Universidade do Minho
Escola de Engenharia

Oregon State University

NOVALINCS
LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

Explaining Spreadsheets with Spreadsheets

[§] Jácome Cunha jacome@di.uminho.pt
[†] Mihai Dan danm@oregonstate.edu
[†] Martin Erwig erwig@oregonstate.edu
[†] Danila Fedorin fedorind@oregonstate.edu
[†] Alex Grejuc grejuca@oregonstate.edu

[§] University of Minho & NOVA LINCS

[†] Oregon State University

GPCE 2018, BOSTON, USA, 5-6 NOVEMBER

1. Introduction

Spreadsheets are Easy to Use (kind of)

Spreadsheets are Multi-Purpose

Spreadsheets are Widely Used

Very Widely Used!

They are the programming language of choice by non-professional programmers, aka end users

In the USA alone, the number of end-user programmers is conservatively estimated at **11 million**, compared to only 2.75 million other, professional programmers

Estimating the numbers of end users and end user programmers,
Christopher Scaffidi, Mary Shaw, and Brad Myers, 2005

So What?

In 2004, RevenueRecognition.com (now Softtrax) had the International Data Corporation (IDC) interview **118** business leaders

IDC found that 85% were using spreadsheets in **financial reporting** and **forecasting**

Sarbanes-Oxley: What About all the Spreadsheets?, Raymond R. Panko and Nicholas Ordway, 2008

http://www.eusprig.org/horror-stories.htm

BARABOO NEWS REPUBLIC

54° Cloudy Weekly Forecast

News | Sports | Opinion | Obituaries | Galleries | Jobs | CarSoup | Homes | Renta

W. Baraboo to pay more for borrowed money than believed

Due to a calculating error by their financial advisors, West Baraboo officials learned Thursday they will be paying about $400,0000 more over the lifetime of their most recent 10-year borrowing plan than originally projected.

During its regular December meeting, the West Baraboo Village Board looked back over last month's decision to sell $1.1 million in general obligation bonds to cover a variety of village projects, said Village Clerk Mary Klingenmeyer. The review was required after the board received a letter from its financial advisory firm, Eheleers of Brookfield.

Ehlers advisor James Mann said "operator error" resulted in a spreadsheet underestimating the total cost of the 10-year bond.

The Salt Lake Tribune | News

Tuesday, April 17, 2012 | Last Updated: 01:01 am

News | Sports | Blogs | Opinion | Money | Lifestyle | Entertainment | Obituaries | Jobs | Homes

Utah | Nation + World | Neighborhood | Politics | Justice | Polygamy | LDS Church | Education | Weather | UtahsRight

Utah education officials make $25M school funding mistake

Economy losses of $10 billion/year!

A miscalculation at the State Office of Education has led to a $25 million mistake in Utah's education budget for next school year — and the resignation of two top finance officials.

Education leaders, however, say they don't expect the potential shortfall to hurt schools or districts. State leaders are considering solutions ranging from using education money expected to be left over at the end of this school year to calling a special legislative session.

Photos

Join the Discussion

» Post a Comment

"We committed to fund [enrollment] growth and this is an important part of growth," said Senate Budget Chairman Lyle Hillyard on Wednesday. "We would hope to get it fixed, and I think that's going to be our first priority."

The $25 million represents less than 1 percent of the state's overall $3 billion-plus education budget.

The problem was that the state office essentially underestimated the number of students expected in schools next school year. The correct number will cost the state $25 million more than anticipated.

State Superintendent Larry Shumway attributed the mistake to "a faulty reference" in a spreadsheet. He emphasized that no money was misappropriated. He called the mistake "significant" but "manageable."

Why? One Reason is...

85% of the participants do not create the spreadsheets they have to work on themselves

Received them from their colleagues

70% of those users have **difficulties understanding** the spreadsheets

Felienne Hermans, Martin Pinzger, and Arie van Deursen. Supporting Professional Spreadsheet Users by Generating Leveled Data FLow Diagrams. ICSE'11

Indeed…

Report of JPMorgan Chase & Co. Management Task Force Regarding 2012 CIO ($6.2bn!) Losses

…

Specifically, after subtracting the old rate from the new rate, the spreadsheet divided by their sum instead of their average, as the **modeler had intended**.

…

2. An Explanation Language for Explaining Spreadsheets

An Explanation Language for Explaining Spreadsheets

We propose to augment spreadsheets with *explanations* written using an explanation language

With (spreadsheet) constructs to abstract spreadsheets' contents

But still within spreadsheets language

An Example

3	Adams	8.9	40	5	=B3*C3	=B3*1.5*D3	=E3+F3
4	Baker	12.55	35	0	=B4*C4	=B4*1.5*D4	=E4+F4
5	Carlton	9.6	40	2	=B5*C5	=B5*1.5*D5	=E5+F5
6	Daniels	10.2	35	0	=B6*C6	=B6*1.5*D6	=E6+F6
7							
8	Totals		=SUM(C3:C6)	=SUM(D3:D6)	=SUM(E3:E6)	=SUM(F3:F6)	=SUM(G3:G6)

- What is being calculated in cell E5?
- How is it being calculated?

An Example

	A	B	C	D	E	F	G
1				Payroll Spreadsheet			
2	Name	Pay Rate	Regular Hours	Overtime Hours	Regular Pay	Overtime Pay	Total
3	Adams	8.9	40	5	=B3*C3	=B3*1.5*D3	=E3+F3
4	Baker	12.55	35	0	=B4*C4	=B4*1.5*D4	=E4+F4
5	Carlton	9.6	40	2	=B5*C5	=B5*1.5*D5	=E5+F5
6	Daniels	10.2	35	0	=B6*C6	=B6*1.5*D6	=E6+F6
7							
8	Totals		=SUM(C3:C6)	=SUM(D3:D6)	=SUM(E3:E6)	=SUM(F3:F6)	=SUM(G3:G6)

- What is being calculated in cell E5?
- How is it being calculated?

Label abstraction

	A	B	C	D	E	F	G
1				Payroll Spreadsheet			
2	Name	Pay Rate	Regular Hours	Overtime Hours	Regular Pay	Overtime Pay	Total
3	Adams	8.9	40	5	=Pay Rate*Regular Hours	=Pay Rate*1.5*Overtime Hours	=Regular Pay+Overtime Pay
4	Baker	12.55	35	0	=Pay Rate*Regular Hours	=Pay Rate*1.5*Overtime Hours	=Regular Pay+Overtime Pay
5	Carlton	9.6	40	2	=Pay Rate*Regular Hours	=Pay Rate*1.5*Overtime Hours	=Regular Pay+Overtime Pay
6	Daniels	10.2	35	0	=Pay Rate*Regular Hours	=Pay Rate*1.5*Overtime Hours	=Regular Pay+Overtime Pay
7							
8	Totals		=SUM(Regular Hours)	=SUM(Overtime Hours)	=SUM(Regular Pay)	=SUM(Overtime Pay)	=SUM(Total)

- What is being calculated in cell F4?
- How is it being calculated?

Loop abstraction

	A	B	C	D	E	F	G
1				Payroll Spreadsheet			
2	Name	Pay Rate	Regular Hours	Overtime Hours	Regular Pay	Overtime Pay	Total
3	*[Adams...Daniels]*	*[8.9...12.55]*	*[35...40]*	*[0...5]*	=Pay Rate*Regular Hours	=Pay Rate*1.5*Overtime Hours	=Regular Pay+Overtime Pay
4							
5	Totals		=SUM(Regular Hours)	=SUM(Overtime Hours)	=SUM(Regular Pay)	=SUM(Overtime Pay)	=SUM(Total)

Does it really matter the number of entries (employees) to understand the spreadsheet?

Guiding the Design of *Explanation Sheets*

- We proposed an initial set of constructs:
  - Label abstraction
  - Loop abstraction
  - Nested loop abstraction
  - Case operator (nested ifs)
  - Group by

- We then manually inspected 40 spreadsheet
  - Randomly selected from a book on how to create spreadsheets
  - And from the repository of spreadsheets from Enron (a bankrupted company)

Real Spreadsheets, Real Language Constructs

- We noticed nested loop occurs rarely
- Nested if constructs also occur rarely
- Group by also occur rarely
- We thus decided to drop them
- They would make the language more complex with little real impact

	A	B
1	**Book ID-moved to NG-Pric**	Don’t move
2	NG-PRICE-BAS	FT-UPSTRM-BAS
3	NG-PRICE-GDI	FT-UPSTRM-GDI
4	NG-PRICE-GDL	GARDEN-STATES-PRC
5	NG-PRICE-PRC	GAS-APEA PREPAY-BAS
6	ARUBA-SPLY-PRC	GAS-APEA PREPAY-IDX
7	ARUBA-TP-PHY	GAS-APEA PREPAY-PRC
8	CES-EAST-PRC	GAS-CAGE-SVC-BAS
9	EAST-GULF1-BAS	FT-UPSTRM-GDL
10	EAST-GULF1-GDI	GAS-CHA-MAH-BAS
11	EAST-GULF1-GDL	GAS-CHA-MAH-IDX
12	EAST-GULF1-IDX	GAS-CAGE-SVC-IDX
13	EAST-GULF1-PRC	GAS-CAGE-SVC-PRC
14	EAST-GULF2-BAS	FT-UPSTRM-IDX
15	EAST-GULF2-GDI	GAS-CHASE-MAHIV-IDX
16	EAST-GULF2-GDL	GAS-CHASE-MAHIV-PRC
17	EAST-GULF2-IDX	GAS-CHASE-MAHV-BAS
18	EAST-GULF2-PRC	GAS-CHASE-MAHV-IDX
19	EAST-GULF3-BAS	GAS-CHASE-MAHV-PRC
20	EAST-GULF3-GDI	GAS-CHA-MAH-PRC
21	EAST-GULF3-GDL	GAS-CHASE-MAHII-BAS
22	EAST-GULF3-IDX	GAS-CHASE-MAHII-IDX
23	EAST-GULF3-PRC	GAS-CHASE-MAHII-PRC
24	EAST-GULF4-BAS	GAS-CHASE-MAHIV-BAS
25	EAST-GULF4-GDI	FT-UPSTRM-PHY
26	EAST-GULF4-GDL	FT-UPSTRM-PRC
27	EAST-GULF4-IDX	IN-UPSTREAM-ST-BAS
28	EAST-GULF4-PRC	IN-UPSTREAM-ST-GDI
29	EAST-GULF5-GDI	IN-UPSTREAM-ST-GDL
30	EAST-MKTEAST-BAS	IN-UPSTREAM-ST-IDX
31	EAST-MKTEAST-GDI	IN-UPSTREAM-ST-PRC
32	EAST-MKTEAST-GDL	CHASE-X-BAS
33	EAST-MKTEAST-IDX	CHASE-X-IDX
34	EAST-MKTEAST-PRC	CHASE-X-PRC

	A	B
1	**Book ID-moved to NG-Price**	Don’t move
2	*[NG-PRICE-BAS .. NG-PRICE-GDI]*	*[FT-UPSTRM-BAS .. FT-UPSTRM-GDI]*

			SoCal		PG&E CtyGate		Malin	
11								
12			Basis		Basis		Basis	
13	**START**	**END**	**Bid**	**Offer**	**Bid**	**Offer**	**Bid**	**Offer**
14								
15								
16	fev/02	fev/02	-0,11	-0,09	-0,02	0,01	-0,175	-0,14
17	mar/02	mar/02	-0,12	-0,1	-0,04	-0,02	-0,2	-0,16
18	abr/02	jun/02	-0,13	-0,11	-0,02	0,01	-0,19	-0,15
19	jul/02	set/02	0,04	0,06	0,14	0,17	-0,13	.-009
20	out/02	out/02	flat	0,03	0,125	0,165	-0,08	-0,03
21	nov/02	mar/03	0,025	0,045	0,21	0,24	0,04	0,08
22	abr/03	out/03	0,1	0,13	0,31	0,39	flat	0,08
23	nov/03	mar/04	0,11	0,14	0,29	0,37	0,08	0,14
24	abr/04	out/04	0,14	0,17	0,36	0,45	0,01	0,09
25	nov/04	mar/05	0,14	0,175	0,37	0,47	0,1	0,16
26	abr/05	out/05	0,15	0,19	0,37	0,49	0,02	0,08
27	nov/05	mar/06	0,155	0,195	0,37	0,49	0,07	0,15
28	abr/06	out/06	0,15	0,2	0,38	0,49	0,02	0,08
29	nov/06	mar/07	0,15	0,2	0,39	0,52	0,07	0,15
30	abr/07	out/07	0,155	0,2	0,4	0,52	0,02	0,08
31	nov/07	mar/08	0,15	0,2	0,4	0,53	0,08	0,15
32	abr/08	out/08	0,155	0,2	0,4	0,52	0,02	0,08
33	nov/08	mar/09	0,15	0,2	0,4	0,53	0,07	0,15
34	abr/09	out/09	0,155	0,21	0,4	0,52	0,02	0,08
35	nov/09	mar/10	0,15	0,22	0,4	0,52	0,07	0,15
36	abr/10	out/10	0,16	0,22	0,4	0,52	0,02	0,08
37	nov/10	mar/11	0,155	0,25	0,4	0,52	0,07	0,15

			Rockies Questar Rockies	
11				
12			Basis	
13	**START**	**END**	**Bid**	**Offer**
14				
15				
16	*[01/02/2002,...]*	*[01/02/2002, ...]*	*[-0,11; ..]*	*[-0,09; ...]*

	A	B	C	D	E	F	G	H	I
1	**SS Kuniang**								
2									
3	**Assumptions**			**Model**					
4		Bid ($M)	=F4		Bid	12		Low Salvage	=MÁXIMO(C10-F4;C7;(
5		P(Low Salvage)	0,3		P(Win)	=(F4-2)/10		High Salvage	=MÁXIMO(C11-F4;C7;(
6									
7		Profit new ship	3,2						
8		Profit tug/barge	1,6					Win? Yes	=C5*I5+(1-C5)*I4
9		Gross profit SSK			Exp. Profit	=F5*I8+(1-F5)*I10			
10		Low Salvage	15,5					Win? No	=MÁXIMO(C7;C8)
11		High Salvage	12,5						

	A	B	C	D	E	F	G	H	I
1	**SS Kuniang**								
2									
3	**Assumptions**			**Model**					
4		Bid ($M)	Bid		Bid	12		Low Salvage	=MÁXIMO(Low Salvage-Bid;Profit new ship;Profit tug/barge)
5		P(Low Salvage)	0,3		P(Win)	=Bid/10		High Salvage	=MÁXIMO(High Salvage-Bid;Profit new ship;Profit tug/barge)
6									
7		Profit new ship	3,2						
8		Profit tug/barge	1,6					Win? Yes	=P(Low Salvage)*High Salvage+(1-P(Low Salvage))*Low Salvage
9		Gross profit SSK			Exp. Profit	=P(Win)*Win? Yes+(1-P(Win))*Win? No			
10		Low Salvage	15,5					Win? No	=MÁXIMO(Profit new ship;Profit tug/barge)
11		High Salvage	12,5						

The Spreadsheets Language

- **Spreadsheets** $s \in S$ are partial mappings from addresses $A = N \times N$ to formulas $f \in Fml$

- **Formulas f** are either
  - **plain values, v** $\in$ Val (e.g. 1, Month)
  - **application of operations, ω** (f,...,f), to other formulas (e.g. SUM(1,A1), and
  - **references** to cells, $a \in A$, (e.g. A1)

$$f \in Fml ::= v \mid \omega(f,\ldots,f) \mid a$$

The Spreadsheets Language

- $⊞_\alpha = A \to \alpha$ represents sheets indexed by addresses and storing values of type α
- A spreadsheet $⊞_{\mathbf{Fml}}$ is then simply a **sheet of formulas**
- Formulas **evaluate** to values Val
- The evaluation of a spreadsheet is a **value sheet**, which is a sheet of values $⊞_{\mathbf{Val}}$
- Semantics of a spreadsheet language are given by a function $[\![\cdot]\!] : ⊞_{Fml} \to ⊞_{Val}$ that maps spreadsheets to value sheets (common Excel behaviour)

Explanation Sheets Language

- **Value range**: $\underline{v} \in \underline{\mathrm{Val}} = \mathrm{Val} \times \mathrm{Val}$
- **Address range**: $\underline{a} \in \underline{A} = A \times A$ (e.g. [A1… B5])
- **Label**: $l \in \mathrm{Lab} = \mathrm{Val} \cup \mathrm{Val} \times \mathrm{Val}$ (e.g. Pay Rate, Adams.Pay Rate)
- **Unexplained**: $\perp$
- **Explanation**: $x \in \mathrm{Xpl} ::= v \mid \underline{v} \mid a \mid \underline{a} \mid l \mid \omega(x,\ldots,x) \mid \perp$
- **Explanation sheet**: $\boxplus_{\mathrm{Xpl}}$

	A	B	C	D	E	F	G
1				Payroll Spreadsheet			
2	Name	Pay Rate	Regular Hours	Overtime Hours	Regular Pay	Overtime Pay	Total
3	[Adams...Daniels]	[8.9...12.55]	[35...40]	[0...5]	=Pay Rate*Regular Hours	=Pay Rate*1.5*Overtime Hours	=Regular Pay+Overtime Pay
4							
5	Totals		=SUM(Regular Hours)	=SUM(Overtime Hours)	=SUM(Regular Pay)	=SUM(Overtime Pay)	=SUM(Total)

Spreadsheet Explanation

A **spreadsheet explanation** is captured by a ***zoom***

$$X \overset{\eta}{\lhd} S$$

which consists of

- an explanation sheet X
- a subject spreadsheet S and
- a total function η
  - that embeds the spreadsheet into the explanation: $\mathrm{dom}(\eta) = \mathrm{dom}(S) \wedge \mathrm{rng}(\eta) = \mathrm{dom}(X)$
  - and whose explanation formulas explain the formulas of the spreadsheet

Spreadsheet Explanation $X \overset{\eta}{\lhd} S$

- A zoom explains a number of similar cells by one cell
- When η^{-1} `(a) = {a`$_1$`,...,a`$_k$`}`, cell `a` summarizes, or explains, all the cells `a`$_1$`,...,a`$_k$
- We formalize this idea through the notion of formula explanation, which is defined as a binary relationship `x` $\lhd$ `f` that says an explanation formula `x` explains a spreadsheet formula `f` (see next slide)

Formula Explanations $X \overset{\eta}{\lhd} S$

Value

$$v \lhd v$$

Value Range

$$\frac{v_1 \leq v \leq v_2}{(v_1, v_2) \lhd v}$$

Address Range

$$\frac{a_1 \leq a \leq a_2}{(a_1, a_2) \lhd a}$$

Formula

$$\frac{x_1 \lhd f_1 \qquad \ldots \qquad x_n \lhd f_n}{\omega(x_1, \ldots, x_n) \lhd \omega(f_1, \ldots, f_n)}$$

Label

$$\frac{L(a) = \ell}{\ell \lhd a}$$

Empty Value

$$(v_1, v_2) \lhd \sqcup$$

Empty Formula

$$\omega(x_1, \ldots, x_n) \lhd \sqcup$$

Unexplained

$$\perp \lhd f$$

Explanation Semantics

Semantics of explanation formulas: $[\![\cdot]\!]$: Xpl → $\boxplus_{\text{Xpl}}$ → $\underline{\text{Val}} \cup \{\perp\}$

$$[\![v]\!]_X = (v, v) \qquad [\![\bar{v}]\!]_X = \bar{v} \qquad [\![a]\!]_X = [\![X(a)]\!]_X$$

$$[\![\bar{a}]\!]_X = \updownarrow\{[\![X(a)]\!]_X \mid a \in \rho(\bar{a})\} \qquad [\![\ell]\!]_X = \updownarrow L^{-1}(\ell)$$

$$\frac{[\![x_i]\!]_X = (v_i^1, v_i^2) \qquad v_i^1 \leq v_i \leq v_i^2}{[\![\omega(x_1, \ldots, x_n)]\!]_X = \updownarrow\{[\![\omega(v_1, \ldots, v_n)]\!]_X\}} \qquad [\![\perp]\!]_X = \perp$$

Semantics of explanation sheets: $[\![\cdot]\!]$: $\boxplus_{\text{Xpl}}$ → $\boxplus_{\underline{\text{Val}} \cup \{\perp\}}$

$[\![\cdot]\!]$ = $\{(a, \underline{v}_\perp) \mid (a,x) \in X \wedge [\![x]\!]_X = \underline{v}_\perp\}$

Zoom Soudness?

Initial Empirical Evaluation

- **Spreadsheets**
  - 4 spreadsheets from 3 different sources (EUSES, a book and Enron)

- **Participants**
  - 10 participants from UMinho and UNL
  - Computer science background
  - Most quite experienced spreadsheet users
  - 2 females and 8 males with ages ranging from 23 to 45

Execution

- Each participant received 2 spreadsheets and 2 explanation sheets (different order)
- **Q1** What is being calculated in row/column/cell X?
- **Q2** How are the values in row/column/cell X calculated?

Results

		average time		average score	
		subject	explanation	subject	explanation
A	Q1	1.3	2.1	2.2	2.4
	Q2	1.1	2.2	3.0	2.8
B	Q1	3.1	2.9	2.0	2.6
	Q2	2.5	3.7	2.0	1.8
C	Q1	2.1	1.8	3.0	1.0
	Q2	1.0	2.9	2.4	1.4
D	Q1	3.6	5.4	1.2	1.4
	Q2	6.8	3.3	1.8	2.0

Few Principles for other Explanation Languages

1. **Structure Preservation**
   - An explanation language should retain key subject language **structures** (e.g. modules)
   - Users are already **familiar** with these structures
   - Reused structures facilitate the **alignment** of explanations with subject programs

2. Abstraction

- An explanation language should aim at **high-level** descriptions
- It should **abstract from details** of the subject language (e.g. syntax)
- Abstraction makes explanations **faster** to read/absorb
- It also allows to **summarize** subject programs

3. Partiality

- An explanation language should support **partial** explanations
- Should not be required to cover all of a subject program
- Partiality supports a **gentle slope** approach to explanations
- Incremental construction of more complete explanations
- Partiality allows one to ignore parts that cannot be explained (e.g. they are not understood, trivial or unimportant)

4. Compositionality

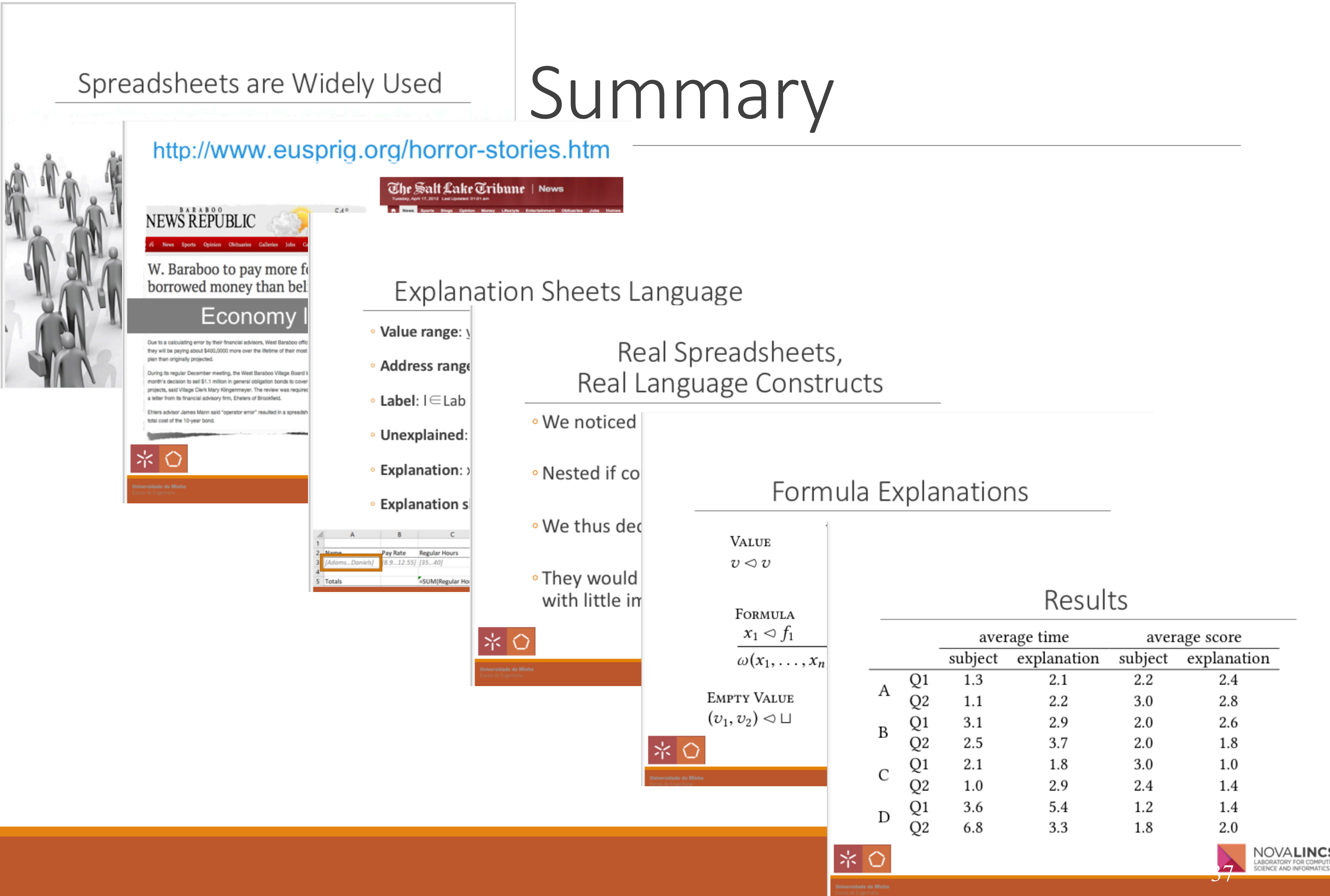
- An explanation language should support constructing bigger explanations from smaller ones

- Requires composition operators

- Supports the systematic construction of explanations and the reuse

- Together with partiality, compositionality supports the distributed creation of explanations by different people who understand different parts of the subject program

Summary



NOVA LINCS LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

Future…

- Implement automatic inference of explanation sheets from existing spreadsheets
- Further empirical evaluation and improvement of language and explanation sheets

Thank you!

Questions?

Universidade do Minho
Escola de Engenharia

Oregon State University

NOVA LINCS
LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

Explaining Spreadsheets with Spreadsheets

[§] Jácome Cunha	jacome@di.uminho.pt
[†] Mihai Dan	danm@oregonstate.edu
[†] Martin Erwig	erwig@oregonstate.edu
[†] Danila Fedorin	fedorind@oregonstate.edu
[†] Alex Grejuc	grejuca@oregonstate.edu

[§] University of Minho & NOVA LINCS

[†] Oregon State University

GPCE 2018, BOSTON, MA, USA, 5-6 NOVEMBER