

# Cognitive Dimensions of Between-Table Context Support in Wrangling Applications

Steve Kasica

Dec. 03, 2019

# Why am I doing the thing that I'm doing?

- More journalists are using the tools and techniques of data science under the term data journalism.
- Little is understood about the unique issues of this group as they related to cleaning, transforming, and otherwise wrangling their data.
- But there's a lot of open-source and commercial wrangling applications available to journalists.
- However, do these interfaces support the kind of tasks and data that journalists actually do in the wild?
- This is an analysis project in the course

# What is the thing that I'm doing?

- Replicate the wrangling workflows done by real journalists working in a programming environment with these GUI-based tools
- Discuss the trade-offs that exist between dimensions in the wrangling activity.
- Compare and contrast the strengths and weaknesses of these two tools in the wrangling of journalistic data.

# Prior Work

- Over the summer, I conducted an artifact-mediated indirect observational study of data wrangling in journalism.
- Identified high-level wrangling actions done by journalists
- Also identified exemplar data and wrangling sequences.

# Workflows and Tools

- Reproduce each wrangling workflow with both tools
- Workflows are abstracted to not a sequence of steps because that would be trivial,
- Workflows are a sequence of intermediate table forms to reproduce
- One workflow-tool combination may include may branches for getting to different table states as there are different means to the same end.

# Workflows to reproduce

- Longterm managed care records in New York
  - Sarah Cohen's CAR 2016 tutorial on data cleaning with OpenRefine
  - Performs the following tasks: extract data from column, remove non-data rows, remove rows that contain notes, remove bad-data rows, remove rows with missing values, aggregate join, resolve entity names
- Water usage over time in California
  - Wrangling performed by Ben Welsh at *Los Angeles Times*
  - Performs the following tasks: configure analysis tools, subset raw data to relevant, string-ify date, filter data, remove rows

# Applications Considered

- There are many applications for wrangling: OpenRefine, Cloud Dataprep, Tableau Prep, Trifacta Wrangler, Workbench
- Focus on OpenRefine and Cloud Dataprep
  - Were recommended by a in the MOOC Data Journalism and Visualization with Free Tools offered from Knight Center for Journalism in the Americas

# OpenRefine

**OpenRefine** all longterm managed care [Permalink](#) Open... Export Help

Facet / Filter Undo / Redo 25 / 25 **3769 rows** Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last »

**Using facets and filters**

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?  
[Watch these screencasts](#)

All	plan name	plan type	report date	county	enrollment
1.	COMPREHENSIVE CARE MGMT CORP	PACE	2009-01-01T00:00:00Z	NEW YORK	2184
2.	COMPREHENSIVE CARE MGMT CORP	PACE	2009-01-01T00:00:00Z	WESTCHESTER	170
3.	EDDY SENIOR CARE	PACE	2009-01-01T00:00:00Z	ALBANY	1
4.	EDDY SENIOR CARE	PACE	2009-01-01T00:00:00Z	SCHENECTADY	108
5.	EDDY SENIOR CARE	PACE	2009-01-01T00:00:00Z	SCHOHAR	1
6.	INDEPENDENT LIVING FOR SENIORS	PACE	2009-01-01T00:00:00Z	MONROE	246
7.	PACE CNY	PACE	2009-01-01T00:00:00Z	CHAUTAUQUA	1
8.	PACE CNY	PACE	2009-01-01T00:00:00Z	ONONDAGA	317
9.	AMERGROUP	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	431
10.	CCM SELECT	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	1040
11.	CCM SELECT	PARTIAL CAPITATION	2009-01-01T00:00:00Z	WESTCHESTER	17
12.	ELANT	PARTIAL CAPITATION	2009-01-01T00:00:00Z	ORANGE	104
13.	ELANT	PARTIAL CAPITATION	2009-01-01T00:00:00Z	ROCKLAND	34
14.	GUILDNET	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NASSAU	519
15.	GUILDNET	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	5810
16.	GUILDNET	PARTIAL CAPITATION	2009-01-01T00:00:00Z	SUFFOLK	217
17.	HHH CHOICES	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	706
18.	HOMEFRST	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	2876
19.	INDEPENDENT CARE SYSTEMS	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	1331
20.	PARTNERS IN COMMUNITY CARE	PARTIAL CAPITATION	2009-01-01T00:00:00Z	ORANGE	157
21.	PARTNERS IN COMMUNITY CARE	PARTIAL CAPITATION	2009-01-01T00:00:00Z	ROCKLAND	91
22.	SENIOR HEALTH PARTNERS INC	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	1375
23.	SENIOR NETWORK HEALTH	PARTIAL CAPITATION	2009-01-01T00:00:00Z	HERKIMER	36
24.	SENIOR NETWORK HEALTH	PARTIAL CAPITATION	2009-01-01T00:00:00Z	ONEIDA	319
25.	TOTAL AGING IN PLACE PROGRAM	PARTIAL CAPITATION	2009-01-01T00:00:00Z	ERE	142
26.	VNS CHOICE	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	6783
27.	WELLCARE	PARTIAL CAPITATION	2009-01-01T00:00:00Z	NEW YORK	158
28.	ARCHCARE SENIOR LIFE	PACE	2010-01-01T00:00:00Z	NEW YORK	10
29.	CHS BUFFALO LIFE	PACE	2010-01-01T00:00:00Z	ERE	9
30.	COMPREHENSIVE CARE MGMT	PACE	2010-01-01T00:00:00Z	NASSAU	8
31.	COMPREHENSIVE CARE MGMT	PACE	2010-01-01T00:00:00Z	NEW YORK	2285
32.	COMPREHENSIVE CARE MGMT	PACE	2010-01-01T00:00:00Z	SUFFOLK	61
33.	COMPREHENSIVE CARE MGMT	PACE	2010-01-01T00:00:00Z	WESTCHESTER	175
34.	EDDY SENIOR CARE	PACE	2010-01-01T00:00:00Z	SCHENECTADY	101
35.	INDEPENDENT LIVING FOR SENIORS	PACE	2010-01-01T00:00:00Z	MONROE	264
36.	PACE CNY	PACE	2010-01-01T00:00:00Z	CHAUTAUQUA	1
37.	PACE CNY	PACE	2010-01-01T00:00:00Z	ONONDAGA	341
38.	TOTAL SENIOR CARE	PACE	2010-01-01T00:00:00Z	CATTARAUGUS	26
39.	AMERGROUP	PARTIAL CAPITATION	2010-01-01T00:00:00Z	NEW YORK	702



# Google Cloud Dataprep

MANAGED LONG TERM CARE FLOW

Managed Long Term Care - 3

Full Data

Run Job

45 Categories

2 Categories

Jan 2009 - Apr 2012

45 Categories

0-24.21k

rec	plan name	rec	plan type	report date	rec	county	#	enrollment
-	COMPREHENSIVE CARE MGT	PACE		JANUARY 2009	NEW YORK			2184
-	COMPREHENSIVE CARE MGT	PACE		JANUARY 2009	WESTCHESTER			170
-	EDDY SENIOR CARE	PACE		JANUARY 2009	ALBANY			1
-	EDDY SENIOR CARE	PACE		JANUARY 2009	SCHENECTADY			108
-	EDDY SENIOR CARE	PACE		JANUARY 2009	SCHOHARIE			1
-	INDEPENDENT LIVING FOR SENIORS	PACE		JANUARY 2009	MONROE			246
-	PACE CNY	PACE		JANUARY 2009	CHAUTAUQUA			1
-	PACE CNY	PACE		JANUARY 2009	ONONDAGA			317
-	AMERIGROUP	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			431
-	CCM SELECT	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			1040
-	CCM SELECT	PARTIAL CAPITATION		JANUARY 2009	WESTCHESTER			17
-	ELANT	PARTIAL CAPITATION		JANUARY 2009	ORANGE			104
-	ELANT	PARTIAL CAPITATION		JANUARY 2009	ROCKLAND			34
-	GUILDNET	PARTIAL CAPITATION		JANUARY 2009	NASSAU			519
-	GUILDNET	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			5810
-	GUILDNET	PARTIAL CAPITATION		JANUARY 2009	SUFFOLK			217
-	HHH CHOICES	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			706
-	HOMEFIRST	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			2976
-	INDEPENDENT CARE SYSTEMS	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			1331
-	PARTNERS IN COMMUNITY CARE	PARTIAL CAPITATION		JANUARY 2009	ORANGE			157
-	PARTNERS IN COMMUNITY CARE	PARTIAL CAPITATION		JANUARY 2009	ROCKLAND			91
-	SENIOR HEALTH PARTNERS INC	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			1375
-	SENIOR NETWORK HEALTH	PARTIAL CAPITATION		JANUARY 2009	HERKIMER			36
-	SENIOR NETWORK HEALTH	PARTIAL CAPITATION		JANUARY 2009	ONEIDA			319
-	TOTAL AGING IN PLACE PROGRAM	PARTIAL CAPITATION		JANUARY 2009	ERIE			142
-	VNS CHOICE	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			6783
-	WELLCARE	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			158
-	ARCHCARE SENIOR LIFE	PACE		JANUARY 2010	NEW YORK			10
-	CHS BUFFALO LIFE	PACE		JANUARY 2010	ERIE			9
-	COMPREHENSIVE CARE MGMT	PACE		JANUARY 2010	NASSAU			8
-	COMPREHENSIVE CARE MGMT	PACE		JANUARY 2010	NEW YORK			2265
-	COMPREHENSIVE CARE MGMT	PACE		JANUARY 2010	SUFFOLK			61
-	COMPREHENSIVE CARE MGMT	PACE		JANUARY 2010	WESTCHESTER			175
-	EDDY SENIOR CARE	PACE		JANUARY 2010	SCHENECTADY			101

5 Columns 3,757 Rows 3 Data Types

Details

RBC county

Unique Values

Search...

CHAUTAUQUA	51
DUTCHESS	50
ULSTER	42
RENSSELAER	41
ALLEGANY	40
NIAGARA	30
COLUMBIA	27
SULLIVAN	23
MONTGOMERY	21
PUTNAM	21
SCHOHARIE	20
FULTON	13
WARREN	13
WASHINGTON	13
CAYUGA	11
CORTLAND	11
ESSEX	11
BROOME	10
TOMPKINS	10
GREENE	10
DELAWARE	9
STEUBEN	7
NEW YORK CITY TOTALS:	7
GRAND TOTALS:	7
TIOGA	5
CHENANGO	5
MADISON	4
LIVINGSTON	3
SARATOGA	2

# Cognitive Dimensions

- There are 13 different dimensions to create a common, interface-independent vocabulary to discuss usability in user interfaces
- Each interface occupies 13-dimensional space, thus improving an interface in one aspect impacts the others
- One goal of this project is to identify these tradeoffs in data wrangling interfaces in general, in addition to compare and contrasting the two tools.

# Viscosity

- “Resistance to change” [Blackwell et al, 2003]

# Visibility

- “Ability to view components easily” [Blackwell et al., 2003]
  - Can we see all components in VPL? [Blackwell et al., 2003; Green, 1996]
- In data wrangling, visibility because an issue as datasets become large
  - Is every part of the relevant data simultaneous visible?
  - In high-dimensional data you have to scroll to view all columns
  - In data with many observations, you have to scroll to view rows.
  - **Focal point:** Would increasing visibility may decrease error-proneness?
  - Visualization may help here. Charts are great at representing data compactly, a.k.a data-ink ratio [Tufte, 1983]

# Visibility in Dataprep

MANAGED LONG TERM CARE FLOW >  
 Managed Long Term Care - 3 ▾  
 Full Data

Run Job

Details

rbc	plan name	rbc	plan type	report date	rbc	county	#	enrollment
48 Categories		2 Categories		Jan 2008 - Apr 2012	45 Categories		0-24 21k	
-	COMPREHENSIVE CARE MGT	PACE		JANUARY 2009	NEW YORK			2184
-	COMPREHENSIVE CARE MGT	PACE		JANUARY 2009	WESTCHESTER			170
-	EDDY SENIOR CARE	PACE		JANUARY 2009	ALBANY			1
-	EDDY SENIOR CARE	PACE		JANUARY 2009	SCHENECTADY			108
-	EDDY SENIOR CARE	PACE		JANUARY 2009	SCHOHARIE			1
-	INDEPENDENT LIVING FOR SENIORS	PACE		JANUARY 2009	MONROE			246
-	PAGE CNY	PACE		JANUARY 2009	CHAUTAUQUA			1
-	PAGE CNY	PACE		JANUARY 2009	ONONDAGA			317
-	AMERIGROUP	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			431
-	OCN SELECT	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			1040
-	OCN SELECT	PARTIAL CAPITATION		JANUARY 2009	WESTCHESTER			17
-	ELANT	PARTIAL CAPITATION		JANUARY 2009	ORANGE			104
-	CLANT	PARTIAL CAPITATION		JANUARY 2009	ROCKLAND			34
-	GUILDNET	PARTIAL CAPITATION		JANUARY 2009	NASSAU			519
-	GUILDNET	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			5810
-	GUILDNET	PARTIAL CAPITATION		JANUARY 2009	SUFFOLK			217
-	HHH CHOICES	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			706
-	HOMEFIRST	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			2976
-	INDEPENDENT CARE SYSTEMS	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			1331
-	PARTNERS IN COMMUNITY CARE	PARTIAL CAPITATION		JANUARY 2009	ORANGE			157
-	PARTNERS IN COMMUNITY CARE	PARTIAL CAPITATION		JANUARY 2009	ROCKLAND			91
-	SENIOR HEALTH PARTNERS INC	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			1375
-	SENIOR NETWORK HEALTH	PARTIAL CAPITATION		JANUARY 2009	HERKIMER			36
-	SENIOR NETWORK HEALTH	PARTIAL CAPITATION		JANUARY 2009	ONEIDA			319
-	TOTAL AGING IN PLACE PROGRAM	PARTIAL CAPITATION		JANUARY 2009	ERIE			142
-	VNS CHOICE	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			6783
-	WELLCARE	PARTIAL CAPITATION		JANUARY 2009	NEW YORK			158
-	ARCHCARE SENIOR LIFE	PACE		JANUARY 2010	NEW YORK			10
-	OHS BUFFALO LIFE	PACE		JANUARY 2010	ERIE			9
-	COMPREHENSIVE CARE MGMT	PACE		JANUARY 2010	NASSAU			8
-	COMPREHENSIVE CARE MGMT	PACE		JANUARY 2010	NEW YORK			2265
-	COMPREHENSIVE CARE MGMT	PACE		JANUARY 2010	SUFFOLK			61
-	COMPREHENSIVE CARE MGMT	PACE		JANUARY 2010	WESTCHESTER			175
-	EDDY SENIOR CARE	PACE		JANUARY 2010	SCHENECTADY			101

5 Columns 3,757 Rows 3 Data Types

RBC county

Unique Values

Search...

CHAUTAUQUA 46

DUTCHESS 50

ULSTER 42

RENSSELAER 41

ALLEGANY 40

NLAGARA 30

COLUMBIA 27

SULLIVAN 23

MONTGOMERY 21

PUTNAM 21

SCHOHARIE 20

FULTON 13

WARREN 13

WASHINGTON 13

CAYUGA 11

CORTLAND 11

ESSEX 11

BROOME 10

TOMPKINS 10

GREENE 10

DELAWARE 9

STEBEN 7

NEW YORK CITY TOTALS: 7

GRAND TOTALS: 7

TIOGA 5


CHENANGO 5

MADISON 4

LIVINGSTON 3

SARATOGA 2

# Visibility in OpenRefine

 **OpenRefine** all longterm managed care [Permalink](#)


Facet / Filter [Undo / Redo](#) 25 / 25

Refresh [Reset All](#) [Remove All](#)

**296 matching rows** (3769 total)


Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) rows

**enrollment** [change](#) [reset](#)



1,000.00 — 5,000.00

**report date** [change](#) [reset](#)



2010-11-30 22:58:30 — 2014-01-31 16:00:00

	All	plan name	plan type	report date	county	enrollment
<a href="#">☆</a> <a href="#">🗨</a>	61.	COMPREHENSIVE CARE MGMT	PACE	2011-01-01T00:00:00Z	NEW YORK	2320
<a href="#">☆</a> <a href="#">🗨</a>	73.	CCM SELECT	PARTIAL CAPITATION	2011-01-01T00:00:00Z	NEW YORK	1843
<a href="#">☆</a> <a href="#">🗨</a>	77.	ELDERPLAN	PARTIAL CAPITATION	2011-01-01T00:00:00Z	NEW YORK	3545
<a href="#">☆</a> <a href="#">🗨</a>	88.	INDEPENDENCE CARE SYSTEMS	PARTIAL CAPITATION	2011-01-01T00:00:00Z	NEW YORK	1645
<a href="#">☆</a> <a href="#">🗨</a>	89.	SENIOR HEALTH PARTNERS INC	PARTIAL CAPITATION	2011-01-01T00:00:00Z	NEW YORK	2535
<a href="#">☆</a> <a href="#">🗨</a>	94.	WELLCARE	PARTIAL CAPITATION	2011-01-01T00:00:00Z	NEW YORK	1226
<a href="#">☆</a> <a href="#">🗨</a>	99.	COMPREHENSIVE CARE MGMT	PACE	2012-01-01T00:00:00Z	NEW YORK	2688
<a href="#">☆</a> <a href="#">🗨</a>	109.	AMERIGROUP	PARTIAL CAPITATION	2012-01-01T00:00:00Z	NEW YORK	1429
<a href="#">☆</a> <a href="#">🗨</a>	110.	CCM SELECT	PARTIAL CAPITATION	2012-01-01T00:00:00Z	NEW YORK	3567
<a href="#">☆</a> <a href="#">🗨</a>	114.	ELDERPLAN	PARTIAL CAPITATION	2012-01-01T00:00:00Z	NEW YORK	4878
<a href="#">☆</a> <a href="#">🗨</a>	116.	ELDERSERVE	PARTIAL CAPITATION	2012-01-01T00:00:00Z	NEW YORK	4097

# Premature Commitment

- “Constraints on the order of doing things” [Blackwell et al., 2003]

# Hidden dependencies

- “Important links between entities are not visible” [Blackwell et al., 2003]
- The output of each transformation step in a wrangling process serves as the input for the next. So in wrangling dependencies are highly sequential
- But often the sequences doesn't matter unless it's a transformation that restructures the dataset.



# Role-Expressiveness

- “The purpose of an entity is readily inferred” [Blackwell et al., 2003]
- In data wrangling, it is already difficult to verbally express table transformations.
- Different tools use different vocabulary to describe the same thing.
  - Entity resolution: “cluster and edit” and “mass edit” in OpenRefine and “standardize” in DataPrep
  - DataPrep does include little icons, which are more helpful than no icons.

# Error-Proneness

- “The notation invites mistakes and the system gives little protection.” [Blackwell et al, 2003]
- In data wrangling, errors often creep in when filtering as Type I vs Type II errors in the gulf of execution and evaluation [Hutchins et al., 1985]
  - Type I / false positive: A row was removed, but it should have been kept.
  - Type II / false negative: A row was kept, but it should have been removed.
- You often have to approve operations on rows that you don't know the values of.

# Abstraction

- “Types and availability of abstraction mechanisms” [Blackwell et al, 2003]
- Wrangling actions may encapsulate many small, low-level actions.

# Secondary notation

- “Extra information in means other than formal syntax” [Blackwell et al, 2003]
- Secondary notation is often used in specifying column extraction methods
  - Python and “index slicing” such as `foo[0:5]`
  - Regular Expressions

# Closeness of mapping

- “Closeness of representation to domain” [Blackwell et al, 2003]
- As examples of direct-manipulation interfaces, both interfaces enjoy a very close mapping between notation and results it’s describing.

# Consistency

- “Similar semantics are expressed in similar syntactic forms”  
[Blackwell et al, 2003]

# Diffuseness

- “Verbosity of language” [Blackwell et al, 2003]
- In some tasks, the notation can be too concise, when you have to specify a sequence of three transformations that might be encapsulated in one transformation.
- Perhaps diffuseness and abstraction are two interrelated dimensions.

# Provisionality

- “Degree of commitment to actions or marks” [Blackwell et al, 2003]
- Both tools support a preview function that addresses provisionality.
- It makes sense that these interfaces may not suffer from pre-mature commitments because they both use this idea.



# Addressing provisionality with previews

### Add column based on column Column 1

New column name

On error  set to blank  store error  copy value from original column

Expression  Language

No syntax error.

**Preview** History Starred Help

row	value	value.find(/[A-Z]+, \s*\d{4}\$/) ...
1.	MANAGED LONG TERM CARE ENROLLMENT	Error: java.lang.ArrayIndexOutOfBoundsException: 0
2.	BY PLAN, COUNTY, AND PROGRAM	Error: java.lang.ArrayIndexOutOfBoundsException: 0
3.	null	Error: java.lang.ArrayIndexOutOfBoundsException: 0
4.	NYS JANUARY, 2009	JANUARY, 2009
5.	null	Error

OK Cancel

# Progressive evaluation

- “Work-to-date can be checked at any time” [Blackwell et al, 2003]
- This may be a barrier to collaboration in OpenRefine as it doesn't support concurrent modifications per project.