AI

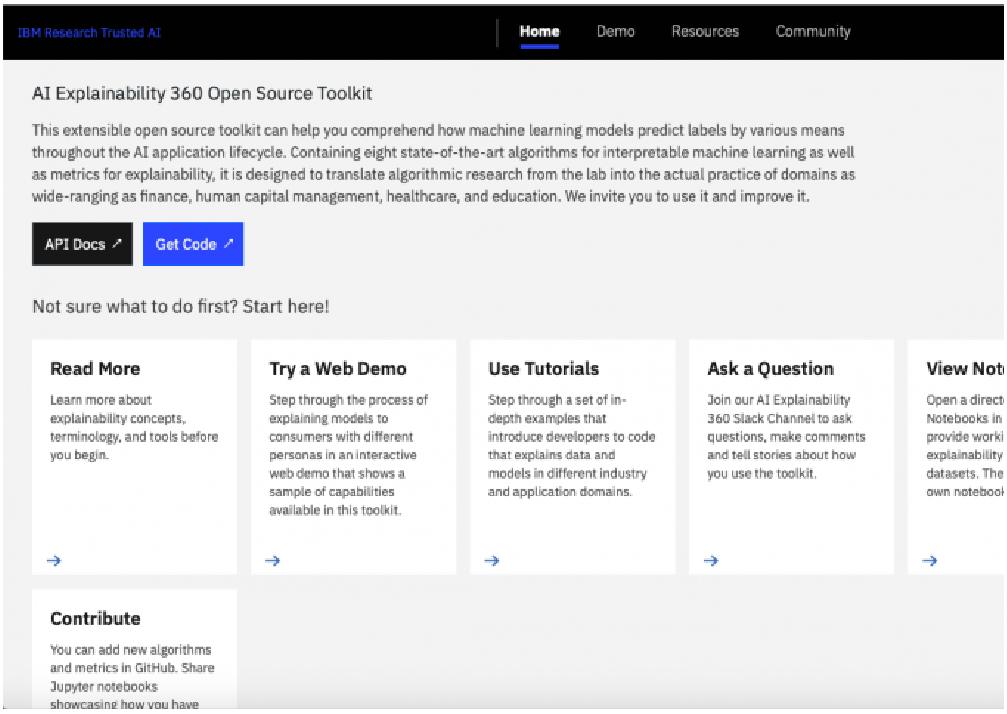
Introducing AI Explainability 360

August 8, 2019 | Written by: Aleksandra Mojsilovic

Categorized: AI

Share this post:

We are pleased to announce AI Explainability 360, a comprehensive open source toolkit of state-
the interpretability and explainability of machine learning models. We invite you to use it and cont
theory and practice of responsible and trustworthy AI.



AI Explainability 360 Toolkit

"explainability" or "interpretability," allows users to gain insight into the machine's decision-maki
things work is essential to how we navigate the world around us and is essential to fostering trust

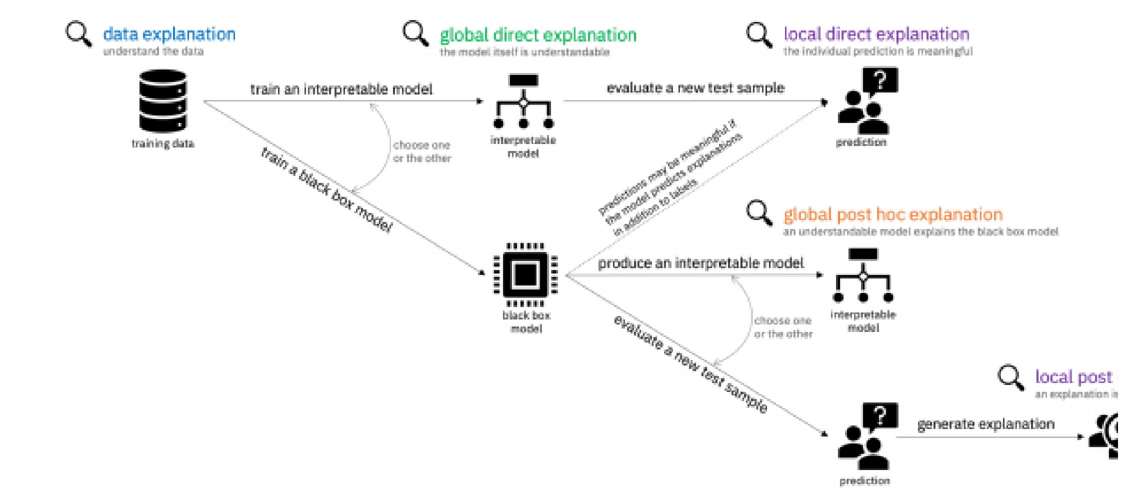
Further, AI explainability is increasingly important among business leaders and policymakers. In f
believe that customers will demand more explainability from AI in the next three years, according
Value survey.

No single approach to explaining algorithms

To provide explanations in our daily lives, we rely on a rich and expressive vocabulary: we use exa
rules and prototypes, and highlight important characteristics that are present and absent.

When interacting with algorithmic decisions, users will expect and demand the same level of expr
diagnosing a patient may benefit from seeing cases that are very similar or very different. An appli
want to understand the main reasons for the rejection and what she can do to reverse the decisio
not probe into only one data point and decision, she will want to understand the behavior of the sy
complies with regulations. A developer may want to understand where the model is more or less c
performance.

As a result, when it comes to explaining decisions made by algorithms, there is no single approac
ways to explain. The appropriate choice depends on the persona of the consumer and the require
pipeline.

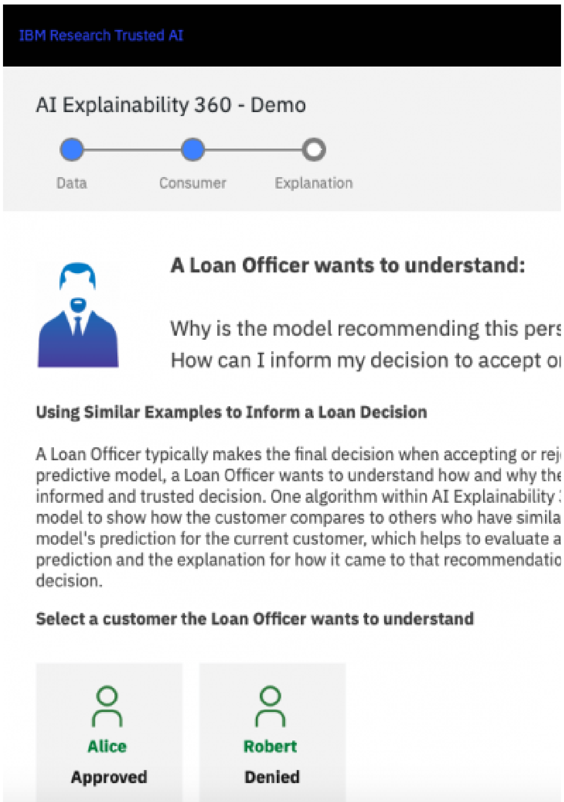


AI Explainability 360 Usage Diagram

AI Explainability 360 tackles explainability in a single interface

different explanation options, we have created helpful resources in a single place:

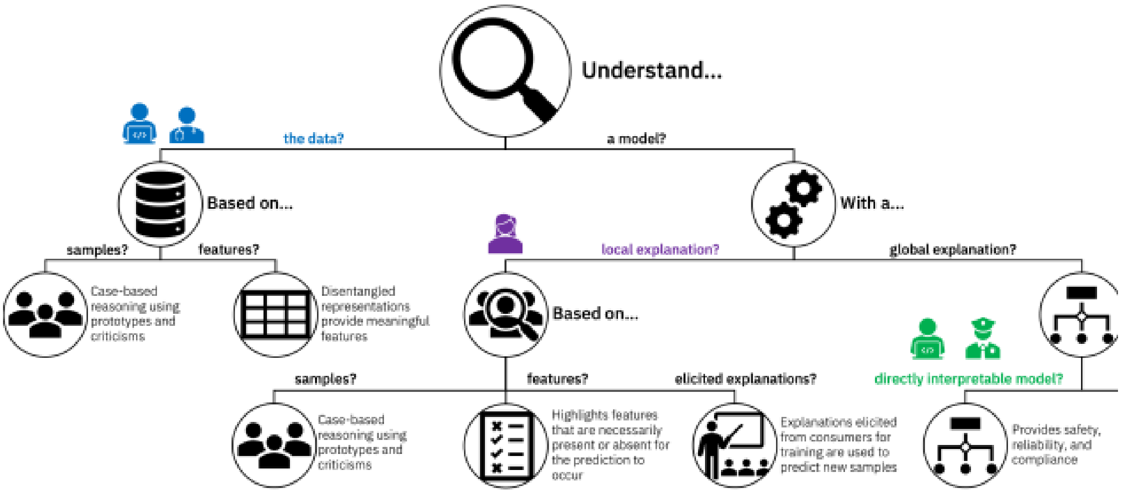
- an interactive experience that provides a gentle introduction through a credit scoring application;
- several detailed tutorials toeducate practitioners on how to inject explainability in other high-stakes applications such as clinical medicine, healthcare management and human resources;
- documentation that guides the practitioner on choosing an appropriate explanation method.



AI Explainability 360 Demo

The toolkit has been engineered with a common interface for all of the different ways of explaining (not an easy feat) and is extensible to accelerate innovation by the community advancing AI explainability. We are open sourcing it to help create a community of practice for data scientists, policymakers, and the general public that need to understand how algorithmic decision making affects them. AI Explainability 360 differs from other open source explainability offerings [1] through the diversity of its methods, focus on educat extensibility via a common framework. Moreover, it interoperates with AI Fairness 360 and Advers open-source toolboxes from IBM Research released in 2018, to support the development of holis pipelines.

The initial release contains eight algorithms recently created by IBM Research, and also includes serve as quantitative proxies for the quality of explanations. Beyond the initial release, we encoura from the broader research community.



AI Explainability 360 Decision Tree

We highlight two of the algorithms in particular. The first, Boolean Classification Rules via Column scalable method of directly interpretable machine learning that won the inaugural FICO Explainab second, Contrastive Explanations Method, is a local post hoc method that addresses the most imp AI that has been overlooked by researchers and practitioners: explaining why an event happened instead of some other event.

AI Explainability 360 complements the ground-breaking algorithms developed by IBM Research t Released last year, the platform helps clients manage AI transparently throughout the full AI lifec applications were built or in which environment they run. OpenScale also detects and addresses b applications, as those applications are being run.

Our team includes members from IBM Research from around the globe. [2] We are a diverse grou scientific discipline, gender identity, years of experience, appetite for vindaloo, and innumerable o belief that the technology we create should uplift all of humanity and ensure the benefits of AI are

The toolkit includes algorithms and metrics from the following papers:

- David Alvarez-Melis and Tommi Jaakkola, "Towards Robust Interpretability with Self-Explainir Neural Information Processing Systems, 2018.
- Sanjeeb Dash, Oktay Günlük, and Dennis Wei, "Boolean Decision Rules via Column Generatior Processing Systems, 2018.
- Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmu Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives", Conferenc Systems, 2018.
- Amit Dhurandhar, Karthikeyan Shanmugam,Ronny Luss, and Peder Olsen, "Improving Simple Conference on Neural Information Processing Systems, 2018.
- Karthik S. Gurumoorthy, Amit Dhurandhar, and Guillermo Cecchi, and Charu Aggarwal. Efficier Prototypes with Importance Weights. IEEE International Conference on Data Mining (ICDM), 2
- Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Ramamurthy, and Kush R. Varshney, "TED: Teaching AI to Explain Its Decisions", AAAI/ACM C Ethics, and Society, 2019.
- Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan, "Variational Inference of Dise Unlabeled Data", International Conference on Learning Representations, 2018.
- Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Karthikeyan Shanmugam, and Contrastive Explanations with Monotonic Attribute Functions", 2019.
- Dennis Wei, Sanjeeb Dash, Tian Gao, and Oktay Günlük, "Generalized Linear Rule Models", Int Learning, 2019.

[2] AI Explainablity 360 includes work from IBM Research – India, the IBM T. J. Watson Research Cambridge in the United States, and the IBM Argentina SilverGate Team.Team members include V Chen, Amit Dhurandhar, Mike Hind, Sam Hoffman, Stephanie Houde, Vera Liao, Ronny Luss, Aleksa Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Moninder Sin Varshney, Dennis Wei, and Yunfeng Zhang.

Aleksandra Mojsilovic
IBM Fellow, AI Science, IBM Research

AI Explainability 360 | explainability | trusted AI

Previous Post

Answering Complex Questions using Neural Program Induction

Bank G

More AI stories

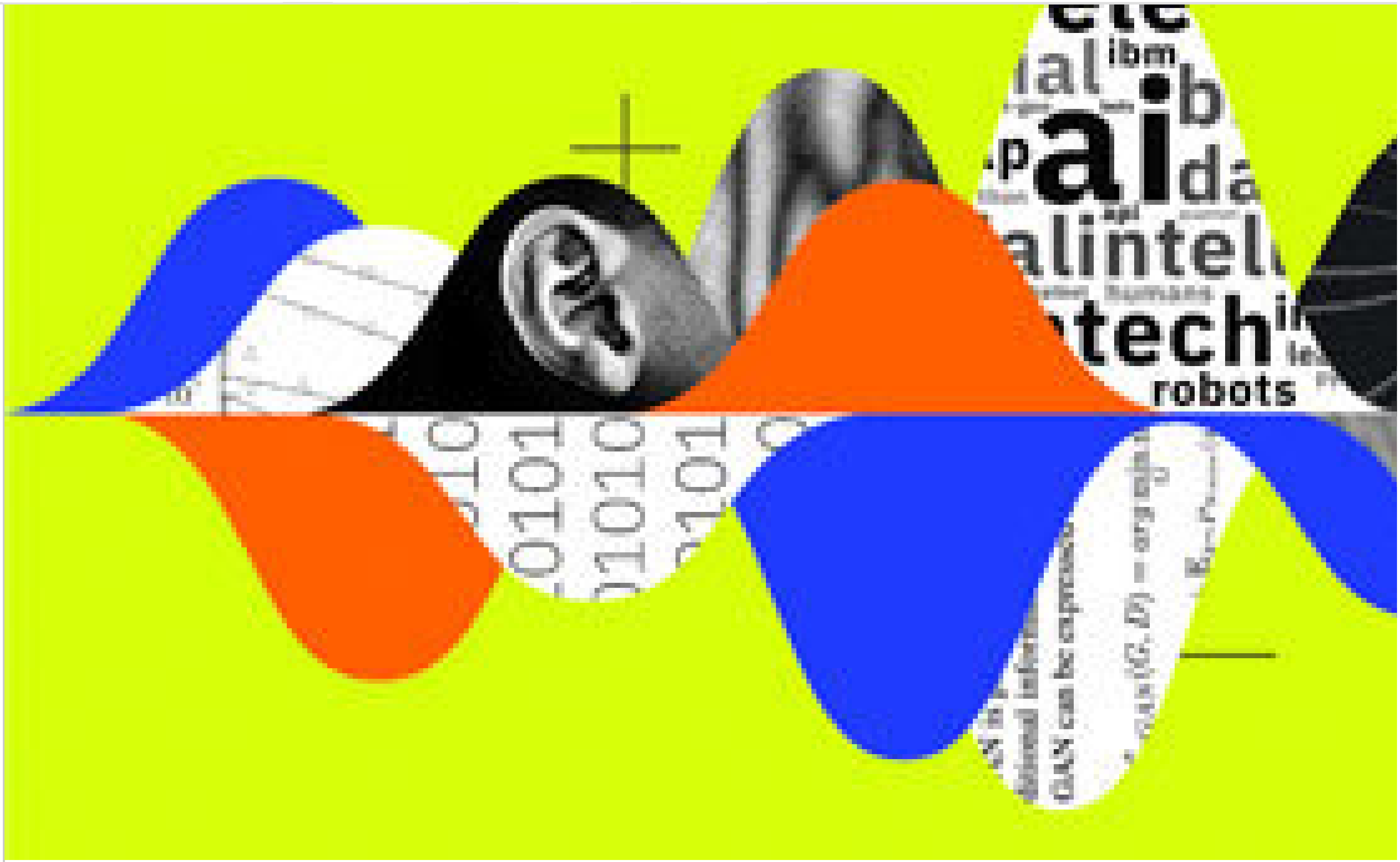
AI

Simplifying data: IBM's AutoAI automates time series foreca

In our recent paper "AutoAI-TS: AutoAI for Time Series Forecasting," which we'll present at AC
Series for Watson Studio incorporates the best-performing models from all possible classes —
technique that performs best across all datasets.

→ Continue reading

AI

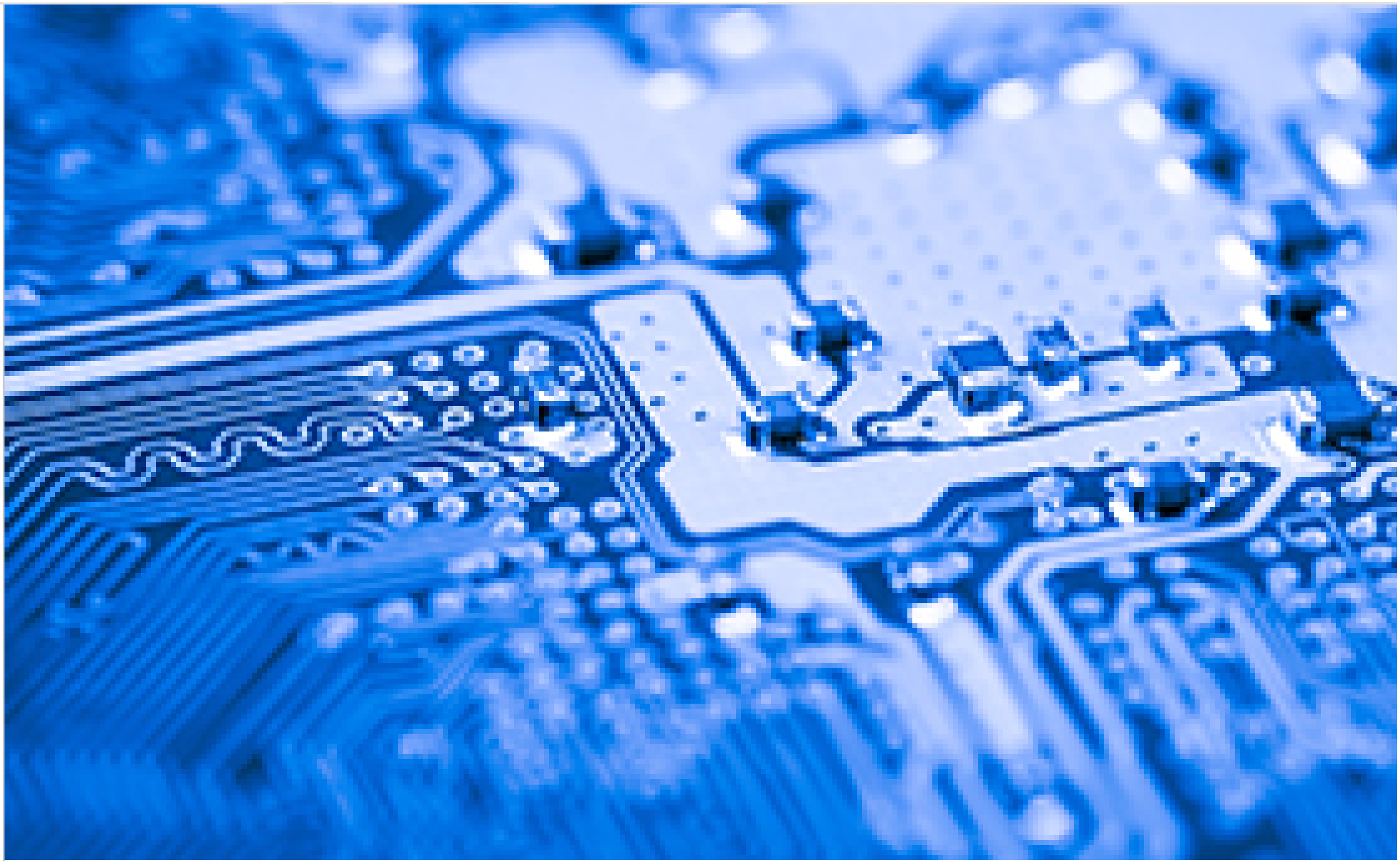


12 new Project Debater AI technologies available as cloud AI

In our recent paper “An autonomous debating system,” published in Nature, we describe Proje evaluate its performance. We also offer free access for academic use to 12 of Project Debater’s APIs, as well as trial and licensing options for developers.

→ Continue reading

AI



Introducing the AI chip leading the world in precision scaling

We've made strides in delivering the next-gen AI computational systems with cutting-edge per energy efficiency.

Continue reading