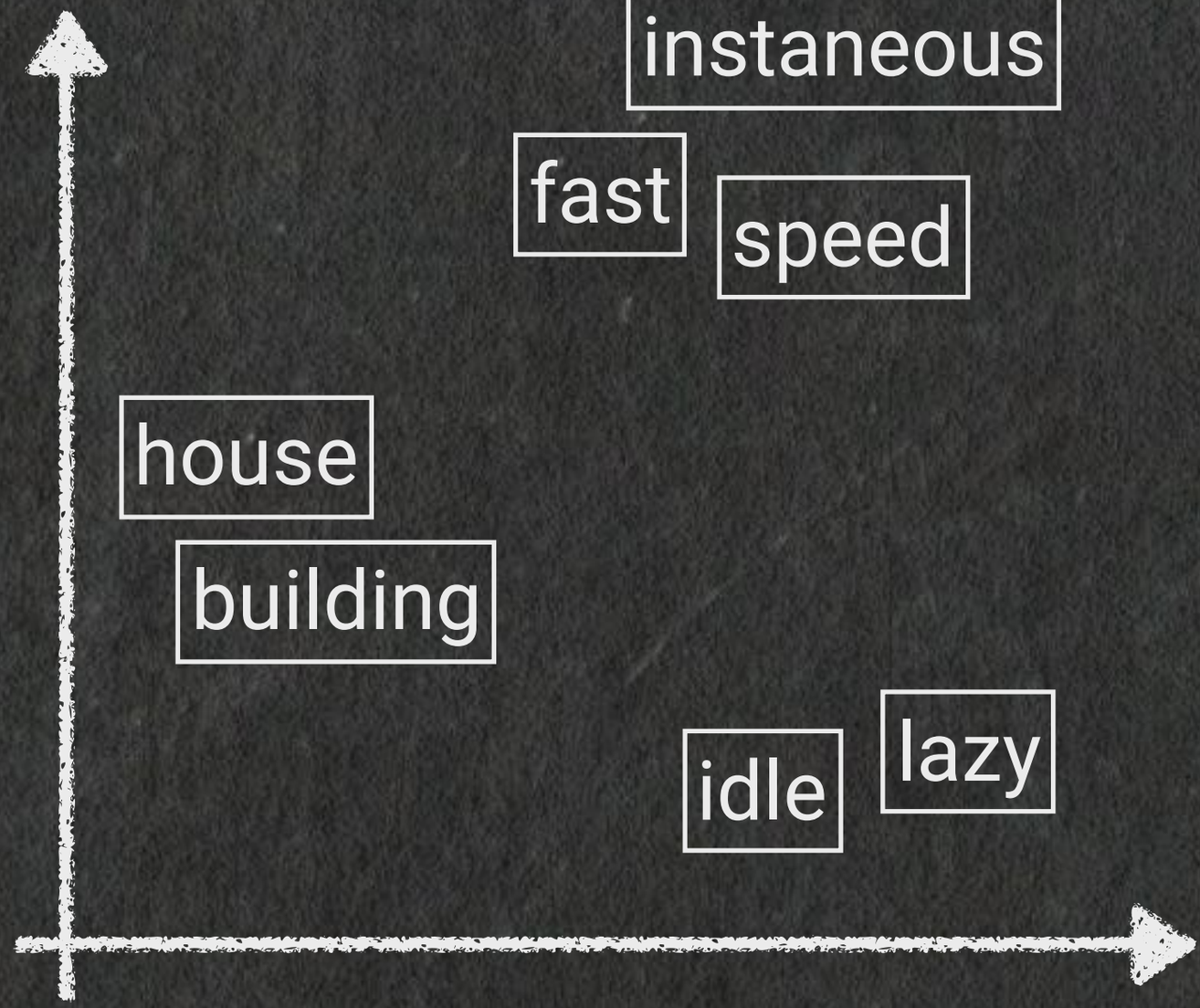
Feeding Word2vec with tens of billions of items, what could possibly go wrong?

Simon Dollé, 12/06/2017

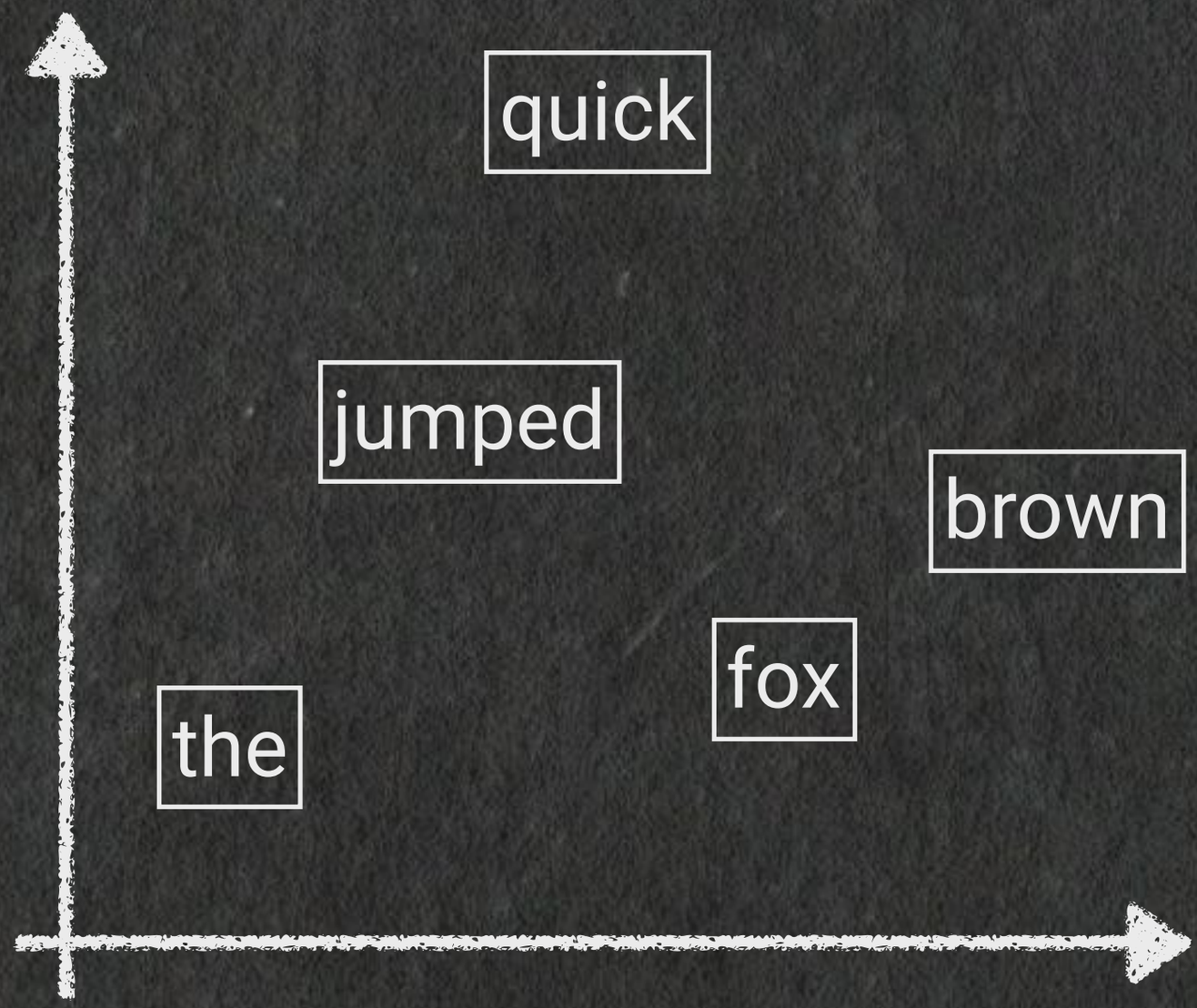
BERLIN BUZZWORDS 2017 JUNE 11-13

criteo


instaneous
fast
speed
house
building
idle
lazy

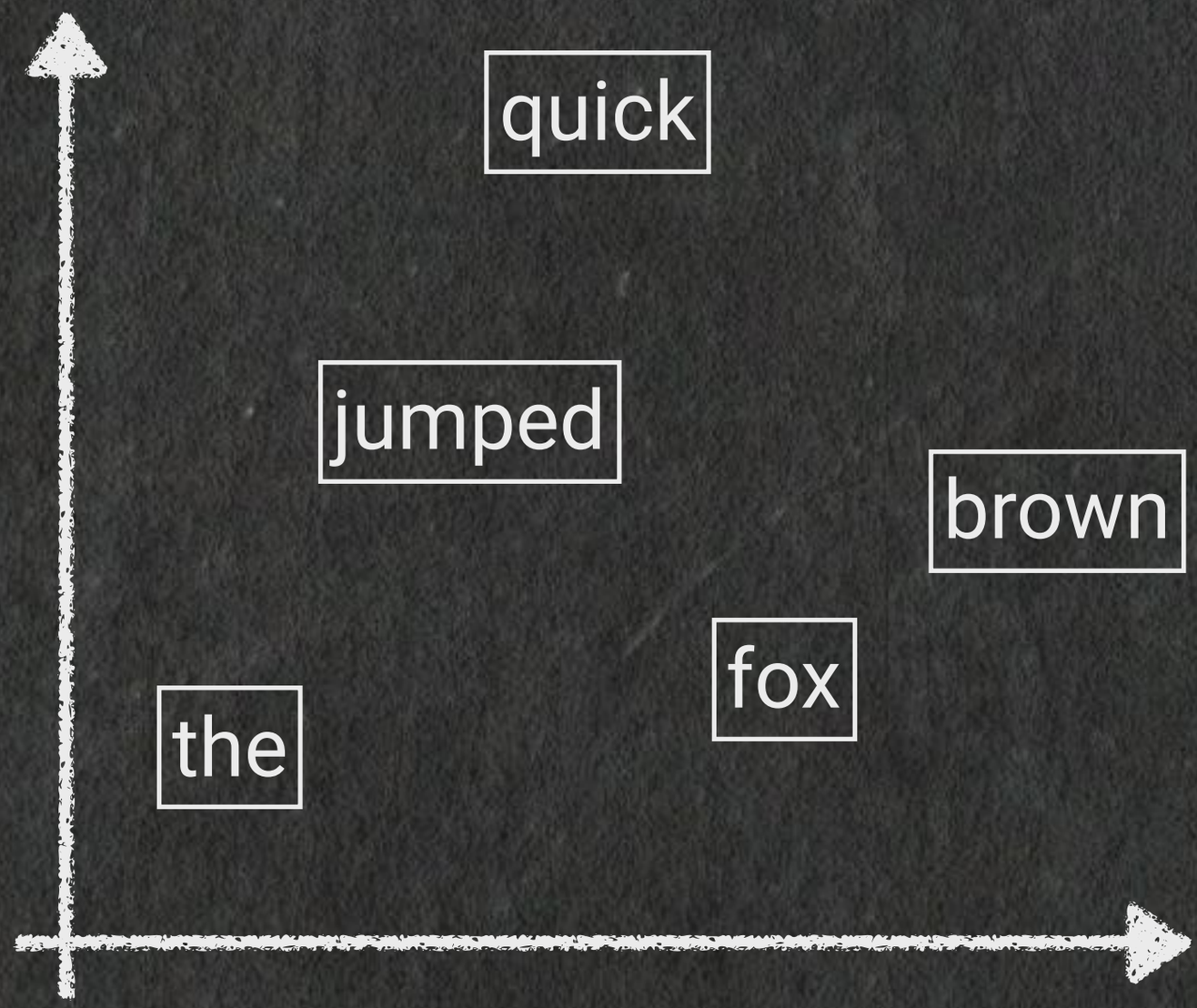
the quick brown fox jumped over the lazy dog

the quick brown fox jumped over the lazy dog



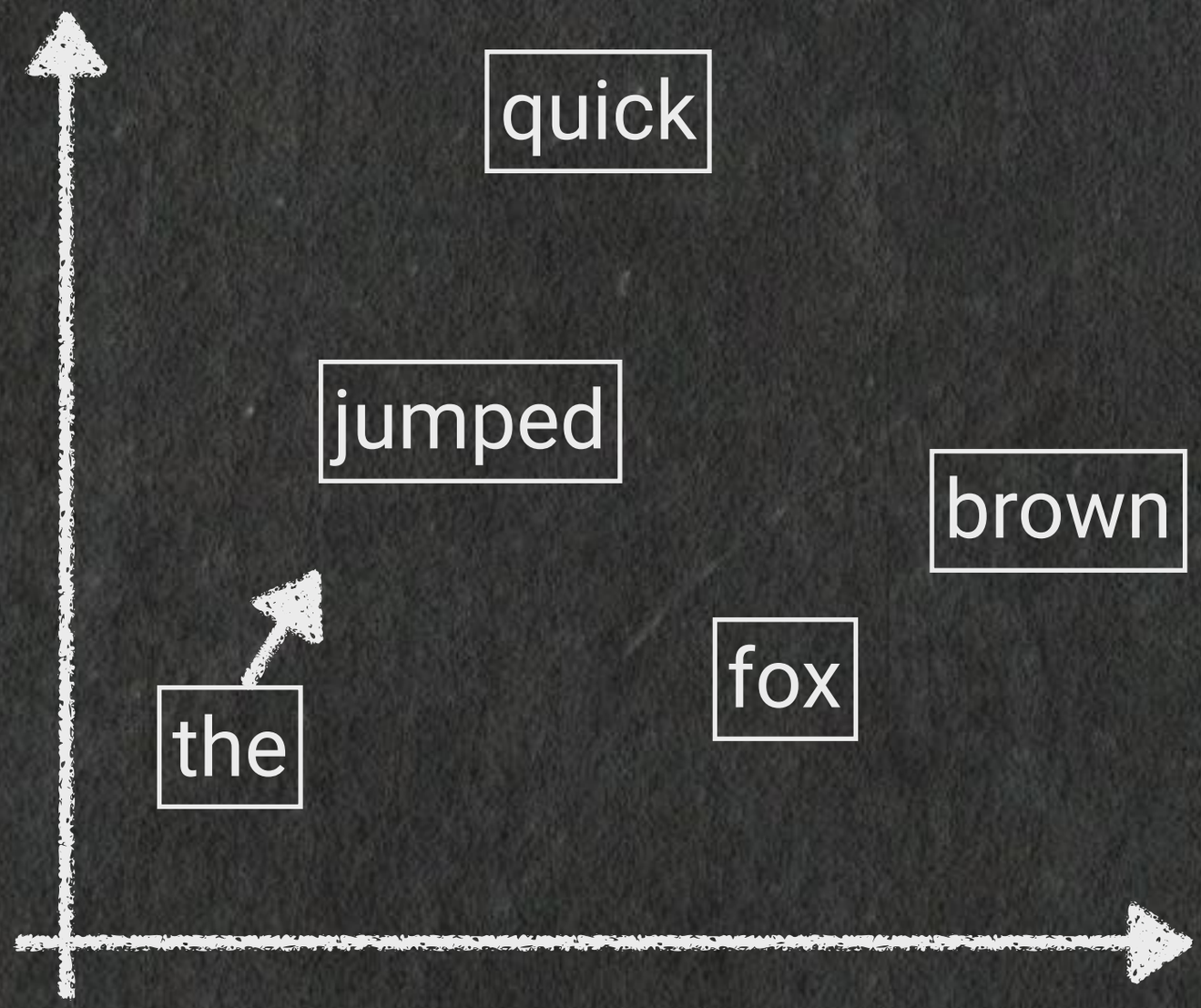
~~the~~ quick ~~brown~~ fox jumped over the lazy dog

? ?



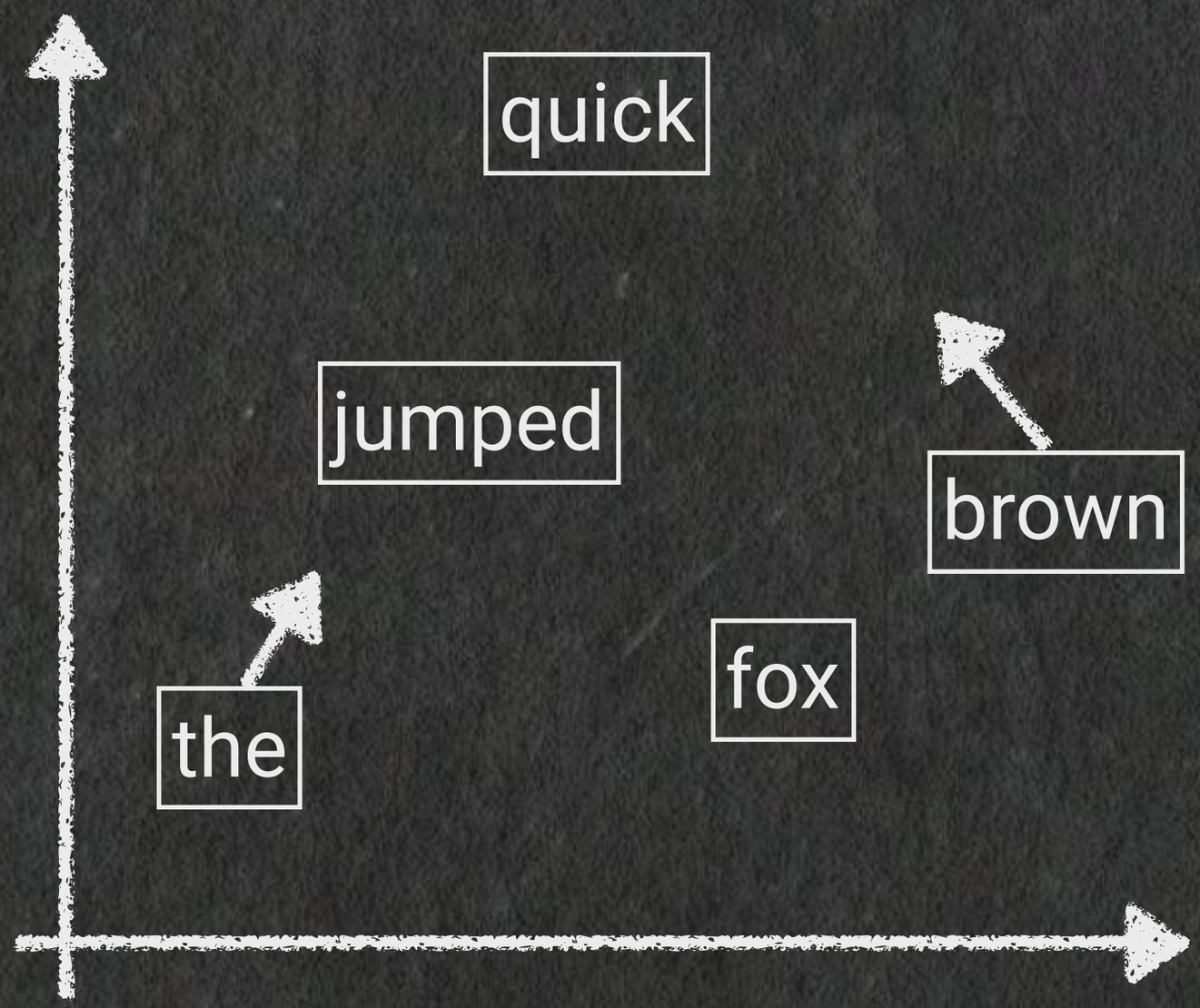
the quick brown fox jumped over the lazy dog

? ?



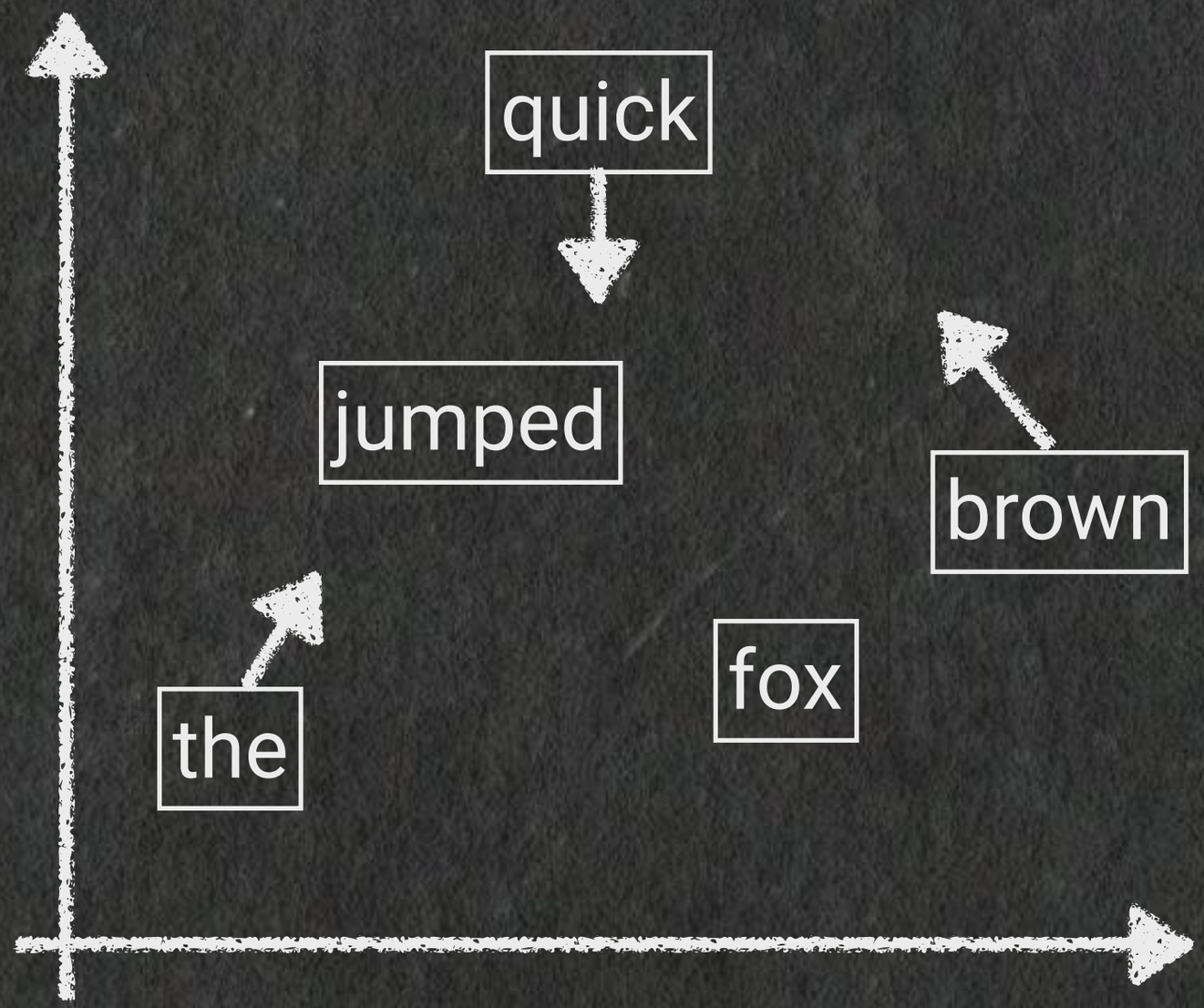
~~the~~ quick ~~brown~~ fox jumped over the lazy dog

? ?



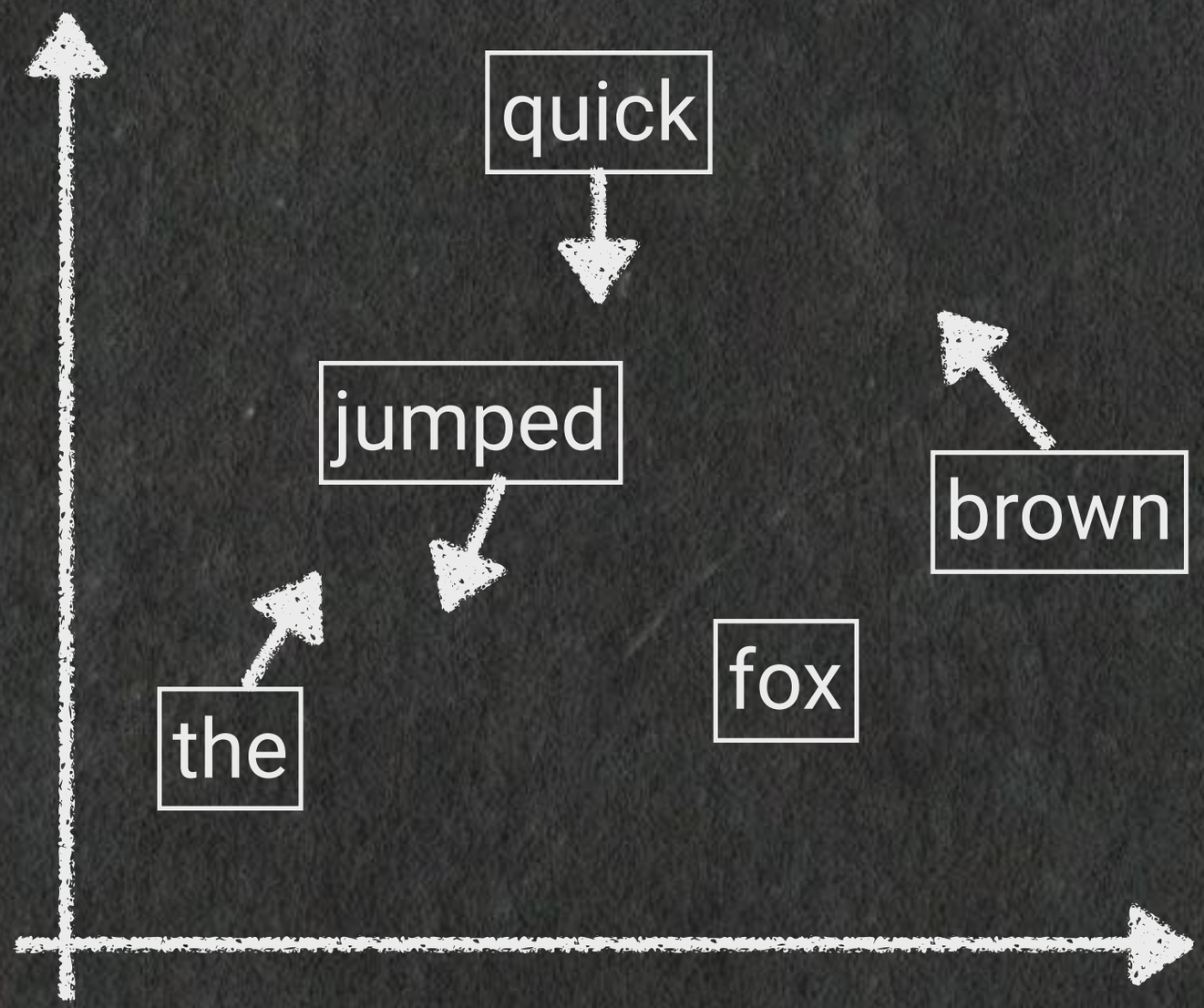
~~the~~ quick ~~brown~~ fox jumped over the lazy dog

? ?

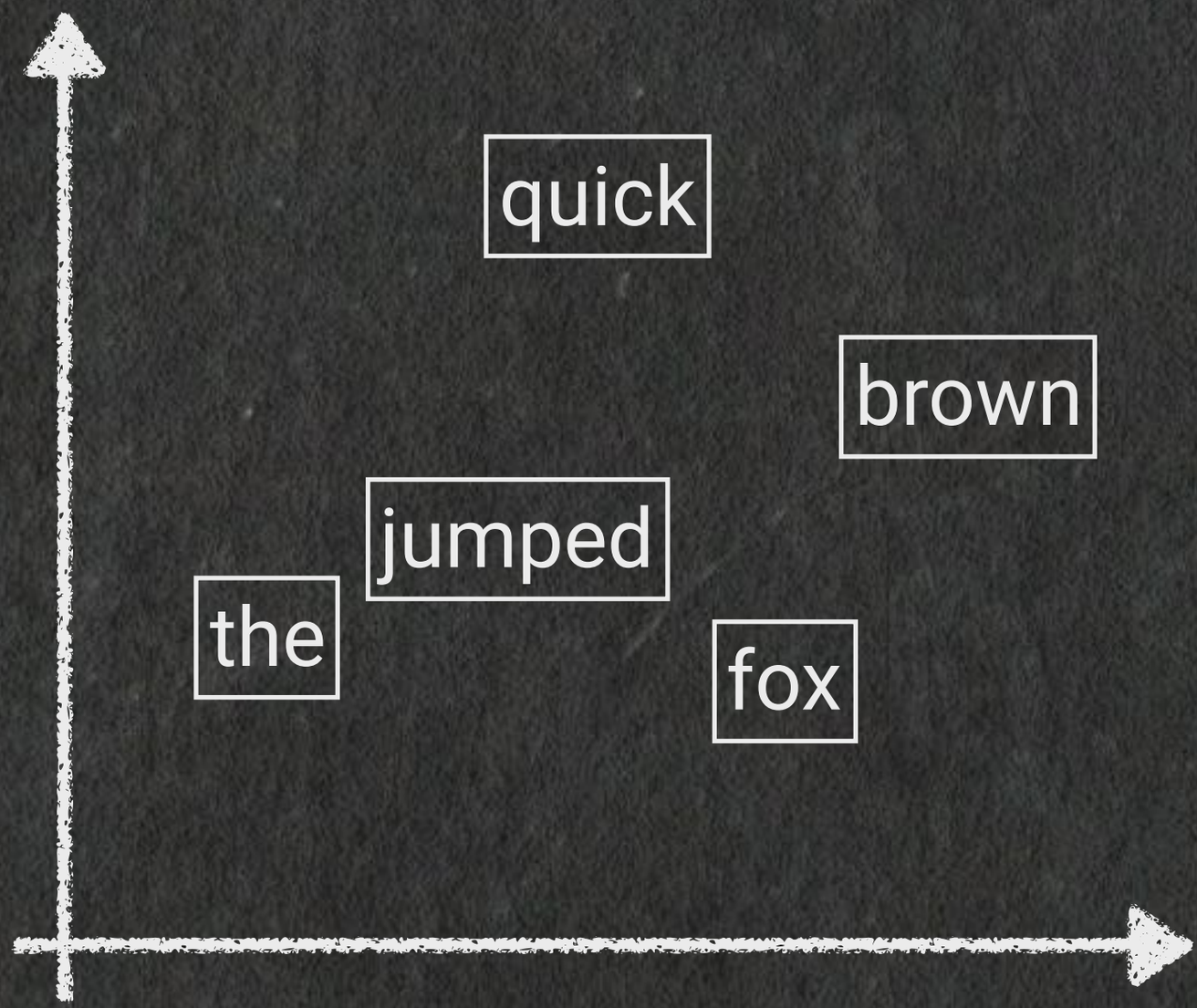


~~the~~ quick ~~brown~~ fox jumped over the lazy dog

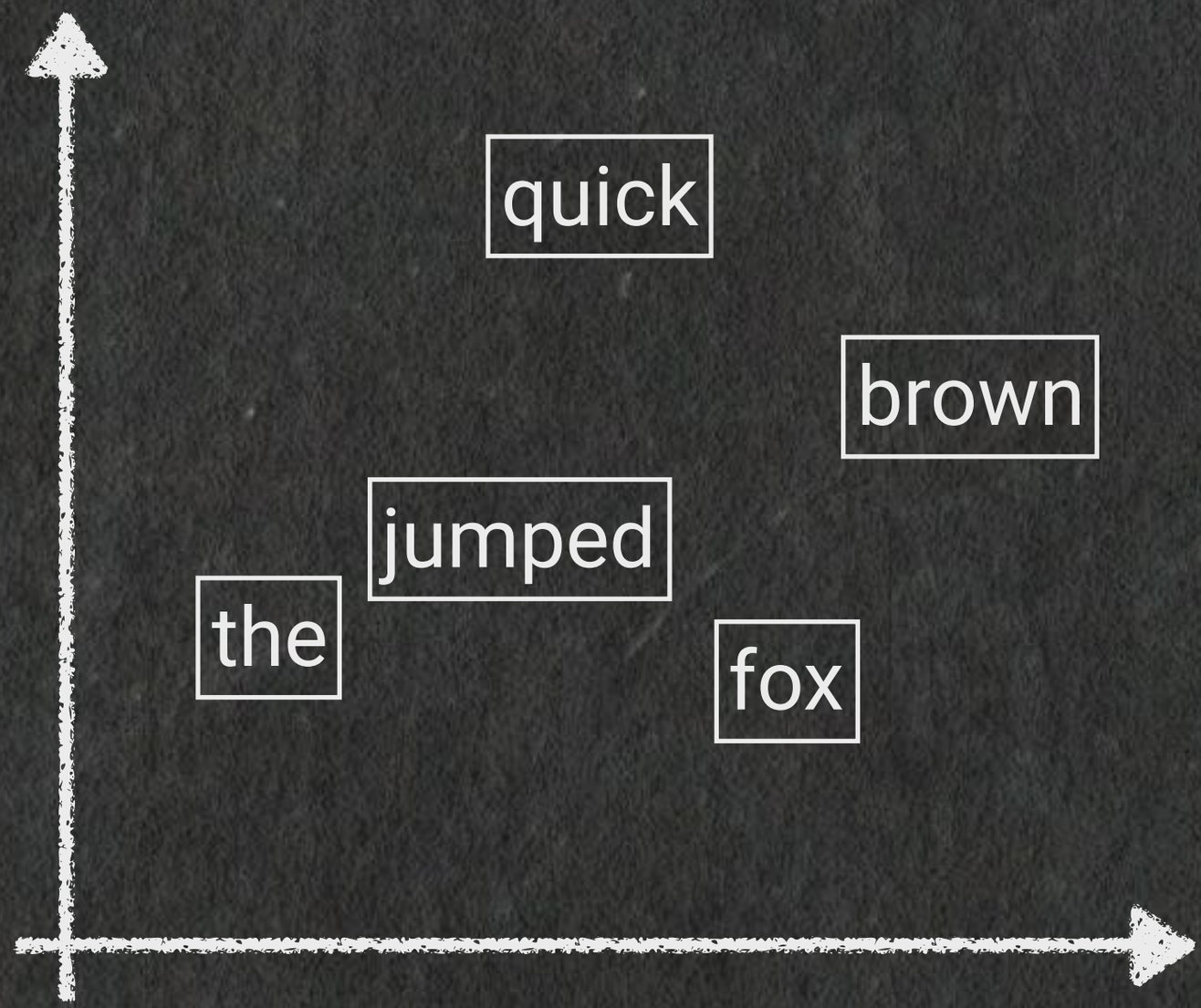
? ?



the quick brown fox jumped over the lazy dog

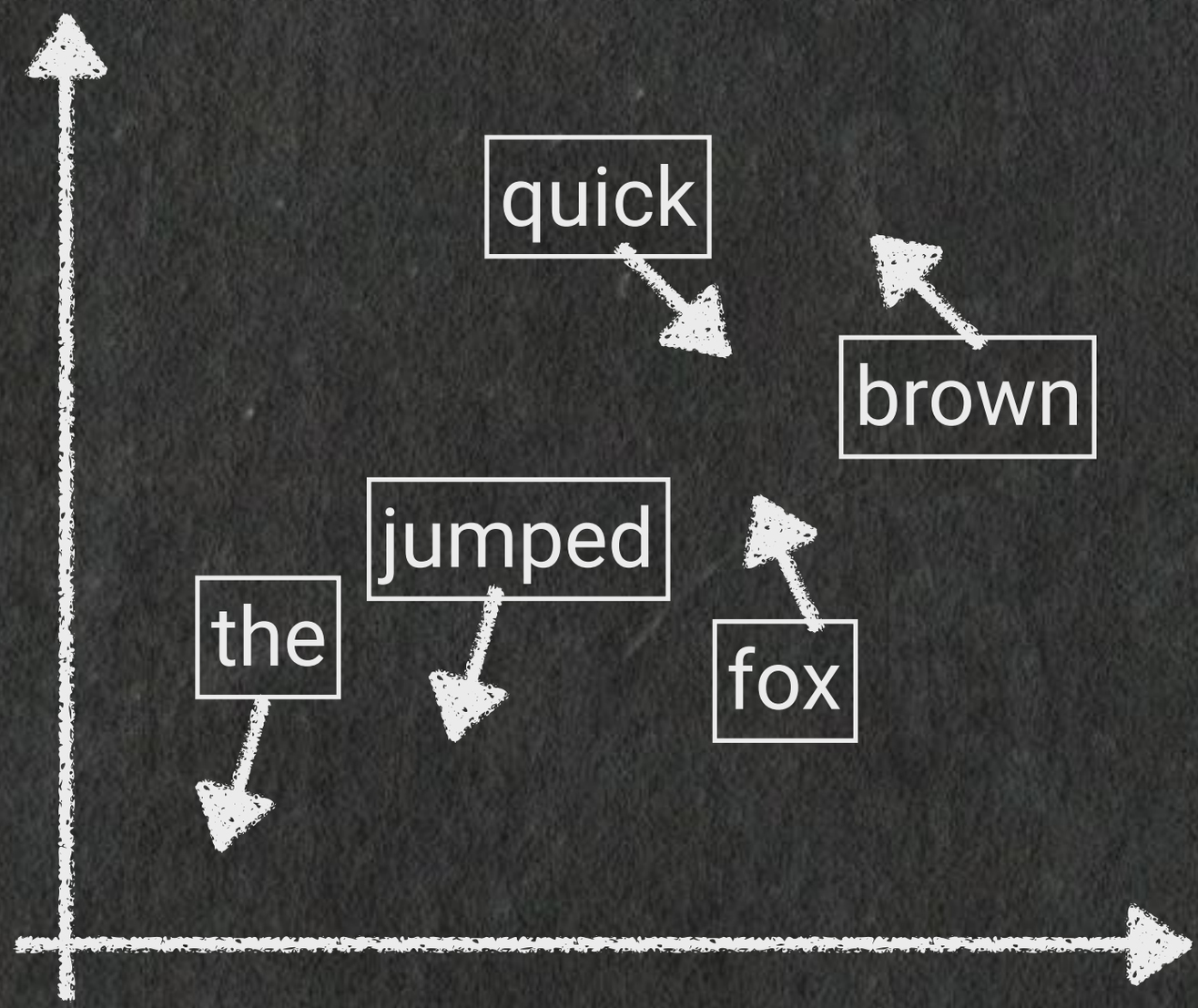


the quick brown fox jumped over the lazy dog

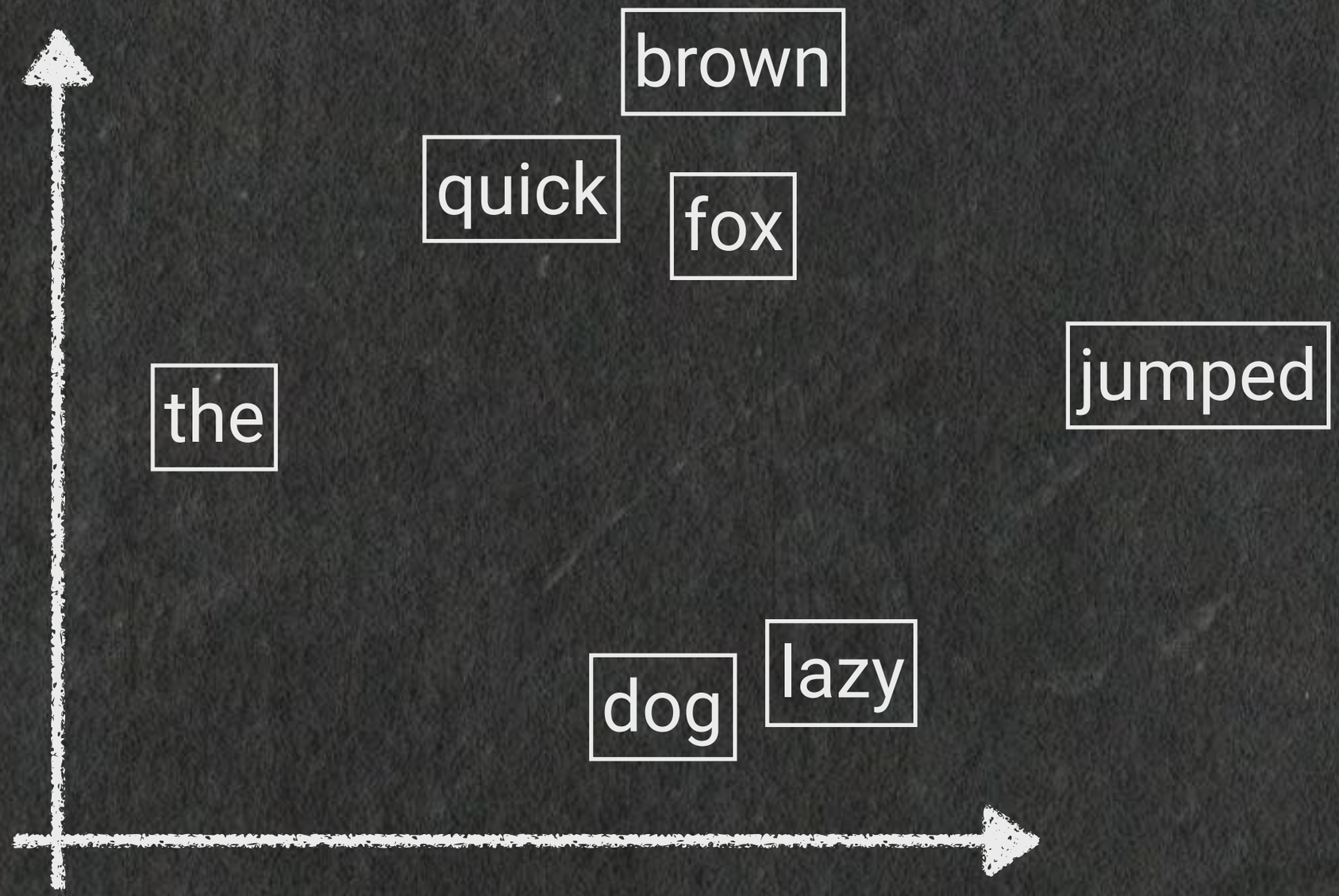


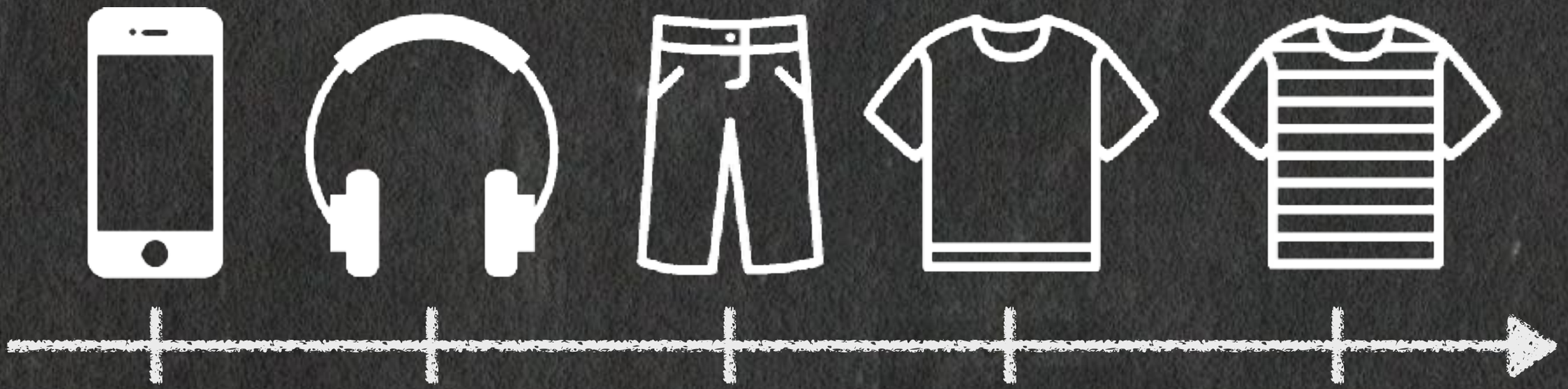
the ~~quick~~ brown ~~fox~~ jumped over the lazy dog

? ?



the quick brown fox jumped over the lazy dog





One size fits all

Criteo

120

billions events

Criteo

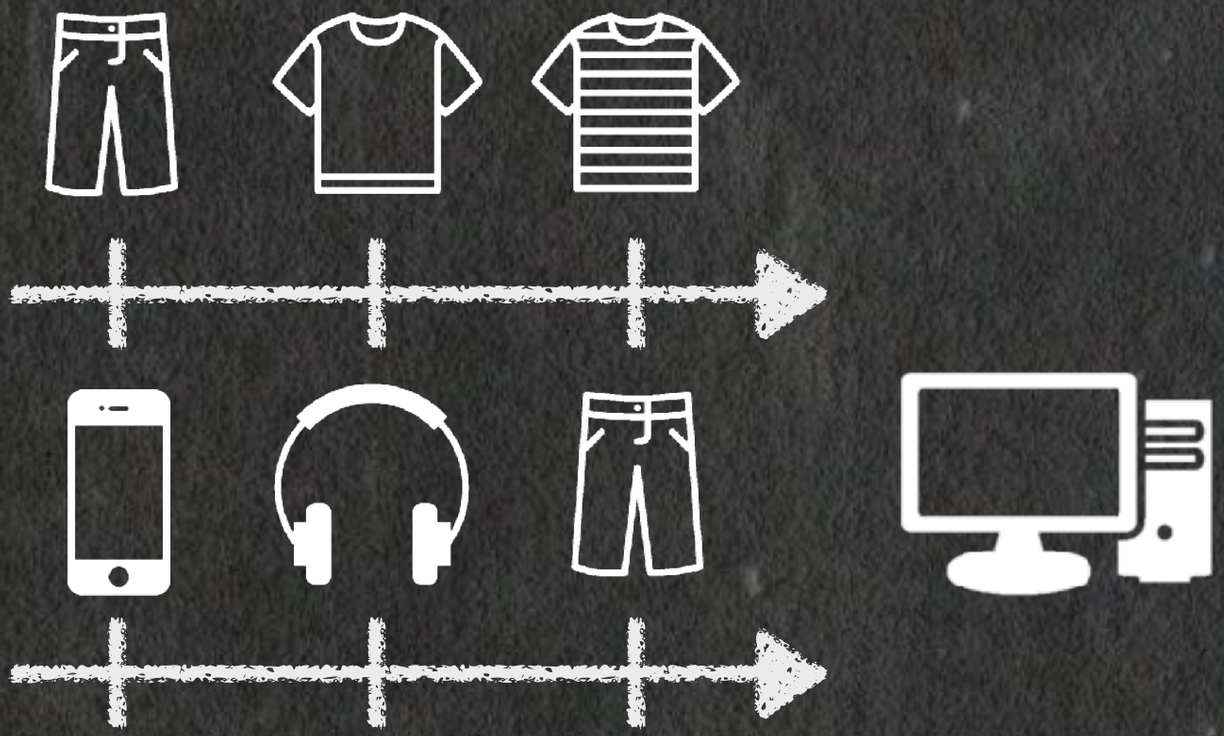
120

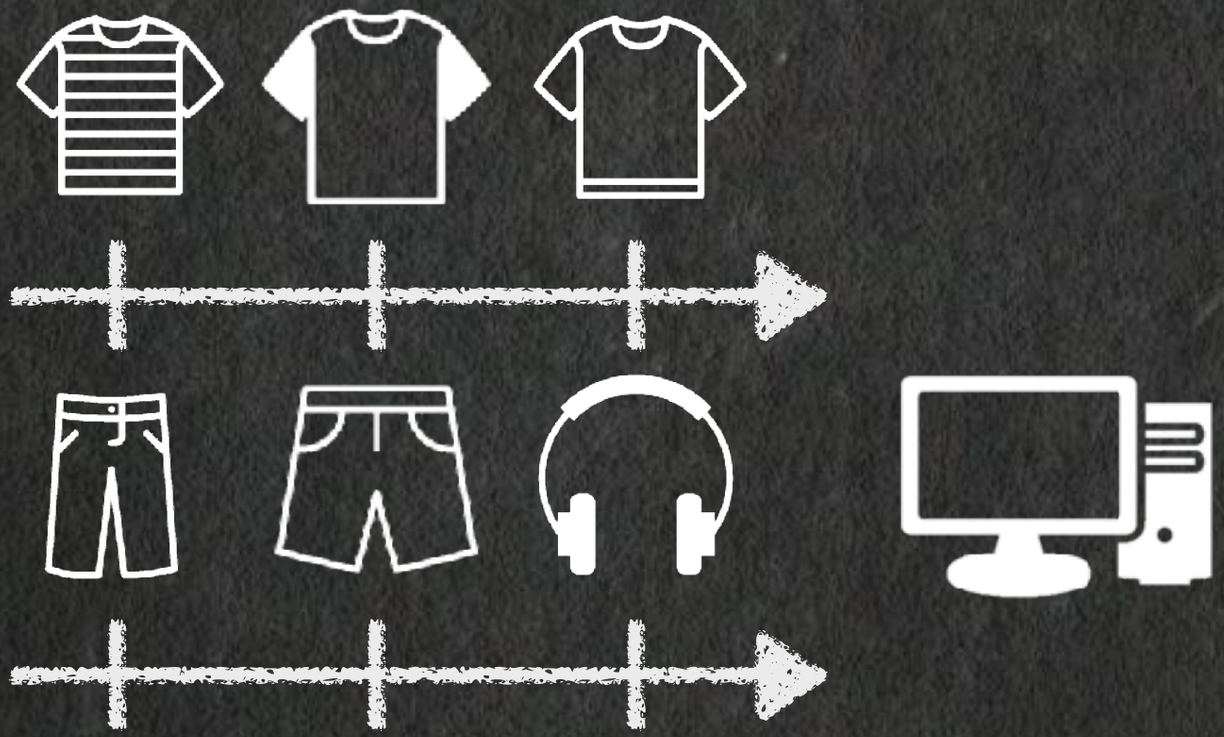
billions events

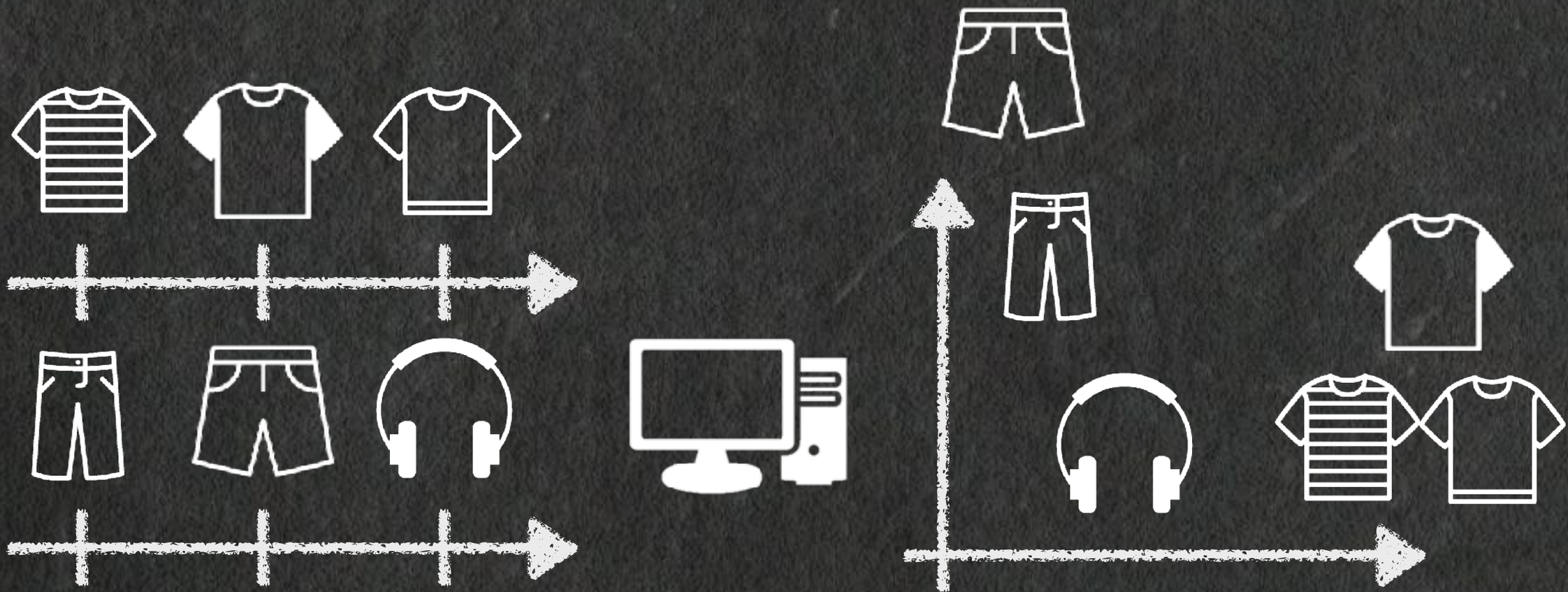
Wikipedia

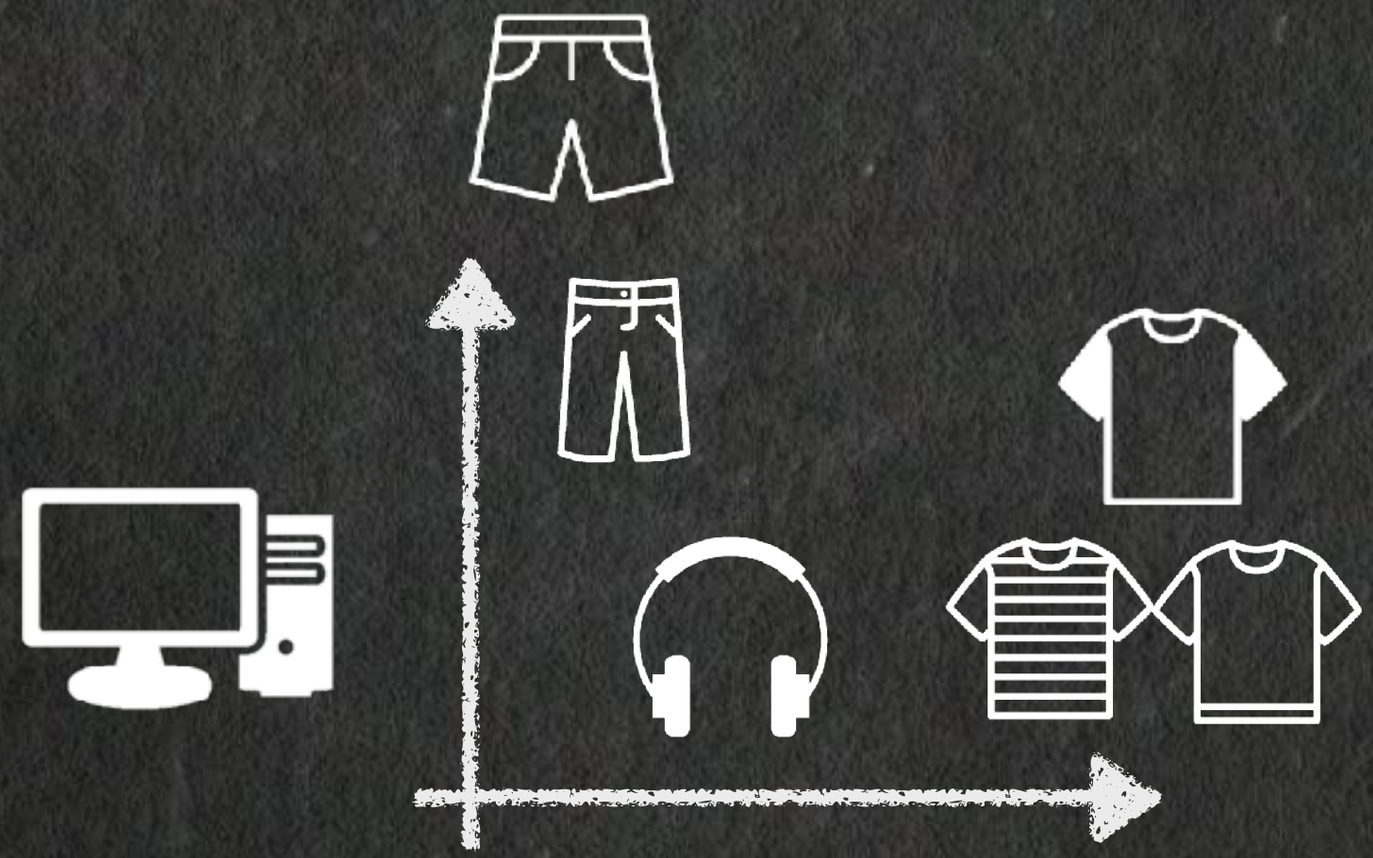
30

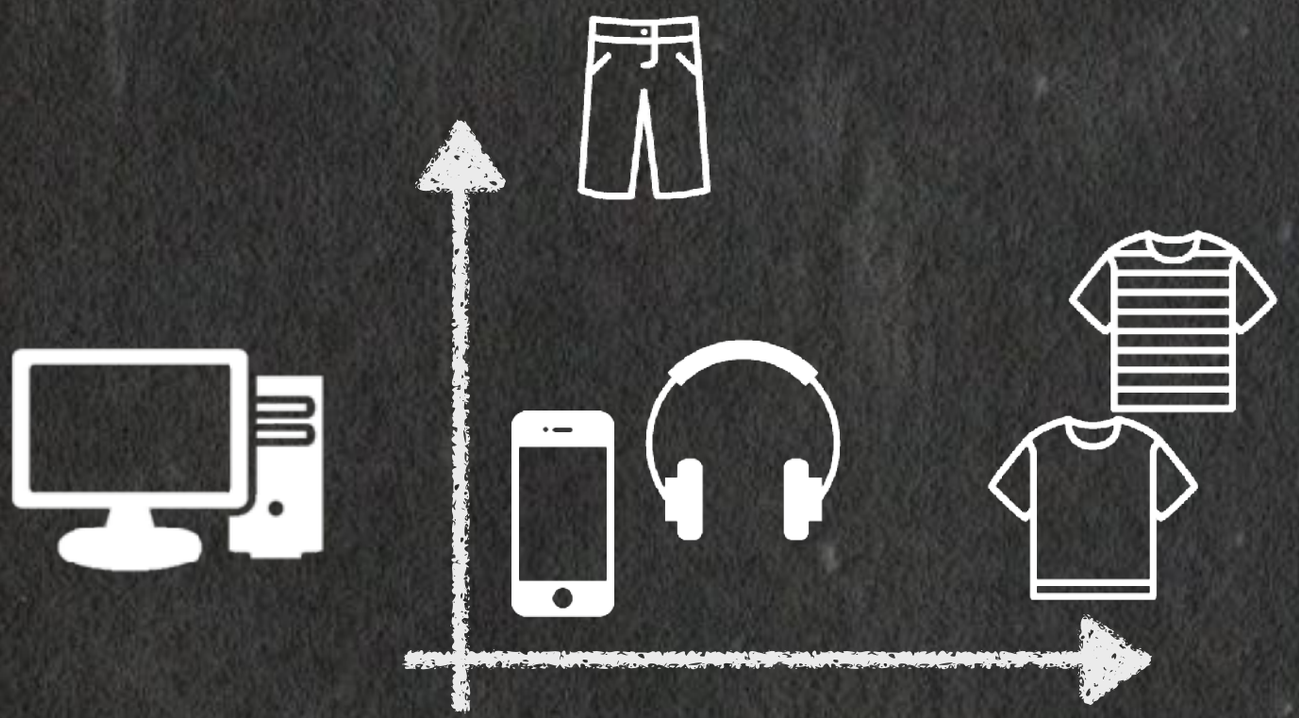
billions words

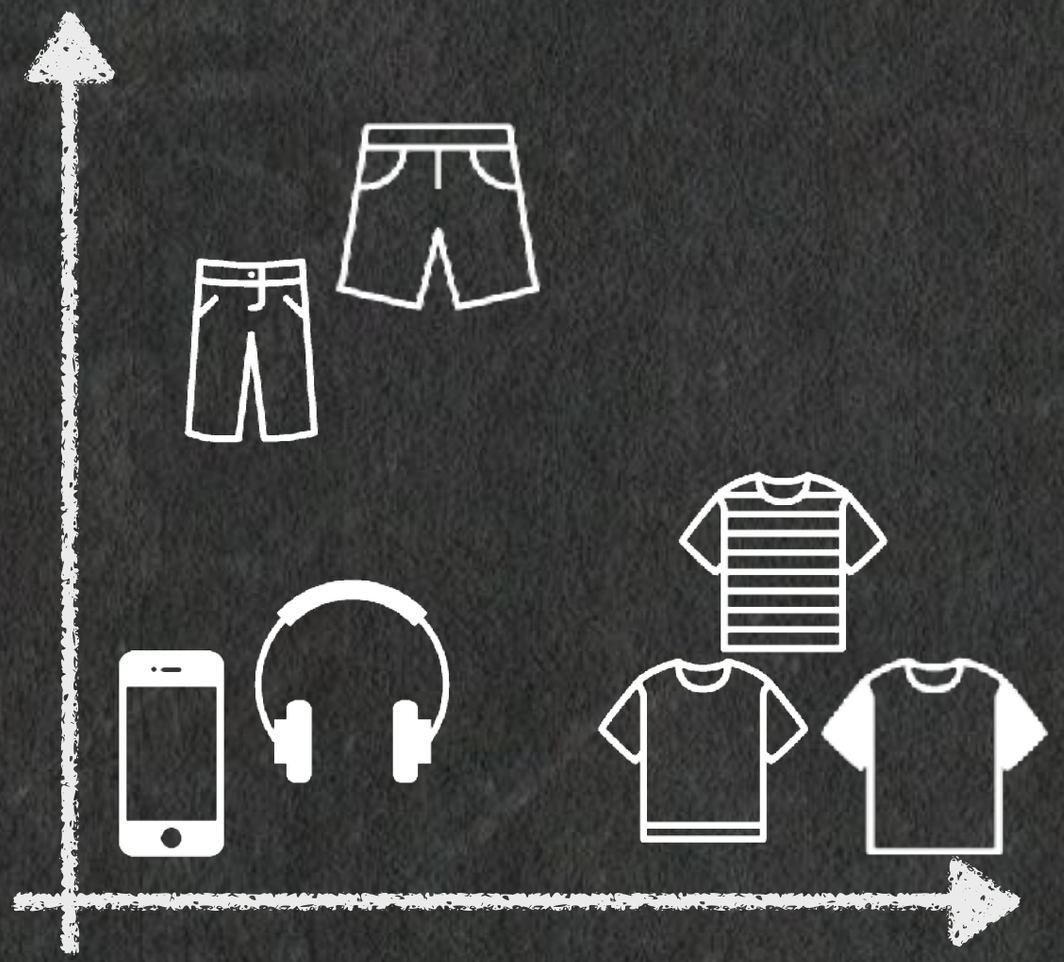


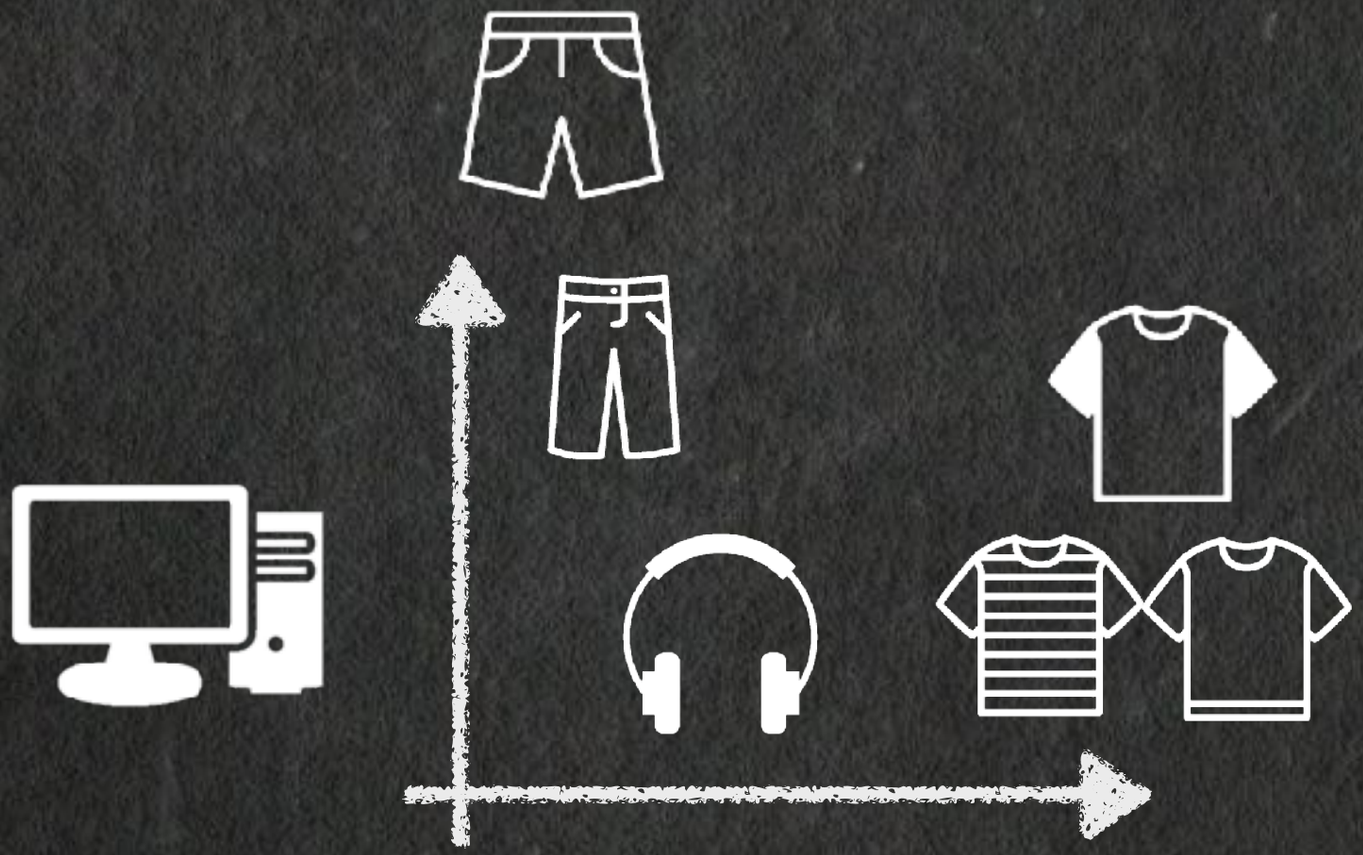


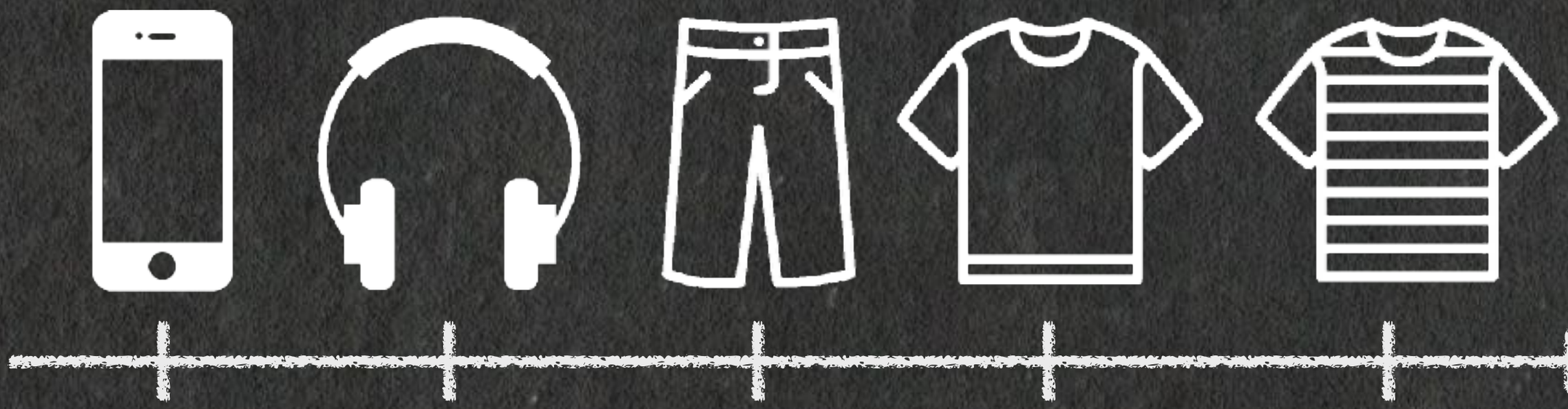


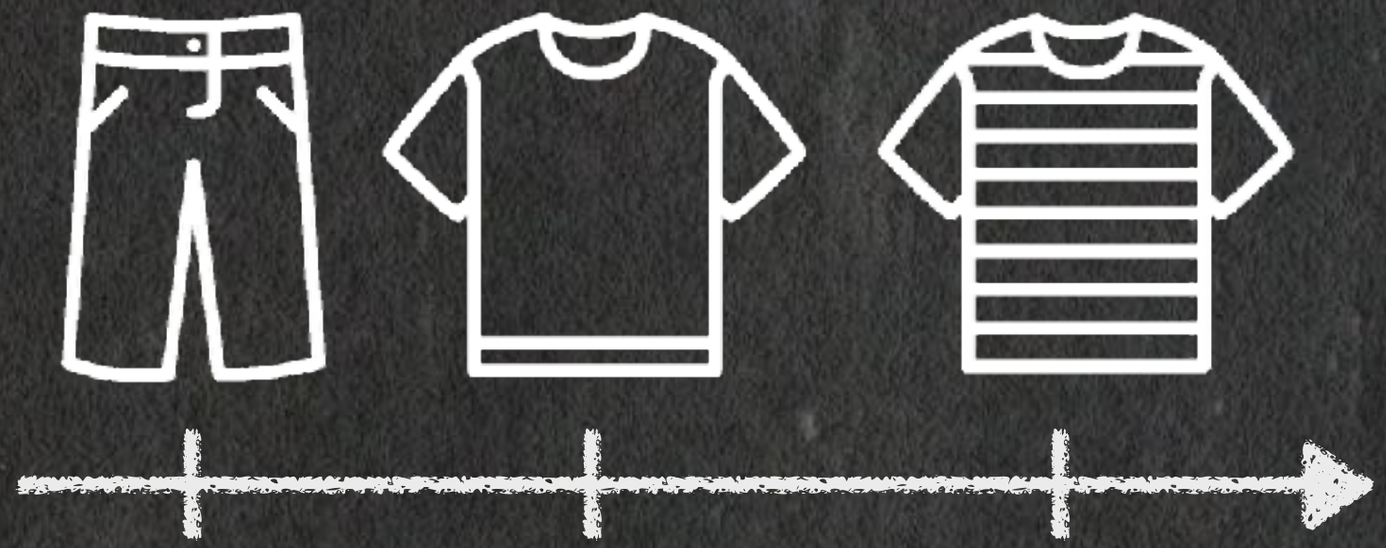


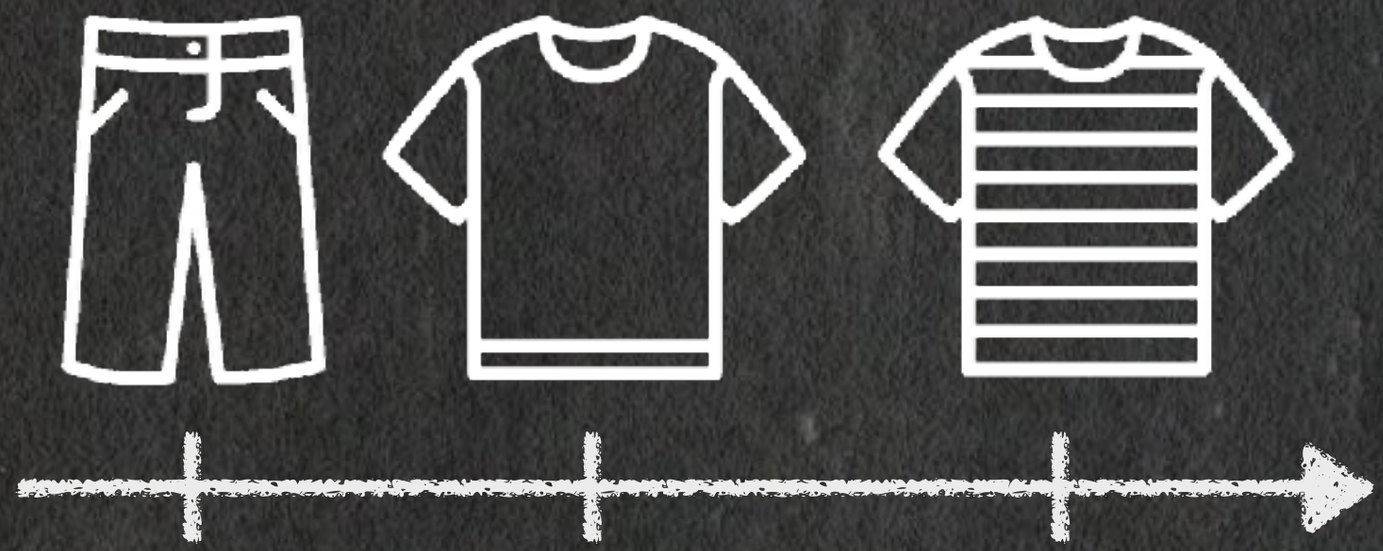


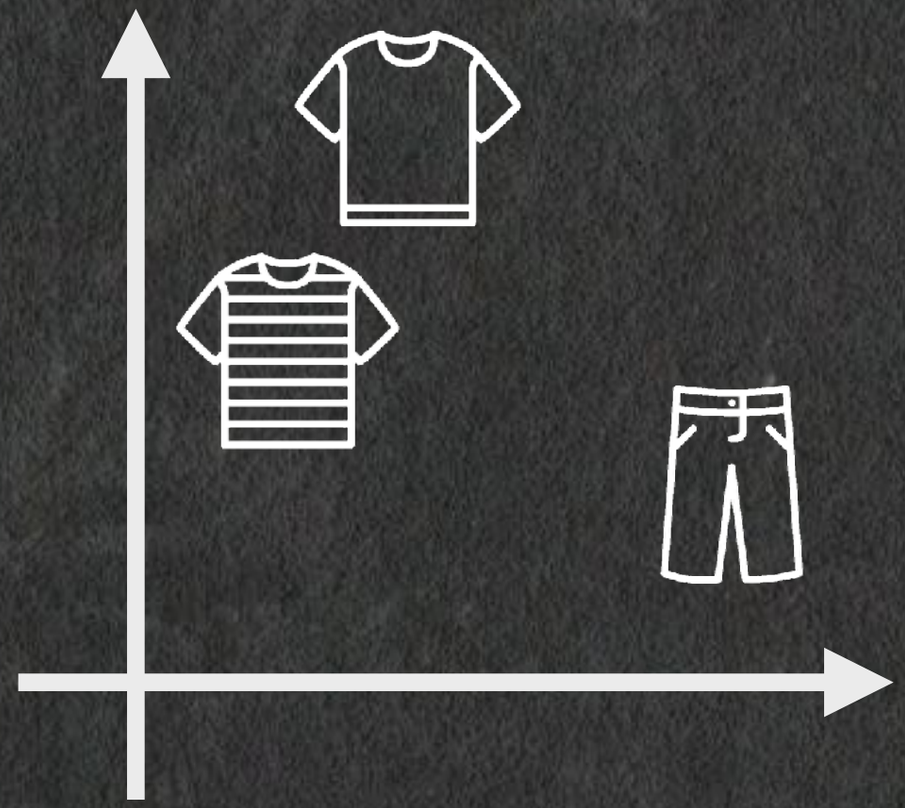


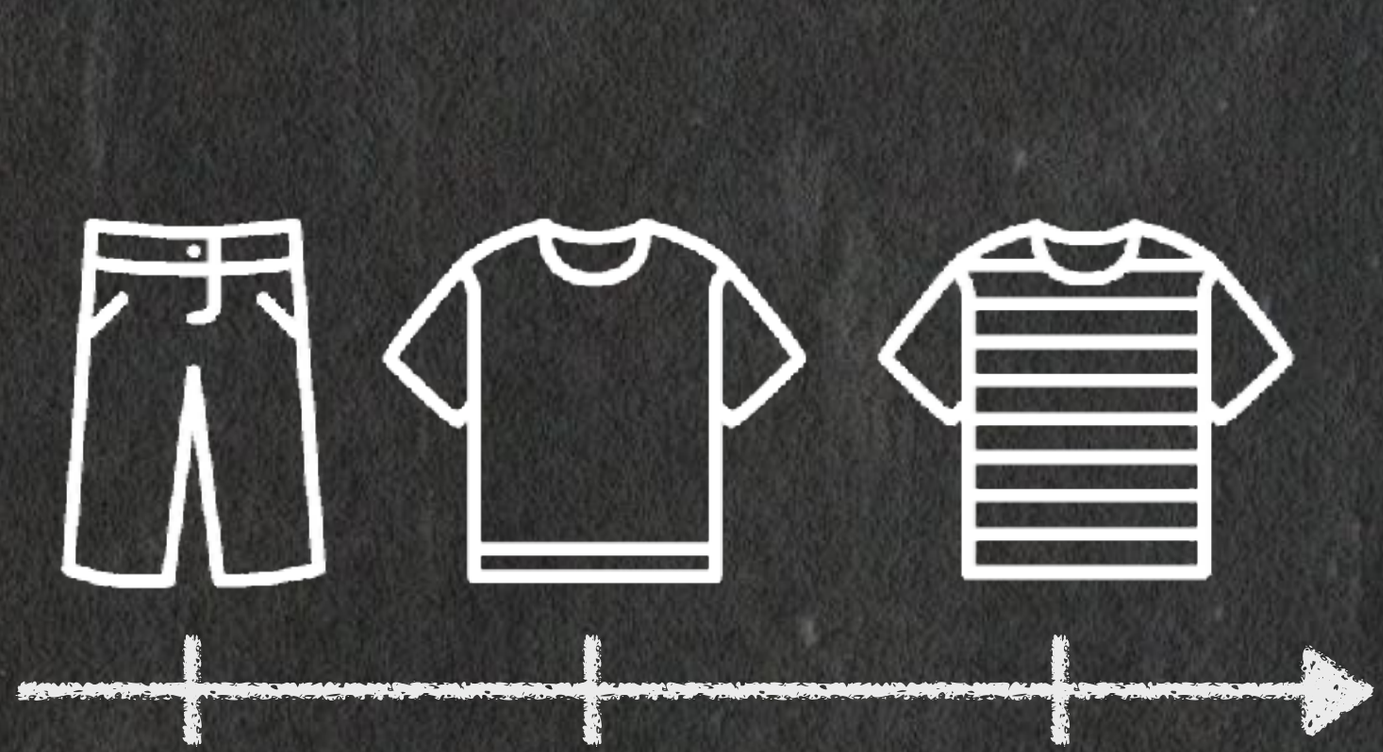


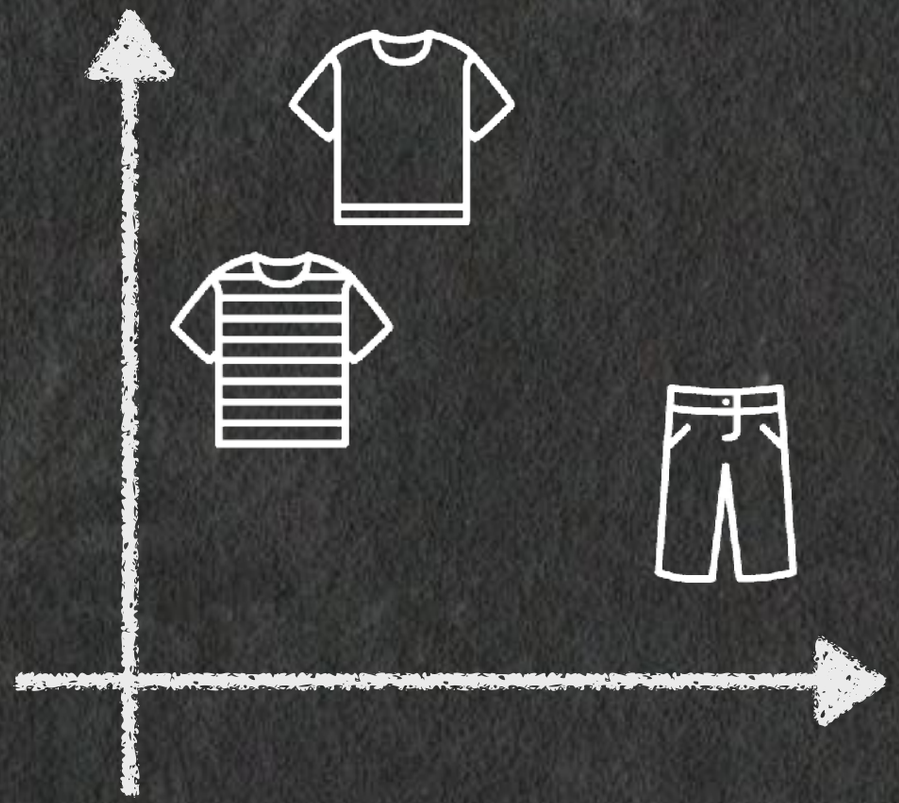


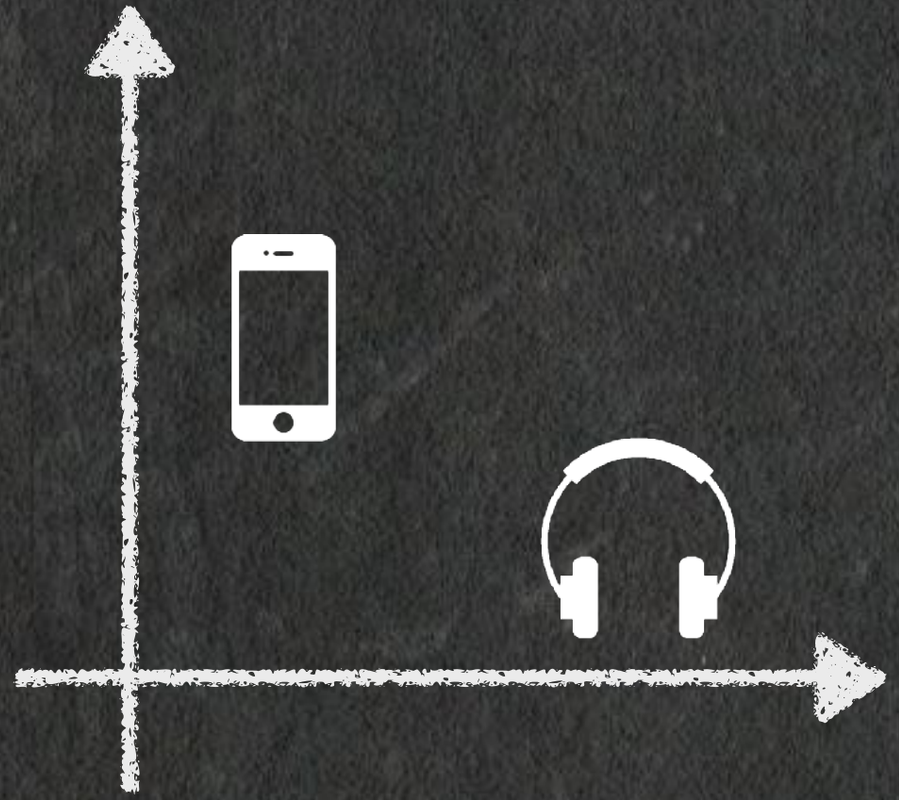












Metrics?
What for?

Llh

measure
probabilities

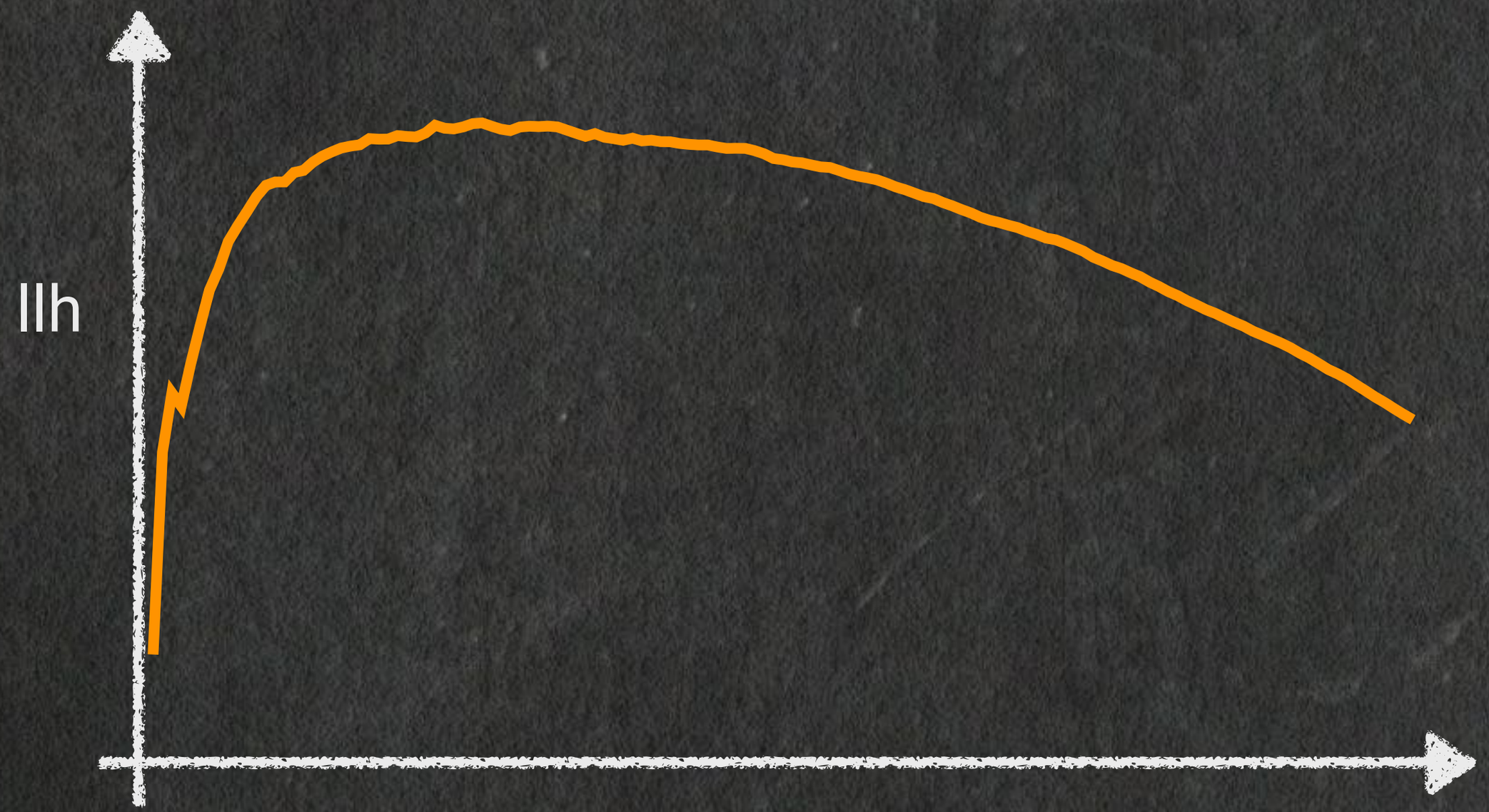
close to what
we optimize

Llh	Recall@K
measure probabilities	measure ranking
close to what we optimize	close to final goal

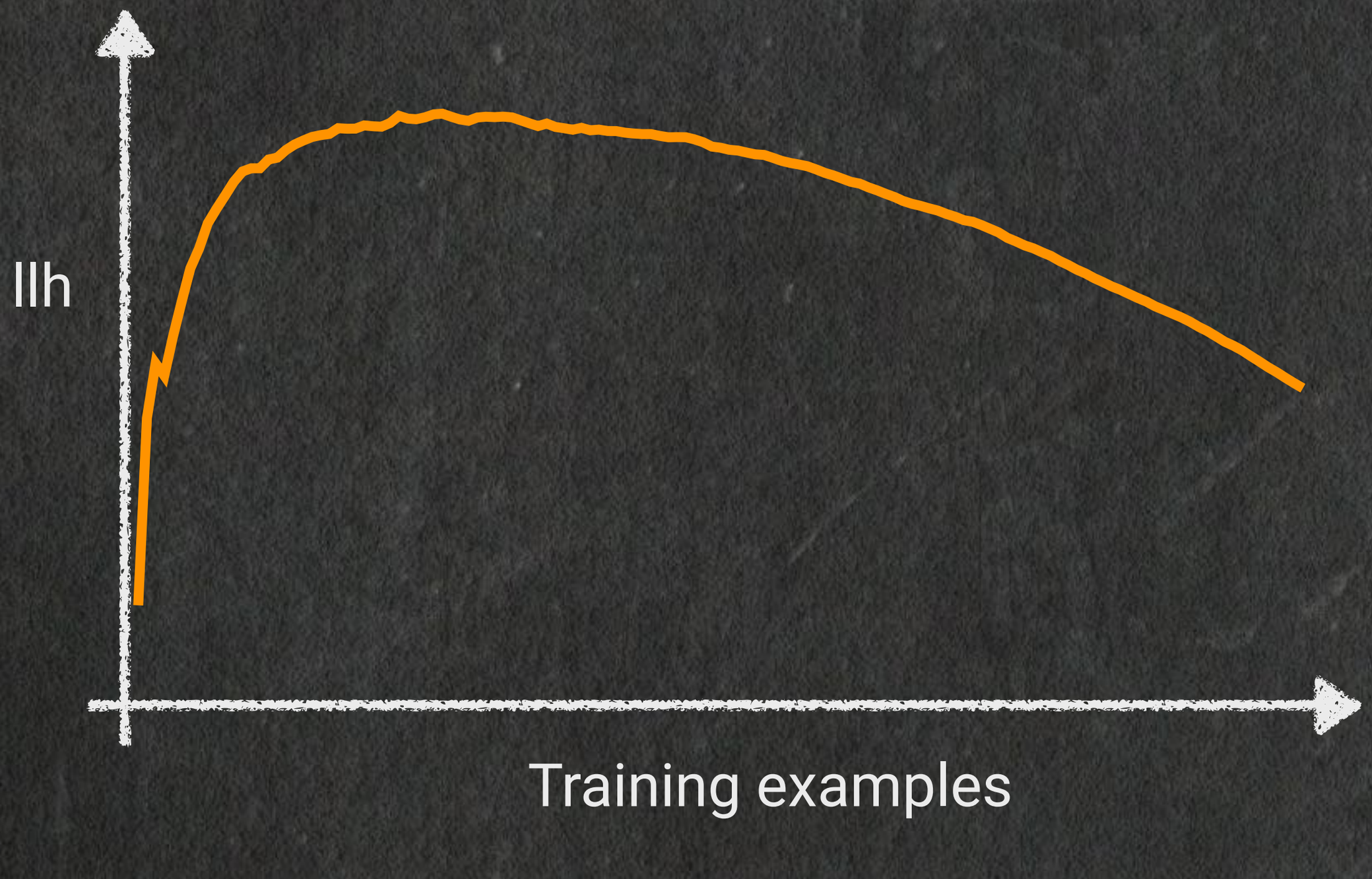
- Number of partitions
- Number of epochs
- Learning rate
- Embedding dimension

What about **clicks**?

Data betrayled us!


llh
Training examples

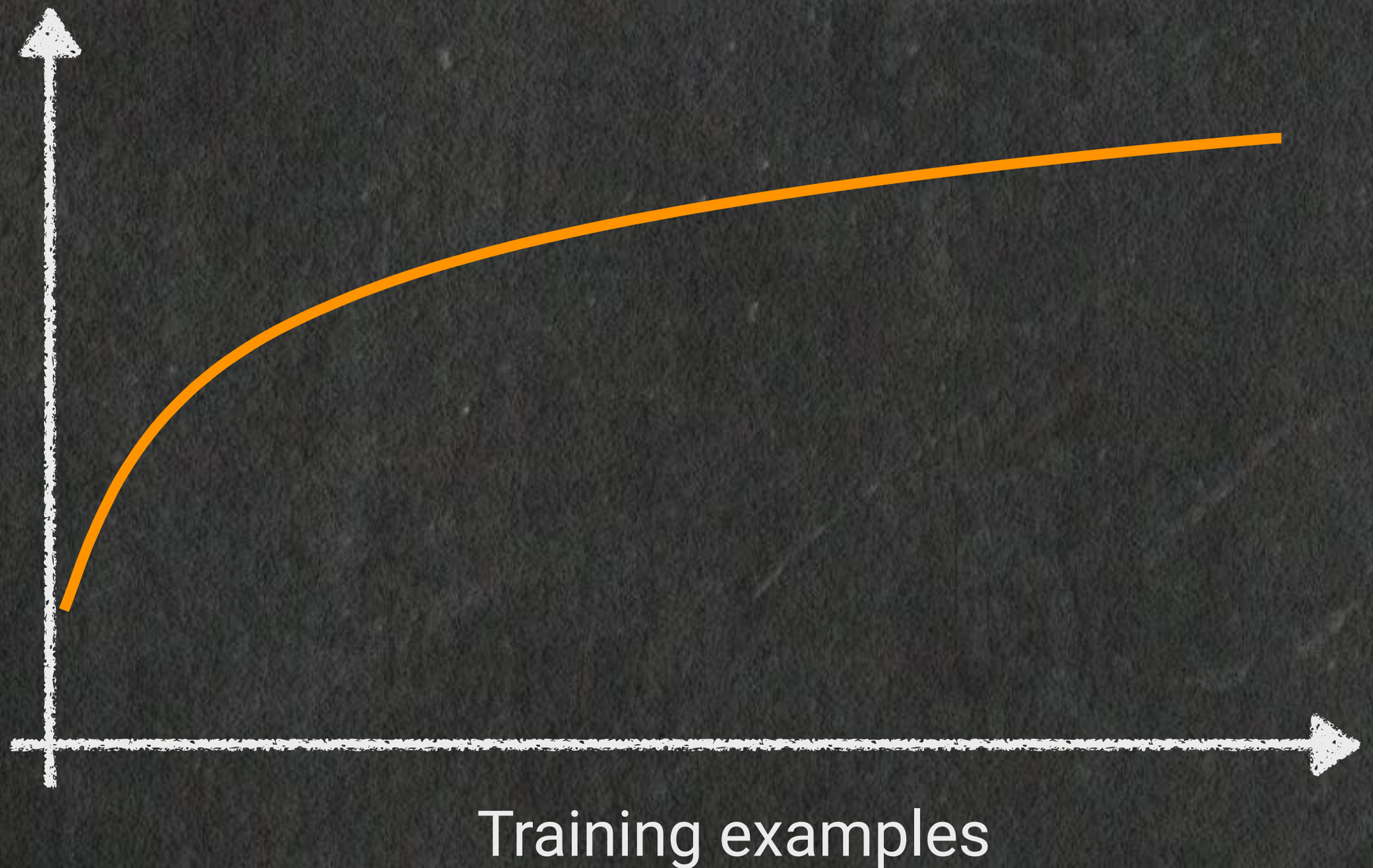
shuffle → process sentence → generate pairs



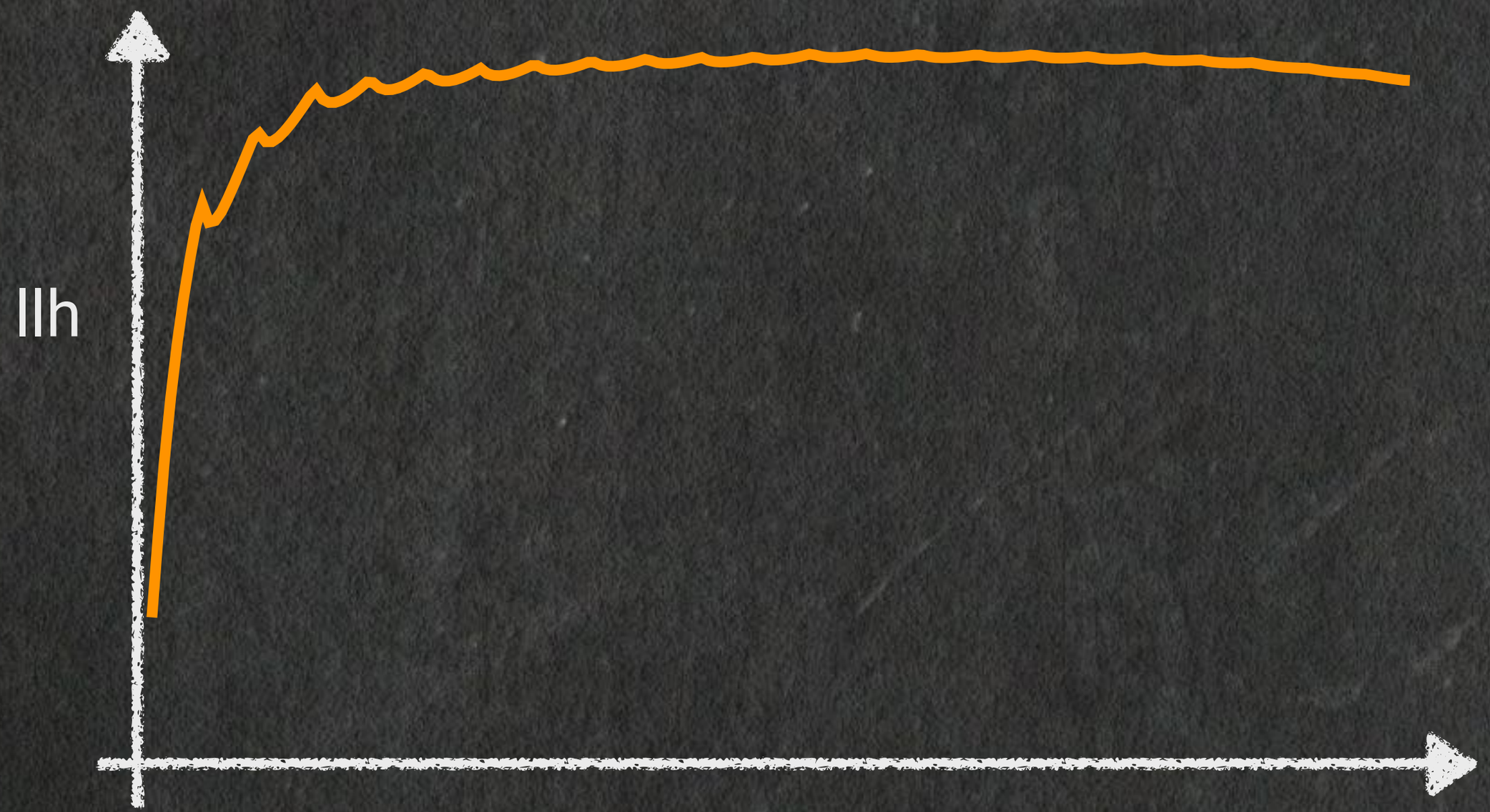
generate pairs

shuffle

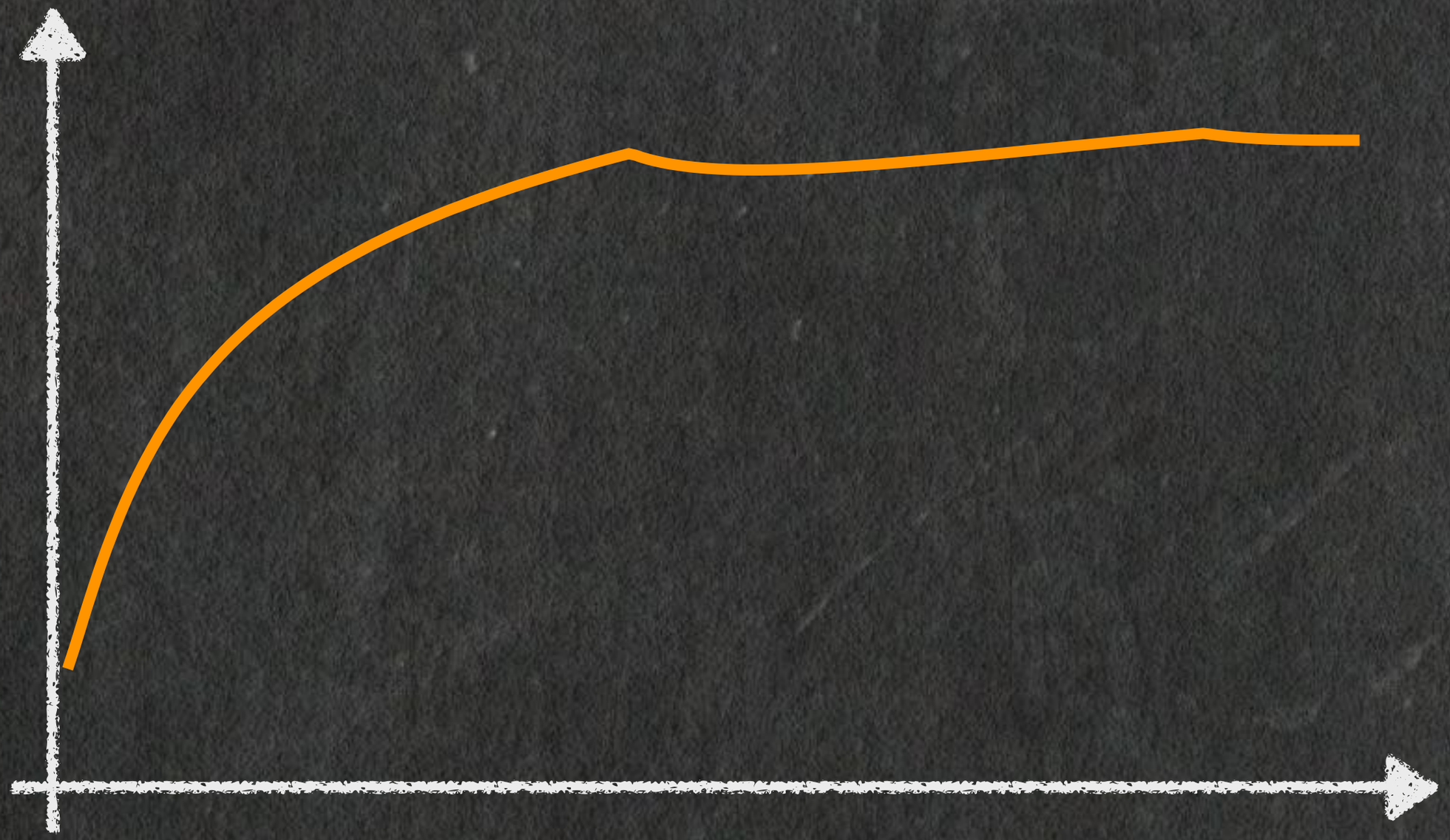
process pairs



shuffle → process batch → generate pairs → shuffle → process pairs



shuffle → process batch → generate pairs → shuffle → process pairs



Embrace **trade-offs**

Embrace **trade-offs**

Measure everything

Embrace **trade-offs**

Measure everything

Play with your data

We're hiring!

@simondolle

http://labs.criteo.com/

Credits: Ralf Schmitzer, Edward Boatman, Bernar Novalyi, Royyan Wijaya, art shop, iconsphere

criteo