Hard

Hard: Outline

- Academia vs industry
- Build an LLM; crawl the Web, build a production inverted index etc.
- Research opportunities; pros/cons; pointers to related work
- Pre-Training

Indexing and crawling

- Small number of web-indexes available
  - English (Google and Bing), Chinese (Baidu), Russian (Yandex)
- Inverted indexing techniques are well understood
- Large scale is a different story
  - MapReduce architectures
- Updating and serving
- Ongoing crawling and data wrangling
- Social networks
  - The Like economy

Relevance

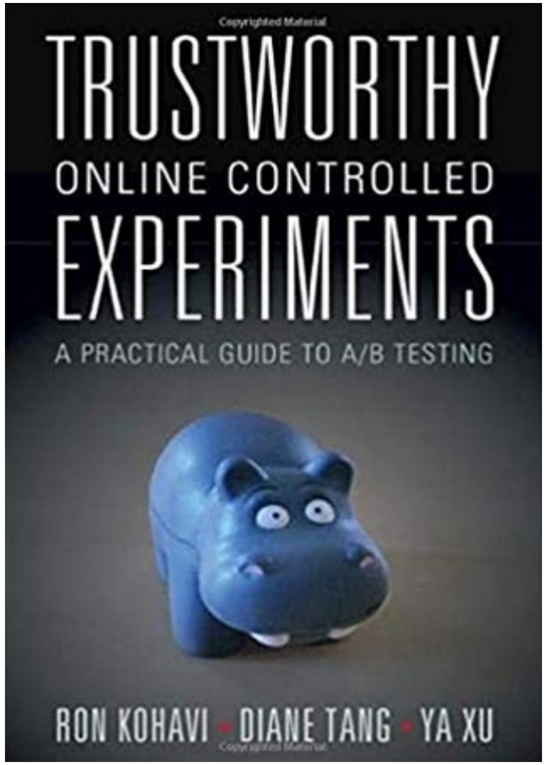
- Many factors affect whether a document satisfies a particular user's information need
- Topicality, novelty, freshness, authority, formatting, prior knowledge, expertise
- Topical relevance: the document is on the same topic as the query
- User relevance: everything else
- Topical relevance
  - Focusing on topical relevance does not mean we're ignoring everything else
  - It only means we're focusing on one criteria by which users judge relevance
- Domain specific features

Large scale experimentation

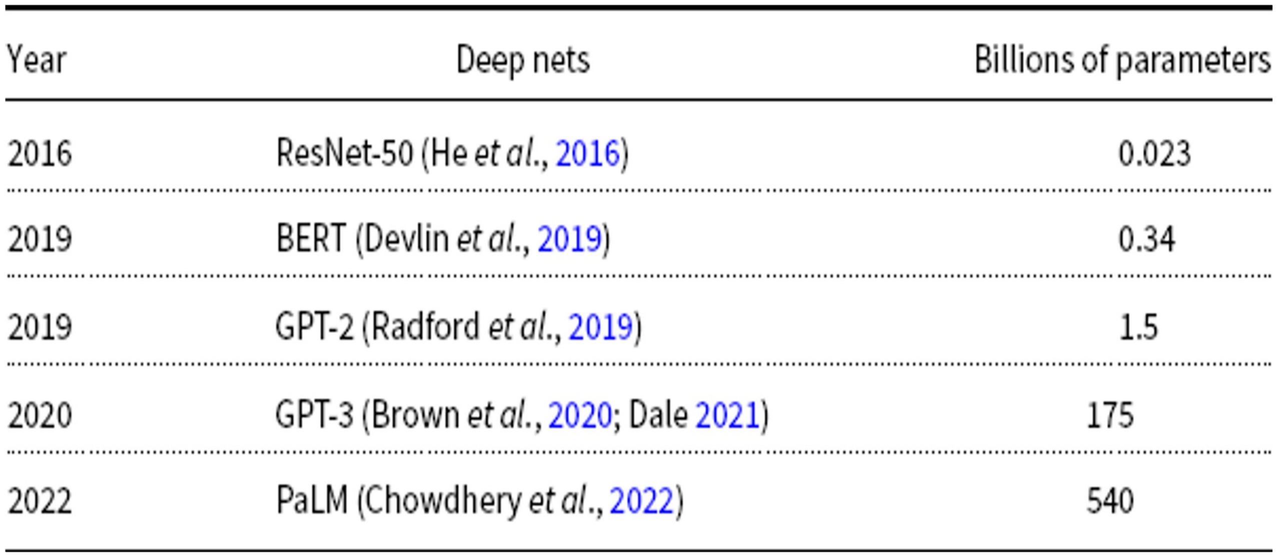
- Principles for designing, running and analyzing experiments
- Imagine an e-commerce website where the user:
  - Visit the homepage -> browse/search for items -> add item to cart -> start purchase process -> complete purchase
- Product team wants to add coupon feature to UX
- We need to evaluate the impact of the change
- Hypothesis: adding new feature will increase revenue
- Two user interfaces:
  - Control (no changes)
  - Treatment (with coupon feature)

Designing experiments (A/B testing)

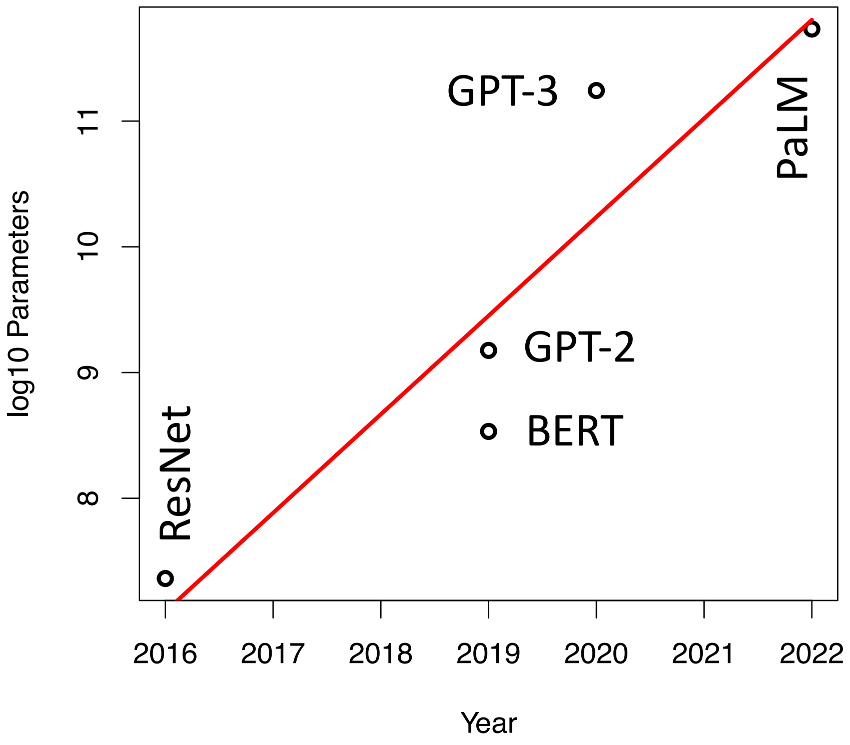
- Randomization unit
- Population
- How large does our experiment need to be
- How long do we run the experiment
- Does the experiment scale well? (1%, 2%, 5%, 10%, …)
- https://experimentguide.com/



Pre-Training Large Language Models (LLMs)



Year	Deep nets	Billions of parameters
2016	ResNet-50 (He *et al.*, 2016)	0.023
2019	BERT (Devlin *et al.*, 2019)	0.34
2019	GPT-2 (Radford *et al.*, 2019)	1.5
2020	GPT-3 (Brown *et al.*, 2020; Dale 2021)	175
2022	PaLM (Chowdhery *et al.*, 2022)	540



Most users should not invest in pretraining because growth (& costs) are out of control

https://github.com/kwchurch/WSDM_2024_tutorial