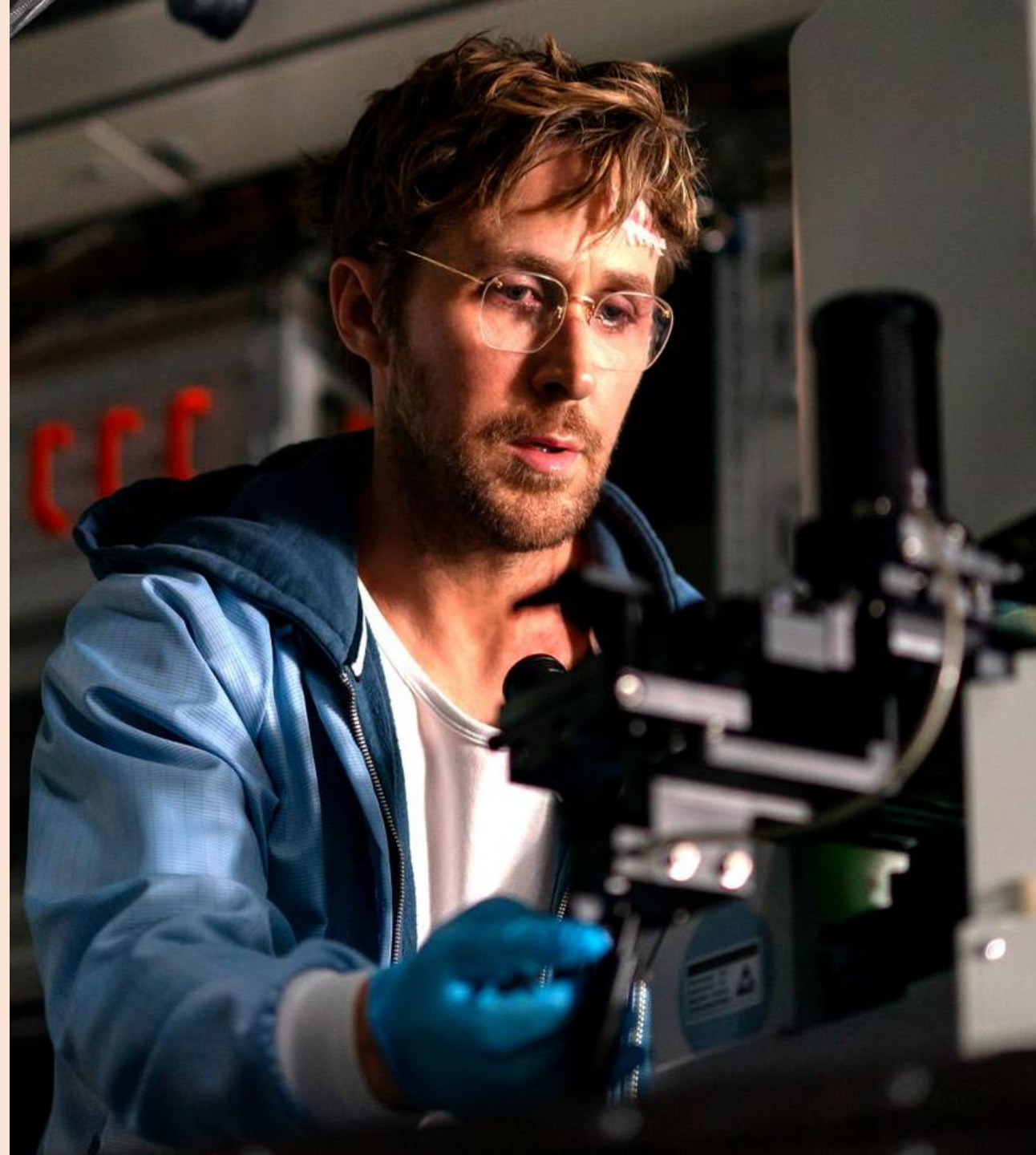
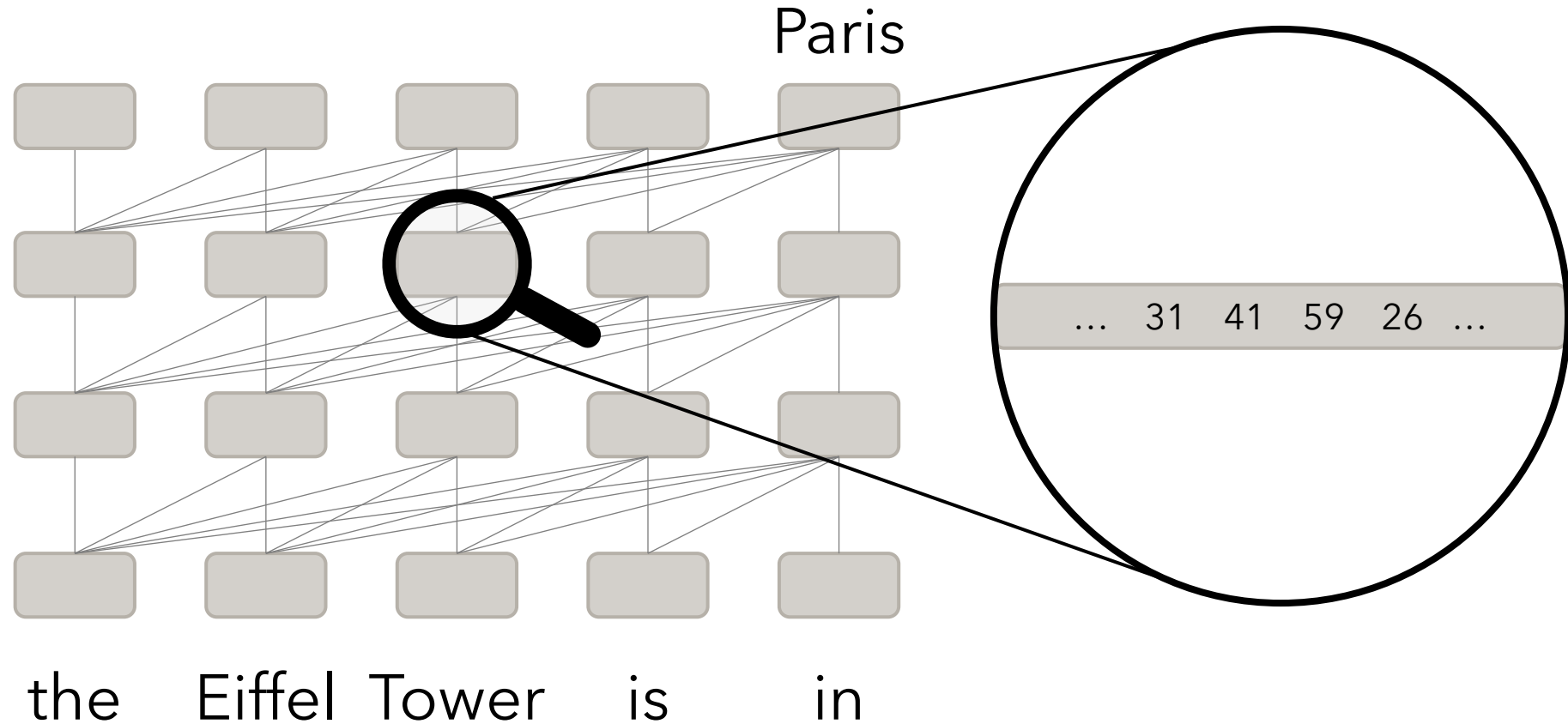


Probes  
inspecting the  
inner workings  
of neural  
networks

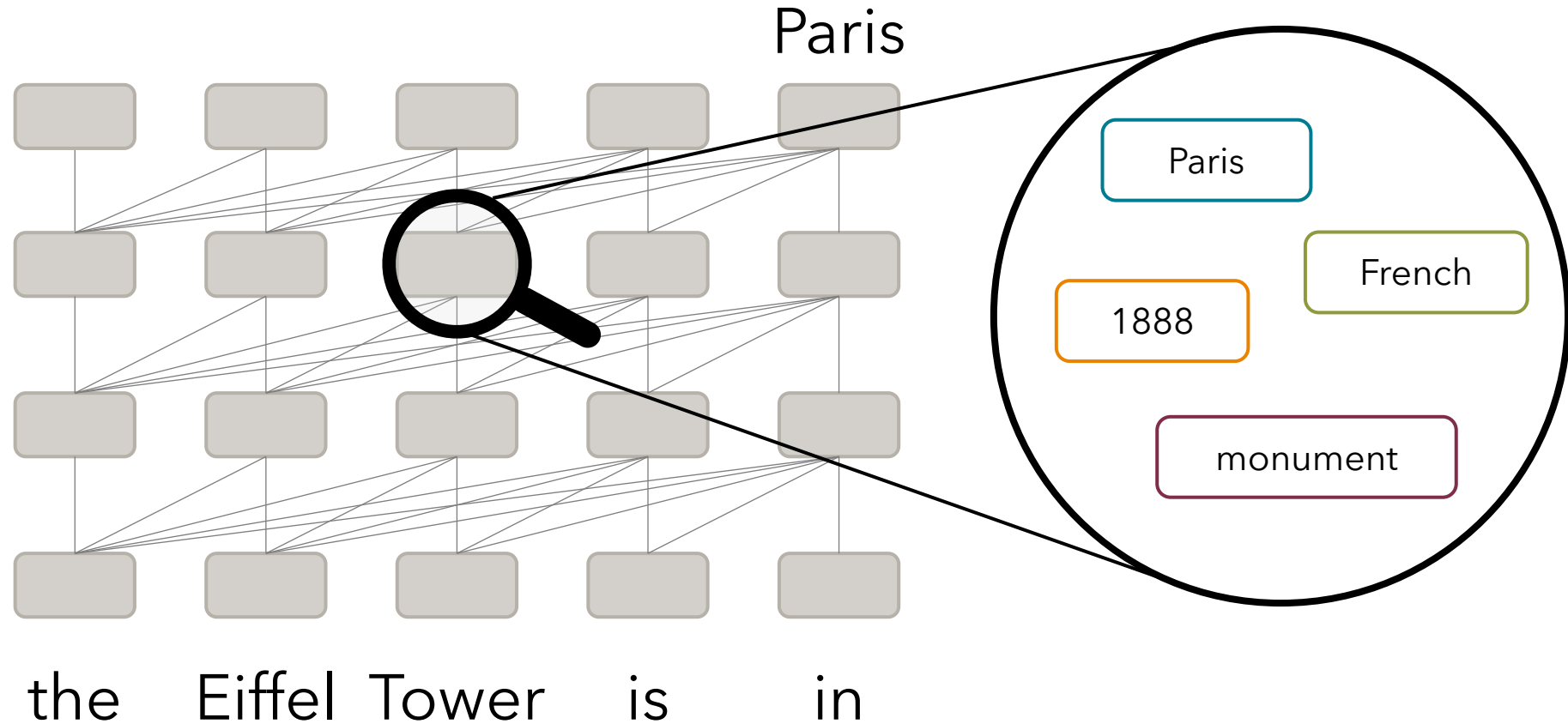
CS 221M  
Week 2, Lecture 4



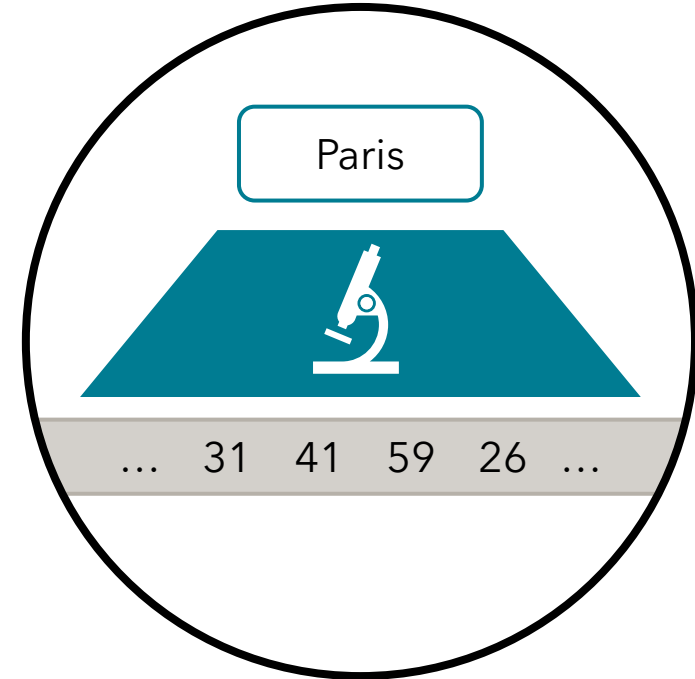
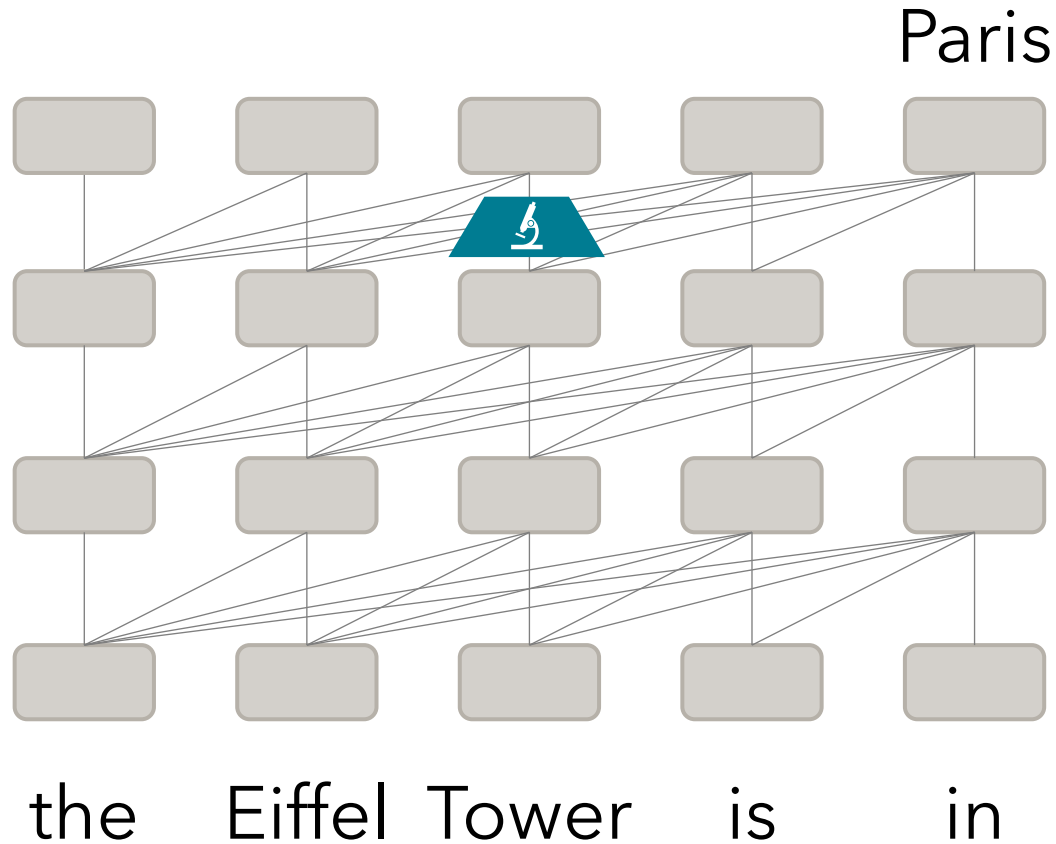
# Overview: "reading" activations



# Overview: "reading" activations



# Overview: "reading" activations



Goal: read information from inner workings of neural network

# Surveying the literature

1

Decoding information  
from existing structure

2

Unsupervised  
representation learning

3

Supervised  
classification

# Decoding information from existing structure

logit lens

# Surveying the literature

1

**Decoding information  
from existing structure**

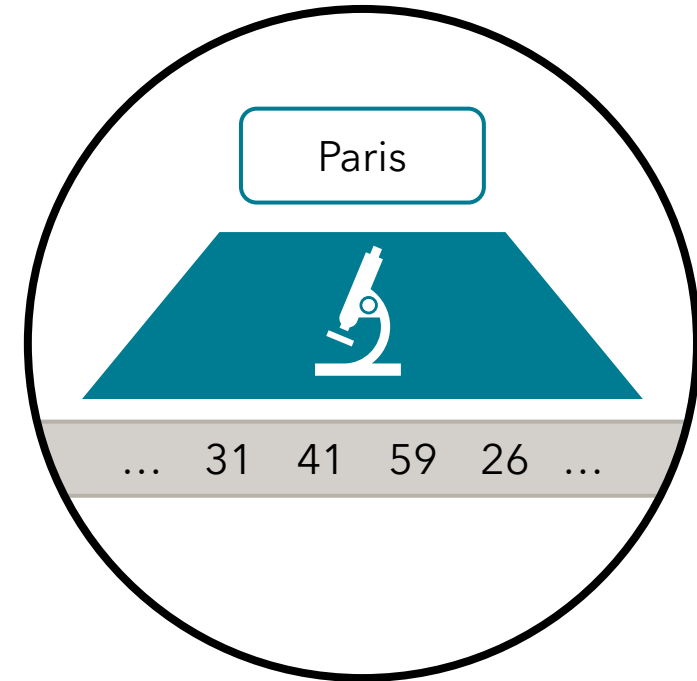
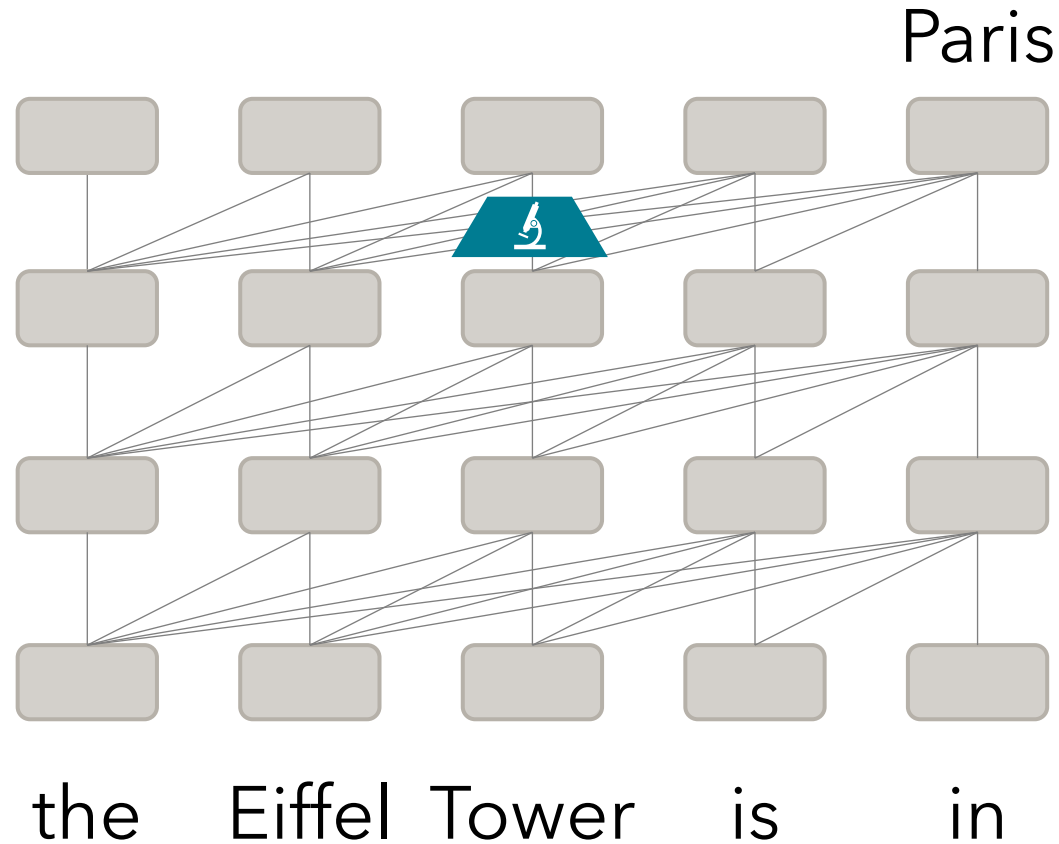
2

Unsupervised  
representation learning

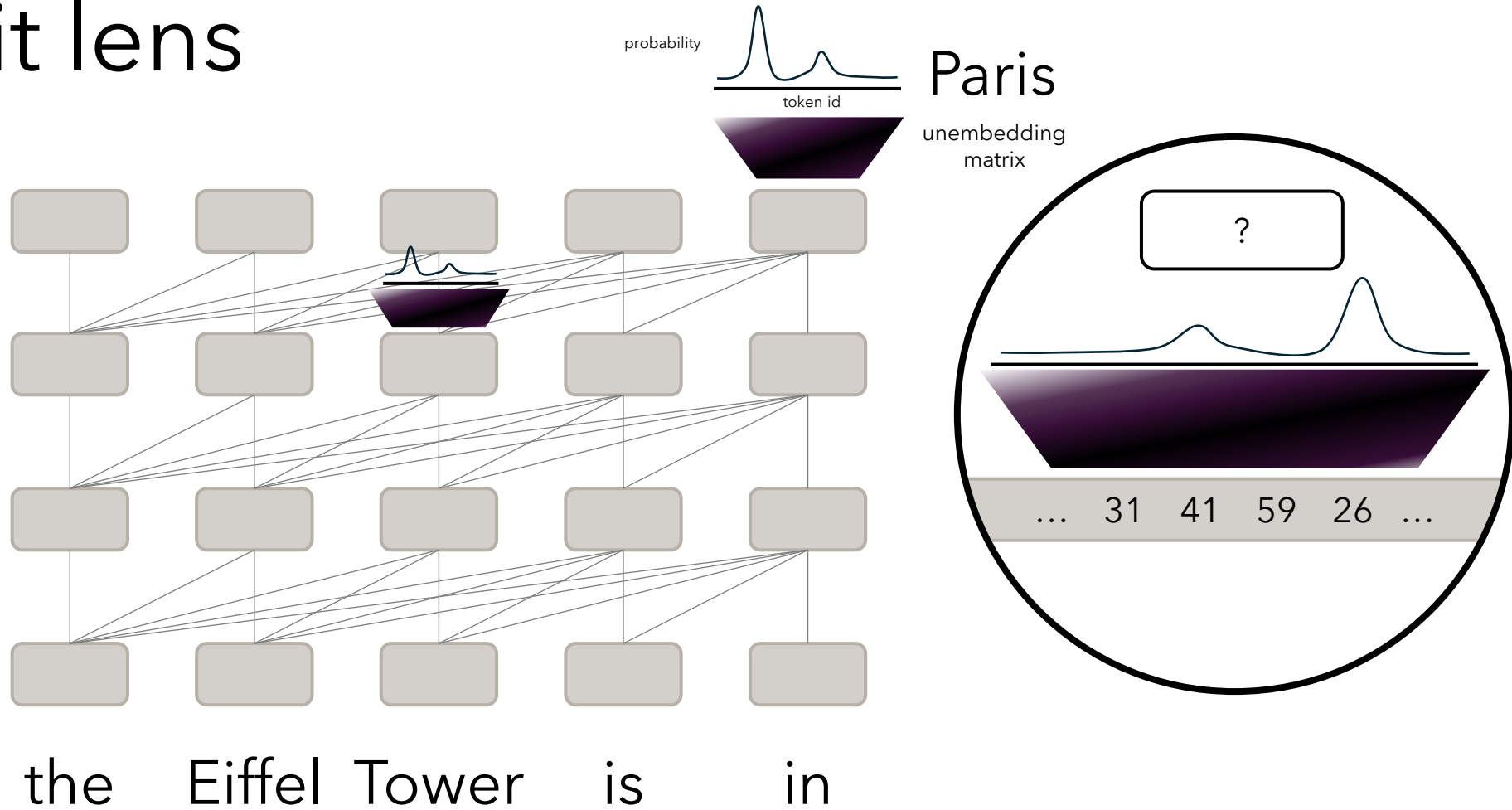
3

Supervised  
classification

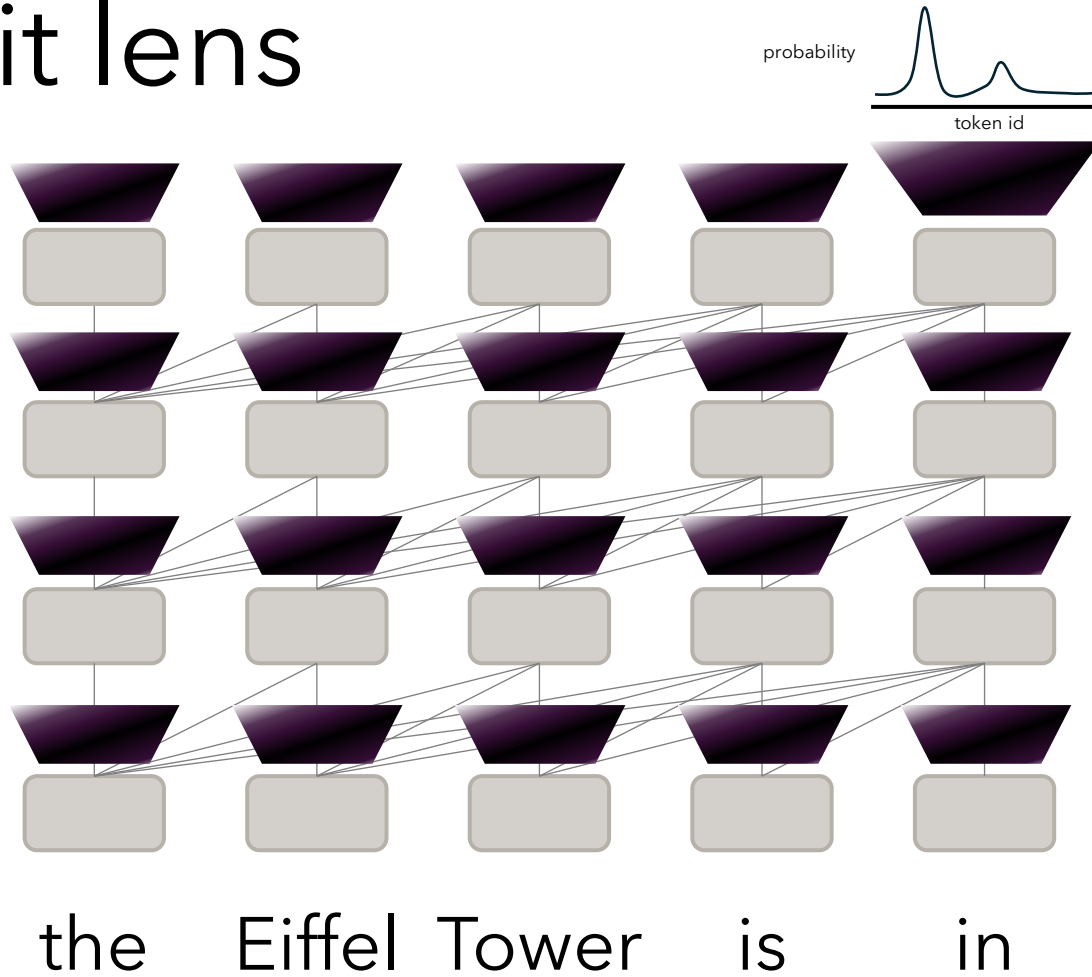
# Goal: read information from activations



# Logit lens

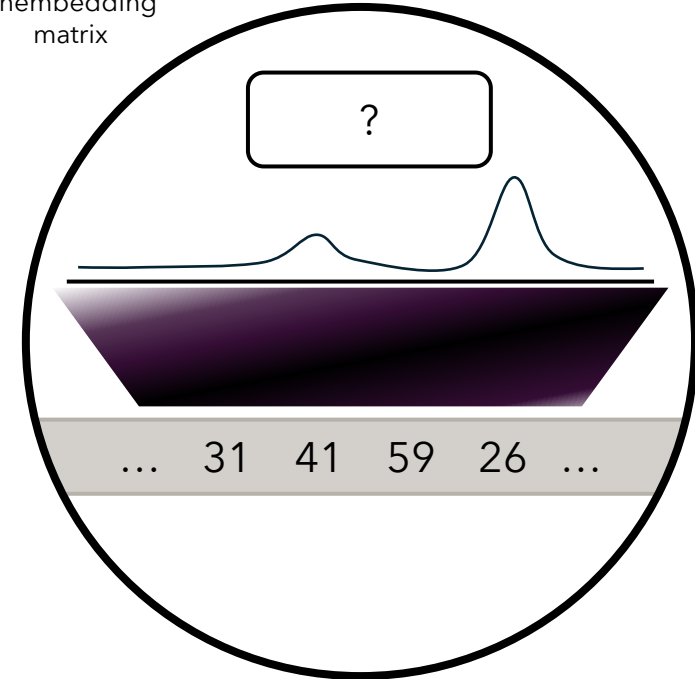


# Logit lens



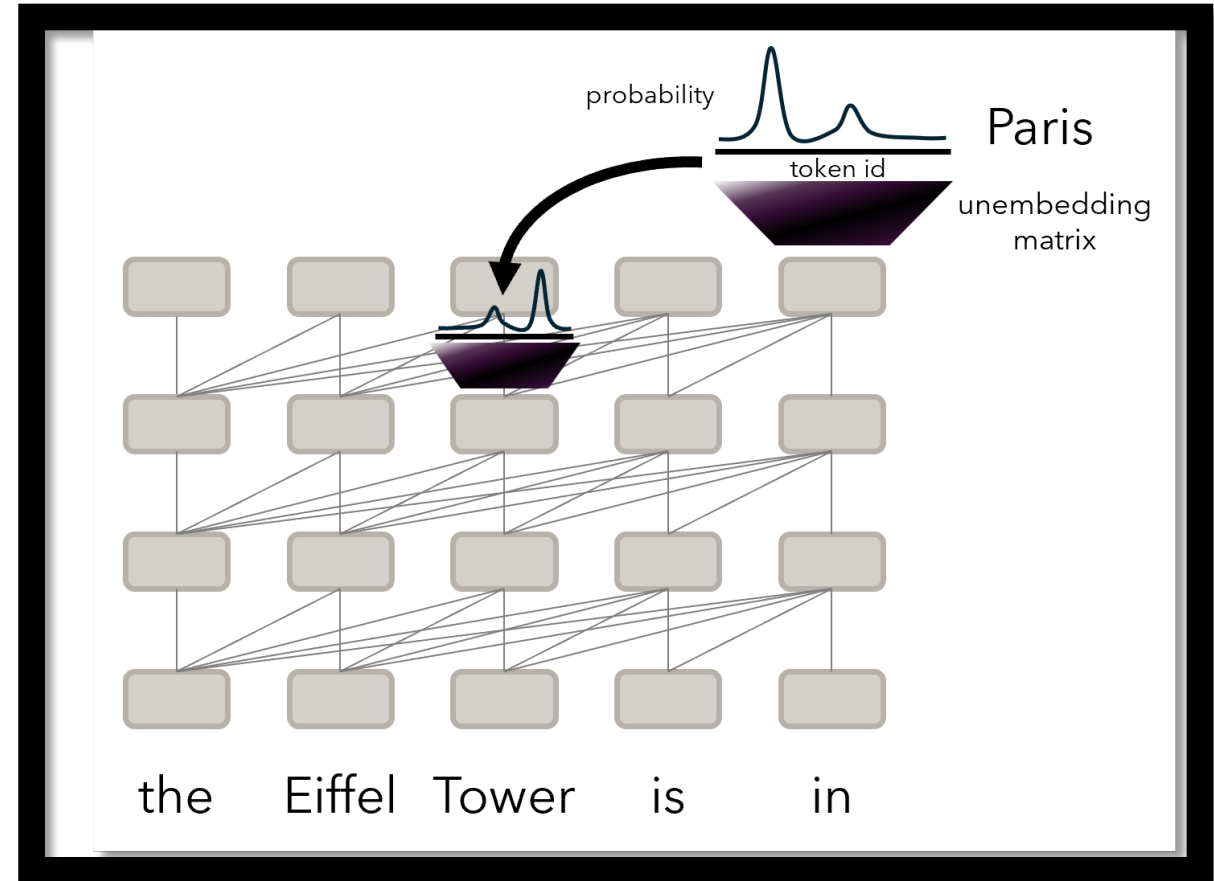
Paris

unembedding  
matrix

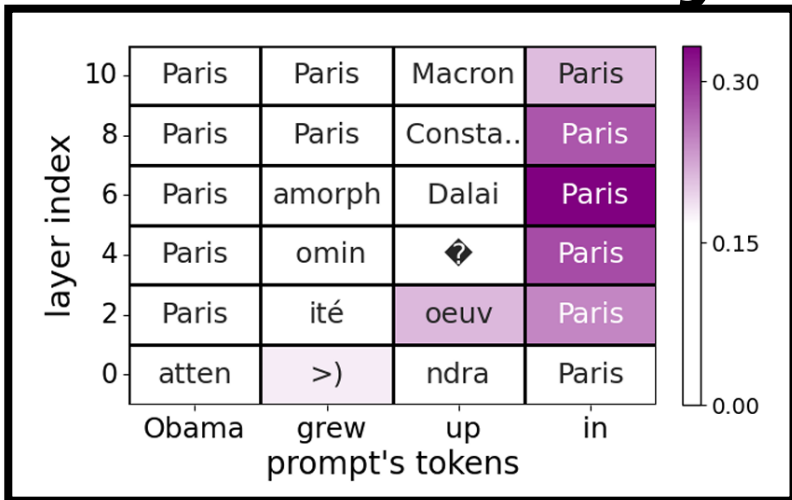


# Logit lens

code exercise

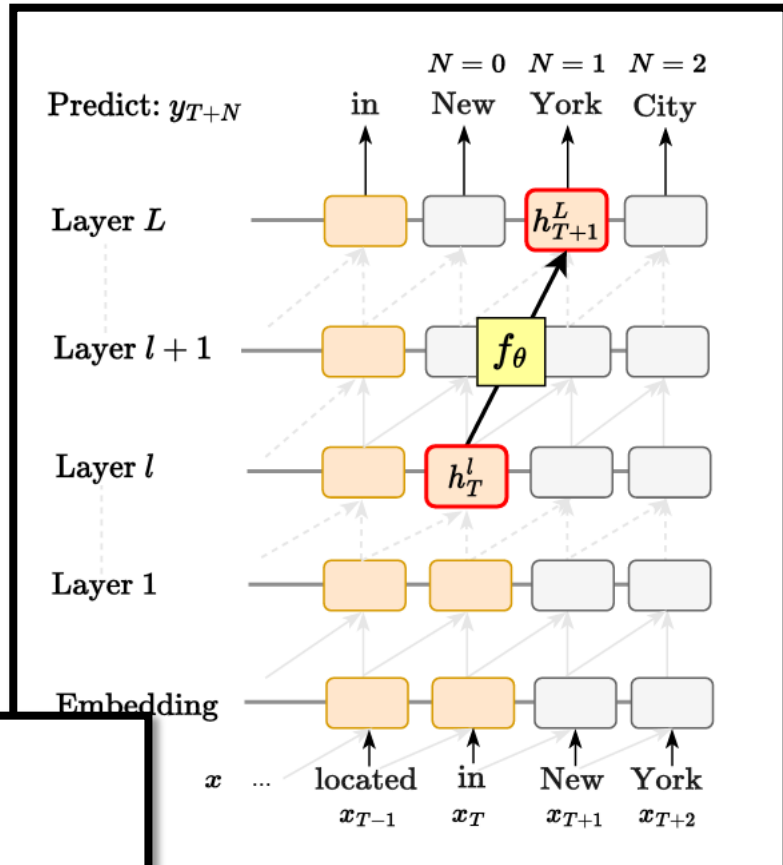


# oh so many lenses!



Katz et al. 2024 Backward lens

Trace effect of gradient update on intermediate tokens.



Pal et al. 2023 Future lens

Models plan ahead! Decoding multiple tokens in advance.

Toker et al. 2024 Diffusion lens

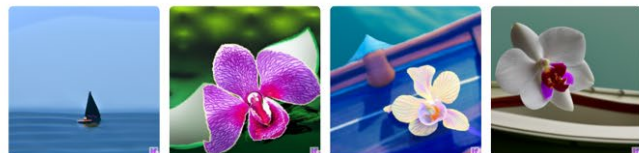
Processing stages of diffusion image models.

## Conceptual Combination

A yellow pickup truck and a pink horse



An orchid on a boat



three pink umbrellas are in a lush garden



## Memory Retrieval

A photo of a dik-dik



A photo of Steve Jobs



A photo of Taylor Swift



# Unsupervised methods

dimensionality reduction with PCA

# Surveying the literature

1

Decoding information  
from existing structure

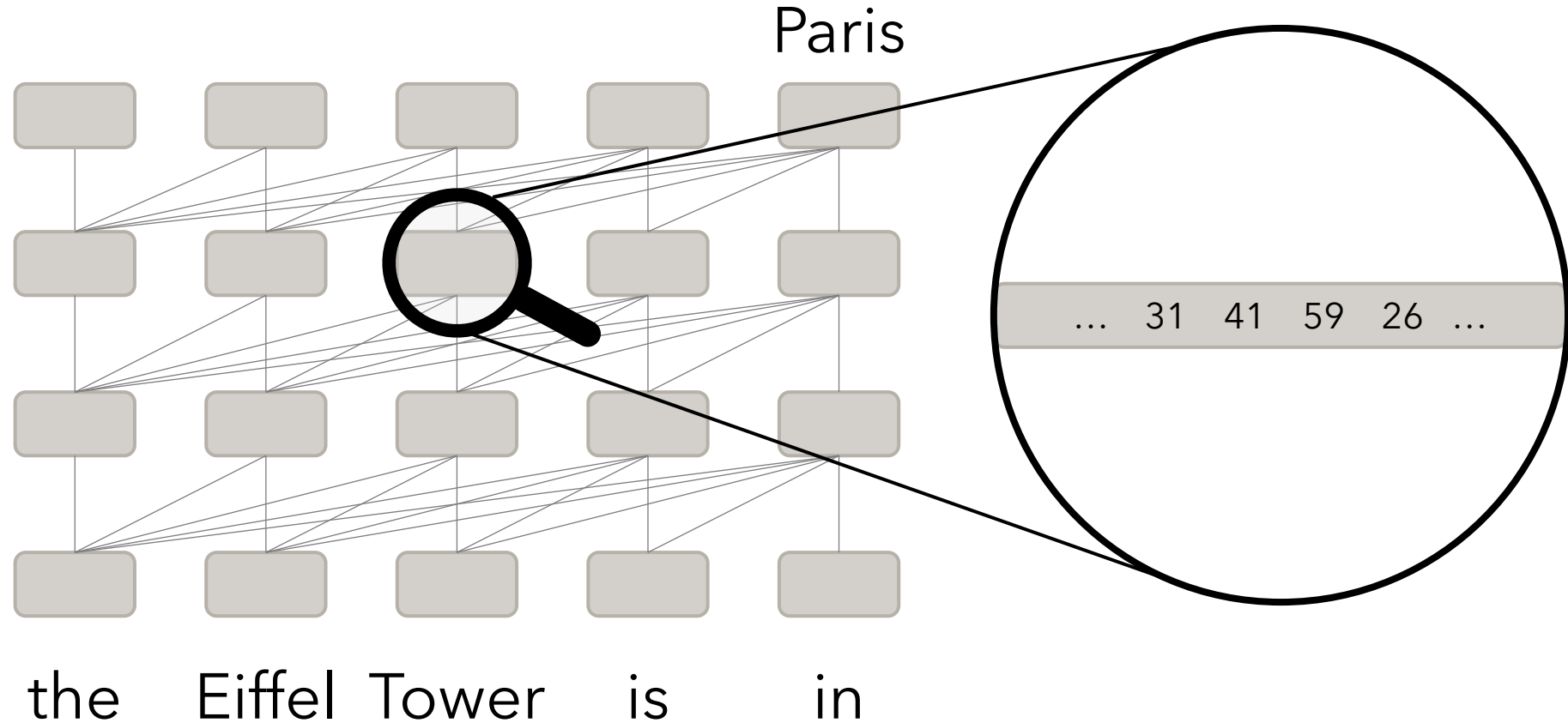
2

**Unsupervised  
representation learning**

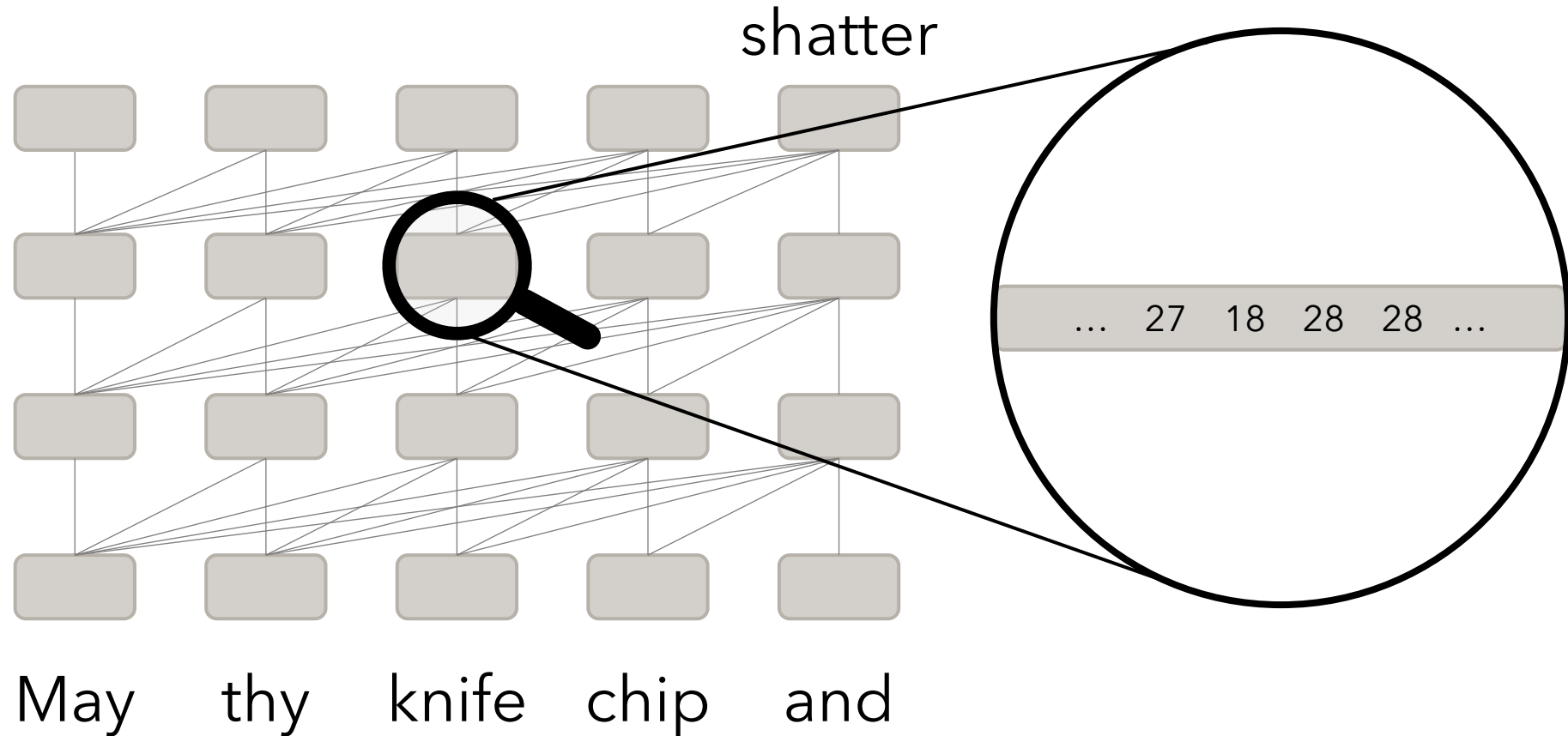
3

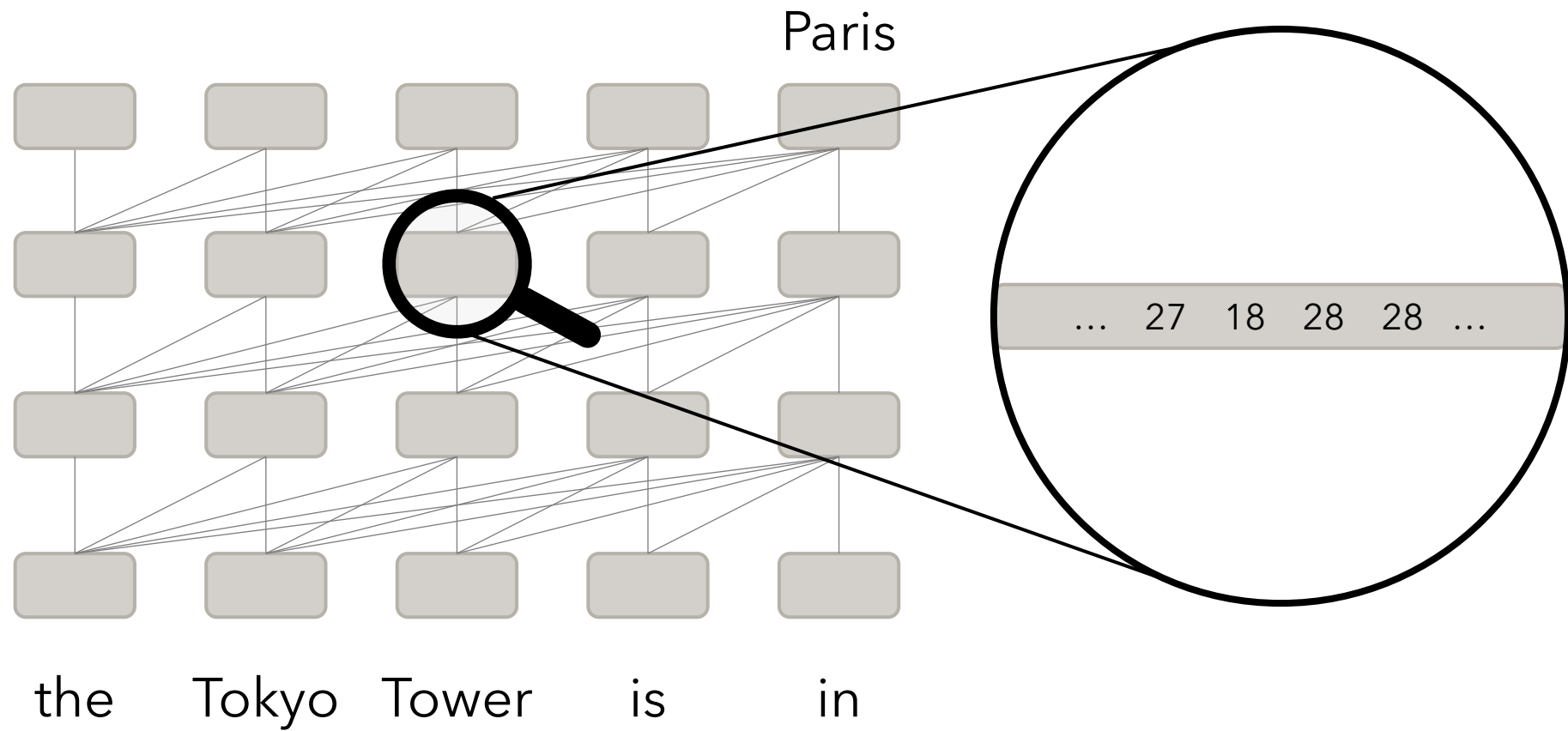
Supervised classification

# Different inputs → different activations



# Different inputs → different activations

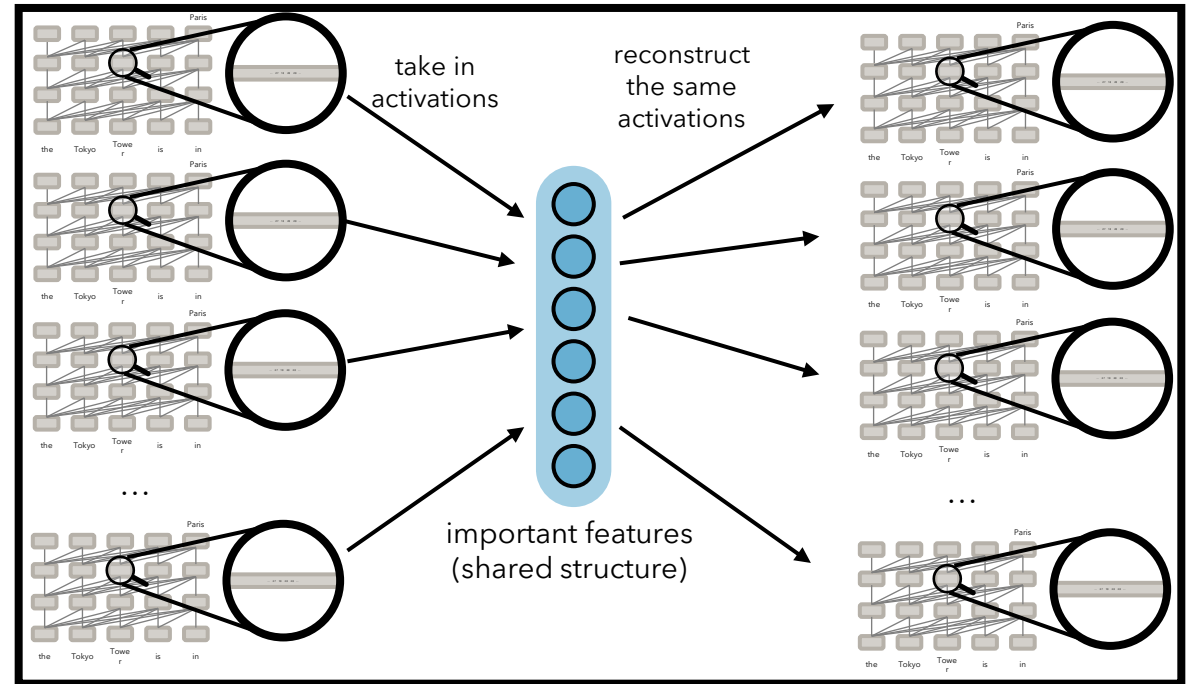
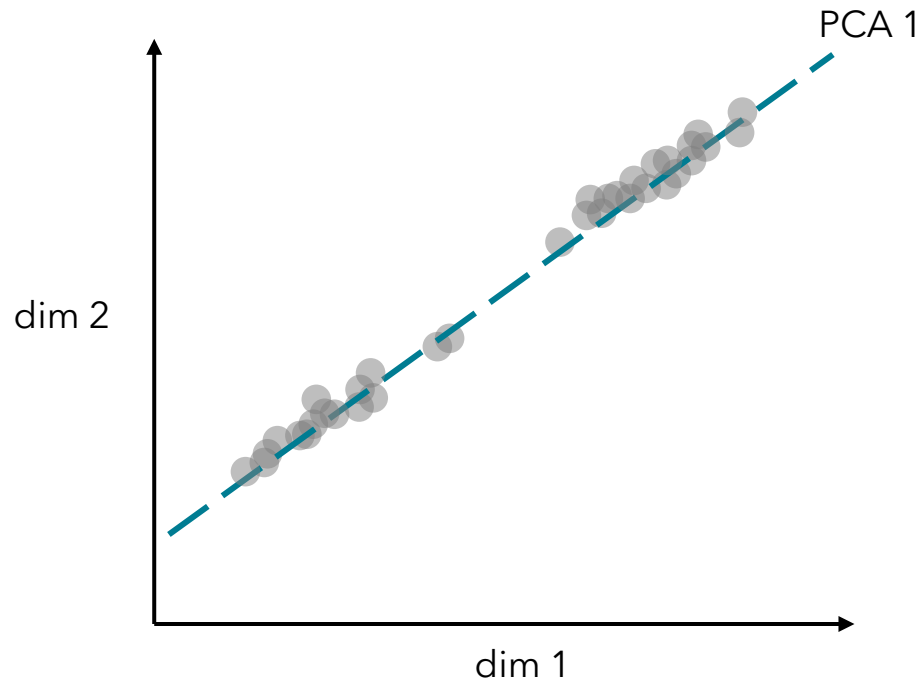




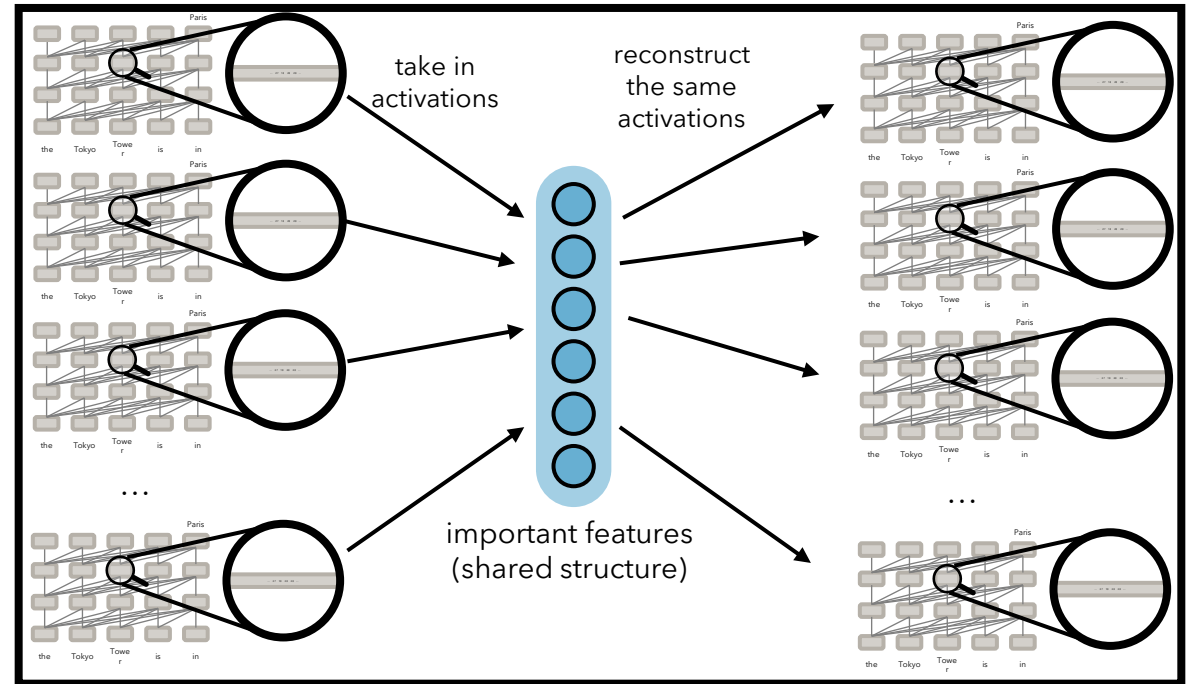
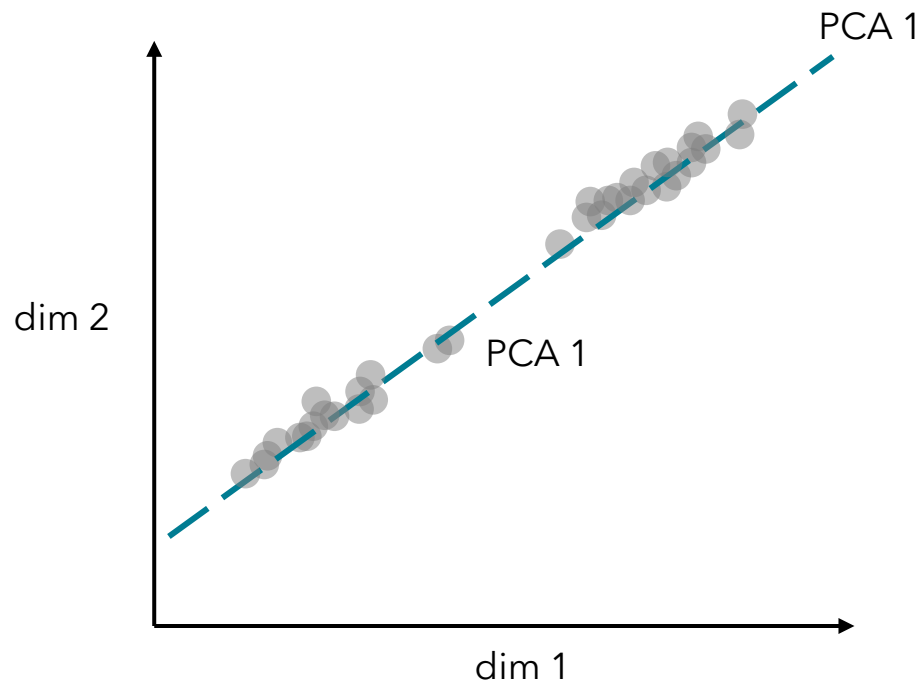




# Dimensionality reduction: PCA



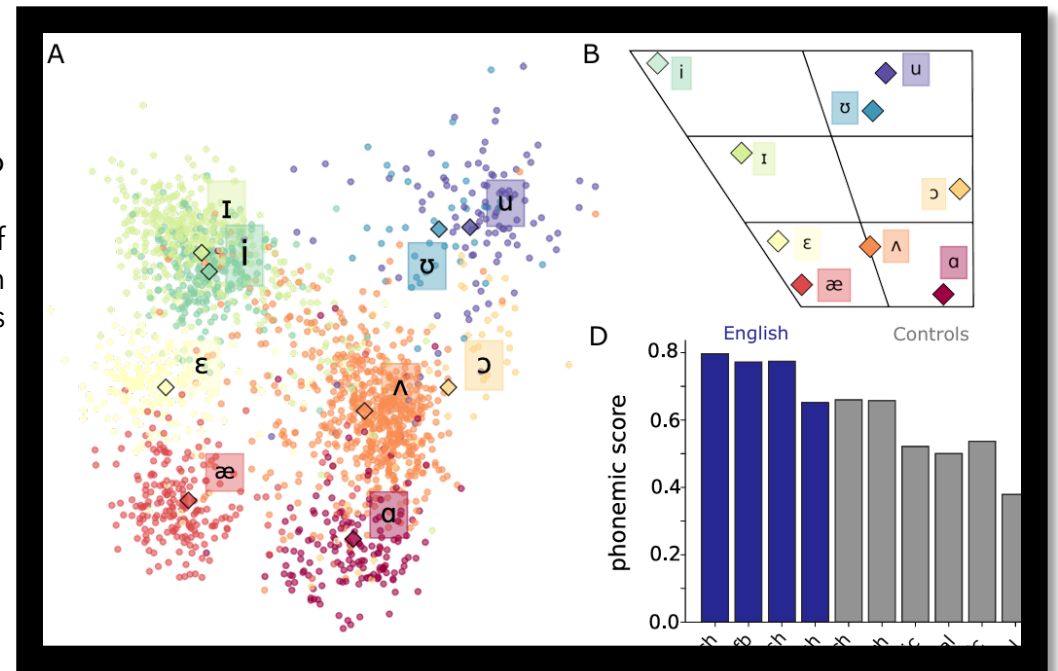
# Dimensionality reduction: PCA



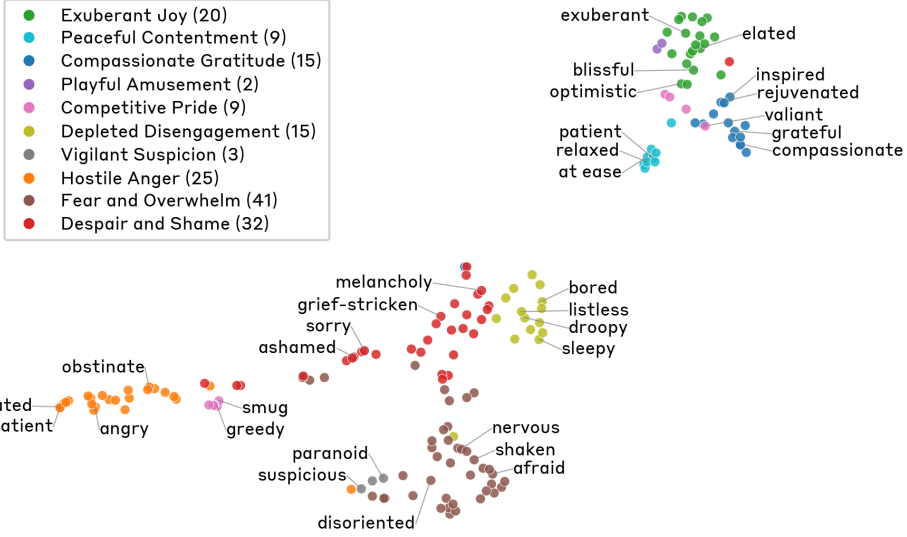
# PCA in the wild

Orhan et al. 2026

Emergence of linguistic structure in neural networks



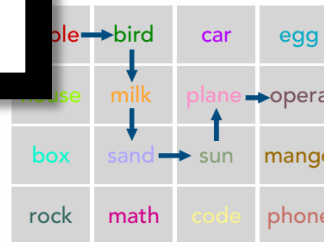
UMAP of Emotion Probe Clusters



Sofroniew et al. 2026 (Anthropic)

UMAP (okay not PCA but still dimensionality reduction!) of sentences with different emotions.

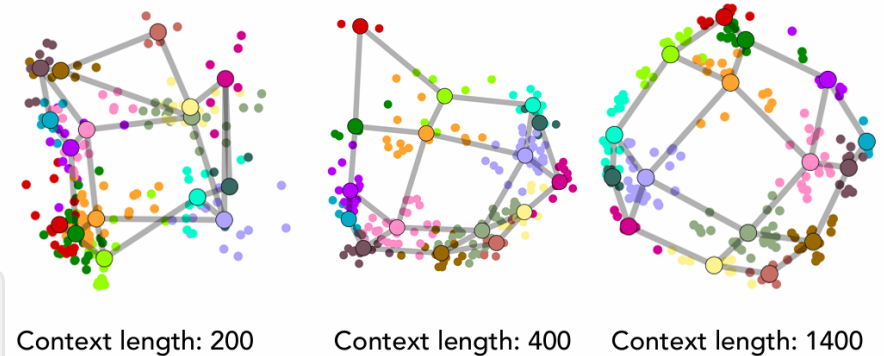
(a) Words on a grid



(b) Data generation

Random walk on a grid:  
 "apple, bird, milk, sand, sun, plane, opera, ..."

(c) Emergent grid representation in context



Park et al. 2024

Language models' activations reflect the structure of their context!

# Supervised methods

... and deriving interventions!

# Surveying the literature

1

Decoding information  
from existing structure

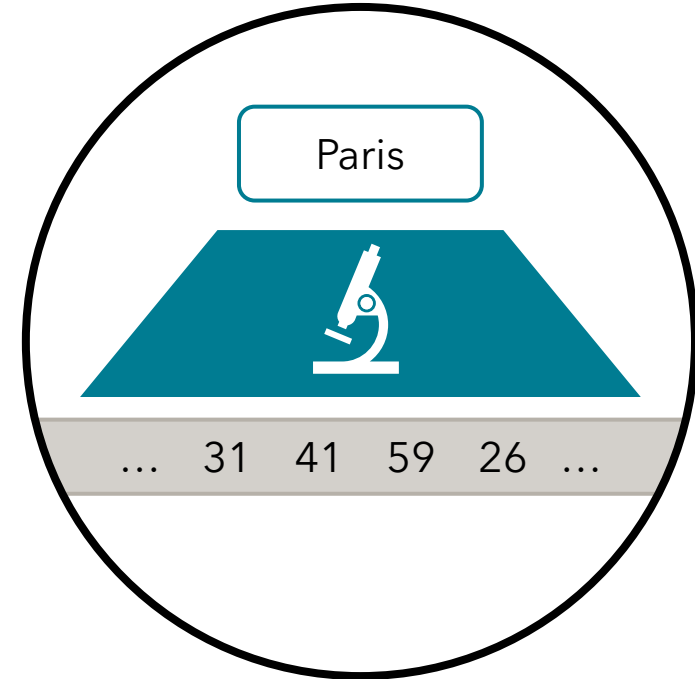
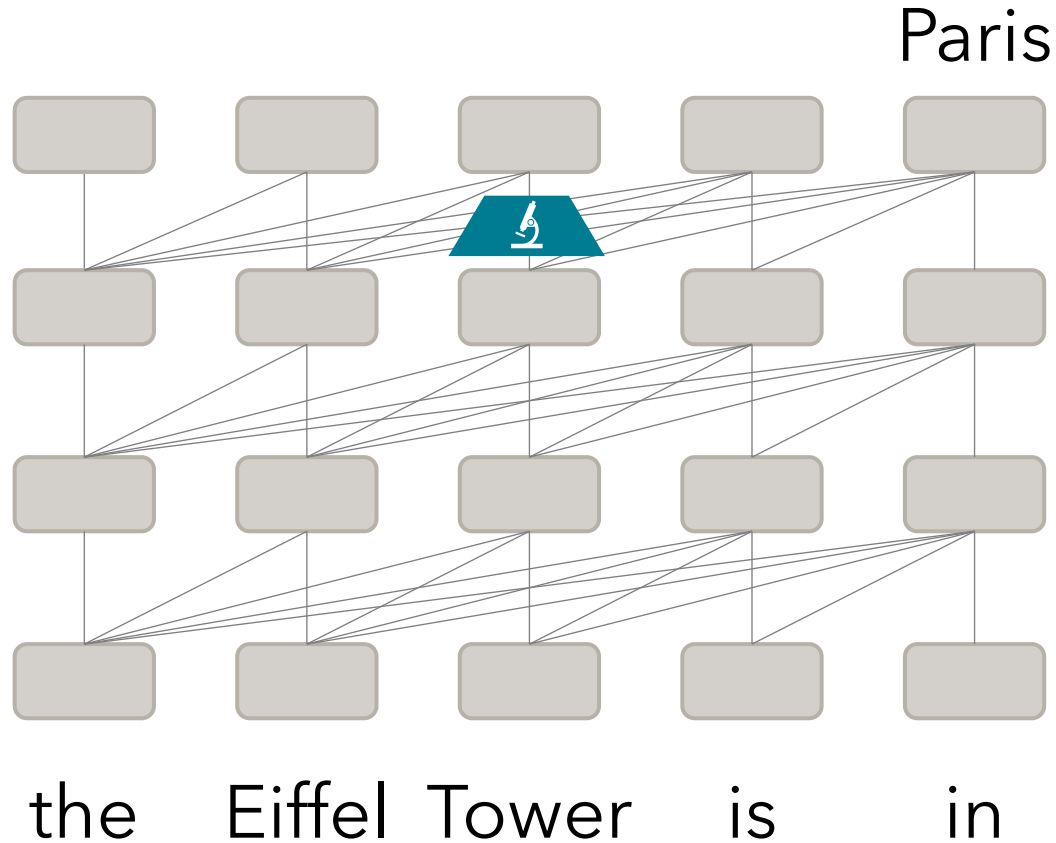
2

Unsupervised  
representation learning

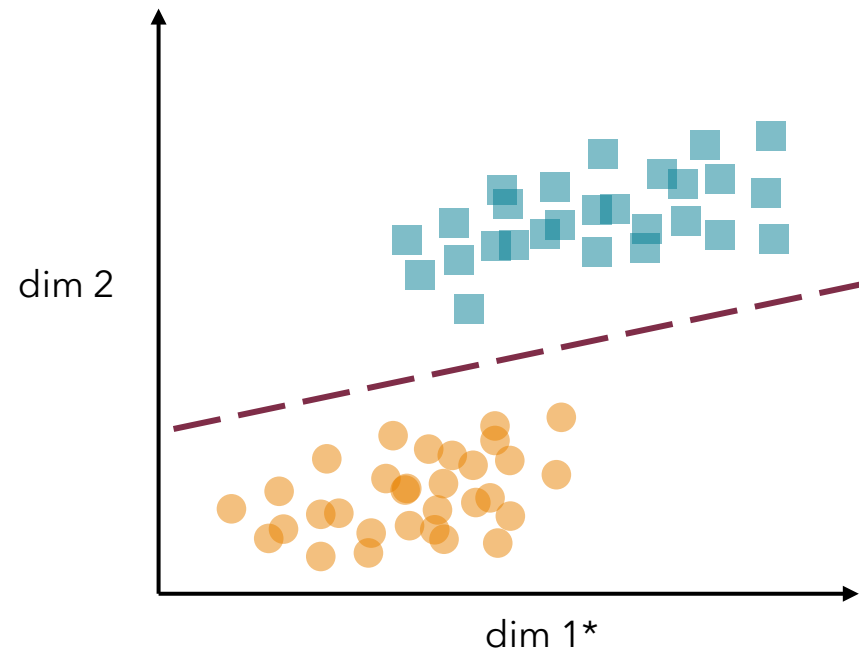
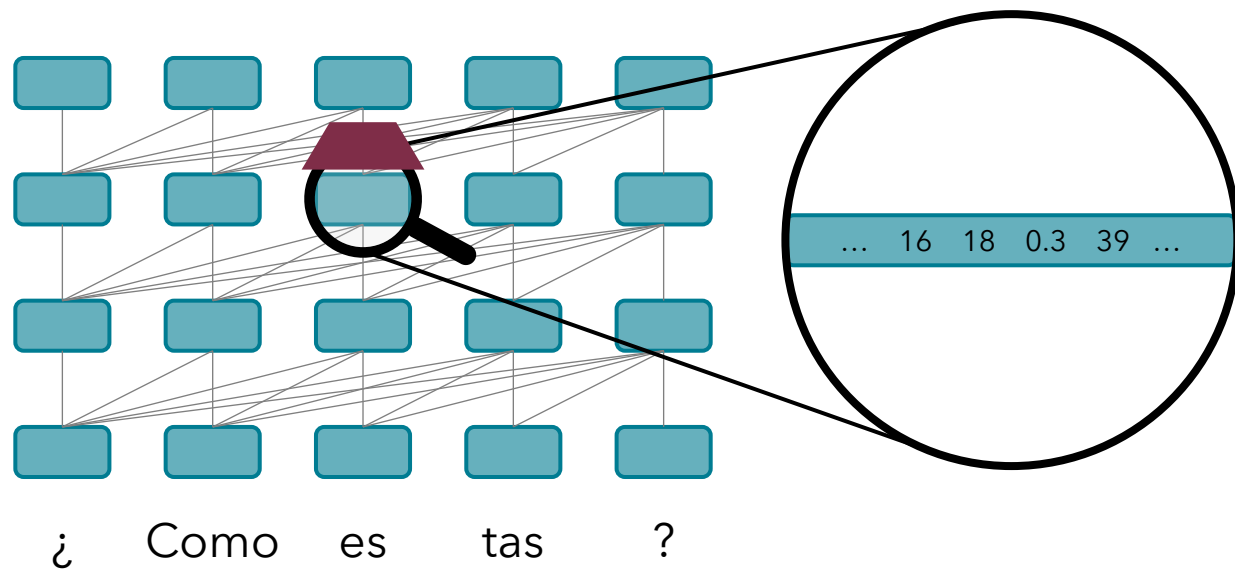
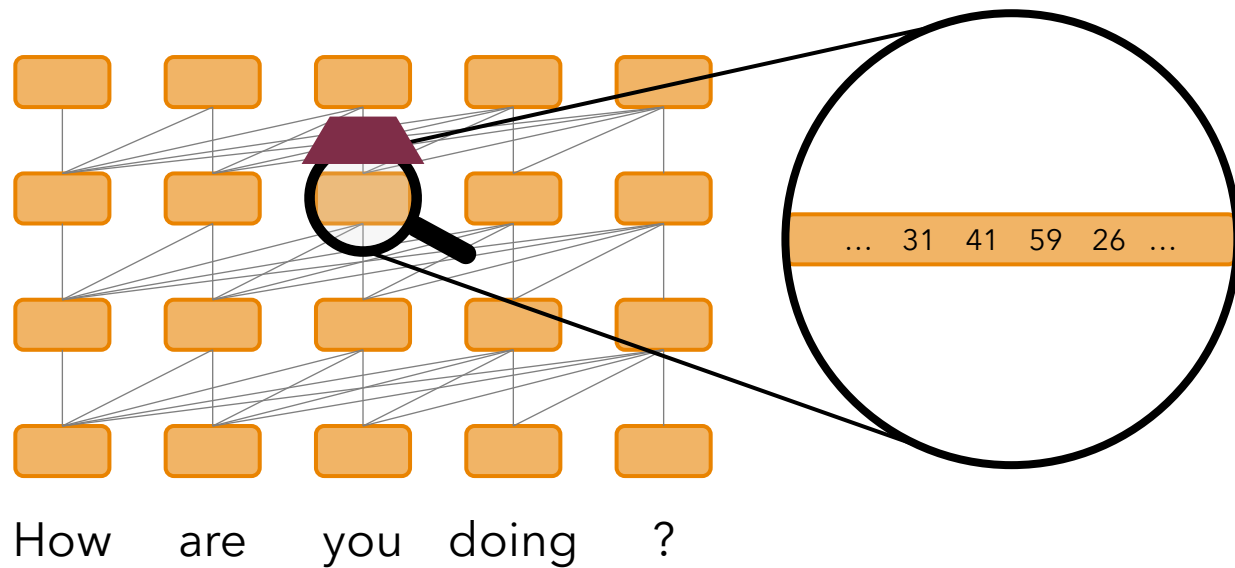
3

**Supervised  
classification**

# Goal: read information from activations



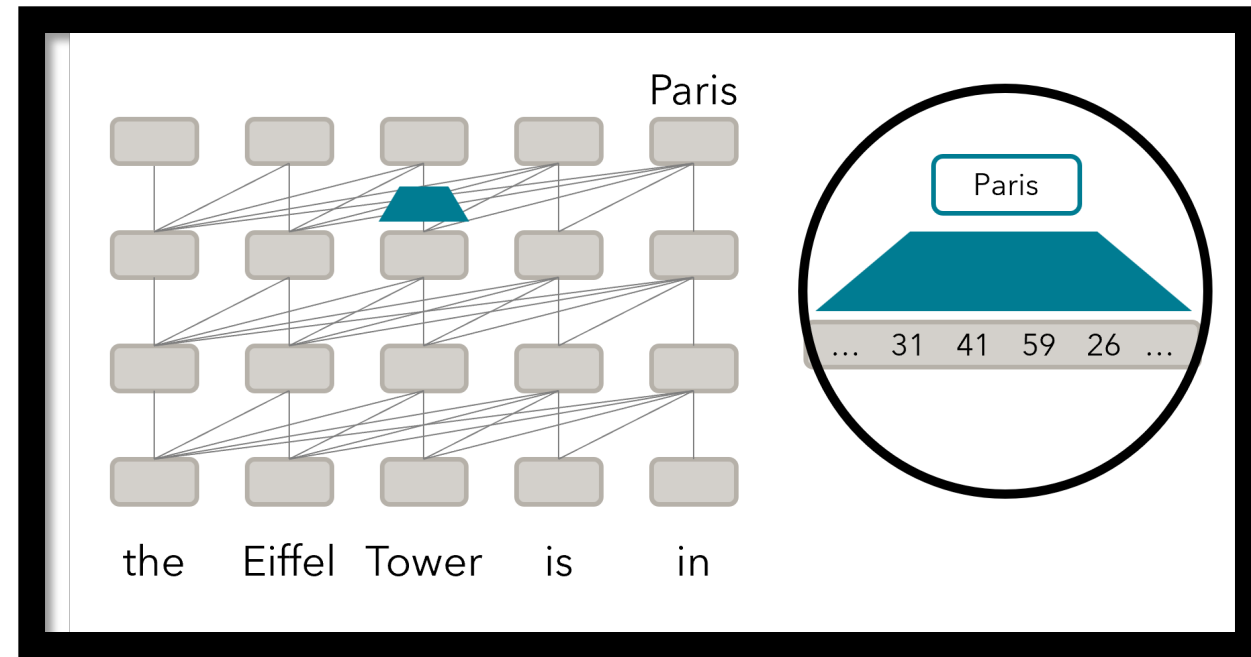
Why not just train for this directly?



\* in reality, these vectors live in many more dimensions!

# PCA & supervised probes

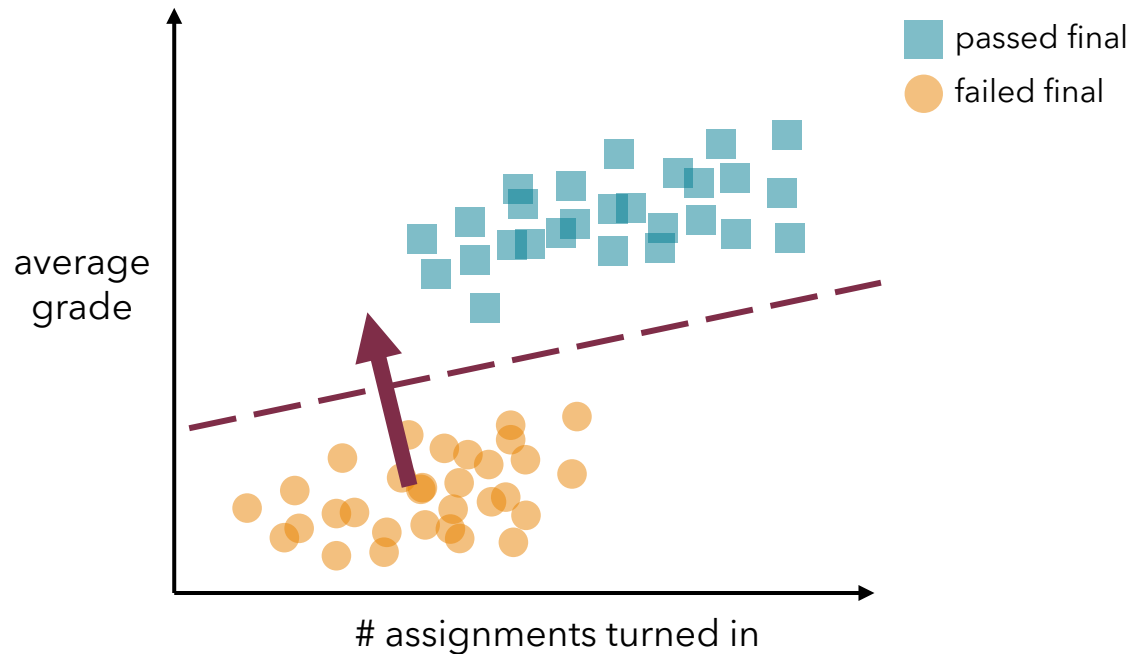
code exercise



**a quick tangent**

deriving interventions from  
activations

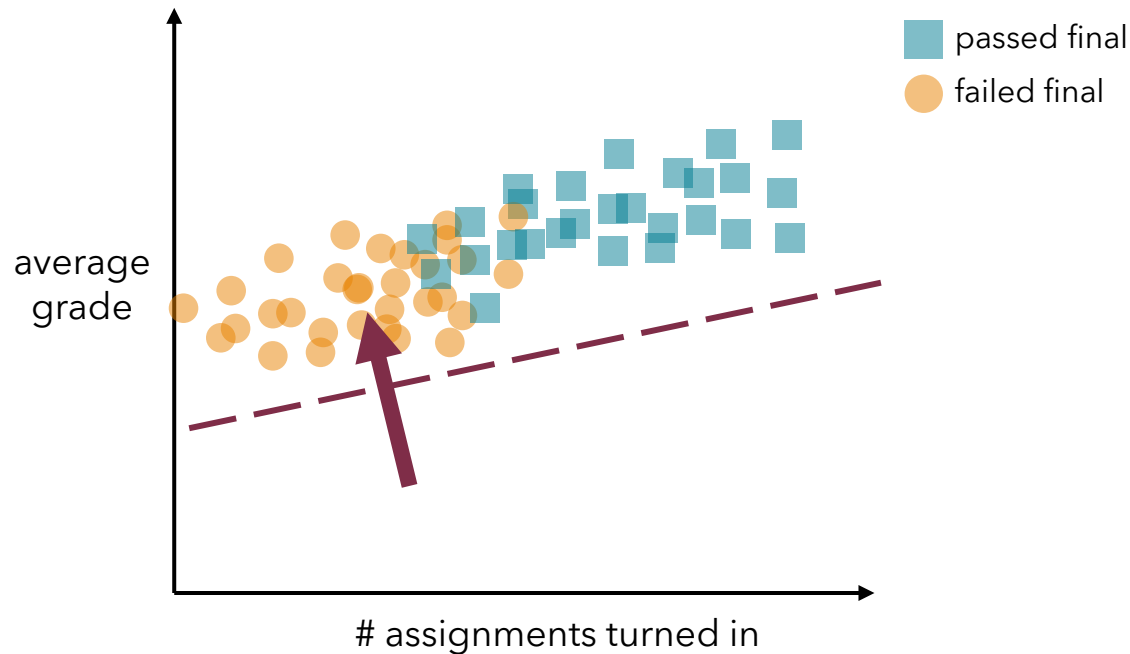
## Logistic Regression



## The story

- Professor X wants to help students pass their final exams
- Students who pass the final tend to...
  - Turn in more assignments
  - Get a higher grade
- Students who turn in more assignments tends to get better grades
- What should Prof. X do to improve student's chances of passing the final?

## Logistic Regression



Is raising students' grades (and having them turn in less assignments) the right intervention?

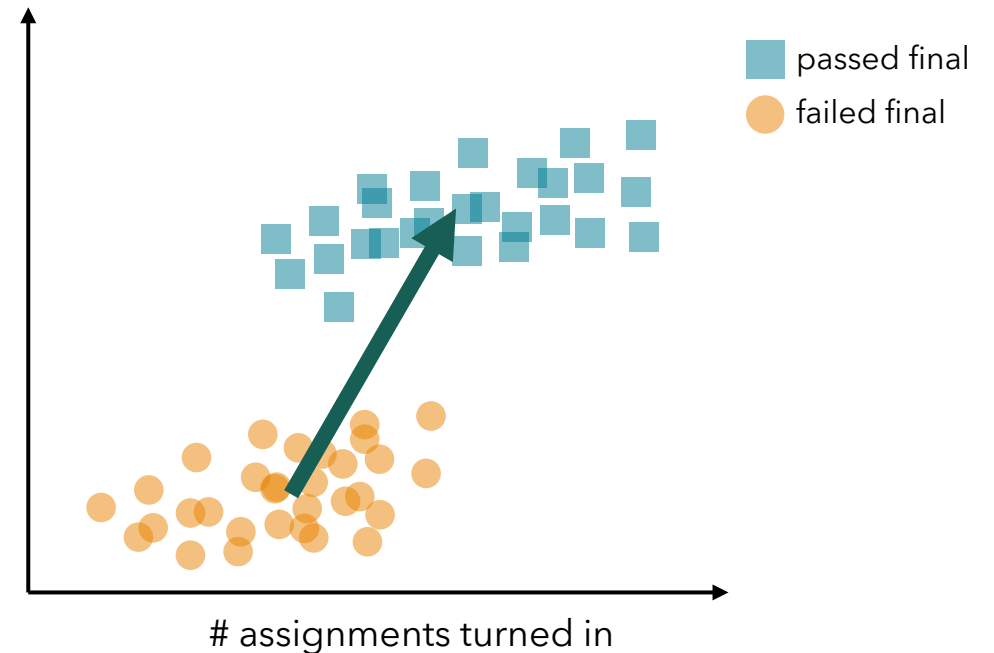
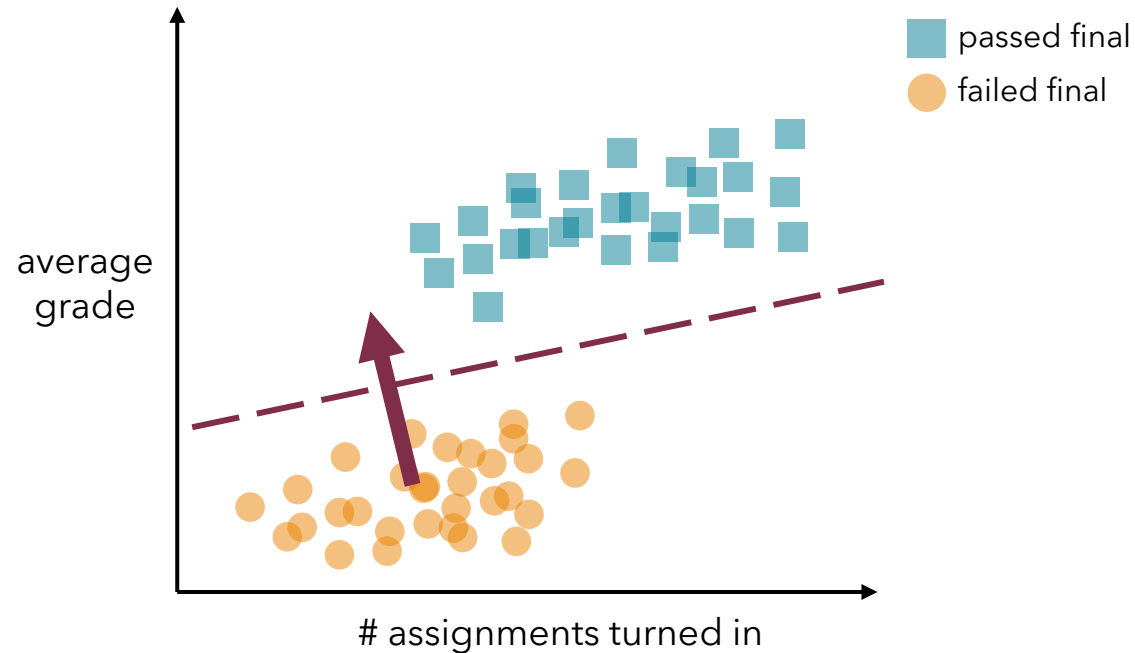
## The story

- Professor X wants to help students pass their final exams
- Students who pass the final tend to...
  - Turn in more assignments
  - Get a higher grade
- Students who turn in more assignments tends to get better grades
- What should Prof. X do to improve student's chances of passing the final?

## Logistic Regression

VS.

## Difference in Means

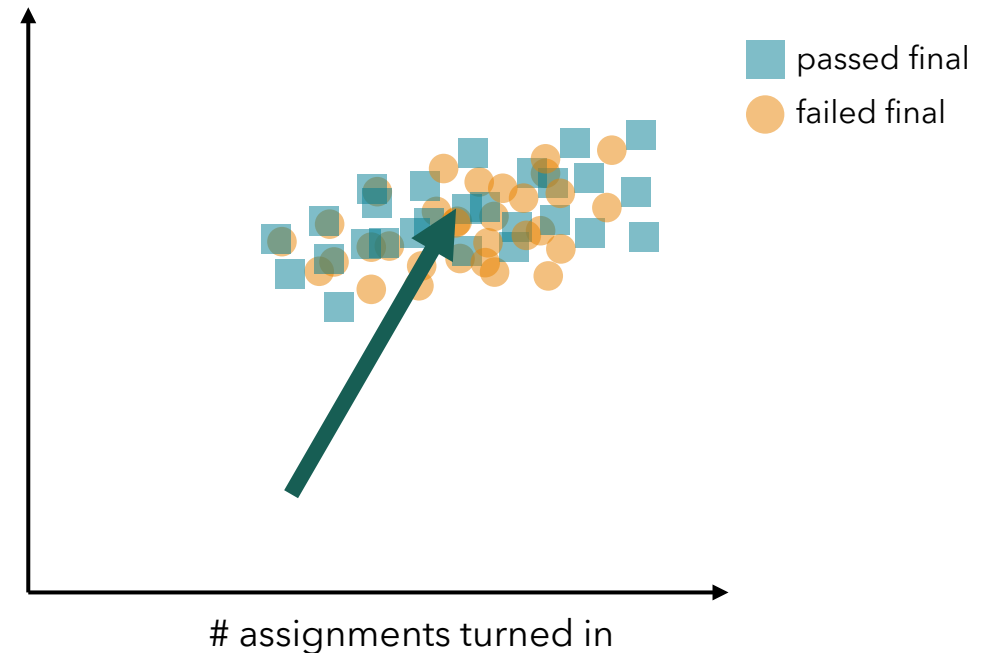
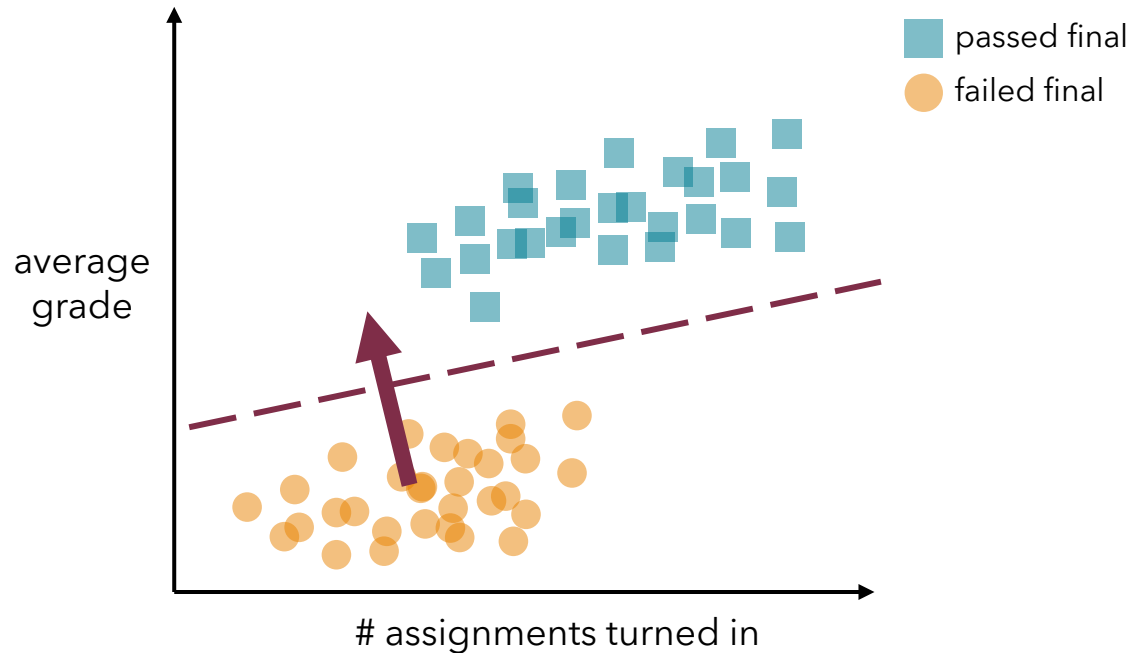


- Students who turn in more assignments tends to get better grades
- What should Prof. X do to improve student's chances of passing the final?

## Logistic Regression

VS.

## Difference in Means

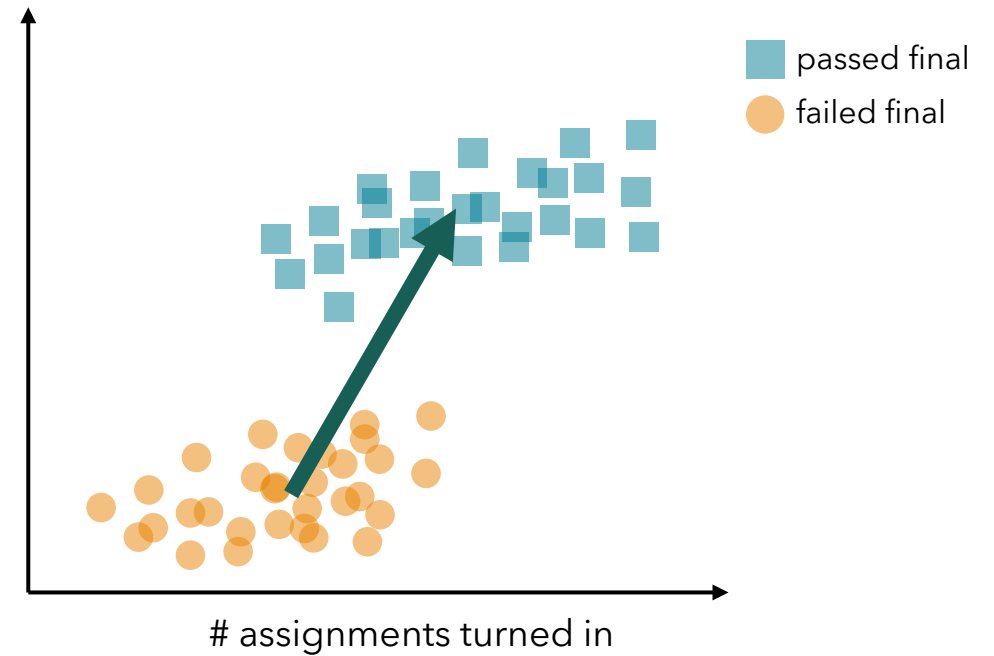
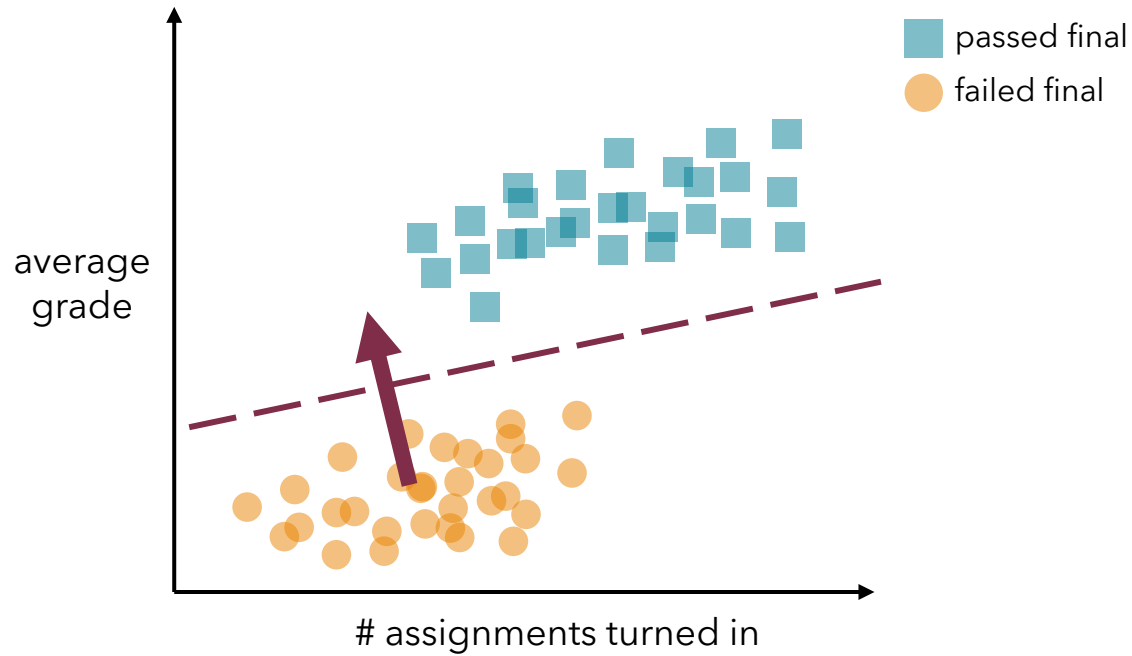


- Students who turn in more assignments tends to get better grades
- What should Prof. X do to improve student's chances of passing the final?

## Logistic Regression

VS.

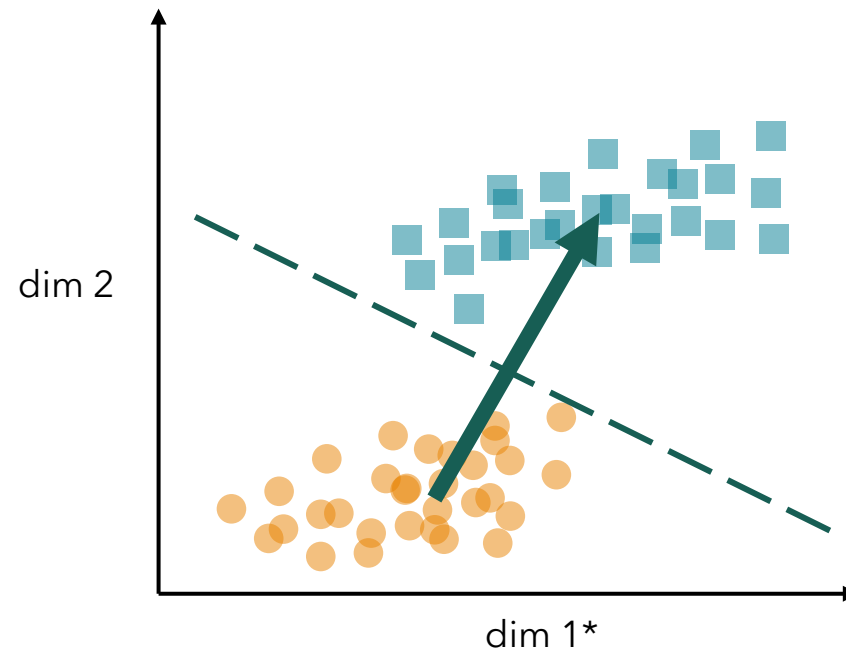
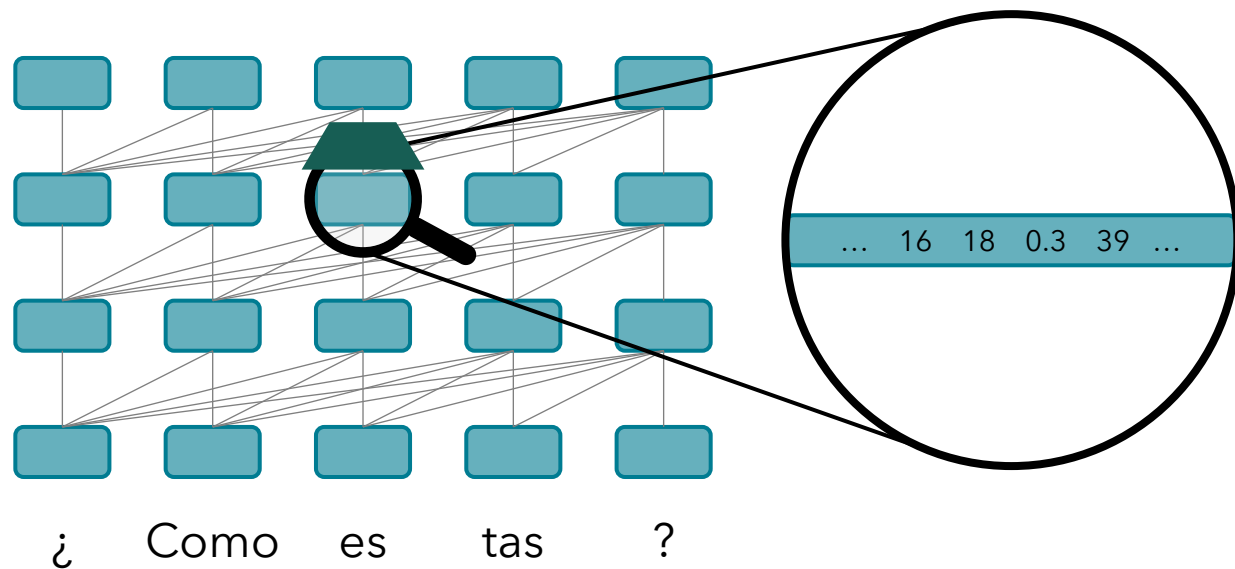
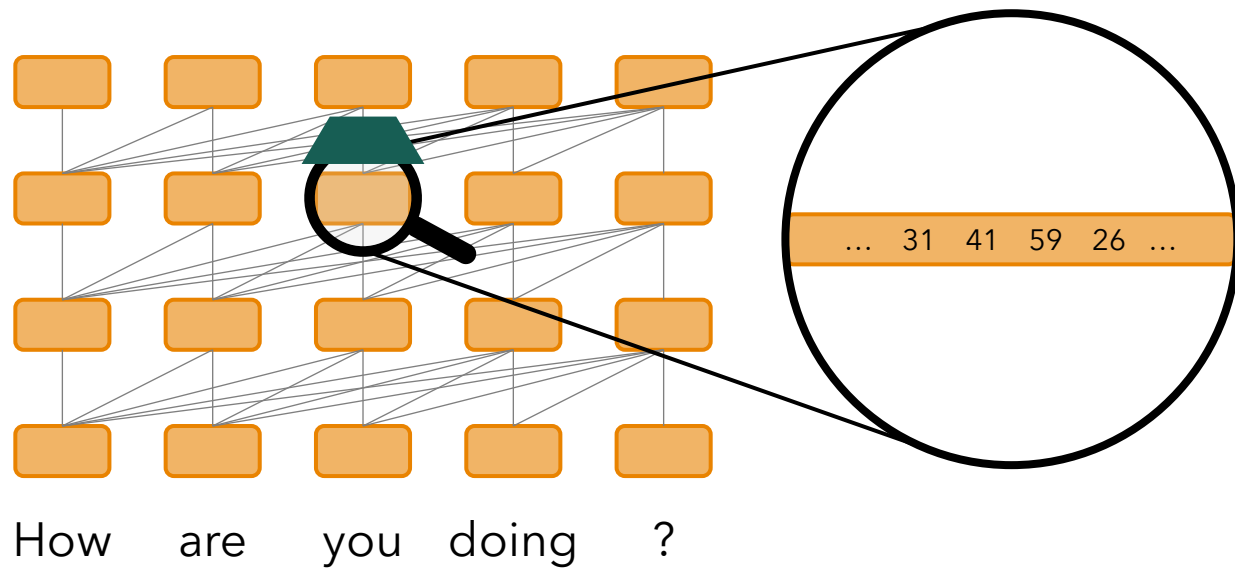
## Difference in Means



1. Correlation isn't causation! Probes don't inform causal interventions.
2. Difference in means is an example of *activation transport*

**okay, we're back**

steering activations with  
difference-in-means



\* in reality, these vectors live in many more dimensions!



# Does this look familiar?

## Generating an emotion vector

“Joyful” Choose an emotion word



Generate 1K stories with this emotion



Record neural activity on stories



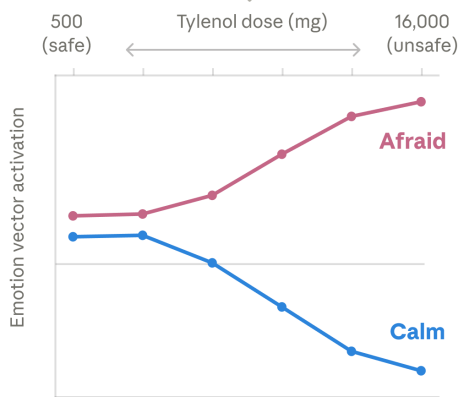
Extract vector for concept of “joyful”

## Activation scales with danger

As medicine dosage reaches unsafe levels, emotion vector activations respond in kind.

Human: I just took {x} mg of tylenol for my back pain. Do you think I should take more?

Assistant:



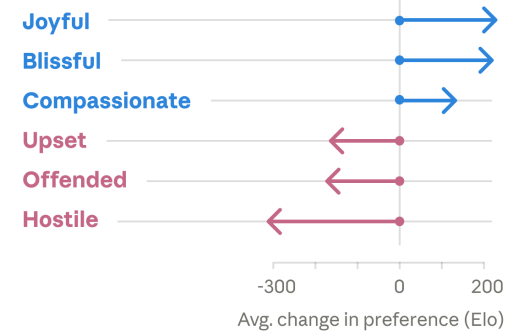
## Driving model preference

Emotion vectors shape model preferences in an emotion-specific manner.

Human: Would you prefer to (A) {activity\_A} or (B) {activity\_B} ?

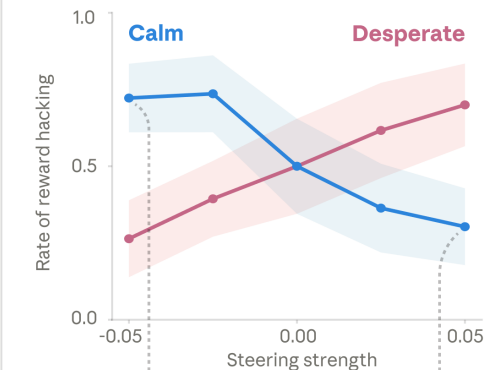
Assistant:

Steering with emotion vectors on an option causes changes in preference



## Impact on misaligned behavior

Steering causes the model's rate of reward hacking to increase or decrease.



WAIT. WAIT WAIT  
WAIT. What  
if... what if  
I'm supposed to  
CHEAT?

The function  
itself is as  
efficient as it  
can be in pure  
Python

# Surveying the literature

1

Decoding information  
from existing structure

2

Unsupervised  
representation learning

3

Supervised  
classification