Explainable AI in Industry

KDD 2019 Tutorial

Sahin Cem Geyik, Krishnaram Kenthapadi & Varun Mithal

LinkedIn

Krishna Gade & Ankur Taly

fiddler

https://sites.google.com/view/kdd19-explainable-ai-tutorial

Agenda

- Motivation
- AI Explainability: **Foundations and Techniques**
  - Explainability concepts, problem formulations, and evaluation methods
  - Post hoc Explainability
  - Intrinsically Explainable models
- AI Explainability: **Industrial Practice**
  - Case Studies from LinkedIn, Fiddler Labs, and Google Research
- Demo
- Key Takeaways and Challenges

Motivation

Third Wave of AI

Symbolic AI

Logic rules represent knowledge

No learning capability and poor handling of uncertainty

Statistical AI

Statistical models for specific domains training on big data

No contextual capability and minimal explainability

Explainable AI

Systems construct explanatory models

Systems learn and reason with new tasks and situations

Factors driving rapid advancement of AI

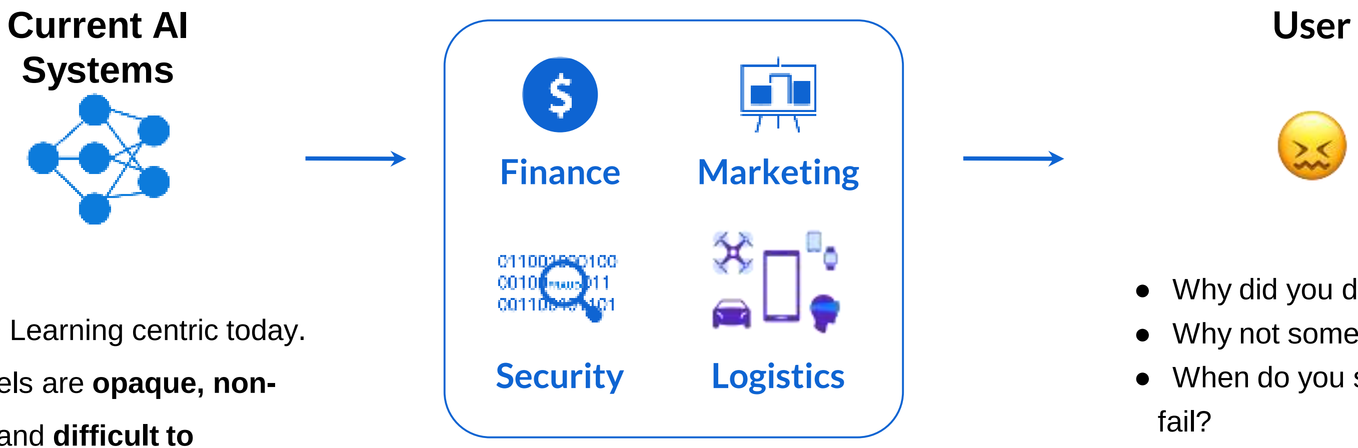
GPUs , On-chip Neural Network

Data Availability

Cloud Infrastructure

New Algorithms

Need for Explainable AI



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

- 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

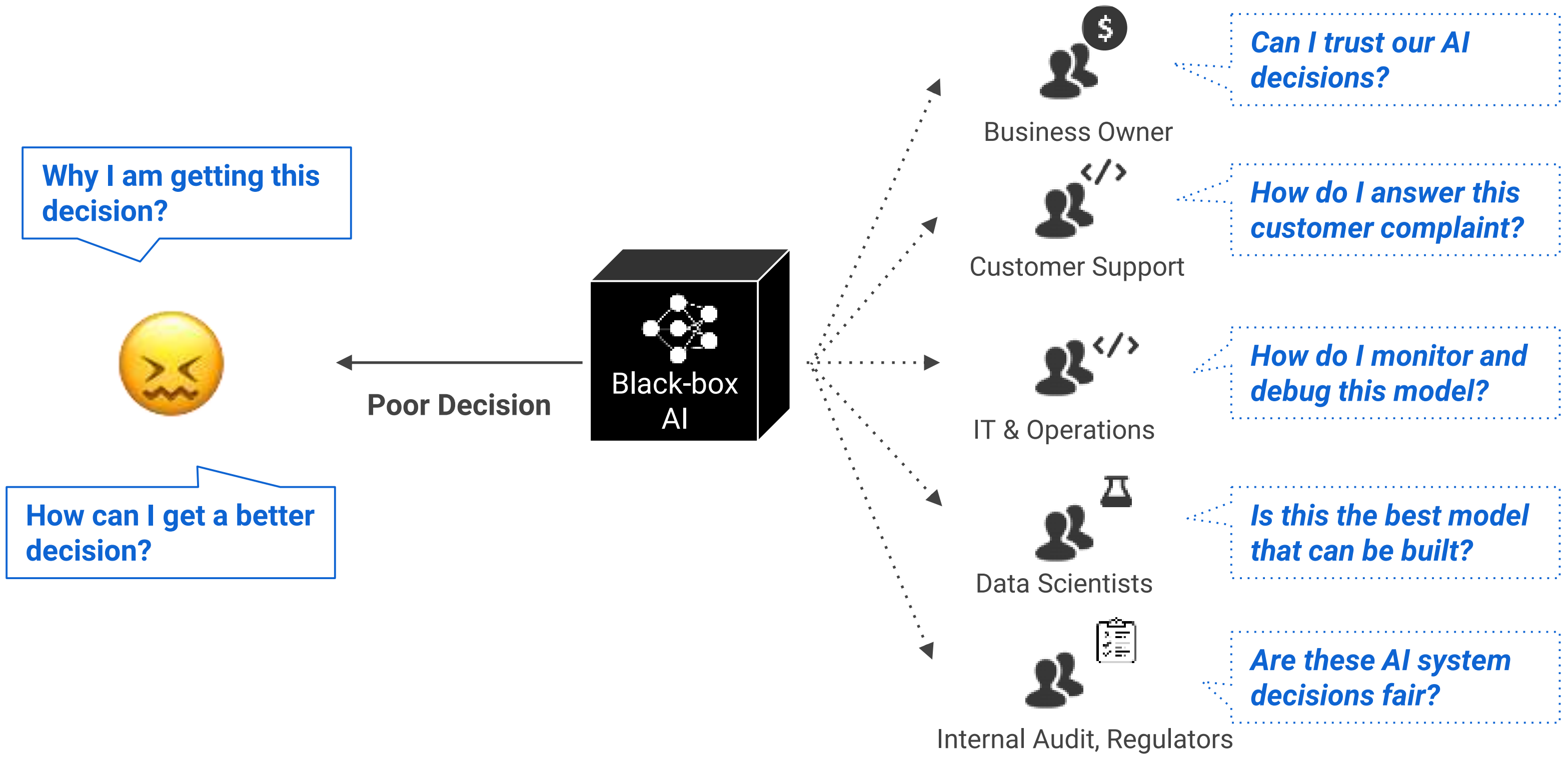
Congressional Leaders Press Zuckerberg On Political Bias

Apr 11, 2018

Guilty! AI Is Found to Perpetuate Biases in Jailing

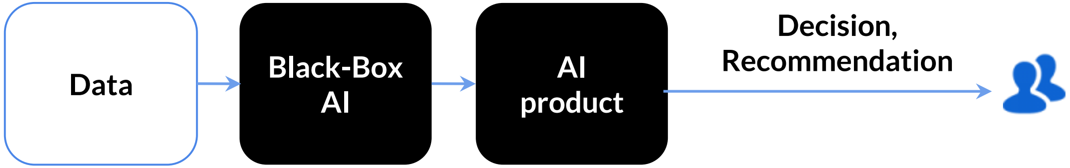
1 day ago

Black-box AI creates confusion and doubt



What is Explainable AI?

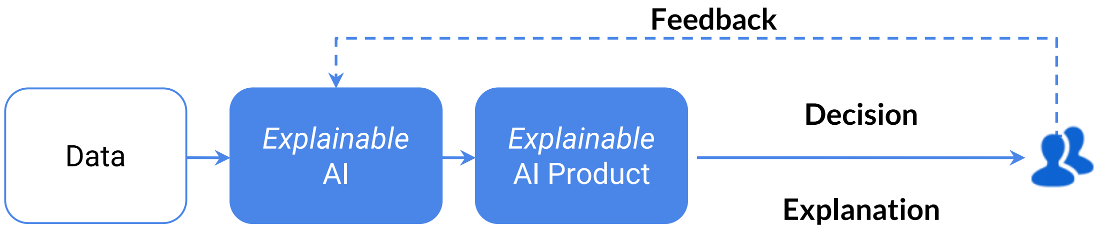
Black Box AI



Confusion with Today's AI Black Box

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

Explainable AI



Clear & Transparent Predictions

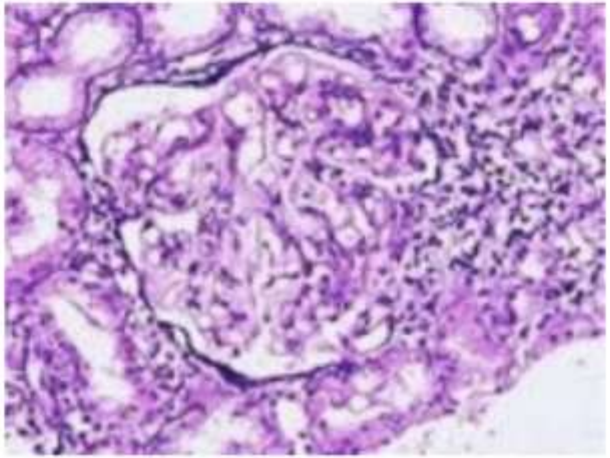
- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Why Explainability: Verify the ML Model / System

Wrong decisions can be costly
and dangerous

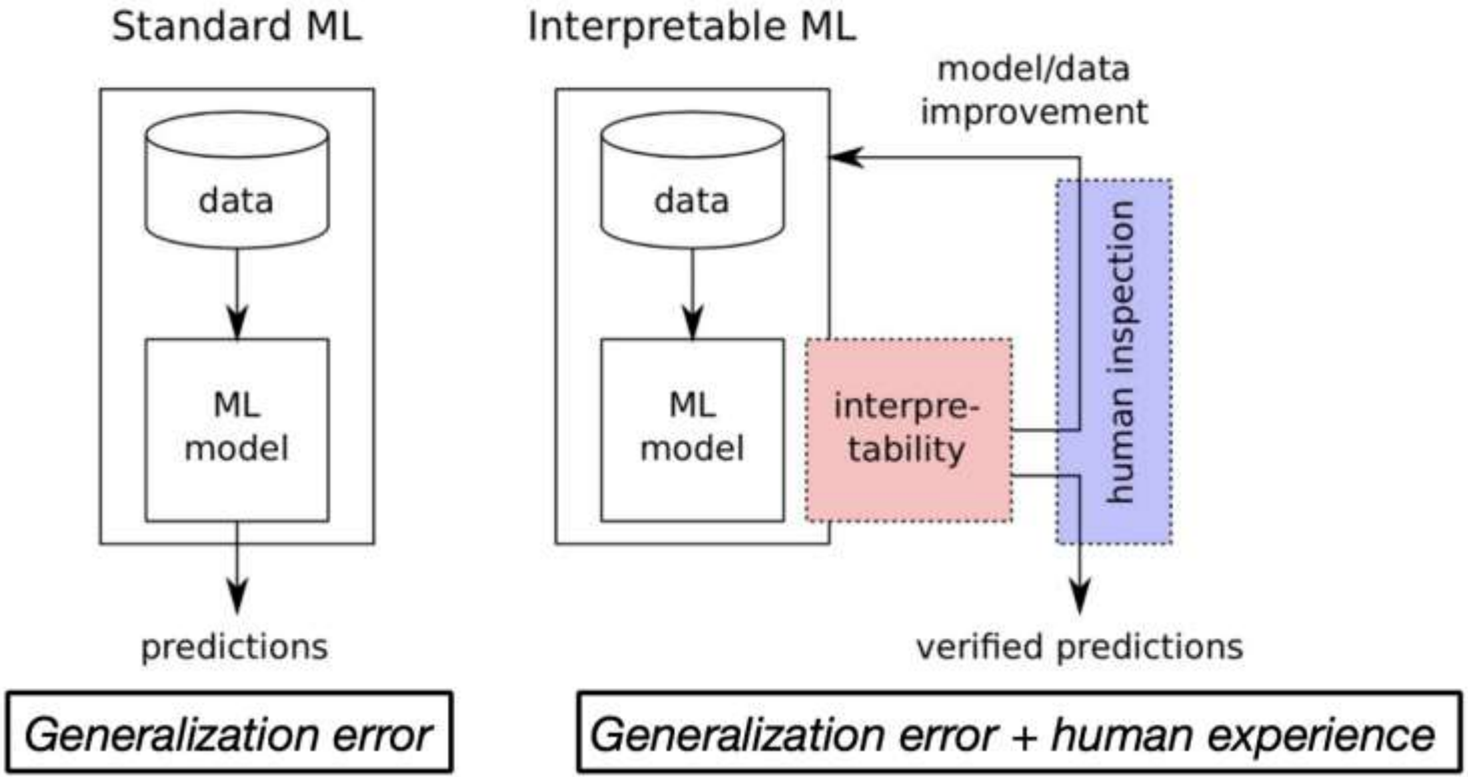
"Autonomous car crashes, because it wrongly recognizes ..."

"AI medical diagnosis system misclassifies patient's disease ..."



Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Improve ML Model

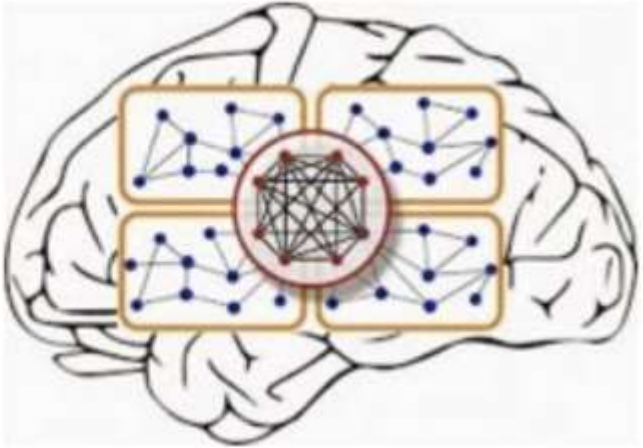


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Learn New Insights

"It's not a human move. I've never seen a human play this move." (Fan Hui)

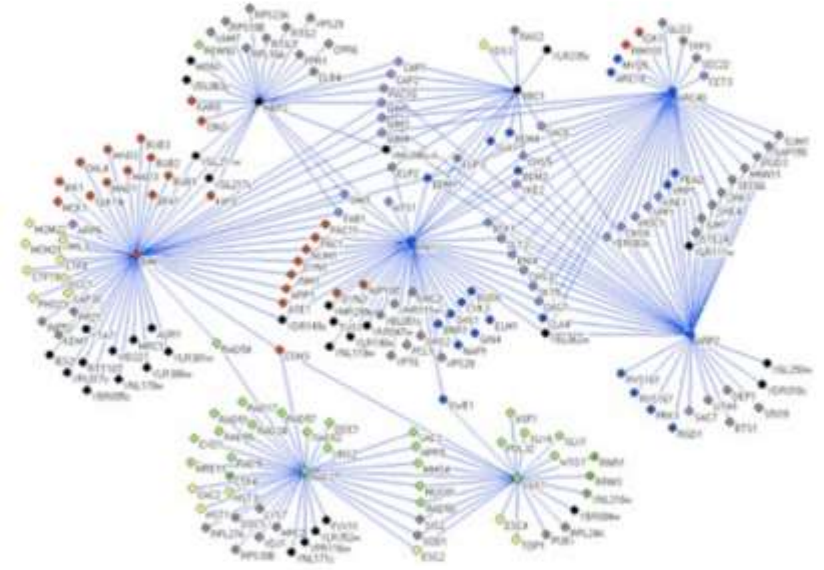
Old promise:
"Learn about the human brain."

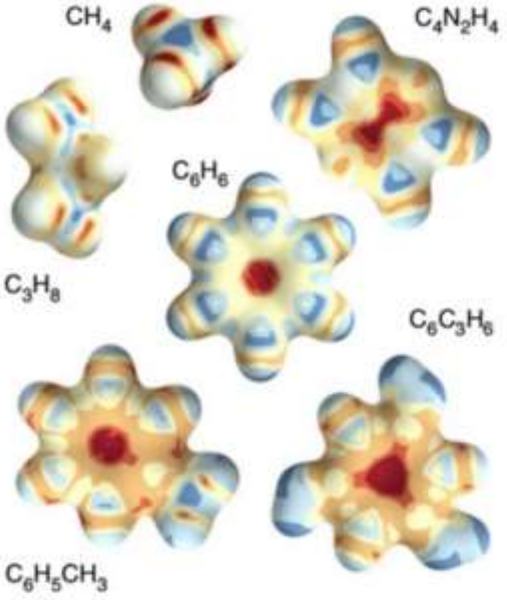


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Learn Insights in the Sciences

Learn about the physical / biological / chemical mechanisms.
(e.g. find genes linked to cancer, identify binding sites …)





Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI’18

Why Explainability: Debug (Mis-)Predictions

Top label: **"clog"**

Why did the network label this image as **"clog"**?

Why Explainability: Laws against Discrimination

Citizenship

Immigration Reform and Control Act

Sex

Equal Pay Act of 1963;
Civil Rights Act of 1964

Age

Age Discrimination in Employment Act of 1967

Race

Civil Rights Act of 1964

Disability status

Rehabilitation Act of 1973;
Americans with Disabilities Act of 1990

And more...

Fairness

Privacy

SR 11-7: Guidance on Model Risk Management

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

GDPR

CALIFORNIA
CONSUMER
PRIVACY
ACT OF 2018

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

VP, European Commision

Article 22. Automated individual decision making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
   (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Recital 71

Profiling*

Fai

cy

[1] The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. [2] Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. [3] However, decision-making based on such processing,

Transparency

Explainability

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
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**.

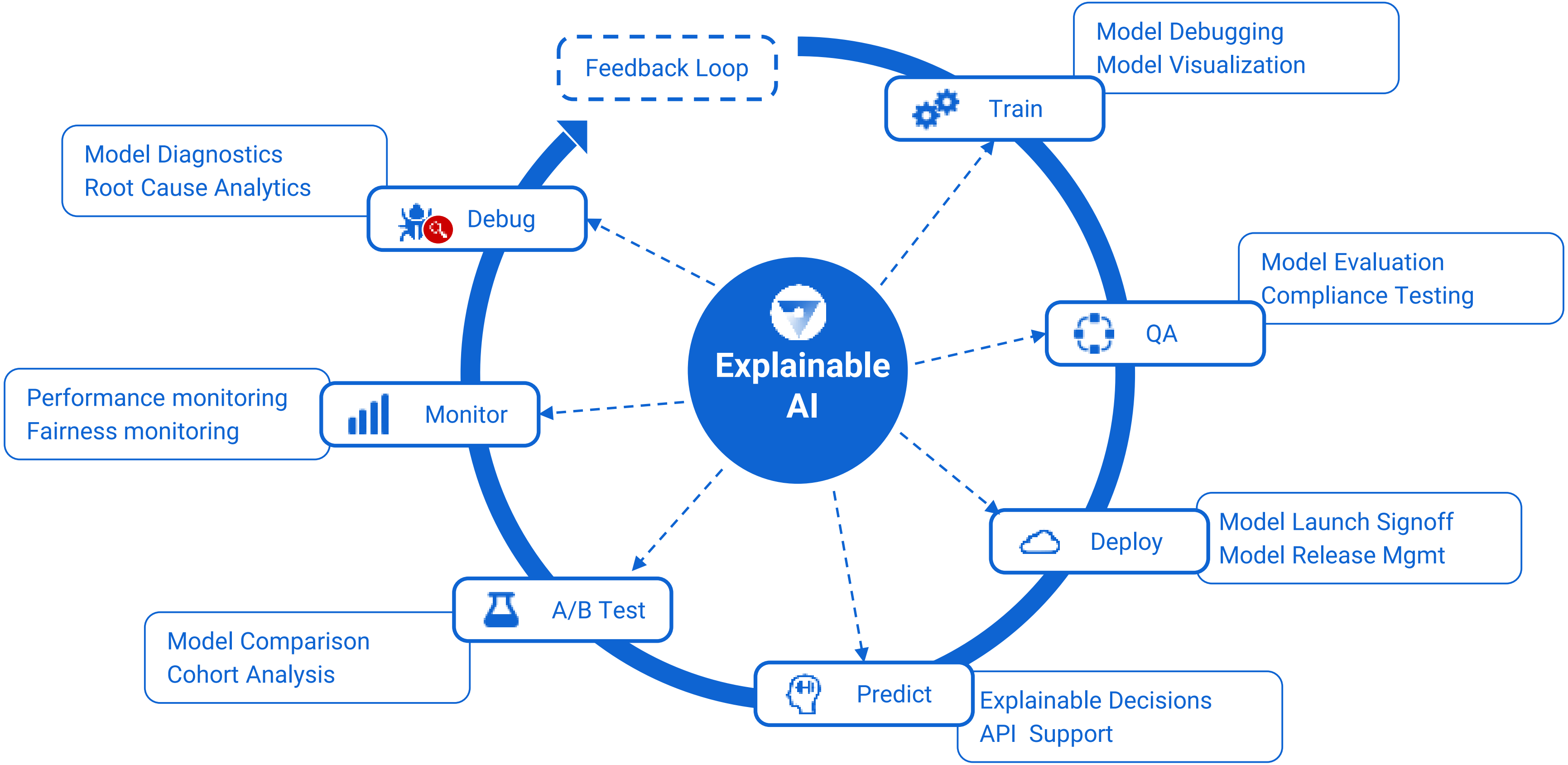
In addition **SR12-07** calls for **incorporating validation or other type of independent review of the stress testing framework to ensure the integrity of stress testing processes and results.**

JOHN HILL

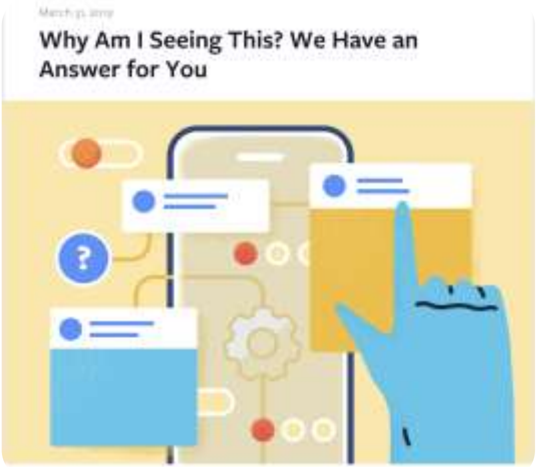
GLOBAL HEAD OF MODEL RISK GOVERNANCE, CREDIT SUISSE

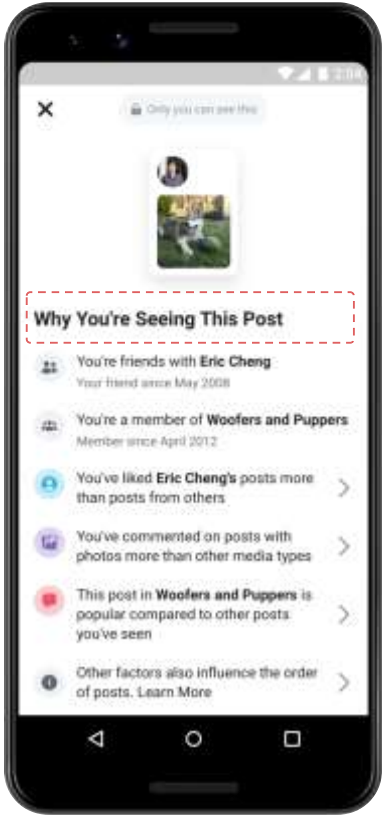
" In the current regulatory environment, model validation policies must be fully compliant with the requirements of SR11-7. While SR11-7 officially applies to US conforming bank and non-US banks doing business in the US, many European financial firms have adopted SR11-7 as their standard as well. "

“Explainability by Design” for AI products



Example: Facebook adds Explainable AI to build Trust





Foundations and Techniques

Achieving Explainable AI

Approach 1: **Post-hoc explain a given AI 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 model**

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

Achieving Explainable AI

Approach 1: **Post-hoc explain a given AI 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 model**

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

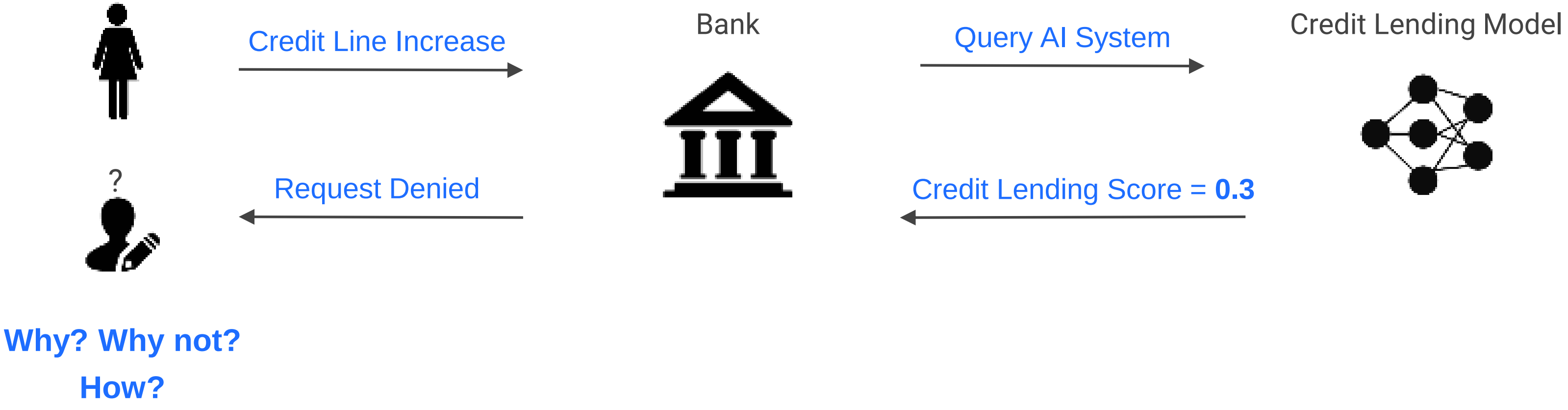
Top label: **"fireboat"**

Why did the network label this image as **"fireboat"**?

Top label: **"clog"**

Why did the network label this image as **"clog"**?

Credit Lending in a black-box ML world



Fair lending laws [ECOA, FCRA] require credit decisions to be explainable

The Attribution Problem

Attribute a model's prediction on an input to features of the input

Examples:

- Attribute an object recognition network's prediction to its pixels
- Attribute a text sentiment network's prediction to individual words
- Attribute a lending model's prediction to its features

A reductive formulation of "why this prediction" but surprisingly useful :-)

Application of Attributions

- Debugging model predictions
  E.g., Attribution an image misclassification to the pixels responsible for it
- Generating an explanation for the end-user
  E.g., Expose attributions for a lending prediction to the end-user
- Analyzing model robustness
  E.g., Craft adversarial examples using weaknesses surfaced by attributions
- Extract rules from the model
  E.g., Combine attribution to craft rules (pharmacophores) capturing prediction logic of a drug screening network

Next few slides

We will cover the following **attribution methods****

- Ablations
- Gradient based methods
- Score Backpropagation based methods
- Shapley Value based methods

**Not a complete list!
See Ancona et al. [ICML 2019], Guidotti et al. [arxiv 2018] for a comprehensive survey

Ablations

Drop each feature and attribute the change in prediction to that feature

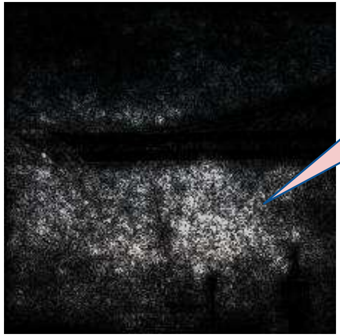
Useful tool but not a perfect attribution method. Why?

- Unrealistic inputs
- Improper accounting of interactive features
- Computationally expensive

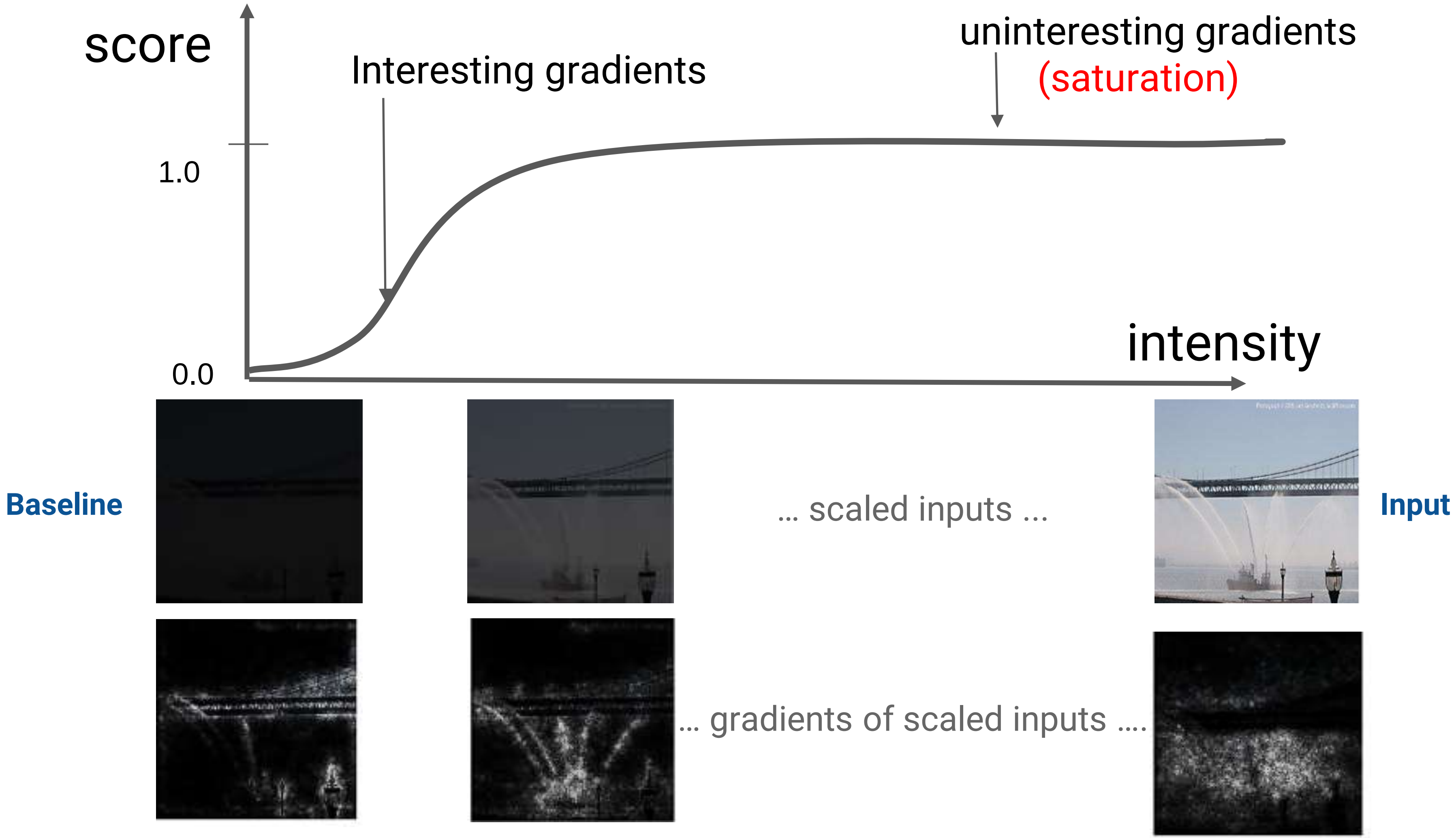
Feature*Gradient

Attribution to a feature is feature value times gradient, i.e., $x_i * \partial y/\partial x_i$

- Gradient captures sensitivity of output w.r.t. feature
- Equivalent to Feature*Coefficient for linear models
    - **First-order Taylor approximation** of non-linear models
- Popularized by SaliencyMaps [NIPS 2013], Baehrens et al. [JMLR 2010]



Gradients in the vicinity of the input seem like noise


score
Interesting gradients
uninteresting gradients
(saturation)
1.0
0.0
intensity
Baseline
... scaled inputs ...
Input
... gradients of scaled inputs ....

Integrated Gradients [ICML 2017]

Integrate the gradients along a **straight-line path from baseline to input**

$$\mathrm{IG}(\mathrm{input}, \mathrm{base}) ::= (\mathrm{input} - \mathrm{base}) * \int_{0\text{-}1} \nabla F(\alpha * \mathrm{input} + (1-\alpha) * \mathrm{base})\, d\alpha$$

Original image

Integrated Gradients

What is a baseline?

- Ideally, the baseline is an **informationless input for the model**
  - E.g., Black image for image models
  - E.g., Empty text or zero embedding vector for text models
- Integrated Gradients explains **F(input) - F(baseline)** in terms of input features

Aside: Baselines (or Norms) are essential to explanations [Kahneman-Miller 86]

- E.g., A man suffers from indigestion. Doctor blames it to a stomach ulcer. Wife blames it on eating turnips. Both are correct relative to their baselines.
- The baseline may also be an important analysis knob.

Why is this image labeled as “clog”?

Original image

“Clog”

Why is this image labeled as “clog”?

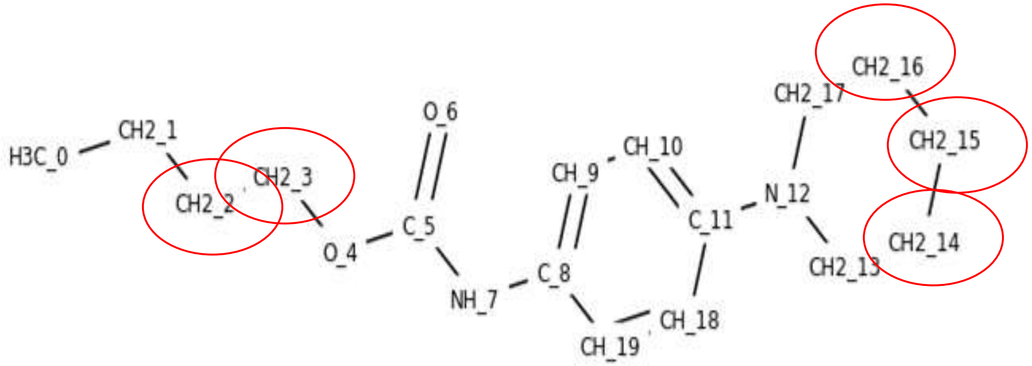
Original image

Integrated Gradients
(for label “clog”)

“Clog”

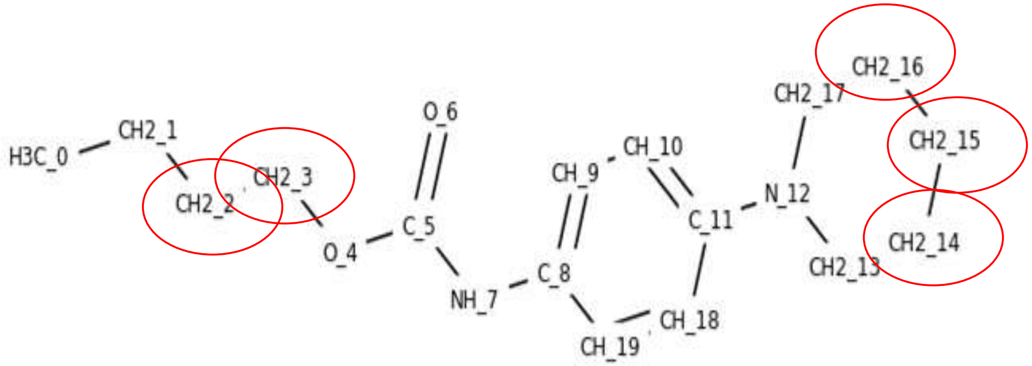
Detecting an architecture bug

- Deep network [Kearns, 2016] predicts if a molecule binds to certain DNA site
- **Finding**: Some atoms had identical attributions despite different connectivity



Detecting an architecture bug

- Deep network [Kearns, 2016] predicts if a molecule binds to certain DNA site
- **Finding**: Some atoms had identical attributions despite different connectivity



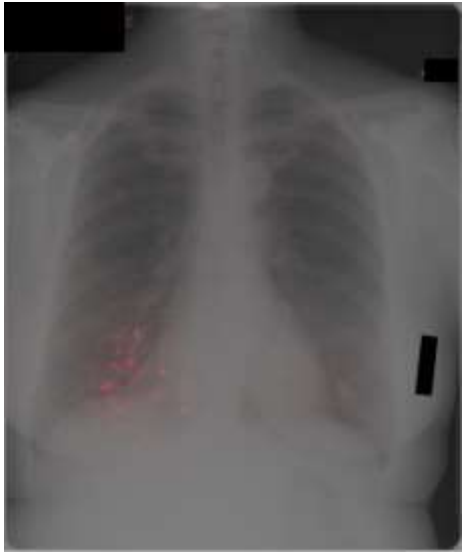
- **Bug**: The architecture had a bug due to which the convolved bond features did not affect the prediction!

Detecting a data issue

- Deep network predicts various diseases from chest x-rays

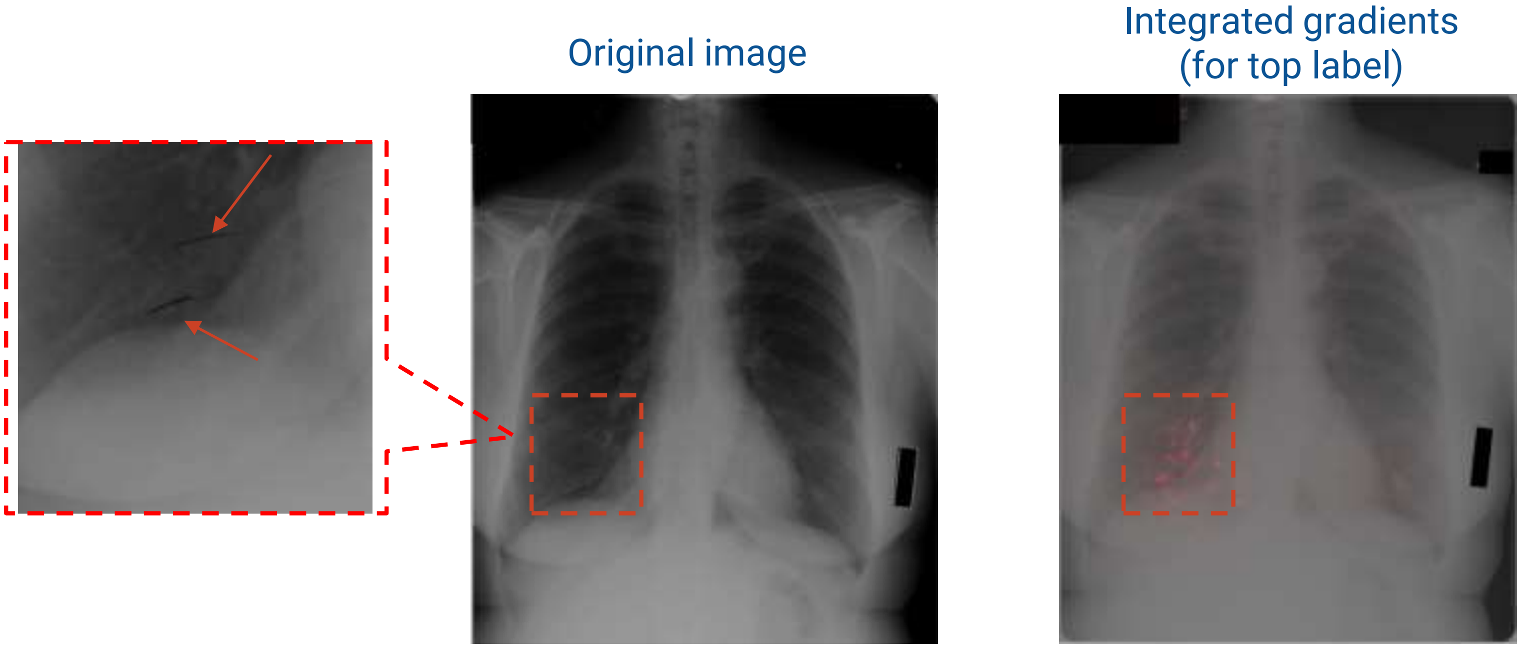
Original image

Integrated gradients
(for top label)



Detecting a data issue

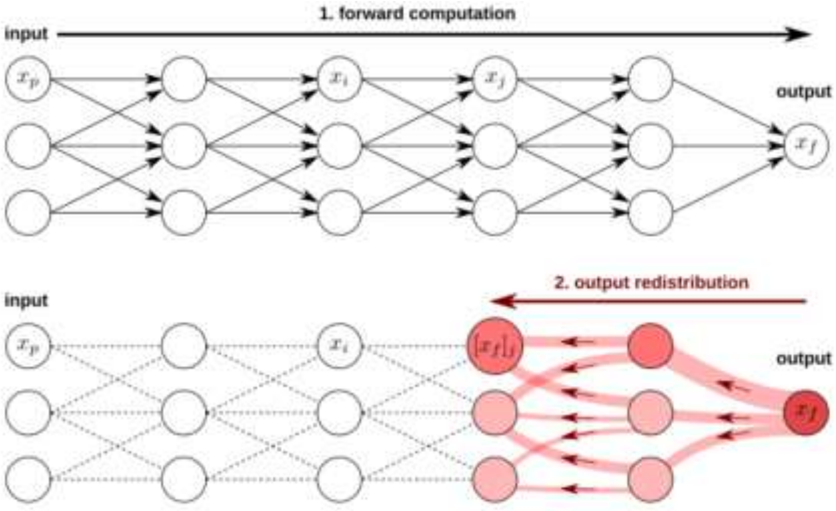
- Deep network predicts various diseases from chest x-rays
- **Finding**: Attributions fell on radiologist's markings (rather than the pathology)



Score Back-Propagation based Methods

Re-distribute the prediction score through the neurons in the network

- LRP [JMLR 2017], DeepLift [ICML 2017], Guided BackProp [ICLR 2014]



Easy case: Output of a neuron is a linear function of previous neurons (i.e., $n_i = \sum w_{ij} * n_j$)
e.g., the logit neuron

- Re-distribute the contribution in proportion to the coefficients w_{ij}

Image credit heatmapping.org

Score Back-Propagation based Methods

Re-distribute the prediction score through the neurons in the network

- LRP [JMLR 2017], DeepLift [ICML 2017], Guided BackProp [ICLR 2014]

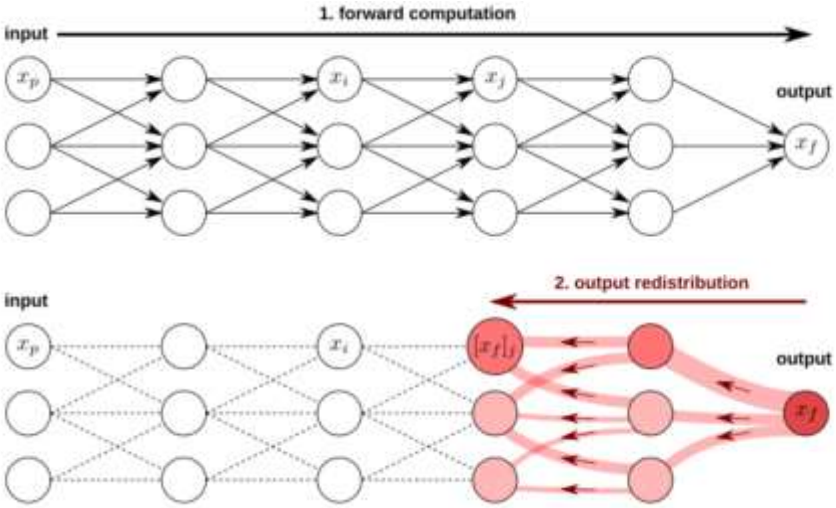


Image credit heatmapping.org

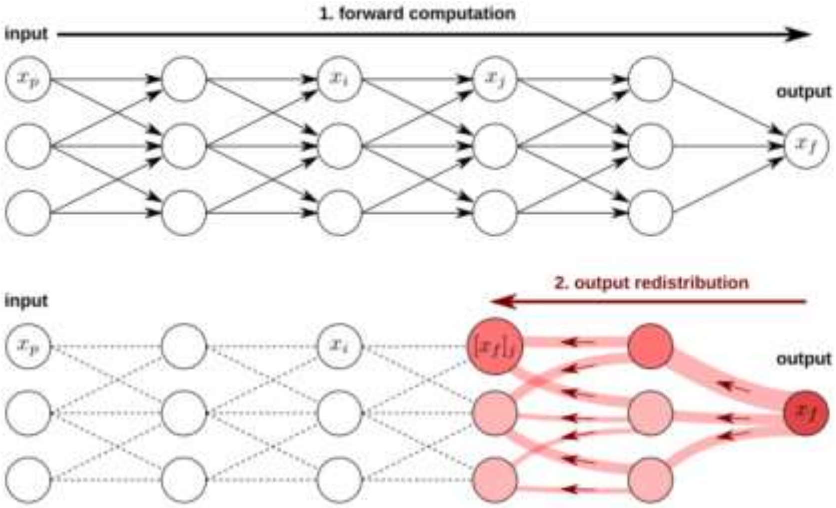
Tricky case: Output of a neuron is a **non-linear** function, e.g., ReLU, Sigmoid, etc.

- **Guided BackProp**: Only consider ReLUs that are on (linear regime), and which contribute positively
- **LRP**: Use first-order Taylor decomposition to linearize activation function
- **DeepLift**: Distribute activation difference relative a reference point in proportion to edge weights

Score Back-Propagation based Methods

Re-distribute the prediction score through the neurons in the network

- LRP [JMLR 2017], DeepLift [ICML 2017], Guided BackProp [ICLR 2014]



Pros:

- Conceptually simple
- Methods have been empirically validated to yield sensible result

Cons:

- Hard to implement, requires instrumenting the model
- **Often breaks implementation invariance**

 Think: F(x, y, z) = x * y *z and

 G(x, y, z) = x * (y * z)

Image credit heatmapping.org

So far we've looked at differentiable models.

But, what about **non-differentiable models**? E.g.,

- Decision trees
- Boosted trees
- Random forests
- etc.

Shapley Value [Annals of Mathematical studies,1953]

Classic result in game theory on distributing gain in a **coalition game**

- **Coalition Games**
  - Players collaborating to generate some **gain** (think: revenue)
  - Set function v(S) determining the gain for any subset S of players

Shapley Value [Annals of Mathematical studies,1953]

Classic result in game theory on distributing gain in a **coalition game**

- **Coalition Games**
  - Players collaborating to generate some **gain** (think: revenue)
  - Set function v(S) determining the gain for any subset S of players
- **Shapley Values** are a fair way to attribute the total gain to the players based on their contributions
  - <u>Concept</u>: **Marginal contribution** of a player to a subset of other players (v(S U {i}) - v(S))
  - Shapley value for a player is a **specific weighted aggregation of its marginal** over all possible subsets of other players

$$\textbf{Shapley Value for player i} = \sum_{S \subseteq N} w(S) * (v(S \cup \{i\}) - v(S))$$

(where $w(S) = N! / |S|! (N - |S| -1)!$)

Shapley Value Justification

Shapley values are unique under four simple axioms

- **Dummy:** If a player never contributes to the game then it must receive zero attribution
- **Efficiency:** Attributions must add to the total gain
- **Symmetry:** Symmetric players must receive equal attribution
- **Linearity:** Attribution for the (weighted) sum of two games must be the same as the (weighted) sum of the attributions for each of the games

Shapley Values for Explaining ML models

SHAP [NeurIPS 2018], QII [S&P 2016], Strumbelj & Konenko [JMLR 2009]

- Define a coalition game for each model input X
  - **Players are the features in the input**
  - **Gain is the model prediction** (output), i.e., gain = F(X)
- Feature attributions are the Shapley values of this game

Shapley Values for Explaining ML models

SHAP [NeurIPS 2018], QII [S&P 2016], Strumbelj & Konenko [JMLR 2009]

- Define a coalition game for each model input X
  - **Players are the features in the input**
  - **Gain is the model prediction** (output), i.e., gain = F(X)
- Feature attributions are the Shapley values of this game

Challenge: Shapley Values require the gain to be defined for all subsets of players

- What is the prediction when **some players (features) are absent**?

  i.e., what is **F(x_1, <absent>, x_3, …, <absent>)**?

Modeling Feature Absence

Key Idea: Take the expected prediction when the (absent) feature is sampled from a certain distribution.

Different approaches choose different distributions

- [SHAP, NIPS 2018] Use conditional distribution w.r.t. the present features
- [QII, S&P 2016] Use marginal distribution
- [Strumbelj et al., JMLR 2009] Use uniform distribution
- [Integrated Gradients, ICML 2017] Use a specific baseline point

Computing Shapley Values

Exact Shapley value computation is exponential in the number of features

- Shapley values can be expressed as an expectation of marginals

$$\boldsymbol{\phi(i) = E_{S \sim \mathcal{D}}\ [\mathrm{marginal}(S, i)]}$$

  - Sampling-based methods can be used to approximate the expectation
  - See: “Computational Aspects of Cooperative Game Theory”, Chalkiadakis et al. 2011
- The method is still computationally infeasible for models with hundreds of features, e.g., image models

Evaluating Attribution Methods

Human Review

Have humans review attributions and/or compare them to (human provided) groundtruth on “feature importance”

Pros:

- Helps assess if attributions are human-intelligible
- Helps increase trust in the attribution method

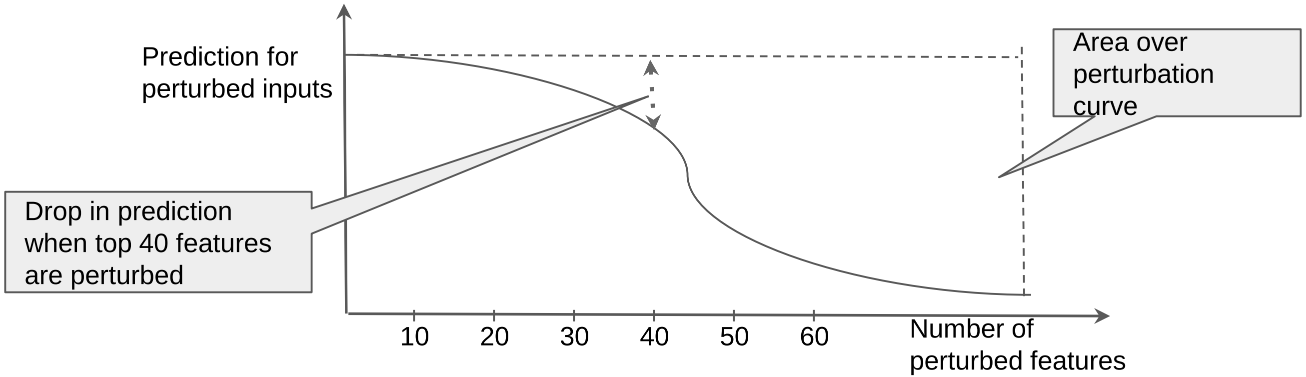
Cons:

- Attributions may appear incorrect because model reasons differently
- **Confirmation bias**

Perturbations (Samek et al., IEEE NN and LS 2017)

Perturb top-k features by attribution and observe change in prediction

- Higher the change, better the method
- Perturbation may amount to replacing the feature with a random value
- Samek et al. formalize this using a metric: **Area over perturbation curve**
    - Plot the prediction for input with top-k features perturbed as a function of k
    - Take the area over this curve



Axiomatic Justification

Inspired by how Shapley Values are justified

- List **desirable criteria (axioms)** for an attribution method
- Establish a uniqueness result: X is the **only** method that satisfies these criteria

Integrated Gradients, SHAP, QII, Strumbelj & Konenko are justified in this manner

Theorem [Integrated Gradients, ICML 2017]: Integrated Gradients is the **unique** path-integral method satisfying: **Sensitivity, Insensitivity, Linearity preservation, Implementation invariance, Completeness, and Symmetry**

Some limitations and caveats

Attributions are pretty shallow

Attributions do not explain:

- Feature interactions
- What training examples influenced the prediction
- Global properties of the model

An instance where attributions are useless:

- A model that predicts TRUE when there are **even number** of black pixels and FALSE otherwise

Attributions are for human consumption

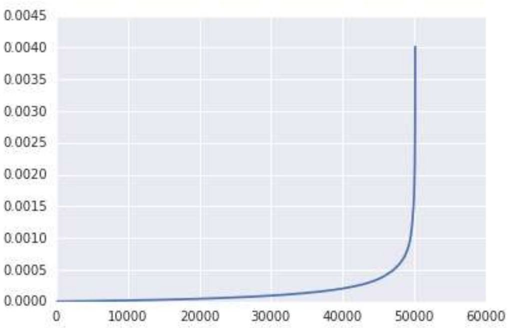
- **Humans** interpret attributions and generate insights
  - Doctor maps attributions for x-rays to pathologies
- **Visualization** matters as much as the attribution technique

Attributions are for human consumption

- **Humans** interpret attributions and generate insights
  - Doctor maps attributions for x-rays to pathologies
- **Visualization** matters as much as the attribution technique

Naive scaling of attributions from 0 to 255

Attributions have a **large range** and **long tail** across pixels



After clipping attributions at 99% to reduce range

Other types of Post-hoc Explanations

Example based Explanations

Learned prototypes and criticisms from Imagenet dataset (two types of dog breeds)

- Prototypes: Representative of all the training data.
- Criticisms: Data instance that is not well represented by the set of prototypes.

Figure credit: Examples are not Enough, Learn to Criticize! Criticism for Interpretability. Kim, Khanna and Koyejo. NIPS 2016

Influence functions

- Trace a model’s prediction through the learning algorithm and back to its training data
- Training points “responsible” for a given prediction

Figure credit: Understanding Black-box Predictions via Influence Functions. Koh and Liang ICML 2017

Local Interpretable Model-agnostic Explanations

(Ribeiro et al. KDD 2016)

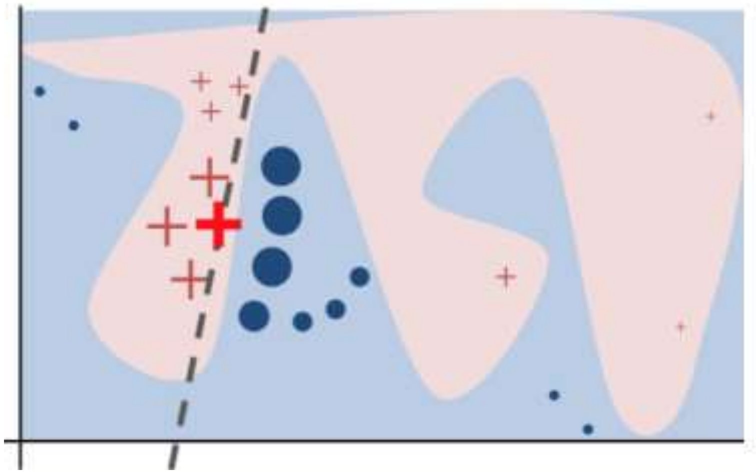


Figure credit: Ribeiro et al. KDD 2016

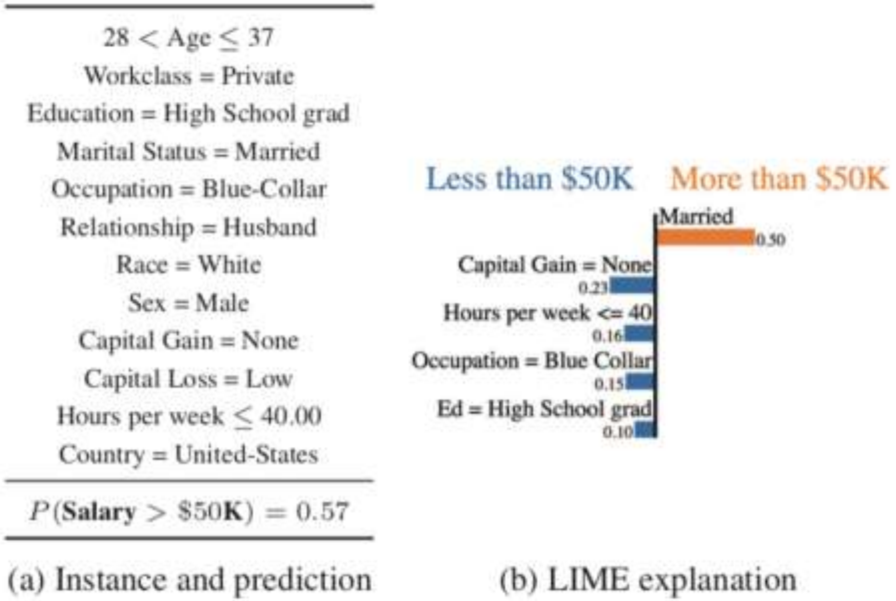


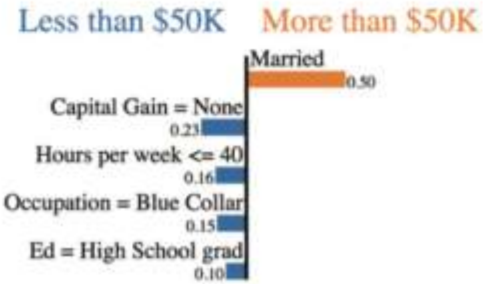
Figure credit: Anchors: High-Precision Model-Agnostic Explanations. Ribeiro et al. AAAI 2018

Anchors

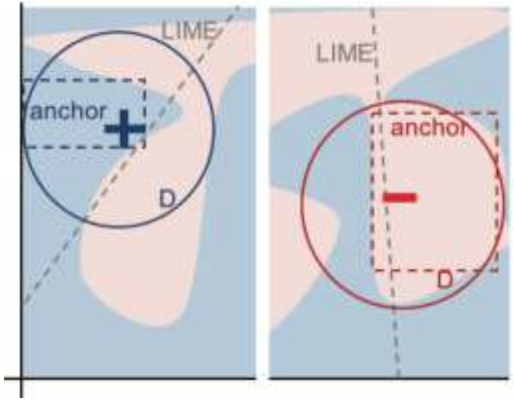
28 < Age ≤ 37
Workclass = Private
Education = High School grad
Marital Status = Married
Occupation = Blue-Collar
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = Low
Hours per week ≤ 40.00
Country = United-States

$P(\textbf{Salary} > \$50\textbf{K}) = 0.57$

(a) Instance and prediction



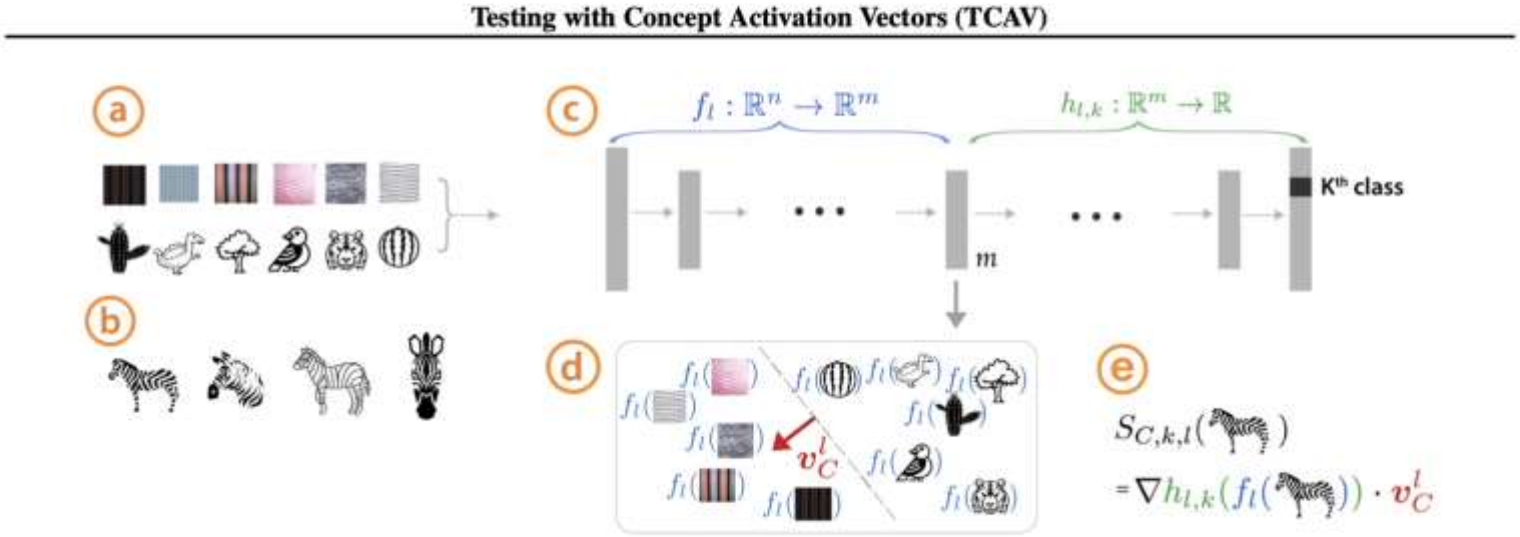
(b) LIME explanation



IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** 28 < Age ≤ 37
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > $50K

(c) An *anchor* explanation

Figure credit: Anchors: High-Precision Model-Agnostic Explanations. Ribeiro et al. AAAI 2018



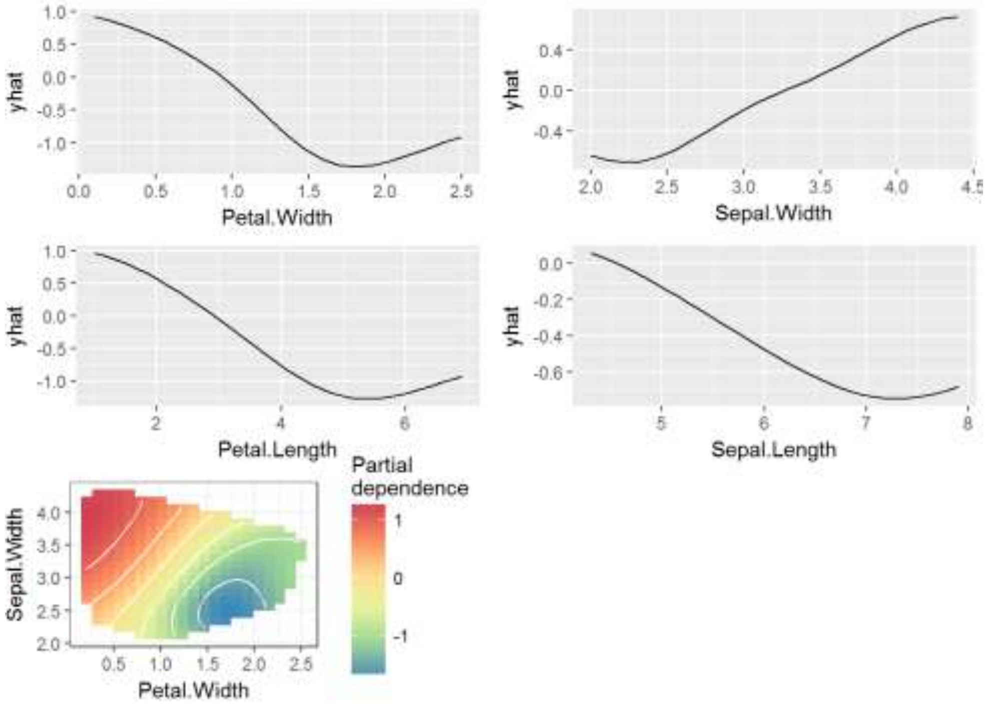
. **Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., 'striped'), and random examples ⓐ, labeled training-data examples for the studied class (zebras) ⓑ, and a trained network ⓒ, TCAV can quantify the model's sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept's examples and examples in any layer ⓓ. The CAV is the vector orthogonal to the classification boundary (v_C^l, red arrow). For the class of interest (zebras), TCAV uses the directional derivative $S_{C,k,l}(\boldsymbol{x})$ to quantify conceptual sensitivity ⓔ.

Figure credit: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) Kim et al. 2018

Global Explanations

Global Explanations Methods

- Partial Dependence Plot: Shows the marginal effect one or two features have on the predicted outcome of a machine learning model



Global Explanations Methods

- Permutations: The importance of a feature is the increase in the prediction error of the model after we permuted the feature’s values, which breaks the relationship between the feature and the true outcome.

	RDSpend	Administration	MarketingSpend	Profit	state_California
1	165349.2	136897.8	471784.1	192261.83	0
2	162597.7	151377.59	443898.53	191792.06	1
3	153441.51	101145.55	407934.54	191050.39	1
...	...	...	...	...	...
48	0	135426.92	0	42559.73	1
49	542.05	51743.15	0	35673.41	0
50	0	116983.8	45173.06	14681.4	1

Random Shuffle of the first feature

Achieving Explainable AI

Approach 1: **Post-hoc explain a given AI model**

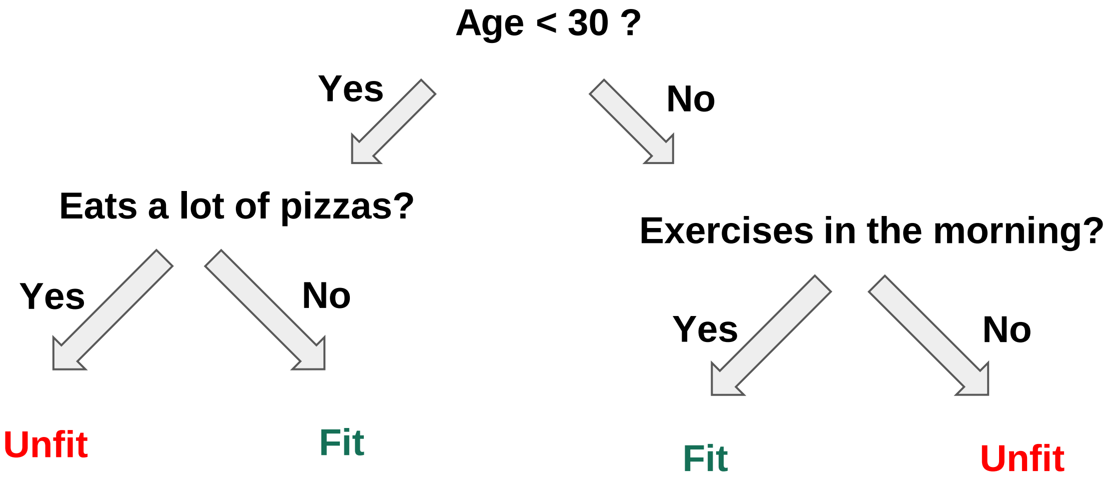
- **Individual prediction explanations** in terms of input features, influential examples, concepts, local decision rules
- **Global prediction explanations** in terms of entire model in terms of partial dependence plots, global feature importance, global decision rules

Approach 2: **Build an interpretable model**

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

Decision Trees

Is the person fit?



Decision List

If Past-Respiratory-Illness $=$Yes **and** Smoker $=$Yes **and** Age $\geq$ 50, **then** Lung Cancer

Else if Allergies $=$Yes **and** Past-Respiratory-Illness $=$Yes, **then** Asthma

Else if Family-Risk-Respiratory $=$Yes, **then** Asthma

Else if Family-Risk-Depression $=$Yes, **then** Depression

Else if Gender $=$Female **and** Short-Breath-Symptoms $=$Yes, **then** Asthma

Else if BMI $\geq$ 0.2 **and** Age$\geq$ 60, **then** Diabetes

Else if Frequent-Headaches $=$Yes **and** Dizziness $=$Yes, **then** Depression

Else if Frequency-Doctor-Visits $\geq$ 0.3, **then** Diabetes

Else if Disposition-Tiredness $=$Yes, **then** Depression

Else if Chest-Pain $=$Yes **and** Nausea **and** Yes, **then** Diabetes

Else Diabetes

Figure credit: Interpretable Decision Sets: A Joint Framework for Description and Prediction, Lakkaraju, Bach, Leskovec

Decision Set

If Allergies =Yes **and** Smoker =Yes **and** Irregular-Heartbeat =Yes, **then** Asthma

If Allergies =Yes **and** Past-Respiratory-Illness =Yes **and** Avg-Body-Temperature $\geq$ 0.1, **then** Asthma

If Smoker =Yes **and** BMI $\geq$ 0.2 **and** Age $\geq$ 60, **then** Diabetes

If Family-Risk-Diabetes =Yes **and** BMI $\geq$ 0.4 =Frequency-Infections $\geq$ 0.2, **then** Diabetes

If Frequency-Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity =Yes **and** Past-Respiratory-Illness =Yes, **then** Diabetes

If Family-Risk-Depression =Yes **and** Past-Depression =Yes **and** Gender =Female, **then** Depression

If BMI $\geq$ 0.3 **and** Insurance-Coverage =None **and** Avg-Blood-Pressure $\geq$ 0.2, **then** Depression

If Past-Respiratory-Illness =Yes **and** Age $\geq$ 50 **and** Smoker =Yes, **then** Lung Cancer

If Family-Risk-LungCancer =Yes **and** Allergies =Yes **and** Avg-Blood-Pressure $\geq$ 0.3, **then** Lung Cancer

If Disposition-Tiredness =Yes **and** Past-Anemia =Yes **and** BMI $\geq$ 0.3 **and** Rapid-Weight-Loss =Yes, **then** Leukemia

If Family-Risk-Leukemia =Yes **and** Past-Blood-Clotting =Yes **and** Frequency-Doctor-Visits $\geq$ 0.3, **then** Leukemia

If Disposition-Tiredness =Yes **and** Irregular-Heartbeat =Yes **and** Short-Breath-Symptoms =Yes **and** Abdomen-Pains =Yes, **then** Myelofibrosis

Figure credit: Interpretable Decision Sets: A Joint Framework for Description and Prediction, Lakkaraju, Bach, Leskovec

GLMs and GAMs

Model	Form	Intelligibility	Accuracy
Linear Model	$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Generalized Linear Model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Additive Model	$y = f_1(x_1) + ... + f_n(x_n)$	++	++
Generalized Additive Model	$g(y) = f_1(x_1) + ... + f_n(x_n)$	++	++
Full Complexity Model	$y = f(x_1, ..., x_n)$	+	+++

Intelligible Models for Classification and Regression. Lou, Caruana and Gehrke KDD 2012

Accurate Intelligible Models with Pairwise Interactions. Lou, Caruana, Gehrke and Hooker. KDD 2013

Case Studies from Industry

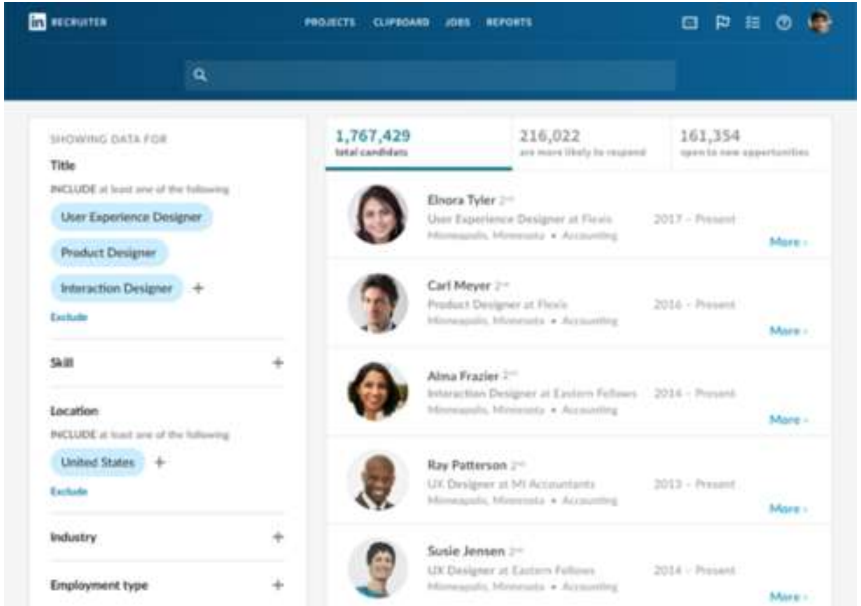
Case Study:

LinkedIn Talent Search

Varun Mithal, Girish Kathalagiri, Sahin Cem Geyik

LinkedIn Recruiter

- Recruiter Searches for Candidates
  - Standardized and free-text search criteria
- Retrieval and Ranking
  - Filter candidates using the criteria
  - Rank candidates in multiple levels using ML models

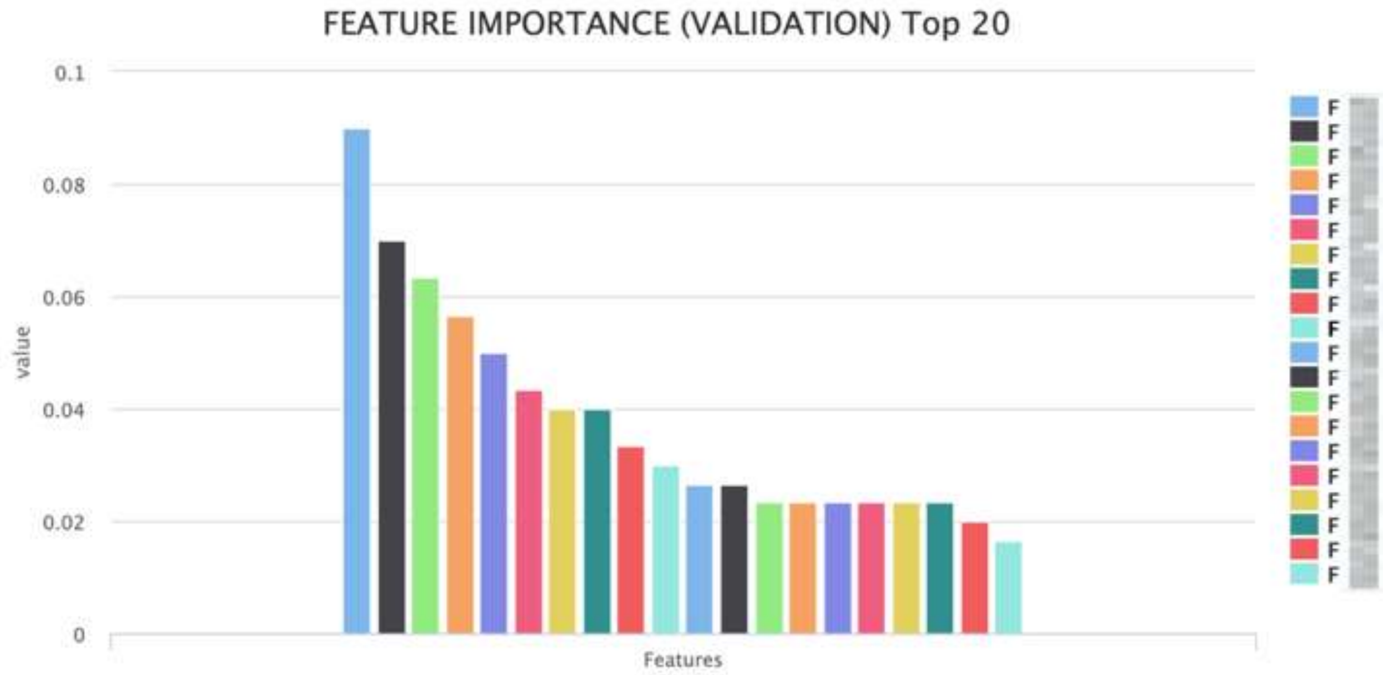


Modeling Approaches

- Pairwise XGBoost
- GLMix
- DNNs via TensorFlow

- Optimization Criteria: inMail Accepts
  - Positive: inMail sent by recruiter, and positively responded by candidate
    - Mutual interest between the recruiter and the candidate

Feature Importance in XGBoost



How We Utilize Feature Importances for GBDT

- Understanding feature digressions
    - Which a feature that was impactful no longer is?
    - Should we debug feature generation?
- Introducing new features in bulk and identifying effective ones
    - An activity feature for last 3 hours, 6 hours, 12 hours, 24 hours introduced (costly to compute)
    - Should we keep all such features?
- Separating the factors for that caused an improvement
    - Did an improvement come from a new feature, or a new labeling strategy, data source?
    - Did the ordering between features change?
- Shortcoming: A global view, not case by case

GLMix Models

- Generalized Linear Mixed Models
  - Global: Linear Model
  - Per-contract: Linear Model
  - Per-recruiter: Linear Model

$$g(\underbrace{P(r,c,re,ca,co)}_{\text{Positive Response Prob.}}) = \underbrace{\beta_{global} \cdot f_{all}}_{\text{Global model}} + \underbrace{\beta_{re} \cdot f_{all}}_{\text{Per-recruiter model}} + \underbrace{\beta_{co} \cdot f_{all}}_{\text{Per-contract model}}$$

- Lots of parameters overall
  - For a specific recruiter or contract the weights can be summed up
- Inherently explainable
  - Contribution of a feature is “weight x feature value”
  - Can be examined in a case-by-case manner as well

TensorFlow Models in Recruiter and Explaining Them

- We utilize the Integrated Gradients [ICML 2017] method

- How do we determine the baseline example?
  - Every query creates its own feature values for the same candidate
  - Query match features, time-based features
  - Recruiter affinity, and candidate affinity features
  - A candidate would be scored differently by each query
  - Cannot recommend a “Software Engineer” to a search for a “Forensic Chemist”
  - There is no globally neutral example for comparison!

Query-Specific Baseline Selection

- For each query:
  - Score examples by the TF model
  - Rank examples
  - Choose one example as the baseline
  - Compare others to the baseline example
- How to choose the baseline example
  - Last candidate
  - Kth percentile in ranking
  - A random candidate
  - Request by user (answering a question like: “Why was I presented candidate x above candidate y?”)

Example

Example - Detailed

Feature	Description	Difference (1 vs 2)	Contribution
Feature……….	Description……….	-2.0476928	-2.144455602
Feature……….	Description……….	-2.3223877	1.903594618
Feature……….	Description……….	0.11666667	0.2114946752
Feature……….	Description……….	-2.1442587	0.2060414469
Feature……….	Description……….	-14	0.1215354111
Feature……….	Description……….	1	0.1000282466
Feature……….	Description……….	-92	-0.085286277
Feature……….	Description……….	0.9333333	0.0568533262
Feature……….	Description……….	-1	-0.051796317
Feature……….	Description……….	-1	-0.050895940

Pros & Cons

- Explains potentially very complex models
- Case-by-case analysis
  - Why do you think candidate x is a better match for my position?
  - Why do you think I am a better fit for this job?
  - Why am I being shown this ad?
  - Great for debugging real-time problems in production
- Global view is missing
  - Aggregate Contributions can be computed
  - Could be costly to compute

Lessons Learned and Next Steps

- Global explanations vs. Case-by-case Explanations
  - Global gives an overview, better for making modeling decisions
  - Case-by-case could be more useful for the non-technical user, better for debugging
- Integrated gradients worked well for us
  - Complex models make it harder for developers to map improvement to effort
  - Use-case gave intuitive results, on top of completely describing score differences
- Next steps
  - Global explanations for Deep Models

Case Study:

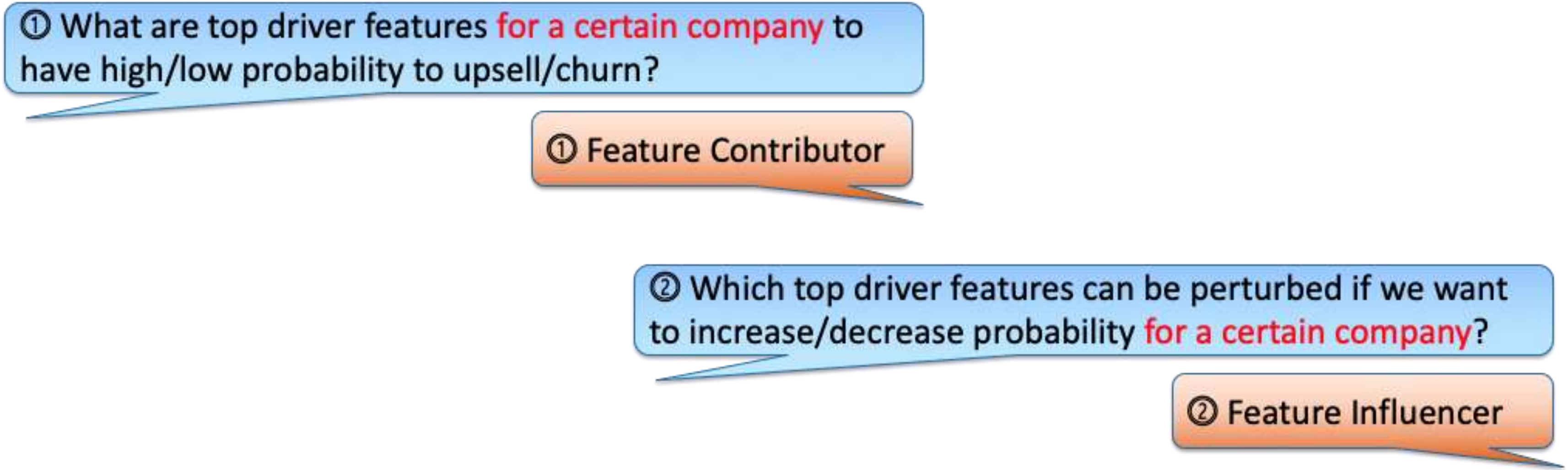
Model Interpretation for Predictive Models in B2B Sales Predictions

Jilei Yang, Wei Di, Songtao Guo

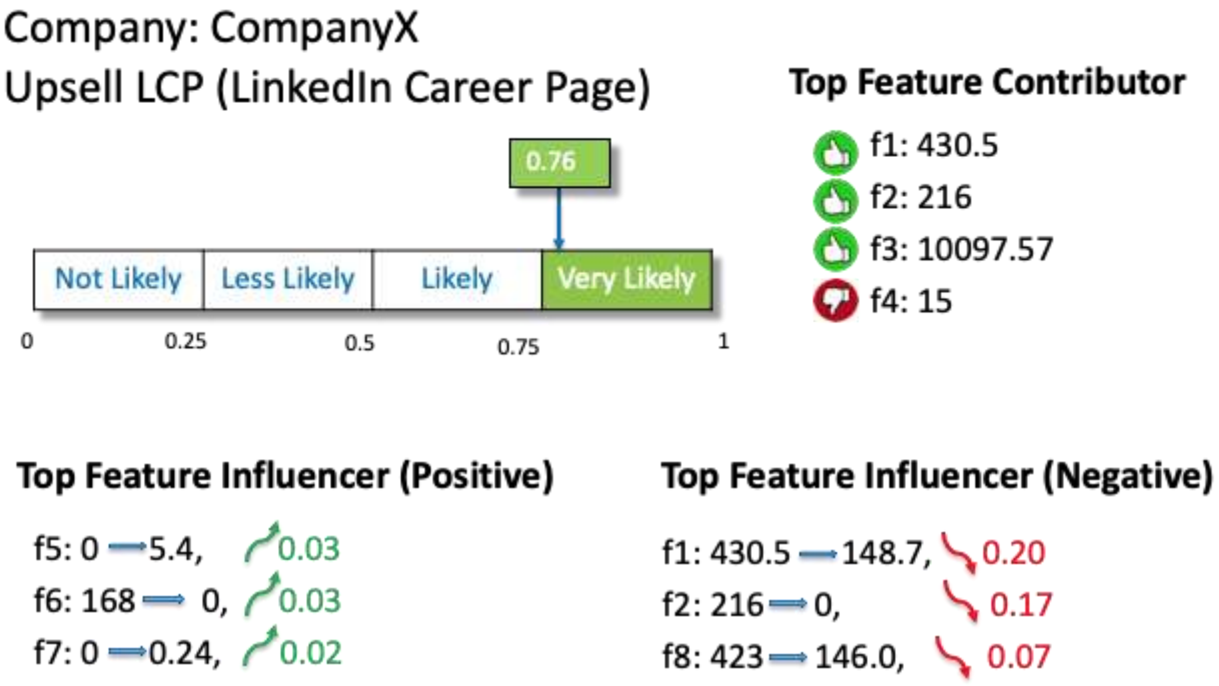
LinkedIn

Problem Setting

- Predictive models in B2B sales prediction
  - E.g.: random forest, gradient boosting, deep neural network, …
  - High accuracy, low interpretability
- Global feature importance → Individual feature reasoning



Example


Company: CompanyX
Upsell LCP (LinkedIn Career Page)
0.76
Not Likely
Less Likely
Likely
Very Likely
0
0.25
0.5
0.75
1
Top Feature Contributor
f1: 430.5
f2: 216
f3: 10097.57
f4: 15
Top Feature Influencer (Positive)
f5: 0 → 5.4, 0.03
f6: 168 → 0, 0.03
f7: 0 → 0.24, 0.02
Top Feature Influencer (Negative)
f1: 430.5 → 148.7, 0.20
f2: 216 → 0, 0.17
f8: 423 → 146.0, 0.07

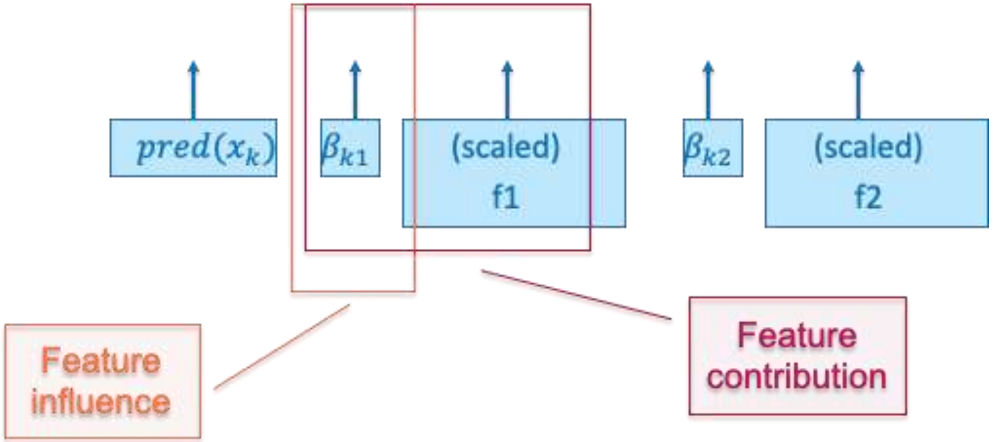
Revisiting LIME

- Given a target sample x_k, approximate its prediction $pred(x_k)$ by building a sample-specific linear model:

$$pred(X) \approx \beta_{k1} X_1 + \beta_{k2} X_2 + \ldots, X \in neighbor(x_k)$$

- E.g., for company CompanyX:

$$0.76 \approx 1.82 * 0.17 \quad + 1.61 * 0.11 + \ldots$$



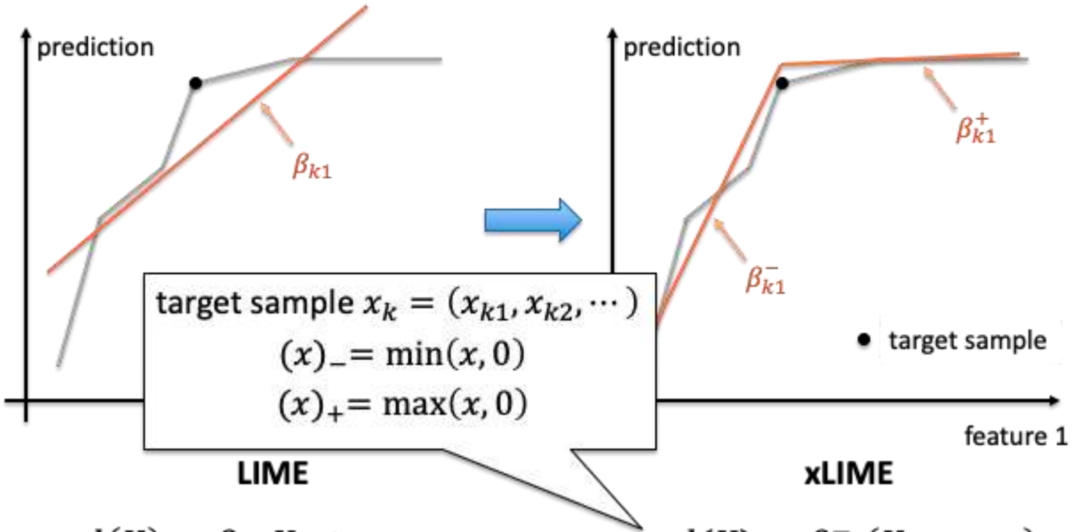
xLIME

Piecewise Linear Regression

Localized Stratified Sampling

Piecewise Linear Regression

Motivation: Separate top positive feature influencers and top negative feature influencers



$$pred(X) \approx \beta_{k1} X_1 + \cdots,\quad X \in neighbor(x_k)$$

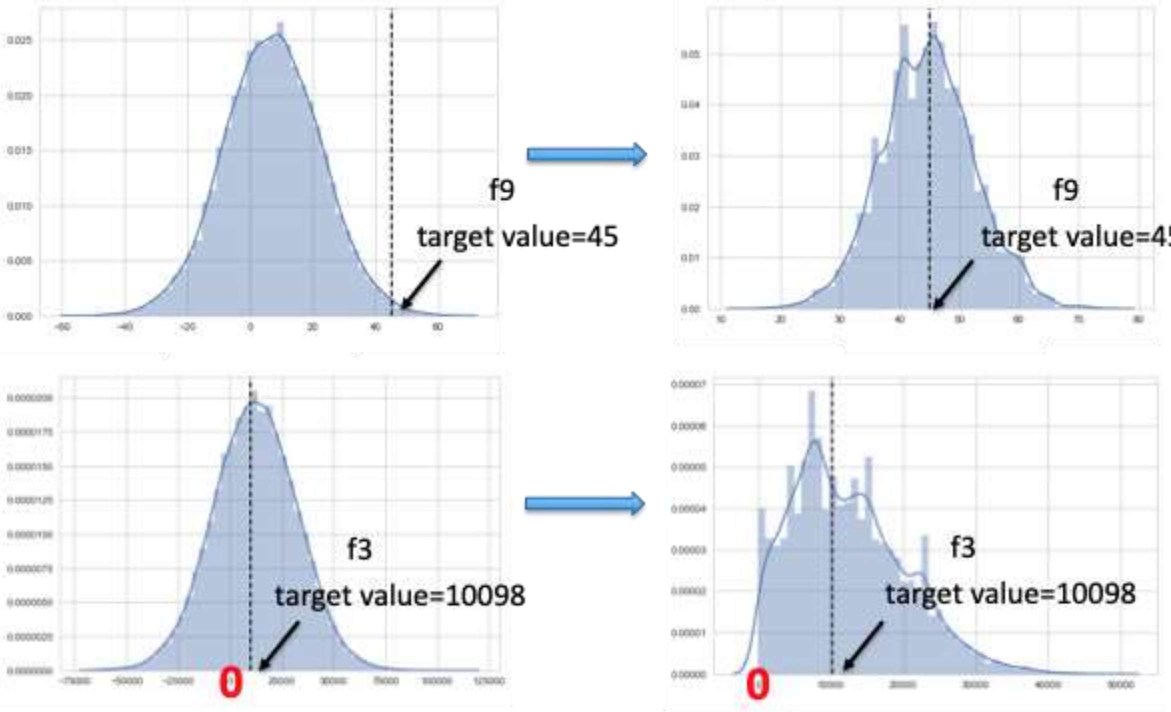
$$pred(X) \approx \beta_{k1}^{-}(X_1 - x_{k1})_- + \beta_{k1}^{+}(X_1 - x_{k1})_+ + \cdots,\quad X \in neighbor(x_k)$$

Impact of Piecewise Approach

- Target sample x_k=(x_{k1}, x_{k2}, $\cdots$)
- Top feature contributor
  - LIME: large magnitude of $\beta_{kj} \cdot x_{kj}$
  - xLIME: large magnitude of $\beta_{kj}^{-} \cdot x_{kj}$
- Top positive feature influencer
  - LIME: large magnitude of β_{kj}
  - xLIME: large magnitude of negative β_{kj}^{-} or positive β_{kj}^{+}
- Top negative feature influencer
  - LIME: large magnitude of β_{kj}
  - xLIME: large magnitude of positive β_{kj}^{-} or negative β_{kj}^{+}

Localized Stratified Sampling: Idea

Method: Sampling based on empirical distribution around target value at each feature level



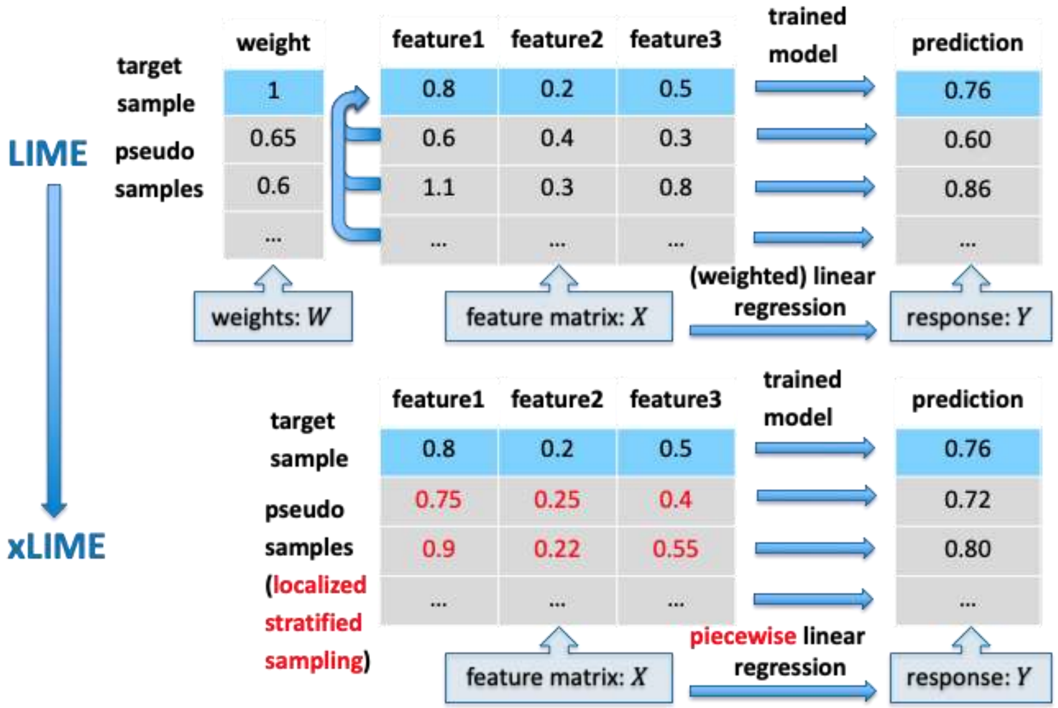
Localized Stratified Sampling: Method

- Sampling based on empirical distribution around target value for each feature
- For target sample $x_k = (x_{k1}, x_{k2}, \cdots)$, sampling values of feature j according to

$$p_j(X_j) \cdot N(x_{kj}, (\alpha \cdot s_j)^2)$$

  - $p_j(X_j)$: empirical distribution.
  - x_{kj} : feature value in target sample.
  - s_j : standard deviation.
  - α : Interpretable range: tradeoff between interpretable coverage and local accuracy.
- In LIME, sampling according to $N(\bar{x}_j, s_j^2)$.

Summary



LIME
xLIME
target sample
pseudo samples
weight
1
0.65
0.6
...
feature1
feature2
feature3
0.8 0.2 0.5
0.6 0.4 0.3
1.1 0.3 0.8
trained model
prediction
0.76
0.60
0.86
weights: W
feature matrix: X
(weighted) linear regression
response: Y
target sample
pseudo samples (localized stratified sampling)
0.8 0.2 0.5
0.75 0.25 0.4
0.9 0.22 0.55
trained model
prediction
0.76
0.72
0.80
feature matrix: X
piecewise linear regression
response: Y

LTS LCP (LinkedIn Career Page) Upsell

- A subset of churn data
  - Total Companies: ~ 19K
  - Company features: 117
- **Problem:** Estimate whether there will be upsell given a set of features about the company’s utility from the product

Top Feature Contributor

Company : CompanyX

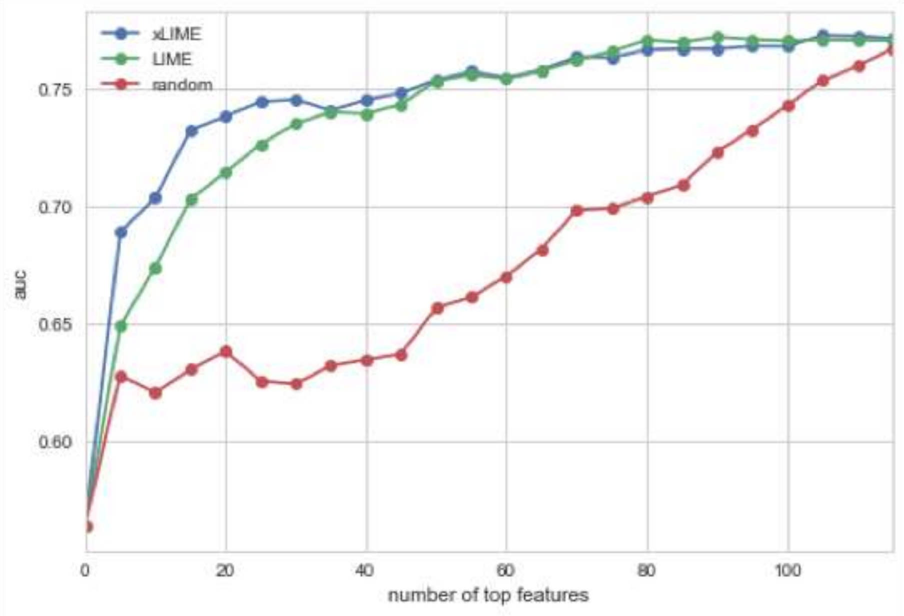
LIME

	name	value	quantile	contribution
	f9	45.0	98	-0.011
	f3	10097.6	66	0.011
	f10	16.5	94	0.010

xLIME

	name	value	quantile	contribution
	f1	430.5	59	0.246
	f2	216.0	40	0.161
	f3	10097.6	66	0.084

- **Explanation curve**: how classification performance varies if one considers only the top ranked feature contributors



Top Feature Influencers

Company: CompanyX

	Positive influencer			Negative influencer		
LIME	f1 +	430.5→712.3	.004	f1 -	430.5→148.7	.004
	f2 +	216.0→435.4	.004	f2 -	216.0→0.0	.004
	f11 +	9.8→13.2	.003	f11 -	9.8→6.3	.003
xLIME	f5 +	0.0→5.4	.032	f1 -	430.5→148.7	.201
	f6 -	168.0→0.0	.031	f2 -	216.0→0.0	.174
	f7 +	0.00→0.24	.016	f8 -	423.0→146.0	.071

Key Takeaways

- Looking at the explanation as contributor vs. influencer features is useful
  - Contributor: Which features end-up in the current outcome case-by-case
  - Influencer: **What needs to be done to improve likelihood, case-by-case**

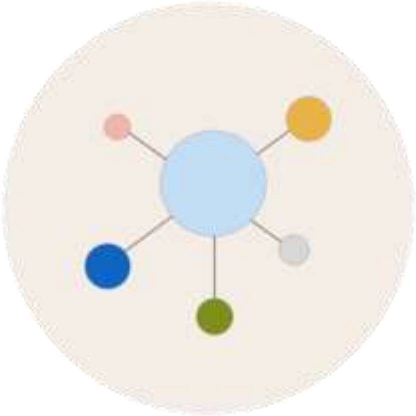
- xLIME aims to improve on LIME via:
  - Piecewise linear regression: More accurately describes local point, helps with finding correct influencers
  - Localized stratified sampling: More realistic set of local points
- Better captures the important features

Case Study:

Relevance Debugging and Explaining @ LinkedIn

Daniel Qiu, Yucheng Qian

Debugging Relevance Models

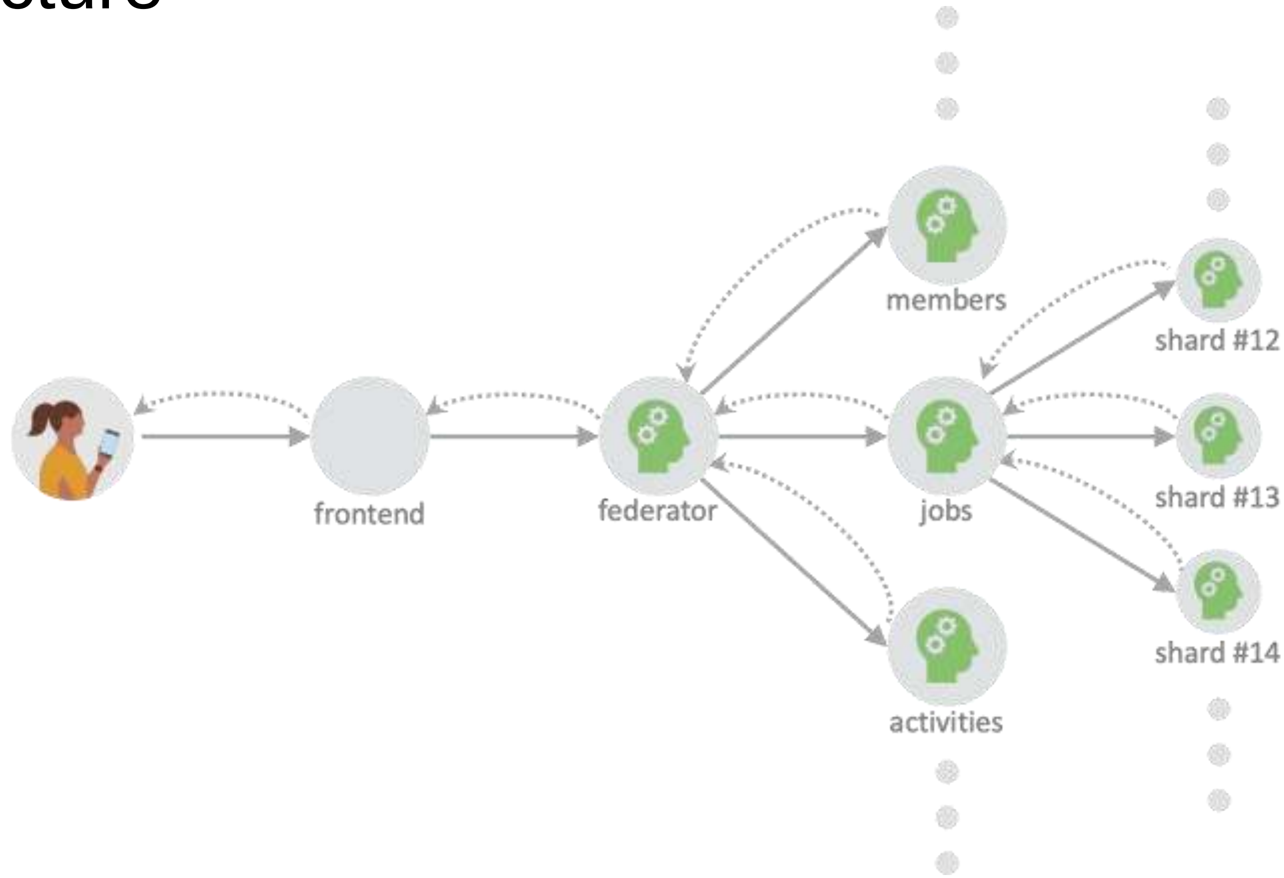


Modeling
Improve the machine learning model

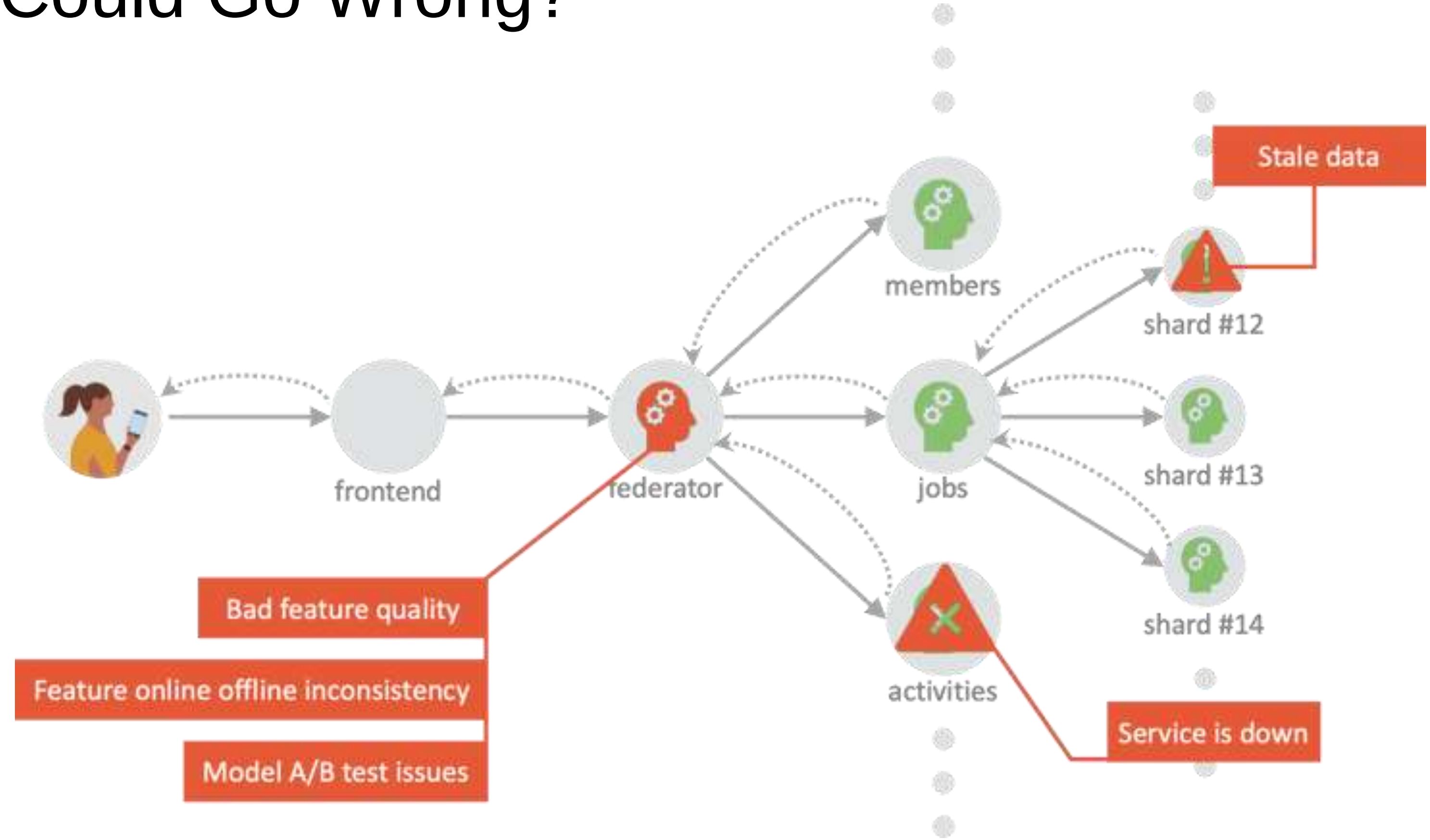
Value
Bring value to our members by providing relevant experience

Trust
Build trust with our members

Architecture

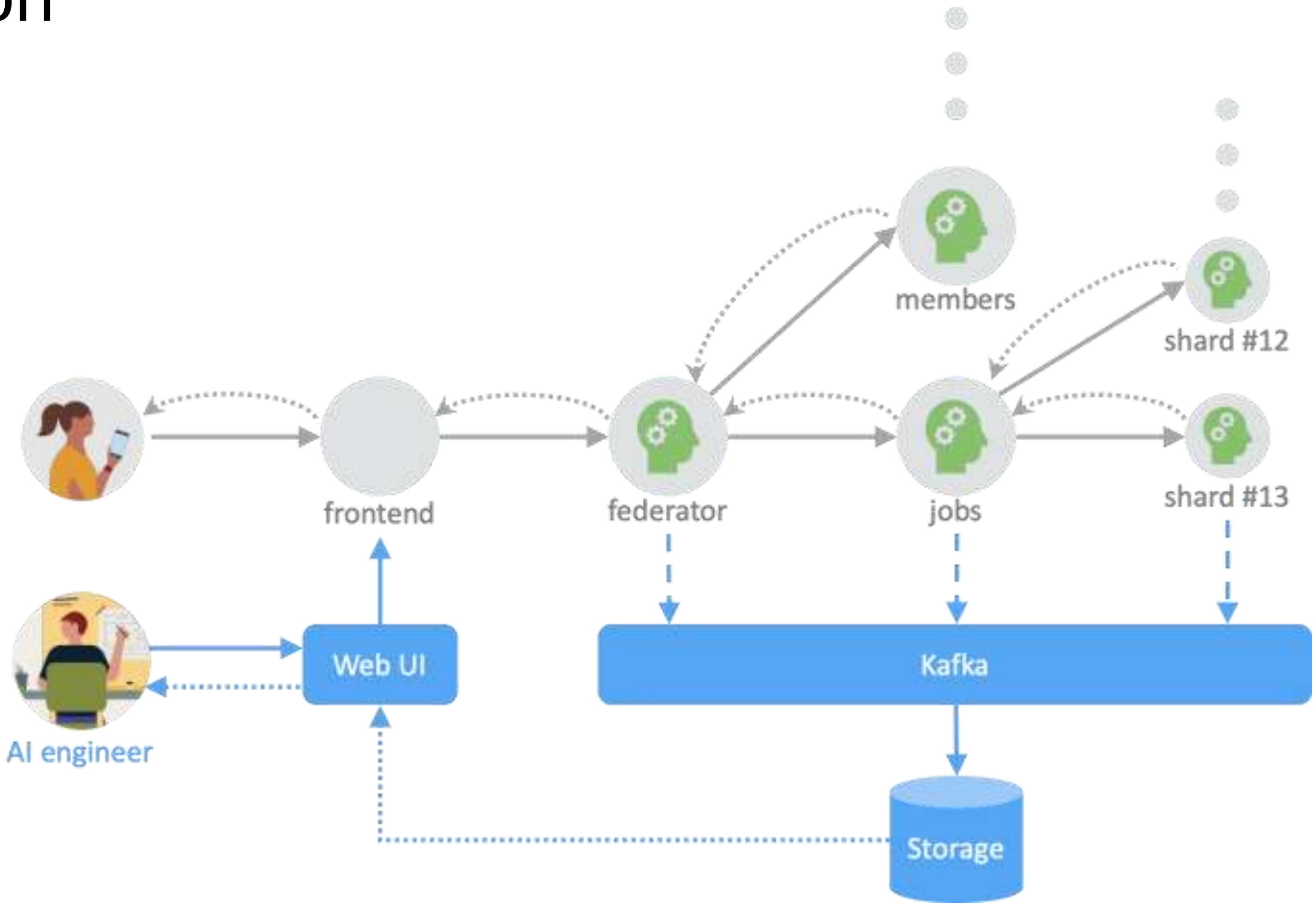


What Could Go Wrong?



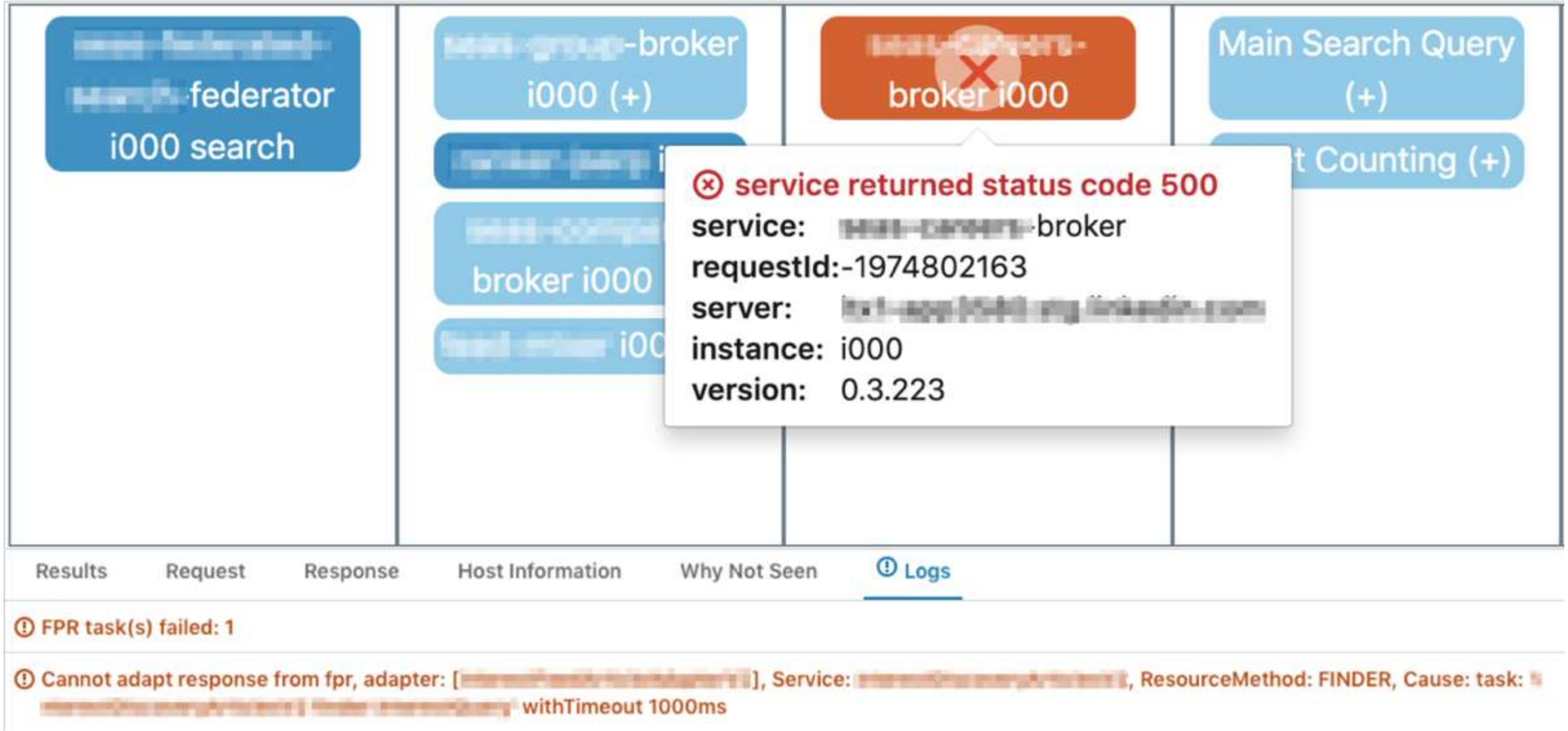
Challenges

Solution



members
shard #12
frontend
federator
jobs
shard #13
Web UI
Kafka
AI engineer
Storage

Call Graph

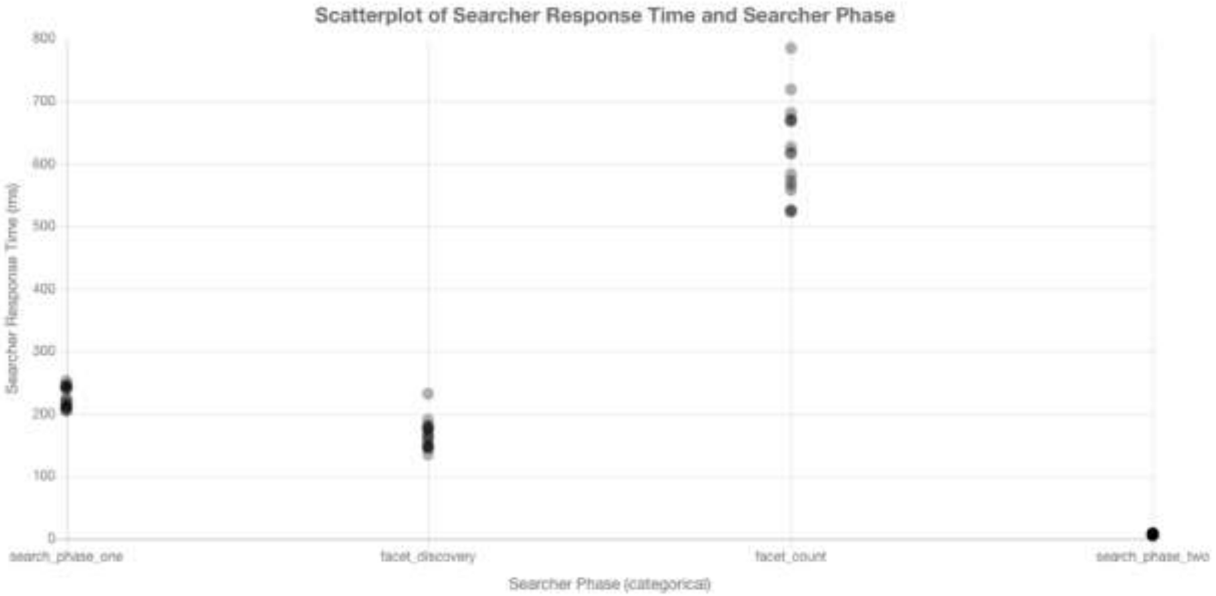

federator
i000 search
-broker
i000 (+)
broker i000
Main Search Query
(+)
t Counting (+)
broker i000
service returned status code 500
service: broker
requestId:-1974802163
server:
instance: i000
version: 0.3.223
Results
Request
Response
Host Information
Why Not Seen
Logs
FPR task(s) failed: 1
Cannot adapt response from fpr, adapter: [ ], Service: , ResourceMethod: FINDER, Cause: task:
withTimeout 1000ms

Timing

Total time (ms): 1041

Number of garbage collection events: 0

	Start Time	End Time	Total Time	Resent?	Partitions	Min	Max	p50	p90
search_phase_one	7	266	259	false	16	205	253	223.0	245.5
facet_discovery	13	240	227	true	16	135	232	164.0	186.0
facet_count	262	1041	779	true	16	523	785	617.0	700.0
search_phase_two	266	274	8	false	15	5	9	8.0	9.0



Features

Group	Feature	Value
SPR	activity_recent_click /	968
SPR	[redacted]	1
SPR	[redacted]	6.8762646
SPR	[redacted]	null
SPR	[redacted]	null
SPR	binary_activity_recent_click / |	1
SPR	[redacted]	null
SPR	log_activity_recent_click /	6.8762646
SPR	[redacted]	0
SPR	[redacted]	0

Advanced Use Cases

Perturbation

Comparison

Replay

Perturbation

1. Inject

Injected as part of the request

- Override A/B test settings
- Model selection
- Feature override

2. Relay

Passed to downstream service

3. Overwrite

Overwrite the system behavior

Comparison

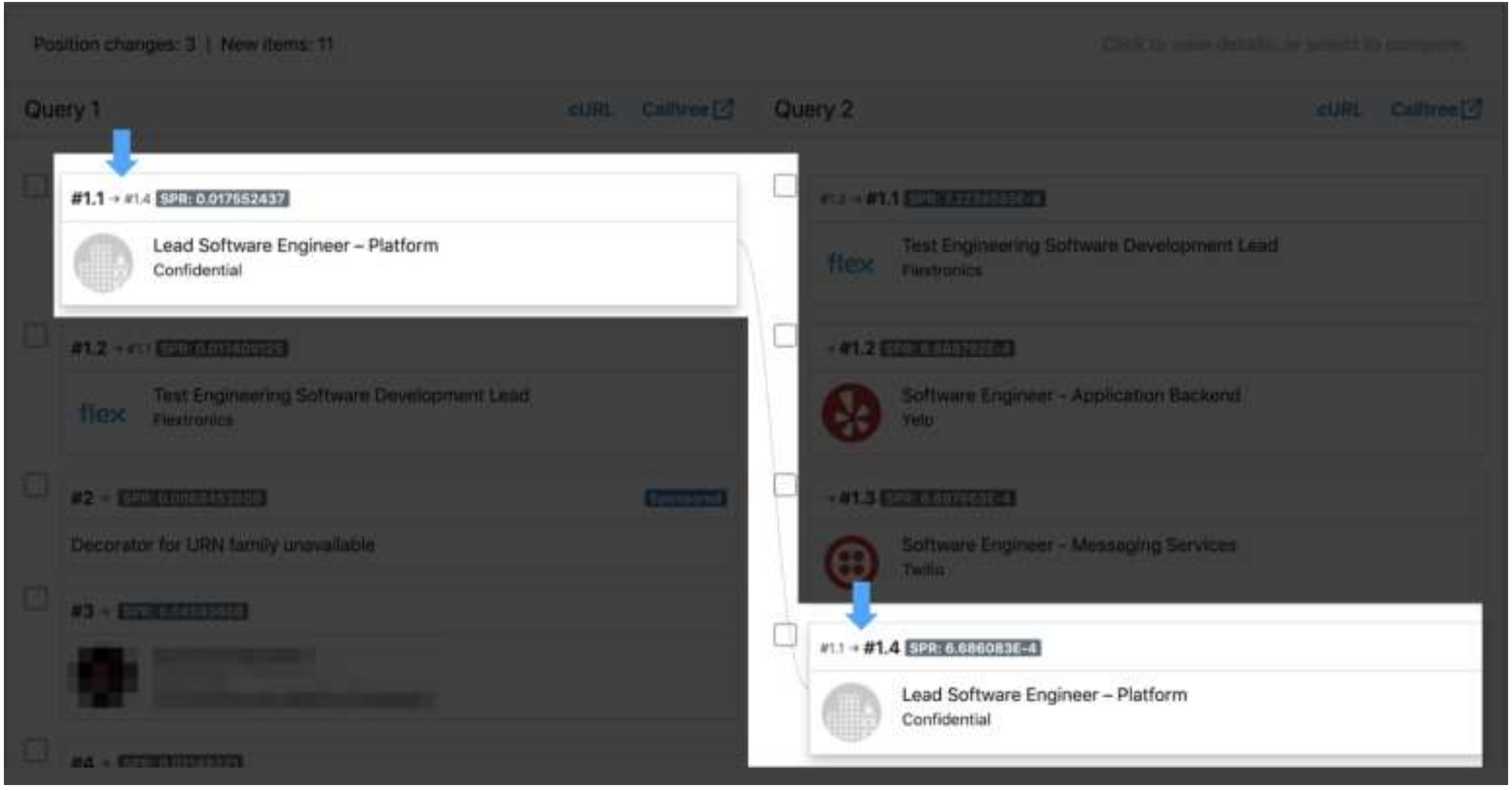
Compare Model

Compare results of 2 different queries/models

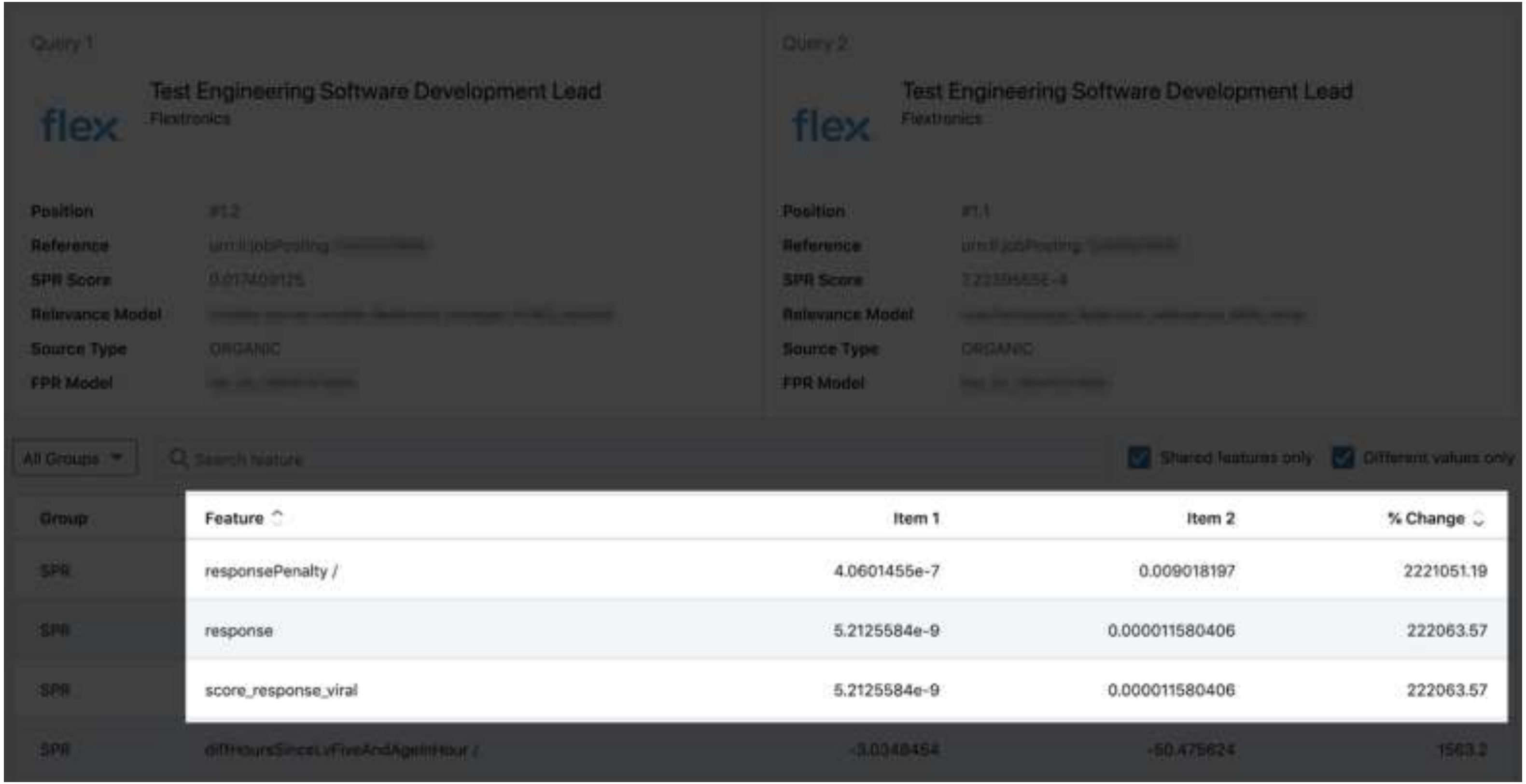
Compare Items

Compare features and scores of 2 different items, from the same query or different queries

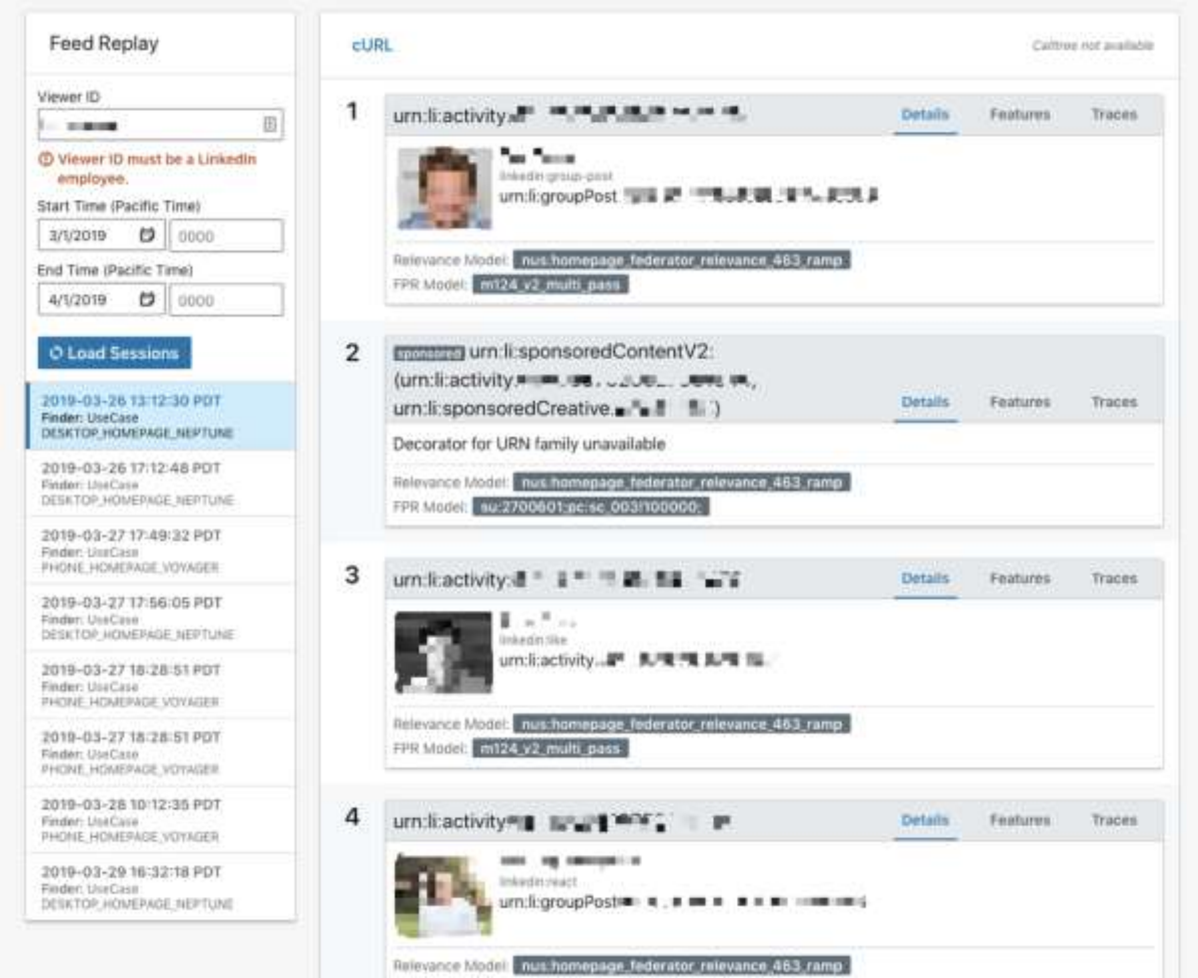
Holistic Comparison



Granular Comparison



Replay


Feed Replay
Viewer ID
Viewer ID must be a LinkedIn employee.
Start Time (Pacific Time)
3/1/2019
0000
End Time (Pacific Time)
4/1/2019
0000
Load Sessions
2019-03-26 13:12:30 PDT
Finder: UseCase
DESKTOP_HOMEPAGE_NEPTUNE
2019-03-26 17:12:48 PDT
Finder: UseCase
DESKTOP_HOMEPAGE_NEPTUNE
2019-03-27 17:49:32 PDT
Finder: UseCase
PHONE_HOMEPAGE_VOYAGER
2019-03-27 17:56:05 PDT
Finder: UseCase
DESKTOP_HOMEPAGE_NEPTUNE
2019-03-27 18:28:51 PDT
Finder: UseCase
PHONE_HOMEPAGE_VOYAGER
2019-03-27 18:28:51 PDT
Finder: UseCase
PHONE_HOMEPAGE_VOYAGER
2019-03-28 10:12:35 PDT
Finder: UseCase
PHONE_HOMEPAGE_VOYAGER
2019-03-29 16:32:18 PDT
Finder: UseCase
DESKTOP_HOMEPAGE_NEPTUNE
cURL
1
urn:li:activity
Details
Features
Traces
urn:li:groupPost
Relevance Model: nus:homepage_federator_relevance_463_ramp
FPR Model: m124_v2_multi_pass
2
sponsored urn:li:sponsoredContentV2:
(urn:li:activity
urn:li:sponsoredCreative
Details
Features
Traces
Decorator for URN family unavailable
Relevance Model: nus:homepage_federator_relevance_463_ramp
FPR Model: su:2700601:pc:sc_003!100000;
3
urn:li:activity
Details
Features
Traces
urn:li:activity
Relevance Model: nus:homepage_federator_relevance_463_ramp
FPR Model: m124_v2_multi_pass
4
urn:li:activity
Details
Features
Traces
urn:li:groupPost
Relevance Model: nus:homepage_federator_relevance_463_ramp

Teams

- Search
- Feed
- Comments
- People you may know
- Jobs you may be interested in
- Notification

Case Study:

Integrated Gradients for Adversarial Analysis of Question-Answering models

Ankur Taly** (Fiddler labs)

(Joint work with Mukund Sundararajan, Kedar Dhamdhere, Pramod Mudrakarta)

**This research was carried out at Google Research

Tabular QA

Rank	Nation	Gold	Silver	Bronze	Total
1	India	102	58	37	197
2	Nepal	32	10	24	65
3	Sri Lanka	16	42	62	120
4	Pakistan	10	36	30	76
5	Bangladesh	2	10	35	47
6	Bhutan	1	6	7	14
7	Maldives	0	0	4	4

Q: How many medals did India win?
A: 197

Neural Programmer (2017) model
33.5% accuracy on WikiTableQuestions

Visual QA

Q: How symmetrical are the white bricks on either side of the building?
A: very

Kazemi and Elqursh (2017) model.
61.1% on VQA 1.0 dataset
(state of the art = 66.7%)

Reading Comprehension

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager

Q: Name of the quarterback who was 38 in Super Bowl XXXIII?
A: John Elway

Yu et al (2018) model.
84.6 F-1 score on SQuAD (state of the art)

Robustness question: Do these models understand the question? :-)

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)

Q: How symmetrical are the white bricks on either side of the building?

A: very

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)

Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)

Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)

Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Q: How fast are the bricks speaking on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)

Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Q: How fast are the bricks speaking on either side of the building?
A: very

Test/dev accuracy does not show us the entire picture. Need to look inside!

Analysis procedure

- Attribute the answer (or answer selection logic) to question words
  - **Baseline**: Empty question, but full context (image, text, paragraph)
    - By design, attribution will **not** fall on the context
- Visualize attributions per example
- Aggregate attributions across examples

Visual QA attributions

Q: How symmetrical are the white bricks on either side of the building?
A: very

How symmetrical are the white bricks on either side of the building?

red: high attribution
blue: negative attribution
gray: near-zero attribution

Over-stability [Jia and Liang, EMNLP 2017]

Jia & Liang note that:

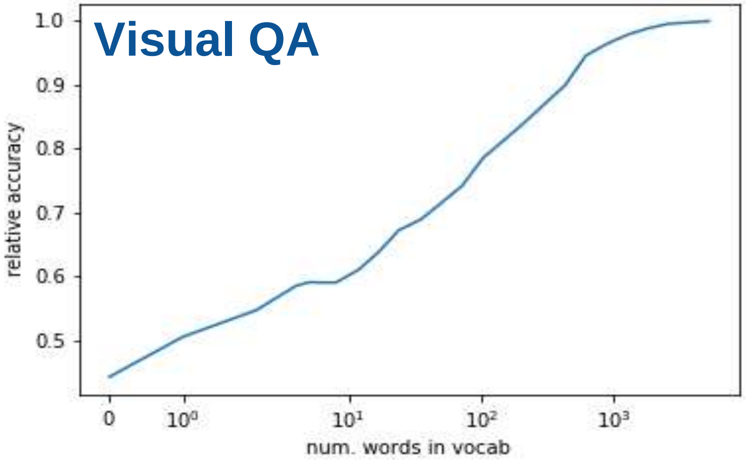
- Image networks suffer from **“over-sensitivity”** to pixel perturbations
- Paragraph QA models suffer from **“over-stability”** to semantics-altering edits

Attributions show how such over-stability manifests in Visual QA, Tabular QA and Paragraph QA networks

Over-stability

During inference, drop all words from the dataset except ones which are frequently top attributions

- E.g. How many ~~red buses are in the picture?~~

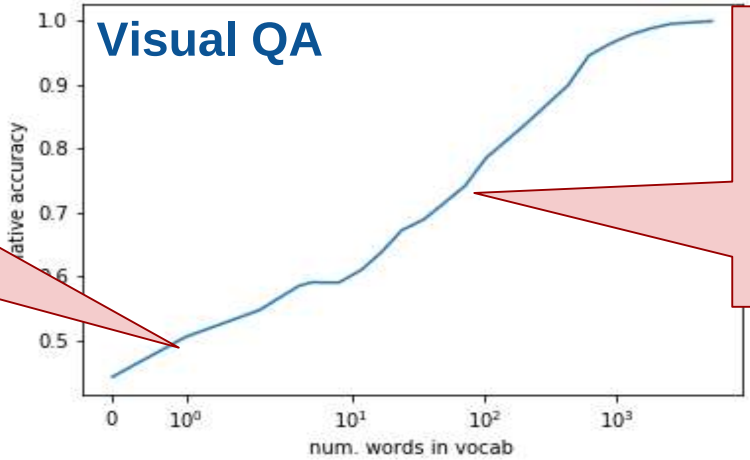


Top tokens: color, many, what, is, how, there, …

Over-stability

During inference, drop all words from the dataset except ones which are frequently top attributions

- E.g. How many ~~red buses are in the picture?~~



50% of final accuracy with just one word **“color”**

80% of final accuracy reached with just 100 words

(Orig vocab size: 5305)

Top tokens: color, many, what, is, how, there, …

Attack: Subject ablation

Replace the subject of a question with a low-attribution noun from the vocabulary

- This **ought to change** the answer but often does not!

Low-attribution nouns
'tweet',
'childhood',
'copyrights',
'mornings',
'disorder',
'importance',
'topless',
'critter',
'jumper',
'fits'

What is the **man** doing? → What is the **tweet** doing?
How many **children** are there? → How many **tweet** are there?

VQA model's response remains the same 75.6% of the time on questions that it originally answered correctly

Many other attacks!

- Visual QA
  - Prefix concatenation attack (accuracy drop: **61.1% to 19%**)
  - Stop word deletion attack (accuracy drop: **61.1% to 52%**)
- Tabular QA
  - Prefix concatenation attack (accuracy drop: **33.5% to 11.4%**)
  - Stop word deletion attack (accuracy drop: **33.5% to 28.5%**)
  - Table row reordering attack (accuracy drop: **33.5 to 23%**)
- Paragraph QA
  - Improved paragraph concatenation attacks of Jia and Liang from [EMNLP 2017]

Paper: Did the model understand the question? **[ACL 2018]**

Fiddler Demo

Fiddler is an explainable AI engine designed for the enterprise

Pluggable Platform

Integrate, **deploy**, **visualize** a wide variety of **custom models**

Explainable AI

Deliver **clear decisions** and **explanations** to your end users

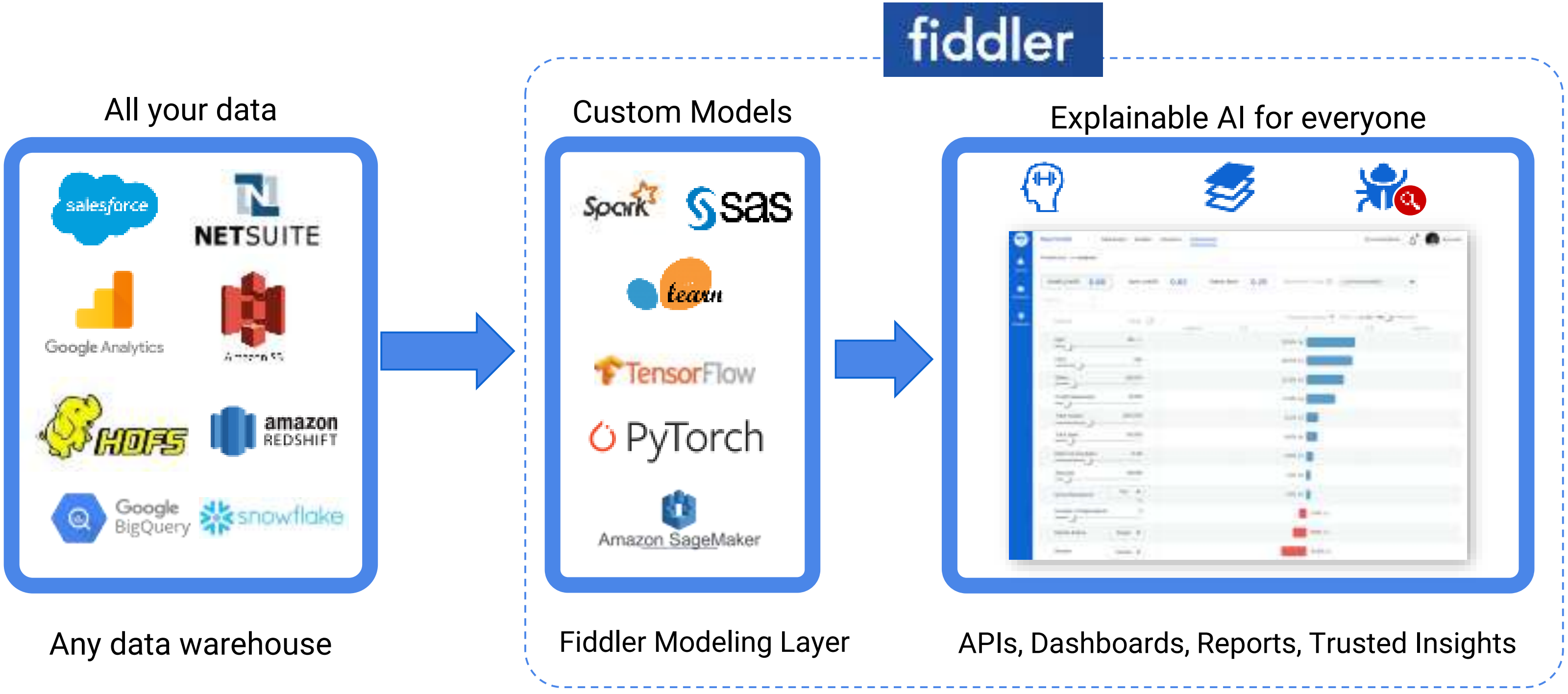
Trust & Governance

Easy **governed access** helps teams build and understand **Trusted AI**

Simplified Setup

Lean and pluggable AI platform with **cloud or on-prem** integrations

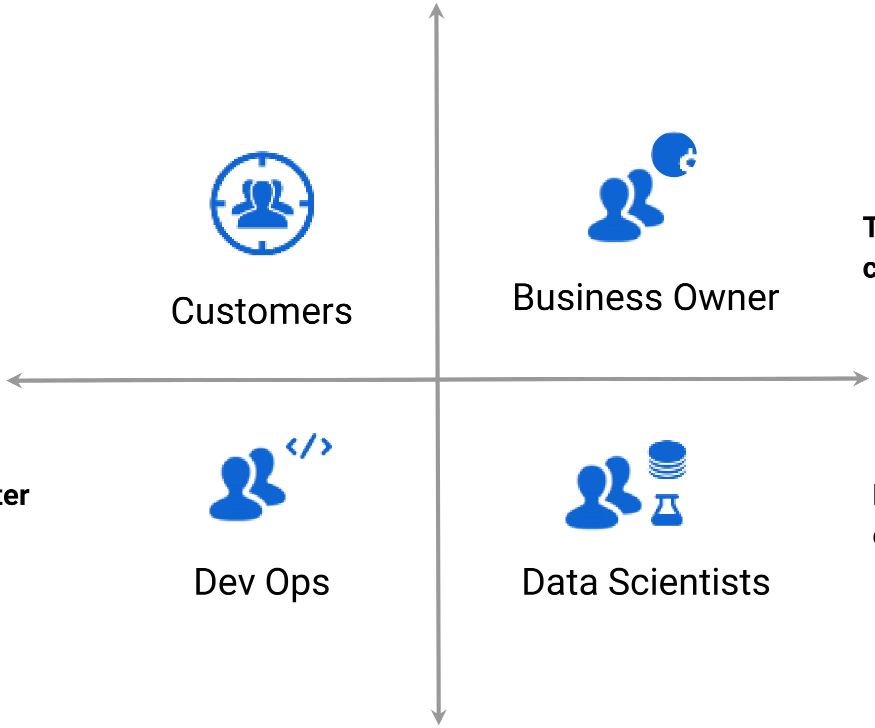
Fiddler - Explainable AI Engine



Benefits Across the Organization

More trust and customer retention

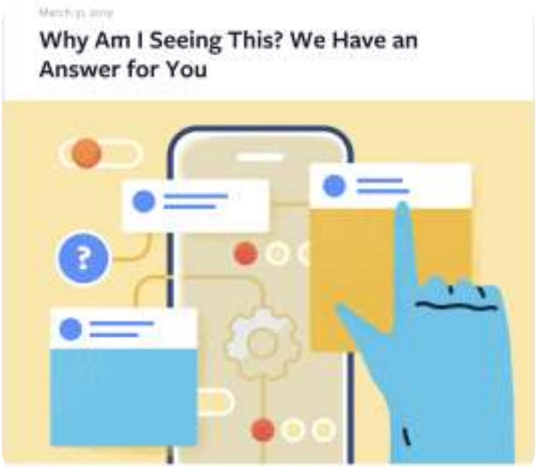
Transparent AI workflow and GDPR compliance

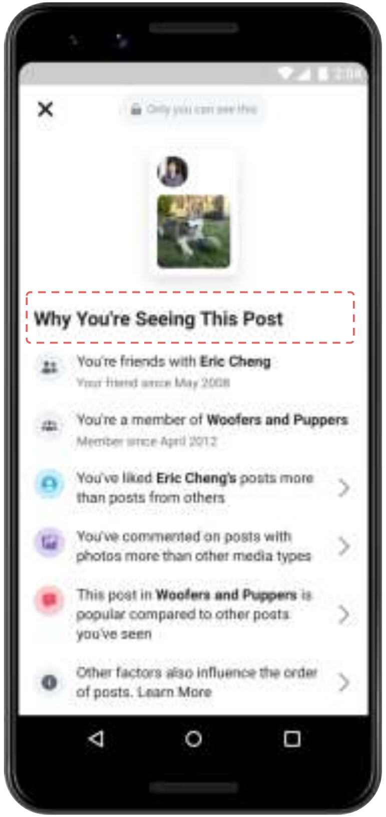


DevOps and IT have greater visibility into AI

Data Scientists are more agile and effective

Fiddler enables building Explainable AI Applications like this!





Can explanations help build Trust?

- Can we know when the model is uncertain?
- Does the model make the same mistake as a human?
- Are we comfortable with the model?

Zachary Lipton, et al. The Mythos of Model Interpretability ICML 2016

Can explanations help identify Causality?

- Predictions vs actions
- Explanations on why this happened as opposed to how

**Zachary Lipton, et al. The Mythos of Model Interpretability ICML 2016*

Can explanations be Transferable?

- Training and test setups often differ from the wild

- Real world data is always changing and noisy

**Zachary Lipton, et al. The Mythos of Model Interpretability ICML 2016*

Can explanations provide more Information?

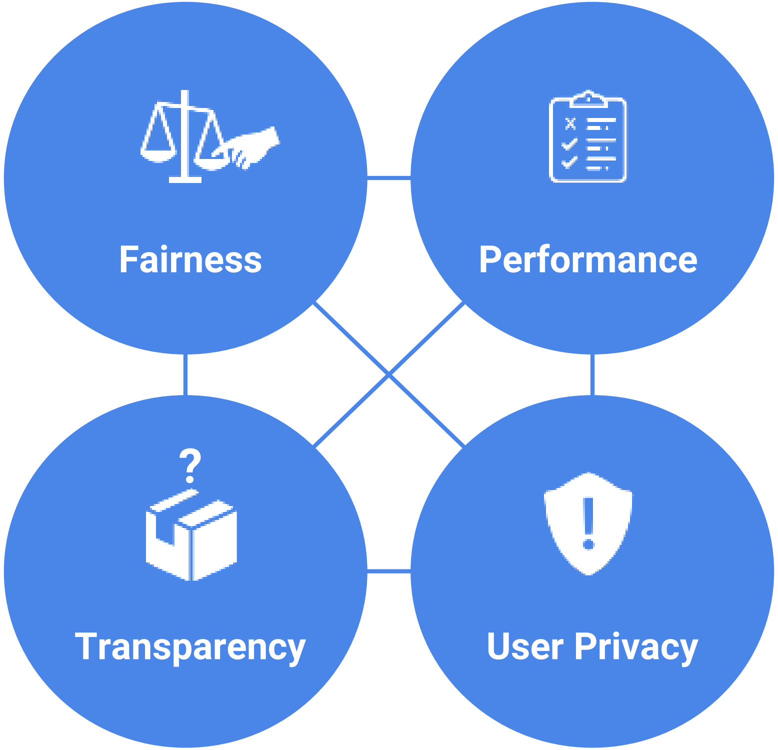
- Often times models aid human decisions

- Extra bits of information other than model decision could be valuable

**Zachary Lipton, et al. The Mythos of Model Interpretability ICML 2016*

Challenges & Tradeoffs

- Lack of standard interface for ML models makes pluggable explanations hard
- Explanation needs vary depending on the type of the user who needs it and also the problem at hand.
- The algorithm you employ for explanations might depend on the use-case, model type, data format, etc.
- There are trade-offs w.r.t. Explainability, Performance, Fairness, and Privacy.



Reflections

- Case studies on explainable AI in practice
- Need “Explainability by Design” when building AI products

Fairness

Privacy

Related KDD'19 sessions:

1. Tutorial: Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned (Sun)
2. Workshop: Explainable AI/ML (XAI) for Accountability, Fairness, and Transparency (Mon)
3. Social Impact Workshop (Wed, 8:15 – 11:45)
4. Keynote: Cynthia Rudin, Do Simpler Models Exist and How Can We Find Them? (Thu, 8 - 9am)
5. Several papers on fairness (e.g., ADS7 (Thu, 10-12), ADS9 (Thu, 1:30-3:30))
6. Research Track Session RT17: Interpretability (Thu, 10am - 12pm)

Transparency

Explainability

Thanks! Questions?

- Feedback most welcome :-)
  - krishna@fiddler.ai, sgeyik@linkedin.com, kkenthapadi@linkedin.com, vamithal@linkedin.com, ankur@fiddler.ai
- Tutorial website: https://sites.google.com/view/kdd19-explainable-ai-tutorial
- To try Fiddler, please send an email to info@fiddler.ai