LFAI & Data
Webinar - Trusted AI Principles – Tools and Techniques

Trusted AI Committee - Principles Working Group (where you will find the slides and materials)

Register https://tinyurl.com/trustedAI

27 October 2021

LF AI & DATA

Trusted AI Principles - Tools and Techniques

Join us at 10am US Eastern October 27, 20211 to meet with Souad Ouali Chair of the Trusted AI Principles Working Group at the LF-AI & Data & members of the Working Group - Hear about tools and techniques for the RREPEATS Principles

Register : https://tinyurl.com/trustedAI

Souad Ouali, Head of interoperators relationships Orange
Co-Chair Trusted AI Committee, LF-AI

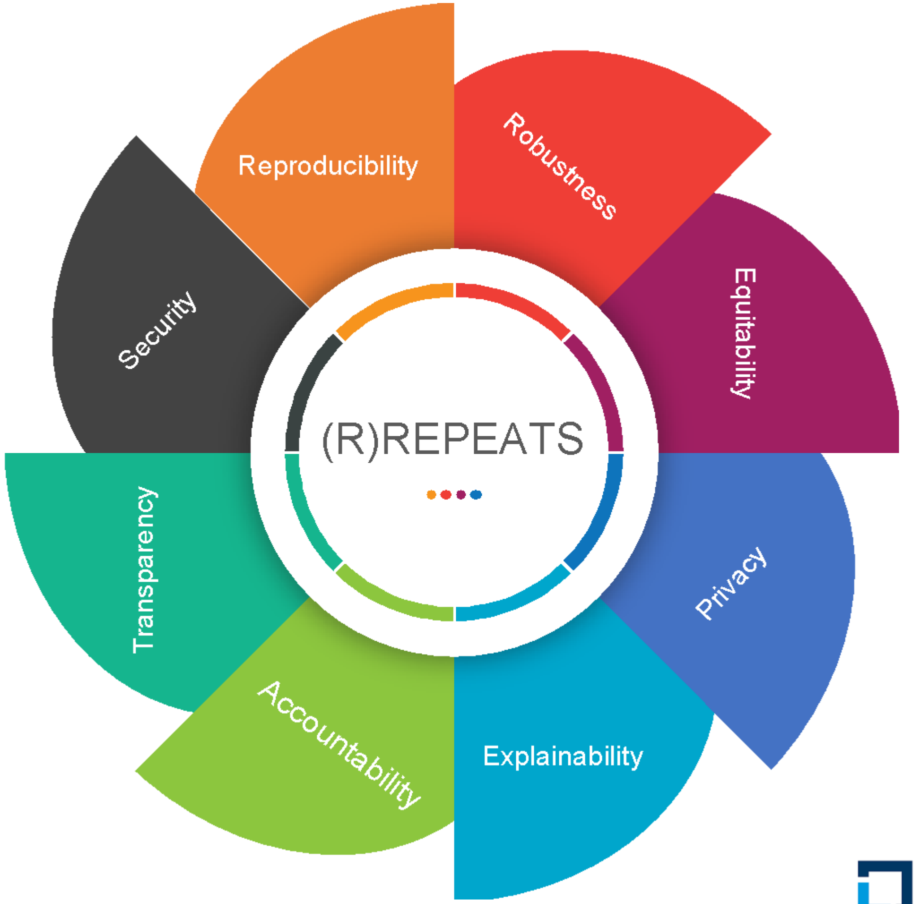
Layla Li, Co-founder & CEO, KOSA

Sarah Luger, Leading strategic AI/ML/NLP startups & technologies engagement, Orange

Sri Krishnamurthy, CEO, QuantUniversity

Animesh Singh, Distinguished Engineer, IBM
Co-Chair Trusted AI Committee, LF-AI

François Jézéquel, Director of Business Development, Orange Fab



LF AI & DATA

Agenda

- Opening & intro to the session, Souad Ouali, Orange
- The RREPEATS Principles Overview, Layla Li, KOSA; Sarah Luger, Orange
- Operationalizing Trusted AI in Finance using the QuSandbox, Sri Krishnamurthy, QuantUniversity
- Trusted AI Tools (AI Fairness, AI Explainability, Adversarial Robustness etc) and RREPEATS, Animesh Singh, IBM
- Emerging DataOps activities at the LF-AI, Trusted AI and RREPEATS, Animesh Singh, IBM
- Summary & Call to Action, François Jézéquel, Orange

Session Host: Souad Ouali
Head of interoperators relationships Orange - Counsel /
Responsable de relations inter opérateurs chez Orange - Conseil

LFAI & Data

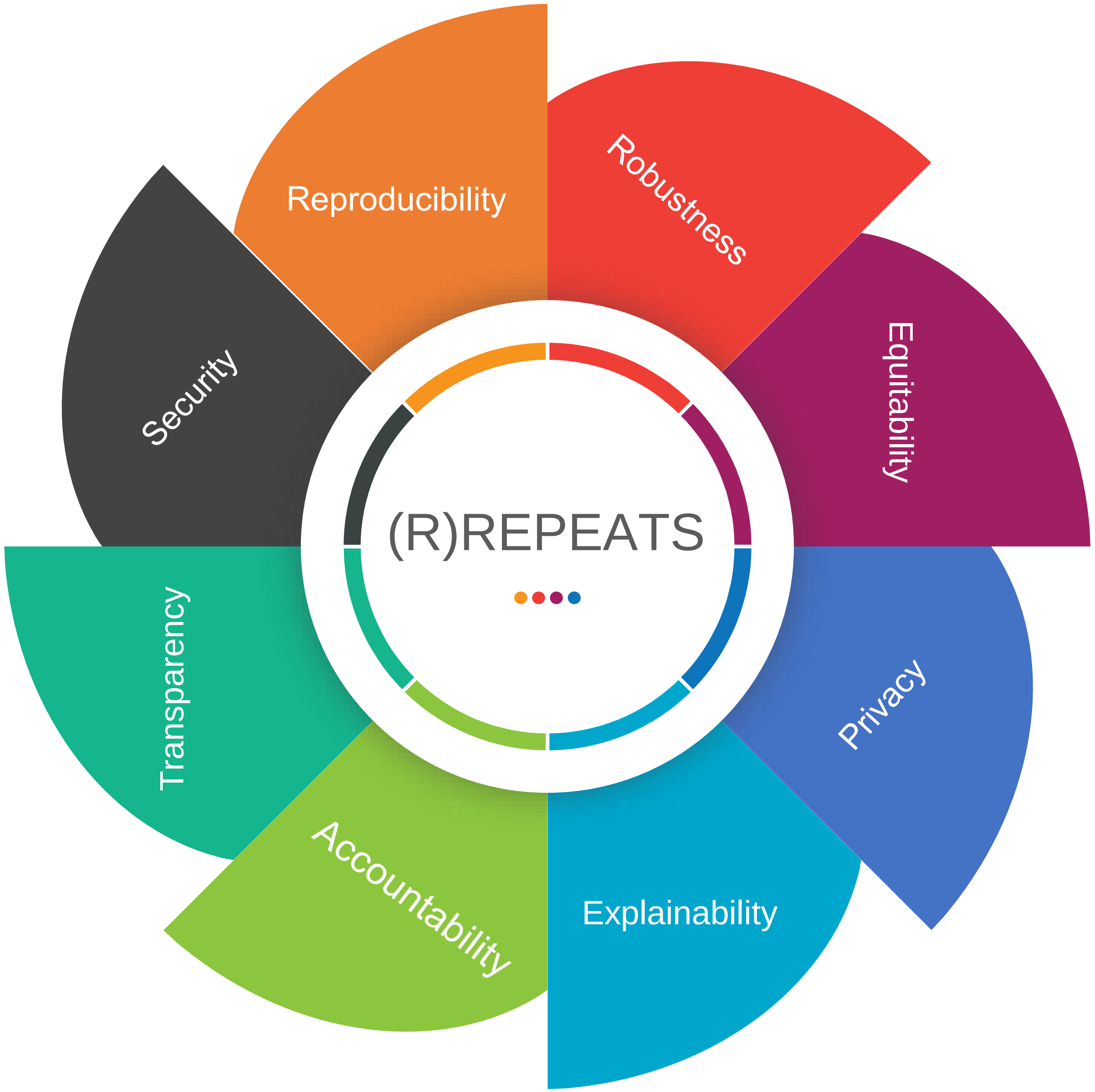
The Trusted AI Principles - Tools and Techniques

The RREPEATS Principles Overview

Layla Li, Co-founder & CEO, KOSA

Sarah Luger, Leading strategic AI/ML/NLP startups & technologies engagement, Orange

LF AI & DATA

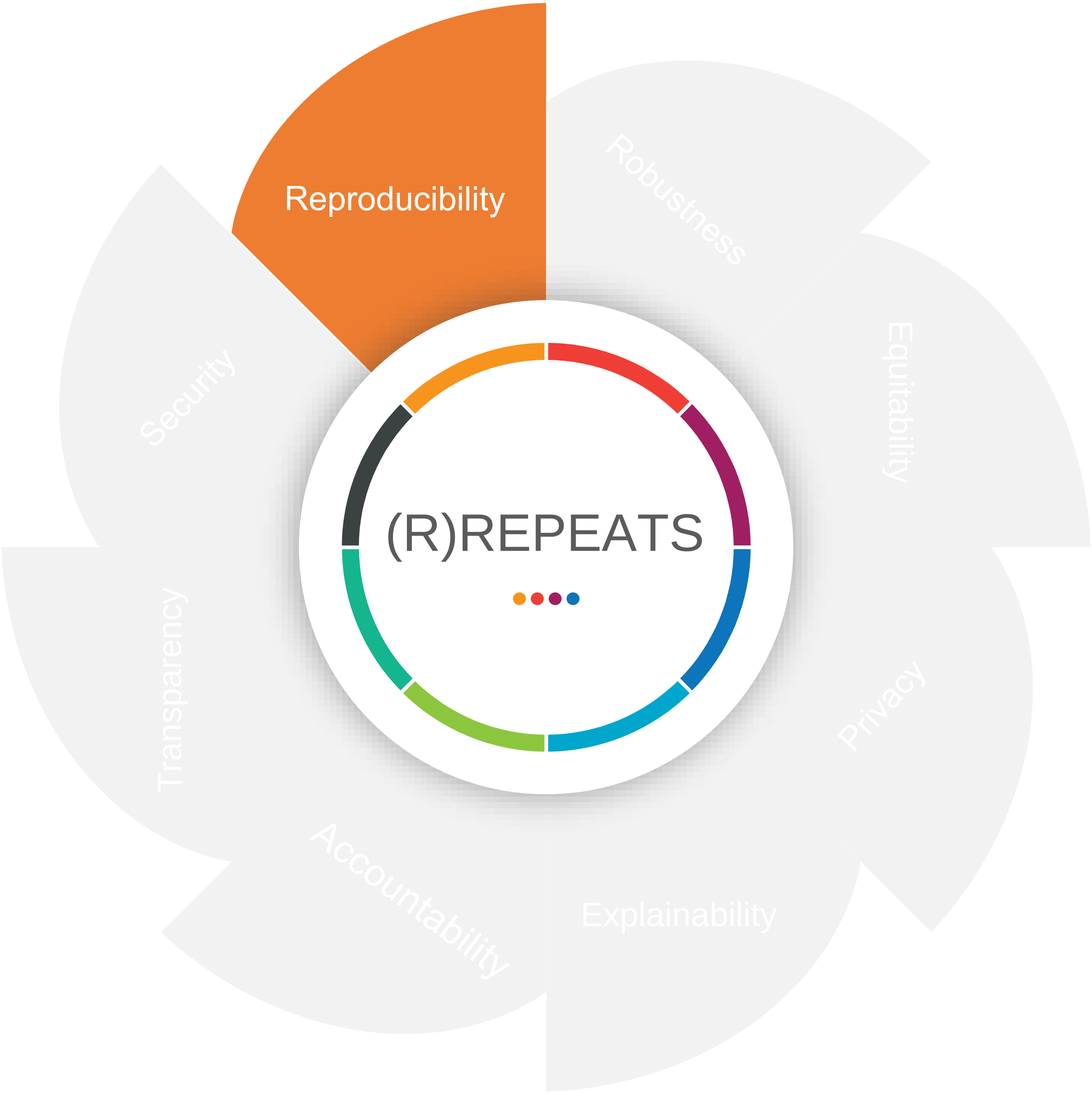


The 8 LFAI Principles for Trusted AI – (R)REPEATS

The principles are of equal importance and value.

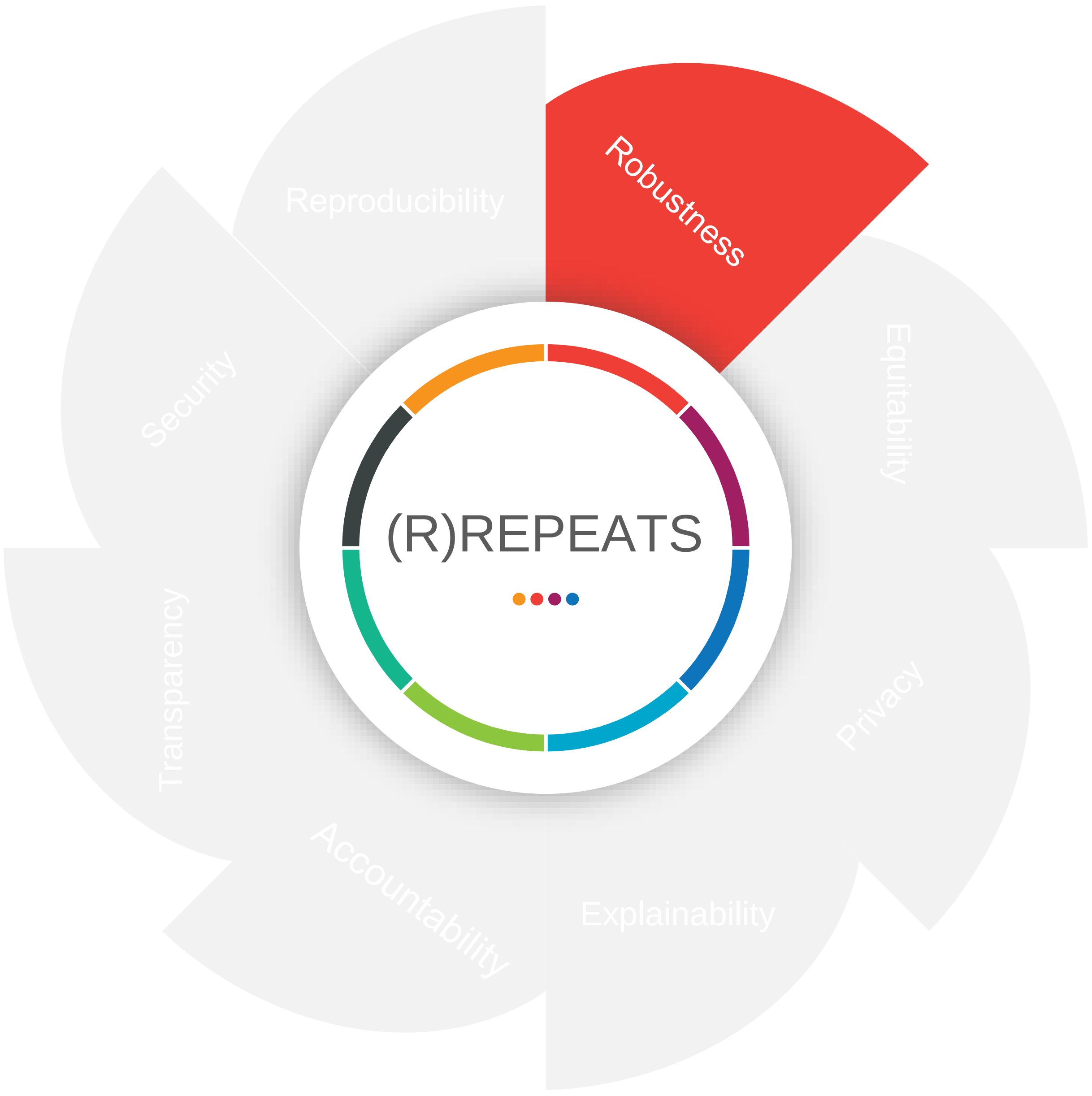
No principle is of higher priority than another.

The principles are related to each other.



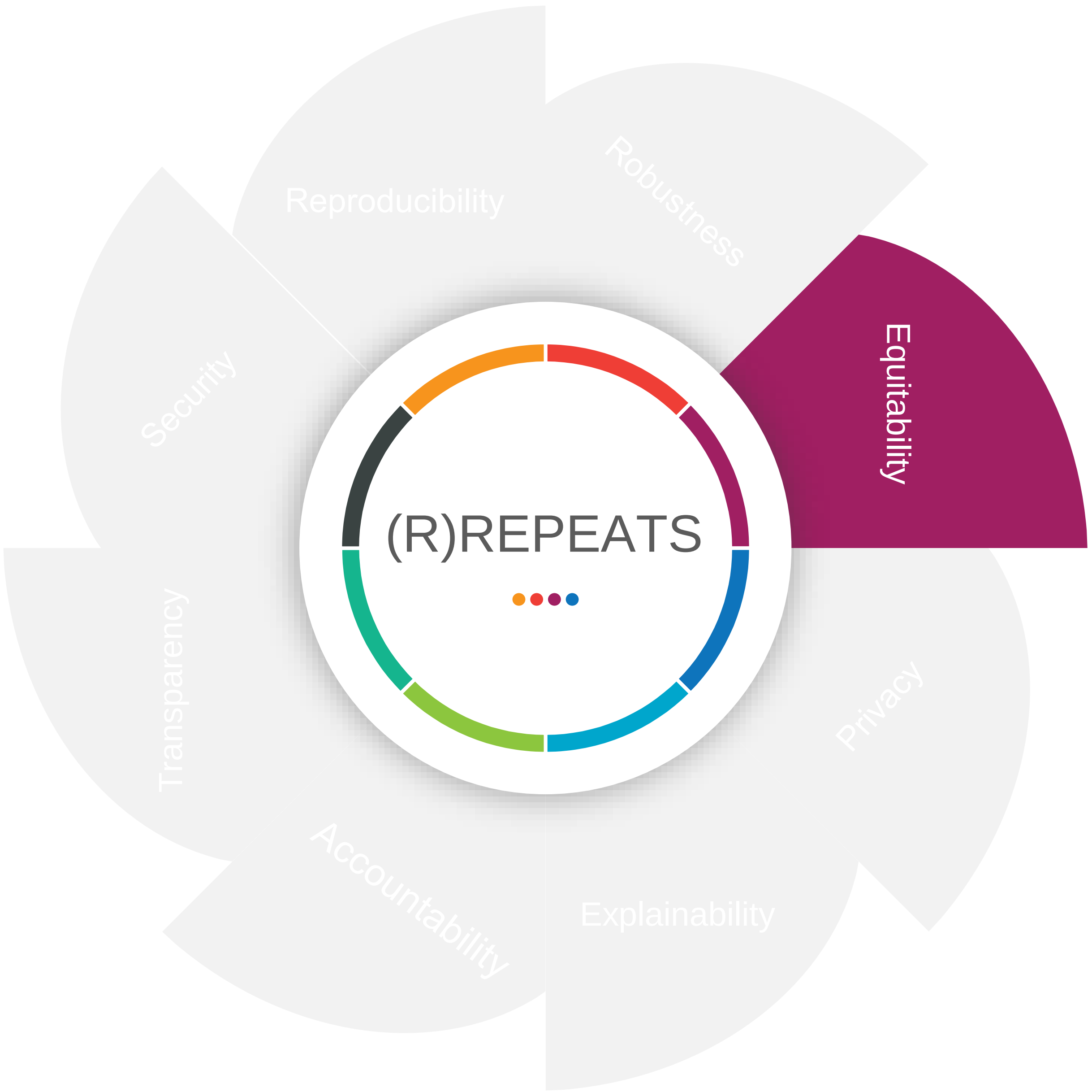
Reproducibility

Adhering to this principle will ensure the reliability of the results or experiences produced by any AI.



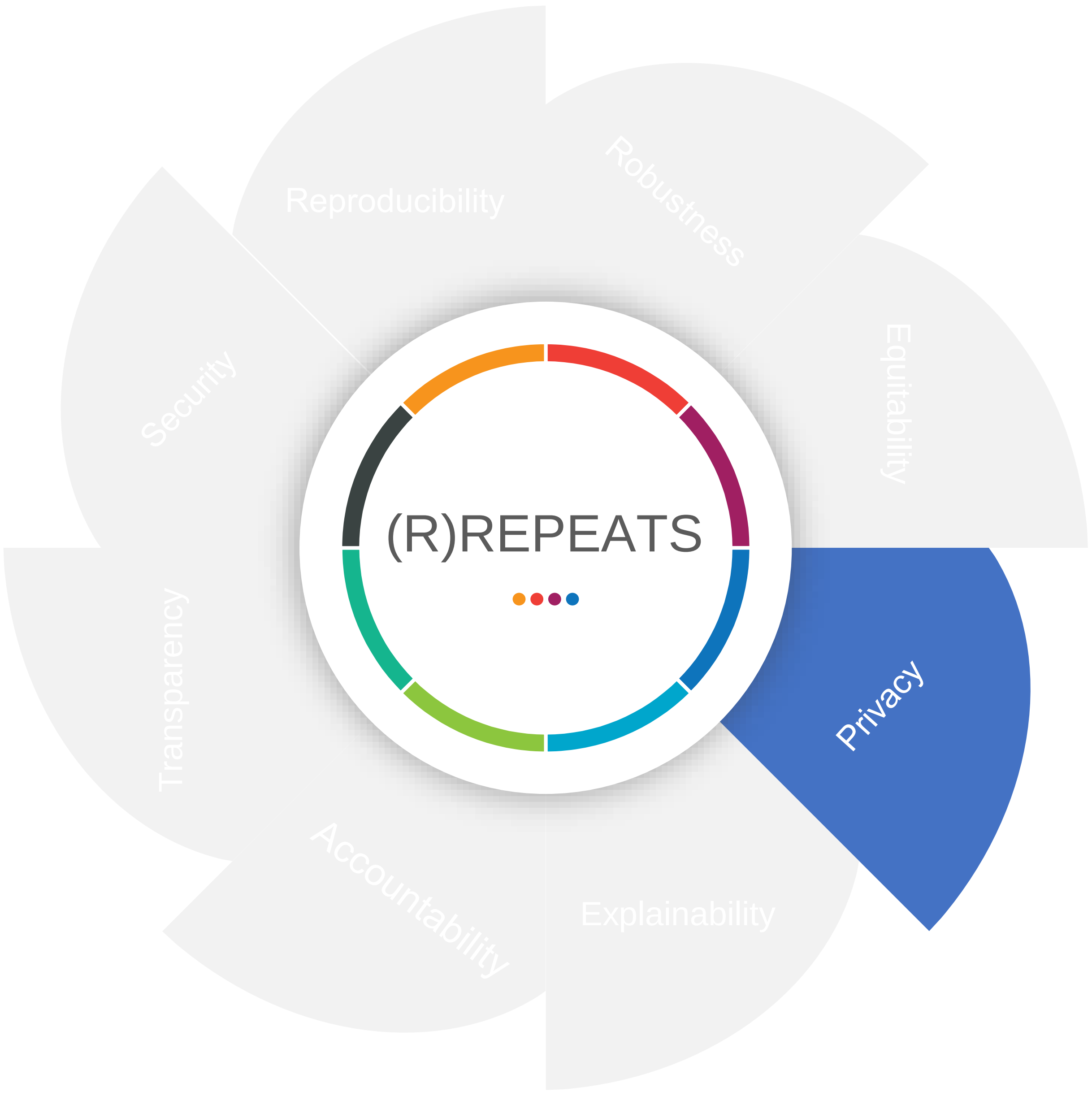
Robustness

Ensure stability, resilience, and performance of the systems in changing ecosystems.



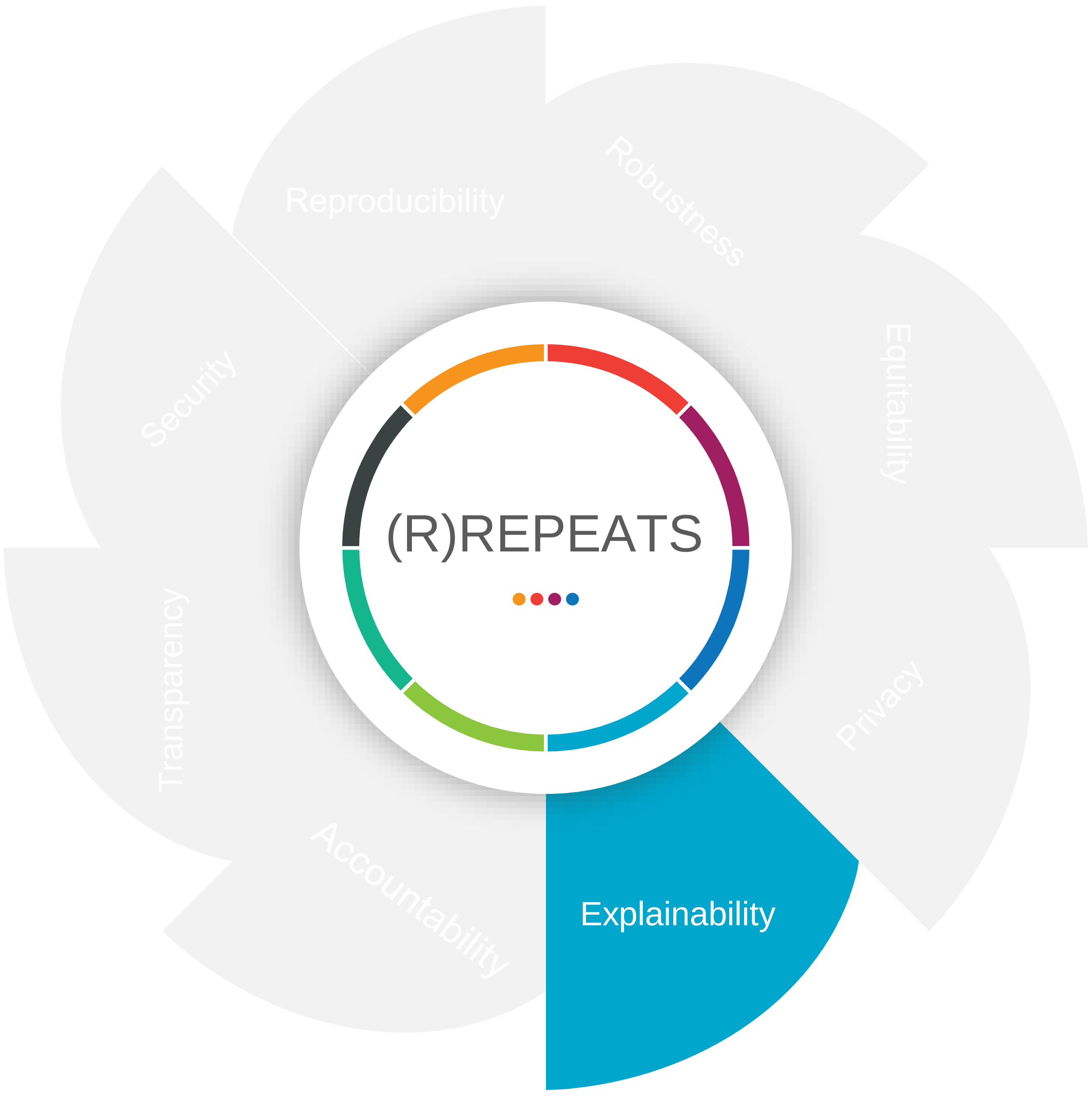
Equitability

Avoid intended or unintended bias and unfairness that would inadvertently cause harm



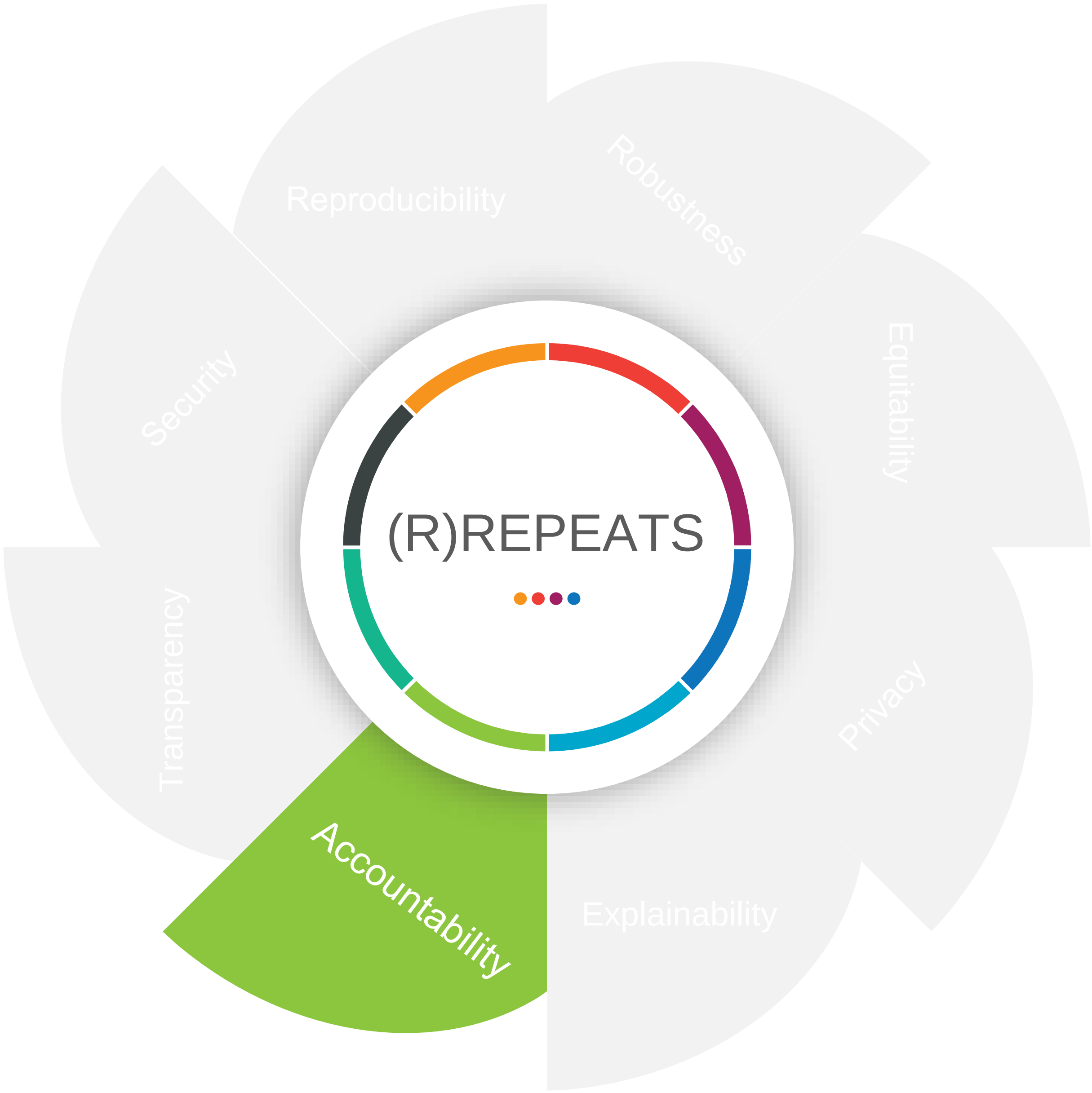
Privacy

Guarantee privacy and data protection throughout a system’s entire lifecycle.



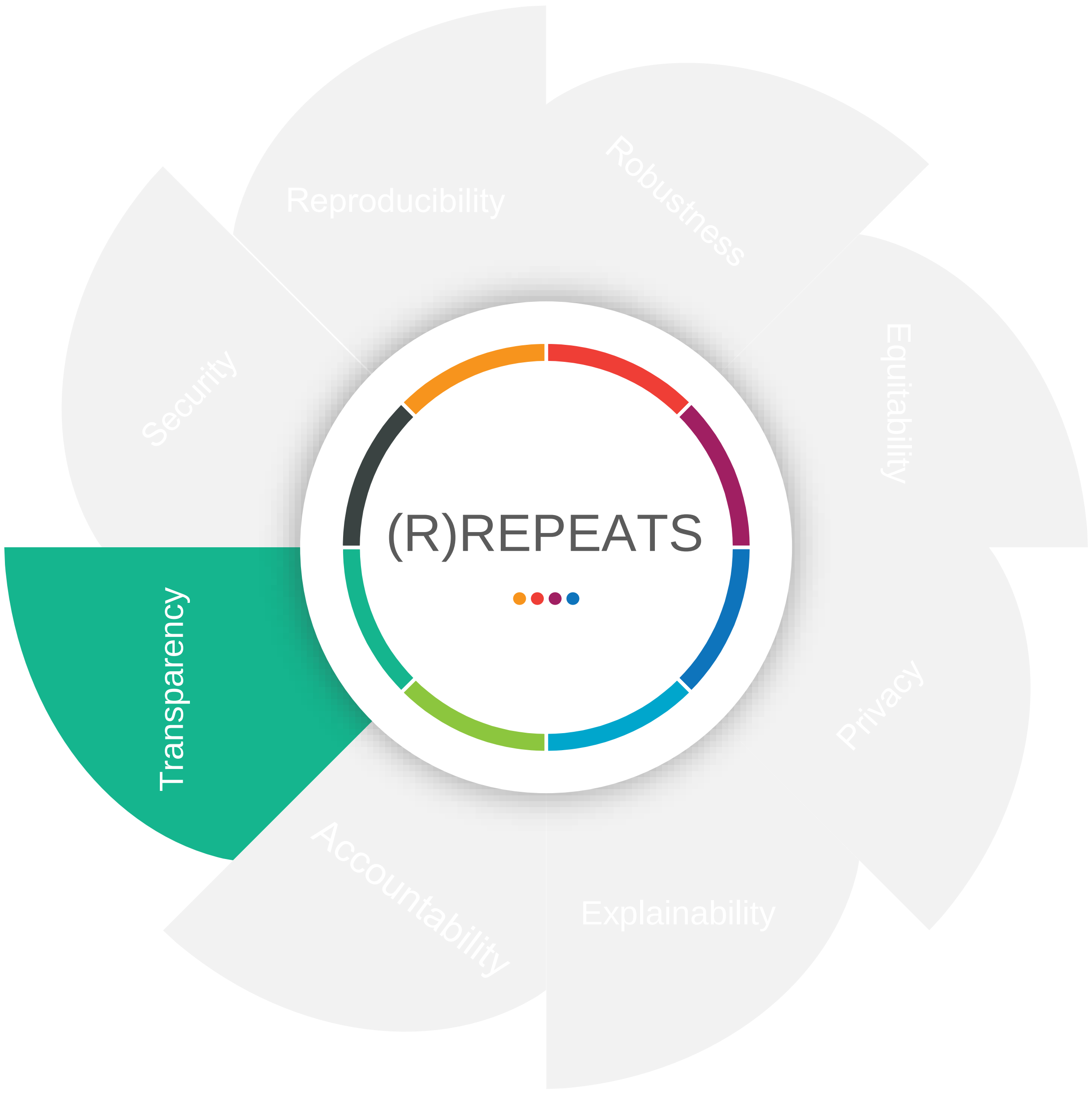
Explainability

Explain how AI “blackbox” make decisions in transparent manners



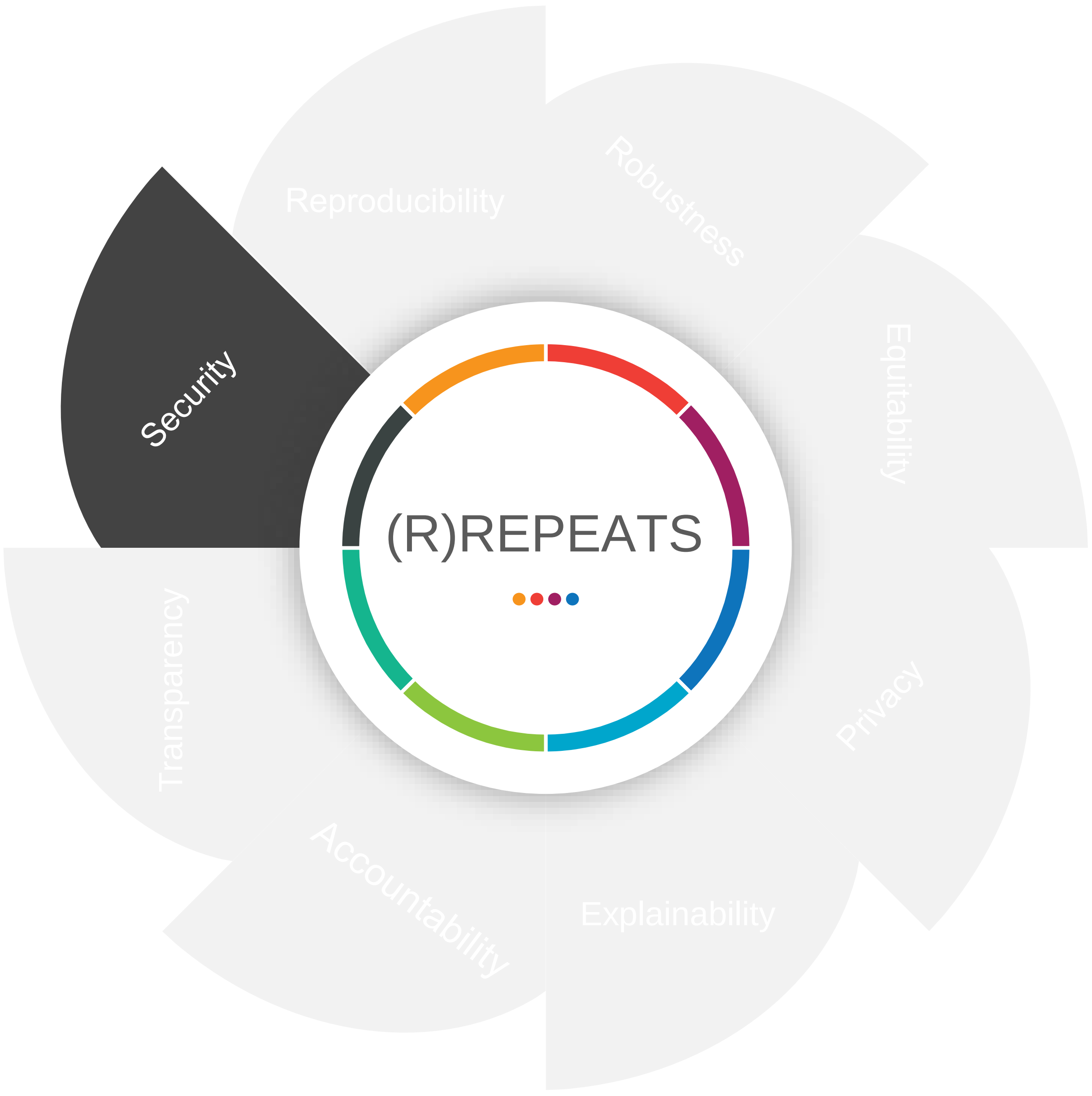
Accountability

“Humans-in-the-loop” to take responsibility and plan for actions of AI



Transparency

Users should be informed of when they interact with AI and understand AI-based outcomes



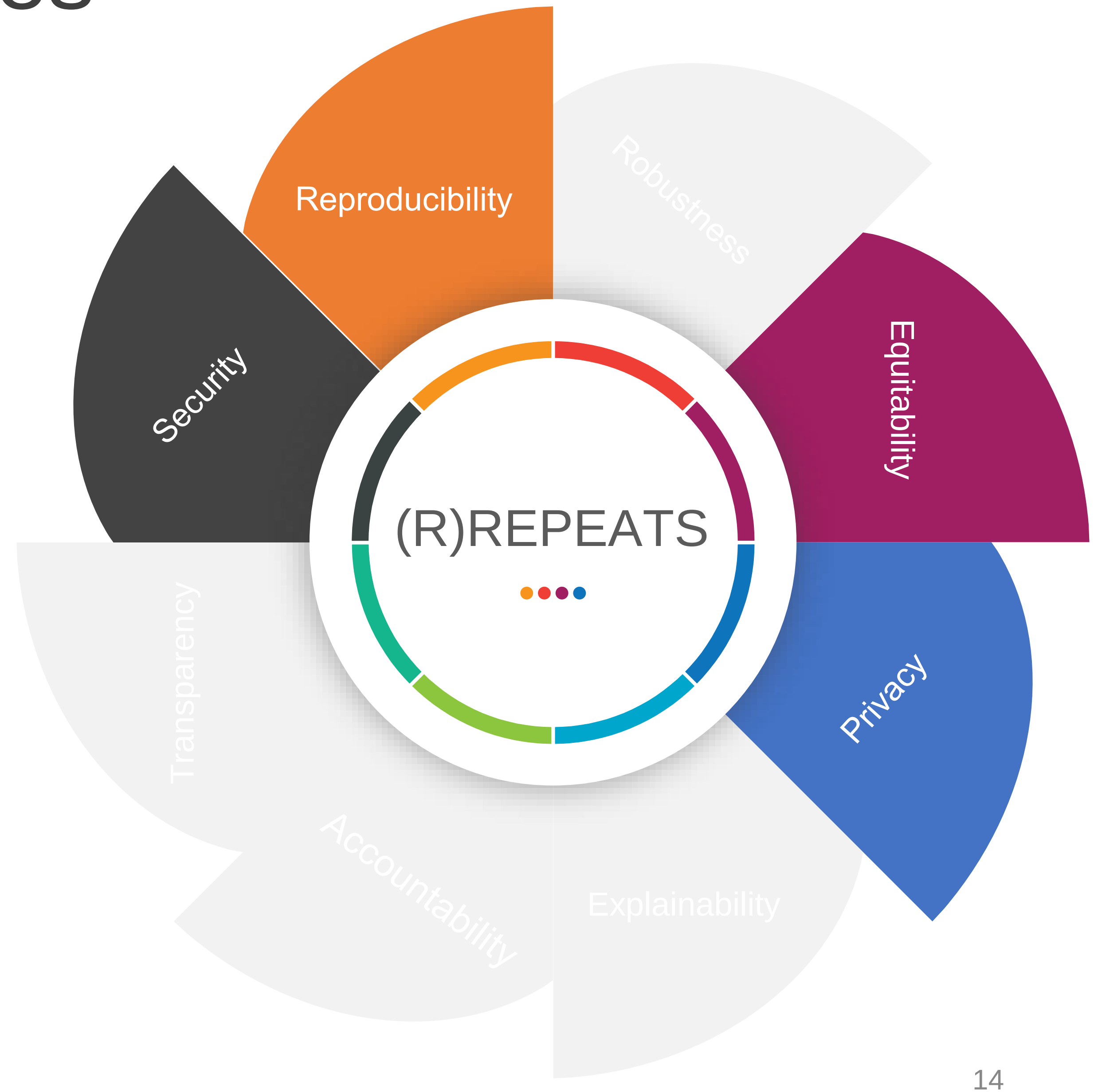
Security

Safety of AI should be tested and assured across the entire lifecycle

The Trusted AI Principles - Case in Point

[Uber sued by drivers over automated “robo-firing”](#).

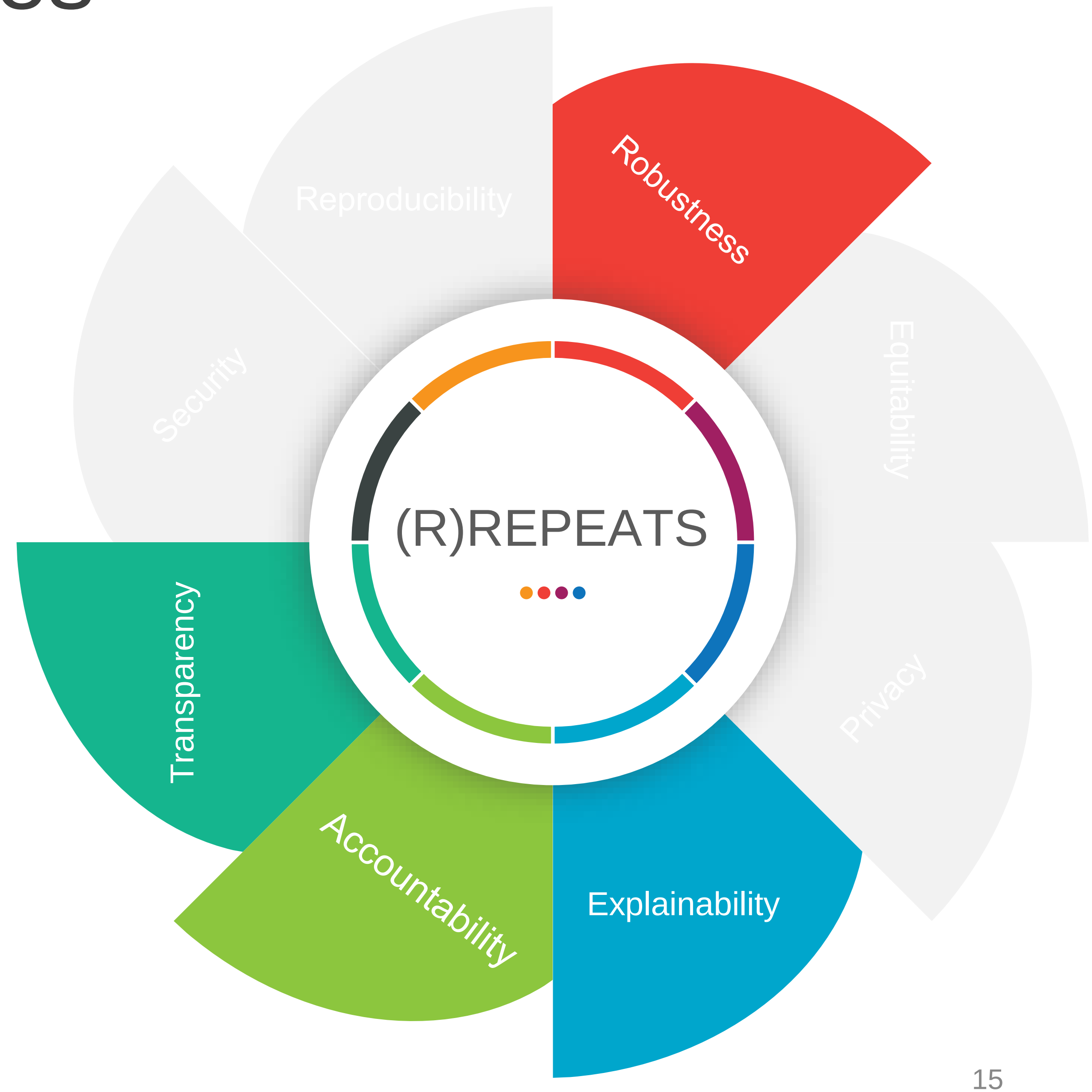
The UK court ruled in favor of the drivers who have allegedly been wrongly accused of fraudulent activity by the company’s algorithms and immediately had their accounts terminated without a right of appeal.



The Trusted AI Principles - Case in Point

Billions were cut off from Facebook, Instagram, WhatsApp for hours.

Changes to its underlying internet infrastructure that coordinates the traffic between its data centers caused one of the biggest tech companies to disappear from the internet for hours.



LFAI & Data
The Trusted AI Principles – Tools and Techniques

Operationalizing Trusted AI in Finance using the QuSandbox

LF AI & DATA

Sri Krishnamurthy, QuantUniversity

Operationalizing Trustworthy AI in Finance with the QuSandbox

Presented By:
Sri Krishnamurthy, CFA, CAP
sri@quantuniversity.com
www.quantuniversity.com

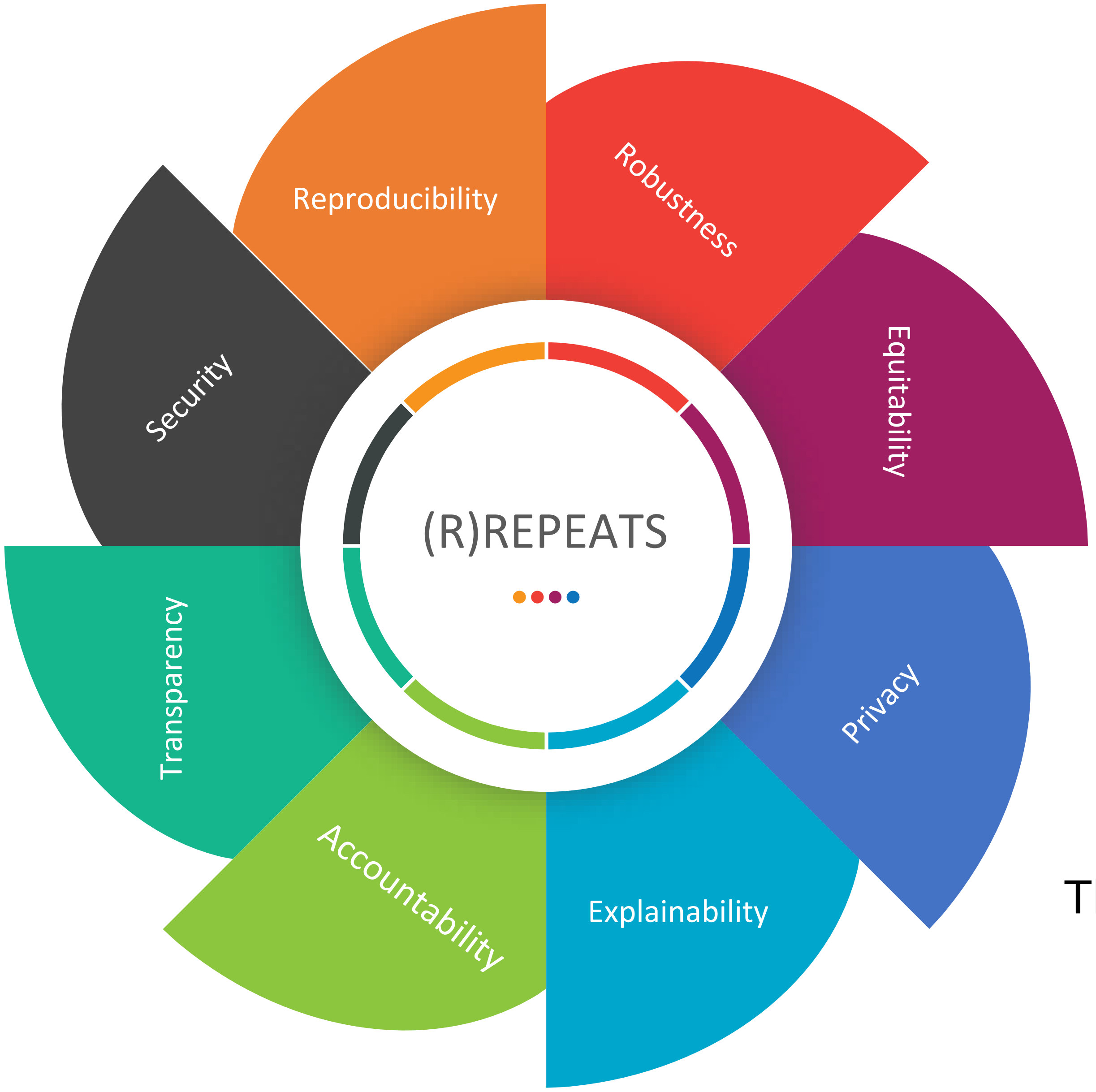
October 27th 2021
LFAI Trustworthy AI principles Meet

QuantUniversity

- Boston-based Data Science, Quant Finance and Machine Learning AI Risk Advisory
- Specialties include Algorithmic audits, Model Risk Management and AI project enablement
- Training programs for more than 1000 students in Quantitative methods, Data Science, Machine Learning and AI Risk Management
- Building QuSandbox, a platform for AI and Machine Learning Governance and Risk Management
- Associate Member of the LFAI since 2021

QuantUniversity, LLC

QuSandbox



The 8 LFAI Principles for Trusted AI – (R)REPEATS

The principles are of equal importance and value.

No principle is of higher priority than another.

The principles are related to each other.

The Implementation GAP!

“Recognizing this, actors across industry, government and civil society have rolled out an expanding array of ethical principles to guide the development and use of AI – over 175 to date.

*While the explosive growth in AI ethics guidelines is welcome, it has created an **implementation gap** – it is easier to define the ethical standards a system should meet than to design and deploy a system to meet them”*

https://www.weforum.org/projects/global-ai-action-alliance

QuantUniversity, LLC
www.quantuniversity.com

Trustworthy AI – Our Approach

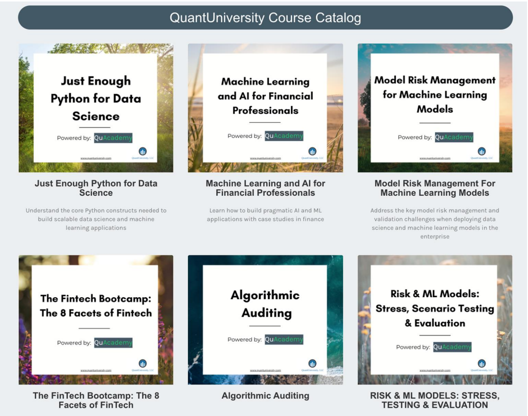
QuSandbox

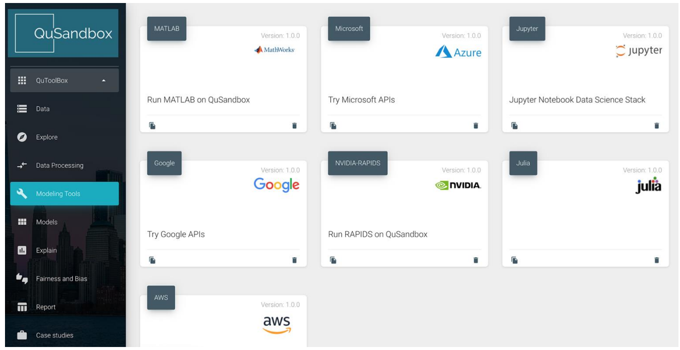
Request DEMO at info@qusandbox.com

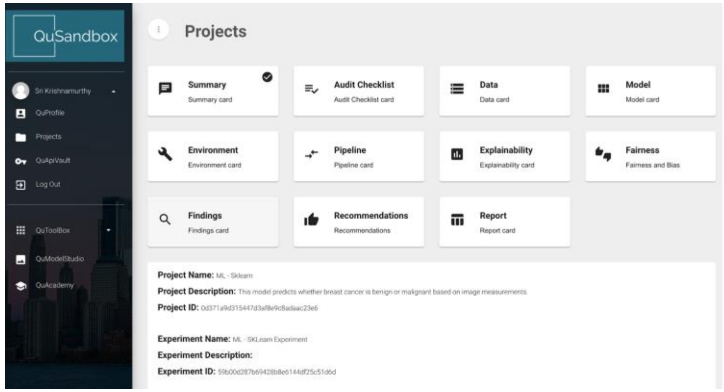
EDUCATION
QUACADEMY

EXPERIMENTATION
QUTOOLBOX

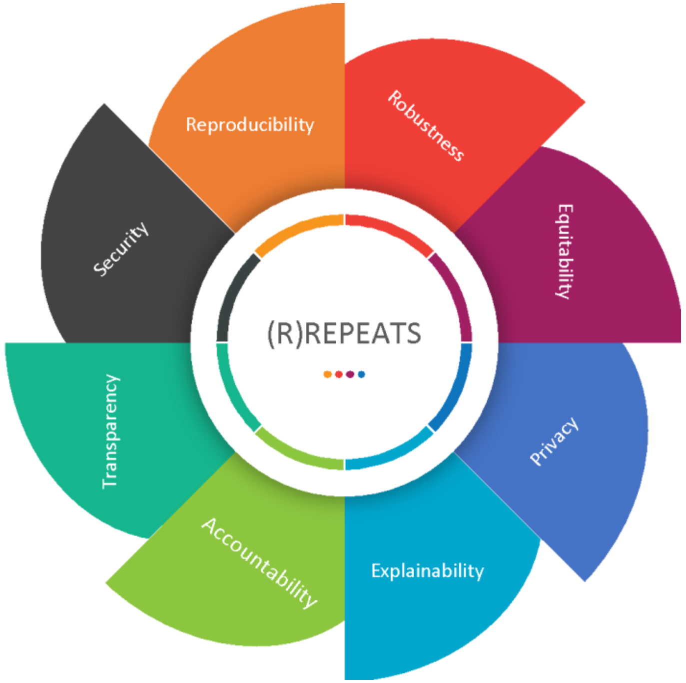
ENABLEMENT
QUSANDBOX

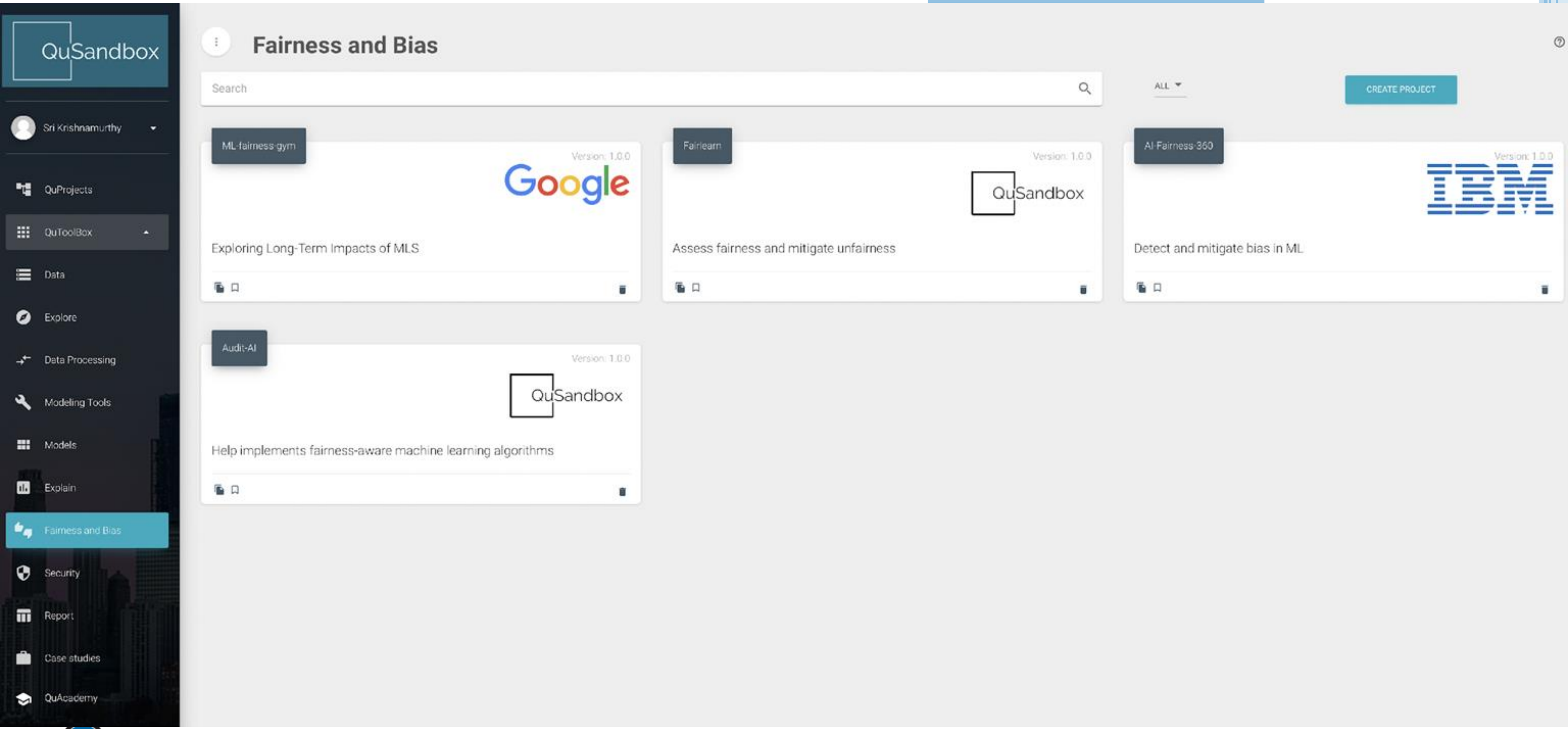






How QuSandbox addresses the LFAI principles





Equitability

QuantUniversity Course Catalog

Just Enough Python for Data Science

Understand the core Python constructs needed to build scalable data science and machine learning applications

Machine Learning and AI for Financial Professionals

Learn how to build pragmatic AI and ML applications with case studies in finance

Model Risk Management For Machine Learning Models

Address the key model risk management and validation challenges when deploying data science and machine learning models in the enterprise

The FinTech Bootcamp: The 8 Facets of FinTech

Algorithmic Auditing

RISK & ML MODELS: STRESS, TESTING & EVALUATION

QuSandbox

- Data
- Explore
- Data Processing
- Modeling Tools
- Models
- Explain
- Fairness and Bias
- Report
- Case studies
- QuAcademy

QuSynthesize_GAN

The QuSandbox Synthesize API aims at generating synthetic data maintaining the features of the original dataset to solve kinds of data problems. This API focuses on a specific use case that is generating synthetic VIX data using GAN, which could be easily implemented on other stock like datasets.

QuSynthesize 0.2 OAS3

/openapi.json

The synthetic data generation API, presented by Quant University

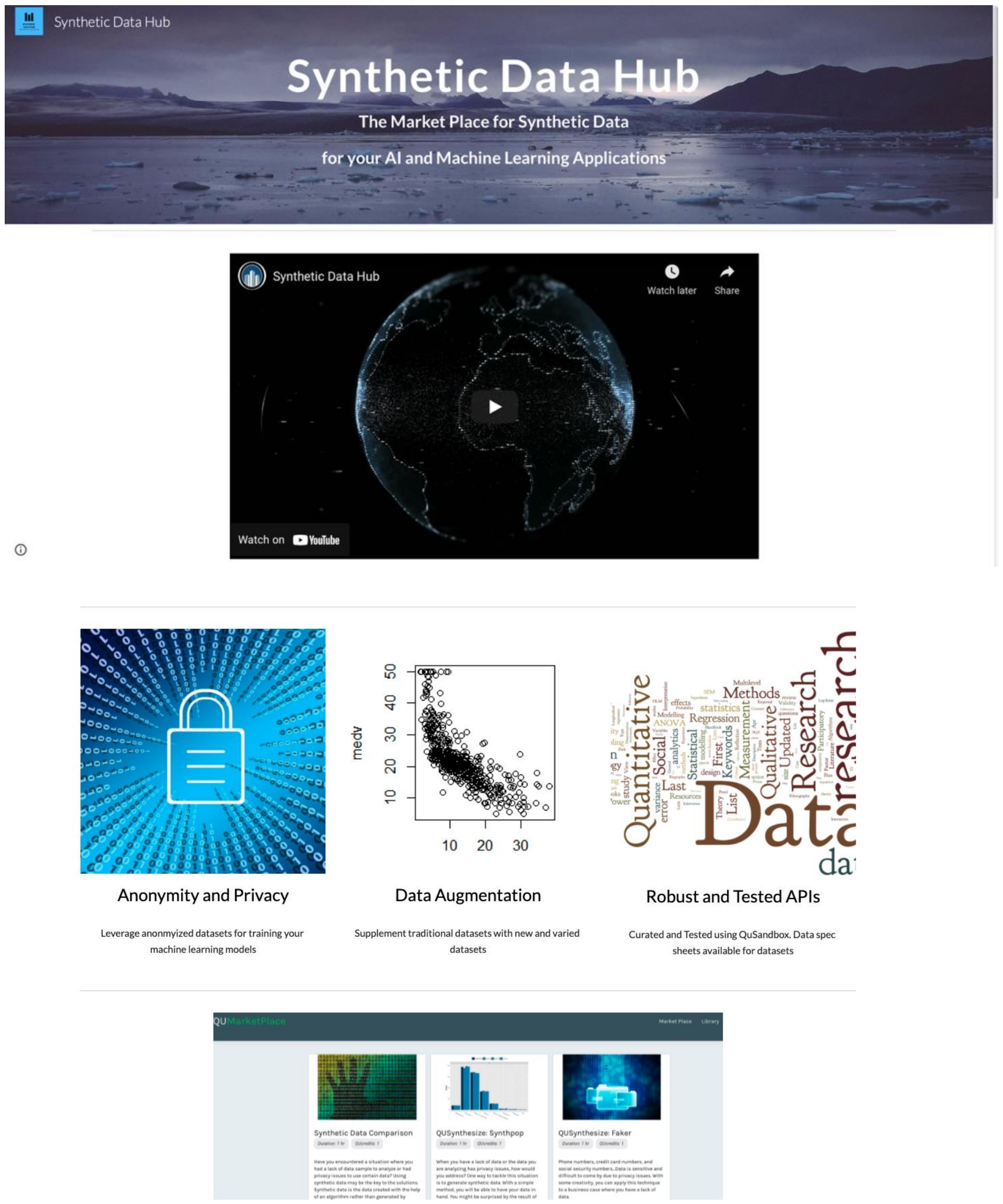
Authorize

test Tests for API access

dataset Operations on the synthesized datasets

POST `/gan/simulation` Simulation

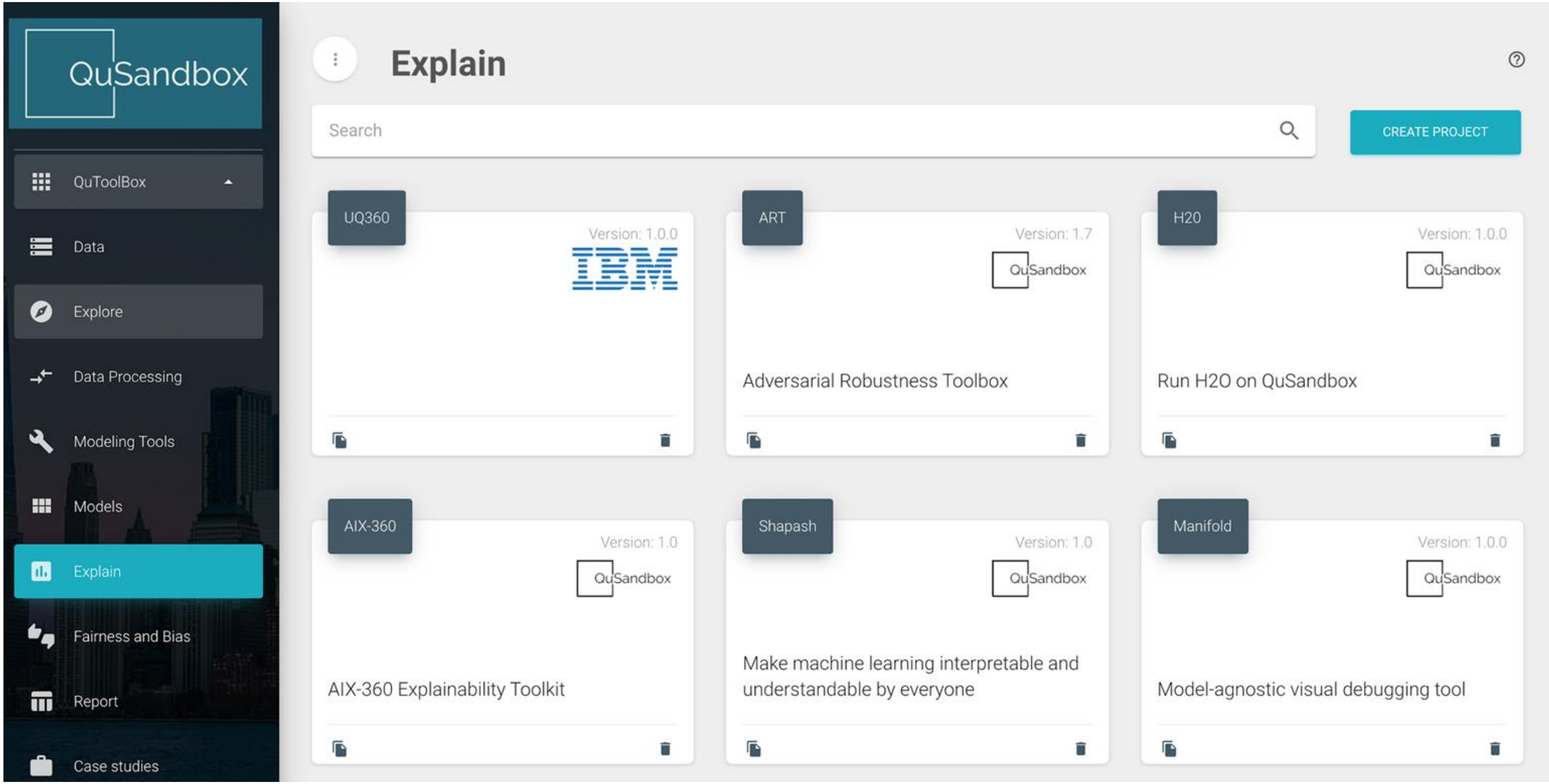
Privacy



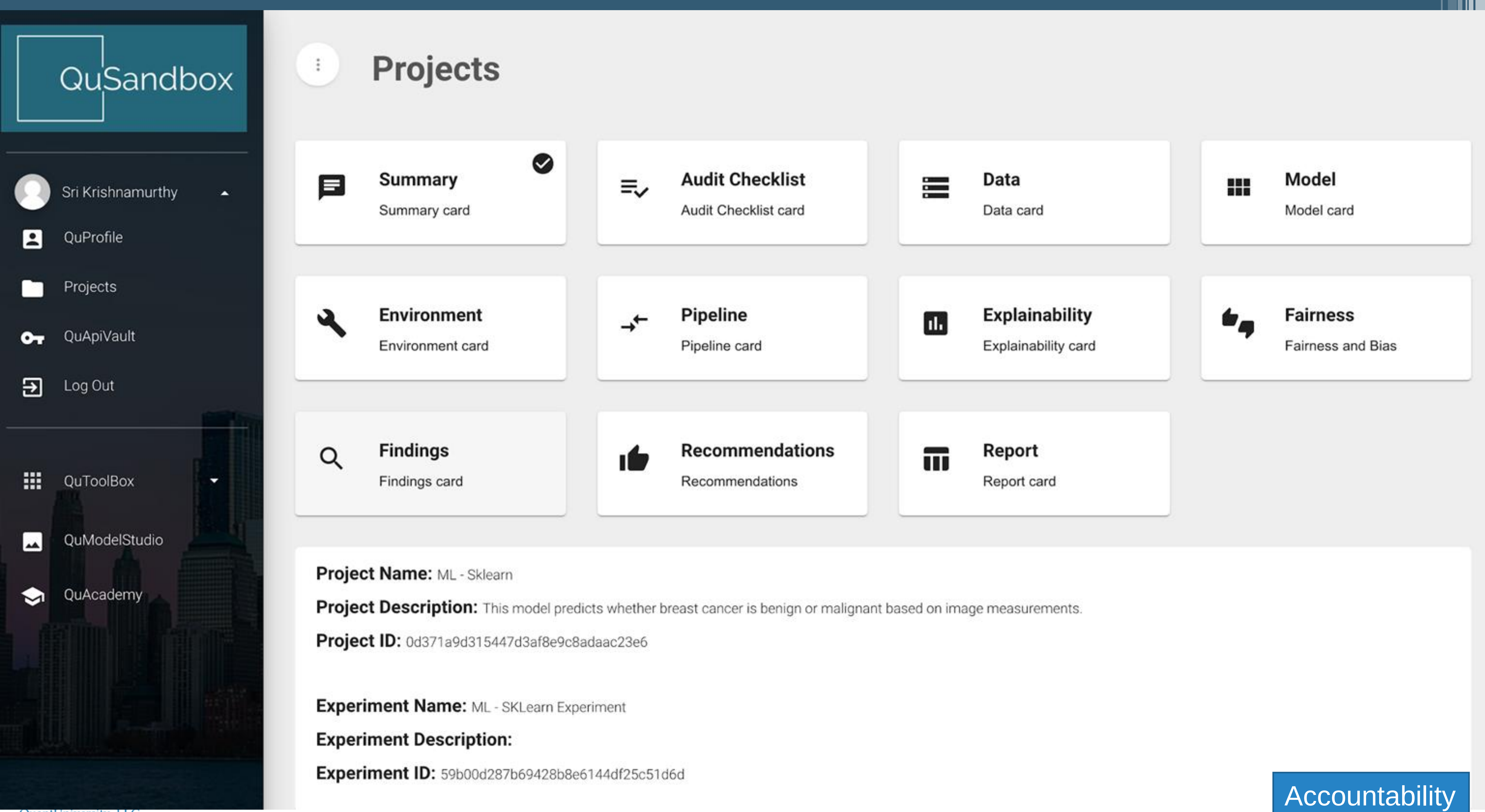
- Synthetic data Hub:
  A synthetic data marketplace to enable testing, benchmarking and replicability.
- Rich visualization and automated reporting and scorecards to quantify and monitor AI risk.
- Testing framework to run stress, scenario tests for ML models ; accelerated by GPUs

SYNTHETIC DATA HUB
WWW.SYNTHETIC DATA HUB.COM

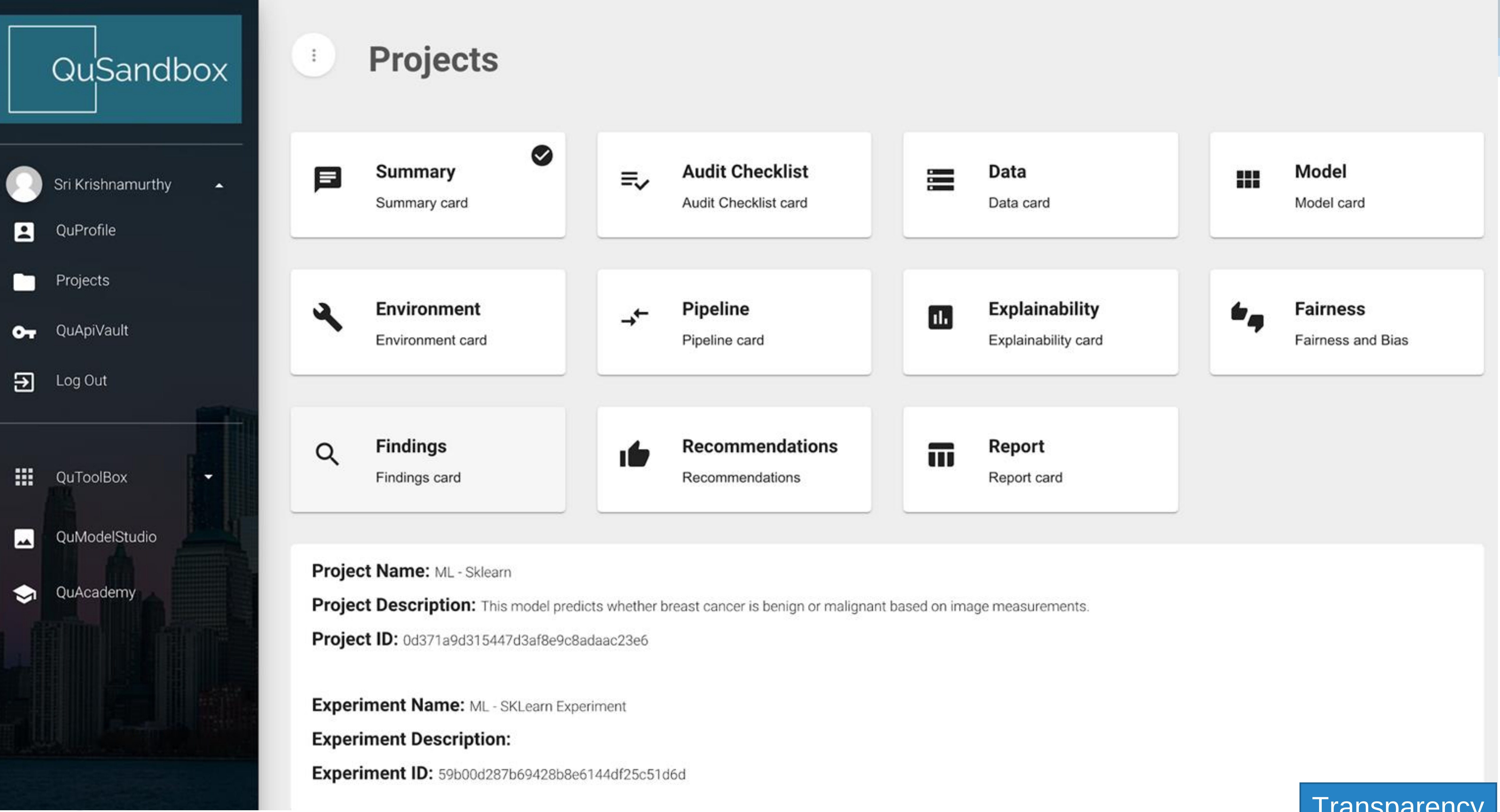
Privacy



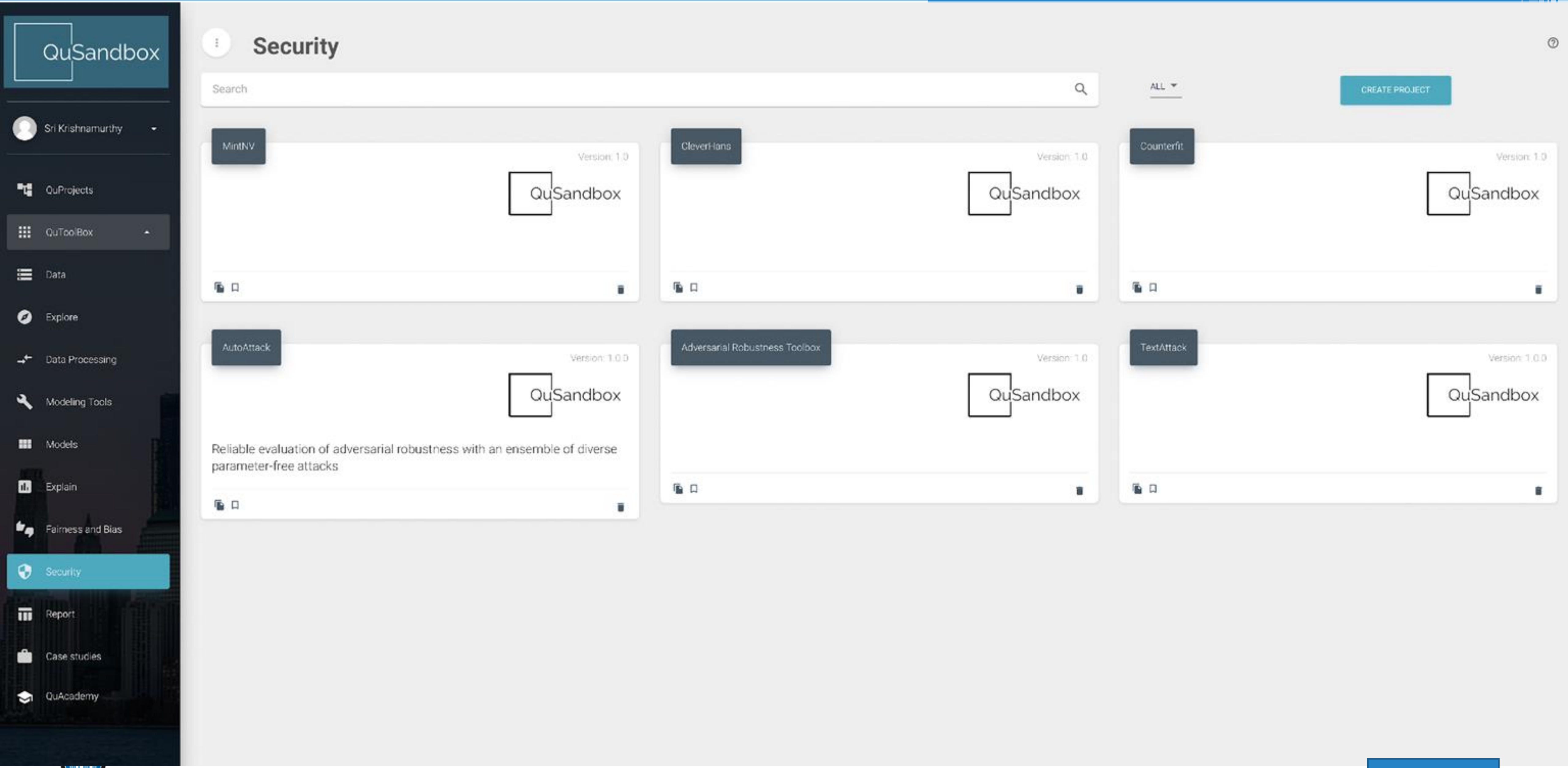
Explainability



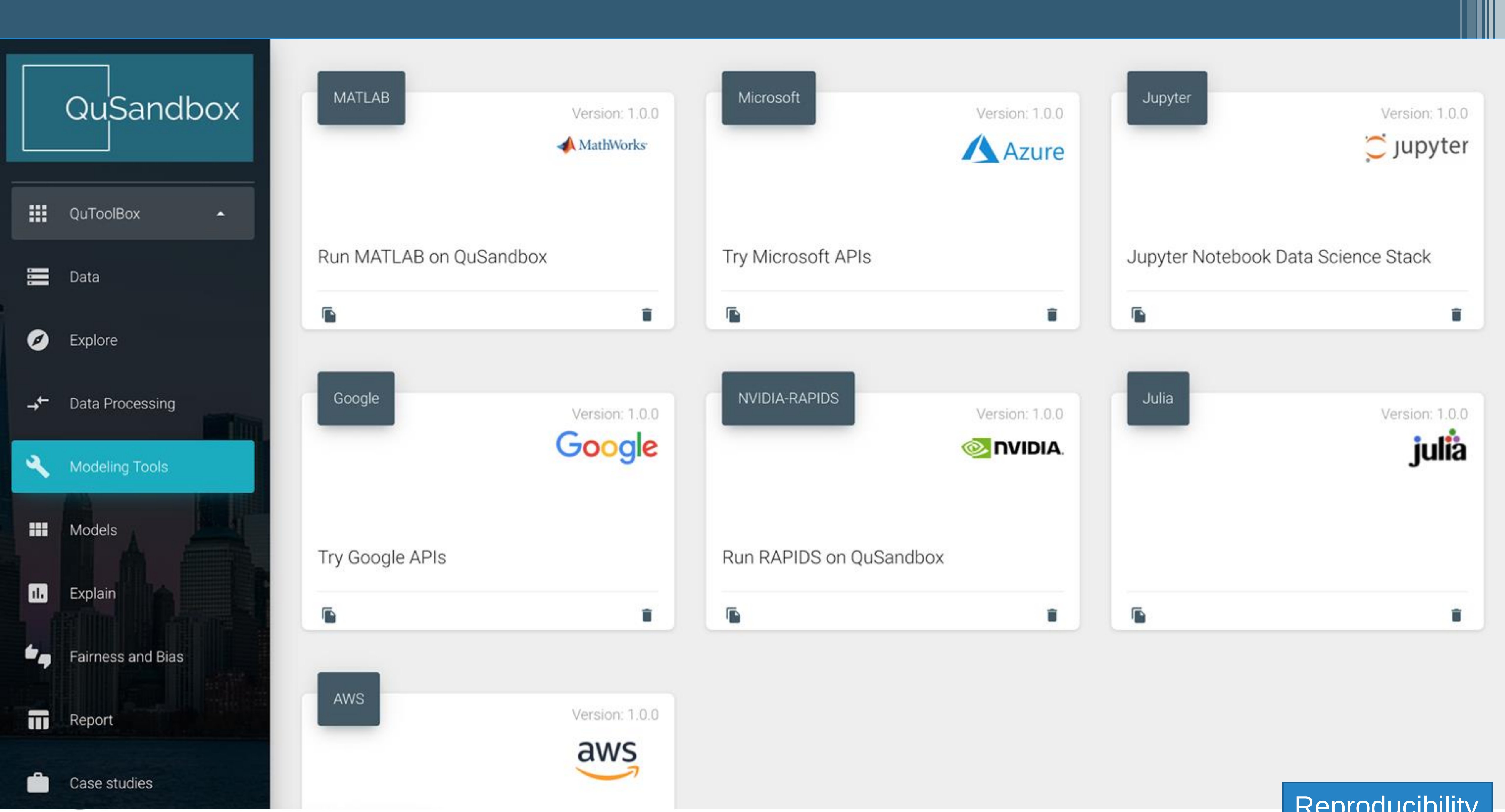
Accountability



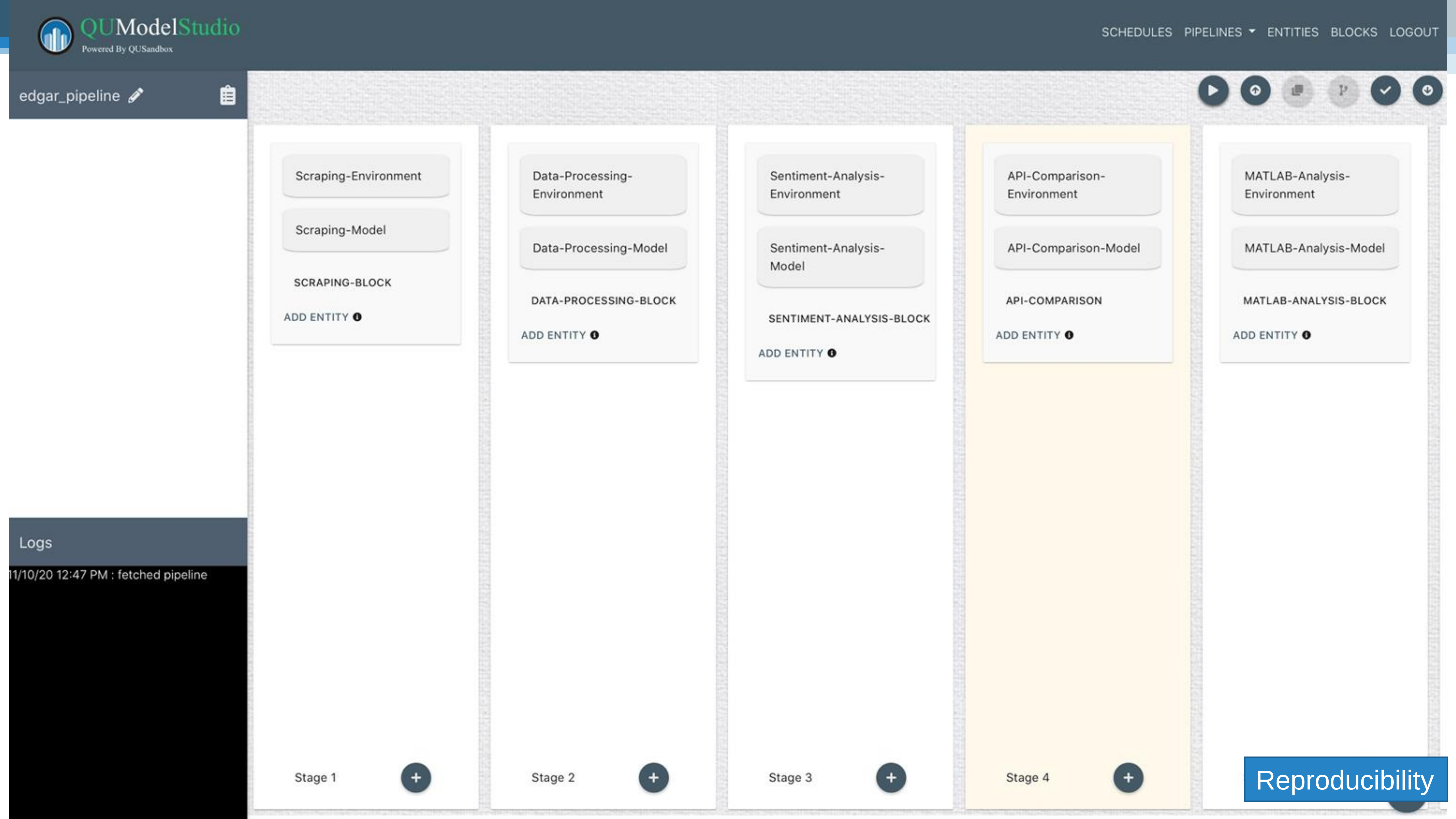
Transparency


QuSandbox
Sri Krishnamurthy
QuProjects
QuToolBox
Data
Explore
Data Processing
Modeling Tools
Models
Explain
Fairness and Bias
Security
Report
Case studies
QuAcademy
Security
Search
ALL
CREATE PROJECT
MintNV
Version: 1.0
QuSandbox
CleverHans
Version: 1.0
QuSandbox
Counterfit
Version: 1.0
QuSandbox
AutoAttack
Version: 1.0.0
QuSandbox
Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks
Adversarial Robustness Toolbox
Version: 1.0
QuSandbox
TextAttack
Version: 1.0.0
QuSandbox

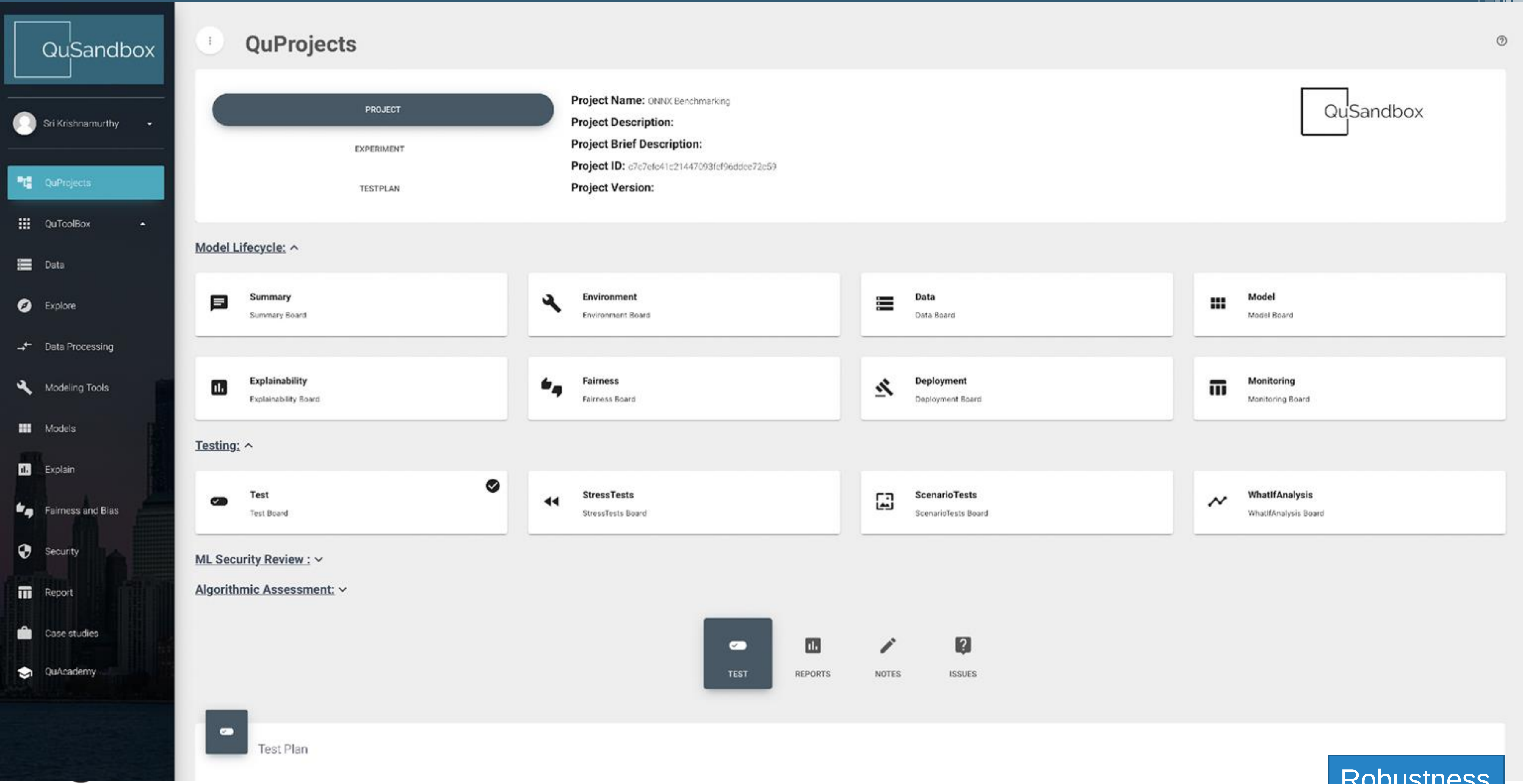
Security



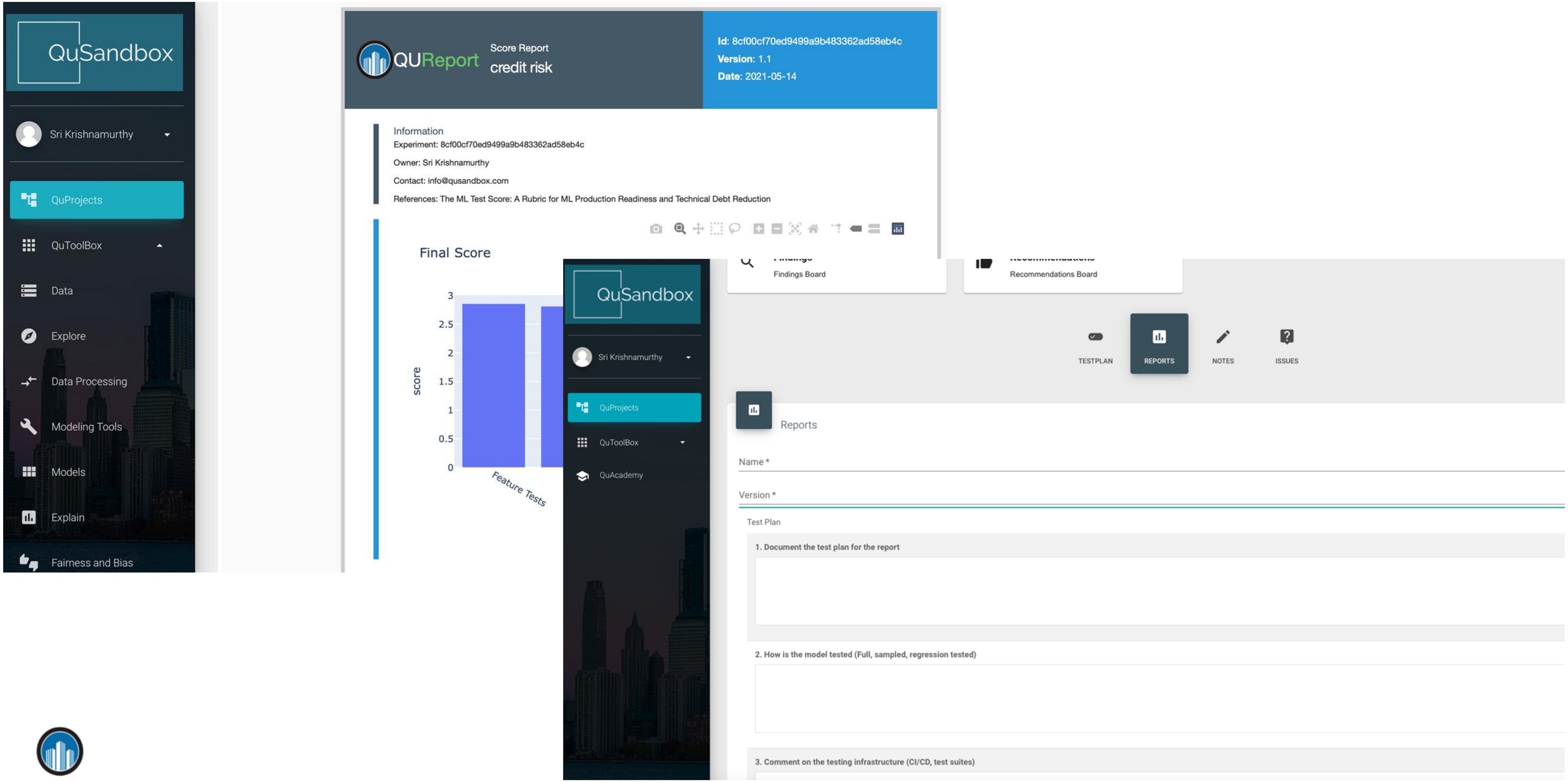
Reproducibility



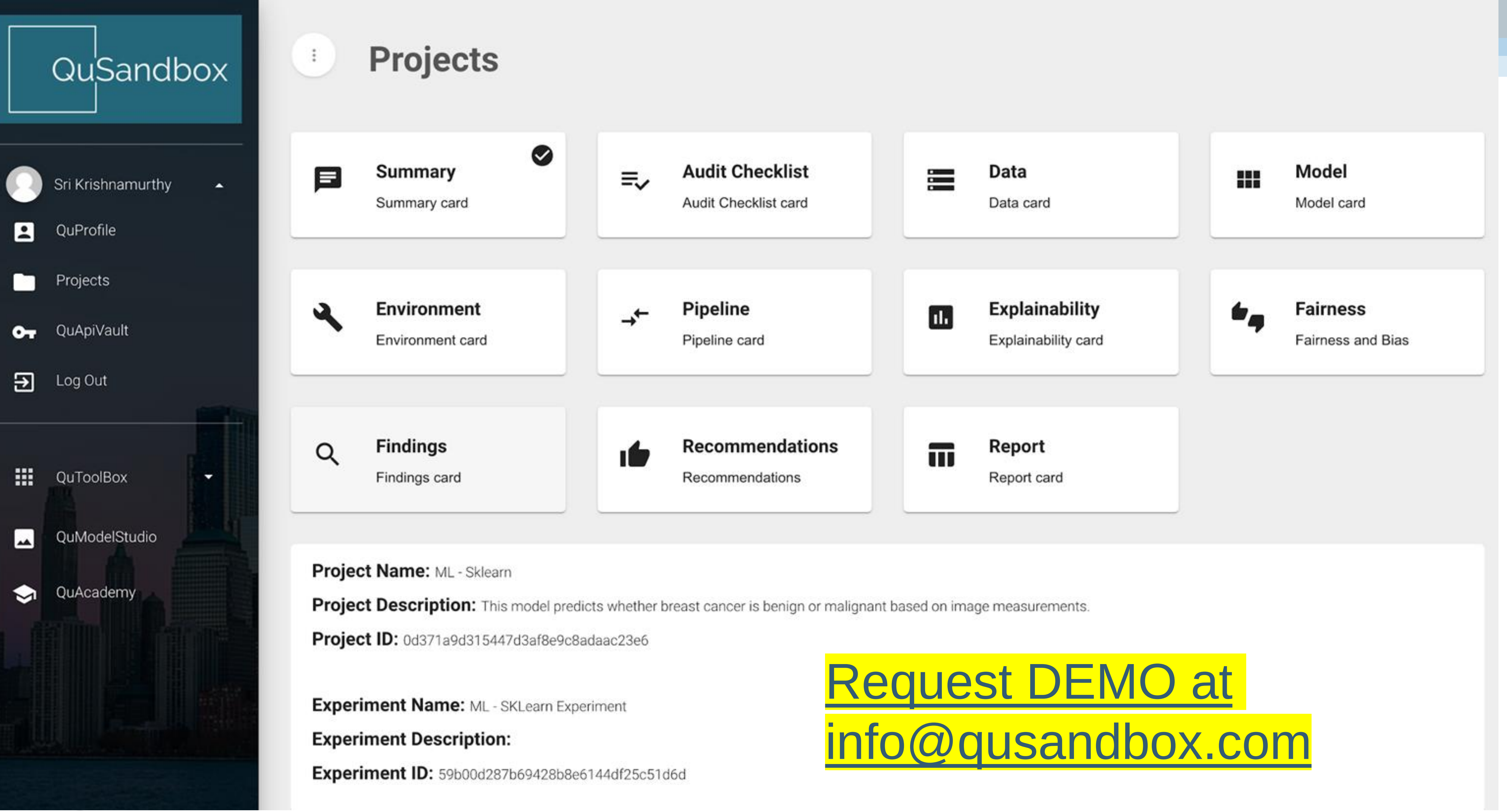
Reproducibility


QuSandbox
Sri Krishnamurthy
QuProjects
QuToolBox
Data
Explore
Data Processing
Modeling Tools
Models
Explain
Fairness and Bias
Security
Report
Case studies
QuAcademy
QuProjects
PROJECT
EXPERIMENT
TESTPLAN
Project Name: ONNX Benchmarking
Project Description:
Project Brief Description:
Project ID: c7c7efc41c21447093fcf96ddce72c59
Project Version:
QuSandbox
Model Lifecycle:
Summary
Summary Board
Environment
Environment Board
Data
Data Board
Model
Model Board
Explainability
Explainability Board
Fairness
Fairness Board
Deployment
Deployment Board
Monitoring
Monitoring Board
Testing:
Test
Test Board
StressTests
StressTests Board
ScenarioTests
ScenarioTests Board
WhatIfAnalysis
WhatIfAnalysis Board
ML Security Review :
Algorithmic Assessment:
TEST
REPORTS
NOTES
ISSUES
Test Plan

Robustness



QuantUniversity, LLC
www.quantuniversity.com



Request DEMO at info@qusandbox.com

Speaker bio

Sri Krishnamurthy
Founder and CEO
QuantUniversity

QuantUniversity, LLC
www.quantuniversity.com

- AI advisory focused on AI Risk, Governance and enablement
- Prior Experience at MathWorks, Citigroup and Endeca and 25+ financial services and energy customers.
- Columnist for the Wilmott Magazine
- Author of forthcoming book "Pragmatic AI and ML in Finance"
- Teaches AI/ML and Fintech Related topics in the MS and MBA programs at Northeastern University, Boston
- Reviewer: Journal of Asset Management

QuantUniversity, LLC

www.quantuniversity.com

Thank you!

Contact

Sri Krishnamurthy, CFA, CAP
Founder and CEO
QuantUniversity LLC.

LinkedIn srikrishnamurthy

www.QuantUniversity.com

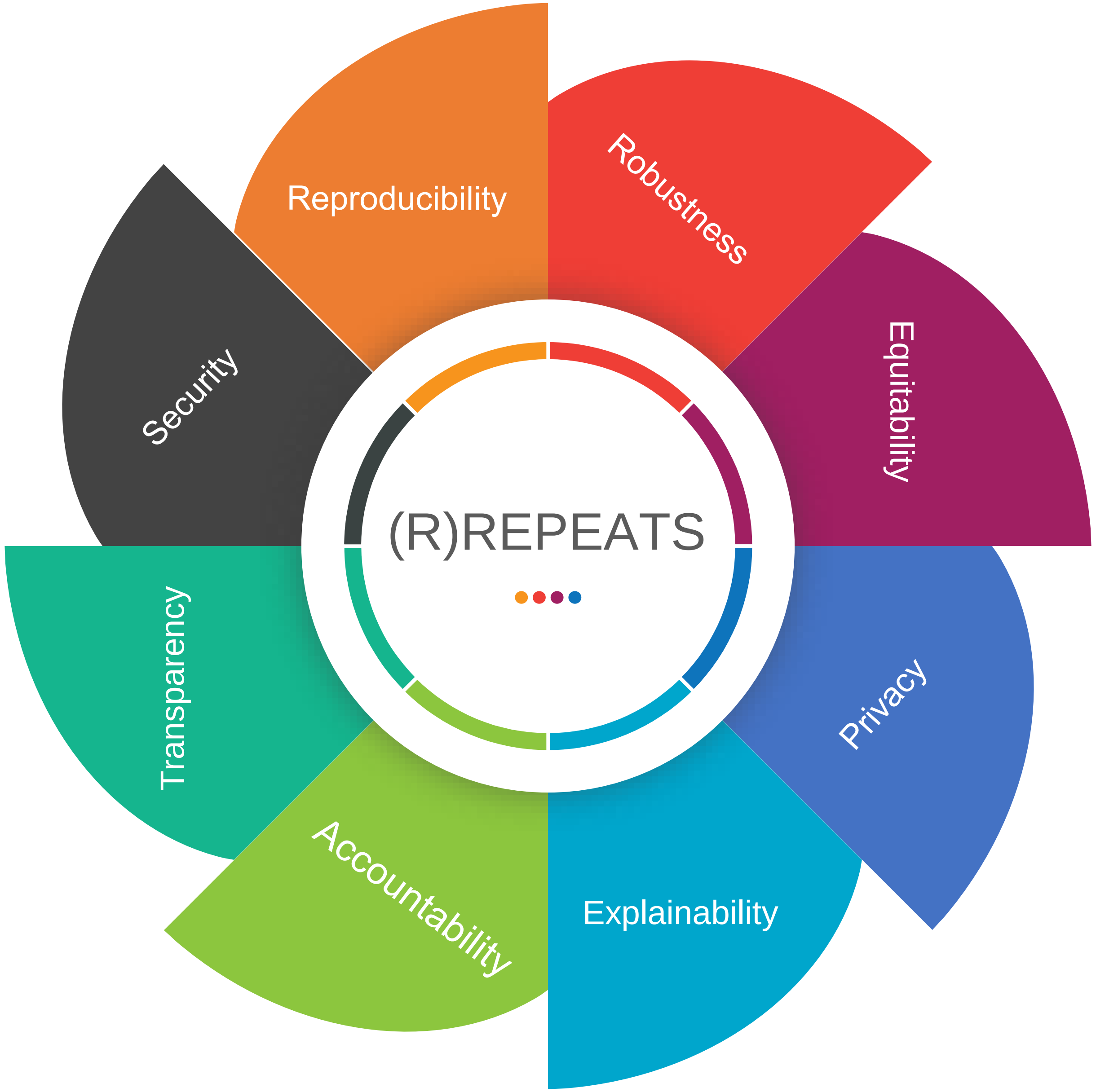
LFAI & Data

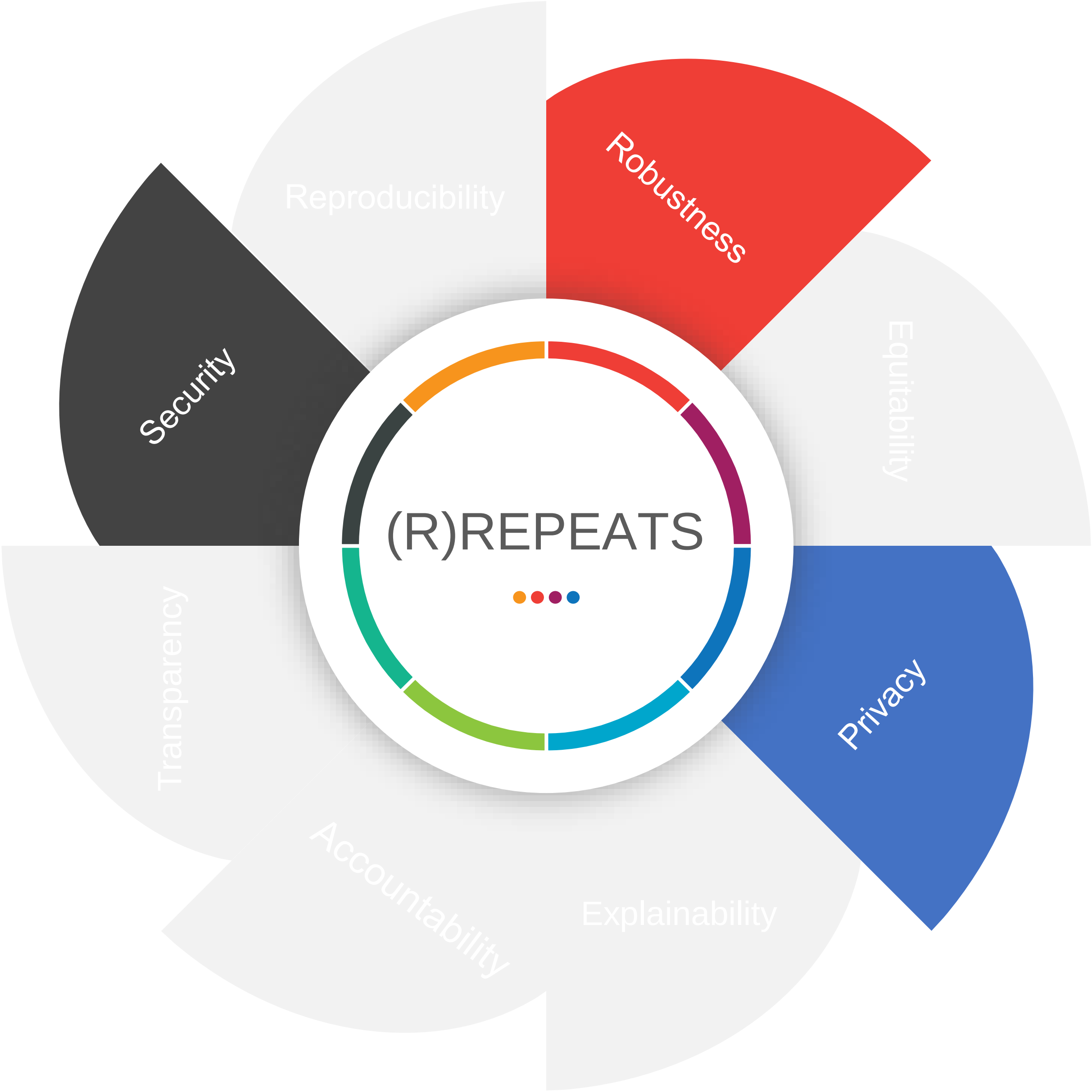
Trusted AI Principles – Tools and Techniques

Trusted AI Tools (AI Fairness, AI Explainability, Adversarial Robustness etc) and RREPEATS
Emerging DataOps activities at the LF-AI, Trusted AI and RREPEATS,
Animesh Singh, IBM

Animesh Singh,
Watson CTO, Open Technology,
Distinguished Engineer,
IBM

LF AI & DATA

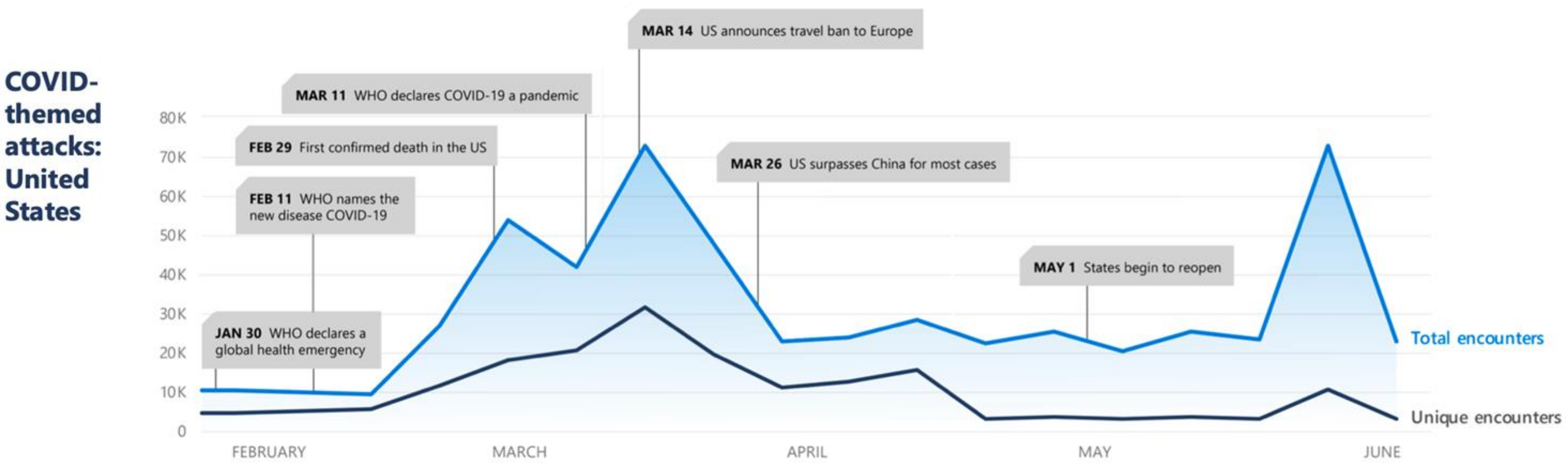

(R)REPEATS
Reproducibility
Robustness
Equitability
Privacy
Explainability
Accountability
Transparency
Security


(R)REPEATS
Reproducibility
Robustness
Equitability
Privacy
Explainability
Accountability
Transparency
Security

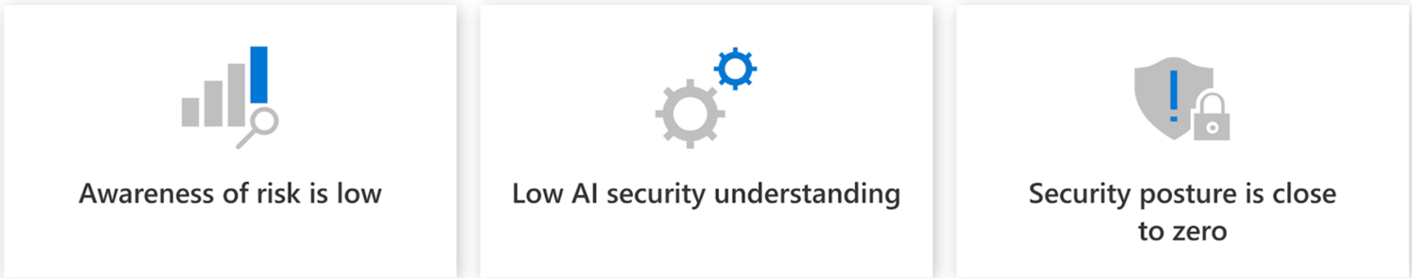
Trusted AI – Increased focus on AI Security

Cybercrime follows the issues of the day

Malware encounters align with news headlines



State of Security for AI



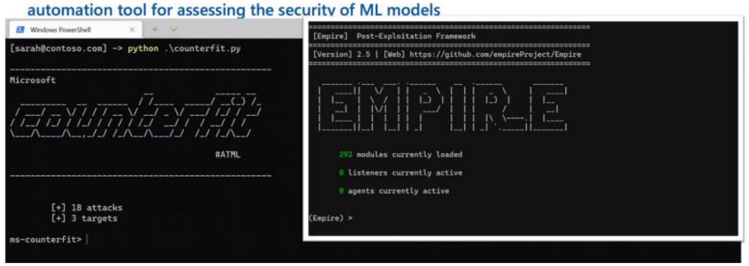
Gartner research

600+ executives say security/privacy is top blocker to using AI

"Machine Learning presents a new attack surface and increases security risks.... Application leaders must anticipate and prepare to mitigate risks of data corruption, model theft, and adversarial examples."

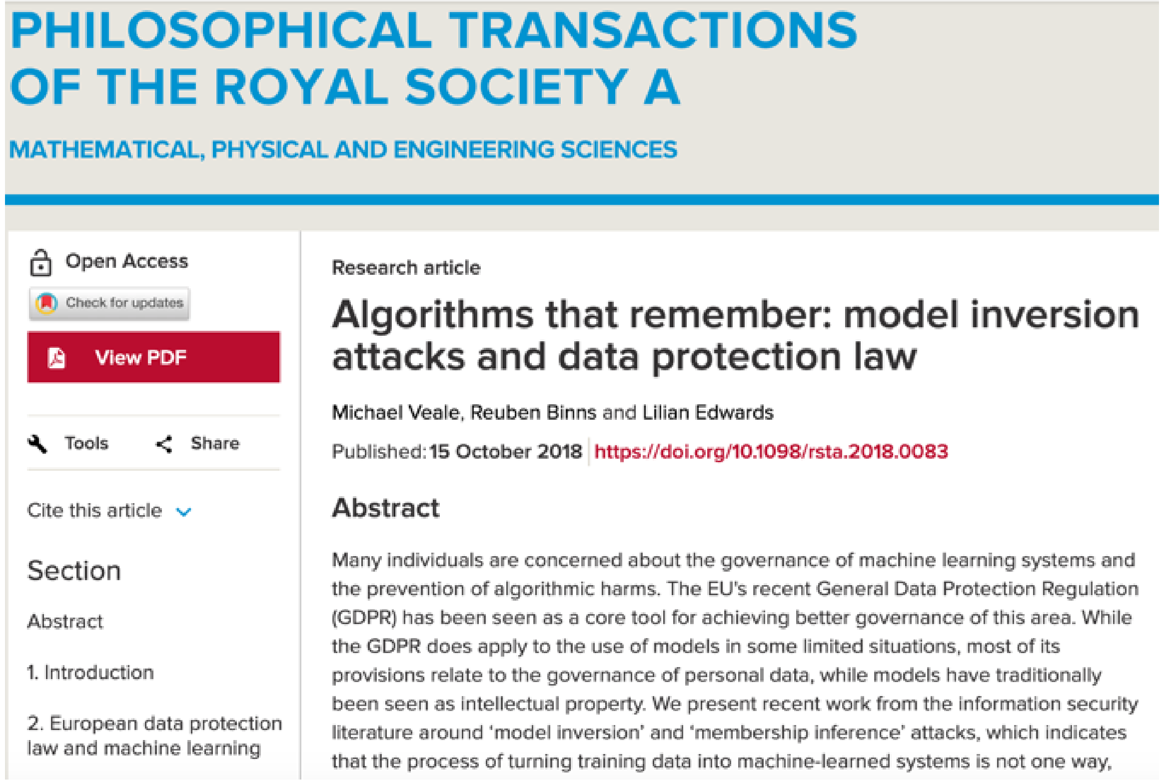
Counterfit

automation tool for assessing the security of ML models



Impact of GDPR on AI

Some AI models will be legally classified as personal data


PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES

Open Access

Check for updates

View PDF

Tools Share

Cite this article

Section

Abstract

1. Introduction

2. European data protection law and machine learning

Research article

Algorithms that remember: model inversion attacks and data protection law

Michael Veale, Reuben Binns and Lilian Edwards

Published: 15 October 2018 https://doi.org/10.1098/rsta.2018.0083

Abstract

Many individuals are concerned about the governance of machine learning systems and the prevention of algorithmic harms. The EU's recent General Data Protection Regulation (GDPR) has been seen as a core tool for achieving better governance of this area. While the GDPR does apply to the use of models in some limited situations, most of its provisions relate to the governance of personal data, while models have traditionally been seen as intellectual property. We present recent work from the information security literature around 'model inversion' and 'membership inference' attacks, which indicates that the process of turning training data into machine-learned systems is not one way,

https://royalsocietypublishing.org/doi/full/10.1098/rsta.2018.0083

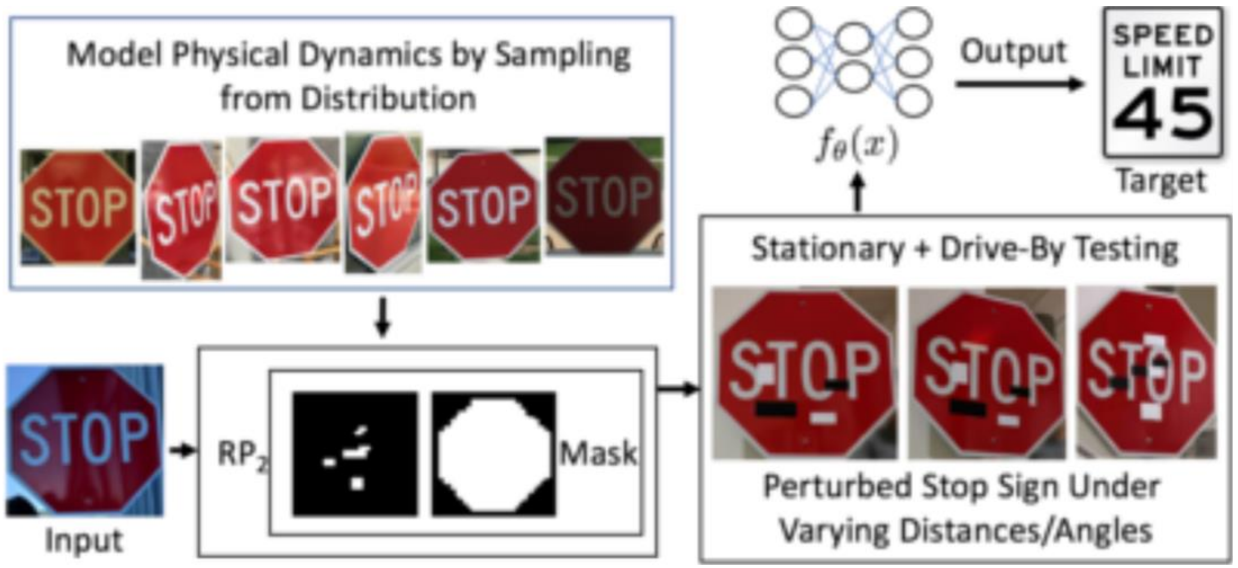
GDPR mostly relates to the governance of personal data.

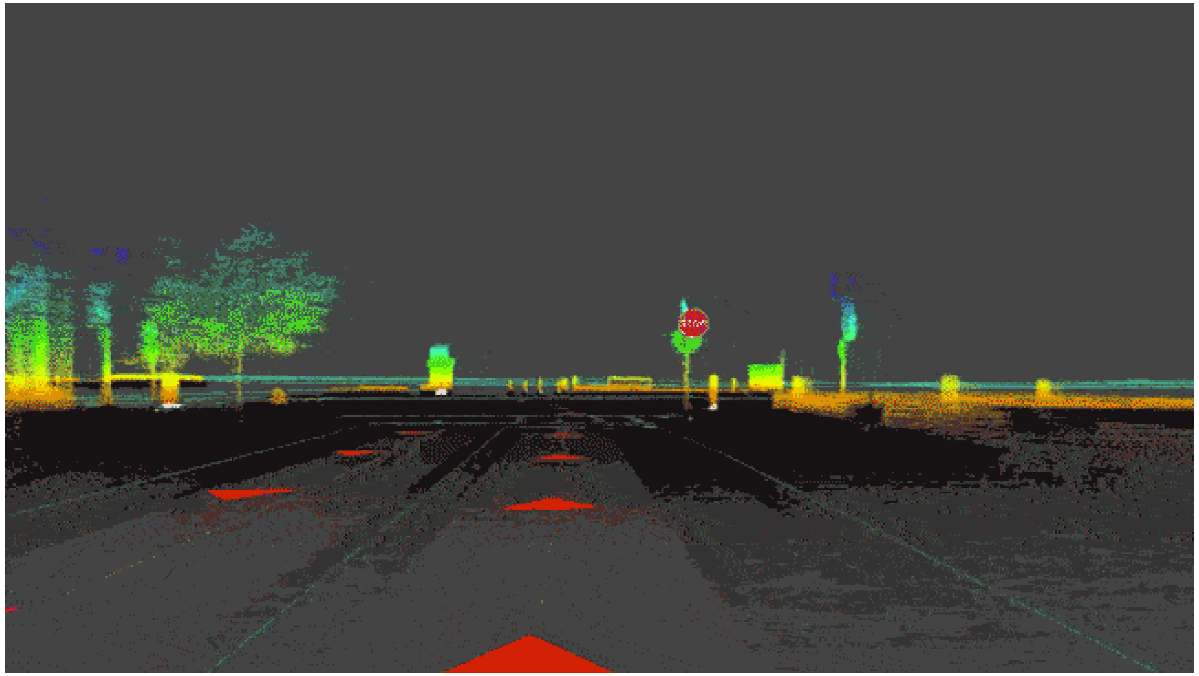
This paper presents recent work from the information security literature around 'model inversion' and 'membership inference' attacks, indicating that the process of turning training data into machine-learned systems is not one way, and demonstrate how this could lead some models to be legally classified as personal data.

They also explore the different rights and obligations this would trigger.

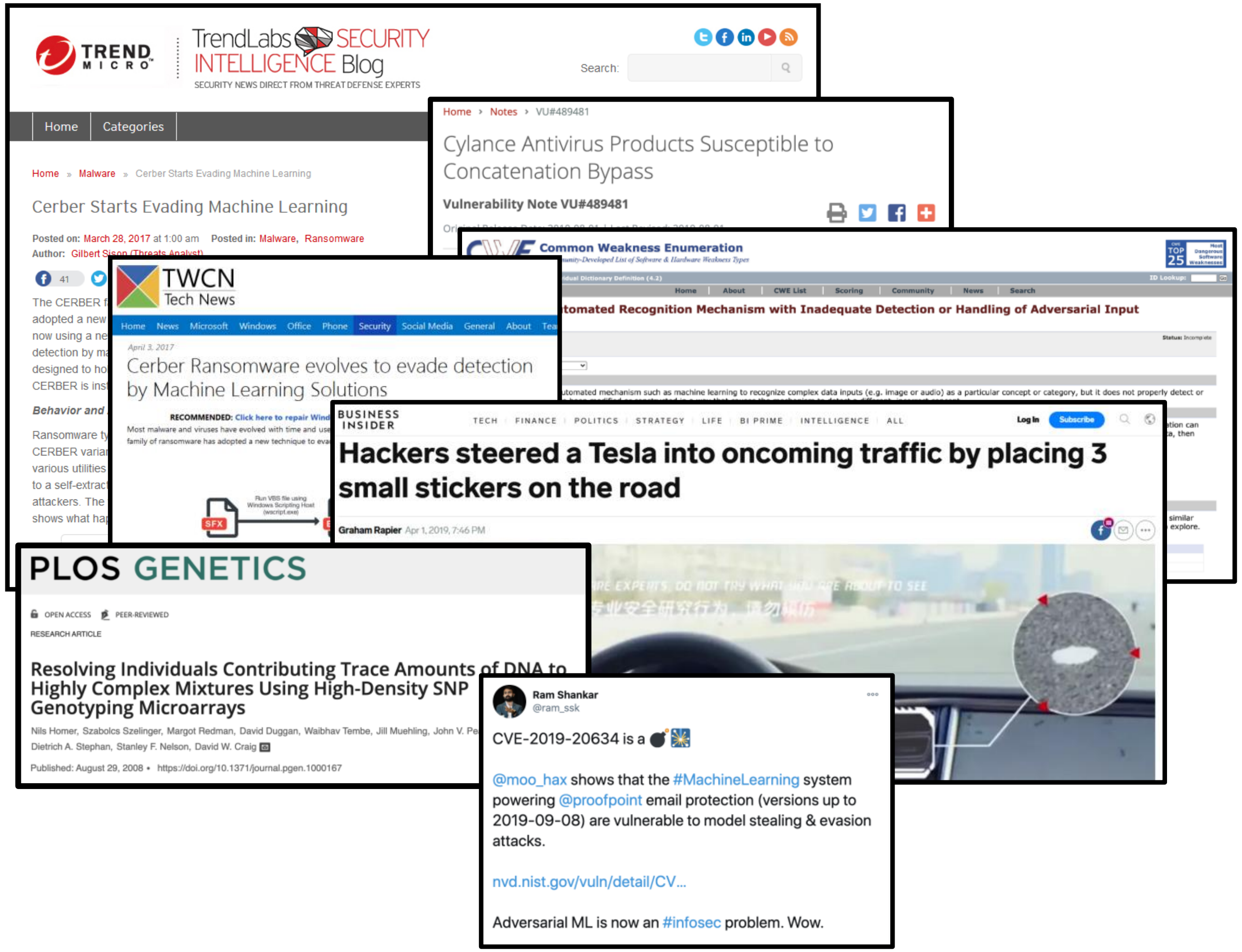
Adversarial Threats to AI

https://arxiv.org/pdf/1707.08945.pdf





Real-world Adversarial Exploits

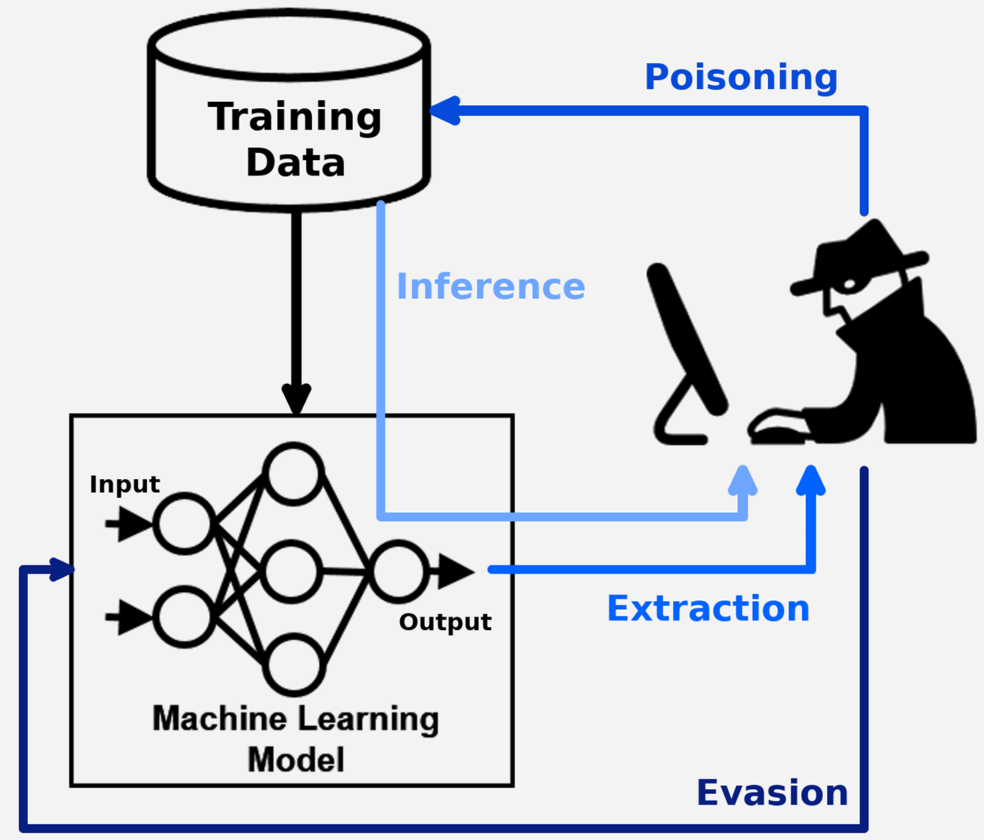


- Evasion of classification in antivirus products
  - Undetected ransomware installs and encrypts your computer
- Real-world adversarial patches for evasion attacks on cars
  - Losing control of autonomous vehicles leads to damages and injury
- Extraction of classification models to stage evasion attack against email protection system
  - Bypassing email security systems increases chances of phishing attacks
- Leaking sensitive private information
  - Revealing a person health condition via membership inference on health-related models

Adversarial Threats to Machine Learning

Adversarial threats against machine learning models and applications have a wide variety of attack vectors.

- **Evasion:** Modifying input to influence model
- **Poisoning:** Modify training data to add backdoor
- **Extraction:** Steal a proprietary model
- **Inference:** Learn information on private data



Adversarial Robustness Toolbox (ART)

Adversarial Robustness Toolbox

ART is a Python library for machine learning security

TensorFlow Keras

PYTORCH mxnet

scikit learn GPy

dmlc XGBoost LightGBM CatBoost

- github.com/Trusted-AI/adversarial-robustness-toolbox
- provide tools to developers and researcher
- Evaluating, Defending, Certifying and Verifying of machine learning models and applications
- **All Tasks:** Classification, Object Detection, Generation, Encoding, Certification, etc.
- **All Frameworks:** TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy
- **All Data:** images, tables, audio, video, etc.
- Contributions and feedback are very welcome!

THE LINUX FOUNDATION

ART Community – Contributors and Tools

ART Adopters and Contributors

- IBM
- Microsoft
- Troj AI
- Two Six Labs, LLC
- Kyushu University
- Intel Corporation
- University of Chicago
- The MITRE Corporation
- General Motors Company
- AGH University of Science and Technology
- Rensselaer Polytechnic Institute (RPI)
- IMT Atlantique

Adversarial Robustness Toolbox

2.5K GitHub Stars

150K Downloads

8K+ Commits

Armory
- Adversarial Robustness Evaluation Test Bed
- Run evaluations with ART locally or scaled in the cloud using Docker containers
- github.com/twosixlabs/armory

Counterfit
- Command line tool to simplify running evaluations with ART in terminals
- github.com/Azure/counterfit

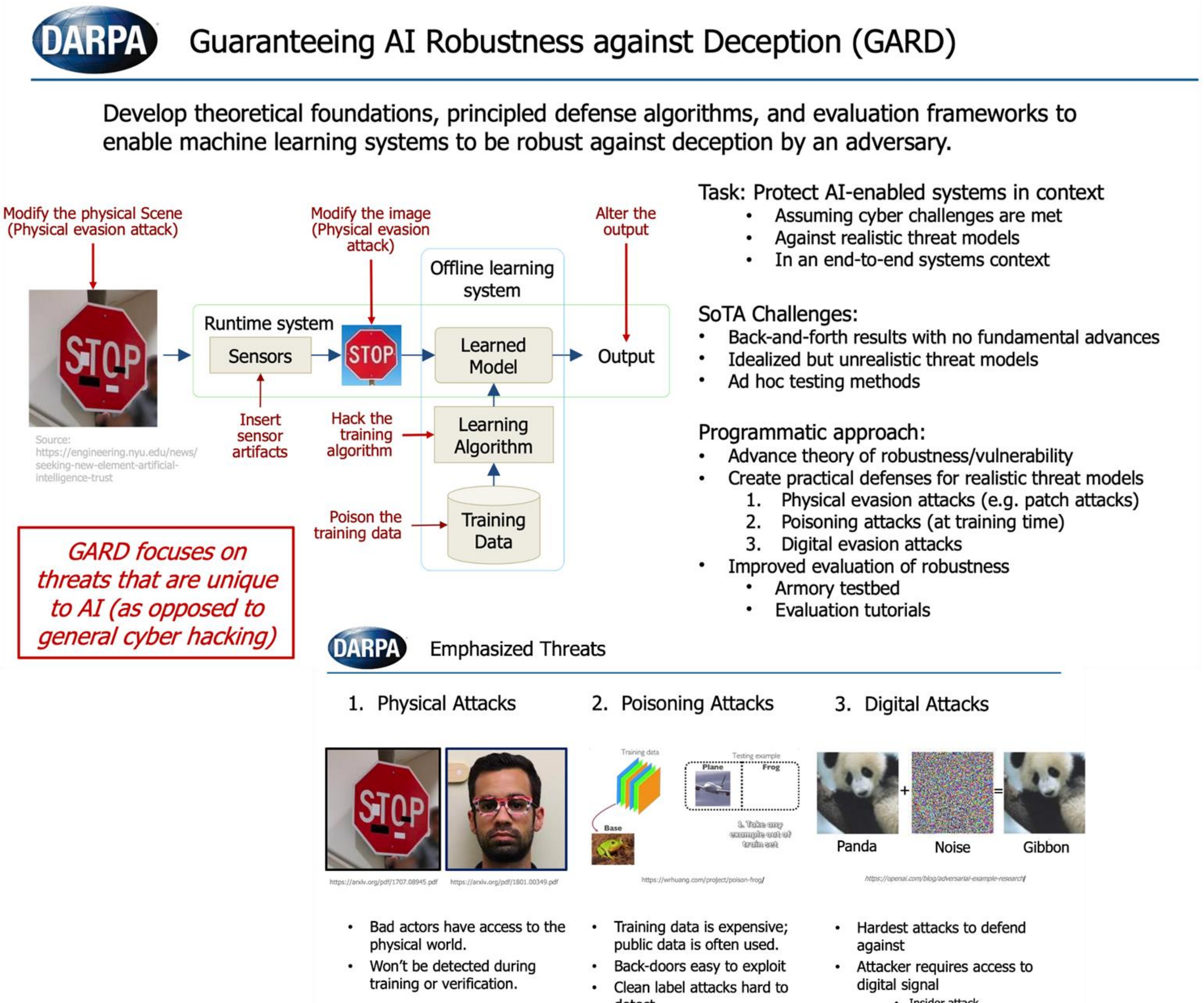
ai-privacy-toolkit
- Tools for privacy and compliance of AI models
- End-to-end privacy evaluation and mitigation of privacy risks
- github.com/IBM/ai-privacy-toolkit

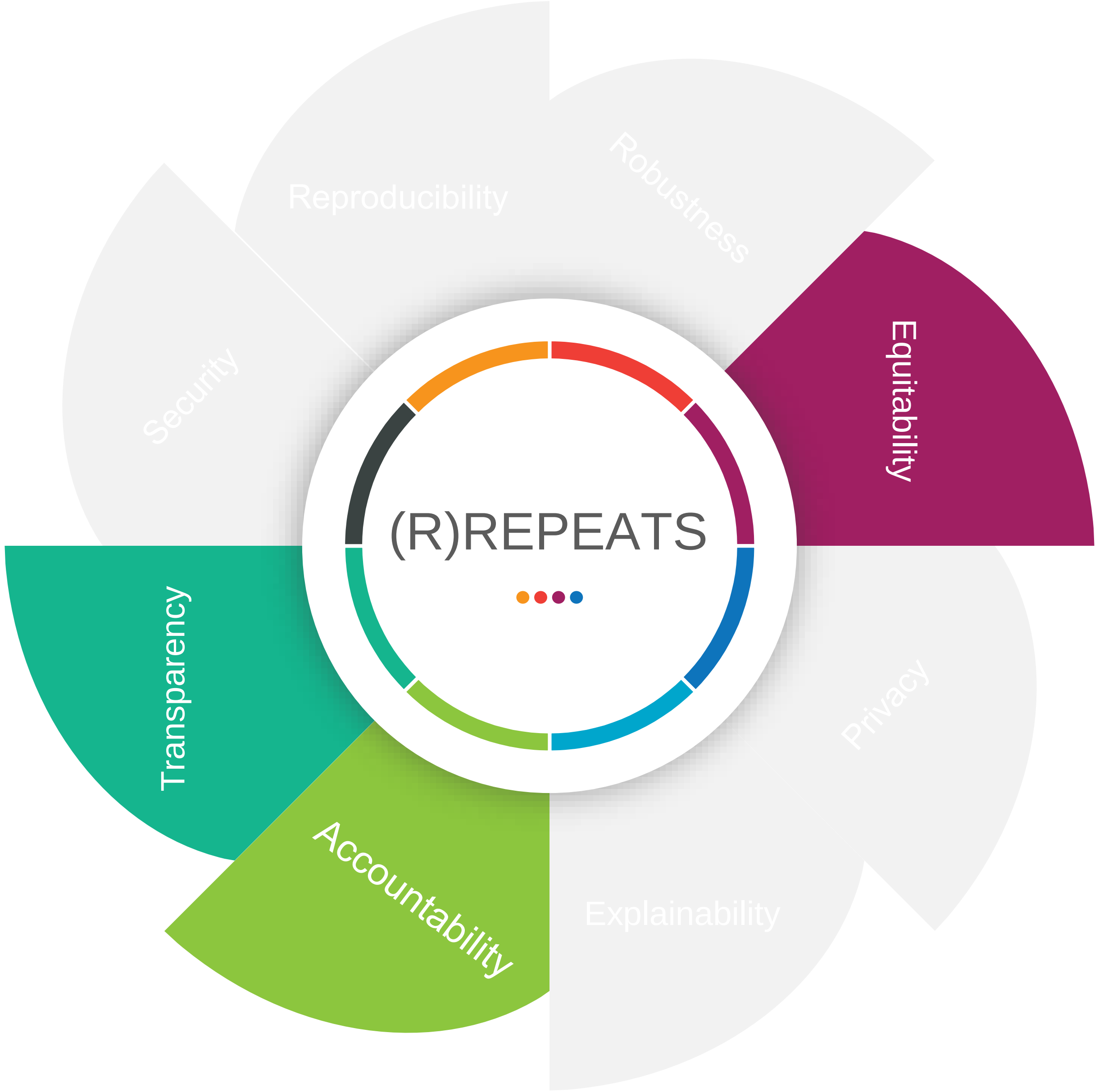
ART Adopter - DARPA

Contributors/Adopters of ART

- IBM
- Microsoft
- Troj AI
- Two Six Labs, LLC
- Kyushu University
- Intel Corporation
- University of Chicago
- The MITRE Corporation
- General Motors Company
- AGH University of Science and Technology
- Rensselaer Polytechnic Institute (RPI)
- IMT Atlantique

Adversarial Robustness Toolbox





(R)REPEATS
Reproducibility
Robustness
Equitability
Privacy
Explainability
Accountability
Transparency
Security

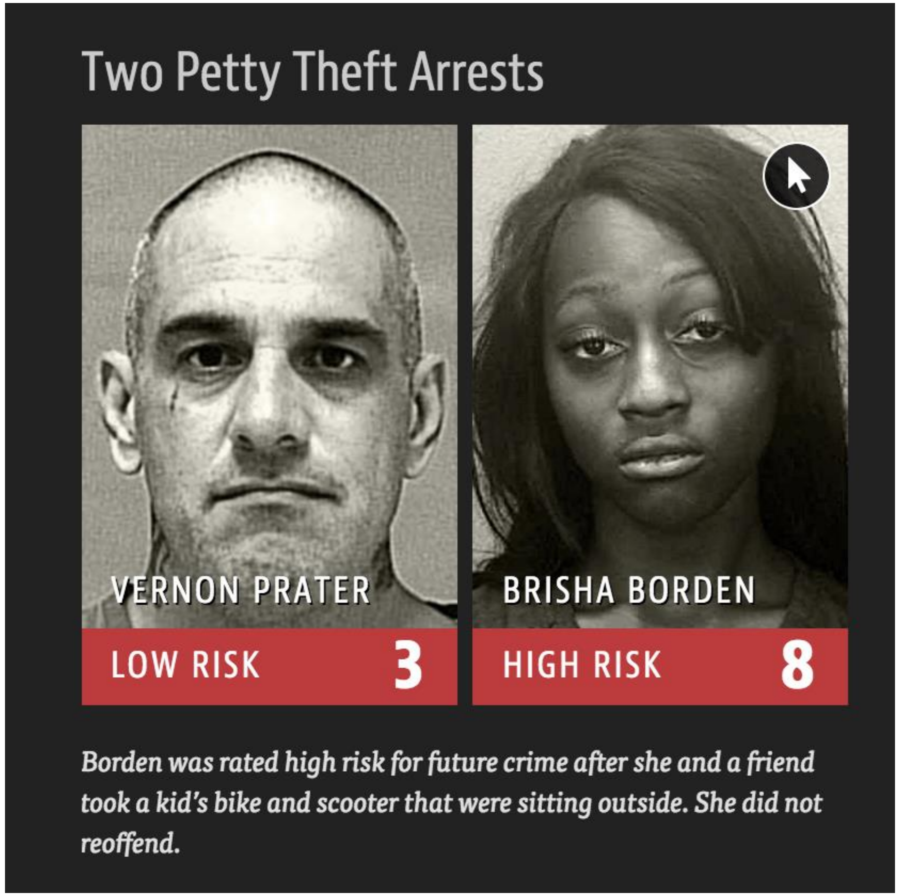
Bias in AI Example: Criminal Justice System

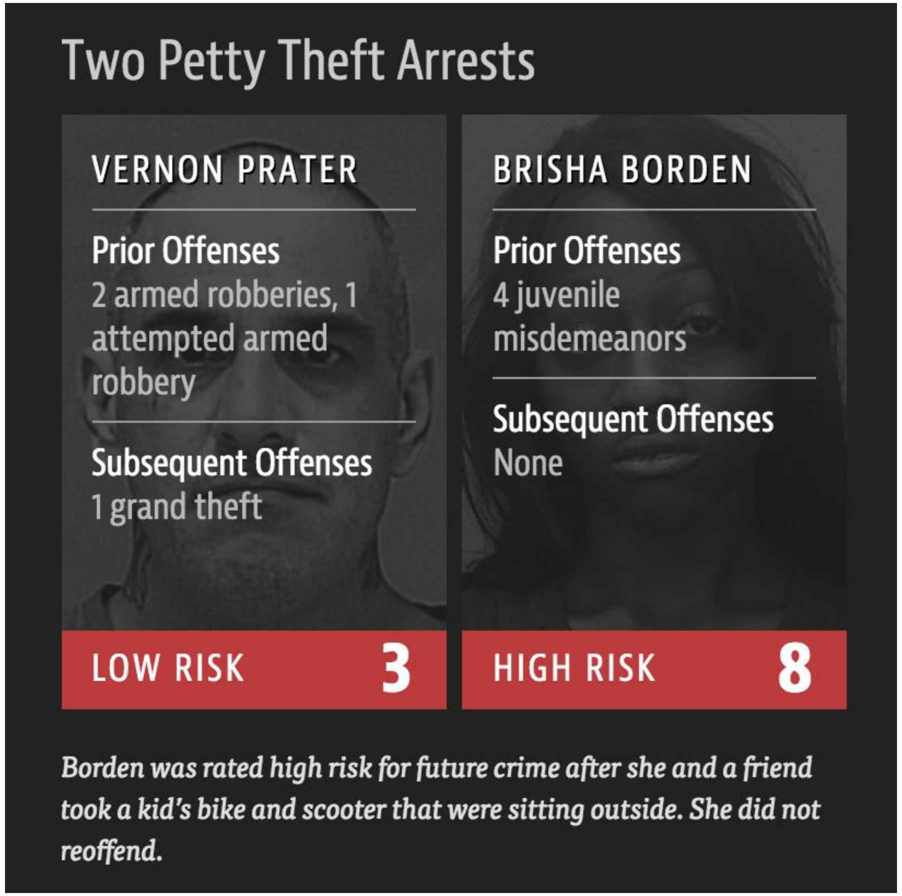
- Since 2008, nearly every arrestee in Broward County, Florida has been assigned a risk score using Northpointe's COMPAS algorithm.
- Defendants with low risk scores are released on bail.
- It falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.

Machine Bias — ProPublica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-

Bias in AI Example: Criminal Justice System

- Since 2008, nearly every arrestee in Broward County, Florida has been assigned a risk score using Northpointe's COMPAS algorithm.
- Defendants with low risk scores are released on bail.
- It falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.

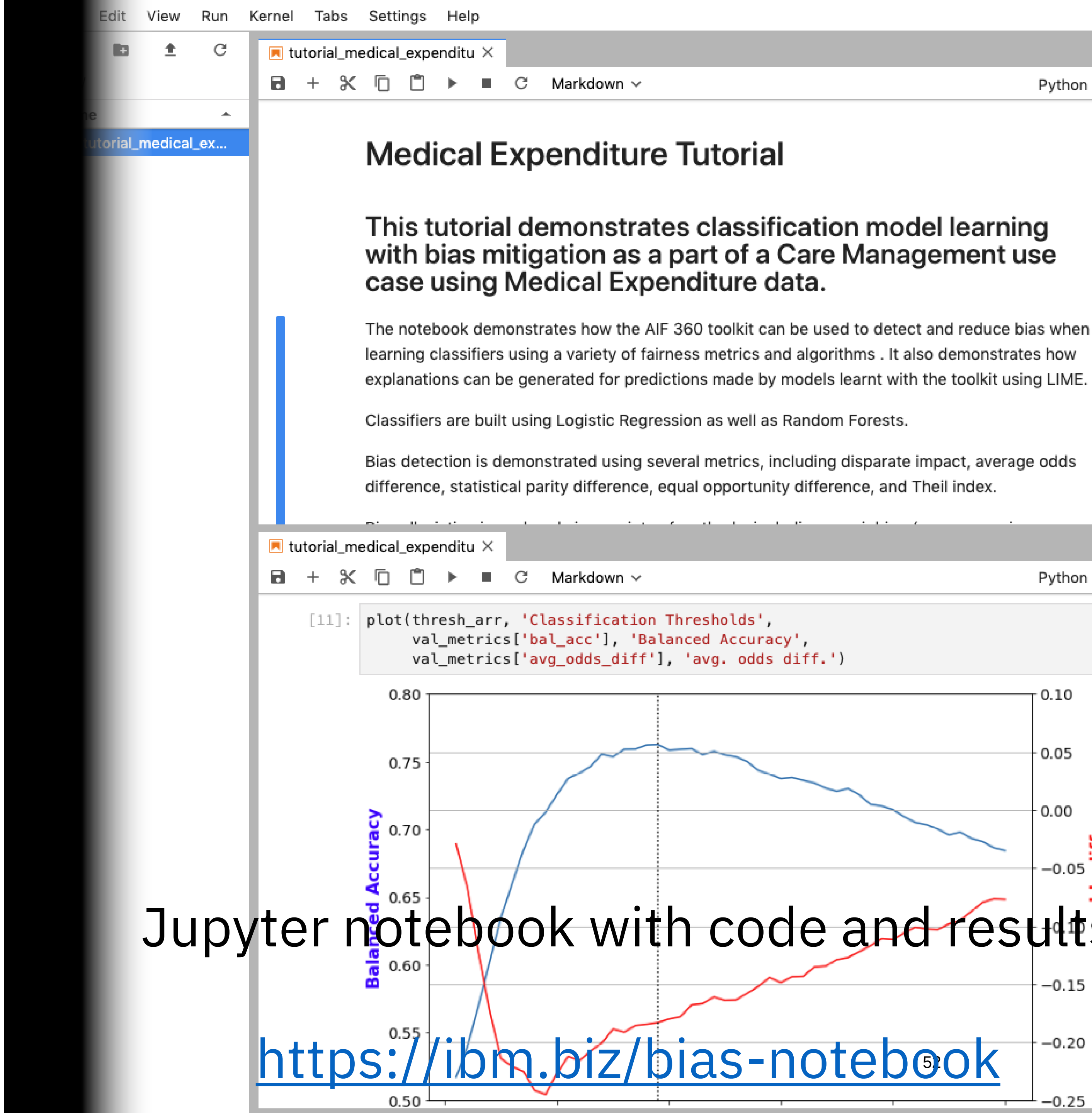




Machine Bias — ProPublica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-

Hypothetical Use Case: Racial Bias in Healthcare

- A healthcare utilization scoring model prioritizes cases for healthcare management.
- IBM Researchers used state-of-the-art techniques to measure and reduce racial bias.
- Problem: Supplemental care decisions can't be predicated on factors such as race of the patient.
- Result: An improved model that is much more fair relative to the original model learned from the original data.



Jupyter notebook with code and results

https://ibm.biz/bias-notebook

Related Real-World Case

(Wall Street Journal October 25, 2019 Page A3)

Some conservative state politicians say Austin is head-

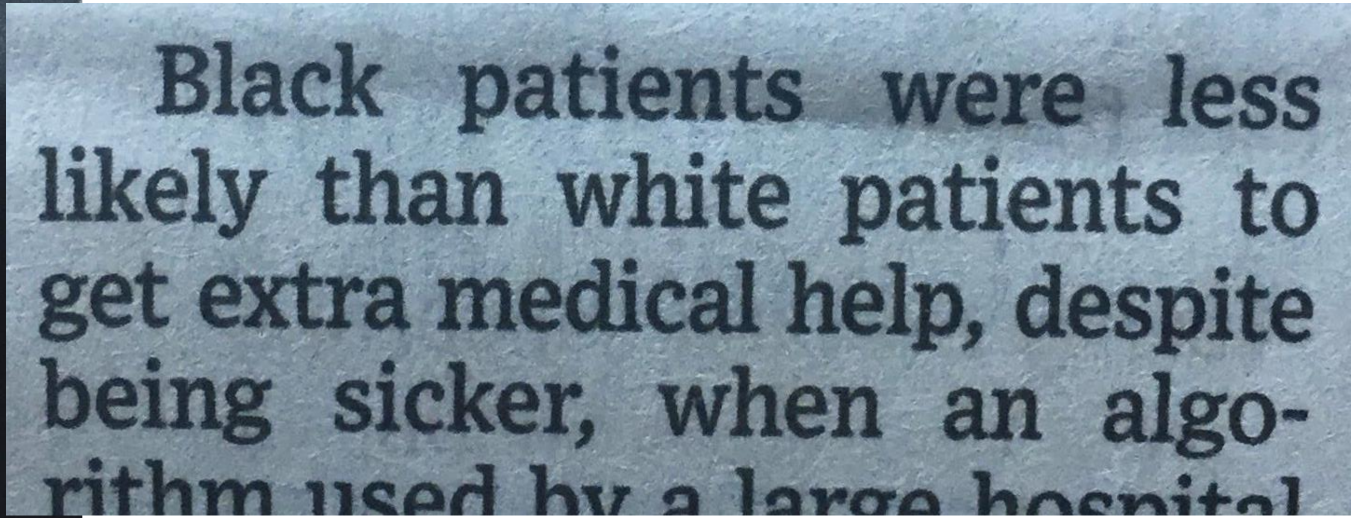
about 1.2 m
has increa

Racial Bias Found in Hospital Algorithm

BY MELANIE EVANS AND ANNA WILDE MATHEWS

Black patients were less likely than white patients to get extra medical help, despite being sicker, when an algorithm used by a large hospital chose who got the additional attention, according to a new study underscoring the risks as technology gains a foothold in medicine.

Hospitals use the algo-



Black patients were less likely than white patients to get extra medical help, despite being sicker, when an algo-

AI Fairness 360
↳ (AIF360)

https://github.com/IBM/AIF360

AIF360 toolkit is an open-source library to help detect and remove bias in machine learning models. **AIF360 translates algorithmic research from the lab into practice**. Applicable domains include finance, human capital management, healthcare, and education.

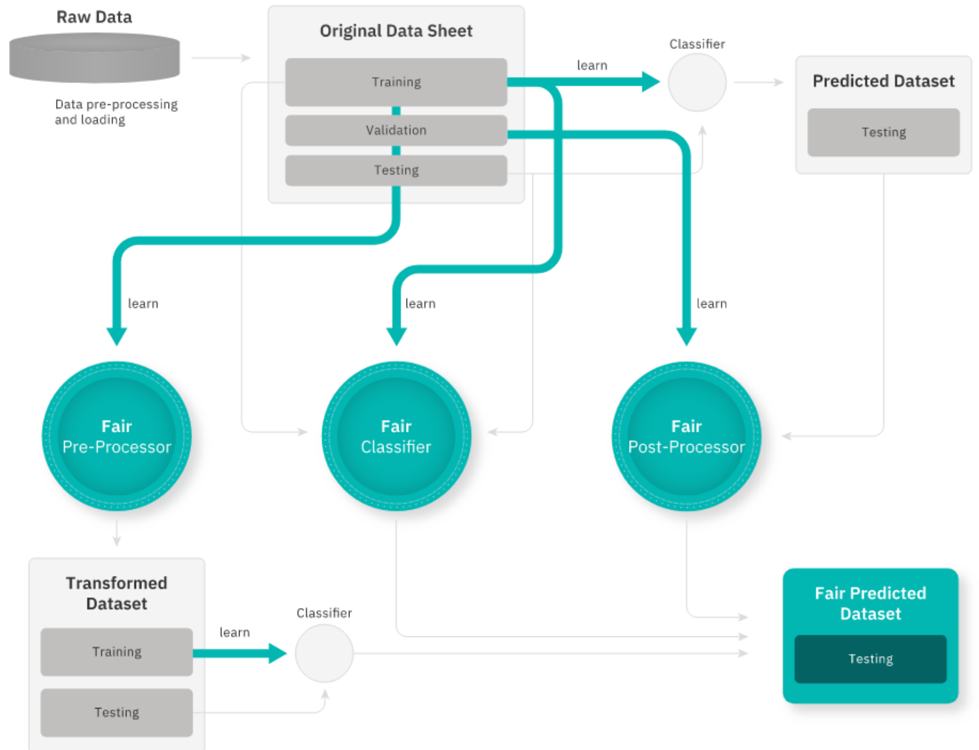
The AI Fairness 360 Python package includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models.

Toolbox

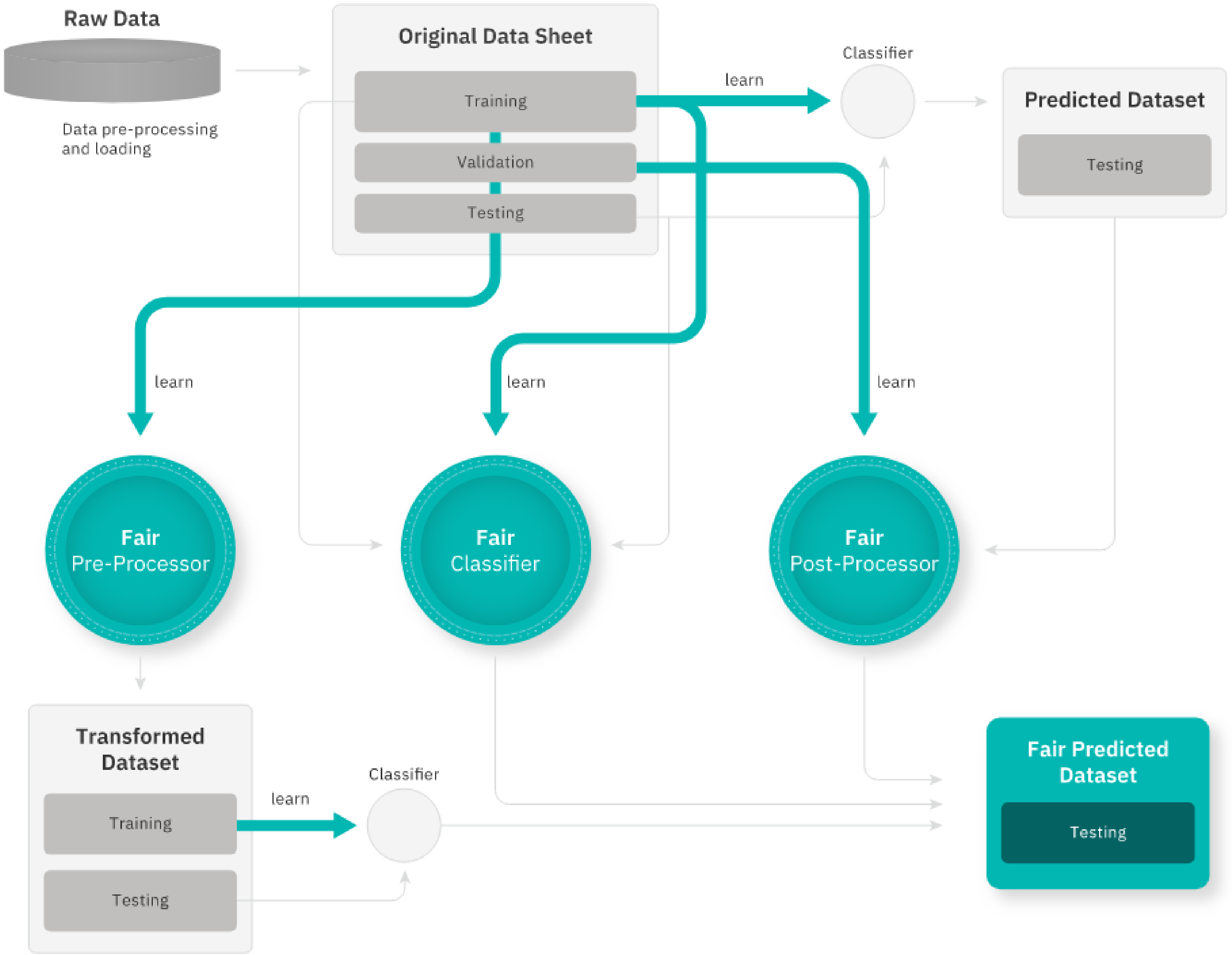
Fairness metrics (70+)
Fairness metric explanations
Bias mitigation algorithms (10+)

http://aif360.mybluemix.net/

- AIF360



AIF 360 detects for fairness in building and deploying models throughout AI Lifecycle



Metrics (70+)

Statistical Parity Difference

The difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.

→

Equal Opportunity Difference

The difference of true positive rates between the unprivileged and the privileged groups.

→

Average Odds Difference

The average difference of false positive rate (false positives/negatives) and true positive rate (true positives/positives) between unprivileged and privileged groups.

→

Disparate Impact

The ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.

→

Theil Index

Measures the inequality in benefit allocation for individuals.

→

Euclidean Distance

The average Euclidean distance between the samples from the two datasets.

→

Mahalanobis Distance

The average Mahalanobis distance between the samples from the two datasets.

→

Manhattan Distance

The average Manhattan distance between the samples from the two datasets.

→

Algorithms (10)

Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.

→

Reweighing

Use to mitgate bias in training data. Modifies the weights of different training examples.

→

Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

→

Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

→

Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.

→

Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.

→

Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

→

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

→

Equalized Odds Post-processing

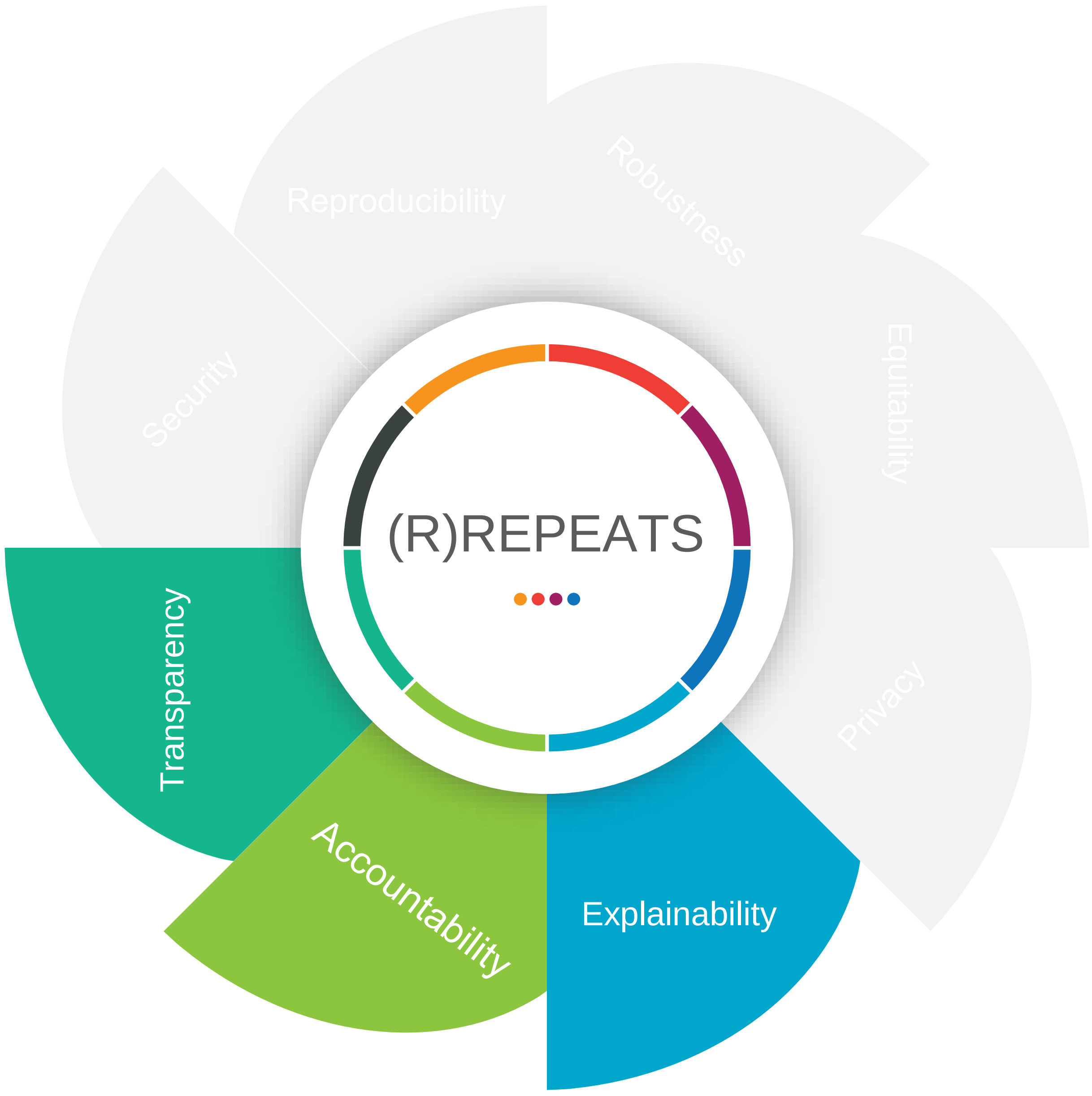
Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

→

Meta Fair Classifier

Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

→



(R)REPEATS
Reproducibility
Robustness
Equitability
Privacy
Explainability
Accountability
Transparency
Security

AI needs to explain its decision, and there are different ways to explain

One explanation does not fit all

Different stakeholders require explanations for different purposes and with different objectives, and explanations will have to be tailored to their needs.

End users/customers (trust)

Doctors: *Why did you recommend this treatment?*

Customers: *Why was my loan denied?*

Teachers: *Why was my teaching evaluated in this way?*

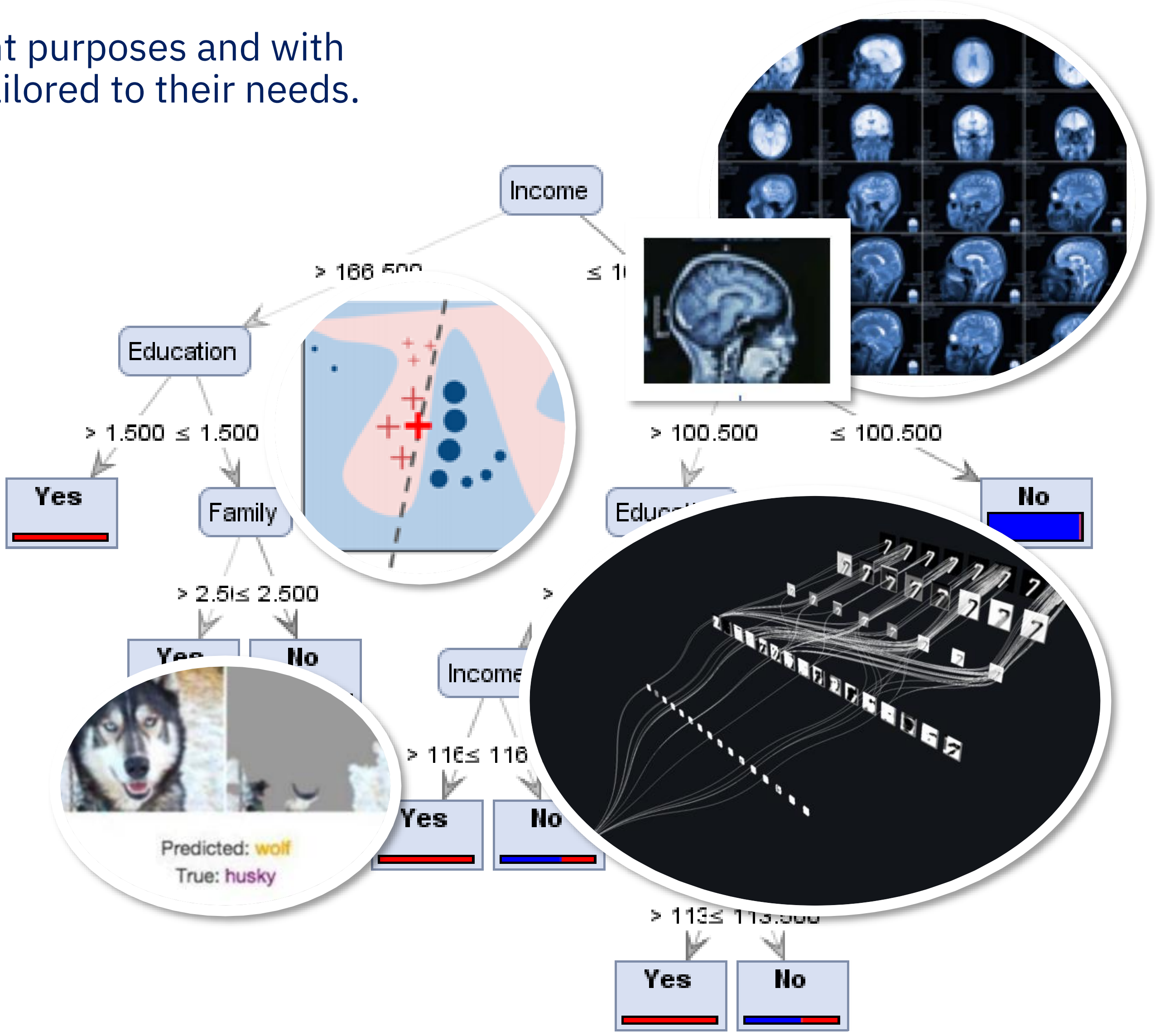
Gov't/regulators (compliance, safety)

Prove to me that you didn't discriminate.

Developers (quality, "debuggability")

Is our system performing well?

How can we improve it?



AI Explainability 360

↳ (AIX360)

https://github.com/IBM/AIX360

AIX360 toolkit is an open-source library to help explain AI and machine learning models and their predictions. This includes three classes of algorithms: local post-hoc, global post-hoc, and directly interpretable explainers for models that use image, text, and structured/tabular data.

The AI Explainability360 Python package includes a comprehensive set of explainers, both at global and local level.

Toolbox

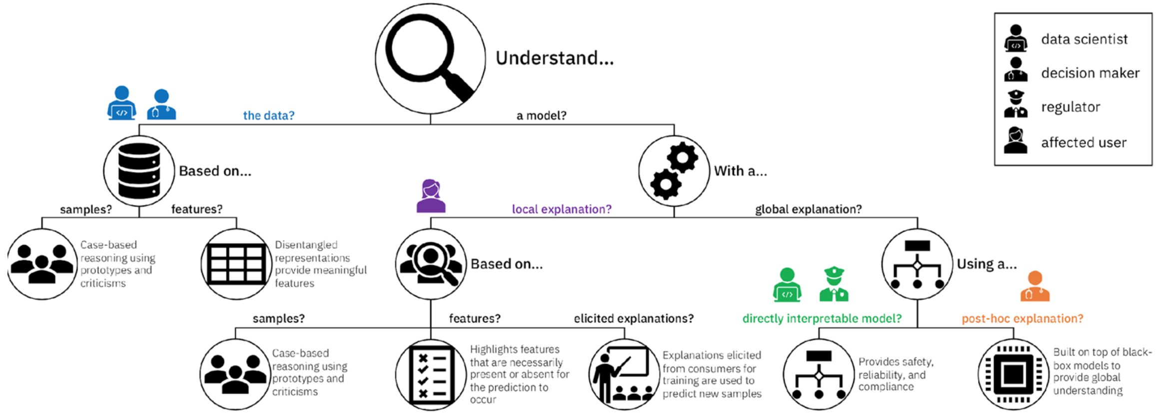
Local post-hoc

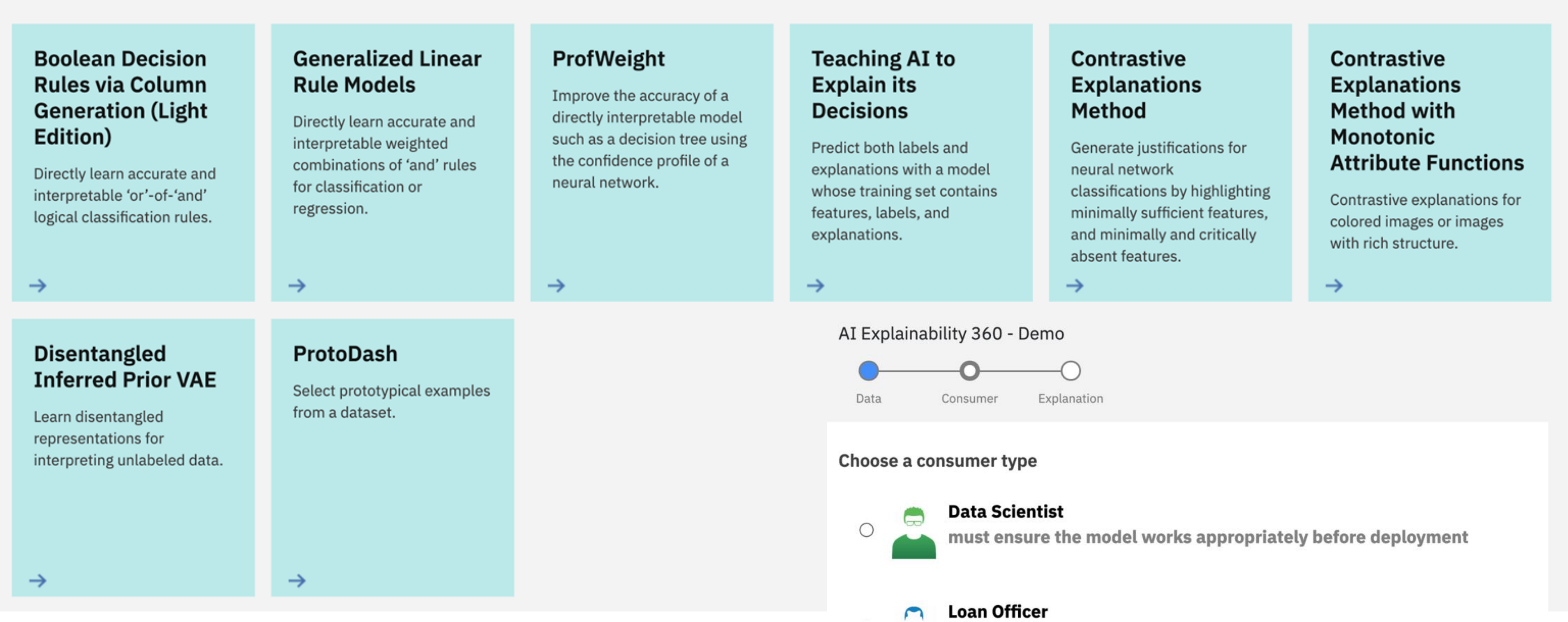
Global post-hoc

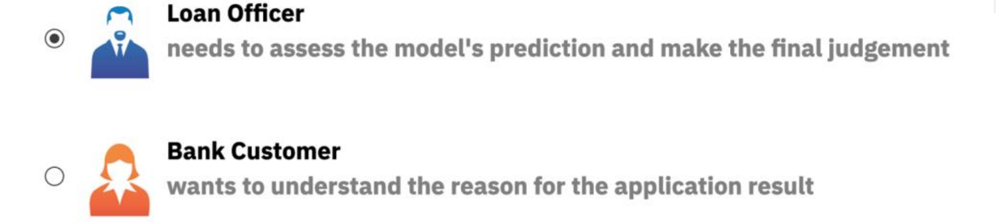
Directly interpretable

http://aix360.mybluemix.net

- AIX360





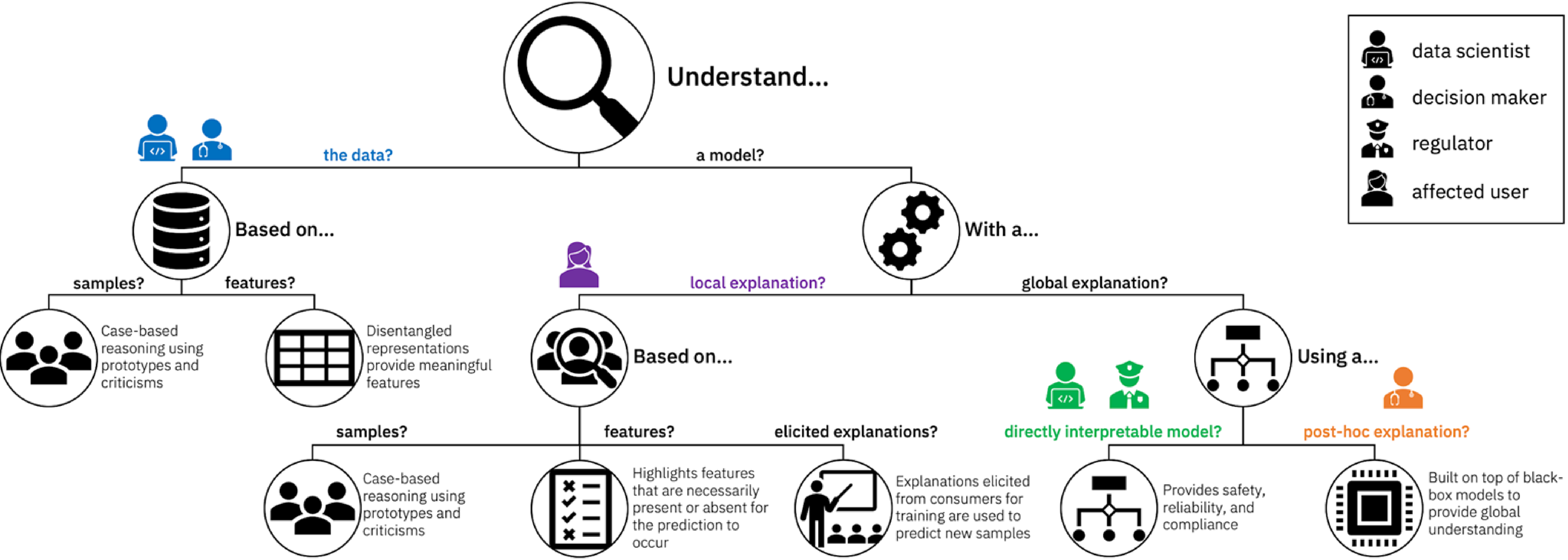


AI Explainability 360

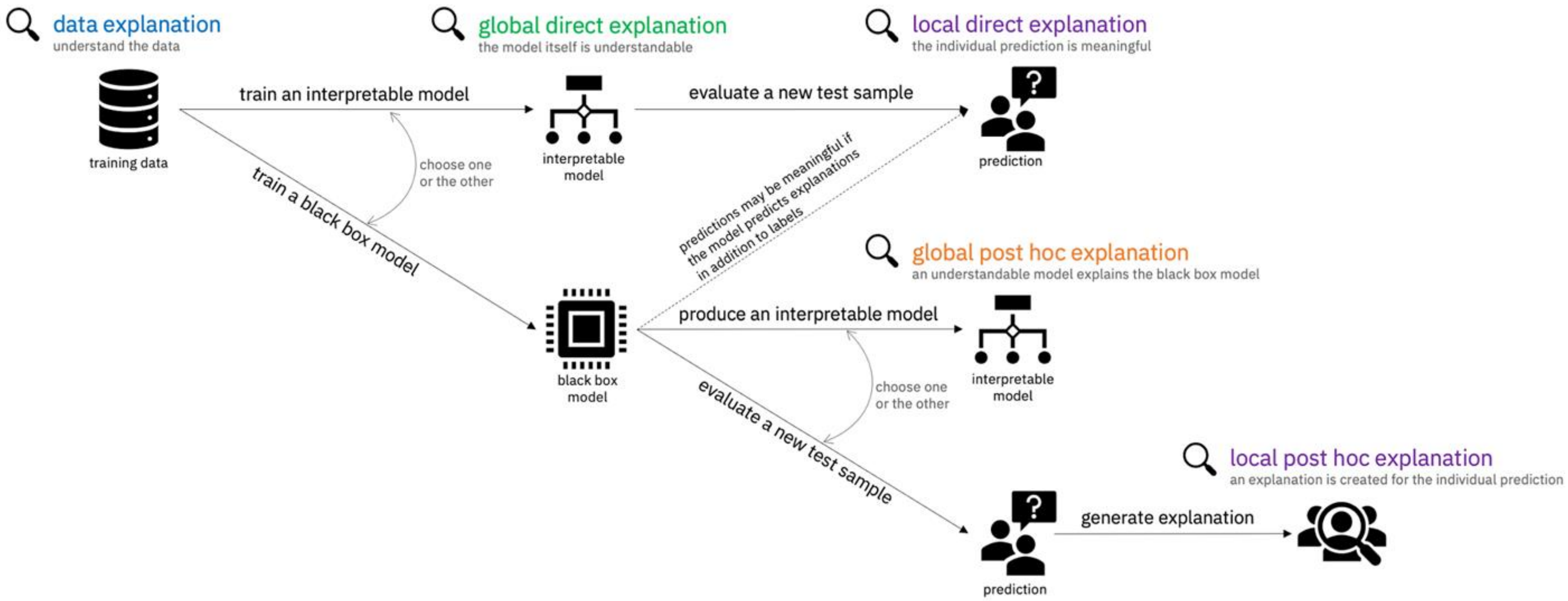
The most comprehensive **open source** toolkit for explaining ML models and data:

- 8 innovated algorithms from IBM Research + 2 from scientific community
- 13 tutorial notebooks covering use cases in finance, healthcare, lifestyle, retention, etc.

AIX360: Multiple dimensions of explainability



AIX360: Multiple dimensions of explainability



LFAI Trusted AI Projects Update

AI Fairness 360

AI Explainability 360

Adversarial Robustness Toolbox

ART	Very strong growth and adoption.	V1.6.2 released.	2.3K GitHub Stars, 150K Downloads, 8K+ Commits
AIF360:	Strong growth trajectory	V0.4.0 released	1.4K GitHub Stars, 330+Commits
AIX360	Used extensively within enterprises	V0.2.1 released	~850 GitHub Stars, 10K downloads/month, 240+Commits

https://ai-fairness-360.org/
https://ai-explainability-360.org/
https://adversarial-robustness-toolbox.org/

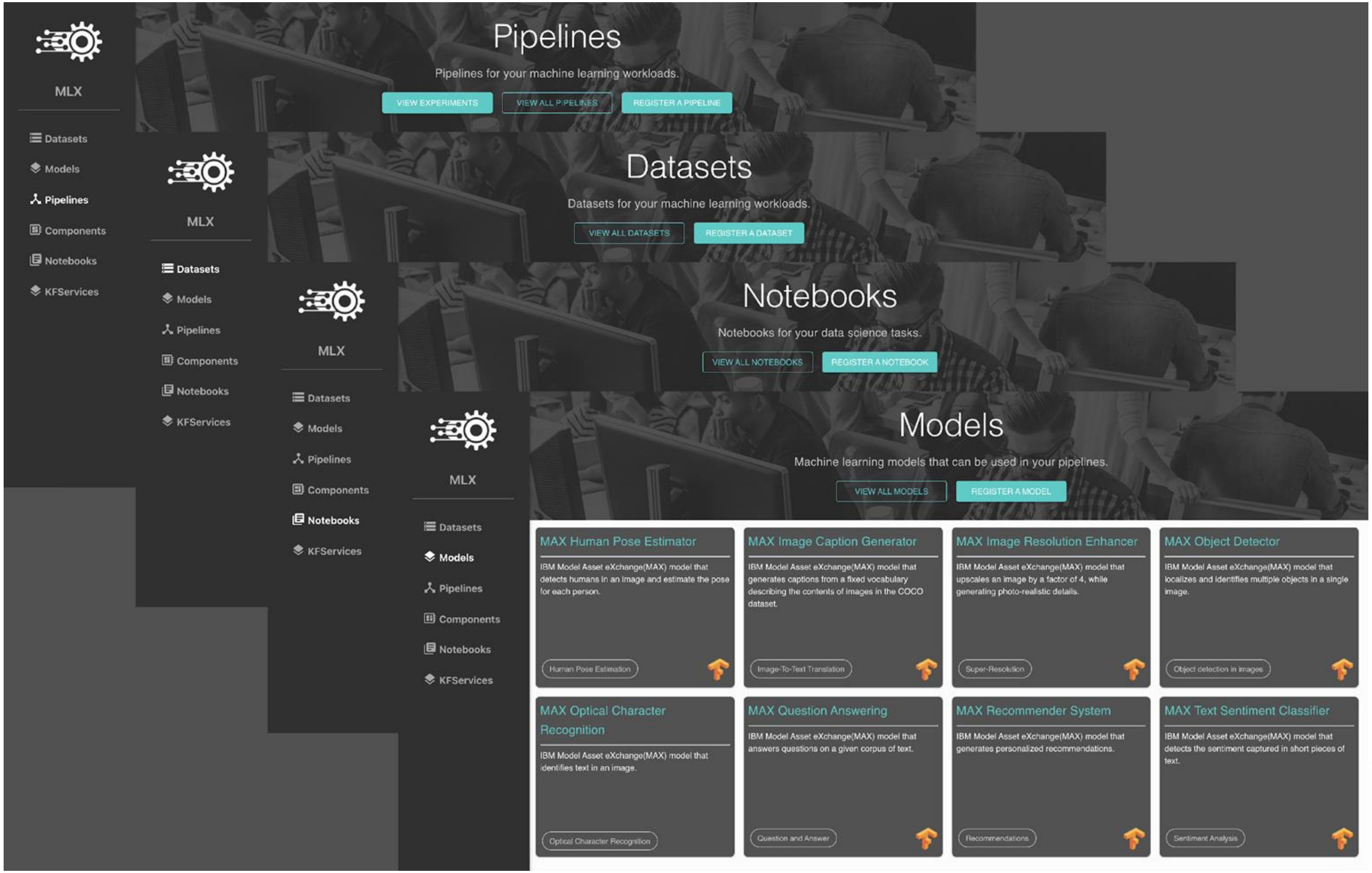
LFAI & Data

Trusted AI and MLOps

LF AI & DATA

Animesh Singh,
Watson CTO, Open Technology,
Distinguished Engineer,
IBM

Machine Learning eXchange (MLX): Data and AI Assets Catalog and Execution Engine



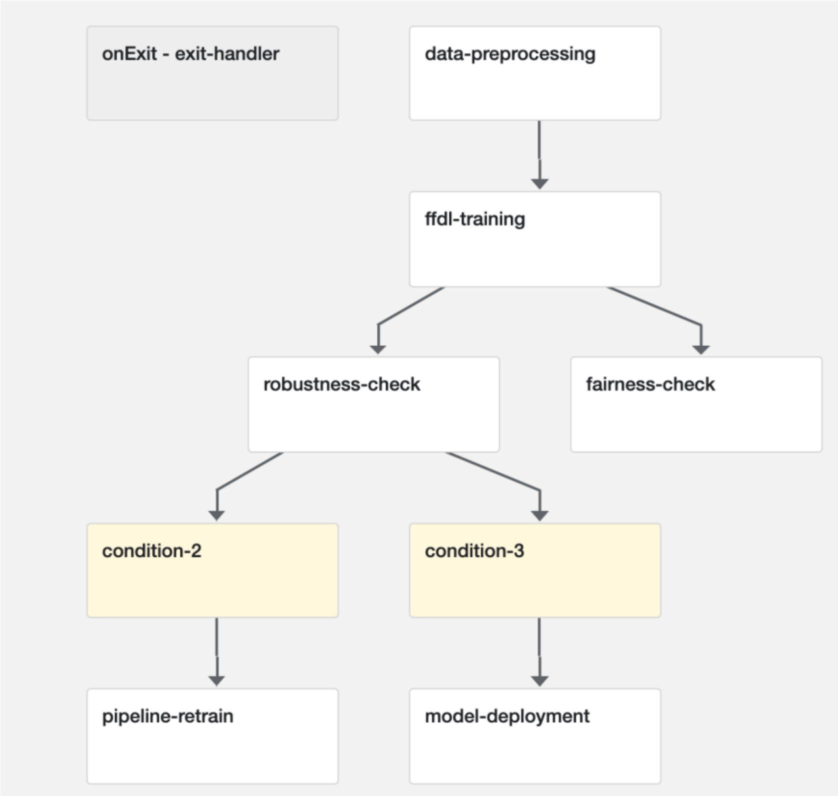
THE LINUX FOUNDATION

Trusted AI Pipelines available in MLX

AI Fairness 360

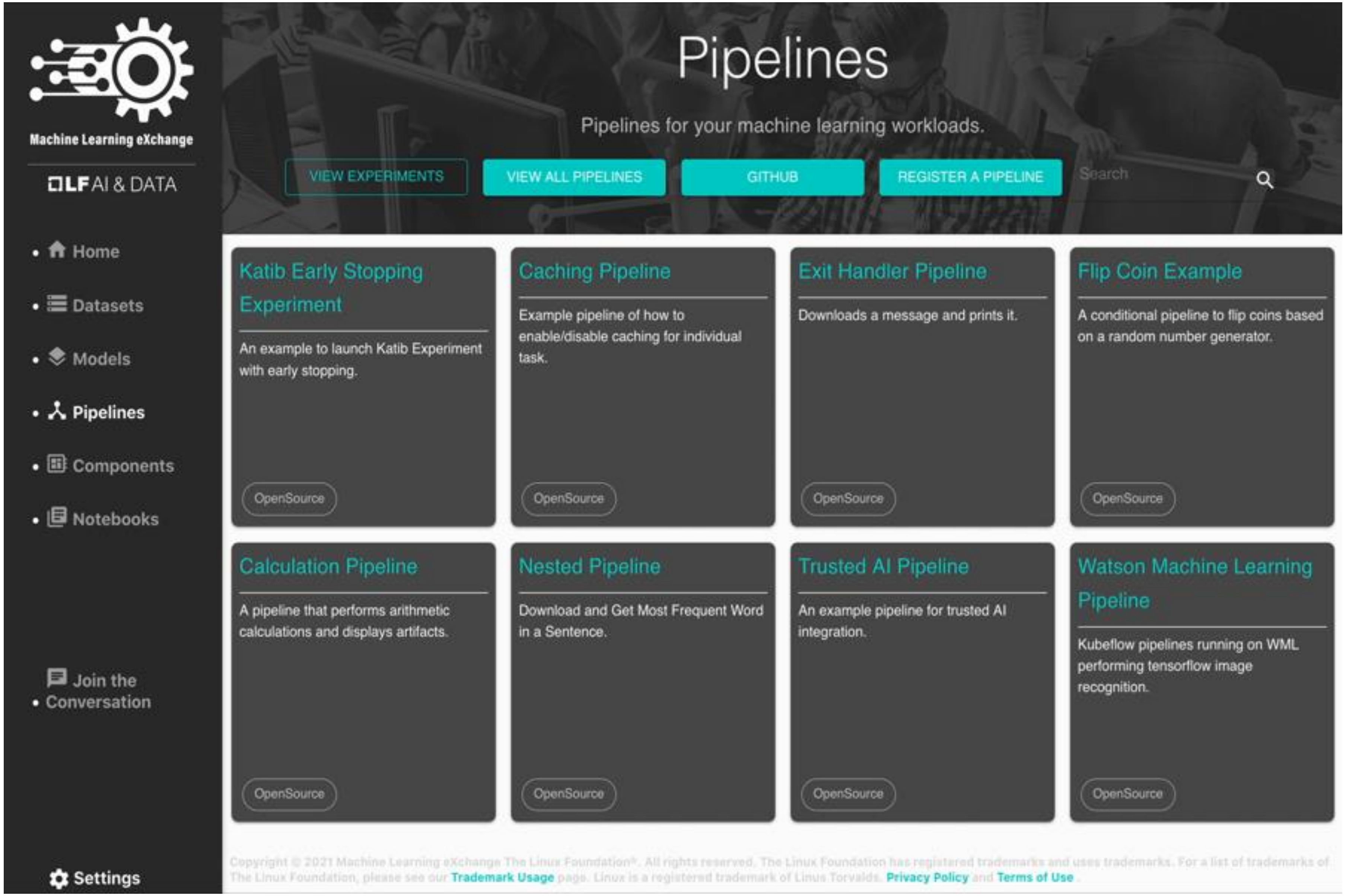
AI Explainability 360

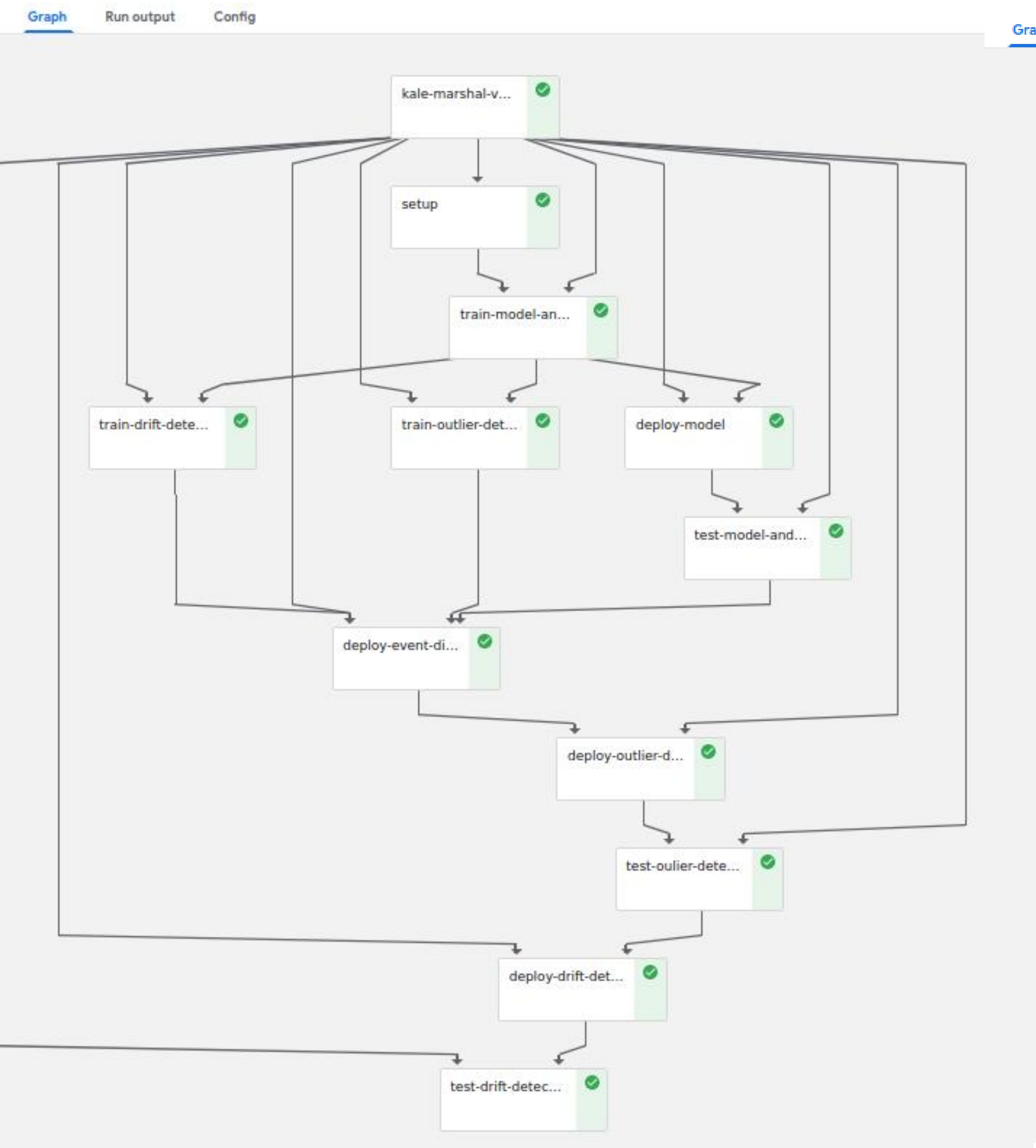
Adversarial Robustness Toolbox



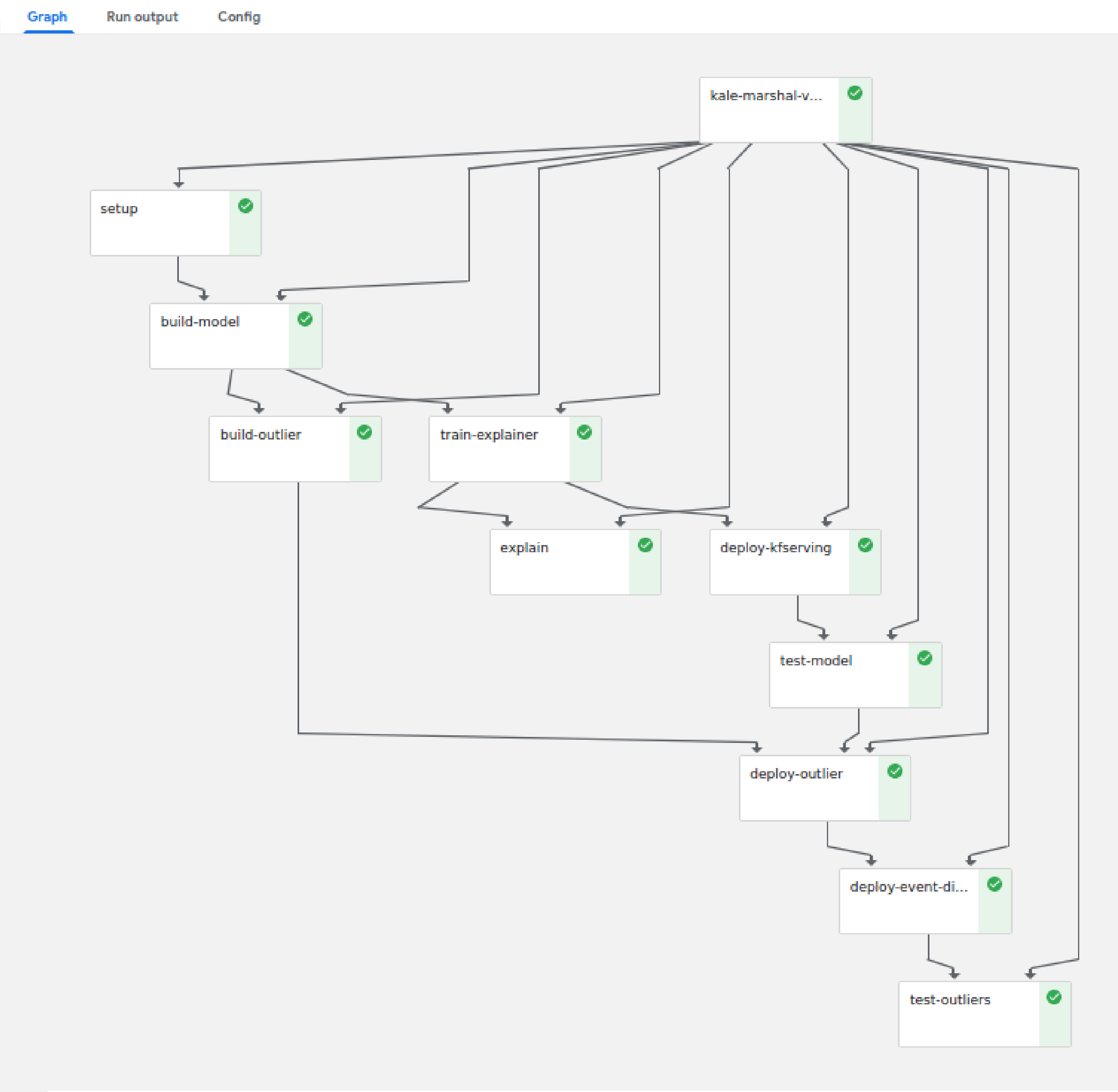
https://ai-fairness-360.org/
https://ai-explainability-360.org/
https://adversarial-robustness-toolbox.org/

Trusted AI Pipeline



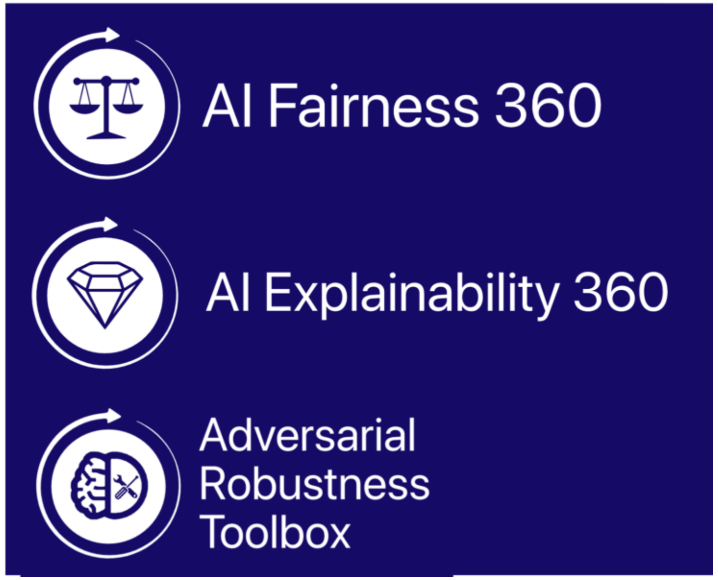


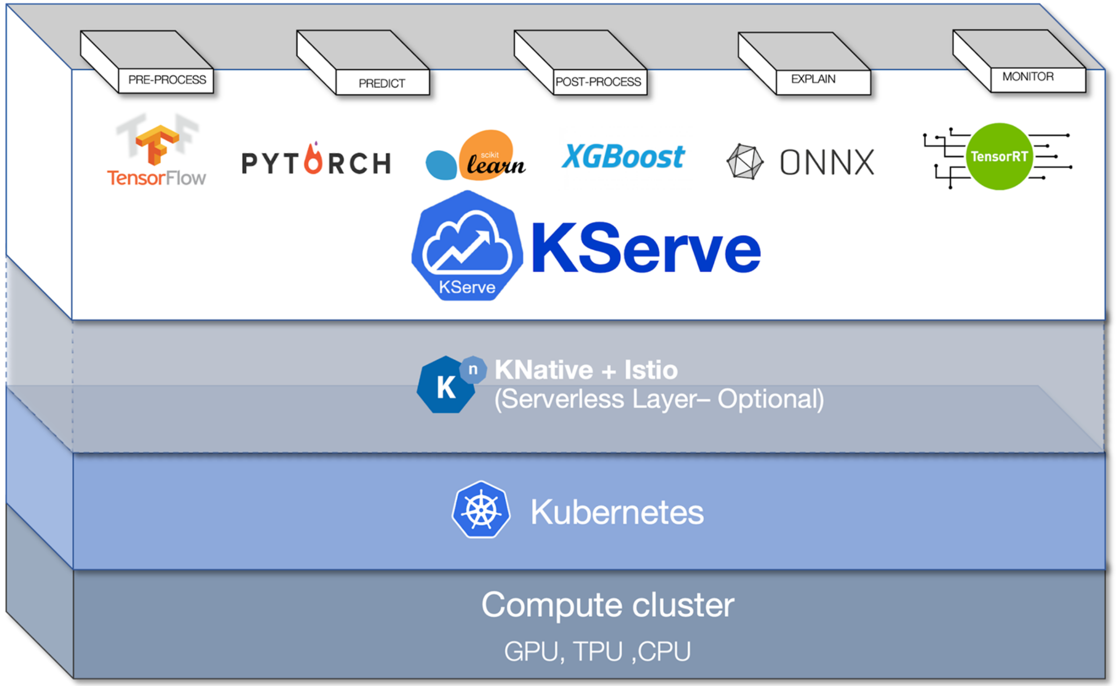
Graph
Run output
Config
kale-marshal-v...
setup
train-model-an...
train-drift-dete...
train-outlier-det...
deploy-model
test-model-and...
deploy-event-di...
deploy-outlier-d...
test-oulier-dete...
deploy-drift-det...
test-drift-detec...



Graph
Run output
Config
kale-marshal-v...
setup
build-model
build-outlier
train-explainer
explain
deploy-kfserving
test-model
deploy-outlier
deploy-event-di...
test-outliers

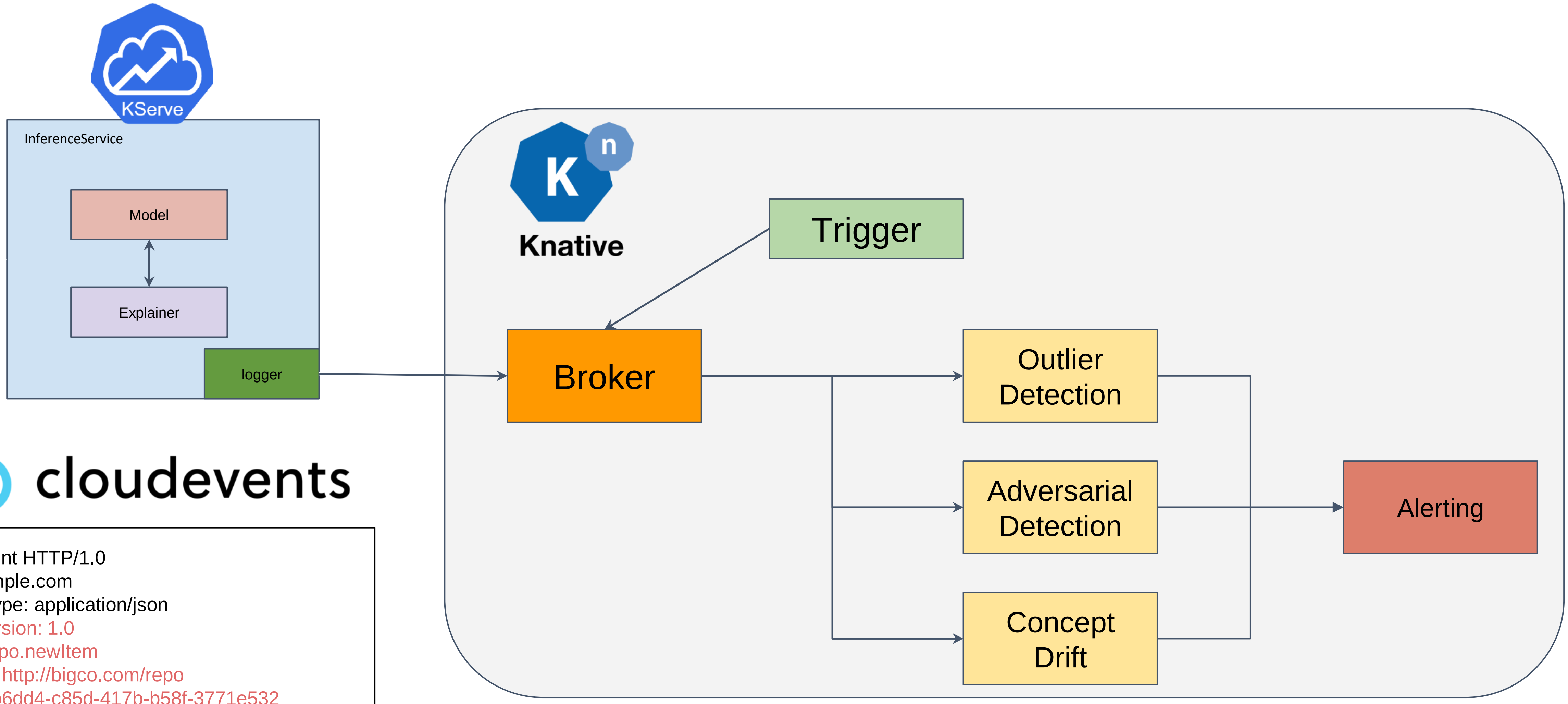
Trusted AI with KServe





https://ai-fairness-360.org/
https://ai-explainability-360.org/
https://adversarial-robustness-toolbox.org/

And more advanced Metrics



cloudevents

```
POST /event HTTP/1.0
Host: example.com
Content-Type: application/json
ce-specversion: 1.0
ce-type: repo.newItem
ce-source: http://bigco.com/repo
ce-id: 610b6dd4-c85d-417b-b58f-3771e532

<payload>
```

LFAI & Data

The Trusted AI Principles - Tools and Techniques

Education Opportunities

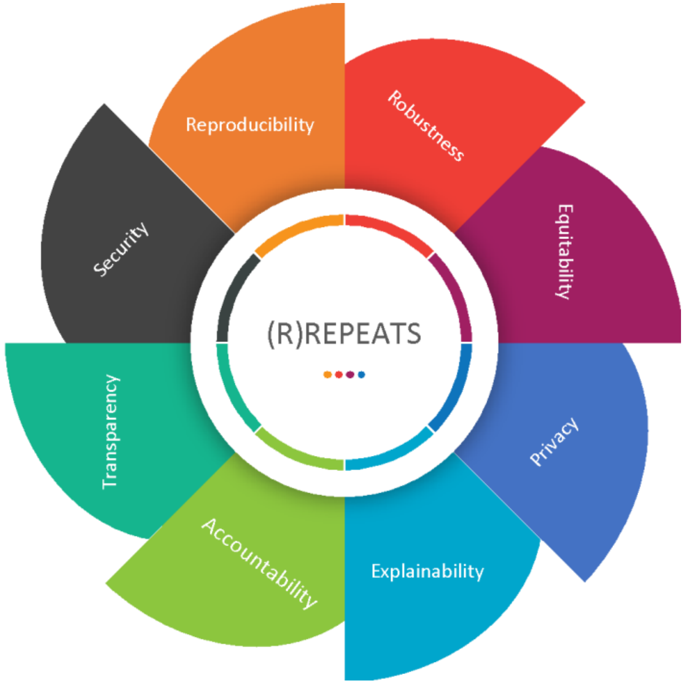
Haluk Demirakan
(University of Washington, Tacoma)

LF AI & DATA

Education Opportunities

Principles Working Group Team @ LF AI & Data is in a process of developing courses to educate our community

- Currently, LF AI & Data is offering a course “Ethics in AI and Data Science” (https://www.edx.org/course/ethics-in-ai-and-data-science) through EdX
- In addition, Principles Working Group Team is developing a series of courses to educate our community about Trusted AI Principles. Please let us know if you would like to join forces to educate our community about importance of Trusted AI.
- Join the mailing list here and participate in an upcoming meeting! Learn more here



LFAI & Data

The Trusted AI Principles - Tools and Techniques

Summary

François Jézéquel, Director of Business Development, Orange Fab

LF AI & DATA

Thank you

Stay connected with the Trusted AI Committee by joining the mailing list here and participate in an upcoming meeting!

Learn more here

References and Resources

- Trusted AI Committee - Principles Working Group (where you will find the slides and materials)
- [LF-AI] The Trusted-AI Principles document bit.ly/lfai-trustedai-principles

- [LF-AI Blog] LF AI & Data Announces Principles for Trusted AI
- [LF-AI Webinar] RREPEATS – An Introduction to the Principles for Trusted AI – Thoughts and Next Steps
- [LF-AI Webinar] RREPEATS Practical Examples Review

- [ACM] ACM Principles for Algorithmic Transparency and Accountability https://www.acm.org/binaries/content/assets/publicpolicy/2017_usacm_statement_algorithms.pdf
- [EU] Ethics Guidelines for Trustworthy AI - High-Level Expert Group on Artificial Intelligence set up by the European Commission https://ec.europa.eu/futurium/en/ai-alliance-consultation
- [EUFeb2020] On Artificial Intelligence -A European approach to excellence and tru https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
- [IEEE] Ethically Aligned Design, IEEE https://ethicsinaction.ieee.org/
- [DoD] AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF
- [OECD] Organisation for Economic Co-operation and Development https://www.oecd.org/going-digital/ai/principles/
- [SoA] State of the Art: Reproducibility in Artificial Intelligence Odd Erik Gundersen, Sigbjørn Kjensmo, Department of Computer Science Norwegian University of Science and Technology https://www.researchgate.net/publication/326450530_State_of_the_Art_Reproducibility_in_Artificial_Intelligence

Contributions

Principles Working Group Team:

- Souad Ouali (Orange)
- Jeff Cao (Tencent)
- Haluk Demirakan (University of Washington, Tacoma)
- Francois Jezequel (Orange)
- Sri Kishnamurthym (Quant University)
- Layla Li (KOSA)
- Sarah Luger (Orange)
- Susan Malaika (IBM)
- Alka Roy (The Responsible Innovation Project/ex-AT&T)
- Alejandro Saucedo (The Institute for Ethical AI / Seldon)
- Animesh Singh (IBM)
- Marta Ziosi (AI for People)