

TOWARDS TRUSTWORTHY AI WITH



DR. BERIL SIRMACEK

GDG UK/IRELAND

August 2021



AGENDA

1. Our need for trust
2. TrustworthyAI
3. Towards ResponsibleAI with Google tools
4. Other helpful tools

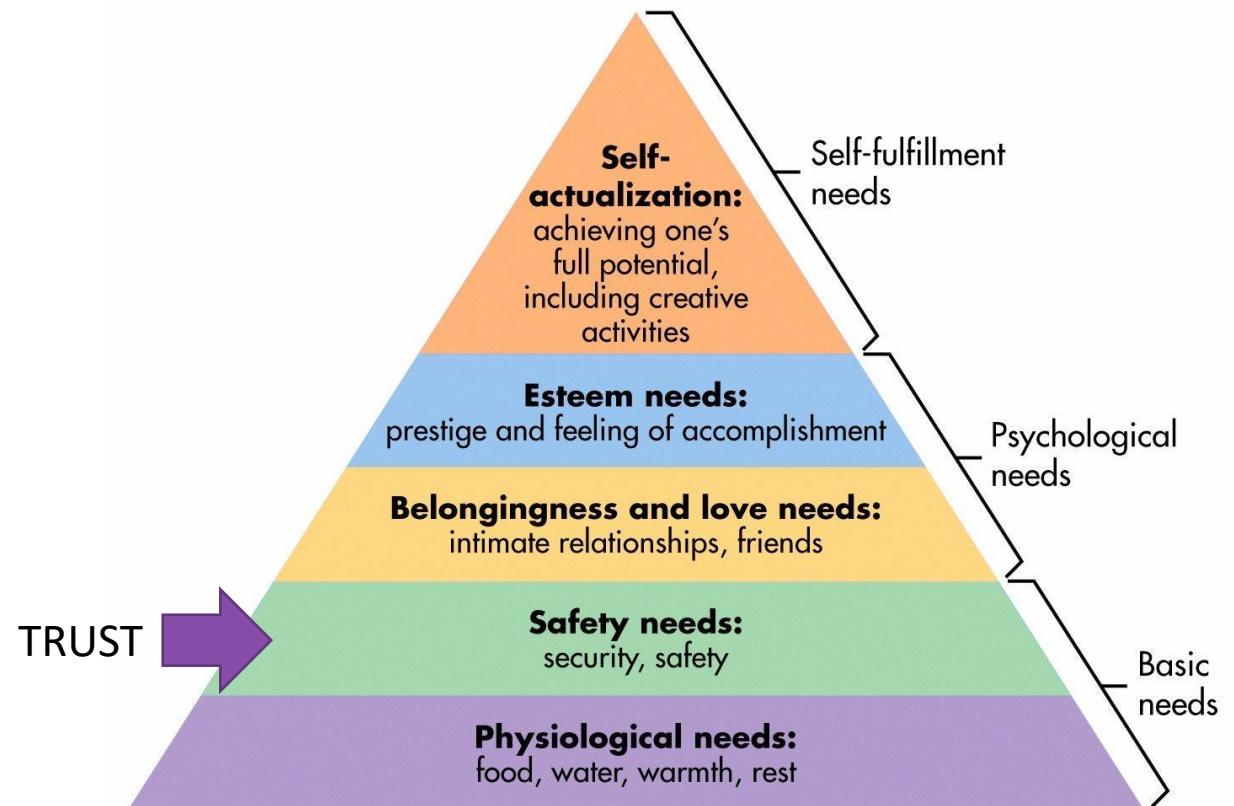
1. OUR NEED FOR TRUST

WHY SHOULD I TRUST?



1. Our need for trust

A basic human need



Maslow's Hierarchy of Needs

<https://www.simplypsychology.org/maslow.html>

1. Our need for trust

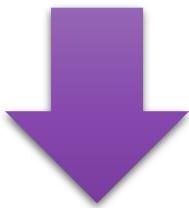
Safety, Security, Equality



Here is the model I made for you!

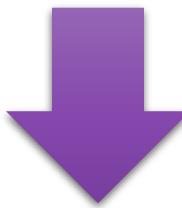
1. Our need for trust

Safety, Security, Equality



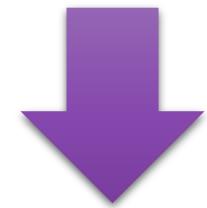
Decision

Why?



Decision

Why?



Decision

Why?

1. Our need for trust

<https://www.technologyreview.com/2013/02/04/253879/racism-is-poisoning-online-ad-delivery-says-harvard-professor/>

Ad related to latanya sweeney ⓘ

Latanya Sweeney Truth
www.instantcheckmate.com/

Looking for Latanya Sweeney? Check Latanya Sweeney's Arrests.

Ads by Google

Latanya Sweeney, Arrested?

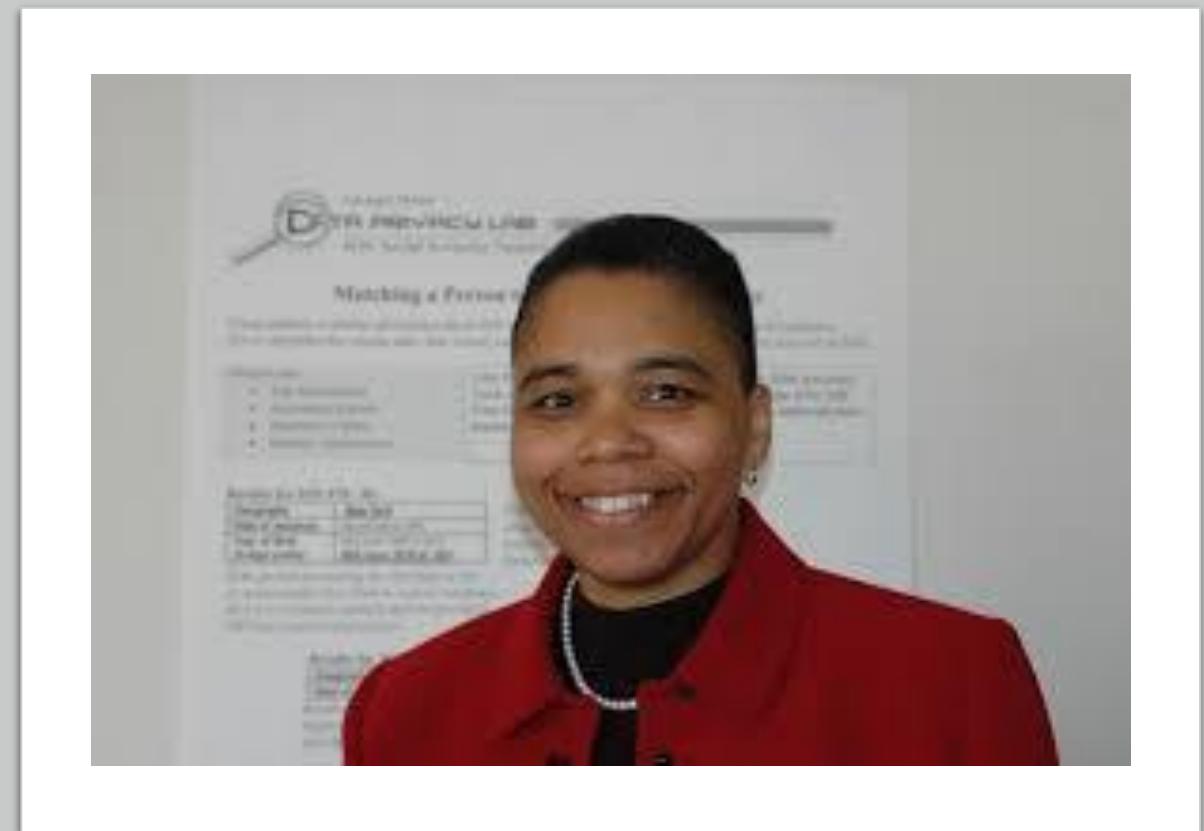
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney

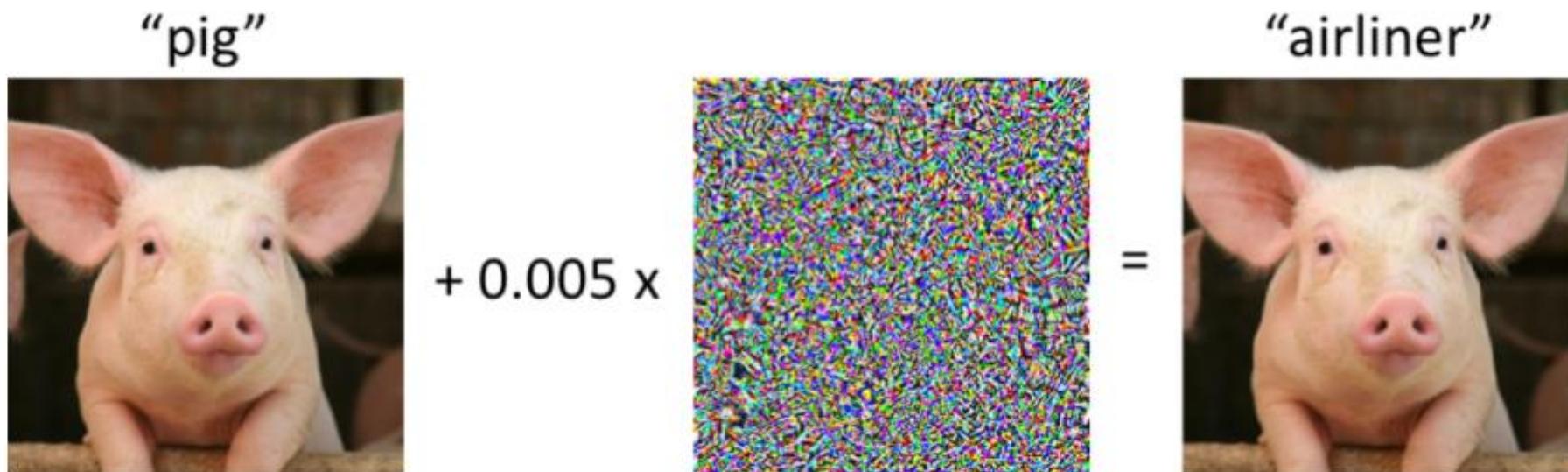
Public Records Found For: Latanya Sweeney. View Now.
www.publicrecords.com/

La Tanya

Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya



1. Our need for trust



<https://doi.org/10.1145/1014052.1014066>

1. Our need for trust

Explain the Prediction



Predicted: Wolf
True: Wolf



Predicted: Husky
True: Husky



Predicted: Husky
True: Husky



Predicted: Wolf
True: Wolf



Predicted: Wolf
True: Wolf



Predicted: Wolf
True: Wolf



Predicted: Husky
True: Wolf



Predicted: Wolf
True: Wolf

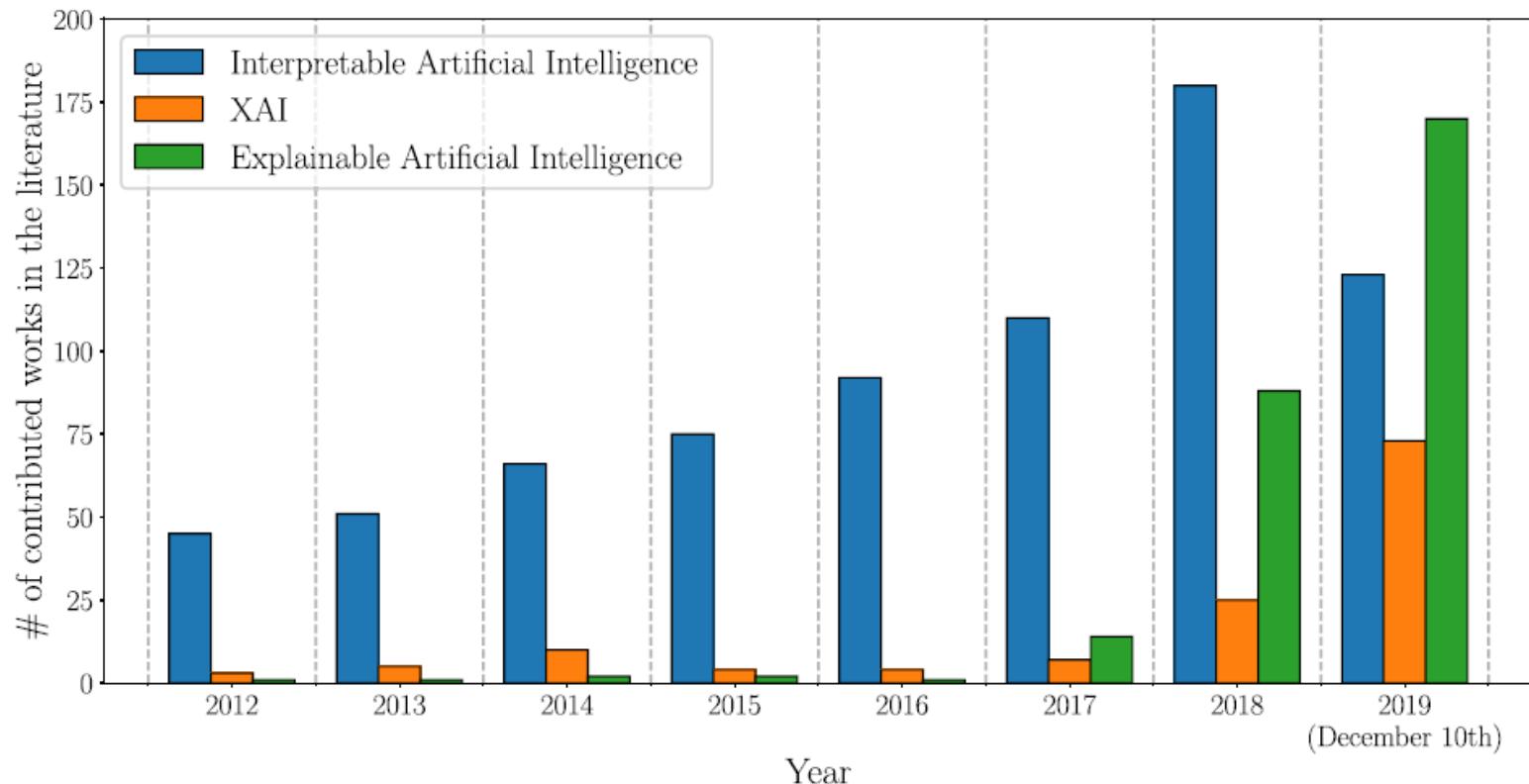


Predicted: Wolf
True: Husky



Predicted: Husky
True: Husky

1. Our need for trust

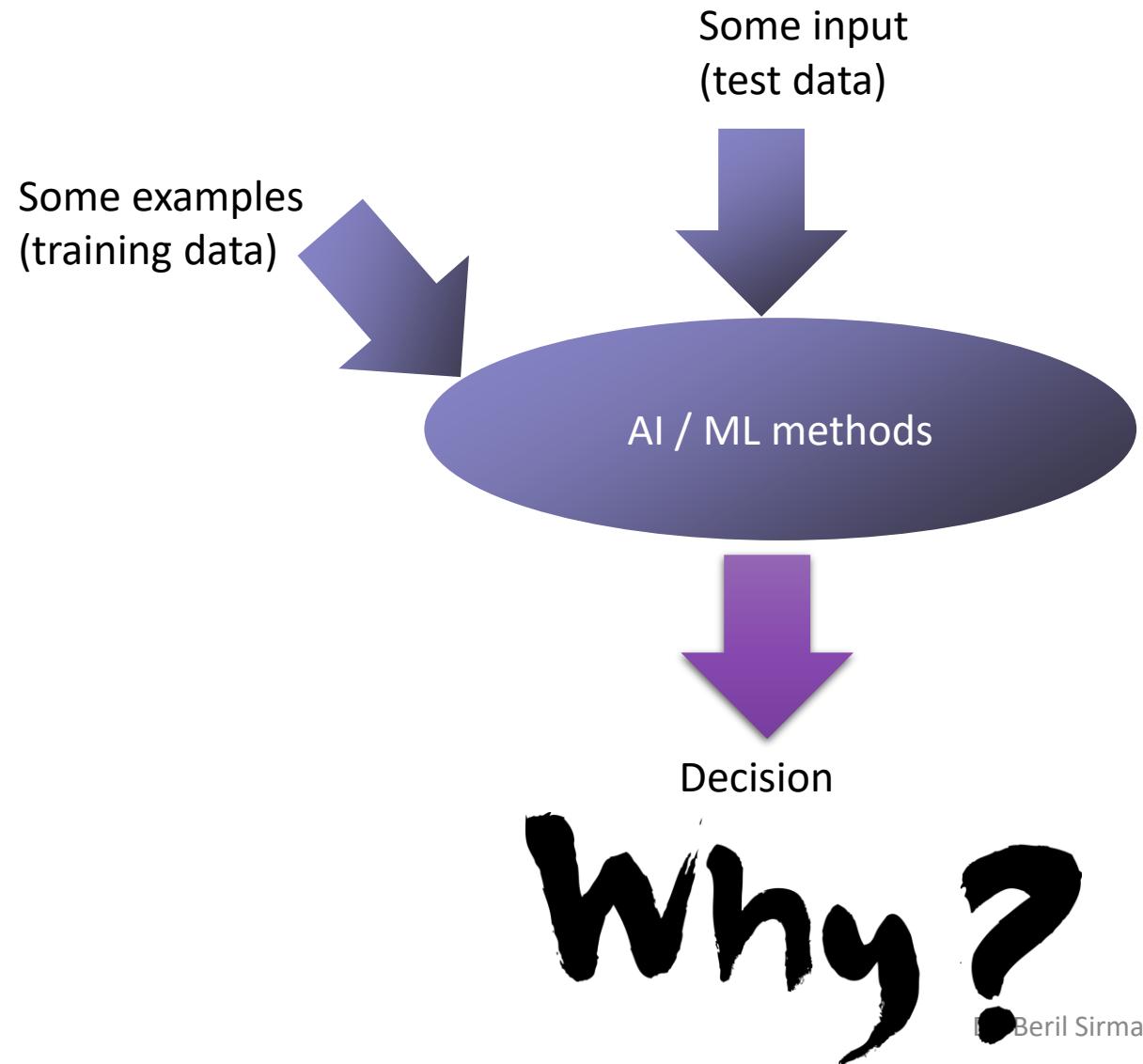


Source: "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI"

2. TrustworthyAI

LET'S DIVE IN

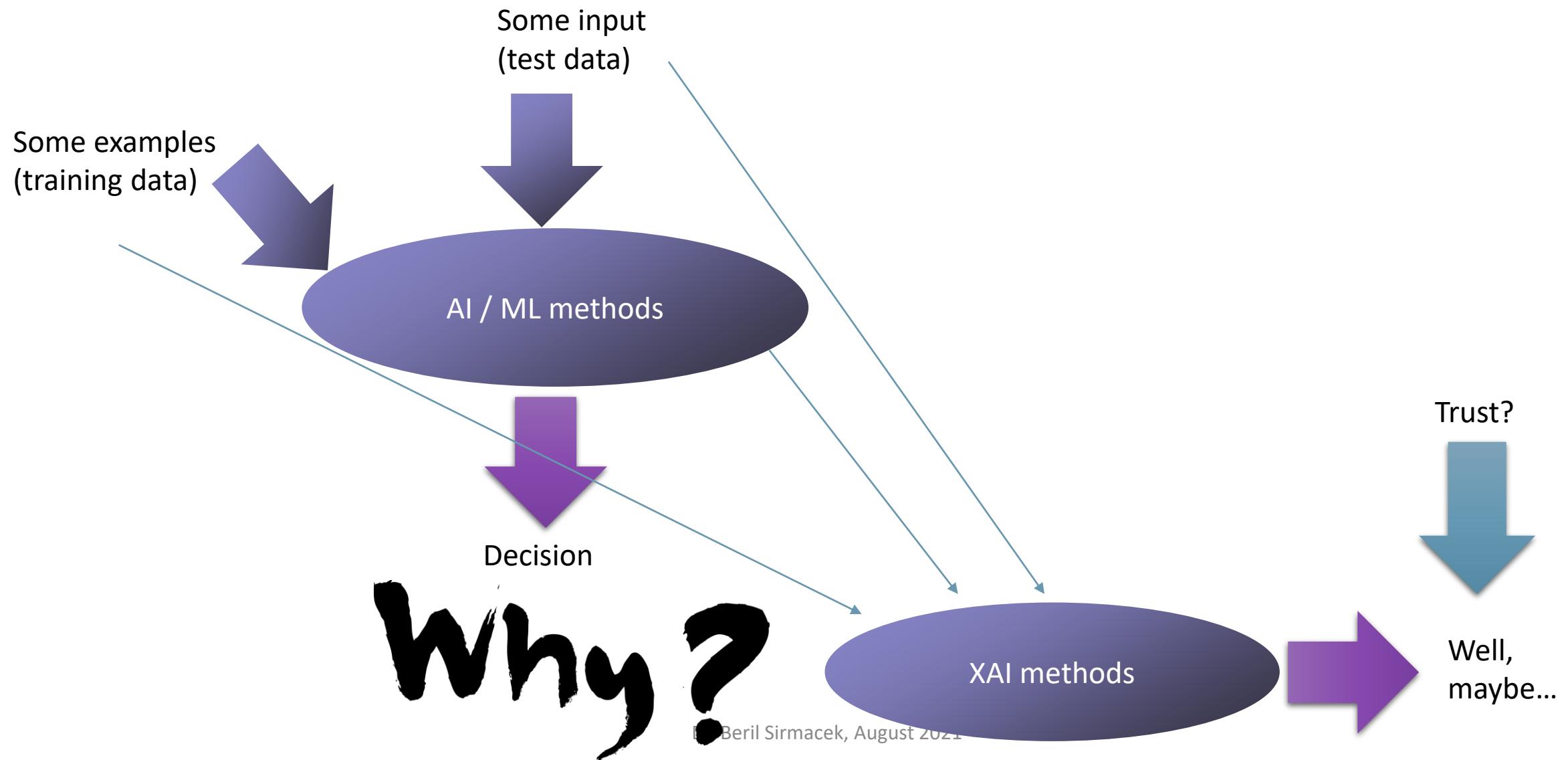
2. TrustworthyAI



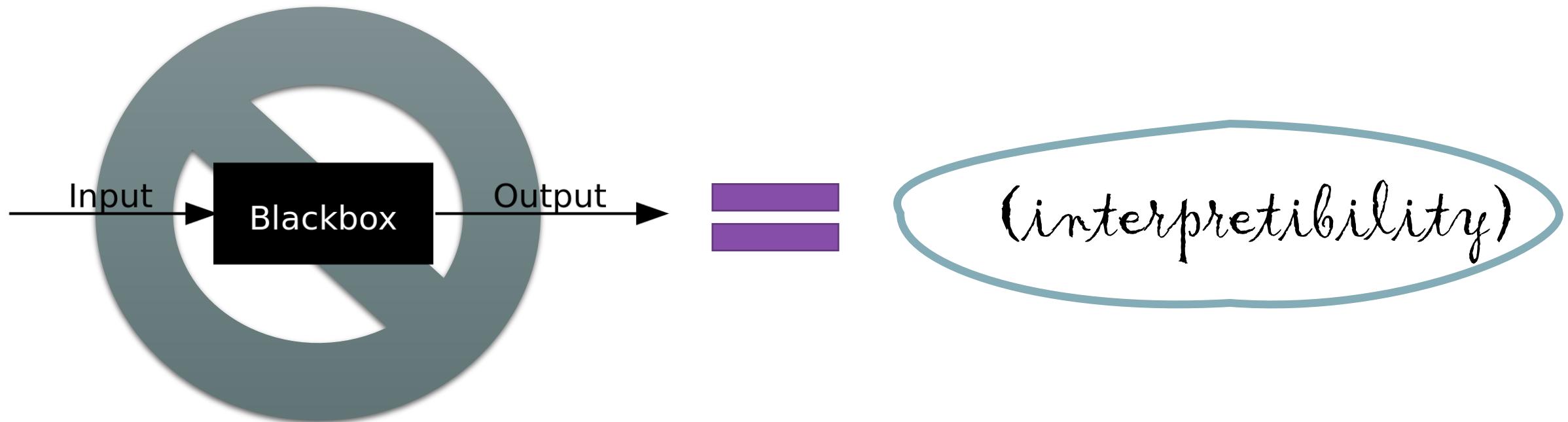
"Given a certain audience, **explainability** refers to the **details** and **reasons** a model gives to make its functioning clear or easy to understand."

"Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI"

2. TrustworthyAI



2. TrustworthyAI



2. TrustworthyAI

1. ***DEPENDING ON THE USER***
2. ***DEPENDING ON THE TRANSPARENCY LEVEL FOCUS***
3. ***DEPENDING ON THE PROBLEMS FOCUSED***
4. ***DEPENDING ON THE MATHEMATICAL METHODS***
5. ***DEPENDING ON WHETHER IT IS NEEDED BEFORE OR AFTER DEVELOPMENT PROCESS***

(interpretability)

2. TrustworthyAI

1. **Simulability** (*entire model*)
2. **Decomposability** (*components, parameters*)
3. **Algorithmic transparency** (*focusing on training*)



(interpretability)

At different levels TRANSPARENCY can be focused on.

2. TrustworthyAI

Summary of the interpretability approaches considering the mathematical methods adopted;

1. Decision Trees (DT)
2. Decision Rules (DR)
3. Features Importance (FI)
4. Saliency Mask (SM)
5. Sensitivity Analysis (SA)
6. Partial Dependence Plot (PDP)
7. Neurons Activation (NA)
8. Prototype Selection (PS)



(interpretability)

2. TrustworthyAI

SHAP (SHapley Additive exPlanations)

- Feature Importance Analysis Method
- Lundberg and Lee (2016) introduced to explain individual contributions



SHAP

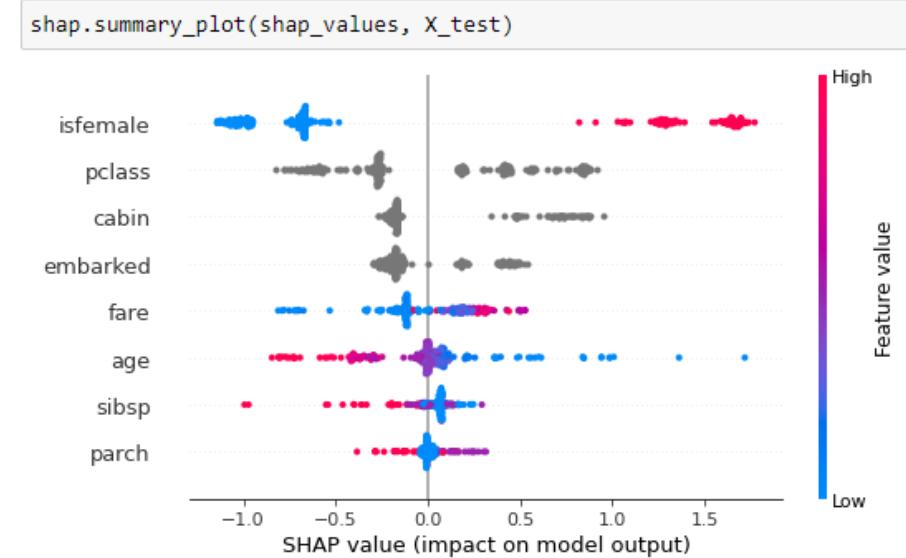
pip install shap

```
# Read Titanic data
titanic_df = pd.read_csv('http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.csv')
print(titanic_df.head())

   pclass  survived          name     sex \
0      1         1  Allen, Miss. Elisabeth Walton  female
1      1         1  Allison, Master. Hudson Trevor    male
2      1         0  Allison, Miss. Helen Loraine  female
3      1         0  Allison, Mr. Hudson Joshua Creighton    male
4      1         0  Allison, Mrs. Hudson J C (Bessie Waldo Daniels)  female

   age  sibsp  parch  ticket      fare  cabin embarked boat  body \
0  29.00     0     0   24160  211.3375      B5        S    2   NaN
1  0.92      1     2   113781  151.5500     C22    C26        S    11   NaN
2  2.00      1     2   113781  151.5500     C22    C26        S   NaN   NaN
3 30.00      1     2   113781  151.5500     C22    C26        S   NaN  135.0
4 25.00      1     2   113781  151.5500     C22    C26        S   NaN   NaN

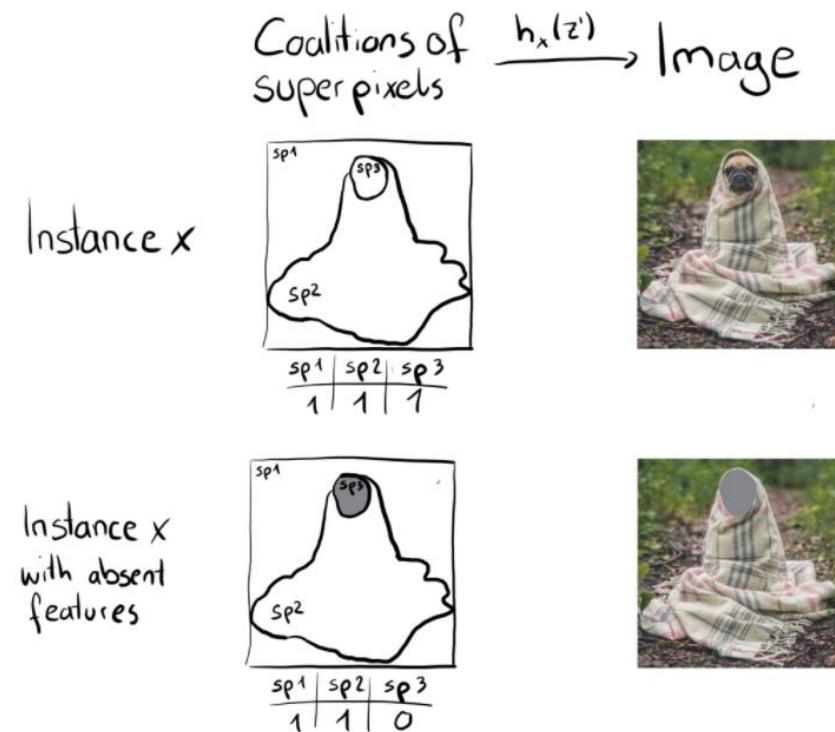
   home.dest
0  St Louis, MO
1  Montreal, PQ / Chesterville, ON
2  Montreal, PQ / Chesterville, ON
3  Montreal, PQ / Chesterville, ON
4  Montreal, PQ / Chesterville, ON
```



2. TrustworthyAI

LIME (Local Interpretable Model-Agnostic Explanations)

- For more theory, please check <https://christophm.github.io/interpretable-ml-book/shap.html>



2. TrustworthyAI

Explain the Prediction



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

[\[1602.04938\] "Why Should I Trust You?": Explaining the Predictions of Any Classifier](#)

Trustworthiness
Transparency
Known Bias
Ethics
Factualness
Constantly improved



SATISFIED AUTHORITIES



A SATISFIED BUSINESS OWNER



A SATISFIED CUSTOMER

TOWARDS RESPONSIBLEAI



3. Towards ResponsibleAI

LOOKING AHEAD

Building with the Responsible AI Toolkit

Speakers



Catherina Xu

Product Lead, Responsible ML Infrastructure



Ludovic Peran

Product Lead, People+AI Research

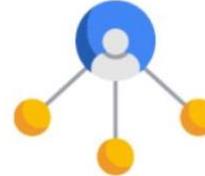
The following Google workflow is edited from:

Google's AI Principles

Google's AI Principles



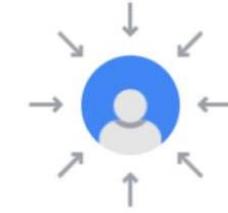
1. Be socially beneficial.



2. Avoid creating or
reinforcing unfair bias.



3. Be built and tested for safety.

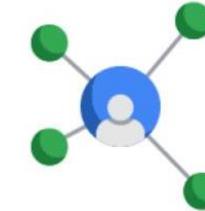


4. Be accountable to people.



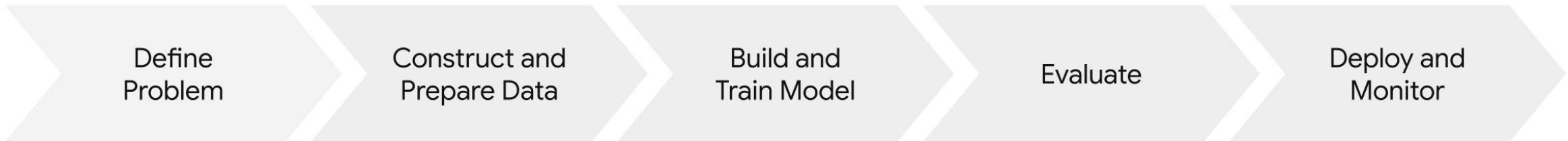
5. Incorporate privacy
design principles.

6. Uphold high standards of
scientific excellence.



7. Be made available for uses that
accord with these principles.

The typical machine learning workflow



Responsible AI Considerations



Who is my ML system for?

Is ML a good fit for the problem?

Responsible AI Considerations



Are there real world
and/or human biases
in my data?

Responsible AI Considerations



What methods
should I use to train
my model in a fair,
interpretable, private,
and secure way?

Responsible AI Considerations



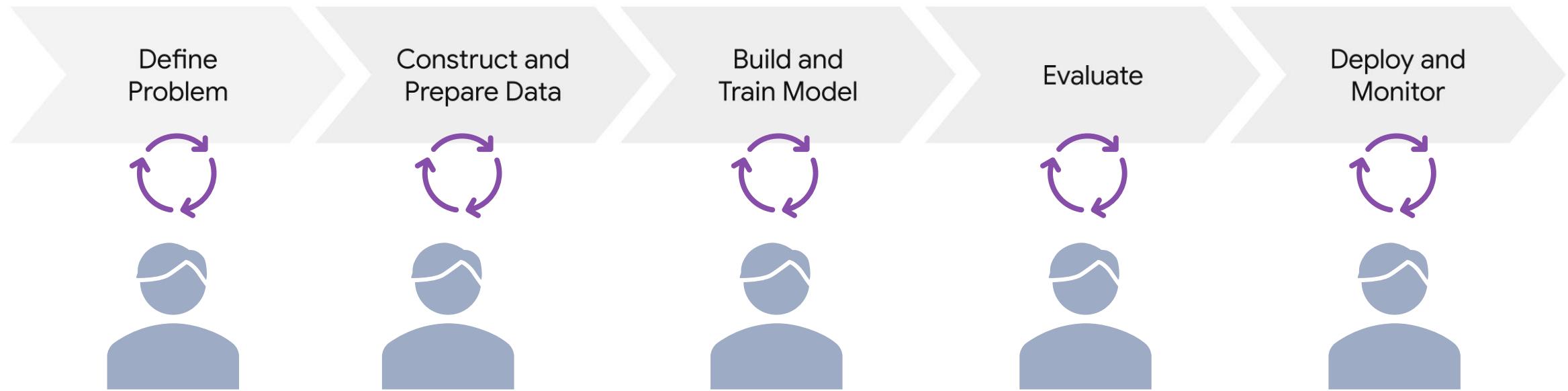
What metrics should
I use to evaluate
my model?

Responsible AI Considerations

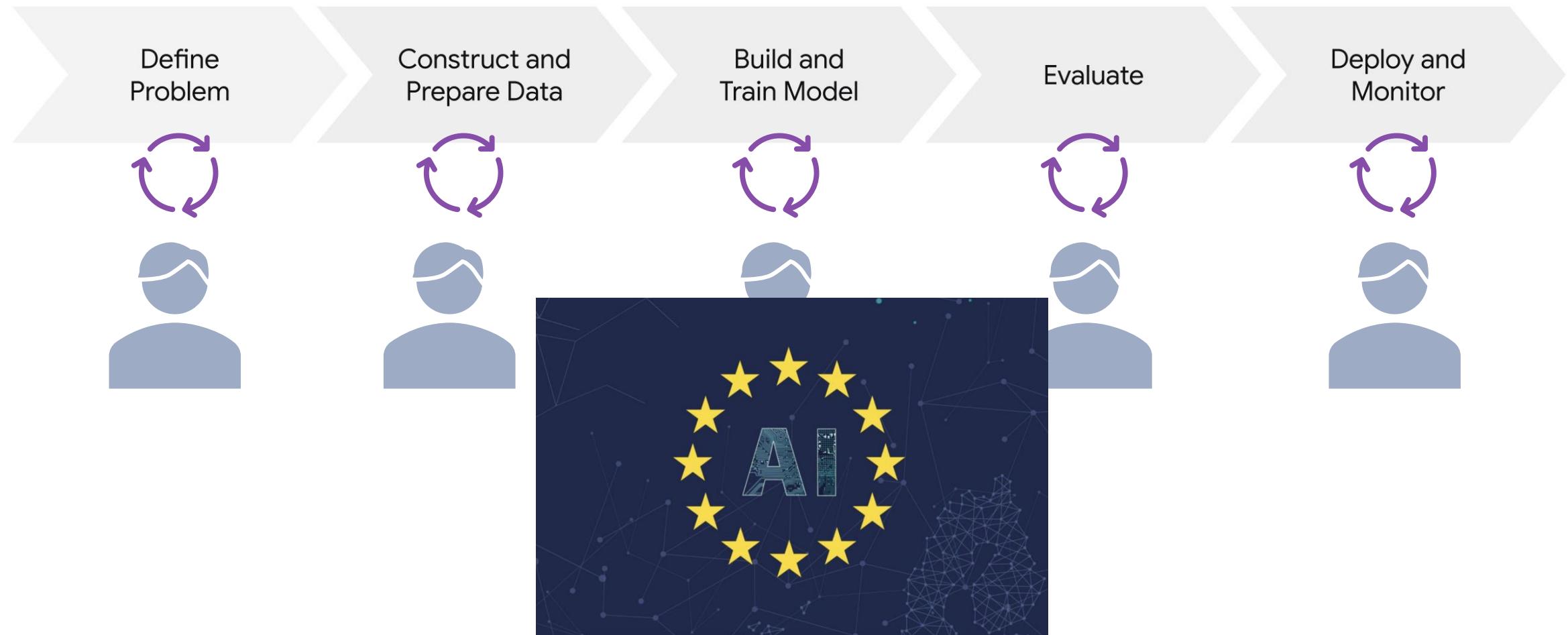


What should key stakeholders know about my model?

The typical machine learning workflow



The typical machine learning workflow



<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Introducing: The Responsible AI Toolkit

A collection of tools to help developers make progress in the responsible development of AI



www.tensorflow.org/responsible_ai

The screenshot shows the TensorFlow website's main navigation bar with options like Install, Learn, API, Resources (which is currently selected), Community, and More. Below the navigation is a sub-navigation for 'Responsible AI' with links for Overview, Guide, Tutorials, and API. A large orange TensorFlow logo is partially visible on the left. On the right, there's a call-to-action button 'Find an event' and a link 'RSVP for your local TensorFlow Everywhere event today!'. Below these, a section titled 'Learn how to integrate Responsible AI practices into your ML workflow using TensorFlow' is displayed, followed by a descriptive paragraph about TensorFlow's commitment to responsible AI development.

TensorFlow

Install Learn API Resources Community More Search

Responsible AI

Overview Guide Tutorials API

RSVP for your local TensorFlow Everywhere event today!

Find an event

Learn how to integrate Responsible AI practices into your ML workflow using TensorFlow

TensorFlow is committed to helping make progress in the responsible development of AI by sharing a collection of resources and tools with the ML community.

Introducing: The Responsible AI Toolkit

A collection of tools to help developers make progress in the responsible development of AI



www.tensorflow.org/responsible_ai

10+ Colab tutorials, API docs, and usage guides!

A screenshot of the TensorFlow website's main navigation bar. A yellow arrow points from the text "10+ Colab tutorials, API docs, and usage guides!" to the "Responsible AI" link in the navigation menu. The "Responsible AI" link is highlighted with a yellow box, and below it, the "Tutorials" link is also highlighted with a yellow box. The rest of the navigation links (Install, Learn, API, Resources, Community, More) are in a grey font. The TensorFlow logo is at the top left. A search bar is at the top right. Below the navigation bar, there is a dark banner with the text "RSVP for your local TensorFlow Everywhere event today!" and a "Find an event" button. The main content area features a large orange graphic element on the left and a stylized illustration of a computer monitor and smartphone on the right.

Learn how to integrate Responsible AI practices into your ML workflow using TensorFlow

TensorFlow is committed to helping make progress in the responsible development of AI by sharing a collection of resources and tools with the ML community.

The typical machine learning workflow



People + AI
guidebook

AI explorables

 PAIR

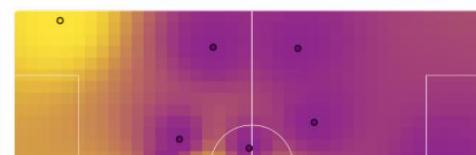
GUIDEBOOK EXPLORABLES TOOLS RESEARCH EVENTS M

AI Explorables

Big ideas in machine learning, simply explained

The rapidly increasing usage of machine learning raises complicated questions: How can we tell if models are fair? Why do models make the predictions that they do? What are the privacy implications of feeding enormous amounts of data into models?

This ongoing series of interactive, formula-free essays will walk you through these important concepts.



The typical machine learning workflow



Define
Problem

Construct and
Prepare Data

Build and
Train Model

Evaluate

Deploy and
Monitor

TF data
validation

Data cards
playbook

Know your data

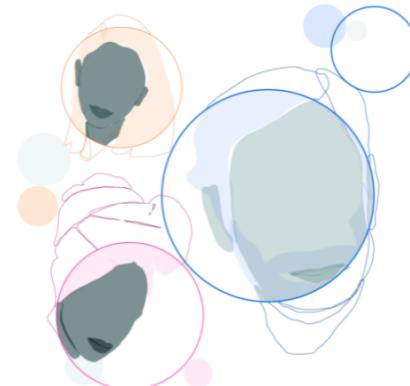
Data Cards Playbook

ABOUT ACTIVITIES RESOURCES PLANNING GET INVOLVED ↗

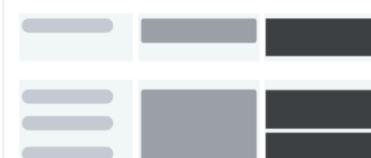
Transparency

Transparency, at its heart, is a clear, easily understandable, and plain language explanation of what something is, what it does, and why it does that.

For dataset documentation, transparency means providing a window into the dataset's quality, validity, reproducibility, and risk—all in a way that's accessible to the various audiences who may interact with it, including developers, business stakeholders, downstream users, and more.



<Your dataset here>



Data Cards

Data Cards are people-centric summaries of transparent dataset documentation.

They offer a structured way to document datasets and encourage informed decisions about the data used in AI systems for product and research. But there is no one-size-fits-all template. That's where the Data Cards Playbook comes in.

The typical machine learning workflow



Define
Problem

Construct and
Prepare Data

Build and
Train Model

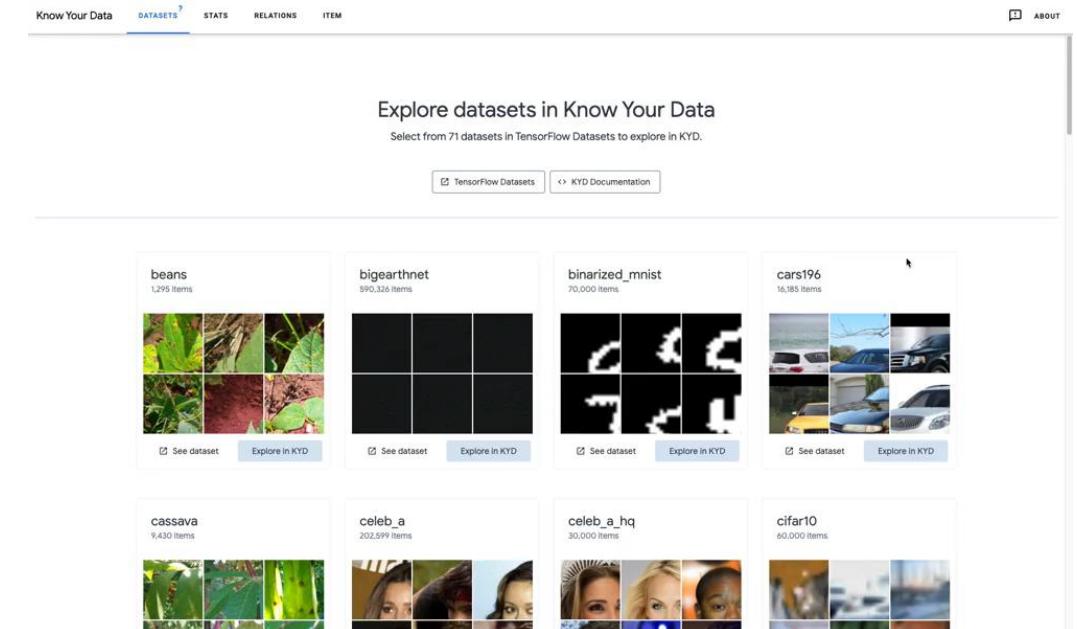
Evaluate

Deploy and
Monitor

TF data
validation

Data cards
playbook

Know your data

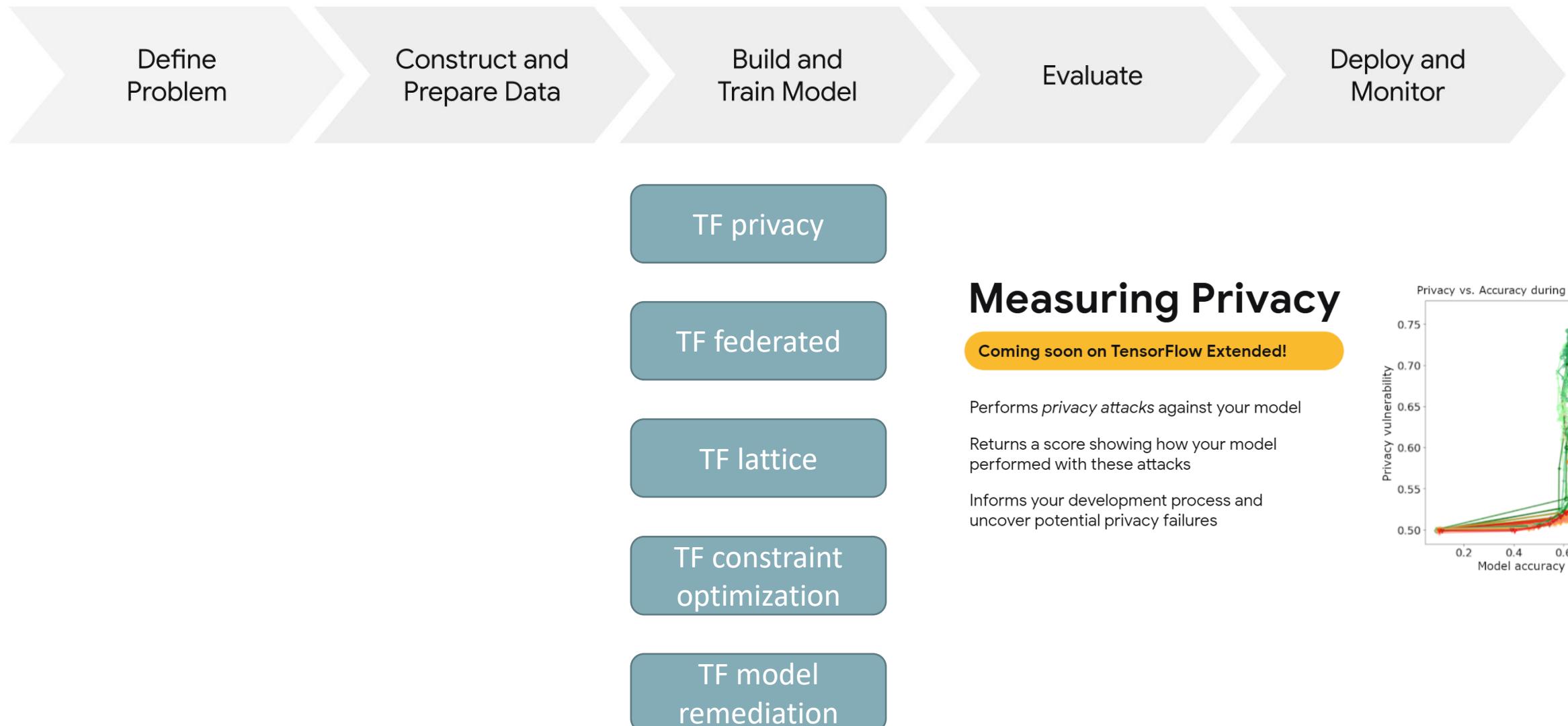


Know Your Data is a tool to help researchers, engineers, product teams and policy teams explore datasets, improve data quality and mitigate bias issues.

KYD aims to answer the following questions:

- Is my data corrupted? (e.g. broken images, garbled text, bad labels, etc).
- Is my data sensitive? (e.g. are there humans, explicit content).
- Does my data have gaps? (e.g. lack of daylight photos).
- Is my dataset balanced across various attributes?

The typical machine learning workflow



The typical machine learning workflow



Define
Problem

Construct and
Prepare Data

Build and
Train Model

Evaluate

Deploy and
Monitor

TF privacy

TF federated

TF lattice

TF constraint
optimization

TF model
remediation

Flexible, controlled and interpretable ML with lattice based models

TensorFlow Lattice is a library that implements constrained and interpretable lattice based models. The library enables you to inject domain knowledge into the learning process through common-sense or policy-driven [shape constraints](#). This is done using a collection of [Keras layers](#) that can satisfy constraints such as monotonicity, convexity and how features interact. The library also provides easy to setup [premade models](#) and [canned estimators](#).

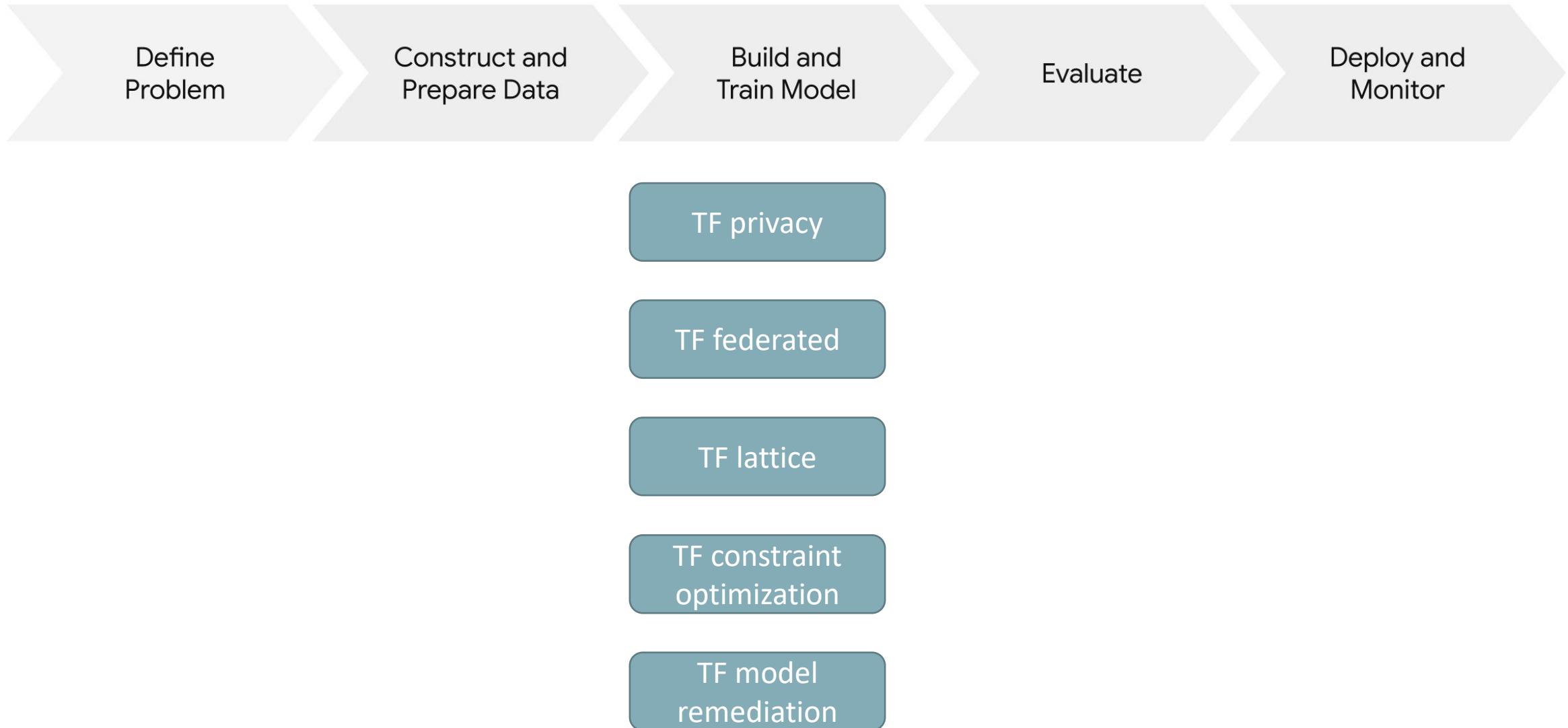
With TF Lattice you can use domain knowledge to better extrapolate to the parts of the input space not covered by the training dataset. This helps avoid unexpected model behaviour when the serving distribution is different from the training distribution.



```
import numpy as np
import tensorflow as tf
import tensorflow_lattice as tfl

model = tf.keras.models.Sequential()
model.add(
    tfl.layers.ParallelCombination([
        # Monotonic piece-wise linear calibration with bounded slopes
        tfl.layers.PWLCalibration(
            monotonicity='increasing',
            input_keypoints=np.linspace(1., 5., num=20),
            output_min=0.0,
            output_max=1.0),
        # Diminishing returns
        tfl.layers.PWLCalibration(
            monotonicity='increasing',
            convexity='concave',
            input_keypoints=np.linspace(0., 200., num=20),
            output_min=0.0,
            output_max=2.0),
        # Partially monotonic categorical calibration: calib(
        tfl.layers.CategoricalCalibration(
            num_buckets=4,
            output_min=0.0,
            output_max=1.0,
            monotonicities=[(0, 1)]),
    )))
model.add(
    tfl.layers.Lattice(
        lattice_sizes=[2, 3, 2],
        monotonicities=['increasing', 'increasing', 'increasing'],
        # Trust: model is more responsive to input 0 if input
        edgeworth_trusts=(0, 1, 'positive')))
model.compile(...)
```

The typical machine learning workflow



The typical machine learning workflow

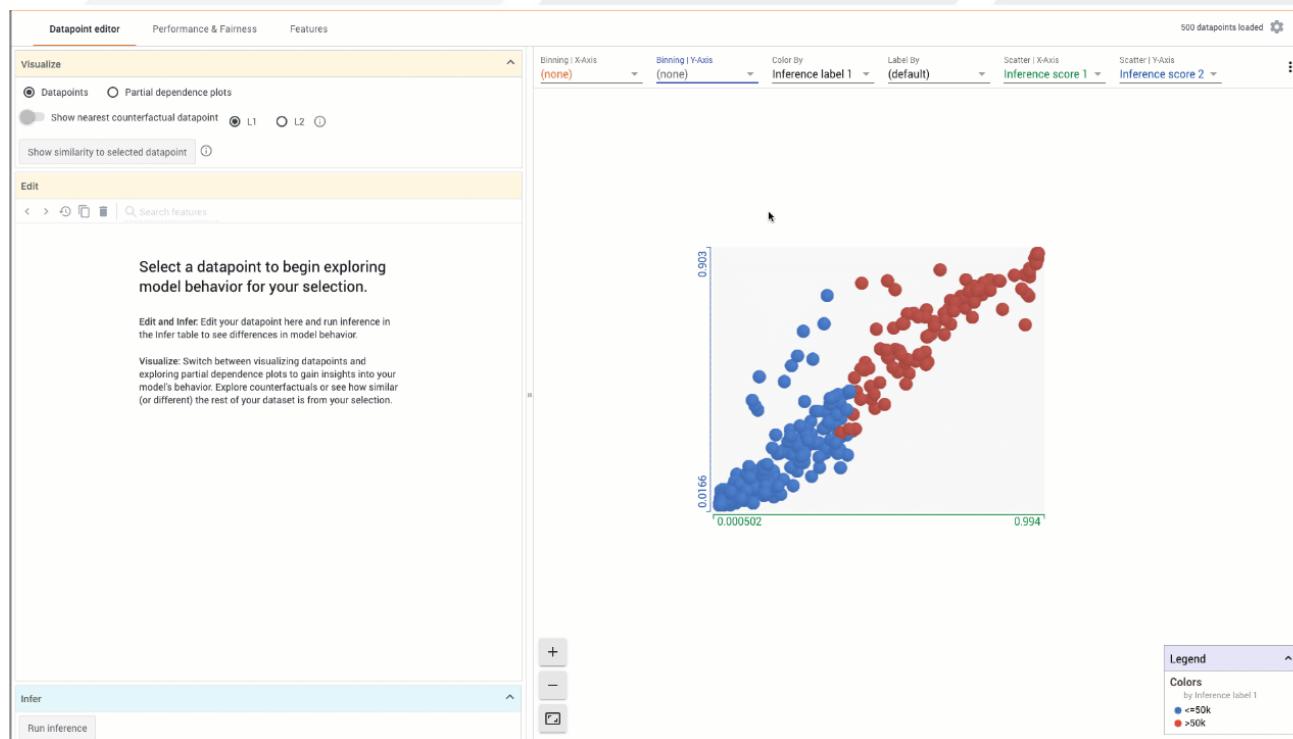
Define Problem

Construct and Prepare Data

Build and Train Model

Evaluate

Deploy and Monitor



Fairness indicators

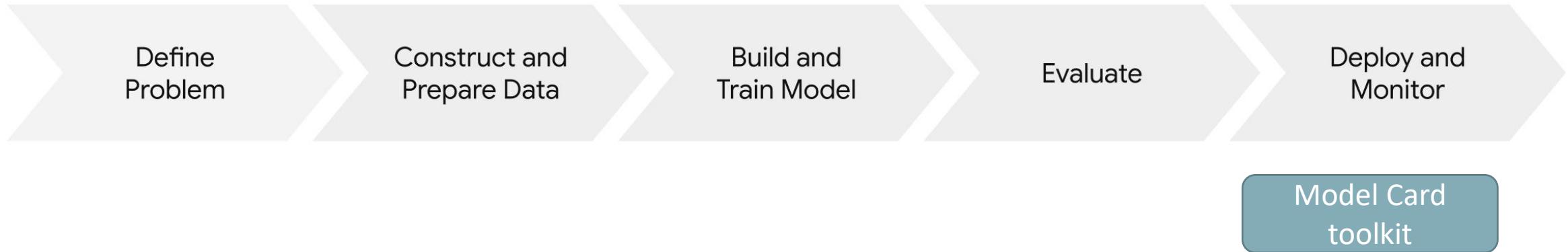
What-if tool

Language interpretability

TF privacy tests

Recommended Lecture:
<https://youtu.be/u96F9BUfH3o>

The typical machine learning workflow



The typical machine learning workflow



Model Card toolkit

EXPLAINABLE AI
Overview
Benefits
Features
Customers
Resources
Pricing

Explainable AI BETA

Tools and frameworks to understand and interpret your machine learning models.

Go to console ▾

View documentation ▾

Understand AI output and build trust

Explainable AI is a set of tools and frameworks to help you understand and interpret predictions made by your machine learning models. With it, you can debug and improve model performance, and help others understand your models' behavior. You can also generate feature attributions for model predictions in



https://github.com/GoogleCloudPlatform/explainable_ai_sdk

4. Other helpful tools

Grad-Cam

https://keras.io/examples/vision/grad_cam/



Recommended Lecture:
<https://youtu.be/DRqMJ-MeBxQ>

Predicted: ['Border_collie']

4. Other helpful tools

<https://convnetplayground.fastforwardlabs.com/>

The screenshot shows the ConvNet Playground interface for performing semantic image search. The top navigation bar includes links for ConvNet Playground, Semantic Search, Model Explorer, and FAQ. The main section is titled "Image Similarity Search".

Search Configuration:

- Select Dataset:** ICONIC200 (200 images), FASHION200 (200 images), TINYIMAGENET (200 images).
- Select Model:** MOBILENET (4.3M params.), EFFICIENTNE.. (5.3M params.), DENSENET121 (8.1M params.), XCEPTION (22.9M params.), INCEPTIONV3 (23.9M params.).
- Select Layer:** layer 1 (864 params.), layer 7 (28.7k params.), layer 64 (725.8k params.), layer 185 (6.5M params.), layer 196 (6.8M params.). A dropdown menu also lists layer 236 (9.8M params.), layer 294 (21.4M params.), and layer 310 (21.8M params.).
- Distance Metric:** COSINE (selected), EUCLIDEAN, SQUARED_EUCLIDEAN.

Visualization of Embeddings (UMAP) for Extracted Features:

Top 15 results based on your search configuration:

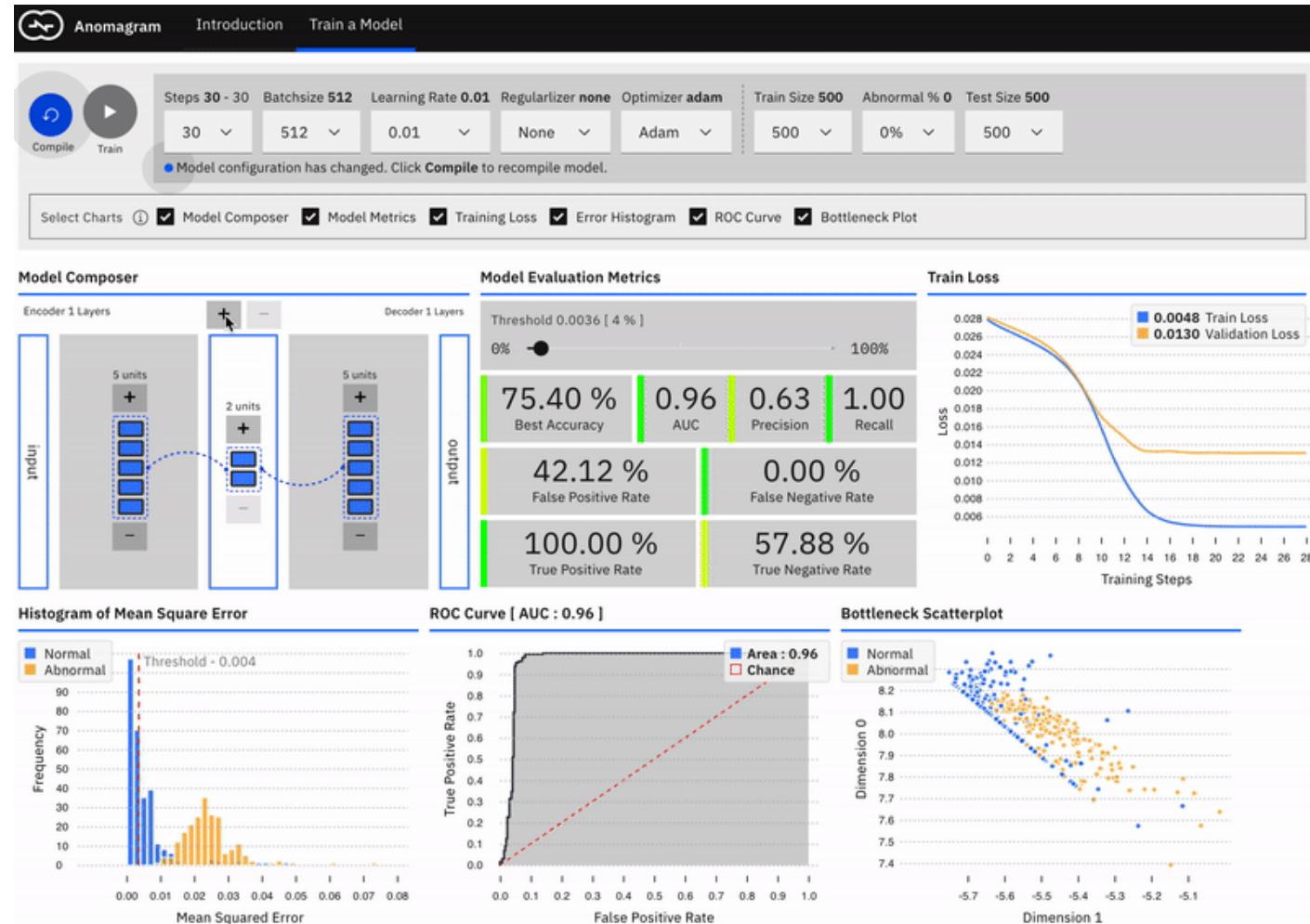
- SELECTED IMAGE:** BEETLE (A colorful Volkswagen Beetle).
- Search result score:** 90.0%.
- Similar Images:** A grid of 15 images showing various cars, with their distance scores listed below them: dst: 0.66, dst: 0.66, dst: 0.64, dst: 0.61, dst: 0.60, dst: 0.60, dst: 0.58, dst: 0.57, dst: 0.57, dst: 0.57, dst: 0.56.

Dataset Information: DATASET: [ICONIC200] A dataset of 200 images across 10 categories (20 images per category) crawled from the Flickr API.

Victor Dibia, GDE ML
@vykthur

4. Other helpful tools

<https://anomagram.fastforwardlabs.com/>



Victor Dibia, GDE ML
@vykthur

THANK YOU



Dr. Beril Sirmacek



@BerilSirmacek



Newsletter, YouTube

www.BerilSirmacek.com