“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”
H.G. Wells

Why Statistical Thinking is Transforming Programming Language Research

And why the dream of usable probabilistic programming in the spreadsheet is getting closer…

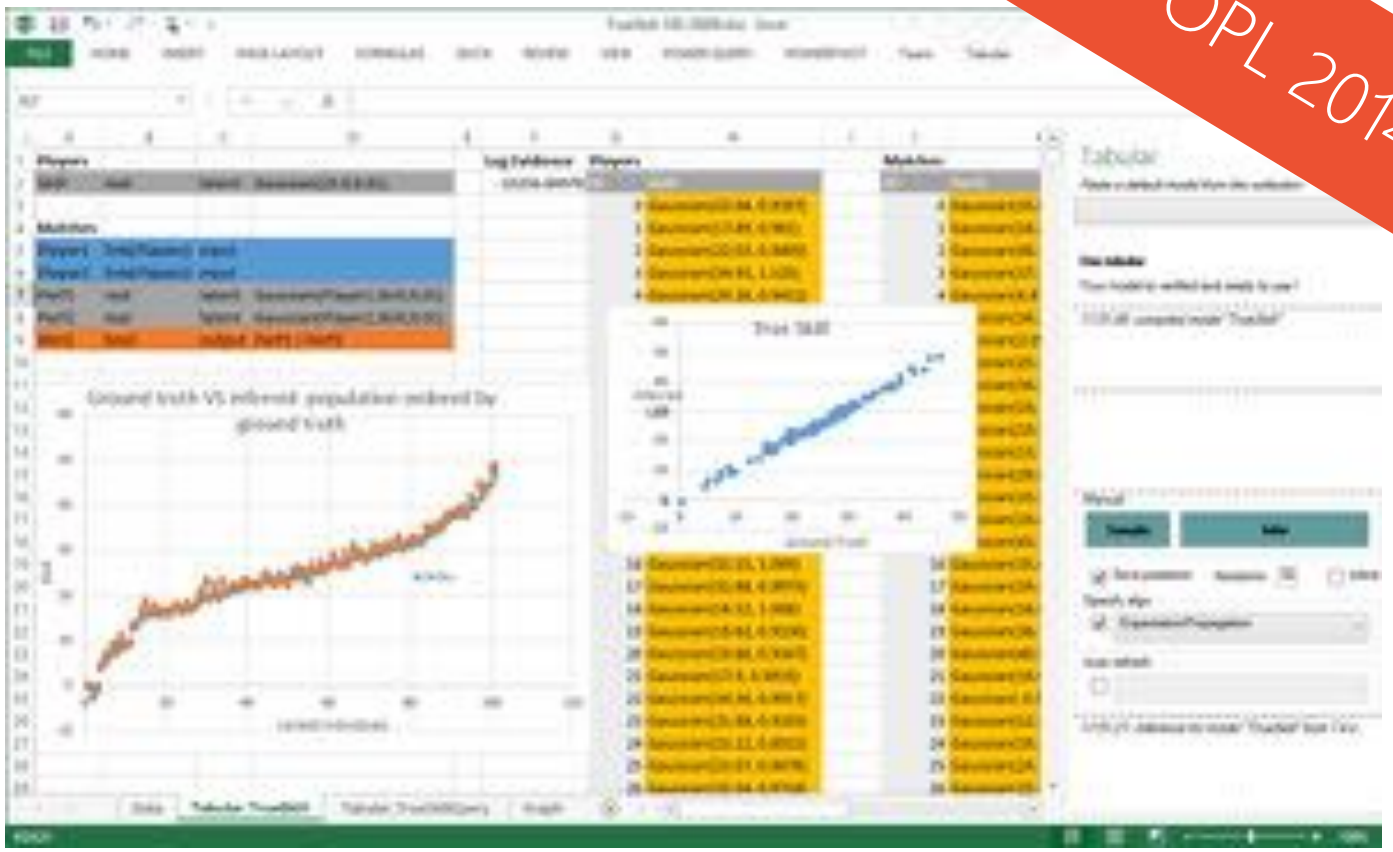
Andrew D. Gordon, Microsoft Research and University of Edinburgh

Based on joint work at Microsoft Research with Judith Borghouts, Dany Fabian, Matt McCutchen, Simon Peyton Jones, Claudio Russo, Advait Sarkar, Sruti Srinivasa Ragavan, Neil Toronto, Jack Williams, Nick Wilson

Colloquium on Probabilistic Programming
Collège de France, June 29 and 30, 2022

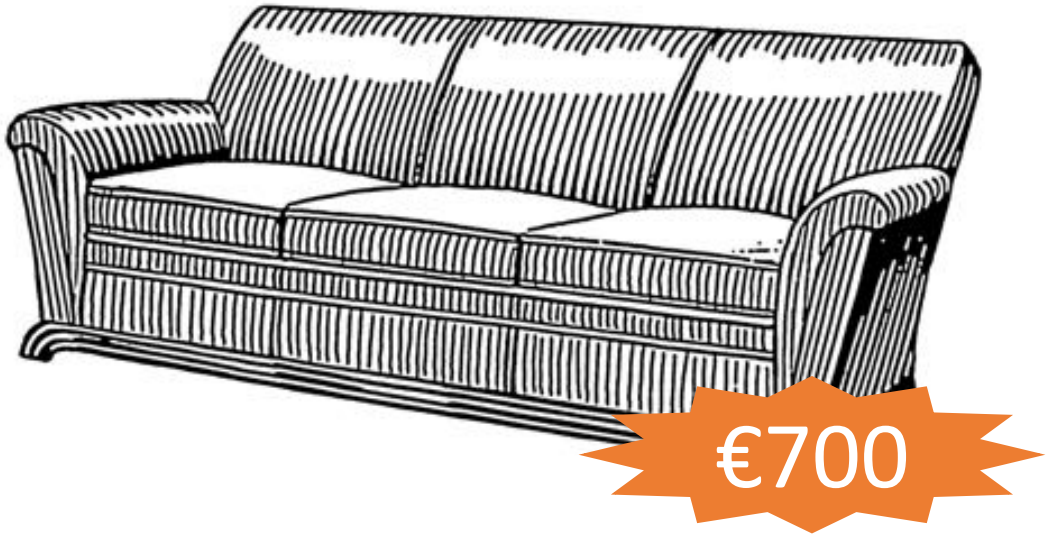
The Dream of Probabilistic Programming in the Spreadsheet

POPL 2014



To empower **more people**
to get **more value** from data
via **statistical thinking**
and **representing uncertainty**

Clara decides whether to buy her sofa



	A	B	C	D	E	F
1					€ 2,000.00	Income
2						
3	Rent			€ 1,100.00		
4	Commute			€ 85.00		
5	Utilities	€ 50.00	€ 150.00	€ 107.05	=range(B5,C5)	
6	Sofa			€ 700.00		
7						
8					€ 1,992.05	Expenses
9					€ 7.95	Balance
10					1	Above zero?

Today

	A	B	C	D	E	F
1					€ 2,000.00	Income
2						
3	Rent			€ 1,100.00		
4	Commute			€ 85.00		
5	Utilities	€ 50.00	€ 150.00	€ 75.46	=range(B5,C5)	
6	Sofa			€ 700.00		
7						
8					€ 1,960.46	Expenses
9					€ 39.54	Balance
10					1	Above zero?
11						
12				1000 3 fields		=Sample(AboveZero,D12)
13					78%	=E12.expectation
14						
15						

New: cell holds a **probability distribution**

New: Clara's formula tells her she's a 75% chance of staying solvent

Bonnie Nardi

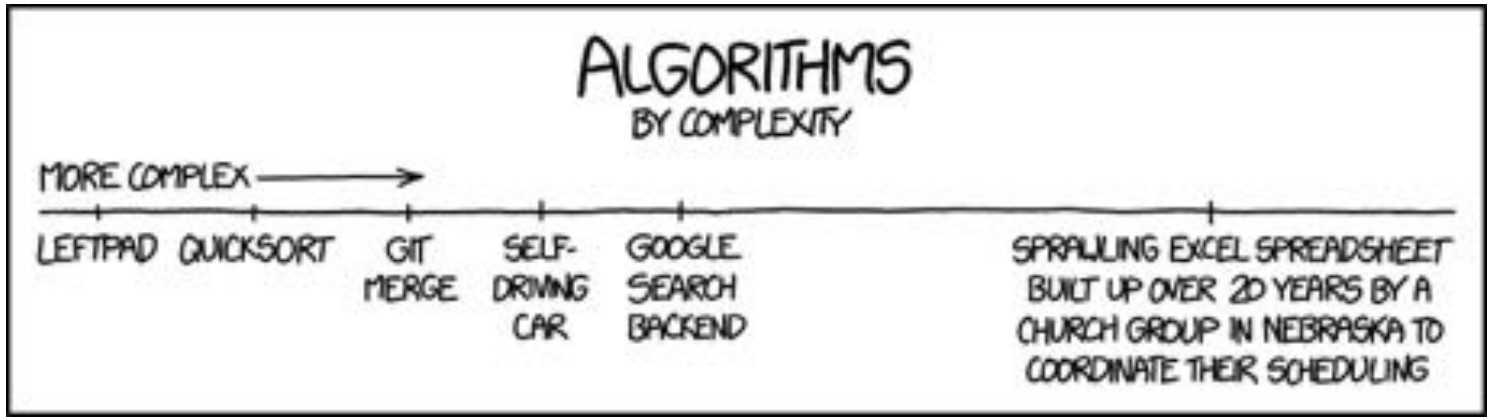
End-user programming

"Programming to achieve the result of a program primarily **for personal, rather than public** use. [Nardi 1993]

End-user programmers might be secretaries, accountants, children, teachers, interaction designers, scientists, or anyone who finds themselves writing programs to support their work or hobbies." [Ko et al, 2011]

Spreadsheets are code!

- Ambitious spreadsheets have thousands of formulas
- Sometimes developed over years



https://xkcd.com/1667/

Sheet defined functions (SDFs)

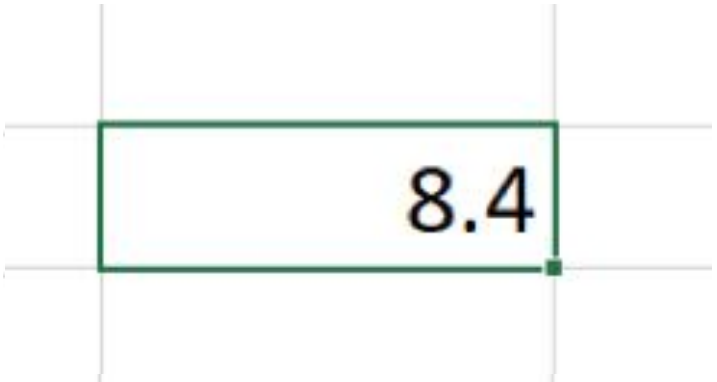
2003

Peyton Jones, S.L., Blackwell, A.F., Burnett, M.M.:
A user-centred approach to functions in Excel. ICFP 165–176 (2003)

1: define function

	A	B	C	D	E	F
1	**Average function**					
2						
3	**Input:**	4		**Count:**	5	=COUNT(B3:B7)
4		5		**Sum:**	42	=SUM(B3:B7)
5		23		**Average:**	8.4	=E4/E3
6		4				
7		6				

2: use like any other

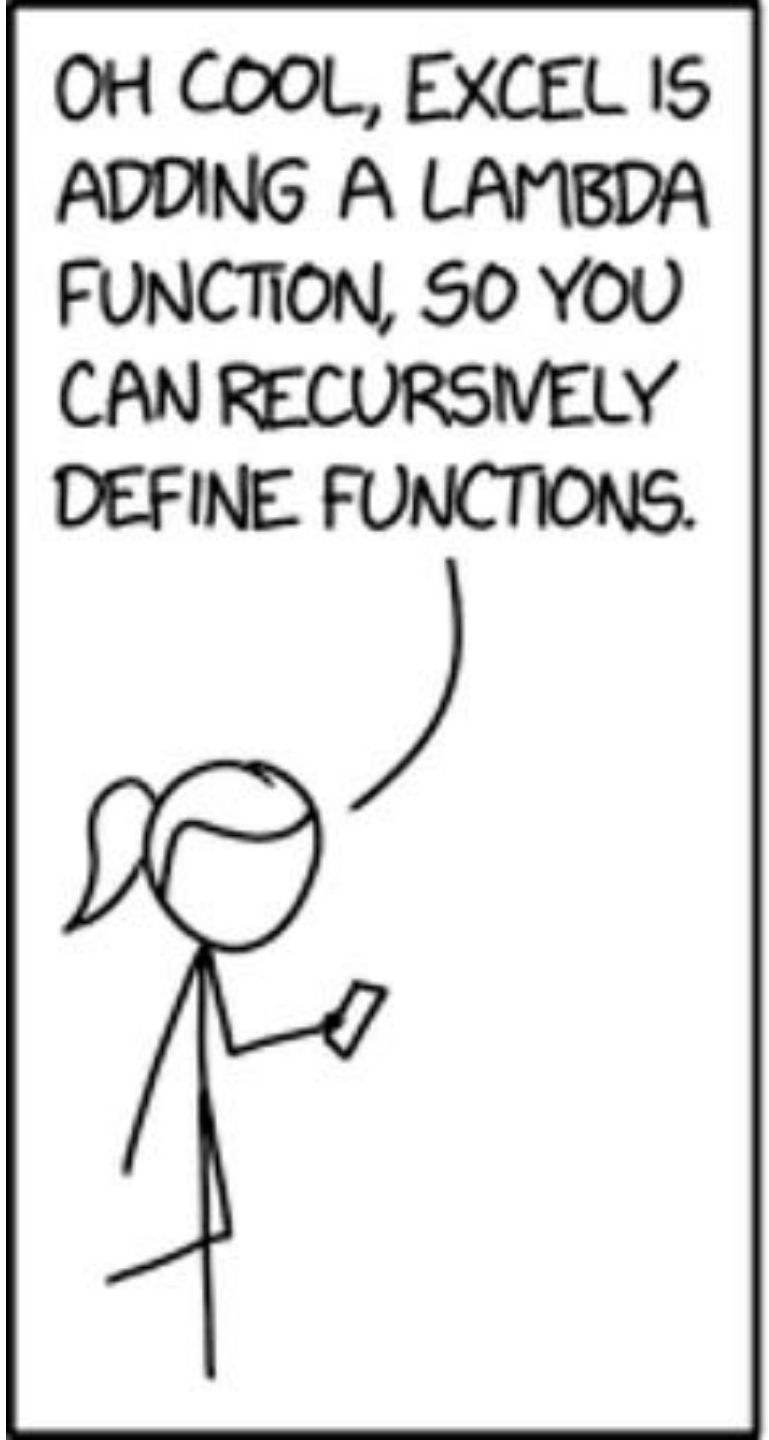


Aim of Microsoft's Project Yellow

Remove the "glass ceiling" that limits the scope and reach of what an end-user programmer can do with Excel:

- **Make Excel functions reflect the abstractions of our end users**, by allowing end-users to define new functions using ordinary formulas.

- **Make Excel's data values reflect the datatypes of our end users' domains**, by adding arrays, vectors, records, and hence represent domain-specific data types such as **probability distributions**.



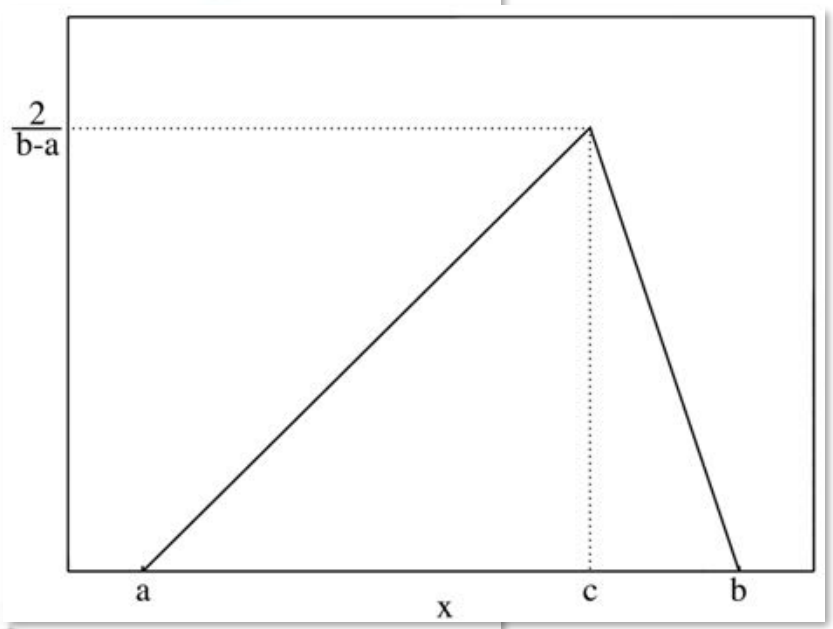
XKCD 2021

https://xkcd.com/2453/

New function *range*(*a*,*b*) via LAMBDA

Workbook Lib + New

```
1   triangular = lambda(a,b,_c,
2     if(NOT(AND(a<b, a<=_c, _c<=b)), "triangular: invalid arguments"
3       , let( U, RAND()
4            , F, (_c-a)/(b-a)
5            , if( U<F
6                , a + SQRT(U * (b-a) * (_c-a))
7                , b - SQRT((1-U) * (b-a) * (b-_c))
8                )
9            )
10      )
11  );
12
13  range = lambda(a,b, triangular(a,b,average(a,b)));
```



€ 50.00	€ 150.00	€ 107.05	=range(B5,C5)

```
AboveZero = LAMBDA(
    LET(
        _E1, 2000,
        _D3, 1100,
        _D4, 85,
        Utilities, 50,
        _C5, 150,
        _D5, range(Utilities, _C5),
        _D6, 700,
        _E8, SUM(
            MAKEARRAY(
                4,
                1,
                LAMBDA(i, j,
                    CHOOSE(
                        (i - 1) * 1 + j,
                        _D3,
                        _D4,
                        _D5,
                        _D6
                    )
                )
            )
        ),
        _E9, _E1 - _E8,
        IF(_E9 > 0, 1, 0)
    )
);
```

AFE compiles probabilistic **sheet-defined function** to a LAMBDA

Join us!

B	C	D	E	F
			€ 2,000.00	Income
		€ 1,100.00		
		€ 85.00		
€ 50.00	€ 150.00	€ 75.46	=range(B5,C5)	
		€ 700.00		
			€ 1,960.46	Expens
			€ 39.54	Balance
			1	Above zer
		1000	3 fields	=Sample(AboveZero,D12)
			78%	=E12.expectation

Simple **inference** coded as a LAMBDA

New Idea: SDFs enable probabilistic programming in spreadsheet

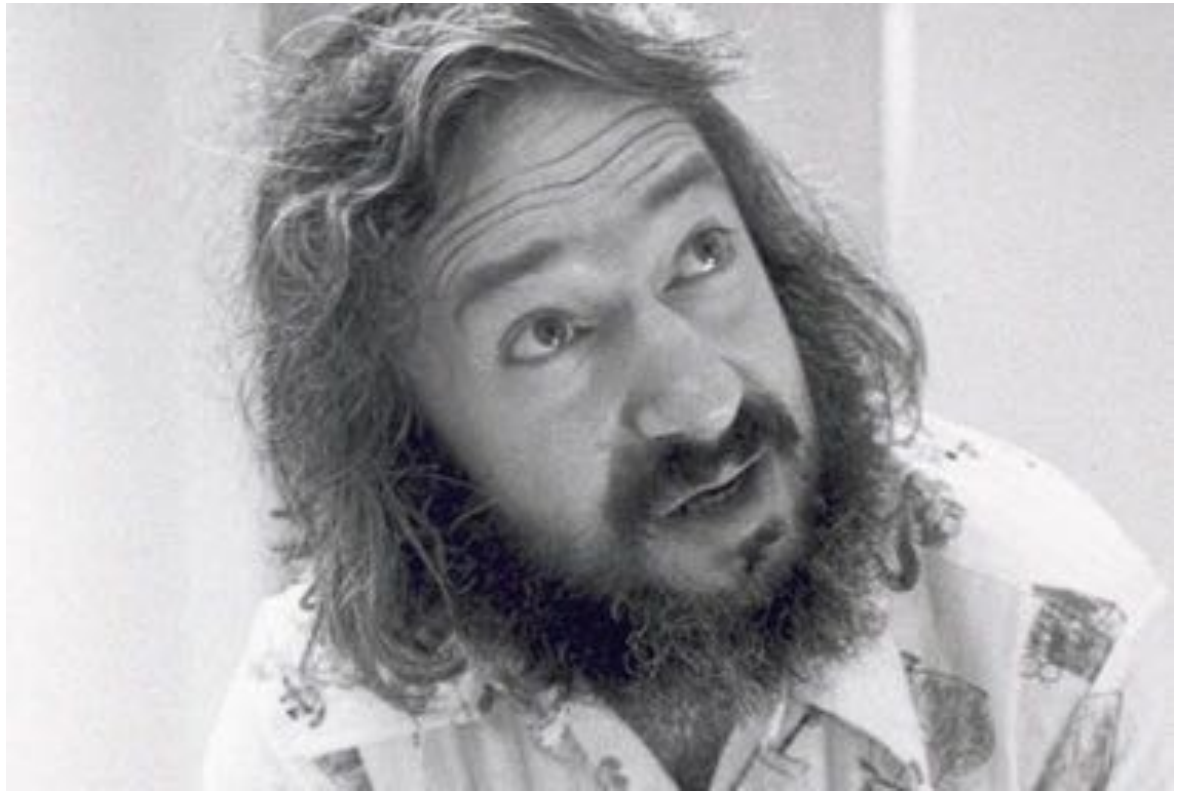
Recap

The dream of usable **probabilistic programming in the spreadsheet**

- was there since the 1980s because of RAND(),
- but getting easier because of transforming the spreadsheet programming language with **general constructs like compound data, LAMBDA, SDF**

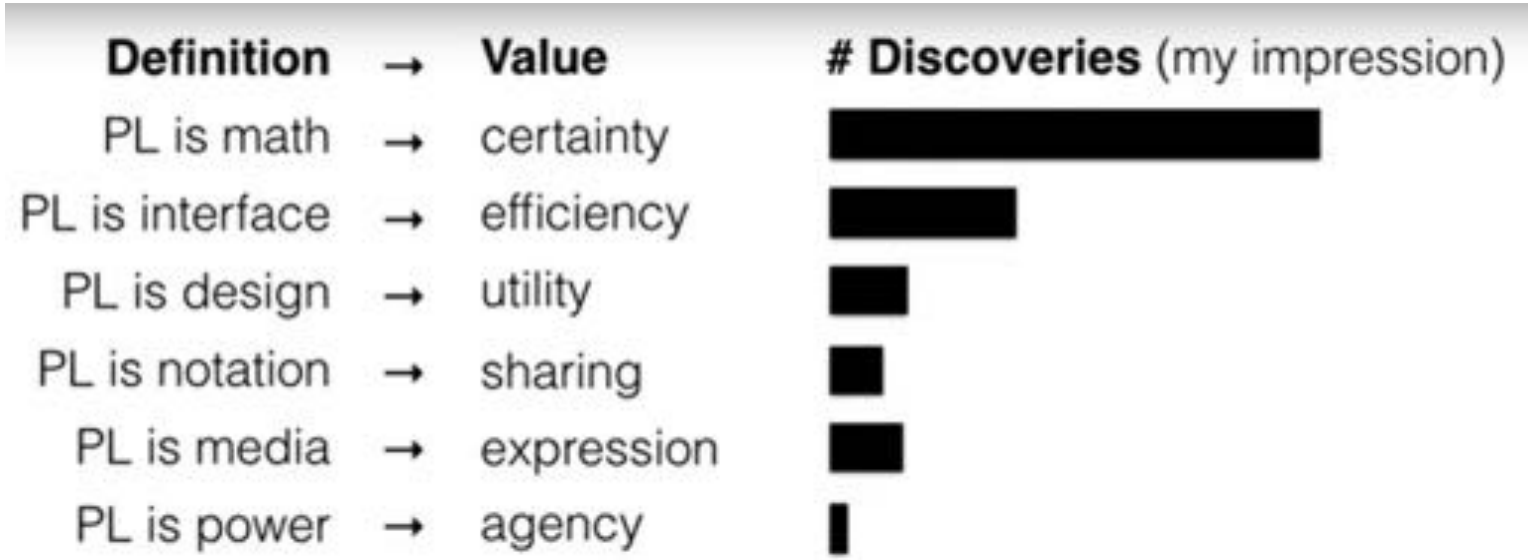
But why is statistical thinking transforming programming language research?


SPLASH 2016
Amy J. Ko, “A human view
of programming languages”



- “…equal access to even the most basic elements of computation requires an epistemological pluralism, **accepting the validity of multiple ways of knowing and thinking**.”

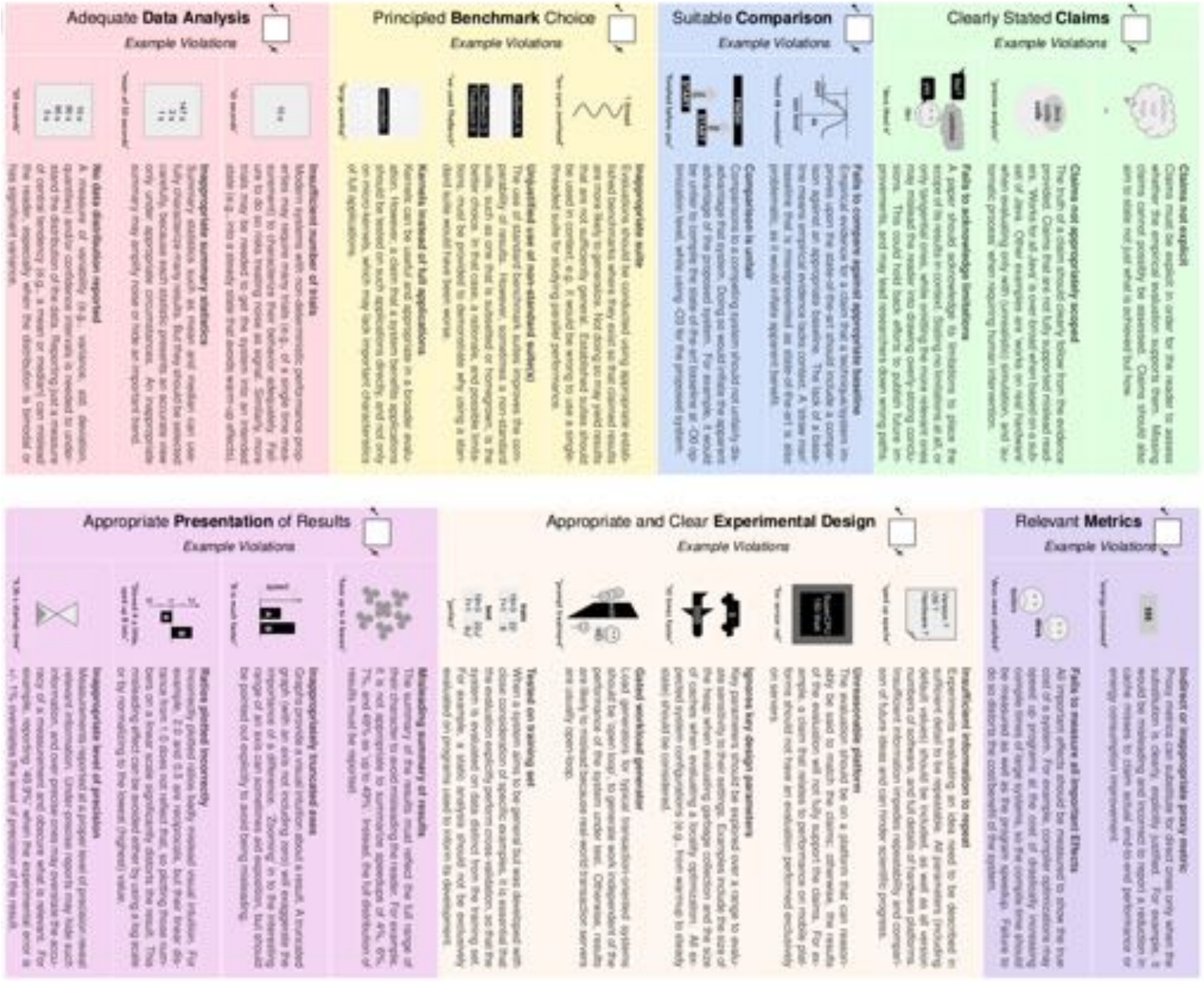
Sherry Turkle and Seymour Papert, "Epistemological pluralism: styles and voices within the computer culture", Signs: Journal of Women in Culture and Society 1990:16(1)



"What values are behind our research? When I look at my impression of where we spend time, most of our discoveries are about certainty, **but there are a lot of values here that we don't explore all that deeply**."

Amy J. Ko

SIGPLAN 2017



A Checklist Manifesto for Empirical Evaluation: A Preemptive Strike Against a Replication Crisis in Computer Science

by Emery D. Berger, Stephen M. Blackburn, Matthias Hauswirth, Michael W. Hicks

https://www.sigplan.org/Resources/EmpiricalEvaluation/

SIGPLAN 2017

"Because **user studies are currently relatively infrequent** in the papers we examined, we have not included them among the category examples."

https://www.sigplan.org/Resources/EmpiricalEvaluation/

A Glorious Dream… but no user study

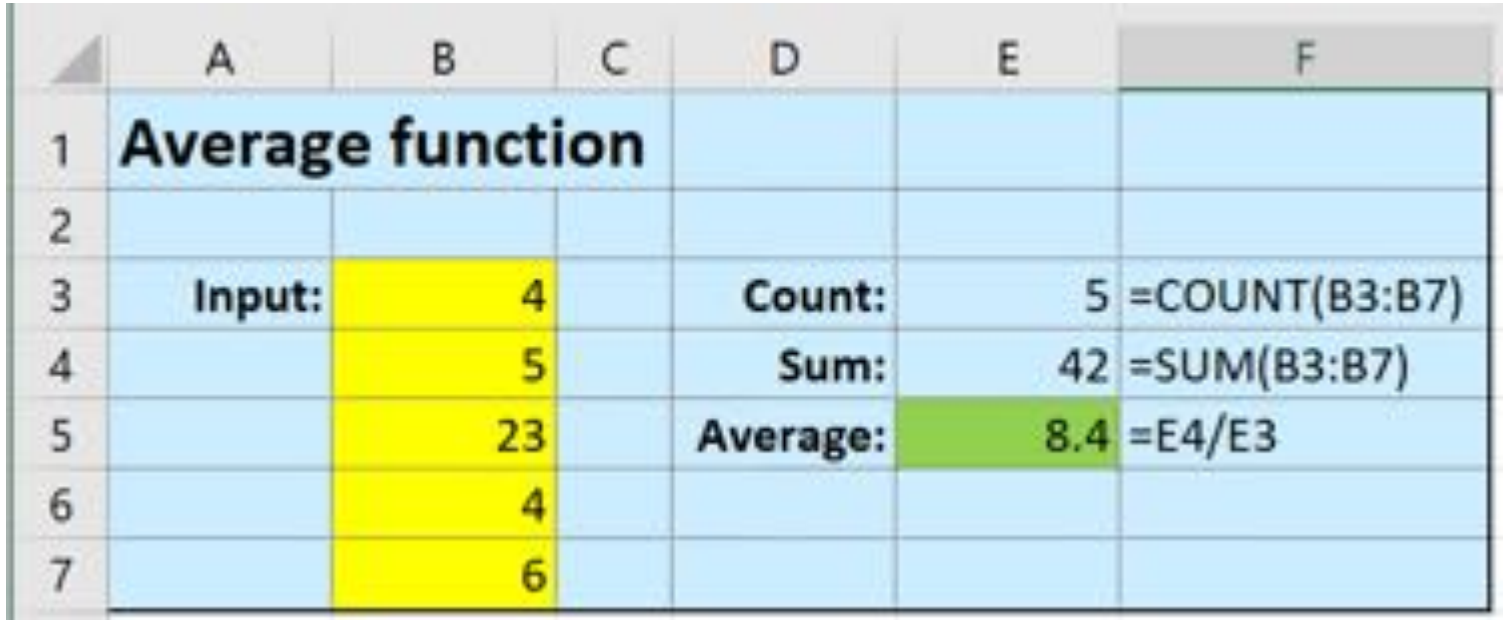
POPL 2014

JFP 2020

Elastic Sheet-Defined Functions

An example of quantitative statistical thinking in programming languages

Elastic SDFs



	A	B	C	D	E	F
1	**Average function**					
2						
3	**Input:**	4		**Count:**	5	=COUNT(B3:B7)
4		5		**Sum:**	42	=SUM(B3:B7)
5		23		**Average:**	8.4	=E4/E3
6		4				
7		6				

```
function AVERAGE( B3:B7 ) returns E5 {
    E3 = COUNT( B3:B7 )
    E4 = SUM( B3:B7 )
    E5 = E4/E3
}
```

Problem:

- =AVERAGE(X5:X7) too small!
- =AVERAGE(G2:G200) too big!

User’s eye view

- Write a function with fixed-size inputs, using familiar copy/paste
- Magic happens
- The function works on input of arbitrary size

Main point: we think that automatically inferred elasticity will dramatically broaden the audience that can use SDFs effectively.

Participants

- 20 participants
- Students
  (Computer Science, Statistics, Management, Mathematics)
- 10 participants industry experience
  (teacher, administrator, economist, consultant)

Procedure

1. Video tutorial of ~10 minutes
2. Elastic SDFs
   - 1 practice task
   - 3 study tasks
   - Cognitive load questionnaire
3. Array SDFs
4. Semi-structured interview

Findings from user study (N=20, 7 female)

People perceived **significantly lower cognitive workload** for elastic SDFs than with SDFs based on map/reduce.

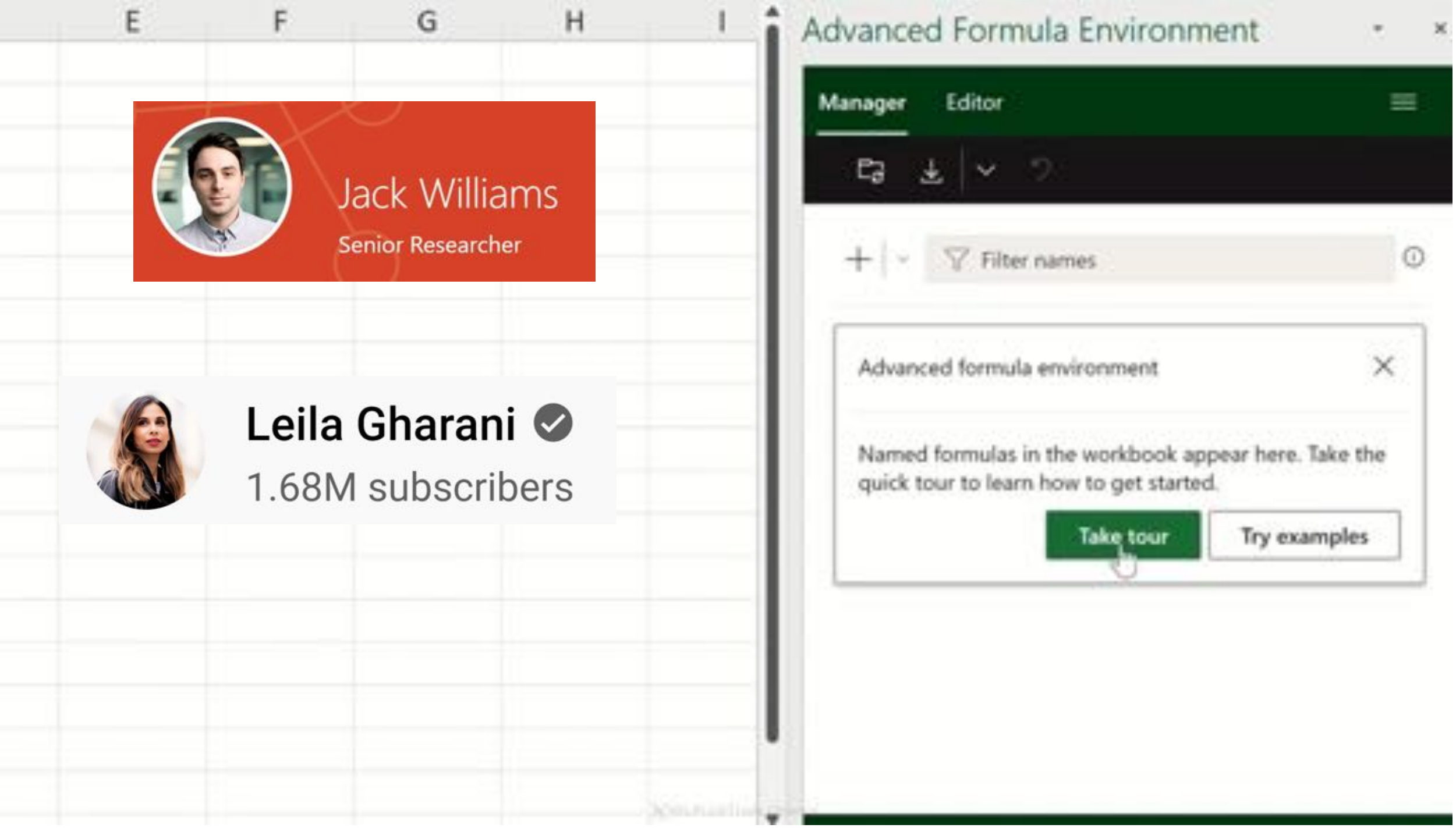
*I think elastic functions are easier to work with, also with the "mental model" that you have of Excel, because you can more just follow your normal Excel **workflow**. – P9*

*It'd be nice to have this kind of **middle ground**, of not having to write the same things over and over again, but not having to persuade someone to make a macro either. – P7*

VL/HCC 2022

Reactions to LAMBDA in Online Forums

An example of **qualitative** statistical thinking in programming languages


E
F
G
H
I
Jack Williams
Senior Researcher
Leila Gharani
1.68M subscribers
Advanced Formula Environment
Manager
Editor
Filter names
Advanced formula environment
Named formulas in the workbook appear here. Take the quick tour to learn how to get started.
Take tour
Try examples


CALCULATE
YEAR TO DATE
+
NEW FUNCTION!
SCAN


RECURSIVE
LAMBDAS


EXCEL λ
LAMBDA
EXPLAINED


BRAND
NEW
FUNCTIONS
BYROW
BYCOL

VL/HCC 2022

A **thematic analysis** of nearly 2,700 comments posted on the Reddit, Hacker News, YouTube, and Microsoft Tech Community online forums, regarding LAMBDA in spreadsheets.

Findings:

- computational abstractions are viewed both as helpful and harmful,
- users encounter learning and understanding barriers to applying them,
- there are deficiencies and opportunities in tooling such as in formula editing, versioning, reuse and sharing.

LAMBDA prompts new debate around:

- whether spreadsheets are code,
- whether writing formulas can be considered programming
- whether spreadsheet users identify themselves as programmers.

Advait Sarkar, Sruti Srinivasa Ragavan, Jack Williams, Andrew D. Gordon, "**End-user encounters with lambda abstraction in spreadsheets: Apollo's bow or Achilles' heel?",** VL/HCC'22, to appear.

V. Braun and V. Clarke, "**Using thematic analysis in psychology**," Qualitative research in psychology, 3(2):77–101, 2006

Embrace the human in programming research

Statistical thinking is transforming programming language research

- Quantitative methods assess effectiveness of new features with users
- Qualitative methods find patterns across population of users

Consider assessing your research with **human-centric methods**

If you review for PL venues, **embrace the diversity of research disciplines**

Empower people with statistical thinking

The dream of usable **probabilistic programming in the spreadsheet**

- was there since the 1980s because of RAND(),
- but getting easier because of transforming the spreadsheet programming language with **general constructs like compound data, LAMBDA, SDF**

Come join us

https://aka.ms/CalcIntel