Advanced Data Management (CSCI 640/490)

Data Wrangling

Dr. David Koop

Types of Dirty Data Problems

- Separator Issues: e.g. CSV without respecting double quotes
  - `12, 13, "Doe, John", 45`
- Naming Conventions: `NYC` vs. `New York`
- Missing required fields, e.g. key
- Different representations: `2` vs. `two`
- Truncated data: `"Janice Keihanaikukauakahihuliheekahaunaele"` becomes `"Janice Keihanaikukauakahihuliheek"` on Hawaii license
- Redundant records: may be exactly the same or have some overlap
- Formatting issues: `2017-11-07` vs. `07/11/2017` vs. `11/07/2017`

[J. Canny et al.]

Dirty Data: Data Scientist's View

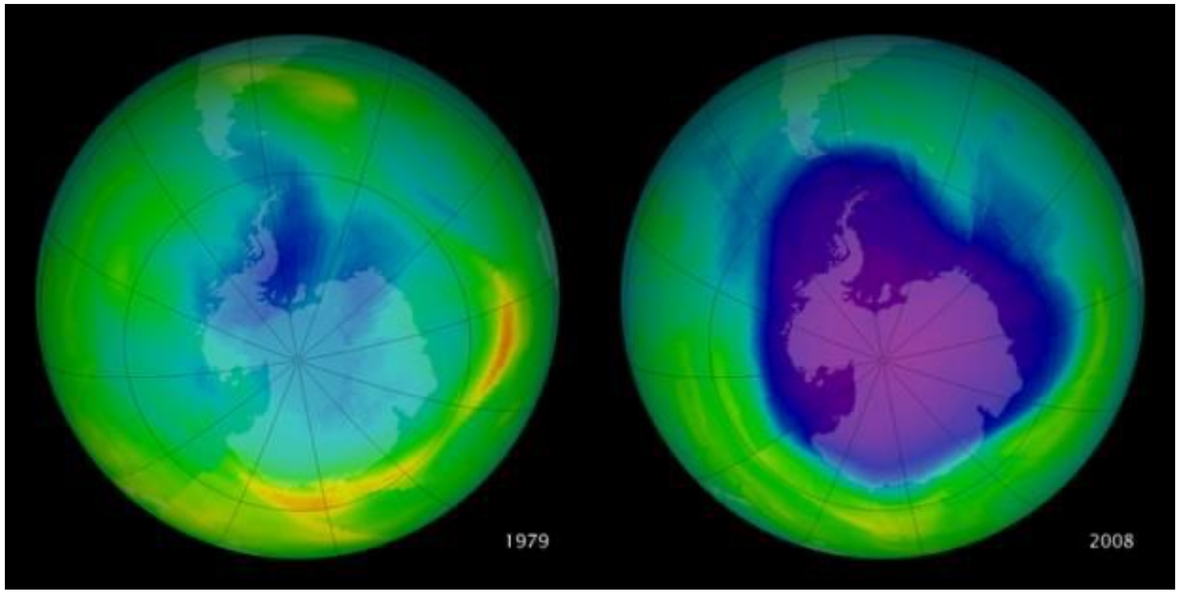
- Combination of:
  - Statistician's View: data has non-ideal samples for model
  - Database Expert's View: missing data, corrupted data
  - Domain Expert's View: data doesn't pass the smell test
- All of the views present problems with the data
- The goal may dictate the solutions:
  - Median value: don't worry too much about crazy outliers
  - Generally, aggregation is less susceptible by numeric errors
  - Be careful, the data may be correct…

[J. Canny et al.]

Be careful how you detect dirty data

- The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning.

  – National Center for Atmospheric Research



[Wikimedia]

CSAN Panel: Real Jobs in the Real World

- Tuesday, Oct. 7, 5:30–7:30pm
- Provides an insight into jobs from NIU alumni
- Food is Provided
- Sponsored by the Computer Science Alumni Network and the NIU Alumni Association

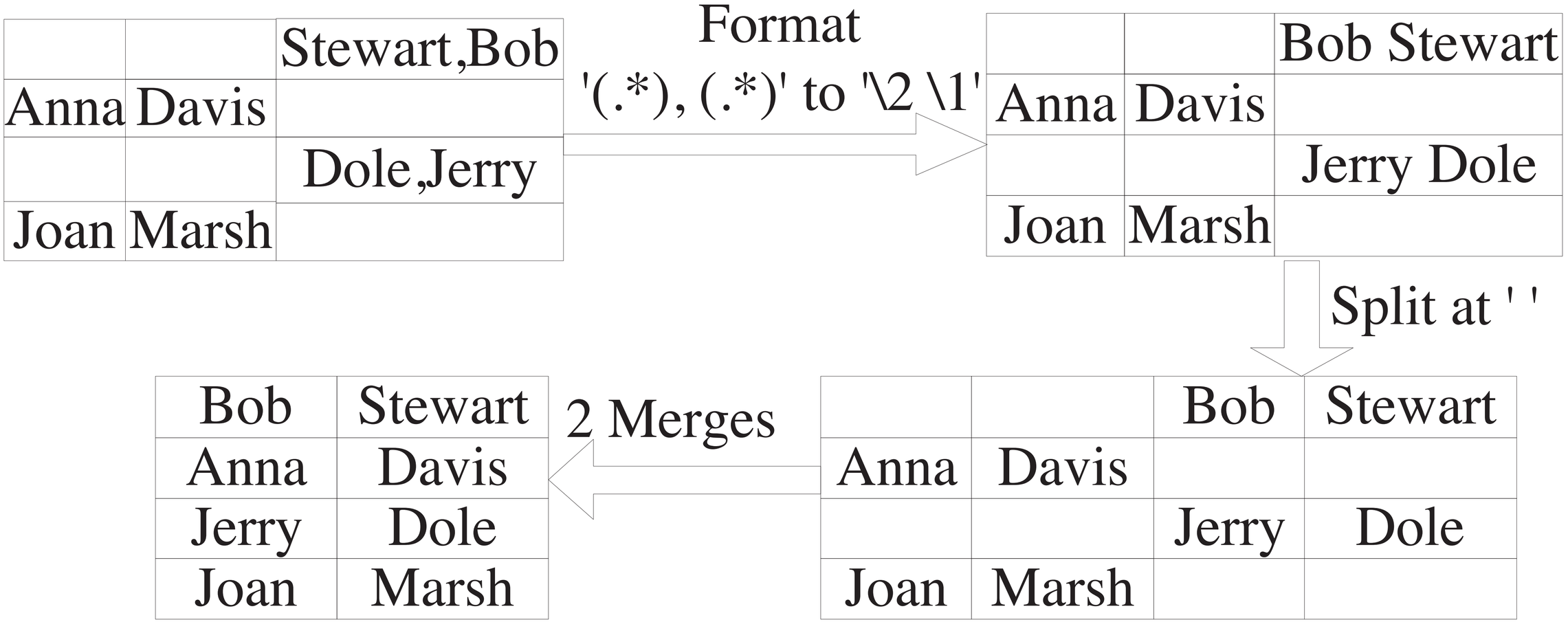
Assignment 2

- Assignment 1 Questions with polars, DuckDB, and pandas
- CS 640 students do all, CS 490 do polars & DuckDB (pandas is EC)
- Can work by framework or by query
- Most questions can be answered with a single statement… but that statement can take a while to write
  - Read documentation
  - Check hints

Wrangler

- Data cleaning takes a lot of **time** and **human effort**
- "Tedium is the message"
- Repeating this process on multiple data sets is even worse!
- Solution:
  - interactive interface (mixed-initiative)
  - transformation language with natural language "translations"
  - suggestions + "programming by demonstration"

Potter's Wheel: Example



[V. Raman and J. Hellerstein, 2001]

Potter's Wheel: Transforms

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

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 glue are values. f is a function mapping values to values. $x \oplus y$ concatenates x and y. 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 splitter. pred is a function returning a boolean.

[V. Raman and J. Hellerstein, 2001]

Potter's Wheel: Inferring Structure from Examples

Example Values Split By User (| is user specified split position)	Inferred Structure	Comments
Taylor, Jane |, $52,072 Blair, John |, $73,238 Tony Smith |, $1,00,533	$(< \xi^* >< \text{',' } \textit{Money} >)$	Parsing is doable despite no good delimiter. A *regular expression* domain can infer a structure of $[0-9,]* for last component.
MAA |to| SIN JFK |to| SFO LAX |–| ORD SEA |/| OAK	$(<\textit{len 3 identifier}>< \xi^* >$ $< \textit{len 3 identifier}>)$	Parsing is possible despite multiple delimiters.
321 Blake #7 |, Berkeley |, CA 94720 719 MLK Road |, Fremont |, CA 95743	$(<\textit{number } \xi^* >< \text{',' } \textit{word}>$ $<\text{',' } \textit{(2 letter word) (5 letter integer)}>)$	Parsing is easy because of consistent delimiter.

[V. Raman and J. Hellerstein, 2001]

Wrangler Transformation Language

- Based on Potter's Wheel
- Map: Delete, Extract, Cut, Split, Update
- Lookup/join: Use external data (e.g. from zipcode→state)
- Reshape: Fold and Unfold (aka pivot)
- Positional: Fill and lag
- Sorting, aggregation, key generation, schema transforms

Interface

- Automated Transformation Suggestions
- Editable Natural Language Explanations

- Visual Transformation Previews
- Transformation History

split	split1	split2	split3	split4
	2004	2004	2004	2003
STATE	Participation Rate 2004	Mean SAT I Verbal	Mean SAT I Math	Participation Rate
New York	87	497	510	82
Connecticut	85	515	515	84
Massachusetts	85	518	523	82
New Jersey	83	501	514	85
New Hampshire	80	522	521	75
D.C.	77	489	476	77
Maine	76	505	501	70
Pennsylvania	74	501	502	73
Delaware	73	500	499	73
Georgia	73	494	493	66

split	fold	fold1	value
New York	2004	Participation Rate 2004	87
New York	2004	Mean SAT I Verbal	497
New York	2004	Mean SAT I Math	510
New York	2003	Participation Rate 2003	82
New York	2003	Mean SAT I Verbal	496
New York	2003	Mean SAT I Math	510
Connecticut	2004	Participation Rate 2004	85
Connecticut	2004	Mean SAT I Verbal	515
Connecticut	2004	Mean SAT I Math	515
Connecticut	2003	Participation Rate 2003	84
Connecticut	2003	Mean SAT I Verbal	512

[S. Kandel et al., 2011]

Automation from past actions

- Infer parameter sets from user interaction
- Generating transforms
- Ranking and ordering transformations:
  - Based on user preferences, difficulty, and corpus frequency
  - Sort transforms by type and diversify suggestions

(a)

Reported crime in Alabama

(b)

before:	{‘in’, ‘ ’}	‘Alabama’ → {‘Alabama’, *word*}
selection:	{‘Alabama’}	‘in’ → {‘in’, *word*, *lowercase*}
after:	∅	‘ ’ → {‘ ’}

(c)

before:	{(‘ ’), (‘in’, ‘ ’), (*word*, ‘ ’), (*lowercase*, ‘ ’)}
selection:	{(‘Alabama’), (*word*)}
after:	∅

(d)

{(),(‘Alabama’),()}	~~{(),(*word*),()}~~
~~{(‘ ’),(),()}~~	~~{(*word*, ‘ ’),(),()}~~
{(‘ ’),(‘Alabama’),()}	{(*word*, ‘ ’),(‘Alabama’),()}
~~{(‘ ’),(*word*),()}~~	~~{(*word*, ‘ ’),(*word*),()}~~
{(‘in’, ‘ ’),(),()}	{(*lowercase*, ‘ ’),(),()}
{(‘in’, ‘ ’),(‘Alabama’),()}	{(*lowercase*, ‘ ’),(‘Alabama’),()}
{(‘in’, ‘ ’),(*word*),()}	~~{(*lowercase*, ‘ ’),(*word*),()}~~

(e) {(*lowercase*, ‘ ’),(‘Alabama’),()} → /[a-z]+ (Alabama)/

[S. Kandel et al., 2011]

Data Wrangler Demo

- http://vis.stanford.edu/wrangler/app/

Transform Script Import Export

- Split **data repeatedly** on **newline** into **rows**
- Split **split repeatedly** on **','**
- Promote **row 0** to header

Text | Columns | Rows | Table | Clear

Delete **row 7**

Delete **empty rows**

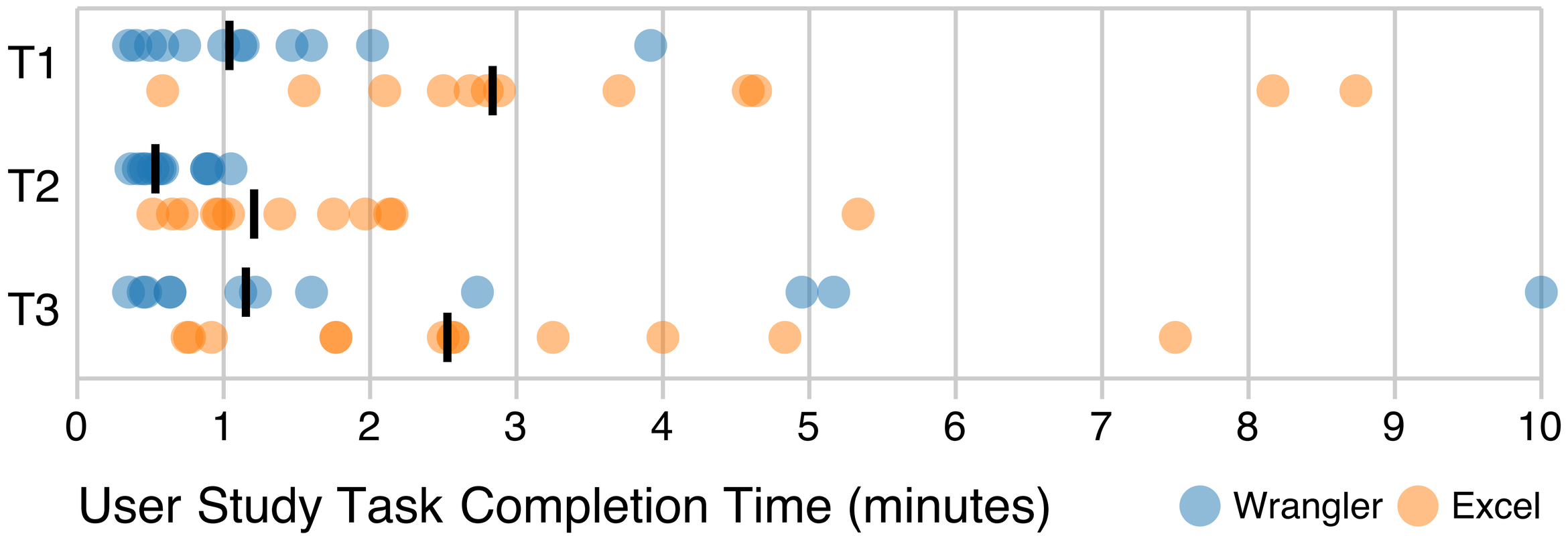
Fill **row 7** by **copying** values from **above**

	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

Evaluation

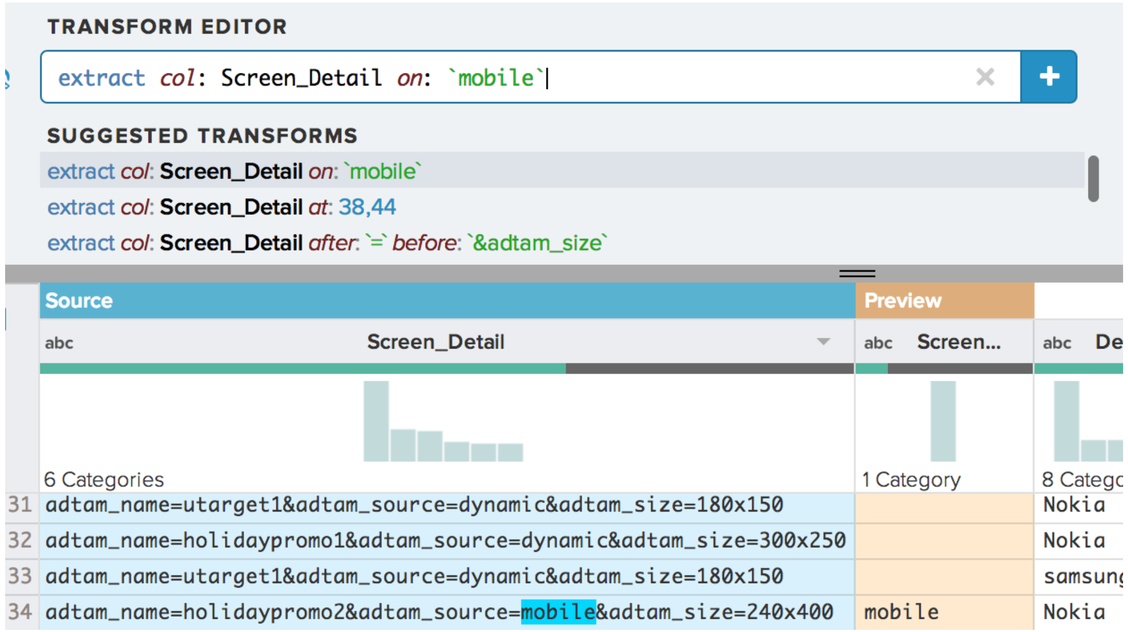
- Compare with Excel
- Tests:
  - Extract text from a single string entry
  - Fill in missing values with estimates
  - Reshape tables
- Allowed users to ask questions about Excel, not Wrangler
- Found significant effect of tool and users found previews and suggestions helpful
- Complaint: No manual fallback, make implications of user choices more obvious for users

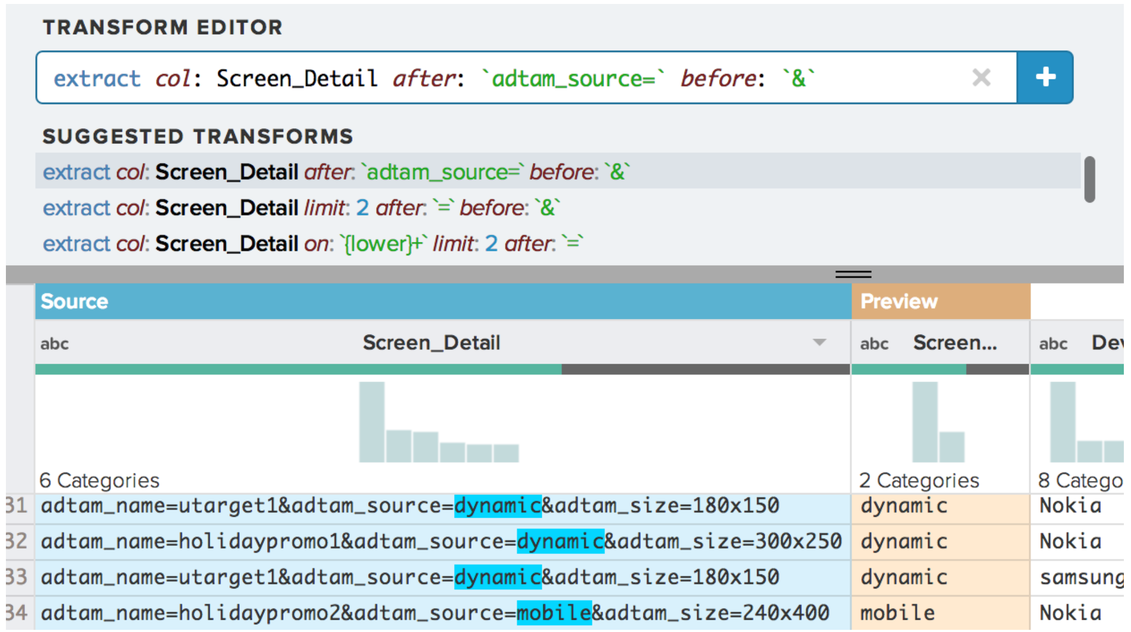
Task Completion Times



[S. Kandel et al., 2011]

Improvements in Prediction





Update suggestions when given more information

[Heer et al., 2015]

Data Wrangling Tasks

- Unboxing: Discovery & Assessment: What's in there? (types, distribution)
- Structuring: Restructure data (table, nested data, pivot tables)
- Cleaning: does data match expectations (often involves user)
- Enriching & Blending: Adding new data
- Optimizing & Publishing: Structure for storage or visualization

[J. M. Hellerstein et al., 2018]

Differences with Extract-Transform-Load (ETL)

- ETL:
  - Who: IT Professionals
  - Why: Create static data pipeline
  - What: Structured data
  - Where: Data centers
- "Modern Data Preparation":
  - Who: Analysts
  - Why: Solve problems by designing recipes to use data
  - What: Original, custom data blended with other data
  - Where: Cloud, desktop

[J. M. Hellerstein et al., 2018]

Evolution of Wrangler

- Authors started a company, Trifacta
- Eventually bought by Alteryx
- Now known as Alteryx Designer Cloud
- Offer Free Student Licenses: Link

Other Tools

- Open Refine
- VS Code Data Wrangler extension
-

Transform Data by Example

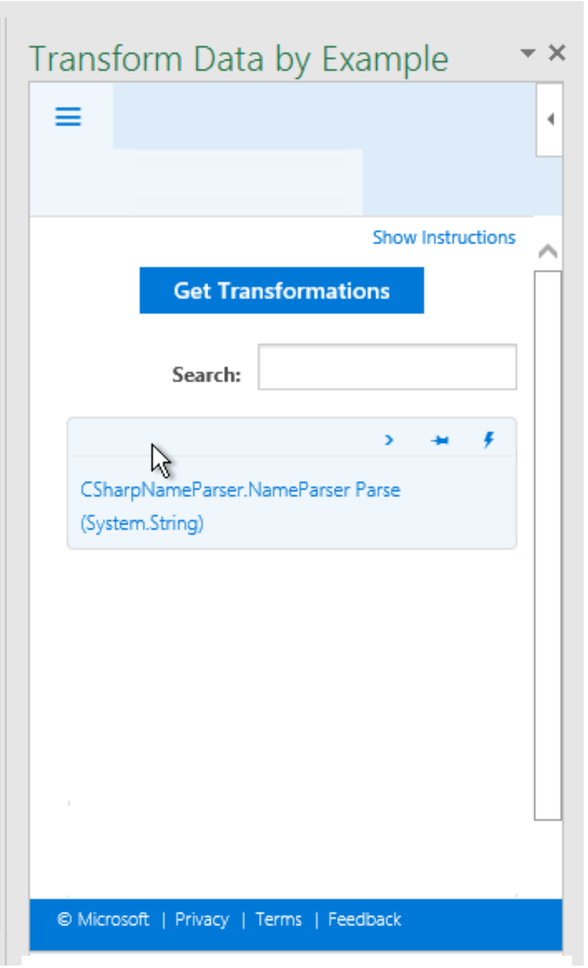
Wrangler

- Have to know what operations to apply
- What about an example-based approach instead?

Microsoft's Transform by Example

TDE: Transform Data by Example

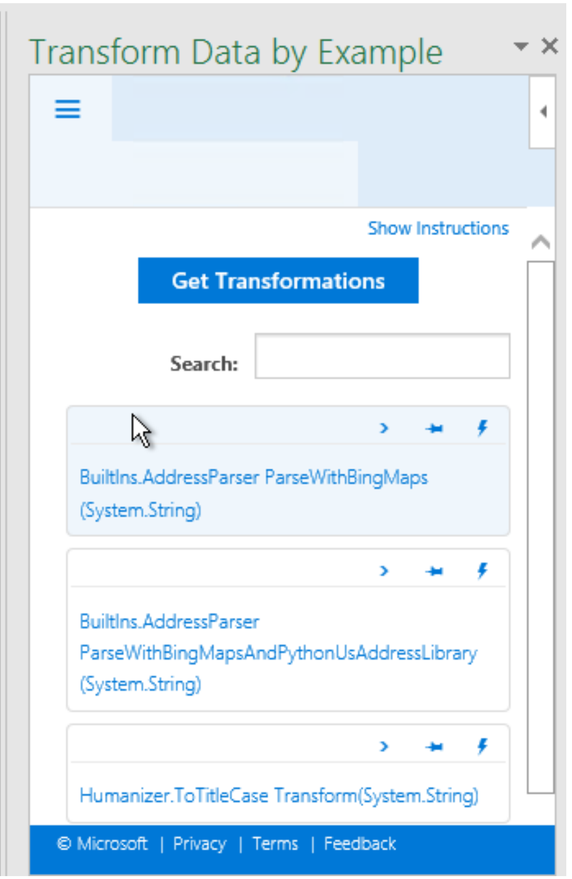
C	D
Customer Name	**Output**
John K. **Doe** Jr.	Doe, John
Mr. **Doe, John**	Doe, John
Jane A. **Smith**	Smith, Jane
MS. **Jane Smith**	Smith, Jane
Smith, Jane	Smith, Jane
Dr **Anthony** R **Von Fange** III	Von Fange, Anthony
Peter Tyson	Tyson, Peter
Dan E. **Williams**	Williams, Dan
James Davis Sr.	Davis, James
James J. **Davis**	Davis, James
Mr. **Donald** Edward **Miller**	Miller, Donald



[Y. He et al., 2018]

TDE: Transform Data by Example

C	D
Address	**Output**
4297 148th Avenue NE L105, Bellevue, WA 98007	Bellevue, WA, 98007
2720 N Mesa St, El Paso, 79902, USA	El Paso, TX, 79902
3524 W Shore Rd APT 1002, Warwick,02886	Warwick, RI, 02886
4740 N 132nd St, Omaha, 68164	Omaha, NE, 68164
10508 Prairie Ln, Oklahoma City	Oklahoma City, OK, 73162
525 1st St, Marysville, WA 95901	Marysville, CA, 95901
211 W Ridge Dr, Waukon,52172	Waukon, IA, 52172
1008 Whitlock Ave NW, Marietta, 30064	Marietta, GA, 30064
602 Highland Ave, Shinnston, 26431	Shinnston, WV, 26431
840 W Star St, Greenville, 27834	Greenville, NC, 27834

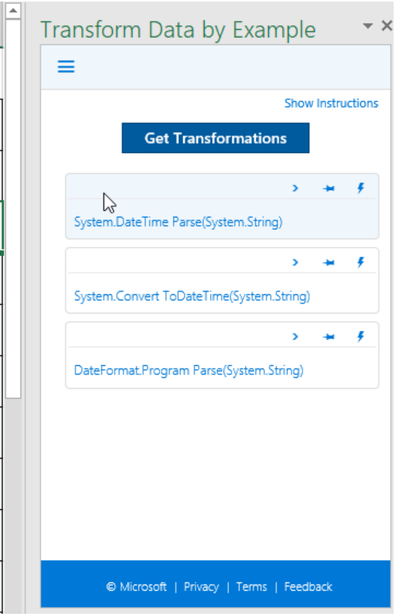


[Y. He et al., 2018]

TDE: Transform Data by Example

C	D
Transaction Date	**output**
Wed, 12 Jan 2011	2011-01-12-Wednesday
Thu, 15 Sep 2011	2011-09-15-Thursday
Mon, 17 Sep 2012	
2010-Nov-30 11:10:41	
2011-Jan-11 02:27:21	
2011-Jan-12	
2010-Dec-24	
9/22/2011	
7/11/2012	
2/12/2012	

C	D
Transaction Date	**output**
Wed, 12 Jan 2011	2011-01-12-Wednesday
Thu, 15 Sep 2011	2011-09-15-Thursday
Mon, 17 Sep 2012	2012-09-17-Monday
2010-Nov-30 11:10:41	2010-11-30-Tuesday
2011-Jan-11 02:27:21	2011-01-11-Tuesday
2011-Jan-12	2011-01-12-Wednesday
2010-Dec-24	2010-12-24-Friday
9/22/2011	2011-09-22-Thursday
7/11/2012	2012-07-11-Wednesday
2/12/2012	2012-02-12-Sunday



[Y. He et al., 2018]

TDE: Synthesized Function

Input Examples
Wed, 12 Jan 2011
Thu, 15 Sep 2011
Mon, 17 Sep 2012
2010-Nov-30 21:10:41
2011-Jan-11 02:27:21
2011-Jan-12

f

Return Object Dump					
Year	Month	Day	Day-of-week	Day-of-Year	...
2011	01	12	Wednesday	12	...
2011	09	15	Thursday	258	...
2012	09	17	Monday	261	...
2010	11	30	Tuesday	334	...
2011	01	11	Tuesday	11	...
2011	01	12	Wednesday	12	...

Member method result dump				
ToLongDateString()	ToTimeStr()	ToUTC()	ToBinary()	...
Wednesday, January 12, 2011	12:00:00 AM	...	...	...
Thursday, September 15, 2011	12:00:00 AM	...	...	...
Monday, September 17, 2012	12:00:00 AM	...	...	...
Tuesday, November 30, 2010	09:10:41 PM	...	...	...
Tuesday, January 11, 2011	02:27:21 AM	...	...	...
Wednesday, January 12, 2012	12:00:00 AM	...	...	...

Synthesize

Desired Output
2011-01-12 (Wed)
2011-09-15 (Thu)

[Y. He et al., 2018]

TDE: Transform Data by Example

- Row-to-row translation only
- Search System, GitHub, and StackOverflow for functions
- Given dataset with examples
  - Use L1 from library
  - Compose synthesized programs (L2)
  - Rank best transformations

TDE Benchmarks

System	Total cases (239)	FF-GR-Trifacta (46)	Head cases (44)	StackOverflow (49)	BingQL-Unit (50)	BingQL-Other (50)
TDE	**72% (173)**	**91% (42)**	**82% (36)**	**63% (31)**	**96% (48)**	**32% (16)**
TDE-NF	53% (128)	87% (40)	41% (18)	35% (17)	96% (48)	10% (5)
FlashFill	23% (56)	57% (26)	34% (15)	31% (15)	0% (0)	0% (0)
Foofah	3% (7)	9% (4)	2% (1)	4% (2)	0% (0)	0% (0)
DataXFormer-UB	38% (90)	7% (3)	36% (16)	35% (17)	62% (31)	**46% (23)**
System-A	13% (30)	52% (24)	2% (1)	10% (5)	0% (0)	0% (0)
OpenRefine-Menu[8]	4% (9)	13% (6)	2% (1)	4% (2)	0% (0)	0% (0)

- TDE and FlashFill focused on row-to-row transformations
- Foofah considers a wider range of transformations (table reformatting)

[Y. He et al., 2018]

Trifacta's Transform by Example

Transform by Pattern (TBP)

- Focus on non-technical users
- More general than Transform by Example
- No need for paired examples
- Use Cases:
  - Auto-Unify: Unify data in different formats
  - Auto-Repair: Fix data quality issues
- Example (Auto-Unify):
  - P_S = `<letter>{3}. <digit>{2}, <digit>{4}`
  - P_T = `<digit>{4}-<digit>{2}-<digit>{2}`

S-timestamp	S-phone	S-coordinates
2019-12-23	(425) 882-8080	(38°57'N, 95°15'W)
2019-12-24	(425) 882-8080	(38°61'N, 95°21'W)
2019-12-23	(206) 876-1800	(39°19'N, 95°18'W)
2019-12-24	(206) 876-1800	(39°26'N, 95°23'W)
2019-12-23	(206) 903-8010	(39°42'N, 96°38'W)

R-timestamp	R-phone	R-coordinates
Nov. 16 2019	650-853-1300	N37°31′ W122°14′
Nov. 17 2019	650-853-1300	N37°18′ W122°19′
Nov. 16 2019	425-421-1225	N37°48′ W122°17′
Nov. 17 2019	425-421-1225	N37°60′ W123°08′
Nov. 16 2019	650-253-0827	N37°01′ W123°72′

[Jin et al.]

TBP Use Cases

- Auto-Unify

S-timestamp	S-phone	S-coordinates
2019-12-23	(425) 882-8080	(38°57'N, 95°15'W)
2019-12-24	(425) 882-8080	(38°61'N, 95°21'W)
2019-12-23	(206) 876-1800	(39°19'N, 95°18'W)
2019-12-24	(206) 876-1800	(39°26'N, 95°23'W)
2019-12-23	(206) 903-8010	(39°42'N, 96°38'W)

R-timestamp	R-phone	R-coordinates
Nov. 16 2019	650-853-1300	N37°31′ W122°14′
Nov. 17 2019	650-853-1300	N37°18′ W122°19′
Nov. 16 2019	425-421-1225	N37°48′ W122°17′
Nov. 17 2019	425-421-1225	N37°60′ W123°08′
Nov. 16 2019	650-253-0827	N37°01′ W123°72′

- Auto-Repair

Date	Opponents
January 12, 1997	Venezuela
February 12, 1997	Peru
April 2, 1997	Colombia
1997-06-04	United States
1997-06-11	Chile
1997-06-14	Ecuador

(a) EN-Wiki: Dates

Year	Artist	Issue Price (BU)
1989	John Mardon	$16.25
1990	D.J. Craig	$16.75
1991	D.J. Craig	$16.75
1992	Karsten Smith	17.50
1993	Stewart Sherwood	$17.50
1994	Ian D. Sparkes	$17.95

(b) EN-Wiki: Currency values

Women's winner	Time
Anikó Kálovics	2:31:24
Lenah Cheruiyot	2:27:02
Lenah Cheruiyot	2:33.44
Emily Kimuria	2:28.42
Jane Ekimat	2:32.08

(c) EN-wiki:time

#	Original air date[1]
12	March 23, 2008
13	March 30, 2008
14	April 6, 2008
15	13 April 2008
16	20 April 2008

(d) EN-Wiki: Date

[Jin et al.]

TBP Programs and Triples

Table 1: An example repository of TBP programs (P_s, P_t, T), where each line is a TBP program. The first three programs can be used to auto-unify the two tables shown in Figure 2.

TBP-id	Source-pattern (P_s)	Target-pattern (P_t)	(T)
TBP-1	<letter>{3}. <digit>{2}, <digit>{4}	<digit>{4}-<digit>{2}-<digit>{2}	...
TBP-2	(<digit>{3}) <digit>{3}-<digit>{4}	<letter>{3}-<digit>{3}-<digit>{4}	...
TBP-3	(<digit>+°<num>'<letter>{1}, <digit>+°<num>'<letter>{1})	<letter>{1}<digit>+°<num>' <letter>{1}<digit>+°<num>'	...
...	...	...	...
TBP-7	<digit>{4}/<digit>{2}/<digit>{2}	<letter>{3} <digit>{2}	...
TBP-8	<num> kg	<num> lb	...
TBP-9	<num> lb	<num> lb <num> oz	...
...	...	...	...
TBP-15	<num> kg	<num>公斤	...
TBP-16	<letter>+ de <digit>{4}	<digit>{4}	...
...	...	...	...

CCT-id	Input-column (C)	Output-column (C')	Program (T)
CCT-1	(C_1) "Born" = {"02/22/1732", "10/30/1735", ... }	(C'_1) "Date of birth" = {"February 22, 1732", ... }	Listing 1
CCT-2	(C_2) "Date of birth" = {"February 22, 1732", ... }	(C'_2) "Born" = {"02/22/1732", "10/30/1735", ... }	...
CCT-3	(C_3) "Died" = {"02/14/1799", "07/04/1826", ... }	(C'_3) "Date of birth" = {"February 22, 1732", ... }	...
CCT-4	(C_4) "Date" = {"11/01/2019", "12/01/2019", ... }	(C'_4) "Date-2" = {"November 01, 2019", ... }	Listing 1
...	...	...	...
CCT-9	(C_9) "Name" = {"Washington, George", "Adam, John", ... }	(C'_9) "Date of birth" = {"February 22, 1732", ... }	∅
...	...	...	...

[Jin et al.]

Learning TBP Programs

- User Logs
  - Similar to Search Engines
  - (Privacy Issues)
- Tables
  - Find common tables whose rows can be linked
  - Link Wikipedia tables across languages
  - Obtain different data formats and abbreviations that can be used as patterns

[Jin et al.]

TBP Learning from Tables

T_1

Name	#	Born	Died
Washington, George	USA President (1)	02/22/1732	12/14/1799
Adams, John	USA President (2), VP (1)	10/30/1735	07/04/1826
Jefferson, Thomas	USA President (3), VP (2)	04/13/1743	07/04/1826
Madison, James	USA President (4)	03/16/1751	06/28/1836
Monroe, James	USA President (5)	04/28/1758	07/04/1851

T_2

Date of birth	President	Birthplace	State† of birth
February 22, 1732	George Washington	Westmoreland County	Virginia†
October 30, 1735	John Adams	Braintree	Massachusetts†

T_3

30.	**George Washington**	–	57y, 10d	22.02.1732	14.12.1799
31.	**John Quincy Adams**	Nat-Rep	57y, 7m, 20d	11.07.1767	23.02.1848
32.	**Thomas Jefferson**	Dem-Rep	57y, 10m, 18d	13.04.1743	04.07.1826
33.	**James Madison**	Dem-Rep	57y, 11m, 15d	16.03.1751	28.06.1836
34.	**James Monroe**	Dem-Rep	58y, 10m, 3d	28.04.1758	04.07.1831

T_4

1.	George Washington	Virginia	Feb. 22, 1732	Dec. 14, 1797
3.	Thomas Jefferson	Virginia	Apr. 13, 1743	July 4, 1826
4.	James Madison	Virginia	Mar. 16, 1751	June 28, 1836
6.	John Quincy Adams	Massachusetts	July 11, 1767	Feb. 23, 1848

T_5

	Name and (party)[1]	Term	State of birth	Born	Died	Religion[2]	Age at inaug.	Age at death
1.	Washington (F)[3]	1789–1797	Va.	2/22/1732	12/14/1799	Episcopalian	57	67
2.	J. Adams (F)	1797–1801	Mass.	10/30/1735	7/4/1826	Unitarian	61	90

T_6

PRESIDENT	BIRTH DATE	BIRTH PLACE	DEATH DATE	LOCATION OF DEATH
George Washington	Feb 22, 1732	Westmoreland Co., Va.	Dec 14, 1799	Mount Vernon, Va.
John Adams	Oct 30, 1735	Quincy, Mass.	July 4, 1826	Quincy, Mass.

[Jin et al.]

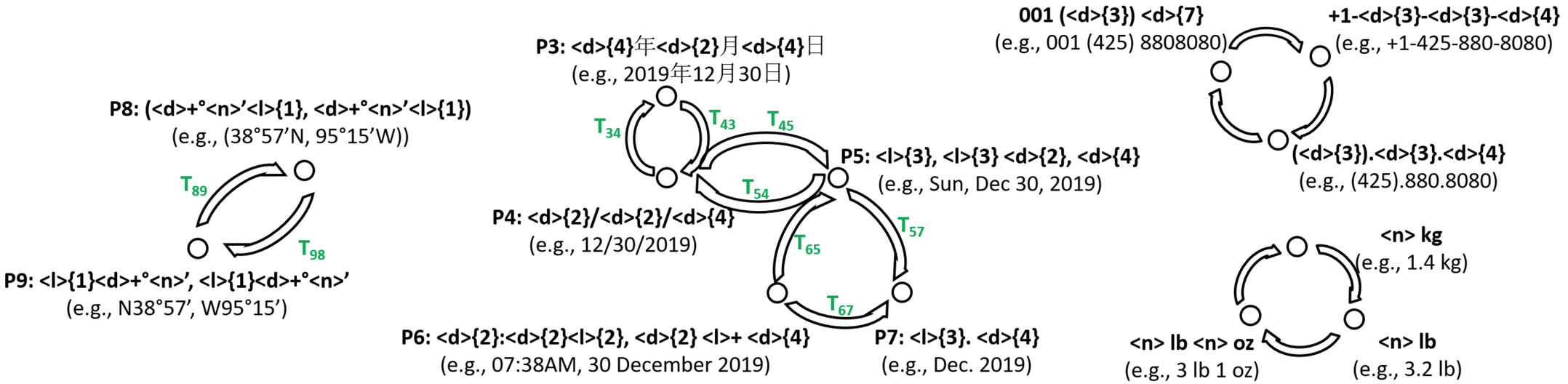
Generating Patterns

- Generate potential regex patterns
- Want more general patterns
- (`<digits>/<digits>/<digits>` vs. `<digits>/<digits>/17<digits>`)
- Can be too general: `<num><symbol><num><symbol><num>`
- Want high "coverage" and high "accuracy"

[Jin et al.]

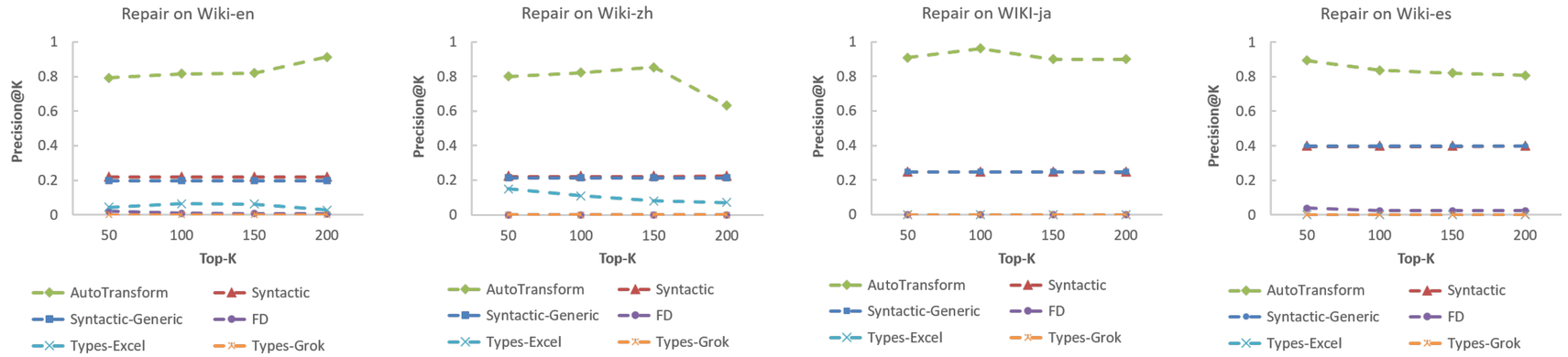
Graph Pattern Relationships

- Lossless inverses: can go back and forth
- Triangular equivalent programs: applying one transformation on a column matches the output of apply two other transformations in sequence



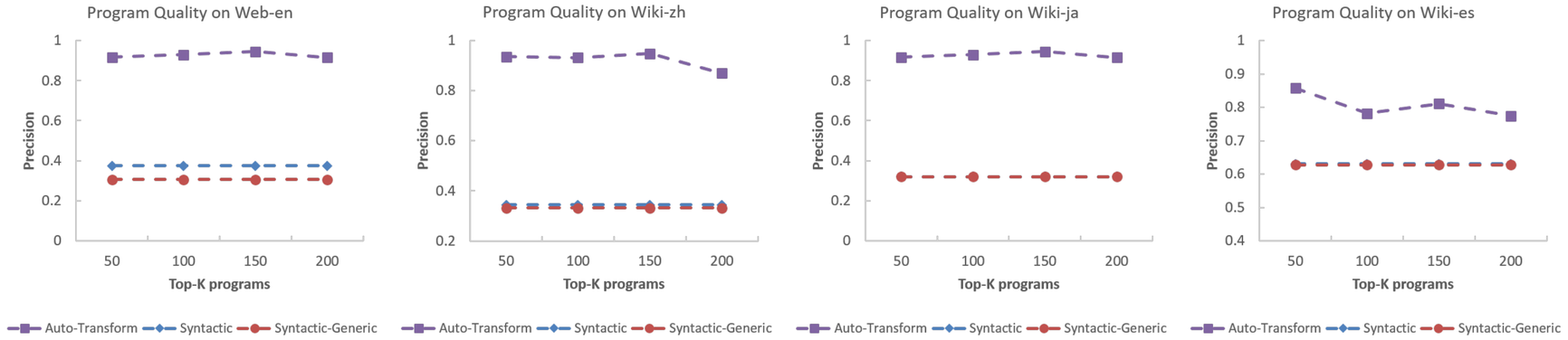
[Jin et al.]

Experiment Results



(a) Wiki-en (b) Wiki-zh (c) Wiki-ja (d) Wiki-es

Figure 10: Quality of repairs on Wiki-en, Wiki-zh, Wiki-ja, Wiki-es, using TBP programs learned from corresponding corpus.



(a) Web-en (b) Wiki-zh (c) Wiki-ja (d) Wiki-es

Figure 11: Quality of TBP programs produced on Web-en, Wiki-zh, Wiki-ja, Wiki-es, respectively.

[Jin et al.]

Data Wrangling and LLMs

- SheetAgent: https://sheetagent.github.io/
- SheetCopilot: https://sheetcopilot.github.io/
- Dango
  - Mixed-initiative system (DSL is also extended from Potter's Wheel)
  - https://site-gold-theta.vercel.app/dango/dango.html
- Table-GPT
  - Ensure LLMs can reason about 2D tables (not just 1D text)

Dango: Mixed-Initiative Data Wrangler using LLM