Static Spreadsheet Analysis

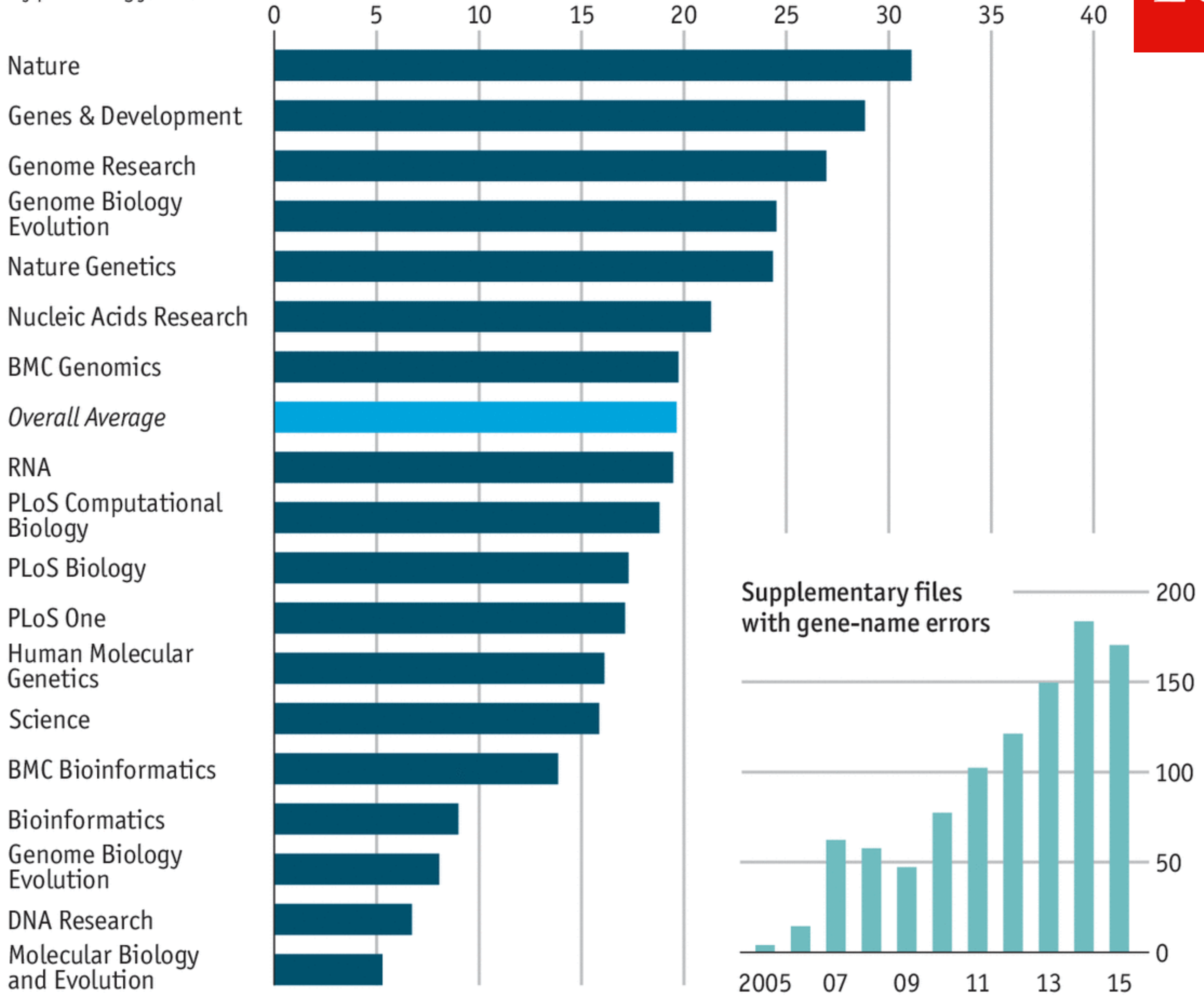
Partick W. Koch, Birgit Hofer, and Franz Wotawa

Debugging of Spreadsheets (DEOS) Project
Funded by the Austrian Science Fund (FWF, contract I2144) and the Deutsche Forschungsgemeinschaft (DFG,contract JA 2095/4-1)

#VALUE! error

Genomics papers with spreadsheet errors in supplementary files, 2005-15, %

By publishing journal



Source: “Gene name errors are now widespread in the scientific literature”, Ziemann, Eren and El-Osta, 2016

Economist.com

The Economist

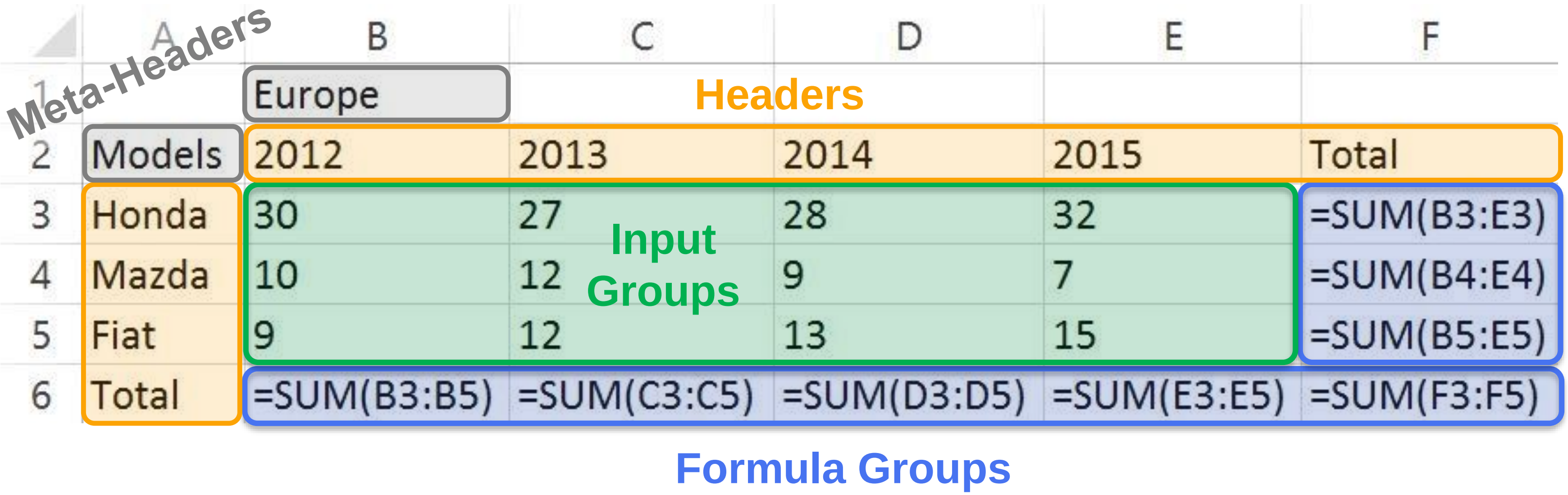
"…the number of genomics papers packaged with error-ridden spreadsheets is **increasing by 15%** a year, far above the 4% annual growth rate in the number of genomics papers published. If we extrapolate current trends …, then **by 2025** every spreadsheet attached to a genetics paper will have an error—unless, of course, there is an error in the spreadsheet we used for this calculation"

The Economist Daily Chart, Sept 7th 2016

Overview

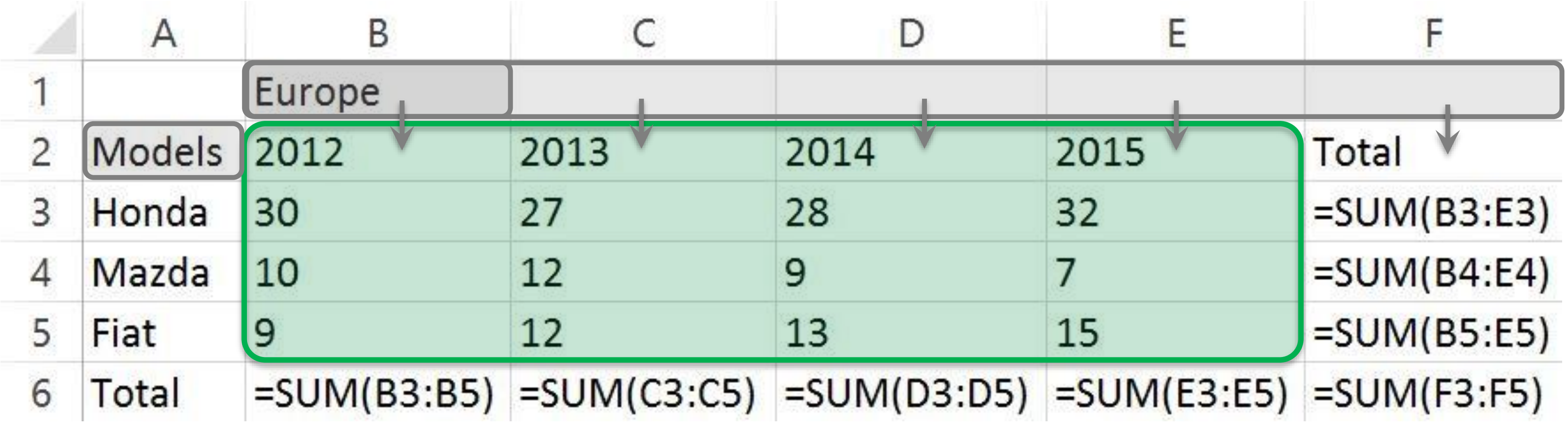
- Running Example
  - Structures that should be detected
- Static Analysis Approach
- Evaluation
- Application
  - Spreadsheet smells

We want to detect …



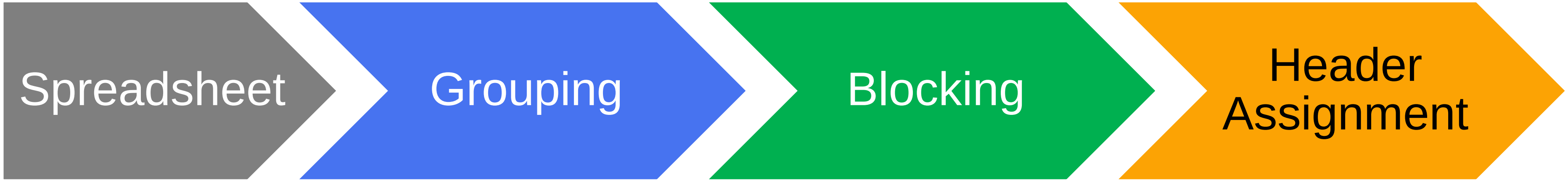
	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

Challenges

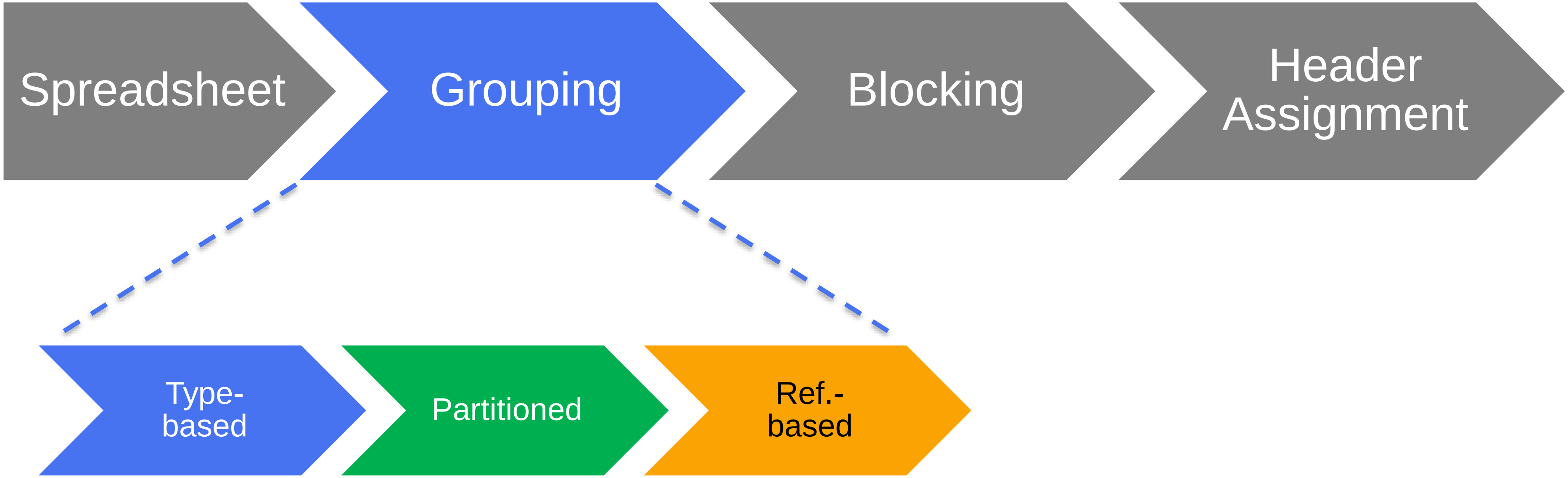


	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

Structural analysis process



Structural analysis process

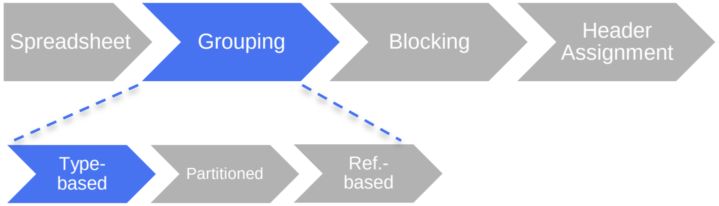


- Cell type
- Neighbors
- Same formula

- 1-dimensional formula groups
- Few groups

- Cells referenced by partitioned group
- Merge overlapping groups with same orientation

Structural analysis process



Spreadsheet

Grouping

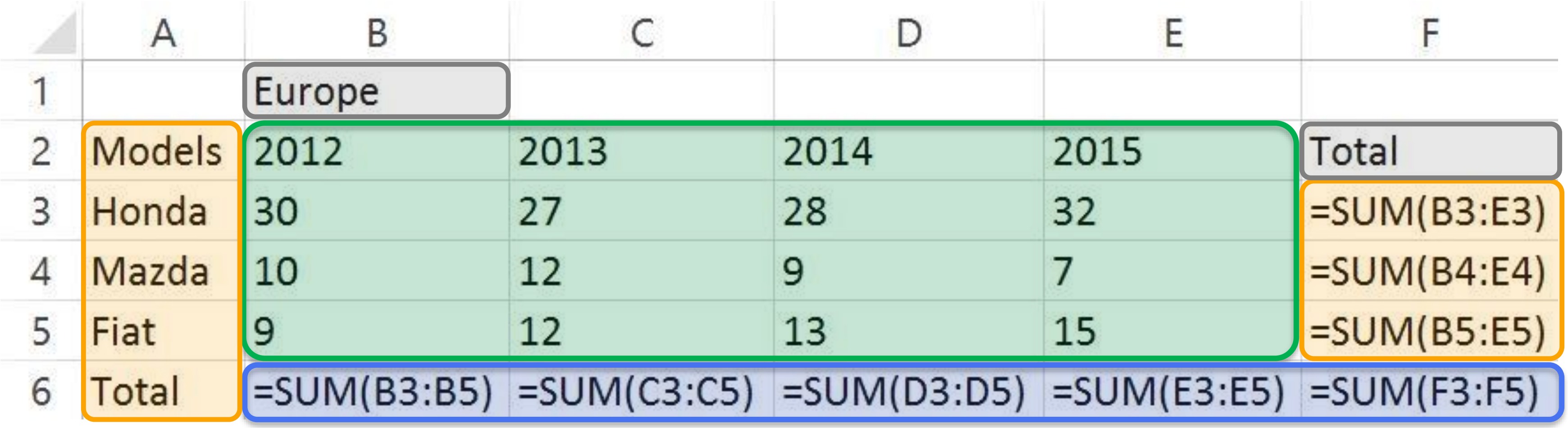
Blocking

Header Assignment

Type-based

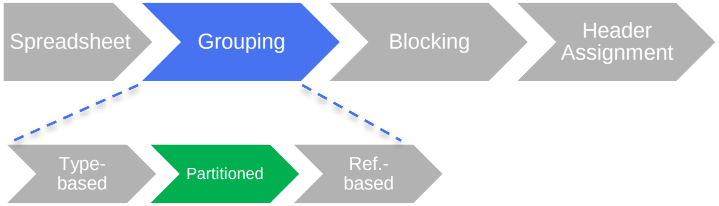
Partitioned

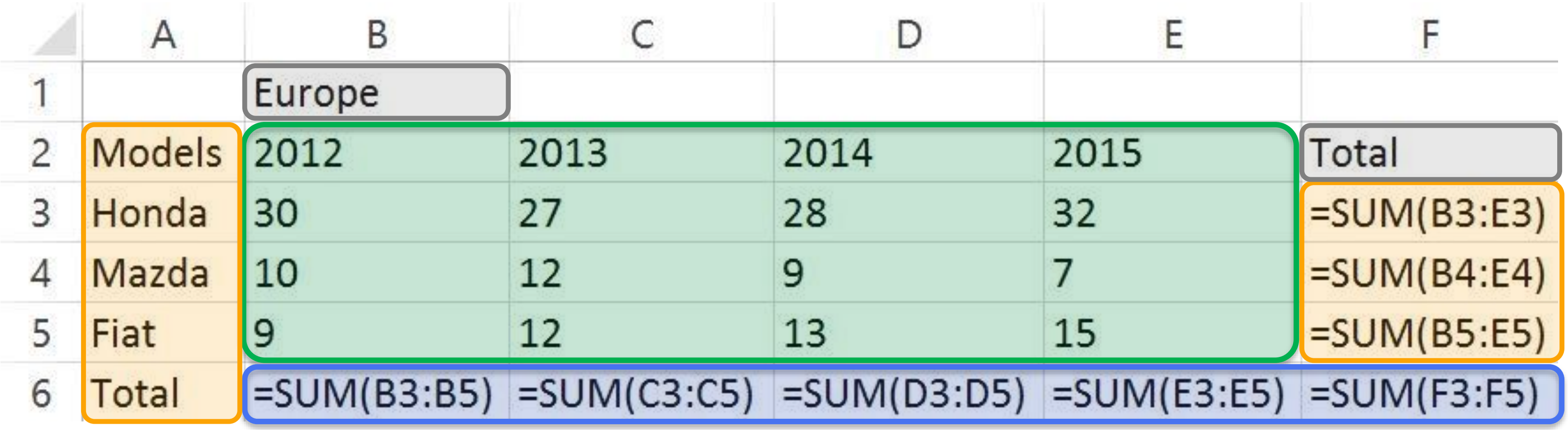
Ref.-based



	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)

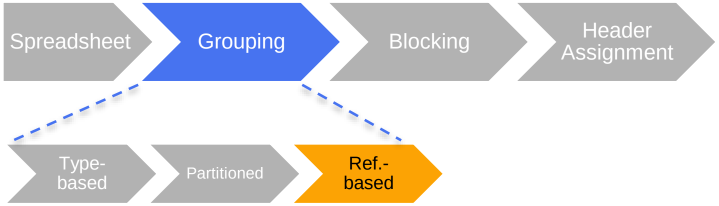
Structural analysis process

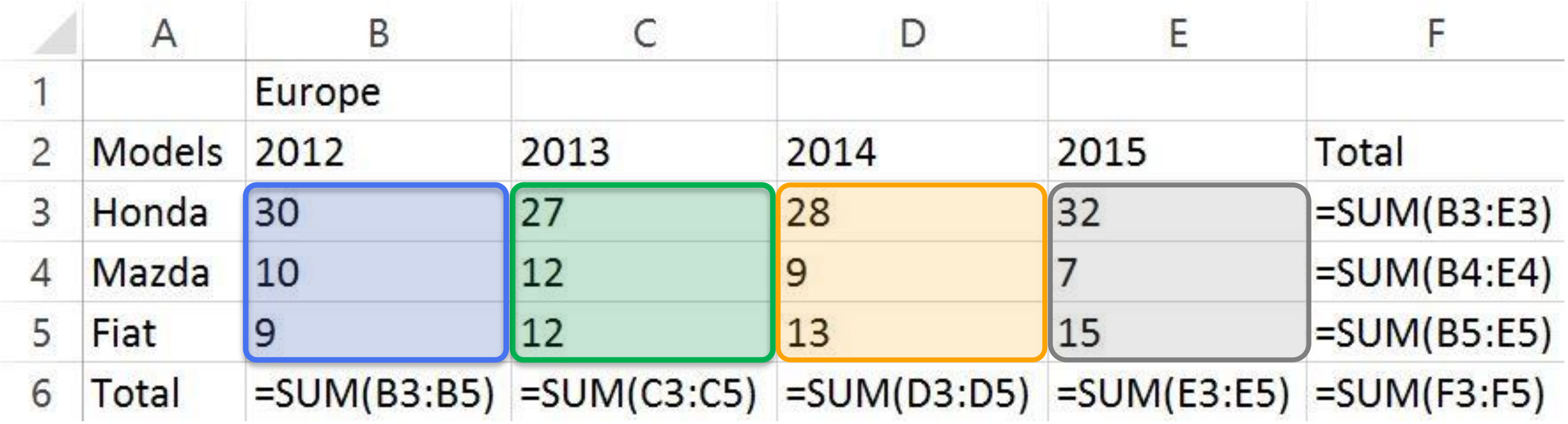




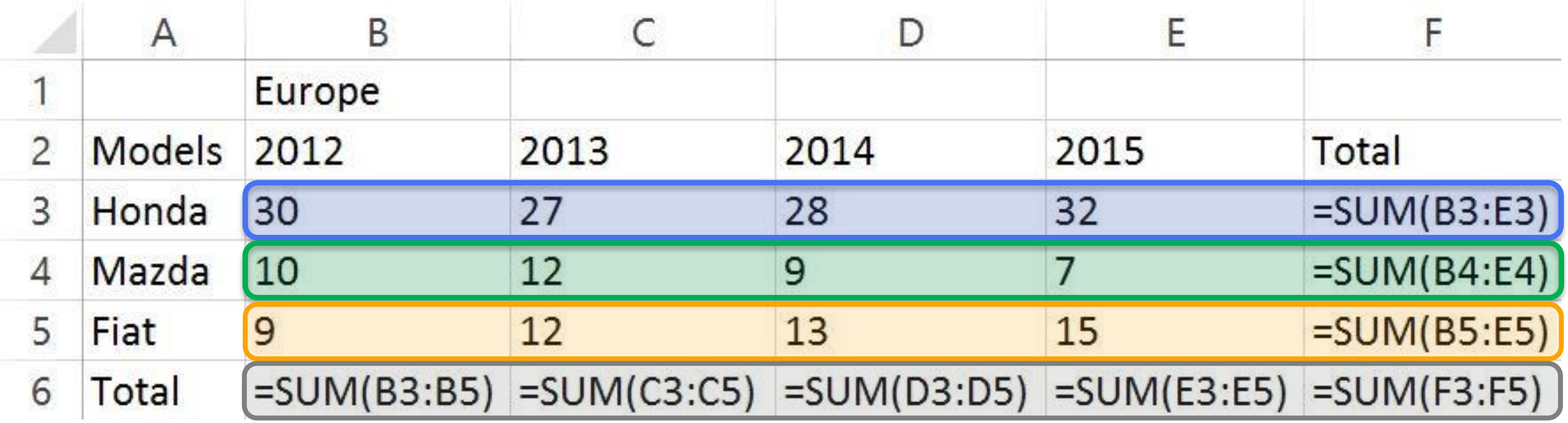
	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)

Structural analysis process



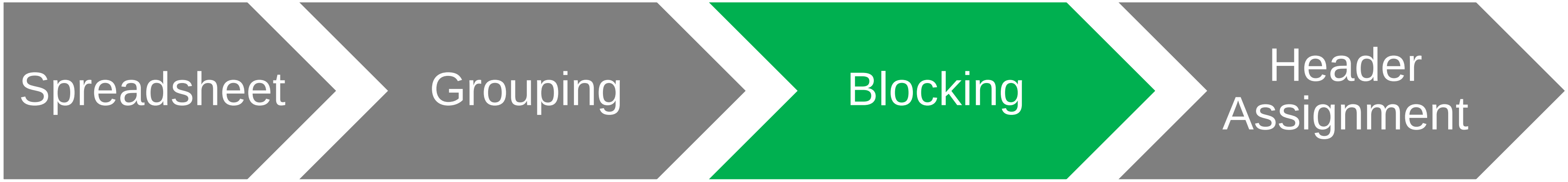


	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)



	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)

Structural analysis process



Spreadsheet

Grouping

Blocking

Header Assignment

What is a block?

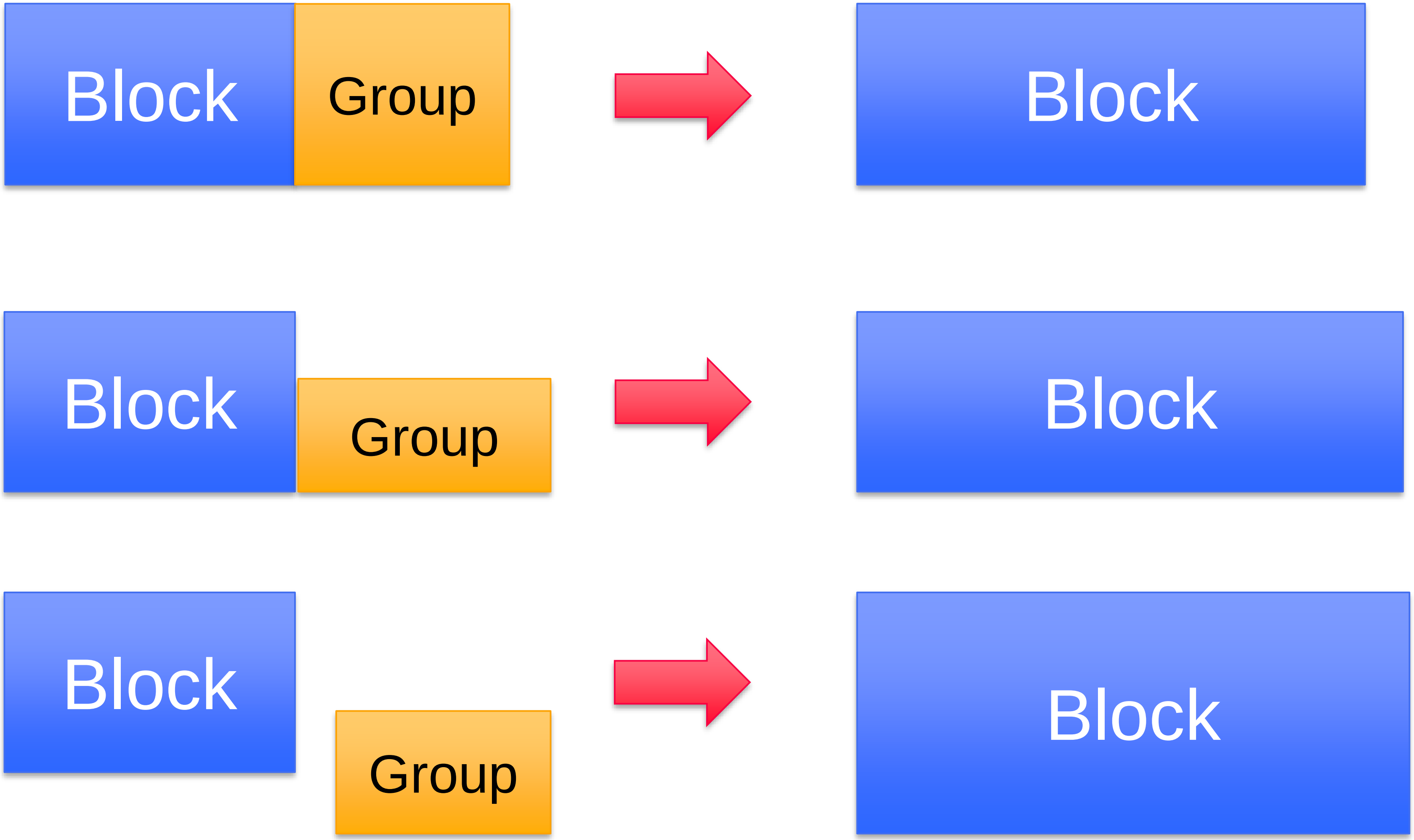
- Rectangular area
- Aggregate input and formula cells, NOT header cells

How to create a block?

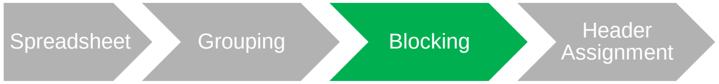
1. Pick formula-group or referenced group
2. Expand

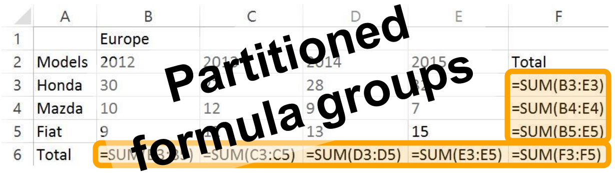
Structural Analysis Process

Spreadsheet → Grouping → Blocking → Header Assignment

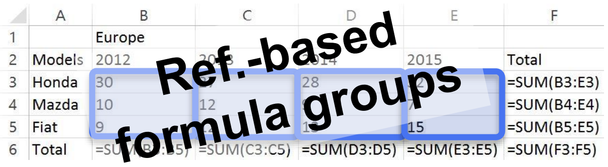


Structural analysis process

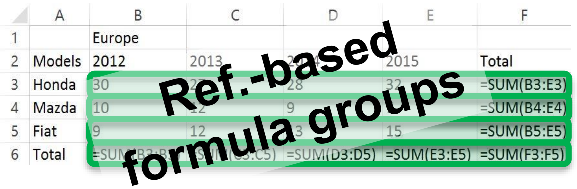




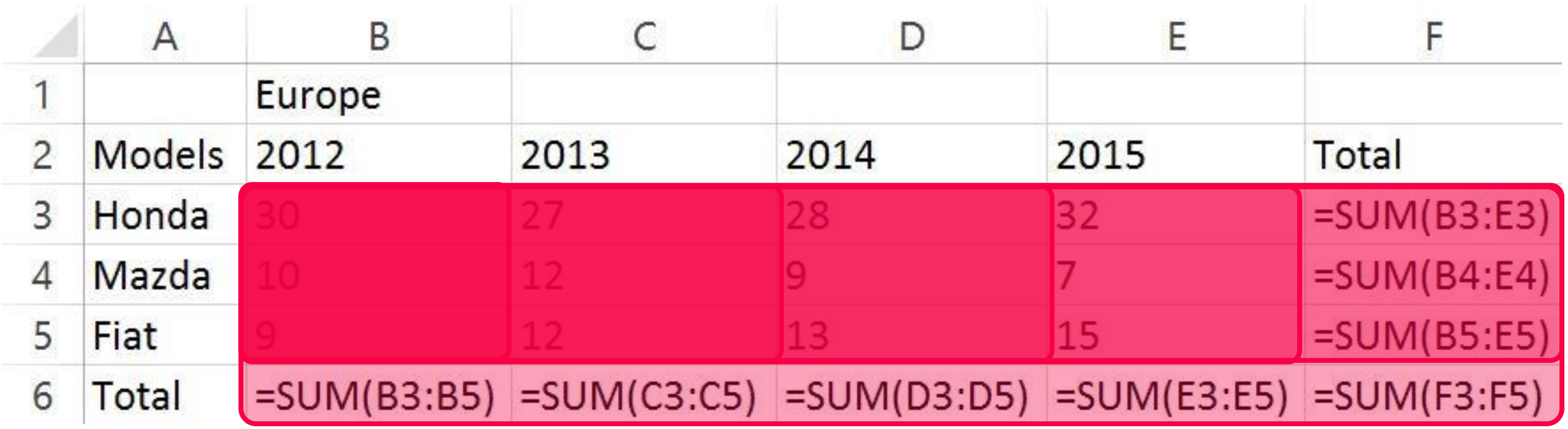
	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)



	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)

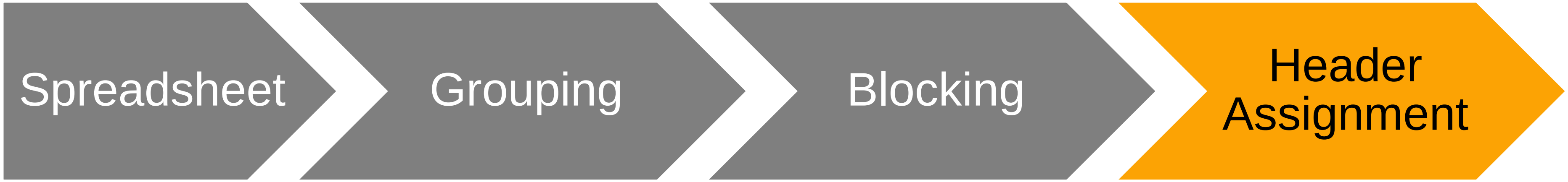


	A	B	C	D	E	F
1		Europe				
2	Models	2012	2013	2014	2015	Total
3	Honda	30	27	28	32	=SUM(B3:E3)
4	Mazda	10	12	9	7	=SUM(B4:E4)
5	Fiat	9	12	13	15	=SUM(B5:E5)
6	Total	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)	=SUM(E3:E5)	=SUM(F3:F5)



	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)

Structural analysis process



Spreadsheet

Grouping

Blocking

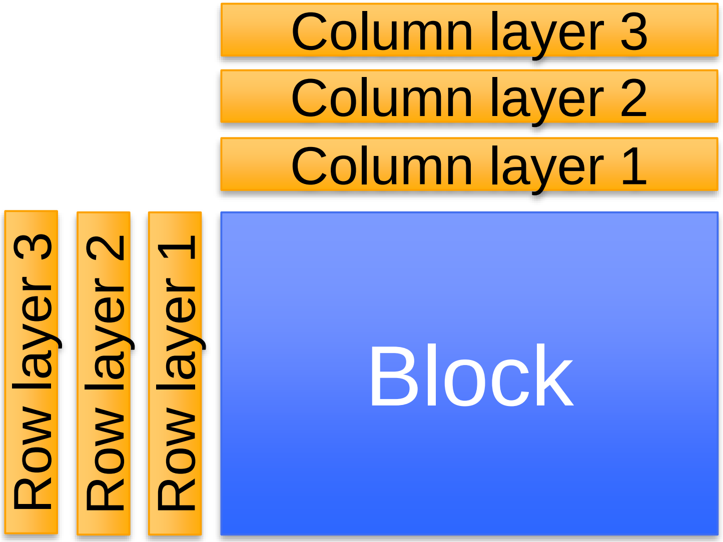
Header Assignment

Position of headers

- Above blocks
- LTR vs. RTL systems
- Column vs. row headers
- Headers of headers

Identification

- Identify header layers



Column layer 3

Column layer 2

Column layer 1

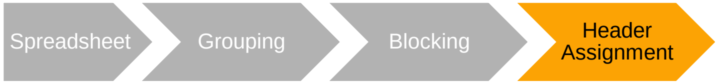
Row layer 3

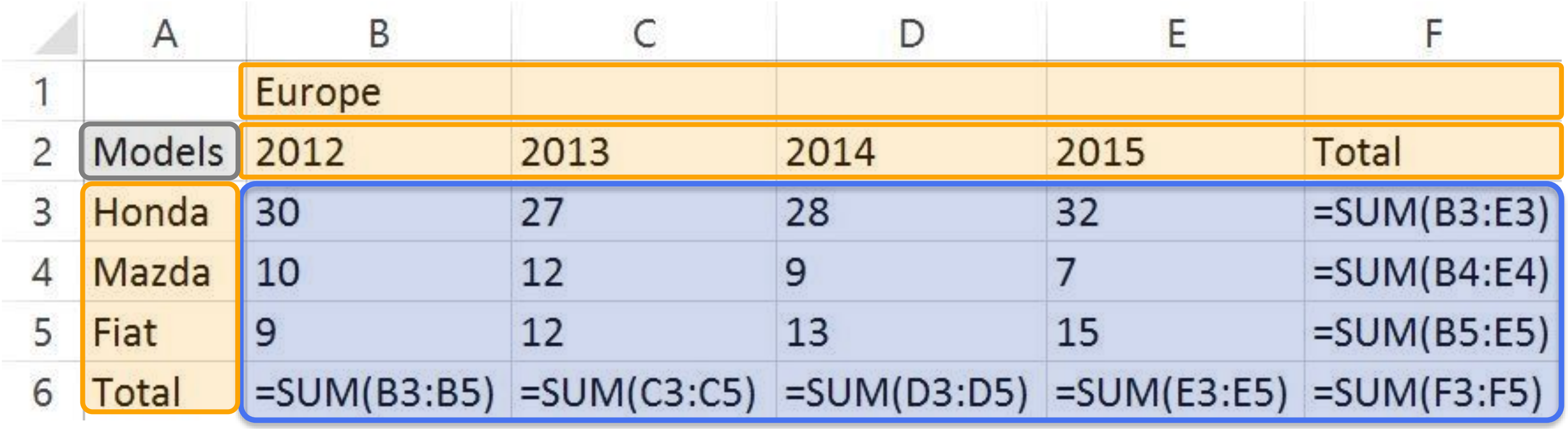
Row layer 2

Row layer 1

Block

Structural analysis process


Spreadsheet
Grouping
Blocking
Header Assignment



	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)

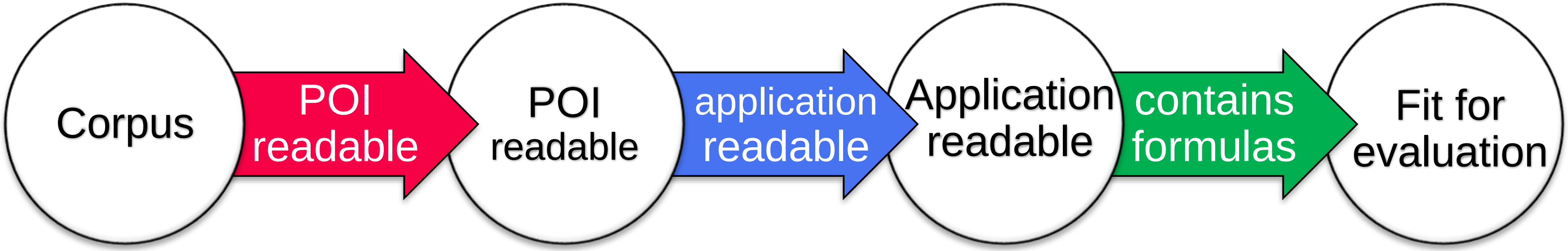
Next step

- Identify remaining meta-header

Future work

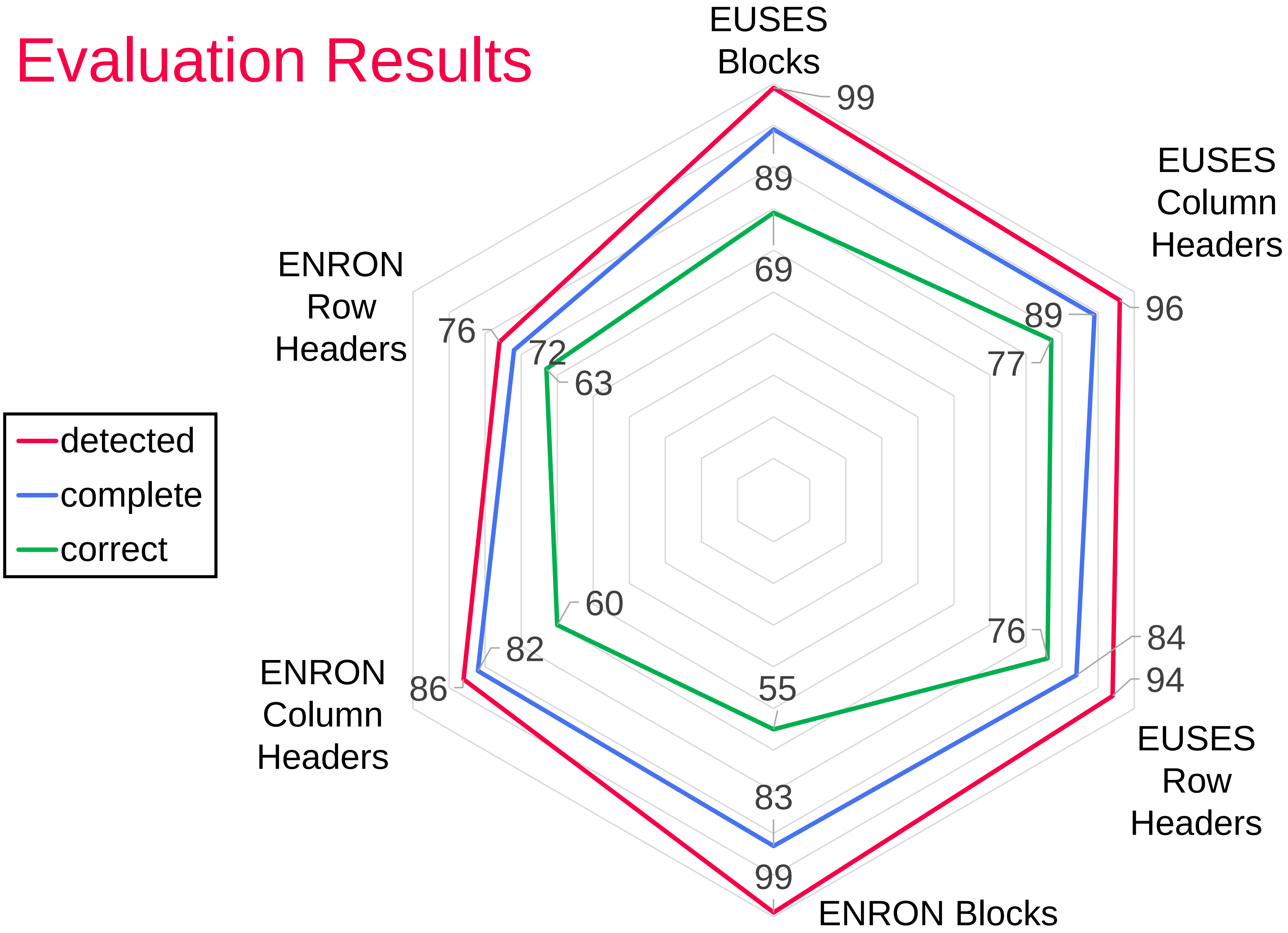
- Better mapping approach for meta headers

Evaluation

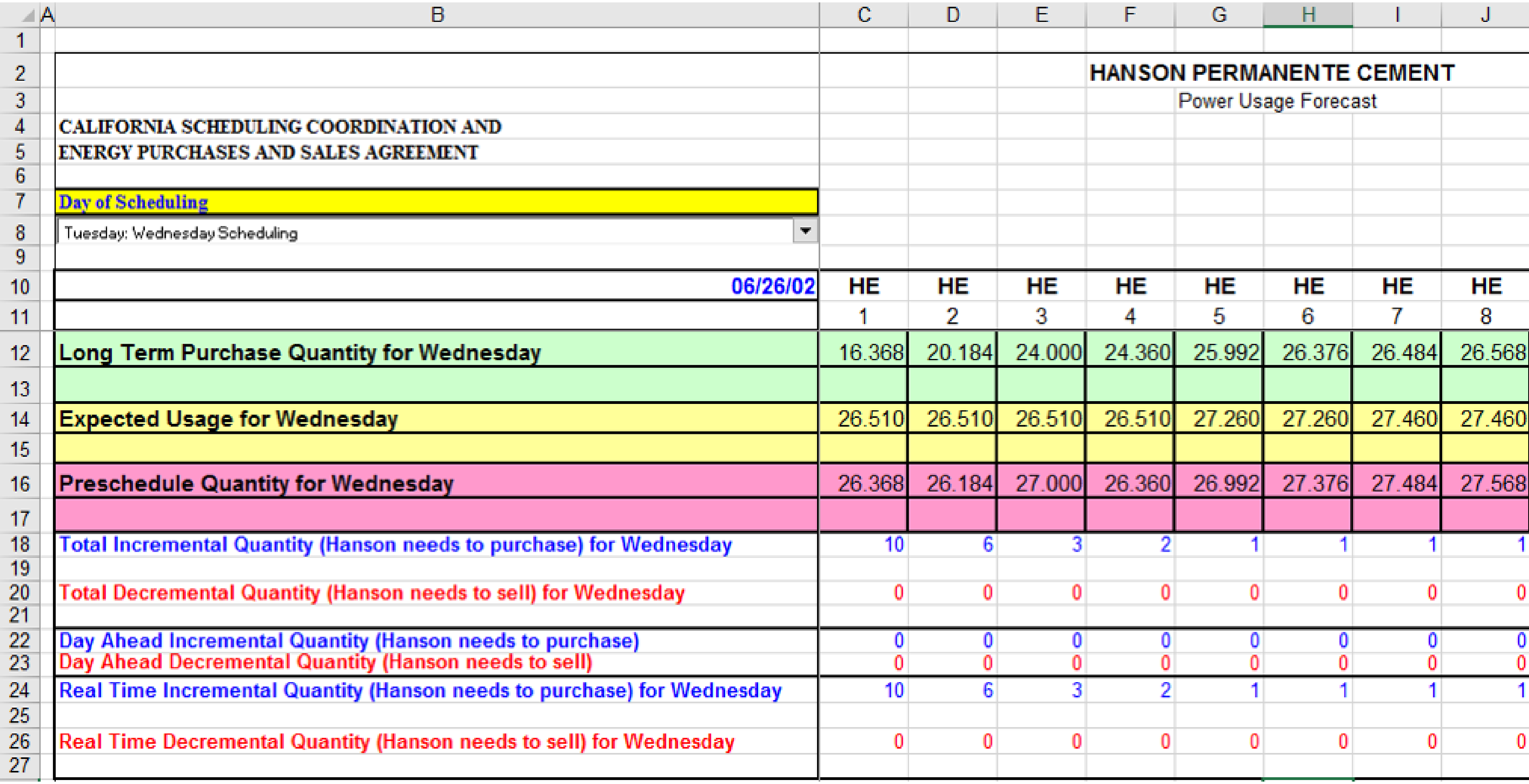


EUSES	4,495	4,172	4,017	1,659
ENRON	15,929	15,929	13,402	6,691

Evaluation Results

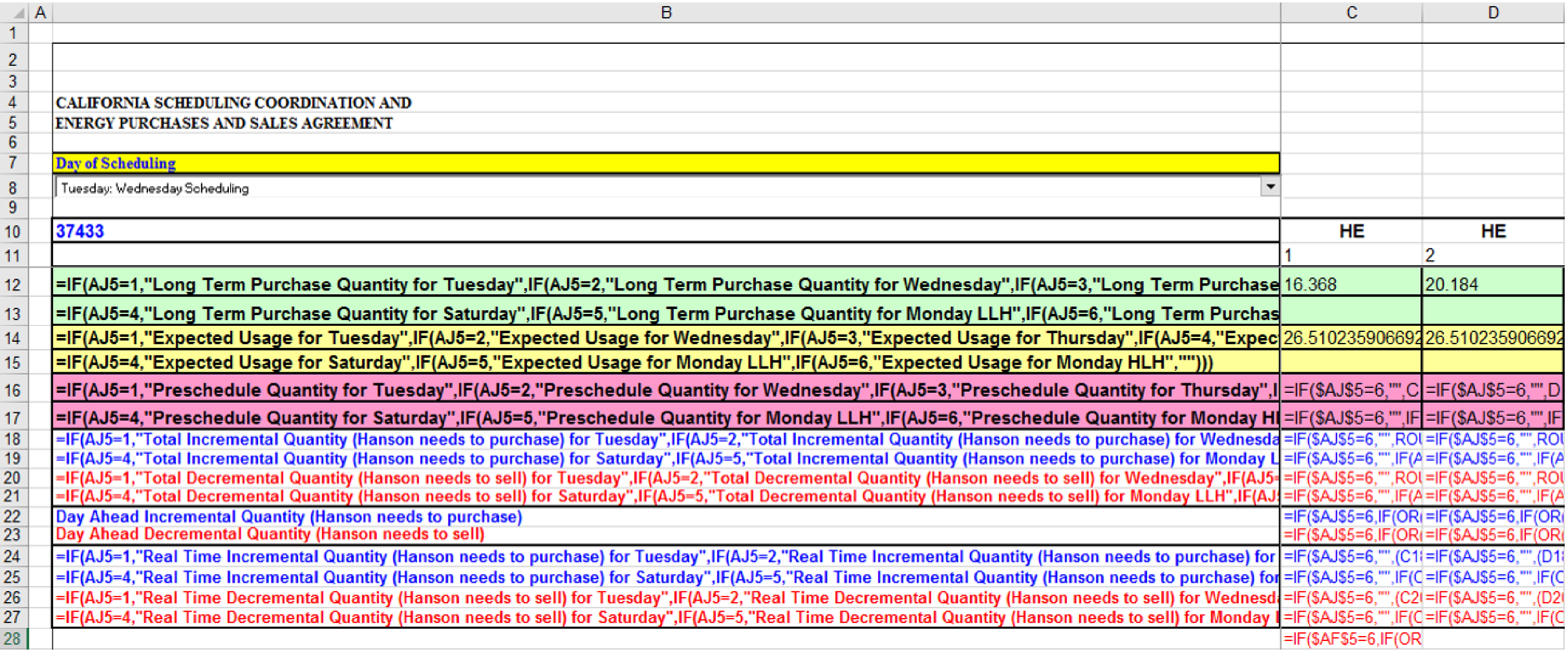


Open Problems



	A	B	C	D	E	F	G	H	I	J
1										
2						HANSON PERMANENTE CEMENT				
3							Power Usage Forecast			
4		CALIFORNIA SCHEDULING COORDINATION AND								
5		ENERGY PURCHASES AND SALES AGREEMENT								
6										
7		Day of Scheduling								
8		Tuesday: Wednesday Scheduling ▾								
9										
10		06/26/02	HE	HE	HE	HE	HE	HE	HE	HE
11			1	2	3	4	5	6	7	8
12		Long Term Purchase Quantity for Wednesday	16.368	20.184	24.000	24.360	25.992	26.376	26.484	26.568
13										
14		Expected Usage for Wednesday	26.510	26.510	26.510	26.510	27.260	27.260	27.460	27.460
15										
16		Preschedule Quantity for Wednesday	26.368	26.184	27.000	26.360	26.992	27.376	27.484	27.568
17										
18		Total Incremental Quantity (Hanson needs to purchase) for Wednesday	10	6	3	2	1	1	1	1
19										
20		Total Decremental Quantity (Hanson needs to sell) for Wednesday	0	0	0	0	0	0	0	0
21										
22		Day Ahead Incremental Quantity (Hanson needs to purchase)	0	0	0	0	0	0	0	0
23		Day Ahead Decremental Quantity (Hanson needs to sell)	0	0	0	0	0	0	0	0
24		Real Time Incremental Quantity (Hanson needs to purchase) for Wednesday	10	6	3	2	1	1	1	1
25										
26		Real Time Decremental Quantity (Hanson needs to sell) for Wednesday	0	0	0	0	0	0	0	0
27										

Open Problems



	A	B	C	D
1				
2				
3				
4		CALIFORNIA SCHEDULING COORDINATION AND		
5		ENERGY PURCHASES AND SALES AGREEMENT		
6				
7		Day of Scheduling		
8		Tuesday: Wednesday Scheduling ▾		
9				
10		37433	HE	HE
11			1	2
12		=IF(AJ5=1,"Long Term Purchase Quantity for Tuesday",IF(AJ5=2,"Long Term Purchase Quantity for Wednesday",IF(AJ5=3,"Long Term Purchase	16.368	20.184
13		=IF(AJ5=4,"Long Term Purchase Quantity for Saturday",IF(AJ5=5,"Long Term Purchase Quantity for Monday LLH",IF(AJ5=6,"Long Term Purchas		
14		=IF(AJ5=1,"Expected Usage for Tuesday",IF(AJ5=2,"Expected Usage for Wednesday",IF(AJ5=3,"Expected Usage for Thursday",IF(AJ5=4,"Expec	26.510235906692	26.510235906692
15		=IF(AJ5=4,"Expected Usage for Saturday",IF(AJ5=5,"Expected Usage for Monday LLH",IF(AJ5=6,"Expected Usage for Monday HLH","")))		
16		=IF(AJ5=1,"Preschedule Quantity for Tuesday",IF(AJ5=2,"Preschedule Quantity for Wednesday",IF(AJ5=3,"Preschedule Quantity for Thursday",I	=IF(AJ5=6,"",C	=IF(AJ5=6,"",D
17		=IF(AJ5=4,"Preschedule Quantity for Saturday",IF(AJ5=5,"Preschedule Quantity for Monday LLH",IF(AJ5=6,"Preschedule Quantity for Monday H	=IF(AJ5=6,"",IF	=IF(AJ5=6,"",IF
18		=IF(AJ5=1,"Total Incremental Quantity (Hanson needs to purchase) for Tuesday",IF(AJ5=2,"Total Incremental Quantity (Hanson needs to purchase) for Wednesda	=IF(AJ5=6,"",ROU	=IF(AJ5=6,"",ROU
19		=IF(AJ5=4,"Total Incremental Quantity (Hanson needs to purchase) for Saturday",IF(AJ5=5,"Total Incremental Quantity (Hanson needs to purchase) for Monday L	=IF(AJ5=6,"",IF(A	=IF(AJ5=6,"",IF(A
20		=IF(AJ5=1,"Total Decremental Quantity (Hanson needs to sell) for Tuesday",IF(AJ5=2,"Total Decremental Quantity (Hanson needs to sell) for Wednesday",IF(AJ5=	=IF(AJ5=6,"",ROU	=IF(AJ5=6,"",ROU
21		=IF(AJ5=4,"Total Decremental Quantity (Hanson needs to sell) for Saturday",IF(AJ5=5,"Total Decremental Quantity (Hanson needs to sell) for Monday LLH",IF(AJ	=IF(AJ5=6,"",IF(A	=IF(AJ5=6,"",IF(A
22		Day Ahead Incremental Quantity (Hanson needs to purchase)	=IF(AJ5=6,IF(OR	=IF(AJ5=6,IF(OR
23		Day Ahead Decremental Quantity (Hanson needs to sell)	=IF(AJ5=6,IF(OR	=IF(AJ5=6,IF(OR
24		=IF(AJ5=1,"Real Time Incremental Quantity (Hanson needs to purchase) for Tuesday",IF(AJ5=2,"Real Time Incremental Quantity (Hanson needs to purchase) for	=IF(AJ5=6,"",(C1	=IF(AJ5=6,"",(D1
25		=IF(AJ5=4,"Real Time Incremental Quantity (Hanson needs to purchase) for Saturday",IF(AJ5=5,"Real Time Incremental Quantity (Hanson needs to purchase) for	=IF(AJ5=6,"",IF(C	=IF(AJ5=6,"",IF(C
26		=IF(AJ5=1,"Real Time Decremental Quantity (Hanson needs to sell) for Tuesday",IF(AJ5=2,"Real Time Decremental Quantity (Hanson needs to sell) for Wednesd	=IF(AJ5=6,"",(C2	=IF(AJ5=6,"",(D2
27		=IF(AJ5=4,"Real Time Decremental Quantity (Hanson needs to sell) for Saturday",IF(AJ5=5,"Real Time Decremental Quantity (Hanson needs to sell) for Monday	=IF(AJ5=6,"",IF(C	=IF(AJ5=6,"",IF(C
28			=IF(AF5=6,IF(OR	

Application: Spreadsheet Smells

- Origin: **Code Smells**
- Current state of spreadsheet smells
  - Too many smells reported
  - Analysis time

→ Use structure Information

- To improve existing smells
- To identify new smells

Improve existing smells

Sliding Window Smells

Similarity-Based Smells

Formula-Based Smells

Long Calculation Chain

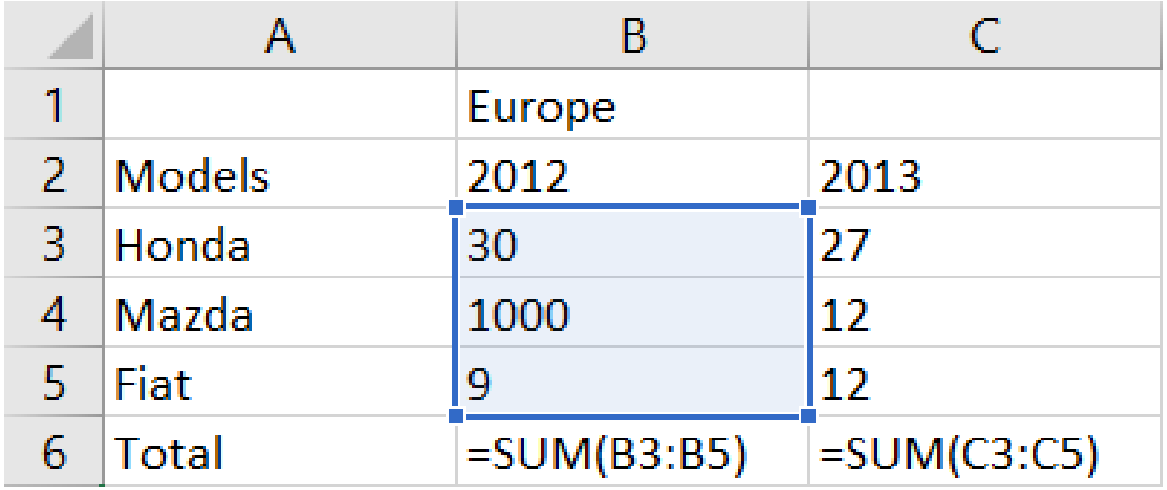
Inter-Worksheet Smells

Update opportunities

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

Example
Sliding Window Smell

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



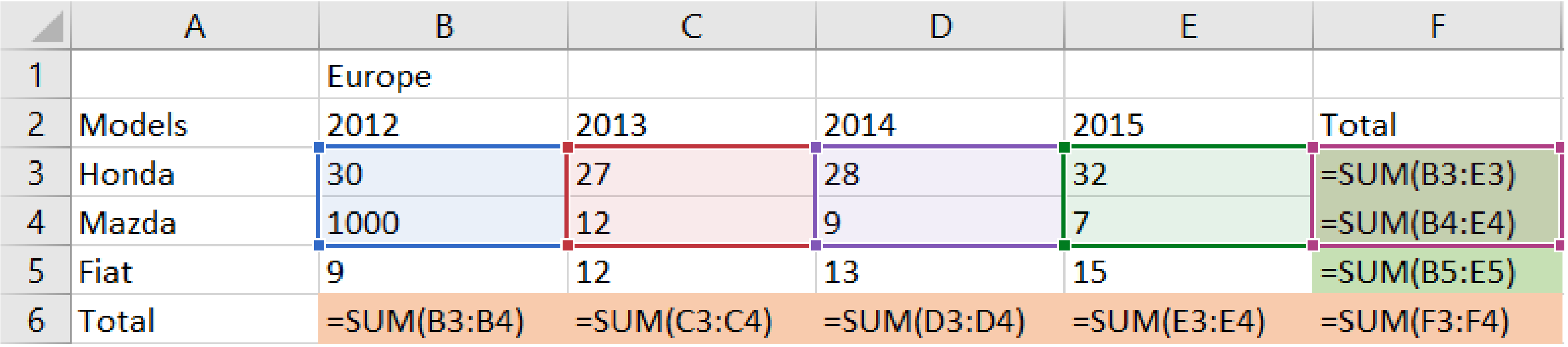
	A	B	C
1		Europe	
2	Models	2012	2013
3	Honda	30	27
4	Mazda	1000	12
5	Fiat	9	12
6	Total	=SUM(B3:B5)	=SUM(C3:C5)

Structure-based Smells: Novel Smells

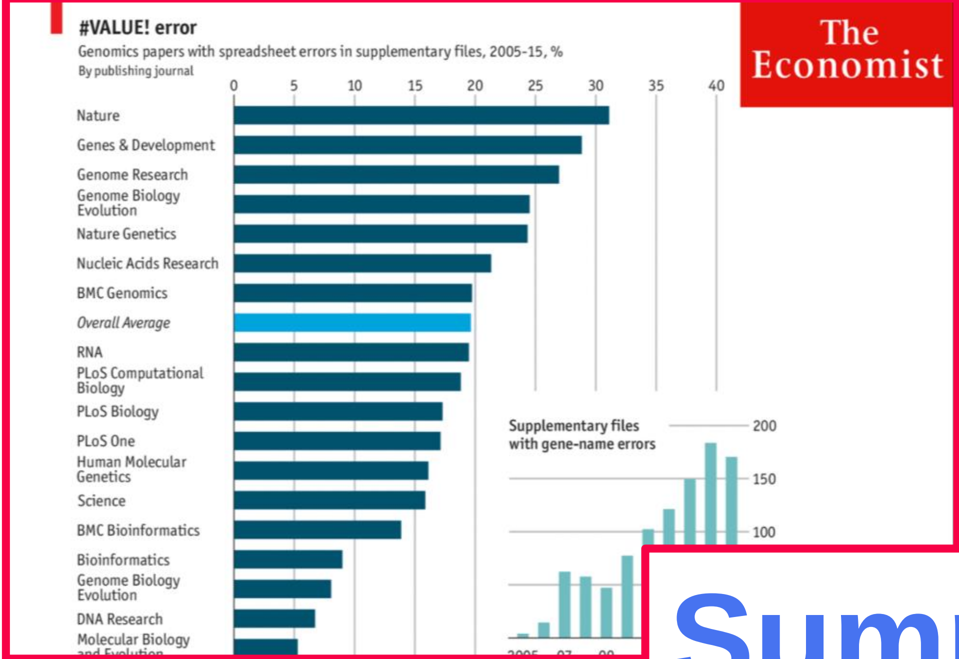
- Duplicated Formula Groups
- Formula Group Distance
- Unrelated Neighbours
- Inconsistent Reference Dimensions
- Inconsistent Formula Group Reference
- Missing Header
- Overburdened Worksheet

Example
Inconsistent Formula Group Reference

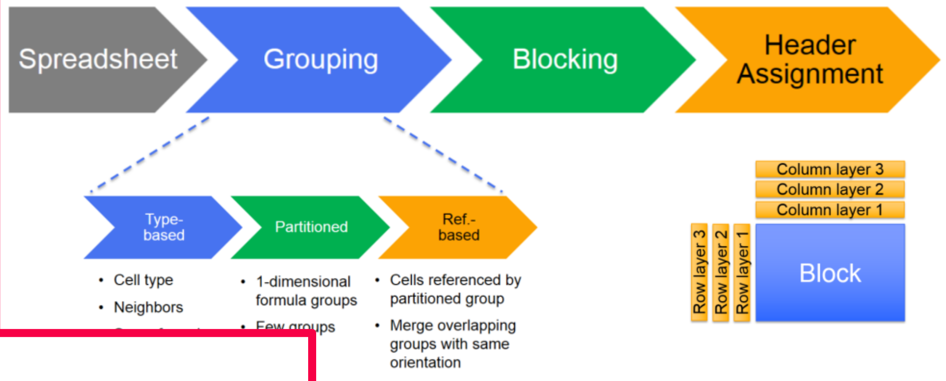
- Size mismatch between groups and group references



	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)

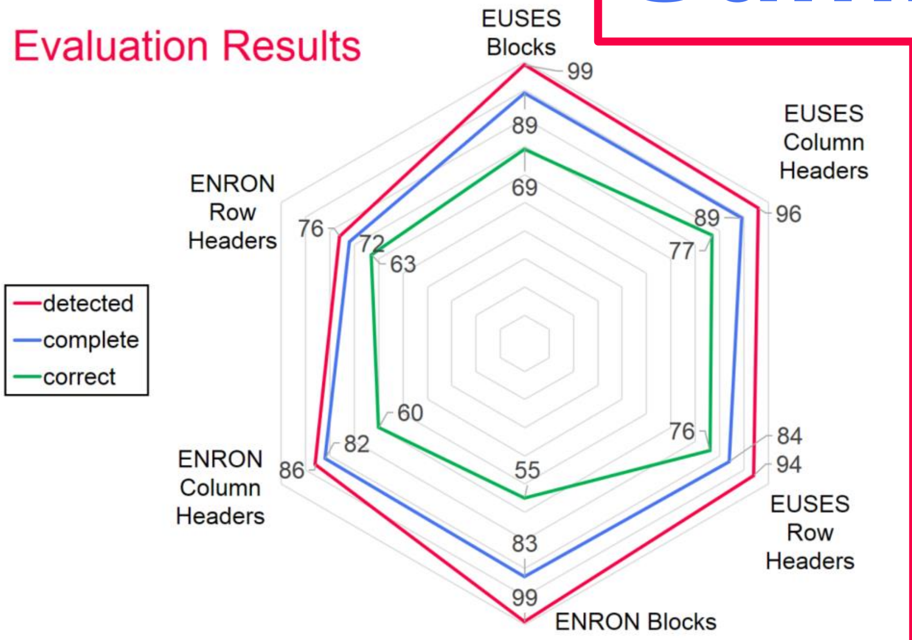


Structural analysis process



Summary

Evaluation Results



Application: Spreadsheet Smells

- Orign: **Code Smells**
- Current state of spreadsheet smells
  - Too many smells reported
  - Analysis time