6.7930/HST.956 — Machine Learning for Healthcare
Spring 2025

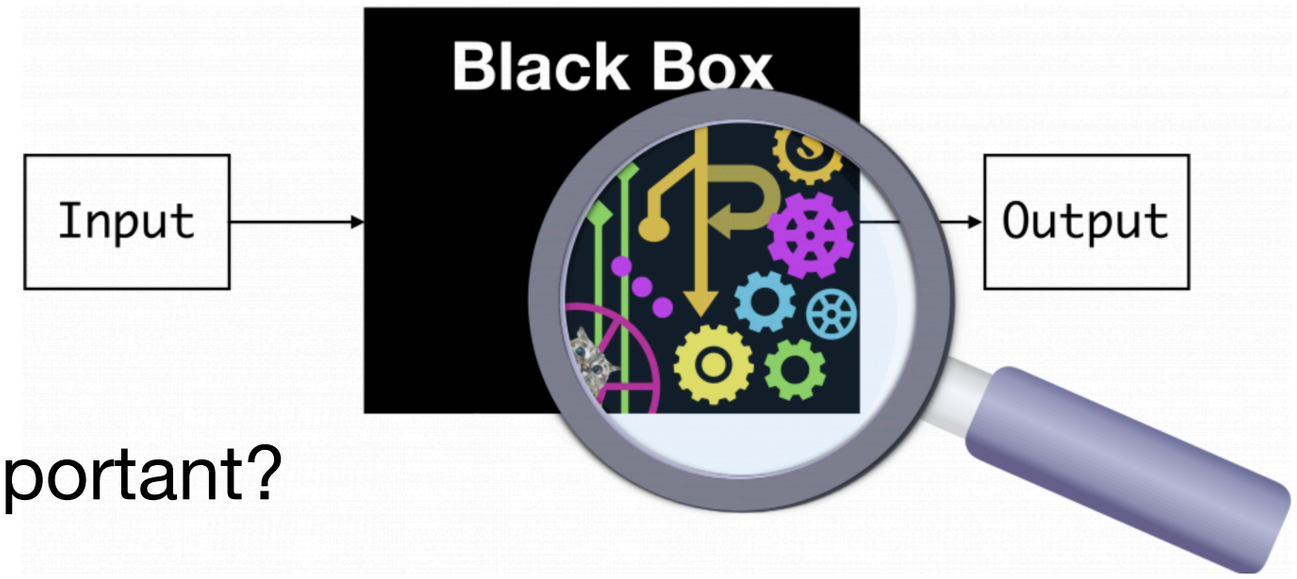
Recitation 8

Interpretability of ML Models

Linear Models
Non-linear Models
Local vs. Global
LIME
Influence Functions
Mechanistic Interpretability

Ilker Demirel, 4/18/2025

What is interpretability?



- Problem specific definition
  - Which features are most important?
  - Which part of the input is the model looking at?
  - ...
- Complex models more performant than simple models
  - Why trade performance with interpretability?

Linear Models

- CoxPH, Logistic Regression, LASSO
  - Most of the time, interpretability is the *end-goal*
- Still not straightforward to interpret

Model	Feature	Target
?	Blood pressure medication (Lisinopril)	Time until heart failure

Linear Models

- CoxPH, Logistic Regression, LASSO
  - Most of the time, interpretability is the *end-goal*
- Still not straightforward to interpret

Model	Feature	Target
CoxPH	Blood pressure medication (Lisinopril)	Time until heart failure

$$\beta = -0.5$$

Linear Models

- CoxPH, Logistic Regression, LASSO
  - Most of the time, interpretability is the *end-goal*
- Still not straightforward to interpret

Model	Feature	Target
CoxPH	Blood pressure medication (Lisinopril)	Time until heart failure

$$\beta = -0.5$$

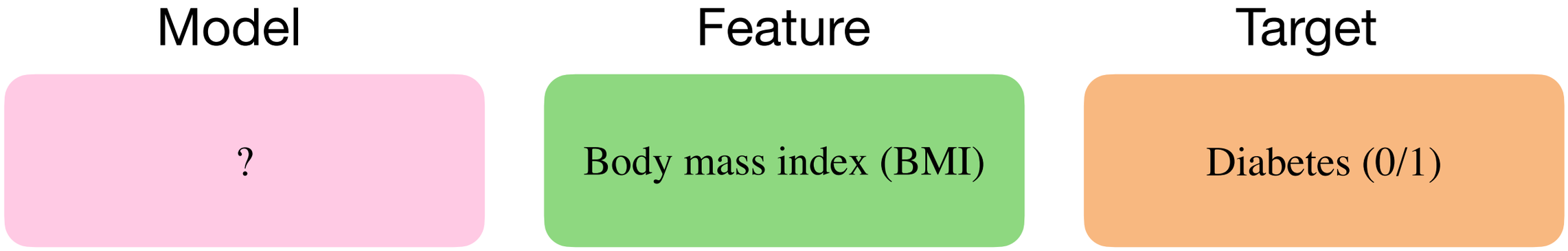
$$\text{HR} = \exp(\beta) = 0.61$$

Causal/interventional interpretation?

Does Lisinopril reduce the HF by 0.39?

Linear Models

- CoxPH, Logistic Regression, LASSO
  - Most of the time, interpretability is the *end-goal*
- Still not straightforward to interpret



Linear Models

- CoxPH, Logistic Regression, LASSO
  - Most of the time, interpretability is the *end-goal*
- Still not straightforward to interpret

Model	Feature	Target
Logistic Regression	Body mass index (BMI)	Diabetes (0/1)

$$\beta = 0.09$$

Linear Models

- CoxPH, Logistic Regression, LASSO
  - Most of the time, interpretability is the *end-goal*
- Still not straightforward to interpret

Model	Feature	Target
Logistic Regression	Body mass index (BMI)	Diabetes (0/1)

$$\beta = 0.09$$

$$\text{OR} = \exp(\beta) = 1.09$$

For each unit increase in BMI, the odds of having diabetes increase by ~9%

Associational

Linear Models

- CoxPH, Logistic Regression, LASSO
  - Most of the time, interpretability is the *end-goal*
- Still not straightforward to interpret

Model fitted on ICU patients

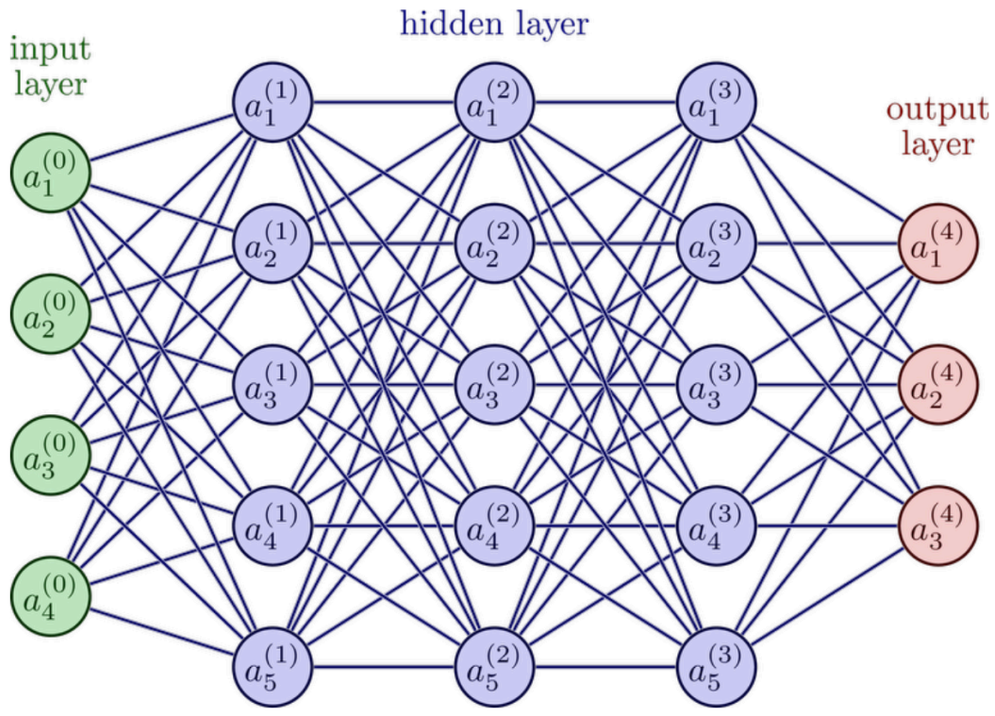
Model	Feature	Target
Logistic Regression	Pneumonia (0/1)	Survival (0/1)

$$\beta = 0.2$$

Does having pneumonia
increases the odds of survival?

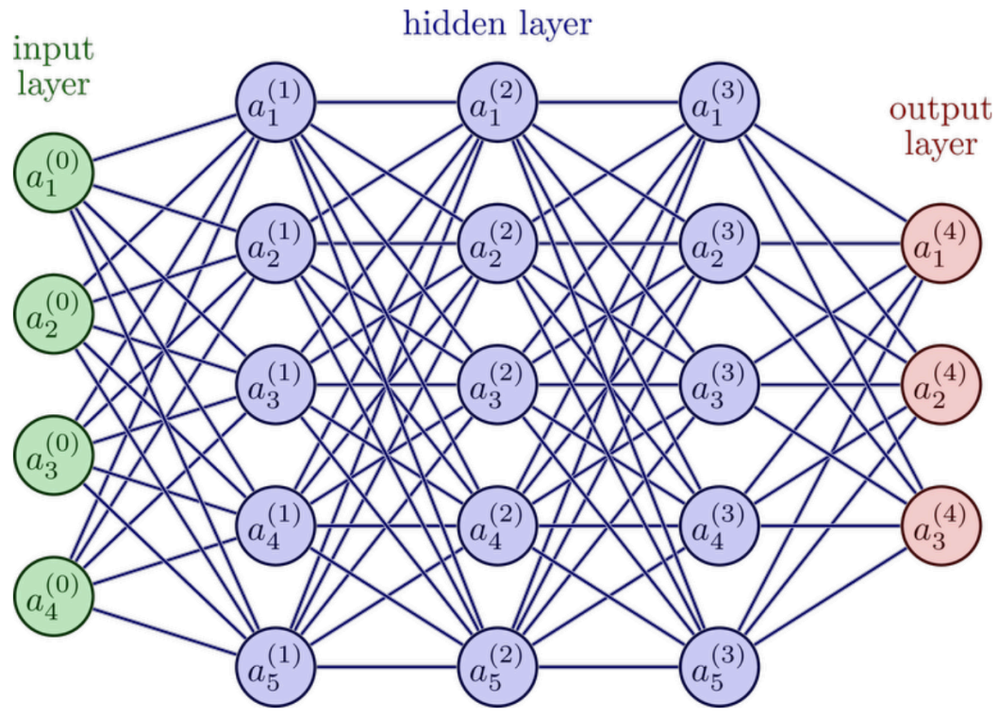
Non-linear Models

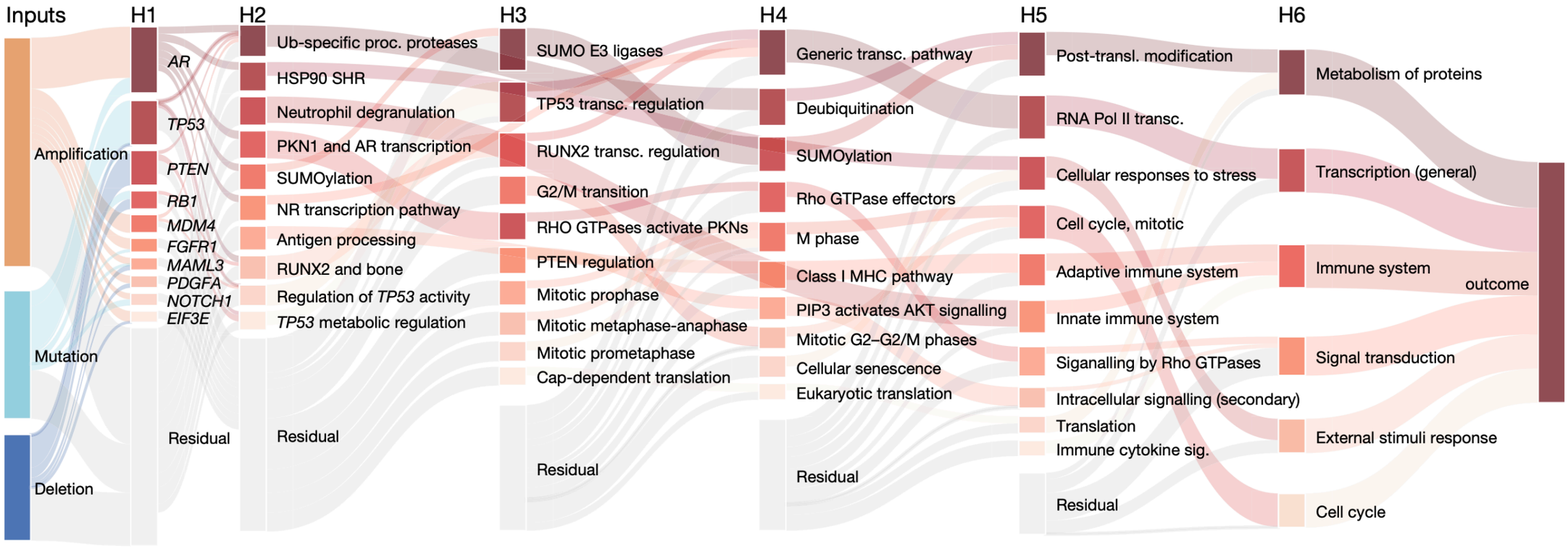
- What does a parameter mean?



Non-linear Models

- What does a parameter mean?





Biologically informed deep neural network for prostate cancer discovery

https://doi.org/10.1038/s41586-021-03922-4

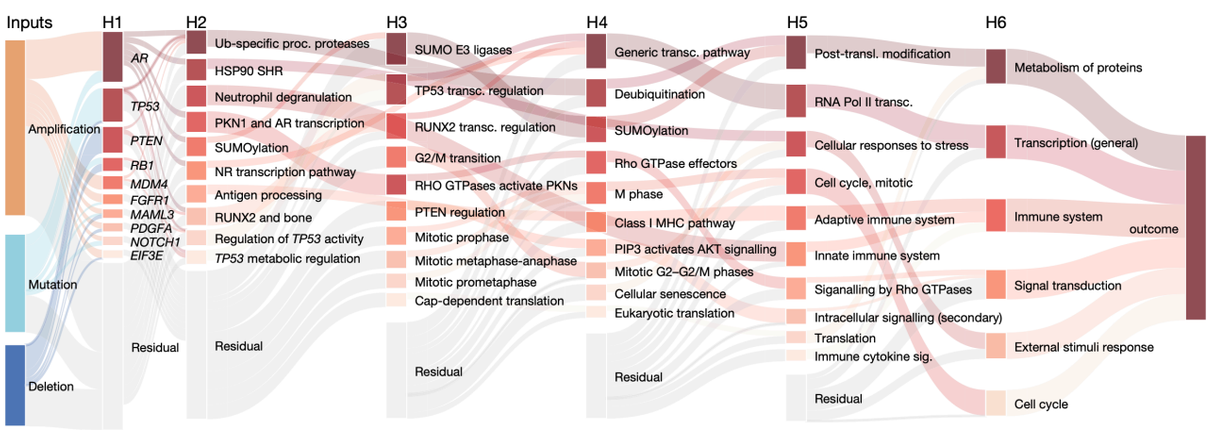
Received: 6 November 2020

Accepted: 17 August 2021

Haitham A. Elmarakeby[1,2,3], Justin Hwang[4], Rand Arafeh[1,2], Jett Crowdis[1,2], Sydney Gang[1], David Liu[1,2], Saud H. AlDubayan[1,2], Keyan Salari[1,2,5], Steven Kregel[6], Camden Richter[1], Taylor E. Arnoff[1,2], Jihye Park[1,2], William C. Hahn[1,2] & Eliezer M. Van Allen[1,2]✉

Local vs. Global

- Global – describe the model entirely



Model

Logistic Regression

Feature

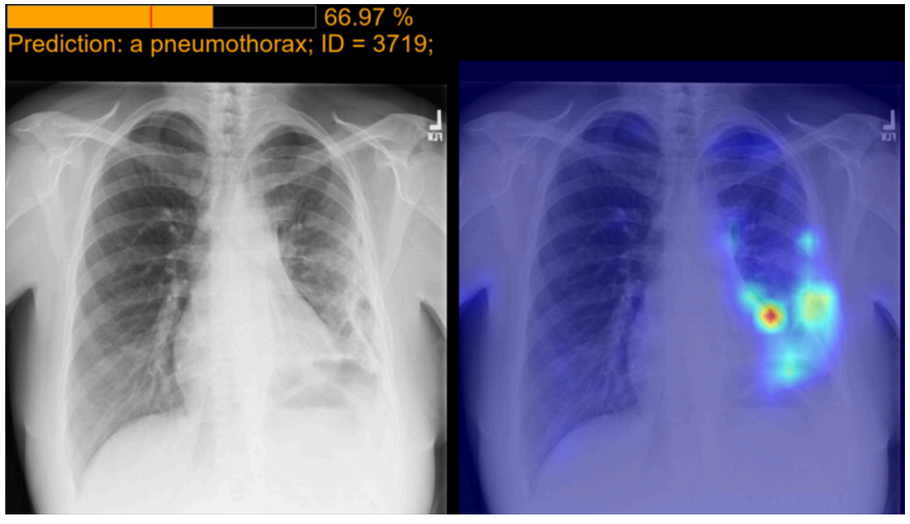
Body mass index (BMI)

Target

Diabetes (0/1)

$\beta = 0.09$

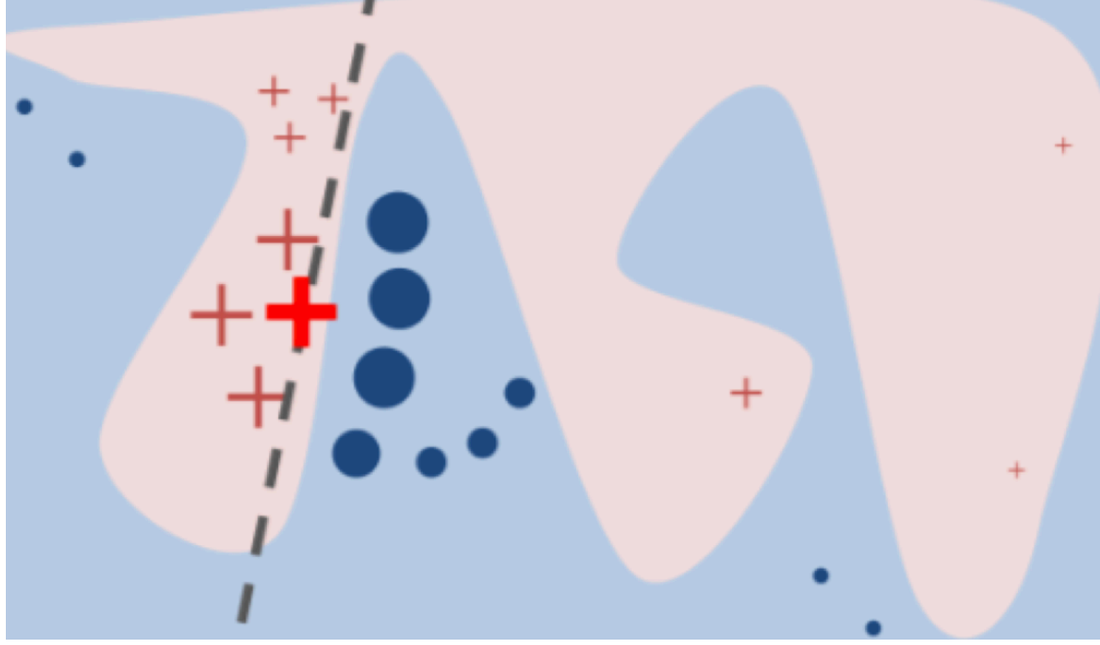
- Local – explain model decision for a specific data point
  - LIME
  - Influence functions
  - Saliency maps, GradCAM



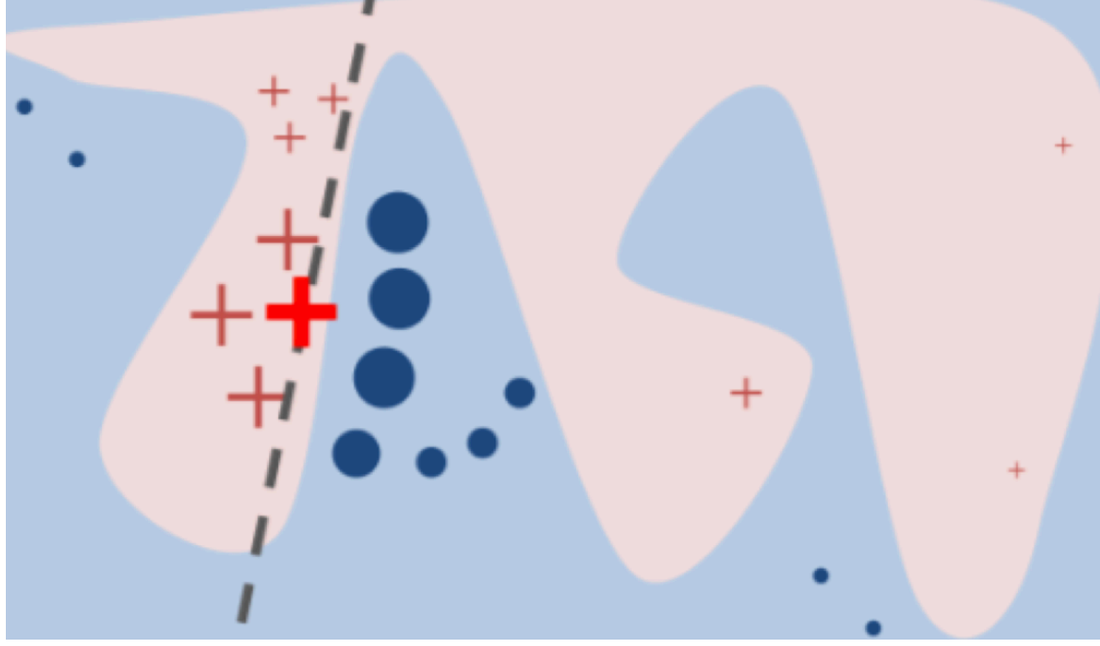
$$\frac{dS(\beta, x)}{dx_i}$$

LIME

- Explain a single prediction x_0
- Explanation: Another *simple* model g
  - Approximation to original model f *around* x_0



LIME

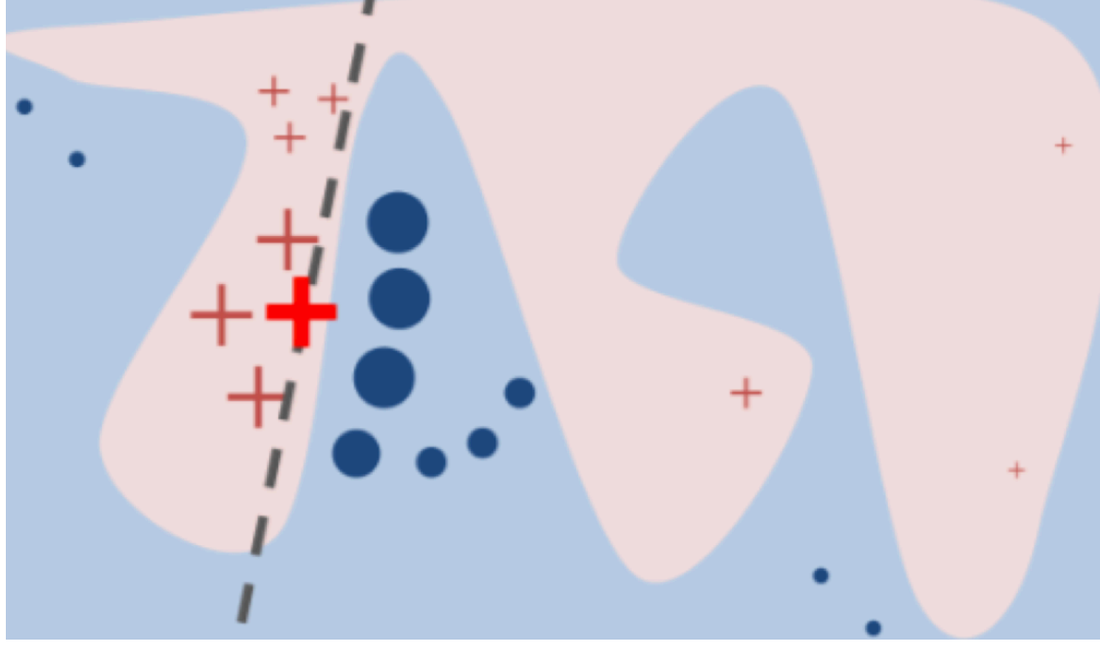


- Explain a single prediction x_0
- Explanation: Another *simple* model g
  - Approximation to original model f *around* x_0

1. Define a neighborhood: $\pi(x_0)$
2. For each $x \in \pi(x_0)$, compute similarity

$$k(x, x_0) = \exp\left(-\left(\frac{\left(x(1) - x_0(1)\right)^2}{\sigma_1^2} + \frac{\left(x(2) - x_0(2)\right)^2}{\sigma_2^2} \right) \right)$$

LIME



- Explain a single prediction x_0
- Explanation: Another *simple* model g
  - Approximation to original model f *around* x_0

1. Define a neighborhood: $\pi(x_0)$

2. For each $x \in \pi(x_0)$, compute similarity

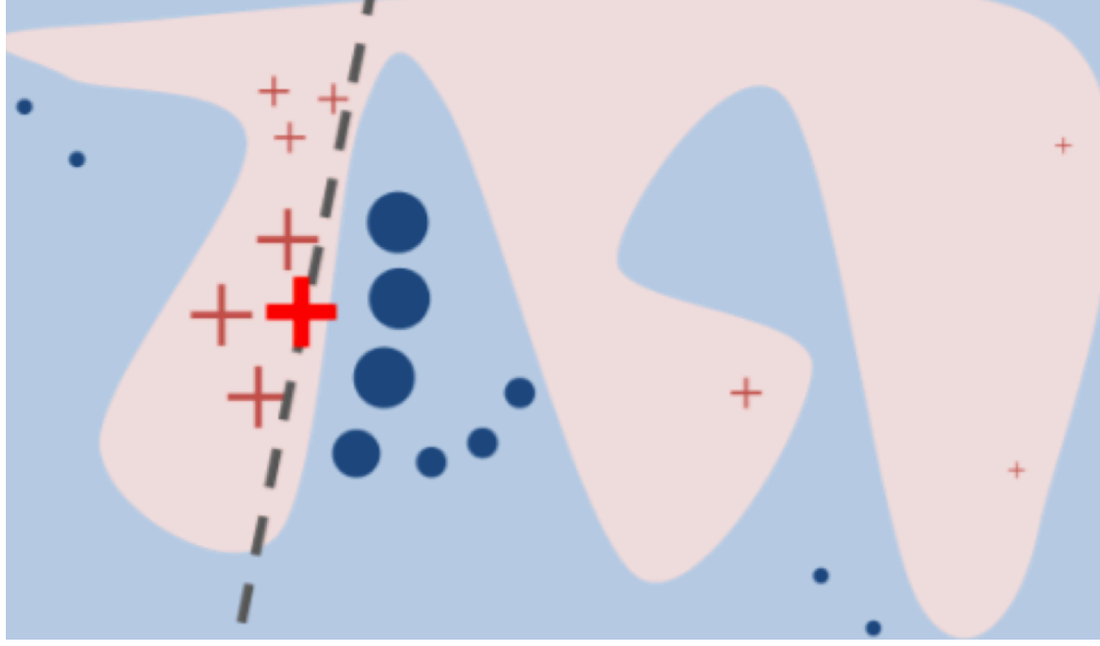
$$k(x, x_0) = \exp\left(-\left(\frac{\left(x(1) - x_0(1)\right)^2}{\sigma_1^2} + \frac{\left(x(2) - x_0(2)\right)^2}{\sigma_2^2} \right) \right)$$

$$\mathscr{L}(f, x_0, g(\theta)) = \sum_{x \in \pi(x_0)} k(x, x_0)(f(x) - g(x))^2 + \lambda \sum_i \theta_i$$

$\text{argmin}_{g \in G} \mathscr{L}(f, g, x_0)$

G is the family of linear functions

LIME



- Explain a single prediction x_0
- Explanation: Another *simple* model g
  - Approximation to original model f *around* x_0

1. Define a neighborhood: $\pi(x_0)$

2. For each $x \in \pi(x_0)$, compute similarity

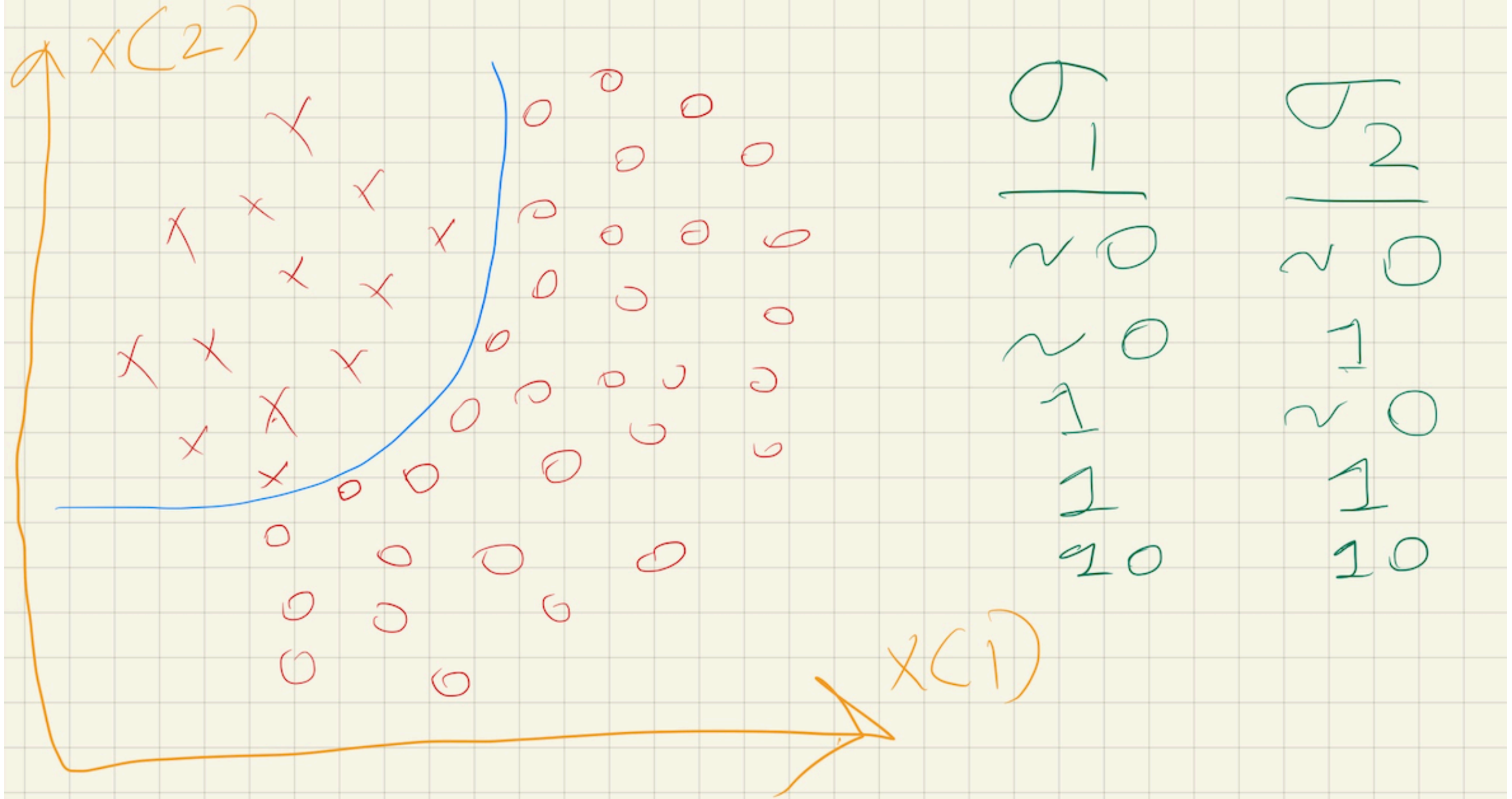
$$k(x, x_0) = \exp\left(-\left(\frac{\left(x(1) - x_0(1)\right)^2}{\sigma_1^2} + \frac{\left(x(2) - x_0(2)\right)^2}{\sigma_2^2} \right) \right)$$

$$\mathcal{L}(f, x_0, g(\theta)) = \sum_{x \in \pi(x_0)} k(x, x_0)(f(x) - g(x))^2 + \lambda \sum_i \theta_i$$

$\text{argmin}_{g \in G} \mathcal{L}(f, g, x_0)$

G is the family of linear functions

How do you choose σ_1, σ_2?



$$k(x, x_0) = \exp\left(-\left(\frac{\left(x(1) - x_0(1)\right)^2}{\sigma_1^2} + \frac{\left(x(2) - x_0(2)\right)^2}{\sigma_2^2} \right) \right)$$

$$\mathscr{L}(f, x_0, g(\theta)) = \sum_{x \in \pi(x_0)} k(x, x_0)(f(x) - g(x))^2 + \lambda \sum_i \theta_i$$

$\text{argmin}_{g \in G} \mathscr{L}(f, g, x_0)$

G is the family of linear functions

Influence Functions

- Find training data points that are most influential on the prediction for a test point

Understanding Black-box Predictions via Influence Functions

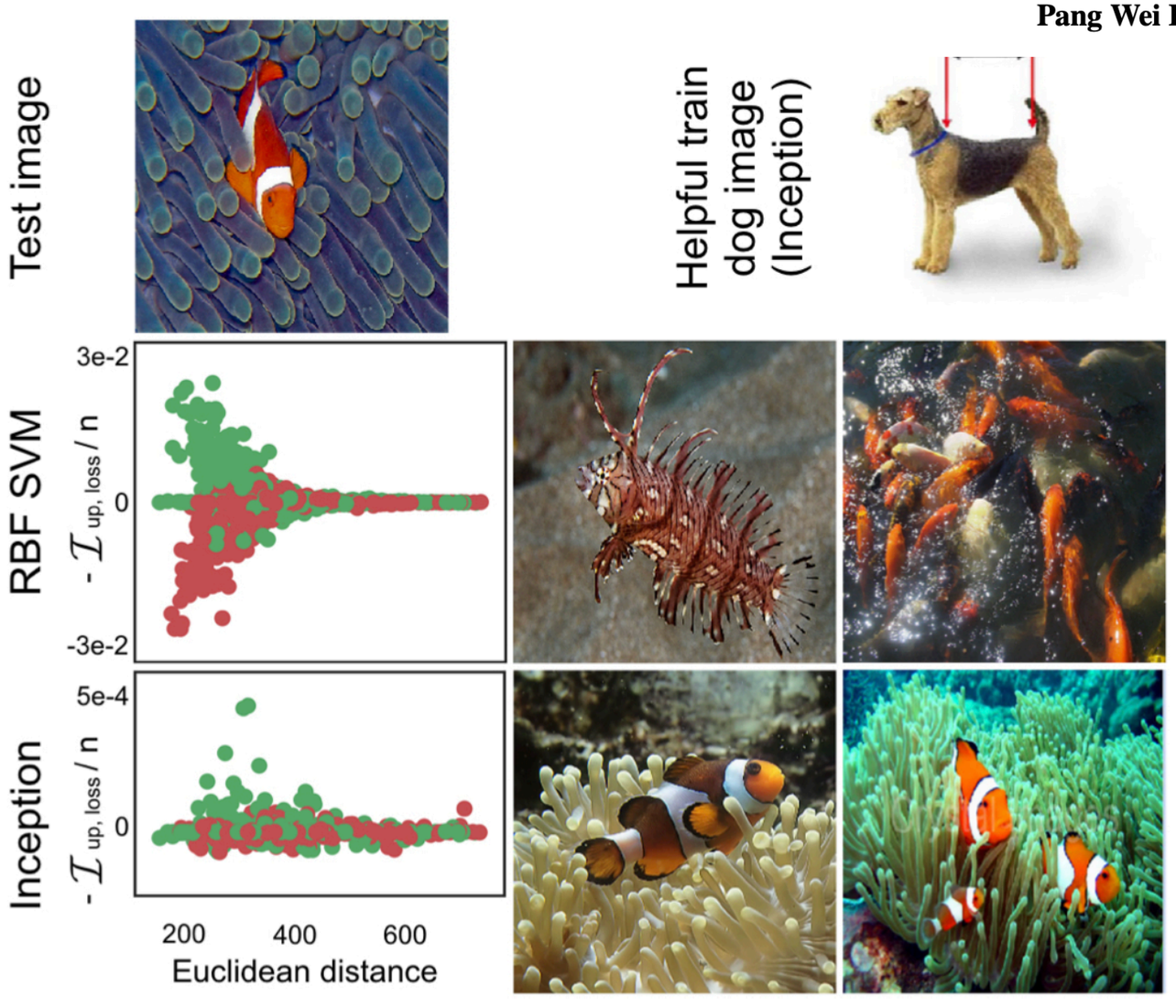
Pang Wei Koh [1] **Percy Liang** [1]

Influence Functions

- Find training data points that are most influential on the prediction for a test point

Understanding Black-box Predictions via Influence Functions

Pang Wei Koh [1] Percy Liang [1]



Mechanistic Interpretability

Which parts of the model get consistently activated in response to a concept?

Feature #34M/31164353 **Golden Gate Bridge** feature example

The feature activates strongly on English descriptions and associated concepts

in the Presidio at the end (that's the
huge park right next to the Golden Gate
bridge), perfect. But not all people

repainted, roughly, every dozen years."
"while across the country in san fran
cisco, the golden gate bridge was

it is a suspension bridge and has similar
coloring, it is often> compared to the
Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋は、ア
メリカ西海岸のサンフランシスコ湾と太平洋が
接続するゴールデンゲート海

골든게이트 교 또는 금문교 는 미국 캘리포니아주
골든게이트 해협에 위치한 현수교이다. 골든게이
트 교는 캘리포니아주 샌프란시

мост золотые ворота — висячий мост
через пролив золотые ворота. он со
единяет город сан-фран

And on relevant images as well

https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html