

# 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.



# The Pain of Every Analytic Exercise

49%

time spent on preparing data for analysis

Unstructured text files

Web

Work Item	Vendor	Labor	Equipment	Materials	Subcont.	Subtotal	Markup %	Markup	Total
Permits/Fees	City of Los Angeles				\$1,500.00	\$1,500.00		\$0.00	\$1,500.00
Excavation		\$6,000.00	\$8,000.00	\$500.00		\$14,500.00	15.00%	\$2,025.00	\$16,675.00
Utilities		\$3,500.00	\$2,500.00	\$2,750.00	\$1,000.00	\$9,750.00	15.00%	\$1,462.50	\$11,212.50
Water Well						\$0.00		\$0.00	\$0.00
Septic Tank						\$0.00		\$0.00	\$0.00
Foundation	Connie's Concrete			\$3,500.00		\$3,500.00	5.00%	\$175.00	\$3,675.00
Concrete Flatwork	Connie's Concrete			\$1,900.00		\$1,900.00	5.00%	\$95.00	\$1,995.00
Form Work	Connie's Concrete			\$15,000.00		\$14,000.00	15.00%	\$2,100.00	\$16,100.00
Form Work	Connie's Concrete			\$3,500.00		\$3,500.00	5.00%	\$175.00	\$3,675.00
Windows/Ext Doors	Wally's Windows			\$8,000.00		\$8,000.00	5.00%	\$400.00	\$8,400.00
Garage Door	Gary's Garage Doors			\$2,250.00		\$2,250.00	5.00%	\$112.50	\$2,362.50
Siding				\$0.00		\$0.00		\$0.00	\$0.00
Electrical	Emie's Electric			\$18,500.00		\$18,500.00			\$18,500.00
Plumbing	Mac's Mechanical			\$16,500.00		\$16,500.00			\$16,500.00
HVAC	Mac's Mechanical			\$23,000.00		\$23,000.00			\$23,000.00
Insulation		\$3,500.00		\$1,000.00		\$4,500.00			\$4,500.00
Masonry	Mason's Masonry			\$14,500.00		\$14,500.00			\$14,500.00
Drywall	Doag's Drywall			\$12,500.00		\$12,500.00			\$12,500.00
Interior Trim	Doag's Drywall			\$9,000.00		\$9,000.00			\$9,000.00
Painting	Carl's Painting			\$13,500.00		\$13,500.00			\$13,500.00
Flooring	Carl's Painting			\$16,500.00		\$16,500.00			\$16,500.00
Lighting	Carl's Painting			\$22,500.00		\$22,500.00			\$22,500.00
Appliances	Abby's Appliances	\$2,500.00		\$11,500.00		\$14,000.00			\$14,000.00
Landscaping	Sonny's Siding			\$2,750.00		\$2,750.00			\$2,750.00
Open/Close/Cont.		\$10,000.00				\$10,000.00			\$10,000.00

Demographic information

47%

time reviewing data for quality and consistency issues

Electronic data

Insight Gained

Business Business Analyst

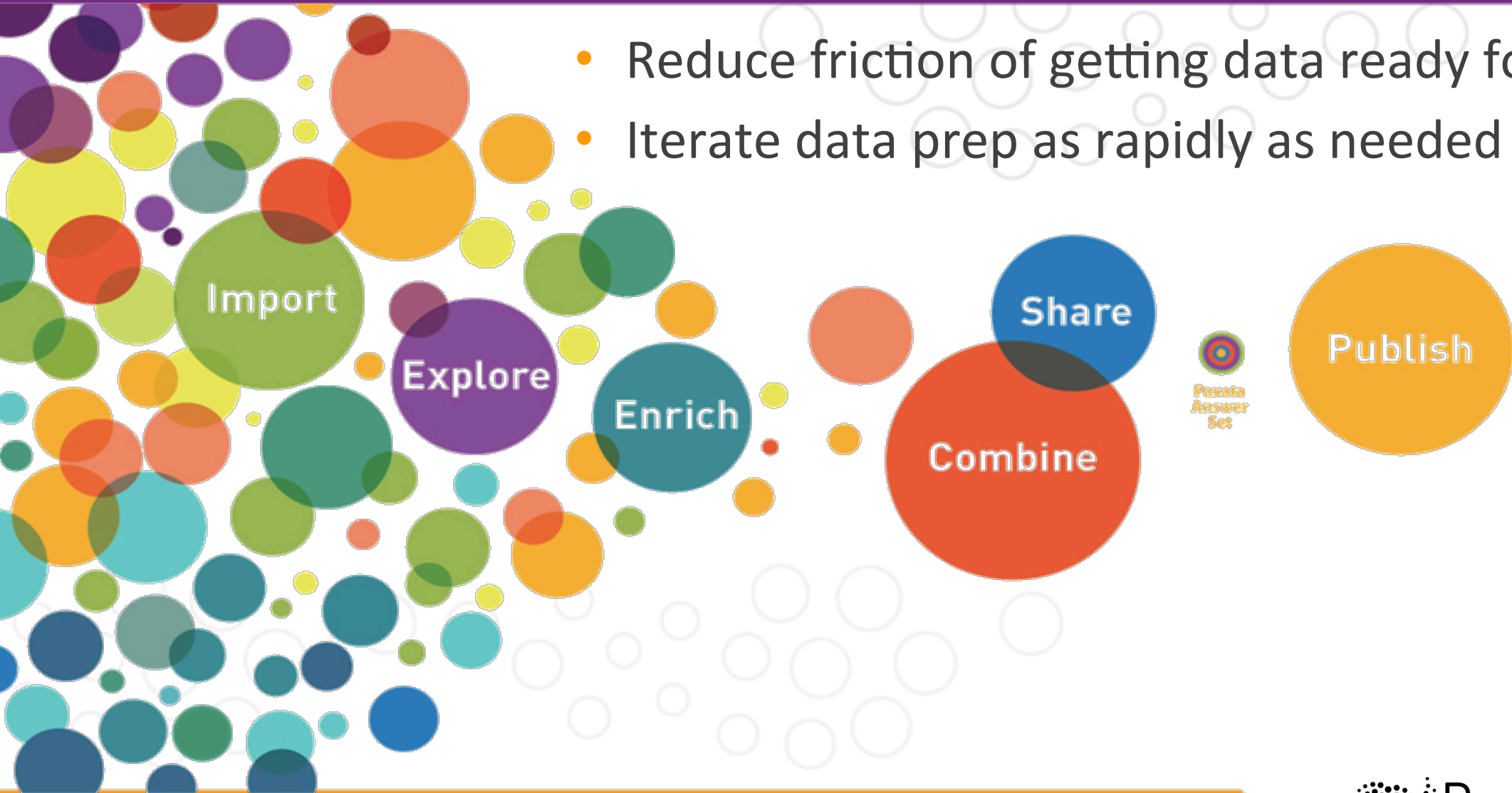


New Requirement Identified

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

# Raw Data to Ready Data in Minutes...Not Months

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



# 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
	Full Name	Location	State	Metro Area	Retail space Sq. feet/Sq. feet(m <sup>2</sup> )	Stores
1	King of Prussia Mall	King of Prussia, PA	Pennsylvania	Philadelphia	2,793,200 square feet (259,500 m <sup>2</sup> )	2
2	Mall of America	Bloomington, MN	Minnesota	Minneapolis	2,779,242 square feet (258,200.0 m <sup>2</sup> )	3
3	Aventura Mall	Aventura, FL	Florida	Miami	2,700,000 square feet (250,000 m <sup>2</sup> )	4
4	South Coast Plaza	Costa Mesa, CA	California	Los Angeles	2,700,000 square feet (250,000 m <sup>2</sup> )	5
5	Del Amo Fashion Center	Torrance, CA	California	Los Angeles	2,500,000 square feet (230,000 m <sup>2</sup> )	6
6	Destiny USA	Syracuse, NY	New York	Syracuse	2,450,000 square feet (228,000 m <sup>2</sup> )	7
7	Sawgrass Mills	Sunrise, FL	Florida	Miami	2,383,906 square feet (221,472.1 m <sup>2</sup> )	8
8	The Galleria	Houston, TX	Texas	Houston	2,298,420 square feet (213,530 m <sup>2</sup> )	9
9	Roosevelt Field	Garden City, NY	New York	New York	2,244,581 square feet (208,528.4 m <sup>2</sup> )	11
10	Woodfield Mall	Schaumburg, IL	Illinois	Chicago	2,224,000 square feet (206,600 m <sup>2</sup> )	12
11	Palisades Center	West Nyack, NY	New York	New York	2,217,322 square feet (205,996.0 m <sup>2</sup> )	13
12	Plaza Las Americas	San Juan, PR	Puerto Rico	San Juan	2,173,000 square feet (201,900 m <sup>2</sup> )	15
13	Westfield Garden State Plaza	Paramus, NY	New Jersey	New York	2,132,112 square feet (198,079.7 m <sup>2</sup> )	16
14	Ala Moana Center	Honolulu, HI	Hawaii	Honolulu	2,100,000 square feet (200,000 m <sup>2</sup> )	17
15	Lakewood Center	Lakewood, CA	California	Los Angeles	2,092,710 square feet (194,419 m <sup>2</sup> )	18
16	Scottsdale Fashion	Scottsdale, AZ	Arizona	Phoenix	2,076,460 square feet (193,070.0 m <sup>2</sup> )	19

The screenshot shows the Paxata interface for 'New Store Location'. It features a table with columns for Full Name, Location, State, Metro Area, Retail space Sq. feet/Sq. feet(m<sup>2</sup>), Stores, and Anchor Stores/Entertainment Venues. The table is filtered to show 'US Top Malls'. A sidebar on the left contains tools for managing columns, highlighting patterns, adding computed columns, viewing history, and downloading data. The 'highlight patterns' tool is currently active, highlighting specific rows in the table.

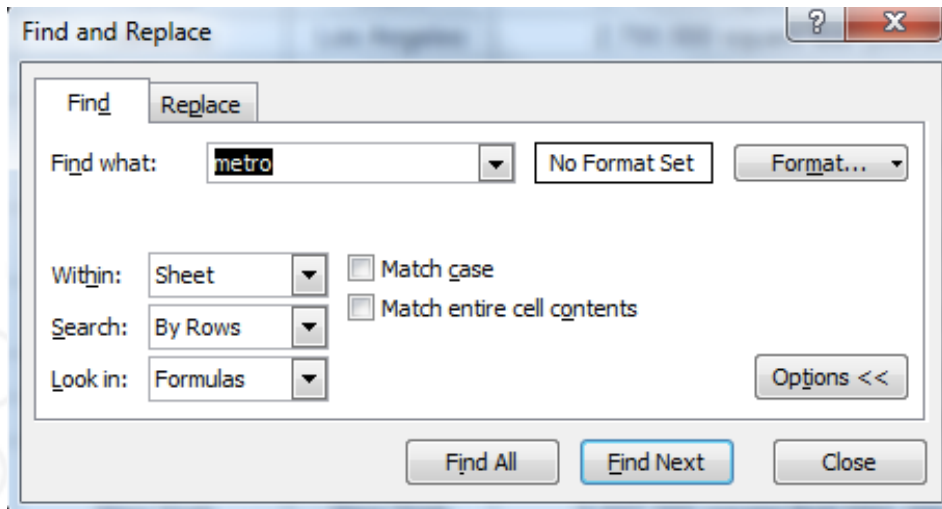
# Use Case 2: Finding Datasets, Columns, Values

## Janitor

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

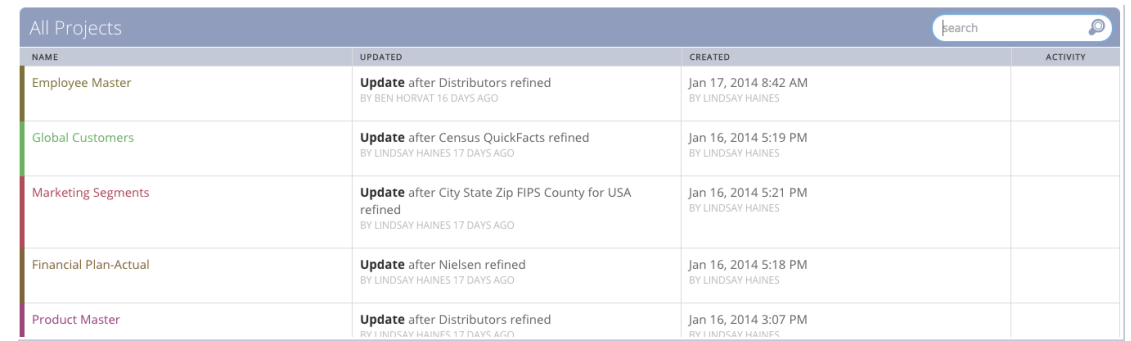
Uses find and replace to find column names and values...but only in a given workbook

Uses imagination to find preparation operations

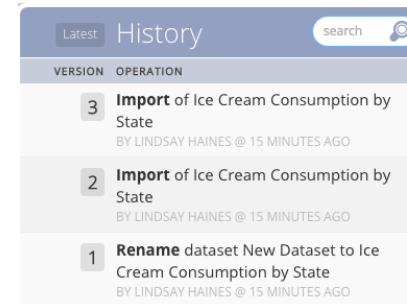


## Scientist

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



NAME	UPDATED	CREATED	ACTIVITY
Employee Master	Update after Distributors refined BY BEN HORVAT 16 DAYS AGO	Jan 17, 2014 8:42 AM BY LINDSAY HAINES	
Global Customers	Update after Census QuickFacts refined BY LINDSAY HAINES 17 DAYS AGO	Jan 16, 2014 5:19 PM BY LINDSAY HAINES	
Marketing Segments	Update after City State Zip FIPS County for USA refined BY LINDSAY HAINES 17 DAYS AGO	Jan 16, 2014 5:21 PM BY LINDSAY HAINES	
Financial Plan-Actual	Update after Nielsen refined BY LINDSAY HAINES 17 DAYS AGO	Jan 16, 2014 5:18 PM BY LINDSAY HAINES	
Product Master	Update after Distributors refined BY LINDSAY HAINES 17 DAYS AGO	Jan 16, 2014 3:07 PM BY LINDSAY HAINES	



VERSION	OPERATION
3	Import of Ice Cream Consumption by State BY LINDSAY HAINES @ 15 MINUTES AGO
2	Import of Ice Cream Consumption by State BY LINDSAY HAINES @ 15 MINUTES AGO
1	Rename dataset New Dataset to Ice Cream Consumption by State BY LINDSAY HAINES @ 15 MINUTES AGO



COLUMN	SOURCE
State	ICE CREAM CONSUM...
Total Area	ICE CREAM CONSUM...
Population (July 1, 2...	ICE CREAM CONSUM...
GDP USD 2010	ICE CREAM CONSUM...
Average Ice Cream ...	ICE CREAM CONSUM...

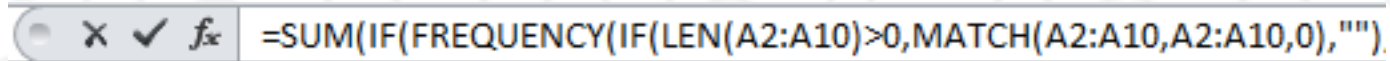
# Use Case 3: Creating Histograms of Values

## 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))
```

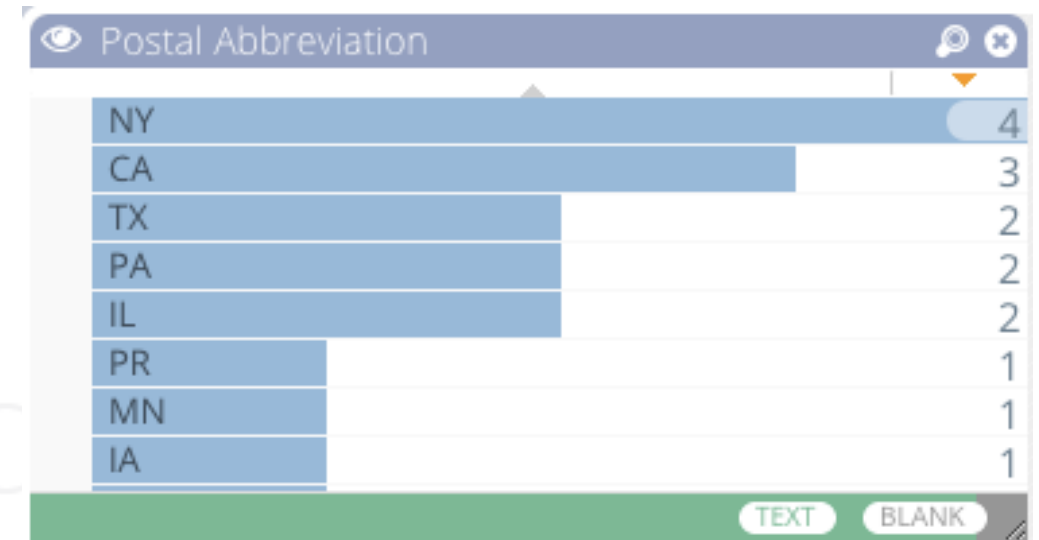
“Count the number of unique text and number values in cells A2:A10 , but do not count blank cells or text values (6)”



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

## 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

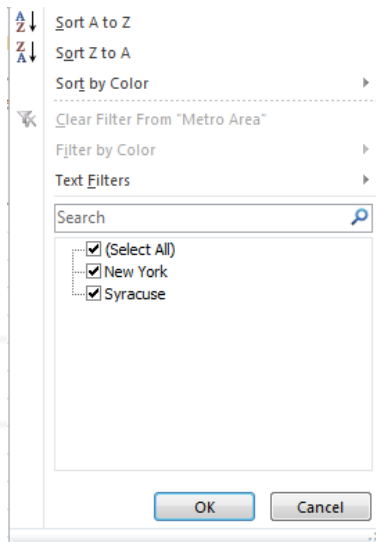
## Janitor

Uses Autofilter!

Makes multiple clicks for each column selection.

Unsure what's filtered or not.

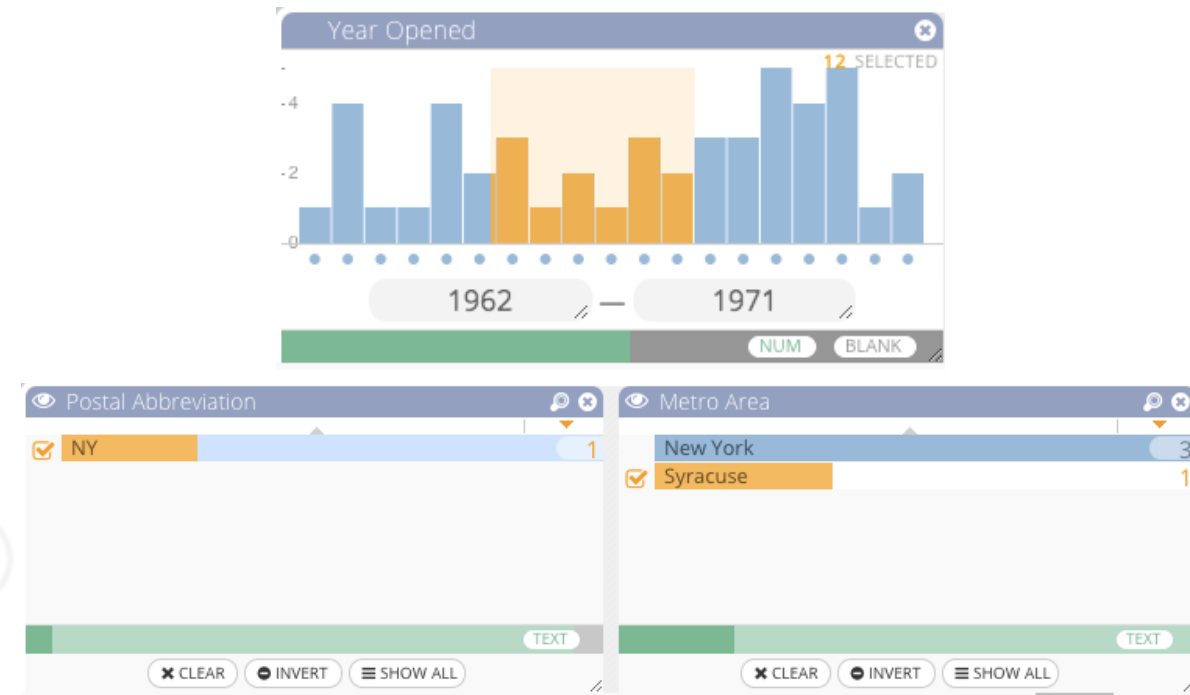
PS. No histograms.



## 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.

Cluster and Edit

Algorithm: Metaphone 0 / 3 clusters selected

MERGE? ALL <input type="checkbox"/>	CLUSTER SIZE	ROW COUNT	VALUES IN CLUSTER	NEW VALUE
<input type="checkbox"/>	4	11	<ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Baskin Robbins 6</li><li><input checked="" type="checkbox"/> Baskin Robins 3</li><li><input checked="" type="checkbox"/> Basken Robbins 1</li><li><input checked="" type="checkbox"/> Baskin Robbinz 1</li></ul>	<input type="text" value="Baskin Robbins"/>
<input type="checkbox"/>	2	4	<ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Carvel Ice Cream 2</li><li><input checked="" type="checkbox"/> Carvel Ice Cream &amp; Bakery 2</li></ul>	<input type="text" value="Carvel Ice Cream"/>
<input type="checkbox"/>	2	2	<ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Haagen Dazs 1</li><li><input checked="" type="checkbox"/> Haagen-Dazs 1</li></ul>	<input type="text" value="Haagen Dazs"/>

First Previous **1** Next Last Page Size: 25 50 100

# Use Case 6: Determine who did what when

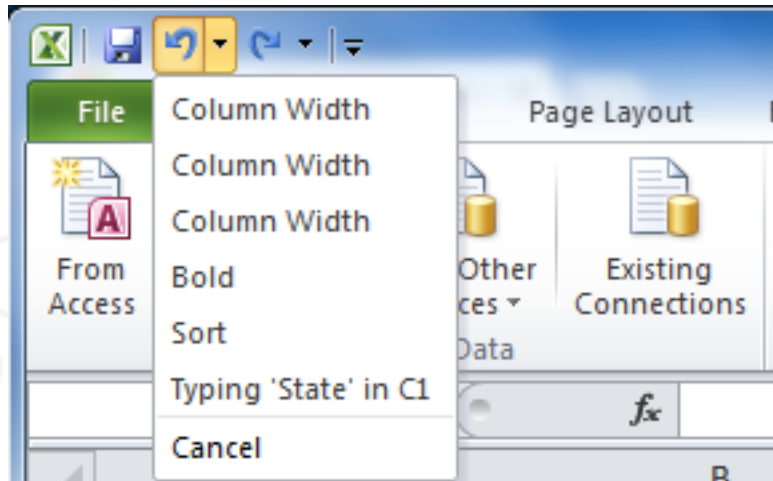
## Janitor

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.

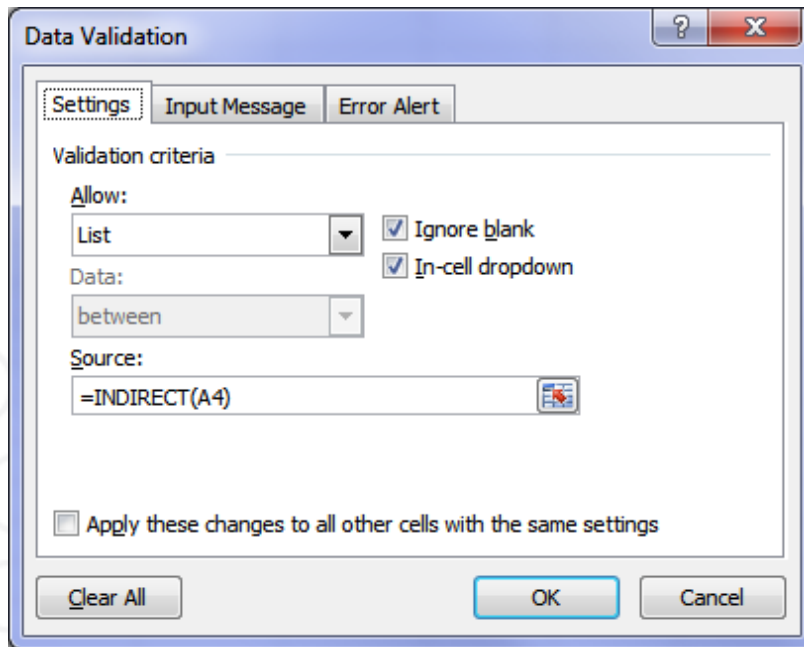
VERSION	OPERATION
8	<b>Edit cell in Postal Abbreviation</b> BY LINDSAY HAINES @ 39 MINUTES AGO
7	<b>Rename column Location_2 to Postal Abbreviation</b> BY LINDSAY HAINES @ ABOUT AN HOUR AGO
6	<b>Split column Location</b> BY LINDSAY HAINES @ ABOUT AN HOUR AGO
5	<b>Edit cell in Retail space Sq. feet/Sq. feet/(m²)</b> BY LINDSAY HAINES @ ABOUT 2 HOURS AGO
4	<b>Edit cell in Retail space Sq. feet/Sq. feet/(m²)</b> BY LINDSAY HAINES @ ABOUT 2 HOURS AGO
3	<b>Import of US Top Malls</b> BY LINDSAY HAINES @ ABOUT 2 HOURS AGO
2	<b>Import of US Top Malls</b> BY LINDSAY HAINES @ ABOUT 2 HOURS AGO
1	<b>Rename dataset New Dataset to US Top Malls</b> BY LINDSAY HAINES @ ABOUT 2 HOURS AGO

# Use Case 7: Validate values in a column

## Janitor

Copies the correct values from trusted source into your workbook.

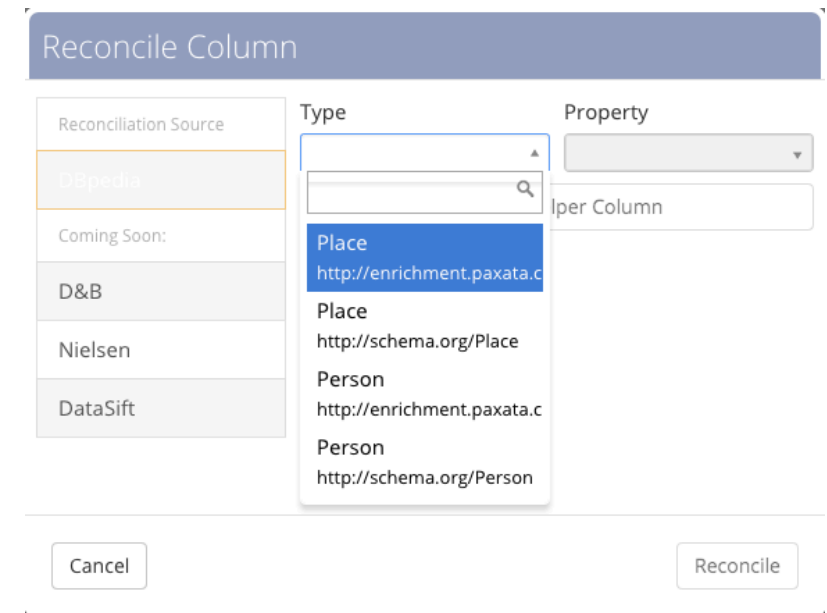
Writes conditional formulas or use the dependent data validation function.



## 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

## 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!

Enrich Column

Select properties to add

Filter Results

totalareaus	http://enrichment.paxata.com/property/totalareaus
pcwater	http://enrichment.paxata.com/property/pcwater
poprank	http://enrichment.paxata.com/property/poprank
densityrank	http://enrichment.paxata.com/property/densityrank
widthus	http://enrichment.paxata.com/property/widthus

Cancel Enrich

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

# Use Case 9: Fill in blank values

## Janitor

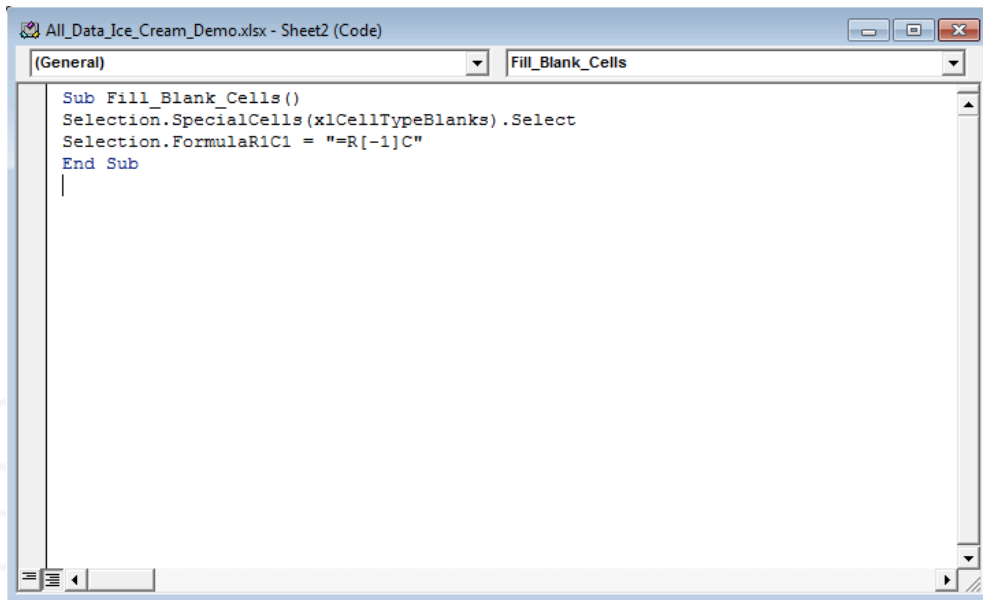
Chooses the column with the blank values

Searches for VBA reference

Clicks Developer > Visual Basic,

Clicks Insert > Module

Inputs the following code into the Module:

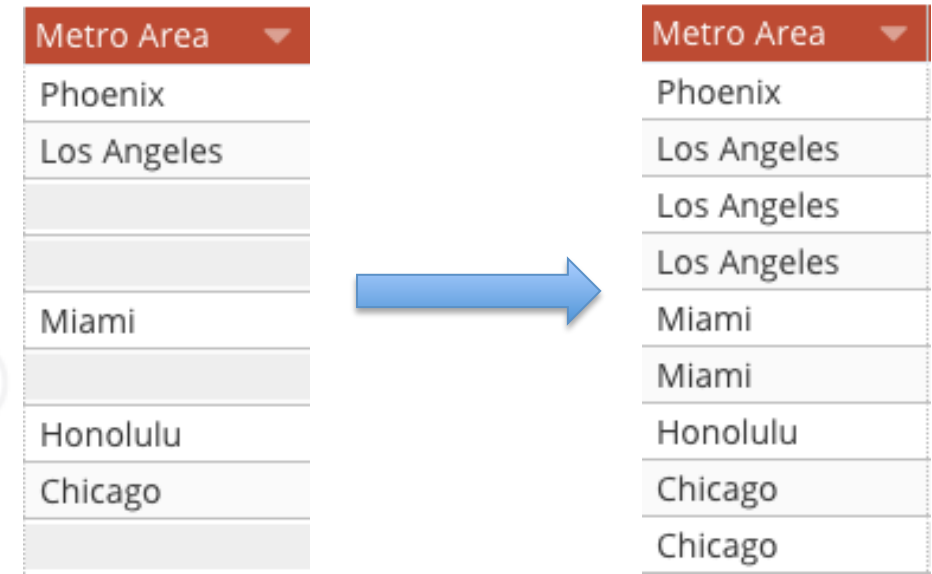


```
Sub Fill_Blank_Cells()  
Selection.SpecialCells(xlCellTypeBlanks).Select  
Selection.FormulaR1C1 = "=R[-1]C"  
End Sub
```

## Scientist

Chooses the column with the blank values

Clicks the “filldown” column operation.



# Use Case 10: Combine data sets

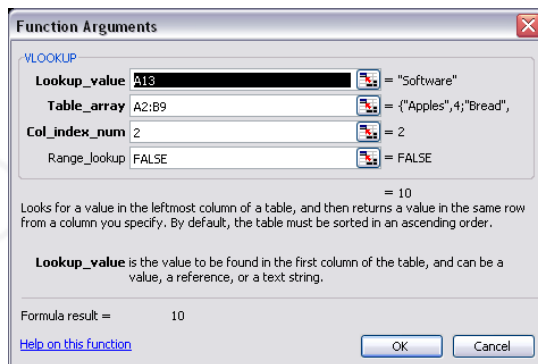
## Janitor

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

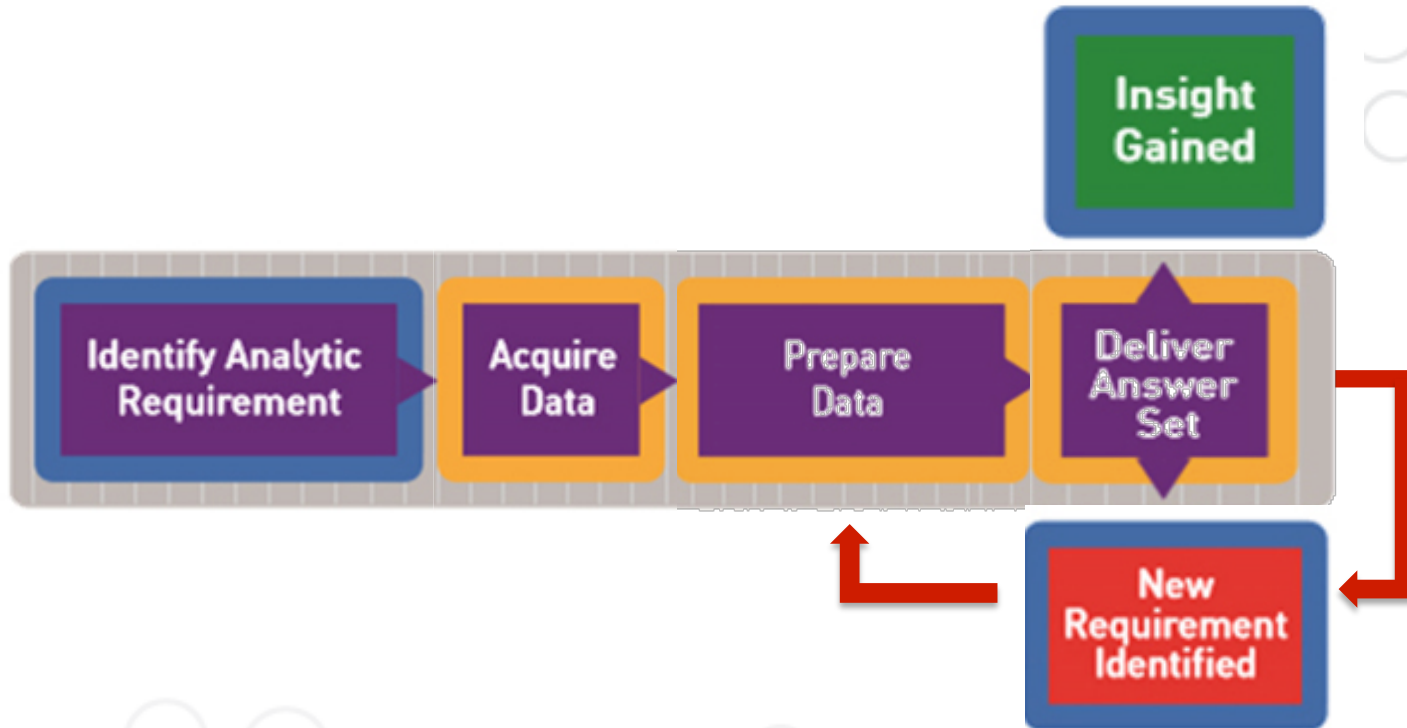
Chooses the columns he wants in his answerset.  
Lets the intelligent combination wizard tell him how to fit the data together, what the cardinality is, and what the data will look like when combined.

The image shows the Paxata data combination interface. It displays a table with columns from two different sources: 'US TOP MALLS' and 'ICE CREAM SHOPS PER MALL'. The table is titled 'COMBINING Ice Cream Shops per Mall' and shows a 100% match rate. The table has columns for 'Retail space Sq. feet/Sq. feet', 'Year Opened', 'Mall', 'Ice Cream Shop', and 'Type'. The data is sorted by 'Year Opened' in descending order.

US TOP MALLS	US TOP MALLS	ICE CREAM SHOPS PER MALL	ICE CREAM SHOPS PER MALL
Retail space Sq. feet/Sq. feet	Year Opened	Ice Cream Shop	Type
1: (200,000 m2)(17)	1959	Ala Moana C...	Sundaes & Cones
2: (250,000 m2)(4)	1983	Aventura Mall	Carvel Ice Cream & Bakery
3: (250,000 m2)(4)	1983	Aventura Mall	Bitter + Sweet
4: (250,000 m2)(4)	1983	Aventura Mall	Baskin Robbins
5: (230,000 m2)(6)	1975	Del Amo Fas...	Baskin Robbins
6: (230,000 m2)(6)	1975	Del Amo Fas...	Orange Tree
7: (230,000 m2)(6)	1975	Del Amo Fas...	Carvel Ice Cream
8: (228,000 m2)(7)	1990	Destiny USA	Ben & Jerry's
9: (228,000 m2)(7)	1990	Destiny USA	Baskin Robbins
10: (228,000 m2)(7)	1990	Destiny USA	Real Ice Cream
11: (190,000 m2)	2004	Jordan Cree...	Piccinotto Italian Ice Cream
12: (259,500 m2)(2)	1963	King of Prus...	Nirvanaah
13: (259,500 m2)(2)	1963	King of Prus...	Cold Stone Creamery
14: (259,500 m2)(2)	1963	King of Prus...	Baskin Robbins
15: (194,419 m2)(18)	1951	Lakewood C...	Shake Shack
16: (258,200.0 m2)(3)	1992	Mall of Ame...	Orange Tree
17: (258,200.0 m2)(3)	1992	Mall of Ame...	Carvel Ice Cream & Bakery
18: (258,200.0 m2)(3)	1992	Mall of Ame...	Baskin Robbins
19: (258,200.0 m2)(3)	1992	Mall of Ame...	Real Ice Cream
20: (190,000 m2)(2)(22)	1965	NorthPark C...	Nirvanaah

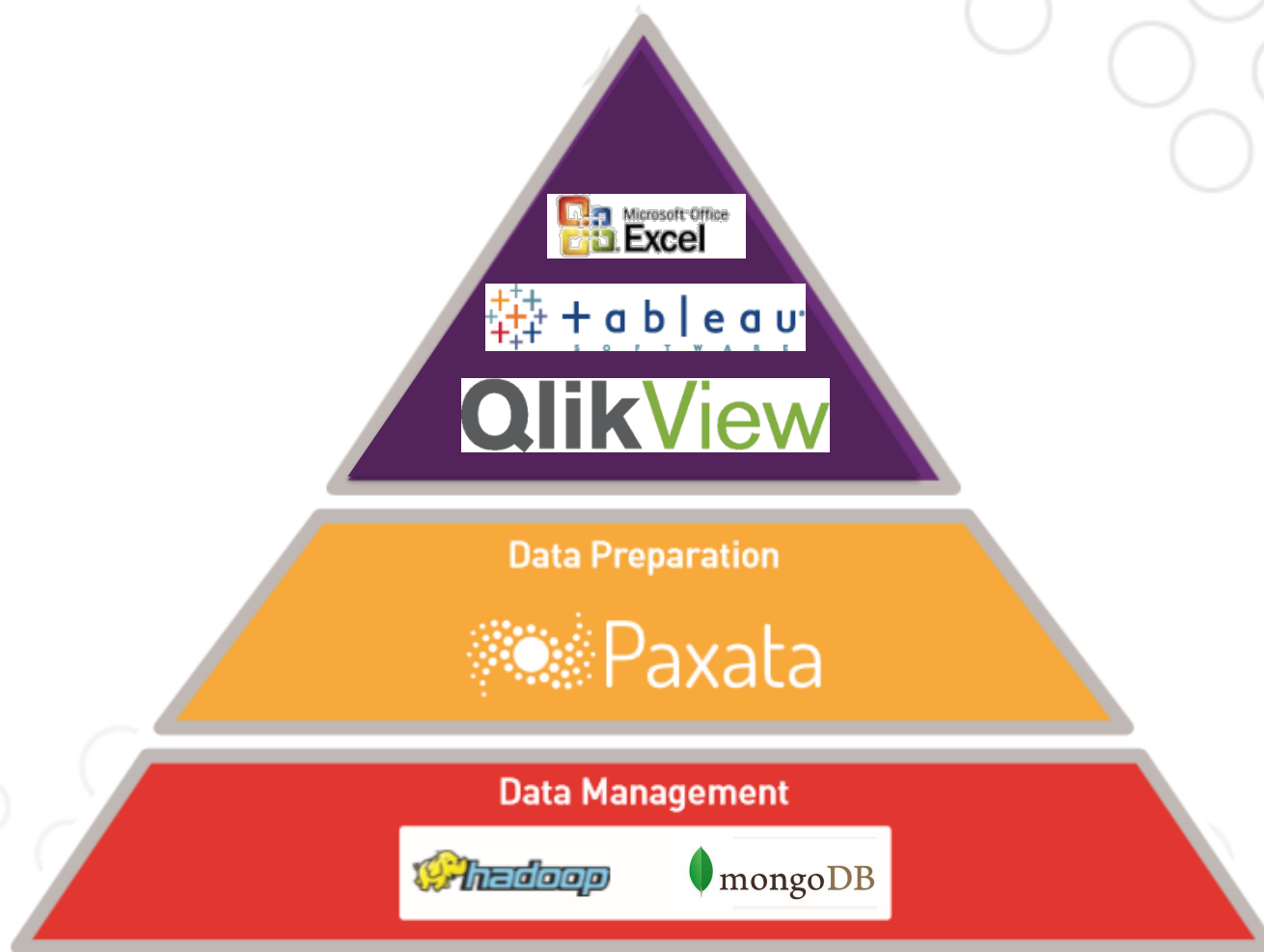
# With Paxata

Time to decision



- 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



# YOU DON'T HAVE TO BE A JANITOR ANYMORE!!!

The screenshot displays the Paxata software interface. At the top, there are navigation links: workspace, admin, help, feedback, logout. Below this, there are three project cards: 'Sales Forecasting', 'Marketing Segmentation', and 'Financial Plan-Actual', each with creation and update timestamps and completion status. A 'Distributors' project is also visible. On the left, a sidebar lists 'All Projects' with categories like Distributors, Marketing Segmentation, Sales Forecasting, and Financial Plan-Actual. The main area shows a data table with columns for CompanyName\_1, CompanyName\_2, State, and ShiptoID. A 'Joining ProductList.xml' window is open, showing a 99% match rate between Retail Sales DC.xml and ProductList.xml. A large 'Paxata' logo is overlaid at the bottom left of the screenshot.

The screenshot shows a report for 'Green Bunny Ice Cream'. It features a map of the United States with a callout for 'Average Ice Cream Consumption' showing a bar chart with 'Below' and 'Above' indicators. A 'Drill Down' button is shown. Below the map, a text box states: 'State Average Ice Cream Consumption is 1-25% Below Average'. Another text box indicates 'Other Brand Ice Cream Stores are 20 across 7 Malls'. At the bottom, there are three columns of data: 'Location', 'Mall', and 'Ice Cream Shop', each with a search icon. The 'Consumption' column shows a bar chart for 'RED BULL' with a value of 135%, marked as 'Below average'.

[www.paxata.com/schedule-a-demo](http://www.paxata.com/schedule-a-demo)



# THANKS!

[www.paxata.com](http://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



\$49B High Tech  
Manufacturer



\$29B  
Financial  
Services Firm



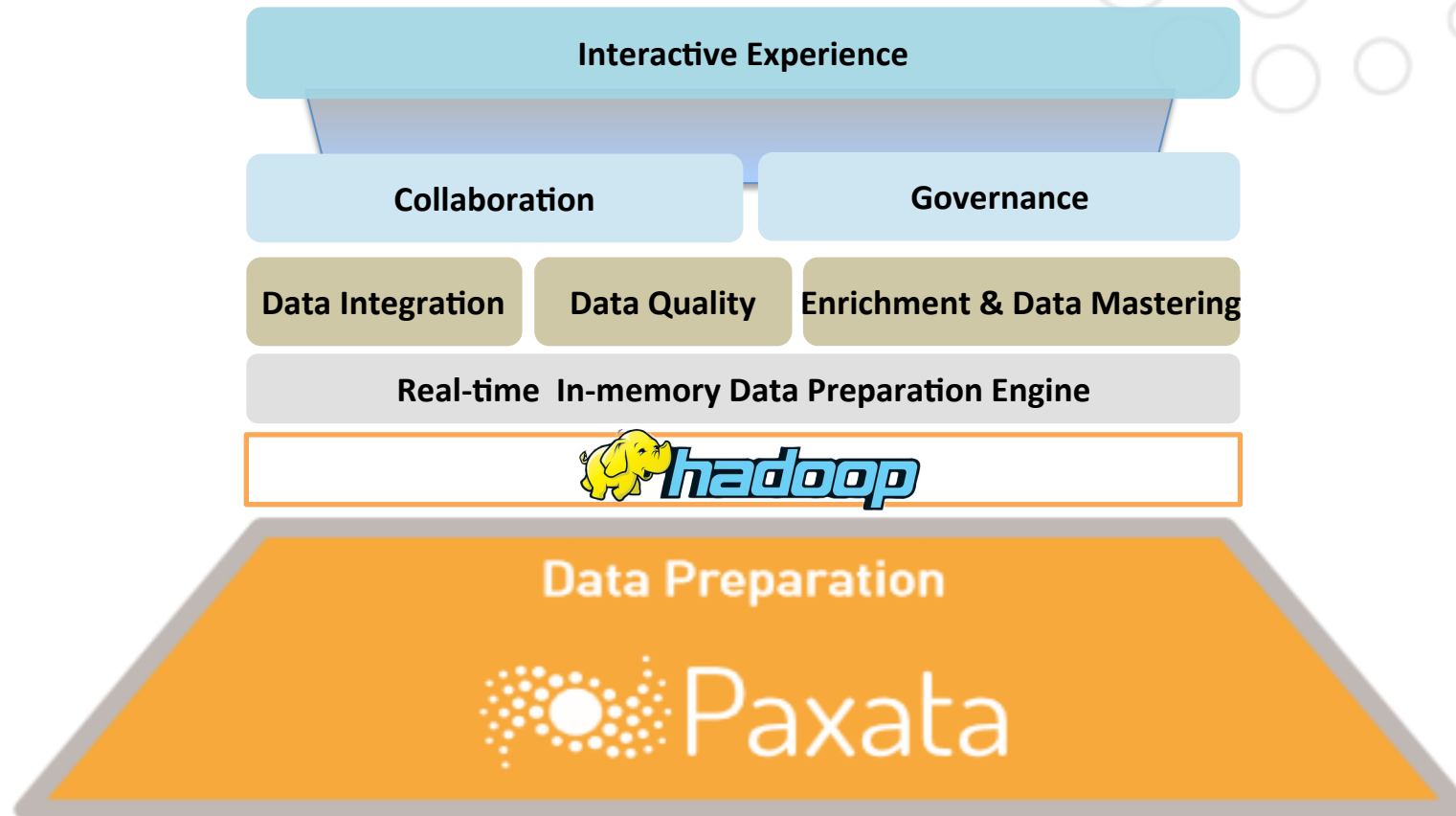
**QlikView**



**able** SOFTWARE

**cloudera**

# The Adaptive Data Preparation Platform



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