How Do We Solve the World's Spreadsheet Problem?

Alex Rasmussen

@alexras

Hi, I'm Alex!

@alexras

alexras.info

freenome

freenome.com

BITS ON DISK

bitsondisk.com

My Background

2009-2013: really fast sorting

2013-2016: data wrangling

2017-2018: cancer-fighting robots

I think/worry a lot about spreadsheets.

Today's focus:

spreadsheet data

(for compute, felienne.com)

This talk:

1. Spreadsheets are great
2. Spreadsheets are a problem
3. How we can fix it

This talk:

1. Spreadsheets are great
2. Spreadsheets are a problem
3. How we can fix it

What's so **great** about spreadsheets?

Spreadsheets are **Ubiquitous**

1.2 billion Office users (~16% of humans)

1.2 billion Office users (~16% of humans)

60 million Office 365 customers

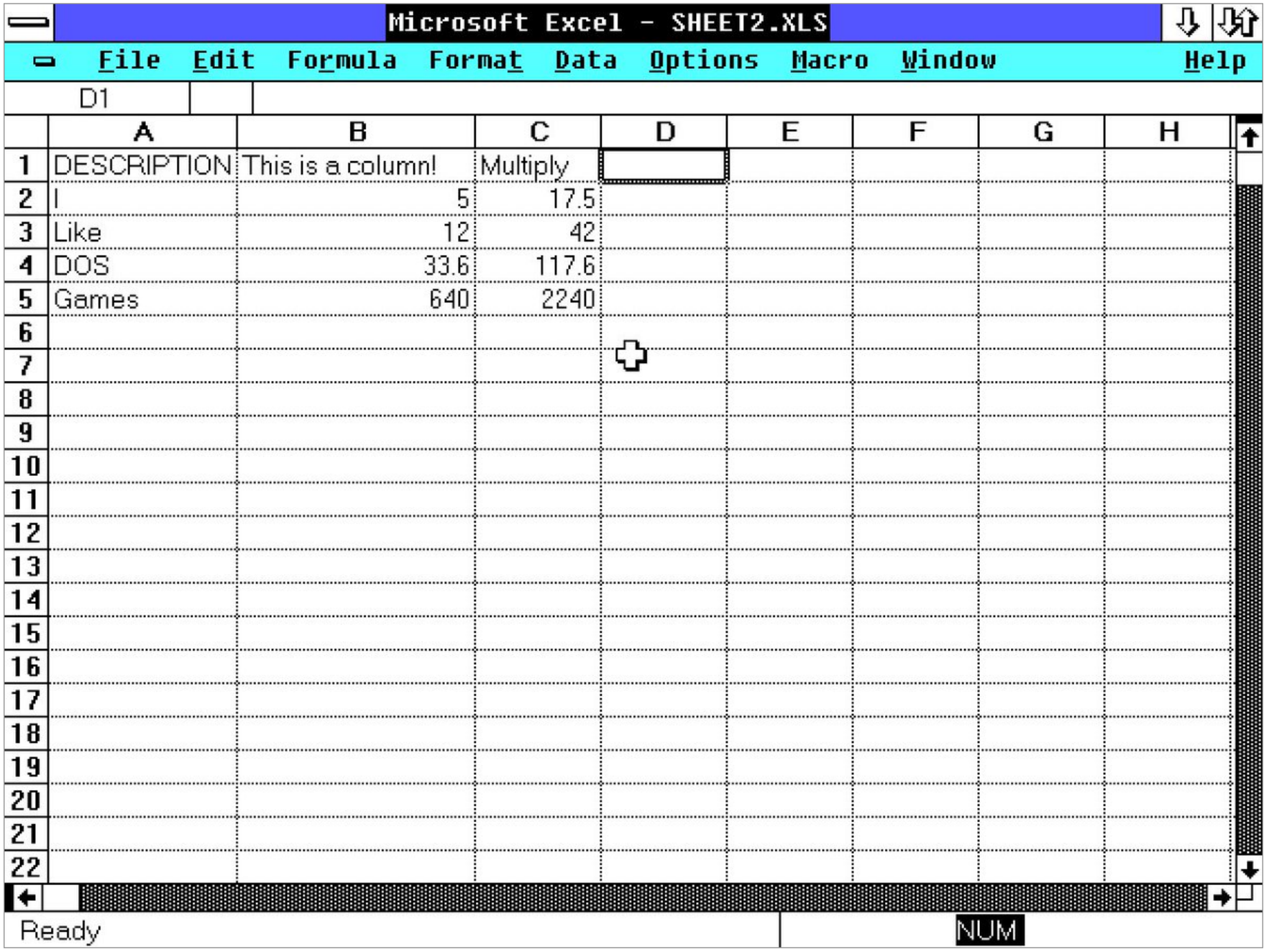
1.2 billion Office users (~16% of humans)

60 million Office 365 customers

>5 million businesses use Google Apps

Spreadsheets are

Approachable


Microsoft Excel - SHEET2.XLS
File Edit Formula Format Data Options Macro Window Help
D1
A B C D E F G H
1 DESCRIPTION This is a column! Multiply
2 I 5 17.5
3 Like 12 42
4 DOS 33.6 117.6
5 Games 640 2240
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Ready
NUM

Spreadsheets are

Flexible

Data grids

Data grids

Graphs

Data grids

Graphs

Anything tabular

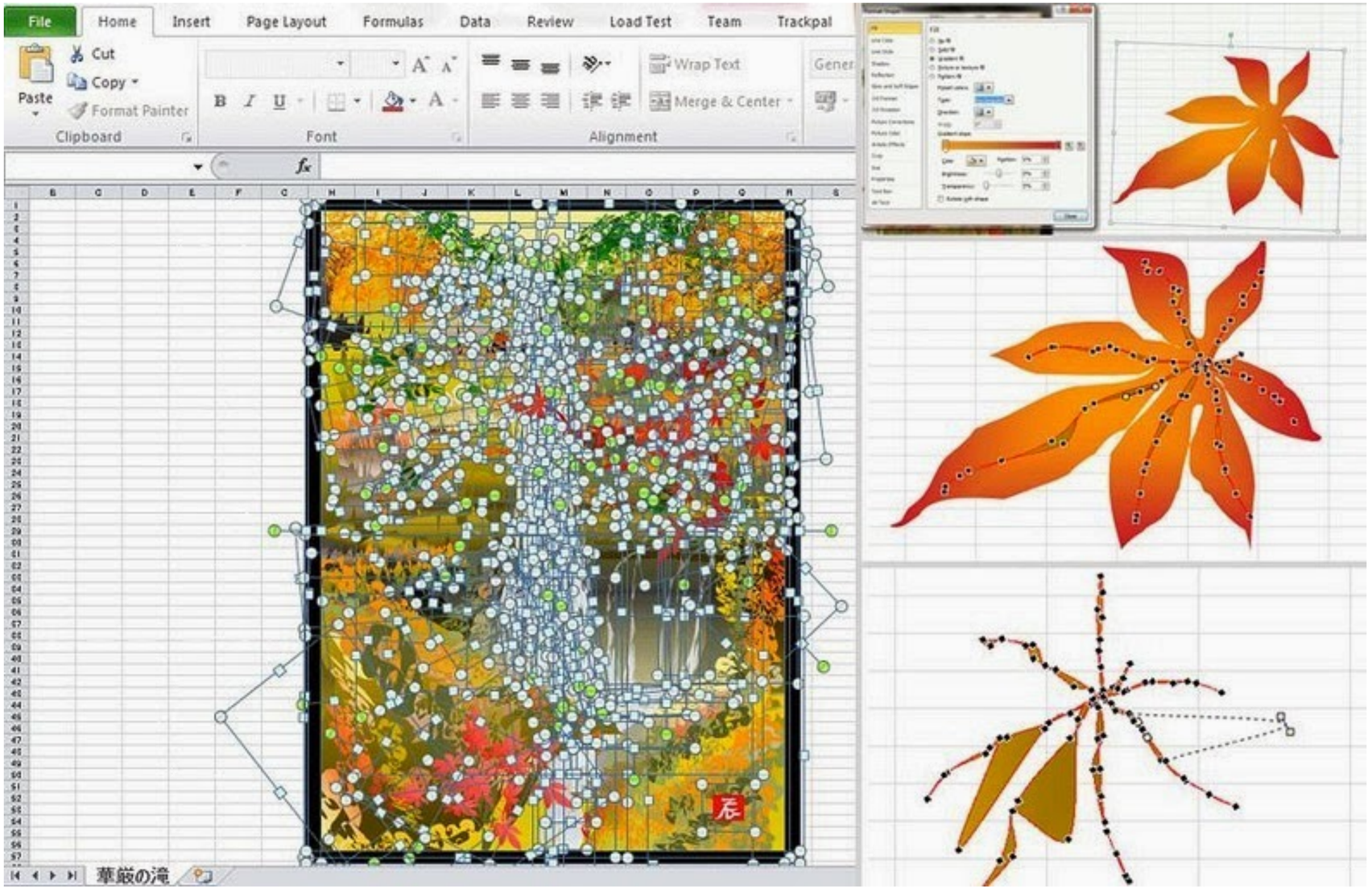
Data grids

Graphs

Anything tabular

Full-scale “apps”

Tatsuo Horiuchi (b. 1940)
Kegon Falls, 2007
AutoShape on canvas



https://pasokonga.com/

So if spreadsheets are
ubiquitous, **approachable**,
and **flexible**,
what's the problem?

This talk:

1. Spreadsheets are great
2. Spreadsheets are a problem
3. How we can fix it

Problem #1:

Data Types

GSV Arson Kite
@Phylan

ME: *makes typo while entering a number*

EXCEL: WAS THAT A DATE

ME: no I meant t-

EXCEL: THAT WAS A 

DATE

ME: it doesn't even make sen-

EXCEL: MAY 12TH 1382. LOOK I EVEN FORMATTED IT. IT IS THIS FOREVER

9:00 AM - Feb 1, 2018

64.8K 17.5K people are talking about this

Automatic type conversion can cause **serious problems**.

DEC1

DEC1

12/1

RIKEN Identifier 2310009E13

RIKEN Identifier

2310009E13

2.31E+13

"We confirmed gene name errors in 987 supplementary files from 704 published articles (**19.6% of all articles**)."

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7

False Equivalence

```
  000123
= 00123
= 123
```

True if they’re integers,
but **what if they’re strings**?

Enumerated Types

Enumerated Types

“Prostate Cancer”

Enumerated Types

“Prostate Cancer”

“prostate cancer”

Enumerated Types

“Prostate Cancer”

“prostate cancer”

“prostatecancer”

Enumerated Types

“Prostate Cancer”

“prostate cancer”

“prostatecancer”

“PC”

Enumerated Types

“Prostate Cancer”

“prostate cancer”

“prostatecancer”

“PC”

“prostate”

Enumerated Types

“Prostate Cancer”

“prostate cancer”

“prostatecancer”

“PC”

“prostate”

“prostrate”

List Validations? Sheet Protection?

😊 Easy to add

😟 Easy to remove by accident

😡 Hard to enforce

Data loss!

False equivalence!

Ontological chaos!

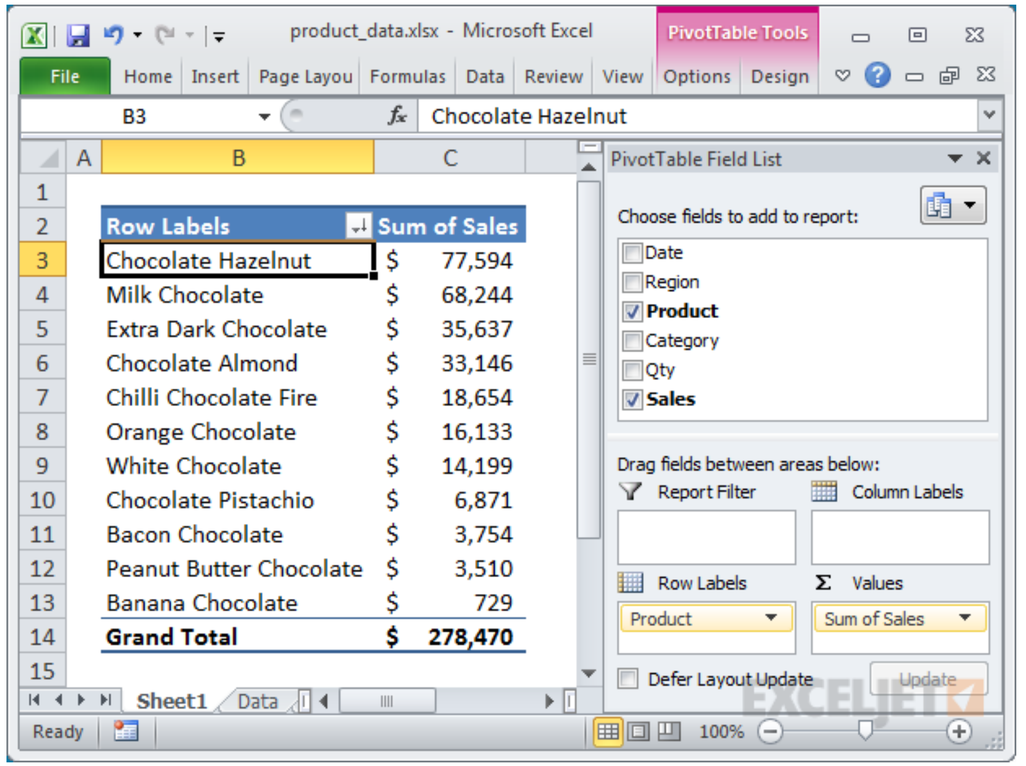
Mass hysteria!

Problem #2:

Queryability

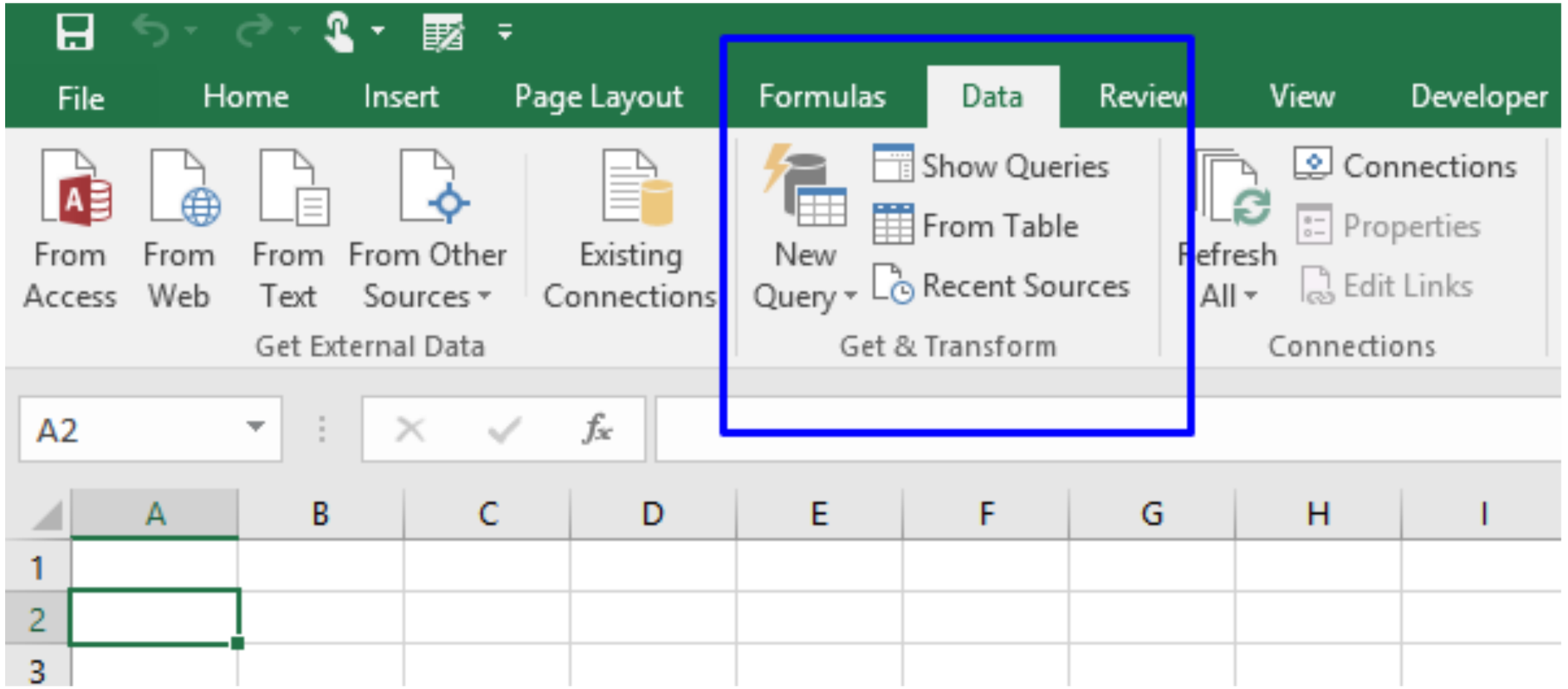
Inside a spreadsheet, things are pretty good!

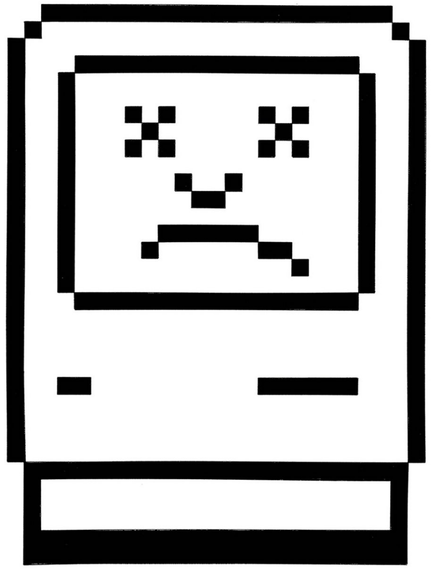
Formulas!
Pivot Tables!
Filters!



What about querying **across** spreadsheets?

Get and Transform





No Mac support.

Structure changes?
Type changes?
Column Renames?
Have fun re-loading.

And what about **joins**?

There's VLOOKUP

```
=VLOOKUP("Product 1",
Prices!$A$2:$B$9,2,FALSE)
```

... but, like, eww.

Data **inside** a spreadsheet is hard to connect to data **outside** that spreadsheet.

Summary:
Spreadsheets are
bad at types and
hard to query

This talk:

1. Spreadsheets are great
2. Spreadsheets are a problem
3. How we can fix it

What about **databases**?

Databases are **great** in ways that spreadsheets **aren’t**.

Databases are great at data type **definition** and **enforcement**.

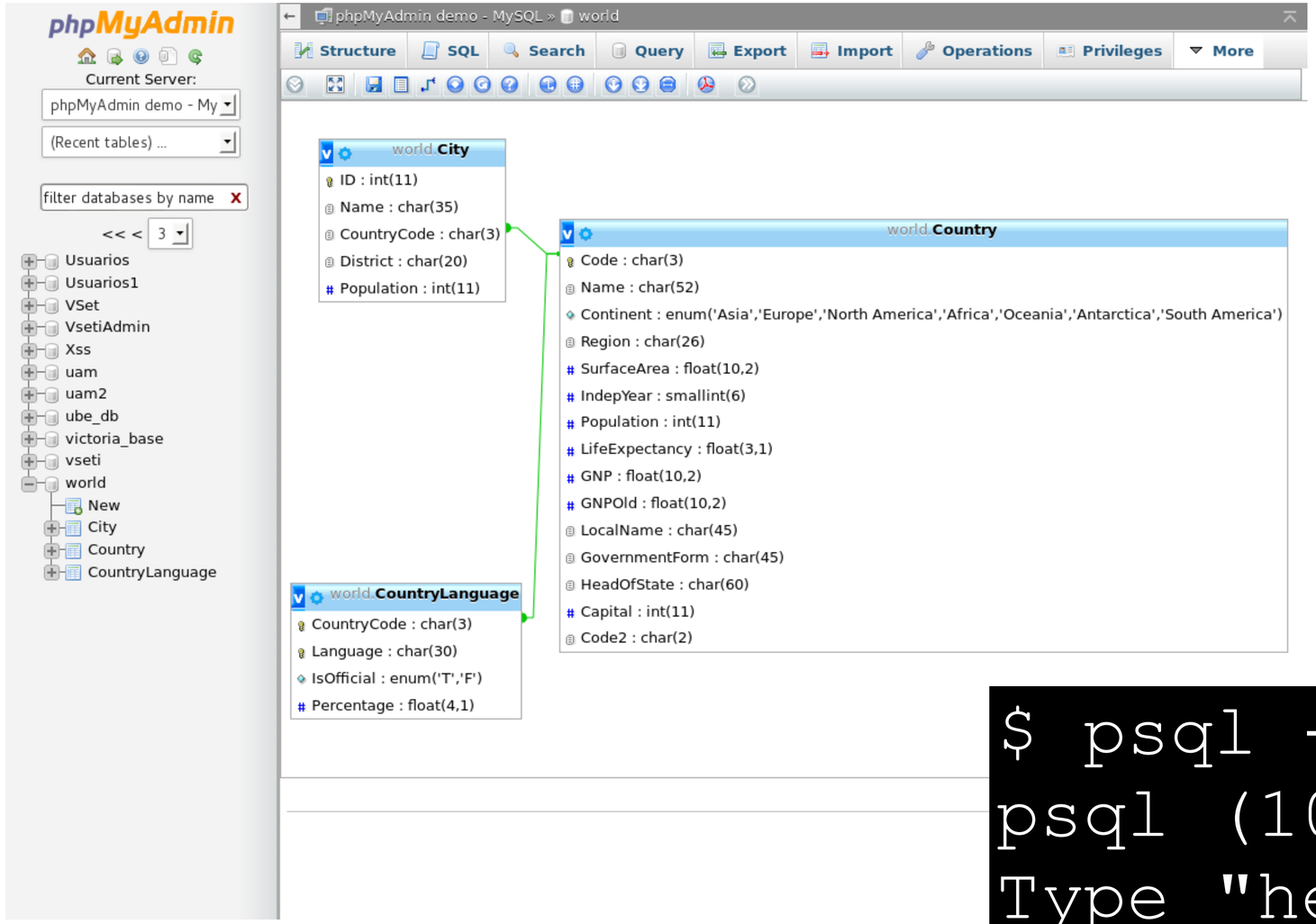
So Many Types of Types!

Numeric	Enumerated	XML
Monetary	Geometric	JSON
Character	Network Address	Arrays
Binary	Bit String	Composite
Date/Time	Text Search	Range
Boolean	UUID	Pseudo-Types

Databases are **purpose-built** for queries and joins.

BUT

Databases
aren't as **approachable**
as spreadsheets.




```
$ psql -d postgres
psql (10.4, server 9.6.9)
Type "help" for help.


postgres=#
```

Databases aren't as **flexible** as spreadsheets.

Databases are good at **storing** and **querying** data.

But that's it.

Spreadsheets and databases have **complementary** skillsets.

So, what do we **do** about it?

How to Solve Your Spreadsheet Problem

1. Identify the use case.
2. Stop the spread.
3. Backfill.

1. Identify the use case.

2. Stop the spread.

3. Backfill.

Every spreadsheet solves a **problem**.
What is that problem?

What's the business need?

How much data is there?

How fast does it change?

How frequent are additions?

1. Identify the use case.

2. Stop the spread.

3. Backfill.

Give new data a **structured home**.

Airtable

FileMaker®
An Apple Subsidiary

Ragic!

No custom apps.

At least at first.

Optimize for

Speed

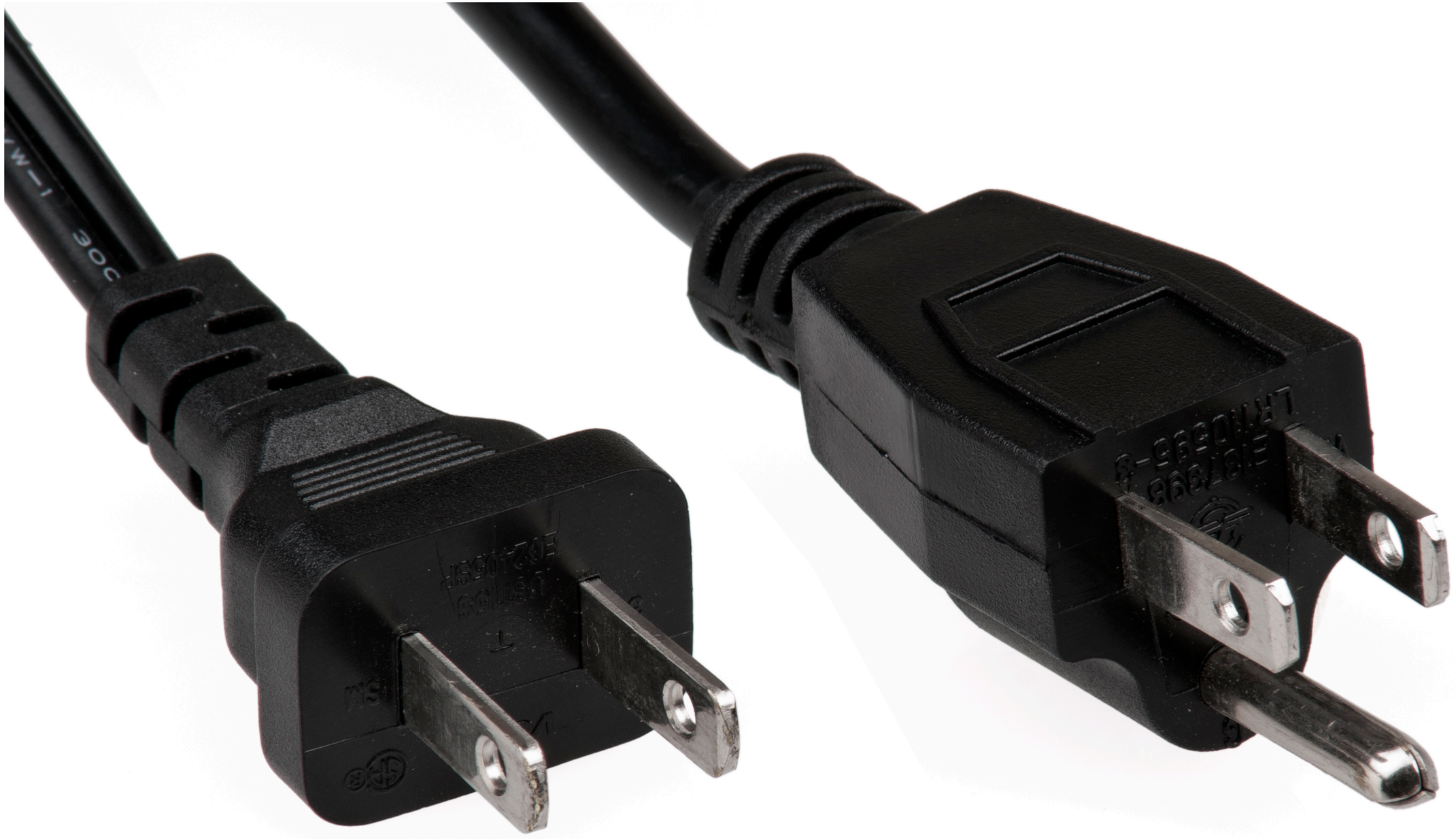
1. Identify the use case.
2. Stop the spread.
3. Backfill.

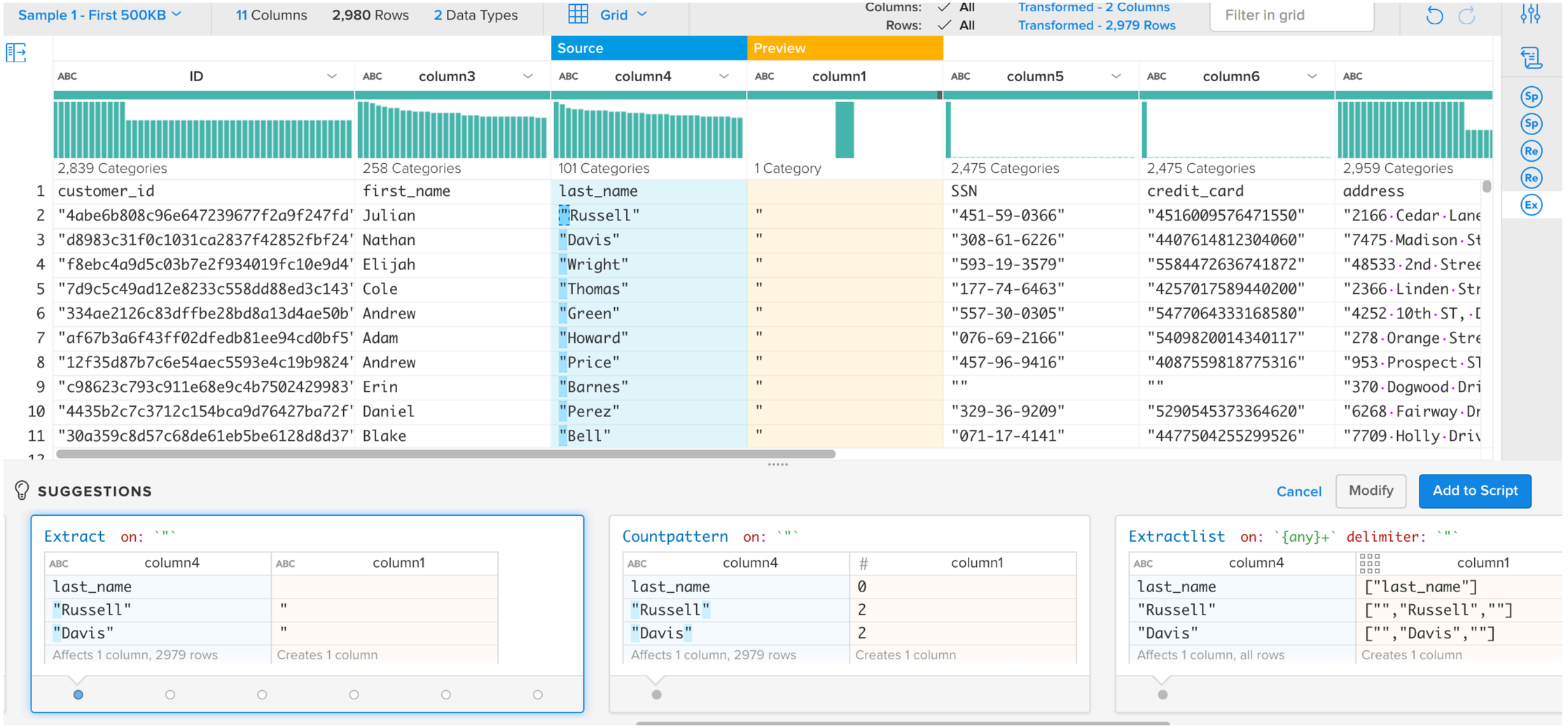
It’s time for some

Data Wrangling.

(yee-haw 🙄)

Writing one-off scripts is *sometimes* the best option.





https://www.trifacta.com/start-wrangling/

Infer wrangle “recipe” from high-level actions.

https://www.trifacta.com/start-wrangling/

Another Option: Programming By Example

FlashRelate

	A	B	C	D	E	. . .	R
1		value	year	value	year		Comments
2	Albania	1,000	1950	930	1981		FRA 1
3	Austria	3,139	1951	3,177	1955		FRA 3
4	Belgium	541	1947	601	1950		
5	Bulgaria	2,964	1947	3,259	1958		FRA 1
6	Czech . . .	2,416	1950	2,503	1960		NC

. . .

(a)

	A	B	C	D
1	Albania	1,000	1950	FRA 1
2	Albania	930	1981	FRA 1
	. . .			
5	Austria	3,139	1951	FRA 3
6	Austria	3,177	1955	FRA 3
	. . .			
9	Belgium	541	1947	
10	Belgium	601	1950	

. . .

(b)

Provide example rows, synthesize *layout* transformations.

https://github.com/microsoft/prose

Foofah

	A	B	C	D	E	F	G
1	Description	9/14/2009	9/15/2009	9/16/2009	9/17/2009	9/18/2009	
2	Item 4	900	0	1800	1800		
3	Item 6					1200	
4	Item 8		1800				
5							

	A	B	C	D
1	Item 4	9/14/2009	900	
2	Item 4	9/15/2009	0	
3	Item 4	9/16/2009	1800	
4	Item 4	9/17/2009	1800	
5	Item 6	9/18/2009	1200	
6	Item 8	9/15/2009	1800	
7				

Provide input/output sample, synthesize *layout* and *syntactic* transformations.

https://github.com/umich-dbgroup/foofah

How to Solve Your Spreadsheet Problem

1. Identify the use case.
2. Stop the spread.
3. Backfill.

What about the **future**?

Spreadsheets aren't going anywhere, **for good reason**.

Learn from the spreadsheet.

Meet the users **where they are**.

Thank you.

@alexras

Consulting Inquiries: contact@bitsondisk.com