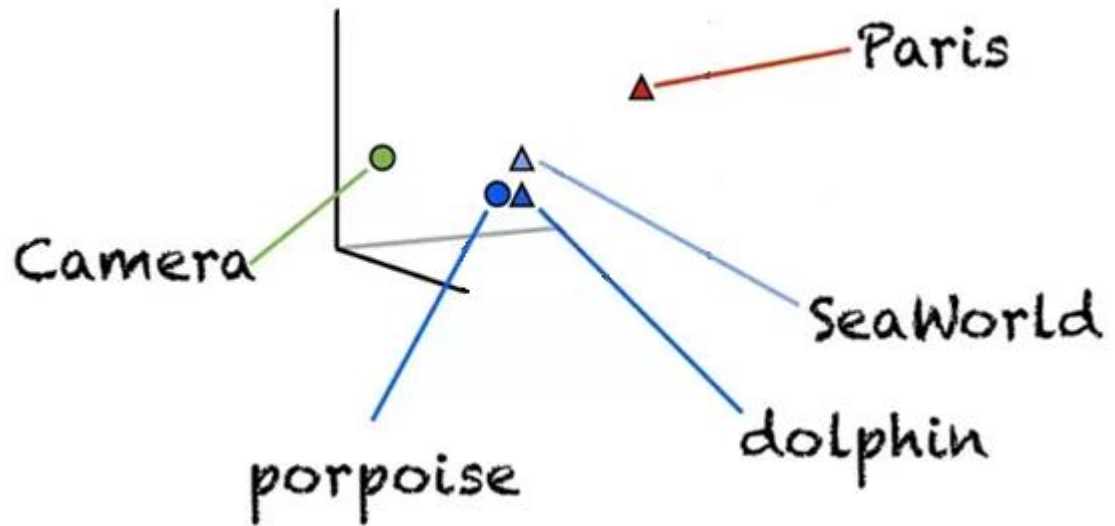


On the Downstream Performance of Compressed Word Embeddings

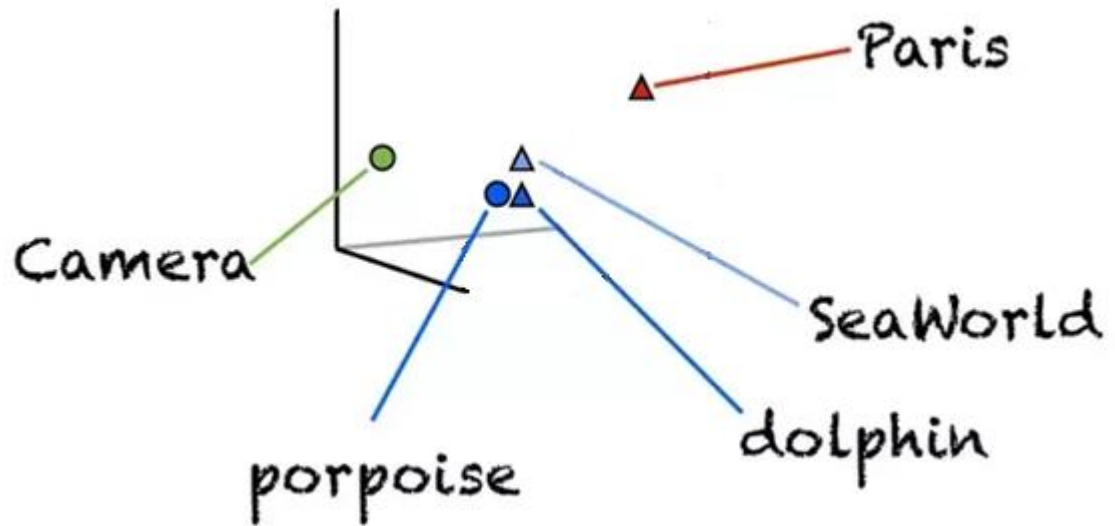
Avner May, Jian Zhang, Tri Dao, Chris Ré
Stanford University



Word Embeddings

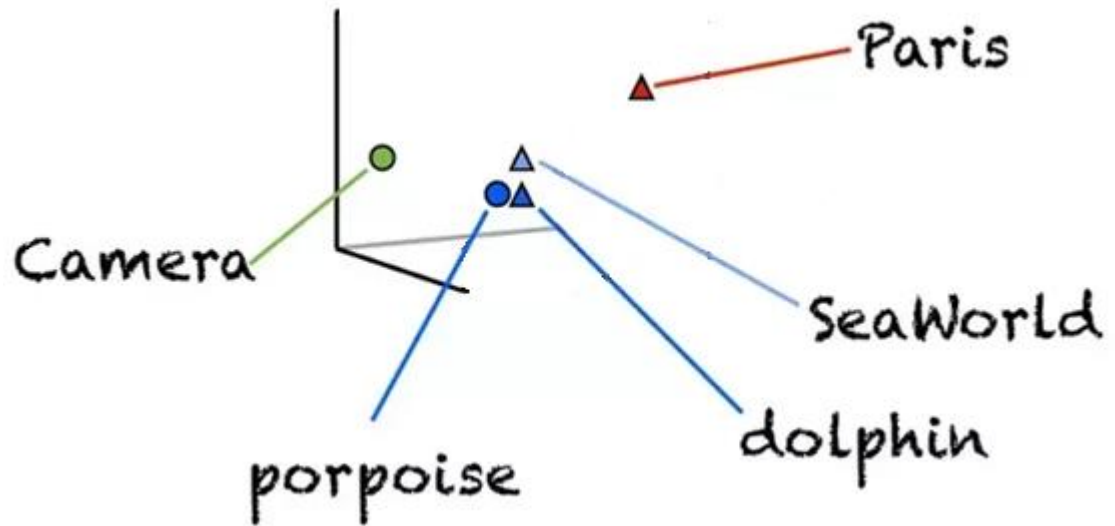


Word Embeddings



Important for strong NLP performance

Word Embeddings

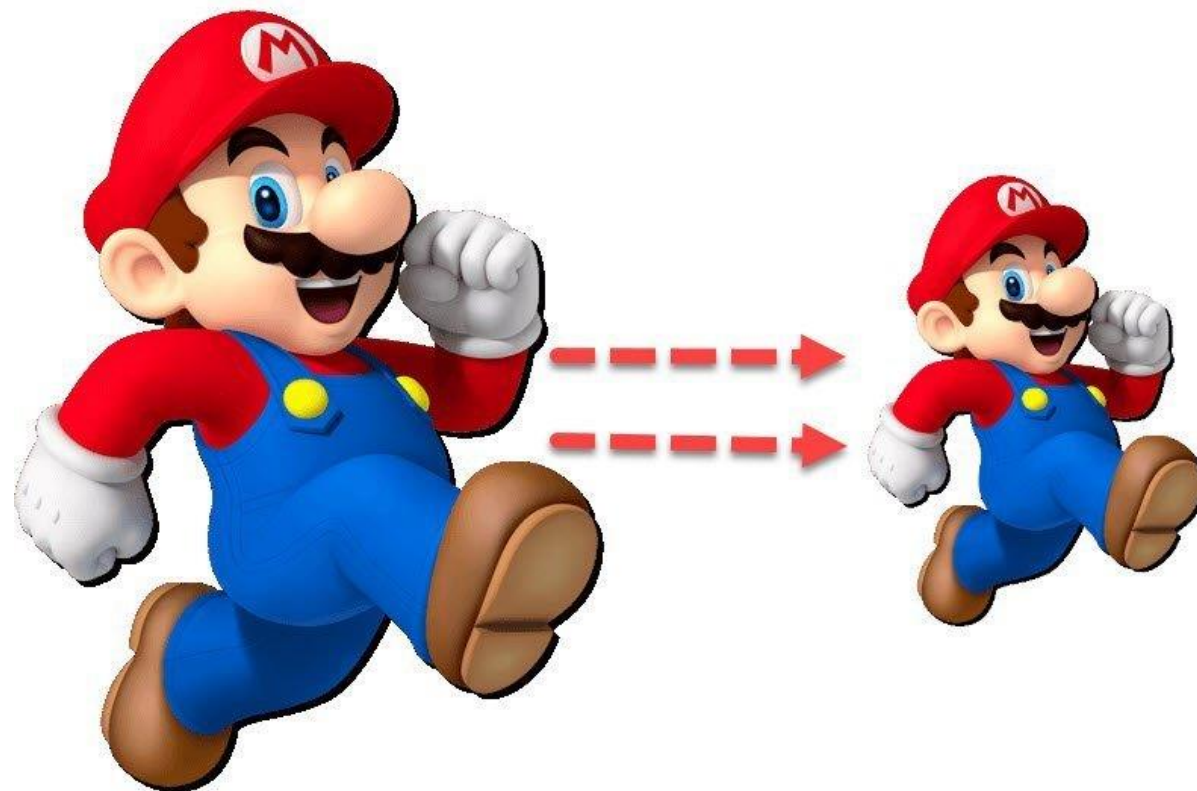


Important for strong NLP performance



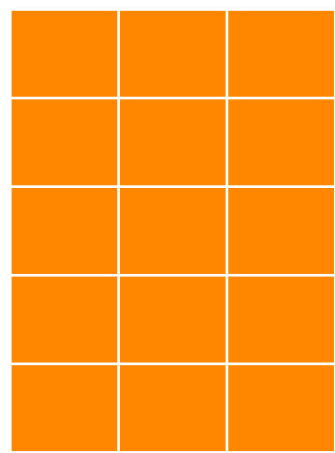
Take a lot of memory

Word Embedding Compression

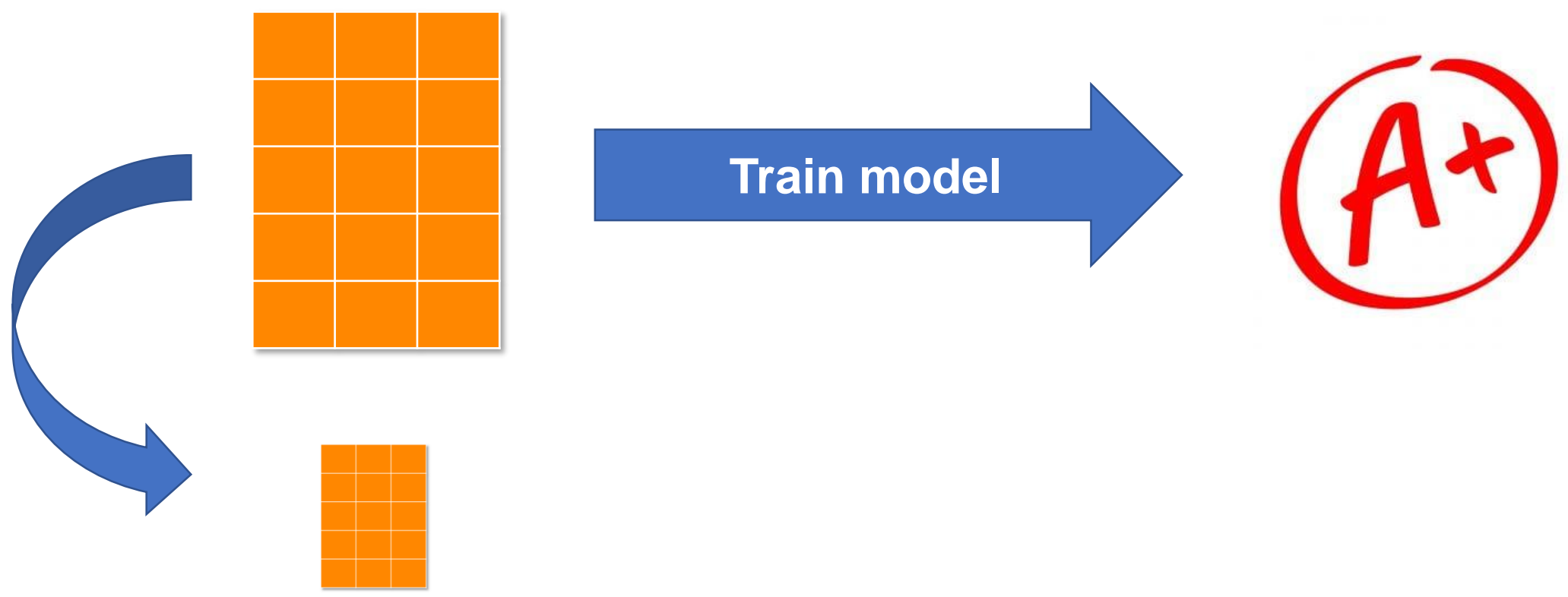


What determines whether a compressed embedding matrix will perform well on downstream tasks?

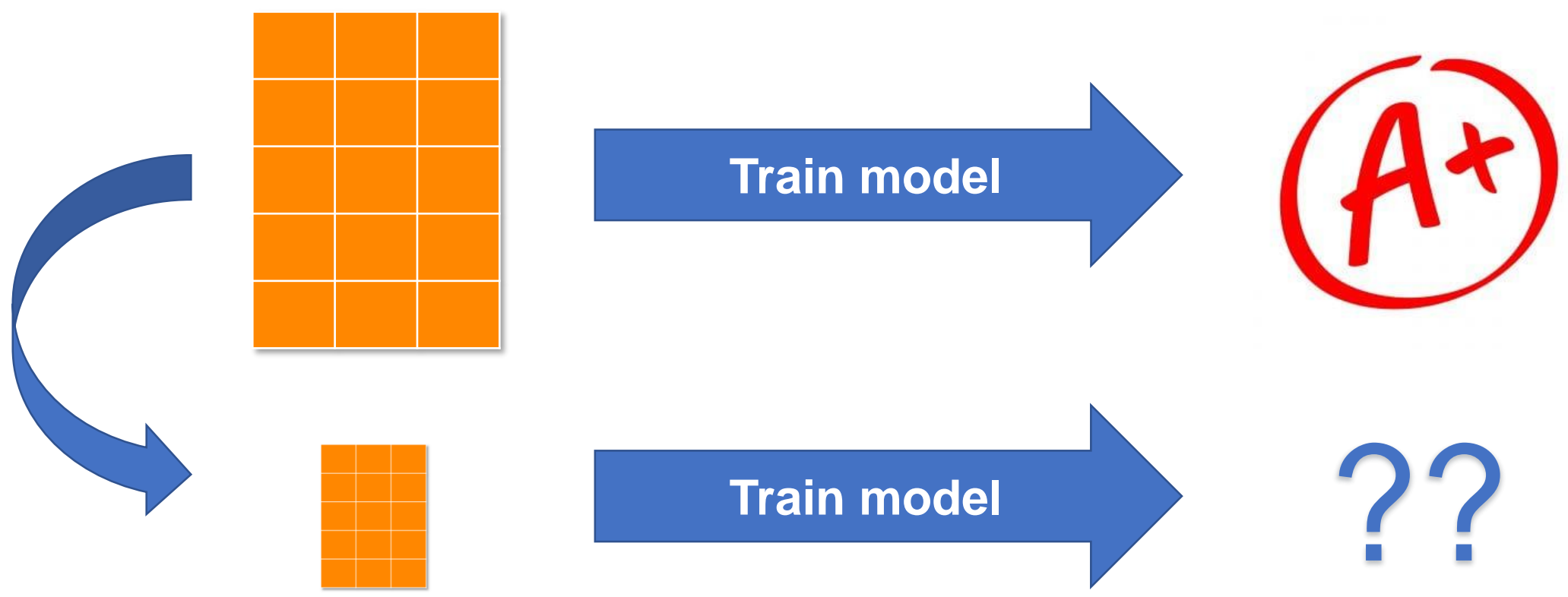
What determines whether a compressed embedding matrix will perform well on downstream tasks?



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



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



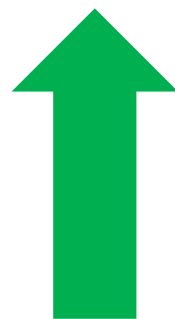
Motivating Observation

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

Motivating Observation

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

Better compression
quality measure



Motivating Observation

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

Better compression
quality measure



Worse downstream
performance

Our Contributions: Outline

Our Contributions: Outline

- ① Define a **new measure** of compression quality.

Our Contributions: Outline

- ① Define a **new measure** of compression quality.
- ② Prove **generalization bounds** using this measure.

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.

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.

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

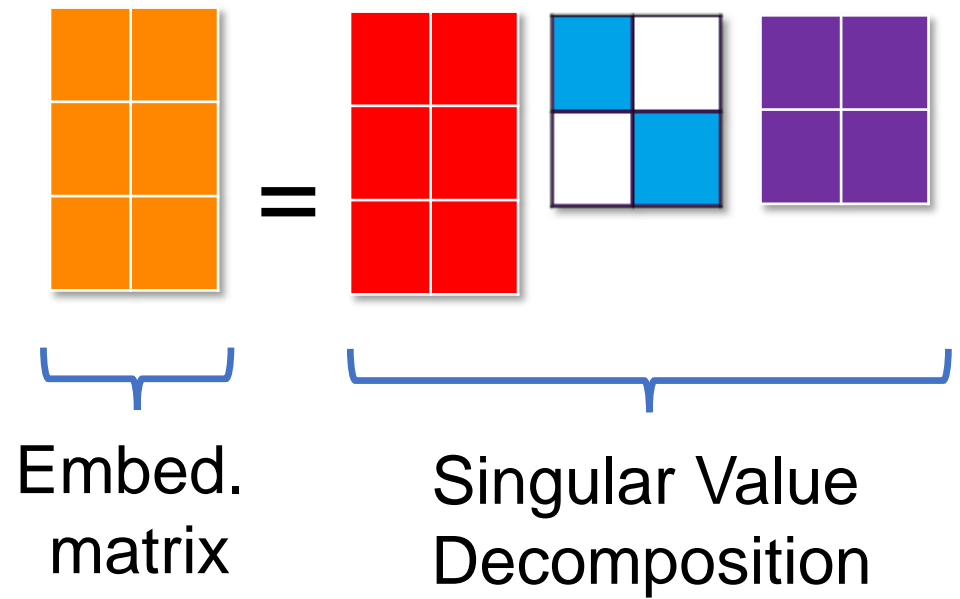
Observation:

Predictions are determined by data matrix's *left singular vectors*.

Defining the Measure: Intuition from Linear Regression

Observation:

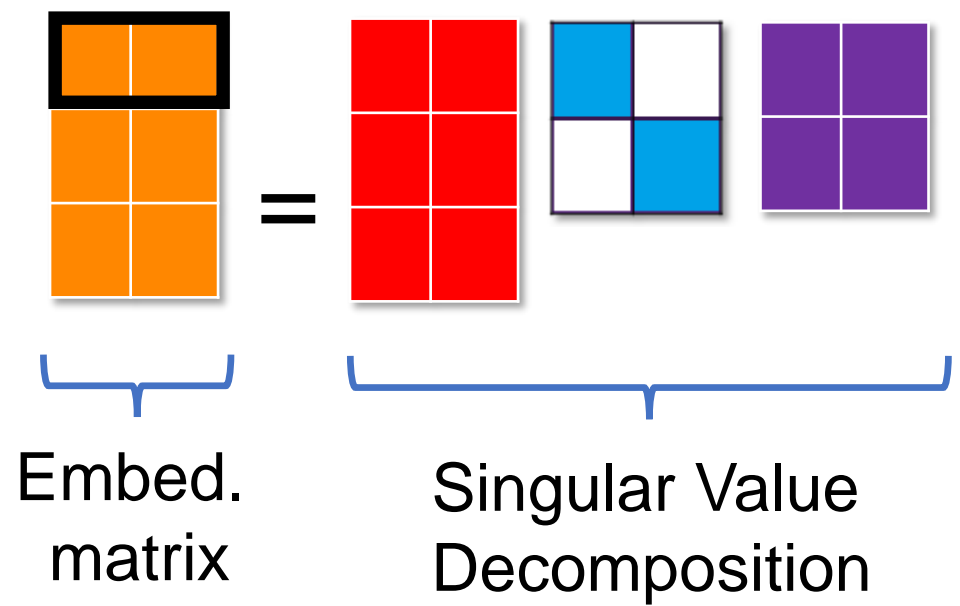
Predictions are determined by data matrix's *left singular vectors*.



Defining the Measure: Intuition from Linear Regression

Observation:

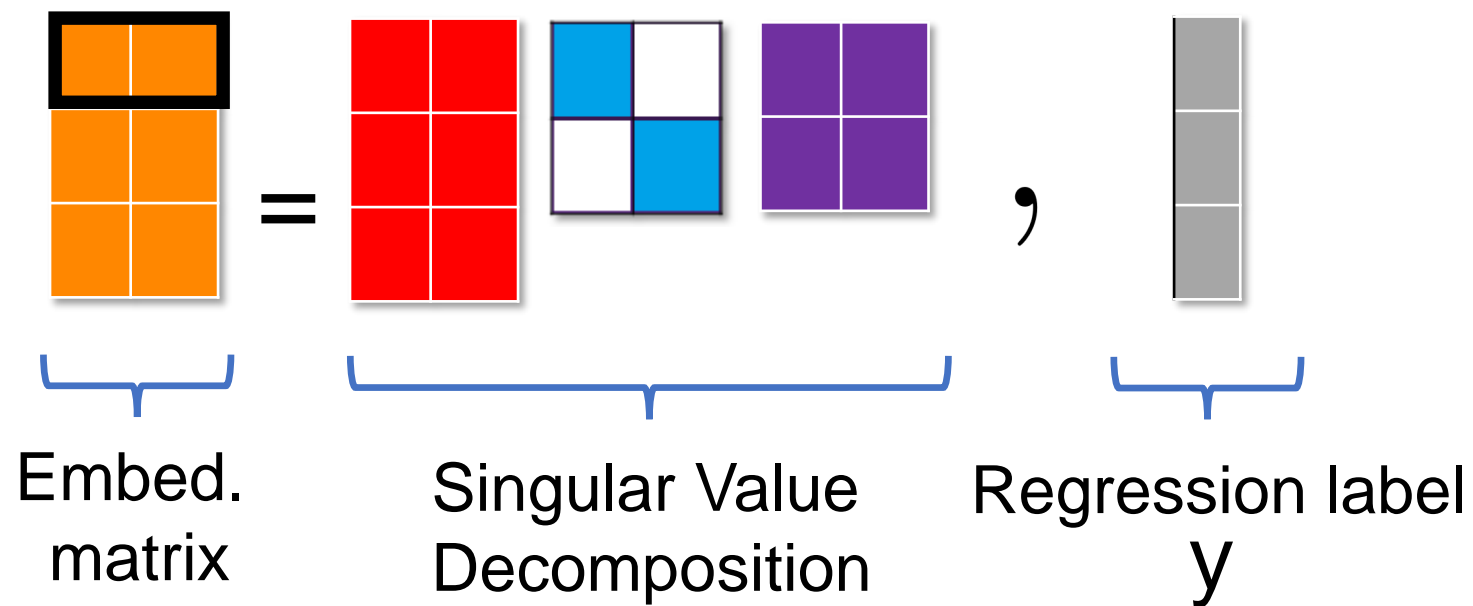
Predictions are determined by data matrix's *left singular vectors*.



Defining the Measure: Intuition from Linear Regression

Observation:

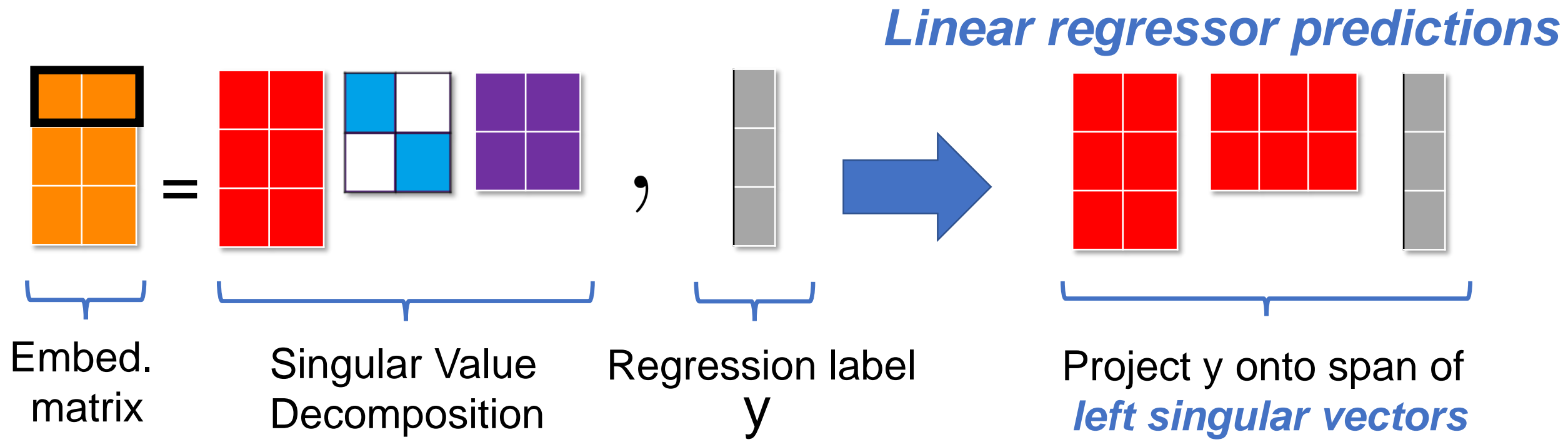
Predictions are determined by data matrix's *left singular vectors*.



Defining the Measure: Intuition from Linear Regression

Observation:

Predictions are determined by data matrix's *left singular vectors*.



Defining the Measure: Eigenspace Overlap Score (EOS)

Intuition:

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

Defining the Measure: Eigenspace Overlap Score (EOS)

Intuition:

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

$$\text{EOS} \left(\underbrace{\begin{matrix} d \\ \begin{matrix} \text{Red} & \text{Red} \\ \text{Red} & \text{Red} \\ \text{Red} & \text{Red} \end{matrix} & \begin{matrix} \text{Blue} & \text{White} \\ \text{White} & \text{Blue} \end{matrix} & \begin{matrix} \text{Purple} & \text{Purple} \\ \text{Purple} & \text{Purple} \end{matrix} \end{matrix}, \underbrace{\begin{matrix} \text{Red} & \text{Red} \\ \text{Red} & \text{Red} \\ \text{Red} & \text{Red} \end{matrix} & \begin{matrix} \text{Light Blue} & \text{White} \\ \text{White} & \text{Light Blue} \end{matrix} & \begin{matrix} \text{Purple} & \text{Purple} \\ \text{Purple} & \text{Purple} \end{matrix} \right) = \frac{1}{d} \left\| \begin{matrix} \text{Red} & \text{Red} & \text{Red} \\ \text{Red} & \text{Red} & \text{Red} \end{matrix} \right\|_F^2$$

Uncompressed embed. SVD **Compressed embed. SVD** **Eigenspace overlap score**

Theoretical Results: Linear Regression

Theorem (informal):

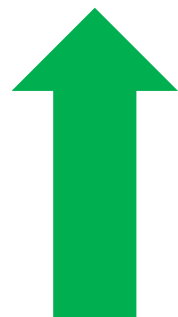
Expected difference in *test mean-squared error* attained by *compressed* vs. *uncompressed* embeddings is *determined by EOS*.

Theoretical Results: Linear Regression

Theorem (informal):

Expected difference in *test mean-squared error* attained by *compressed* vs. *uncompressed* embeddings is *determined by EOS*.

Higher EOS

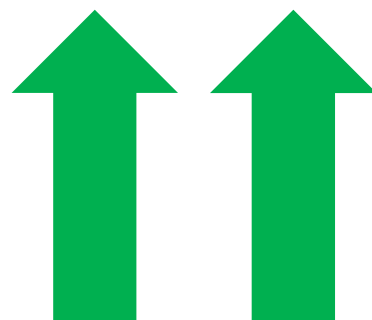


Theoretical Results: Linear Regression

Theorem (informal):

Expected difference in *test mean-squared error* attained by *compressed* vs. *uncompressed* embeddings is *determined by EOS*.

Higher EOS



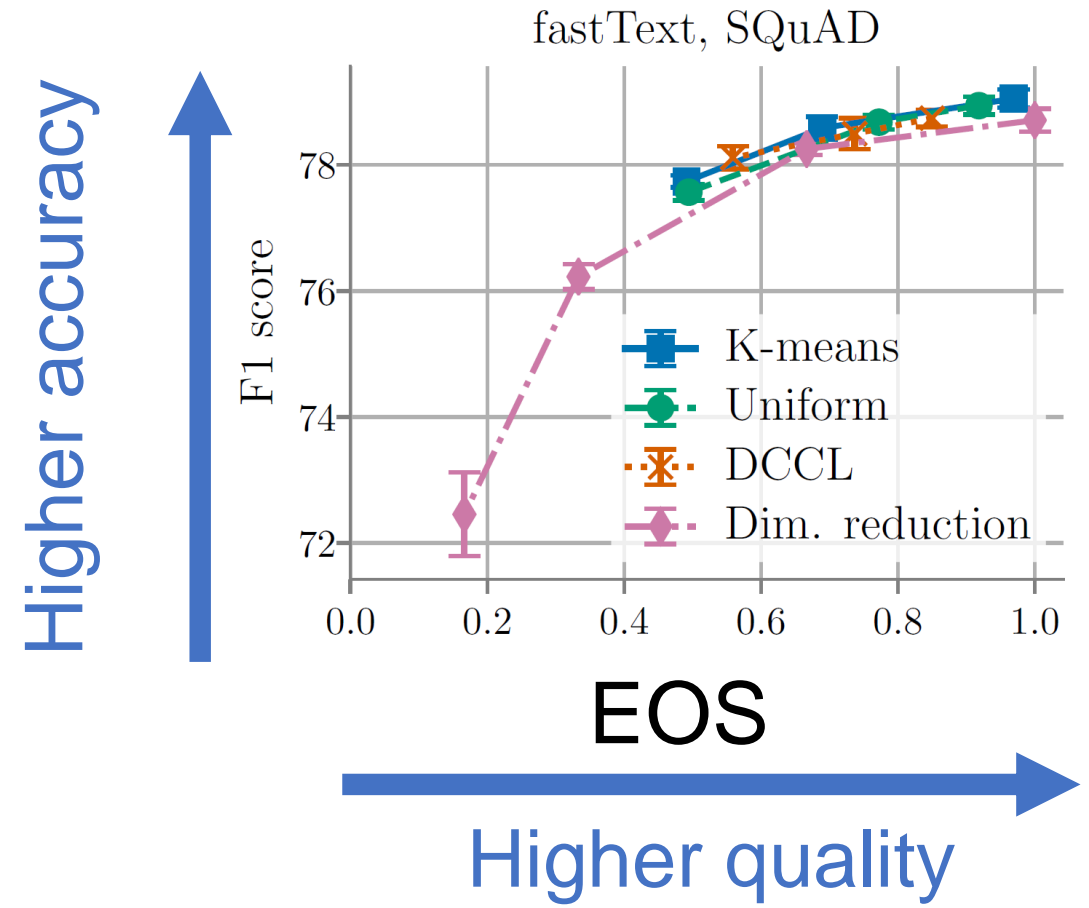
Better downstream performance

Empirical Correlation: Beyond Linear Regression

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

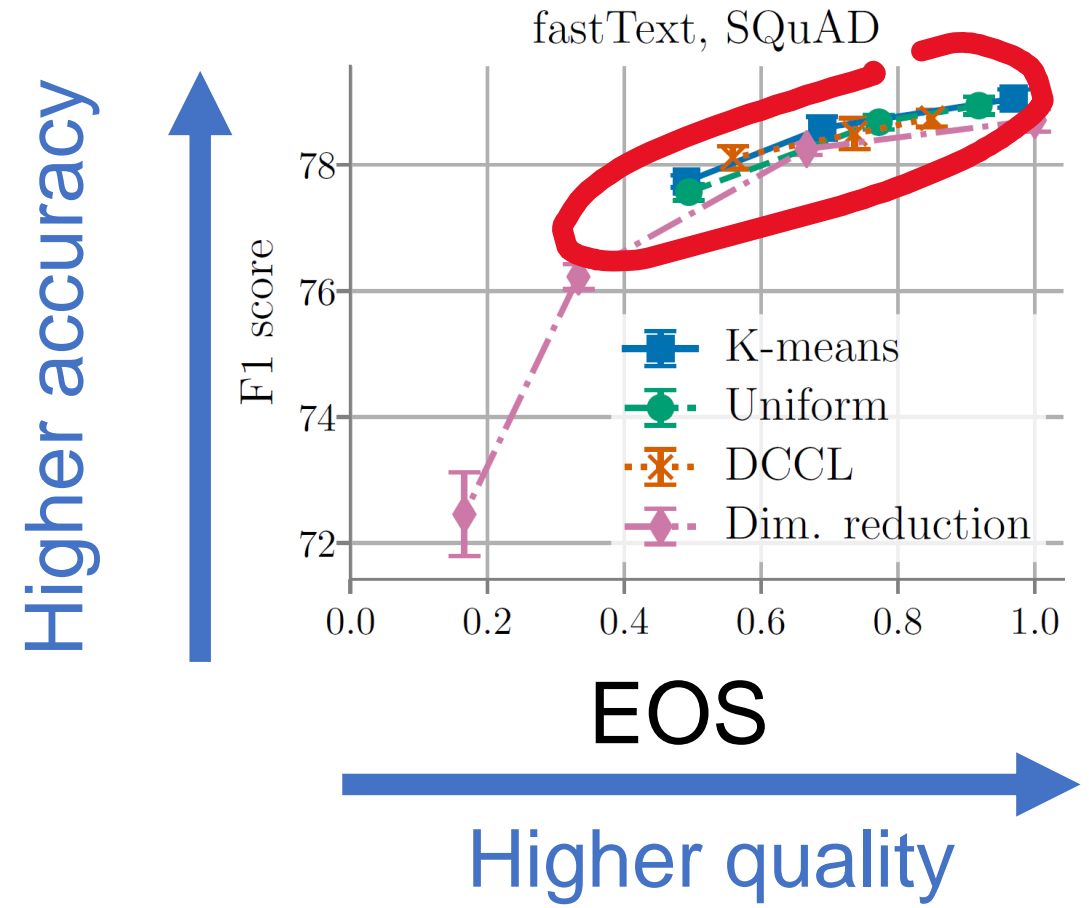
Empirical Correlation: Beyond Linear Regression

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



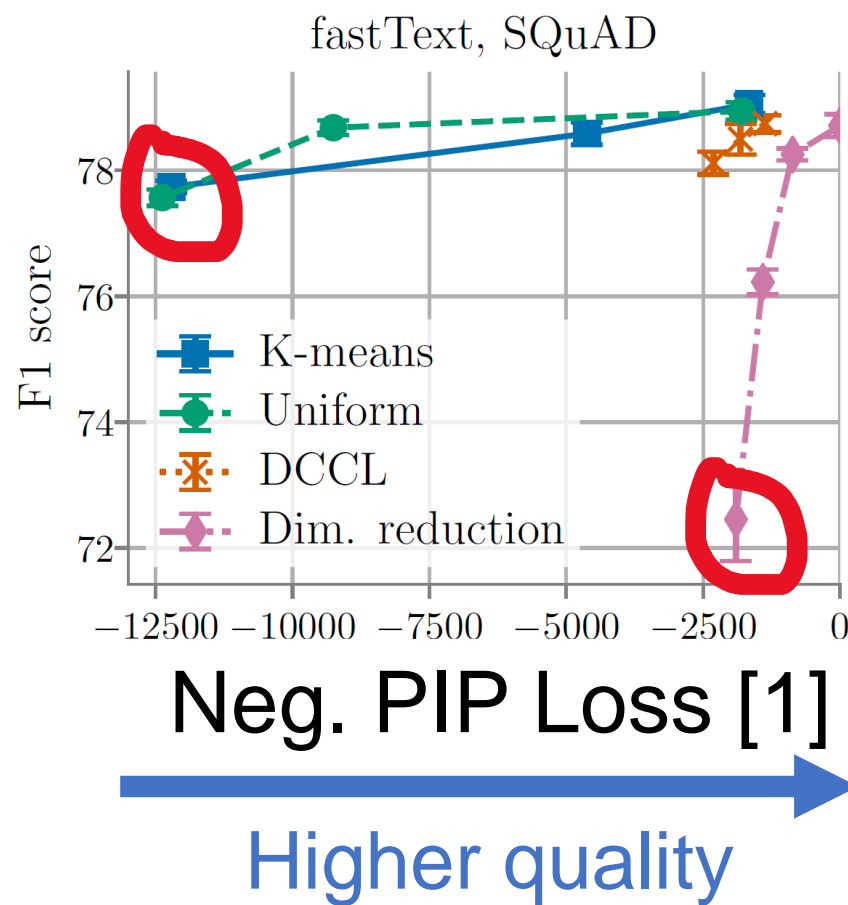
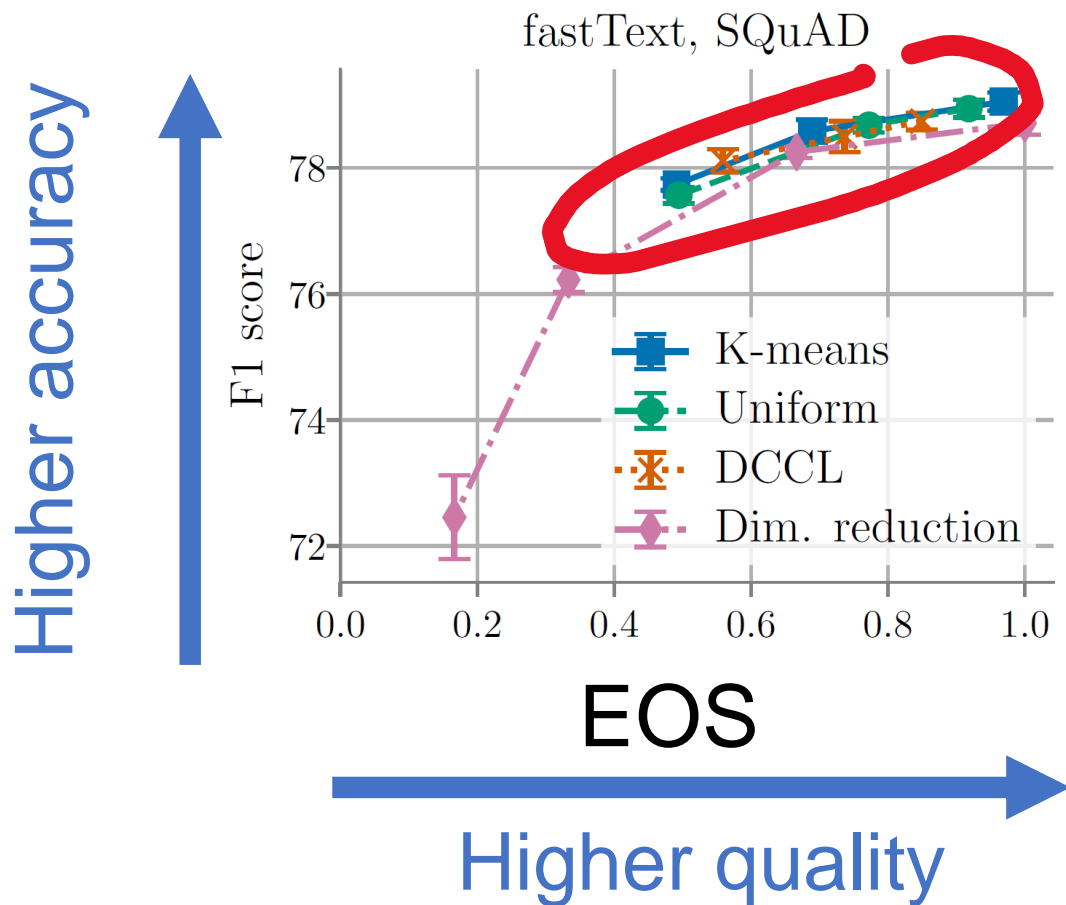
Empirical Correlation: Beyond Linear Regression

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



Empirical Correlation: Beyond Linear Regression

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



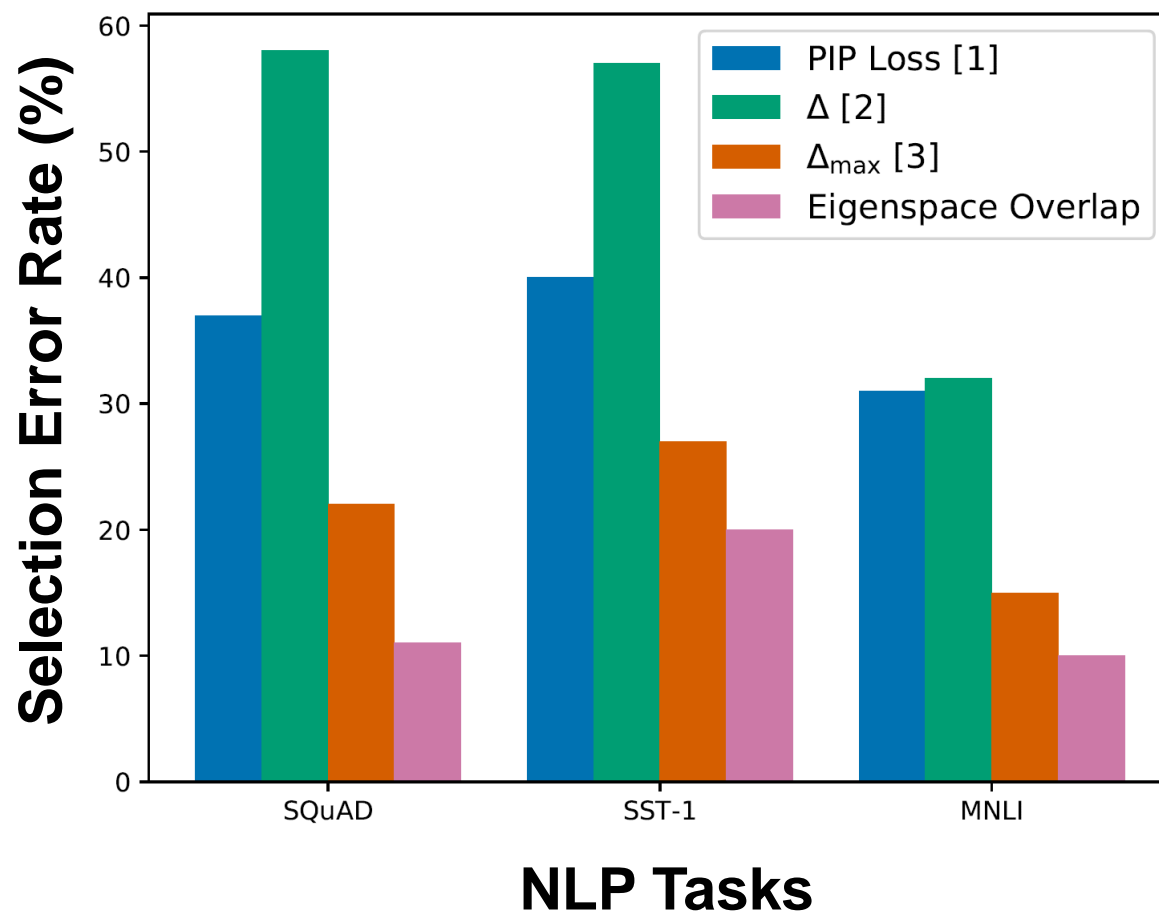
[1] Yin and Shen, *On the Dimensionality of Word Embeddings*. NeurIPS 2018.

EOS as a Selection Criterion

EOS attains *up to 2x lower selection* error rates than 2nd best.

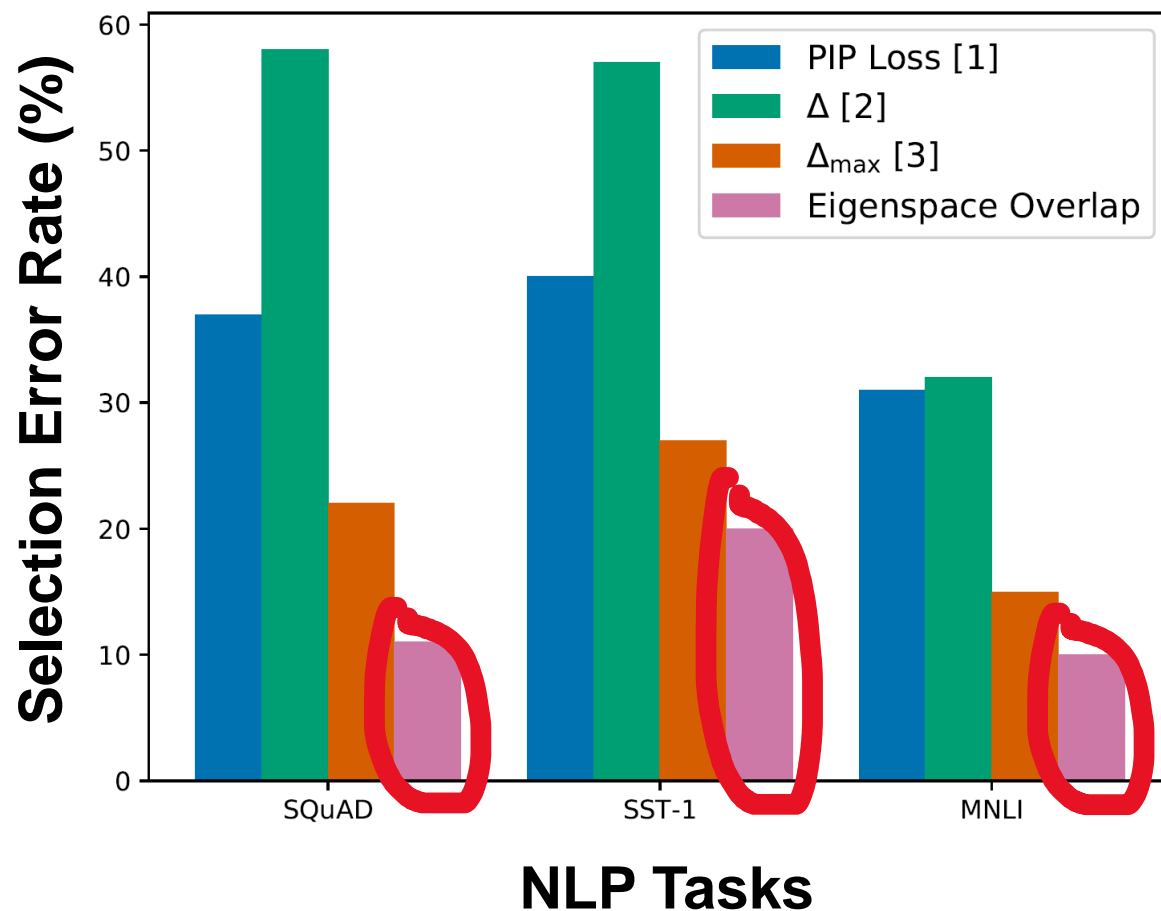
EOS as a Selection Criterion

EOS attains *up to 2x lower selection* error rates than 2nd best.



EOS as a Selection Criterion

EOS attains *up to 2x lower selection* error rates than 2nd best.



Our Contributions: Summary

Our Contributions: Summary

- ① Defined a **new measure** of compression quality.

Our Contributions: Summary

- ① Defined a **new measure** of compression quality.
- ② Proved **generalization bounds** using this measure.

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.

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>

Code: <https://github.com/HazyResearch/smallfry>

E-mail: avnermay@cs.stanford.edu