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

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**

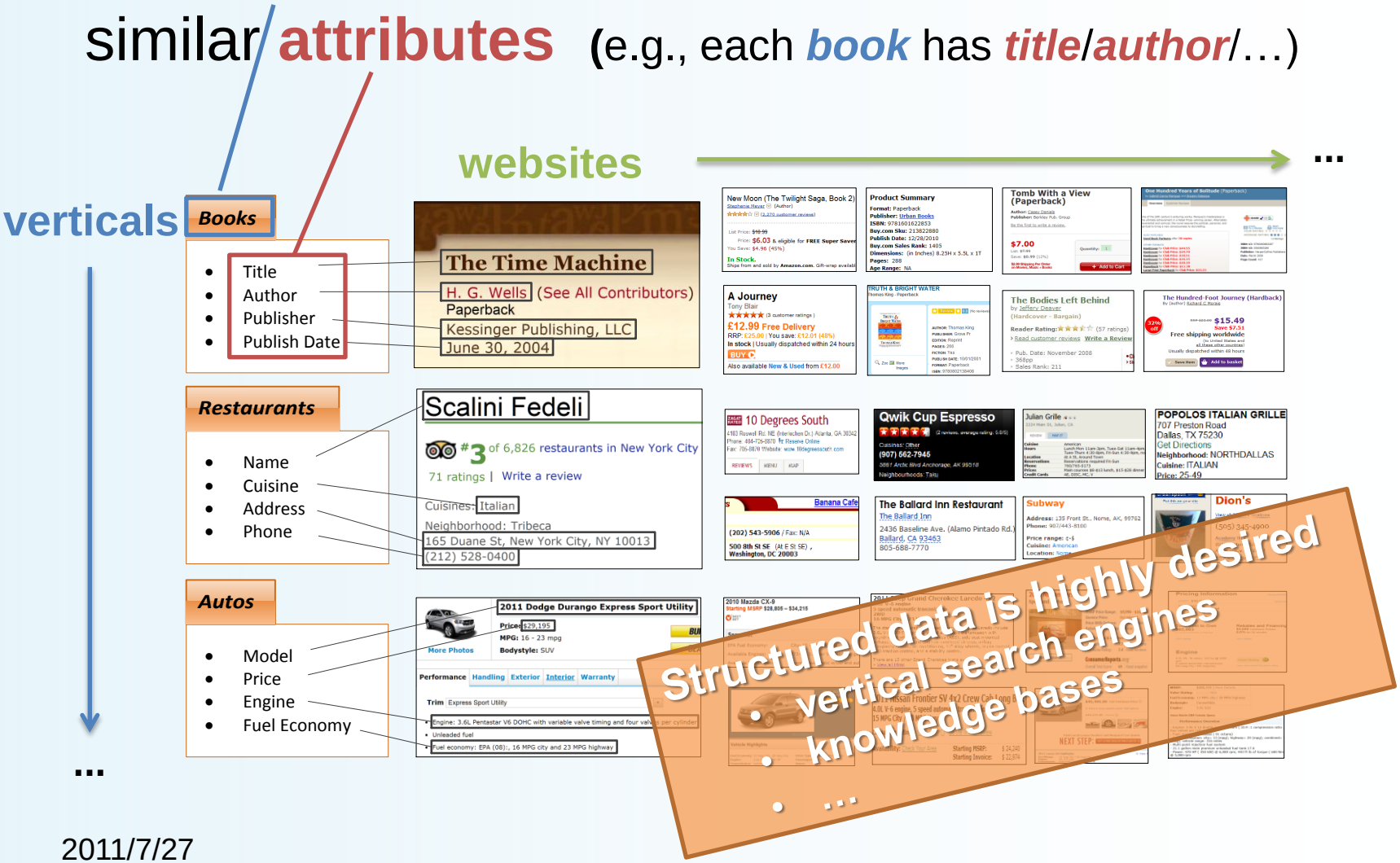
attributes

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***/…)



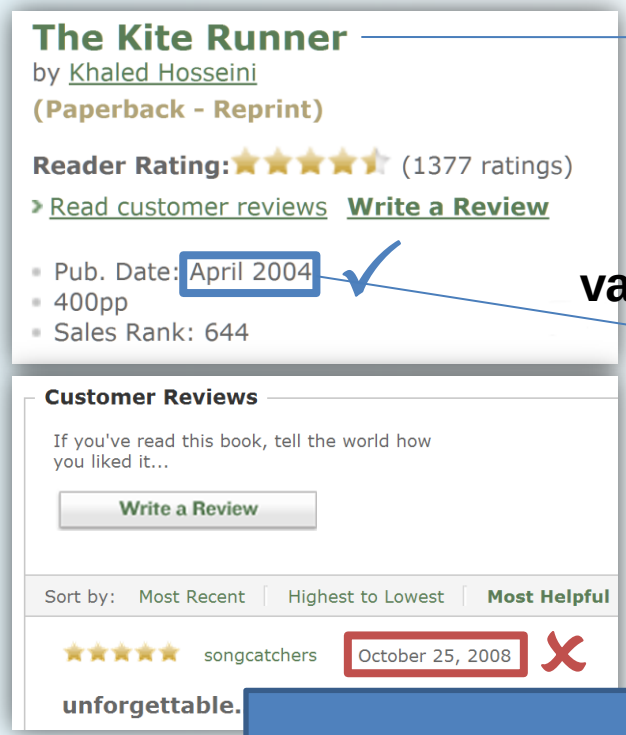
Challenges

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

a page from site 1

a page from site 2

the same entity

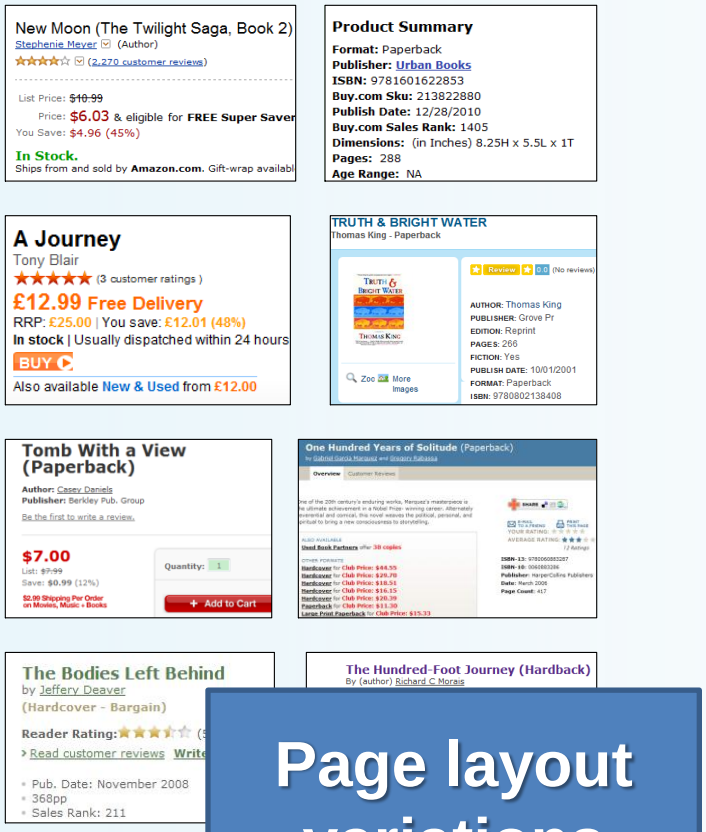


different value formats

Noisy page contents

Attribute value variations across sites

Challenges (contd.)



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

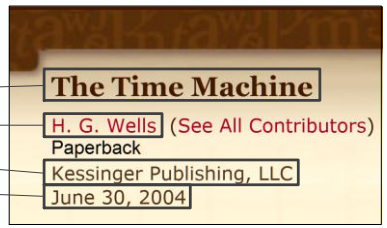
- A unified solution for extracting structured data with:

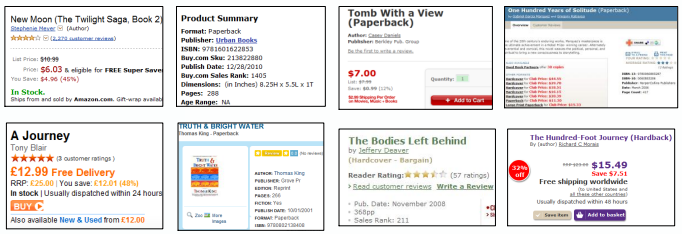
Minimal human effort

- 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

Autos

- Model
- Price
- Engine
- Fuel Economy

Books

- Title
- Author
- Publisher
- Publish Date

Jobs

- Title
- Company
- Location
- Date

Restaurants

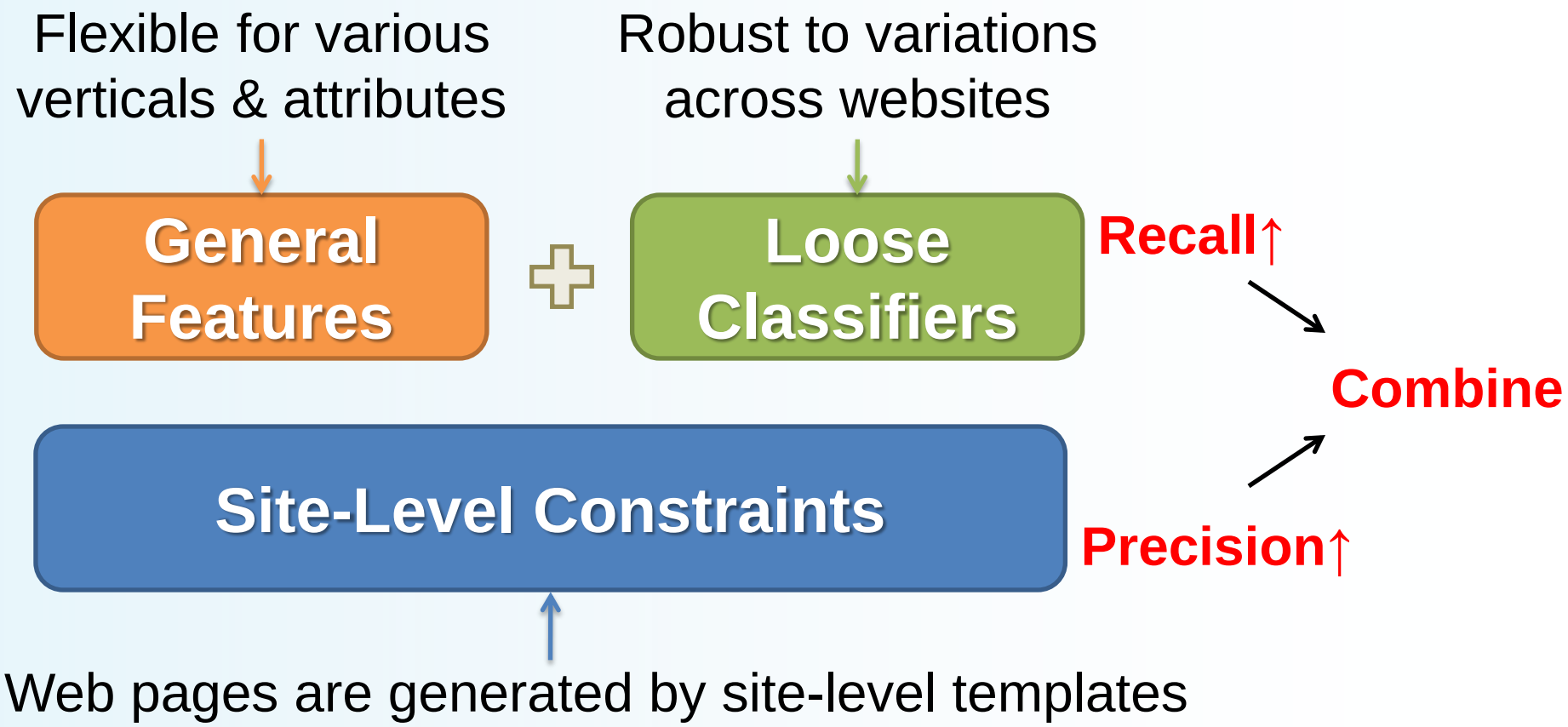
- Name
- Cuisine
- Address
- Phone

...

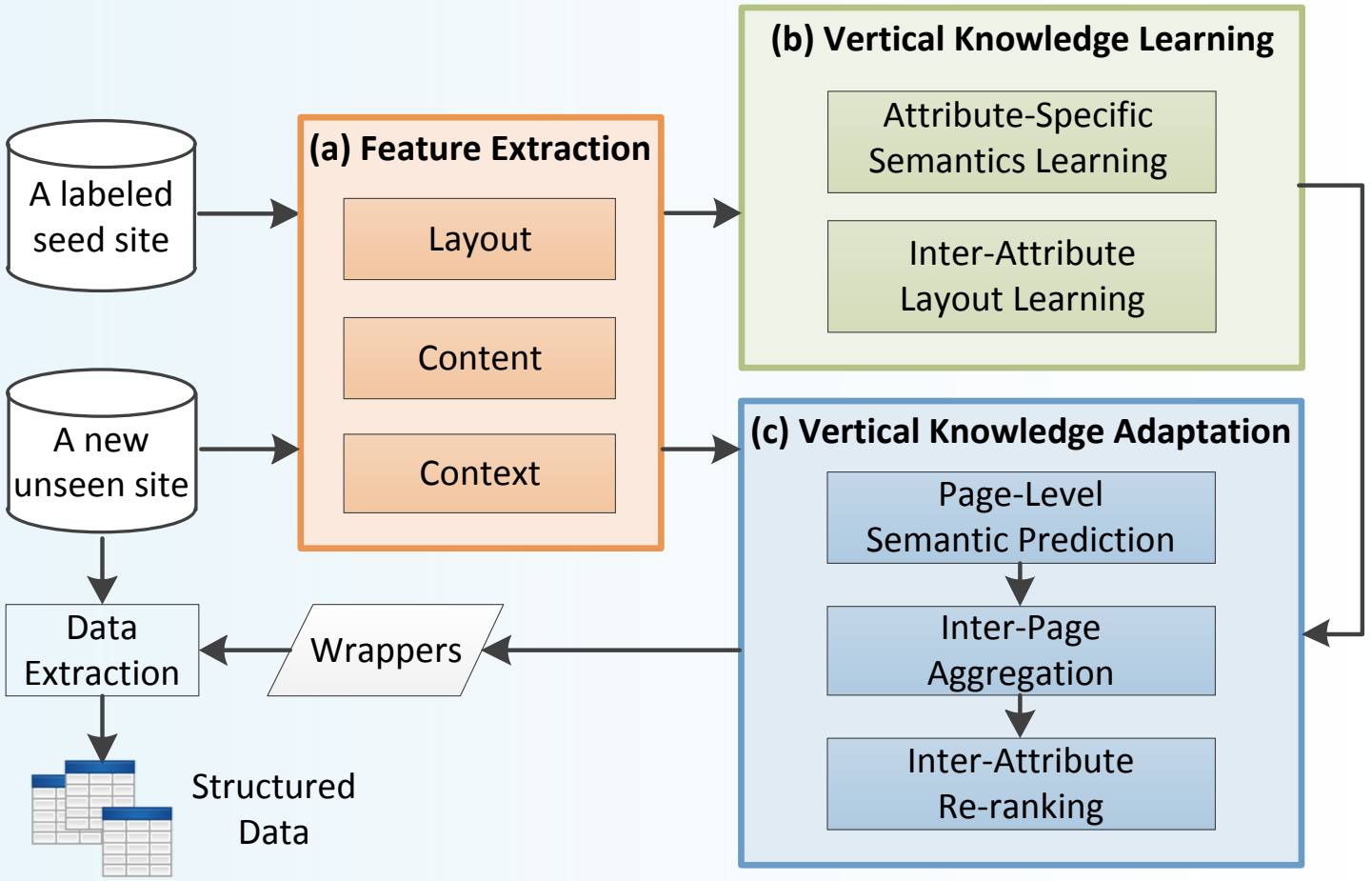
Outline

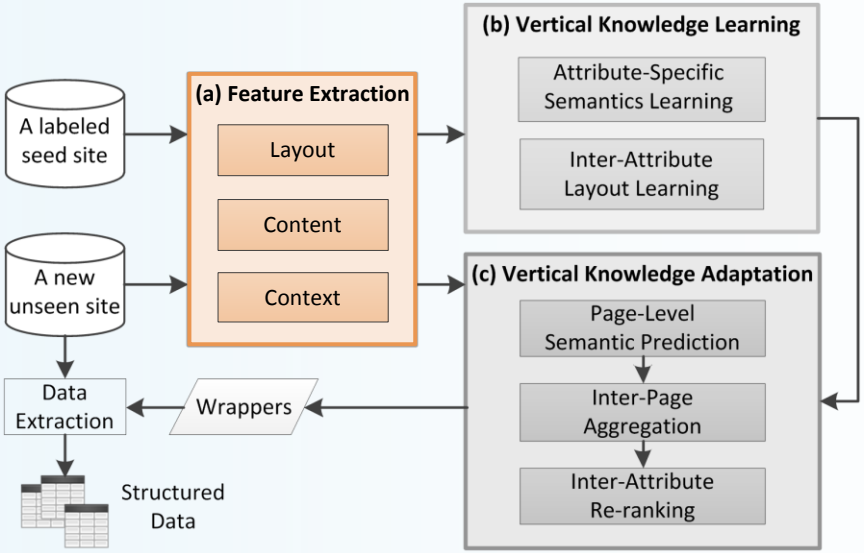
- Motivation & Challenges
- Our Solution
  - Main Idea & Framework Overview
  - Feature Extraction
  - Vertical Knowledge Learning
  - Vertical Knowledge Adaptation
- Experimental Results
- Summary

Our Solution: Main Idea



Framework Overview



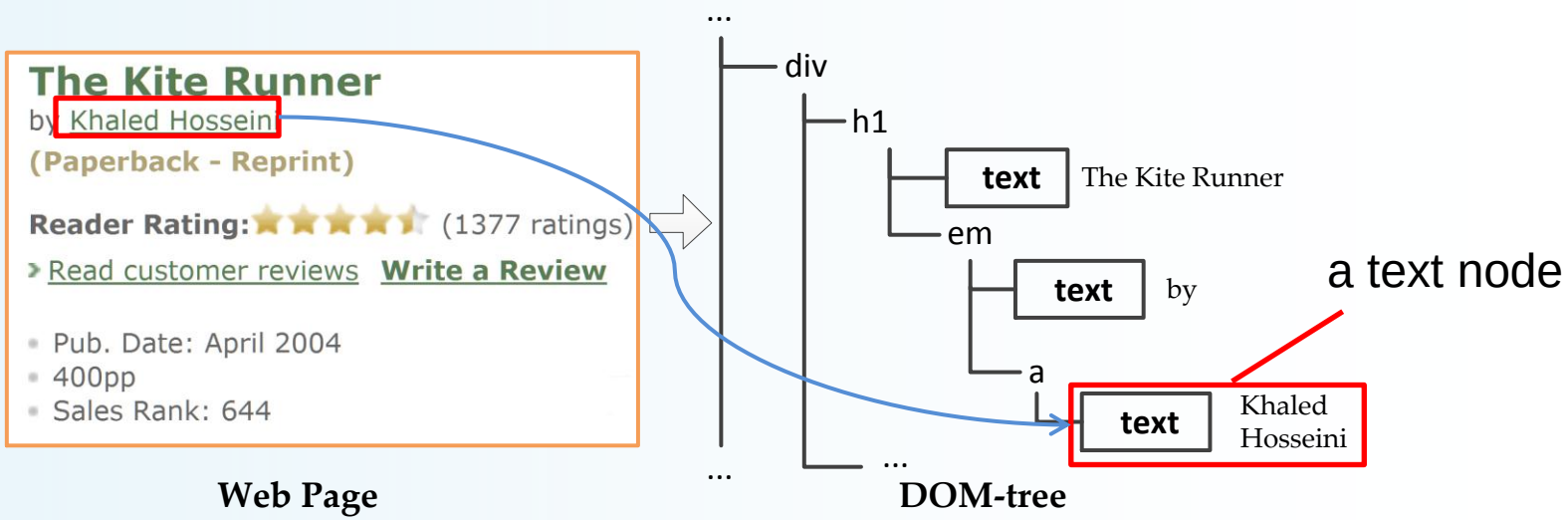


Feature Extraction

- 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)

Visual Position

position in a rendered page
= coordinates to the top left

DOM Path

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

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

Content Features

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

Extracted at page level

Unigram	Set of unique tokens
Length	Number of tokens/characters
Character Type	Proportion of letters/digits/symbols

Example:

Compared with specific features like (for ***Price***) ***Contains***('$', '.', '0-9'), more flexible for any mixture of characters

Site-level statistics

Page Redundancy	Proportion of pages containing text node with the same value

Example:

In ***Restaurant*** vertical, ***Cuisines*** are much redundant than ***Names***

General enough to characterize various attributes

Context Features

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

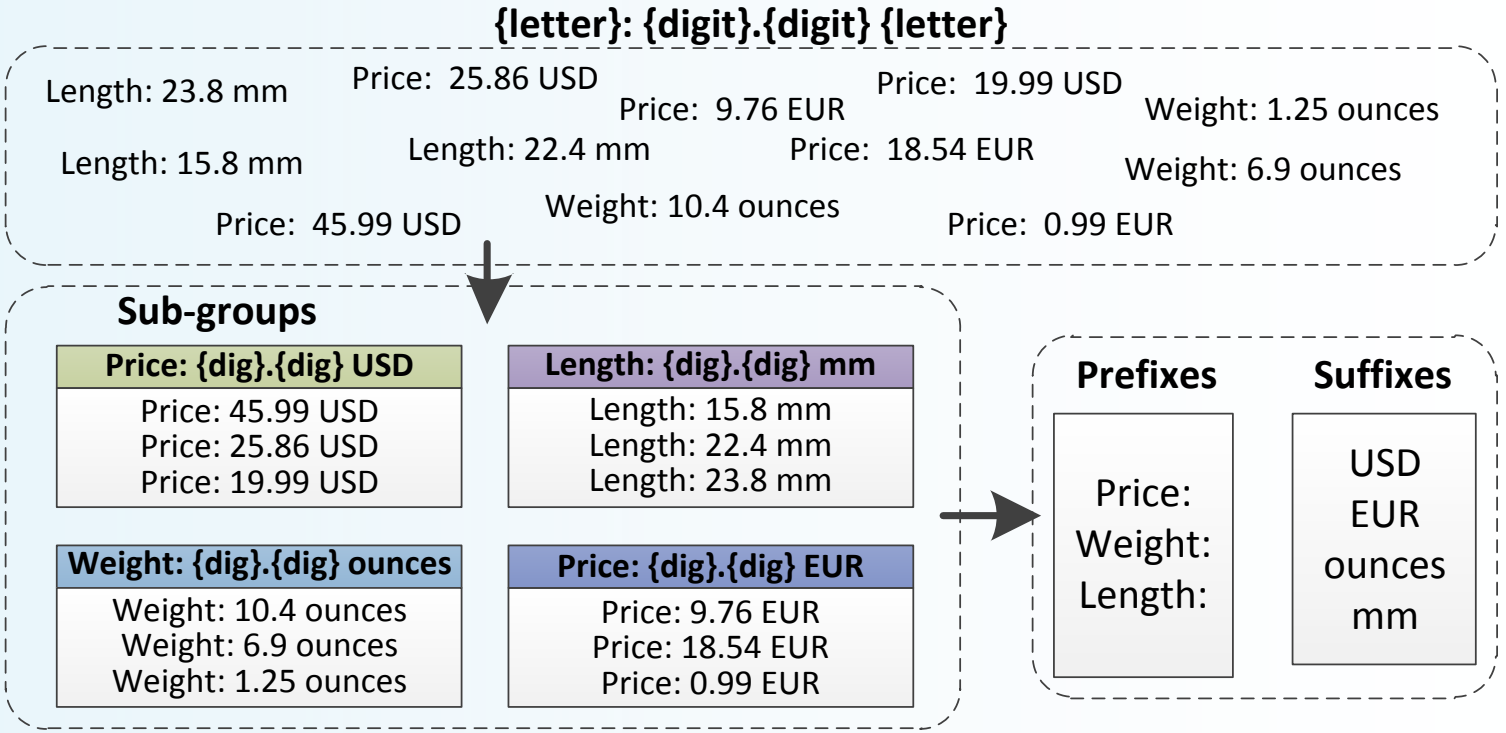
Extracted at page level	**Preceding Text**	values of visually preceding text nodes
Site-level statistics	**Prefix & Suffix**	static (across pages) sub-strings of text node values

Extracted automatically

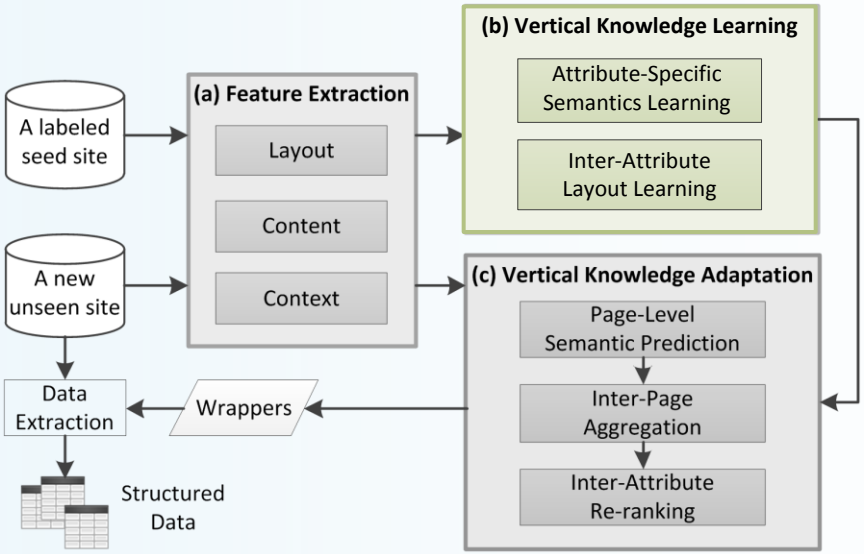
Illustrative example of prefix/suffix extraction

Context Features

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



Illustrative example of prefix/suffix extraction

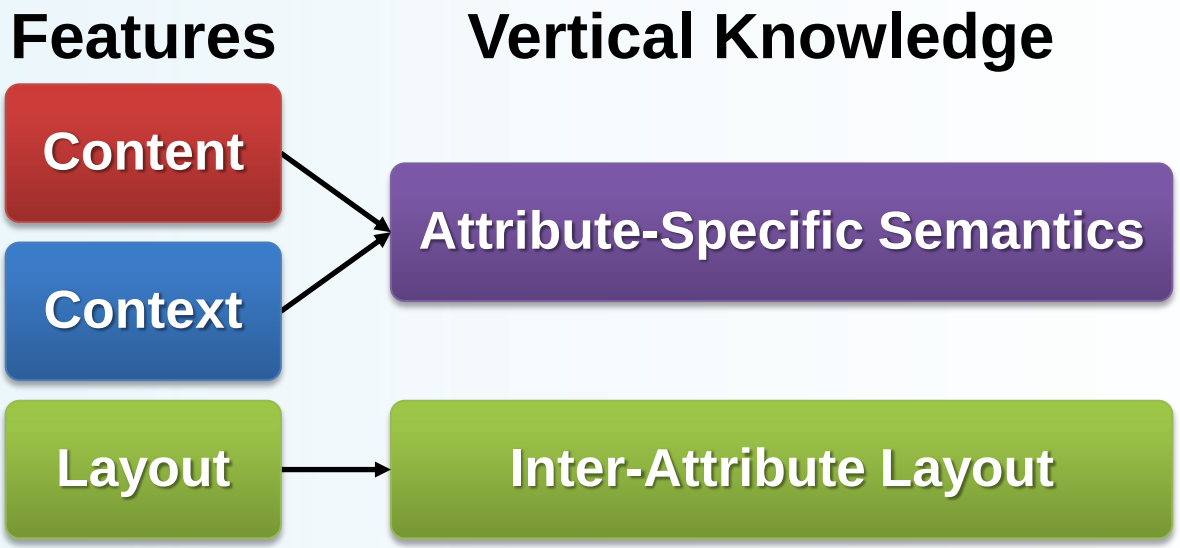


Vertical Knowledge Learning

- Attribute-Specific Semantics
- Inter-Attribute Layout

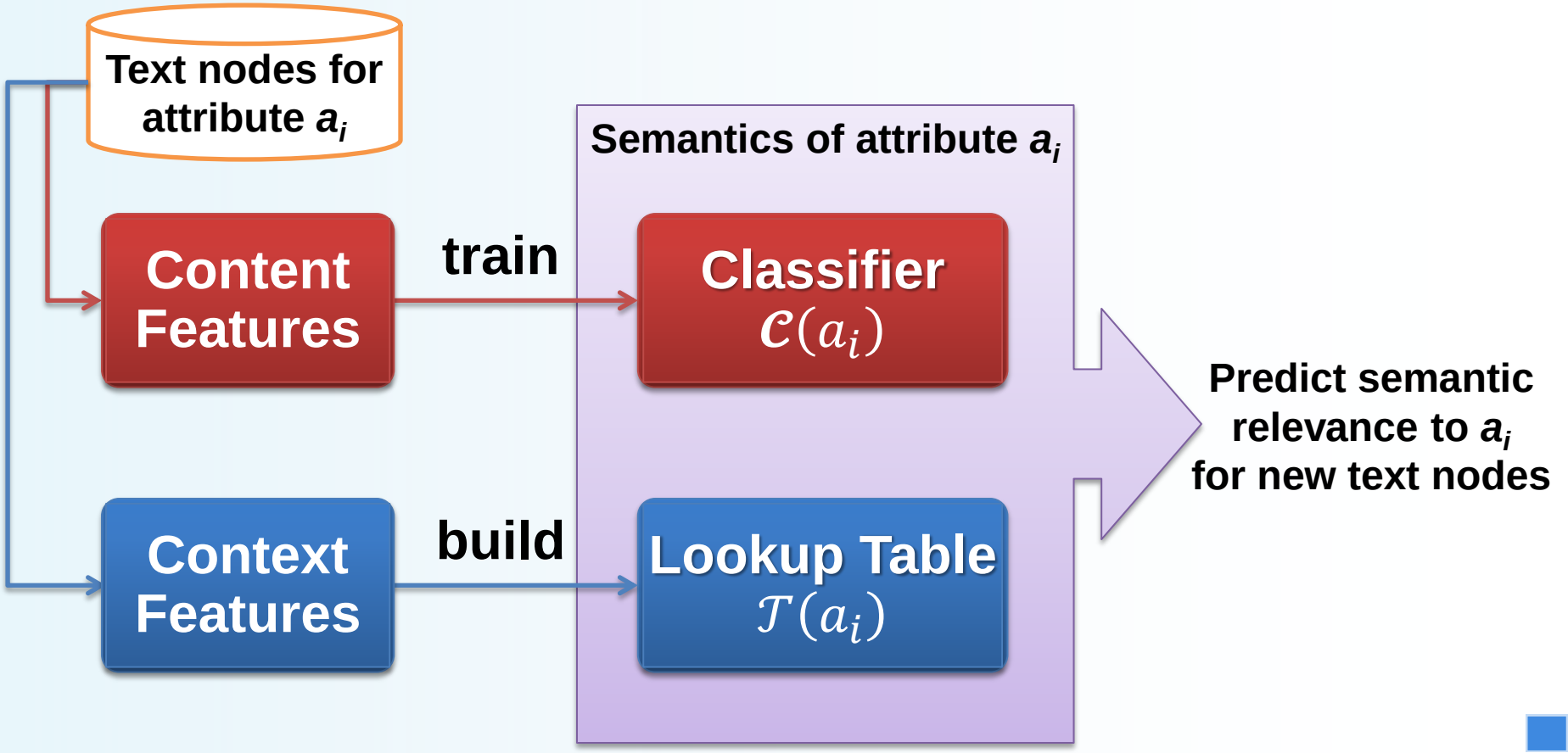
Features to Vertical Knowledge

- 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



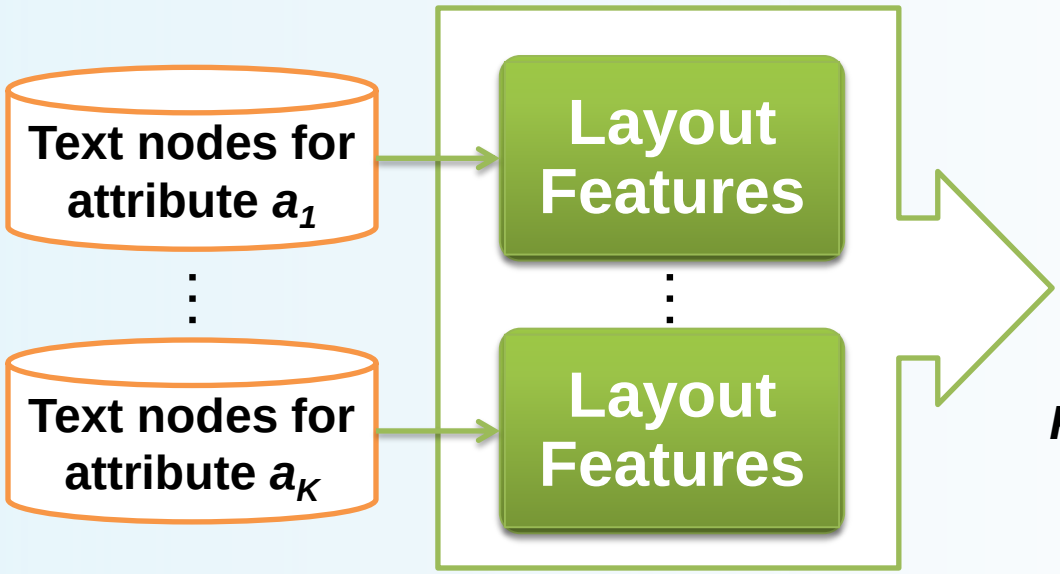
Attribute-Specific Semantics

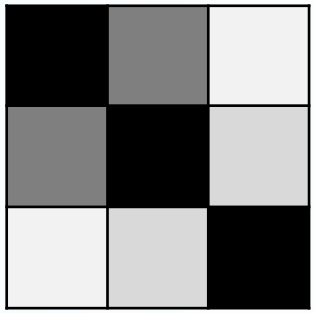
- **Content** features → ***classifiers*** (e.g., SVMs)
- **Context** features → (token-score) ***lookup tables***



Inter-Attribute Layout

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





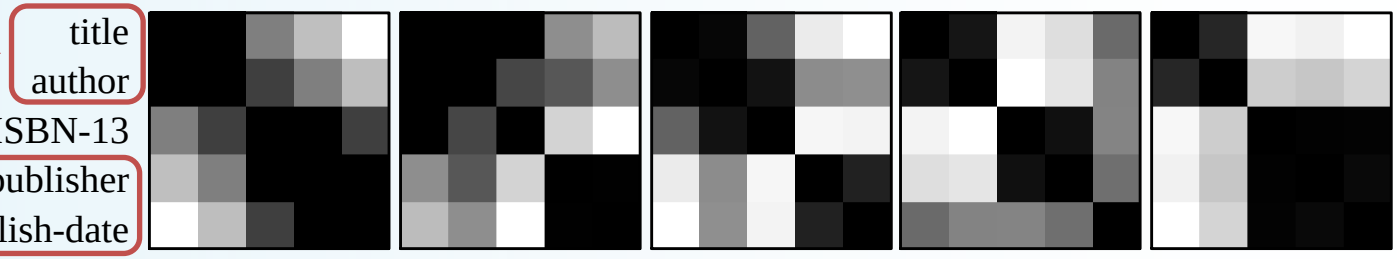
$K \times K$ layout matrix

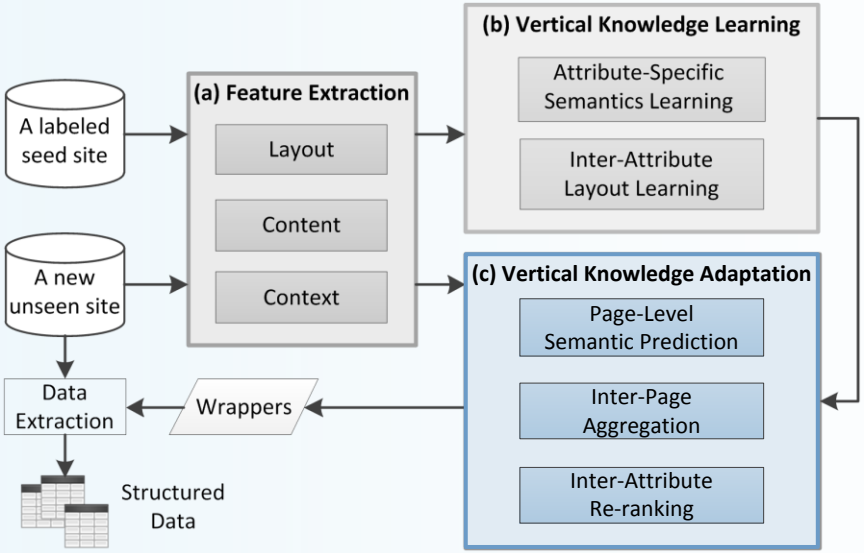
(darker → closer)

→ **Help verify combinations of attributes**

Tend to be close

Example: layout matrices from 5 websites in ***Book*** vertical



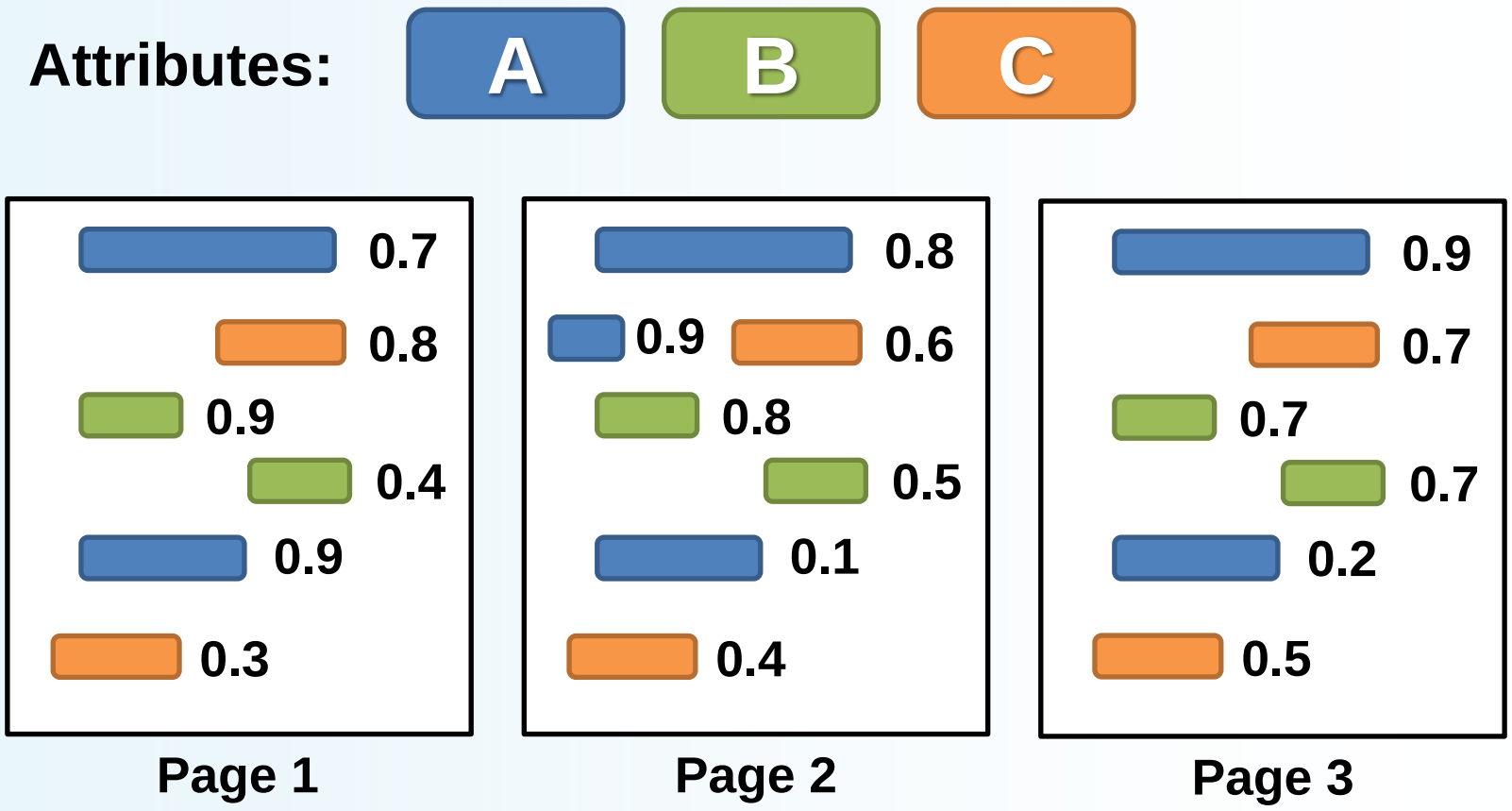


Vertical Knowledge Adaptation

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

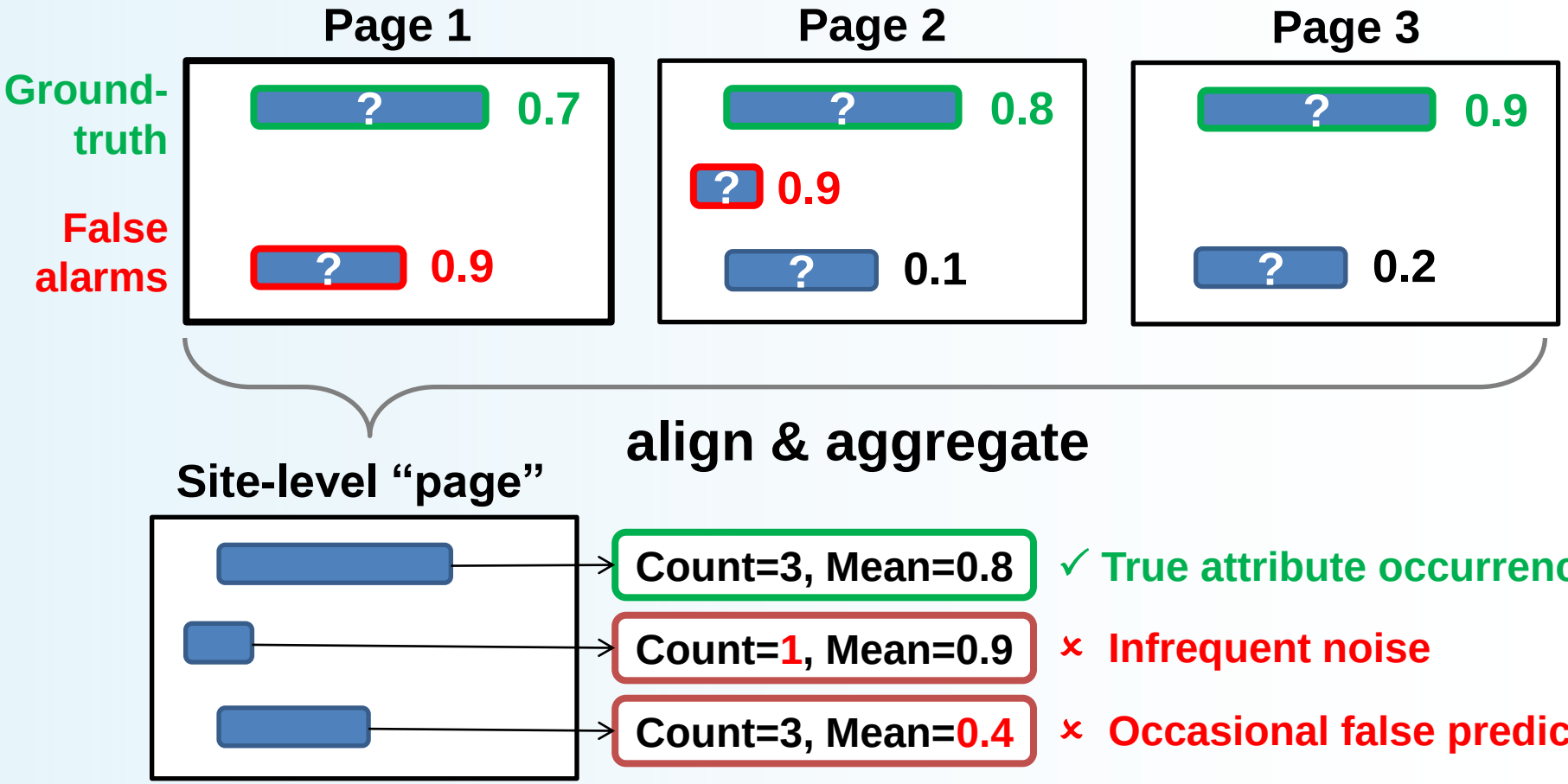
Page-Level Semantic Prediction

Attribute-specific semantics → page-level candidates



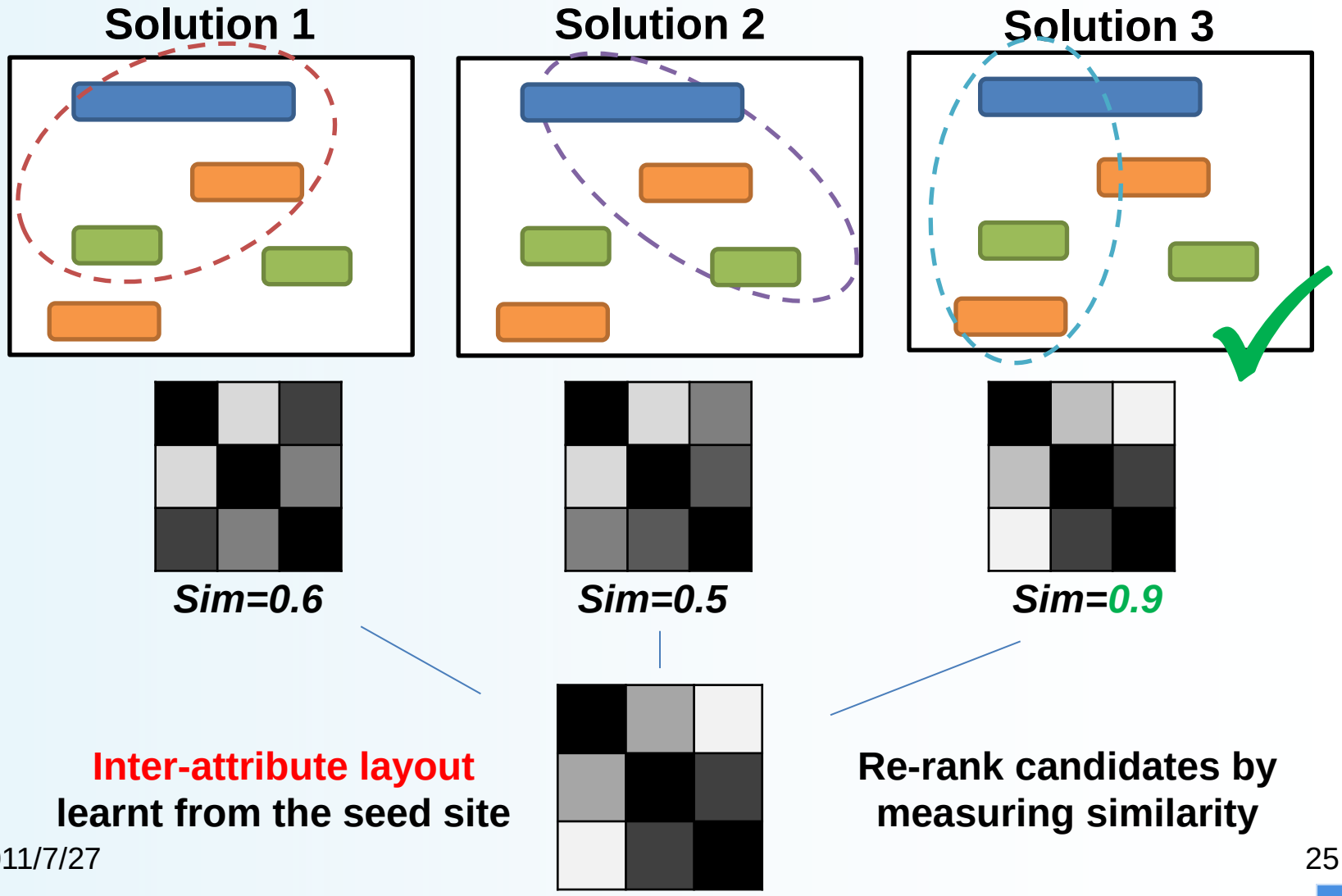
Inter-Page Aggregation

- For each attribute: multiple candidates per page →

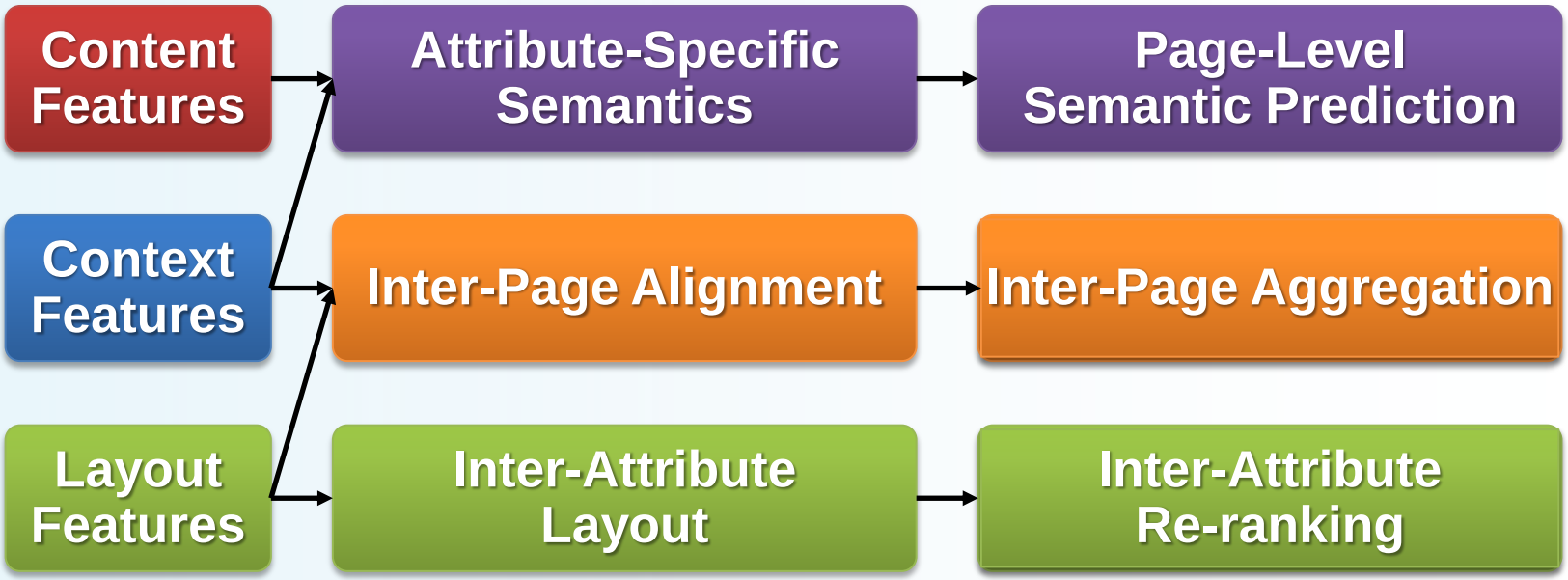


Inter-Attribute Re-ranking

- Multiple possible solutions (attribute combinations)



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

Experimental Settings

- Methods

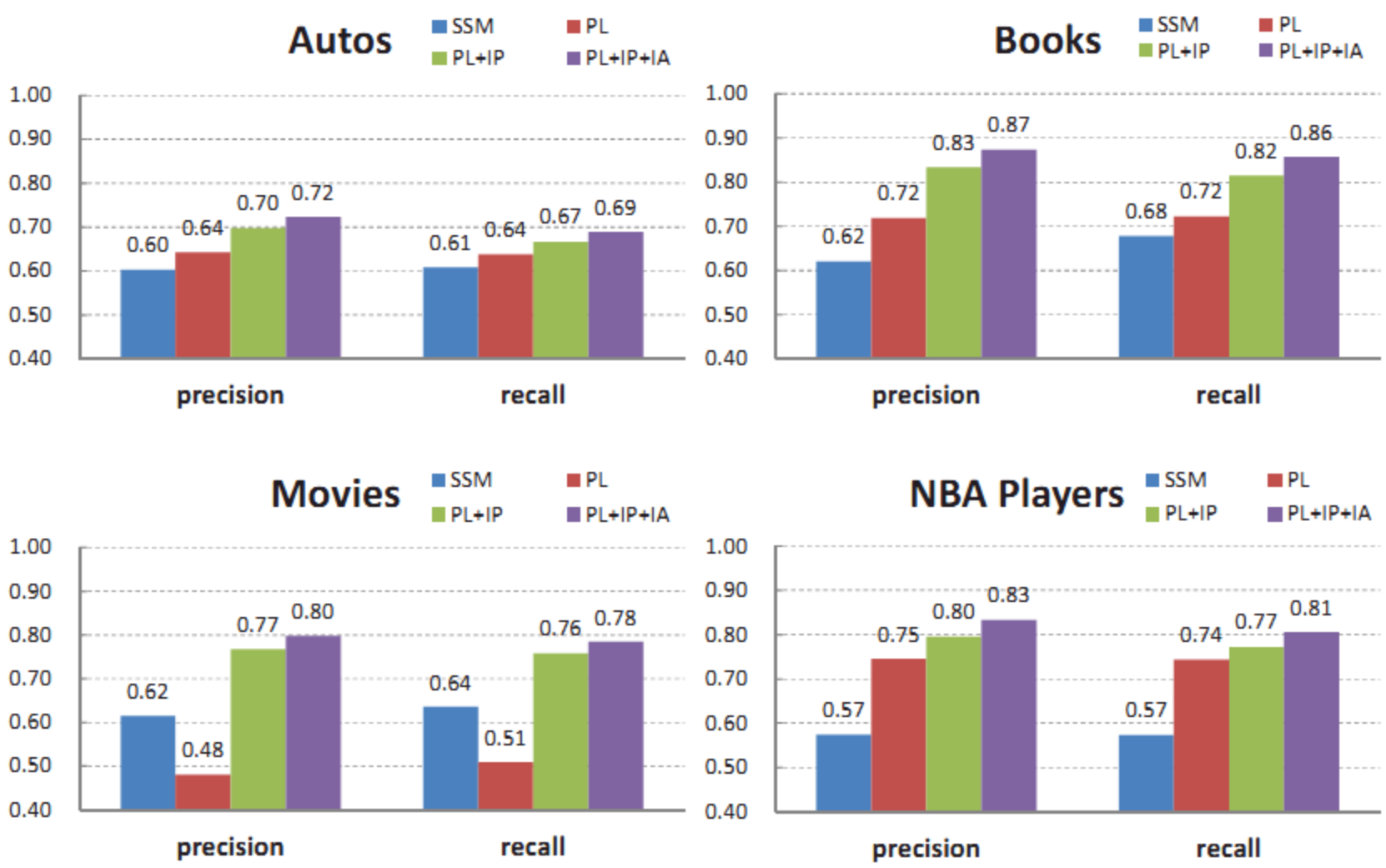
  <Baselines>

  1. **SSM** (Stacked Skews Model) *Carlson et al. ECML'08*
  2. **PL** (page-level semantic prediction)
  3. **PL + IP** (inter-page aggregation)

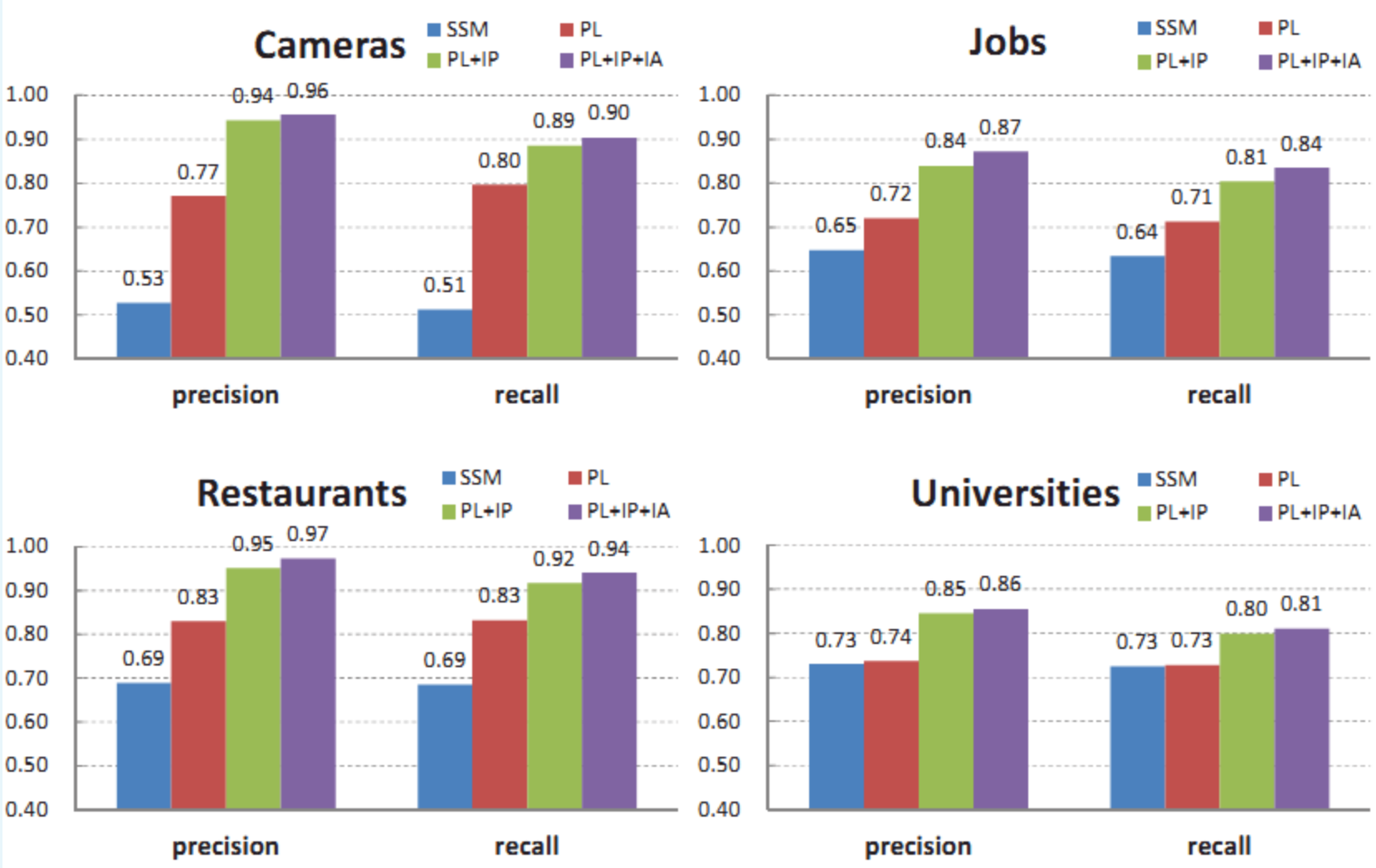
  <Full solution>

  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 ≥ 80%, Recall ≥ 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