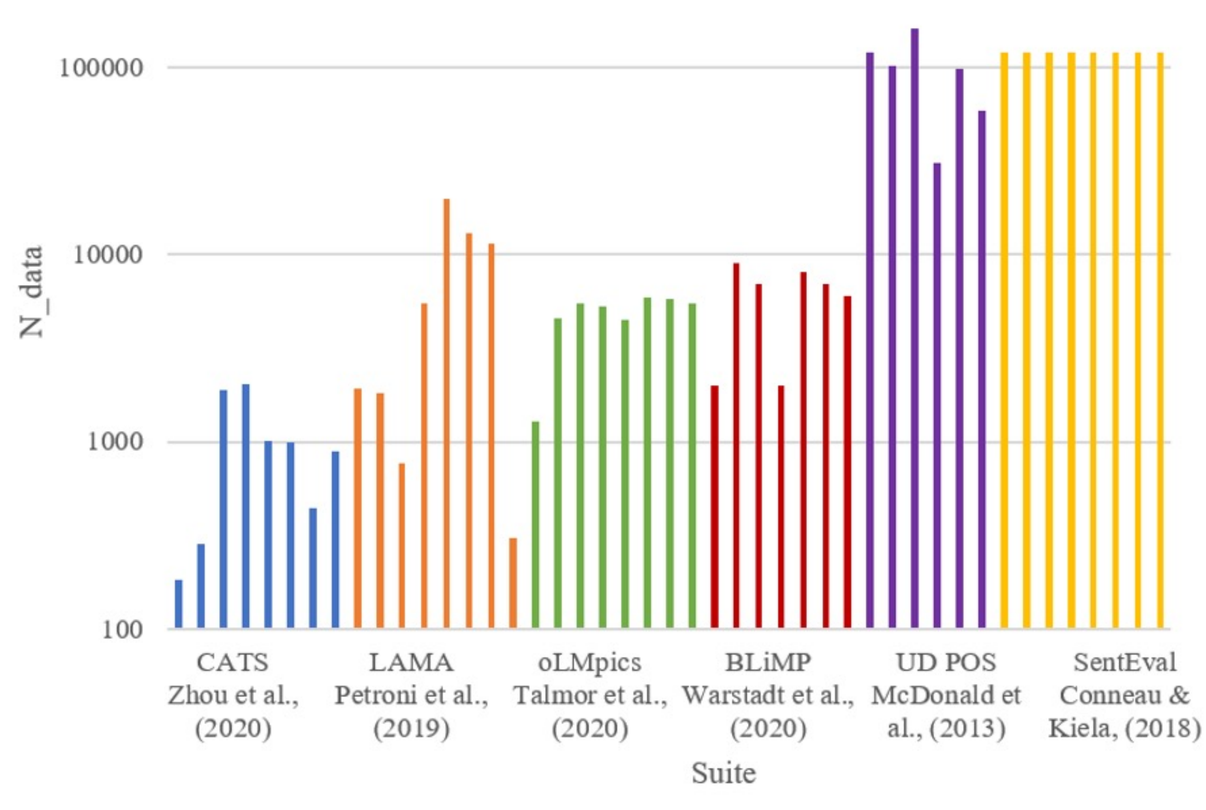
On the Data Requirements of Probing

Zining Zhu, Jixuan Wang, Bai Li, Frank Rudzicz

Probing dataset sizes:
Small: ~100 samples
Large: 100,000+ samples

How large should they be, so we can **reliably replicate** probing findings?

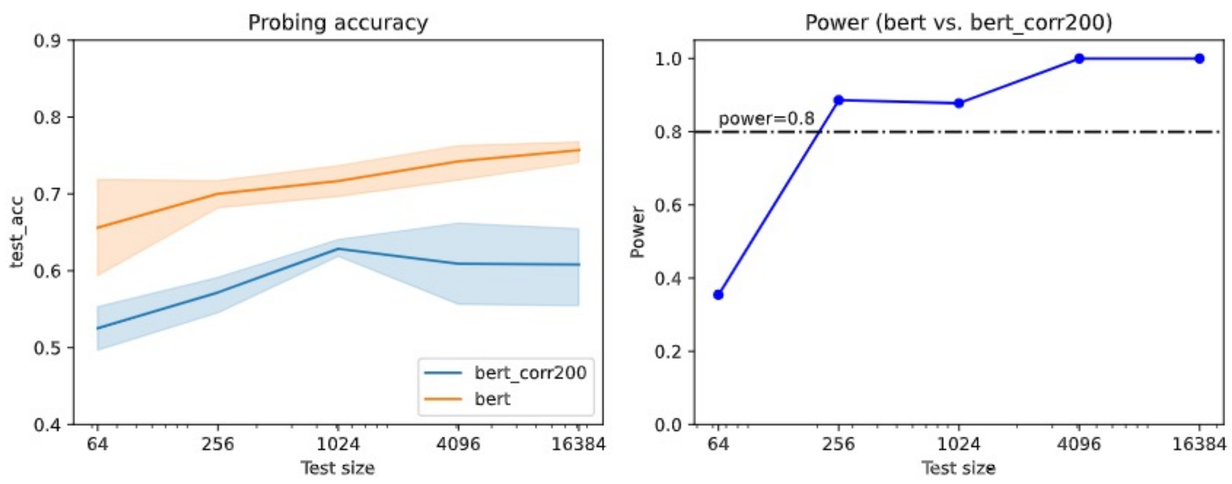


Propose a method that recommends the size of probing dataset for comparing C_A and C_B:

$$\mathbb{P}\left(|R(\hat{f}) - R(f_*)| > B\sqrt{\frac{2\log\frac{2|\mathcal{F}|}{\delta}}{n}}\right) < \delta$$

By setting $\left|R(\hat{f}) - R(f_*)\right| = \frac{|R_A - R_B|}{2}$, we can solve for the recommended train data size n, With more than n data samples, the comparison results between C_A and C_B won't be changed by excess risks (w/ prob of at least $1 - \delta$)

Config C_A	Config C_B
Task: POS probing	Task: POS probing
Encoder A: BERT	Encoder B: InferSent
Probe A: LogReg	Probe B: LogReg



Example case study:
While bert outperforms bert_corr200, there is not enough power until $N_{\text{test}} = 256$ ($N_{\text{train}} = 1{,}024$). Our recommended N_{train} have enough statistical powers.

Subsampled N_{test}	Mean $\lvert R_1 - R_2 \rvert$	Recommended N_{train}
64	.1313	22,263
256	.1281	23,362
1,024	.0879	49,647
4,096	.1331	21,662
16,384	.1488	17,331