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

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

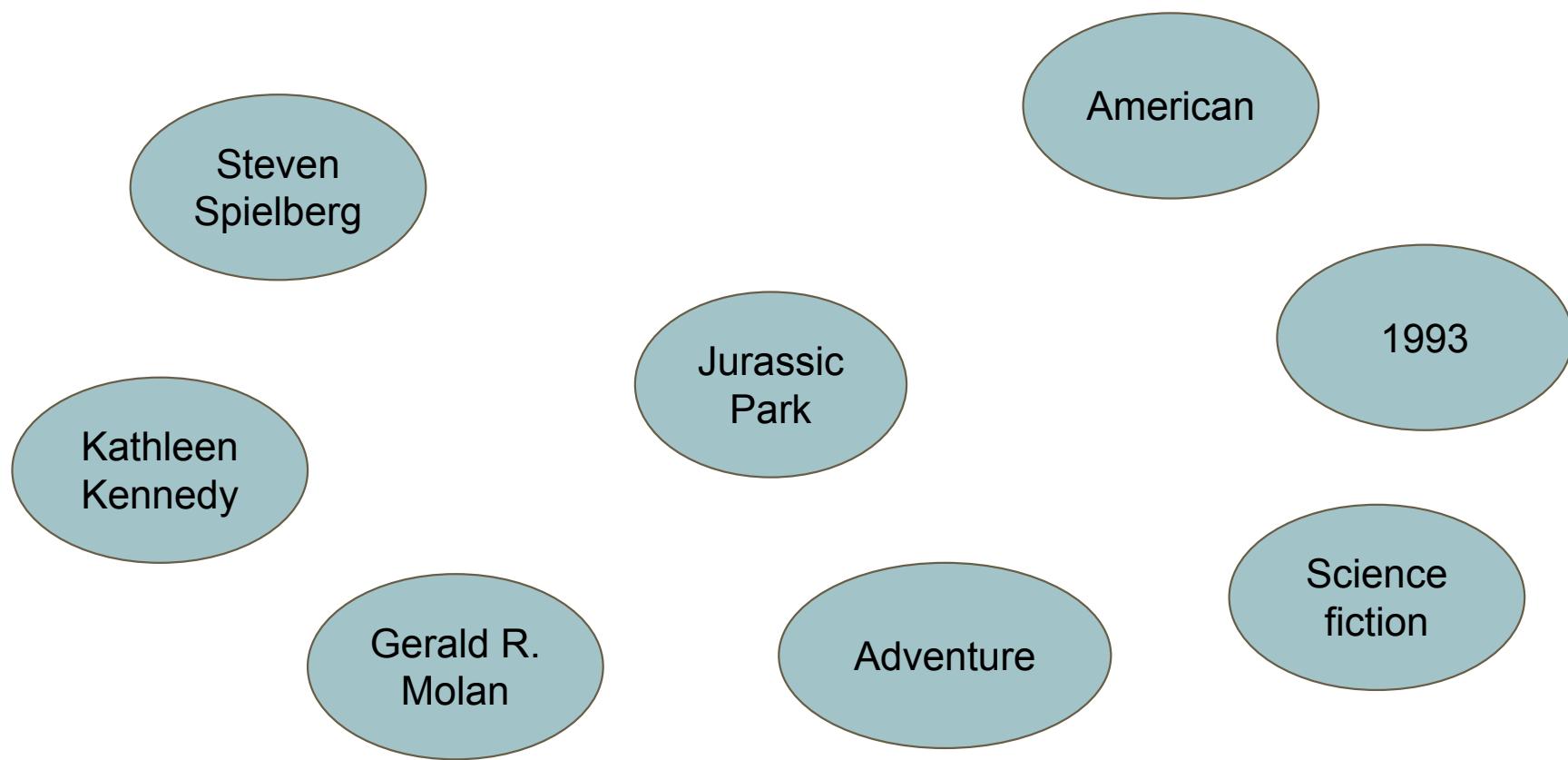
<https://sites.google.com/view/acl-2020-multi-modal-ie>



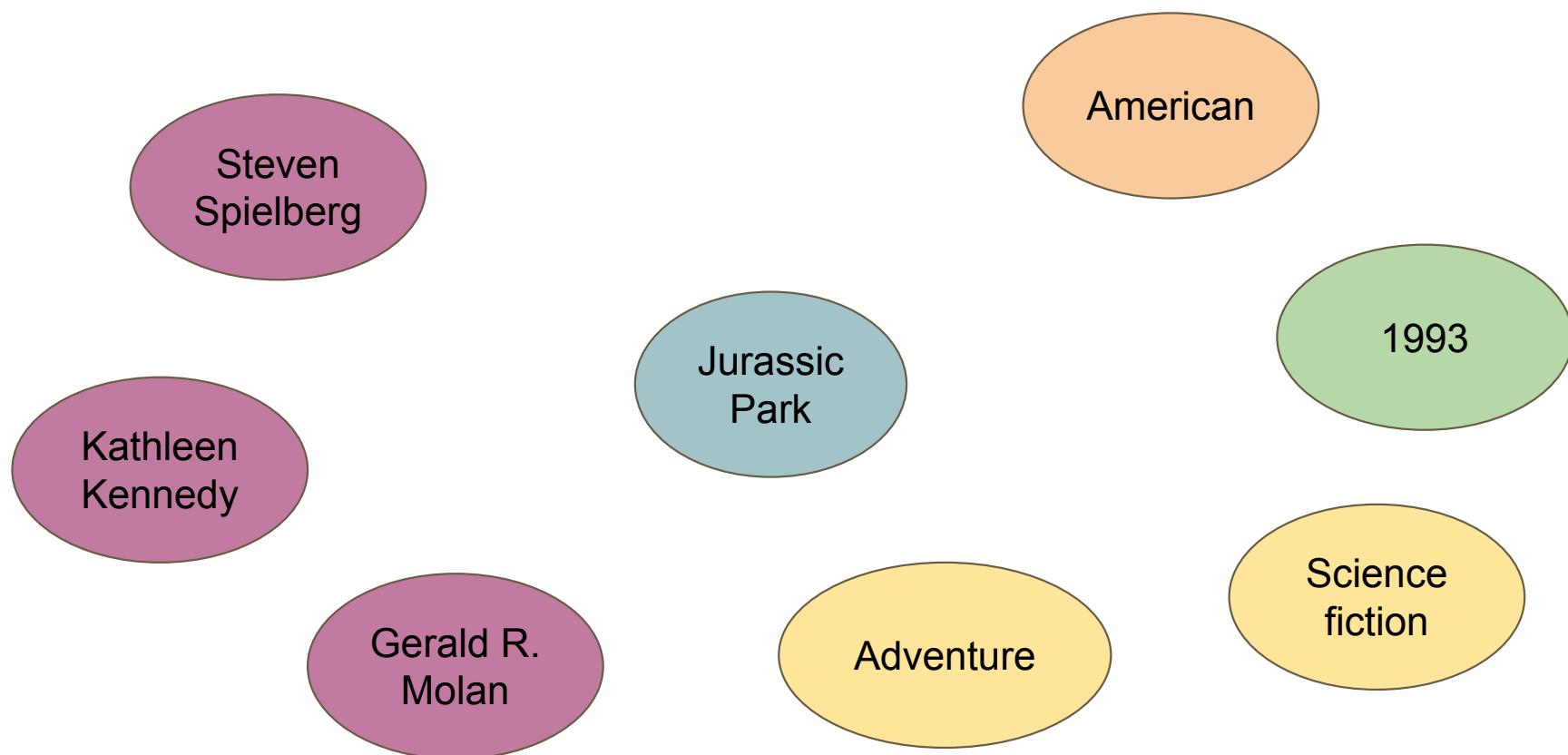
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

What is Knowledge Graph?

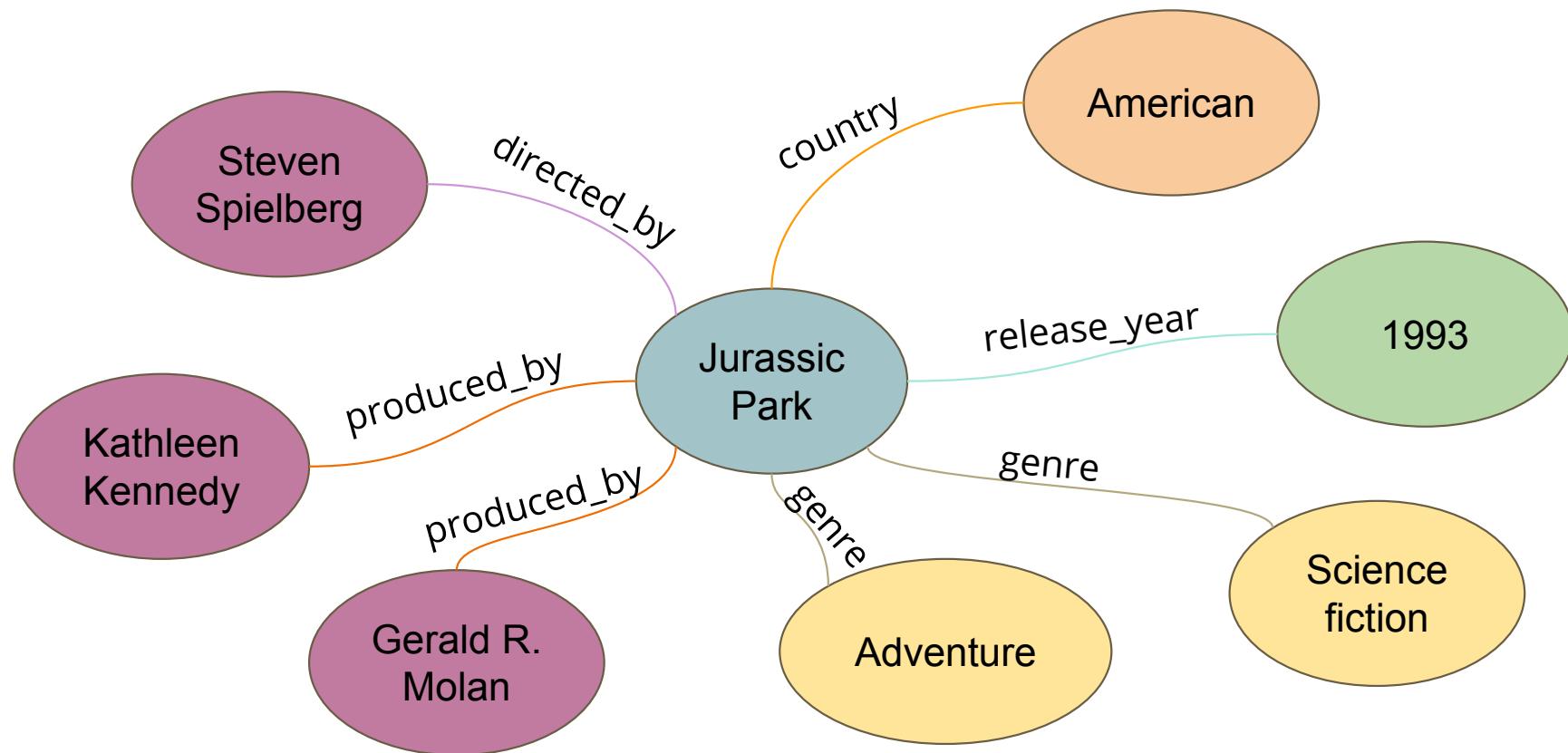
Knowledge graph: entities and relationships



Knowledge graph: entities and relationships



Knowledge graph: entities and relationships



Application 1. Web search

bruce croft

All Videos News Images Shopping More Settings Tools

About 13,700,000 results (0.59 seconds)

ciir.cs.umass.edu › croft ▾
[W. Bruce Croft | Center for Intelligent Information Retrieval ...](#)
W. Bruce Croft Distinguished University Professor, College of Information and Computer Sciences, and Director, Center for Intelligent Information Retrieval.

scholar.google.com › citations ▾
[W. Bruce Croft - Google Scholar Citations](#)
W. Bruce Croft. Distinguished Professor of ... JM Ponte, WB Croft. Proceedings of the 21st annual ... WB Croft, D Metzler, T Strohman. Addison-Wesley, 2010.

en.wikipedia.org › wiki › W._Bruce_Croft ▾
[W. Bruce Croft - Wikipedia](#)
W. Bruce Croft is a distinguished professor of computer science at the University of Massachusetts Amherst whose work focuses on information retrieval. He is the founder of the Center for Intelligent Information Retrieval and served as the editor-in-chief of ACM Transactions on Information Systems from 1995 to 2002.

W. Bruce Croft



Professor

W. Bruce Croft is a distinguished professor of computer science at the University of Massachusetts Amherst whose work focuses on information retrieval. He is the founder of the Center for Intelligent Information Retrieval and served as the editor-in-chief of ACM Transactions on Information Systems from 1995 to 2002. [Wikipedia](#)

h-index: 105

Affiliation: University of Massachusetts, Amherst

Books: [Search Engines: Information Retrieval in Practice](#)

Education: [University of Cambridge](#), [Monash University](#)

Awards: [ACM Fellow](#)

Application 2: Question answering

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



Speakers

Keynote Speakers

Bin Yu
University of California, Berkeley
Veridical Data Science
Tuesday, Feb 4th

Bin Yu is Chancellor's Professor in the Departments of Statistics and of Electrical Engineering & Computer Sciences at the University of California at Berkeley and a former chair of Statistics at UC Berkeley. Her research focuses on practice,

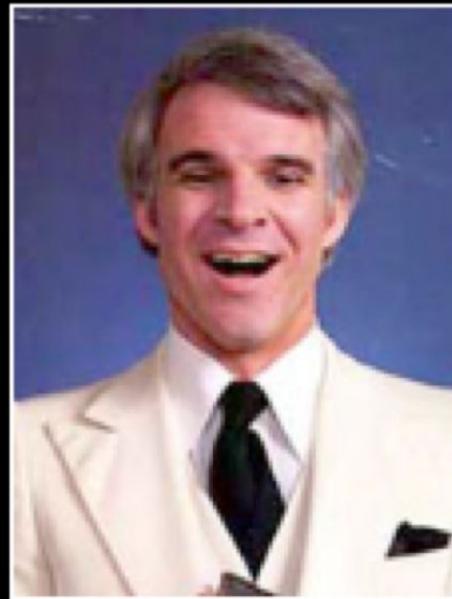
This year's keynotes are from Bin Yu, Ed H. Chi, Kristen Grauman...



Application 3: Recommendation

Discover similarities between entities





Steve Martin

[Main](#)[Jokes](#)[Reviews](#)[Video](#)[Tour Dates](#)

Born: August 14, 1945

Blue Meter: Tame

43 W

Like this comedian? ...



Michelle Wolf

Main

Jokes

Reviews

Video

Tour Dates

Born: June 21, ?????

Blue Meter: Risqué

1 1

Like this comedian?

WHO



Tracy Morgan

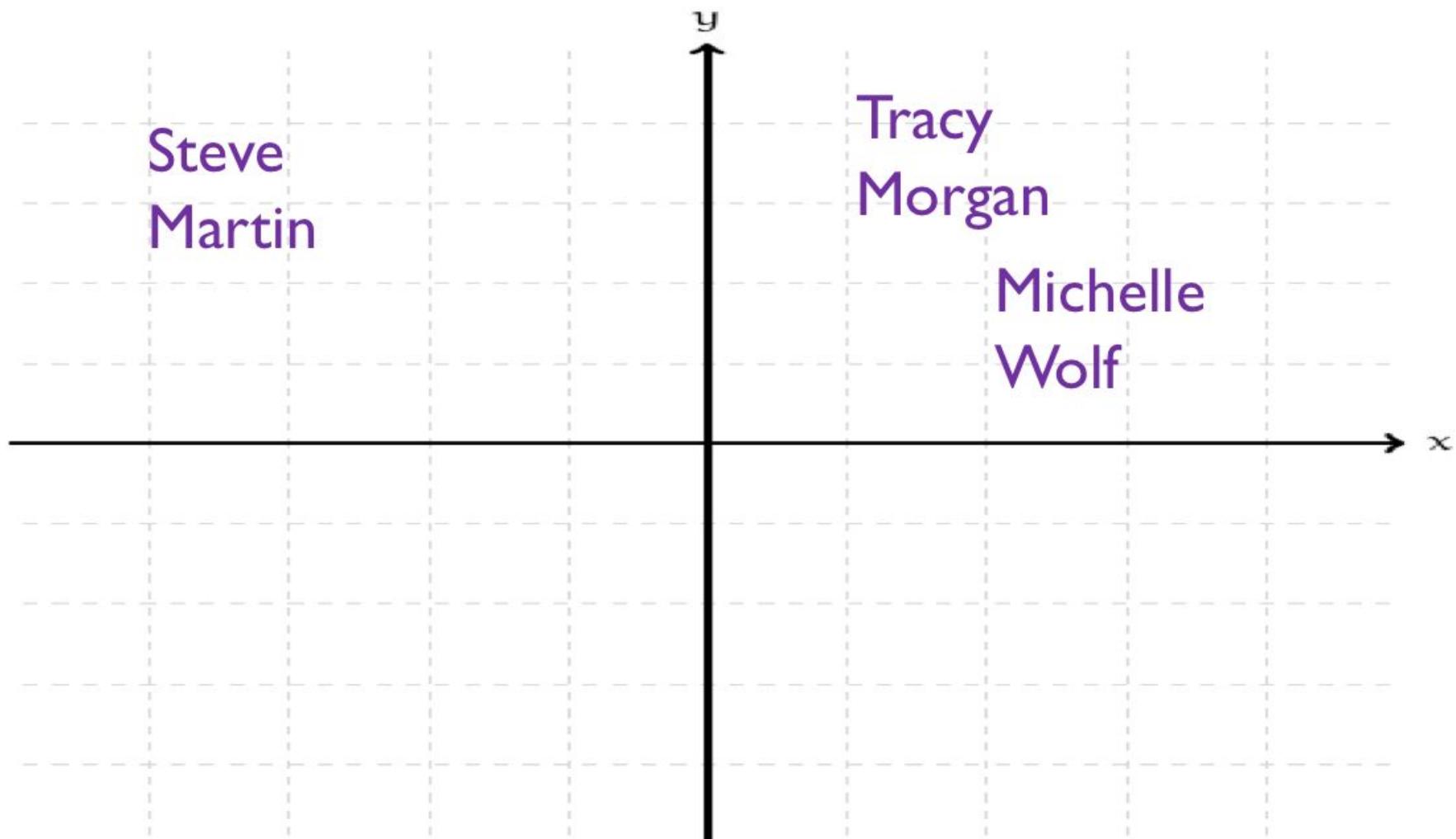
[Main](#)[Jokes](#)[Reviews](#)[Video](#)[Tour](#)

Born: November 10, 1968

Blue Meter: Risqué

13

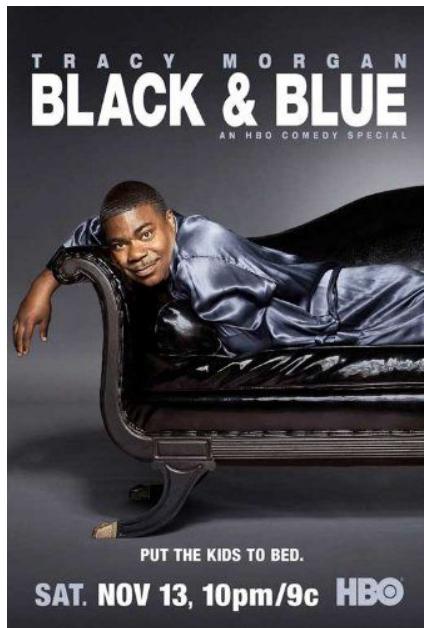
I like this comic



Better embeddings

= Better cold start recommendations

Because you watched...



you might like...



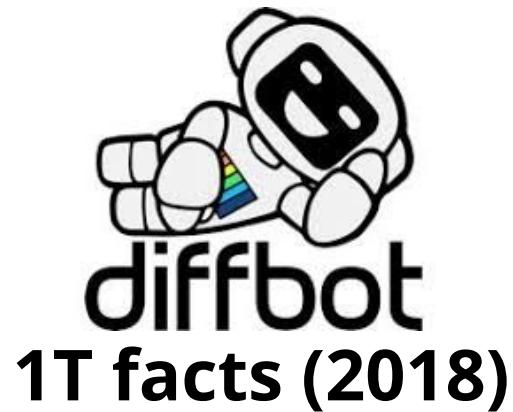
Industry knowledge graphs



70B facts (2016)

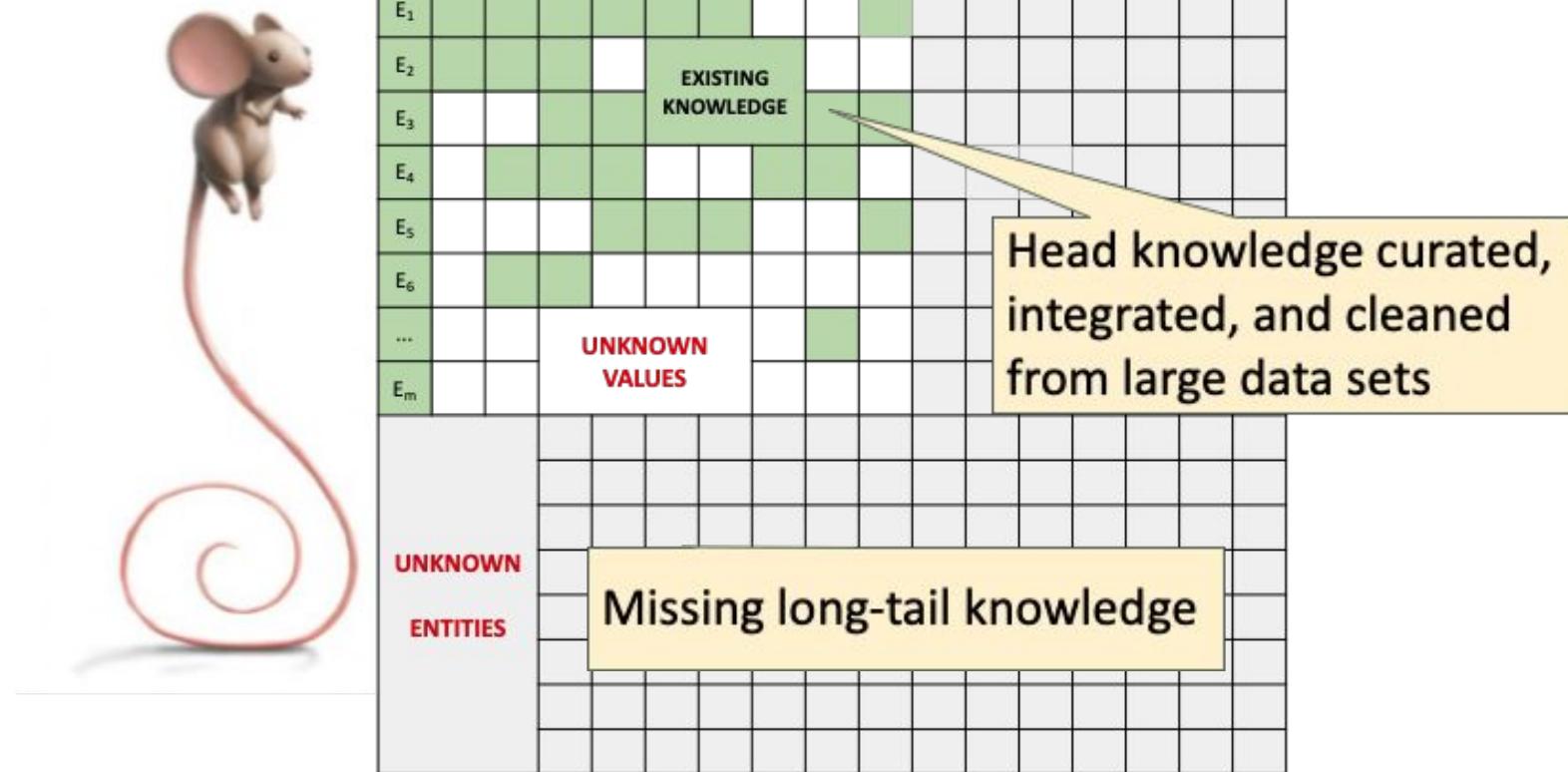


50B facts (2018)



Why Web-Scale Knowledge Collection?

Still Missing A Lot of Long-Tail Knowledge



Still Missing A Lot of Long-Tail Knowledge

	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A _n	UNKNOWN ATTRIBUTES
E ₁										
E ₂										
E ₃										
E ₄										
E ₅										
E ₆										
...										
E _m										
UNKNOWN VALUES										
UNKNOWN ENTITIES										

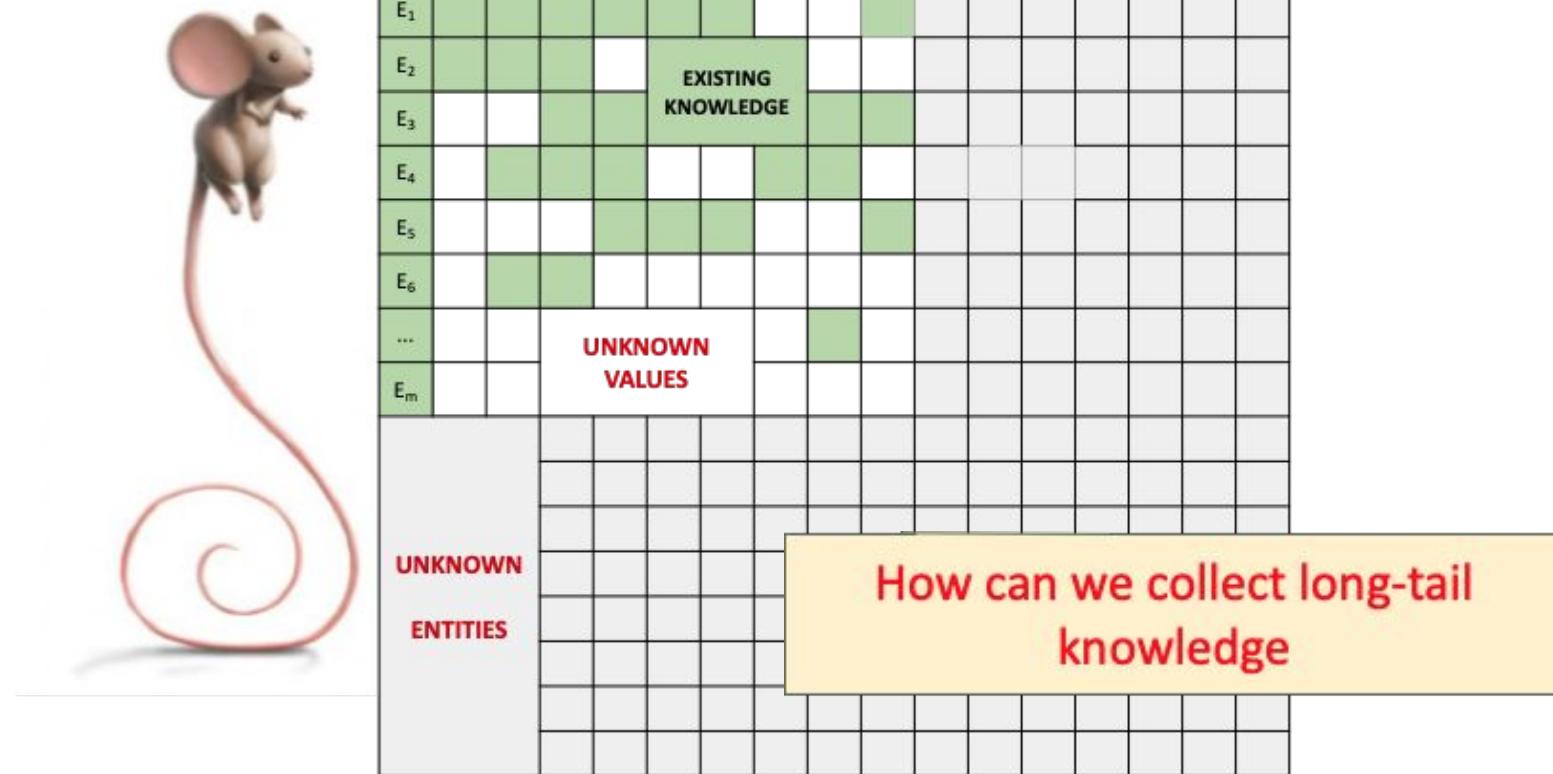
- Alexa, when did Van Gogh live in Paris?
- Sorry, I'm not sure.

- 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
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 Tournaments

Year: 2020

2020 Stats

Prize Money	Singles Titles
\$46,600	1

2020 Tournaments

- *AUSTRALIAN OPEN - Melbourne, Australia
January 19, 2020 to February 1, 2020
- ASB Classic 2020 - Auckland, New Zealand
January 5, 2020 to January 11, 2020

ROUND OPPONENT

Round	Opponent
1st	Anastasia Potapova
2nd	Tamara Zidansek

Tools

- What links here
- Related changes
- Upload file
- Special pages
- Permanent link
- Page information
- Wikidata item
- Cite this page

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

Ada Lovelace

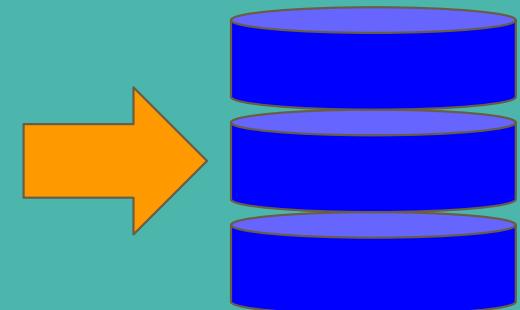
From Wikipedia, the free encyclopedia

Augusta Ada King, Countess of Lovelace (née Byron; 10 December 1815 – 27 November 1852) was a mathematician 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, and thus is considered to be the first to recognise the full potential of a "computing machine" and one of the first to recognise the full potential of a "computing machine" and one of the first to recognise the full potential of a "computing machine".

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





57th

Tg: Storytelling | Advances in NLP | Learning
ACL 2019 ANNUAL MEETING July 28th - August 2nd
Florence of the Association for Computational Linguistics

[f FACEBOOK](#)

[t TWITTER](#)

[i INSTAGRAM](#)

[c 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

📍 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.



ACL 2018

56th Annual Meeting of the Association for Computational Linguistics

15-20 July 2018 Melbourne

Image from A Canvas of Light

Tutorials

Tutorials will be held on July 15th, 2018. All tutorials will run for a half-day at the times noted below

T1: 100 Things You Always Wanted to Know about Semantics & Pragmatics But Were Afraid to Ask

Emily M. Bender

09:00 – 12:30

Location: 216, MCEC

NeurIPS | NIPS | NIP | NIPS 2016 Schedule | NIPS 2015 Schedule

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

Login Login Login Search Schedule Filter Day Filtering for

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

Year (2019) ▾

Help ▾

My Registrations

Profile ▾

Contact NeurIPS

Sponsor Info

Publications

Future Meetings

Diversity & Inclusion

Code of Conduct

About Us

Press

News

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 Aw

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 » »

NIPS 2006 Schedule

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

Login Search Schedule Filter Day Filtering for Tutorials

NeurIPS | 2006

Twentyith 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

Dec 2010

6

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 (DJVu) » 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) »

Past tutorials - Admin Wiki

aclweb.org/adminwiki/index.php?title=Past_tutorials

Page Discussion Read View source View history Search

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-Tzuna 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 2019 Schedule

nips.cc/Conferences/2019/Schedule?type=Tutorial

Login Search Schedule

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
- Record 2019

Filter Day Filtering for Tutorials

Past tutorials - Admin Wiki

Mon Dec 9th 08:30 -- 10:30 **Human Behavior**
Nuria M Oliver · Albert Ali

Mon Dec 9th 08:30 -- 10:30 **Imitation Learning**
Kyunghyun Cho · Hal Daumé III

Mon Dec 9th 08:30 -- 10:30 **Deep Learning with Reinforcement Learning**
Mohammad Emtiyaz Khan · Doina Precup

Mon Dec 9th 11:15 AM -- 12:30 **Machine Learning**
Anna Goldengberg · Barbara Webb

Mon Dec 9th 11:15 AM -- 12:30 **Interpretable Computation**
Wittawat Jitkrittum · Doug Erwin

Mon Dec 9th 11:15 AM -- 12:30 **Efficient Processing**
Vivienne Sze

Mon Dec 9th 02:45 -- 04:15 **Representation Learning**
Moustapha Cisse · Samim Webb

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

ACL Anthology

Past tutorials

This page belongs to the COLING conference.

Contents [hide]

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

2019 tutorial

Title
Latent Structure Processing
Graph-Based Processing
Discourse Analysis
Computational Research Evaluation
Wikipedia as a Retrieval System
Deep Bayesian Networks

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)

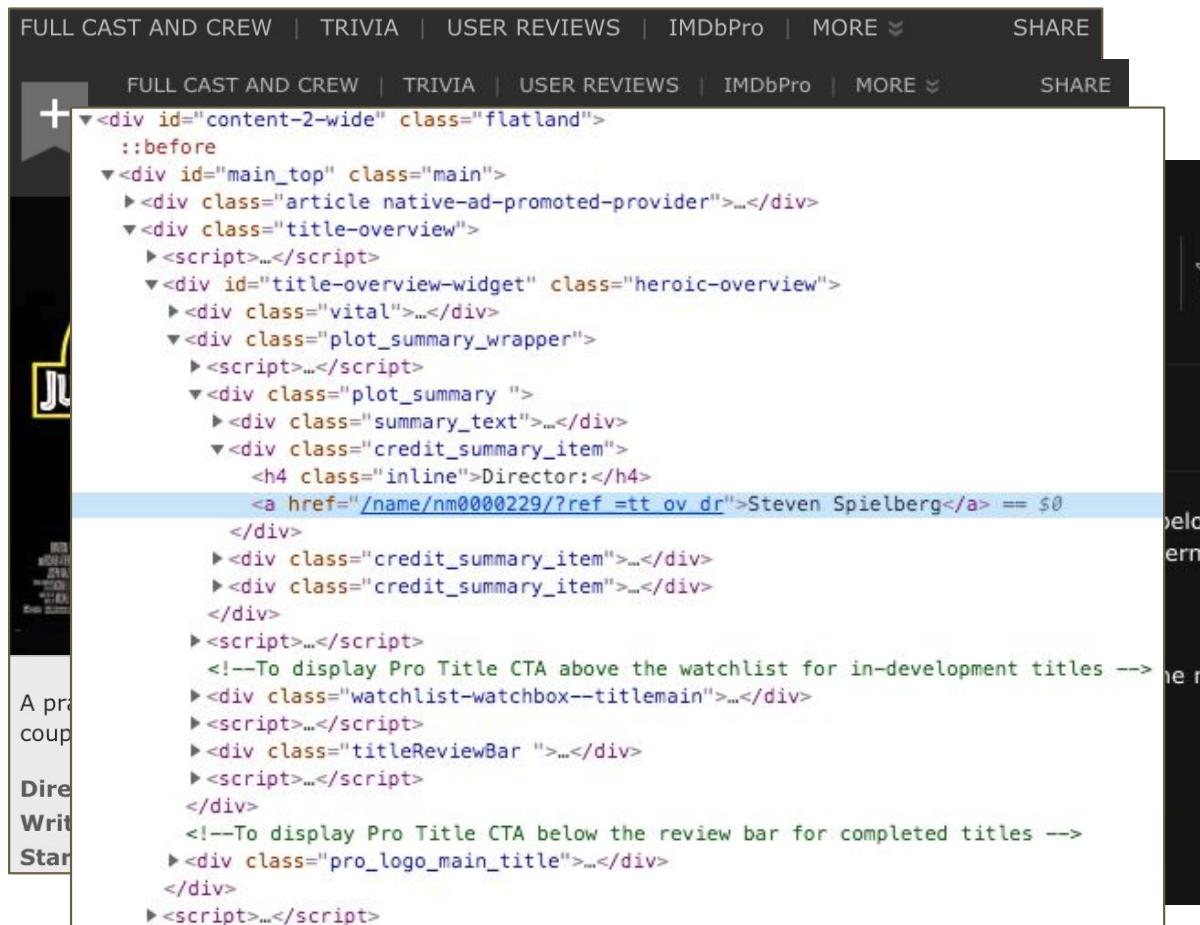
From Wikipedia, the free encyclopedia

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

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 screenshot shows a portion of the IMDB website for the movie "Flatland". The page includes navigation links like "FULL CAST AND CREW", "TRIVIA", "USER REVIEWS", "IMDbPro", and "MORE". Below the header, there's a title overview section with a "plot_summary" div containing a summary text and credit summary items for the director (Steven Spielberg) and other crew members. The code is displayed with syntax highlighting for HTML tags and CSS classes.

```
+<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/nm000229/?ref=tt_overview">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 ²)
2	Kielder Reservoir	3.86 sq mi (10.0 km ²)
3	Ullswater	3.44 sq mi (8.9 km ²)
4	Bassenthwaite Lake	2.06 sq mi (5.3 km ²)
5	Derwent Water	2.06 sq mi (5.3 km ²)

(a) Relational Table

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

(b) Entity Table

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

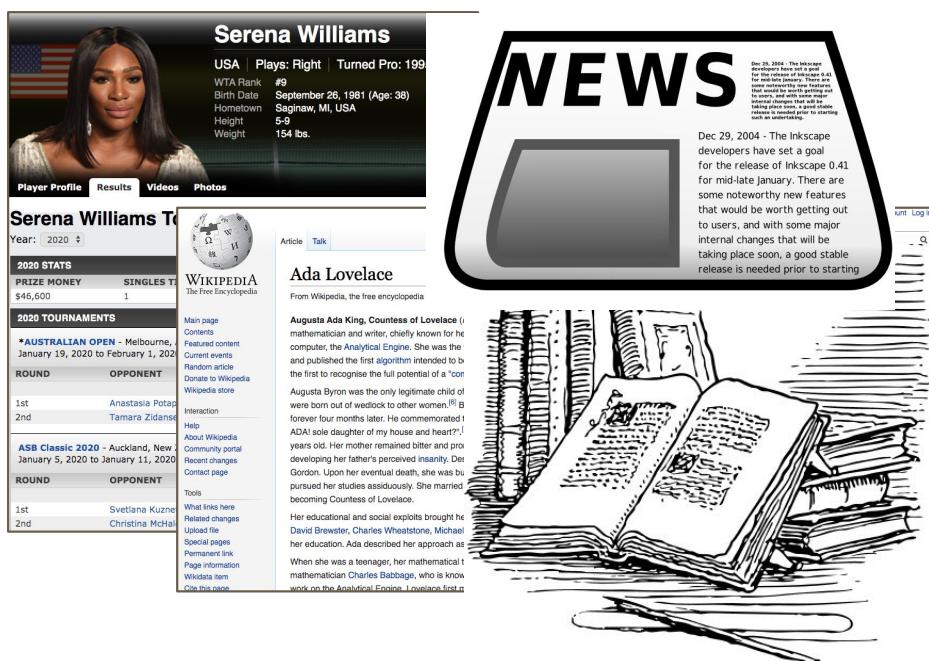
(c) Matrix Table

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

Image via <http://webdatacommons.org/webtables/>, Braunschweig et al, 2015

What is Information Extraction?

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



(Serena Williams, born in, Florida)

Subject

Object

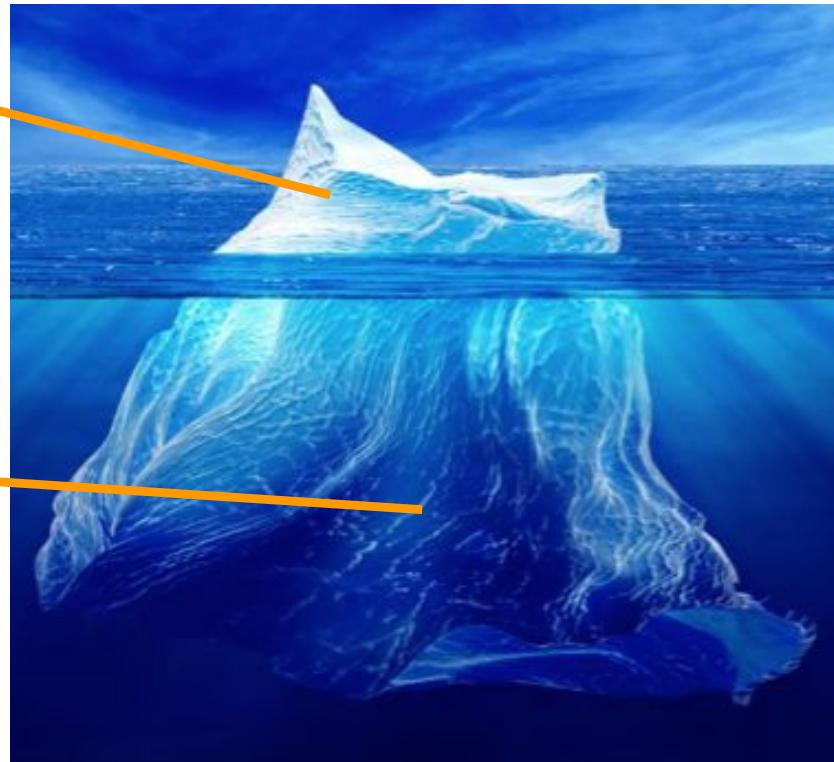
Predicate
(Relation)

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

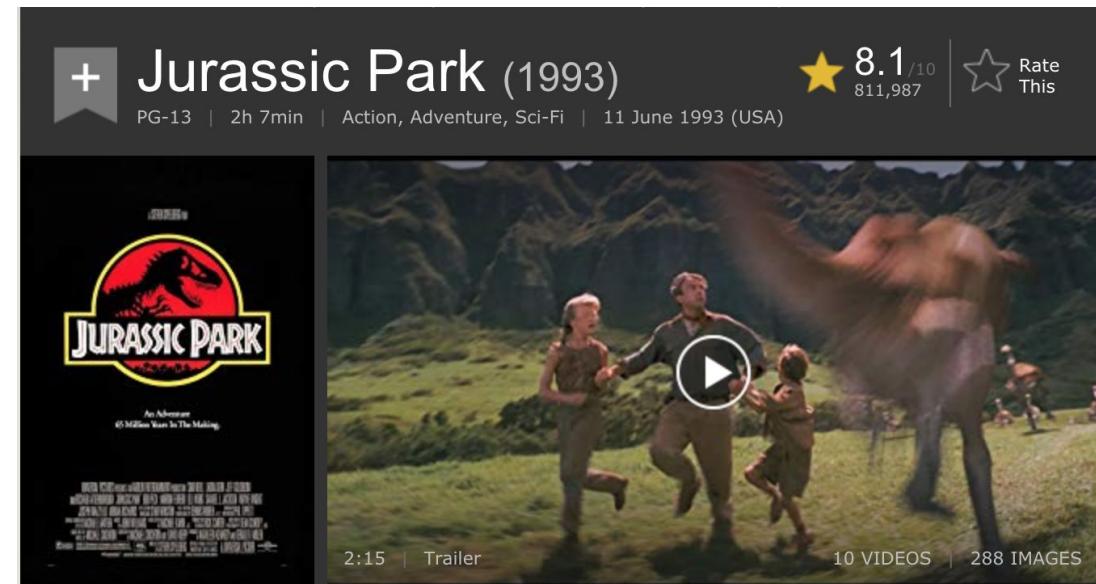
This article is about the 1993 film. For the franchise, see [Jurassic Park \(disambiguation\)](#).

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

The screenshot shows the movie page for "Jurassic Park (1993)". At the top, there is a navigation bar with a plus sign icon, the movie title "Jurassic Park (1993)", its rating of "8.1 / 10" (based on 811,987 reviews), and a "Rate This" button. Below the title, the movie's details are listed: PG-13, 2h 7min, Action, Adventure, Sci-Fi, and the release date of 11 June 1993 (USA). The main image on the left is the movie poster, which features a T-Rex in a red and yellow logo. To the right is a scene from the movie showing a large Tyrannosaurus Rex running towards two people in a lush green landscape. A play button icon is overlaid on the scene. Below the image, it says "2:15 | Trailer". At the bottom, there is a summary of the plot: "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." Below the summary, there are sections for "Director: Steven Spielberg", "Writers: Michael Crichton (novel), Michael Crichton (screenplay) | 1 more credit »", and "Stars: Sam Neill, Laura Dern, Jeff Goldblum | See full cast & crew »".

Semi-structured data vs. semi-structured data

Directed by	Steven Spielberg
Produced by	Kathleen Kennedy Gerald R. Molen
Screenplay by	Michael Crichton David Koepp
Based on	<i>Jurassic Park</i> by Michael Crichton
Starring	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 ^[154]	International Film	<i>Jurassic Park</i>	Won
	66th Academy Awards ^[155]	Best Sound Editing	Gary Rydstrom and Richard Hymns	Won
		Best Sound Mixing	Gary Summers, Gary Rydstrom, Shawn Murphy and Ron Judkins	Won
		Best Visual Effects	Dennis Muren, Stan Winston, Phil Tippett and Michael Lantieri	Won
		Best Director	Steven Spielberg	Won
	Saturn Awards ^[147]	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

Directed by Steven Spielberg

Produced by Kathleen Kennedy

Gerald R. Molen

Screenplay by Michael Crichton

David Koepp

Based on *Jurassic Park*

by Michael Crichton

Starring 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

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 diverse data

Extracting from multiple languages = More diverse data

Extracting from multiple subject domains = More diverse data

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
- We need 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 model
- But labeling data is expensive
- We need to:
 - Label data cheaply
 - Label data automatically
 - Learn from limited data
 - Learn from noisy data



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
 - 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

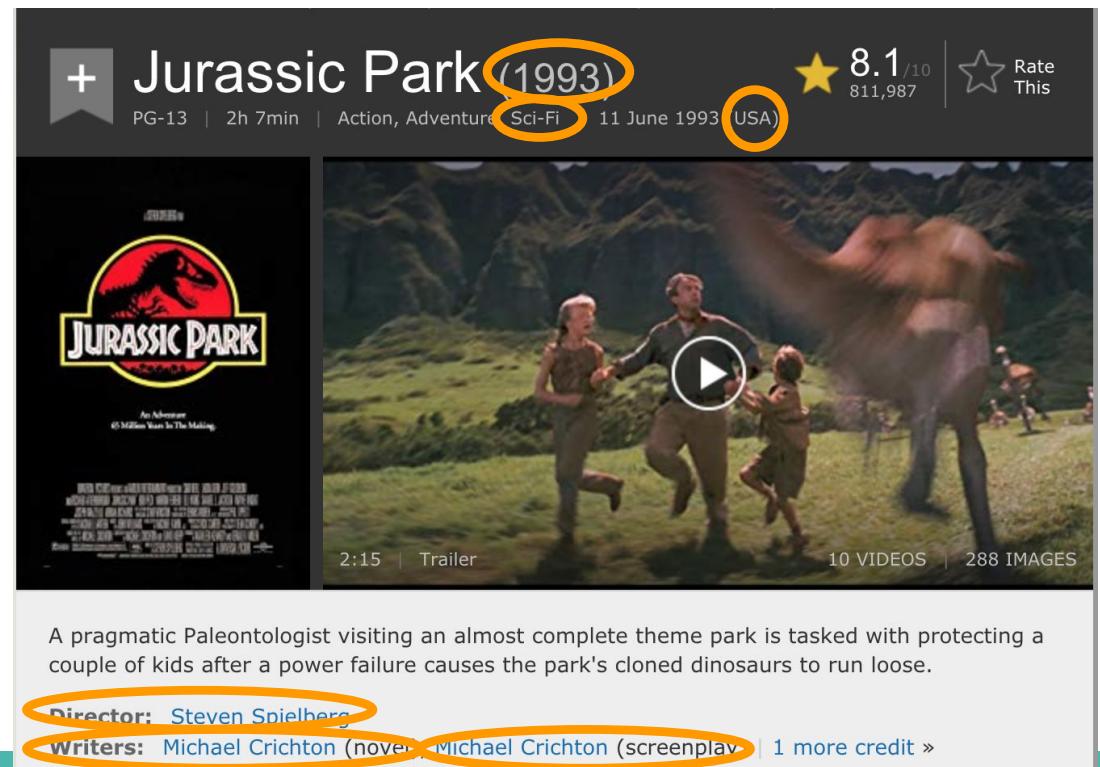
- Diversity→Identifying consistent patterns

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



Key Intuitions

- Diversity→Identifying consistent patterns

노다지
A Bonanza (Nodaji)

1961년 · 대한민국 · 127분 · 1961-06-01 (개봉)

제작사 화성영화주식회사

감독 정창화

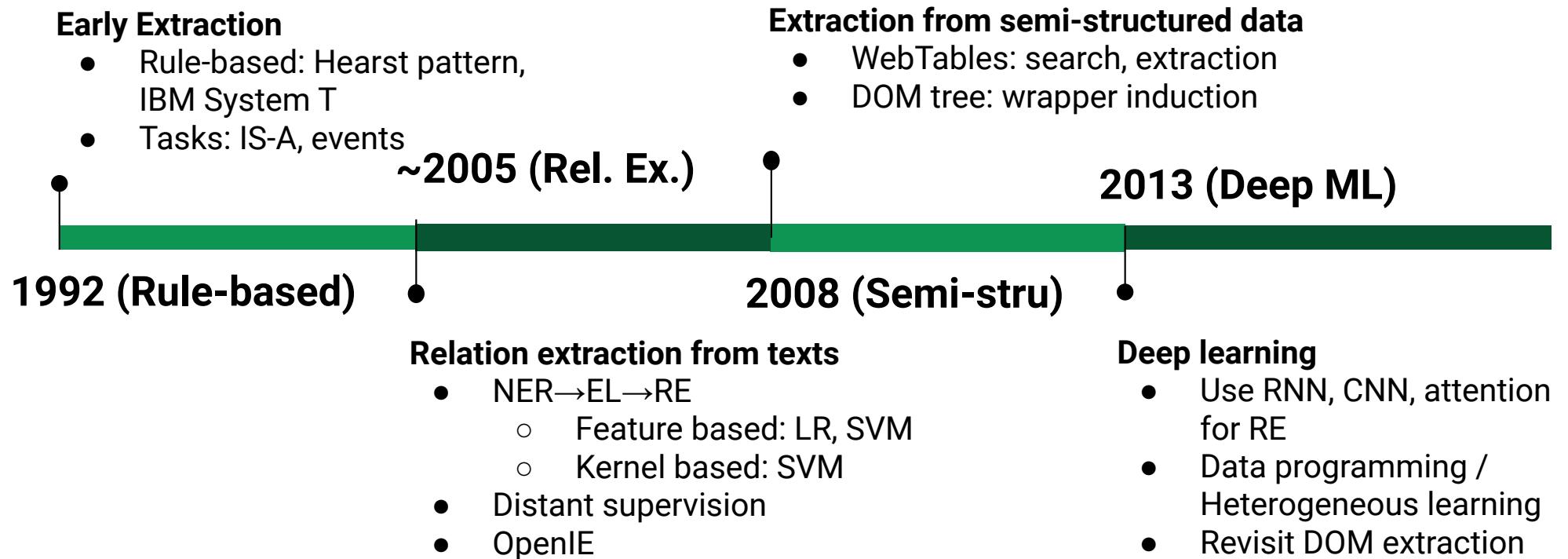
출연 김승호, 황해, 엄앵란, 조미령, 허장강 [더보기](#)

 [스크랩하기](#)

Key Intuitions

- Diversity→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





© 1999 Kluwer Academic Publishers. Manufactured in The Netherlands.

Machine Learning 34, 233–272 (1999)

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

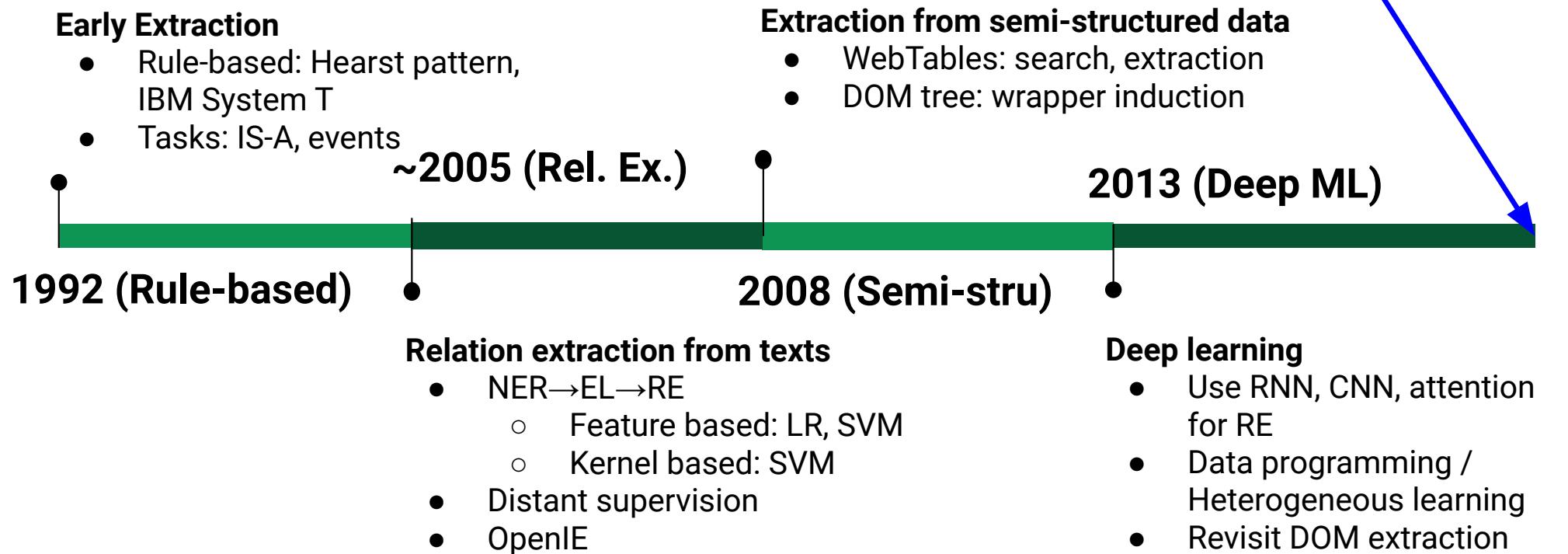
Editors: Claire Cardie and Raymond Mooney

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?

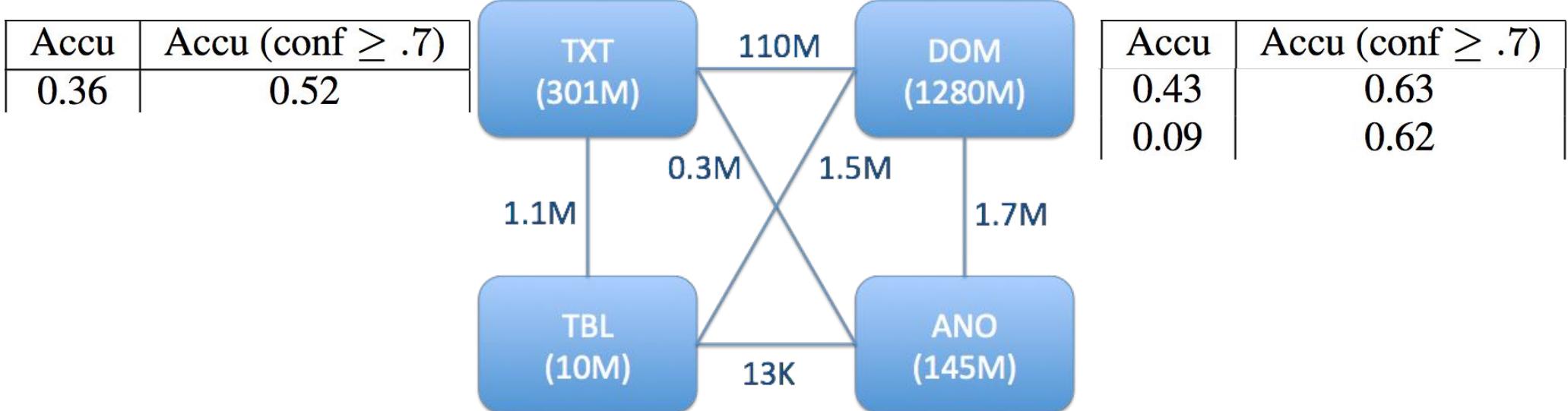
In the multi-modal setting, we will consider methods that jointly address unstructured, semi-structured, and tabular text and bring in **visual** information

No real full-fledged systems in practice yet



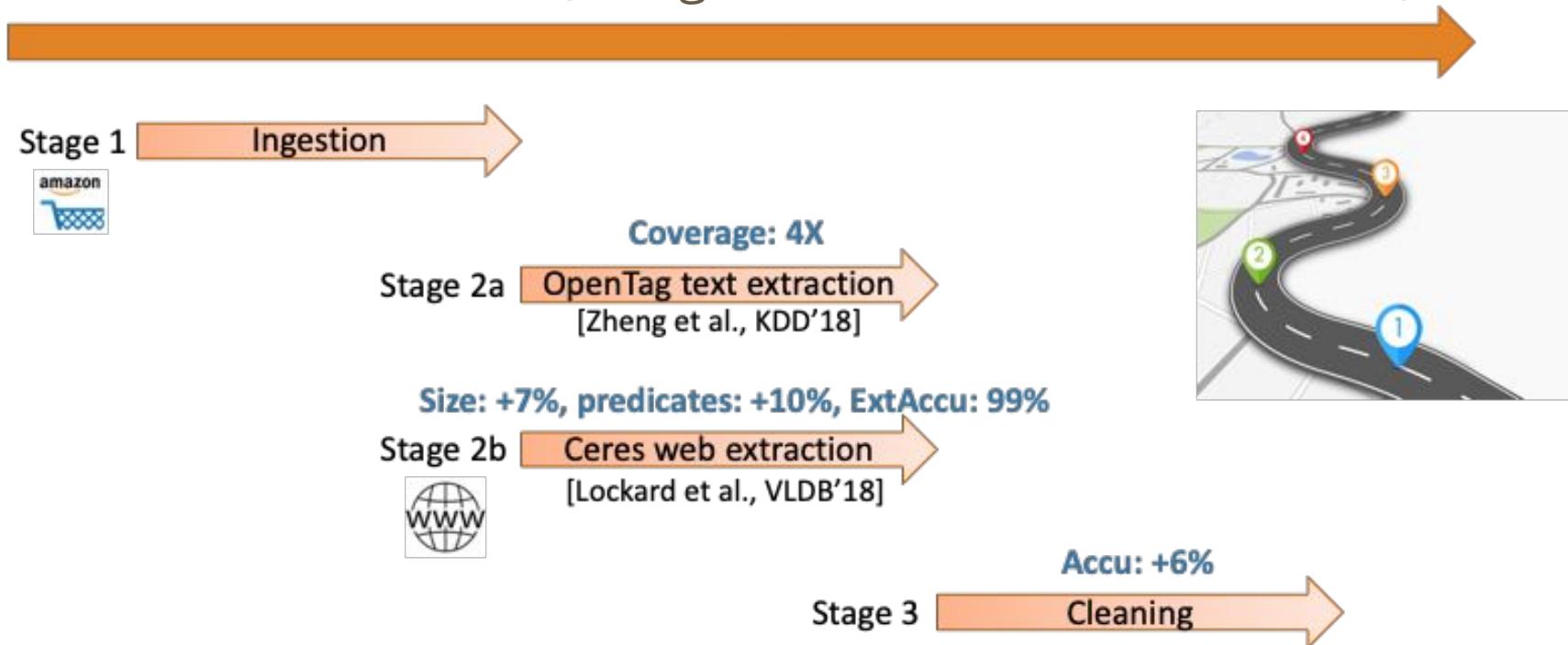
Example 1. Google Knowledge Vault

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



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 (30 minutes)
 - Part Ia: Unstructured text (30 minutes)
 - Break (30 minutes)
 - Part Ib: Unstructured text: Methods (15 minutes)
 - Part II: Semi-structured text (45 minutes)
 - Part III: Tabular text (15 minutes)
 - Part IV: Multi-modal extraction (30 minutes)
 - Conclusion and future directions (15 minutes)
-

Knowledge Collection from Unstructured Text

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



Outline

- Introduction (30 minutes)
- **Part Ia: Unstructured text: Overview** (30 minutes)
- Break (30 minutes)
- Part Ib: Unstructured text: Methods (15 minutes)
- Part II: Semi-structured text (45 minutes)
- Part III: Tabular text (15 minutes)
- Part IV: Multi-modal extraction (30 minutes)
- Conclusion and future directions (15 minutes)



Questions we will answer in this section

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

Questions we will answer in this section

What are the tasks and subtasks of extraction from unstructured text?

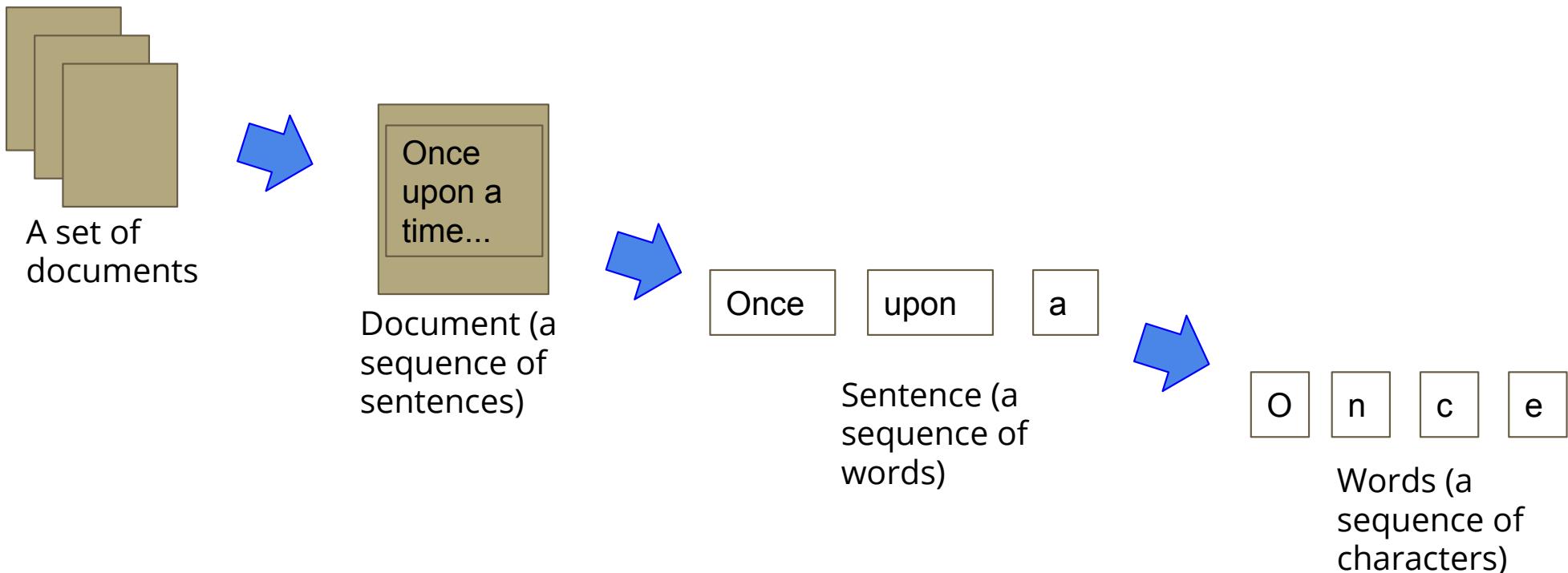
What are the models and algorithms used to approach this task?

What are the central challenges in implementing a system in practice?

How can those challenges be overcome?



What Is Unstructured Text?





WIKIPEDIA
The Free Encyclopedia

Article Talk

Oprah Winfrey

From Wikipedia, the free encyclopedia

“Oprah Gail Winfrey (born Orpah Gail Winfrey;^[1] January 29, 1954) is an American media executive, actress, talk show host, television producer, and philanthropist.”

The New York Times

Bumblebee Vomit: Scientists Are No Longer Ignoring It

https://www.nytimes.com/2020/01/22/science/bees-vomit-nectar.html?algo=identity&fallback=false&imp_id=571335236&imp_id=309859981&action=click&module=Science%20%20Technology&pgtype=Homepage

Illustrated Edition

LITTLE WOMEN

LOUISA MAY ALCOTT



*Little Women | Good Wives
Little Men | Jo's Boys*

"She's coming! Strike up, Beth! Open the door, Amy! Three cheers for Marmee!" cried Jo, prancing about while Meg went to conduct Mother to the seat of honor.'



Ariana Grande 
@ArianaGrande



“guys i can’t tell u why yet but i’m so excited for tonight i’ve never felt this way goodbye”

facebook



Cristiano Ronaldo

January 1 at 9:54 AM · 

“Feliz Ano, meu Amor! ❤ Que 2020 seja um ano repleto de amor, saúde, paz e sucesso para todos! Happy New Year to all! 🎉”

[https://www.facebook.com/Cristiano/posts/10157949836957164?_xts_\[0\]=68.ARAbIf0bWtRKzH5XU4C-ExoUZZ3OmxNWEsNJCIciAZRTvptLB1onHoTUsNbhUBW6N4-3GnPmPcH7Wt2m-HtWE-sneckG0zKOdzT5b-k1ZW-8SShEkUvkk62FF_AZH99sJlHKXHvg1sdGPN1LB4kbGQgda_EaKYpap00Bm607HxgL9j5sYxgcvkm7_3iFfxu86TSIGRn7HeIdIWMYatcotHEEaX_G63MMHd4H0wSq05RmmDsqaCPYvx6wlPrqlvKlxdlPPdxBbyP0WEjFvmmLuaa98ZpG97O6ffWH0915I8PnwkpWzZQbsSpT9No-vgxpRL0xTdk0Bm8i-ljiAA&__tn__=-R](https://www.facebook.com/Cristiano/posts/10157949836957164?_xts_[0]=68.ARAbIf0bWtRKzH5XU4C-ExoUZZ3OmxNWEsNJCIciAZRTvptLB1onHoTUsNbhUBW6N4-3GnPmPcH7Wt2m-HtWE-sneckG0zKOdzT5b-k1ZW-8SShEkUvkk62FF_AZH99sJlHKXHvg1sdGPN1LB4kbGQgda_EaKYpap00Bm607HxgL9j5sYxgcvkm7_3iFfxu86TSIGRn7HeIdIWMYatcotHEEaX_G63MMHd4H0wSq05RmmDsqaCPYvx6wlPrqlvKlxdlPPdxBbyP0WEjFvmmLuaa98ZpG97O6ffWH0915I8PnwkpWzZQbsSpT9No-vgxpRL0xTdk0Bm8i-ljiAA&__tn__=-R)

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



What is extraction from texts

- Input: Paragraphs (about a particular entity, or general paragraphs)
- Output
 - Binary relationship: IS-A
 - Triple relationship: (subject, predicate, object)
 - Event: 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.

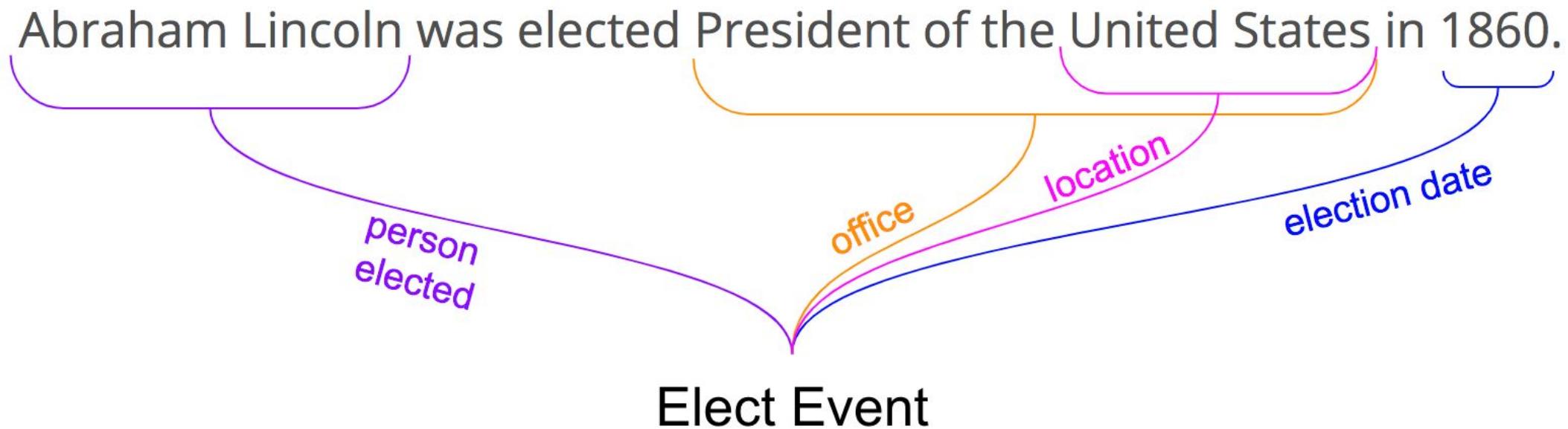
Person

Person

is friends with

Event Extraction

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



Why is extraction from text hard?

- Diversity

Bill Gates founded Microsoft in 1975.

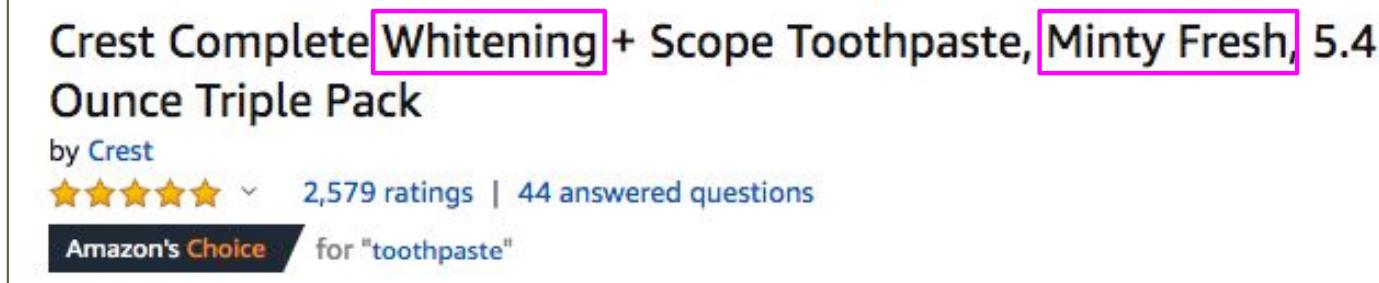
Bill Gates, founder of Microsoft, ...

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?

- Fuzzy language with weak structure



History of Amazon - Wikipedia

Amazon was founded in the garage of **Bezos**' rented home in Bellevue, Washington. **Bezos**' parents invested almost \$250,000 in the start-up. In July 1995, the company began service as an online bookstore.

Why is extraction from text hard?

- 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

SCIENCE \ FOSSILS \ BOOKS

How Jurassic Park led to the modernization of dinosaur paleontology

It led to an explosion of interest in dinosaurs, and by extension, people interested in researching them

Why is extraction from text hard?

- Lack of training data

SACODEYL Annotator 2.9

Corpus Document Config Help

Category

Add Cat. Delete cat. Up cat. Down cat.

Logical view XML View Document metadata

Se... Applied categories Speaker Transcription

1

- Instituto
- Vacaciones
- Viajar
- Sección de interés para su explotación
- Pues a principio cl apoyar
- Características propias del lenguaje oral
- A nivel de texto
- Preterito Perfecto
- Preterito imperfecto
- Pasiva Refleja
- B1
- Drag and Drop template
- Pretérito indefinido
- Colocados

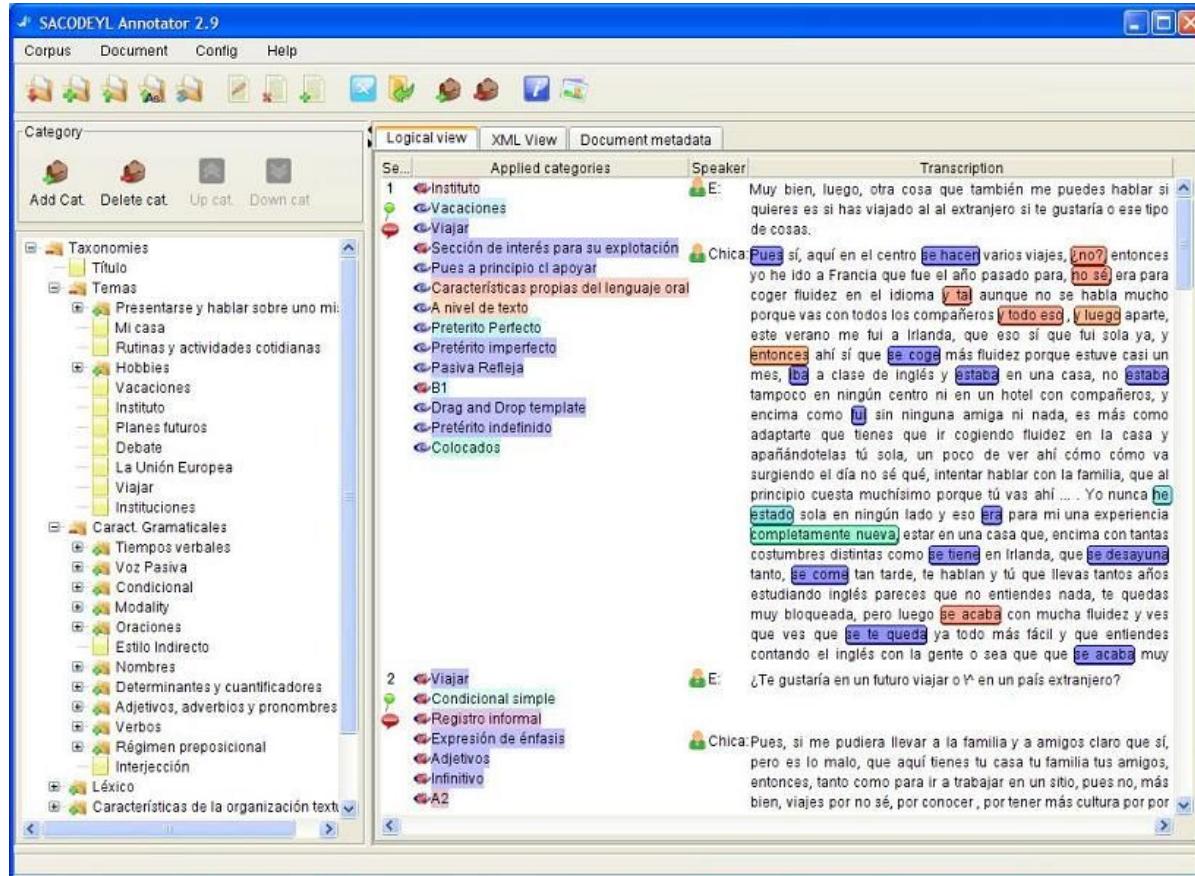
Chica: Pues sí, aquí en el centro se hacen varios viajes, ¿no? entonces yo he ido a Francia que fue el año pasado para, no sé era para coger fluidez en el idioma y ta aunque no se habla mucho porque vas con todos los compañeros y todo eso, y luego aparte, este verano me fui a Irlanda, que eso sí que fui sola ya, y entonces ahí sí que se coge más fluidez porque estuve casi un mes, a la clase de inglés y estaba en una casa, no estaba tampoco en ningún centro ni en un hotel con compañeros, y encima como tu sin ninguna amiga ni nada, es más como adaptarte que tienes que ir cogiendo fluidez en la casa y apañándotelas tú sola, un poco de ver ahí cómo cómo va surgiendo el día no sé qué, intentar hablar con la familia, que al principio cuesta muchísimo porque tú vas ahí ... Yo nunca he estado sola en ningún lado y eso es para mí una experiencia completamente nueva

2

- Viajar
- Condicional simple
- Registro informal
- Expresión de énfasis
- Adjetivos
- Infinitivo
- A2

E: ¿Te gustaría en un futuro viajar o ya en un país extranjero?

Chica: Pues, si me pudiera llevar a la familia y a amigos claro que sí, pero es lo malo, que aquí tienes tu casa tu familia tus amigos, entonces, tanto como para ir a trabajar en un sitio, pues no, más bien, viajes por no sé, por conocer, portener 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

Opportunities

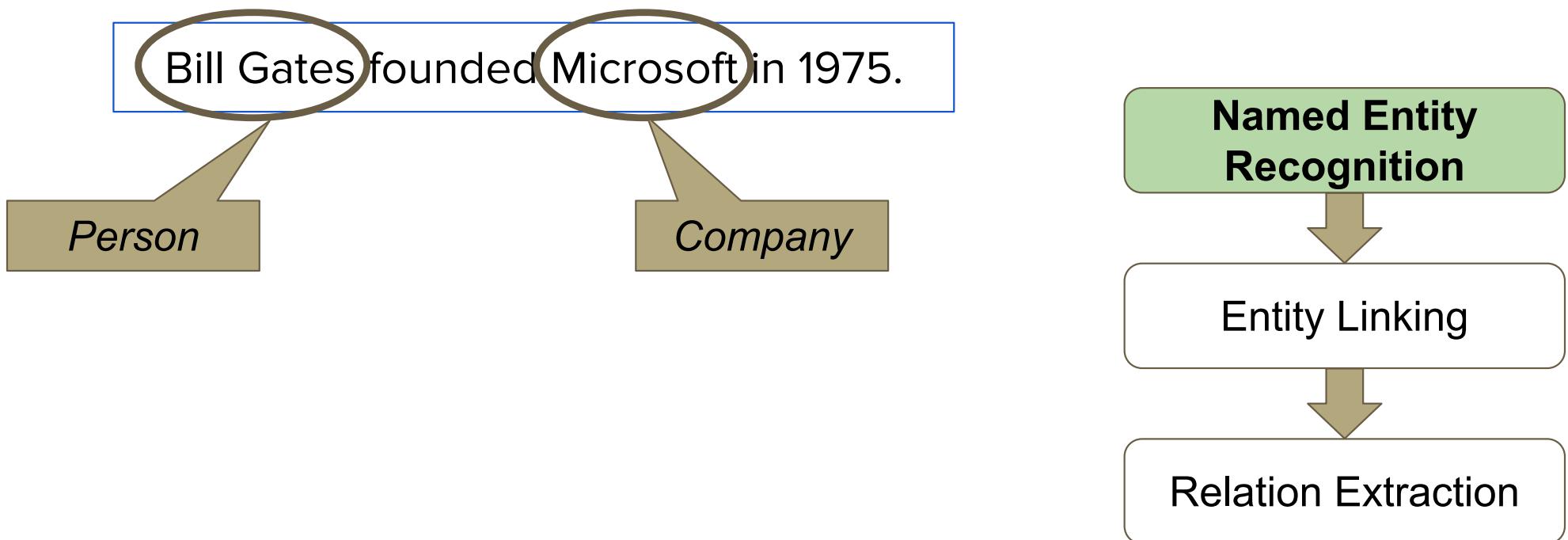
- **Consistency:** Same grammar and word semantics
- **Redundancy:** Same fact is often repeated in different articles, in various ways



Short Answers

- **Consistency**
 - Model problem as text span classification and relationships between spans
 - Word embedding models help capture text semantics
- **Training data**
 - 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.



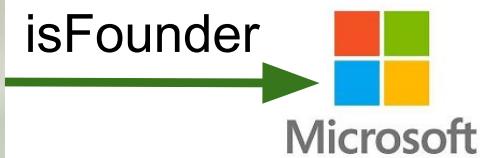
Named Entity
Recognition

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

Classification models

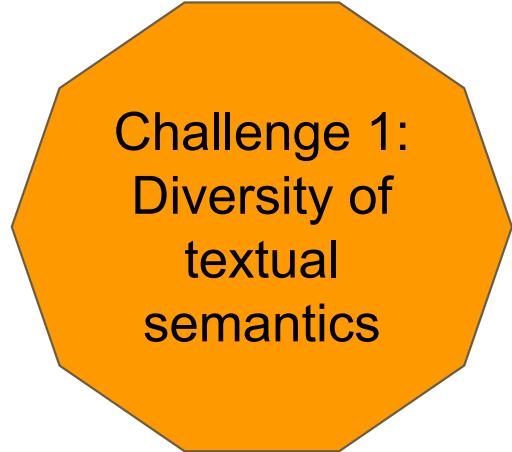
Machine learning classifiers take in a set of **features** that describe a data point and output a **prediction** of that datapoint's class.

We'll need to:

1. Select **features** to represent our raw text
 2. **Combine** those features for larger units (e.g. spans) if necessary
 3. Select a **model** to take in these features and make a prediction
 4. **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 desiderata

- 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

Now: Pre-trained embedding models



Word Embeddings

Dense vector representation of a word or sub-word part

Large corpus: Learn to predict nearby words

Word2Vec (Mikolov et al, 2013), GloVe (Pennington et al, 2014)

Contextual Word Embeddings

- BERT (Devlin et al, 2019): Biggest revolution in NLP of last few years
 - Builds contextual representation of each token in a sentence
 - Training objective: Learn to predict missing words in a sentence
 - Transformer neural net architecture
 - Also builds representation of entire sentence
- Pre-trained BERT available from Google



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



Challenge 1:
Diversity of
language

Faster alternatives to BERT

- Active area of research
 - ALBERT (Lan et al, 2019)
- Alternative embedding models
 - FastText (Grave et al, 2016)



Outline

- Introduction (30 minutes)
 - Part Ia: Unstructured text (30 minutes)
 - **Break (30 minutes)**
 - Part Ib: Unstructured text: Methods (15 minutes)
 - Part II: Semi-structured text (45 minutes)
 - Part III: Tabular text (15 minutes)
 - Part IV: Multi-modal extraction (30 minutes)
 - Conclusion and future directions (15 minutes)
-

Knowledge Collection from Unstructured Text

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



Outline

- Introduction (30 minutes)
 - Part Ia: Unstructured text: Overview (30 minutes)
 - Break (30 minutes)
 - **Part Ib: Unstructured text: Methods** (15 minutes)
 - Part II: Semi-structured text (45 minutes)
 - Part III: Tabular text (15 minutes)
 - Part IV: Multi-modal extraction (30 minutes)
 - Conclusion and future directions (15 minutes)
-

Questions we will answer in this section

How can we extract 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

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

Methods

**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)**



- 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." Each word is followed by a small orange circle representing its tag. Brackets below the words group them into entities: "Winfrey's" is grouped by a purple bracket labeled "B-Person"; "best friend since their early twenties" is grouped by an orange bracket; "is" is a single orange circle; "Gayle" is a single orange circle; and "King" is grouped by a purple bracket labeled "I-Person".

B-Person O 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.

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) 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

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

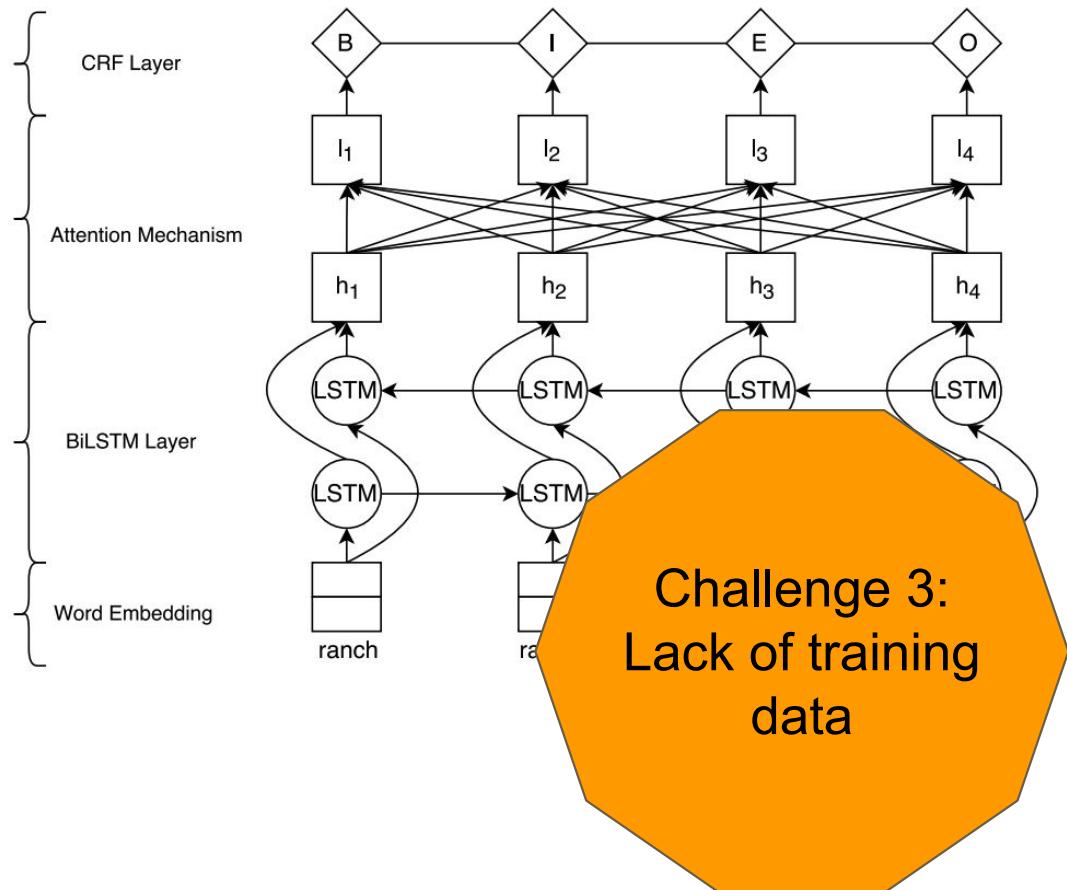
○ ○ B-Flavor I-Flavor ○ B-Flavor I-Flavor ○ ○ ○ ○ ○ ○

(ASIN B0001234567, has_flavor, "Filet Mignon")

(ASIN B0001234567, has_flavor, "Porterhouse Steak")

Span classification: OpenTag

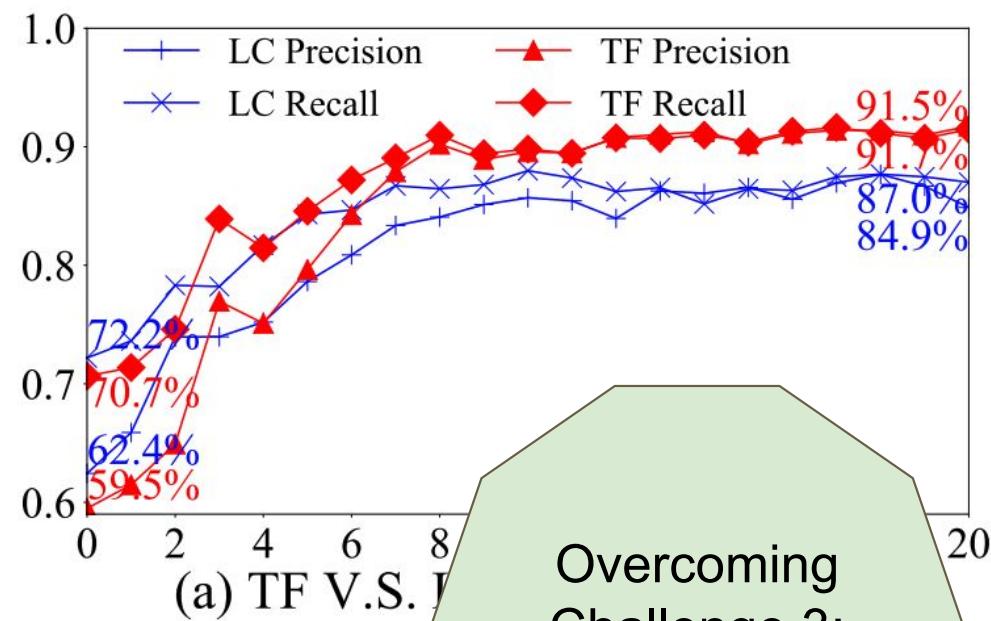
- Word embeddings capture word meaning
- LSTM layer captures word sequence information
- Attention layer allows interaction across sequence
- CRF layer enforces consistency



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	86.6	85.9	86.3
Camera: Title Attribute: Brand name	BiLSTM	94.7	88.8	91.8
	BiLSTM-CRF	91.9	93.8	92.9
	OpenTag	94.9	93.4	94.1
Detergent: Title Attribute: Scent	BiLSTM	81.3	82.2	81.7
	BiLSTM-CRF	85.1	82.6	83.8
	OpenTag	84.5	88.2	86.4

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

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

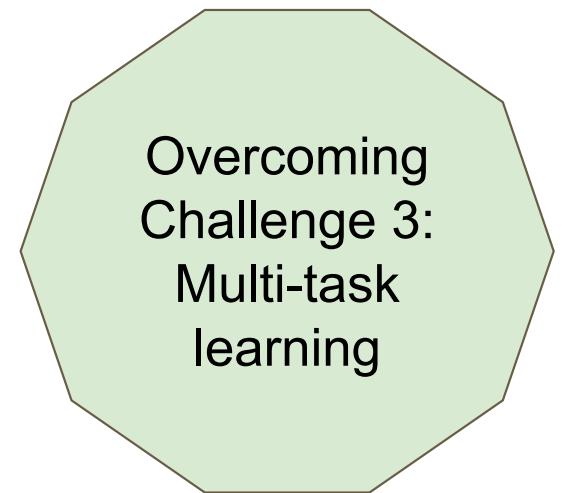
DyGIE (Luan et al, 2019)

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

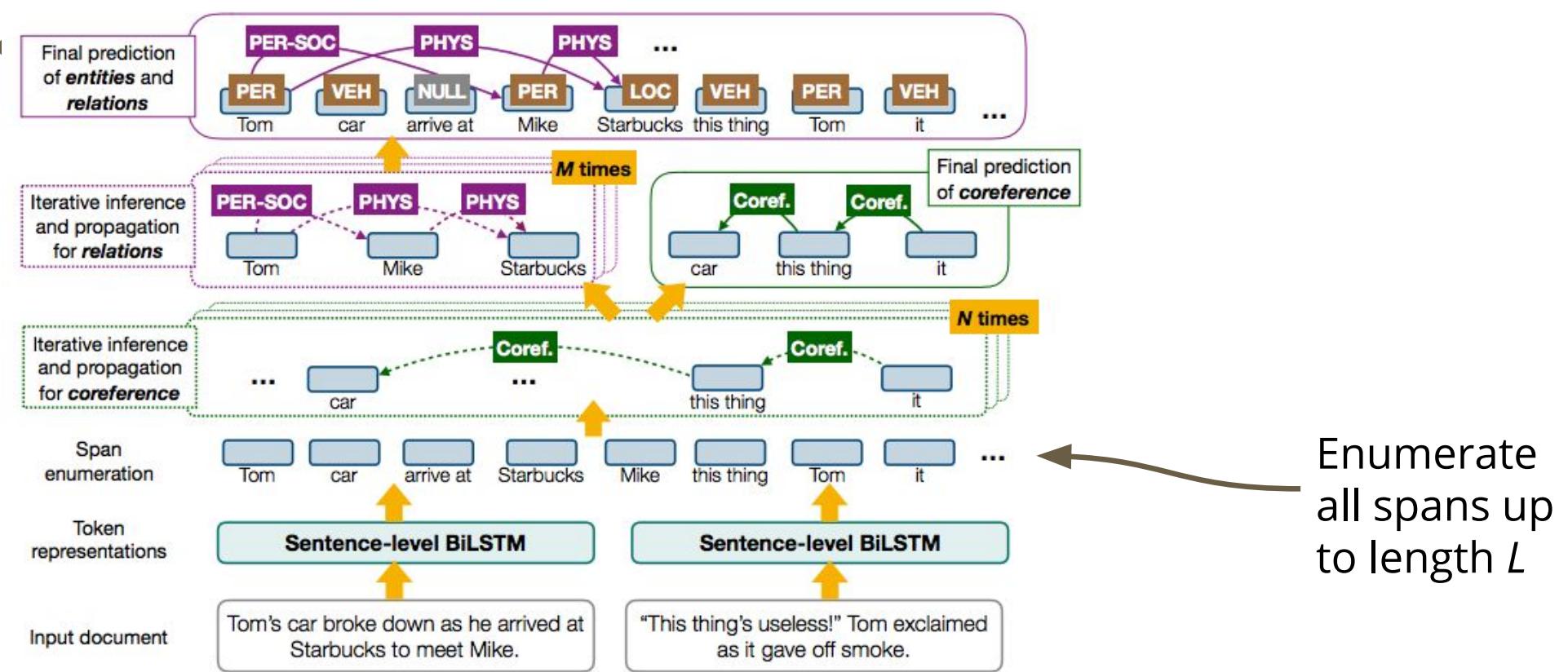


DyGIE (Luan et al, 2019)

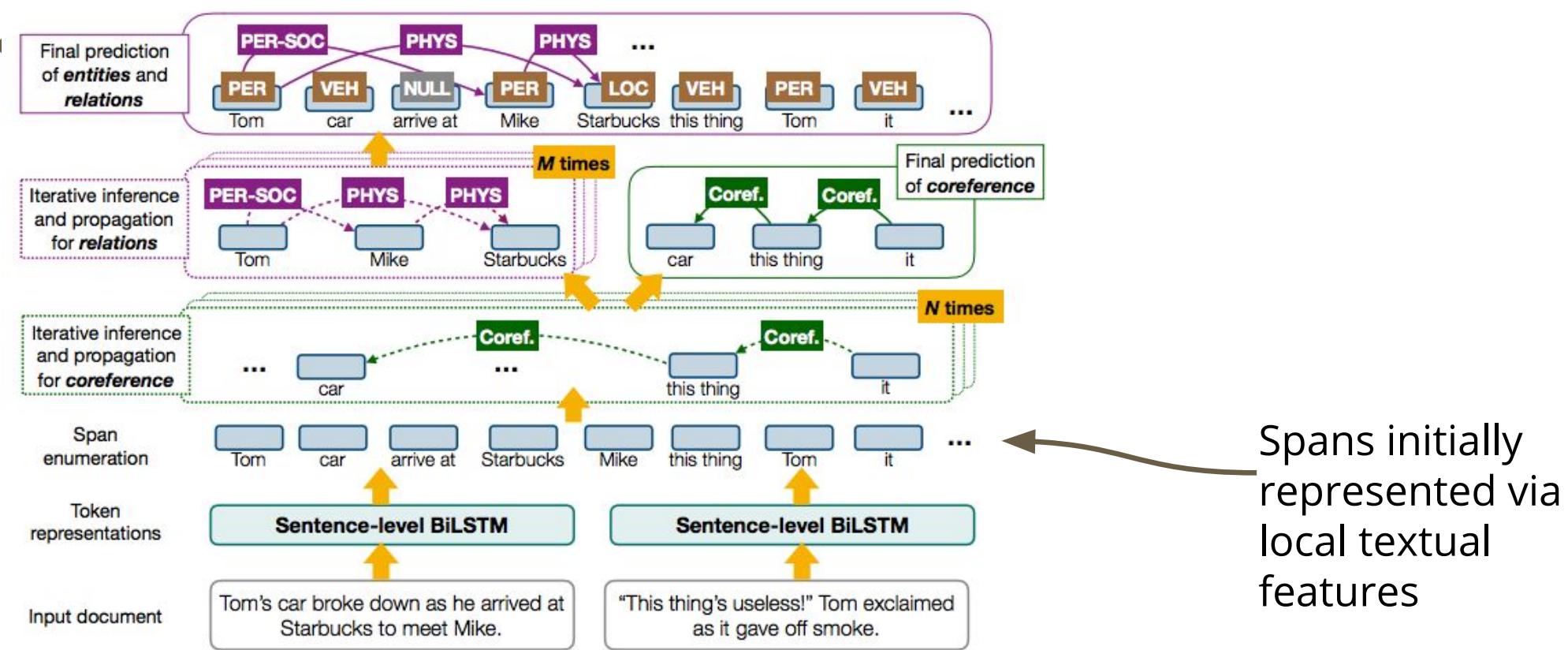
- Single model for NER, co-reference, relation extraction
 - Multi-task learning objective



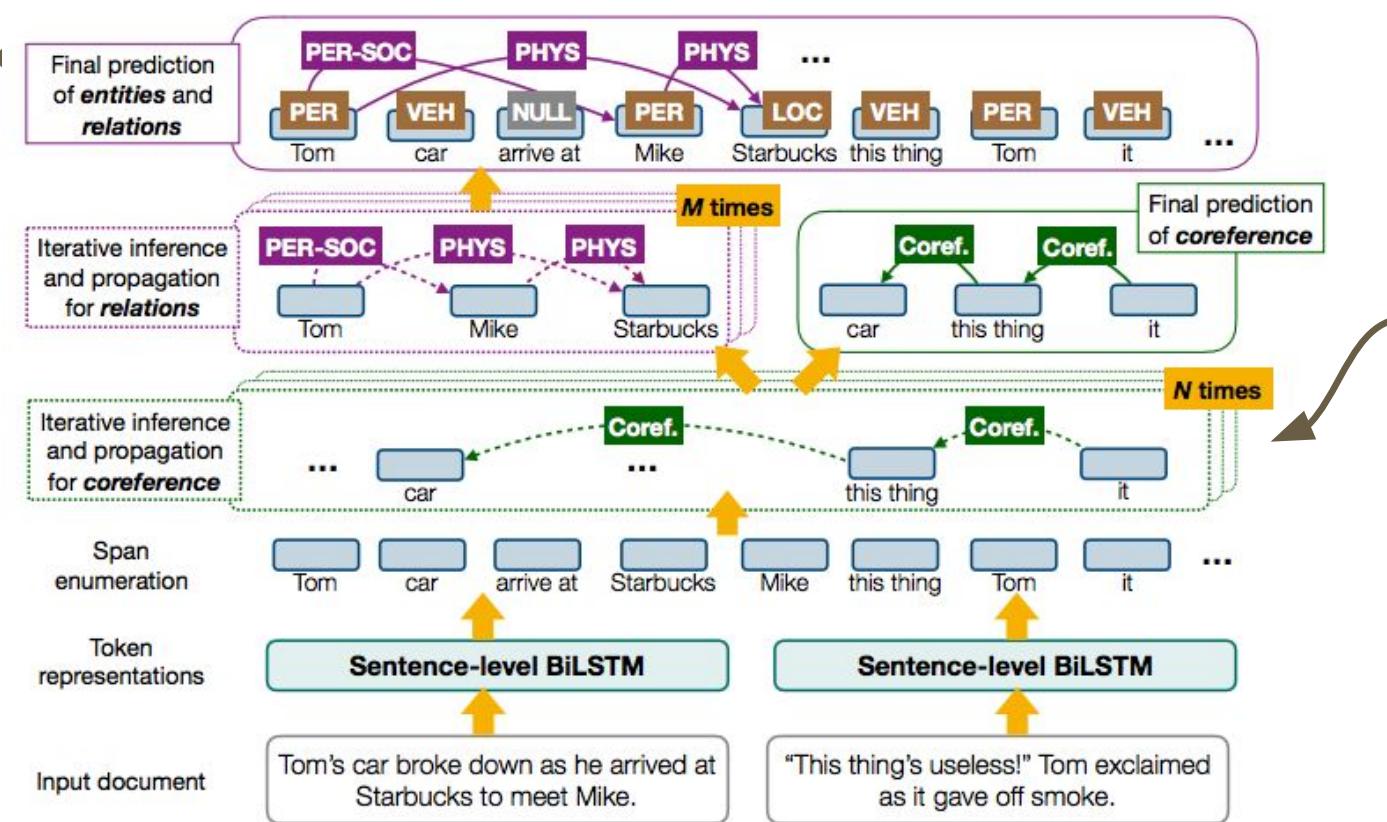
DyGIE (Luan et al, 2019)



DyGIE (Luan et al, 2019)



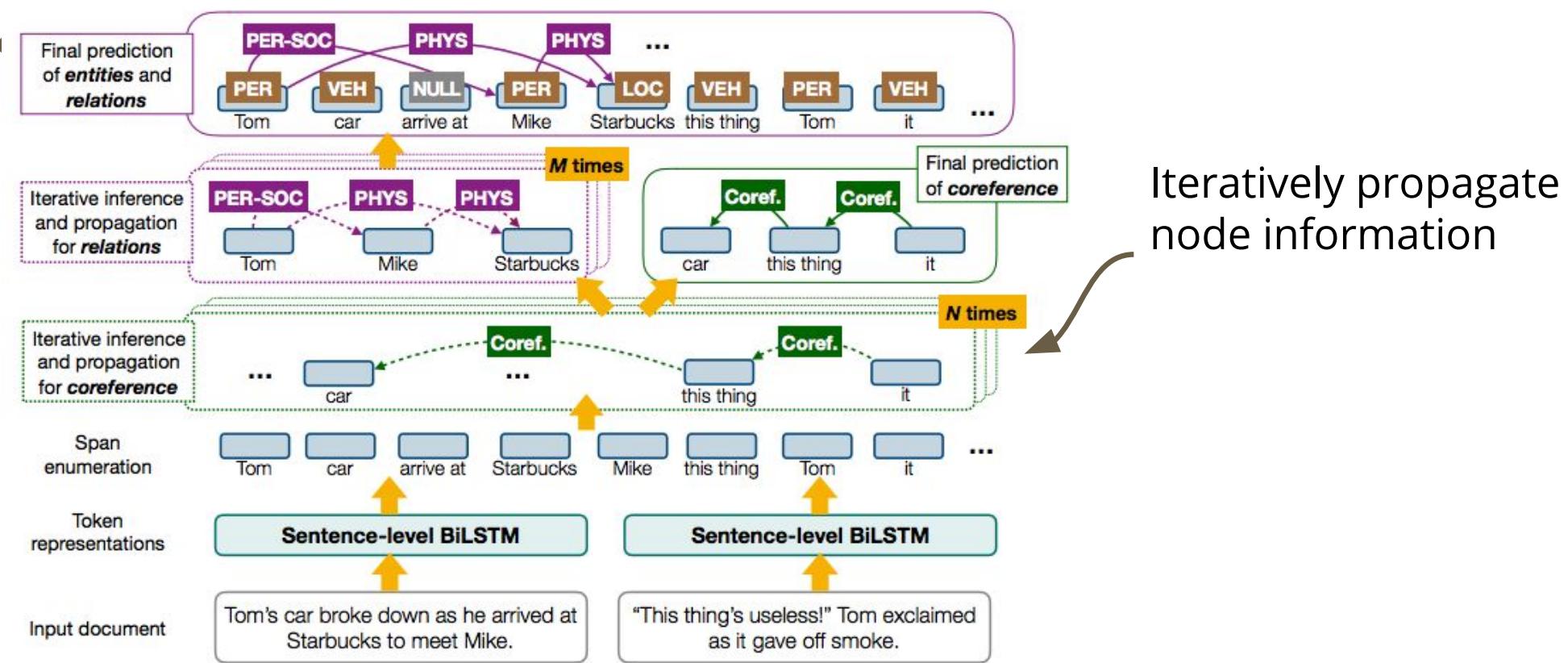
DyGIE (Luan et al, 2019)



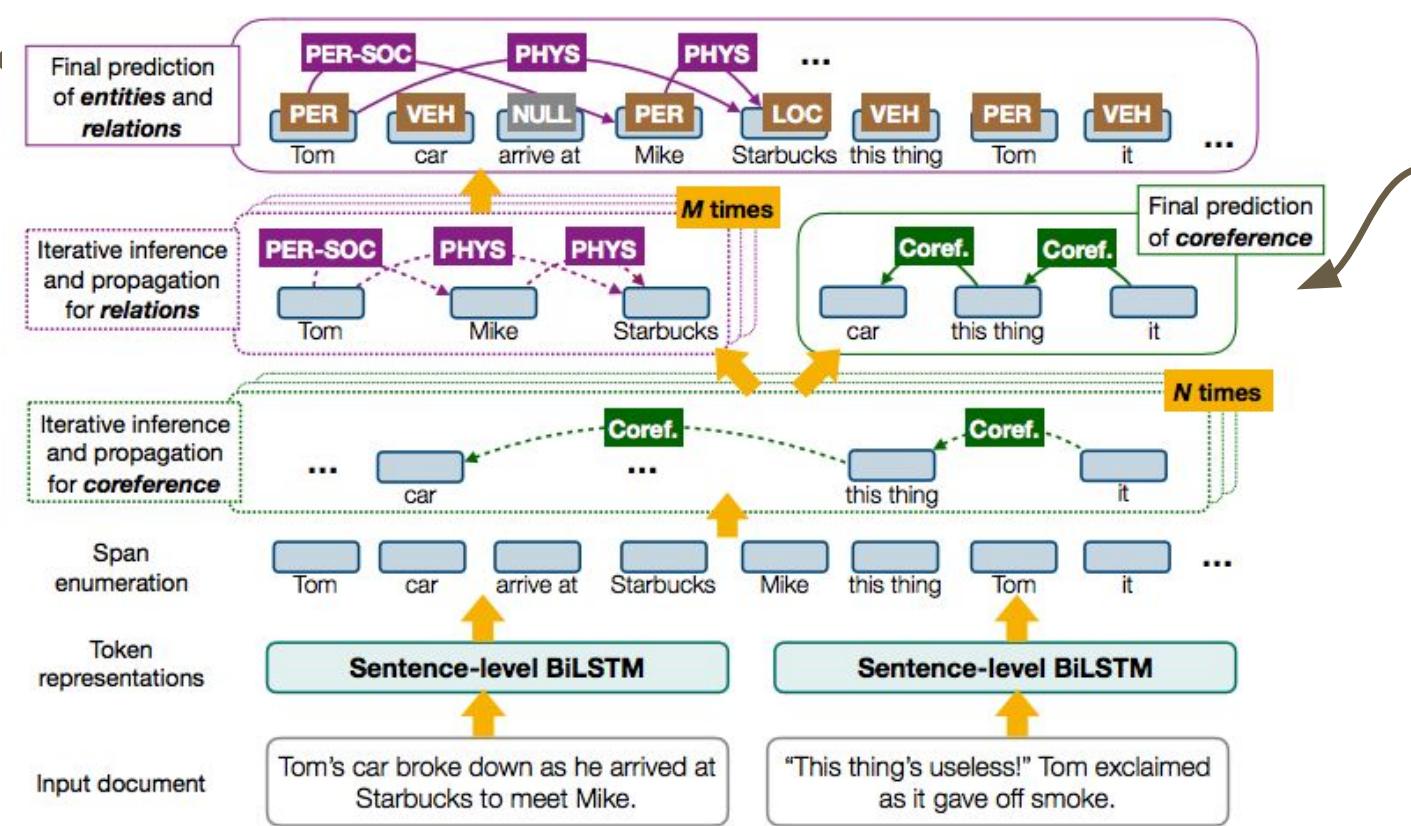
Construct graph:

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

DyGIE (Luan et al, 2019)



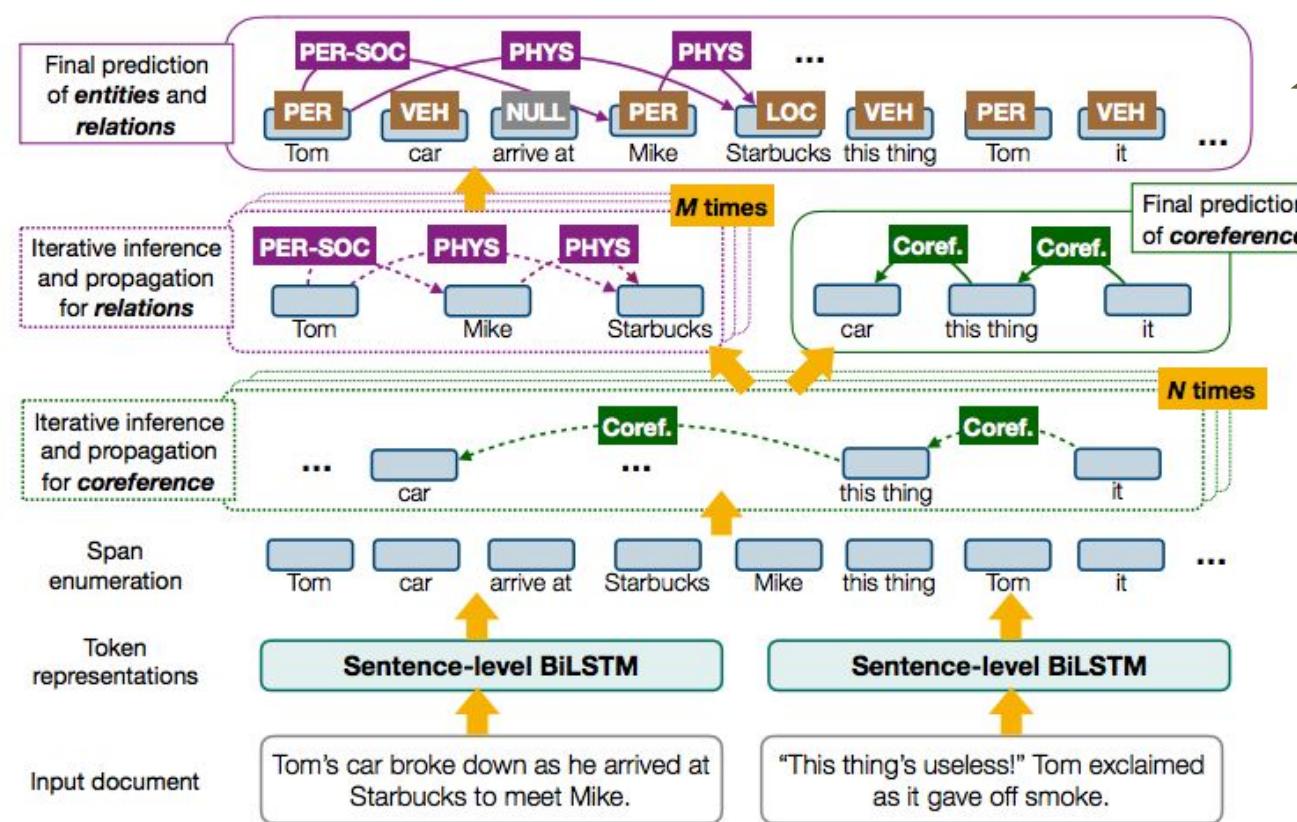
DyGIE (Luan et al, 2019)



Repeat process for relations:

- Edges now indicate relation types

DyGIE (Luan et al, 2019)

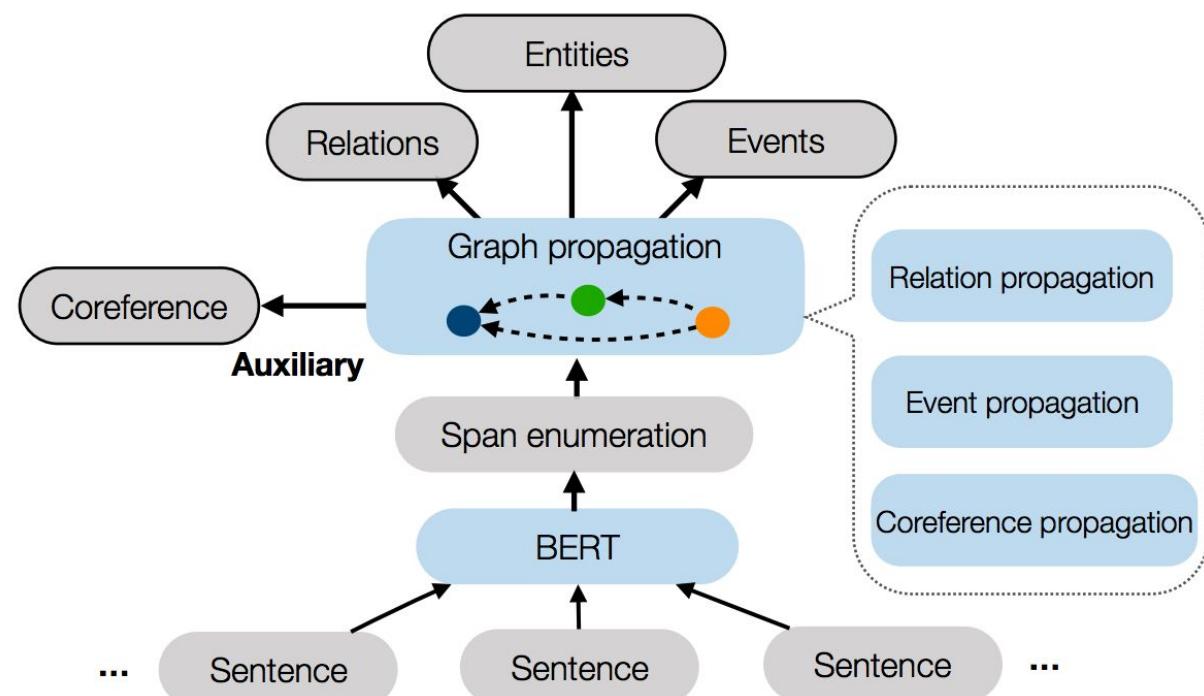


Use final representations to predict entity types and relations

DyGIE++ (Wadden et al, 2019)

DyGIE++ adds events

Replaces word embeddings and LSTM with BERT word representations



DyGIE++

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

State-of-the-art
results across many
datasets

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)

How can we extract without manually labeling data?

Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

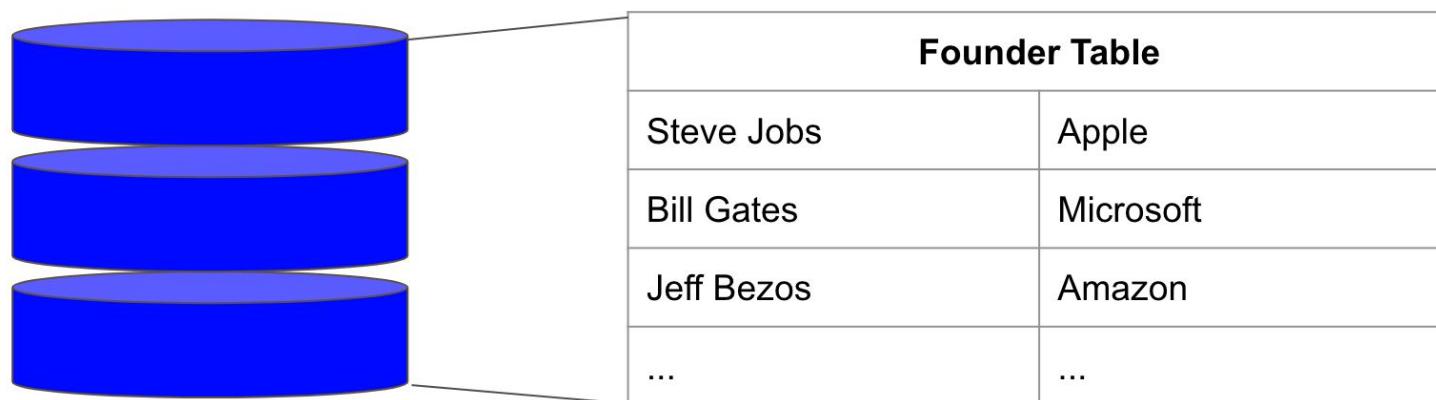
Steve Jobs was the founder of Apple.



Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

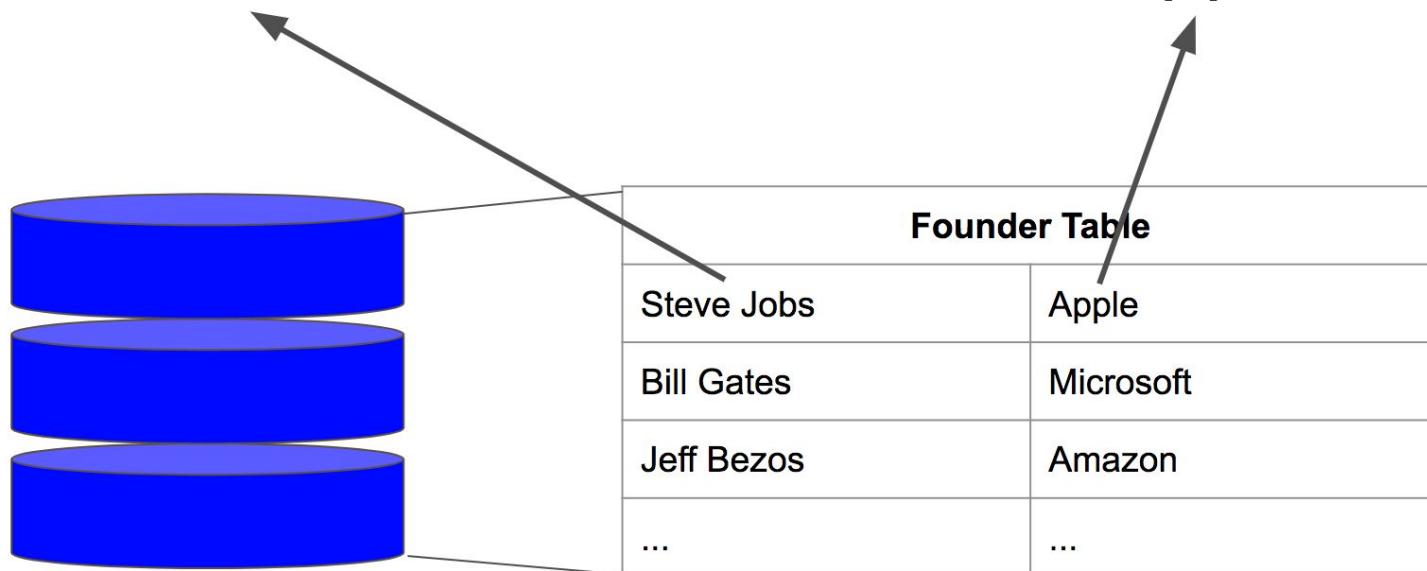
Steve Jobs was the founder of Apple.



Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

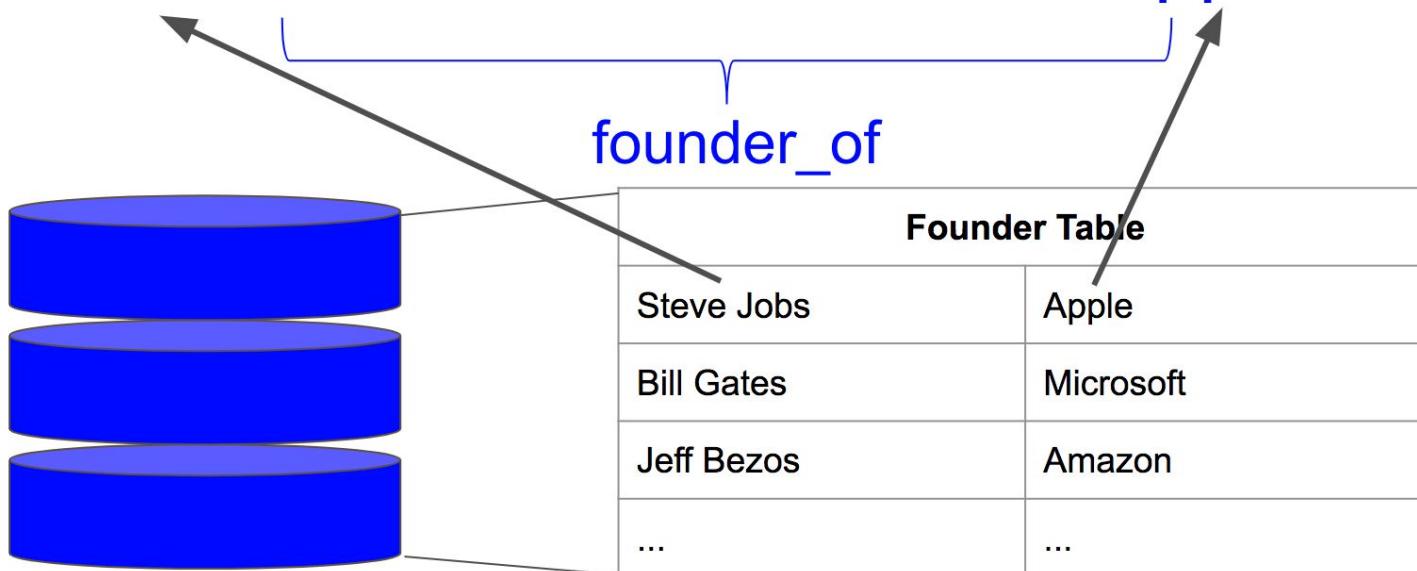
Steve Jobs was the founder of Apple.



Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

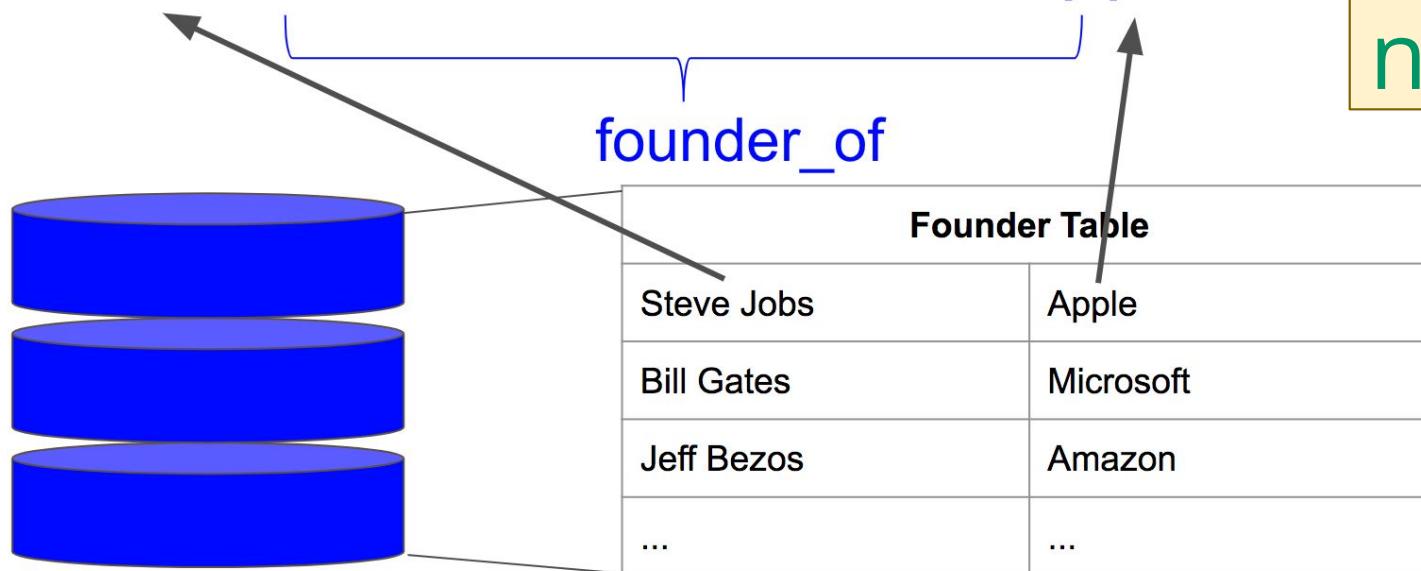
Steve Jobs was the founder of Apple.



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”

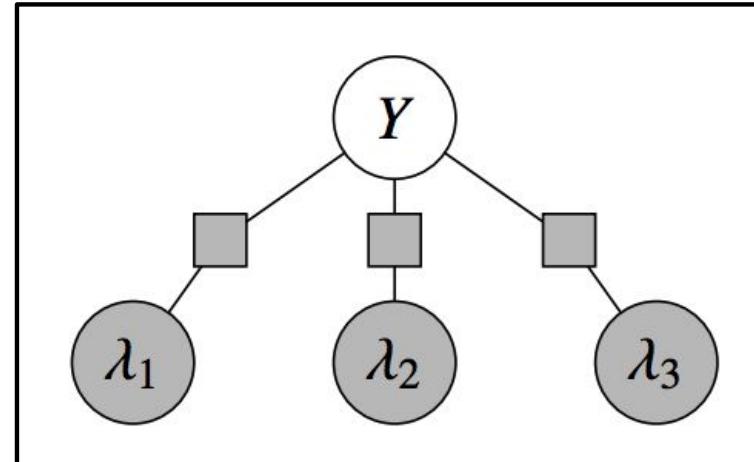
Generative model to “de-noise” training data

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

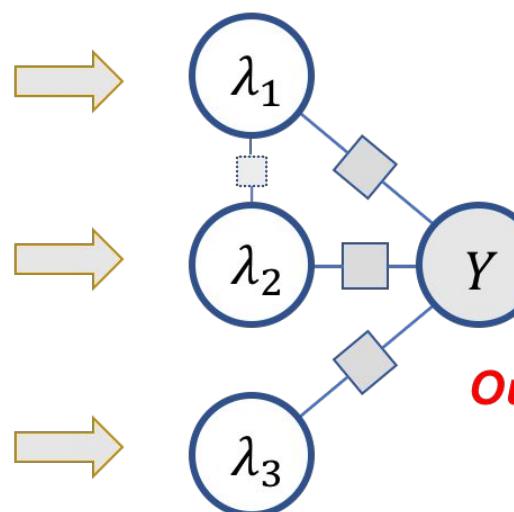


```
def lf1(x):
    cid = (x.chemical_id,
    x.disease_id)
    return 1 if cid in KB else 0
```

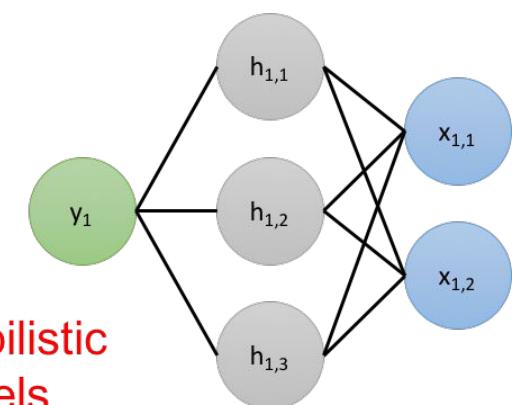
```
def lf2(x):
    m = re.search(r'.*cause.*',
    x.between)
    return 1 if m else 0
```

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

Generative Model



**Noise-Aware
Discriminative Model**



Output: Probabilistic
Training Labels

Snorkel

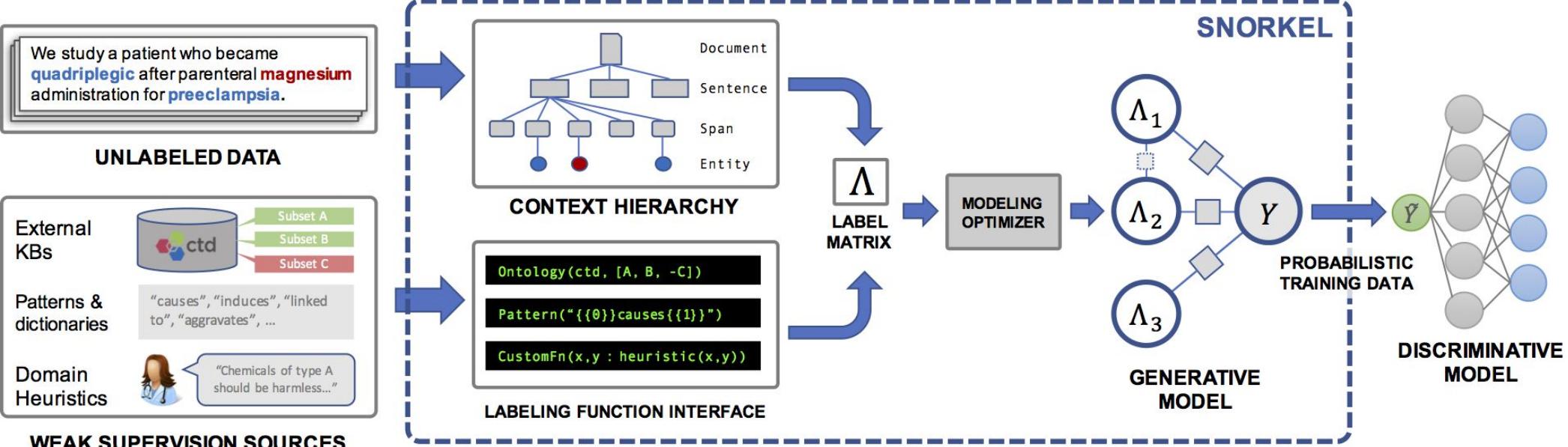


Image via https://hazyresearch.github.io/snorkel/blog/snorkel_programming_training_data.html

```
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

How can we discover new relations?

OpenIE (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 4:
Unknown
unknowns

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

Knowledge Collection from Semi-structured Text

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



Outline

- Introduction (30 minutes)
- Part I: Unstructured text (45 minutes)
- Break (30 minutes)
- Part Ib: Unstructured text: Methods (15 minutes)
- **Part II: Semi-structured text** (45 minutes)
- Part III: Tabular text (15 minutes)
- Part IV: Multi-modal extraction (30 minutes)
- Conclusion and future directions (15 minutes)



Questions we will answer in this section

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 /10 990,317 Rate This

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

NOTHING CAN SEPARATE
COPPER COULD BETWEEN THEM

LEONARDO DICAPRIO KATE WINSLET

TITANIC

2:11 | Trailer

25 VIDEOS 638 IMAGES

A seventeen-year-old aristocrat falls in love with a kind but poor artist aboard the luxurious, ill-fated R.M.S. Titanic.

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

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

The screenshot displays two product cards side-by-side on a website.

Samsung 1TB T5 Portable Solid-State Drive (Black)
B&H # SAMUPA1T0BAM • 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

Add to Cart

Add to Wish List

SmartGift Available

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

Add to Cart

Add to Wish List

SmartGift Available

Updated Model Available

What is a semi-structured website?

Personal Details

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 | 1 Article | 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?

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

Edit

Hide ▲

2018

15-2018

Edit

2017/I

Topic entity

Relations as key-value pairs

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 the BollywoodMDB.com homepage with a search bar for 'Movies / Celebrities'. Below it, a movie card for 'Made In China' (2019) is displayed, listing cast and crew information.

Made In China (2019)

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

Bollywood films

The screenshot shows the NMDB (Now Playing Movie Database) website. A movie card for 'Twisted (Short film)' is displayed, showing a trailer thumbnail and basic movie details.

Twisted (Short film)

Daniel Ademinokan's "TWISTED" Trailer

Year of production: 2014
Running Time: 212 mins
Written by: Daniel Ademinokan
Produced by: Daniel Ademinokan
Directed by: Daniel Ademinokan
Starring: Stella Damasus, Rob Byrnes, Matt Meissen and David Ademinokan

Nigerian films

... and many many long-tail websites

The screenshot shows the NFB (National Film Board of Canada) website. A movie card for 'Abegweit' is displayed, showing a thumbnail, basic movie details, and a detailed credits section.

Abegweit

Serge Morin
1998 | 1 h 11 min
cc
CAMPUS

CREDITS

DIRECTOR	SCRIPT	PRODUCER	CAMERA
Serge Morin	Serge Morin	Pierre Bernier	Marc Poulin
SOUND	EDITING	Diane Poitras	SOUND EDITING
Georges Hannan	Fernand Bélanger	Serge Bovin	Fernand Bélanger
NARRATOR	MUSIC	Jean Paul Viard	Claude Longlois
Alex Madsen	Richard Gibson		

A day-to-day record of the construction of the Confederation Bridge reveals what one of the bridge-- to be part of and their changed by the ill have on eeting of .

Canadian films

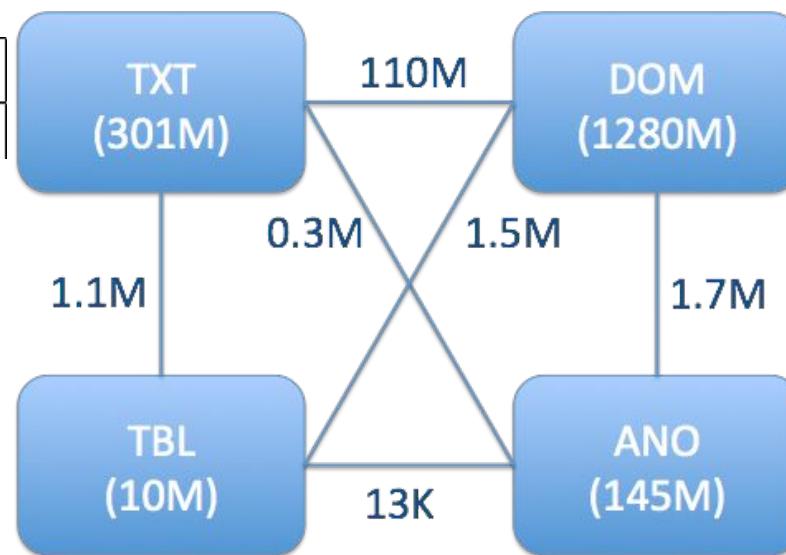
Characteristics of semi-structured websites

- **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

↑
Accuracy is still low

The screenshot shows the IMDb profile page for Rita Moreno. At the top, there's a navigation bar with links for 'Movies, TV & Showtimes', 'Celebs, Events & Photos', 'News & Community', and 'Watchlist'. Below the navigation is a search bar and a 'Top 5000' badge. The main content area features a large photo of Rita Moreno, her name 'Rita Moreno' in a prominent box, and her title 'Actress | Soundtrack'. Below her name are links to 'View Resume' and 'Official Photos'. A detailed bio follows, mentioning her acting career spanning six decades, her Oscar win, and her birthplace in Humacao, Puerto Rico. Her birth date is listed as 'December 11, 1931'. The page also includes sections for 'Official Sites', 'Alternate Names', 'Height' (5' 2 1/2" (1.59 m)), 'Did You Know?' (with a personal quote), 'Trivia' (about being the mother of Fernanda Gordon), and 'Star Sign' (Sagittarius). The entire page is framed by a light gray border.

What is semi-structured website extraction?

Extraction of structured data records 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



Horizontal vs.
vertical layout

2 1998 Academy Award® Nominations! BEST ACTRESS, Fernanda Montenegro
CENTRAL STATION (1998)
Cast
Vinícius De Oliveira as Josue
Fernanda Montenegro as Dora
Soia Lira as Ana
Marilia Pera 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 is Tom Cruise's profile:

Tom Cruise

Highest Rated: 97% **Mission: Impossible - Ghost Protocol**
Lowest Rated: 5% **Cocktail (1988)**

Birthday: Jul 3, 1962
Birthplace: Syracuse, New York

Attended 15 schools in 14 years. Studied for Franciscan seminary. Joined the Church of Sc...
Newman got him interested in motor racing w... Money, and he went on to race on Newman's More

The screenshot shows Tom Cruise's IMDBPro profile:

Tom Cruise
Actor | Producer | Soundtrack

0:50 | Trailer

In 1976, if you had told fourteen year-old Francisco Mapother IV that one day in the not too distant future he would be one of the 100 movie stars of all time, he would have probably...
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

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

Central Station (1998) ★ 8.0 /10 33,811 Rate This

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) ★ 8.1 /10 22,171 Rate This

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



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 »

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

Star Wars: Episode VIII - The Last Jedi (2017) ★ 7.0 /10 522,260 Rate This

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



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

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

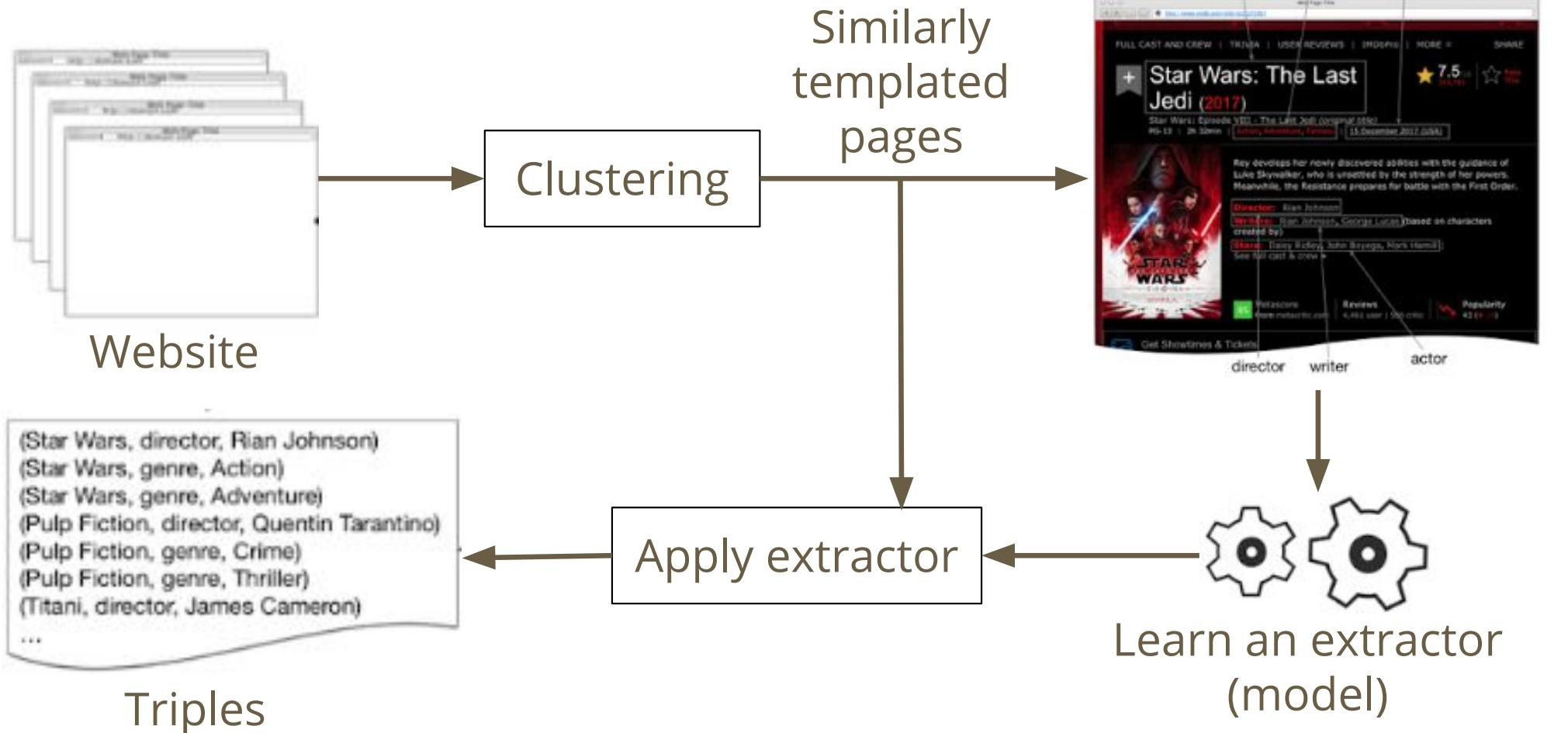


Short Answers

- **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.



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 may correspond to different DOM tree nodes

The image shows the movie poster for "Central do Brasil" (1998) on the left, featuring a woman and a young boy. To the right is a large, close-up still from the film showing a young boy's face. A play button icon is overlaid on the still. Below the poster, the text "1:53 | Trailer" is visible. At the top of the page, there are navigation links: "FULL CAST AND CREW", "TRIVIA", "USER REVIEWS", "IMDbPro", "MORE", and "SHARE". On the right side, there is a star rating of 8.0 from 33,811 reviews, a "Rate This" button, and a red box highlighting the "Stars" section.

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 »

Challenges to wrapper induction

Optional/missing sections: Same DOM node may correspond to different relations

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
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

From the Director of *The Thin Red Line*, *Jarhead*, *Flags of Our Fathers*
"IF THERE IS ONE MOVIE THAT OUGHT TO BE SEEN BY EVERY LEADER OF COUNTRIES AROUND THE WORLD AT THIS TREACHEROUS HISTORICAL MOMENT, IT IS 'THE FOG OF WAR' ERROL MORRIS'S DOCUMENTARY ON THE FORMER DEFENSE SECRETARY, ROBERT S. MCNAMARA." —Audrey Parker, *THE NEW YORK TIMES*

"SMART AND COMPLETELY FASCINATING!" —Peter Travers, *ROLLING STONE*

"STUNNING! SUPERBLY MADE!" —Peter Travers, *ROLLING STONE*

"SPLendid! A Masterpiece!" —Peter Travers, *ROLLING STONE*

THE FOG OF WAR An Errol Morris Film

2:07 | Trailer 2 VIDEOS | 11 IMAGES

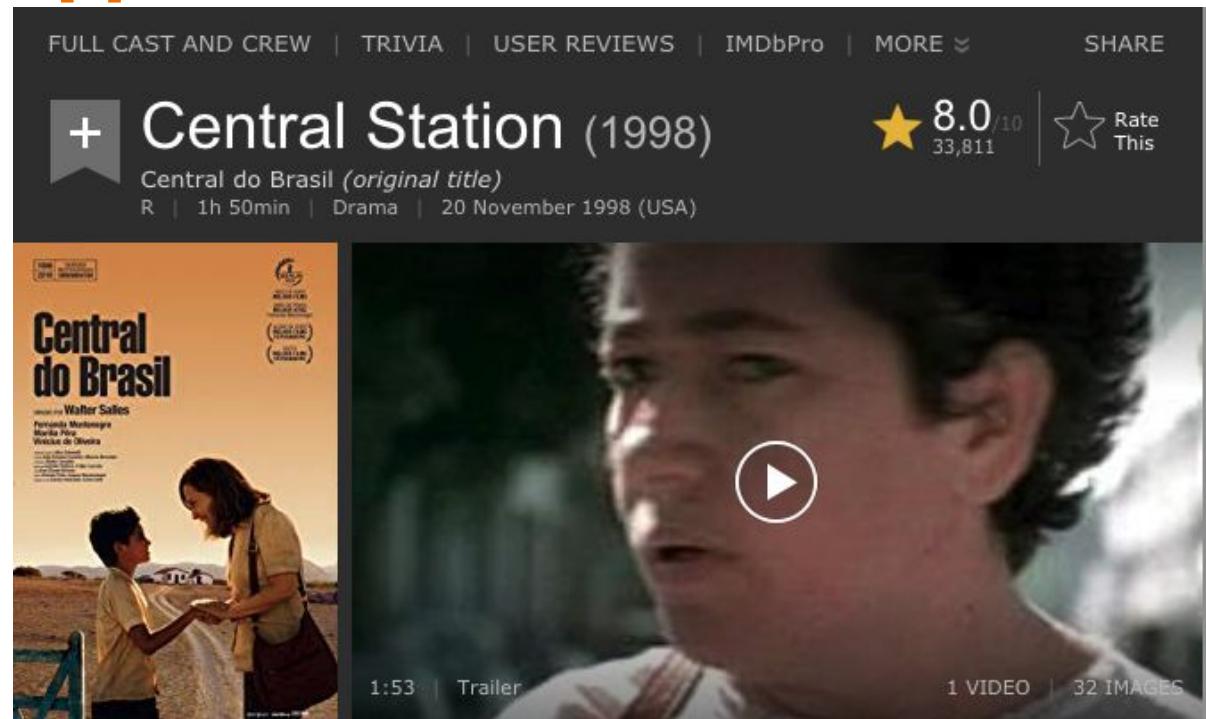
The story of America as seen through the eyes of the former Secretary of Defense under President John F. Kennedy and President Lyndon Baines Johnson, Robert McNamara.

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



The image shows a screenshot of the IMDb movie page for "Central Station (1998)". The page includes the title, rating (8.0/10), plot summary, director information (Walter Salles), and other credits. A red box highlights the "Director" section.

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

Walter Salles

Fernanda Montenegro
Marcos Bernstein
Vinícius de Oliveira

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.

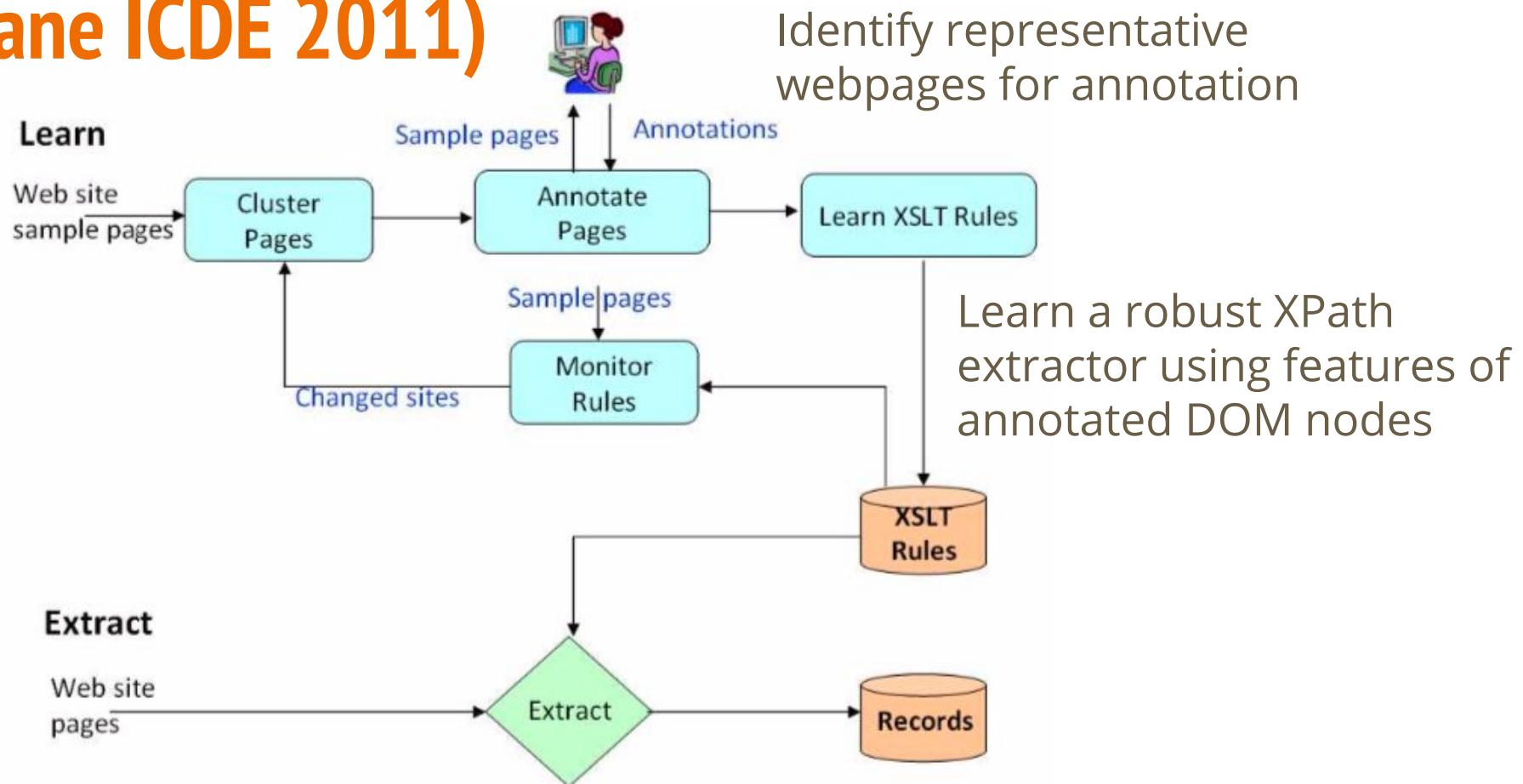
Director: Walter Salles

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

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

1:53 | Trailer | 1 VIDEO | 32 IMAGES

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

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 XPaths as rules

'Price' on www.amazon.com

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

'Forum title' on www.city-data.com

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

'Address' on www.hotels.com

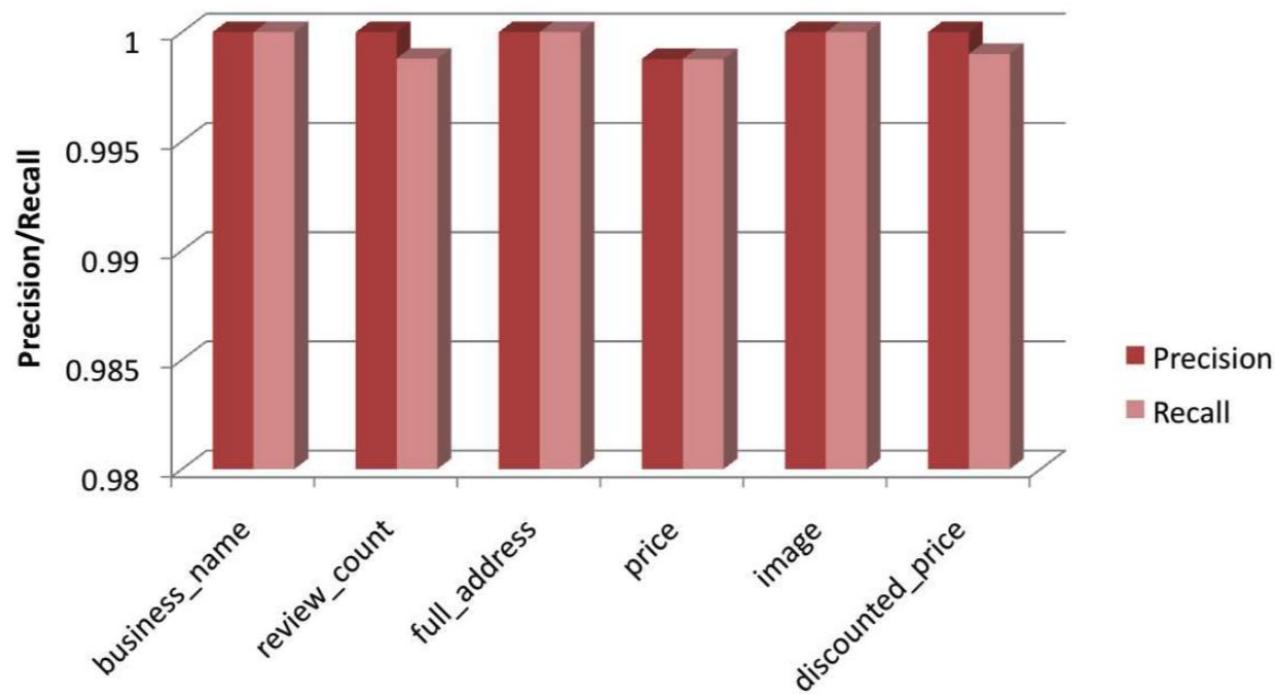
```
//node()[@class="adr"]
```

'Image' on www.alibaba.com

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

Performance

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

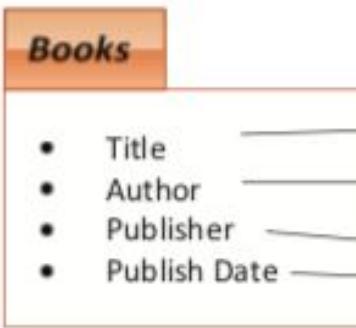
Cons:

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

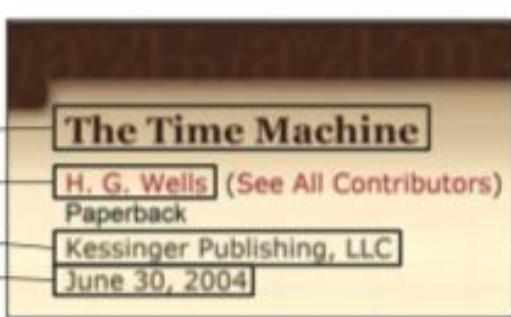
How can we extract from ALL websites in a domain given ONE or few labeled websites?

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

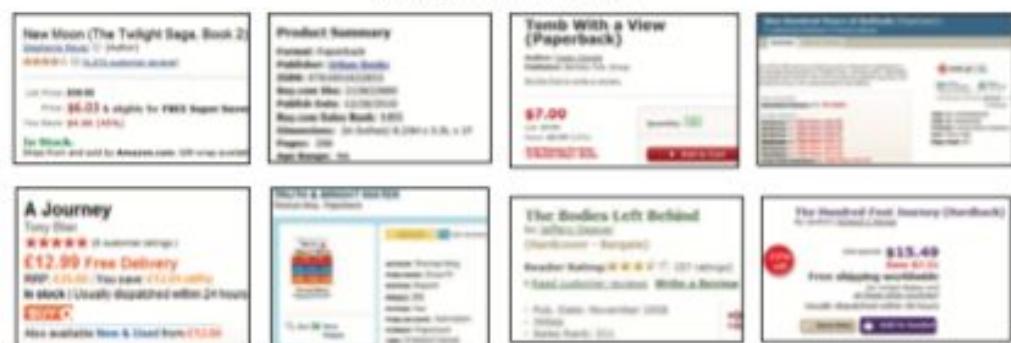
Verticals and Attributes



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'

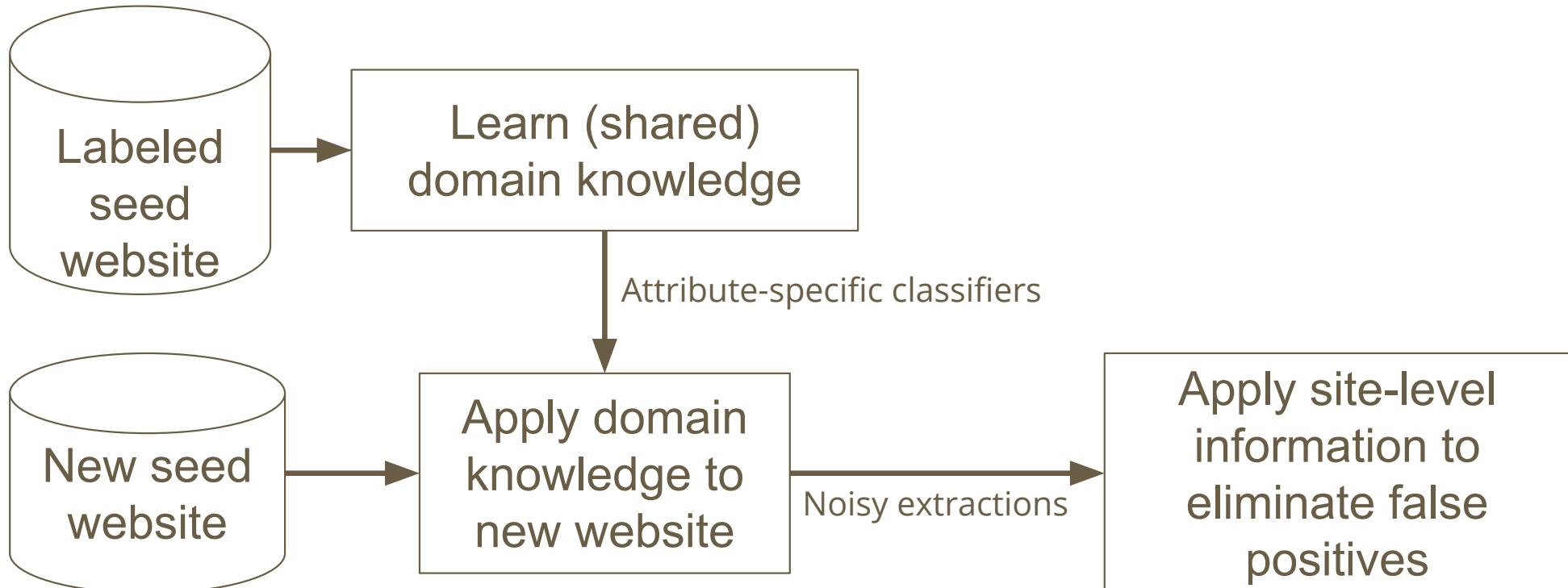
Inter-attribute layout consistency

Some attributes are often close to each other on the page
e.g. title and author

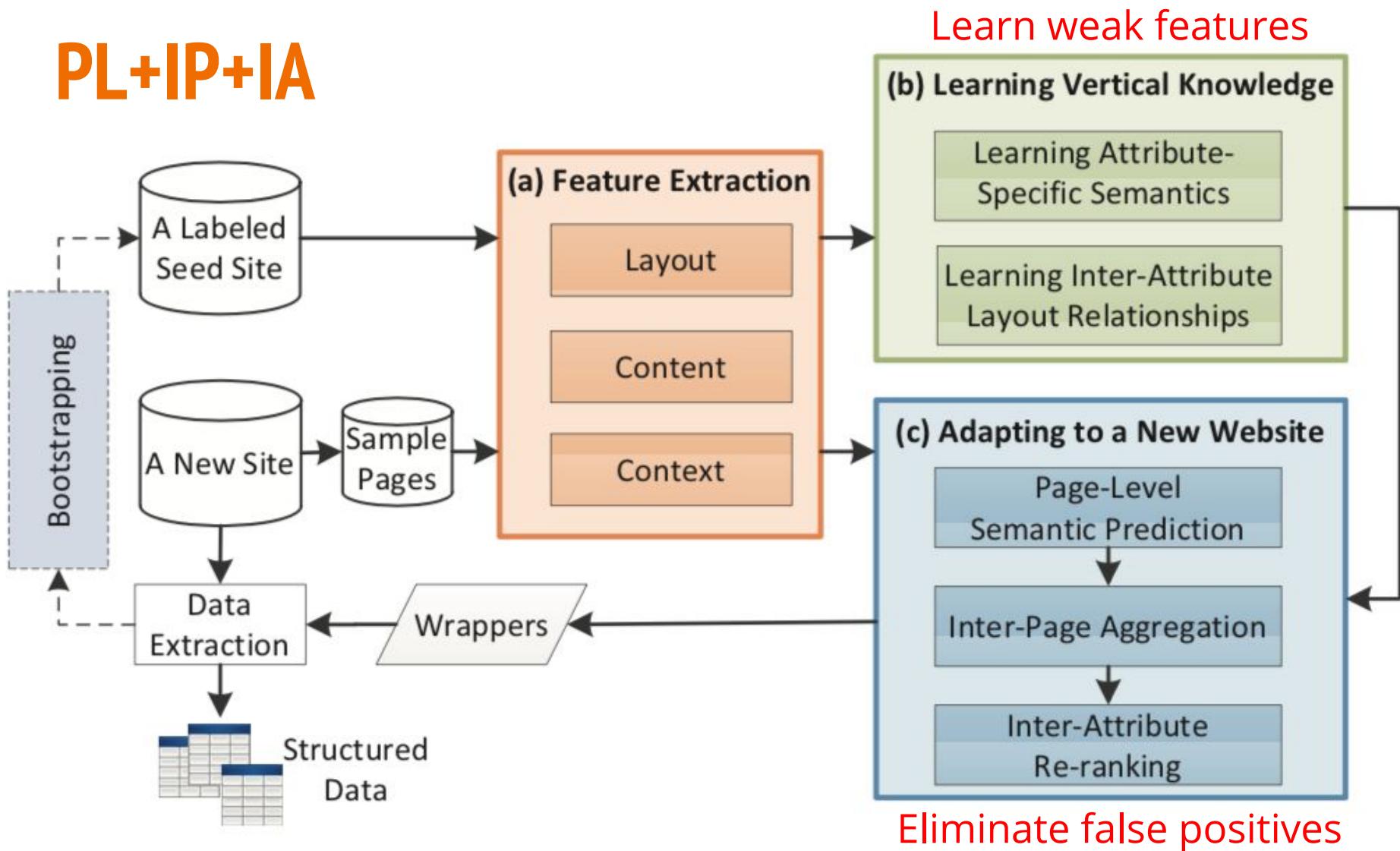


the darker cells indicate attributes are in close vicinity

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
Our Solution	0.843	0.860	0.868	0.884	0.886
Our Solution (Bootstrap)	0.843	0.856	0.861	0.859	0.865
SSM	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?**

How 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.



Caveat: The annotation may be noisy.

Rita Moreno

Actress | Soundtrack

Born: December 11, 1931 in Humacao, Puerto Rico

Height: 5' 2½" (1.59 m)

Star Sign: Sagittarius

Automatic annotations of KB predicates

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

The screenshot shows the IMDb movie page for 'Do the Right Thing' (1989). It includes the movie poster, a trailer thumbnail, and a summary: '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.' Below the summary, there are credits: Director: Spike Lee, Writer: Spike Lee, Stars: Danny Aiello, Ossie Davis, Ruby Dee. A green box highlights the 'Director: Spike Lee' entry. An arrow points from this box to the 'Real mention' label.

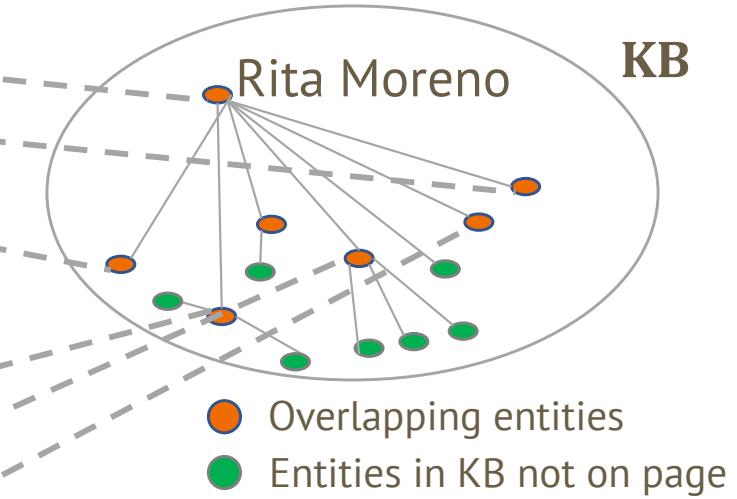
Below this, the 'More Like This' section is shown for the movie 'Crooklyn'. It features the movie poster and a summary: 'Spike Lee's vibrant semi-autobiographical portrait of a school teacher, her stubborn jazz musician husband and their five kids living in Brooklyn in 1973.' Below the summary, there are credits: Director: Spike Lee, Stars: Alfre Woodard, Delroy Lindo, ... A green box highlights the 'Director: Spike Lee' entry. Another arrow points from this box to the 'Spurious mention' label.

Challenges

- Entity linking problem
- Distant supervision would require examination of all entity mention pairs as candidates for annotation
 - ⇒ computationally infeasible
 - ⇒ can lead to spurious annotations
- Disambiguating relations involving same entity pair
- Distinguishing between real and spurious relation mentions

Topic entity annotation

The screenshot shows Rita Moreno's profile on IMDb. At the top, there's a navigation bar with links for 'Trailer / Trailers', '& Showtimes', '& Photos', 'Community', and 'Watchlist'. Below the navigation is a large photo of Rita Moreno. Her name, 'Rita Moreno', is highlighted with a green box, and 'Actress' and 'Soundtrack' are listed below it. A 'Top 5000' badge is visible. The bio text discusses her acting career, mentioning she won an Oscar, Emmy, Tony, and Grammy. It also notes her birth on December 11, 1931, in Humacao, Puerto Rico. A 'See full bio' link is present. Below the bio, 'Born: December 11, 1931 in Humacao, Puerto Rico' is highlighted with an orange box. There are also links for 'More at IMDbPro' and 'Contact Info'. The 'Filmography' section is titled 'Actress (155 credits)' and includes a list of roles, such as 'Nina's World' (TV Series) and 'Abuelita', with some entries highlighted by orange boxes. A 'Show all' button is available.



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

Relation annotation

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

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

Rita Moreno Top 5000

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 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)

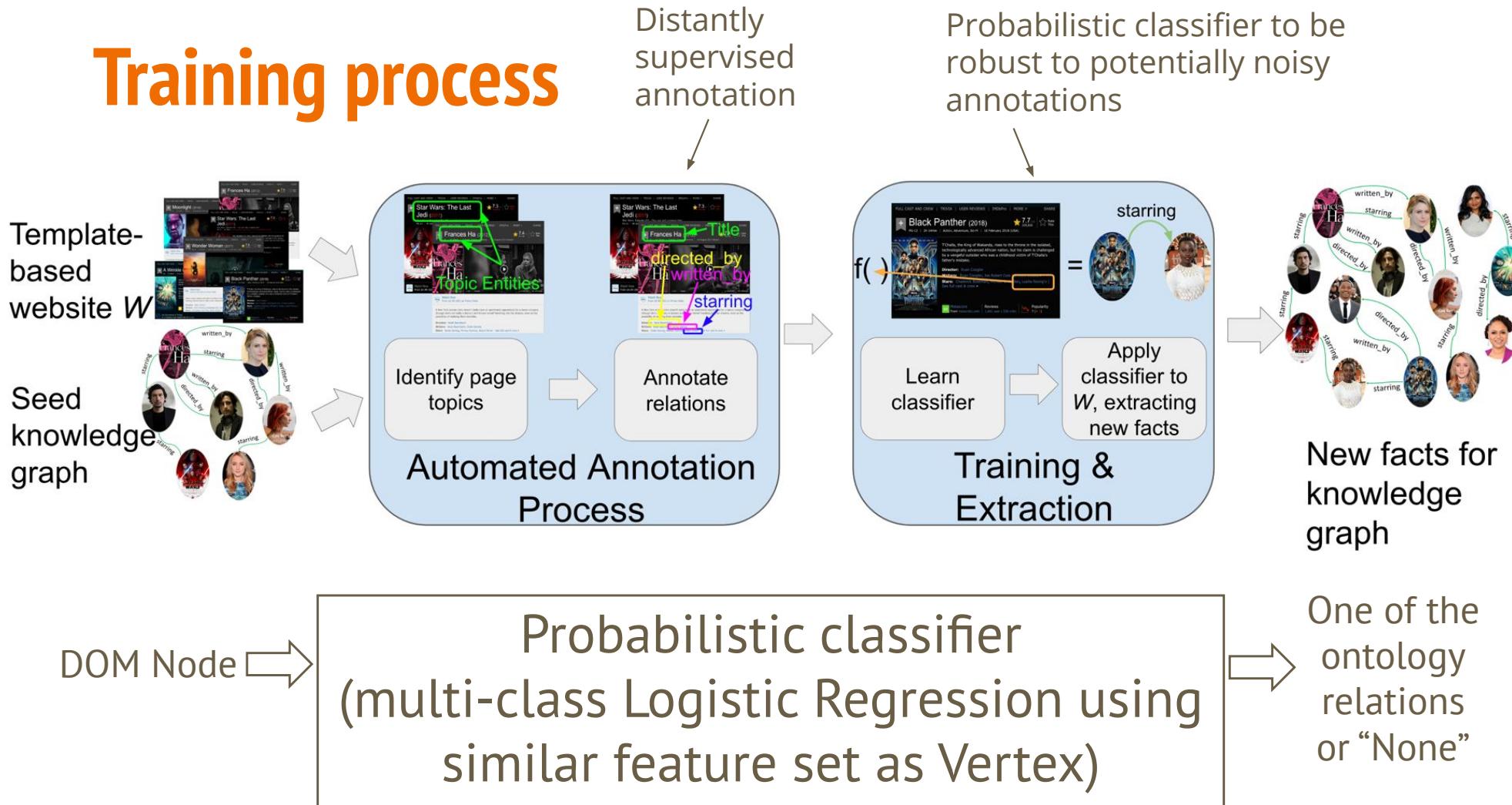
Did You Know?

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](#)

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	0.98	0.98	0.97
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 ^a	0.91 ^a	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	1.00	0.94
CERES-Baseline	no	NA ^b	0.78	0.72	0.27
CERES-Topic	no	0.99^a	0.97	0.96	0.72
CERES-Full	no	0.99^a	0.98	0.94	0.76

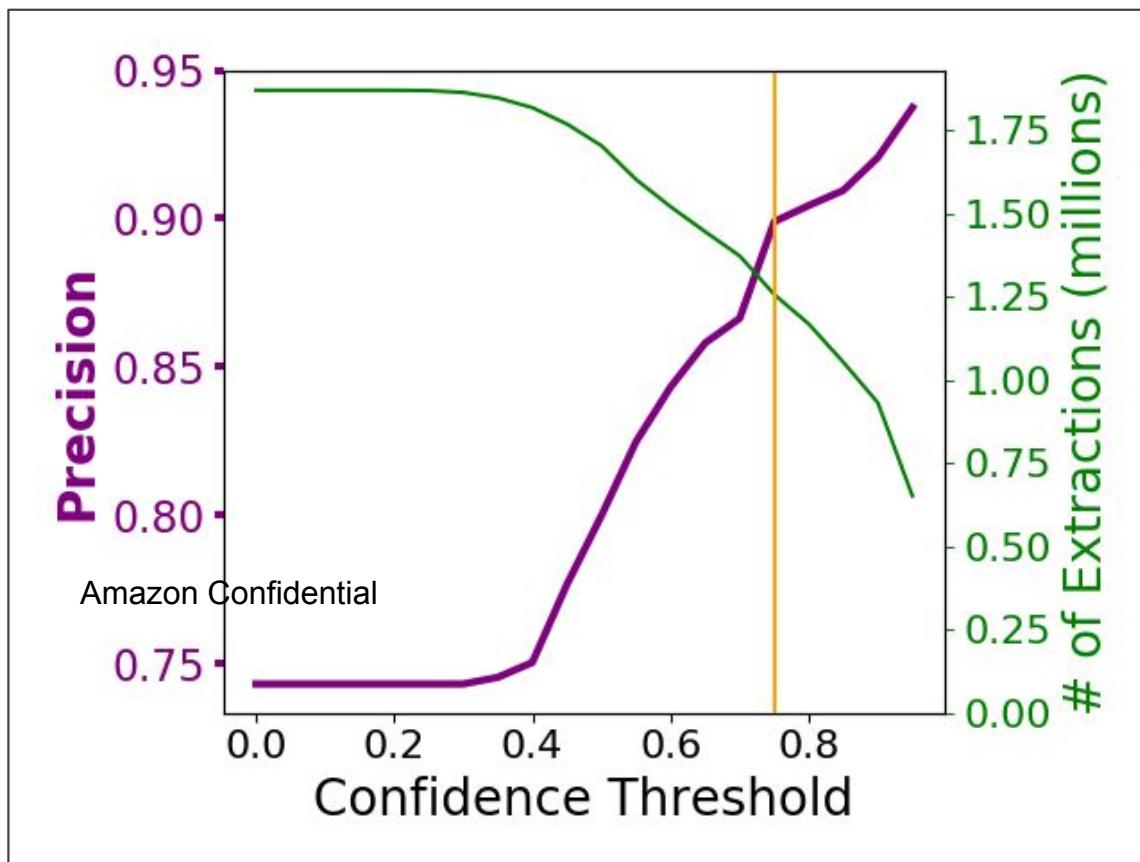
Domain having low overlap with seed data performs suboptimally

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 »](#)

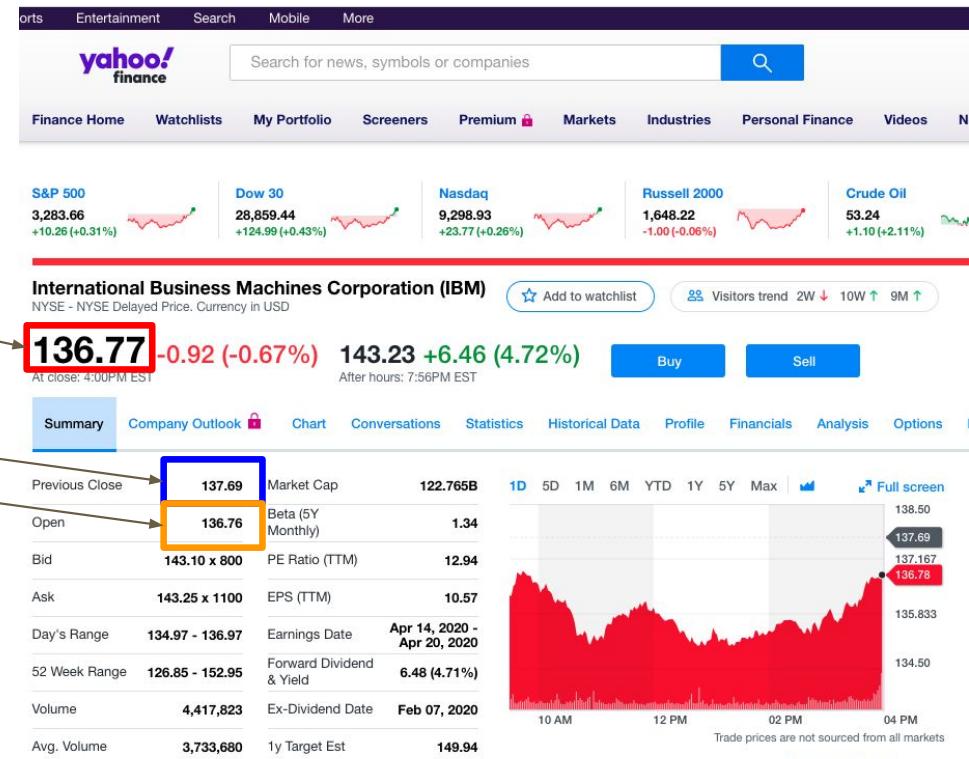
Box Office

Edit

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 correct extractors?

International Business Machines Corp. IBM.N



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)

S_G : Google



TICKER	HIGH	LOW	...
AAPL	259.4	253.4	...
CSCO	21.09	20.73	...
...	

$e_G(\pi_G(\sigma_G(\mathcal{H})))$

S_R : Reuters



TICKER	PRICE	VOLUME	CAP	...
AAPL	256.9	29,129,000	506B	...
CSCO	20.91	34,770,000	226B	...
...	

$e_R(\pi_R(\sigma_R(\mathcal{H})))$

S_Y : Yahoo!



TICKER	VOLUME	CAP	...
AAPL	29,1M	506B	...
CAT	7,7M	56B	...
...	

$e_Y(\pi_Y(\sigma_Y(\mathcal{H})))$

\mathcal{H}

TICKER	PRICE	HIGH	LOW	VOLUME	CAP	...
AAPL	256.88	259.40	253.35	29,129,032	506B	...
CSCO	20.91	21.09	20.72	34,777,673	111.5B	...
CAT	60.76	62.42	60.05	7,709,405	56B	...

How do we design an extractor that leverages the redundancy of data?

Key intuition:

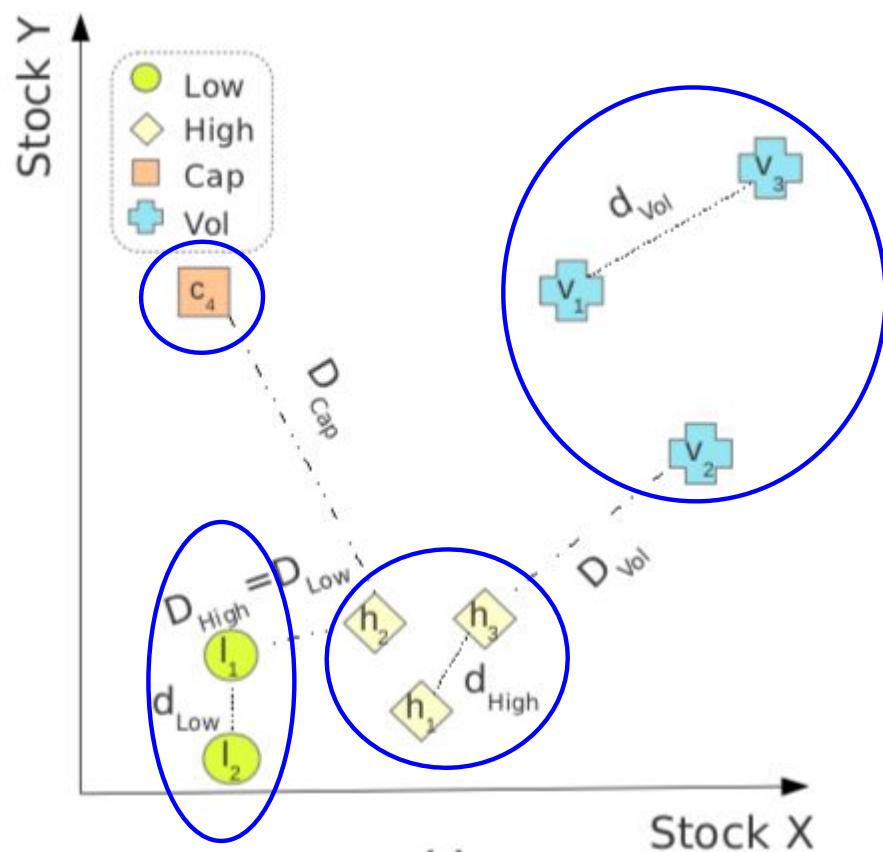
Assuming we had extractors for different overlapping websites, a *correct* extractor will likely extract data that match with those extracted from at least one other *correct* extractor from a different website

Challenges:

- How do generate extractors in the first place?
- How to differentiate extractors of different attributes?

Leverage key properties of semi-structured websites

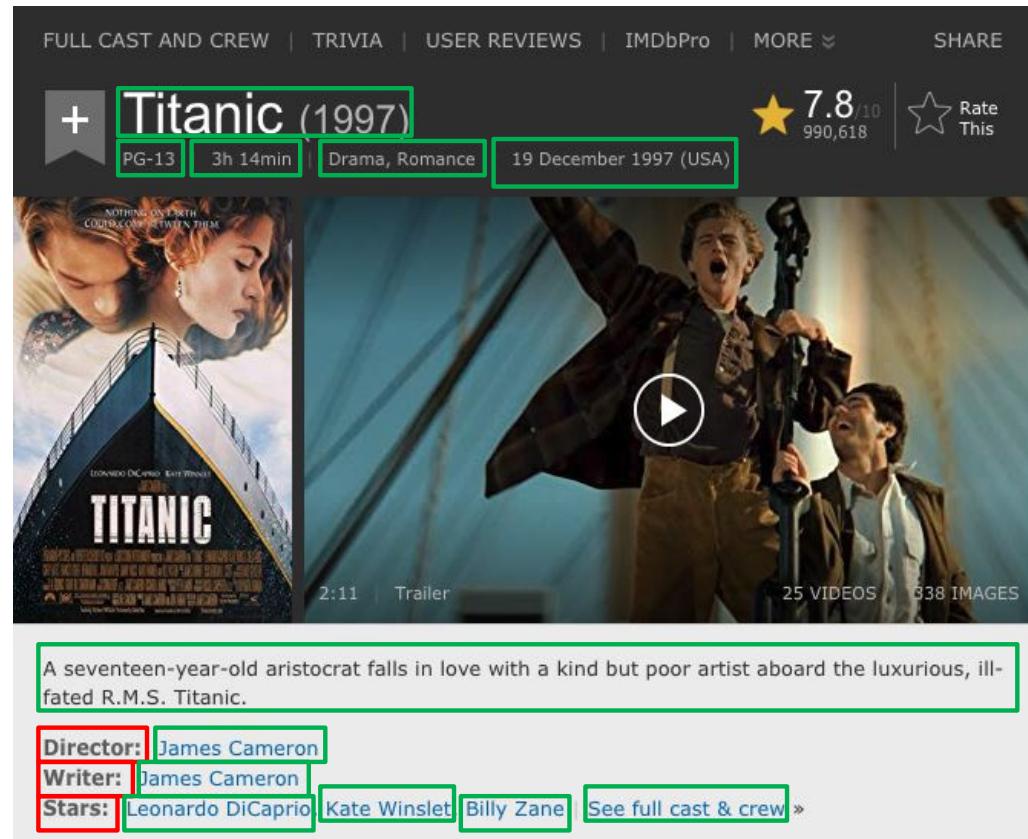
1. **Local consistency:** A website does not publish different values for the same attribute
2. **Separable semantics:** Attributes with similar semantics are *closer* than attributes with different semantics



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



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

Domain	#p	#o	d	P	R	F	Time
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)

how to extract
these new
relations?

The screenshot shows a movie details page with the following sections and extracted relations:

- 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

Challenges in Open IE from semi-structured website

The screenshot shows a semi-structured website for a movie. It includes sections for Genres, Motion Picture Rating (MPAA), Parents Guide, Details (Country, Language, Release Date, Also Known As, Filming Locations), and Box Office (Budget, Gross USA). The 'Drama' genre is highlighted in yellow.

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

Ceres distant supervision enables us to match **objects**, but ..

1. How do we identify their **relation strings?**
2. How do we identify **new relation strings** and their **objects?**

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. “Language” vs. “English”

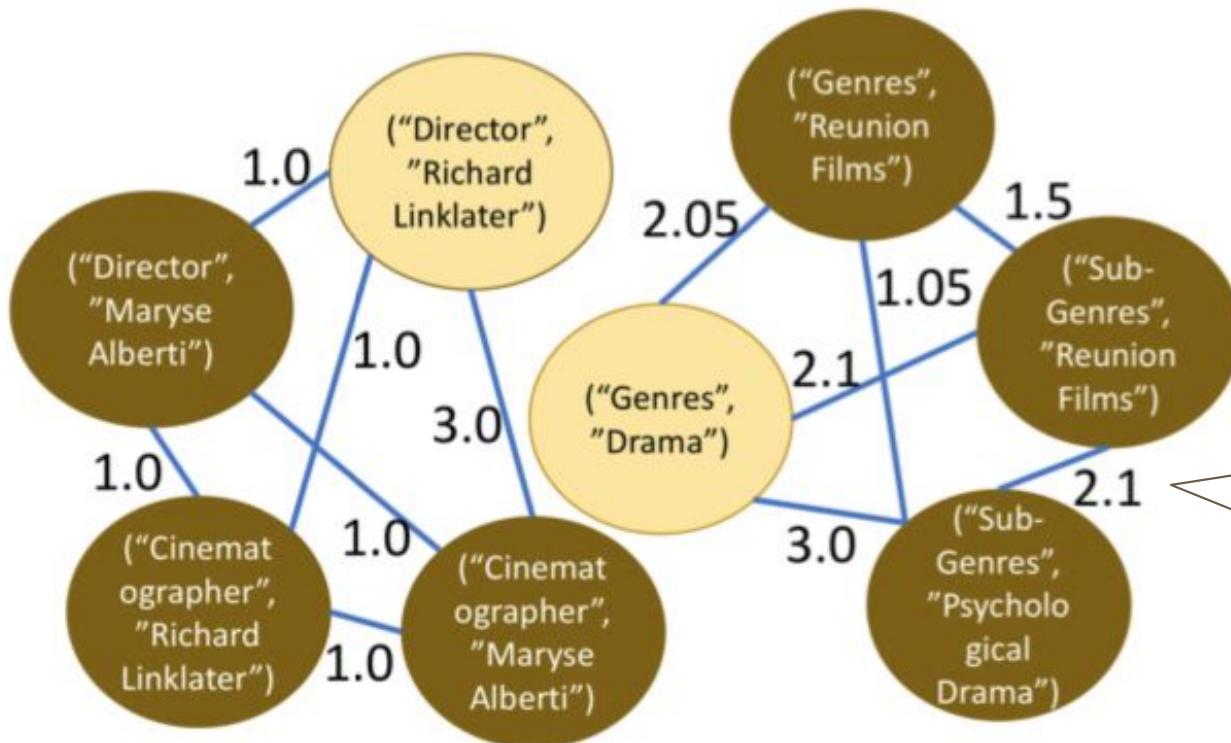
Two main steps:

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



How do we identify new predicate strings? --

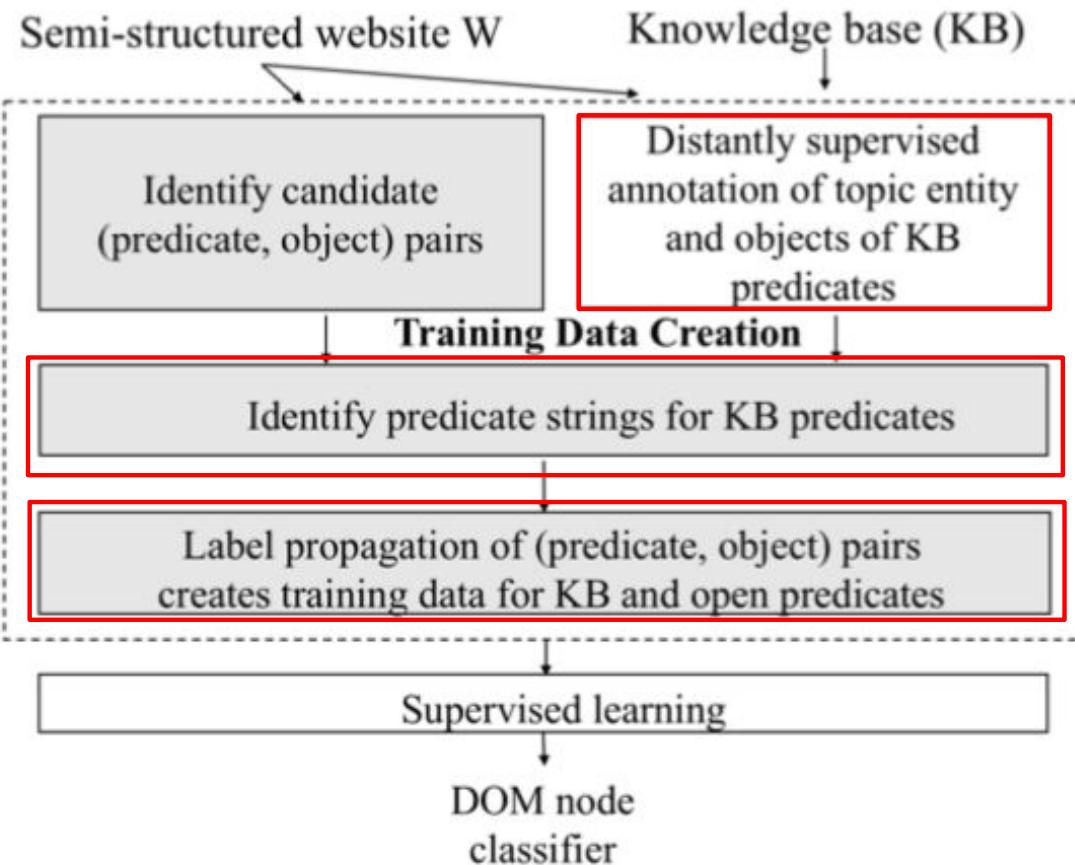
Graph-based label propagation (Lockard, NAACL'19)



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 detail page with various metadata fields. Several fields are highlighted with colored boxes: "Genres" (yellow), "Country" (purple), "Language" (green), "Release Date" (red), "Also Known As" (red), and "Filming Locations" (green). The "Genres" section lists "Comedy", "Drama", and "Romance". The "Release Date" is listed as "25 June 1993 (USA)" with a "See more" link. The "Filming Locations" field contains the address "1517 Pike Place, Seattle, Washington, US".

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	0.31
OpenCeres	0.77	0.68	0.74	0.48	0.65	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

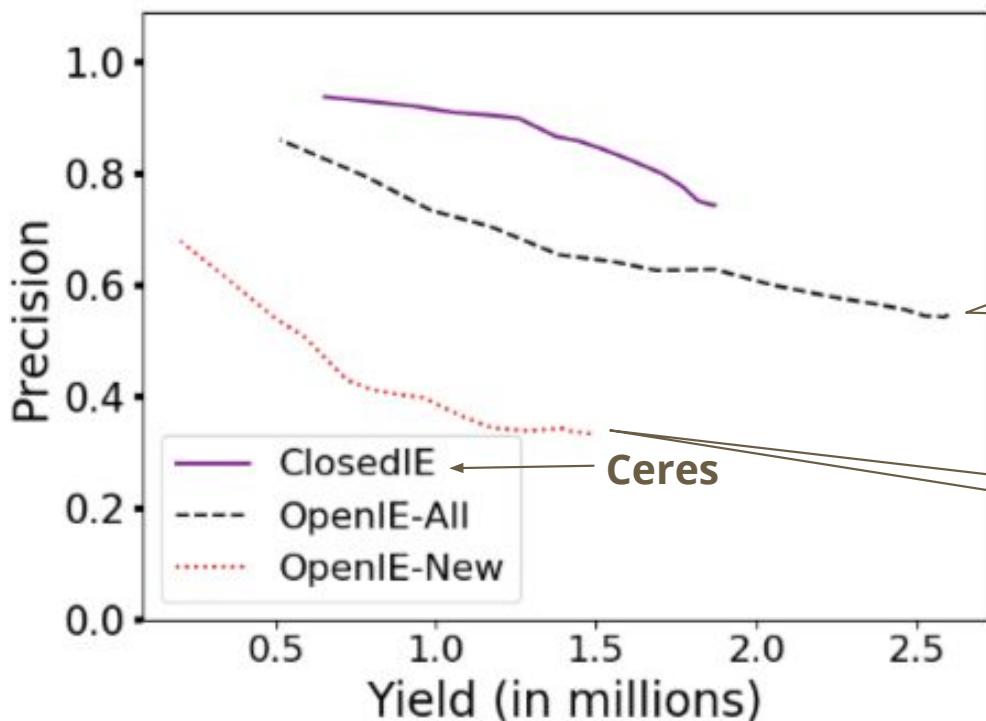
Performance

- **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

	Movie	NBA Player	University
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 on a large Common Crawl dataset



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).

Outline

- Introduction (30 minutes)
- Part Ia: Unstructured text (30 minutes)
- Break (30 minutes)
- Part Ib: Unstructured text: Methods (15 minutes)
- Part II: Semi-structured text (45 minutes)
- **Part III: Tabular text** (15 minutes)
- Part IV: Multi-modal extraction (30 minutes)
- Conclusion and future directions (15 minutes)



Knowledge Collection from Tabular Text

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



Outline

- Introduction (30 minutes)
- Part I: Unstructured text (45 minutes)
- Break (30 minutes)
- Part Ib: Unstructured text: Methods (15 minutes)
- Part II: Semi-structured text (45 minutes)
- **Part III: Tabular text** (15 minutes)
- Part IV: Multi-modal extraction (30 minutes)
- Conclusion and future directions (15 minutes)



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 ^[a]	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 ^[a]	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 ^[a]	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 ^[a]	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 ^[b]	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 the URL "http://t...". Below the title, it says "The 50 Greatest Cartoons" and "from *The 50 Greatest Cartoons* by Jerry Beck (ed.) (Atlanta, Turner Publishing, 1994) ISBN# 1-878685-49-X". A numbered list of 14 cartoon titles follows:

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)

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 ^[a]	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 ^[a]	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 ^[a]	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 ^[a]	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

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	Downington, Pennsylvania
Baton Rouge	Steak	Imvescor	Montreal, Quebec

We assume horizontal tables in this tutorial

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)

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

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

93% of web tables have 2-9 attributes

Very few tables have large number of attributes

Tables have much greater diversity in row counts

Why extract from web tables?

- Table search based on keywords

Google | STATA - city population - Mozilla Firefox

File Edit View History Bookmarks Tools Help

city population Submit

done!

Tables 1 - 10 of 15057 found. (22.005 seconds elapsed)

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

export to... show entire table

Rank(1)	City / Urban area(2)	Country	Population(3)	Land area (in sqKm)(4)	Density (people per sqKm)(5)
1(1)	Tokyo Yokohama	Japan	33,200,000(3.32E7)	8,091,000(3)	4,100(4.10E3)
2(2)	New York Metro	USA	17,800,000(1.78E7)	8,091,000(3)	2,198(2.19E3)
3(3)	Sao Paulo	Brazil	17,700,000(1.77E7)	1,065,000(2)	16,500(1.65E4)
4(4)	Seoul Incheon	South Korea	17,600,000(1.76E7)	1,046,000(2)	16,800(1.68E4)
5(5)	Mexico City	Mexico	17,400,000(1.74E7)	2,073,000(2)	8,400(8.40E3)
6(6)	Oakland/SF Bay	USA	16,425,000(1.64E7)	2,584,000(2)	6,400(6.40E3)
7(7)	Manila	Philippines	14,700,000(1.47E7)	1,386,000(2)	10,600(1.06E4)
8(8)	Mumbai	India	14,350,000(1.46E7)	4,040,000(3)	3,550(3.55E3)
9(9)	Delhi	India	14,300,000(1.46E7)	1,286,000(2)	11,000(1.10E4)
10(10)	Jakarta	Indonesia	14,250,000(1.46E7)	1,386,000(2)	10,400(1.04E4)
11(11)	Lagos	Nigeria	13,400,000(1.34E7)	7,987,000(3)	1,700(1.70E3)
12(12)	Kolkata	India	13,700,000(1.37E7)	8,109,000(3)	1,680(1.68E3)
13(13)	Cairo	Egypt	12,200,000(1.22E7)	1,286,000(2)	9,400(9.40E3)
14(14)	Los Angeles	USA	11,780,000(1.178E7)	4,320,000(2)	2,700(2.70E3)
15(15)	Buenos Aires	Argentina	11,200,000(1.12E7)	2,286,000(2)	4,800(4.80E3)
16(16)	Rio de Janeiro	Brazil	10,800,000(1.08E7)	1,386,000(2)	7,800(7.80E3)
17(17)	Moscow	Russia	10,800,000(1.08E7)	2,186,000(2)	4,900(4.90E3)
18(18)	Shanghai	China	10,200,000(1.02E7)	7,467,000(3)	1,360(1.36E3)
19(19)	Karachi	Pakistan	9,050,000(9.05E6)	9,162,000(3)	980(9.80E2)
20(20)	Paris	France	9,440,000(9.44E6)	2,724,000(2)	3,480(3.48E3)
21(21)	Istanbul	Turkey	9,100,000(9.10E6)	1,196,000(2)	7,600(7.60E3)
22(22)	Nagoya	Japan	8,230,000(8.23E6)	2,874,000(2)	2,870(2.87E3)
23(23)	Beijing	China	8,110,000(8.11E6)	7,467,000(3)	1,080(1.08E3)
24(24)	Chicago	USA	8,050,000(8.05E6)	5,486,000(3)	1,440(1.44E3)
25(25)	London	UK	8,278,000(8.278E6)	1,823,000(2)	4,540(4.54E3)

ESTIMATING CITY POPULATIONS

export to... show entire table

REGION	People per Hectare	Margin of Error
Cities of Antiquity	100	10-70%
Cities of Islam	280	30-25%
Cities of Europe (Greek and Roman)	100-110	20%
Total	100-110	10%

Google

city population

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?

Why extract from web tables?

- Table search based on keywords
- Schema autocomplete tool for database designers
 - Suggest ‘company’, ‘rank’ and ‘sales’ as attributes to add to a schema for ‘stock-symbol’ as an input



Why extract from web tables?

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



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

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

How do I 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. μ
- cell strlen stddev. σ
- 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.

However, we are far from this goal!

Aspects of semantics recovery pursued thus far:

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	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	1991	4000	

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	
Bubba Gump Shrimp Co.	Seafood	Bubba Gump Shrimp Co.	Columbia, Georgia	1984	4000	

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 IJDAR'04)
- Table extraction (Gatterbauer WWW'07)
- WebTables (Cafarella WebDB'08, VLDB'08)

Metadata recovery

- Subject column discovery (Venetis VLDB'11)
- 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	Fraction of cells with unique content
2	Fraction of cells with numeric content
3	Average number of letters in each cell
4	Average number of numeric tokens in each cell
5	Variance in the number of date tokens in each cell
6	Average number of data tokens in each cell
7	Average number of special characters in each cell
8	Average number of words in each cell
9	Column index from the left
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%



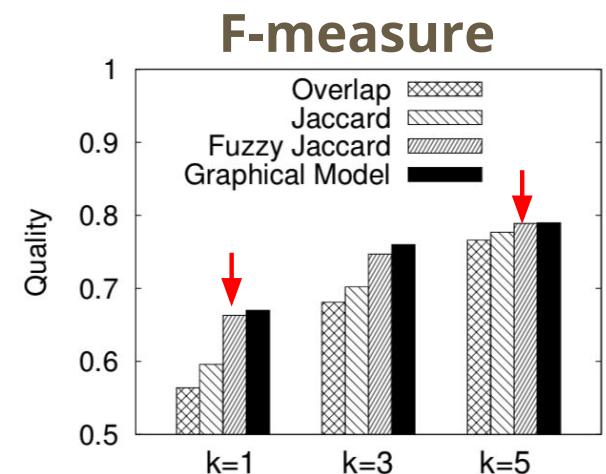
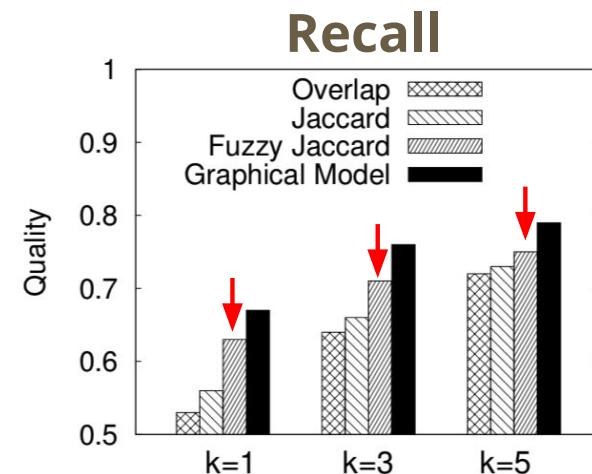
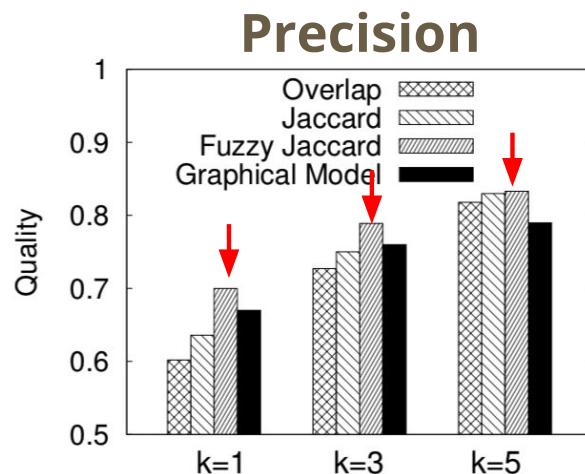
Fairly high performance

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

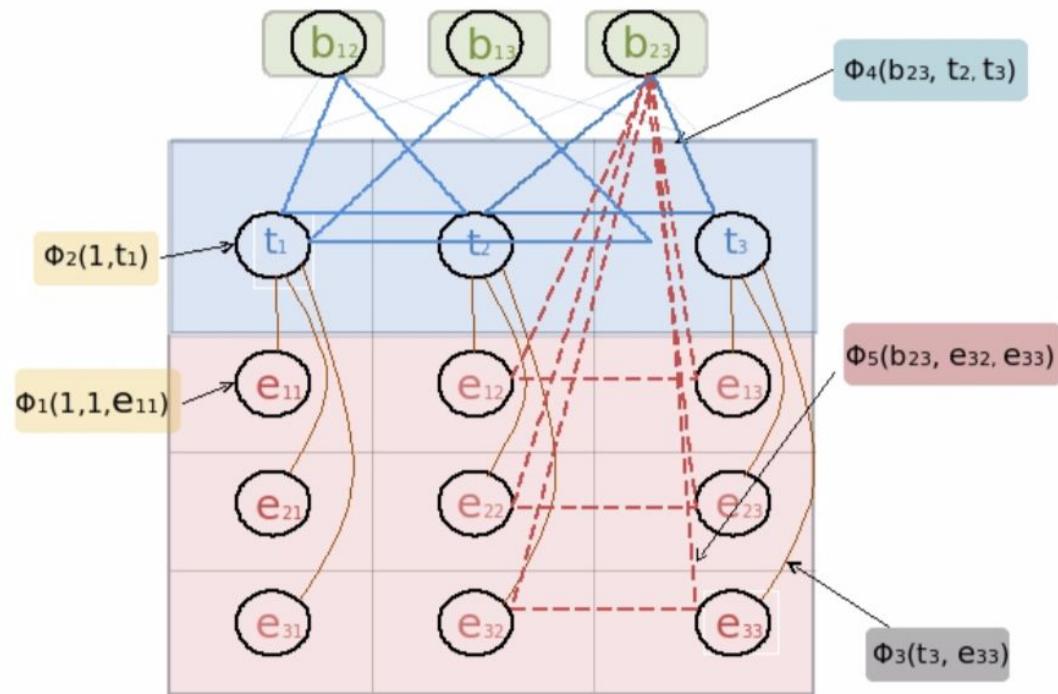
Performance: 45% Precision, 70% Recall (low performance)

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

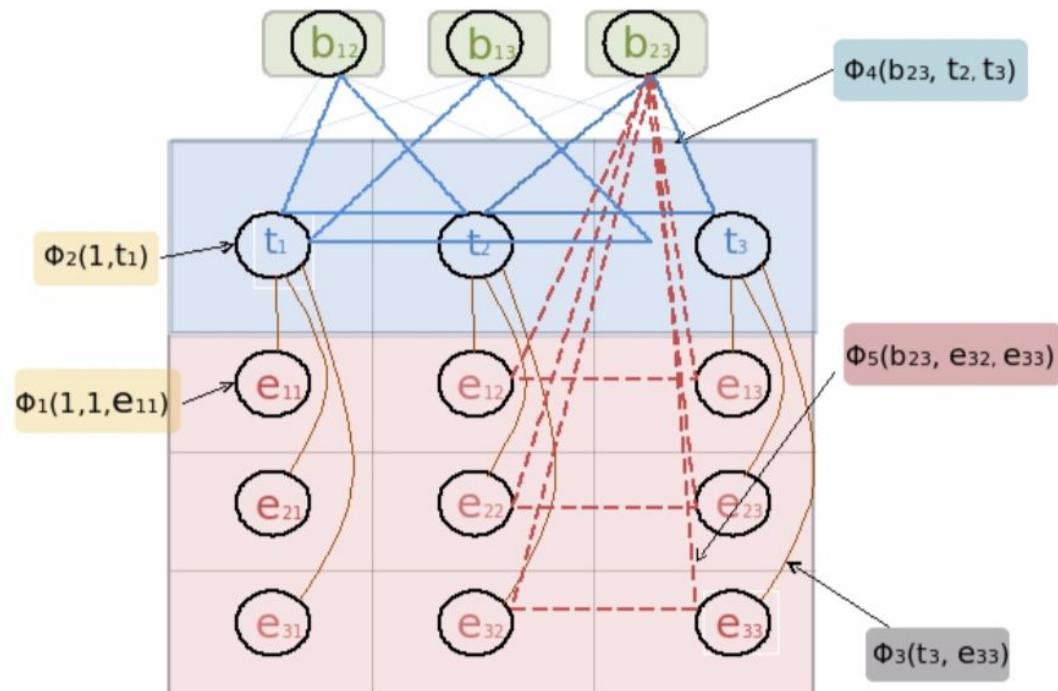
- Cell text (in Web table) and entity label (in catalog)
- Column header (in Web table) and type label (in catalog)
- Column type and cell entity (in Web table)



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

- Pair of column types (in Web table) and relation (in catalog)
- Entity pairs (in Web table) and relation (in catalog)



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	83.92
Web_Manual	59.68	75.87	81.37
Wiki_Link	67.92	77.63	84.28

Type annotation accuracy			
Dataset	LCA	MAJORITY	COLLECTIVE
Wiki_Manual	8.63	44.60	56.12
Web_Manual	15.16	31.45	43.23

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

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

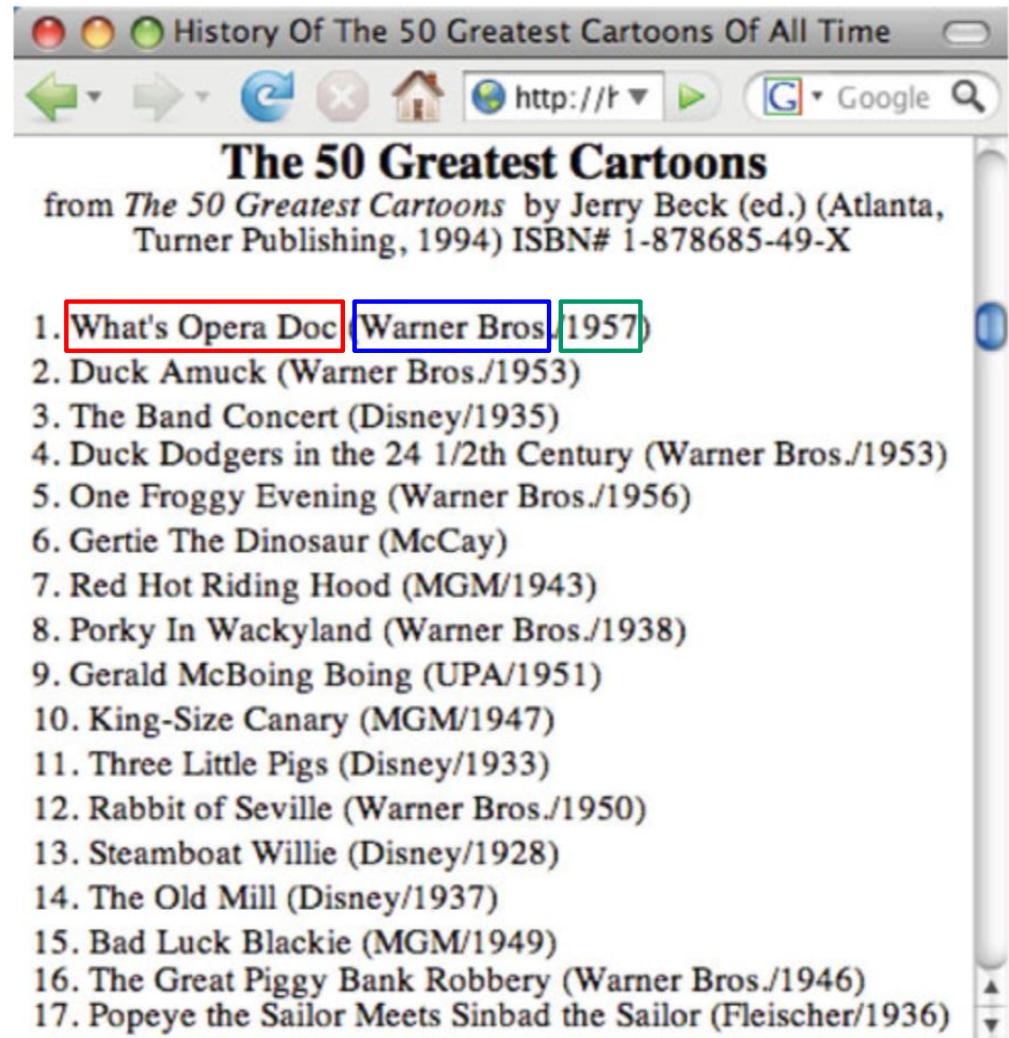
How can we extract from a web list?

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 titled "History Of The 50 Greatest Cartoons Of All Time". The page header reads "The 50 Greatest Cartoons" and provides information from "The 50 Greatest Cartoons" by Jerry Beck (ed.) (Atlanta, Turner Publishing, 1994) ISBN# 1-878685-49-X. Below the header is a numbered list of 17 cartoon titles, each with a small image and some text. The first item, "1. What's Opera Doc (Warner Bros./1957)", has its title and year highlighted with colored boxes: "What's Opera Doc" is red, "(Warner Bros./1957)" is green, and the entire row is bordered in blue.

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

Slide from: Ella Bolshinsky

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

Slide from: Ella Bolshinsky

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

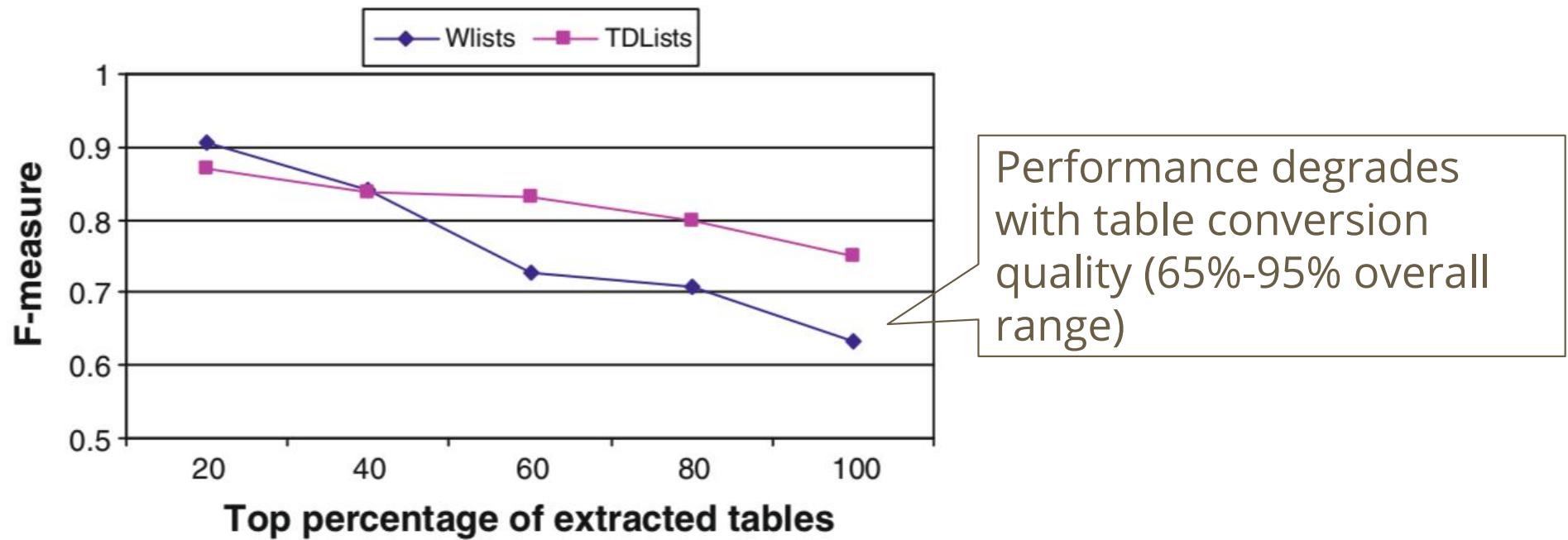
Idea: Transform a list into table

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).

Outline

- Introduction (30 minutes)
- Part Ia: Unstructured text (30 minutes)
- Break (30 minutes)
- Part Ib: Unstructured text: Methods (15 minutes)
- Part II: Semi-structured text (45 minutes)
- Part III: Tabular text (15 minutes)
- **Part IV: Multi-modal extraction** (30 minutes)
- Conclusion and future directions (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

	Unstructured	Semi-structured	Tabular	Multi-modal
Input data	Raw text (sentence, paragraph, or document)	Detail page HTML	Rows and columns	HTML + Rendered visuals
Diversity Challenges	Languages and dialects, diversity of expression	Templates, topic domain, relation strings	Topic domains	All: Language, template, topic
Consistent Patterns	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.

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 CO₂ emission and 105 MPGe fuel consumption.

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%.

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

	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

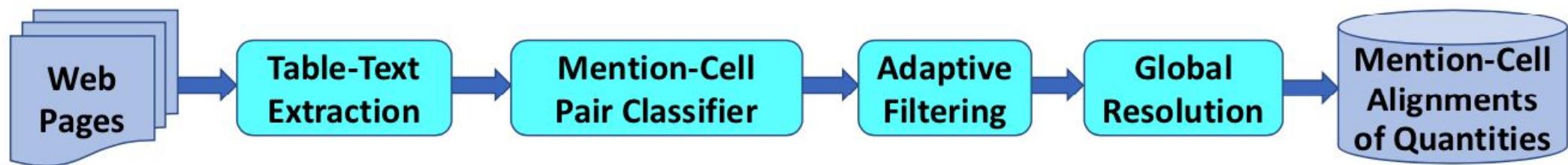
	Income gains (in Mio)	2013	2012	2011
Total Revenue	3,263	3,193	2,911	
Gross income	1,069	1,053	0,877	
Income taxes	179	177	160	
Income	890	876	849	

a) Example about Health

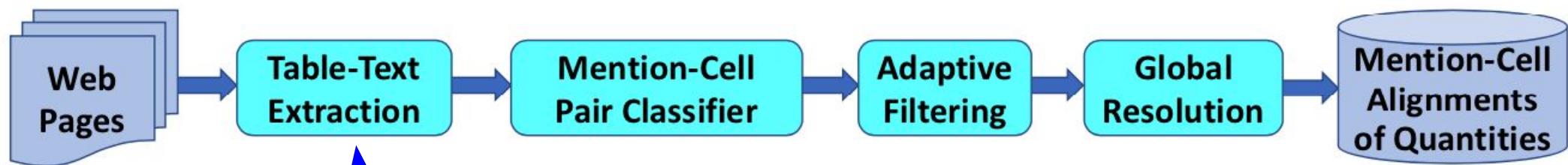
b) Example about Environment

c) Example about Finance

BriQ

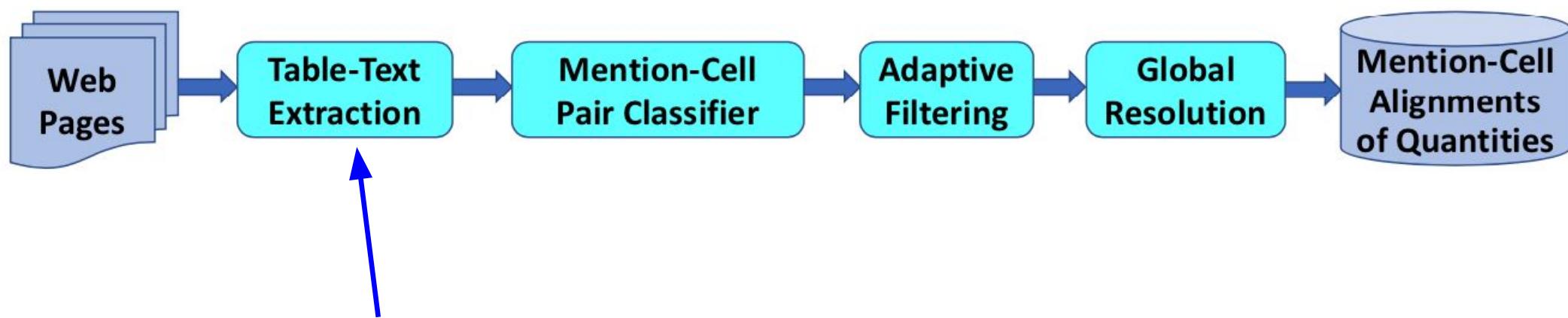


BriQ



Get text and
tables from
webpage, find
numeric mentions

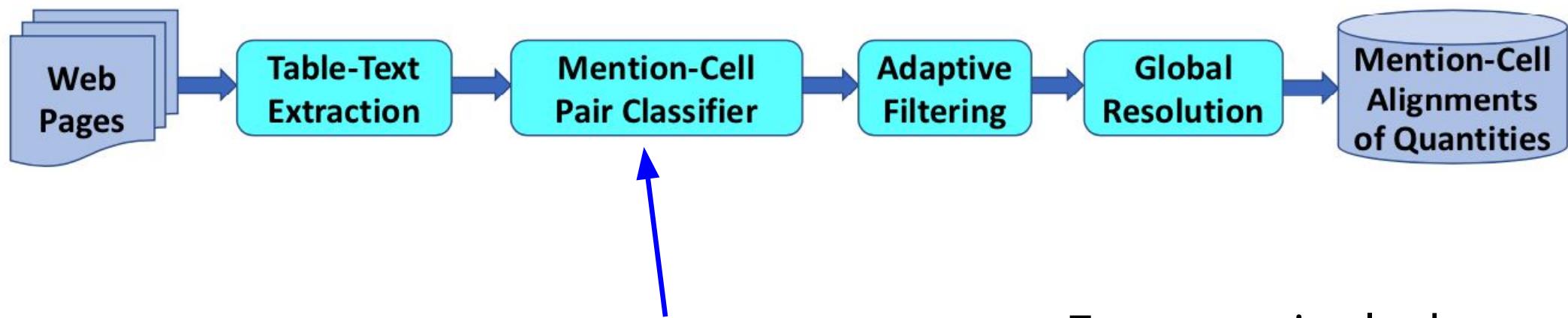
BriQ



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

- Sum
- Difference
- Percentage
- Change ratio

BriQ

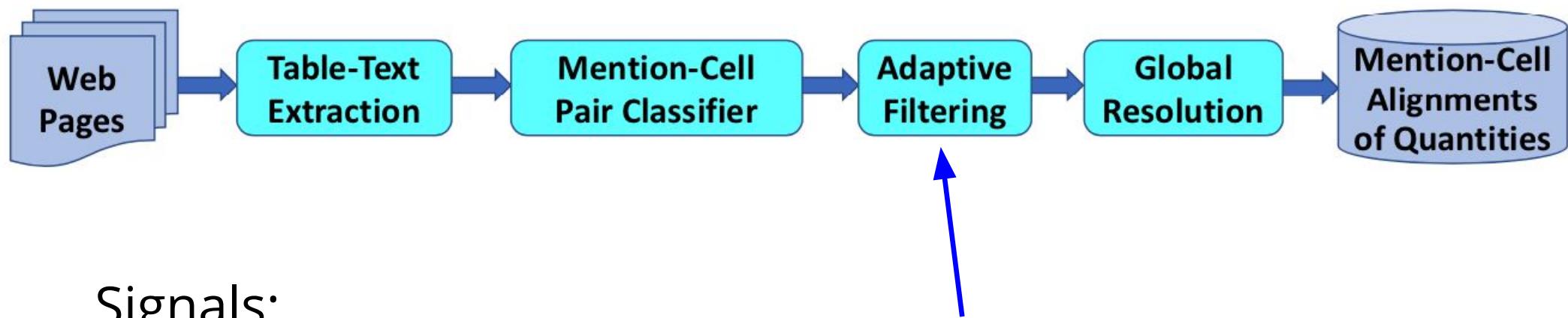


Binary classification of text/table quantity pairs as being likely/unlikely to indicate same quantity

Features include:

- Scale diff
- Precision diff
- Unit match
- Text context

BriQ

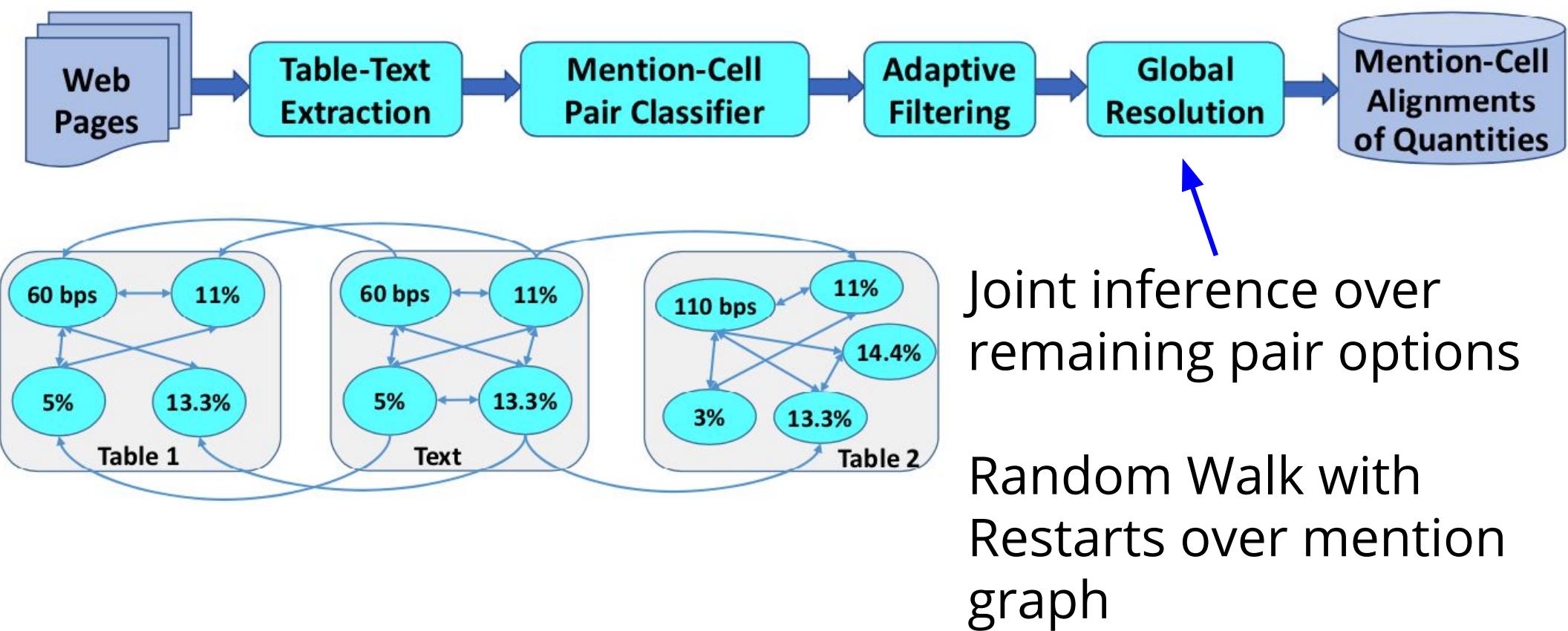


Signals:

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

Prune to best options

BriQ



BriQ

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

- 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

**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

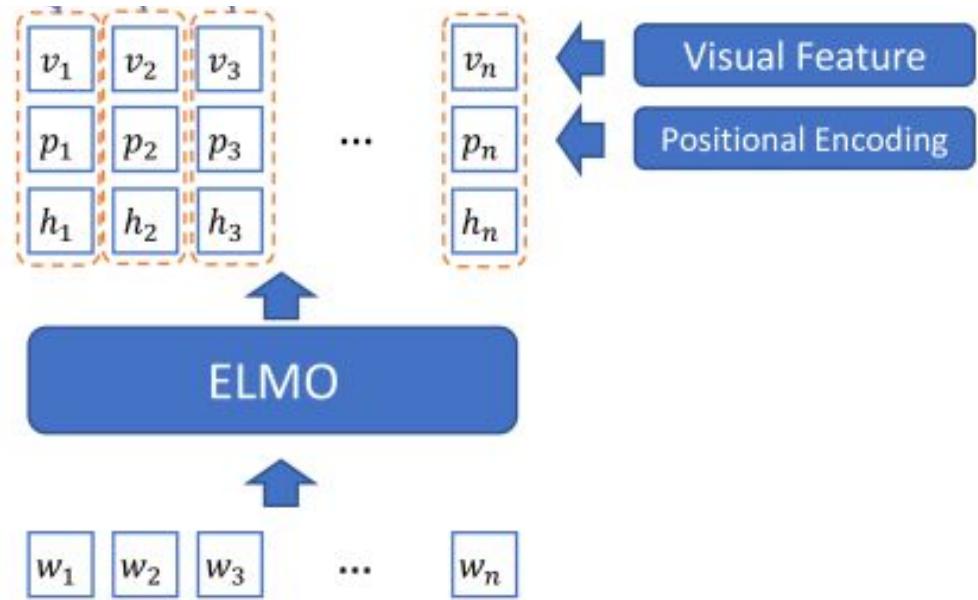
Typical approach: Use only unstructured text

This method: Incorporate visual features



Bling-KPE

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

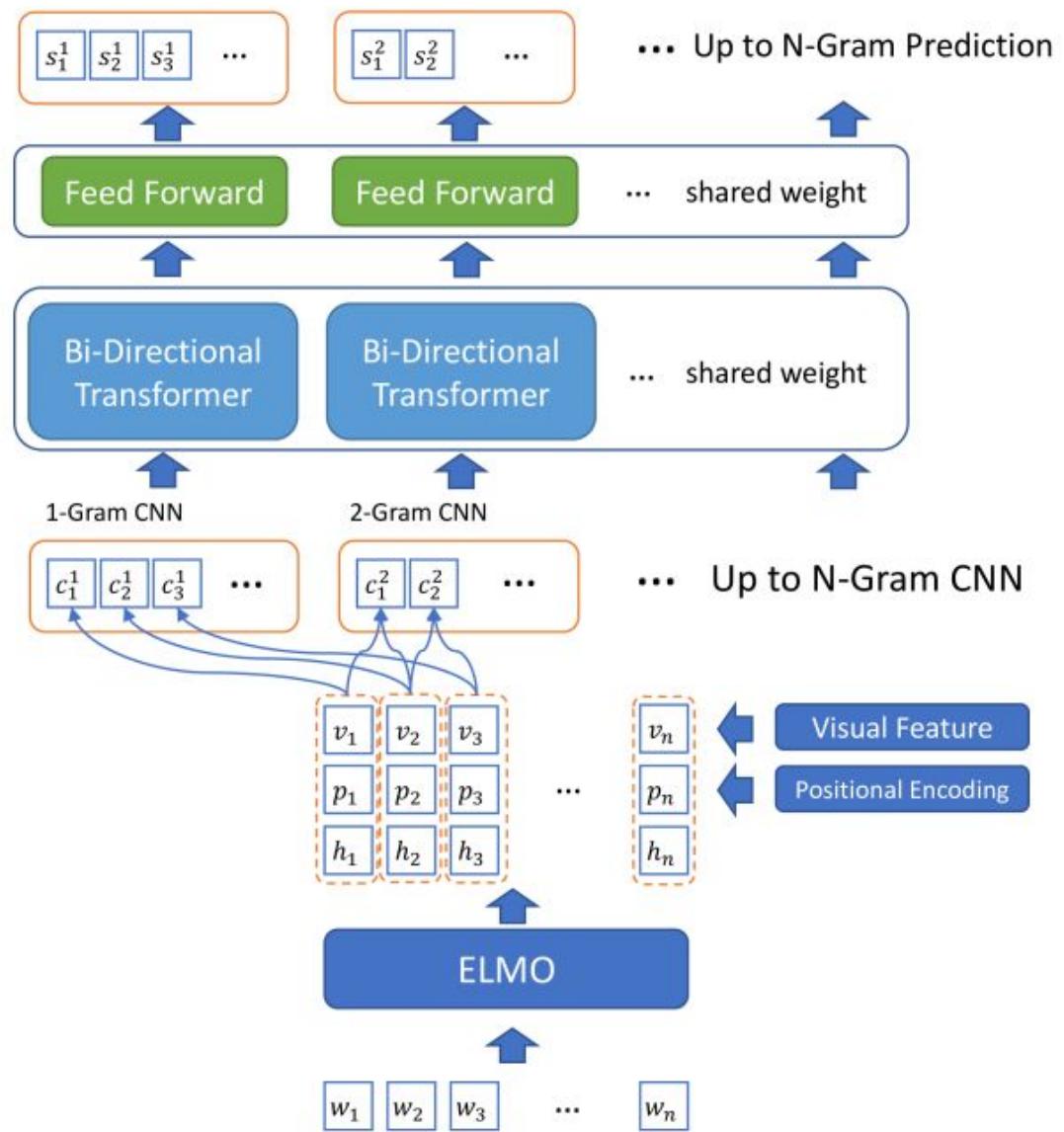


Name	Dimension
Font Size	1×2
Text Block Size	2×2
Location in Rendered Page	2×2
Is Bold Font	1×2
Appear In Inline	1×2
Appear In Block	1×2
Appear In DOM Tree Leaf	1×2

Bling-KPE

Convolution over n-grams models potential keyphrases

Weak supervision from search logs





Shop by category ▾

Search for anything

eBay > Consumer Electronics >

Radio Communication > Parts & Accessories > Manuals & 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 Staple Gun is a reliable fire engine that produces 16 gauge staples at a rate of 50 staples per second. It is ideal for applications of 16-gauge wire and staples."

[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	0.404	0.220

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

Submission Date: 22.05.2018

Preserves
2D layout

Operates on
pixels

Document as image

Title: Chargrid

paperID
0245 conference
EMNLP

Submission Date: 22.05.2018

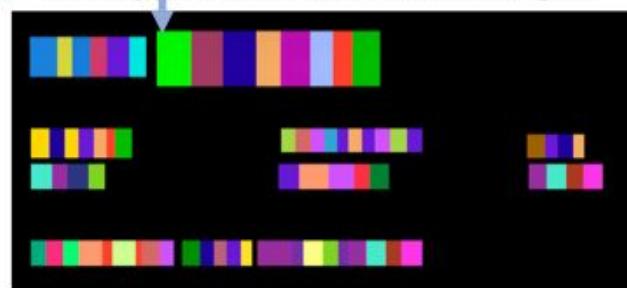
Operates on text
Ignores 2D layout

Document as serialized string

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

Operates on text
Preserves 2D layout

Our Approach: Document as Chargrid



Chargrid zoom-in

0	0	0	0	0	0	0	0	0	0
0	0	3	3	3	3	3	8	8	8
0	0	3	3	3	3	3	8	8	8
0	0	3	3	3	3	3	8	8	8
0	0	3	3	3	3	3	8	8	8
0	0	3	3	3	3	3	8	8	8
0	0	3	3	3	3	3	8	8	8
0	0	3	3	3	3	3	8	8	8
0	0	0	0	0	0	0	0	0	0

Image via <https://medium.com/sap-machine-learning-research/chargrid-77aa75e6d605>

CharGrid

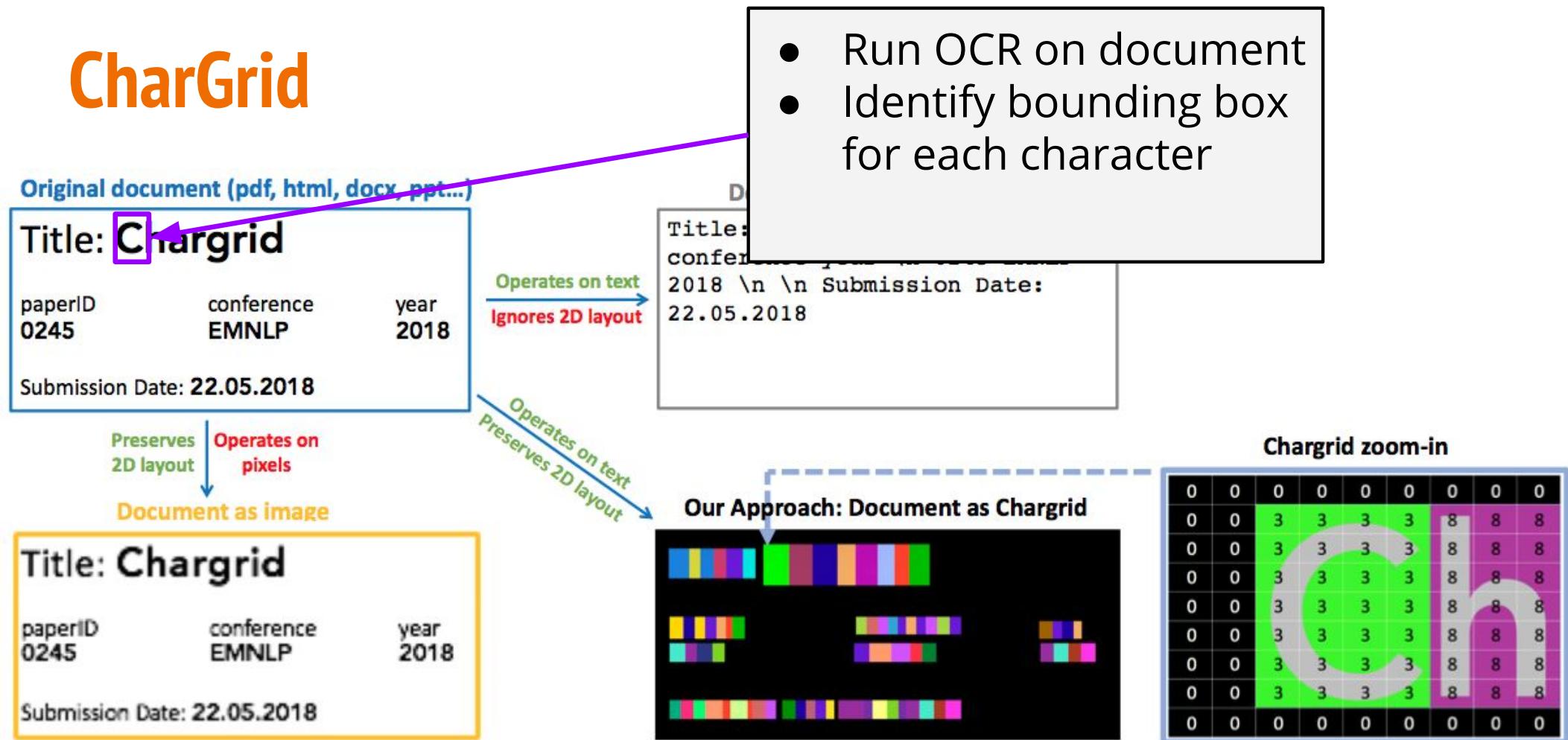


Image via <https://medium.com/sap-machine-learning-research/chargrid-77aa75e6d605>

CharGrid

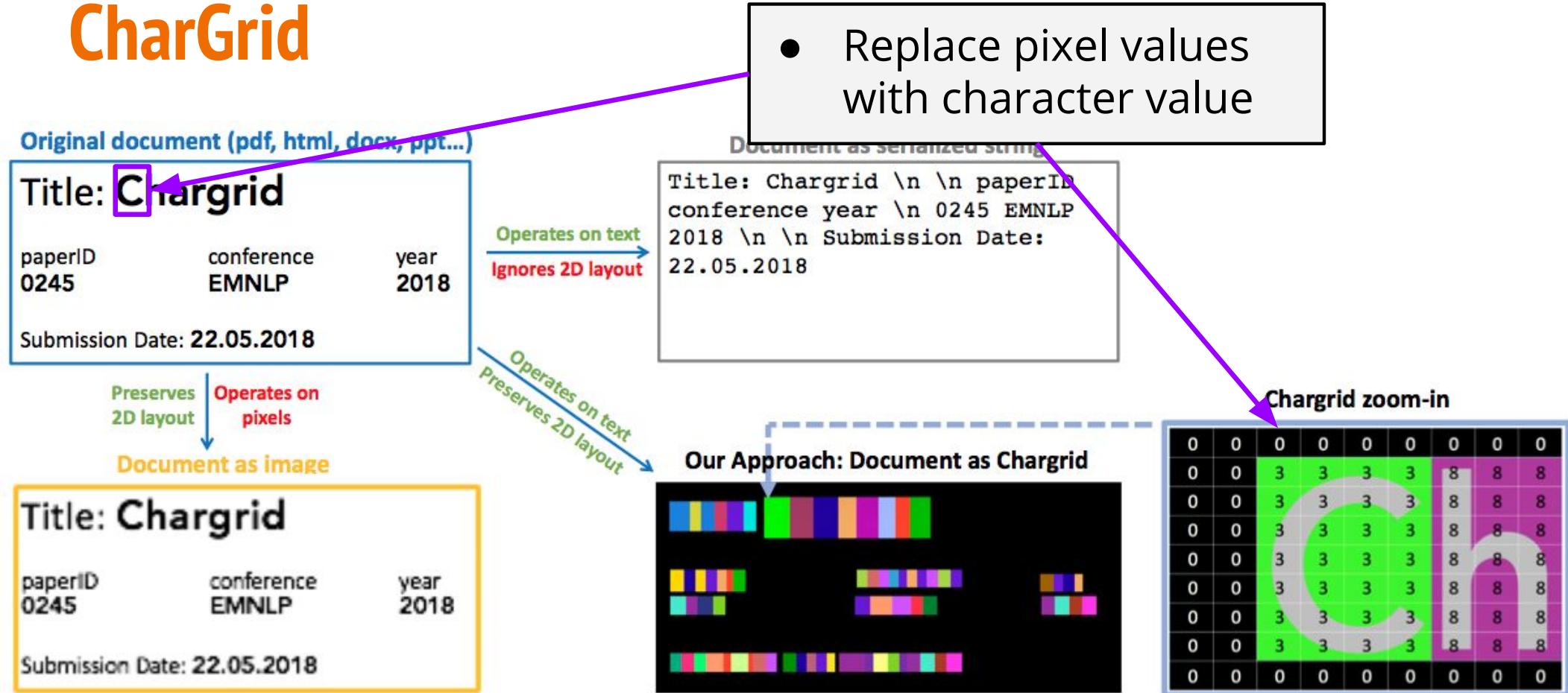


Image via <https://medium.com/sap-machine-learning-research/chargrid-77aa75e6d605>

CharGrid

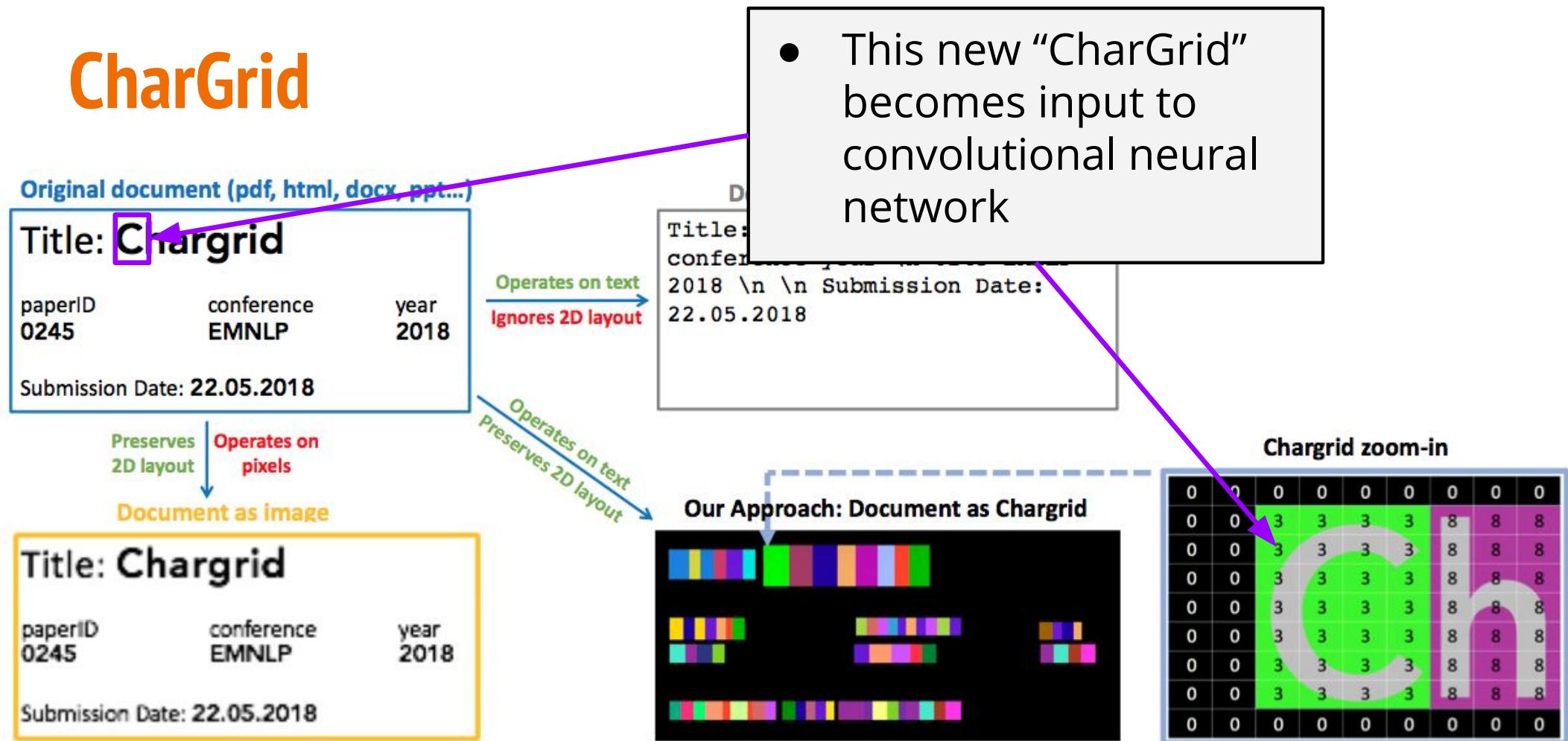
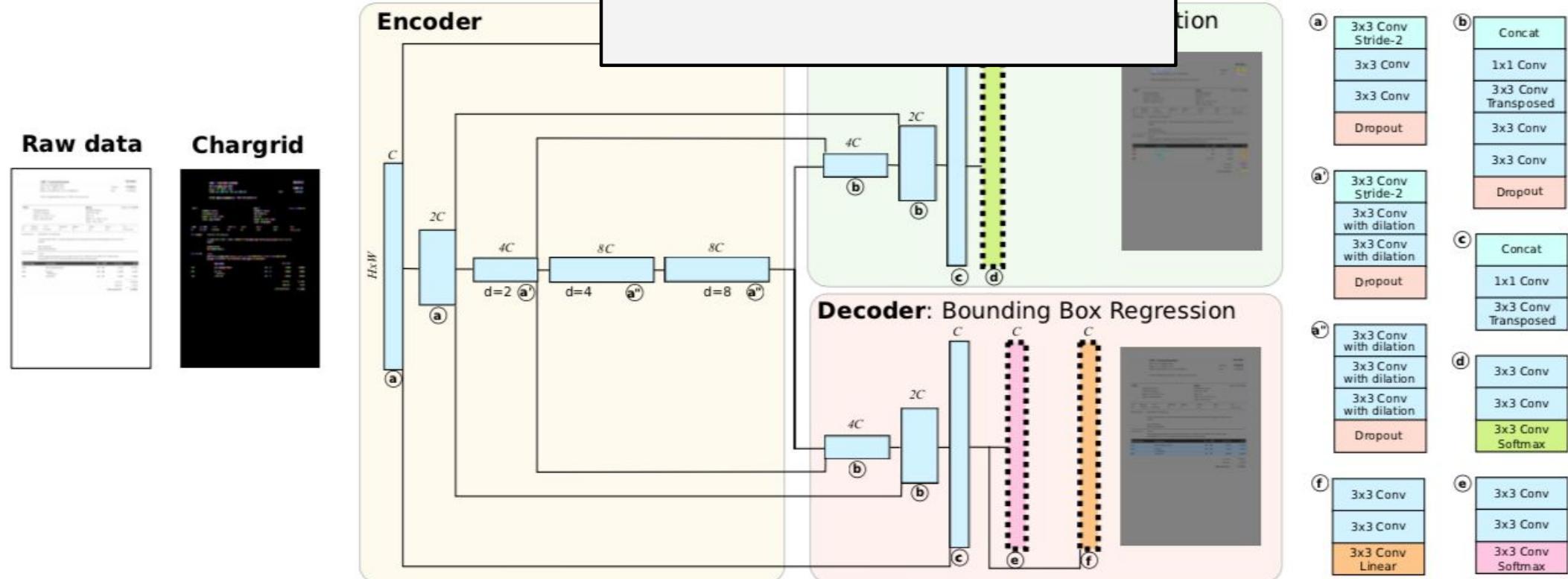


Image via <https://medium.com/sap-machine-learning-research/chargrid-77aa75e6d605>

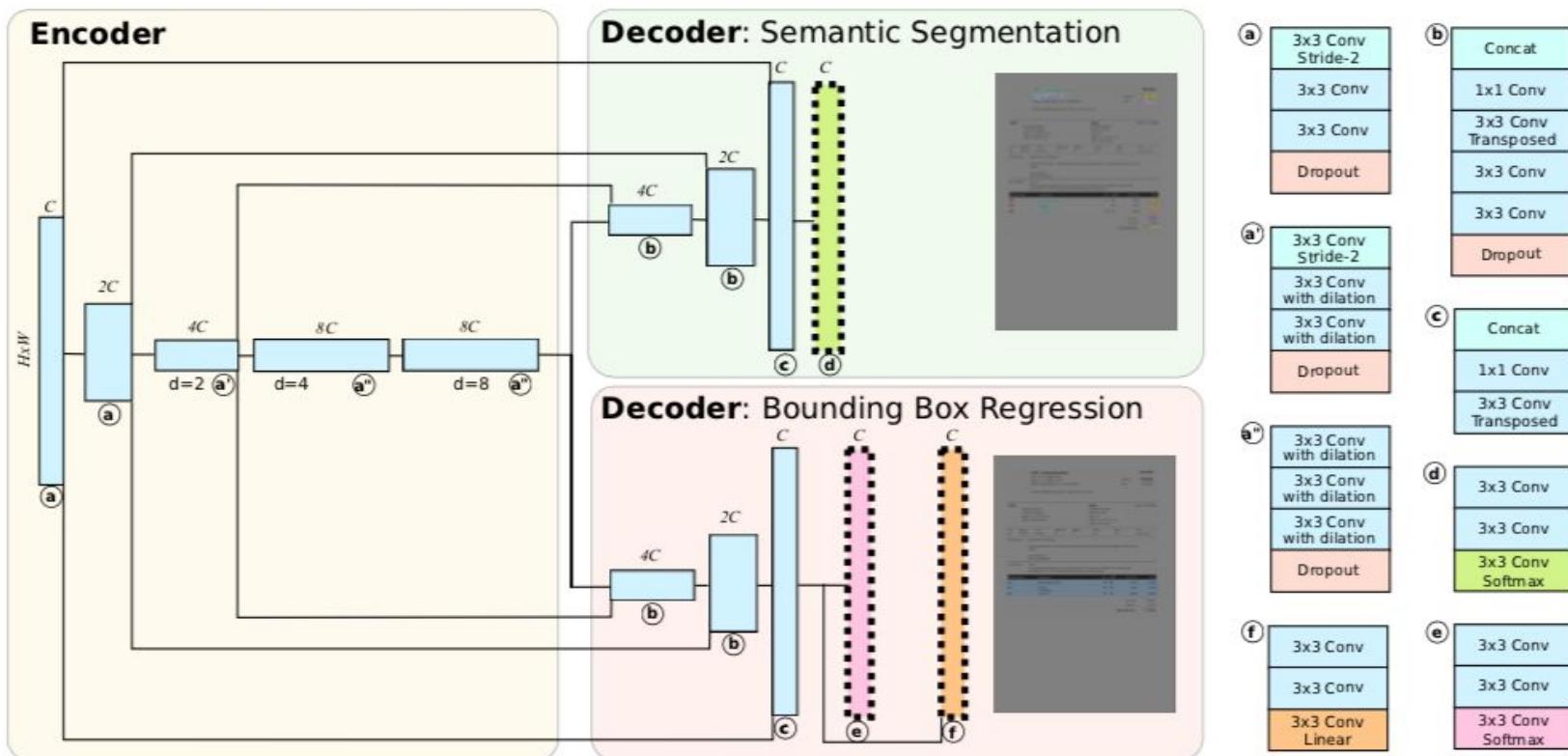
CharGrid

- VGGNet Convolutional Neural Network encodes CharGrid



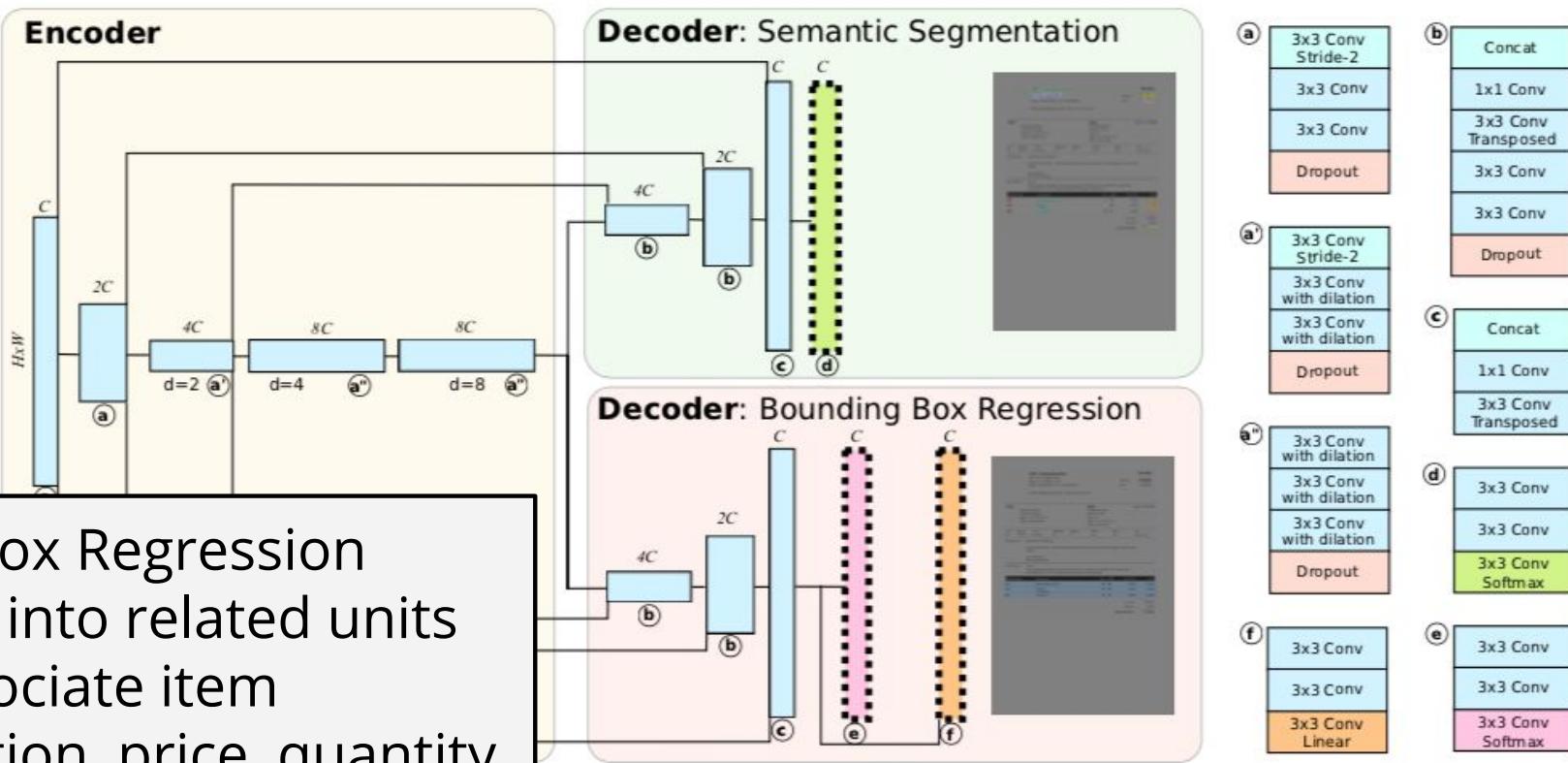
CharGrid

- 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%

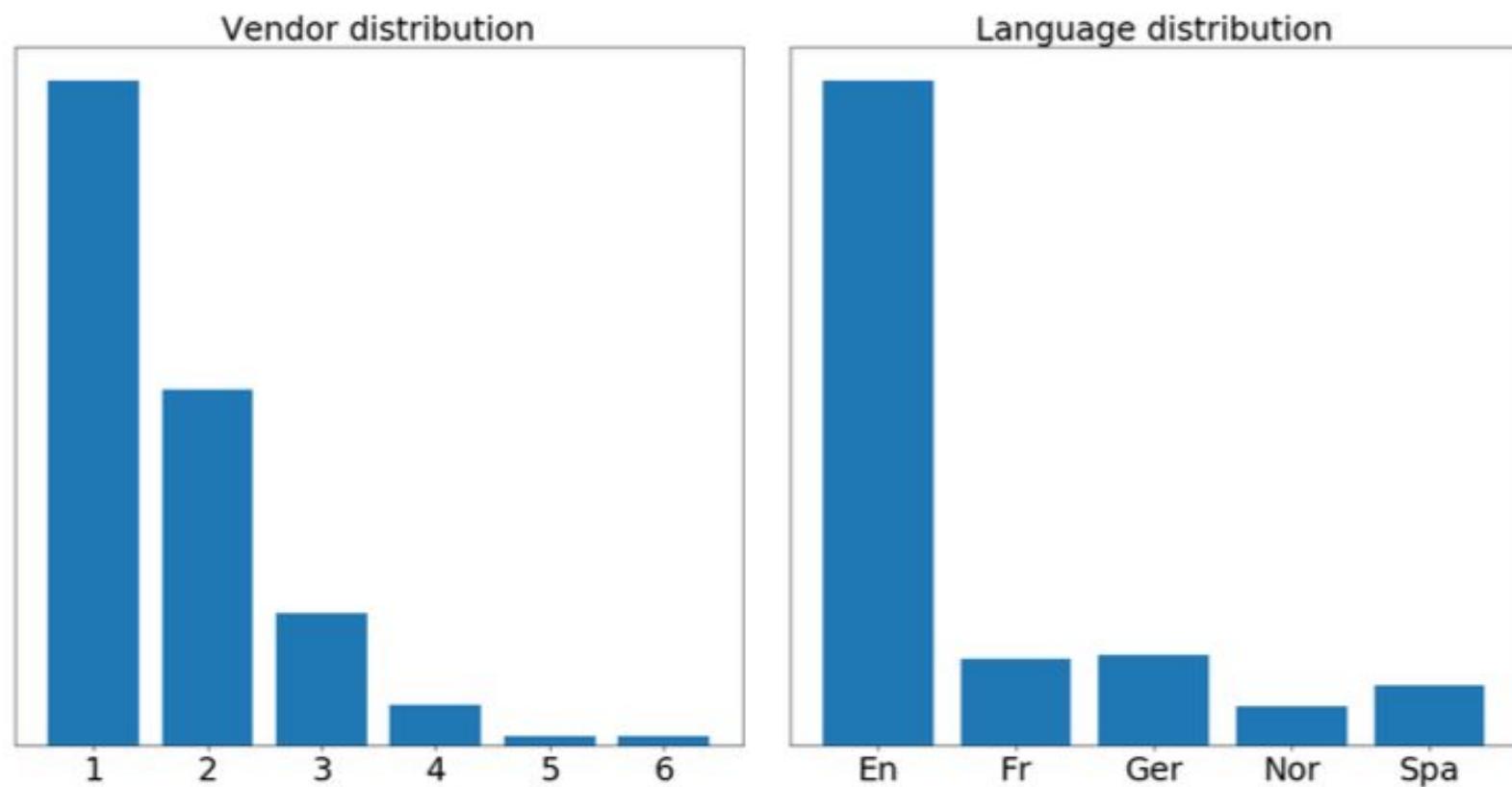
Line-item values require associating multiple text fields.
Bounding box detection makes this possible for
CharGrid.

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

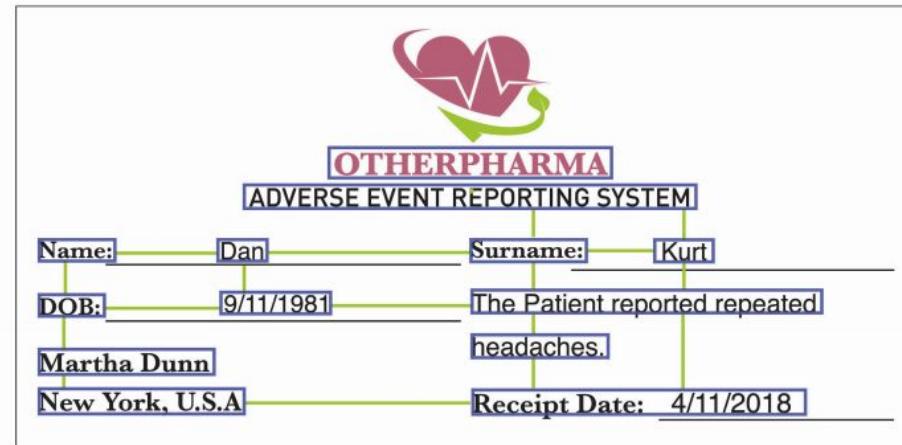
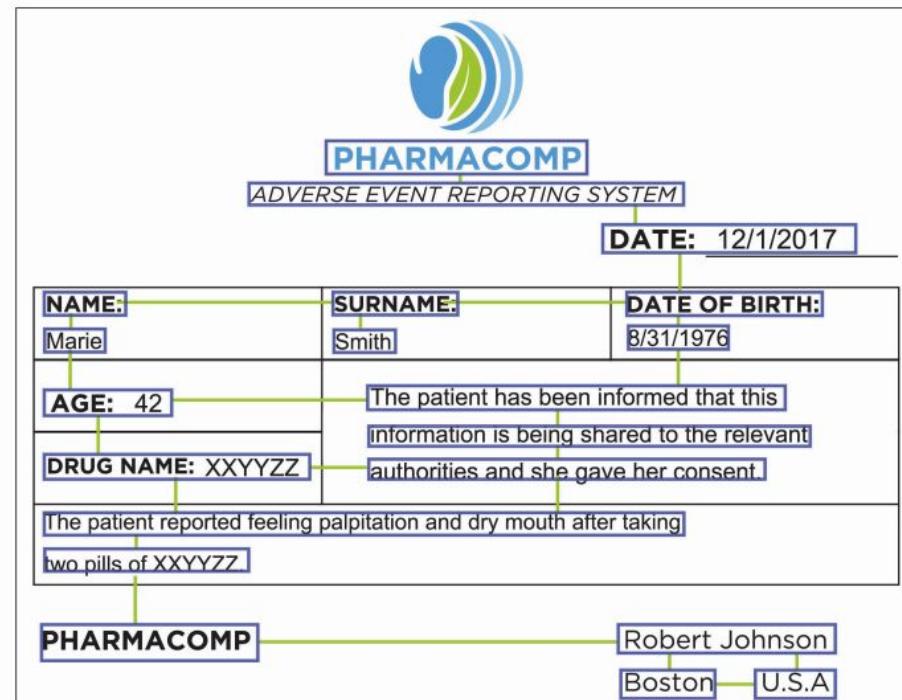


CharGrid

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

GraphIE (Qian et al, 2019)

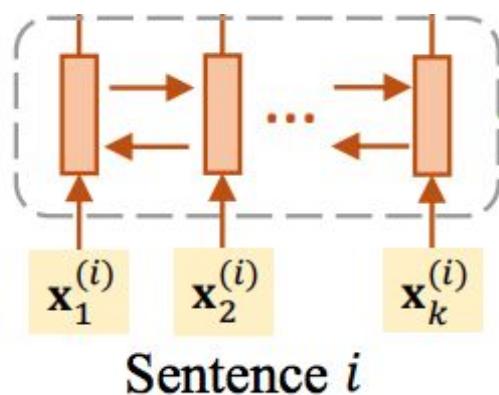
- Combine textual and layout information of semi-structured documents
- Model documents as a graph
 - Nodes are text fields
 - Edges indicate horizontal/vertical adjacency between pair of text fields



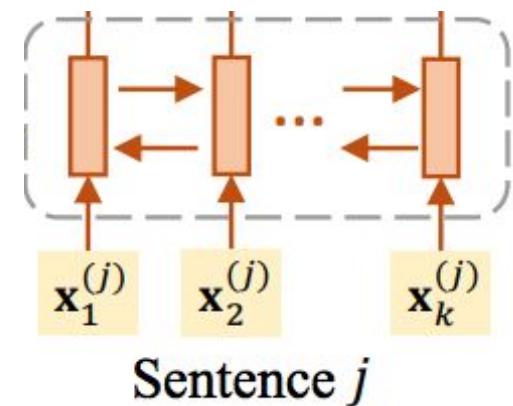
GraphIE

Encode text in each text field.
(They use LSTM. Could also use BERT)

Encoder
(BiLSTM)

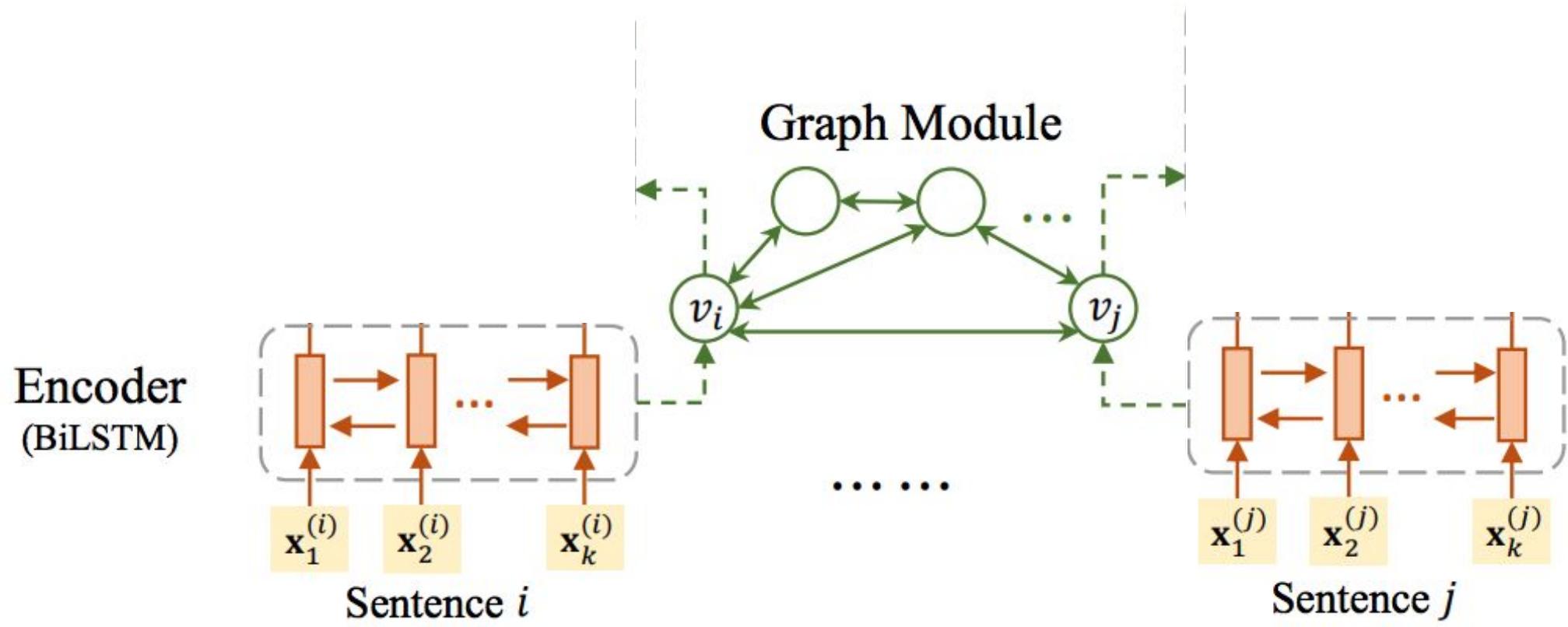


...



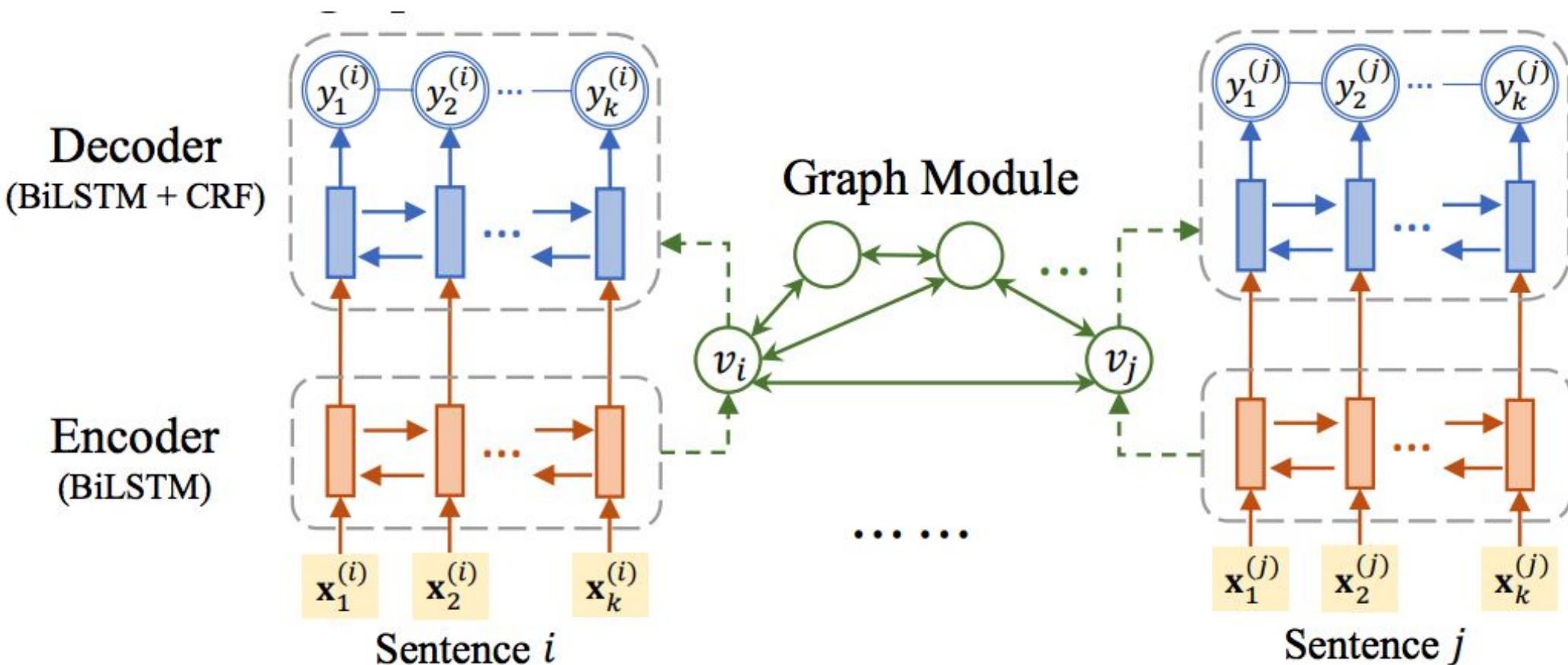
GraphIE

Apply Graph Convolutional Network to page layout graph



GraphIE

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



GraphIE

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

Table 6: Extraction accuracy on the AECR dataset (Task 3). Scores are the average of 5 runs. *P.* is the abbreviation for *Patient*, and *R.* for *Reporter*. [†] 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	92.9	93.6	91.9	92.8
<i>P. Age</i>	94.0	91.6	92.8	94.8	91.1	92.9
<i>P. Birthday</i>	96.6	96.0	96.3	96.9	94.7	95.8
<i>Drug Name</i>	71.2	51.2	59.4	78.5	50.4	61.4
<i>Event</i>	62.6	65.2	63.9	64.1	68.7	66.3
<i>R. First Name</i>	78.3	95.7	86.1	79.5	95.9	86.9
<i>R. Last Name</i>	84.5	68.4	75.6	85.6	68.2	75.9
<i>R. City</i>	88.9	65.4	75.4	92.1	66.3	77.1
Avg. (macro)	83.7	78.2	80.3	85.7	78.4	81.1[†]
Avg. (micro)	78.5	73.8	76.1	80.3	74.6	77.3[†]

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

On templates unseen during training, the layout graph helps tremendously

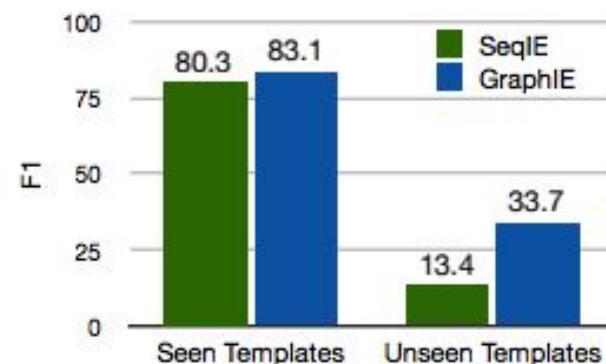


Figure 4: Micro average F1 scores tested on *seen* and *unseen* templates (Task 3).

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

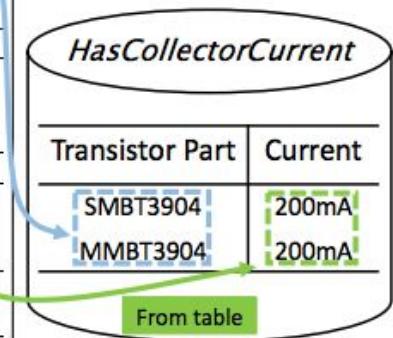
Extraction model uses multimodal features

Transistor Datasheet

SMBT3904..MMBT3904			
NPN Silicon Switching Transistors			
• High DC current gain: 0.1 mA to 100 mA			
• Low collector-emitter saturation voltage			
From header			
Maximum Ratings			
Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	← Aligned →		200 mA
Total power dissipation $T_S \leq 60^\circ\text{C}$	P_{tot}	330	mW
$T_S \leq 115^\circ\text{C}$		250	
Junction temperature	T_J	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Font: Arial; Size: 12; Style: Bold

Knowledge Base



3 labeling functions

Each is informative on different examples

```
def has_research_in_row(c):
    if 'research' in row_ngrams(c.amount):
        return 1
    else:
        return 0
```

```
def vertical_aligned(c):
    if c.quarter.x == c.amount.x:
        return 1
    else:
        return 0
```

```
def magnitude_too_small(c):
    if c.amount < 1000:
        return -1
    else:
        return 0
```

(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
```

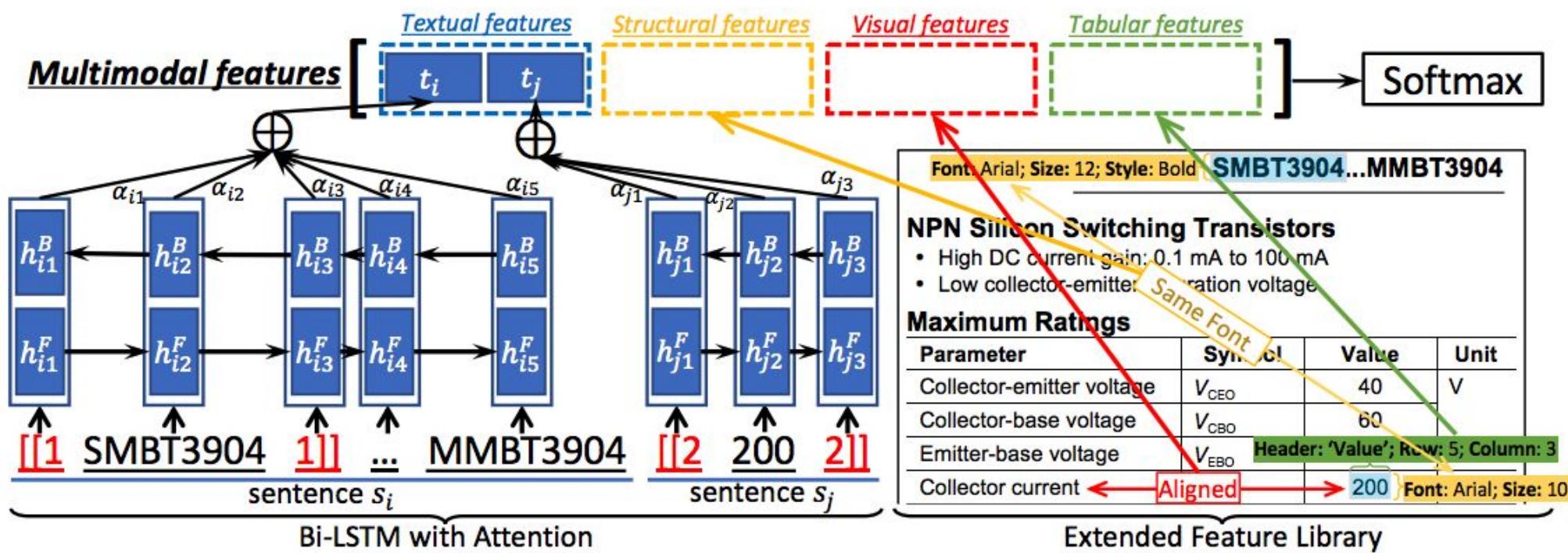
```
def value_in_column_header(cand):
    if 'Value' in header_ngrams(cand.current):
        return 1
    else:
        return 0
```

```
# 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
```



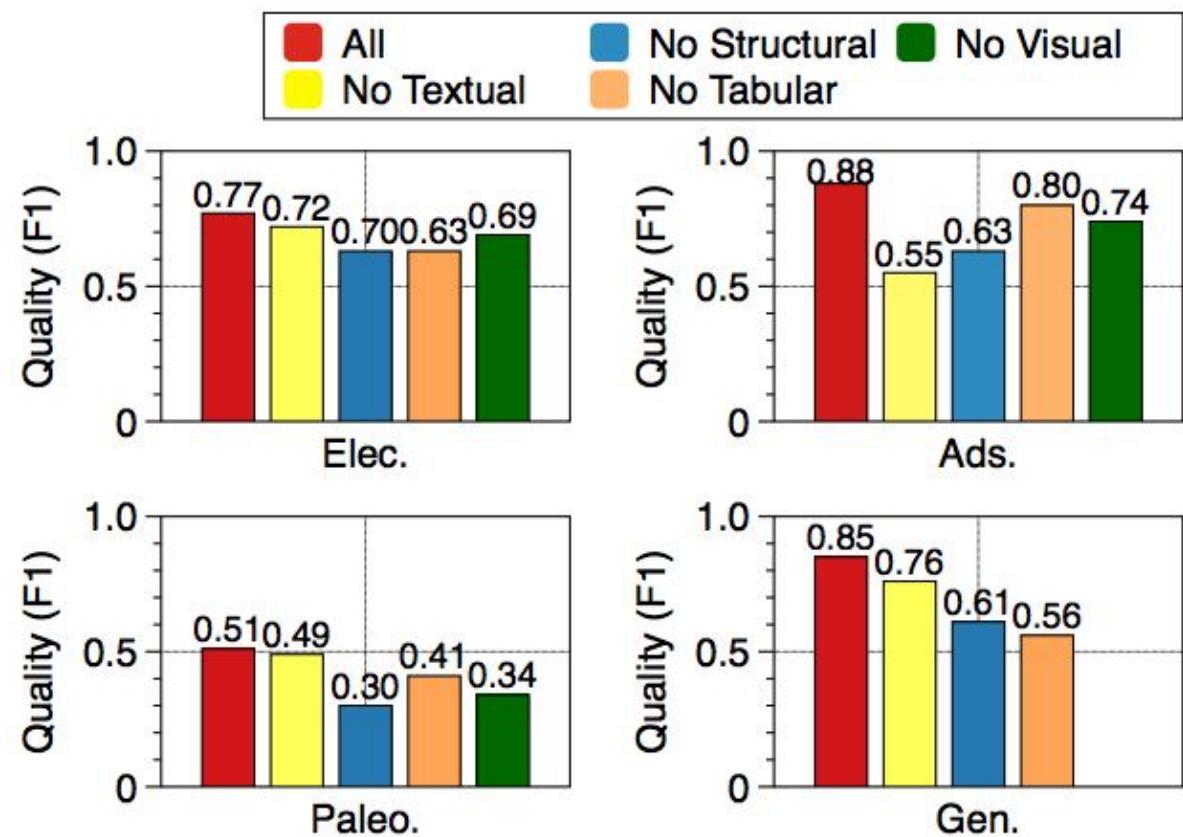
Fonduer

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	0.77
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	0.88
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	0.51
GEN.	Prec.	0.00	0.00	0.00	0.89
	Rec.	0.00	0.00	0.00	0.81
	F1	0.00 [#]	0.00 [#]	0.00 [#]	0.85

Huge gains in recall
with small loss of
precision

Fonduer

Different datasets benefit from different features



Fondue

- 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

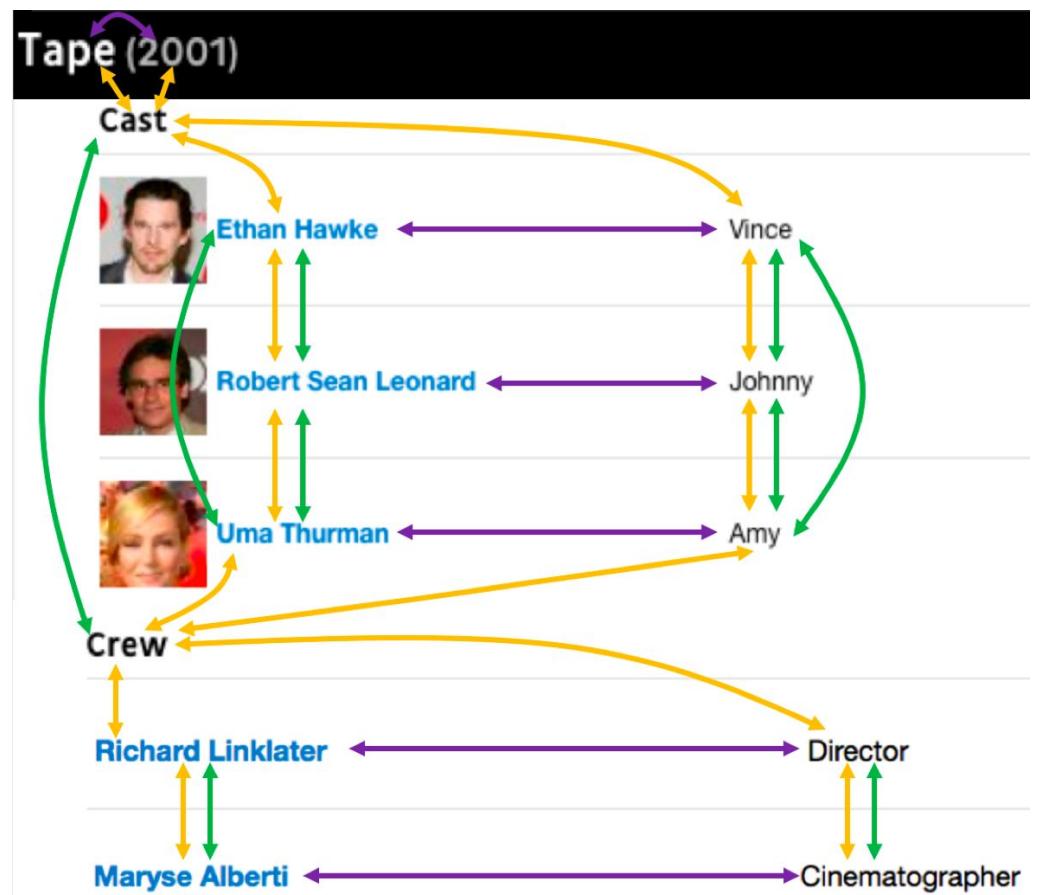


ZeroShotCeres

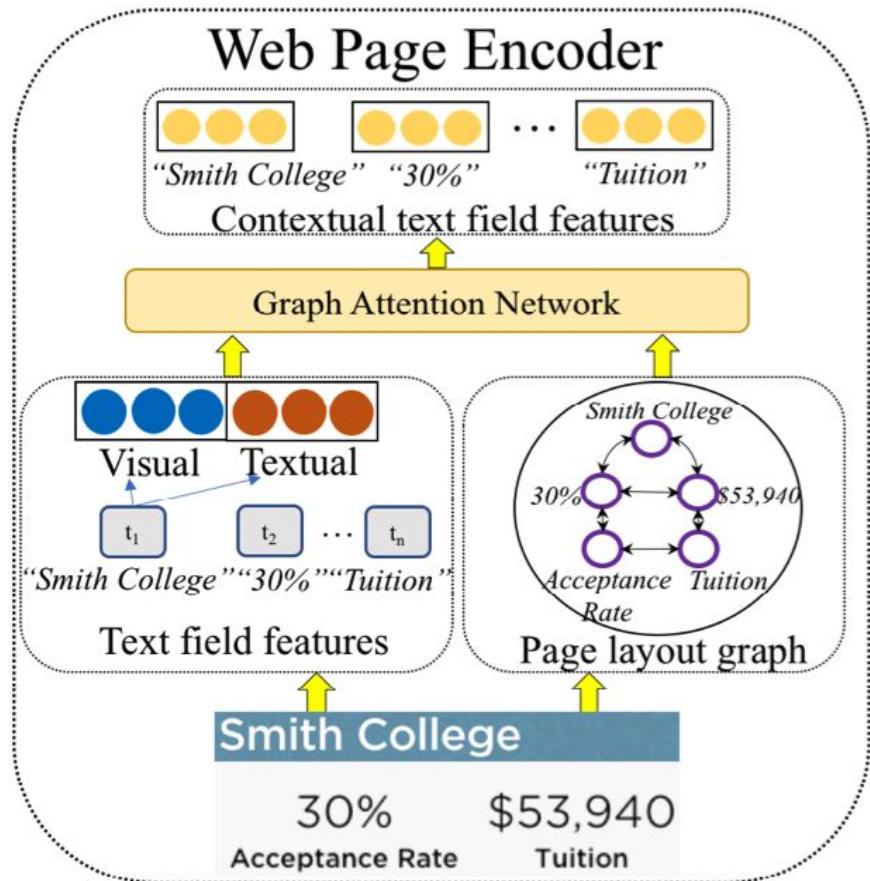
Horizontal edges

Vertical edges

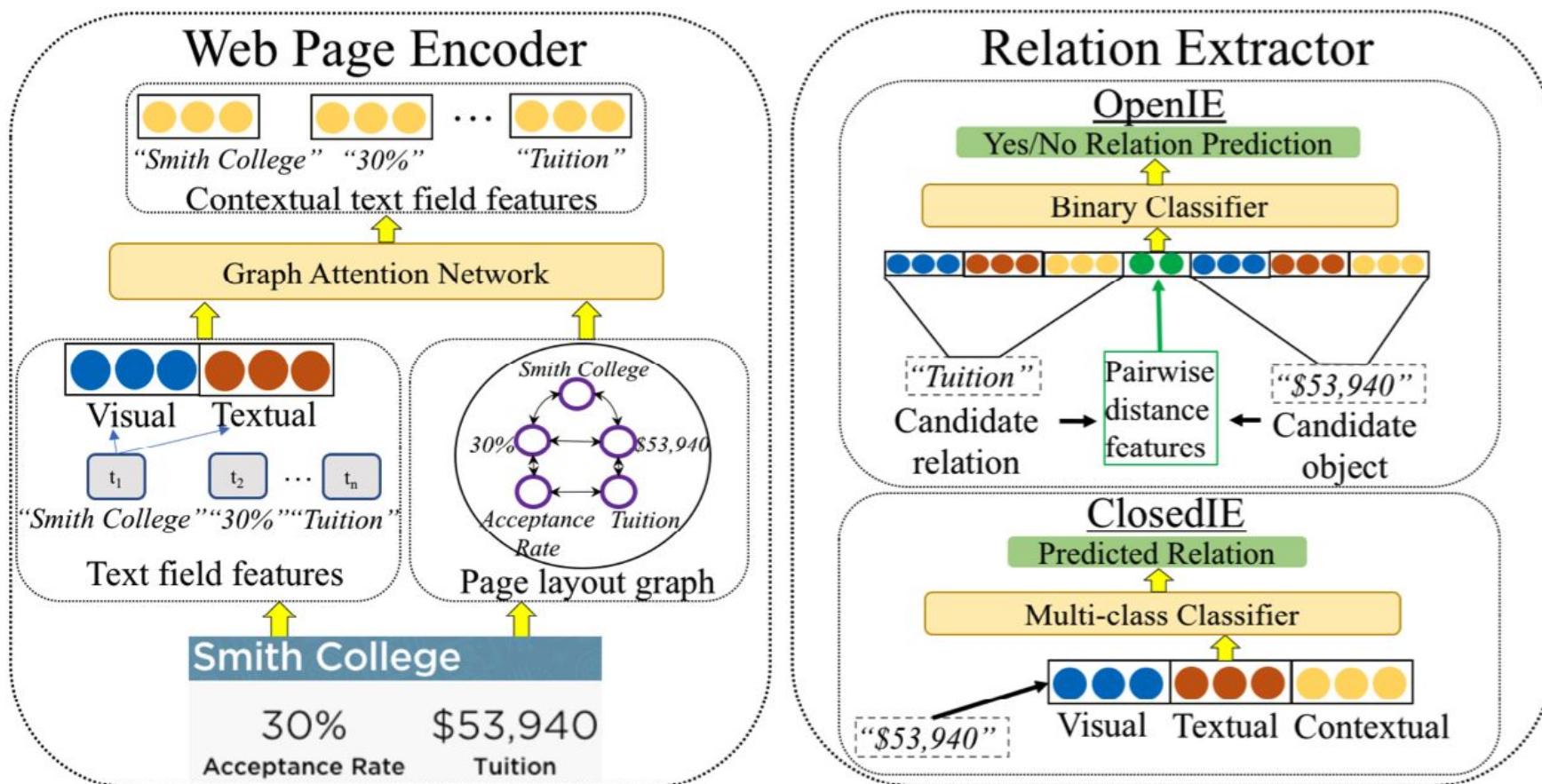
DOM edges connect nodes
that are siblings/cousins in
DOM tree



ZeroShotCeres



ZeroShotCeres



ZeroShotCeres

System	Site-specific Model	Level	Movie			NBA			University		
			P	R	F1	P	R	F1	P	R	F1
OpenCeres	Yes	III	0.71	0.84	0.77	0.74	0.48	0.58	0.65	0.29	0.40
WEIR	Yes	III	0.14	0.10	0.12	0.08	0.17	0.11	0.13	0.18	0.15
SWPR-FFNN All-Domain	No	II	0.37	0.5	0.45	0.35	0.49	0.41	0.47	0.59	0.52
SWPR-GNN All-Domain	No	II	0.49	0.51	0.50	0.47	0.39	0.42	0.50	0.49	0.50
Colon Baseline	No	I	0.47	0.19	0.27	0.51	0.33	0.40	0.46	0.31	0.37
SWPR-FFNN New Domain	No	I	0.42	0.38	0.40	0.44	0.46	0.45	0.50	0.45	0.48
SWPR-GNN New Domain	No	I	0.43	0.42	0.42	0.48	0.49	0.48	0.49	0.45	0.47

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

ZeroShotCeres

System	Site-specific Model	Level	Movie			NBA			University		
			P	R	F1	P	R	F1	P	R	F1
OpenCeres	Yes	III	0.71	0.84	0.77	0.74	0.48	0.58	0.65	0.29	0.40
WEIR	Yes	III	0.14	0.10	0.12	0.08	0.17	0.11	0.13	0.18	0.15
SWPR-FFNN All-Domain	No	II	0.37	0.5	0.45	0.35	0.49	0.41	0.47	0.59	0.52
SWPR-GNN All-Domain	No	II	0.49	0.51	0.50	0.47	0.39	0.42	0.50	0.49	0.50
Colon Baseline	No	I	0.47	0.19	0.27	0.51	0.33	0.40	0.46	0.31	0.37
SWPR-FFNN New Domain	No	I	0.42	0.38	0.40	0.44	0.46	0.45	0.50	0.45	0.48
SWPR-GNN New Domain	No	I	0.43	0.42	0.42	0.48	0.49	0.48	0.49	0.45	0.47

With zero prior knowledge on University,
more accurate than OpenCeres

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
SWPR	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

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 and Xin Luna Dong. "OpenCeres: When Open Information Extraction Meets the Semi-Structured Web." *NAACL-HLT* (2019).

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 (30 minutes)
- Part Ia: Unstructured text (30 minutes)
- Break (30 minutes)
- Part Ib: Unstructured text: Methods (15 minutes)
- Part II: Semi-structured text (45 minutes)
- Part III: Tabular text (15 minutes)
- Part IV: Multi-modal extraction (30 minutes)
- **Conclusion and future directions (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
 - 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 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 (Jia et al, 2019)
 - 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.

References

Jia, Robin, Cliff Wong and Hoifung Poon. "Document-Level N-ary Relation Extraction with Multiscale Representation Learning." *NAACL-HLT* (2019).



Thank you!

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

Enjoy the rest of ACL!
