Smelly Spreadsheet Structures:

Structural Analysis of Spreadsheets to enhance Smell Detection

Defense of Master's Thesis

Patrick Koch

Supervisor
Univ.-Prof. Dipl.-Ing. Dr.techn. Franz Wotawa

2016-05-19

Outline


Motivation
Structural Analysis Process
Evaluation
Structure-based Smells
Future Work

Motivation

Why Structural Analysis of Spreadsheets?

Let us restate the question.

Motivation

Why spreadsheets?

- Intuitive to use
- Versatile
- Everyone used them once
- Everyone has access

Motivation

So that is why…

- 60% of PC-users use spreadsheets
- 7.4% of PC-time spent using spreadsheets
- Used by 95% of businesses for financial accounting

However…

- Only 15% of users are professionals
- Spreadsheets contain (a lot of) faults

Source:
Panko and Port: “End user computing: The dark matter (and dark energy) of corporate IT”, HICSS 2012
Kevin Taylor: “An Analysis of Computer Use across 95 Organisations in Europe, North America and Australasia”, 2007

A grave Example

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1				Number of observations					Average							
2									Real GDP growth				Inflation			
3					Debt/GDP				Debt/GDP				Debt/GDP			
4		Country	Coverage	Total	30 or less	30 to 60	60 to 90	90 or abov	30 or less	30 to 60	60 to 90	90 or abov	30 or less	30 to 60	60 to 90	90 or abov
5	1	US	1791-2009		129	58	23	5	4.0	3.4	3.3	-1.8	1.1	1.8	2.3	6.1
6	2	UK	1831-2009		3	68	27	75	2.5	2.1	2.1	1.8	0.8	4.1	1.4	1.6
7	3	Sweden	1880-2009		79	40	11	0	2.9	2.9	2.7	n.a.	2.8	4.6	4.2	n.a.
8	4	Spain	1851-2009		26	53	47	29	1.5	3.2	1.3	2.2	10.5	5.5	2.3	0.5
9	5	Portugal	1880-2009		42	10	39	0	4.8	2.5	1.4	n.a.	8.8	3.3	0.9	n.a.
10	6	Norway	1880-2009		98	25	1	0	2.9	4.4	10.2	n.a.	4.4	-0.1	n.a.	n.a.
11	7	New Zeala	1932-2009		9	33	17	19	2.5	2.9	3.9	3.7	2.6	7.4	5.0	2.8
12	8	Netherland	1880-2009		17	50	32	8	4.1	2.8	2.4	2.0	6.4	1.5	0.0	-2.2
13	9	Japan	1886-2009		47	42	11	11	5.2	3.7	3.9	0.7	6.3	2.1	3.2	-1.1
14	10	Italy	1880-2009		26	12	39	49	5.4	4.9	1.9	0.7	5.6	11.1	10.6	13.1
15	11	Ireland	1949-2009		8	14	32	7	4.4	4.5	4.0	2.4	2.9	4.8	7.3	5.3
16	12	Greece	1884-2009		13	11	11	56	4.0	3.3	4.8	2.5	13.3	20.8	12.3	2.8
17	13	Germany	1880-2009		96	11	0	0	3.6	0.9	n.a.	n.a.	1.8	1.5	n.a.	n.a.
18	14	France	1880-2009		26	21	19	36	5.1	2.7	2.8	1.9	5.2	5.0	1.5	0.9
19	15	Finland	1914-2009		69	18	6	3	3.2	3.0	4.3	1.9	10.6	5.4	13.2	32.7
20	16	Denmark	1880-2009		57	16	17	0	3.2	1.7	2.4	n.a.	2.5	4.7	3.3	n.a.
21	17	Canada	1925-2009		3	52	23	7	1.9	4.5	3.0	2.2	2.2	4.1	0.6	6.0
22	18	Belgium	1836-2009		37	60	32	33	3.0	2.4	2.1	3.3	0.9	1.5	3.0	3.2
23	19	Austria	1880-200		43	32	35	0	4.3	3.0	2.3	n.a.	5.3	2.4	0.7	n.a.
24	20	Australia	1902-2009		38	33	28	9	3.1	4.1	2.9	3.5	5.9	2.9	4.7	3.3
25																
26				1767	688	466	315	=SUM(H5	3.7	3.1	3.5	1.6	5.5	5.3	4.9	5.7
27		Minimum							1.5	0.9	1.3	-1.8	0.8	-0.1	0.0	-2.2
28		Maximum							5.4	4.9	10.2	3.7	13.3	20.8	13.2	32.7

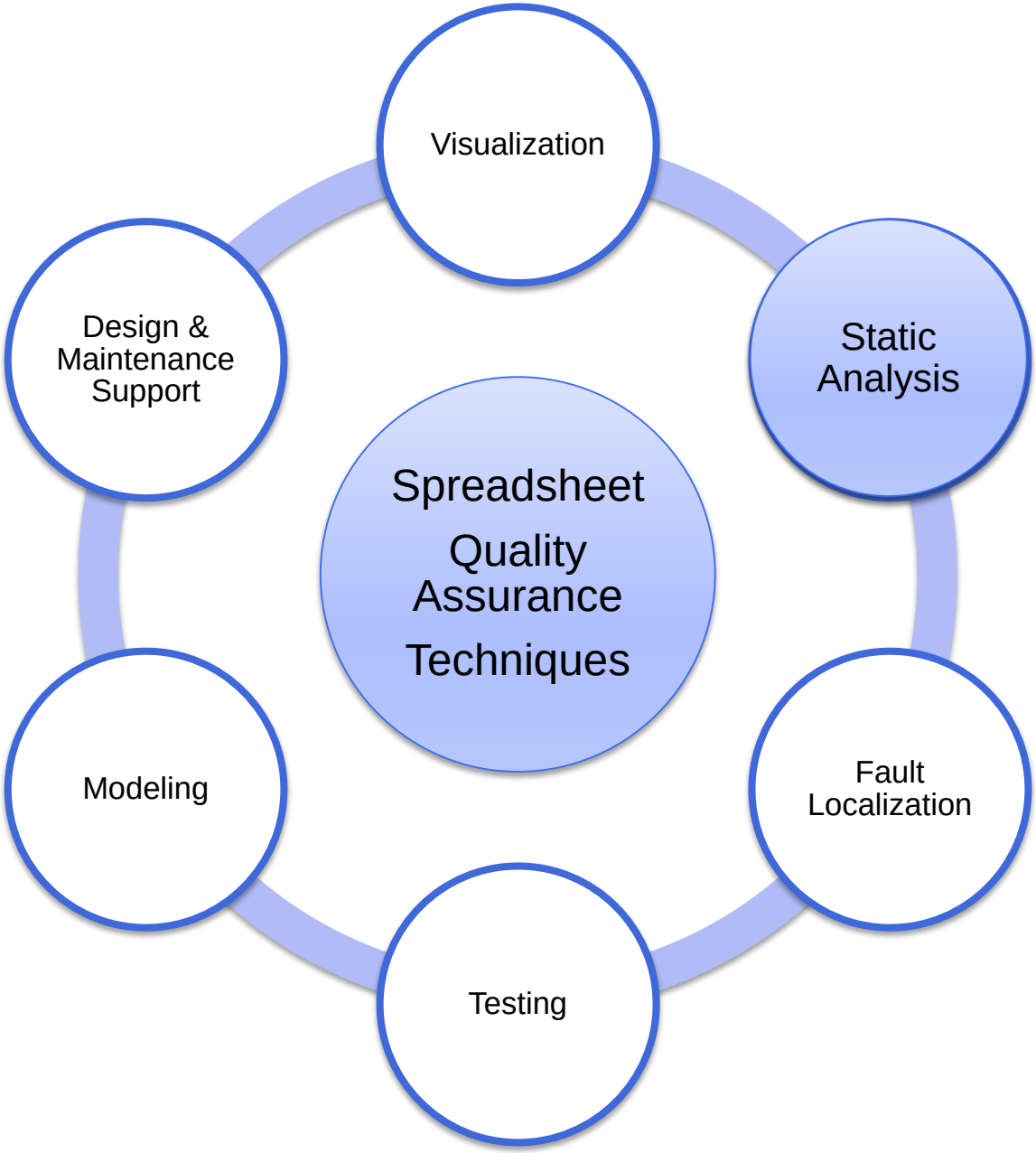
Source: Reinhart, Carmen M, and Kenneth Rogoff.: “Growth in a Time of Debt", 2010.

Motivation

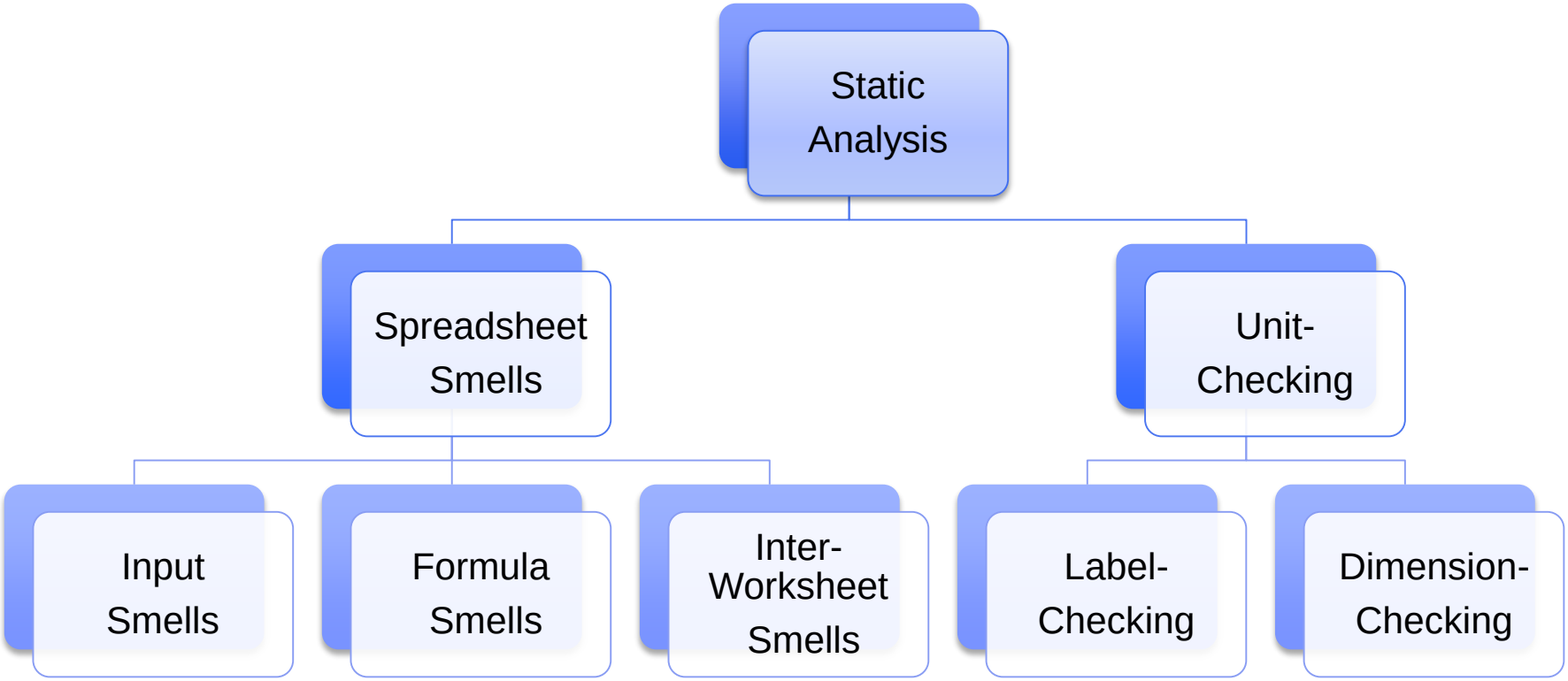
What can we do about it?

Techniques to automatically ...

- Avoid Faults
- Find Faults
- Fix Faults


Visualization
Static Analysis
Fault Localization
Testing
Modeling
Design & Maintenance Support
Spreadsheet Quality Assurance Techniques

Spreadsheet QA Techniques



Outline


Motivation
Structural Analysis Process
Evaluation
Structure-based Smells
Future Work

Structural Analysis Process

We want to automatically...

- Detect structures
- Infer relations
- Provide abstractions

In order to...

- Improve QA techniques
- Introduce new techniques

Structural Analysis Process

Which structures to detect?

Structural Analysis Process: Example Worksheet

Values

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	117
4	Mazda	10	12	9	7	38
5	Fiat	9	12	13	15	49
6	Total	49	51	50	54	204

Formulas

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

Structural Analysis Process

Which structures to detect?

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)
7						
8				formula		

Structural Analysis Process

Which structures to detect?

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)
7						
8				input		

Structural Analysis Process

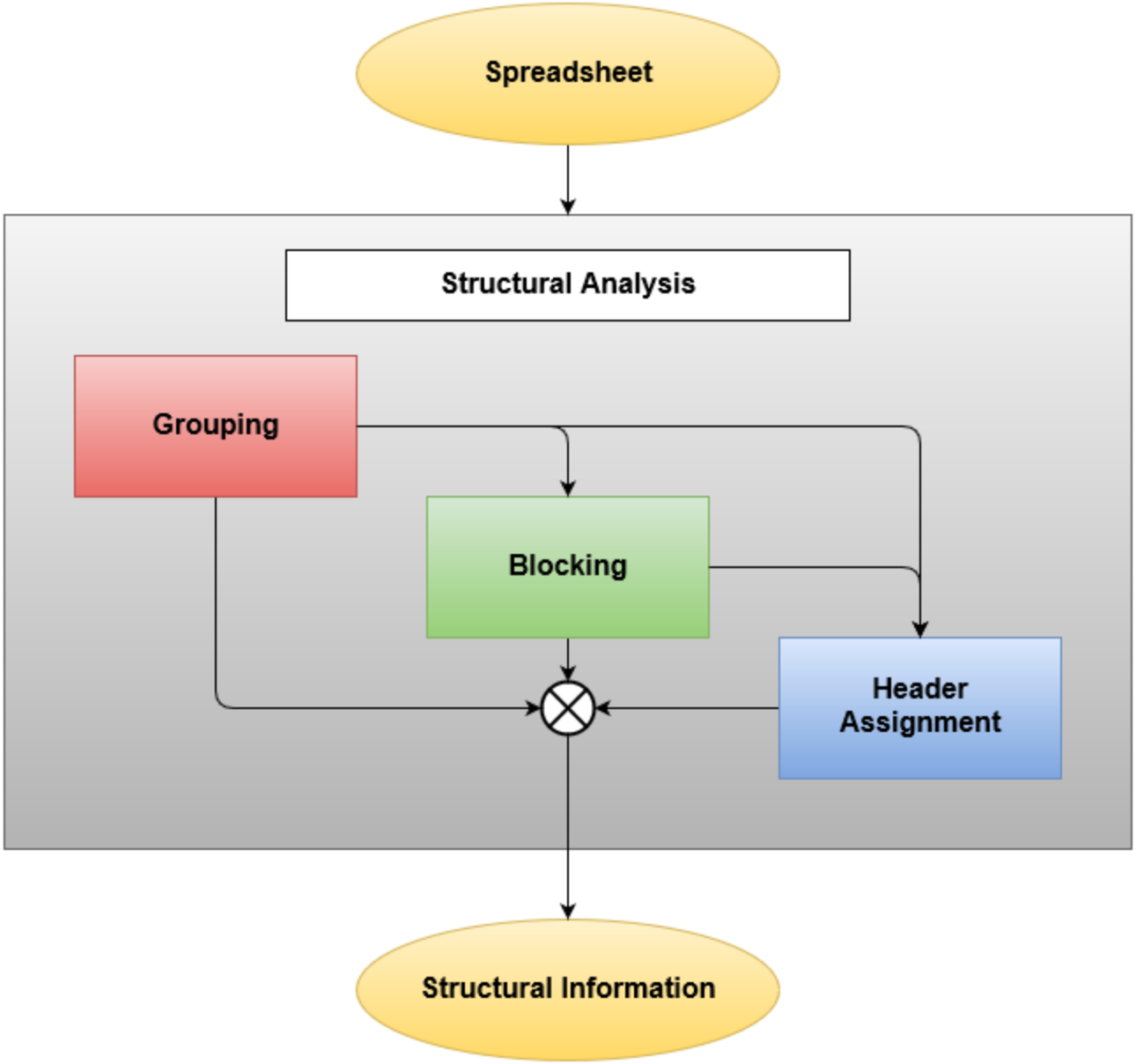
Which structures to detect?

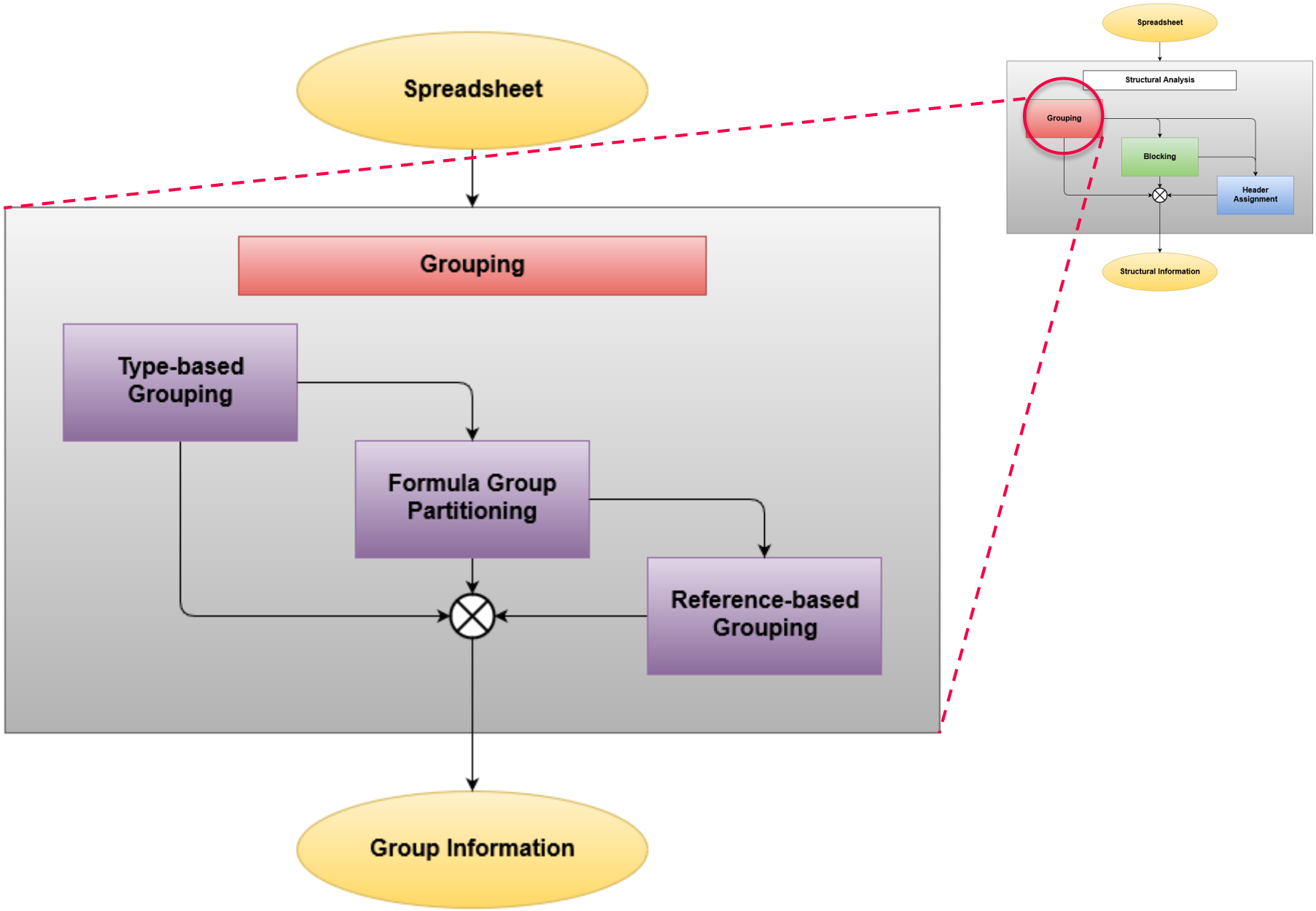
	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)
7						
8				block		

Structural Analysis Process

Which structures to detect?

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)
7						
8				header		


Spreadsheet
Structural Analysis
Grouping
Blocking
Header Assignment
Structural Information


Spreadsheet
Grouping
Type-based Grouping
Formula Group Partitioning
Reference-based Grouping
Group Information
Spreadsheet
Structural Analysis
Grouping
Blocking
Header Assignment
Structural Information

Structural Analysis Process: Grouping

Resulting Formula Groups

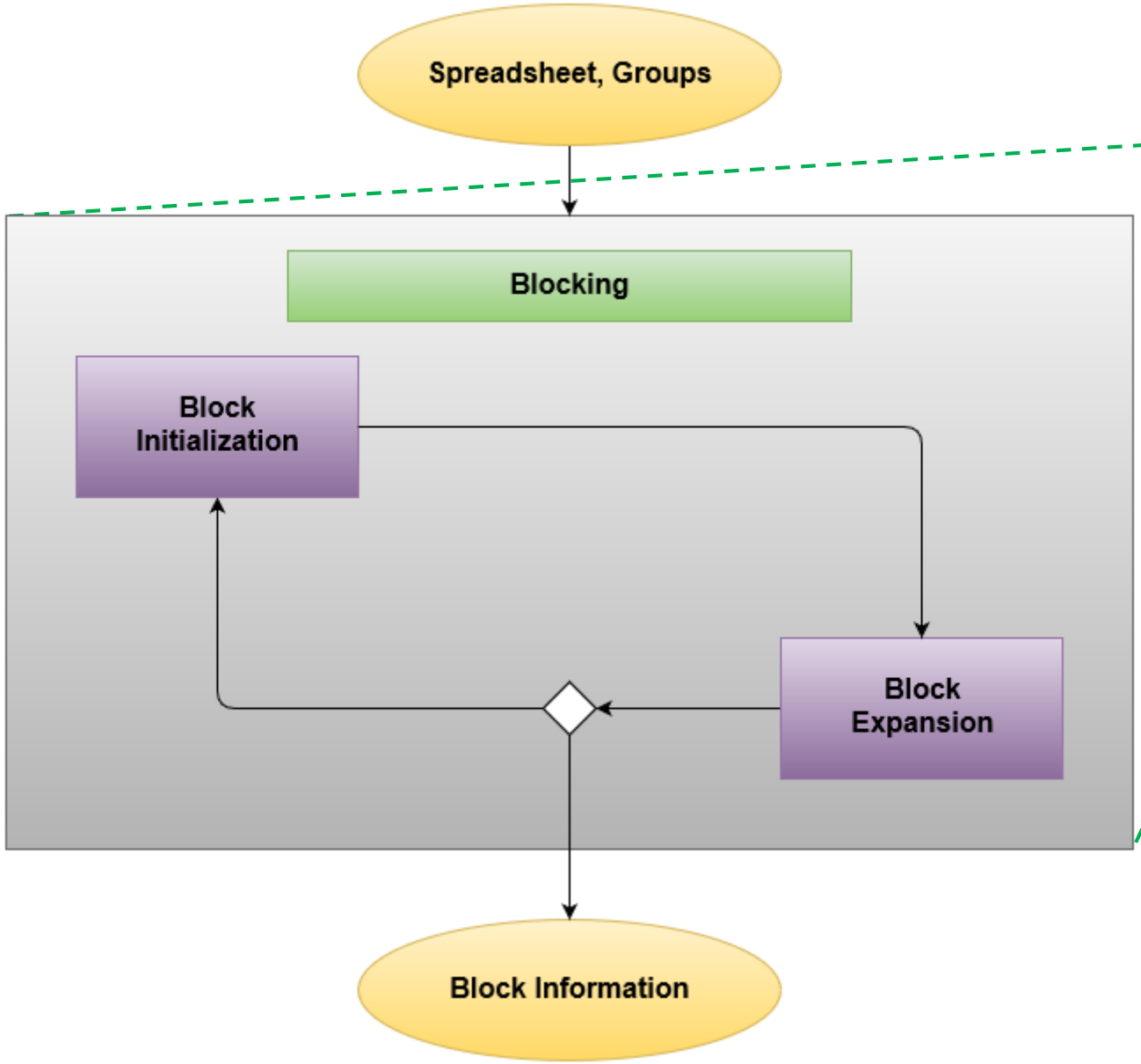
	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

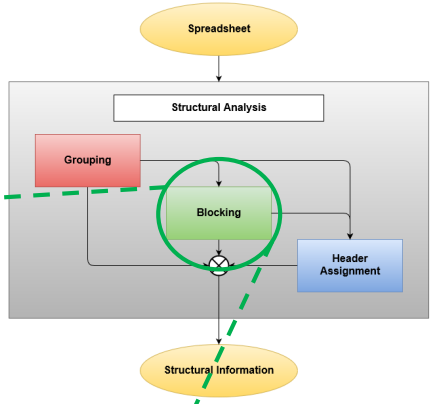
Structural Analysis Process: Grouping

Resulting Reference-based Groups

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

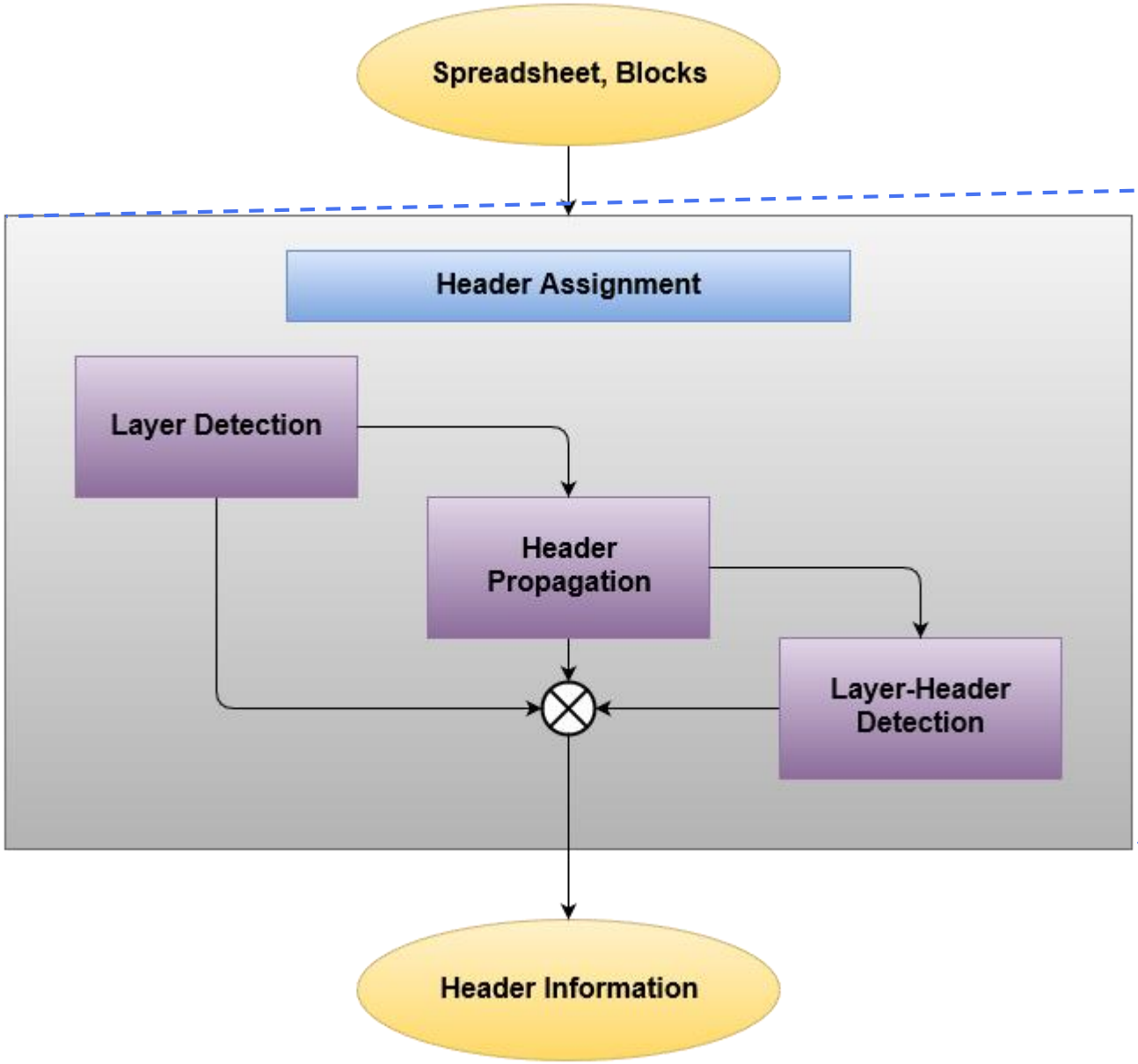

Spreadsheet, Groups
Blocking
Block Initialization
Block Expansion
Block Information

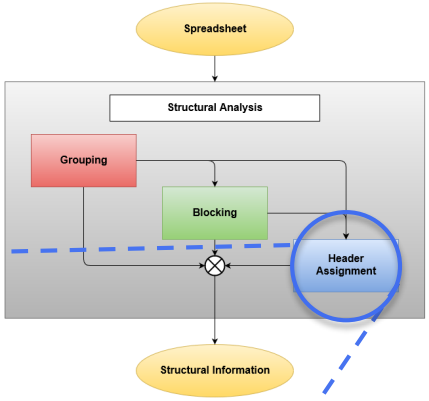

Spreadsheet
Structural Analysis
Grouping
Blocking
Header Assignment
Structural Information

Structural Analysis Process: Blocking

Resulting Block

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)


Spreadsheet, Blocks
Header Assignment
Layer Detection
Header Propagation
Layer-Header Detection
Header Information


Spreadsheet
Structural Analysis
Grouping
Blocking
Header Assignment
Structural Information

Structural Analysis Process: Header Assignation

Resulting Header-Layers

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

Resulting Layer Headers

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

Structural Analysis Process: Challenges

1. Prerequisites for analysis not fulfilled
   - No formula cells
   - No available header cells

2. No consistent guideline for spreadsheet layout
   - Ambiguous spreadsheet structures
   - Inconsistent positioning of individual components
   - Hidden columns and rows

3. Inconsistent use of spreadsheet elements
   - Formula used as header
   - References to valid header cells

Outline


Motivation
Structural Analysis Process
Evaluation
Structure-based Smells
Future Work

Evaluation

- Based on EUSES and ENRON spreadsheet corpora
- Filtering procedure to determine valid candidates

- Resulting filtered evaluation corpora

	Total	Fit for evaulation	Difference
EUSES	4,495	1,659	2,836
ENRON	16,929	6,691	9,238

Evaluation

- Evaluation via comparison with manual inspection
- Detection rates of high-level structures (Blocks, Headers)
- Sample selection based on processed and available metrics
  - EUSES: five typical spreadsheets from each of nine different categories
  - ENRON: 30 most relevant spreadsheets based on version history

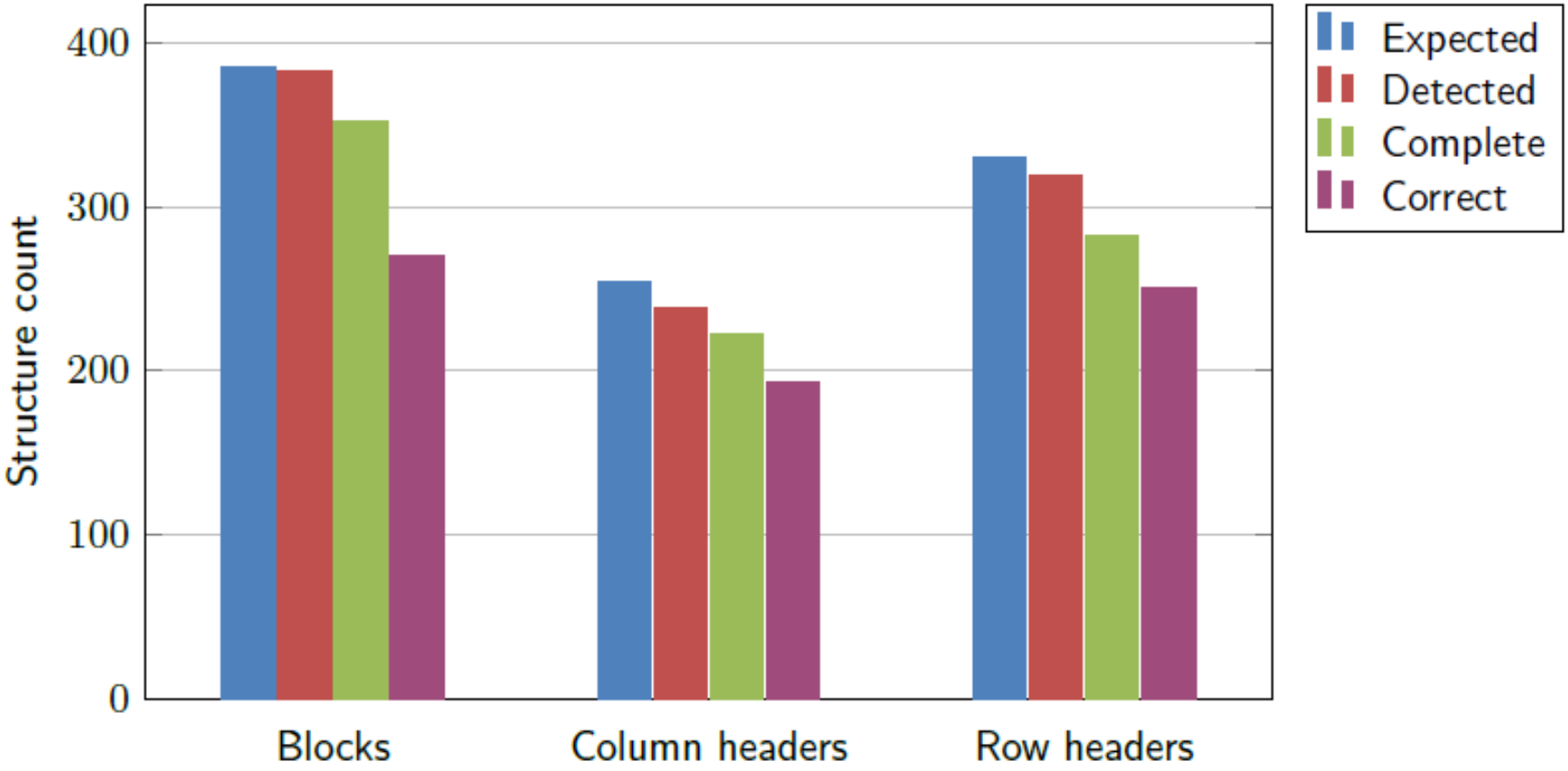
Evaluation: Evaluation Process

Manual inspection

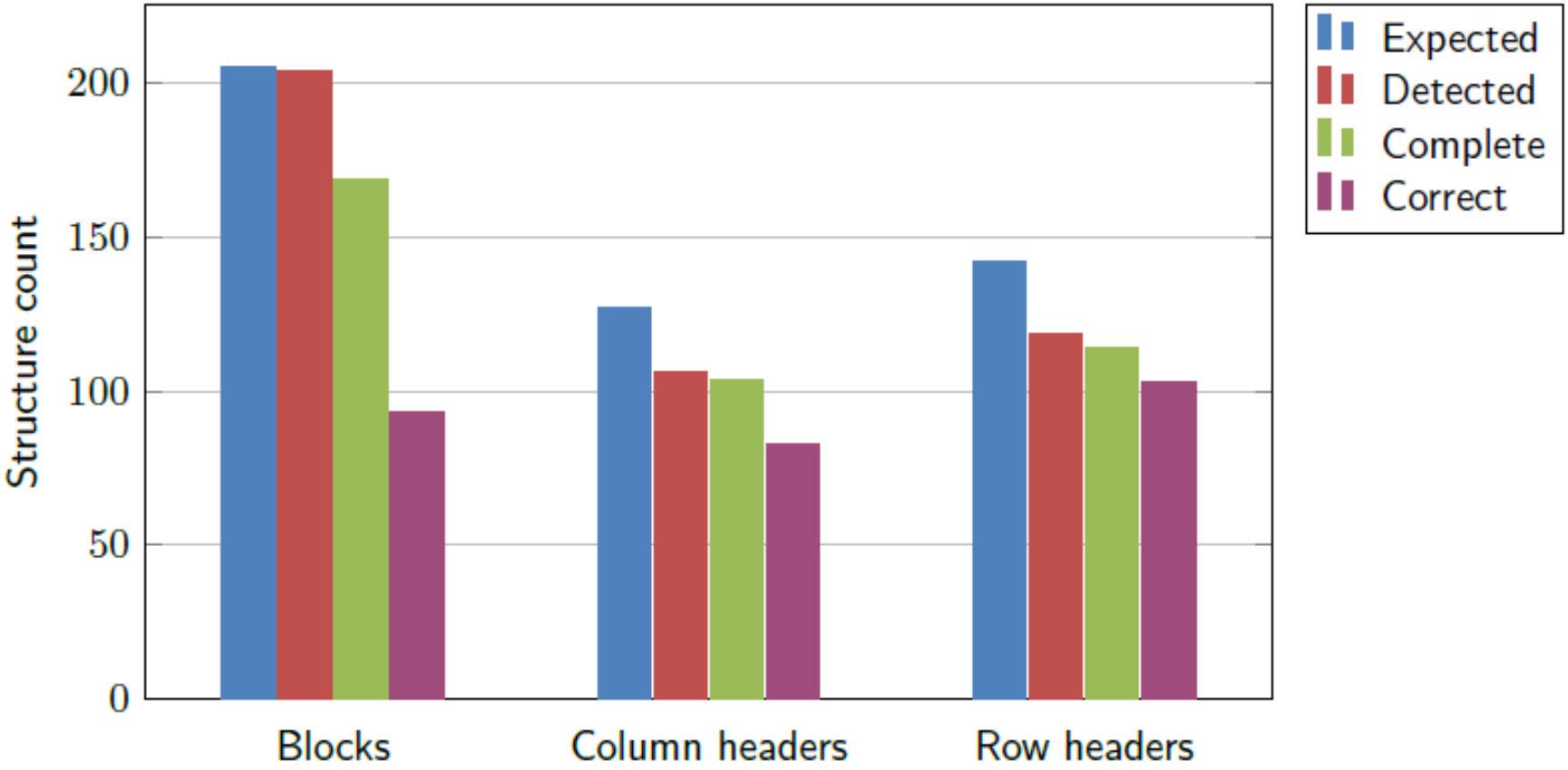
Automatic detection by prototype implementation

Comparison of expected and detected structures

Evaluation: Results EUSES



Evaluation: Results ENRON



Evaluation: Main Findings

1. 99% of the blocks could be detected
2. >80% of expected headers could be detected
3. Most of the detected structures inferred featuring complete dimensions

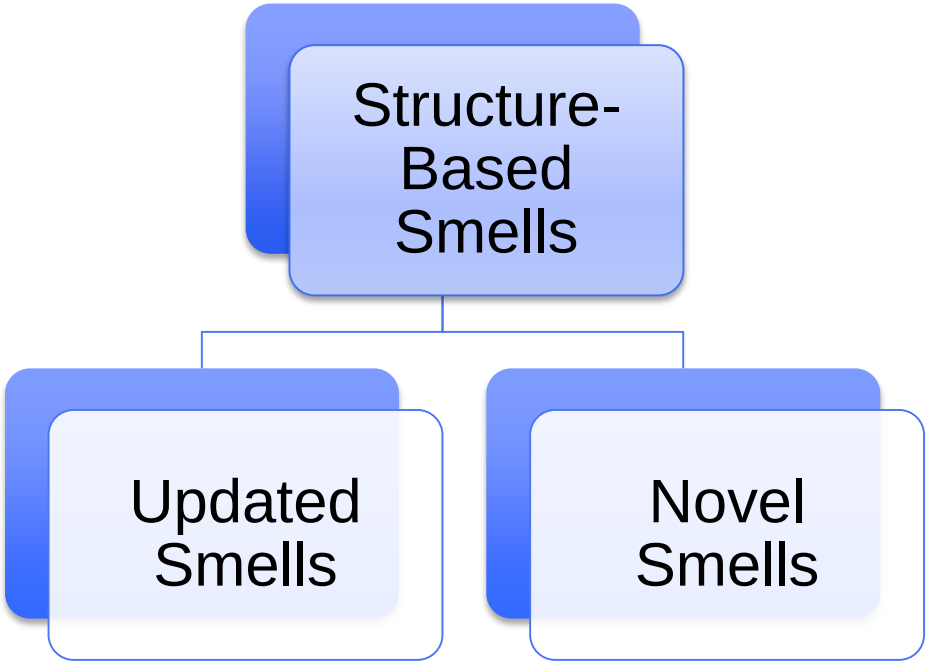
Outline


Motivation
Structural Analysis Process
Evaluation
Structure-based Smells
Future Work

Structure-based Smells

- Application of structural information

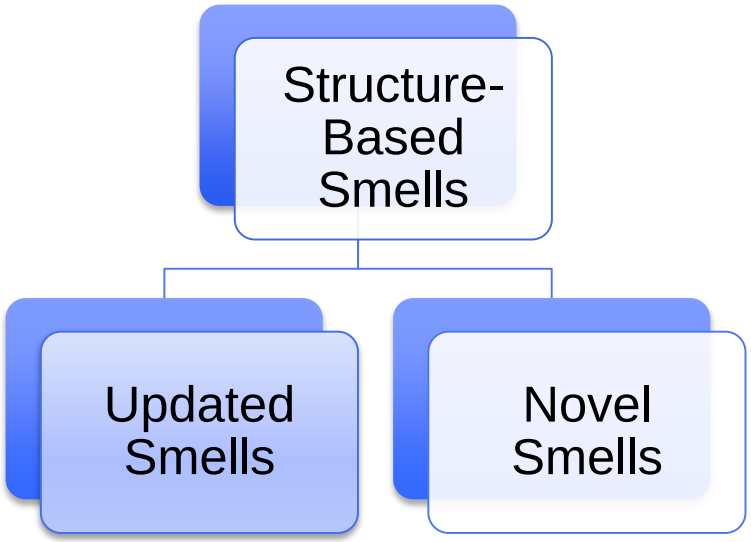
- Focus on enhancement of Spreadsheet Smells



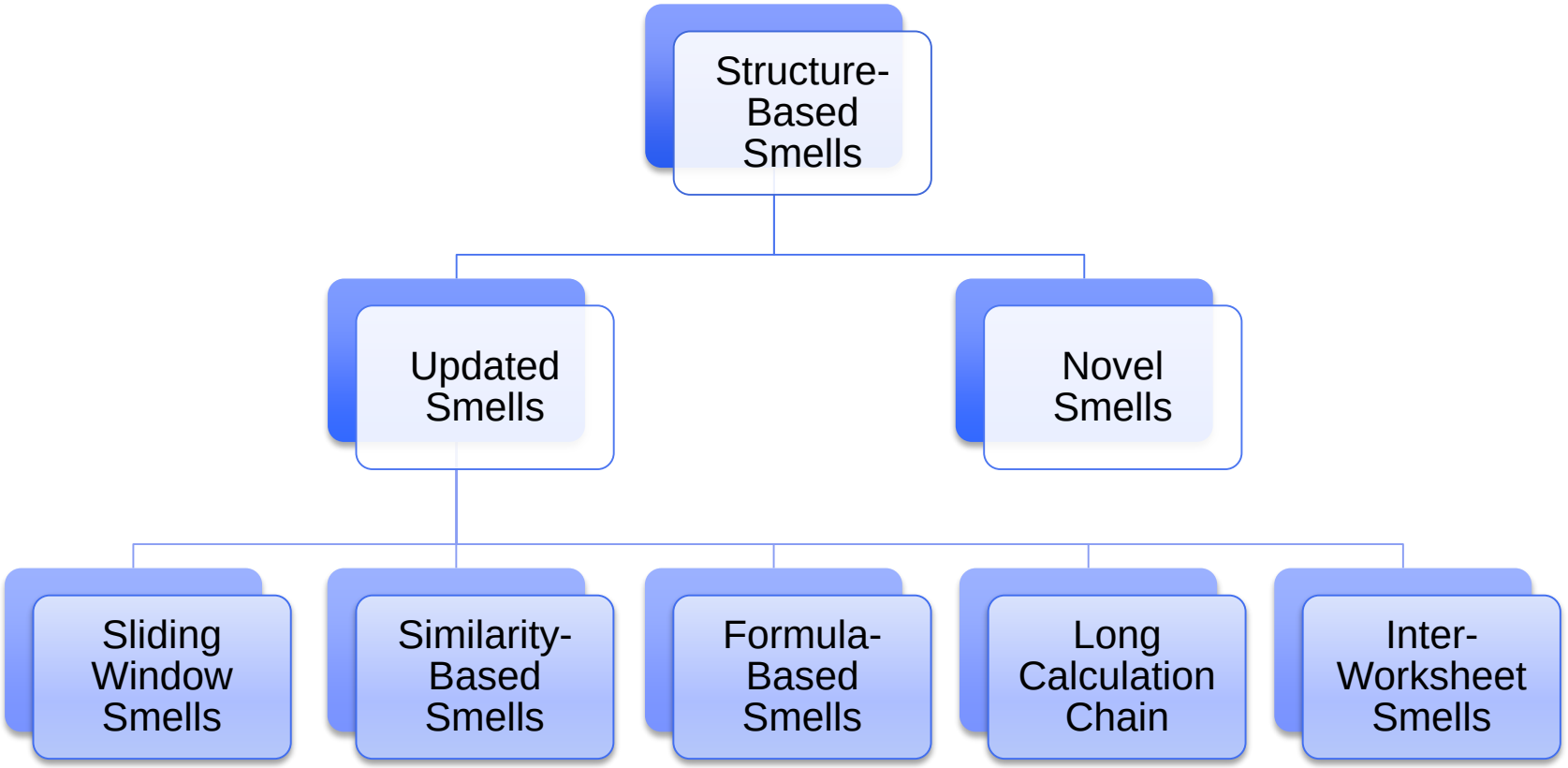
Structure-based Smells: Updated Smells

Update opportunities:

- Focus analysis methods on spreadsheet structures
- Analyse group formulas instead of cell fromulas
- Analyse group references instead of cell references



Structure-based Smells: Updated Smells



Structure-based Smells: Updated Smells

Example: Sliding-Window Smells

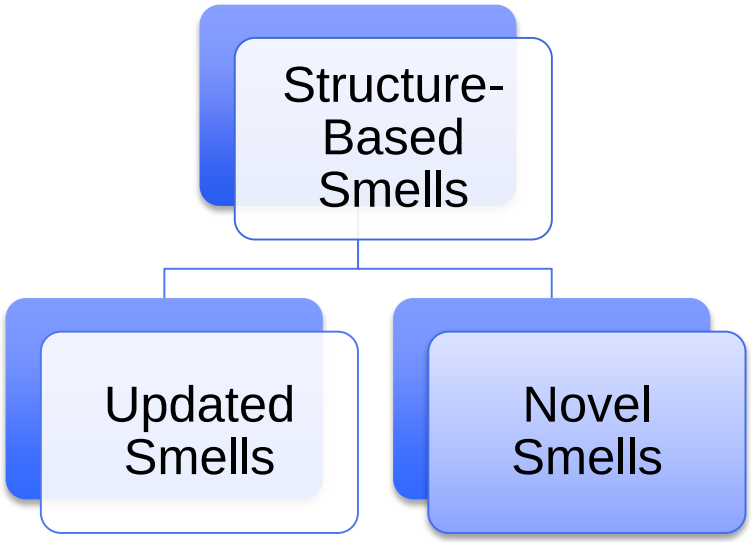
- Detects anomalies in sliding windows
- Update: limit windows to structures

	A	B	C
1		Europe	
2	Models	2012	2013
3	Honda	30	27
4	Mazda	1000	12
5	Fiat	9	12
6	Total	=SUM(B3:B5)	=SUM(C3:C5)

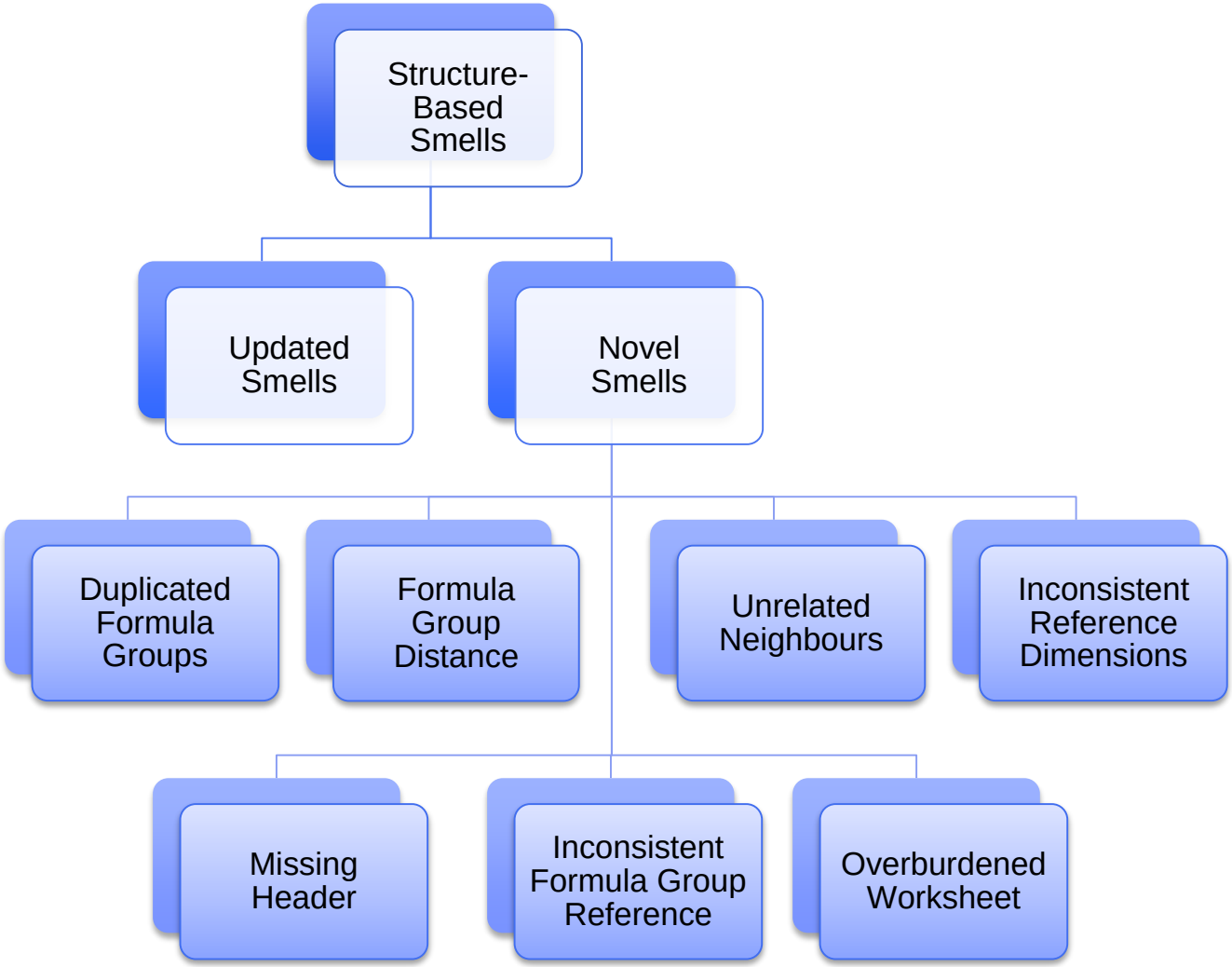
Structure-based Smells: Novel Smells

Smell origins:

- Similar existing smells
- Analyse group formulas instead of cell fromulas
- Analyse group references instead of cell references



Structure-based Smells: Novel Smells



Spreadsheet Smells: Formula Smells

Example: Inconsistent Formula Group Reference

- Size mismatch between groups and group references

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	1000	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B4)	=SUM(C3:C4)	=SUM(D3:D4)	=SUM(E3:E4)	=SUM(F3:F4)

Outline


Motivation
Structural Analysis Process
Evaluation
Structure-based Smells
Future Work

Future Work

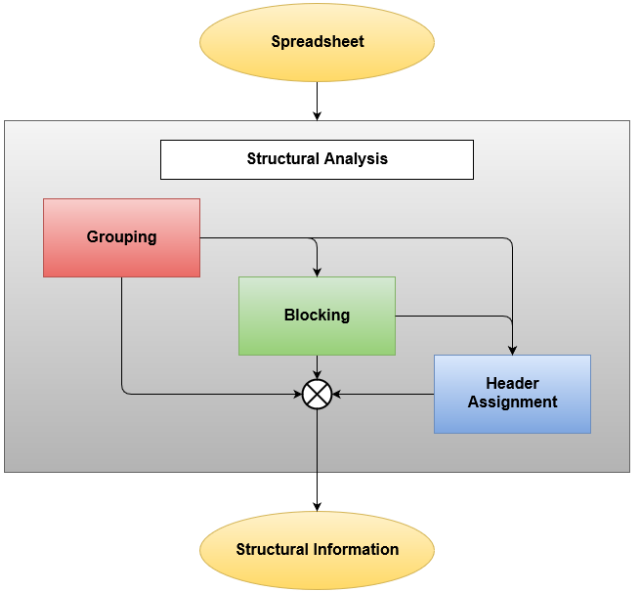
- Further improve spreadsheet QA techniques
  - Unit inference approaches
  - Introduce group-editing operations
  - Introduce knowledge-based approaches

- Expand structural analysis approach
  - Better adapt to special circumstances
  - Introduce new group types
  - Further analyze inter-group dependencies

Structural Analysis Process

What we want to detect

	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)
7						
8	header		input		formula	



Summary

Evaluation: Results EUSES