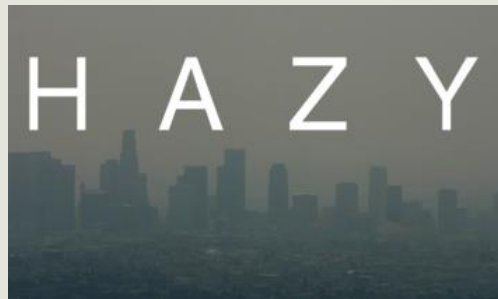


DeepDive

Christopher Ré

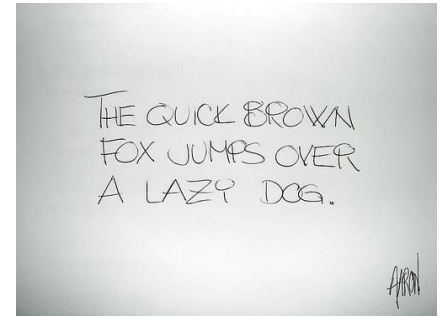
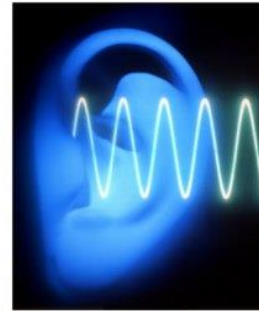
<http://hazy.cs.wisc.edu> and [youtube.com/HazyResearch](https://www.youtube.com/HazyResearch)

University of Wisconsin-Madison



DeepDive's Motivation

1. **Valuable data** in a wide range of formats



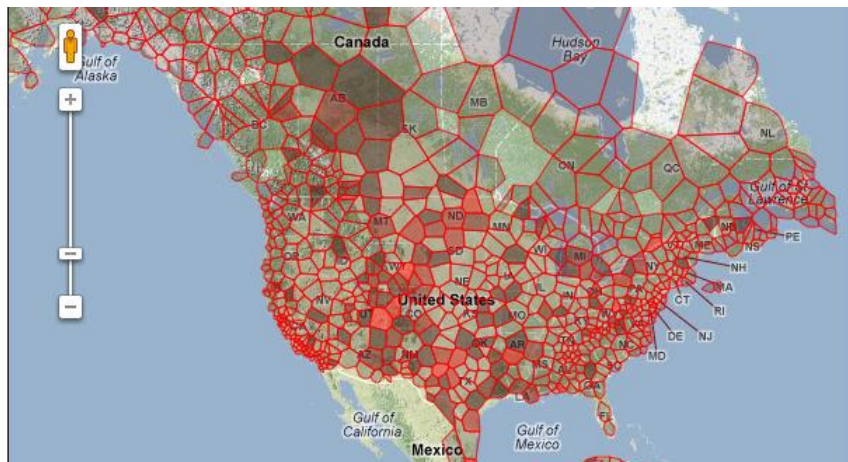
2. **Arms race** to more deeply understand data

DeepDive is a framework to help
(1) **acquire** data, (2) make **inferences**, and
(3) **maintain** it over time

Statistical reasoning used in all three stages
(consider all alternatives is expensive!)



GeoDeepDive



A macro view of
North America

Facts: How much carbon
in North America?

Inference-based Query:
Recoverable Shale?

The Atlantic

MAY 2013

What If We Never Run Out of Oil?

New technology and a little-known energy source suggest that fossil fuels ma

Key Challenge: High-enough-quality inference
for **science** ($p > 95\%$)



IBM's DeepQA



Microsoft's
Entity Cube



Distributed
Bayesian
Learner



MPI's Yago



UW's
TextRunner



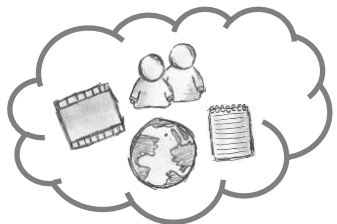
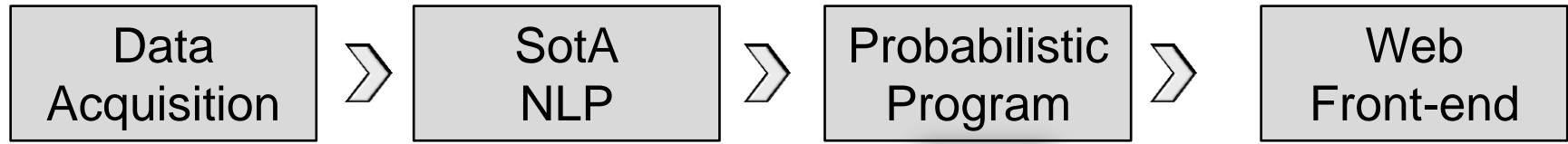
CMU's NELL

The DeepDive Framework

Applications of our DeepDive Framework

1. **DeepDive**: Web-scale KB
2. **GeoDeepDive**: Geology
3. **PaleoDeepDive**: Crawling the Fossil Record
4. **AncientBooks**: English literature (Valenza, UW English)

The DeepDive Framework (#s from GeoDeepDive)



MaltParser



Magic Happens!

GEODEEPDIVE

- 36822 Units
- 9049 Strats
- 1593 Columns
- 466 Intervals
- 122149 Documents

~ 4M CPU Hours (~450 years)

~1M Docs (1TB)
(122k curated)

240M sentences

400M Mention.
5M Measurements.

Acquisition:
VLDB 12, PODS 10

Distant Supervision:
ACL 2012

Inference:
SIGMOD13, VLDB11a

Maintenance:
ICDE12, VLDB11b



X 1000 @ UW-Madison
X 100K @ US Open
Science Grid

200 Nodes
250 TB

X 2 High-end Servers

Raw Compute Infrastructure

Storage Infrastructure

Inference Infrastructure

Backend: WebDB10, VLDB11c, PODS 12

DeepDive Scale (Web-KB)

WISCI Page Discussion Read Edit

Barack Obama

From Hazy@Wisconsin, powered by Felix (learn more about Wisci)

Stats Text Mentions Related Entities Video Metadata Mentions Video Content Mentions

Felix found that Barack Obama was mentioned in...

Mentions	210446	documents
Entities	1783910	sentences
Related entities	13003	videos (by metadata)
Timelines	266	videos (by content)
Buzz over time		

Barack Obama

44th President of the United States
Incumbent
Assumed office
January 20, 2009

256	Entities
138	Relations
12K	Books
3	Courses
10	Lectures
2M	Sentences
285	Videos (Trend)

Tasks

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




Usage Statistics

- **50TB** Data
- **1Bn** Webpages
- **400K** Videos
- **20K** Books
- **7Bn** Entity Mentions
- **114M** Relationship Mentions

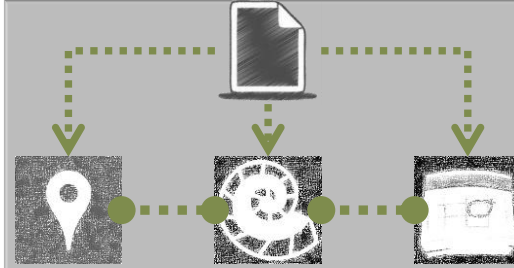
graphical models at Web scale in a tiny lab.

DeepDive Quality (PaleoDeepDive)

Paleobiology?

-  Biodiversity crisis
-  Climate change?
-  rate of evolution?

PaleoDeepDive



Input: Geology journal articles, books, museum catalogs...

Output:
Fossils, location, geologic time

Construction Comparison

176 Papers
(16 in Science
or Nature)

PaleoDB

Human-built



329 scientists



13 years



46K docs



126K fossils

PaleoDeepDive

Machine-built



2000 machines



78 machine years



300K docs



3M fossils

A couple of
weeks!

$P > 0.9$ PaleoDB as ground truth. Often better!

Concluding Thoughts

DeepDive helps with acquisition, inference, and maintenance of data.

Condor is *invaluable* for the acquisition phase—and potentially inference!

Code, data, VMs, papers, and videos!
hazy.cs.wisc.edu