Responsible AI and ModelOps in Industry: Practical Challenges and Lessons Learned

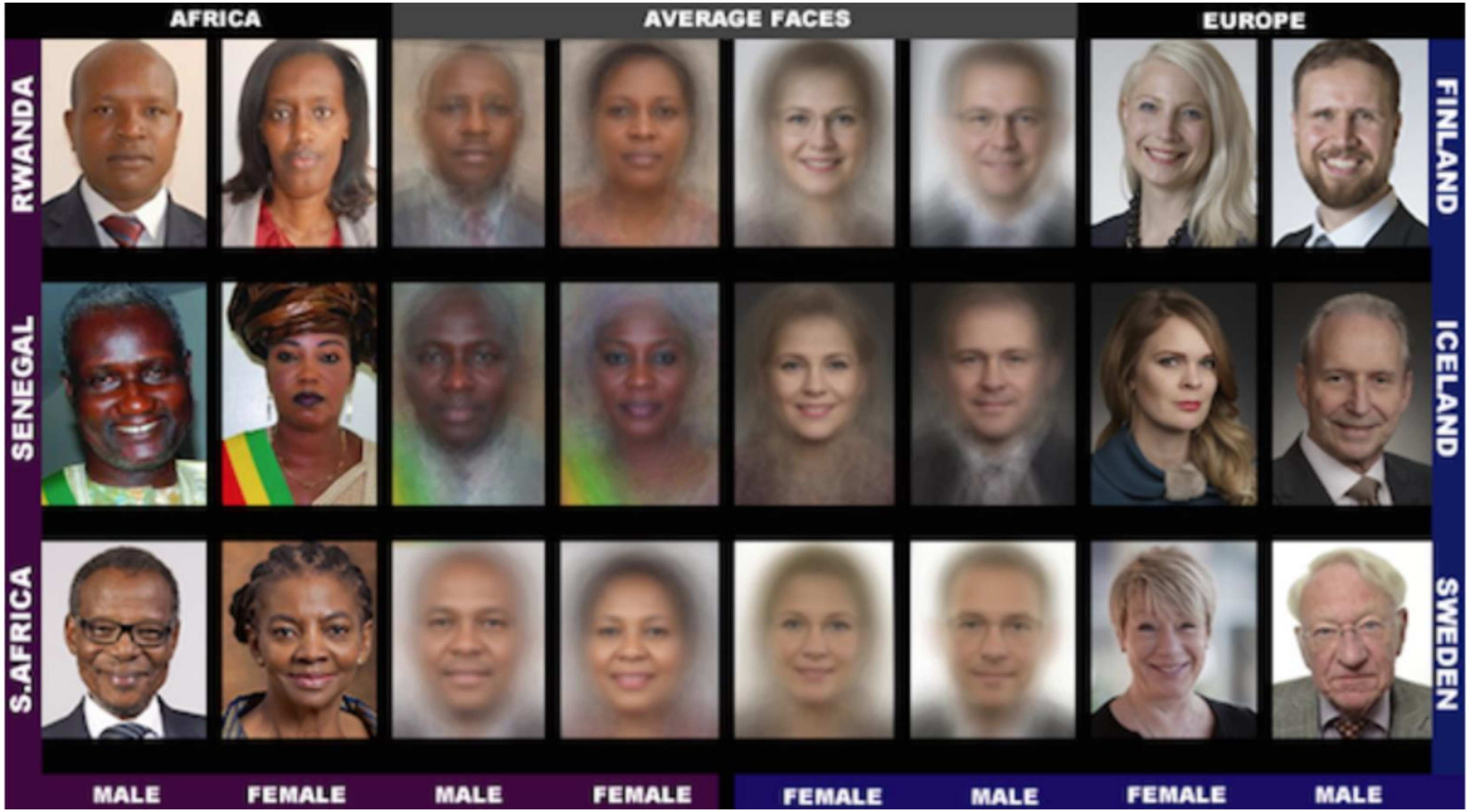
Krishnaram Kenthapadi
Chief Scientist
Fiddler AI

The Coded Gaze [Joy Buolamwini 2016]

Face detection software: Fails for some darker faces

https://youtu.be/162VzSzzoPs

Gender Shades
[Joy Buolamwini & Timnit Gebru, 2018]



- **Facial analysis software:**

  Higher accuracy for light skinned men

- **Error rates for dark skinned women: 20% - 34%**

Algorithmic Bias

When Algorithms Discriminate

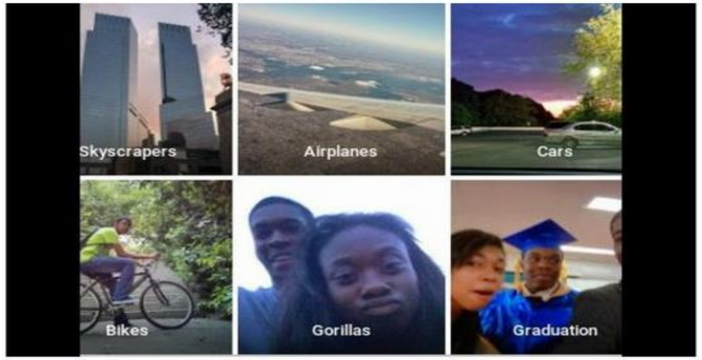
The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

Technology

Google apologises for Photos app's racist blunder

1 July 2015 | Technology



Do Google's 'unprofessional hair' results show it is racist?

Leigh Alexander

Search term brings back mainly results of black women, which some say is evidence of bias. But algorithms may just be reflecting the wider social landscape

There results of image searches for 'unprofessional hair for work' (left) and 'professional hair for work' (right) on Google. Photograph: Google

Machine Bias

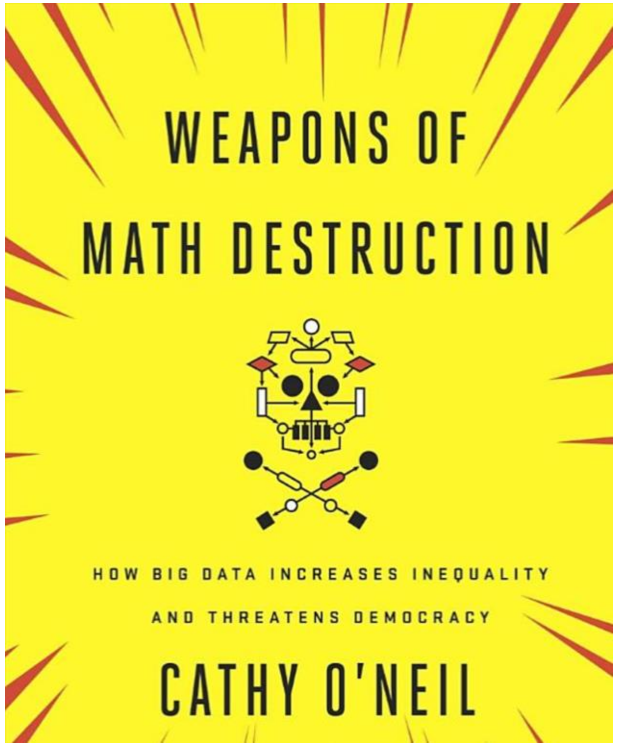
There's software used across the country to predict future criminals. And it's biased against blacks.

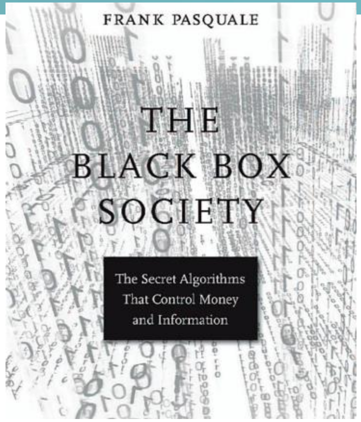
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

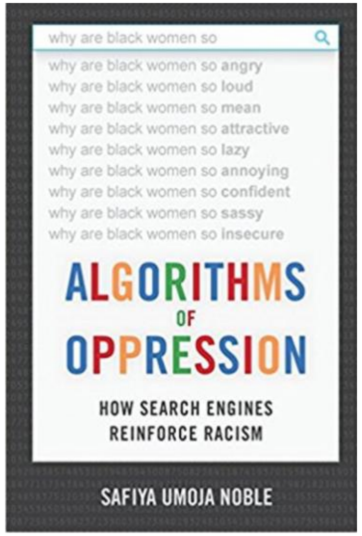
ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Algorithmic Bias

- Ethical challenges posed by AI systems
- Inherent biases present in society
- Reflected in training data
- AI/ML models prone to amplifying such biases







A History of Privacy Failures …

- Re-identification [Sweeney '00, …]
  - GIC data, health data, clinical trial data, DNA, Pharmacy data, text data, registry information, …
- Blatant non-privacy [Dinur, Nissim '03], …
- Auditors [Kenthapadi, Mishra, Nissim '05]
- AOL Debacle '06
- Genome-Wide association studies (GWAS) [Homer et al. '08]
- Netflix award [Narayanan, Shmatikov '09]
- Social networks [Backstrom, Dwork, Kleinberg '11]
- Genetic research studies [Gymrek, McGuire, Golan, Halperin, Erlich '11]
- Microtargeted advertising [Korolova 11]
- Recommendation Systems [Calandrino, Kiltzer, Naryanan, Felten, Shmatikov 11]
- Israeli CBS [Mukatren, Nissim, Salman, Tromer '14]
- Attack on statistical aggregates [Homer et al.'08] [Dwork, Smith, Steinke, Vadhan '15]

fiddler

Credit: Kobbi Nissim, Or Sheffet

Recent Privacy Attack on Large Language Models

Extracting Training Data from Large Language Models

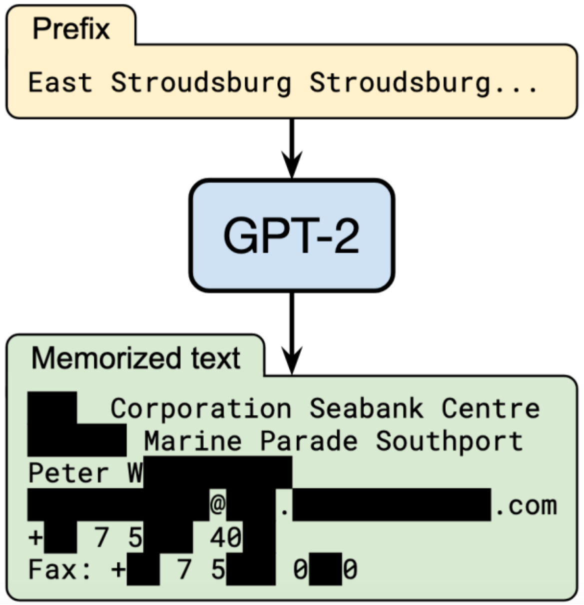
Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Abstract

It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a *training data extraction attack* to recover individual training examples by querying the language model.

We demonstrate our attack on GPT-2, a language model trained on scrapes of the public Internet, and are able to extract hundreds of verbatim text sequences from the model's training data. These extracted examples include (public) personally identifiable information (names, phone numbers, and email addresses), IRC conversations, code, and 128-bit UUIDs. Our attack is possible even though each of the above sequences are included in just *one* document in the training data.

We comprehensively evaluate our extraction attack to un-



AI Teams Lack Visibility into Their Models

- **Model Transparency**

MIT Technology Review

Facebook whistleblower Frances Haugen's testimony at the Senate today ***raised serious questions about how Facebook's algorithms work***...

- **Model Decay**

FORTUNE

NEWSLETTERS • EYE ON A.I.

This is not a drill: The coronavirus pandemic is testing A.I.'s ability to handle extreme events

- **Model Bias**

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

- **Model Compliance**

*"On Artificial Intelligence, **trust is a must**, not a nice to have.*

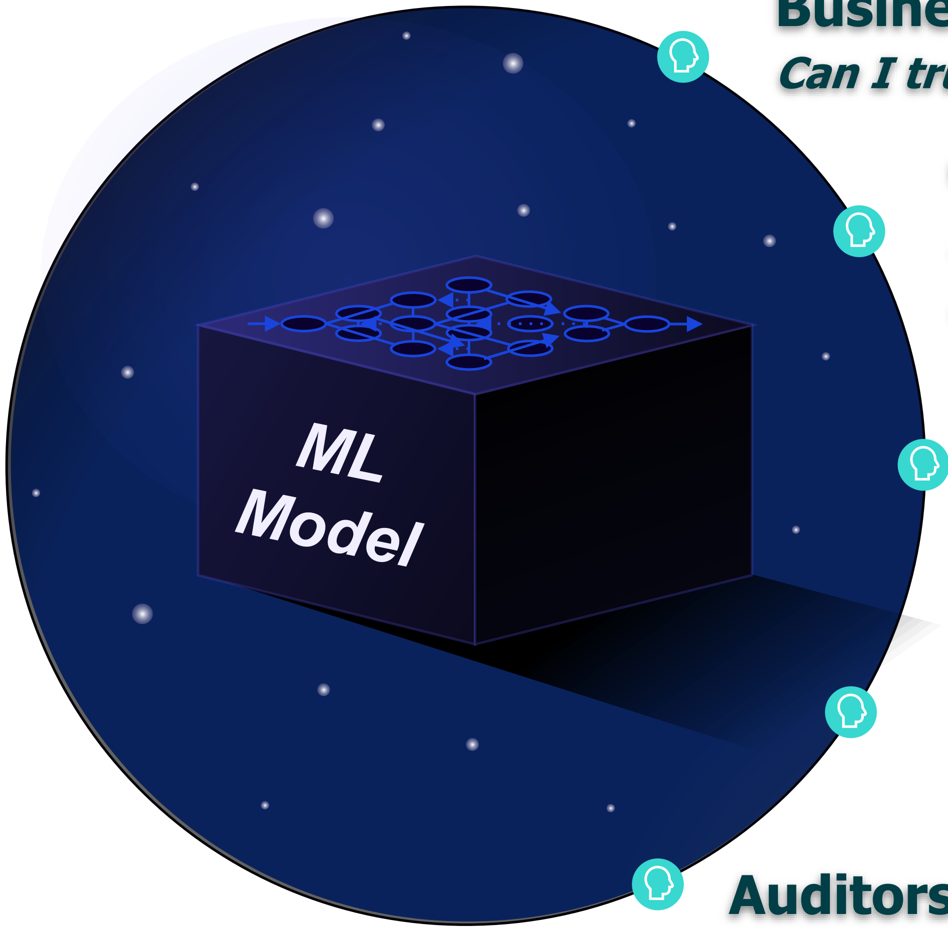
- EU Commission

Most ML Models are Opaque

No Explanations
of model behavior

No Understanding
of feature impact and fairness

No Monitoring
to catch potential bias or drift



Business User
Can I trust our AI?

Customer Support
How do I answer this complaint?

IT & Operations
How do I monitor & debug?

Data Scientists
How does this model work?

Auditors & Regulators
Are these decisions fair?

Challenges with Operationalizing AI/ML Models

"We had a **model drift** over the weekend that **cost $500,000**"

— **Chief Data Scientist**

"When something goes wrong, it takes our data scientist **2 weeks to troubleshoot the problem.**"

— **Data Science Director**

"My team spends **70%** of their time **identifying errors** and **debugging** instead of generating new models"

— **VP, Data Science**

"It takes my team **2-3 months to validate a model**"

— **Head of Model Validation**

"As we automate transportation & the lives of people are in our hands, **model explainability is a must have**"

— **CTO**

"The last thing I want to do is have to explain our AI models while **testifying in front of our parliament.**"

— **CTO**

"Our internal monitoring tools are **costing us a fortune to maintain**"

— **IT Leader**

"Monitoring and drift detection have a **direct impact on the bottom line** for us"

— **ML Platform Lead**

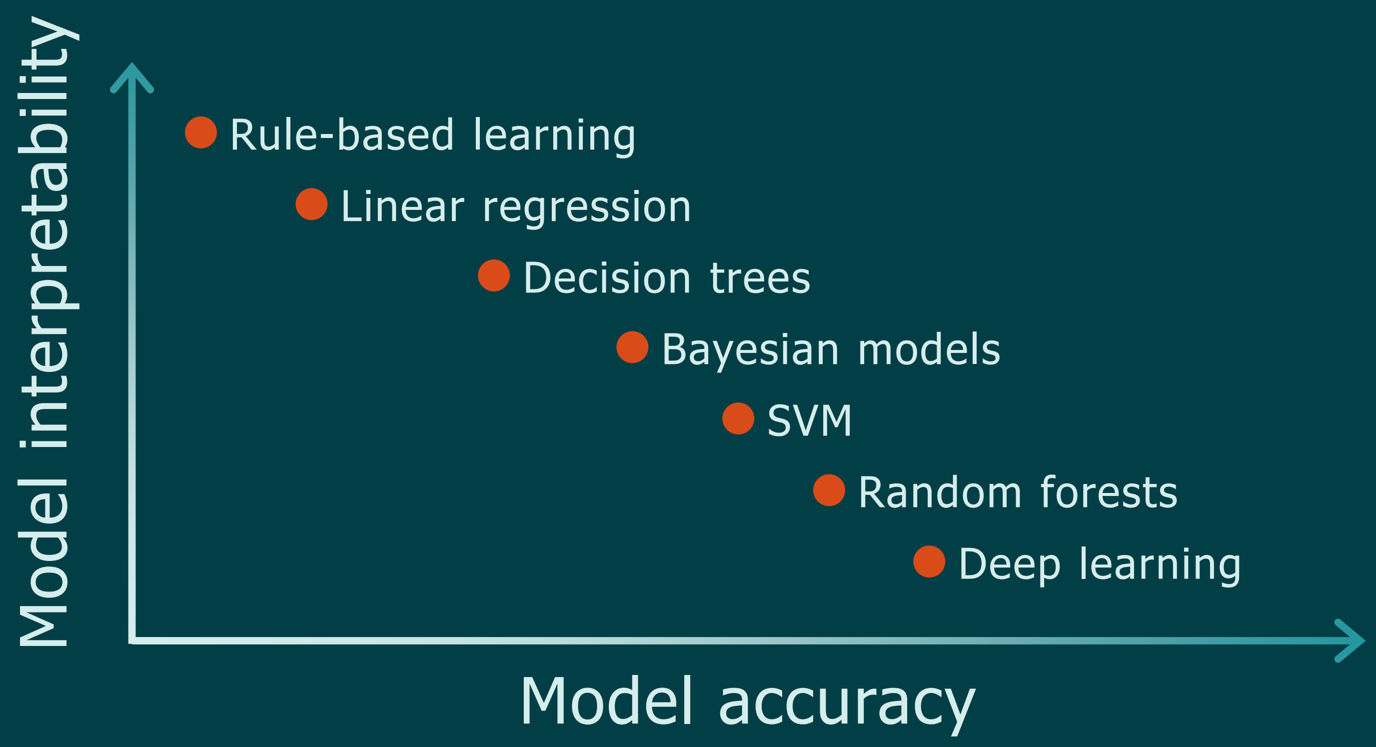
"We **don't have checks for data drift and performance** in real time"

— **Data Science Lead**

Explainable AI: Overview & Case Study

fiddler

Trade-off between model accuracy and interpretability



Explainable AI Methods

Approach 1: **Post-hoc explain a given opaque ML model**

- **Individual prediction explanations** in terms of input features, influential examples, concepts, local decision rules
- **Global prediction explanations** in terms of entire model in terms of partial dependence plots, global feature importance, global decision rules

Approach 2: **Build an interpretable ML model**

- Logistic regression, Decision trees, Decision lists and sets, Generalized Additive Models (GAMs)

Integrated Gradients for Explaining Diabetic Retinopathy Predictions

Retinal Fundus Image

Prediction: **"proliferative"** DR[1]

- Proliferative implies vision-threatening

Can we provide an explanation to the doctor with supporting evidence for **"proliferative"** DR?

[1]**Diabetic Retinopathy (DR)** is a diabetes complication that affects the eye. Deep networks can predict DR grade from retinal fundus images with high accuracy (AUC ~0.97) [JAMA, 2016].

Work done by Ankur Taly, Mukund Sundararajan, Kedar Dhamdhere, Pramod Mudrakarta (Sourced from Explainable AI in Industry WWW'20 Tutorial)

Integrated Gradients for Explaining Diabetic Retinopathy Predictions

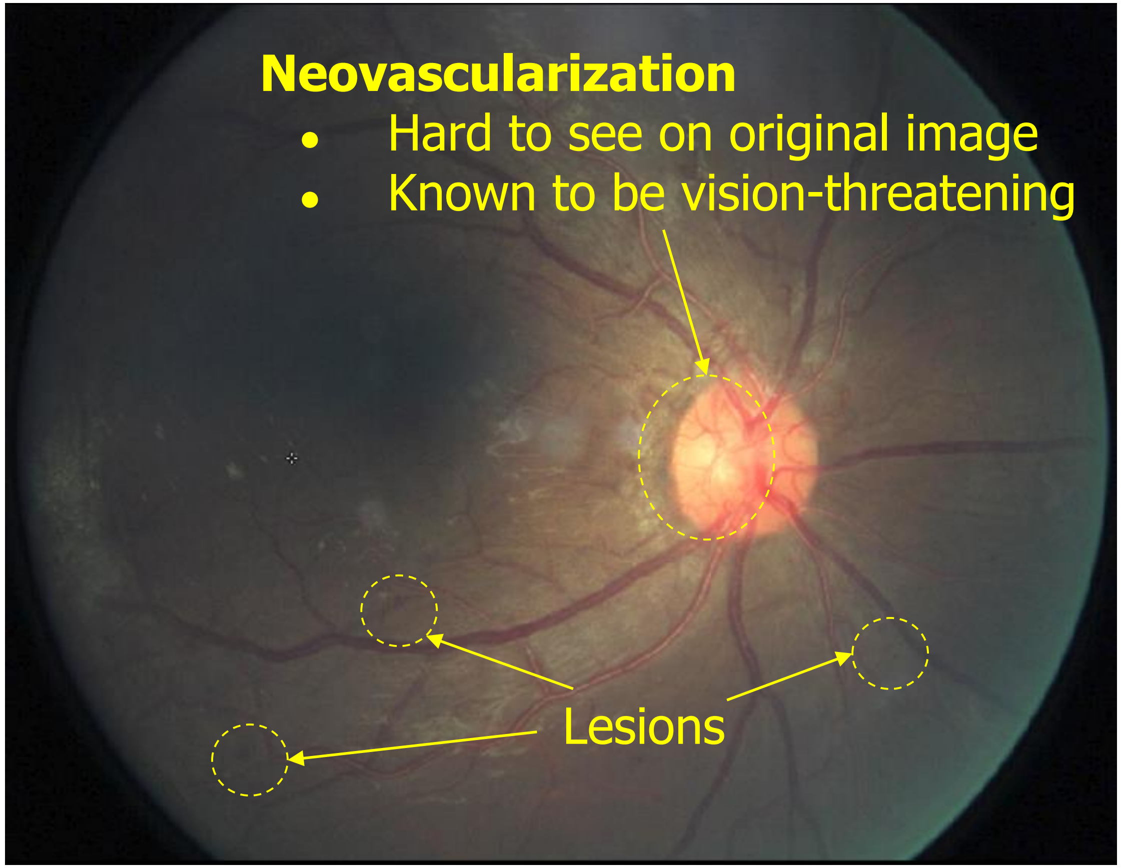
Retinal Fundus Image

Integrated Gradients for label: “proliferative”

Visualization: Overlay heatmap on green channel

Integrated Gradients for Explaining Diabetic Retinopathy Predictions

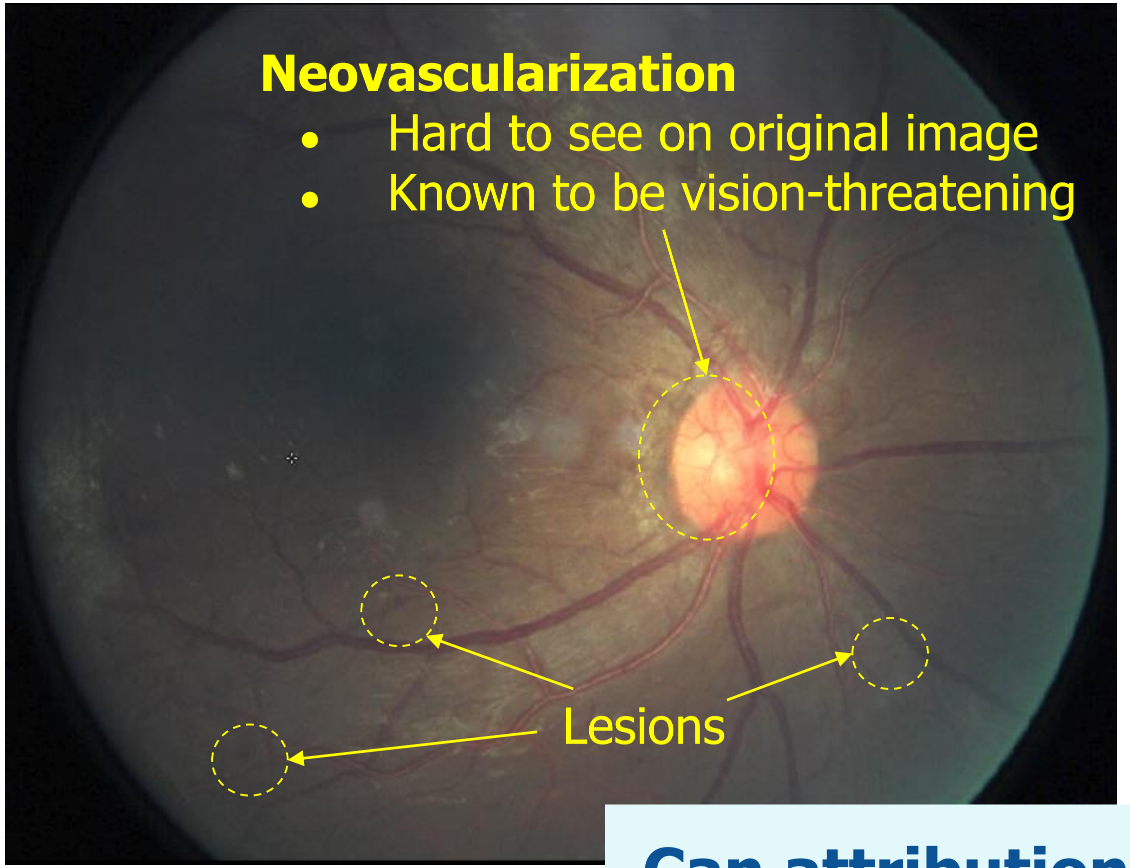
Retinal Fundus Image



Integrated Gradients for label: "proliferative"
Visualization: Overlay heatmap on green channel

Integrated Gradients for Explaining Diabetic Retinopathy Predictions

Retinal Fundus Image



Integrated Gradients for label: "proliferative"
Visualization: Overlay heatmap on green channel

Can attributions help doctors better diagnose diabetic retinopathy?

Integrated Gradients for Explaining Diabetic Retinopathy Predictions

embedded VISION summit

9 doctors graded 2000 images under three different conditions

A. Image only
B. Image + Model's prediction scores
C. Image + Model's prediction scores + Explanation (Integrated Gradients)

Findings:

- Model's predictions (B) significantly improve accuracy vs. image only (A) ($p < 0.001$)
- Both forms of assistance (B and C) improved sensitivity without hurting specificity
- Explanations (C) improved accuracy of cases with DR ($p < 0.001$) but hurt accuracy of cases without DR ($p = 0.006$)
- Both B and C increase doctor ↔ model agreement

Paper: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy --- Journal of Ophthalmology [2018]

fiddler

Model Performance Management: Overview & Case Study

fiddler

Model Performance Management (MPM)

Offline | **Online**

Validate
Ingest model, explain performance
Discover slices of low performance
Create model dashboards and reports

Train
Log training and test datasets
Check for bias and feature quality

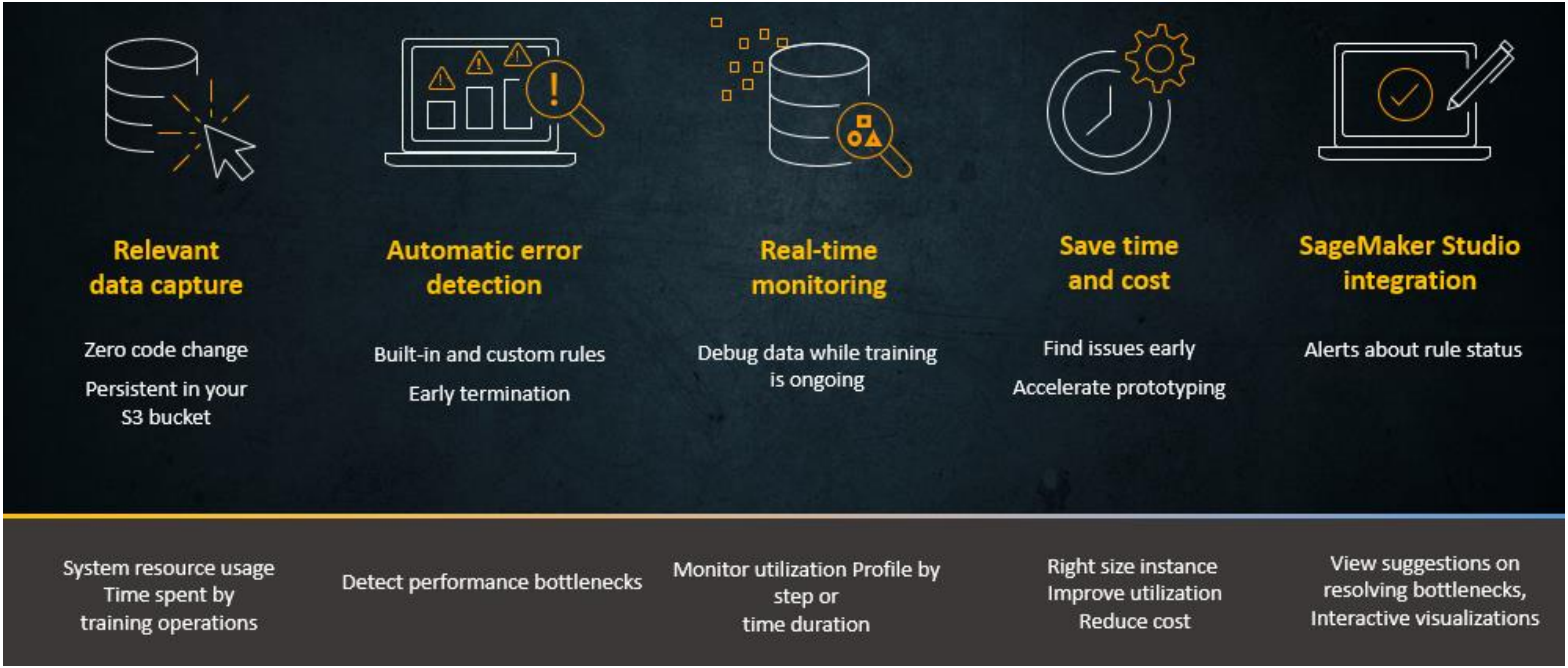
Improve
Collate insights to address data or training gaps

Deploy
Record traffic and metadata
Compare challenger and champion models on performance and bias

Monitor
Observe performance, drift, outliers, bias and errors
Alert on conditions and slices

Analyze
Root-cause performance issues
Slice & dice prediction log data

Amazon SageMaker Debugger



Debugging Model Predictions using Amazon SageMaker Debugger & Model Monitor

- Model Monitor
  - Captures inference requests & predictions
  - Raises an alarm if data drift is detected

- Debugger
  - Captures relevant tensors
  - Get visual explanations (saliency maps) for incoming requests

Source: AWS ML Blog by N. Rauschmayr, S. Bhattacharjee, and V. Kumar, July'20

Reference: Rauschmayr, et al. **Amazon SageMaker Debugger: A system for real-time insights into machine learning model training**, MLSys'21

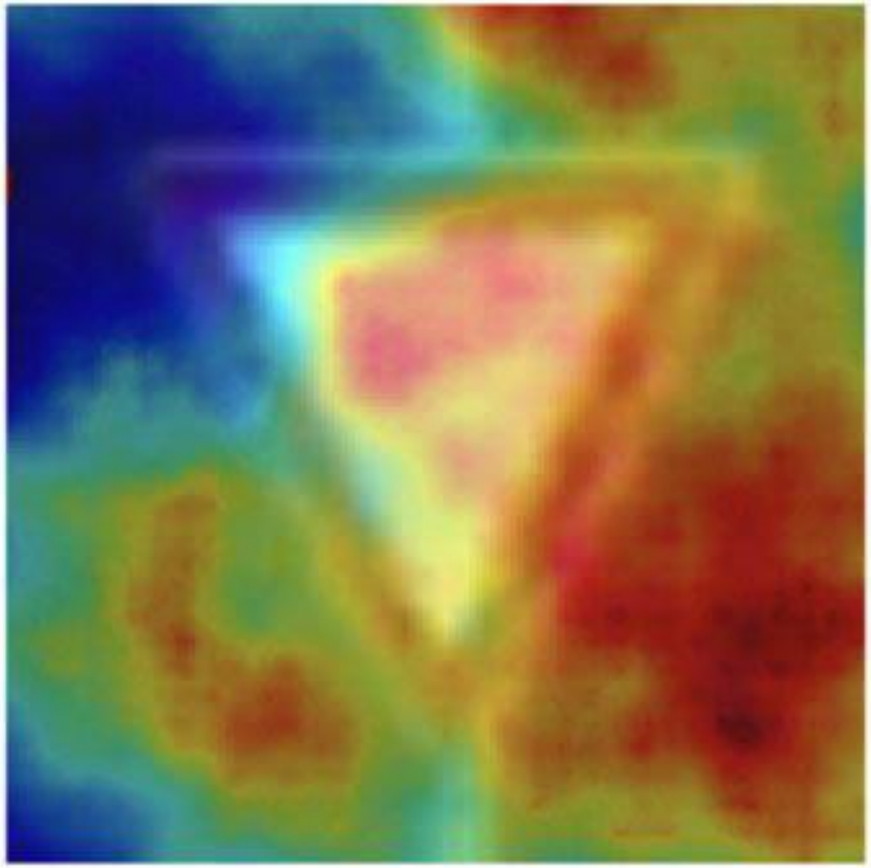
Debugging Model Predictions using Amazon SageMaker Debugger & Model Monitor

fiddler

Debugging Model Predictions using Amazon SageMaker Debugger & Model Monitor

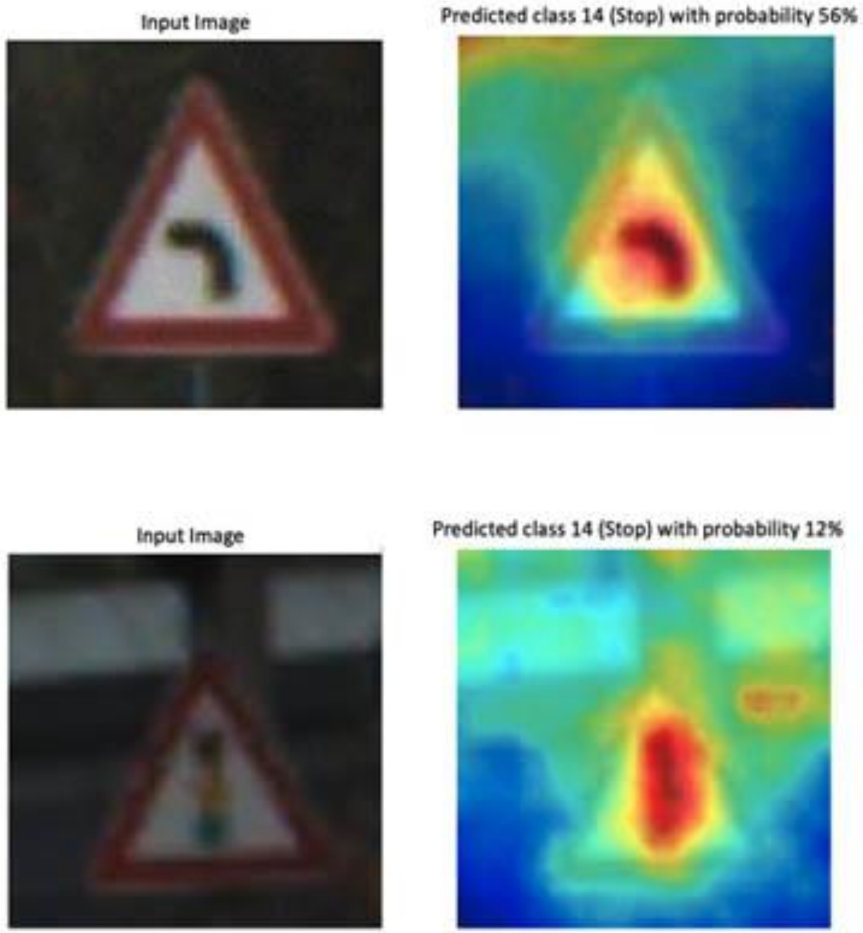
Input Image

Predicted class 14 (Stop) with probability 69%

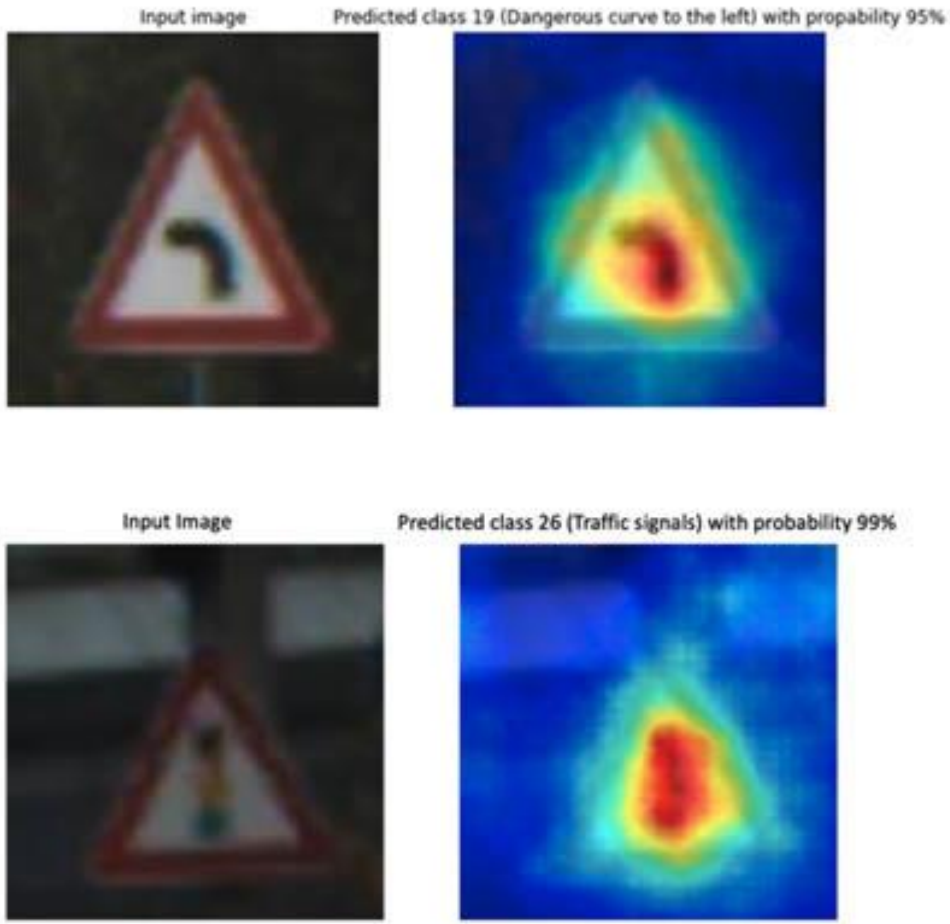


Debugging Model Predictions using Amazon SageMaker Debugger & Model Monitor

Adversarial image

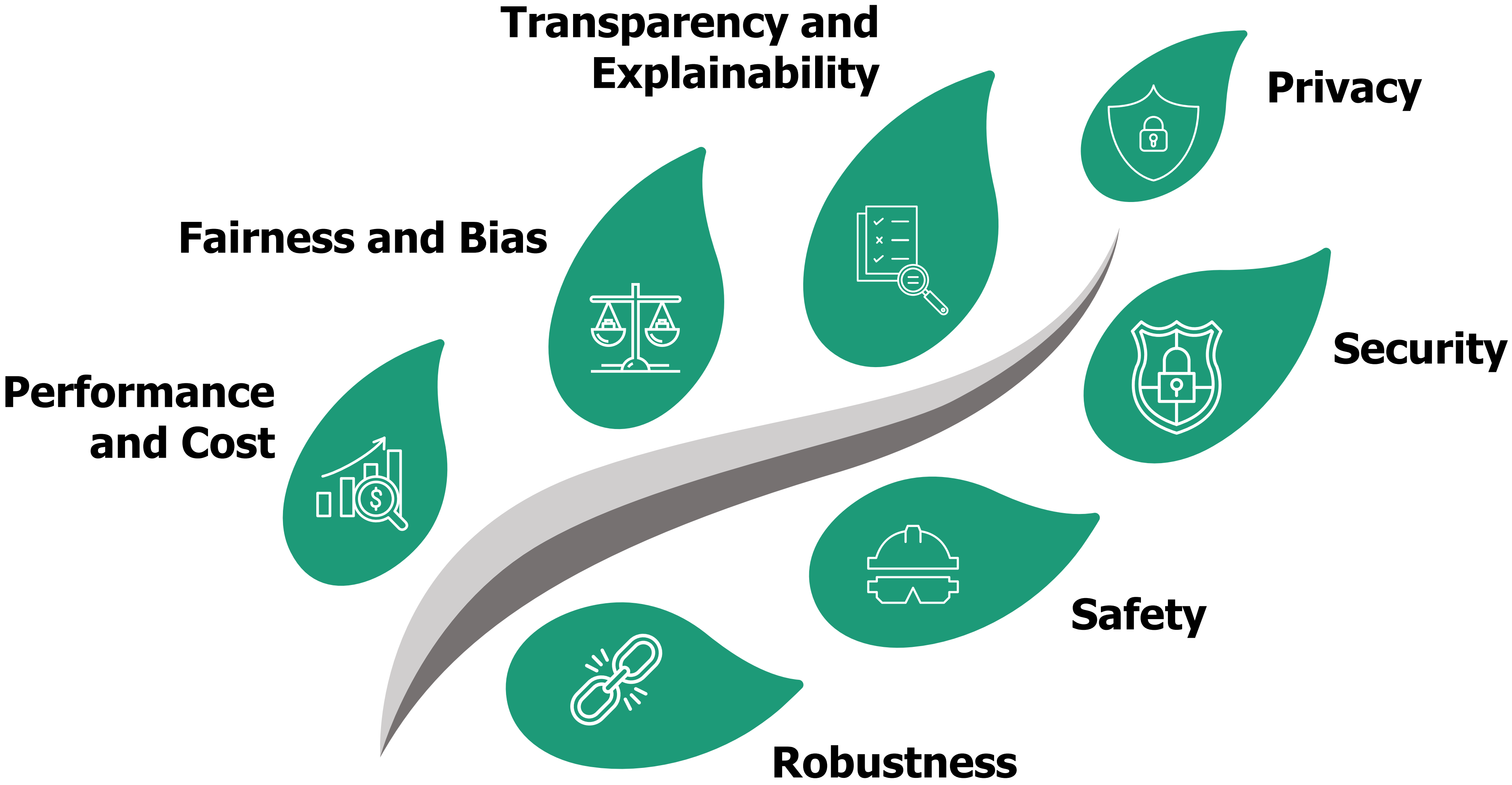


Original image



fiddler

Beyond Accuracy



Process Best Practices

- Identify product goals
- Get the right people in the room
- Identify stakeholders

- Select a fairness approach
- Analyze and evaluate your system
- Mitigate issues
- Monitor Continuously and Escalation Plans
- Auditing and Transparency

Responsible AI: Opportunities

Fairness by Design in the ML Lifecycle

Problem Formation

Is an algorithm an ethical solution to the problem?

Dataset Construction

Is the training data representative of different groups?

Are there biases in labels or features?

Does the data need to be modified to mitigate bias?

Algorithm Selection

Do fairness constraints need to be included in the objective function?

Training Process

Testing Process

Has the model been evaluated using relevant fairness metrics?

Deployment

Is the model deployed on a population for which it was not trained or evaluated?

Are there unequal effects across users?

Monitoring/ Feedback

Does the model encourage feedback loops that can produce increasingly unfair outcomes?

Explainability in ML

- Actionable explanations
- Balance between explanations & model secrecy
- Robustness of explanations to failure modes (Interaction between ML components)
- Application-specific challenges
- Tools for explanations across AI lifecycle
  - Pre & post-deployment for ML models
  - Model developer vs. End user focused

Privacy in ML

- Privacy for highly sensitive data: model training & analytics using secure enclaves, homomorphic encryption, federated learning / on-device learning, or a hybrid

- Privacy-preserving model training, robust against adversarial membership inference attacks (Dynamic settings + Complex data / model pipelines)

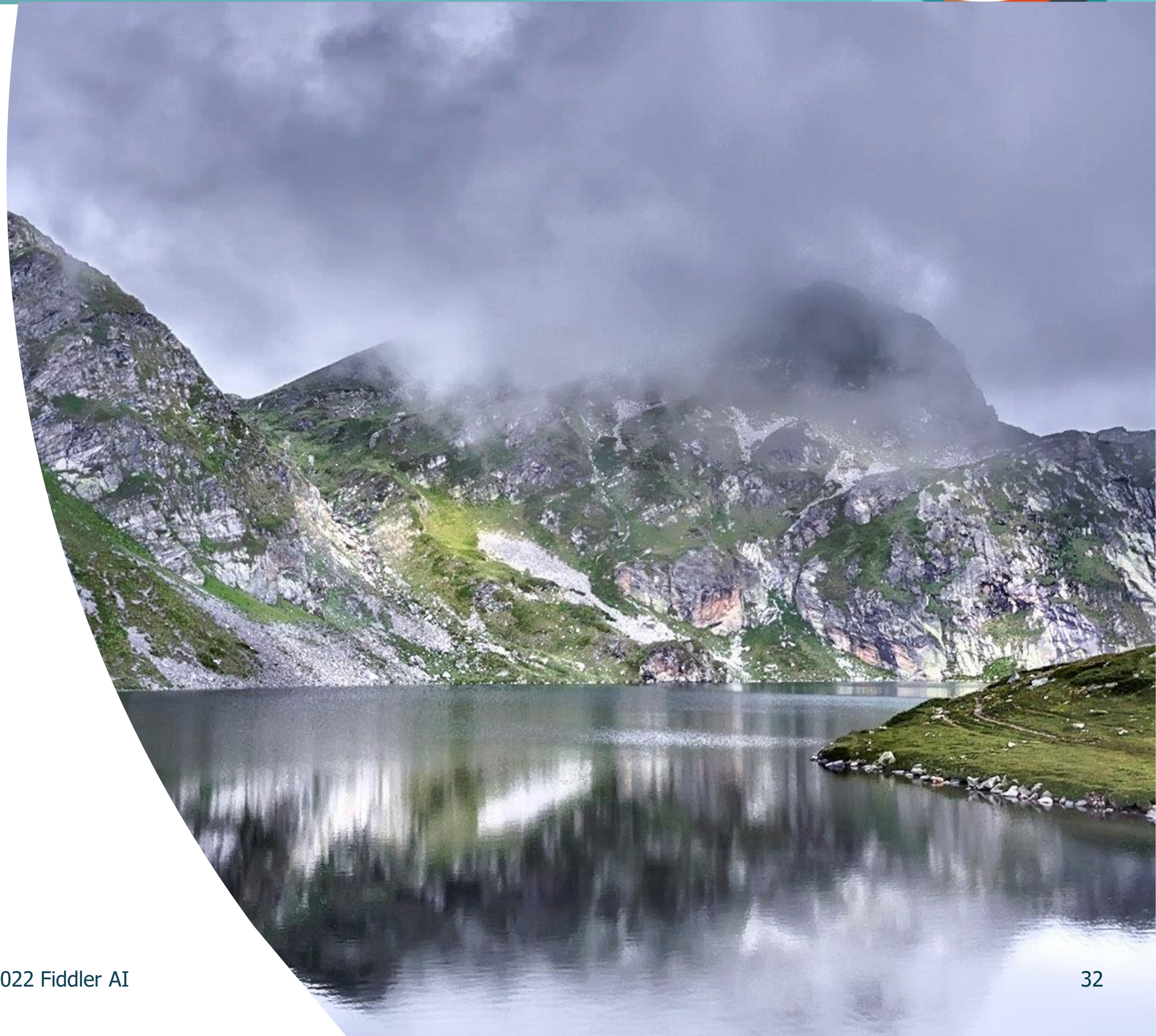
- Privacy-preserving mechanisms for data marketplaces

Reflections

"Responsible AI by Design" when building AI products

Collaboration/consensus across key stakeholders

NYT / WSJ / ProPublica test :)



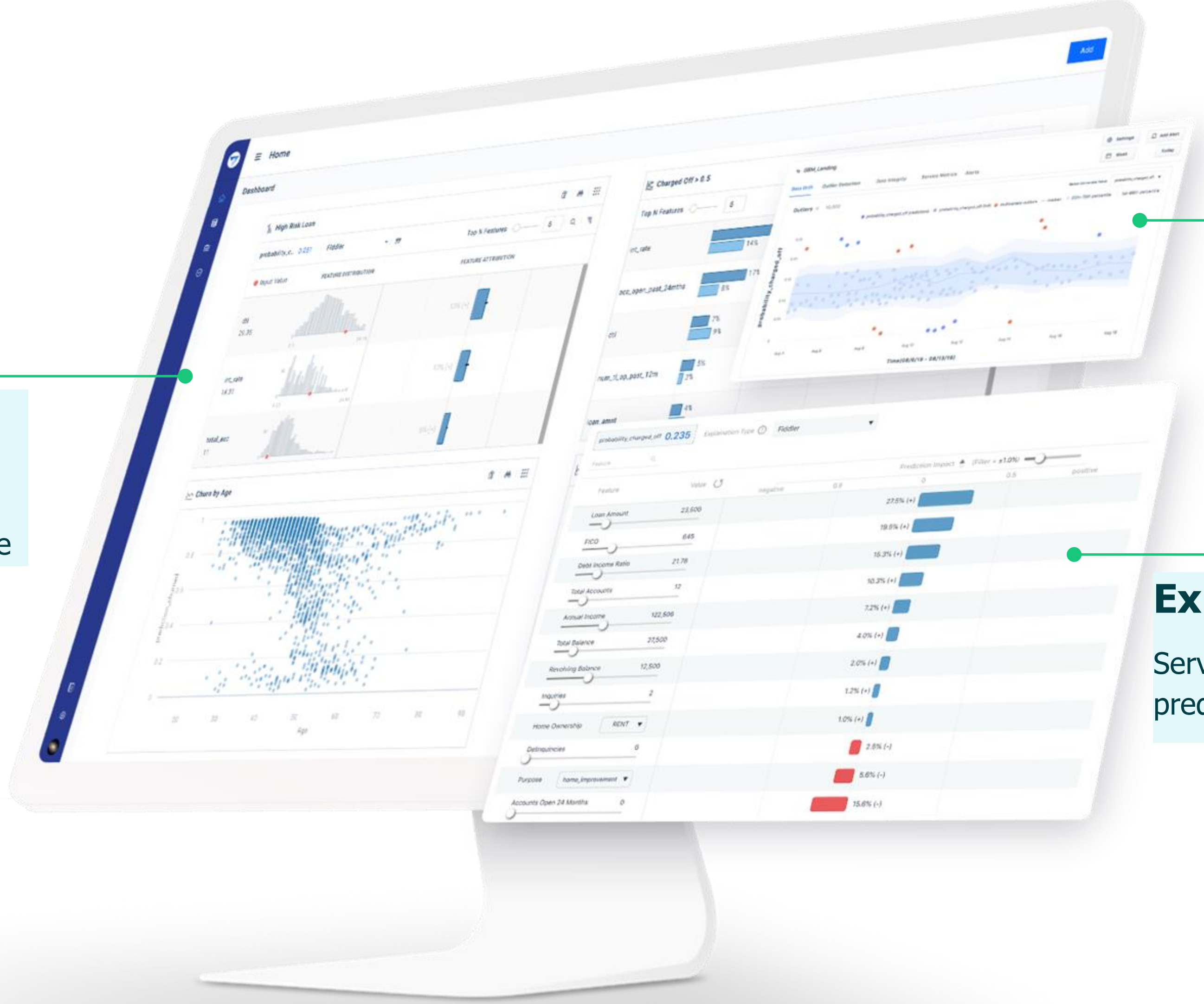
Related Tutorials / Resources

- ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)
- AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)
- Sara Hajian, Francesco Bonchi, and Carlos Castillo, Algorithmic bias: From discrimination discovery to fairness-aware data mining, KDD Tutorial, 2016.
- Solon Barocas and Moritz Hardt, Fairness in machine learning, NeurIPS Tutorial, 2017.
- Kate Crawford, The Trouble with Bias, NeurIPS Keynote, 2017.
- Arvind Narayanan, 21 fairness definitions and their politics, FAccT Tutorial, 2018.
- Sam Corbett-Davies and Sharad Goel, Defining and Designing Fair Algorithms, Tutorials at EC 2018 and ICML 2018.
- Ben Hutchinson and Margaret Mitchell, Translation Tutorial: A History of Quantitative Fairness in Testing, FAccT Tutorial, 2019.
- Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, and Jean Garcia-Gathright, Translation Tutorial: Challenges of incorporating algorithmic fairness into industry practice, FAccT Tutorial, 2019.

Related Tutorials / Resources

- Sarah Bird, Ben Hutchinson, Krishnaram Kenthapadi, Emre Kiciman, Margaret Mitchell, Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned, Tutorials at WSDM 2019, WWW 2019, KDD 2019.
- Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, Ankur Taly, Explainable AI in Industry, Tutorials at KDD 2019, FAccT 2020, WWW 2020.
- Himabindu Lakkaraju, Julius Adebayo, Sameer Singh, Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities, NeurIPS 2020 Tutorial.
- Kamalika Chaudhuri, Anand D. Sarwate, Differentially Private Machine Learning: Theory, Algorithms, and Applications, NeurIPS 2017 Tutorial.
- Krishnaram Kenthapadi, Ilya Mironov, Abhradeep Guha Thakurta, Privacy-preserving Data Mining in Industry, Tutorials at KDD 2018, WSDM 2019, WWW 2019.
- Krishnaram Kenthapadi, Ben Packer, Mehrnoosh Sameki, Nashlie Sephus, Responsible AI in Industry, Tutorials at AAAI 2021, FAccT 2021, WWW 2021, ICML 2021.

Fiddler's Model Performance Management Platform



Control Pane

Providing a cockpit view of all models and performance

- **Data /AI Teams**
- **IT Operations**
- **Business Users**
- **Legal/Regulatory**

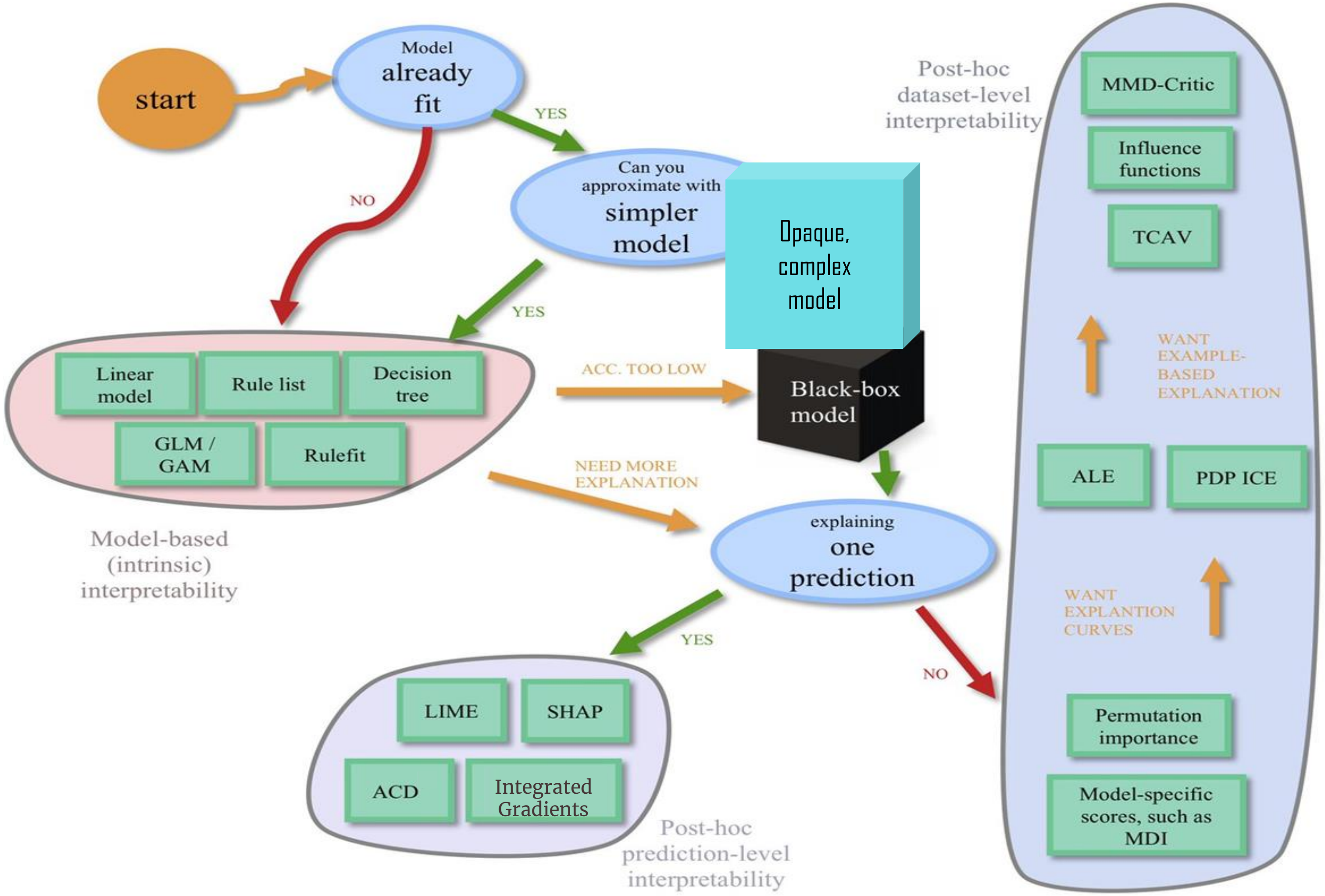
Monitoring with Alerts

Scanning for outliers and tools to investigate bias and drift

Explainability in Context

Serving up the rationale for each prediction to key stakeholders

Backup



interpretability cheat-sheet

View on github

Based on this interpretability review
and the sklearn cheat-sheet.
More in this book + these slides.

Summaries and links to code

RuleFit – automatically add features extracted from a small tree to a linear model

LIME – linearly approximate a model at a point

SHAP – find relative contributions of features to a prediction

ACD – hierarchical feature importances for a DNN prediction

Text – DNN generates text to explain a DNN's prediction (sometimes not faithful)

Permutation importance – permute a feature and see how it affects the model

ALE – perturb feature value of nearby points and see how outputs change

PDP ICE – vary feature value of all points and see how outputs change

TCAV – see if representations of certain points learned by DNNs are linearly separable

Influence functions – find points which highly influence a learned model

MMD-CRITIC – find a few points which summarize classes

Credit: Slide based on https://twitter.com/chandan_singh96/status/1138811752769101825

Model Performance Management is a Framework for Operationalizing AI/ML

Monitoring + Explainability + Fairness = MPM **Model Performance Mgmt**

Similar

MPM is to MLOps as **APM is to DevOps**

Application Performance Mgmt.

Amazon SageMaker Debugger

Relevant data capture

Zero code change

Persistent in your S3 bucket

Automatic error detection

Built-in and custom rules

Early termination

Real-time monitoring

Debug data while training is ongoing

Save time and cost

Find issues early

Accelerate prototyping

SageMaker Studio integration

Alerts about rule status

System resource usage Time spent by training operations	Detect performance bottlenecks	Monitor utilization Profile by step or time duration	Right size instance Improve utilization Reduce cost	View suggestions on resolving bottlenecks, Interactive visualizations

Fairness in ML

- Application specific challenges
- Tools for ensuring fairness (measuring & mitigating bias) in AI lifecycle
  - Pre-processing (representative datasets; modifying features/labels)
  - ML model training with fairness constraints
  - Post-processing
  - Experimentation & Post-deployment