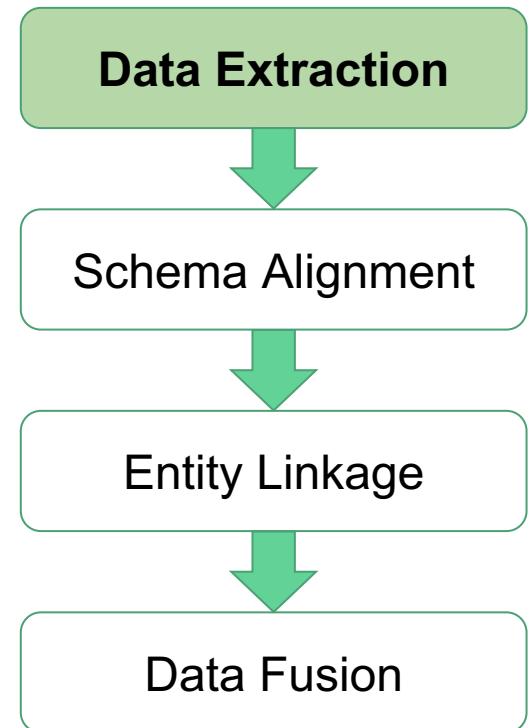


# Outline

- Part I. Introduction
- Part II. ML for DI
  - ML for entity linkage
  - ML for data extraction
  - ML for data fusion
  - ML for schema alignment
- Part III. DI for ML
- Part IV. Conclusions and research direction



# What is Data Extraction?

- Definition: Extract structured information, e.g., (entity, attribute, value) triples, from semi-structured data or unstructured data.

**Web tables & Lists**

	Name and (party) <sup>1</sup>	Term
1.	Washington (F) <sup>3</sup>	1789–1797
2.	J. Adams (F)	1797–1801
3.	Jefferson (DR)	1801–1809
4.	Madison (DR)	1809–1817

**DOM Trees**

yelp

Welcome About Me Write a Review Find Friends

Shana Thai Restaurant

54 reviews Rating Details

Category: Thai (16)

311 Moffett Blvd  
Sunnyvale, CA 94085  
(855) 940-9999  
<http://www.shanathai.com>

Explore the menu

Hours:

Mon–Sun 11 am – 9 pm  
Mon–Sun 5 pm – 10 pm  
Good for Kids: Yes  
Accepts Credit Cards: Yes  
Parking: Private Lot  
Atmos: Casual  
Good for Groups: Yes

Price Range: \$  
Takes Reservations: No  
Delivery: No  
Take-out: Yes  
Washer Service: Yes  
Outdoor Seating: No  
Wi-Fi: No  
Good For: Dinner

**Free texts**

Synopsis

Born on April 15, 1452, in Vinci, Italy, Leonardo da Vinci was a man concerned with the laws of science and nature. He informed his work as a painter, sculptor, engineer, and scientist. His ideas and body of work -- which include the Vitruvian Man, the Last Supper, Leda and the Swan, and the Vitruvian Horse -- influenced countless artists and made him one of the most influential figures of the Italian Renaissance.

**Diagram**

Biological level	Examples	Pre-amputation	Post-amputation	Regenerate
Whole body	Regeneration from a small body fragment			
Structure	Limb, fin, tail, head, tentacle, siphon, arm, stalk			
Internal organ	Heart, liver, lens			
Tissue	Epidermis, gut lining			
Cell	Axon, muscle fiber			

Regeneration

TRENDS in Ecology & Evolution

# What is Data Extraction?

- Definition: Extract structured information, e.g., (entity, attribute, value) triples, from **semi-structured** data or unstructured data.

**Focus of this tutorial**

The diagram illustrates the focus of the tutorial on semi-structured data. It features four main sections: 'Web tables & Lists', 'DOM Trees', 'Free texts', and 'Diagram'. A green arrow points from the text 'Focus of this tutorial' to the 'Free texts' section, which contains examples of semi-structured data like Yelp reviews and historical synopsis text. Below the 'Free texts' section is a detailed diagram of biological regeneration levels from whole body down to cell, showing pre- and post-amputation states and regenerate processes.

**Web tables & Lists**

	Name and (party) <sup>1</sup>	Term
1.	Washington (F) <sup>3</sup>	1789–1797
2.	J. Adams (F)	1797–1801
3.	Jefferson (DR)	1801–1809
4.	Madison (DR)	1809–1817

**DOM Trees**

yelp

Welcome About Me Write a Review Find Friends

Shana Thai Restaurant

Category: Thai (16) 740 reviews Rating Details

311 Moffett Blvd  
Sunnyvale, CA 94085  
(855) 940-9999  
<http://www.shanathai.com>

Explore the menu

Hours:

Price Range: \$

Takes Reservations: Yes

Delivery: No

Takes-out: Yes

Waiter Service: Yes

Outdoor Seating: Yes

Wi-Fi: No

Good For Groups: Yes

Good For: Dinner

**Free texts**

**Synopsis**

Born on April 15, 1452, in Vinci, Italy, Leonardo da Vinci was a polymath who concerned with the laws of science and nature. He informed his work as a painter, sculptor, architect, engineer, and scientist. His ideas and body of work -- which include the Vitruvian Man, the Last Supper, and Leda and the Swan -- influenced countless artists and made him one of the most influential figures of the Italian Renaissance.

**Diagram**

Biological level	Examples	Pre-amputation	Post-amputation	Regenerate
Whole body	Regeneration from a small body fragment			
Structure	Limb, fin, tail, head, tentacle, siphon, arm, stalk			
Internal organ	Heart, liver, lens			
Tissue	Epidermis, gut lining			
Cell	Axon, muscle fiber			

TRENDS in Ecology & Evolution

# Three Types of Data Extraction

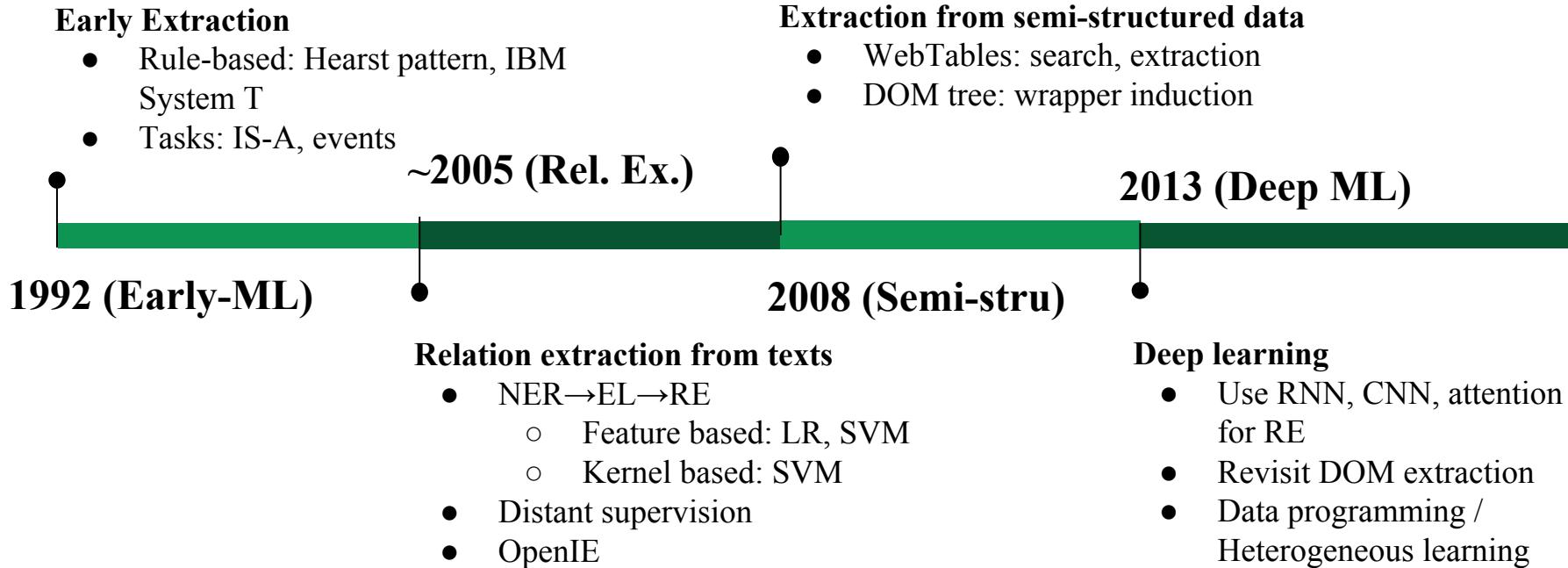
- **Closed-world extraction:** align to existing entities and attributes; e.g.,  
("ID\_Obama", place\_of\_birth, ID\_USA)
- **ClosedIE:** align to existing attributes, but extract new entities; e.g.,  
("Xin Luna Dong", place\_of\_birth, "China")
- **OpenIE:** not limited by existing entities or attributes; e.g.,  
("Xin Luna Dong", "was born in", "China"),  
("Luna", "is originally from", "China")

# Three Types of Data Extraction

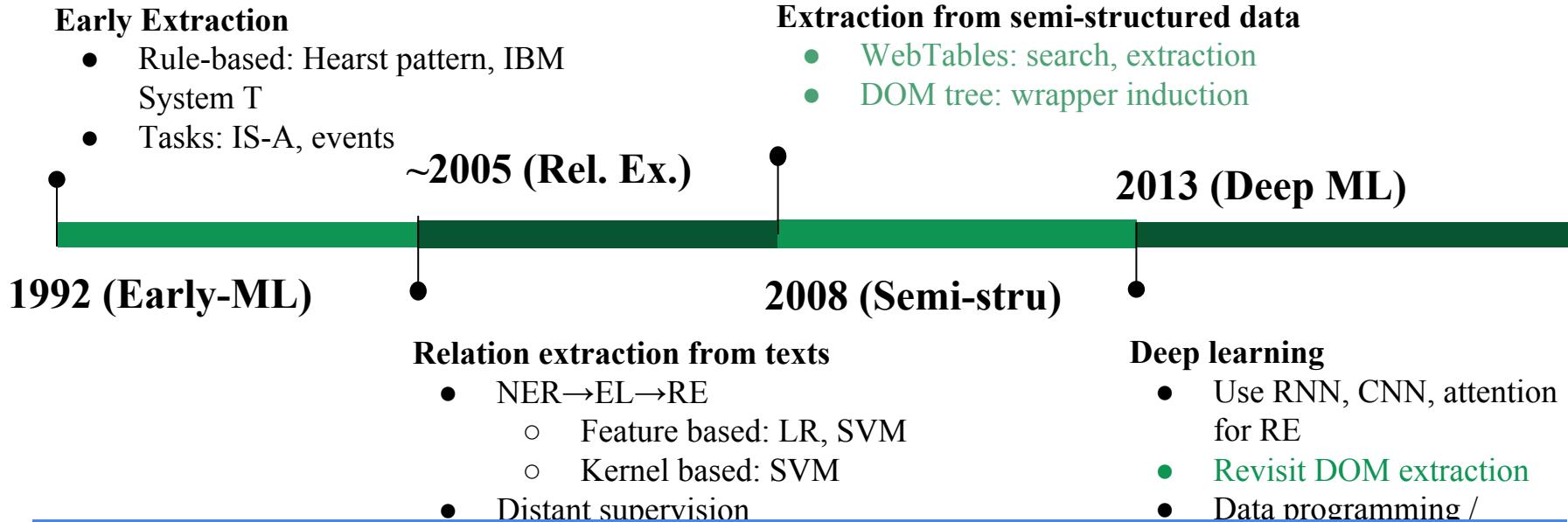
- **Closed-world extraction:** align to existing entities and attributes; e.g.,  
("ID\_Obama", place\_of\_birth, ID\_USA)
- **ClosedIE:** align to existing attributes, but extract new entities; e.g.,  
("Xin Luna Dong", place\_of\_birth, "China")
- **OpenIE:** not limited by existing entities or attributes; e.g.,  
("Xin Luna Dong", "was born in", "China"),  
("Luna", "is originally from", "China")

Focus of this tutorial

# 35 Years of Data Extraction



# 35 Years of Data Extraction

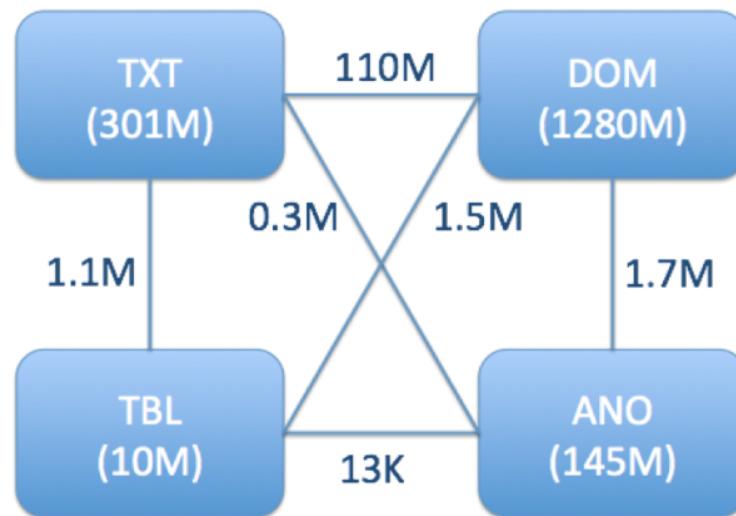


Come to our VLDB tutorial for text extraction and OpenIE!!

# Why Semi-Structured Data?

- Knowledge Vault @ Google showed big potential from DOM-tree extraction  
[Dong et al., KDD'14][Dong et al., VLDB'14]

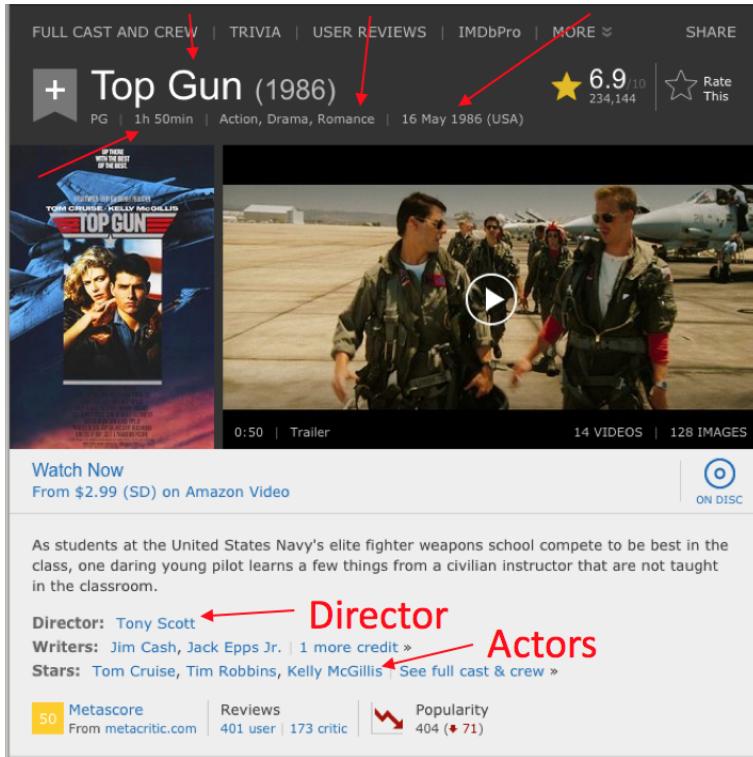
Accu	Accu (conf $\geq .7$ )
0.36	0.52



Accu	Accu (conf $\geq .7$ )
0.43	0.63
0.09	0.62

# Wrapper Induction--Vertex [Gulhane et al., ICDE'11]

Runtime



## Extracted relationships

- (Top Gun, type.object.name, "Top Gun")
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed\_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, "1h 50min")
- (Top Gun, film.film.release\_Date\_s, "16 May 1986")

# Wrapper Induction--Vertex [Gulhane et al., ICDE'11]

- Solution: find XPaths from DOM Trees

Filmography

Show all | Show by... | Edit

Jump to: Actor | Producer | Soundtrack | Director | Writer | Thanks | Self | Archive footage

**Actor (46 credits)** Hide ▲

<b>Top Gun: Maverick</b> ( <i>pre-production</i> ) Maverick	2019
<b>M:I 6 - Mission Impossible</b> ( <i>filming</i> ) Ethan Hunt	2018
<b>American Made</b> ( <i>completed</i> ) Barry Seal	2017
<b>Luna Park</b> ( <i>announced</i> )	
<b>The Mummy</b> Nick Morton	2017
<b>Jack Reacher: Never Go Back</b> Jack Reacher	2016
<b>Mission: Impossible - Rogue Nation</b> Ethan Hunt	2015
<b>Edge of Tomorrow</b> Cage	2014
<b>Oblivion</b> Jack	2013/I
<b>Jack Reacher</b> Reacher	2012
<b>Rock of Ages</b> Stacee Jaxx	2012
<b>Mission: Impossible - Ghost Protocol</b> Ethan Hunt	2011
<b>Knight and Day</b> Roy Miller	2010
<b>Valkyrie</b> Colonel Claus von Stauffenberg	2008
<b>Tropic Thunder</b>	2008

```
<div id="filmography"> = $0
  ><div id="filmo-head-actor" class="head" data-category="actor" onclick="toggleFilmoCategory(this);"></div>
  ><div class="filmo-category-section">
    ><div class="filmo-row odd" id="actor-tt1745960">
      <span class="year_column">
        &nbsp;2019
      </span>
      <b>
        <a href="/title/tt1745960/?ref=nm_flm_act_1">Top Gun: Maverick</a>
      </b>
      "
      (
      <a href="/r/legacy-inprod-name/title/tt1745960" class="in_production">pre-production</a>
      )
      "
      <br>
      <a href="/character/ch0085702/?ref=nm_flm_act_1">Maverick</a>
    </div>
    ><div class="filmo-row even" id="actor-tt4912910"></div>
    ><div class="filmo-row odd" id="actor-tt3532216"></div>
    ><div class="filmo-row even" id="actor-tt1123441"></div>
    ><div class="filmo-row odd" id="actor-tt2345759">
      <span class="year_column">
        &nbsp;2017
      </span>
      <b>
        <a href="/title/tt2345759/?ref=nm_flm_act_5">The Mummy</a>
      </b>
      <br>
      <a href="/character/ch0573416/?ref=nm_flm_act_5">Nick Morton</a>
    </div>
    ><div class="filmo-row even" id="actor-tt3393786"></div>
    ><div class="filmo-row odd" id="actor-tt2381249"></div>
    ><div class="filmo-row even" id="actor-tti1631867"></div>
    ><div class="filmo-row odd" id="actor-tt1483013"></div>
    ><div class="filmo-row even" id="actor-tt0790724"></div>
    ><div class="filmo-row odd" id="actor-tt1336608"></div>
```

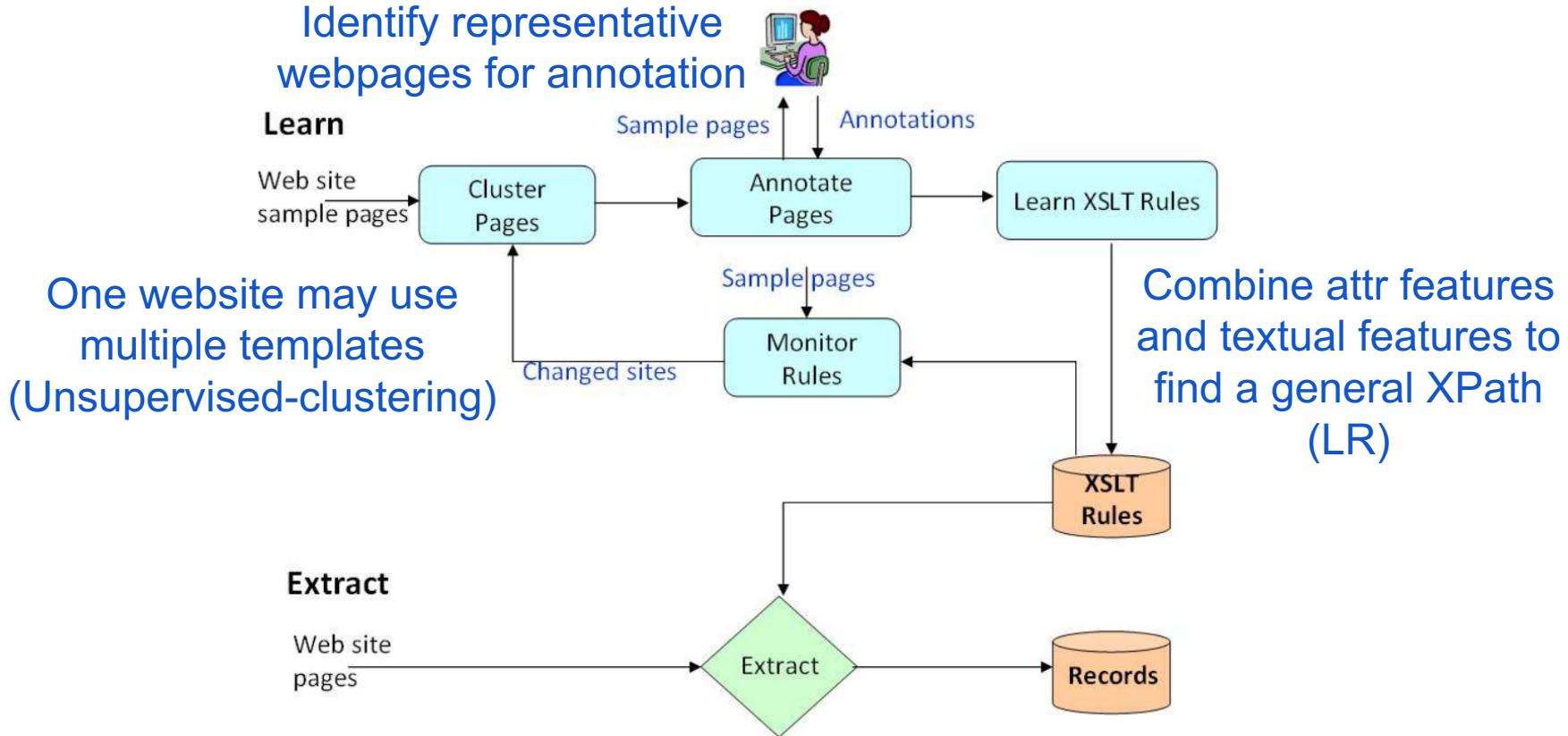
# Wrapper Induction--Vertex [Gulhane et al., ICDE'11]

- Challenge: slight variations from page to page

```
/html/body/div[1]/div/div[4]/div[3]/div[3]/div[3]/div[4]/div[26]/b/a  
/html/body/div[1]/div/div[4]/div[3]/div[3]/div[3]/div[3]/div[2]/div[10]/b/a
```

**Figure 2: Example of XPaths corresponding to the *acted in* predicate on two IMDb pages. They differ at two node indices, and the second path corresponds to the *producer of* predicate from the first page.**

# Wrapper Induction--Vertex [Gulhane et al., ICDE'11]



# Wrapper Induction--Vertex [Gulhane et al., ICDE'11]

- Sample learned XPaths on IMDb

- `//*[@itemprop="name"]`
- `//*[@class="bp_item bp_text_only"]/*/*/*[@class="bp_heading"]`
- `//*[following-sibling::*[position()=3][@class="subheading"]]/*[following-sibling::*[position()=1][@class="attribute"]]`
- `//*[preceding-sibling::node()[normalize-space(.)!=""][text()="Language:"]`

Ensure high recall

Ensure high precision

# Distantly Supervised Extraction

- **Annotation-based extraction**
  - Pros: high precision and recall
  - Cons: does not scale--annotation per cluster per website
- **Distantly-supervised extraction**
  - Step 1. Use seed data to automatically annotate
  - Step 2. Use the (noisy) annotations for training
  - E.g., DeepDive, Knowledge Vault

# Distant Supervision [Mintz et al., ACL'09]

Corpus Text

Bill Gates founded Microsoft in 1975.  
Bill Gates, founder of Microsoft, ...  
Bill Gates attended Harvard from ...  
Google was founded by Larry Page ...

Training Data

Freebase

(Bill Gates, Founder, Microsoft)  
(Larry Page, Founder, Google)  
(Bill Gates, CollegeAttended, Harvard)

[Adapted example from Luke Zettlemoyer]

# Distant Supervision [Mintz et al., ACL'09]

Corpus Text

Bill Gates founded Microsoft in 1975.  
Bill Gates, founder of Microsoft, ...  
Bill Gates attended Harvard from ...  
Google was founded by Larry Page ...

Training Data

(Bill Gates, Microsoft)  
Label: Founder  
Feature: X founded Y

Freebase

(Bill Gates, Founder, Microsoft)  
(Larry Page, Founder, Google)  
(Bill Gates, CollegeAttended, Harvard)

[Adapted example from Luke Zettlemoyer]

# Distant Supervision [Mintz et al., ACL'09]

## Corpus Text

Bill Gates founded Microsoft in 1975.

Bill Gates, founder of Microsoft, ...

Bill Gates attended Harvard from ...

Google was founded by Larry Page ...

## Training Data

(Bill Gates, Microsoft)

Label: Founder

Feature: X founded Y

Feature: X, founder of Y

## Freebase

(Bill Gates, Founder, Microsoft)

(Larry Page, Founder, Google)

(Bill Gates, CollegeAttended, Harvard)

[Adapted example from Luke Zettlemoyer]

# Distant Supervision [Mintz et al., ACL'09]

## Corpus Text

Bill Gates founded Microsoft in 1975.  
Bill Gates, founder of Microsoft, ...  
**Bill Gates attended Harvard from ...**  
Google was founded by Larry Page ...

## Freebase

(Bill Gates, Founder, Microsoft)  
(Larry Page, Founder, Google)  
**(Bill Gates, CollegeAttended, Harvard)**

## Training Data

(Bill Gates, Microsoft)  
Label: Founder  
Feature: X founded Y  
Feature: X, founder of Y

(Bill Gates, Harvard)  
Label: CollegeAttended  
Feature: X attended Y

For negative examples, sample  
unrelated pairs of entities.

[Adapted example from Luke Zettlemoyer]

# Distantly Supervised Extraction--Ceres [Lockard et al., VLDB'18]



## Movie entity



Runtime

## Genre      Release Date



As students at the United States Navy's elite fighter weapons school compete to be best in the class, one daring young pilot learns a few things from a civilian instructor that are not taught in the classroom.  
Director: Tony Scott  
Writers: Jim Cash, Jack Epps Jr. | 1 more credit  
Stars: Tom Cruise, Tim Robbins, Kelly McGillis | See full cast & crew  
Metascore  
Reviews  
Popularity

Director  
Actors

## Extracted triples

- (Top Gun, type.object.name, "Top Gun")
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed\_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, "1h 50min")
- (Top Gun, film.film.release\_Date\_s, "16 May 1986")

# Distantly Supervised Extraction--Ceres [Lockard et al., VLDB'18]

- Extraction experiments on SWDE benchmark

Vertical	Predicate	Vertex++			CERES-Full		
		P	R	F1	P	R	F1
Movie	Title	1.00	1.00	1.00	1.00	1.00	1.00
	Director	0.99	0.99	0.99	0.99	0.99	0.99
	Genre	0.88	0.87	0.87	0.93	0.97	0.95
	MPAA Rating	1.00	1.00	1.00	NA	NA	NA
	Average	0.97	0.97	0.97	0.97	0.99	0.98
NBAPlayer	Name	0.99	0.99	0.99	1.00	1.00	1.00
	Team	1.00	1.00	1.00	0.91	1.00	0.95
	Weight	1.00	1.00	1.00	1.00	1.00	1.00
	Height	1.00	1.00	1.00	1.00	0.90	0.95
	Average	1.00	1.00	1.00	0.98	0.98	0.98

Vertical	Predicate	Vertex++			CERES-Full			
		P	R	F1	P	R	F1	
University	Name	1.00	1.00	1.00	1.00	1.00	1.00	
	Type	1.00	1.00	1.00	0.72	0.80	0.76	
	Phone	0.97	0.92	0.94	0.85	0.95	0.90	
	Website	1.00	1.00	1.00	0.90	1.00	0.95	
	Average	0.99	0.98	0.99	0.87	0.94	0.90	
Book	Title	0.99	0.99	0.99	1.00	0.90	0.95	
	Author	0.97	0.96	0.96	0.72	0.88	0.79	
	Publisher	0.85	0.85	0.85	0.97	0.77	0.86	
	Publication Date	0.90	0.90	0.90	1.00	0.40	0.57	
	ISBN-13	0.94	0.94	0.94	0.99	0.19	0.32	
		Average	0.93	0.93	0.93	0.94	0.63	0.70

Very high precision

Competent w. Wrapper induction w. manual annotation

# Distantly Supervised Extraction--Ceres [Lockard et al., VLDB'18]

- Extraction on long-tail movie websites

#Websites / #Webpages	33 / 434K
Language	English and 6 other languages
Domains	Animated films, Documentary films, Financial performance, etc.
# Annotated pages	70K (16%)
Annotated : Extracted #entities	1 : 2.6
Annotated : Extracted #triples	1 : 3.0
# Extractions	1.25 M
Precision	90%

# Distantly Supervised Extraction--Ceres [Lockard et al., VLDB'18]

- Which model is the best?
  - Logistic regression: best results (20K features on one website)
  - Random forest: lower precision and recall
  - Deep learning??

# Challenges in Applying Deep Learning on Extracting Semi-structured Data

- Web layout is neither 1D sequence nor regular 2D grid, so CNN or RNN does not directly apply

The screenshot shows a section of a movie's production credits and technical specifications. The credits are listed in a horizontal scrollable bar. The technical specs include runtime, sound mix options, color, and aspect ratio.

**Company Credits**

**Production Co:** Lucasfilm, Walt Disney Pictures, Allison Shearmur Productions [See more »](#)

Show more on [IMDbPro »](#)

---

**Technical Specs**

**Runtime:** 135 min

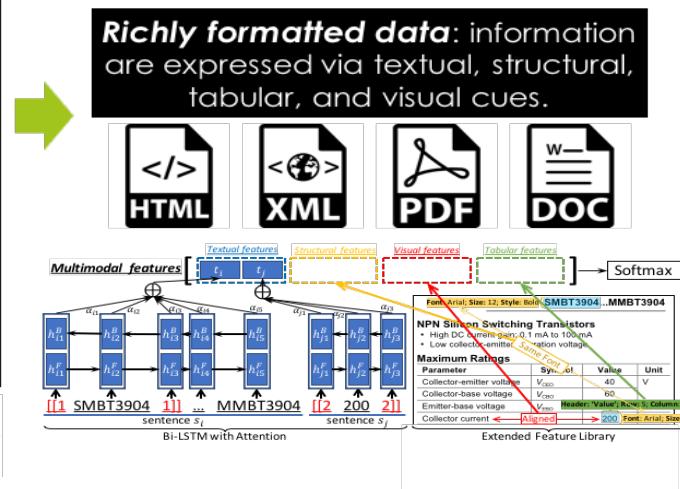
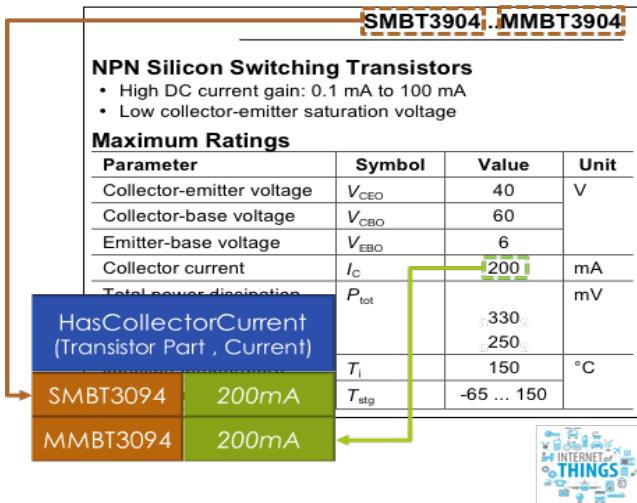
**Sound Mix:** Dolby Atmos | DTS (DTS: X) | 12-Track Digital Sound | Auro 11.1 | Dolby Digital  
Dolby Surround 7.1

**Color:** Color

**Aspect Ratio:** 2.39 : 1

See [full technical specs »](#)

# Example System: Fonduer [Wu et al., SIGMOD'18]



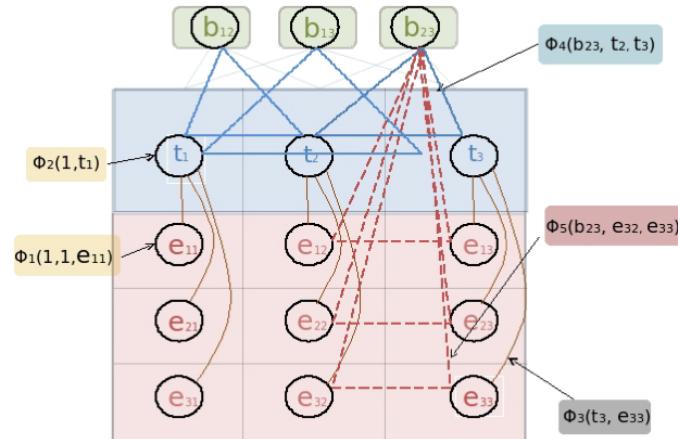
Fonduer combines a new **bi-directional LSTM with multimodal features** and weak supervision (specifically **data programming**).

Attend the talk in Research Session 13!

New version of code coming soon: <https://github.com/HazyResearch/fonduer>

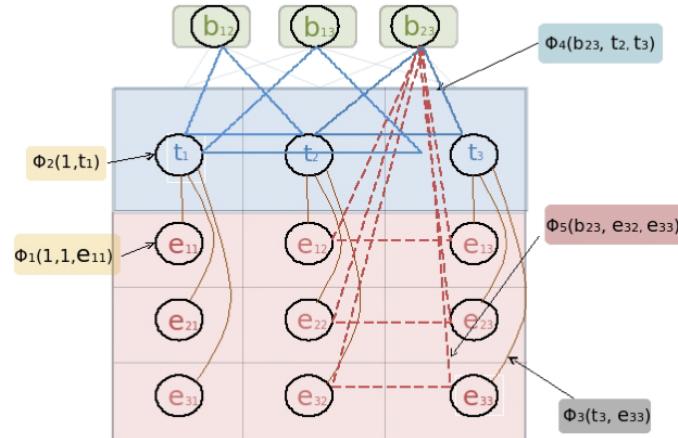
# WebTable Extraction [Limaye et al., VLDB'10]

- Model table annotation using interrelated random variables, represented by a probabilistic graphical model
  - Cell text (in Web table) and entity label (in catalog)
  - Column header (in Web table) and type label (in catalog)
  - Column type and cell entity (in Web table)



# WebTable Extraction [Limaye et al., VLDB'10]

- Model table annotation using interrelated random variables, represented by a probabilistic graphical model
  - Pair of column types (in Web table) and relation (in catalog)
  - Entity pairs (in Web table) and relation (in catalog)



# Challenges in Applying ML on DX

- Automatic data extraction cannot reach production quality requirement. How to improve precision?
- Every web designer has her own whim, but there are underlying patterns across websites. How to learn extraction patterns on different websites, especially for semi-structured sources?
- ClosedIE throws away too much data. How to apply OpenIE on all kinds of data?

# Recipe for Data Extraction

- Problem definition: Extract structure from semi- or un-structured data
- Short answers
  - Wrapper induction has high prec/rec
  - Distant supervision is critical for collecting training data
  - LR is often effective; more research is needed for DL

