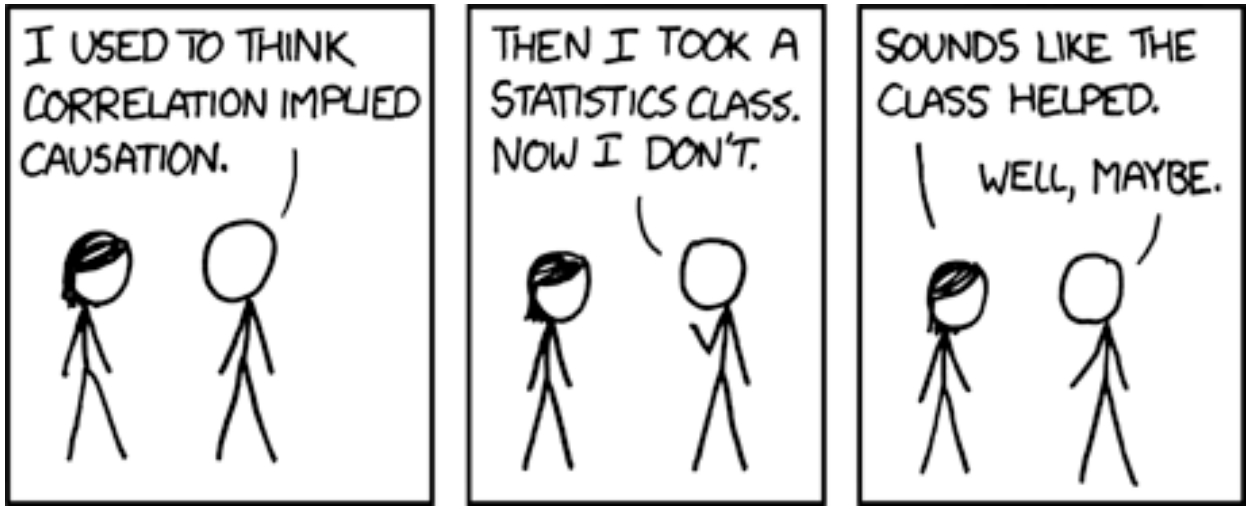
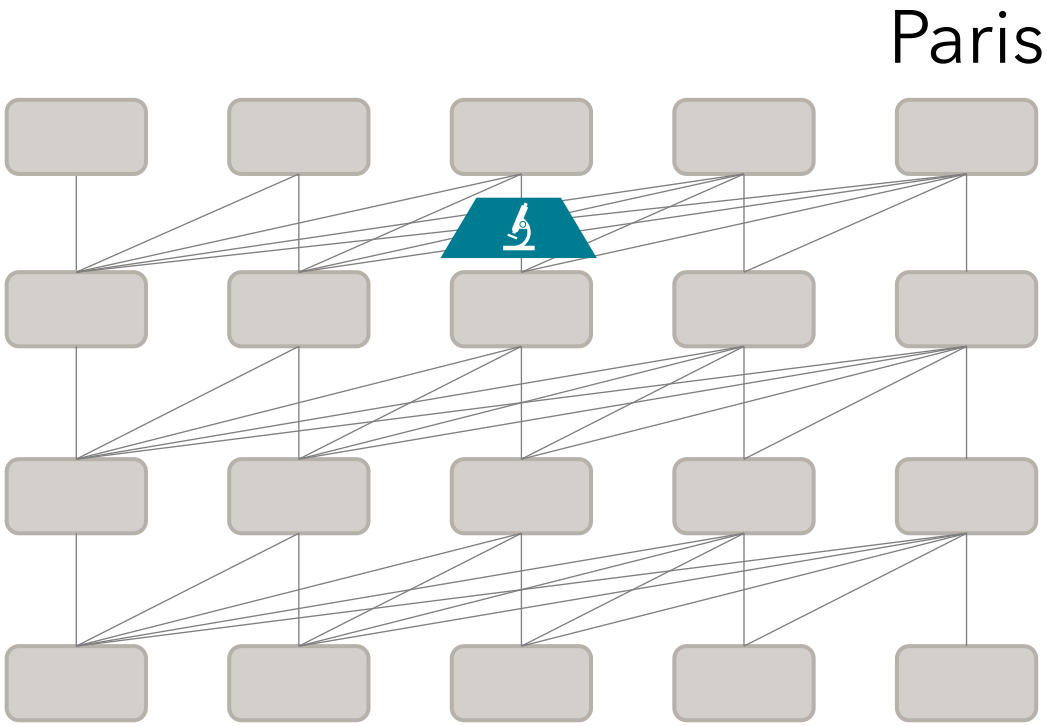
interventions surgically editing the internals of neural networks

CS 221M
Week 3, Lecture 5



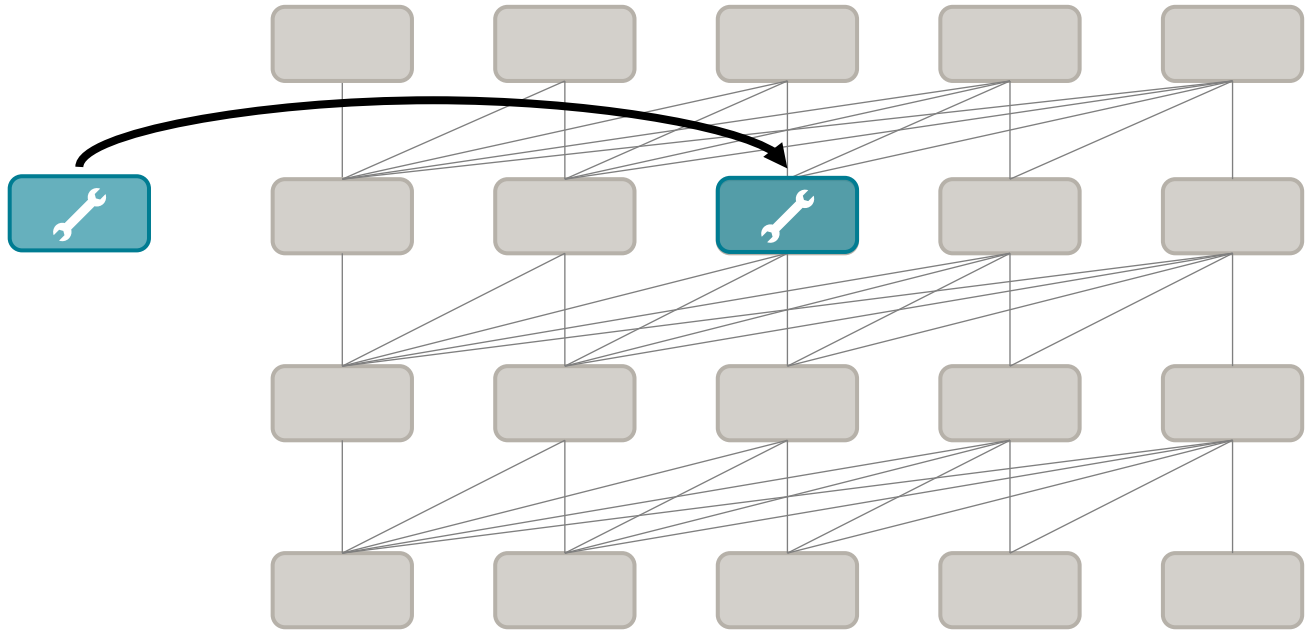
Previously: "reading" activations

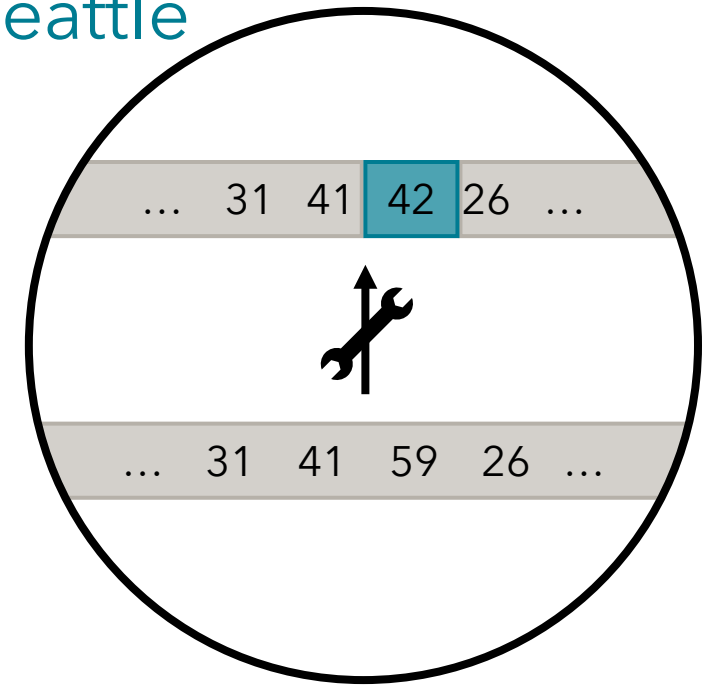


Goal: read information from inner workings of neural network

Today: "editing" activations

~~Paris~~ Seattle





the Eiffel Tower is in

Goal: edit information within inner workings of neural network

Surveying the literature

1

Interventions for removing information

2

Interventions for identifying causes

3

Interventions for controlling behavior

Interventions for removing information

iterative nullspace projection

Surveying the literature

1

Interventions for removing information

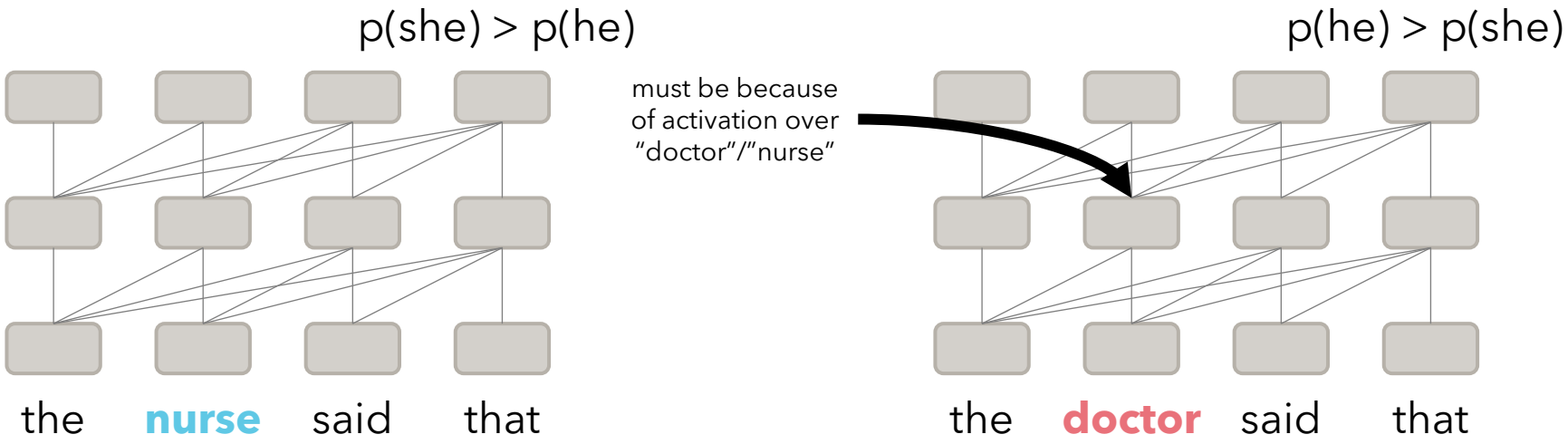
2

Interventions for identifying causes

3

Interventions for controlling behavior

motivating example: model bias

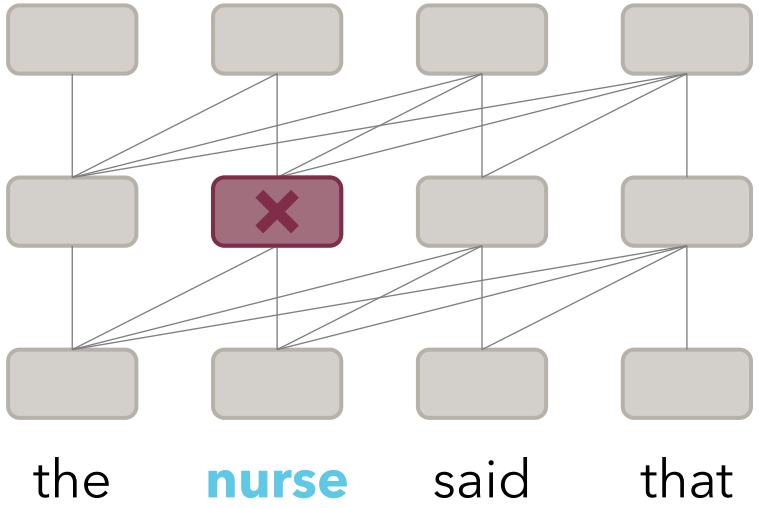


minimal pair: occupation changes predicted pronoun

goal: remove bias from activation

p(she) = p(he)

~~p(she) > p(he)~~

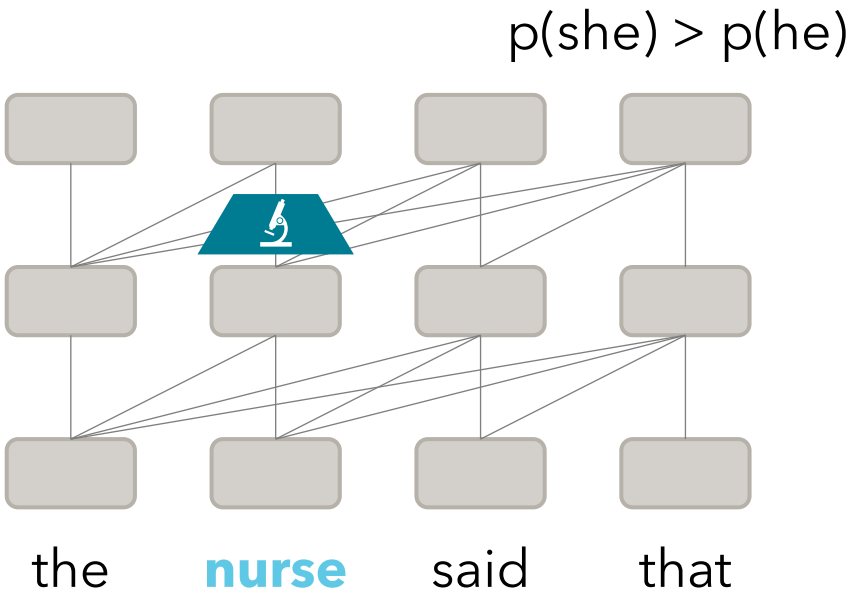


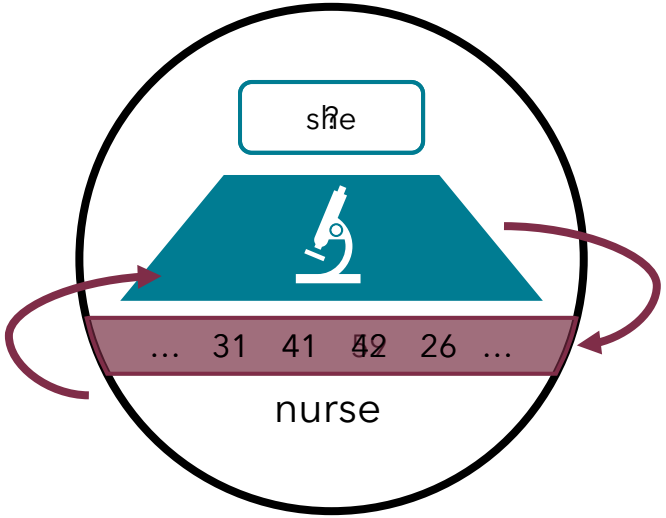
more recently, focus on harmful behavior (e.g., sycophancy) and verbatim memorization (e.g., for copyright)

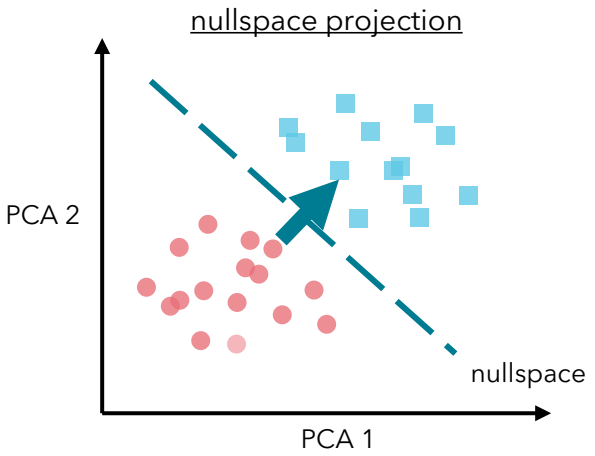
iterative nullspace projection (INLP

Goal: fool all probes

1. Train probe to predict "he" vs. "she"
2. Project activation to probe's null-space
3. Repeat with a new probe



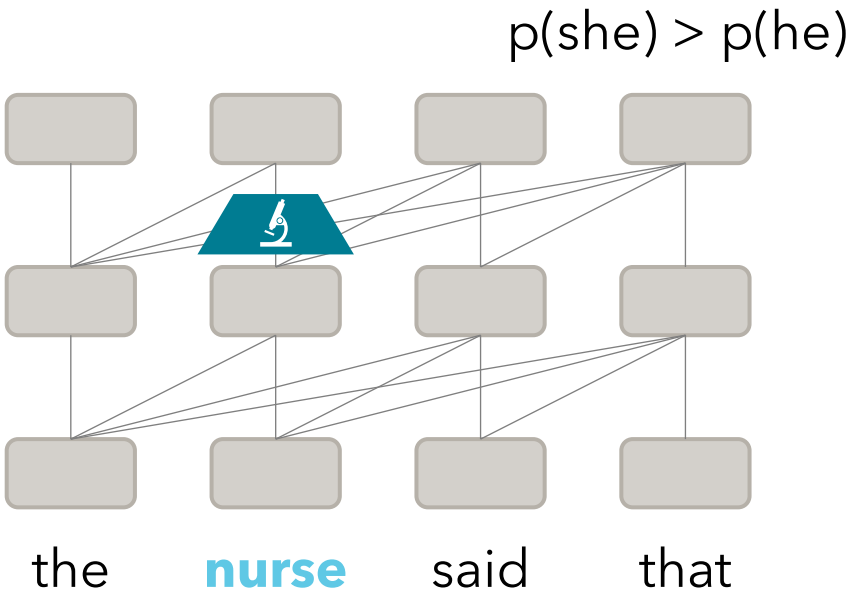


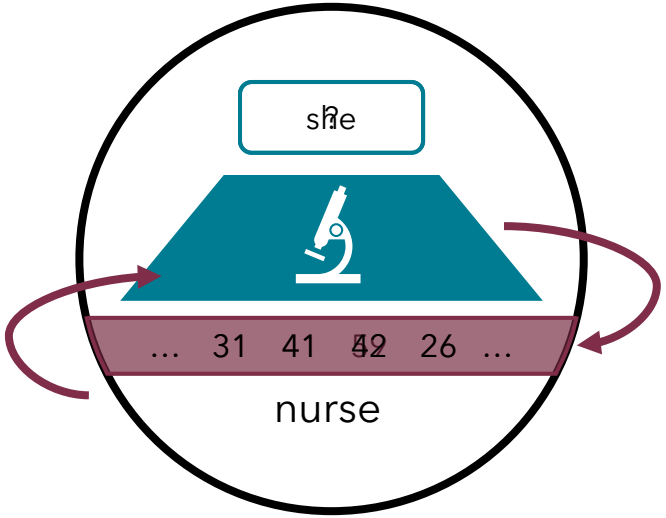


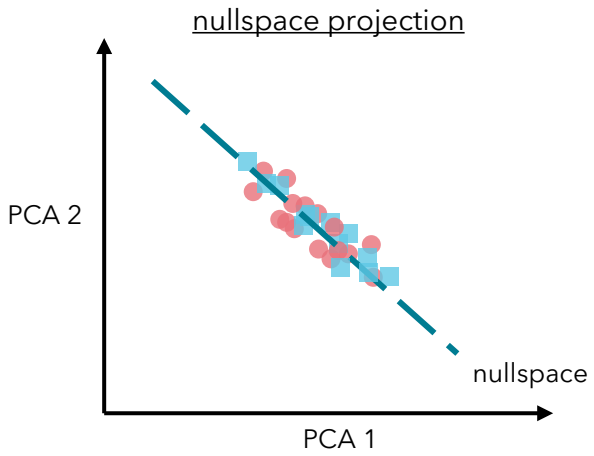
iterative nullspace projection (INLP

Goal: fool all probes

1. Train probe to predict "he" vs. "she"
2. Project activation to probe's null-space
3. Repeat with a new probe

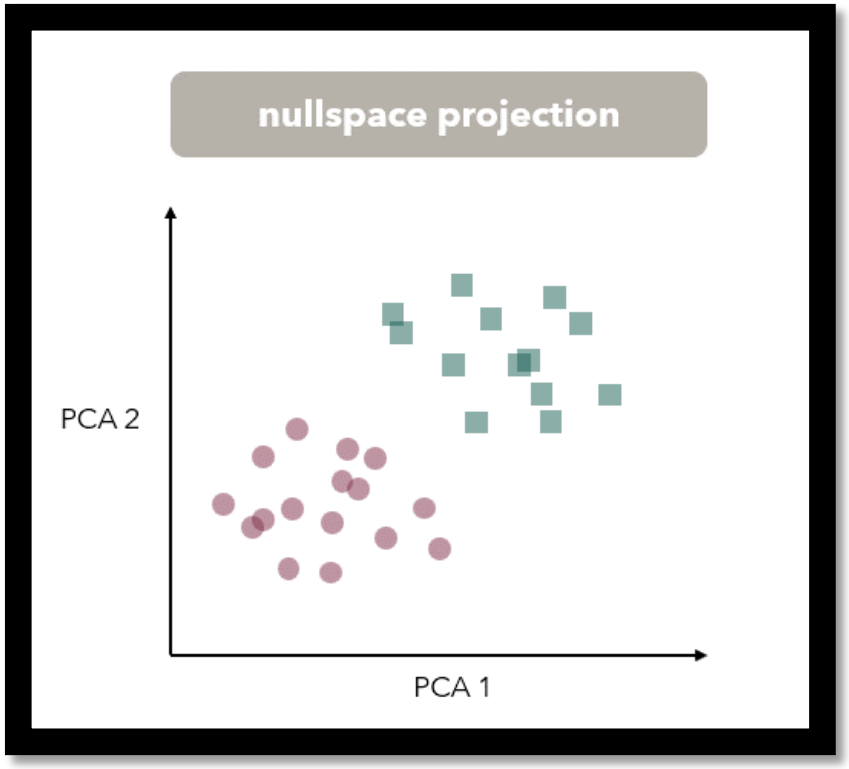




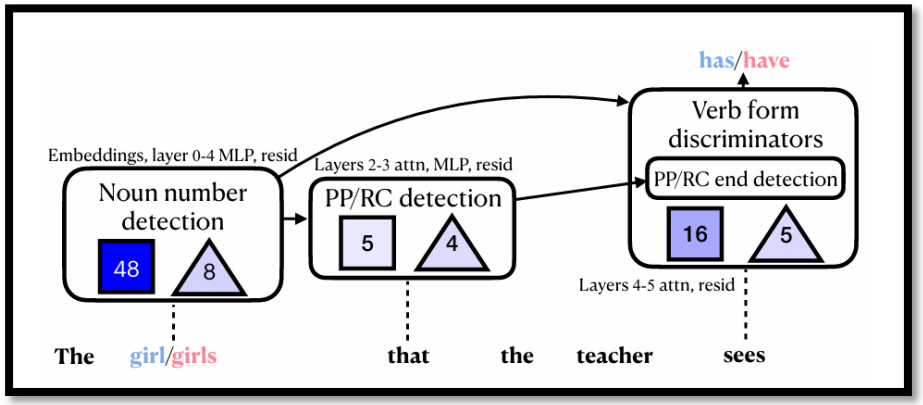


iterative nullspace projection (INLP)

code exercise

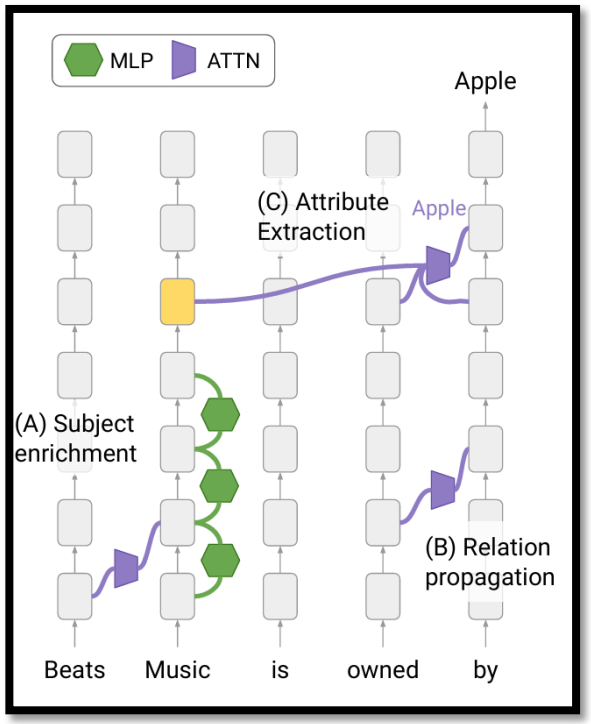


types of ablation



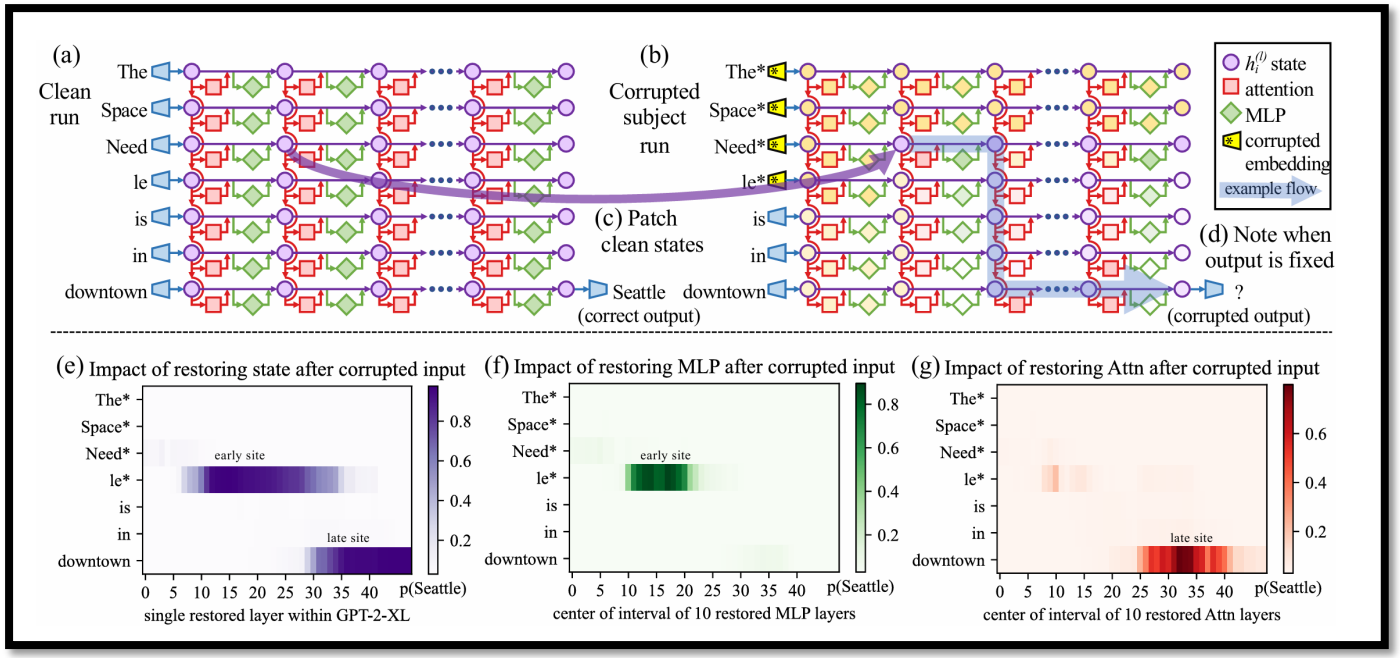
Mean ablation
set value to mean of activations

- Marks et al. 2024 sparse feature circuits
- Li et al. 2023 circuit breaking



Zero ablation
set activation value to 0

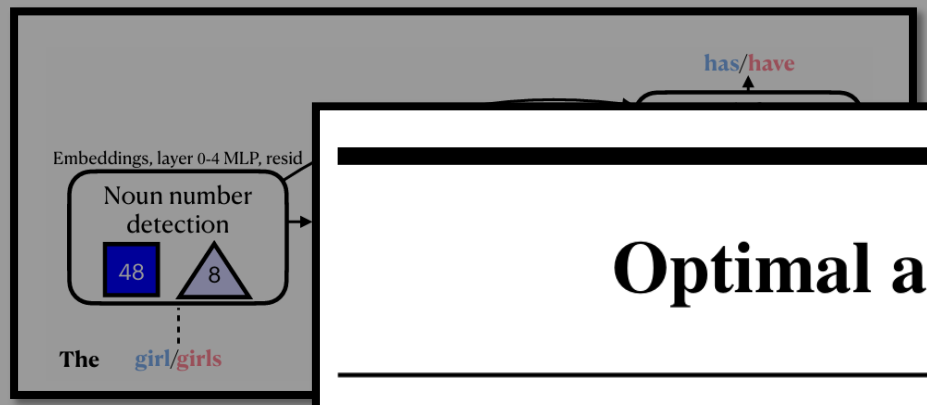
- Geva et al. 2023 attention knockout
- Merullo et al. 2023 LMs implement word2vec
- Gurnee et al. 2024 universal neurons in gpt2



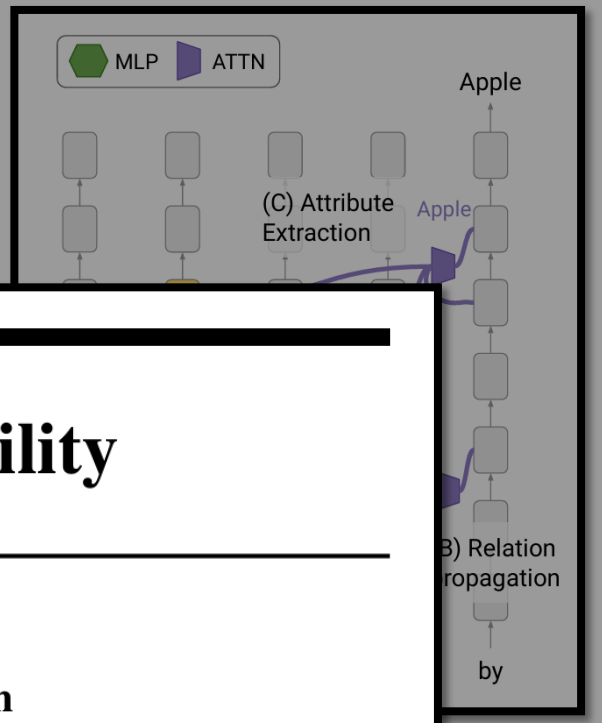
Noise ablation
add random noise to activation

- Meng et al. 2022 editing facts

types of ablation



Mean ablation
set value to mean of activations

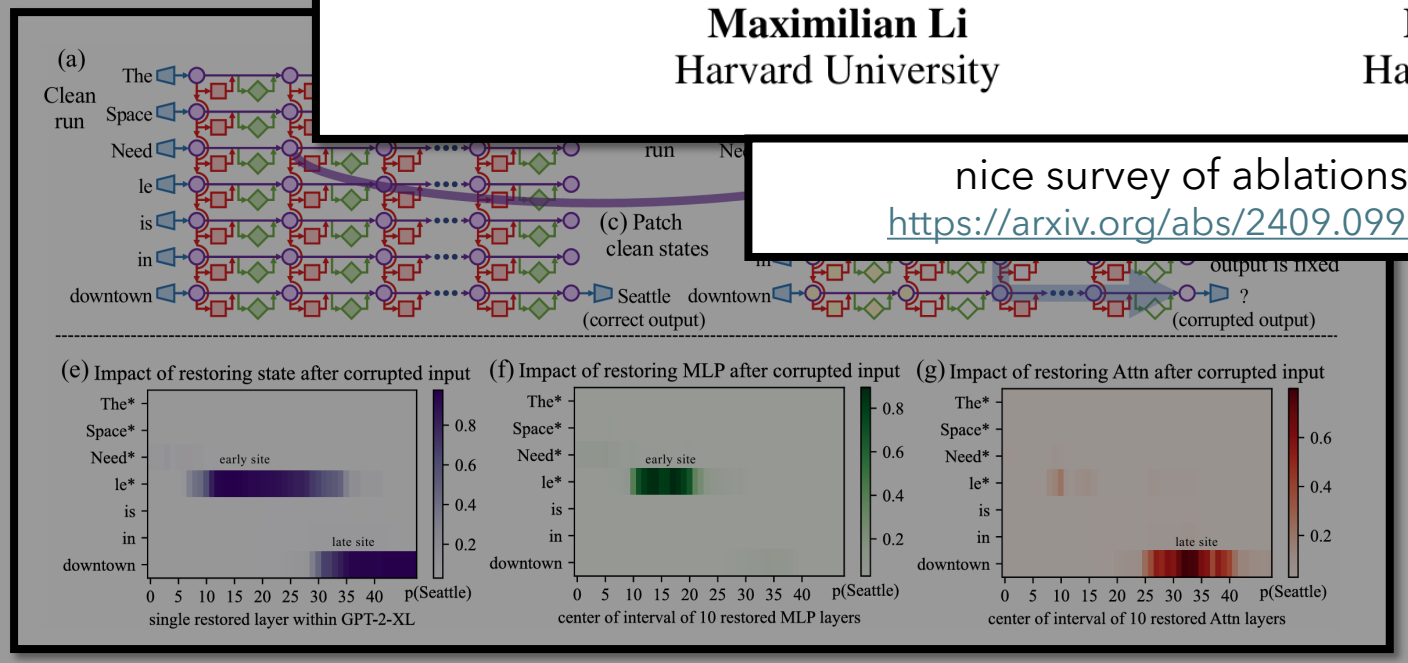


Optimal ablation for interpretability

Maximilian Li
Harvard University

Lucas Janson
Harvard University

nice survey of ablations!
https://arxiv.org/abs/2409.09951v1



- Geva et al. 2023 attention knockout
- Merullo et al. 2023 LMs implement word2vec
- Gurnee et al. 2024 universal neurons in gpt2

Noise ablation
add random noise to activation

- Meng et al. 2022 editing facts

Interventions for identifying causally relevant factors

interchange interventions

Surveying the literature

1

Interventions for removing information

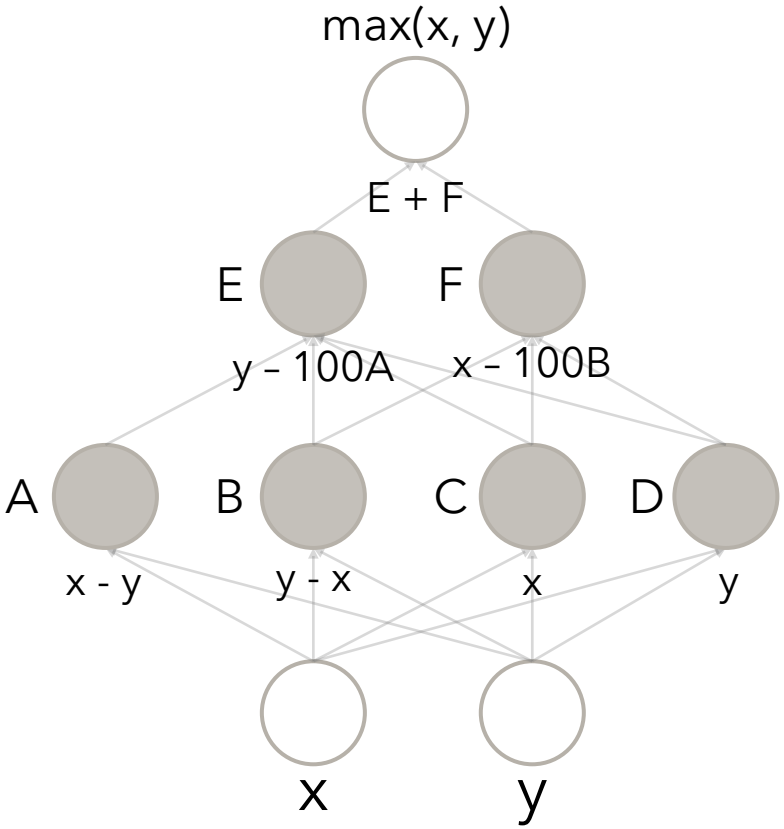
2

Interventions for identifying causes

3

Interventions for controlling behavior

Non-linear representations



goal

$$\max(x, y) = \begin{cases} x & x > y \\ y & \text{otherwise} \end{cases}$$

implementation

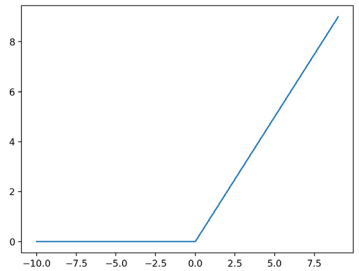
$$h_1 = \text{ReLU}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} [x \quad y]\right)$$

$$h_2 = \text{ReLU}\left(\begin{bmatrix} -100 & 0 & 0 & 1 \\ 0 & -100 & 1 & 0 \end{bmatrix} h_1\right)$$

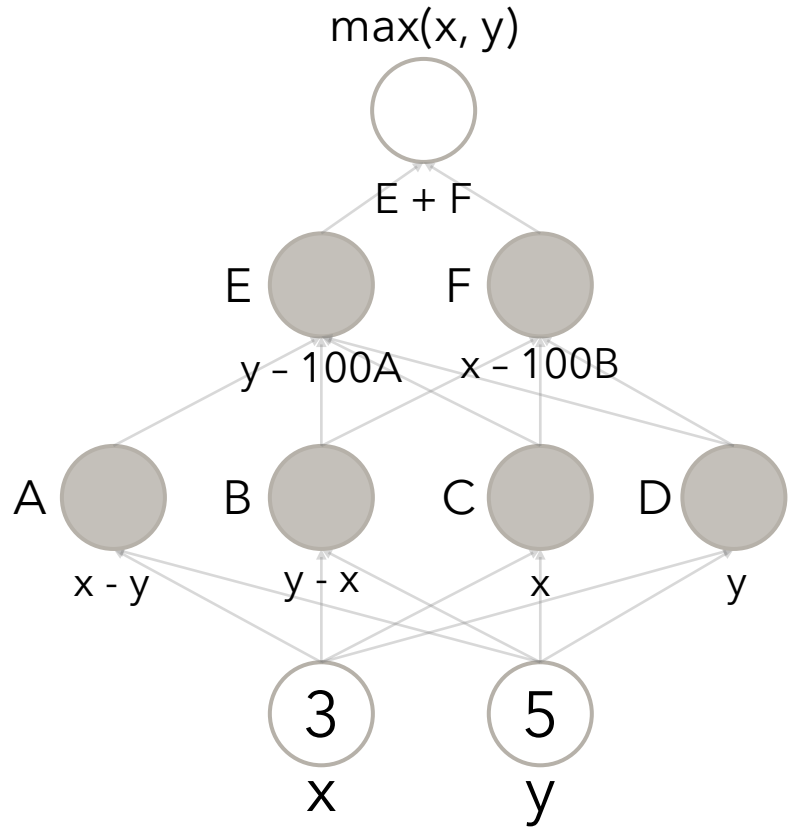
$$\text{output} = h_3 = [1 \quad 1] h_2$$

reference

$$\text{ReLU}(n) = \begin{cases} 0 & n < 0 \\ n & \text{otherwise} \end{cases}$$



Non-linear representations



goal

$$\max(x, y) = \begin{cases} x & x > y \\ y & \text{otherwise} \end{cases}$$

implementation

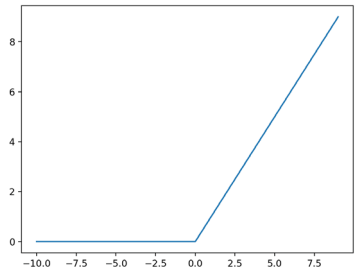
$$h_1 = \text{ReLU}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} [x \quad y]\right)$$

$$h_2 = \text{ReLU}\left(\begin{bmatrix} -100 & 0 & 0 & 1 \\ 0 & -100 & 1 & 0 \end{bmatrix} h_1\right)$$

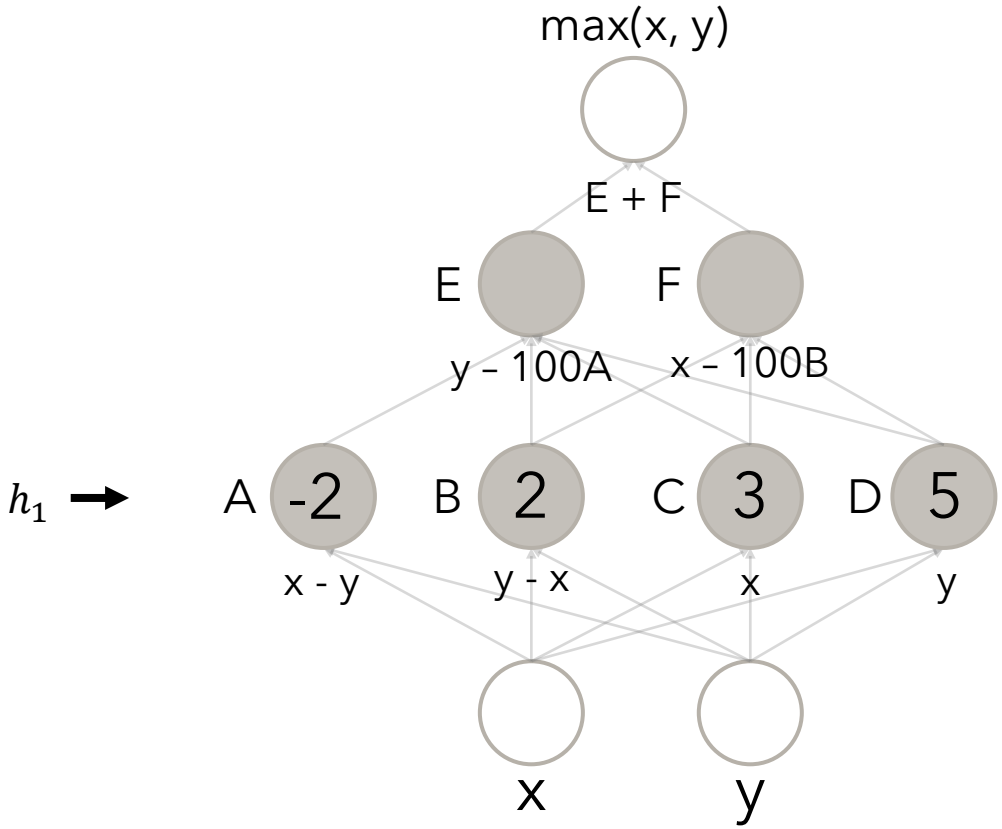
$$\text{output} = h_3 = [1 \quad 1] h_2$$

reference

$$\text{ReLU}(n) = \begin{cases} 0 & n < 0 \\ n & \text{otherwise} \end{cases}$$



Non-linear representations



goal

$$\max(x, y) = \begin{cases} x & x > y \\ y & \text{otherwise} \end{cases}$$

implementation

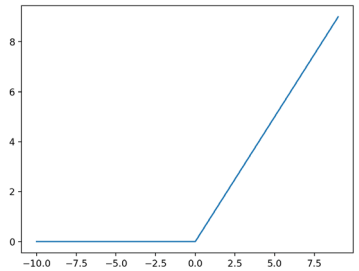
$$h_1 = \text{ReLU}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} [x \quad y]\right)$$

$$h_2 = \text{ReLU}\left(\begin{bmatrix} -100 & 0 & 0 & 1 \\ 0 & -100 & 1 & 0 \end{bmatrix} h_1\right)$$

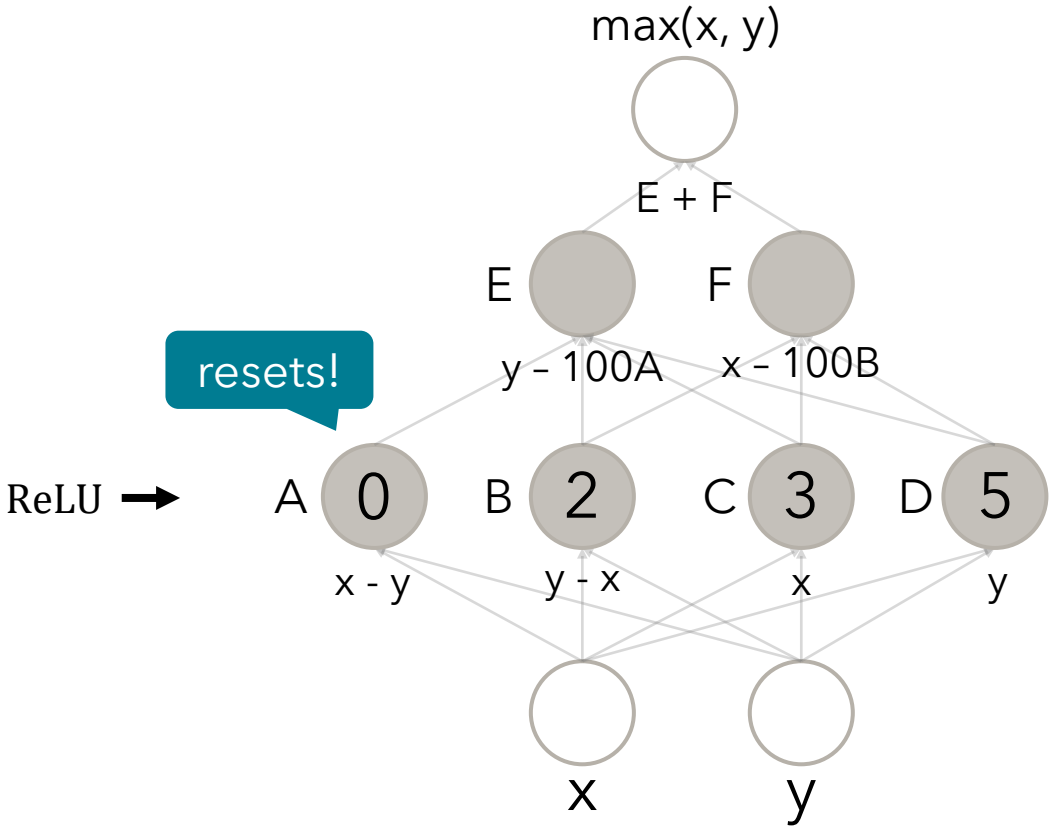
$$\text{output} = h_3 = [1 \quad 1] h_2$$

reference

$$\text{ReLU}(n) = \begin{cases} 0 & n < 0 \\ n & \text{otherwise} \end{cases}$$



Non-linear representations



goal

$$\max(x, y) = \begin{cases} x & x > y \\ y & \text{otherwise} \end{cases}$$

implementation

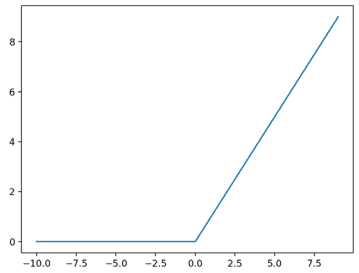
$$h_1 = \text{ReLU}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} [x \quad y]\right)$$

$$h_2 = \text{ReLU}\left(\begin{bmatrix} -100 & 0 & 0 & 1 \\ 0 & -100 & 1 & 0 \end{bmatrix} h_1\right)$$

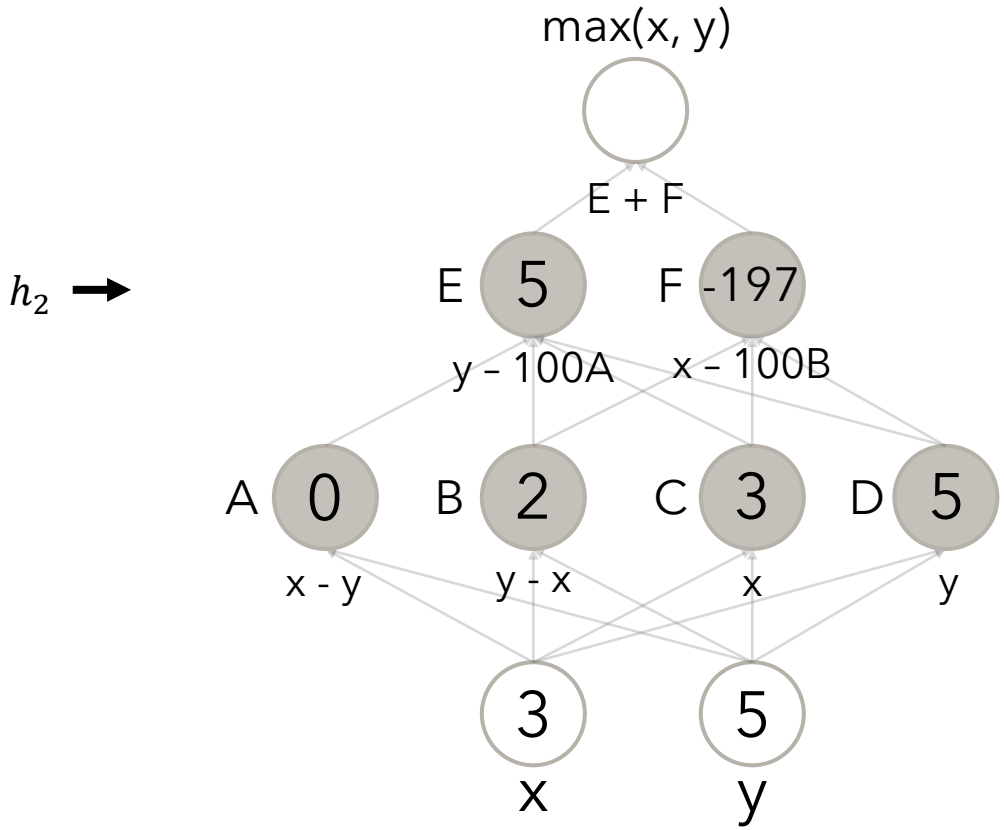
$$\text{output} = h_3 = [1 \quad 1] h_2$$

reference

$$\text{ReLU}(n) = \begin{cases} 0 & n < 0 \\ n & \text{otherwise} \end{cases}$$



Non-linear representations



goal

$$\max(x, y) = \begin{cases} x & x > y \\ y & \text{otherwise} \end{cases}$$

implementation

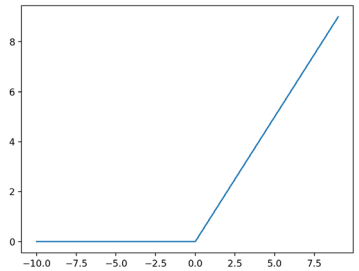
$$h_1 = \text{ReLU}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} [x \quad y]\right)$$

$$h_2 = \text{ReLU}\left(\begin{bmatrix} -100 & 0 & 0 & 1 \\ 0 & -100 & 1 & 0 \end{bmatrix} h_1\right)$$

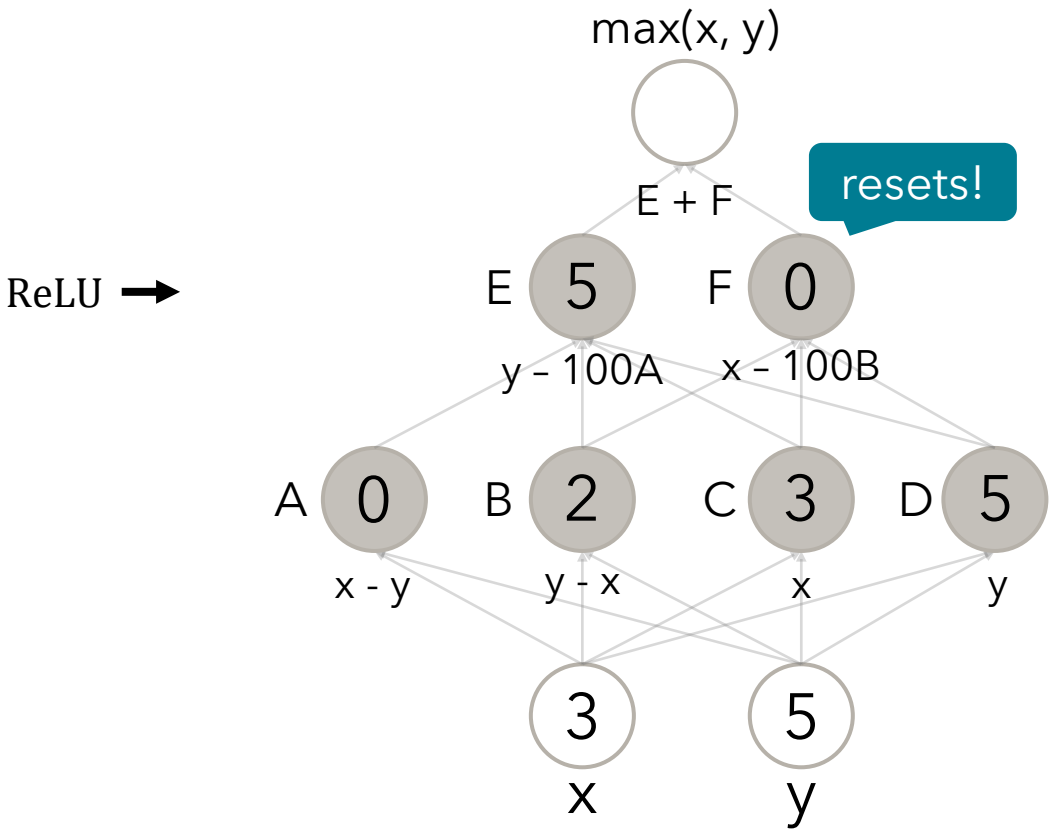
$$\text{output} = h_3 = [1 \quad 1] h_2$$

reference

$$\text{ReLU}(n) = \begin{cases} 0 & n < 0 \\ n & \text{otherwise} \end{cases}$$



Non-linear representations



goal

$$\max(x, y) = \begin{cases} x & x > y \\ y & \text{otherwise} \end{cases}$$

implementation

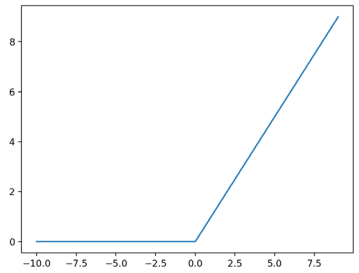
$$h_1 = \text{ReLU}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} [x \quad y]\right)$$

$$h_2 = \text{ReLU}\left(\begin{bmatrix} -100 & 0 & 0 & 1 \\ 0 & -100 & 1 & 0 \end{bmatrix} h_1\right)$$

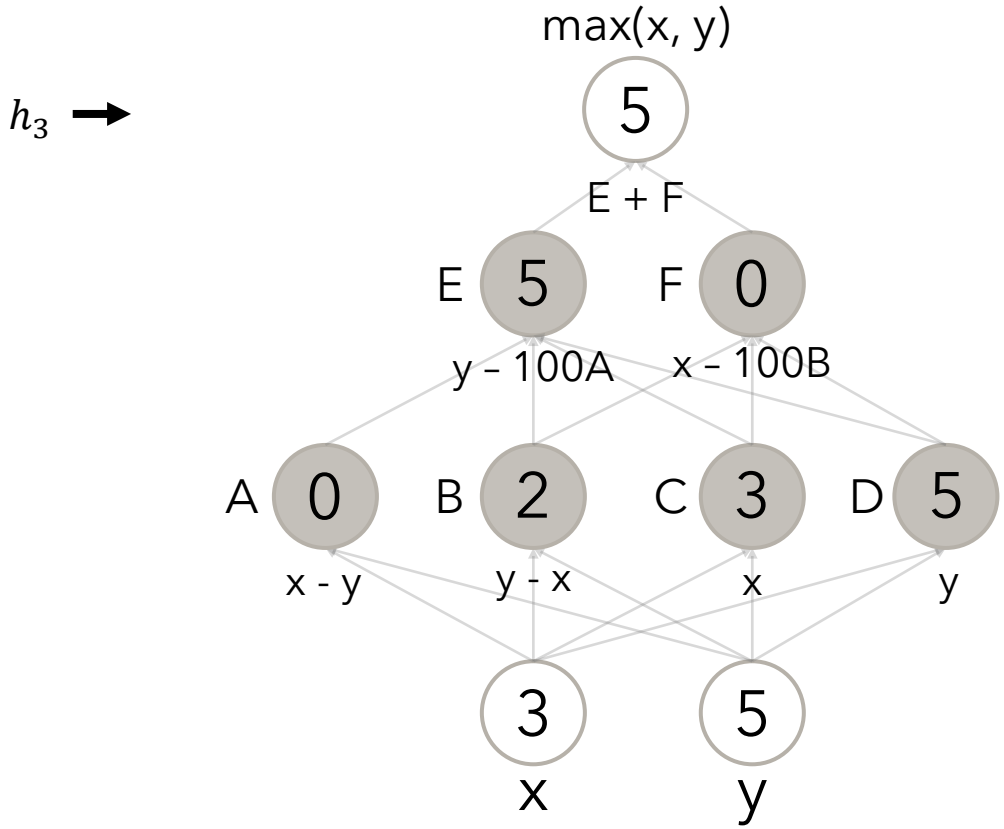
$$\text{output} = h_3 = [1 \quad 1] h_2$$

reference

$$\text{ReLU}(n) = \begin{cases} 0 & n < 0 \\ n & \text{otherwise} \end{cases}$$



Non-linear representations



goal

$$\max(x, y) = \begin{cases} x & x > y \\ y & \text{otherwise} \end{cases}$$

implementation

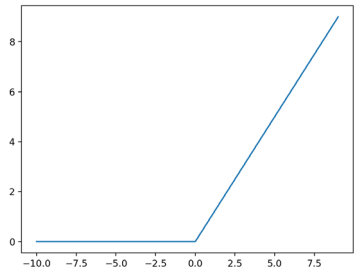
$$h_1 = \text{ReLU}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} [x \quad y]\right)$$

$$h_2 = \text{ReLU}\left(\begin{bmatrix} -100 & 0 & 0 & 1 \\ 0 & -100 & 1 & 0 \end{bmatrix} h_1\right)$$

$$\text{output} = h_3 = [1 \quad 1] h_2$$

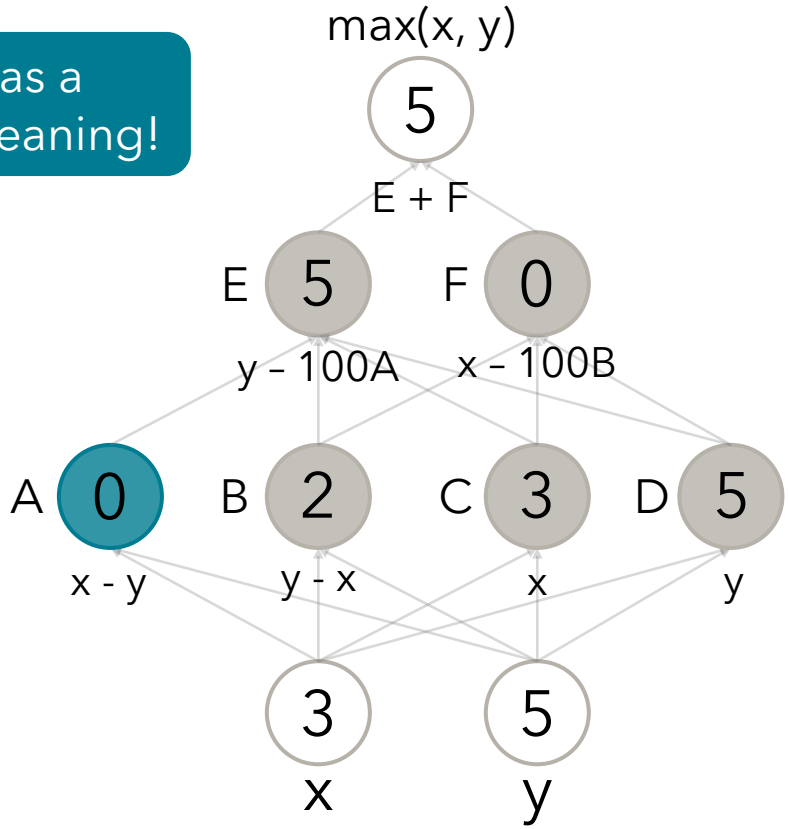
reference

$$\text{ReLU}(n) = \begin{cases} 0 & n < 0 \\ n & \text{otherwise} \end{cases}$$



Non-linear representations

zero has a special meaning!



goal

$$\max(x, y) = \begin{cases} x & x > y \\ y & \text{otherwise} \end{cases}$$

implementation

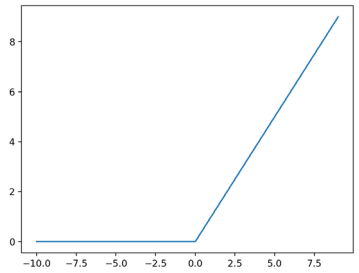
$$h_1 = \text{ReLU}\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} [x \quad y]\right)$$

$$h_2 = \text{ReLU}\left(\begin{bmatrix} -100 & 0 & 0 & 1 \\ 0 & -100 & 1 & 0 \end{bmatrix} h_1\right)$$

$$\text{output} = h_3 = [1 \quad 1] h_2$$

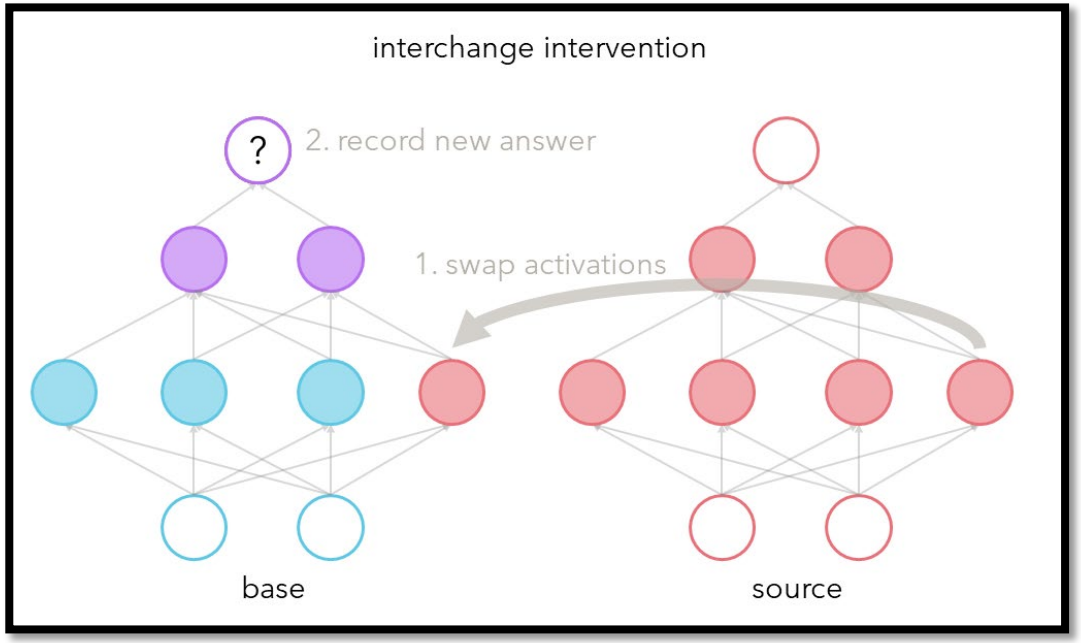
reference

$$\text{ReLU}(n) = \begin{cases} 0 & n < 0 \\ n & \text{otherwise} \end{cases}$$



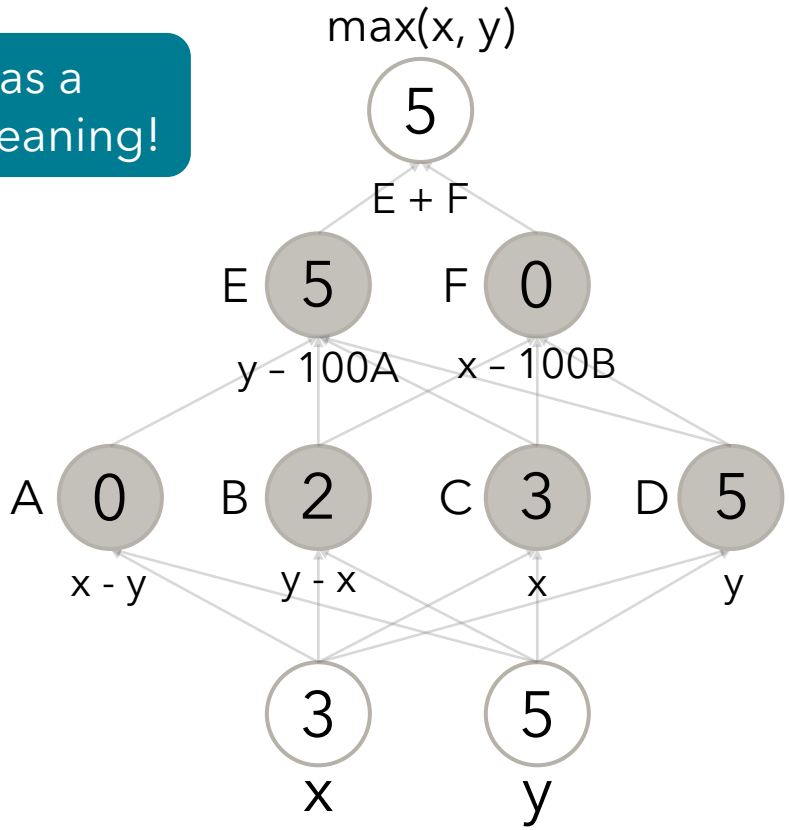
towards interchange interventions

code exercise



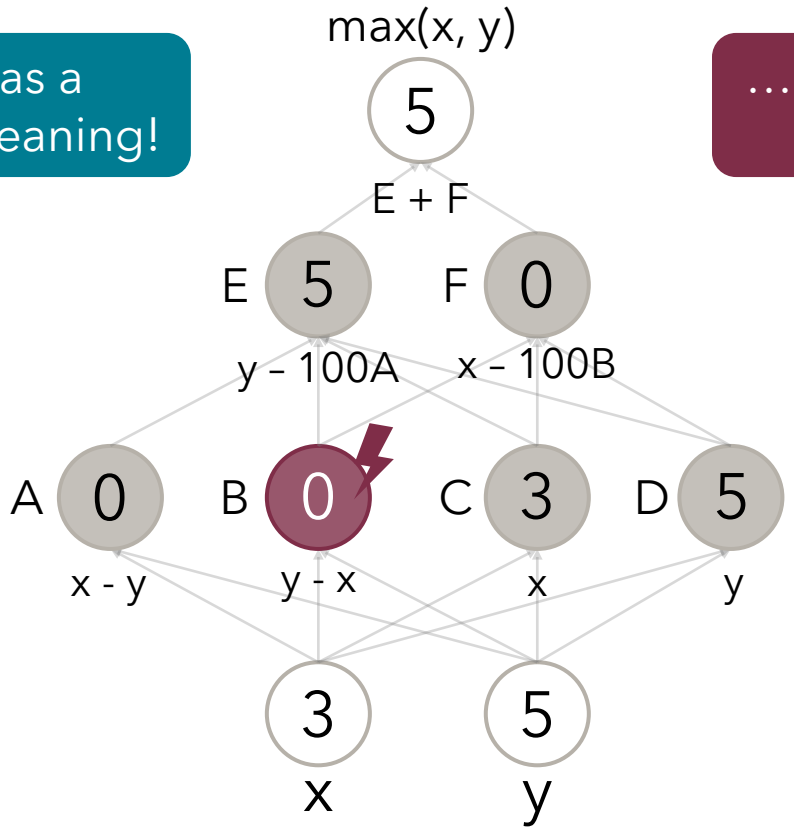
interventions can make assumptions!

zero has a special meaning!



interventions can make assumptions!

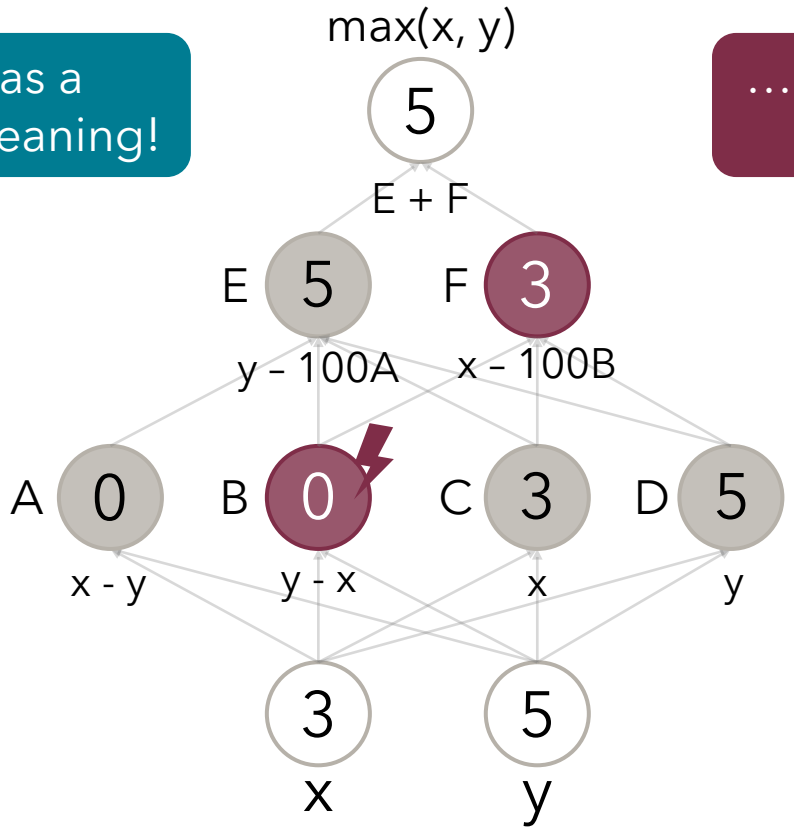
zero has a special meaning!



… so zero ablation gets weird

interventions can make assumptions!

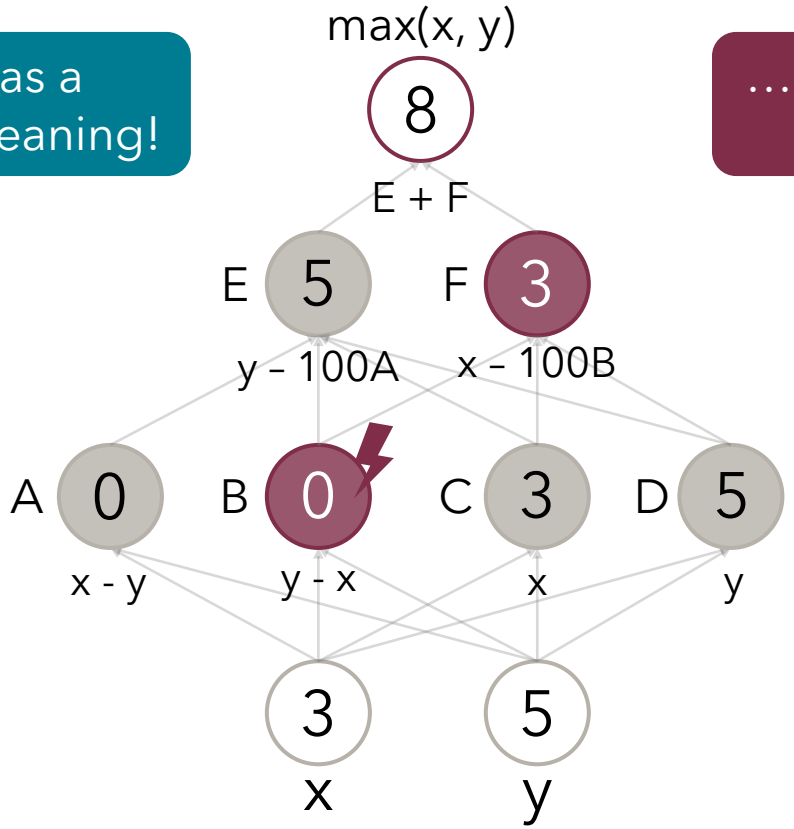
zero has a special meaning!



… so zero ablation gets weird

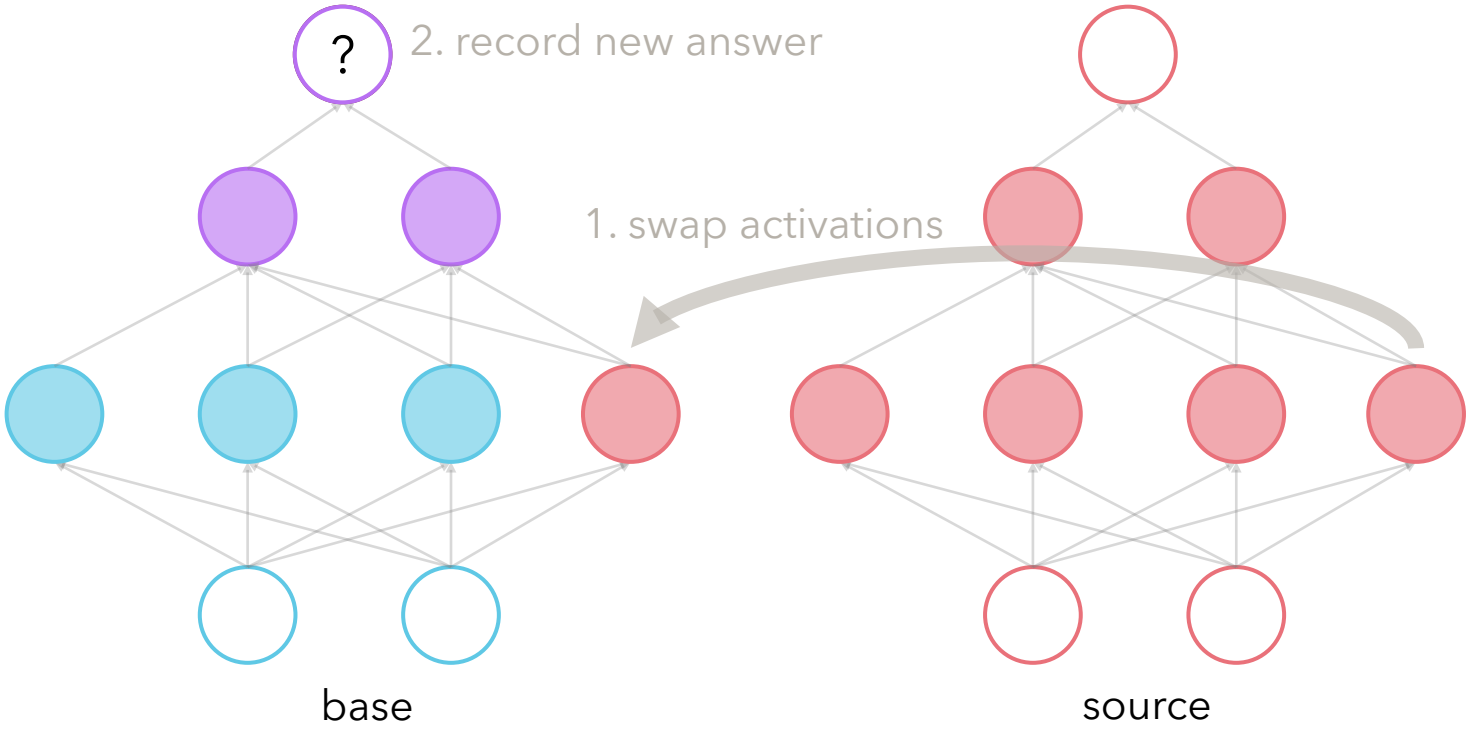
interventions can make assumptions!

zero has a special meaning!

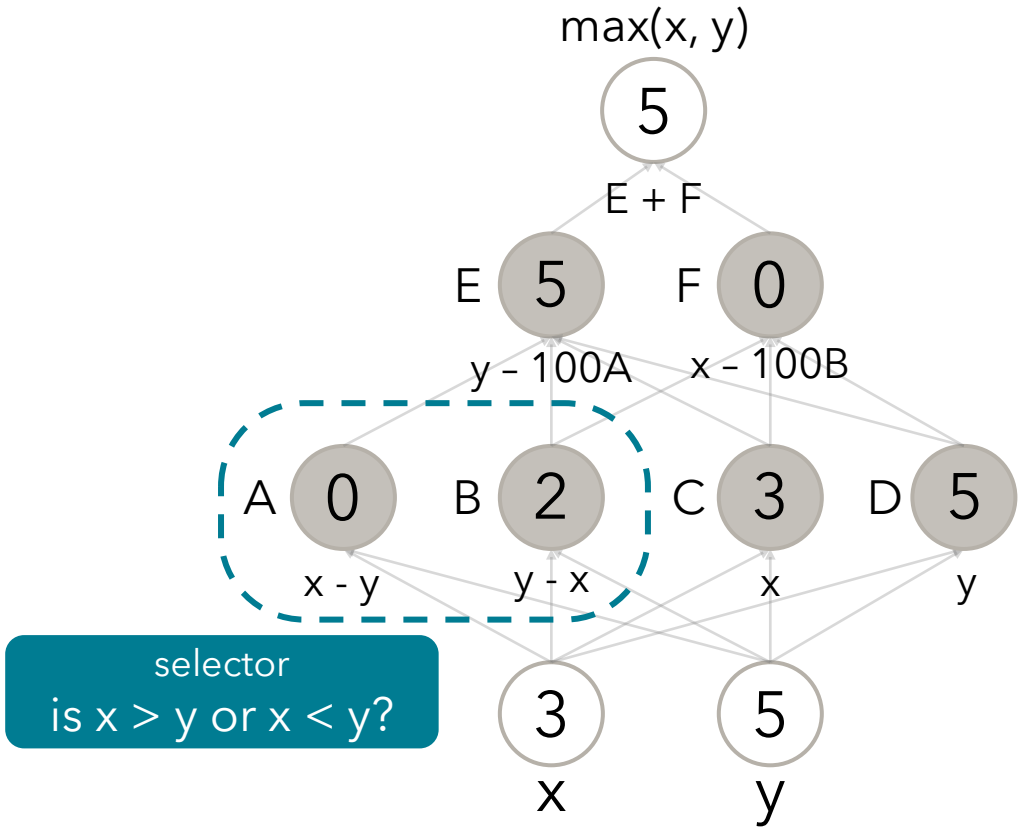


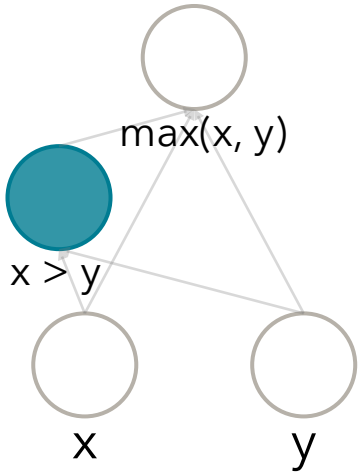
… so zero ablation gets weird

interchange interventions: swap values

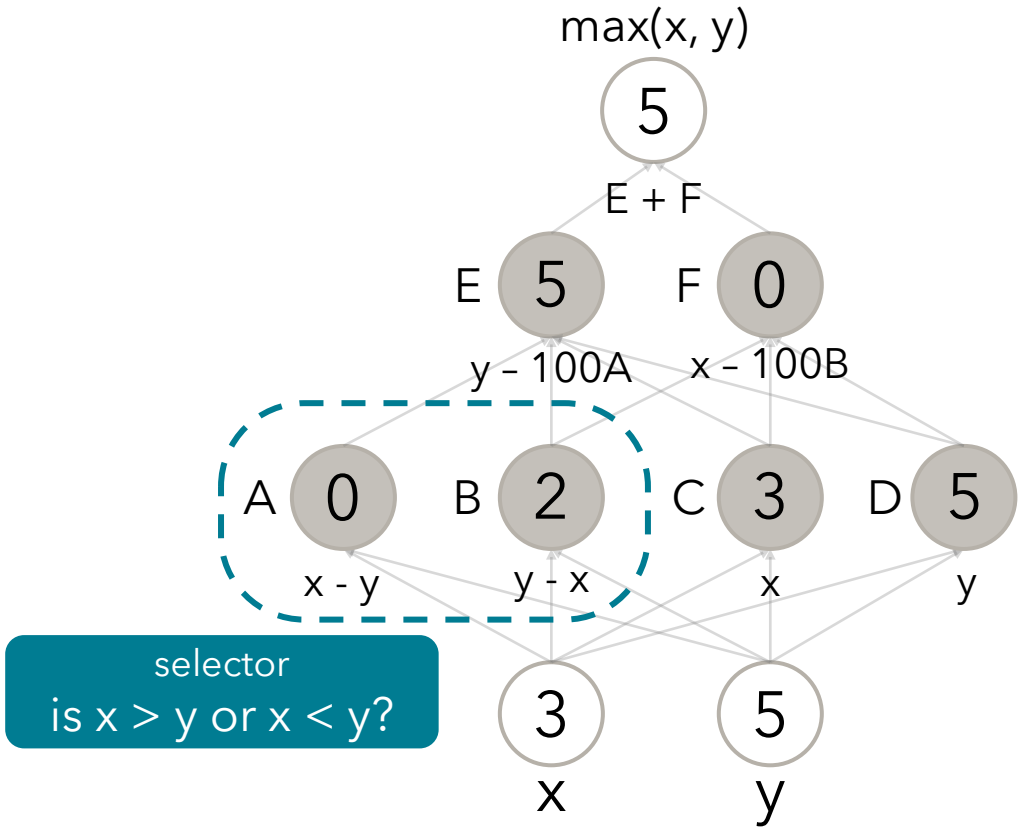


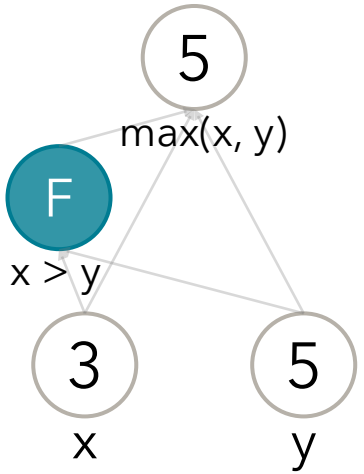
towards causal abstraction



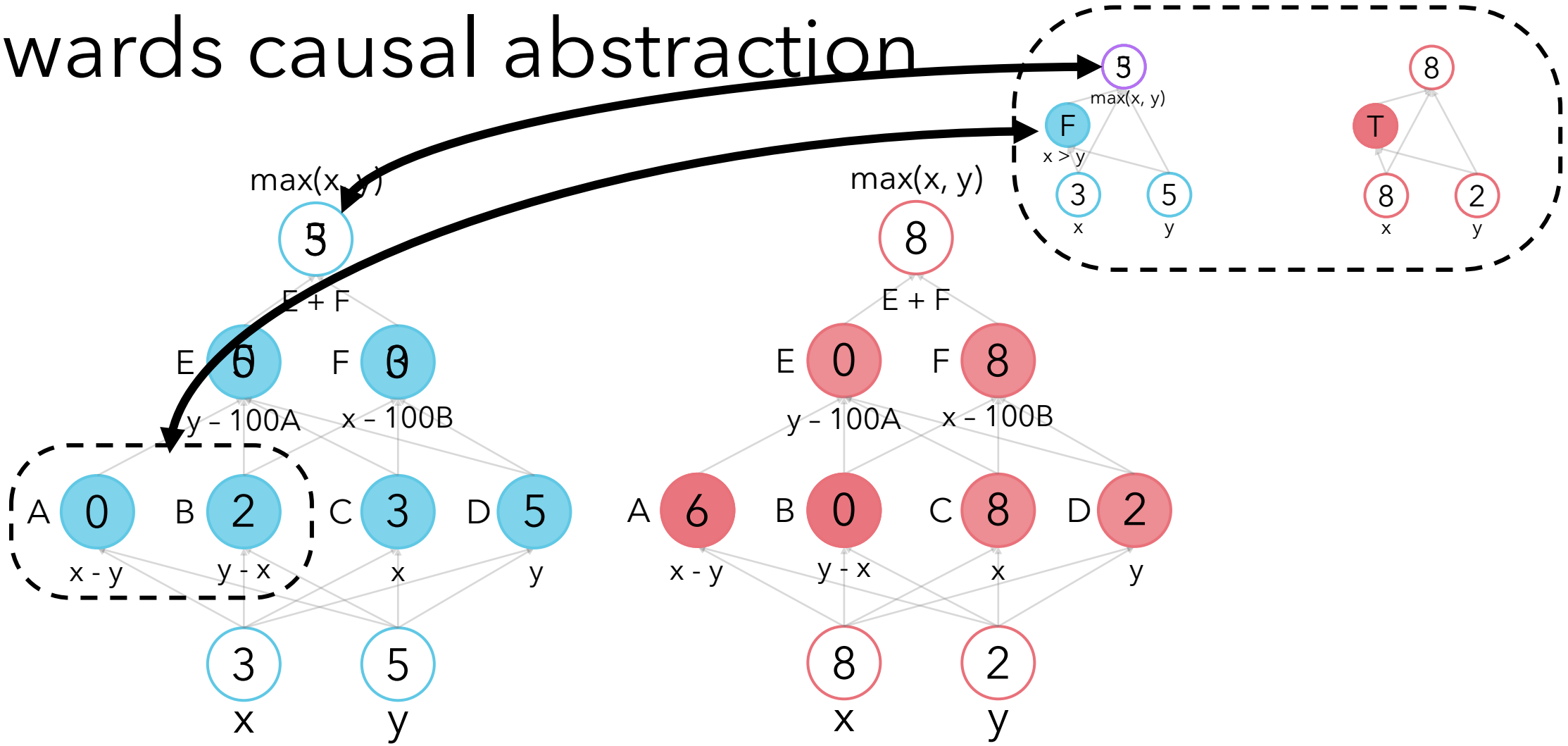


towards causal abstraction





towards causal abstraction



Supervised steering

what if we just cared about
controlling model behavior?

Surveying the literature

1

Interventions for removing information

2

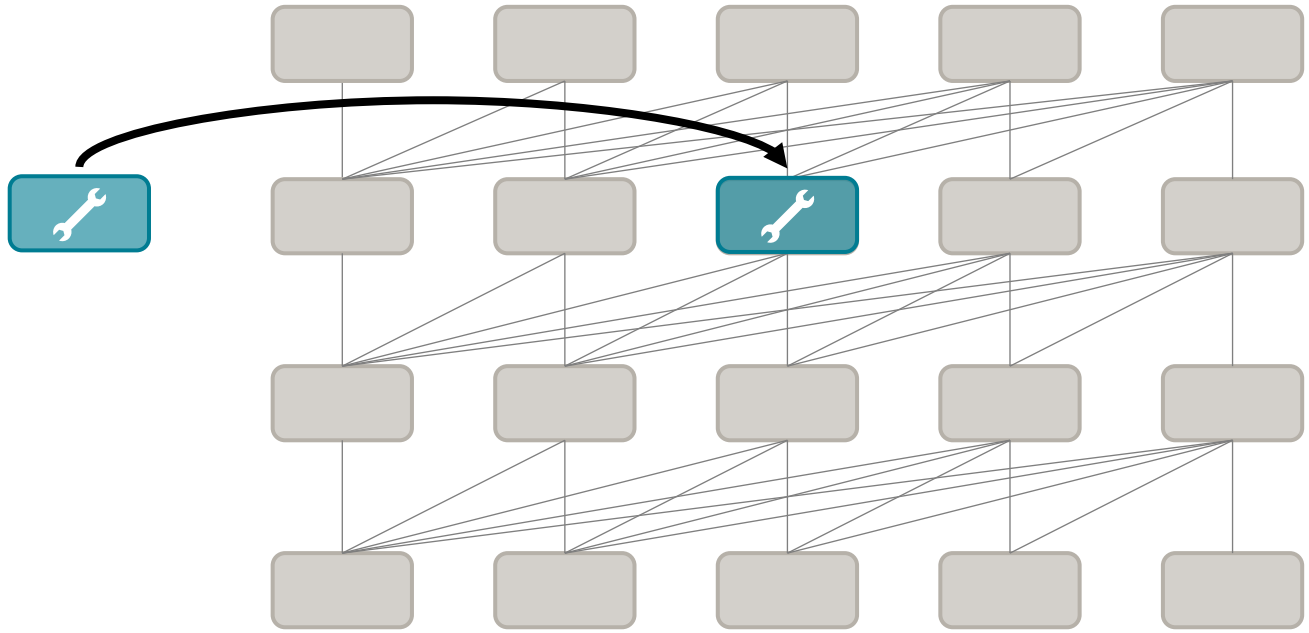
Interventions for identifying causes

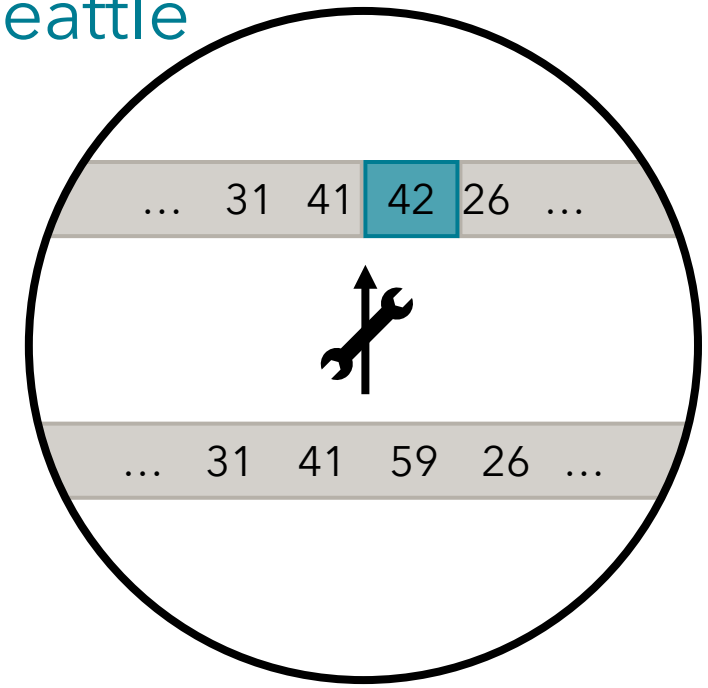
3

Interventions for controlling behavior

Today: "editing" activations

~~Paris~~ Seattle





the Eiffel Tower is in

Goal: edit information within inner workings of neural network

Surveying the literature

1

Interventions for removing information

2

Interventions for identifying causes

3

Interventions for controlling behavior