Table Augmentation

SIGIR 2019 tutorial - Part V

Shuo Zhang and Krisztian Balog

University of Stavanger

Motivation

- Working with tables/spreadsheets is a labour-intensive task
- Table augmentation aims to provide smart assistance for users who are working with tables

Outline for this Part

Definition

Table augmentation refers to the task of extending a seed table with more data.

1. Row extension
2. Column extension
3. Data completion

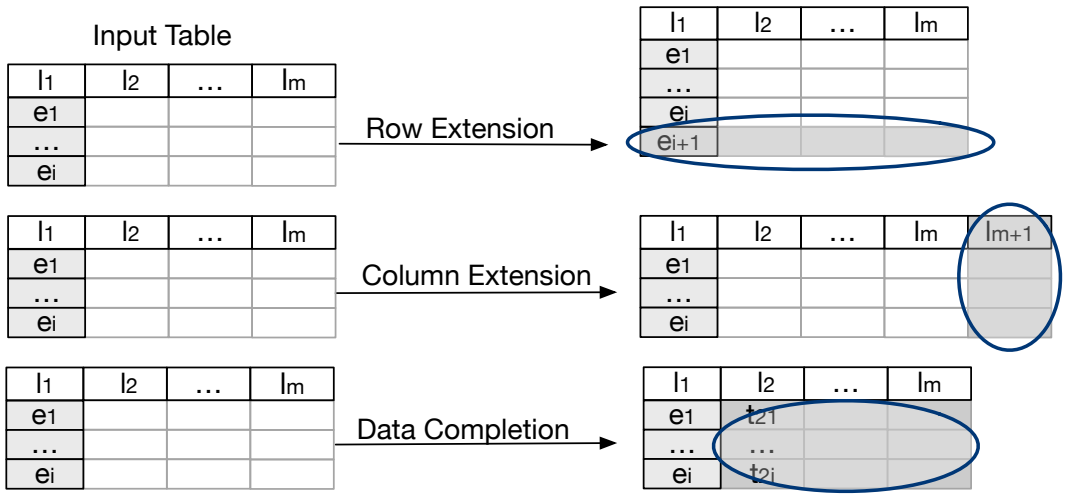
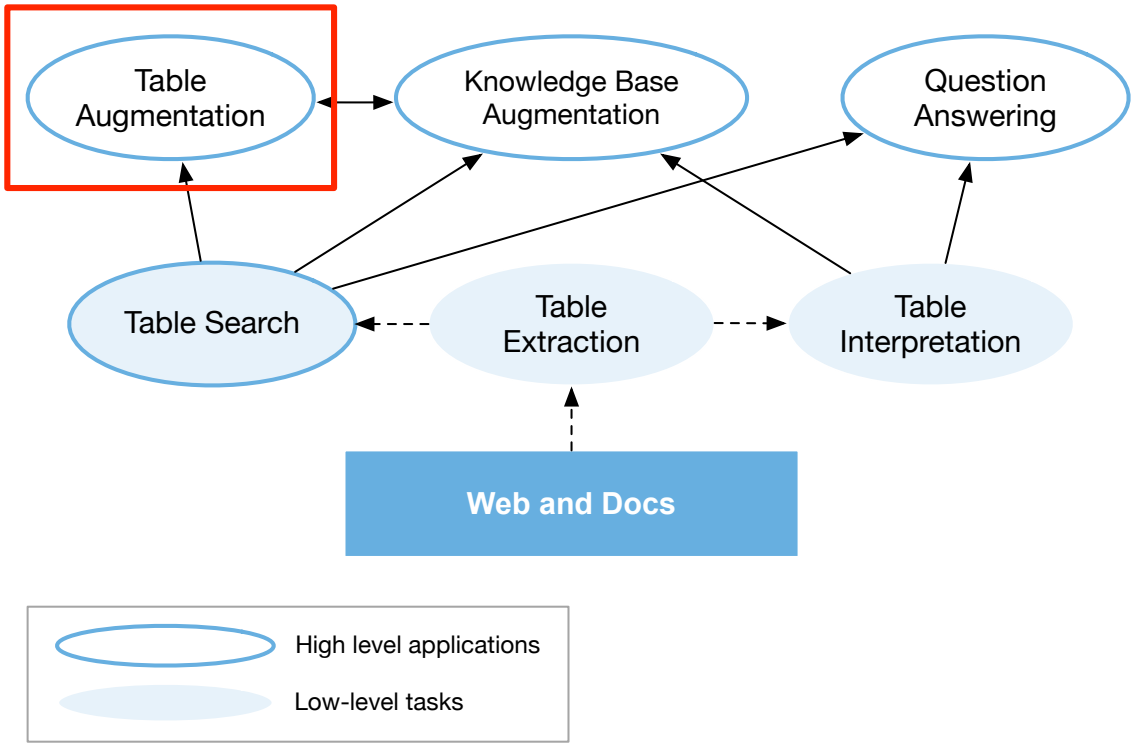


Table Augmentation VS Search by Table



- Search by table is a key block for table augmentation
- Search by table can be for many other purposes
- Table augmentation could rely on other sources as well

Data Sources

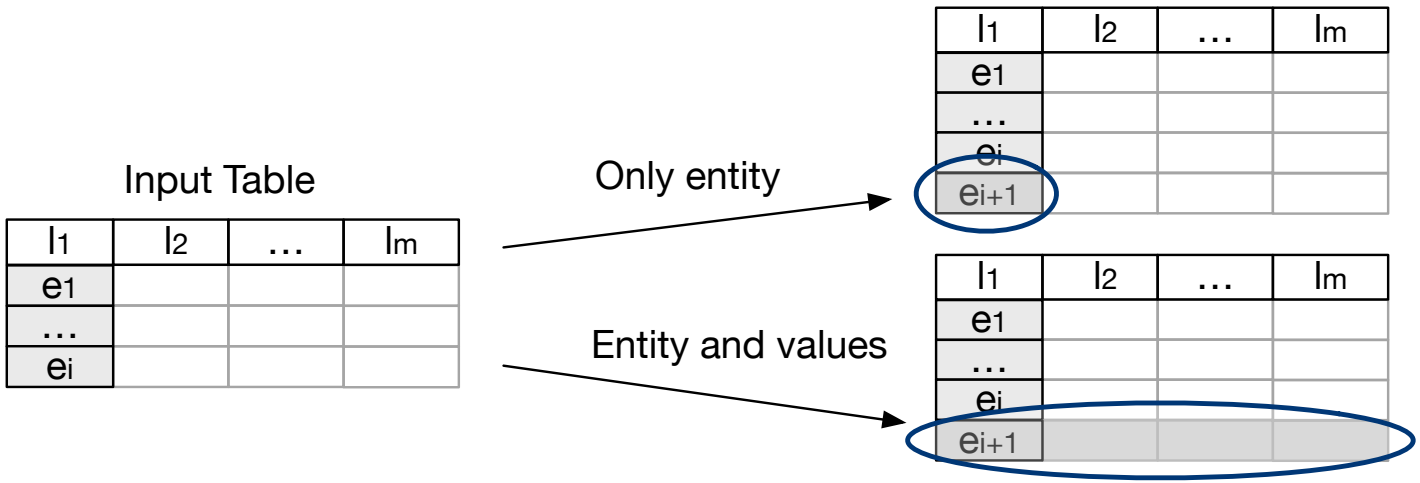
We can predict tabular values from:

1. Other tables
2. Knowledge bases
3. Unstructured data

Row Extension

Definition

Row extension aims to extend a given table with more rows or row elements.



Overview of Row Extension

Reference	Data		Tasks	
	KB	Tables	Table search	Row population
Wang et al. (2015)		✓	✓	✓*
Das Sarma et al. (2012)		✓	✓	
Yakout et al. (2012)		✓	✓	✓
Zhang and Balog (2017)	✓	✓	✓	✓

* Originally developed for concept expansion, but can be used for row population.

Finding Related Tables (Das Sarma et al., 2012)

1. They search for *entity complement* tables that are semantically related to entities in the input table (as we have already discussed in Part-4)
2. Then, the top-k related tables could be used for populating the input table (however, they stop at the table search task)

Entity Consistency and Expansion (Das Sarma et al., 2012)

1. **Knowledge base types:** Das Sarma et al. (2012) would like a related table to have the same type of entities as the seed table
2. **Table co-occurrence:** Co-occurrence is an important signal to tell if a new entity should be added to the seed table

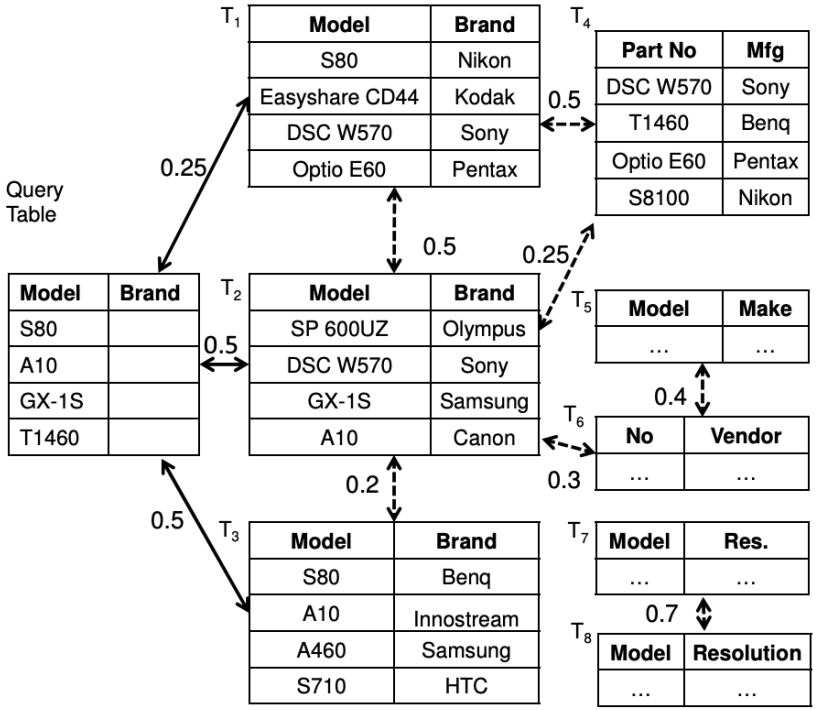
InfoGather (Yakout et al., 2012)

S80	Nikon
A10	Canon
GX-1S	
T1460	

S80	Nikon
A10	Canon
GX-1S	Samsung
T1460	Benq

- *Augmentation by example* operation in InfoGather (Yakout et al., 2012)

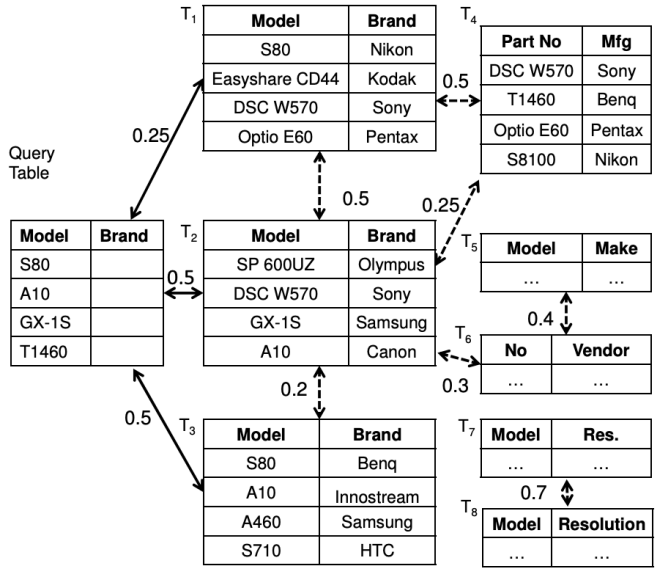
InfoGather (Yakout et al., 2012)



1. First search for related tables, then consider entities from these tables, weighted by the table relatedness scores
2. A schema matching graph among web tables (SMW graph) is built based on pairwise table similarity

Take-away Points from InfoGather (Yakout et al., 2012)

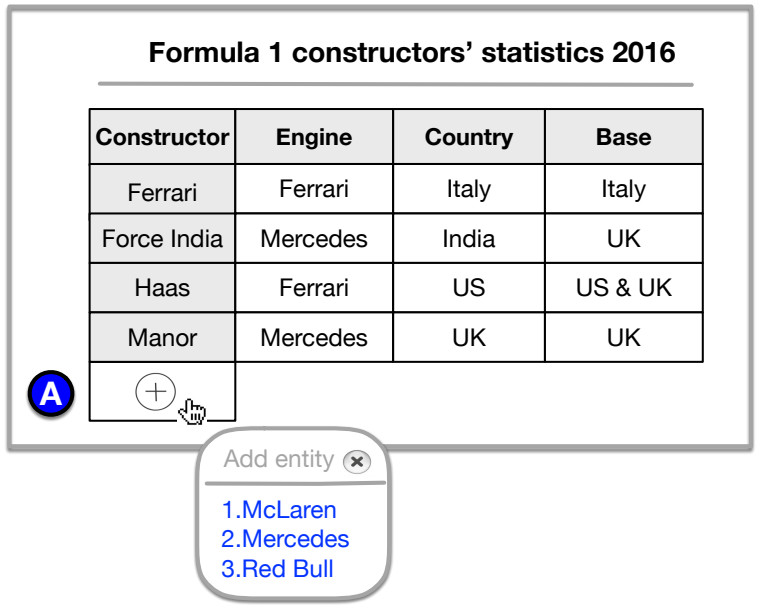
1. Despite the use of scalable techniques, this remains to be computationally very expensive, which is a main limitation of the approach
2. Relying only on tables



Row Population (Zhang and Balog, 2017)

- Zhang and Balog (2017) propose the task of row population
- Instead of relying only on related tables from a table corpus, they also consider a knowledge base (DBpedia) for identifying candidate entities

Use-case



- We assume a user, working with a table, at some intermediate stage in the process
- The user has already set the caption of the table and entered some data into the table
- The table is assumed to have a column header

Candidate Selection (Zhang and Balog, 2017)

Find candidates from both a knowledge base (DBpedia) and the table corpus:

1. **DBpedia**: focus on entities share the same **types** and **categories** as the seed entities (knowledge base types)
2. Search related **tables** (contain any seed entities, similar table caption, etc) and take their entities as candidates (co-occurrence)

Entity Ranking (Zhang and Balog, 2017)

They employ a generative probabilistic model for the subsequent ranking of candidate entities:

$$P(e|E, L, c) \propto P(e|E)P(L|e)P(c|e).$$

Components:

Entity similarity:	$P(e\|E) = \lambda_E P_{KB}(e\|E) + (1 - \lambda_E)P_{TC}(e\|E)$
Heading label likelihood:	$P(L\|e) = \sum_{l \in L} \left(\lambda_L \left(\prod_{t \in l} P_{LM}(t\|\theta_e)\right) + \frac{(1-\lambda_L)}{\|L\|} P_{EM}(l\|e)\right)$
Caption Likelihood:	$P(c\|e) = \prod_{t \in c} \left(\lambda_c P_{KB}(t\|\theta_e) + (1 - \lambda_c)P_{TC}(t\|e)\right)$

Evaluation

Data:

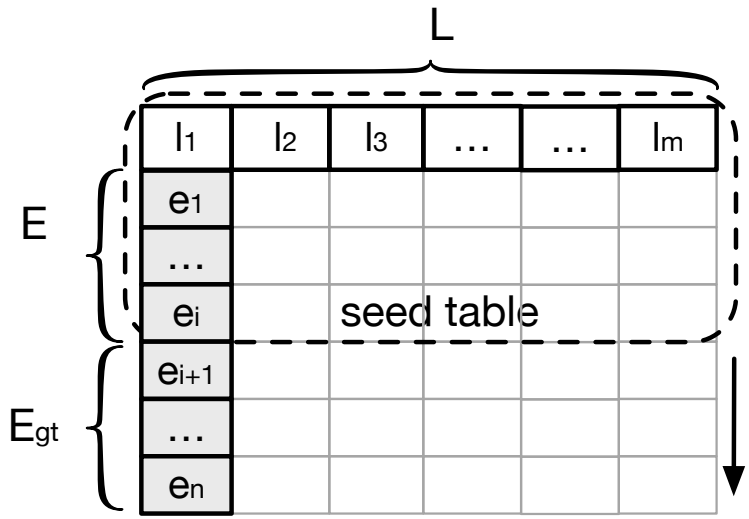
- **Table corpus**: Wikipedia tables
- **Knowledge baes**: DBpedia

Test set and validation set from the table corpus (Wikipedia tables)

- 1000 entity tables each
- Each table has at least 6 rows and 4 columns

Evaluation

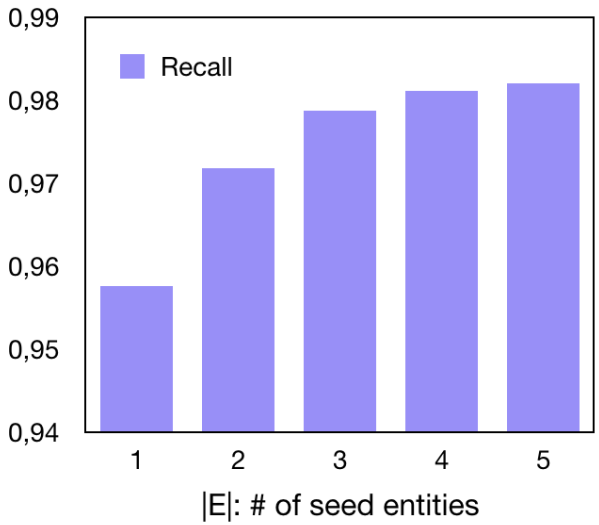
- For each table, use the first $|E|$ rows of the table as input ($|E| = 1..5$)
- The rest of the table is considered as the ground truth
- Evaluation metrics (averaged over 1000 tables):
  - Candidate selection: Recall
  - Entity ranking: MAP, MRR



Row population

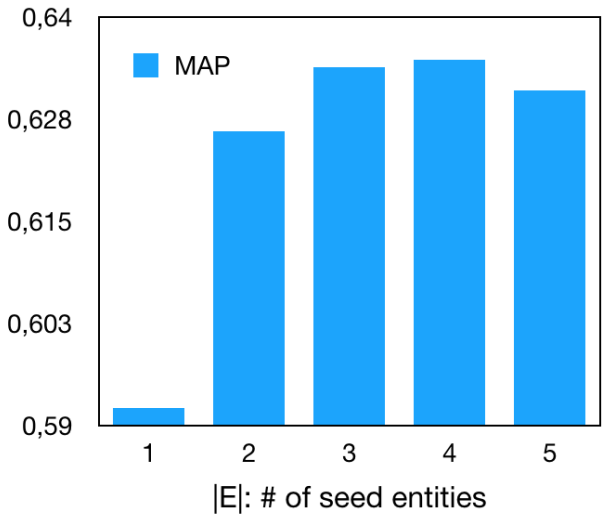
Candidate Selection Results

Method	#Seed entities ($\|E\|$)	
	1	
	Recall	#cand
(A1) Categories (k=256)	0.6470	1721
(A2) Types (k=4096)	0.0553	7703
(B) Table caption (k=256)	0.3966	987
(C) Table entities (k=256)	0.6643	312
(B) & (C) (k=256)	0.7090	1250
(A1) & (B) (k=256)	0.7642	2671
(A1) & (C) (k=256)	0.8434	1962
(A1) & (B) & (C) (k=256)	0.8662	2880
(A1) & (B) & (C) (k=4096)	**0.9576**	28733



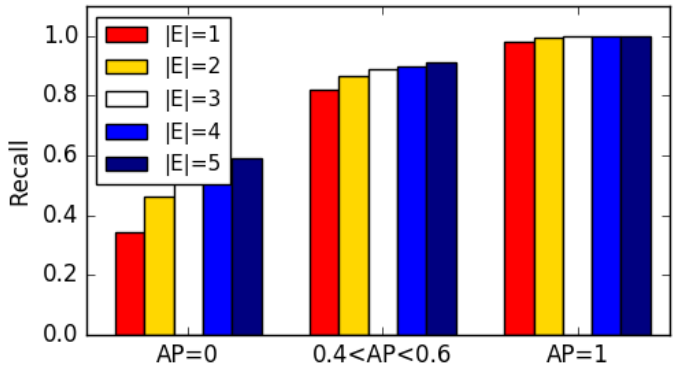
Entity Ranking Results

Method	#Seed entities ($\|E\|$)	
	1	
	Recall	#cand
(A1) $P(e\|E)$ Relations ($\lambda = 0.5$)	0.4962	0.6857
(A2) $P(e\|E)$ WLM ($\lambda = 0.5$)	0.4674	0.6246
(A3) $P(e\|E)$ Jaccard ($\lambda = 0.5$)	0.4905	0.6731
(B) $P(L\|e)$	0.2857	0.3558
(C) $P(c\|e)$	0.2348	0.2656
(A3) & (B)	0.5726	0.7593
(A3) & (C)	0.5743	0.7467
(B) & (C)	0.3677	0.4521
(A3) & (B) & (C)	**0.5922**	**0.7729**



Take-away Points for Row Population

1. Both tables and KBs are useful for this task
2. Candidate selection:
   - Category > Type
   - Entity > Caption > Headings
   - All complement each other
3. Entity ranking
   - Entity > Headings > Caption
   - All complement each other
   - Highly relevant to candidate selection
4. Code and data: https://github.com/iai-group/sigir2017-table/



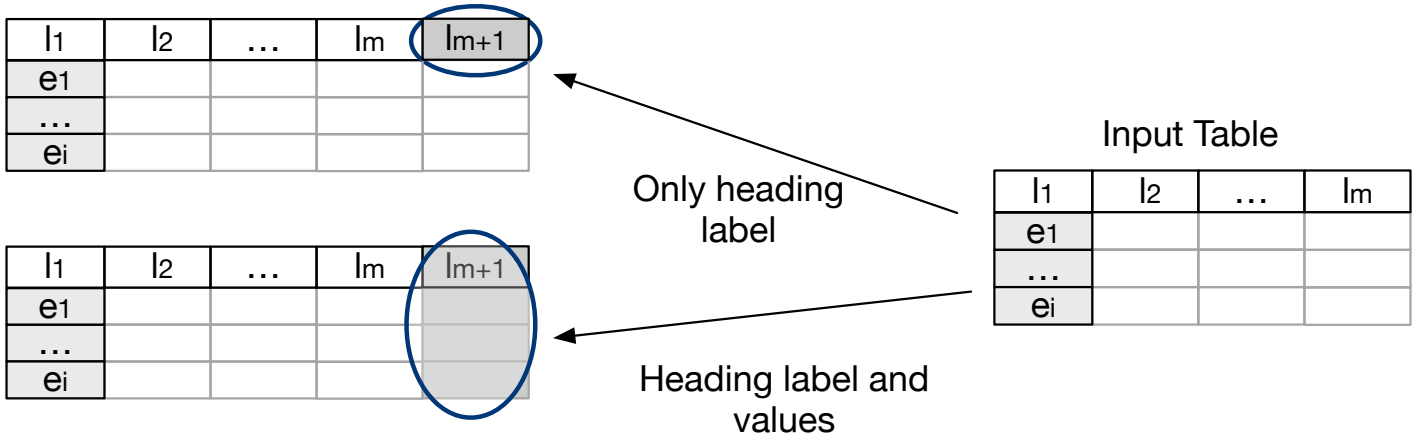
Outline for this Part

1. ~~Row extension~~
2. **Column extension**
3. Data completion

Column Extension

Definition

Column extension aims to extend a table with additional columns.



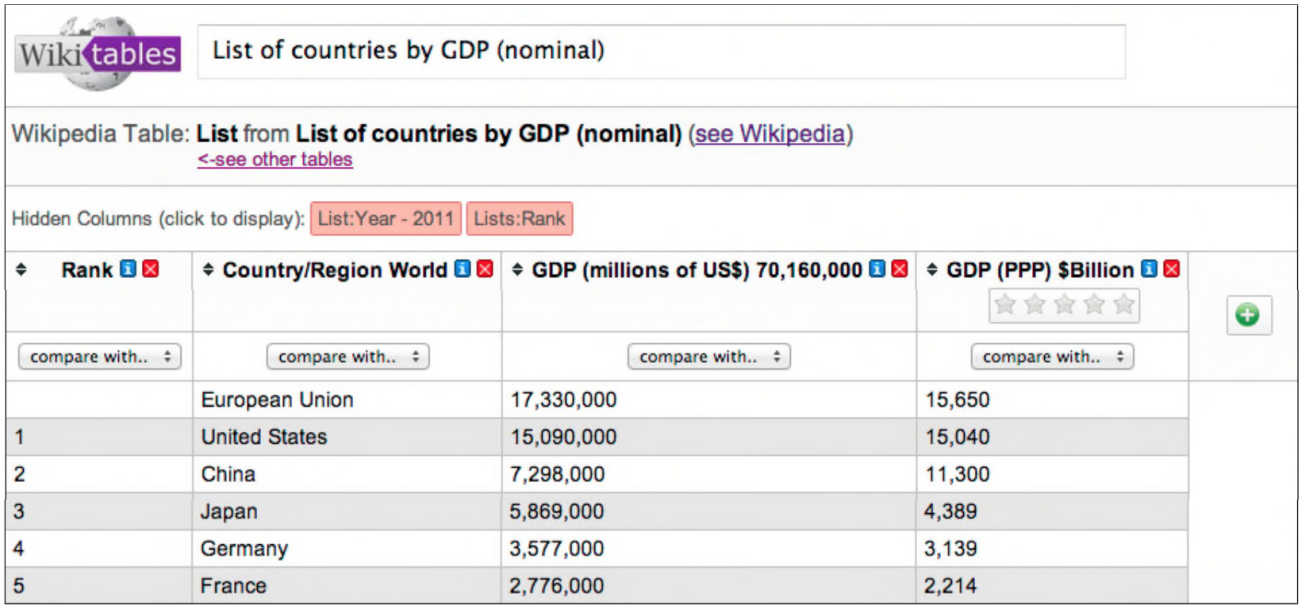
Overview of Column Extension

Reference	Tasks	
	Table search	**Column population**
Relation join (Lehmberg et al., 2015)	✓	✓
Schema complement (Das Sarma et al., 2012)	✓	
InfoGather (Yakout et al., 2012)	✓	✓
Column population (Zhang and Balog, 2017)	✓	✓

OCTOPUS (Cafarella et al., 2009)

1. OCTOPUS combines search, extraction, data cleaning and integration
2. It enables users to add more columns to a table by performing a join
3. Any new columns do not necessarily come from the same single source table
   - Keyword table search
   - Schema matching (publications vs. papers)
   - Reference reconciliation problem (Alon Halevy vs. Alon Levy)

WikiTables (Bhagavatula et al., 2013)



Wikitables — List of countries by GDP (nominal)

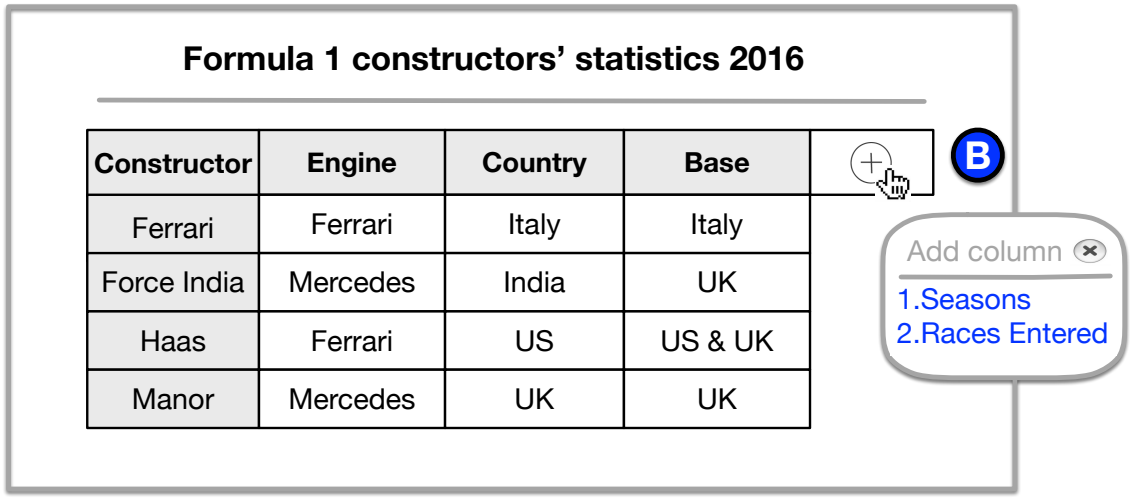
Wikipedia Table: **List** from **List of countries by GDP (nominal)** (see Wikipedia)
<-see other tables

Hidden Columns (click to display): List:Year - 2011 | Lists:Rank

Rank	Country/Region World	GDP (millions of US$) 70,160,000	GDP (PPP) $Billion
compare with..	compare with..	compare with..	compare with..
	European Union	17,330,000	15,650
1	United States	15,090,000	15,040
2	China	7,298,000	11,300
3	Japan	5,869,000	4,389
4	Germany	3,577,000	3,139
5	France	2,776,000	2,214

- http://downey-n1.cs.northwestern.edu/wikiTables/
- Bhagavatula et al. (2013) utilize the Milne-Witten Semantic Relatedness measure for estimating the relatedness between the input tables and candidate columns

Zhang and Balog (2017)



Zhang and Balog (2017) try to find the headings that can be added as columns to an input table.

Column Population (Zhang and Balog, 2017)

A two-step pipeline:

1. Candidate Selection:
   - Search related tables (contain any seed column labels, table entities, similar table caption)
   - Take their column labels as candidates
2. Column label ranking

Column Label Ranking (Zhang and Balog, 2017)

They employ a generative probabilistic model for the subsequent ranking of candidate labels:

$$P(l|E,c,L) = \sum_{T} P(l|T)P(T|E,c,L).$$

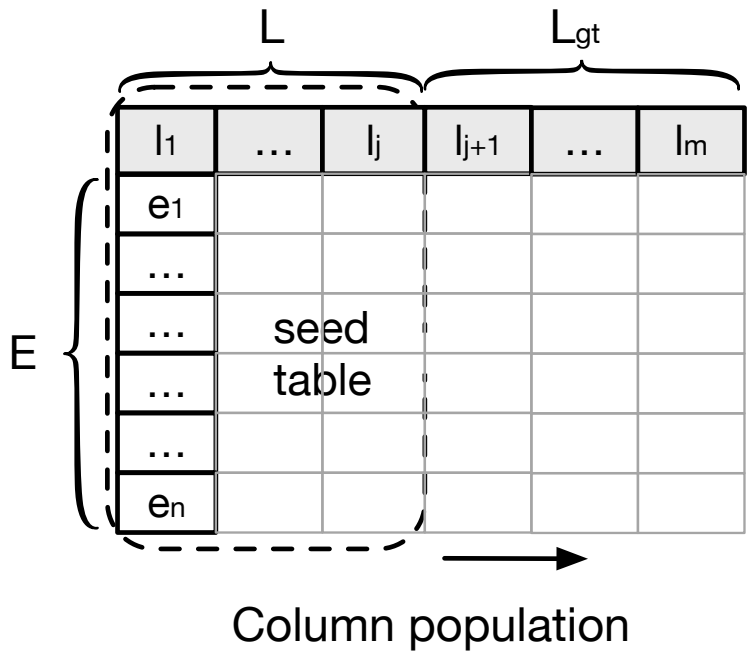
It is based on the similarity to:

Tabel Likelihood: $$P(l|T) = \begin{cases} 1, & \text{if } l \text{ appears in } T \\ 0, & \text{otherwise .} \end{cases}$$

Table Relevance Estimation: $$P(T|E,c,L) = \frac{P(T|E)P(T|c)P(T|L)}{P(T)^2}$$

Evaluation

- For each table, use the first $|L|$ columns of the table as input ($|L| = 1..3$)
- The rest of the table is considered as the ground truth
- Evaluation metrics (averaged over 1000 tables):
  - Candidate selection: Recall
  - Entity ranking: MAP, MRR



Candidate Selection Results (Zhang and Balog, 2017)

Method	#Seed column labels ($\|L\|$)					
	1		2		3	
	Recall	#cand	Recall	#cand	Recall	#cand
(A) Table caption (k=256)	0.7177	232	0.7115	232	0.7135	231
(B) Column labels (k=256)	0.2145	115	0.5247	235	0.7014	357
(C) Table entities (k=64)	0.7617	157	0.7544	156	0.7505	155
(A) (k=256) & (B) (k=256) & (C) (k=64)	0.8799	467	0.8961	572	0.9040	682
(A) (k=4096) & (B) (k=4096) & (C) (k=4096)	**0.9211**	2614	**0.9292**	3309	**0.9351**	3978

Column Label Ranking Results (Zhang and Balog, 2017)

Method	#Seed column labels ($\|L\|$)					
	1		2		3	
	MAP	MRR	MAP	MRR	MAP	MRR
(A) Table caption	0.2584	0.3496	0.2404	0.2927	0.2161	0.2356
(B) Column labels	0.2463	0.3676	0.3145	0.4276	0.3528	0.4246
(C) Table entities	0.3878	0.4544	03714	0.4187	0.3475	0.3732
(A) & (B)	0.4824	0.5896	0.4929	0.5837	0.4826	0.5351
(A) & (C)	0.5032	0.5941	0.4909	0.5601	0.4724	0.5132
(B) & (C)	0.5060	0.5954	0.5410	0.6178	0.5323	0.5802
(A) & (B) & (C)	**0.5863**	**0.6854**	**0.5847**	**0.6690**	**0.5696**	**0.6201**

Take-away Points for Column Population

1. Entity > Caption > Heading
2. All table elements complement each other
3. Code and data:
   https://github.com/iai-group/sigir2017-table/

Table2vec (Deng et al., 2019)

Region	Release Date	Label	Release Format
United Kingdom	22 September 2008	Super Records	DVD
Ireland		cords	DVD
Japan		rax	DVD
Argentina	18 May 2009	EMI Music	Digital Download
Singapore	12 June 2009	Warner Music	DVD
Spain	1 December 2009	EMI Music Spain	Digital Download

pgTitle: Radio:Active
secondTitle: Release history
caption: Release history

(a) *Table2VecW*

Region	Release_Date	Label	Release_Format
United Kingdom	22 September 2008	Super Records	DVD
Ireland	22 September 2008	Super Records	DVD
Japan	11 February 2009	Avex Trax	DVD
Argentina	18 May 2009	EMI Music	Digital Download
Singapore	12 June 2009	Warner Music	DVD
Spain	1 December 2009	EMI Music Spain	Digital Download

(b) *Table2VecH*

Region	Release Date	Label	Release Format
United_Kingdom	22 September 2008	Super Records	DVD
Ireland	22 September 2008	Super Records	DVD
Japan	11 February 2009	Avex_Trax	DVD
Argentina	18 May 2009	EMI	Music_Download
Singapore	12 June 2009	Warner_Music_Group	DVD
Spain	1 December 2009	EMI Music Spain	Music_Download

(c) *Table2VecE*

Region	Release Date	Label	Release Format
United_Kingdom	22 September 2008	Super Records	DVD
Ireland	22 September 2008	Super Records	DVD
Japan	11 February 2009	Avex_Trax	DVD
Argentina	18 May 2009	EMI	Music_Download
Singapore	12 June 2009	Warner_Music_Group	DVD
Spain	1 December 2009	EMI Music Spain	Music_Download

(d) *Table2VecE

Welcome to our poster 1-08 at **Session 2A**!

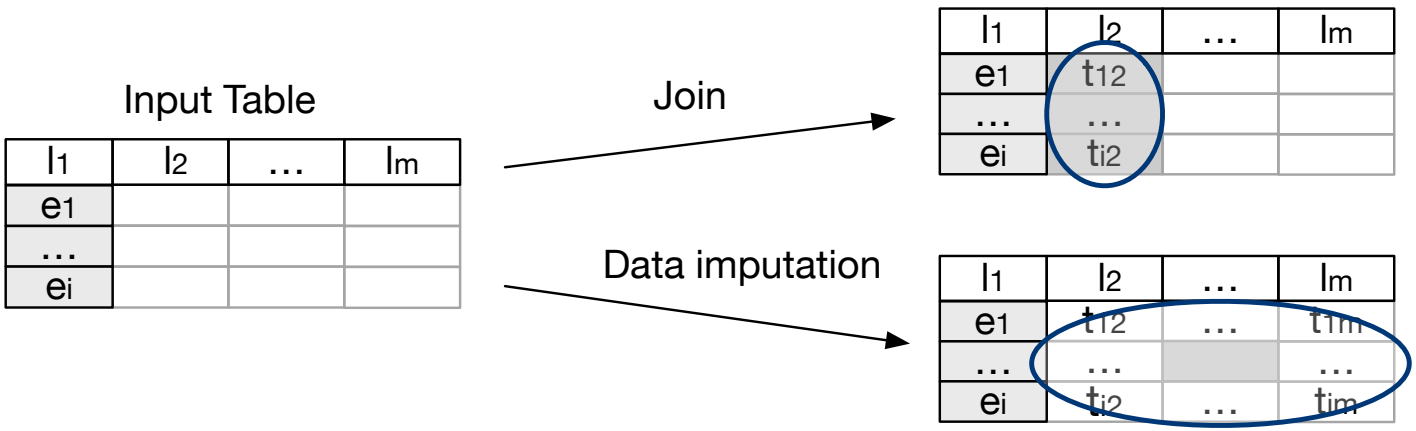
Outline for this Part

1. ~~Row extension~~
2. ~~Column extension~~
3. **Data completion**

Data Completion

Definition

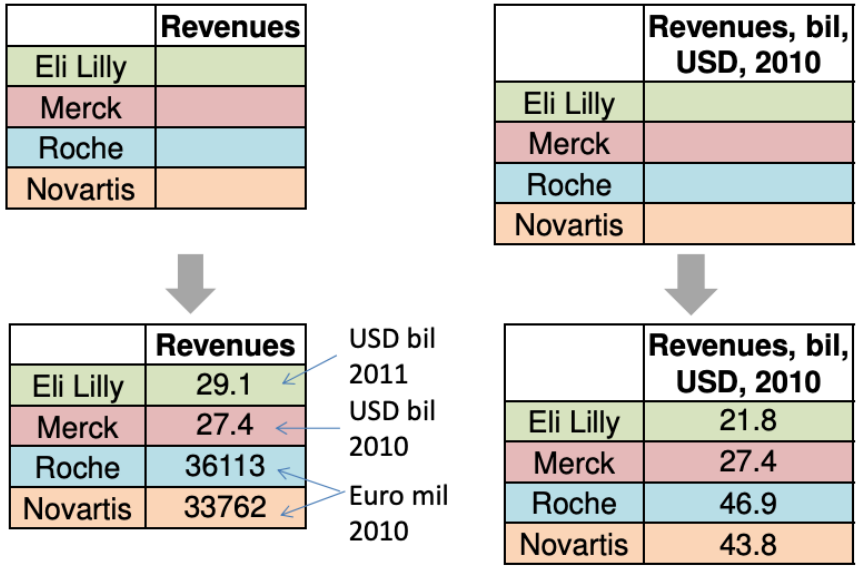
Data completion for tables refers to the task of filling in the empty table cells.



Overview of Data Completion Methods

Reference	Data		Output	
	Tables	Web	$T_{[:,j]}$	$T_{[i,j]}$
Yakout et al. (2012)	✓		✓	
Zhang and Chakrabarti (2013)	✓		✓	
Cafarella et al. (2009)	✓		✓	
Ahmadov et al. (2015)	✓	✓	✓	✓

InfoGather+ (Zhang and Chakrabarti, 2013)



- InfoGather (Yakout et al., 2012) focuses on finding values that are entities
- InfoGather+ (Zhang and Chakrabarti, 2013), focuses on numerical and time-varying attributes

InfoGather+ (Zhang and Chakrabarti, 2013)

- They use undirected graphical models and build a semantic graph that labels columns with units, scales, and timestamps, and computes semantic matches between columns
- The experiments are conducted on three types of tables: company (revenue and profit), country (area and tax rate), and city (population)
- They find that the conversion rules (manually designed unit conversion mapping) achieve higher coverage than string-based schema matching methods

Summary of this Part

1. Row extension could rely on multiple sources
2. Column extension mainly deals with tables
3. End-to-end applications (apply to spreadsheets?)
4. How to use unstructured data for extracting evidence?

Bibliography I

Ahmad Ahmadov, Maik Thiele, Julian Eberius, Wolfgang Lehner, and Robert Wrembel. Towards a hybrid imputation approach using web tables. In *Proceedings of the IEEE 2nd International Symposium on Big Data Computing*, BDC '15, pages 21–30. IEEE, 2015. ISBN 978-0-7695-5696-3.

Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Methods for exploring and mining tables on wikipedia. In *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics*, IDEA '13, pages 18–26, New York, NY, USA, 2013. ACM.

Michael J. Cafarella, Alon Halevy, and Nodira Khoussainova. Data integration for the relational web. *Proc. VLDB Endow.*, 2(1):1090–1101, August 2009. ISSN 2150-8097.

Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. Finding related tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 817–828, New York, NY, USA, 2012. ACM.

Li Deng, Shuo Zhang, and Krisztian Balog. Table2vec: Neural word and entity embeddings for table population and retrieval. In *Proc. of SIGIR '19*, 2019.

Oliver Lehmberg, Dominique Ritze, Petar Ristoski, Robert Meusel, Heiko Paulheim, and Christian Bizer. The mannheim search join engine. *Web Semant.*, 35(P3):159–166, December 2015. ISSN 1570-8268.

Bibliography II

Chi Wang, Kaushik Chakrabarti, Yeye He, Kris Ganjam, Zhimin Chen, and Philip A. Bernstein. Concept expansion using web tables. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1198–1208, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. Infogather: Entity augmentation and attribute discovery by holistic matching with web tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 97–108, New York, NY, USA, 2012. ACM.

Meihui Zhang and Kaushik Chakrabarti. Infogather+: Semantic matching and annotation of numeric and time-varying attributes in web tables. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 145–156, New York, NY, USA, 2013. ACM.

Shuo Zhang and Krisztian Balog. Entitables: Smart assistance for entity-focused tables. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 255–264, New York, NY, USA, 2017. ACM.