Sayan Dey

Data Scientist | BI Consultant | Machine Learning Engineer | AI | Deep Learning | Speaker | Corporate Trainer

Sayan is a Data Science and Analytics Professional with around a decade's worth of rich experience across the analytics technology stack.

He has worked across a multitude of roles spanning corporate trainer, individual contributor, developer, consultant, project manager, scrum master and client engagement manager.

His USP is having a unique blend of extensive production experience on cutting edge AI problems and excellent training experience through his association with several of the world’s top training vendors both in the online and offline formats.

He has successfully trained tens of thousands of IT professionals spanning 10000+ hours across experience levels ranging from 0 to 30+ years including directors and founders.

He continues to be actively involved in global technology events/seminars as guest speaker and a passionate evangelist of advanced technologies in the data science technology stack.

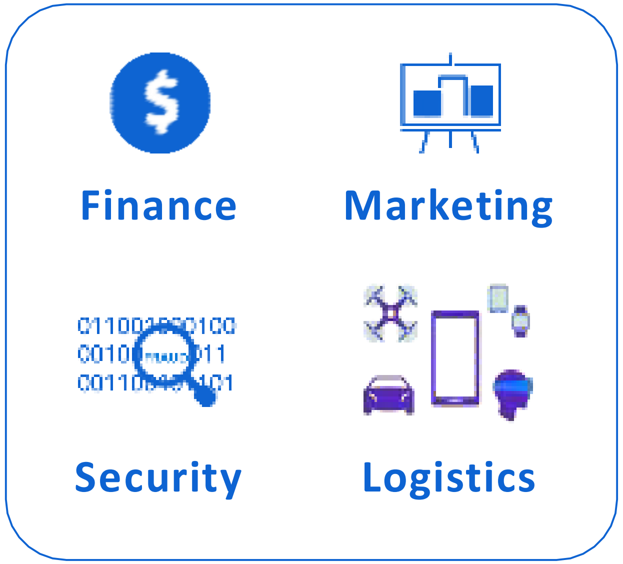
He is a regular visiting faculty for some of the best institutes like Great Lakes and IIIT Bangalore catering to the AI/ML technology stack to name a few.

He is extremely passionate in this domain and besides consulting with organisations like Walmart and Intuit, he is also actively working with a few startups on high end projects in Computer Vision and Natural Language Processing.

Need for Explainable AI

Current AI Systems

- Machine Learning centric today.
- ML Models are **opaque, non-intuitive** and **difficult to understand**

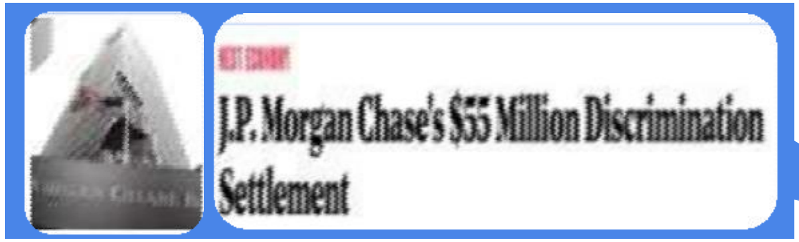


User

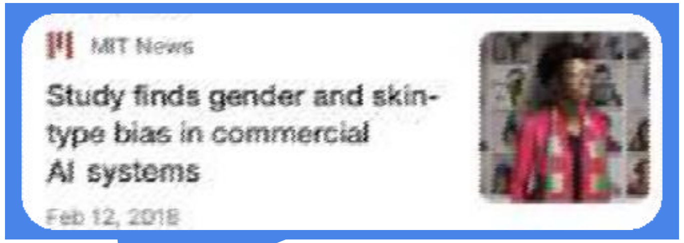
- Why did you do that?
- Why not something else?
- When do you succeed or fail?
- How do I correct an error?
- When do I trust you?

Explainable AI and ML *is essential for future customers to understand, trust, and effectively manage the emerging generation of AI applications*

Black-box AI creates business risk for Industry



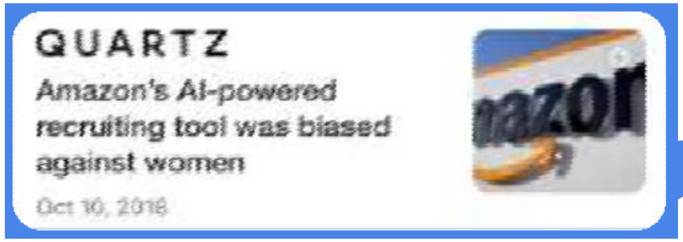
J.P. Morgan Chase's $55 Million Discrimination Settlement



MIT News

Study finds gender and skin-type bias in commercial AI systems

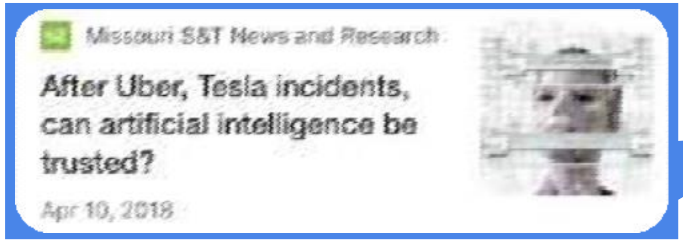
Feb 12, 2018



QUARTZ

Amazon's AI-powered recruiting tool was biased against women

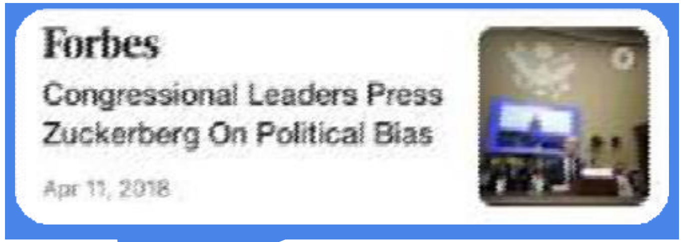
Oct 10, 2018



Missouri S&T News and Research

After Uber, Tesla incidents, can artificial intelligence be trusted?

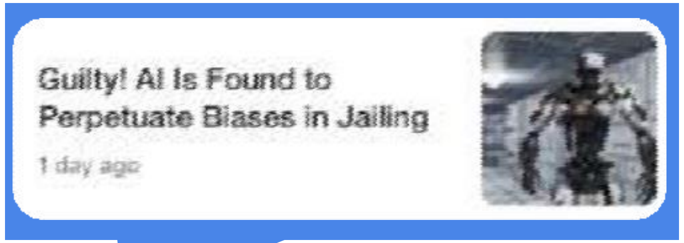
Apr 10, 2018



Forbes

Congressional Leaders Press Zuckerberg On Political Bias

Apr 11, 2018

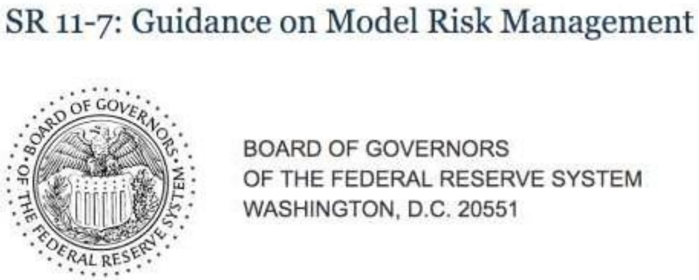


Guilty! AI Is Found to Perpetuate Biases in Jailing

1 day ago

Impact of Bias in ML Systems

Fairness


SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551

Privacy

Transparency

Explainability

GDPR Concerns Around Lack of Explainability in AI

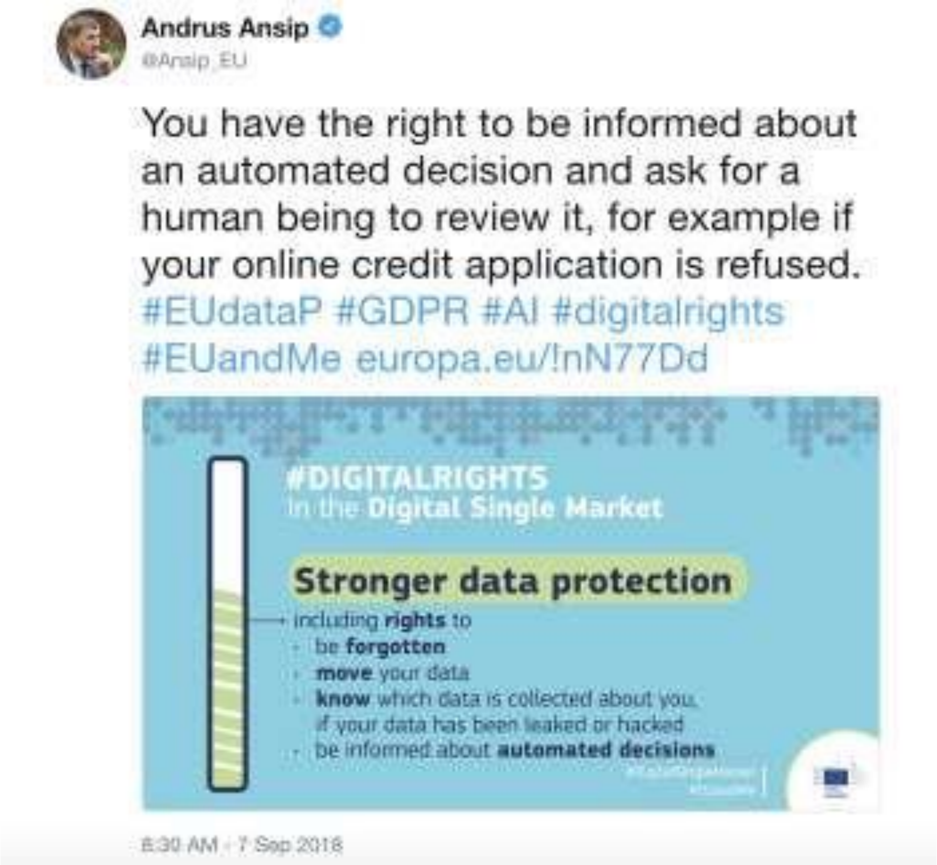
"

*Companies should commit to ensuring systems that could fall under GDPR, including AI, will be compliant. The threat of **sizeable fines of €20 million or 4% of global turnover** provides a sharp incentive.*

*Article 22 of GDPR empowers individuals with the **right to demand an explanation of how an AI system made a decision that affects them.***

"

- European Commision



Andrus Ansip
@Ansip_EU

You have the right to be informed about an automated decision and ask for a human being to review it, for example if your online credit application is refused. #EUdataP #GDPR #AI #digitalrights #EUandMe europa.eu/!nN77Dd

#DIGITALRIGHTS
in the **Digital Single Market**

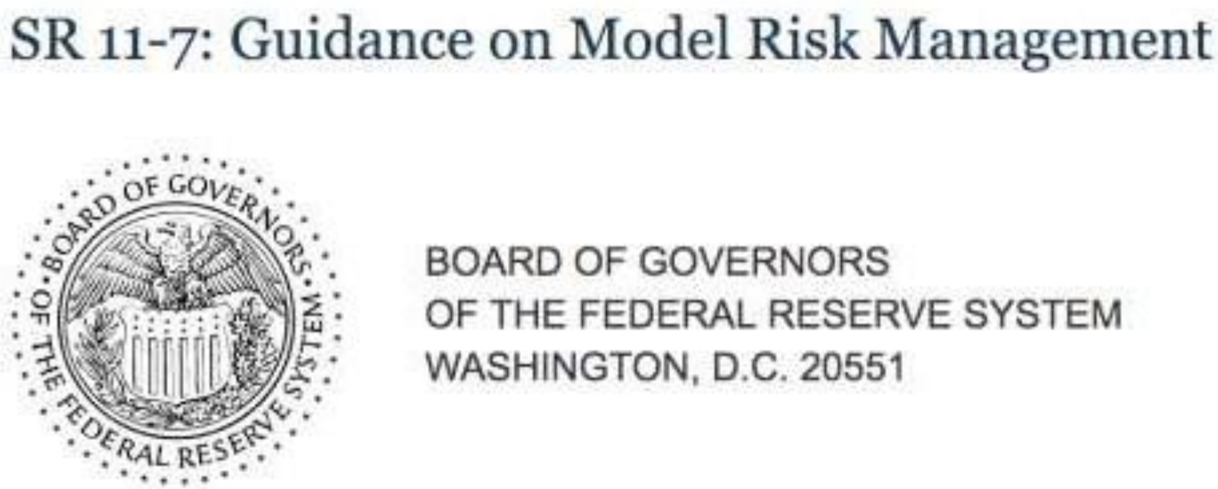
Stronger data protection

including **rights** to
- be **forgotten**
- **move** your data
- **know** which data is collected about you, if your data has been leaked or hacked
- be informed about **automated decisions**

8:30 AM - 7 Sep 2018

VP, European Commision

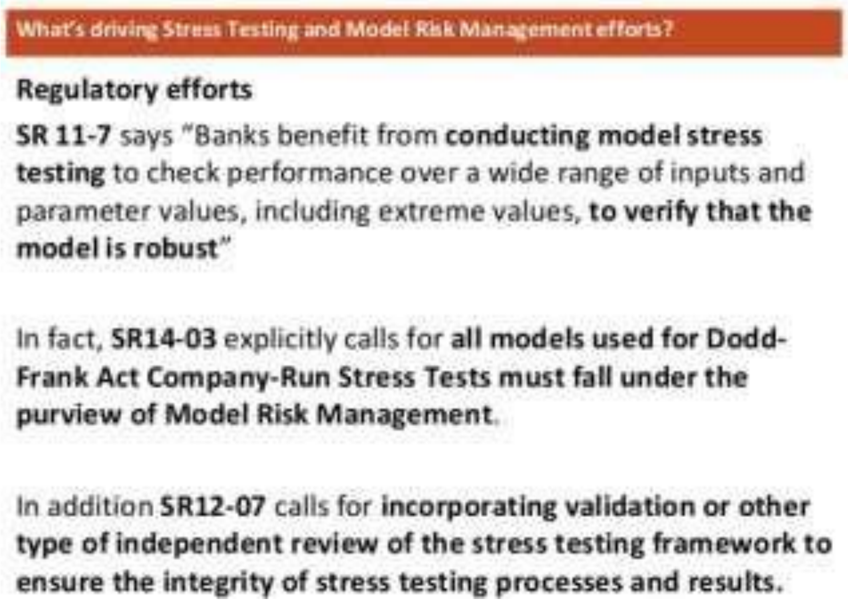
SR 11-7 and OCC regulations for Financial Institutions



SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551



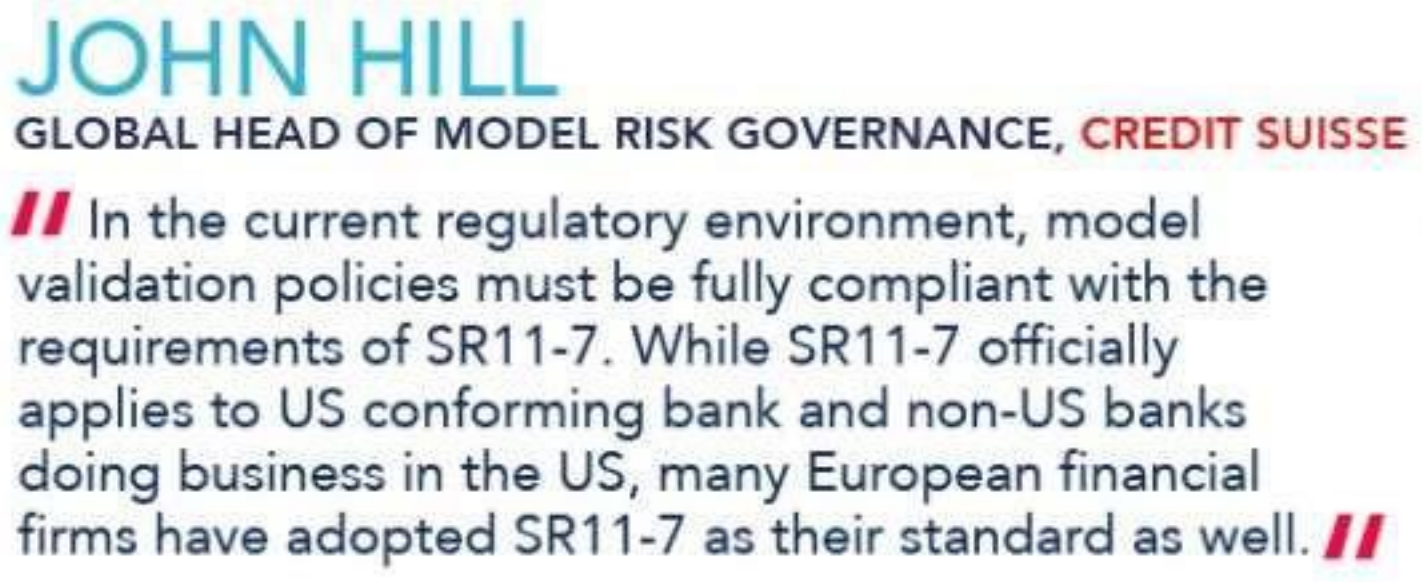
What's driving Stress Testing and Model Risk Management efforts?

Regulatory efforts

SR 11-7 says "Banks benefit from **conducting model stress testing** to check performance over a wide range of inputs and parameter values, including extreme values, **to verify that the model is robust**"

In fact, **SR14-03** explicitly calls for **all models used for Dodd-Frank Act Company-Run Stress Tests must fall under the purview of Model Risk Management**.

In addition **SR12-07** calls for **incorporating validation or other type of independent review of the stress testing framework to ensure the integrity of stress testing processes and results.**

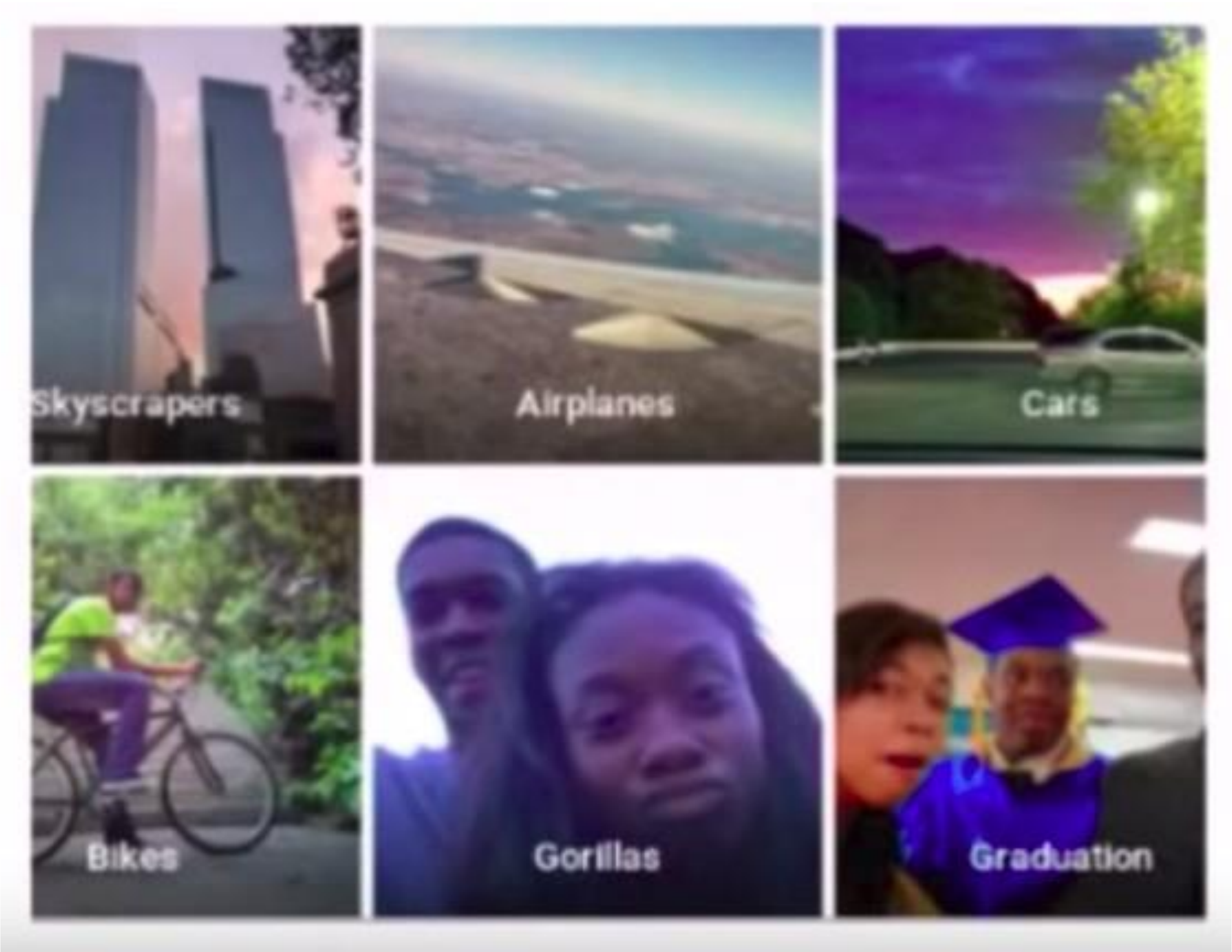


JOHN HILL

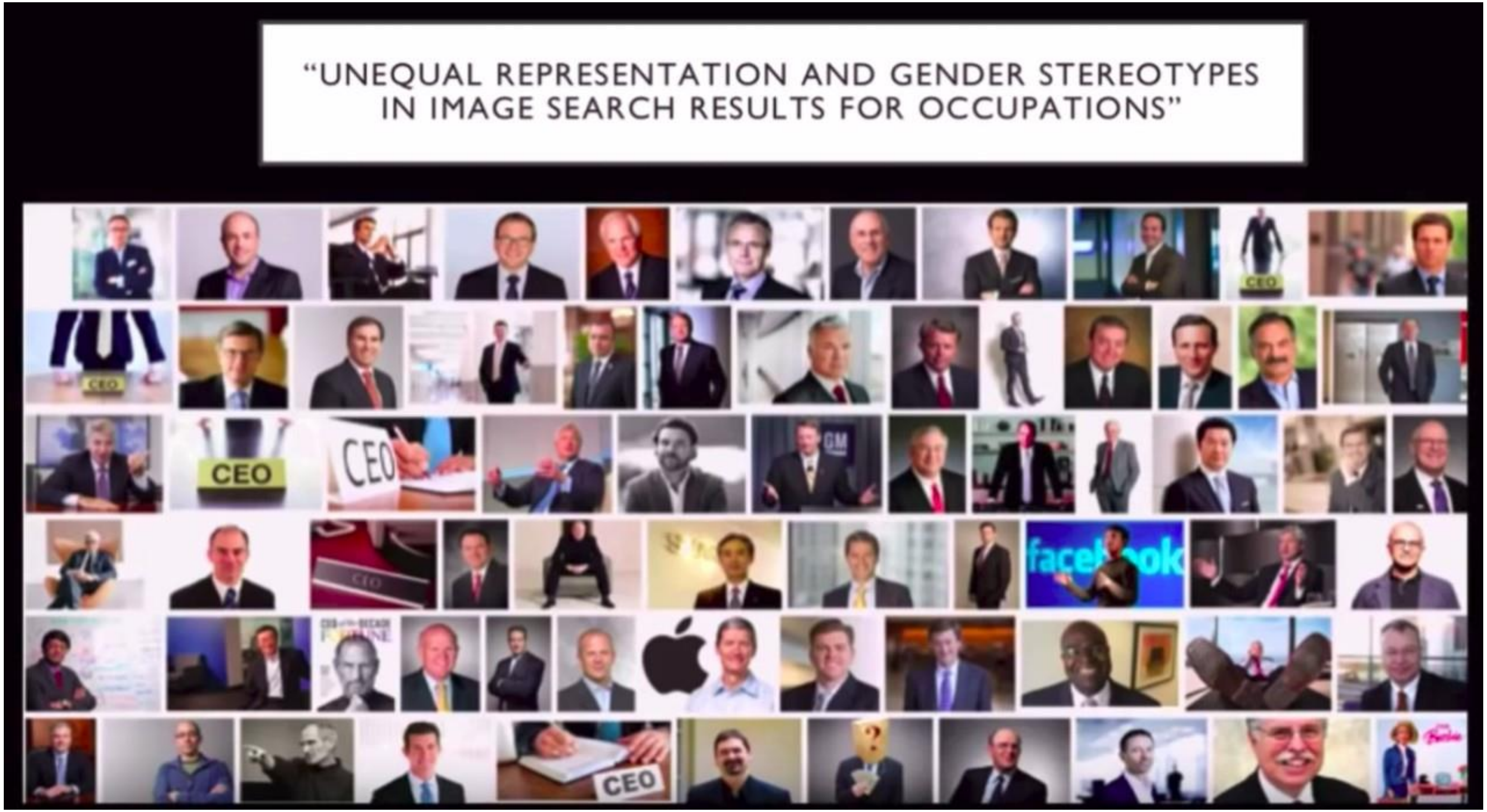
GLOBAL HEAD OF MODEL RISK GOVERNANCE, CREDIT SUISSE

" In the current regulatory environment, model validation policies must be fully compliant with the requirements of SR11-7. While SR11-7 officially applies to US conforming bank and non-US banks doing business in the US, many European financial firms have adopted SR11-7 as their standard as well. "

Impact of Bias in ML Systems

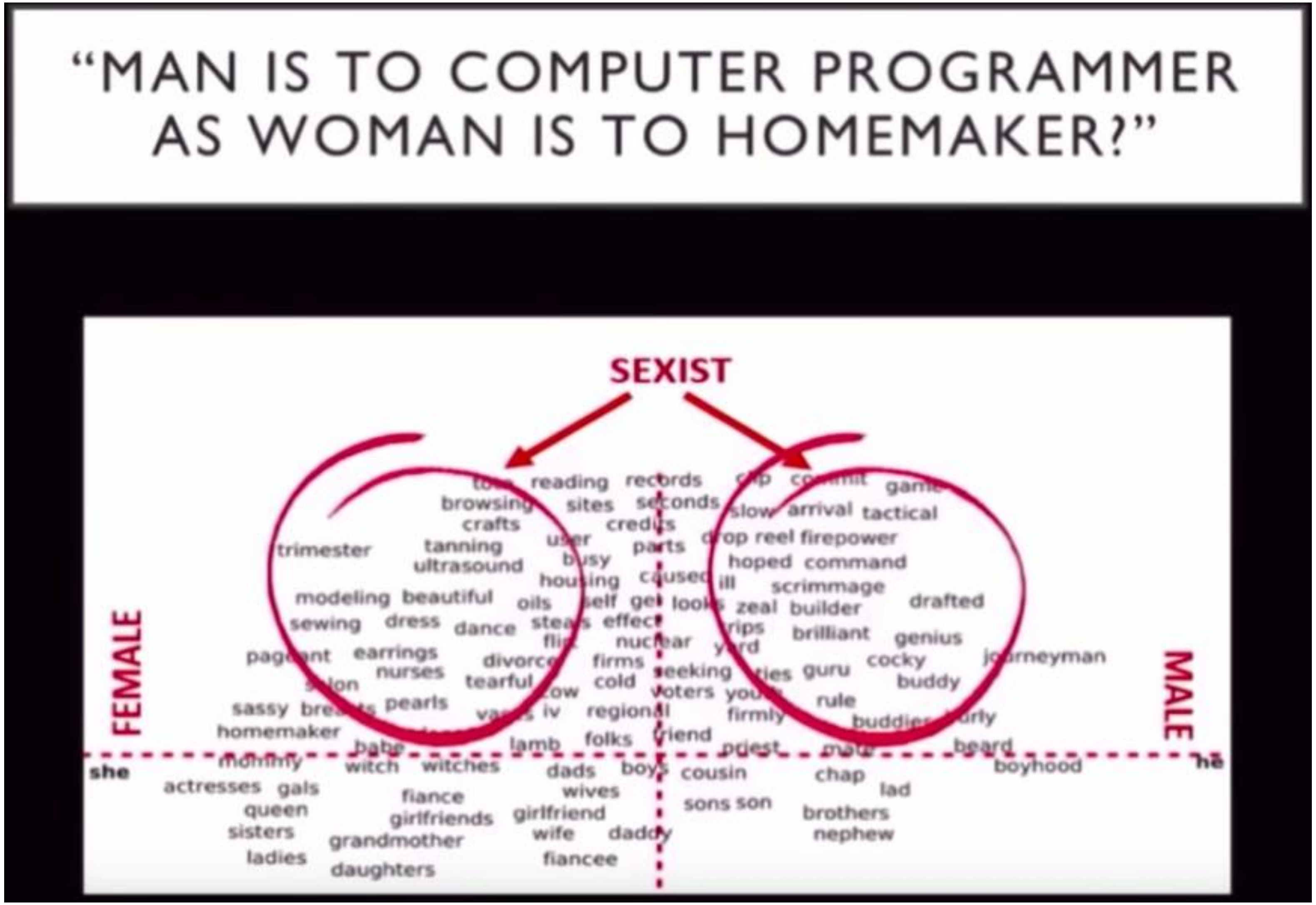


Impact of Bias in ML Systems



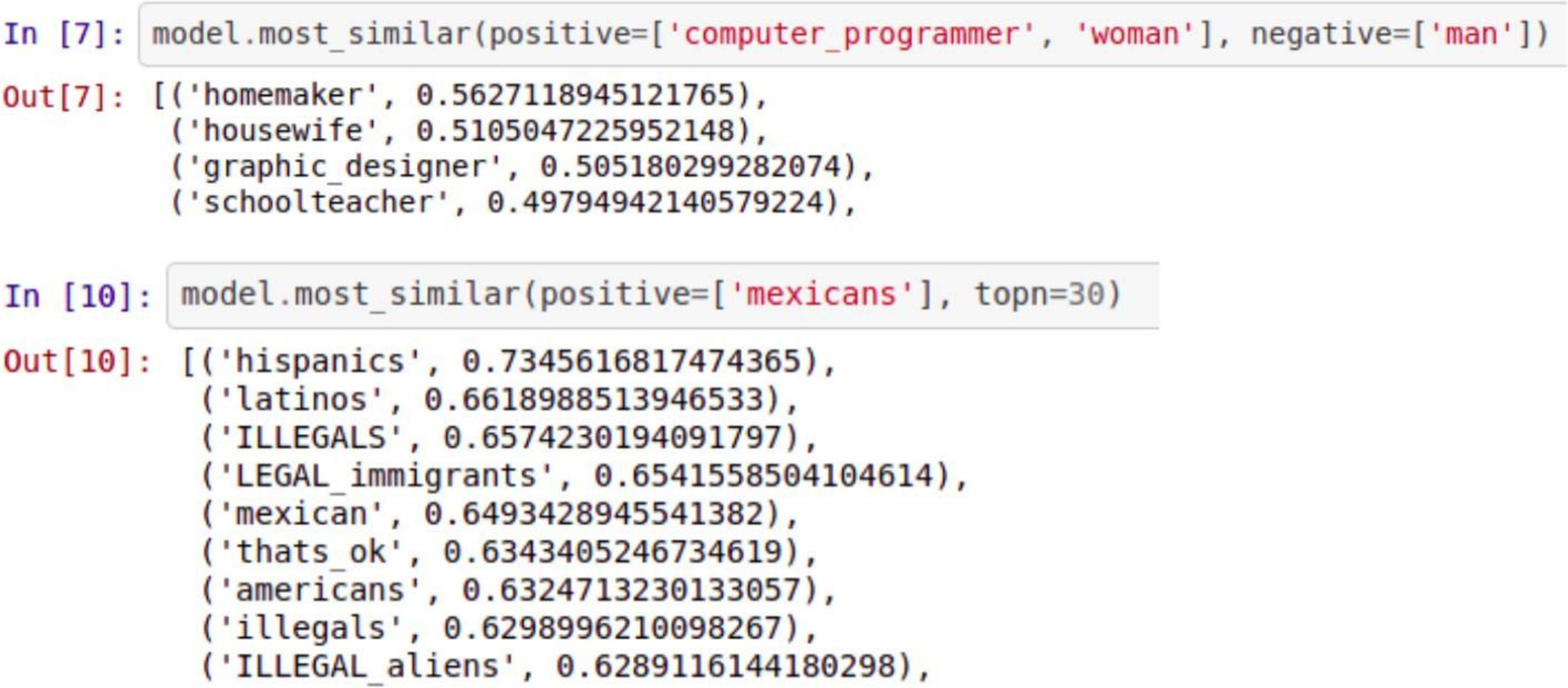
"UNEQUAL REPRESENTATION AND GENDER STEREOTYPES IN IMAGE SEARCH RESULTS FOR OCCUPATIONS"

Impact of Bias in ML Systems



"MAN IS TO COMPUTER PROGRAMMER AS WOMAN IS TO HOMEMAKER?"

Impact of Bias in ML Systems



```
In [7]:  model.most_similar(positive=['computer_programmer', 'woman'], negative=['man'])

Out[7]:  [('homemaker', 0.5627118945121765),
          ('housewife', 0.510504722595148),
          ('graphic_designer', 0.505180299282074),
          ('schoolteacher', 0.49794942140579224),
```

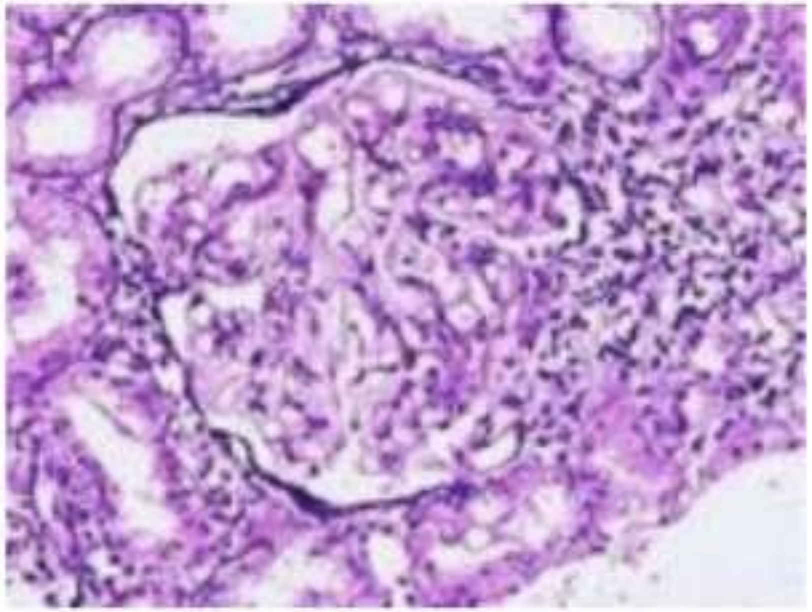
```
In [10]:  model.most_similar(positive=['mexicans'], topn=30)

Out[10]:  [('hispanics', 0.7345616817474365),
           ('latinos', 0.6618988513946533),
           ('ILLEGALS', 0.6574230194091797),
           ('LEGAL_immigrants', 0.6541558504104614),
           ('mexican', 0.6493428945541382),
           ('thats_ok', 0.634340524673619),
           ('americans', 0.6324713230133057),
           ('illegals', 0.6298996210098267),
           ('ILLEGAL_aliens', 0.6289116144180298),
```

Why Explainability: Verify the ML Model / System

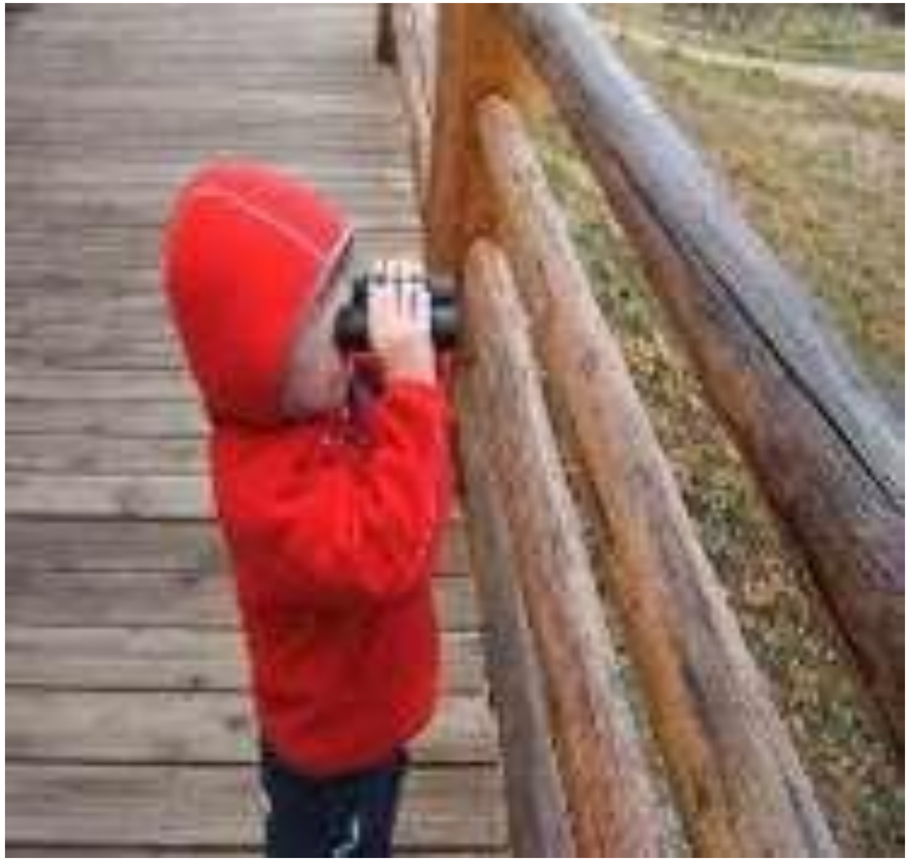
Wrong decisions can be costly and dangerous

“Autonomous car crashes, because it wrongly recognizes …”

“AI medical diagnosis system misclassifies patient’s disease …”

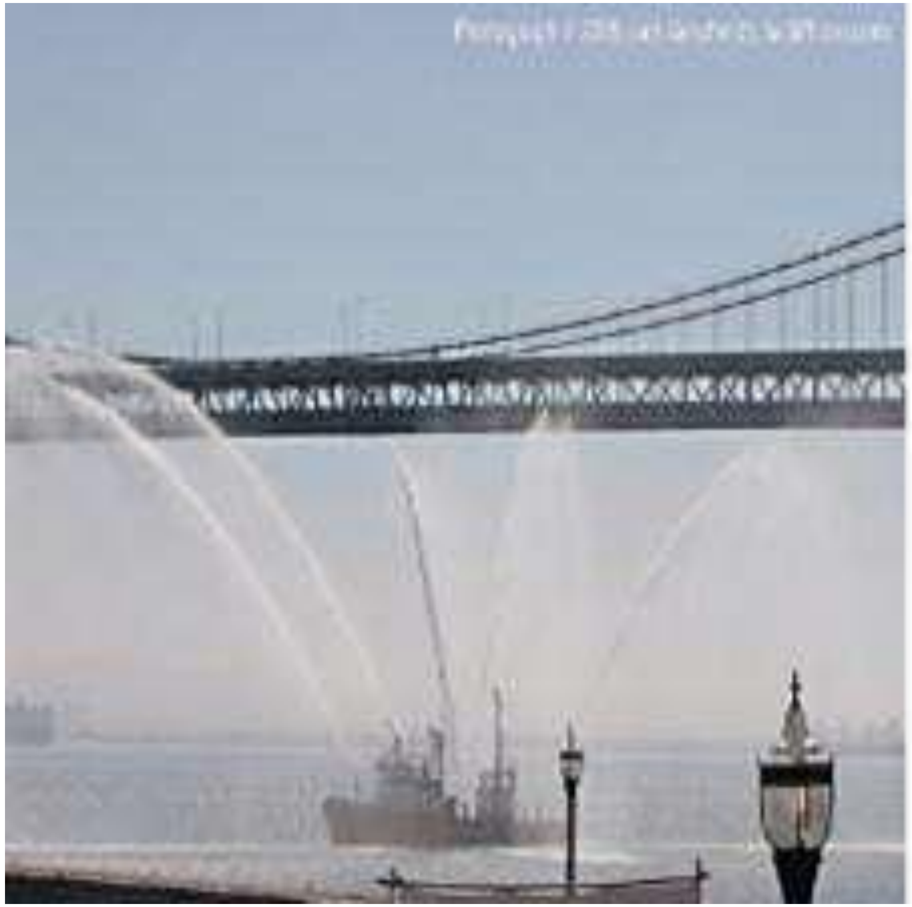


Why Explainability: Debug (Mis-)Predictions



Top label: **“clog”**

Why did the network label this image as **“clog”**?

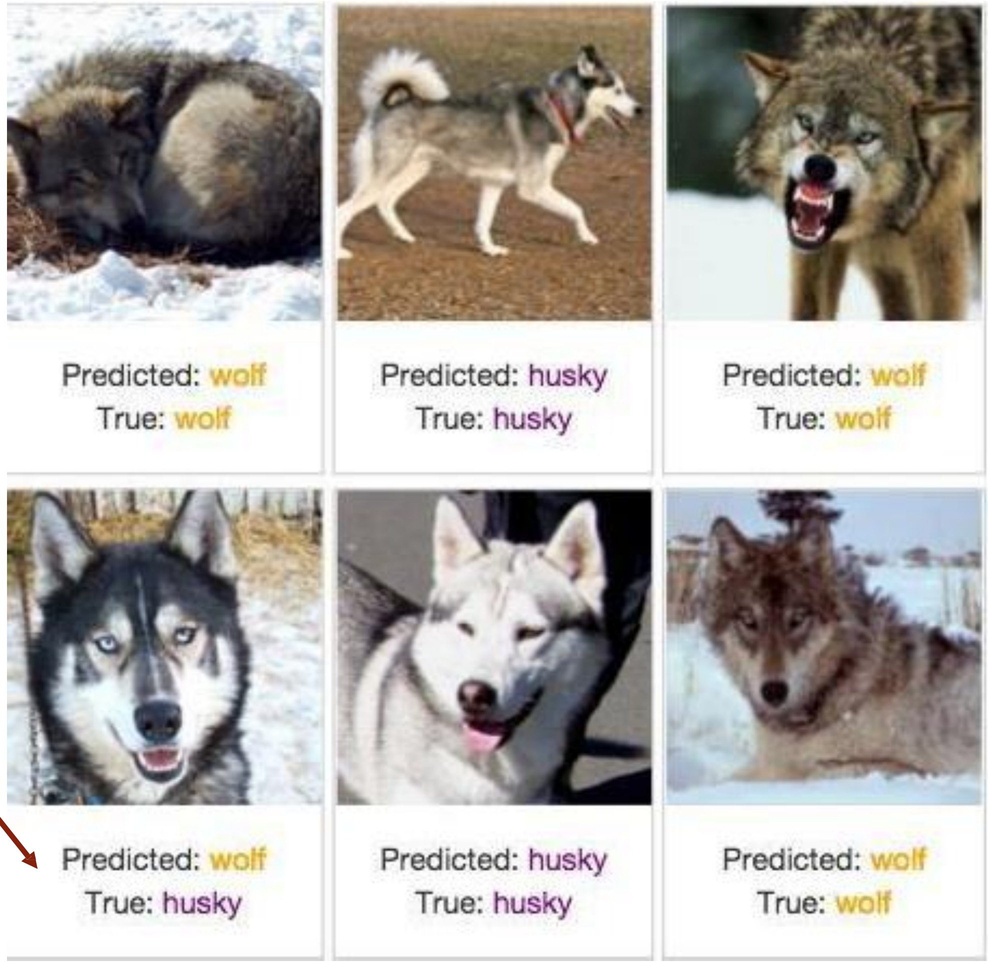


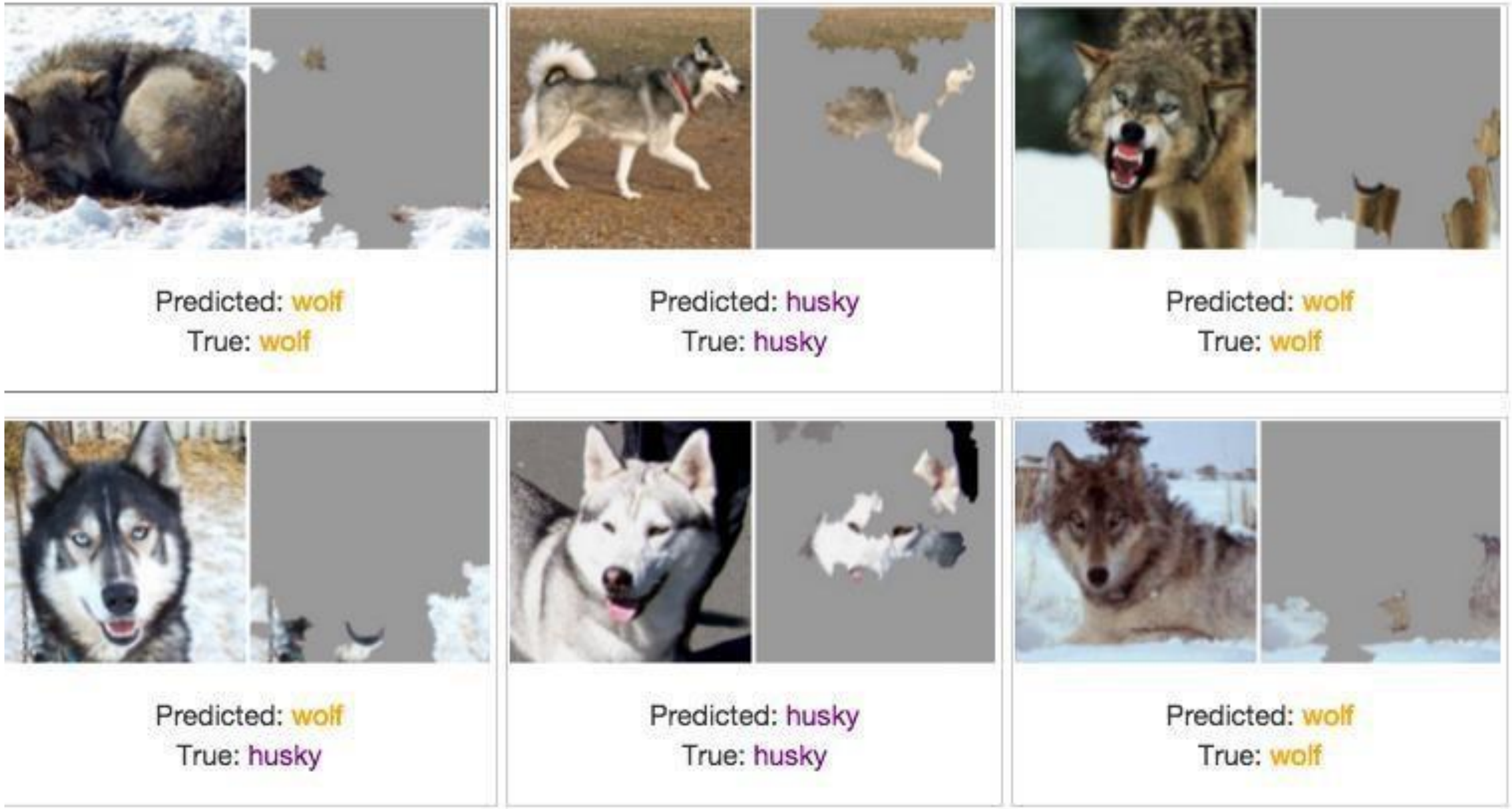
Top label: **“fireboat”**

Why did the network label this image as **“fireboat”**?

CAN YOU BUILD YOUR TRUST BASED ON ACCURACY?

Only 1 mistake!



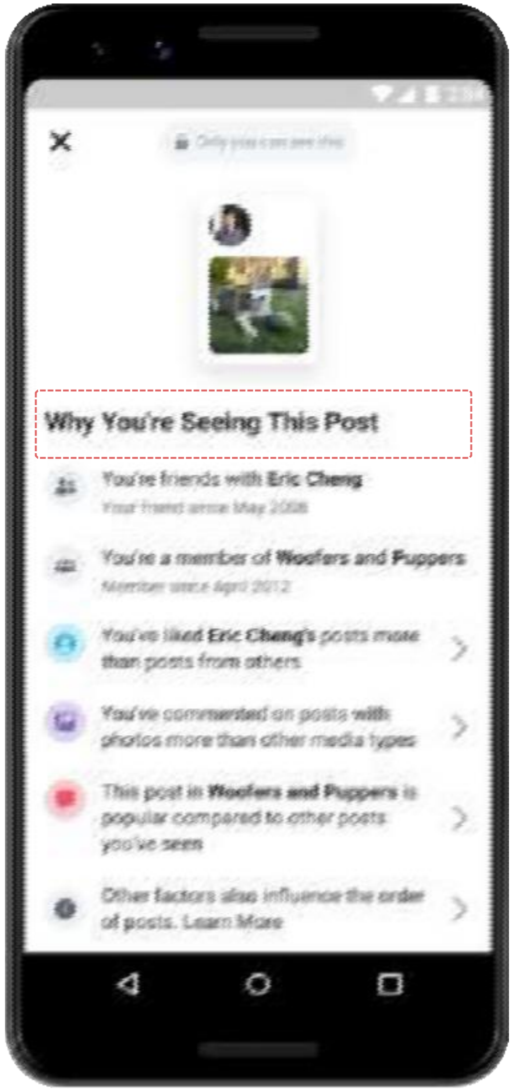


Predicted: wolf
True: wolf
Predicted: husky
True: husky
Predicted: wolf
True: wolf
Predicted: wolf
True: husky
Predicted: husky
True: husky
Predicted: wolf
True: wolf

Example: Facebook adds Explainable AI to build Trust

March 31, 2019

Why Am I Seeing This? We Have an Answer for You



Why You're Seeing This Post

You're friends with **Eric Cheng**
Your friend since May 2008

You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with photos more than other media types

This post in **Woofers and Puppers** is popular compared to other posts you've seen

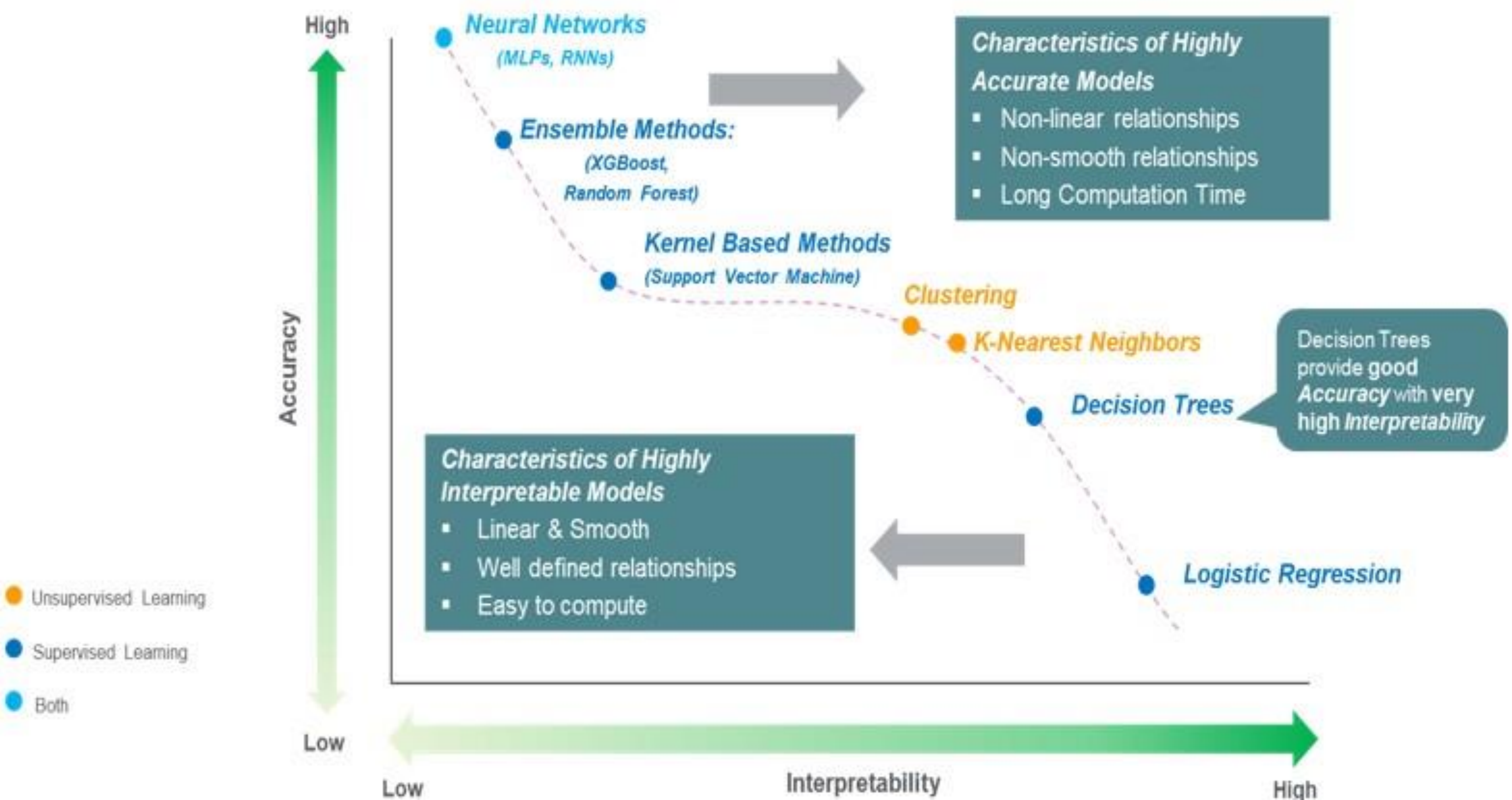
Other factors also influence the order of posts. Learn More

Criteria for Model Interpretation Methods

- ## Intrinsic or post hoc?
  - **Intrinsic interpretability** is all about leveraging a machine learning model which is intrinsically interpretable in nature (like linear models, parametric models or tree based models).
  - **Post hoc interpretability** means selecting and training a black box model (ensemble methods or neural networks) and applying interpretability methods after the training
- ## Model-specific or model-agnostic?
  - **Model-specific interpretation tools** are very specific to intrinsic model interpretation methods which depend purely on the capabilities and features on a per-model basis.
  - **Model-agnostic tools** are more relevant to post hoc methods and can be used on any machine learning model. These agnostic methods usually operate by analyzing (and perturbations of inputs) feature input and output pairs.
- ## Local or global?

  This classification of interpretation talks about if the interpretation method explains a single prediction or the entire model behavior?

The Accuracy vs. Interpretability Trade-off



Characteristics of Highly Accurate Models

- Non-linear relationships
- Non-smooth relationships
- Long Computation Time

Decision Trees provide **good** ***Accuracy*** with **very high** ***Interpretability***

Characteristics of Highly Interpretable Models

- Linear & Smooth
- Well defined relationships
- Easy to compute

Techniques for Interpreting ML Models - Structured Data

1. **Using Interpretable Models**

   Use models which are interpretable like linear models or decision trees or RuleFit models

2. **Model Feature Importances**

   Feature importance is generic term for the degree to which a predictive model relies on a particular feature. This can be model specific or model agnostic

3. **Partial Dependence Plots**

   Partial Dependence describes the average marginal impact of a feature on model prediction, holding other features in the model constant and perturbing the feature value

4. **Individual Conditional Expectation Plots**

   An ICE plot visualizes the dependence of the prediction on a feature for each instance separately, resulting in one line per instance, compared to one line overall in partial dependence plots

5. **Global Surrogate Models**

   Model-agnostic surrogate models approximate the predictions of the underlying model as closely as possible while being interpretable by fitting interpretable models on the source model predictions

6. **Local Interpretable Model-agnostic Explanations (LIME)**

   LIME focuses on fitting local surrogate models to explain how single prediction decisions were made.

7. **Shapley Values and SHapley Additive exPlanations (SHAP)**

   SHAP values try to explain the output of a model (function) as a sum of the effects of each feature being introduced into a conditional expectation (avg over different orderings)