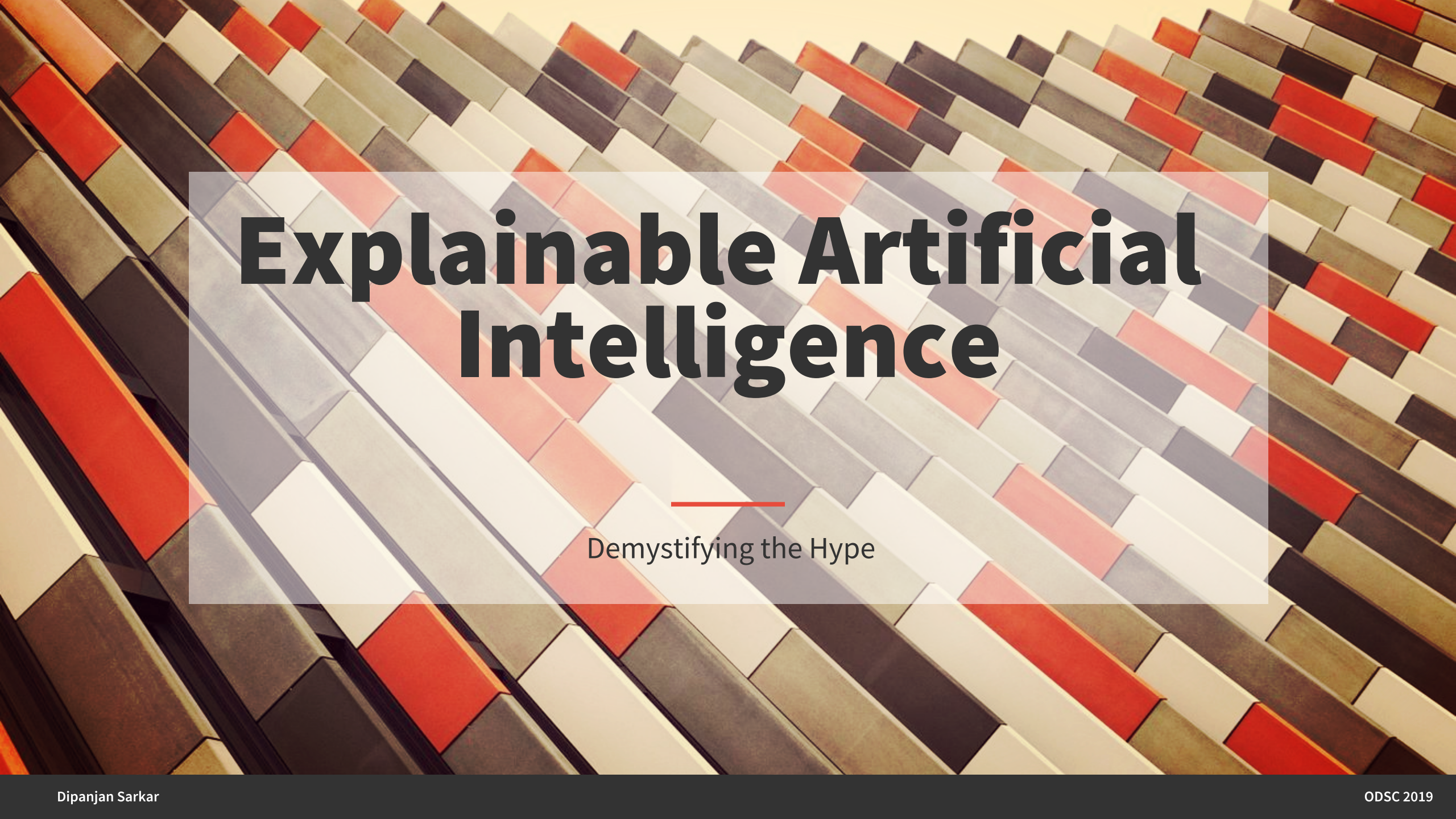


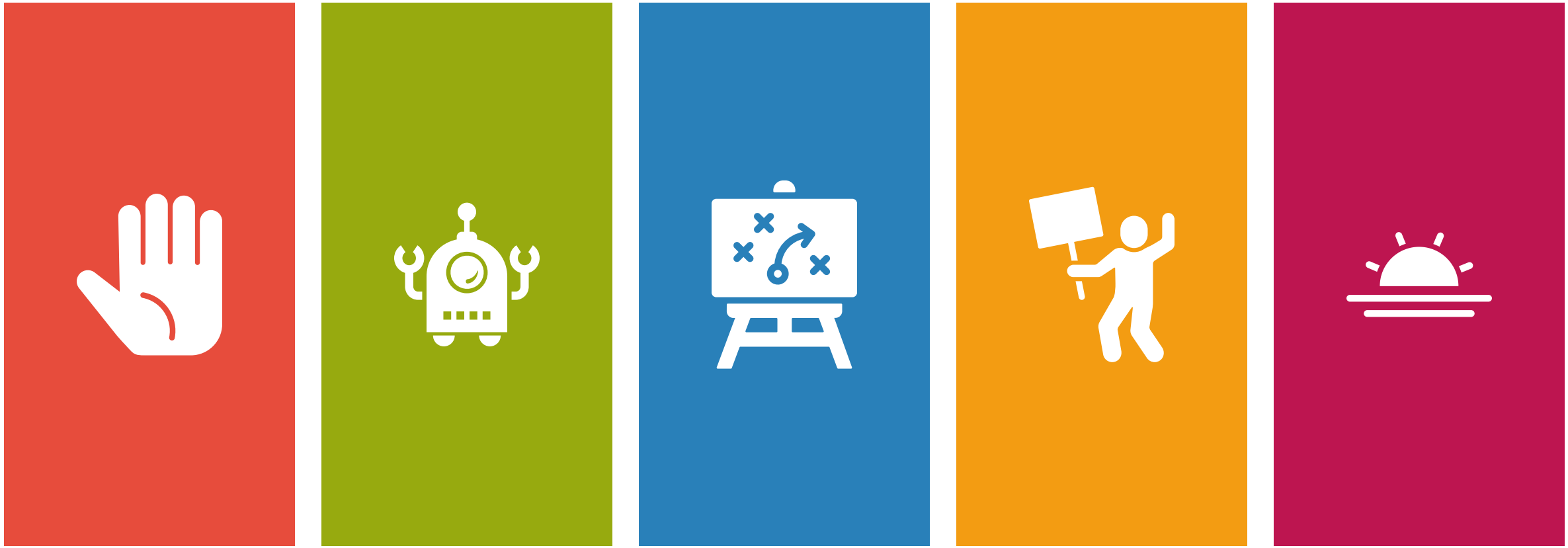
Explainable Artificial Intelligence

Demystifying the Hype

Slides and Code

http://bit.ly/xai_odsc19

Session Agenda



Introduction

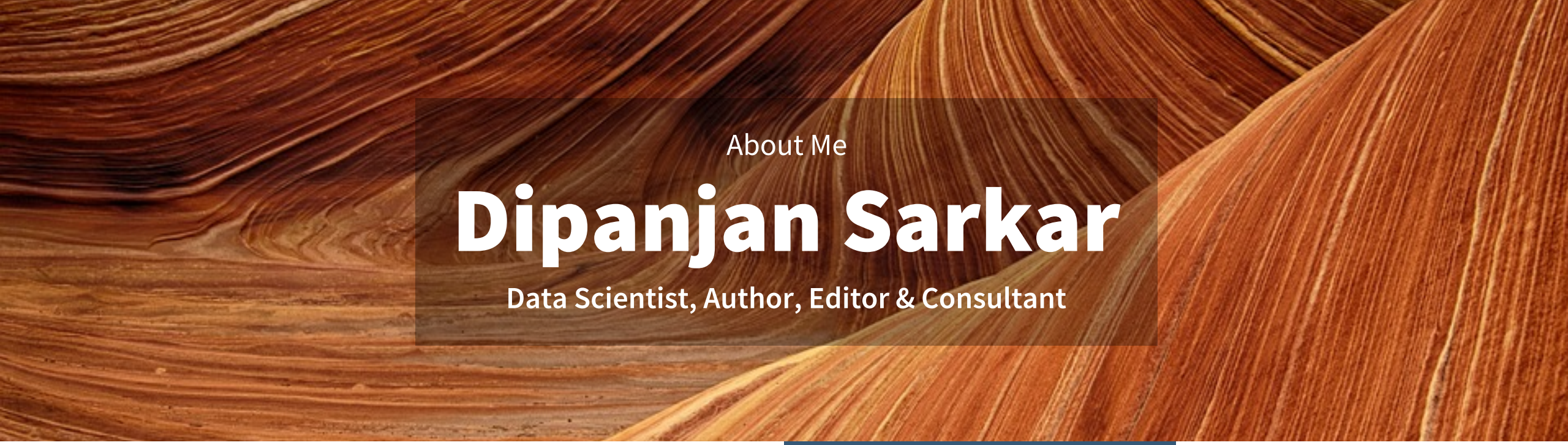
Importance of Explainable AI

Model Interpretation Strategies

Hands-on Model Interpretation

Final Words


Introduction



About Me

Dipanjan Sarkar

Data Scientist, Author, Editor & Consultant

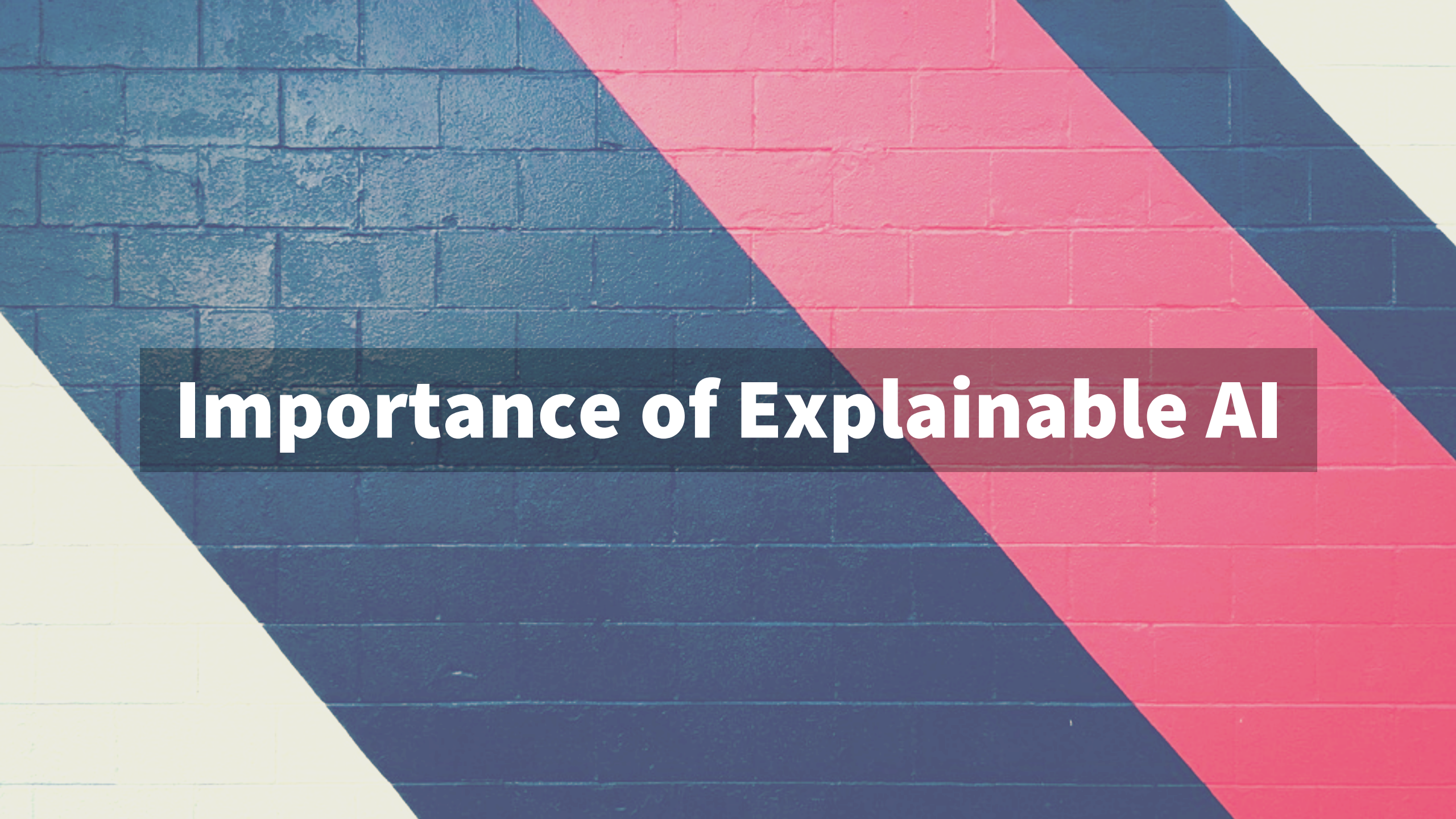
Red Hat

Springboard

Towards Data Science

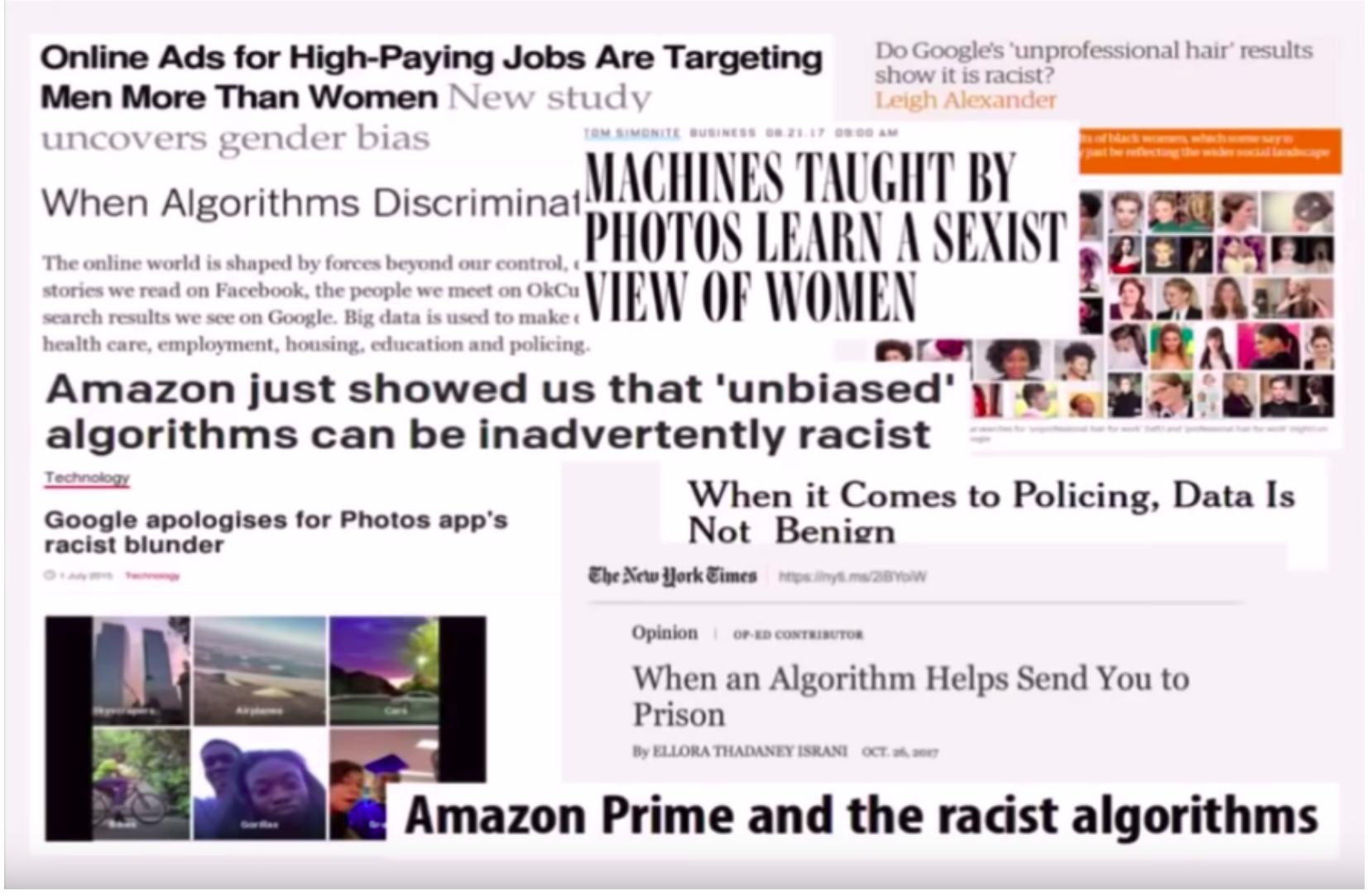
Experts 2019

Machine Learning GDE

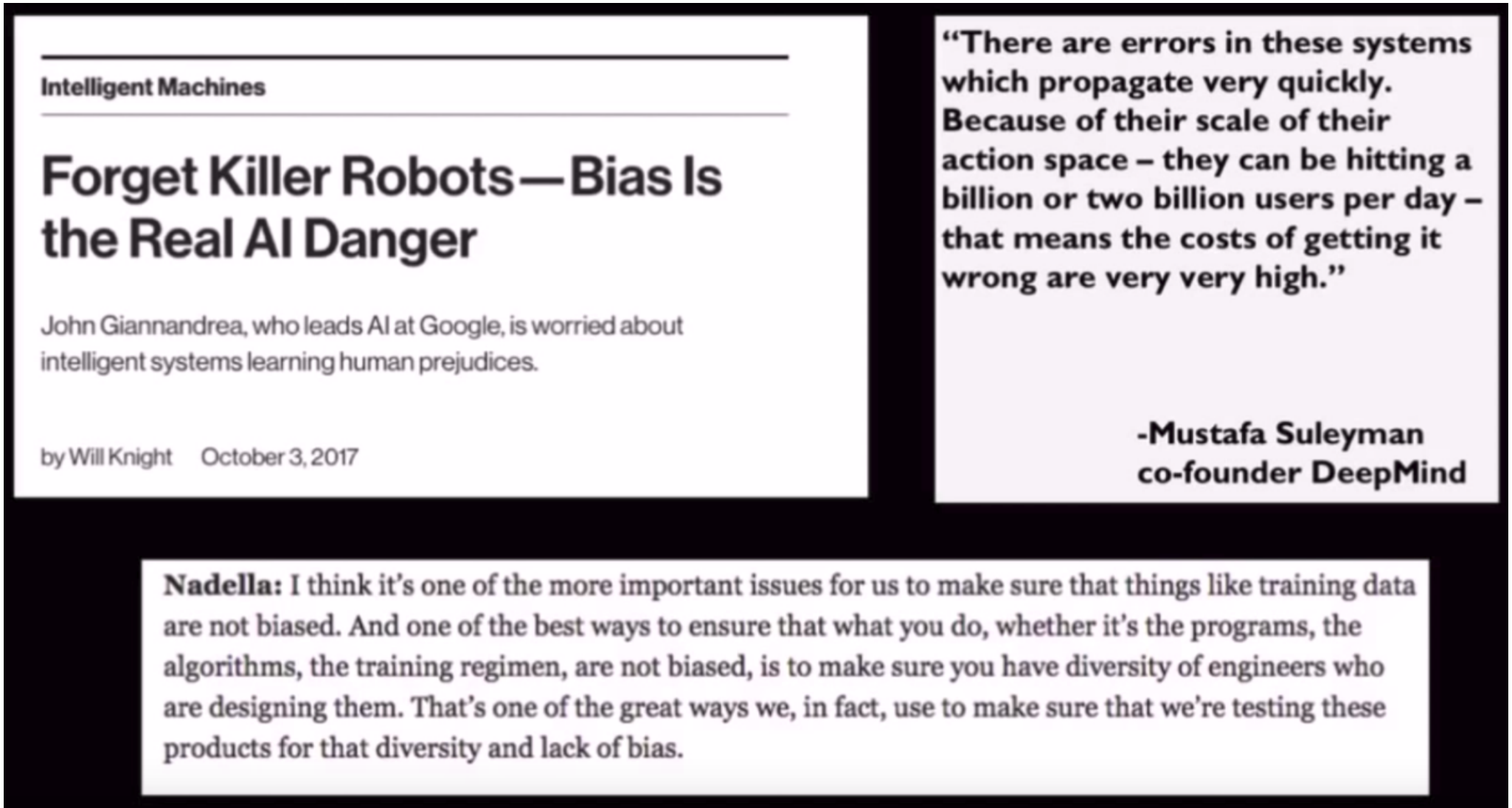

Importance of Explainable AI

The Problem with Machine Learning

The Problem with Machine Learning



The Trouble with Bias in ML Systems



Intelligent Machines

Forget Killer Robots—Bias Is the Real AI Danger

John Giannandrea, who leads AI at Google, is worried about intelligent systems learning human prejudices.

by Will Knight October 3, 2017

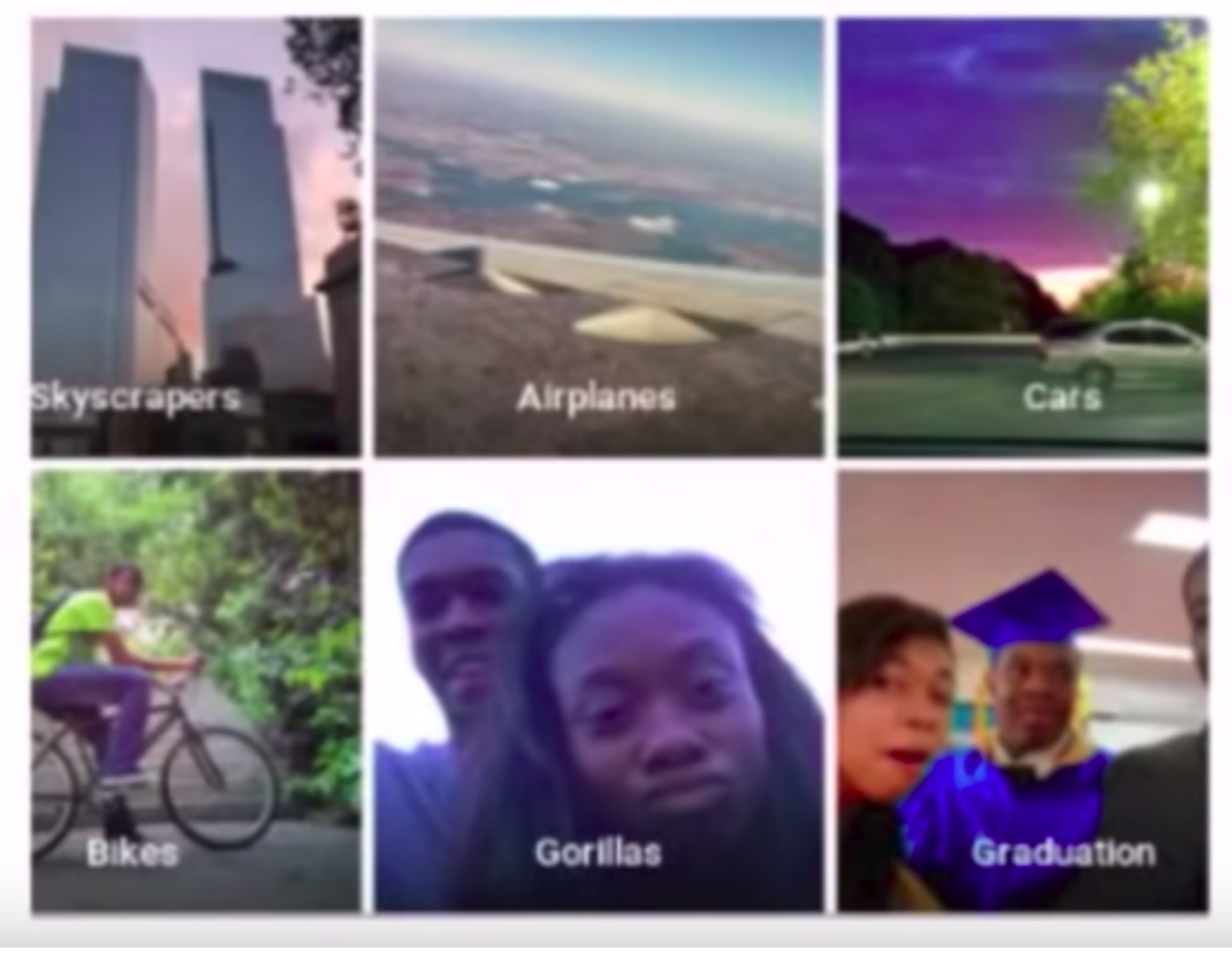
"There are errors in these systems which propagate very quickly. Because of their scale of their action space – they can be hitting a billion or two billion users per day – that means the costs of getting it wrong are very very high."

-Mustafa Suleyman
co-founder DeepMind

Nadella: I think it's one of the more important issues for us to make sure that things like training data are not biased. And one of the best ways to ensure that what you do, whether it's the programs, the algorithms, the training regimen, are not biased, is to make sure you have diversity of engineers who are designing them. That's one of the great ways we, in fact, use to make sure that we're testing these products for that diversity and lack of bias.

Impact of Bias in ML Systems

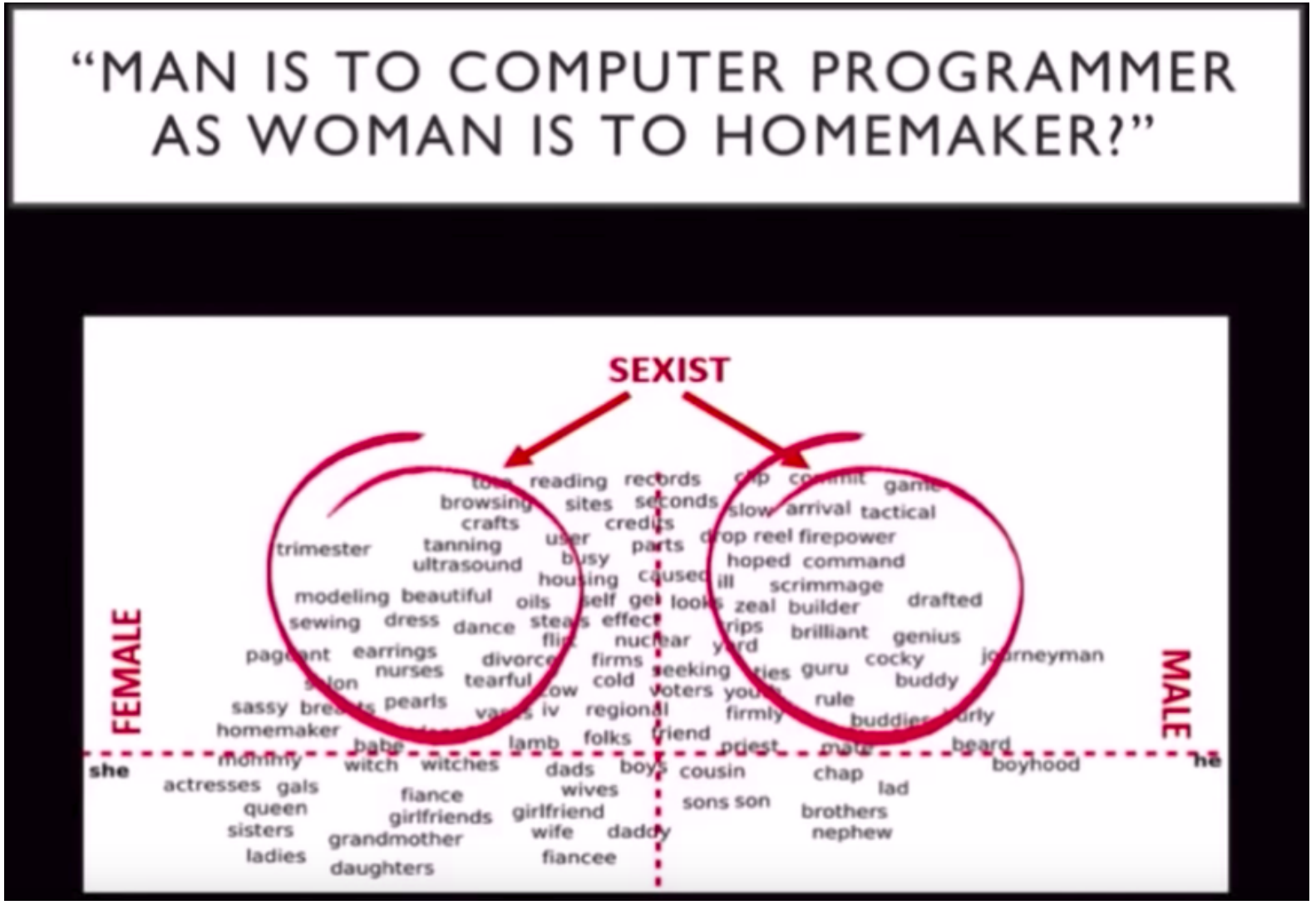
Impact of Bias in ML Systems



Impact of Bias in ML Systems

Impact of Bias in ML Systems

Impact of Bias in ML Systems



Impact of Bias in ML Systems

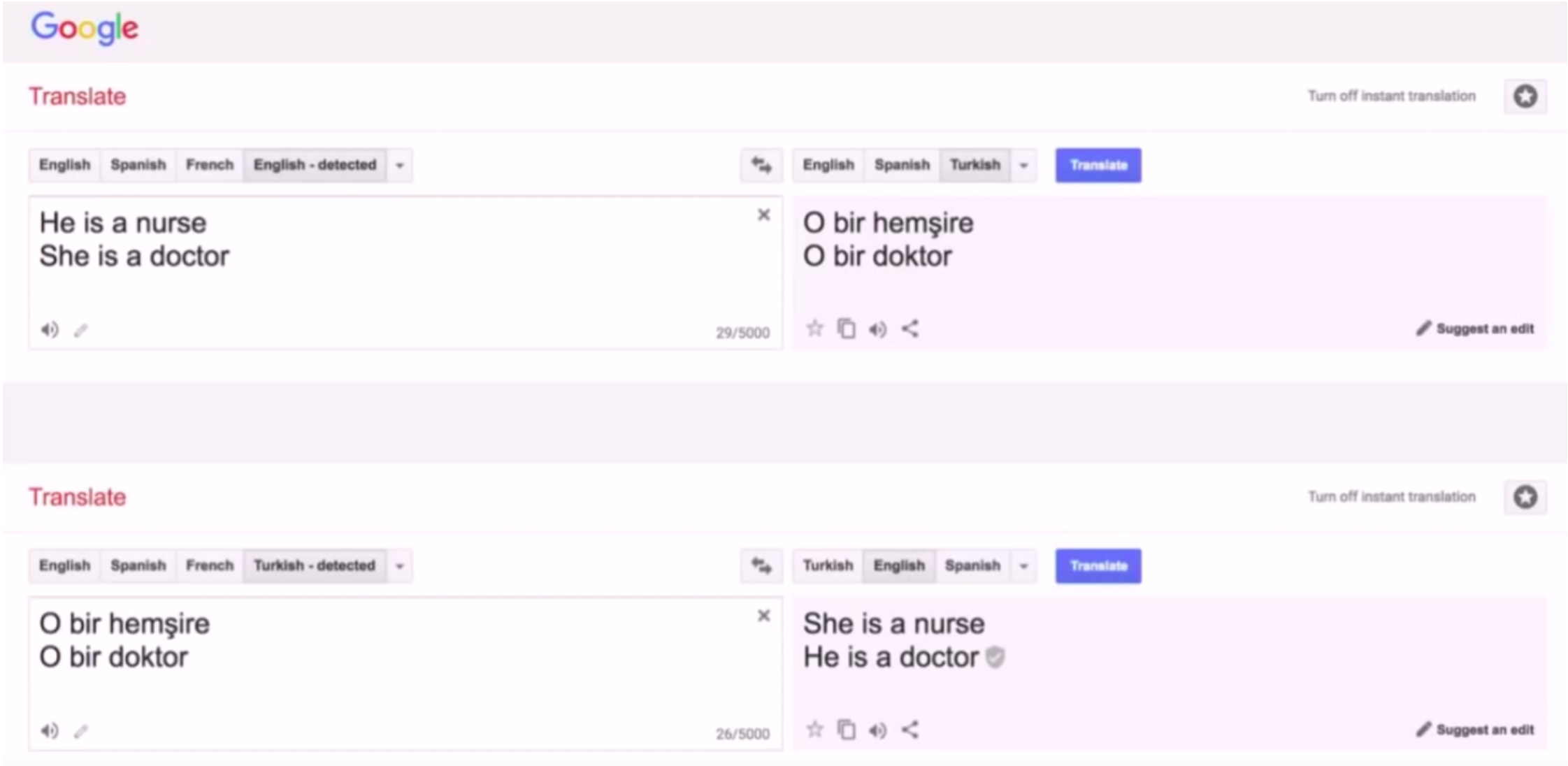
```
In [7]: model.most_similar(positive=['computer_programmer', 'woman'], negative=['man'])

Out[7]: [('homemaker', 0.562711894512176 5),
         ('housewife', 0.510504722595214 8),
         ('graphic_designer', 0.505180299282074),
         ('schoolteacher', 0.4979494214057922 4),
```

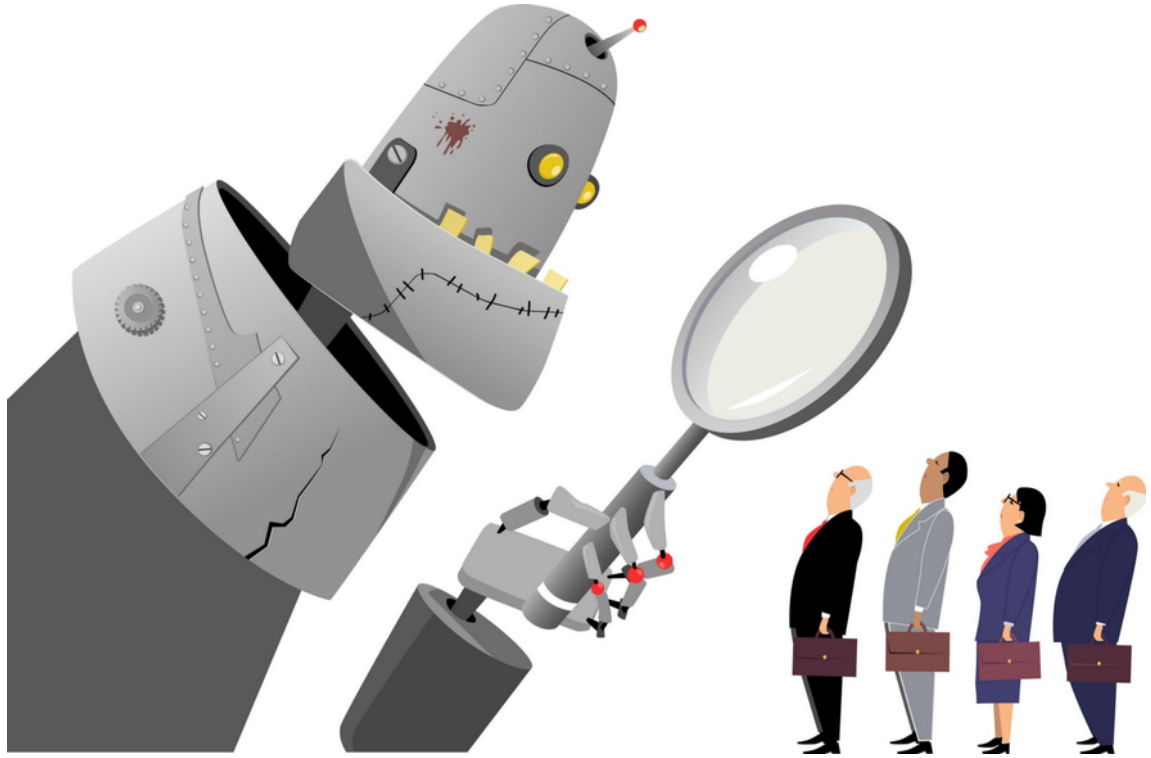
```
In [10]: model.most_similar(positive=['mexicans'], topn=30)

Out[10]: [('hispanics', 0.7345616817474365),
          ('latinos', 0.661898851394653 3),
          ('ILLEGALS', 0.657423019409179 7),
          ('LEGAL_immigrants', 0.654155850410461 4),
          ('mexican', 0.649342894554138 2),
          ('thats_ok', 0.634340524673461 9),
          ('americans', 0.632471323013305 7),
          ('illegals', 0.629899621009826 7),
          ('ILLEGAL_aliens', 0.628911614418029 8),
```

Impact of Bias in ML Systems

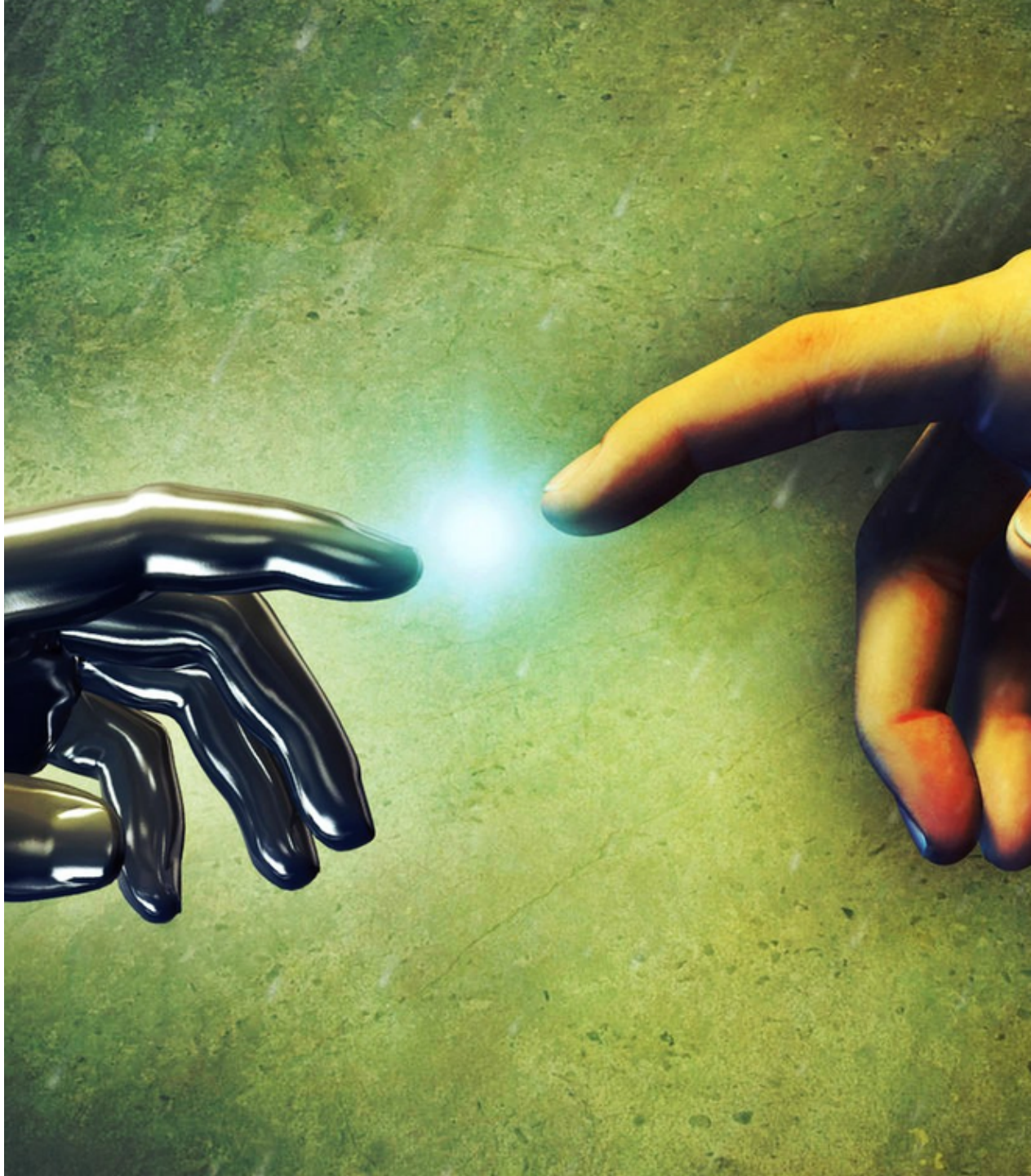


- While some might argue that if something is working well, why question how it works?
- Being humans, logic and reasoning is something we adhere to for most of our decisions
- Hence, in the paradigm shift towards artificial intelligence (AI) making decisions will no doubt be questioned.
- There are a lot of real-world scenarios where biased models might have really adverse effects
- But you don't need explainable models for everything!



Model Interpretation Strategies

Interpretability also popularly known as **human-interpretable interpretations (HII)** of a machine learning model is the extent to which a human (including non-experts in machine learning) can understand the choices taken by models in their decision-making process (the **how**, **why** and **what**).



Understanding Model Interpretation

- ## What drives model predictions?

  We should have the ability to query our model and find out latent feature interactions to get an idea of which features might be important in the decision-making policies of the model. This ensures **fairness** of the model.

- ## Why did the model take a certain decision?

  We should also be able to validate and justify why certain key features were responsible in driving certain decisions taken by a model during predictions. This ensures **accountability** and **reliability** of the model.

- ## How can we trust model predictions?

  We should be able to evaluate and validate any data point and how a model takes decisions on it. This should be demonstrable and easy to understand for key stakeholders that the model works as expected. This ensures **transparency** of the model.

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?

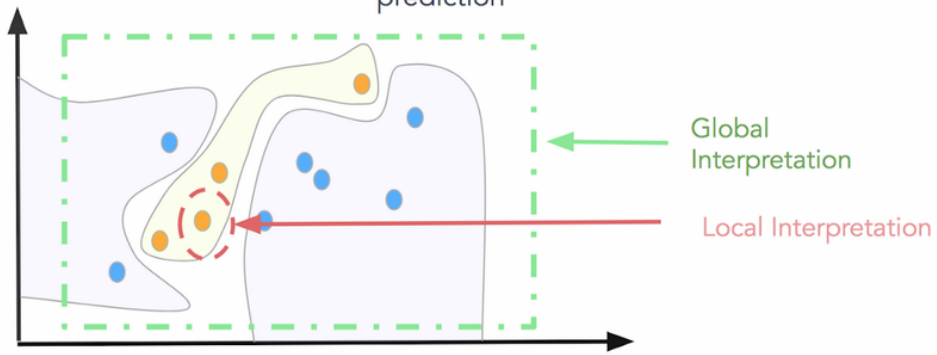
Scope of Model Interpretation

Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction



- **Global Interpretations**

  This is all about trying to understand **“How does the model make predictions?”** and **”How do comprehend and interpret the whole model at once?”**

- **Local Interpretations**

  This is all about trying to understand **“Why did the model make specific decisions for a single instance?”** and **”Why did the model make specific decisions for a group of instances?”**

Traditional Techniques for Model Interpretation

- **Exploratory analysis and visualization**

  These include techniques like clustering and dimensionality reduction.

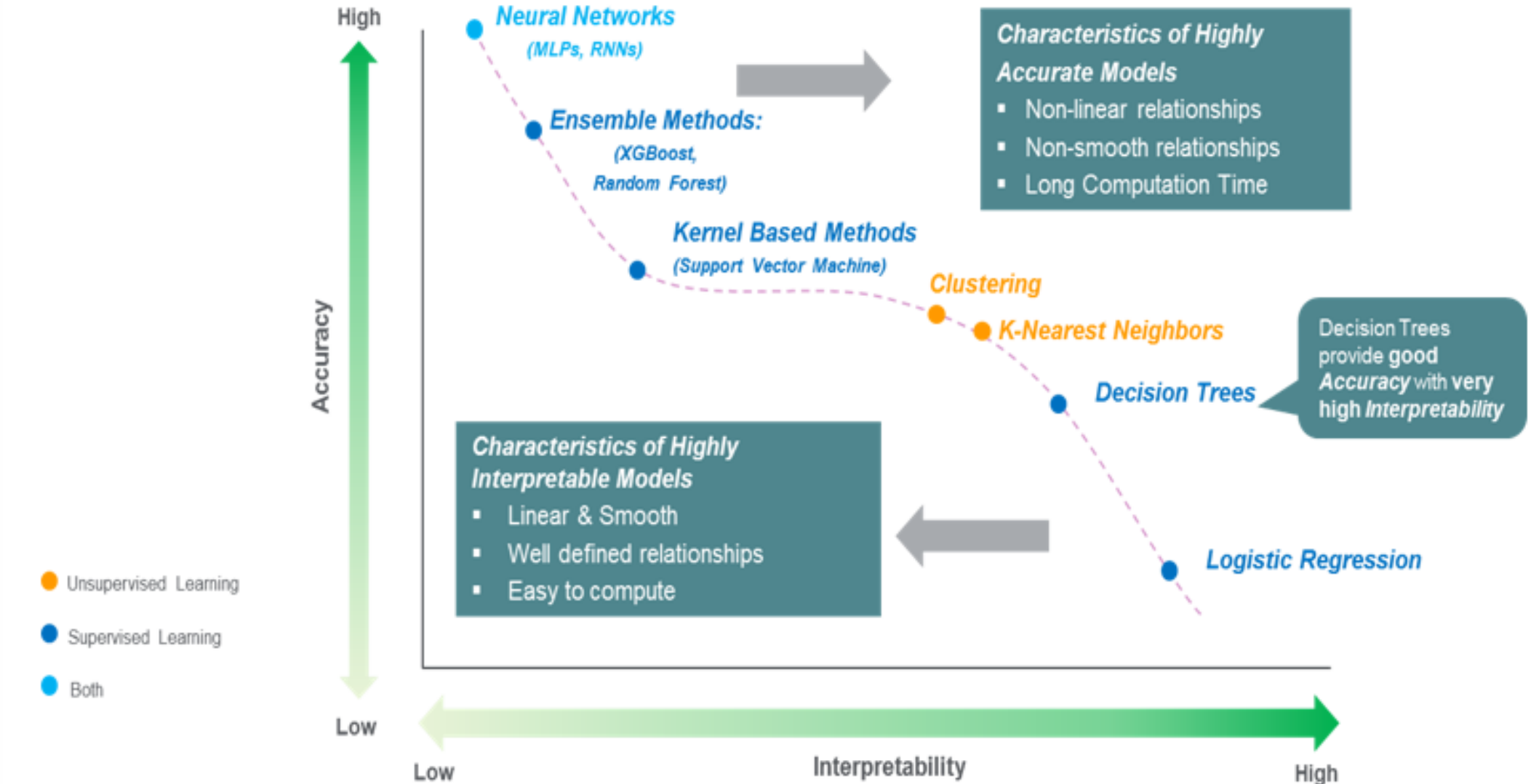
- **Model performance evaluation metrics**

  These include techniques like precision, recall, accuracy, ROC curve and the AUC (for classification models) and the coefficient of determination (R-square), root mean-square error, mean absolute error (for regression models)

- **Use Interpretable Models**

  These include linear models like linear or logistic regression and non-linear models like decision trees

The Accuracy vs. Interpretability Trade-off



Hands-on Sessions for Model Interpretation

1. Model Interpretation Techniques on Structured Data
2. Model Intepretation on Unstructured Text Data
3. Model Interpretation for Computer Vision

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)

Techniques for Interpreting DL Models - Structured Data & Text Data

1

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)

2

Deep SHAP

Decomposes the output prediction of a neural network on a distribution of input samples by backpropagating the contributions of all neurons in the network to every feature of the input samples.

Techniques for Interpreting DL Models - Image Data

1. ## SHAP Gradient Explainer

   Combines ideas from Integrated Gradients, SHAP, and SmoothGrad into a single expected value equation

2. ## Visualizing Activation Layers

   Visualize how a given input comes out of specific activation layers. Explores which feature maps are getting activated in the model

3. ## Occlusion Sensitivity

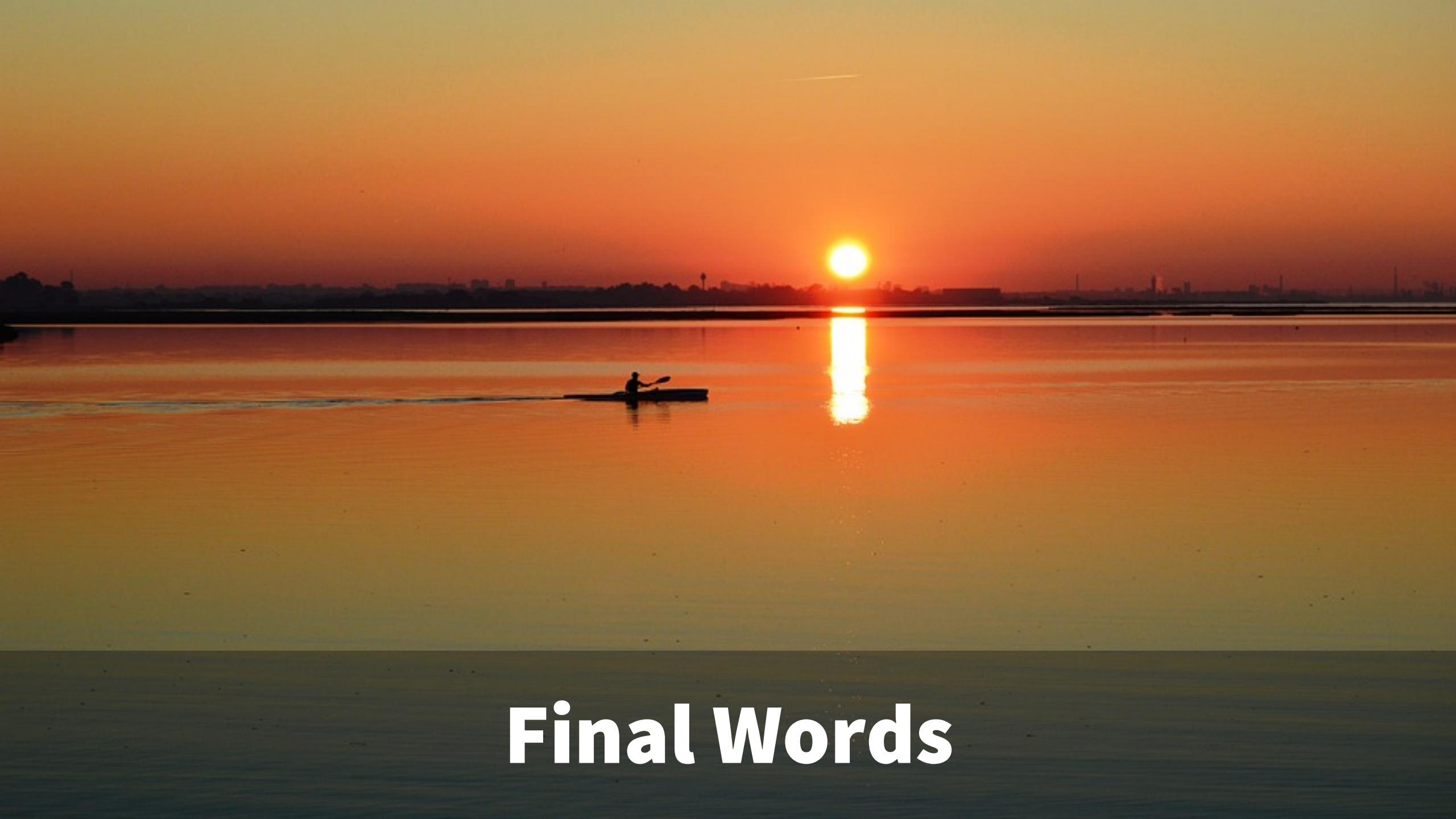
   Visualize how parts of the image affects neural network's confidence by occluding \ hiding parts of the image iteratively

4. ## Grad-CAM

   Visualize how parts of the image affects neural network's output by looking into the gradients backpropagated to the class activation maps

5. ## SmoothGrad

   Averaging gradient sensitivity maps for an input image to identify pixels of interest



Final Words

Final Words

- Models are not biased by themselves, the roots lie in bias in data and humans
- Need unified tools and processes for XAI
- Most interpretable methods are not scalable
- Don't force-fit XAI if it may not be necessary (depends on your problem objective!)
- Frequent Model and Data Auditing will be necessary
- With fairness and transparency fueled by explainability in AI, we can make the world a better place

Contact Me

Interested in collaborating, research or writing?

LinkedIn

https://www.linkedin.com/in/dipanzan

GitHub

https://github.com/dipanjanS

Towards Data Science

https://towardsdatascience.com/@dipanzan.sarkar