



From One Tree to a Forest – A Unified Solution for Structured Web Data Extraction

Qiang Hao[‡], Rui Cai[†], Yanwei Pang[‡], Lei Zhang[†]

[†] Microsoft Research Asia

[‡] Tianjin University

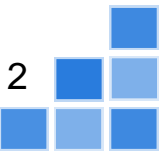
Microsoft
Research
微软亚洲研究院



Outline



- Motivation & Challenges
- Our Solution
 - Main Idea & Framework Overview
 - Feature Extraction
 - Vertical Knowledge Learning
 - Vertical Knowledge Adaptation
- Experimental Results
- Summary



What's Structured Data Extraction



- Extracting structured data records from web pages
= identifying values of **attributes**

The Kite Runner
by [Khaled Hosseini](#)
(Paperback - Reprint)
Reader Rating: ★★★★★ (1377 ratings)
> [Read customer reviews](#) [Write a Review](#)

- Pub. Date: [April 2004](#)
- 400pp
- Sales Rank: 644

9780307264237
Mercy
[Toni Morrison](#)

ISBN 10: 0307264238 / 0-307-26423-8
ISBN 13: 9780307264237
Publisher: **Alfred a Knopf Inc**
Publication Date: **2008**
Binding: **Hardcover**

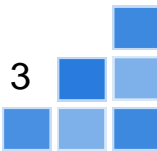
The Time Machine
H. G. Wells (See All Contributors)
Paperback
Kessinger Publishing, LLC
June 30, 2004



attributes

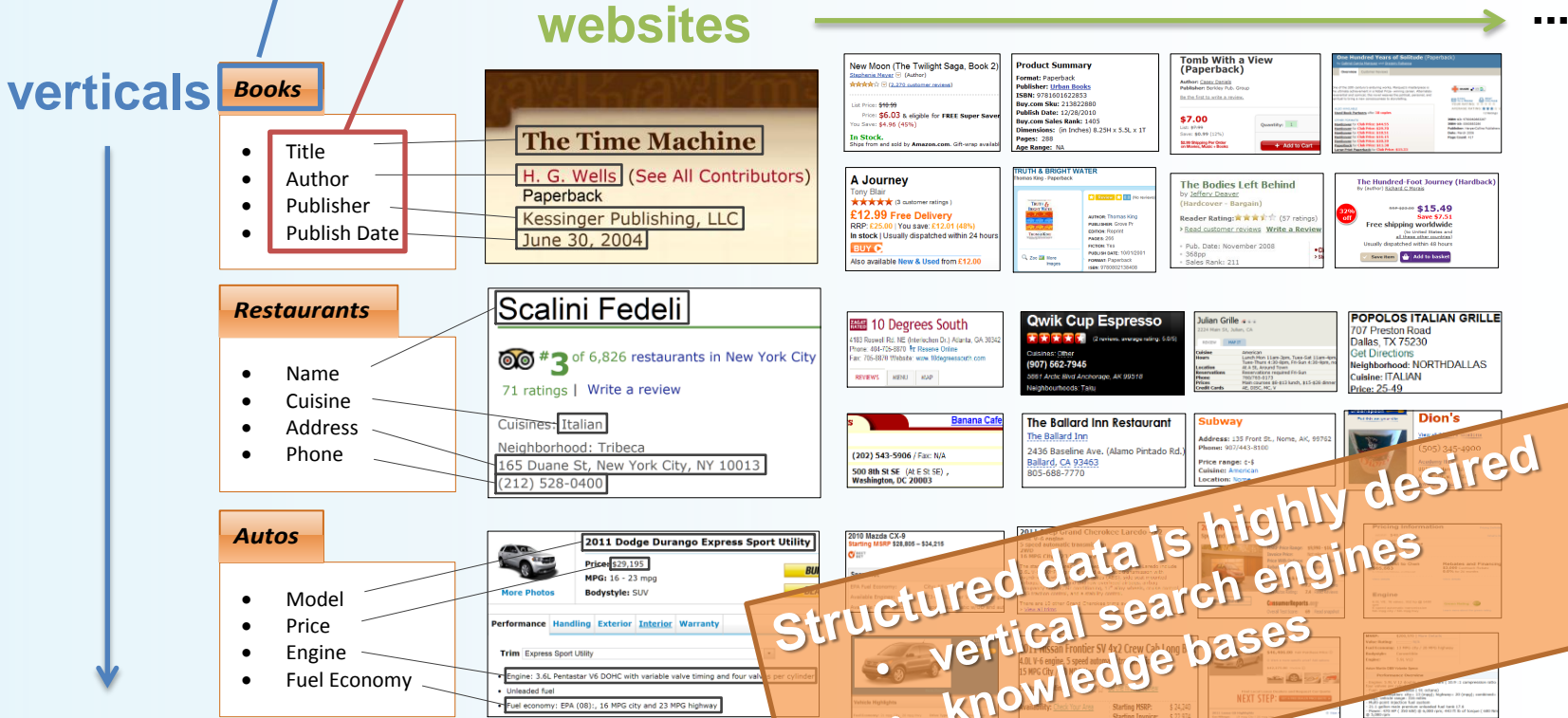
Title	Author	Publish Date	...
The Kite Runner	Khaled Hosseini	April 2004	...
Mercy	Toni Morrison	2008	...
The Time Machine	H. G. Wells	June 30, 2004	...

a data record = attribute values of an entity



We Need Structured Data

- A **vertical** is a category of entities associated with similar **attributes** (e.g., each **book** has **title/author/...**)



Challenges

- Example: Vertical = **Book**, Attribute = **Pub. Date**

a page from site 1

a page from site 2

the same entity

The Kite Runner
by [Khaled Hosseini](#)
(Paperback - Reprint)
Reader Rating: ★★★★★ (1377 ratings)
> [Read customer reviews](#) [Write a Review](#)

- Pub. Date: April 2004 ✓
- 400pp
- Sales Rank: 644

Customer Reviews

If you've read this book, tell the world how you liked it...

[Write a Review](#)

Sort by: Most Recent | Highest to Lowest | **Most Helpful**

★★★★★ songcatchers **October 25, 2008** ✗
unforgettable.

About the The Kite Runner

The timely and critically acclaimed debut novel phenomenon...

[more](#)

Publishing Details
Publisher: Penguin Group (USA) Incorporated
Date: April 01, 2004
ISBN13: 9781594480003
ISBN: 1594480001
BINC: 7524841
Age:18 and up

different value formats

Attribute value variations across sites

Noisy page contents

Challenges (contd.)



New Moon (The Twilight Saga, Book 2)
 Stephenie Meyer (Author)
 ★★★★★ (2,270 customer reviews)
 List Price: \$10.99
 Price: **\$6.03** & eligible for **FREE Super Saver**
 You Save: \$4.96 (45%)
In Stock.
 Ships from and sold by Amazon.com. Gift-wrap available

Product Summary
 Format: Paperback
 Publisher: Urban Books
 ISBN: 9781501522853
 Buy.com SKU: 213822880
 Publish Date: 12/28/2010
 Buy.com Sales Rank: 1405
 Dimensions: (in Inches) 8.25H x 5.5L x 1T
 Pages: 288
 Age Range: NA

A Journey
 Tony Blair
 ★★★★★ (3 customer ratings)
£12.99 Free Delivery
 RRP: £25.00 | You save: £12.01 (48%)
In stock | Usually dispatched within 24 hours
BUY
 Also available **New & Used** from **£12.00**

TRUTH & BRIGHT WATER
 Thomas King - Paperback
 AUTHOR: Thomas King
 PUBLISHER: Grove Pr
 EDITION: Reprint
 PAGES: 260
 FICTION: Yes
 PUBLISH DATE: 10/01/2001
 FORMAT: Paperback
 ISBN: 9780802138408

Tomb With a View (Paperback)
 Author: Casey Daniels
 Publisher: Berkley Pub. Group
 Be the first to write a review.
\$7.00
 List: \$7.99
 Save: \$0.99 (12%)
\$2.99 Shipping Per Order on Movies, Music & Books
 Add to Cart

One Hundred Years of Solitude (Paperback)
 Gabriel Garcia Marquez
 Overview Customer Reviews
 One of the 20th century's enduring masterworks, Marquez's masterpiece is a sublime achievement in novel form, weaving magic, allegory, metaphor and comedy into a novel whose mythical, personal, and social meanings are inseparable from its storytelling.
Hardcover for \$15.99
 ISBN-13: 9780000000000
 Hardcover to Club Price: \$14.50
 ISBN-10: 0000000000
 Publisher: HarperCollins Publishers
 Date: March 2005
 Hardcover to Club Price: \$18.54
 ISBN-13: 9780000000000
 Hardcover to Club Price: \$20.39
 ISBN-10: 0000000000
 Hardcover to Club Price: \$15.33

The Bodies Left Behind
 by Jeffrey Deaver
 (Hardcover - Bargain)
 Reader Rating: ★★★★★ (5)
 > Read customer reviews Write a review
 • Pub. Date: November 2008
 • 368pp
 • Sales Rank: 211

The Hundred-Foot Journey (Hardback)
 By (author) Richard C. Morais

Page layout variations across sites

Autos

- Model
- Price
- Engine
- Fuel Economy

Books

- Title
- Author
- Publisher
- Publish Date

Jobs

- Title
- Company
- Location
- Date

Movies

- Title
- Director
- Genre
- Rating

Restaurants

- Name
- Cuisine
- Address
- Phone

Universities

- Name

Various verticals & attributes

Existing Solutions



Manual solutions

- **Pros:** highly accurate
- **Cons:** labor-intensive; difficult to scale up

Kushmerick(PhD thesis '97)
Muslea et al.(AGENTS'99)
Soderland(Mach.Learn.'99)
Zheng et al.(KDD'07)
...

Semi-automatic solutions

- **Pros:** automatically locate data in templates
- **Cons:** need to annotate semantics manually

Crescenzi et al.(VLDB'01)
Arasu et al.(SIGMOD'03)
Liu et al.(KDD'03)
Zhai et al.(WWW'05)
...

Automatic solutions

- **Pros:** extract data with specified semantics
- **Cons:** need strong features and/or abundant training data

Zhu et al.(ICML'05,KDD'06)
Carlson et al.(ECML'08)
Wong et al.(SIGIR'08 & '09)
Yang et al.(WWW'09)
...

Our Goal

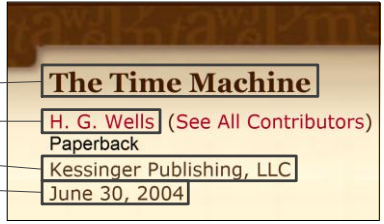
- A unified solution for extracting structured data with:

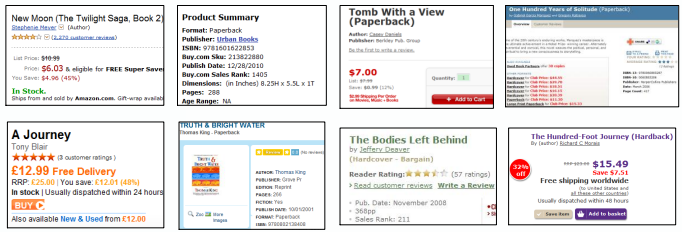
Minimal human effort

- Label **one** seed site for each vertical → many unseen sites

Books

- Title
- Author
- Publisher
- Publish Date





Flexibility for verticals

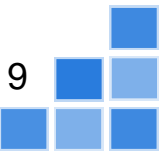
- Handle **various** verticals & attributes without redesign

<p>Autos</p> <ul style="list-style-type: none"> Model Price Engine Fuel Economy 	<p>Books</p> <ul style="list-style-type: none"> Title Author Publisher Publish Date
<p>Jobs</p> <ul style="list-style-type: none"> Title Company Location Date 	<p>Restaurants</p> <ul style="list-style-type: none"> Name Cuisine Address Phone

Outline



- Motivation & Challenges
- **Our Solution**
 - Main Idea & Framework Overview
 - Feature Extraction
 - Vertical Knowledge Learning
 - Vertical Knowledge Adaptation
- Experimental Results
- Summary



Our Solution: Main Idea



Flexible for various verticals & attributes

Robust to variations across websites

General Features

Loose Classifiers



Recall↑

Combine

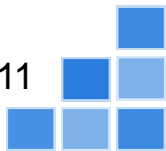
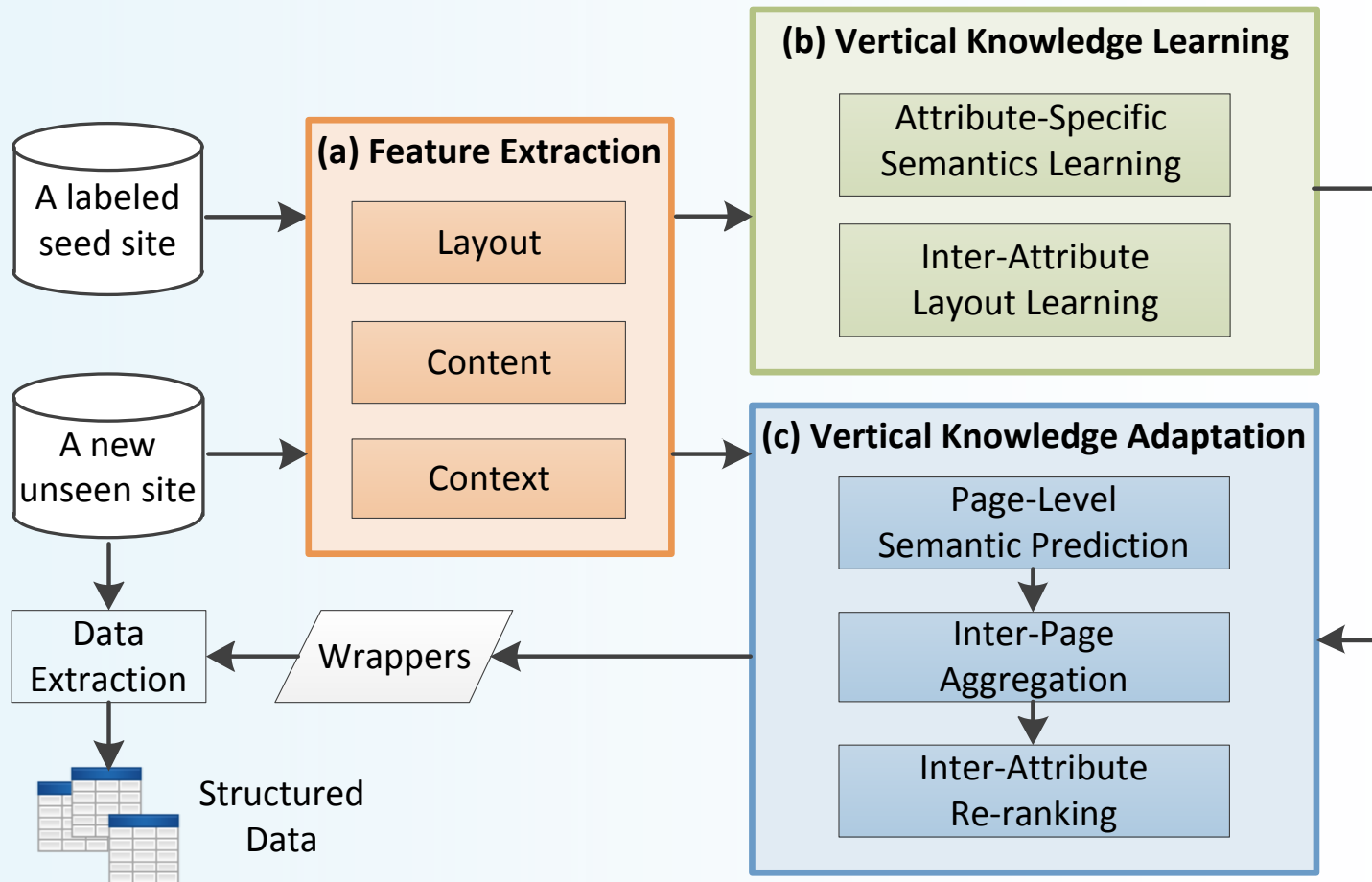
Site-Level Constraints

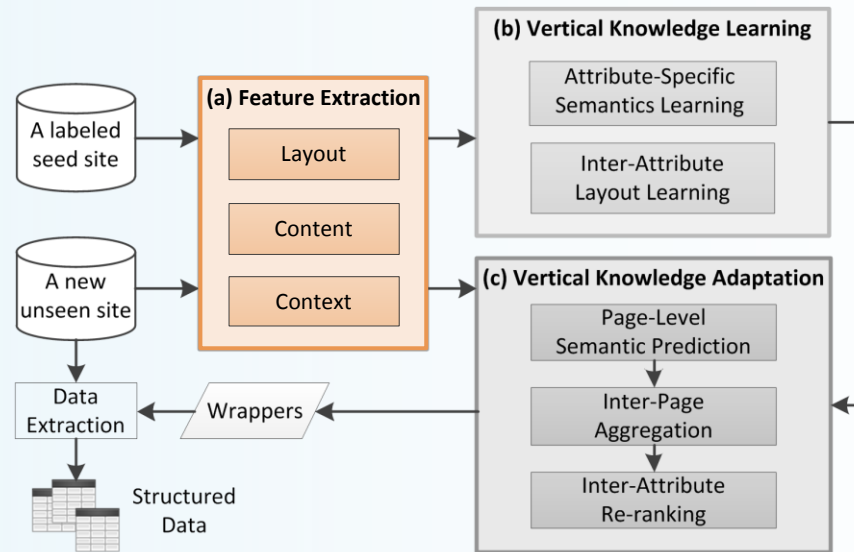
Precision↑

Web pages are generated by site-level templates



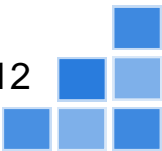
Framework Overview





Feature Extraction

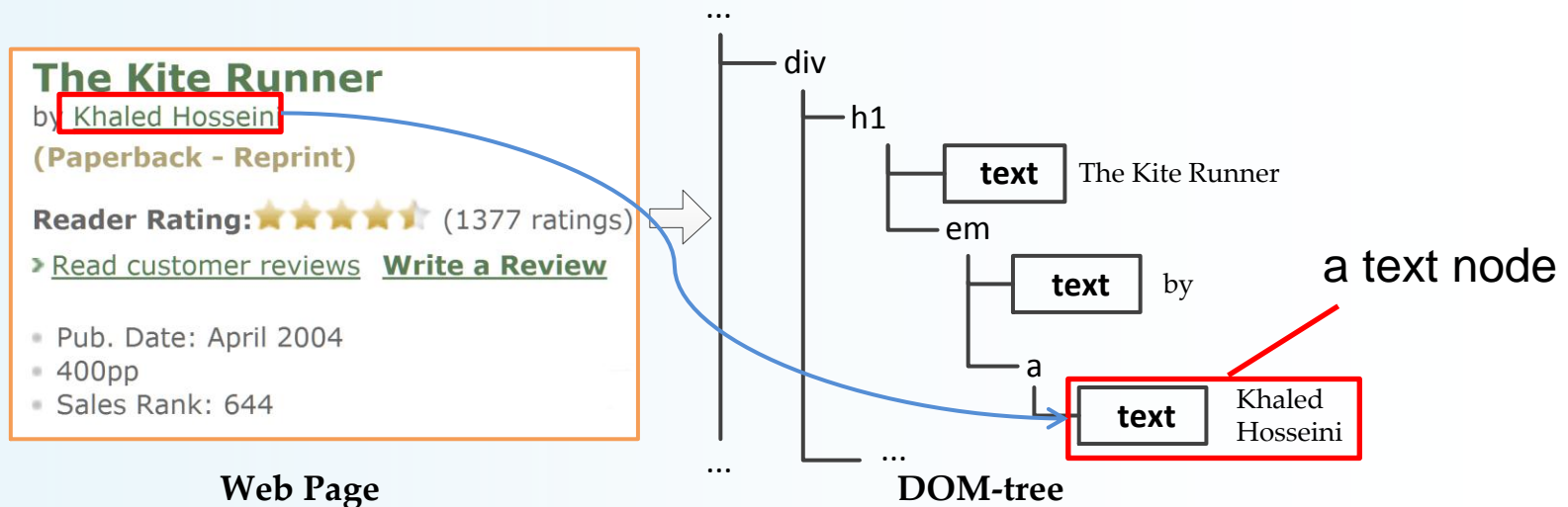
- Layout
- Content
- Context



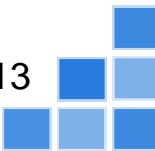
General Features of Web Pages



- Extract features from *text nodes* in *DOM trees* of web pages



Three types of features: *layout, content, context*



Layout Features

- Goal: characterize the *position* of a text node

Visual Position = (24, 798)



The Kite Runner
by **Khaled Hosseini**
(Paperback - Reprint)

Reader Rating: ★★★★★ (1377 ratings)
[Read customer reviews](#) [Write a Review](#)

- Pub. Date: April 2004
- 400pp
- Sales Rank: 644

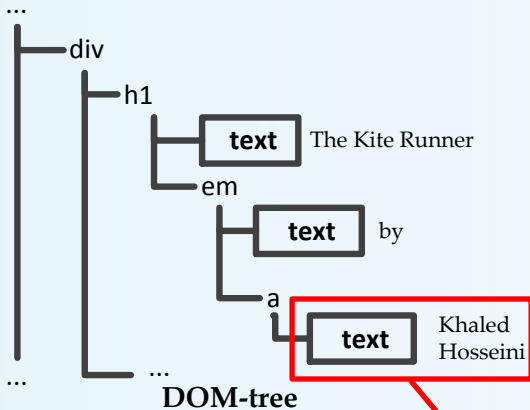
Visual Position

position in a rendered page
= coordinates to the top left

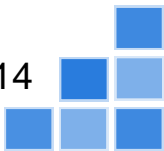
DOM Path

position in a DOM tree
= root-to-leaf tag path

Web Page



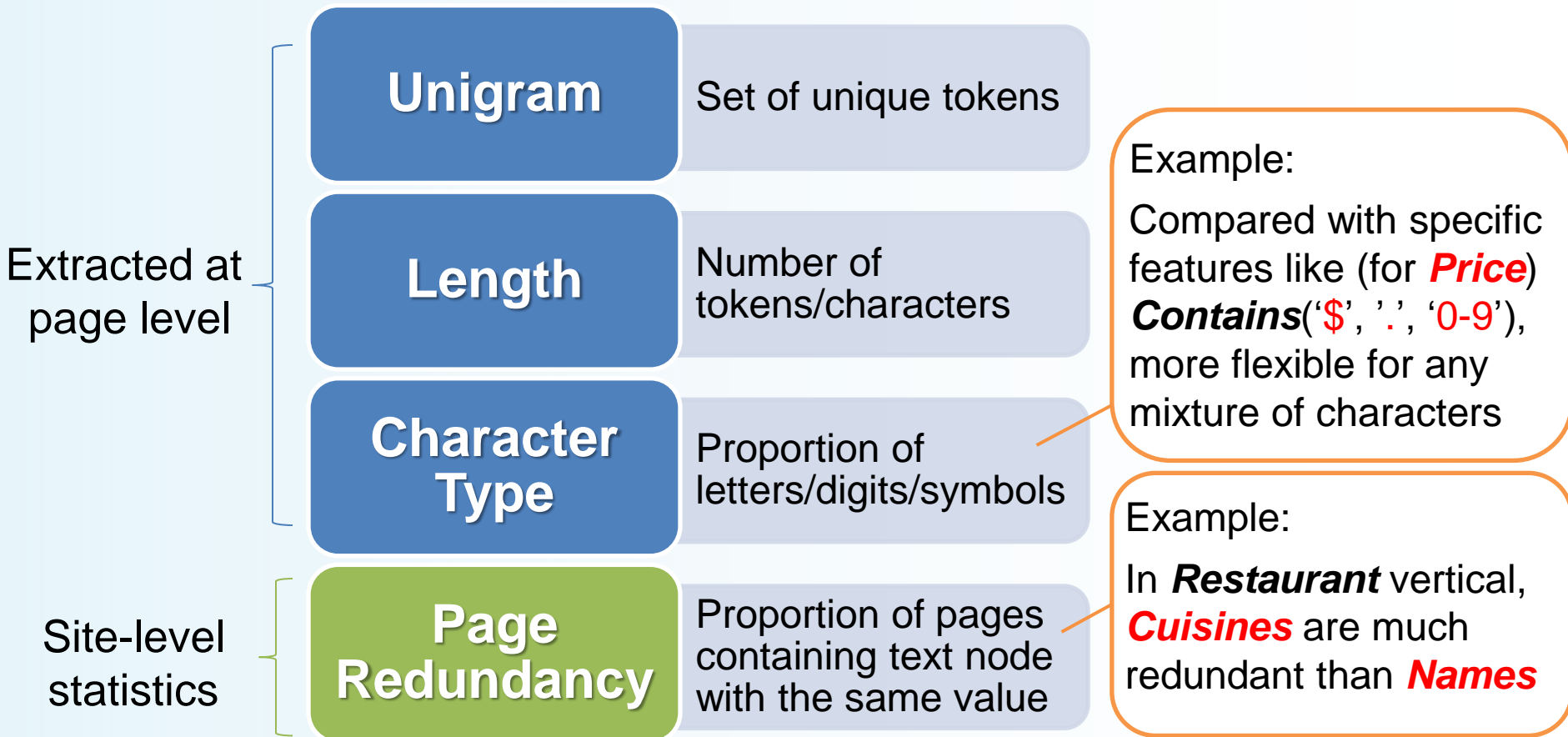
DOM Path = /html/body/div/div/div/div/h1/em/a/text



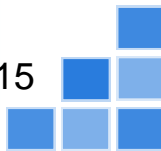
Content Features



- Derived from the *value* contained in a text node



General enough to characterize various attributes



Context Features



- Motivation
 - Surrounding text indicates semantics of text nodes
 - Text nodes with identical context → similar semantics

Extracted at
page level

**Preceding
Text**

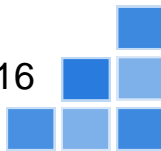
values of visually
preceding text nodes

Site-level
statistics

**Prefix &
Suffix**

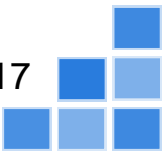
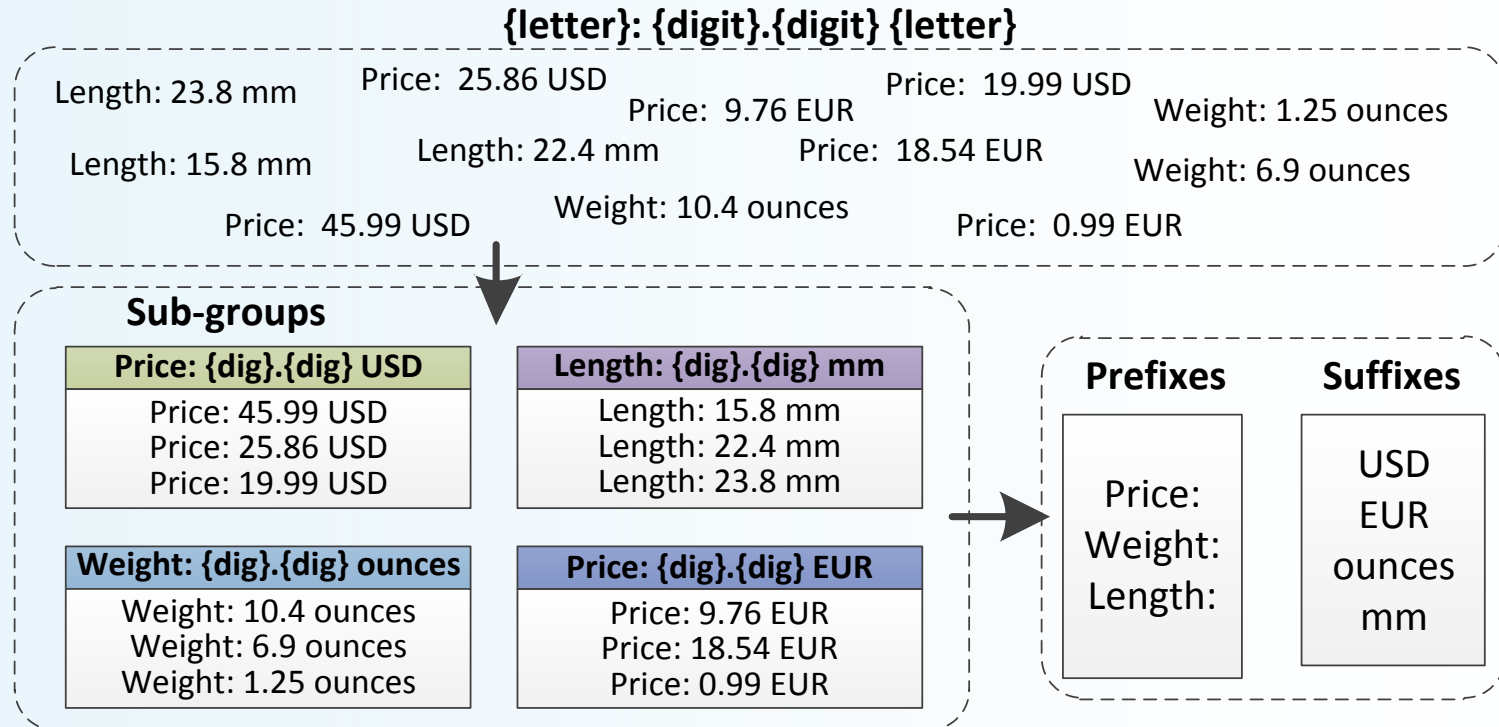
static (across pages)
sub-strings of text node values

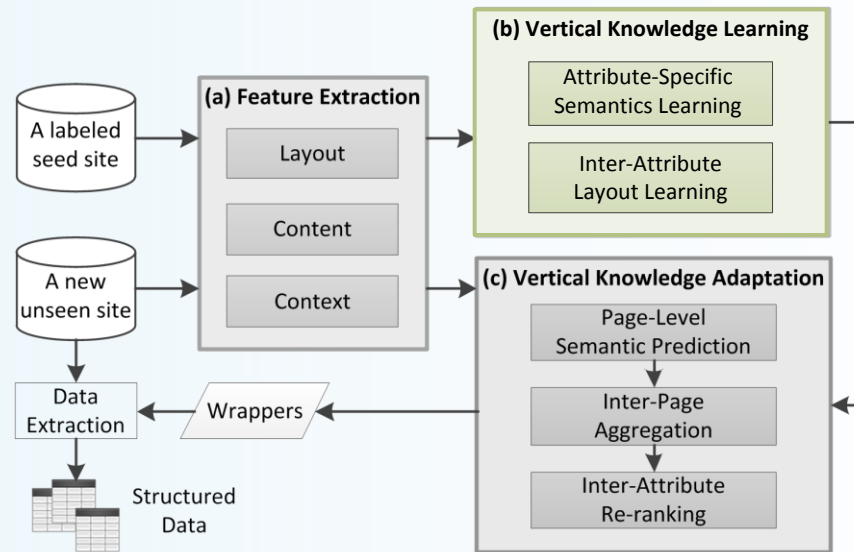
Extracted automatically



Context Features

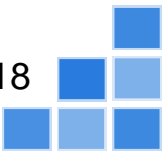
- Motivation
 - Surrounding text indicates semantics of text nodes
 - Text nodes with identical context → similar semantics





Vertical Knowledge Learning

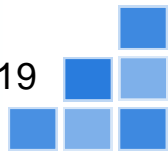
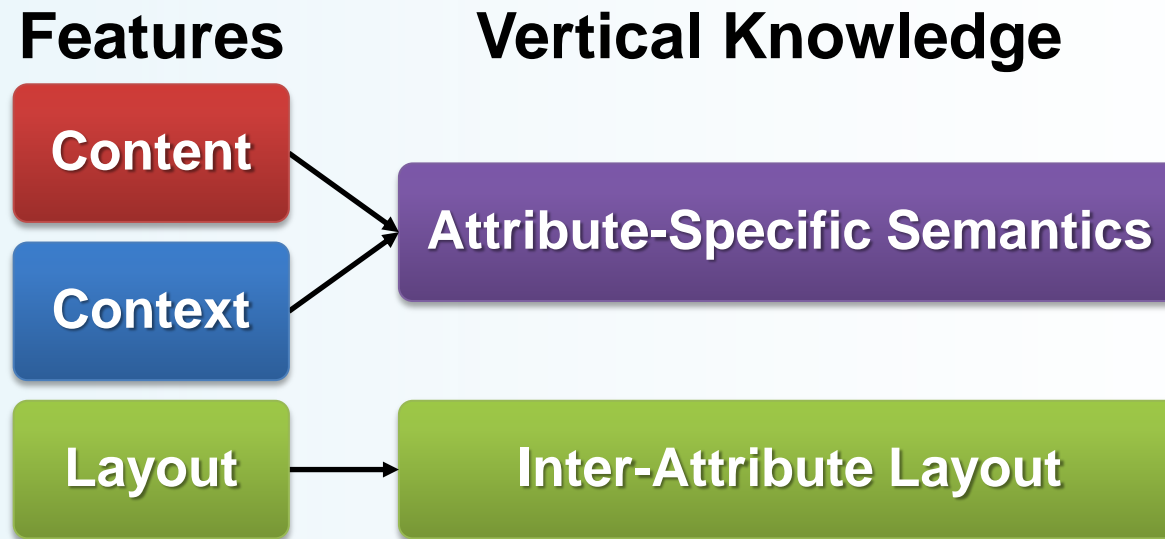
- Attribute-Specific Semantics
- Inter-Attribute Layout



Features to Vertical Knowledge

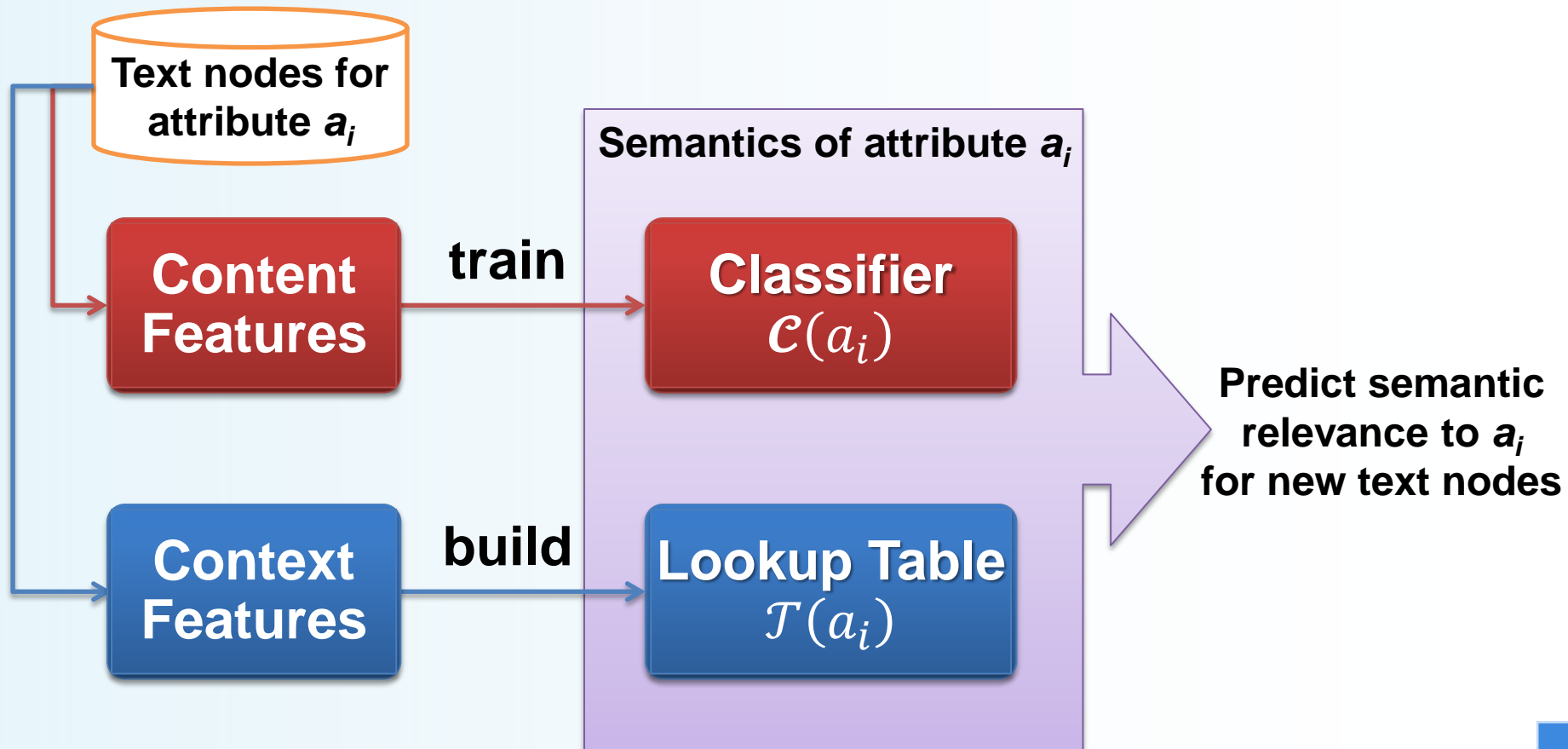


- Goal
 - Learn knowledge from a labeled seed site based on features extracted from text nodes
 - Guide data extraction from unseen sites
- Two types of vertical knowledge



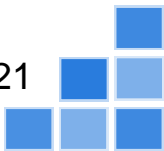
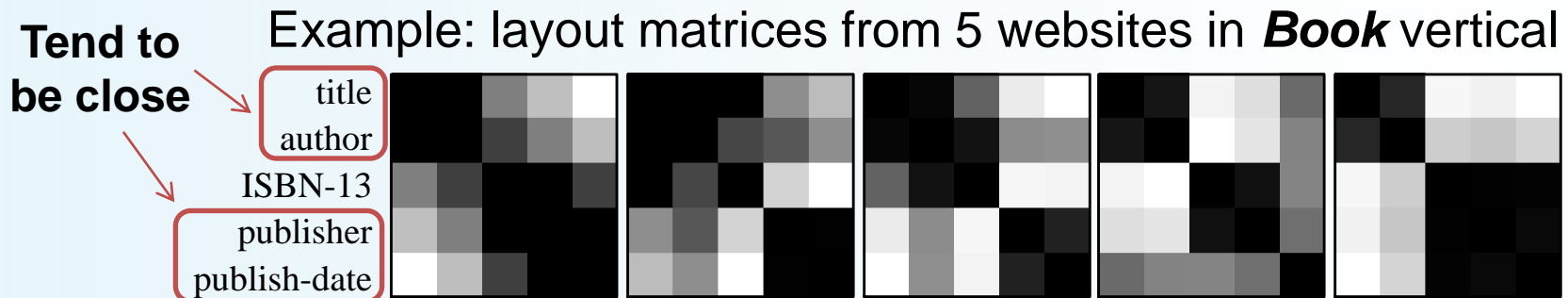
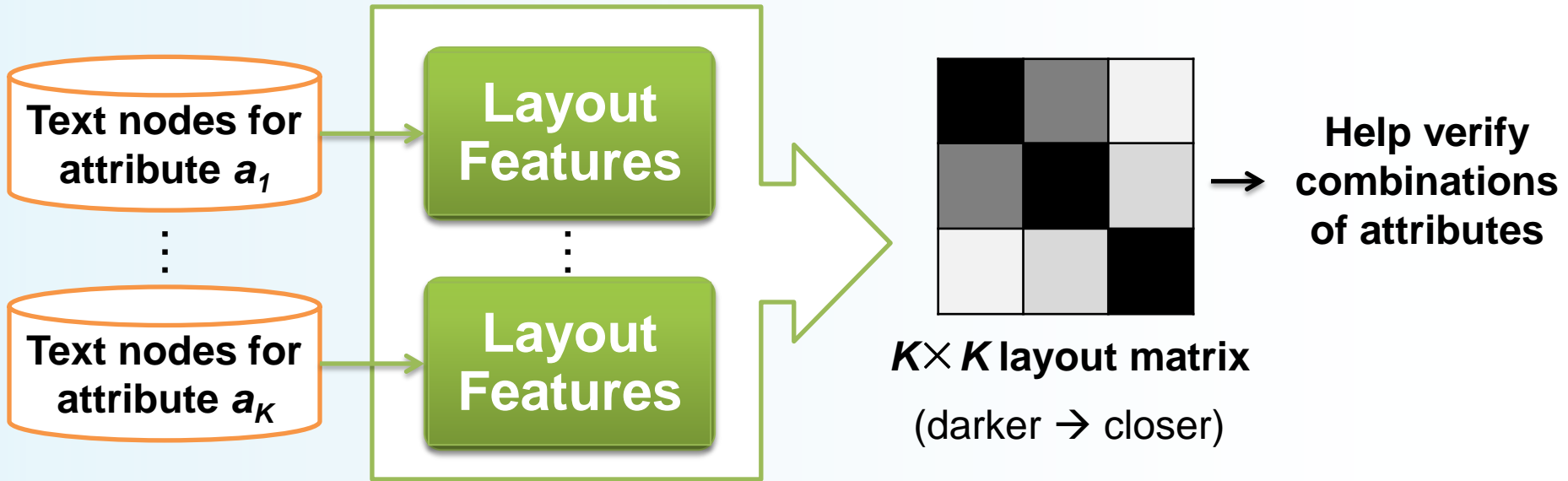
Attribute-Specific Semantics

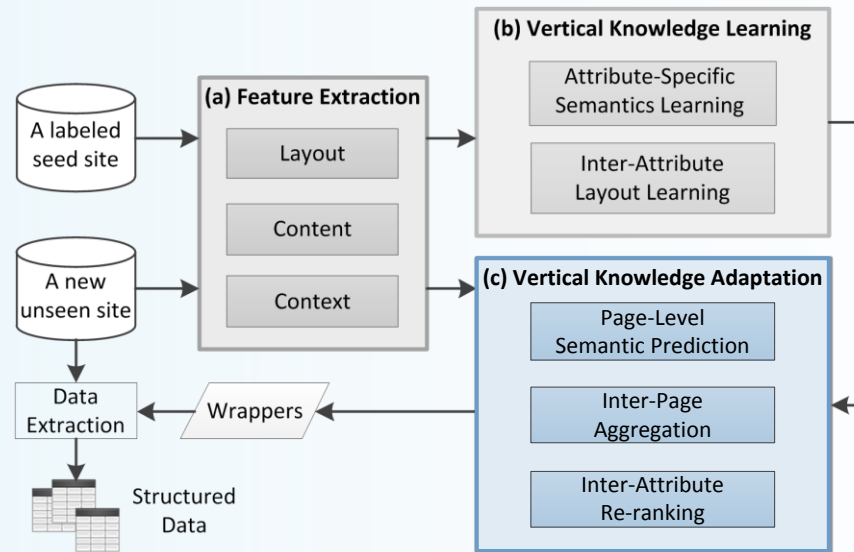
- **Content** features \rightarrow *classifiers* (e.g., SVMs)
- **Context** features \rightarrow (token-score) *lookup tables*



Inter-Attribute Layout

- Construct a $K \times K$ layout matrix from **layout** features
- Encode **pairwise** distances between K attributes





Vertical Knowledge Adaptation

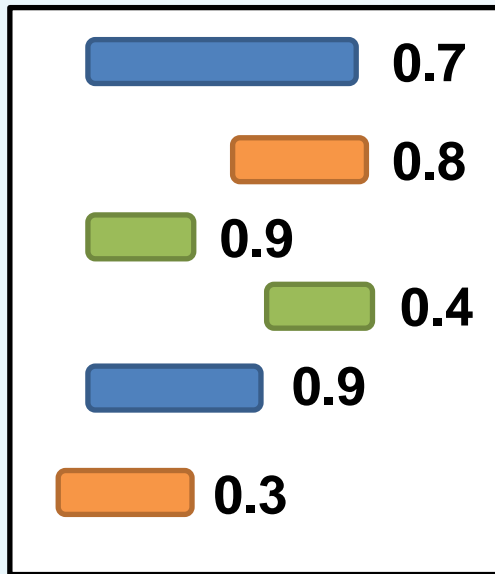
- Page-Level Semantic Prediction
- Inter-Page Aggregation
- Inter-Attribute Re-ranking



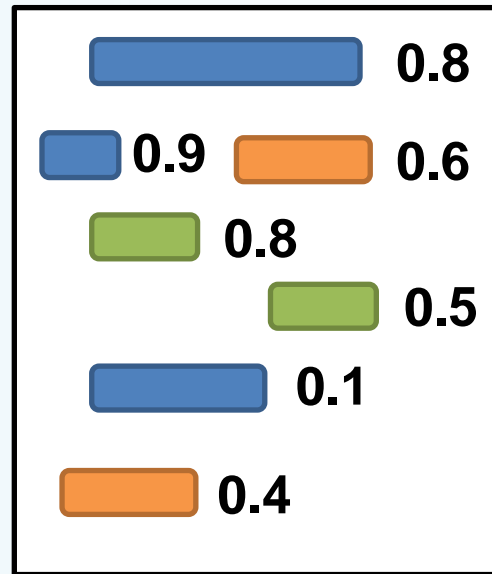
Page-Level Semantic Prediction

Attribute-specific semantics → page-level candidates

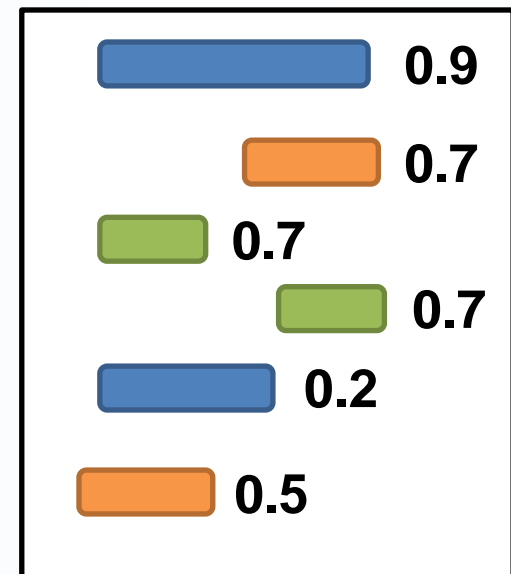
Attributes: A B C



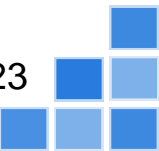
Page 1



Page 2



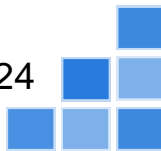
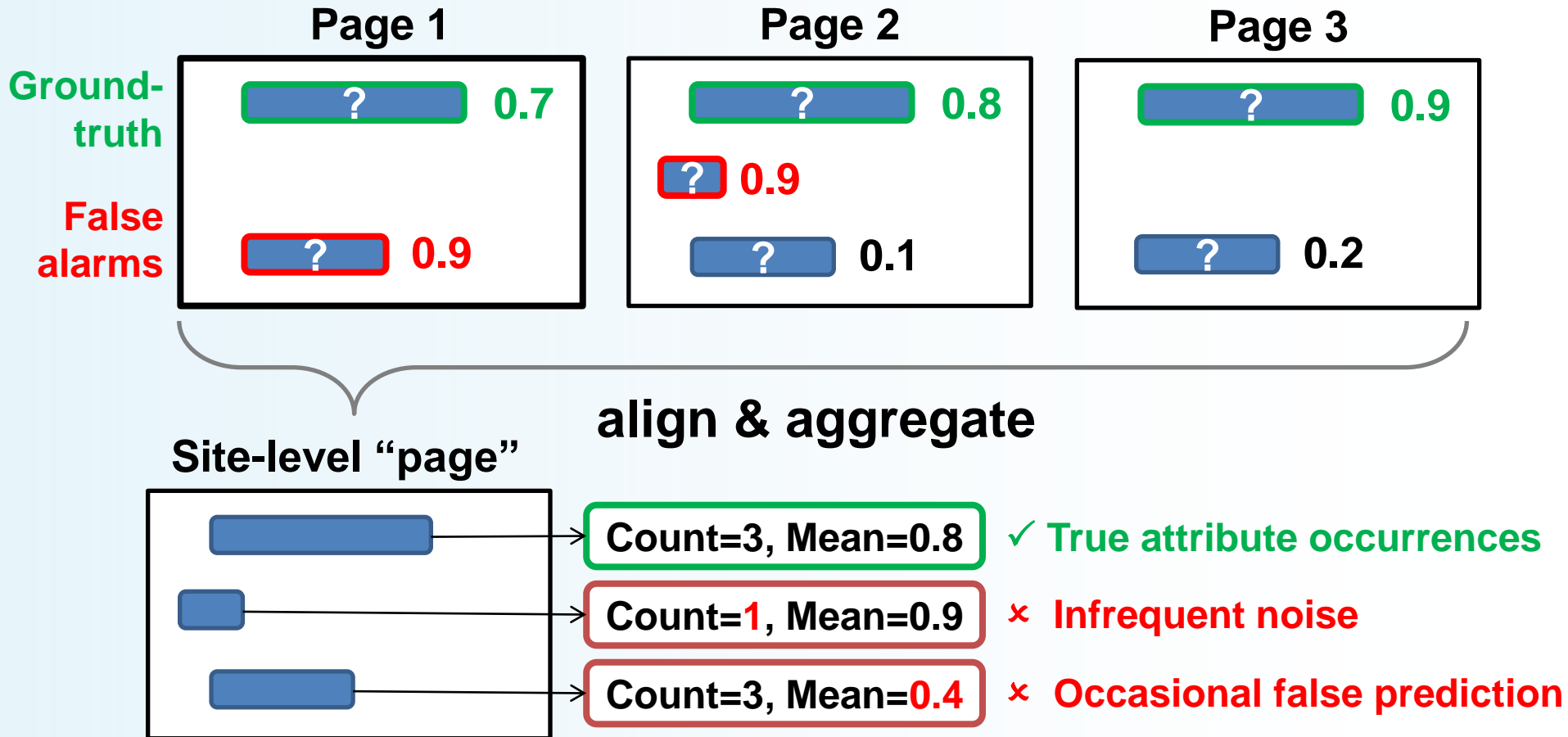
Page 3



Inter-Page Aggregation

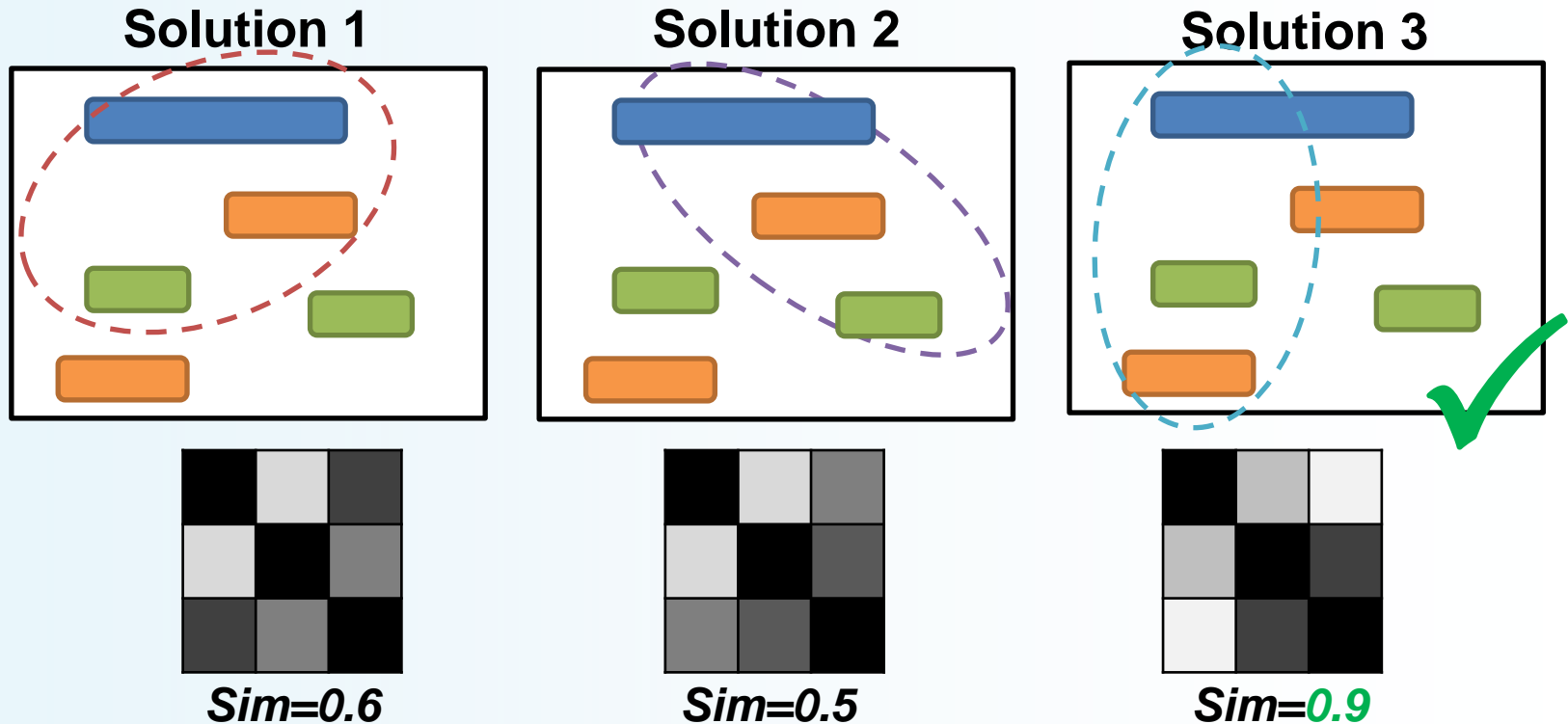


- For each attribute: multiple candidates per page →



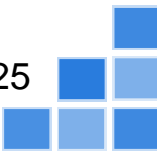
Inter-Attribute Re-ranking

- Multiple possible solutions (attribute combinations)

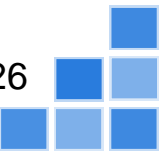
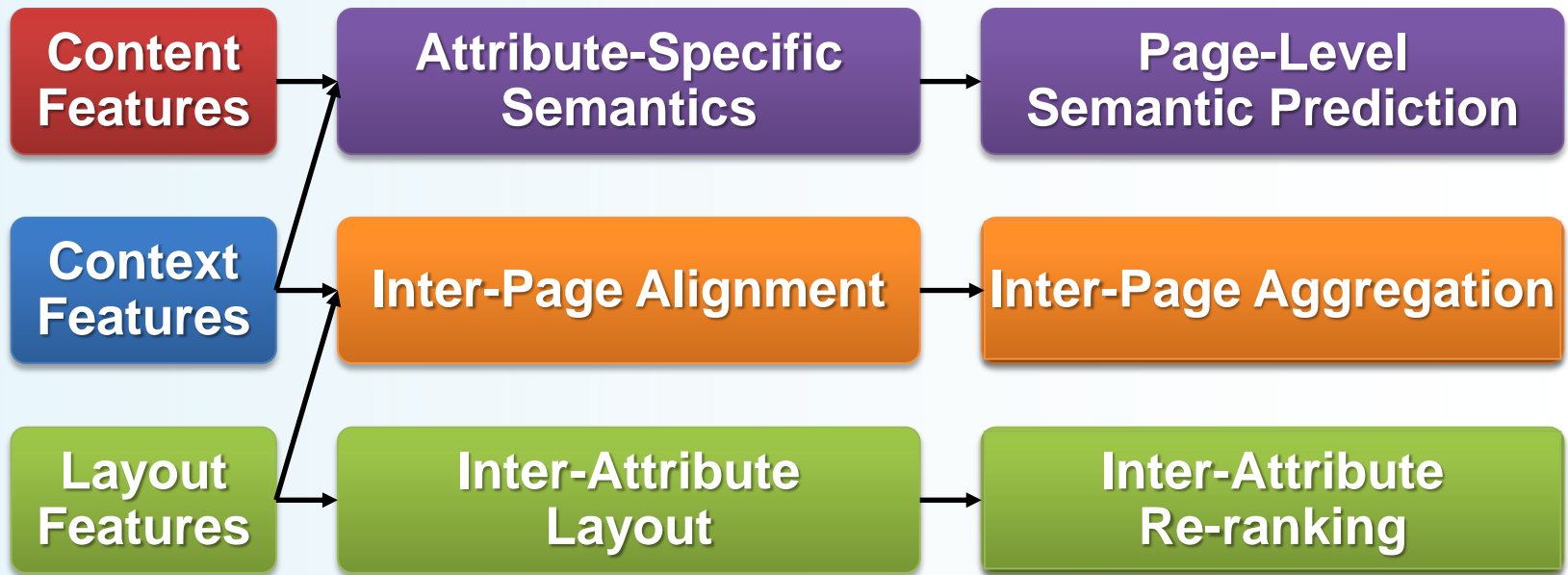


Inter-attribute layout
learnt from the seed site

Re-rank candidates by
measuring similarity



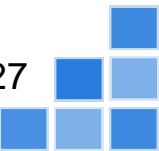
Summary: Flowchart of Features



Outline



- Motivation & Challenges
- Our Solution
 - Main Idea & Framework Overview
 - Feature Extraction
 - Vertical Knowledge Learning
 - Vertical Knowledge Adaptation
- **Experimental Results**
- Summary



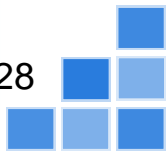
A Large-Scale Dataset



(Publicly available at <http://swde.codeplex.com>)

- **8 verticals** with diverse semantics
- **80 websites** (10 per vertical)
- **124,291 pages** (200~2,000 per website)
- **32 attributes** (3~5 per vertical) with labeled ground-truth

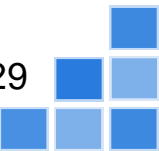
Vertical	Attributes
<i>Autos</i>	model, price, engine, fuel_economy
<i>Books</i>	title, author, ISBN, publisher, pub_date
<i>Cameras</i>	model, price, manufacturer
<i>Jobs</i>	title, company, location, date
<i>Movies</i>	title, director, genre, rating
<i>NBA players</i>	name, team, height, weight
<i>Restaurants</i>	name, address, phone, cuisine
<i>Universities</i>	name, phone, website, type



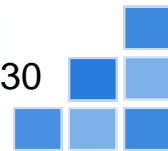
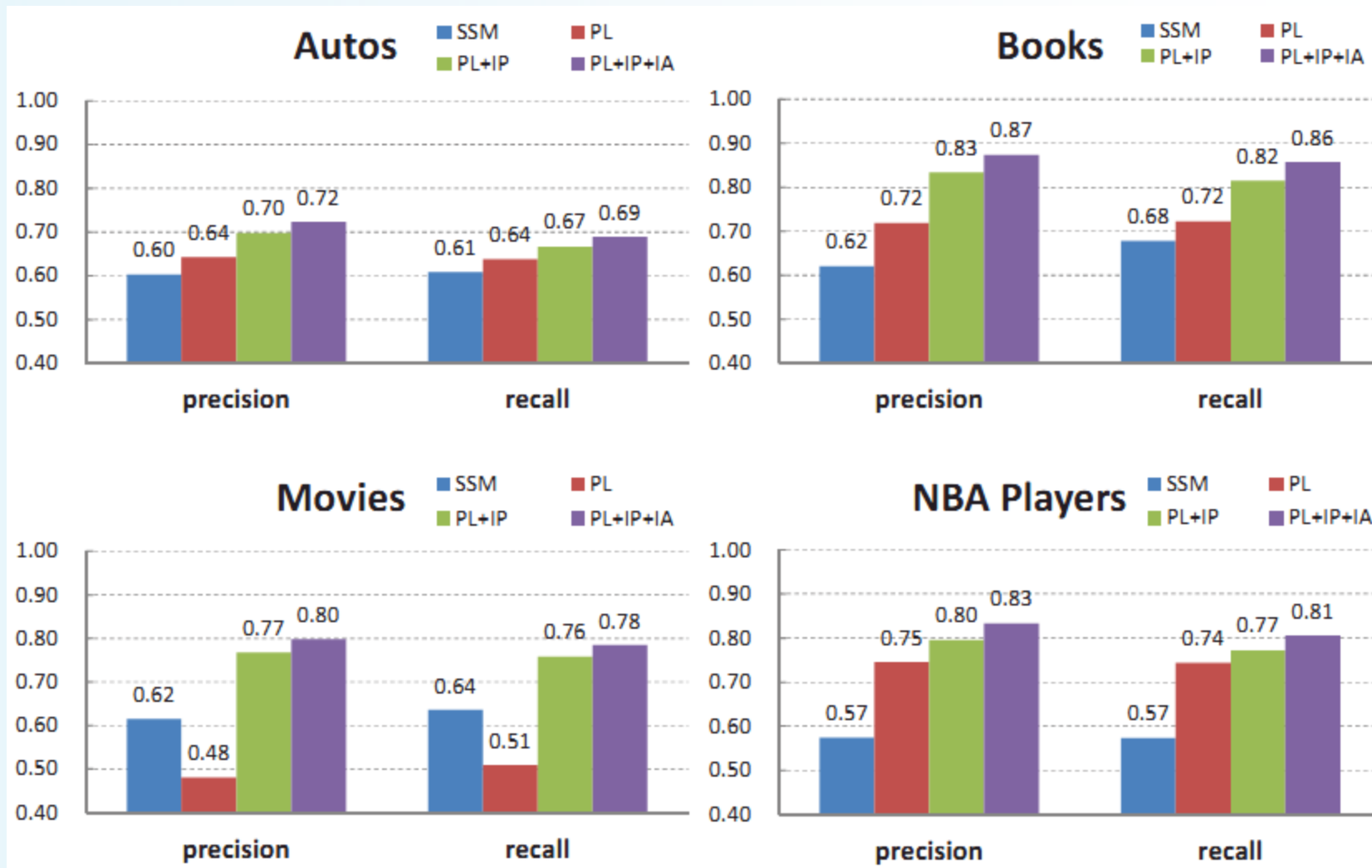
Experimental Settings



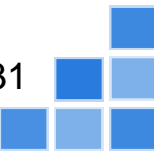
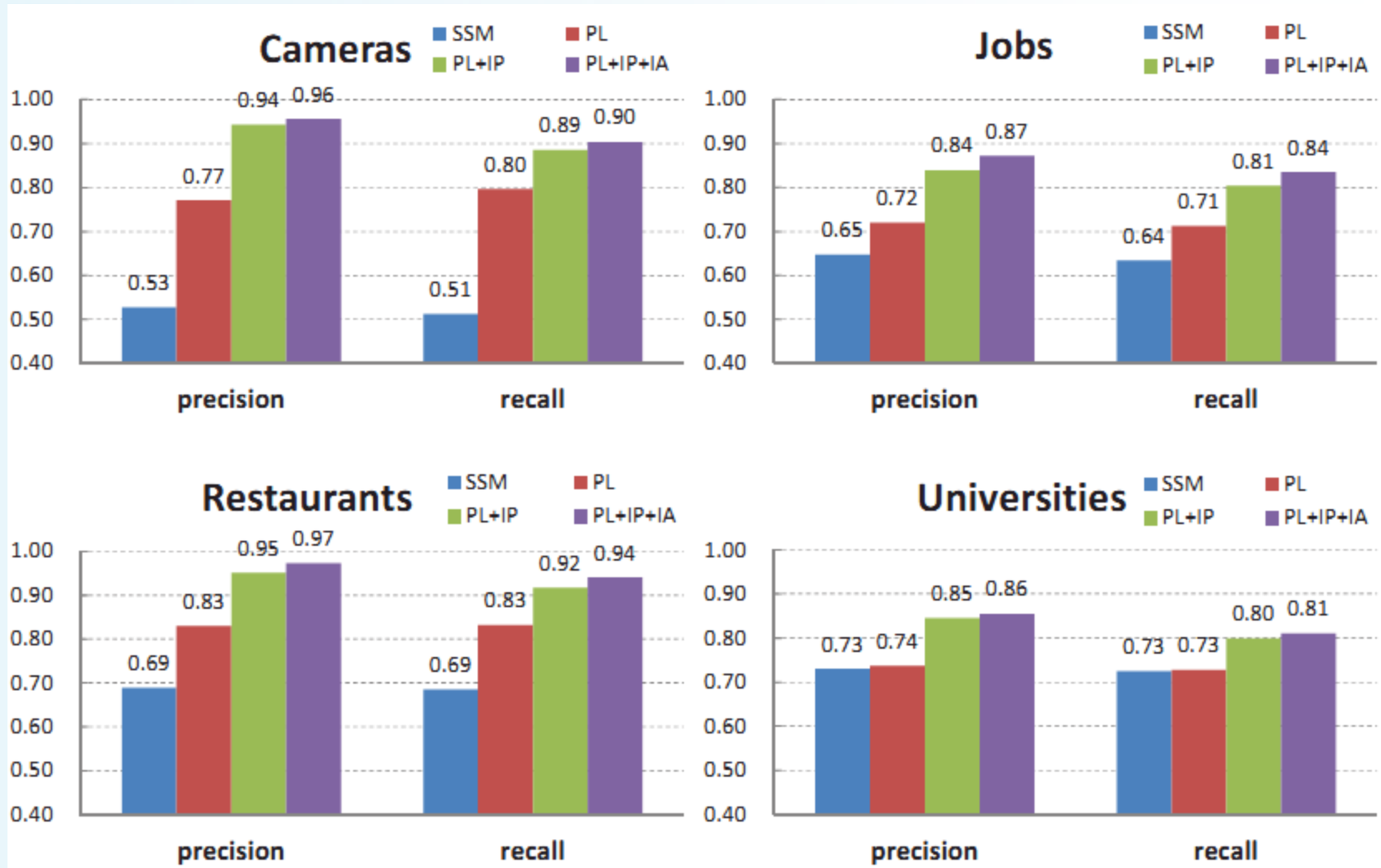
- Methods
 - <Baselines>
 1. **SSM** (Stacked Skews Model) *Carlson et al. ECML'08*
 2. **PL** (page-level semantic prediction)
 3. **PL + IP** (inter-page aggregation)
 - <Full solution>
 4. **PL + IP + IA** (inter-attribute re-ranking)
- One seed site (by turns), test on other sites
- Performance metrics: precision & recall



Performance



Performance (contd.)

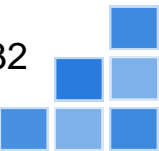


Performance: Multiple Seed Sites

- Our solution with multiple seed sites
 - Take the solution with highest confidence score
- Our solution with (one seed + bootstrapping seeds)
- SSM with multiple seed sites

Table 3: Average F-scores of the proposed solution based on multiple seed sites.

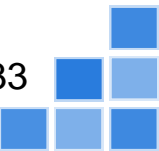
#Seeds	1	2	3	4	5
Our Solution	0.843	0.860	0.868	0.884	0.886
Our Solution (Bootstrap)	0.843	0.856	0.861	0.859	0.865
SSM	0.630	0.645	0.692	0.719	0.741



Summary



- A unified solution for structured data extraction
 - Minimal human effort: labeling one site per vertical
 - Flexible for various verticals & attributes
- A large-scale dataset (*has been published online*)
 - 124K pages from 80 websites in 8 verticals
- Promising performance
 - Precision $\geq 80\%$, Recall $\geq 80\%$ for most verticals
- Future work
 - Bootstrapping: accumulate vertical knowledge incrementally





Thank you!

Dataset available at
<http://swde.codeplex.com>

Microsoft
Research
微软亚洲研究院





Q & A

Dataset available at
<http://swde.codeplex.com>

Microsoft
Research
微软亚洲研究院

