Lecture 10: Explainable ML

CMSC 25910

Spring 2024

The University of Chicago

THE UNIVERSITY OF CHICAGO

The Evolution of the Right to an Explanation

US Equal Credit Opportunity Act (1974)

- ECOA requires creditors to provide notification about specific actions taken, as well as to provide an explanation
- “(2) *Statement of specific reasons.* The statement of **reasons for adverse action** required by paragraph (a)(2)(i) of this section **must be specific and indicate the principal reason(s)** for the adverse action. Statements that the adverse action was based on the creditor's internal standards or policies or that the applicant, joint applicant, or similar party failed to achieve a qualifying score on the creditor's credit scoring system are insufficient.”

See https://www.ecfr.gov/cgi-bin/text-idx?SID=1bb14fdfd1afc7d3f2d4756d223aeadb&mc=true&node=se12.8.1002_19

GDPR <u>May</u> Provide Such a Right

- “The data subject should have the right not to be subject to a decision…which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her...”
- “In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, **to obtain an explanation of the decision reached after such assessment** and to challenge the decision...”
- “…the controller should use **appropriate mathematical or statistical procedures** for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised...”

See https://www.privacy-regulation.eu/en/recital-71-GDPR.htm

Example Approaches to "Explain" Ads

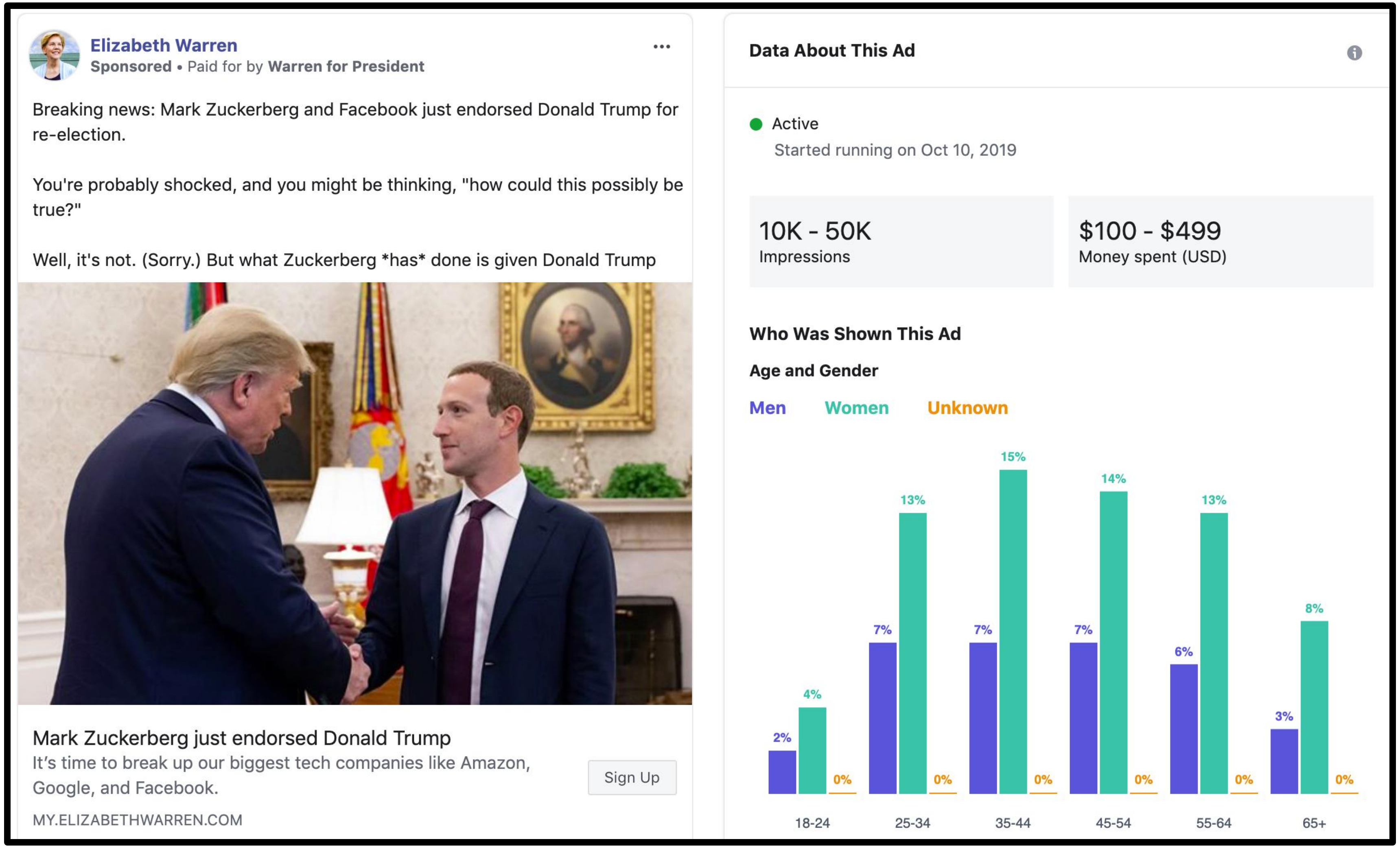
Disclose That An Ad is an Ad (Web)

Ad Consumer Attention

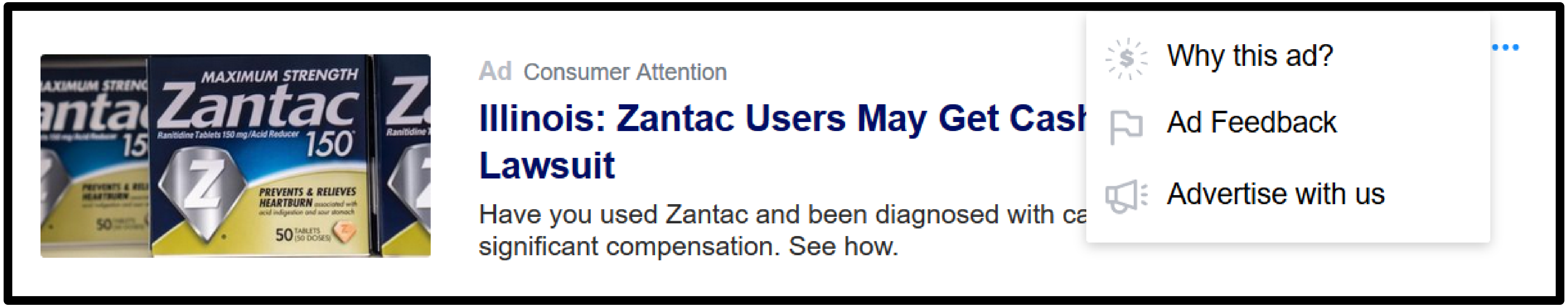
Illinois: Zantac Users May Get Cash Settlement In Lawsuit

Have you used Zantac and been diagnosed with cancer? You may qualify for significant compensation. See how.

Explaining Ad Campaigns (Global)

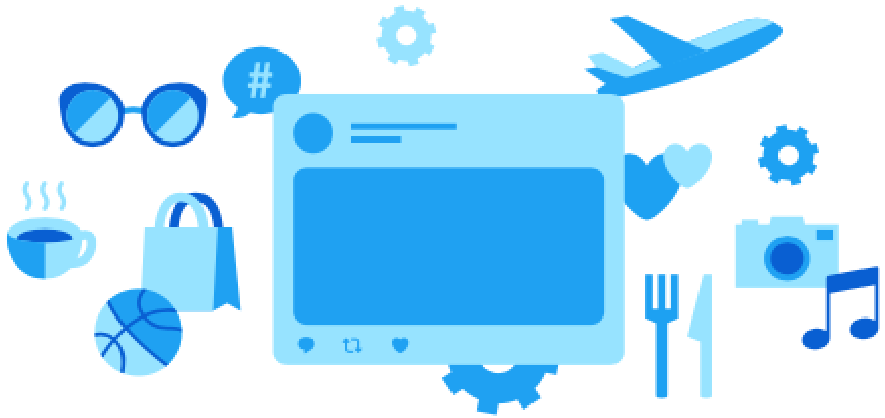


Explaining One Ad (Local)



Explaining One Ad (Local)

Why am I seeing this ad?



One reason you may be seeing this ad is that **Bud Light** wants to reach people interested in **NFL Season 2020-2021**. There may be other reasons you're seeing this ad, including that **Bud Light** wants to reach **people between the ages of 21 and 49 and located here: United States.**

You can view and manage information connected to your account that Twitter may use for ads purposes. See your Twitter data.

Twitter also personalizes ads using information received from partners as well as app and website visits. You can control these interest-based ads using the "Personalize ads" setting.

Explaining One Ad (Local)

Why This Ad?

verizon media

For Consumers

- The sites and apps you use work with online advertising companies to provide you with advertising that is as relevant and useful as possible. Personalization may be informed by various factors such as the content of the site or app you are using, information you provide, historical searches you conduct, what your friends or contacts recommend to you, apps on your device, or based on your other interests. Read about Verizon Media's privacy and advertising practices to learn more about how Verizon Media selects the ads you see.

Who placed this ad?

- This ad was served by Verizon Media or one of Verizon Media's advertising partners.

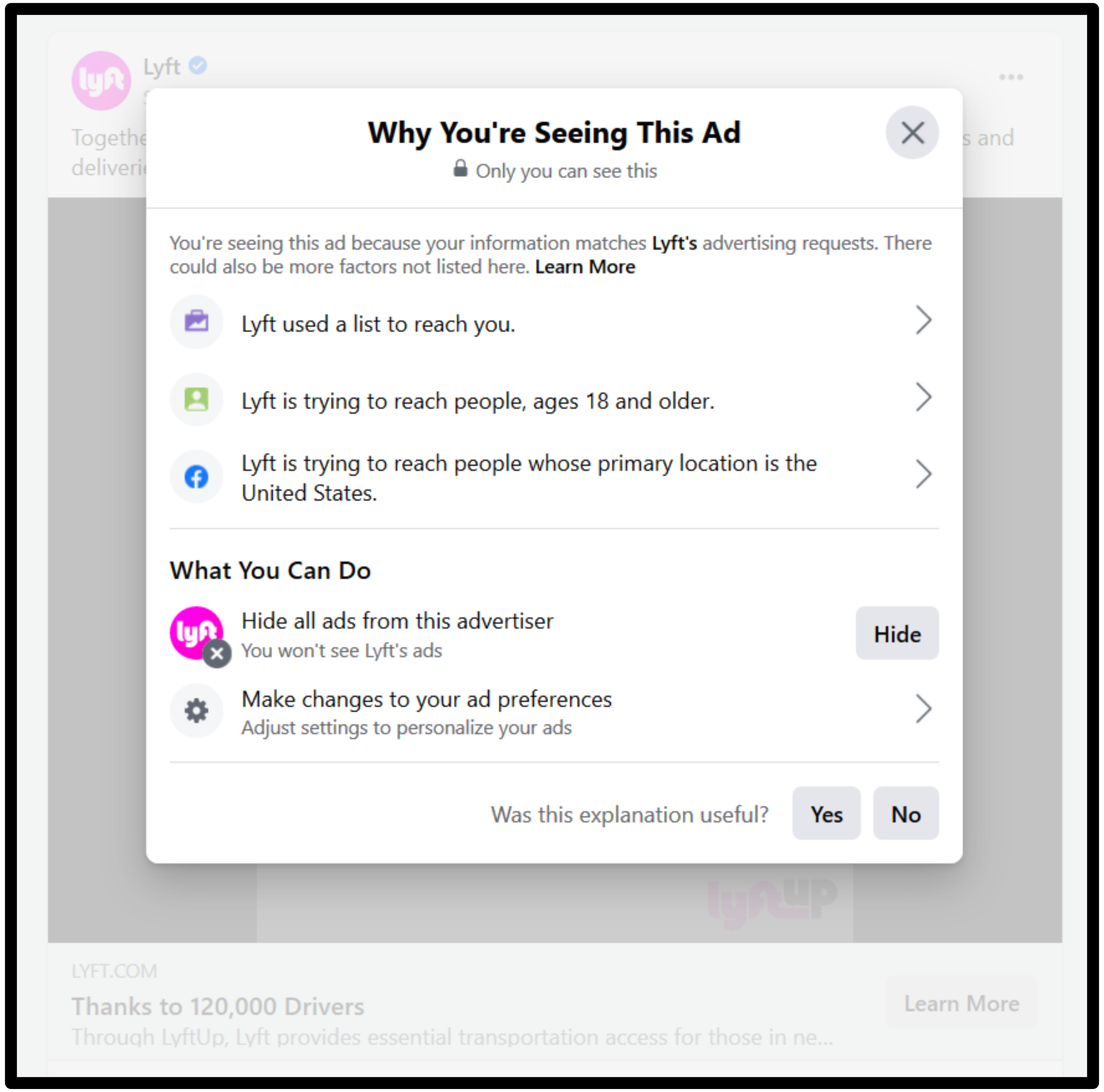
Why was this ad served?

- Certain factors like your activity, searches, demographic data, apps on your device, and location information may be used to select the ads you see.

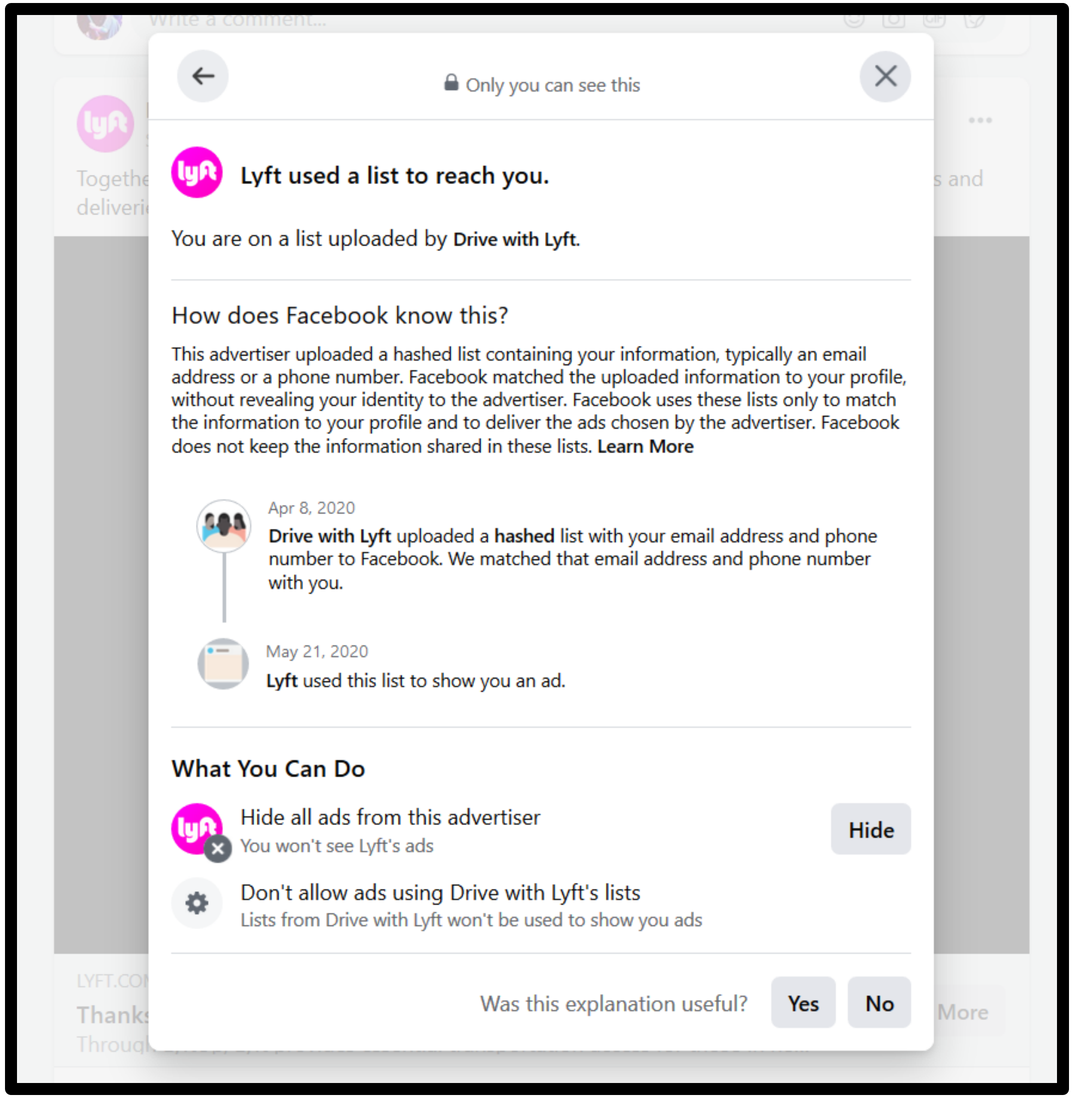
What choices do I have?

- Manage interest-based advertising categories, or opt-out of all categories, from Verizon Media.
- View our other privacy controls.
- Visit the Network Advertising Initiative (US) and the Digital Advertising Alliance Ad Choices - DAA (US), EDAA (EU), DAAC (Canada), ADAA (AU/NZ) to see your opt-out choices from other participating companies.
- Explore other controls and tools to help set and maintain your privacy choices.
- If you are using Safari or a browser enabled with Intelligent Tracking Protection (ITP) or similar cookie-blocking technology, if you wish to opt out of receiving personalized ads, you will need to do so directly via the Verizon Media Privacy Center.

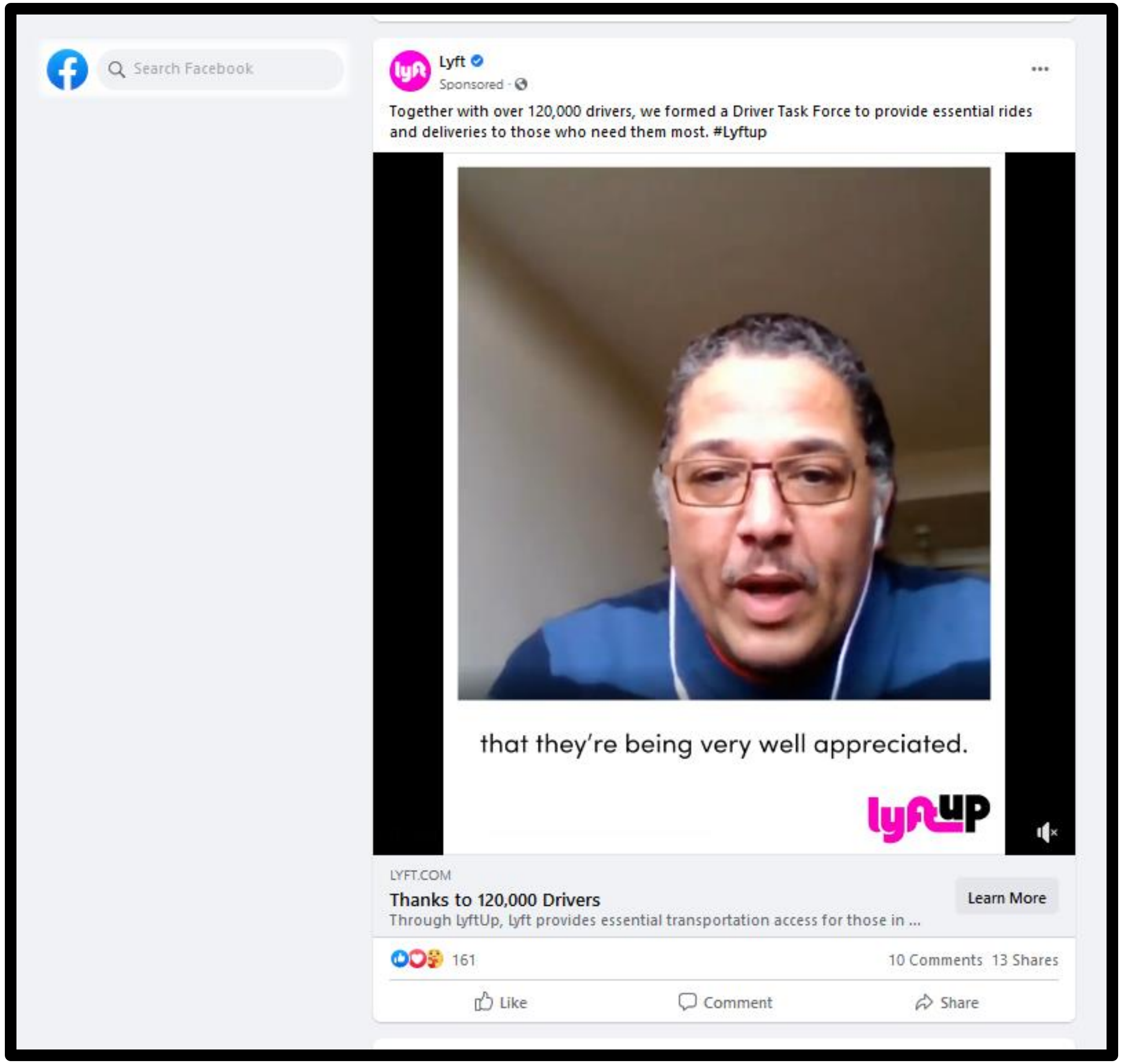
Explaining One Ad (Local)



Explaining One Ad (Local)

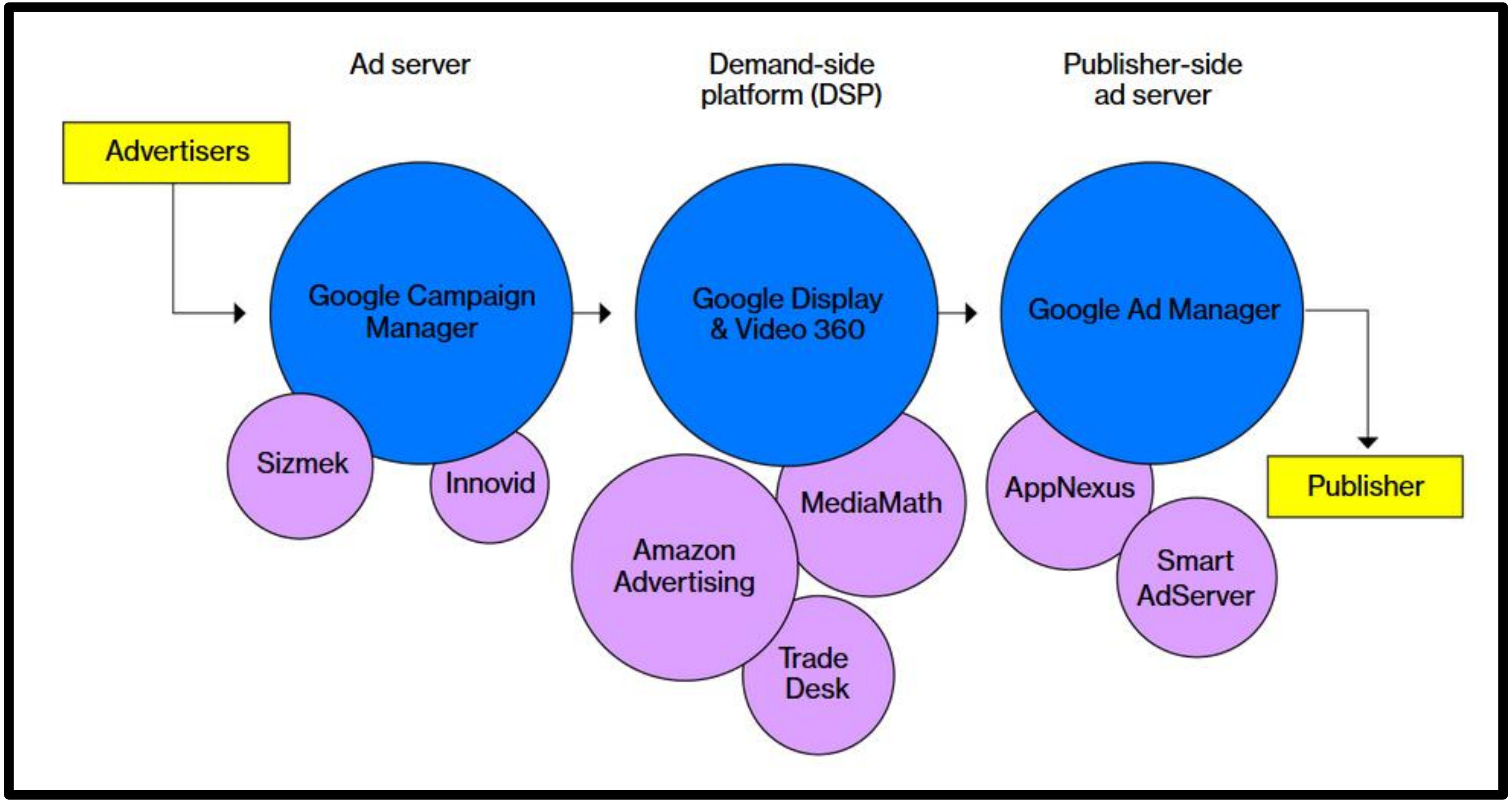


Disclose That An Ad is an Ad (Facebook)



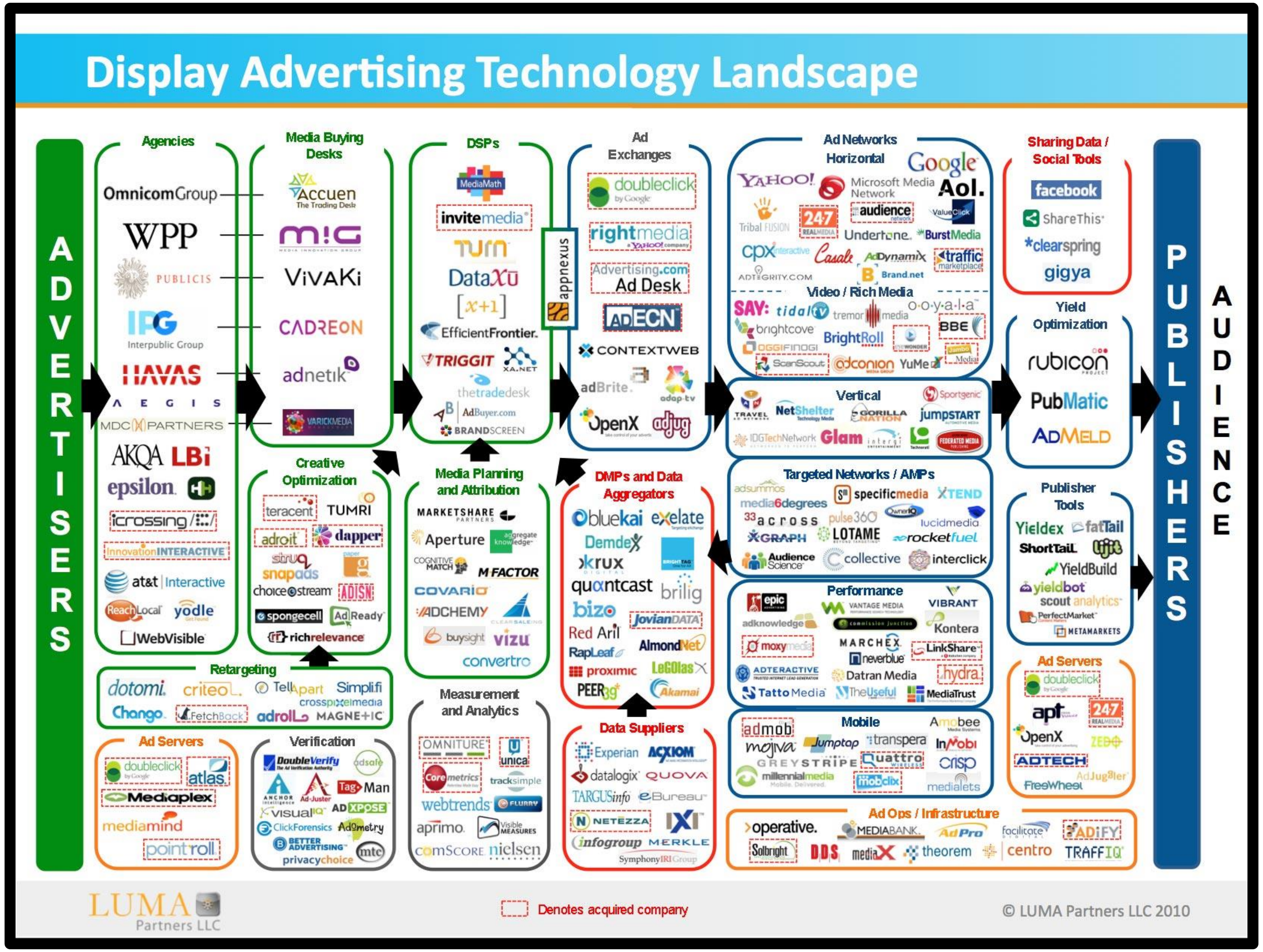
Approaches to “Explain” Models

Is This Enough? (Ad Ecosystem)



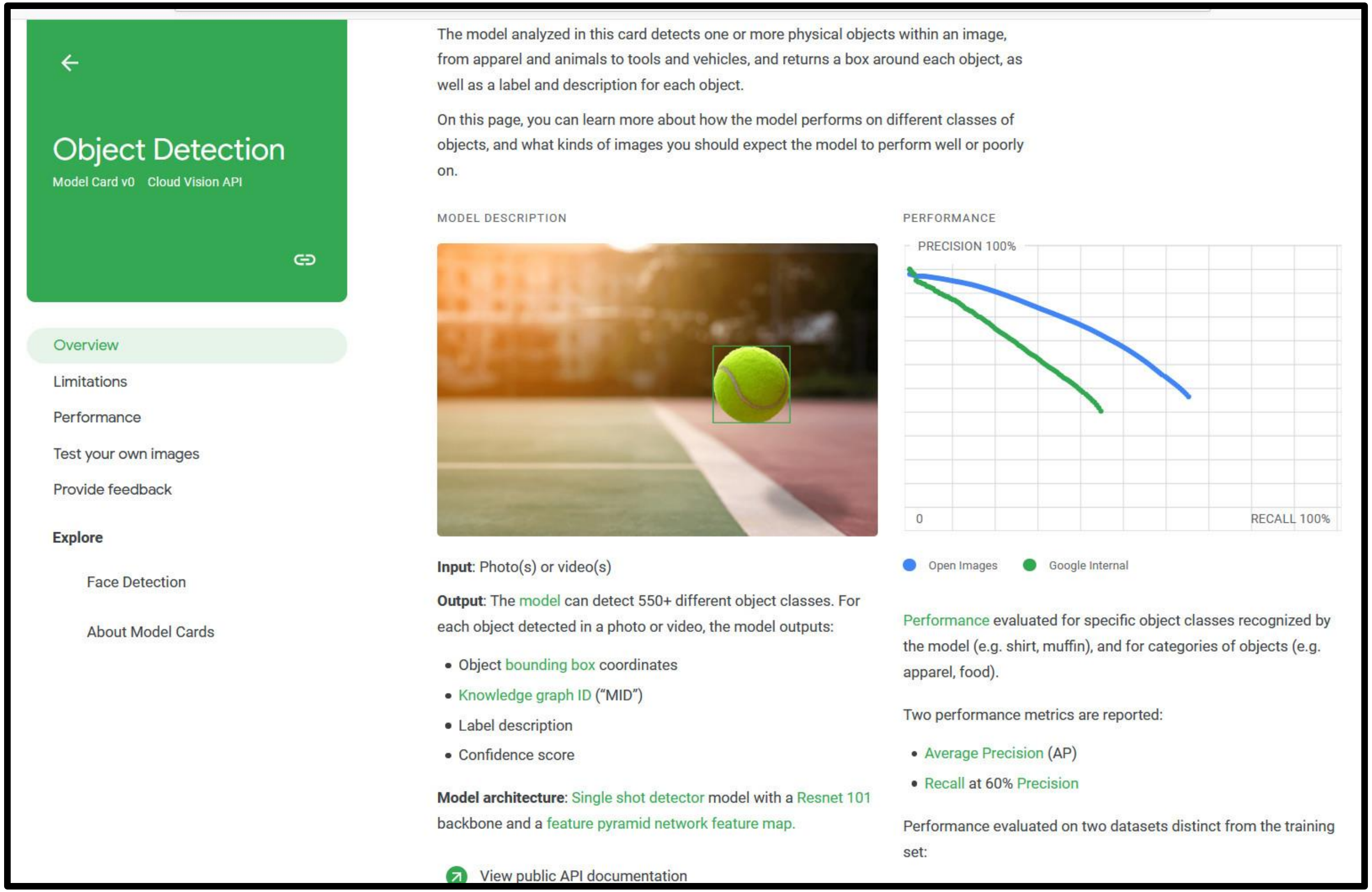
Taken from https://www.bloomberg.com/news/features/2019-10-24/how-google-s-ad-ecosystem-works

Is This Enough? (Ad Ecosystem)



Taken from https://www.adexchanger.com/venture-capital/luma-partners-ad-tech-ecosystem-map-the-december-2010-update/

Explaining a Model (Global)



Object Detection

Model Card v0 Cloud Vision API

- Overview
- Limitations
- Performance
- Test your own images
- Provide feedback

Explore

- Face Detection
- About Model Cards

The model analyzed in this card detects one or more physical objects within an image, from apparel and animals to tools and vehicles, and returns a box around each object, as well as a label and description for each object.

On this page, you can learn more about how the model performs on different classes of objects, and what kinds of images you should expect the model to perform well or poorly on.

MODEL DESCRIPTION

Input: Photo(s) or video(s)

Output: The model can detect 550+ different object classes. For each object detected in a photo or video, the model outputs:

- Object bounding box coordinates
- Knowledge graph ID ("MID")
- Label description
- Confidence score

Model architecture: Single shot detector model with a Resnet 101 backbone and a feature pyramid network feature map.

View public API documentation

PERFORMANCE

Performance evaluated for specific object classes recognized by the model (e.g. shirt, muffin), and for categories of objects (e.g. apparel, food).

Two performance metrics are reported:

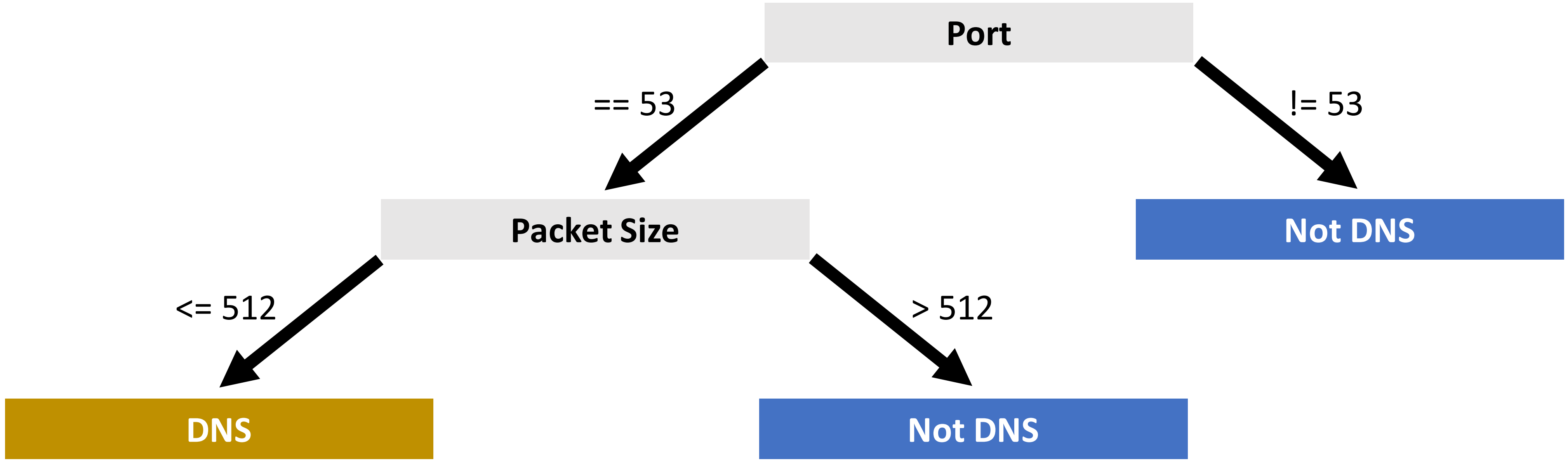
- Average Precision (AP)
- Recall at 60% Precision

Performance evaluated on two datasets distinct from the training set:

https://modelcards.withgoogle.com/ and https://arxiv.org/abs/1810.03993

Inherently Interpretable Models

- Example: decision trees



https://modelcards.withgoogle.com/ and https://arxiv.org/abs/1810.03993

Inherently Interpretable Models

- Example: regression models
  - Linear regression: coefficients
  - Logistic regression: odds ratio

Table 2: Linear regression with the number of scenarios the participant answered correctly as the dependent variable. Higher numbers correspond to more correct answers.

Factor	β	SE	t	p
(Intercept)	9.53	0.40	24.0	**<.001**
Condition: Brave	0.79	0.49	1.61	.108
Condition: Brave-Mobile	0.49	0.49	0.08	.940
Condition: Chrome	1.07	0.49	2.16	**.032**
Condition: Chrome-Mobile	0.62	0.50	1.25	.211
Condition: Chrome-Old	1.09	0.50	2.20	**.028**
Condition: Edge	0.05	0.50	0.10	.923
Condition: Firefox	0.88	0.59	1.80	.073
Condition: Firefox-Mobile	-0.30	0.50	-0.60	.550
Condition: Opera	0.57	0.50	1.15	.252
Condition: Opera-Mobile	-0.34	0.49	-0.70	.484
Condition: Safari	0.78	0.51	1.53	.127
Condition: Safari-Mobile	0.95	0.49	1.93	.055
Gender: Male	0.50	0.20	2.51	**.013**
Technical: Yes	0.49	0.31	1.60	.111
Age Range	-0.65	0.52	-1.24	.216
Browsing in Private Mode (%)	-0.00	0.01	-0.26	.792
Reopened Disclosure (#)	0.19	0.16	1.19	.236

Taken from Wu et al. "Your Secrets Are Safe: How Browsers' Explanations Impact Misconceptions About Private Browsing Mode," WWW 2018. https://www.blaseur.com/papers/www18privatebrowsing.pdf

Highlight (Globally) Important Features

Category	Collection Method	List of Features
Metadata	Google Drive/Dropbox API	account size, used space, file size, file type (img, doc, etc.), extension (jpg, txt, etc.), last modified date, last modifying user, access type (owner, editor, etc.), sensitive filename, sharing status
Documents	Local text processing	bag of words for top 100 content keywords, LDA topic models, TF-IDF vectors, word2vec representations, table schemas for spreadsheets
Images	Google Vision API [20]	image object labels, adult, racy, medical, violent, logos, dominant RGB values, average RGB value
Sensitive Identifiers	Google DLP API [18]	***counts*** of the following identifiers in a file: name, gender, ethnic group, address, email, date of birth, drivers license #, passport #, credit card, SSN, bank account #, VIN

Table 3: A list of the features we automatically collected for each file using multiple APIs and custom code.

Taken from Khan et al. "Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage," USENIX Security 2021. https://www.blaseur.com/papers/usenix21-aletheia.pdf

Highlight (Globally) Important Features

Task		Features
Sensitivity	Documents	gender; fraction of ethnic/VIN/location files; credit card; date of birth; email
	Images	fraction of gender/SSN/ethnic/location files; adult; credit card; racy; passport
Usefulness	Documents	access type; last modifying user; finance keywords; report & journal keywords
	Images	file size; finance keywords; access type; last modifying user; medical keywords
File Management	All Files	usefulness; sensitivity; spoof; account size; used space; finance keywords; medical keywords

Table 8: Top features for prediction tasks. Italicized *keywords* were top terms identified via the bag of words collections.

Taken from Khan et al. "Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage," USENIX Security 2021. https://www.blaseur.com/papers/usenix21-aletheia.pdf

Retrospective Explanations

Explaining Text Classification

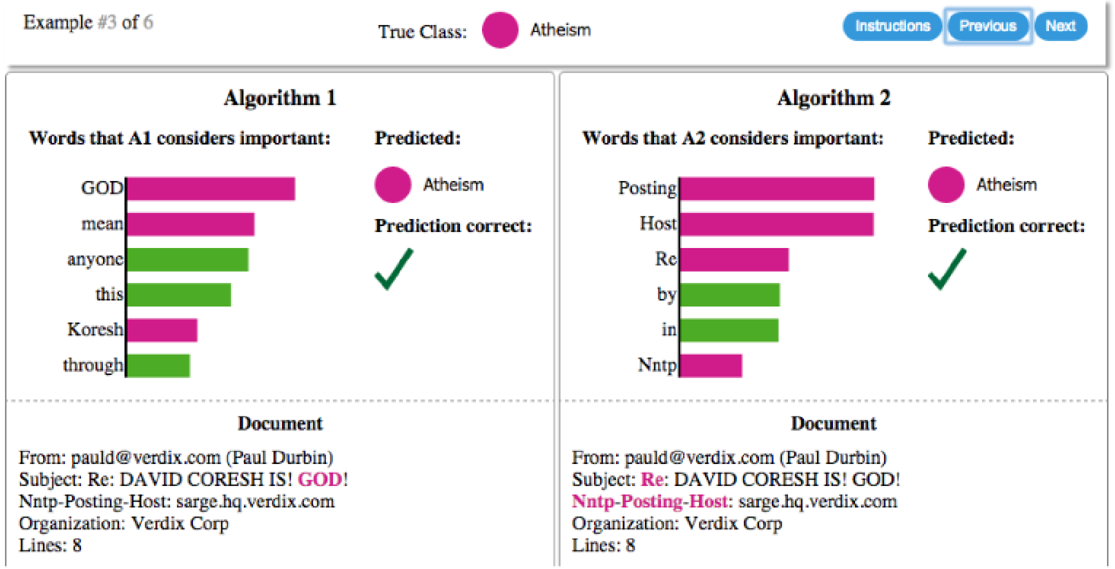
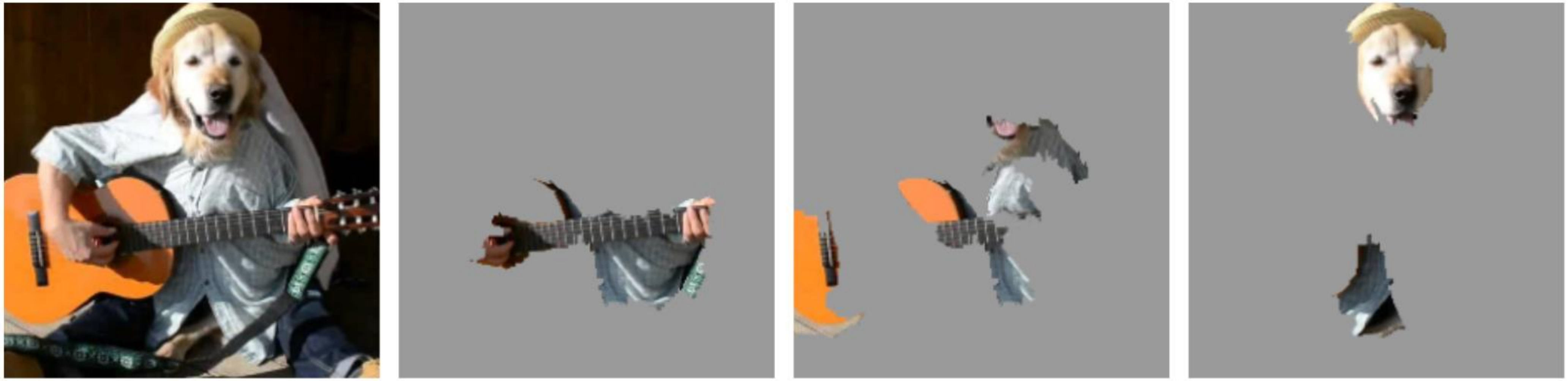


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

Taken from Ribiero et al. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD 2016
https://dl.acm.org/doi/pdf/10.1145/2939672.2939778

Explaining Image Classifications

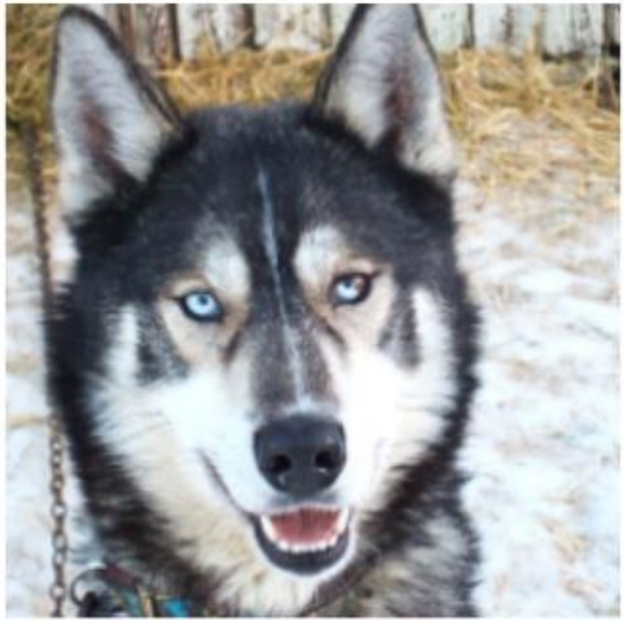


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$)**, "Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

Taken from Ribiero et al. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD 2016
https://dl.acm.org/doi/pdf/10.1145/2939672.2939778

Explaining Image Classifications



(a) Husky classified as wolf

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Taken from Ribiero et al. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD 2016
https://dl.acm.org/doi/pdf/10.1145/2939672.2939778

Explaining Image Classifications

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Taken from Ribiero et al. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD 2016
https://dl.acm.org/doi/pdf/10.1145/2939672.2939778

LIME

- Local Interpretable Model-agnostic Explanations
- Overall goal: “identify an interpretable model over the interpretable representation that is locally faithful to the classifier.”
- Distinguishes between **features** (used by the model) and **interpretable representation** (used to explain to a human)
- “We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way… We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks).”

Ribiero et al. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD 2016
https://dl.acm.org/doi/pdf/10.1145/2939672.2939778

LIME Overview

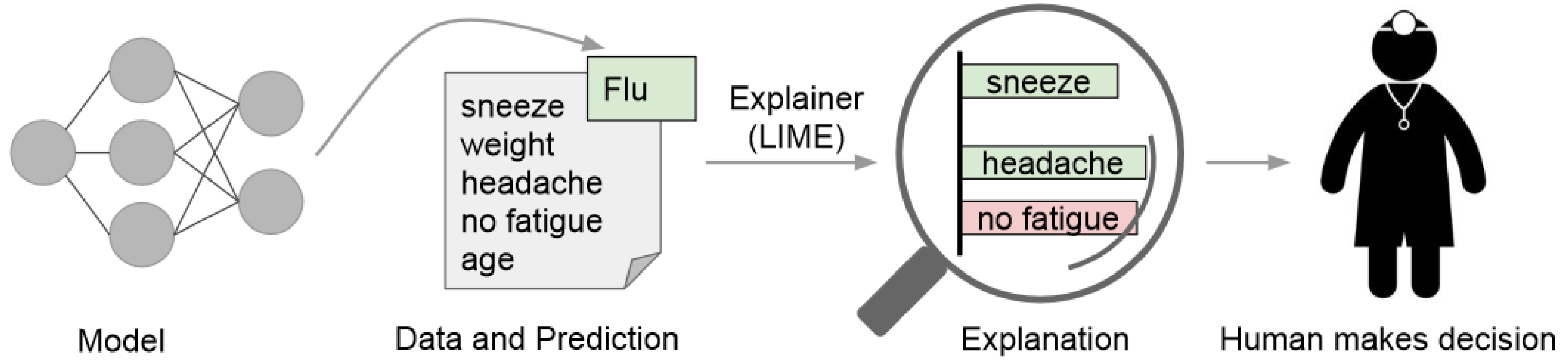


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

Taken from Ribiero et al. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD 2016
https://dl.acm.org/doi/pdf/10.1145/2939672.2939778

Local vs. Global Explanations

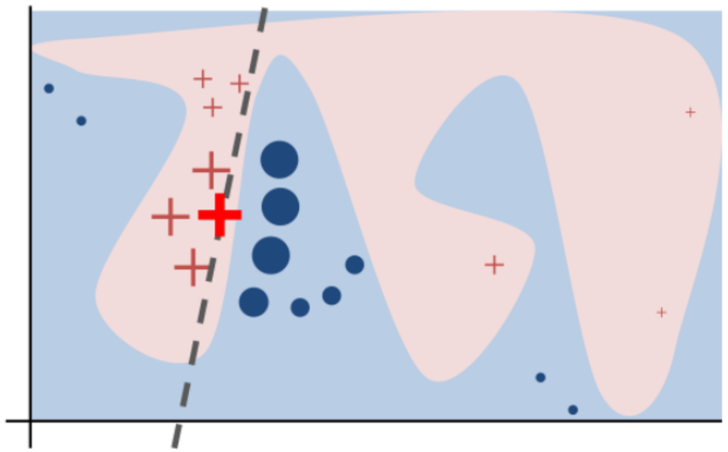


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Taken from Ribiero et al. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD 2016
https://dl.acm.org/doi/pdf/10.1145/2939672.2939778

Recap: The Form an Explanation Could Take

Potential Audiences

- Individual data subjects
- All data subjects
- Regulators / policymakers
- Third parties who receive explanations from individual data subjects (e.g., journalists)

- **Global** (model, all data subjects) vs. **Local** (one data subject)

Potential Information

- The general approach taken
- The precise factors taken into account
  - Do we provide access to the data subject's own values?
  - Do we explain how those factors are combined?
  - Do we explain the ML model used?
  - Do we open-source the ML model?
- Counterfactuals – "If *X* had not occurred, *Y* would not have occurred"
  - What could a person have changed for a different classification?
  - Can we define some distance function and show the *smallest* change?

See https://docs.seldon.io/projects/alibi/en/stable/methods/CF.html

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model (**Global**)	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	**How**
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	**How**, Why, Why not, What if
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	**How**, Why, Why not, What if
Explain a prediction (**Local**)	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	**Why**
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	**Why**, **How to still be this**
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	**What if**, How to be that, How to still be this
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	**Why**, **Why not**, **How to be that**
Example based	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	**Why**, How to still be this
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	**Why**, **Why not**, How to be that

Table 1. Taxonomy of XAI methods mapping to user question types. Questions in bold are the primary ones that the XAI method addresses. Questions in regular font are ones that only a subset of cases the XAI method can address. For example, while a global decision tree approximation can potentially answer *Why, Why not, and What if* questions for individual instances [58], the approximation may not cover certain instances.

Taken from Liao et al. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," CHI 2020.
https://dl.acm.org/doi/pdf/10.1145/3313831.3376590

Input

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

Figure 1. XAI Question Bank, with leading questions in bold, and new questions identified from the interviews with *

Taken from Liao et al. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," CHI 2020. https://dl.acm.org/doi/pdf/10.1145/3313831.3376590

Do we have a right to an explanation from automated decision-making systems?

Should we?