Spreadsheet Understandig

A Survey

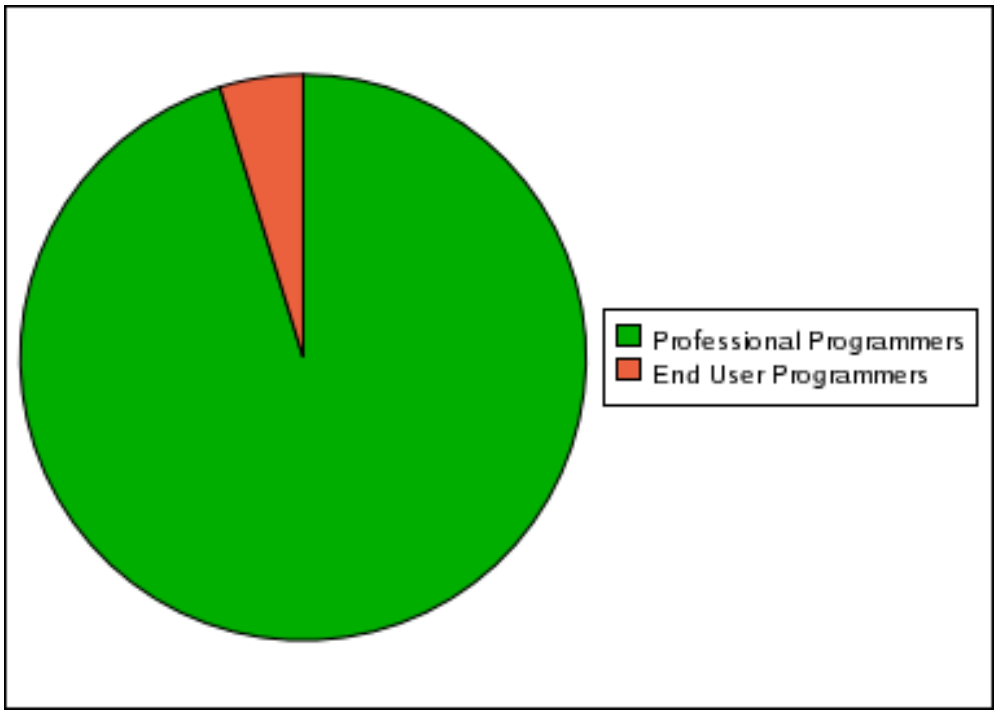
Alexandra Silva

Departamento de Informática
Universidade do Minho

PURe Workshop, October 2005

Big Reasons

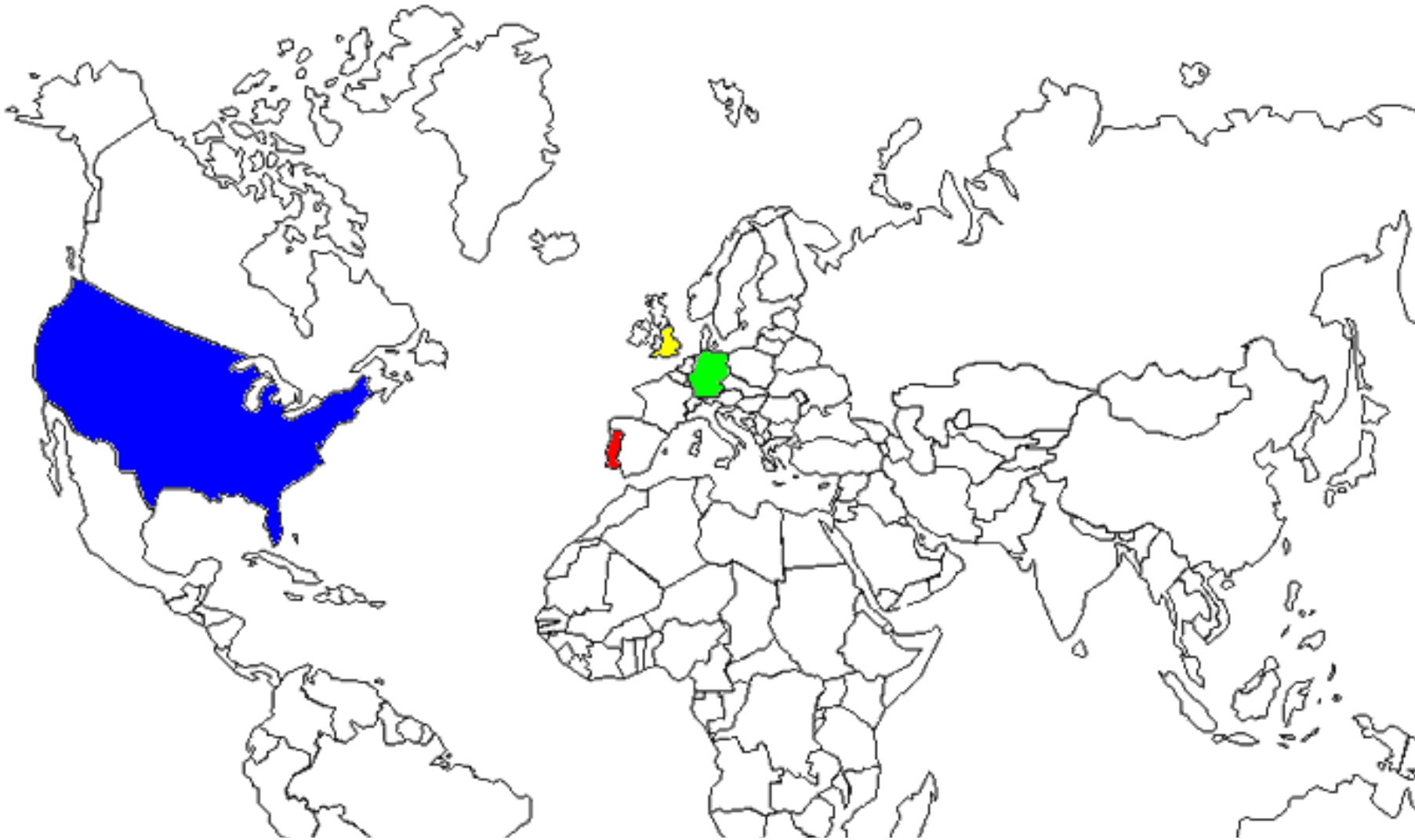
- End user programers



Big Reasons

- End user programers
- Money – errors in spreadsheets cost millions of dollars to enterprises.

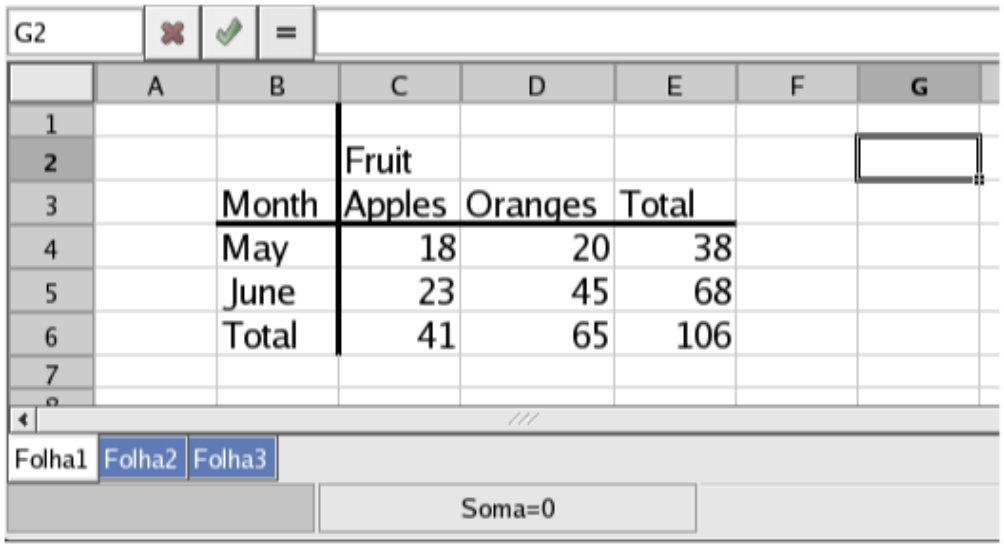
Research all over the world



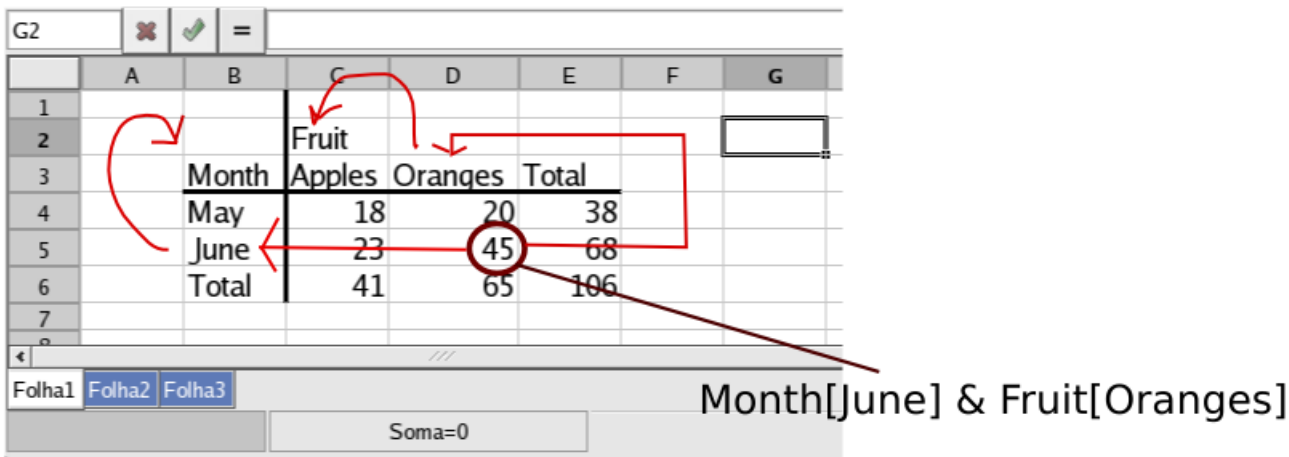
- Martin Erwig and Margaret Burnett

- System of units in spreadsheets;
- Identification of errors based on bad combination of units;
- Automatic generation and mainteinance of spreadsheets using *templates*.

Adding apples and oranges

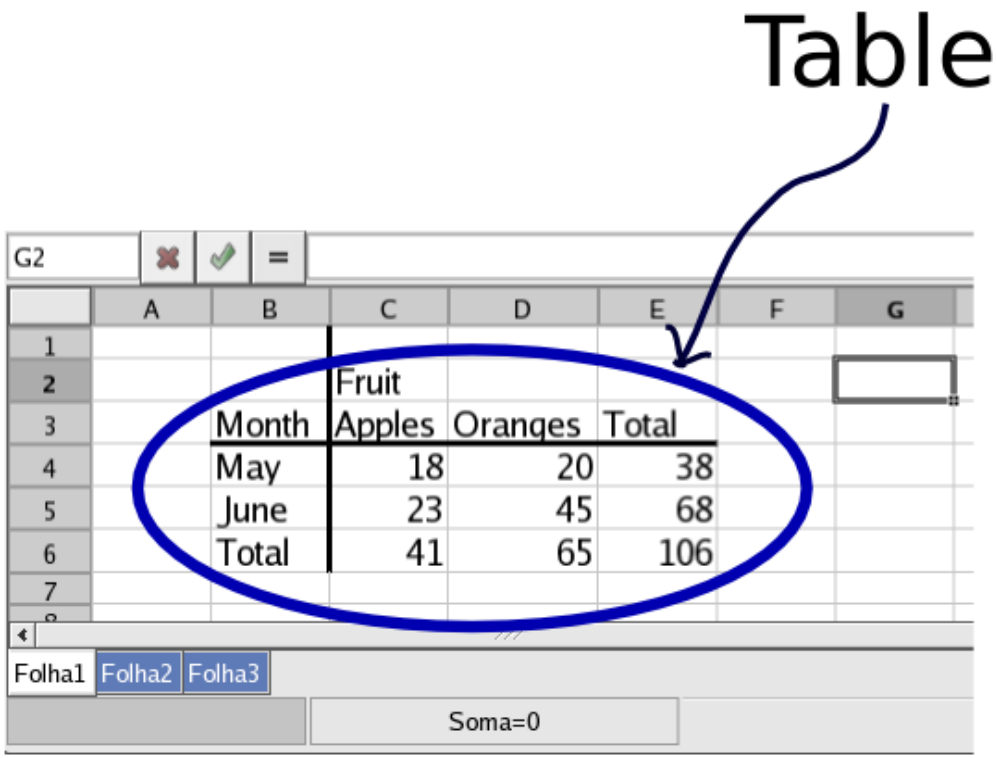


Unit system

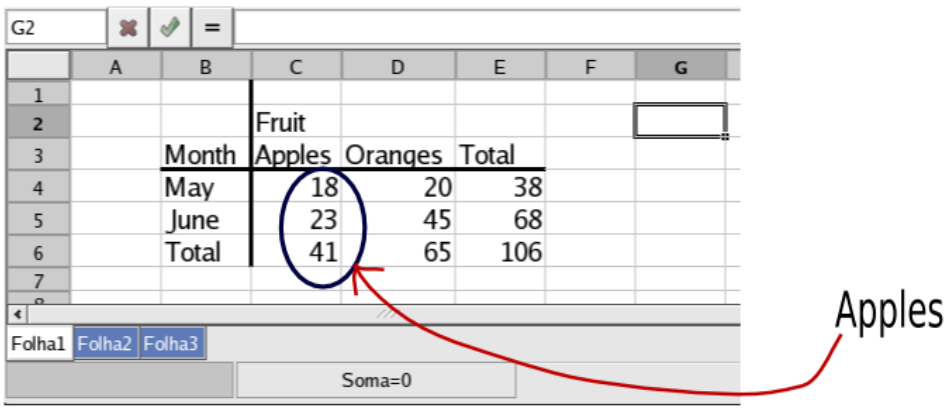

G2
=
A
B
C
D
E
F
G
1
2
3
4
5
6
7
Fruit
Month
Apples
Oranges
Total
May
18
20
38
June
23
45
68
Total
41
65
106
Folha1
Folha2
Folha3
Soma=0
Month[June] & Fruit[Oranges]

Visually Customizing Inference Rules About Apples and Oranges

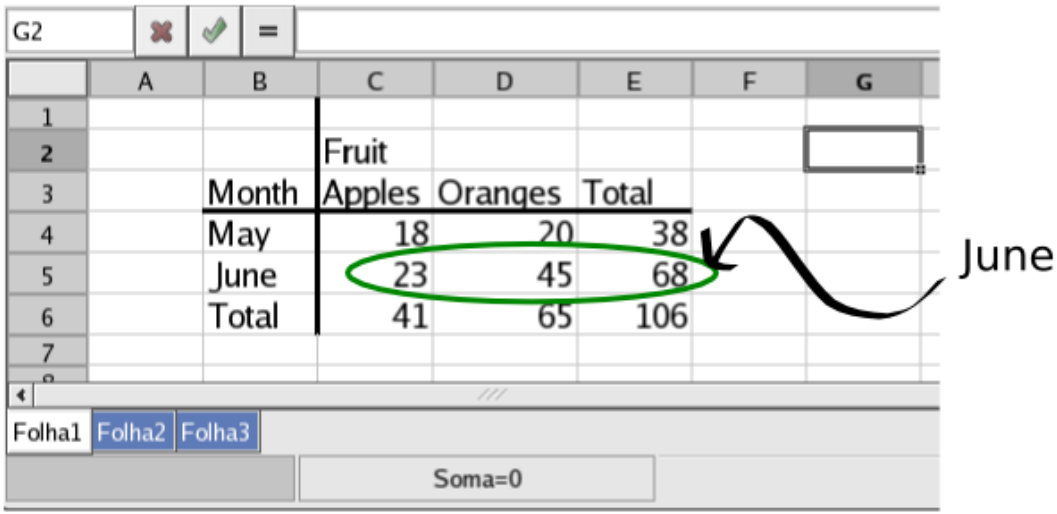
- Borders identified



- Groups data


G2
A
B
C
D
E
F
G
Fruit
Month
Apples
Oranges
Total
May
18
20
38
June
23
45
68
Total
41
65
106
Folha1
Folha2
Folha3
Soma=0
Apples

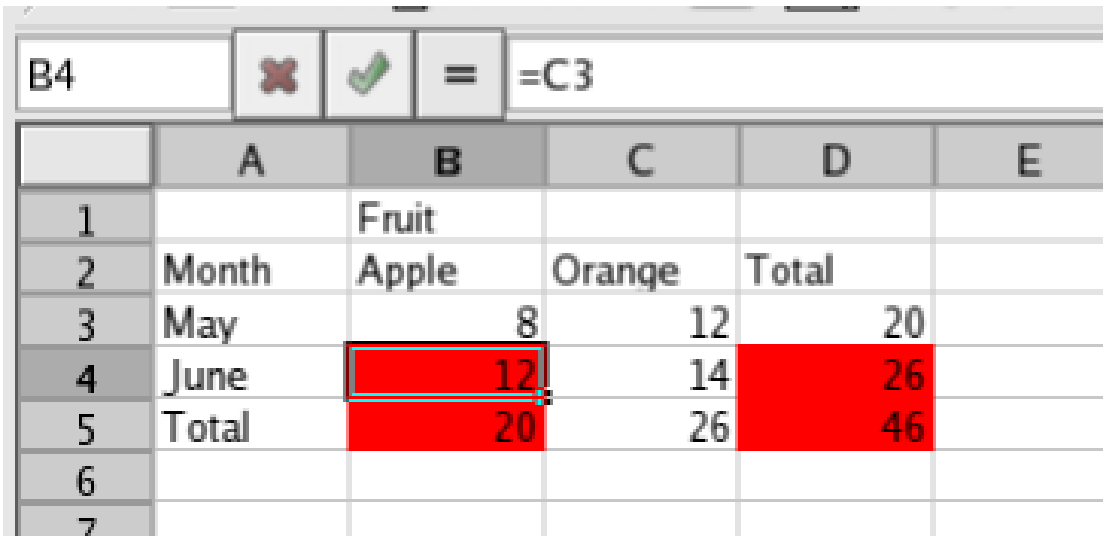
- Groups data


G2
A
B
C
D
E
F
G
Fruit
Month
Apples
Oranges
Total
May
18
20
38
June
23
45
68
Total
41
65
106
June
Folha1
Folha2
Folha3
Soma=0

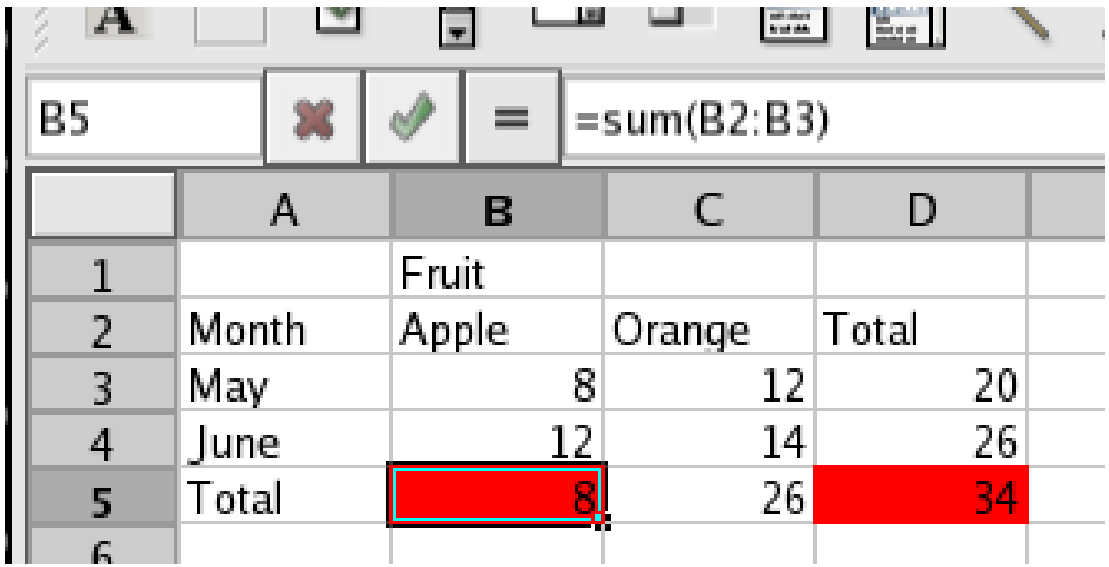
Header and Unit Inference for Spreadsheets Trough Spatial Analysis

- Integrated into Excel
- Identifies reference errors, range errors, etc

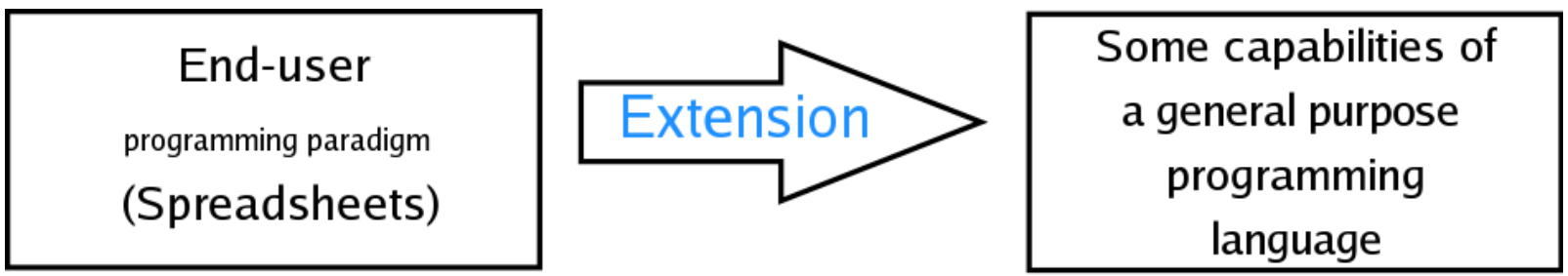
- Identifies **reference errors**, range errors, etc



- Identifies reference errors, **range errors**, etc



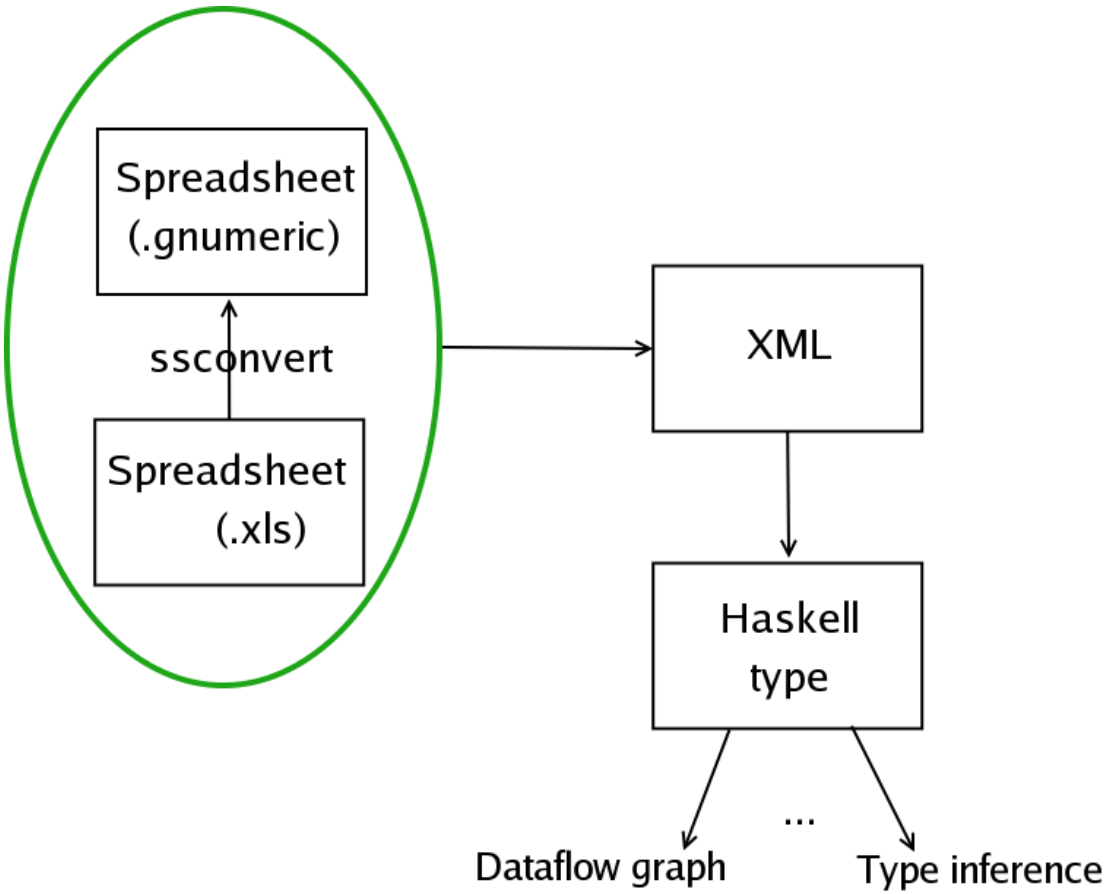
Simon Peyton Jones *et al.*

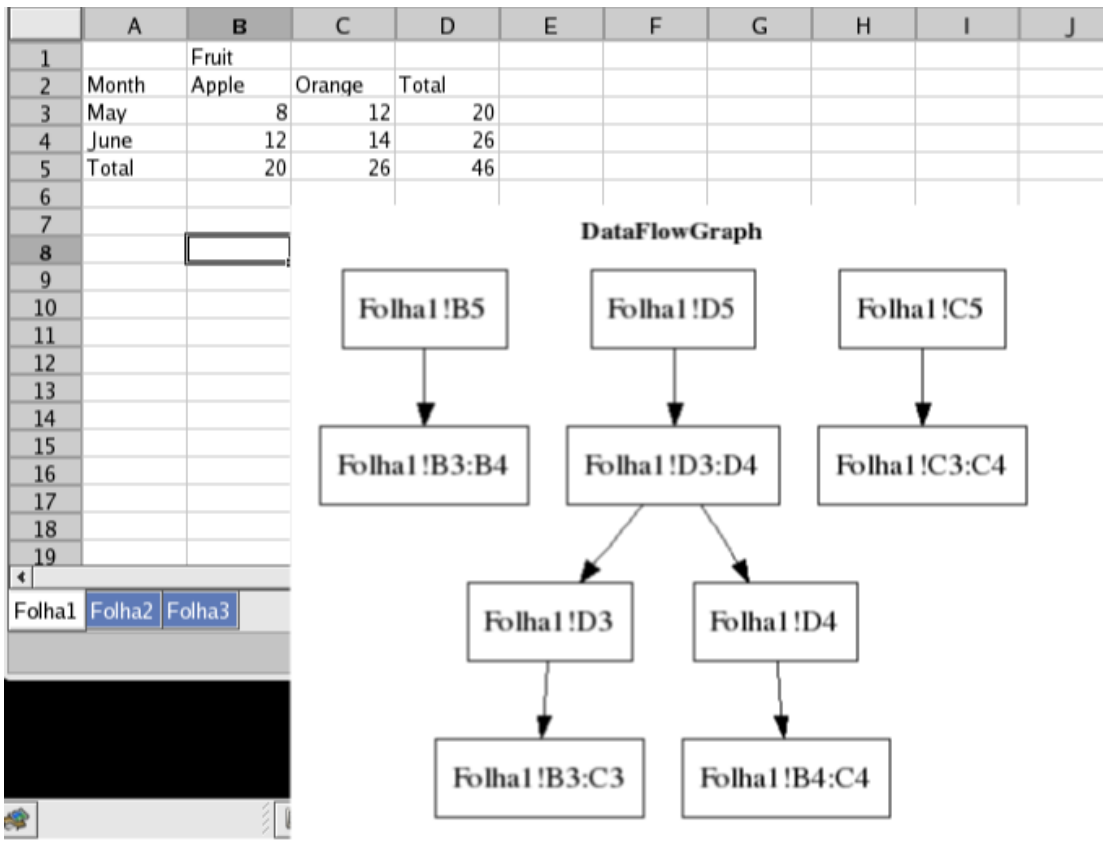

End-user
programming paradigm
(Spreadsheets)
Extension
Some capabilities of
a general purpose
programming
language

- User defined functions
- Matrices

- Unit System
- Identification of tables in spreadsheets (**Ucheck**)
- User defined functions in Excel

- Dataflow inside spreadsheets
- Database pointfree theory
- Functional Dependencies


Spreadsheet
(.gnumeric)
ssconvert
Spreadsheet
(.xls)
XML
Haskell
type
...
Dataflow graph
Type inference


A
B
C
D
E
F
G
H
I
J
Fruit
Month
Apple
Orange
Total
May
8
12
20
June
12
14
26
Total
20
26
46
Folha1
Folha2
Folha3
DataFlowGraph
Folha1!B5
Folha1!D5
Folha1!C5
Folha1!B3:B4
Folha1!D3:D4
Folha1!C3:C4
Folha1!D3
Folha1!D4
Folha1!B3:C3
Folha1!B4:C4

. . . and theory becomes simple and concise

Relational database table $\equiv$ set of $n-$tuples

Relational database table $\equiv$ set of $n-$tuples

Number	Name	Mark
35820	"Ana Quintela Castro Ferreira"	18
33567	"David Filipe Oliveira Costa"	17
35672	"Joana Figueiredo Martins"	12
...	...	...

$\equiv$

{ (35820, "Ana Quintela Castro Ferreira", 18), (33567, "David Filipe Oliveira Costa", 17), (35672, "Joana Figueiredo Martins", 12)}

S = {Number, Name, Mark } is the *schema* of the table.
Number uniquely determines Name.
Name is **functionally dependent** of Number.

Functional dependency – Traditional definition

Given subsets $x, y \subseteq S$ of the relation scheme S of a relation R, this relation is said to satisfy functional dependency $x \rightharpoonup y$ iff all pairs of tuples $t, t' \in R$ which "agree" on x also "agree" on y, that is,

$$\langle \forall t, t' : t, t' \in R : t[x] = t'[x] \Rightarrow t[y] = t'[y] \rangle$$

(Notation $t[x]$ meaning "the values in t of the attributes in x" will be scrutinized in the sequel)

Functional dependency – Pointfree definition

We say that relation $R : B \leftarrow A$ satisfies the $f \rightharpoonup g$ functional dependency iff $g \leq f.R^{\circ}$ or, equivalently, $ker(f.R^{\circ}) \subseteq kerg$.

What's next

- To define $n - ary$ tuples (HList)
- Functional dependencies pointfree calculus
- Conceptual analysis(?)

The team




Contributers are welcome:)

Questions?