

Smells & Units:

An Overview of Selected Static Analysis Methods for Spreadsheets

Patrick Koch

2015-03-18



Outline

- Motivation
- Spreadsheet QA Techniques
- Spreadsheet Smells
- Unit-Checking Techniques
- Future Work

Outline

- Motivation
- Spreadsheet QA Techniques
- Spreadsheet Smells
- Unit-Checking Techniques
- Future Work

Motivation

Why spreadsheet Quality Assurance (QA) ?

Let us ask another question first.



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
- Often regarded as „mission-critical“

However...

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

Source:

Panko and Port: “End user computing: The dark matter (and dark energy) of corporate IT”, HICSS 2012

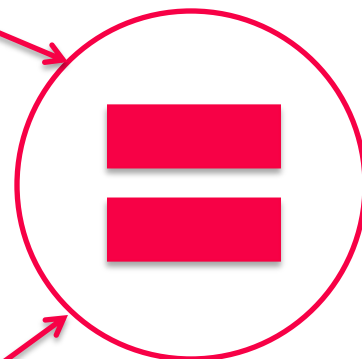
A safe, flexible, and stable Example

	A	S	T	U
3	Multiplier:	1000000	1000000	1000000
4	Currency:	USD	USD	USD
791	2008-05	67335.2	3709.3	481814.0
792	2008-06	67235.6	3729.5	482395.8
793	2008-07	66830.9	3723.3	479220.7
794	2008-08	66386.1	3807.6	477509.8
795	2008-09	65837.3	3798.3	465290.8
796	2008-10	63850.8	3802,0	455194.2
797	2008-11	62661.0	3964.4	448951.2
798	2008-12	62027.9	4247.9	449985.0
799	2009-01	50073.8	4003.6	448257.9
800	2009-02	44605.6	3819.5	438314.7
801	2009-03	45190.7	3670,0	445854.8
802	2009-04	44535.3	3701.1	440687.9
803	2009-05	44090.4	3730.8	437418.2
804	2009-06	43533.4	3729.5	446933.6
805	2009-07	42829.6	3723.3	443879.9
806	2009-08	39669.7	3807.6	441348.2
807	2009-09	38926.6	3798.3	443552.5
808	2009-10	38988.5	3802,0	434859.1
809	2009-11	39019.8	3964.4	429285.0

Source: Tyler Durden: "Blatant Data Error At The Federal Reserve", www.zerohedge.com, 2010.

A not so safe and stable Example

	A	S	T	U
3	Multiplier:	1000000	1000000	1000000
4	Currency:	USD	USD	USD
791	2008-05	67335.2	3709.3	481814.0
792	2008-06	67235.6	3729.5	482395.8
793	2008-07	66830.9	3723.3	479220.7
794	2008-08	66386.1	3807.6	477509.8
795	2008-09	65837.3	3798.3	465290.8
796	2008-10	63850.8	3802,0	455194.2
797	2008-11	62661.0	3964.4	448951.2
798	2008-12	62027.9	4247.9	449985.0
799	2009-01	50073.8	4003.6	448257.9
800	2009-02	44605.6	3819.5	438314.7
801	2009-03	45190.7	3670,0	445854.8
802	2009-04	44535.3	3701.1	440687.9
803	2009-05	44090.4	3730.8	437418.2
804	2009-06	43533.4	3729.5	446933.6
805	2009-07	42829.6	3723.3	443879.9
806	2009-08	39669.7	3807.6	441348.2
807	2009-09	38926.6	3798.3	443552.5
808	2009-10	38988.5	3802,0	434859.1
809	2009-11	39019.8	3964.4	429285.0



Outline

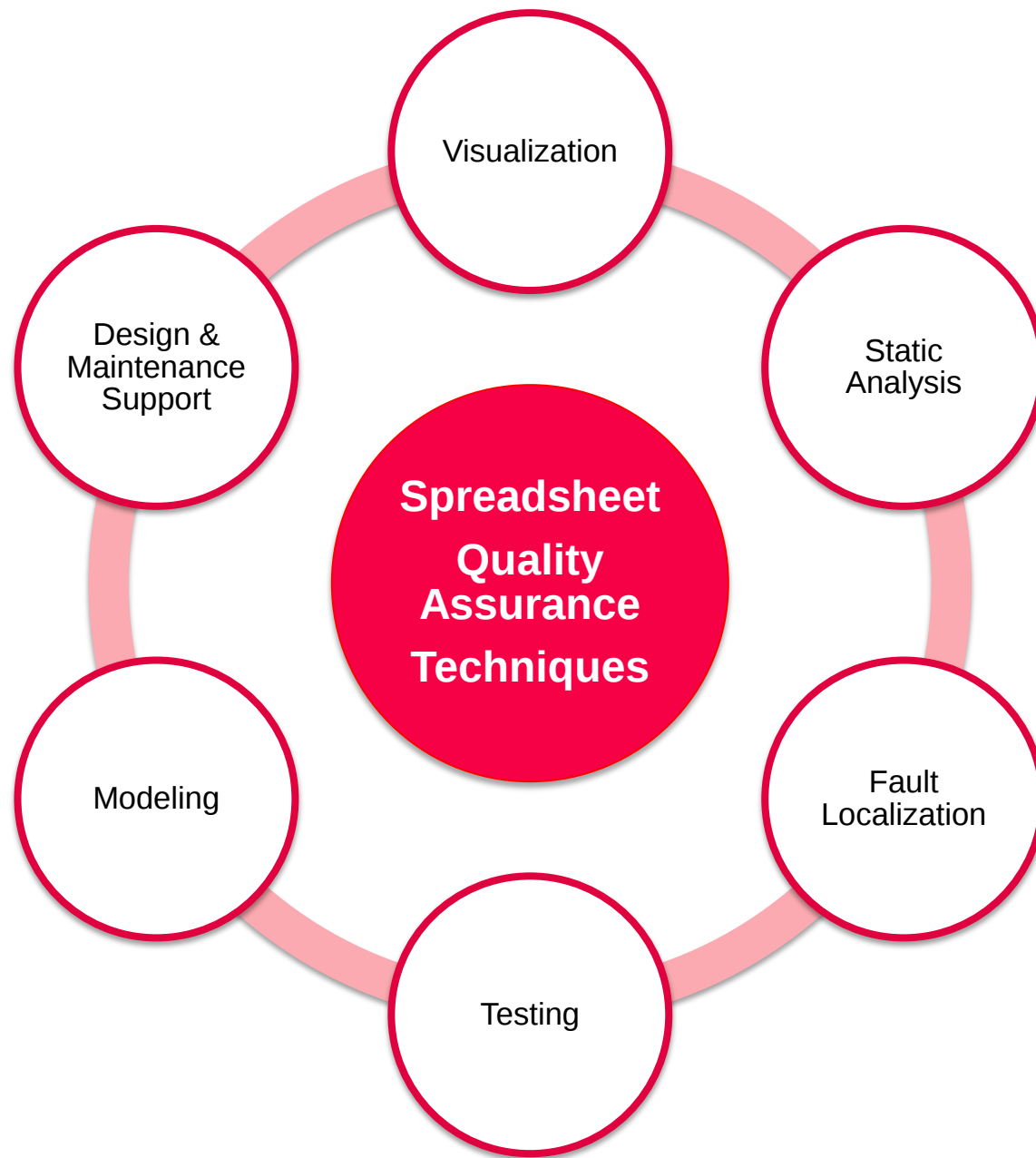
- Motivation
- Spreadsheet QA Techniques
- Spreadsheet Smells
- Unit-Checking Techniques
- Future Work

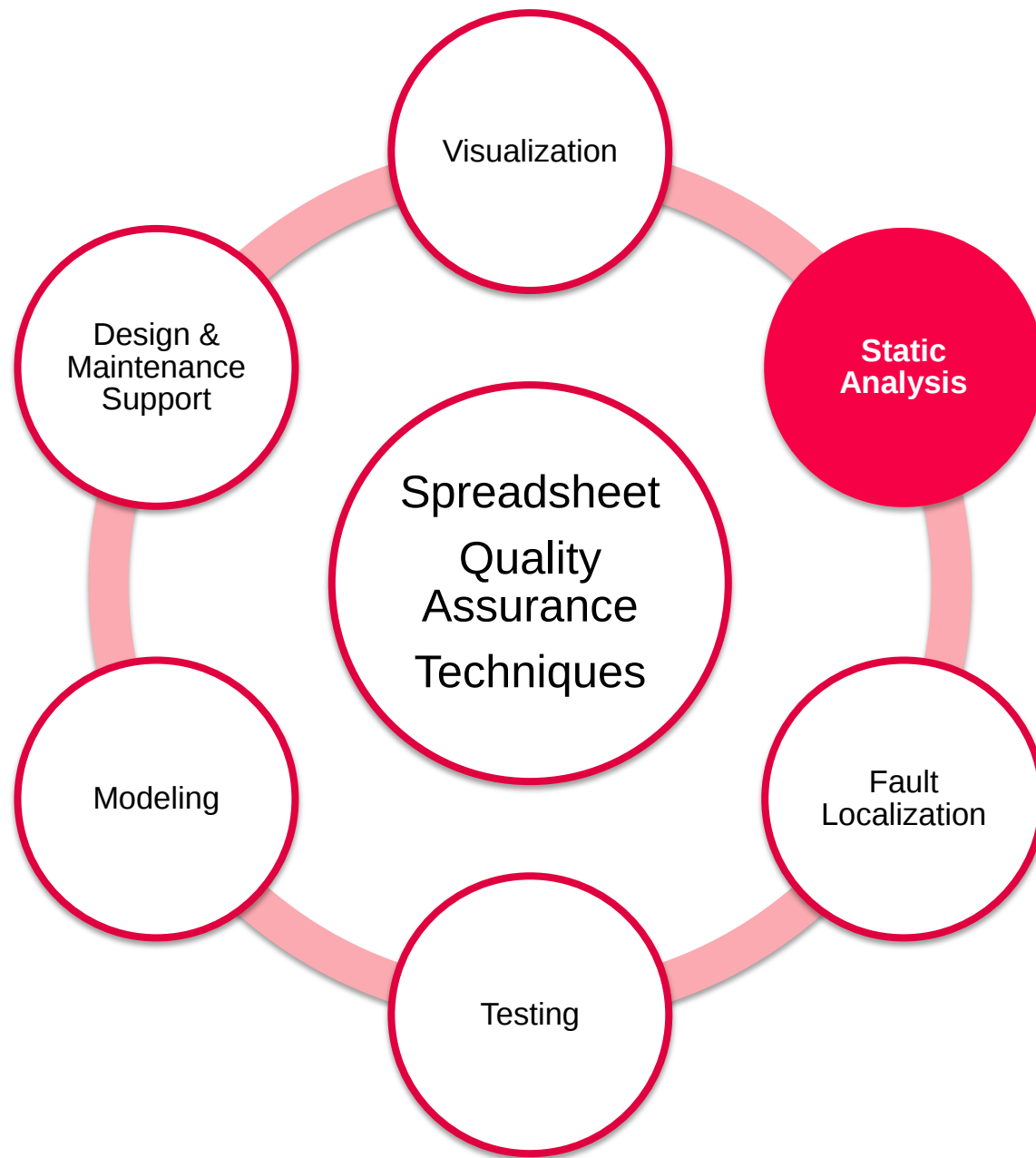
Spreadsheet QA Techniques

We want to...

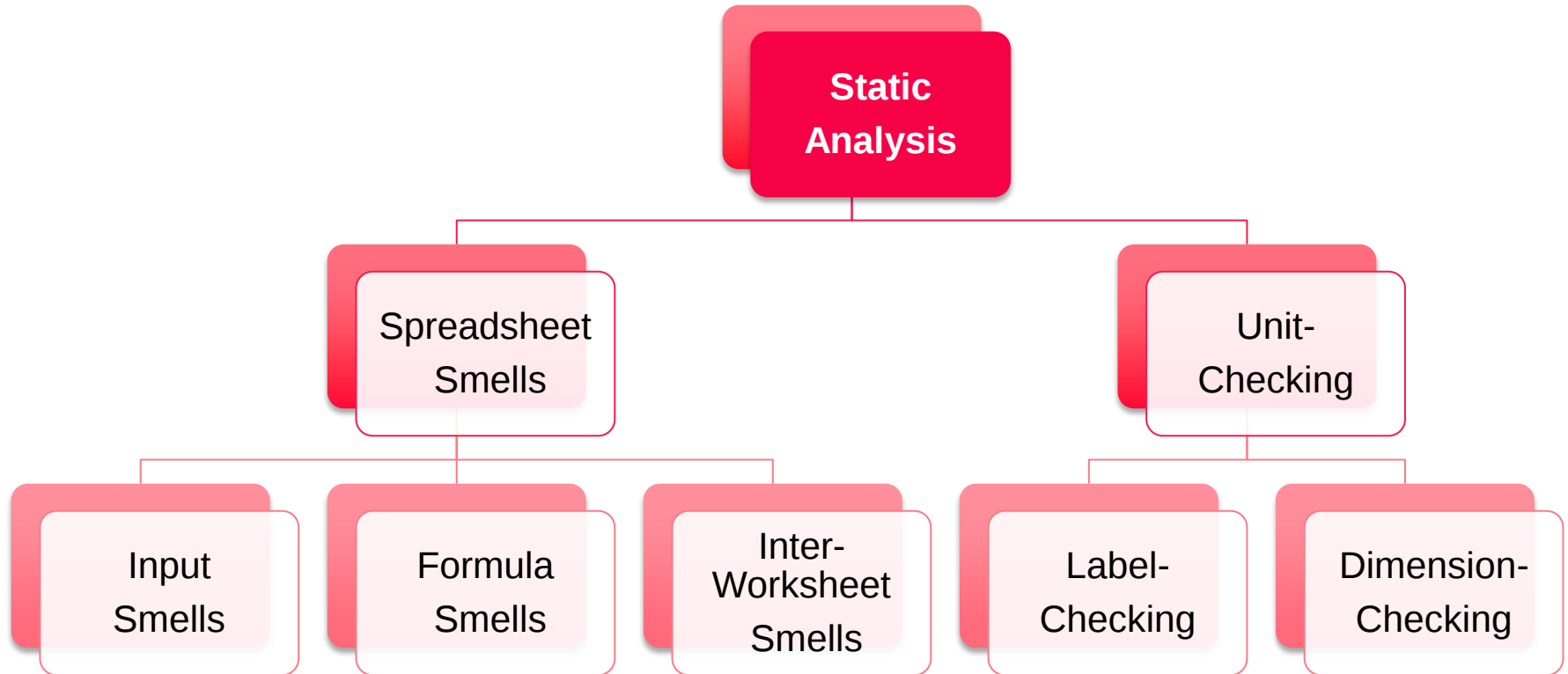
- Avoid Faults
- Find Faults
- Fix Faults
- Automatically







Spreadsheet QA Techniques



Outline

- Motivation
- Spreadsheet QA Techniques
- **Spreadsheet Smells**
- Unit-Checking Techniques
- Future Work

Spreadsheet Smells

What is a smell?

- Something stinks in daily life:
 - Call quality of good into question
- We perceive a spreadsheet smell:
 - Call quality of affected part into question
 - Hard to comprehend
 - Hard to maintain
 - Error-prone



Spreadsheet Smells: Smell Origins

- Based on Code Smells
 - Basic idea by Martin Fowler
 - Indicate required Refactoring
 - Detected via Quality Metrics
- Smells for Spreadsheets
 - Proposed by various Research Groups
 - Adapted existing Code Smells
 - Introduced new Quality Metrics

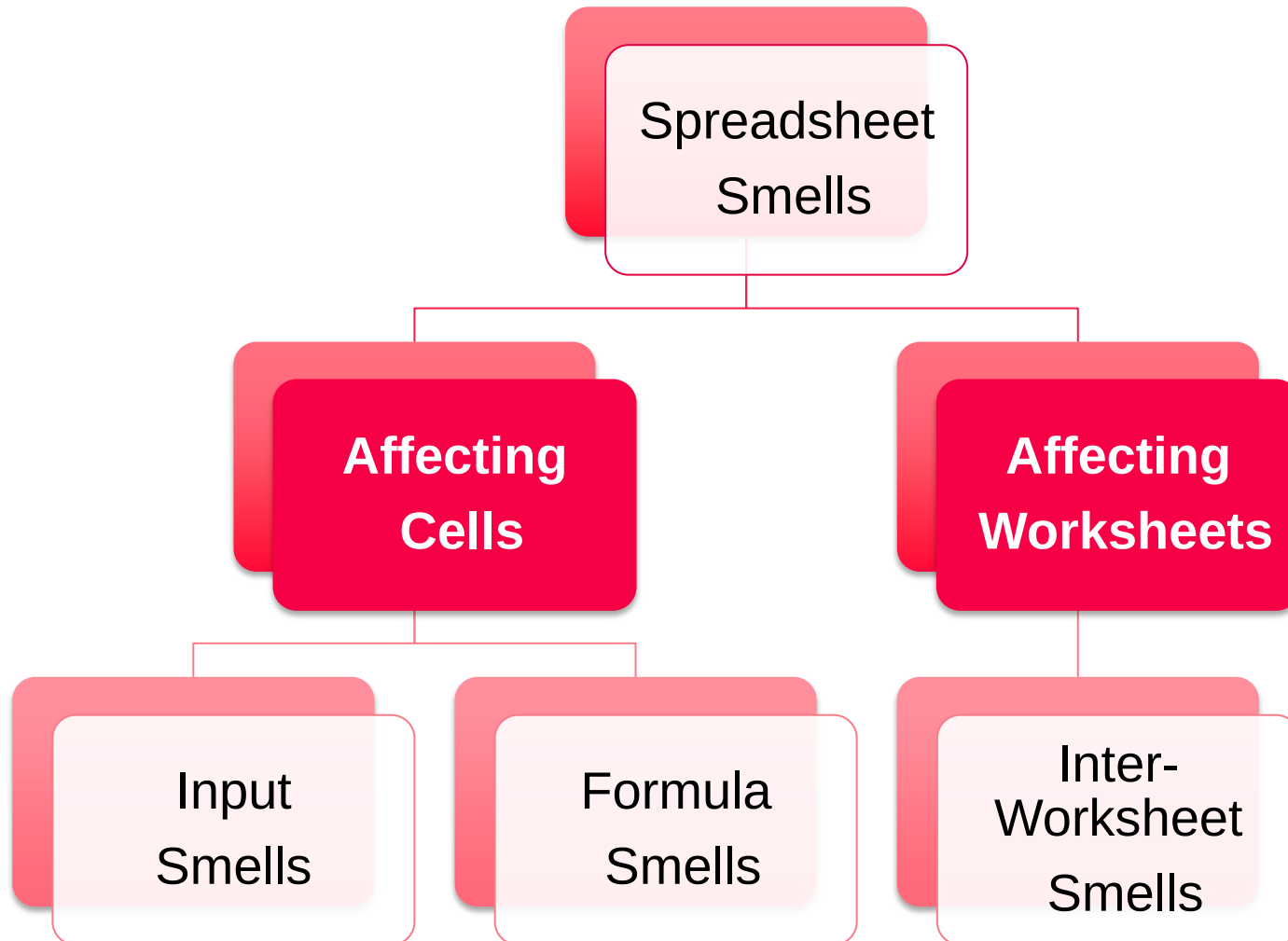


Spreadsheet Smells: Catalogue

- Catalogue of 15 refined Spreadsheet Smells
- 3 types
- 4 performance characteristics
- Basis for further research

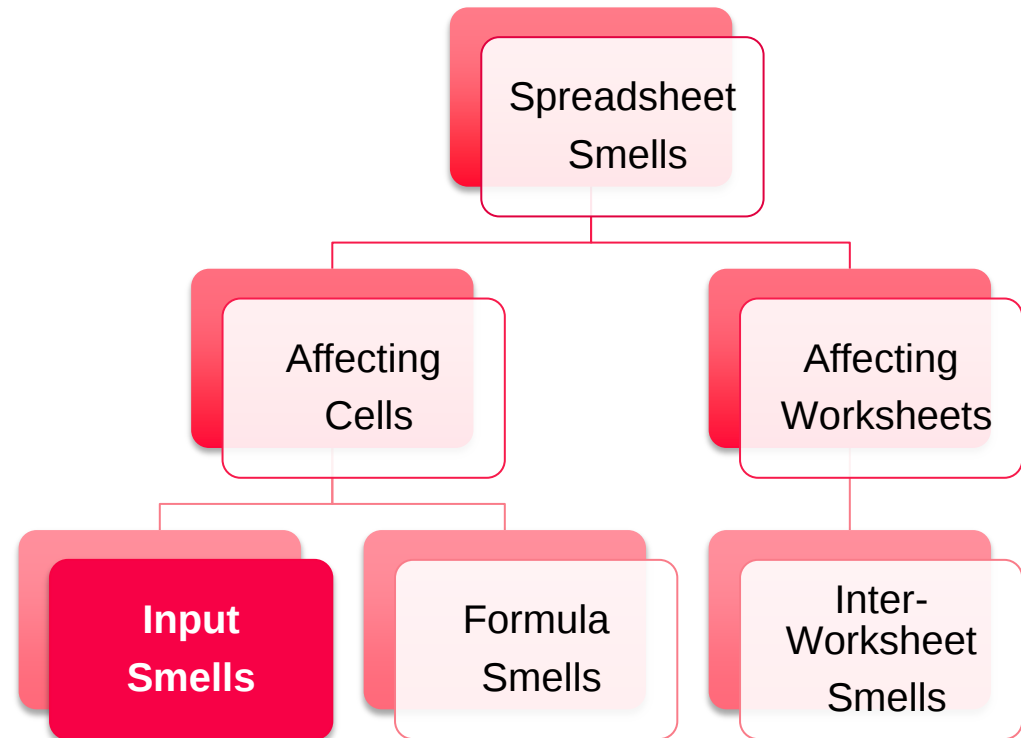


Spreadsheet Smells: Smell Types



Spreadsheet Smells: Input Smells

- Affect input cells
- Used for detection:
 - Cell type
 - Numeric values
 - String values



Spreadsheet Smells: Input Smells

Example: Standard Deviation

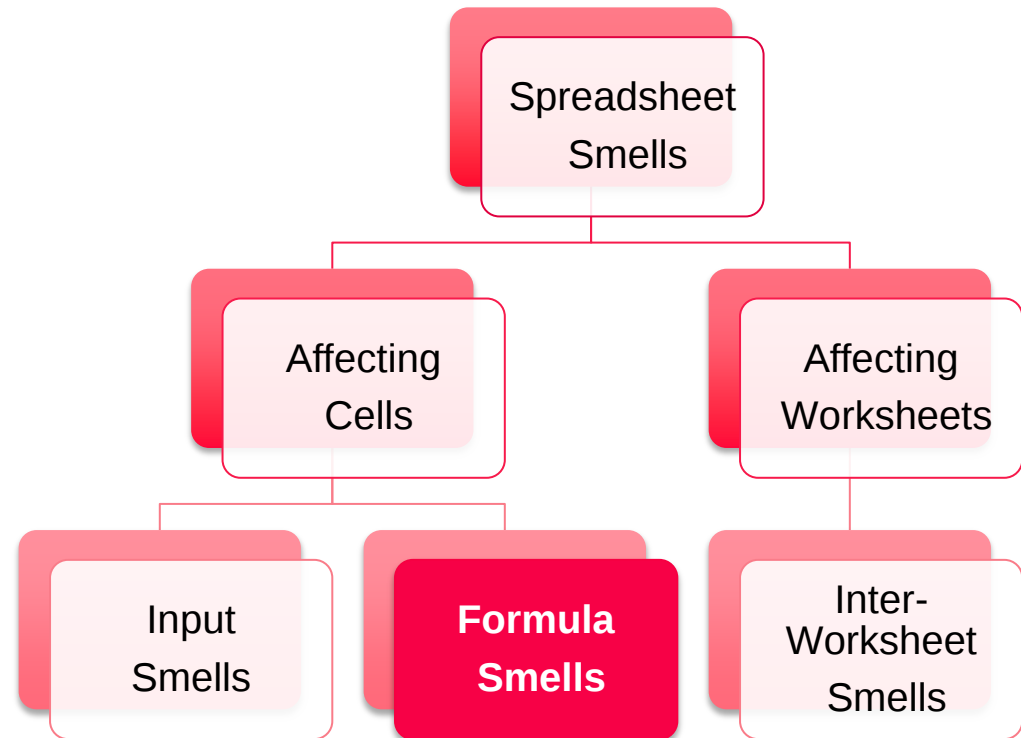
- Detects statistical outliers in values
- Works by:
 - Building groups of numeric cells
 - Calculating normal distribution model

	A	B	C
1	id	reference	description
2	5	1007993410	Product One
3	5	1079	Product One
4	5	1007993410	Product One
5	5	1007993410	Product One
6	6	1002394514	Product Two
7	6	1002394514	Product Two

Source: Cunha *et al.*: "Towards a catalog of spreadsheet smells", ICCSA 2012.

Spreadsheet Smells: Formula Smells

- Affect formula cells
- Used for detection:
 - Cell references
 - Formula Operators
 - Formula Constants



Spreadsheet Smells: Formula Smells

Example: Reference to Empty Cells

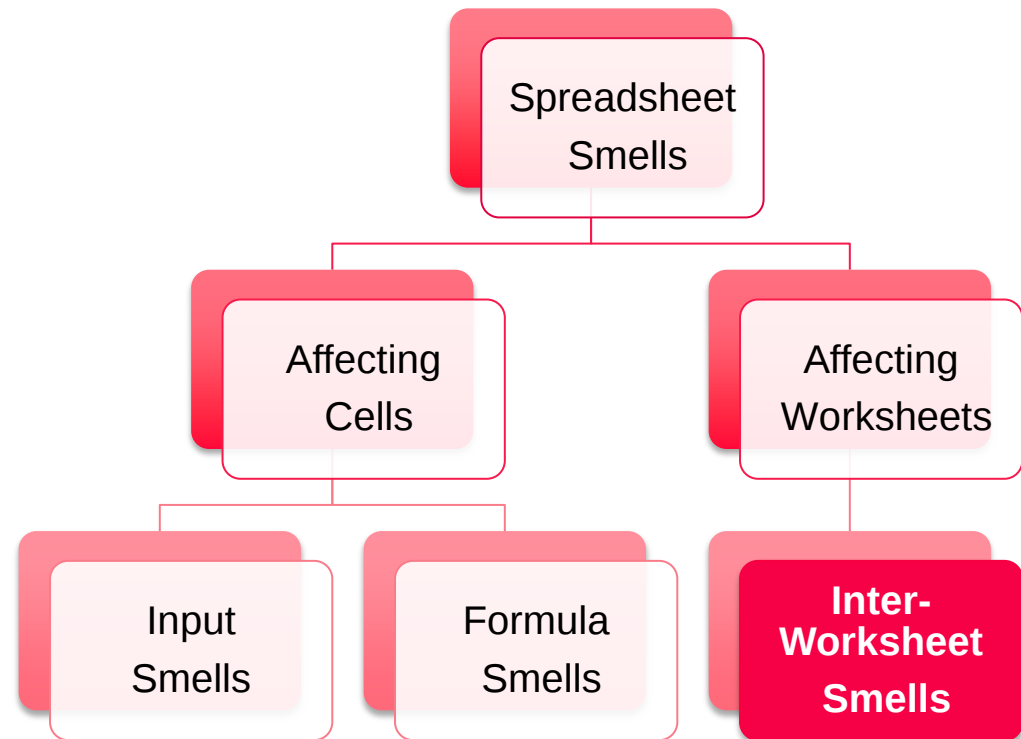
- Detects references to empty cells
- Works by:
 - Analyzing formulas
 - Checking references

	A	F	G
1	id	price	revenue
2	5	\$19.99	=F2*E2
3	5	\$19.99	=F3*E3
4	5	\$19.99	=F4*E4
5	5	\$19.99	=F5*E5
6	6		=F6*E6
7	6	\$29.99	=F7*E7

Source: Cunha *et al.*: "Towards a catalog of spreadsheet smells", ICCSA 2012.

Spreadsheet Smells: Inter-Worksheet Smells

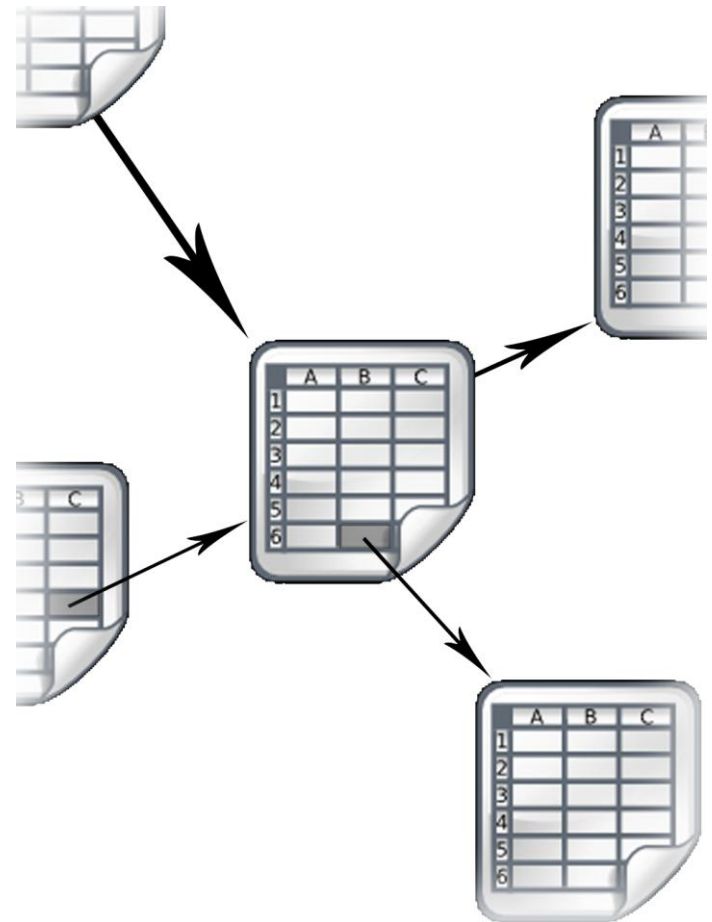
- Affect worksheets
- Used for detection:
 - References to other worksheets
 - References from other worksheets



Spreadsheet Smells: Inter-Worksheet Smells

Example: Shotgun Surgery

- Detects excessively referenced worksheets
- Works by calculating:
 - # references from other formulas to worksheet
 - # references by other worksheets to worksheet



Source:
Hermans *et al.*: “Detecting and visualizing inter-worksheet smells in spreadsheets“, ICSE 2012.

Spreadsheet Smells: Smell Interactions

- Smells within the same category interact
- Interaction of Quality-Metrics
 - Similar metrics within each category
 - May affect each other
 - Do not have to affect each other



Spreadsheet Smells: Smell Interactions

Example: Formula Smells

Multiple Operations, Multiple References

	A	B	D	F	G
1	Ingredient A	Ingredient B	Profit Margin	Distribution Fee	Price
2	10	5	20%	1	$= (A2 + B2) * (100\% + D2) + F2$



Long Calculation Chain

	A	B	C	D	E	F	G
1	Ingredient A	Ingredient B	Sum	Profit Margin	Cost	Distribution Fee	Price
3	10	5	$= A3 + B3$	20%	$= C3 * (100\% + D3)$	1	$= E3 + F3$



Source: Hermans *et al.*: "Detecting code smells in spreadsheet formulas", ICSM 2012.

Spreadsheet Smells: Further Properties

Introduction

- Creation
- Data Entry
- Expansion

Consequences

- Erroneous Result
- Impeded Quality

Alleviation

- Manual
- Assisted
- Automated

Spreadsheet Smells: Catalogue

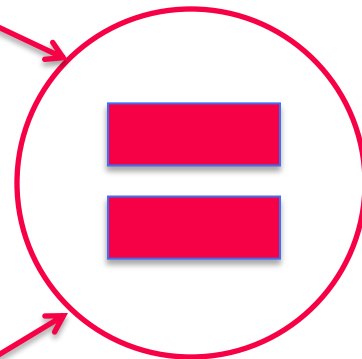
- Catalogue is open for Expansion
- Example: Data Clones by Hermans *et al.*
 - Type: Input smell
 - Introduction: Data Entry, Expansion
 - Consequence: Erroneous result
 - Alleviation: Manual



Source: Hermans *et al.*: "Data clone detection and visualization in spreadsheets", ICSE 2013.

Spreadsheet Smells: Data Clones

	A	S	T	U
3	Multiplier:	1000000	1000000	1000000
4	Currency:	USD	USD	USD
791	2008-05	67335.2	3709.3	481814.0
792	2008-06	67235.6	3729.5	482395.8
793	2008-07	66830.9	3723.3	479220.7
794	2008-08	66386.1	3807.6	477509.8
795	2008-09	65837.3	3798.3	465290.8
796	2008-10	63850.8	3802,0	455194.2
797	2008-11	62661.0	3964.4	448951.2
798	2008-12	62027.9	4247.9	449985.0
799	2009-01	50073.8	4003.6	448257.9
800	2009-02	44605.6	3819.5	438314.7
801	2009-03	45190.7	3670,0	445854.8
802	2009-04	44535.3	3701.1	440687.9
803	2009-05	44090.4	3730.8	437418.2
804	2009-06	43533.4	3729.5	446933.6
805	2009-07	42829.6	3723.3	443879.9
806	2009-08	39669.7	3807.6	441348.2
807	2009-09	38926.6	3798.3	443552.5
808	2009-10	38988.5	3802,0	434859.1
809	2009-11	39019.8	3964.4	429285.0



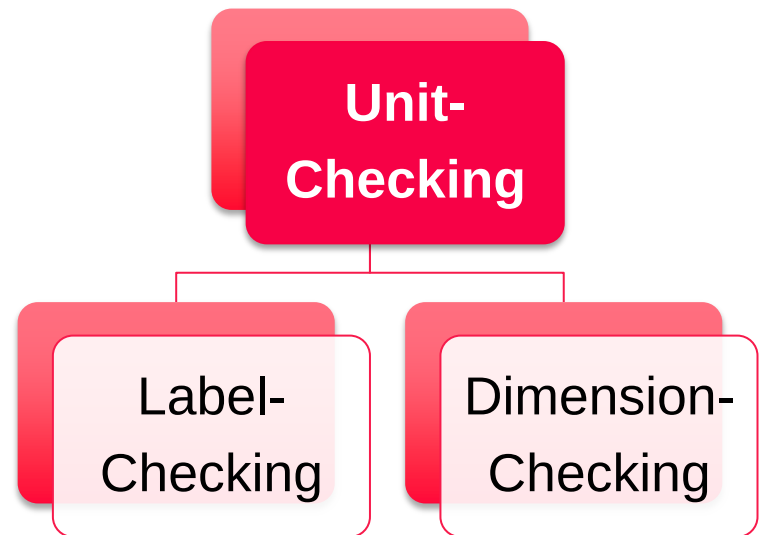
Could have prevented this

Outline

- Motivation
- Spreadsheet QA Techniques
- Spreadsheet Smells
- **Unit-Checking Techniques**
- Future Work

Unit-Checking Techniques

- Reason about cell units and dimensions
 - Detect semantic faults in formulas
- Require unit information
 - Manual annotation
 - Header inference
- 2 main approaches
 - Share common concepts
- Further information in seminar paper



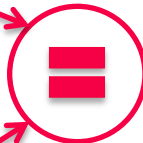
Outline

- Motivation
- Spreadsheet QA Techniques
- Spreadsheet Smells
- Unit-Checking Techniques
- **Future Work**

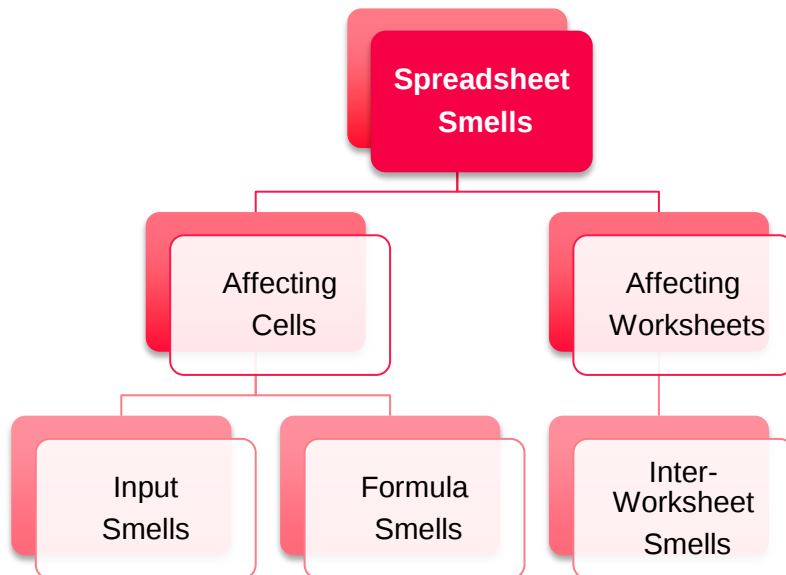
Future Work

- Spreadsheet Smells
 - Investigate characteristics & interactions
 - Introduce further smells
 - Find inter-worksheet smell counterbalance
 - Improve tool support for automated refactoring
- Unit-Checking Techniques for Spreadsheets
 - Improve header inference techniques
 - Improve false-positive detection rates

	A	S	T	U
3	Multiplier:	1000000	1000000	1000000
4	Currency:	USD	USD	USD
791	2008-05	67335.2	3709.3	481814.0
792	2008-06	67235.6	3729.5	482395.8
793	2008-07	66830.9	3723.3	479220.7
794	2008-08	66386.1	3807.6	477509.8
795	2008-09	65837.3	3798.3	465290.8
796	2008-10	63850.8	3802.0	455194.2
797	2008-11	62661.0	3964.4	448951.2
798	2008-12	62027.9	4247.9	449985.0
799	2009-01	50073.8	4003.6	448257.9
800	2009-02	44605.6	3819.5	438314.7
801	2009-03	45190.7	3670.0	445854.8
802	2009-04	44535.3	3701.1	440687.9
803	2009-05	44090.4	3730.8	437418.2
804	2009-06	43533.4	3729.5	446933.6
805	2009-07	42829.6	3723.3	443879.9
806	2009-08	39669.7	3807.6	441348.2
807	2009-09	38926.6	3798.3	443552.5
808	2009-10	38988.5	3802.0	434859.1
809	2009-11	39019.8	3964.4	429285.0

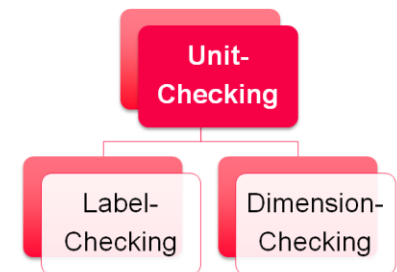


Summary



Unit-Checking Techniques

- Reason about cell units and dimensions
 - Detect semantic faults in formulas
- Require unit information
 - Manual annotation
 - Header inference
- 2 main approaches
 - Share common concepts
- Further information in seminar paper



Unit-Checking Techniques: Coffee Example

Values

	A	B	C	D	E	F
1	Coffee	Price	Weight	Price per weight	Standard weight	Total Price
2	Jacobs	260	25	10.4	100	1040
3	HAG	360	30	12	100	1200
4	Illy	240	25	265	100	26500
5						1040

Formulas

	A	B	C	D	E	F
1	Coffee	Price	Weight	Price per weight	Standard weight	Total Price
2	Jacobs	260	25	=B2/C2	100	=D2*E3
3	HAG	360	30	=B3/C3	100	=D3*E3
4	Illy	240	25	=B4+C4	100	=D4*E4
5						=min(F2:F4)

Unit-Checking: Label-Checking

Formulas

	A	B	C	D	E	F
1	Coffee	Price	Weight	Price per weight	Standard weight	Total Price
2	Jacobs	260	25	=B2/C2	100	=D2*E3
3	HAG	360	30	=B3/C3	100	=D3*E3
4	Illy	240	25	=B4+C4	100	=D4*E4
5						=min(F2:F4)

Label Inference

	A	B	C	D	E	F
1	Coffee	Price	Weight	Price per weight	Standard weight	Total Price
2	Jacobs	Price & Jacobs	Weight & Jacobs	Price & Jacobs	Standard weight & Jacobs	ERROR
3	HAG	Price & HAG	Weight & HAG	Price & HAG	Standard weight & HAG	HAG
4	Illy	Price & Illy	Weight & Illy	(Price Weight) & Illy	Standard weight & Illy	Illy
5						ERROR

Unit-Checking: Dimension-Checking

Formulas

	A	B	C	D	E	F
1	Coffee	Price	Weight	Price per weight	Standard weight	Total Price
2	Jacobs	260	25	=B2/C2	100	=D2*E3
3	HAG	360	30	=B3/C3	100	=D3*E3
4	Illy	240	25	=B4+C4	100	=D4*E4
5						=min(F2:F4)

Dimension Inference

	A	B	C	D	E	F
1	Coffee	Price	Weight	Price per weight	Standard weight	Total Price
2	Jacobs	\$	kg	\$ / kg	kg	\$
3	HAG	\$	kg	\$ / kg	kg	\$
4	Illy	\$	kg	ERROR: \$ + kg	kg	\$
5						\$

Name	Target					Oo-pendant	Introduction			Consequence		Alleviation		
	Empty	Numeric	String	Formula	Worksheet		Creation	Data entry	Expansion	Erroneous Result	Impeded Quality	Manual	Assisted	Automated
Std. Deviation		●				-		●		●		●		
Empty Cell	●					-	●	●		●			●	
Pattern Finder	●	●	●	●		-	●	●		●			●	
String Distance			●			-		●		●				●
Ref. to empty Cells				●		-	●		●	●			●	
QFD		●	●			-		●		●				●
Mult. Operations				●		Long Method			●		●			●
Multi. References				●		Many Params.			●		●			●
Cond. Complexity				●		-	●		●		●			●
Long Calc.-Chain				●		-	●		●		●			●
Dupl. Formulas				●		Duplication			●		●			●
Inappr. Intimacy					●	Inappr. Intimacy	●		●		●	●		
Feature Envy				●		Feature Envy	●		●		●	●		
Middle Man					●	Middle Man			●		●	●		●
Shotgun Surgery					●	Shotgun Surgery	●		●		●	●		

References: Spreadsheet Smells

- Abreu *et al.*: „Smelling faults in spreadsheets“, ICSME 2014.
- Abreu *et al.*: „FaultySheet detective: When smells meet fault localization“, ICSME 2014.
- Badame and Dig: „Refactoring meets spreadsheet formulas“, ICSM 2012.
- Cunha *et al.*: „SmellSheet detective: A tool for detecting bad smells in spreadsheets“, VL/HCC 2012.
- Cunha *et al.*: „Towards a catalog of spreadsheet smells“, ICCSA 2012.
- Hermans *et al.*: „Detecting and visualizing inter-worksheet smells in spreadsheets“, ICSE 2012.
- Hermans *et al.*: „Detecting code smells in spreadsheet formulas“, ICSM 2012.
- Hermans *et al.*: „Data clone detection and visualization in spreadsheets“, ICSE 2013.
- Hermans *et al.*: „Detecting and refactoring code smells in spreadsheet formulas“, ESE 2014.
- Martin Fowler: „Refactoring: improving the design of existing code“, 1999.

References: Unit-Checking Techniques

- Abraham and Erwig: „Header and unit inference for spreadsheets through spatial analyses“, VL / HCC 2004.
- Abraham and Erwig: „Ucheck: A spread-sheet type checker for end users“, 2007.
- Ahmed *et al.*: „A type system for statically detecting spreadsheet errors“, ASE 2003.
- Antoniu *et al.*: „Validating the unit correctness of spreadsheet programs“, ICSE 2004.
- Burnett and Erwig: „Visually customizing inference rules about apples and oranges“, HCC 2002.
- Chambers and Erwig: „Dimension inference in spreadsheets“, VL/HCC 2008.
- Chambers and Erwig: „Automatic detection of dimension errors in spreadsheets“, 2009.
- Chambers and Erwig: „Reasoning about spreadsheets with labels and dimensions“, 2010.
- Coblenz *et al.*: „Using objects of measurement to detect spreadsheet errors“, VL/HCC 2005.
- Erwig and Burnett: „Adding apples and oranges“, PADL 2002.

References: General

- Jannach *et al.*: „Avoiding, finding and fixing spreadsheet errors – a survey of automated approaches for spreadsheet QA“, 2014.
- Panko and Port: „End user computing: The dark matter (and dark energy) of corporate IT“, HICSS 2012.