

---

# **Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web**

Colin Lockard (UW, Amazon), Prashant Shiralkar (Amazon),  
**Xin Luna Dong** (Amazon), Hannaneh Hajishirzi (UW, AI2)

---

<https://sites.google.com/view/acl-2020-multi-modal-ie>

---



PAUL G. ALLEN SCHOOL  
OF COMPUTER SCIENCE & ENGINEERING

# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- Part 1b: Unstructured Text: Methods (10 minutes)
- **Live Q&A & Discussion (15 minutes)**
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- Part 3: Multi-modal Extraction (30 minutes)
- Conclusion (5 minutes)
- **Live Q&A & Discussion (15 minutes)**

Please ask questions over  
RocketChat!

**# tutorial-2**

We will monitor and respond  
during our presentation!

---

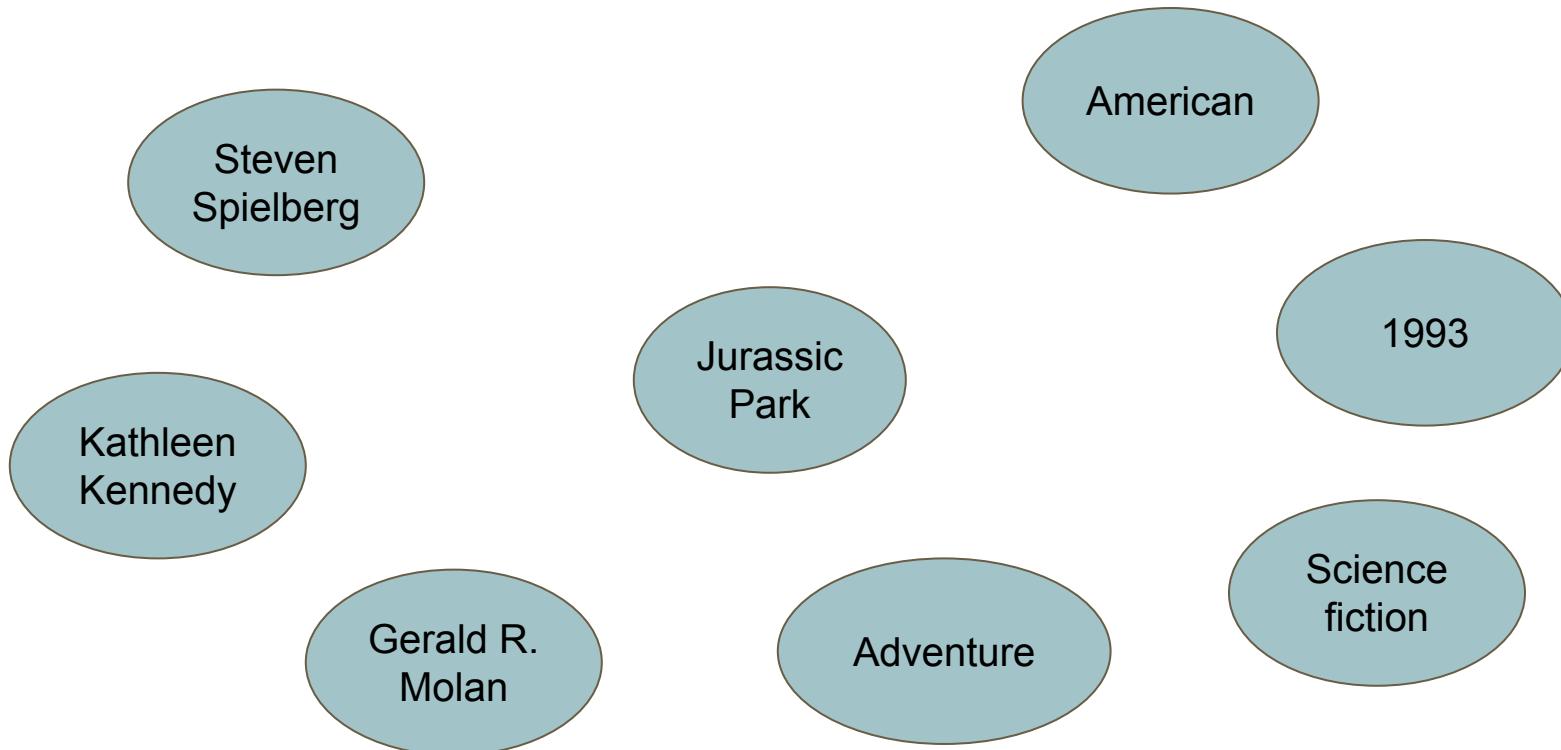
---

# What is Knowledge Graph?

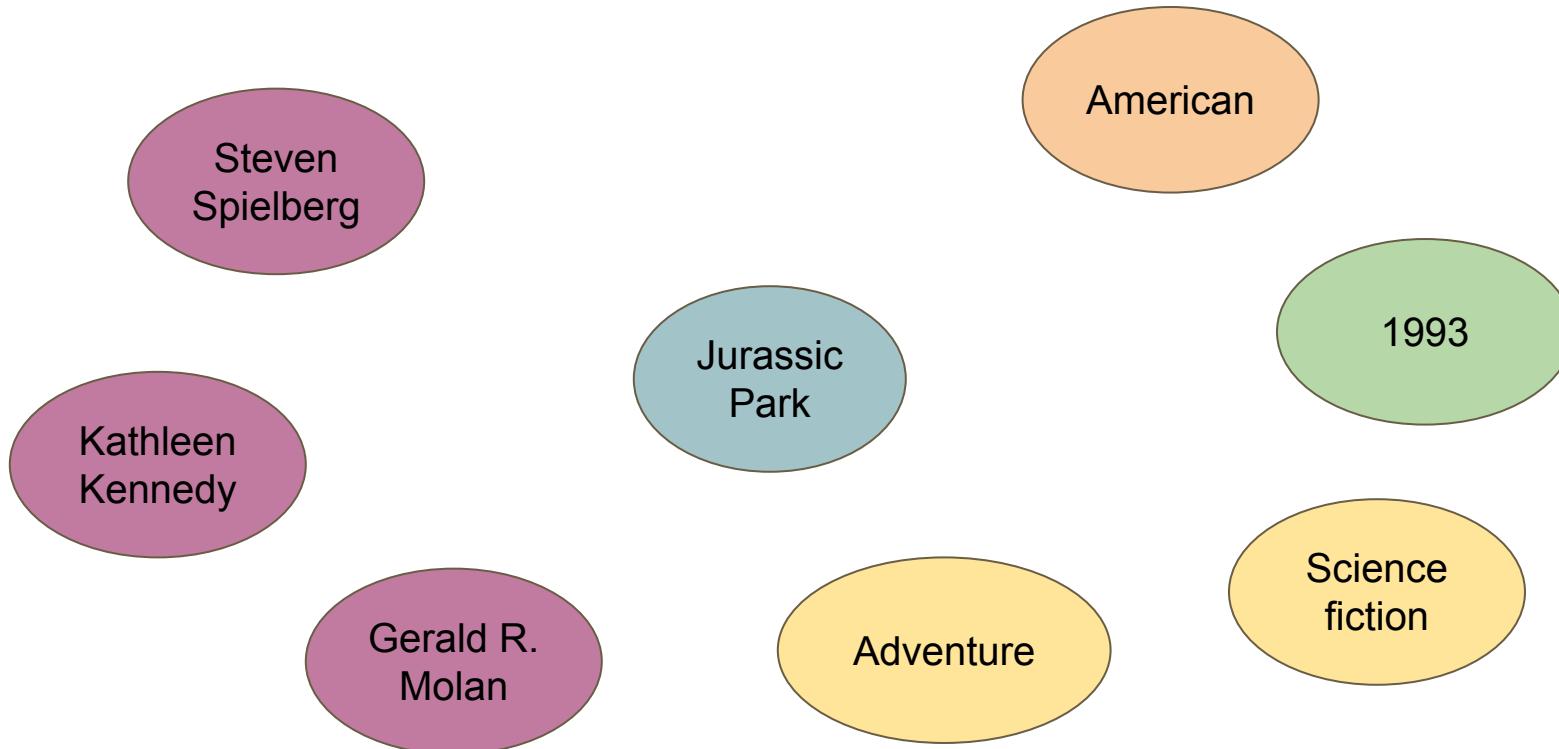
---

---

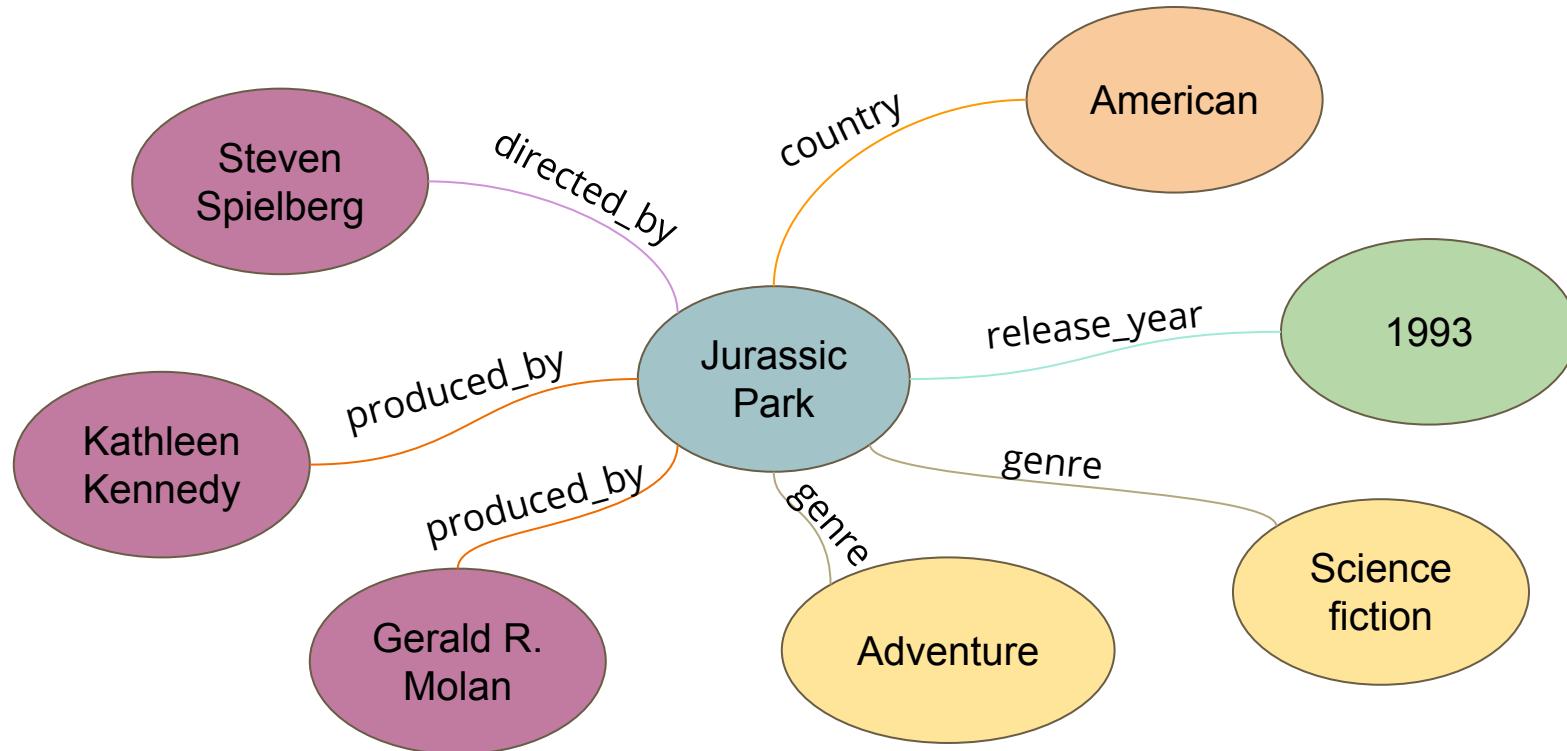
# Knowledge graph: entities and relationships



# Knowledge graph: entities and relationships



# Knowledge graph: entities and relationships



# Application 1. Web search

andrew mccallum

All News Images Videos Shopping More Settings Tools

About 4,160,000 results (0.68 seconds)

people.cs.umass.edu › ~mccallum

**Andrew McCallum Homepage**

Research. The main goal of my research is to dramatically increase our ability to mine actionable knowledge from unstructured text. I am especially interested in ...

Bio · Publications · Personal · Teaching

You've visited this page 2 times. Last visit: 7/3/19

people.cs.umass.edu › ~mccallum › pubs

**Andrew McCallum Publications**

Ari Kobren, Nicholas Monath, Andrew McCallum. Automated Knowledge Base Construction (AKBC), 2019. The Materials Science Procedural Text Corpus: ...

scholar.google.com › citations

**Andrew McCallum - Google Scholar Citations**

A comparison of event models for naive bayes text classification. A McCallum, K Nigam. AAAI-98 workshop on learning for text categorization 752 (1), 41-48, ...

Images for andrew mccallum

Andrew McCallum

Professor



Andrew McCallum is a professor and researcher in the computer science department at University of Massachusetts Amherst. His primary specialties are in machine learning, natural language processing, information extraction, information integration, and social network analysis. [Wikipedia](#)

**Place of birth:** Massachusetts

**h-index:** 106

**Known for:** Conditional random field

**Education:** Dartmouth College, University of Rochester

**Awards:** Best 10-year Paper Award of the ICML

## Books



An  
Introducti...  
to Condit...



Creativity  
and  
Learning...  
2012



The  
Complete  
Citizens...  
2011

## Publications

A comparison of event models for naive bayes text...

A McCallum, K Nigam  
AAAI-98 workshop on learning for  
text categorization

Text classification from label  
and unlabeled documents us  
K Nigam, AK McCallum, S Thrun,  
Mitchell  
Machine learning

# Application 2: Question answering

Alexa, who are the keynote speakers at this year's ACL?

This year's keynotes are from Kathleen McKeown and Josh Tenenbaum



# Application 3: Recommendation



# Application 3: Recommendation

Which char more important?



What would people ask about boy's clothing



Trend of Darth Vader lamps?

Why people who bought this lamp also bought this chair?



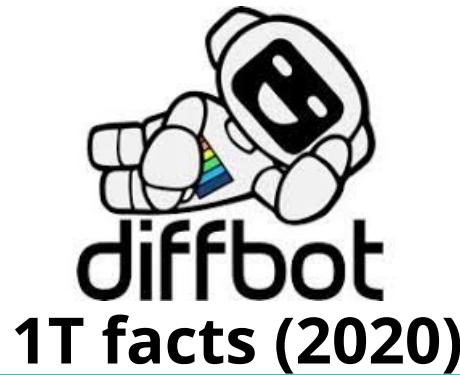
# Industry knowledge graphs



500B facts (2020)



50B facts (2018)



---

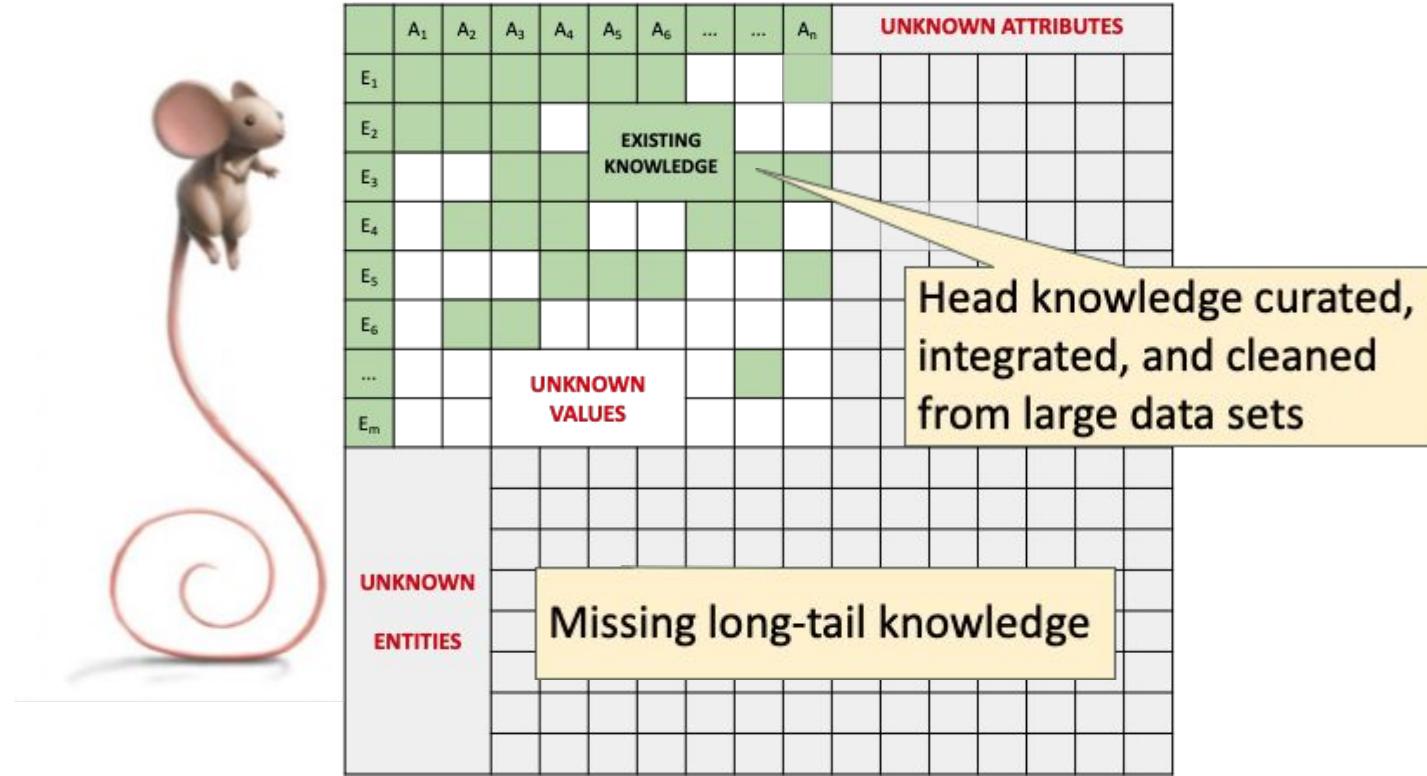
---

# Why Web-Scale Knowledge Collection?

---

---

# Still Missing A Lot of Long-Tail Knowledge



# Still Missing A Lot of Long-Tail Knowledge

	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	A <sub>6</sub>	...	...	A <sub>n</sub>	UNKNOWN ATTRIBUTES
E <sub>1</sub>										
E <sub>2</sub>										
E <sub>3</sub>										
E <sub>4</sub>										
E <sub>5</sub>										
E <sub>6</sub>										
...										
E <sub>m</sub>										
UNKNOWN VALUES										
UNKNOWN ENTITIES										

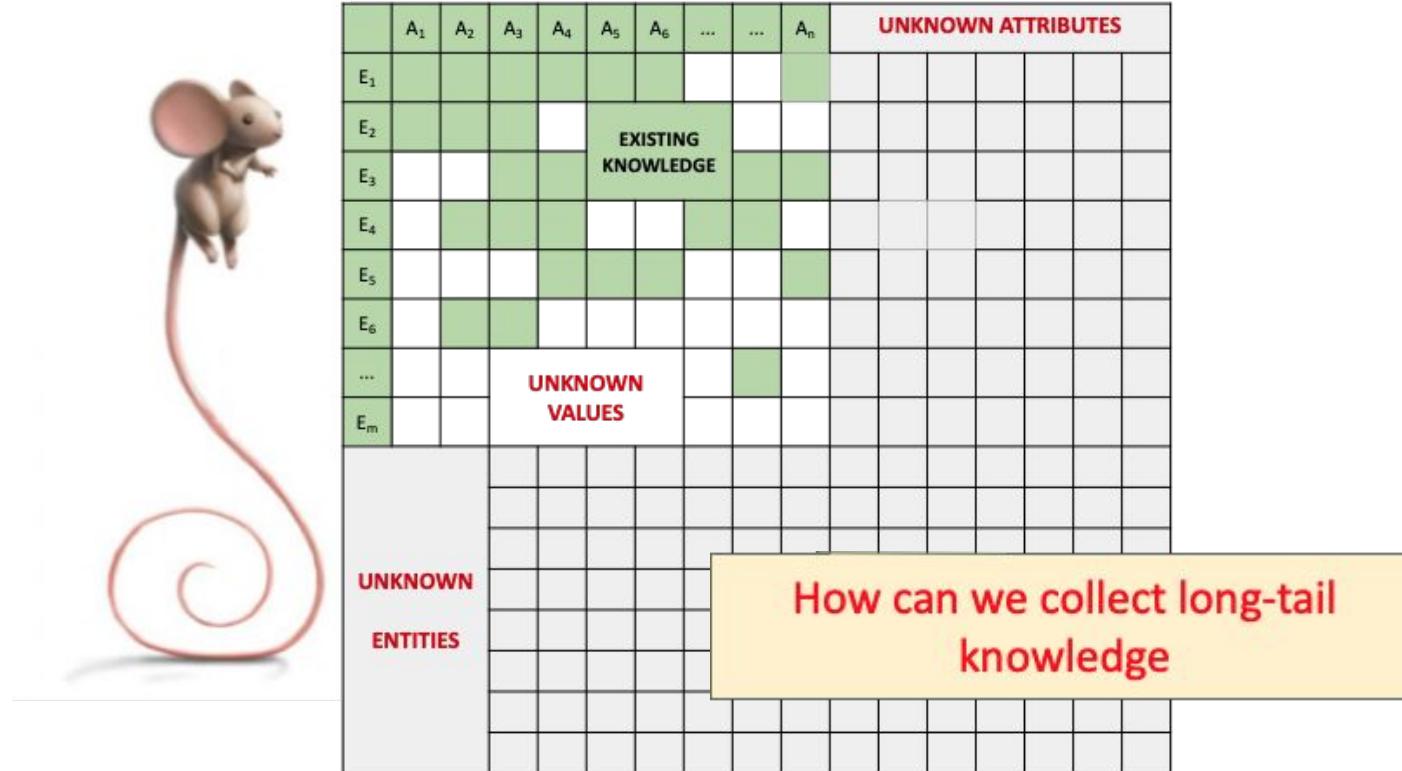
- Alexa, when did Van Gogh live in Paris?  
- Sorry, I'm not sure.

- Alexa, tell me the recent movies by Ziyi Zhang  
- Sorry, I don't know that

- Alexa, does Taylor Swift have a pet?  
- Yes, Taylor Swift has at least one nickname

- Alexa, which body part does the lotus position in yoga stretch?  
- Here's something I found on Wikipedia: Lotus position is...

# Still Missing A Lot of Long-Tail Knowledge



# How can we take advantage of the vast quantity of information on the web and convert it into useful information?

Serena Williams

USA Plays: Right Turned Pro: 1995

WT Ranking #9 Birth Date September 26, 1981 (Age: 38)

Hometown Saginaw, MI, USA Height 5'9 Weight 154 lbs.

Player Profile Results Videos Photos

Serena Williams Tournaments Year: 2020

2020 Stats Prize Money \$46,600 Singles Titles 1

2020 Tournaments Australian Open - Melbourne, Australia January 19, 2020 to February 1, 2020

ROUND	OPPONENT	WOMEN
1st	Anastasija Potapova	
2nd	Tamara Zidansek	

ASB Classic 2020 - Auckland, New Zealand January 5, 2020 to January 11, 2020

ROUND	OPPONENT	WOMEN
1st	Svetlana Kuznetsova	
2nd	Christina McHale	

WikiPEDIA The Free Encyclopedia

Ada Lovelace

From Wikipedia, the free encyclopedia

Augusta Ada King, Countess of Lovelace (née Byron; 10 December 1815 – 27 November 1852) was an English polymath and writer, chiefly known for her work on Charles Babbage's proposed mechanical general-purpose computer, the Analytical Engine. She was the first to recognise that the machine could be used for calculations other than those intended by its designer, and published the first algorithm intended to be carried out by such a machine.<sup>[1]</sup> Augusta Lovelace was the only legitimate child of poet Lord Byron and his wife Anne Isabella Milner. She was born out of wedlock to other women.<sup>[2]</sup> Byron separated from his wife a month before Ada was born.<sup>[3]</sup> Ada was born in London, England, and was brought up in France, where she spent most of her childhood. Her mother died when Ada was two years old. Her mother remained bitter and promoted Ada's interest in mathematics, developing her father's perceived insanity. Despite this, Ada remained interested in poetry and literature throughout her life. Upon her eventual death, she was buried next to him at her request. Ada pursued her studies assiduously. She married William King in 1835. King was made a baronet in 1851, and Ada became Countess of Lovelace.

Her educational and social exploits brought her into contact with scientists such as David Brewster, Charles Wheatstone, Michael Faraday and the author Charles Dickens, contacts which she used to further her education. Ada described her approach as "poetical science"<sup>[4]</sup> and herself as an "Analyst (& Metaphysician)".<sup>[5]</sup> When she was a teenager, her mathematical talents led her to a long working relationship and friendship with fellow British mathematician Charles Babbage, who is known as "the father of computers". She was in particular interested in Babbage's work on the Analytical Engine. Lovelace first met him in June 1833, through their mutual friend, and her private tutor, Mary

PAUL G. ALLEN SCHOOL  
OF COMPUTER SCIENCE & ENGINEERING

About Us Contact Us

NEWS & EVENTS PEOPLE ACADEMICS RESEARCH & INNOVATION OUTREACH SUPPORT #UWALLEN

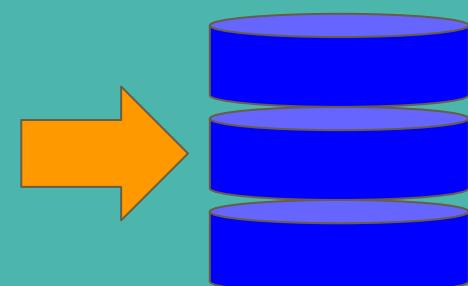
## Magdalena Balazinska and Paul Beame named Fellows of the ACM

The Association for Computing Machinery honored Balazinska for her contributions to scalable distributed data systems, and Beame for his contributions in computational and proof complexity and for service to the computing community.

More

Born The Hon. Augusta Ada Byron  
10 December 1815  
London, England

Died 27 November 1852 (aged 36)  
Marylebone, London, England





## PROGRAM DETAILS

Keynote Speakers  
Tutorials  
Workshops  
Accepted Papers

## Tutorials

**NEW:** Checkout the new blogpost by tutorial chairs describing [detailed modalities of ACL2020 tutorials.](#)

The following tutorials will be held on Sunday, July 5th, 2020. Checkout [this blog post](#) by tutorial chairs to learn more about the new virtual format of tutorials.

### Tutorial Schedule

Note that the time slots are mentioned in Seattle, Pacific Daytime Time (UTC/GMT-7). Some tutorials are held twice and some are held once.

#### 6:00 AM to 9:30 AM (Seattle Time)

[T1](#): Interpretability and Analysis in Neural NLP  
[T5](#): Achieving Common Ground in Multi-modal Dialogue  
[T7](#): Integrating Ethics into the NLP Curriculum

#### 10:30 AM to 2:00 PM (Seattle Time)

[T1](#): Interpretability and Analysis in Neural NLP  
[T3](#): Reviewing Natural Language Processing Research  
[T4](#): Stylized Text Generation: Approaches and Applications  
[T7](#): Integrating Ethics into the NLP Curriculum

#### 3:00 PM to 6:30 PM (Seattle Time)

[T2](#): Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web  
[T5](#): Achieving Common Ground in Multi-modal Dialogue  
[T6](#): Commonsense Reasoning for Natural Language Processing  
[T8](#): Open-Domain Question Answering

#### On this page

##### Tutorial Schedule

6:00 AM to 9:30 AM (Seattle Time)

10:30 AM to 2:00 PM (Seattle Time)

3:00 PM to 6:30 PM (Seattle Time)

##### Tutorial Details

T1: Interpretability and Analysis in Neural NLP (cutting-edge)

T2: Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web (Cutting-edge)

T3: Reviewing Natural Language Processing Research (Introductory)

T4: Stylized Text Generation: Approaches and Applications (Cutting-edge)

T5: Achieving Common Ground in Multi-modal Dialogue (Cutting-edge)

T6: Commonsense Reasoning for Natural Language Processing (Introductory)

T7: Integrating Ethics into the NLP Curriculum (Introductory)

T8: Open-Domain Question Answering (Cutting-edge)



57<sup>th</sup>

Advances in Language  
Learning  
Tg: Storytelling  
ANNUAL MEETING  
July 28<sup>th</sup> - August 2<sup>nd</sup>  
Florence  
of the Association for Computational Linguistics

[f FACEBOOK](#)

[Twitter](#)

[INSTAGRAM](#)

[CHAIRS BLOG](#)

[HOMEPAGE](#)

[CHECK THE PROGRAM](#)

[COMMITTEES](#)

[CALL FOR PAPERS](#)

[CALL FOR NOMINATIONS](#)

[NOMINATIONS FOR ACL 2019 BEST PAPER AWARDS](#)

[WINNERS OF ACL 2019 BEST PAPER AWARDS](#)

[TUTORIALS](#)

[INSTRUCTIONS FOR REVIEWERS](#)

[INSTRUCTIONS FOR PRESENTERS](#)

## SUNDAY JULY 28TH 2019 - MORNING

### T1: Latent Structure Models for Natural Language Processing

André F. T. Martins, Tsvetomila Mihaylova, Nikita Nangia and Vlad Niculae

9. HALL 1 + HALL 3 Tutorial Materials

Latent structure models are a powerful tool for modeling compositional data, discovering linguistic structure, and building NLP pipelines. They are appealing for two main reasons: they allow incorporating structural bias during training, leading to more accurate models; and they allow discovering hidden linguistic structure, which provides better interpretability.

This tutorial will cover recent advances in discrete latent structure models. We discuss their motivation, potential, and limitations, then explore in detail three strategies for designing such models: gradient approximation, reinforcement learning, and end-to-end differentiable methods. We highlight connections among all these methods, enumerating their strengths and weaknesses. The models we present and analyze have been applied to a wide variety of NLP tasks, including sentiment analysis, natural language inference, language modeling, machine translation, and semantic parsing.

Examples and evaluation will be covered throughout. After attending the tutorial, a practitioner will be better informed about which method is best suited for their problem.

Past tutorials - Admin Wiki

aclweb.org/adminwiki/index.php?title=Past\_tutorials

Page Discussion Read View source View history Search



Main page  
Recent changes  
Random page  
Help  
  
Tools  
What links here  
Related changes  
Special pages  
Permanent link  
Page information

Print/export  
Create a book  
Download as PDF  
Printable version

## Past tutorials

This page belongs to the [tutorial chair handbook](#). It summarizes data on tutorials which took place at some recent ACL, EACL, NAACL, EMNLP and COLING conferences.

### Contents [hide]

- 1 2019 tutorials
- 2 2018 tutorials
- 3 2017 tutorials
- 4 2016 tutorials

### 2019 tutorials

Title	Trainers	Conference	Conference link	ACL Anthology link
Latent Structure Models for Natural Language Processing	André F. T. Martins, Tsvetomila Mihaylova, Nikita Nangia and Vlad Niculae	ACL 2019	[1] ↗	
Graph-Based Meaning Representations: Design and Processing	Alexander Koller, Stephan Oepen and Weiwei Sun	ACL 2019	[2] ↗	
Discourse Analysis and Its Applications	Shafiq Joty, Giuseppe Carenini, Raymond Ng and Gabriel Murray	ACL 2019	[3] ↗	
Computational Analysis of Political Texts: Bridging Research Efforts Across Communities	Goran Glavaš, Federico Nanni and Simone Paolo Ponzetto	ACL 2019	[4] ↗	
Wikipedia as a Resource for Text Analysis and Retrieval	Marius Pasca	ACL 2019	[5] ↗	
Deep Bayesian Natural Language Processing	Jen-Tzung Chien	ACL 2019	[6] ↗	



# Scalable Construction and Reasoning of Massive Knowledge Bases

Xiang Ren, Nanyun Peng, William Yang Wang

## Abstract

In today's information-based society, there is abundant knowledge out there carried in the form of natural language texts (e.g., news articles, social media posts, scientific publications), which spans across various domains (e.g., corporate documents, advertisements, legal acts, medical reports), which grows at an astonishing rate. Yet this knowledge is mostly inaccessible to computers and overwhelming for human experts to absorb. How to turn such massive and unstructured text data into structured, actionable knowledge, and furthermore, how to teach machines learn to reason and complete the extracted knowledge is a grand challenge to the research community. Traditional IE systems assume abundant human annotations for training high quality machine learning models, which is impractical when trying to deploy IE systems to a broad range of domains, settings and languages. In the first part of the tutorial, we introduce how to extract structured facts (i.e., entities and their relations for types of interest) from text corpora to construct knowledge bases, with a focus on methods that are weakly-supervised and domain-independent for timely knowledge base construction across various application domains. In the second part, we introduce how to leverage other knowledge, such as the distributional statistics of characters and words, the annotations for other tasks and other domains, and the linguistics and problem structures, to combat the problem of inadequate supervision, and conduct low-resource information extraction. In the third part, we describe recent advances in knowledge base reasoning. We start with the gentle introduction to the literature, focusing on path-based and embedding based methods. We then describe DeepPath, a recent attempt of using deep reinforcement learning to combine the best of both worlds for knowledge base reasoning.

PDF

BibTeX

Search

Video

**Anthology ID:** N18-6003

**Volume:** Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts

**Month:** June

NeurIPS | NIPS | NIPS | NIPS 2016 Sched | NIPS 2015 Schedule

nips.cc/C... nips.cc/Conferences/2015/Schedule?type=Tutorial

Login Login Login Search Schedule Filter Day Filtering for

NeurIPS | NeurIPS | NeurIPS | NeurIPS | 2016 NeurIPS | 2015

Thirty-third Conference on Neural Information Processing Systems

Thirty-second Conference on Neural Information Processing Systems

Thirty-first Conference on Neural Information Processing Systems

Thirtieth Conference on Neural Information Processing Systems

Twenty-ninth Conference on Neural Information Processing Systems

Year (2019) ▾

Help ▾

My Registrations

Profile ▾

Contact NeurIPS

Sponsor Info

Publications

Future Meetings

Diversity & Inclusion

Code of Conduct

About Us

Press

News

Board 2019

Year (2018) ▾

Help ▾

My Registrations

Profile ▾

Contact NeurIPS

Sponsor Info

Publications

Future Meetings

Diversity & Inclusion

Code of Conduct

About Us

Press

News

Year (2017) ▾

Help ▾

My Registrations

Profile ▾

Contact NeurIPS

Sponsor Info

Publications

Future Meetings

Diversity & Inclusion

Code of Conduct

About Us

Press

News

Year (2016) ▾

Help ▾

My Registrations

Profile ▾

Contact NeurIPS

Sponsor Info

Publications

Future Meetings

Diversity & Inclusion

Code of Conduct

About Us

Press

News

Year (2015) ▾

Help ▾

My Registrations

Profile ▾

Contact NeurIPS

Sponsor Info

Publications

Future Meetings

Diversity & Inclusion

Code of Conduct

About Us

Press

News

Dates Calls Program Books Schedule Committees

Mon Dec 7th 09:30 - 11:30 AM @ Level 2 room 210 AB  
**Deep Learning**  
Geoffrey E Hinton · Yoshua Bengio · Yann LeCun  
[Slides »](#) [Slides \(vision\) »](#)

Mon Dec 7th 09:30 - 11:30 AM @ Level 2 room 210 E,F  
**Large-Scale Distributed Systems for Training Neural Networks**  
Jeff Dean · Oriol Vinyals  
[Slides »](#)

Mon Dec 7th 01:00 - 03:00 PM @ Level 2 room 210 AB  
**Monte Carlo Inference Methods**  
Iain Murray  
[Slides »](#)

Mon Dec 7th 01:00 - 03:00 PM @ Level 2 room 210 E,F  
**Probabilistic Programming**  
Frank Wood  
[Slides »](#)

Mon Dec 7th 03:30 - 05:30 PM @ Level 2 room 210 AB  
**Introduction to Reinforcement Learning with Function Approximation**  
Richard S Sutton  
[Slides »](#)

21

NIPS 2006 Schedule

nips.cc/Conferences/2006/Schedule?type=Tutorial

Login Search Schedule Filter Day Filtering for Tutorials

# NeurIPS | 2006

Twentieth Conference on Neural Information Processing Systems

Year (2006) ▾

Help ▾

My Registrations

Profile ▾

Contact NeurIPS

Sponsor Info

Publications

Future Meetings

Diversity & Inclusion

Code of Conduct

About Us

Press

News

Board 2019

6

Program Highli

Mon Dec 4th 09:30 - 11:30 AM @ Regency F

Tutorial

## Machine Learning for Natural Language Processing: New Developments and Challenges

Dan Klein

[Slides \(PDF\)](#) » [Part 2 - QuickTime Movie \(900x600\)](#) » [Part 1 - QuickTime Movie \(900x600\)](#) » [Part 1 - QuickTime Movie \(640x480\)](#) »  
[Part 2 - QuickTime Movie \(320x240\)](#) » [Part 1 - QuickTime Movie \(320x240\)](#) » [Part 2 - QuickTime Movie \(640x480\)](#) »

Mon Dec 4th 09:30 - 11:30 AM @ Regency E

Tutorial

## Advances in Gaussian Processes

Carl Rasmussen

[Slides \(PDF\)](#) » [QuickTime Movie \(900x600\)](#) » [QuickTime Movie \(640x480\)](#) » [QuickTime Movie \(320x240\)](#) »

Mon Dec 4th 01:00 - 03:00 PM @ Regency F

Tutorial

## The Role of Computational Methods in Creating a Systems Level View from Biological Data

Maya Schuldiner · Nir Friedman

[Slides \(PowerPoint\)](#) » [Slides \(PDF\)](#) » [QuickTime Movie \(900x600\)](#) » [QuickTime Movie \(640x480\)](#) » [QuickTime Movie \(320x240\)](#) »

Mon Dec 4th 01:02 - 03:00 PM @ Regency E

Tutorial

## Bayesian Models of Human Learning and Inference

Josh Tenenbaum

[Slides \(PowerPoint\)](#) » [QuickTime Movie \(900x600\)](#) » [QuickTime Movie \(640x480\)](#) » [QuickTime Movie \(320x240\)](#) »

Mon Dec 4th 03:30 - 05:30 PM @ Regency F

Tutorial

## Energy-Based Models: Structured Learning Beyond Likelihoods

Yann LeCun

[Slides \(D|Vu\)](#) » [Slides \(PDF\)](#) » [QuickTime Movie \(900x600\)](#) » [QuickTime Movie \(640x480\)](#) » [QuickTime Movie \(320x240\)](#) »

Mon Dec 4th 03:30 - 05:30 PM @ Regency E

Tutorial

## Diffusion Tensor Imaging and Fiber Tracking of Human Brain Pathways

Brian A Wandell

[Slides \(PDF\)](#) » [QuickTime Movie \(900x600\)](#) » [QuickTime Movie \(640x480\)](#) » [QuickTime Movie \(320x240\)](#) »

**NeurIPS | 2019**  
Thirty-third Conference on Neural Information Processing Systems

**Year (2019) ▾**

- Help ▾
- My Registrations
- Profile ▾
- Contact NeurIPS
- Sponsor Info
- Publications
- Future Meetings
- Diversity & Inclusion
- Code of Conduct
- About Us
- Press
- News
- Board 2019

**Past tutorials**

This page belongs to the [tutorial](#) COLING conferences.

**Contents [hide]**

- 1 2019 tutorials
- 2 2018 tutorials
- 3 2017 tutorials
- 4 2016 tutorials

**2019 tutorials**

- Title
- Latent Structure Models for Natural Language Processing
- Graph-Based Meaning Representation Processing
- Discourse Analysis and Its Applications
- Computational Analysis of Political Research Efforts Across Communities
- Wikipedia as a Resource for Text Retrieval
- Deep Bayesian Natural Language Processing

**ACL Anthology**

## Scalable Construction and Reasoning of Massive Knowledge Bases

Xiang Ren, Nanyun Peng, William Yang Wang

### Abstract

In today's information-based society, there is abundant knowledge out there carried in the form of natural language texts (e.g., news articles, social media posts, scientific publications), which spans across various domains (e.g., corporate documents, advertisements, legal acts, medical reports), which grows at an astonishing rate. Yet this knowledge is mostly inaccessible to computers and overwhelming for human experts to absorb. How to turn such massive and unstructured text data into structured, actionable knowledge, and furthermore, how to teach machines learn to reason and complete the extracted knowledge is a grand challenge to the research community. Traditional IE systems assume abundant human annotations for training high quality machine learning models, which is impractical when trying to deploy IE systems to a broad range of domains, settings and languages. In the first part of the tutorial, we introduce how to extract structured facts (i.e., entities and their relations for types of interest) from text corpora to construct knowledge bases, with a focus on methods that are weakly-supervised and domain-independent for timely knowledge base construction across various application domains. In the second part, we introduce how to leverage other knowledge, such as the distributional statistics of characters and words, the annotations for other tasks and other domains, and the linguistics and problem structures, to combat the problem of inadequate supervision, and conduct low-resource information extraction. In the third part, we describe recent advances in knowledge base reasoning. We start with the gentle introduction to the literature, focusing on path-based and embedding based methods. We then describe DeepPath, a recent attempt of using deep reinforcement learning to combine the best of both worlds for knowledge base reasoning.

**Anthology ID:** N18-6003  
**Volume:** Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts  
**Month:** June

**PDF**

**BibTeX**

**Search**

**Video**

# What Is Unstructured Text?

## *Jurassic Park* (film)

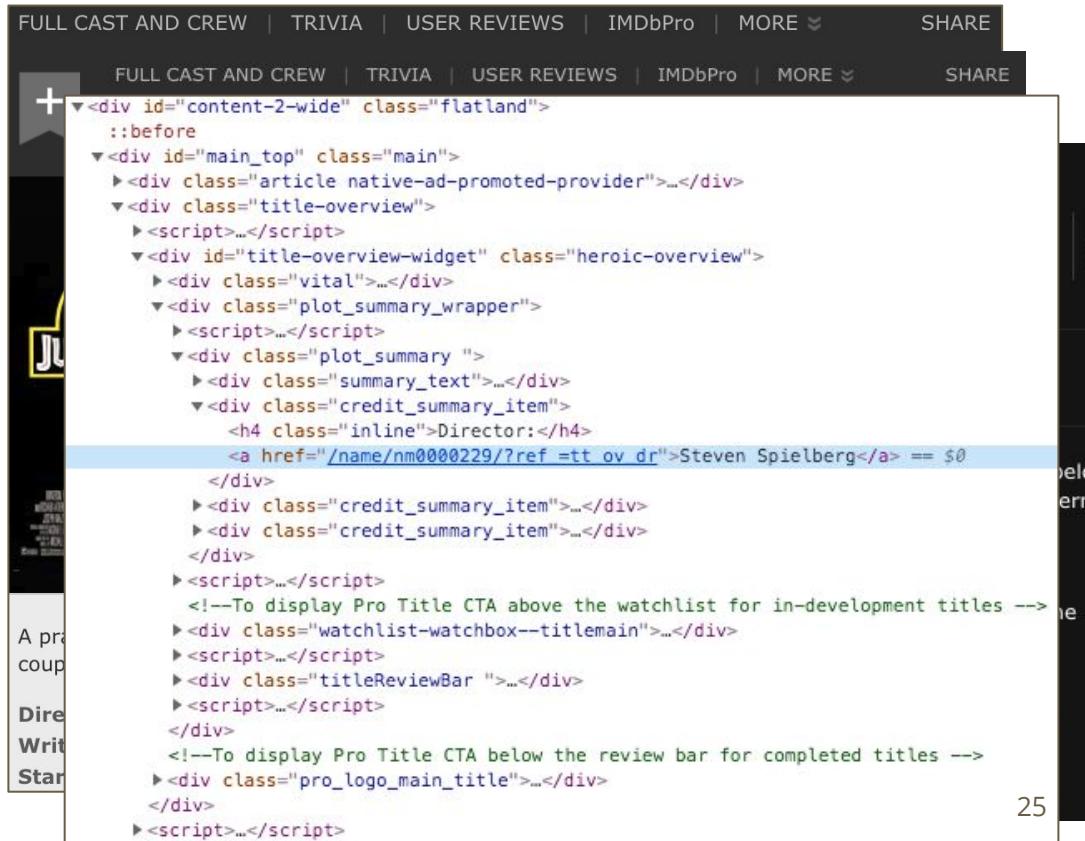
From Wikipedia, the free encyclopedia

*This article is about the 1993 film. For the franchise, see [Jurassic Park](#). For other uses, see [Jurassic Park](#).*

**Jurassic Park** is a 1993 American science fiction adventure film directed by [Steven Spielberg](#) and produced by [Kathleen Kennedy](#) and [Gerald R. Molen](#). It is the first installment in the [Jurassic Park](#) franchise, and is based on the [1990](#) novel [of the same name](#) by [Michael Crichton](#) and a screenplay written by Crichton and [David Koepp](#). The film is set on the fictional island of [Isla Nublar](#), located off Central America's Pacific Coast near [Costa Rica](#). There, billionaire philanthropist John Hammond and a small team of genetic scientists have created a [wildlife park](#) of [de-extinct dinosaurs](#). When industrial sabotage leads to a catastrophic

# What Is Semi-structured Text?

- Consistent layout/template
- Facts in specific position
  - (or specific relative to some constant piece of text)



The image shows a screenshot of the movie 'Jaws' on the IMDb website. The page includes navigation links like 'FULL CAST AND CREW', 'TRIVIA', 'USER REVIEWS', 'IMDbPro', and 'MORE'. Below the header, there's a large image of the movie poster. To the right of the poster, there's a sidebar with sections for 'Awards', 'Casts', 'Couples', 'Directors', 'Writers', and 'Stars'. The main content area displays the movie's title, plot summary, and credits. A blue box highlights a specific line of code in the page source: '[Steven Spielberg](/name/nm0000229/?ref_=tt_ov_dr) == \$0'. The page source code is displayed as a tree structure with various HTML tags and class names.

```
FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE SHARE  
FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE SHARE  
+<div id="content-2-wide" class="flatland">  
  ::before  
  <div id="main_top" class="main">  
    <div class="article native-ad-promoted-provider">...</div>  
    <div class="title-overview">  
      <script>...</script>  
      <div id="title-overview-widget" class="heroic-overview">  
        <div class="vital">...</div>  
        <div class="plot_summary_wrapper">  
          <script>...</script>  
          <div class="plot_summary ">  
            <div class="summary_text">...</div>  
            <div class="credit_summary_item">  
              <h4 class="inline">Director:</h4>  
              <a href="/name/nm0000229/?ref_=tt_ov_dr">Steven Spielberg</a> == $0  
            </div>  
            <div class="credit_summary_item">...</div>  
            <div class="credit_summary_item">...</div>  
          </div>  
          <script>...</script>  
          <!--To display Pro Title CTA above the watchlist for in-development titles -->  
          <div class="watchlist-watchbox--titlemain">...</div>  
          <script>...</script>  
          <div class="titleReviewBar ">...</div>  
          <script>...</script>  
        </div>  
        <!--To display Pro Title CTA below the review bar for completed titles -->  
        <div class="pro_logo_main_title">...</div>  
      </div>  
      <script>...</script>
```

# What Is Tabular Text?

	Lake	Area
1	Windermere	5.69 sq mi (14.7 km <sup>2</sup> )
2	Kielder Reservoir	3.86 sq mi (10.0 km <sup>2</sup> )
3	Ullswater	3.44 sq mi (8.9 km <sup>2</sup> )
4	Bassenthwaite Lake	2.06 sq mi (5.3 km <sup>2</sup> )
5	Derwent Water	2.06 sq mi (5.3 km <sup>2</sup> )

(a) Relational Table

Government <sup>[3]</sup>	
• Type	Mayor–Council
• Body	New York City Council
• Mayor	Bill de Blasio (D)
Area <sup>[2]</sup>	
• Total	468.9 sq mi (1,214 km <sup>2</sup> )
• Land	304.8 sq mi (789 km <sup>2</sup> )
• Water	164.1 sq mi (425 km <sup>2</sup> )
• Metro	13,318 sq mi (34,490 km <sup>2</sup> )
Elevation <sup>[4]</sup>	
	33 ft (10 m)

(b) Entity Table

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

(c) Matrix Table

On web, defined by <table>, <tr>, <td> tags

# What is Information Extraction?

Information extraction is to identify facts from **documents** or **semi-structured form** and convert them into **structured form**.

Serena Williams

USA | Plays: Right | Turned Pro: 1995

WTA Rank #9  
Birth Date September 26, 1981 (Age: 38)  
Hometown Saginaw, MI, USA  
Height 5'9  
Weight 154 lbs.

Player Profile Results Videos Photos

Serena Williams T  
Year: 2020

2020 STATS  
PRIZE MONEY \$46,600 SINGLES T 1

2020 TOURNAMENTS  
\*AUSTRALIAN OPEN - Melbourne, January 19, 2020 to February 1, 2020

ROUND OPPONENT  
1st Anastasia Potapova  
2nd Tamara Zidansek

Interaction Help About Wikipedia Community portal Recent changes Contact page Tools What links here Related changes Special pages Permanent link Page information Cite this page

**NEWS**

Dec 29, 2004 - The Inkscape development team have set a goal for the release of version 0.41 for late January. There are some noteworthy new features that would be worth getting out to users, and with some major internal changes that will be taking place soon, a good stable release is needed prior to starting Dev 0.42 development.

Article Talk

Ada Lovelace

From Wikipedia, the free encyclopedia

Augusta Ada King, Countess of Lovelace (–) was the first computer programmer. She was the first to recognize the full potential of a "computer" and published the first algorithm intended to be run by a machine. Her work laid the foundation for modern computing.

Augusta Byron was the only legitimate child of Lord Byron and his wife, Anne Isabella Milner. She was born out of wedlock to other women, forever four months later. He commemorated ADA's sole daughter of my house and heart! years old. Her mother remained bitter and poor developing her father's perceived insanity. Des Gordon. Upon her eventual death, she was buried in a family vault.

Her education and social exploits brought her into contact with Charles Babbage, who became her mentor. Ada described her approach as:

When she was a teenager, her mathematical studies were encouraged by Charles Babbage, who knew work on the Analytical Engine. Lovelace first met

(Serena Williams, born in, Florida)

Subject

Object

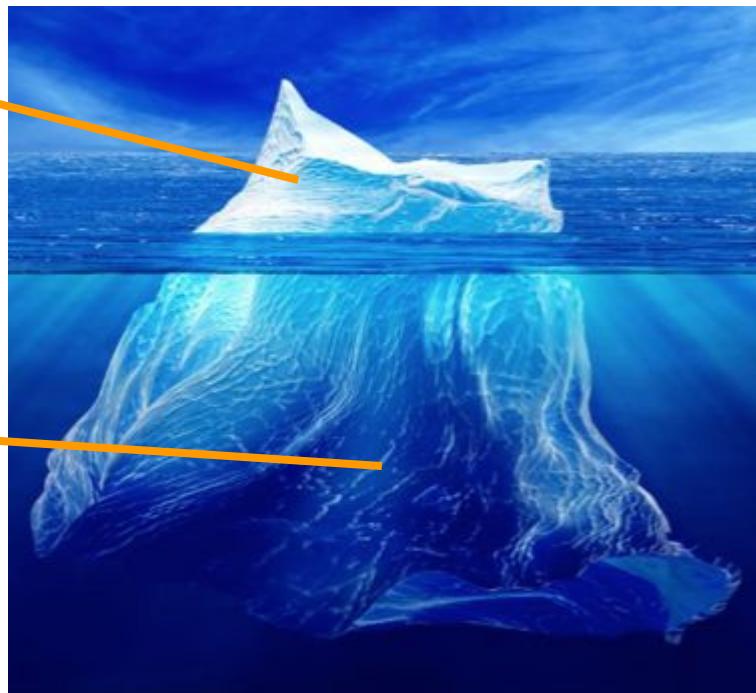
Predicate  
(Relation)

Knowledge Triple



# ClosedIE vs OpenIE

- ClosedIE: Known unknowns  
Align to existing attributes  
("Trump", place\_of\_birth, "USA")
- OpenIE: Unknown unknowns  
Not limited by existing attributes  
("Trump", "likes most", "Trump tower")



# Where Are We in Web-Scale Knowledge Extraction

- Collected mostly from a few web sources
- Automatic collection has fairly low precision and recall
- Cover only known unknowns
- Collected knowledge cannot be easily aligned w. existing knowledge



---

---

# Why Is This Hard?

---

---

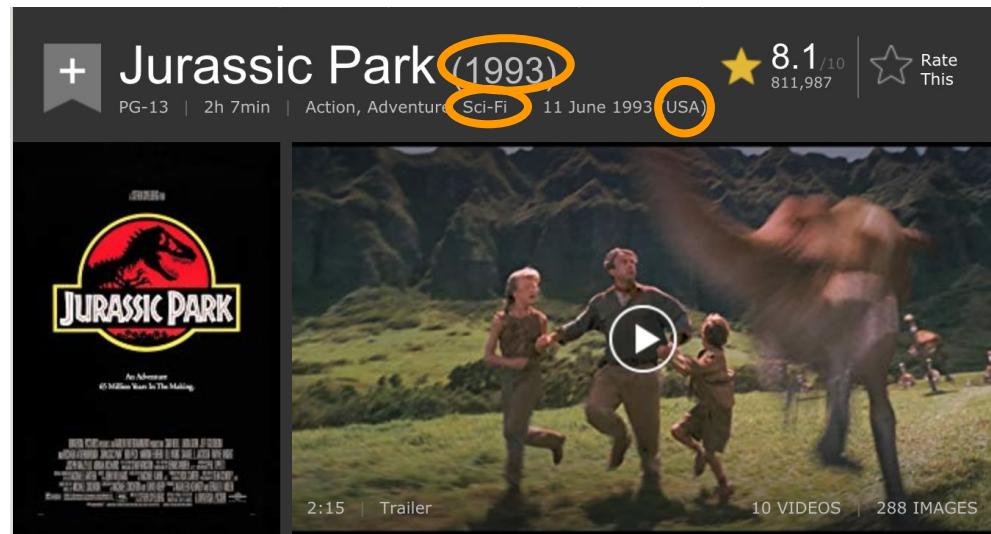
# Text vs. semi-structured data

## Jurassic Park (film)

From Wikipedia, the free encyclopedia

*This article is about the 1993 film. For the franchise, see [Jurassic Park \(disambiguation\)](#).*

**Jurassic Park** is a 1993 American science fiction adventure film directed by Steven Spielberg and produced by Kathleen Kennedy and Gerald R. Molen. It is the first installment in the *Jurassic Park* franchise, and is based on the 1990 novel of the same name by Michael Crichton and a screenplay written by Crichton and David Koepp. The film is set on the fictional island of Isla Nublar, located off Central America's Pacific Coast near Costa Rica. There, billionaire philanthropist John Hammond and a small team of genetic scientists have created a *wildlife park* of de-extinct dinosaurs. When



A pragmatic Paleontologist visiting an almost complete theme park is tasked with protecting a couple of kids after a power failure causes the park's cloned dinosaurs to run loose.

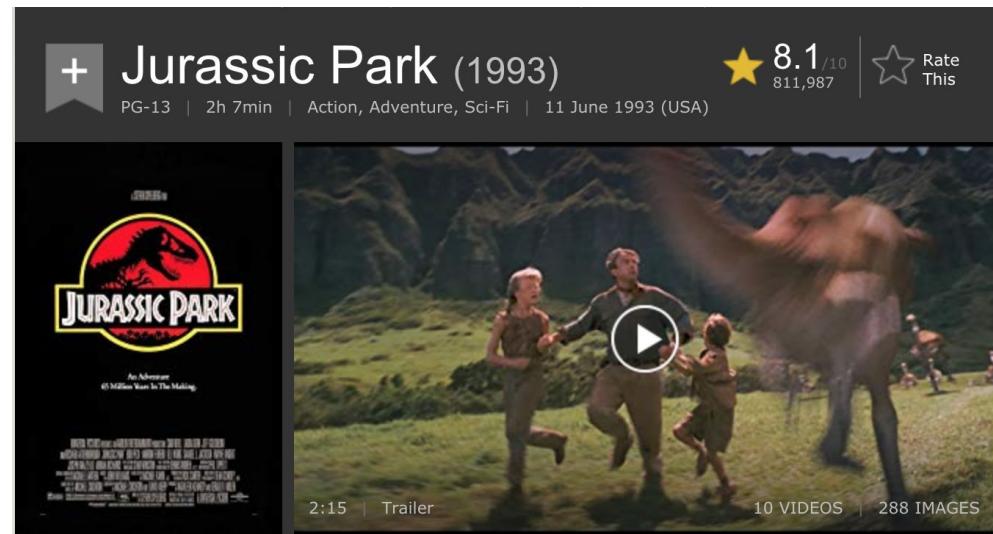
Director: Steven Spielberg

Writers: Michael Crichton (novel) | Michael Crichton (screenplay) | 1 more credit »

Stars: Sam Neill, Laura Dern, Jeff Goldblum | See full cast & crew »

# Semi-structured data vs. semi-structured data

<b>Directed by</b>	Steven Spielberg
<b>Produced by</b>	Kathleen Kennedy Gerald R. Molen
<b>Screenplay by</b>	Michael Crichton David Koepp
<b>Based on</b>	<i>Jurassic Park</i> by Michael Crichton
<b>Starring</b>	Sam Neill Laura Dern Jeff Goldblum Richard Attenborough Bob Peck Martin Ferrero



A pragmatic Paleontologist visiting an almost complete theme park is tasked with protecting a couple of kids after a power failure causes the park's cloned dinosaurs to run loose.

**Director:** Steven Spielberg

**Writers:** Michael Crichton (novel), Michael Crichton (screenplay) | 1 more credit »

**Stars:** Sam Neill, Laura Dern, Jeff Goldblum | See full cast & crew »

# Text vs. web table

surpass \$1 billion in ticket sales. The film won more than twenty awards, including three [Academy Awards](#) for its technical achievements in visual effects and sound design. *Jurassic Park* is considered a

Year	Award	Category	Nominees	Result
1993	Bambi Awards <sup>[154]</sup>	International Film	<i>Jurassic Park</i>	Won
66th Academy Awards <sup>[155]</sup>	Best Sound Editing		Gary Rydstrom and Richard Hymns	Won
			Gary Summers, Gary Rydstrom, Shawn Murphy and Ron Judkins	Won
			Dennis Muren, Stan Winston, Phil Tippett and Michael Lantieri	Won
		Best Visual Effects		
	Saturn Awards <sup>[147]</sup>	Best Director	Steven Spielberg	Won
		Best Science Fiction Film	<i>Jurassic Park</i>	Won
		Best Special Effects	Dennis Muren, Stan Winston, Phil Tippett and Michael Lantieri	Won
		Best Writing	Michael Crichton and David Koepp	Won
		Best Actress	Laura Dern	Nominated
		Best Costumes		Nominated
		Best Music	John Williams	Nominated
		Best Performance by a Young Actor	Joseph Mazzello	Nominated

# Language vs. language

<b>Directed by</b>	Steven Spielberg
<b>Produced by</b>	Kathleen Kennedy Gerald R. Molen
<b>Screenplay by</b>	Michael Crichton David Koepp
<b>Based on</b>	<i>Jurassic Park</i> by Michael Crichton
<b>Starring</b>	Sam Neill Laura Dern Jeff Goldblum Richard Attenborough Bob Peck Martin Ferrero

## 侏罗纪公园 Jurassic Park (1993)



导演: 史蒂文·斯皮尔伯格

编剧: 迈克尔·克莱顿 / 大卫·凯普

主演: 山姆·尼尔 / 劳拉·邓恩 / 杰夫·高布伦 / 理查德·阿滕伯勒 / 鲍勃·佩克 / 更多...

类型: 科幻 / 惊悚 / 冒险

官方网站: [jurassicpark.com](http://jurassicpark.com)

制片国家/地区: 美国

语言: 英语 / 西班牙语

上映日期: 2013-08-20(中国大陆 3D) / 1993-06-11(美国) / 2013-04-05(美国)

片长: 127 分钟

又名: Jurassic Park 3D

IMDb链接: [tt0107290](https://www.imdb.com/title/tt0107290)

# Challenge 1: Diversity of Data

- Different languages
- Different subject domains
- Different entity and relation types
- Different lexical/syntactic phrases
- Different website templates
- Different textual modalities

# Challenge 1: Diversity of Data

Extracting from more websites = More diversity

Extracting from multiple languages = More diversity

Extracting from multiple subject domains = More diversity

More detail = More diversity

# Challenge 2: Multiple Modality of Text

- Facts about an entity may be expressed in unstructured text, semi-structured fields, and tables
- How to--
  - Extract from all kinds of text
  - Link values between different kinds of text
  - Benefit from signals expressed in different modalities

# Challenge 3: Lack of Training Data

- More data → Better models
- But labeling data is expensive
- We need tons of labels!!
  - Labels for different domains
  - Labels for different relations
  - Labels for different modalities
  - Labels for different templates

# Challenge 4: Unknown Unknowns

- New Relationships
  - On 10 semi-structured movie websites, the IMDb ontology covers only 7% of relations.
- New Domains
  - Jurassic Park ride?
  - Video game?
  - Broadway show?
- Interesting? Not interesting?

## *Jurassic Park* (film)

---

From Wikipedia, the free encyclopedia

*This article is about the 1993 film. For the franchise, see [Jurassic Park \(disambiguation\)](#).*

**Jurassic Park** is a 1993 American science fiction adventure film directed by Steven Spielberg and produced by Kathleen Kennedy and Gerald R. Molen. It is the first installment in the *Jurassic Park* franchise, and is based on the 1990 novel of the same name by Michael Crichton and a screenplay written by Crichton and David Koepp. The film is set on the fictional island of Isla Nublar, located off Central America's Pacific Coast near Costa Rica. There, billionaire philanthropist John Hammond and a small team of genetic scientists have created a wildlife park of de-extinct dinosaurs. When

# Summary: Four Challenges

1. Diversity of data
2. Multiple modalities of text
3. Lack of training data
4. Unknown unknowns

Can we build a single extractor to find **consistent signals** across these diverse elements of data **from all modalities of text**?

---

---

# How to Do Web-Scale Information Extraction

---

---

# Key Intuitions

- Diversity → Identifying consistent patterns

# Key Intuitions

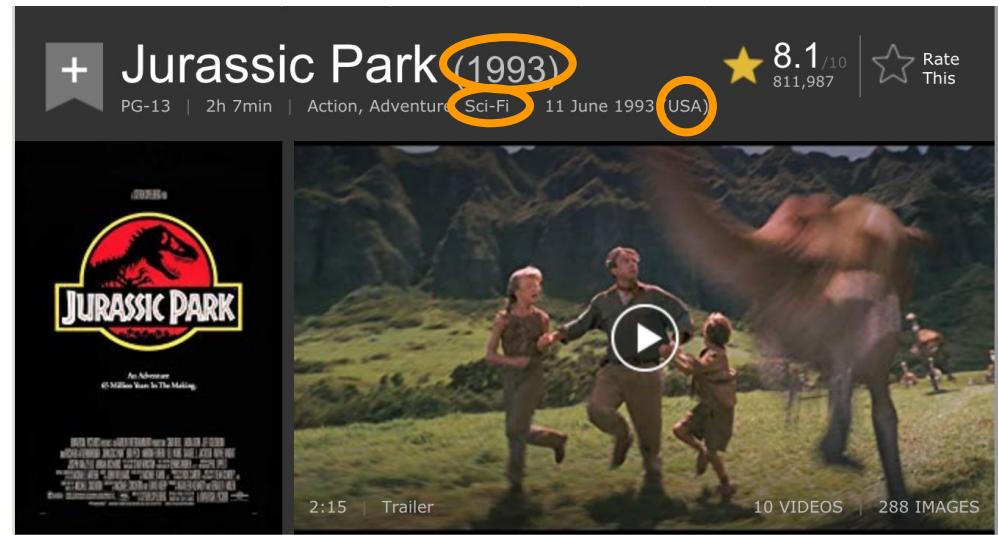
- Diversity→Identifying consistent patterns

## Jurassic Park (film)

From Wikipedia, the free encyclopedia

This article is about the 1993 film. For the franchise, see [Jurassic Park \(disambiguation\)](#).

**Jurassic Park** is a 1993 American science fiction adventure film directed by Steven Spielberg and produced by Kathleen Kennedy and Gerald R. Molen. It is the first installment in the [Jurassic Park](#) franchise, and is based on the 1990 novel of the same name by Michael Crichton and a screenplay written by Crichton and David Koepp. The film is set on the fictional island of Isla Nublar, located off Central America's Pacific Coast near Costa Rica. There, billionaire philanthropist John Hammond and a small team of genetic scientists have



A pragmatic Paleontologist visiting an almost complete theme park is tasked with protecting a couple of kids after a power failure causes the park's cloned dinosaurs to run loose.

Director: Steven Spielberg

Writers: Michael Crichton (novel), Michael Crichton (screenplay) | 1 more credit »

# Key Intuitions

- Diversity→Identifying consistent patterns



A Bonanza ( Nodaji )

1961년 · 대한민국 · 127분

1961-06-01 (개봉)

제작사      화성영화주식회사

감독

정창화

출연

김승호

황해, 엄앵란

조미령

허장강

[더보기](#)

[스크랩하기](#)

# Key Intuitions

- Diversity / Modality → Identifying consistent patterns
  - Leverage consistency in model/representation
  - Leverage redundancy across the web (make scale an advantage)
  - Combining information from multiple modalities can give more consistent signals

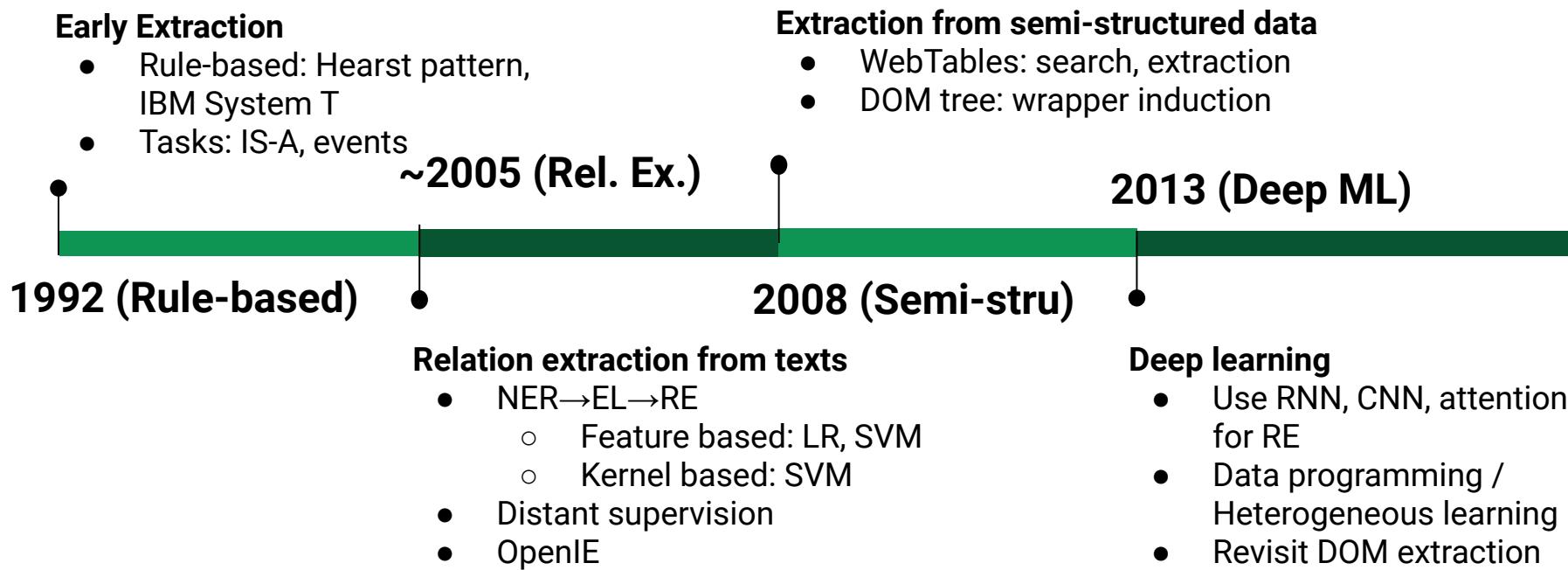
# Key Intuitions

- Diversity / Modality→Identifying consistent patterns
  - Leverage consistency in model/representation
  - Leverage redundancy across the web (make scale an advantage)
  - Combining information from multiple modalities can give more consistent signals
- Lack of training data→Learning with limited labels
  - Find automated ways to label data
  - Employ weak learning or semi-supervision

# Key Intuitions

- Diversity / Modality→Identifying consistent patterns
  - Leverage consistency in model/representation
  - Leverage redundancy across the web (make scale an advantage)
  - Combining information from multiple modalities can give more consistent signals
- Lack of training data→Learning with limited labels
  - Find automated ways to label data
  - Employ weak learning or semi-supervision
- Unknown unknowns→OpenIE
  - Identifying similarity between known predicates and unknown predicates

# 35 Years of Information Extraction





# Learning Information Extraction Rules for Semi-Structured and Free Text

STEPHEN SODERLAND

soderlan@cs.washington.edu

*Department Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350*

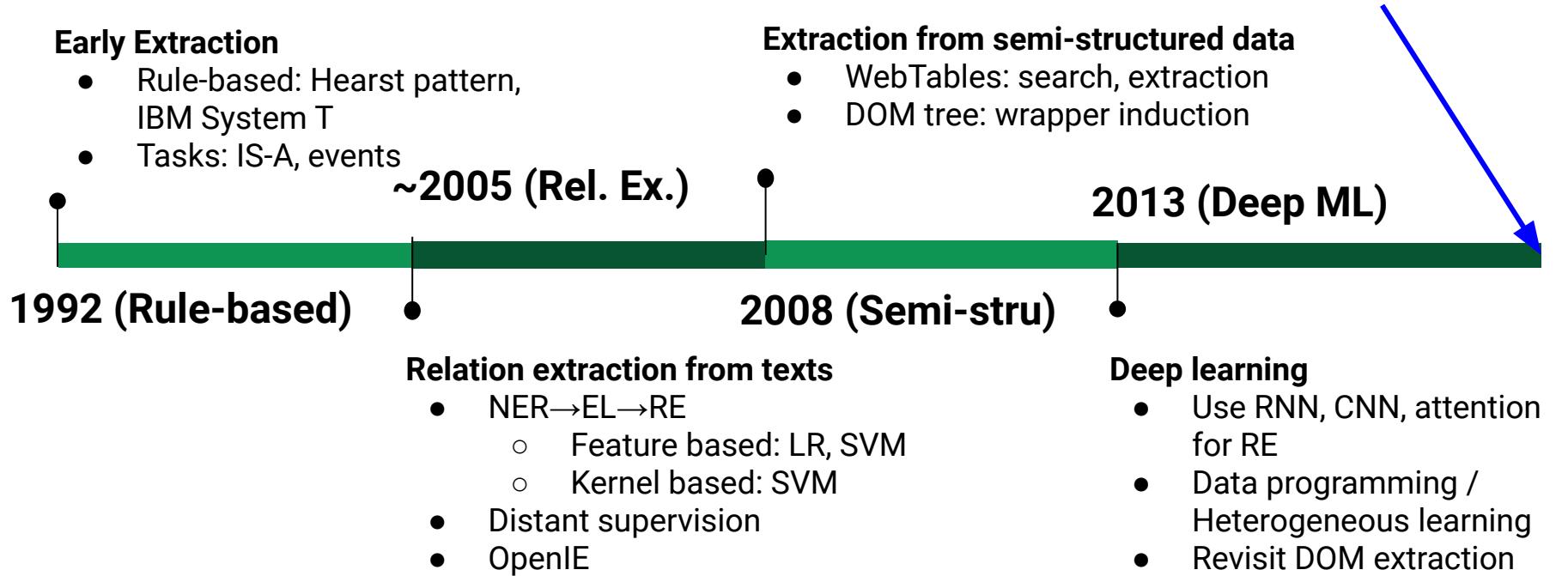
**Editors:** Claire Cardie and Raymond Mooney

**Abstract.** A wealth of on-line text information can be made available to automatic processing by information extraction (IE) systems. Each IE application needs a separate set of rules tuned to the domain and writing style. WHISK helps to overcome this knowledge-engineering bottleneck by learning text extraction rules automatically.

WHISK is designed to handle text styles ranging from highly structured to free text, including text that is neither rigidly formatted nor composed of grammatical sentences. Such semi-structured text has largely been beyond the scope of previous systems. When used in conjunction with a syntactic analyzer and semantic tagging, WHISK can also handle extraction from free text such as news stories.

**Keywords:** natural language processing, information extraction, rule learning

# 35 Years of Information Extraction



# What is multi-modal extraction?

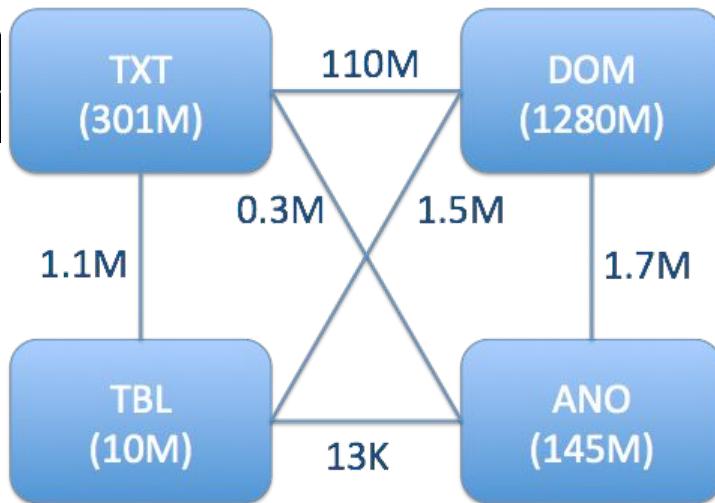
In the multi-modal setting, we will consider methods that jointly address unstructured, semi-structured, and tabular text and bring in **visual** information

No real full-fledged systems in practice yet

# Example 1. Google Knowledge Vault

Knowledge extraction from four types of web data (Dong et al., KDD 2014, VLDB 2014)

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



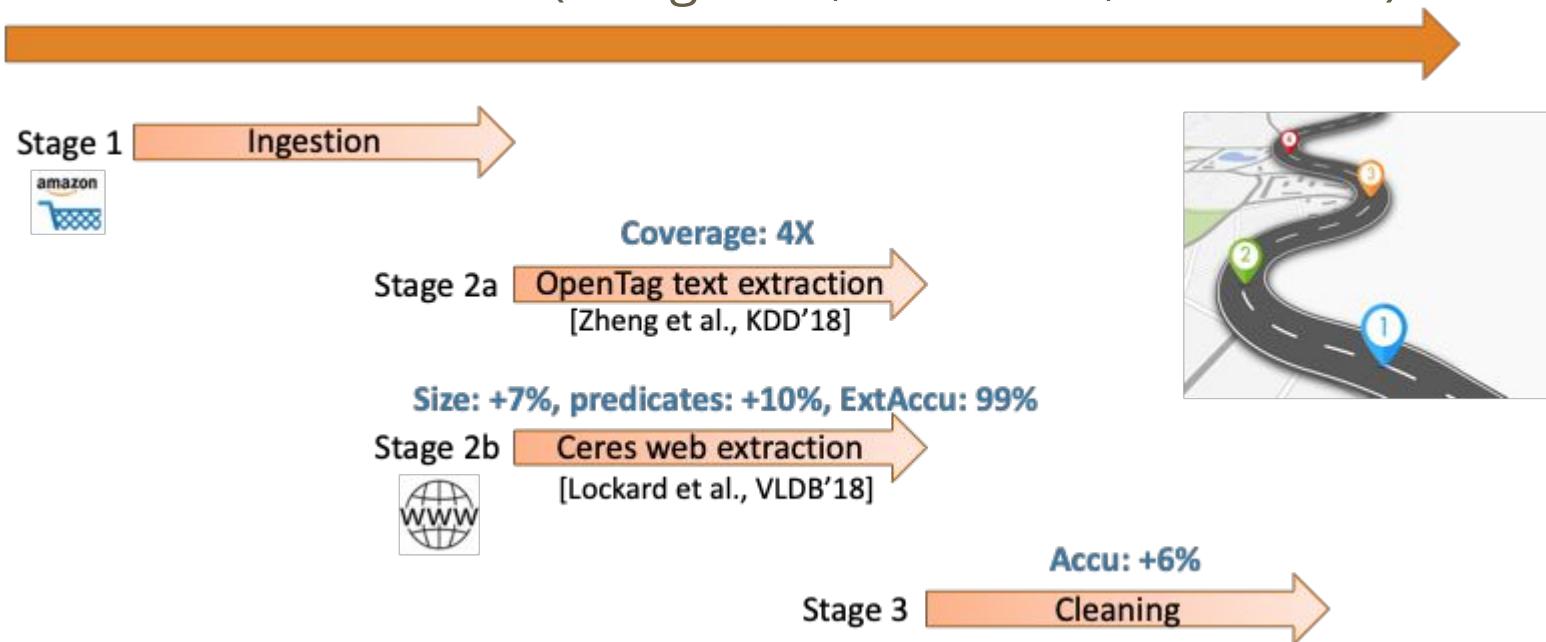
Accu	Accu (conf $\geq .7$ )
0.43	0.63
0.09	0.62

\*Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). Knowledge vault: a web-scale approach to probabilistic knowledge fusion. KDD '14.

\*Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., & Zhang, W. (2014). From Data Fusion to Knowledge Fusion. VLDB.

# Example 2. Amazon Product Graph

Product knowledge extraction from Catalog product profiles and semi-structured websites (Dong et al., KDD 2018, ICDE 2019)



# In this tutorial, we will cover...

- Information extraction techniques for unstructured, semi-structured, and tabular text
- Overview of common challenges facing any extraction project (and suggested solutions)
- State-of-the-art approaches from academia and industry that consider all types of text
- A look to the future of knowledge collection from the web

# In this tutorial, we will NOT cover...

- Web crawling
- Machine translation
- Entity linking
- Knowledge base cleaning
- Knowledge fusion
- Automated question answering
- ...

# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- Part 1b: Unstructured Text: Methods (10 minutes)
- **Live Q&A & Discussion (15 minutes)**
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- Part 3: Multi-modal Extraction (30 minutes)
- Conclusion (5 minutes)
- **Live Q&A & Discussion (15 minutes)**

Please ask questions over  
RocketChat!

**# tutorial-2**

We will monitor and respond  
during our presentation!

---

# Knowledge Acquisition from Unstructured Text

---

— Colin Lockard, Prashant Shiralkar,  
Xin Luna Dong, **Hannaneh Hajishirzi**

---



PAUL G. ALLEN SCHOOL  
OF COMPUTER SCIENCE & ENGINEERING

# Outline

- Introduction (40 minutes)
- **Part 1a: Unstructured Text (25 minutes)**
- Part 1b: Unstructured Text: Methods (10 minutes)
- Live Q&A (15 minutes)
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- Part 3: Multi-modal Extraction and Conclusion (35 minutes)
- Live Q&A (15 minutes)

# Questions we will answer in this section

- Task: knowledge acquisition from text
  - Unstructured text
  - Tasks and sub-tasks
  - Challenges
- Models
  - General overview
  - Specific methods

# How can we extract knowledge from texts?

## Jurassic Park (film)

From Wikipedia, the free encyclopedia

*This article is about the 1993 film. For the franchise, see [Jurassic Park \(disambiguation\)](#).*

**Jurassic Park** is a 1993 American science fiction adventure film directed by Steven Spielberg and produced by Kathleen Kennedy and Gerald R. Molen. It is the first installment in the *Jurassic Park* franchise, and is based on the 1990 novel of the same name by Michael Crichton and a screenplay written by Crichton and David Koepp. The film is set on the fictional island of Isla Nublar, located off Central America's Pacific Coast near Costa Rica. There, billionaire philanthropist John Hammond and a small team of genetic scientists have created a [wildlife park](#) of de-extinct dinosaurs. When

Crest Complete Whitening + Scope Toothpaste, Minty Fresh, 5.4 Ounce Triple Pack

by Crest

★★★★★ 2,579 ratings | 44 answered questions

Amazon's Choice for "toothpaste"

List Price: \$8.77

Price: **\$5.59** (\$0.35 / Ounce) FREE Shipping on orders over \$25.00 shipped by Amazon or get Fast, Free Shipping with Amazon Prime & FREE Returns

You Save: **\$3.18 (36%)**

In Stock.

Want it Friday, Jan. 24? Order within **6 hrs 31 mins** and choose **Two-Day Shipping** at checkout. Details Ships from and sold by Amazon.com.

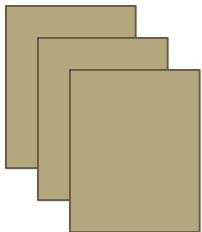
This item is returnable

Style Name: Minty Fresh Toothpaste (Triple Pack)



- Leaves mouth and breath feeling refreshed
- Whitens teeth by gently removing surface stains
- Fights cavities
- Fights tartar build-up

# What Is Unstructured Text?



A set of documents



Document (a sequence of sentences)



Once

upon

a

Sentence (a sequence of words)



O   n   c   e

Words (a sequence of characters)

# Unstructured text comes in many different forms

The screenshot shows the Wikipedia article page for "Oprah Winfrey". At the top left is the Wikipedia logo (a globe made of puzzle pieces). Below it, the word "WIKIPEDIA" is written in large, bold, black letters, with "The Free Encyclopedia" in smaller text underneath. In the top navigation bar, there are two tabs: "Article" (which is selected) and "Talk". The main title "Oprah Winfrey" is prominently displayed in a large serif font. Below the title, a subtitle reads "From Wikipedia, the free encyclopedia".

The screenshot displays a scientific paper titled "Scientific Information Extraction with Semi-supervised Neural Tagging". The authors listed are Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. They are from the Department of Electrical Engineering, University of Washington, with their email addresses: {yuanly, ostendorf, hannaneh}@uw.edu. The paper is divided into sections: "Task", "Method", and "Abstract". The "Abstract" section contains several paragraphs of text, with some parts highlighted in red, such as "This paper addresses the problem of extracting keyphrases from scientific articles and categorizing them as corresponding to a task, process, or material." and "annotation performance on the 2017 IAAI Task 10 SciencE test".

The screenshot shows a news article from The New York Times. The headline is "Bumblebee Vomit: Scientists Are No Longer Ignoring It". The article is presented in a clean, modern layout with a large title at the top and a detailed description below.

The screenshot shows the front cover of the book "Little Women" by Louisa May Alcott. The cover features a painting of four young women (the March sisters) looking down at a book. The title "LITTLE WOMEN" is at the top, and the author's name "LOUISA MAY ALCOTT" is at the bottom. The background of the slide includes some text and images related to the book, such as "Illustrated Edition" and a small illustration of a woman reading.

The screenshot shows a tweet from Cristiano Ronaldo (@Cristiano). The profile picture is a portrait of him. The tweet text is: "Feliz Ano, meu Amor! ❤ Que 2020 seja um ano repleto de amor, saúde, paz e sucesso para todos! Happy New Year to all! 🎉". The timestamp indicates it was posted on Jan 1.

The screenshot shows the front cover of the book "Little Women" by Louisa May Alcott. The cover features a painting of four young women (the March sisters) looking down at a book. The title "LITTLE WOMEN" is at the top, and the author's name "LOUISA MAY ALCOTT" is at the bottom. The background of the slide includes some text and images related to the book, such as "Illustrated Edition" and a small illustration of a woman reading. A maroon overlay covers the bottom right corner of the slide.

# Characteristics of unstructured texts

- **Completely free form:** paragraphs, sentences, phrases
- **Common grammar and words:** different articles can have different styles, but grammar and words are similar
- **Rich information from text:** human language possibly has the highest expressiveness
- **Typically not much of layout:** normally just paragraphs with hyperlinks
- **A lot of information is not factual:** subjective, emotions, fictional, etc.

# Why extracting from unstructured texts

- Text is the fundamental way for people to communicate and pass on knowledge



- Acquired knowledge is useful for real-world applications
  - Search engines, question answering, healthcare

# What is extraction from texts

- Input:
  - Sentences
  - Paragraphs
  - Documents
- Output
  - **Entity** Extraction: Binary relationship (IS-A)
  - **Relation** Extraction: (subject, predicate, object) triple
  - **Event** Extraction: When, Where, Who, What, How



# Extraction output

Abraham Lincoln was elected President of the United States in 1860.

Person

Job Title

held job

Winfrey's best friend since their early twenties is Gayle King.

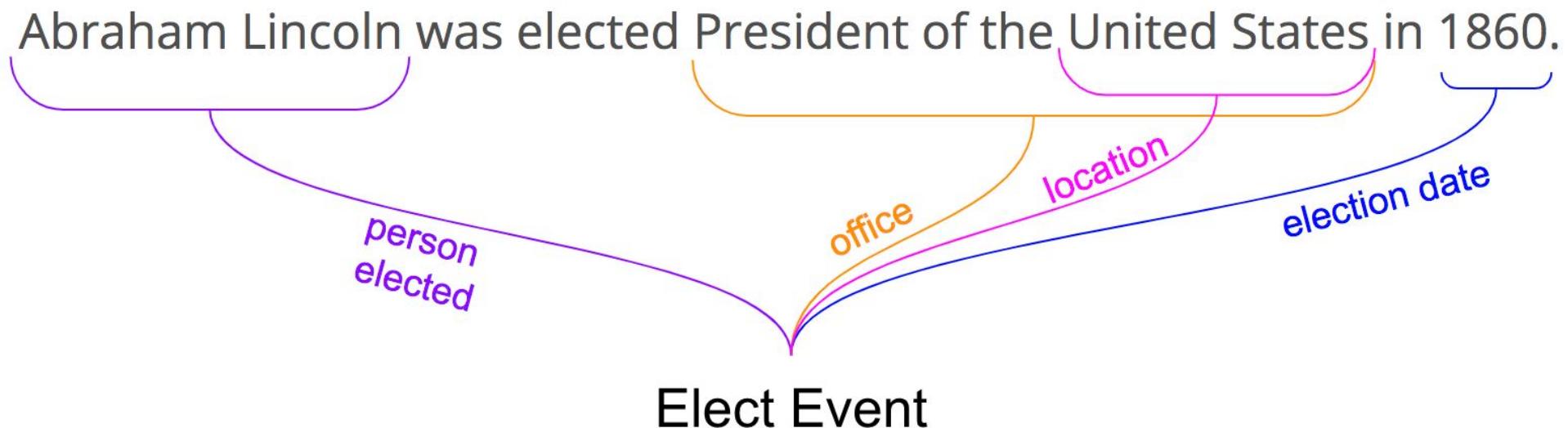
Person

Person

is friends with

# Event Extraction

“Events” are relationships that occur at specific time and place.



# Questions we will answer in this section

- Task: knowledge acquisition from text
  - Unstructured text
  - Tasks and sub-tasks
  - Challenges
- Models
  - General overview
  - Specific methods

# Why is extraction from text hard?

- Diversity

Bill Gates founded Microsoft in 1975.

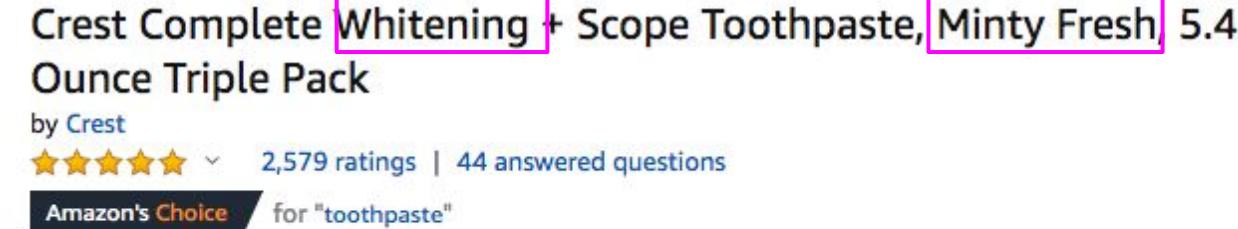
Bill Gates, founder of Microsoft, ...

Google was founded by Larry Page ...

Amazon was founded in the garage of Bezos' rented home in Bellevue, Washington

# Why is extraction from text hard?

- Language can be fuzzy, ambiguous
- Implicit mention of the relations



- Relations might exist across sentences

To reduce ambiguity, the **MORphological PArser MORPA** is provided with a PCFG ...

MORPA is a fully implemented parser developed for a [text-to-speech system](#).

**MORphological PArser MORPA** → **Used-for** → **[text-to-speech system](#)**

# Why is extraction from text hard?

- Salient Entities and Relations: What to extract and what not to?

## Jurassic Park (film)

From Wikipedia, the free encyclopedia

This article is about the 1993 film. For the franchise, see [Jurassic Park \(disambiguation\)](#).

Jurassic Park is a 1993 American science fiction adventure film directed by Steven Spielberg and produced by Kathleen Kennedy and Gerald R. Molen. It is the first installment in the Jurassic Park franchise, and is based on the 1990 novel of the same name by Michael Crichton and a screenplay written by Crichton and David Koepp. The film is set on the fictional island of Isla Nublar, located off Central America's Pacific Coast near Costa Rica. There, billionaire philanthropist John Hammond and a small team of genetic scientists have created a wildlife park of de-extinct dinosaurs. When

We evaluate our model on the task of question answering using

....

### Section : Dataset

SQuAD is a machine comprehension dataset on a large set of Wikipedia articles , ..... Two metrics are used to evaluate models : Exact Match ( EM ) and a softer metric , F1 score

....

### Section: Model Details .

... Each paragraph and question are tokenized by a regular - expression - based word tokenizer ( PTB Tokenizer ) and fed into the model .

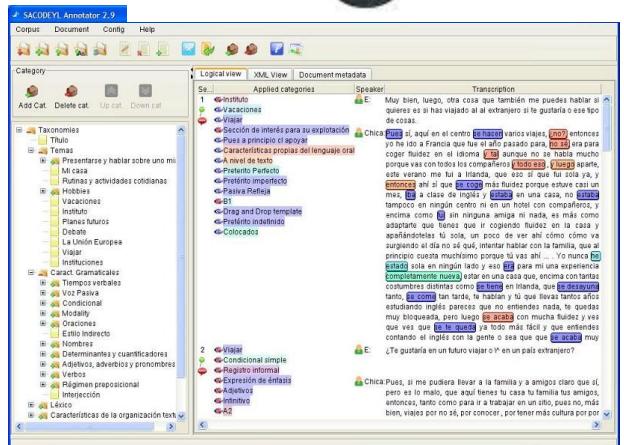
....

### Section : Results .

The results of our model and competing approaches on the hidden test are summarized in Table [ reference ]. BiDAF ( ensemble ) achieves an EM score of 73.3 and an F1 score of 81.1 , outperforming all previous approaches .

# Why is extraction from text hard?

- Lack of training data
  - Annotation is challenging
  - Need to define the ontology
  - Domains require expert annotators
    - E.g., scientific domains



The screenshot shows the SACODEYI Annotator 2.9 application. On the left, there's a sidebar with various categories and tools:

- Taxonomías:** Título, Temas, Interesante y hablar sobre uno mismo, Música, Runas y actividades cotidianas, Hobbies, Vacaciones, Interés, Planes futuros, Declarar, La Unión Europea, Visitar, Instituciones.
- Categorías Gramaticales:** Voz Pasiva, Voz Pasiva, Condicional, Modality, Clases, Estilo Indirecto, Nombres, Clases y cuantificadores, Adjetivos, adverbios y pronombres, Verbos, Régimen preposicional, Interacción, Léxico, Características de la organización textual.
- Sel:** Inifinitivo, Visceral, Visión, Visión de interés para su explotación, Pues a principio ci aci al, Características propias del lenguaje oral, Características propias del lenguaje escrito, Práctico Perfecto, Práctico Imperfecto, Paliva Reflexa, Sel 1, Drag and Drop template, Práctico Indistinto, Colocados.
- Applied categories:** Speaker, Transcription.
- Speaker:** E.
- Transcription:**

Muy bien, luego, otra cosa que también me puedes hablar si quieres o si has viajado al extranjero te gustaría o ese tipo de cosas.  
Chica Pues aquí en el centro [baja voz] entiendo, [baja voz] entonces vía a Francia que he oido pasado, [baja voz] era para coger hielos en el sistema [baja voz] aunque no se habla mucho porque vas con todos los compañeros [baja voz] ya, ya, ya, este verano, ma que irá a Irlanda, [baja voz] al que tu has ido, [baja voz] y [baja voz] que es que [baja voz] fuiste a vivir a Irlanda con tu novio, [baja voz] a la clase de inglés y [baja voz] en una casa, no [baja voz] tampoco en ningún centro ni en un hotel con compañeros, y [baja voz] que [baja voz] un día te quedaste en Irlanda, [baja voz] como asustarte que [baja voz] que el cogieron [baja voz] en la cara y [baja voz] te soltó un poco de ver año como [baja voz] como va surgiendo el día no se quedó, [baja voz] interactuando con la familia, que es [baja voz] que [baja voz] te quedaste en Irlanda, [baja voz] yo [baja voz] sola en ningún lado y [baja voz] para mi una experiencia completamente nueva, [baja voz] estar en una casa que, encima con tantas costumbres distintas como [baja voz] Irlanda, que [baja voz] tanto, [baja voz] que [baja voz] te quedaste en Irlanda, [baja voz] estudiando inglés pareces que no entiendes nada, [baja voz] quedas muy bloqueada, pero luego [baja voz] con mucha fuerza y ves que [baja voz] ya todo más fácil y que entiendes contando el inglés con el gente y [baja voz] que [baja voz] muy [baja voz] te gustaría un futuro viajar o no?

Chica Pues, [baja voz] lo más que pasa es la forma y a amigos dice que si, [baja voz] que [baja voz] es lo malo, que [baja voz] te pasa que [baja voz] amigas, entonces, tanto como para ir a trabajar en un sitio, pues no, más bien, viajes por no sé, por conocer, [baja voz] tener más cultura por por

# Why is extraction from text hard?

- **Diversity**
  - Different ways of expressing the same entity, relationship, etc.
  - Language can be fuzzy, ambiguous
  - Different languages
- **Lack of training data**
- **Unknown unknowns**
  - factual and interesting vs. factual but not interesting  
vs. subjective

# Questions we will answer in this section

- Task: knowledge acquisition from text
  - Unstructured text
  - Challenges
  - Tasks and sub-tasks
- Solutions
  - General overview
  - Specific methods

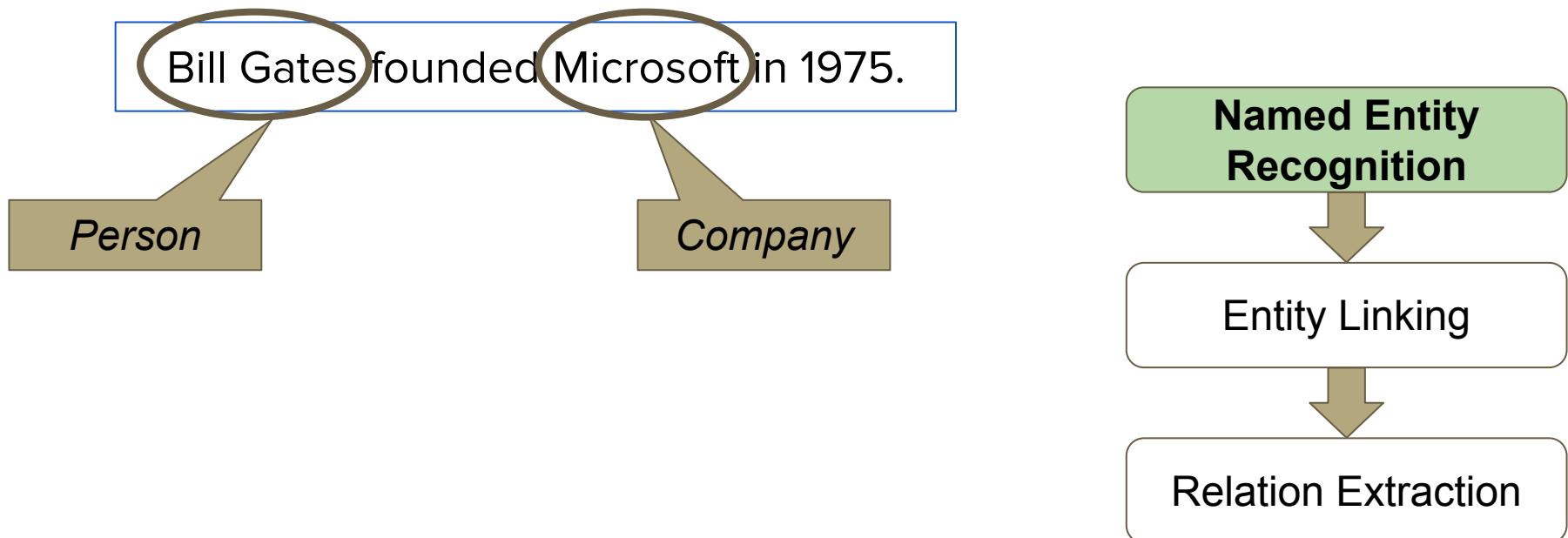
# Opportunities

- **Consistency:** Same grammar and word semantics
- **Redundancy:** Same fact is often repeated in different articles, in various ways

# Short Answers

- **Consistency**
  - Model problem as text span classification and relationships between spans
  - Word embedding models help capture text semantics
- **Training data**
  - Small-scale training data and semi-supervision
  - Weak supervision gives cheap training data
- **OpenIE**
  - Discovery of new types and relationships

# High-level approach for extraction



# High-level approach for extraction

Bill Gates founded Microsoft in 1975.



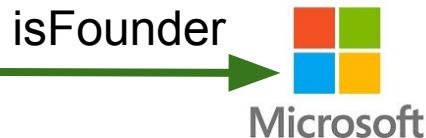
Named Entity  
Recognition

Entity Linking

Relation Extraction

# High-level approach for extraction

Bill Gates founded Microsoft in 1975.



We focus on Relation Extraction in the rest of the tutorial.

Named Entity Recognition

Entity Linking

Relation Extraction

# Generic Information Extraction Method

Machine learning classifiers take in a set of **features** that describe a data point and output a **prediction** of that datapoint's class.

We'll need to:

1. Select **features** to represent our raw text
  - a. **Span features: Combine** those features for larger units
2. Select a **model** to take in these features and make a prediction
3. **Train** that model

# Generic Information Extraction Method

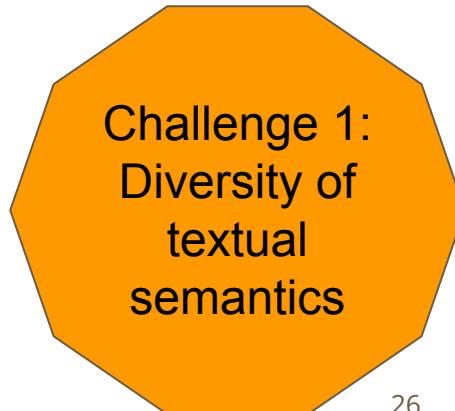
Machine learning classifiers take in a set of **features** that describe a data point and output a **prediction** of that datapoint's class.

We'll need to:

1. Select **features** to represent our raw text
  - a. **Span features: Combine** those features for larger units
2. Select a **model** to take in these features and make a prediction
3. **Train** that model

# Representing words is hard

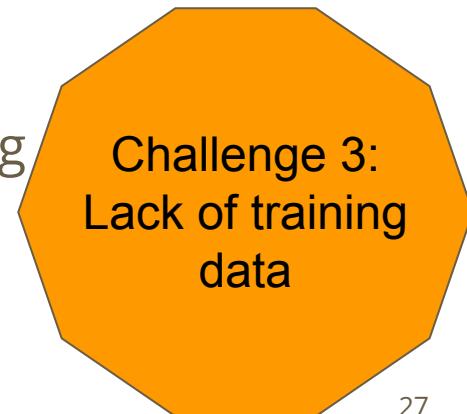
- Different words can mean the same thing.
  - Dog, pup, pooch, hound, canine can all refer to the same animal
- The same word can mean different things.
  - “by the river bank” and “by the Chase bank”



Challenge 1:  
Diversity of  
textual  
semantics

# Representing words is hard

- Different words can mean the same thing.
  - Dog, pup, pooch, hound, canine can all refer to the same animal
- The same word can mean different things.
  - “by the river bank” and “by the Chase bank”
- There are a lot of words.
  - Some words appear rarely/never during training

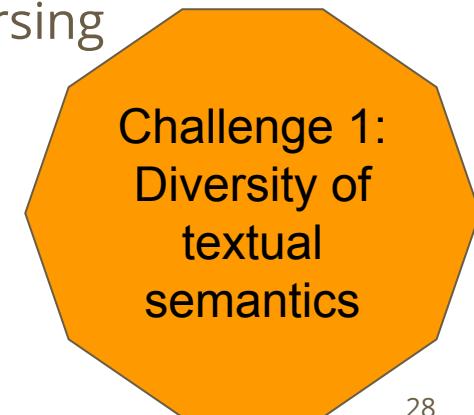


# Text features

- Understand the meaning of each word
- Understand the meaning of each word in its context
- Understand the meaning of multiple words in a sequence

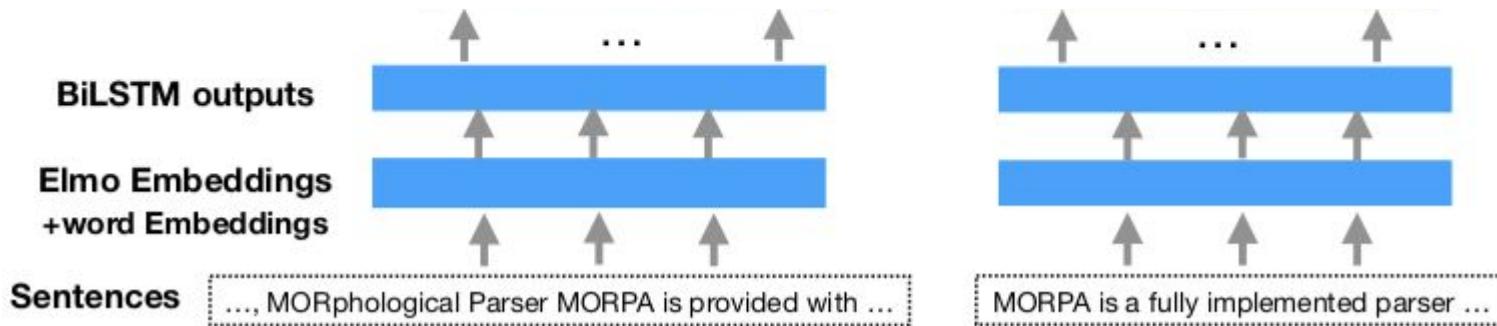
# Featurizing text

- A few years ago: Bag-of-words, POS tags, syntactic parsing
- Word embeddings: Word2Vec (Mikolov et al, 2013), GloVe (Pennington et al, 2014)
- Now: Pre-trained contextual embedding models



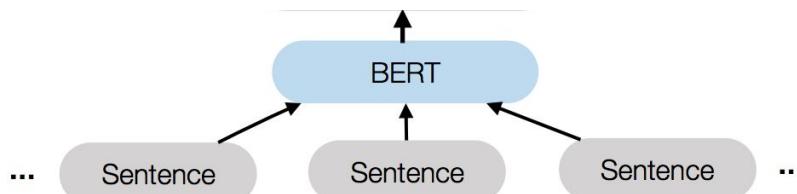
# Word Embeddings and LSTMs

- Dense vector representation of a word
  - Bi-LSTMs to encode context



# Contextual Word Embeddings

- BERT (Devlin et al, 2019), etc.
  - Builds contextual representation of each token in a sentence
  - Training objective: Learn to predict masked words in a sentence
  - Transformer neural net architecture
  - Also builds representation of entire sentence
  - Pre-trained on a large text corpus



# Problems with BERT

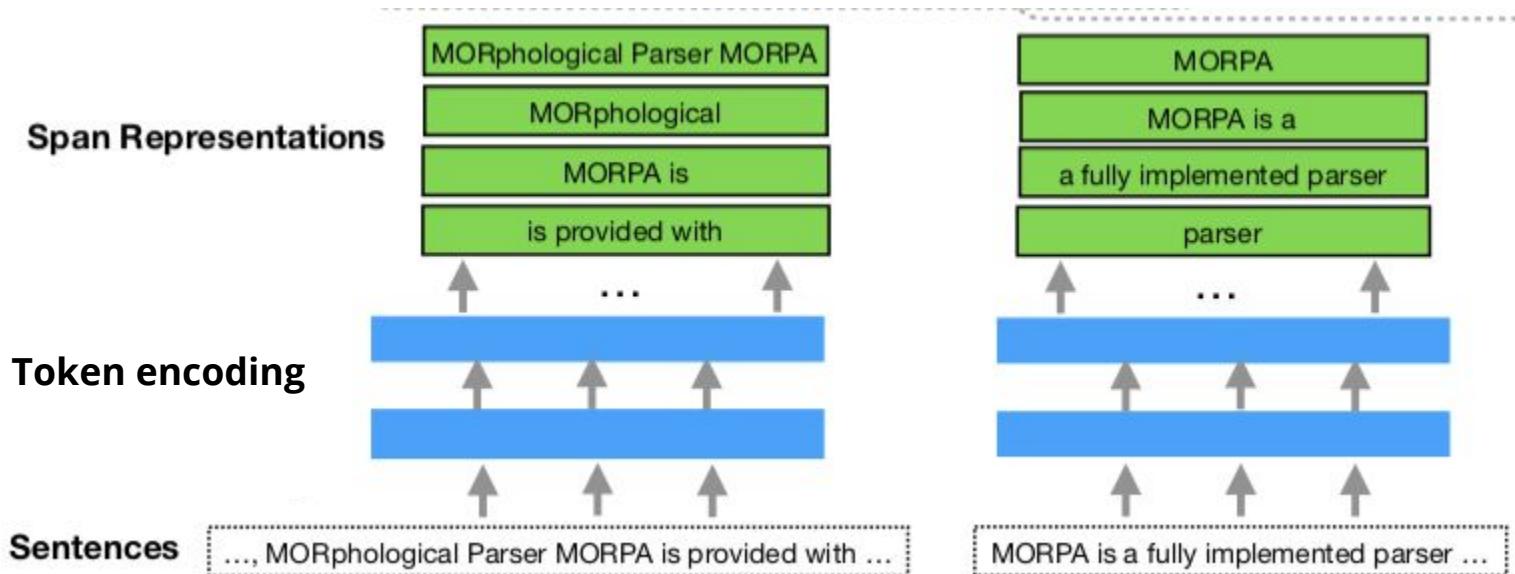
- Less effective if text is very different from “normal” English
  - Train model specific to your text
  - E.g. SciBERT (Beltagy et al, 2019) for scientific documents
- Computationally expensive
  - 1-30 seconds per webpage on GPU
  - Faster alternatives: ALBERT (Lan et al, 2019)
- Mostly limited to a limited number of tokens (e.g., 512 tokens)
  - How to capture longer sequences like documents?

\*Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. EMNLP/IJCNLP.

\*Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ICLR.

# Span Features

- **Encoding:** Integrate encodings of individual tokens in the span
  - Use Long Short Term Memories
  - **Simple, but effective:** Concatenate start and end token encodings
- **Challenges:** How to enumerate and encode spans for long sequences?



# Generic Information Extraction Method

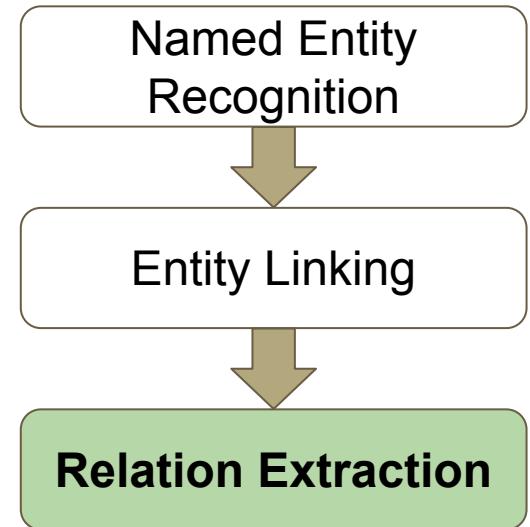
Machine learning classifiers take in a set of **features** that describe a data point and output a **prediction** of that datapoint's class.

We'll need to:

1. Select **features** to represent our raw text
  - a. **Span features: Combine** those features for larger units
2. Select a **model** to take in these features and make a prediction
3. **Train** that model

# Classification Models

- Pipeline approach:
  - Cast as a sequence tagging problem  
(Lin et al, 2018)
- End-to-end training:
  - Joint entity and relation extraction  
(Miwa and Bansal, 2016, Luan et al. 2018)



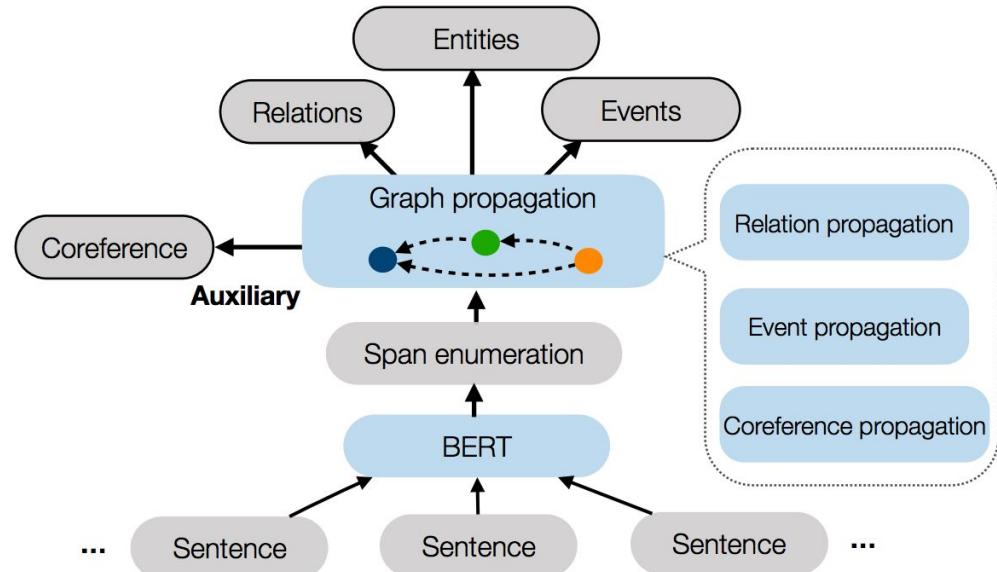
\*Lin, Y., Yang, S., Stoyanov, V., & Ji, H. (2018). A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling. ACL.

\*Miwa, M., & Bansal, M. (2016). End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. ACL.

\*Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. EMNLP.

# Multi-Task Learning

- A general-purpose information extraction model  
**DyGIE++** (Wadden et al.' 2019)
- Shared span representations between all tasks
  - Uses BERT
  - Updates span representations given other spans in the document



\*Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019). Entity, Relation, and Event Extraction with Contextualized Span Representations. EMNLP/IJCNLP.

# Generic Information Extraction Method

Machine learning classifiers take in a set of **features** that describe a data point and output a **prediction** of that datapoint's class.

We'll need to:

1. Select **features** to represent our raw text
  - a. **Span features: Combine** those features for larger units
2. Select a **model** to take in these features and make a prediction
3. **Train that model**

# Data annotation

Winfrey's best friend since their early twenties is Gayle King.

The diagram illustrates the concept of data annotation. It shows a sentence: "Winfrey's best friend since their early twenties is Gayle King." Two entities are identified: "Winfrey" and "Gayle King", each underlined by a purple bracket and labeled "Person". A green bracket below them spans the entire phrase "best friend since their early twenties" and is labeled "is friends with".

- What is the ontology?
- Where should the label for the “is friends with” relation go?
  - On the word “friend”?
  - On the span between the entity pair?
  - On the sentence?
  - On the paragraph?
  - On the document?

Challenge 2:  
Lack of training  
data

# Data annotation

*“We next expressed ALK F1174L, ALK F1174L/L1198P, ALK F1174L/G1123S, and ALK F1174L/G1123D in the original SH-SY5Y cell line.”*

(... 15 sentences spanning 3 paragraphs ...)

*“The 2 mutations that were only found in the neuroblastoma resistance screen (G1123S/D) are located in the glycine-rich loop, which is known to be crucial for ATP and ligand binding and are the first mutations described that induce resistance to TAE684, but not to PF02341066.”*

This drug-gene-mutation relationship example from Jia et al (2019) spans 3 paragraphs.

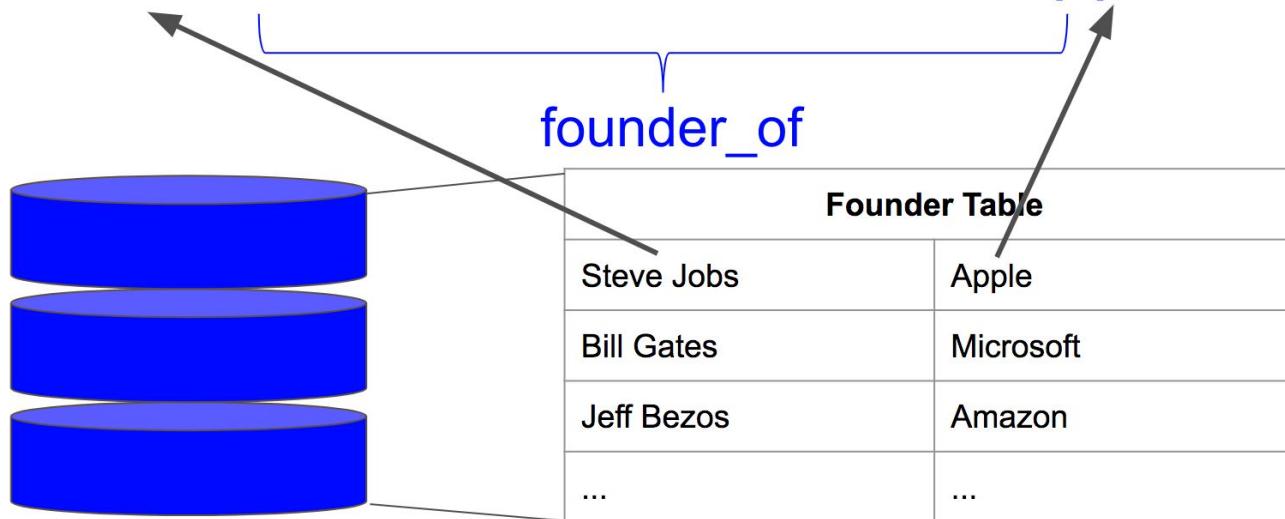
# Some Datasets

Dataset	Task	Source type	Size	Year	Annotation level
ACE	NER, relations, events	Newswire, web, transcripts	Thousands	2005	Sentence
SciERC	NER, relations, coref	Scientific paper abstracts	Thousands	2018	Cross-sentence
SciREX	NER, relations	Scientific papers	Thousands	2020	Document

# Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

Steve Jobs was the founder of Apple.

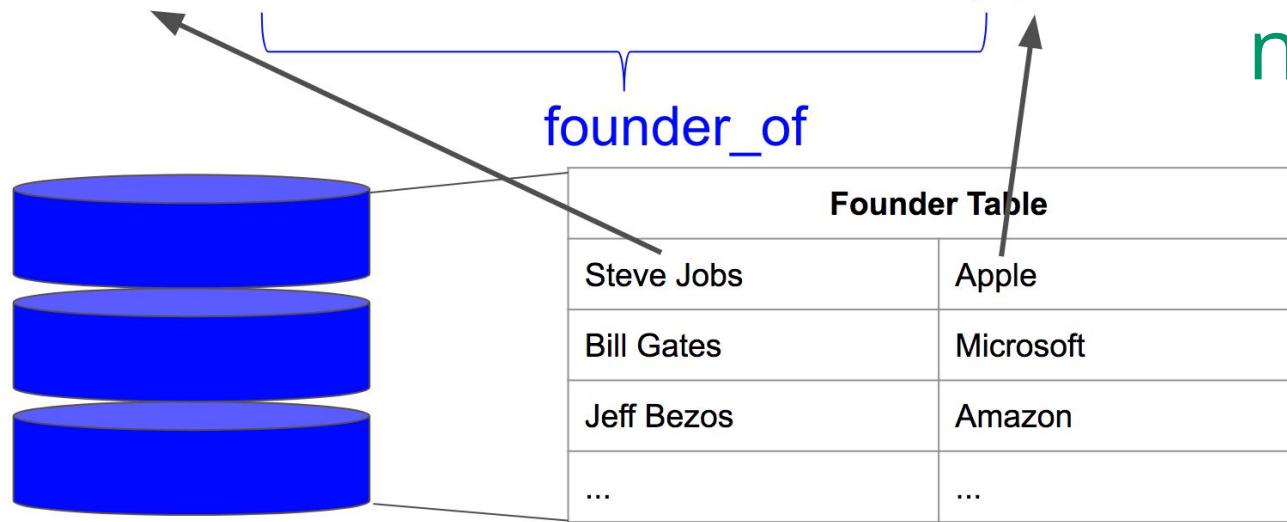


# Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

Steve Jobs was the CEO of Apple.

Matching  
to KB is  
noisy!



# Weak Supervision

- Use external data as distant supervision

Dataset	Task	Source type	Size	Year	Annotation level
NYT	relations	Newswire	>100,000	2010	Sentence (distantly supervised)
TACRED	relations	Newswire, Web	100,000	2017	Sentence
NewsSpike	relations/events	Newswire, web	>100,000	2017	Sentence

**How can we discover new relations?**

# Open Information Extraction (Banko et al, 2008)

All of the prior work requires a defined set of entity and relation types

Open Information Extraction: Extract arguments with a string representing the relationship



Challenge 3:  
Unknown  
unknowns

\*Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M.A., & Etzioni, O. (2008). Open Information Extraction from the Web. CACM.

# Discussions

- Unsupervised methods
  - Few-shot learning from entity descriptions in Wikipedia (Ling et al. 2019)
- Knowledge from pre-trained language models
  - Language models as knowledge bases (Petroni et al., 2019)

# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- **Part 1b: Unstructured Text: Methods (10 minutes)**
- Live Q&A (15 minutes)
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- Part 3: Multi-modal Extraction and Conclusion (35 minutes)
- Live Q&A (15 minutes)

---

# Knowledge Collection from Unstructured Text: Methods

---

— Colin Lockard, Prashant Shiralkar,  
Xin Luna Dong, Hannaneh Hajishirzi

---



# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- **Part 1b: Unstructured Text: Methods (10 minutes)**
- Live Q&A (15 minutes)
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- Part 3: Multi-modal Extraction (35 minutes)
- Live Q&A (15 minutes)

# Applied Methods

## Method: Sequence Tagging

## Applied Example: OpenTag

How can we extract attributes and relationships from detail pages with a known subject?

# Detail page

## Crest Complete Whitening + Scope Toothpaste, Minty Fresh, 5.4 Ounce Triple Pack

by Crest



2,579 ratings | 44 answered questions

Amazon's Choice

for "toothpaste"

List Price: \$8.77

Price: **\$5.59** (\$0.35 / Ounce) FREE Shipping on orders over \$25.00 shipped by Amazon or get Fast, Free Shipping with Amazon Prime & FREE Returns

You Save: \$3.18 (36%)

In Stock.

Want it Friday, Jan. 24? Order within **6 hrs 31 mins** and choose **Two-Day Shipping** at checkout. Details  
Ships from and sold by Amazon.com.

This item is returnable

Style Name: **Minty Fresh Toothpaste (Triple Pack)**



\$5.59



--

- Leaves mouth and breath feeling refreshed
- Whitens teeth by gently removing surface stains
- Fights cavities
- Fights tartar build-up

# Span classification

Winfrey's best friend since their early twenties is Gayle King.

Person

Person

- Sequence tagging problem
- “BIO Tagging”
  - “Beginning”
  - “Inside”
  - “Outside”

# Sequence tagging

Winfrey's best friend since their early twenties is Gayle King.

The diagram illustrates sequence tagging for the sentence "Winfrey's best friend since their early twenties is Gayle King." Below the sentence, each word is followed by a small orange circle representing a tag. Brackets above the words group them into entities: "Winfrey's" is bracketed together with the tag "B-Person" below it; "best friend since their early twenties" is grouped together with the tag "O" below it; "is" is grouped with the tag "O"; and "Gayle King." is grouped together with the tag "B-Person I-Person" below it. The "B-Person" tag indicates the start of a person entity, while the "I-Person" tag indicates the continuation of the entity from the previous tag.

B-Person O O O O O O B-Person I-Person

Typically used for Named Entity Recognition

# OpenTag (Zheng et al, 2018)

- Span classification for relation extraction
- Data is product detail pages
  - No need to extract product
- Extracts product attributes such as brand and flavor from product title/description



In stock.

Get it as soon as Wednesday, Feb. 14 when you choose Two-Day Shipping at checkout.  
Ships from and sold by [Cunningham Collective](#).

Variety Pack Filet Mignon and Porterhouse Steak Dog Food (12 Count)  
Price: **\$92.60** & FREE Shipping

[Be the first to review this item](#)

- 6 trays of Filet Mignon flavor in meaty juices
- Cesar pet food has an irresistible taste with exceptional palatability to tempt even the fussiest dogs
- Formulated to meet the nutritional levels established by the AAFCO Dog Food Nutrient Profiles for maintenance

#### Product description

Variety pack includes: 6 trays of Filet Mignon flavor in meaty juices 6 trays of Porterhouse Steak flavor in meaty juices  
Cesar pet food has an irresistible taste with exceptional palatability to tempt even the fussiest dogs Formulated to meet the nutritional levels established by the AAFCO Dog Food Nutrient Profiles for maintenance Complete & balanced nutrition for small adult dogs Fortified with vitamins and minerals Packaged in convenient feeding trays with no-fuss, peel-away freshness seals Includes 6 Each Chicken & Liver

Variety Pack Filet Mignon and Porterhouse Steak Dog Food (12 count)

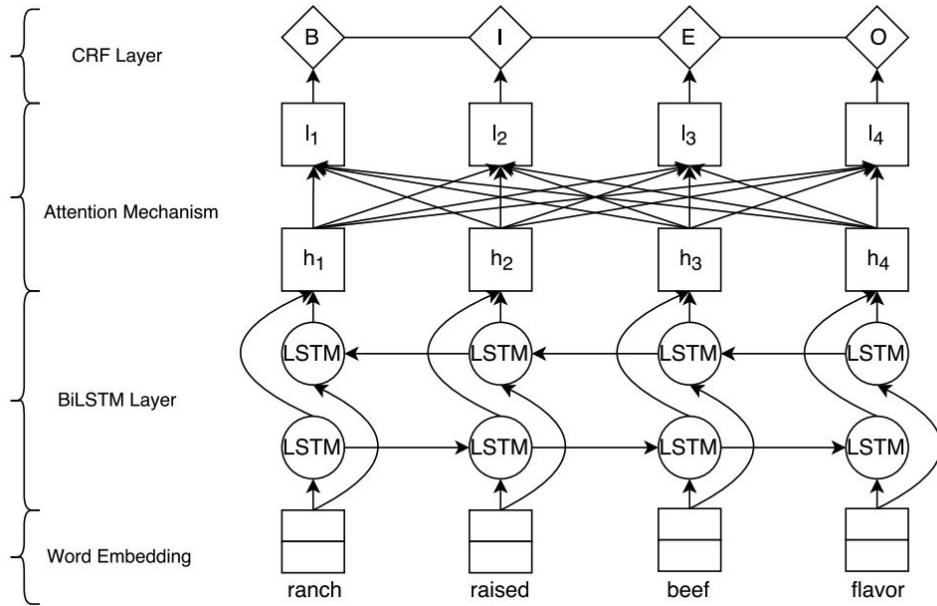
O O B-Flavor I-Flavor O B-Flavor I-Flavor O O O O

(ASIN B0001234567, has\_flavor, "Filet Mignon")

(ASIN B0001234567, has\_flavor, "Porterhouse Steak")

# Span classification: OpenTag

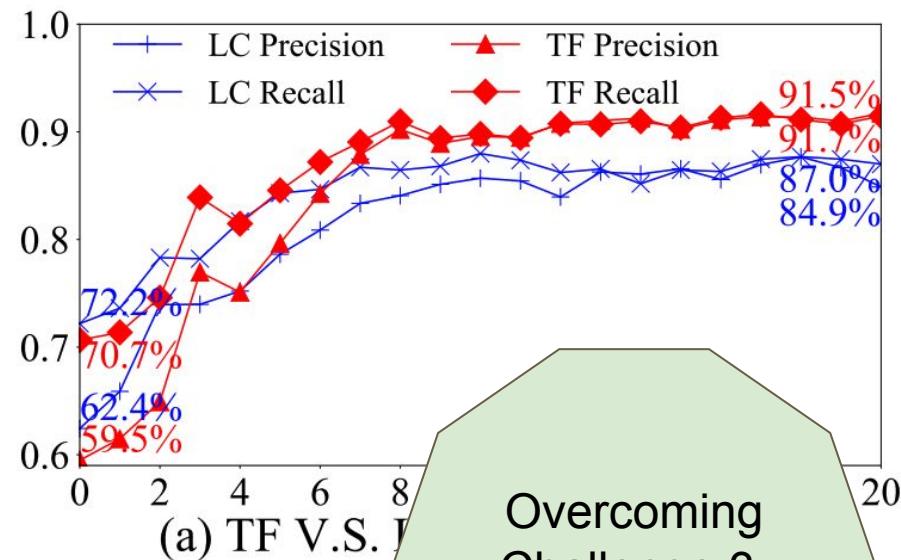
- Word embeddings capture word meaning
- LSTM layer captures word sequence information
- Attention layer allows interaction across sequence
- CRF layer enforces consistency



# Active Learning with OpenTag

Start with small amount of labeled data

Ask human to selectively label most informative datapoints



Overcoming  
Challenge 3:  
Active learning

# OpenTag Results

Datasets/Attribute	Models	Precision	Recall	Fscore
Dog Food: Title Attribute: Flavor	BiLSTM	83.5	85.4	84.5
	BiLSTM-CRF	83.8	85.0	84.4
	OpenTag	<b>86.6</b>	<b>85.9</b>	<b>86.3</b>
Camera: Title Attribute: Brand name	BiLSTM	94.7	88.8	91.8
	BiLSTM-CRF	91.9	<b>93.8</b>	92.9
	OpenTag	<b>94.9</b>	93.4	<b>94.1</b>
Detergent: Title Attribute: Scent	BiLSTM	81.3	82.2	81.7
	BiLSTM-CRF	<b>85.1</b>	82.6	83.8
	OpenTag	84.5	<b>88.2</b>	<b>86.4</b>

Relation extraction results with ~90% accuracy

# OpenTag: Summary

- Relation extraction as span classification via BiLSTM-CRF
- Pros:
  - Reduces relation extraction from span pair classification to single span classification
  - Active learning
- Cons:
  - Only works on text from detail page

# Method: Multi-task learning of span relationships

## Applied Example: DyGIE

How can we extract jointly extract entities, relationships, and events from any unstructured text?

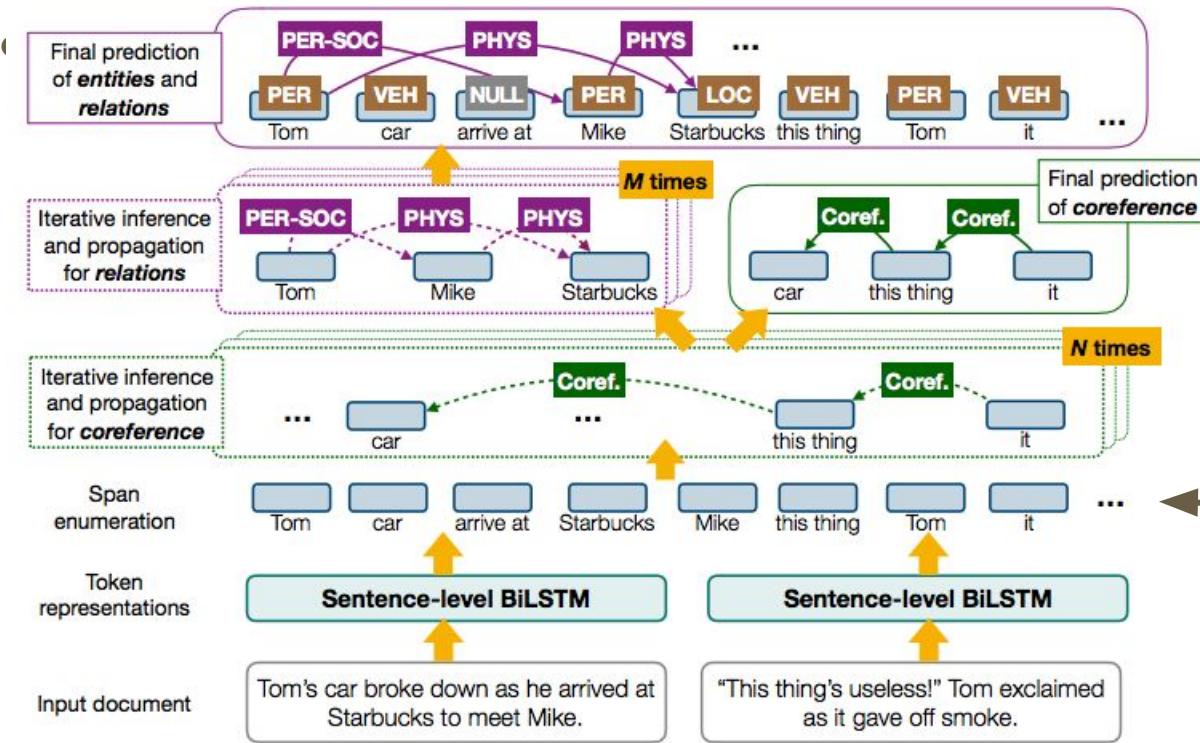
# DyGIE (Luan et al, 2019)

- Single model for NER, co-reference, relation extraction

# DyGIE (Luan et al, 2019)

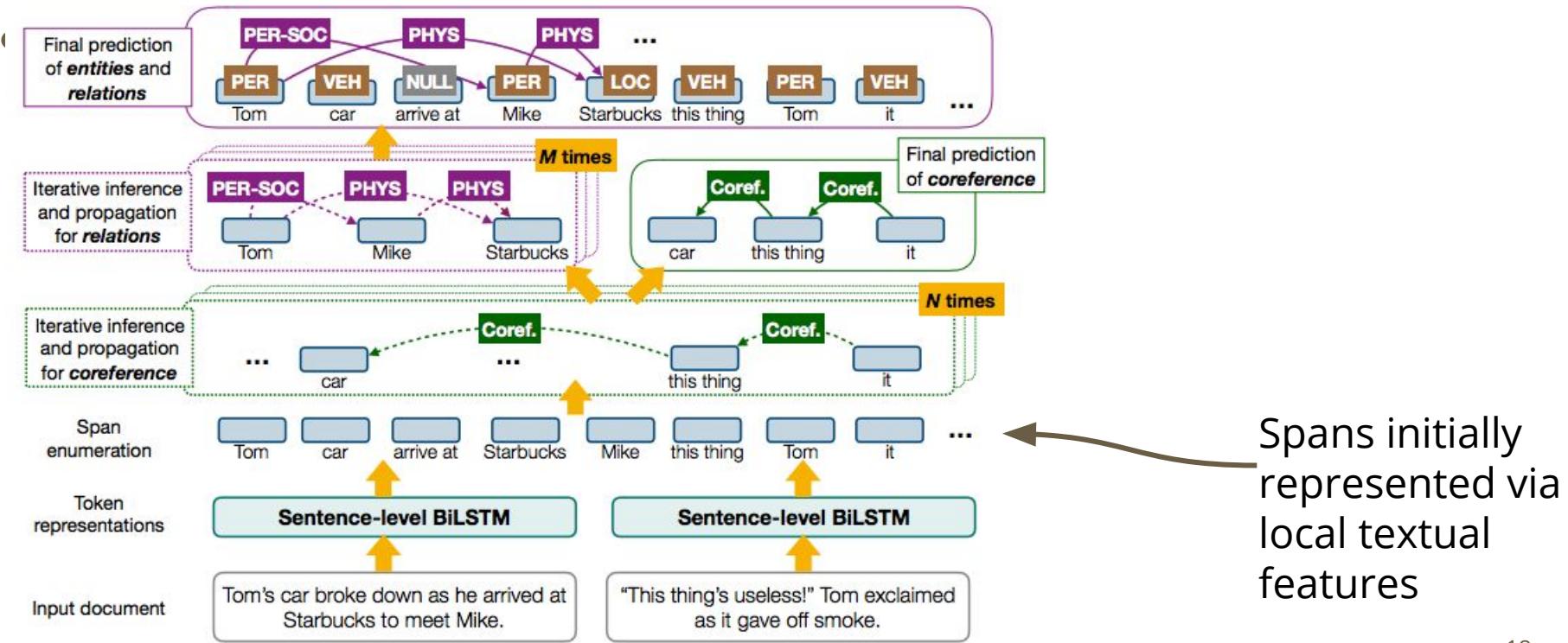
- Single model for NER, co-reference, relation extraction
  - Multi-task learning objective

# DyGIE (Luan et al, 2019)

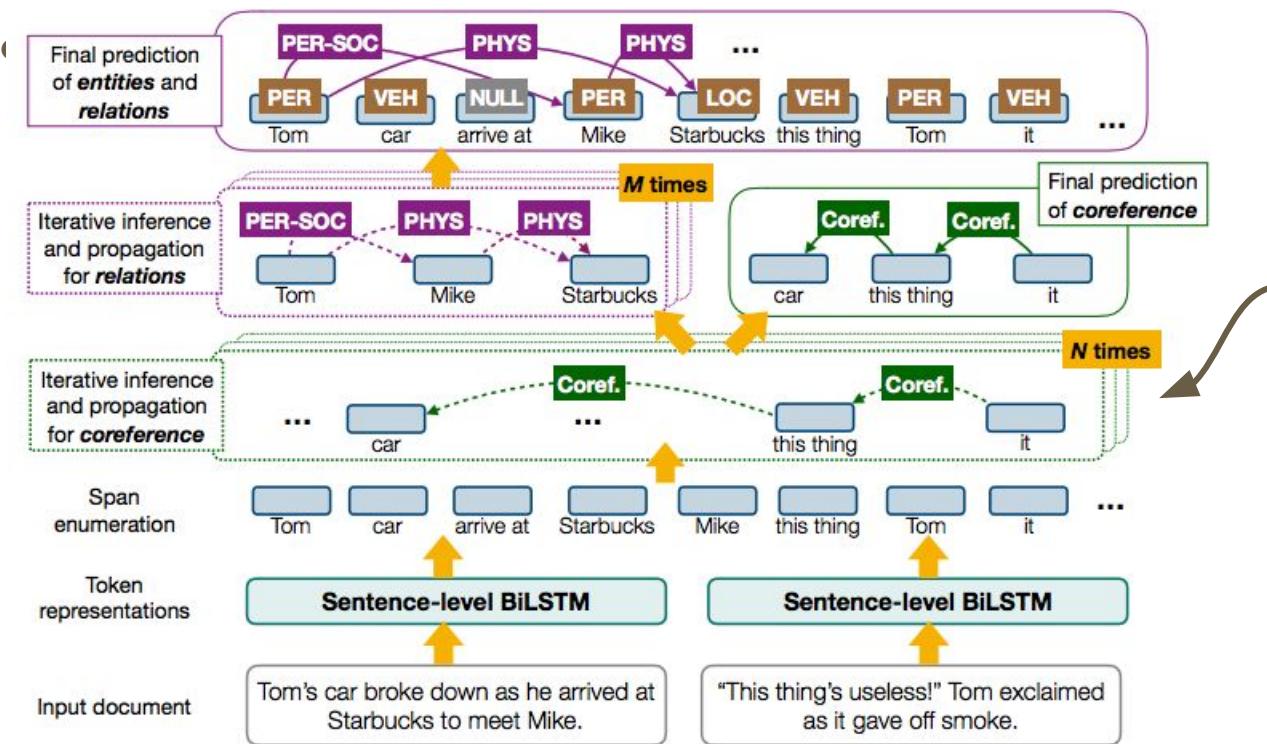


Enumerate  
all spans up  
to length  $L$

# DyGIE (Luan et al, 2019)



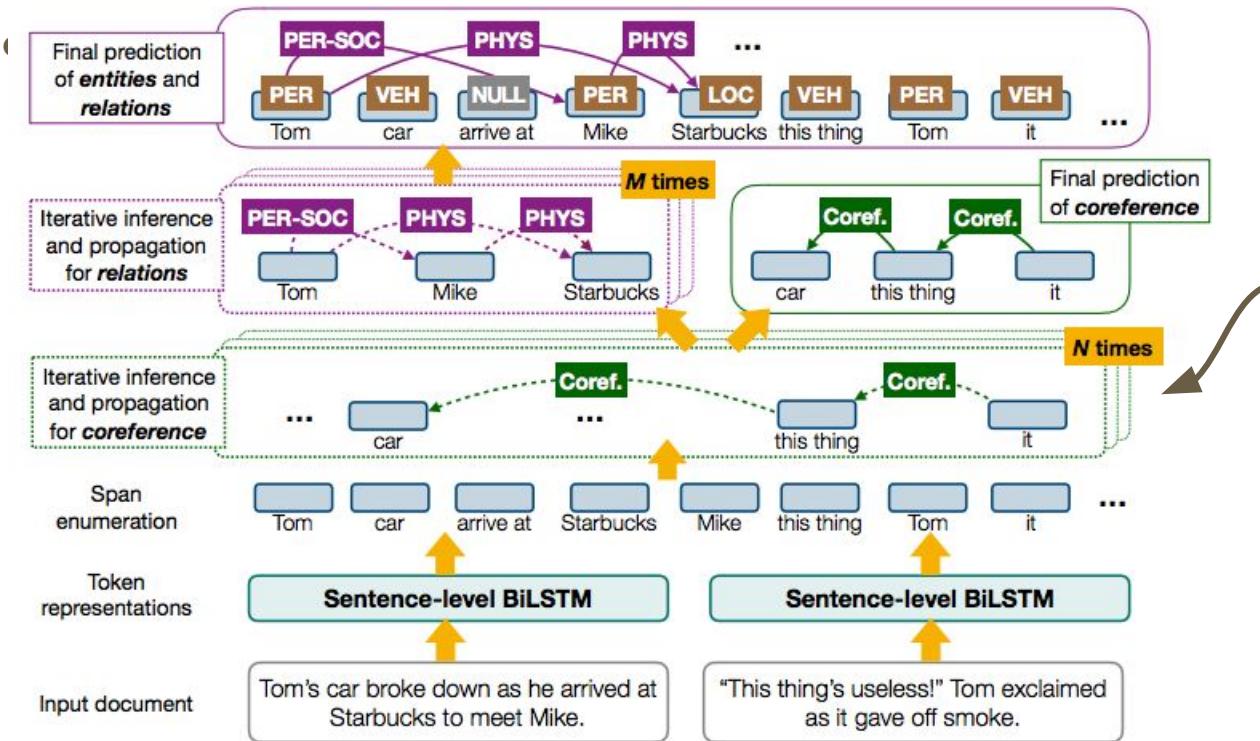
# DyGIE (Luan et al, 2019)



Construct graph:

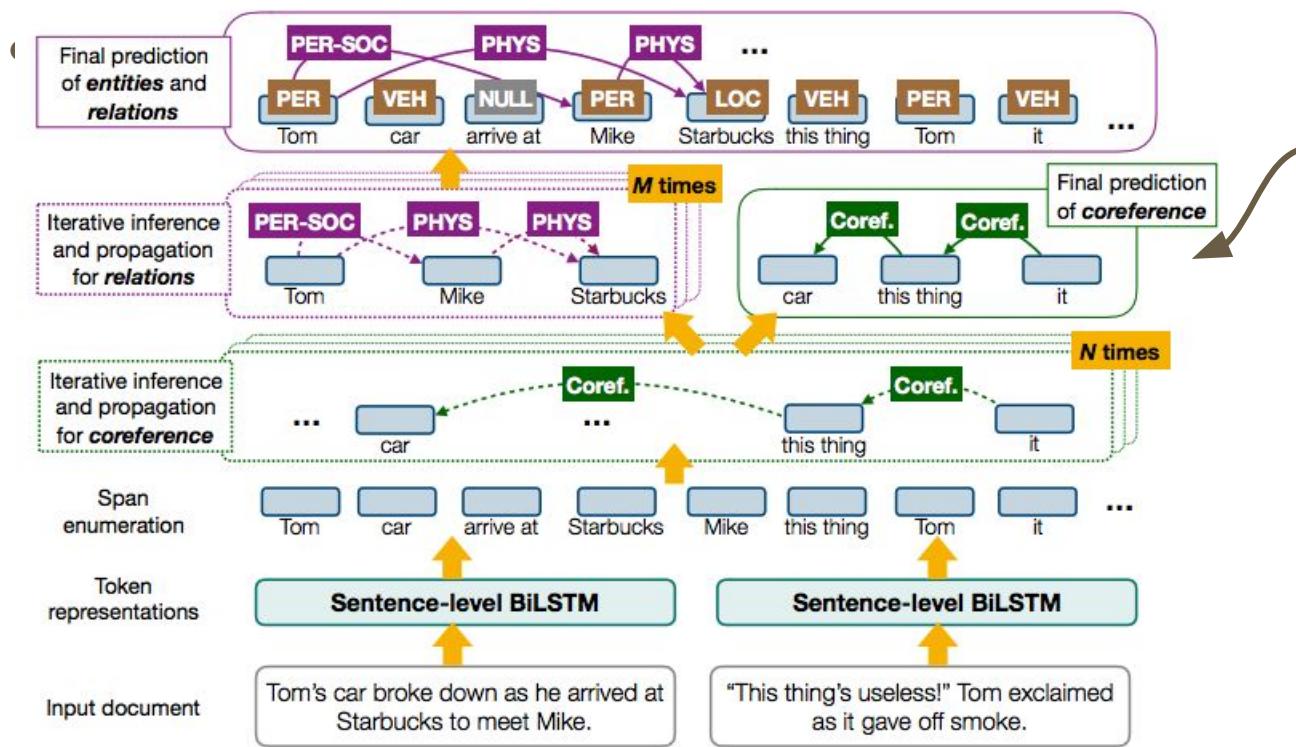
- Spans are nodes
- Edges are (potential) coreferences
- Edge weight indicates confidence

# DyGIE (Luan et al, 2019)



Iteratively propagate node information

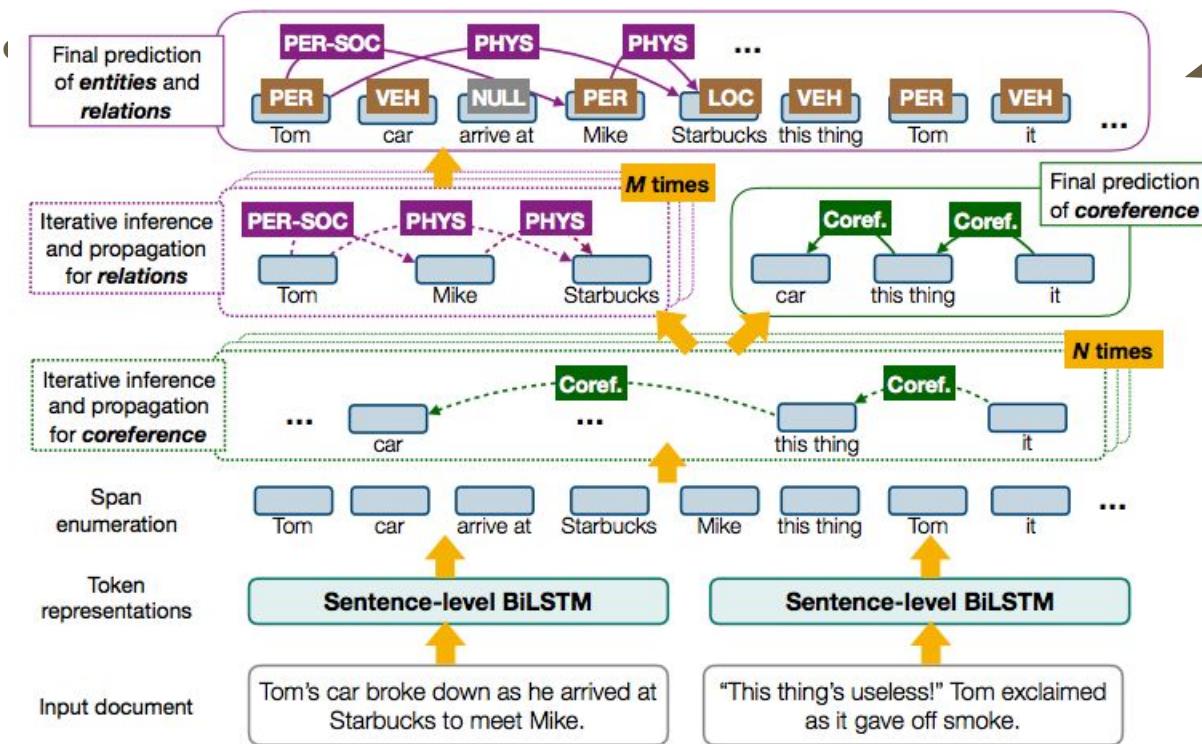
# DyGIE (Luan et al, 2019)



Repeat process for relations:

- Edges now indicate relation types

# DyGIE (Luan et al, 2019)

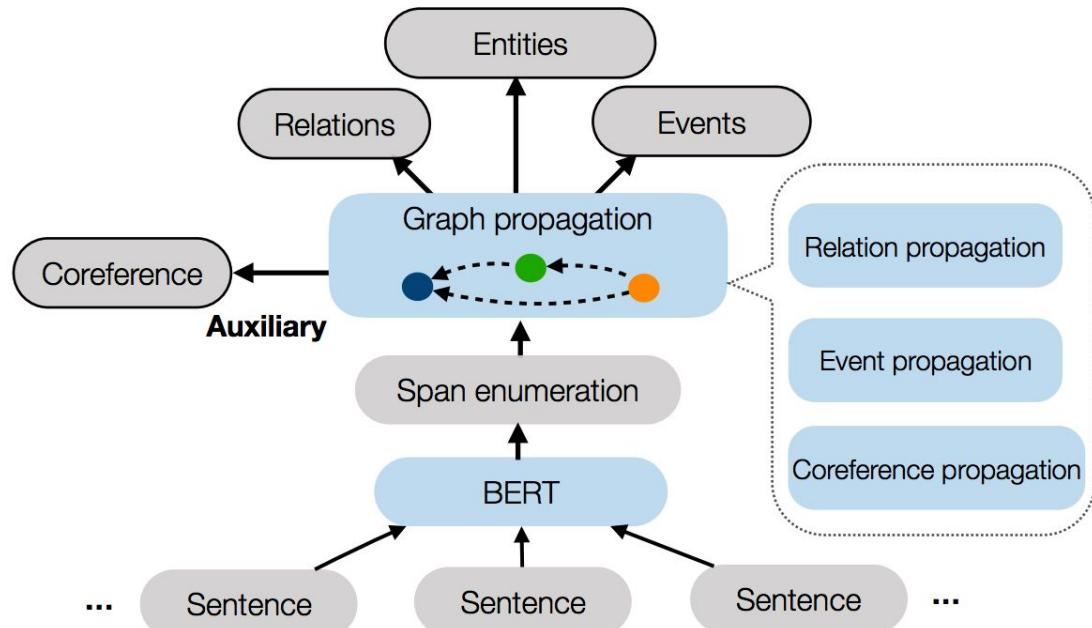


Use final representations to predict entity types and relations

# DyGIE++ (Wadden et al, 2019)

DyGIE++ adds events

Replaces word embeddings and LSTM with BERT word representations



# DyGIE++

State-of-the-art  
results across many  
datasets

Dataset	Task	SOTA	Ours	$\Delta\%$
ACE05	Entity	88.4	<b>88.6</b>	1.7
	Relation	63.2	<b>63.4</b>	0.5
ACE05-Event*	Entity	87.1	<b>90.7</b>	27.9
	Trig-ID	73.9	<b>76.5</b>	9.6
	Trig-C	72.0	<b>73.6</b>	5.7
	Arg-ID	<b>57.2</b>	55.4	-4.2
	Arg-C	52.4	<b>52.5</b>	0.2
SciERC	Entity	65.2	<b>67.5</b>	6.6
	Relation	41.6	<b>48.4</b>	11.6
GENIA	Entity	76.2	<b>77.9</b>	7.1
WLPC	Entity	79.5	<b>79.7</b>	1.0
	Relation	64.1	<b>65.9</b>	5.0

More accurate on  
newswire data

Scientific/medical  
text is more  
challenging

# DyGIE takeaways

- Builds span representations via graph propagation over span graph
- Pros:
  - Multi-task learning finds signal from different sources
  - Single model for all IE tasks
  - Handles overlapping spans
- Cons:
  - Still requires manually labeled training data
  - Still relatively small scale (single paragraph)

## **Method: Weak supervision**

## **Applied Example: Snorkel**

**How can we extract without manually labeling data?**

# Distant Supervision (Mintz et al, 2009)

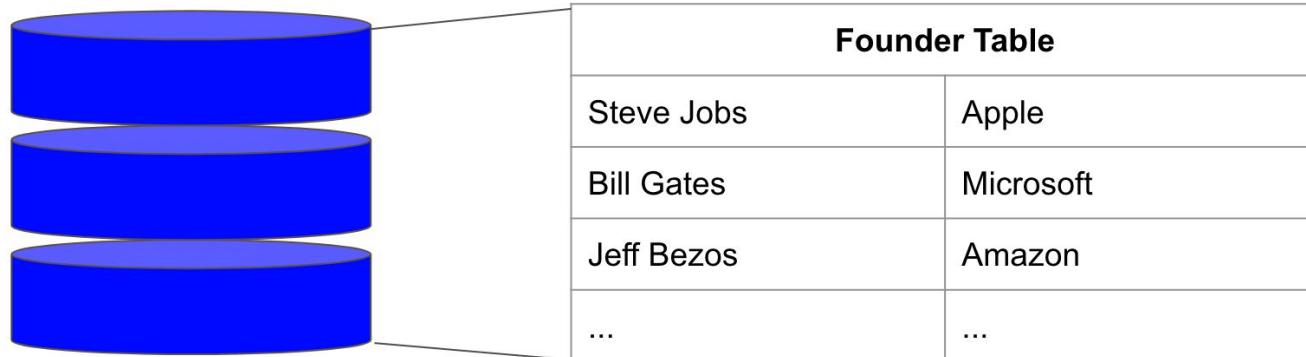
Automatically generate training data using existing knowledge

**Steve Jobs was the founder of Apple.**

# Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

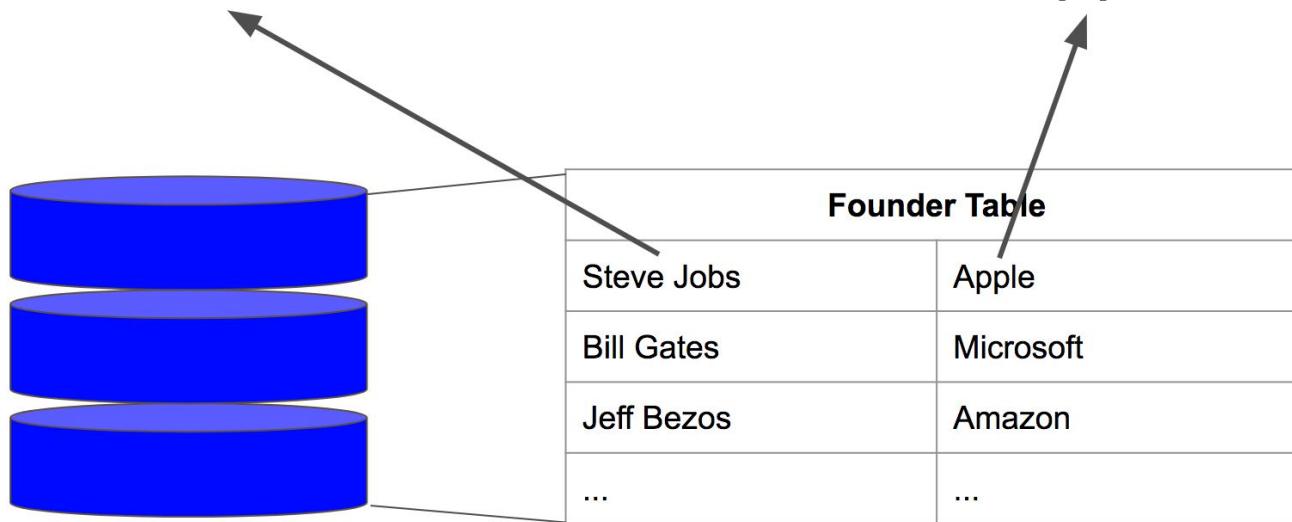
Steve Jobs was the founder of Apple.



# Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

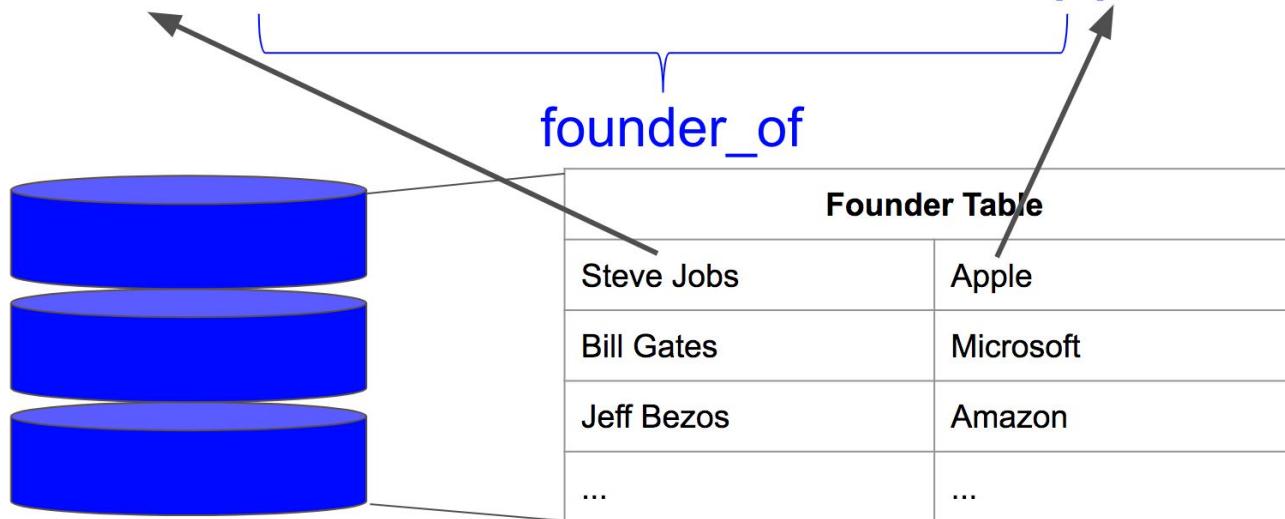
Steve Jobs was the founder of Apple.



# Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

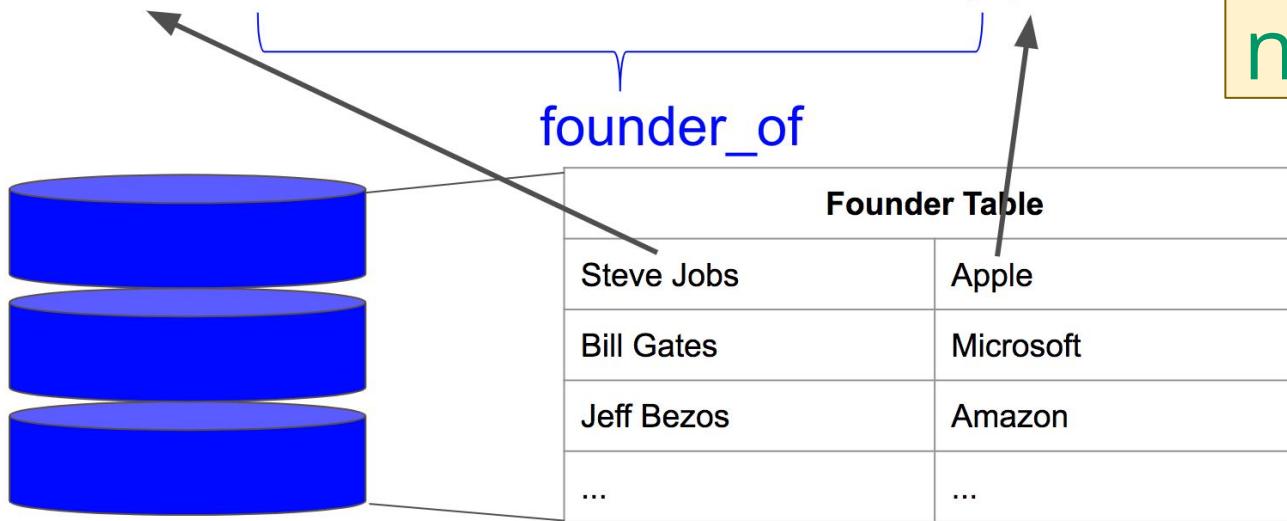
Steve Jobs was the founder of Apple.



# Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

Steve Jobs was the CEO of Apple.



Matching  
to KB is  
noisy!

# Distant Supervision (Mintz et al, 2009)

- Automatically create training data based on existing knowledge
- Pros:
  - Free training data
- Cons:
  - Training data is noisy
  - Assumes existing knowledge base

# Data Programming (Ratner et al, 2016)

- Often may have multiple sources of weak supervision
  - Distant supervision from a Knowledge Base
  - Heuristics / regular expressions
  - Noisy crowd-labeled data
  - Manually defined constraints
  - Extractions from an existing (and imperfect) IE system
- How can we most effectively learn from noisy data from different sources?

# Data Programming (Ratner et al, 2016)

Noisy labels from multiple “labeling functions”

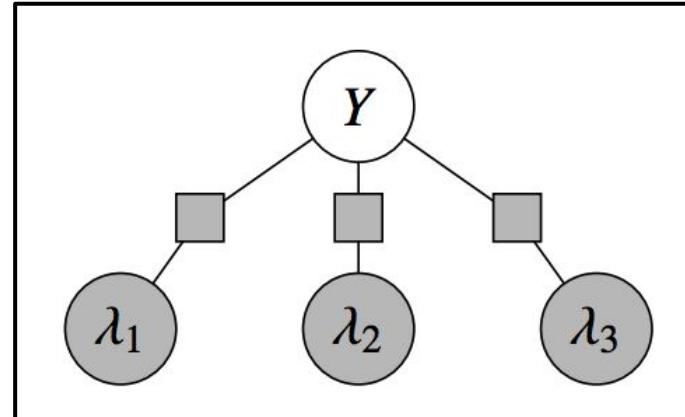
Generative model to “de-noise” training data

Learns which labeling functions are best for which data points

```
def lambda_1(x):
    return 1 if (x.gene,x.pheno) in KNOWN_RELATIONS_1 else 0

def lambda_2(x):
    return -1 if re.match(r'.*not\_cause.*', x.text_between) else 0

def lambda_3(x):
    return 1 if re.match(r'.*associated.*', x.text_between)
        and (x.gene,x.pheno) in KNOWN_RELATIONS_2 else 0
```



# Snorkel (Ratner et al, 2017)

Open source system implementing Data Programming paradigm

Interface allows user to easily create labeling functions

# Snorkel (Ratner et al, 2017)

**Input:** Labeling Functions,  
*Unlabeled data*

DOMAIN  
EXPERT



def lf1(x):  
 cid = (x.chemical\_id,  
 x.disease\_id)  
 return 1 if cid in KB else 0

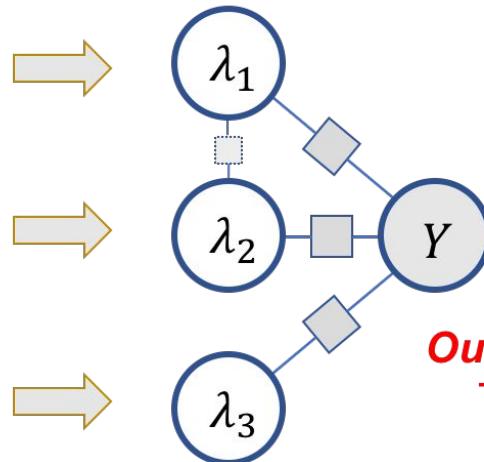
  

def lf2(x):  
 m = re.search(r'.\*cause.\*',  
 x.between)  
 return 1 if m else 0

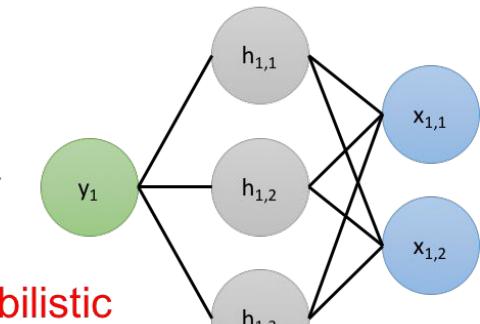
def lf3(x):  
 m = re.search(r'.\*not  
 cause.\*', x.between)  
 return 1 if m else 0

**Generative Model**

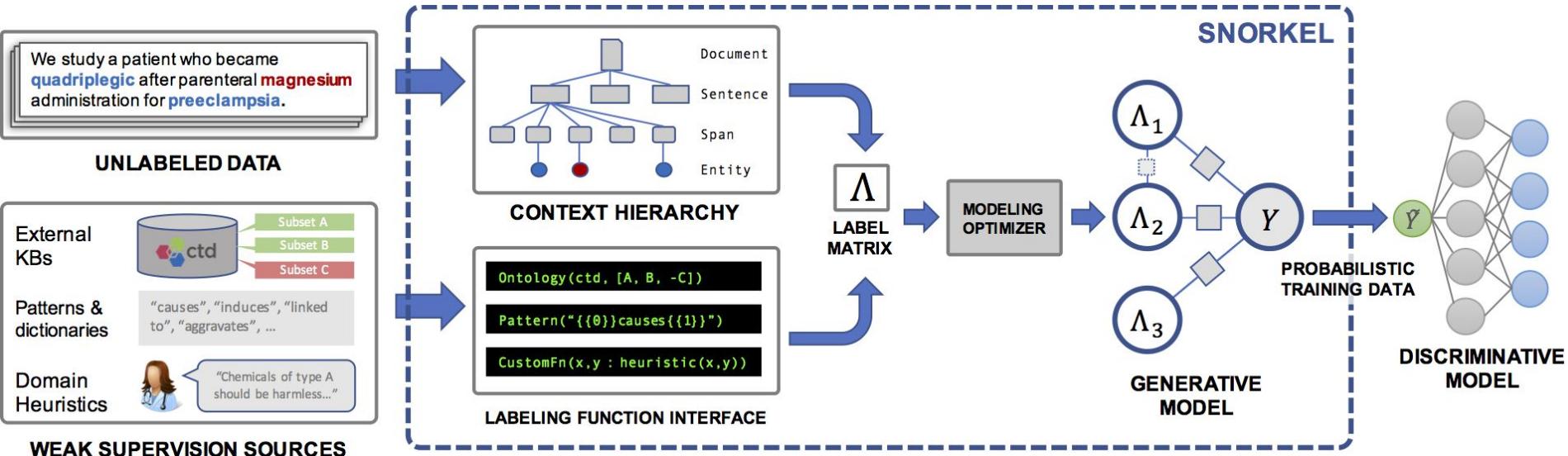


**Output:** Probabilistic  
Training Labels

**Noise-Aware  
Discriminative Model**



# Snorkel



```
from snorkel.labeling import labeling_function

@labeling_function()
def lf_contains_link(x):
    # Return a label of SPAM if "http" in comment text, otherwise ABSTAIN
    return SPAM if "http" in x.text.lower() else ABSTAIN
```

```
import re

@labeling_function()
def regex_check_out(x):
    return SPAM if re.search(r"check.*out", x.text, flags=re.I) else ABSTAIN
```

```
def keyword_lookup(x, keywords, label):  
    if any(word in x.text.lower() for word in keywords):  
        return label  
    return ABSTAIN
```

```
@labeling_function()  
def short_comment(x):  
    """Ham comments are often short, such as 'cool video!'"""  
    return HAM if len(x.text.split()) < 5 else ABSTAIN
```

# Snorkel

Relatively small number  
of labeling functions

Task	# LFs	% Pos.	# Docs	# Candidates
Chem	16	4.1	1,753	65,398
EHR	24	36.8	47,827	225,607
CDR	33	24.6	900	8,272
Spouses	11	8.3	2,073	22,195
Radiology	18	36.0	3,851	3,851

# Snorkel

Relatively small number  
of labeling functions

Task	# LFs	% Pos.	# Docs	# Candidates
Chem	16	4.1	1,753	65,398
EHR	24	36.8	47,827	225,607
CDR	33	24.6	900	8,272
Spouses	11	8.3	2,073	22,195
Radiology	18	36.0	3,851	3,851

Up to 39 point F1  
improvement over  
distant supervision

Task	Distant Supervision			Snorkel (Gen.)				Snorkel (Disc.)				Hand Supervision		
	P	R	F1	P	R	F1	Lift	P	R	F1	Lift	P	R	F1
Chem	11.2	41.2	17.6	78.6	21.6	33.8	+16.2	87.0	39.2	54.1	+36.5	-	-	-
EHR	81.4	64.8	72.2	77.1	72.9	74.9	+2.7	80.2	82.6	81.4	+9.2	-	-	-
CDR	25.5	34.8	29.4	52.3	30.4	38.5	+9.1	38.8	54.3	45.3	+15.9	39.9	58.1	47.3
Spouses	9.9	34.8	15.4	53.5	62.1	57.4	+42.0	48.4	61.6	54.2	+38.8	47.8	62.5	54.2

# Snorkel

Relatively small number  
of labeling functions

Task	# LFs	% Pos.	# Docs	# Candidates
Chem	16	4.1	1,753	65,398
EHR	24	36.8	47,827	225,607
CDR	33	24.6	900	8,272
Spouses	11	8.3	2,073	22,195
Radiology	18	36.0	3,851	3,851

Competitive with manual  
training labels

Task	Distant Supervision			Snorkel (Gen.)				Snorkel (Disc.)				Hand Supervision		
	P	R	F1	P	R	F1	Lift	P	R	F1	Lift	P	R	F1
Chem	11.2	41.2	17.6	78.6	21.6	33.8	+16.2	87.0	39.2	54.1	+36.5	-	-	-
EHR	81.4	64.8	72.2	77.1	72.9	74.9	+2.7	80.2	82.6	81.4	+9.2	-	-	-
CDR	25.5	34.8	29.4	52.3	30.4	38.5	+9.1	38.8	54.3	45.3	+15.9	39.9	58.1	47.3
Spouses	9.9	34.8	15.4	53.5	62.1	57.4	+42.0	48.4	61.6	54.2	+38.8	47.8	62.5	54.2

# Snorkel

- Tool for creating labeling functions to automatically create training data
- Pros:
  - Cheaply create lots of training data
  - More accurate than distant supervision
- Cons:
  - Still need to create well defined ontology

# OpenIE from Texts (Etzioni et al, 2011)

Bill Gates founded Microsoft in 1975.

Where are predicates from?

- Predicate: longest sequence of words as light verb construction
- Subject: learn left and right boundary
- Object: learn right boundary
- LR for triple confidence

# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- Part 1b: Unstructured Text: Methods (10 minutes)
- **Live Q&A (15 minutes)**
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- Part 3: Multi-modal Extraction (35 minutes)
- Live Q&A (15 minutes)

Please join us in the  
Zoom chat.

---

# Knowledge Collection from Semi-structured Text

---

Colin Lockard, **Prashant Shiralkar**,  
Xin Luna Dong, Hannaneh Hajishirzi

---



# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- Part 1b: Unstructured Text: Methods (10 minutes)
- Live Q&A (15 minutes)
- Break (30 minutes)
- **Part 2: Semi-structured and Tabular Text (40 minutes)**
- Part 3: Multi-modal Extraction and Conclusion (35 minutes)
- Live Q&A (15 minutes)

# **Section A. Semi-structured Text Extraction**

# Questions we will answer in this section

- What is a semi-structured source?
- How can we extract from semi-structured websites?

Semi-structured website pages

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ | SHARE

**Titanic** (1997)

PG-13 | 3h 14min | Drama, Romance | 19 December 1997 (USA)

7.8 990,317 Rate This

NOTHING CAN SEPARATE US  
COURTSHIP, ROMANCE, DREAM

LEONARDO DICAPRIO KATE WINSLET

**TITANIC**

Director: James Cameron  
Writer: James Cameron  
Stars: Leonardo DiCaprio, Kate Winslet, Billy Zane | See full cast & crew »

2:11 | Trailer | 25 VIDEOS | 838 IMAGES

See Showtimes & Tickets | ... | + Add to Watchlist

75 Metascore | From metacritic.com | Reviews 2,744 user | 325 critic | Popularity 131 (• 21)

# Questions we will NOT answer in this section

## Semi-structured records



★★★★★ (233)

Add to Compare

**Samsung 1TB T5 Portable Solid-State Drive (Black)**  
B&H # SAMUP1T5OBAM • MFR # MU-PA1T0B/AM

**KEY FEATURES**

- 1TB Storage Capacity
- USB 3.1 Type-C and Type-A Connections
- Up to 540 MB/s Data Transfer Rate
- USB Type-C & USB Type-A Cables Included

[More Information](#)

**In Stock**  
Order by 6pm to ship today

Free 2-Day Shipping

**\$169.99**

1 **Add to Cart**

Add to Wish List

 SmartGift Available



★★★★★ (367)

Add to Compare

**LaCie 2TB Rugged Mini USB 3.0 External Hard Drive**  
B&H # LARM02 • MFR # LAC9000298

**KEY FEATURES**

- 2TB Storage Capacity
- USB 3.0/3.1 Gen 1 Interface
- Up to 130 MB/s Data Transfer Speed
- Bus Powered

[More Information](#)

**In Stock**  
Order by 6pm to ship today

Free 2-Day Shipping

**\$99.99**

1 **Add to Cart**

Add to Wish List

 SmartGift Available

⚡ Updated Model Available

# What is a semi-structured website?

The screenshot shows the IMDb profile page for Rita Moreno. At the top, there's a search bar and navigation links for 'Movies, TV & Showtimes', 'Celebs, Events & Photos', 'News & Community', and 'Watchlist'. A promotional banner for Prime Video is visible. The main content features a large photo of Rita Moreno on the left, with her name 'Rita Moreno' in a green box above it. Below her name are two orange boxes labeled 'Actress' and 'Soundtrack'. There are links to 'View Resume' and 'Official Photos'. A large orange box highlights her biography, which reads: 'Rita Moreno has had a thriving acting career for the better part of six decades. One of the very few performers (and the very first) to win an Oscar, an Emmy, a Tony and a Grammy, she was born Rosita Dolores Alverio in Humacao, Puerto Rico, on December 11, 1931, to seamstress Rosa Maria (Marcano) and farmer Francisco Jose "Paco" Alverio. She and ... See full bio ». Another orange box highlights her birth information: 'Born: December 11, 1931 in Humacao, Puerto Rico'. At the bottom, there are links to 'More at IMDbPro »' and 'Contact Info: View agent, publicist, legal on IMDbPro'. A footer indicates '250 photos | 30 videos »'.

**Topic entity:** a real-world entity that is the focus of the detail page

Image

Short unstructured text

Relations as key-value pairs

## Actress (156 credits)

Hide 

### Elena of Avalor (TV Series)

Queen Camila

- Song of the Sirenas (2018) ... Queen Camila (voice)

2018

### Nina's World (TV Series)

Abuelita

- The Best Ending Ever! (2018) ... Abuelita (voice)
- Carlos' Winning Shirt (2018) ... Abuelita (voice)
- Nina Live (2018) ... Abuelita (voice)
- Nina in Charge (2018) ... Abuelita (voice)
- Nina's Seaside Rescue (2018) ... Abuelita (voice)

2015-2018

Show all 78 episodes

### One Day at a Time (TV Series)

Lydia Riera

- What Happened (2018) ... Lydia Riera
- Exclusive (2018) ... Lydia Riera
- To Zir, With Love (2018) ... Lydia Riera
- Storage Wars (2018) ... Lydia Riera
- Citizen Lydia (2018) ... Lydia Riera

2017-2018

Show all 26 episodes

### Torch

Aunt Francine

2017/I

## Information in a list

## Personal Details

Edit

**Other Works:** Stage: Appeared (Broadway debut) in "Skydrift" on Broadway. Written by Harry Kleiner. Scenic Design / Costume Design by Motley. Directed by Roy Hargrave. Belasco Theatre: 13 Nov 1945-17 Nov 1945 (7 performances). Cast: Wolfe Barzell (as "Mr. Bucelli"), William Chambers (as "Pvt. Edward Freling"), Zachary A. Charles (as "Pvt. Mario ... See more »

**Publicity Listings:** 1 Print Biography | 1 Interview | 7 Articles | 2 Pictorials | 6 Magazine  
Cover Photos | See more »

**Official Sites:** Official Site | Twitter

**Alternate Names:** Rita Moreno Gordon | Rosita Moreno

**Height:** 5' 2½" (1.59 m)

## Did You Know?

Edit

**Personal Quote:** [Her Oscar acceptance speech] I can't believe it! Good Lord! I'll leave you with that. See more »

**Trivia:** Awarded a Kennedy Center Honor in 2015. See more »

**Star Sign:** Sagittarius

Rich relationships in key-value pair format

IMDb is an example, with millions of such semi-structured pages about celebrities and movies.

# Semi-structured websites are everywhere!

40-50% of content on the Web is templates (Gibson WWW'05)

The screenshot shows a movie page for 'Made in China' (2019) on the BollywoodMDB.com website. The page includes a banner, director information, and a cast list.

**BollywoodMDB.com**  
Movies / Celebrities

HOME MOVIES ▾ Movie Calendar 2018 REVIEWS INTERVIEWS BOX OFFICE VIDEOS ▾  
Home > Movies > Made In China

**Made in China** (2019)

Banner: Miodock Films  
Director: Mikhil Musole  
Producer: Dinesh Vijan  
Star: Rajkummar Rao, Mouni Roy, Boman Irani ... see full cast & crew

Bollywood films

The screenshot shows a movie page for 'Twisted (Short film)' on the NMDB website. The page includes a trailer, production details, and a cast list.

**NMDB** MOVIES ▾ TV SHOWS ACTORS ▾ CREW ▾ EVENTS

**Twisted (Short film)**

Daniel Ademinokan's "TWISTED" Trailer

Year of production: 2014  
Running Time: 2:12 mins  
Written by: Daniel Ademinokan  
Produced by: Daniel Ademinokan  
Directed by: Daniel Ademinokan  
Starring: Stella Damasus, Rob Byrnes, Matt Meisen and David Ademinokan

Nigerian films

The screenshot shows a movie page for 'Abegweit' on the ONF NFB website. The page includes credits and a synopsis.

**ONF NFB**

**Abegweit**

Serge Morin  
1998 | 1 h 11 min  
CC CAMPUS

AVAILABLE ON DVD

SYNOPSIS EDUCATION

A day-to-day record of the construction of the Confederation Bridge

**CREDITS**

DIRECTOR	SCRIPT	PRODUCER	CAMERA
Serge Morin	Serge Morin	Pierre Bernier Diane Poitras	Marc Paulin
SOUND	EDITING	RE-RECORDING	SOUND EDITING
Georges Hannan	Fernand Bélanger	Serge Boivin Jean Paul Vilard	Fernand Bélanger Claude Langlois
NARRATOR	MUSIC		PARTICIPATION
Alex Madsen	Richard Gibson		Francine Blais Peter Briden Ralph Murray Guy Cormier Jim Feltham Kim Gallant Maurice Gallant Joe Ghiz Aldeene Giannelia Alexis Giannelia Paul Giannelia Pat Hepditch Betty Howatt Hubert Jacquin Ronnie-Gilles LeBlanc

Canadian films

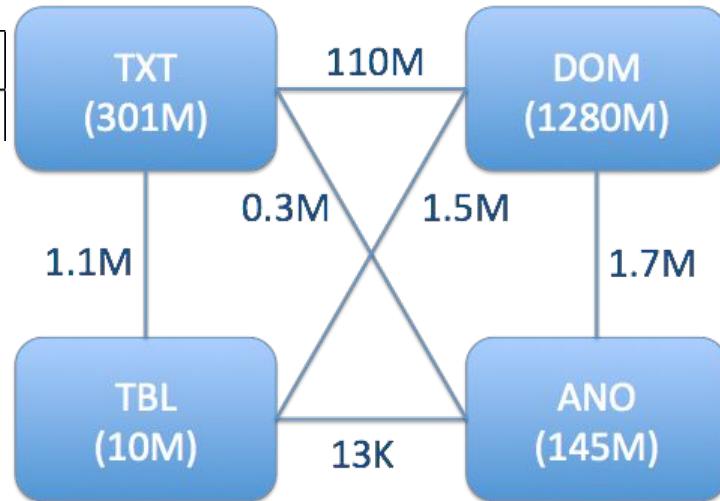
# Characteristics of semi-structured websites

- **Data rich:** websites are HTML templates populated by underlying database records
- **Distinct page per domain entity:** each detail page is about a distinct topic entity in the domain
- **Attributes as key-value pairs:** attribute names and values are often found in key-value format
- **DOM tree:** Each page can be represented as a DOM tree
- **Text extraction:** Each textual value can be located by applying an XPath to the DOM tree page representation

# Why extract from semi-structured websites?

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

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



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

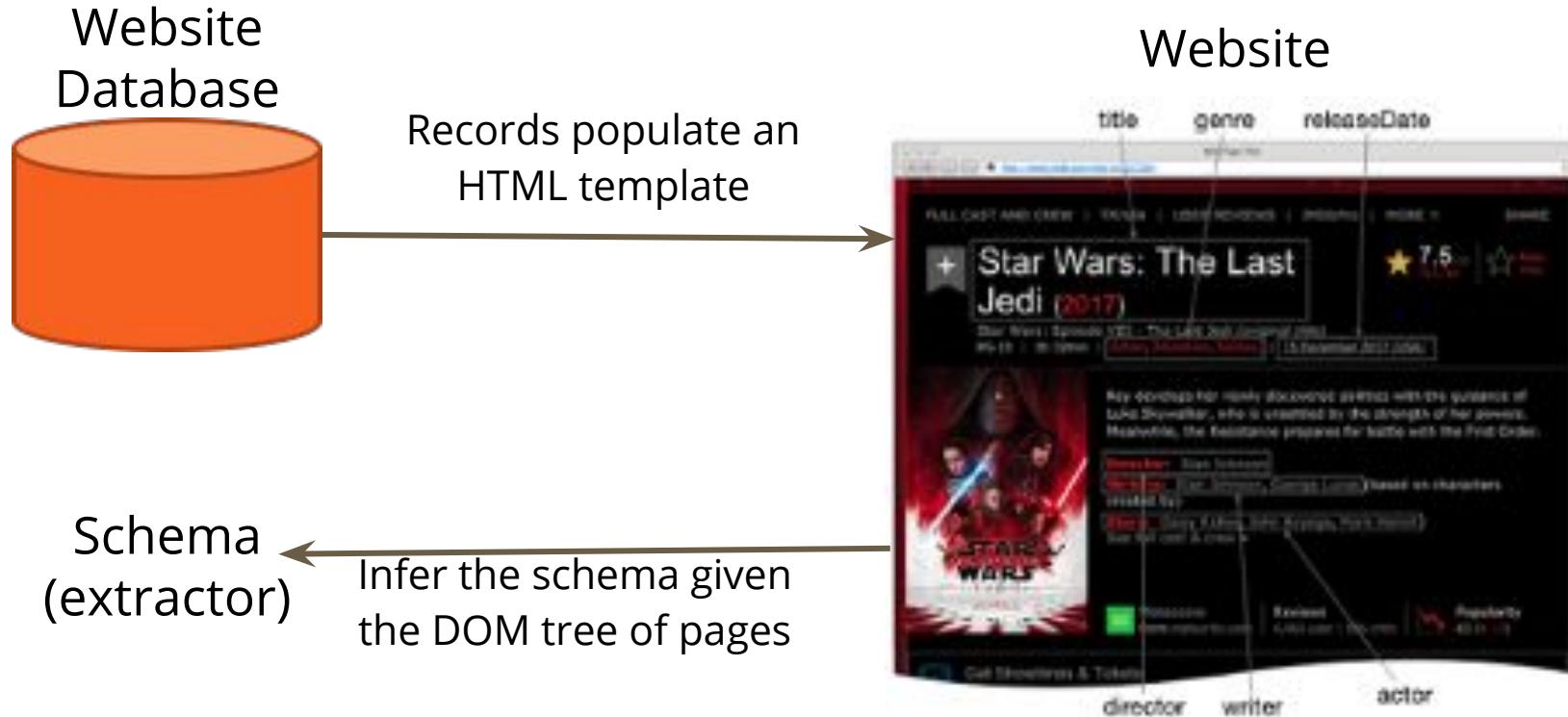


Low accuracy of various  
DOM extractors

\*Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). Knowledge vault: a web-scale approach to probabilistic knowledge fusion. KDD '14.

\*Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., & Zhang, W. (2014). From Data Fusion to Knowledge Fusion. VLDB.

# What is semi-structured website extraction?



# Entity detail page extraction problem

## Input:

A semi-structured website (same HTML template)

Optionally, a set of attributes of interest

## Extract:

The text indicating the attribute values



## Rita Moreno



Actress | Soundtrack

[View Resume](#) | [Official Photos](#) »

Rita Moreno has had a thriving acting career for the better part of six decades. One of the very few performers (and the very first) to win an Oscar, an Emmy, a Tony and a Grammy, she was born Rosita Dolores Alverio in Humacao, Puerto Rico, on December 11, 1931, to seamstress Rosa María (Marcano) and farmer Francisco José "Paco" Alverio. She and ... [See full bio](#)

Born: [December 11, 1931](#) at [Humacao, Puerto Rico](#)

[More at IMDbPro](#) »

Official Sites: [Official Site](#) | [Twitter](#)

Alternate Names: Rita Moreno Gordon | Rosita Moreno

Height: [5' 2 1/2" \(1.59 m\)](#)

### Did You Know?

Edit

**Personal Quote:** There was nobody that I could look up and say "That's somebody like me". Which is probably why I'm now known in my community as 'La Pionera', or the Pioneer. I really don't think of myself as a role model. But it turns out that I am to a lot of the Hispanic community. Not just in show business, but in life. But that's what happens when you're first, right? [See more](#) »

**Trivia:** Mother of Fernanda Gordon. [See more](#) »

**Star Sign:** [Sagittarius](#)

Extraction of (subject, predicate, object) triples from given semi-structured webpages.

## Records as triples

("Rita Moreno", birthDate, "December 11, 1931")

("Rita Moreno", birthPlace, "Humacao, Puerto Rico")

("Rita Moreno", height, "5' 2 1\2" (1.59 m))

("Rita Moreno", starsign, "Sagittarius")

....

# Why is semi-structured website extraction hard?

- Diversity:
  - Layout: key-value pairs, tables, lists, records



Vertical layout

Horizontal layout

A screenshot of the movie 'CENTRAL STATION (1998)'. It shows the 1998 Academy Award Nominations for Best Actress, featuring Fernanda Montenegro. The page lists the cast members: 'Vincius De Oliveira as Josue', 'Fernanda Montenegro as Dora', 'Soia Lira as Ana', and 'Marilia Pêra as Irene'. The 'Written by' section is highlighted with a red box and contains 'João Emanuel Carneiro' and 'Marcos Bernstein'. The 'Directed by' section contains 'Walter Salles'. The entire image is labeled 'Horizontal layout'.

Cast
Vincius De Oliveira as Josue
Fernanda Montenegro as Dora
Soia Lira as Ana
Marilia Pêra as Irene
Written by
João Emanuel Carneiro
Marcos Bernstein
Directed by
Walter Salles

# Why is semi-structured website extraction hard?

- Diversity:
  - Terms: “Birthday” and “Birthplace” (Site 1) vs. “Born” (Site 2)

The screenshot shows the Rotten Tomatoes homepage with a search bar and trending sections. Below, Tom Cruise's profile is displayed. His photo is on the left, and his name "Tom Cruise" is in bold. Underneath, it says "Highest Rated: 🍅 97% Mission: Impossible -" and "Lowest Rated: 🍂 5% Cocktail (1988)". Two specific fields, "Birthday" and "Birthplace", are highlighted with red boxes. The "Birthday" field contains "Jul 3, 1962" and the "Birthplace" field contains "Syracuse, New York". A large block of text below discusses his education and racing interests, ending with a "More" link.

The screenshot shows a movie profile for Tom Cruise. At the top, it lists "Actor | Producer | Soundtrack". Below is a large photo of Tom Cruise smiling. To the right is a smaller thumbnail for a trailer. The trailer thumbnail shows Tom Cruise in a flight suit. The text "0:50 | Trailer" is at the bottom of the thumbnail. A bio at the bottom starts with "In 1976, if you had told fourteen year-old Franciscar Mapother IV that one day in the not too distant futur". A "Born" field is highlighted with a red box, containing "July 3, 1962 in Syracuse, New York, USA".

In 1976, if you had told fourteen year-old Franciscar Mapother IV that one day in the not too distant future 100 movie stars of all time, he would have probably was to join the priesthood. Nonetheless, this sensitiv  
Born: July 3, 1962 in Syracuse, New York, USA

More at IMDbPro »  
Contact Info: View agent, publicist, legal on IMDB

# Why is semi-structured website extraction hard?

- **Diversity:**
  - Layout: key-value pairs, tables, lists, records
  - Terms: “Birthday” and “Birthplace” (Site 1) vs. “Born” (Site 2)
  - Format: fonts, abbreviations, e.g. “T. Cruise” vs. “Tom Cruise”
  - Language: “place of birth” (English) vs. “출생지” (Korean)
  - Domain: music, movies, books, sports, ..
- **Mismatch in values:**
  - “Aug 4” (imprecise) vs. “Aug 4, 1961” (complete)
  - B. Obama’s birthplace as “Kenya” (false) vs. “Hawaii” (true)
- **Training data scarcity:** no training data for each website template

# Opportunities

- **Consistency within a website template:**
    - Topic entities have their own page with similar format
    - Key-value pairs corresponding to (relation, object) pairs have similar layout

The image shows the movie poster for "Central Station" (1998) on the left, featuring a woman holding a child. To the right is a large, close-up photograph of a woman's face, looking slightly to the side with a serious expression. A play button icon is overlaid on the right side of the image.

An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

Director: Walter Salles

Writers: Marcos Bernstein, João Emanuel Carneiro | 1 more credit »

**Stars:** Fernanda Montenegro, Vinícius de Oliveira, Marília Pêra | See full cast & crew »

The story of America as seen through the eyes of the former Secretary of Defense under President John F. Kennedy and President Lyndon Baines Johnson, Robert McNamara

Director: Errol Morris

**Stars:** Robert McNamara, John F. Kennedy, Fidel Castro | See full cast & crew

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ SHARE

# + Star Wars: Episode VIII - The Last Jedi (2017)

PG-13 | 2h 32min | Action, Adventure, Fantasy | 15 December 2017 (USA)



7.0 10  
522,260

Rate This

0:27 | Trailer

63 VIDEOS | 810 IMAGES

Rey develops her newly discovered abilities with the guidance of Luke Skywalker, who is unsettled by the strength of her powers. Meanwhile, the Resistance prepares for battle with the First Order.

Director: Rian Johnson

**Writers:** Rian Johnson, George Lucas (based on characters created by)

**Stars:** Daisy Ridley, John Boyega, Mark Hamill | See full cast & crew »

# Opportunities

- **Consistency within a website template:**
  - Topic entities have their own page with similar format
  - Key-value pairs corresponding to (relation, object) pairs have similar layout
- **Informativeness:**
  - Multiple attributes per entity
  - Diverse attribute values across entities

# Opportunities

- **Consistency within a website template:**
  - Topic entities have their own page with similar format
  - Key-value pairs corresponding to (relation, object) pairs have similar layout
- **Informativeness:**
  - Multiple attributes per entity
  - Diverse attribute values across entities
- **Uniqueness:** only one or at most two detail pages per entity
- **Redundancy across websites:**
  - Instance-level: attribute values
  - Ontology/schema-level: attributes

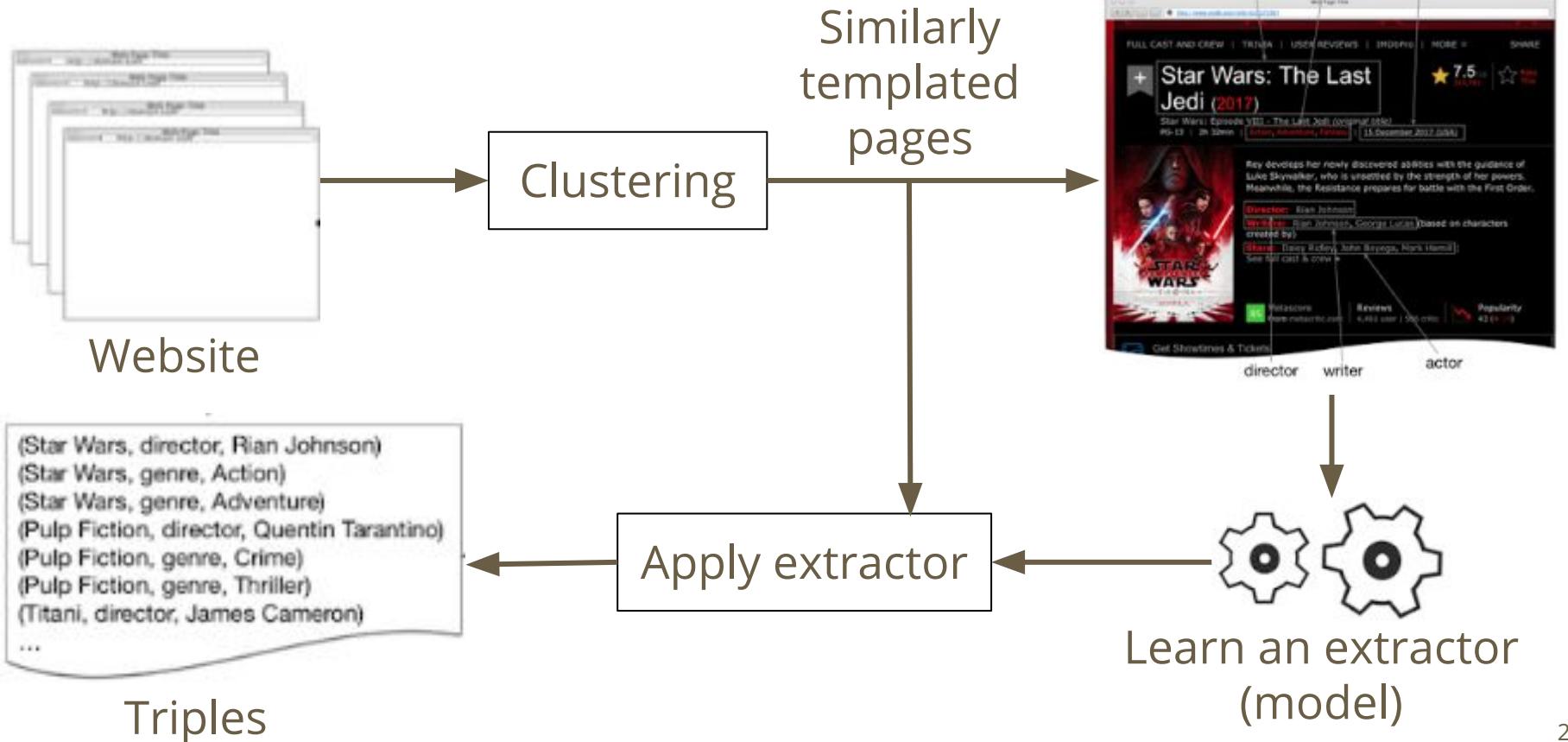
# Key differences with text

Dimension	Unstructured text	Semi-structured websites
Input unit	Sentence or page	Entity page
Consistency	Grammatical pattern	Page template
Entity pair relation	Explicit within a sentence or paragraph	Explicit to the left/top/right of object
NER tools available?	Yes	No
Context	Rich, often ambiguous	Short, clean

# Short Answers

- **Consistency**
  - Leverage general key-value pair consistency universal in templates
  - Leverage site-level consistency in layout and presentation
- **Training data**
  - Use distant supervision to generate cheap, but noisy training data
- **OpenIE**
  - Discover new relations by label propagation

# High-level approach for extraction



# Methods for semi-structured website extraction

- **Closed IE:** extraction for a closed, pre-determined set of relations
- **Open IE:** extraction for open, unseen set of relations on the Web

## Closed IE

- Wrapper induction (IJCAI'97, VLDB'01, SIGMOD'09, ICDE'11, VLDB'14...)
- Distant supervision
  - Labeled seed sites (SIGIR'11)
  - Linked Open Data (AAAI'15)
  - Knowledge base (VLDB'18)

## Open IE

- WEIR (VLDB'13)
- Label propagation (NAACL'19)

**How do we build a high-quality extractor for a website template?**

# Wrapper induction (Kushmerick, IJCAI'97)

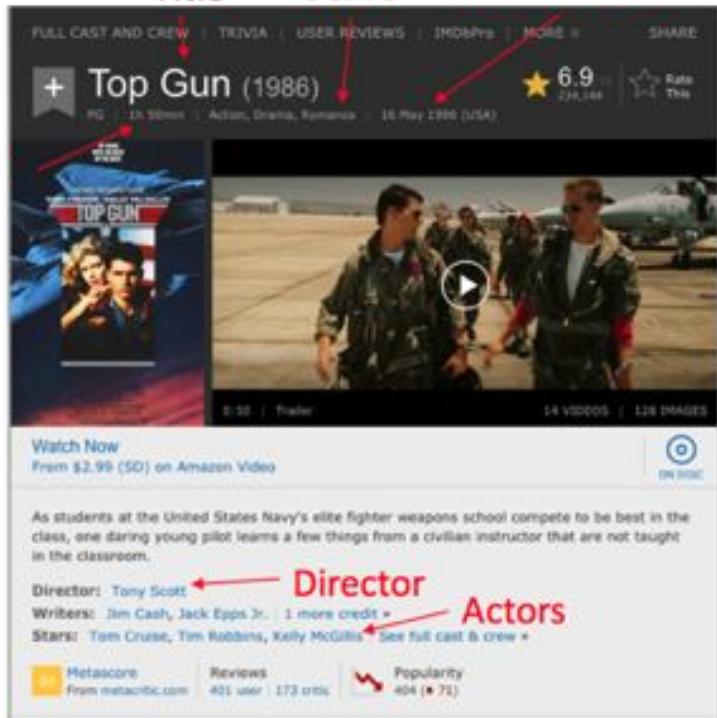
## What is wrapper induction?

Semi-structured webpages are created by populating an HTML template with records from an underlying database.

Wrapper induction is the task of inferring the schema (rules) for each relation in the database given the DOM tree of pages, and the rules learned are called *wrappers*.

# Wrapper induction

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")

# Challenges to wrapper induction

**Minor variations:** Same relation present in different DOM tree node

Page 1: //\*[@id="title-overview-widget"]/div[2]/div[1]/div[4]/a

Page 2: //\*[@id="title-overview-widget"]/div[2]/div[1]/div[3]/a

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ SHARE

## Central Station (1998)

Central do Brasil (original title)  
R | 1h 50min | Drama | 20 November 1998 (USA)



8.0 /10 33,811 Rate This

1:53 | Trailer 1 VIDEO | 32 IMAGES

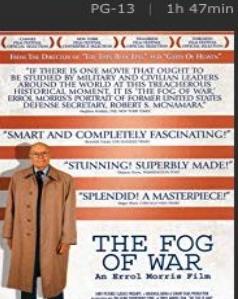
An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

**Director:** Walter Salles  
**Writer:** Marcos Bernstein, João Emanuel Carneiro | 1 more credit »  
**Stars:** Fernanda Montenegro, Vinícius de Oliveira, Marília Pêra | See full cast & crew »

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ SHARE

## The Fog of War: Eleven Lessons from the Life of Robert S. McNamara (2003)

PG-13 | 1h 47min | Documentary, Biography, History | 5 March 2004 (USA)



8.1 /10 22,171 Rate This

2:07 | Trailer 2 VIDEOS | 11 IMAGES

The story of America as seen through the eyes of the former Secretary of Defense under President John F. Kennedy and President Lyndon Baines Johnson, Robert McNamara.

**Director:** Errol Morris  
**Stars:** Robert McNamara, John F. Kennedy, Fidel Castro | See full cast & crew »

# Challenges to wrapper induction

**Optional/missing sections:** Same DOM node may correspond to different relations. E.g. "Writers:" vs. "Stars:"

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ SHARE

## Central Station (1998)

Central do Brasil (original title)

R | 1h 50min | Drama | 20 November 1998 (USA)



8.0 /10  
33,811 Rate This

+

Central Station (1998)

Central do Brasil (original title)

R | 1h 50min | Drama | 20 November 1998 (USA)

1:53 | Trailer

1 VIDEO | 32 IMAGES

An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

Director: [Walter Salles](#)

Writers: [Marcos Bernstein](#), [João Emanuel Carneiro](#) | 1 more credit »

Stars: [Fernanda Montenegro](#), [Vinícius de Oliveira](#), [Marília Pêra](#) | See full cast & crew »

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ SHARE

## The Fog of War: Eleven Lessons from the Life of Robert S. McNamara (2003)

PG-13 | 1h 47min | Documentary, Biography, History | 5 March 2004 (USA)



8.1 /10  
22,171 Rate This

+

The Fog of War: Eleven Lessons from the Life of Robert S. McNamara (2003)

PG-13 | 1h 47min | Documentary, Biography, History | 5 March 2004 (USA)

“IF THERE IS ONE MONTH THAT OUGHT TO BE STUDIED BY MILITARY AND CIVILIAN LEADERS AROUND THE WORLD AT THIS TREACHEROUS TIME IN HISTORY, IT IS APRIL 1968. SEE ESROL MORRIS'S PORTRAIT OF FORMER UNITED STATES DEFENSE SECRETARY ROBERT S. McNAMARA.”

“SMART AND COMPLETELY FASCINATING!”

“STUNNING! SUPERBLY MADE!”

“SPLendid! A Masterpiece!”

THE FOG OF WAR  
An Errol Morris Film

2:07 | Trailer

2 VIDEOS | 11 IMAGES

The story of America as seen through the eyes of the former Secretary of Defense under President [John F. Kennedy](#) and President [Lyndon Baines Johnson](#), [Robert McNamara](#).

Director: [Errol Morris](#)

Stars: [Robert McNamara](#), [John F. Kennedy](#), [Fidel Castro](#) | See full cast & crew »

# How do we learn a wrapper for a relation?

Key intuition:

Capture locally consistent features around an attribute's values to learn a rule that is robust to minor page variations

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ SHARE

**Central Station (1998)**  
Central do Brasil (original title)  
R | 1h 50min | Drama | 20 November 1998 (USA)

★ 8.0 / 10 33,811 | Rate This

**Central do Brasil**  
Directed by Walter Salles  
Written by Fernanda Montenegro, Marcos Bernstein, João Emanuel Carneiro  
Starring Fernanda Montenegro, Vinícius de Oliveira, Marília Pêra

An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

**Director:** Walter Salles

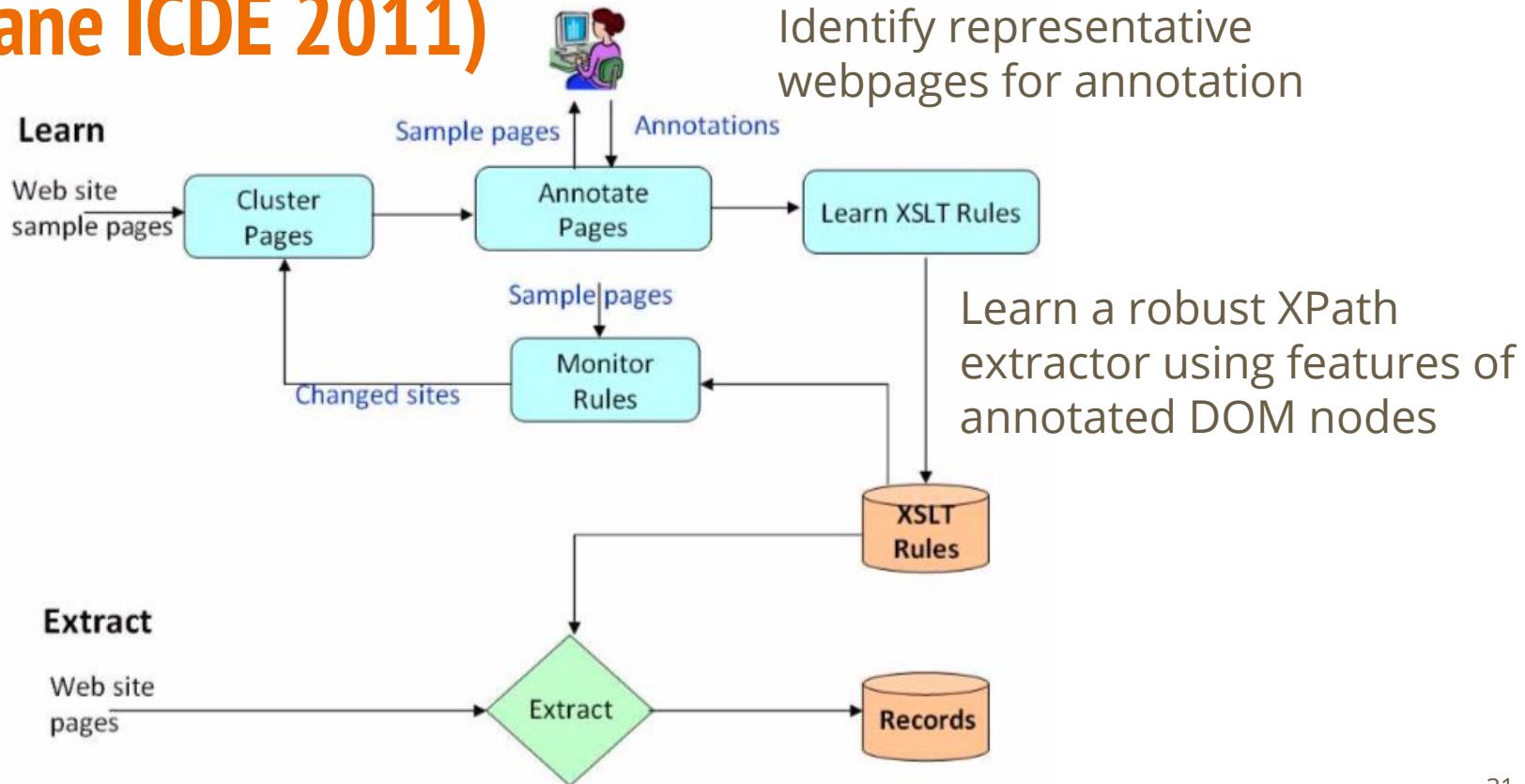
**Writers:** Marcos Bernstein, João Emanuel Carneiro | 1 more credit »

**Stars:** Fernanda Montenegro, Vinícius de Oliveira, Marilia Pêra | See full cast & crew »

1:53 | Trailer | 1 VIDEO | 32 IMAGES

30

# Vertex - A wrapper induction method (Gulhane ICDE 2011)



# Example annotation

<https://www.allmusic.com/album/tring-a-ling-mw0000895190>

```
"annotations": {  
    "hasReleaseFeature": {  
        "//*[@id=\"cmn_wrap\"]/div[1]/div[1]/section[3]/div[4]/div/a": "Post-Bop"  
    },  
    "hasMainPerformer": {  
        "//*[@id=\"cmn_wrap\"]/div[1]/div[2]/header/hgroup/h2/span/a": "Joanne Brackeen"  
    },  
    "hasRecordingDate": {  
        "//*[@id=\"cmn_wrap\"]/div[1]/div[1]/section[3]/div[5]/div": "1977"  
    },  
    "hasTitle": {  
        "//*[@id=\"cmn_wrap\"]/div[1]/div[2]/header/hgroup/h1": "Tring-A-Ling"  
    },  
    "hasGenre": {  
        "//*[@id=\"cmn_wrap\"]/div[1]/div[1]/section[3]/div[3]/div/a": "Jazz"  
    },  
    "hasDuration": {  
        "//*[@id=\"cmn_wrap\"]/div[1]/div[1]/section[3]/div[2]/span": "57:31"  
    },  
    "hasStudioInformation": {  
        "//*[@id=\"cmn_wrap\"]/div[1]/div[1]/section[3]/div[6]/ul/li": "MacDonald Studio"  
    },  
    "hasOriginalReleaseDate": {  
        "//*[@id=\"cmn_wrap\"]/div[1]/div[1]/section[3]/div[1]/span": "March 20, 1977"  
    }  
}
```

Specifies location  
and value for a  
predicate

# Learning a robust XPath

**Features of annotated DOM nodes:**

- HTML tag features (id, class, HTML attributes)
- Siblings and ancestors of annotated nodes
- Path to template strings (e.g., “Director:”)
- Textual features

**Training data:** annotated sample + unannotated sample

**Learning:**

1. Enumerate XPaths for each feature
2. Iteratively combine, evaluate and rank each XPath by its “fitness” based on annotated and unannotated sample
3. Stop when the best, robust XPath is found

# Example learned XPaths as rules

'Price' on [www.amazon.com](http://www.amazon.com)

```
//node() [@class="listprice"]/node()
```

'Forum title' on [www.city-data.com](http://www.city-data.com)

```
//td[@class="navbar"]/*/text()
```

'Address' on [www.hotels.com](http://www.hotels.com)

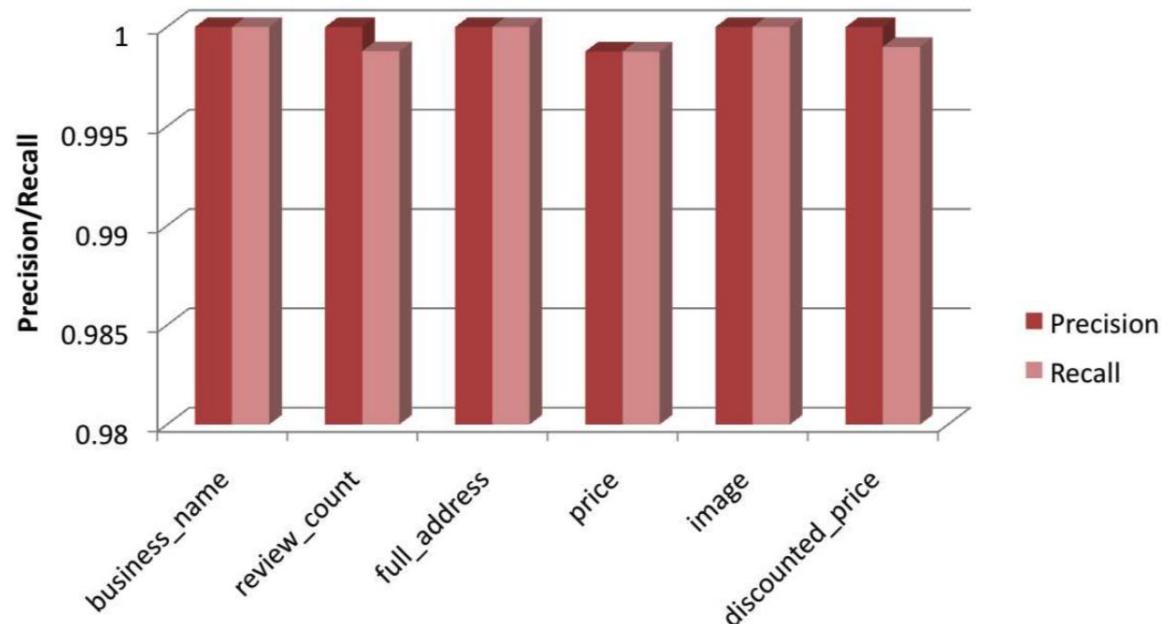
```
//node() [@class="adr"]
```

'Image' on [www.alibaba.com](http://www.alibaba.com)

```
//node() [@class="detailImage" or @class="detailMain hackBorder"]/*/img
```

# Performance (Vertex)

Very accurate extractors: ~100% F1-score



# Summary of Vertex

A semi-supervised, closed IE approach that learns attribute rules using layout context features of manually annotated DOM nodes.

## Pros:

- High performance: very accurate ~100% F-score
- Robust to local diversity
- Expressive rule space to handle diverse layout

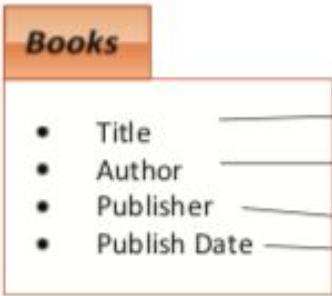
## Cons:

- Requires accurate, manually labeled data limiting its scalability
- Operates on a template-by-template basis

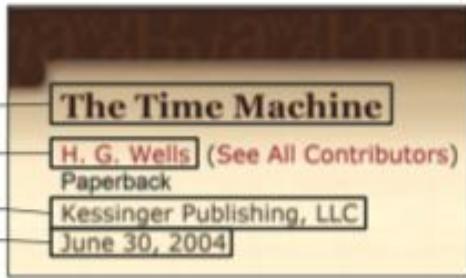
**How can we extract from ALL websites in a domain given ONE or few labeled websites?**

# Extracting from all websites in a domain given a single labeled website -- PL+IP+IA ( Hao, SIGIR 2011)

Verticals and Attributes



One Labeled Seed Site



Many Unseen Sites



Given:

- A set of domain attributes
- A labeled seed website

Task:

Extract from a new unseen website

# Key problem for PL+IP+IA

**Given:**

a DOM tree representation of pages of a new website

**Determine:**

Text values for each attribute in the domain

# Challenges in moving from ONE to ALL websites

- **Variation of attribute values:** multiple values, abbrev. vs. full values
- **Variation of layout:** different page layout structures
  - E.g. optional/missing sections, tables vs key-value pairs
- **Noisy page content:** extraneous content intertwined with target attribute values
  - E.g. other date-type values besides true value for 'publish-date'

# What is shared domain knowledge among websites?

## 1. Attribute-specific semantics

“Birthplace” (Site 1) vs. “Place of birth” (Site 2)

## 2. Inter-attribute layout consistency

Book title and author generally appear together

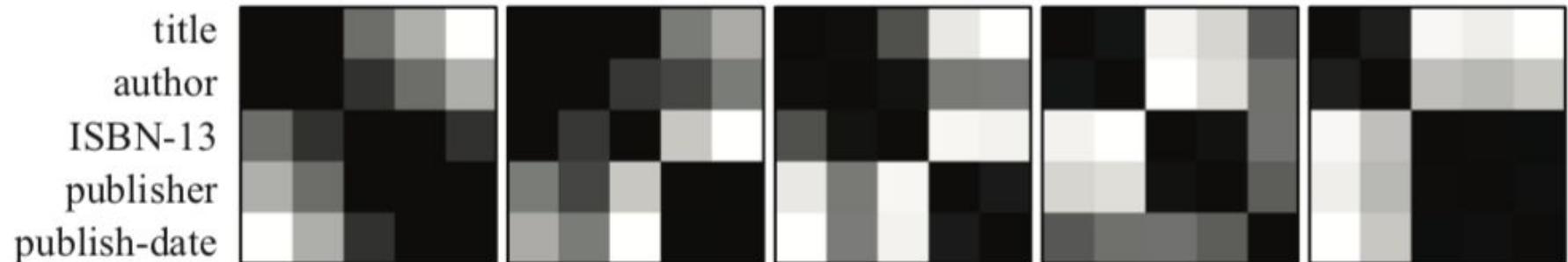
# Attribute-specific semantics

- **Unigrams:** some terms indicate presence of the attribute
  - e.g. 'press' help identify a book 'publisher'
- **Token/Character count:** attribute values typically have 2-4 terms and are often fixed length e.g. ISBN-13
- **Character type:** values often only contain certain characters
  - e.g. 'price' has digits and symbols (\$, Rs.)
- **Redundancy:**
  - Some attributes have a fixed set e.g. 'cuisine'
  - Other attributes have unique values e.g. 'name'
- **Context:** prefix/suffix indicate presence of attribute value
  - e.g. 'Publisher:', 'Pub. date'

# Inter-attribute layout consistency

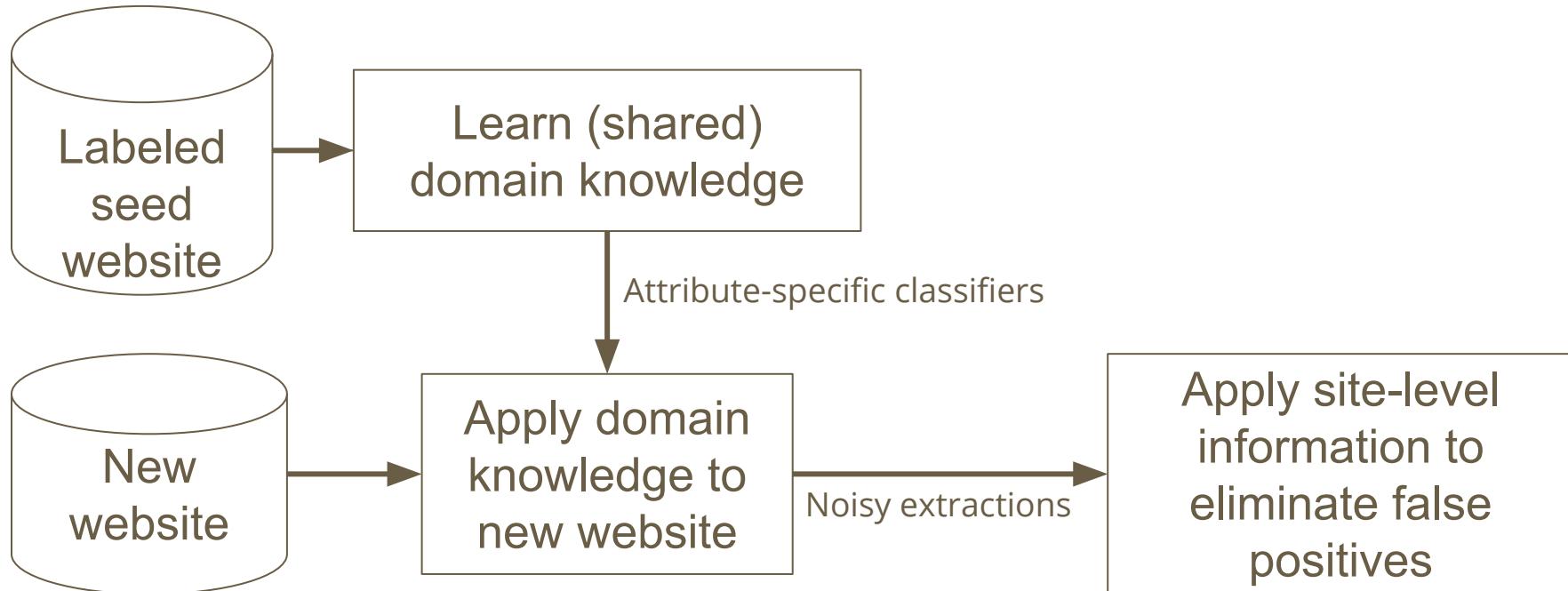
Some attributes are often close to each other on the page

e.g. title and author

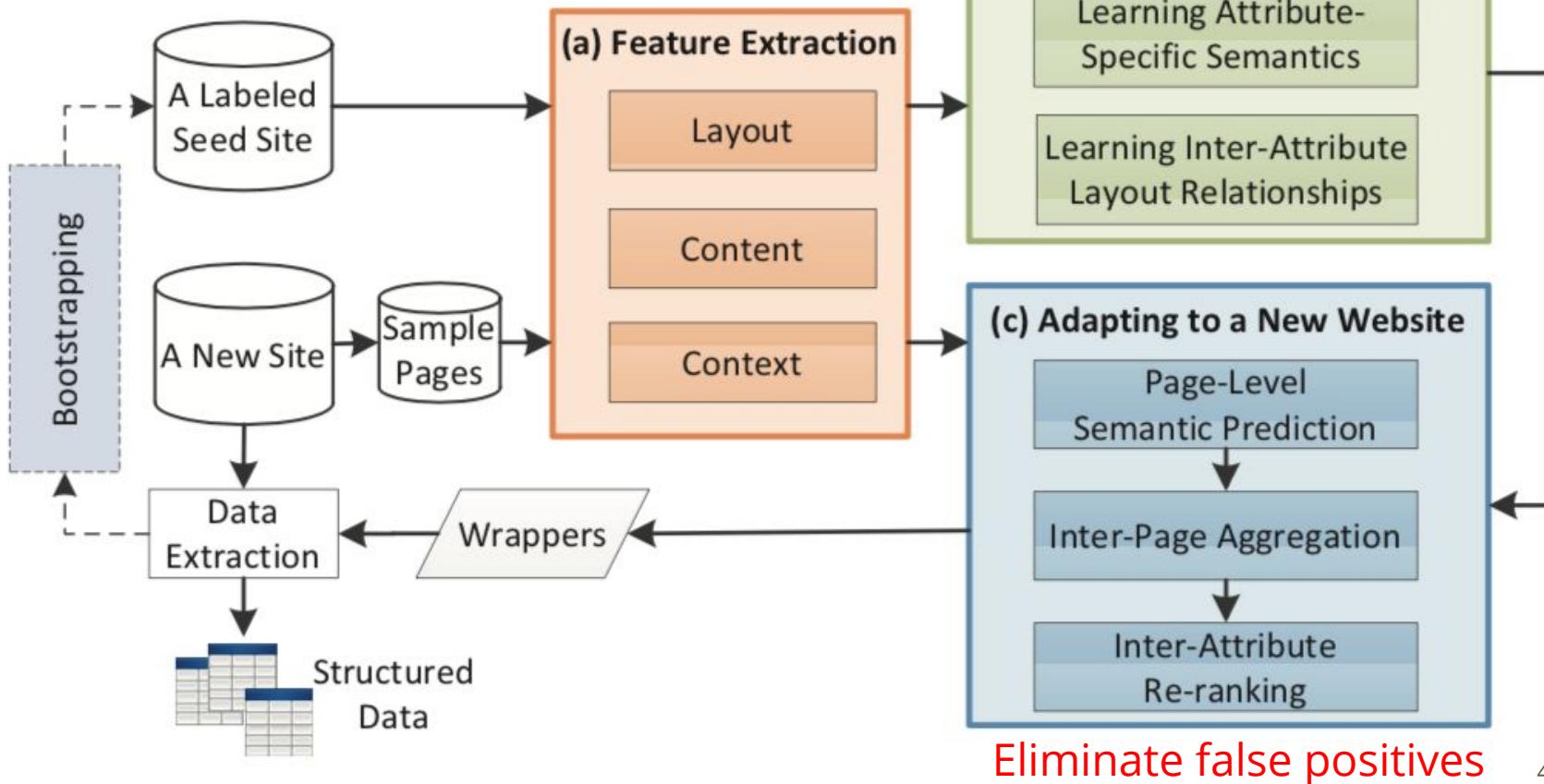


the darker cells indicate attributes are in close vicinity

# High-level idea



# PL+IP+IA



# Performance

Good overall performance

## Limitations:

- Variety of content (e.g. 1.96m, 6 ft 5 in, 6'5" for height)
- No standard attribute definition (e.g. model)
- Disambiguating between true and other relevant content (e.g. recommended movie titles)

Vertical	Attribute	Precision	Recall	F-score
Autos	model	0.46 ± 0.27	0.41 ± 0.26	0.43 ± 0.26
	price	0.80 ± 0.19	0.79 ± 0.19	0.80 ± 0.19
	engine	0.82 ± 0.14	0.82 ± 0.14	0.82 ± 0.14
	fuel-economy	0.81 ± 0.20	0.73 ± 0.18	0.77 ± 0.19
Books	title	0.89 ± 0.13	0.87 ± 0.14	0.88 ± 0.14
	author	0.95 ± 0.04	0.89 ± 0.04	0.92 ± 0.04
	ISBN-13	0.84 ± 0.19	0.84 ± 0.18	0.84 ± 0.18
	publisher	0.81 ± 0.06	0.81 ± 0.06	0.81 ± 0.06
	publish-date	0.88 ± 0.08	0.88 ± 0.08	0.88 ± 0.08
Cameras	model	0.93 ± 0.07	0.88 ± 0.06	0.90 ± 0.07
	price	0.98 ± 0.04	0.90 ± 0.05	0.94 ± 0.05
	manufacturer	0.96 ± 0.06	0.93 ± 0.06	0.94 ± 0.06
Jobs	title	0.99 ± 0.03	0.93 ± 0.04	0.95 ± 0.04
	company	0.84 ± 0.24	0.80 ± 0.22	0.82 ± 0.22
	location	0.87 ± 0.07	0.84 ± 0.07	0.85 ± 0.07
	date	0.79 ± 0.20	0.77 ± 0.19	0.78 ± 0.20
Movies	title	0.71 ± 0.25	0.68 ± 0.25	0.69 ± 0.25
	director	0.75 ± 0.11	0.80 ± 0.12	0.77 ± 0.12
	genre	0.96 ± 0.04	0.91 ± 0.04	0.93 ± 0.04
	rating	0.78 ± 0.23	0.75 ± 0.23	0.76 ± 0.23
NBA Players	name	0.84 ± 0.24	0.82 ± 0.23	0.83 ± 0.23
	team	0.82 ± 0.09	0.82 ± 0.09	0.82 ± 0.09
	height	0.76 ± 0.19	0.67 ± 0.17	0.71 ± 0.18
	weight	0.91 ± 0.10	0.91 ± 0.10	0.91 ± 0.10
Restaurants	name	0.95 ± 0.08	0.89 ± 0.07	0.92 ± 0.07
	address	0.97 ± 0.02	0.96 ± 0.02	0.96 ± 0.02
	phone	1.00 ± 0.00	0.98 ± 0.01	0.99 ± 0.00
	cuisine	0.98 ± 0.07	0.94 ± 0.06	0.96 ± 0.06
Universities	name	0.97 ± 0.05	0.95 ± 0.06	0.96 ± 0.06
	phone	0.79 ± 0.12	0.78 ± 0.12	0.79 ± 0.12
	website	0.96 ± 0.09	0.83 ± 0.08	0.89 ± 0.08
	type	0.70 ± 0.29	0.68 ± 0.27	0.69 ± 0.28

# Performance

More labeled seed websites lead to improved performance

Average F-scores

#Seeds	1	2	3	4	5
<b>Our Solution</b>	0.843	0.860	0.868	0.884	0.886
<b>Our Solution (Bootstrap)</b>	0.843	0.856	0.861	0.859	0.865
<b>SSM</b>	0.630	0.645	0.692	0.719	0.741

# Summary of PL+IP+IA

A semi-supervised, closed IE approach that is able to extract from all websites in a domain given a single or few seed websites

## Pros:

- First approach to use domain knowledge as "labeled data"
- Moderately high performance 84% F-score

## Cons:

- Weak generalizable knowledge (high diversity in content format, lack of available context)
- Requires manual labels for at least one website/template

**How can we avoid manual annotations to scale  
to the large number of websites on the Web?**

# Can we automatically annotate? -- Distant supervision

Idea: Use a seed KB of a domain as source for distant supervision

**Distant supervision assumption:** A sentence that contains a pair of entities that participate in a known KB relation is likely to express that relation in some way.

film.release\_year

Central  
Station

1998



**Caveat:** The annotation may be noisy.



Rita Moreno

Actress | Soundtrack



[View Resume](#) | [Official Site](#) | [Photos](#) »

Rita Moreno has had a thriving acting career for the better part of six decades. One of the very few performers (and the very first) to win an Oscar, an Emmy, a Tony and a Grammy, she was born Rosita Dolores Alverio in Humacao, Puerto Rico, on December 11, 1931, to seamstress Rosa María (Marcano) and farmer Francisco José "Paco" Alverio. She and ... [See full bio](#)

Born: **December 11, 1931** in **Humacao, Puerto Rico**

[More at IMDbPro](#) »

Official Sites: [Official Site](#) | [Twitter](#)

Alternate Names: Rita Moreno Gordon | Rosita Moreno

Height: **5' 2½" (1.59 m)**

### Did You Know?

**Personal Quote:** There was nobody that I could look up and say "That's somebody like me". Which is probably why I'm now known in my community as 'La Pionera', or the Pioneer. I really don't think of myself as a role model, but it turns out that I am to a lot of the Hispanic community. Not just in show business, but in life. But that's what happens when you're first, right? [See more](#) »

**Trivia:** Mother of Fernanda [Don. See more](#) »

**Star Sign:** **Sagittarius**

Automatic annotations  
of KB predicates [Edit](#)

# Ceres (Lockard, VLDB 2018)

Input:

- Seed KB

Output:

- Triples from all pages

("R. Moreno", rdf:type, Person)

("R. Moreno", birthday, "Dec 11, 1931")

("R. Moreno", birthplace, "Humacao, Puerto Rico")

("R. Moreno", height, "5' 2½" (1.59 m))

("R. Moreno", star\_sign, "Sagittarius")

... likewise, from all other pages

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾

**Do the Right Thing** (1989) ★ 7.9 70,044 Rate This

R | 2h Comedy, Drama | 21 July 1989 (USA)

**Do the Right Thing**

Watch Now From \$2.99 (SD) on Prime Video

ON DISC

On the hottest day of the year on a street in the Bedford-Stuyvesant section of Brooklyn, everyone's hate and bigotry smolders and builds until it explodes into violence.

Director: **Spike Lee**

Writer: Spike Lee

Stars: Danny Aiello, Ossie Davis, Ruby Dee | See full cast & crew »

More Like This

**Crooklyn** (1994) PG-13 Comedy | Drama

★★★★★ 6.9/10

Spike Lee's vibrant semi-autobiographical portrait of a school teacher, her stubborn jazz musician husband and their five kids living in Brooklyn in 1973.

Add to Watchlist

Next »

Director: **Spike Lee**

Stars: Alfre Woodard, Delroy Lindo, ...

◀ Prev 6 Next 6 ▶



# Challenges

- Entity linking problem
- Distant supervision applied naively requires  $n^2$  entity mention pair comparisons
  - computationally infeasible
  - can lead to spurious annotations
- Disambiguate relations involving same entity pair
- Distinguish true relation mentions from spurious mentions

# Topic entity annotation

Rita Moreno (Actress Soundtrack) Top 5000

[View Resume](#) | [Official Photos](#) »

Rita Moreno has had a thriving acting career for the better part of six decades. One of the very few performers (and the very first) to win an Oscar, an Emmy, a Tony and a Grammy, she was born Rosita Dolores Alverío in Humacao, Puerto Rico, on December 11, 1931, to seamstress Rosa María (Marcano) and farmer Francisco José "Paco" Alverío. She and ... [See full bio](#) »

**Born:** December 11, 1931 in Humacao, Puerto Rico

[More at IMDbPro](#) »

**Contact Info:** View agent, publicist, legal on [IMDbPro](#)

## Filmography

Jump to: [Actress](#) | [Soundtrack](#) | [Self](#) | [Archive footage](#)

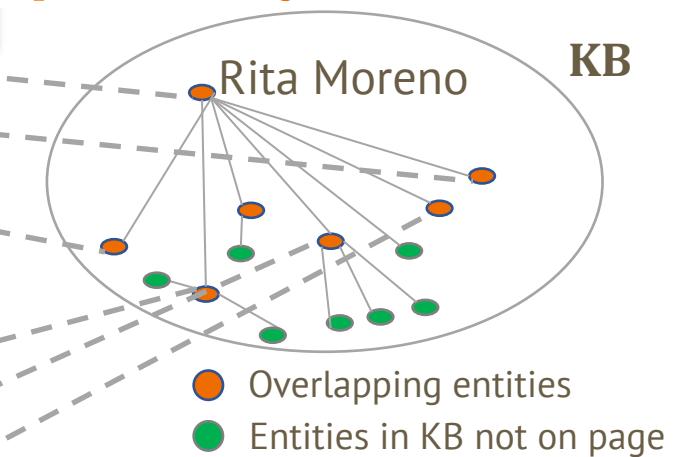
**Actress (155 credits)**

Nina's World (TV Series)

Abuelita

- The Best Ending Ever! (2018) ... Abuelita (voice)
- *Cardel Wining Shirt* (2018) ... Abuelita (voice)
- Nina Live (2018) ... Abuelita (voice)
- *Nina in Charge* (2018) ... Abuelita (voice)

2015-2018



- Local consistency:** The topic entity should be associated with many entities on the page.
- Global consistency:** The topic entity's name should be in a consistent location on each page. 53

# Relation annotation

Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

Rita Moreno (I)

Actress | Soundtrack

Top 5000

[View Resume](#) | [Official Photos](#) »

Rita Moreno has had a thriving acting career for the better part of six decades. One of the very few performers (and the very first) to win an Oscar, an Emmy, a Tony and a Grammy, she was born Rosita Dolores Alverio in Humacao, Puerto Rico, on December 11, 1931, to seamstress Rosa María (Marcano) and farmer Francisco José "Paco" Alverio. She and ... [See full bio](#) »

Born: **December 11, 1931** in **Humacao, Puerto Rico**

More at [IMDbPro](#) »

Official Sites: [Official Site](#) | [Twitter](#)

Alternate Names: Rita Moreno Gordon | Rosita Moreno

Height: **5' 2 1/2" (1.59 m)**

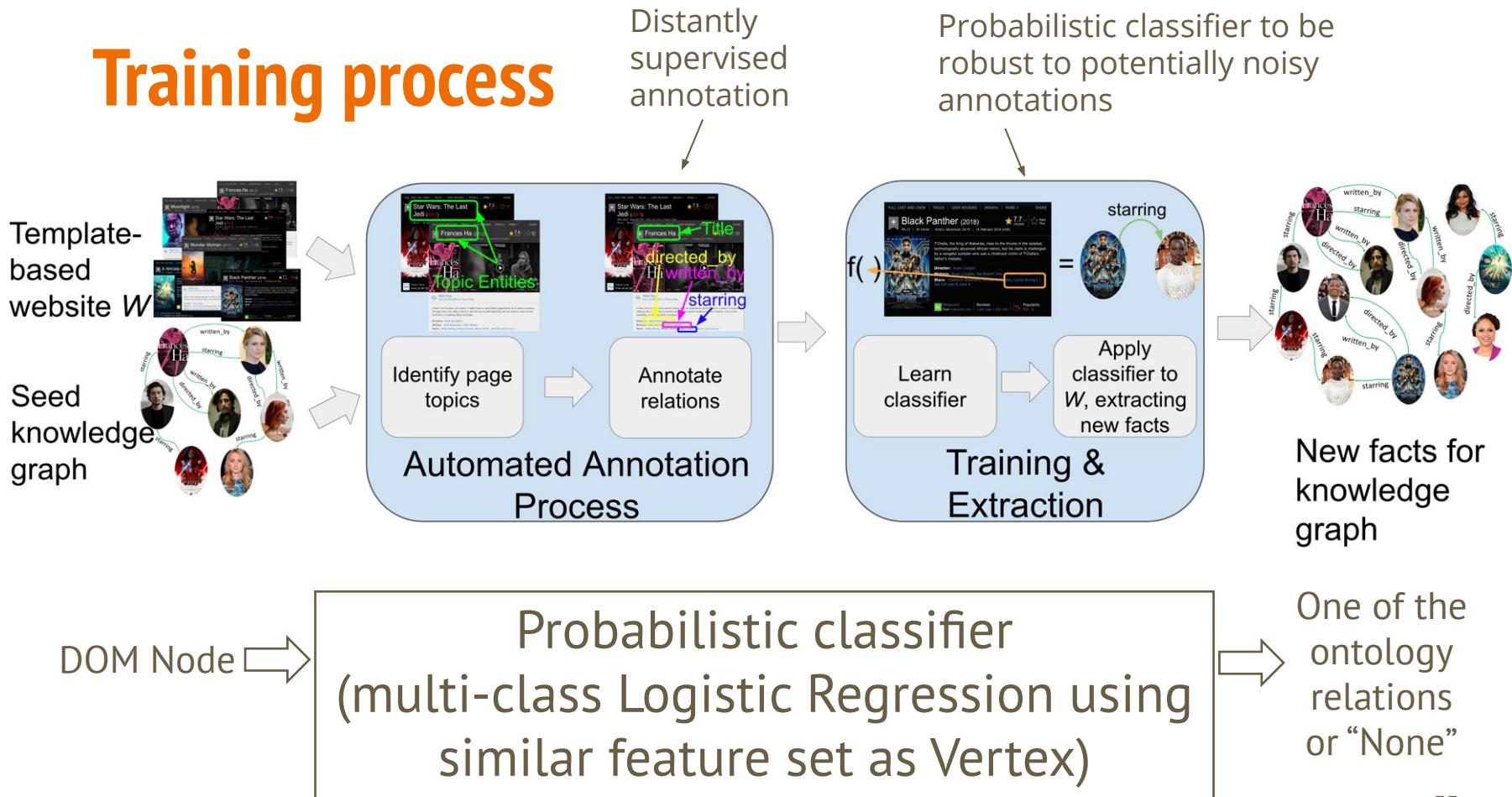
Did You Know?

Personal Quote: There was nobody that I could look up and say "That's somebody like me". Which is probably why I'm now known in my community as 'La Pionera', or the Pioneer. I really don't think of myself as a role model. But it turns out that I am to a lot of the Hispanic community. Not just in show business, but in life. But that's what happens when you're first, right? [See more](#) »

Automatic annotations of KB predicates

1. Annotate entity mention pairs using known factual relations from the KB.
2. **Local consistency:** KB objects of the same predicate should be in the same section of page.
3. **Global consistency:** Predicates should be in a *similar* location on all pages. Cluster all potential mentions of a relation across site and choose the most common location.

# Training process



# Performance

PL+IP+IA

Another distant supervision method using instances from Linked Open Data (LOD) for supervision

Ceres delivers highest F-measure on two domains

System	Manual Labels	Movie	NBA Player	University	Book
Hao <i>et al.</i> [19]	yes	0.79	0.82	0.83	0.86
XTPath [7]	yes	0.94	<b>0.98</b>	0.98	<b>0.97</b>
BigGrams [26]	yes	0.74	0.90	0.79	0.78
LODIE-Ideal [15]	no	0.86	0.9	0.96	0.85
LODIE-LOD [15]	no	0.76	0.87 <sup>a</sup>	0.91 <sup>a</sup>	0.78
RR+WADaR [29]	no	0.73	0.80	0.79	0.70
RR+WADaR 2 [30]	no	0.75	0.91	0.79	0.71
Bronzi <i>et al.</i> [4]	no	0.93	0.89	0.97	0.91
Vertex++	yes	0.90	0.97	<b>1.00</b>	0.94
CERES-Baseline	no	NA <sup>b</sup>	0.78	0.72	0.27
CERES-Topic	no	<b>0.99<sup>a</sup></b>	0.97	0.96	0.72
CERES-Full	no	<b>0.99<sup>a</sup></b>	<b>0.98</b>	0.94	0.76

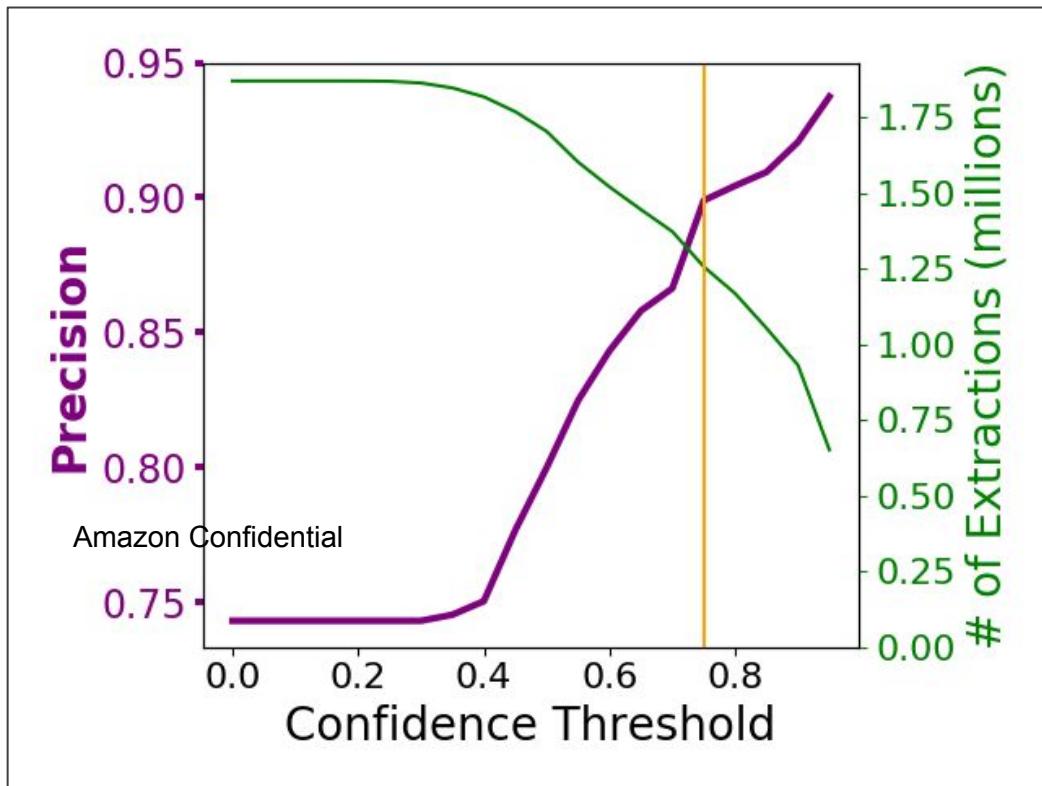
Domain having low overlap with seed data performs suboptimally

# Ceres -- distant supervision extraction

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%

# Performance on long-tail movie websites



Unlike rules, you can tune your classifier to emphasize precision or recall

1.25M triples extracted at 90% precision using 0.75 as confidence threshold

# Summary of Ceres

A fully automatic, closed IE approach that extracts data by learning a robust relation classifier using layout context features of distantly annotated DOM nodes (labels).

## Pros:

- Automatic labeling process through distant supervision by a seed knowledge base
- Fairly high performance (~90% precision)

## Cons:

- Assumes availability of a domain-specific knowledge base
- Low recall of attributes due to inherently being a closed IE method

# How do we extract MORE relations on the Web?

# OpenIE for harvesting new relations

**Closed IE:** We have fully automatic extraction methods for a few relations

**Open IE:** How do we expand the set of relations to include **new relations** on the Web?

## Storyline

Edit

Jedi Master-in-hiding Luke Skywalker unwillingly attempts to guide young hopeful Rey in the ways of the force, while Leia, former princess turned general, attempts to lead what is left of the Resistance away from the ruthless tyrannical grip of the First Order.

Written by [Danny Moniz](#)

[Plot Summary](#) | [Plot Synopsis](#)

**Plot Keywords:** [wisecrack humor](#) | [one liner](#) | [sabotage](#) | [asiatic](#) | [chubby](#) | [See All \(570\)](#) »

**Taglines:** Always in Motion is the Future [See more](#) »

**Genres:** [Action](#) | [Adventure](#) | [Fantasy](#) | [Sci-Fi](#)

**Motion Picture Rating (MPAA)**

Rated PG-13 for sequences of sci-fi action and violence. | [See all certifications](#) »

**Parents Guide:** [View content advisory](#) »

## Details

Edit

**Official Sites:** [Official Facebook](#) | [Official Site](#) | [See more](#) »

**Country:** USA

**Language:** English

**Release Date:** 15 December 2017 (USA) [See more](#) »

**Also Known As:** Star Wars: Episode VIII - The Last Jedi [See more](#) »

**Filming Locations:** Pinewood Studios, Iver Heath, Buckinghamshire, England, UK [See more](#) »

# WEIR -- The first open IE method (Bronzi, VLDB'13)

- Data-rich websites overlap at the schema and instance level
- Why not leverage the **data redundancy** to learn extractors?

International Business Machines Corp. IBM.N

LATEST TRADE  
**136.77** USD  
As of 3:30 PM PST Jan 30 on the New York Stock Exchange · Minimum 15 minute delay

Profile News Key Developments Charts People Financials Key Metrics Events All Listings

Pricing

Previous Close

Open

Volume

3M AVG Volume

Today's High

Today's Low

52 Week High

52 Week Low

Shares Out (MIL)

Market Cap (MIL)

121,943.40

Reuters

CHANGE  
**-0.92 (-0.67%)**  
VOLUME  
1,074,881

TODAY'S RANGE  
134.98 - 136.97

LATEST TRADE  
**136.77** USD  
As of 3:30 PM PST Jan 30 on the New York Stock Exchange · Minimum 15 minute delay

Profile News Key Developments Charts People Financials Key Metrics Events All Listings

Pricing

Previous Close

Open

Volume

3M AVG Volume

Today's High

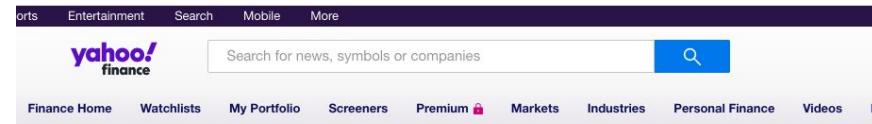
Today's Low

52 Week High

52 Week Low

Shares Out (MIL)

Market Cap (MIL)



International Business Machines Corporation (IBM)

NYSE - NYSE Delayed Price. Currency in USD

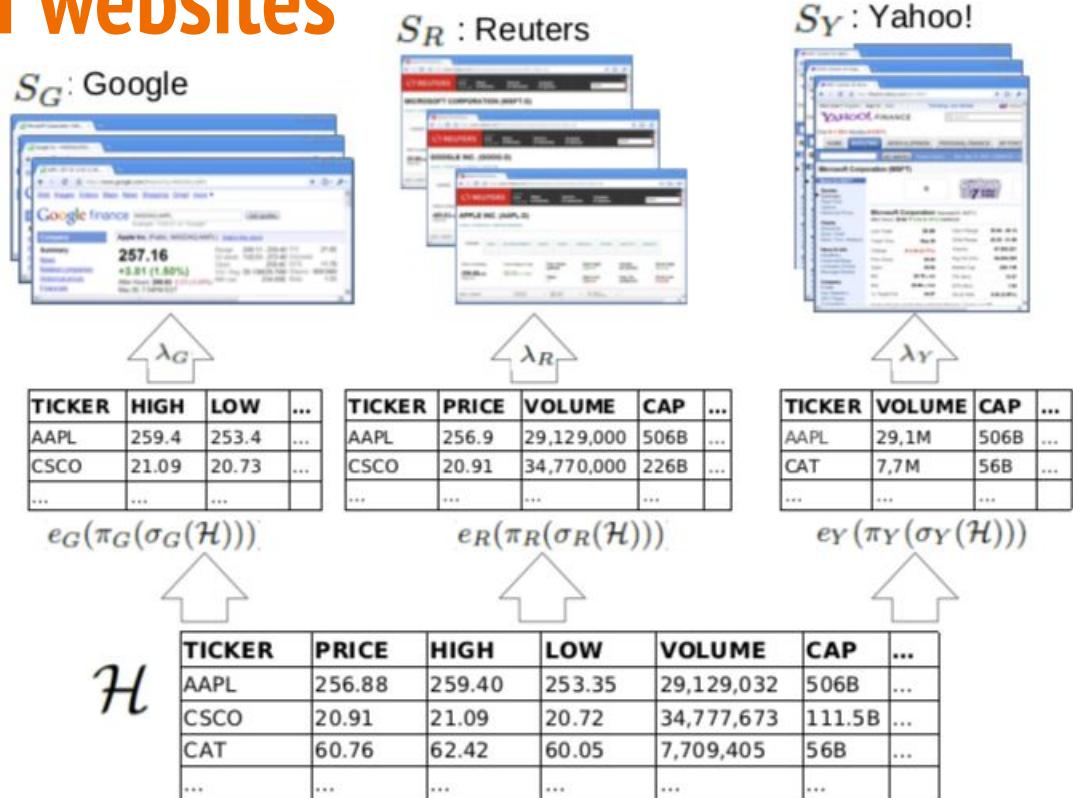
At close: 4:00PM EST	<b>136.77</b>	-0.92 (-0.67%)	143.23 +6.46 (4.72%)	After hours: 7:56PM EST
Summary	Company Outlook	Chart	Conversations	Statistics
Previous Close	137.69	Market Cap	122.765B	1D 5D 1M 6M YTD 1Y 5Y Max Full screen
Open	136.76	Beta (5Y Monthly)	1.34	
Bid	143.10 x 800	PE Ratio (TTM)	12.94	
Ask	143.25 x 1100	EPS (TTM)	10.57	
Day's Range	134.97 - 136.97	Earnings Date	Apr 14, 2020 - Apr 20, 2020	
52 Week Range	126.85 - 152.95	Forward Dividend & Yield	6.48 (4.71%)	
Volume	4,417,823	Fx-Dividend Date	Feb 07, 2020	

# Generative model of websites

Overlapping websites

Partial views over  $\mathcal{H}$

Abstract relation: a set of abstract attributes



**Extraction & integration = inverting the generation process (i.e. discover the abstract relation)**

# Key intuition: Leverage data redundancy

IF we had extractors for multiple websites in the same domain,

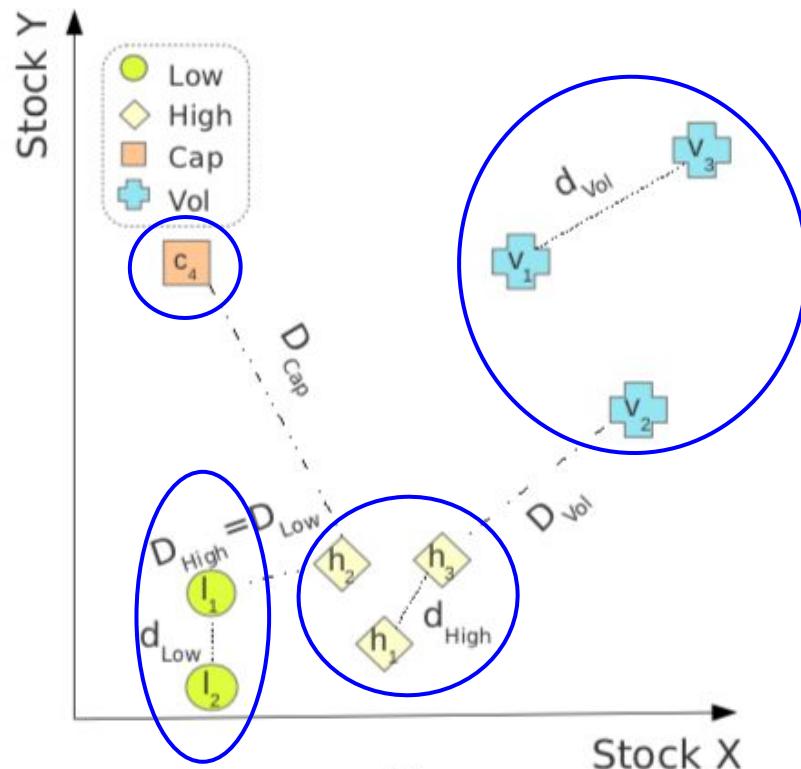
- Data (instances) should be similar (e.g. 5ft. 9in. Site 1 vs. 5' 9" Site 2)
- Semantics of one attribute are different than that of another

a *correct* extractor will likely extract data that match with those extracted from at least one other *correct* extractor from a different website

⇒ Find extractors that maximize overlap semantically similar data from different sites

# Leverage key properties of semi-structured websites

1. **Local consistency:** A website does not publish different values for the same attribute
2. **Separable semantics:** Attributes with similar semantics are *closer* than attributes with different semantics



# Recipe

1. Eliminate obvious non-attribute values
2. Enumerate data-type aware extractors as XPath rules for all candidate attribute values
3. Filter out useless and “weak” rules
4. Cluster extractors that match data having similar semantics while obeying the “separable semantics” constraint

Template values  
Candidate attribute values

The screenshot shows the IMDb movie page for 'Titanic' (1997). At the top, there's a navigation bar with links for 'FULL CAST AND CREW', 'TRIVIA', 'USER REVIEWS', 'IMDbPro', and 'MORE'. Below the title 'Titanic (1997)' are several green-highlighted boxes containing movie details: 'PG-13', '3h 14min', 'Drama, Romance', and '19 December 1997 (USA)'. To the right is a yellow star rating of '7.8/10' with '990,618' reviews, and a 'Rate This' button. The main image is a composite of two scenes: a close-up of Leonardo DiCaprio and Kate Winslet, and a wider shot of Leonardo DiCaprio shouting. Below the image is a movie poster for 'TITANIC' with a play button icon. A video player interface shows '2:11' and 'Trailer'. At the bottom, there's a green-highlighted plot summary: 'A seventeen-year-old aristocrat falls in love with a kind but poor artist aboard the luxurious, ill-fated R.M.S. Titanic.' Below the plot summary is a cast list where 'Director:' and 'Writer:' are in red boxes, while 'Stars:' and the names 'Leonardo DiCaprio', 'Kate Winslet', and 'Billy Zane' are in green boxes, with a link 'See full cast & crew'.

# WEIR kills two birds with one stone!

Tackles two problems simultaneously:

1. **Data extraction problem:** generate attribute extraction rules for a given set of websites
2. **Data integration problem:** unify the diversity of relation terms used on different websites by integrating them into a unified schema

# Performance

Instance-level overlap between sources  
 $d = 1 \Rightarrow$  all sources have shared instances  
 $d > 1 \Rightarrow$  many source pairs do not share instances

#Pages #instances

<i>Domain</i>	<i>#p</i>	<i>#o</i>	<i>d</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>Time</i>
soccer players	5,850	4,178	3	0.90	0.93	0.91	80 s
stock quotes	4,656	573	1	0.90	0.81	0.85	67 s
video games	12,339	5,364	2	0.93	0.90	0.91	204 s
books	1,318	196	1	0.94	0.78	0.84	15 s

Fairly high precision (~90%)

# Summary of WEIR

The first open IE, unsupervised approach that exploits data redundancy to extract and integrate information from multiple websites.

## Pros:

- Fairly high performance (precision 90%+)
- Solves data extraction and schema alignment problem simultaneously

## Cons:

- Requires availability of multiple websites within a domain for data redundancy (each instance on at least 5 websites)
- Limits the recall of all relations on the websites due to needed data redundancy

**How can we push the recall of relations?**

# OpenCeres (Lockard, NAACL 2019)

We have **object** annotations via  
distant  
supervision

The screenshot shows a movie details page with the following annotations:

- Genres:** Comedy, Drama, Romance (Predicate, Object)
- Motion Picture Rating (MPAA):** Rated PG for some language | See all certifications »
- Parents Guide:** View content advisory
- Details:** Country: USA (Predicate, Object)
- Language:** English
- Release Date:** 25 June 1993 (USA) See more »
- Also Known As:** Sintonía de amor See more »
- Filming Locations:** 1517 Pike Place, Seattle, Washington, USA See more »
- Box Office:** Budget: \$21,000,000 (estimated), Gross USA: \$126,533,006

Annotations are highlighted with colored boxes:

- Genres: Comedy, Drama, Romance
- Parents Guide: View content advisory
- Country: USA
- Filming Locations: 1517 Pike Place, Seattle, Washington, USA
- Budget: \$21,000,000 (estimated)
- Gross USA: \$126,533,006

We want to  
extract these  
new relations

# Challenges in Open IE from semi-structured website

The screenshot shows a semi-structured website for a movie. At the top, there's a navigation bar with 'Genres' (highlighted in purple), 'Comedy', 'Drama' (highlighted in yellow), and 'Romance'. Below that, it says 'Motion Picture Rating (MPAA)' and 'Rated PG for some language | See all certifications ». Under 'Parents Guide', there's a button 'View content advisory' which is highlighted with a red box.

---

**Details**

Country: USA (highlighted in yellow)  
Language: English  
Release Date: 25 June 1993 (USA) [See more »](#)  
Also Known As: Sintonía de amor [See more »](#)  
Filming Locations: 1517 Pike Place, Seattle, Washington, US (highlighted with a red box)

---

**Box Office**

Budget: \$21,000,000 (estimated)  
Gross USA: \$126,533,006

Ceres distant supervision enables us to match **object mentions** of a *known* relation.

How can we similarly annotate **object mentions** of an *unseen relation*? -- training data for relations not in seed KB

**Idea:** Leverage visual similarity between (relation, object) pairs

# How to identify relation string for matching objects?

**Intuition:** Relation strings are generally more common across a website than their related objects, e.g. “Genres:” vs. “Drama”

## Two main steps:

1. Enumerate candidate relation strings
2. Select closest similar string: string that is lexically or semantically similar to a dictionary of pre-defined terms known for the relation

Genres: [Comedy](#) | [Drama](#) | [Romance](#)

### Motion Picture Rating ([MPAA](#))

Rated PG for some language | [See all certifications »](#)

Parents Guide: [View content advisory »](#)

## Details

Country: [USA](#)

Language: [English](#)

Release Date: 25 June 1993 (USA) [See more »](#)

Also Known As: Sintonía de amor [See more »](#)

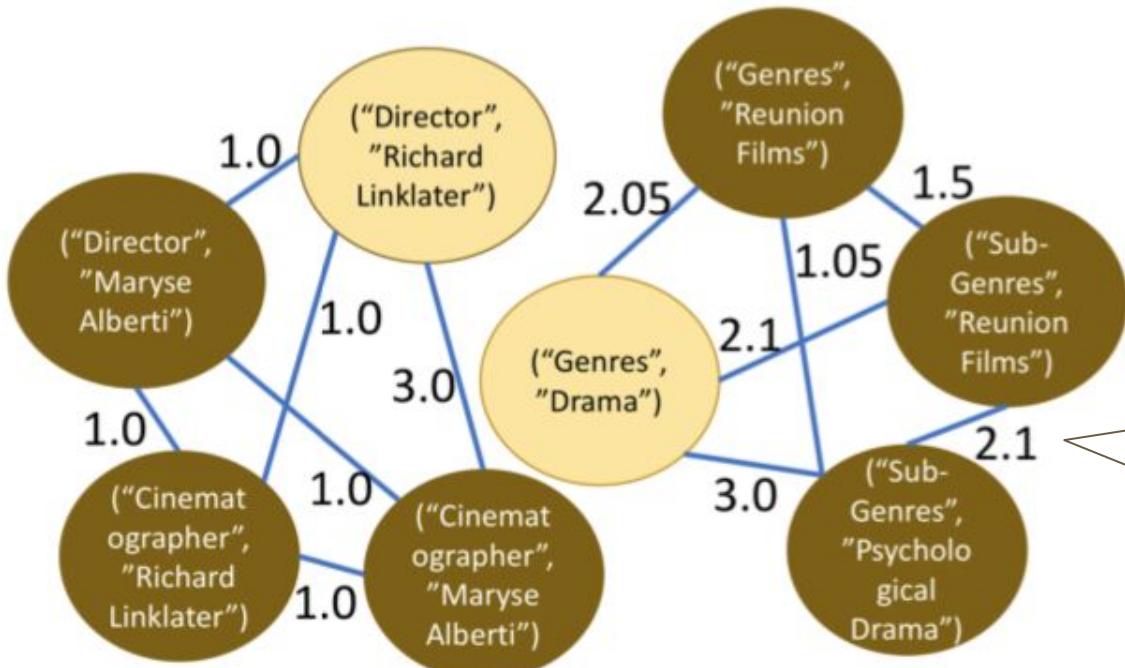
Filming Locations: 1517 Pike Place, Seattle, Washington, U

## Box Office

Budget: \$21,000,000 (estimated)

Gross USA: \$126,533,006

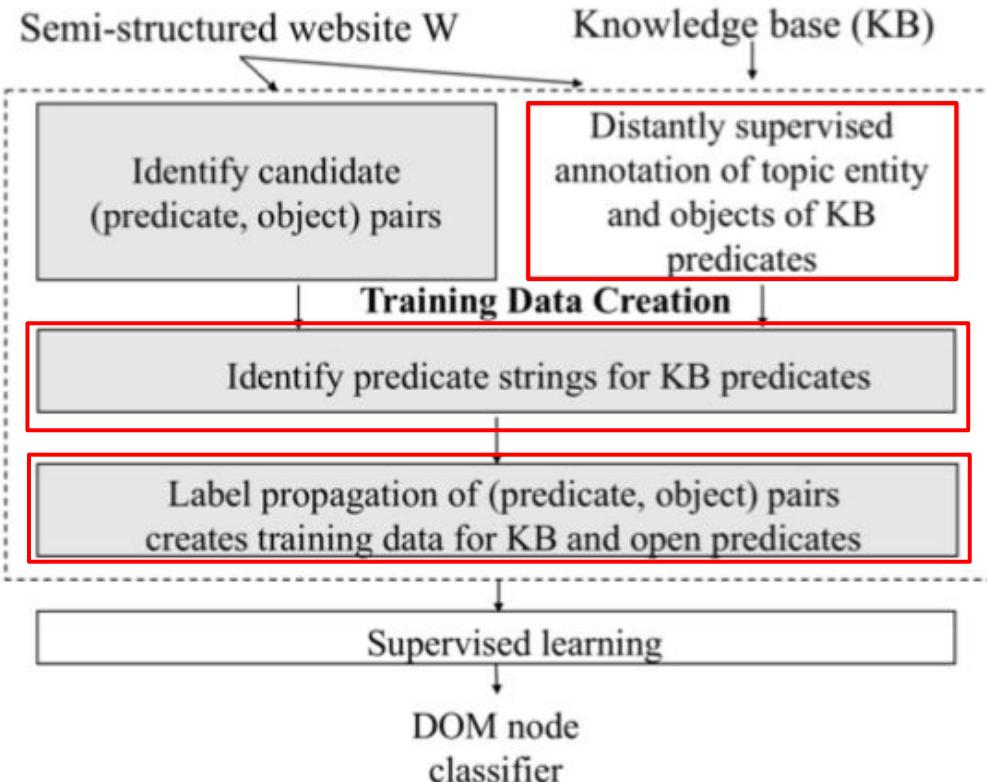
# How do we identify *new* predicate strings? -- Graph-based label propagation



**Seed pairs, New pairs** that are visually similar form a graph

Weights capture how visually similar new (relation, object) pairs are to seed pairs.

# Learning OpenCeres model



Genres: Comedy | Drama | Romance

Motion Picture Rating (MPAA)  
Rated PG for some language | See all certifications »

Parents Guide: View content advisory »

Details

Country: USA

Language: English

Release Date: 25 June 1993 (USA) See more »

Also Known As: Sintonía de amor See more »

Filming Locations: 1517 Pike Place, Seattle, Washington, US

Box Office

Budget: \$21,000,000 (estimated)

Gross USA: \$126,533,006

This screenshot shows a movie details page. At the top, genres are listed: Comedy (highlighted in purple), Drama (highlighted in yellow), and Romance. Below that, motion picture ratings and parents guides are mentioned. The main section is titled "Details". It lists the country (USA), language (English), release date (25 June 1993), and other known names (Sintonía de amor). Filming locations are also listed. The bottom section is titled "Box Office" and provides budget and gross earnings information.

# Performance

Average improvement of 36% precision, 88% recall over baseline

System	Movie		NBA		University	
	P	R	P	R	P	R
WEIR (Bronzi et al., 2013)	0.23	0.17	0.08	0.17	0.13	0.18
Colon Baseline	0.63	0.21	0.51	0.33	0.46	<b>0.31</b>
OpenCeres	<b>0.77</b>	<b>0.68</b>	<b>0.74</b>	<b>0.48</b>	<b>0.65</b>	0.29
OpenCeres-Gold	0.99	0.74	0.98	0.80	0.99	0.60

OpenCeres  
outperforms WEIR  
and a naive baseline

Ceres with manually labeled data for all relations

# Performance

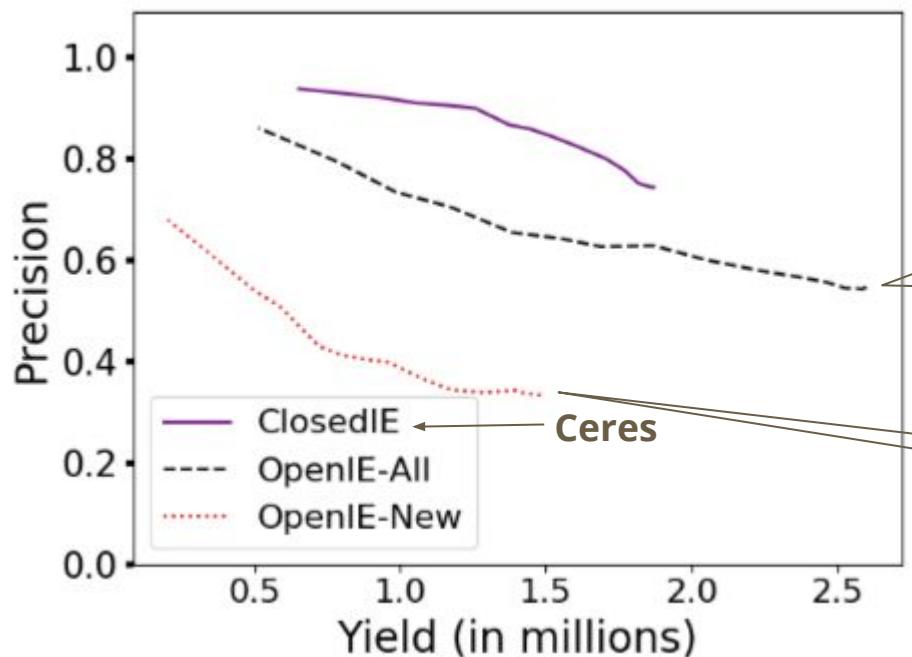
- **Triple-level performance:** 68% F1(lenient), 61% F1 (strict)
- **Predicate-level performance:** avg. 74% precision, 39% recall
- **New relations:** Avg. of 10.5 new relations for every relation in the seed ontology using label propagation

	<b>Movie</b>	<b>NBA Player</b>	<b>University</b>
Triple-level F1	0.72 (0.65)	0.58 (0.58)	0.41 (0.36)
Pred-level Prec	0.55 (0.52)	0.86 (0.86)	0.81 (0.76)
Pred-level Rec	0.35 (0.32)	0.46 (0.46)	0.37 (0.35)
Pred-level F1	0.43 (0.40)	0.60 (0.60)	0.51 (0.48)
New:Existing-pred ratio	4.4 : 1	4.3 : 1	23.0 : 1

Numbers in parentheses indicate strict scoring (vs. lenient otherwise)

OpenCeres boosts recall of relations

# OpenCeres on a large Common Crawl dataset



Conf. thresh	Prec.	#Triples	#Triples w. new relations
0.5	58%	2.5M	1.17 (51%)
0.8	70%	1.17M	0.58 (50%)

Open IE added significant amount of knowledge

Still need improvement on new relations

# Examples of OpenIE relations

## Movie

**Seed:** Director, Writer, Producer, Actor, Release Date, Genre, Alternate Title

**New:** Country, Filmed In, Language, MPAA Rating, Set In, Reviewed by, Studio, Metascore, Box Office, Distributor, Tagline, Budget, Sound Mix

## NBA Player

**Seed:** Height, Weight, Team

**New:** Birth Date, Birth Place, Salary, Age, Experience, Position, College

## University

**Seed:** Phone Number, Web address, Type (public/private)

**New:** Calendar System, Enrollment, Highest Degree, Local Area, Student Services, President

# Summary of OpenCeres

A fully automatic, open IE extraction approach that leverages visual similarity between seed and new (relation, object) pairs to discover new relationships.

## Pros:

- Automatic labeling process for new relations using label prop.
- Improved recall of predicates (7x predicates than baselines)

## Cons:

- Low to moderate precision
- Operates only at single template level for a given domain.

# State of the art for semi-structured data extraction

Method	#Sites	Learning paradigm	Supervision	Manual supervision	Features	Model type
RoadRunner 2001	Single	Neither closed nor open IE	Unsupervised	N	Layout context	Union-free regex
Vertex 2011	Single	Closed IE	Semi-supervised	Y	Layout context	XPath rule
PL+IP+IA 2011	Multiple	Closed IE	Semi-supervised	Y	Textual content + context	Text classifier + ranking
Ceres 2018	Single	Closed IE	Distantly supervised	N	Layout context	Relation classifier
WEIR 2013	Multiple	Open IE	Unsupervised	N	Layout context + text redundancy	XPath rules
OpenCeres 2019	Single	Open IE	Distant sup. + Label prop.	N	Text-based visual + layout context	(rel, obj) pair classifier

# Recipe for semi-structured website extraction

- **Problem definition:** Extract structured attribute data from homogenous set of webpages belonging to a template.
- **Short answers:**
  - Wrapper induction has high precision and recall
  - Distant supervision is critical for creating training data
  - Graph-based label propagation is effective at extracting new relations

# References

Kushmerick, Nicholas, Daniel S. Weld and Robert B. Doorenbos. "Wrapper Induction for Information Extraction." IJCAI (1997).

Gulhane, Pankaj, Amit Madaan, Rupesh R. Mehta, Jeyashankher Ramamirtham, Rajeev Rastogi, Sandeepkumar Satpal, Srinivasan H. Sengamedu, Ashwin Tengli and Charu Tiwari. "Web-scale information extraction with vertex." 2011 IEEE 27th International Conference on Data Engineering (2011): 1209-1220.

Hao, Qiang, Rui Cai, Yanwei Pang and Lei Zhang. "From one tree to a forest: a unified solution for structured web data extraction." SIGIR '11 (2011).

Lockard, Colin, Xin Luna Dong, Arash Einolghozati and Prashant Shiralkar. "CERES: Distantly Supervised Relation Extraction from the Semi-Structured Web." ArXiv abs/1804.04635 (2018): n. pag.

# References

Bronzi, Mirko, Valter Crescenzi, Paolo Merialdo and Paolo Papotti. "Extraction and Integration of Partially Overlapping Web Sources." PVLDB 6 (2013): 805-816.

Lockard, Colin, Prashant Shiralkar and Xin Dong. "OpenCeres: When Open Information Extraction Meets the Semi-Structured Web." NAACL-HLT (2019).

Gibson, David, Kunal Punera and Andrew Tomkins. "The volume and evolution of web page templates." WWW '05 (2005).

---

# Knowledge Collection from Tabular Text

---

Colin Lockard, **Prashant Shiralkar**,  
Xin Luna Dong, Hannaneh Hajishirzi

---



# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- Part 1b: Unstructured Text: Methods (10 minutes)
- Live Q&A (15 minutes)
- Break (30 minutes)
- **Part 2: Semi-structured and Tabular Text (40 minutes)**
- Part 3: Multi-modal Extraction and Conclusion (35 minutes)
- Live Q&A (15 minutes)

# Section B. Web Table Text Extraction

# Questions we will answer in this section

## How can we extract from web tables and web lists?

Web table

#	President	Born	Age at start of presidency	Age at end of presidency	Post-presidency timespan	Lifespan	
						Died	Age
1	George Washington	Feb 22, 1732 <sup>[a]</sup>	57 years, 67 days Apr 30, 1789	65 years, 10 days Mar 4, 1797	2 years, 285 days	Dec 14, 1799	67 years, 295 days
2	John Adams	Oct 30, 1735 <sup>[a]</sup>	61 years, 125 days Mar 4, 1797	65 years, 125 days Mar 4, 1801	25 years, 122 days	Jul 4, 1826	90 years, 247 days
3	Thomas Jefferson	Apr 13, 1743 <sup>[a]</sup>	57 years, 325 days Mar 4, 1801	65 years, 325 days Mar 4, 1809	17 years, 122 days	Jul 4, 1826	83 years, 82 days
4	James Madison	Mar 16, 1751 <sup>[a]</sup>	57 years, 353 days Mar 4, 1809	65 years, 353 days Mar 4, 1817	19 years, 116 days	Jun 28, 1836	85 years, 104 days
5	James Monroe	Apr 28, 1758	58 years, 310 days Mar 4, 1817	66 years, 310 days Mar 4, 1825	6 years, 122 days	Jul 4, 1831	73 years, 67 days
6	John Quincy Adams	Jul 11, 1767	57 years, 236 days Mar 4, 1825	61 years, 236 days Mar 4, 1829	18 years, 356 days	Feb 23, 1848	80 years, 227 days
7	Andrew Jackson	Mar 15, 1767	61 years, 354 days Mar 4, 1829	69 years, 354 days Mar 4, 1837	8 years, 96 days	Jun 8, 1845	78 years, 85 days
8	Martin Van Buren	Dec 5, 1782	54 years, 89 days Mar 4, 1837	58 years, 89 days Mar 4, 1841	21 years, 142 days	Jul 24, 1862	79 years, 231 days
9	William Henry Harrison	Feb 9, 1773	68 years, 23 days Mar 4, 1841	68 years, 54 days Apr 4, 1841 <sup>[b]</sup>	0 days	Apr 4, 1841	68 years, 54 days
10	John Tyler	Mar 29, 1790	51 years, 6 days Apr 4, 1841	54 years, 340 days Mar 4, 1845	16 years, 320 days	Jan 18, 1862	71 years, 295 days

Web list

The screenshot shows a web browser window with the title "History Of The 50 Greatest Cartoons Of All Time". The address bar shows "http://t". The main content area displays the title "The 50 Greatest Cartoons" and a subtitle "from *The 50 Greatest Cartoons* by Jerry Beck (ed.) (Atlanta, Turner Publishing, 1994) ISBN# 1-878685-49-X". Below this, a numbered list of 15 cartoon titles is shown.

1.	What's Opera Doc (Warner Bros./1957)
2.	Duck Amuck (Warner Bros./1953)
3.	The Band Concert (Disney/1935)
4.	Duck Dodgers in the 24 1/2th Century (Warner Bros./1953)
5.	One Froggy Evening (Warner Bros./1956)
6.	Gertie The Dinosaur (McCay)
7.	Red Hot Riding Hood (MGM/1943)
8.	Porky In Wackyland (Warner Bros./1938)
9.	Gerald McBoing Boing (UPA/1951)
10.	King-Size Canary (MGM/1947)
11.	Three Little Pigs (Disney/1933)
12.	Rabbit of Seville (Warner Bros./1950)
13.	Steamboat Willie (Disney/1928)
14.	The Old Mill (Disney/1937)
15.	Red-Legged Blackie (MGM/1940)

# What is a web table? - (Cafarella VLDB'08 WebDB'08)

- A small relational database embedded in an HTML page. E.g. “List of U.S. presidents by age” on Wikipedia
- Different from tables for page layout, calendars and other non-relational reasons

#	President	Born	Age at start of presidency	Age at end of presidency	Post-presidency timespan	Lifespan	
						Died	Age
1	George Washington	Feb 22, 1732 <sup>[a]</sup>	57 years, 67 days Apr 30, 1789	65 years, 10 days Mar 4, 1797	2 years, 285 days	Dec 14, 1799	67 years, 295 days
2	John Adams	Oct 30, 1735 <sup>[a]</sup>	61 years, 125 days Mar 4, 1797	65 years, 125 days Mar 4, 1801	25 years, 122 days	Jul 4, 1826	90 years, 247 days
3	Thomas Jefferson	Apr 13, 1743 <sup>[a]</sup>	57 years, 325 days Mar 4, 1801	65 years, 325 days Mar 4, 1809	17 years, 122 days	Jul 4, 1826	83 years, 82 days
4	James Madison	Mar 16, 1751 <sup>[a]</sup>	57 years, 353 days Mar 4, 1809	65 years, 353 days Mar 4, 1817	19 years, 116 days	Jun 28, 1836	85 years, 104 days
			58 years 310 days	66 years 310 days			

\*Cafarella, M.J., Halevy, A.Y., Wang, D.Z., Wu, E., & Zhang, Y. (2008). WebTables: exploring the power of tables on the web. VLDB.

\*Cafarella, M.J., Halevy, A.Y., Zhang, Y., Wang, D.Z., & Wu, E. (2008). Uncovering the Relational Web. WebDB.

# Characteristics of web tables

- Unlike pure relational tables, no uniform schema
  - No column types, primary key, or foreign key
- Horizontal tables vs. vertical tables

Horizontal table

Name	Known for	Parent company	First store location
Applebee's	American	DineEquity	Decatur, Georgia
Arby's	Sandwiches	Roark Capital Group (majority)	Boardman, Ohio
Auntie Anne's	Baked goods	Focus Brands	Downingtown, Pennsylvania
Baton Rouge	Steak	Imvescor	Montreal, Quebec

Vertical table

Author	J. K. Rowling
Country	United Kingdom
Language	English
Genre	Fantasy, drama, young adult fiction, mystery, thriller, Bildungsroman
Publisher	Bloomsbury Publishing (UK) Pottermore (e-books; all languages)
Published	26 June 1997 – 21 July 2007 (initial publication)

We assume horizontal tables in this tutorial

# Characteristics of web tables

- Unlike pure relational tables, no uniform schema
  - No column types, primary key, or foreign key
- Horizontal tables vs. vertical tables
  - Horizontal tables: attribute along columns, tuples along rows
  - Vertical tables: attribute along rows, values along columns
- Diverse tables
  - Different tables may use different column names for the same underlying class
- Subject-like column vs. attributes of the subject entities

# Web contains large number of web tables!

By 2008 estimate, 154 million HTML tables are web tables (Cafarella, WebDB'08)

Cols	Raw %	Recovered %
0	1.06	0
1	42.50	0
2-9	55.00	93.18
10-19	1.24	6.17
20-29	0.19	0.46
30+	0.02	0.05

93% of web tables have 2-9 attributes (cols)

Very few tables have large number of attributes

Rows	Raw %	Recovered %
0	0.88	0
1	62.90	0
2-9	33.06	64.07
10-19	1.98	15.83
20-29	0.57	7.61
30+	0.61	12.49

Tables have much greater diversity in row counts

# What is web table extraction? -- (Cafarella VLDB'18)

Two key problems to solve:

1. **Relation recovery**: How do I detect a web table?
2. **Metadata recovery**: How I understand the semantics of a web table to extract its records?

We focus on 'Metadata recovery' in this tutorial

# Why extract from web tables?

- Table search based on keywords

Google Table - city population +阮桂芳

city population

Submit

Tables 1 - 10 of 13057 found. (22.005 seconds elapsed)

City Mayors: Largest cities in the world by population (1 to 125)

export to... show entire table

Rank(t)	City / Urban area(s)	Country	Population(t)	Land area (in sqKm)(t, 2)	Density (people per sqKm)(t, 2)
1(1.0)	Tokyo/Yokohama	Japan	33,200,000(3,3287)	8,065,093(3,3)	4,118(4,118)
2(2.0)	New York Metro	USA	17,800,000(1,7867)	8,065,093(2,2)	2,198(2,198)
3(3.0)	Sao Paulo	Brazil	17,790,000(1,7787)	1,085,169(2,2)	16,317(16,317)
4(4.0)	Seoul/Inchon	South Korea	17,000,000(1,7867)	1,046,104(2,2)	16,345(16,345)
5(5.0)	Mexico City	Mexico	17,400,000(1,7487)	2,373,057(2,2)	7,364(7,364)
6(6.0)	Osaka/Kobe/Himeji	Japan	16,420,000(1,84087)	2,394,429(4,4)	6,880(6,880)
7(7.0)	Manila	Philippines	14,755,000(1,4767)	1,386,136(9,9)	10,630(10,630)
8(8.0)	Mumbai	India	14,200,000(1,4967)	404,644(2,2)	35,183(35,183)
9(9.0)	Delhi	India	14,300,000(1,4967)	1,286,120(5,5)	11,000(11,000)
10(10.0)	Jakarta	Indonesia	14,270,000(1,4967)	1,386,136(2,2)	10,440(10,440)
11(11.0)	Lagos	Nigeria	13,400,000(1,3487)	7367,08(2,2)	18,130(18,130)
12(12.0)	Kolkata	India	13,790,000(1,3787)	81,071(1,1)	169,911(169,911)
13(13.0)	Cairo	Egypt	12,200,000(1,2287)	1,386,136(2,2)	8,880(8,880)
14(14.0)	Los Angeles	USA	11,790,000(1,17867)	4,320,403(2,2)	2,700(2,700)
15(15.0)	Buenos Aires	Argentina	11,230,000(1,1287)	2,296,029(9,9)	4,880(4,880)
16(16.0)	Rio de Janeiro	Brazil	10,800,000(1,0867)	1,386,136(2,2)	7,880(7,880)
17(17.0)	Moscow	Russia	10,200,000(1,0567)	2,182,160(2,2)	4,680(4,680)
18(18.0)	Shanghai	China	10,300,000(1,0867)	7367,08(2,2)	13,980(13,980)
19(19.0)	Karachi	Pakistan	9,600,000(0,9400000)	6160718(2,2)	154,400(154,400)
20(20.0)	Paris	France	9,540,000(0,9400000)	2,723,027(2,2)	3,480(3,480)
21(21.0)	Istanbul	Turkey	9,030,000(0,8800000)	1,386,136(2,2)	6,610(6,610)
22(22.0)	Nagoya	Japan	8,230,000(0,8000000)	2,872,027(2,2)	2,872,027(2,2)
23(23.0)	Beijing	China	8,114,000(0,8140000)	7367,08(2,2)	10,880(10,880)
24(24.0)	Chicago	USA	8,106,000(0,8080000)	5,460,949(2,2)	1,480(1,480)
25(25.0)	London	UK	8,279,000(0,8278000)	1,828,162(2,2)	4,540(4,540)
...	...	...	...	...	...
ESTIMATING CITY POPULATIONS	export to...	show entire table			
REGION	People per Hectare	Margin of Error			
Cities of Antiquity	100	10-10%			
Cities of Islam	200	20-20%			
Cities of Europe (Ireland and France)	100-110	20%			
(1000-10000)	100-110	10%			

city population

Google

All Images News Maps Videos More Settings Tools

About 798,000,000 results (0.76 seconds)

The largest US cities: Cities ranked 1 to 100

Rank	City; State	2010 population
1	New York City; New York	8,175,133
2	Los Angeles; California	3,792,621
3	Chicago; Illinois	2,695,598
4	Houston; Texas	2,099,451
...	...	...

88 more rows

City Mayors: Largest 100 US cities  
www.citymayors.com/gratis/uscities\_100.html

About this result Feedback

People also ask

What are the 10 largest cities in the world?

What are the 10 most populated cities in the world?

10

# Why extract from web tables?

- Table search based on keywords
- Schema autocomplete tool for database designers
  - For 'stock-symbol' as an input, suggest 'company', 'rank' and 'sales' as attributes to add to a schema
- Attribute synonym finding tool
  - Automatically find 'hr' = 'home run' for baseball data
- Question answering

# Key differences with text & semi-structured websites

Dimension	Unstructured text	Semi-structured websites	Web tables
Input unit	Sentence	Entity page	Table row
Consistency	Grammatical pattern	Page template	Similar-ranged values across rows
Entity pair relation	Explicit within a sentence or paragraph	Explicit to the left/top/right of object	Column semantics
NER tools available?	Yes	No	No
Context	Rich, often ambiguous	Short, clean	Short, ambiguous

# How do we detect a web table?

# Challenges in relation recovery

- HTML tables vs. other HTML structures that look like tables
- Relational vs. non-relational (“relational” in an informal sense)
- Detecting presence of a header row

# Relation recovery -- (Cafarella, WebDB'08)

**Idea:** Use generic features that discriminate a relation table from a non-relational one to create a classifier

## Features

- # rows
- # cols
- % rows w/mostly NULLS
- # cols w/non-string data
- cell strlen avg.  $\mu$
- cell strlen stddev.  $\sigma$
- cell strlen  $\frac{\mu}{\sigma}$

## Performance: Focus on recall

true class	Precision	Recall
relational	0.41	0.81
non-relational	0.98	0.87

154M relational tables  
(1.1% of raw HTML tables)

**How I understand the semantics of a web table  
to extract its records?**

# What is metadata (semantics) recovery?

**Goal:** Ideally, we want to transform a web table into a pure relational database table to reap the latter's benefits.

## Aspects of semantics recovery:

1. Subject column detection
2. Column class detection
3. Relation extraction between a column pair

# What is subject column detection?

75% of web tables have a column containing subject entities describing each row, enhancing table search quality (Venetis, VLDB'11)

**Task:** Annotate which column represents the subject entities.

Name	Known for	Parent company	First store location	Founded	Locations worldwide	Employees
Applebee's	American	DineEquity	Decatur, Georgia	1980	1830	31,500
Arby's	Sandwiches	Roark Capital Group (majority)	Boardman, Ohio	1964	3472	26,788
Auntie Anne's	Baked goods	Focus Brands	Downingtown, Pennsylvania	1988	1500+	12,000
Baton Rouge	Steak	Imvescor	Montreal, Quebec	1992	29	
BeaverTails	Baked goods		Ottawa, Ontario	1978	119	
Big Smoke Burger	Hamburgers		Toronto, Ontario	2007	19	
Bonchon Chicken	Chicken	Bonchon Chicken Inc.	Busan, South Korea	2002	64	18
Buffalo Wild Wings	Chicken	Buffalo Wild Wings, Inc.	Columbus, Ohio	1981	1228	

# What is column class (concept) detection?

'Name' or 'Restaurant' ?

**Task:** Annotate a column with its class label from an ontology.



Name	Known for	Parent company	First store location	Founded	Locations worldwide	Employees
Applebee's	American	DineEquity	Decatur, Georgia	1980	1830	31,500
Arby's	Sandwiches	Roark Capital Group (majority)	Boardman, Ohio	1964	3472	26,788
Auntie Anne's	Baked goods	Focus Brands	Downington, Pennsylvania	1988	1500+	12,000
Baton Rouge	Steak	Imvescor	Montreal, Quebec	1992	29	
BeaverTails	Baked goods		Ottawa, Ontario	1978	119	
Big Smoke Burger	Hamburgers		Toronto, Ontario	2007	19	
Bonchon Chicken	Chicken	Bonchon Chicken Inc.	Busan, South Korea	2002	64	19
Buffalo Wild Wings	Chicken	Buffalo Wild Wings, Inc.	Columbus, Ohio	1981	1228	

# What is relation extraction between a column pair?

What is the relation between (Name, Parent company) columns?

**Task:** Annotate the ontology relation between two columns

Name	Known for	Parent company	First store location	Founded	Locations worldwide	Employees
Applebee's	American	DineEquity	Decatur, Georgia	1980	1830	31,500
Arby's	Sandwiches	Roark Capital Group (majority)	Boardman, Ohio	1964	3472	26,788
Auntie Anne's	Baked goods	Focus Brands	Downington, Pennsylvania	1988	1500+	12,000
Baton Rouge	Steak	Imvescor	Montreal, Quebec	1992	29	
BeaverTails	Baked goods		Ottawa, Ontario	1978	119	
Big Smoke Burger	Hamburgers		Toronto, Ontario	2007	19	
Bonchon Chicken	Chicken	Bonchon Chicken Inc.	Busan, South Korea	2002	64	
Buffalo Wild Wings	Chicken	Buffalo Wild Wings, Inc.	Columbus, Ohio	1981	1238	

# Main challenge in metadata recovery

## Limited contextual clues

- **Subject column detection:** In absence of any additional text, how do we infer the correct column describing subject entities?
- **Column class detection:** How to assign a class label to a column when each cell can map to multiple classes/types?
- **Relation extraction between column pair:** How do we infer a relation between columns given that there is no intrinsic clue?

# Methods for web table extraction

## Relation discovery

- Table detection (Wang WWW'02, Zanibbi IJ DAR'04)
- Table extraction (Gatterbauer WWW'07)
- WebTables (Cafarella WebDB'08, VLDB'08)

## Metadata recovery

- Subject column discovery (Venetis VLDB'11, Hulsebos KDD'19)
- Column class detection (Wang ICER'12, Deng VLDB'13)
- Relation extraction (Venetis VLDB'11, Limaye VLDB'10, Gupta VLDB'14)

# Short Answers

- **Subject column detection**
  - Leverage generic features of subject entities such as value uniqueness, string type, number of characters and words
- **Column class detection**
  - Leverage external data -- web extracted triples, knowledge graph
- **Relation extraction between column pair**
  - Measure similarity between a column and entities of a type in a knowledge base

# Subject column detection as binary classification -- (Venetis, VLDB'11)

Use generic features of subject column to train a classifier

No.	Feature Description
1	<b>Fraction of cells with unique content</b>
2	<b>Fraction of cells with numeric content</b>
3	Average number of letters in each cell
4	Average number of numeric tokens in each cell
5	<b>Variance in the number of date tokens in each cell</b>
6	Average number of data tokens in each cell
7	Average number of special characters in each cell
8	<b>Average number of words in each cell</b>
9	<b>Column index from the left</b>
10	Column index excluding numbers and dates

# Performance

**Naive assignment:** Scan the table from left to right and select the first non-numeric and non-date column as the subject column

Method	Accuracy
Naive assignment	83%
SVM classifier	94%

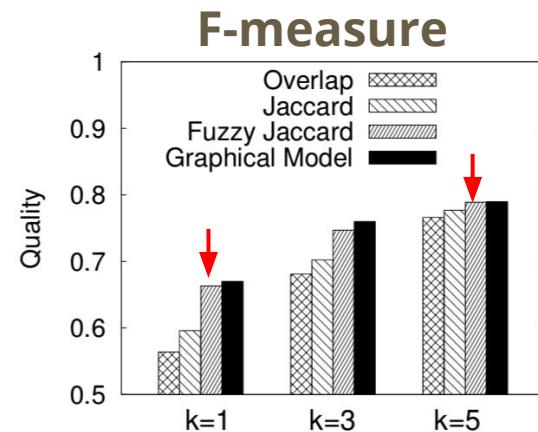
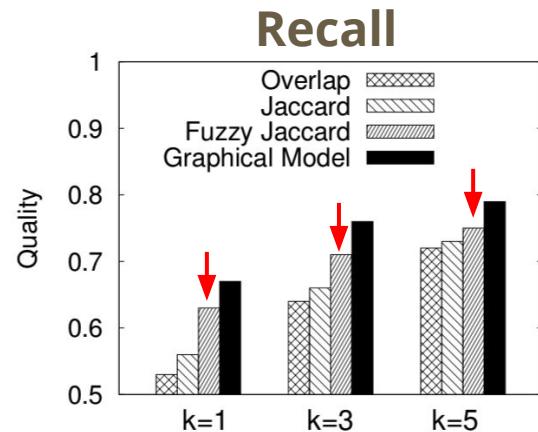
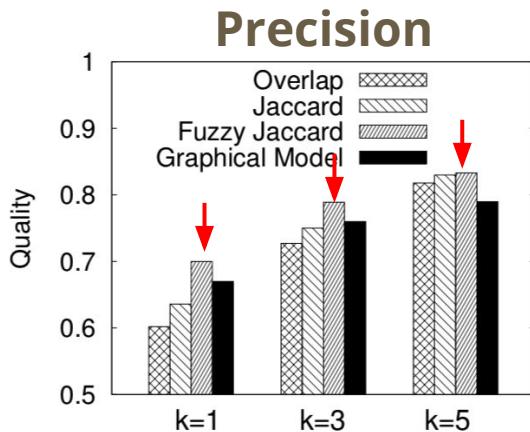
Fairly high performance

**75% of tables on the Web have a subject column**

# Column class detection -- (Deng, VLDB'13)

**Idea:** A column  $C$  can be described by a type  $T$  from an ontology, if  $T$  shares significant similarity with  $C$ .

$\text{Similarity}(T, C)$ : cell contents of  $C$  and entities of  $T$  in a knowledge base



Better precision than Graphical model  
(Limaye VLDB'10 -- coming up)

Performance for top- $k$  types ~65% F1

# Relation extraction between a column pair -- Maximum likelihood model (Venetis, VLDB'11)

**Key idea:** Look for evidence of support for column pair values in an external database of relations or knowledge base

**Intuition:** If a relation exists in external data for many rows of the table, the relation is the likely label for the column pair

$$l(A) = \arg \max_{l_i} \{ \Pr [v_1, \dots, v_n \mid l_i] \}$$

↑                              ↑  
A pair of values              relation

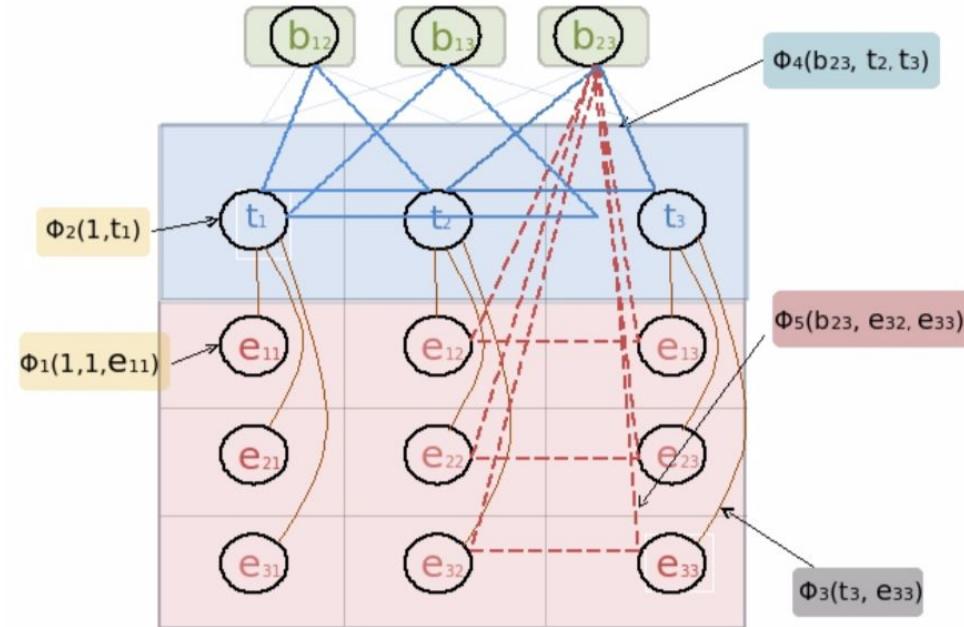
**Performance:** 45% Precision, 70% Recall (low performance)

**How can we perform all the three tasks using a single model?**

# Performing all the three tasks jointly -- probabilistic graphical model (Limaye, VLDB 2010)

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)



# Performance

0/1 loss for entity annotation accuracy

F1 score for type and relation annotation accuracy

Entity annotation accuracy			
Dataset	LCA	MAJORITY	COLLECTIVE
Wiki_Manual	59.75	74.24	<b>83.92</b>
Web_Manual	59.68	75.87	<b>81.37</b>
Wiki_Link	67.92	77.63	<b>84.28</b>

Type annotation accuracy			
Dataset	LCA	MAJORITY	COLLECTIVE
Wiki_Manual	8.63	44.60	<b>56.12</b>
Web_Manual	15.16	31.45	<b>43.23</b>

Relation annotation accuracy			
Dataset	LCA	MAJORITY	COLLECTIVE
Wiki_Manual	-	62.50	<b>68.97</b>
Web_Relations	-	60.87	<b>63.64</b>
Web_Manual	-	50.30	<b>51.50</b>

Performance is better than baselines, but the problem is still far from solved

# Recipe for web table extraction

- **Problem definition:** Extract semantics of a web table by identifying the subject column, column class, and ontological relation for pairs of columns.
- **Short answers:**
  - Catalog or external data is needed to add context to a table
  - Probabilistic graphical models solve the three annotation tasks jointly
  - Subject column detection has fairly high performance (~94%), while column type detection and relation extraction have relatively lower performance (50-70%)
  - Problem is far from solved

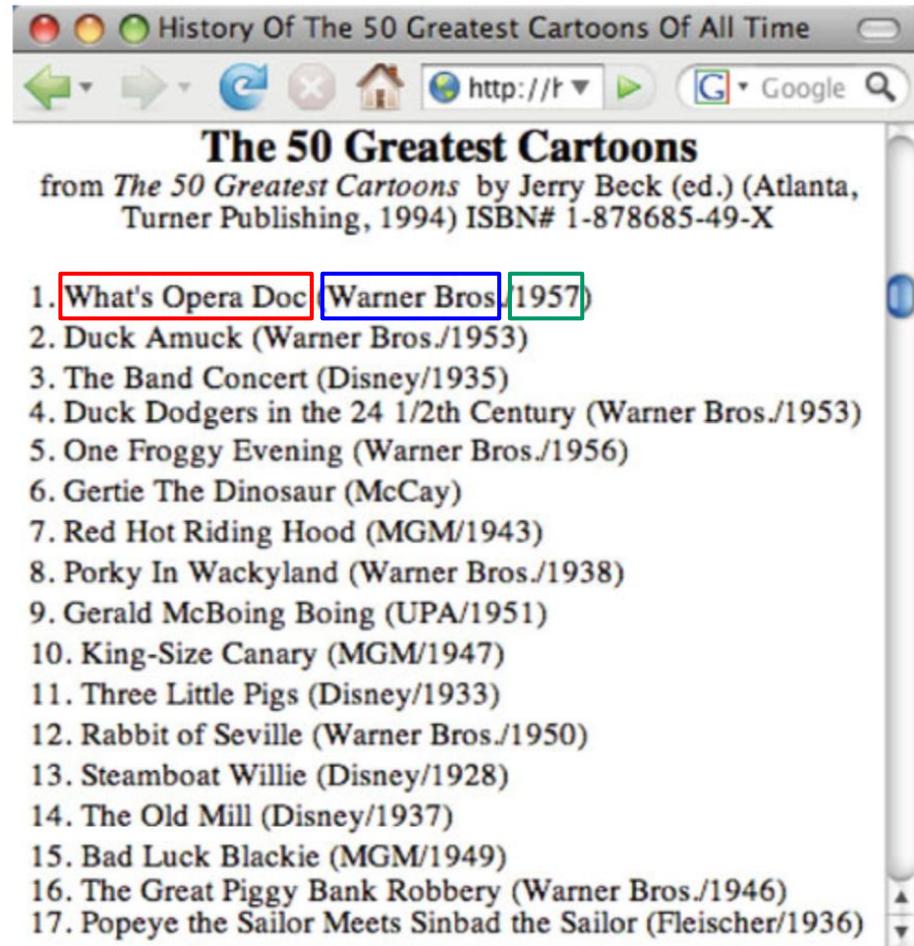
# How can we extract from a web list?

# What is a web list?

A web list is a data structure containing semi-structured data in the form of manually generated HTML list.

Not as rich a source as web tables, but large nevertheless

~100K lists (Elmeleegy VLDB'11)



The screenshot shows a web browser window with the title "History Of The 50 Greatest Cartoons Of All Time". The page content is titled "The 50 Greatest Cartoons" and includes a subtitle from "The 50 Greatest Cartoons" by Jerry Beck (ed.) (Atlanta, Turner Publishing, 1994) ISBN# 1-878685-49-X. Below the subtitle is a numbered list of 17 cartoon titles, each with its studio and year. The first item, "What's Opera Doc", is highlighted with a red box. The second item, "Warner Bros./1957", is highlighted with a blue box. The last item, "Popeye the Sailor Meets Sinbad the Sailor (Fleischer/1936)", is highlighted with a green box.

1. What's Opera Doc (Warner Bros./1957)
2. Duck Amuck (Warner Bros./1953)
3. The Band Concert (Disney/1935)
4. Duck Dodgers in the 24 1/2th Century (Warner Bros./1953)
5. One Froggy Evening (Warner Bros./1956)
6. Gertie The Dinosaur (McCay)
7. Red Hot Riding Hood (MGM/1943)
8. Porky In Wackyland (Warner Bros./1938)
9. Gerald McBoing Boing (UPA/1951)
10. King-Size Canary (MGM/1947)
11. Three Little Pigs (Disney/1933)
12. Rabbit of Seville (Warner Bros./1950)
13. Steamboat Willie (Disney/1928)
14. The Old Mill (Disney/1937)
15. Bad Luck Blackie (MGM/1949)
16. The Great Piggy Bank Robbery (Warner Bros./1946)
17. Popeye the Sailor Meets Sinbad the Sailor (Fleischer/1936)

# Challenges in extracting a web list

- Largely unstructured, inconsistent delimiters

Missing delimiter?



- Ella Koon, Hong Kong singer
- Ella Maillart (1903–1997), Swiss adventurer, travel writer, photographer and sportswoman
- Ella Mae Morse (1924–1999), American popular singer from the 1940s
- Ella Pamfilova (born 1953), Russian politician
- Ella (singer) (born 1966), popular Malaysian rock singer

# Challenges in extracting a web list

- Missing information

□ Ella Koon, Hong Kong singer

- Name, city, job

□ Ella Maillart (1903-1997), Swiss adventurer, travel writer, photographer and sportswoman

- Name, birth date, death date, jobs

□ Ella Pamfilova (born 1953), Russian politician

- Name, birth date, job

# Extracting from web lists -- (Elmeleegy VLDB'11)

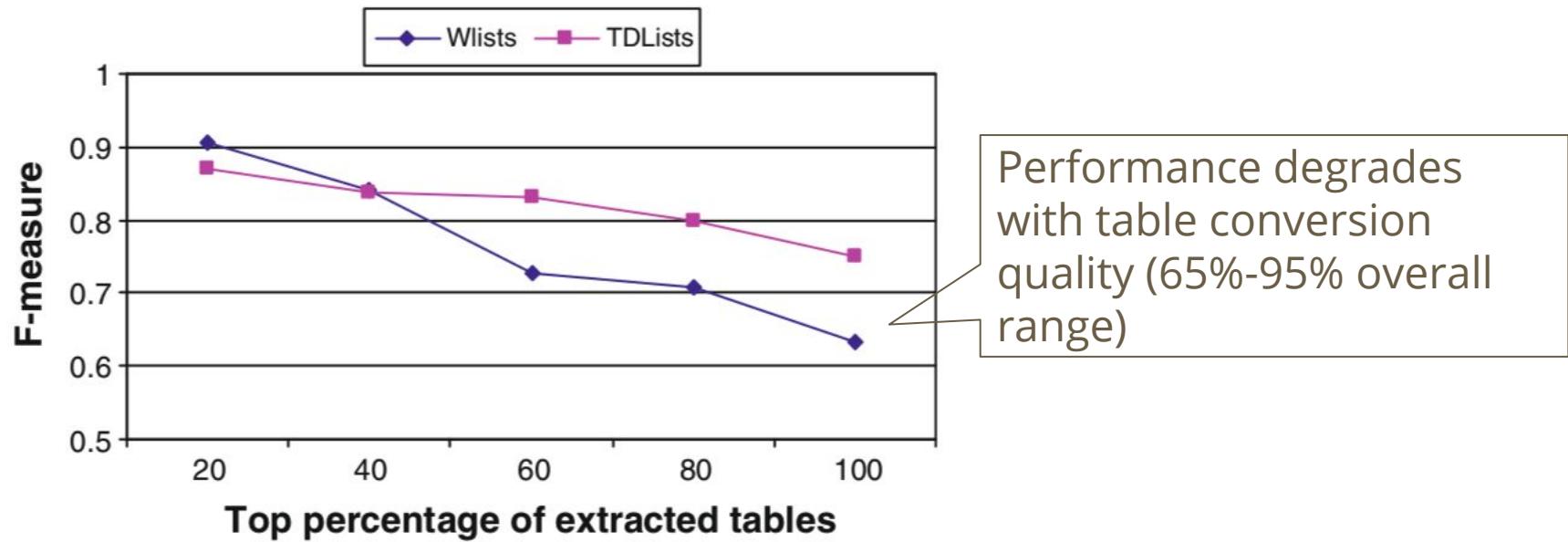
**Idea:** Transform a list into table

**Recipe:**

- 1. Independent splitting:** split each line in the list
- 2. Alignment:** align fields into columns
- 3. Refinement:** detect and fix incorrect fields

1		What's Opera Doc		Warner Bros		1957
2		Duck Amuck		Warner Bros		1953
3		The Band Concert		Disney		1935
4.	Duck Dodgers in the 24 1/2th Century	(Warner Bros		1953		
5		One Froggy Evening		Warner Bros		1956
6		Gertie The Dinosaur		McCay		
7		Red Hot Riding Hood		MGM		1943
8		Porky In Wackyland		Warner Bros		1938
9		Gerald McBoing Boing		UPA		1951
10		King-Size Canary		MGM		1947
11		Three Little Pigs		Disney		1933
12		Rabbit of Seville		Warner Bros		1950
13		Steamboat Willie		Disney		1928
14		The Old Mill		Disney		1937
15		Bad Luck Blackie	(MGM		1949	
16		The Great Piggy Bank Robbery		Warner Bros		1946
17		Popeye the Sailor		Meets		Sinbad the Sailor
						Fleischer
						1936

# Extracting from web lists -- (Elmeleegy VLDB'11)



# Recipe for web list extraction

- **Problem definition:** Extract semantics of a web list by creating structured records from semi-structured lines.
- **Short answers:**
  - Convert a web list into a web table
  - Performance depends on table conversion ability

# References

- Cafarella, Michael J., Alon Y. Halevy, Daisy Zhe Wang, Eugene Wu and Yang Zhang. "WebTables: exploring the power of tables on the web." *VLDB* 1 (2008): 538-549.
- Cafarella, Michael J., Alon Y. Halevy, Yang Zhang, Daisy Zhe Wang and Eugene Wu. "Uncovering the Relational Web." *WebDB* (2008).
- Limaye, Girija, Sunita Sarawagi and Soumen Chakrabarti. "Annotating and Searching Web Tables Using Entities, Types and Relationships." *VLDB* 3 (2010): 1338-1347.
- Venetis, Petros, Alon Y. Halevy, Jayant Madhavan, Marius Pasca, Warren Shen, Fei Wu, Gengxin Miao and Chung Wu. "Recovering Semantics of Tables on the Web." *VLDB* 4 (2011): 528-538.
- Elmeleegy, Hazem, Jayant Madhavan and Alon Y. Halevy. "Harvesting Relational Tables from Lists on the Web." *VLDB* 2 (2009): 1078-1089.

# References

- Wang, Jingjing, Haixun Wang, Zhongyuan Wang and Kenny Q. Zhu. "Understanding Tables on the Web." ER (2012).
- Cafarella, Michael J., Alon Y. Halevy, Hongrae Lee, Jayant Madhavan, Cong Yu, Daisy Zhe Wang and Eugene Wu. "Ten Years of WebTables." PVLDB 11 (2018): 2140-2149.
- Deng, Dong, Yu Jiang, Guoliang Li, Jian Li and Cong Yu. "Scalable Column Concept Determination for Web Tables Using Large Knowledge Bases." PVLDB 6 (2013): 1606-1617.
- Gupta, Rahul, Alon Y. Halevy, Xuezhi Wang, Steven Euijong Whang and Fei Wu. "Biperpedia: An Ontology for Search Applications." PVLDB 7 (2014): 505-516.
- Zanibbi, Richard, Dorothea Blostein and James R. Cordy. "A survey of table recognition." Document Analysis and Recognition 7 (2004): 1-16.

# References

- Gatterbauer, Wolfgang, Paul Bohunsky, Marcus Herzog, Bernhard Krüpl and Bernhard Pollak.  
"Towards domain-independent information extraction from web tables." WWW '07 (2007).
- Wang, Yalin and Jianying Hu. "A machine learning based approach for table detection on the web." WWW '02 (2002).
- Hulsebos, Madelon, et al. "Sherlock: A deep learning approach to semantic data type detection." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019.

# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- Part 1b: Unstructured Text: Methods (10 minutes)
- Live Q&A (15 minutes)
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- **Part 3: Multi-modal Extraction and Conclusion (35 minutes)**
- Live Q&A (15 minutes)

---

# Knowledge Collection with Multi-modal Signals

---

— Colin Lockard, Prashant Shiralkar,  
Xin Luna Dong, Hannaneh Hajishirzi

---



# What is multi-modal extraction?

- Methods that jointly consider text found in different modalities on a webpage
  - e.g. An entity mentioned both in unstructured and tabular text
- Methods that combine signals from more than one modality to improve extraction
  - Including textual semantics, table position, layout, visual features

# Why consider multi-modal signals?

## 登録情報

スタイル名: mineo SIMエントリーパッケージ(紙版)

注意事項 [354 KB PDF]

商品重量: 9.07 g

発送重量: 9.1 g

メーカー型番: 511015

ASIN: B00UT26M0Q

Amazon.co.jp での取り扱い開始日: 2015/3/27

おすすめ度: ★★★★☆ 2,247件のカスタマーレビュー

Amazon 売れ筋ランキング: 家電&カメラ - 206位 (家電&カメラの売れ筋ランキングを見る)

3位 – 定期契約SIMカード

さらに安い価格について知らせる

Subsection of page with consistent formatting

Horizontal alignment suggests (relation, object) pair

Textual semantics tell us "9.07 g" is likely object, not predicate

# Short answers

- **Diversity**
  - Textual, layout, and visual signals can combine to form consistent patterns
- **Training data**
  - Multi-modal signals allow for accurate and easy creation of training data with Data Programming
- **OpenIE**
  - Visual semantics help make OpenIE extractions from semi-structured documents without prior knowledge of the subject domain

	<b>Unstructured</b>	<b>Semi-structured</b>	<b>Tabular</b>	<b>Multi-modal</b>
<b>Input data</b>	Raw text (sentence, paragraph, or document)	Detail page HTML	Rows and columns	HTML + Rendered visuals
<b>Diversity Challenges</b>	Languages and dialects, diversity of expression	Templates, topic domain, relation strings	Topic domains	All: Language, template, topic
<b>Consistent Patterns</b>	Lexical/syntactic, textual semantics	Absolute or relative DOM location	Entity types, entity linking	Textual, Layout, and Visual semantics

**How can we connect values found in different modalities of text?**

# BriQ (Ibrahim et al, 2019)

Align mentions in unstructured text with mentions in tabular text

Focused on quantities

May differ in units, aggregation, rounding

# BriQ (Ibrahim et al, 2019)

A total of 123 patients who undergo the drug trials reported side effects, of which there were 69 female patients and 54 male patients. The most common side affect is depression, reported by 38 patients; and the least common side affect is eye disorder, reported by 5 patients.

side effects	male	female	total
Rash	15	20	35
Depression	13	25	38
Hypertension	19	15	34
Nausea	5	6	11
Eye Disorders	2	3	5

The final ratings are dominated by the PHEV from Audi (2.67) and ICE from Volkswagen (2.67). Audi A3 e-tron is the least affordable option with 37K EUR in Germany and 39K USD in the US. The Ford Focus Electric, lowest rating (1.33), is a 2K EUR (2.3K USD) cheaper alternative with 0 CO<sub>2</sub> emission and 105 MPGe fuel consumption.

	BEV Focus E	PHEV A3	ICE VW Golf
German MSRP	34900	36900	33800
American MSRP	29120	38900	29915
Emission (g/km)	0	105	122
Fuel Economy	105	70.6	61.4
Final rating	1.33	2.67	2.67

a) Example about Health

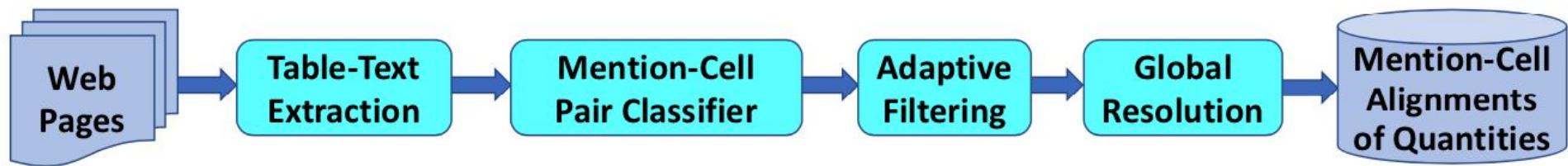
\*Ibrahim, Y., Riedewald, M., Weikum, G., & Zeinalipour-Yazti, D. (2019). Bridging Quantities in Tables and Text. ICDE.

In 2013 revenue of \$3.26 billion CDN was up \$70 million CDN or 2% from the previous year. The net income of 2013 was \$0.9 billion CDN. Compared to the revenue of 2012, it increased by 1.5%.

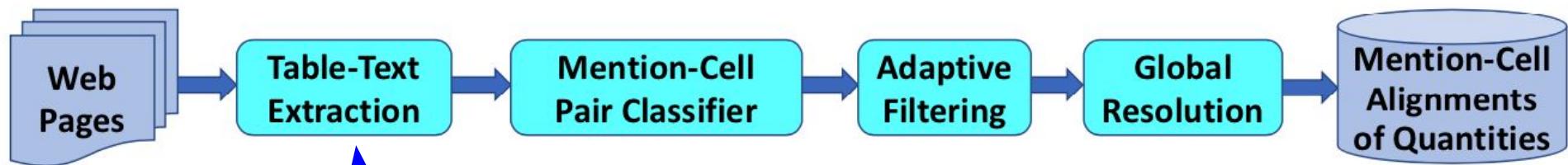
	Income gains (in Mio)	2013	2012	2011
Total Revenue	3,263	3,193	2,911	
Gross income	1,069	1,053	0,877	
Income taxes	179	177	160	
Income	890	876	849	

c) Example about Finance

# BriQ (Ibrahim et al, 2019)

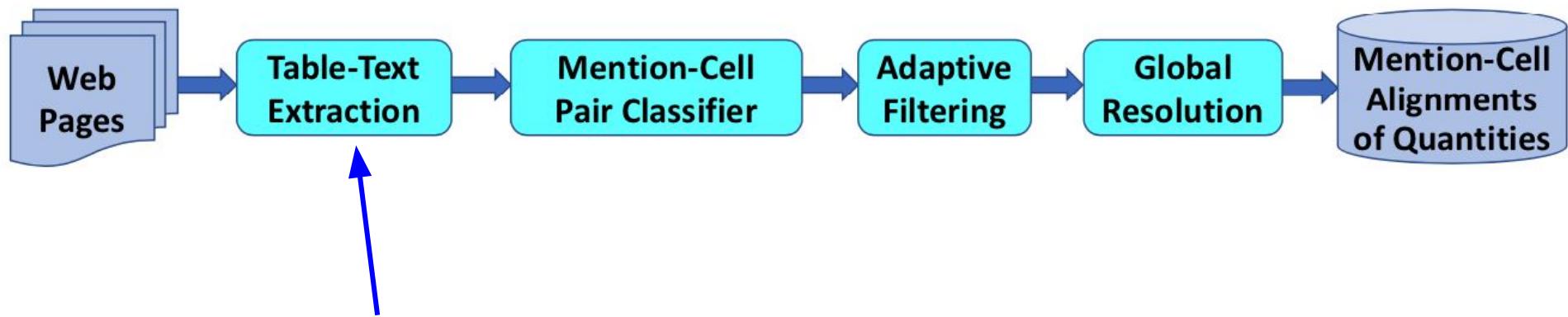


# BriQ (Ibrahim et al, 2019)



Get text and  
tables from  
webpage, find  
numeric mentions

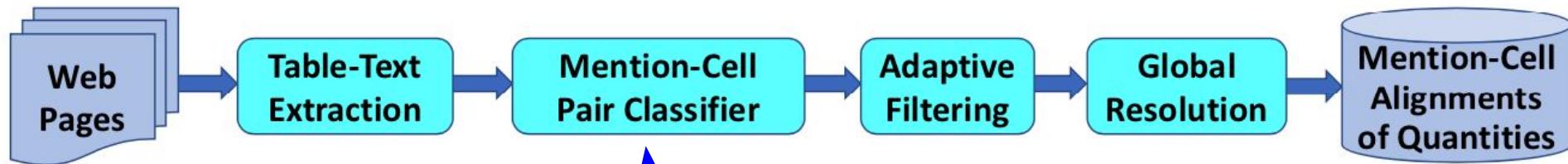
# BriQ (Ibrahim et al, 2019)



Additionally, create  
“virtual” table cells with  
aggregations of  
row/column quantities

- Sum
- Difference
- Percentage
- Change ratio

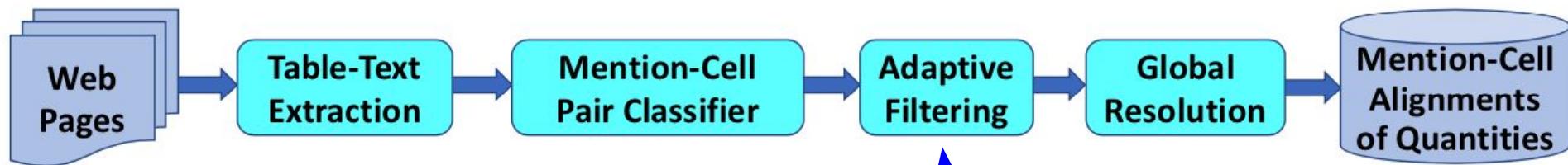
# BriQ (Ibrahim et al, 2019)



Binary classification of text/table quantity pairs as being likely/unlikely to indicate same quantity

- Features include:
- Scale diff
  - Precision diff
  - Unit match
  - Text context

# BriQ (Ibrahim et al, 2019)

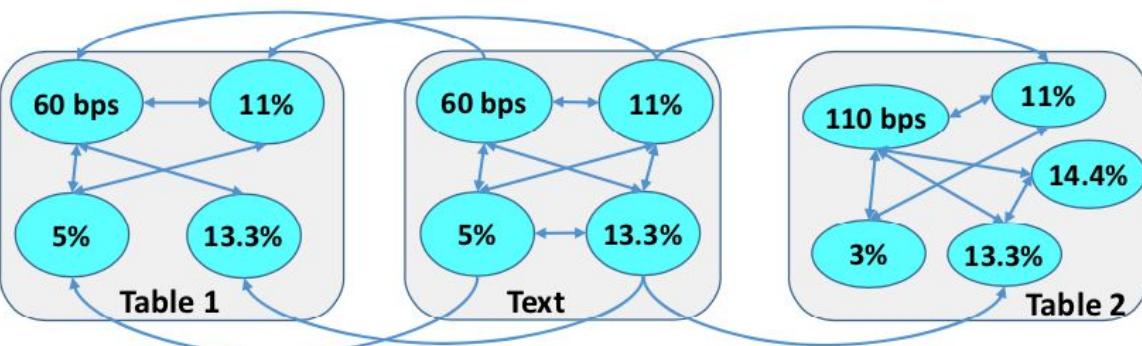
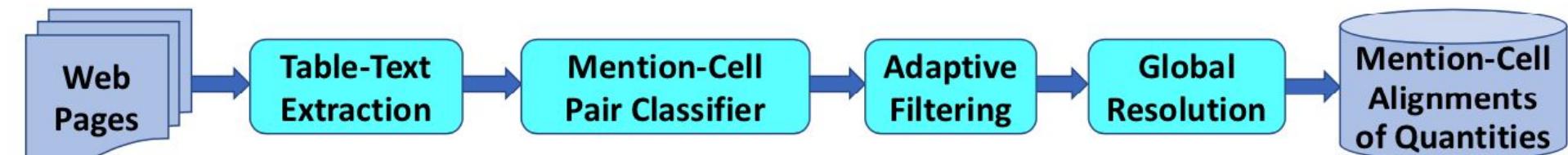


## Signals:

- Classifier confidence
- Text context mentions
- aggregation function
- Value difference

Prune to best options

# BriQ (Ibrahim et al, 2019)



Joint inference over remaining pair options

Random Walk with Restarts over mention graph

# BriQ (Ibrahim et al, 2019)

RESULTS FOR *original, truncated and rounded* TEXT MENTIONS.

	Original			Truncated			Rounded		
	RF	RWR	BriQ	RF	RWR	BriQ	RF	RWR	BriQ
recall	0.43	0.52	0.68	0.27	0.42	0.58	0.13	0.34	0.49
prec.	0.37	0.53	0.79	0.25	0.44	0.63	0.10	0.35	0.52
F1	0.40	0.53	0.73	0.26	0.43	0.60	0.11	0.34	0.51

Rounded values increase the difficulty of the task

# BriQ (Ibrahim et al, 2019)

- Link quantity values in unstructured text and tables
- Pros:
  - Allows for matching when values are aggregated/rounded/truncated
- Cons:
  - Only works for quantities
  - Doesn't perform extraction

# ACL 2020 Shoutout: TAPAS

## TAPAS: Weakly Supervised Table Parsing via Pre-training

Jonathan Herzig<sup>1,2</sup>, Paweł Krzysztof Nowak<sup>1</sup>, Thomas Müller<sup>1</sup>,  
Francesco Piccinno<sup>1</sup>, Julian Martin Eisenschlos<sup>1</sup>

<sup>1</sup>Google Research

<sup>2</sup>School of Computer Science, Tel-Aviv University

{jherzig, pawelnow, thomasmueller, piccinno, eisenjulian}@google.com

- QA over tables from webpages.
- Use unstructured text in pretraining process to build table representations.

**How can we combine signals from diverse multi-modal features?**

# How can we combine signals from diverse multi-modal features?

- Emerging research problem
- Shallow combination: Concatenate together features of different types
  - Bling-KPE
- Deep combination: Build multi-modal interactions into structure of model
  - CharGrid (Convolutional Neural Networks)
  - GraphIE (Graph Neural Networks)

# Bling-KPE (Xiong et al, 2019)

Goal: “Keyphrase” extraction from webpages

Typical approach: Use only unstructured text

# Bling-KPE (Xiong et al, 2019)

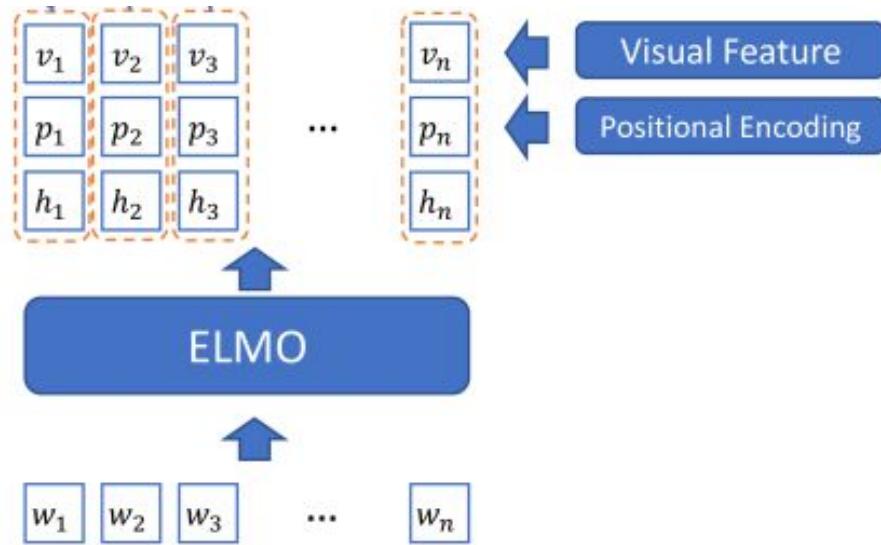
Goal: “Keyphrase” extraction from webpages

Typical approach: Use only unstructured text

This method: Incorporate visual features

# Bling-KPE

- Start with ELMO word embedding method
- Visual features capture size, location, font, and DOM info

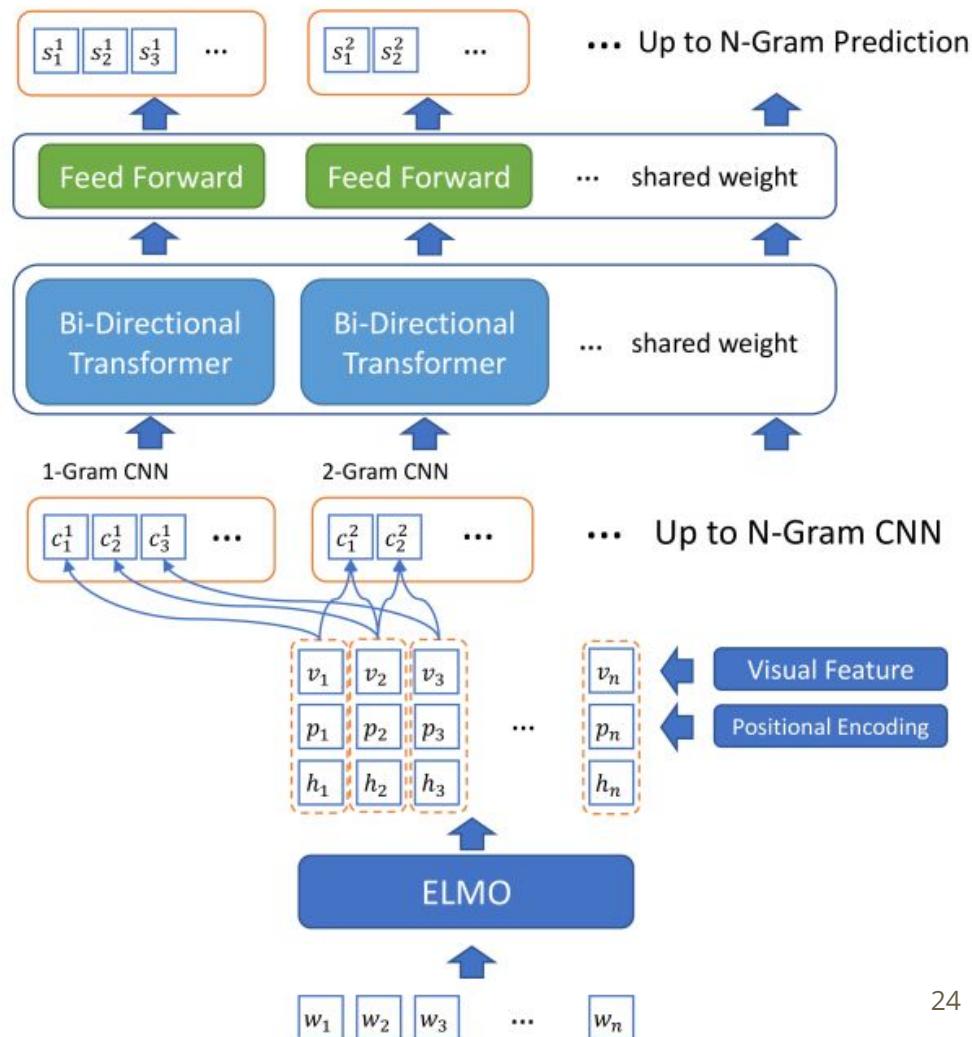


Name	Dimension
Font Size	$1 \times 2$
Text Block Size	$2 \times 2$
Location in Rendered Page	$2 \times 2$
Is Bold Font	$1 \times 2$
Appear In Inline	$1 \times 2$
Appear In Block	$1 \times 2$
Appear In DOM Tree Leaf	$1 \times 2$

# Bling-KPE

Convolution over n-grams models potential keyphrases

Weak supervision from search logs



eBay &gt; Consumer Electronics

&gt; Radio Communication &gt; Parts &amp; Accessories &gt; Manuals &amp; Magazines

## Bostitch 651S5 7/16-inch by 2-inch Stapler



3 product ratings | About this product



Brand new: lowest

**\$185.00**

Free Shipping

Get it by Monday, Mar 11 frc

- New condition
- 30 day returns - Buyer p

*"The 16 GA 7/16" Construction fire engine that produces 1 is ideal for applications of i*

[See details](#)

# Bling-KPE results

Method	P@1	R@1
TFIDF	0.283	0.150
TextRank	0.077	0.041
LeToR	0.301	0.158
PROD	0.353	0.188
PROD (Body)	0.214	0.094
CopyRNN	0.288	0.174
BLING-KPE	<b>0.404</b>	<b>0.220</b>

Significant improvement  
over strong TFIDF  
baseline

# Bling-KPE ablation study

Method	P@1	R@1
No ELMo	0.270	0.145
No Transformer	0.389	0.211
No Position	0.394	0.213
No Visual	0.370	0.201
No Pretraining	0.369	0.198
Full Model	0.404	0.220

Textual semantics are  
biggest contributor

Visual features also  
help

# Bling-KPE

- Combines textual and visual semantics
- Pros:
  - Weak supervision from search logs
  - Uses visual features
- Cons:
  - Single extraction class, no relations between text fields
  - Shallow feature interaction

# CharGrid (Katti et al, 2018)

- IE from semi-structured and visually rich documents such as invoices
- Motivation:
  - Approach IE as computer vision task
  - Problem: Learning from raw pixels forces learning language from scratch
  - Solution: Model as 2D grid of pixels, but pixel value is character, not color
- Used in production in SAP Concur

# CharGrid

Original document (pdf, html, docx, ppt...)

Title: Chargrid

paperID  
0245

conference  
EMNLP

year  
2018

Submission Date: 22.05.2018

Preserves  
2D layout

Operates on  
pixels

Document as image

Title: Chargrid

paperID  
0245

conference  
EMNLP

year  
2018

Submission Date: 22.05.2018

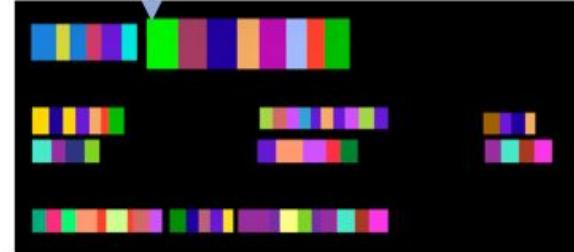
Operates on text  
Ignores 2D layout

Operates on text  
Preserves 2D layout

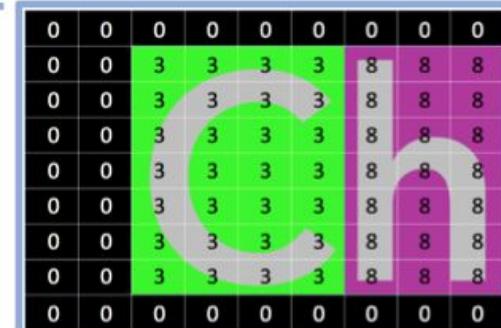
Document as serialized string

Title: Chargrid \n \n paperID  
conference year \n 0245 EMNLP  
2018 \n \n Submission Date:  
22.05.2018

Our Approach: Document as Chargrid



Chargrid zoom-in



# CharGrid

Original document (pdf, html, docx, ppt...)

Title: **Chargrid**

paperID  
0245      conference  
EMNLP      year  
2018

Submission Date: 22.05.2018

Preserves  
2D layout  
Operates on  
pixels  
Document as image

Title: **Chargrid**

paperID  
0245      conference  
EMNLP      year  
2018

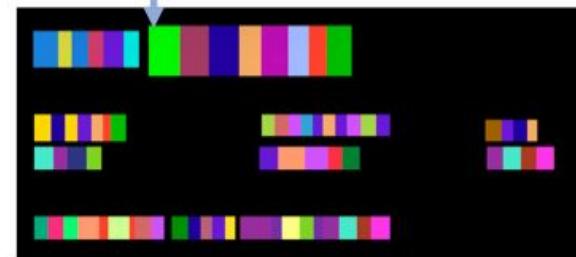
Submission Date: 22.05.2018

Operates on text  
Ignores 2D layout

Operates on text  
Preserves 2D layout

D  
Title:  
confer  
2018 \n \n Submission Date:  
22.05.2018

Our Approach: Document as Chargrid



# CharGrid

Original document (pdf, html, docx, ppt...)

Title: **Chargrid**  
paperID 0245 conference EMNLP year 2018  
Submission Date: 22.05.2018

Preserves 2D layout  
Operates on pixels  
Document as image

Title: **Chargrid**  
paperID 0245 conference EMNLP year 2018  
Submission Date: 22.05.2018

Operates on text  
Ignores 2D layout

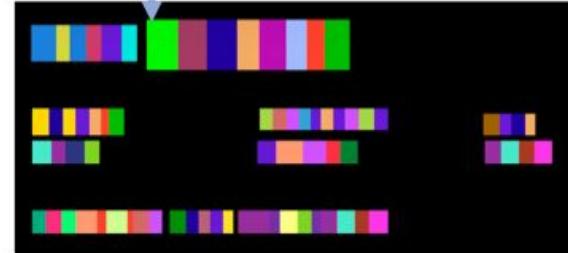
Operates on text  
Preserves 2D layout

Document as serialized string

Title: Chargrid \n \n paperID  
conference year \n 0245 EMNLP  
2018 \n \n Submission Date:  
22.05.2018

- Replace pixel values with character value

Our Approach: Document as Chargrid



Chargrid zoom-in



# CharGrid

Original document (pdf, html, docx, ppt...)

Title: **Chargrid**

paperID  
0245      conference  
EMNLP      year  
2018

Submission Date: 22.05.2018

Preserves 2D layout  
Operates on pixels  
Document as image

Title: **Chargrid**

paperID  
0245      conference  
EMNLP      year  
2018

Submission Date: 22.05.2018

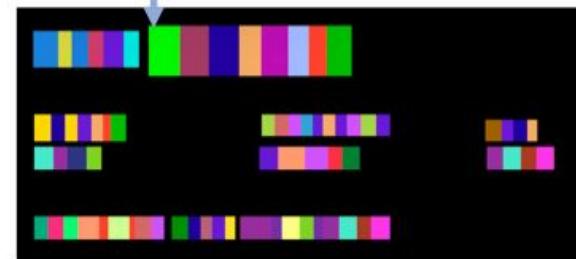
Operates on text  
Ignores 2D layout

Operates on text  
Preserves 2D layout

D  
Title: confer  
2018 \n \n Submission Date:  
22.05.2018

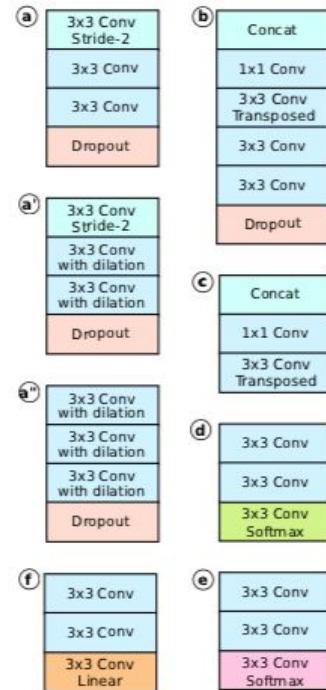
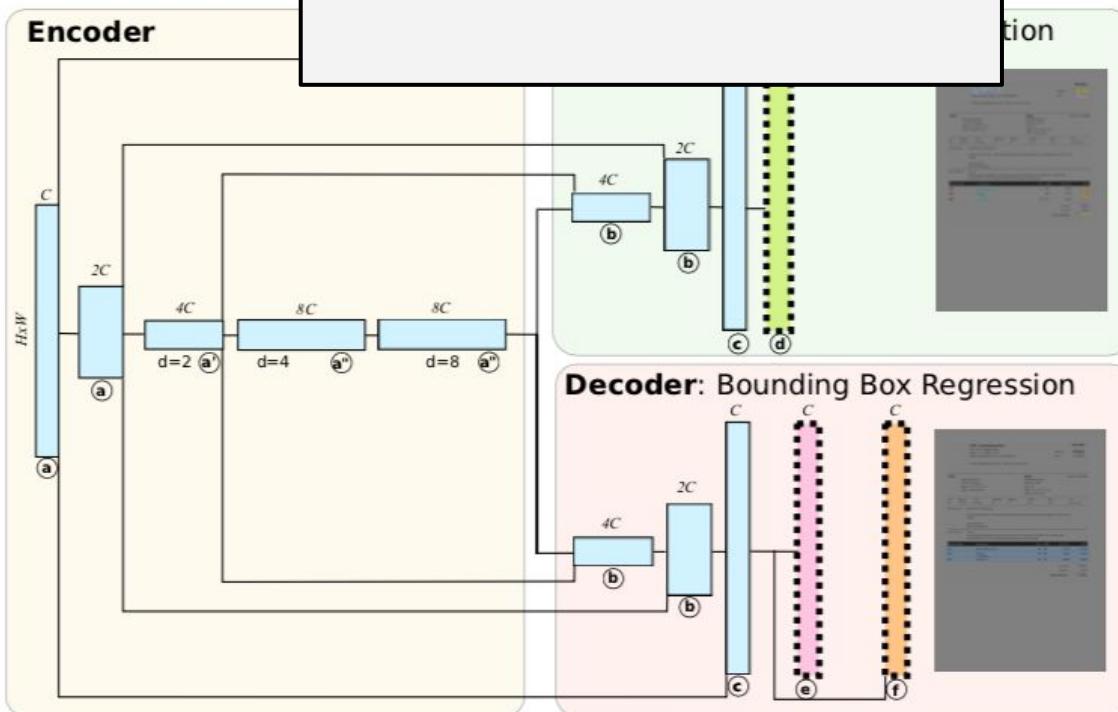
- This new “CharGrid” becomes input to convolutional neural network

Our Approach: Document as Chargrid



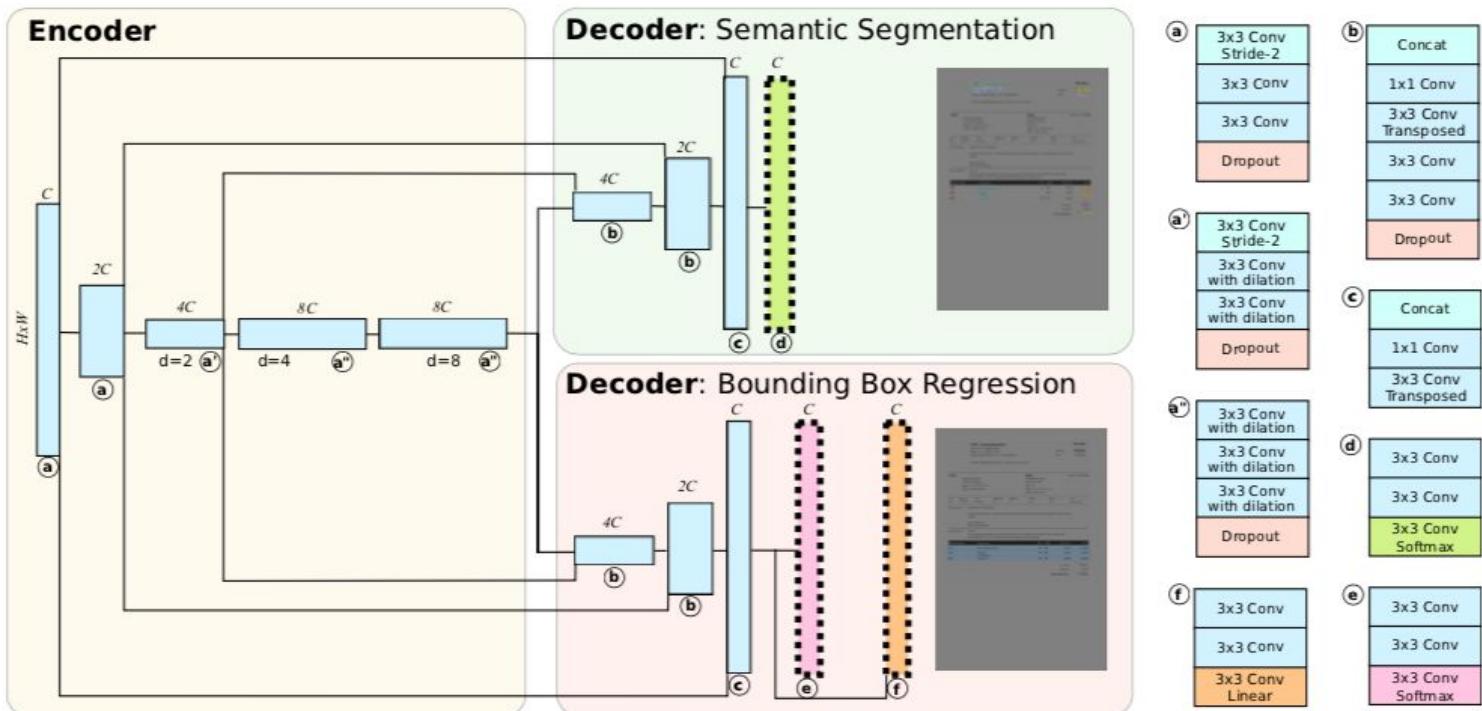
# CharGrid

- VGGNet Convolutional Neural Network encodes CharGrid



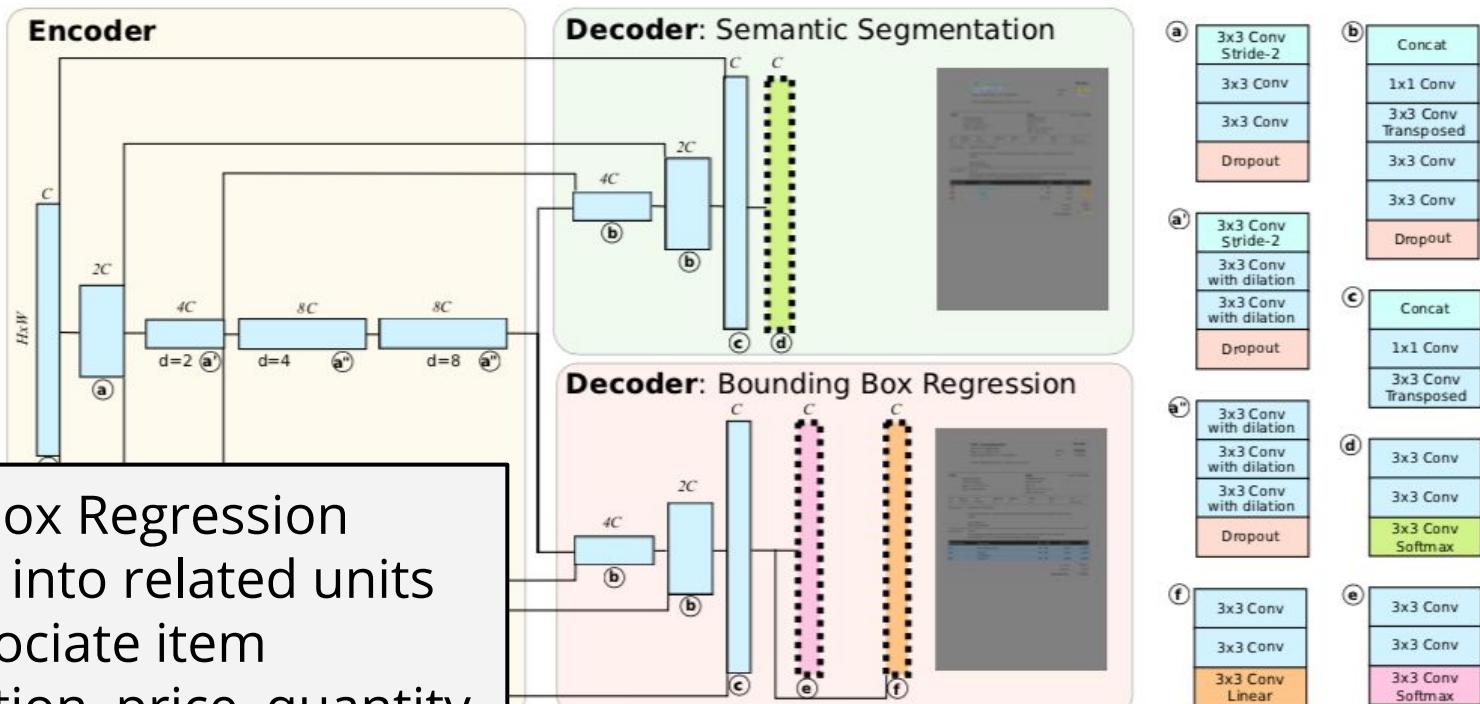
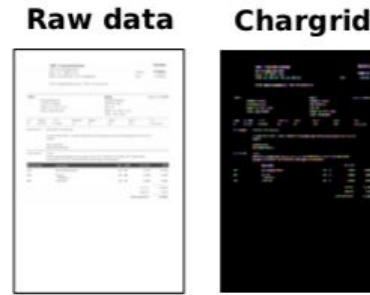
# CharGrid

- Semantic segmentation assigns each character to a class



# CharGrid

- Semantic segmentation assigns each character to a class



- Bounding Box Regression groups text into related units
  - e.g. associate item description, price, quantity

# CharGrid

Model/Field	Invoice Number	Invoice Amount	Invoice Date	Vendor Name	Vendor Address	Line-item Description	Line-item Quantity	Line-item Amount
sequential	80.98%	79.13%	83.98%	28.97%	16.94%	-0.01%	-0.18%	0.22%
image-only	47.79%	68.91%	45.67%	19.68%	13.99%	49.50%	46.79%	63.49%
chargrid-net	80.48%	80.74%	83.78%	36.00%	39.13%	52.80%	65.20%	65.57%
chargrid-hybrid-C32	74.85%	77.93%	80.40%	32.00%	31.48%	46.27%	64.04%	63.25%
chargrid-hybrid-C64	82.49%	80.14%	84.28%	34.27%	36.83%	48.81%	64.59%	64.53%

CharGrid is similar to text-only for invoice number, amount, date

- Text values very informative

# CharGrid

Model/Field	Invoice Number	Invoice Amount	Invoice Date	Vendor Name	Vendor Address	Line-item Description	Line-item Quantity	Line-item Amount
sequential	80.98%	79.13%	83.98%	28.97%	16.94%	-0.01%	-0.18%	0.22%
image-only	47.79%	68.91%	45.67%	19.68%	13.99%	49.50%	46.79%	63.49%
chargrid-net	80.48%	80.74%	83.78%	36.00%	39.13%	52.80%	65.20%	65.57%
chargrid-hybrid-C32	74.85%	77.93%	80.40%	32.00%	31.48%	46.27%	64.04%	63.25%
chargrid-hybrid-C64	82.49%	80.14%	84.28%	34.27%	36.83%	48.81%	64.59%	64.53%

Names and addresses have more textual diversity.  
CharGrid wins here.

# CharGrid

Model/Field	Invoice Number	Invoice Amount	Invoice Date	Vendor Name	Vendor Address	Line-item Description	Line-item Quantity	Line-item Amount
sequential	80.98%	79.13%	83.98%	28.97%	16.94%	-0.01%	-0.18%	0.22%
image-only	47.79%	68.91%	45.67%	19.68%	13.99%	49.50%	46.79%	63.49%
chargrid-net	80.48%	80.74%	83.78%	36.00%	39.13%	52.80%	65.20%	65.57%
chargrid-hybrid-C32	74.85%	77.93%	80.40%	32.00%	31.48%	46.27%	64.04%	63.25%
chargrid-hybrid-C64	82.49%	80.14%	84.28%	34.27%	36.83%	48.81%	64.59%	64.53%

Line-item values require associating multiple text fields.  
Bounding box detection makes this possible for  
CharGrid.

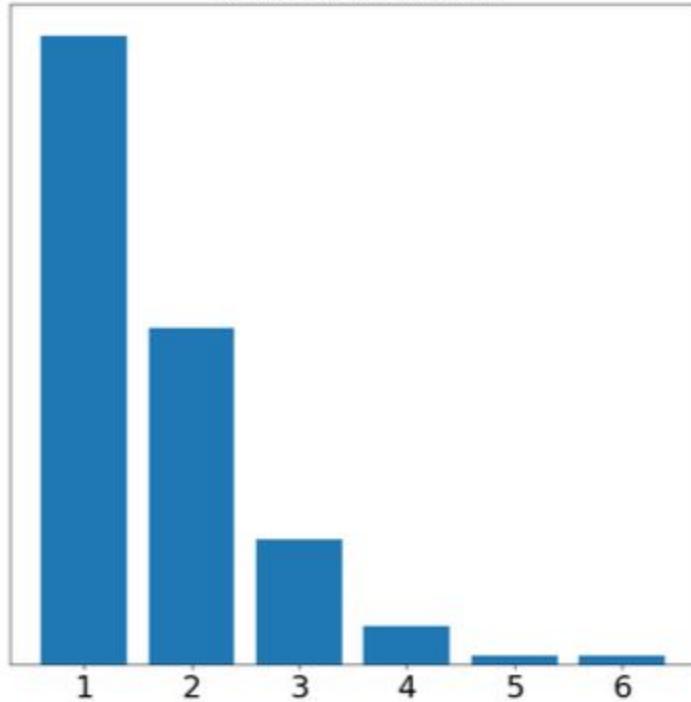
# CharGrid

Model/Field	Invoice Number	Invoice Amount	Invoice Date	Vendor Name	Vendor Address	Line-item Description	Line-item Quantity	Line-item Amount
sequential	80.98%	79.13%	83.98%	28.97%	16.94%	-0.01%	-0.18%	0.22%
image-only	47.79%	68.91%	45.67%	19.68%	13.99%	49.50%	46.79%	63.49%
chargrid-net	80.48%	80.74%	83.78%	36.00%	39.13%	52.80%	65.20%	65.57%
chargrid-hybrid-C32	74.85%	77.93%	80.40%	32.00%	31.48%	46.27%	64.04%	63.25%
chargrid-hybrid-C64	82.49%	80.14%	84.28%	34.27%	36.83%	48.81%	64.59%	64.53%

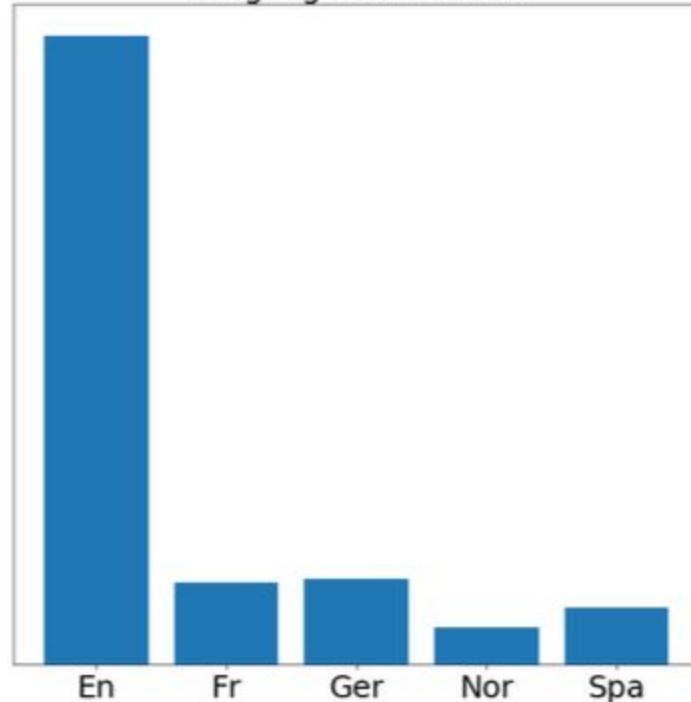
Hybrid models add image-only features to CharGrid.  
They provide little improvement.

# CharGrid

Vendor distribution



Language distribution

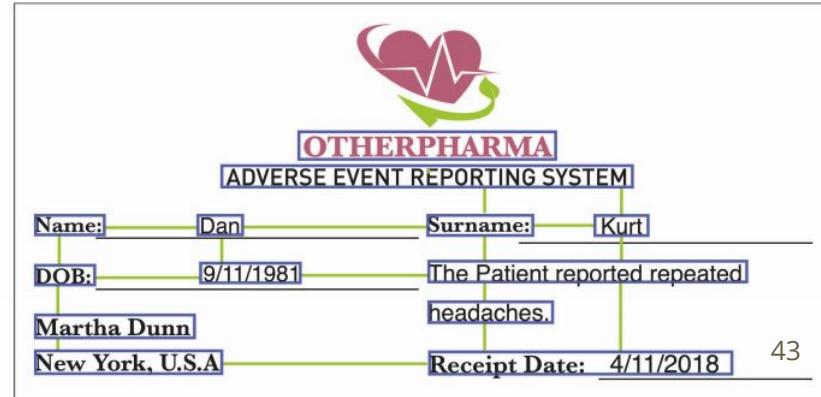
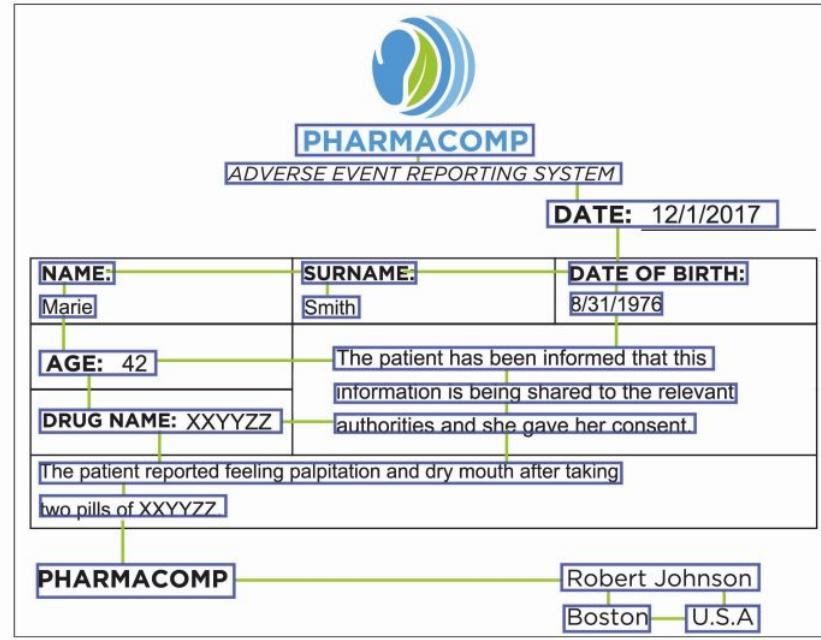


# CharGrid

- Convert image into 2D grid of characters, process with CNN
- Pros:
  - Learns layout semantics
- Cons:
  - No language priors

# GraphIE (Qian et al, 2019)

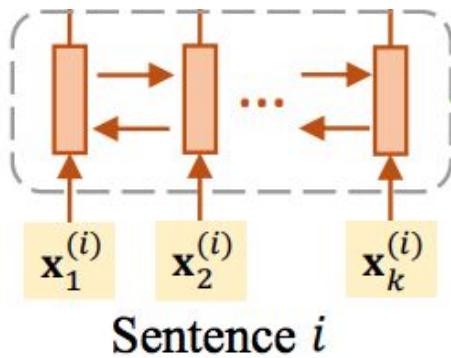
- Combine textual and layout information of semi-structured documents
- Model documents as a graph
  - Nodes are text fields
  - Edges indicate horizontal/vertical adjacency between pair of text fields



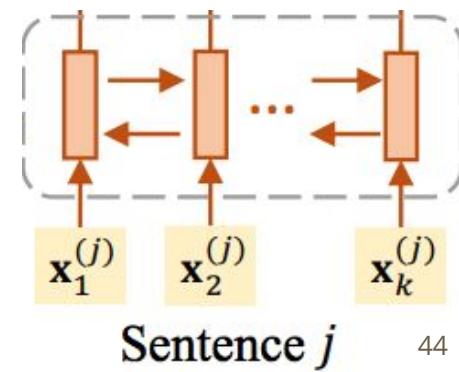
# GraphIE

Encode text in each text field with LSTM.

Encoder  
(BiLSTM)

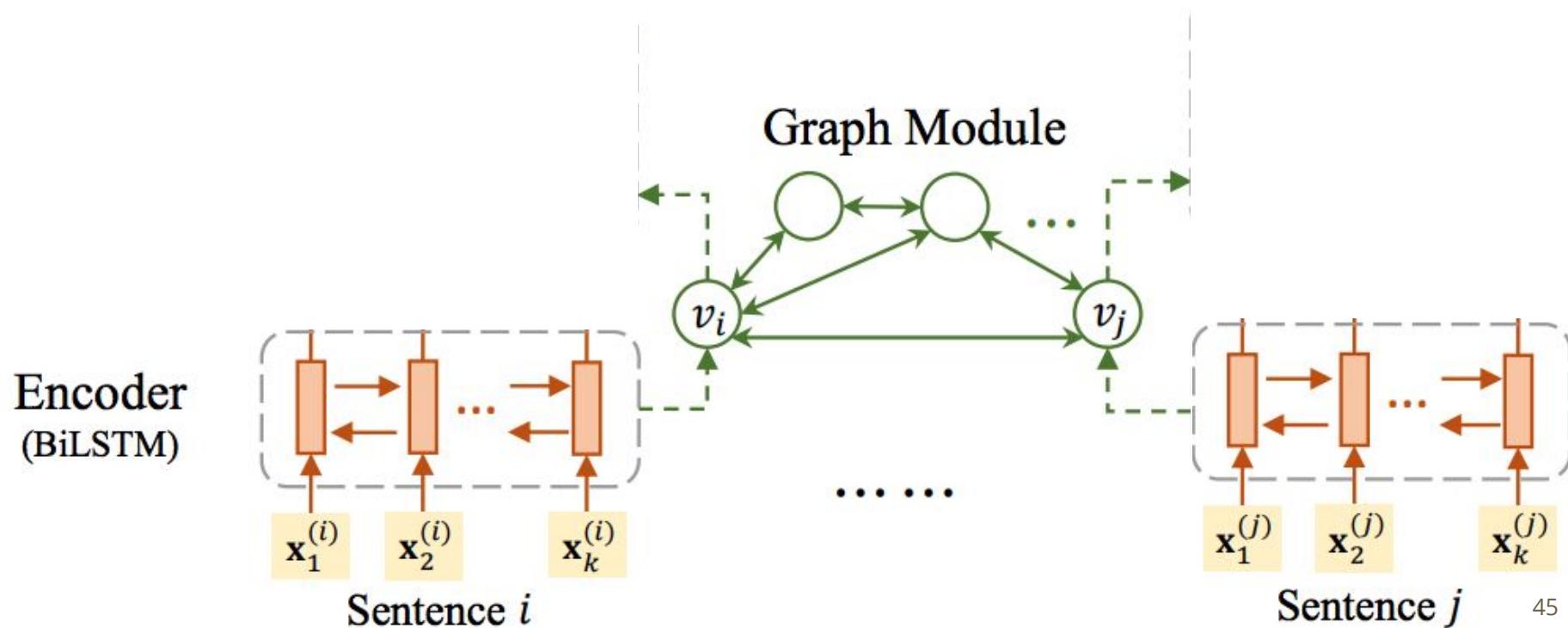


...



# GraphIE

Apply Graph Convolutional Network to page layout graph

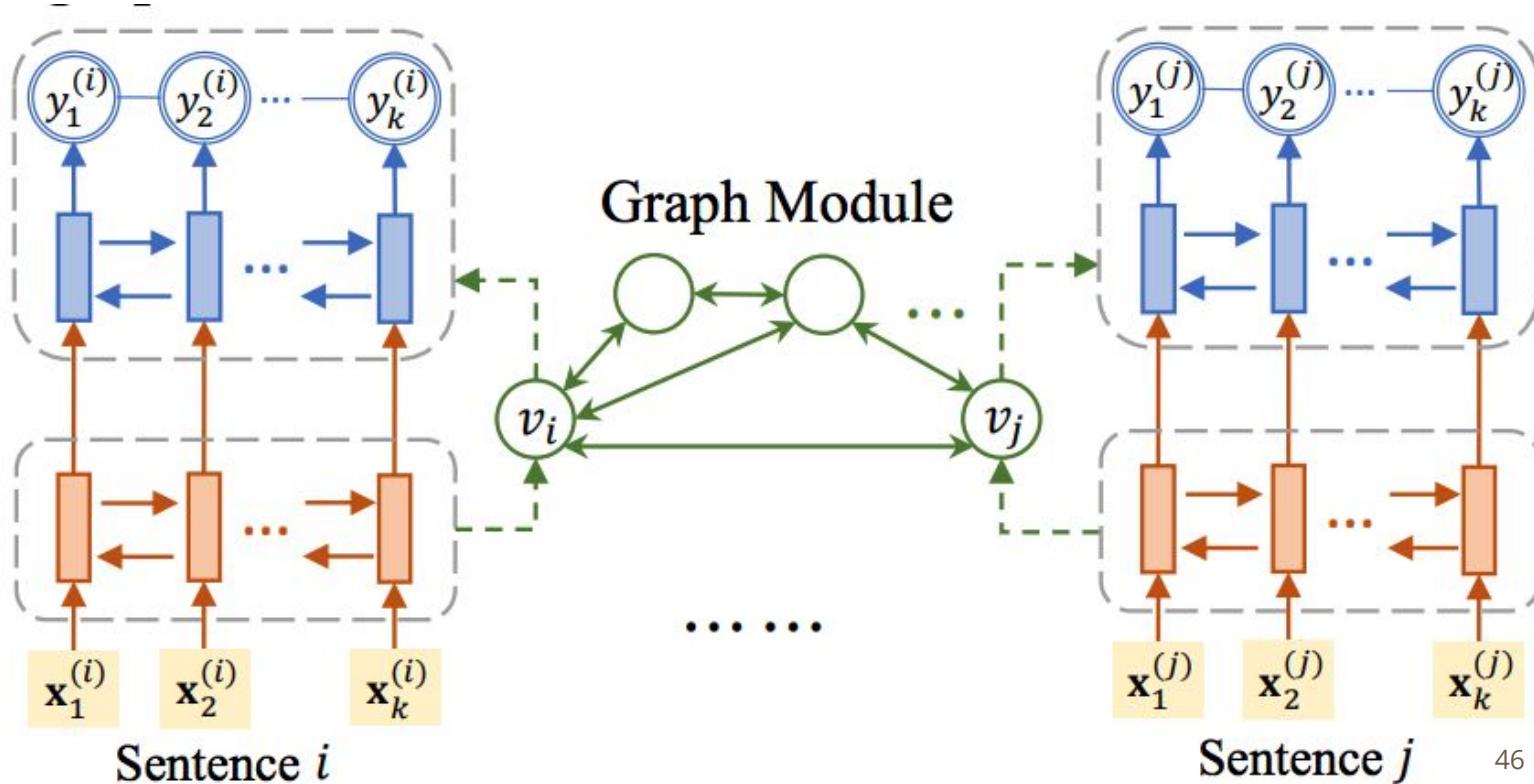


# GraphIE

Run NER-style LSTM model over sentence with graph representation as initial state

Decoder  
(BiLSTM + CRF)

Encoder  
(BiLSTM)



# GraphIE

ATTRIBUTE	SeqIE			GraphIE		
	P	R	F1	P	R	F1
<i>P. Initials</i>	93.5	92.4	<b>92.9</b>	93.6	91.9	92.8
<i>P. Age</i>	94.0	91.6	92.8	94.8	91.1	<b>92.9</b>
<i>P. Birthday</i>	96.6	96.0	<b>96.3</b>	96.9	94.7	95.8
<i>Drug Name</i>	71.2	51.2	59.4	78.5	50.4	<b>61.4</b>
<i>Event</i>	62.6	65.2	63.9	64.1	68.7	<b>66.3</b>
<i>R. First Name</i>	78.3	95.7	86.1	79.5	95.9	<b>86.9</b>
<i>R. Last Name</i>	84.5	68.4	75.6	85.6	68.2	<b>75.9</b>
<i>R. City</i>	88.9	65.4	75.4	92.1	66.3	<b>77.1</b>
Avg. (macro)	83.7	78.2	<b>80.3</b>	85.7	78.4	<b>81.1<sup>†</sup></b>
Avg. (micro)	78.5	73.8	76.1	80.3	74.6	<b>77.3<sup>†</sup></b>

Table 6: Extraction accuracy on the AEGR dataset (Task 3). Scores are the average of 5 runs. *P.* is the abbreviation for *Patient*, and *R.* for *Reporter*. <sup>†</sup> indicates statistical significance of the improvement over SeqIE ( $p < 0.05$ ).

On a dataset of medical PDFs, graph information adds about a point of F1 compared to an unstructured text extractor

# GraphIE

ATTRIBUTE	SeqIE			GraphIE		
	P	R	F1	P	R	F1
<i>P. Initials</i>	93.5	92.4	<b>92.9</b>	93.6	91.9	92.8
<i>P. Age</i>	94.0	91.6	92.8	94.8	91.1	<b>92.9</b>
<i>P. Birthday</i>	96.6	96.0	<b>96.3</b>	96.9	94.7	95.8
<i>Drug Name</i>	71.2	51.2	59.4	78.5	50.4	<b>61.4</b>
<i>Event</i>	62.6	65.2	63.9	64.1	68.7	<b>66.3</b>
<i>R. First Name</i>	78.3	95.7	86.1	79.5	95.9	<b>86.9</b>
<i>R. Last Name</i>	84.5	68.4	75.6	85.6	68.2	<b>75.9</b>
<i>R. City</i>	88.9	65.4	75.4	92.1	66.3	<b>77.1</b>
Avg. (macro)	83.7	78.2	80.3	85.7	78.4	<b>81.1<sup>†</sup></b>
Avg. (micro)	78.5	73.8	76.1	80.3	74.6	<b>77.3<sup>†</sup></b>

Table 6: Extraction accuracy on the AEGR dataset (Task 3). Scores are the average of 5 runs. *P.* is the abbreviation for *Patient*, and *R.* for *Reporter*. <sup>†</sup> indicates statistical significance of the improvement over SeqIE ( $p < 0.05$ ).

Layout graph wins big on unseen templates!  
(though overall numbers are still low)

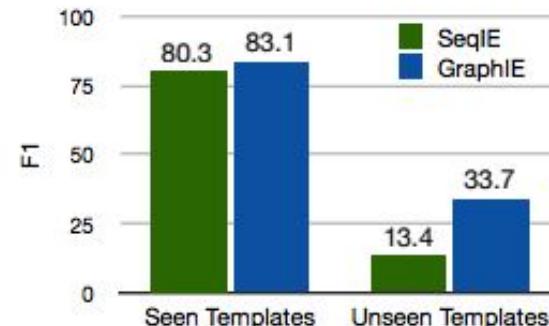


Figure 4: Micro average F1 scores tested on *seen* and *unseen* templates (Task 3).

# GraphIE

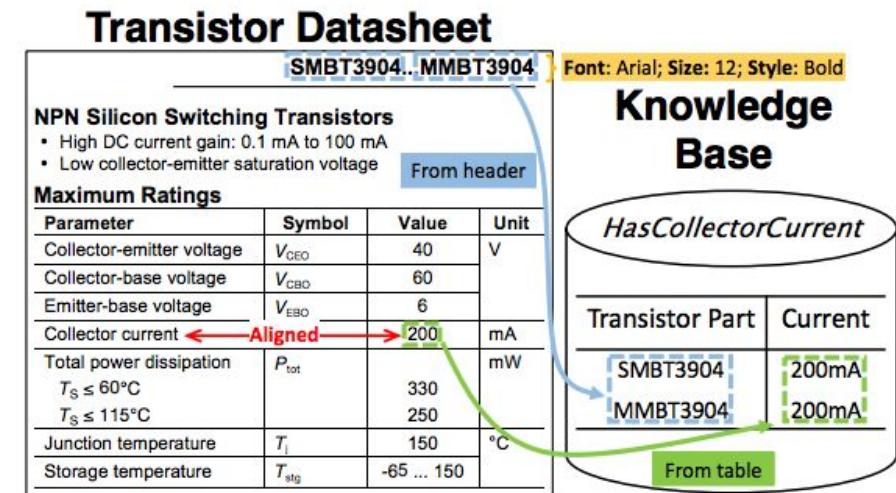
- Textual features propagated over page layout graph
- Pros:
  - Combines rich textual information with abstract template representation
- Cons:
  - Weak generalization to new templates
  - Uses layout relationship, but not other visual features
  - Requires defined ontology
  - Manually labeled training data

**How can the multi-modal setting help us with  
Data Programming?**

# Fonduer (Wu et al, 2018)

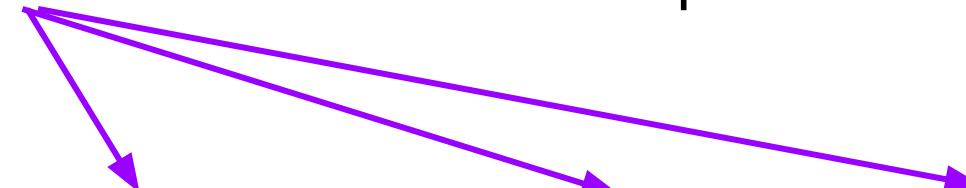
Extends Snorkel (Ratner et al, 2017) to focus on richly formatted documents

Extraction model uses multimodal features



# 3 labeling functions

Each is informative on different examples



(2014 Q2, 41520K)		[Three Months Ended June 30]	
	2015	2014	
Revenues			
Automotive	\$ 878,090	\$ 727,829	
Services and other	76,886	41,520	



(2015 Q2, 181712K)		[Three Months Ended June 30]	
	2015	2014	
Operating expenses			
Research and development	181,712	107,717	
Selling, general and administrative	201,846	134,031	



(2014 Q2, 247K)		[Three Months Ended June 30]	
	2015	2014	
Interest income	247	467	
Interest expense	(24,352)	(31,238)	
Other income (expense), net	13,233	(1,226)	



```
# Rule-based LF based on tabular content
def has_current_in_row(cand):
    if 'current' in row_ngrams(cand.current):
        return 1
    else:
        return 0
```

```
def value_in_column_header(cand):
    if 'Value' in header_ngrams(cand.current):
        return 1
    else:
        return 0
```

```
# Rule-based LF based on visual information
def y_axis_aligned(cand):
    return 1 if cand.part.y == cand.current.y else 0
```

```
# Rule-based LF based on tabular content
def has_current_in_row(cand):
    if 'current' in row_ngrams(cand.current):
        return 1
    else:
        return 0
```

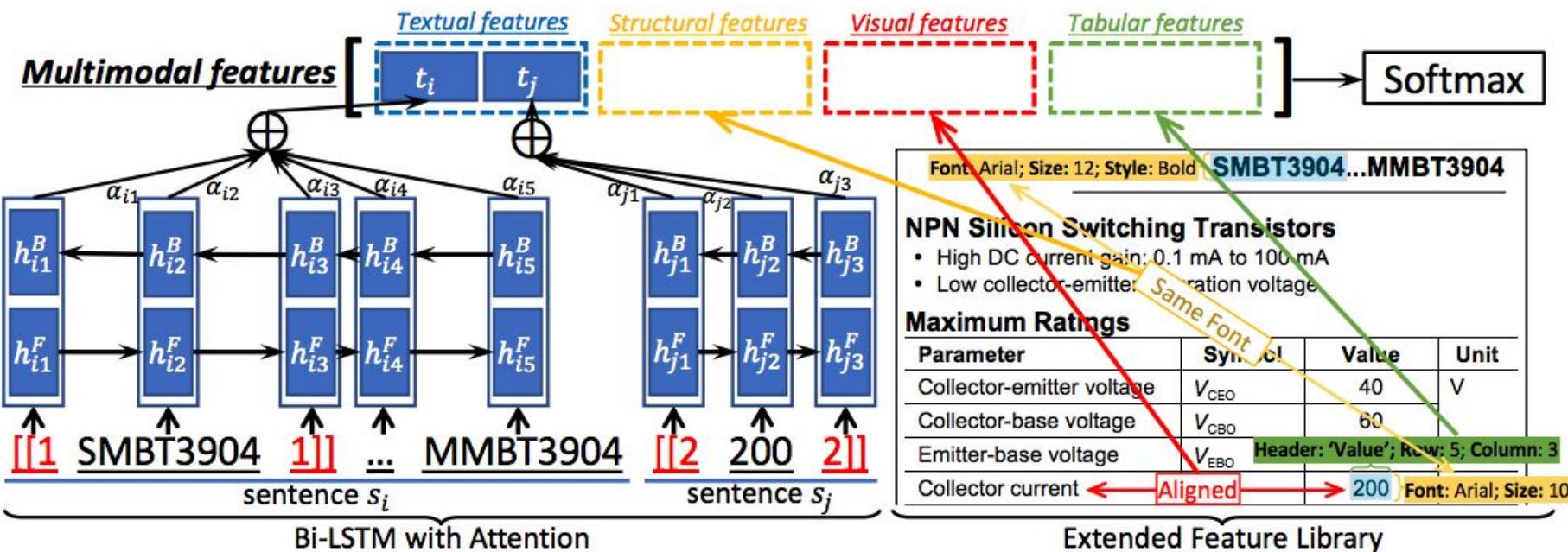
```
def value_in_column_header(cand):
    if 'Value' in header_ngrams(cand.current):
        return 1
    else:
        return 0
```

```
# Rule-based LF based on visual information
def y_axis_aligned(cand):
    return 1 if cand.part.y == cand.current.y else 0
```

```
# Rule-based LF based on tabular content
def has_current_in_row(cand):
    if 'current' in row_ngrams(cand.current):
        return 1
    else:
        return 0
```

```
def value_in_column_header(cand):
    if 'Value' in header_ngrams(cand.current):
        return 1
    else:
        return 0
```

```
# Rule-based LF based on visual information
def y_axis_aligned(cand):
    return 1 if cand.part.y == cand.current.y else 0
```



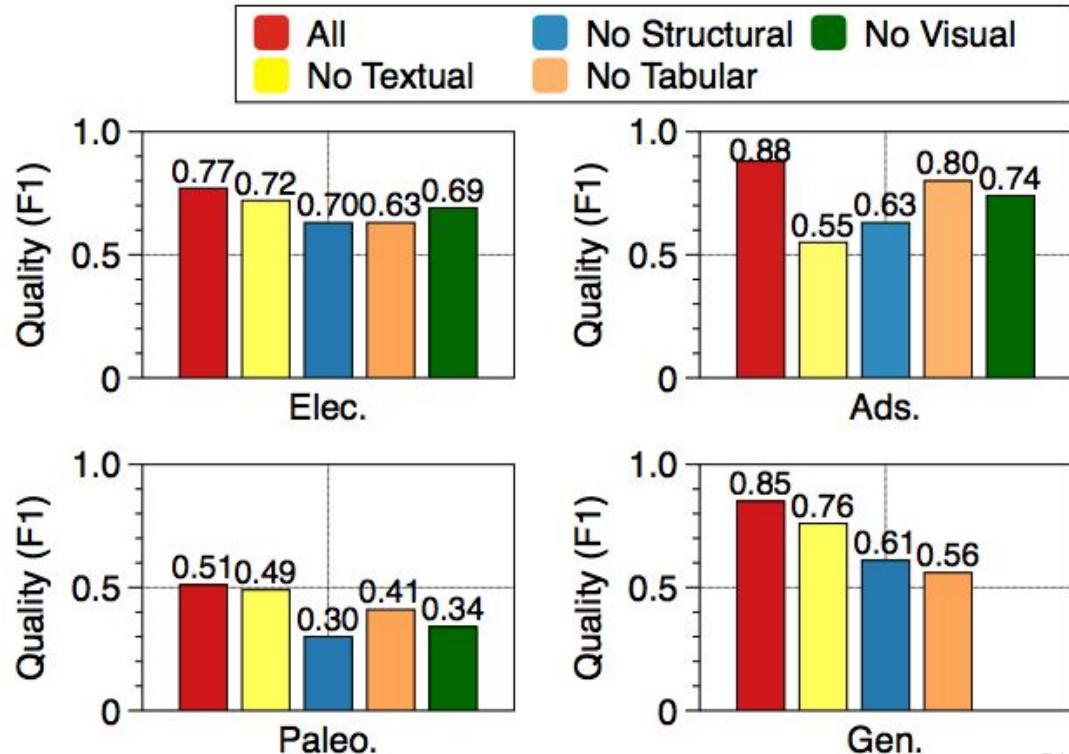
# Fonduer

Sys.	Metric	Text	Table	Ensemble	Fonduer
ELEC.	Prec.	1.00	1.00	1.00	0.73
	Rec.	0.03	0.20	0.21	0.81
	F1	0.06	0.40	0.42	<b>0.77</b>
ADS.	Prec.	1.00	1.00	1.00	0.87
	Rec.	0.44	0.37	0.76	0.89
	F1	0.61	0.54	0.86	<b>0.88</b>
PALEO.	Prec.	0.00	1.00	1.00	0.72
	Rec.	0.00	0.04	0.04	0.38
	F1	0.00*	0.08	0.08	<b>0.51</b>
GEN.	Prec.	0.00	0.00	0.00	0.89
	Rec.	0.00	0.00	0.00	0.81
	F1	0.00 <sup>#</sup>	0.00 <sup>#</sup>	0.00 <sup>#</sup>	<b>0.85</b>

Huge gains in recall  
with small loss of  
precision

# Fonduer

Different datasets benefit from different features



# Fonduer

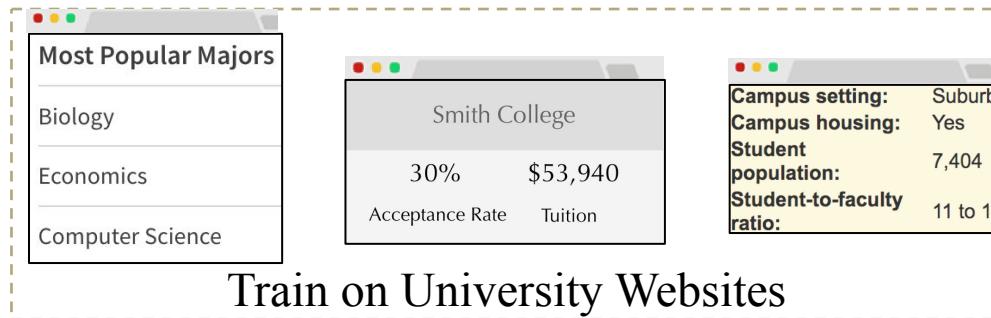
- Cheaply create training data for multi-modal extraction
- Pros:
  - Good accuracy for low price
  - Multi-modal labeling functions
  - Combines all textual modalities
- Cons:
  - Requires manual work for each subject domain
  - Requires ontology

**How can the multi-modal setting help us with  
OpenIE?**

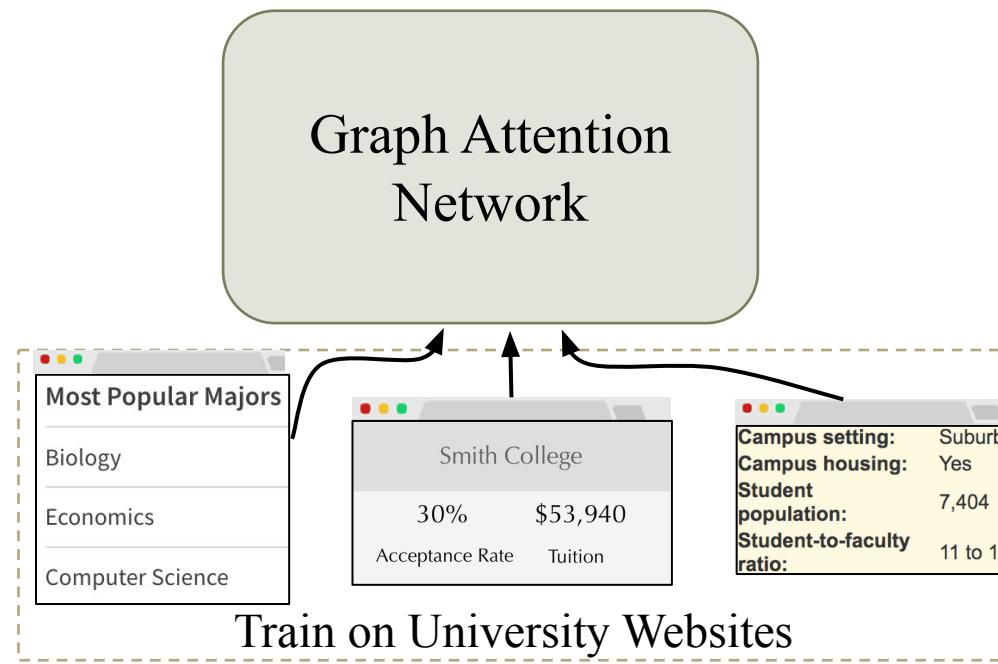
# ZeroShotCeres (Lockard et al, 2020)

- Page layout graph similar to GraphIE
  - Also includes DOM relationships
- OpenIE: Extracts predicates and objects
- Zero-shot generalization to unseen templates
- Zero-shot generalization to unseen subject domains

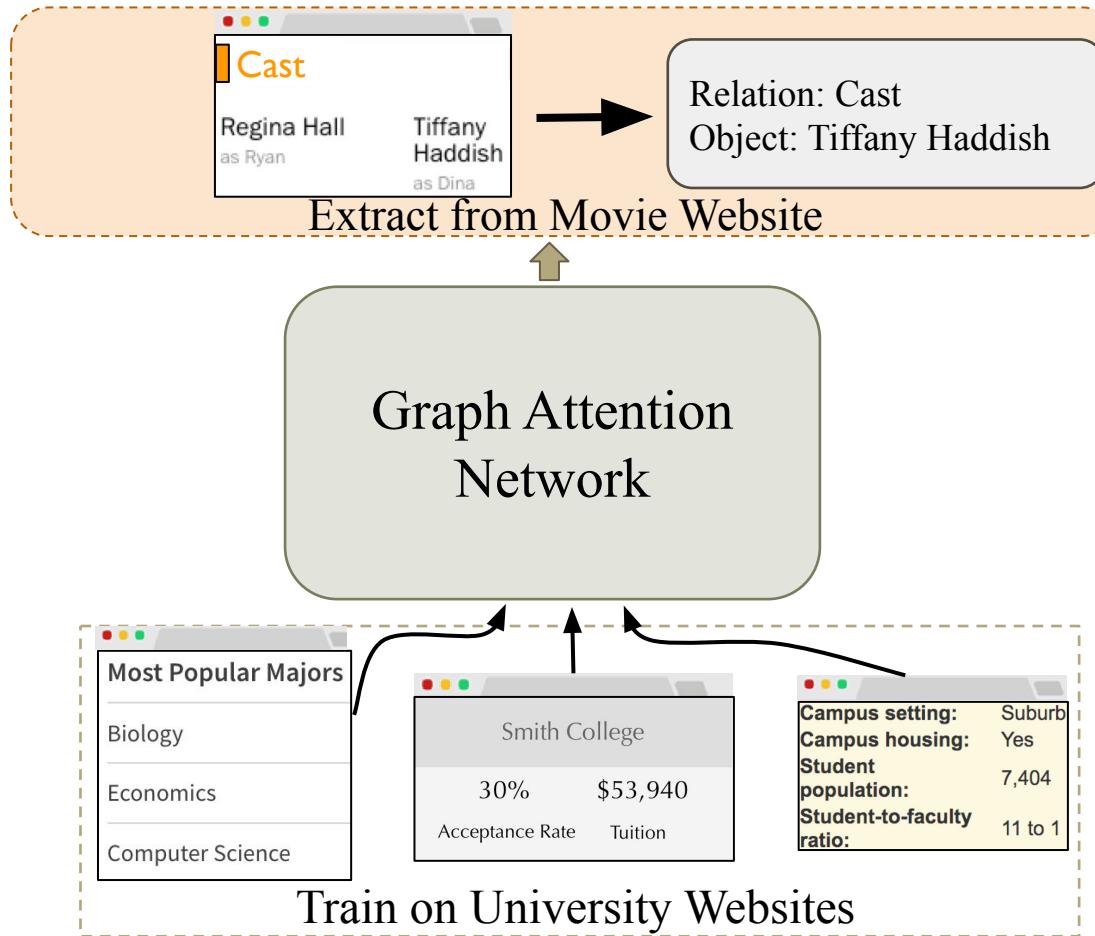
# Zero-shot OpenIE



# Zero-shot OpenIE



# Zero-shot OpenIE

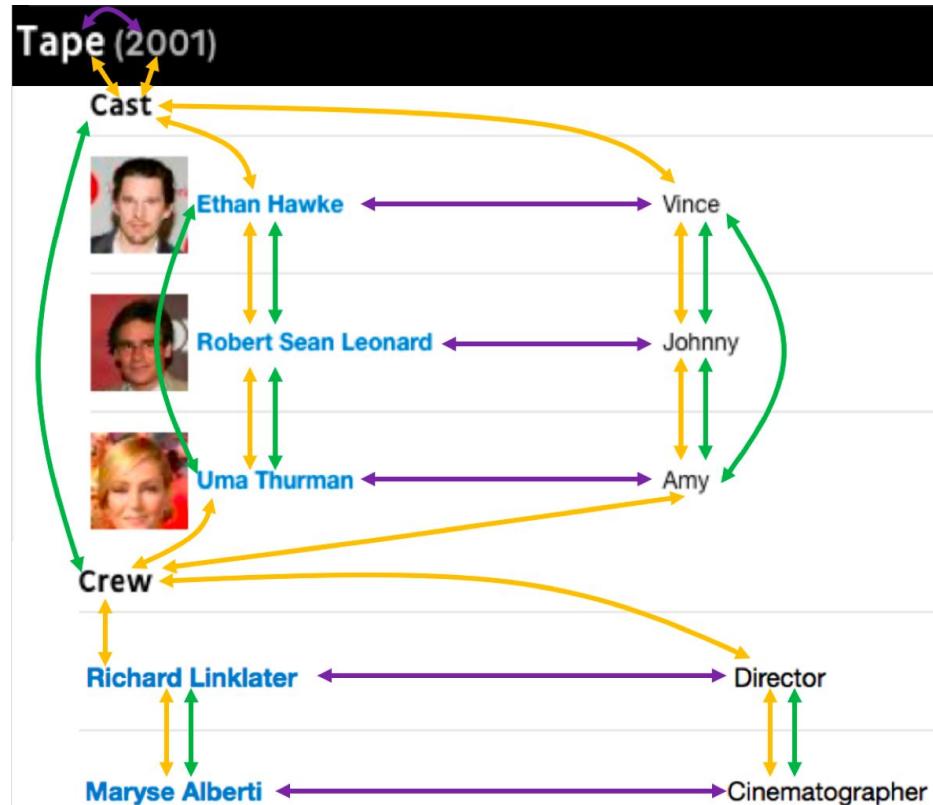


# ZeroShotCeres

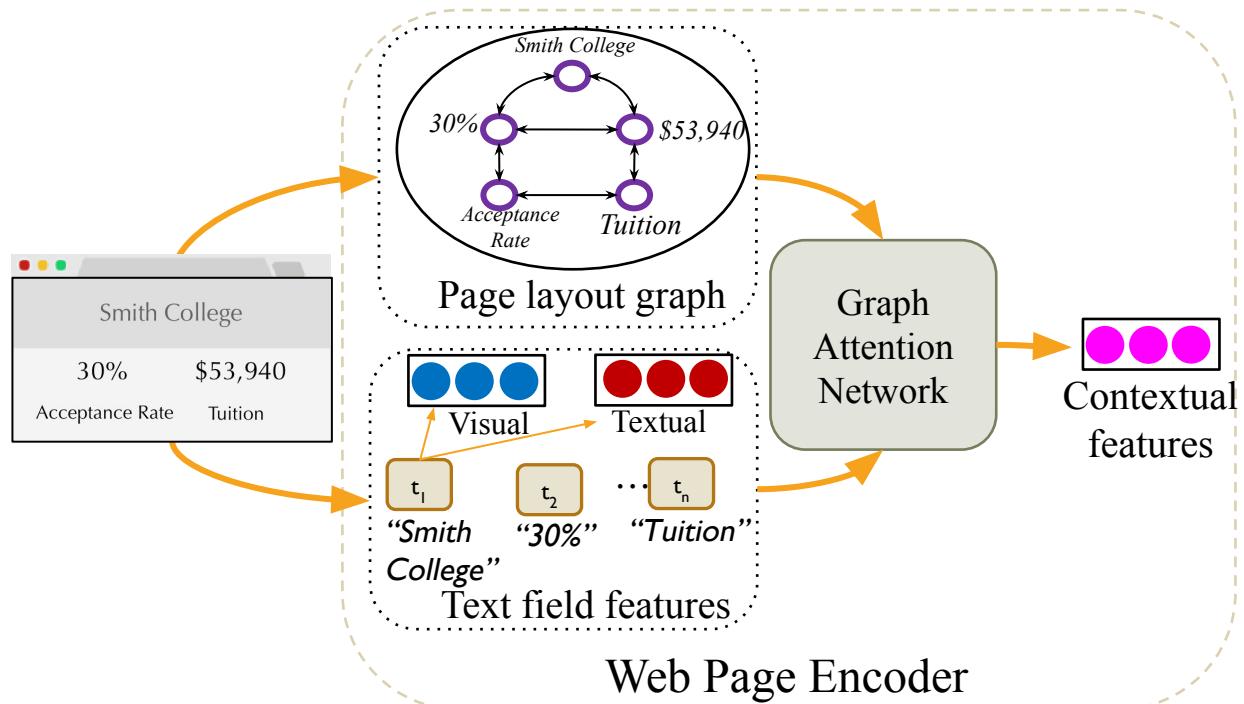
Horizontal edges

Vertical edges

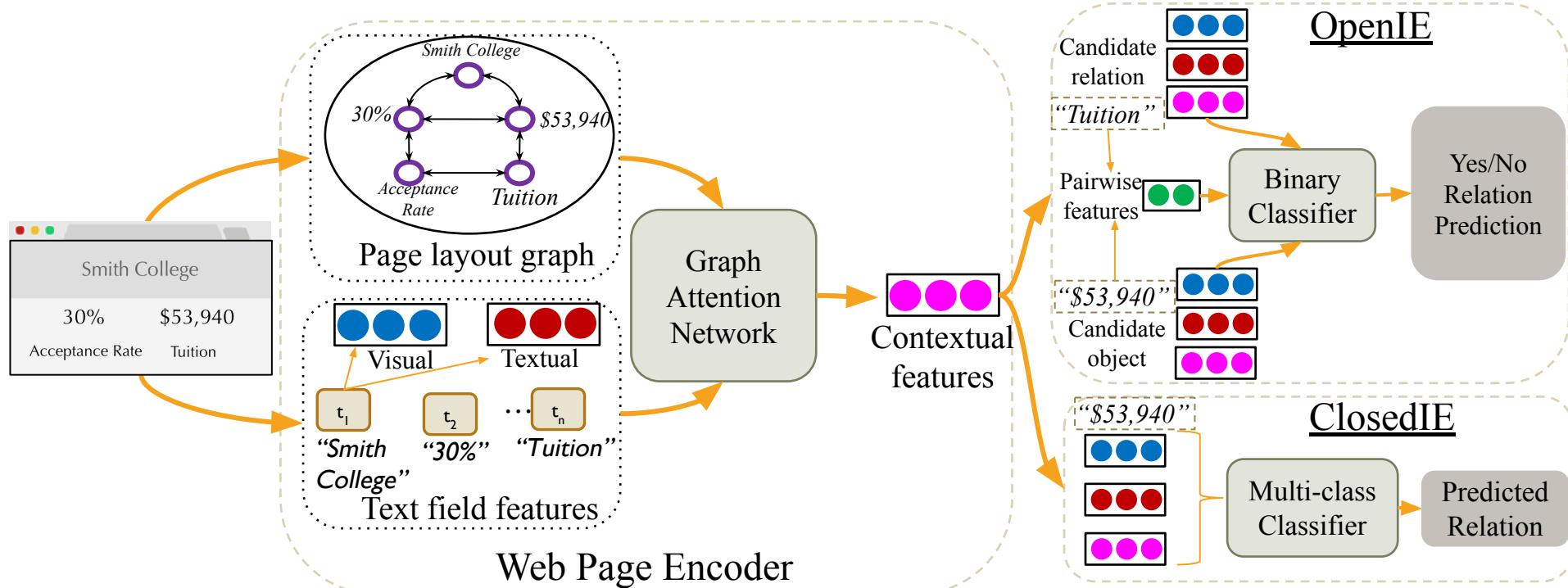
DOM edges connect nodes  
that are siblings/cousins in  
DOM tree



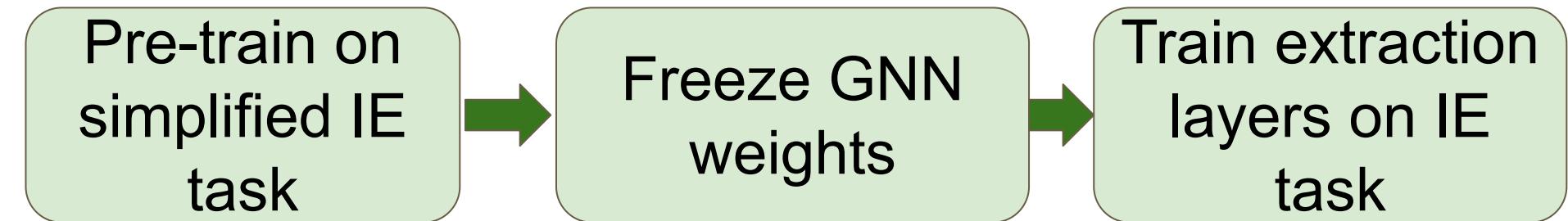
# ZeroShotCeres



# ZeroShotCeres



# ZeroShotCeres Training



3-way classification in  
{relation string, object  
string, other}

# OpenIE Evaluation

Method	Movie	NBA	University	Average
Unsupervised Baseline	0.27	0.40	0.37	0.35
Site-specific Knowledge Baseline (OpenCeres)	0.77	0.58	0.40	0.58
ZeroShotCeres	0.42	0.48	0.47	0.46

OpenIE training on 2 subject domains  
Extract from 3rd (unseen) domain

# OpenIE Evaluation

Method	Movie	NBA	University	Average
Unsupervised Baseline	0.27	0.40	0.37	0.35
Site-specific Knowledge Baseline (OpenCeres)	0.77	0.58	0.40	0.58
ZeroShotCeres	0.42	0.48	0.47	0.46

Without pretraining, 5 point drop in F1

# ZeroShotCeres Overview

- OpenIE on zero-shot websites and subject domains
- Pros:
  - Learns layout/visual semantics of key-value relationships
- Cons:
  - Still room for improvement in accuracy

# State of the art for multi-modal text extraction

Method	Extraction Type	Supervision	Requires ontology	Features	Model type
Bling-KPE	Single Span	Weak Supervision	N	Text, position, font visuals	Transformer
CharGrid	Grouped spans	Supervised	Y	Character-aligned pixel map	CNN
GraphIE	Single span	Supervised	Y	Text, layout graph	GNN
Fonduer	Single span	Weak Supervision	Y	Text, DOM, font visuals, table location	LSTM
ZeroShotCeres	Span pairs	Supervised	N	Text, layout graph, font visuals	GNN

# Short answers

- **Diversity**
  - Textual, layout, and visual signals can combine to form consistent patterns
- **Training data**
  - Multi-modal signals allow for accurate and easy creation of training data with Data Programming
- **OpenIE**
  - Visual semantics help make OpenIE extractions from semi-structured documents without prior knowledge of the subject domain

# New work at ACL 2020!

## Representation Learning for Information Extraction from Form-like Documents

Bodhisattwa Prasad Majumder<sup>†♣</sup> Navneet Potti<sup>♣</sup> Sandeep Tata<sup>♣</sup>  
James B. Wendt<sup>♣</sup> Qi Zhao<sup>♣</sup> Marc Najork<sup>♣</sup>

<sup>♣</sup>Department of Computer Science and Engineering, UC San Diego

bmajumde@eng.ucsd.edu

<sup>♣</sup>Google Research, Mountain View

{navsan, tata, jwendt, zhaqi,  
najork}@google.com

### TAPAS: Weakly Supervised Table Parsing via Pre-training

Jonathan Herzig<sup>1,2</sup>, Paweł Krzysztof Nowak<sup>1</sup>, Thomas Müller<sup>1</sup>,  
Francesco Piccinno<sup>1</sup>, Julian Martin Eisenschlos<sup>1</sup>

<sup>1</sup>Google Research

<sup>2</sup>School of Computer Science, Tel-Aviv University

{jherzig, pawelnow, thomasmueller, piccinno, eisenjulian}@google.com

### INFOTABS: Inference on Tables as Semi-structured Data

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, Vivek Srikumar

School of Computing, University of Utah

{vgupta, maitrey, pnokhiz, svivek}@cs.utah.edu

# References

- Ibrahim, Yusra, Mirek Riedewald, Gerhard Weikum and Demetrios Zeinalipour-Yazti. "Bridging Quantities in Tables and Text." *ICDE* (2019): 1010-1021.
- Katti, Anoop R., Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne and Jean Baptiste Faddoul. "Chagrid: Towards Understanding 2D Documents." *EMNLP* (2018).
- Lockard, Colin, Prashant Shiralkar, Xin Dong and Hannaneh Hajishirzi. "ZeroShotCeres: Zero-Shot Relation Extraction from Semi-Structured Webpages." *ACL* (2020).
- Qian, Yujie, Enrico Santus, Zhijing Jin, Jiang Guo and Regina Barzilay. "GraphIE: A Graph-Based Framework for Information Extraction." *NAACL-HLT* (2019).

# References

Ratner, Alexander, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu and Christopher Ré. "Snorkel: Rapid Training Data Creation with Weak Supervision." *PVLDB* 11 3 (2017): 269-282 .

Xiong, Lee, Chuan Hu, Chenyan Xiong, Daniel Campos and Arnold Overwijk. "Open Domain Web Keyphrase Extraction Beyond Language Modeling." *EMNLP/IJCNLP* (2019).

Wu, Sen, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas, Philip Levis and Christopher Ré. "Fonduer: Knowledge Base Construction from Richly Formatted Data." Proceedings. *SIGMOD* 2018 (2018): 1301-1316 .

# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- Part 1b: Unstructured Text: Methods (10 minutes)
- Live Q&A (15 minutes)
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- Part 3: Multi-modal Extraction and Conclusion (35 minutes)
- Live Q&A (15 minutes)

---

---

# Conclusion

---

---

— Colin Lockard, Prashant Shiralkar,  
Xin Luna Dong, Hannaneh Hajishirzi —



# Four Challenges

1. Diversity of data
2. Multiple modalities of text
3. Lack of training data
4. Unknown unknowns

Can we build a single extractor to find **consistent signals** across these diverse elements of data **across all modalities of text?**

# Key Intuitions

- Diversity: Identifying consistent patterns
  - Leverage consistency in model/representation
  - Combining information from multiple modalities can give more consistent signals
- Lack of training data: Learning with limited labels
  - Find automated ways to label data
  - Employ weak or semi-supervision in limited labeled data settings
- Unknown unknowns: Stay open--Sacrificing granularity of knowledge representation allows for easier scaling

# Unstructured Text: Short Answers

- **Consistency**
  - Model problem as text span classification and relationships between spans
  - Word embedding models help capture text semantics
- **Training data**
  - Weak supervision gives cheap training data
- **OpenIE**
  - Discovery of new types and relationships

# Semi-Structured Text: Short Answers

- **Consistency**
  - Leverage general key-value pair consistency universal in templates
  - Leverage site-level consistency in layout and presentation
- **Training data**
  - Use distant supervision to generate cheap, but noisy training data
- **OpenIE**
  - Discover new relations by label propagation

# Tabular text - Short Answers

- **Subject column detection**
  - Leverage generic features of subject entities such as value uniqueness, string type, number of characters and words
- **Column class detection**
  - Leverage external data -- web extracted triples, knowledge graph
- **Relation extraction between column pair**
  - Measure similarity between a column and entities of a type in a knowledge base

# Multi-modal extraction: Short answers

- **Diversity**
  - Textual, layout, and visual signals can combine to form consistent patterns
- **Training data**
  - Multi-modal signals allow for accurate and easy creation of training data with Data Programming
- **OpenIE**
  - Visual semantics help make OpenIE extractions from semi-structured documents without prior knowledge of the subject domain

# Future Directions - Unstructured text

- Full document understanding
  - Relation extraction beyond single sentence/paragraph
- Faster embedding models for scalability
- Non-English languages

# Future Directions - Semi-structured text

- N-ary relations
- Relations not involving page topic

# Future Directions - Tabular text

- Direct extraction (not relying on existing knowledge)

# Future Directions - Multi-modal extraction

- Combine all signals from a document
- Make use of images
- Operate from jpgs, scanned pdfs
- Pre-training webpage representations
- Automated ontology construction
- Reproducible research
  - Webpage visual features depend on browser, CSS/JS availability, etc.

# Outline

- Introduction (40 minutes)
- Part 1a: Unstructured Text (25 minutes)
- Part 1b: Unstructured Text: Methods (10 minutes)
- Live Q&A (15 minutes)
- Break (30 minutes)
- Part 2: Semi-structured and Tabular Text (40 minutes)
- Part 3: Multi-modal Extraction and Conclusion (35 minutes)
- **Live Q&A (15 minutes)**

# Thank you!

[https://sites.google.com/view/  
acl-2020-multi-modal-ie](https://sites.google.com/view/acl-2020-multi-modal-ie)

Please join us in  
the Zoom Chat!

---