DeepDive

Christopher Ré
http://hazy.cs.wisc.edu and 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!)
**Facts:** How much carbon in North America?

**Inference-based Query:** Recoverable Shale?

**A macro view of North America**

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

Joint work with Ce Zhang and others and Jude Shavlik
The DeepDive Framework
(#s from GeoDeepDive)

Data Acquisition

SotA NLP

Probabilistic Program

Web Front-end

~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

Magic Happens!

~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

~ 4M CPU Hours (~450 years)

Condor
High Throughput Computing

Acquisition:
VLDB 12, PODS 10

Distant Supervision:
ACL 2012

Inference:
SIGMOD13, VLDB11a

Maintenance:
ICDE12, VLDB11b

Raw Compute Infrastructure

X 1000 @ UW-Madison

X 100K @ US Open

Science Grid

Storage Infrastructure

Inference Infrastructure

200 Nodes
250 TB

X 2 High-end Servers

Backend: WebDB10, VLDB11c, PODS 12
DeepDive Scale (Web-KB)

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)

<table>
<thead>
<tr>
<th>Paleobiology?</th>
<th>PaleoDeepDive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biodiversity crisis?</td>
<td><strong>Input</strong>: Geology journal articles, books, museum catalogs…</td>
</tr>
<tr>
<td>Climate change?</td>
<td><strong>Output</strong>: Fossils, location, geologic time</td>
</tr>
<tr>
<td>rate of evolution?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Construction</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PaleoDB</strong></td>
<td><strong>PaleoDeepDive</strong></td>
</tr>
<tr>
<td><em>Human</em>-built</td>
<td><em>Machine</em>-built</td>
</tr>
<tr>
<td>329 scientists</td>
<td>2000 machines</td>
</tr>
<tr>
<td>13 years</td>
<td>78 machine years</td>
</tr>
<tr>
<td>46K docs</td>
<td>300K docs</td>
</tr>
<tr>
<td>126K fossils</td>
<td>3M fossils</td>
</tr>
</tbody>
</table>

P>0.9 PaleoDB as ground truth. Often better!

176 Papers (16 in Science or Nature)

A couple of weeks!
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