Research and Education with MapReduce/Hadoop:
Data-Intensive Text Processing and Beyond

Jimmy Lin
The iSchool
University of Maryland

(with Tamer Elsayed, Chris Dyer, Philip Resnik, and Doug Oard)

Monday, October 5, 2009

Why large data?

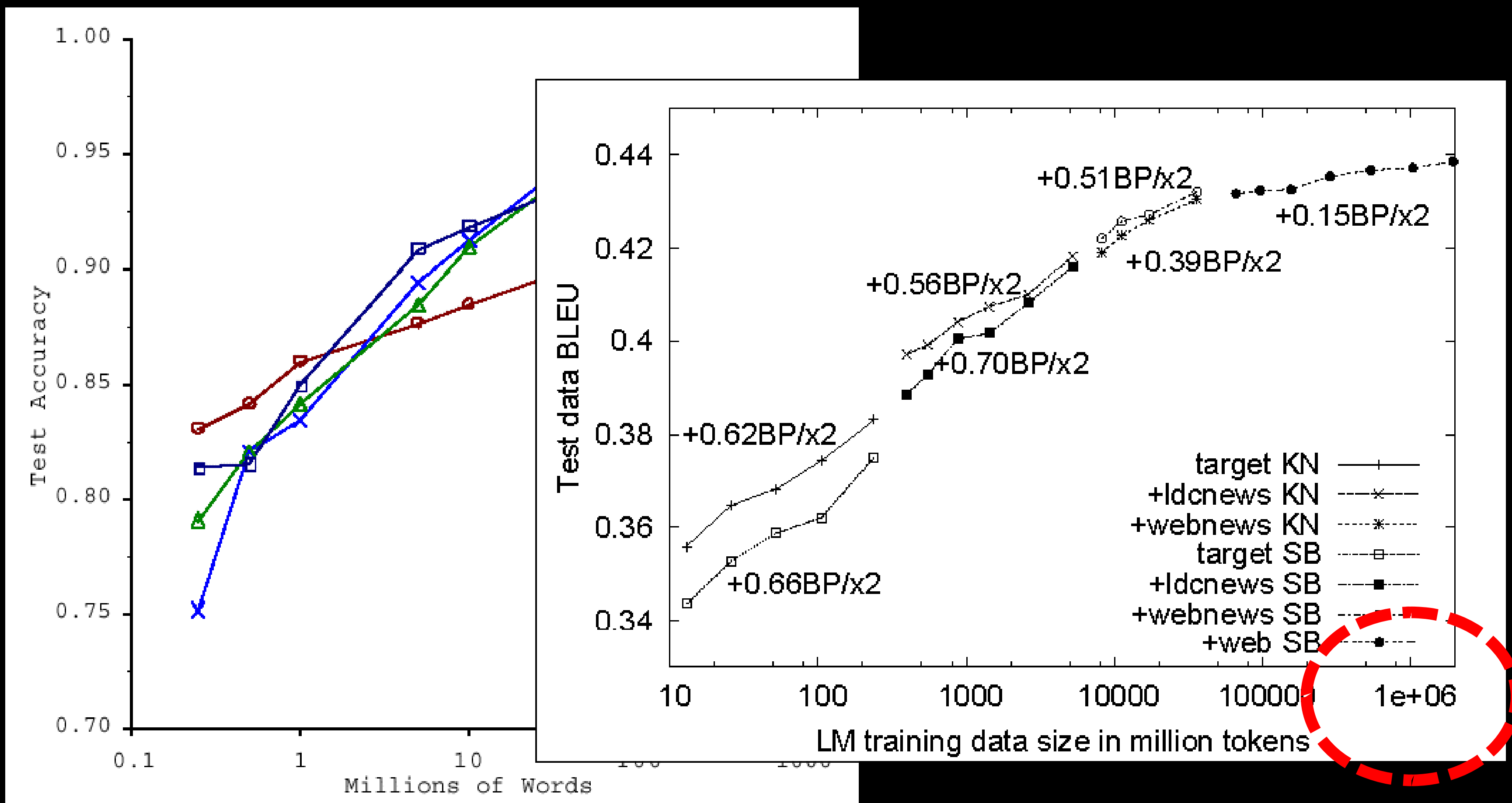
Because they're there...

How much data?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN's LHC will generate 15 PB a year (??)

Unreasonable Effectiveness of Data

by Alon Halevy, Peter Norvig, and Fernando Pereira



(Banko and Brill, ACL 2001)
(Brants et al., EMNLP 2007)

How do we move the entire field forward?
How do we educate future computer scientists?

Google/IBM ACCI
NSF CLuE

Maryland participation since Fall 2007

MapReduce

- MapReduce provides an abstraction for large-scale distributed algorithms
  - Programmer writes mappers and reducers… that's it!
- MapReduce "runtime" takes care of distributed execution
  - Handles scheduling, data distribution, synchronization, faults…
  - Seamless scaling out: 1000's of nodes+, 100's of TB+
- Vibrant open-source software ecosystem
  - Hadoop: implementation of MapReduce
  - Pig: higher-level dataflow language
  - Hive: data warehouse application
  - …

Focus on NLP/IR algorithms, not system-level details!

Cloud Computing (Spring 2008)

- Explicit goal: integration of research and education
  - Basic idea: Ph.D. students leading teams of masters and undergraduate students
  - Goal: tackle “Web-scale” research problems and generate publishable results (and they did!)
- Organization:
  - 3 week Hadoop “boot camp”; rest of the time spent on projects
  - Half a dozen teams on a variety of projections
  - Course used early configuration of the CLuE Cluster
- Follow-on course in Fall 2009

Similar efforts at U. Washington, Berkeley, etc.

Research/Education Integration
Case study #1:
DNA Sequence Alignment

Spring 2008: **Michael Schatz*** (Ph.D. student, Computer Science)
Fall 2008: **Ben Langmead*** (M.S. student, Computer Science)

*Advised by Mihai Pop and Steven Salzberg

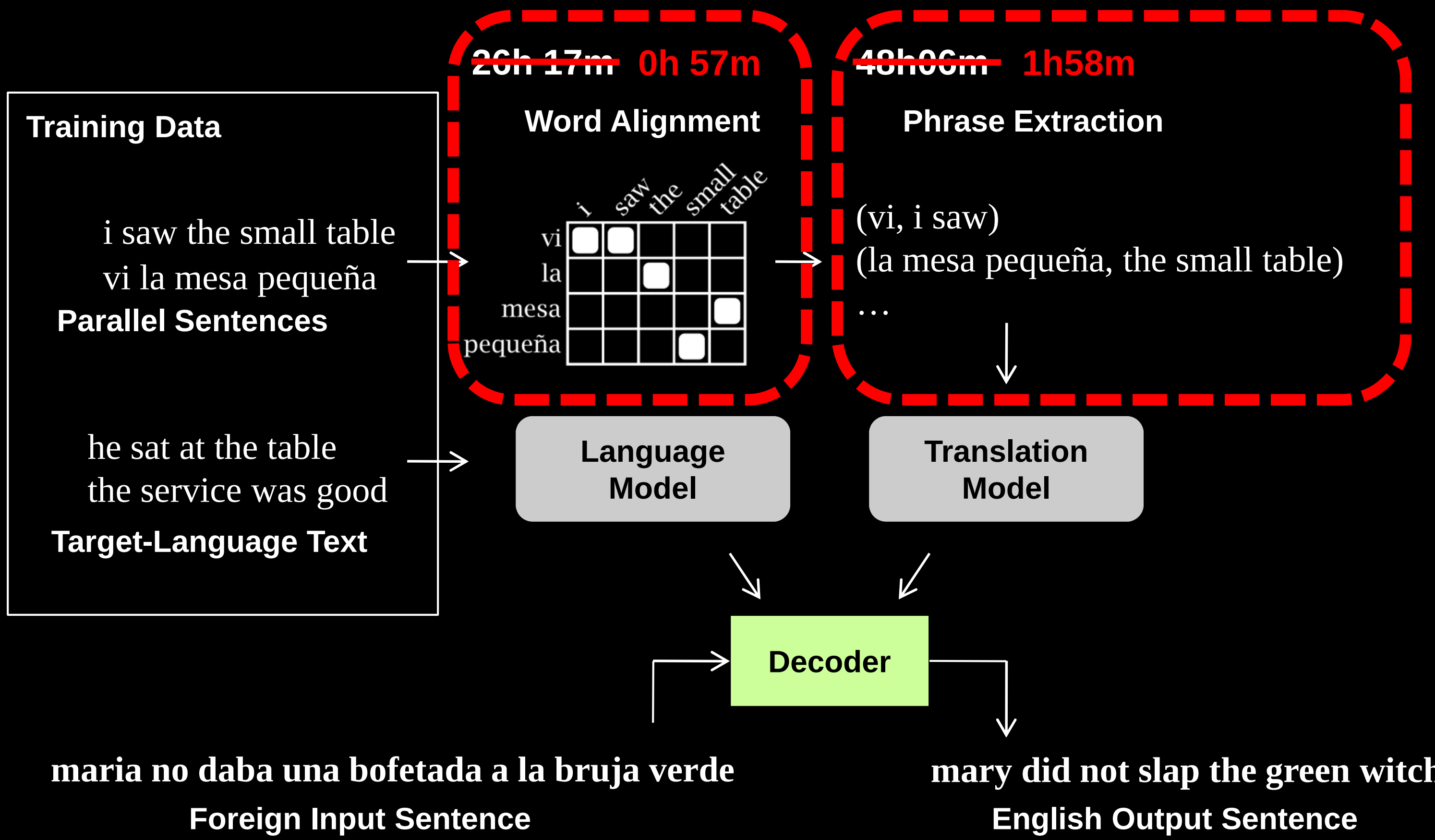
Research/Education Integration
Case study #2:
Statistical Machine Translation

Chris Dyer* (Ph.D. student, Linguistics)
Aaron Cordova (undergraduate, Computer Science)
Alex Mont (undergraduate, Computer Science)

*Advised by Philip Resnik

SMT with MapReduce

We've built MapReduce Implementations of these two components!



Chris Dyer, Aaron Cordova, Alex Mont, and Jimmy Lin. Fast, Easy, and Cheap: Construction of Statistical Machine Translation Models with MapReduce. Proceedings of the Third Workshop on Statistical Machine Translation at ACL 2008

Experiments on earlier 38-core configuration of the CLuE Cluster

Research/Education Integration
Case study #3:
Identity Resolution in Email

Tamer Elsayed * (Ph.D. student, Computer Science)
Greg Jablonski (MLS student, the iSchool)
Alan Jackoway (undergraduate, Computer Science)

*Advised by Doug Oard

Identity Resolution in Email

Date: Wed Dec 20 08:57:00 EST 2000

From: Kay Mann <kay.mann@enron.com>

To: Suzanne Adams <suzanne.adams@enron.com>

Subject: Re: GE Conference Call has be rescheduled

Who dat?

Did Sheila want Scott to participate? Looks like the call will be too late for him.

Is this a real problem?

Identity Resolution in Email

Date: Wed Dec 20

From: Kay Mann <

To: Suzanne Adar .com>

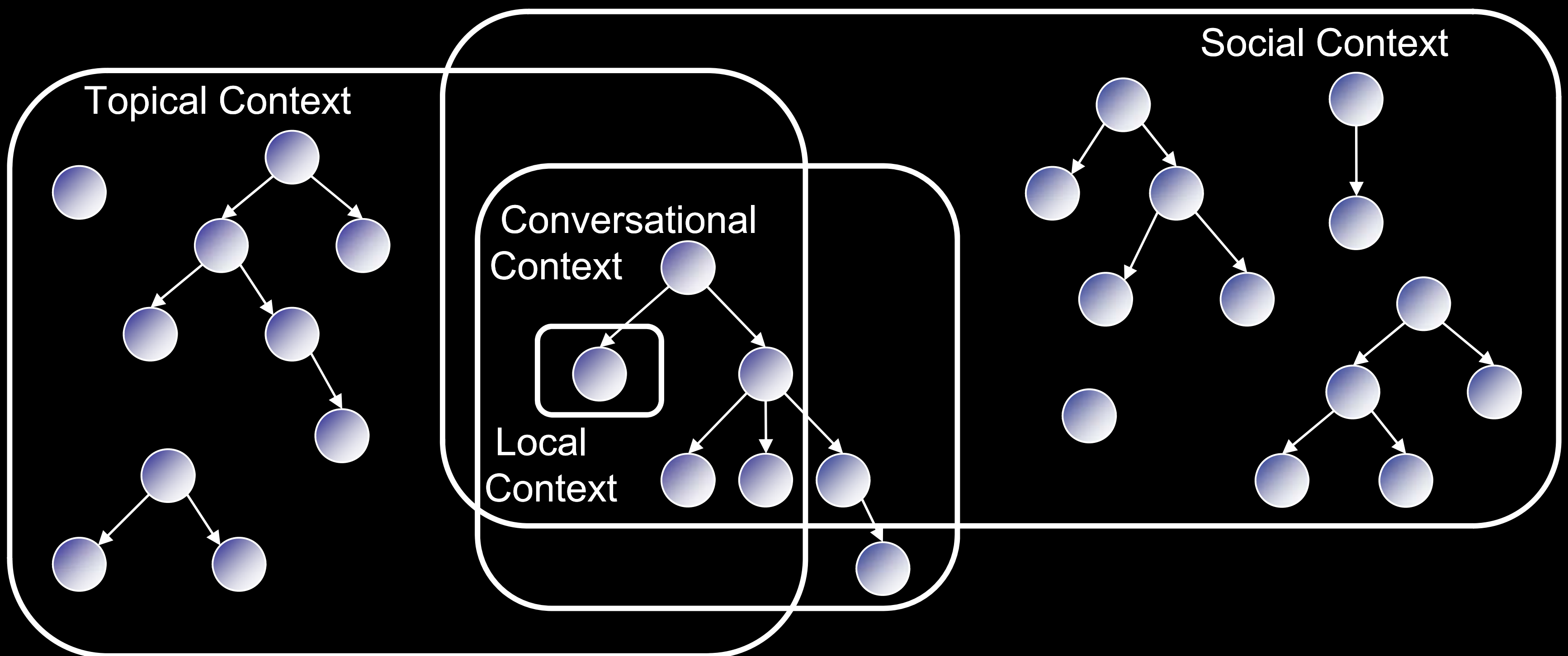
Subject: Re: GE C heduled

Did Sheila want s like the call will be too la

Sheila …

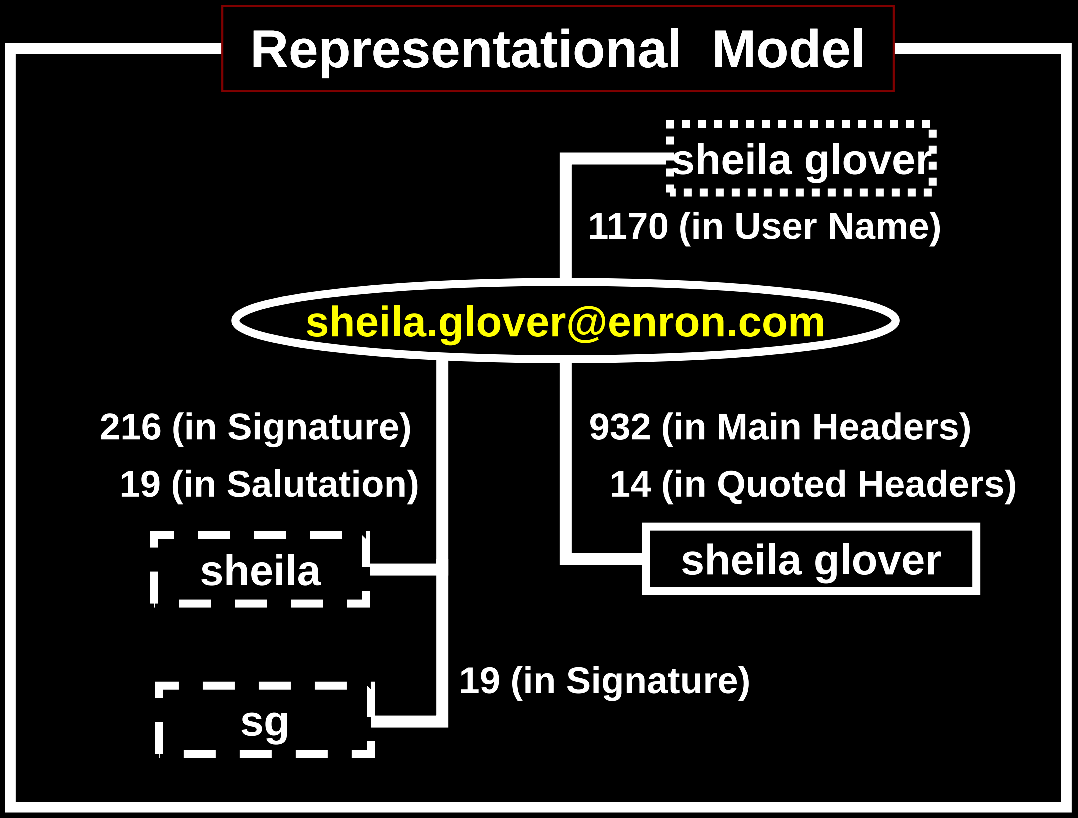
Weisman	Maynes	Jarnot
Pardo	Nacey	Kirby
Glover	Ferrarini	Knudsen
Rich	Dey	Boehringer
Jones	Macleod	Lutz
Breeden	Howard	Wollam
Huckaby	Darling	Jortner
Tweed	Watson	Neylon
Mcintyre	Perlick	Qhanger
Chadwick	Advani	Nagel
Birmingham	Hester	Graves
Kahanek	Kenner	Mclaughlin
Foraker	Lewis	Venville
Tasman	Walton	Rappazzo
Fisher	Whitman	Miller
Petitt	Berggren	Swatek
Dombo	Osowski	Hollis
Robbins	Kelly	Chang

Sources of Evidence



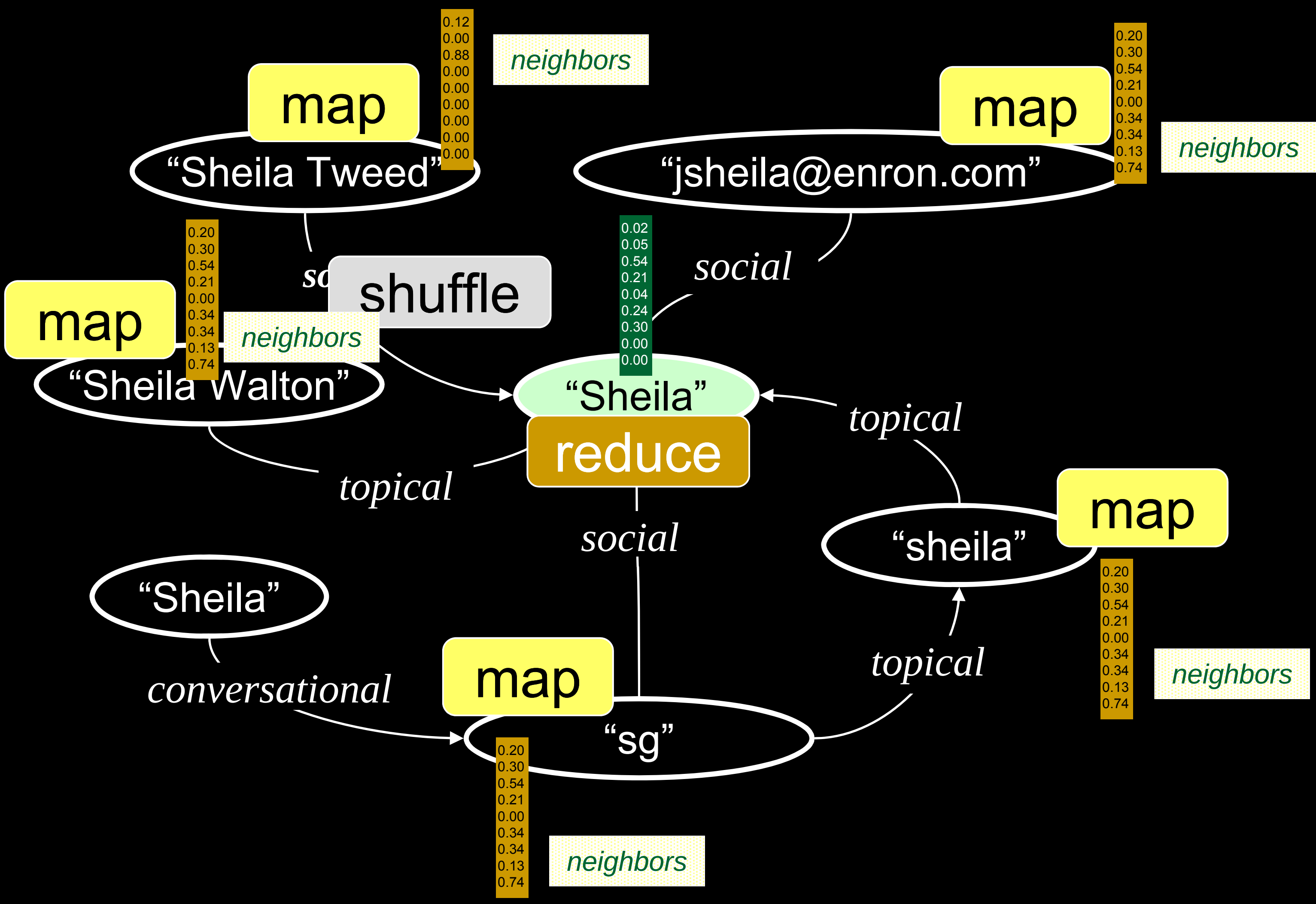
Identity Models

Identity resolution = identify correct email address



77,240 models

Mention Resolution with MapReduce



Constructing the Mention Graph

- Local and conversational context: easy
- Topical and social context: more difficult
  - Boils down to a pairwise similarity comparison problem
  - Topical context: bag of words
  - Social context: bag of participants
- We've developed an efficient MapReduce for pairwise similarity
  - Basic idea: build inverted index, cross each positing with itself
  - MapReduce does all the heavy lifting

Tamer Elsayed, Jimmy Lin, and Douglas Oard. Pairwise Document Similarity in Large Collections with MapReduce. ACL 2008, Companion Volume.

Jimmy Lin. Brute Force and Indexed Approaches to Pairwise Document Similarity Comparisons with MapReduce. SIGIR 2009.

The Future?

- "Web-scale" processing will become a necessity, not merely a luxury
  - Commoditization of cluster computing → different models of accessing cloud resources
  - Education will remain critical to ensuring progress
- We've only just begun…
  - Richer distributed programming models
  - Innovations in system architectures
  - Breakthroughs in applications
- Continued need for academic/industrial partnerships!

Questions?