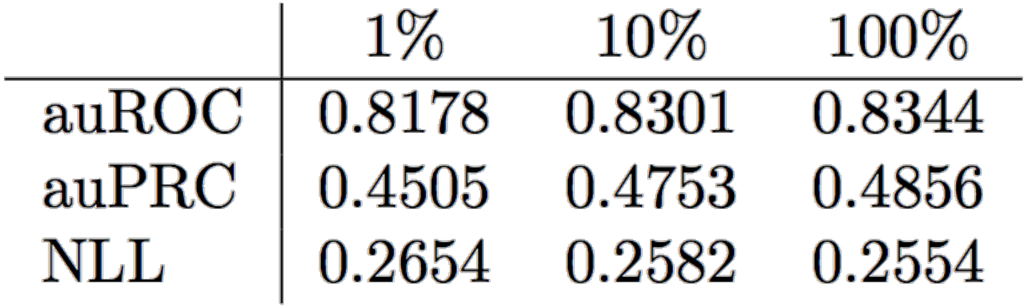
Why the size matters

Why do we need such huge datasets?

- It turns out you can learn better models using the same simple linear classifier

Ad click prediction

- https://arxiv.org/pdf/1110.4198.pdf
- Trillions of features, billions of training examples
- Data sampling hurts the model



	1%	10%	100%
auROC	0.8178	0.8301	0.8344
auPRC	0.4505	0.4753	0.4856
NLL	0.2654	0.2582	0.2554

Sampling rate