Hasso Plattner Institut
Digital Engineering · Universität Potsdam

```
<!DOCTYPE html>
<html>
 <head>
  .
 </head>
 <body>
  <table>
   <tr>
    <td>..</td>
    <td>..</td>
   </tr>
  </table>
 </body>
</html>
```

Processing Web Tables

Prof. Dr. Felix Naumann , Hazar Harmouch and Leon Bornemann

SoS-2019

Agenda

1. Chair Introduction
2. Organisational Information + Grading
3. The Research Area of Webtables
   a. History
   b. Typical Problems
   c. Challenges
   d. Use Cases
   e. What do we need Webtables for?
   f. What Datasets/Toolkits exist?
4. Your Research Topics

Agenda

1. Chair Introduction
2. Organisational Information + Grading
3. The Research Area of Webtables
   a. History
   b. Typical Problems
   c. Challenges
   d. Use Cases
   e. What do we need Webtables for?
   f. What Datasets/Toolkits exist?
4. Your Research Topics

Information Systems Team


Dr. Thorsten **Papenbrock**


Diana **Stephan**


Prof. Felix **Naumann**


Dr. Ralf **Krestel**


Leon **Bornemann**


Hazar **Harmouch**


Konstantina **Lazaridou**


Tim **Repke**


Julian **Risch**


Michael **Loster**


John **Koumarelas**


Tobias **Bleifuß**


Nitisha **Jain**


Lan **Jiang**


Gerardo **Vitagliano**

project **DuDe**

Duplicate Detection

Data Change

Data Fusion

project **Stratosphere**

Entity Search

Data Profiling

Information Integration

project **DataKnoller**

Web Science

project **DataChEx**

Data Scrubbing

Data as a Service

Information Quality

Data Cleansing

Text Mining

Web Data

Linked Open Data

RDF Data Mining

Dependency Detection

ETL Management

Service-Oriented Systems

project **Janus**

Entity Recognition

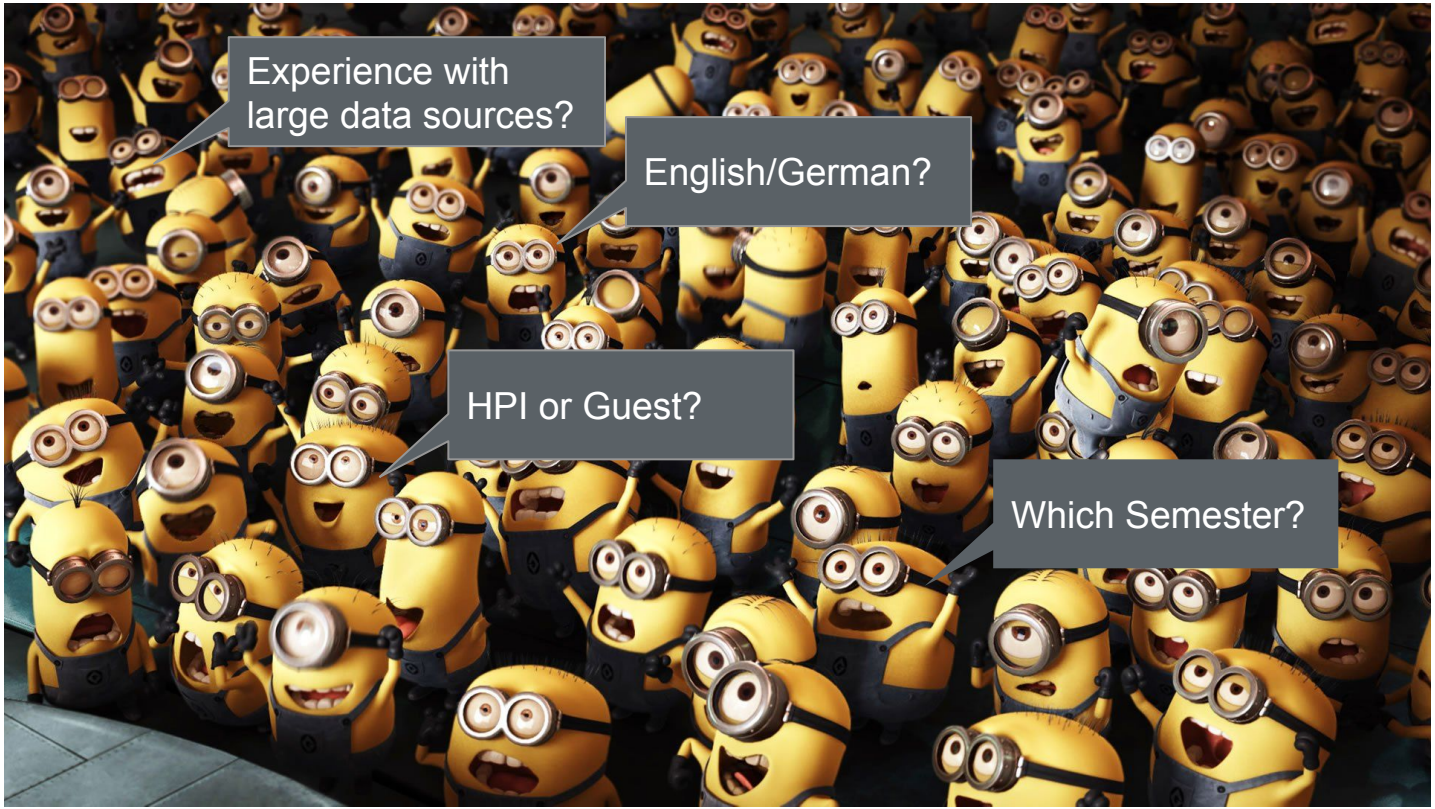
Opinion Mining

Data Preparation

project **Metanome**

Change Exploration

What about you?



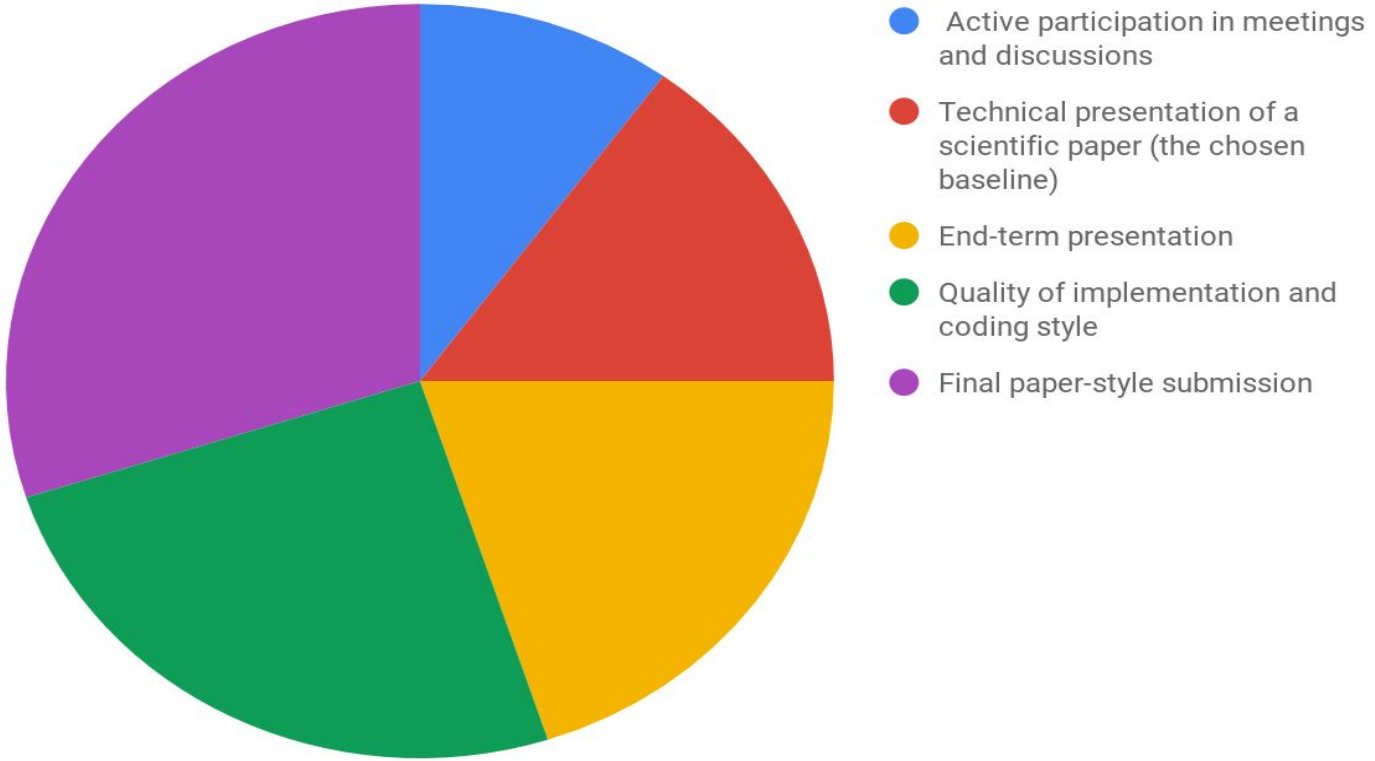
Agenda

1. Chair Introduction
2. Organisational Information + Grading
3. The Research Area of Webtables
   a. History
   b. Typical Problems
   c. Challenges
   d. Use Cases
   e. What do we need Webtables for?
   f. What Datasets/Toolkits exist?
4. Your Research Topics

Organization

<table>
<tr><td colspan="2">Group allocation</td><td>6 participants,
3 teams of 2 students</td></tr>
<tr><td>Present a summary of related work</td><td rowspan="2">Implement your baseline</td><td rowspan="2">Run experiments and describe results</td></tr>
<tr><td>Technical presentation about the baseline solution</td></tr>
<tr><td colspan="2">Mid-term presentation</td><td></td></tr>
<tr><td>Design your solution or an improvement over baseline</td><td>Implement your solution</td><td>Run experiments and compare results</td></tr>
<tr><td colspan="2">End-term presentation</td><td></td></tr>
<tr><td colspan="3">Final paper writing</td></tr>
</table>

Grading


Active participation in meetings and discussions
Technical presentation of a scientific paper (the chosen baseline)
End-term presentation
Quality of implementation and coding style
Final paper-style submission

Further Procedure

- To apply for this seminar (bindingly):
  - Send an email to leon.bornemann@hpi.de
  - Deadline: Monday 15.4.2019 23:59
  - We will inform you about the results in Tuesday 16.4.2019
- In case of too many applications, we need to choose randomly.
- Group allocation deadline: 18.4.2019
- Meeting next week: at Campus II, Building F, Room F-2-11.

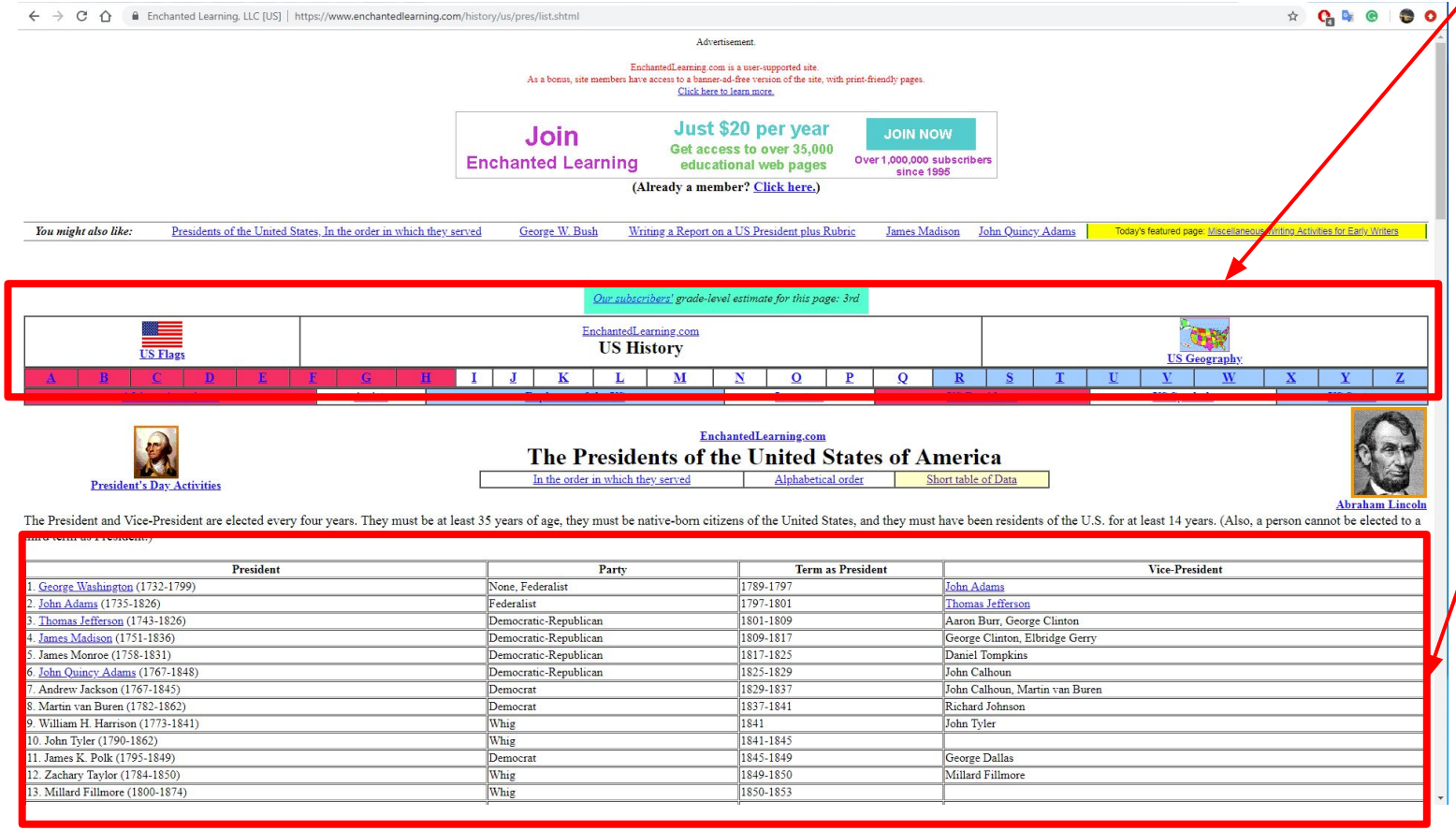
Agenda

1. Chair Introduction
2. Organisational Information + Grading
3. The Research Area of Webtables
   a. History
   b. Typical Problems
   c. Challenges
   d. Use Cases
   e. What do we need Webtables for?
   f. What Datasets/Toolkits exist?
4. Your Research Topics

History

- **WebTables** project started 2007 at Google and "still ongoing".
- Goal
  - Exploit the large and diverse set of informal online structured data in the form of HTML tables.
  - Processing Web tables helps in producing machine-understandable knowledge to power another tasks.
- Contributions:
  - Largest collection of databases and schemas
  - Large-scale extracted schema data for first time enables novel applications

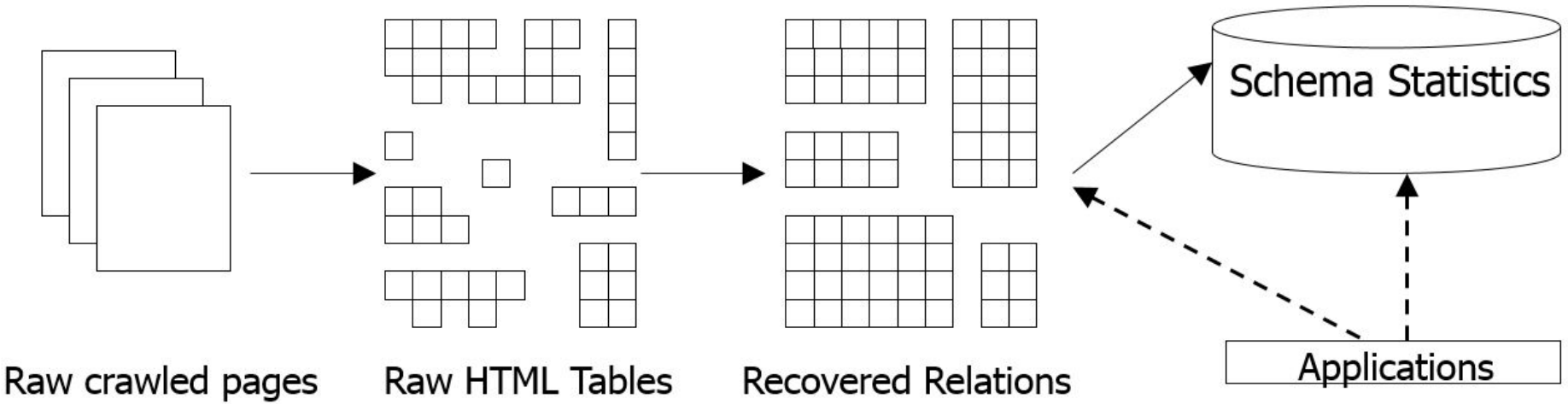
16 distinct HTML tables, but only one relational database



President	Party	Term as President	Vice-President
1. George Washington (1732-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1735-1826)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins
6. John Quincy Adams (1767-1848)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1767-1845)	Democrat	1829-1837	John Calhoun, Martin van Buren
8. Martin van Buren (1782-1862)	Democrat	1837-1841	Richard Johnson
9. William H. Harrison (1773-1841)	Whig	1841	John Tyler
10. John Tyler (1790-1862)	Whig	1841-1845	
11. James K. Polk (1795-1849)	Democrat	1845-1849	George Dallas
12. Zachary Taylor (1784-1850)	Whig	1849-1850	Millard Fillmore
13. Millard Fillmore (1800-1874)	Whig	1850-1853	

https://www.enchantedlearning.com/history/us/pres/list.shtml

WebTables extraction pipeline



- Automatically extracts dbs from web crawl
- A relation is one table+labeled columns

How the corpus was built: Step1: **Is a table relational?**

- **Relational Filtering:** Classifier based on human judgment
- Relational tables:
  - rows represent separate tuple-like objects
  - Columns represent different dimensions of each tuple.
- Recall 81%, Precision 41%

- ➔ #rows
- ➔ #cols
- ➔ % rows w/mostly NULLs
- ➔ # cols w/non-string data
- ➔ Cell strlen
- ➔ ...

How the corpus was built: Step 2: **Has header?**

- **Metadata Detection:** detect header row of attributes labels if recovered.
- Recall 85%, Precision 89%

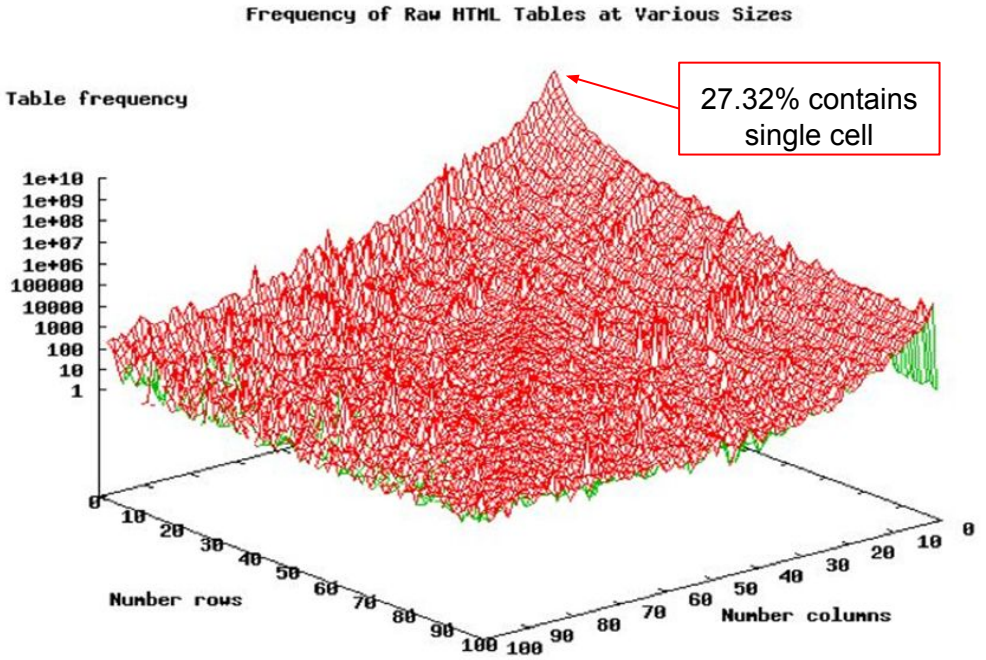
- #rows
- #cols
- % cols w/ lower-case in row1
- ...

How the corpus was built:
Step 3: **Schema statistics**

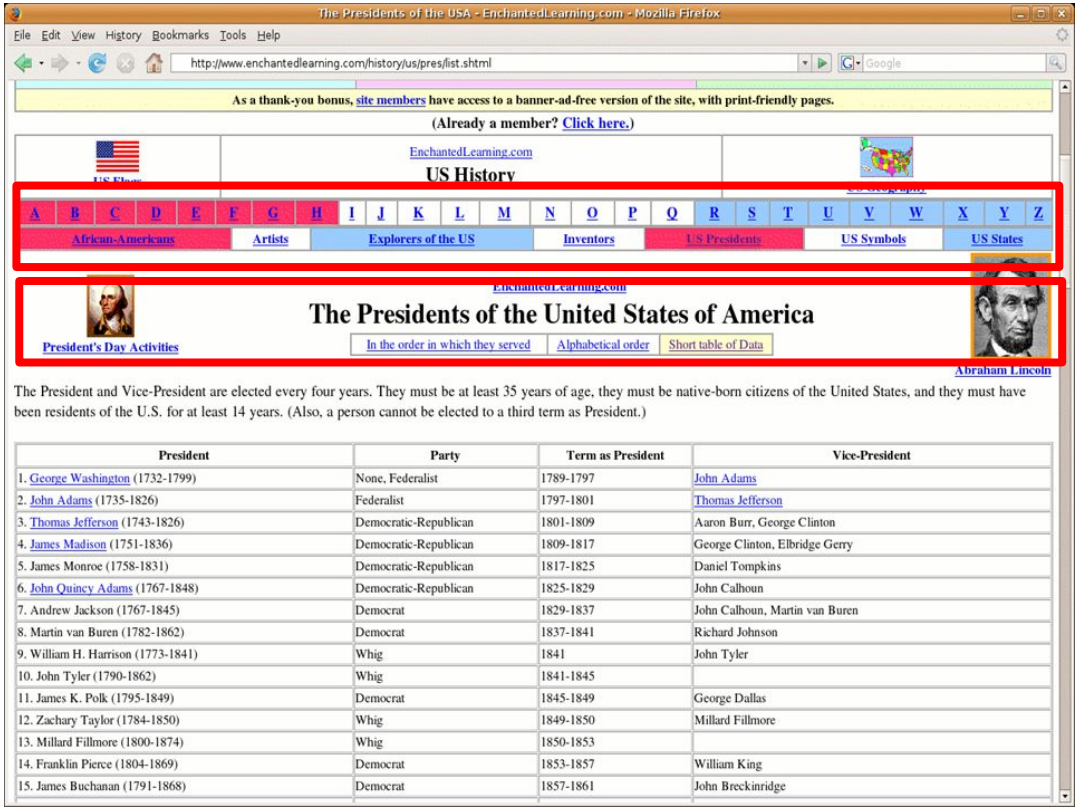
- **ACSDb:** attribute correlation statistics database.
- 5.4M attributes and 2.6M schemata
- Pairs (S,C):
  - S : unique schema/attribute
  - C: how many relations contain S
- Enable the computation of the probability of seeing various attributes in schema and detect relationship between attributes.

WebTables corpus Statistics

- Crawl of 14.1B raw HTML tables
- Tables categories:
  - 88.06% small tables
  - 1.34% HTML forms
  - 0.04% calendars
- Relational?
  - 98.9% non-relational
  - 1.10% relational (154.15M Tables)
- Schemas
  - 2.6M unique relational schemas.



Challenges: bad tables (layout and navigation)



President	Party	Term as President	Vice-President
1. George Washington (1732-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1735-1826)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins
6. John Quincy Adams (1767-1848)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1767-1845)	Democrat	1829-1837	John Calhoun, Martin van Buren
8. Martin van Buren (1782-1862)	Democrat	1837-1841	Richard Johnson
9. William H. Harrison (1773-1841)	Whig	1841	John Tyler
10. John Tyler (1790-1862)	Whig	1841-1845	
11. James K. Polk (1795-1849)	Democrat	1845-1849	George Dallas
12. Zachary Taylor (1784-1850)	Whig	1849-1850	Millard Fillmore
13. Millard Fillmore (1800-1874)	Whig	1850-1853	
14. Franklin Pierce (1804-1869)	Democrat	1853-1857	William King
15. James Buchanan (1791-1868)	Democrat	1857-1861	John Breckinridge

Challenges: Different layouts

	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

Challenges: orientation of a table

Paper Number	20206-PA
DOI What's this?	10.2118/20206-PA
Title	**Theoretical Study of Water Blocking in Miscible Flooding**
Authors	Muller, Thomas, BEB Erdgas and Erdol GmbH; Lake, Larry W., U. of Texas
Journal	SPE Reservoir Engineering
Volume	Volume 6, Number 4
Date	November 1991
Pages	445-451
Copyright	1991. Society of Petroleum Engineers
Language	English

Challenges: Sub-Header Rows

PLANT	COLOR	HEIGHT	BLOOM PERIOD
SHRUBS			
Azalea	variable	shrub	spring
Buddleia	blue, pink, white	shrub	midsummer-fall
Lilac	lavender, white, pink	shrub	spring
Sumac	white	shrub	spring
Vaccinium spp.	white, pink	low shrubs	spring-early summer
Viburnums	white	shrubs	spring
CULTIVATED ANNUALS			
Alyssum	violet, white	4 inches	summer-fall
Candytuft	white, pink	8-10 inches	spring-summer
Cosmos	white, lilac, red, yellow	1-3 feet	late summer

Challenges: No context

Men's open [edit]

Year	Athlete	Country/State or Province	Time	Notes
1897	John J. McDermott	United States(NY)	2:55:10	
1898	Ronald J. MacDonald	Canada (NS)	2:42:00	
1899	Lawrence Brignolia	United States (MA)	2:54:38	
1900	John "Jack" Caffery	Canada (ON)	2:39:44	
1901	John "Jack" Caffery	Canada (ON)	2:29:23	2nd victory
1902	Sammy Mellor	United States (NY)	2:43:12	
1903	John Lorden	United States (MA)	2:41:29	
1904	Michael Spring	United States (NY)	2:38:04	
1905	Frederick Lorz	United States (NY)	2:38:25	
1906	Tim Ford	United States (MA)	2:45:45	
1907	Thomas Longboat	Canada (ON)	2:24:24	
1908	Thomas Morrissey	United States (NY)	2:25:43	
1909	Henri Renaud	United States (NH)	2:53:36	
1910	Fred Cameron	Canada (NS)	2:28:52	
1911	Clarence DeMar	United States (MA)	2:21:39	
1912	Michael J. Ryan	United States (NY)	2:21:18	
1913	Fritz Carlson	United States (MN)	2:25:14	
1914	James Duffy	Canada (ON)	2:25:14	
1915	Édouard Fabre	Canada (PQ)	2:31:41	
1916	Arthur Roth	United States (MA)	2:27:16	
1917	Bill Kennedy	United States (NY)	2:28:37	

- List of winners of the Boston Marathon???

https://en.wikipedia.org/wiki/List_of_winners_of_the_Boston_Marathon

Challenges: context is subtle

COFFEE PRODUCTION BY COUNTRY IN 2006

The following table lists the total coffee production of each coffee exporting country in the year 2006[1].

Country	60 kilogram bags	Kilograms	Pounds
Brazil	42,512,000	2,550,720,000	5,611,584,000
Vietnam	15,000,000	900,000,000	1,980,000,000
Colombia	11,600,000	696,000,000	1,531,200,000
Indonesia	6,850,000	411,000,000	904,200,000
Ethiopia	5,500,000	330,000,000	726,000,000
India	5,005,000	300,300,000	660,660,000
Mexico	4,500,000	270,000,000	594,000,000
Guatemala	4,000,000	240,000,000	528,000,000
Peru	3,500,000	210,000,000	462,000,000
Honduras	2,700,000	162,000,000	356,400,000
Uganda	2,500,000	150,000,000	330,000,000
Ivory Coast	2,350,000	141,000,000	310,200,000
Costa Rica	1,808,000	108,480,000	238,656,000
El Salvador	1,374,000	82,440,000	181,368,000
Nicaragua	1,300,000	78,000,000	171,600,000

https://coffee.fandom.com/wiki/Coffee_production_by_country_in_2006

Applications/use cases

- **Keyword search**: Returns a ranked list of tables to answer a keyword query. (later integrated in search engine).
- On top of ACSDb:
  - **Schema auto-completion**: suggest most-likely next attribute to add to a schema.

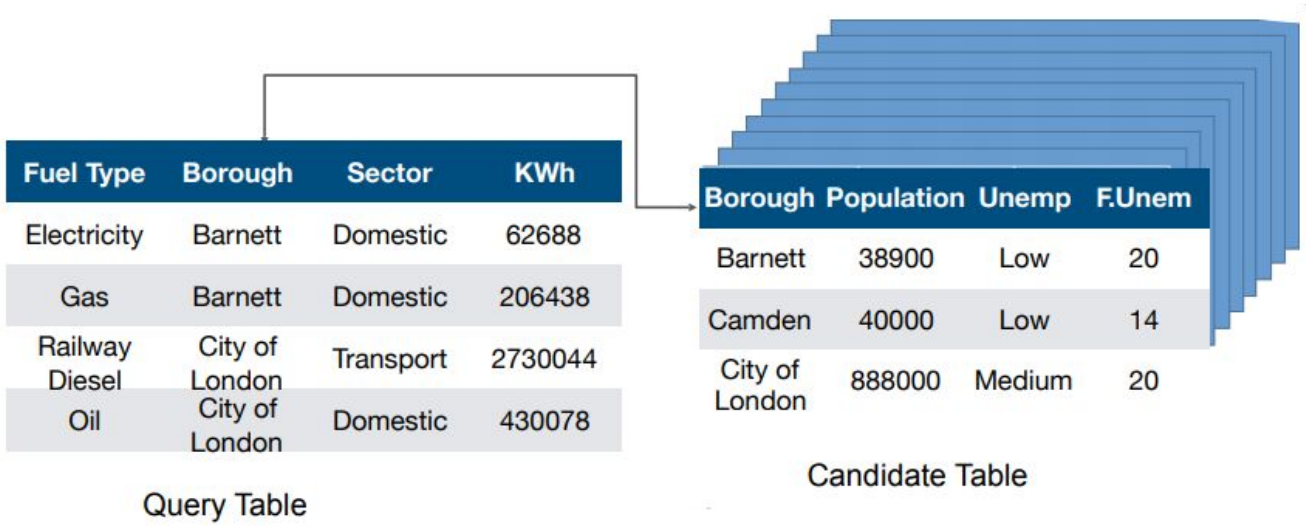
Input attribute	Auto-completer output
name	name, size, last-modified, type
instructor	instructor, time, title, days, room, course
elected	elected, party, district, incumbent, status, opponent, description

  - **Attribute synonym finding**: find pairs does not appear in same schema bur share co-attributes.

Input context	Synonym-finder outputs
name	e-mail|email, phone|telephone, e-mail address|email address, date|last-modified
instructor	course-title|title, day|days, course|course-#, course-name|course-title
elected	candidate|name, presiding-officer|speaker

Applications/use cases

- **Table join:**
  - find the set of tables joinable with a query table.



Fuel Type	Borough	Sector	KWh
Electricity	Barnett	Domestic	62688
Gas	Barnett	Domestic	206438
Railway Diesel	City of London	Transport	2730044
Oil	City of London	Domestic	430078

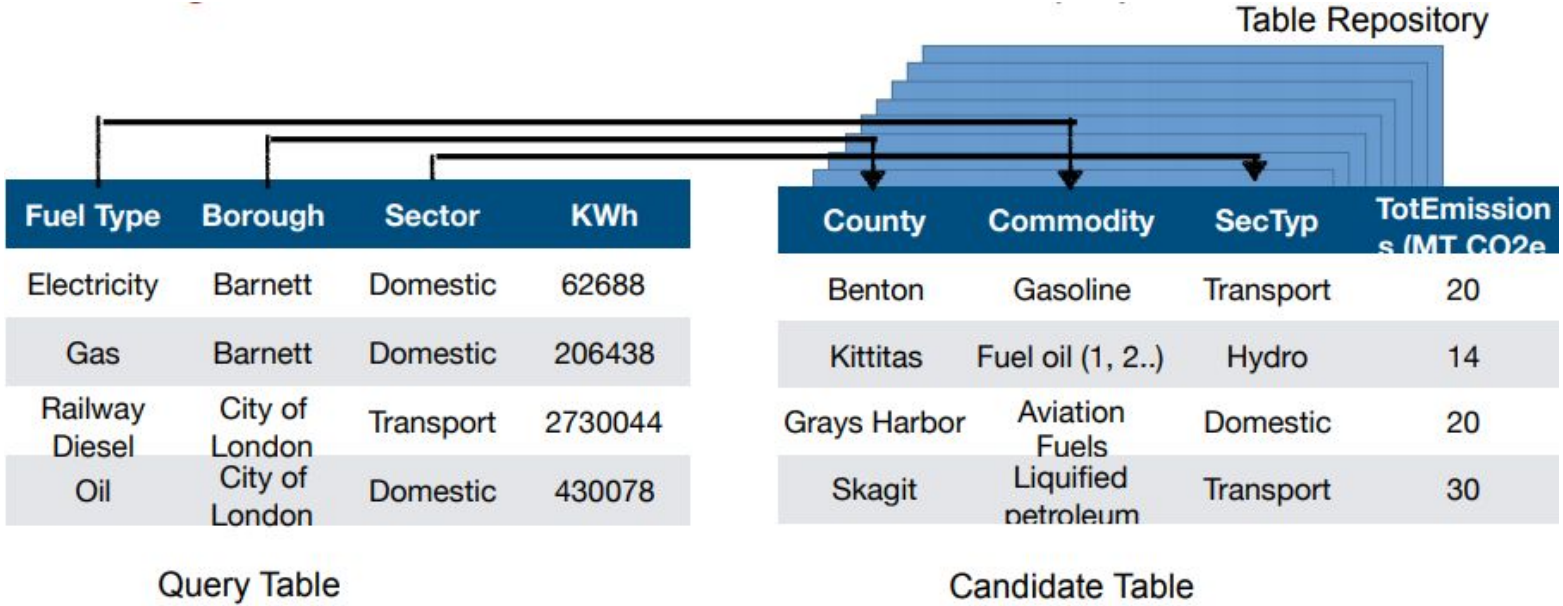
Query Table

Borough	Population	Unemp	F.Unem
Barnett	38900	Low	20
Camden	40000	Low	14
City of London	888000	Medium	20

Candidate Table

Applications/use cases

- **Table union:**
  - find the set of tables that can be unioned with a query table.

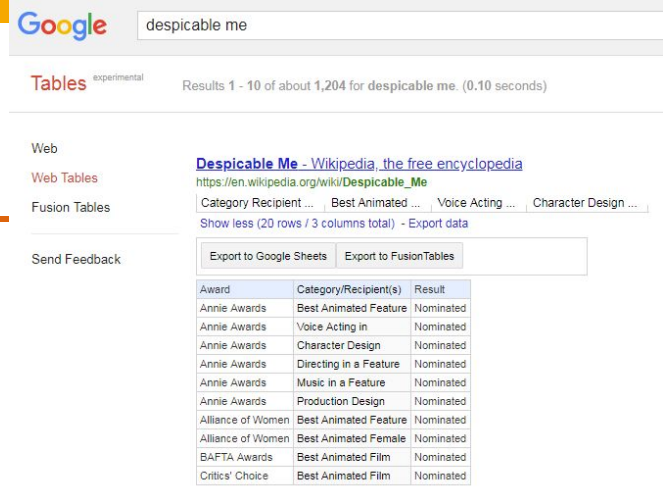


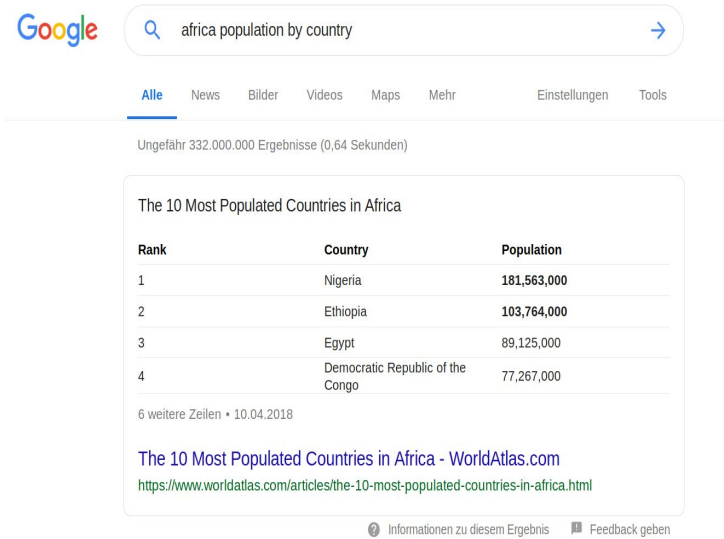
Applications/use cases

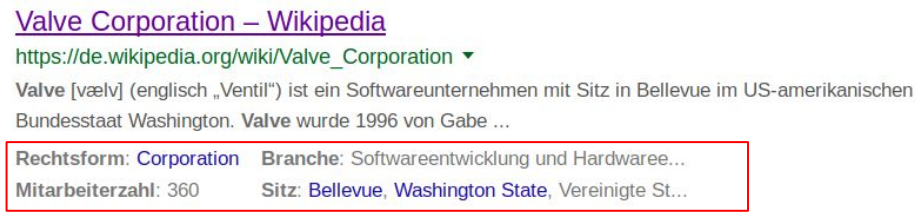
- **Extended class of Table search queries**:
  - *Row-subset query*: return a subset of rows in a larger table
    - Example: Largest software companies in USA
  - *Entity-attribute query*: match a specific attribute in an entity
    - Example: Abeedeen population
- **Entity Linking**
- **Knowledge base completion**
- **Table Enhancement:**
  - *Augment by attribute name, by example..*
- **…**

Products

- Google Tables
  - https://research.google.com/tables
- Google Search Results
  - Tables in Featured Snippets
  - Structured Snippets
- Finding Synonyms
  - Bing synonym API
- Table Fusion API







What do we need Webtables for?

- Change Exploration
  - Track entities and their data over time
  - Why?
    - Curiosity
    - Data Quality (increases trust in data)
    - Performance Optimization (find useless regular updates)
- Wikipedia Tables as a data Source
  - Edit-History is available
  - Tables need to be matched to their successor (done)
  - We need to identify what changed in the table → Key Discovery (my current research topic)

What do we need Webtables for?
Missing Headers

- Missing Headers in Open Data tables:
  - About 28% of the tables have missing header rows in CSV files from the Austrian Open Government and the European Open Data portals [Neumaier16].
  - Around 11k documents have no detectable header row in a data corpus from 232 Open Data portals [Mitlöhner16].
- Missing Headers in Web Tables
  - The majority of tables on the web have missing header row [Balakrishnan15]
  - Around 29% of true web relations extracted by [Cafarella08] suffered from this problem

What do we need Webtables for?

Missing Headers ≈ un-interpreted headers

noe_pop_age_sex_2012_201...

Sort fields | Data source order

Abc noe_pop_age_sex... Nuts1	Abc noe_pop_age_sex... Nuts2	Abc noe_pop_age_sex... Nuts3	# noe_pop_age_sex_2... Lau2 Code
AT1	AT12	AT124	30101
AT1	AT12	AT124	30101
AT1	AT12	AT124	30101
AT1	AT12	AT124	30101
AT1	AT12	AT124	30101
AT1	AT12	AT124	30101
AT1	AT12	AT124	30101

Bevölkerung nach Alter und Geschlecht

Entdecke

Veröffentlichende Stelle	Land Niederösterreich
Kontaktseite der veröffentlichenden Stelle	http://www.noe.gv.at
Datenverantwortliche Stelle	Abteilung Raumordnung und Regionalpolitik-Statistik
Lizenz	Creative Commons Namensnennung 3.0 Österreich
Link zur Lizenz	http://creativecommons.org/licenses/by/3.0/at/legalcode
Link zu den Nutzungsbedingungen	http://data.noe.gv.at/nutzungsbedingungen
Attributbeschreibung	NUTS1: Land NUTS2: Bundesland NUTS3: Gruppen von Bezirken LAU2_CODE: Gemeindekennzahl LAU2_NAME: Gemeindename AGE_GROUP: Altersgruppe (5-jährig) POP_TOTAL: Bevölkerung POP_MALE: männliche Bevölkerung POP_FEMALE: weibliche Bevölkerung YEAR: Referenzjahr

https://www.data.gv.at/katalog/dataset/fde11447-12cf-4333-bed9-c37056551e0c

Agenda

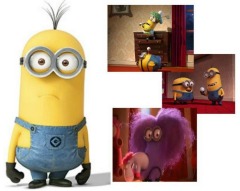
1. Chair Introduction
2. Organisational Information + Grading
3. The Research Area of Webtables
   a. History
   b. Typical Problems
   c. Challenges
   d. Use Cases
   e. What do we need Webtables for?
   f. What Datasets/Toolkits exist?
4. Your Research Topics

Your Research Topics

1. Table Header Detection
2. Relational Webtable Detection
3. Table Normalization

Bob

Kevin



Stuart

Table Header Detection

- Many tables come without explicit headers
  - Understanding the data becomes difficult (and thus further processing)
  - Mixing header and data is obviously bad

Year	Africa	Americas	Asia & Pacific	Europe
2014	2,300	8,950	*9,325*	4,200
2015	2,725	*9,200*	8,850	4,775

```
<table>
        <tr>
                <td> <b> Year </b> </td> ...
        </tr>
        ...
</table>
```

Club	Season	League		Cup	
		Apps	Goals	Apps	Goals
Bayern Munich II	2005–06	1	0	—	
	2006–07	31	2	—	
	2007–08	10	3	—	
	Totals	**42**	**5**	**—**	

Table Header Detection

- What is the State of the Art?
  - Machine Learning to the rescue!
  - Use tables with explicit headers (th-tags) as training data
  - Create model with handcrafted features
  - Use the created model to find headers in tables without th-tags
- Other potential Approaches using Machine Learning
  - Manually engineer different/more features
  - Use Deep Learning
  - Render html and do Image Recognition

Relational Webtable Recognition

There are many different types of tables

Year	Africa	Americas	Asia & Pacific	Europe
2014	2,300	8,950	*9,325*	4,200
2015	2,725	*9,200*	8,850	4,775

Relational Table

Layout Table

44th President of the United States	
	In office
	January 20, 2009 – January 20, 2017
Vice President	Joe Biden
Preceded by	George W. Bush
Succeeded by	Donald Trump
United States Senator from Illinois	

Key-Value Pairs (Infobox)

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

Matrix

Relational Webtable Recognition

- What is the State of the Art?
  - Original Approach
    - A few handcrafted rules to filter out large amounts of non-relational tables
    - Training a classifier with the help of human labelling
- Other potential Approaches using Machine Learning
  - Manually engineer different/more features
  - Use Deep Learning
  - Render html and do Image Recognition ← probably difficult to get enough training data

Webtable Normalization

Many tables might not be normalized (Join-Tables)

Team	Location	Stadium	Capacity
FC Augsburg	Augsburg	WWK ARENA	30,660
Bayer Leverkusen	Leverkusen	BayArena	30,210
Bayern Munich	Munich	Allianz Arena	75,000
Borussia Dortmund	Dortmund	Signal Iduna Park	81,359
Borussia Mönchengladbach	Mönchengladbach	Stadion im Borussia-Park	59,724

Key/Subject Column

Extracted RDF-Triples:

- <FC-Augsburg, Location, Augsburg>
- <FC-Augsburg, Stadium, WWK ARENA>
- <FC-Augsburg, Capacity, 30,660>
- ...

Wrong Fact!

Webtable Normalization

- Table should be split on Functional Dependency Stadium → Capacity
  - (Team,Location,Stadium)
  - (Stadium,Capacity)
- What is the State of the Art?
  - Detect Functional Dependencies (FDs) and Subject Columns
  - Rank the resulting FDs and split the tables accordingly
- Other Potential Approaches
  - Come up with better scores
  - Use Machine Learning (?)
  - ???

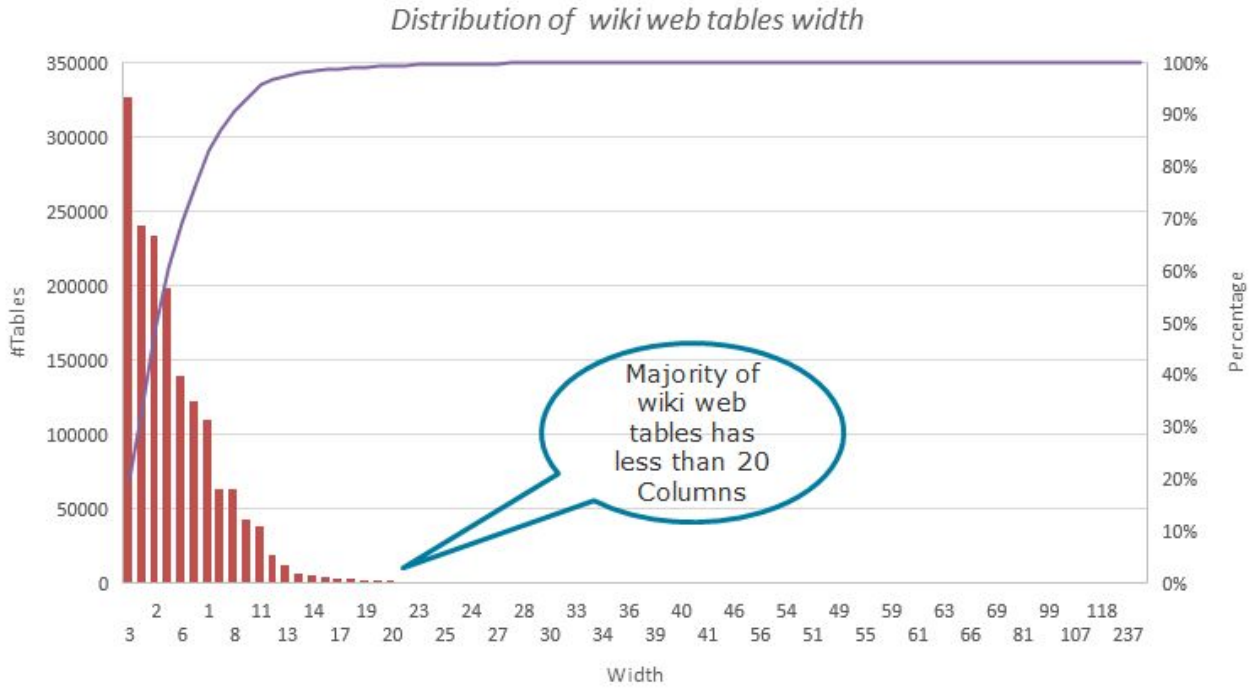
Datasets

- Wikipedia web tables: http://websail-fe.cs.northwestern.edu/TabEL/

- Web data common:http://webdatacommons.org/webtables/index.html

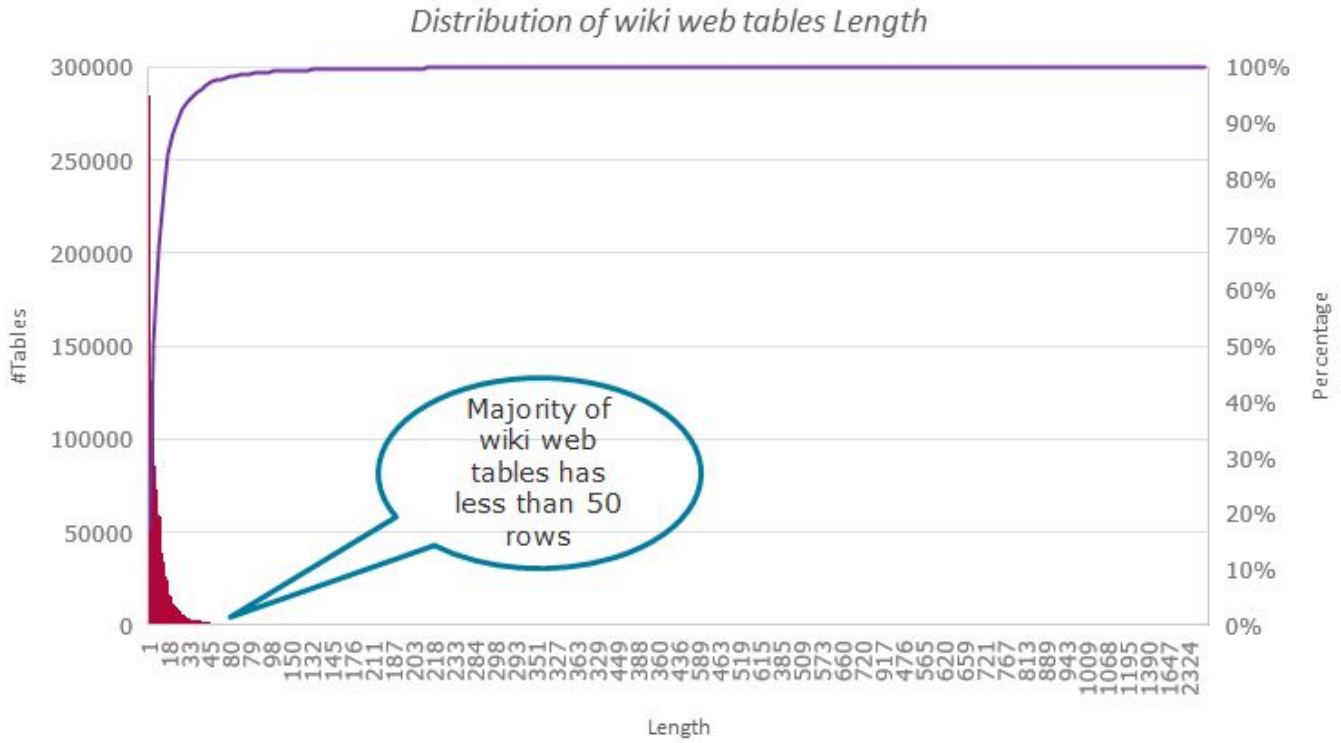
Wikipedia web tables

- A dataset of 1.6M Wikipedia Tables in JSON format



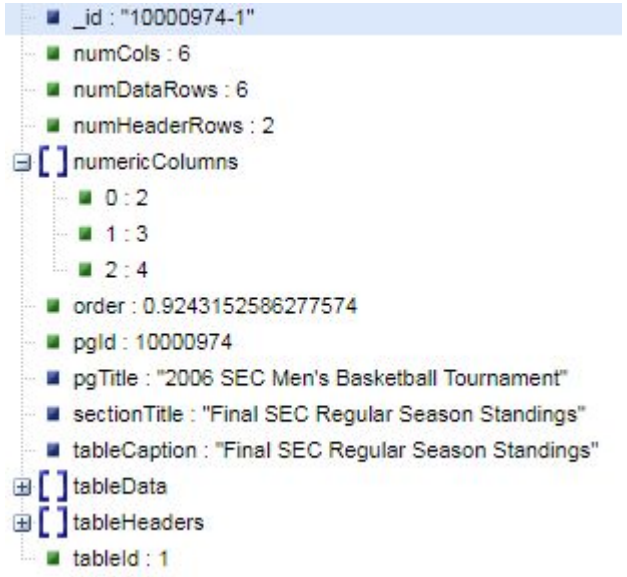
Wikipedia web tables

- A dataset of 1.6M Wikipedia Tables in JSON format



Wikipedia web tables

- A dataset of 1.6M Wikipedia Tables in JSON format



Final SEC Regular Season Standings [edit]

SEC East					
School	Coach	W	L	Pct	Seed
Tennessee	Bruce Pearl	12	4	.750	E1
Florida	Billy Donovan	10	6	.625	E2
Kentucky	Tubby Smith	9	7	.563	E3
Vanderbilt	Kevin Stallings	7	9	.438	E4
South Carolina	Dave Odom	6	10	.375	E5
Georgia	Dennis Felton	5	11	.313	E6

WDC Web Table Corpus 2015

- 233M Web tables, each table has:
  - one of the categories: Relational,Entity, Matrix
  - metadata including table orientation, header rows, key columns
  - context information such as the title of the HTML page, the caption of the table, the text before and after the table, and timestamps from the page

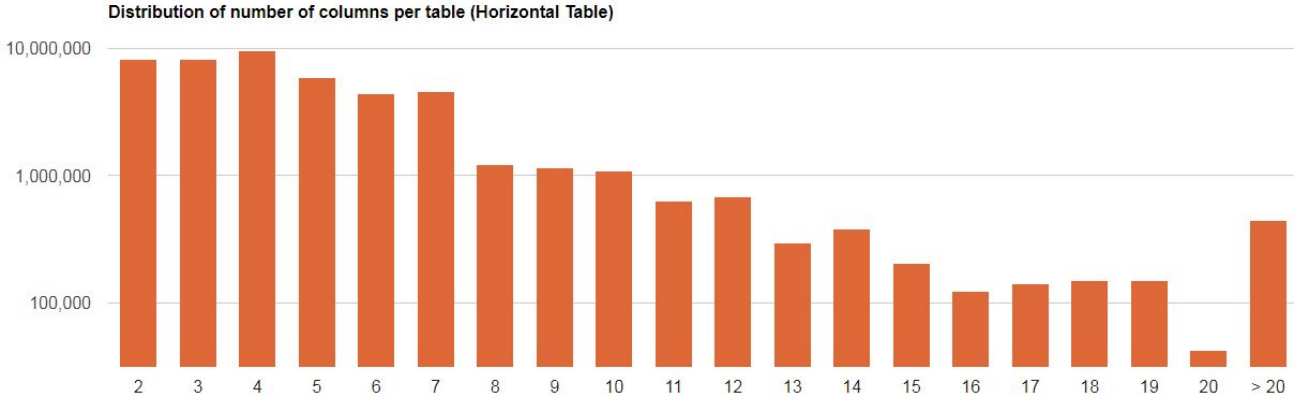
Data Set	Size	#Files
Complete Corpus 2015	165 GB	99 (.tar)
Relational Corpus 2015	69 GB	99 (.tar)
English-Language Relational Web Tables 2015	69 GB	51 (.tar)

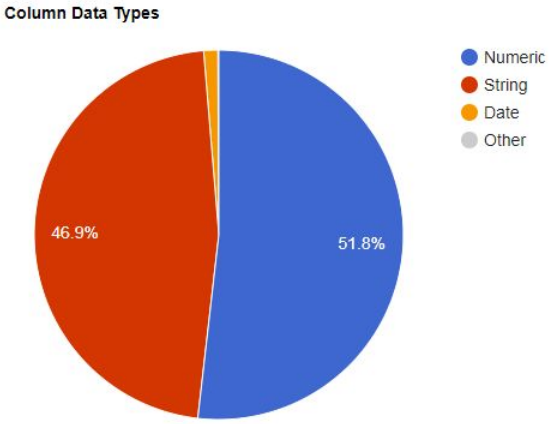
WDC Web Table Corpus 2015-example

{ "relation": [["#","1","2","3"], ["Club","Barcelona","Real Madrid","Bayern München"],
["Country","ESP","ESP","GER"],
["Points","2037","2008","1973"]],
"Title": "",
"hasHeader": true,
"headerPosition": "FIRST_ROW",
"tableType": "RELATION",
"tableOrientation": "HORIZONTAL",
"hasKeyColumn": true,
"keyColumnIndex": 1,
"headerRowIndex": 0,
.
.
.
.
}

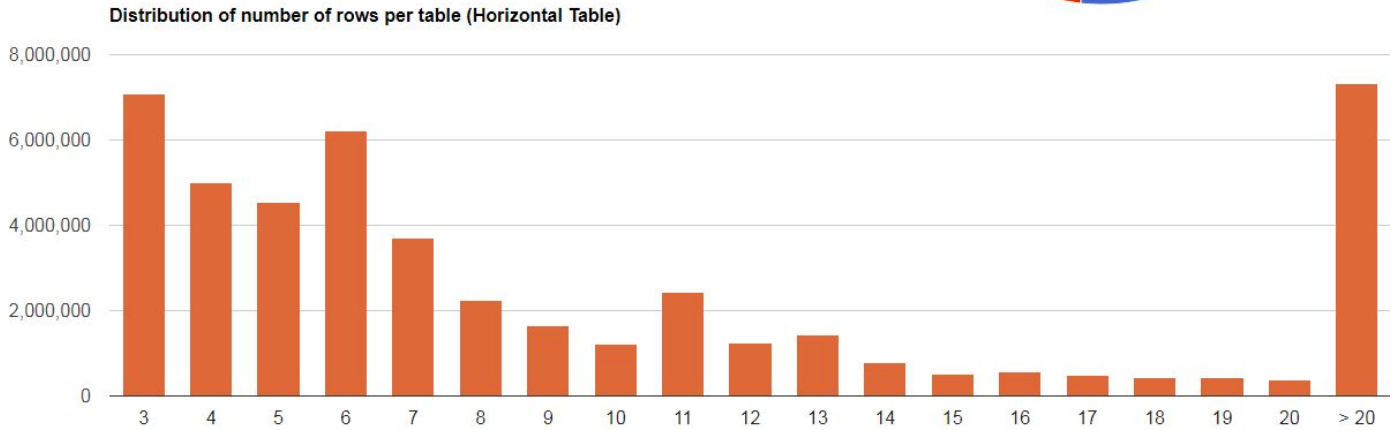
#	Club	Country	Points
1	Barcelona	ESP	2037
2	Real Madrid	ESP	2008
3	Bayern München	GER	1973

English-Language Relational Web Tables 2015-statistics





horizontal	47,669,450
vertical	3,150,715
sum	50,820,165



References

- [Neumaier16] Neumaier, S., Umbrich, J., Parreira, J. X., & Polleres, A. (2016, October). Multi-level semantic labelling of numerical values. In International Semantic Web Conference (pp. 428-445). Springer, Cham.
- [Mitlöhner16] Mitlöhner, J., Neumaier, S., Umbrich, J., & Polleres, A. (2016, August). Characteristics of open data csv files. In Open and Big Data (OBD), International Conference on (pp. 72-79). IEEE.
- [Balakrishnan15] S. Balakrishnan, A. Y. Halevy, B. Harb, H. Lee, J. Madhavan, A. Rostamizadeh, W. Shen, K. Wilder, F. Wu, and C. Yu. Applying webtables in practice. In CIDR, 2015.
- [Cafarella08] M. J. Cafarella, A. Y. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. Uncovering the relational web. In WebDB, 2008.
- [Cafarella18] Cafarella, M., Halevy, A., Lee, H., Madhavan, J., Yu, C., Wang, D. Z., & Wu, E. (2018). Ten years of webtables. Proceedings of the VLDB Endowment, 11(12), 2140-2149.