Table Identification and Information extraction in Spreadsheets

Elvis Koci[1,2], Maik Thiele[1], Oscar Romero[2], and Wolfgang Lehner[1]

[1]Technische Universität Dresden, Germany

[2]Universitat Politècnica de Catalunya, Spain

Spreadsheets in the Realm of Big Data

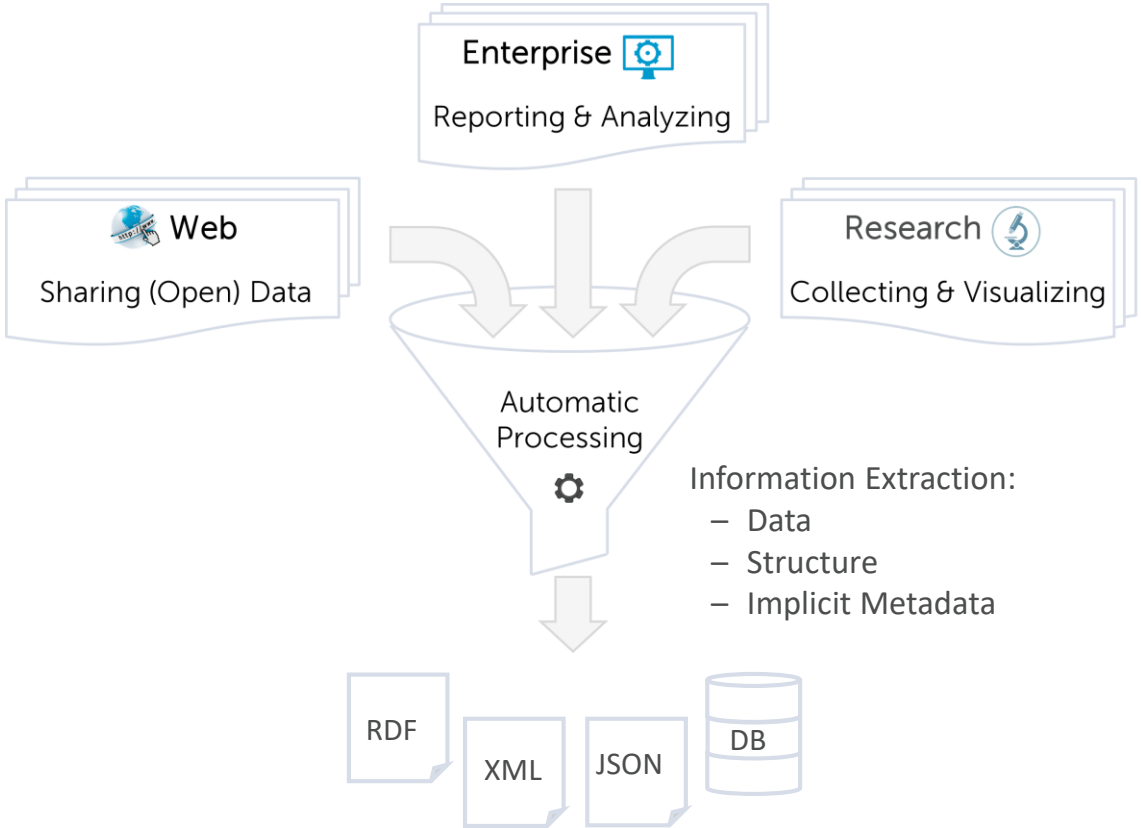
Dresden Database Systems Group

Important source of information

- Used in various domains
- Content generated by experts

Challenges

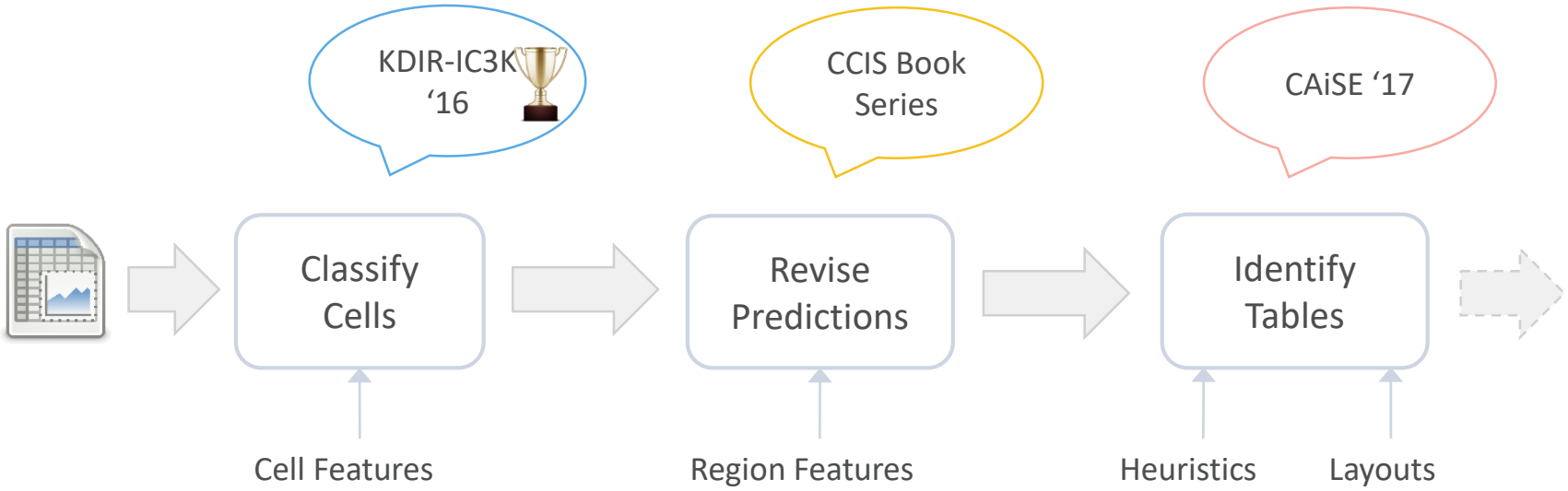
- No description of the structure
- High degree of autonomy
- Implicit Information



Proposed Solution

Dresden Database Systems Group

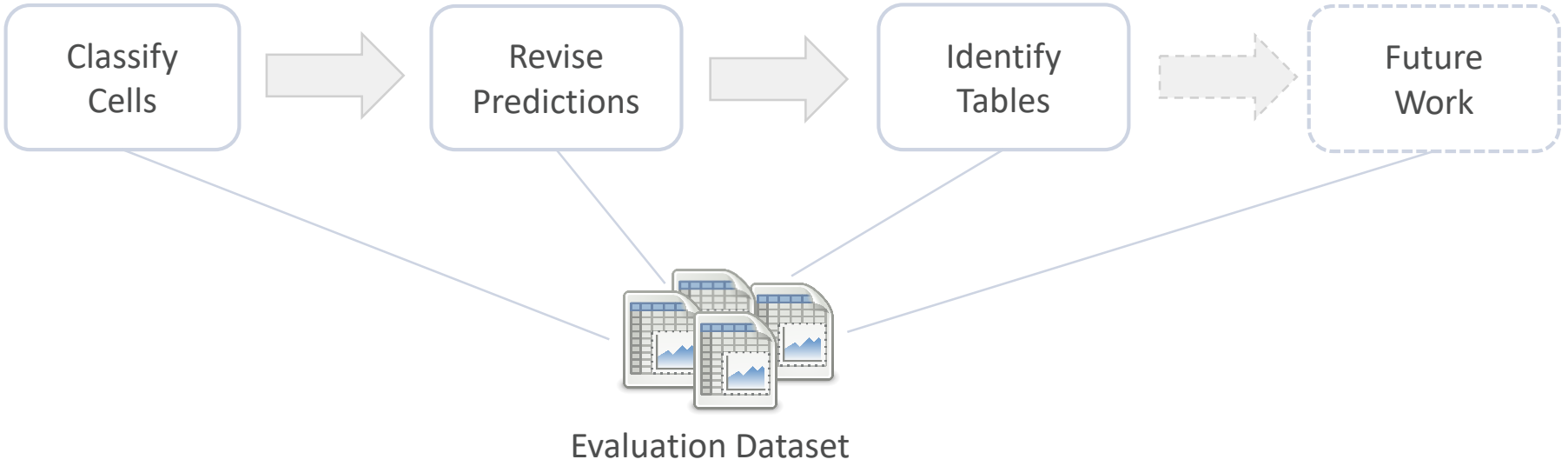
A PROCESSING PIPELINE



Proposed Solution

Dresden Database Systems Group

A PROCESSING PIPELINE



Cell Classification*

Dresden Database Systems Group

UNDERSTANDING THE DATA

- Work at the cell granularity:
  - Omit empty cells
  - Omit floating objects
- Use **5** layout categories (labels):

Header

Data

Derived

Attributes

Metadata

	A	B	C	D		E
1		**Number of Items (in Units) Sold per Region**				
2						
3		Location	Item			Total
4			Mouse	Monitor	Adapter	
5					VGA	HDMI
6		Europe				
7		Spain	500	200	85	61	846
8		France	465	169	80	80	794
9		Asia				
10		China	422	163	90	44	719
11		Vietnam	473	182	74	55	784
12		Item “Keyboard” is omitted. Check next sheet.				

* Koci et al.: A machine learning approach for layout inference in spreadsheets. KDIR IC3K’16

Cell Classification*

Dresden Database Systems Group

UNDERSTANDING THE DATA

- Work at the cell granularity:
  - Omit empty cells
  - Omit floating objects
- Use **5** layout categories (labels):

Header
Data
Derived
Attributes
Metadata

	A	B	C	D		E
1		**Number of Items (in Units) Sold per Region**				
2						
3		Location	Item			Total
4			Mouse	Monitor	Adapter	
5					VGA	HDMI
6		Europe				
7		Spain	500	200	85	61	846
8		France	465	169	80	80	794
9		Asia				
10		China	422	163	90	44	719
11		Vietnam	473	182	74	55	784
12		Item “Keyboard” is omitted. Check next sheet.				

* Koci et al.: A machine learning approach for layout inference in spreadsheets. KDIR IC3K’16

Cell Classification*

Dresden Database Systems Group

UNDERSTANDING THE DATA

- Work at the cell granularity:
  - Omit empty cells
  - Omit floating objects
- Use **5** layout categories (labels):

Header

Data

Derived

Attributes

Metadata

	A	B	C	D		E
1		**Number of Items (in Units) Sold per Region**				
2						
3		Location	Item			Total
4			Mouse	Monitor	Adapter	
5					VGA	HDMI
6		Europe				
7		Spain	500	200	85	61	846
8		France	465	169	80	80	794
9		Asia				
10		China	422	163	90	44	719
11		Vietnam	473	182	74	55	784
12		Item “Keyboard” is omitted. Check next sheet.				

* Koci et al.: A machine learning approach for layout inference in spreadsheets. KDIR IC3K’16

Cell Classification*

Dresden Database Systems Group

UNDERSTANDING THE DATA

- Work at the cell granularity:
  - Omit empty cells
  - Omit floating objects
- Use **5** layout categories (labels):

Header
Data
Derived
Attributes
Metadata

	A	B	C	D		E
1		**Number of Items (in Units) Sold per Region**				
2						
3		Location	Item			Total
4			Mouse	Monitor	Adapter	
5					VGA	HDMI
6		Europe				
7		Spain	500	200	85	61	846
8		France	465	169	80	80	794
9		Asia				
10		China	422	163	90	44	719
11		Vietnam	473	182	74	55	784
12		Item “Keyboard” is omitted. Check next sheet.				

* Koci et al.: A machine learning approach for layout inference in spreadsheets. KDIR IC3K’16

Cell Classification*

Dresden Database Systems Group

UNDERSTANDING THE DATA

- Work at the cell granularity:
  - Omit empty cells
  - Omit floating objects
- Use **5** layout categories (labels):

Header

Data

Derived

Attributes

Metadata

	A	B	C	D		E
1		**Number of Items (in Units) Sold per Region**				
2						
3		Location	Item			Total
4			Mouse	Monitor	Adapter	
5					VGA	HDMI
6		Europe				
7		Spain	500	200	85	61	846
8		France	465	169	80	80	794
9		Asia				
10		China	422	163	90	44	719
11		Vietnam	473	182	74	55	784
12		Item “Keyboard” is omitted. Check next sheet.				

* Koci et al.: A machine learning approach for layout inference in spreadsheets. KDIR IC3K’16

Cell Classification*

Dresden Database Systems Group

UNDERSTANDING THE DATA

- Work at the cell granularity:
  - Omit empty cells
  - Omit floating objects
- Use **5** layout categories (labels):

Header
Data
Derived
Attributes
Metadata

	A	B	C	D		E
1		**Number of Items (in Units) Sold per Region**				
2						
3		Location	Item			Total
4			Mouse	Monitor	Adapter	
5					VGA	HDMI
6		Europe				
7		Spain	500	200	85	61	846
8		France	465	169	80	80	794
9		Asia				
10		China	422	163	90	44	719
11		Vietnam	473	182	74	55	784
12		Item “Keyboard” is omitted. Check next sheet.				

* Koci et al.: A machine learning approach for layout inference in spreadsheets. KDIR IC3K’16

Supervised Classifications of Cells

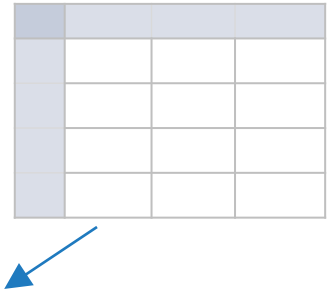
Dresden Database Systems Group

Collecting features

- Binary and numerical **>200**
- Selected **43** for evaluation

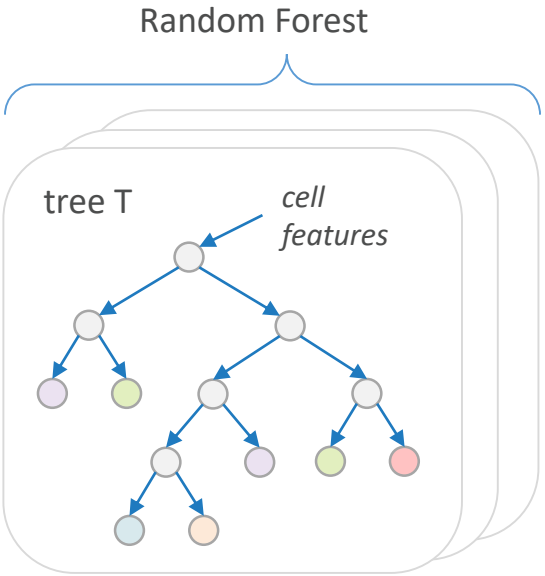
Classification

- Evaluated **5** classifiers
- Random Forest most accurate



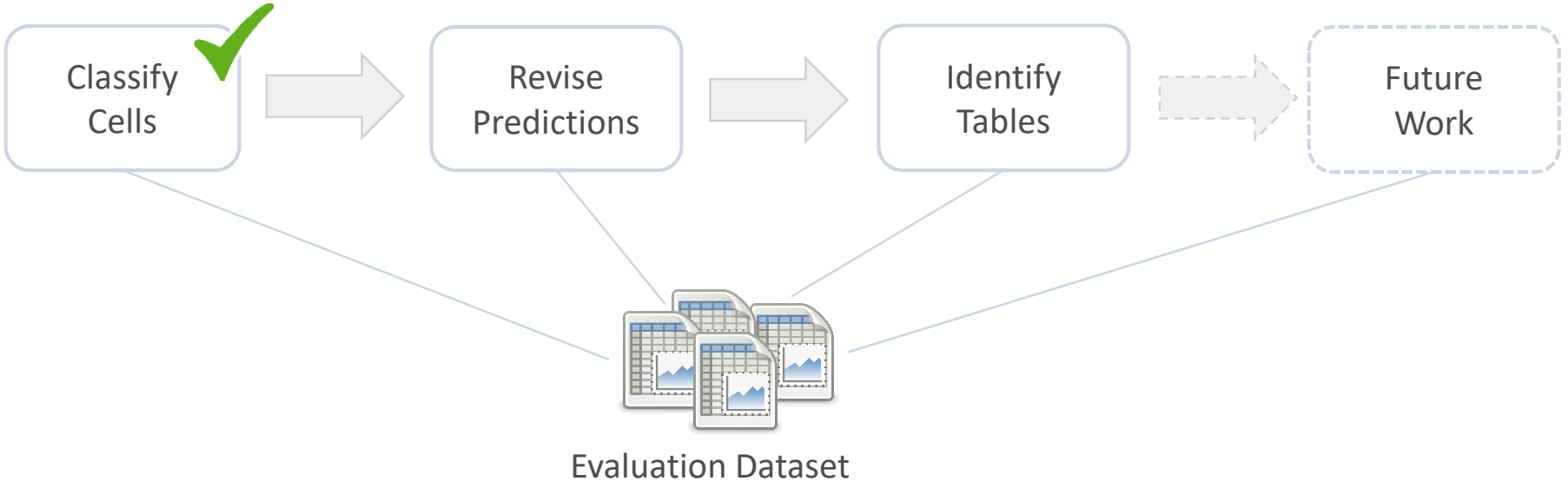
Features

– *Type String*	– *Thick Left Border*
– *Is Bold Font*	– *Not Aggregation*
– *No Underline*	– *Not Referenced*
– *Center Aligned*	– *Right is Empty*
– *Default Fill Color*	– *Top Same Style*
…	…



Best Results for Data and Header

A PROCESSING PIPELINE



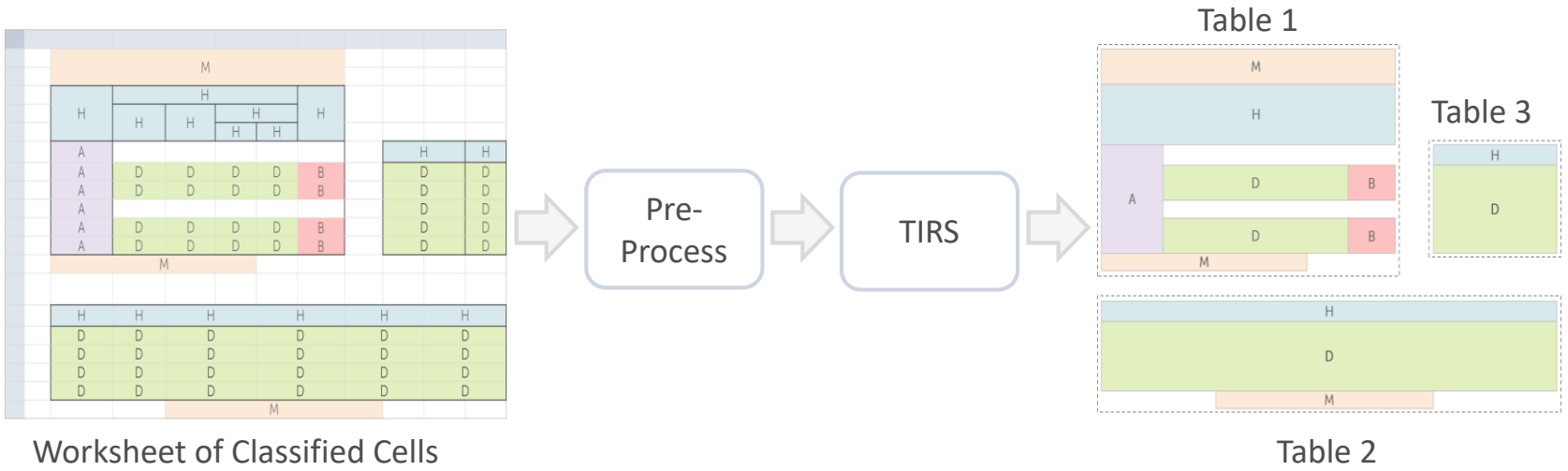
TIRS Framework*

Table identification and Reconstruction

Dresden Database Systems Group

Heuristic Procedure

- Identify tables from the results of cell classification
- A procedure of **8** heuristics tasks



* Koci et al.: Table Identification and Reconstruction in Spreadsheets. CAiSE '17

Grouping Cells

Dresden Database Systems Group

Notions

- Work with larger structures
- Decrease complexity

Label Regions (LRs)

- Adjacent Cells
- Same Label

<table>
<tr><td colspan="6">M</td></tr>
<tr><td rowspan="2">H</td><td colspan="4">H</td><td rowspan="2">H</td></tr>
<tr><td rowspan="2">H</td><td rowspan="2">H</td><td colspan="2">H</td></tr>
<tr><td>H</td><td>H</td></tr>
<tr><td>A</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>A</td><td>D</td><td>D</td><td>D</td><td>D</td><td>B</td></tr>
<tr><td>A</td><td>D</td><td>D</td><td>D</td><td>D</td><td>B</td></tr>
<tr><td>A</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>A</td><td>D</td><td>D</td><td>D</td><td>D</td><td>B</td></tr>
<tr><td>A</td><td>D</td><td>D</td><td>D</td><td>D</td><td>B</td></tr>
<tr><td colspan="4">M</td><td></td><td></td></tr>
</table>

Grouping Cells

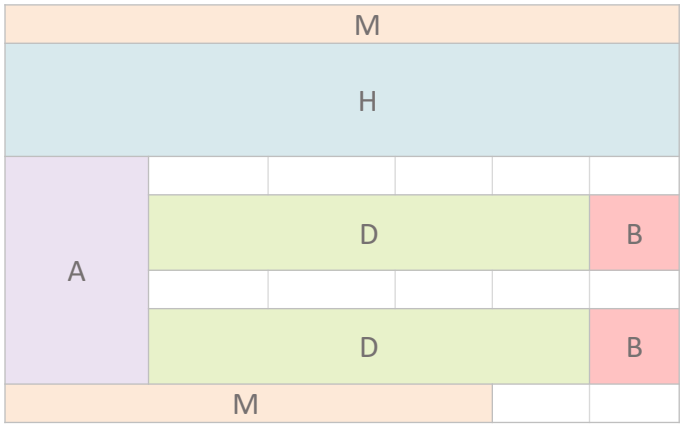
Dresden Database Systems Group

Notions

- Work with larger structures
- Decrease complexity

Label Regions (LRs)

- Adjacent Cells
- Same Label



Grouping Cells

Dresden Database Systems Group

Notions

- Work with larger structures
- Decrease complexity

Label Regions (LRs)

- Adjacent Cells
- Same Label

M					
H	H		H		H
	H	H	H	H	
A					
A	D	D	D	D	B
A	A!	D	D	D	B
A					
D!	D	D	D	D	B
A	D	D	D	D	B
M					

Grouping Cells

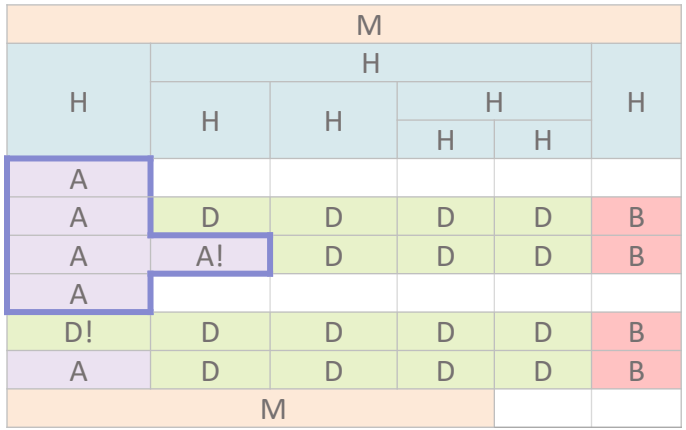
Dresden Database Systems Group

Notions

- Work with larger structures
- Decrease complexity

Label Regions (LRs)

- Adjacent Cells
- Same Label
- Rectilinear polygons



Grouping Cells

Dresden Database Systems Group

Notions

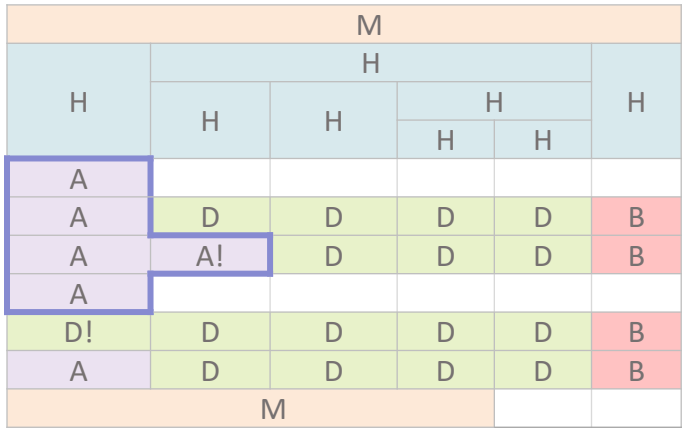
- Work with larger structures
- Decrease complexity

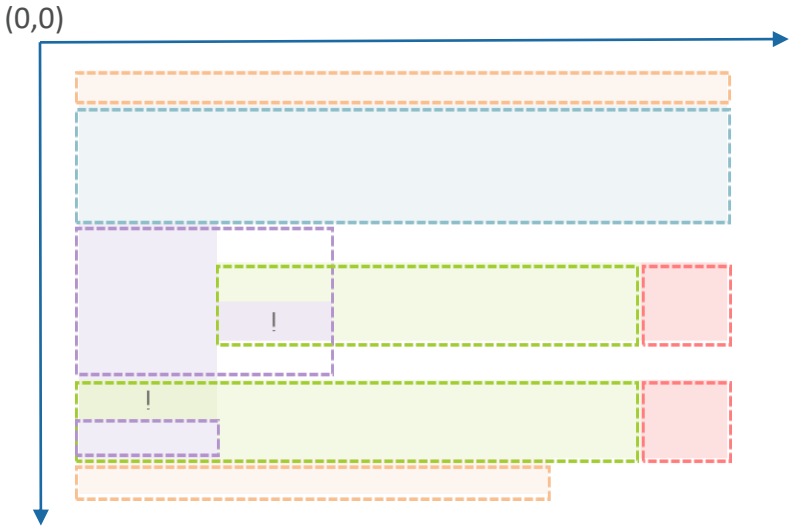
Label Regions

- Adjacent Cells
- Same Label
- Rectilinear polygons

Minimum bounding Rectangle

- An approximation for LRs
- Extensively used for spatial problems





Constructing Tables - Preliminaries

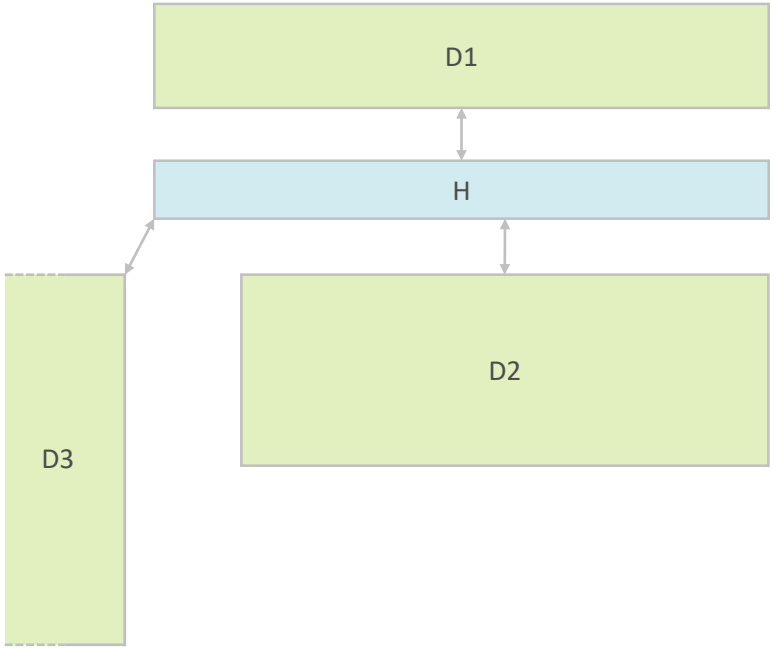
Dresden Database Systems Group

Constrains For Pairing Header with Data

1) A Header is on the top of Data
2) The distance between them is the smallest
3) Their projections overlap significantly:

$$\frac{overlap(projection(D), projection(F))}{max(projection\_length(D), projection\_length(F))} > \theta$$

- Overlap ratio measures this significance
- For the evaluation ϑ (empirically) set to **0.5**



Constructing Tables - Preliminaries

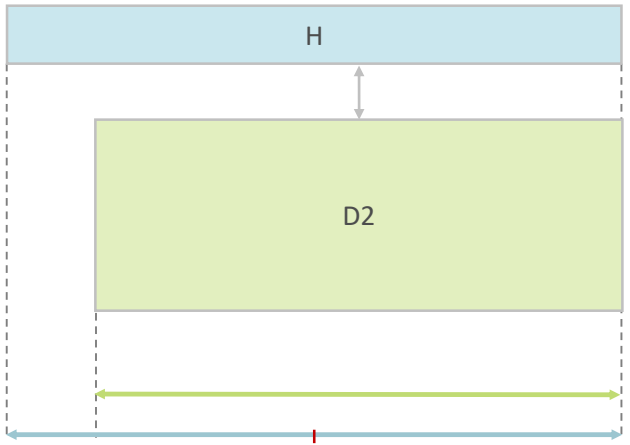
Dresden Database Systems Group

Constrains For Pairing Header with Data

1) A Header is on the top of Data
2) The distance between them is the smallest
3) Their projections overlap significantly:

$$\frac{overlap(projection(D), projection(F))}{max(projection\_length(D), projection\_length(F))} > \theta$$

- Overlap ratio measures this significance
- For the evaluation ϑ (empirically) set to **0.5**



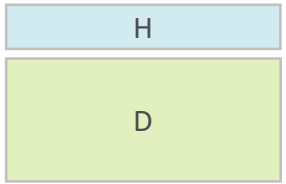
First 5 Heuristic Steps

Dresden Database Systems Group

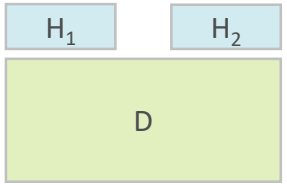
TABLE PATTERNS

- There are **5** patterns of tables
- Each corresponds to one heuristic task

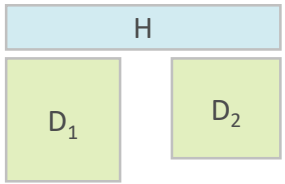
$$\frac{\sum_i^N \sum_j^M overlap(projection(H_i),\ projection(Dj))}{max\left(\sum_i^N projection\_length(Hi),\ \sum_j^M projection\_length(Dj)\right)} > \theta$$



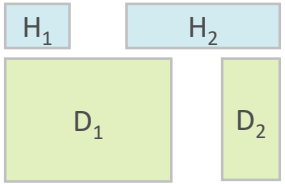
One to One



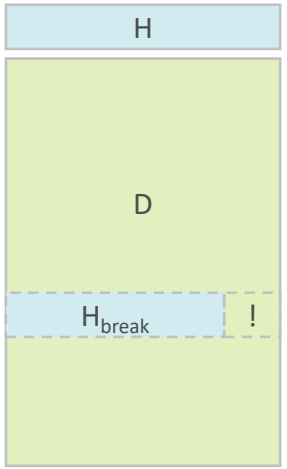
Many to One



One to Many



Many to Many



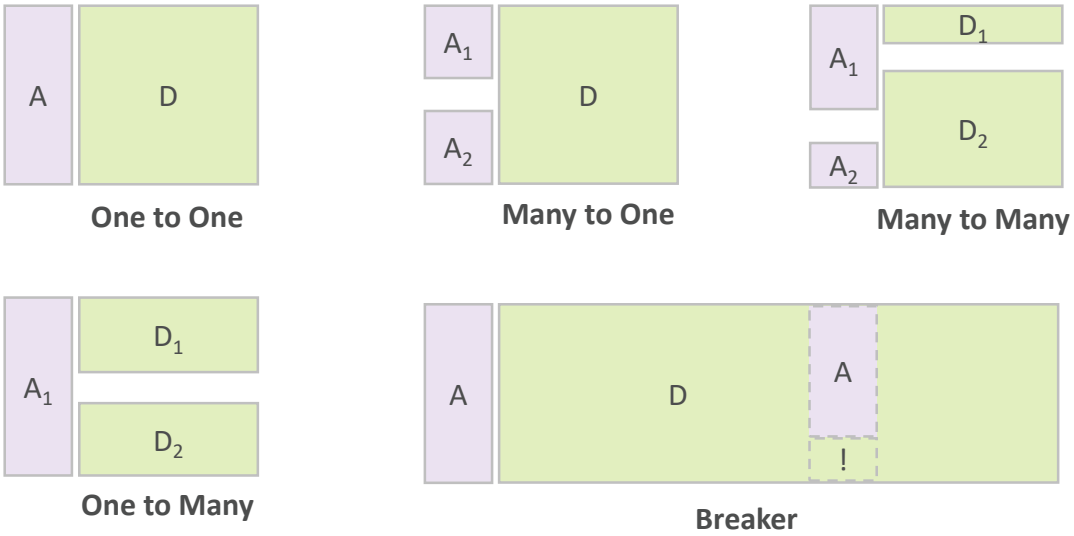
Breaker

First 5 Heuristic Steps

Dresden Database Systems Group

TABLE PATTERNS

- There are **5** patterns of tables
- Each corresponds to one heuristic task

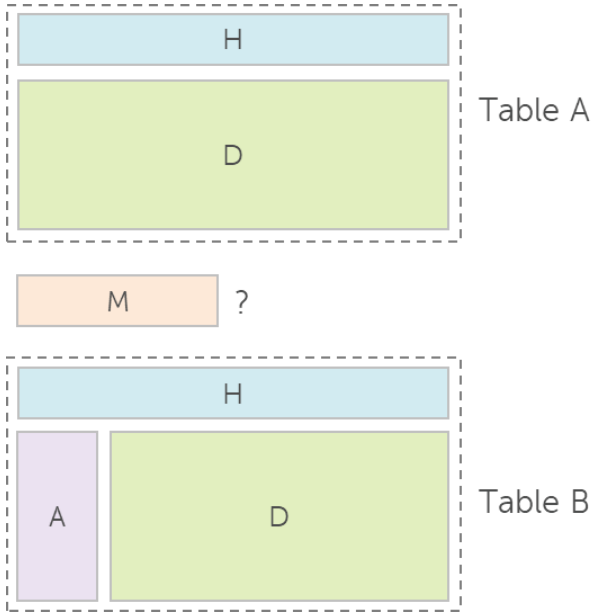


Heuristics Steps 6-8

Dresden Database Systems Group

Handle unpaired & Remove Inconsistencies

- Assign regions In-between tables

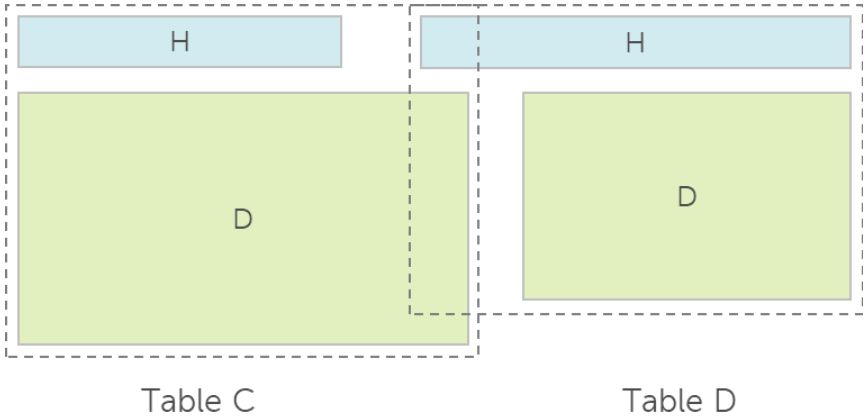


Heuristics Steps 6-8

Dresden Database Systems Group

Handle unpaired & Remove Inconsistencies

- Assign regions In-between tables
- Detect overlapping tables



Heuristics Steps 6-8

Dresden Database Systems Group

Handle unpaired & Remove Inconsistencies

- Assign regions In-between tables
- Detect overlapping tables
- Give regions to nearest table

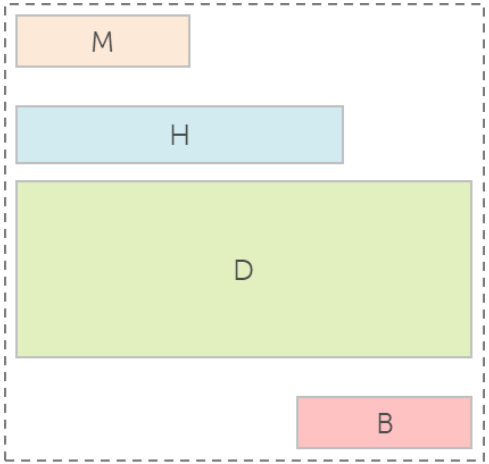


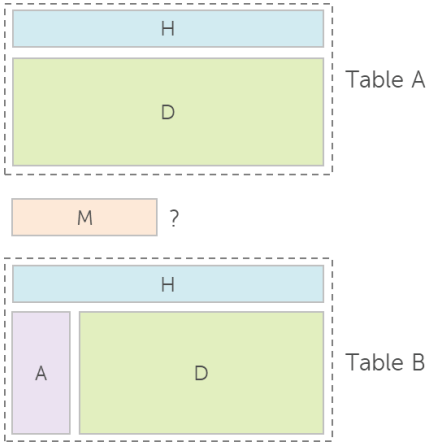
Table E

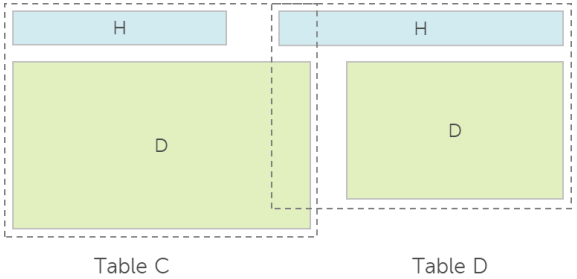
Heuristics Steps 6-8

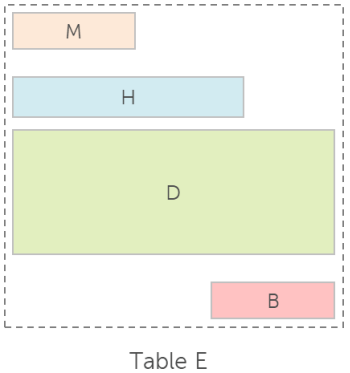
Dresden Database Systems Group

Handle unpaired & Remove Inconsistencies

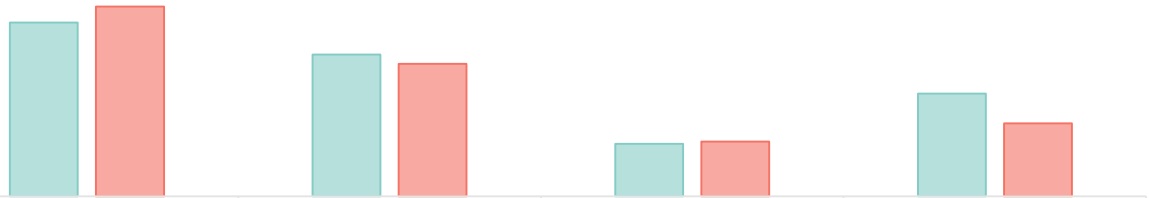
- Assign regions In-between tables
- Detect overlapping tables
- Give regions to nearest table







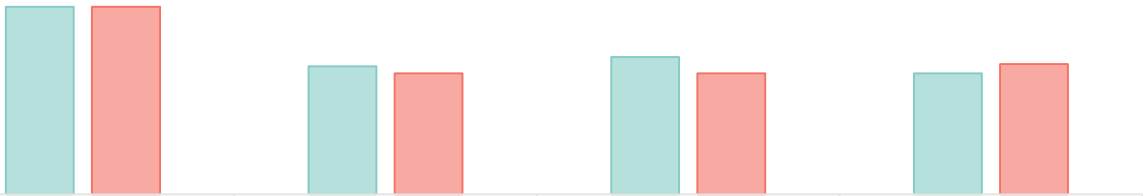
Evaluation: Table Orientations



	One Table	Vertical	Horizontal	Both
Precision	76%	62%	23%	45%
Recall	83%	58%	24%	32%

Evaluation: Number of Misclassified Cells

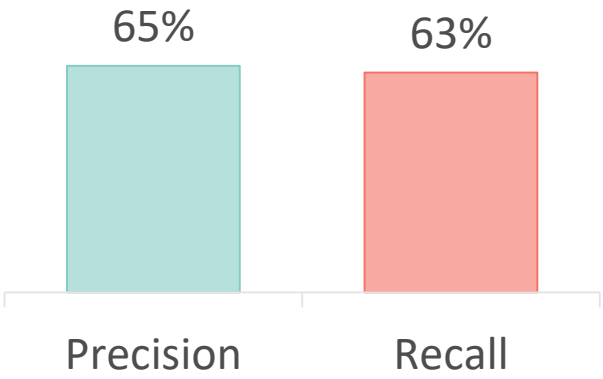
Dresden Database Systems Group



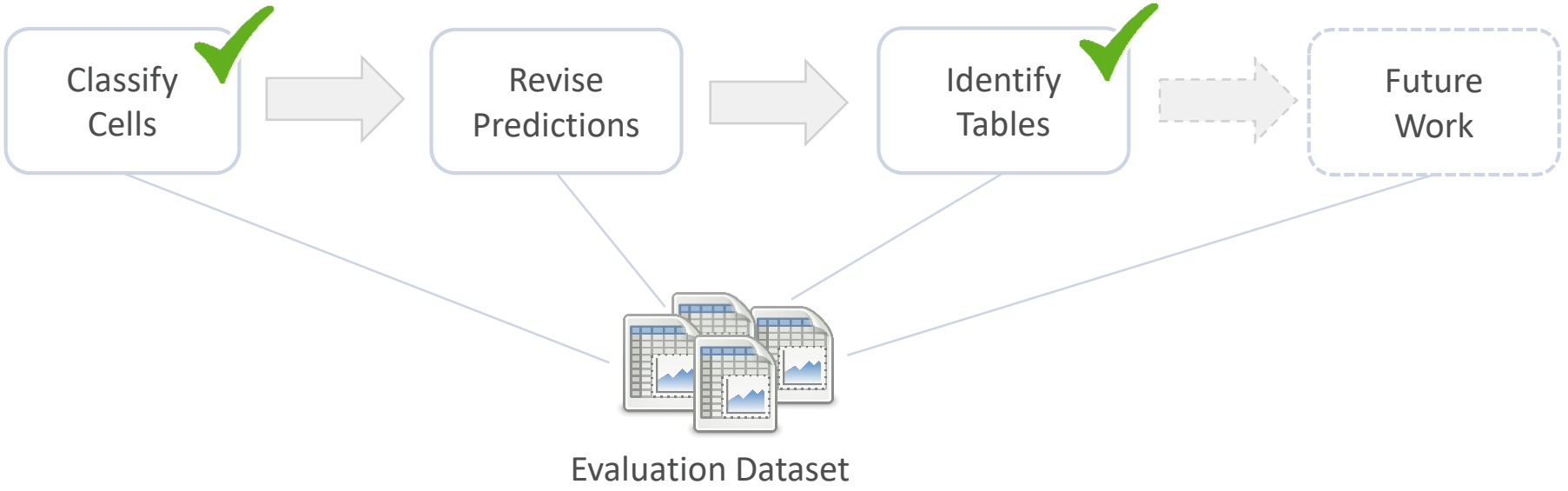
	None	Few(1-3)	Bunch(4-7)	Many(8-∞)
Precision	82%	56%	60%	53%
Recall	82%	53%	53%	57%

Overall Results

Dresden Database Systems Group

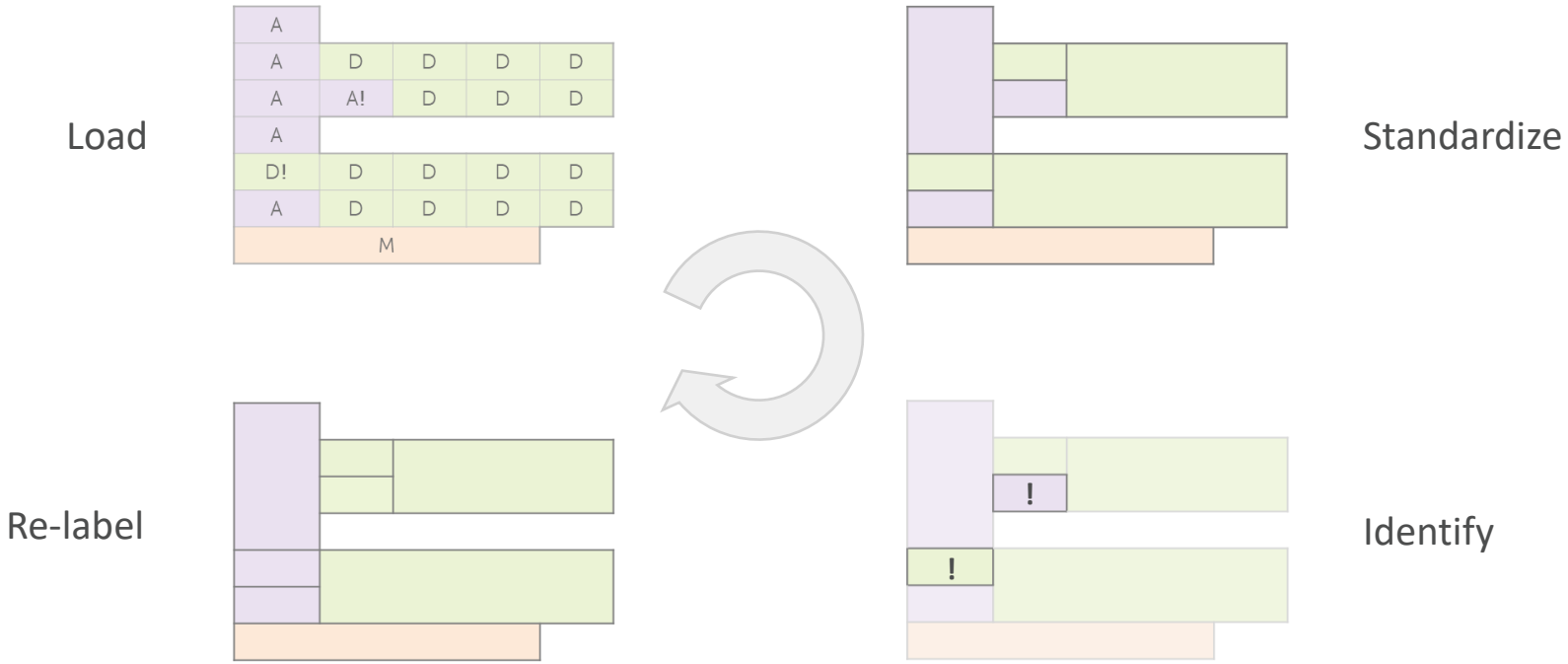


A PROCESSING PIPELINE



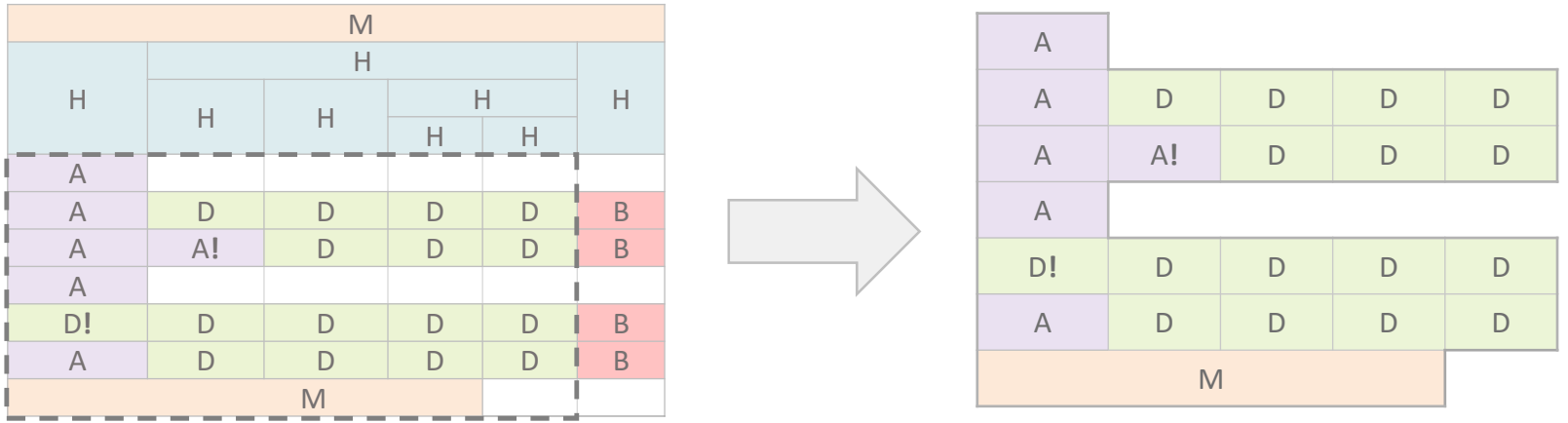
Revising Classification Prediction*

Dresden Database Systems Group



* Koci et al.: Cell Classification for Layout Recognition in Spreadsheets. Chapter at Communications in Computer and Information Science (CCIS) book series. (Currently In Press)

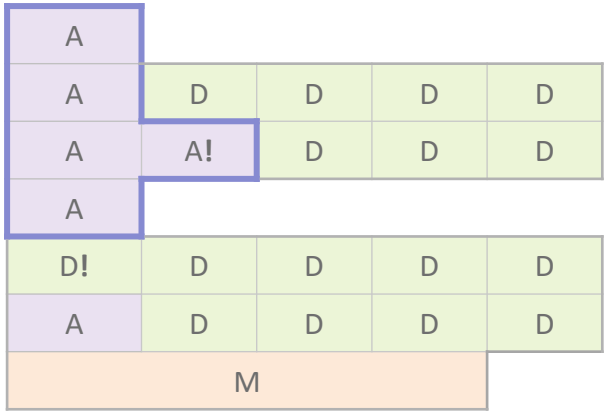
Revising Classification Prediction



Revising Classification Prediction

Dresden Database Systems Group

IRREGULAR REGIONS HINT MISCLASSIFICATIONS

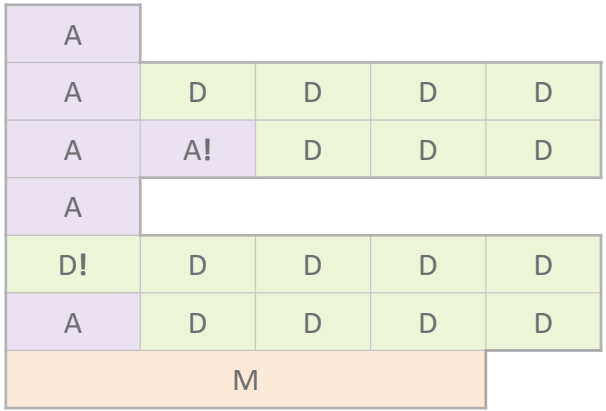


Revising Classification Prediction

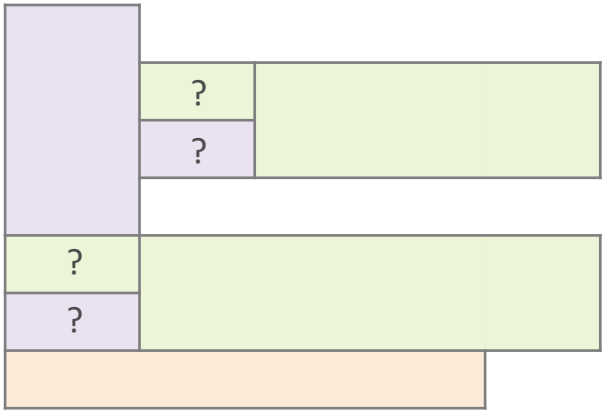
Dresden Database Systems Group

Irregular regions hint misclassifications

- Make all strictly rectangular
- Small regions are suspicious



Standardize!

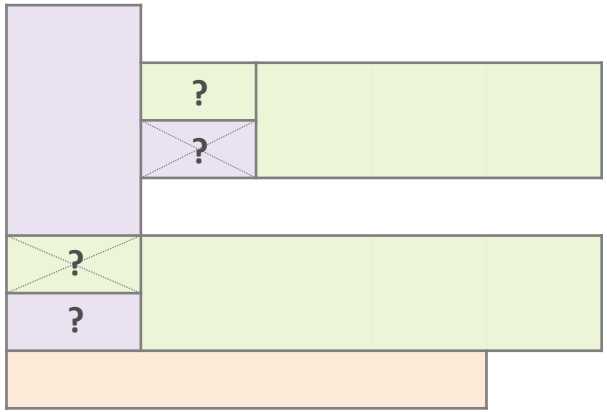


Revising Classification Prediction

Dresden Database Systems Group

IDENTIFY TRUE NEGATIVES

- A supervised machine learning task
- Use the context from the neighborhood



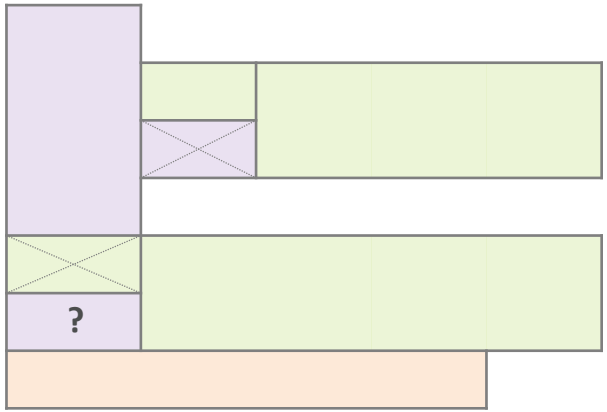
- Similarity
- Dissimilarity

Revising Classification Prediction

Dresden Database Systems Group

IDENTIFY TRUE NEGATIVES

- A supervised machine learning task
- Use the context from the neighborhood



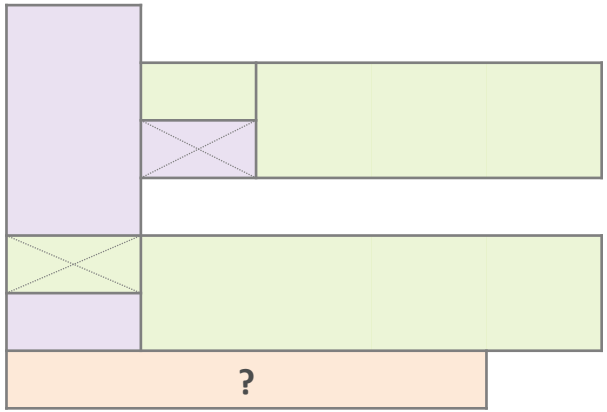
- Similarity
- Dissimilarity
- Influence

Revising Classification Prediction

Dresden Database Systems Group

IDENTIFY TRUE NEGATIVES

- A supervised machine learning task
- Use the context from the neighborhood



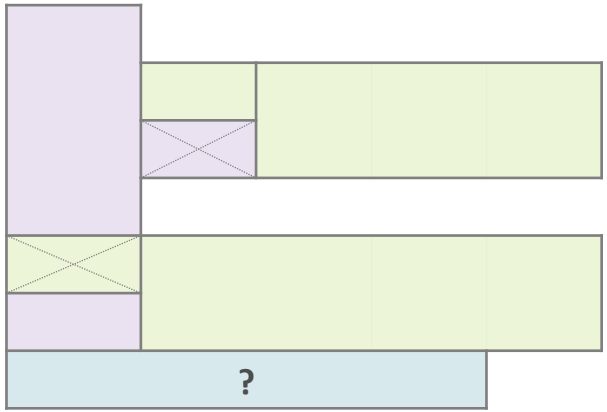
- Similarity
- Dissimilarity
- Influence
- Absence

Revising Classification Prediction

Dresden Database Systems Group

IDENTIFY TRUE NEGATIVES

- A supervised machine learning task
- Use the context from the neighborhood



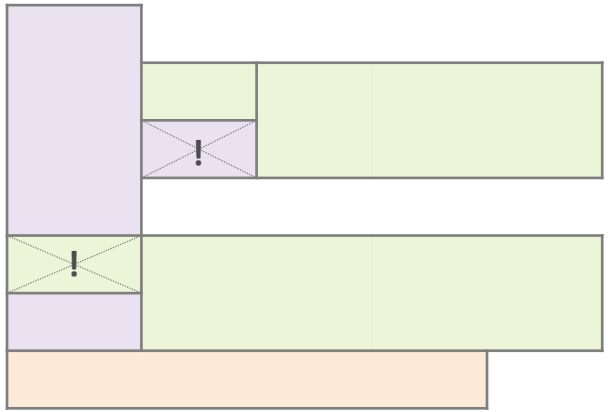
- Similarity
- Dissimilarity
- Influence
- Absence

Revising Classification Prediction

Dresden Database Systems Group

IDENTIFY TRUE NEGATIVES

- A supervised machine learning task
- Use the context from the neighborhood



$$weight_i = OverlapRatio(r, n_i, d) \cdot \frac{1}{1 + Distance(r, n_i)}$$

$$similarity_i = \begin{cases} 0 & Label(r) \neq Label(n_i) \ \vee \ n_i \notin Nearest(r, d) \\ weight_i & otherwise \end{cases}$$

$$dissimilarity_i = \begin{cases} 0 & Label(r) = Label(n_i) \ \vee \ n_i \notin Nearest(r, d) \\ weight_i & otherwise \end{cases}$$

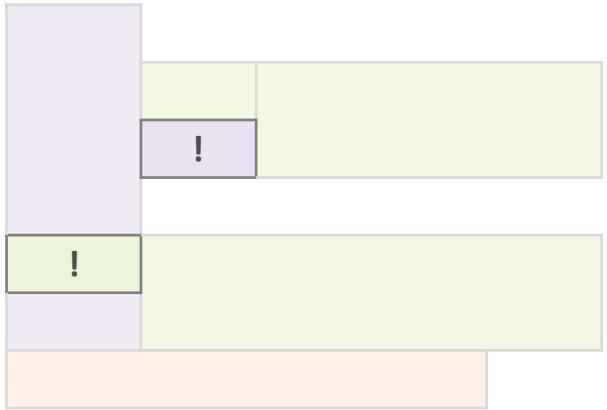
$$influence_i = \begin{cases} 0 & Label(n_i) \neq l \ \vee \ n_i \notin Nearest(r, d, l) \\ weight_i & otherwise \end{cases}$$

Revising Classification Prediction

Dresden Database Systems Group

PREDICT TRUE LABEL (I.E., RELABELING)

- A supervised machine learning task
- Use the context from the neighborhood



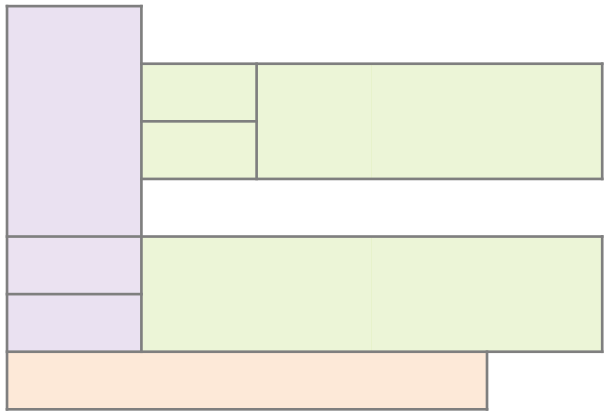
- Similarity
- Dissimilarity
- Influence
- Absence

Revising Classification Prediction

Dresden Database Systems Group

Predict true label (i.e., relabeling)

- A supervised machine learning task
- Use the context from the neighborhood



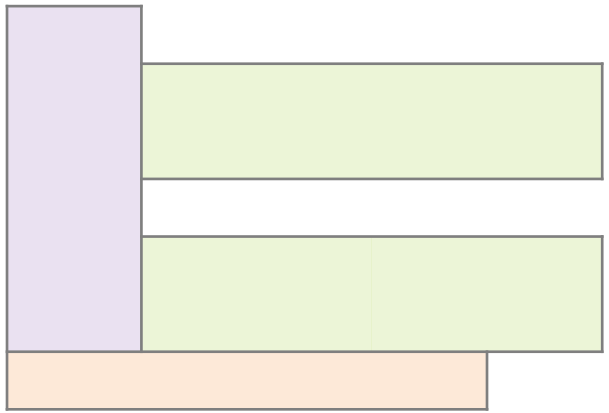
- Similarity
- Dissimilarity
- Influence
- Absence

Revising Classification Prediction

Dresden Database Systems Group

Predict true label (i.e., relabeling)

- A supervised machine learning task
- Use the context from the neighborhood

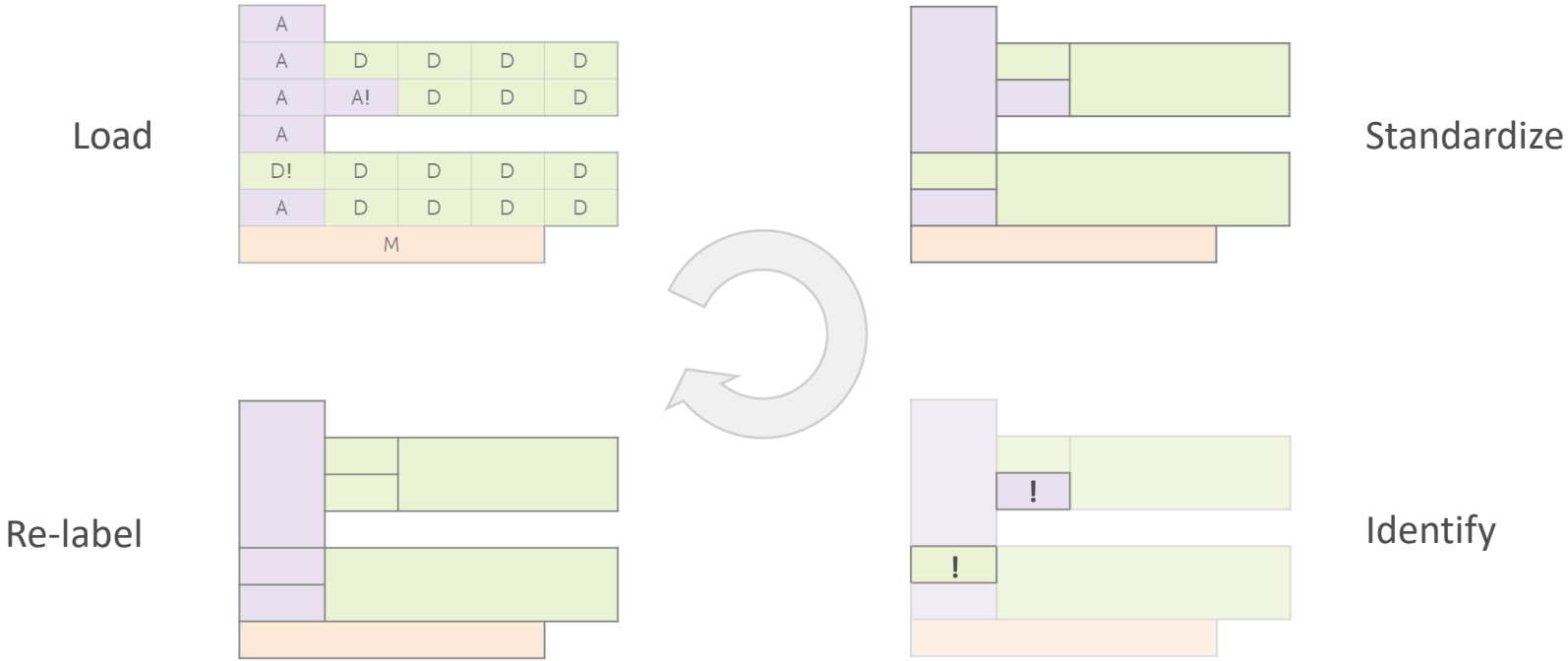


- Similarity
- Dissimilarity
- Influence
- Absence

Revising Classification Prediction

Dresden Database Systems Group

OVERALL PROCESS

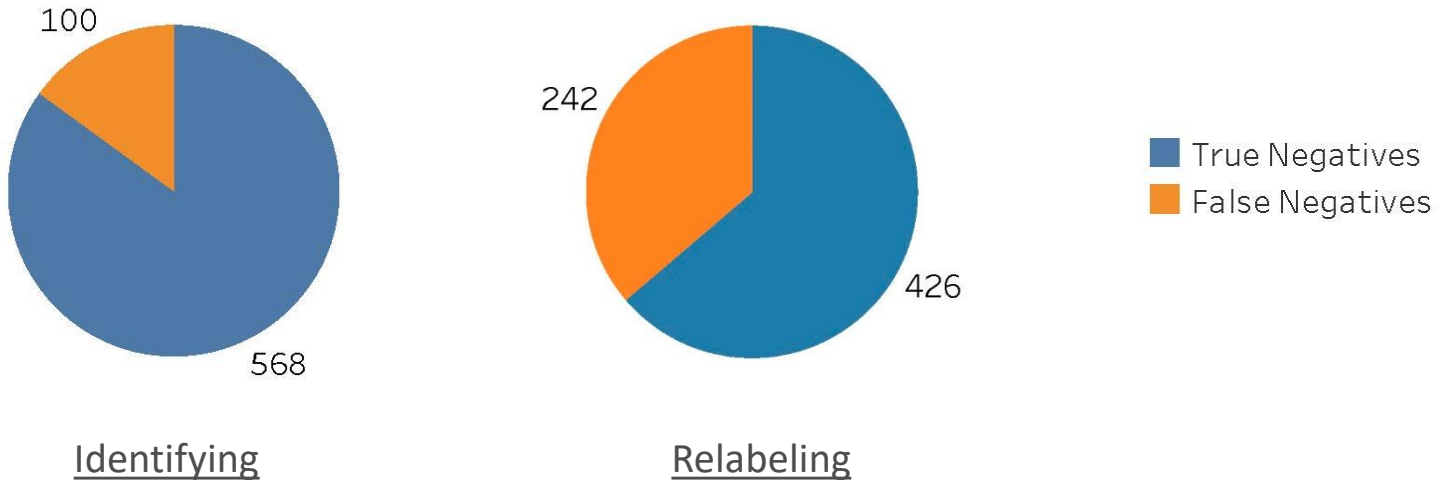


Revising Classification Prediction

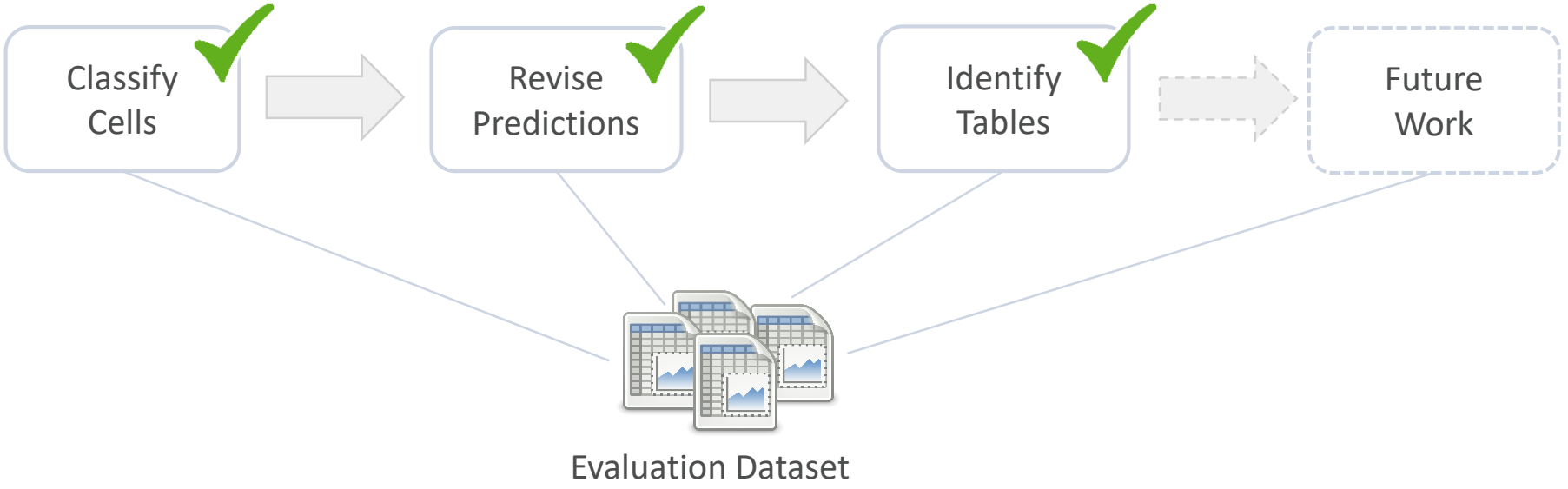
Dresden Database Systems Group

Evaluation Results

- Successfully identified almost half of misclassified (total **1,237**)
- Relabeling approach needs further improvements



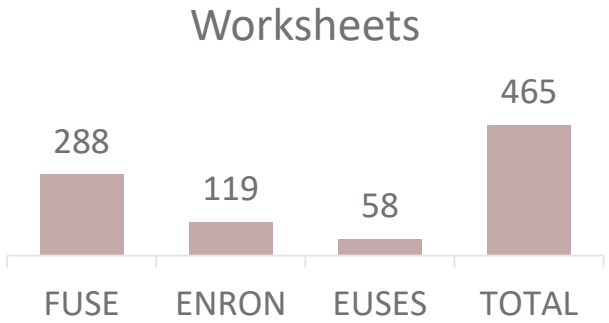
A PROCESSING PIPELINE

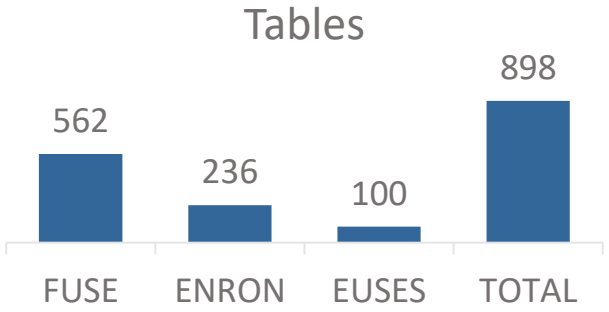


The Evaluation Dataset

Dresden Database Systems Group

THREE CORPORA OF SPREADSHEETS



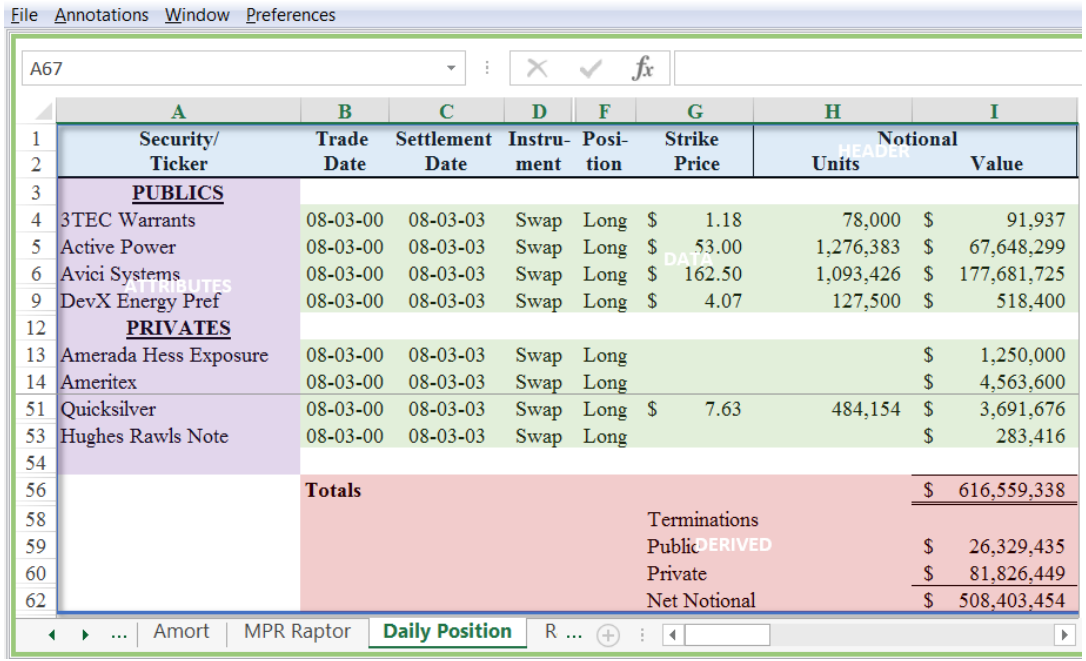


- Barik, T., Lubick, K., Smith, J., Slankas, J., & Murphy-Hill, E. FUSE: A Reproducible, Extendable, Internet-Scale Corpus of Spreadsheets. MSR 2015
- Hermans, F., & Murphy-Hill, E. Enron's Spreadsheets and Related Emails: A Dataset and Analysis. ICSE 2015
- Fisher, M., & Rothermel, G. The EUSES Spreadsheet Corpus: a Shared Resource for Supporting Experimentation With Spreadsheet Dependability Mechanisms. ACM SIGSOFT Software Engineering Notes 2005 (Vol. 30, No. 4, pp. 1-5)

The Evaluation Dataset

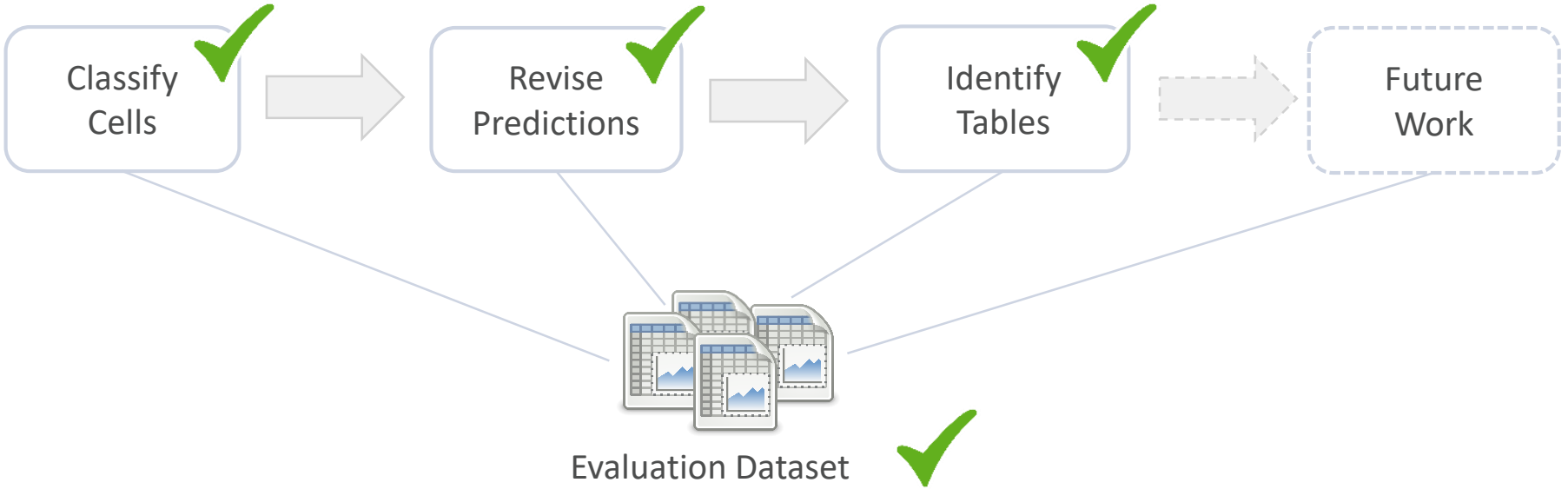
Dresden Database Systems Group

ANNOTATION TOOL



	A	B	C	D	F	G	H	I
1	Security/	Trade	Settlement	Instru-	Posi-	Strike	Notional	
2	Ticker	Date	Date	ment	tion	Price	Units	Value
3	**PUBLICS**							
4	3TEC Warrants	08-03-00	08-03-03	Swap	Long	$ 1.18	78,000	$ 91,937
5	Active Power	08-03-00	08-03-03	Swap	Long	$ 53.00	1,276,383	$ 67,648,299
6	Avici Systems	08-03-00	08-03-03	Swap	Long	$ 162.50	1,093,426	$ 177,681,725
9	DevX Energy Pref	08-03-00	08-03-03	Swap	Long	$ 4.07	127,500	$ 518,400
12	**PRIVATES**							
13	Amerada Hess Exposure	08-03-00	08-03-03	Swap	Long			$ 1,250,000
14	Ameritex	08-03-00	08-03-03	Swap	Long			$ 4,563,600
51	Quicksilver	08-03-00	08-03-03	Swap	Long	$ 7.63	484,154	$ 3,691,676
53	Hughes Rawls Note	08-03-00	08-03-03	Swap	Long			$ 283,416
54								
56		**Totals**						$ 616,559,338
58						Terminations		
59						Public		$ 26,329,435
60						Private		$ 81,826,449
62						Net Notional		$ 508,403,454

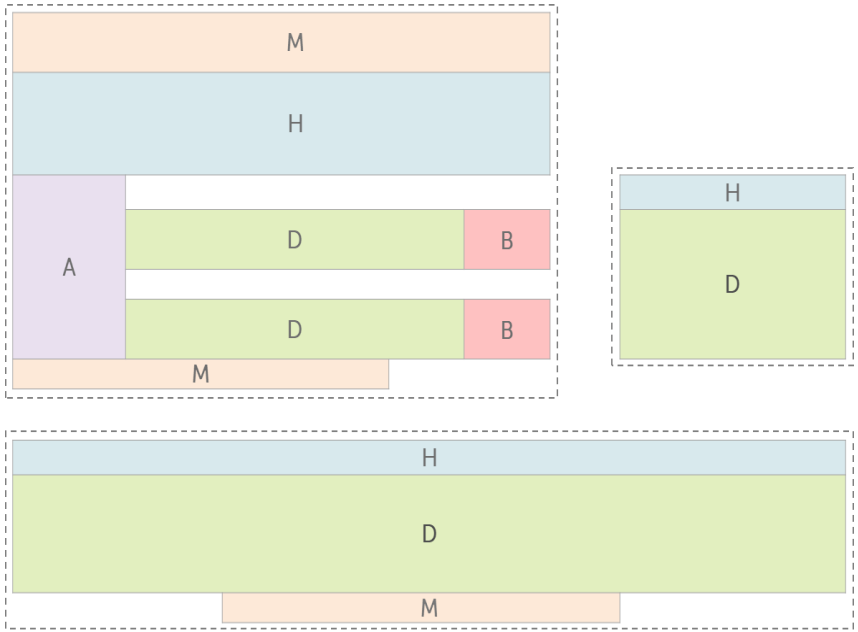
A PROCESSING PIPELINE

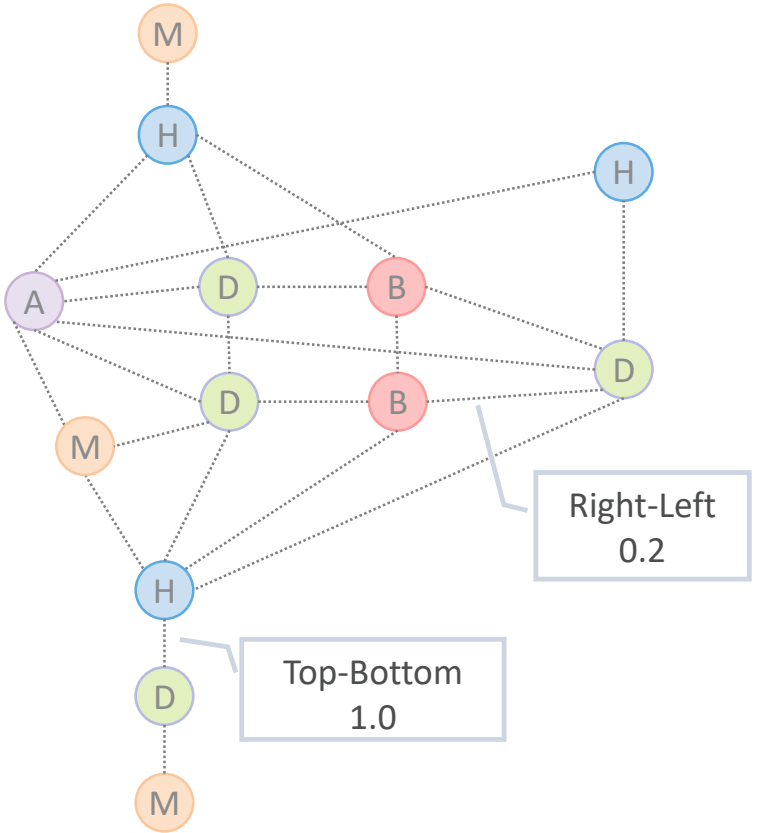


Future Work

Dresden Database Systems Group

Graph Representation

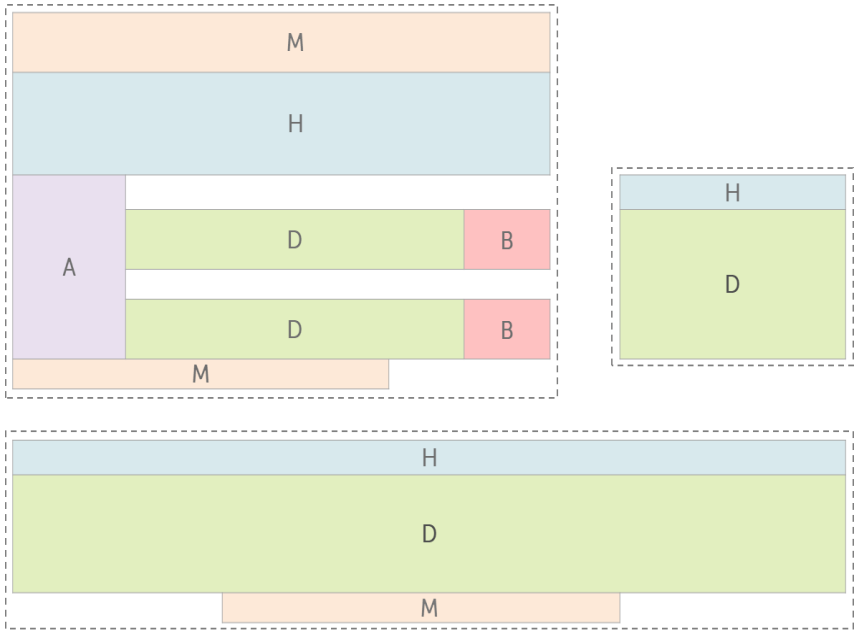


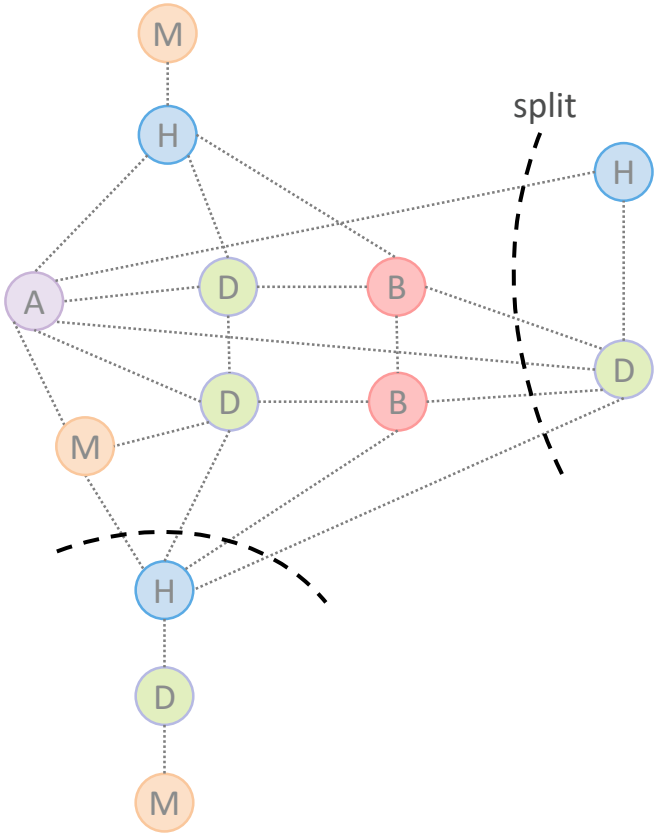


Future Work

Dresden Database Systems Group

Graph Representation

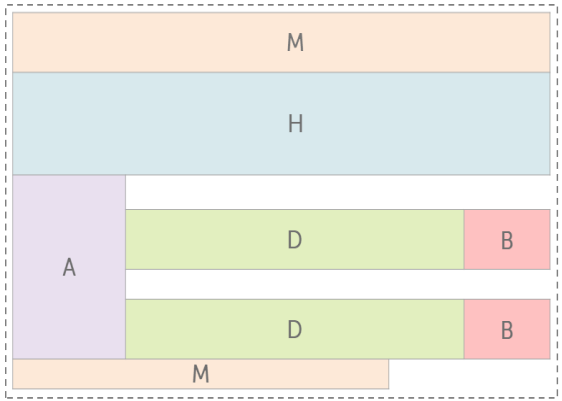


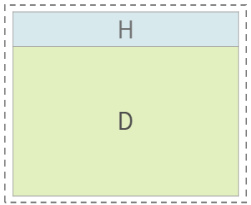


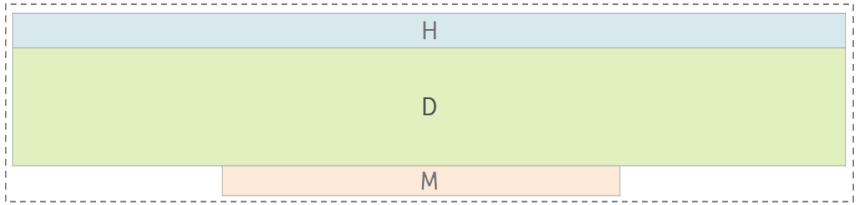
Future Work

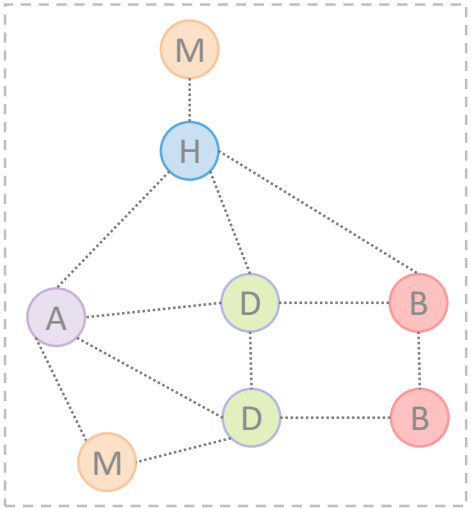
Dresden Database Systems Group

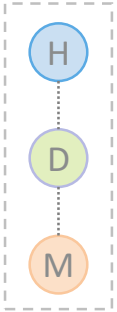
Graph Representation







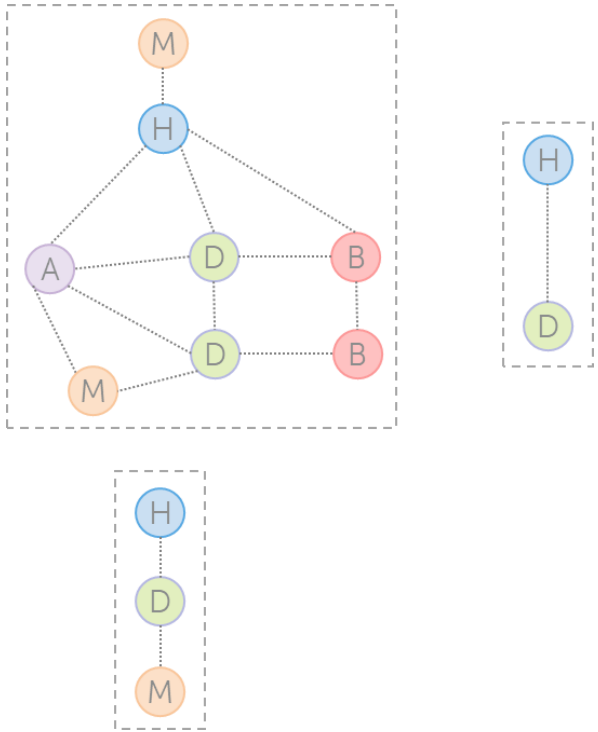




Future Work

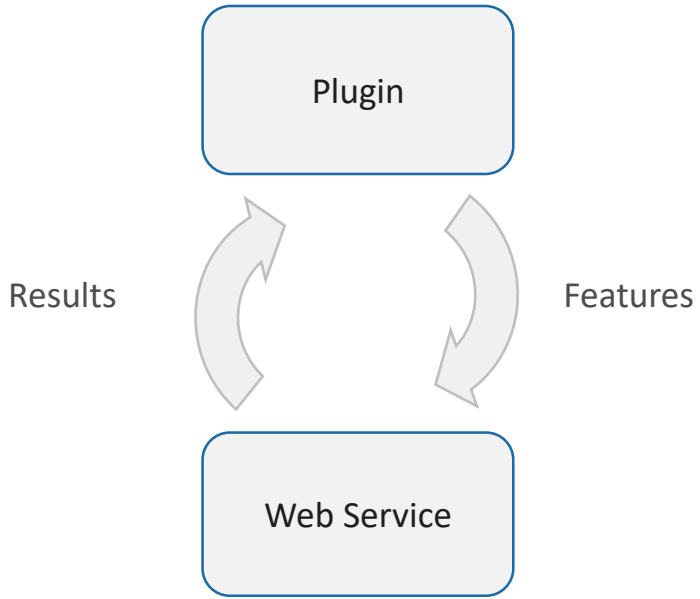
Dresden Database Systems Group

- Graph-based Table Identification (target **ECIR**)
  - Node clustering
  - Guided minimum-cut
  - Fitness function



Future Work

- Graph-based Table Identification (target **ECIR**)
  - Node clustering
  - Guided minimum-cut
  - Fitness function
- Demo: a plugin for Excel to extract information

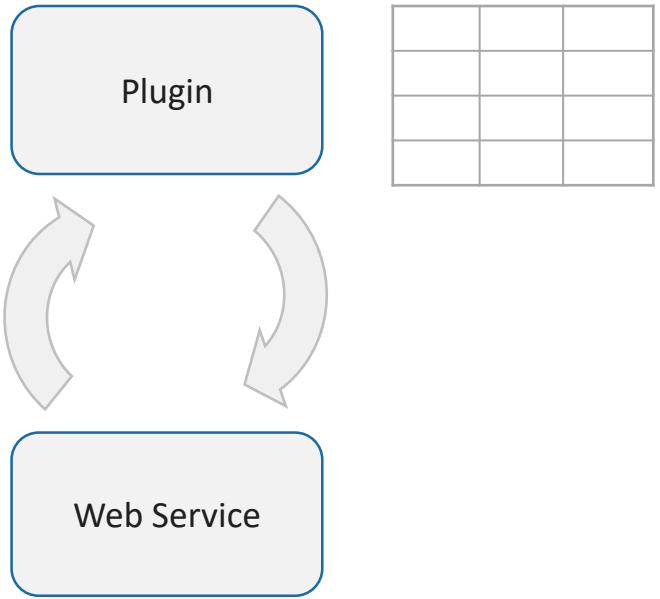


Future Work

Dresden Database Systems Group

Future Work

- Graph-based Table Identification (target **ECIR**)
  - Node clustering
  - Guided minimum-cut
  - Fitness function
- Demo: a plugin for Excel to extract information

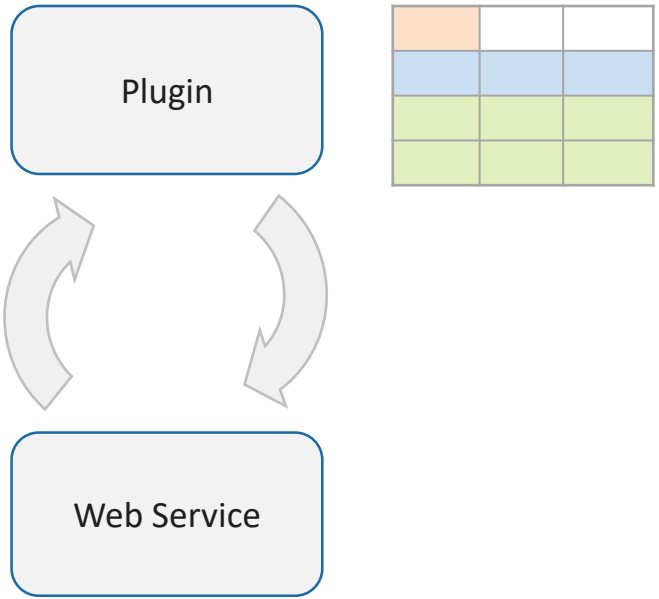


Future Work

Dresden Database Systems Group

Future Work

- Graph-based Table Identification (target **ECIR**)
  - Node clustering
  - Guided minimum-cut
  - Fitness function
- Demo: a plugin for Excel to extract information



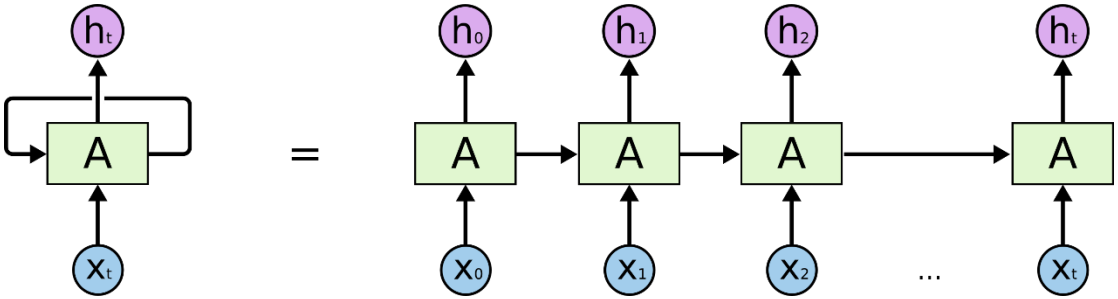
Future Work

Dresden Database Systems Group

Future Work

- Graph-based Table Identification (target **ECIR**)
  - Node clustering
  - Guided minimum-cut
  - Fitness function
- Demo: a plugin for Excel to extract information
- Cell classification using Recurrent Neural Networks

	A	B	C	D	E
1		Number of Items (in Units) Sold per Region			
2					
3		Location	Item		Total
4			Mouse	Monitor	Adapter	
5					VGA	HDMI	
6		Europe					
7		Spain	500	200	85	61	846
8		France	465	169	80	80	794
9		Asia					
10		China	422	163	90	44	719
11		Vietnam	473	182	74	55	784
12		Item "Keyboard" is omitted. Check next sheet.					



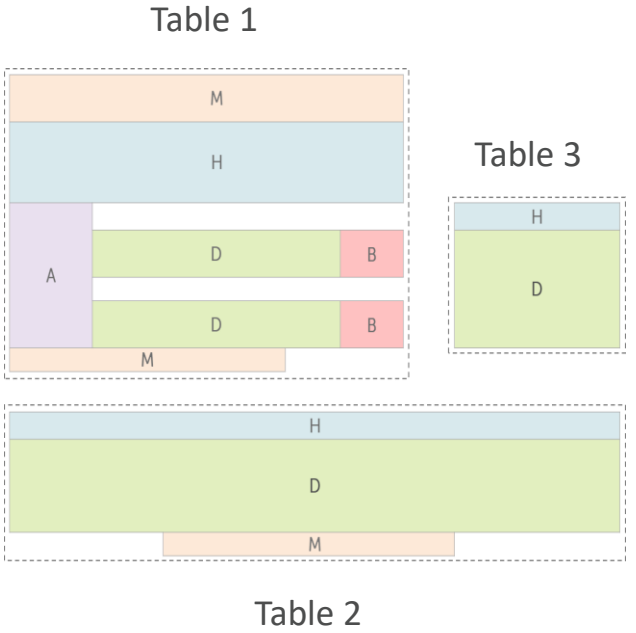
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Conclusions & Future Work

Dresden Database Systems Group

Future Work

- Graph-based Table Identification (target **ECIR**)
  - Node clustering
  - Guided minimum-cut
  - Fitness function
- Demo: a plugin for Excel to extract information
- Cell classification using Recurrent Neural Networks
- Schema Extraction from identified tables

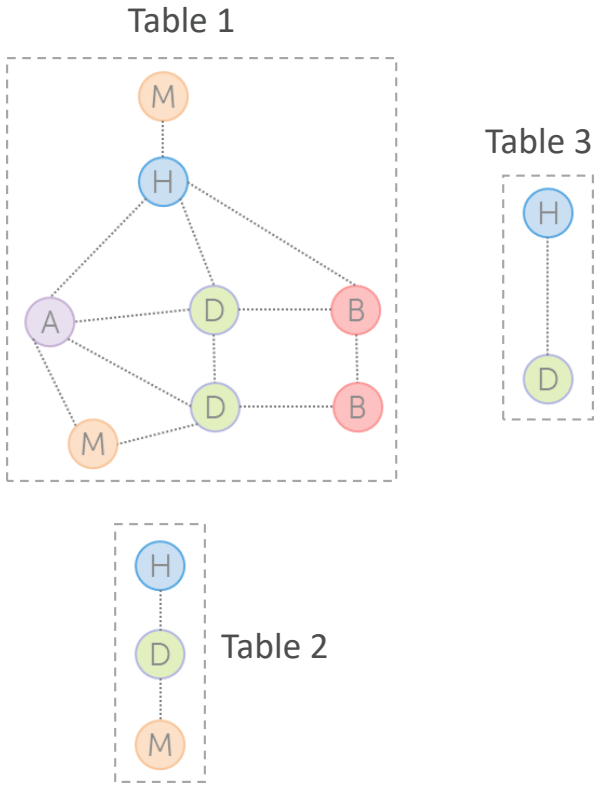


Conclusions & Future Work

Dresden Database Systems Group

Future Work

- Graph-based Table Identification (target **ECIR**)
  - Node clustering
  - Guided minimum-cut
  - Fitness function
- Demo: a plugin for Excel to extract information
- Cell classification using Recurrent Neural Networks
- Schema Extraction from identified tables



Related Work

1. Chen, Z., & Cafarella, M. (2013, August). “Automatic Web Spreadsheet Data Extraction”. In Proceedings of the 3rd International Workshop on Semantic Search Over the Web (p. 1). ACM.

2. Eberius, J., Werner, C., Thiele, M., Braunschweig, K., Dannecker, L., & Lehner, W. (2013, October). “DeExcelerator: A Framework for Extracting Relational Data from Partially Structured Documents”. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (pp. 2477-2480). ACM.

3. Adelfio, M. D. and Samet, H. (2013). “Schema Extraction for Tabular Data on the Web”. VLDB’13, 6(6): 421–432.

4. Cunha, J., Saraiva, J., & Visser, J. (2009, January). “From Spreadsheets to Relational Databases and Back”. In Proceedings of the 2009 ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation (pp. 179-188). ACM.

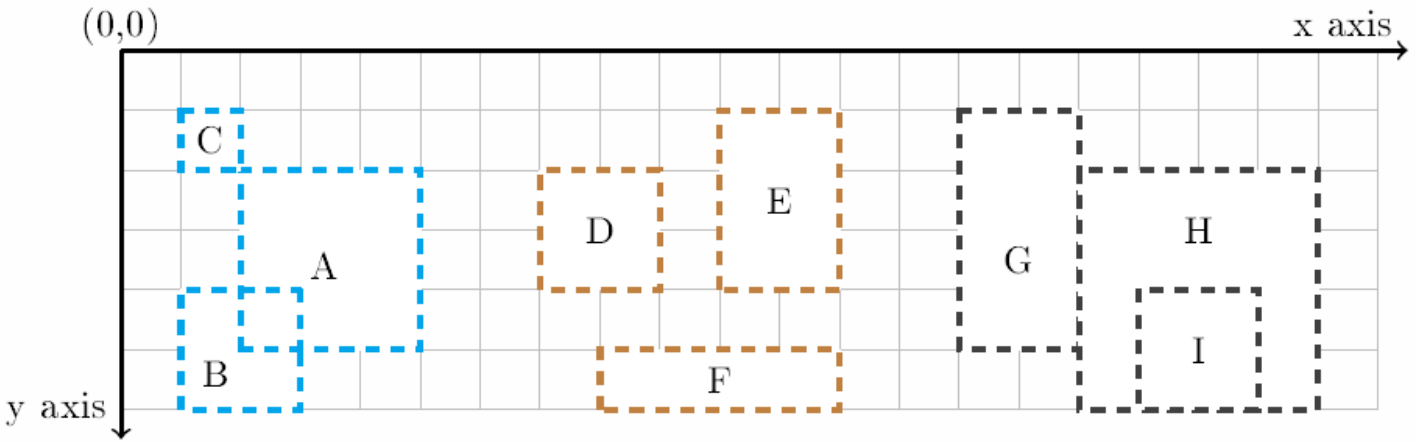
5. Abraham, R., & Erwig, M. (2007). “UCheck: A Spreadsheet Type Checker for End Users”. Journal of Visual Languages & Computing, 18(1), 71-95.

6. Shigarov, Alexey O. (2015). "Table Understanding Using a Rule Engine“. Expert Systems with Applications, 42(2): 929-937.

7. Swidan, A., & Hermans, F. (2017, June). Semi-automatic Extraction of Cross-Table Data from a Set of Spreadsheets. In International Symposium on End User Development (pp. 84-99). Springer, Cham.

Thank You!

Region Spatial Arrangements



Constructing Tables - Preliminaries

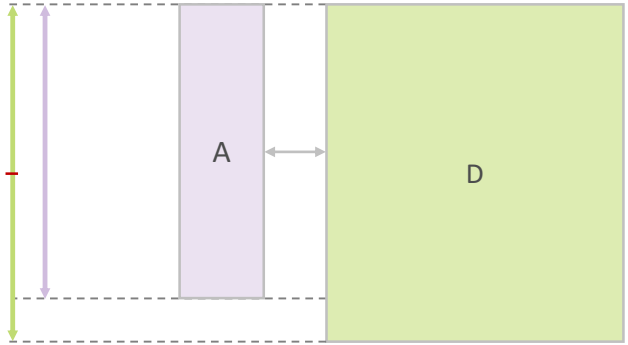
Dresden Database Systems Group

Constrains For Pairing Attributes with Data

1) **Attributes** are on the **left** of Data
2) The distance between them is the smallest
3) Their projections overlap significantly:

$$\frac{overlap(projection(D), projection(F))}{max(projection\_length(D), projection\_length(F))} > \theta$$

- Overlap ratio measures this significance
- For the evaluation ϑ (empirically) set to **0.5**



Constructing Tables - Preliminaries

Dresden Database Systems Group

Constrains For Pairing Attributes with Data

1) Attributes are on the left of Data
2) The distance between them is the smallest
3) Their projections overlap significantly:

$$\frac{overlap(projection(D), projection(F))}{max(projection\_length(D), projection\_length(F))} > \theta$$

- Overlap ratio measures this significance
- For the evaluation ϑ (empirically) set to **0.5**