UNIVERSITY OF TWENTE.

BIG DATA

Big Data & Education, KIVI, 13 April 2016

Djoerd Hiemstra

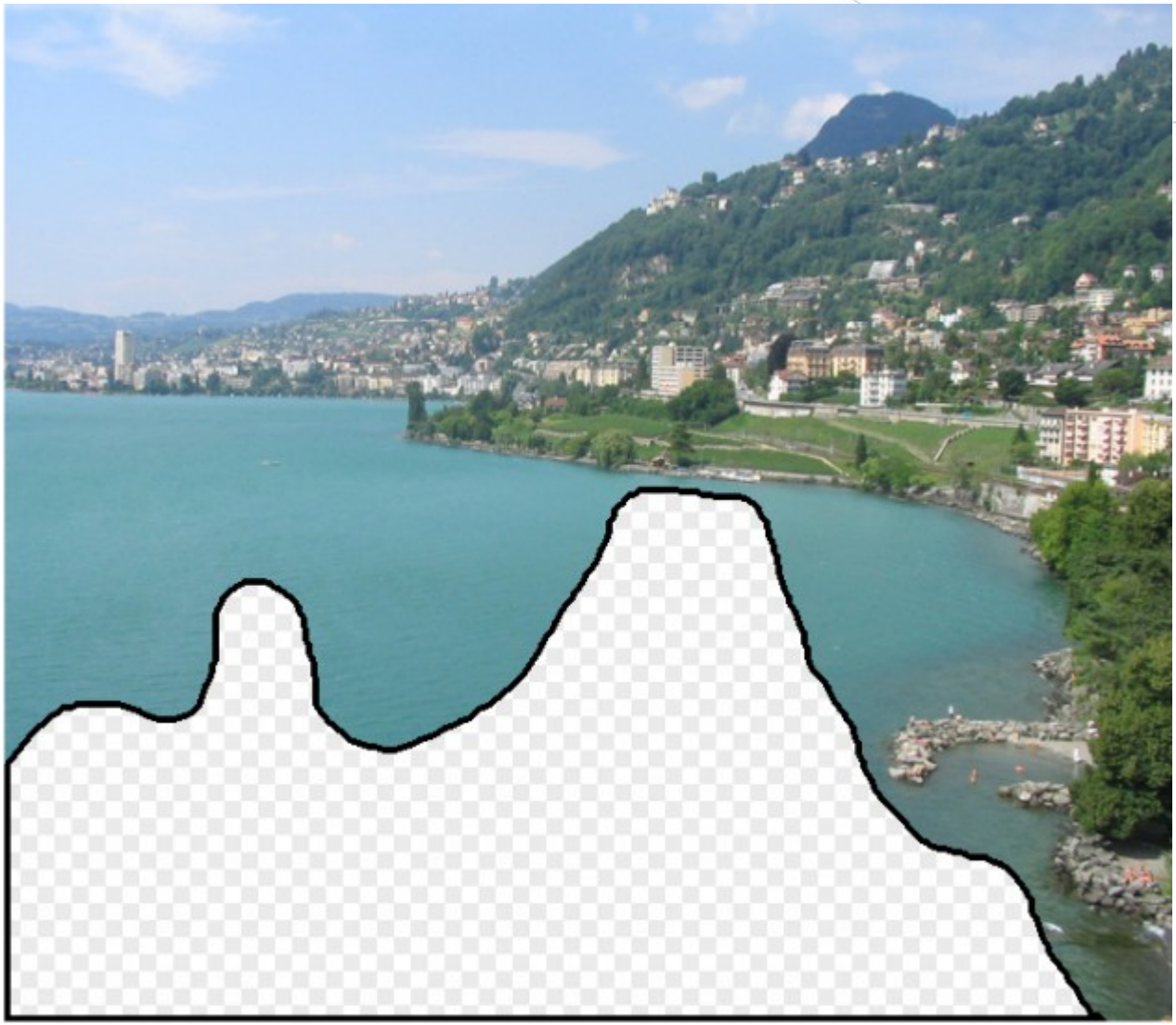
http://www.cs.utwente.nl/~hiemstra

WHY BIG DATA?

Source: http://en.wikipedia.org/wiki/Mount_Everest

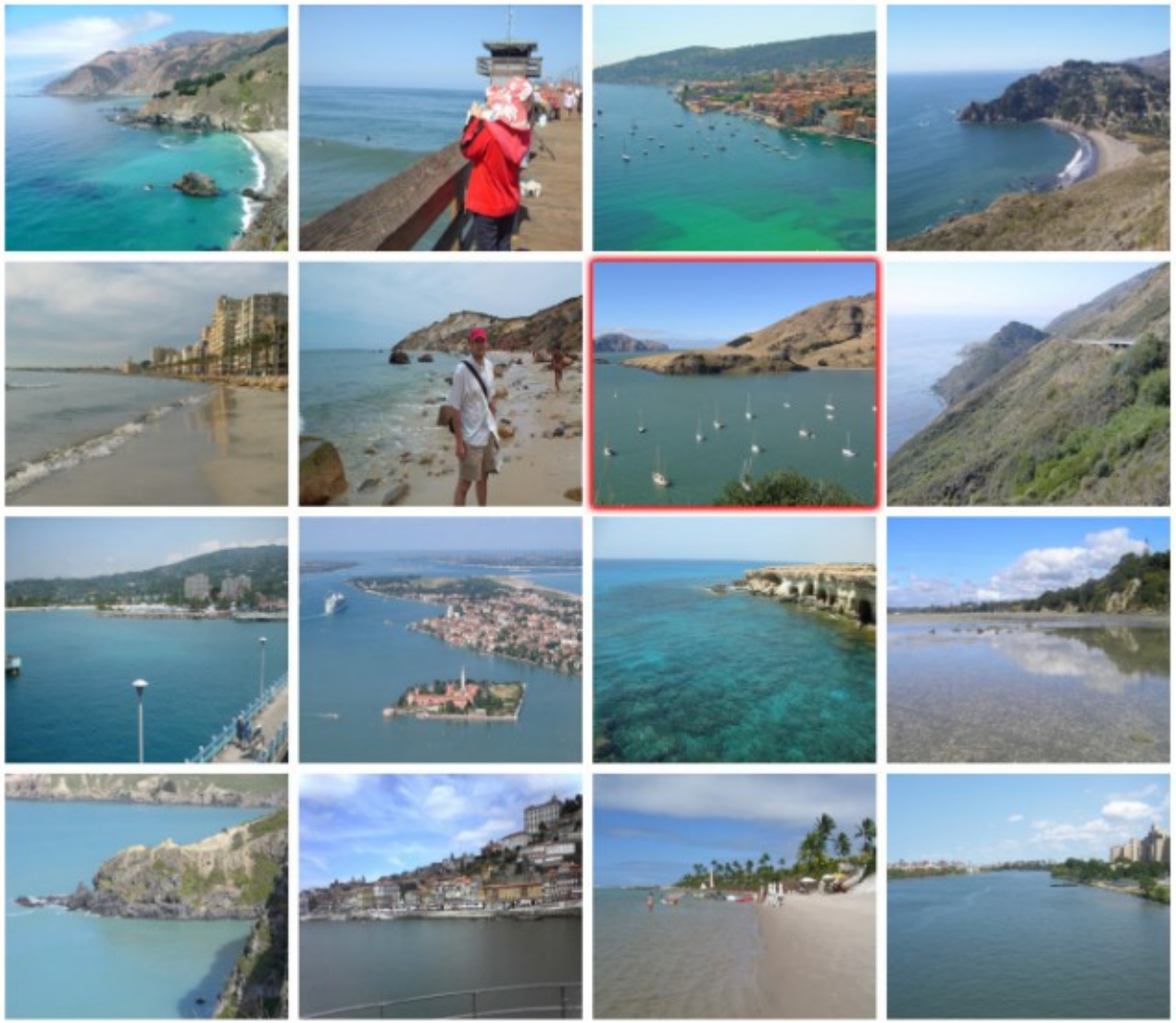

NATIONAL
GEOGRAPHIC
MAGAZINE
19 May 2012:
234 people
reach the top

James Hays and Alexei Efros. Scene Completion Using Millions of Photographs. ACM Transactions on Graphics (SIGGRAPH), 26(3), 2007.



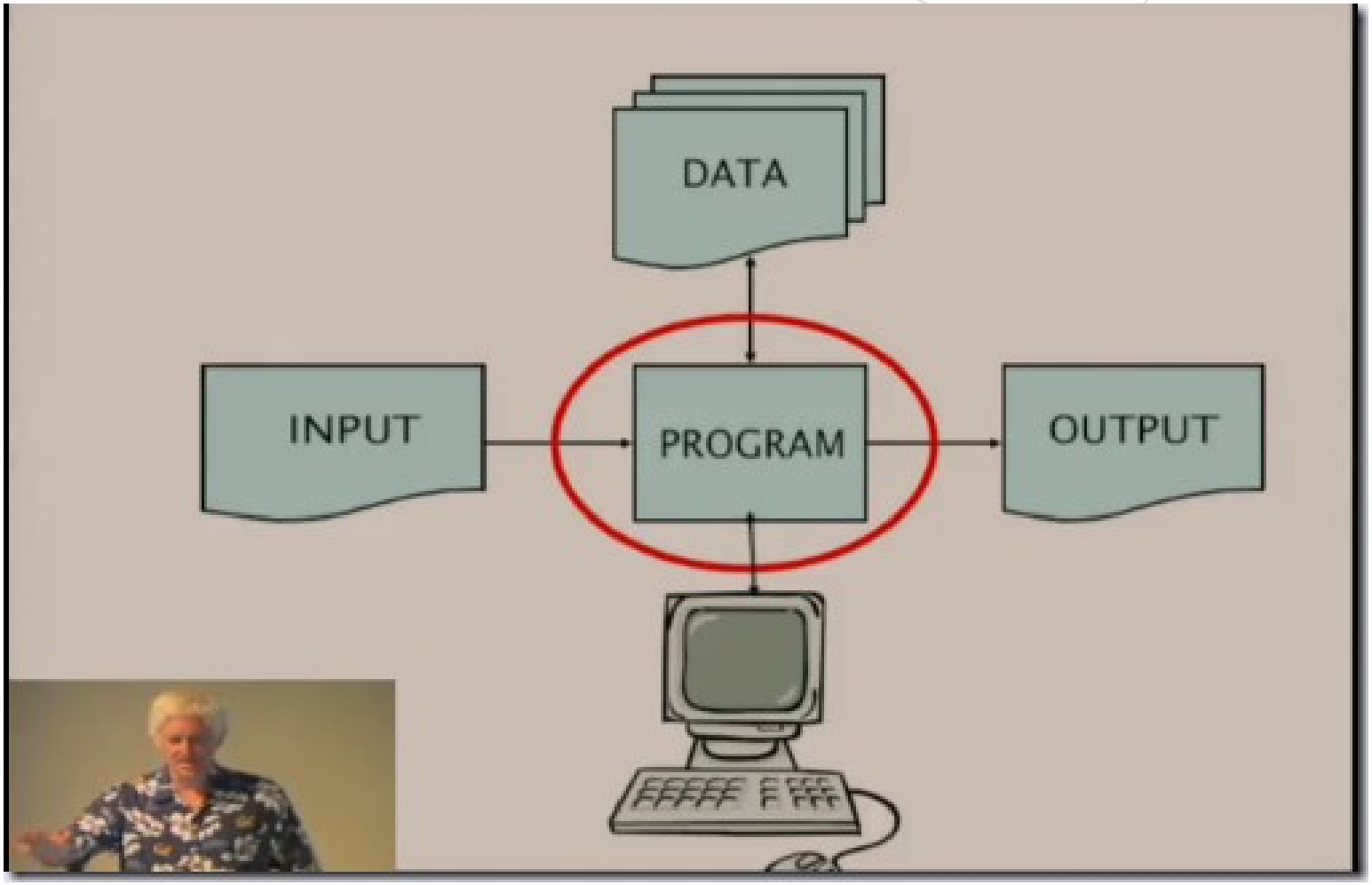
James Hays and Alexei Efros. Scene Completion Using Millions of Photographs. ACM Transactions on Graphics (SIGGRAPH), 26(3), 2007.

James Hays and Alexei Efros. Scene Completion Using Millions of Photographs. ACM Transactions on Graphics (SIGGRAPH), 26(3), 2007.



James Hays and Alexei Efros. Scene Completion Using Millions of Photographs. ACM Transactions on Graphics (SIGGRAPH), 26(3), 2007.

THE PROGRAM
VS.
THE DATA...



Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), 2009

SOME HISTORY

1980'S SPEECH RECOGNITION

"Every time I fire a linguist, the performance of the speech recognizer goes up"

(Frederick Jelinek, 1932 – 2010)

When Linguists Left the Group

Task: New Raleigh Language

- Acoustic model 1:
  - phonetic baseforms: three ←→ ?rí
  - Model statistics estimated by experts (35% accuracy)
- Acoustic model 2:
  - phonetic baseforms: three ←→ ?rí
  - Model statistics estimated automatically from data (75% accuracy)

http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf

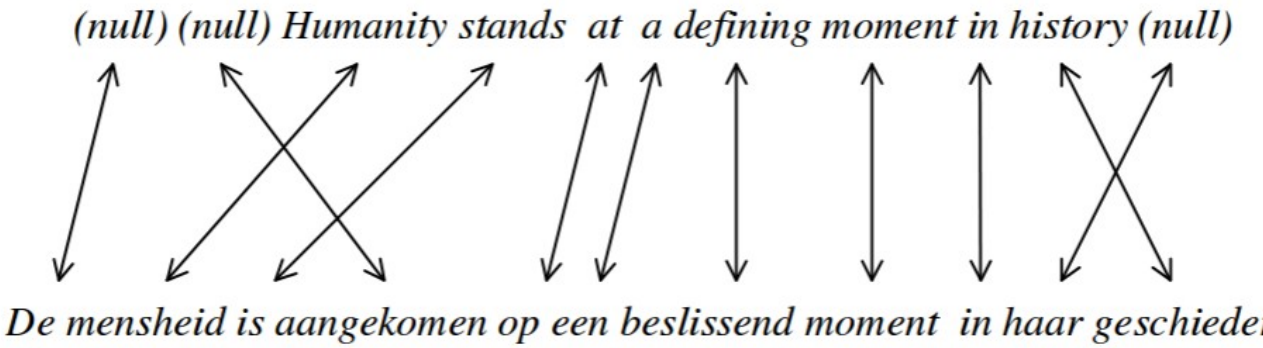
1990's: ALL NLP

"There is no data like more data"

(Robert L. Mercer, IBM in 1985)

DATA-DRIVEN MACHINE TRANSLATION

- Data-driven machine translation



First sentence of the Agenda 21 corpus.

MY MASTER THESIS

- Data-driven machine translation

local	
plaatselijke	0.51
lokale	0.24
lokaal	0.15
plaatselijk	0.09
maken	0.01

can	
kunnen	0.58
kan	0.33
dit	0.03
leveren	0.03
brede	0.01

dieren	
animal	0.50
animals	0.40
(null)	0.08
such	0.01

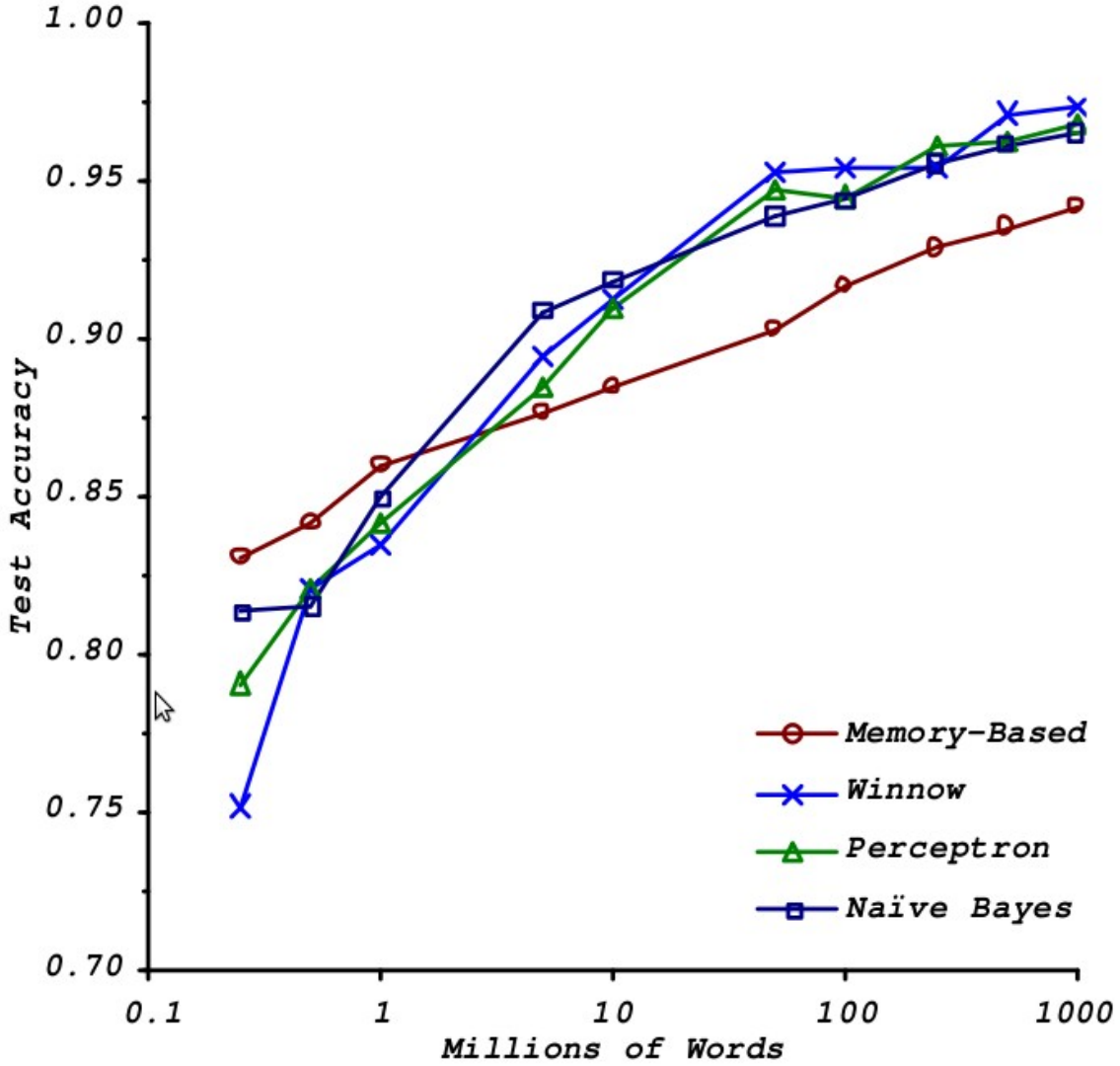
verbetering	
improving	0.31
improvement	0.28
improve	0.16
improved	0.06
enhancing	0.03
(null)	0.02

Djoerd Hiemstra, Using statistical methods to create a bilingual dictionary, Master's Thesis, University of Twente, 1996

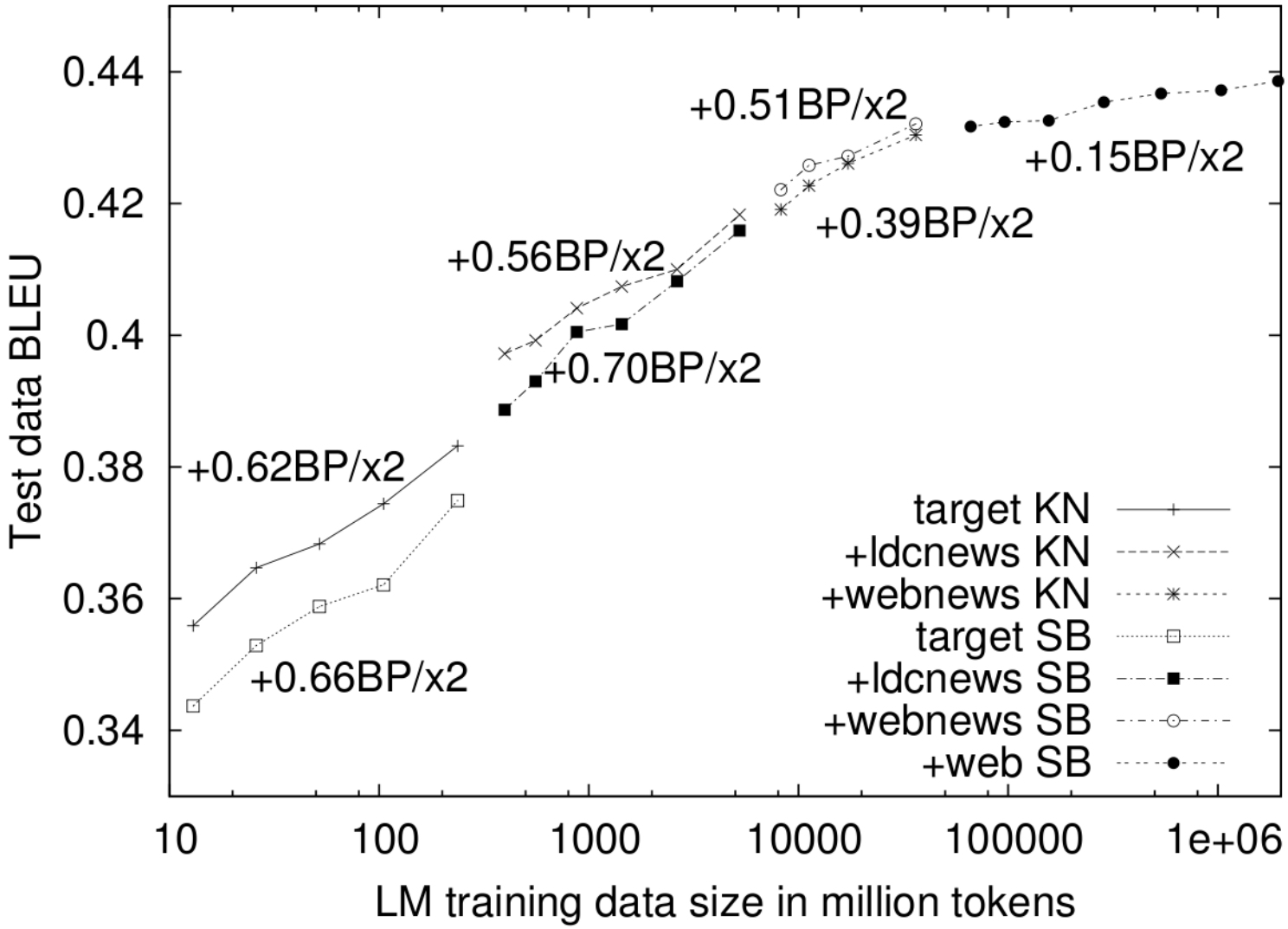
2000's DATA-DRIVEN METHODS IN PRODUCTION

"More data is more important than better algorithms"

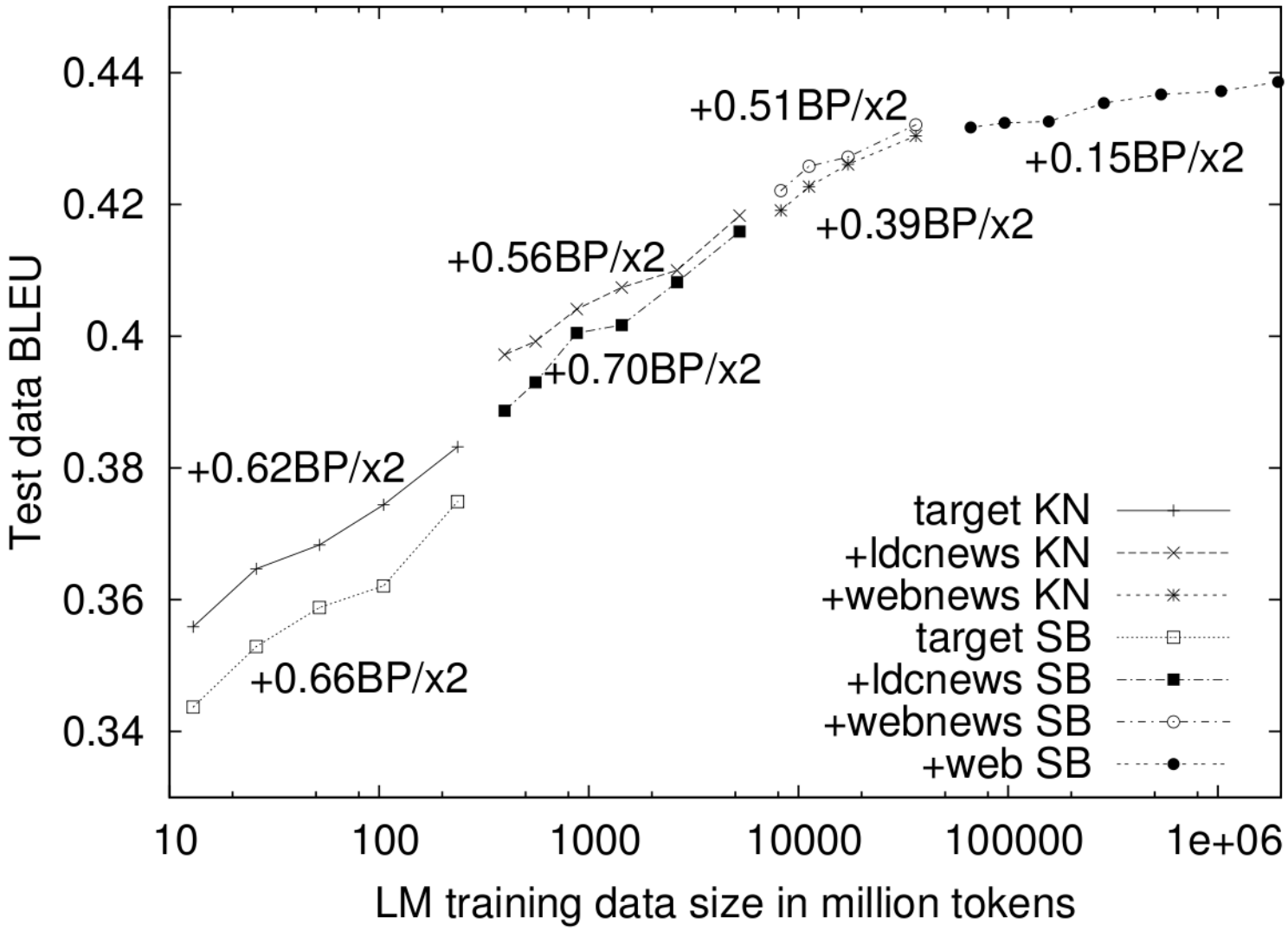
(Michele Banko & Eric Brill, Microsoft)



Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In Proceedings of the 39th Annual Meeting of the ACL, 2001.



Thorsten Brants, Ashok Popat, Peng Xu, Franz Och, Jeffrey Dean. Large Language Models in Machine Translation. In: Proceedings of EMNLP, 2007



How to get here if you are (not) Google?

Thorsten Brants, Ashok Popat, Peng Xu, Franz Och, Jeffrey Dean. Large Language Models in Machine Translation. In: Proceedings of EMNLP, 2007

HOW? TEACH A COURSE
(and get a DIY datacenter)

COURSE: MANAGING BIG DATA

- M.Sc. Course Computer Science
- First edition: Nov. 2009 – Feb. 2010
- with Maarten Fokkinga and Robin Aly

COURSE: MANAGING BIG DATA

- File systems (Google File System)
- New Storage model (BigTable)
- Programming paradigm (MapReduce)
- Programming languages (Haskell, Java,...)
- New Query languages (Yahoo Pig,...)
- Queuing and streams (Kafka, Twitter Storm, …)


YAHOO!


CASE STUDY: CLUEWEB09

- Web crawl of 1 billion pages (25 TB)
  - crawled in Jan. – Feb. 2009
  - using only the English pages (0.5 billion)

CASE STUDY: CLUEWEB09

- Rebuild Google's experimental infrastructure
  - Jeffrey Dean. Challenges in building large-scale information retrieval systems. In *WSDM* 2009
  - Using Hadoop

SEQUENTIAL SEARCH

- 50 test queries take less than 30 minutes on Anchor Text representation
- Language model, no smoothing, length prior
- Expected Precision at 5, 10 and 20 documents (MTC method):

0.42	0.39	0.35	
(0.44	0.42	0.38	U. Amsterdam)
(0.43	0.38	0.38	Microsoft Asia)
(0.42	0.40	0.39	Microsoft UK)

EXPERIMENTAL RESULTS

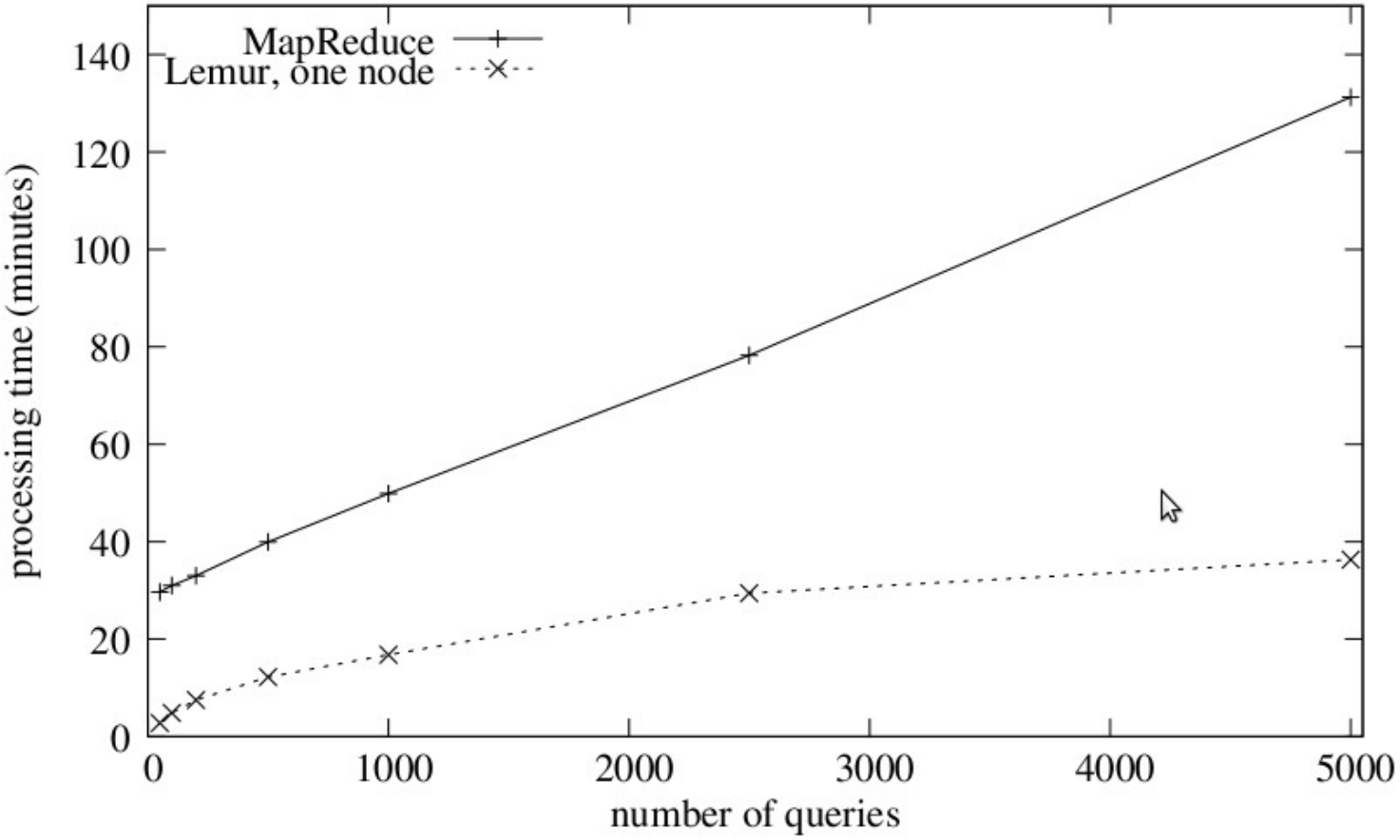


Figure 2: Processing time for query set sizes

BRUTE-FORCE MAP/REDUCE INSTEAD OF PRODUCTION SYSTEM

1. Less time coding and debugging
2. Easy to include new information that is not in the engine's standard inverted index
3. Oversee all the code in the experiment
4. Large-scale experiments in reasonable time

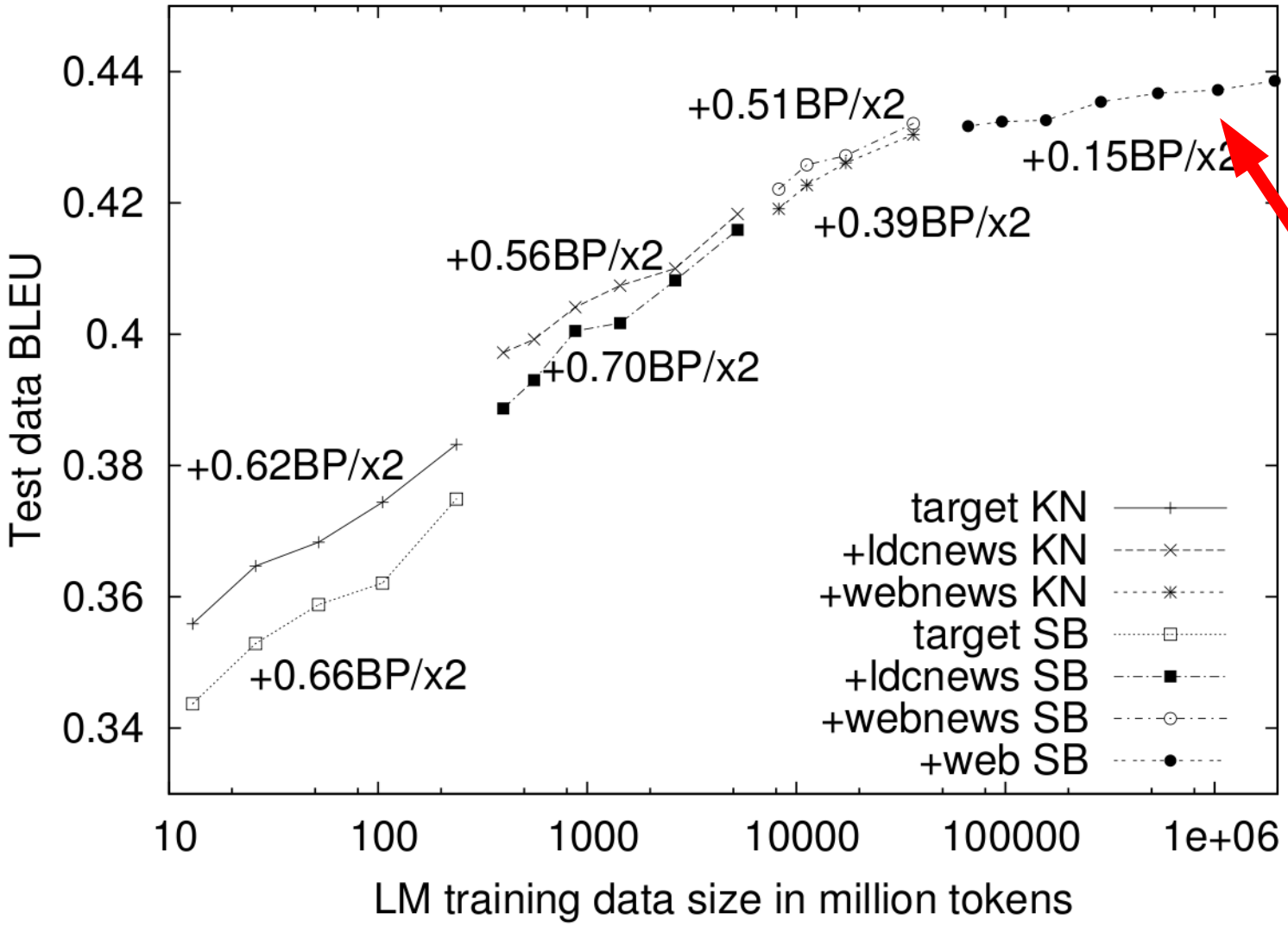
BIG DATA FOR RESEARCHERS

- Minimize coding a proof-of-concept
- Faster turnaround of the experimental cycle:
  - = more experiments on more data
  - = more improvement of search quality
  - = better system!

MORE INFO

- Djoerd Hiemstra and Claudia Hauff. MapReduce for information retrieval evaluation. In: CLEF, Multilingual and Multimodal Information Access Evaluation, pages 64-69, 2010
- Software open source at http://mirex.sourceforge.net

Wait: where about is this on Google's graph?



We were here in 2010!

YOU CAN DO
WHAT GOOGLE DOES!
(in 3 to 4 years)

ACKNOWLEDGEMENTS

- Yahoo Research

- Netherlands Organization for Scientific Research (NWO).

- COMMIT, a public-private ICT research community