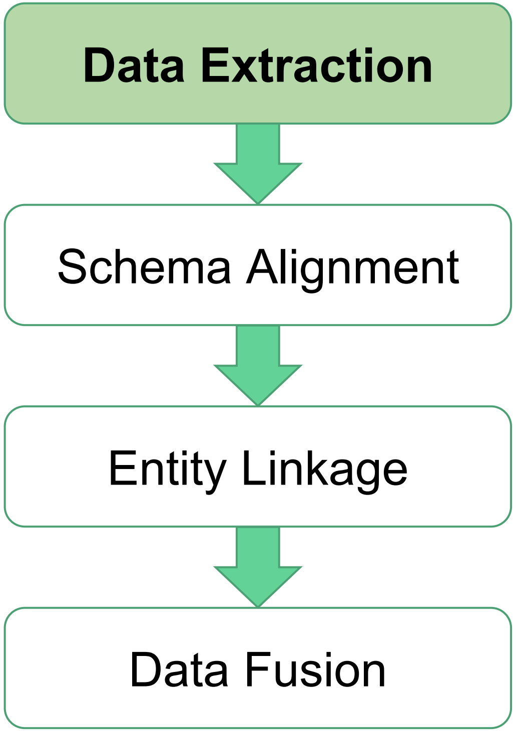
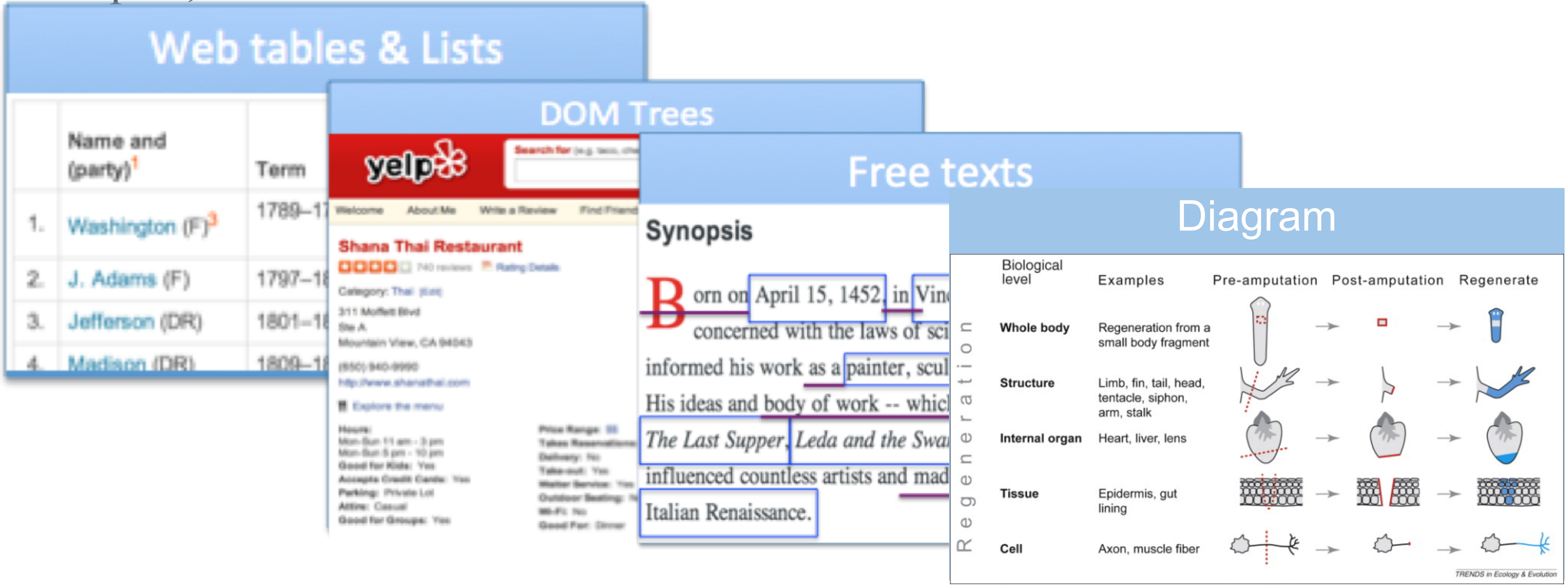
Outline

- 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?

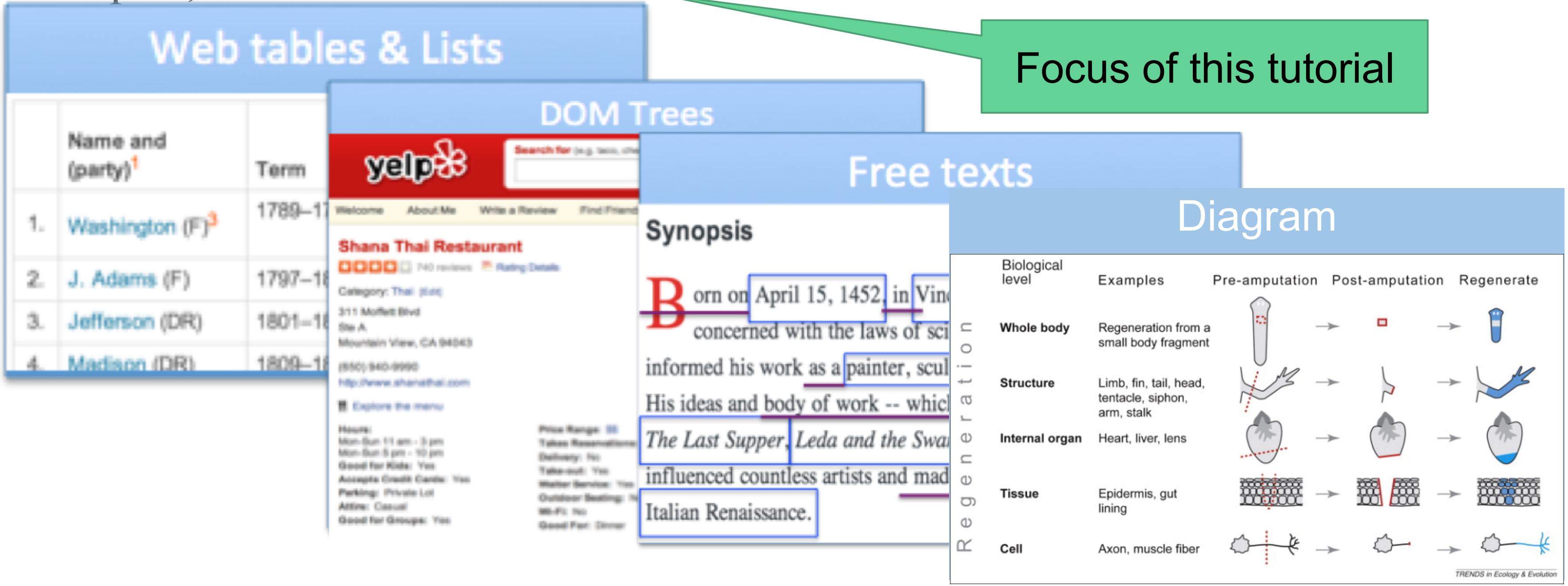
- Definition: Extract structured information, e.g., (entity, attribute, value) triples, from semi-structured data or unstructured data.



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



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")

Focus of this tutorial

- **OpenIE**: not limited by existing entities or attributes; e.g., ("Xin Luna Dong", "was born in", "China"), ("Luna", "is originally from", "China")

35 Years of Data Extraction

1992 (Early-ML)

Early Extraction

- Rule-based: Hearst pattern, IBM System T
- Tasks: IS-A, events

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

2008 (Semi-stru)

Extraction from semi-structured data

- WebTables: search, extraction
- DOM tree: wrapper induction

2013 (Deep ML)

Deep learning

- Use RNN, CNN, attention for RE
- Revisit DOM extraction
- Data programming / Heterogeneous learning

35 Years of Data Extraction

Early Extraction
- Rule-based: Hearst pattern, IBM System T
- Tasks: IS-A, events

1992 (Early-ML)

~2005 (Rel. Ex.)

Relation extraction from texts
- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision

2008 (Semi-stru)

Extraction from semi-structured data
- WebTables: search, extraction
- DOM tree: wrapper induction

2013 (Deep ML)

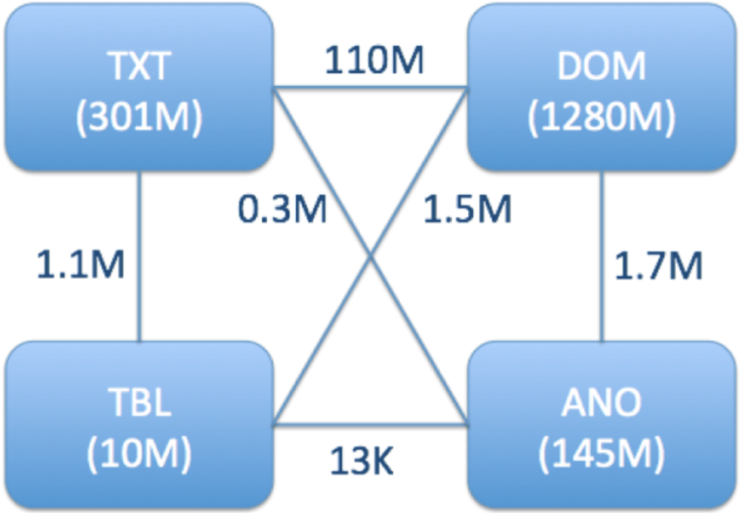
Deep learning
- Use RNN, CNN, attention for RE
- Revisit DOM extraction
- Data programming /

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

Why Semi-Structured Data?

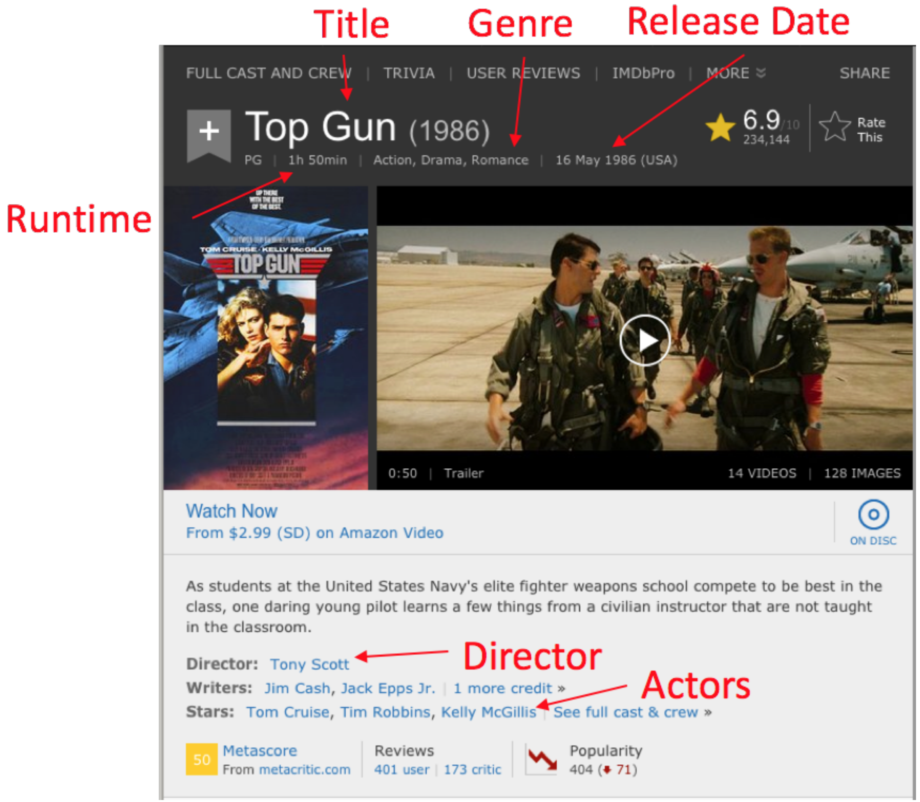
- Knowledge Vault @ Google showed big potential from DOM-tree extraction [Dong et al., KDD'14][Dong et al., VLDB'14]

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]



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

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

- Challenge: slight variations from page to page

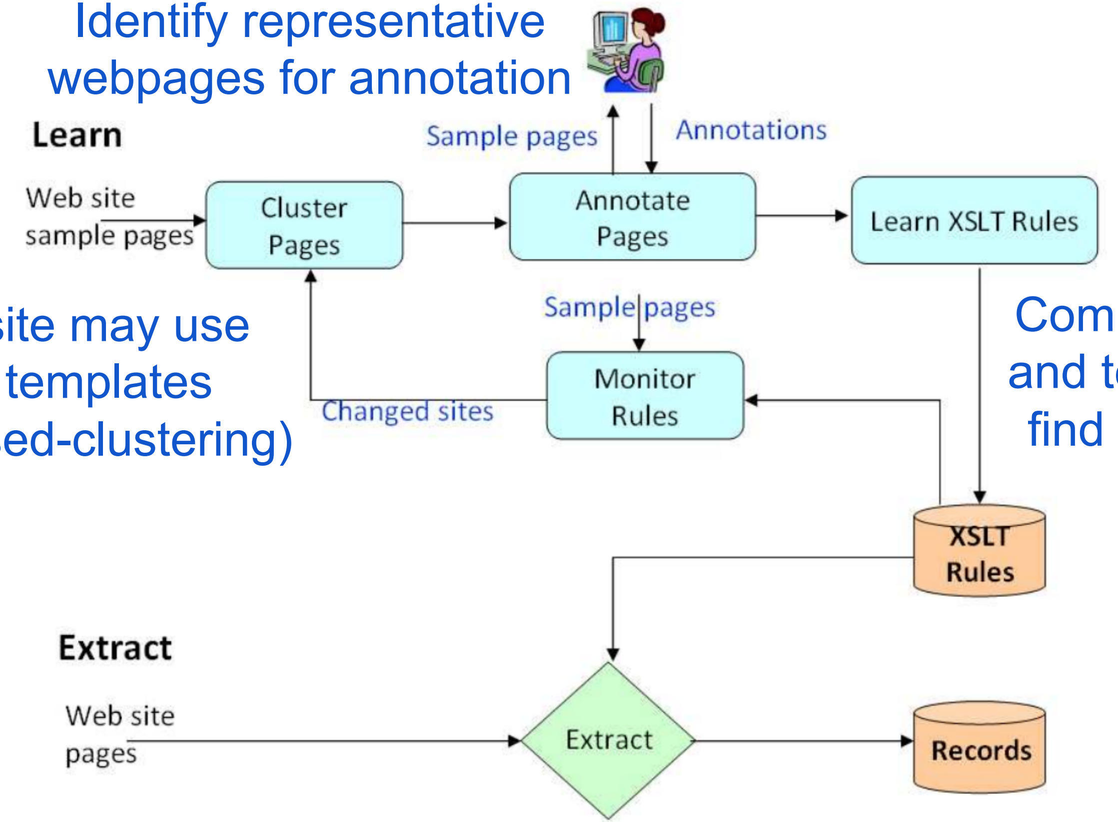
```
/html/body/div[1]/div/div[4]/div[3]/div[3]/div[3]/div[3]/div[4]/div[26]/b/a

/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]

Identify representative webpages for annotation



One website may use multiple templates (Unsupervised-clustering)

Combine attr features and textual features to find a general XPath (LR)

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

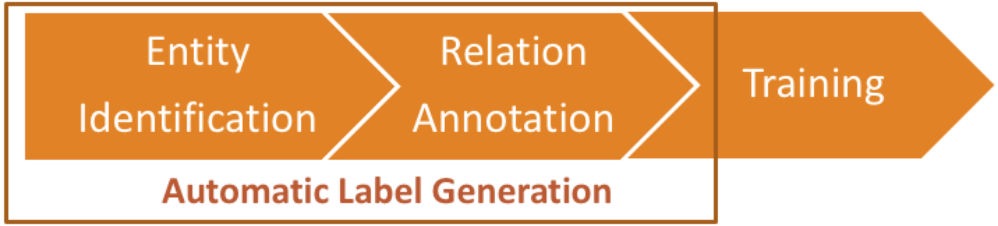
(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

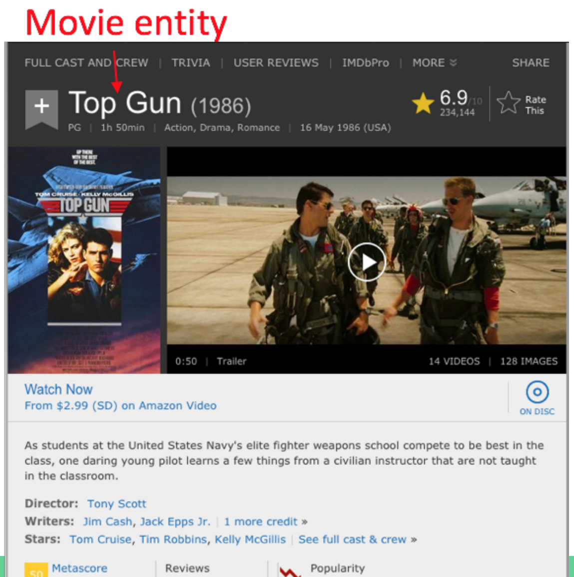
(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

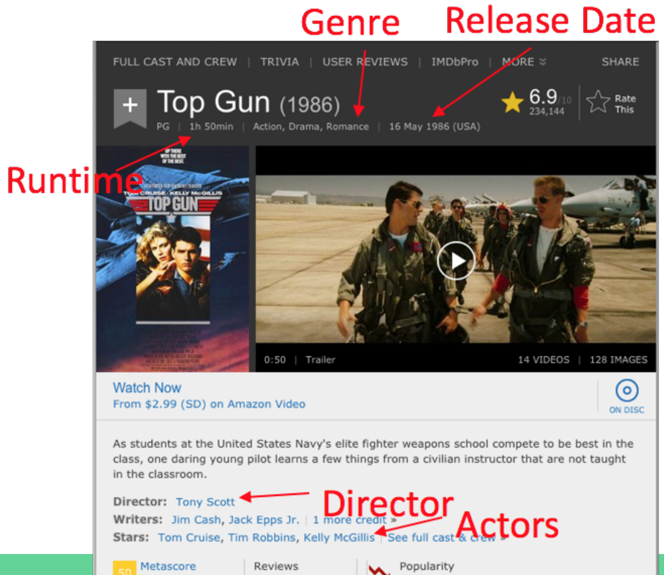
For negative examples, sample unrelated pairs of entities.

[Adapted example from Luke Zettlemoyer]

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







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

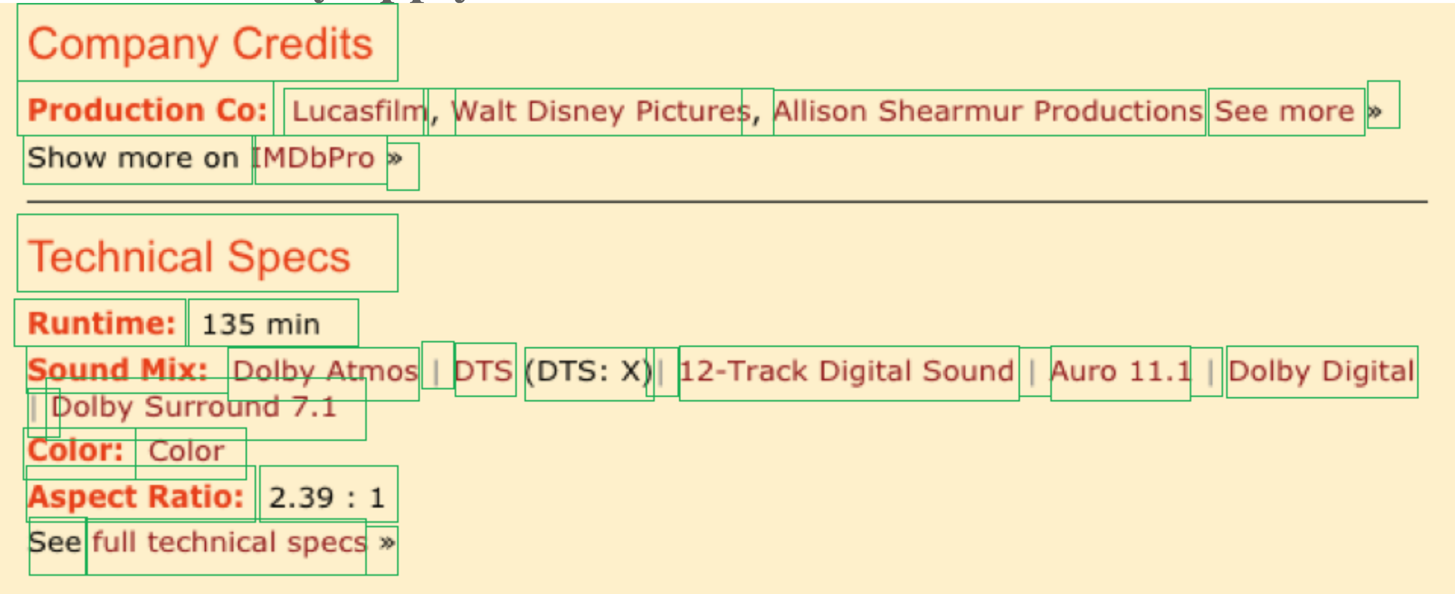
#Websites / #Webpages	33 / 434K
Language	English and 6 other languages
Domains	Animated films, Documentary films, Financial performance, etc.
# Annotated pages	70K (16%)
Annotated : Extracted #entities	1 : **2.6**
Annotated : Extracted #triples	1 : **3.0**
# Extractions	1.25 M
Precision	90%

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

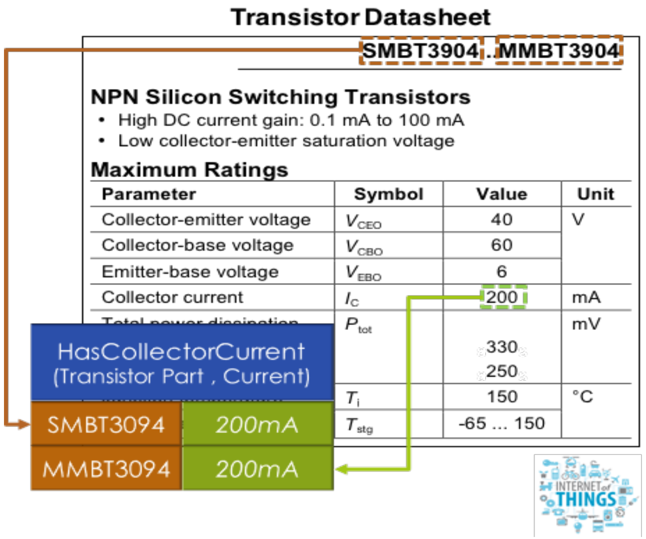
- **Which model is the best?**
  - Logistic regression: best results (20K features on one website)
  - Random forest: lower precision and recall
  - Deep learning??

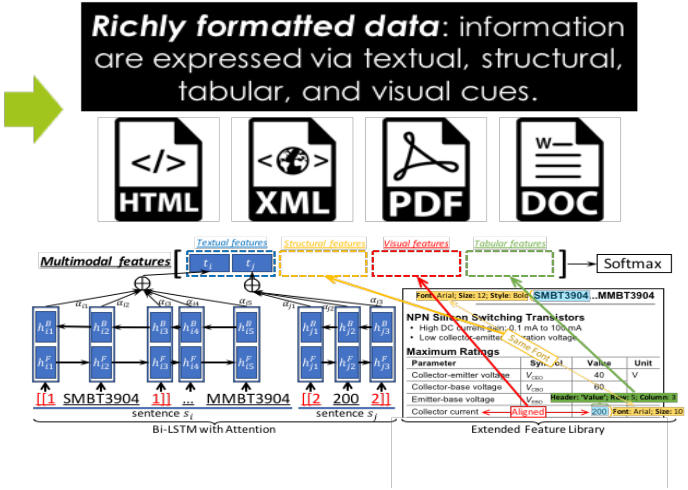
Challenges in Applying Deep Learning on Extracting Semi-structured Data

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



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





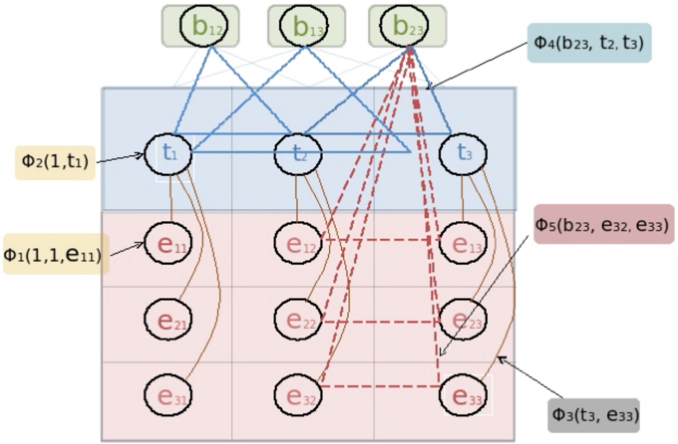
Fonduer combines a new **bi-directional LSTM with multimodal features** and weak supervision (specifically **data programming**).

Attend the talk in Research Session 13!

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

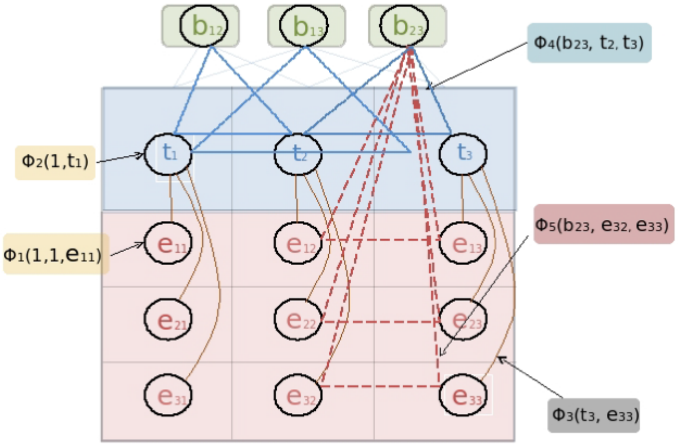
WebTable Extraction [Limaye et al., VLDB'10]

- 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**