



# Data Wrangling

John Meehan

Jeff Rasley

# Working with raw data sucks.

- Data comes in all shapes and sizes
  - CSV files, PDFs, stone tablets, .jpg...
- Different files have different formatting
  - Spaces instead of NULLs, extra rows
- “Dirty” data
  - Unwanted anomalies
  - Duplicates

# Current Tools

- Focus on specific problems
  - Resolving entities
  - Removing duplicates
  - Schema matching
- Most systems are non-interactive
  - Inaccessible to general audience
- A lot of people just use Excel or regular expressions...

# Data Wrangling

- Goal: extract and standardize the raw data
  - Combine multiple data sources
  - Clean data anomalies
- Combine automation with interactive visualizations to aid in cleaning
- Improve efficiency and scale of data importing
- Lower the threshold for broader audiences

## A typical

Three Data Sources:  
Database, PDF, CSV

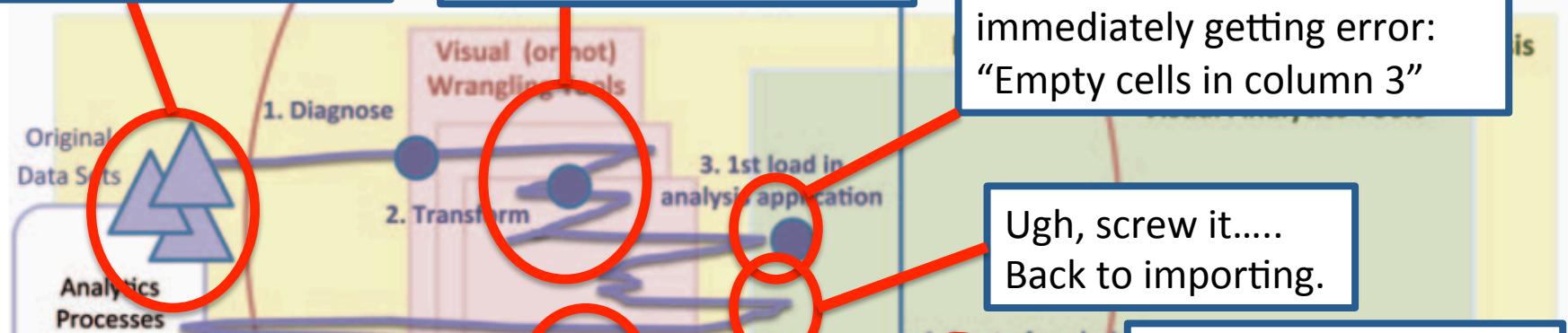
## Wrangling

- Missing headers,
- Multiple date formats,
- Merged columns

## Analysis

- Visualize
- Analyze

Success! .....but  
immediately getting error:  
“Empty cells in column 3”



Data reimported,  
More cleaning

Analysis possible, but  
data quality still sucks

SUCCESS!!!!

Repeat data loading  
less painful, but still  
annoying

usable data  
Triangle icon  
Trail of data  
transformations

usable data + findings  
Triangle icon and star icon  
Trail of analysis =  
Insight Provenance

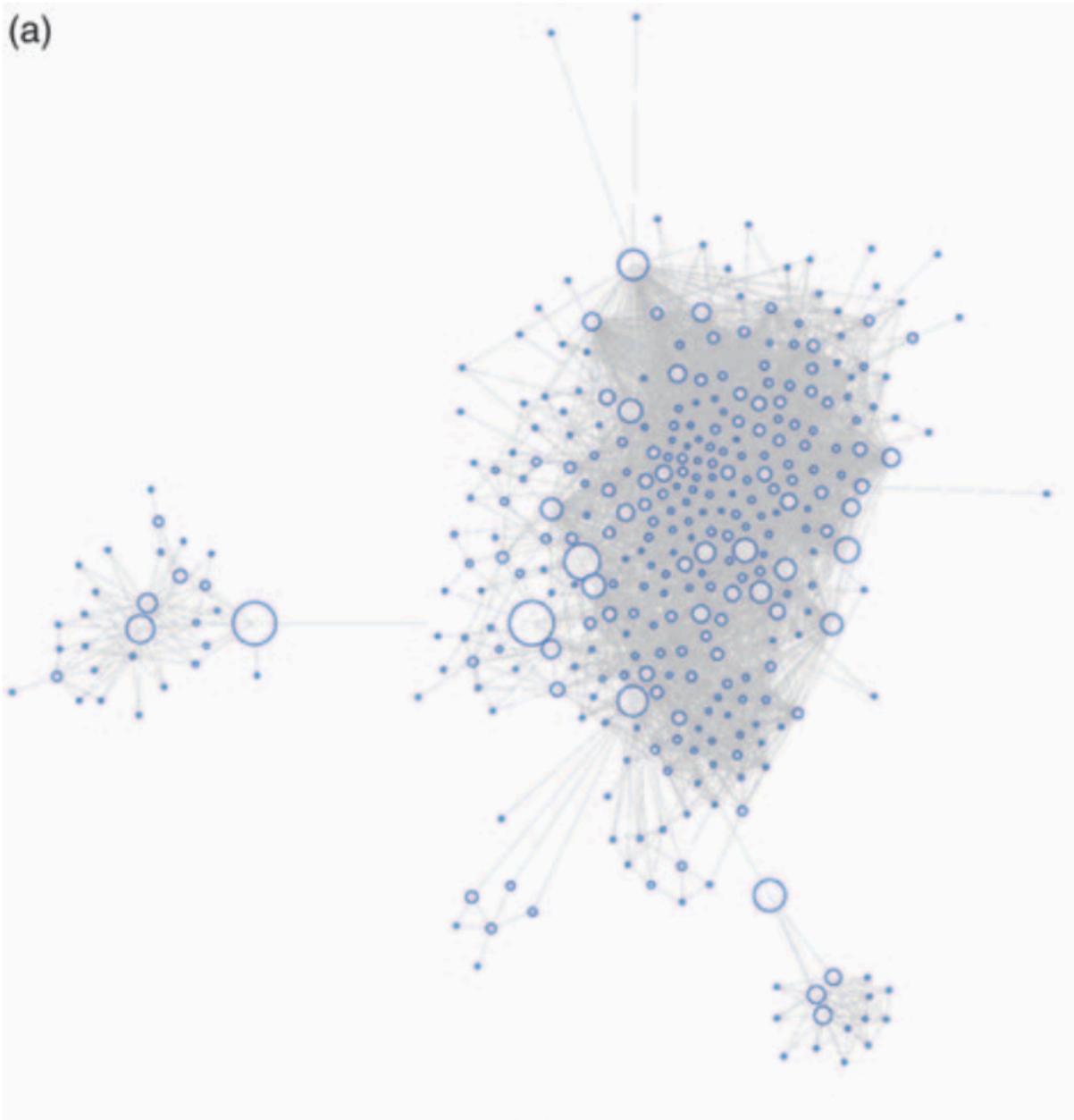
# Types of Data Problems

- Missing data
- Incorrect data
- Inconsistent representations of the same data
- About 75% of data problems require human intervention
- Cleaning data vs overly-sanitizing data

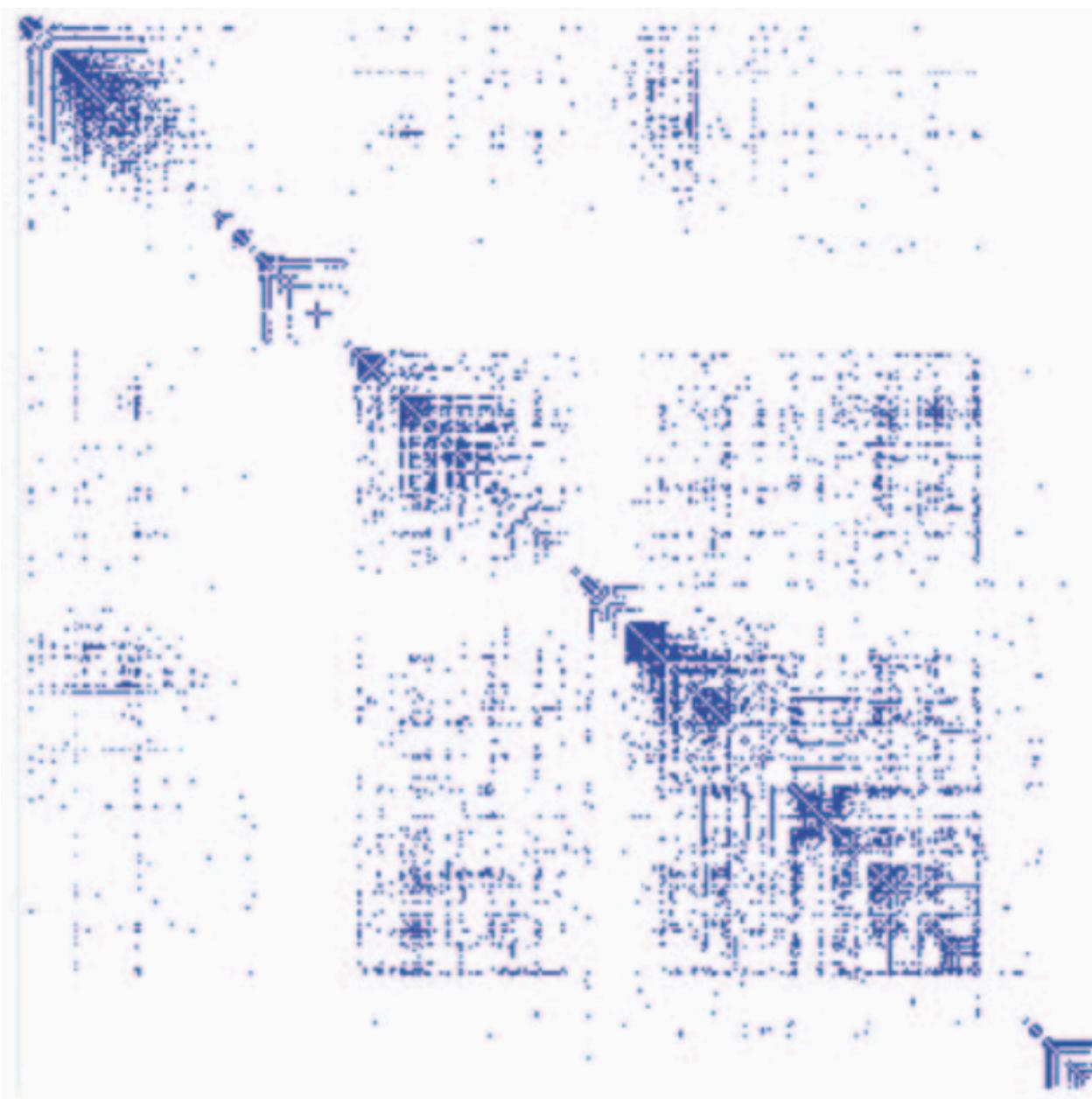
# Diagnosing data problems

- Visualizations can convey “raw” data
- Different visual representations highlight different types of data issues
  - Outliers often stand out in a plot
  - Missing data will cause gap or zero value
- Becomes increasingly difficult as data gets larger
  - Visual design coupled with interaction
  - Sampling

## Node-Link Diagram

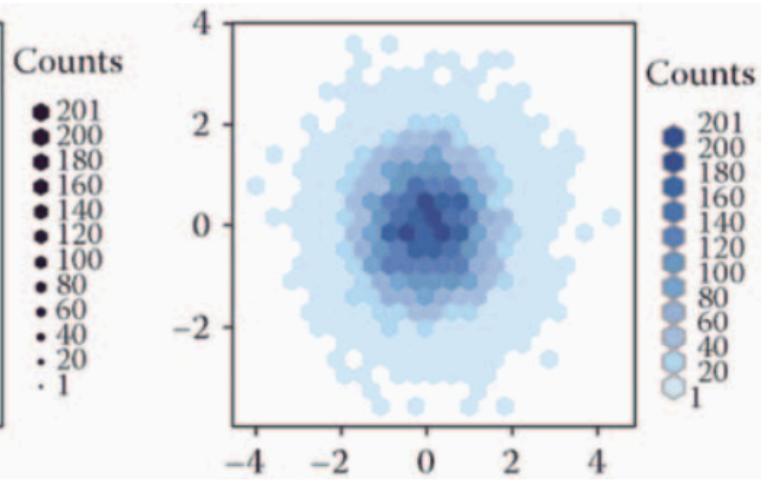
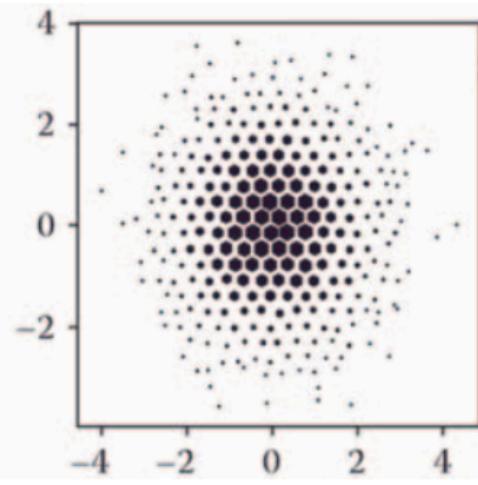
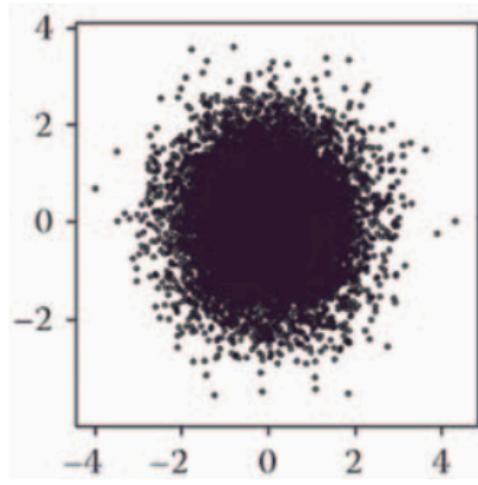


## Matrix View



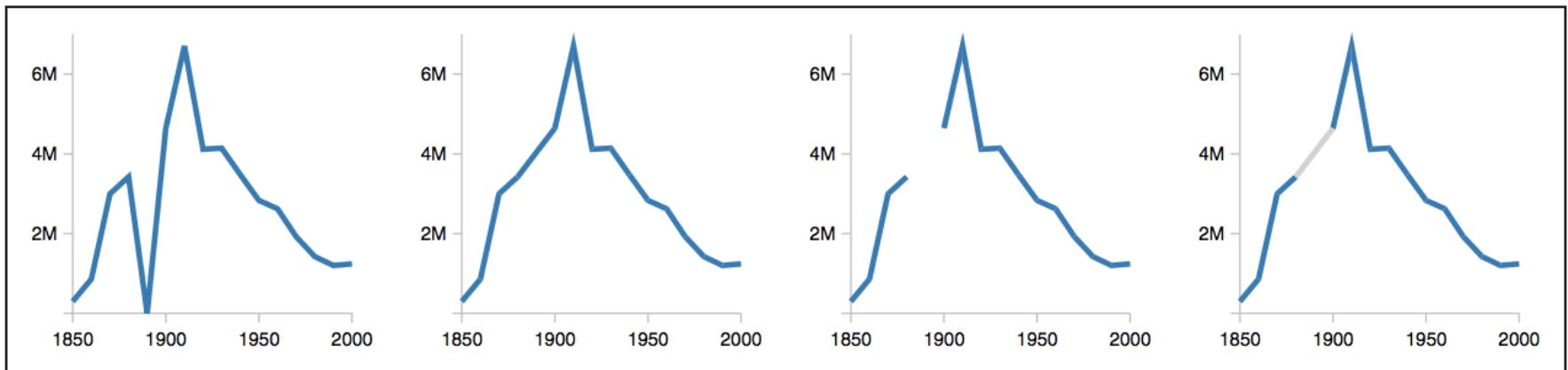
## Sorted Matrix View

(c)



# Visualizing Missing Data

- Set values to zero?
- Interpolate based on existing data?
- Omit missing data?



# Visualizing Uncertain Data

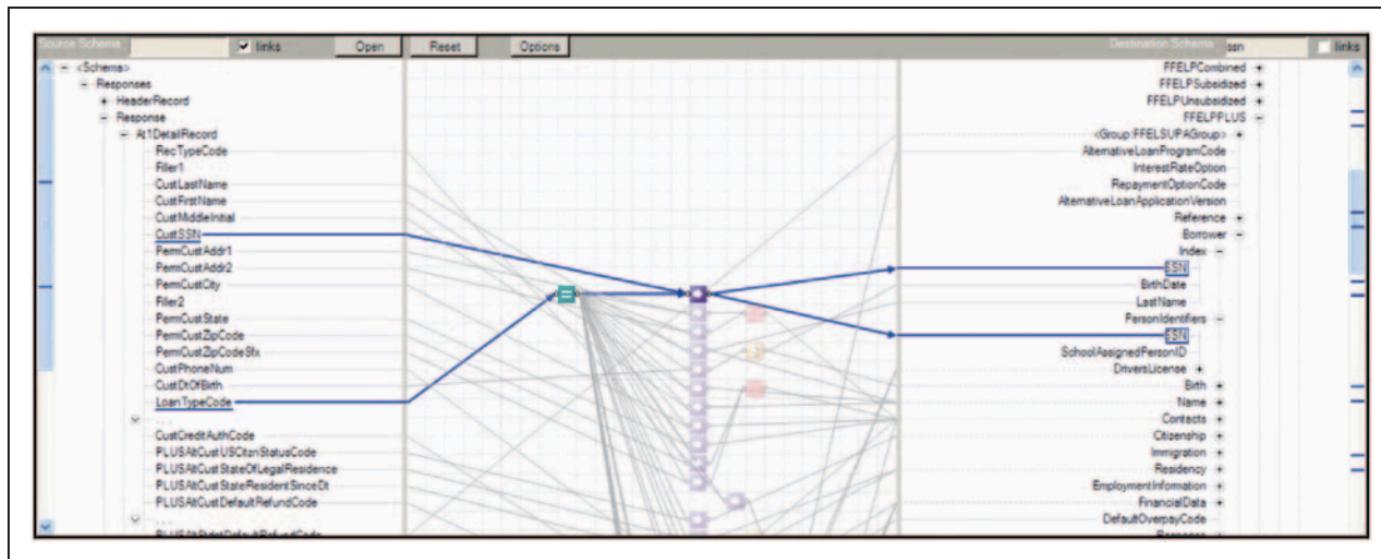
- Can arise from:
  - Measurement errors
  - Missing data
  - Sampling
- Visualization must
  - Consider all components of uncertainty
  - Depict multiple kinds of uncertainty
  - Interact with uncertainty depictions

# Transforming Data

- Splitting columns, converting into meaningful records
- Typical methods: regular expressions, programming by demonstration
  - Prone to errors, tedious
- Interactive tools simplify the process
  - Guide user through setting automated constraints
  - Generates scripts for the user

# Transforming Data (cont)

- Data formatting, extraction, and conversion
- Correcting erroneous values
- Integrating multiple data sets



# Editing and Auditing Transformations

- Data Provenance
  - Maintaining the data history
  - Track the lineage of a specific item's origins
- Used for the modification, reuse, and understanding of a transformation
- What transformation language should be used?
  - Extend existing languages?

# Wrangling in the Cloud

- Allows the sharing of data transformations
- Mining records of wrangling
  - Better automatic suggestions
- User-defined data types
- Feedback from downstream analysts
  - Crowdsourcing the final result
  - Allow users to annotate or correct the data

# Checkpoint-Conclusion

- Data wrangling is often a second-class citizen
- Common problems, very time-consuming
  - Manual approach is no longer a viable option
- Future work: extend visual approaches into the data wrangling phase
- Plenty of research directions

# Secure Data Analytics

- Private/sensitive data
  - SSN, Medical, Classified, etc.
- Cloud-base analysis currently doesn't work
- Local analysis is key
- Research area?

# Potter's Wheel: An Interactive Data Cleaning System

- Vijayshankar Raman and Joseph Hellerstein (VLDB '01)
- Provide a graphical interactive tool to support various data transformations with suggestions

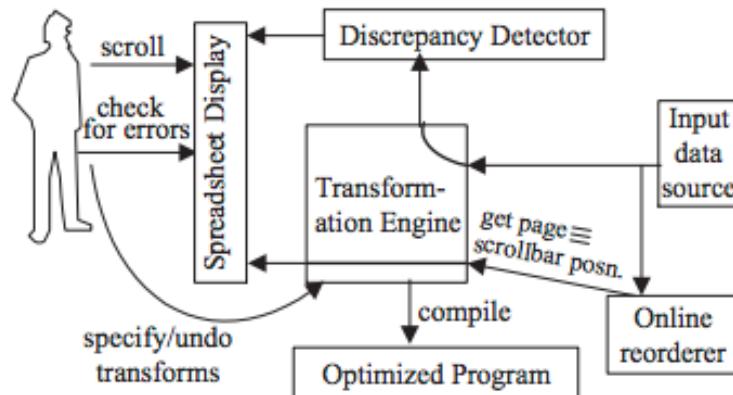


Figure 2: Potter's Wheel Architecture

# Transformations

Transform		Definition
Format	$\phi(R, i, f)$	$\{(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n, f(a_i)) \mid (a_1, \dots, a_n) \in R\}$
Add	$\alpha(R, x)$	$\{(a_1, \dots, a_n, x) \mid (a_1, \dots, a_n) \in R\}$
Drop	$\pi(R, i)$	$\{(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n) \mid (a_1, \dots, a_n) \in R\}$
Copy	$\kappa((a_1, \dots, a_n), i)$	$\{(a_1, \dots, a_n, a_i) \mid (a_1, \dots, a_n) \in R\}$
Merge	$\mu((a_1, \dots, a_n), i, j, \text{glue})$	$\{(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_{j-1}, a_{j+1}, \dots, a_n, a_i \oplus \text{glue} \oplus a_j) \mid (a_1, \dots, a_n) \in R\}$
Split	$\omega((a_1, \dots, a_n), i, \text{splitter})$	$\{(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n, \text{left}(a_i, \text{splitter}), \text{right}(a_i, \text{splitter})) \mid (a_1, \dots, a_n) \in R\}$
Divide	$\delta((a_1, \dots, a_n), i, \text{pred})$	$\{(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n, a_i, \text{null}) \mid (a_1, \dots, a_n) \in R \wedge \text{pred}(a_i)\} \cup$ $\{(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n, \text{null}, a_i) \mid (a_1, \dots, a_n) \in R \wedge \neg \text{pred}(a_i)\}$
Fold	$\lambda(R, i_1, i_2, \dots, i_k)$	$\{(a_1, \dots, a_{i_1-1}, a_{i_1+1}, \dots, a_{i_2-1}, a_{i_2+1}, \dots, a_{i_k-1}, a_{i_k+1}, \dots, a_n, a_{i_l}) \mid$ $(a_1, \dots, a_n) \in R \wedge 1 \leq l \leq k\}$
Select	$\sigma(R, \text{pred})$	$\{(a_1, \dots, a_n) \mid (a_1, \dots, a_n) \in R \wedge \text{pred}((a_1, \dots, a_n))\}$

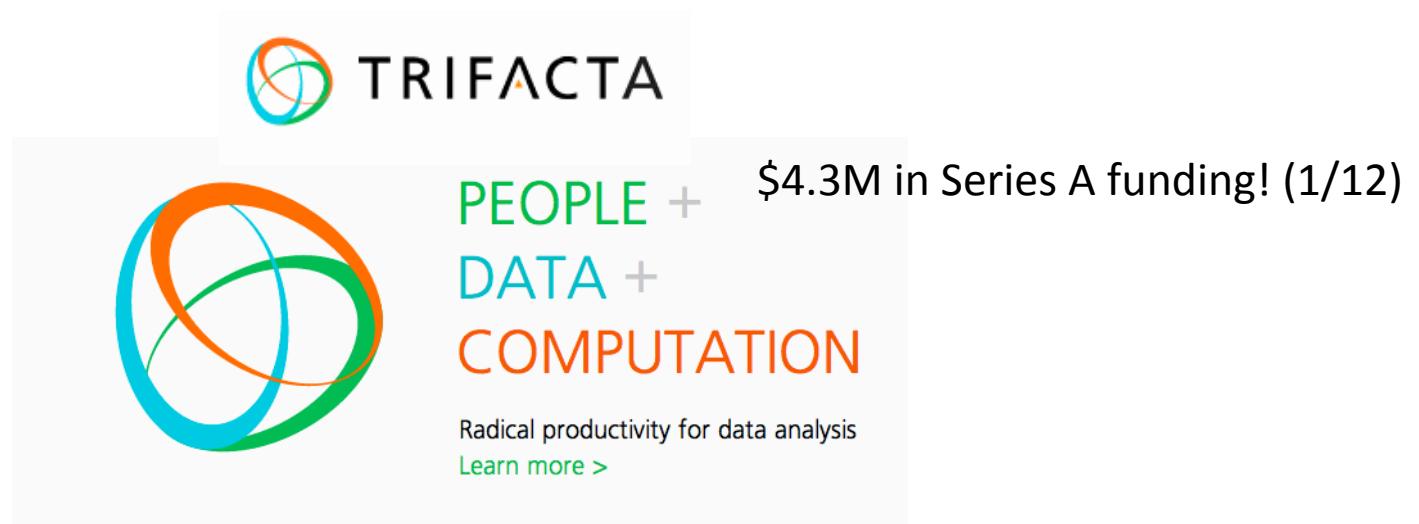
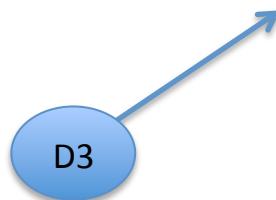
**Notation:**  $R$  is a relation with  $n$  columns.  $i, j$  are column indices and  $a_i$  represents the value of a column in a row.  $x$  and  $\text{glue}$  are values.  $f$  is a function mapping values to values.  $x \oplus y$  concatenates  $x$  and  $y$ .  $\text{splitter}$  is a position in a string or a regular expression,  $\text{left}(x, \text{splitter})$  is the left part of  $x$  after splitting by  $\text{splitter}$ .  $\text{pred}$  is a function returning a boolean.

# Download!

- Potter's Wheel A-B-C: An Interactive Tool for Data Analysis, Cleansing, and Transformation
  - <http://control.cs.berkeley.edu/abc/>
- (But you probably don't want to, last release was Oct 10, 2000)

# Data Wrangler

- Wrangler: Interactive Visual Specification of Data Transformation Scripts (CHI '11)
  - Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer (Stanford Vis Group + Berkeley)



# What's new?

- Similar goals as Potter's Wheel
- Improved UI for the web
- Python & JavaScript libraries
- Additional transformations such as fill

# Demo

(~10min)

Transform Script		Import	Export
▶ Split data repeatedly on <code>newline</code> into rows			
▶ Split <code>split</code> repeatedly on <code>;</code>			
▶ Promote row 0 to header			
Text	Columns	Rows	Table
<a href="#">Clear</a>			
<a href="#">Delete row 7</a>			
<a href="#">Delete empty rows</a>			
<a href="#">Fill row 7 by copying values from above</a>			

#	Year	Property_crime_rate
0	Reported crime in Alabama	
1		
2	2004	4029.3
3	2005	3900
4	2006	3937
5	2007	3974.9
6	2008	4081.9
7		
8	Reported crime in Alaska	
9		
10	2004	3370.9
11	2005	3615
12	2006	3582

# 7 Command-Line Tools for Data Science

- [http://jeroenjanssens.com/2013/09/19/  
seven-command-line-tools-for-data-  
science.html](http://jeroenjanssens.com/2013/09/19/seven-command-line-tools-for-data-science.html)