Responsible Data Science

Transparency in Practice

April 3, 2025

Professor Umang Bhatt

Center for Data Science &
Computer Science and Engineering
New York University

This week's reading

The imperative of interpretable machines

As artificial intelligence becomes prevalent in society, a framework is needed to connect interpretability and trust in algorithm-assisted decisions, for a range of stakeholders.

Julia Stoyanovich, Jay J. Van Bavel and Tessa V. West

We are in the midst of a global trend to regulate the use of algorithms, artificial intelligence (AI) and automated decision systems (ADS). As reported by the *One Hundred Year Study on Artificial Intelligence*[1]: "AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human-aware and trustworthy." Major cities, states and national governments are establishing task forces, passing laws and issuing guidelines about responsible development and use of technology, often starting with its use in government itself, where there is, at least in theory, less friction between organizational goals and societal values.

In the United States, New York City has made a public commitment to opening the black box of the government's use of technology: in 2018, an ADS task force was convened, the first of such in the nation, and charged with providing recommendations to New York City's government agencies for how to become transparent and accountable in their use of ADS. In a 2019 report, the task force recommended using ADS where they are beneficial, reduce potential harm and promote fairness, equity, accountability and transparency[2]. Can these principles become policy in the face of the apparent lack of trust in the government's ability to manage AI in the interest of the public? We argue that overcoming this mistrust hinges on our ability to engage in substantive multi-stakeholder conversations around ADS, bringing with it the imperative of interpretability — allowing humans to understand and, if necessary, contest the computational process and its outcomes.

Remarkably little is known about how humans perceive and evaluate algorithms and their outputs, what makes a human trust or mistrust an algorithm[3], and how we can empower humans to exercise agency — to adopt or challenge an algorithmic decision. Consider, for example, scoring and ranking — data-driven algorithms that prioritize entities such as individuals, schools, or products and services. These algorithms may be used to determine credit worthiness, and desirability for college admissions or employment. Scoring and ranking are as ubiquitous and powerful as they are opaque. Despite their importance, members of the public often know little about why one person is ranked higher than another by a résumé screening or a credit scoring tool, how the ranking process is designed and whether its results can be trusted.

As an interdisciplinary team of scientists in computer science and social psychology, we propose a framework that forms connections between interpretability and trust, and develops actionable explanations for a diversity of stakeholders, recognizing their unique perspectives and needs. We focus on three questions (Box 1) about making machines interpretable: (1) what are we explaining, (2) to whom are we explaining and for what purpose, and (3) how do we know that an explanation is effective? By asking — and charting the path towards answering — these questions, we can promote greater trust in algorithms, and improve fairness and efficiency of algorithm-assisted decision making.

Box 1 | Research questions

- **What are we explaining?** Do people trust algorithms more or less than they would trust an individual making the same decisions? What are the perceived trade-offs between data disclosure and the privacy of individuals whose data are being analysed, in the context of interpretability? Which potential sources of bias are most likely to trigger distrust in algorithms? What is the relationship between the perceptions about a dataset's fitness for use and the overall trust in the algorithmic system?
- **To whom are we explaining and why?** How do group identities shape perceptions about algorithms? Do people lose trust in algorithmic decisions when they learn that outcomes produce disparities? Is this only the case when these disparities harm their in-group? Are people more likely to see algorithms as biased if members of their own group were not involved in algorithm construction? What kinds of transparency will promote trust, and when will transparency decrease trust? Do people trust the moral cognition embedded within algorithms? Does this apply to some domains (for example, pragmatic decisions, such as clothes shopping) more than others (for example, moral domains, such as criminal sentencing)? Are certain decisions taboo to delegate to algorithms (for example, religious advice)?
- **Are explanations effective?** Do people understand the label? What kinds of explanations allow individuals to exercise agency: make informed decisions, modify their behaviour in light of the information, or challenge the results of the algorithmic process? Does the nutrition label help create trust? Can the creation of nutrition labels lead programmers to alter the algorithm?

What are we explaining?

Existing legal and regulatory frameworks, such as the US's Fair Credit Reporting Act and the EU's General Data Protection Regulation, differentiate between two kinds of explanations. The first concerns the outcome: what are the results for an individual, a demographic group or the population as a whole? The second concerns the logic behind the decision-making process: what features help an individual or group get a higher score, or, more generally, what are the rules by which the score is computed? Selbst and Barocas[4] argue for an additional kind of an explanation that considers the justification: why are the rules what they are? Much has been written about explaining outcomes[5], so we focus on explaining and justifying the process.

Procedural justice aims to ensure that algorithms are perceived as fair and

Nutritional Labels for Data and Models *

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

Bill Howe
University of Washington
Seattle, WA, USA
billhowe@uw.edu

Abstract

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often used outside of the original context for which they were intended. In response, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a nutritional label, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Nutritional labels are derived automatically or semi-automatically as part of the complex process that gave rise to the data or model they describe, embodying the paradigm of interpretability-by-design. In this paper we further motivate nutritional labels, describe our instantiation of this paradigm for algorithmic rankers, and give a vision for developing nutritional labels that are appropriate for different contexts and stakeholders.

1 Introduction

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often repurposed — used outside of the original context for which they were intended.[1] In response, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a *nutritional label*, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Short of setting up a chemistry lab, the consumer would otherwise

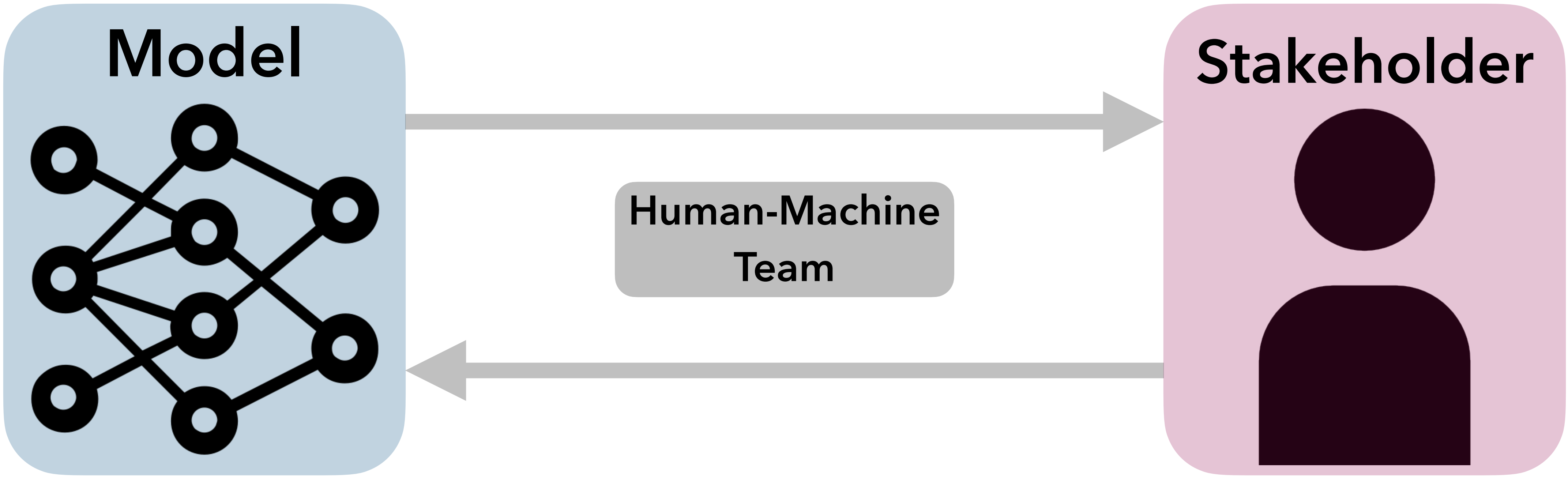
Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

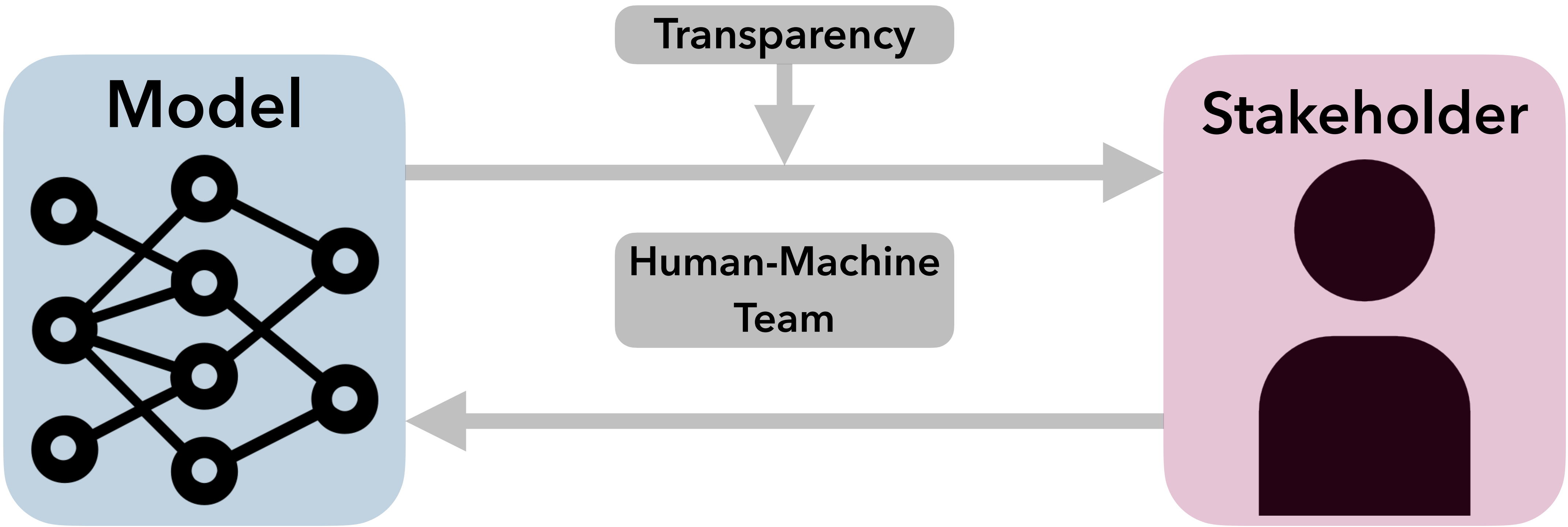
*This work was supported in part by NSF Grants No. 1926250, 1916647, and 1740996.

[1] See Section 1.4 of Salganik's "Bit by Bit" [24] for a discussion of data repurposing in the Digital Age, which he aptly describes as "mixing readymades with custommades."

r/ai

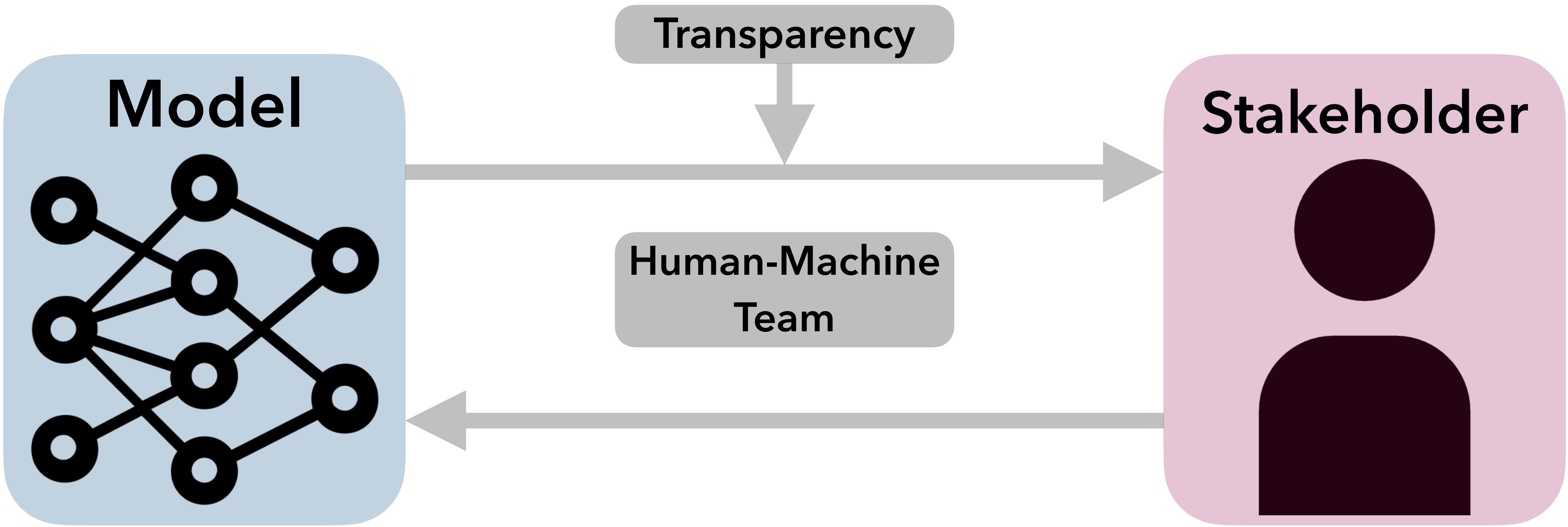
transparency in practice


Model
Human-Machine
Team
Stakeholder



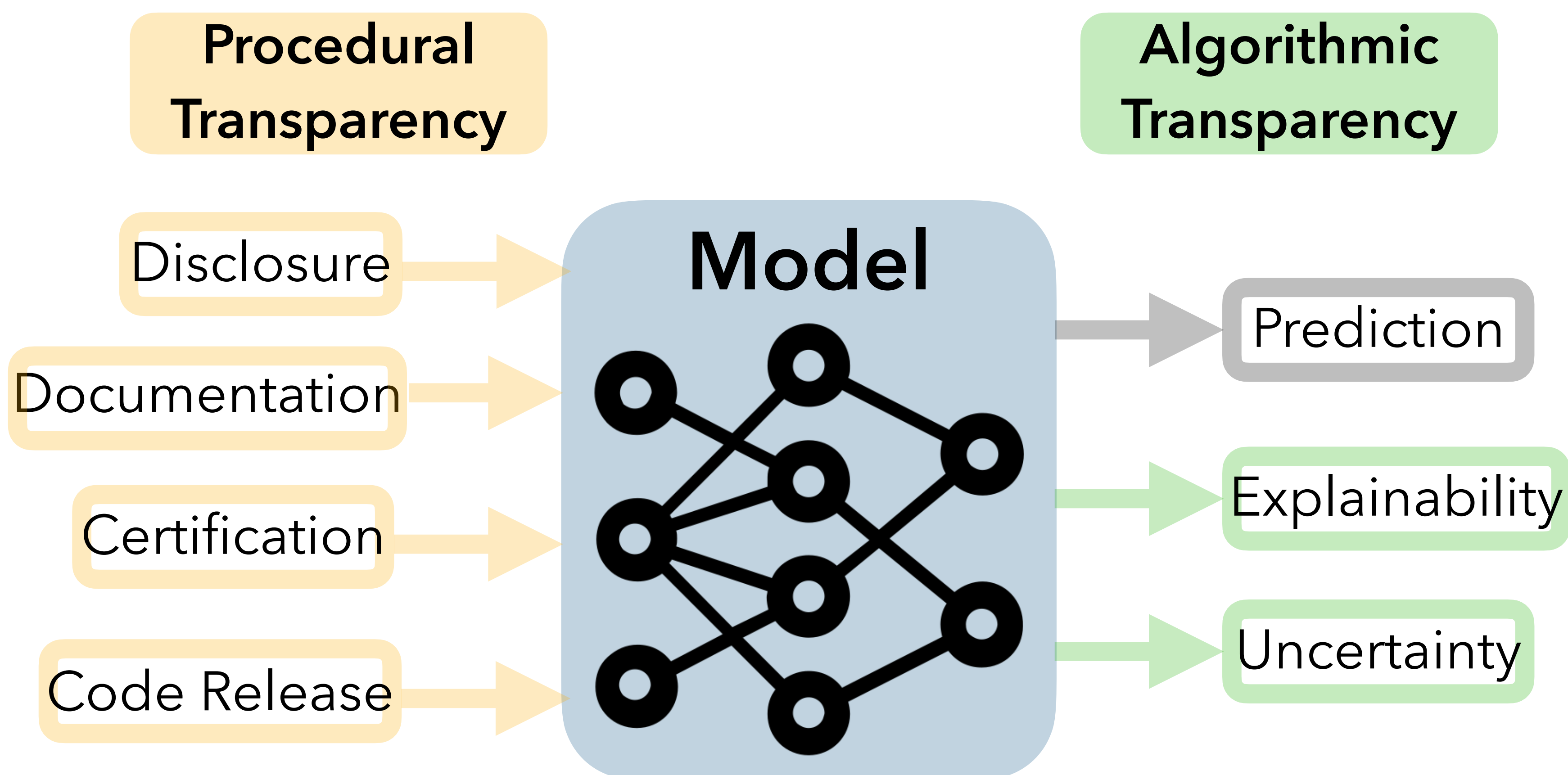
Transparency means providing stakeholders with *relevant* information about how a model works

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment*. ACM FAccT. 2020.

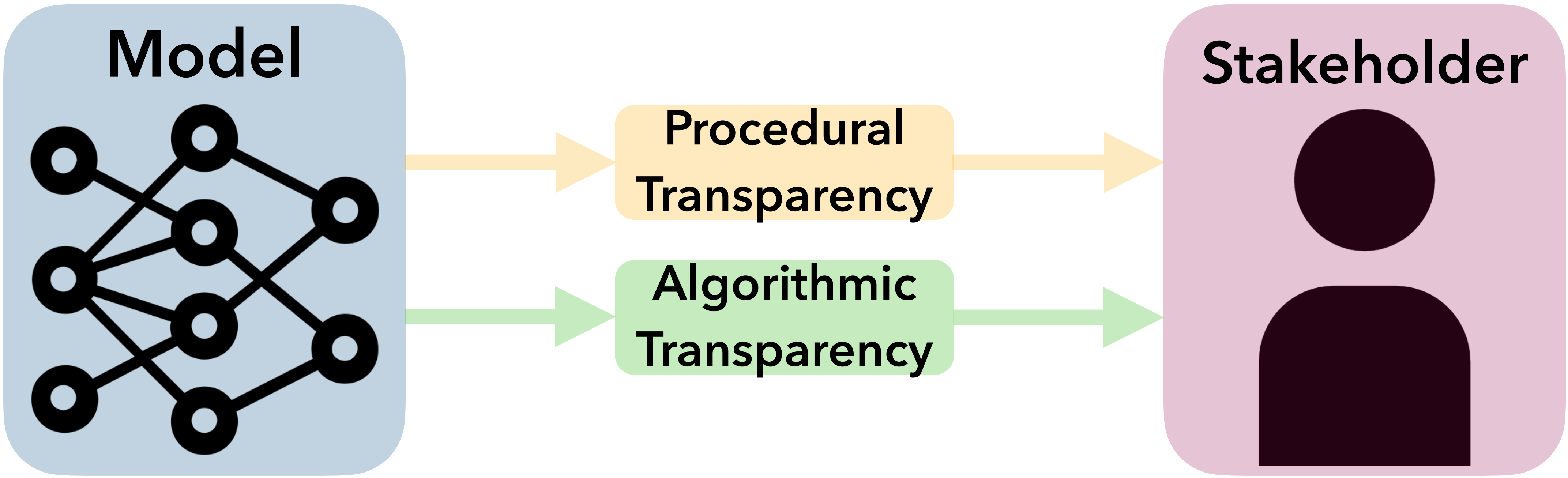


Transparency means providing stakeholders with *relevant* information about how a model works

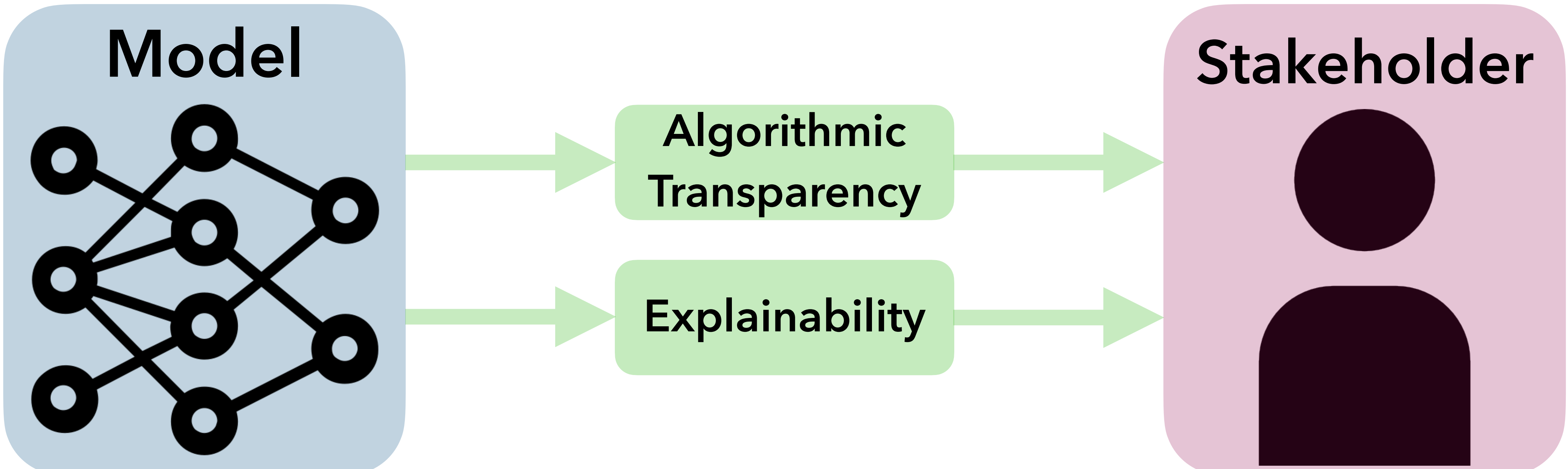
B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.



B, Shams. *Trust in Artificial Intelligence: Clinicians Are Essential.* Chapter 10 in Healthcare Information Technology for Cardiovascular Medicine. 2021.

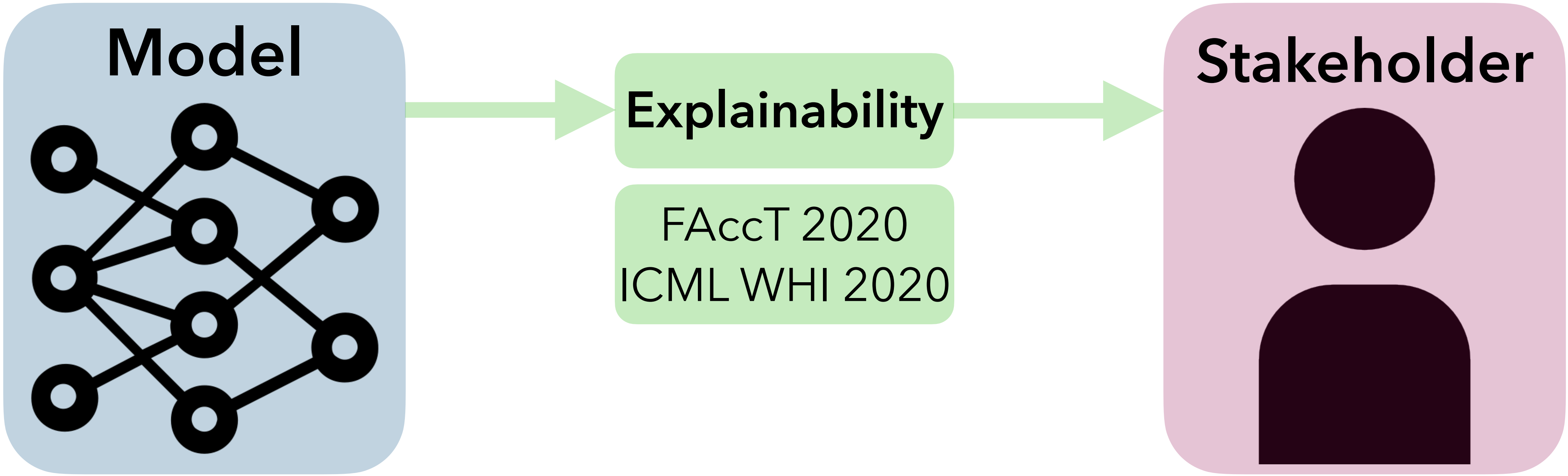


Model
Procedural Transparency
Algorithmic Transparency
Stakeholder

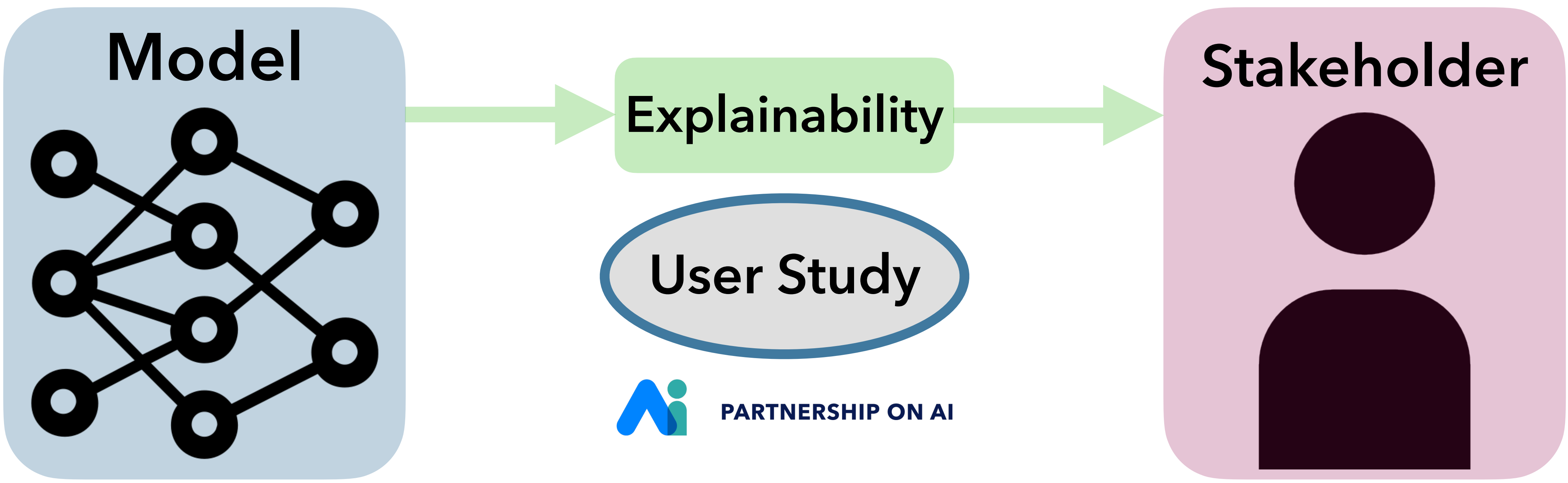


Explainability means providing insight into a model's behavior for specific datapoint(s)

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.


Model
Explainability
FAccT 2020
ICML WHI 2020
Stakeholder

r/ai

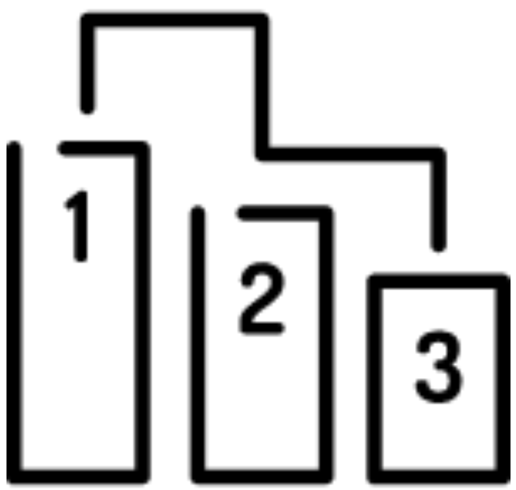


Goal: understand how explainability methods are used in *practice*

Approach: 30min to 2hr *semi-structured* interviews with 50 individuals from 30 organizations

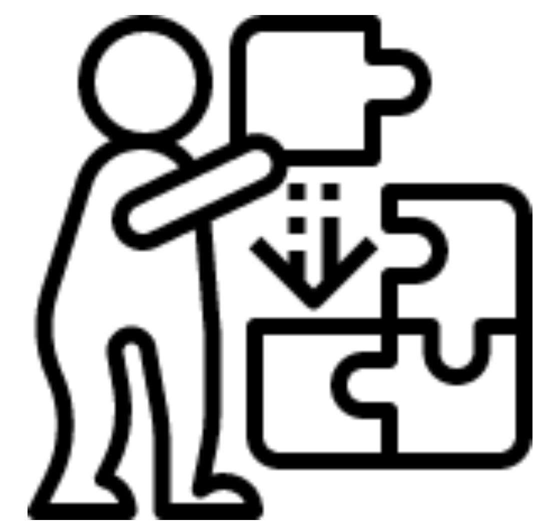
B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

Popular Explanation Styles



Feature Importance

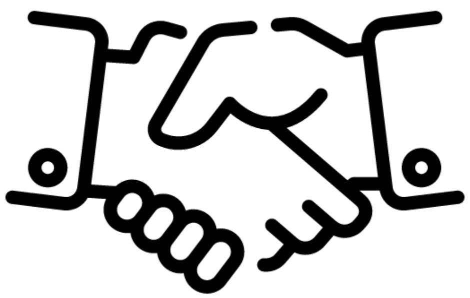
Sample Importance



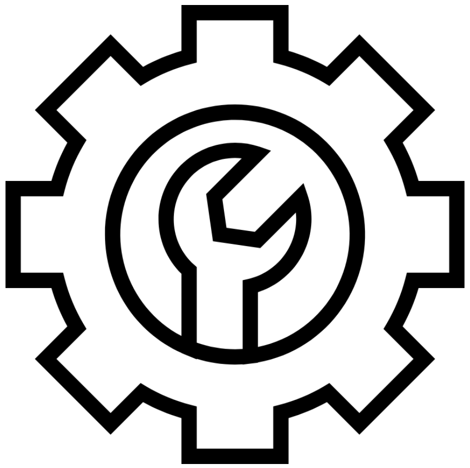
Counterfactuals

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

Common Explanation Stakeholders



Executives



Engineers

End Users

Regulators

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

Findings

1. Explainability is used for **debugging** internally
2. **Goals** of explainability are not clearly defined within organizations
3. Technical **limitations** make explainability hard to deploy in real-time

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

r/ai

Use cases

Domain	Model Purpose	Explainability Technique	Stakeholders	Evaluation Criteria
Finance	Loan Repayment	Feature Importance	Loan Officers	Completeness [34]
Insurance	Risk Assessment	Feature Importance	Risk Analysts	Completeness [34]
Content Moderation	Malicious Reviews	Feature Importance	Content Moderators	Completeness [34]
Finance	Cash Distribution	Feature Importance	ML Engineers	Sensitivity [69]
Facial Recognition	Smile Detection	Feature Importance	ML Engineers	Faithfulness [7]
Content Moderation	Sentiment Analysis	Feature Importance	QA ML Engineers	ℓ_2 norm
Healthcare	Medicare access	Counterfactual Explanations	ML Engineers	normalized ℓ_1 norm
Content Moderation	Object Detection	Adversarial Perturbation	QA ML Engineers	ℓ_2 norm

Table 1: Summary of select deployed local explainability use cases

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

r/ai

Establishing Explainability Goals

1. Identify stakeholders
   Who will consume the explanation?
2. Engage stakeholders
   What purpose with the explanation serve?
3. Devise workflow
   How will the explanation be used in practice?

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

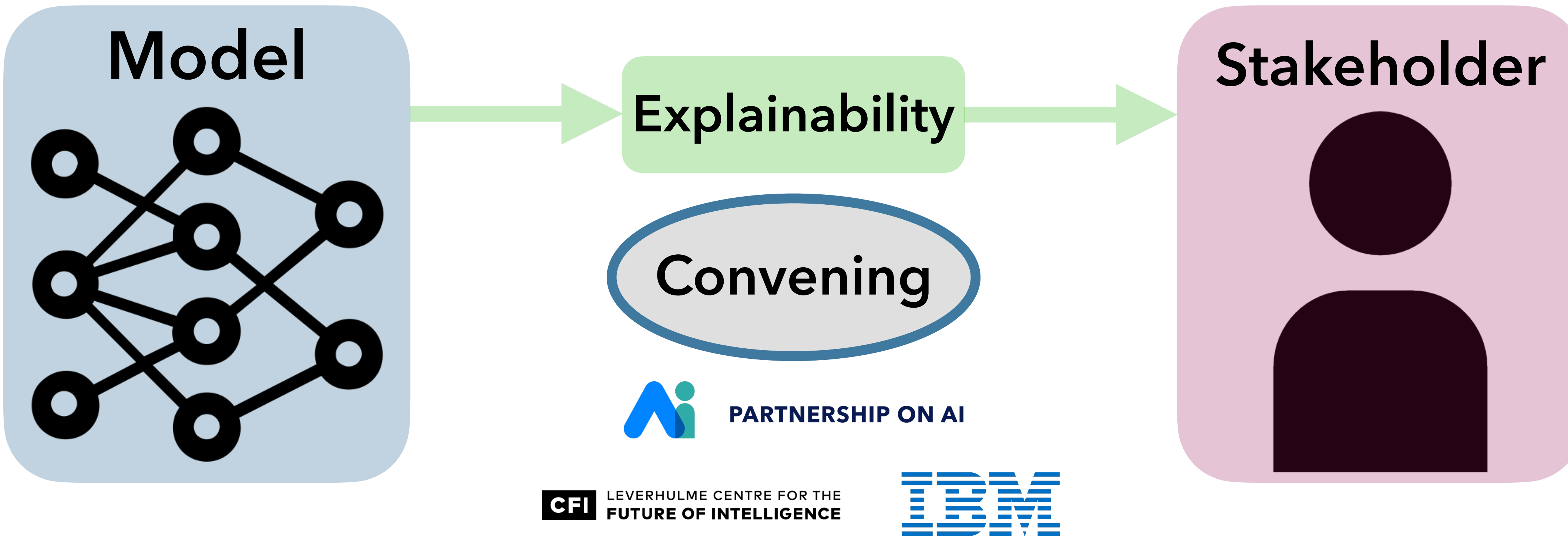
Technical Limitations

1. **Spurious correlations** exposed by feature level explanations

2. **Sample importance** is computationally infeasible to deploy at scale

3. Privacy concerns of **model inversion**

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.



Goal: facilitate an *inter-stakeholder* conversation around explainability

Conclusion: *Community engagement* and *context consideration* are important factors in deploying explainability thoughtfully

B, Andrus, Xiang, Weller. *Machine Learning Explainability for External Stakeholders.* ICML WHI. 2020.

Community Engagement

1. *In which **context** will this explanation be used?*

2. *How should the explanation be **evaluated**? Both quantitatively and qualitatively…*

3. *Can we prevent data misuse and preferential treatment by involving **affected groups** in the development process?*

4. *Can we **educate** stakeholders regarding the functionalities and limitations of explainable machine learning?*

B, Andrus, Xiang, Weller. *Machine Learning Explainability for External Stakeholders.* ICML WHI. 2020.

r/ai

Deploying Explainability

1. *How does **uncertainty** in the model's predictions and explanation technique affect the resulting explanations?*

2. *How can stakeholders **interact** with the resulting explanations?*

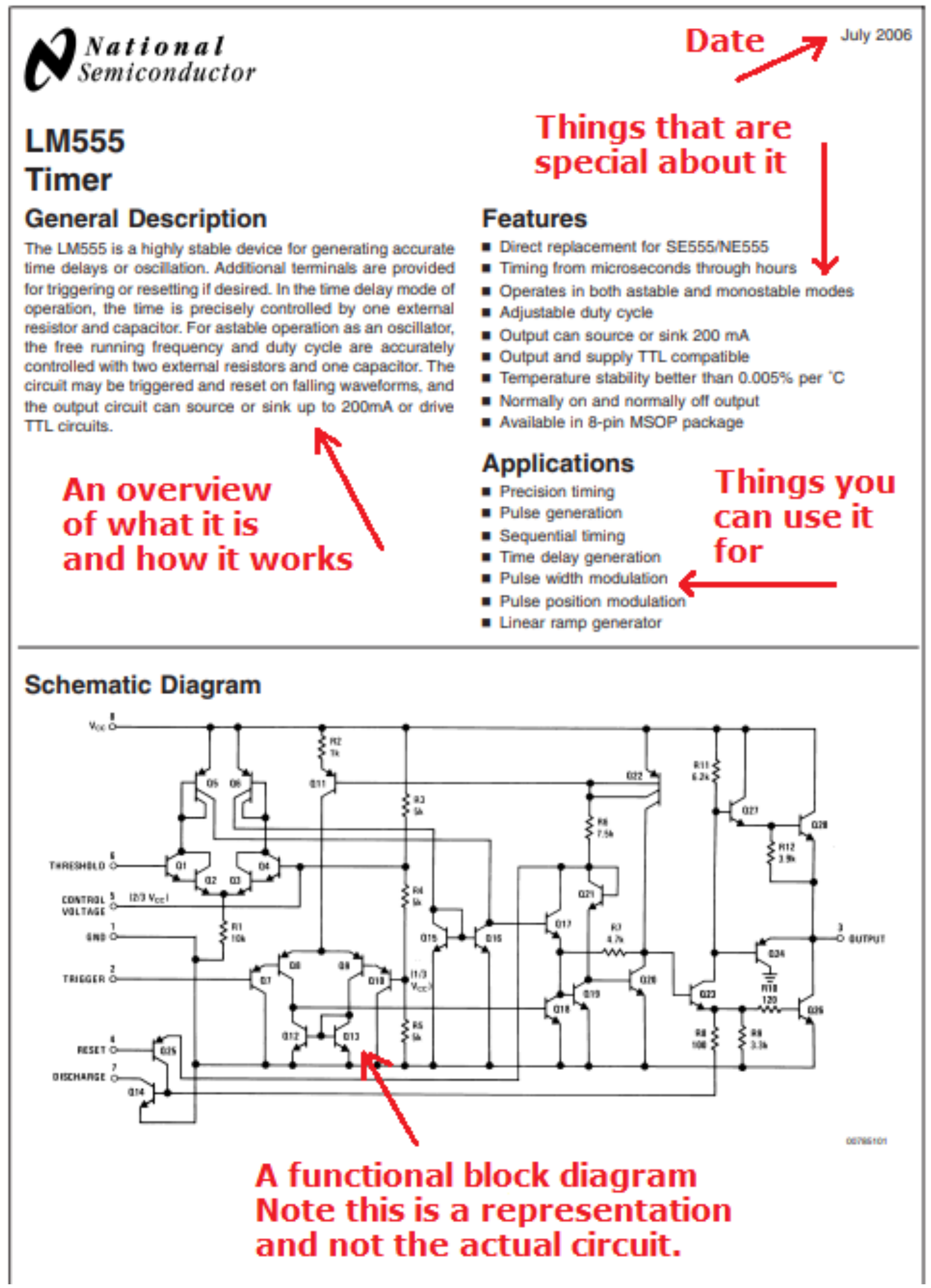
3. *How, if at all, will stakeholder **behavior** change as a result of the explanation shown?*

4. *Over **time**, how will the model and explanations adapt to changes in stakeholder behavior?*

B, Andrus, Xiang, Weller. *Machine Learning Explainability for External Stakeholders.* ICML WHI. 2020.

case for documentation

Datasheets for Electronics


National Semiconductor

Date

July 2006

LM555 Timer

Things that are special about it

General Description

The LM555 is a highly stable device for generating accurate time delays or oscillation. Additional terminals are provided for triggering or resetting if desired. In the time delay mode of operation, the time is precisely controlled by one external resistor and capacitor. For astable operation as an oscillator, the free running frequency and duty cycle are accurately controlled with two external resistors and one capacitor. The circuit may be triggered and reset on falling waveforms, and the output circuit can source or sink up to 200mA or drive TTL circuits.

An overview of what it is and how it works

Features

- Direct replacement for SE555/NE555
- Timing from microseconds through hours
- Operates in both astable and monostable modes
- Adjustable duty cycle
- Output can source or sink 200 mA
- Output and supply TTL compatible
- Temperature stability better than 0.005% per °C
- Normally on and normally off output
- Available in 8-pin MSOP package

Applications

- Precision timing
- Pulse generation
- Sequential timing
- Time delay generation
- Pulse width modulation
- Pulse position modulation
- Linear ramp generator

Things you can use it for

Schematic Diagram

A functional block diagram
Note this is a representation and not the actual circuit.

Datasheets for Electronics

onsemi

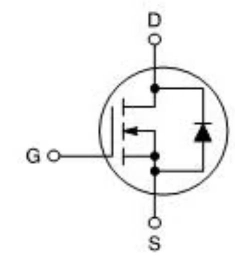
DATA SHEET
www.onsemi.com

MOSFET - SiC Power, Single N-Channel, TO247-3L

650 V, 57 mΩ, 38 A

NVHL075N065SC1

$V_{(BR)DSS}$	$R_{DS(ON)}$ MAX	I_D MAX
650 V	85 mΩ @ 18 V	38 A

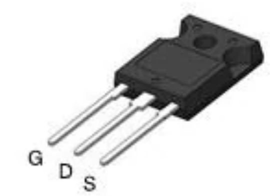


Features

- Typ. $R_{DS(on)}$ = 57 mΩ @ V_{GS} = 18 V
  Typ. $R_{DS(on)}$ = 75 mΩ @ V_{GS} = 15 V
- Ultra Low Gate Charge ($Q_{G(tot)}$ = 61 nC)
- Low Output Capacitance (C_{oss} = 107 pF)
- 100% Avalanche Tested
- AEC−Q101 Qualified and PPAP Capable
- This Device is Pb−Free and is RoHS Compliant

Typical Applications

- Automotive On Board Charger
- Automotive DC/DC Converter for EV/HEV



TO−247 Long Leads
CASE 340CX

MAXIMUM RATINGS (T_J = 25°C unless otherwise noted)

Parameter			Symbol	Value	Unit
Drain−to−Source Voltage			V_{DSS}	650	V
Gate−to−Source Voltage			V_{GS}	−8/+22	V
Recommended Operation Values of Gate−to−Source Voltage		T_C < 175°C	V_{GSop}	−5/+18	V
Continuous Drain Current (Note 1)	Steady State	T_C = 25°C	I_D	38	A
Power Dissipation (Note 1)			P_D	148	W
Continuous Drain Current (Note 1)	Steady State	T_C = 100°C	I_D	26	A
Power Dissipation (Note 1)			P_D	74	W

MARKING DIAGRAM

https://www.mouser.com/datasheet/lrg/308/1/NVHL075N065SC1_D-3326430.jpg

Datasheets for Datasets

nvironments **Labeled Faces in the Wild**

Property	**Value**
Database Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Demographic Characteristic	**Value**
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

Table 2. Demographic characteristics of the LFW dataset as measured by Han, Hu, and Anil K. Jain. *Age, gender and race estimation from unconstrained face images.* Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

- Document the *dataset* properties
- Disclose (1) motivation for dataset creation, (2) dataset composition, (3) data collection process, (4) data preprocessing, (5) dataset distribution, (6) dataset maintenance, (7) legal/ethical considerations
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford. *Datasheets for Datasets.* CACM 2021.

Datasheets for Datasets

DATASET OVERVIEW

BASICS: CONTACT, DISTRIBUTION, ACCESS

1. Dataset name
2. Dataset version number or date
3. Dataset owner/manager contact information, including name and email
4. Who can access this dataset (e.g., team only, internal to the company, external to the company)?
5. How can the dataset be accessed?

DATASET CONTENTS

6. What are the contents of this dataset? Please include enough detail that someone unfamiliar with the dataset who might want to use it can understand what is in the dataset.

   Specifically, be sure to include:

   - What does each item/data point represent (e.g., a document, a photo, a person, a country)?
   - How many items are in the dataset?
   - What data is available about each item (e.g., if the item is a person, available data might include age, gender, device usage, etc.)? Is it raw data (e.g., unprocessed text or images) or features (variables)?
   - *For static datasets:* What timeframe does the dataset cover (e.g., tweets from January 2010–December 2020)?

INTENDED & INAPPROPRIATE USES

7. What are the intended purposes for this dataset?
8. What are some tasks/purposes that this dataset is not appropriate for?

- Encourage data documentation but hard to operationalize
- http://aka.ms/datadoc

r/ai

Model Cards for Model Reporting

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

- Document the *model* properties
- Disclose (1) model details, (2) intended use, (3) factors, (4) metrics, (5) evaluation data, (6) training data, (7) qualitative analyses, (8) ethical considerations
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. *Model Cards for Model Reporting.* ACM FAccT 2019.

r/ai

Model Cards for Model Reporting

DATA FOCUSED

- Data Sheets
- Data Statements
- Data Nutrition Labels
- Data Cards for NLP
- Dataset Development Lifecycle Documentation Framework
- Data Cards

MODELS & METHODS FOCUSED

- Model Cards
- Value Cards
- Method Cards
- Consumer Labels for Models

SYSTEMS FOCUSED

- System Cards
- FactSheets
- ABOUT ML

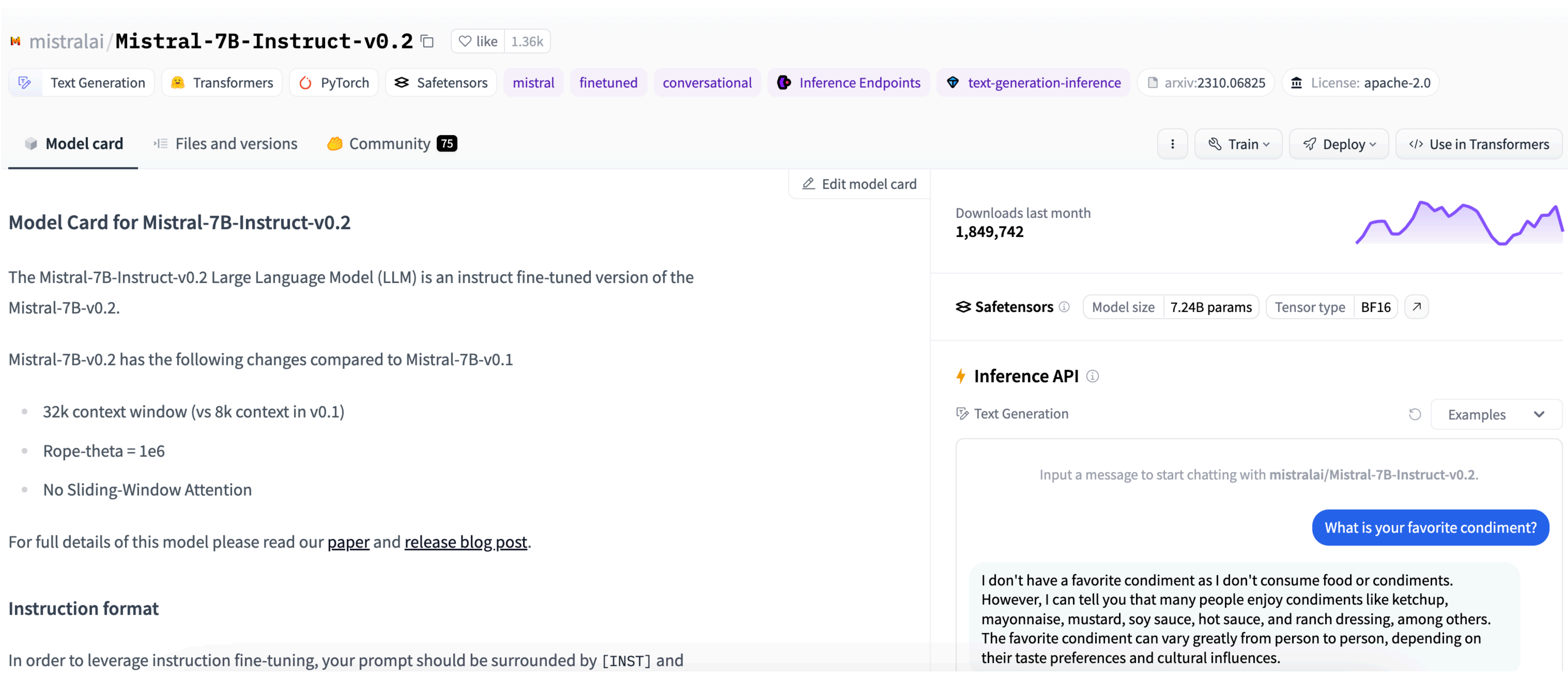
SAMPLE OF POTENTIAL AUDIENCES

- ML Engineers
- Model Developers/Reviewers
- Students
- Policymakers
- Ethicists
- Data Scientists/Business Analysts
- Impacted Individuals

- Encourage model card generation as part of development best practices
- https://huggingface.co/blog/model-cards

r/ai

Model Cards for Model Reporting


mistralai/**Mistral-7B-Instruct-v0.2** like 1.36k

Text Generation | Transformers | PyTorch | Safetensors | mistral | finetuned | conversational | Inference Endpoints | text-generation-inference | arxiv:2310.06825 | License: apache-2.0

Model card | Files and versions | Community 75 | Train | Deploy | Use in Transformers

Edit model card

Model Card for Mistral-7B-Instruct-v0.2

The Mistral-7B-Instruct-v0.2 Large Language Model (LLM) is an instruct fine-tuned version of the Mistral-7B-v0.2.

Mistral-7B-v0.2 has the following changes compared to Mistral-7B-v0.1

- 32k context window (vs 8k context in v0.1)
- Rope-theta = 1e6
- No Sliding-Window Attention

For full details of this model please read our paper and release blog post.

Instruction format

In order to leverage instruction fine-tuning, your prompt should be surrounded by `[INST]` and

Downloads last month
1,849,742

Safetensors | Model size 7.24B params | Tensor type BF16

Inference API

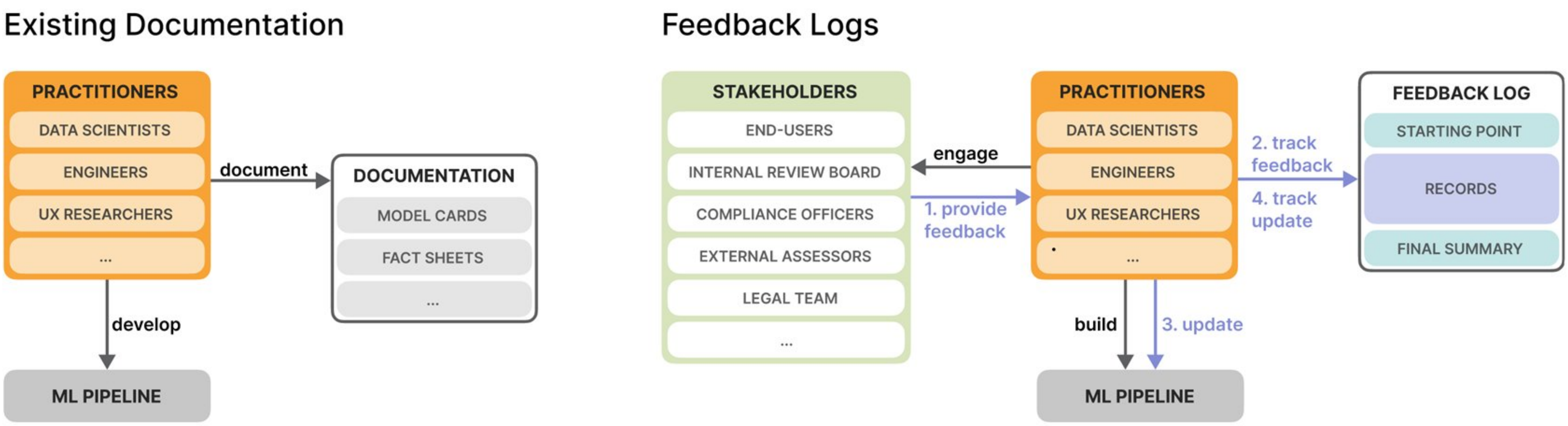
Text Generation | Examples

Input a message to start chatting with **mistralai/Mistral-7B-Instruct-v0.2**.

What is your favorite condiment?

I don't have a favorite condiment as I don't consume food or condiments. However, I can tell you that many people enjoy condiments like ketchup, mayonnaise, mustard, soy sauce, hot sauce, and ranch dressing, among others. The favorite condiment can vary greatly from person to person, depending on their taste preferences and cultural influences.

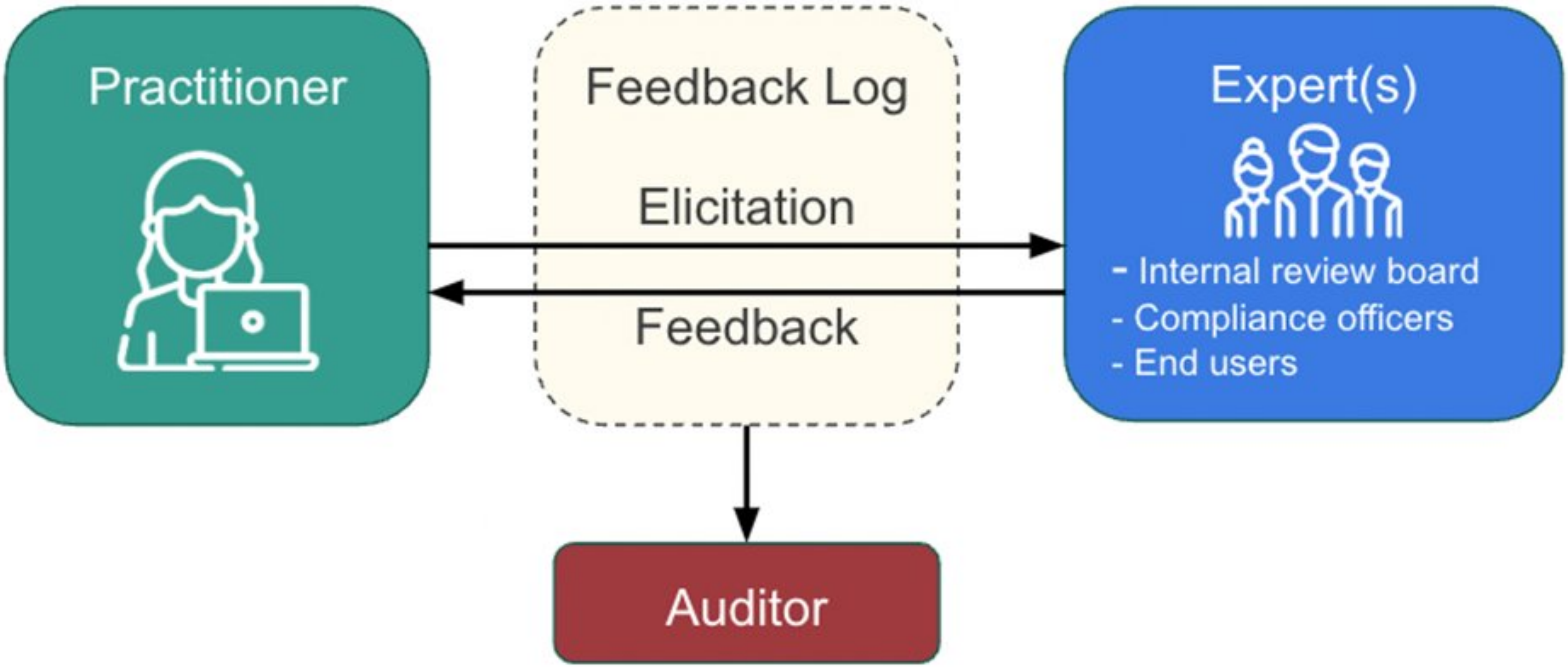
Feedback Logs



Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback into Machine Learning Pipelines*. ACM EAAMO. 2023.

r/ai

Feedback Logs



Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback into Machine Learning Pipelines*. ACM EAAMO. 2023.

Feedback Logs

Starting Point
Data: *Description of the dataset(s) used to train/test/validate the model.*
Models: *Description of the model(s) used and any existing design decisions.*
Metrics: *Description of the metrics used to evaluate the model(s) and their performance.*

Record 1

Elicitation
Who and why? *Which stakeholder(s) are being consulted? What prompted the request for feedback? e.g. legal requirements, poor performance on metrics.*
How? *How is the relevant information presented to them? e.g. model metrics, predictions, prototype.*

Feedback
What? *What insights have been provided by the stakeholder(s)?*

Incorporation

Which?	Where?	When?	Why?	Effect?
Which updates are considered?	*Where in the pipeline did the update occur?*	*When in the pipeline did the update occur?*	*Why has this update been selected?*	*What effect(s) did the update have on the metrics?*
Update 1	x	x	x	x
Update 2	x	x	x	x
...	...	...	...	...

Summary
What? *Summary of the update(s) chosen and their effect(s) on the metric(s).*

Record 2

...

Final Summary
Data: *Description of the dataset(s) used to train/test/validate the model after all updates have been applied.*
Model: *Description of model(s) used and any design changes resulting from the updates.*
Metric performance: *Description of the metrics to evaluate the model(s) and their performance after the above updates.*

Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback into Machine Learning Pipelines*. ACM EAAMO. 2023.

Feedback Logs

Image Recognition `FeedbackLog`

Starting Point
Data: Imagenet1K for training and validation datasets, consisting of 1000 image classes.
Model: Convolutional Neural Network (ResNet50).
Metrics: None defined yet.

Record 1: Elicitation
Who and why? Hypothetical external assessor vested in the model. Require regulatory approval to use image recognition model in practice.
How? Asked for minimum benchmark performance, similar to the 80 percent disparate impact rule.

Feedback
What? Received a dataset containing adversarial examples of automotive vehicles, along with a minimum accuracy required for this dataset to test the model's robustness.

Incorporation

Which?	Where?	When?	Why?	Effect?
Imagenet-A with relevant automotive classes	Dataset	Pre-Training	Tests model robustness	Testing dataset for model
Minimum accuracy $> 50\%$	Ecosystem & Metrics	Training	Required for regulatory approval	Benchmark when testing model

Summary
What? Dataset update: provided new dataset to test the model's robustness when recognising automotive vehicles. Ecosystem update as part of metrics: added requirement that model should achieve $> 50\%$ accuracy (robustness) on test dataset.

Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback into Machine Learning Pipelines*. ACM EAAMO. 2023.

Feedback Logs

Record 2: Elicitation
Who and why? Hypothetical compliance team. Need to ensure model meets external requirements set by industry regulators, as well as internal company policies.
How? Presented with current performance on testing dataset recommended by regulator, along with example predictions.

Feedback
What? Current robustness (34%) isn't sufficient to meet requirements. In addition, the model is overconfident in its predictions which may cause serious accidents that are unacceptable under company policy.

Incorporation

Which?	Where?	When?	Why?	Effect?
ResNet-101	Parameter space	Before training	Identify complex features	Robustness 39%
MEAL V2	Loss function	During training	Soften labels	Robustness: 47%
CutMix	Dataset	Before training	Background invariance	Robustness: 48%

Summary
What? Used ResNet-101 model with CutMix for data augmentation, since when both updates are used the robustness is 55%, which exceeds the minimum requirement of 50%.

Final Summary
Data: Imagenet1K augmented with CutMix for training, Imagenet-A with relevant automotive classes for testing.
Model: Convolutional Neural Network (ResNet-101).
Metric performance: 55% robustness on Imagenet-A testing dataset.

Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback into Machine Learning Pipelines*. ACM EAAMO. 2023.

r/ai

EU AI Act

EU AI Act

Article 11

Technical documentation

The technical documentation of a high-risk AI system shall be drawn up before that system is placed on the market or put into service and shall be kept up-to date.

Article 12

Record-keeping

High-risk AI systems shall technically allow for the automatic recording of events ('logs') over the duration of the lifetime of the system.

https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

Article 13

Transparency and provision of information to deployers

High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable deployers to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Chapter 3 of this Title.

Article 14

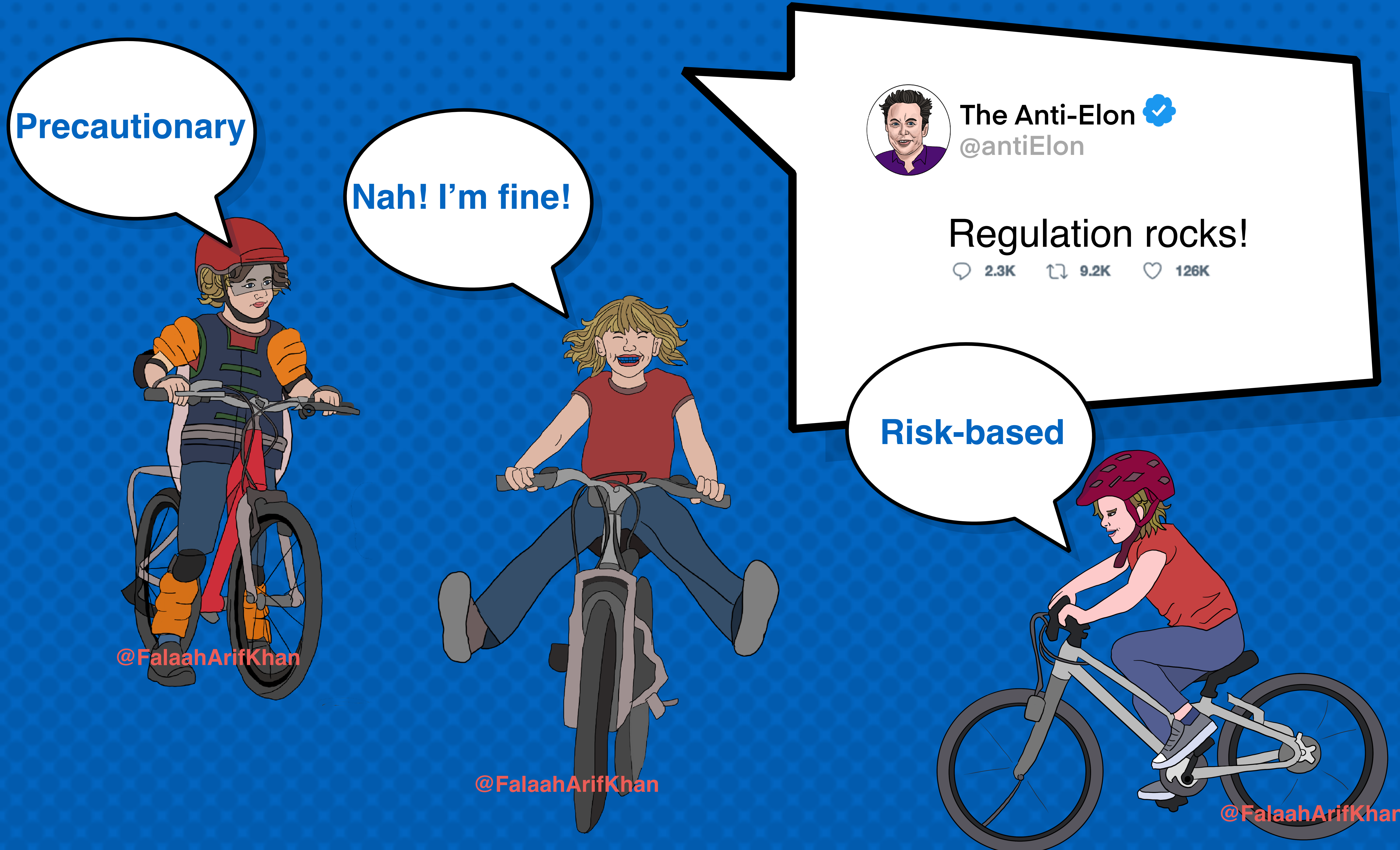
Human oversight

High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.

from data to impacts: algorithmic impact statements

r/ai

Regulating ADS?



Setting the stage: “Big Data Policing”

“Despite its growing popularity, predictive policing is in its relative infancy and is still mostly hype. Current prediction is akin to early weather forecasting, and, like Big Data approaches in other sectors, mixed evidence exists about its effectiveness.

Cities such as Los Angeles, Atlanta, Santa Cruz, and Seattle have enlisted the predictive policing software company PredPol to predict where property crimes will occur. Santa Cruz reportedly “saw burglaries drop by 11% and robberies by 27% in the first year of using [PredPol’s] software.” Similarly, Chicago’s Strategic Subject List—or “heat list”—of people most likely to be involved in a shooting had, as of mid-2016, predicted more than 70% of the people shot in the city, according to the police.

But two rigorous academic evaluations of predictive policing experiments, one in Chicago and another in Shreveport, have shown no benefit over traditional policing. **A great deal more study is required to measure both predictive policing’s benefits and its downsides**. “

what are the potential benefits?

what are the potential downsides?

[Andrew Selbst, 2017]

r/ai

How to regulate "Big Data Policing"

"While policing is just one of many aspects of society being upended by machine learning, and potentially exacerbating disparate impact in a hidden way as a result, it is a particularly useful case study because of how little our legal system is set up to regulate it."

The Fourth Amendment: *The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.*

"[…] the Fourth Amendment's reasonable suspicion requirement is inherently a "small data doctrine," rendering it impotent in even its primary uses when it comes to data mining."

new legal strategies are needed

[Andrew Selbst, 2017]

r/ai

How to regulate “Big Data Policing”

“ Regarding predictive policing specifically, **society lacks basic knowledge and transparency about both the technology’s efficacy and its effects on vulnerable populations**. Thus, this Article proposes a regulatory solution designed to fill this knowledge gap—to **make the police do their homework** and show it to the public before buying or building these technologies.”

Main contribution: Algorithmic Impact Statements (AISs)

“Impact statements are designed to **force consideration of the problem at an early stage**, and to document the process so that the public can learn what is at stake, **perhaps as a precursor to further regulation**. The primary problem is that no one, including the police using the technology, yet knows what the results of its use actually are.”

[Andrew Selbst, 2017]

Algorithmic Impact Statements (AISs)

- Modeled on the Environmental Impact Statements (EISs) of the 1969 National Environmental Policy Act (NEPA)
- GDPR requires “data protection impact assessments (DPIAs) whenever data processing “is likely to result in a high risk to the rights and freedoms of natural persons”
- Privacy impact statements (PIAs) are used to assess the risks of using personally identifiable information by IT systems

The gist:

- Explore and evaluate all reasonable alternatives
- Include the alternative of “No Action”
- Include appropriate mitigation measures
- Provide opportunities for public comment

[Andrew Selbst, 2017]

r/ai

Canadian ADS directive

Government of Canada | Gouvernement du Canada

Home → How government works → Policies, directives, standards and guidelines

Directive on Automated Decision-Making

The Government of Canada is increasingly looking to utilize artificial intelligence to make, or assist in making, administrative decisions to improve service delivery. The Government is committed to doing so in a manner that is compatible with core administrative law principles such as transparency, accountability, legality, and procedural fairness. Understanding that this technology is changing rapidly, this Directive will continue to evolve to ensure that it remains relevant.

Date modified: 2019-02-05

- Took effect on **April 1, 2019**, compliance by **April 1, 2020**
- Applies to any ADS developed or procured after April 1, 2020
- Reviewed automatically every 6 months

Definitions

Appendix A: Definitions

- **Administrative Decision** Any decision that is made by an authorized official of an institution as identified in section 9 of this Directive pursuant to powers conferred by an Act of Parliament or an order made pursuant to a prerogative of the Crown that affects legal rights, privileges or interests.
- **Algorithmic Impact Assessment** A framework to help institutions better understand and reduce the risks associated with Automated Decision Systems and to provide the appropriate governance, oversight and reporting/audit requirements that best match the type of application being designed.
- **Automated Decision System** Includes any technology that either assists or replaces the judgement of human decision-makers. These systems draw from fields like statistics, linguistics, and computer science, and use techniques such as rules-based systems, regression, predictive analytics, machine learning, deep learning, and neural nets.

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Objectives

Section 4: Objectives and Expected Results

- **4.1** The objective of this Directive is to ensure that Automated Decision Systems are deployed in a manner that **reduces risks** to Canadians and federal institutions, and **leads to more efficient, accurate, consistent, and interpretable decisions** made pursuant to Canadian law.
- **4.2** The expected results of this Directive are as follows:
  - Decisions made by federal government departments are data-driven, responsible, and complies with procedural fairness and due process requirements.
  - Impacts of algorithms on administrative decisions are assessed and negative outcomes are reduced, when encountered.
  - Data and information on the use of Automated Decision Systems in federal institutions are made available to the public, when appropriate.

Requirements

Section 6.1: Algorithmic Impact Assessment (excerpt)

- **6.1.1 Completing** an Algorithmic Impact Assessment **prior to the production** of any Automated Decision System.
- **6.1.2** …
- **6.1.3 Updating** the Algorithmic Impact Assessment when system functionality or the scope of the Automated Decision System changes.
- **6.1.4 Releasing the final results of Algorithmic Impact Assessments** in an accessible format via Government of Canada websites and any other services designated by the Treasury Board of Canada Secretariat pursuant to the Directive on Open Government.

Requirements

Section 6.2: Transparency

- providing notice **before** decisions
- providing explanations **after** decisions
- access to components
- release of source code, unless it's classified Secret, Top Secret or Protected C

Impact Assessment Levels

Decisions classified w.r.t. impact on:

- the rights of individuals or communities,
- the health or well-being of individuals or communities,
- the economic interests of individuals, entities, or communities,
- the ongoing sustainability of an ecosystem.

Level I: no impact: impacts are reversible and brief

Level II: moderate: impacts are likely reversible and short-term

Level III: high: impacts are difficult to reversible and ongoing

Level IV: very high: impacts are irreversible and perpetual

higher impact levels lead to more stringent requirements

so what’s algorithmic transparency?

r/ai

Point 1

algorithmic transparency is not synonymous with releasing the source code

publishing source code helps, but it is sometimes unnecessary and often insufficient

Point 2

algorithmic transparency requires data transparency

data is used in training, validation, deployment

validity, accuracy, applicability can only be understood in the data context

data transparency is necessary for all ADS, not only for ML-based systems

Point 3

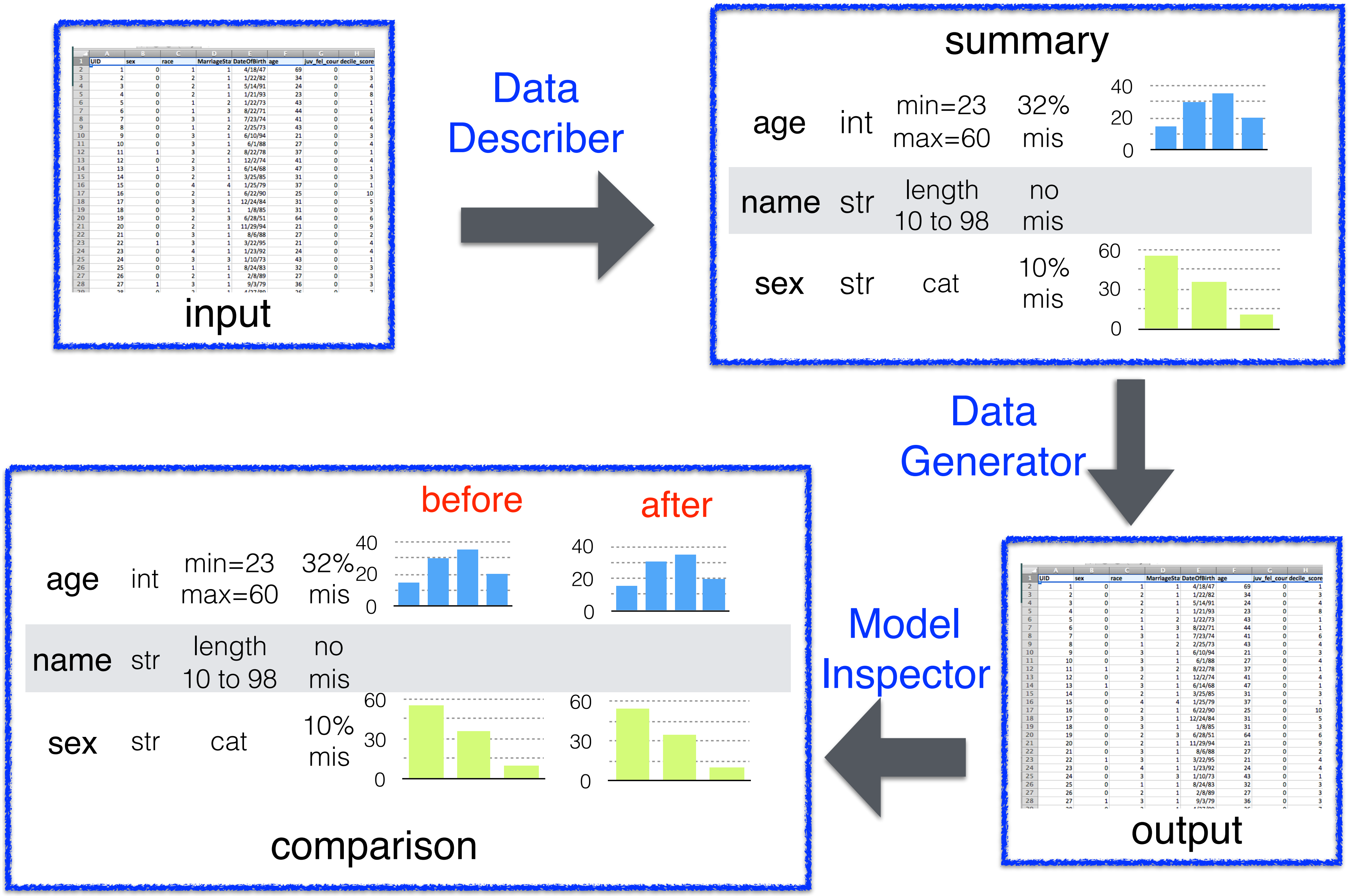
data transparency is not synonymous with making all data public

release data whenever possible;

also release:

data selection, collection and pre-processing methodologies; data provenance and quality information; known sources of bias; privacy-preserving statistical summaries of the data

Data Synthesizer



[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

Point 4

actionable transparency requires
interpretability

explain assumptions and effects, not details of operation

engage the public - technical and non-technical

“Nutritional labels” for data and models

Recipe

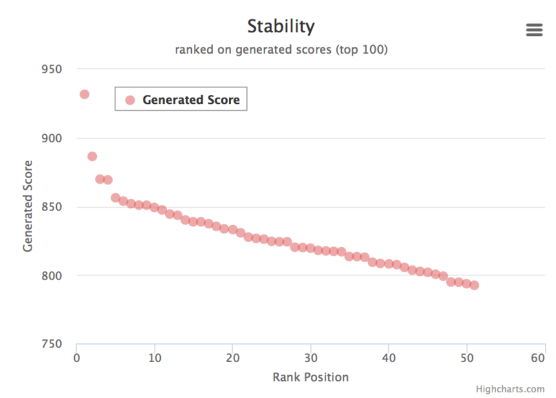
Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
Faculty	122	52.5	45
GRE	800.0	796.3	771.9

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
Faculty	122	32.0	14
GRE	800.0	790.0	757.8

Stability



Slope at top-10: -6.91. Slope overall: -1.61.

Unstable when absolute value of slope of fit line in scatter plot <= 0.25 (slope threshold). Otherwise it is stable.

Ranking Facts

Recipe

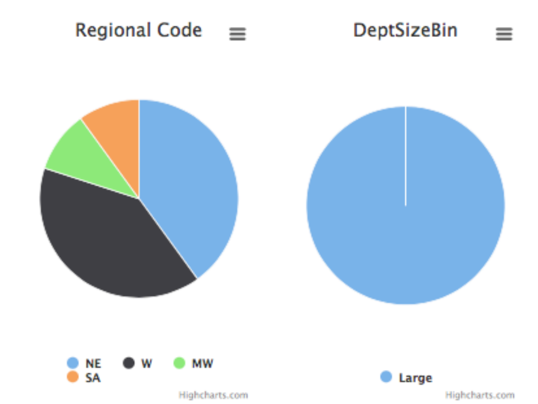
Attribute	Weight
PubCount	1.0
Faculty	1.0
GRE	1.0

Ingredients

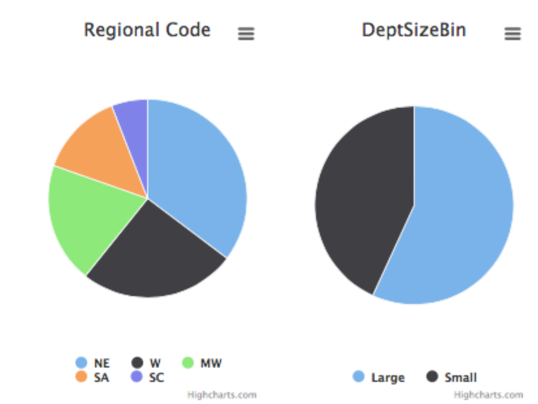
Attribute	Correlation
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Correlation strength is based on its absolute value. Correlation over 0.75 is high, between 0.25 and 0.75 is medium, under 0.25 is low.

Diversity at top-10



Diversity overall



Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

Fairness

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair	Fair	Fair
Small	Unfair	Unfair	Unfair

Unfair when p-value of corresponding statistical test <= 0.05.

Ingredients

Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
CSRankingAllArea	13	6.5	1
Faculty	122	52.5	45

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
CSRankingAllArea	48	26.0	1
Faculty	122	32.0	14

Fairness

	FA*IR		Pairwise		Proportion	
DeptSizeBin	p-value	adjusted α	p-value	α	p-value	α
Large	1.0	0.87	0.99	0.05	1.0	0.05
Small	0.0	0.71	0.0	0.05	0.0	0.05

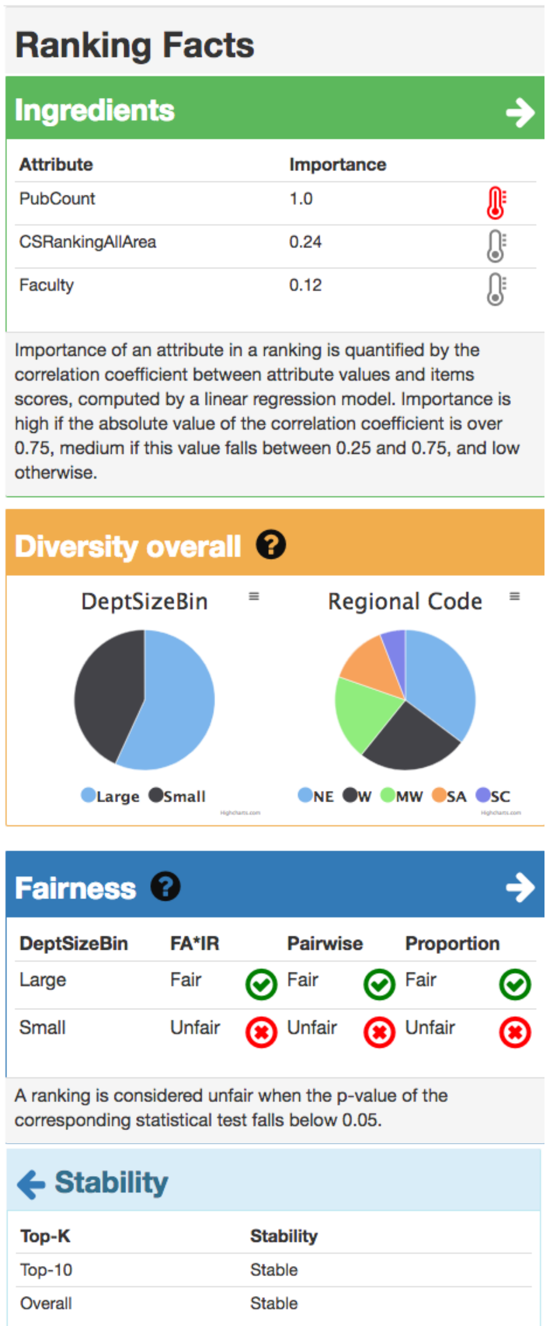
Top K = 26 in FA*IR and Proportion oracles. Setting of top K: In FA*IR and Proportion oracle, if N > 200, set top K =100. Otherwise set top K = 50%N. Pairwise oracle takes whole ranking as input. FA*IR is computed as using code in FA*IR codes. Proportion is implemented as statistical test 4.1.3 in Proportion paper.

http://demo.dataresponsibly.com/rankingfacts/nutrition_facts/

[K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, HV Jagadish, G. Miklau; 2018]

r/ai

Properties of a nutritional label



comprehensible: short, simple, clear

consultative: provide actionable info

comparable: implying a standard

concrete: helps determine a dataset's fitness for use for a given task

computable: produced as a "by-product" of computation - interpretability-by-design

[Stoyanovich and Howe, 2019]

r/ai

Point 5

transparency / interpretability by design, not as an afterthought

provision for transparency and interpretability at every stage of the data lifecycle

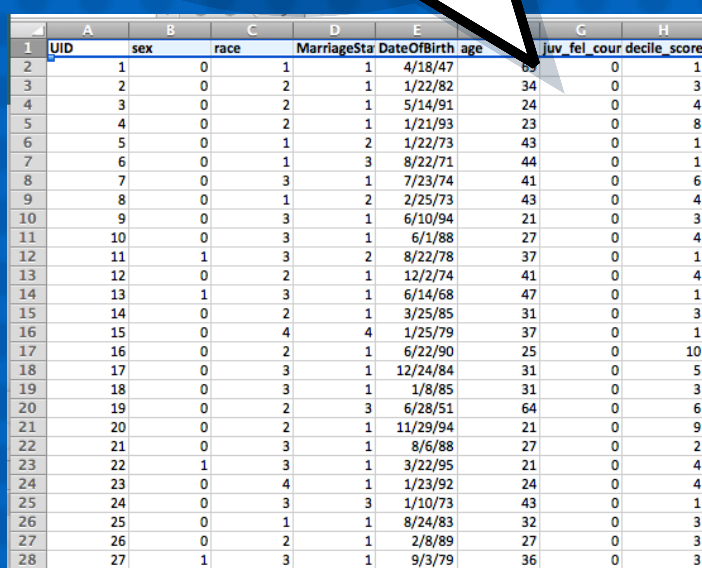
useful internally during development, for communication and coordination between agencies, and for accountability to the public

Frog's eye view

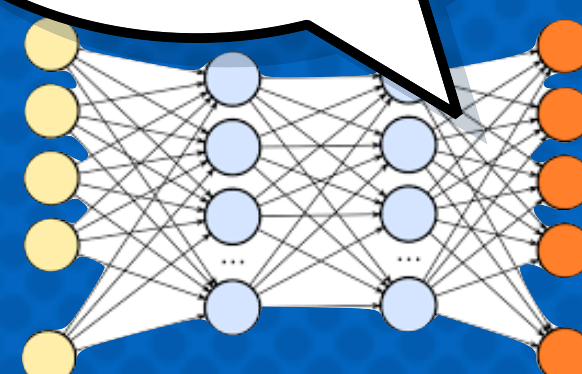
where did the data come from?

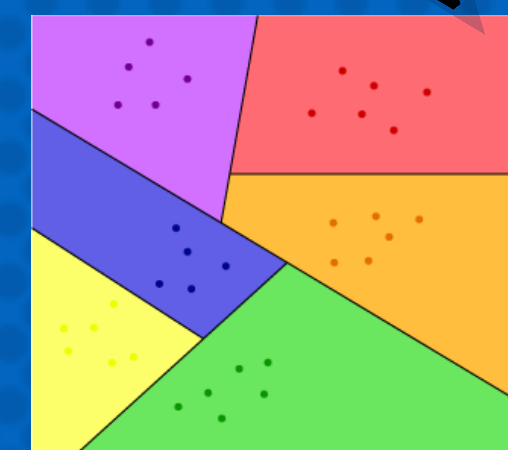
what happens inside the box?

how are results used?

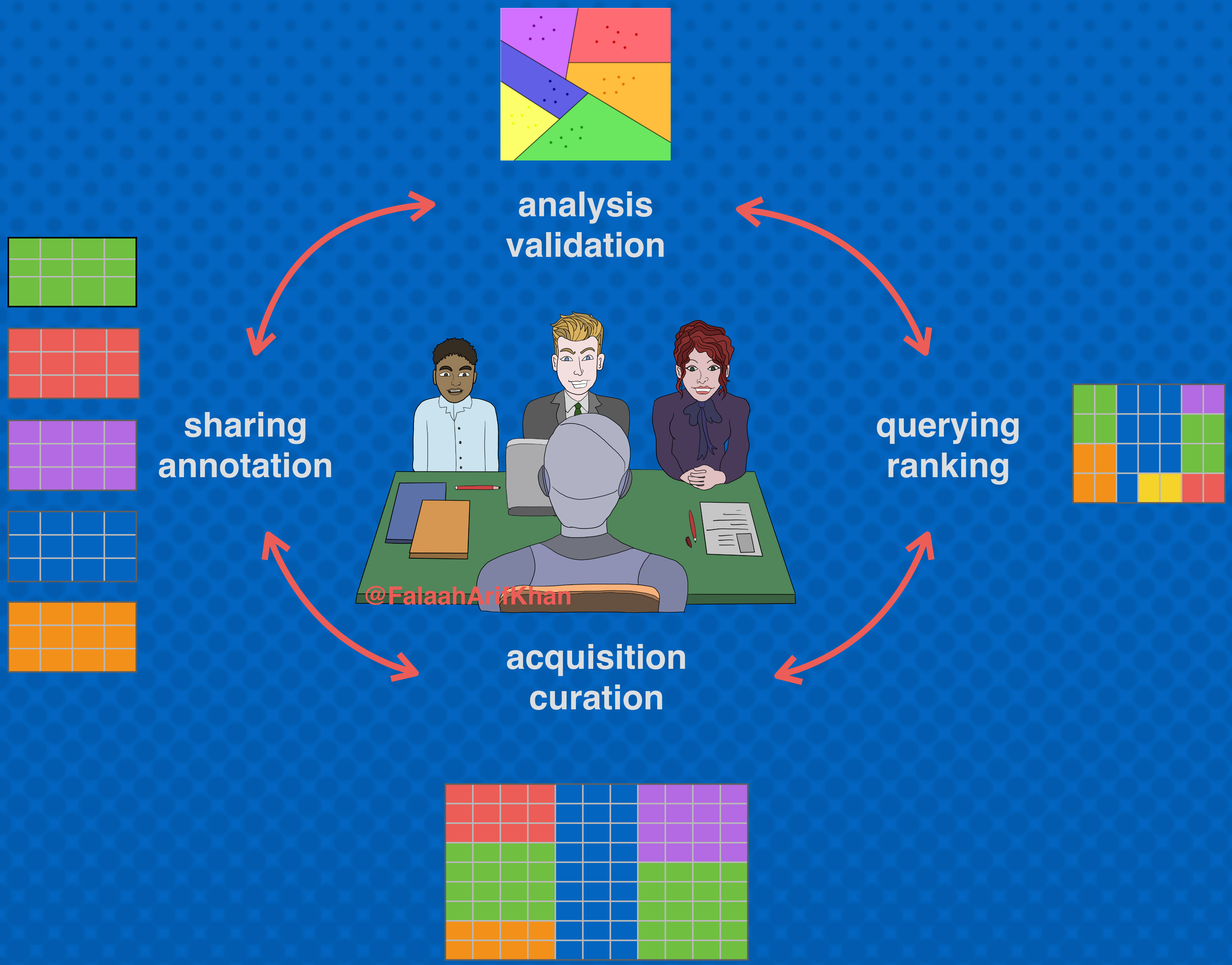


	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	[illegible]	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3





Data lifecycle of an ADS



interpretability in the eye of the stakeholder

What are we explaining?

process (same for everyone? why is this the process?) vs. outcome

procedural justice aims to ensure that algorithms are perceived as fair and legitimate

data transparency is unique to algorithm-assisted decision-making, relates to the justification dimension of interpretability

[J. Stoyanovich, J. Van Bavel, T. West, 2020]

r/ai

To whom are we explaining and why?

accounting for the needs of different stakeholders

social identity - people trust their in-group members more

moral cognition - is a decision or outcome morally right or wrong?

[J. Stoyanovich, J. Van Bavel, T. West, 2020]

r/ai

How do we know that we explained well?

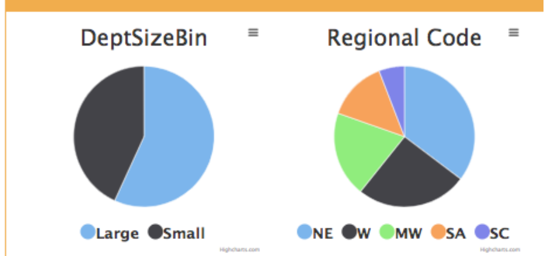
Ranking Facts

Ingredients

Attribute	Importance
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Importance of an attribute in a ranking is quantified by the correlation coefficient between attribute values and items scores, computed by a linear regression model. Importance is high if the absolute value of the correlation coefficient is over 0.75, medium if this value falls between 0.25 and 0.75, and low otherwise.

Diversity overall



Fairness

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair	Fair	Fair
Small	Unfair	Unfair	Unfair

A ranking is considered unfair when the p-value of the corresponding statistical test falls below 0.05.

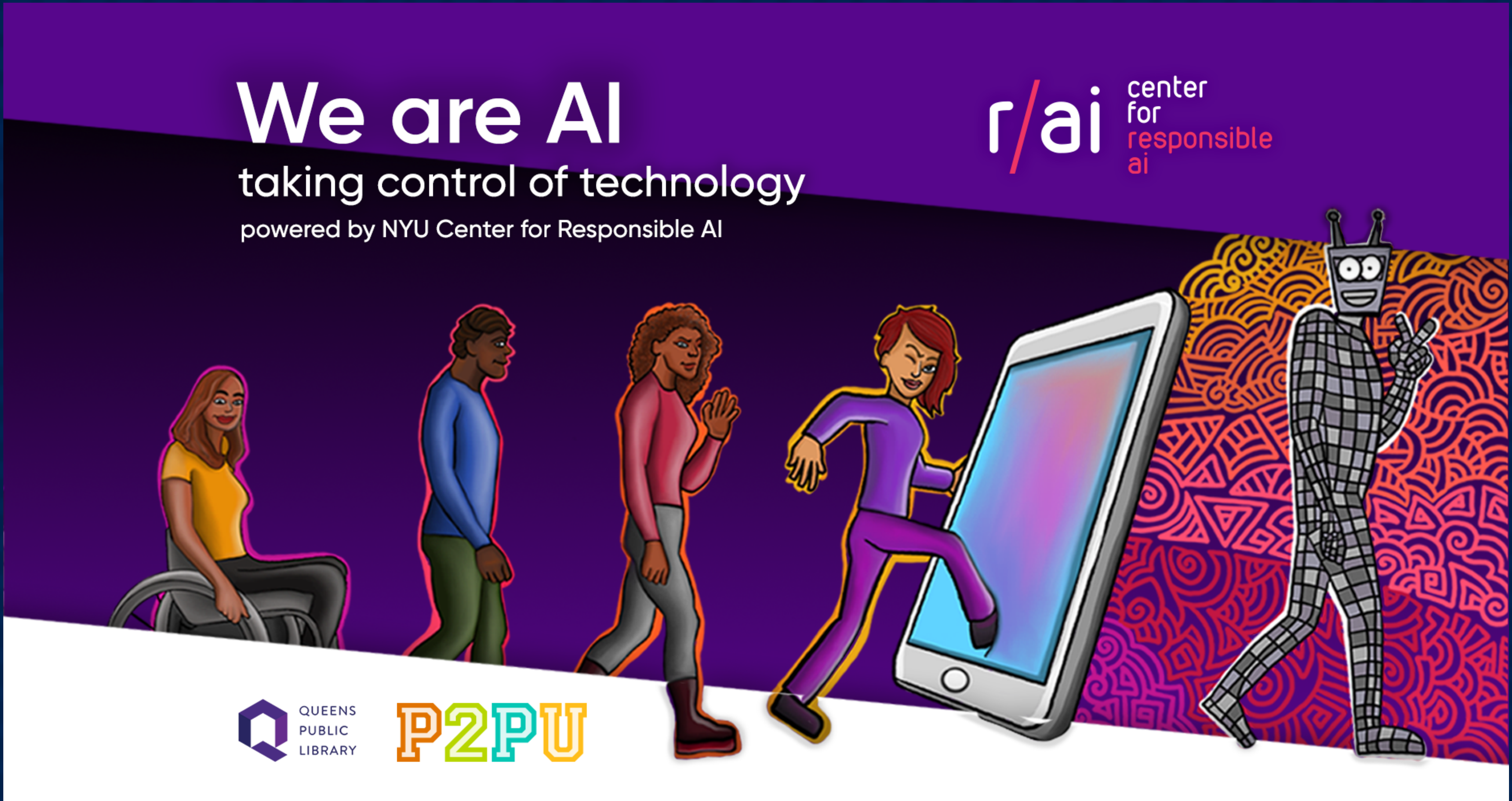
Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

nutritional labels! :)

… but do they work?

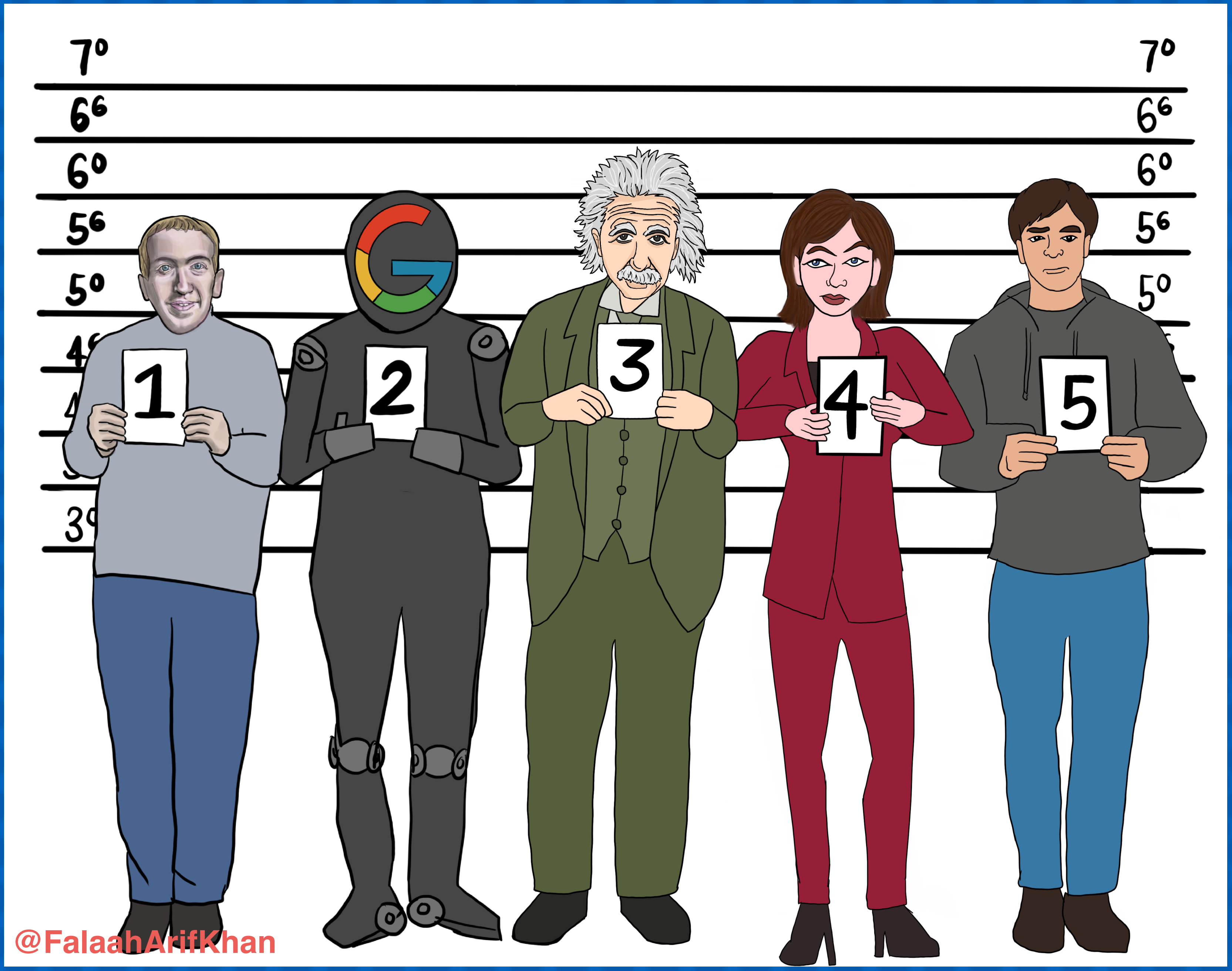
[J. Stoyanovich, J. Van Bavel, T. West, 2020]



https://dataresponsibly.github.io/we-are-ai/

r/ai

We all are responsible



Tech rooted in people

Responsible Data Science

Thank you!