TOWARDS TRUSTWORTHY AI WITH

DR. BERIL SIRMACEK

GDG UK/IRELAND

August 2021

AGENDA

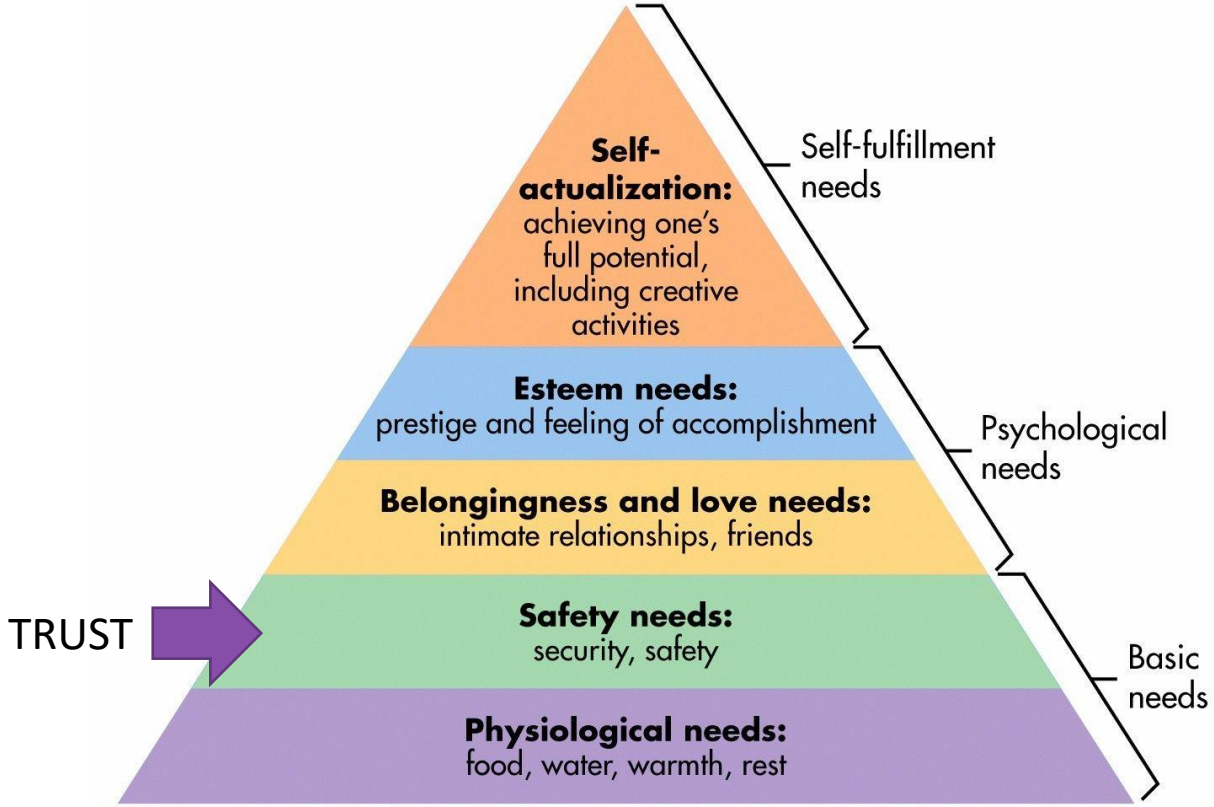
1. Our need for trust
2. TrustworthyAI
3. Towards ResponsibleAI with Google tools
4. Other helpful tools

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?

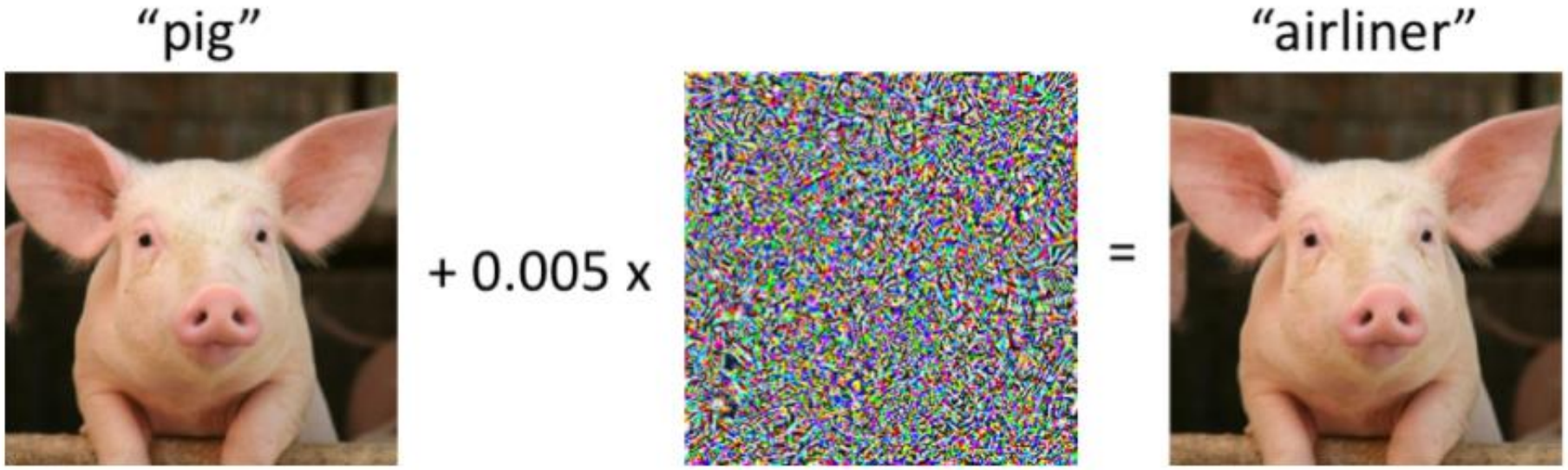
Decision

Why?

1. Our need for trust

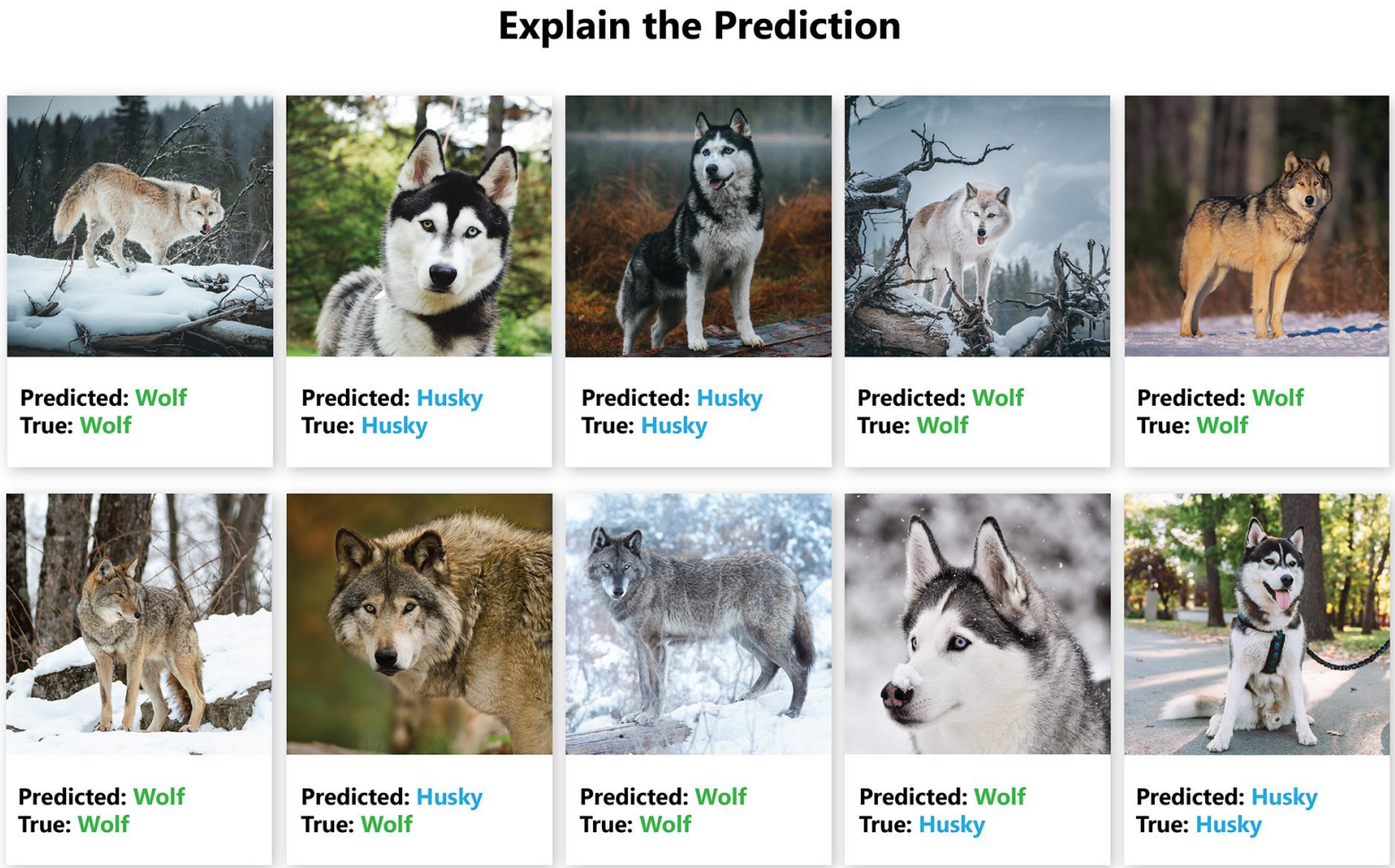
https://www.technologyreview.com/2013/02/04/253879/racism-is-poisoning-online-ad-delivery-says-harvard-professor/

1. Our need for trust

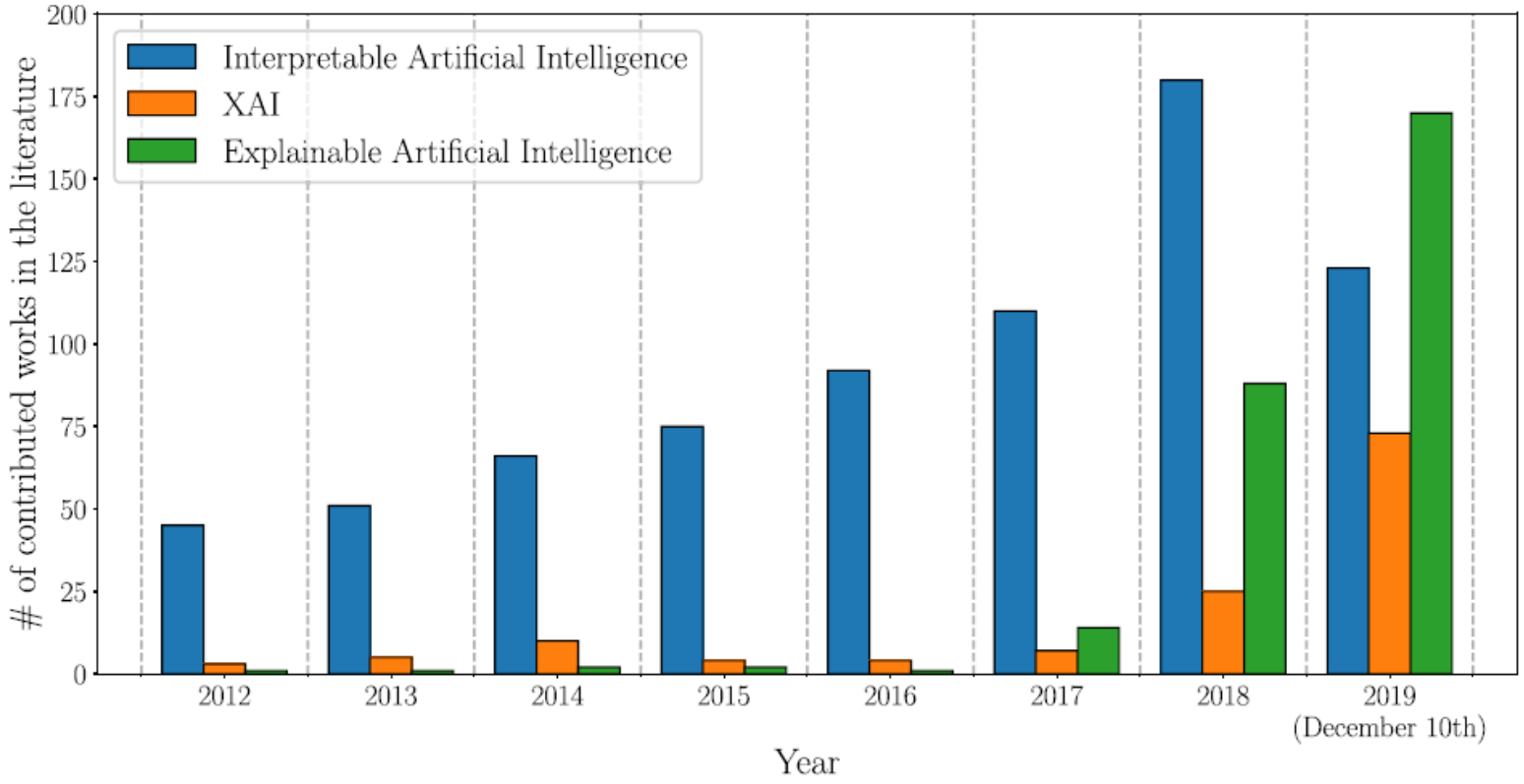


https://doi.org/10.1145/1014052.1014066

1. Our need for trust



1. Our need for trust

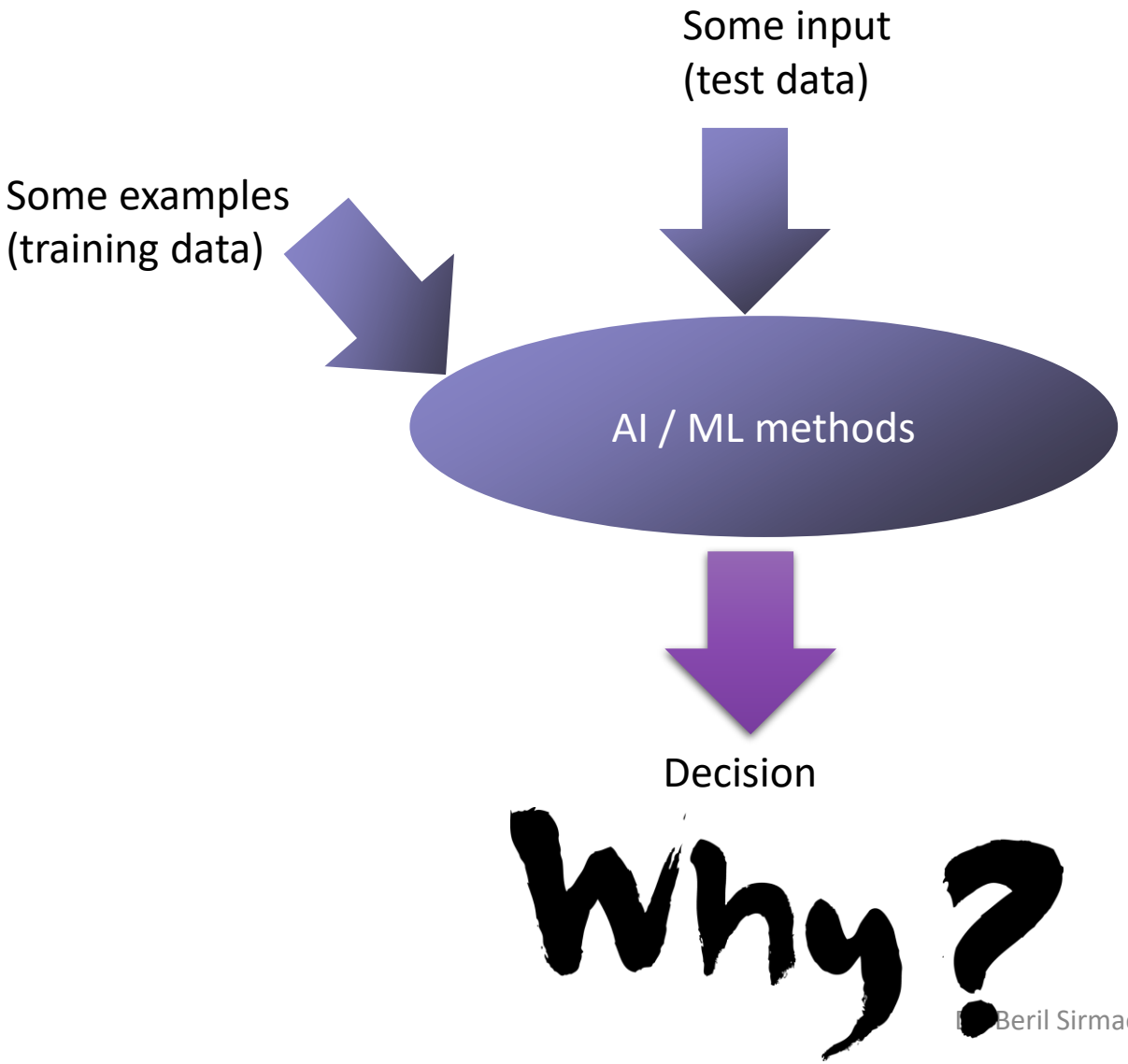


Source: "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI"

2. TrustworthyAI

LET'S DIVE IN

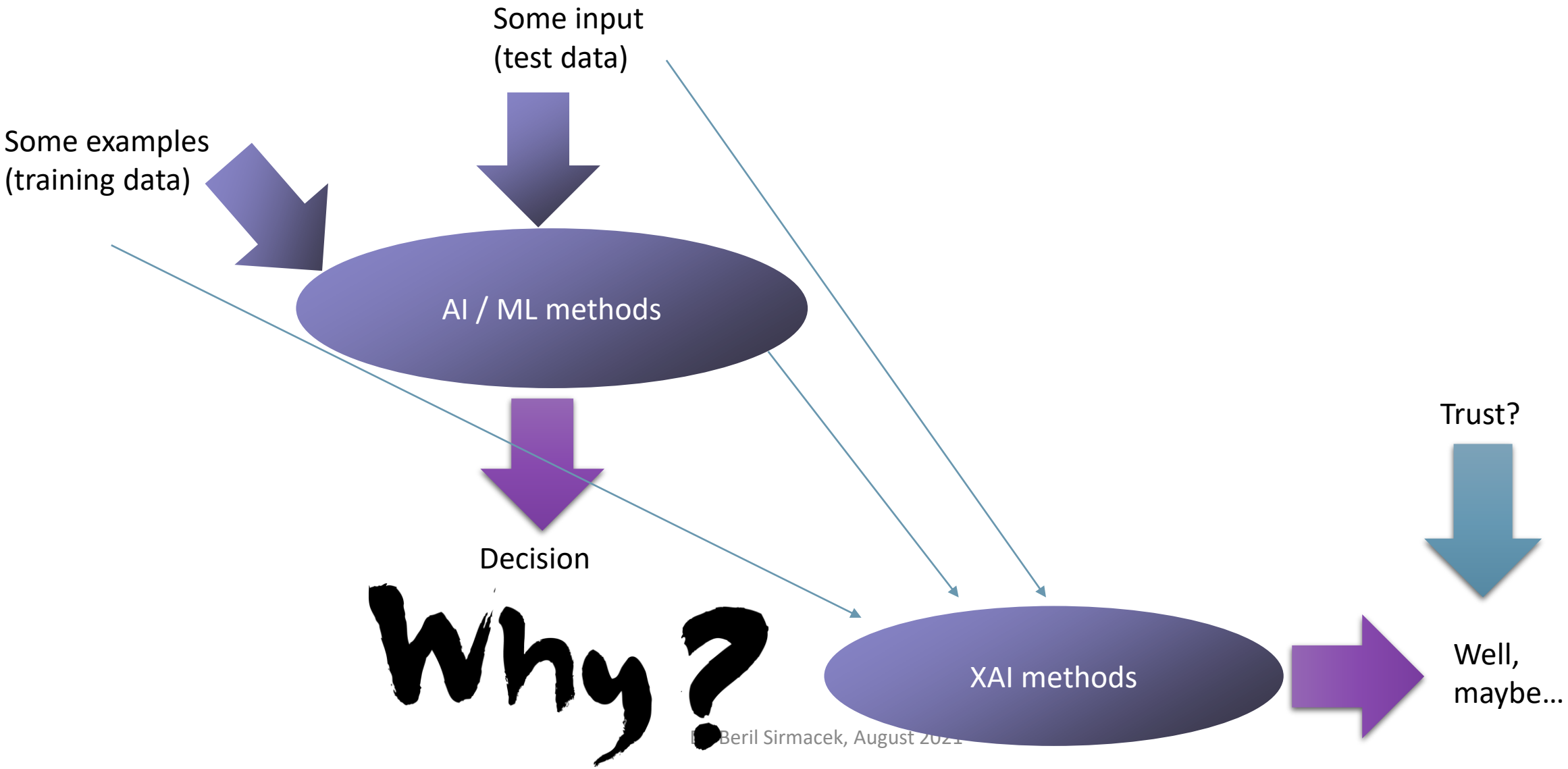
2. TrustworthyAI



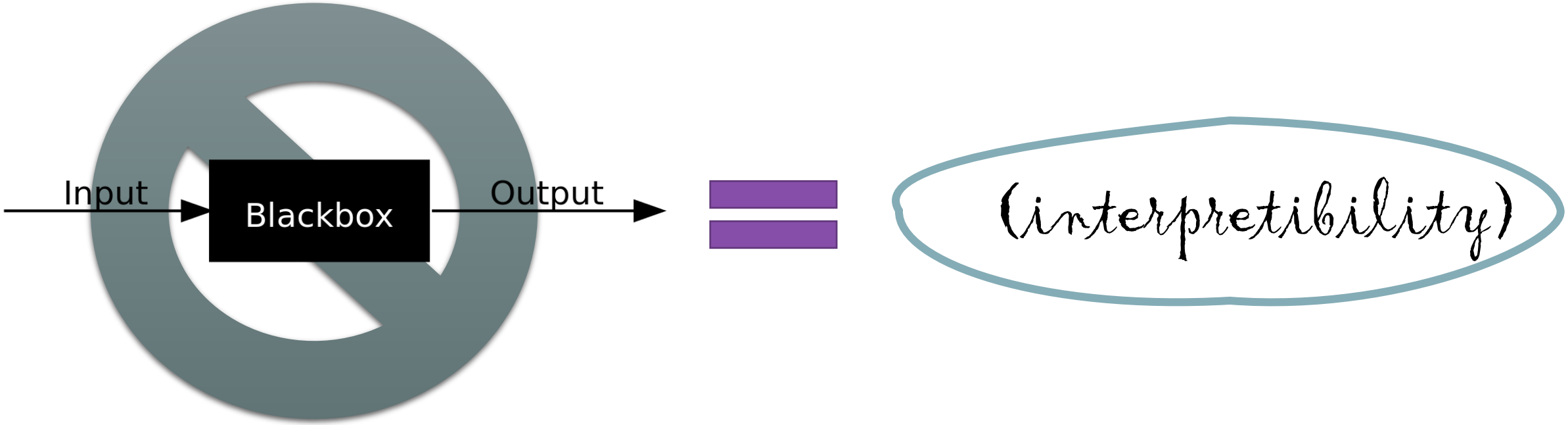
”Given a certain audience, ***explainability*** refers to the ***details*** and ***reasons*** a model gives to make its functioning clear or easy to understand.”

”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*

2. TrustworthyAI

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

At different levels TRANSPARENCY can be focused on.

2. TrustworthyAI

Summary of the interpretibility 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)

2. TrustworthyAI

SHAP (SHapley Additive exPlanations)

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

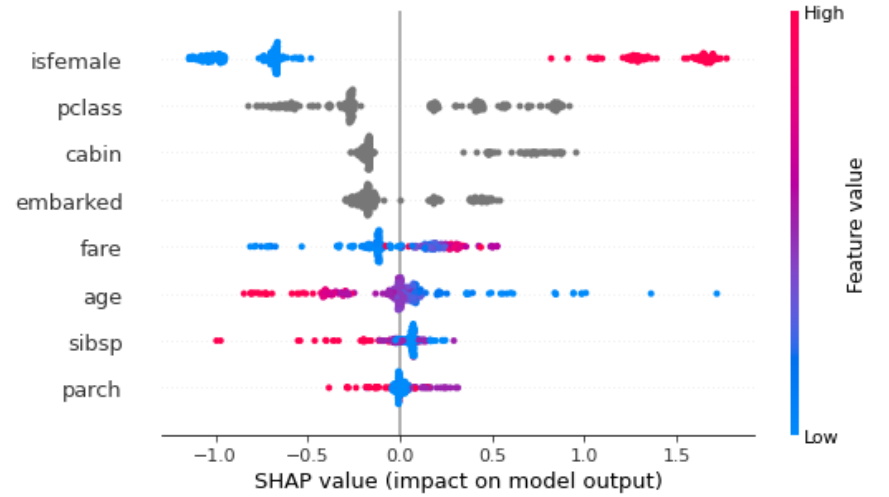
SHAP

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
```

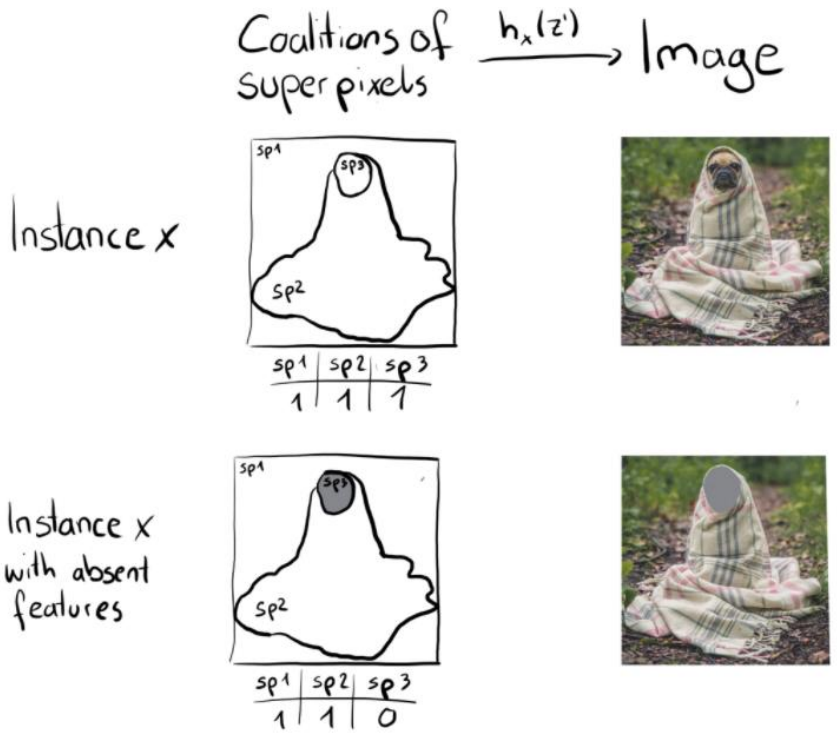
```
shap.summary_plot(shap_values, X_test)
```



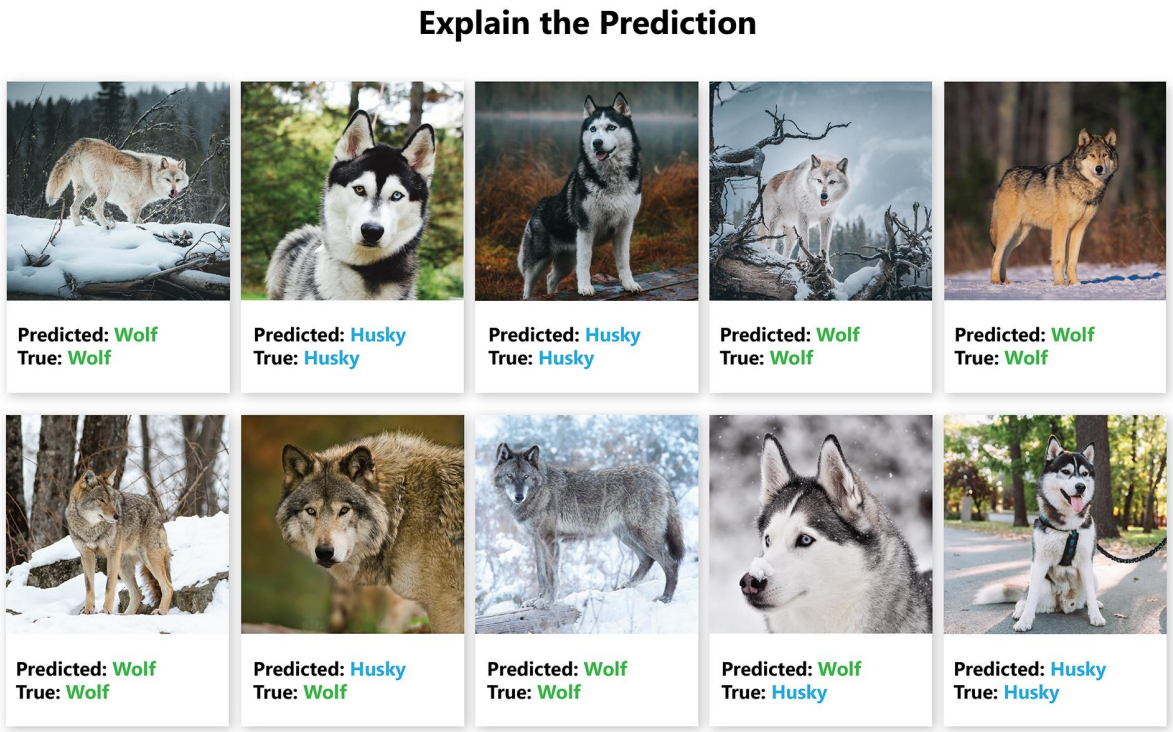
2. TrustworthyAI

LIME (Local Interpretable Model-Agnostic Explanations)

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



2. TrustworthyAI



(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

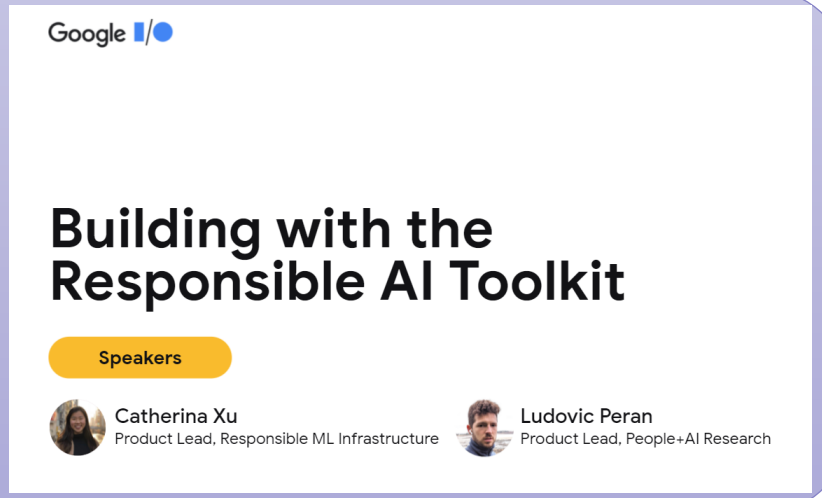
TOWARDS RESPONSIBLEAI

Dr. Beril Sirmacek, August 2021

3. Towards ResponsibleAI

LOOKING AHEAD

The following Google workflow is edited from:

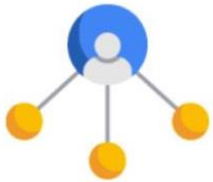


Google's AI Principles

Dr. Beril Sirmacek, August 2021

Google's AI Principles

1. Be socially beneficial.

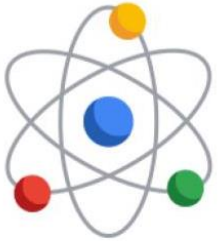


2. Avoid creating or reinforcing unfair bias.

3. Be built and tested for safety.

4. Be accountable to people.

5. Incorporate privacy design principles.

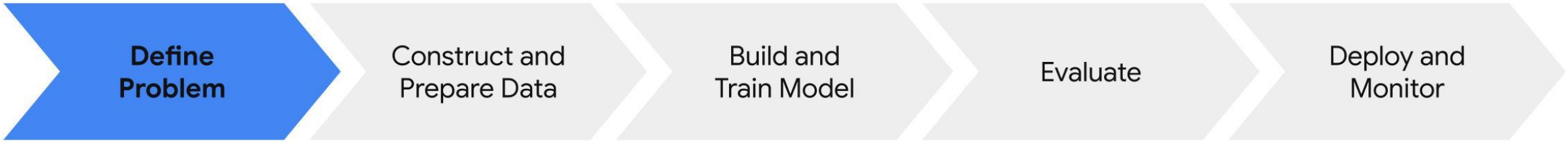


6. Uphold high standards of scientific excellence.

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

The typical machine learning workflow

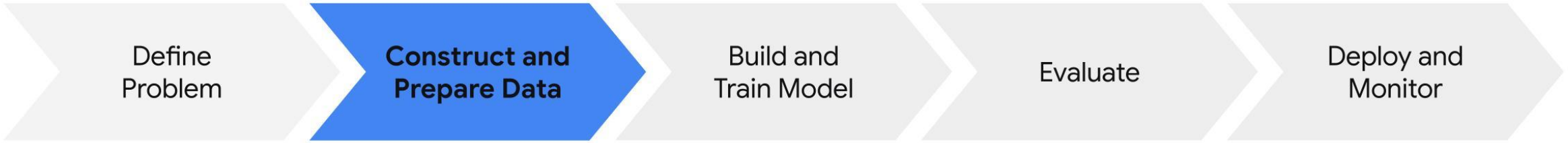
Responsible AI Considerations



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?

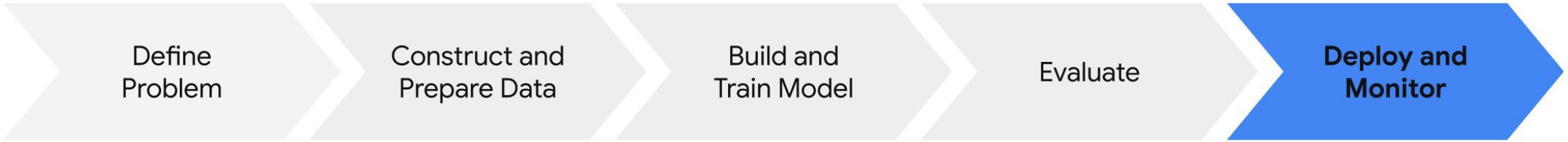
Responsible AI Considerations

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

Responsible AI Considerations

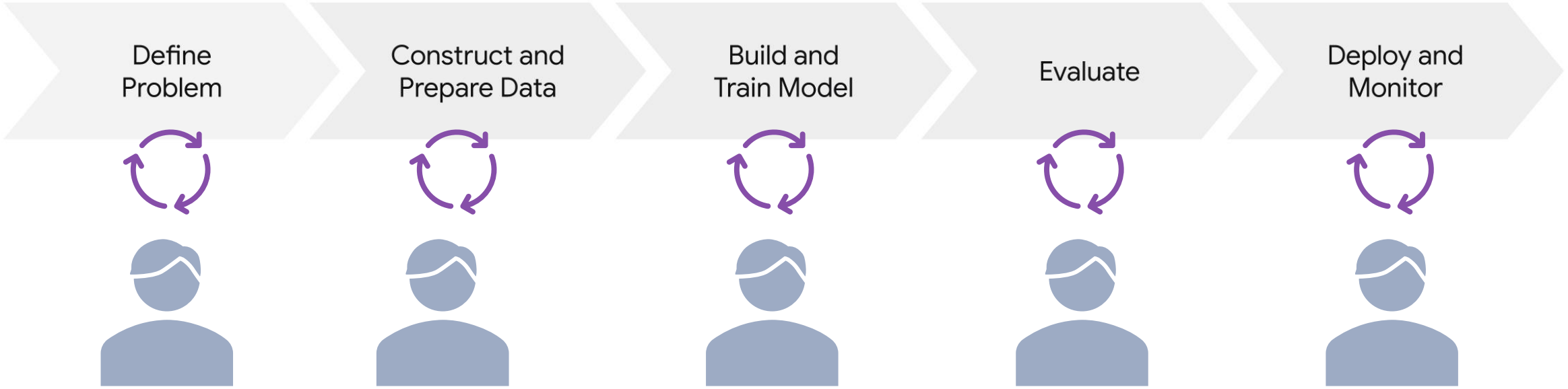
What metrics should I use to evaluate my model?

Responsible AI Considerations

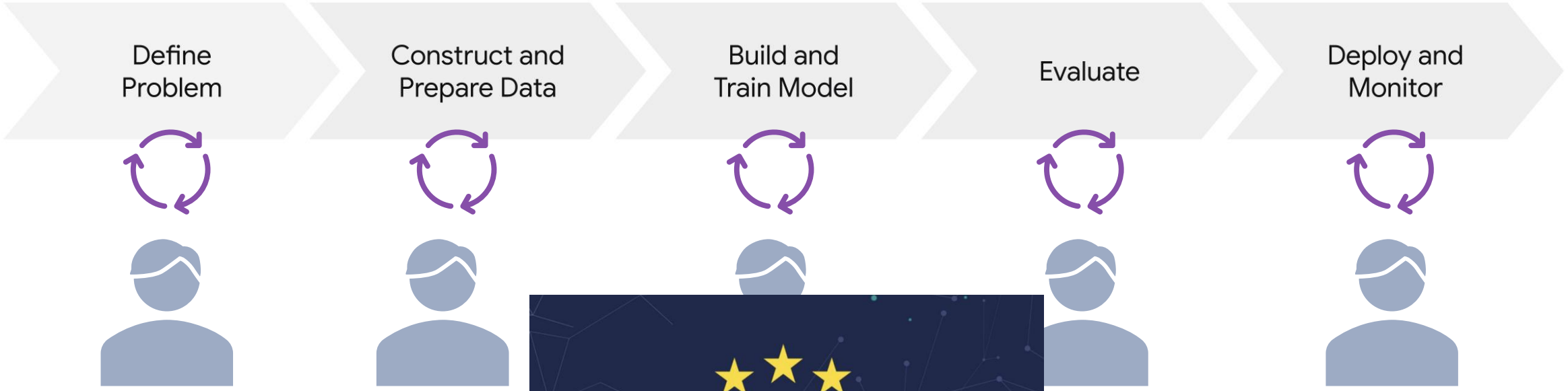


What should key stakeholders know about my model?

The typical machine learning workflow



The typical machine learning workflow

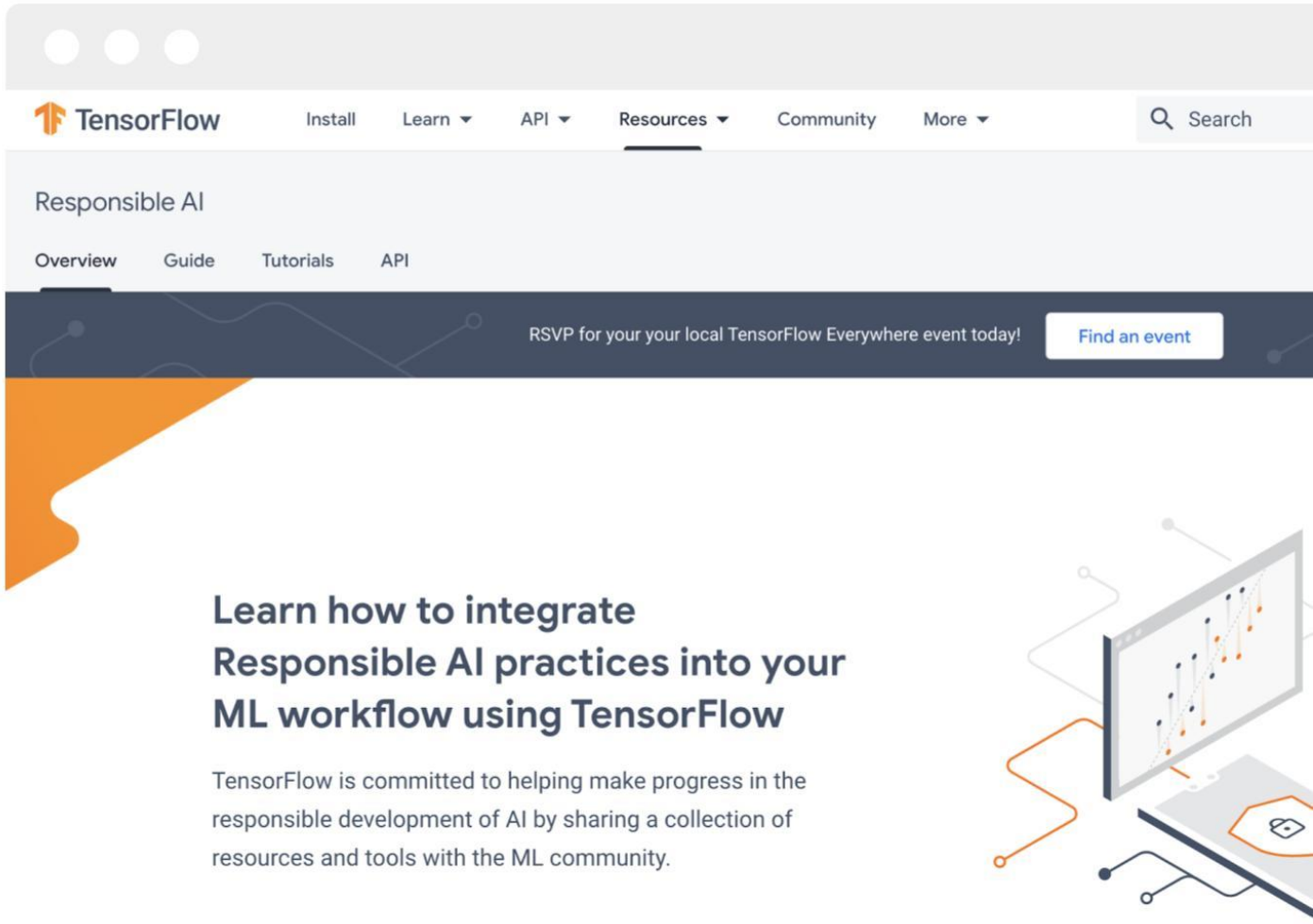


https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Introducing: The Responsible AI Toolkit

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

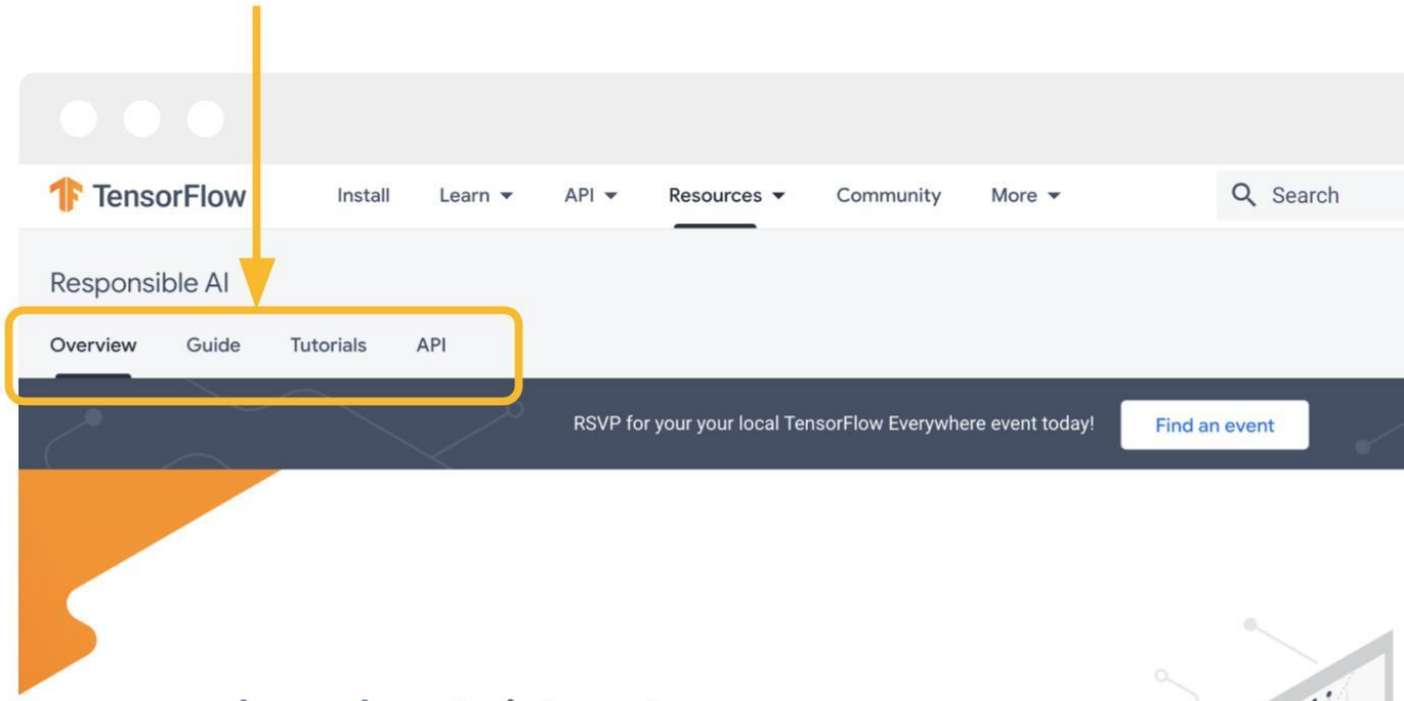
www.tensorflow.org/responsible_ai



Introducing: The Responsible AI Toolkit

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

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



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.

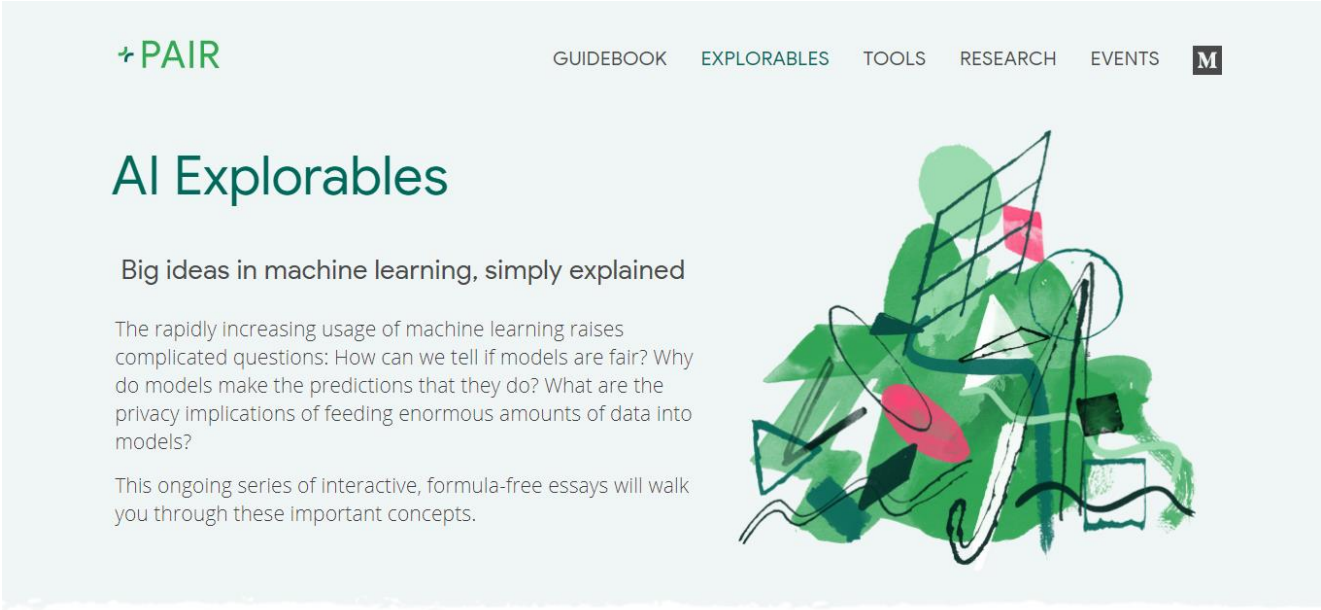
www.tensorflow.org/responsible_ai

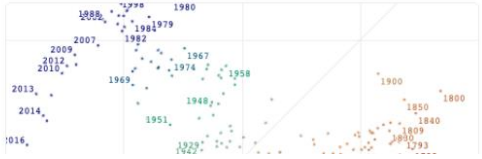
The typical machine learning workflow

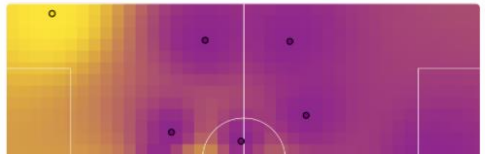
Define Problem → Construct and Prepare Data → Build and Train Model → Evaluate → Deploy and Monitor

People + AI guidebook

AI explorables







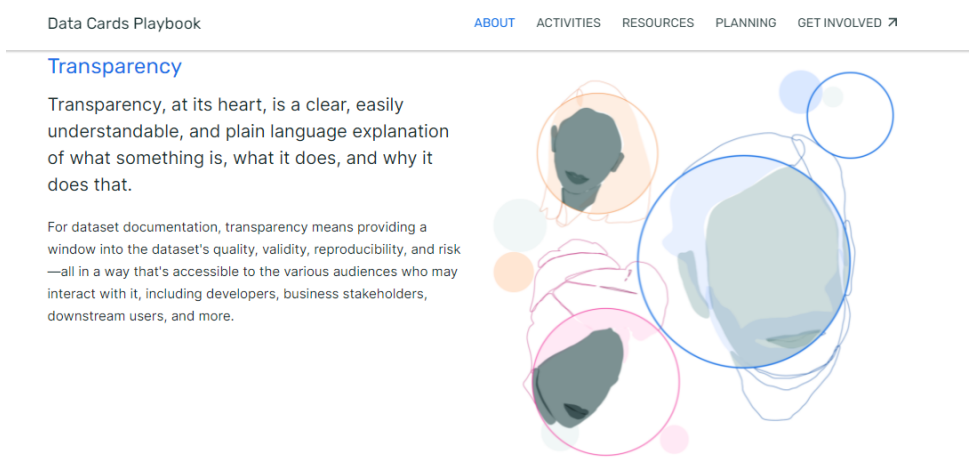
The typical machine learning workflow

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

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.

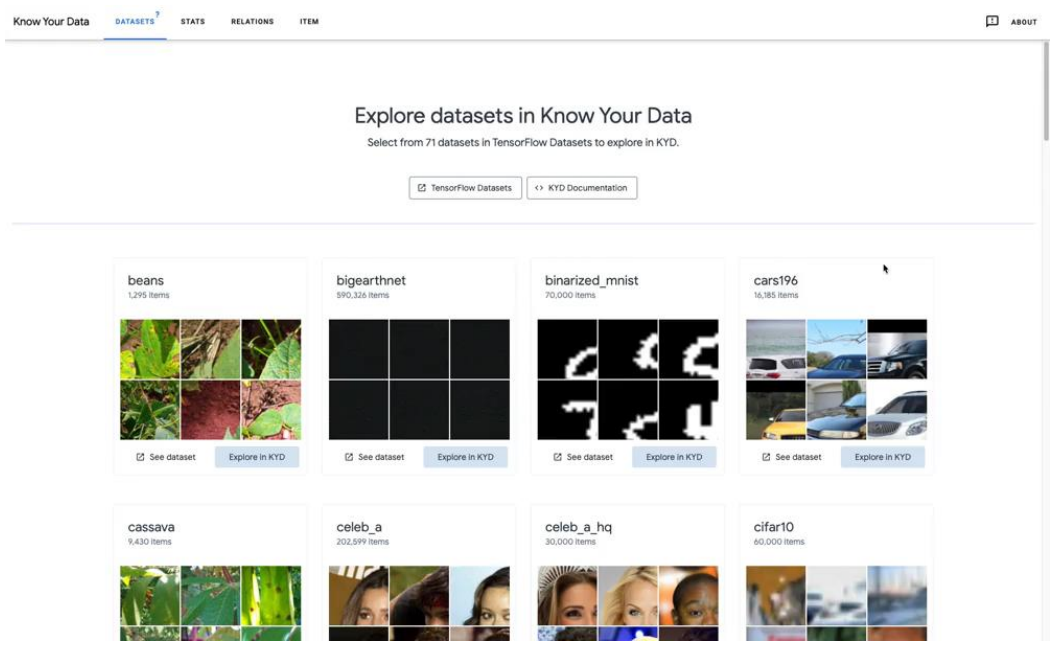
The typical machine learning workflow

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

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

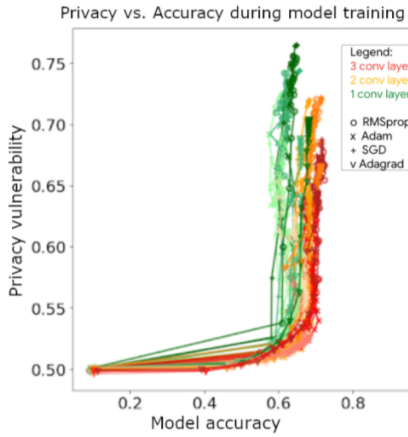
Measuring Privacy

Coming soon on TensorFlow Extended!

Performs *privacy attacks* against your model

Returns a score showing how your model performed with these attacks

Informs your development process and uncover potential privacy failures



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.

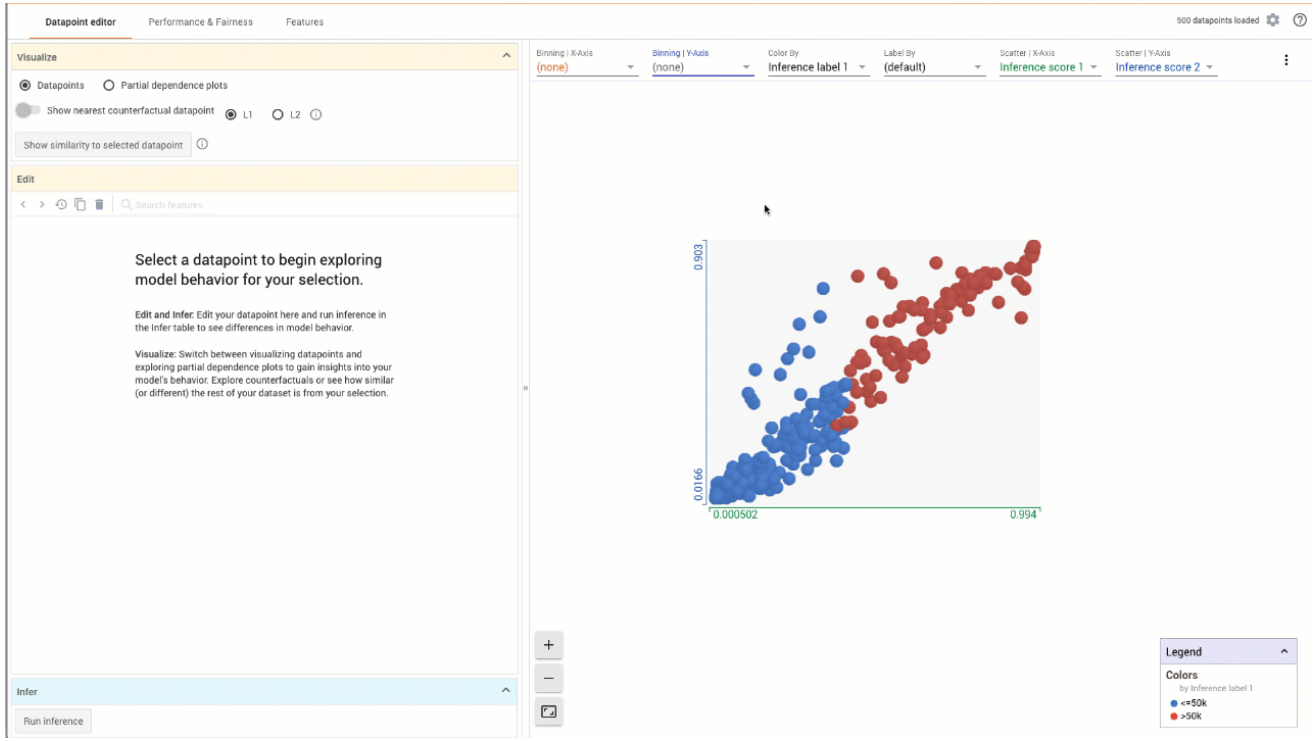

```
import numpy as np
import tensorflow as tf
import tensorflow_lattice as tfl

model = tf.keras.models.Sequential()
model.add(
    tfl.layers.ParallelCombination([
        # Monotonic piece-wise linear calibration with bounde
        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', 'increasi
        # Trust: model is more responsive to input 0 if input
        edgeworth_trusts=(0, 1, 'positive')))
model.compile(...)
```

The typical machine learning workflow

- TF privacy
- TF federated
- TF lattice
- TF constraint optimization
- TF model remediation

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

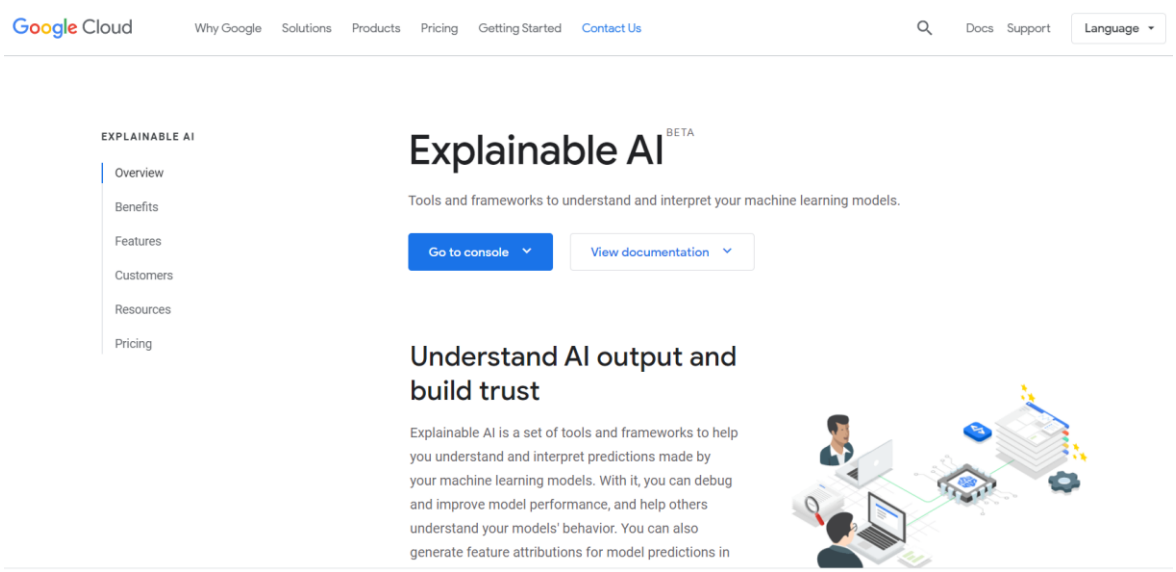
Recommended Lecture:
https://youtu.be/u96F9BUfH3o

The typical machine learning workflow

Model Card toolkit

The typical machine learning workflow

Define Problem → Construct and Prepare Data → Build and Train Model → Evaluate → Deploy and Monitor



Model Card toolkit

https://github.com/GoogleCloudPlatform/explainable_ai_sdk

4. Other helpful tools

Grad-Cam

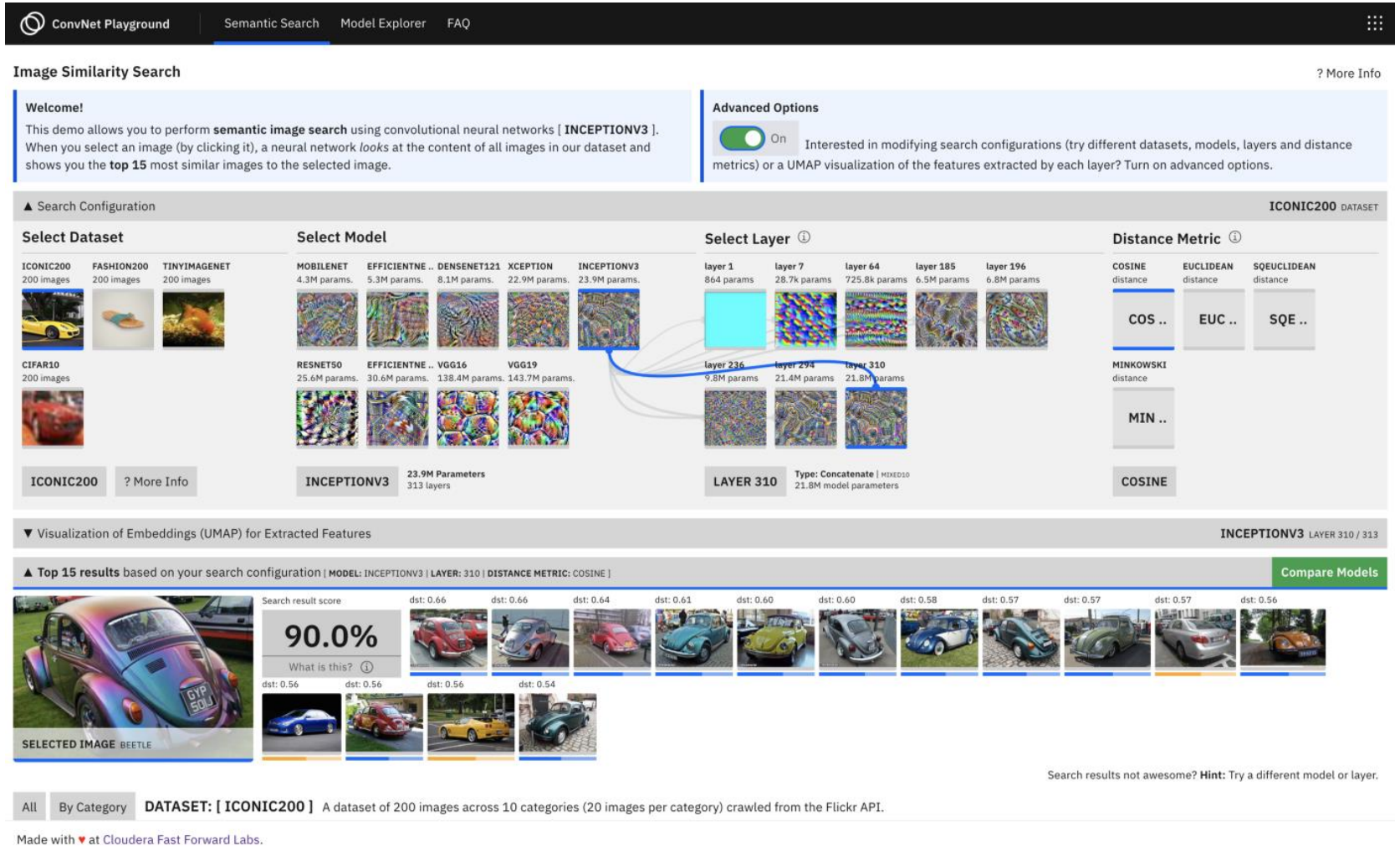
https://keras.io/examples/vision/grad_cam/

Predicted: ['Border_collie']

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

4. Other helpful tools

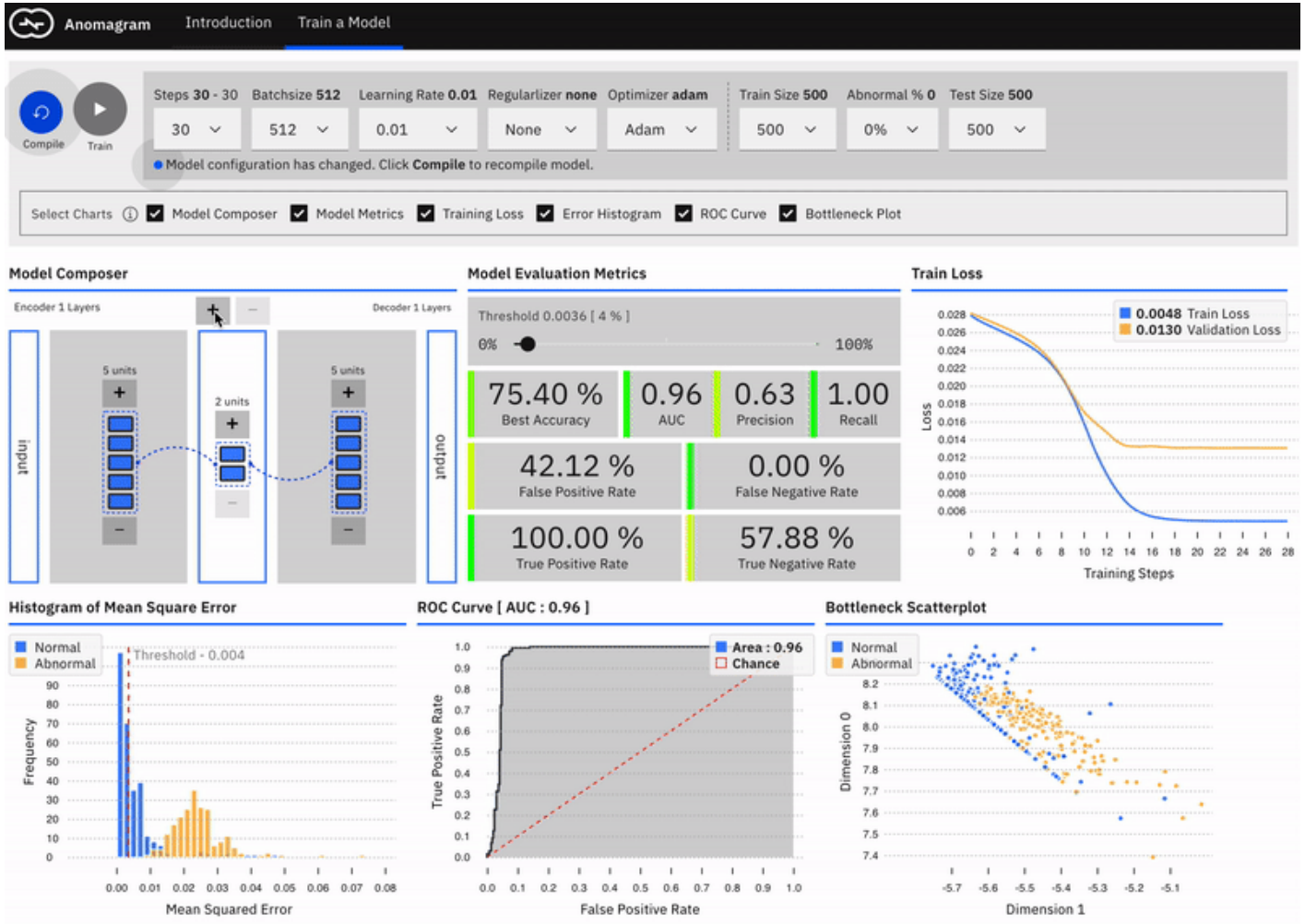
https://convnetplayground.fastforwardlabs.com/



Victor Dibia, GDE ML
@vykthur

4. Other helpful tools

https://anomagram.fastforwardlabs.com/



Victor Dibia, GDE ML
@vykthur

Dr. Beril Sirmacek, August 2021

THANK YOU

Dr. Beril Sirmacek

@BerilSirmacek

Newsletter, YouTube

www.BerilSirmacek.com