A Conceptual Model for Measuring the Complexity of Spreadsheets

Thomas Reschenhofer, Bernhard Waltl, Klym Shumaiev, Florian Matthes

Technical University of Munich (TUM)
Department of Informatics
Chair of Software Engineering for Business Information Systems

wwwmatthes.in.tum.de

Agenda

1. Introduction & Motivation
2. Some Related Work
3. Research Method
4. A Conceptual Model for Spreadsheet Complexity
5. A Selection of Complexity Metrics
6. Application of Metrics to Spreadsheet Corpora
7. Conclusion

Introduction

sebis

- Spreadsheets are widely used in industry
  - E.g., financial reporting, workload planning, and general administration
  - Critical for many business processes

- Spreadsheets are **error prone**
  - Errors have significant impact on business operations

- Several attempts to avoid, identify, classify and fix spreadsheet errors
  - Numerous studies analyzing the causes of errors

- Typical errors
  - typing and copying mistakes
  - errors in logic and formulas
  - erroneous cell references
  - misplacement of data
  - ….

Introduction II

sebis

- Measuring complexity of spreadsheets as an **indicator** for the **risk** of errors
- There is already some research approaches on spreadsheet complexity and risk

- However, there is no conceptual model as a common foundation for those approaches which
  - formally **captures** potential (structural) **drivers** for spreadsheet complexity
  - facilitates the **identification** and **definition** of new complexity metrics or the **adaption** from metrics from other domains
  - enhances **reproducibility** of the application of complexity measures
  - establishes a **shared understanding** of (structural) spreadsheet complexity

1. What is a **spreadsheet model** capturing potential complexity drivers for spreadsheets, and which enables the formal definition of complexity metrics?

2. How can **metrics form software engineering and linguistics** be defined based on the proposed conceptual model?

3. According to those indicators, how complex are today's spreadsheets, and how do those metrics **correlate** to each other?

Related Work

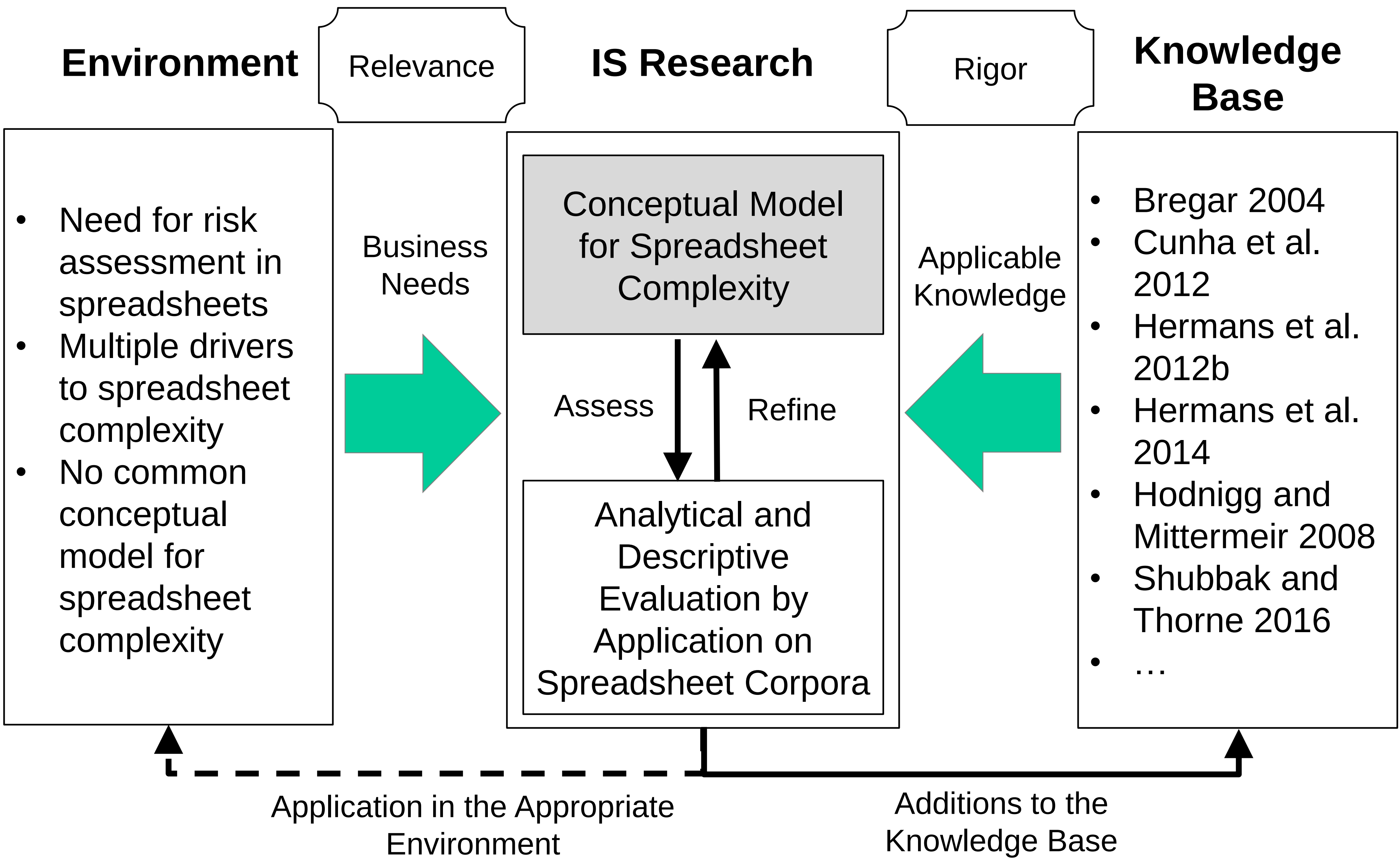
- **Bregar 2004**
  - Mathematical definition of complexity metrics, mostly adapted from the SE domain
- **Cunha et al. 2012**
  - Quality model of spreadsheets based on common software engineering standard
- **Hermans et al. 2010b**
  - Correlation of risk and complexity of spreadsheets with understandability of formulas
- **Hermans et al. 2014**
  - Adaption of the concept of code smells to spreadsheets in order to generate risk maps and locate “high-risk areas” of spreadsheets
- **Hodnigg and Mittermeir 2014**
  - Complexity metrics based on mathematical and graph-based notations
- **Shubbak and Thorne 2016**
  - Assessment of risk by a spreadsheet’s nature, importance, use, and complexity

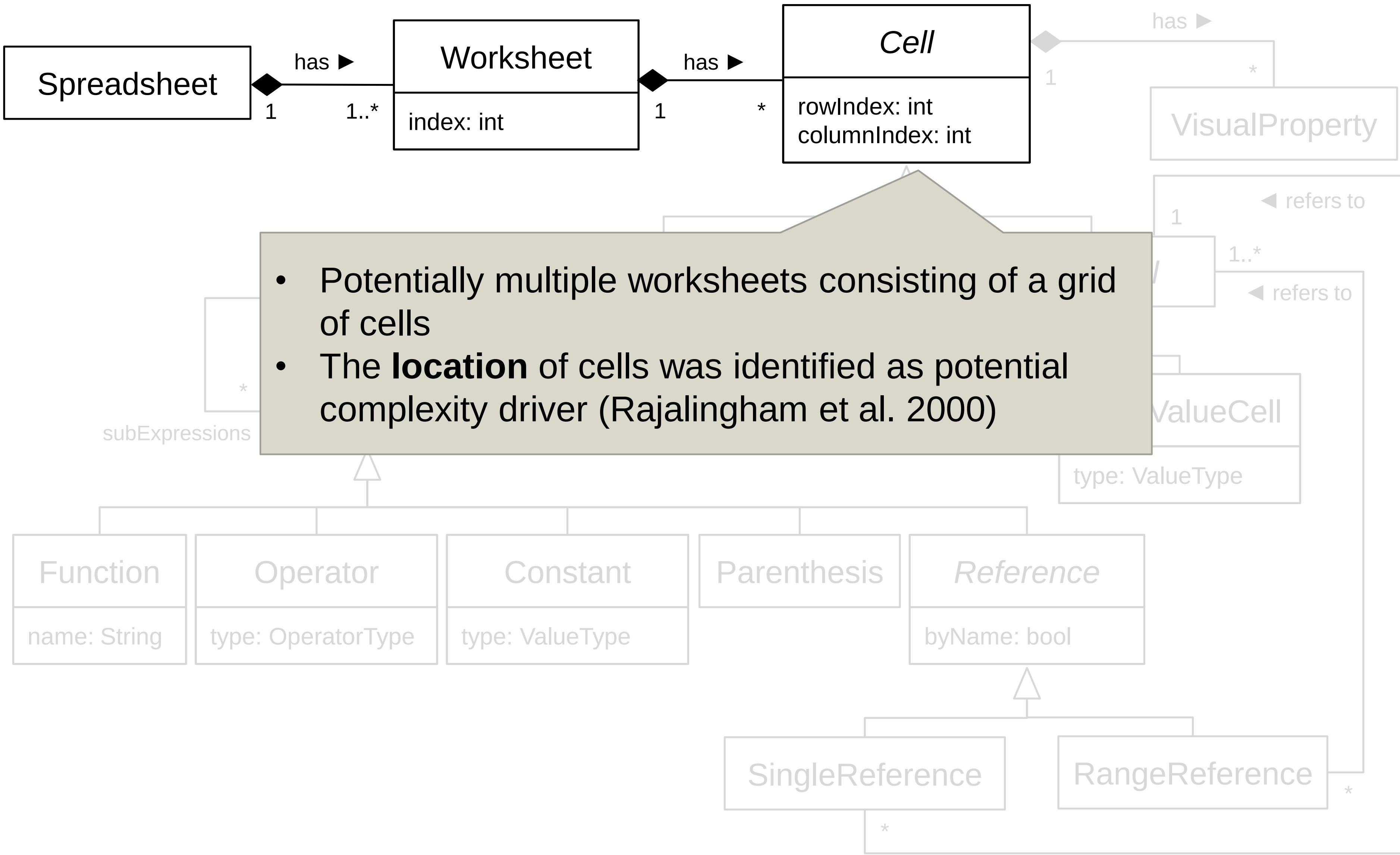
➢ No common conceptual model of spreadsheet complexity

➢ Different aspects which are considered to be drivers of complexity

Research Method
Hevner et al. (2004)

sebis

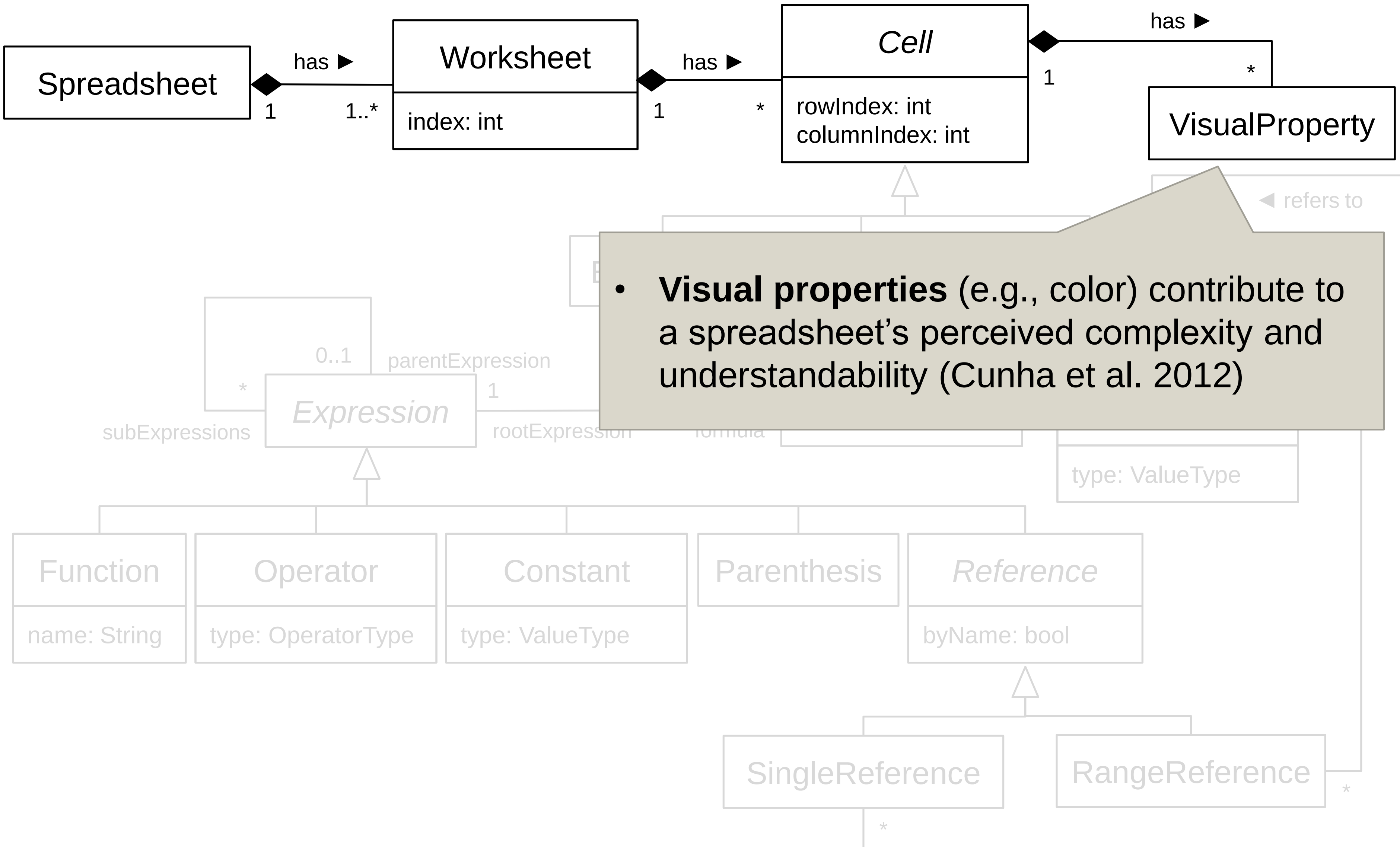




- Potentially multiple worksheets consisting of a grid of cells
- The **location** of cells was identified as potential complexity driver (Rajalingham et al. 2000)

A Conceptual Model for Spreadsheet Complexity
Visual Properties

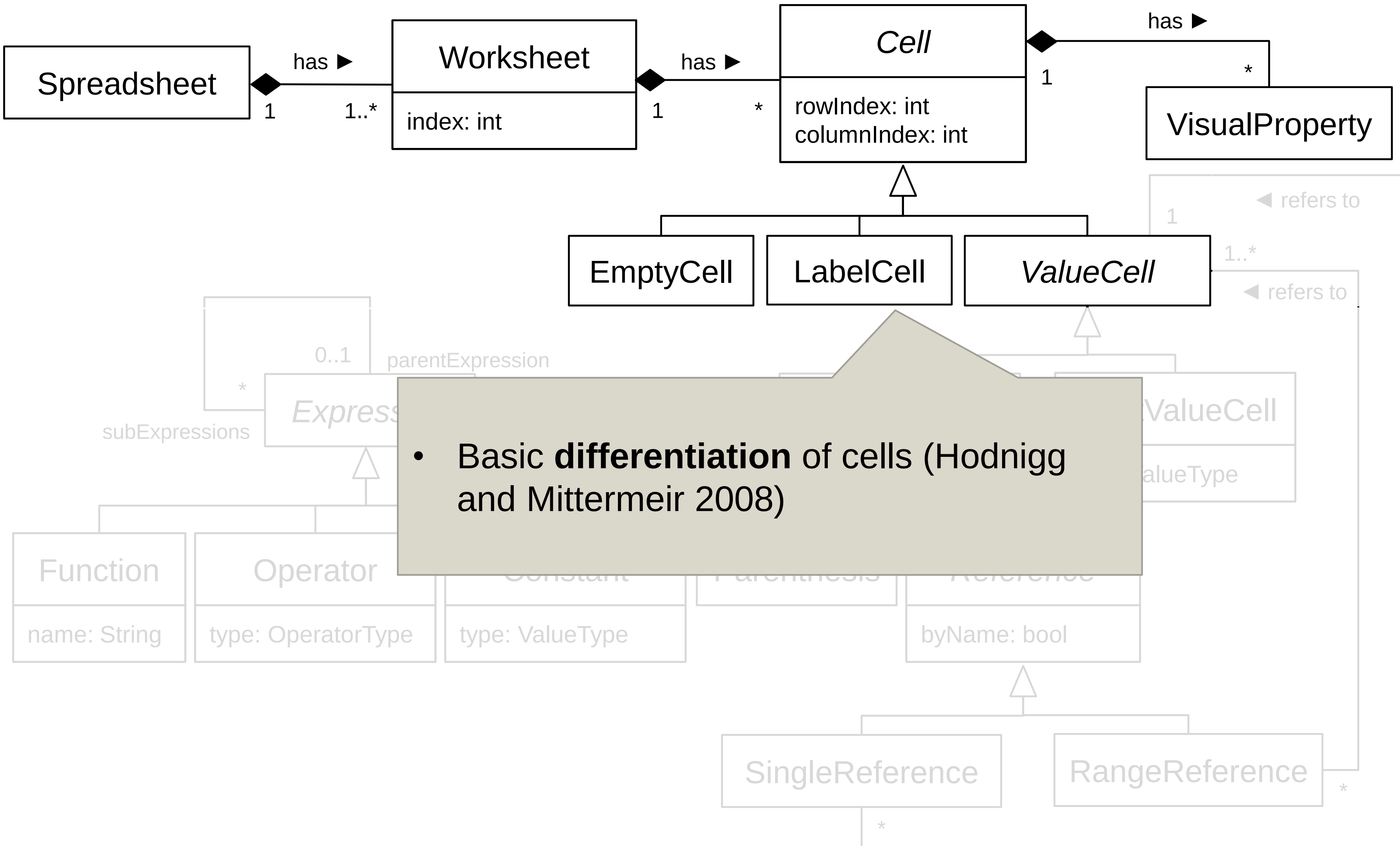
sebis



- **Visual properties** (e.g., color) contribute to a spreadsheet's perceived complexity and understandability (Cunha et al. 2012)

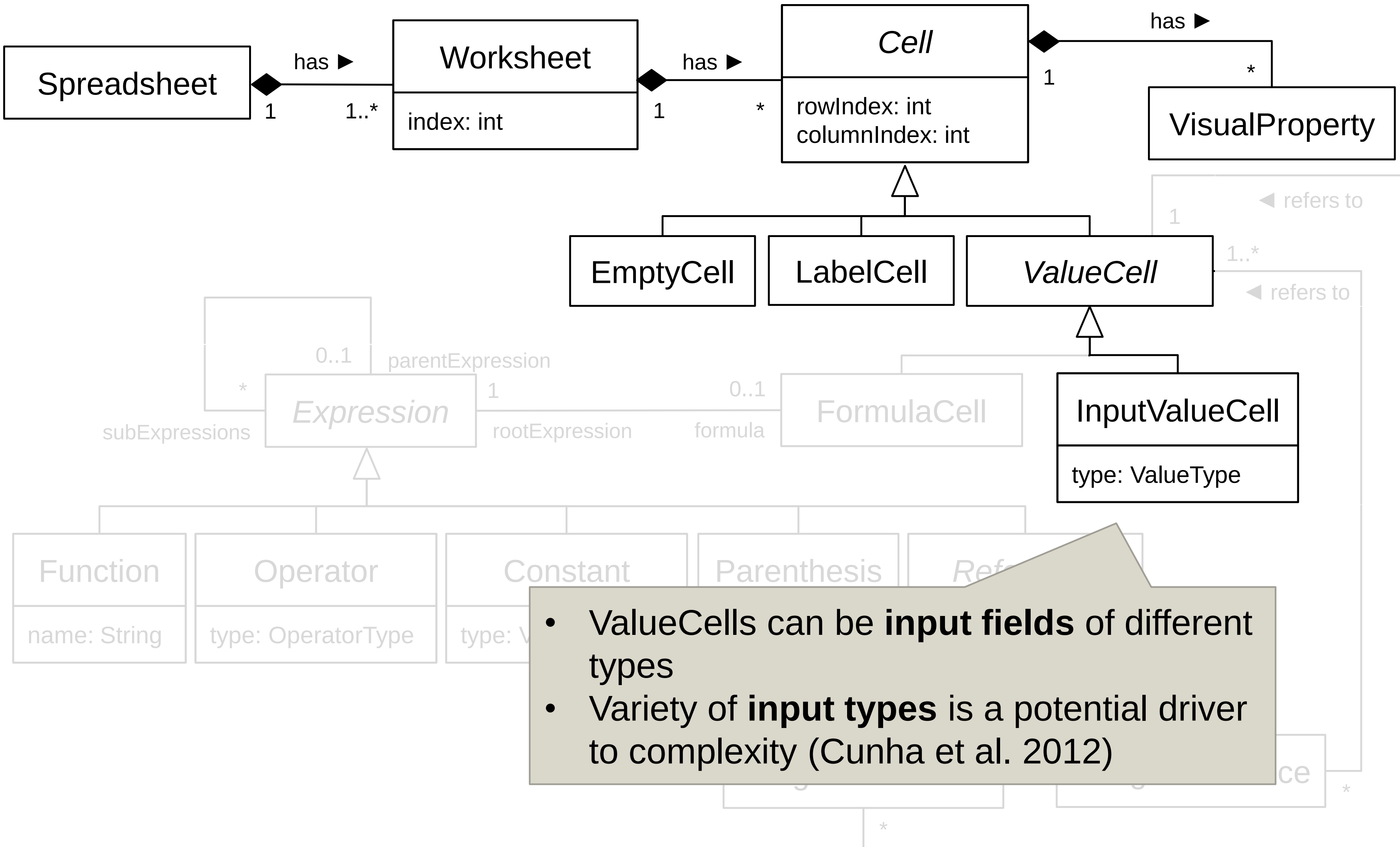
A Conceptual Model for Spreadsheet Complexity
Basic Differentiation of Cells

sebis



- Basic **differentiation** of cells (Hodnigg and Mittermeir 2008)

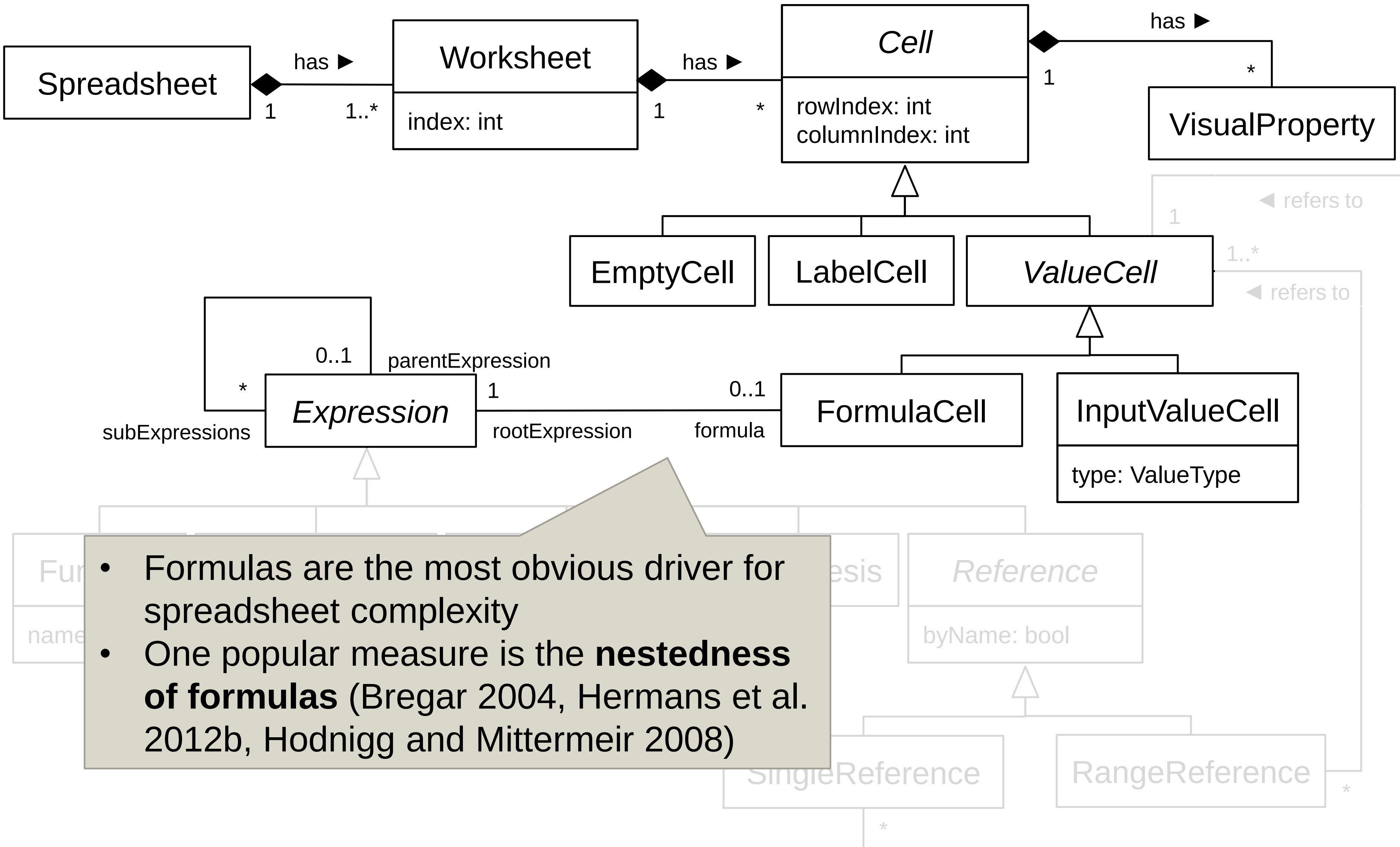
A Conceptual Model for Spreadsheet Complexity
InputValueCells



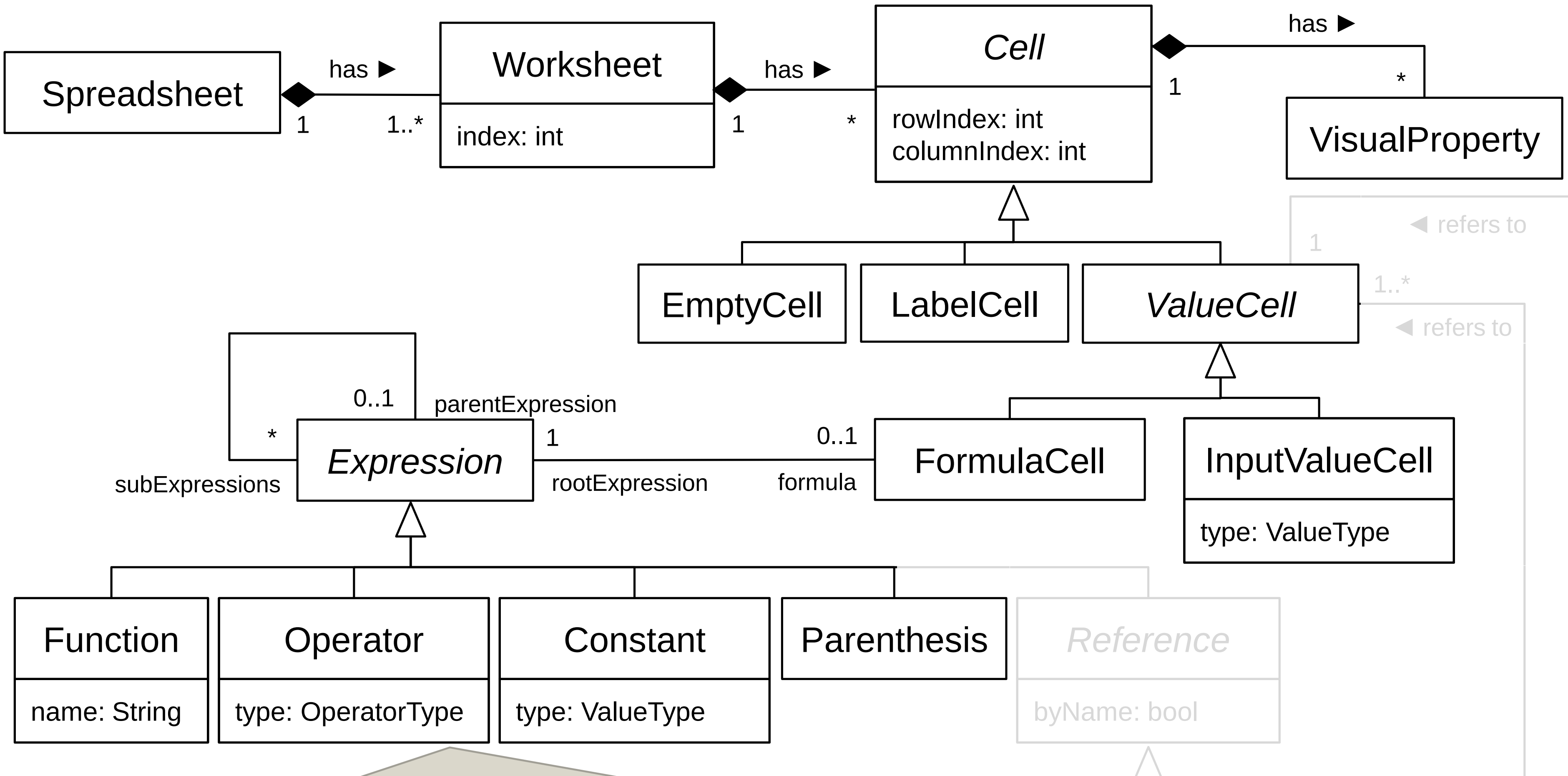
- ValueCells can be **input fields** of different types
- Variety of **input types** is a potential driver to complexity (Cunha et al. 2012)

A Conceptual Model for Spreadsheet Complexity
Formulas

sebis



- Formulas are the most obvious driver for spreadsheet complexity
- One popular measure is the **nestedness of formulas** (Bregar 2004, Hermans et al. 2012b, Hodnigg and Mittermeir 2008)



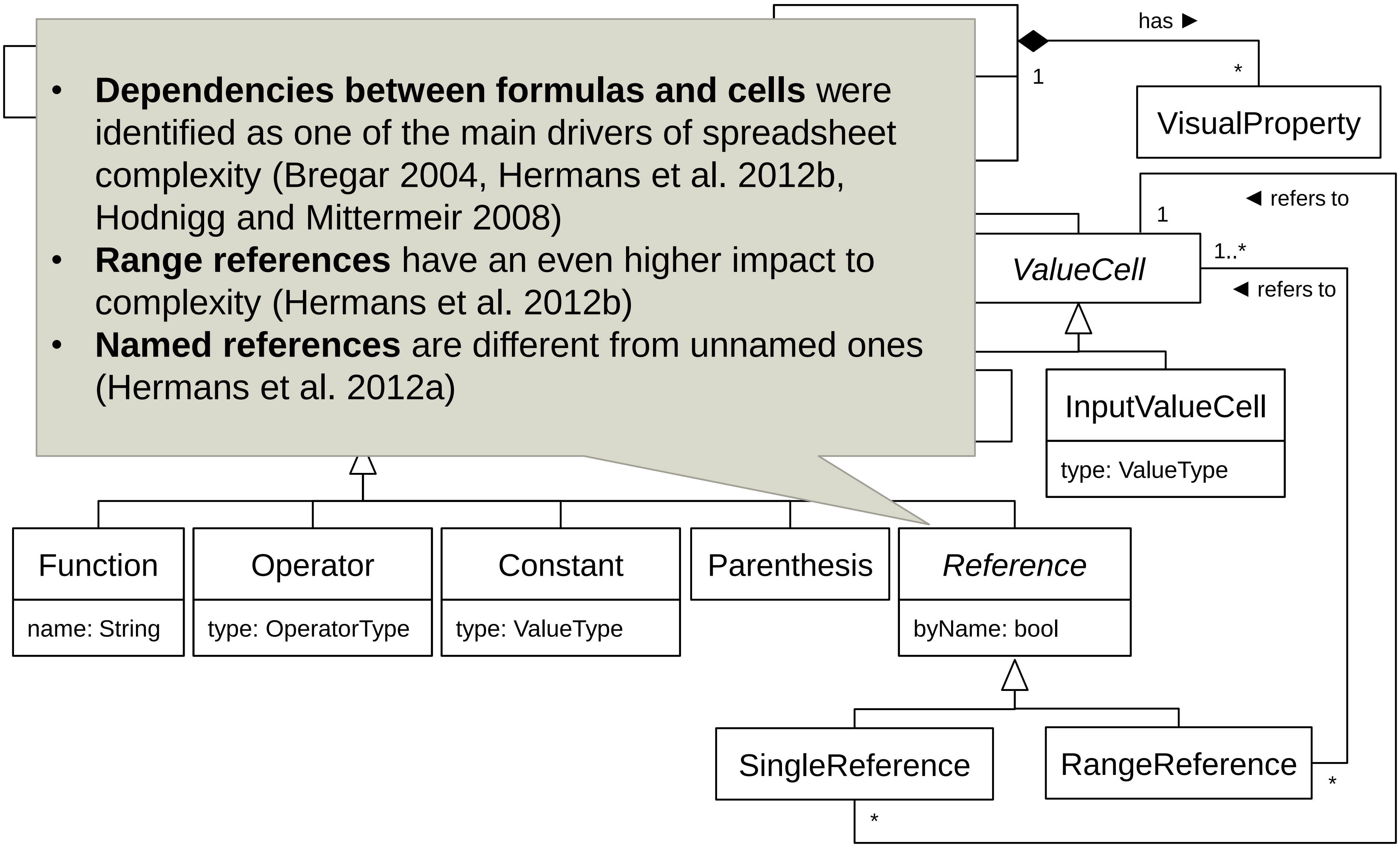
- More **concrete complexity measures regarding formulas** are the number of conditionals, occurrence of functions (e.g., lookup), and diversity of operators (Bregar 2004, Hodnigg and Mittermeir 2008)

A Conceptual Model for Spreadsheet Complexity

Dependencies Between Formulas and Cells

sebis

- **Dependencies between formulas and cells** were identified as one of the main drivers of spreadsheet complexity (Bregar 2004, Hermans et al. 2012b, Hodnigg and Mittermeir 2008)
- **Range references** have an even higher impact to complexity (Hermans et al. 2012b)
- **Named references** are different from unnamed ones (Hermans et al. 2012a)



A Selection of Complexity Metrics

- We selected metrics from related literature originating from the domain of Software Engineering and Linguistics

Software Engineering

- Average/Max AST depth per formula
- Number/Ratio of formula cells (to non-empty cells)
- Number/Ratio of input cells (to non-empty cells)
- Number of distinct formulas
- Average/Max fan-out per formula
- Average/Max fan-in per formula
- Average/Max number of conditionals per formula
- Average/Max spreading factor per formula

Linguistics

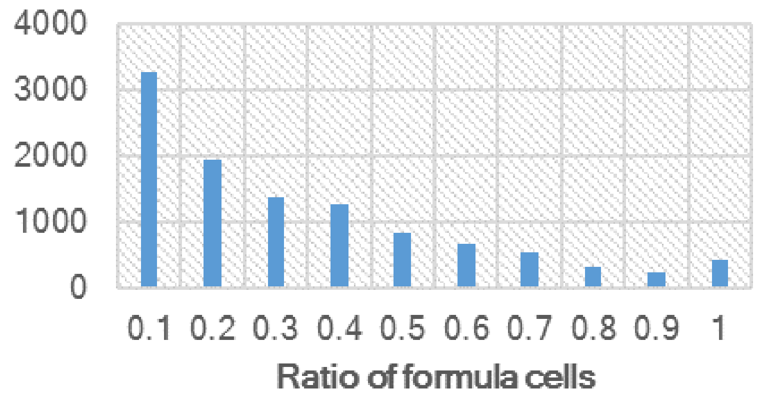
- Average/Max number of functions per formula
- Average/Max number of distinct functions per formula
- Average/Max number of elements per formula

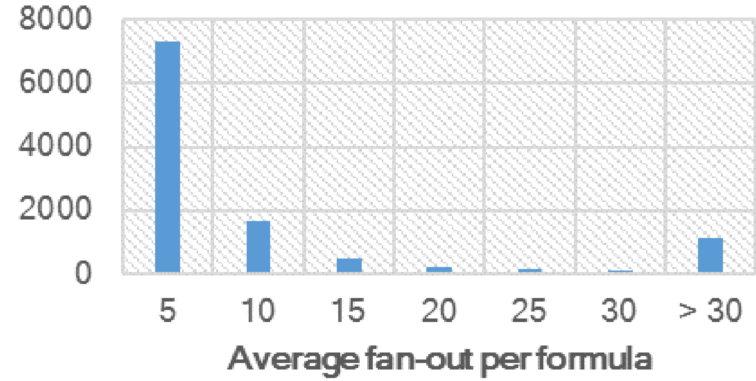
- **Each aspect/concept** as defined by the conceptual model is **captured** by at least one metric

Application of Metrics to Spreadsheet Corpora

- We applied the metrics to two spreadsheet corpora
  - **EUSES**: > 4 000 spreadsheets
  - **Enron**: > 15 000 spreadsheets

	EUSES	Enron
Ratio of spreadsheets with formulas	43 %	58 %
Number of formula cells	350	2107.53
Number of input cells	4931.90	11170.50
Ratio of input cells to non-empty cells	1.55	5.38
Ratio of formula cells to input cells	3.63	2.54
Number of distinct formulas	3.13	10.50
Average fan-out per formula	167.94	473.27
Max fan-out per formula	476.79	4709.88
Average fan-in per formula	0.93	7.70
Max fan-in per formula	9.20	50.53
Average spreading factor per formula	148.13	374.80
Max spreading factor per formula	350.94	1522.60





- **Important finding**: Only metrics capturing the same aspects of the conceptual model correlate to each other

Conclusion

- Conceptual model of spreadsheet complexity which
  - captures all **aspects** which were **identified by related work** as potential complexity drivers
  - serves as **foundation** for the **definition** or **adaption** of new metrics or metrics from other domains
  - formalizes **structural aspects** which are relevant to measure a spreadsheet‘s complexity

- Applying metrics capturing different aspects of the conceptual model shows that...
  - ... spreadsheets of the **Enron corpus** seem to be **more complex** than those of the **EUSES corpus**
  - ... **most spreadsheets** seem to have a rather **low complexity**, but that there is still a considerable amount of very complex spreadsheets
  - ... the aspects captured by the conceptual model are **independent from each other**

Thank you for your attention!

Thomas Reschenhofer
M.Sc.

Technical University of Munich (TUM)
Department of Informatics
Chair of Software Engineering for
Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17100
Fax +49.89.289.17136

thomas.reschenhofer@tum.de
wwwmatthes.in.tum.de