Using Scalable Inference to Build DeepDive: A Declarative Dark Data System.

Christopher Ré
Stanford University
Computer Science

DeepDive

Dark Data System: ETL on Steroids

Quality that can exceed paid human annotators and volunteers

DeepDive

Extraction, Integration, & Cleaning are ***inference problems***

Focus on what matters:

“Amdahl’s law” for quality

DeepDive

Declarative Inference:

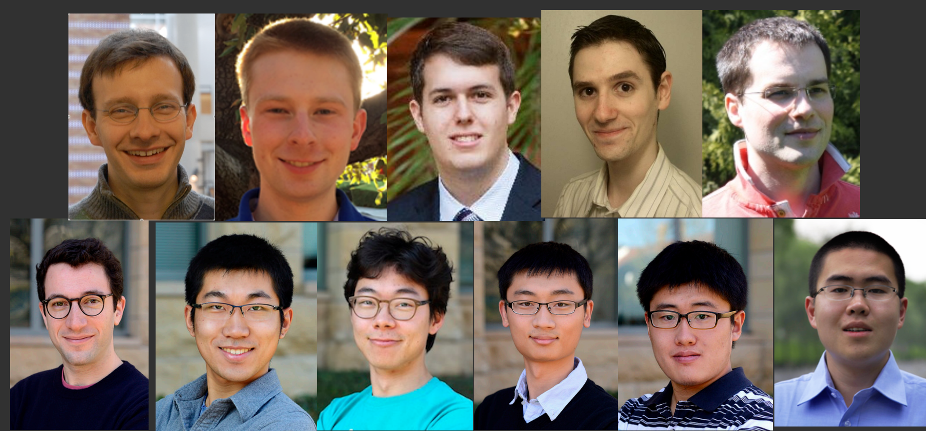
*Think about **features** not **algorithms.***

Enables non-CS users, but

scale is a challenge.

First, some **thank yous**

DeepDive



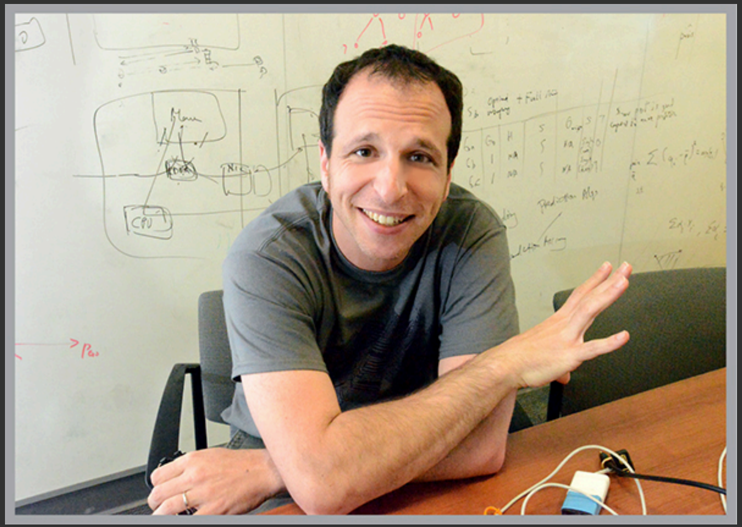
The DeepDive Team

http://deepdive.stanford.edu/

INCOMPLETE THANK YOU

Special Thank You to my optimization advisors

Stephen J. Wright
Wisconsin.
God of Optimization.



Ben Recht
Berkeley.
Patriots Fan.

My Actual Advisor

My Actual Boss

Many Existing Systems with the similar goal

Unstructured Input → **Structured Knowledge Base**

Many Amazing Industrial Systems

IBM Research

Google

An Incomplete Set of Awesome Research

- TextRunner & ReVerb (Washington)
- NELL (CMU)
- Knowledge Vault (Google)
- ProbKB (Florida)
- Lixto
- SystemT (IBM)
- YAGO-NAGA, SOFIE (MPI)
- DBLife, xLog (Wisconsin)
- StatSnowBall (Tshinghua & MSRA)

Many more

Back to our

regularly scheduled programming…

The world's scientific knowledge is **accessible**, but not **readable**.

Today, some pressing problems require **macroscopic knowledge**

Climate & Biodiversity

Health

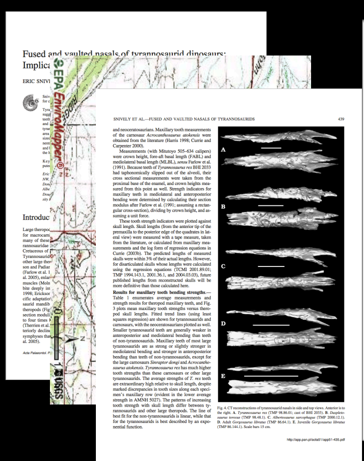
Financial Markets

Could we build a machine to **read** for us?

PaleoDeepDive

The Goal

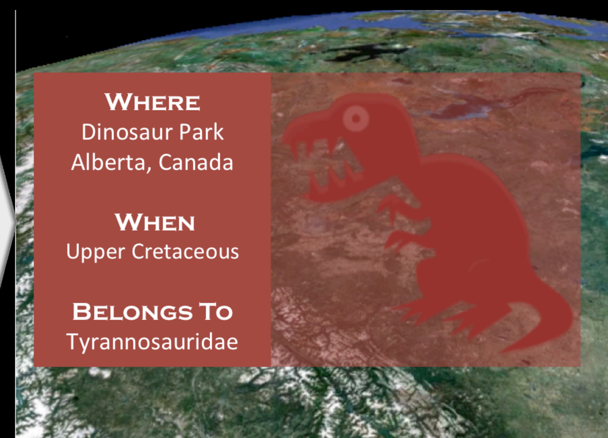
Extract paleobiological facts to build **high coverage** fossil record.



T. Rex are found dating to the upper Cretaceous.

Statistical Inference

Appears("T. Rex", "Cretaceous")



Aggressive Approach

Every character, word, part of speech is a variable

Statistical inference *on billions of variables.*

Data are buried in tables, but not in a self-contained way

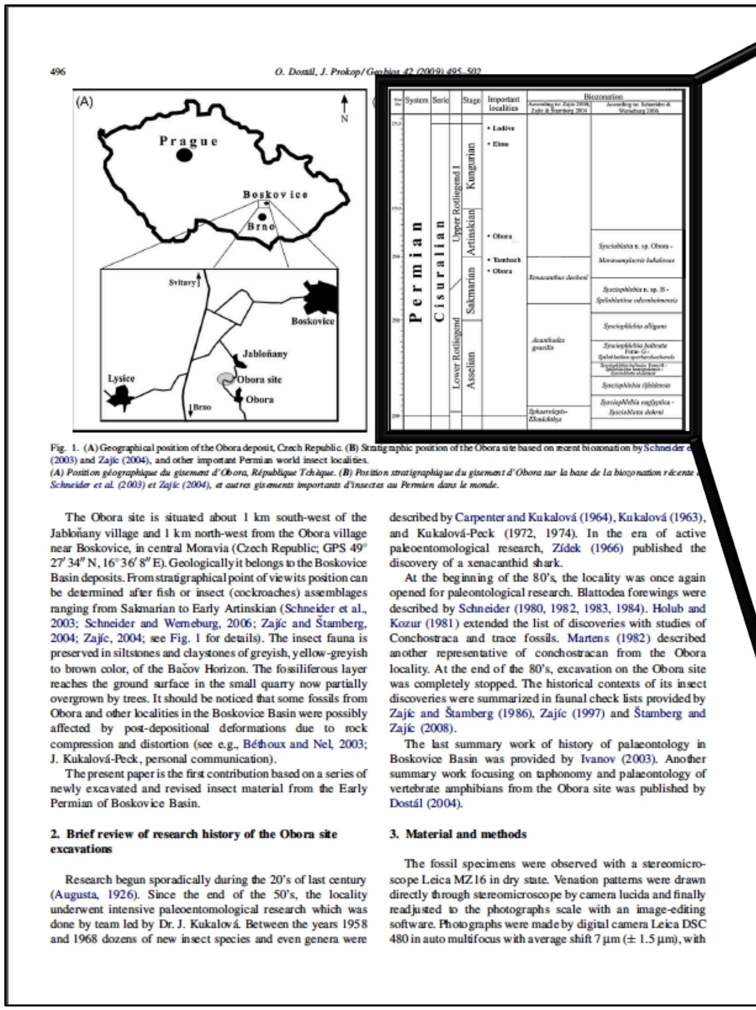


Fig. 1. (A) Geographical position of the Obora deposit, Czech Republic. (B) Stratigraphic position of the Obora site based on recent biozonation by Schneider e (2003) and Zajíc (2004), and other important Permian world insect localities.
(A) Position géographique du gisement d'Obora, République Tchèque. (B) Position stratigraphique du gisement d'Obora sur la base de la biozonation récente Schneider et al. (2003) et Zajíc (2004), et autres gisements importants d'insectes au Permien dans le monde.

The Obora site is situated about 1 km south-west of the Jabloňany village and 1 km north-west from the Obora village near Boskovice, in central Moravia (Czech Republic; GPS 49° 27′ 34″ N, 16° 36′ 8″ E). Geologically it belongs to the Boskovice Basin deposits. From stratigraphical point of view its position can be determined after fish or insect (cockroaches) assemblages ranging from Sakmarian to Early Artinskian (Schneider et al., 2003; Schneider and Werneburg, 2006; Zajíc and Štamberg, 2004; Zajíc, 2004; see Fig. 1 for details). The insect fauna is preserved in siltstones and claystones of greyish, yellow-greyish to brown color, of the Bačov Horizon. The fossiliferous layer reaches the ground surface in the small quarry now partially overgrown by trees. It should be noticed that some fossils from Obora and other localities in the Boskovice Basin were possibly affected by post-depositional deformations due to rock compression and distortion (see e.g., Béthoux and Nel, 2003; J. Kukalová-Peck, personal communication).

The present paper is the first contribution based on a series of newly excavated and revised insect material from the Early Permian of Boskovice Basin.

2. Brief review of research history of the Obora site excavations

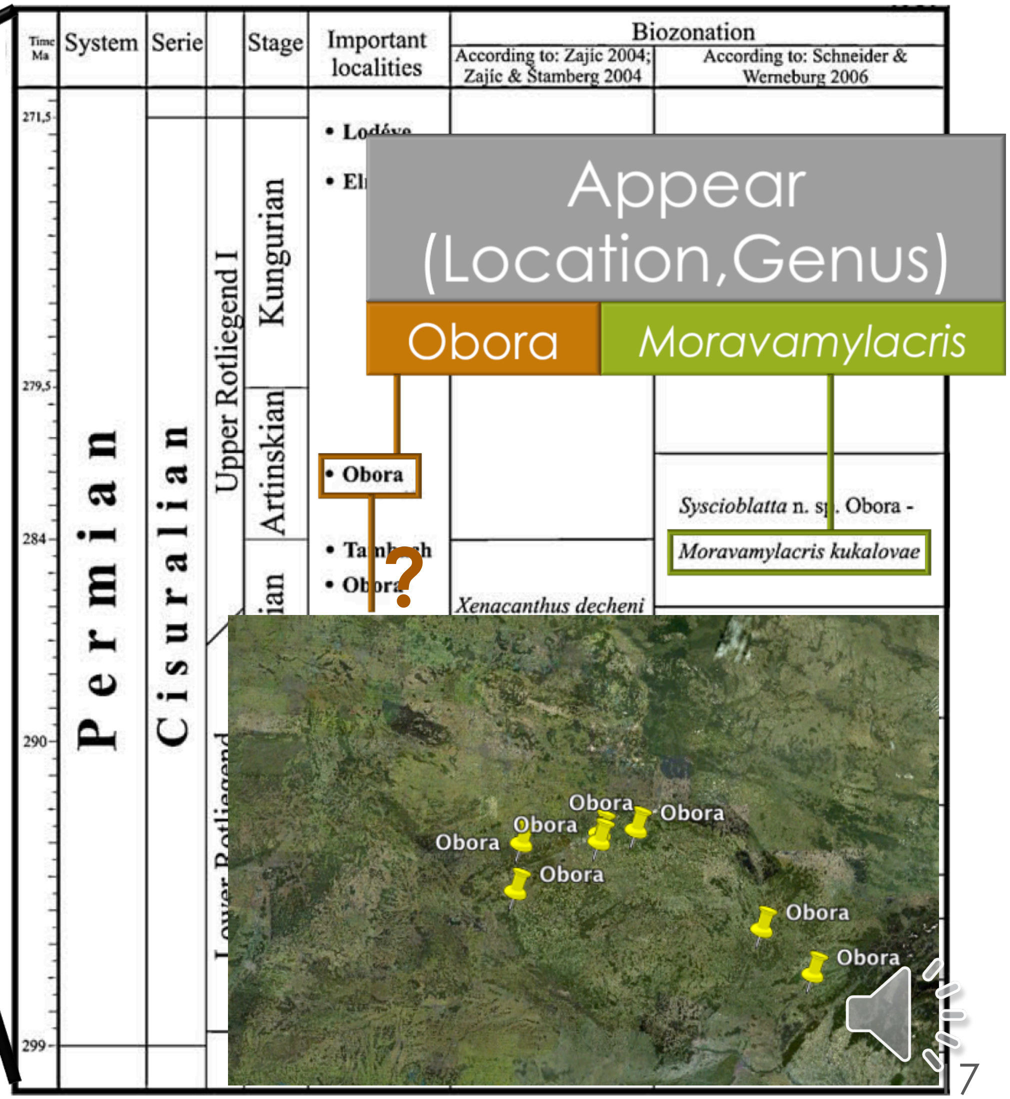
Research begun sporadically during the 20's of last century (Augusta, 1926). Since the end of the 50's, the locality underwent intensive paleoentomological research which was done by team led by Dr. J. Kukalová. Between the years 1958 and 1968 dozens of new insect species and even genera were described by Carpenter and Kukalová (1964), Kukalová (1963), and Kukalová-Peck (1972, 1974). In the era of active paleoentomological research, Zídek (1966) published the discovery of a xenacanthid shark.

At the beginning of the 80's, the locality was once again opened for paleontological research. Blattodea forewings were described by Schneider (1980, 1982, 1983, 1984). Holub and Kozur (1981) extended the list of discoveries with studies of Conchostraca and trace fossils. Martens (1982) described another representative of conchostracan from the Obora locality. At the end of the 80's, excavation on the Obora site was completely stopped. The historical contexts of its insect discoveries were summarized in faunal check lists provided by Zajíc and Štamberg (1986), Zajíc (1997) and Štamberg and Zajíc (2008).

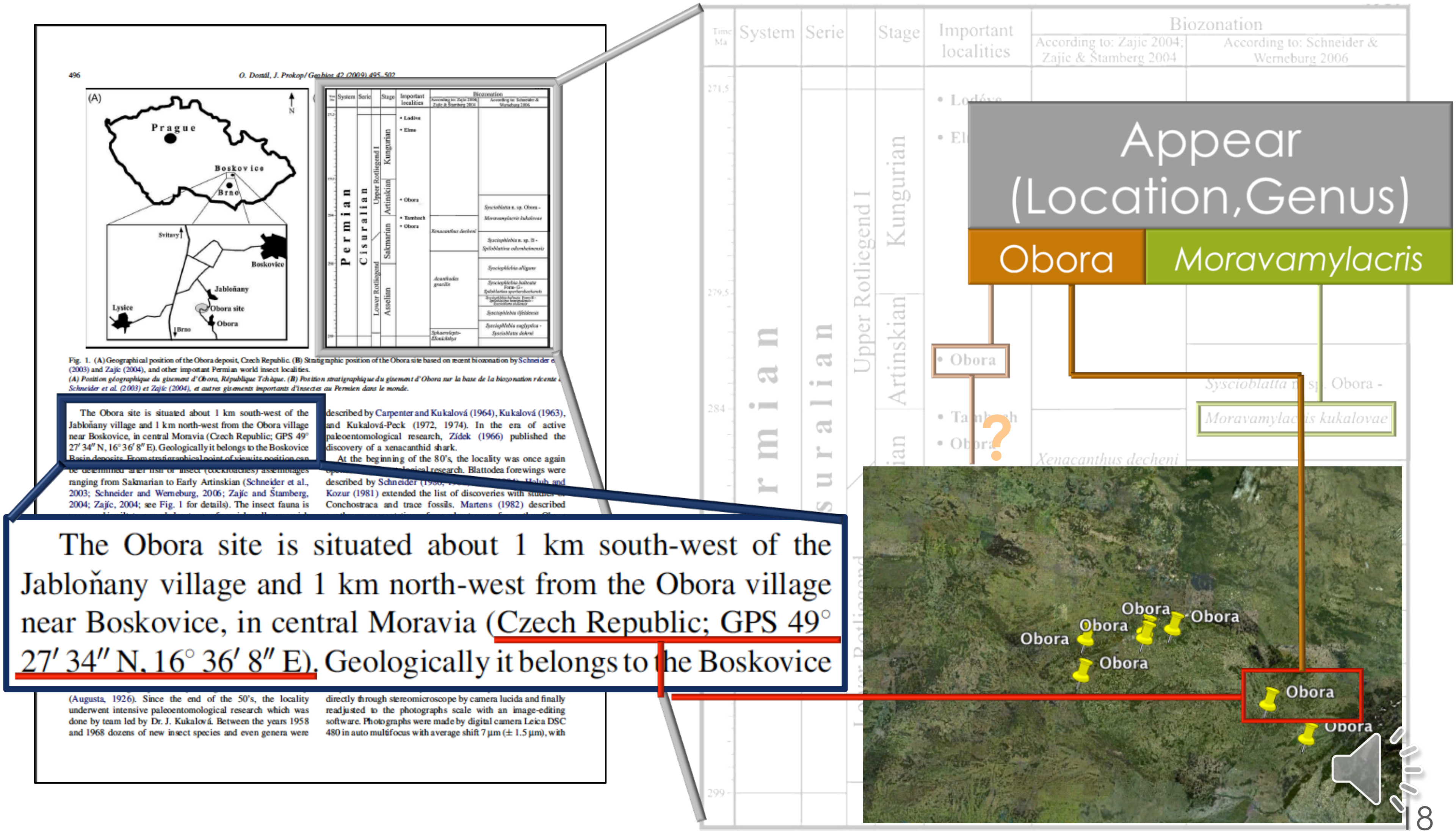
The last summary work of history of palaeontology in Boskovice Basin was provided by Ivanov (2003). Another summary work focusing on taphonomy and palaeontology of vertebrate amphibians from the Obora site was published by Dostál (2004).

3. Material and methods

The fossil specimens were observed with a stereomicroscope Leica MZ16 in dry state. Venation patterns were drawn directly through stereomicroscope by camera lucida and finally readjusted to the photographs scale with an image-editing software. Photographs were made by digital camera Leica DSC 480 in auto multifocus with average shift 7 µm (± 1.5 µm), with



Data are buried in tables, but not in a self-contained way



Joint Probabilistic Inference Matters.

PaleoDeepDive

Shanan Peters (Geo) and Miron Livny (CS)
DeepDive.Stanford.edu (**Ce Zhang** et al.)

nature International weekly journal of science

PaleoDB

Human-created

329 volunteers

13 years

46K documents

200+ Papers,
17 Nature/Science

PaleoDeepDive

Machine-created

10x **documents.**

100x **extractions.**

Preliminary Precision

Formation **Precision**	
	PaleoDB Volunteers: **0.84**
	PaleoDeepDive: **0.94**

Peters, S., Zhang, C, Livny, M., and Ré, C. A New Machine-Aggregated Empirical History of Life on Earth. ***PLOS ONE***, *2014 featured in* ***Nature*** *July 1, 15*

Hope: knowledge bases can help **accelerate science**.

Tree of Life

Drug Repurposing

Genomics

Used by a number of companies with quality that best ***professional*** *human annotators; winner of TACKBP14.*

Human Trafficking on the (Dark) Web…

DARPA MEMEX

Hypothesis: Trafficked individuals offer ***lower cost*** and ***riskier*** sexual services.

In Plain sight: Web ads for such services

Challenges:

1. Need **high-resolution information** to build model.
   - *services for what rate, ethnicity, location, etc.*
2. Scientific papers are **clear**—*dark web is* **obfuscated**.

Human Trafficking on the (Dark) Web…

DARPA MEMEX

Web Text

NASA

NYU VIDA
Where data comes to life

DeepDive

Structured Info: Phone #, Rates of Service, …

Normal two call agency. Jessica was ready when I arrived at body. A couple of tattoos but not in visual-annoying locations pleasing the client - definitely not GFE. She does the basics. recommending - I may repeat but only because of her looks.

Drug_Abuse?
Forced_Prostitution?

Use: *law enforcement/ NGOs*

In Use by Law Enforcement

New York DA use MEMEX Data for all trafficking investigations this year. ***Real Arrests***

For DARPA MEMEX, we were operational in 6 months

- Processed >35M documents (~26M records)
- Tens of columns (location, phone #, price, etc)
- With compute times of less than a day
- >90% Precision for most relations

Declarative language allows **algorithmic independence**

Example: Extracting Spouse Relations

Corpus (Dark Data)

U.S. President **Barack Obama**’s wife **Michelle Obama** honored all mothers on Mother’s Day and offered her thoughts …

How do we produce tuples like

`Married(Barack Obama, Michelle Obama)`

And all other married couples in text?

Examples on Github!

0. Data Preprocessing

Corpus (Dark Data)

text

U.S. President Barack Obama's wife Michelle Obama honored all mothers on Mother's Day and offered her thoughts …

DDlog: **Declarative** inspired by Datalog, MLNs

```
Sentences :- !nlp(Corpus).
function nlp over (text)
          returns (words,pos,ner,sid)
   implementation "udf/corenlp.sh …" .
```

Sentences

words	POS	NER	SID
[U.S.,President,Barack,Obama,'s,wife,Michelle,Obama,…]	[NNP,NNP,NNP,NNP,POS,NN,NNP,NNP,…]	[LOC,O,PER,PER,O,O,PER,PER,…]	S1

1. Candidate Mappings

Sentences

words	POS	NER	SID
[U.S.,President,Barack,Obama,'s,wife,Michelle,Obama,…]	[NNP,NNP,NNP,NNP,POS,NN,NNP,NNP,…]	[LOC,O,PER,PER,O,O,PER,PER,…]	S1

```
Mentions :- !ext_person(Sentences).
function ext_person over (words,pos,ner,sid)
                 returns (sid,mid,words)
          implementation "udf/find_person.py".
```

```
MarriedCandidate(s,p1,p2) :-
  Mentions(s,p1,_), Mentions(s,p2,_).
```

Mentions

SID	MID	words
S1	M1	[Barack,Obama]
S1	M2	[Michelle,Obama]

MarriedCandidate

SID	MID	MID
S1	M1	M2

2. Feature Extraction

Sentences

words	POS	NER	SID
[U.S.,President,Barack,Obama,'s,wife,Michelle,Obama,…]	[NNP,NNP,NNP,NNP,POS,NN,NNP,NNP,…]	[LOC,O,PER,PER,O,O,PER,PER,…]	S1

```
Features :- !ext_features(Sentences, MarriedCandidate).
function ext_features …
  implementation "udf/ext_features.py".
```

Mentions

SID	MID	words
S1	M1	[Barack,Obama]
S1	M2	[Michelle,Obama]

MarriedCandidate

SID	MID	MID
S1	M1	M2

Features

MID	MID	feature
M1	M2	's wife

3. Inference Rules

Sentences

words	POS	NER	SID
[U.S.,President,Barack,Obama,'s,wife,Michelle,Obama,...]	[NNP,NNP,NNP,NNP,POS,NN,NNP,NNP,...]	[LOC,O,PER,PER,O,O,PER,PER,...]	S1

```
Married(p1,p2) :-
  MarriedCandidate(_,p1,p2),
  Features(p1,p2,f)
weight = f.
```

Just defined a Binary classifier!

Married is an (incomplete) set of examples

Mentions

SID	MID	words
S1	M1	[Barack,Obama]
S1	M2	[Michelle,Obama]

MarriedCandidate

SID	MID	MID
S1	M1	M2

Features

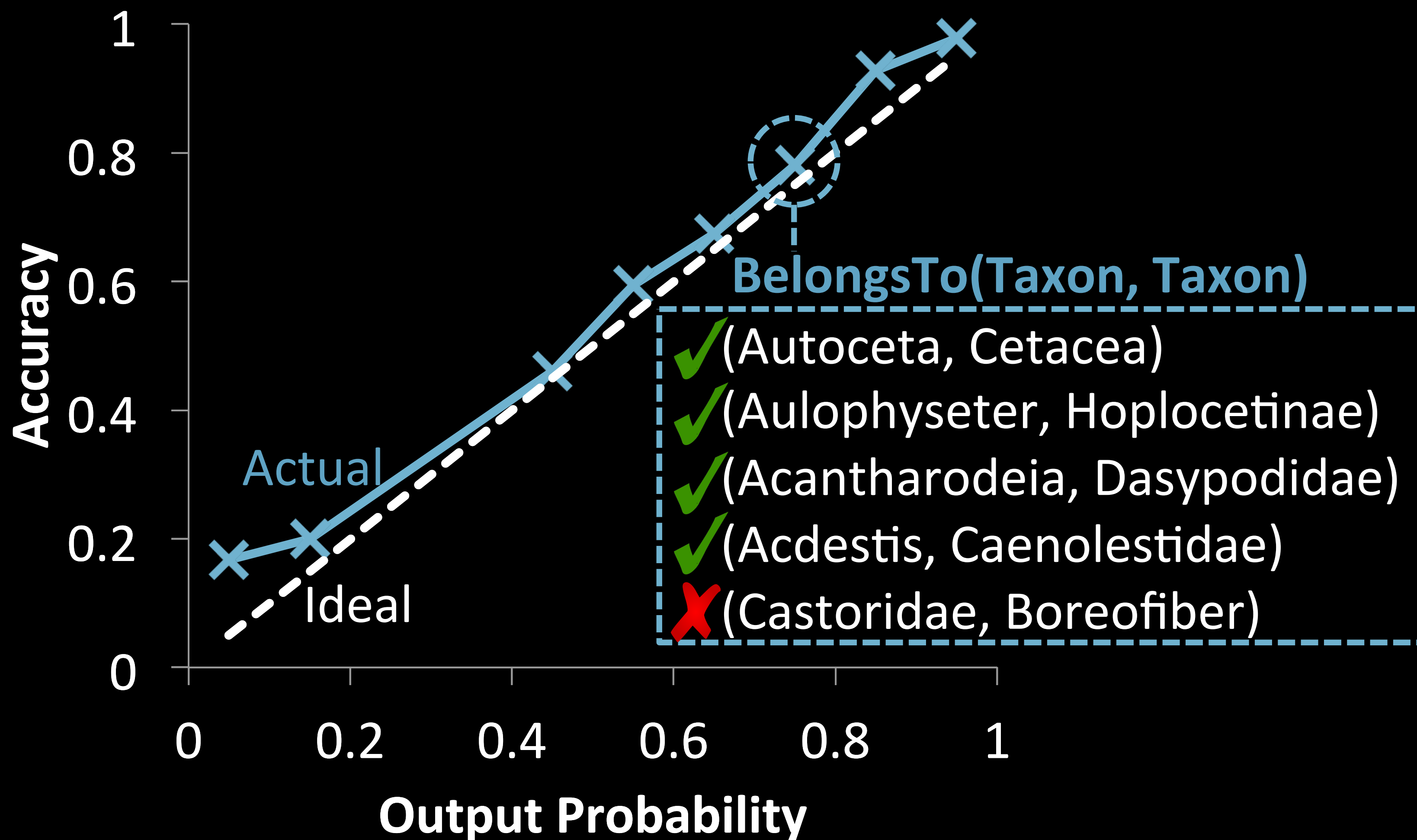
MID	MID	feature
M1	M2	's wife

Users write Features & Transformation in **DDlog** (Inspired by MLNs) & Python.

```
Sentences :- !nlp(Corpus).
function nlp over (text) returns (words,pos,ner,sid)
   implementation "udf/corenlp.sh …" .

Mentions :- !ext_person(Sentences).
function ext_person over (words,pos,ner,sid) returns (sid,mid,words)
          implementation "udf/find_person.py".
MarriedCandidate(s,p1,p2) :- Mentions(s,p1,_), Mentions(s,p2,_).

Features :- !ext_features(Sentences, MarriedCandidate).
function ext_features … implementation "udf/ext_features.py".

Married(p1,p2) :- MarriedCandidate(_,p1,p2), Features(p1,p2,f)
weight = f.
```

No reference to algorithms: Just features. DD does the rest. *(Demos online)*

Meaningful probability, **not just** scores



Algorithmic Independence

Define the *meaning* **independently** of *the algorithm used to compute it*.

algorithmic independences,
requires a *fast* engine….

The Key Balance

Key Issue: Balance
<u>Statistical</u> *versus* *<u>Hardware</u>* *Efficiency.*

Statistical Analytics Crash Course

Staggering amount of machine learning/stats can be written as:

$$\min_x \sum_{i=1}^{N} f(x, y_i)$$

N (number of y_is, data) typically in the billions

Ex: Classification, Recommendation, Deep Learning.

De facto iteration to solve large-scale problems: **SGD**.

$$x^{k+1} = x^k - \alpha \boxed{N \nabla f(x^k, y_j)}$$

Select one term, j, and estimate gradient.

Billions of tiny iterations.

How do we run SGD in Parallel?

Data Systems Perspective of SGD.

$$x^{k+1} = x^k - \alpha N \nabla f(x^k, y_j)$$

Insane conflicts: Billions of tiny (~100 instructions) jobs, RW conflicts on x, which is called **the model.**

How can we hope to speed this up with parallelism?

Serializability seems hopeless…

How do we run SGD in Parallel?

Thm: If we do ***no locking***, SGD still converges to right answer—at essentially the same theoretical rate!

Hogwild! [Niu, Recht, **Ré,** Wright NIPS11]
AsySCD [Liu, Wright et al. ICML14, JMLR14]
Buckwild! [DeSa et al. ICML15**]**

Technical conditions on ratio of processors, delays, (semantic) sparsity.

High-level idea: Go Hogwild! answer is only *statistically* correct.

Cortana & Project Adam
Microsoft's Digital Assistant

AI breakthrough: Microsoft's 'Project Adam' identifies dog breeds, points to future of machine learning

WIRED

"…using a technology called, of all things, ***Hogwild!"***

I ♥ Microsoft

http://www.wired.com/2014/07/microsoft-adam/

http://www.geekwire.com/2014/artificial-intelligence-breakthrough-microsofts-project-adam-identifies-dog-breeds/

A larger trend?

****NB**: There is theory here SGD [NIPS11,NIPS12], SCD [ICML14,ICML15], more soon and systems work [SIGMOD13, SIGMOD14, VLDB14]*

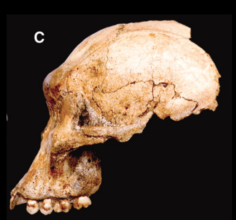
Relaxing **consistency** to be **architecturally aware** can be a big performance win.

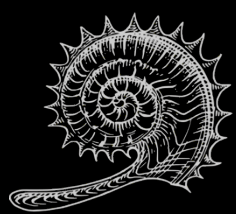
Dark(er) Data Systems

Integrative: Varied Data

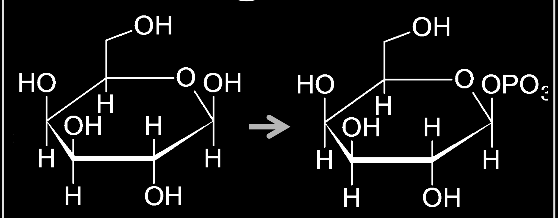
Text&Tables

Images

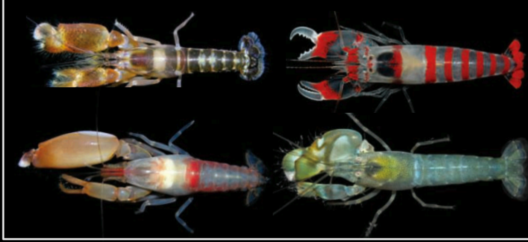




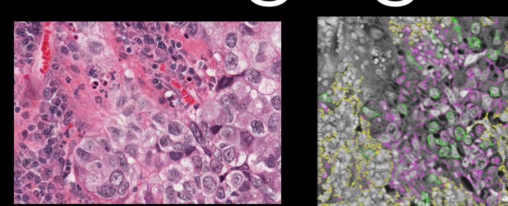
Diagrams



Taxonomy



Imaging



Drug Repurposing

Paleobiology

Cancer Imaging

Everyone: Broadly Usable

PaleoDeepDive

2 years

1 CS Student

PharmaDeepDive

6 months

1 BioE Student

How do we make building a KB
easier and cheaper?

Think about **features**, not **algorithms.**

A **framework** for feature engineering.
[SIGMOD14: train 100 models as quickly as 1]

Conclusion

1. **Dark Data** to help with **macroscopic questions**

2. Probabilistic inference = **algorithmic** independence

3. **Hardware v. Statistical** Efficiency.