DeepDive

Deep Linguistic Processing with Condor

Feng Niu, Christopher Ré, and Ce Zhang

Hazy Research Group

University of Wisconsin-Madison

http://www.cs.wisc.edu/hazy/

(see for students who did the real work)

Overview

Our research group's hypothesis:

"The next breakthrough in data analysis may not be in individual algorithms…

But may be in the ability to rapidly combine, deploy, and maintain existing algorithms."

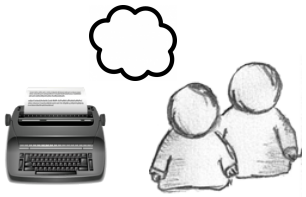
With Condor's help, use state-of-the-art NLU tools and statistical inference to read the web.

Today's talk, demos.

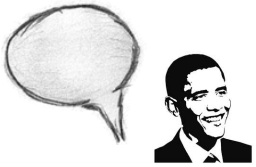
Enhance Wikipedia with the Web

***What* about Barack Obama?**

- wife is Michelle Obama
- went to Harvard Law School
- …

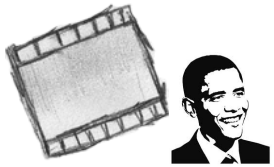


Billions of webpages

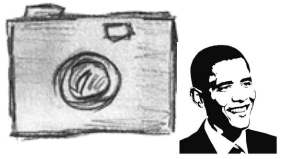


Billions of tweets

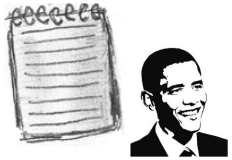
Billions of events



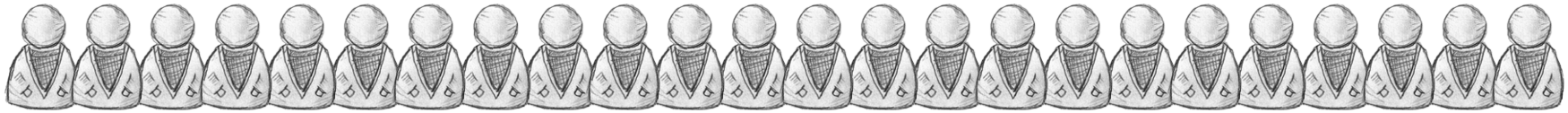
Billions of videos



Billions of photos



Billions of blogs

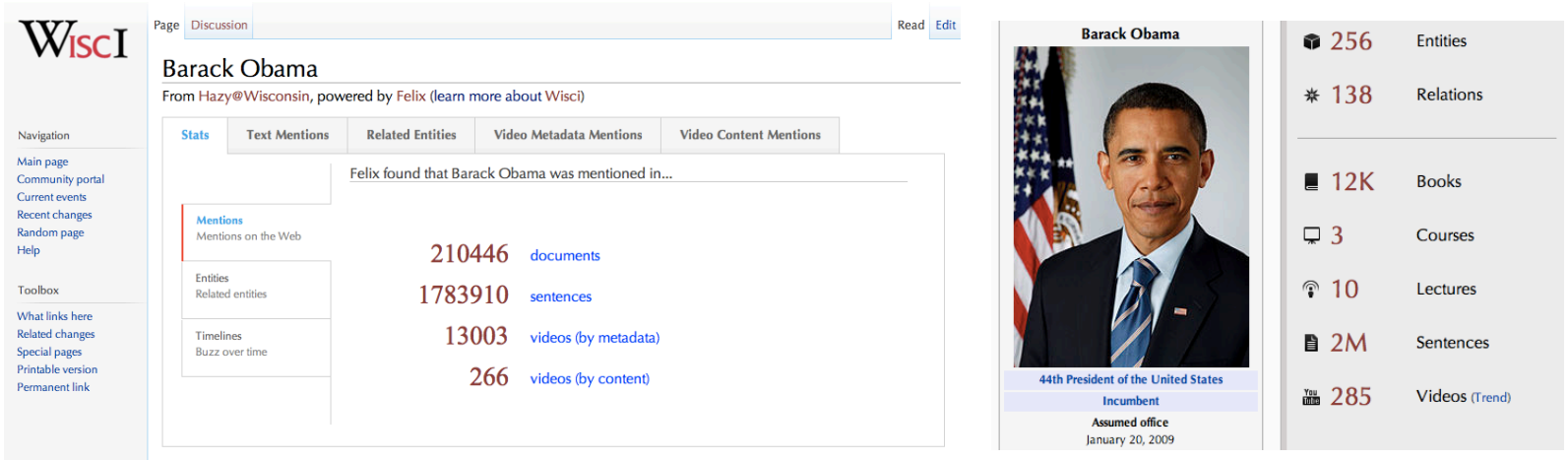


Demo

http://research.cs.wisc.edu/hazy/wisci/

Key to demo: Ability to combine and maintain
(1) structured & unstructured data and
(2) statistical tools (e.g., NLP and inference).

Some Statistics

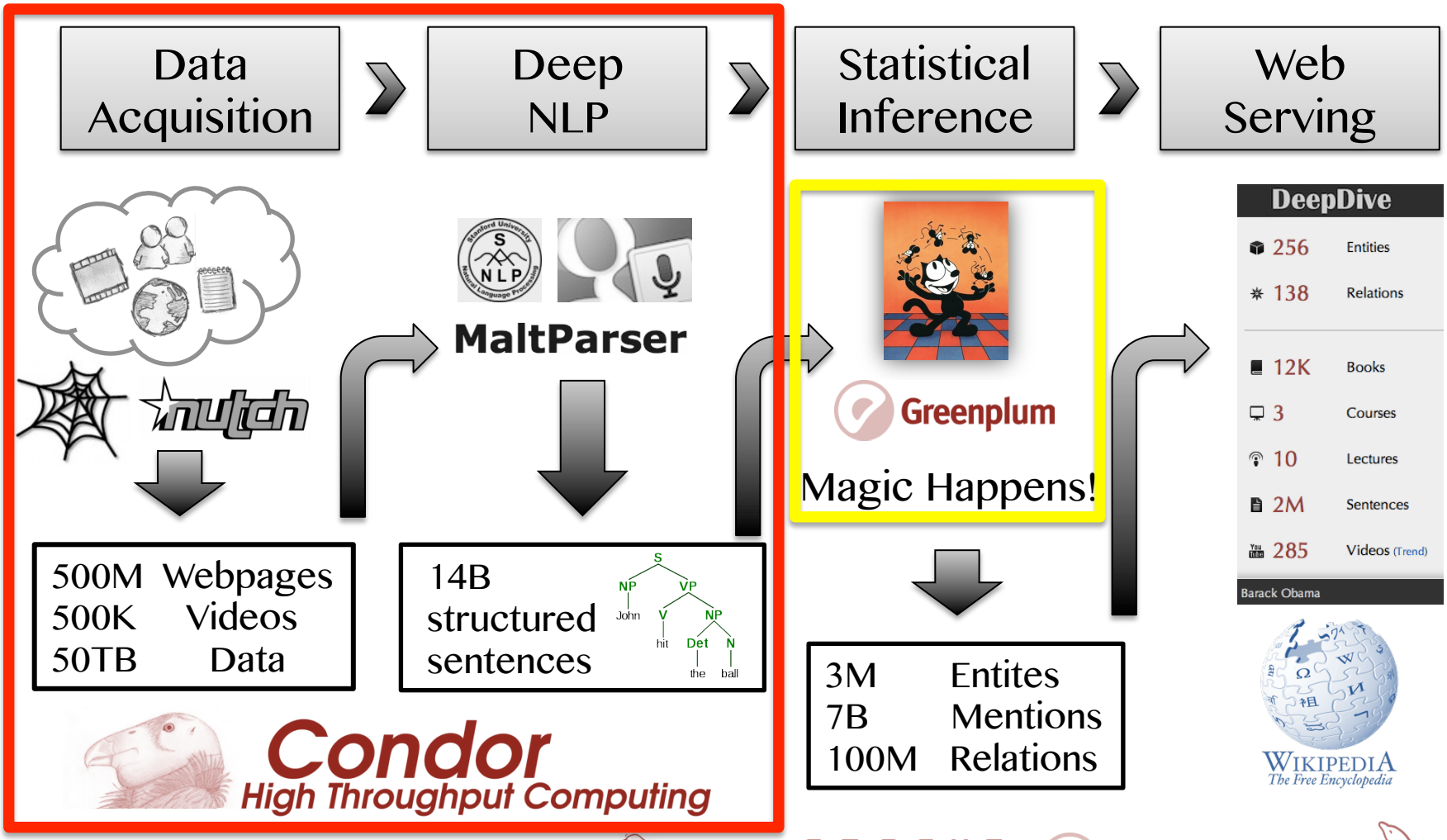


Tasks we perform:

- Web Crawling
- Information Extraction
- Deep Linguistic Processing
- Audio/Video Transcription
- Tera-byte Parallel Joins

Some Information

- **50TB** Data
- **500K** Machine hours
- **500M** Webpages
- **400K** Videos
- **20K** Books
- **7Bn** Entity Mentions
- **114M** Relationship Mentions



X 1000 @ UW-Madison

X 100K @ US Open Science Grid

100 Nodes
100 TB

X 10 High-end Servers

Raw Compute Infrastructure

Storage Infrastructure

Stats. Infrastructure

Data Acquisition with Condor

We overlay
an **ad hoc MapReduce cluster**
with **several hundred nodes**
to perform a **daily** web crawl
of **millions of web pages**

Crawl **400K** Youtube Videos,
and invoke Google's Speech API to
perform **video transcription**
in **3 days**

Deep NLP with Condor

We finish deep linguistic processing
(Stanford NLP, Coreference, POS)
on **500M web pages** (2TB text)
within **10 days**
Using **150K machine hours**

We leverage **thousands of OSG nodes**
to do deep semantic analysis
of **2TB of web pages**
within **24 hours**

High Throughput Data Processing with Condor

We run **parallel SQL join** (using Python)
over **8TB of TSV data**
with **5X higher throughput than a 100-node parallel database**

The Next Demos and Projects

A Glimpse at the Next Demos and Projects

Demo: GeoDeepDive

- Help Shanan Peters, Assoc. Prof., Geoscience, enhance a rock formation Database

1. (Number of mentions in journal articles)
2. (Number of mentions in all documents)

- Canyon Formation (2393 in 457) (2437 in 477)
- Canyon Limestone Formation (2393 in 457) (2437 in 477)
- Delaware Limestone Formation (1769 in 267) (1822 in 295)
- Williston Formation of Ocala Group (1598 in 255) (1663 in 277)
- Wilcox Formation of Wilcox Group (1453 in 310) (1463 in 315)
- Michigan Formation of Grand Rapids Group (1415 in 211) (1454 in 227)
- Dakota Formation of Cloverly Group (1359 in 286) (1392 in 296)

Condor:

- Acquire Articles
- Feature Extraction
- Measurement Extraction

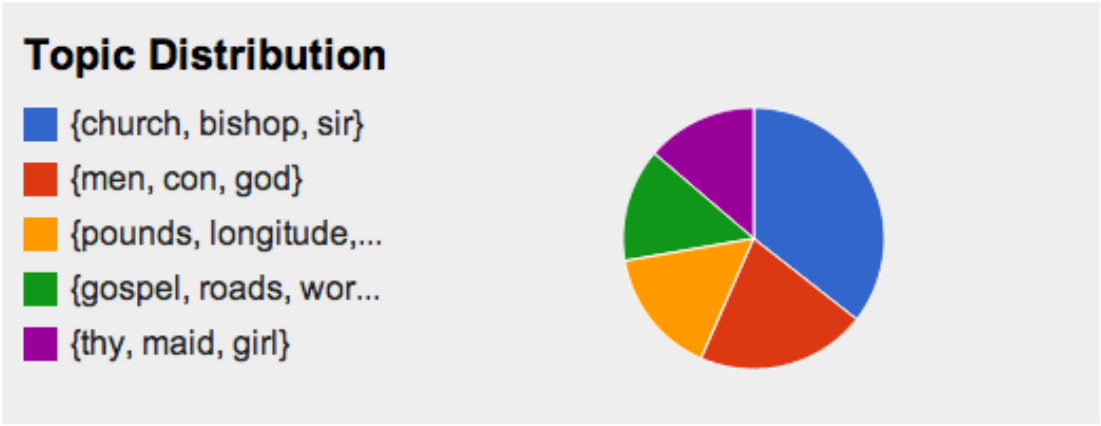
***We Hope to Answer**: What is the carbon record of North America?*

Demo: AncientText

- Help Robin Valenza, Assoc. Prof., English understand 140K books from UK 1700-1900

AncientText

[0030100700] Sir Isaac Newton's tables for renewing and purchasing the leases of cathedral-churches and colleges, according to the several rates of interest: ... To which is added, the value of church and college leases consider'd, ... By a late Bishop of Chichester. (by Newton, Isaac, Sir, 1642-1727)

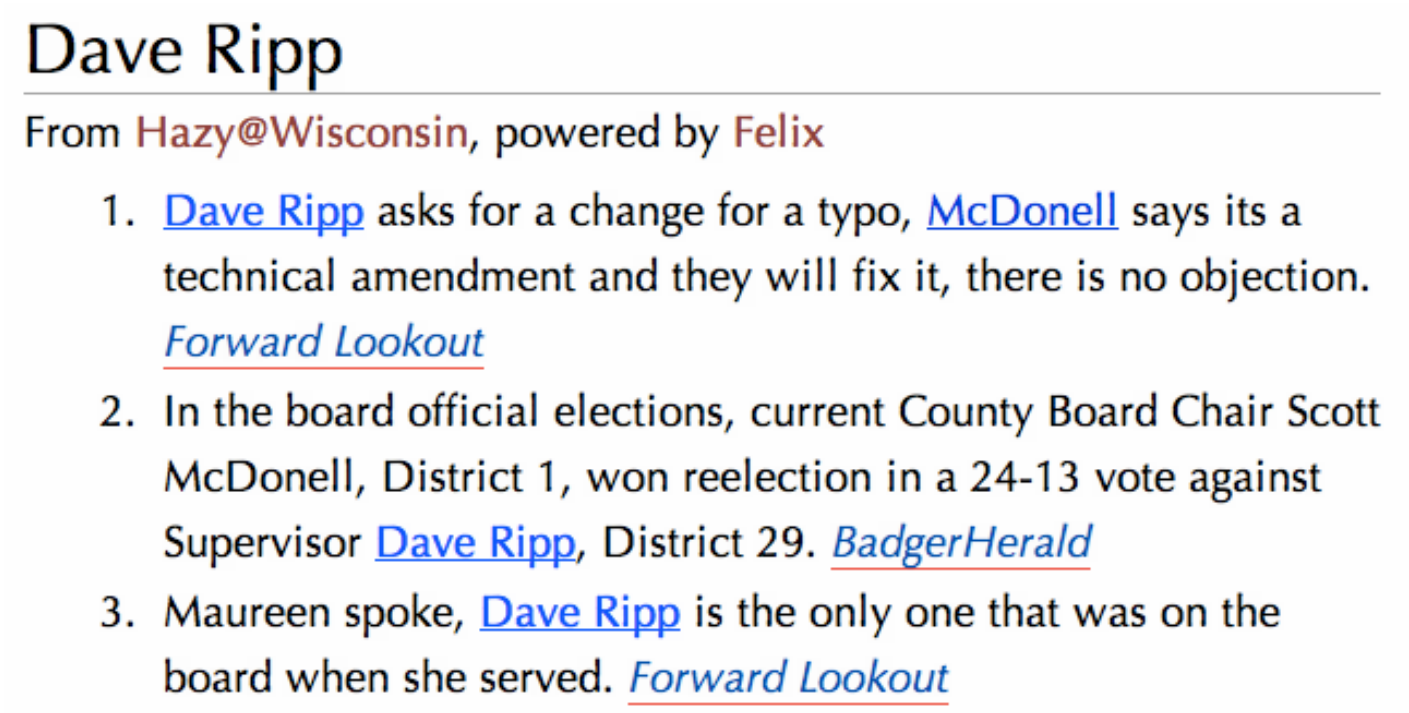


Condor Helps:

- Building Topic Models
  - Slice and Dice!
  - By Year, Author, …
- Advanced OCR
  - **Challenge** how many alternatives to store?

Demo: MadWiki

- Machine-powered Wiki on Madison people with Erik Paulsen, Computer Sciences.



Dave Ripp

From Hazy@Wisconsin, powered by Felix

1. Dave Ripp asks for a change for a typo, McDonell says its a technical amendment and they will fix it, there is no objection. *Forward Lookout*
2. In the board official elections, current County Board Chair Scott McDonell, District 1, won reelection in a 24-13 vote against Supervisor Dave Ripp, District 29. *BadgerHerald*
3. Maureen spoke, Dave Ripp is the only one that was on the board when she served. *Forward Lookout*

Conclusion

- Condor is the *key enabling tech* across a large number of our projects
  - Crawling, Feature Extraction, and Data Processing, …. and even Statistical Inference

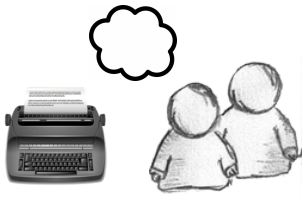
- We started with a Hadoop-based infrastructure but are gradually killing it off.

Thank you to Condor and CHTC!
Miron, Bill, Brooklin, Ken, Todd, Zach,
and the Condor and CHTC Teams

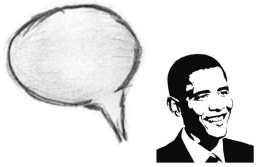
Idea: Machine-Curated Wikipedia

***What* about Barack Obama?**

- wife is Michelle Obama
- went to Harvard Law School
- …

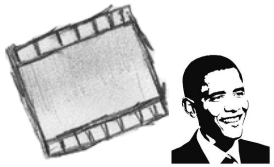


Billions of webpages

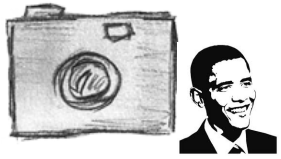


Billions of tweets

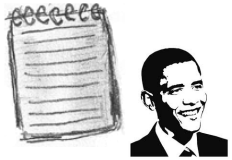
Billions of events



Billions of videos



Billions of photos



Billions of blogs