COLLEGE OF ENGINEERING
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

Integrating Spreadsheet Data via Accurate and Low-Effort Extraction

Zhe (Shirley) Chen
Michael Cafarella

University of Michigan

Spreadsheets are Everywhere

- Microsoft estimates *more than 400 million* people using Excel worldwide. [1]
- The U.S. government published thousands of spreadsheets about a variety of topics.

[1] http://www.cutimes.com/2013/07/31/rethinking-spreadsheets-and-performance-management?ref=hp

What is the strength of the connection between smoking and lung cancer for the 50 U.S. states ?

Smoking Spreadsheet From census.gov

Table 200. Current Cigarette Smoking by Sex and State: 2007

See notes.

State	2007 Total	Male	Female
United States, total \1	**19.8**	**21.3**	**18.4**
Alabama	22.5	25.7	19.7
Alaska	22.2	24.6	19.7
Arizona	19.8	23.4	16.3
Arkansas	22.4	24.8	20.2
California	14.3	18.1	10.6
Colorado	18.7	19.7	17.7
Connecticut	15.5	16.6	14.5
Delaware	19.0	17.6	20.3
District of Columbia	17.3	19.1	15.7
Florida	19.3	21.3	17.5
Georgia	19.3	21.2	17.5
Hawaii	17.0	19.8	14.3
Idaho	19.2	20.9	17.4
Illinois	20.2	22.1	18.4
Indiana	24.1	25.9	22.4
Iowa	19.8	21.4	18.3
Kansas	17.9	18.7	17.1
Kentucky	28.3	28.8	27.8
Louisiana	22.6	26.4	19.1

Lung Cancer Spreadsheet From census.gov

Table 177. Cancer--Estimated New Cases and Deaths by State: 2009

See notes.

State	New cases \1 Total \2	Female breast	Lung & bronchus	Deaths Total \2
United States	**1,479.4**	**192.4**	**219.4**	**562.3**
Alabama	24.1	3.0	4.0	9.9
Alaska	2.5	0.4	0.4	0.8
Arizona	27.6	3.5	4.0	10.3
Arkansas	14.8	1.8	2.6	6.2
California	152.2	21.7	17.9	54.6
Colorado	20.3	2.8	2.2	6.7
Connecticut	20.7	2.8	2.7	7.0
Delaware	4.7	0.6	0.8	1.9
District of Columbia	2.6	0.3	0.4	1.0
Florida	102.2	12.7	17.8	41.3
Georgia	39.1	5.4	6.2	15.0
Hawaii	6.4	0.9	0.7	2.3
Idaho	6.8	0.8	0.8	2.5
Illinois	61.0	7.6	9.2	23.2
Indiana	31.3	3.7	5.4	12.8
Iowa	16.7	2.1	2.6	6.4
Kansas	13.1	1.8	2.1	5.3
Kentucky	24.1	2.8	4.7	9.4
Louisiana	22.2	2.7	3.7	8.8
Maine	9.0	1.1	1.4	3.2
Maryland	26.7	3.7	4.1	10.3

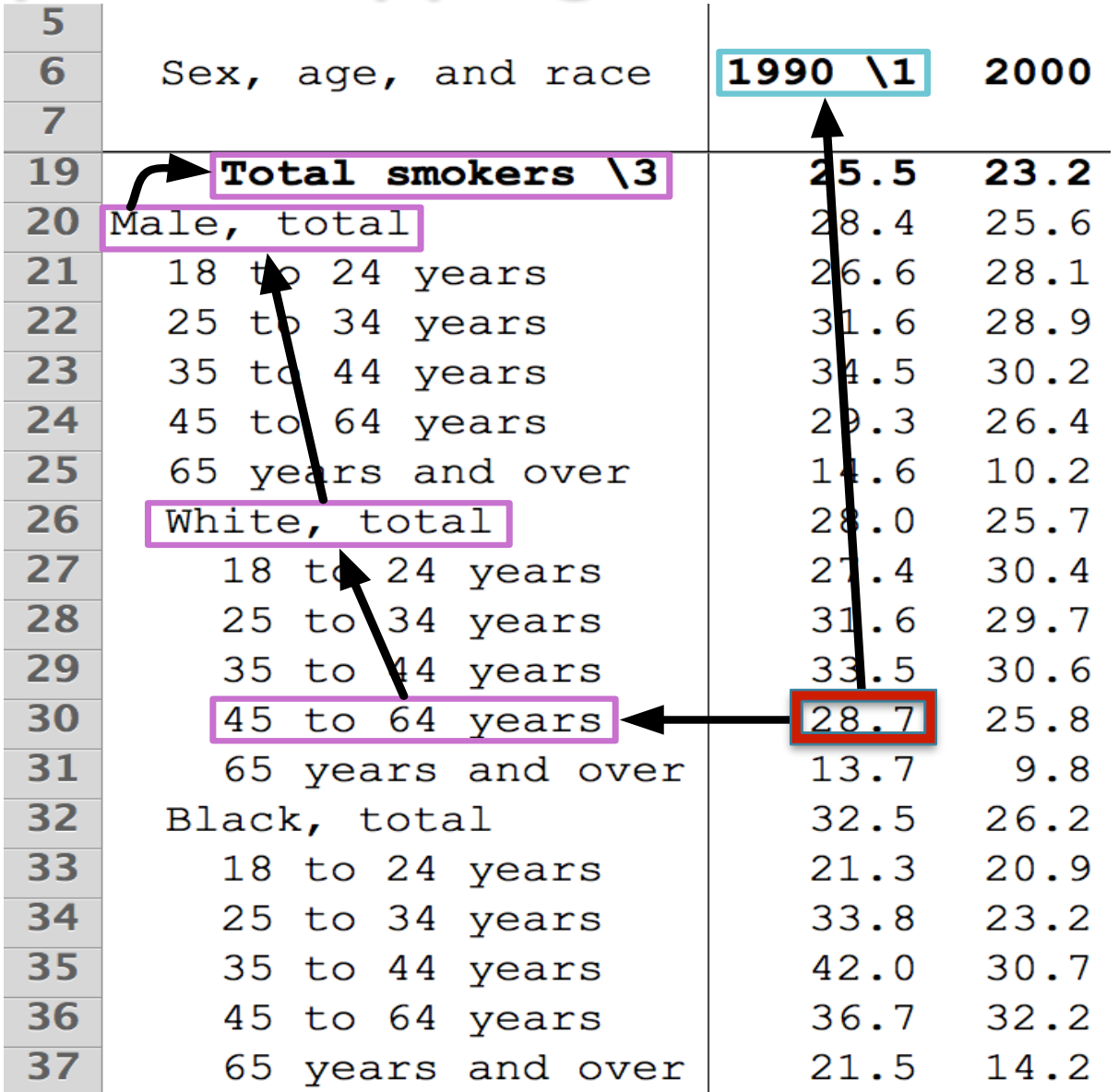
OUR GOAL: Integration

- Spreadsheet integration applications are appealing.
- If we can convert spreadsheets into *relational format*, the integration will be made possible.
- But spreadsheet schema is entirely *implicit*.

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

Implicit Mapping Structures



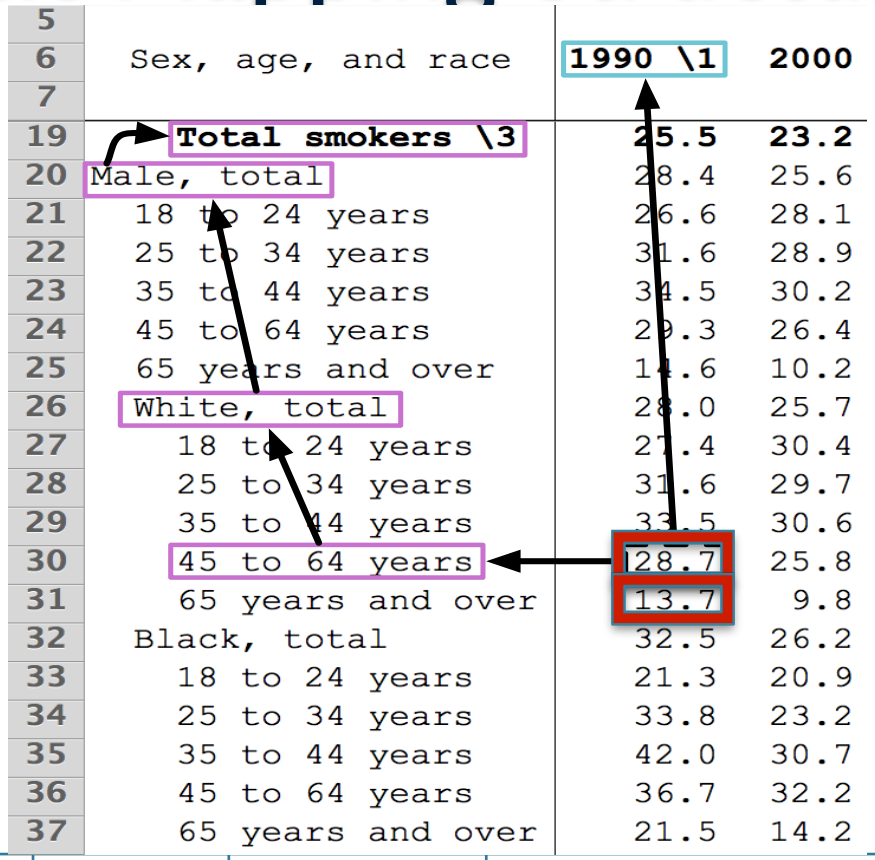
	Sex, age, and race	1990 \1	2000
19	**Total smokers \3**	**25.5**	**23.2**
20	Male, total	28.4	25.6
21	18 to 24 years	26.6	28.1
22	25 to 34 years	31.6	28.9
23	35 to 44 years	34.5	30.2
24	45 to 64 years	29.3	26.4
25	65 years and over	14.6	10.2
26	White, total	28.0	25.7
27	18 to 24 years	27.4	30.4
28	25 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

Implicit Mapping Structures

	Sex, age, and race	1990 \1	2000
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

1990	Male	White	45 to 64 years	28.7

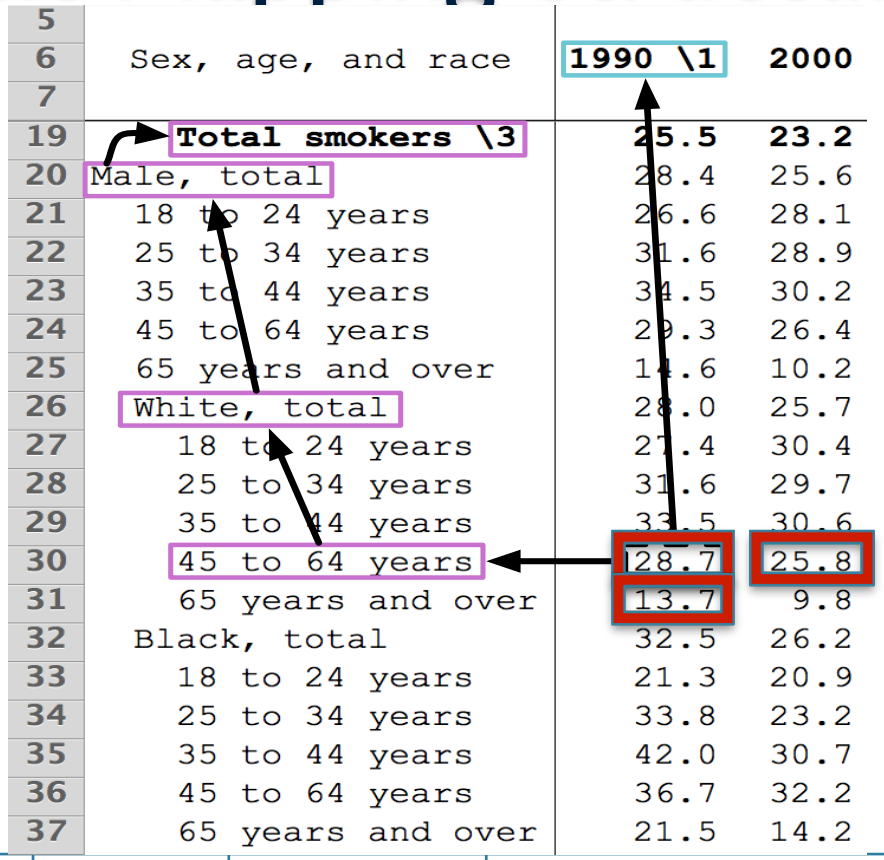
Implicit Mapping Structures



Sex, age, and race	1990 \1	2000
Total smokers \3	**25.5**	**23.2**
Male, total	28.4	25.6
18 to 24 years	26.6	28.1
25 to 34 years	31.6	28.9
35 to 44 years	34.5	30.2
45 to 64 years	29.3	26.4
65 years and over	14.6	10.2
White, total	28.0	25.7
18 to 24 years	27.4	30.4
25 to 34 years	31.6	29.7
35 to 44 years	33.5	30.6
45 to 64 years	28.7	25.8
65 years and over	13.7	9.8
Black, total	32.5	26.2
18 to 24 years	21.3	20.9
25 to 34 years	33.8	23.2
35 to 44 years	42.0	30.7
45 to 64 years	36.7	32.2
65 years and over	21.5	14.2

1990	Male	White	45 to 64 years	28.7
1990	Male	White	65 years and over	13.7

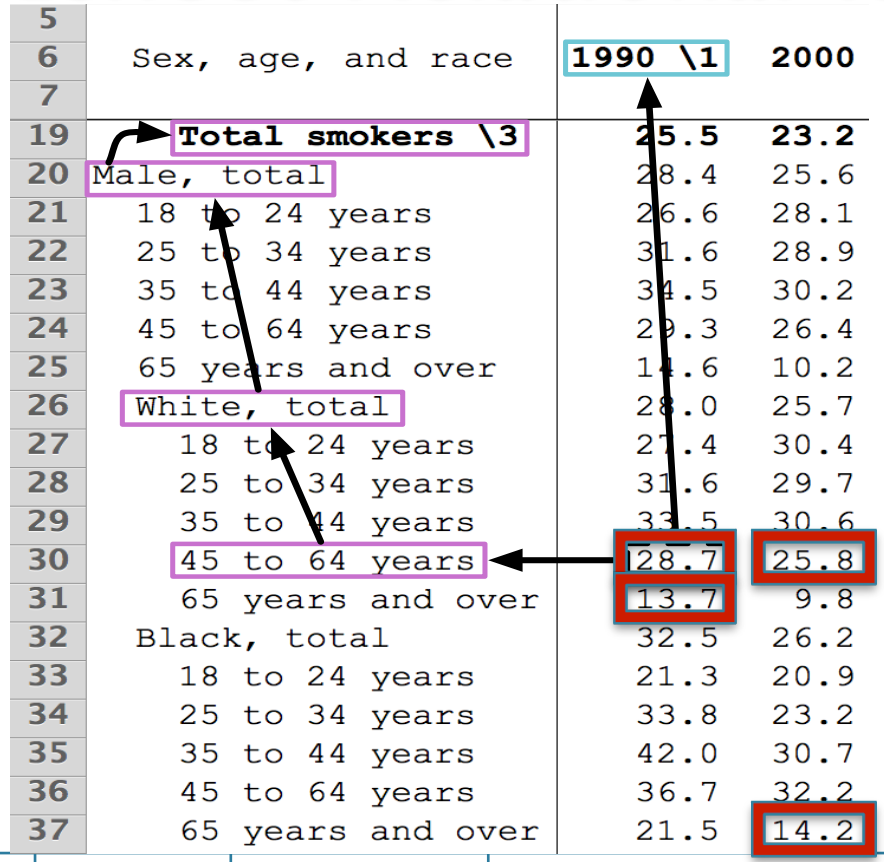
Implicit Mapping Structures



Sex, age, and race	1990 \1	2000
Total smokers \3	**25.5**	**23.2**
Male, total	28.4	25.6
18 to 24 years	26.6	28.1
25 to 34 years	31.6	28.9
35 to 44 years	34.5	30.2
45 to 64 years	29.3	26.4
65 years and over	14.6	10.2
White, total	28.0	25.7
18 to 24 years	27.4	30.4
25 to 34 years	31.6	29.7
35 to 44 years	33.5	30.6
45 to 64 years	28.7	25.8
65 years and over	13.7	9.8
Black, total	32.5	26.2
18 to 24 years	21.3	20.9
25 to 34 years	33.8	23.2
35 to 44 years	42.0	30.7
45 to 64 years	36.7	32.2
65 years and over	21.5	14.2

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

Spreadsheet Relational Table



Sex, age, and race	1990 \1	2000
Total smokers \3	25.5	23.2
Male, total	28.4	25.6
18 to 24 years	26.6	28.1
25 to 34 years	31.6	28.9
35 to 44 years	34.5	30.2
45 to 64 years	29.3	26.4
65 years and over	14.6	10.2
White, total	28.0	25.7
18 to 24 years	27.4	30.4
25 to 34 years	31.6	29.7
35 to 44 years	33.5	30.6
45 to 64 years	28.7	25.8
65 years and over	13.7	9.8
Black, total	32.5	26.2
18 to 24 years	21.3	20.9
25 to 34 years	33.8	23.2
35 to 44 years	42.0	30.7
45 to 64 years	36.7	32.2
65 years and over	21.5	14.2

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

Hierarchical Spreadsheets are Popular

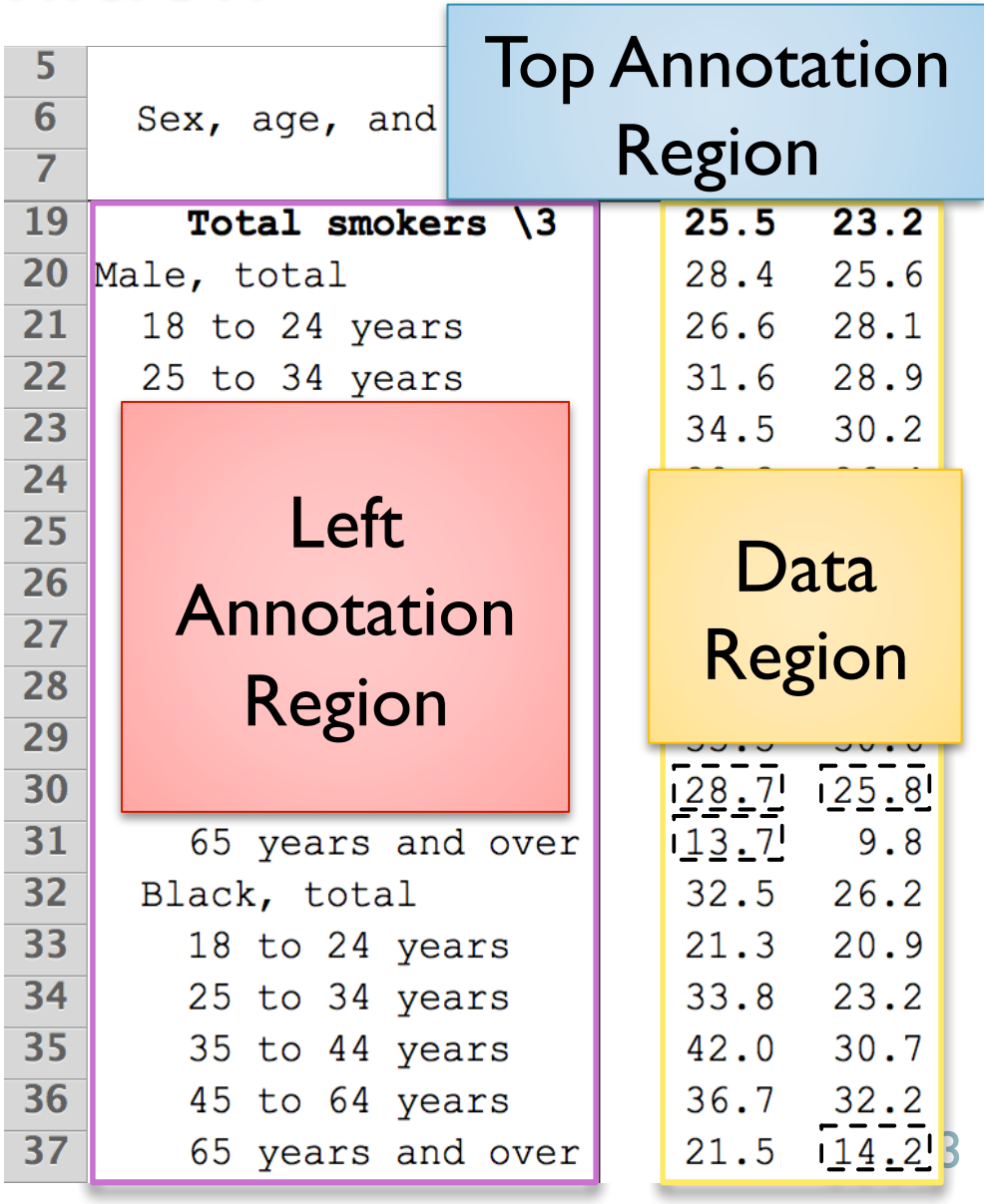
- *32.5%* of the Web spreadsheets are hierarchical
- *More than 60%* in the top 10 Internet domains that publish the largest number of spreadsheets are also hierarchical.

Outline

- Introduction
- Problem Definition
- Our Approach
- Experiments
- Conclusions

Problem Definition

- We focus on spreadsheets that consists of annotation and data regions.



Problem Definition

- We study the ***hierarchy extraction*** problem, which recovers the hierarchical structure of annotations in spreadsheets.
  - **Input**: An annotation region
  - **Output**: Hierarchies for annotations
- The **Hierarchy Extraction** problem is ***one of the most critical components*** towards building the spreadsheet converted relational table.

Hierarchy Extraction is Challenging

- Different spreadsheets have different formatting style.
- We want to obtain accurate extraction results using very little users' feedback.

We propose a novel two-phase semi-automatic approach to extract spreadsheet hierarchies accurately and with low effort.

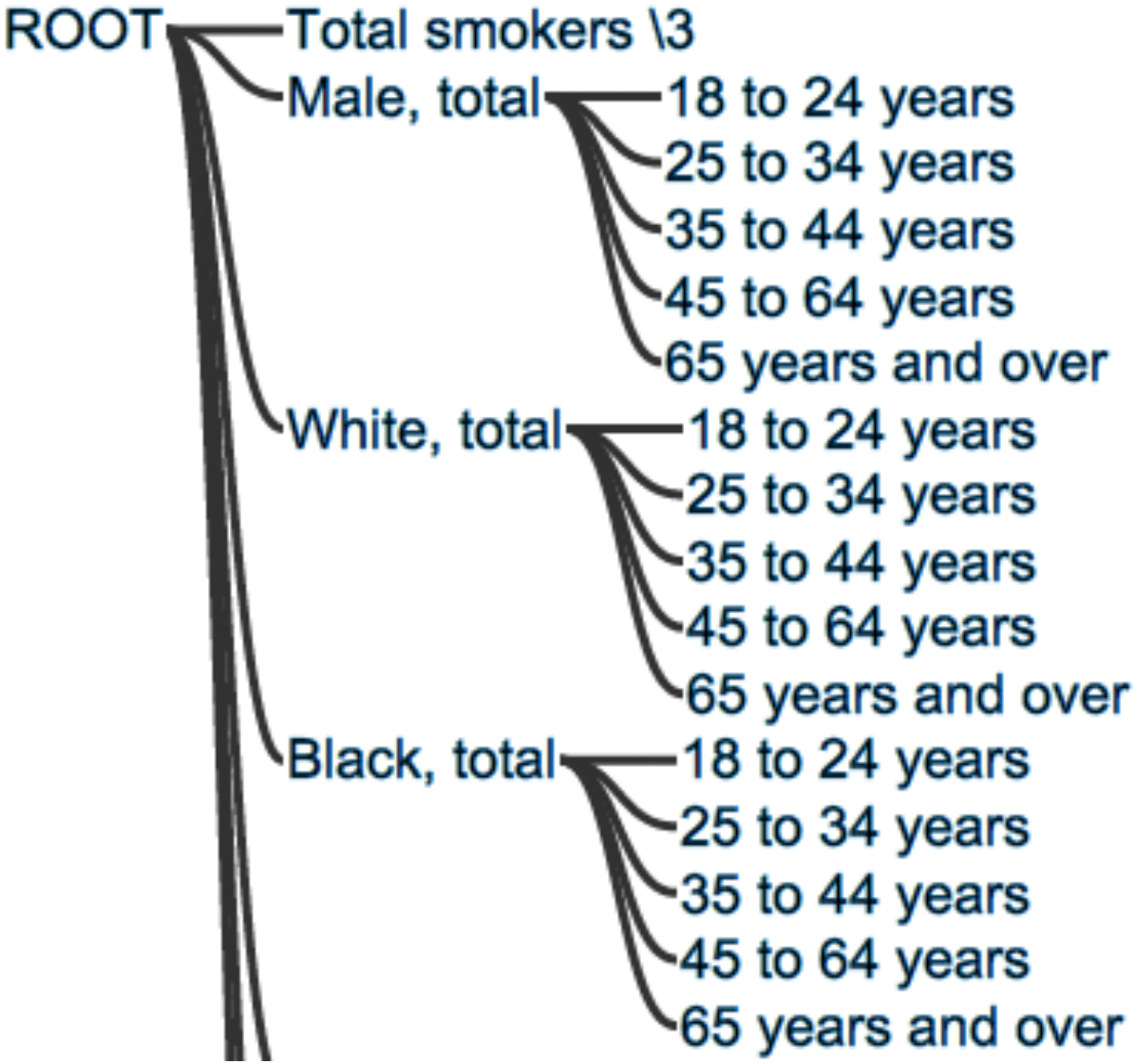
Outline

- Introduction
- Problem Definition
- Our Approach
- Experiments
- Conclusions

User Workflow

- **Phase1 --- Automatic Extraction.**
  - The system presents the initial extraction results computed by the automatic extraction.
- **Phase 2 --- Interactive Repair.**
  - A user iteratively reviews and repairs the results repair via our interactive interface.
  - The system spreads the user's repair.

Phase 1: Interface

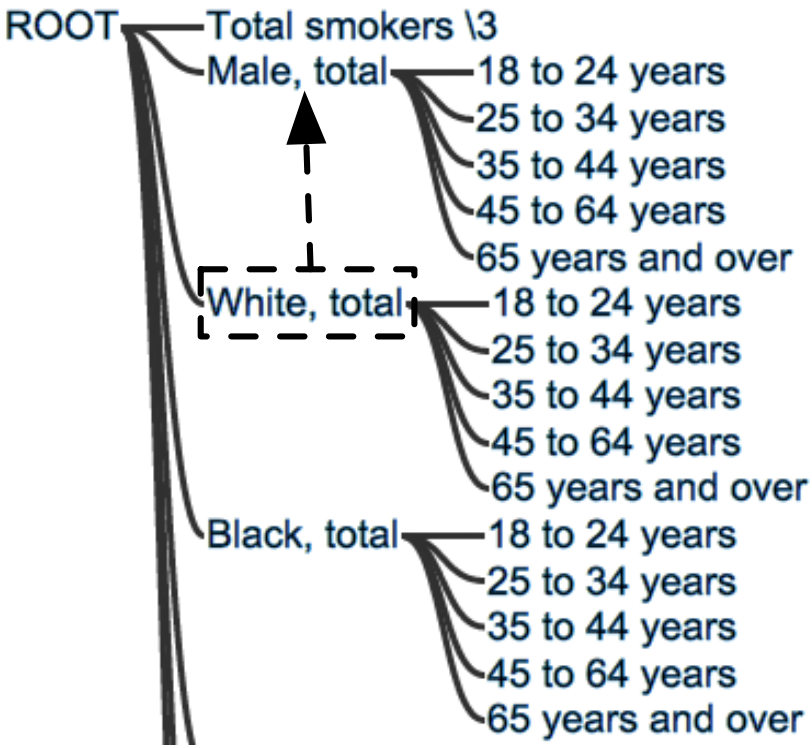


User Workflow

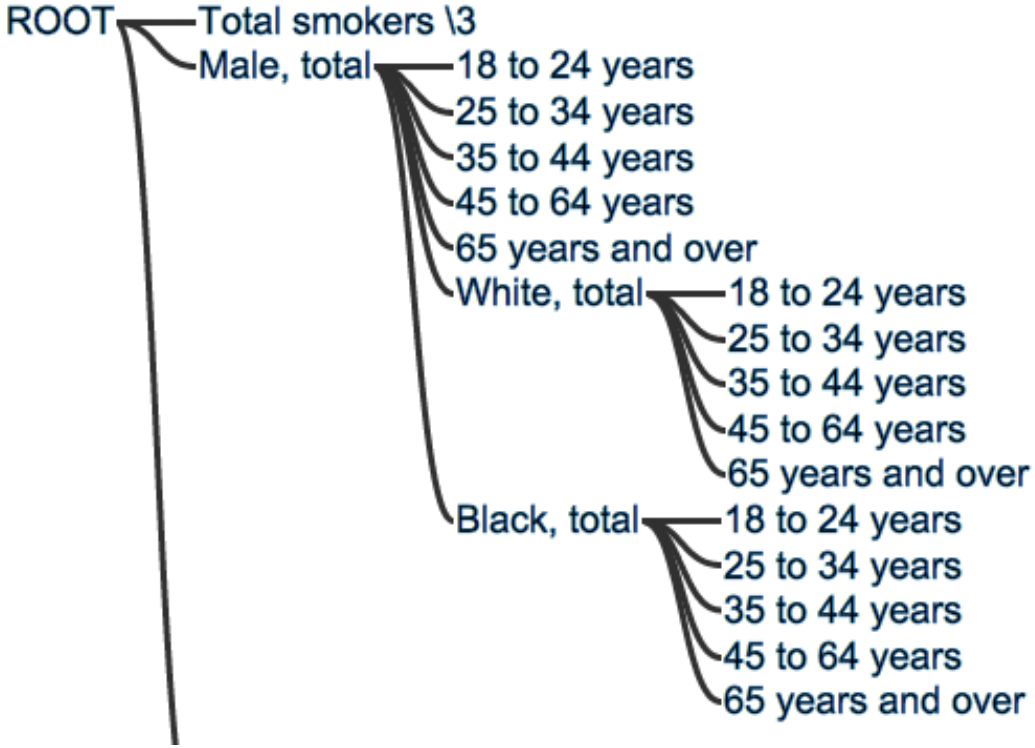
- **Phase 1 --- Automatic Extraction.**
  - The system presents the initial extraction results computed by the automatic extraction.
- **Phase 2 --- Interactive Repair.**
  - A user iteratively reviews and repairs the results repair via our interactive interface.
  - The system spreads the user's repair to other similar mistakes.

Phase 2: A Repair Operation

Before Repair: "White, total"



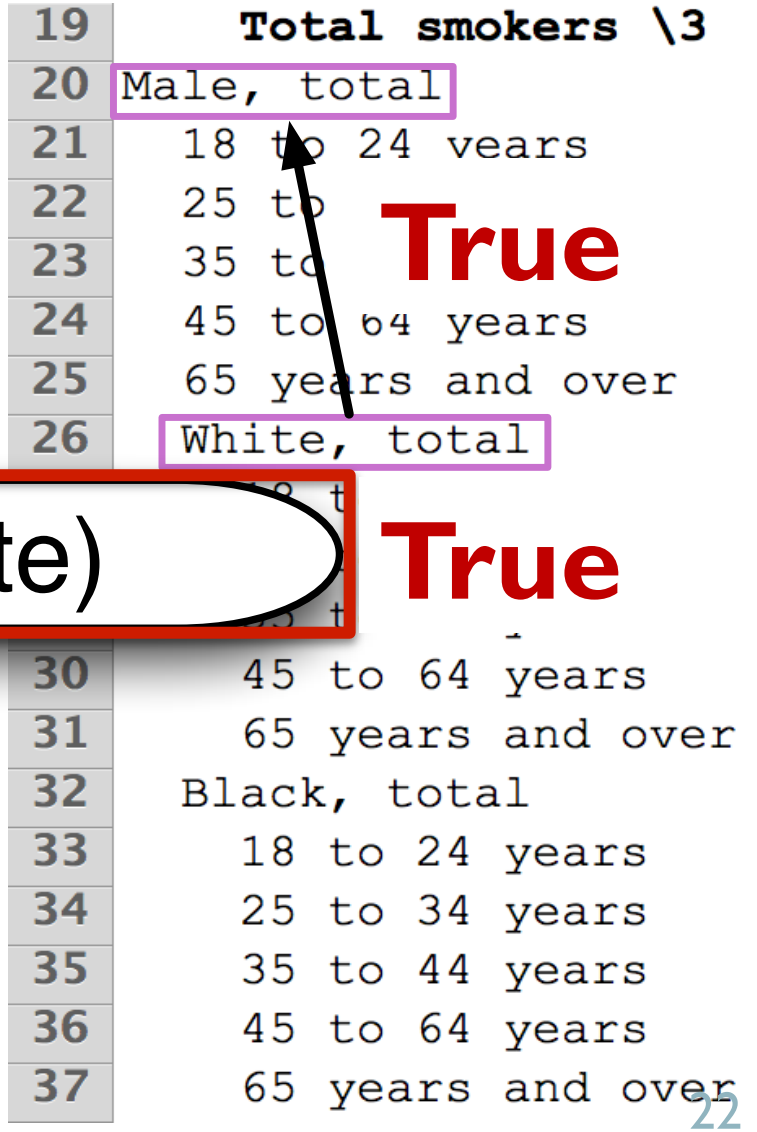
After Repair: "White, total"



An Undirected Graphical Model Based Approach

The Hierarchy Extraction Problem

- We model this hierarchy extraction problem as a classification task.
  - We enumerate all the annotations and annotation pairs as ParentChild pair candidates, and each of the candidate takes a label from {true, false}.



The Hierarchy Extraction Problem

- We model this hierarchy extraction problem as a classific
  - We enumerate all the annotation pairs in an annotation region as a ParentChild pair candidates, and each of the candidate takes a label from {true, false}.

(Black, White)

False

False

```
19      Total smokers \3
20  Male, total
21    18 to 24 years
25    65 years and over
26  White, total
27      18 to 24 years
28      25 to
29      35 to
30      45 to
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
```

Building the Graphical Model

	Sex, age, and race	1990 \1	2000
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

(Male, 18 to 24 years)

(Male, 25 to 34 years)

(18 to 24 years, Male)

(18 to 24 years, 25 to 34 years)

(25 to 34 years, Male)

(25 to 34 years, 18 to 24 years)

Building the Graphical Model

- Node potentials
- Edge potentials
- Global potentials
- Repair potentials.

(Male, 18 to 24 years)

(Male, 25 to 34 years)

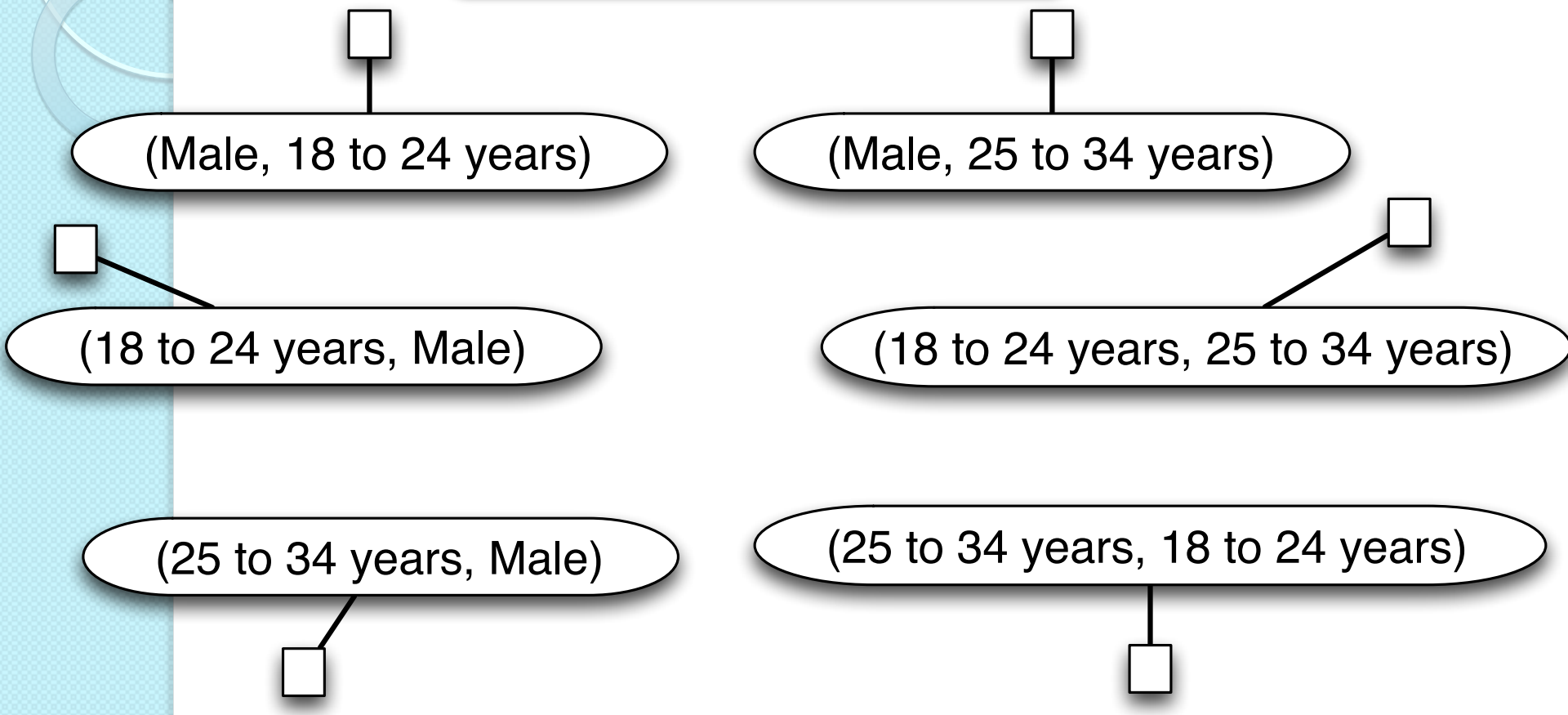
(18 to 24 years, Male)

(18 to 24 years, 25 to 34 years)

(25 to 34 years, Male)

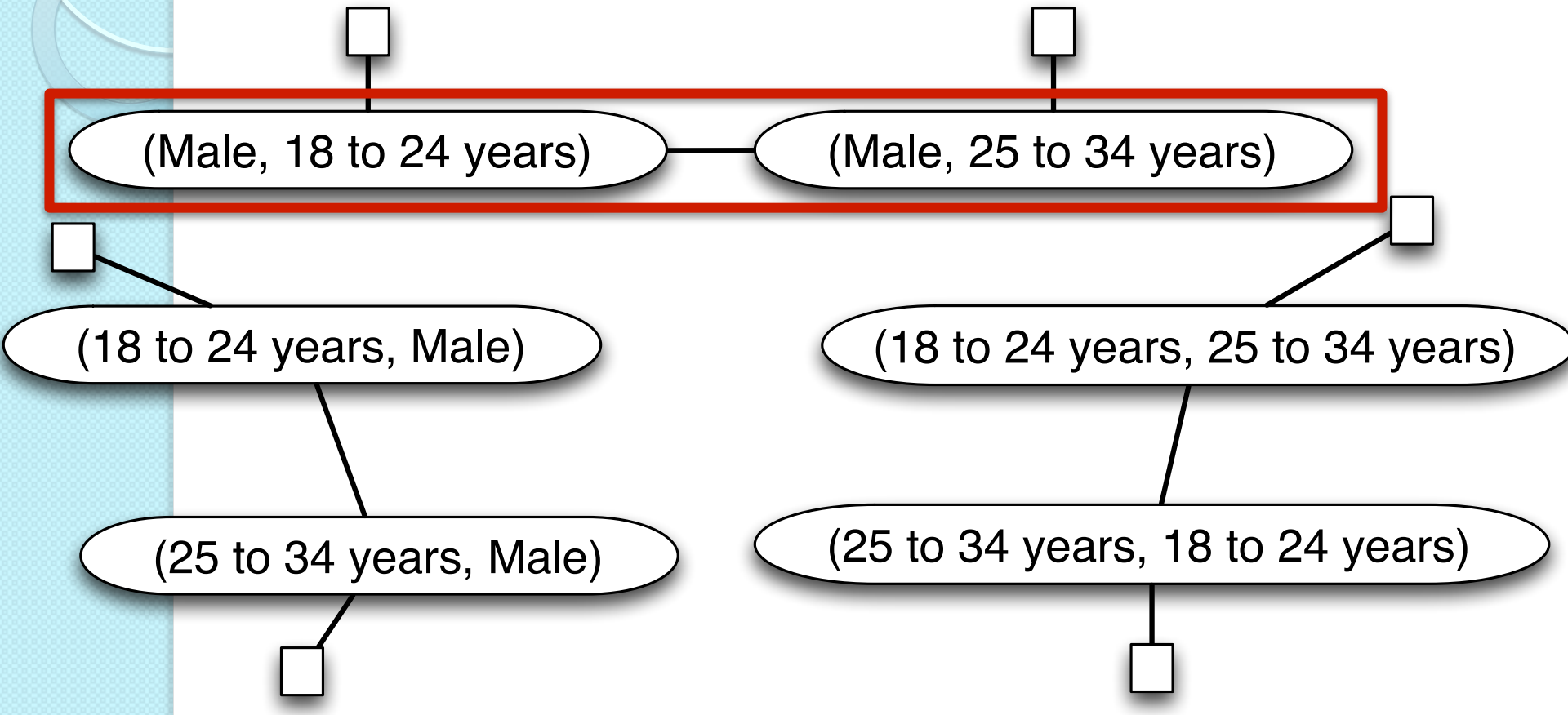
(25 to 34 years, 18 to 24 years)

Node Potentials



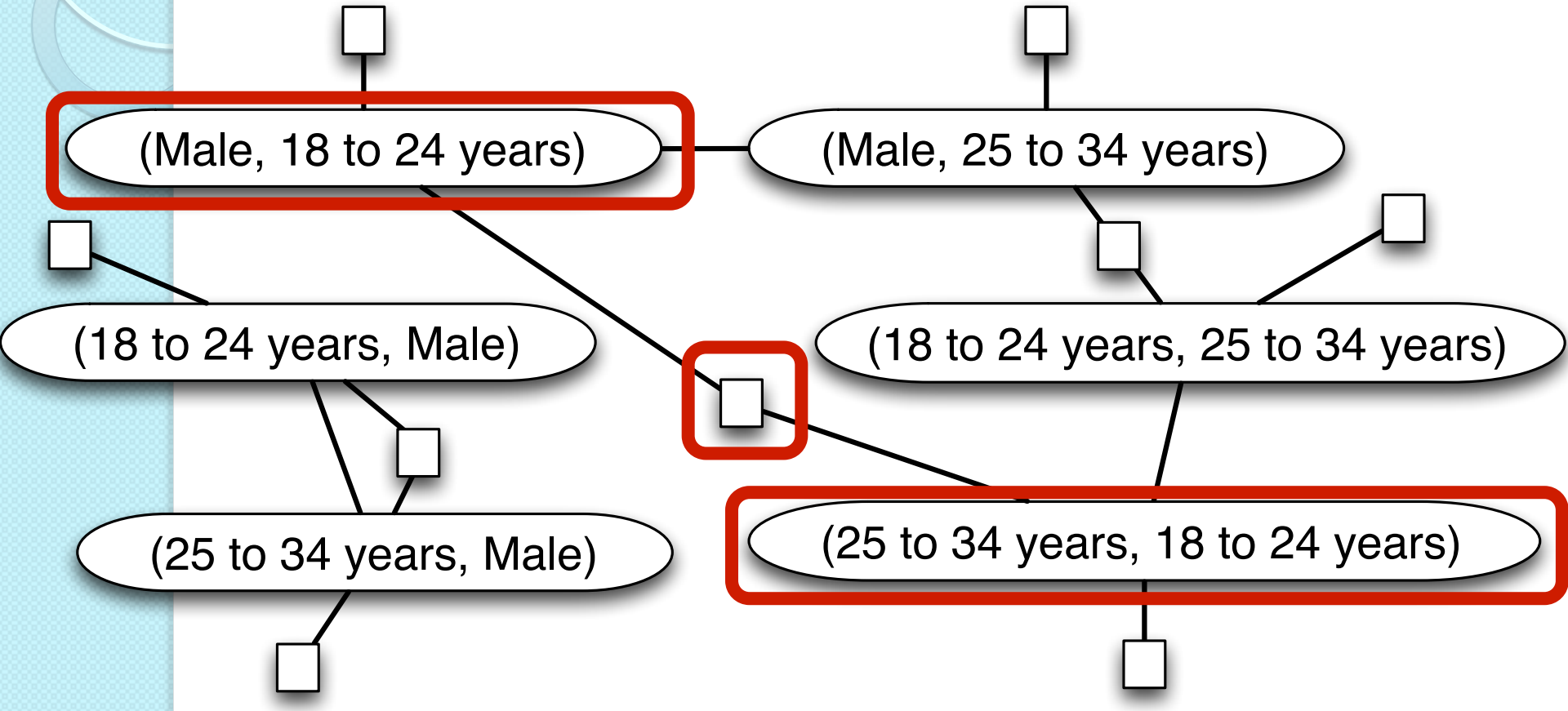
- Node potentials encode features such as indentations, alignment, and so on.

Edge Potentials



- Edge potentials encode stylistic affinity, metadata affinity, and adjacent dependency.

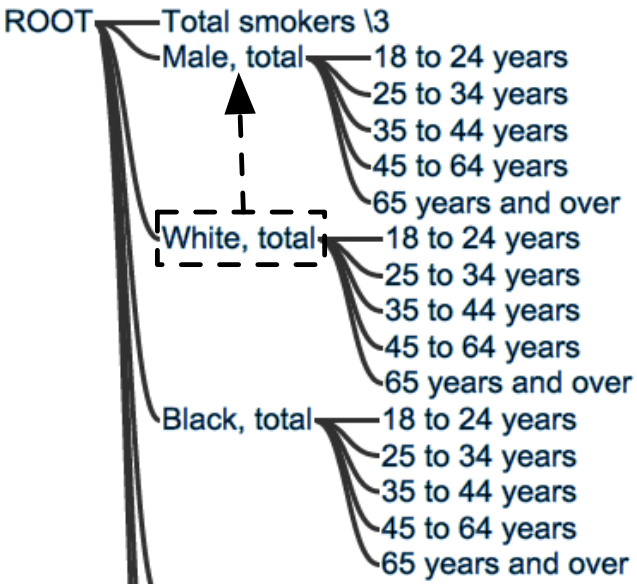
Global Potentials



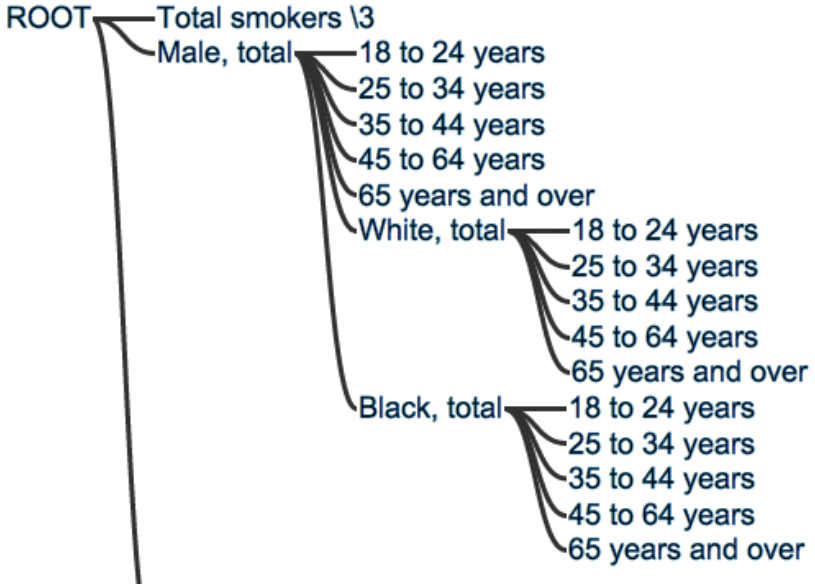
- Global potentials encode orientation constraints and one-parent constraints.

Encoding Interactive Repair

Before Repair: "White, total"



After Repair: "White, total"



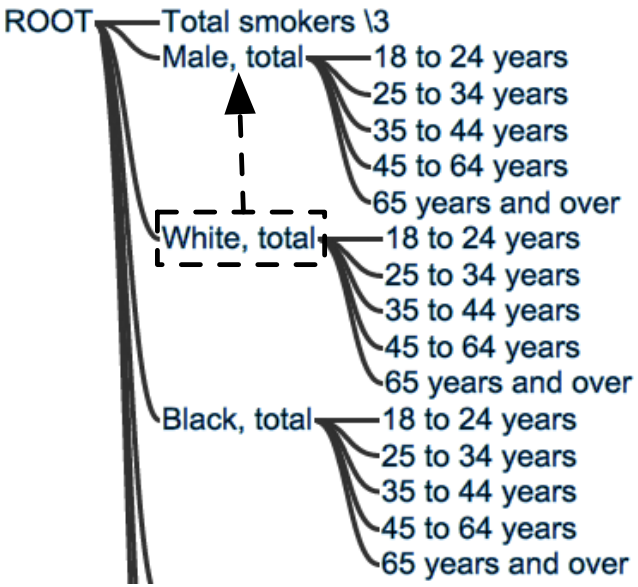
(Male, White) **True**

(Male, White) **True**

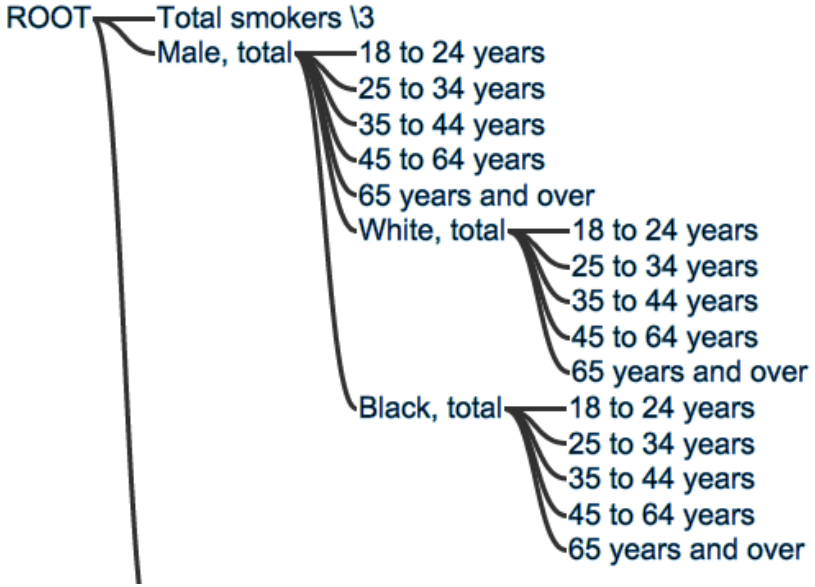
(Male, Black) **True**

Encoding Interactive Repair

Before Repair: "White, total"



After Repair: "White, total"



- The **backtracking** problem.

Repair Potentials

Outline

- Introduction
- Problem Definition
- Our Approach
- Experiments
- Conclusions

Experiment Setup

- We use two spreadsheet corpora:
  - *SAUS* --- The 2010 Statistical Abstract of the United States consists of 1369 spreadsheet files totaling 70MB.
  - *WEB* --- Our Web dataset consists of 410,554 Microsoft Excel files from 51252 distinct Internet domains, totaling 101GB.

Automatic Extraction Evaluation

- We select 200 random hierarchical spreadsheets from both SAUS and WEB.
- We split equally for training and testing, and repeat 20 times.
- Report the average *F1*.

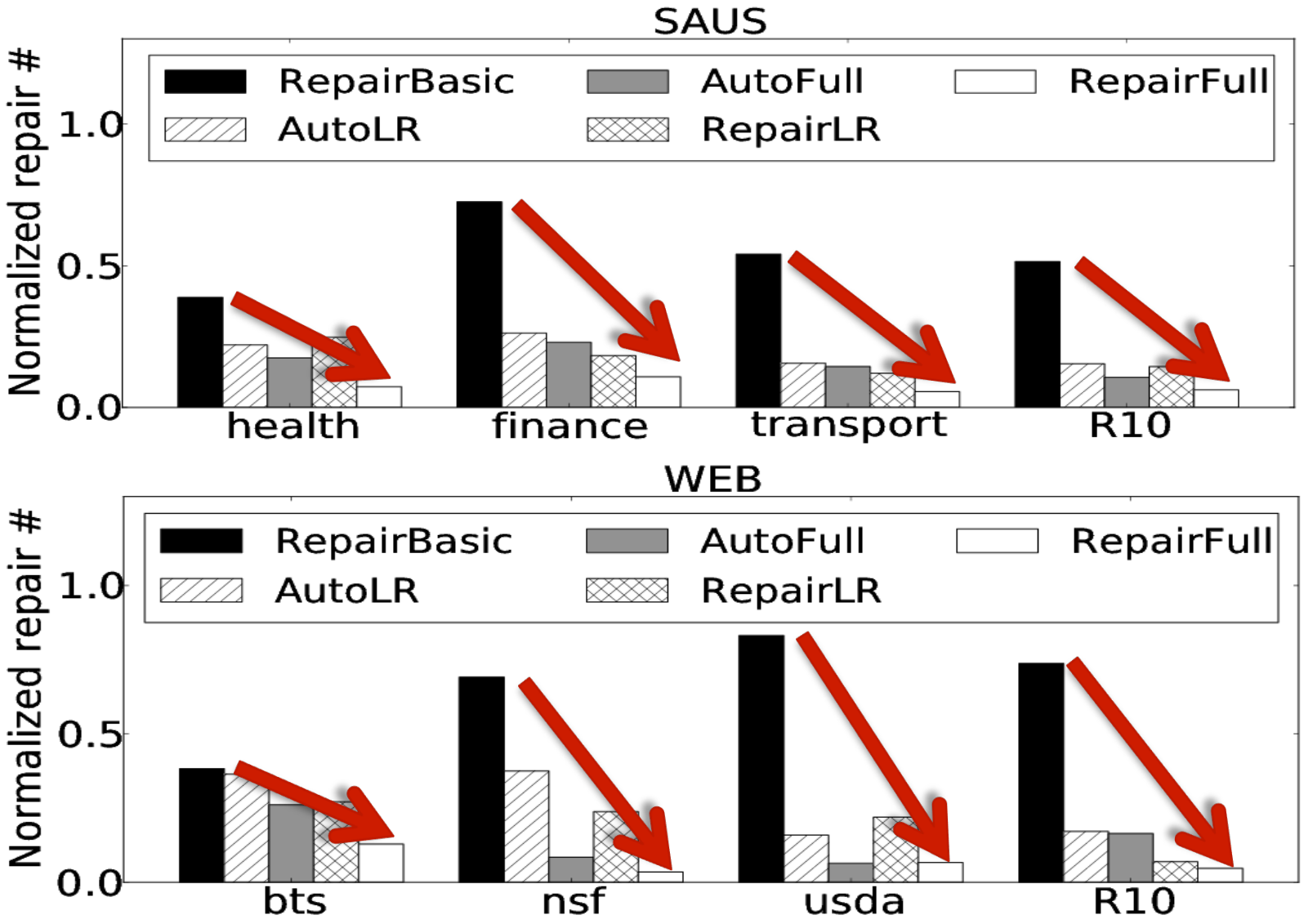
Automatic Extraction Evaluation

Dataset	Methods	F1
SAUS	AutoBasic	0.4641
	AutoLR	0.8751
	AutoEdge	0.8794
	AutoGlobal	0.8834
	AutoFull	*0.8860*
WEB	AutoBasic	0.4736
	AutoLR	0.7892
	AutoEdge	0.7973
	AutoGlobal	0.8122
	AutoFull	*0.8327*

Interactive Repair Evaluation

- We report *the required number of repair operations* to fix all the extraction errors in an annotation hierarchy.

Interactive Repair Evaluation



Outline

- Introduction
- Problem Definition
- Our Approach
- Experiments
- Conclusions

Conclusions

- A ***semi-automatic*** framework for ***extracting relational data*** from spreadsheets.

- New opportunities for ***data integration*** with spreadsheets.

Q & A