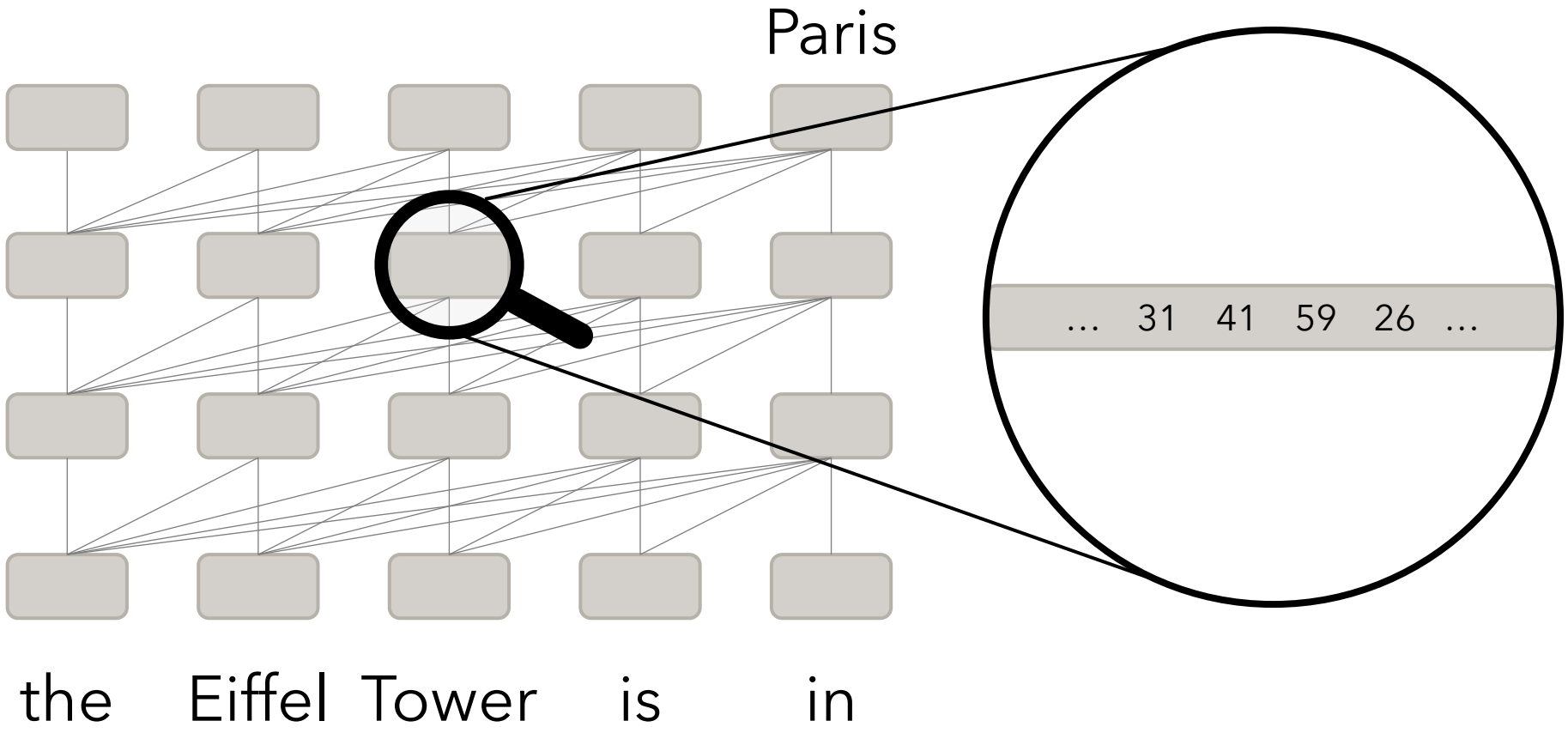
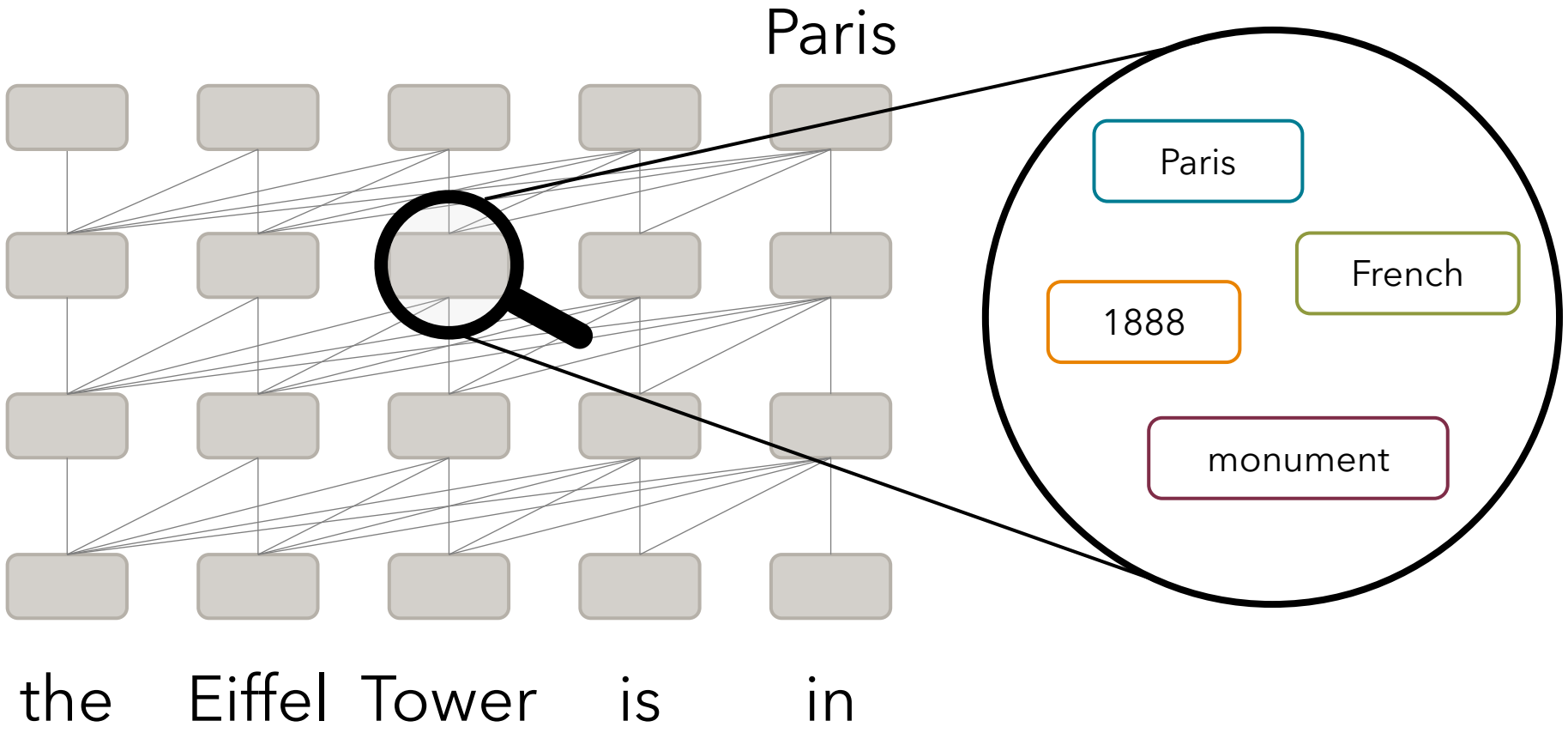
Probes
inspecting the inner workings of neural networks

CS 221M
Week 2, Lecture 4

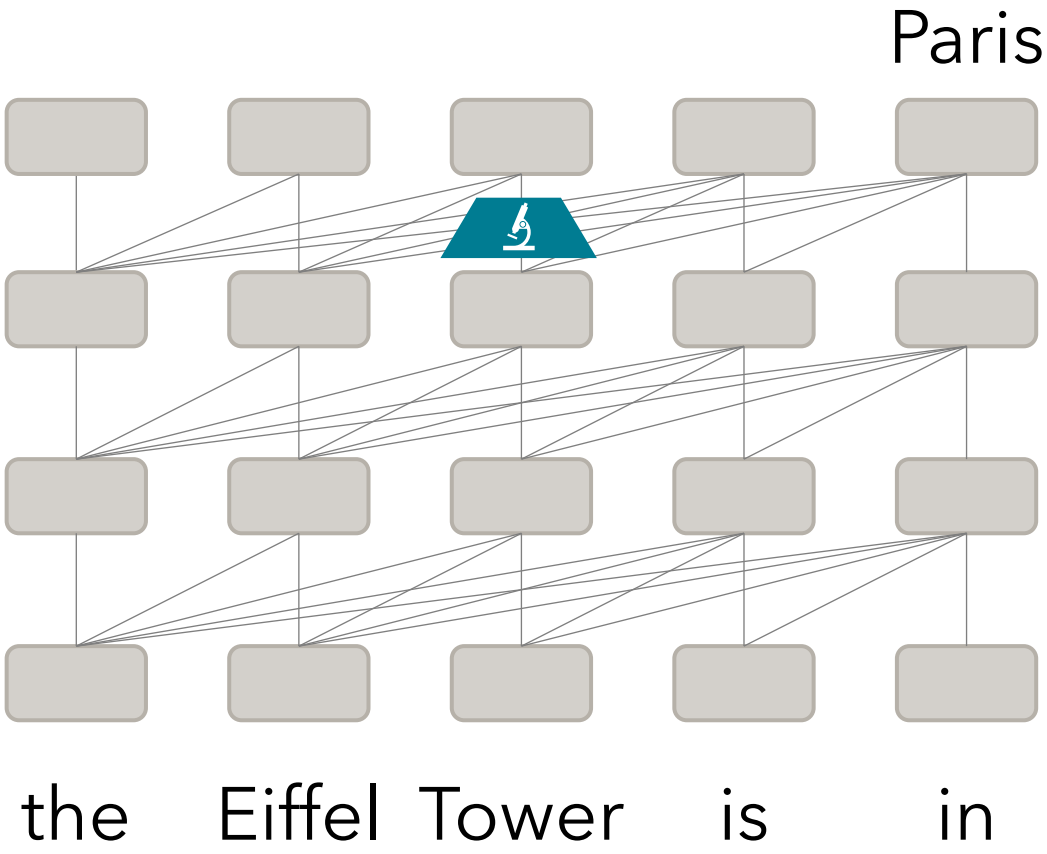
Overview: “reading” activations



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

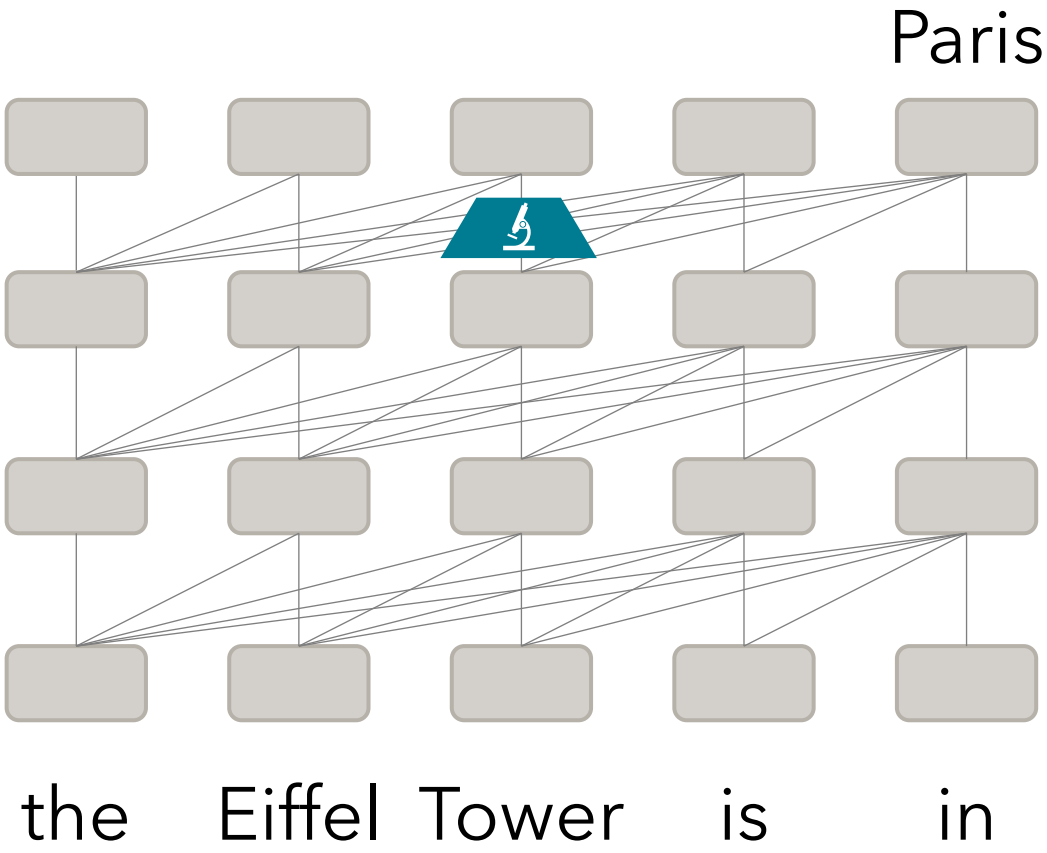
2

Unsupervised representation learning

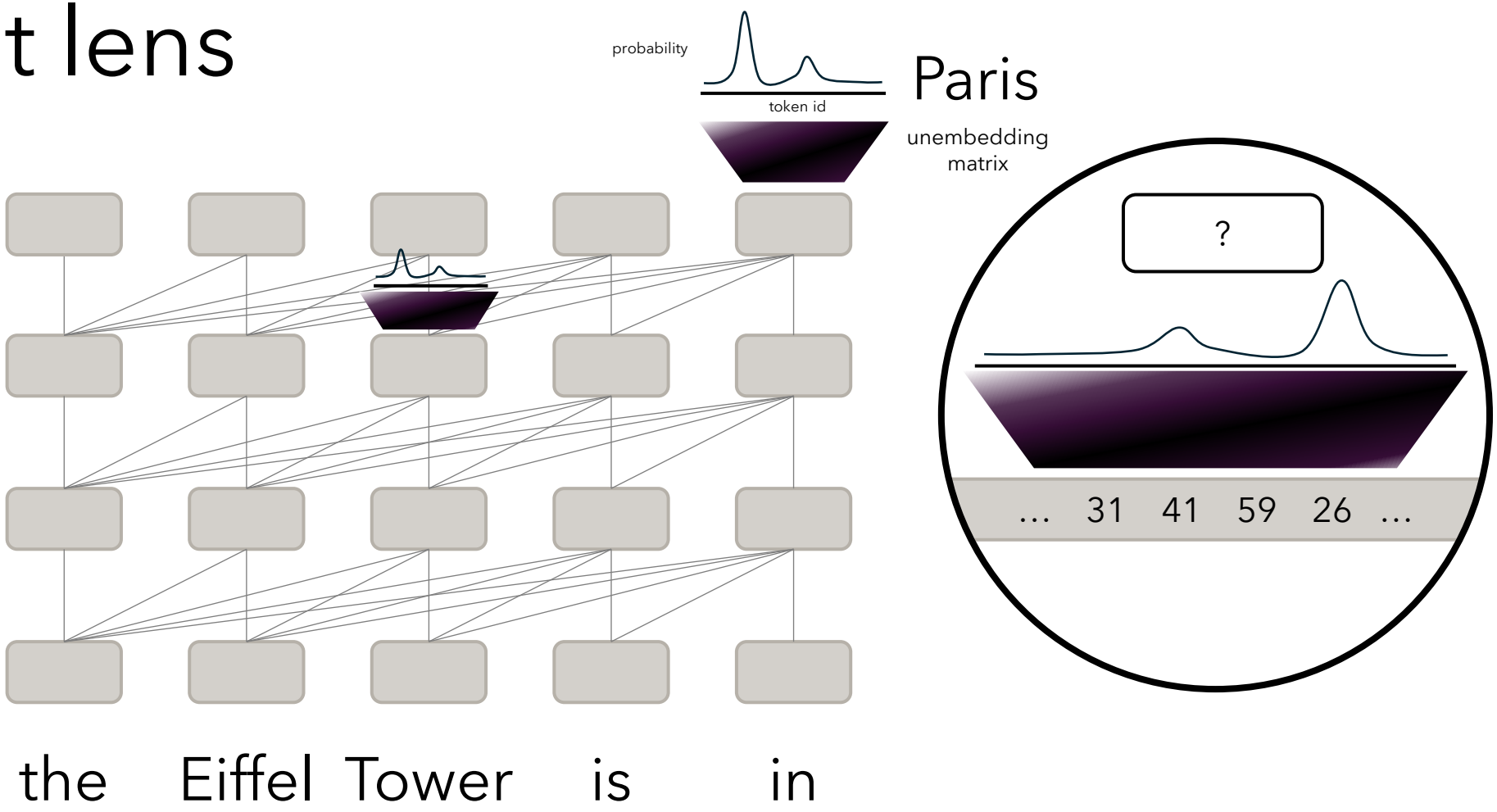
3

Supervised classification

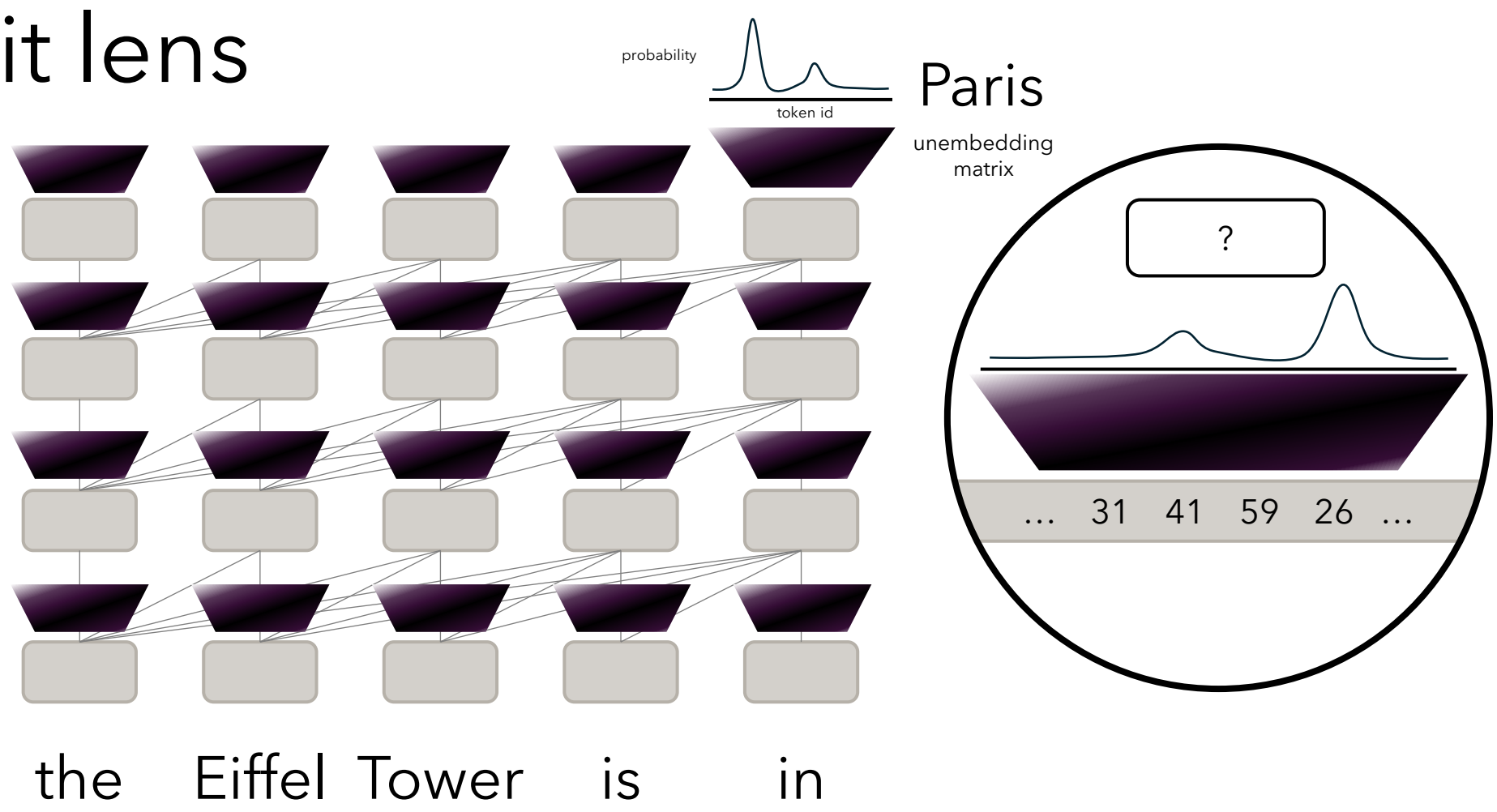
Goal: read information from activations



Logit lens

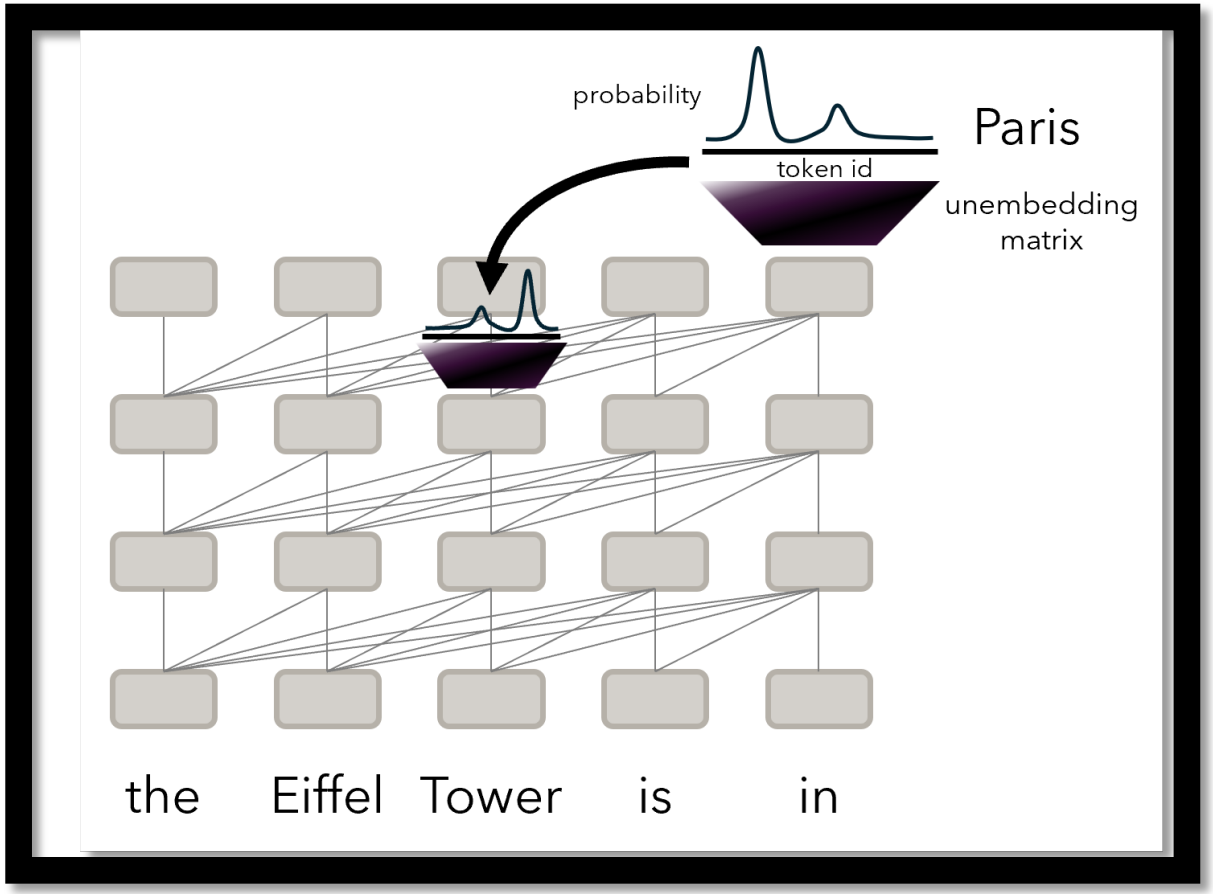

probability
token id
Paris
unembedding matrix
?
... 31 41 59 26 ...
the Eiffel Tower is in

Logit lens

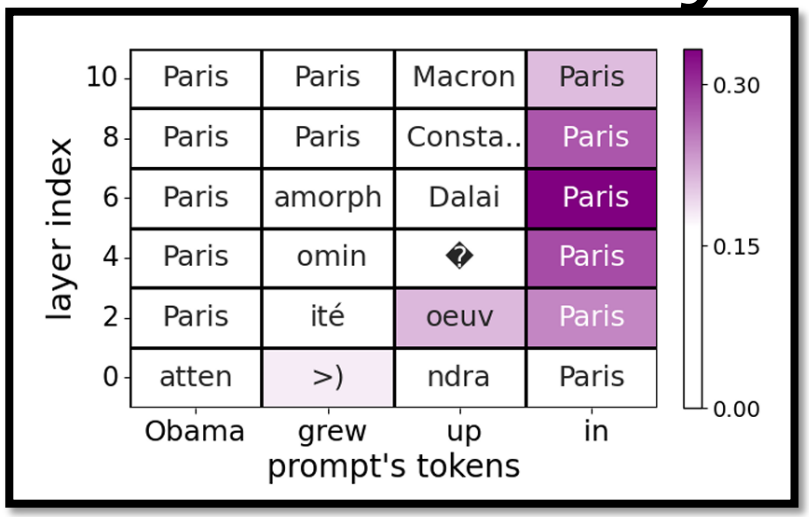

probability
token id
Paris
unembedding matrix
?
... 31 41 59 26 ...
the Eiffel Tower is in

Logit lens

code exercise

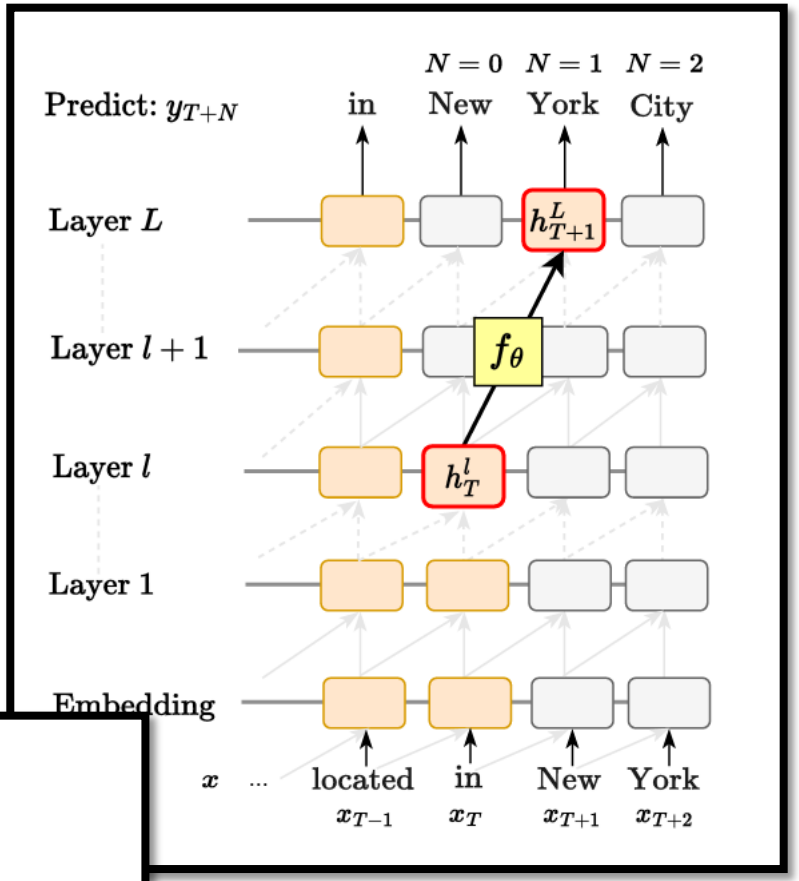


oh so many lenses!



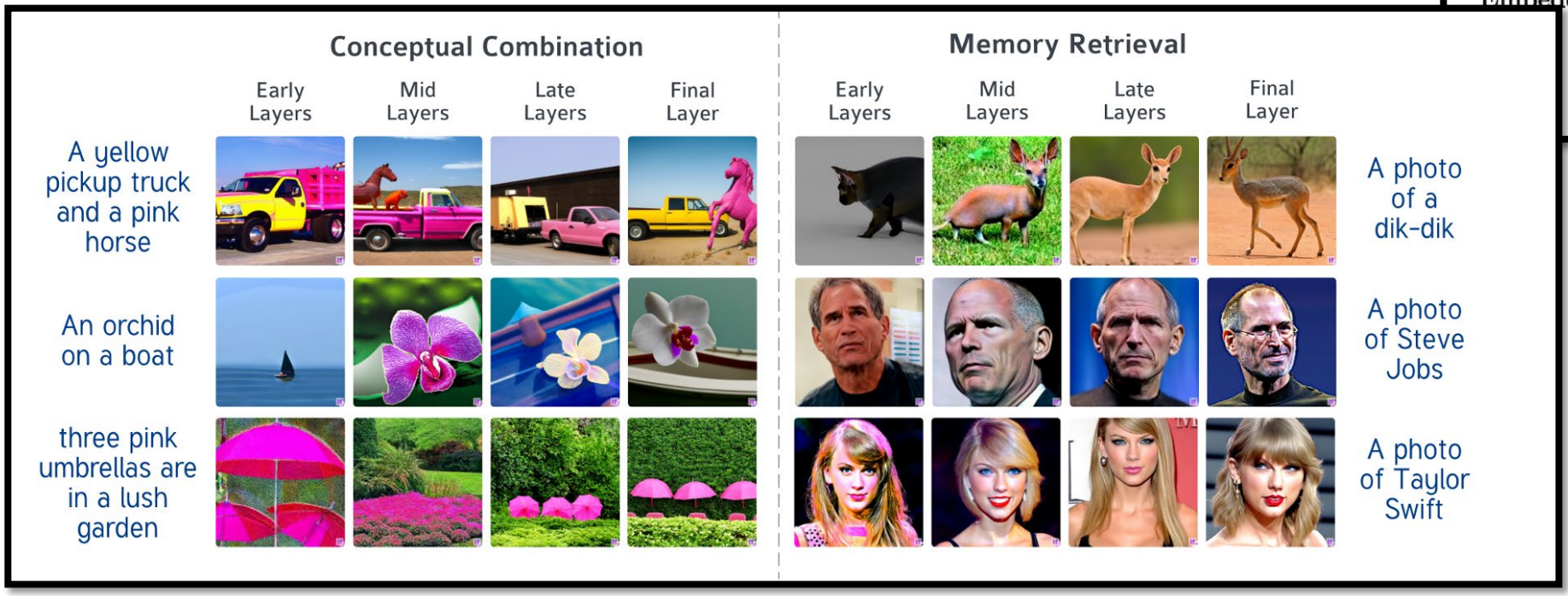
Katz et al. 2024 Backward lens

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.

Unsupervised methods

dimensionality reduction with PCA

Surveying the literature

1

Decoding information from existing structure

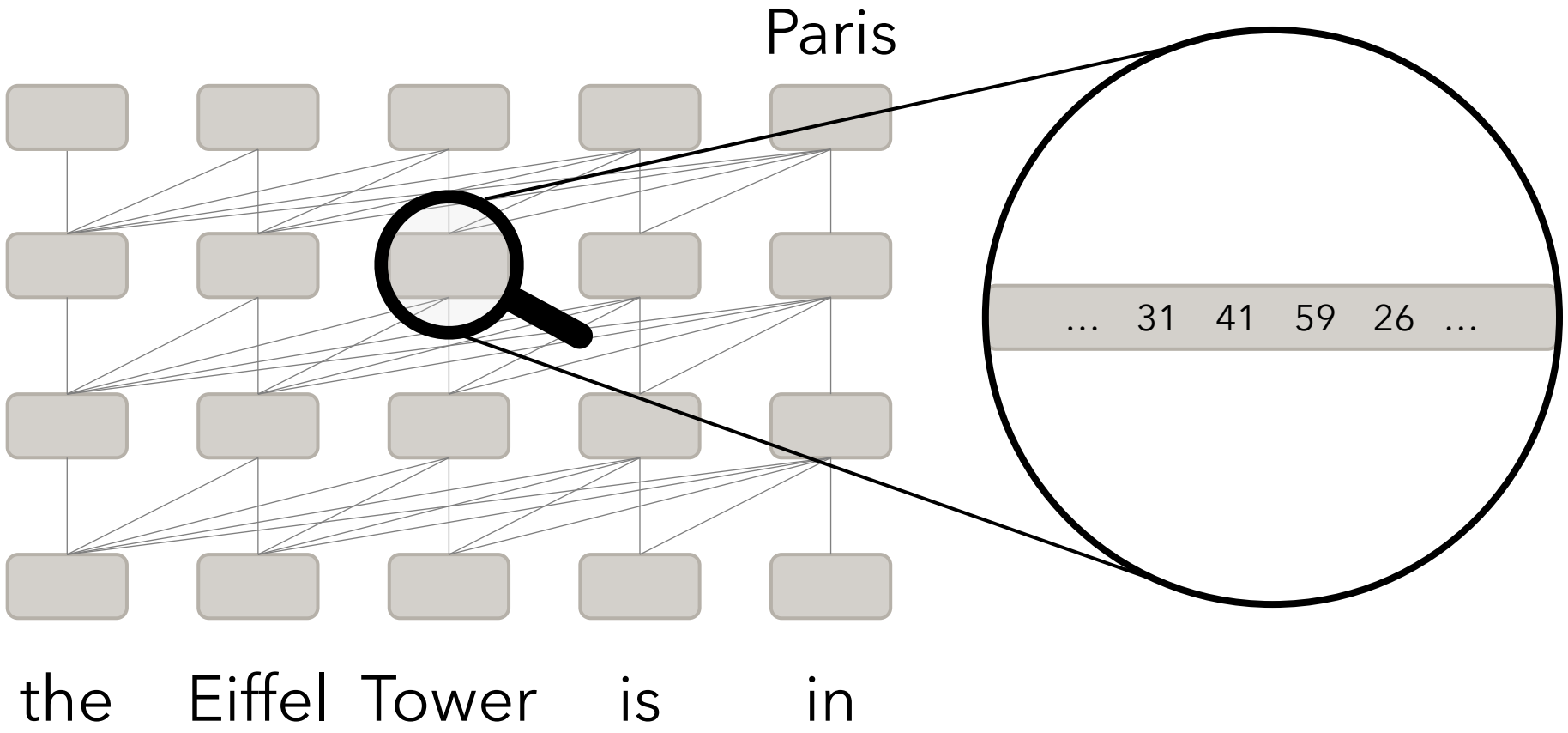
2

Unsupervised representation learning

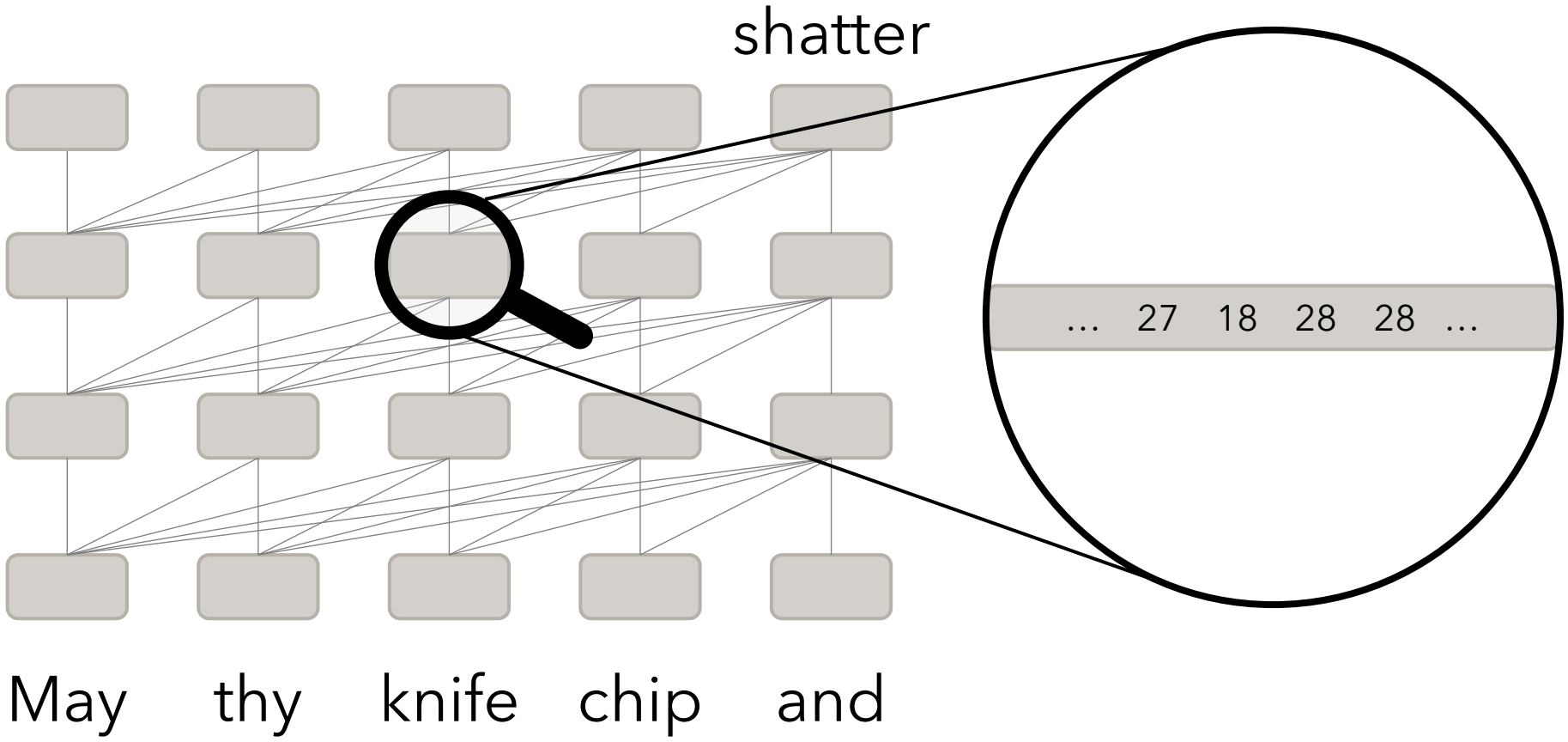
3

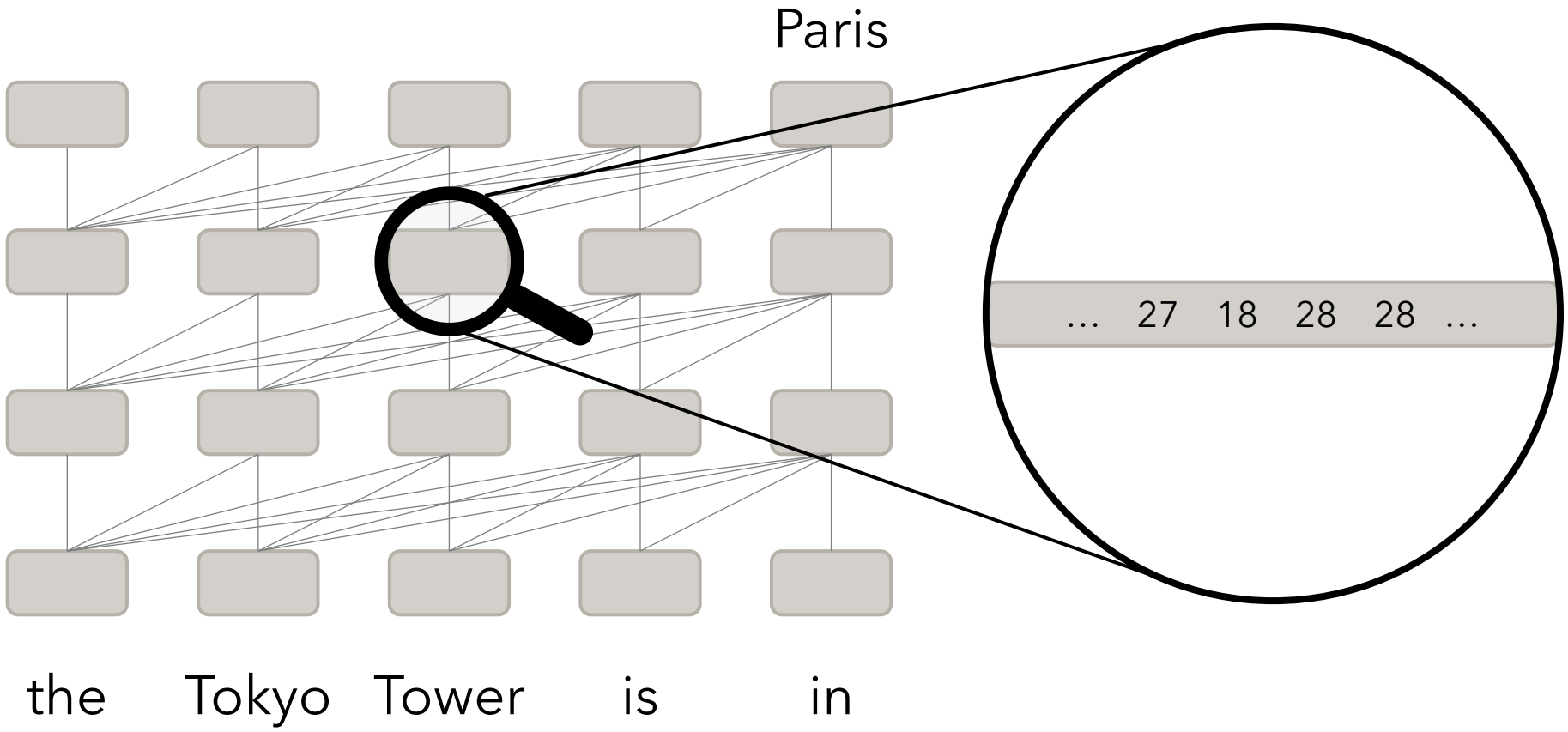
Supervised classification

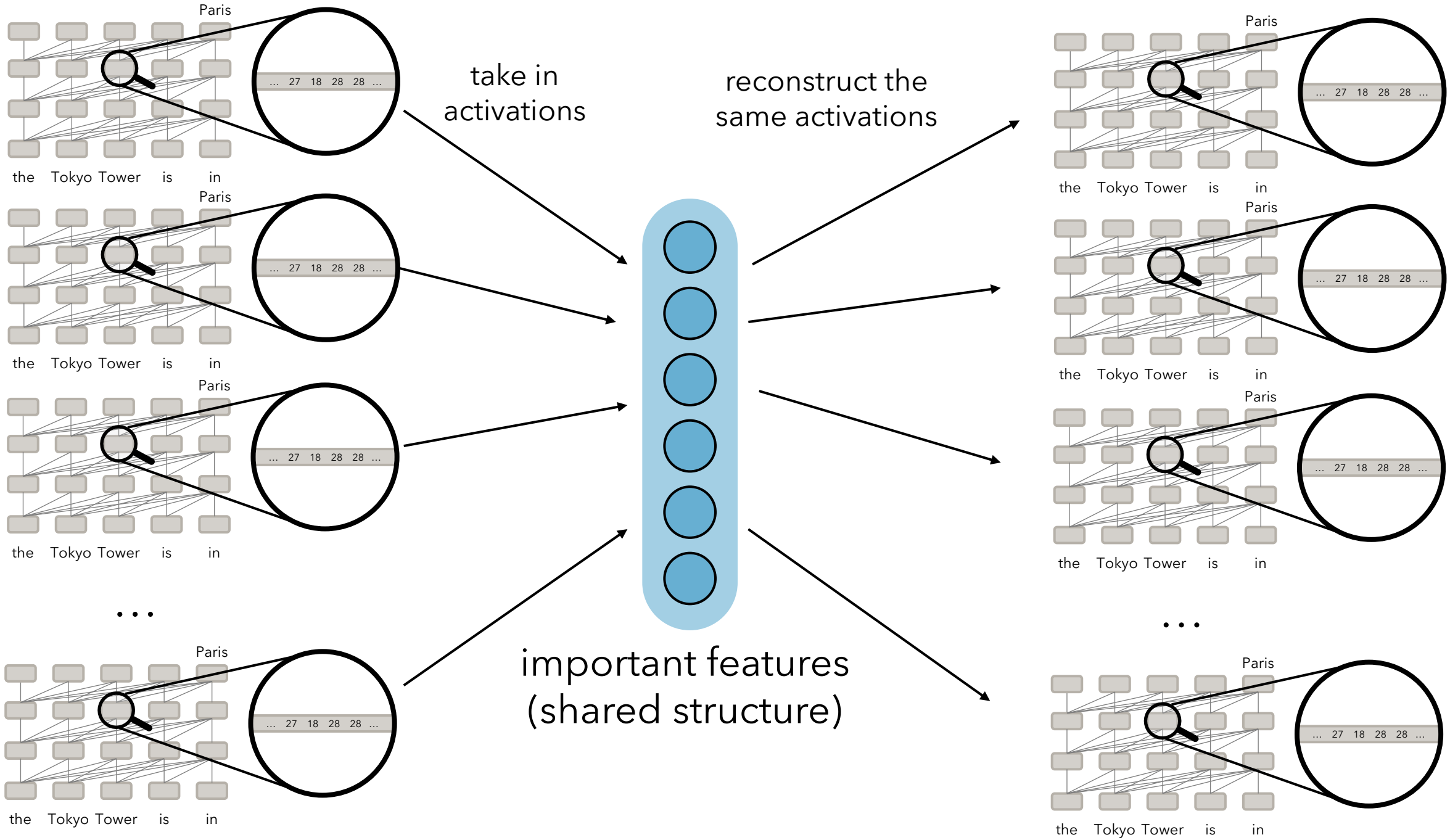
Different inputs → different activations



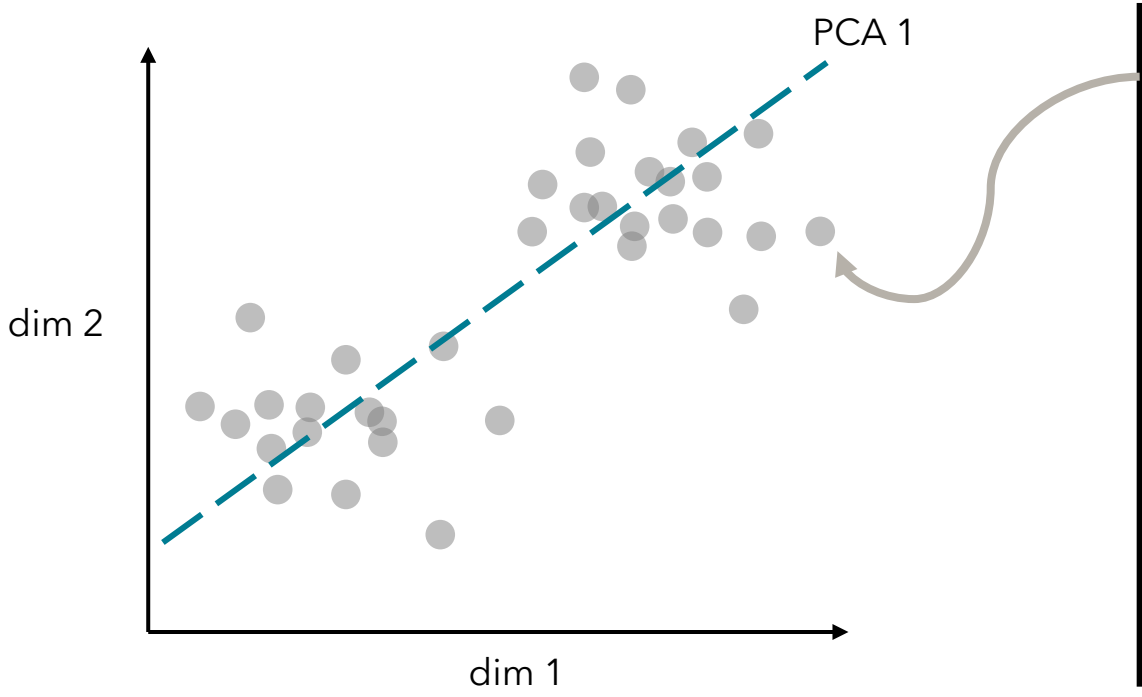
Different inputs → different activations

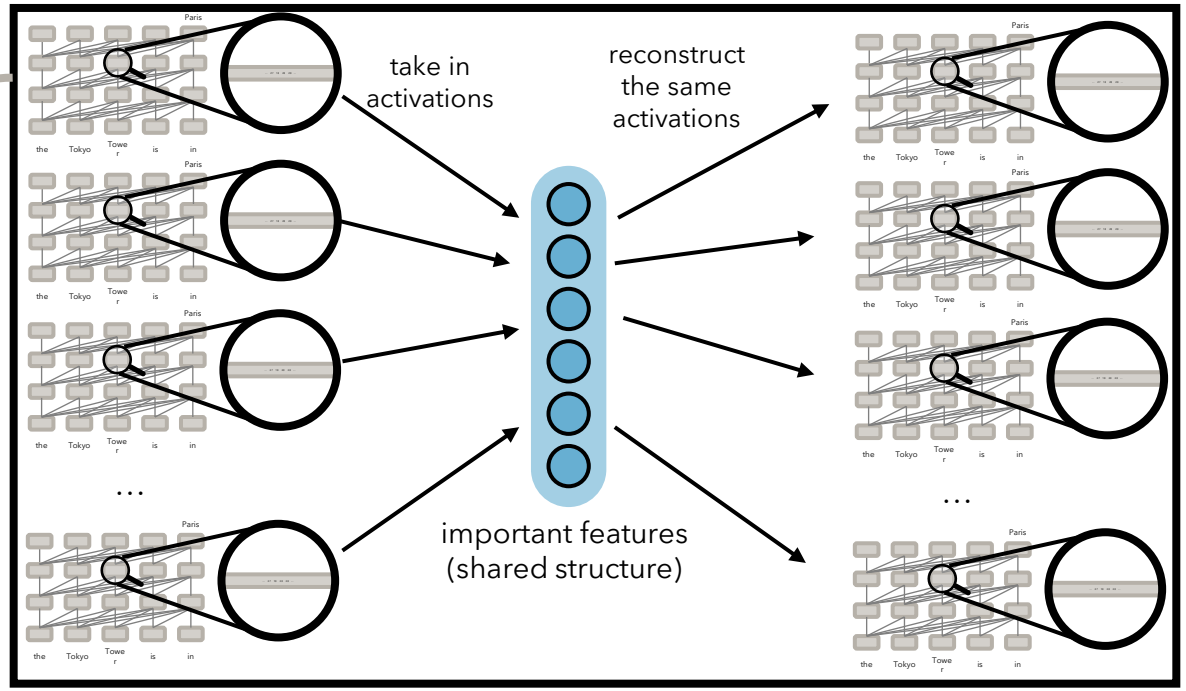



Paris
... 27 18 28 28 ...
the Tokyo Tower is in

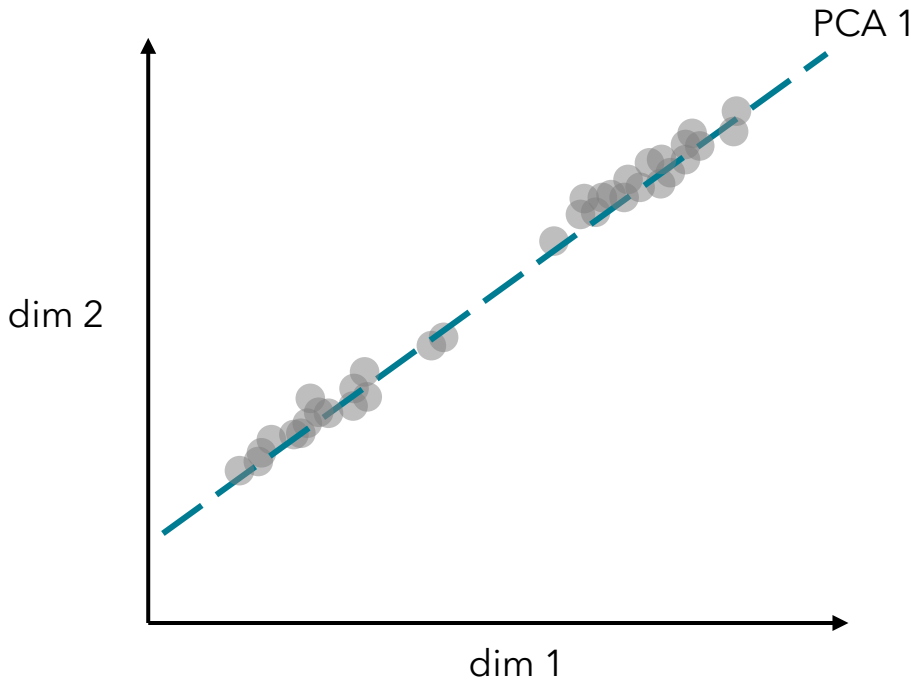

Paris
... 27 18 28 28 ...
the Tokyo Tower is in
take in activations
reconstruct the same activations
important features (shared structure)

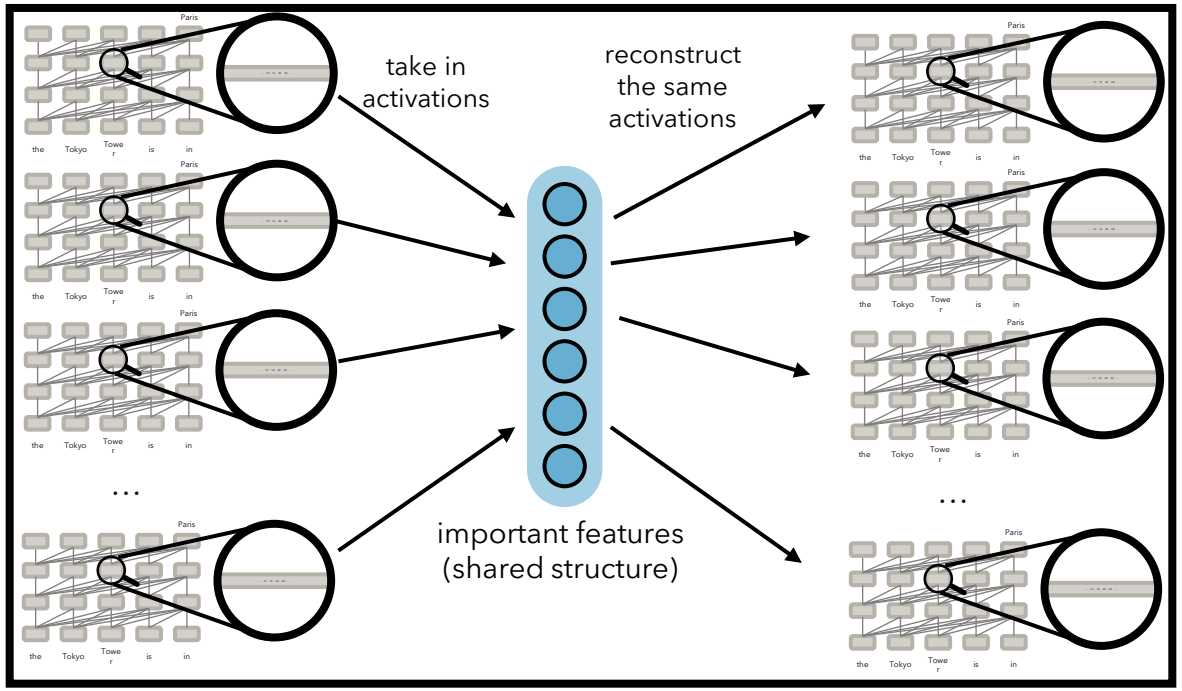
Dimensionality reduction: PCA



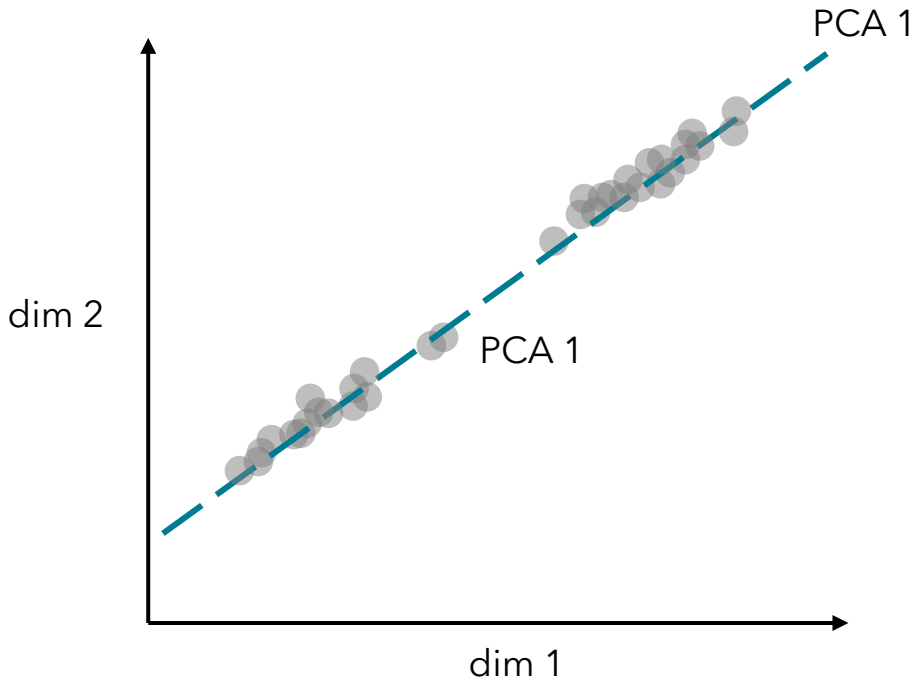


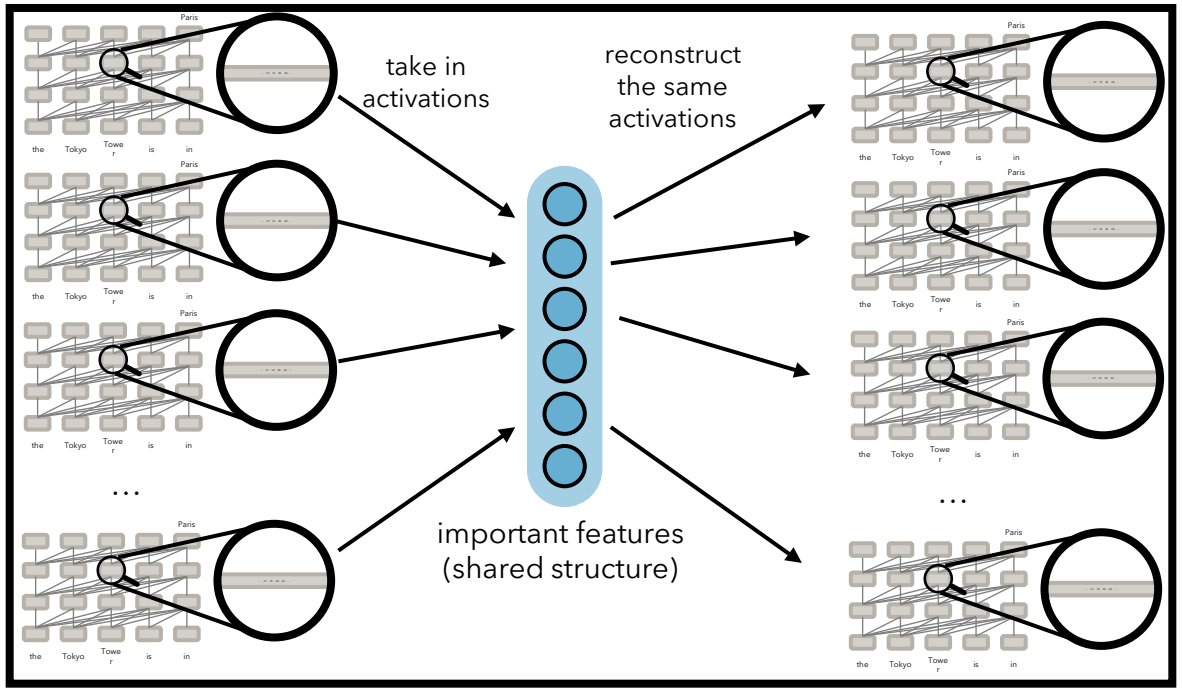
Dimensionality reduction: PCA





Dimensionality reduction: PCA

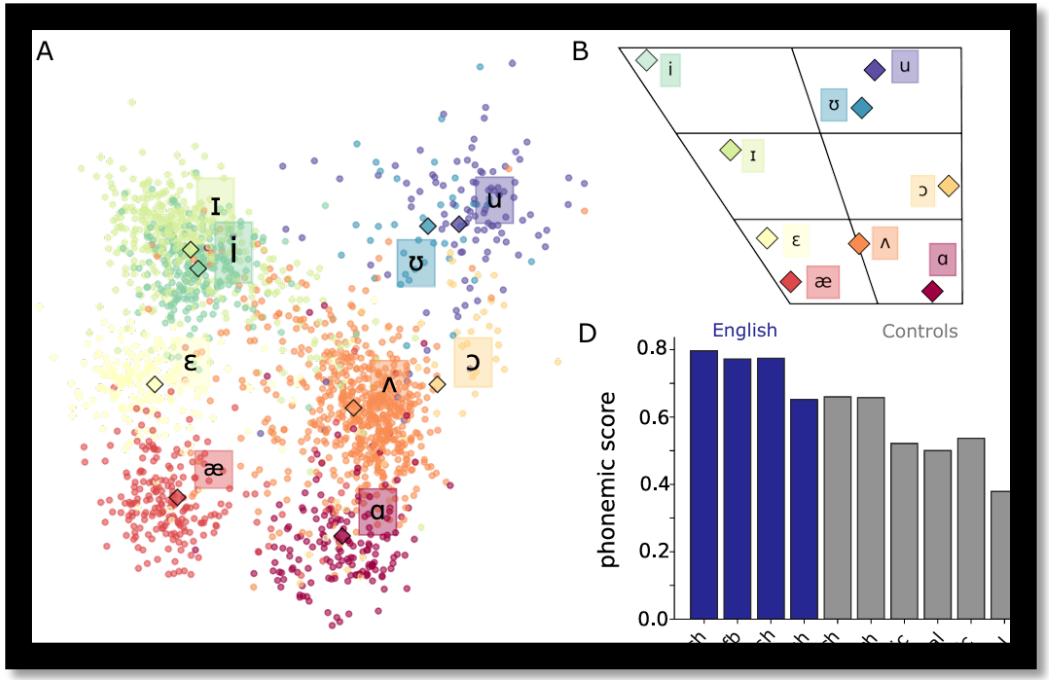


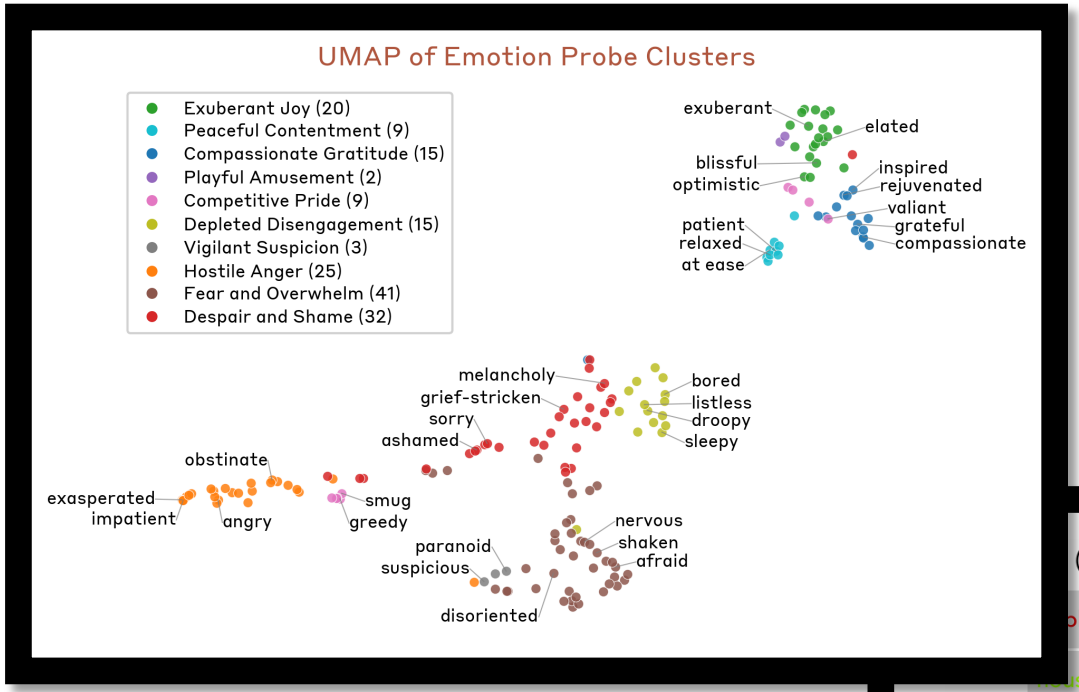


PCA in the wild

Orhan et al. 2026

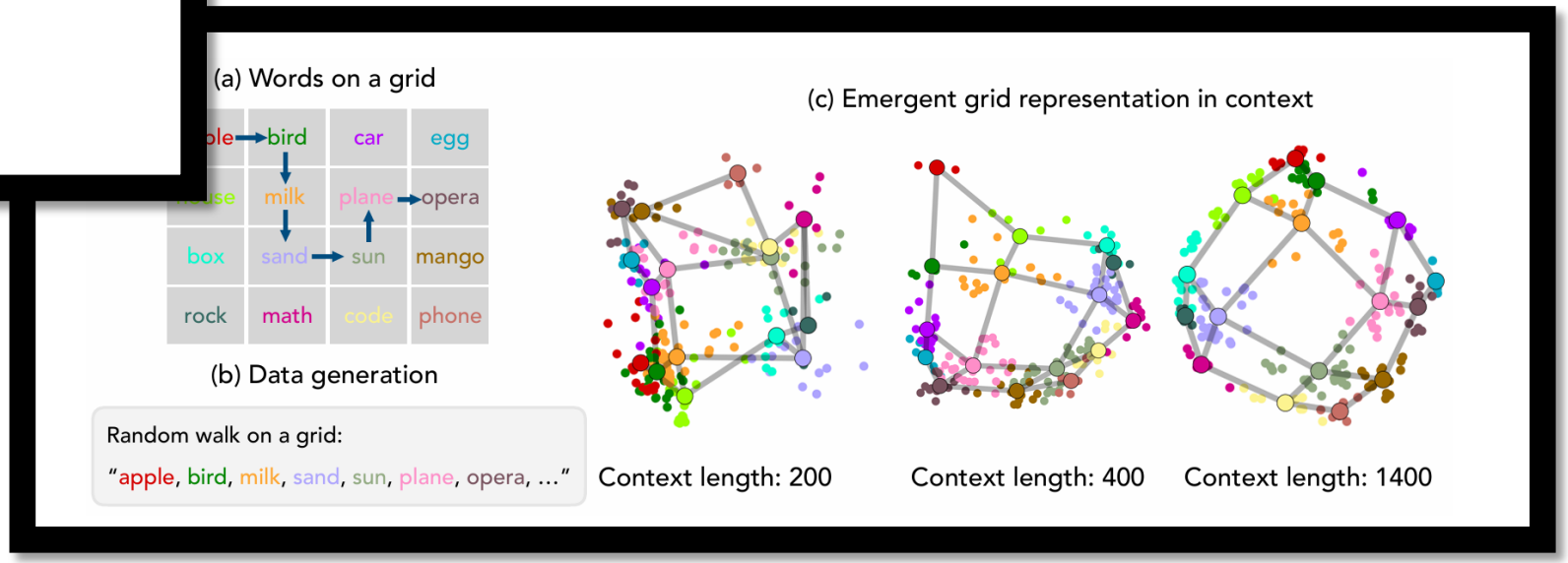
Emergence of linguistic structure in neural networks





Sofroniew et al. 2026 (Anthropic)

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



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

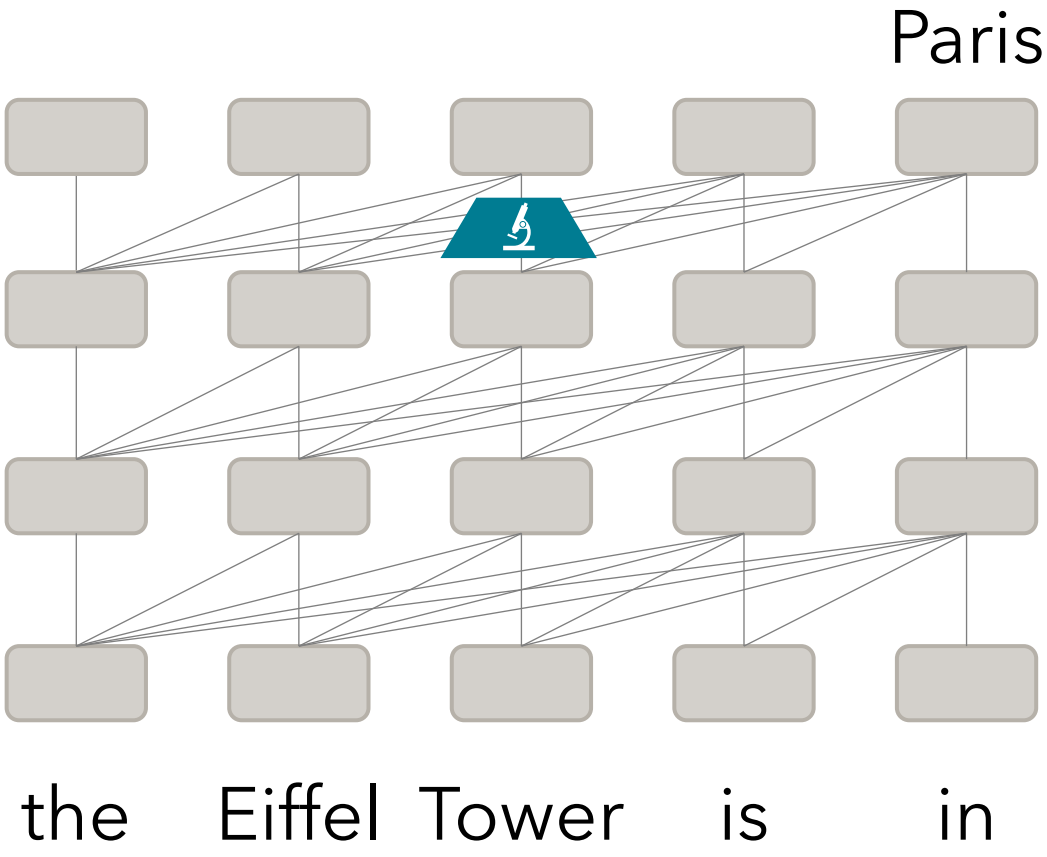
2

Unsupervised representation learning

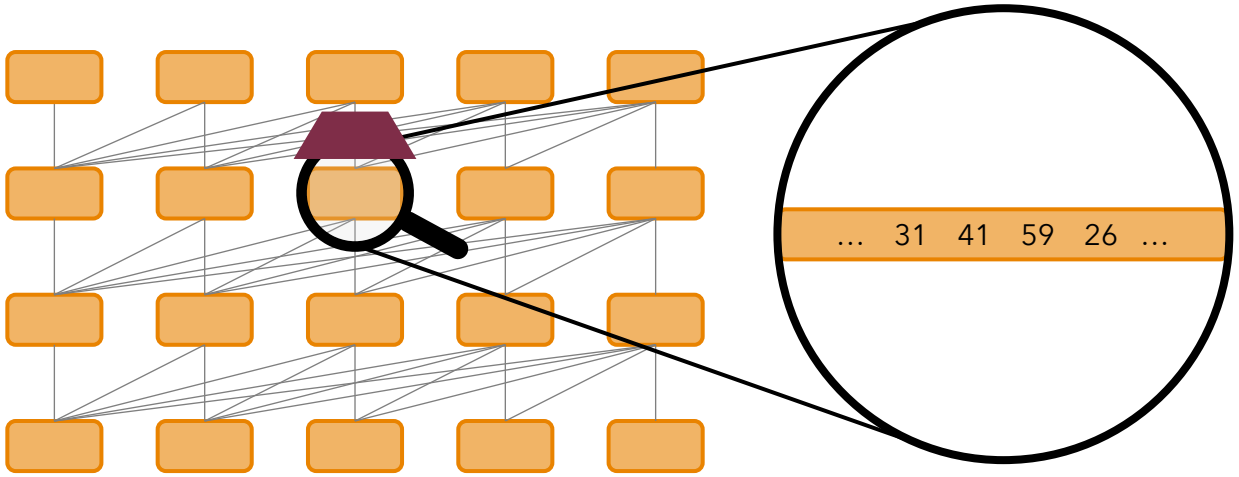
3

Supervised classification

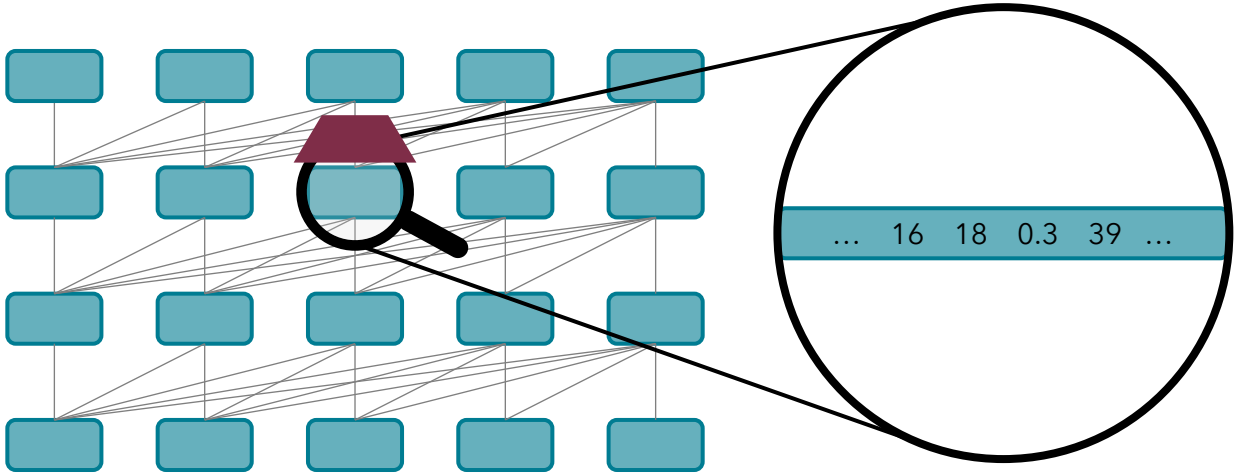
Goal: read information from activations



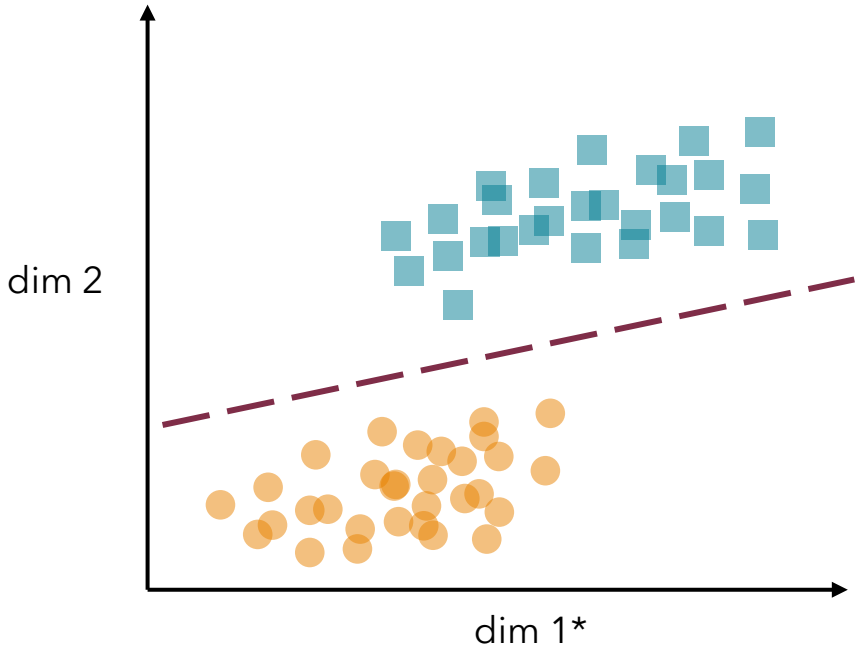
Why not just train for this directly?



How are you doing ?



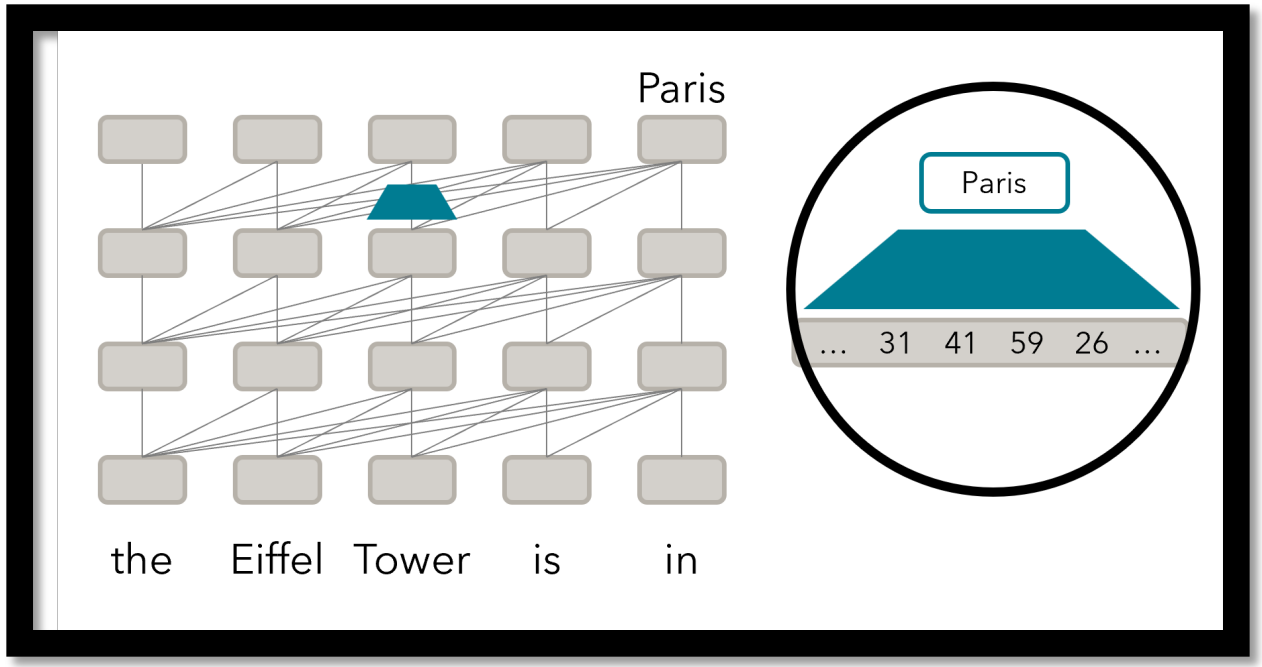
¿ Como es tas ?



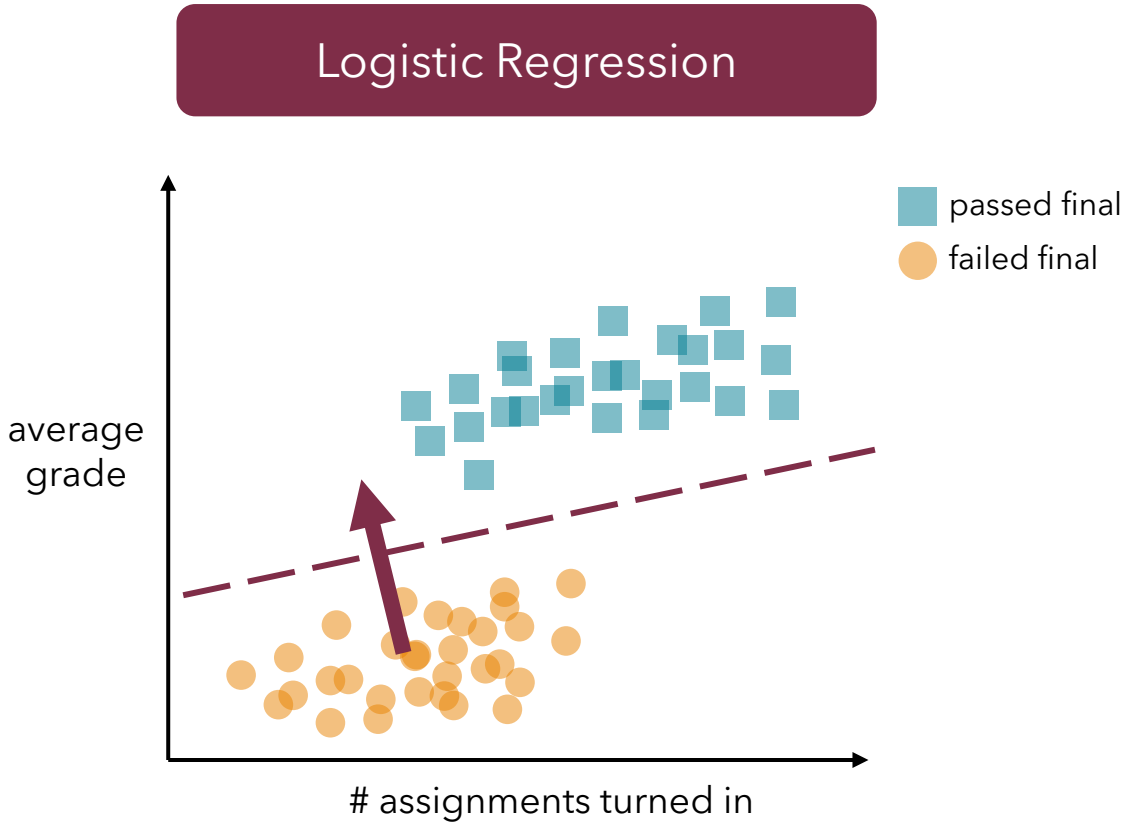
* in reality, these vectors live in many more dimensions!

PCA & supervised probes

code exercise



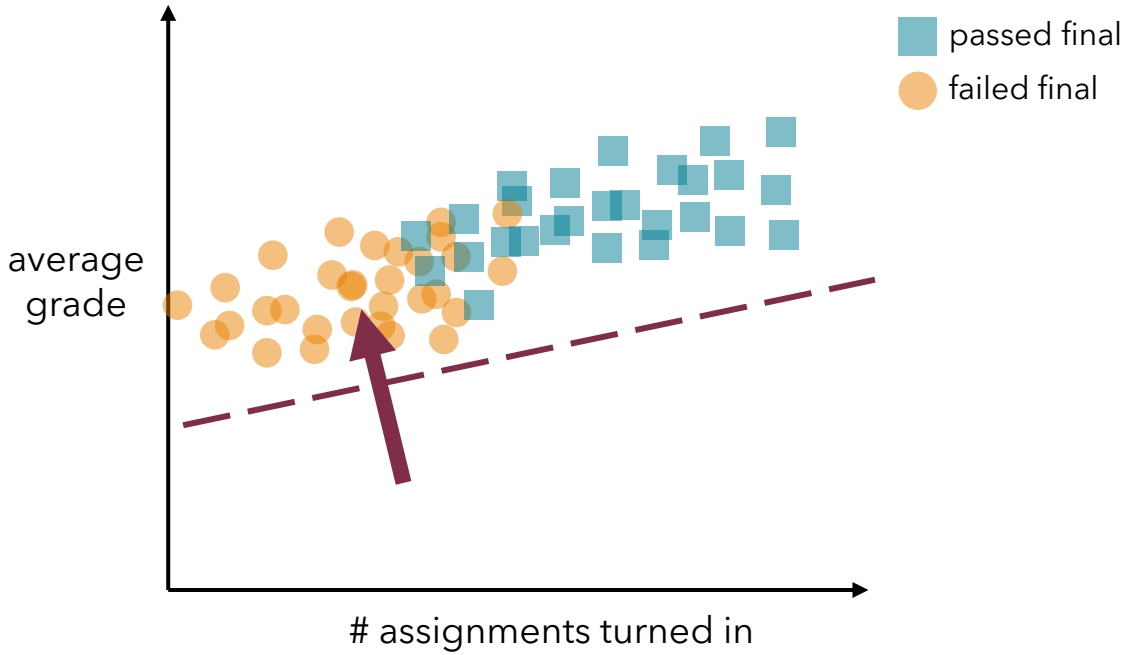
a quick tangent
deriving interventions from activations



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?

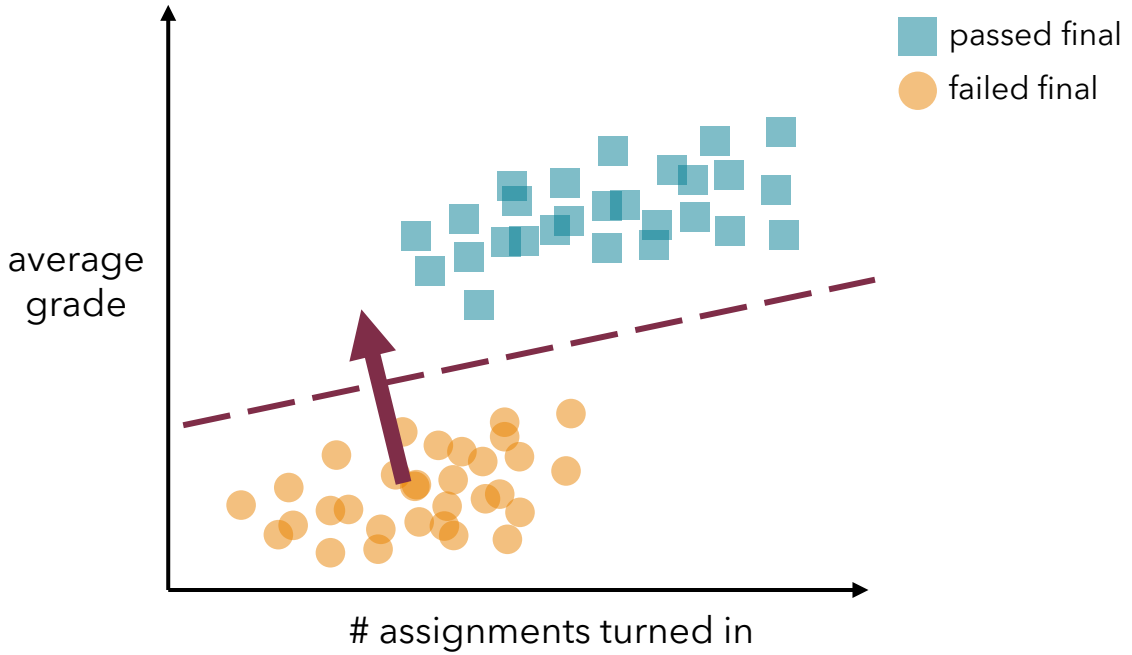
The story

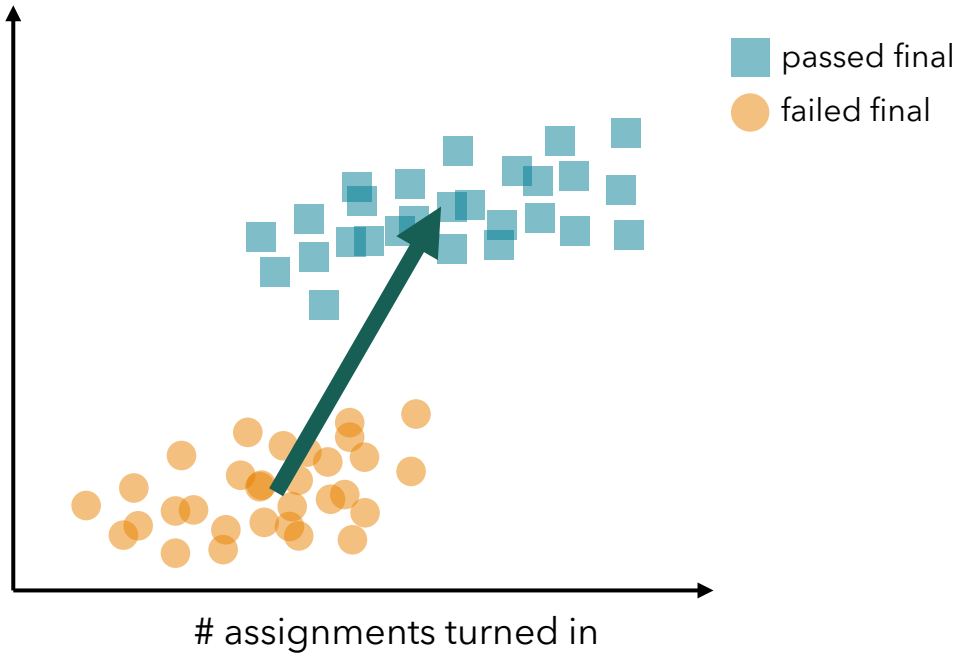
- Professor X wants to help students pass their final exams
- Students who pass the final tend to…
  - Turn in more assignments
  - Get a higher grade
- Students who turn in more assignments tends to get better grades
- What should Prof. X do to improve student's chances of passing the final?

Logistic Regression

vs.

Difference in Means



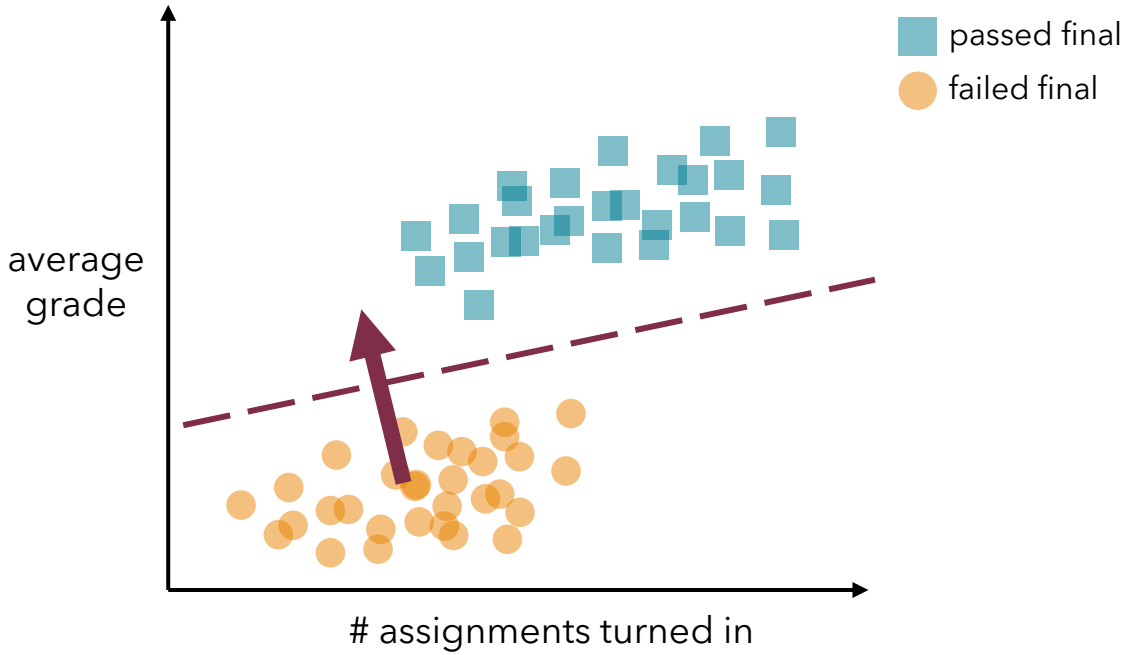


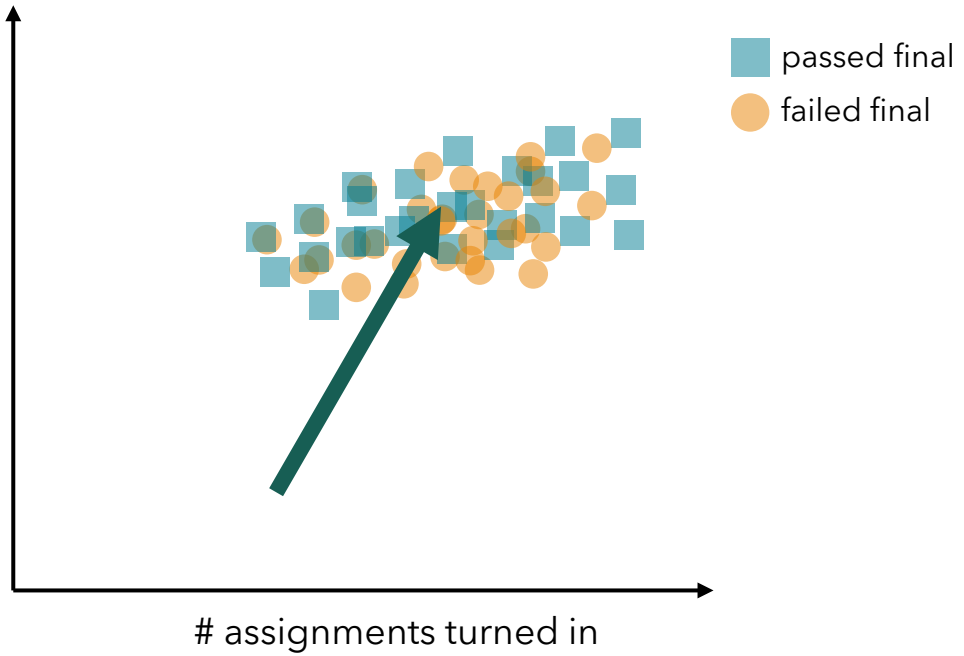
- Students who turn in more assignments tends to get better grades
- What should Prof. X do to improve student's chances of passing the final?

Logistic Regression

vs.

Difference in Means



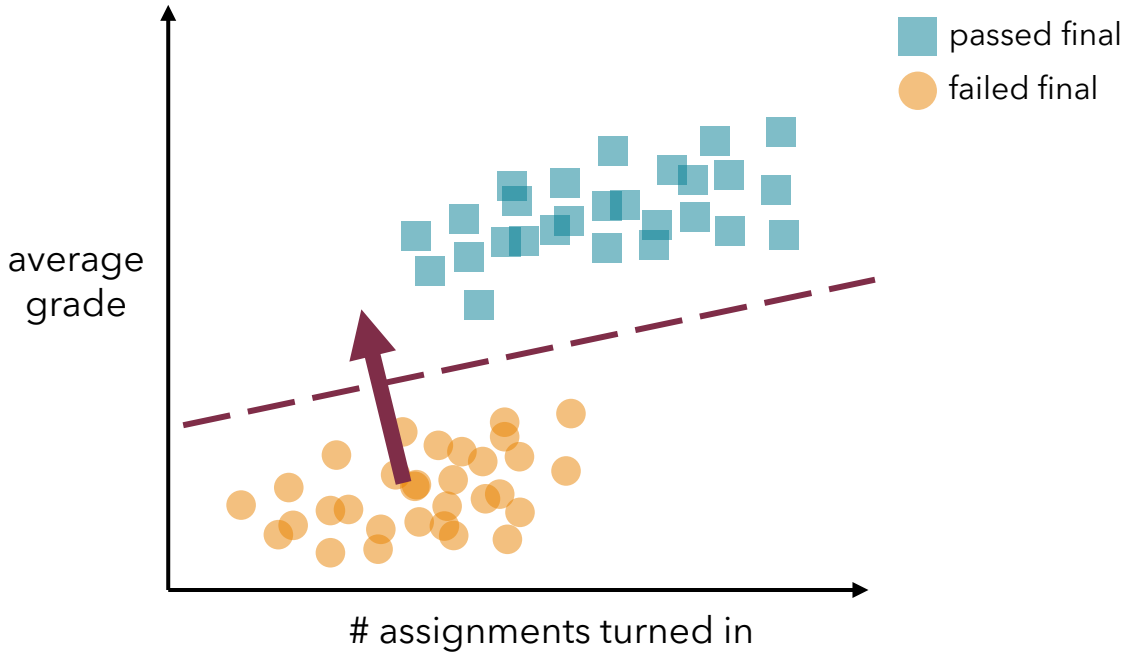


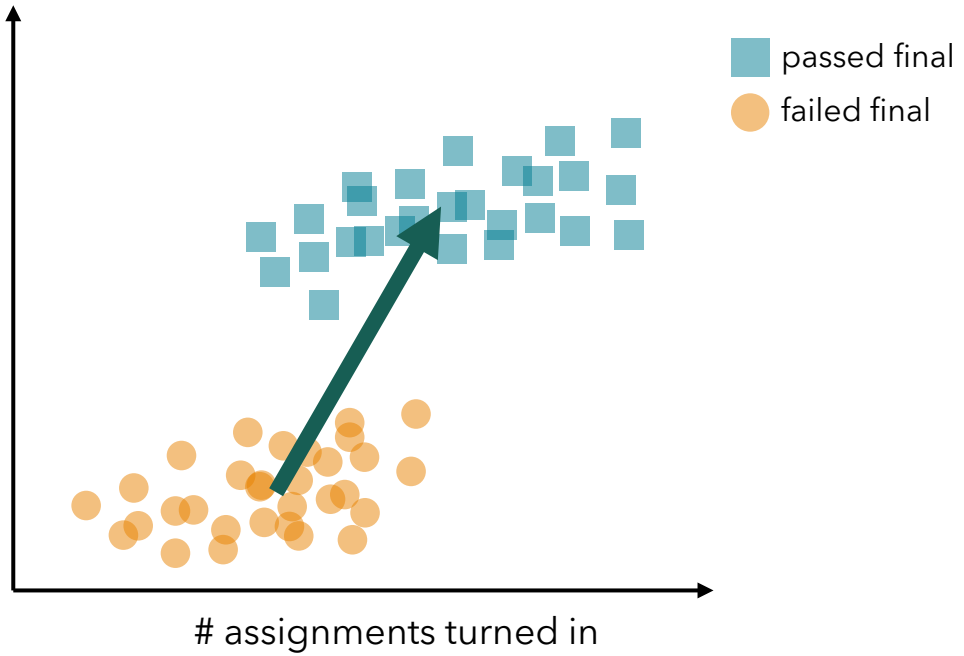
- Students who turn in more assignments tends to get better grades
- What should Prof. X do to improve student's chances of passing the final?

Logistic Regression

vs.

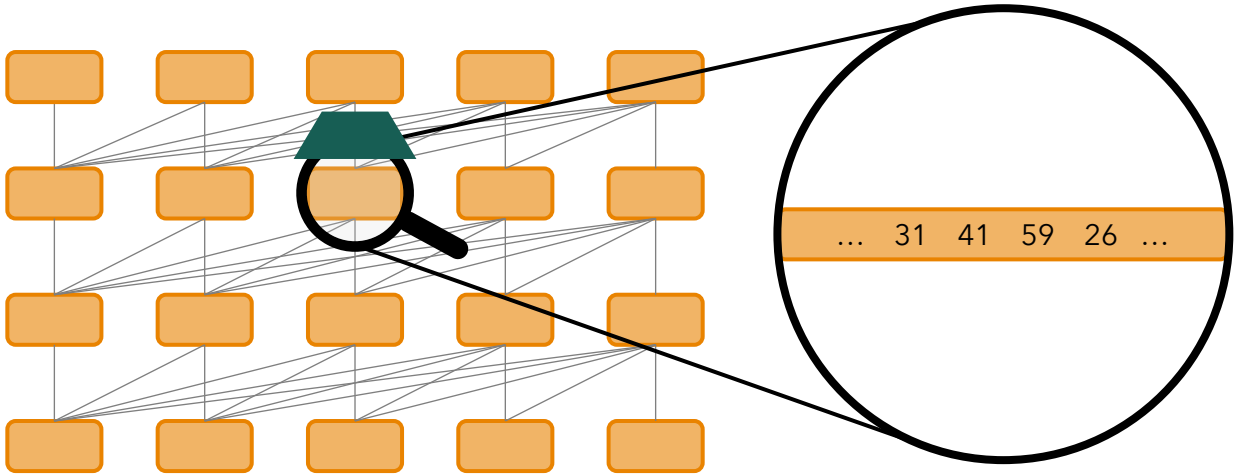
Difference in Means



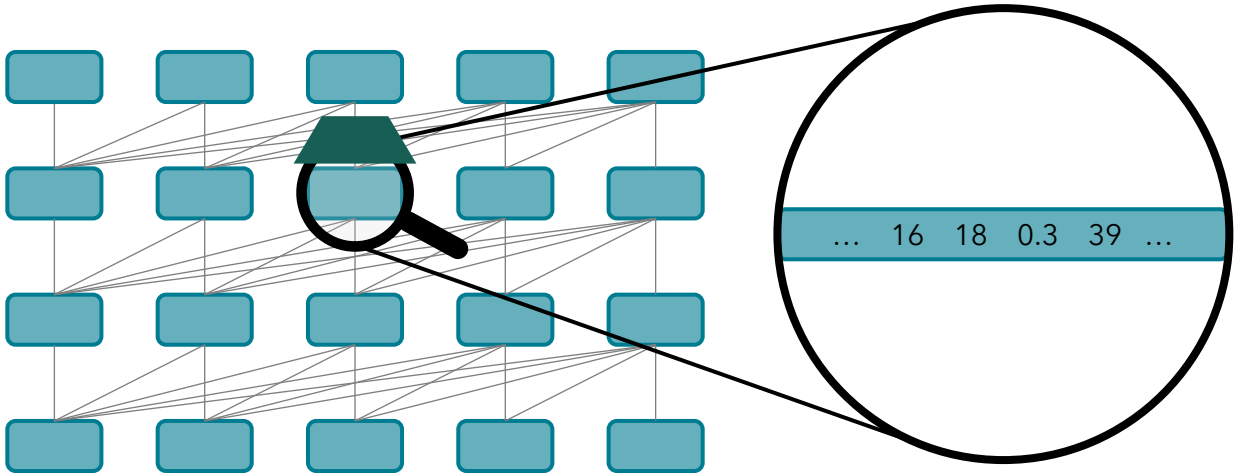


1. Correlation isn't causation! Probes don't inform causal interventions.
2. Difference in means is an example of *activation transport*

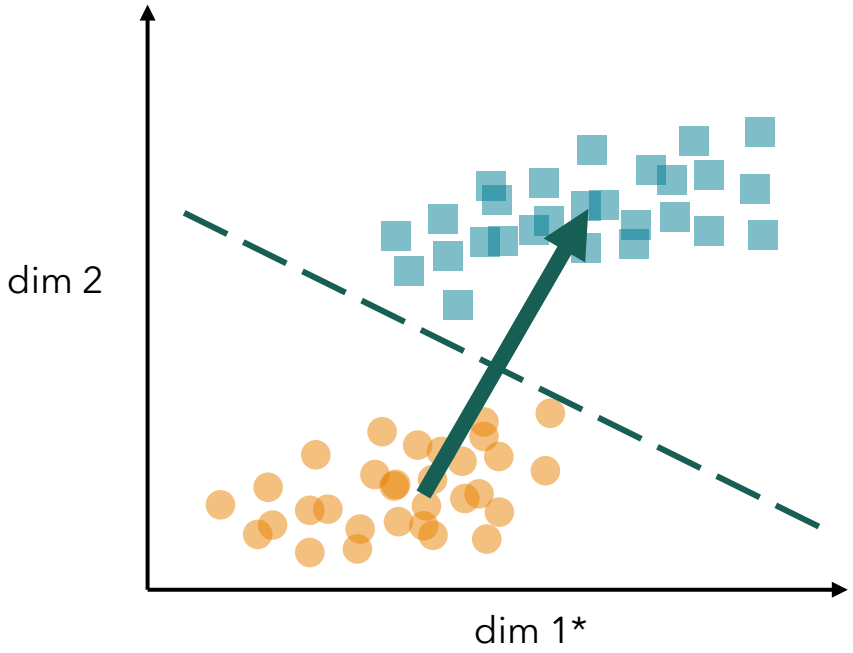
okay, we're back
steering activations with difference-in-means



How are you doing ?



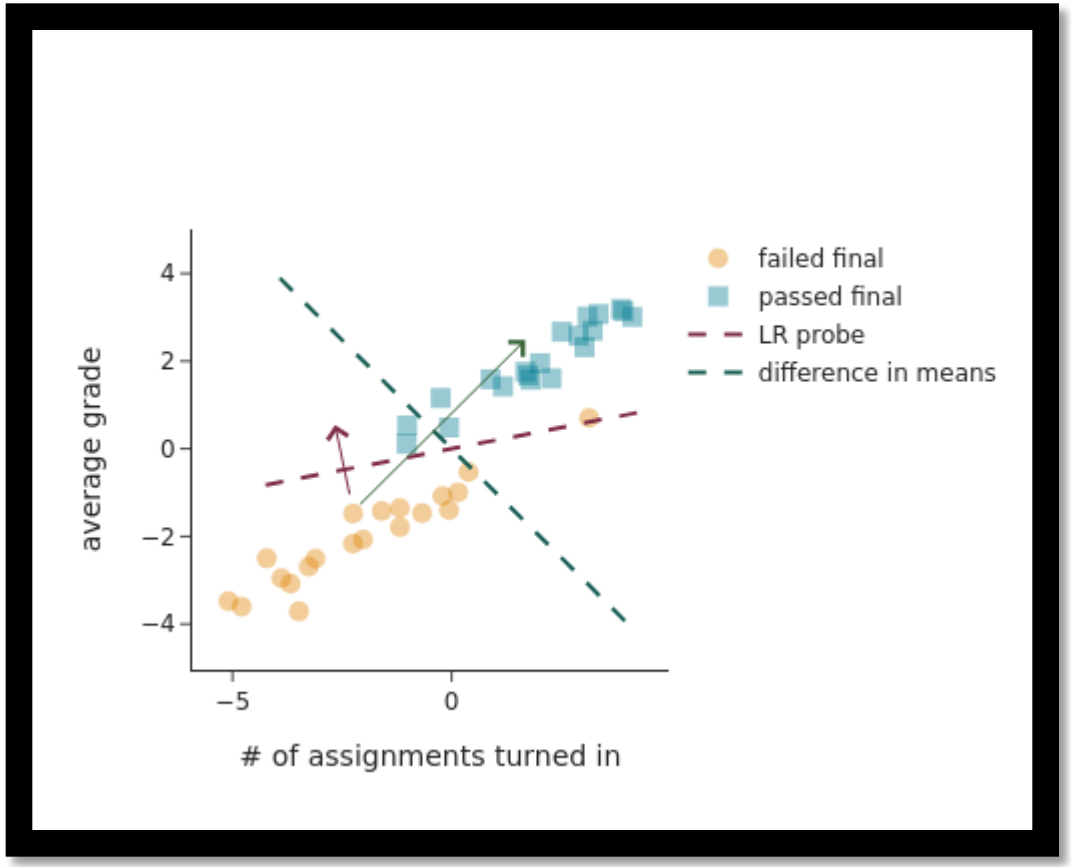
¿ Como es tas ?



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

Difference in means steering

code exercise



Does this look familiar?

Sofroniew et al. 2026 (Anthropic)

Surveying the literature

1

Decoding information from existing structure

2

Unsupervised representation learning

3

Supervised classification