Are We Data Scientists or Data Janitors?

Nenshad Bardoliwalla

February 13, 2014

Are We Data Scientists or Data Janitors?

To change the world around us, we first need to understand it. That understanding comes from data—but data is dirty, incomplete and complicated. As any analyst will tell you, much of what passes for data science is janitorial work. And it's only getting worse: rather than preparing and augmenting data once, today's exploratory approach to ad-hoc analytics means that every query, every new question, is yet another round of scrubbing, joining, and augmenting.

In other words, if you want to be data-driven, you first need to drive through the bottleneck of data preparation—often many times over.

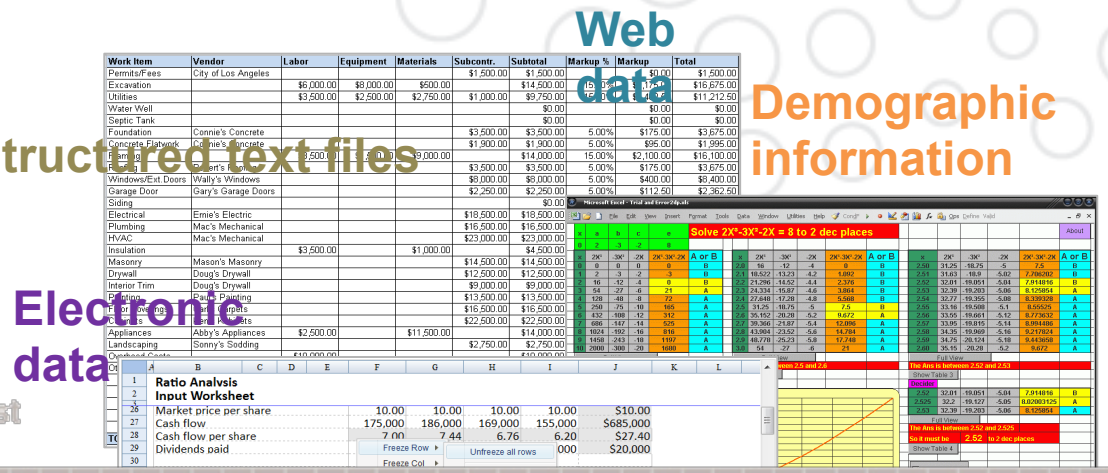
That's changing. A new wave of data preparation tools blend human insight and machine automation to edit huge amounts of data in real time. Join Paxata's Nenshad Bardoliwalla for a look at the new breed of data preparation tools that use semantic algorithms to detect data types, apply machine learning to find hidden patterns, and link related columns of data automatically. The result is more than just a reduction in preparation time—it's an entirely new perspective on what's in your data, how fast you can understand it and what more you can do with it.

Paxata

The Pain of Every Analytic Exercise

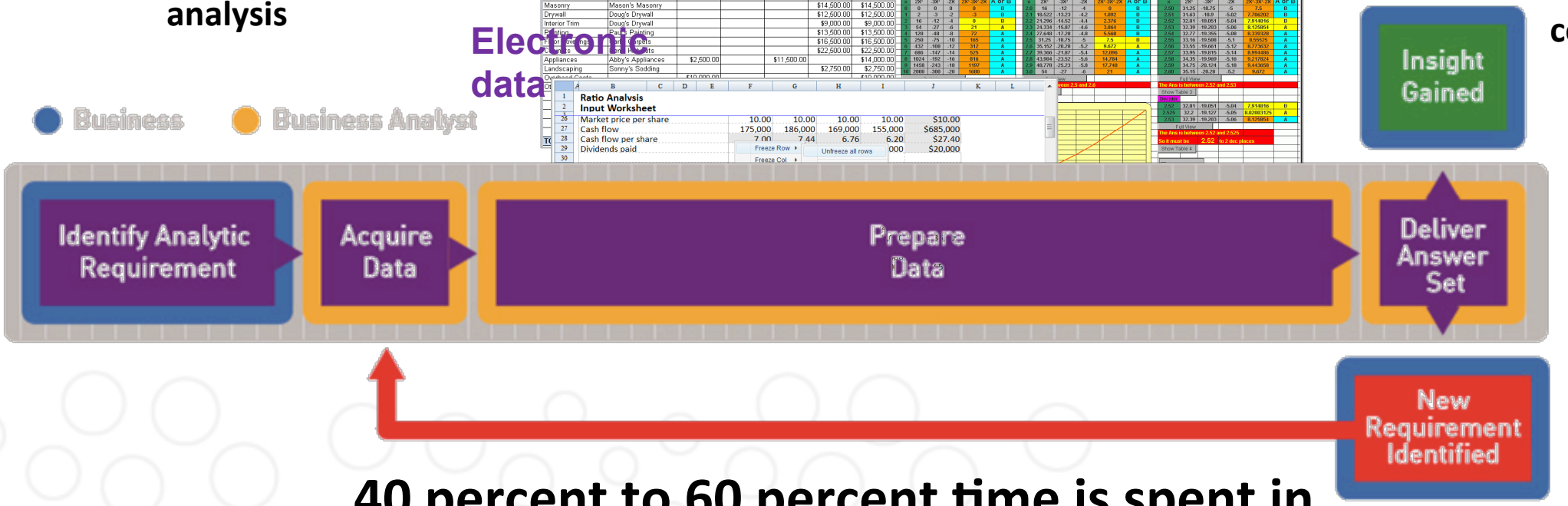
49%

time spent on preparing data for analysis



47%

time reviewing data for quality and consistency issues

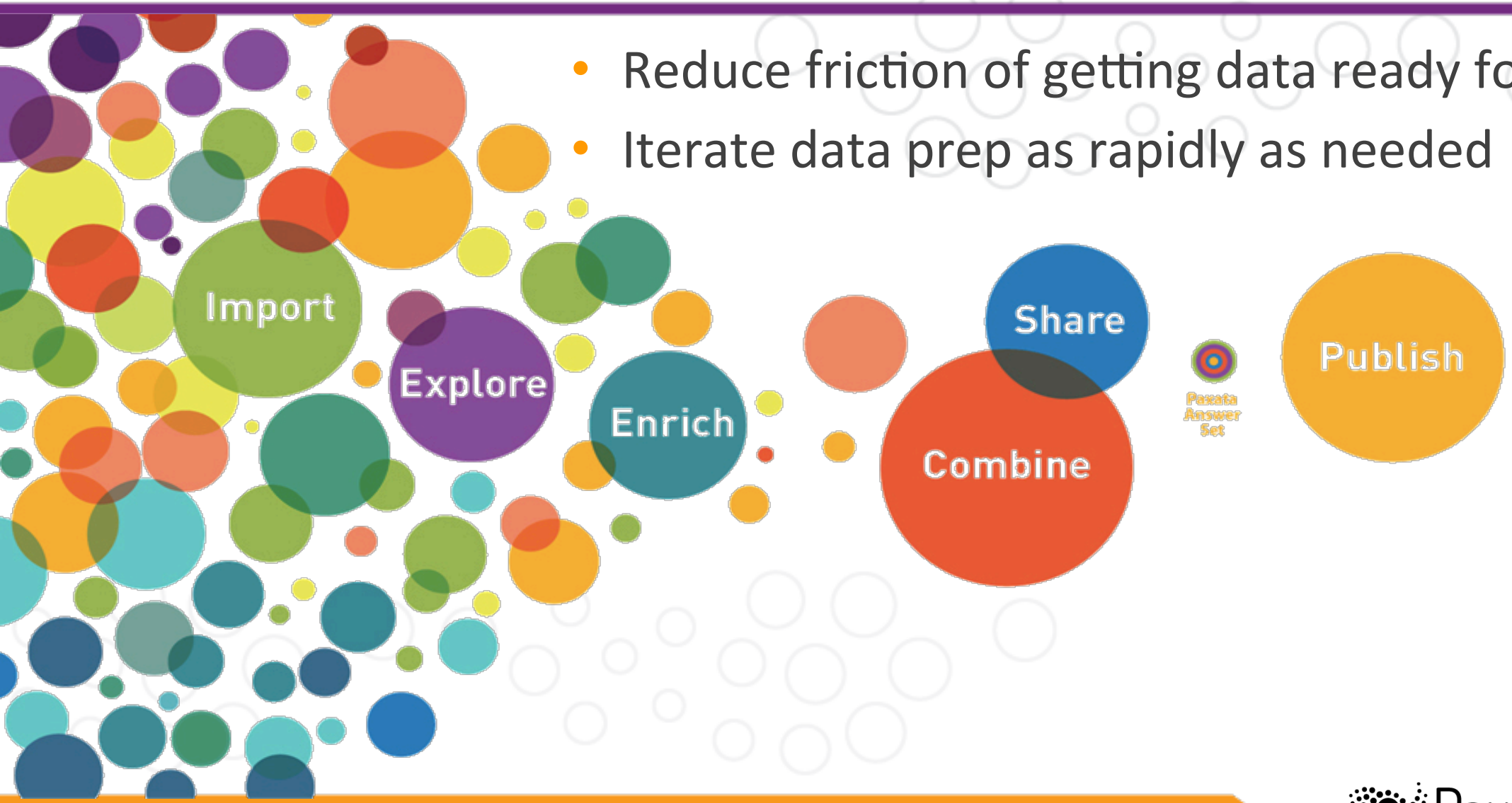


40 percent to 60 percent time is spent in data preparation phase that *precedes* actual analysis of the data

Source: Ventana's benchmark research into Information Optimization

Raw Data to Ready Data in Minutes…Not Months

- Reduce friction of getting data ready for analytics
- Iterate data prep as rapidly as needed



Paxata

Use Case 1: Detecting Data Problems

Janitor

Scrolls through every row, and every column

Tries to find white spaces, nulls, and occurrences of repeat values

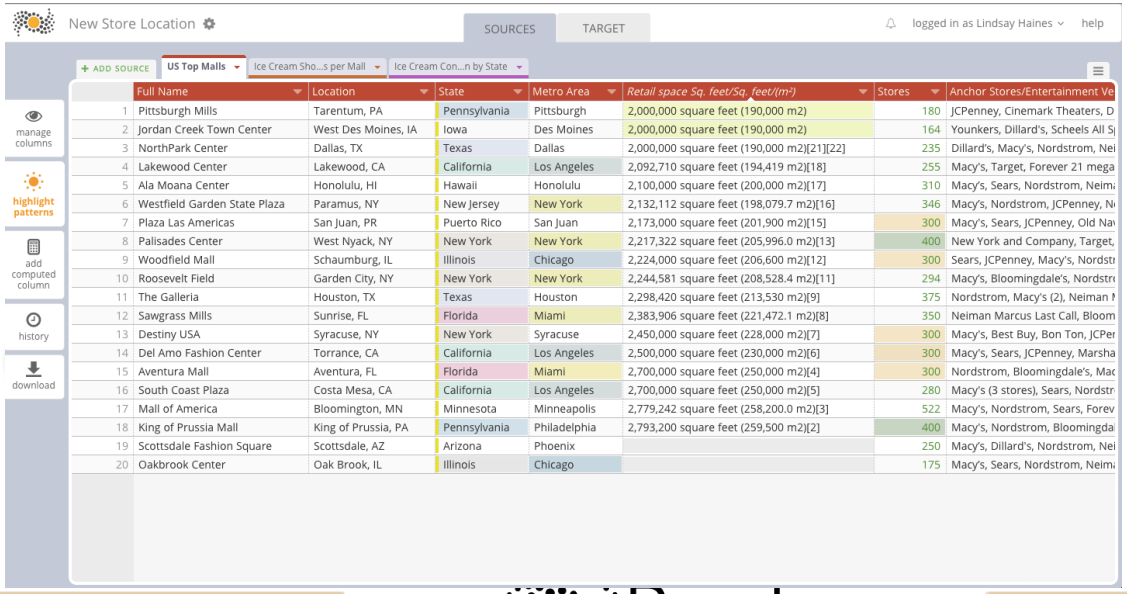
Gets a call from the boss, who found them first

Scientist

Presses a button that highlights white spaces, nulls, and occurrences of repeat values

PS. He fixes them fast too (as we'll see later!)

	A	B	C	D	E	F
1	**Full Name**	**Location**	**State**	**Metro Area**	**Retail space Sq. feet/Sq. feet/(m²)**	**Stores**
2	King of Prussia Mall	King of Prussia, PA	Pennsylvania	Philadelphia	2,793,200 square feet (259,500 m2)[2]	400
3	Mall of America	Bloomington, MN	Minnesota	Minneapolis	2,779,242 square feet (258,200.0 m2)[3]	522
4	Aventura Mall	Aventura, FL	Florida	Miami	2,700,000 square feet (250,000 m2)[4]	300
5	South Coast Plaza	Costa Mesa, CA	California	Los Angeles	2,700,000 square feet (250,000 m2)[5]	280
6	Del Amo Fashion Center	Torrance, CA	California	Los Angeles	2,500,000 square feet (230,000 m2)[6]	300
7	Destiny USA	Syracuse, NY	New York	Syracuse	2,450,000 square feet (228,000 m2)[7]	300
8	Sawgrass Mills	Sunrise, FL	Florida	Miami	2,383,906 square feet (221,472.1 m2)[8]	350
9	The Galleria	Houston, TX	Texas	Houston	2,298,420 square feet (213,530 m2)[9]	375
10	Roosevelt Field	Garden City, NY	New York	New York	2,244,581 square feet (208,528.4 m2)[11]	294
11	Woodfield Mall	Schaumburg, IL	Illinois	Chicago	2,224,000 square feet (206,600 m2)[12]	300
12	Palisades Center	West Nyack, NY	New York	New York	2,217,322 square feet (205,996.0 m2)[13]	400
13	Plaza Las Americas	San Juan, PR	Puerto Rico	San Juan	2,173,000 square feet (201,900 m2)[15]	300
14	Westfield Garden State Plaza	Paramus, NY	New Jersey	New York	2,132,112 square feet (198,079.7 m2)[16]	346
15	Ala Moana Center	Honolulu, HI	Hawaii	Honolulu	2,100,000 square feet (200,000 m2)[17]	310
16	Lakewood Center	Lakewood, CA	California	Los Angeles	2,092,710 square feet (194,419 m2)[18]	255
	Scottsdale Fashion					



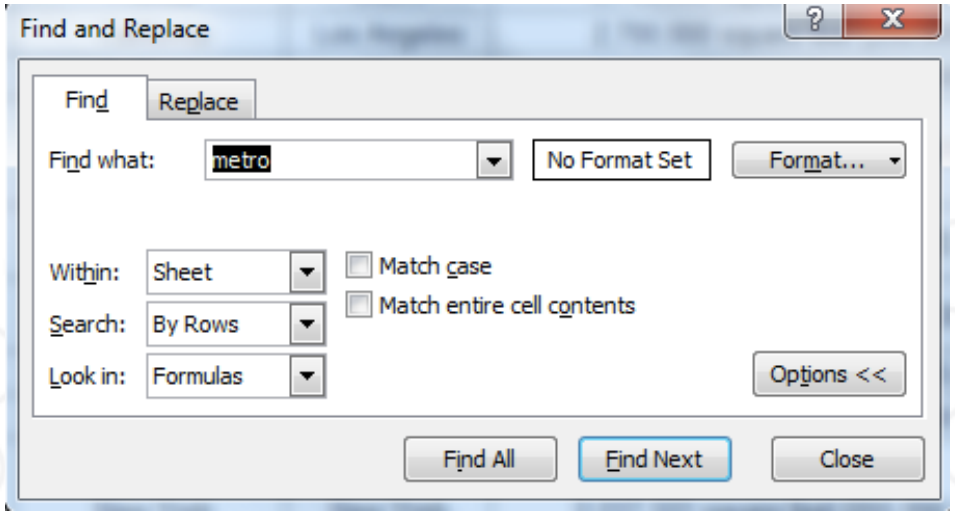
Use Case 2: Finding Datasets, Columns, Values

Janitor

Uses desktop, e-mail, fileshare search to find projects?

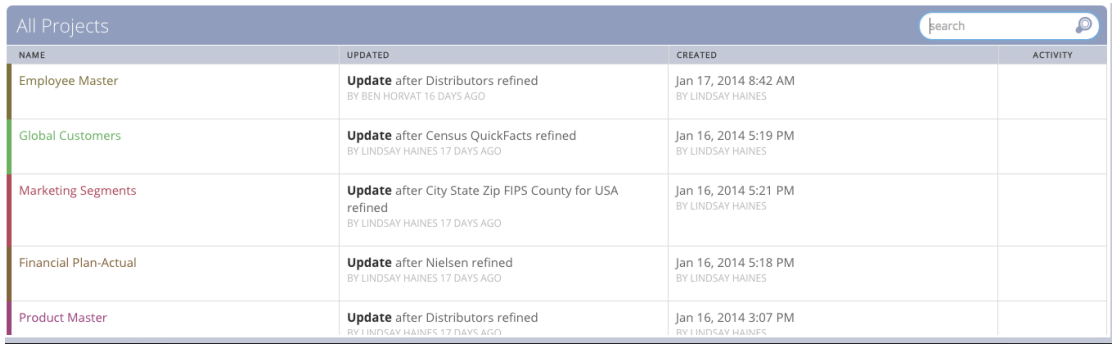
Uses find and replace to find column names and values…but only in a given workbook

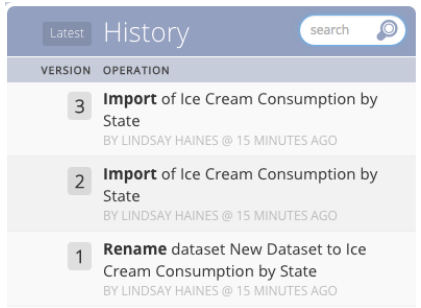
Uses imagination to find preparation operations



Scientist

Uses pervasive search to find projects, preparation tasks, column names, and column values.





Paxata

Use Case 3: Creating Histograms of Values

Janitor

Writes exciting formulas…again and again!

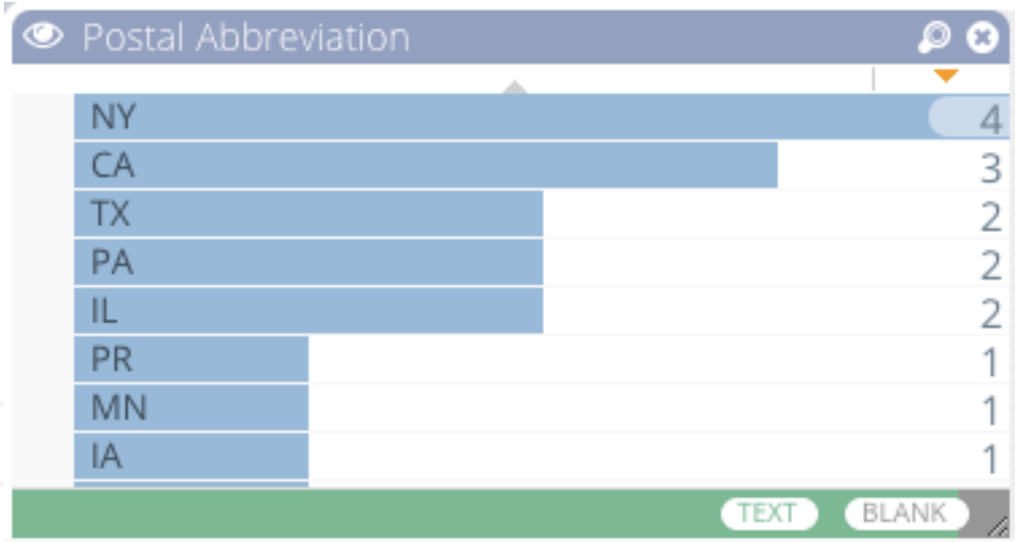
=SUM(IF(FREQUENCY(IF(LEN(A2:A10)>0,MATCH(A2:A10,A2:A10,0),""),IF(LEN(A2:A10)>0,MATCH(A2:A10,A2:A10,0),""))>0,1))

"Count the number of unique text and number values in cells A2:A10 , but do not count blank cells or text values (6)"

```
=SUM(IF(FREQUENCY(IF(LEN(A2:A10)>0,MATCH(A2:A10,A2:A10,0),"")
```

Scientist

Uses filtergrams to count unique values, see them in a histogram. Click a button to remove errors or blanks



Use Case 4: Exploring with Crossfilters

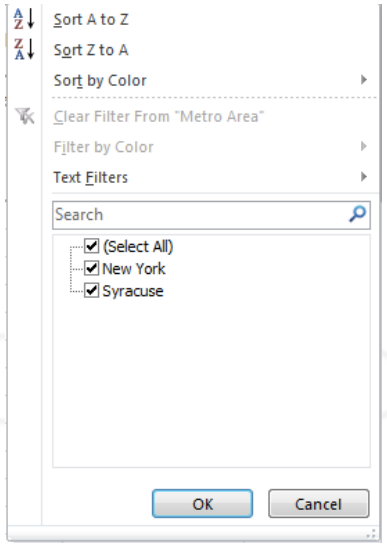
Janitor

Uses Autofilter!

Makes multiple clicks for each column selection.

Unsure what's filtered or not.

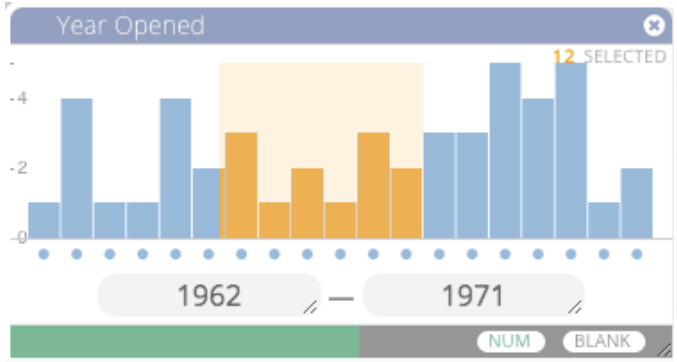
PS. No histograms.

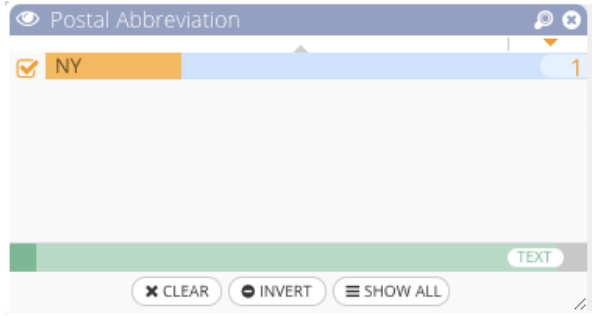


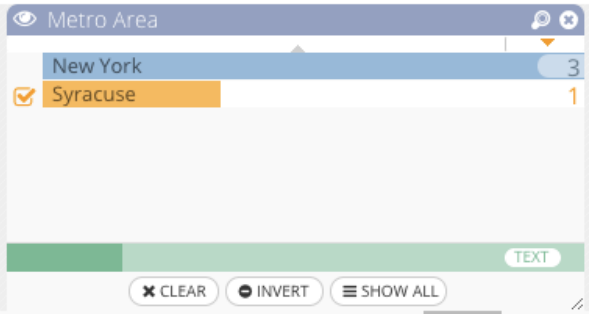
Scientist

Creates N filtergrams

Clicks on whichever values she wants to filter on







Use Case 5: Find & Fix Spelling Variations

Janitor

Sorts

Scans

Standardizes

Repeats

CRIES. A LOT.

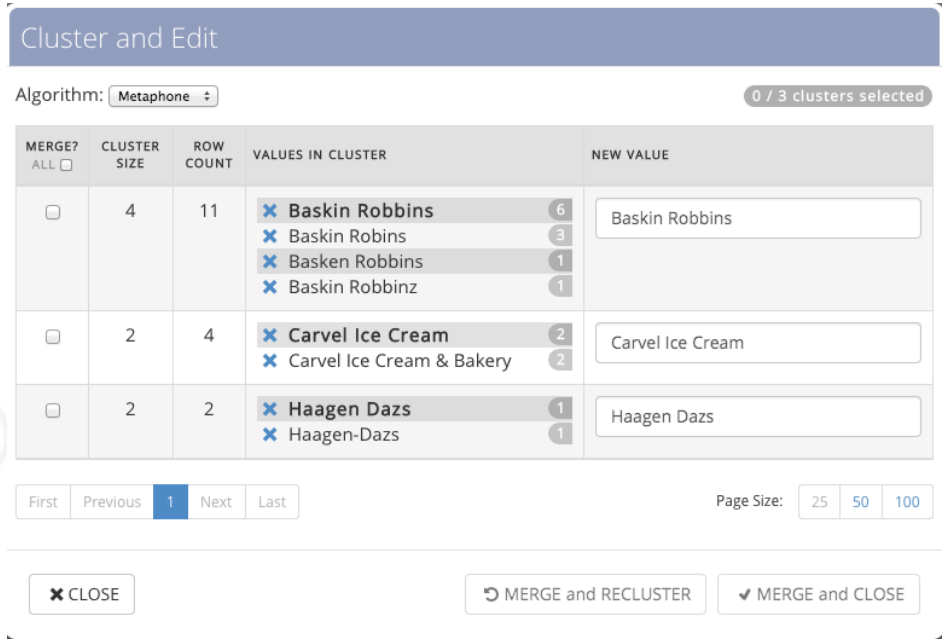
Baskin Robbins
Baskin Robbins
Baskin Robbinz
Baskin Robins

PS. There is an unsupported Fuzzy Lookup tool for Excel 2010

Scientist

Clicks the "cluster and edit" menu option.

Uses a drop down to pick an algorithm and get a preview of the results.



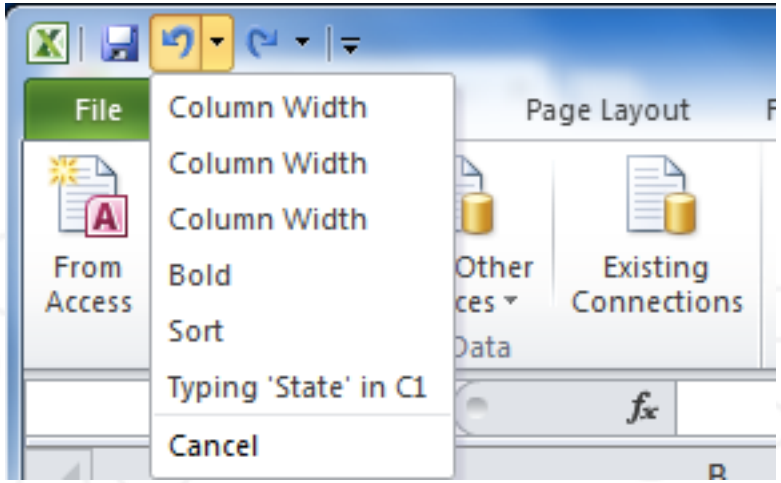
Use Case 6: Determine who did what when

Janitor

Clicks on the change log.

Sees every action is recorded.

Doesn't see any time stamps. Or user names. Or search. Guess she'll be on IM for a while.

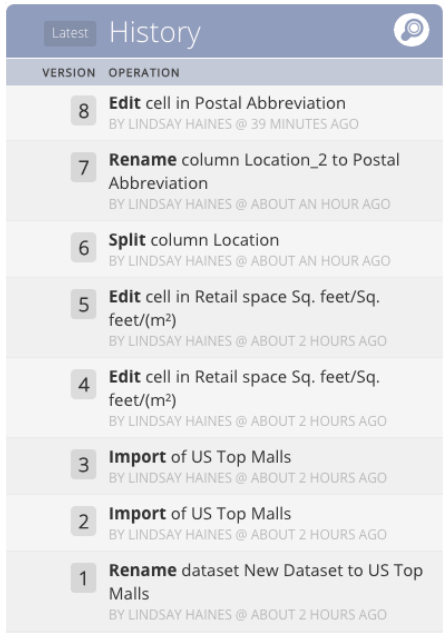


Scientist

Click the history button.

Sees every action is stamped by time and who did them.

Searches for the offending op.

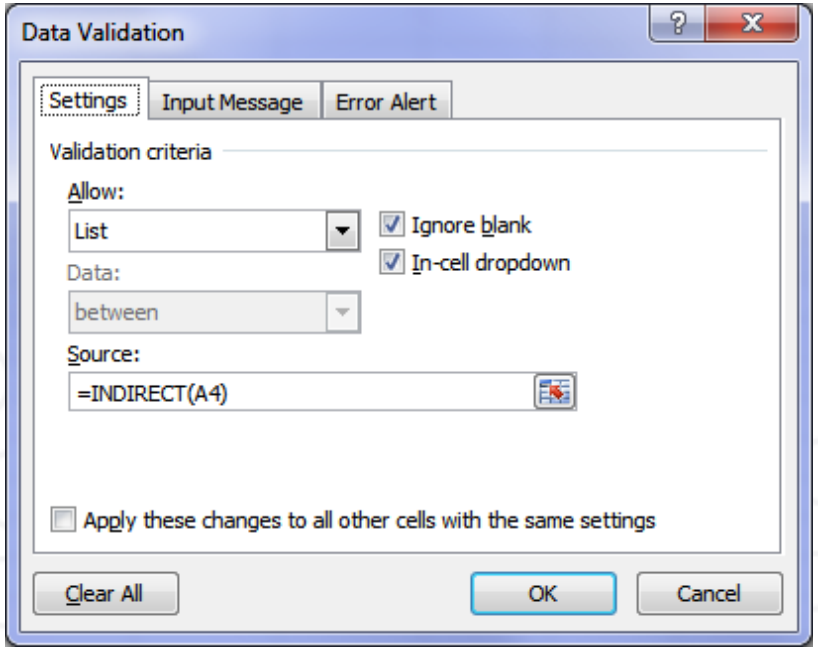


Use Case 7: Validate values in a column

Janitor

Copies the correct values from trusted source into your workbook.

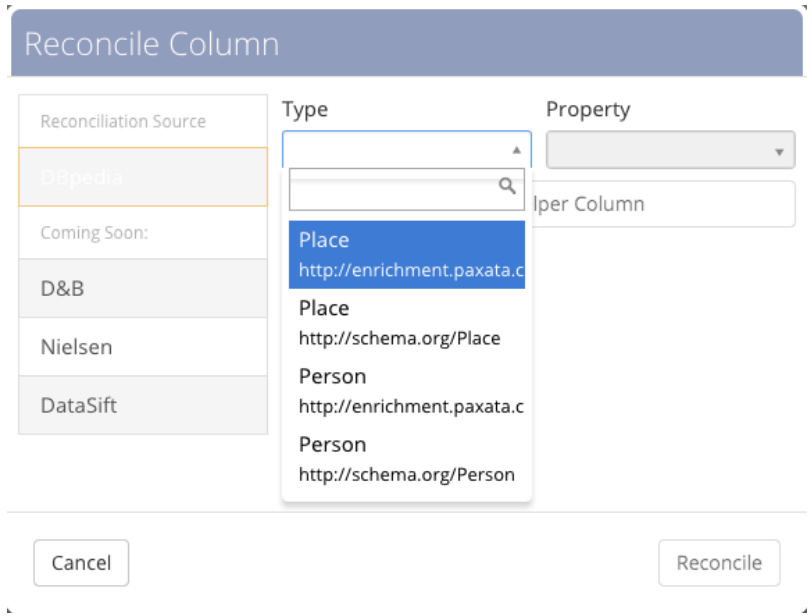
Writes conditional formulas or use the dependent data validation function.



Scientist

Clicks reconcile column. The system guesses what type of data it's looking at based on the trusted source.

Automatically matches the values and shows you the exceptions.



Use Case 8: Enrich data with additional info

Janitor

Obtains the data she wants to enrich with, maybe from data.gov

Copies the enrichment values from data source into her workbook.

Writes multiple VLOOKUPS.

Put all of the data together column by column

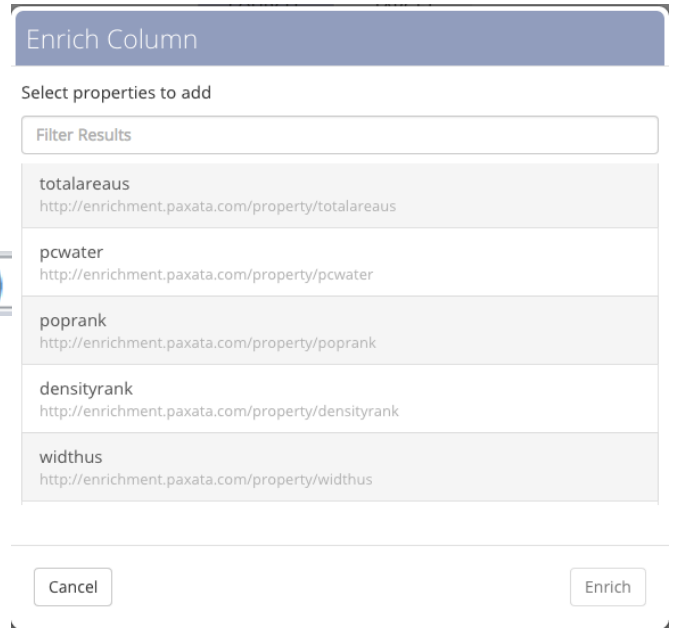
```
=IF(ISNA(VLOOKUP(5,A2:E7,2,FALSE)) = TRUE, "State not found", VLOOKUP(5,A2:E7,2,FALSE))
```

Scientist

Clicks the enrich column option.

Gets a list of recommended attributes to pull in…on the fly.

Picks those she wants to add. Enjoys!



Use Case 9: Fill in blank values

Janitor

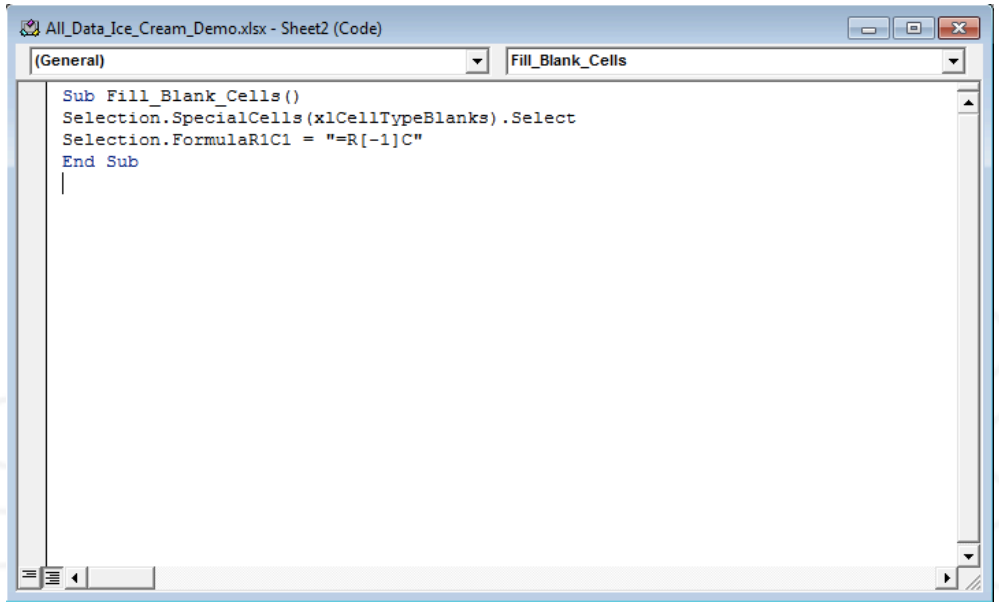
Chooses the column with the blank values

Searches for VBA reference

Clicks Developer > Visual Basic,

Clicks Insert > Module

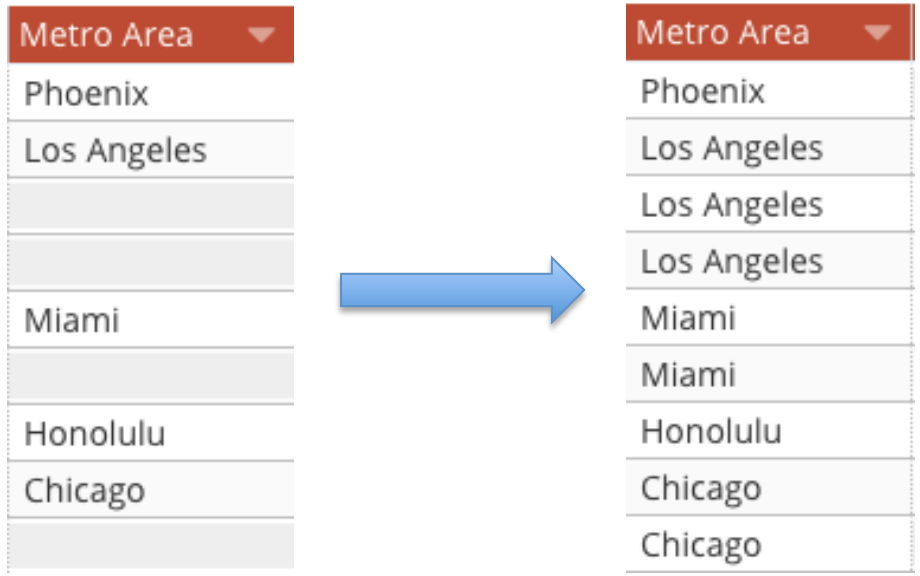
Inputs the following code into the Module:



Scientist

Chooses the column with the blank values

Clicks the “filldown” column operation.



Use Case 10: Combine data sets

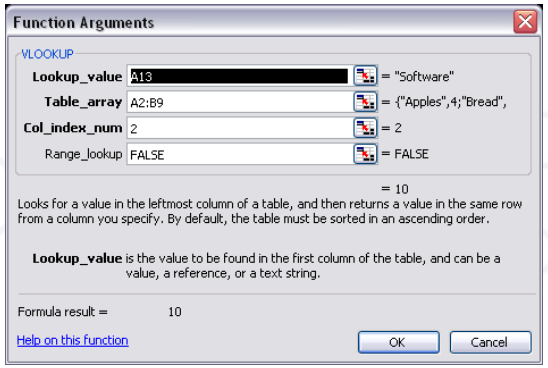
Janitor

Scans the column names of every dataset for obvious candidates

Scans all the row values in every column in every dataset

Tries VLOOKUPs on every possible set of columns that might match

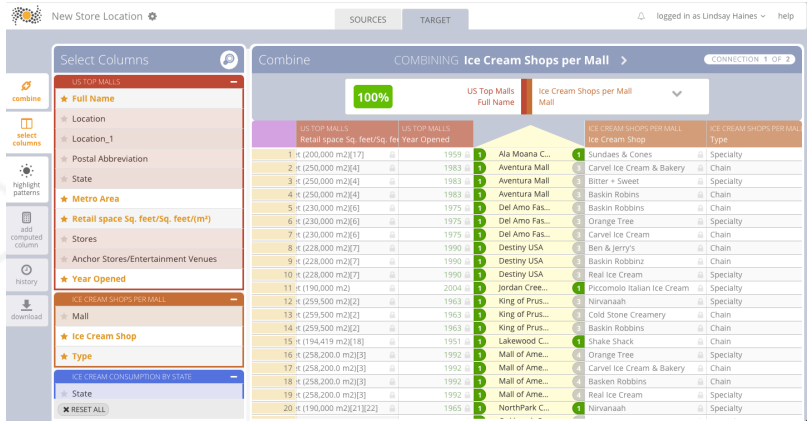
Chooses the VLOOKUPs that look right for every additional column he wants to add



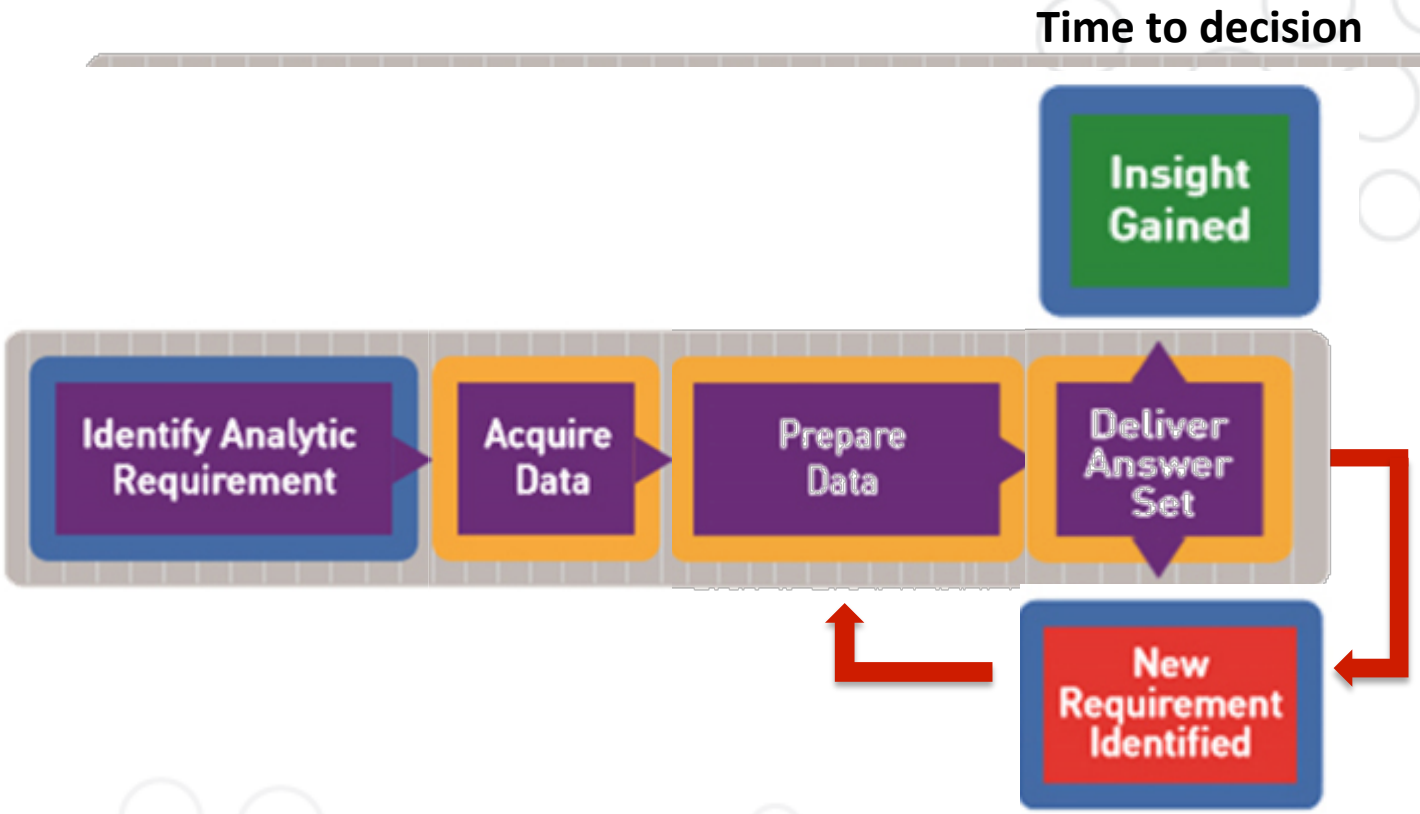
Scientist

Chooses the columns he wants in his answerset.

Lets the intelligent combination wizard tell him how to fit the data together, what the cardinality is, and what the data will look like when combined.

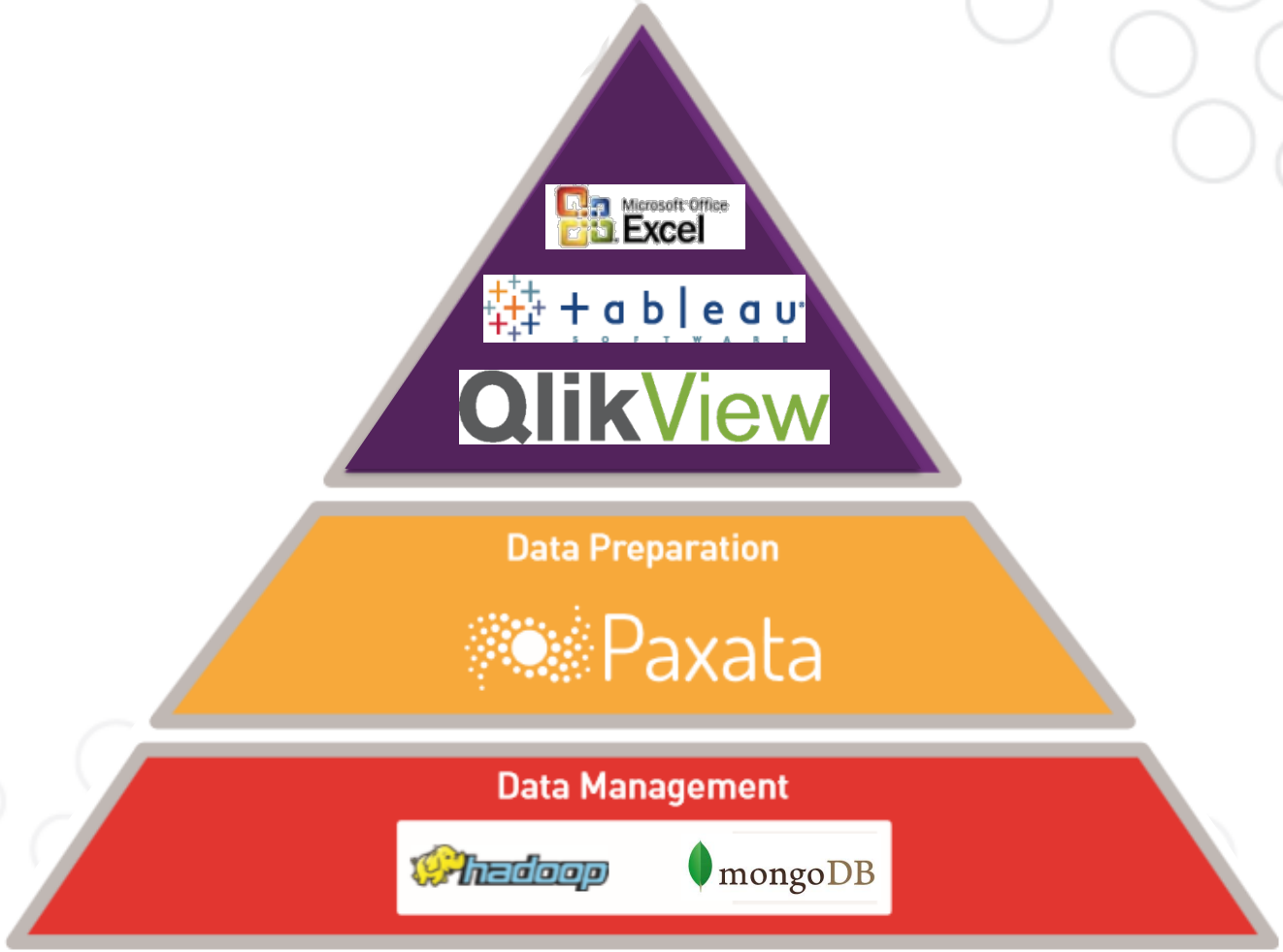


With Paxata



- Reduce the friction of getting answer sets delivered to data discovery tools
- Repeat the data prep process as rapidly as business iterates needs or asks more questions

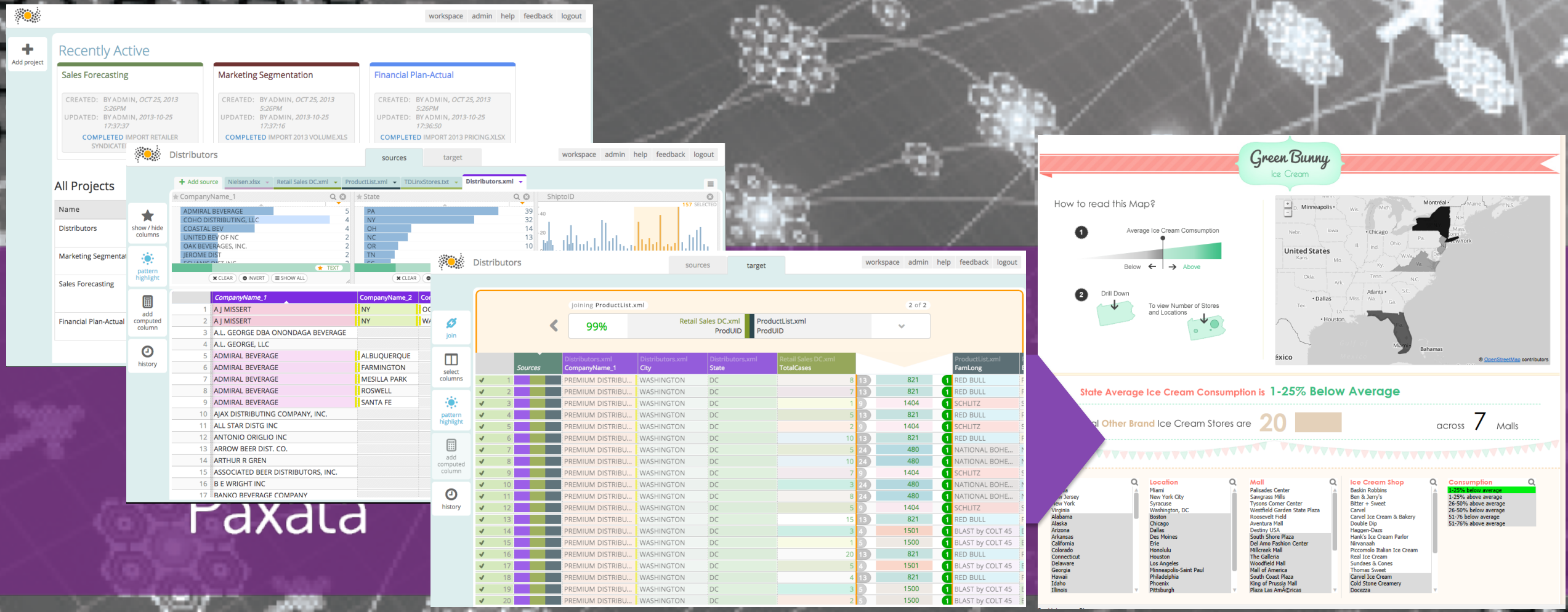
The Missing Link in the Analytics Triangle



- Paxata bridges the gap between next generation data management solutions and data discovery and visual analytic tools
- Paxata delivers an adaptive data preparation platform built for the business analyst

Paxata

YOU DON’T HAVE TO BE A JANITOR ANYMORE!!!



www.paxata.com/schedule-a-demo

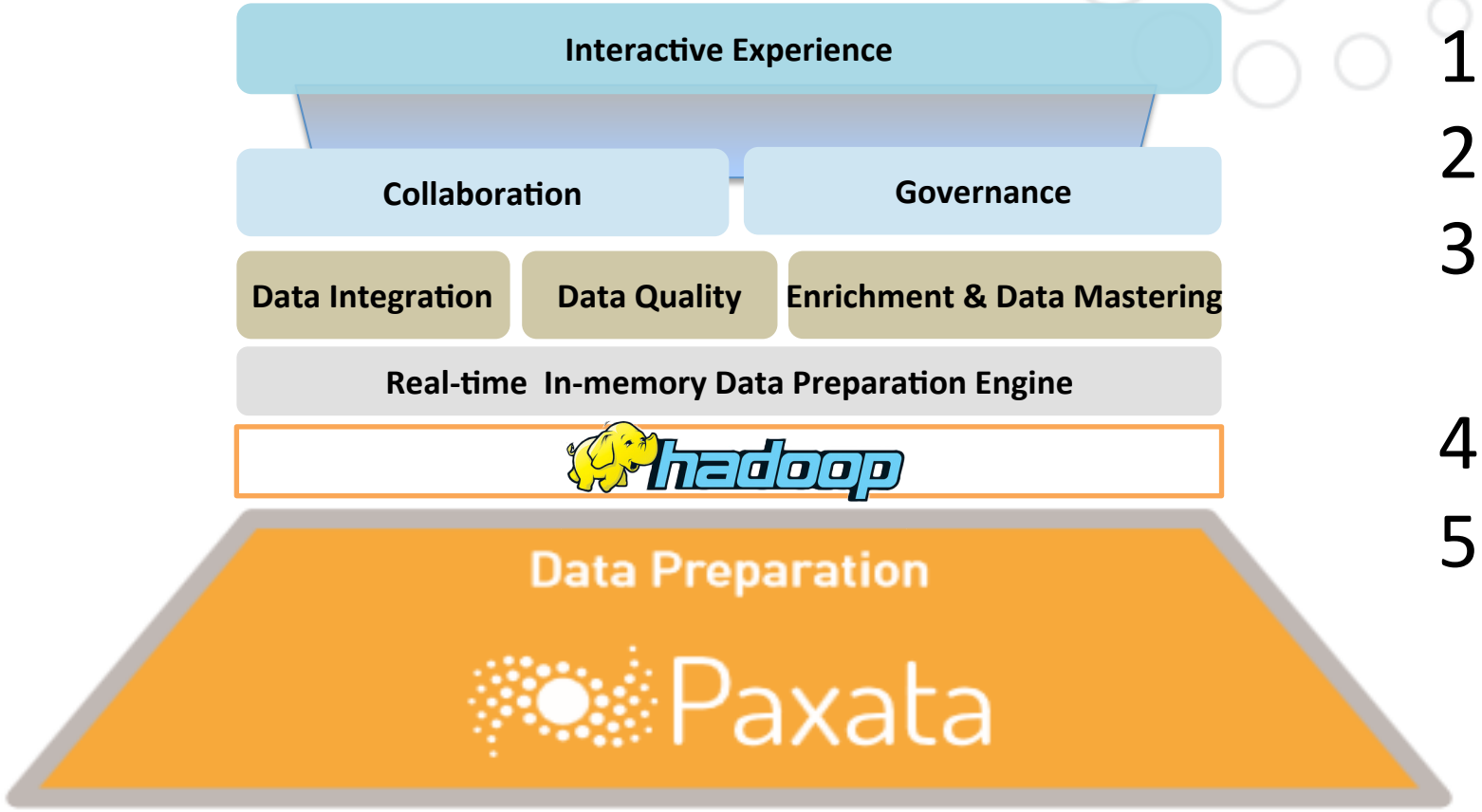
THANKS!

www.paxata.com
@PaxataInc
Facebook/Paxata
LinkedIn/company/Paxata
Youtube/PaxataTV

A Quick Look at Paxata

- Founded in 2012; Headquarters: Redwood City, CA
- Seasoned team
- Proven success in technology delivery – Public & Private Cloud
- Happy, deployed customers of every size
- Partner ecosystem
- Venture backed: ACCEL Partners India and Walden | Riverwood
- IQT Investment: December 2013

The Adaptive Data Preparation Platform



1. Data integration
2. Data quality
3. Enrichment and data mastering
4. Collaboration
5. Governance

Paxata