OREGON PROGRAMMING LANGUAGES SUMMER SCHOOL

# End-User Probabilistic Programming

Andrew D. Gordon

Based on joint work with Judith Borghouts, Advait Sarkar, and Neil Toronto

Part 3 of “Empowering Spreadsheet Users with Probabilistic Programming”

# Pop Quiz

Your project is comprised of four independent, parallel subprojects. Each subproject will take between 1 and 2 months to complete, and each has an average completion time of 1.5 months.

On average, how long will your project take to complete?

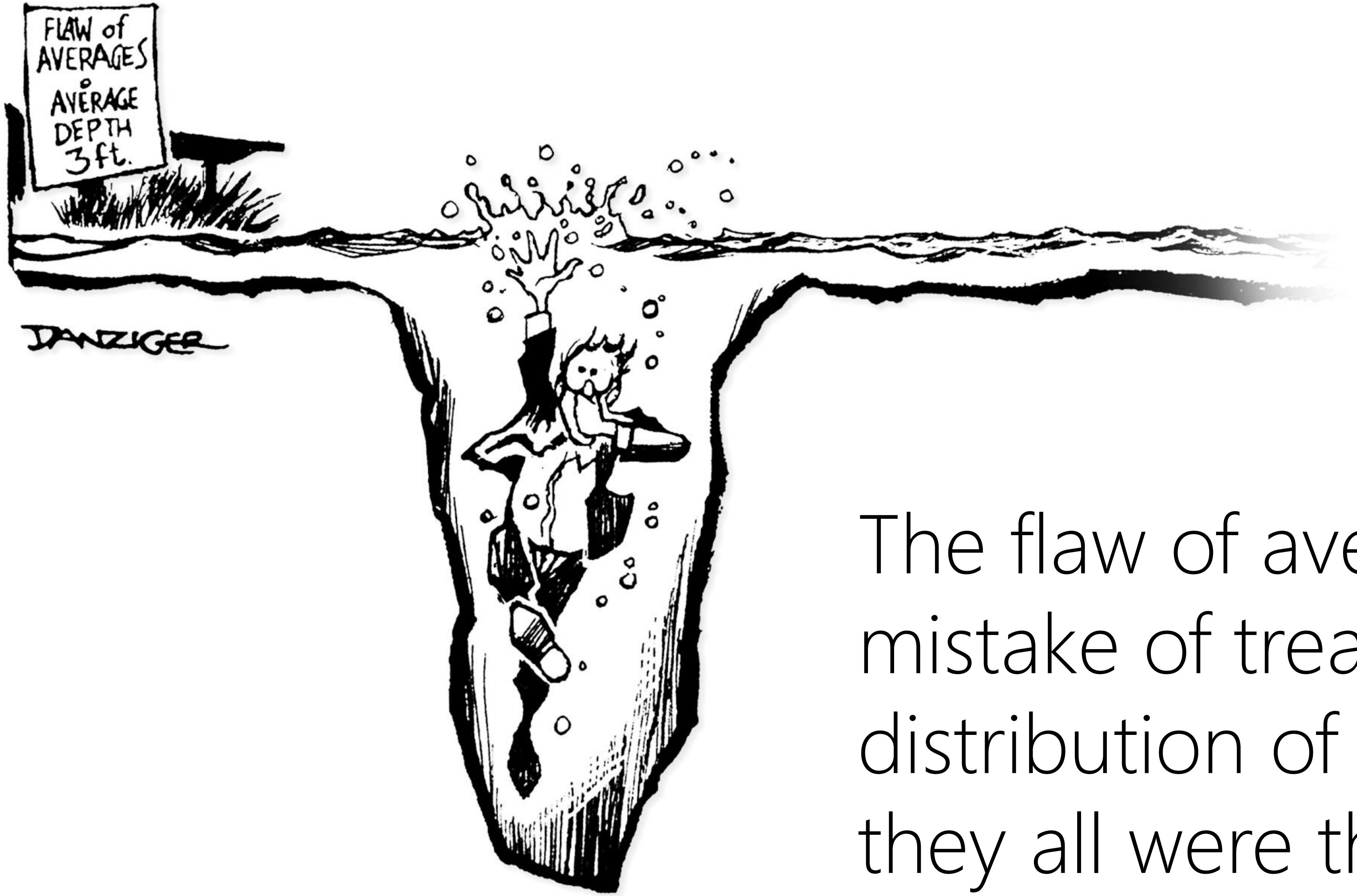
A. Less than 1.5 months
B. 1.5 months
C. More than 1.5 months

# Quiz Answer

**C. More than 1.5 months**

Why? Mathematical answer: Jensen's inequality says so.

Why? Intuitive account: You can't finish *your project* until *all of your subprojects* are finished. Finishing all of them in ≤ 1.5 months is as likely as flipping four coins and having them all come up heads (1 in 16).




The flaw of averages is the mistake of treating a distribution of things as if they all were the average

# Spreadsheets, since 1979

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | =1+RAND() | =1+RAND() | =1+RAND() | =1+RAND() | =MAX(A2:D2) |
| 2 | 1.104 | 1.340 | 1.947 | 1.221 | 1.947 |
| 3 | 1.937 | 1.020 | 1.321 | 1.163 | 1.937 |
| 4 | 1.113 | 1.491 | 1.638 | 1.121 | 1.638 |
| 5 | 1.056 | 1.586 | 1.939 | 1.781 | 1.939 |
| 6 | 1.194 | 1.159 | 1.329 | 1.800 | 1.800 |
| 7 | 1.343 | 1.444 | 1.489 | 1.602 | 1.602 |
| 8 | 1.658 | 1.944 | 1.025 | 1.332 | 1.944 |
| 9 | 1.089 | 1.376 | 1.163 | 1.446 | 1.446 |
| 10 | 1.836 | 1.637 | 1.601 | 1.695 | 1.836 |
| 11 | | | | | =AVERAGE(E2:E10) |
| 12 | | | | | 1.787 |
| 13 | | | | | |
| 14 | =AVERAGE(A2:A10) | =AVERAGE(B2:B10) | =AVERAGE(C2:C10) | =AVERAGE(D2:D10) | =MAX(A15:D15) |
| 15 | 1.370 | 1.444 | 1.495 | 1.462 | 1.495 |

# Spreadsheets, since 1979

## Using RAND() for Monte Carlo simulation is not quite probabilistic programming

## But add-ins do support simple probabilistic programming

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | =1+RAND() | =1+RAND() | =1+RAND() | =1+RAND() | =MAX(A2:D2) |
| 2 | 1.104 | [illegible] | 1.947 | 1.221 | 1.947 |
| 3 | 1.937 | 1.020 | 1.321 | 1.163 | 1.937 |
| 4 | 1.113 | 1.491 | 1.638 | 1.121 | 1.638 |
| 5 | 1.056 | 1.586 | 1.939 | 1.781 | 1.939 |
| 6 | 1.194 | 1.159 | 1.329 | 1.800 | 1.800 |
| 7 | 1.343 | 1.444 | 1.489 | 1.602 | 1.602 |
| 8 | 1.658 | 1.944 | 1.025 | 1.332 | 1.944 |
| 9 | 1.089 | 1.376 | 1.163 | 1.446 | 1.446 |
| 10 | 1.836 | 1.637 | 1.501 | 1.595 | 1.836 |
| 11 | | | | | =AVERAGE(E2:E10) |
| 12 | | | | | 1.787 |
| 13 | | | | | |
| 14 | =AVERAGE(A2:A10) | =AVERAGE(B2:B10) | =AVERAGE(C2:C10) | =AVERAGE(D2:D10) | =MAX(A15:D15) |
| 15 | 1.370 | 1.444 | 1.495 | 1.462 | 1.495 |

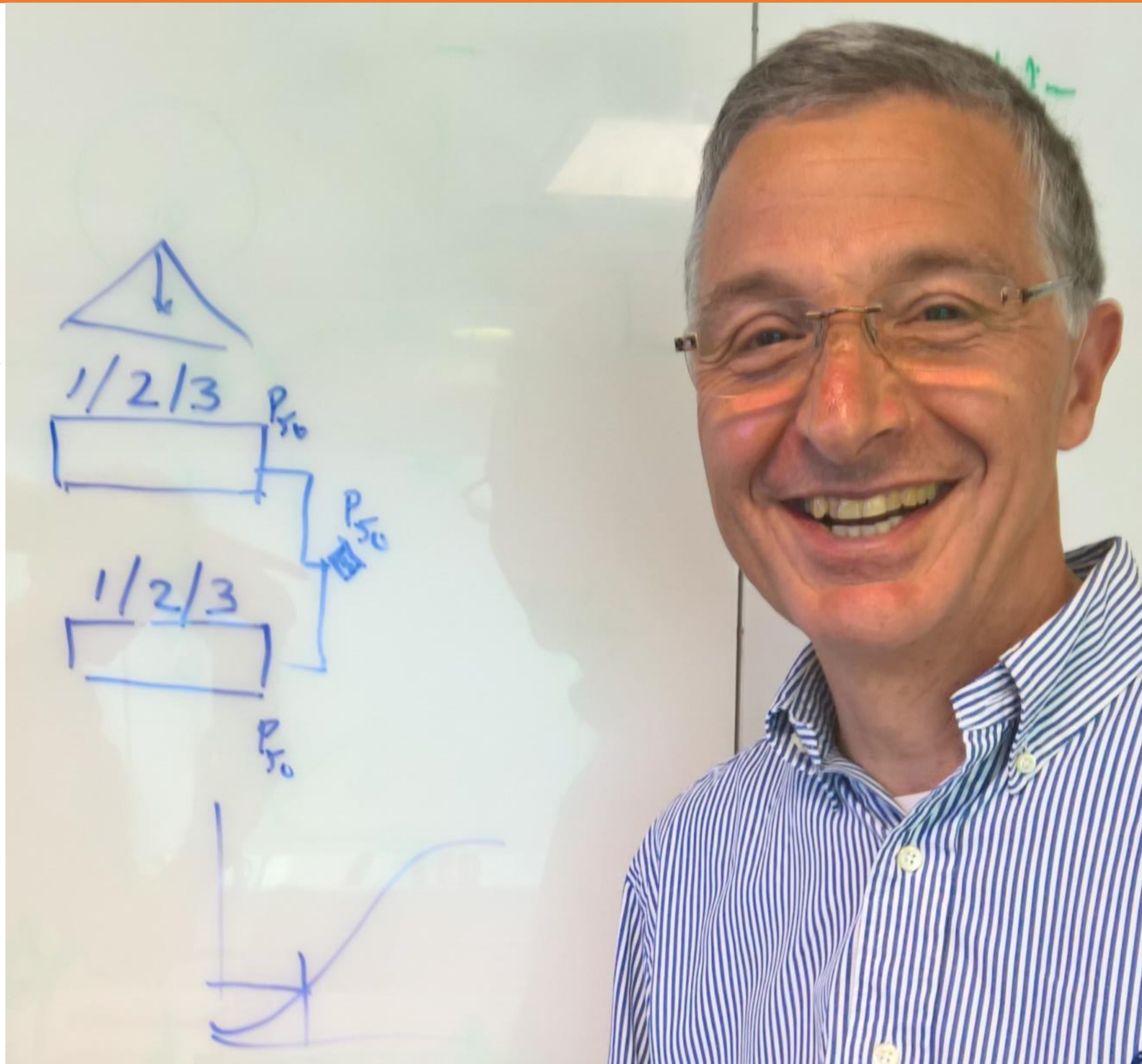
# Insights from @Risk Users

We interviewed Peter Adlington, a consultant who helps companies make risk assessments using @Risk.

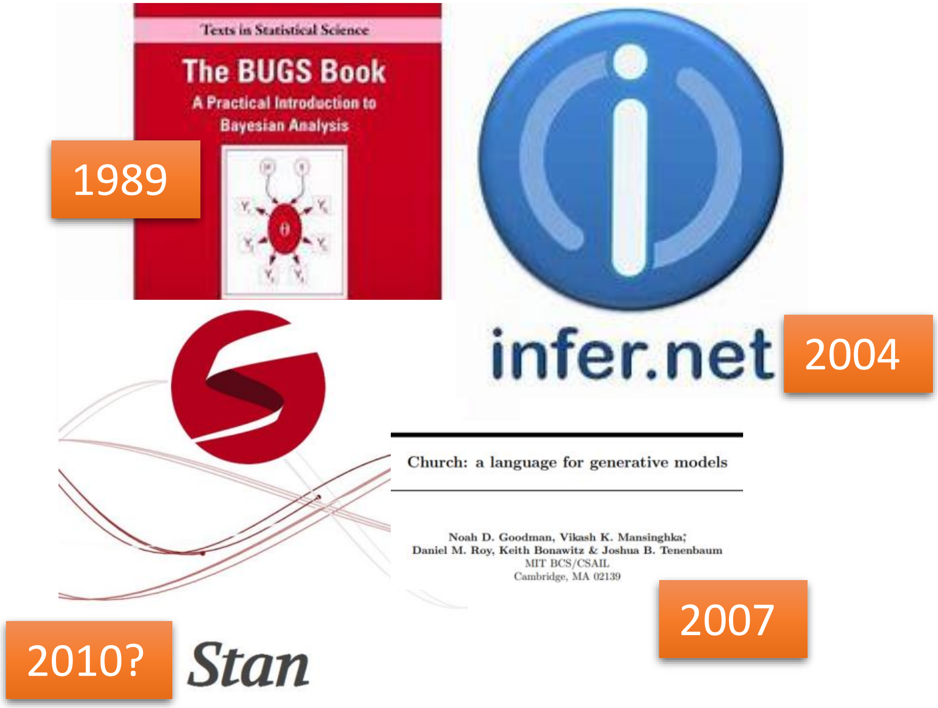
One key idea is the use of triples such as 1/2/3 to specify **triangular** distributions, with 1 being the **minimum**, 2 the **most likely**, and 3 the **maximum value**.

He explained that his customers don't initially appreciate that average inputs don't result in average outputs – the **Flaw of Averages**.

@Risk said to have ½ million users.



# Probabilistic Programming since the 80s


Spreadsheets since 1987 (@RISK launched)

Languages since 1989 (BUGS started)

# Estimated Usage – Highly Uncertain

Programming (writing formulas) in spreadsheets
(tens to hundreds of millions of people)

>>>

Probabilistic Programming (without conditioning) in spreadsheets
(hundreds of thousands of people)

>

Probabilistic Programming in probabilistic programming languages
(tens of thousands of people)

# This Talk

Probabilistic programming aims to help decisions under uncertainty.

We examine **probabilistic programming for end-users**, people like spreadsheet users who write code (formulas) for their own use.

We examine spreadsheet extensions for uncertainty: Streit's **uncertain values**, but also Peyton Jones and others' **sheet-defined functions**.

We examine the types of uncertainty actually encountered by users, and their coping mechanisms, via an interview study.

Hence, we draw implications for spreadsheet design.

# Extension 1: Uncertain Values

Based on Streit’s proposal of uncertain values that propagate through a sheet

Alexander Streit, “Encapsulation and abstraction for modeling and visualizing information uncertainty” (March 2008), Queensland University of Technology

# Clara's Budget

Clara is an end-user confident to model with simple formulas

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | | | | **Budget for Jan 2018** | |
| 2 | | | | | | |
| 3 | | | | | **Income** | 2000 |
| 4 | | | | | | |
| | | | | | **Expenses** | |
| | | | | | Rent | 1100 |
| | | | | | Commute | 85 |
| | | | | | Sofa! | 700 |
| | | | | | Utilities | |
| | | | | | **Total expenses** | 1885 |
| | | | | | | |
| | | | | | **Balance** | 115 |

E1 = "Budget for Jan 2018"
E3 = "Income"; F4 = 2000
E5 = "Expenses"
E6 = "Rent"; F6 = 1100
E7 = "Commute"; F7 = 85
E8 = "Sofa!"; F8 = 700
E9 = "Utilities"; F9 = ???
E10 = "Total expenses"; F10 = SUM(F6:F9)
E12 = "Balance"; F12 = F4−F10

# Streit's Uncertain Values in Cells

1. The user can input uncertainty information into cells.
   - Input can be via a textual notation, such as a numeric range, or as parameters of a probability distribution.
   - Assisted by some UI that constructs an underlying formula.
2. Uncertainty information propagates through formula evaluation.
3. The presence of uncertainty information is indicated within cell.
   - A most likely value might be displayed, for example.

Essentially, a computational effect for uncertainty in spreadsheets

Streit considered estimates, intervals, parametric distributions

# Qualitative

The function ESTIMATE(V) returns the value V tagged as uncertain.

We assume some UI to help user insert the formula ESTIMATE(100)

```
E1 = "Budget for Jan 2018"
E3 = "Income"; F4 = 2000
E5 = "Expenses"
E6 = "Rent"; F6 = 1100
E7 = "Commute"; F7 = 85
E8 = "Sofa!"; F8 = 700
E9 = "Utilities"; F9 = ESTIMATE(100)
E10 = "Total expenses"; F10 = SUM(F6:F9)  // value: ESTIMATE(1985)
E12 = "Balance"; F12 = F4−F10  // value: ESTIMATE(15)
```

# Possibilistic (sets of values)

The function SCENARIOS(V1,…,VN) returns an uncertain value representing N scenarios.

```
E1 = "Budget for Jan 2018"
E3 = "Income"; F4 = 2000
E5 = "Expenses"
E6 = "Rent"; F6 = 1100
E7 = "Commute"; F7 = 85
E8 = "Sofa!"; F8 = 700
E9 = "Utilities"; F9 = SCENARIOS(50,100,150)
E10 = "Total expenses"; F10 = SUM(F6:F9)  // value: SCENARIOS(1935,1985,2035)
E12 = "Balance"; F12 = F4−F10  // value: SCENARIOS(65,15,−35)
```

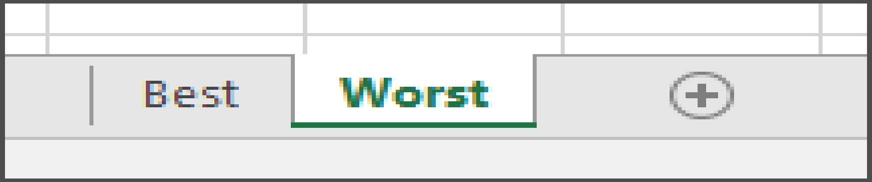
# Possibilistic (sets of values)

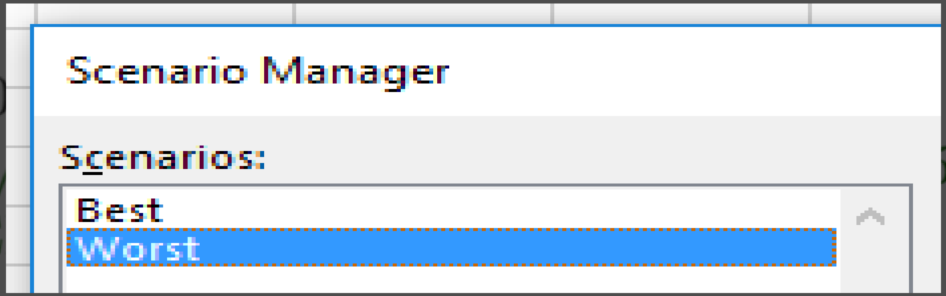
The function SCENARIOS(V1,...,VN) re
representing N scenarios.

## Alternative to existing techniques:

E1 = "Budget for Jan 2018"
E3 = "Income"; F4 = 2000
E5 = "Expenses"
E6 = "Rent"; F6 = 1100
E7 = "Commute"; F7 = 85
E8 = "Sofa!"; F8 = 700
E9 = "Utilities"; F9 = SCENARIOS(50,100,150
E10 = "Total expenses"; F10 = SUM(F6:F9) // 5)
E12 = "Balance"; F12 = F4−F10 // value: SCE

| | A | B | C |
|---|---|---|---|
| 1 | | *Best* | *Worst* |
| 2 | **Widgets** | 75000 | 125000 |
| 3 | **Cost** | $6.50 | $7.50 |
| 4 | **Total** | $487,500.00 | $937,500.00 |
| 5 | **Budget** | $750,000.00 | $750,000.00 |
| 6 | **Over?** | FALSE | TRUE |



# Probabilistic (weighted sets of values)

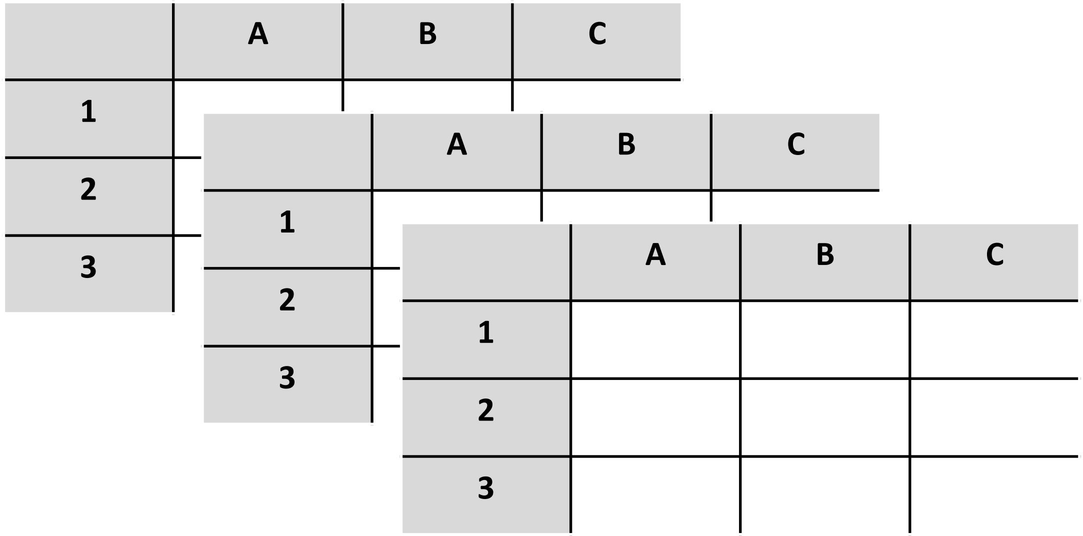
The function DIST.TRIANG(*a*,*b*,*c*) constructs a value distributed according to a **triangular distribution**, with lower bound *a*, upper bound *b*, and mode *c*, where $a < b$ and $a \leq c \leq b$.

```
E1 = "Budget for Jan 2018"
E3 = "Income"; F4 = 2000
E5 = "Expenses"
E6 = "Rent"; F6 = 1100
E7 = "Commute"; F7 = 85
E8 = "Sofa!"; F8 = 700
E9 = "Utilities"; F9 = DIST.TRIANG(50,150,100)
E10 = "Total expenses"; F10 = SUM(F6:F9)  // value: DIST.TRIANG(1935,2035,1985)
E12 = "Balance"; F12 = F4−F10  // value: DIST.TRIANG(−35,65,15)
E13="Chance of overdraft"; F13=F12<0 // value: 25%
```

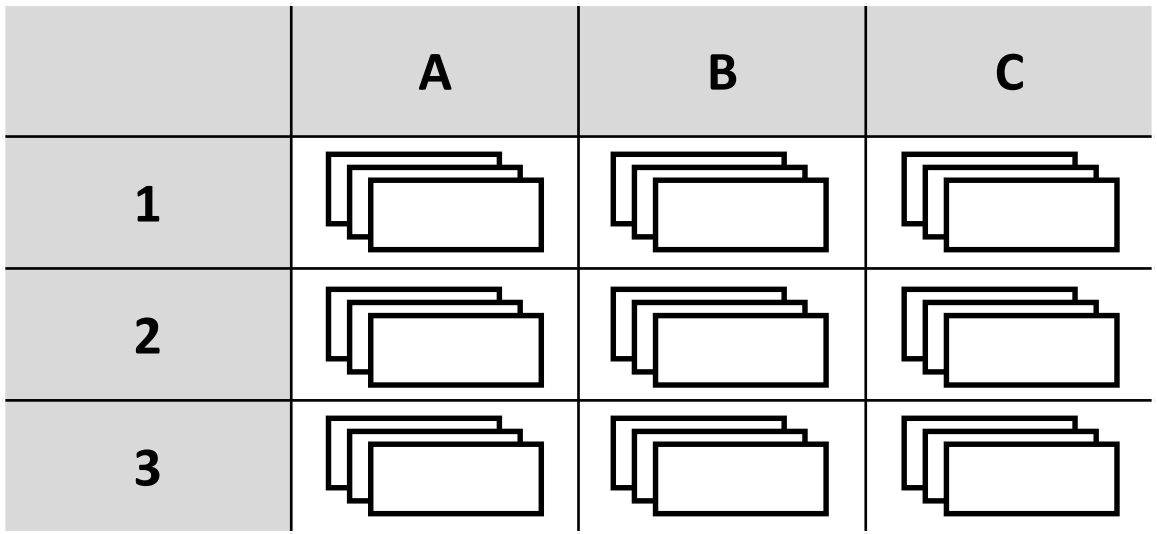
# Idea: Store Array of Scenarios in Each Cell

Means the same as mashing F9 on a spreadsheet with random values…

Acts like this




Stored like this




…except each sequence of "random" numbers is repeatable

# How Many Valentine Cards to Order?

| Year | Sold | | | Cost per card | Price per card | |
|---|---|---|---|---|---|---|
| 2010 | 833924 | | | $1.47 | $1.99 | |
| 2011 | 1299434 | | | | | |
| 2012 | 760148 | | **Ordered** | **Cost** | **Gross Profit** | **Net Profit** |
| 2013 | 1083010 | | 400000 | $588,000.00 | $796,000.00 | $208,000.00 |
| 2014 | 923113 | | 550000 | $808,500.00 | $1,094,500.00 | $286,000.00 |
| 2015 | 1216995 | | 700000 | $1,029,000.00 | $1,393,000.00 | $364,000.00 |
| 2016 | 861598 | | 850000 | $1,249,500.00 | $1,691,500.00 | $442,000.00 |
| | | | 1000000 | $1,470,000.00 | $1,983,808.83 | $513,808.83 |
| **Average** | 996889 | | 1150000 | $1,690,500.00 | $1,983,808.83 | $293,308.83 |
| | | | 1300000 | $1,911,000.00 | $1,983,808.83 | $72,808.83 |

Today, we'd model 2017 demand by exactly the average of previous years' demand.

# How Many Valentine Cards to Order?

| | | | | | | |
|---|---|---|---|---|---|---|
| **Year** | **Sold** | | **Cost per card** | **Price per card** | | |
| 2010 | 833924 | | $1.47 | $1.99 | | |
| 2011 | 1299434 | | | | | |
| 2012 | 760148 | **Ordered** | **Cost** | **Gross Profit** | **Net Profit** | |
| 2013 | 1083010 | 400000 | $588,000.00 | ~$795,792.39 | ~$207,792.39 | |
| 2014 | 923113 | 550000 | $808,500.00 | ~$1,092,363.04 | ~$283,863.04 | |
| 2015 | 1216995 | 700000 | $1,029,000.00 | ~$1,379,405.37 | ~$350,405.37 | |
| 2016 | 861598 | 850000 | $1,249,500.00 | ~$1,634,464.79 | ~$384,964.79 | After |
| | | 1000000 | $1,470,000.00 | ~$1,823,660.99 | ~$353,660.99 | Before |
| **Average** | 996889 | 1150000 | $1,690,500.00 | ~$1,929,667.49 | ~$239,167.49 | |
| **Stdev** | 205648 | 1300000 | $1,911,000.00 | ~$1,971,125.46 | ~$60,125.46 | |
| **Demand** | ~996899 | =DIST.NORM(N,B12,B13) | | | | |

Expected net profits are lower when we model distributions.
Surprisingly, the optimal decision is to order below the average demand.

# Visualizing Uncertainty (Streit 2008)

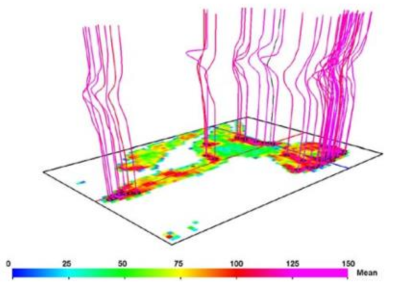
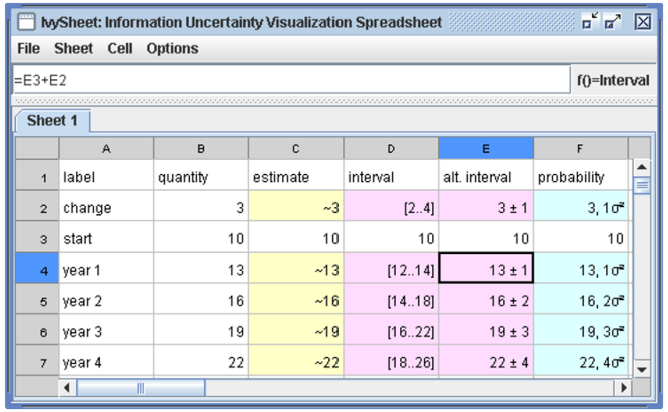



Figure 2.17: A visualization that draws the probability density function over associated data points




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | label | quantity | estimate | interval | alt. interval | probability |
| 2 | change | 3 | ~3 | [2..4] | 3 ± 1 | $3, 1\sigma^2$ |
| 3 | start | 10 | 10 | 10 | 10 | 10 |
| 4 | year 1 | 13 | ~13 | [12..14] | 13 ± 1 | $13, 1\sigma^2$ |
| 5 | year 2 | 16 | ~16 | [14..18] | 16 ± 2 | $16, 2\sigma^2$ |
| 6 | year 3 | 19 | ~19 | [16..22] | 19 ± 3 | $19, 3\sigma^2$ |
| 7 | year 4 | 22 | ~22 | [18..26] | 22 ± 4 | $22, 4\sigma^2$ |

Figure 3.4: Screenshot of the Prototype System

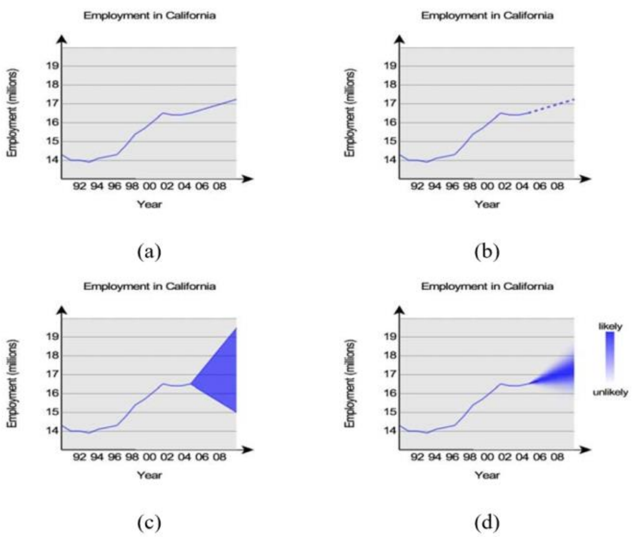



Figure 3.1: Visualizations of Employment Numbers in California. Years 2005-2010 are predicted. (a) Assuming Average Growth (b) Indicating Growth is Estimated (c) Possible Growth (d) Likely Growth. (Data Source: California Employment Development Department)

| salary growth | 7% | | |
|---|---|---|---|
| year | salary | tax | net income |
| 1 | 60000 | 10200 | 49800 |
| 2 | 64200 | 10914 | 53286 |
| 3 | 68694 | 11677.98 | 57016.02 |
| 4 | 73502.58 | 12495.44 | 61007.14 |

(a)

| salary rate lo | 6% | salary rate hi | 8% | | | |
|---|---|---|---|---|---|---|
| year | salary (lo) | salary (hi) | tax (lo) | tax (hi) | net (lo) | net (hi) |
| 1 | 60000 | 60000 | 10200 | 10200 | 49800 | 49800 |
| 2 | 63600 | 64800 | 10812 | 11016 | 52788 | 53784 |
| 3 | 67416 | 69984 | 11460.72 | 11897.28 | 55955.28 | 58086.72 |
| 4 | 71460.96 | 75582.72 | 12148.36 | 12849.06 | 59312.60 | 62733.66 |

(b)

| salary growth | 7% | | |
|---|---|---|---|
| year | salary | tax | net income |
| 1 | 60000 | 10200 | 49800 |
| 2 | 64200 | 10914 | 53286 |
| 3 | 68700 | 11679 | 57021 |
| 4 | 73521.84 | 12498.71 | 61023.13 |

(c)

| salary growth | [6, 8]% | | |
|---|---|---|---|
| year | salary | tax | net income |
| 1 | 60000 | 10200 | 49800 |
| 2 | [63600, 64800] | [10812, 11016] | [52584, 53988] |
| 3 | [67416, 69984] | [11460.72, 11897.28] | [55518.72, 58523.28] |
| 4 | [71460.96, 75582.72] | [12148.36, 12849.06] | [58611.90, 63434.36] |

(d)

Figure 4.6: Interval Modeling Example: (a) Original Model (b) Traditional Spreadsheet (c) Prototype System Uncertainty Hidden (d) Prototype System Uncertainty Shown

| Type | Format | Example |
|---|---|---|
| Formula | = *Expression* | =A1+6 |
| Label | '*String* | '12.06 |
| Quantity | *Number* | 12.06 |
| KnownQty | *Number*# | 12.06# |
| Estimate | ~*Number* | ~12.06 |
| Interval | *Number* +- *Number* | 12 +- 5.0e-10 |
| | [ *Number* , *Number* ] | [1, 2.1] |
| | [ *Number* .. *Number* ] | [-15..15] |
| Gaussian | *Number* @ *Number* | 10.1 @ 2.178 |

Table 4.1: Format of Cells in the Prototype System

An important topic but beyond our scope

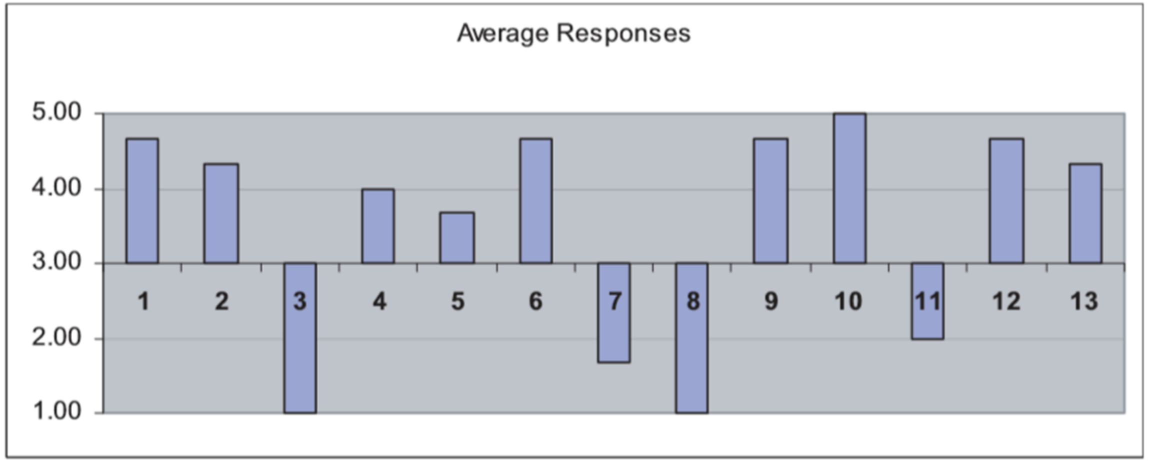
# Streit's User Study

## A Spreadsheet Approach to Facilitate Visualization of Uncertainty in Information

Alexander Streit, Binh Pham, Ross Brown

Source: IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS 2007

Positive evaluation from rather small (N=3) user study – first study we are aware of. Three financial experts volunteered to complete tasks with the system.




1. Overall, I found the system to be intuitive
2. Placing an interval in a cell made sense to me
3. I would need more training to complete the tasks
4. I found it easy to follow the effect that an interval had on the rest of the spreadsheet
5. The graph helped me to understand the effects better
6. I would definitely use the graph if there were many numbers
7. The spreadsheet became too cluttered or hard to read
8. This tool makes it easier to make mistakes
9. I would use a tool like this if it were available to me
10. I would use intervals in Microsoft Excel if they were supported
11. My current tools are adequate for modeling and visualizing uncertainty
12. I would use normal probability distributions more often in my work if I had a tool like this
13. I would model uncertainty more often if I had these features available

# Two Potential Concerns

**Participant Response Bias**: "We find that respondents are about 2.5x more likely to prefer a technological artifact they believe to be developed by the interviewer, even when the alternative is identical."

Session: The Tools of the Trade
CHI 2012, May 5–10, 2012, Austin, Texas, USA

**"Yours is Better!" Participant Response Bias in HCI**

Nicola Dell† Vidya Vaidyanathan‡ Indrani Medhi§ Edward Cutrell§ William Thies§

†University of Washington
nixdell@cs.washington.edu

‡San Jose State University
vidya.vn@gmail.com

§Microsoft Research India
{indranim,cutrell,thies}@microsoft.com

**Showing just one design can be misleading**: "We found that when presented with a single design, users give significantly higher ratings and were more reluctant to criticize than when presented with the same design in a group of three"

CHI 2006 Proceedings • Usability Methods
April 22-27, 2006 • Montréal, Québec, Canada

**Getting the Right Design and the Design Right: Testing Many Is Better Than One**

Maryam Tohidi
University of Toronto
Toronto, Canada
mtohidi@dgp.toronto.edu

William Buxton
Microsoft Research
Toronto, Canada
bill@billbuxton.com

Ronald Baecker
University of Toronto
Toronto, Canada
rmb@kmdi.utoronto.ca

Abigail Sellen
Microsoft Research
Cambridge, UK
asellen@microsoft.com

# Extension 2:
# Sheet-Defined Functions

Some benefits of uncertain values obtained from a general-purpose extension:

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

# Background: Sheet-Defined Functions

A sheet-defined function *f* is specified by a **body**, a piece of grid,
with **inputs** identified by ranges in the body,
and an **output**, also identified by a range in the body.

A function call *f*(V1,…,Vn) is computed by:

- making a copy of the body,
- pasting the values V1,…,Vn into the input ranges,
- evaluating the body,
- and then returning the value from the result range.

# Clara’s budget as an SDF

```
function Budget( F9 ) returns F12:G12 {
  E1 = "Budget for Jan 2018"
  E3 = "Income"; F4 = 2000
  E5 = "Expenses"
  E6 = "Rent"; F6 = 1100  // A certain cost
  E7 = "Commute"; F7 = 85 // Another certain cost
  E8 = "Sofa!"; F8 = 700  // Clara is deciding whether to buy this item
  E9 = "Utilities"; F9 = 100  // This cell is the parameter to the SDF
  E10 = "Total expenses"; F10 = SUM(F6:F9)
  E12 = "Balance"; F12 = F4−F10; G12 = IF(F12<0,1,0)
}
```

# Possibilistic Model (Multiple Scenarios)

$$\text{B54}{=}50;\ \ \text{C54} = \text{BUDGET}(\text{B54})$$
$$\text{B55}{=}100;\ \text{C55} = \text{BUDGET}(\text{B55})$$
$$\text{B56}{=}150;\ \text{C56} = \text{BUDGET}(\text{B56})$$

| Utilities | Balance | Overdrawn |
| --- | --- | --- |
| 50 | 65 | 0 |
| 100 | 15 | 0 |
| 150 | -35 | 1 |

# Probabilistic Model (SIPs)

$$B58 = \text{SIP.TRIANGULAR}(B54, B56, B55, 1000)$$
$$C58 = \text{VMAP}(B58, \text{Budget})$$

| Utilities | Balance | Overdrawn |
|---|---|---|
| 99.6382583 | 17.856223 | 0 |
| 99.3063957 | 14.738061 | 0 |
| 91.4256927 | -7.472543 | 1 |
| [illegible] | [illegible] | 1 |
| 132.734228 | -21.861 | 1 |
| 99.4700682 | 10.093057 | 0 |
| [illegible] | [illegible] | 0 |
| 94.6155694 | 13.13844 | 0 |

| Balance | |
|---|---|
| N | 1000 |
| min | -32.3987 |
| mean | 14.946 |
| median | [illegible] |
| max | 63.12915 |

| Overdrawn | |
|---|---|
| N | 1000 |
| min | 0 |
| mean | 0.26 |
| median | 0 |
| max | 1 |

Like using RAND() for Monte Carlo simulation, this is not quite probabilistic programming

It's a new idiom for SDFs to help users avoid flaw of averages

# Interview study

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

Borghouts, J., Gordon, A.D., Sarkar, A., O'Hara, K.P., Toronto, N.: **Somewhere around that number: An interview study of how spreadsheet users manage uncertainty.** arXiv preprint arXiv:1905.13072 (2019)

# Research Questions and Method

What types of uncertainty do spreadsheet users deal with?

How do they manage these various types of uncertainty?

We conducted semi-structured interviews of 11 participants, who walked us through their spreadsheets.

We analysed the audio transcripts of these interviews, identifying six qualitatively different **types of uncertainty**, and six **categories of strategies** used to cope with the uncertainty.

(Thematic analysis is not quantitative.)

# Participants

We interviewed 11 participants who worked in **finance, construction, IT consulting, the oil and gas industry, business administration, and academic research**.

The size of participants' spreadsheets ranged from 40 rows to thousands of rows

Participants were recruited using convenience sampling via email and were eligible to participate if they used spreadsheets that contained "uncertain data".

We did not filter participants based on whether they used spreadsheets for work or personal use, but all participants we recruited dealt with uncertainty in spreadsheets for work purposes.

To elicit participation, the invitation gave several examples of spreadsheet tasks which could involve uncertainty, such as **budgeting, planning, business forecasting, data collection and analysis, scientific modelling, and making predictions**.

Interviews lasted on average 60 minutes, and participants were reimbursed with a £30 voucher for an online store.

# Procedure

We asked participants to bring one or more of their own spreadsheets which contained uncertain data to the interview session.
They were instructed to remove any sensitive information.

1. In the **first part of the session**, participants were asked to talk about their work, and how uncertainty and spreadsheets are a part of this work.
2. In the **second part of the session**, we discussed if participants gain insights from uncertainty, what tools or strategies they use to gain this insight, and what challenges they perceive in doing so.
3. In the **final part**, we asked participants to walk us through their spreadsheets, and explain how these spreadsheets were constructed, and what they were used for.

The session was audio recorded, and participants' walk-through of their spreadsheets was screen recorded.

# Data Analysis

The audio recordings were transcribed verbatim and analysed using iterative coding based on an inductive approach of **thematic analysis**.

There was no pre-existing **coding scheme**, but we did approach the data with a specific focus to uncover uncertainty types, and user strategies for managing uncertainty.

Through a detailed analysis of the transcripts, we identified key features of the spreadsheets and work practices that related to participant concerns with uncertainty.

# Spreadsheet Roles

**Database:** *'The vast bulk of users use it as a way of storing, analysing, capturing data. (...) it becomes the record of the corporate memory'* (P1).

**Template to repeat analysis:** *'Even sometimes when you use past examples, you still think: There's still something not quite right about that figure'* (P4).

**Calculator and analysis tool:** P2 and P5 used spreadsheets to build prediction models with formulas.

**Notepad:** P4 used spreadsheets to get a better understanding of how long a project would approximately take.

| Element | Code | Nature of Work | Time Days | Dependencies | Confidence % |
|---|---|---|---|---|---|
| **Main Site** | MS::C | Content | 2 | None | 80 |
| **Vocabulary Builder** | C::VB::C | Content | 3 | None | 10 |

**Data exploration tool:** P10 and P11 dealt with large datasets and used statistical analysis tools to analyse these.

# Findings: Types of Uncertain Data

| | Definition | Example |
| --- | --- | --- |
| Estimate | Data approximated by the user of which the precise value is not known | expected number of attendees to an event |

“We’re talking about the future. We don’t know exactly what’s going to happen. All we can do is make best **estimates**” (P8)

| | | |
| --- | --- | --- |
| Unfindable | Data actually contained in the spreadsheet that cannot easily be found by the user | total amount of hours worked on the weekend in an employee timesheet |
| Untraceable | Data whose original source, or how it is calculated by formulas, cannot be traced | subjective estimates made by other people, or complex and inaccessible formulas |

# Findings: Categories of Coping Strategy

| | Definition | Example |
|---|---|---|
| Add | The user adds additional uncertainty to a spreadsheet | by using an approximated value, such as an estimate or an interval or an average |
| Communicate | The user communicates uncertainty | to others verbally, through reports and presentations or to themselves by highlighting cells |
| Exploit | The user uses the amount of uncertainty in a spreadsheet as a valuable piece of information | by adding weights to missing values during model building, based on how often they were missing |
| Ignore | The user ignores uncertainty | by removing it from the spreadsheet or replacing it with other values |
| Minimise | The user minimises uncertainty | by acquiring more data, or discussing it with colleagues |
| Understand | The user tries to understand uncertainty | by reading literature, discussing it with colleagues, plotting data, evaluating the data source, and analysing a subset of possible scenarios |

# Implications for design

How well do our three formalisms deal with the six types?

# Formalisms versus Types of Uncertainty

| | Dynamic data | Estimates | Errors | Missing data | Unfindable data | Untraceable data |
|---|---|---|---|---|---|---|
| | 15% | 50% | 13% | 12% | 7% | 3% |
| **Qualitative** | Some | Good* | Good* | Weak | None | None |
| **Possibilistic** | Some* | Good* | None | Some | None | None |
| **Probabilistic** | Good* | Weak | Some* | Some* | None | None |

How well do we think a formalism would support a type of uncertainty?
Scale: None, Weak, Some, Good

Percen

An ast
that ty

Estimates like "somewhere around 9" are subjective and our users had little knowledge how it might be distributed, nor indeed how a probabilistic model could be interpreted

# Most Promising Combinations

Qualitat
Possibili
Probabil
Possibili … in.
Probabil … ata.
Probabil … ta and

**Probabilistic:Errors:** `At every stage, a lot of the time we have: if we know we've got however many unknown values in spreadsheet 1, by the time we get to spreadsheet 10, there's a debate about whether the data's of any use at all, because the errors propagate, and multiply. Now I don't have a measure of how that affects it. I'd love to be able to measure that, but I have no idea how to do that within the tools.' (P11)

# Recommendations

No single formalism can solve most user problems

- Qualitative tagging is most applicable, but has low value
- We recommend offering features from multiple formalisms

If we were to pick one to exclude, it would be Possibilistic

- Spreadsheets already have effective coping strategies

Our extensions for representing uncertainty unlikely to help with unfindable and untraceable data

# Reflections

1. There is already more probabilistic programming in spreadsheets than in probabilistic programming languages.
2. **But,** few instances of uncertainty encountered by end users in our study could easily be dealt with by probabilistic modelling.
3. Serendipity: the new proposal of SDFs as a general tool for possibilistic and probabilistic uncertainty.
4. Are end-users capable of understanding Bayesian update? Invrea?
5. Thematic analysis of human experience with PL messier theory. Complementary, but probably better guide to what's useful!