

Automatic Web Spreadsheet Data Extraction

Shirley Zhe Chen
Michael Cafarella

Spreadsheets Are Everywhere

More than 400 million Excel users worldwide.

—Microsoft

50 to 80% of enterprises use standalone spreadsheets for critical applications.

—Forester

Spreadsheets Are Everywhere

- Our Web crawl obtained **410,554** Microsoft Excel Files from **51,252** distinct Internet domains.

Domains	# of spreadsheets	% of total
www.bts.gov	12435	3.03%
www.census.gov	7862	1.91%
www.stat.co.jp	6633	1.62%
www.bankofengland.co.uk	5520	1.34%
www.ers.usda.gov	4328	1.05%
www.agr.gc.ca	4186	1.02%
www.wto.org	3863	0.94%
www.doh.wa.gov	3579	0.87%
www.nsf.gov	2770	0.67%
nces.ed.gov	2177	0.53%

OUR GOAL: Integration

- Spreadsheets often contain data that are roughly relational, but the schema is entirely implicit.
- **Example:** An analyst may want to combine a spreadsheet about *company sales* with a government produced spreadsheet about *economic performance* to predict future sales.

Relational Data Enables Integration

Web browser interface showing a search for "smoking" on the Senbazuru website. The interface includes a search bar, navigation buttons (Previous, Next), and a result count (Result 5 of 9). The data is displayed in a table format with columns for demographic factors and a numerical value.

Female, total,	Black, total,	18 to 24 years,	1965.0,	37.1
Female, total,	Black, total,	18 to 24 years,	1974.0,	35.6
Female, total,	Black, total,	18 to 24 years,	1979.0,	31.8
Female, total,	Black, total,	18 to 24 years,	1983.0,	32.0
Female, total,	Black, total,	18 to 24 years,	1985.0,	23.7
Female, total,	Black, total,	18 to 24 years,	1987.0,	20.4
Female, total,	Black, total,	18 to 24 years,	1988.0,	21.8
Female, total,	Black, total,	18 to 24 years,	1990 \1,	10.0
Female, total,	Black, total,	18 to 24 years,	1991.0,	11.8
Female, total,	Black, total,	18 to 24 years,	1992.0,	10.3

Tablet interface showing a "Join" operation. The screen displays two tables side-by-side, each with columns for demographic factors and a numerical value. The tables are titled "Table 100: Current Cigarette Smoking by Sex and State 2007" and "Table 111: Current Cigarette Smoking by Sex and State 2008".

Table 100: Current Cigarette Smoking by Sex and State 2007	Table 111: Current Cigarette Smoking by Sex and State 2008
Alabama	Alabama
Alaska	Alaska
Arizona	Arizona
Arkansas	Arkansas
California	California
Colorado	Colorado
Connecticut	Connecticut
Delaware	Delaware
District of Columbia	District of Columbia
Florida	Florida
Georgia	Georgia
Hawaii	Hawaii
Idaho	Idaho
Illinois	Illinois
Indiana	Indiana
Iowa	Iowa
Kansas	Kansas
Kentucky	Kentucky
Louisiana	Louisiana
Maine	Maine
Maryland	Maryland
Massachusetts	Massachusetts
Michigan	Michigan
Minnesota	Minnesota
Mississippi	Mississippi
Missouri	Missouri
Montana	Montana
Nebraska	Nebraska
Nevada	Nevada
New Hampshire	New Hampshire
New Jersey	New Jersey
New Mexico	New Mexico
New York	New York
North Carolina	North Carolina
North Dakota	North Dakota
Ohio	Ohio
Oklahoma	Oklahoma
Oregon	Oregon
Pennsylvania	Pennsylvania
Rhode Island	Rhode Island
South Carolina	South Carolina
South Dakota	South Dakota
Tennessee	Tennessee
Texas	Texas
Utah	Utah
Vermont	Vermont
Virginia	Virginia
Washington	Washington
West Virginia	West Virginia
Wisconsin	Wisconsin
Wyoming	Wyoming

How to Extract Relational Data from Spreadsheets to Enable Integration?

Related Work

- Transform spreadsheets into relational format but require user specified rules, e.g.
 - Hung et al. (transformation languages)
- Automatic extraction on a simple and specific type of spreadsheets, e.g.
 - Cunha et al. (focus on the type of spreadsheets with relational tables)
 - Ahmad et al. (detect spreadsheet errors)

Challenges: Implicit Structures

5			
6	Sex, age, and race	1990 \1	2000
7			
19	Total smokers \3	25.5	23.2
20	Male, total	28.4	25.6
21	18 to 24 years	26.6	28.1
22	25 to 34 years	31.6	28.9
23	35 to 44 years	34.5	30.2
24	45 to 64 years	29.3	26.4
25	65 years and over	14.6	10.2
26	White, total	28.0	25.7
27	18 to 24 years	27.4	30.4
28	25 to 34 years	31.6	29.7
29	35 to 44 years	33.5	30.6
30	45 to 64 years	28.7	25.8
31	65 years and over	13.7	9.8
32	Black, total	32.5	26.2
33	18 to 24 years	21.3	20.9
34	25 to 34 years	33.8	23.2
35	35 to 44 years	42.0	30.7
36	45 to 64 years	36.7	32.2
37	65 years and over	21.5	14.2

Challenges: Implicit Structures

5			
6	Sex, age, and race	1990 \1	2000
7			
19	Total smokers \3	25.5	23.2
20	Male, total	28.4	25.6
21	18 to 24 years	26.6	28.1
22	25 to 34 years	31.6	28.9
23	35 to 44 years	34.5	30.2
24	45 to 64 years	29.3	26.4
25	65 years and over	14.6	10.2
26	White, total	28.0	25.7
27	18 to 24 years	27.4	30.4
28	25 to 34 years	31.6	29.7
29	35 to 44 years	33.5	30.6
30	45 to 64 years	28.7	25.8
31	65 years and over	13.7	9.8
32	Black, total	32.5	26.2
33	18 to 24 years	21.3	20.9
34	25 to 34 years	33.8	23.2
35	35 to 44 years	42.0	30.7
36	45 to 64 years	36.7	32.2
37	65 years and over	21.5	14.2

Challenges: Implicit Structures

5			
6	Sex, age, and race	1990 \1	2000
7			
19	Total smokers \3	25.5	23.2
20	Male, total	28.4	25.6
21	18 to 24 years	26.6	28.1
22	25 to 34 years	31.6	28.9
23	35 to 44 years	34.5	30.2
24	45 to 64 years	29.3	26.4
25	65 years and over	14.6	10.2
26	White, total	28.0	25.7
27	18 to 24 years	27.4	30.4
28	25 to 34 years	31.6	29.7
29	35 to 44 years	33.5	30.6
30	45 to 64 years	28.7	25.8
31	65 years and over	13.7	9.8
32	Black, total	32.5	26.2
33	18 to 24 years	21.3	20.9
34	25 to 34 years	33.8	23.2
35	35 to 44 years	42.0	30.7
36	45 to 64 years	36.7	32.2
37	65 years and over	21.5	14.2

Relational Tuples:

1990	Male	White	45 to 64 years	28.7
1990	Male	White	65 years and over	13.7
2000	Male	White	45 to 64 years	25.8
2000	Male	Black	65 years and over	14.2

Outline

1. Introduction

2. System Framework

3. Experiments

4. Conclusion

Spreadsheet Terminologies

- A *data frame* is a three-part spreadsheet structure, consisting of *attribute* and *value* regions.

5			
6	Sex, age, and ra		
7			
19	Total smokers \3	25.5	23.2
20	Male, total	28.4	25.6
21	18 to 24 years	26.6	28.1
22	25 to 34 years	31.6	28.9
23	35 to 44 years	34.5	30.2
24	45 to 64 years		
25	65 years and over		
26			
27			
28			
29			
30			
31	65 years and over	28.7	25.8
32	Black, total	13.7	9.8
33	18 to 24 years	21.3	20.9
34	25 to 34 years	33.8	23.2
35	35 to 44 years	42.0	30.7
36	45 to 64 years	36.7	32.2
37	65 years and over	21.5	14.2

Top Attributes

Left Attributes

Value Region

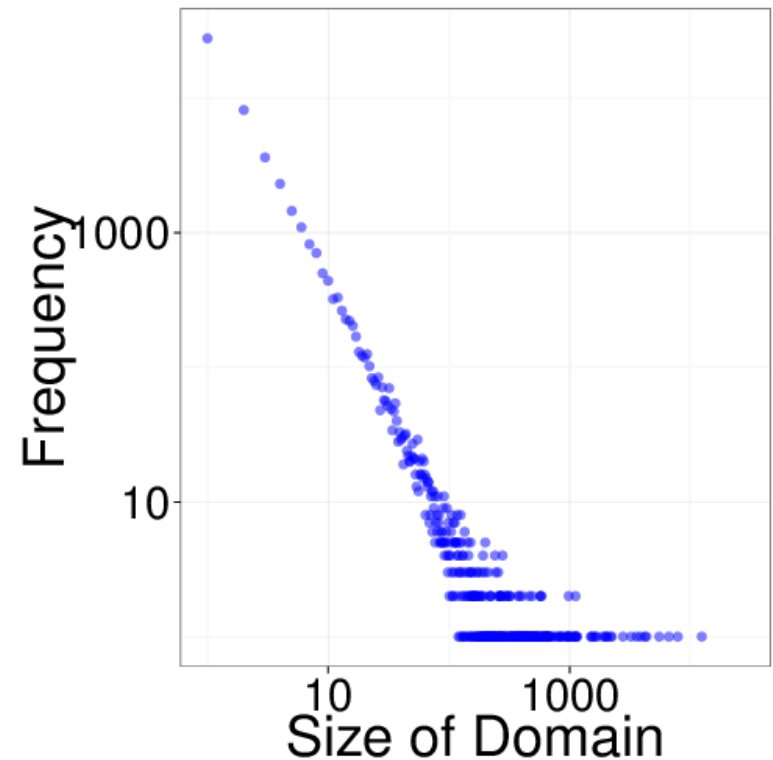
Spreadsheet Terminologies

- A *hierarchical spreadsheet* is a data frame spreadsheet with either a hierarchical *left* or *top* attribute region.

5			
6	Sex, age, and race	1990 \1	2000
7			
19	Total smokers \3	25.5	23.2
20	Male, total	28.4	25.6
21	18 to 24 years	26.6	28.1
22	25 to 34 years	31.6	28.9
23	35 to 44 years	34.5	30.2
24	45 to 64 years	29.3	26.4
25	65 years and over	14.6	10.2
26	White, total	28.0	25.7
27	18 to 24 years	27.4	30.4
28	25 to 34 years	31.6	29.7
29	35 to 44 years	33.5	30.6
30	45 to 64 years	28.7	25.8
31	65 years and over	13.7	9.8
32	Black, total	32.5	26.2
33	18 to 24 years	21.3	20.9
34	25 to 34 years	33.8	23.2
35	35 to 44 years	42.0	30.7
36	45 to 64 years	36.7	32.2
37	65 years and over	21.5	14.2

Web Spreadsheets Observations

- *Strongly skewed distribution*



Web Spreadsheets Observations

- *Strongly skewed distribution.*
- *32.5%* of the Web spreadsheets are hierarchical.
- *More than 60%* spreadsheets in the top 10 Internet domains are hierarchical.

Our system requirements

- The system has to process data frame spreadsheets, especially the hierarchical ones.
- The system has to be automatic.

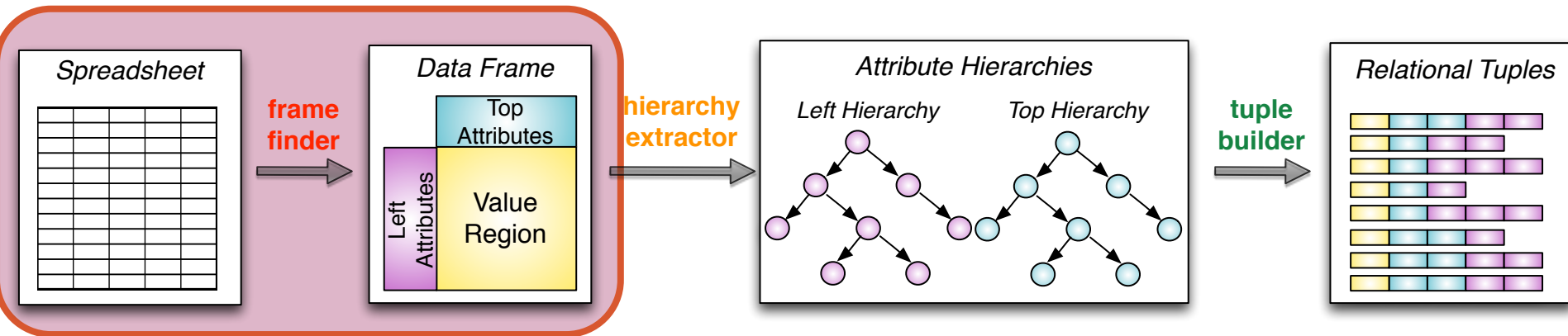
Problem Formulation

We present the first automatic, domain-independent spreadsheet extractor, the first step in building a spreadsheet integration tool.

- **Input:** A *data frame* spreadsheet
- **Output:** The *relational tuples* for the spreadsheet

Three-Stage Pipeline

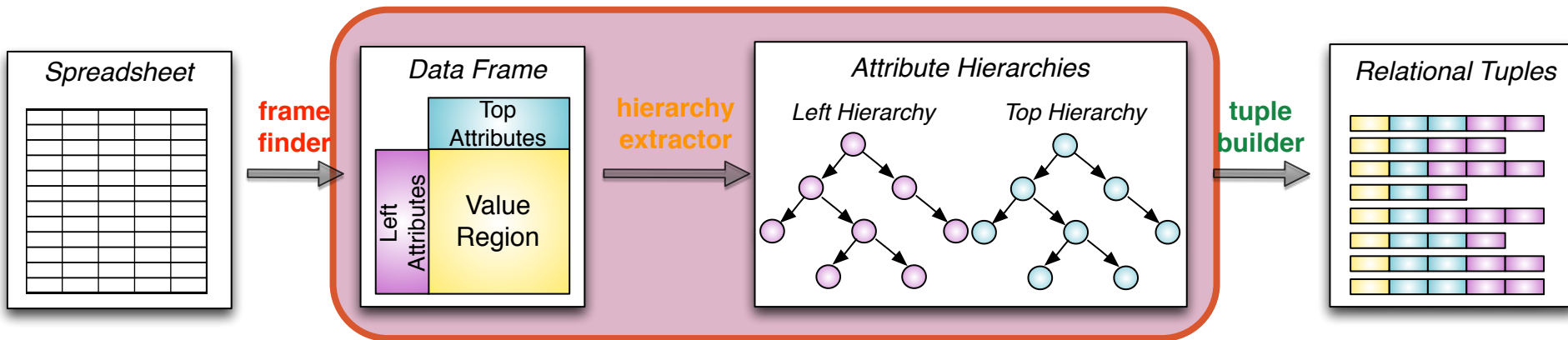
1. Frame Finder
2. Hierarchy Extractor
3. Tuple Builder



Frame Finder

Three-Stage Pipeline

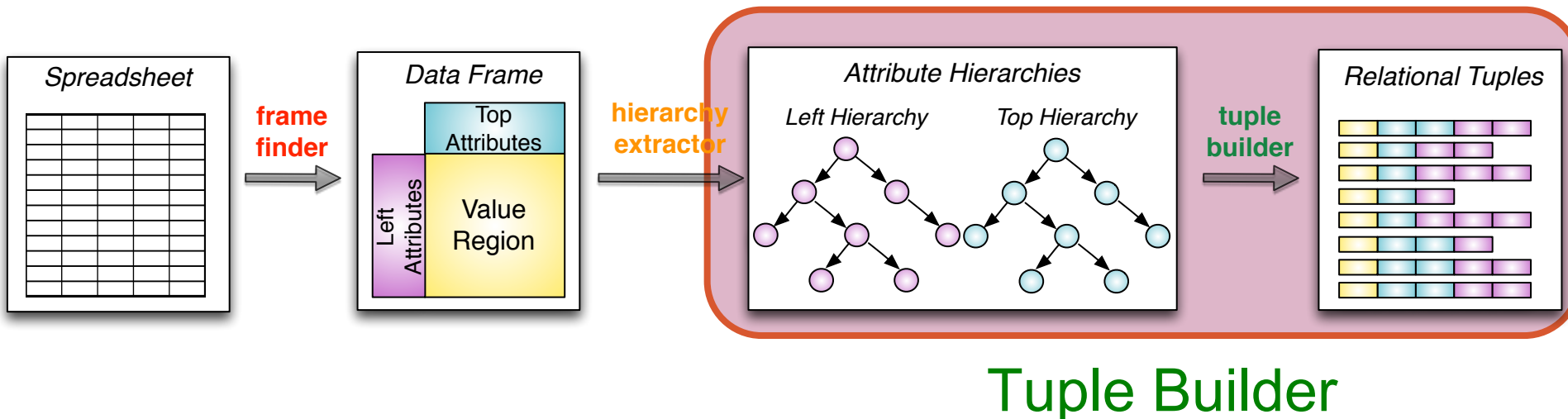
1. Frame Finder
2. Hierarchy Extractor
3. Tuple Builder



Hierarchy Extractor

Three-Stage Pipeline

1. Frame Finder
2. Hierarchy Extractor
3. Tuple Builder



Step 1: Frame Finder

- The *frame finder* detects the three semantic regions in a spreadsheet.
- Simplify the task as a *row labeler* task: For each row in a spreadsheet, assign a label in *{title, header, data, footnote}*.

Step 1: Frame Finder

- The *row labeler* is based on a CRF to encode two types of observations:
 - The properties of each row indicate its semantic label.
 - The labels assigned to adjacent rows are highly related.

Step 1: Frame Finder

title	1	Table 199. Current Cigarette Smoking		
	2			
footnote	3	See notes.		
	4			
	5			
header	6	Sex, age, and race	1990 \1	2000
	7			
data	19	Total smokers \3	25.5	23.2
data	20	Male, total	28.4	25.6
data	21	18 to 24 years	26.6	28.1
data	22	25 to 34 years	31.6	28.9
data	23	35 to 44 years	34.5	30.2
data	24	45 to 64 years	29.3	26.4
data	25	65 years and over	14.6	10.2
data	26	White, total	28.0	25.7
data	27	18 to 24 years	27.4	30.4
data	28	25 to 34 years	31.6	29.7
data	29	35 to 44 years	33.5	30.6
data	30	45 to 64 years	28.7	25.8
data	31	65 years and over	13.7	9.8

Step 1: Frame Finder

title	1	Table 199. Current Cigarette Smoking			
	2				
footnote	3	See notes.			
	4				
	5				
header	6	Sex,	Top Attribute Region		2000
	7				
data	19	Total smokers \3		25.5	23.2
data	20	Male, total		28.4	25.6
data	21	18 to 24 years		26.6	28.1
data	22	25 to 34 years		31.6	28.9
data	23	35 to 44 years			
data	24	45 to 64 years			
data	25	65 years and over			
data	26	White		28.0	25.7
data	27	18 to 24 years		27.4	30.4
data	28	25 to 34 years		31.6	29.7
data	29	35 to 44 years		33.5	30.6
data	30	45 to 64 years		28.7	25.8
data	31	65 years and over		13.7	9.8

Step 2: Hierarchy Extractor

- The *hierarchy extractor* recovers the attribute hierarchy for the left or top attribute region.
- It identifies all the annotation attribute pairs in the attribute region, thus recovering the attribute hierarchy.

Step 2: Hierarchy Extractor

- Algorithm 1: Classification
 - We enumerate all the attribute pairs in an attribute region as the *annotation attribute pair* candidates, and each of the candidate takes a label from $\{true, false\}$.
 - E.g. (White, Male) = *true* and (White, Black) = *false*.

19	Total smokers \3
20	Male, total
21	18 to 24 years
22	25 to 34 years
23	35 to 44 years
24	45 to 64 years
25	65 years and over
26	White, total
27	18 to 24 years
28	25 to 34 years
29	35 to 44 years
30	45 to 64 years
31	65 years and over
32	Black, total
33	18 to 24 years
34	25 to 34 years
35	35 to 44 years
36	45 to 64 years
37	65 years and over

Step 2: Hierarchy Extractor

- Algorithm 1: Classification
 - We enumerate all the attribute pairs in an attribute region as the *annotation attribute pair* candidates, and each of the candidate takes a label from $\{true, false\}$.
 - E.g. (White, Male) = *true* and (White, Black) = *false*.

19	Total smokers \3
20	Male, total
21	18 to 24 years
22	25 to 34 years
23	35 to 44 years
24	45 to 64 years
25	65 years and over
26	White, total
27	18 to 24 years
28	25 to 34 years
29	35 to 44 years
30	45 to 64 years
31	65 years and over
32	Black, total
33	18 to 24 years
34	25 to 34 years
35	35 to 44 years
36	45 to 64 years
37	65 years and over

False

Step 2: Hierarchy Extractor

- Algorithm 2: Enforced-tree Classification
 - Obtain the probability associated with each annotation pair during the classification.

19	Total smokers \3
20	Male, total
21	18 to 24 years
22	25 to 34 years
23	35 to 44 years
24	45 to 64 years
25	65 years and over
26	White, total
27	18 to 24 years
28	25 to 34 years
29	35 to 44 years
30	45 to 64 years
31	65 years and over
32	Black, total
33	18 to 24 years
34	25 to 34 years
35	35 to 44 years
36	45 to 64 years
37	65 years and over

P = 0.8

P = 0.7

Step 2: Hierarchy Extractor

- Algorithm 2: Enforced-tree Classification
 - Obtain the probability associated with each annotation pair during the classification.
 - For a child, find the parent with the highest probability.

19	Total smokers \3
20	Male, total
21	18 to 24 years
22	25 to 34 years
23	35 to 44 years
24	45 to 64 years
25	65 years and over
26	White, total
27	18 to 24 years
28	25 to 34 years
29	35 to 44 years
30	45 to 64 years
31	65 years and over
32	Black, total
33	18 to 24 years
34	25 to 34 years
35	35 to 44 years
36	45 to 64 years
37	65 years and over

The diagram illustrates the hierarchy extractor algorithm. It shows a list of 19 annotations (rows 19-37) and their parent-child relationships. The parent-child relationships are indicated by arrows and probabilities:

- Annotation 21 (18 to 24 years) is a child of Annotation 20 (Male, total) with a probability of $P = 0.8$.
- Annotation 25 (65 years and over) is a child of Annotation 24 (45 to 64 years) with a probability of $P = 0.7$.

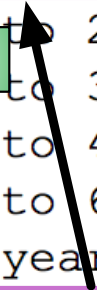
The annotations are grouped into three main categories: Total smokers, White, and Black, each with a total annotation and five age-based sub-annotations.

Step 2: Hierarchy Extractor

- Algorithm 2: Enforced-tree Classification
 - Obtain the probability associated with each annotation pair during the classification.
 - For a child, find the parent with the highest probability.

19	Total smokers \3
20	Male, total
21	18 to 24 years
22	25 to 34 years
23	35 to 44 years
24	45 to 64 years
25	65 years and over
26	White, total
27	18 to 24 years
28	25 to 34 years
29	35 to 44 years
30	45 to 64 years
31	65 years and over
32	Black, total
33	18 to 24 years
34	25 to 34 years
35	35 to 44 years
36	45 to 64 years
37	65 years and over

P = 0.8



Step 3: Tuple Builder

- The *tuple builder* generates relational tuples for each value in the value region.

5			
6	Sex, age, and race	1990 \1	2000
7			
19	Total smokers \3	25.5	23.2
20	Male, total	28.4	25.6
21	18 to 24 years	26.6	28.1
22	25 to 34 years	31.6	28.9
23	35 to 44 years	34.5	30.2
24	45 to 64 years	29.3	26.4
25	65 years and over	14.6	10.2
26	White, total	28.0	25.7
27	18 to 24 years	27.4	30.4
28	25 to 34 years	31.6	29.7
29	35 to 44 years	33.5	30.6
30	45 to 64 years	28.7	25.8
31	65 years and over	13.7	9.8

1990	Total smokers	Male, total	White, total	45 to 64 years	28.7
------	---------------	-------------	--------------	----------------	------

Outline

1. Introduction

2. System Framework

3. Experiments

4. Conclusion

Experiment Setup

- Our WEB dataset has 410,554 Excel files from 51,252 distinct Internet domains.
 - We randomly selected 100 random *hierarchical* spreadsheets (with hierarchical top or left attributes).
 - Average depth of the *top* hierarchy is 2.14 with the maximum 5.
 - Average depth of the *left* hierarchy is 2.61 with the maximum 9.

Experiment Setup

- For all the experiments, we split the 100 spreadsheets into 50 training and 50 testing for 10 times. conducted equal-sized training and testing for 10 times and obtained the average value for the following metrics:
 - *Precision* and *Recall*
 - *Error per sheet* = (false positive + false negative)/N where N is the # of sheets.

Frame Finder

- The row labeler assigns a label in {title, header, data, footnote} for each non-empty row in a spreadsheet.
- Comparison methods:
 - Base-CRF: textual features
 - Full-CRF: textual features and layout features

Frame Finder

Metric	Methods	title	header	data	footnote
F1	Base-CRF	0.582	0.615	0.982	0.647
	Full-CRF	0.774	0.774	0.994	0.834
Error per sheet	Base-CRF	3.534	2.348	6.526	4.208
	Full-CRF	0.872	1.316	1.528	1.208

Frame Finder

Metric	Methods	title	header	data	footnote
F1	Base-CRF	0.582	0.615	0.982	0.647
	Full-CRF	0.774	0.774	0.994	0.834
Error per sheet	Base-CRF	3.534	2.348	6.526	4.208
	Full-CRF	0.872	1.316	1.528	1.208

A diagram below the table shows two red boxes containing the values 1.316 and 1.528. Arrows from these two boxes point to a third red box below them containing the value 2.844, indicating that 2.844 is the sum of 1.316 and 1.528.

Hierarchy Extractor

- The hierarchy extractor detects all the annotation attribute pairs in an attribute region.
- Comparison methods:
 - Human
 - SVM: classification method
 - EN-SVM: tree-enforced classification method

Hierarchy Extractor

Metric	Methods	Top	Left
F1	SVM	0.919	0.769
	EN-SVM	0.920	0.811

Metric	Methods	Top	Left
Error per sheet	Human	22.469	58.598
	SVM	1.834	19.554
	EN-SVM	1.829	16.154

Outline

1. Introduction
2. System Framework
3. Experiments
4. Conclusion

Conclusion

- We present a novel system to extract spreadsheet relational data automatically, which makes it possible for downstream integration applications.
- The system parses spreadsheets to detect different semantic regions, recognizes the implicit hierarchical structures of the attributes and then constructs relational tuples.

Questions?

