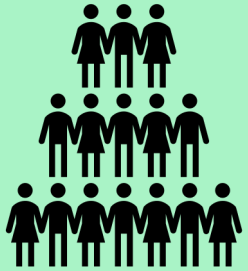
Building Trustworthy AI Systems - Understanding needs, impact, process, & tools



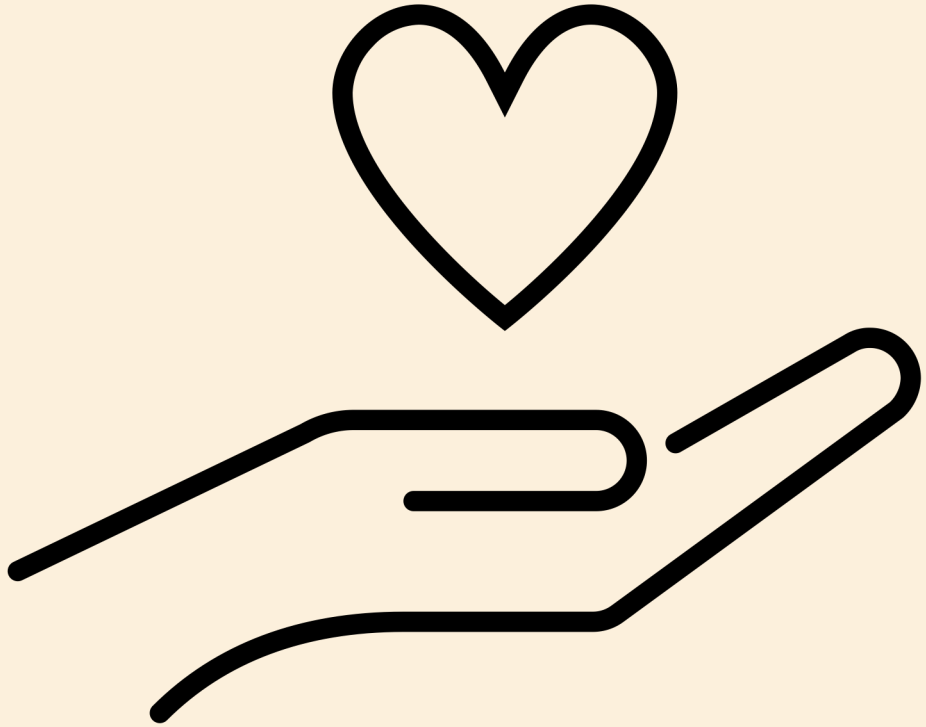
We are in this together

Safe Space With

Saishruthi Swaminathan

Trustworthy AI Advocate

Why should we care?



There Is a Racial Divide in Speech-Recognition Systems, Researchers Say

https://www.nytimes.com/2020/03/23/technology/speech-recognition-bias-apple-amazon-google.html

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Twitter investigates racial bias in image previews

21 September 2020

https://www.bbc.com/news/technology-54234822

ARTIFICIAL INTELLIGENCE

LinkedIn's job-matching AI was biased. The company's solution? More AI.

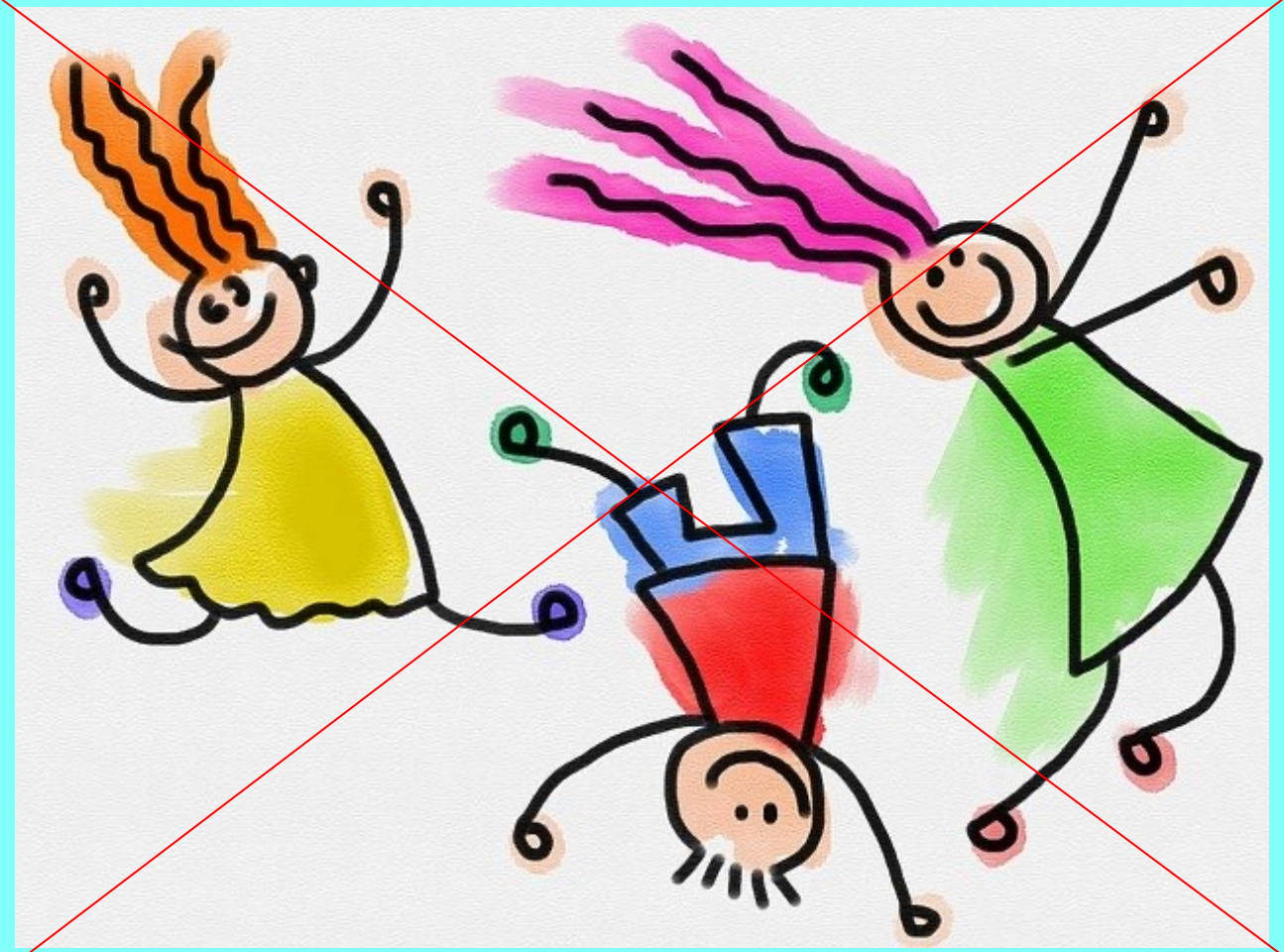
ZipRecruiter, CareerBuilder, LinkedIn—most of the world's biggest job search sites use AI to match people with job openings. But the algorithms don't always play fair.

By Sheridan Wall & Hilke Schellmann June 23, 2021

https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/

How can AI systems earn Trust?

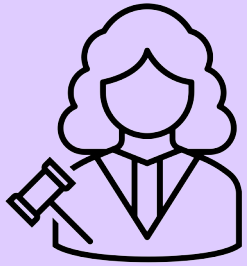
My system has 99% accuracy



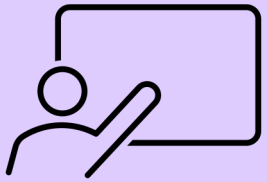
No more joy with just 99% accuracy

Image from pixabay

What we need is more than model accuracy!



Is the model fair ?

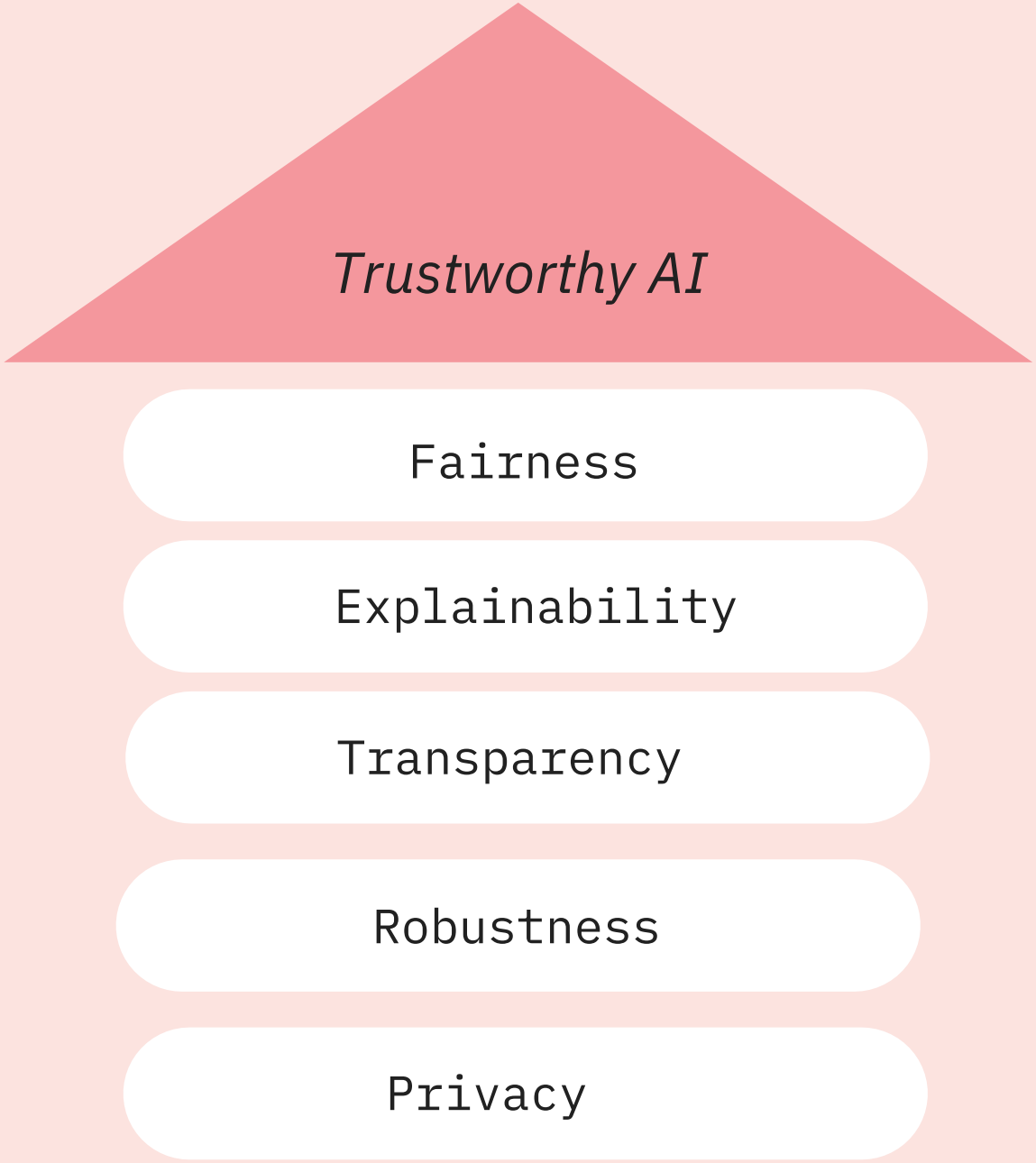


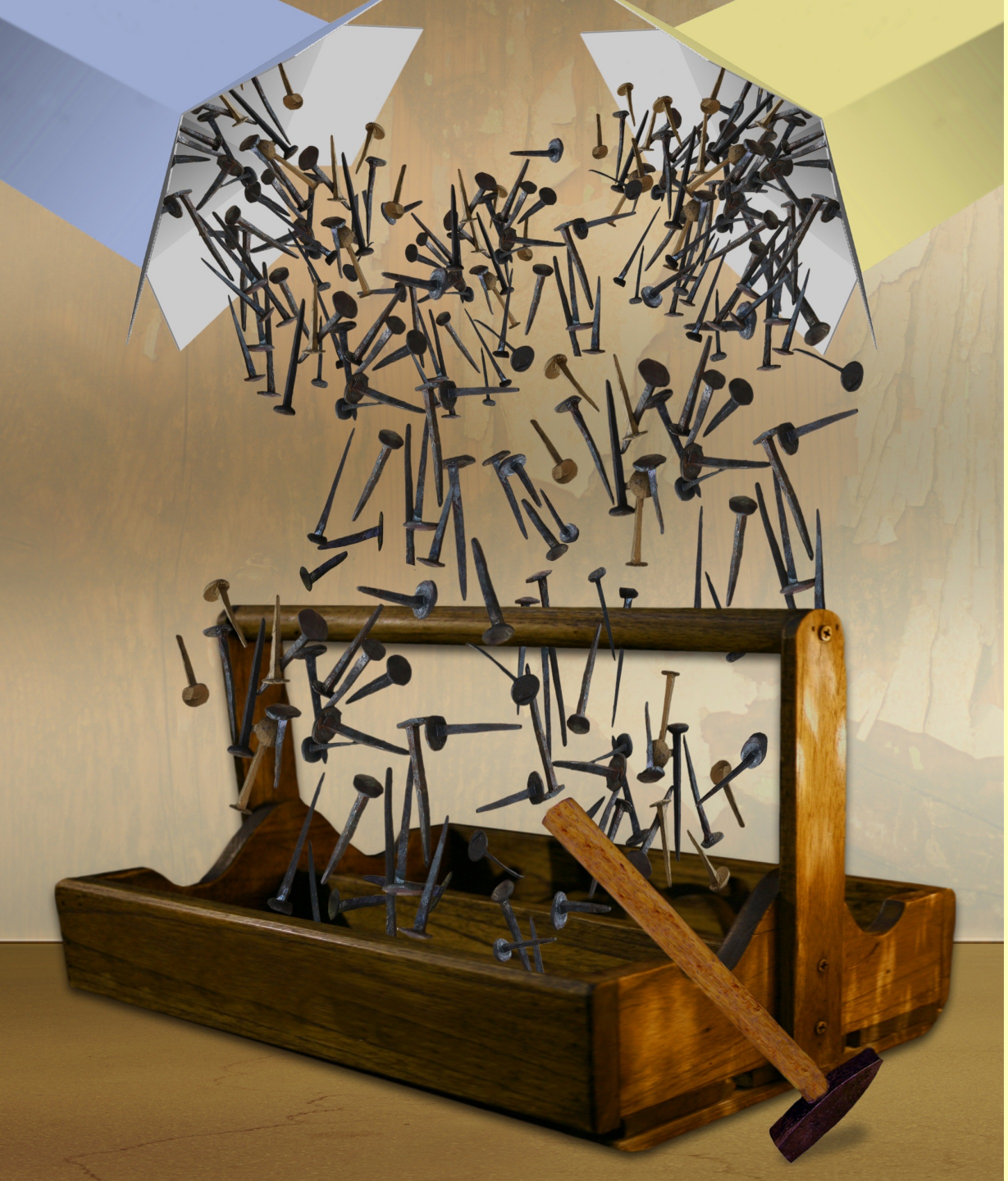
Is model explainable?

Is data protected?

Is it transparent?

Is it robust?


Trustworthy AI
Fairness
Explainability
Transparency
Robustness
Privacy



Can tools alone solve trust issue?

https://get.pxhere.com/photo/wood-tool-hammer-box-nail-sculpture-art-sketch-drawing-illustration-design-assembly-tourist-attraction-modern-art-tinker-ancient-history-1068993.jpg

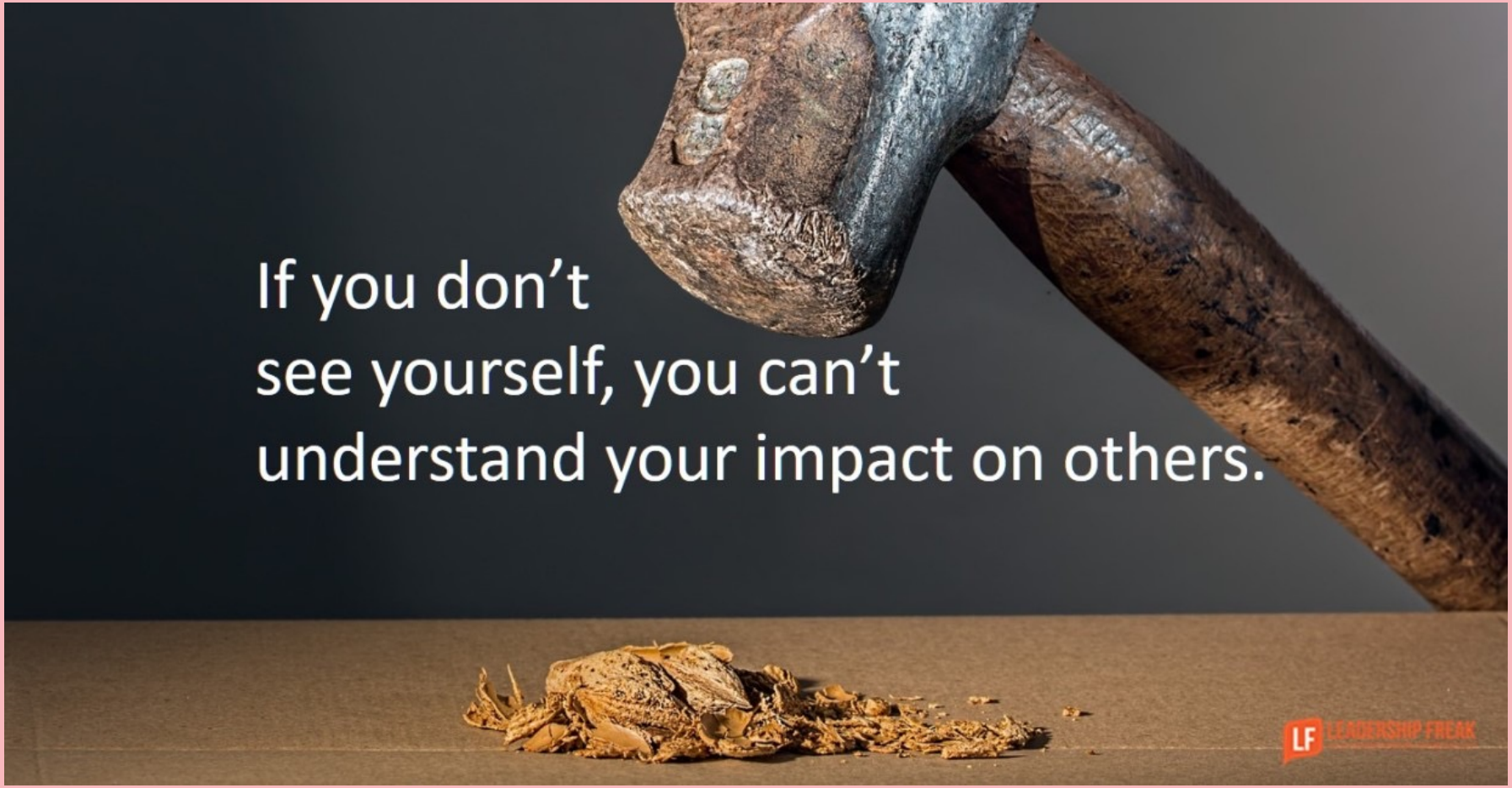
Socio-technological Challenge

Culture and Diversity

Governance

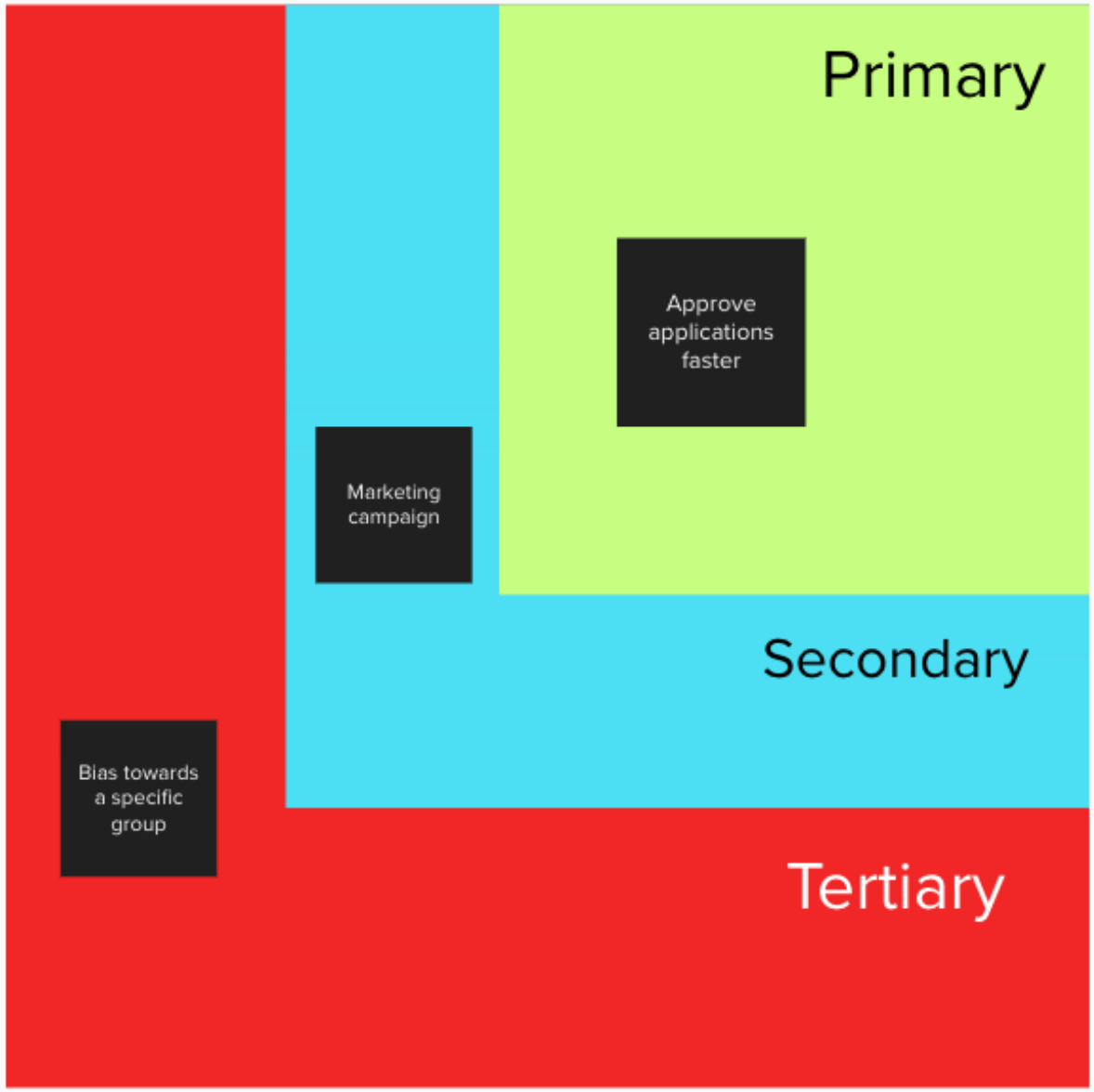
Tools

Self-awareness




PAUSE

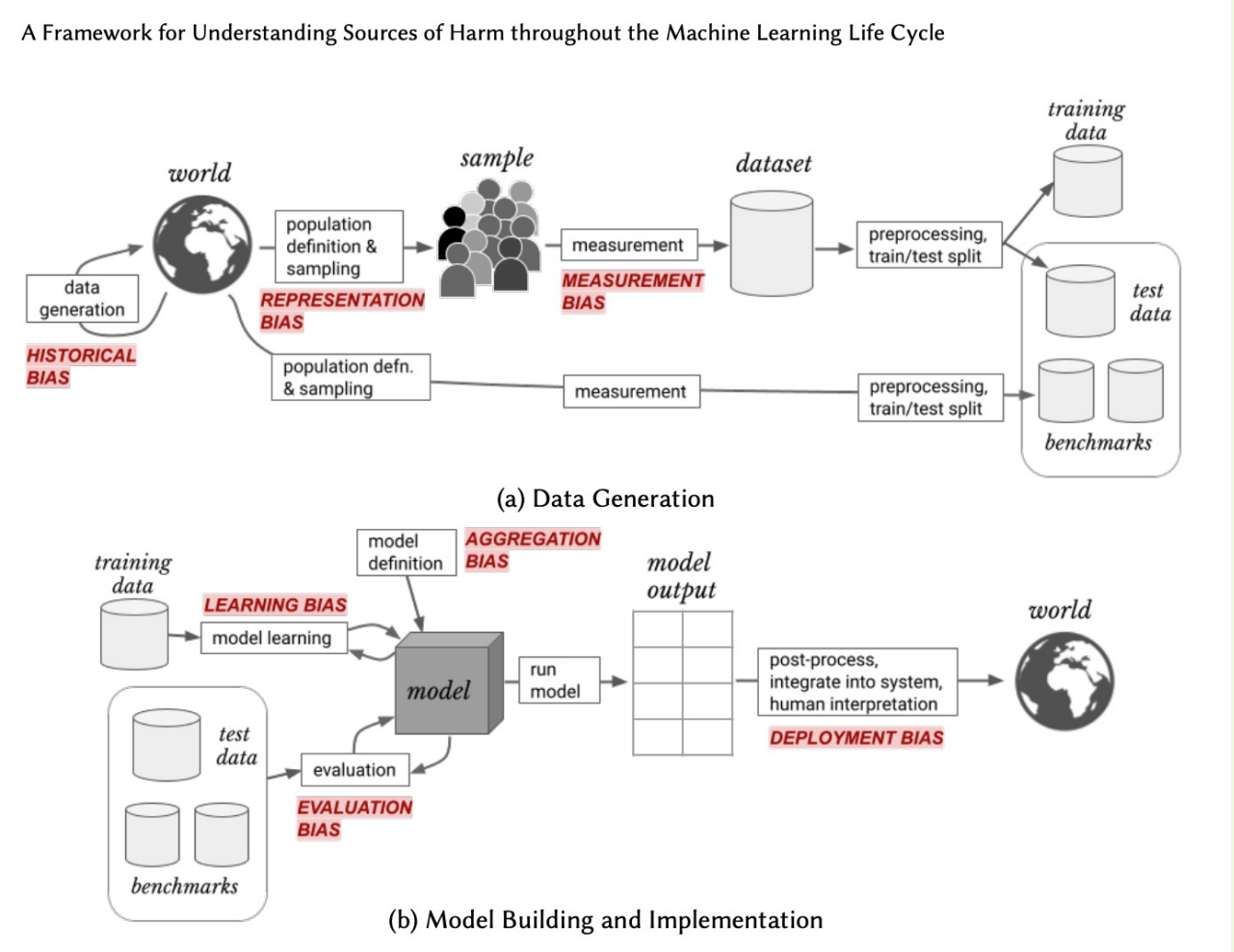
Layers of Effect



https://www.designethically.com/layers

Fairness

Types of Bias



https://arxiv.org/pdf/1901.10002.pdf

Open Source Tools for Fairness

IBM AI Fairness 360 Toolkit

Microsoft Fairlearn

Google’s What-if Tool

Aequitas

Scikit Fairness

LinkedIn Fairness Toolkit (LiFT)

AI Fairness 360 Toolkit

An extensible, open source toolkit for measuring, understanding, and reducing AI bias. It combines the top bias metrics, bias mitigation algorithms, and metric explainers from fairness researchers across industry and academia

Implement techniques from eight published papers across the greater AI fairness research community

Available in Python and R

Terminologies

Favorable label: A label whose value corresponds to an outcome that provides an advantage to the recipient (such as receiving a loan, being hired for a job, not being arrested)

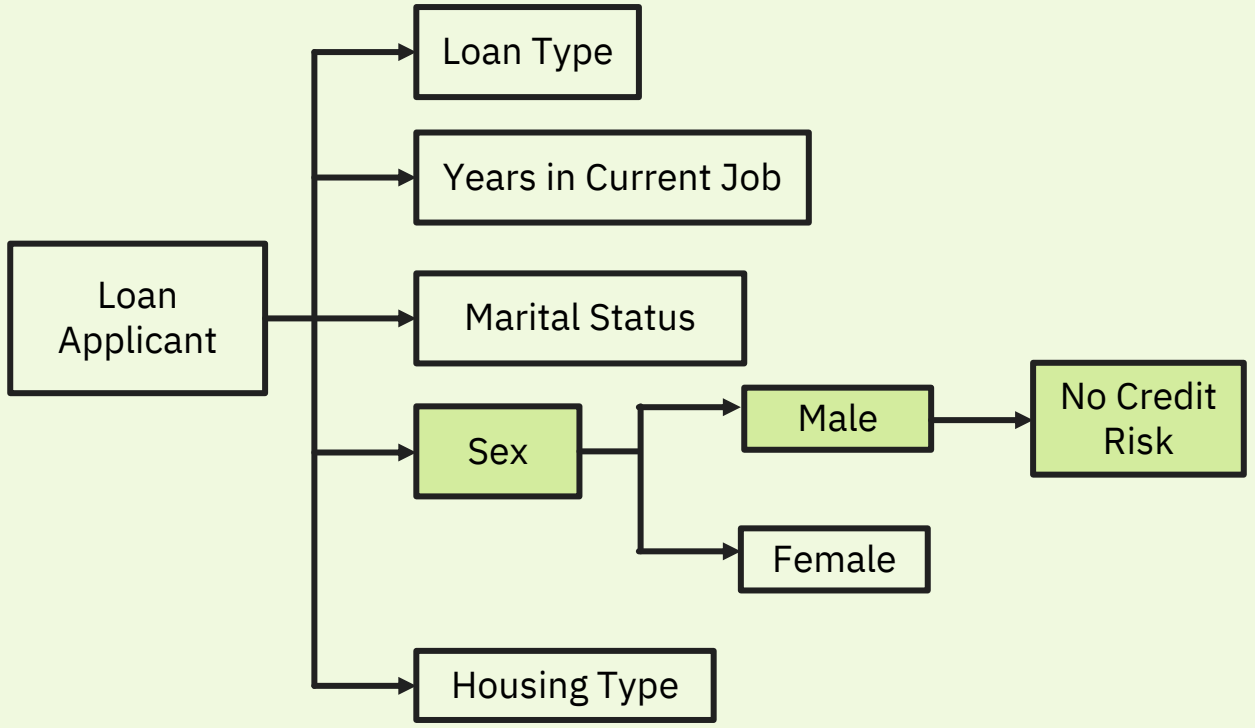
Protected attribute: An attribute that partitions a population into groups whose outcomes should have parity (such as race, sex, caste, and religion)

Privileged value (of a protected attribute): A protected attribute value indicating a group that has historically been at a systemic advantage

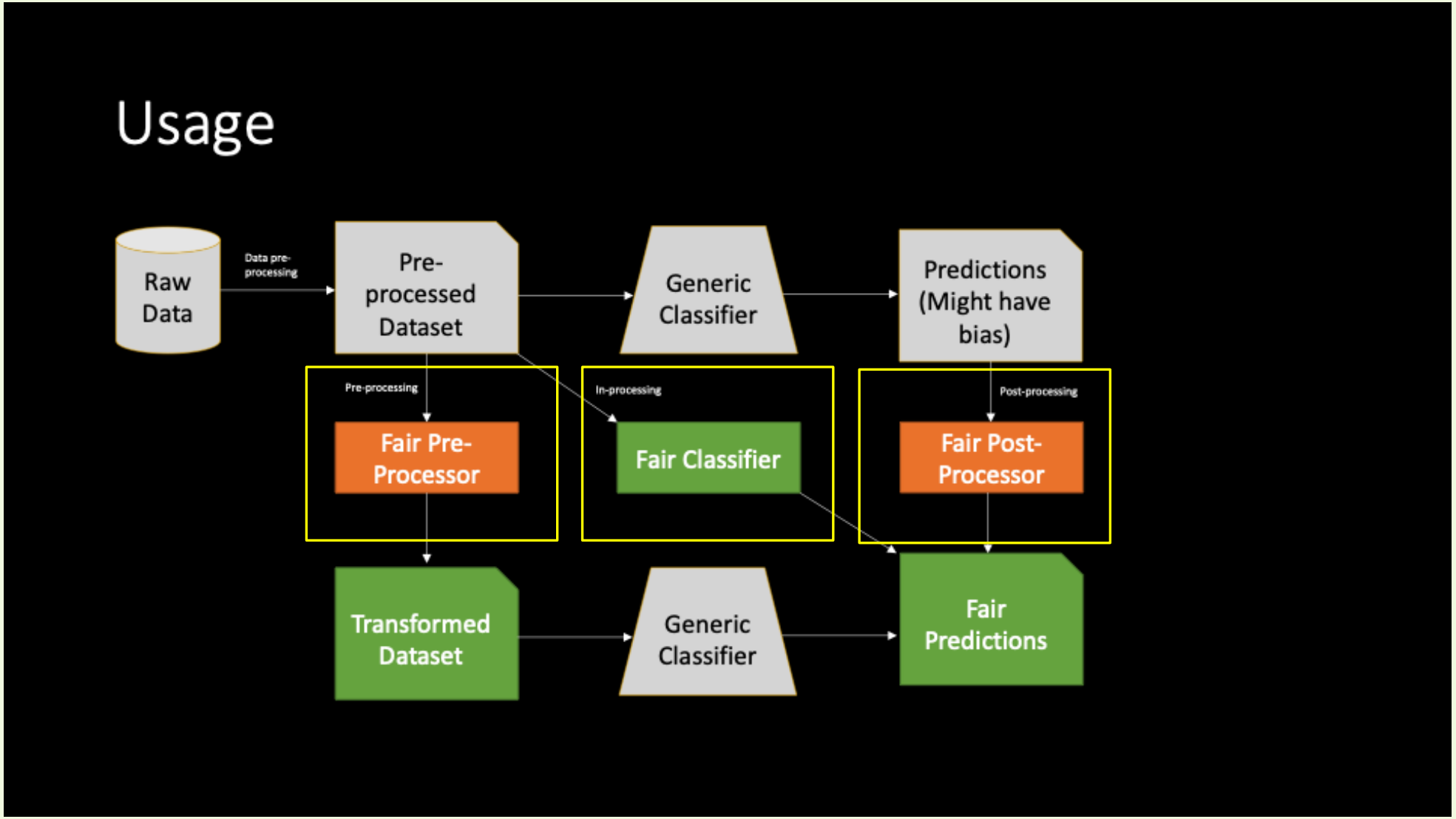
Discrimination/unwanted bias: When specific privileged groups are placed at a systematic advantage and specific unprivileged groups are placed at a systematic disadvantage. This relates to attributes such as race, sex, age, and sexual orientation.

Terminologies

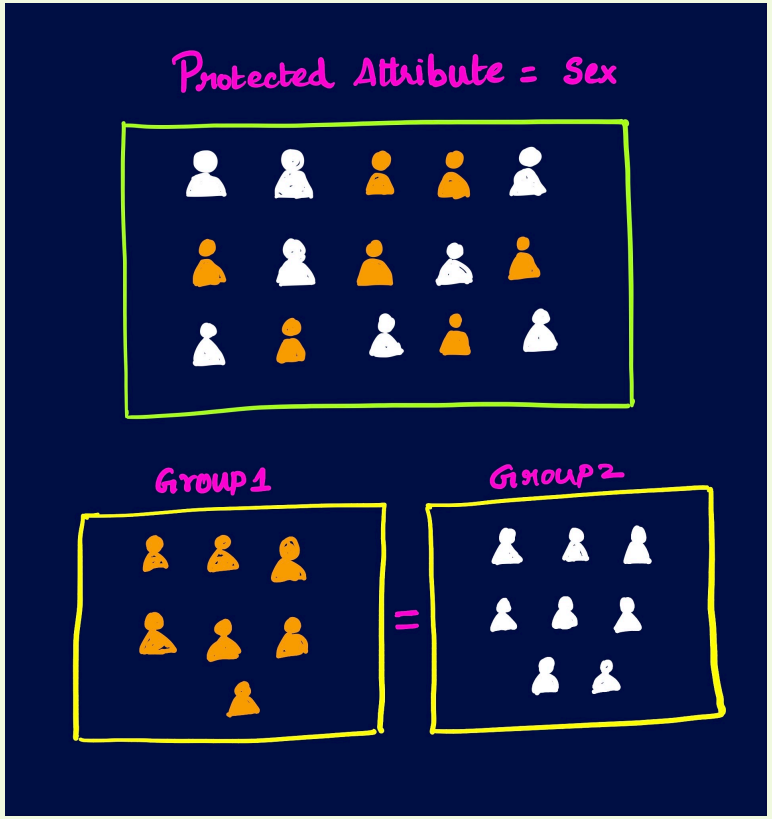
- Favorable label: No Credit Risk
- Unfavorable label: Credit Risk
- Protected Attribute: Sex
- Privileged Protected Attribute: Male



Where can you use the toolkit?



Metrics to Detect Bias



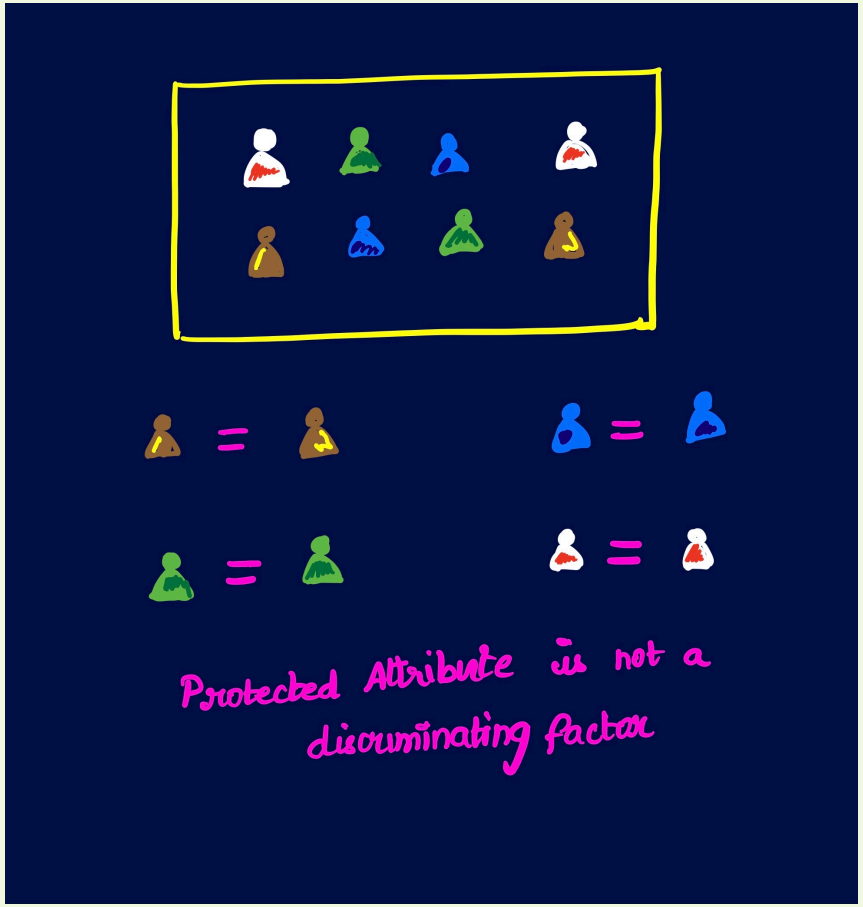
Group fairness

Partitions a population into groups defined by protected attributes & seeks for some statistical measure to be equal across groups.

Metrics to Detect Bias

Individual fairness

Seeks for similar individuals to be treated similarly.

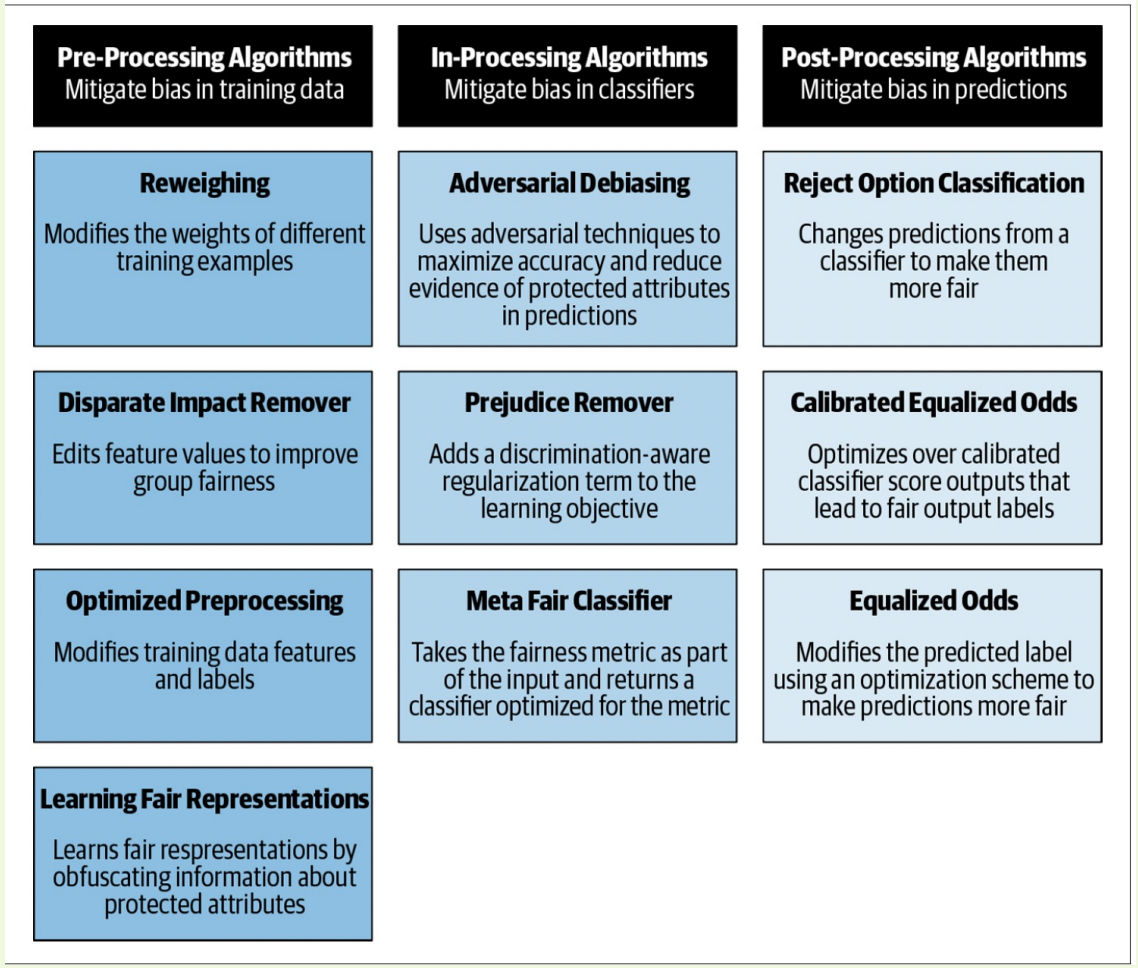


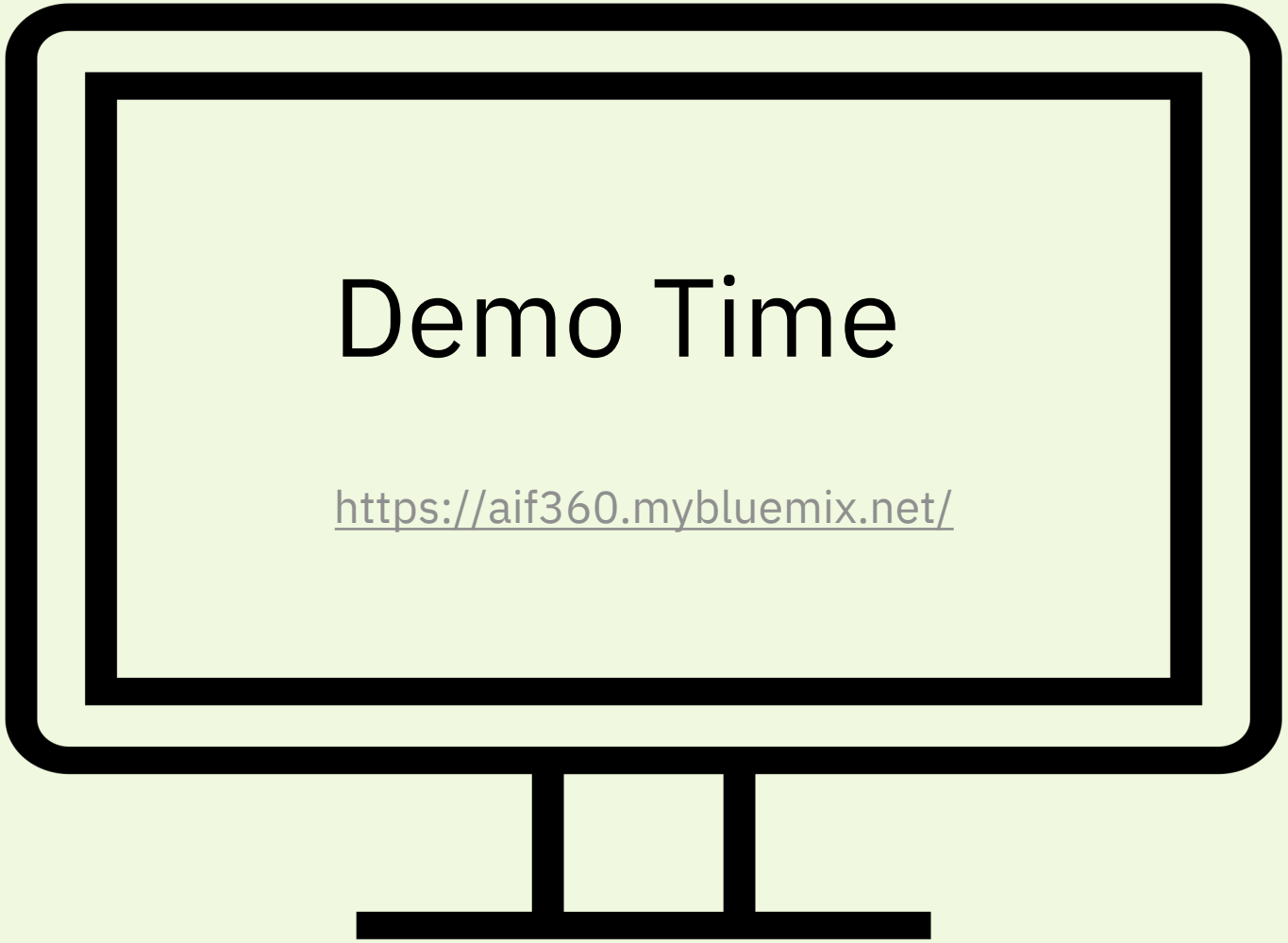
Algorithms

- Bias mitigation algorithms attempt to improve the fairness metrics by modifying the training data, the learning algorithm, or the predictions.

- These algorithm categories are known as pre-processing, in-processing, and post-processing, respectively.

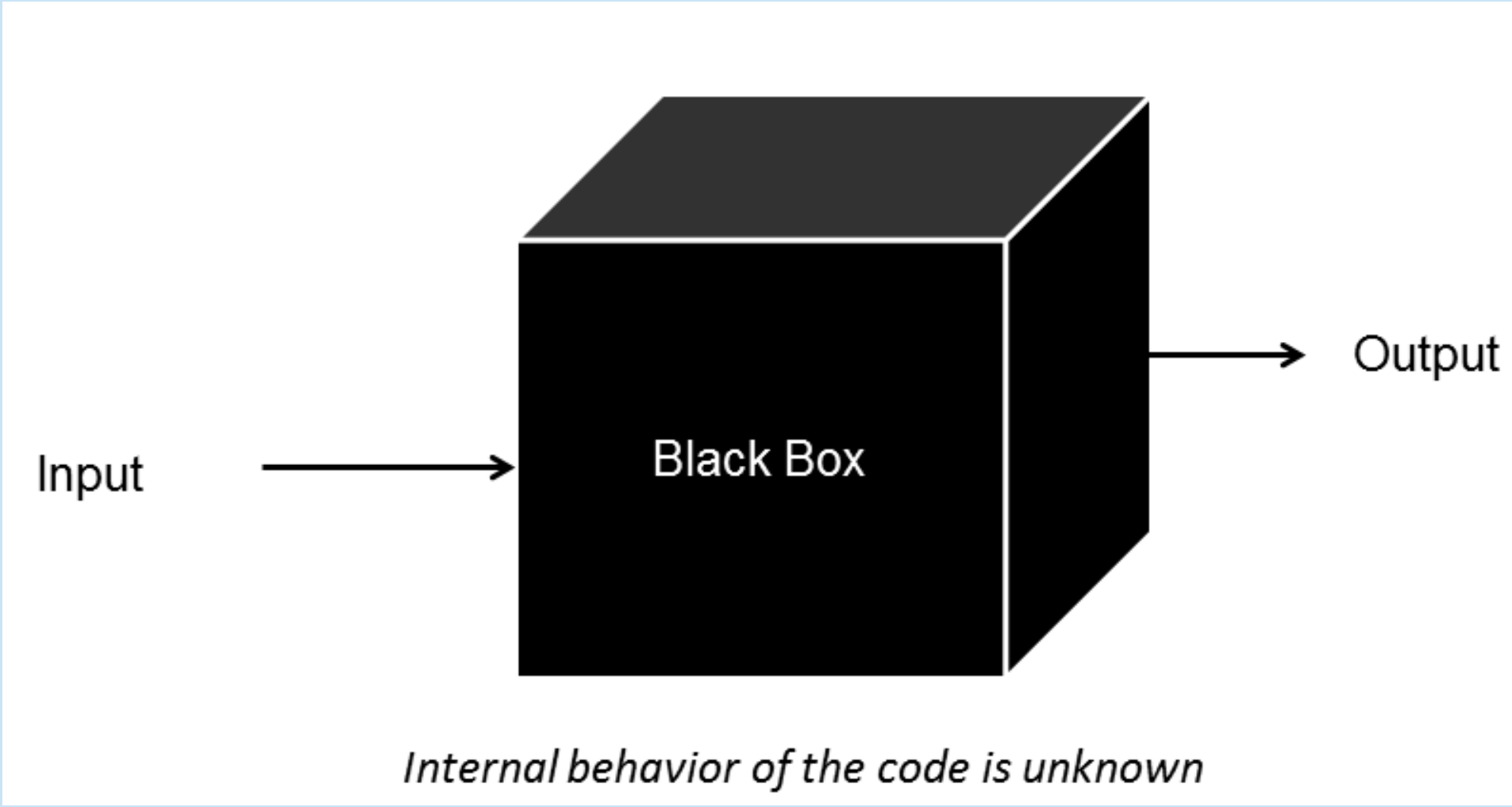
Algorithms

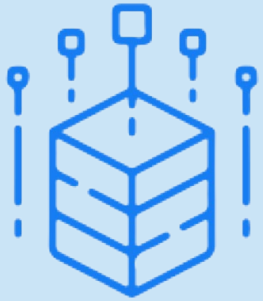

Pre-Processing Algorithms
Mitigate bias in training data
In-Processing Algorithms
Mitigate bias in classifiers
Post-Processing Algorithms
Mitigate bias in predictions
Reweighing
Modifies the weights of different training examples
Adversarial Debiasing
Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions
Reject Option Classification
Changes predictions from a classifier to make them more fair
Disparate Impact Remover
Edits feature values to improve group fairness
Prejudice Remover
Adds a discrimination-aware regularization term to the learning objective
Calibrated Equalized Odds
Optimizes over calibrated classifier score outputs that lead to fair output labels
Optimized Preprocessing
Modifies training data features and labels
Meta Fair Classifier
Takes the fairness metric as part of the input and returns a classifier optimized for the metric
Equalized Odds
Modifies the predicted label using an optimization scheme to make predictions more fair
Learning Fair Representations
Learns fair respresentations by obfuscating information about protected attributes


Demo Time
https://aif360.mybluemix.net/

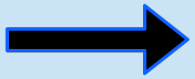
Explainability

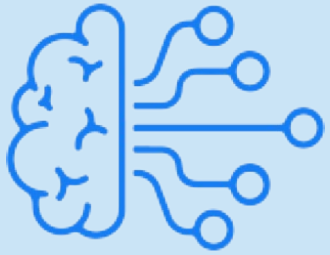
The black box problem



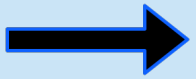


Dataset





AI/ML Model



Decision
or Action

- Why does the model decide or do that?
- How sure is the model about this decision?
- How can I claim in case of error?
- How can I be sure there are no biases?
- Why should I trust the model?

How does a model work?

What is driving the decisions?

Can I trust the model?

Key Stakeholders

Data Scientist

- Understand the model
- Debug it
- Improve its performance

Business owner

- Understand the model
- Evaluate fit for purpose
- Agree to use

Model Risk

- Challenge the model
- Ensure its robustness
- Approve it

Regulator

- Check its impact on consumers
- Verify reliability

Consumer

- “What is the impact on me?”
- “What actions can I take?“

AI Explainability 360
↳ (AIX360)
https://github.com/IBM/AIX360

AIX360

AIX360 toolkit is an open-source library to help **explain** AI and ML models and their predictions.

This includes 3 classes of algorithms: local post-hoc, global post-hoc, and directly interpretable explainers for models that use image, text, and structured/tabular data.

Toolbox
Local post-hoc
Global post-hoc
Directly interpretable

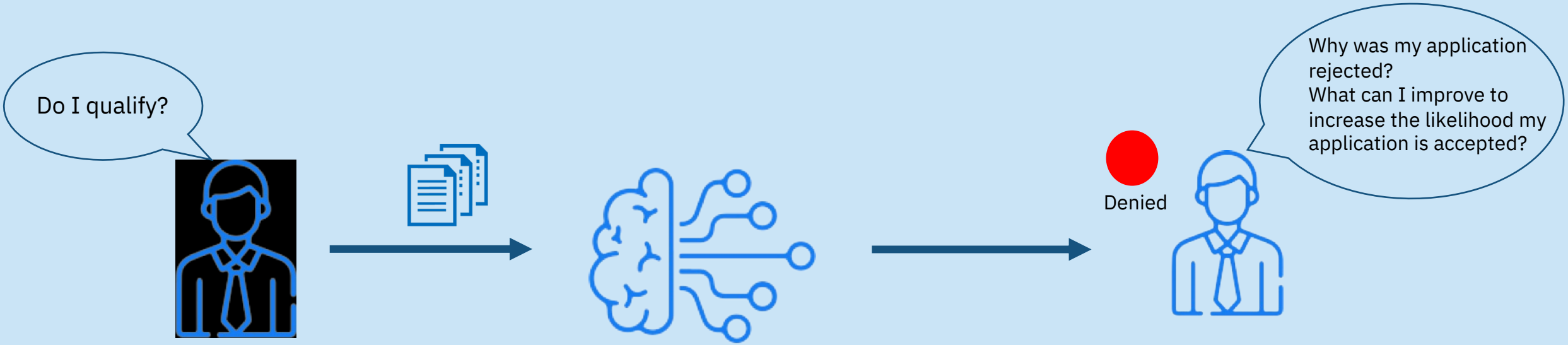
FICO Explainable Machine Learning Challenge

Use Case:

The customers in this dataset have requested a credit line in the range of $5,000 - $150,000.

The fundamental task is to use the information about the applicant in their credit report to predict whether they will make timely payments over a two-year period. This is the machine learning task that we focus on.

The machine learning prediction is then used by loan officers to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.



Transparency

Open Source Tools for Transparency

IBM AI Factsheet 360 Toolkit

Google Model Cards

Transparent reporting mechanisms are the basis for trust in many industries and applications

Nutrition Facts

Serving Size 8 oz
Servings Per Container 1.5

Amount Per Serving

Calories 23

	% Daily Value*
Total Fat 0g	**0%**
Saturated Fat 0g	**0%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 0mg	**0%**
Total Carbohydrate 5g	**2%**
Dietary Fiber 0g	**0%**
Sugars 6g	
Protein 1g	**2%**

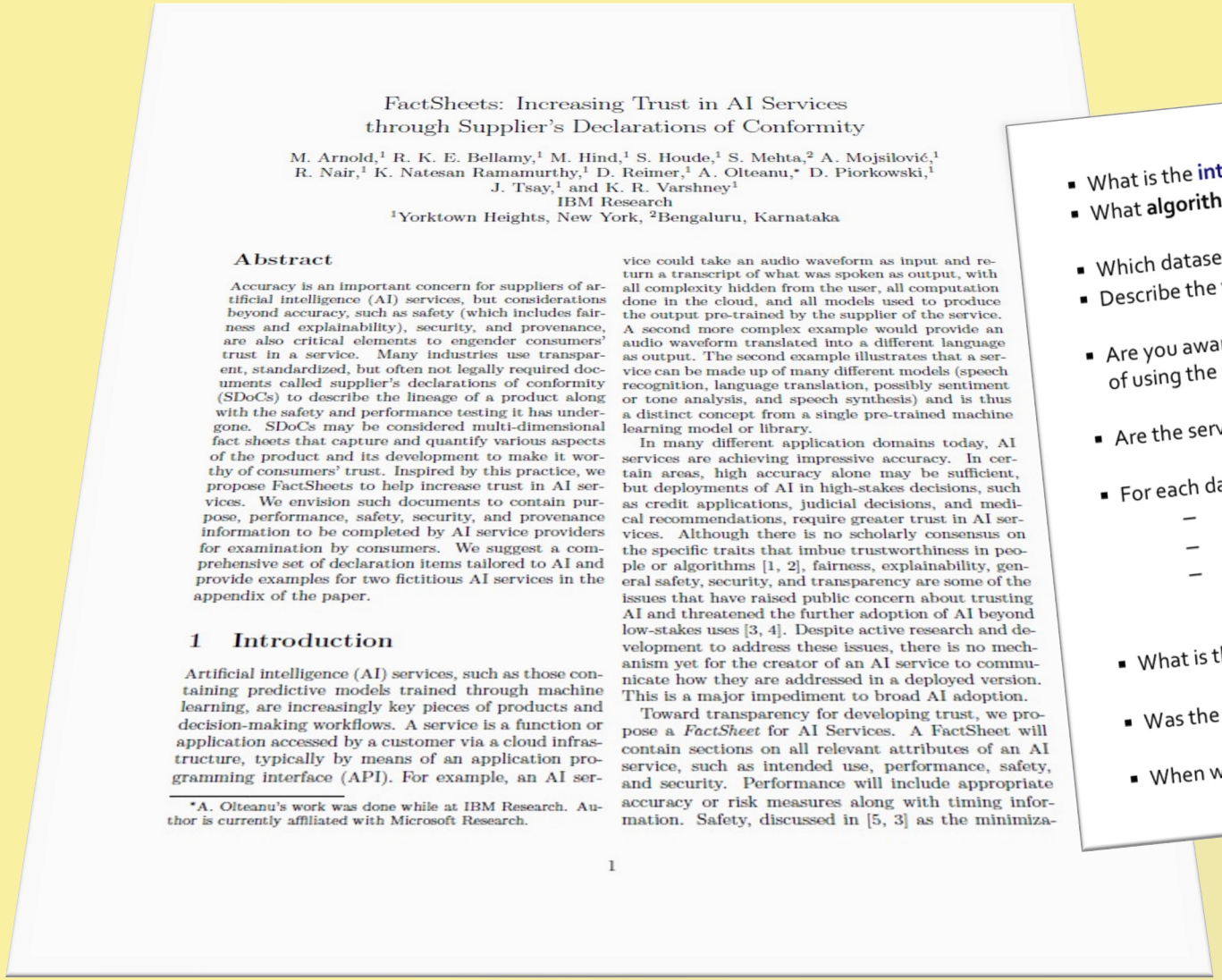
*Percent Daily Values are based on a 2,000 calorie diet.

ENERGY STAR

Moody's		S&P		Fitch		Rating description	
Long-term	Short-term	Long-term	Short-term	Long-term	Short-term		
Aaa	P-1	AAA	A-1+	AAA	F1+	Prime	Investment-grade
Aa1		AA+		AA+		High grade	
Aa2		AA		AA			
Aa3		AA−		AA−			
A1		A+	A-1	A+	F1	Upper medium grade	
A2		A		A			
A3	P-2	A−	A-2	A−	F2		
Baa1		BBB+		BBB+		Lower medium grade	
Baa2	P-3	BBB	A-3	BBB	F3		
Baa3		BBB−		BBB−			
Ba1	Not prime	BB+	B	BB+	B	Non-investment grade speculative	Non-investment grade aka high-yield bonds aka junk bonds
Ba2		BB		BB			
Ba3		BB−		BB−			
B1		B+		B+		Highly speculative	
B2		B		B			
B3		B−		B−			
Caa1		CCC+	C	CCC	C	Substantial risks	
Caa2		CCC				Extremely speculative	
Caa3		CCC−				Default imminent with little prospect for recovery	
Ca		CC					
		C					
C		D	/	DDD	/	In default	
/				DD			
				D			

IEEE STANDARDS ASSOCIATION

In 2018, IBM Research proposed the concept of AI FactSheets



FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity

M. Arnold,[1] R. K. E. Bellamy,[1] M. Hind,[1] S. Houde,[1] S. Mehta,[2] A. Mojsilović,[1] R. Nair,[1] K. Natesan Ramamurthy,[1] D. Reimer,[1] A. Olteanu,* D. Piorkowski,[1] J. Tsay,[1] and K. R. Varshney[1]
IBM Research
[1]Yorktown Heights, New York, [2]Bengaluru, Karnataka

Abstract

Accuracy is an important concern for suppliers of artificial intelligence (AI) services, but considerations beyond accuracy, such as safety (which includes fairness and explainability), security, and provenance, are also critical elements to engender consumers' trust in a service. Many industries use transparent, standardized, but often not legally required documents called supplier's declarations of conformity (SDoCs) to describe the lineage of a product along with the safety and performance testing it has undergone. SDoCs may be considered multi-dimensional fact sheets that capture and quantify various aspects of the product and its development to make it worthy of consumers' trust. Inspired by this practice, we propose FactSheets to help increase trust in AI services. We envision such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers. We suggest a comprehensive set of declaration items tailored to AI and provide examples for two fictitious AI services in the appendix of the paper.

1 Introduction

Artificial intelligence (AI) services, such as those containing predictive models trained through machine learning, are increasingly key pieces of products and decision-making workflows. A service is a function or application accessed by a customer via a cloud infrastructure, typically by means of an application programming interface (API). For example, an AI service could take an audio waveform as input and return a transcript of what was spoken as output, with all complexity hidden from the user, all computation done in the cloud, and all models used to produce the output pre-trained by the supplier of the service. A second more complex example would provide an audio waveform translated into a different language as output. The second example illustrates that a service can be made up of many different models (speech recognition, language translation, possibly sentiment or tone analysis, and speech synthesis) and is thus a distinct concept from a single pre-trained machine learning model or library.

In many different application domains today, AI services are achieving impressive accuracy. In certain areas, high accuracy alone may be sufficient, but deployments of AI in high-stakes decisions, such as credit applications, judicial decisions, and medical recommendations, require greater trust in AI services. Although there is no scholarly consensus on the specific traits that imbue trustworthiness in people or algorithms [1, 2], fairness, explainability, general safety, security, and transparency are some of the issues that have raised public concern about trusting AI and threatened the further adoption of AI beyond low-stakes uses [3, 4]. Despite active research and development to address these issues, there is no mechanism yet for the creator of an AI service to communicate how they are addressed in a deployed version. This is a major impediment to broad AI adoption.

Toward transparency for developing trust, we propose a *FactSheet* for AI Services. A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance, safety, and security. Performance will include appropriate accuracy or risk measures along with timing information. Safety, discussed in [5, 3] as the minimiza-

*A. Olteanu's work was done while at IBM Research. Author is currently affiliated with Microsoft Research.

1

- What is the **intended use** of the service output?
- What **algorithms** or techniques does this service implement?
- Which datasets was the service **tested** on?
- Describe the **testing methodology** and **test results**.
- Are you aware of possible examples of **bias, ethical** issues, or other **safety risks** as a result of using the service?
- Are the service outputs **explainable** and/or interpretable?
- For each dataset used by the service:
  - Was the dataset checked for **bias**?
  - What efforts were made to ensure that it is **fair** and **representative**?
  - Does the service implement and perform any **bias detection and remediation**?
- What is the **expected performance** on unseen data or data with different distributions?
- Was the service checked for **robustness against adversarial attacks**?
- When were the models last updated?

VB

AI

IBM researchers propose 'factsheets' for AI transparency

https://arxiv.org/abs/1808.07261

In 2019, IBM Research described experiences completing AI FactSheets

There is no magic set of questions for everyone for a FactSheet

The appropriate questions vary from :

- person to person
- project to project
- role to role
- regulations in US , EU guidelines, industry consortia
- transparency vs governance needs

https://arxiv.org/abs/1911.08293

Experiences with Improving the Transparency of AI Models and Services

Michael Hind, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic,
David Piorkowski, John Richards, and Kush R. Varshney
IBM Research

Abstract

AI models and services are used in a growing number of high-stakes areas, resulting in a need for increased transparency. Consistent with this, several proposals for higher quality and more consistent documentation of AI data, models, and systems have emerged. Little is known, however, about the needs of those who would produce or consume these new forms of documentation. Through semi-structured developer interviews, and two document creation exercises, we have assembled a clearer picture of these needs and the various challenges faced in creating accurate and useful AI documentation. Based on the observations from this work, supplemented by feedback received during multiple design explorations and stakeholder conversations, we make recommendations for easing the collection and flexible presentation of AI facts to promote transparency.

1 Introduction

AI models and services are being used in a growing number of high-stakes areas such as financial risk assessment (Goyal 2018), medical diagnosis and treatment planning (Strickland 2019), hiring and promotion decisions (Alsever 2017), social services eligibility determination (Fishman, Eggers, and Kishnani 2019), predictive policing (Ensign et al. 2017), and probation and sentencing recommendations (Larson et al. 2016).

While most models are created for bespoke purposes, some are also being packaged in model catalogs for use by others. For many models there will be risk, compliance, and/or regulatory needs for information covering the nature and intended uses of the model, its overall accuracy, its ability to explain particular decisions, its fairness with respect to protected classes, and at least high-level information about the provenance of training data and assurances that suitable privacy protections have been maintained. In reviewing models within a catalog for suitability in a particular application context, there may be an additional need to easily compare multiple candidates.

Recent work has outlined the need for increased transparency in AI for data sets (Gebru et al. 2018; Bender and Friedman 2018; Holland et al. 2018), models (Mitchell et al. 2019), and services (Arnold et al. 2019). Proposals in support of ethical and trusted AI are also emerging (High-Level Expert Group on Artificial Intelligence 2019; Partnership On AI 2019; IEEE 2017). While details differ, all are driving towards a common set of attributes that capture essential "facts" about a model. We are not yet aware of developers adopting these ideas for regular use. Neither are we aware of published work describing developers' needs or the difficulties they face when producing or consuming AI documentation.

In this paper we discuss formative research with developers and other stakeholders to better understand their documentation needs. We also report on a study in which developers in several application areas created AI documentation in the form of a *FactSheet*.

A FactSheet, as proposed by (Arnold et al. 2019), is a collection of relevant information about an AI model or service that is created during the machine learning life cycle. It includes information from the business owner (e.g., intended use and business justification), from the data gathering/feature selection/data cleaning phase (e.g., data sets, features used or created, cleaning operations), from the model training phase (e.g., bias, robustness, and explainability information), and from the model validation and deployment phase (e.g., key performance indicators). A FactSheet is associated with a model (or service) and is meant to be write once, i.e., an update to a model would trigger a new FactSheet for the updated model. FactSheets can be consumed by any role in the ML life cycle to confirm process governance adherence or model performance, or by the ultimate users of a model to provide increased transparency. Of course, the diversity of model types, and the range of possible application domains, makes the specification of a common FactSheet schema difficult. We hope the work reported here provides useful guidance going forward.

The contributions of this paper include

- summaries of semi-structured interviews with AI developers on their documentation needs and practices,
- observations on documentation requirements and difficulties of AI developers in creating FactSheets,
- additional requirements from feedback on prototype FactSheet designs and unstructured interviews with stakeholders involved throughout the AI life cycle, and

FactSheets and Different Flavors of Trust

AI Transparency

AI Marketplace

Enabling AI consumers to find the right trusted AI technology for their needs

AI Governance

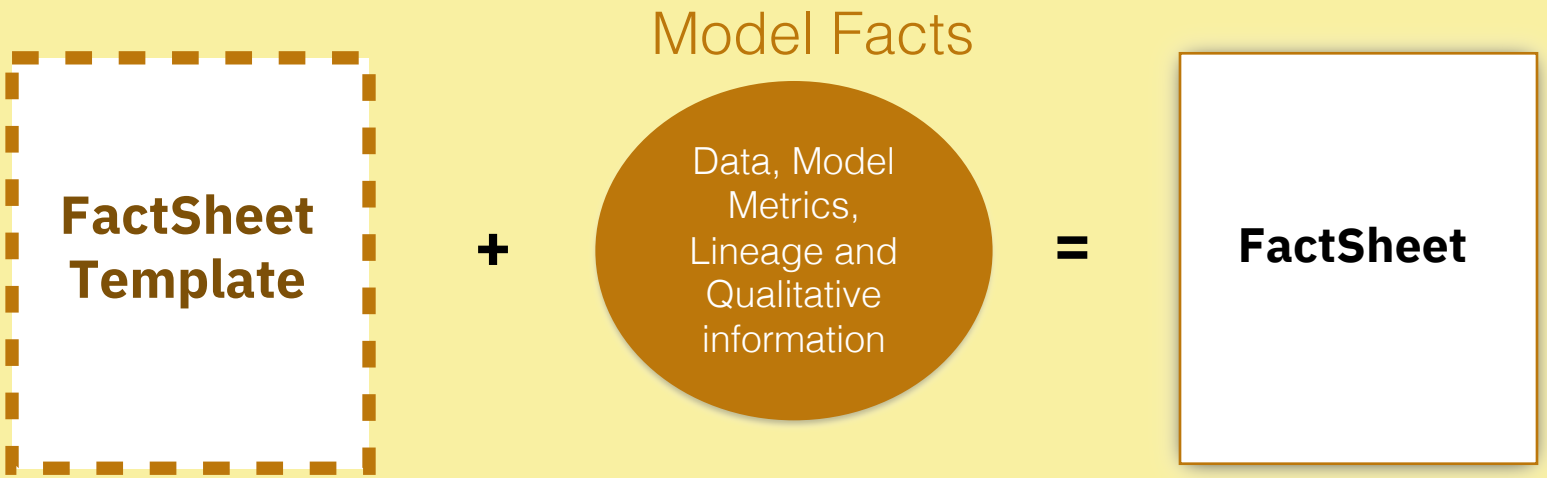
Enterprise AI Documentation

Enabling enterprises to easily document key AI technology characteristics to facilitate subsequent enterprise validation or external regulation

DevOps Expectations

Enabling the enlightened monitoring of AI systems using development-time performance characteristics

A FactSheet Template



Defines what information and metrics are to be collected for Models

By Enterprise, organization, industry, use case, etc.

Questions

All information about the model that can be collected.

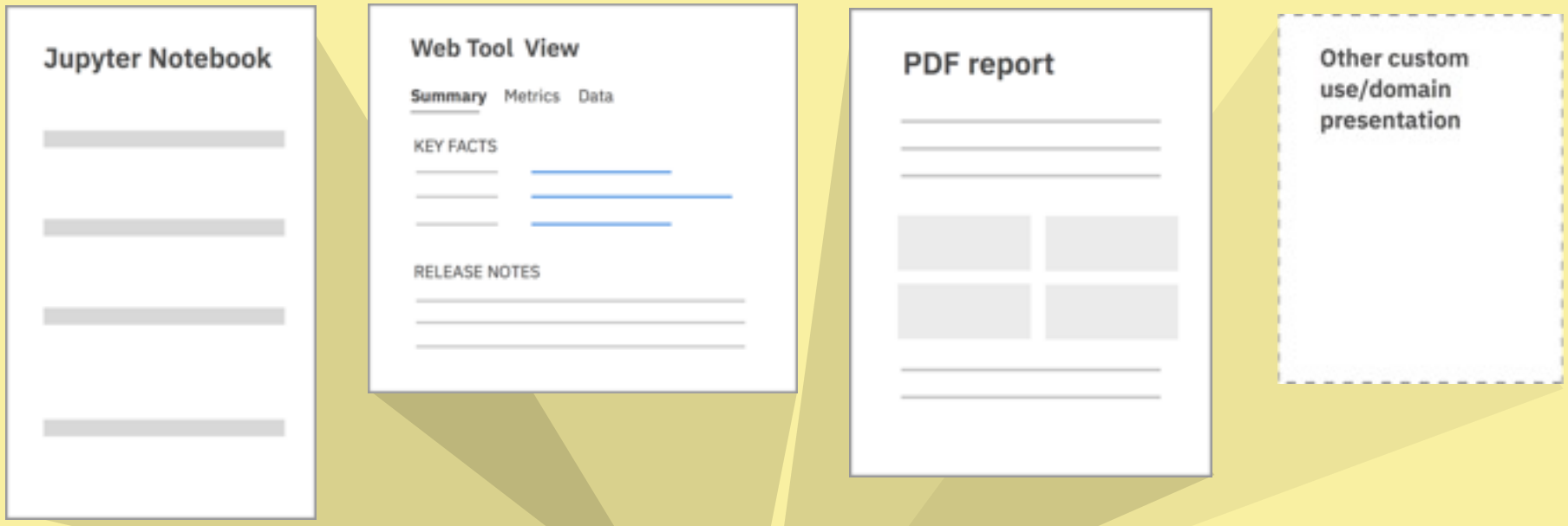
Answers

FactSheet is created from Model Facts based on schema from FactSheet Template.

**(When the model is built, when the model is updated,
When the model is running in production etc.)**

Questions + Answers

FactSheets in various formats have different purposes



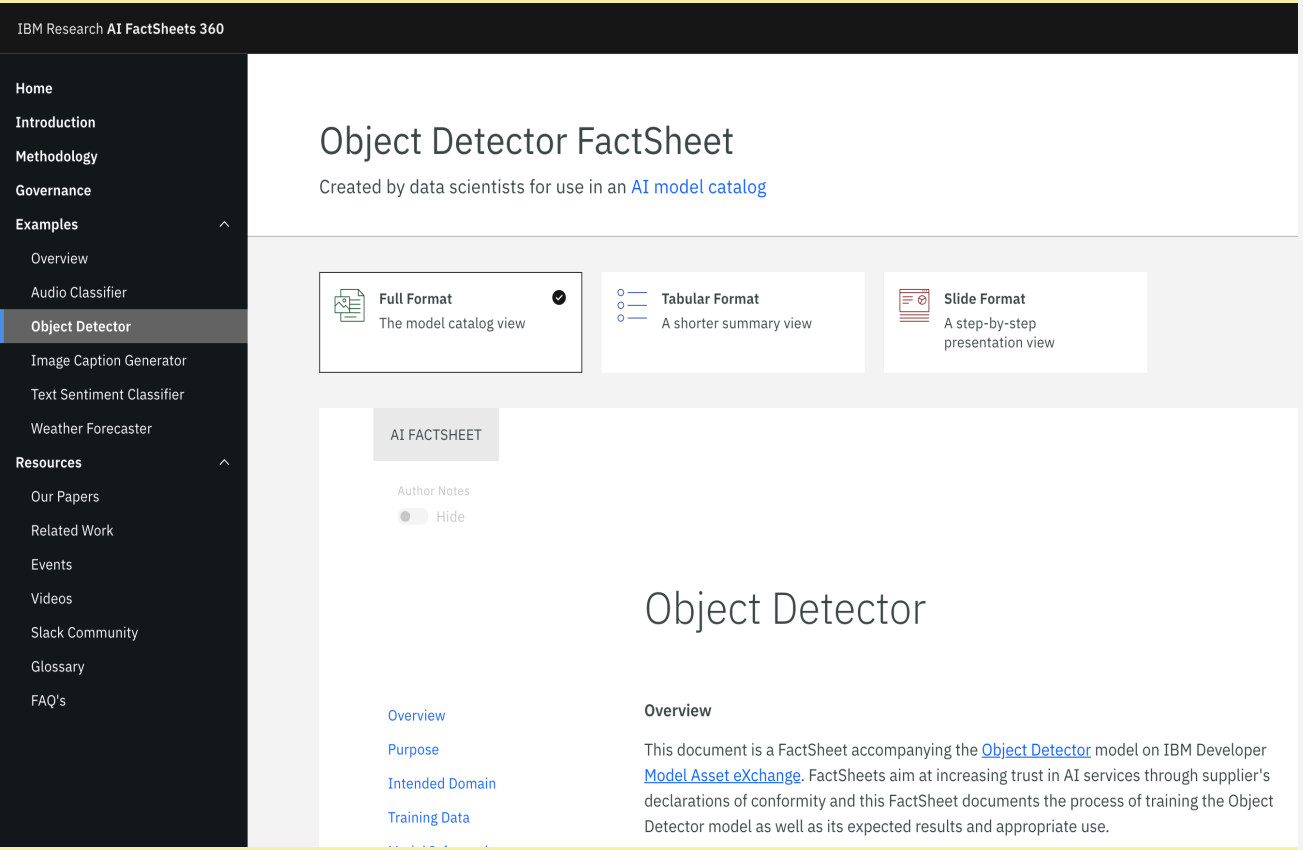
Metrics generated automatically, throughout the lifecycle of the model/service/app

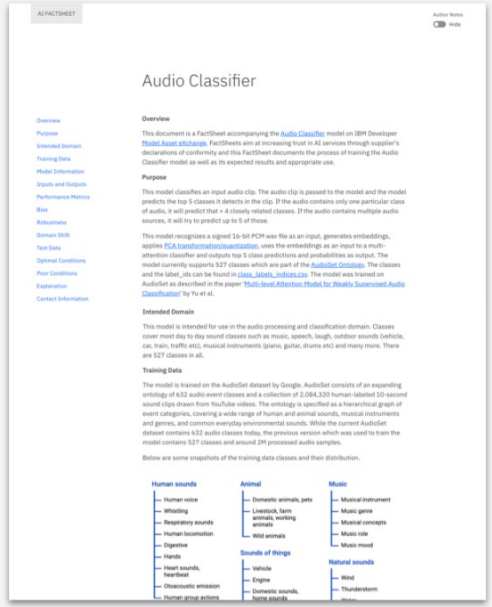
Model Facts

JSON file

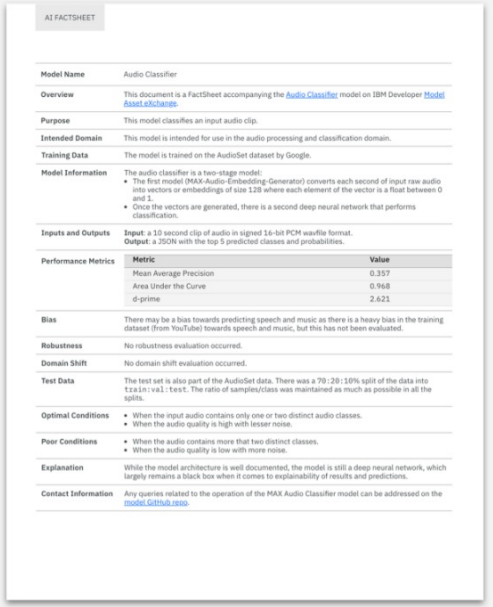
Factsheet Template

https://aifs360.mybluemix.net/

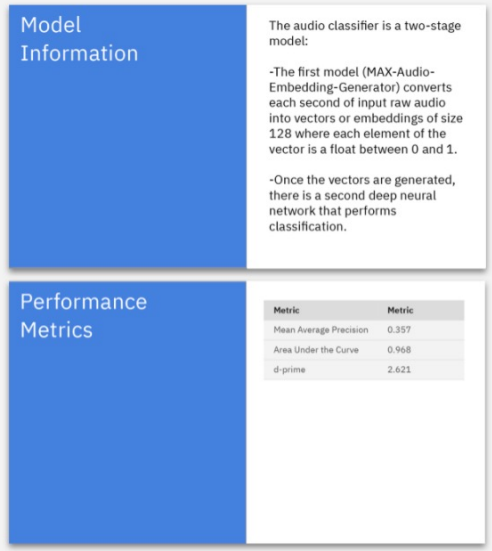




Full Report Format



Tabular Format

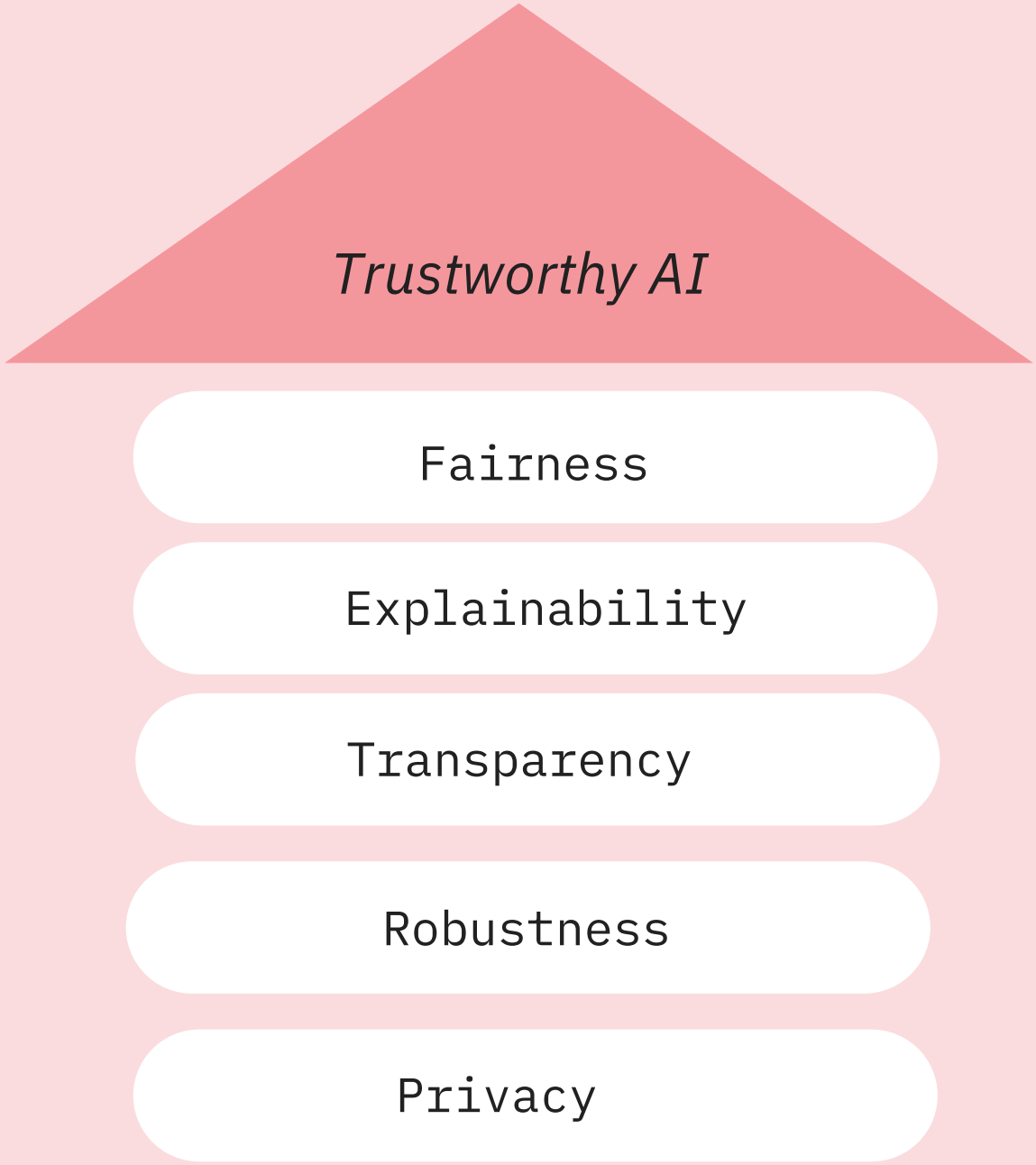


Slide Format

Factsheet Layout

- Overview
- Purpose
- Intended Domain
- Training Data
- Model Information
- Inputs and Outputs
- Performance Metrics
- Bias
- Robustness
- Domain Shift
- Test Data
- Optimal Conditions
- Poor Conditions
- Explanation
- Contact Information

https://aifs360.mybluemix.net/


Trustworthy AI
Fairness
Explainability
Transparency
Robustness
Privacy

Thank You

What is Trustworthy AI for you?