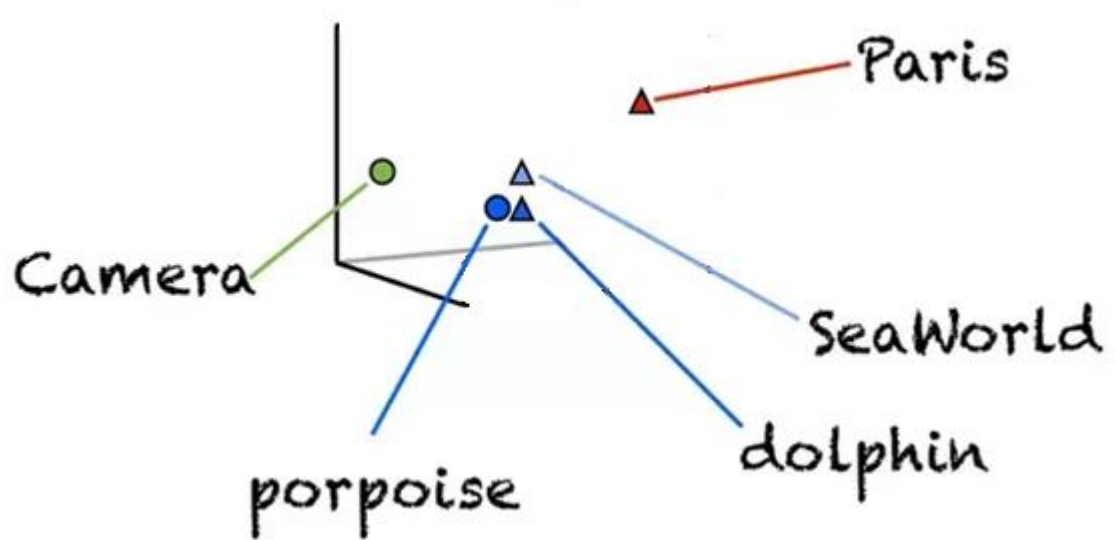


# On the Downstream Performance of Compressed Word Embeddings

Avner May, Jian Zhang, Tri Dao, Chris Ré  
Stanford University



# Word Embeddings

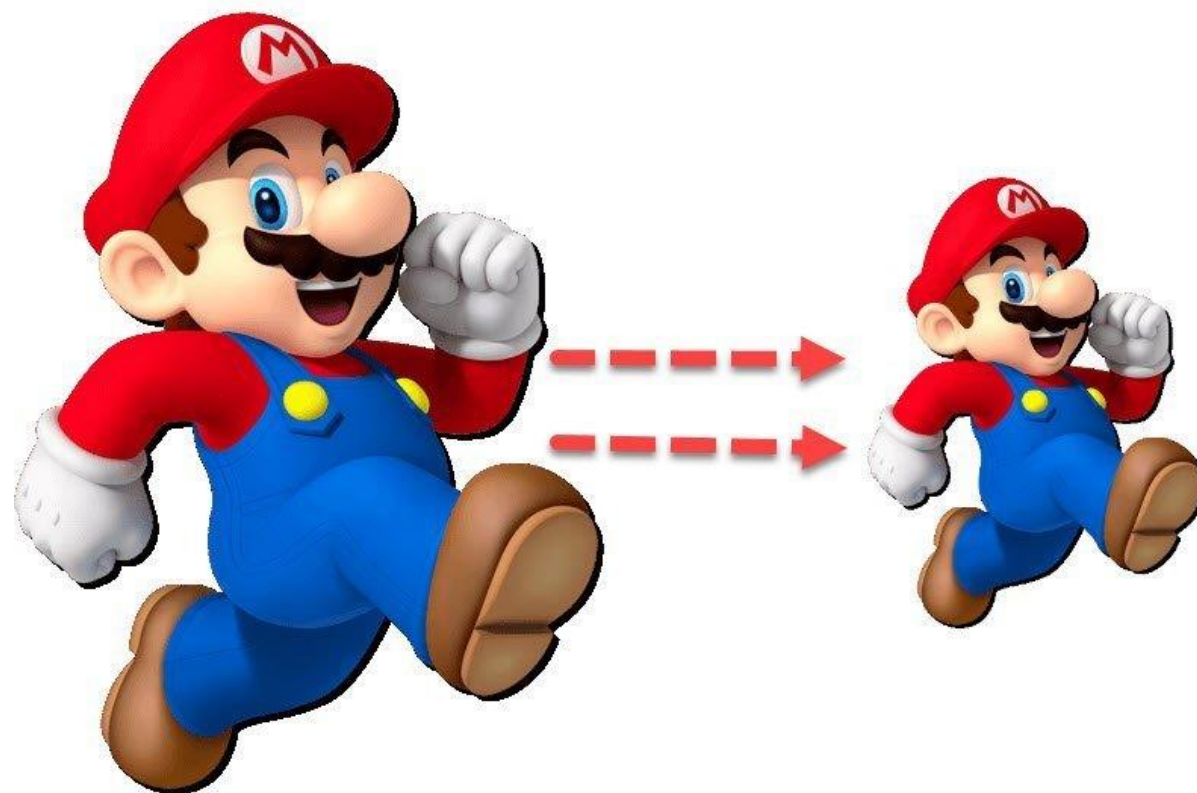


**Important for strong NLP performance**

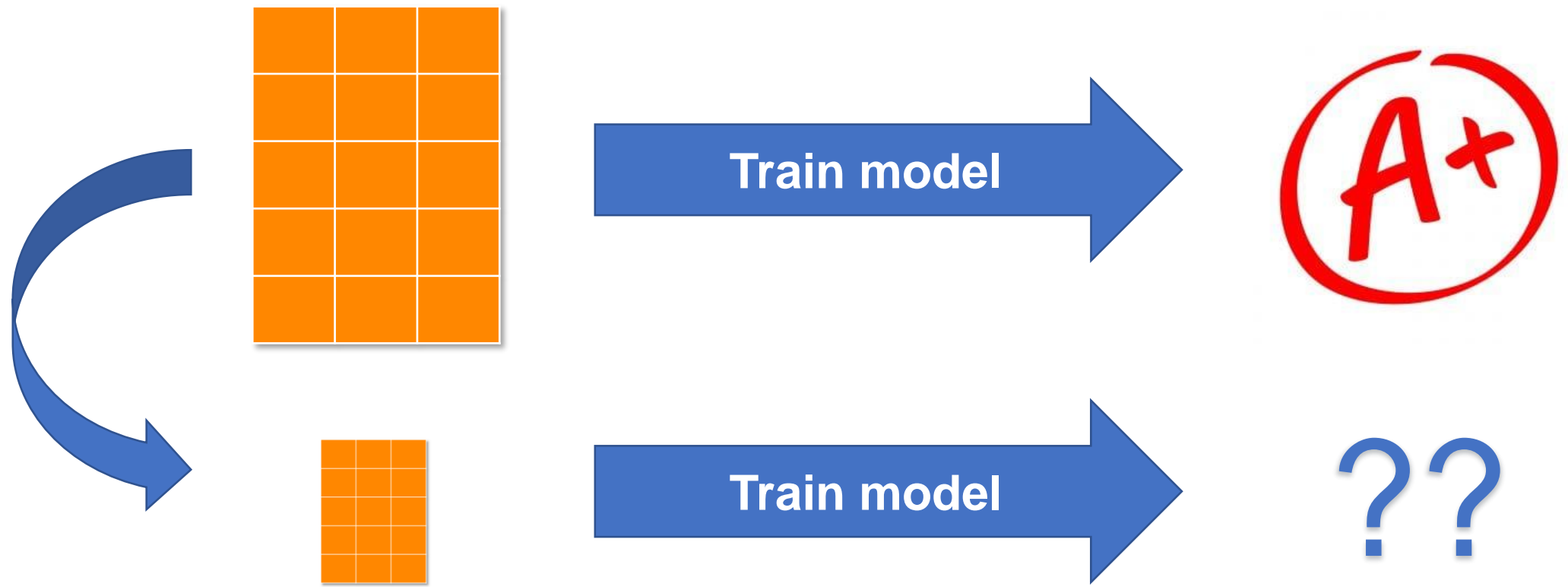


**Take a lot of memory**

# Word Embedding Compression

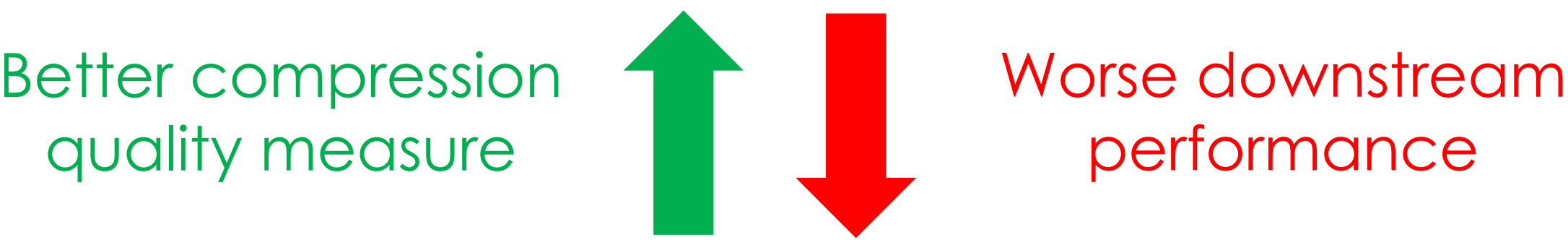


# What determines whether a compressed embedding matrix will perform well on downstream tasks?



# Motivating Observation

Existing ways of measuring compression quality often fail to explain relative downstream performance.



# Our Contributions: Outline

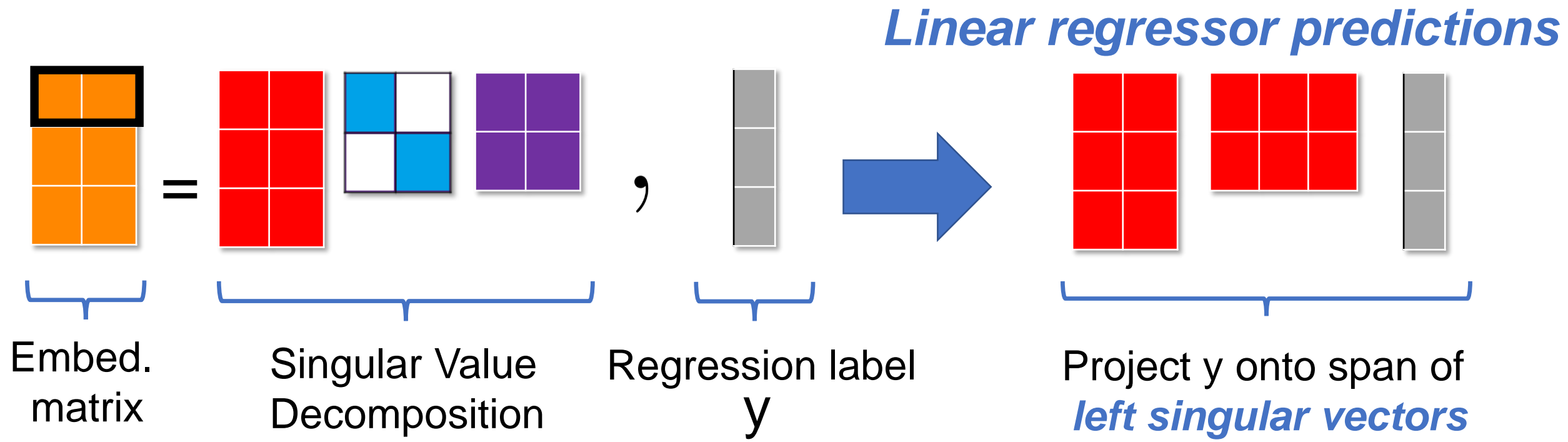
- ① Define a **new measure** of compression quality.
- ② Prove **generalization bounds** using this measure.
- ③ Show strong **empirical correlation** w. downstream performance.
- ④ Use measure to **select** compressed embeddings.

**Up to 2x lower selection error rates  
than the next best measure.**

# Defining the Measure: Intuition from Linear Regression

## Observation:

Predictions are a projection onto the span of *left singular vectors*.



# Defining the Measure: Eigenspace Overlap Score (EOS)

*Intuition:*  
Measures similarity between the span of *left singular vectors*.

$$\mathcal{E} \left( \underbrace{\begin{matrix} d \\ \text{[Red 2x2 grid]} \quad \text{[Blue 2x2 grid]} \quad \text{[Purple 2x2 grid]} \end{matrix}}_{\text{Uncompressed embed. SVD}}, \underbrace{\begin{matrix} \text{[Red 2x2 grid]} \quad \text{[Light Blue 2x2 grid]} \quad \text{[Purple 2x2 grid]} \end{matrix}}_{\text{Compressed embed. SVD}} \right) = \underbrace{\frac{1}{d} \left\| \begin{matrix} \text{[Red 2x2 grid]} & \text{[Red 2x2 grid]} \end{matrix} \right\|_F^2}_{\text{Eigenspace overlap score}}$$



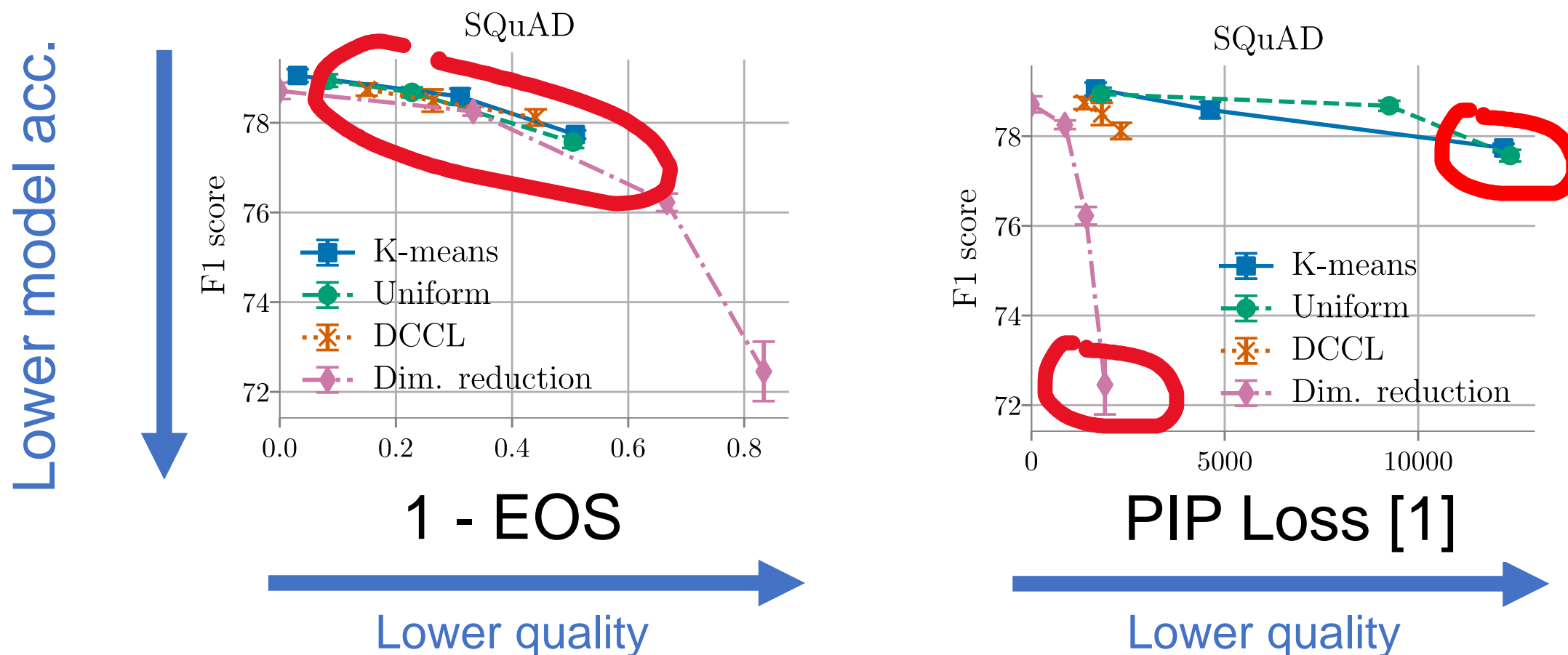
# Theoretical Results (informal): Linear Regression

Expected difference between  
*compressed* and *uncompressed*  
test mean-squared error  
is **determined by EOS.**



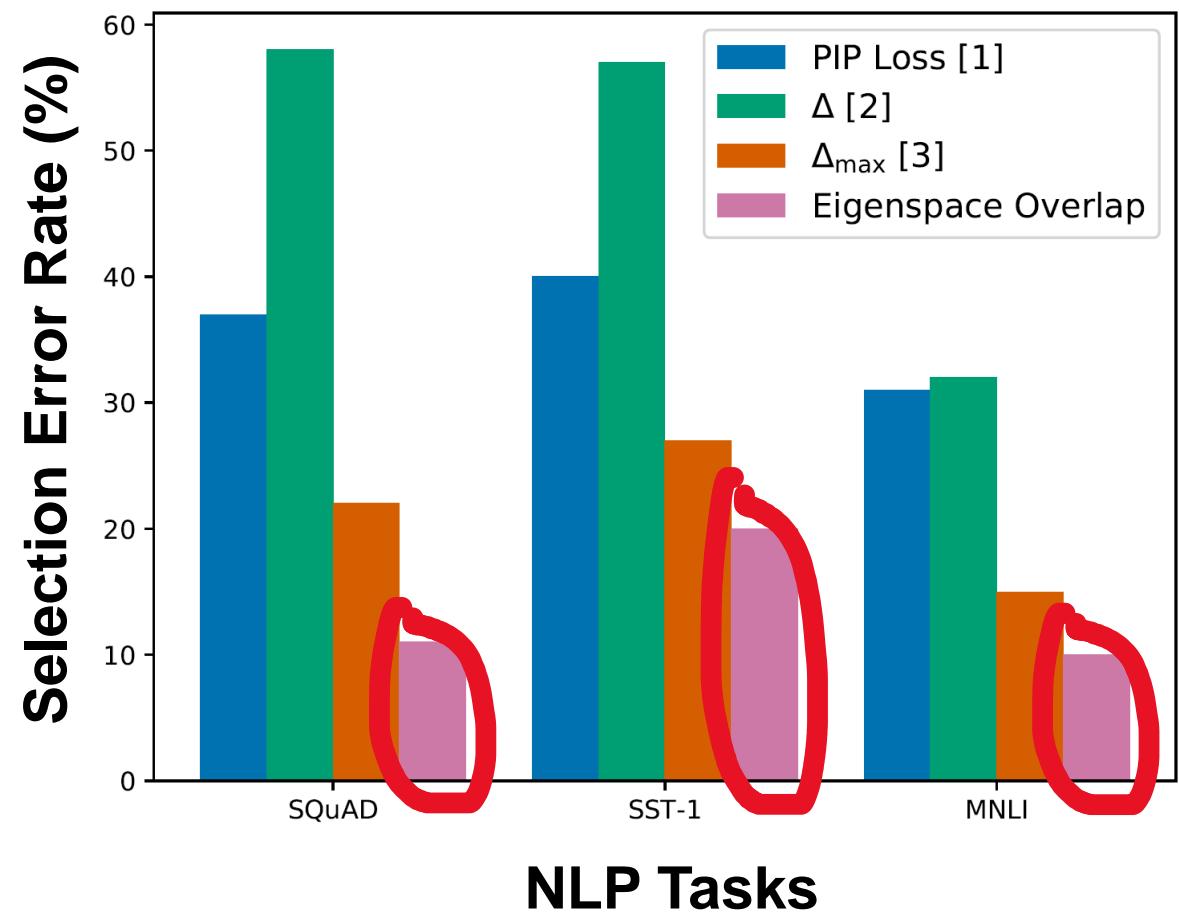
# Empirical Correlation: Beyond Linear Regression

EOS attains **strong correlation** with downstream **model accuracy**.



# Empirical Correlation: Beyond Linear Regression

EOS attains *up to 2x lower selection* error rates than 2<sup>nd</sup> best.



[1] Avron et al., ICML 2017. [2] Yin and Shen. NeurIPS 2018. [3] Zhang et al., AISTATS 2019.

# Our Contributions: Summary

- ① Defined a **new measure** of compression quality.
- ② Proved **generalization bounds** using this measure.
- ③ Showed strong **empirical correlation** w. downstream perf.
- ④ Used measure to **select** compressed embeddings.



# THANK YOU!

## Poster #185, 5-7 pm today!

Paper: <https://arxiv.org/pdf/1909.01264.pdf>

E-mail: [avnermay@cs.stanford.edu](mailto:avnermay@cs.stanford.edu)