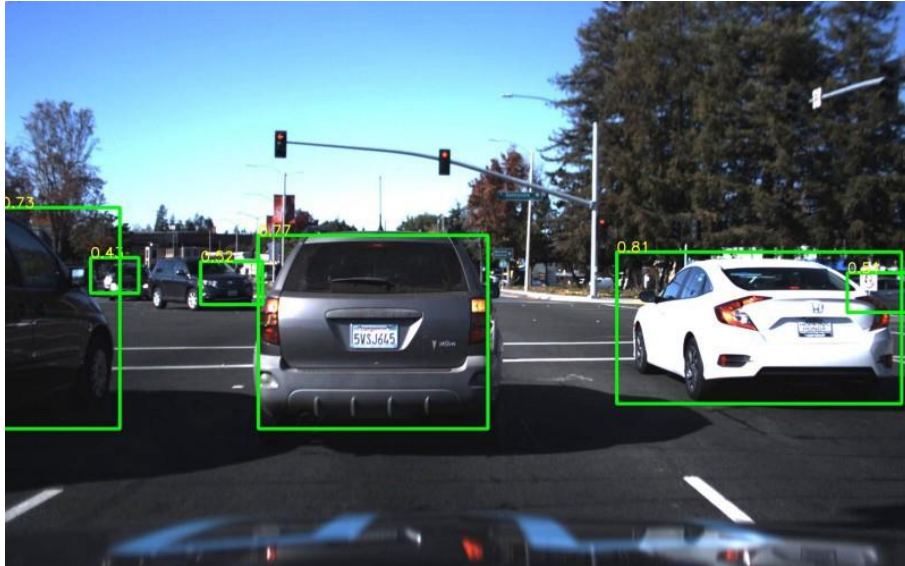


Google Translate



When learning works, and fails...

Self-driving cars



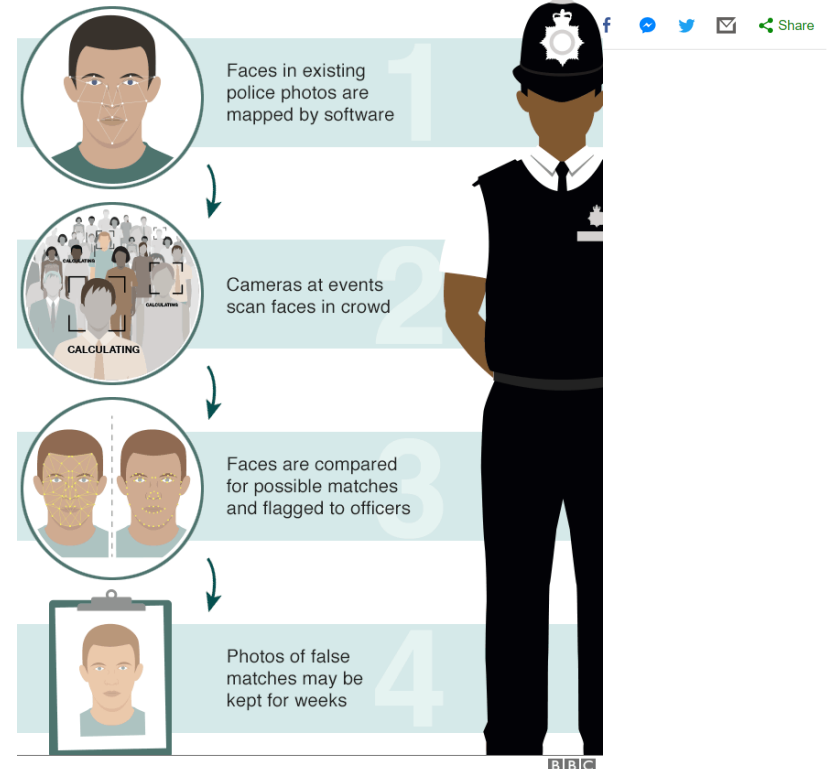
Facial recognition



Technology

Face recognition police tools 'staggeringly inaccurate'

How does live facial recognition work?



- ❖ General Books – Mainly Stats, Probability and Supervised Learning
 - ❖ Trevor, H., Robert, T., & JH, F. (2009). The elements of statistical learning: data mining, inference, and prediction.
(<http://web.stanford.edu/~hastie/ElemStatLearn/>)
 - ❖ Efron, B., & Hastie, T. (2016). Computer age statistical inference (Vol. 5). Cambridge University Press. (<https://web.stanford.edu/~hastie/CASI/>)
- ❖ Practical Books
 - ❖ Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.". (<https://github.com/ageron/handson-ml>)
 - ❖ Vishnu Subramanian (2018). Deep Learning with PyTorch: A Practical Approach to Building Neural Network Models Using PyTorch. " O'Reilly Media, Inc.".

❖ Specific Books

- ❖ Theory: Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT press. (<https://cs.nyu.edu/~mohri/mlbook/>)
- ❖ Kernels: Shawe-Taylor, J., & Cristianini, N. (2004). Kernel methods for pattern analysis. Cambridge university press.
- ❖ Bayesian: Barber, D. (2012). Bayesian reasoning and machine learning. Cambridge University Press. (<http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf>)
- ❖ Reinforcement Learning: Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
(<https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf>)

❖ Tutorials and other content

- ❖ <https://www.youtube.com/watch?v=iOh7QUZGyiU>
- ❖ <https://www.edx.org/course/deep-learning-with-python-and-pytorch>
- ❖ <https://www.coursera.org/learn/python>
- ❖ <http://playground.tensorflow.org/>

- ❖ **Introduction to AI & Machine Learning (Algorithms)**
- ❖ **Key Components of Algorithmic Impact Assessment**
- ❖ **Algorithmic Explainability**
- ❖ **Algorithmic Fairness**
- ❖ **Algorithmic Robustness**

- ❖ What do we mean by Algorithmic Impact (AI) Assessment ?
- ❖ Assessment vs by Design in AI Assessment
- ❖ Quick view of AI Assessment Canvas
- ❖ Further reading

- ❖ **Algorithmic Impact Assessment** focus on **evaluating** an **Automated Decision-making system** mainly from a **Robustness, Fairness and Explainability** point of view
- ❖ The **goals** of AI Assessment are
 - ❖ Set the boundary, usage and shelf-life of a system
 - ❖ Build trust between the stakeholders of a system
 - ❖ Be the entry point to hold the system's creators accountable of the results of its decision-making
- ❖ We should also mention **other areas** of AI Assessment, such as **Transparency, Accountability**, etc.

What do we mean by AI Assessment

In a nutshell

- ❖ **Robustness**: systems should be safe and secure, not vulnerable to tampering or compromising of the data they are trained on.
- ❖ **Fairness**: systems should use training data and models that are free of bias, to avoid unfair treatment of certain groups.
- ❖ **Explainability**: systems should provide decisions or suggestions that can be understood by their users and developers.

To avoid these cases

In the news



Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours
Telegraph.co.uk - 5 hours ago

To chat with Tay, you can tweet or DM her by finding @tayandyou on Twitter, or add her as a ...

Microsoft Releases AI Twitter Bot That Immediately Learns How To Be Racist
Kotaku - 3 hours ago

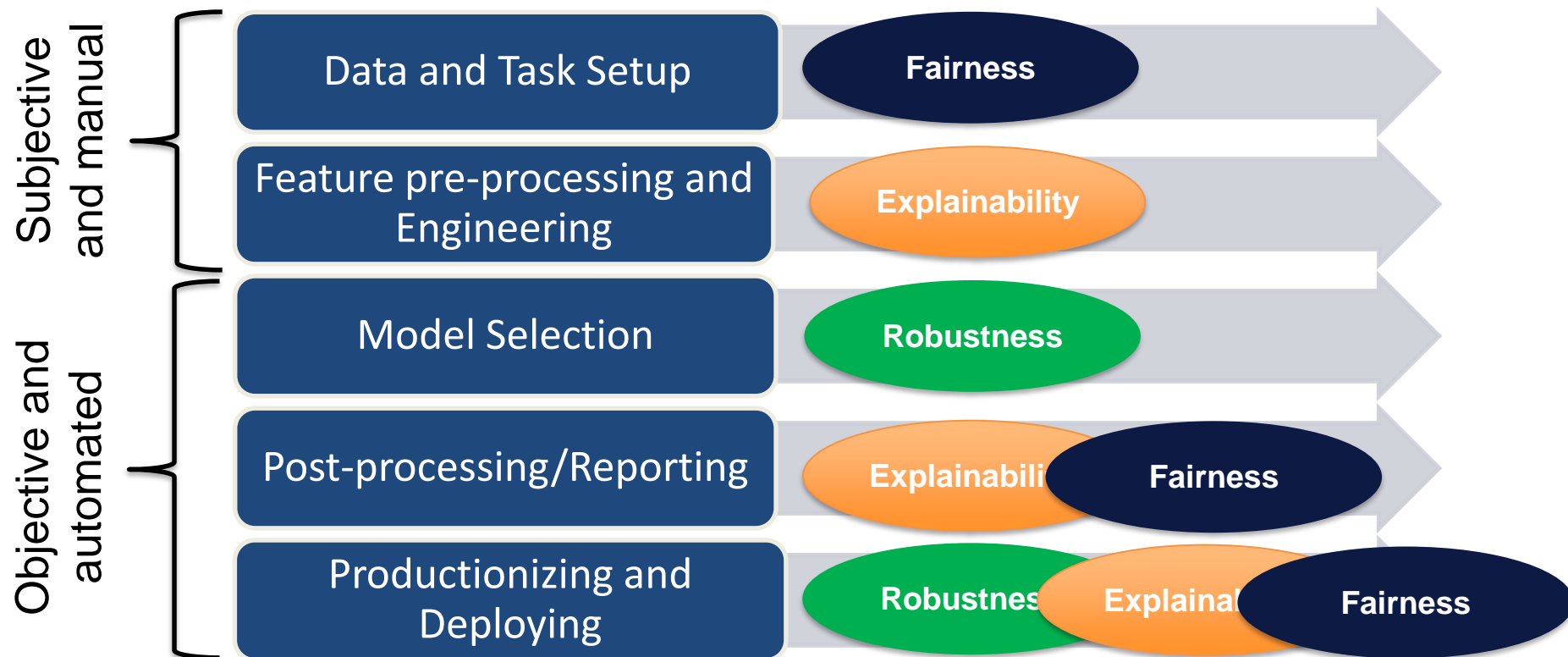
Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.
New York Times - 3 hours ago



JAMES RIVELLI	ROBERT CANNON
Prior Offenses 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking	Prior Offense 1 petty theft
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	MEDIUM RISK 6

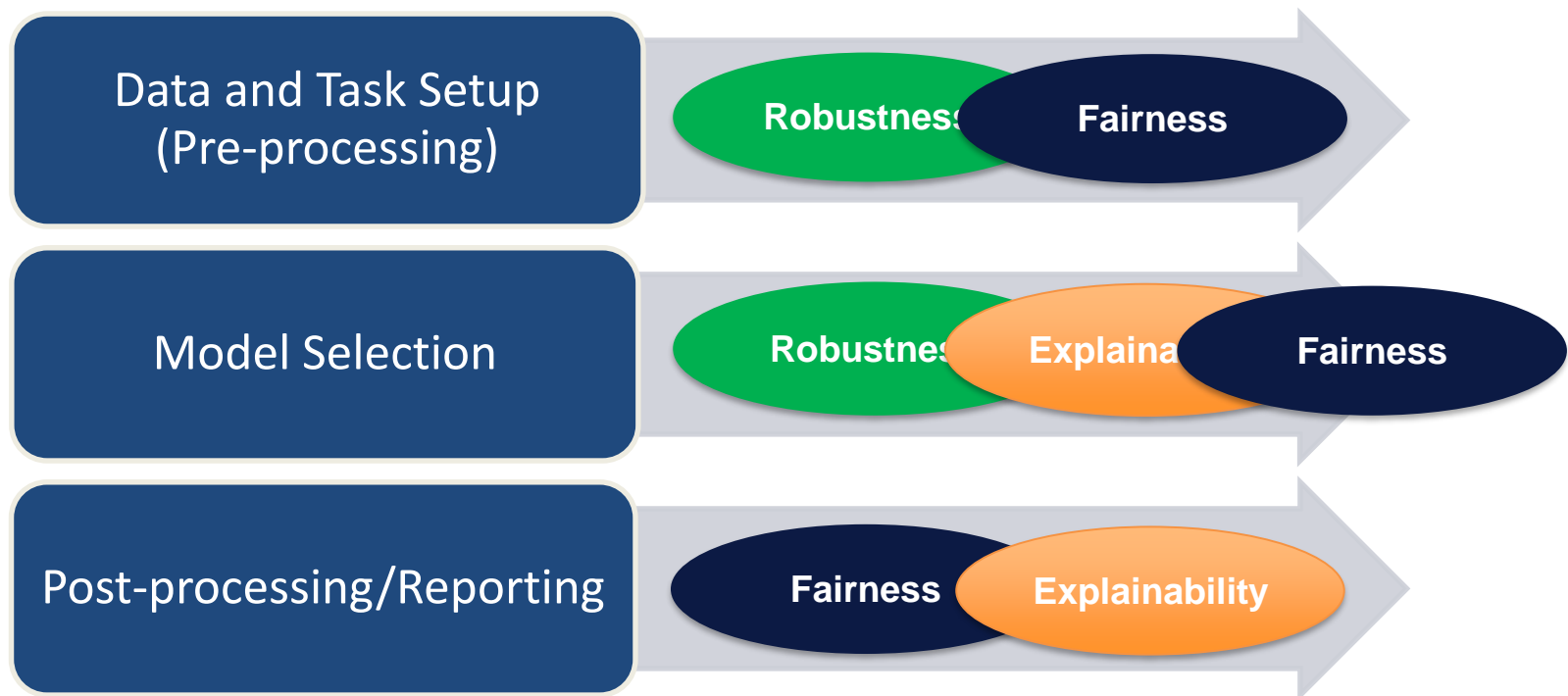
From an Assessment point of view

- ❖ Areas in the modelling pipeline where a AI Assessment Analyst (AI²) should analyse using the different criteria

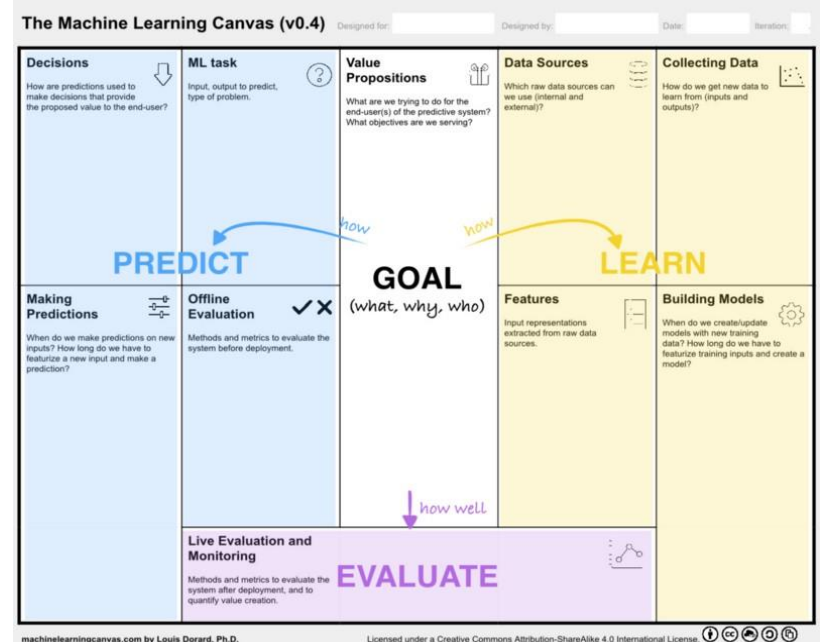


From a by Design point of view

- ❖ It is possible to have systems that by design are able to increase or fulfil the stakeholders demand for Fairness, Robustness and Explainability



- ❖ The Algorithmic Impact Assessment Canvas is a great tool for planning, communication and project tracking
- ❖ However, its focus is on how the **problem** will be solved, and not what **questions** the solution need to address
- ❖ Hence, we need to recreate this Canvas, moving it from
 - ❖ **value-centric**
 - to
 - ❖ **safety-centric**
 - decision-making



An AI Assessment Canvas

The AI Assessment Canvas (v1.0)

Designed for:

Designed by:

Date:

Iteration:

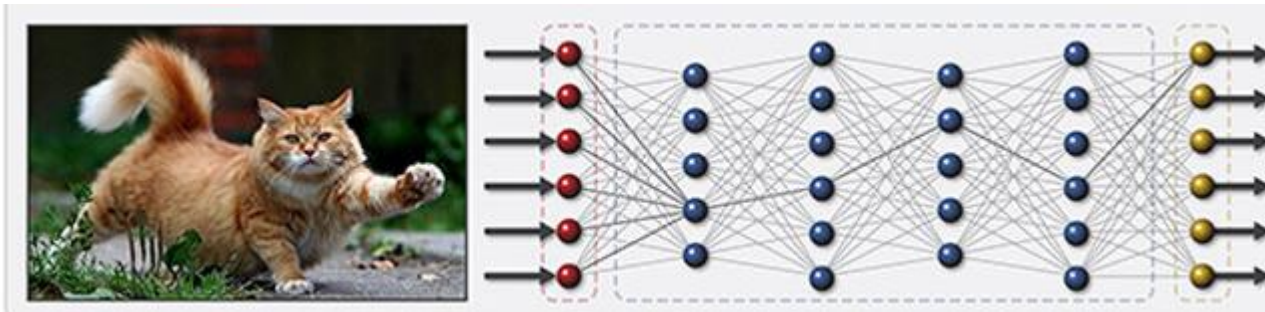


- ❖ **Introduction to AI & Machine Learning (Algorithms)**
- ❖ **Key Components of Algorithmic Impact Assessment**
- ❖ **Algorithmic Explainability**
- ❖ **Algorithmic Fairness**
- ❖ **Algorithmic Robustness**

- ❖ What do we mean by an explainable decision
- ❖ Why and what type of Explainability
- ❖ Legal basis for Explainability
- ❖ Different types of Explainability
- ❖ Technological solutions for Explainability
- ❖ Explainability: an AI Assessment checklist
- ❖ Further reading

What do we mean by an explainable decision

Object recognition



This is a cat.

This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:



Healthcare

predict
breast cancer

NHS

Home About Predict Predict Tool Contact Legal Earlier versions

What is Predict?

Predict is an online tool that helps patients and clinicians see how different treatments for early invasive breast cancer might improve survival rates after surgery.

It is endorsed by the American Joint Committee on Cancer (AJCC).

Start Predict

Did you mean to visit Predict Prostate?



Finance



Sorry, your loan application has been rejected.

If instead you had the following values, your application would have been approved:

- MSinceOldestTradeOpen: **161**
- NumSatisfactoryTrades: **36**
- NetFractionInstallBurden: **38**
- NumRevolvingTradesWBalance: **4**
- NumBank2NatlTradesWHighUtilization: **2**



(b) Counterfactual explanation

- ❖ Explicability is **crucial for building and maintaining users' and designers' trust** in AI-based decisions
 - ❖ Users: contest decisions, learning
 - ❖ Creators: knowledge discovery, debugging systems, uncover unfair decisions
- ❖ Hence, the capabilities and purpose of AI systems should be
 - ❖ **openly communicated**
 - ❖ decisions **explainable** to those **directly and indirectly affected**
 - ❖ **timely and adapted to the expertise** of the stakeholder concerned (e.g. layperson, regulator or researcher)

- ❖ **Credit Scoring in the US** have a well-established right to explanation
 - ❖ The Equal Credit Opportunity Act (1974)
- ❖ Credit agencies and data analysis firms such as FICO comply with this regulation by providing a list of reasons (generally at most 4, per interpretation of regulations)
- ❖ **From an AI standpoint**, there are new regulations that gives the system's user the right (?) to know why a certain automated decision was taken in a certain form
 - ❖ **Right to an Explanation – EU General Data Protection Regulation (2018)**

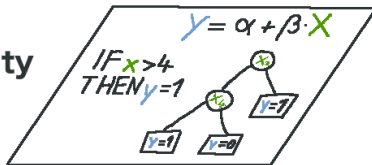
Different types of Explainability

Humans



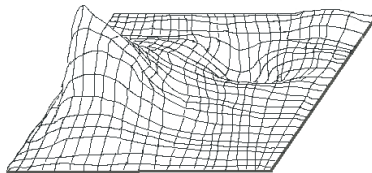
↑ inform

Interpretability Methods



↑ extract

Black Box Model



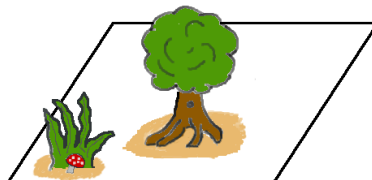
↑ learn

Data

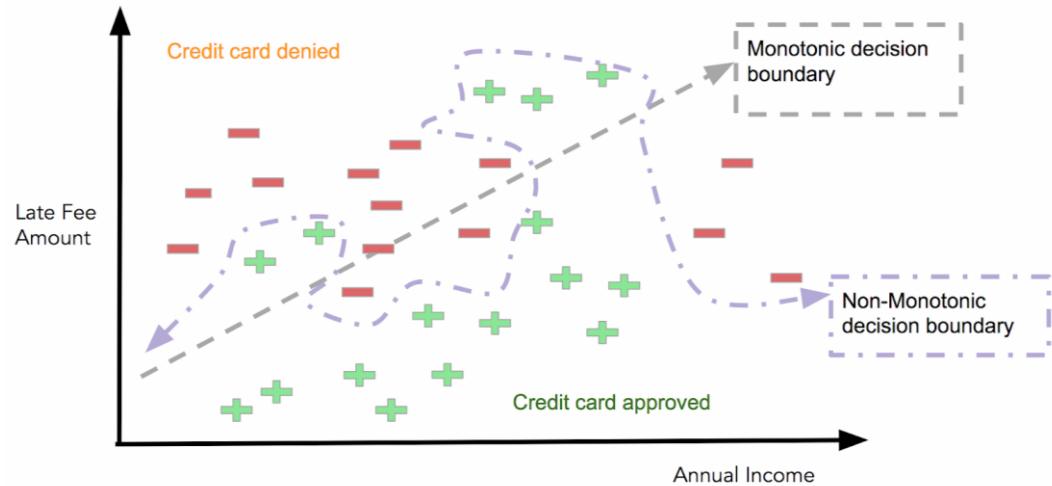
X ₁	X ₂	X ₃	...	X _n
1	5	10	...	100
0	2	0	...	0
1	1	0	...	0

↑ capture

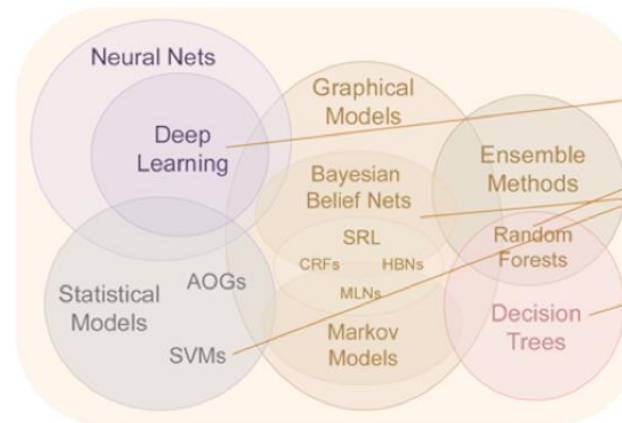
World



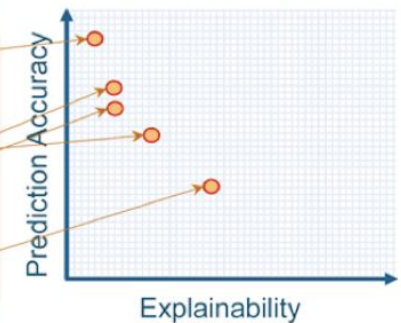
PERFORMANCE VS. INTERPRETABILITY



Learning Techniques (today)



Explainability (notional)



Different types of Explainability



Sorry, your loan application has been rejected.

If instead you had the following values, your application would have been approved:

- MSinceOldestTradeOpen: **161**
- NumSatisfactoryTrades: **36**
- NetFractionInstallBurden: **38**
- NumRevolvingTradesWBalance: **4**
- NumBank2NatlTradesWHighUtilization: **2**

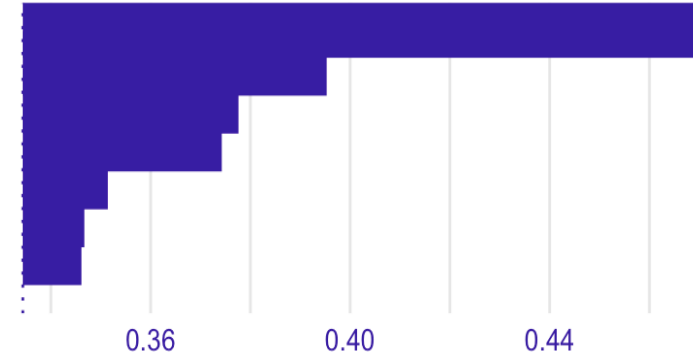


(b) Counterfactual explanation


Model-agnostic

Random Forest

gender
class
age
fare
embarked
sibsp
parch
country



local



global

amazon.com

Recommended for You

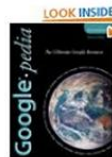
Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



[Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop](#)

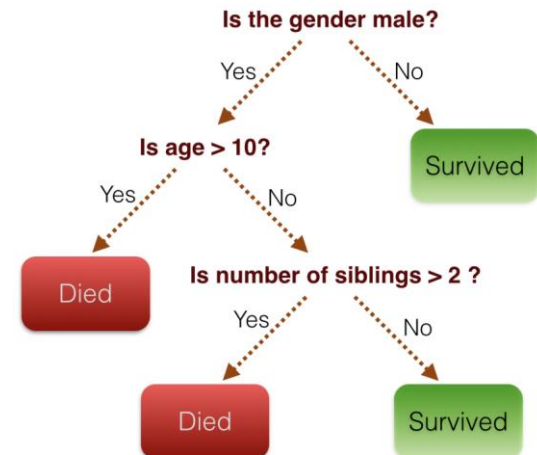


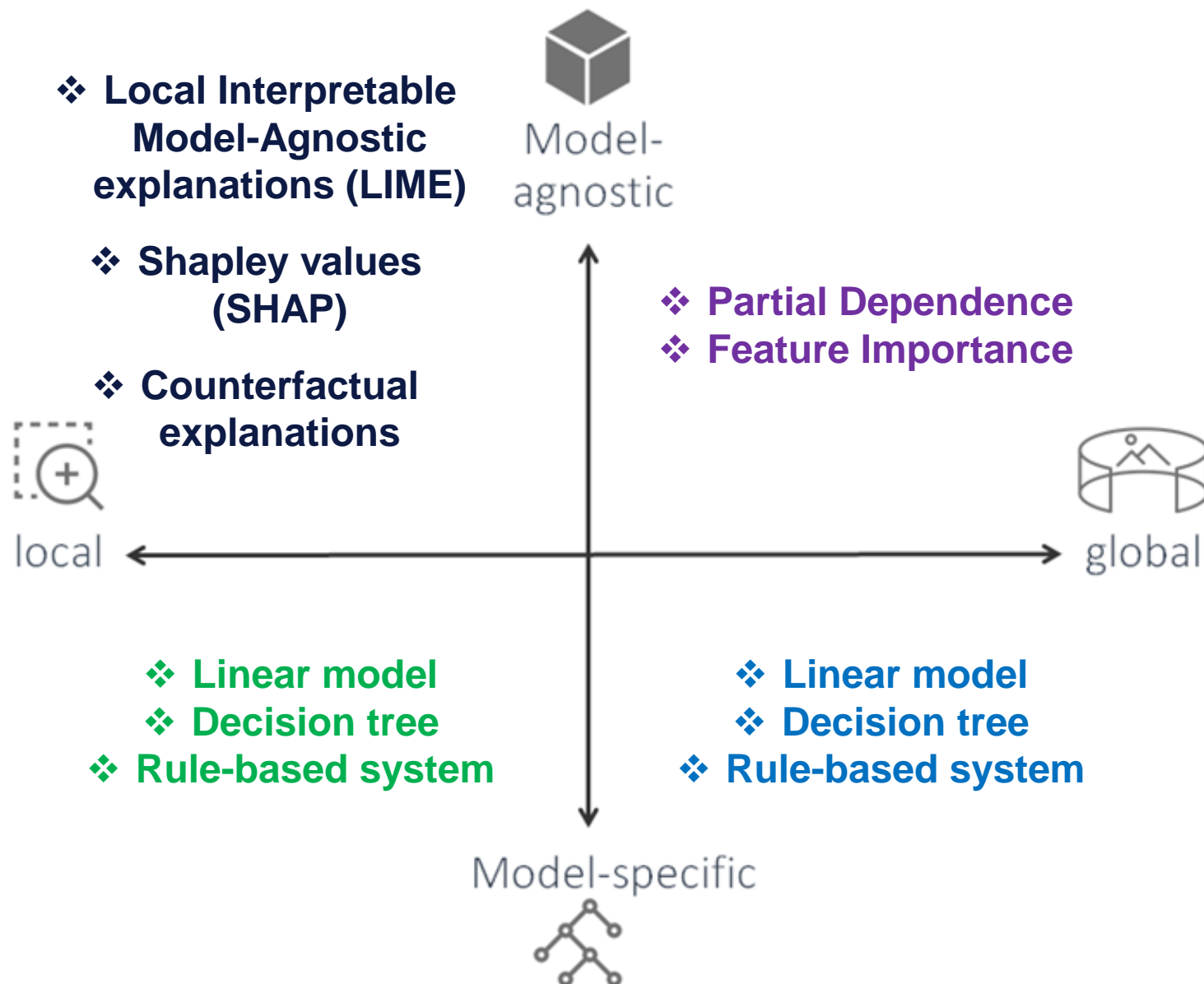
[Google Apps Administrator Guide: A Private-Label Web Workspace](#)



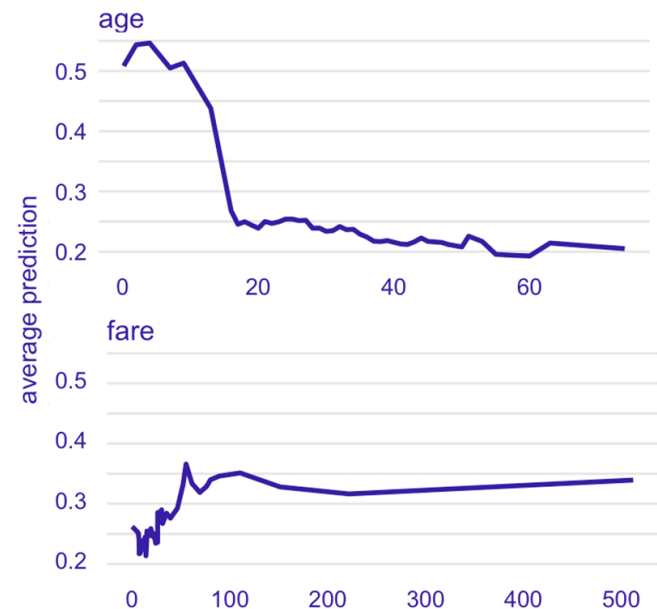
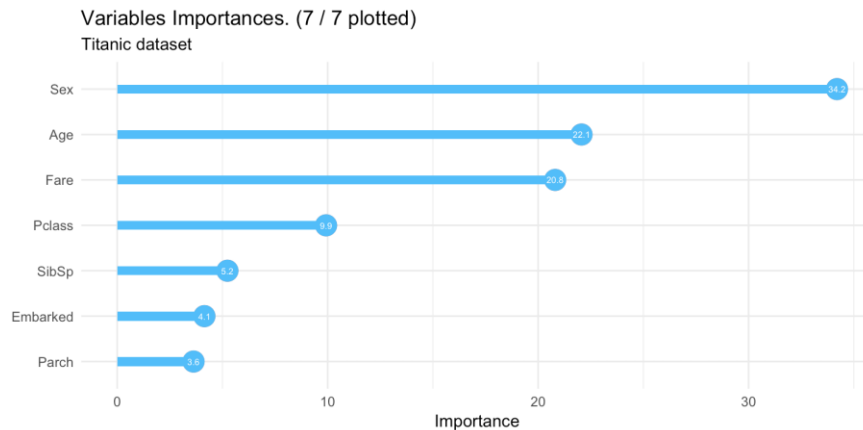
[Googlepedia: The Ultimate Google Resource \(3rd Edition\)](#)

Model-specific





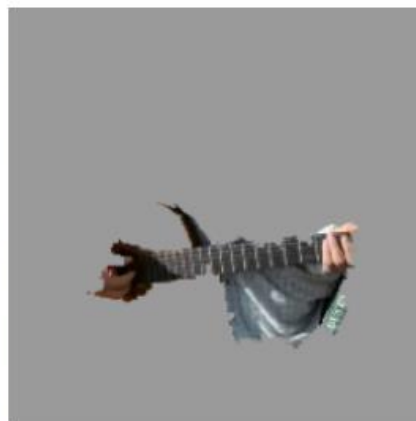
- ❖ Some examples of how it works
 - ❖ Feature and partial dependency plots



- ❖ Local Interpretable Model-Agnostic explanations



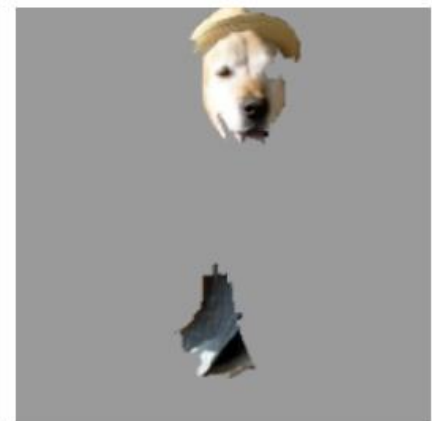
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

Explainability

- ❖ Did you assess to what extent the decisions and hence the outcome made by the AI system can be understood?
- ❖ Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?
- ❖ Did you design the AI system with interpretability in mind from the start?
- ❖ Did you research and try to use the simplest and most interpretable model possible for the application in question?
- ❖ Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model?

❖ Legislation

- ❖ The Equal Credit Opportunity Act: <https://www.justice.gov/crt/equal-credit-opportunity-act-3>
- ❖ GDPR: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/>

❖ Papers and books

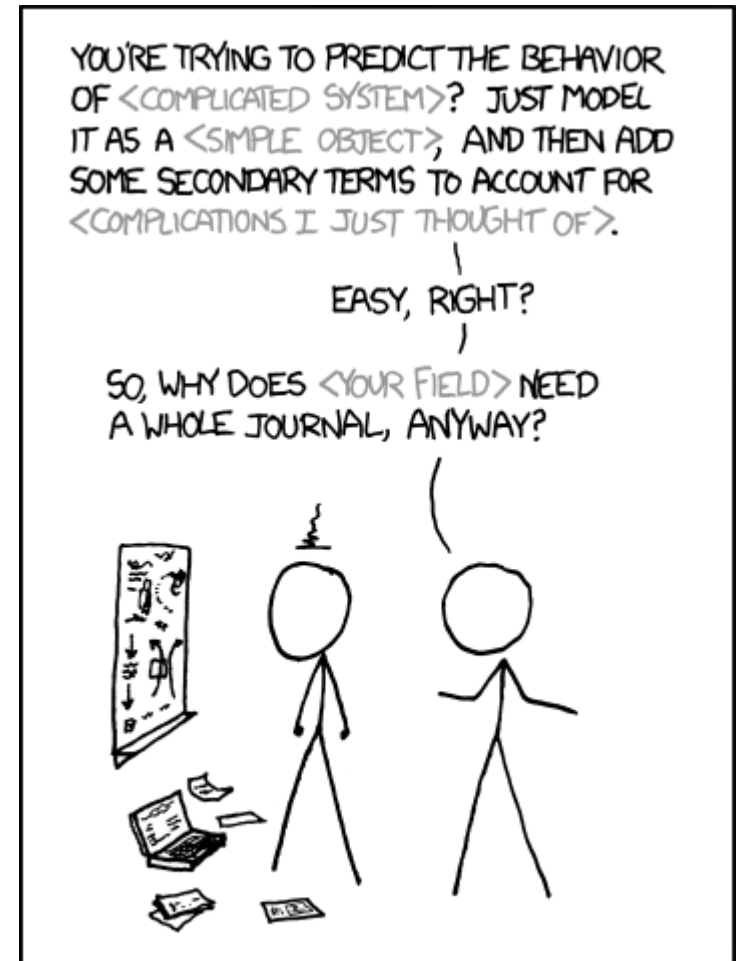
- ❖ Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).
- ❖ Hall, Patrick. "On the Art and Science of Machine Learning Explanations." arXiv preprint arXiv:1810.02909 (2018).
- ❖ Hall, Patrick, and Navdeep Gill. Introduction to Machine Learning Interpretability. O'Reilly Media, Incorporated, 2018.
- ❖ Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harvard Journal of Law & Technology 31, no. 2 (2017): 2018.

❖ Tools

- ❖ <https://pair-code.github.io/what-if-tool/>
- ❖ <https://github.com/marcotcr/lime>
- ❖ <https://github.com/microsoft/interpret>
- ❖ <https://github.com/slundberg/shap>

❖ Other good online resources

- ❖ <https://christophm.github.io/interpretable-ml-book/>
- ❖ <https://distill.pub/2018/building-blocks/>
- ❖ <https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning>



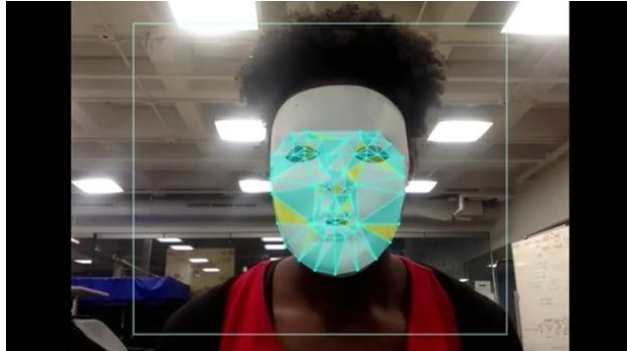
LIBERAL-ARTS MAJORS MAY BE ANNOYING SOMETIMES, BUT THERE'S *NOTHING* MORE OBNOXIOUS THAN A PHYSICIST FIRST ENCOUNTERING A NEW SUBJECT.

- ❖ **Introduction to AI & Machine Learning (Algorithms)**
- ❖ **Key Components of Algorithmic Impact Assessment**
- ❖ **Algorithmic Explainability**
- ❖ **Algorithmic Fairness**
- ❖ **Algorithmic Robustness**

- ❖ When a decision not fair
- ❖ Guiding principles of Fairness
- ❖ Legal basis for Fairness
- ❖ Mathematical description of Fairness
- ❖ Technical solutions for Fairness
- ❖ Fairness: an AI Assessment checklist
- ❖ Further reading

When a decision is not fair

Facial recognition



Recidivism Model



JAMES RIVELLI	ROBERT CANNON
Prior Offenses 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking	Prior Offense 1 petty theft
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	MEDIUM RISK 6

Recruitment

Facebook's algorithm perpetuates employment discrimination

The fraction of men in the ad's audience

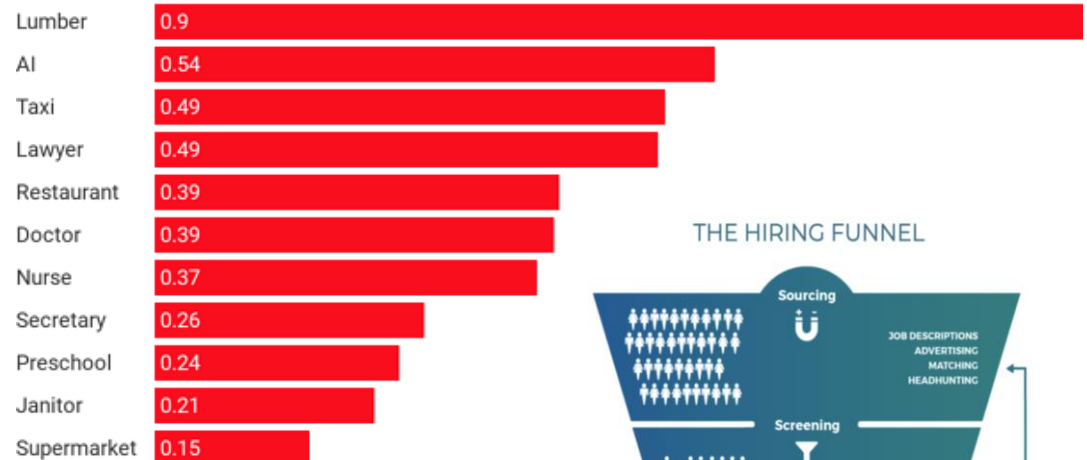
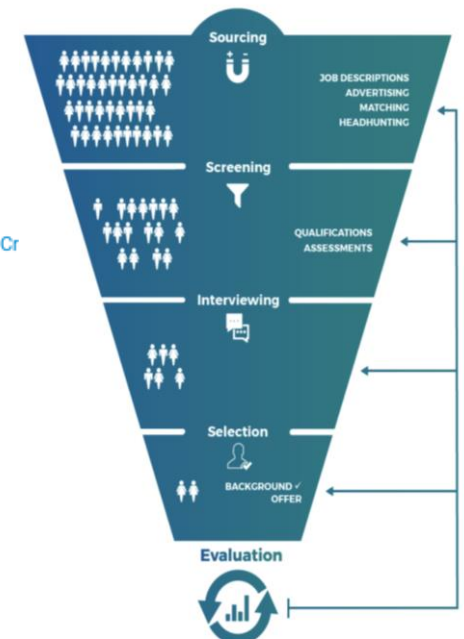


Chart: MIT Technology Review • Source: [Ali & Sapiezynski et al.](#) • Cr



THE HIRING FUNNEL



- ❖ Fairness as an ideal have been present in different **manifestos and charters** during humankind history, gradually amplifying its outreach across the population
 - ❖ *Magna Carta* (1215)
 - ❖ France's *Declaration of the Rights of Man and of the Citizens* (1789)
 - ❖ UN *Universal Declaration of Human Rights* (1948)

- ❖ **In AI, fairness** has both a substantive and a procedural connotation
 - ❖ **Substantive**: ensuring just distribution of both benefits and costs, and ensuring that individuals and groups are free from discrimination
 - ❖ **Procedural**: entails the ability to contest and seek effective redress against decisions made by systems and the humans operating them

- ❖ Most of the **legal basis was developed after multiple public demonstrations, civil rights movements**, etc. and are in many situations set or uphold at Constitutional level
- ❖ **UK**: Equal Pay Act (1970), Sex Discrimination Act (1975), Race Relations Act (1976), Disability Discrimination Act (1995), Equality Act (2010)
- ❖ **US**: Civil Rights Act (1957 and 1964), Americans with Disability Act (1990)
- ❖ **Others**: France, German, Brazil, etc. Constitutions
- ❖ **From an AI standpoint**, there are emerging principles that should be followed in order to develop fair algorithmic decision-making
 - ❖ IEEE Ethically Aligned Design (<https://ethicsinaction.ieee.org/>)
 - ❖ European Commission Ethics Guidelines for Trustworthy AI (<https://ec.europa.eu/futurium/en/ai-alliance-consultation>)

- ❖ There are multiple sources of bias that explain how an automated decision-making process becomes unfair
 - ❖ **Tainted examples:** Any ML system keeps the bias existing in the old data caused by human bias (e.g. recruitment).
 - ❖ **Skewed sample:** future observations confirm predictions, which create a perverse feedback loop (e.g. police record).
 - ❖ **Limited features:** features may be less informative or reliably collected for minority group(s).
 - ❖ **Sample size disparity:** training data coming from the minority group is much less than those coming from the majority group.
 - ❖ **Proxies:** even if protected attributes is not used for training a system, there can always be other proxies of the protected attribute (e.g. neighbourhood).

- ❖ We first need to differentiate between Individual and Group level fairness
 - ❖ **Individual**: seeks for similar individuals to be treated similarly
 - ❖ **Group**: split a population into groups defined by protected attributes and seeks for some measure to be equal across groups
- ❖ Which mathematically translate to (though there are multiple other definitions...)

Individual (Consistency)	Group (Statistical parity)
$1 - \frac{1}{n \times N(x_i) } \sum_i^n f(x_i) - \sum_{j \in N(x_i)} f(x_j) $	$\mathbb{P}(f(x) v = \textit{unprivileged}) - \mathbb{P}(f(x) v = \textit{privileged})$

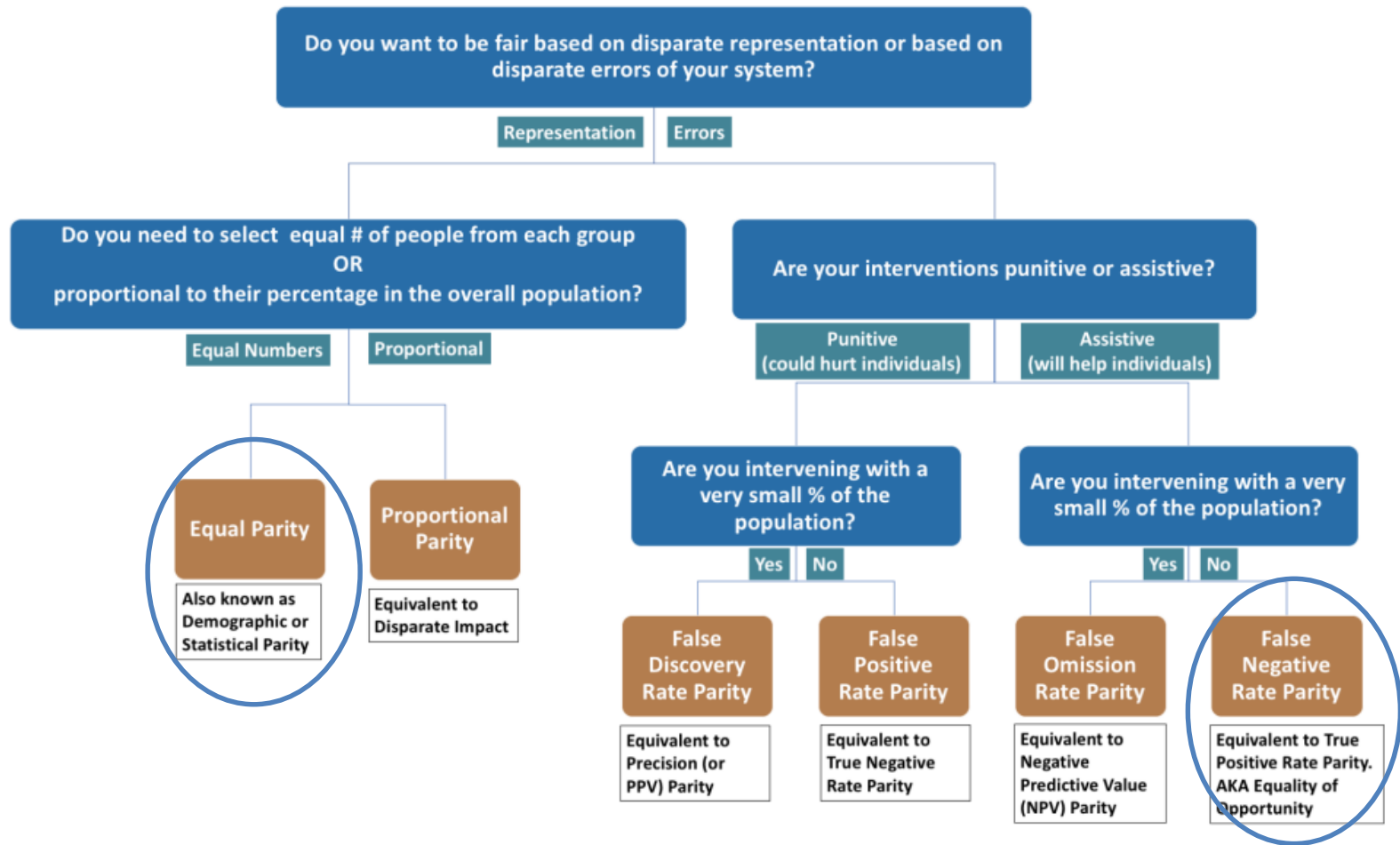
* x : input variable; $f(x)$: algorithm prediction/decision; v : protected attribute, divided into unprivileged vs privileged categories; \mathbb{P} : probability measure; $N(x_i)$: set of neighbours of x_i

- ❖ Also, within Group fairness, it is possible to distinguish between the aim of Equality of Opportunity and Outcome. For example, using SAT score as a feature for predicting success in college:
 - ❖ the Opportunity worldview says that the score correlates well with future success and that there is a way to use the score to correctly compare the abilities of applicants
 - ❖ the Outcome worldview says that the SAT score may contain structural biases so its distribution being different across groups should not be mistaken for a difference in distribution in ability.
- ❖ Opportunity (Avg odds difference):
$$\frac{1}{2} \left([FPR_{v=unpriv} - FPR_{v=priv}] + [TPR_{v=unpriv} - TPR_{v=priv}] \right)$$
- ❖ Outcome (Statistical parity): $\mathbb{P}(f(x)|v = unpriv) - \mathbb{P}(f(x)|v = priv)$

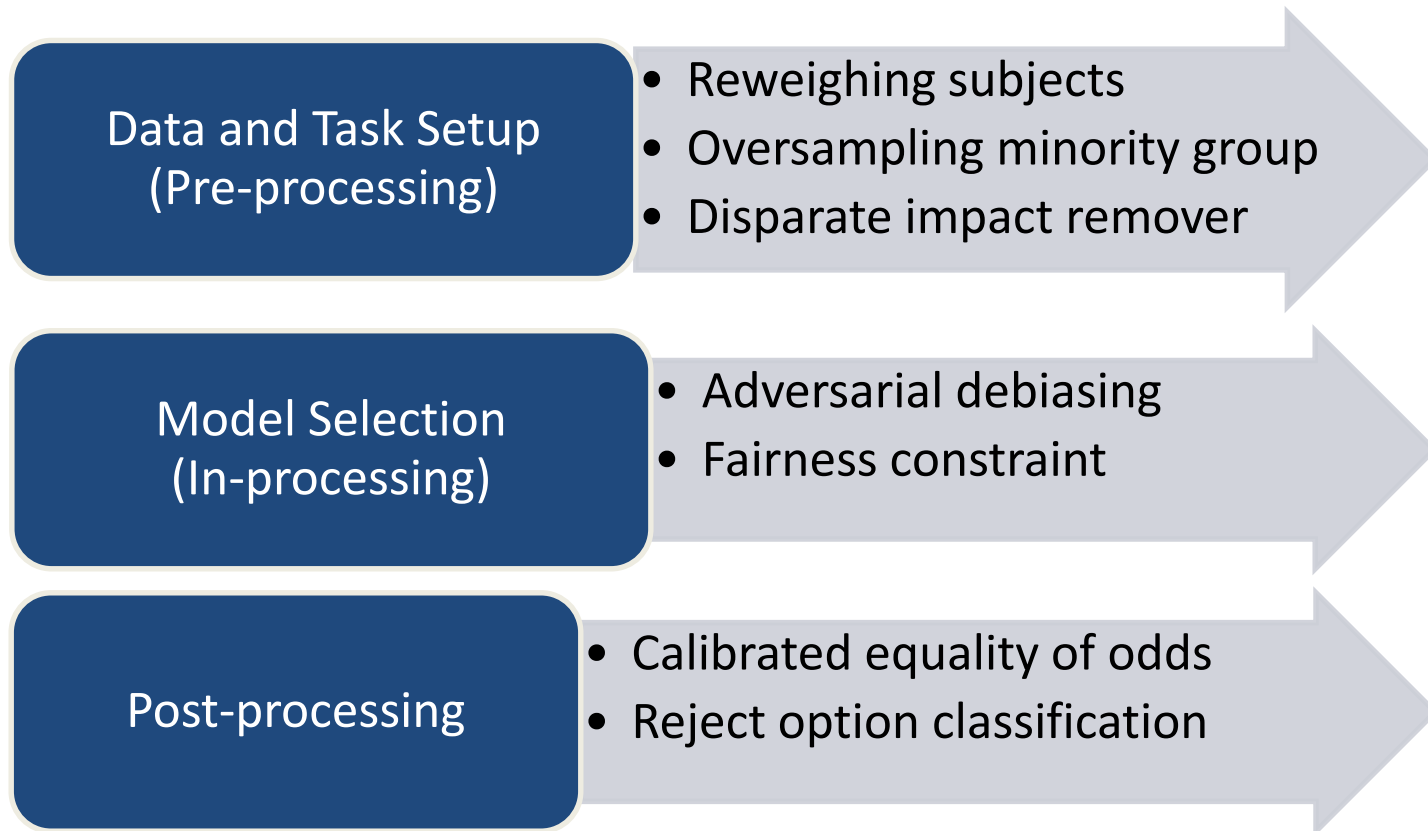
Mathematical definition of Fairness

- ❖ As we said, there are multiple ways to assess fairness, which can be broadly grouped by the outcomes the system's designer is aiming for:

FAIRNESS TREE



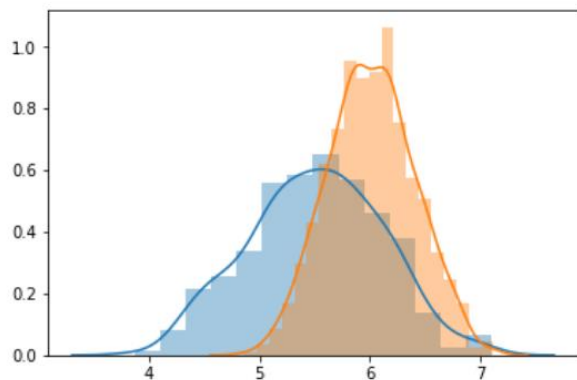
- ❖ Regardless of the measure used, unFairness can be mitigated at different points in a modelling pipeline: pre-processing, in-processing, and post-processing



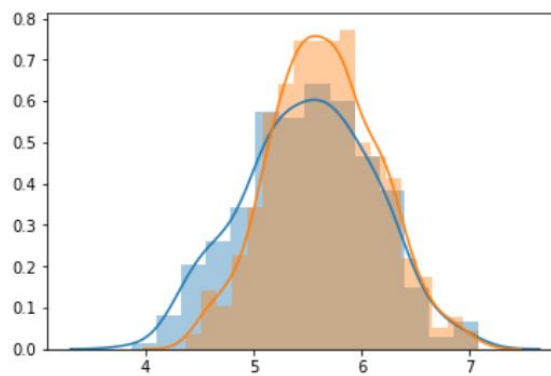
❖ Some examples of how it works

❖ Disparate Impact Remover

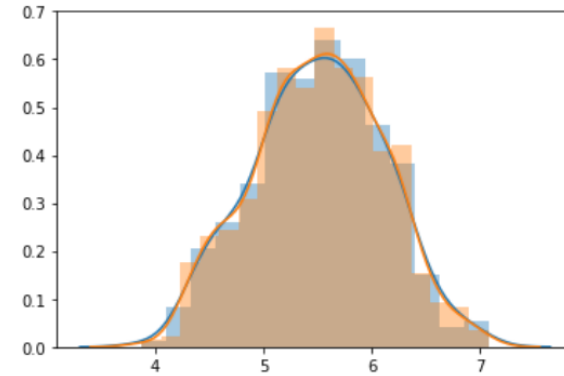
Observed Values



Repair value = 0.8 (4/5 rule)

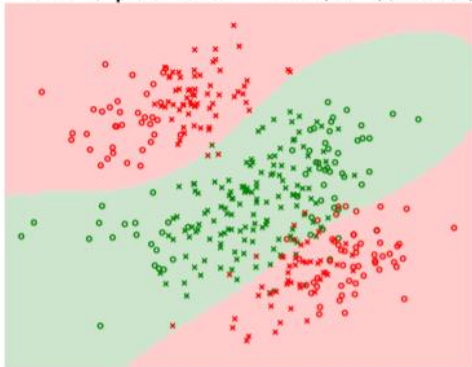


Repair value = 1.0



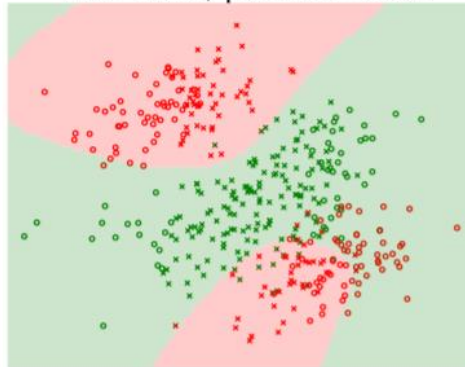
❖ RBF Kernel SVM with fairness constraints

Acc=0.94; p% rule=42%($\pi/4$), 12%($\pi/8$)



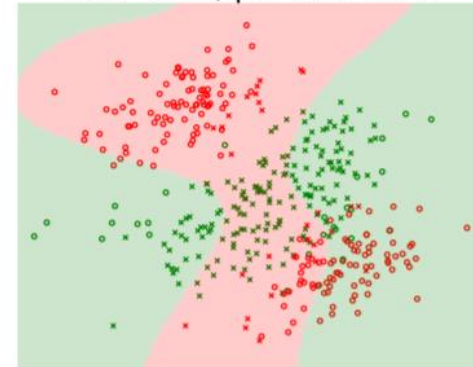
(a) Unconstrained

Acc=0.83; p% rule=95%



(b) $\phi = \pi/4$

Acc=0.60; p% rule=97%



(c) $\phi = \pi/8$

Unfair bias avoidance

- ❖ Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
- ❖ Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system?
- ❖ Did you assess whether there is any possible decision variability that can occur under the same conditions?
- ❖ Did you ensure an adequate working definition of “fairness” that you apply in designing AI systems?

❖ Papers

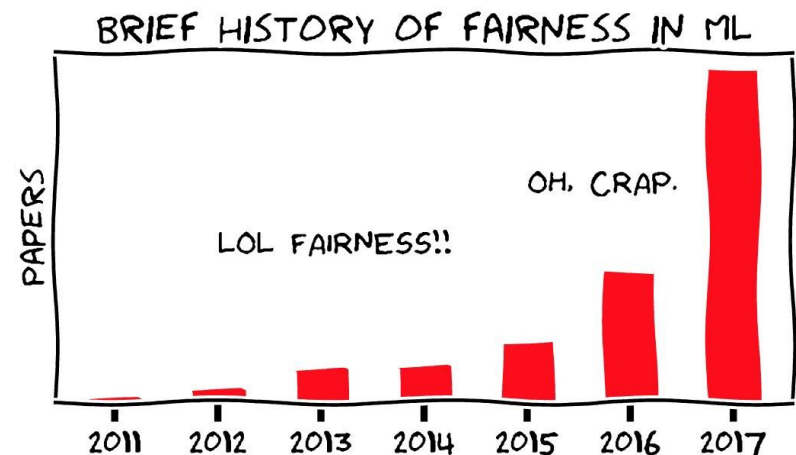
- ❖ R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," International Conference on Machine Learning, 2013.
- ❖ B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2018.
- ❖ G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration," Conference on Neural Information Processing Systems, 2017
- ❖ Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. "Fairness constraints: Mechanisms for fair classification." arXiv preprint arXiv:1507.05259 (2015).
- ❖ F. Kamiran, A. Karim, and X. Zhang, "Decision Theory for Discrimination-Aware Classification," IEEE International Conference on Data Mining, 2012.
- ❖ Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and removing disparate impact." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259-268. ACM, 2015.
- ❖ Donini, Michele, Luca Oneto, Shai Ben-David, John S. Shawe-Taylor, and Massimiliano Pontil. "Empirical risk minimization under fairness constraints." In *Advances in Neural Information Processing Systems*, pp. 2791-2801. 2018.
- ❖ Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness." In *Advances in Neural Information Processing Systems*, pp. 4066-4076. 2017.

❖ Online resources

- ❖ Tutorial on fairness: <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>
- ❖ Review on fairness: <https://arxiv.org/pdf/1810.08810.pdf>
- ❖ FAT Conferences: <https://fatconference.org/>
- ❖ NIPS Tutorial: <https://nips.cc/Conferences/2017/Schedule?showEvent=8734>

❖ Online tools

- ❖ <http://aif360.mybluemix.net/resources#glossary>
- ❖ <https://dsapp.uchicago.edu/projects/aequitas/>



- ❖ **Introduction to AI & Machine Learning (Algorithms)**
- ❖ **Key Components of Algorithmic Impact Assessment**
- ❖ **Algorithmic Explainability**
- ❖ **Algorithmic Fairness**
- ❖ **Algorithmic Robustness**

- ❖ When a system is not robust
- ❖ Legal basis for robust algorithms
- ❖ Guiding principles of Robustness
- ❖ Mathematical definition of Robustness
- ❖ Technological solutions for Robustness
- ❖ Robustness: an AI Assessment checklist
- ❖ Further reading

Chatbots

In the news



Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours
Telegraph.co.uk - 5 hours ago
To chat with Tay, you can tweet or DM her by finding @tayandyou on Twitter, or add her as a ...

Microsoft Releases AI Twitter Bot That Immediately Learns How To Be Racist
Kotaku - 3 hours ago

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.
New York Times - 3 hours ago

Automated Diagnosis

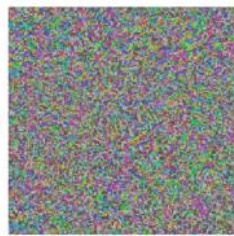
Original image



Dermoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network.



Adversarial noise



Perturbation computed by a common adversarial attack technique. See (7) for details.

Adversarial example



Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.



Facial recognition

BBC

Your account



News

Sport

Weather

iPlayer

Sounds

NEWS

Home

UK

World

Business

Politics

Tech

Science

Health

Family & Education

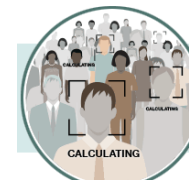
Technology

Face recognition police tools 'staggeringly inaccurate'

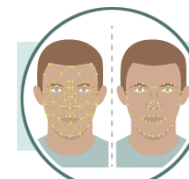
How does live facial recognition work?



1
Faces in existing police photos are mapped by software



2
Cameras at events scan faces in crowd



3
Faces are compared for possible matches and flagged to officers



4
Photos of false matches may be kept for weeks

BBC

- ❖ Robustness as a technical concept is closely linked to **the principle of prevention of harm**
- ❖ AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. Preventing harm also entails consideration of the natural environment and all living beings.
- ❖ The idea behind this principle is present in several moments and with different phrasings
 - ❖ France's *Declaration of the Rights of Man and of the Citizens* (1789)
 - ❖ John Stuart Mill's *On Liberty* (1859)
 - ❖ UN *Universal Declaration of Human Rights* (1948)

- ❖ Most of the legal basis is established by an interaction between Regulatory Agencies, Professional Associations and Industry Trade Groups, where standards, rules and code of conducts are created
 - ❖ **Financial algorithms:** FCA, FSB, BBA
 - ❖ **Power systems:** Fed Ener Reg Com, IEEE
 - ❖ **Electrical appliances:** NIST, Nat Fire Prote Assoc, State Legislation
 - ❖ **Automotive sector:** Nat Trans Saft Board, Soc Auto Engineers
- ❖ Apart from sector/application-specific laws, there are general guidelines that have been proposed by governments and institutions
 - ❖ IEEE Ethically Aligned Design (<https://ethicsinaction.ieee.org/>)
 - ❖ European Commission Ethics Guidelines for Trustworthy AI (<https://ec.europa.eu/futurium/en/ai-alliance-consultation>)

- ❖ We can rate an algorithm's robustness using four key criteria
 - ❖ **Resilience to attack and security:** AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, such as data poisoning, model leakage or the infrastructure, both software and hardware.
 - ❖ **Fallback plan and general safety:** AI systems should have safeguards that enable a fallback plan in case of problems. Also, the level of safety measures required depends on the magnitude of the risk posed by an AI system.
 - ❖ **Accuracy:** pertains to an AI system's ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models.
 - ❖ **Reliability and Reproducibility:** a reliable AI system is one that works properly with a range of inputs and in a range of situations, whilst reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions.

- ❖ We can map each criteria in a math/technical concept as well as a provide a layman interpretation for it

Criteria	Math/Tech Concept	Interpretation
Accuracy	Expected generalization performance	In general, the algorithm works? (e.g. in 7 out of 10 cases, the algorithm makes the right decision)
Resilience to attack and security; Fallback plan and general safety	Adversarial robustness	How the algorithm performed in the worst-case scenario? (e.g. how the algorithm would react during the 2008 Financial Crisis?)
Fallback plan and general safety; Reliability and reproducibility	Formal verification	The algorithm attends the problem specifications and constraints? (e.g. respect physical laws)
Reliability and reproducibility	Continuous integration	Is the algorithm auditable? (e.g. reliably reproduce its decisions)

- ❖ For some of the concepts we can write down a math formula:

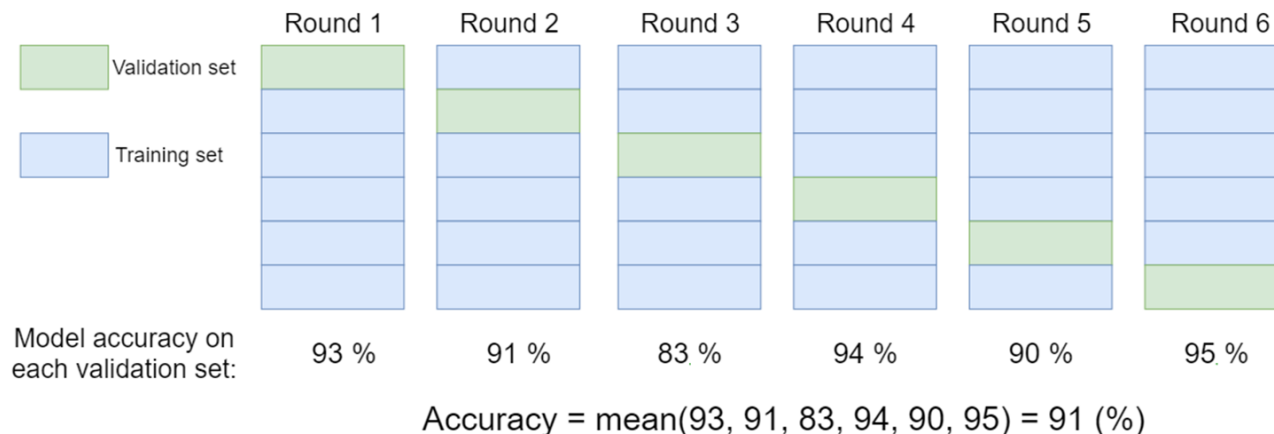
Math/Tech Concept	Mathematical Description*
Expected generalization performance	Expected loss: $\mathbb{E}_{(x,y) \sim p}[L(y; f(x))] \approx \text{mean}_{(x,y) \in D^{val}}[L(y; f(x))]$
Adversarial robustness	Adversarial risk: $\mathbb{E}_{(x,y) \sim p} \left[\max_{\delta \in \Delta(x)} L(y; f(x + \delta)) \right] \approx$ $\text{mean}_{(x,y) \in D^{val}} \left[\max_{\delta \in \Delta(x)} L(y; f(x + \delta)) \right]$
Formal verification	Verification bound: $\mathbb{P}(F(x; f(x)) \leq 0) \approx \frac{\#(F(x^{nom}; f(x)) \leq 0)}{ S_{in}(x^{nom}, \delta) }$

* L : loss function; \mathbb{E} : expectation operator; y : output variable; x : input variable; $f(x)$: algorithm prediction/decision; p : sampling distribution of (x, y) ; D^{val} : holdout set of (x, y) ; $\Delta(x)$: set of feasible perturbations (δ) of x ; F : specification mapping x and $f(x)$ in a real number, if $F(x; f(x)) \leq 0$ then, we say it is satisfied; $S_{in}(x^{nom}, \delta)$: the set of all input x that are at most δ distant from x^{nom} ($S_{in}(x^{nom}, \delta) = \{x: ||x - x^{nom}||_{\infty} \leq \delta\}$); \mathbb{P} : probability measure

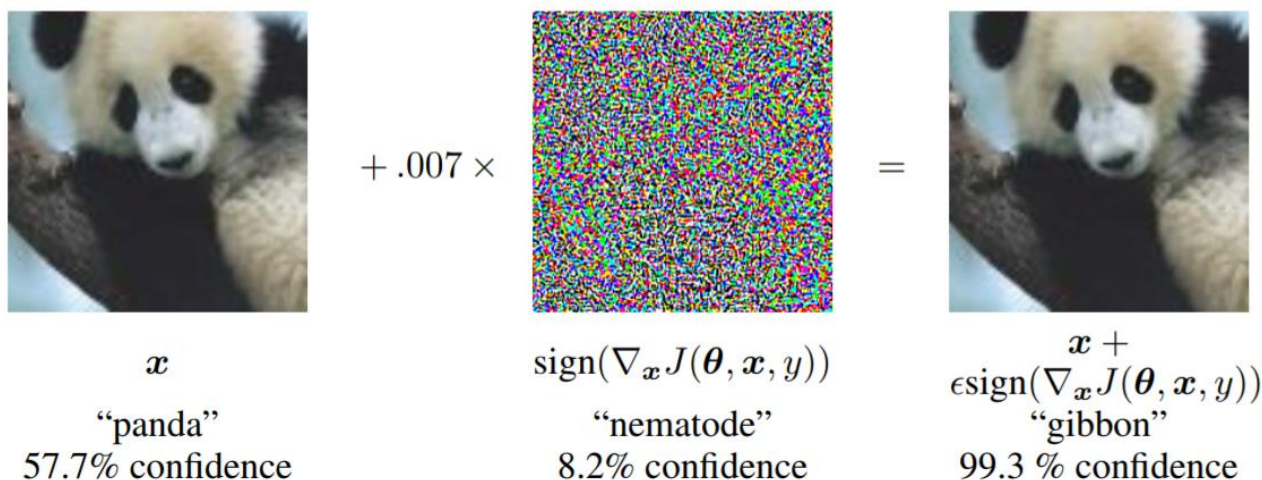
- ❖ Similarly, for each different criteria we have an algorithm-agnostic solution to assess its robustness level

Criteria	Technical Solution
Expected generalization performance	<ul style="list-style-type: none">- Cross-validation: k-fold cv, bootstrap, etc.- Covariance-penalty: Mallow's C_p, Stein Unbiased Risk Estimator, bootstrap approximation, etc.
Adversarial robustness	<ul style="list-style-type: none">- Evasion attacks: fast gradient sign method, DeepFool, etc.- Defence: label smoothing, variance minimization, etc.
Formal verification	<ul style="list-style-type: none">- Complete: Satisfiability Modulo Theory, Mixed Prog, etc.- Incomplete: Propagating bounds, Convex Opt, etc.
Reliability and reproducibility	<ul style="list-style-type: none">- Code versioning: Git (Github), Mercurial (BitBucket), etc.- Reproducible analysis: Binder, Docker, etc.- Automated testing: Travis CI, Scrutinizer CI, etc.

- ❖ Some examples of how it works
 - ❖ Generalization performance: 6-fold cross-validation



- ❖ Adversarial robustness: Fast gradient sign method



Resilience to attack and security

- Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?
- Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?
- Did you verify how your system behaves in unexpected situations and environments?

Fall-back plan and general safety

- Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)?
- Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?

Accuracy

- Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?
- Did you verify what harm would be caused if the AI system makes inaccurate predictions?
- Did you put in place ways to measure whether your system is making an unacceptable amount of inaccurate predictions?

Reliability and reproducibility

- Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications?
- Did you put in place verification methods to measure and ensure different aspects of the system's reliability and reproducibility?
- Did you clearly document and operationalise these processes for the testing and verification of the reliability of AI systems?

❖ Papers

- ❖ Bryson, Joanna, and Alan Winfield. "Standardizing ethical design for artificial intelligence and autonomous systems." *Computer* 50, no. 5 (2017): 116-119.
- ❖ Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete problems in AI safety." *arXiv preprint arXiv:1606.06565* (2016).
- ❖ Leike, Jan, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. "AI safety gridworlds." *arXiv preprint arXiv:1711.09883* (2017).
- ❖ Arlot, Sylvain, and Alain Celisse. "A survey of cross-validation procedures for model selection." *Statistics surveys* 4 (2010): 40-79.
- ❖ Qin, Chongli, Brendan O'Donoghue, Rudy Bunel, Robert Stanforth, Sven Gowal, Jonathan Uesato, Grzegorz Swirszcz, and Pushmeet Kohli. "Verification of non-linear specifications for neural networks." *arXiv preprint arXiv:1902.09592* (2019).
- ❖ Carlini, Nicholas, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. "On evaluating adversarial robustness." *arXiv preprint arXiv:1902.06705*(2019).

❖ Online resources

❖ DeepMind:

<https://deepmind.com/blog/robust-and-verified-ai/>

❖ Technical robustness and safety checklist:

https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477

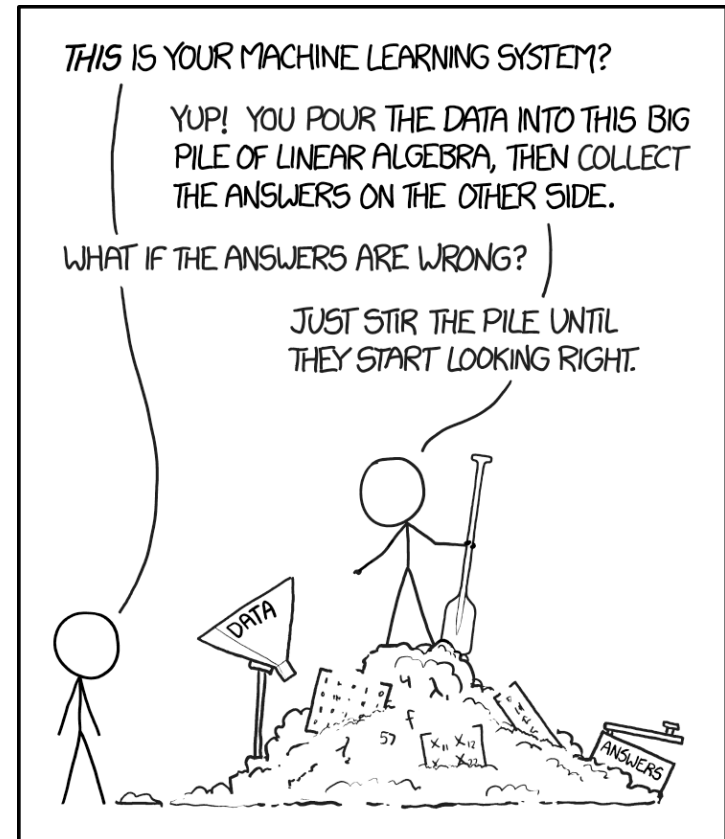
❖ Turing reproducibility:

<https://www.turing.ac.uk/research/research-projects/turing-way-handbook-reproducible-data-science>

❖ Online tools

❖ <https://adversarial-ml-tutorial.org/introduction/>

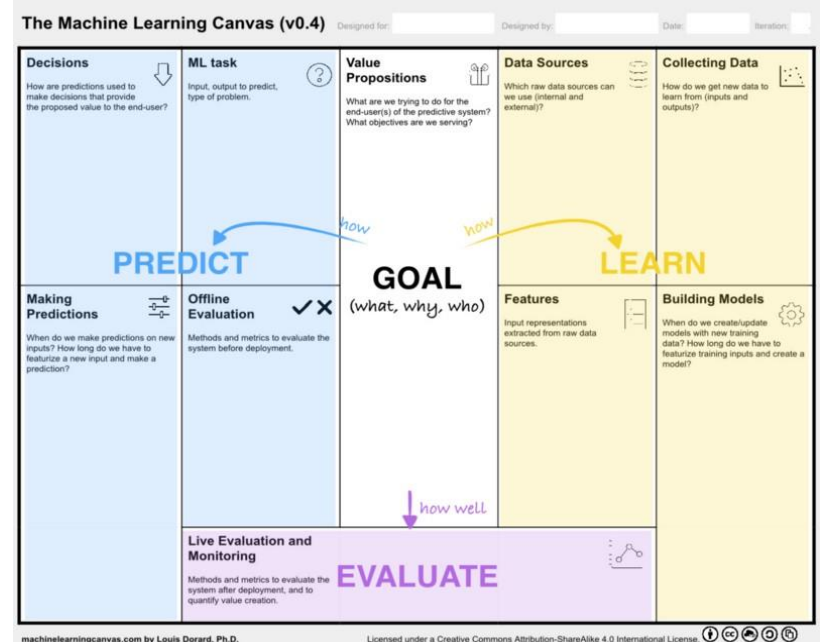
❖ <https://github.com/IBM/adversarial-robustness-toolbox/tree/master/examples>



AI Assessment Canvas

An AI Assessment Canvas

- ❖ The AI Assessment Canvas is a great tool for planning, communication and ML project tracking
- ❖ However, its focus is on how the **problem** will be solved, and not what **questions** the solution need to address
- ❖ Hence, we need to recreate this Canvas, moving it from
 - ❖ **value-centric**
 - to
 - ❖ **safety-centric**
 - decision-making



An AI Assessment Canvas

The AI Assessment Canvas (v1.0)

Designed for:

Designed by:

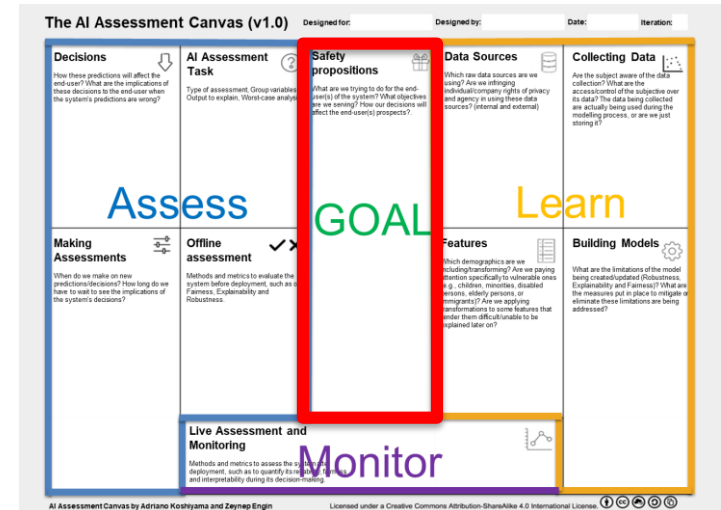
Date:

Iteration:



❖ Section and content of ML Canvas

- ❖ Value propositions: What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?



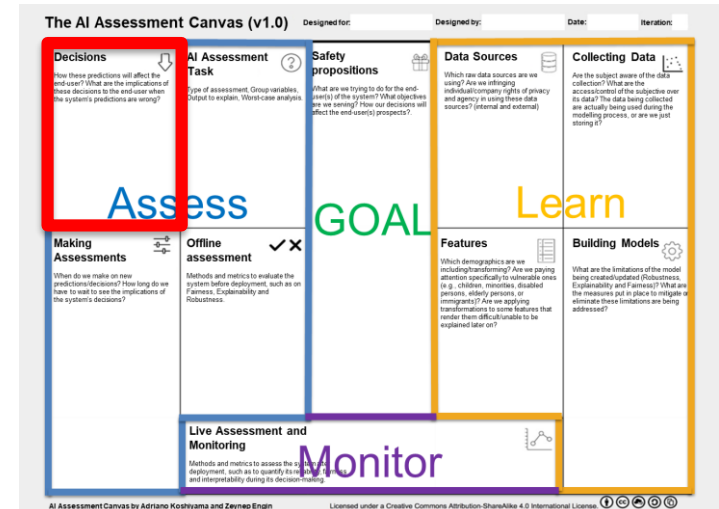
❖ Section and content of AI Assessment Canvas

- ❖ Safety propositions:
 - ❖ What are we trying to do for the end-user(s) of the system?
 - ❖ What objectives are we serving?
 - ❖ How our decisions will affect the end-user(s) prospects?

An AI Assessment Canvas

❖ Section and content of ML Canvas

- ❖ Decisions: How are predictions used to make decisions that provide the proposed value to the end-user?



❖ Section and content of AI Assessment Canvas

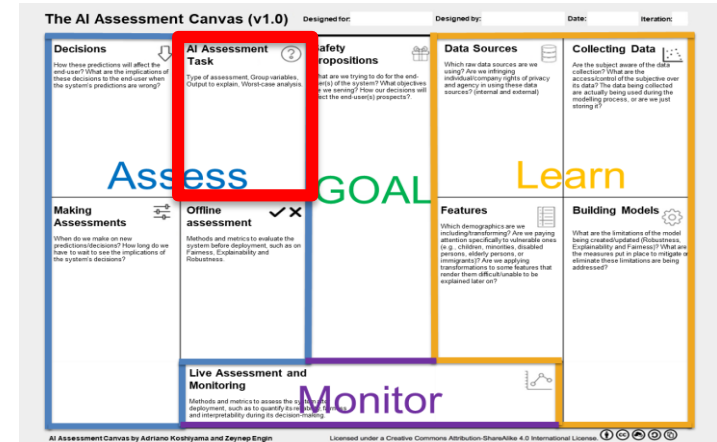
- ❖ Decisions:
 - ❖ How these predictions will affect the end-user?
 - ❖ What are the implications of these decisions to the end-user when the system's predictions are wrong?

An AI Assessment Canvas

❖ Section and content of ML Canvas

❖ ML Task:

- ❖ input: individual transactions
- ❖ output to predict: default
- ❖ type of problem: classification/credit-scoring



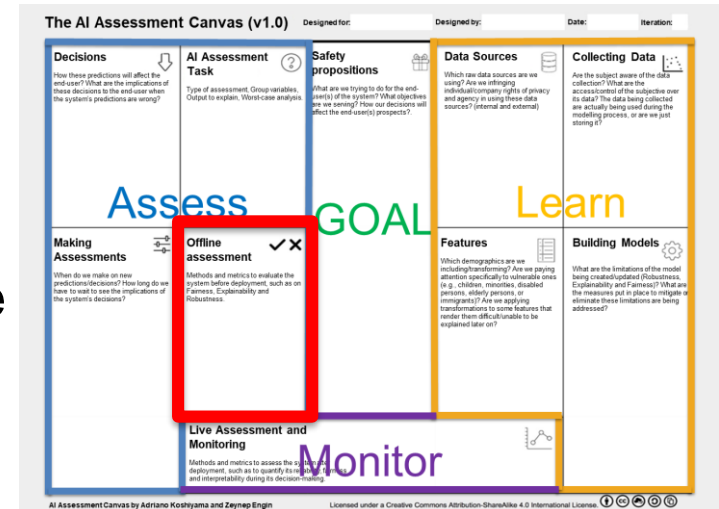
❖ Section and content of AI Assessment Canvas

❖ AI Assessment Task:

- ❖ Type of assessment: fairness/discrimination in credit scoring
- ❖ Group variables: gender, ethnicity, etc.
- ❖ Output to explain: default ratio
- ❖ Worst-case analysis: not applicable

An AI Assessment Canvas

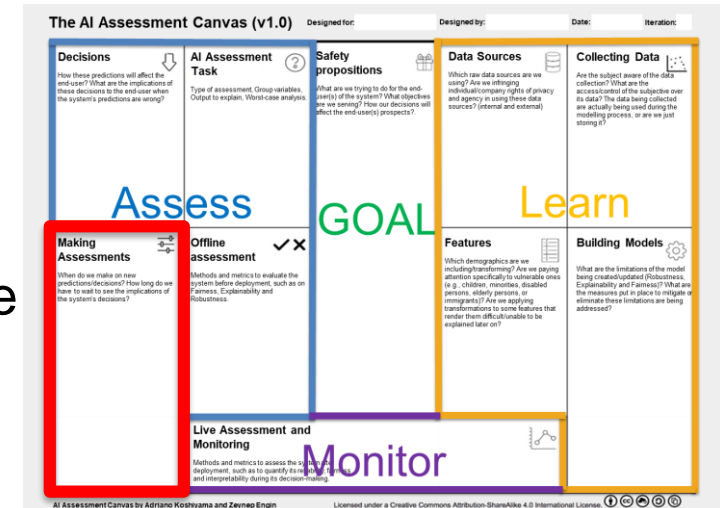
- ❖ Section and content of ML Canvas
- ❖ Offline evaluation: Methods and metrics to evaluate the system before deployment.



- ❖ Section and content of AI Assessment Canvas
- ❖ Offline assessment:
 - ❖ Fairness: describe here the tools used (e.g., fairness constraints)
 - ❖ Explainability: describe here the tools used (e.g., Shapley-values)
 - ❖ Robustness: describe here the tools used (e.g., cross-validation)

An AI Assessment Canvas

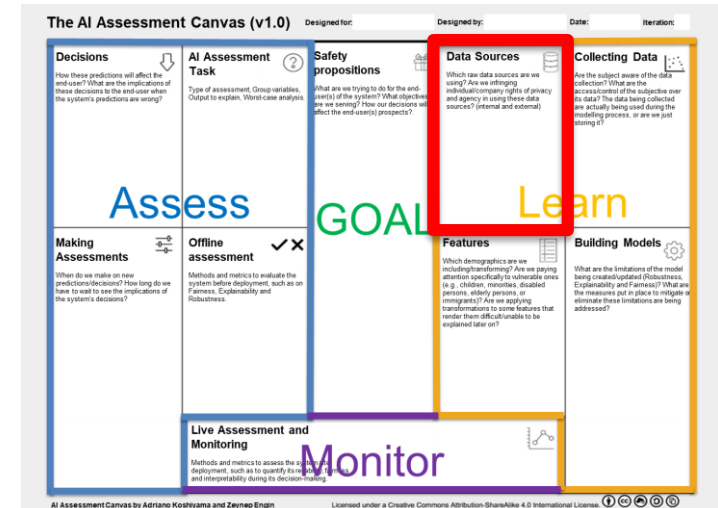
- ❖ Section and content of ML Canvas
 - ❖ Making predictions: When do we make on new inputs? How long do we have to featurize a new input and make a prediction?
-
- ❖ Section and content of AI Assessment Canvas
 - ❖ Making assessments:
 - ❖ When do we make on new predictions/decisions?
 - ❖ How long do we have to wait to see the implications of the system's decisions?



An AI Assessment Canvas

❖ Section and content of ML Canvas

- ❖ Data sources: which raw data sources can we use (internal and external)?



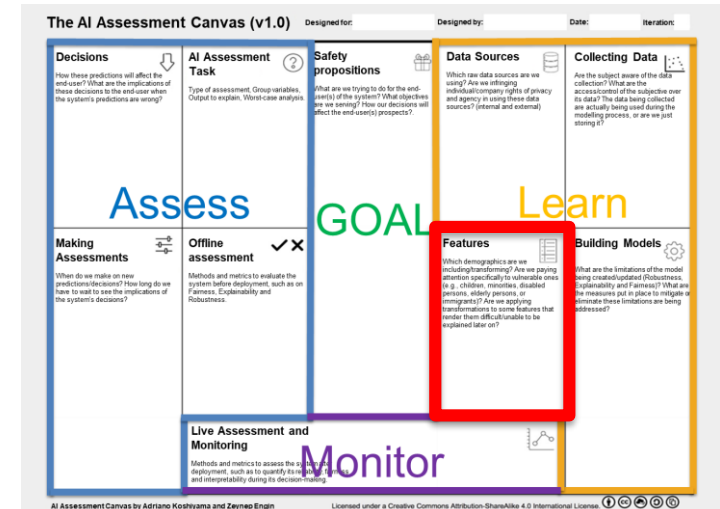
❖ Section and content of AI Assessment Canvas

- ❖ Data sources (internal and external):
 - ❖ Which raw data sources are we using?
 - ❖ Are we infringing individual/company rights of privacy and agency in using these data sources?

An AI Assessment Canvas

❖ Section and content of ML Canvas

- ❖ Features: Input representations extracted from raw data sources.

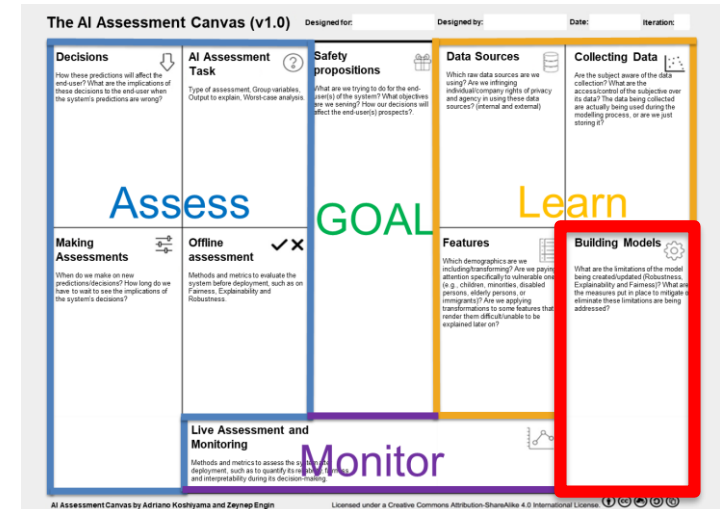


❖ Section and content of AI Assessment Canvas

- ❖ Features:
 - ❖ Which demographics are we including/transforming? Are we paying attention specifically to vulnerable ones (e.g., children, minorities, disabled persons, elderly persons, or immigrants)?
 - ❖ Are we applying transformations to some features that render them difficult/unable to be explained later on?

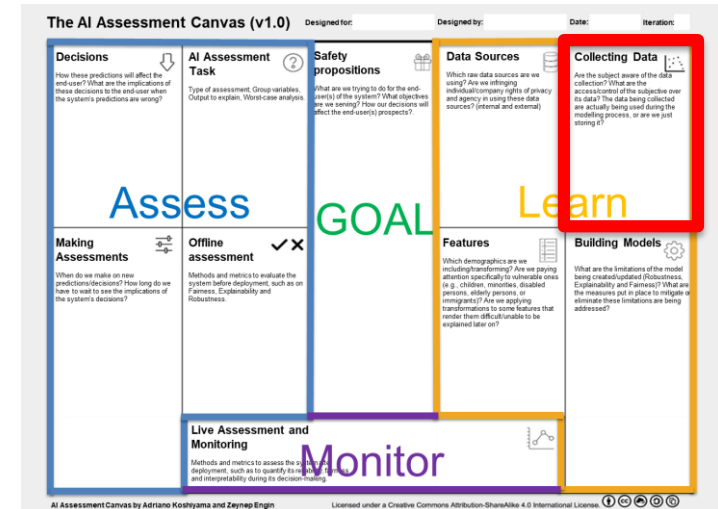
An AI Assessment Canvas

- ❖ Section and content of ML Canvas
- ❖ Building models: When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?
- ❖ Section and content of AI Assessment Canvas
 - ❖ Building models:
 - ❖ What are the limitations of the model being created/updated (Robustness, Explainability and Fairness)?
 - ❖ What are the measures put in place to mitigate or eliminate these limitations are being addressed?



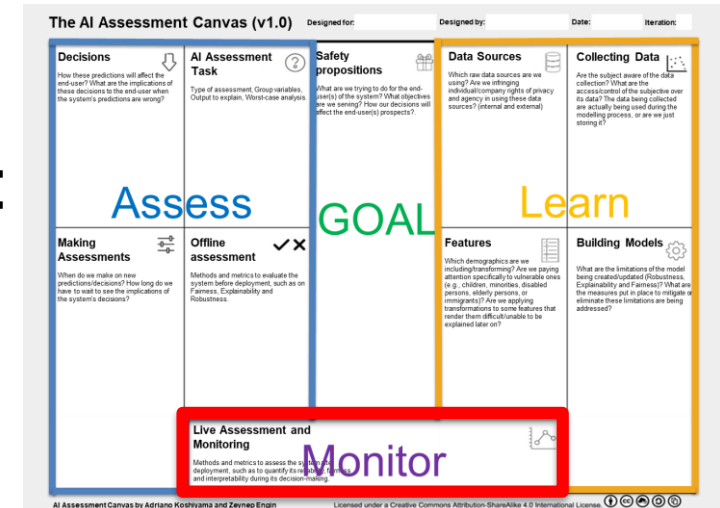
An AI Assessment Canvas

- ❖ Section and content of ML Canvas
- ❖ Collecting data: How do we get new data to learn from (inputs and outputs)?
- ❖ Section and content of AI Assessment Canvas
- ❖ Collecting data:
 - ❖ Are the subject aware of the data collection?
 - ❖ What are the access/control of the subjective over its data?
 - ❖ The data being collected are actually being used during the modelling process, or are we just storing it?



An AI Assessment Canvas

- ❖ Section and content of ML Canvas
- ❖ Live evaluation and monitoring:
Methods and metrics to evaluate the system after deployment, and to quantify value creation.
- ❖ Section and content of AI Assessment Canvas
- ❖ Live evaluation and monitoring:
 - ❖ Methods and metrics to assess the system after deployment, such as to quantify its reliability, fairness and interpretability during its decision-making.



An AI Assessment Canvas











The AI Assessment Canvas (v1.0)

Designed for:

Designed by:

Date:

Iteration:

Decisions  <p>How these predictions will affect the end-user? What are the implications of these decisions to the end-user when the system's predictions are wrong?</p>	AI Assessment Task  <p>Type of assessment, Group variables, Output to explain, Worst-case analysis.</p>	Safety propositions  <p>What are we trying to do for the end-user(s) of the system? What objectives are we serving? How our decisions will affect the end-user(s) prospects?.</p>	Data Sources  <p>Which raw data sources are we using? Are we infringing individual/company rights of privacy and agency in using these data sources? (internal and external)</p>	Collecting Data  <p>Are the subject aware of the data collection? What are the access/control of the subjective over its data? The data being collected are actually being used during the modelling process, or are we just storing it?</p>
Making Assessments  <p>When do we make on new predictions/decisions? How long do we have to wait to see the implications of the system's decisions?</p>	Offline assessment  <p>Methods and metrics to evaluate the system before deployment, such as on Fairness, Explainability and Robustness.</p>		Features  <p>Which demographics are we including/transforming? Are we paying attention specifically to vulnerable ones (e.g., children, minorities, disabled persons, elderly persons, or immigrants)? Are we applying transformations to some features that render them difficult/unable to be explained later on?</p>	Building Models  <p>What are the limitations of the model being created/updated (Robustness, Explainability and Fairness)? What are the measures put in place to mitigate or eliminate these limitations are being addressed?</p>
	Live Assessment and Monitoring  <p>Methods and metrics to assess the system after deployment, such as to quantify its reliability, fairness and interpretability during its decision-making.</p>			

An AI Assessment Canvas










The AI Assessment Canvas (v1.0)

Designed for:

Designed by:

Date:

Iteration:

Decisions 	AI Assessment Task 	Safety propositions 	Data Sources 	Collecting Data 
Making Assessments 	Offline assessment 		Features 	Building Models 
	Live Assessment and Monitoring. 