DATA SCIENCE
&
OUR WORK@ FUM

Faezeh Ensan

Lots of Data is being Generated and Collected

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs of log data every day

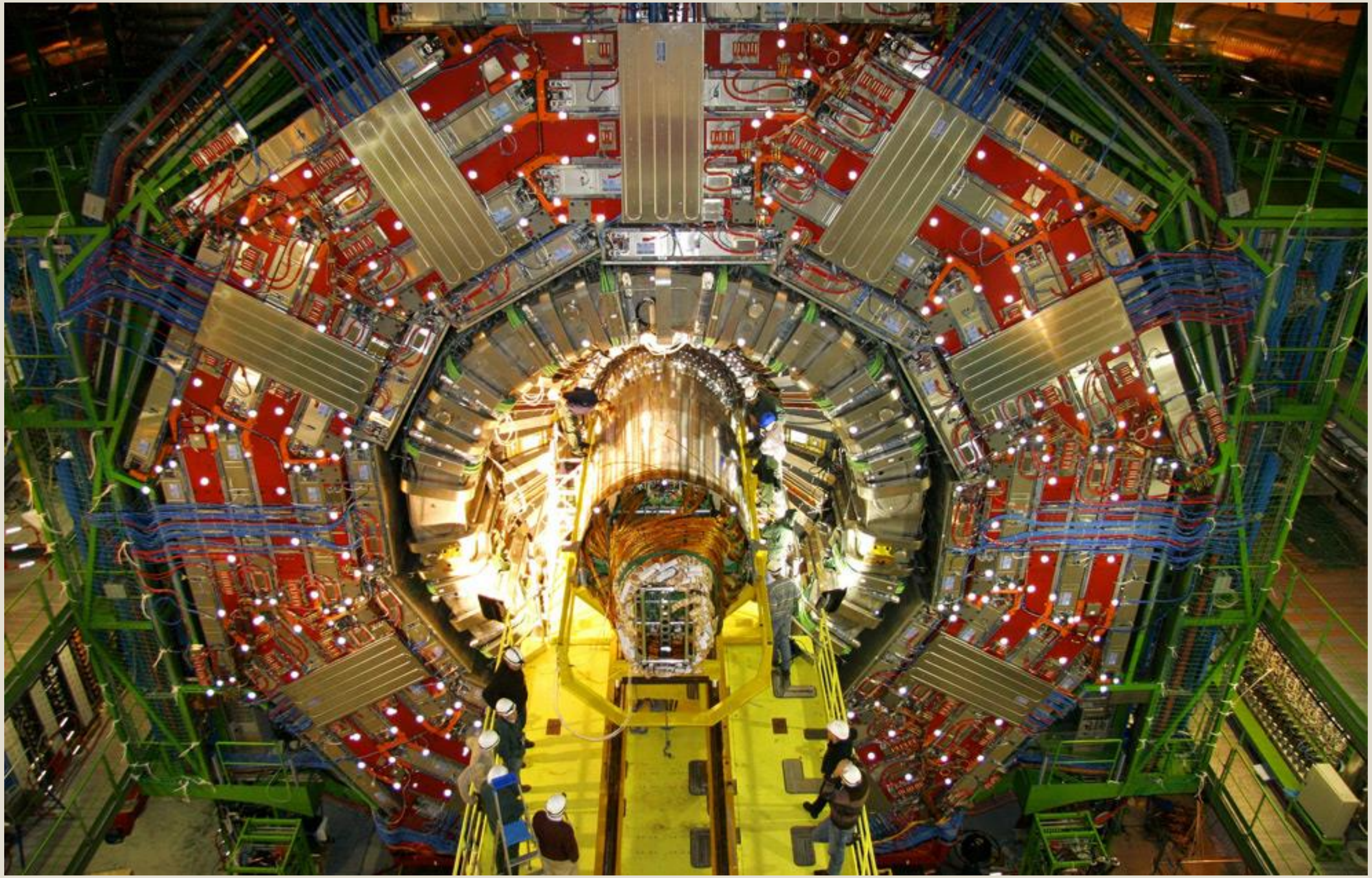
30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

2+ billion people on the Web by end 2011

76 million smart meters in 2009… 200M by 2014



CERN’s Large Hydron Collider (LHC) generates 15 PB a year

The Earthscop

- Earthscope
  - the world's largest science project.
- To track North America's geological evolution
  - records data over 3.8 million square miles, amassing 67 terabytes of data.

What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Semantic search
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Data Science: Why?

e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC
(Center for disease and prevention)
data

New models are estimating
which cities are most at risk
for spread of the Ebola virus.

Data Science: Why?

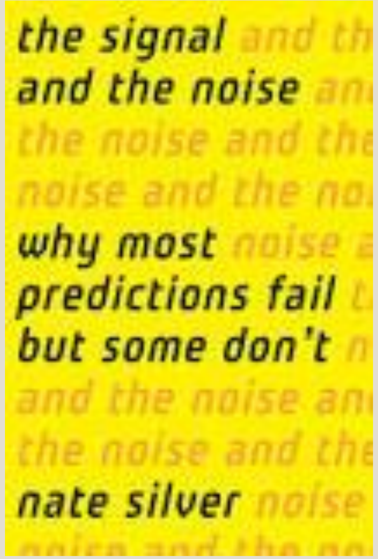
elections2012

Live results | President | Senate | House | Governor | Choose your

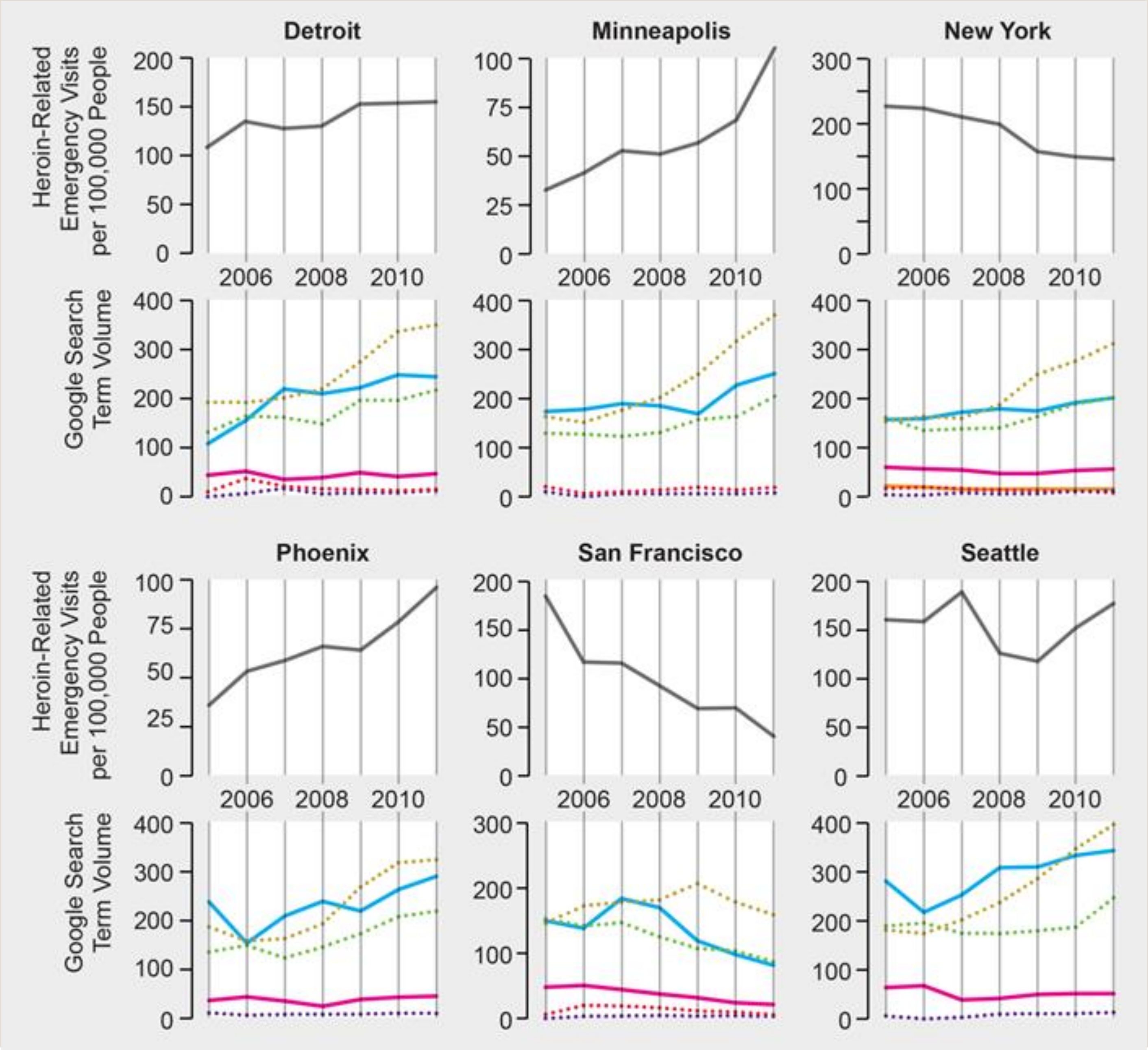
Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST

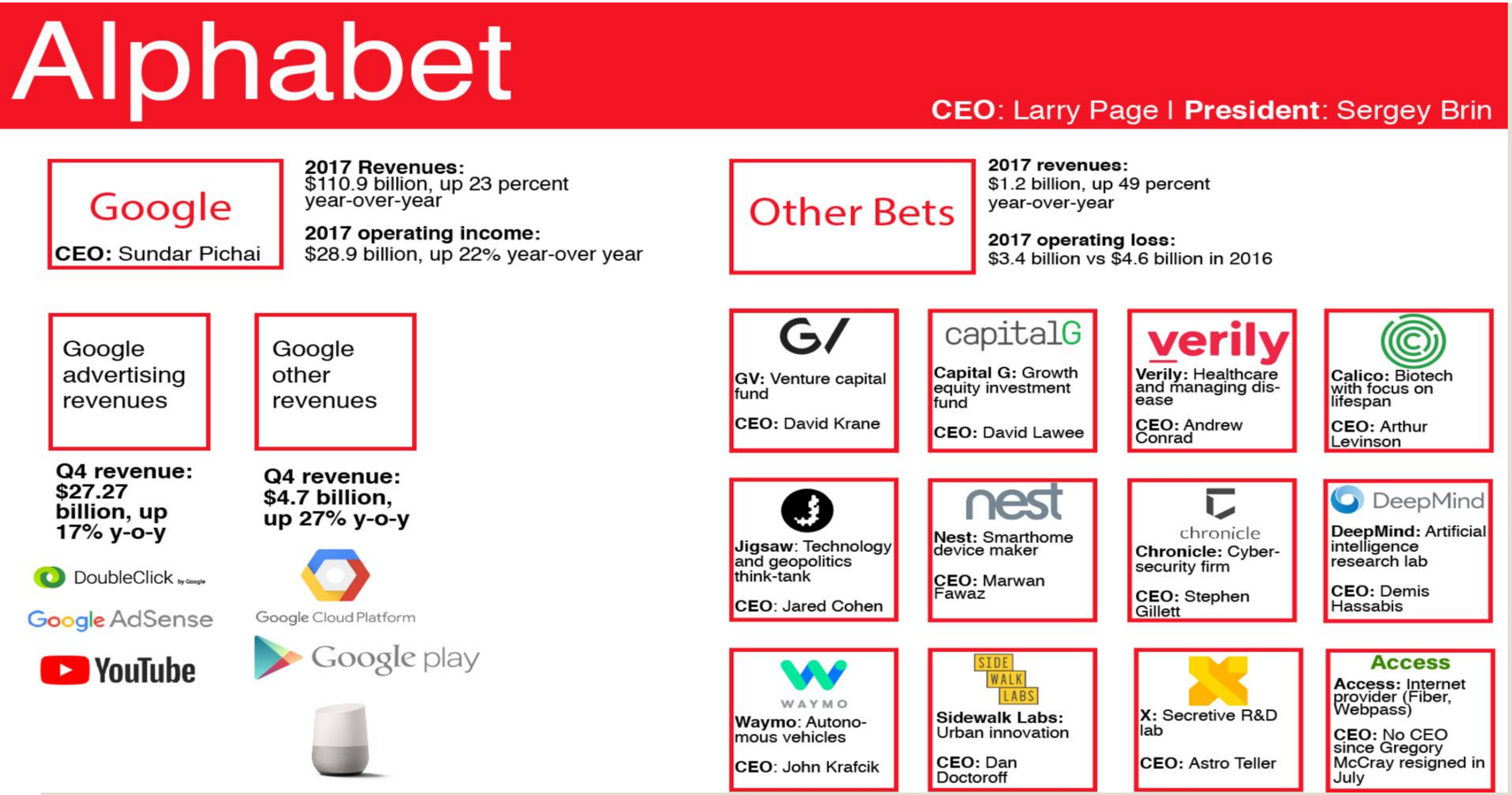


Data Science: Why?



Google Searches Could Predict Heroin Overdoses

- Relations: opioid-related keywords, metropolitan income inequality and total number of emergency room visits.
- Findings: regional differences (*graphic*) in where and how people searched for such information and found that more overdoses were associated with a greater number of searches per keyword.
- The best-fitting model, explained about 72 percent
- Brown Sugar?



The most economically important application

The company reported fourth quarter earnings : which revealed that it still makes 84 percent of its revenue from advertising, with 14.5 percent coming from the likes of its cloud unit and hardware, and 1.2 percent coming from its so-called Other Bets, like its Fiber internet service and Nest smart-home products

Our work at FUM

Semantic Recommendation, Semantic Search, and Social Networks

Semantic Recommendation

- TREC Competition:
  - The TREC Contextual Suggestion Trac: a personalized point of interest (POI) recommendation task, related to a profile and a context.
  - Data: 1,235,844 URLs. This crawl includes web pages from different domains like yelp, tripadvisor and foursquare.
- Our approach: Category-based and semantic-based models

- *Ranked #4 in the competition*

Ensan, et.al. A Context Based Recommender System through Collaborative Filtering and Word Embedding Techniques. In *TREC 2016*

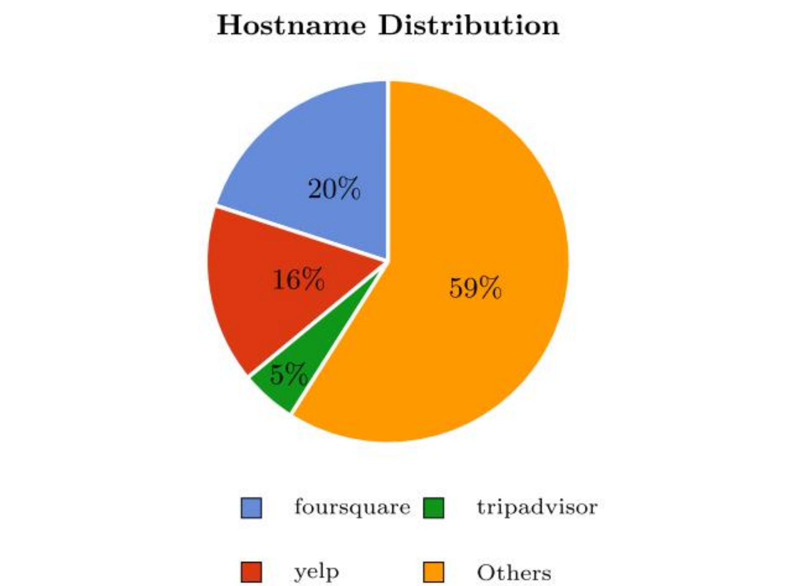


Figure 1: Most popular domains in the TREC Contextual Suggestion Web Corpus.

Semantic Search [Language model]

$$P(Q_{q_j}|D_d) =$$
$$\begin{cases} (1-\lambda)P_{selm}(Q_{q_j}|D_d) + \lambda P(Q_{q_j}|Col) & \textit{similar concept found} \\ \lambda P(Q_{q_j}|Col) & \textit{Otherwise} \end{cases} \quad (2)$$

Based on this model, we wish to find $P_{selm}(Q_{q_j}|D_d)$, the probability of a given query concept based on a given document. According to [16], we have:

$$P_{selm}(Q_{q_j}|D_d) = \frac{1}{Z(D_d)} exp(\sum_{i=1}^{i=k} f_i(C_i, q_j, D_d)) \quad (3)$$

where $C_i \subseteq V$ is a clique over G and $C_i \not\subset D$, f_i is a feature function defined over C_i. $Z(d)$ is a normalization factor and is defined as:

$$Z(D_d) = \sum_j exp(\sum_{i=1}^{i=k} f_i(C_i, Q_{q_j}, D_d)) \quad (4)$$

- Documents and Queries: Sets of Entities
- Entity: Entries of Wikipedia

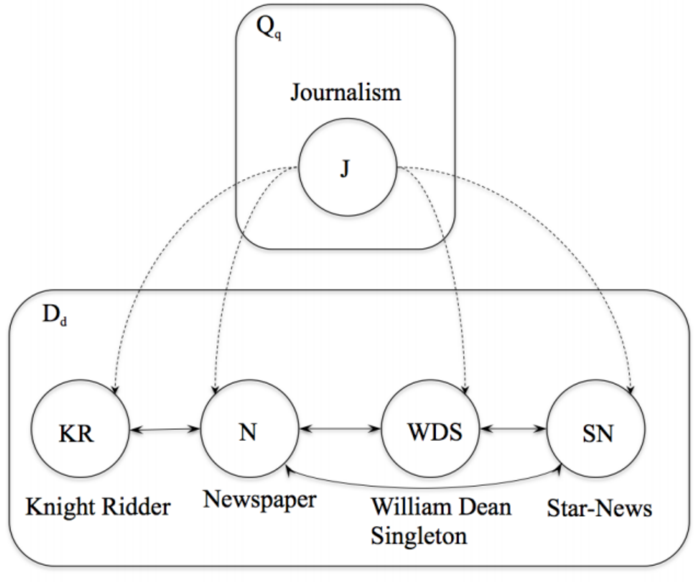
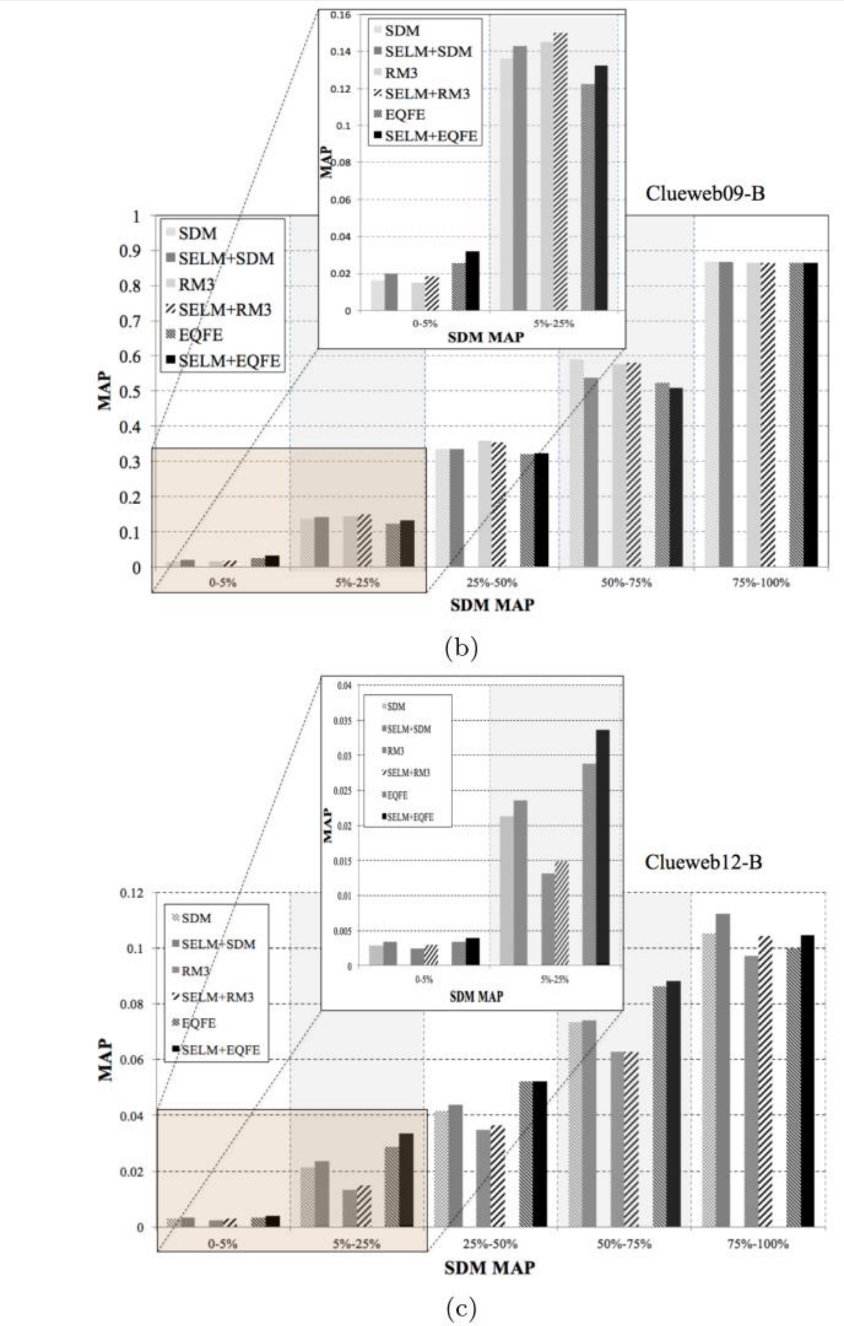


Figure 1: Sample query and document relationship model.



Results and Insight on Future Work

Hard vs. Soft Queries
MSc. Thesis

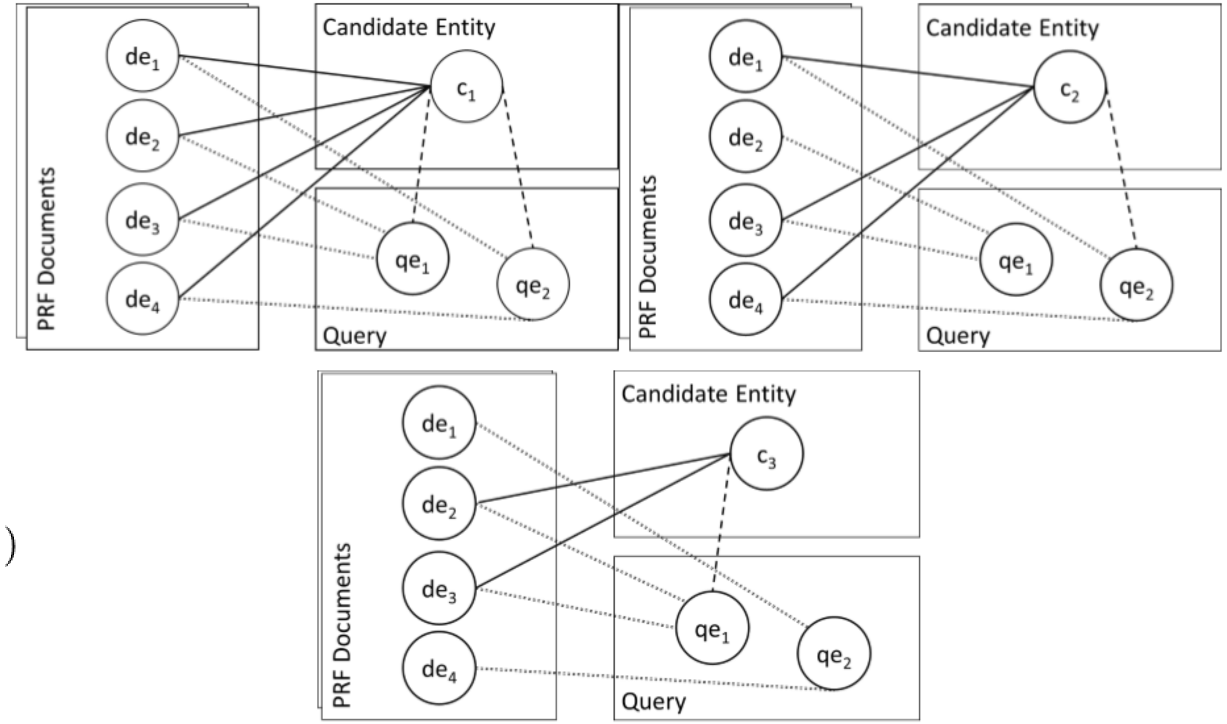
- Ensan, et.al . "Ad hoc retrieval via entity linking and semantic similarity." ***Knowledge and Information Systems*** (2018): DOI: https://doi.org/10.1007/s10115-018-1190-1
- Ensan, et.al . Document retrieval model through semantic linking." *Proceedings of the tenth ACM international conference on web search and data mining*. ACM, 2017. **(WSDM 2017)**

Semantic Search [Query Expansion]

$$f_{rank}(c|q) \approx \Sigma_{d\in R} P(c,q|d)P(d)$$

$$P(qc|d) = \frac{1}{Z(d)} exp\left(\sum_{i=1}^{i=k} f_k(Cl_i, qc, d)\right)$$

$$f_k(Cl_i, qc, d) = \sum_{d_j \in d} ef(d_j, d) \times Sim(Cl_i, qc, d_j)$$



Results and Insight on Future Work

	ClueWeb09B				ClueWeb12B			
	MAP	ΔMAP	NDCG@20	ΔNDCG@20	MAP	ΔMAP	NDCG@20	ΔNDCG@20
RM	0.1994†	-0.0260 (-13.06%)	0.2554†	-0.0723 (-28.29%)	0.0357†	-0.0215 (-60.16%)	0.1085†	-0.0670 (-61.80%)
SDM	0.1916†	-0.0339 (-17.69%)	0.2488†	-0.0789 (-31.70%)	0.0417†	-0.0155 (-37.24%)	0.1239†	-0.0516 (-41.66%)
EQFE	0.1814†	-0.0440 (-24.26%)	0.2384†	-0.0893 (-37.48%)	0.0454†	-0.0118 (-25.99%)	0.1430†	-0.0325 (-22.75%)
LES-COL	0.1053†	-0.0273 (-25.88%)	0.2834†	-0.0442 (-15.61%)	n/a		n/a	
LES-FB	0.1129†	-0.0196 (-17.36%)	0.2998†	-0.0278 (-9.29%)	n/a		n/a	
SELM	0.2002†	-0.0253 (-12.63%)	0.2691†	-0.0586 (-21.79%)	0.0443†	-0.0129 (-29.12%)	0.1315†	-0.0440 (-33.49%)
Duet	0.1797†	-0.0458 (-25.49%)	0.3213	-0.0064 (-1.99%)	0.0472†	-0.01 (-21.08%)	0.1724	-0.0031 (-1.77%)
RESS	**0.2255** (0.1326**)		**0.3277**		**0.0572**		**0.1756**	

Entity VS Words for expansion? MSc. Thesis

- Ensan, et.al " Relevance-based Entity Selection for Ad hoc Retrieval." Submitted and revision requested: **Information Processing & Management - Journal – Elsevier**
- Ensan, et.al Query expansion using pseudo relevance feedback on wikipedia. **J. Intell. Inf. Syst. 50(3)**: 455-478 (2018)

Semantic similarities by Embeddings

Feature Type		Feature Description
Word Embedding	Word	(Average/Max) Cosine similarity between pairs of vector of words in document body/title/keyword/description and vector of words that appear in the query
	Entity	(Average/Max) Cosine similarity between pairs of vector of words in the abstracts of entities that appear in document body/title/keyword/description and vectors of words in the abstracts of entities in the query
Document Embedding	Word	Cosine similarity between the vector for document body/title/keyword/description and the vector for the query
	Entity	(Average/Max) Cosine similarity between the vector for entity abstracts in the document body/title/keyword/description and the vector for entity abstracts in the query
	Chunk	(Average/Max) Cosine similarity between vectors of chunks (non-overlapping windows of size 10/30/50) from document body and the vector for the query
Baseline		LETOR 4.0 ranking datasets features

		Listwise			
		AdaRank		ListNet	
		All Queries	Hard Queries	All Queries	Hard Queries
Baseline		0.4215	0.1256	0.4564	0.1001
Word Embedding	Word (Embedding)	0.3876 (-8.03%)▽	0.1457 (15.93%)▲	0.3861 (-15.42%)▽	0.126 (25.88%)▲
	Word (Interpolation)	0.4578 (8.61%)▲	0.1529 (21.69%)▲	0.4672 (2.37%)	0.1222 (22.03%)▲
	Entity (Embedding)	0.3766 (-10.65%)▽	0.1669 (32.82%)▲	0.3845 (-15.77%)▽	0.1294 (29.27%)
	Entity (Interpolation)	0.4547 (7.9%)▲	0.1613 (28.35%)▲	0.4716 (3.33%)▲	0.1175 (17.42%)▲
Document Embedding	Word (Embedding)	0.402 (-4.61%)	0.1535 (22.15%)▲	0.4042 (-11.43%)▽	0.1364 (36.22%)▲
	Word (Interpolation)	0.4641 (10.13%)▲	0.1592 (26.72%)▲	0.4563 (-0.4%)	0.1085 (8.39%)▲
	Entity (Embedding)	0.3861 (-8.4%)	0.1584 (26.09%)	0.3924 (-14.03%)▽	0.1536 (53.43%)▲
	Entity (Interpolation)	0.4671 (10.84%)▲	0.1708 (35.92%)▲	0.4637 (1.58%)	0.1082 (8.05%)▲
	Chunk (Embedding)	0.4114 (-2.38%)	0.1719 (36.8%)▲	0.4186 (-8.69%)	0.1636 (63.42%)▲
	Chunk (Interpolation)	0.4564 (8.29%)▲	0.1673 (33.15%)▲	0.463 (1.43%)	0.1205 (20.32%)▲

Finished MSc Thesis (Learning to rank with semantic features including embeddings)

- Ensan, et.al . Neural word and entity embeddings for ad hoc retrieval. **Inf. Process. Manage. 54(4):** 657-673(2018)
- Ensan, et.al . Impact of Document Representation on Neural Ad hoc Retrieval. **CIKM 2018:** 1635-1638
- Ensan, et.al . An Empirical Study of Embedding Features in Learning to Rank. **CIKM 2017**: 2059-2062

Semantics: Entity Extraction, Semantic Annotation, Indexing

- **An Analysis of the Semantic Annotation Task on the Linked Data Cloud**. CoRR abs/1811.05549 (2018)
- The state of the art in semantic relatedness: a framework for comparison. **Knowledge Eng. Review 32:** e10 (2017)
- Semantic tagging and linking of software engineering social content. **Autom. Softw. Eng. 23(2)**: 147-190 (2016)
- Efficient indexing for semantic search. **Expert Syst. Appl**. 73: 92-114 (2017)

Social Networks

Lots to Do!

- **Mining Actionable Insights from Social Networks at WSDM 2017.** WSDM 2017

- **Foreword to the special issue on mining actionable insights from social networks.** Inf. Syst.78: 162-163 (2018)

Conclusion

- A lot of potentionals
  - Recommenders, E-commerce, Tourist
    - Aparat, Dijikala, Takhfifan, Alibaba, ….
  - Big data technologies:
    - Cloud, Map-reduce Techniques, ..
- A Lot of Open Research Question
- "In my view, success for data science professionals relies on becoming trained and able data scientists with the ability to perform data processing and computation at a massive scale. To achieve this, **professionals must invest time in ongoing education through institutions with multidisciplinary programs** that include elements from engineering, mathematical sciences, and social sciences. Converting big data into meaningful information begins with skilled professionals who are educated in all disciplines to be both data scientists and statisticians."
- *—Devavrat Shah, Professor at MIT's Department of Electrical Engineering and Computer Science*