



Programming by Example: Challenges and Opportunities

Anish Doshi

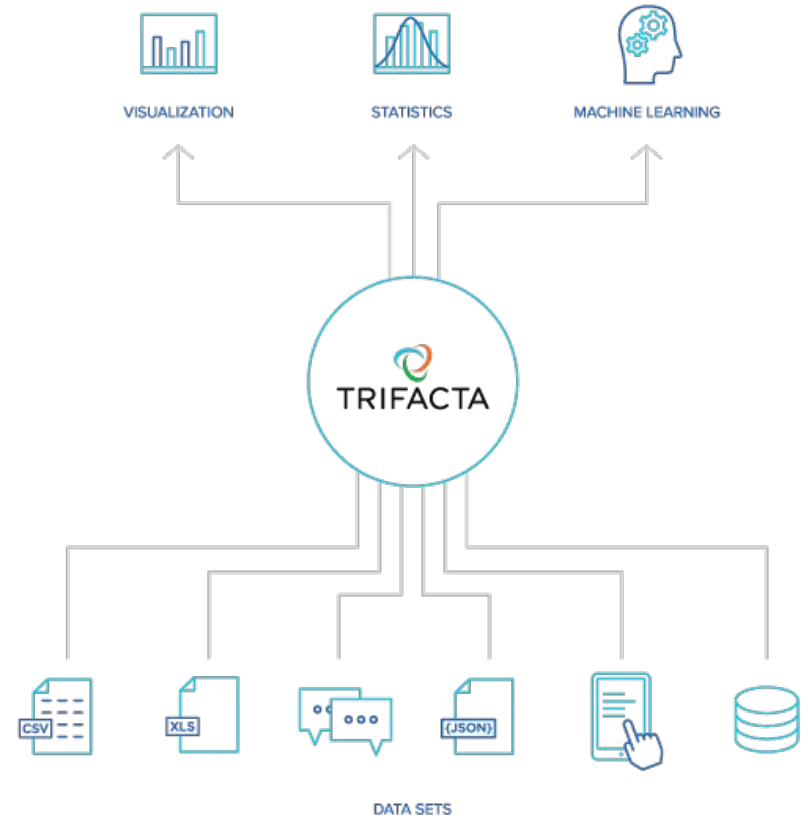
What this talk will cover

- What **programming by example** (PBE) is
- **Algorithms** for solving the PBE problem
- **Integrating** it into Trifacta, a production data application
- How we enable PBE to become a **user data-driven** feature



What Trifacta Is

- Data Preparation Platform - Focus on *Data Cleaning* for analytics/ML
- Data scientists can spend 80% of their time cleaning, validating, and preparing their data



What Trifacta Is

→ Interactive, "Excel Like" page for seeing, visualizing, and transforming data

CUSTOMER ANALYSIS > customer > Random

Run Job

New Step Recipe

#	IMSI	CONTRACT_END	CONTRACT_START	#	SUBSCRIBER_AGE	RBC	STATUS
310T - 310.26T		Jan 2013 - Dec 2016	Jan 2008 - Dec 2014	0 - 7		2 Categories	
310170226812721		6/4/16	7/29/09				ACTIVE
310160900766700		3/28/15	10/6/13	1			ACTIVE
310026939721905		9/11/15	9/18/10	4			ACTIVE
310026261822800		8/13/13	11/23/08	4			INACTIVE
310005667082048		8/4/16	10/22/14				ACTIVE
310170836020164		1/22/15	10/19/14	0			ACTIVE
310160772267782		11/21/15	12/28/14				ACTIVE
310170116249240		27-Sep-2011	2/9/09				INACTIVE
310004436630316		9/15/16	7/24/11				ACTIVE
310120423699542		2/27/15	6/29/11	3			ACTIVE
310030295859214		2/7/15	3/24/12	2			ACTIVE
310120387060694		10/1/16	10/25/11	3			ACTIVE
310026629156410		6/5/15	2/22/11	3			ACTIVE
310170433696080		10/1/14	5/15/13	1			INACTIVE
310260755452501		4/4/15	10/6/13	1			ACTIVE
310012031947706		10/29/16	11/26/14	0			ACTIVE
310120713199565		1/9/16	10/23/12	2			ACTIVE
310004657046981		4/28/16	2/16/13	1			ACTIVE
310150886959506		7/9/14	12/5/12	1			INACTIVE
310006069950078		12/26/15	2/5/12				ACTIVE
310026816434782		8/14/15	4/23/11	3			ACTIVE
310030452190559		5/1/15	5/22/11	3			ACTIVE
310006042012224		6/25/15	8/6/12	2			ACTIVE
310005594700906		6/19/15	2/8/14	0			ACTIVE
310012850050116		8/6/15	9/4/12	2			ACTIVE
310030990202436		6/19/15	2/24/13	1			ACTIVE

19 Columns 9,322 Rows 8 Data Types

- 1 Create CONTRACT_END_copy from CONTRACT_END
- 2 Move CONTRACT_END_copy after CONTRACT_END
- 3 Set CONTRACT_END_copy to IF(CONTRACT_END_copy > DATE(2013, 1, 1), NULL(), \$col)
- 4 Set CONTRACT_END_copy to Change date format of CONTRACT_END_copy to dd-MMM-yyyy
- 5 Set CONTRACT_END to IF(ISMISSING([CONTRACT_END_copy]), CONTRACT_END, CONTRACT_END_copy)
- 6 Delete CONTRACT_END_copy
- 7 Keep rows where (DATE(2008, 1, 1) <= CONTRACT_START) && (CONTRACT_START < DATE(2015, 1, 1))



Data cleaning involves...Stuff with Strings

- Dates, Phone Numbers, Addresses, Currencies, Floats, Emails, URLs
- User often wants to *standardize* a column to a single format
- Existing solution is in regex transformations / limited pattern standardization

The screenshot displays a data analysis application interface. The main window shows a data table with columns: #, IMSI, CONTRACT_END, CONTRACT_START, #, SUBSCRIBER_AGE, and STATUS. The data is grouped by time periods: Jan 2013 - Dec 2016 and Jan 2008 - Dec 2014. A modal window is open over the 'CONTRACT_END' column, showing a list of dates and a 'Convert' button. The 'Patterns' sidebar on the right contains several sections: 'Keep rows' with a regex pattern, 'Convert' with a 'to pattern format' field, 'Delete rows' with a regex pattern, and 'Set' with a 'to' field. The bottom status bar indicates 19 Columns, 9,322 Rows, and 8 Data Types.

#	IMSI	CONTRACT_END	CONTRACT_START	#	SUBSCRIBER_AGE	RBC	STATUS
3107 - 310.26T		6/4/16	7/29/09				
310160900766700		3/28/15	10/6/13				
310026261822380		9/11/15	9/10/08				
310026261822380		8/13/13	11/23/08				
310005667002048		8/4/16	10/22/14				
310170036020164		1/22/15	10/19/14				
310160772267782		11/21/15	12/28/14				
310170116249240		27-Sep-2011	2/9/09				
310004436630316		9/15/16	7/24/11				
310120423699542		2/27/15	6/29/11				
31003029589214		2/7/15	3/24/12				
310120387060694		10/1/16	10/25/11				
310026629156410		6/5/15	2/22/11				
310170433696080		10/1/14	5/15/13				
310260755452501		4/4/15	10/6/13				
310012031947706		10/29/16	11/26/14				
310120713199565		1/9/16	10/23/12				
310004657046981		4/28/16	2/16/13				
310150080909506		7/9/14	12/5/12				
310006069950070		12/26/15	2/5/12				
310026816447482		8/14/15	4/23/11				
310030452190559		5/11/15	5/22/11				
310006402012224		6/25/15	8/6/12				
310005504780906		6/19/15	2/8/14				
310012850050116		8/6/15	9/4/12				
310030990202436		6/19/15	2/24/13				



Cleaning messy data: Standardization

'949-727-6490'

'282-854-3328'

'972-399-4667'

'4046985388'

'6177590287'

'8133663103'

'(315)·732-3363'

'(315)·317-2248'



+1·949·727·6490

+1·282·854·3328

+1·972·399·4667

+1·404·698·5388

+1·617·759·0287

+1·813·366·3103

+1·315·732·3363

+1·315·317·2248

```

CREATE FUNCTION [dbo].[ufn_FormatPhone]
(@PhoneNumber VARCHAR(32))
RETURNS VARCHAR(32)
AS
BEGIN
    DECLARE @Phone CHAR(32)

    SET @Phone = @PhoneNumber

    -- cleanse phone number string
    WHILE PATINDEX('%[^0-9]%',@PhoneNumber) > 0
        SET @PhoneNumber = REPLACE(@PhoneNumber,
            SUBSTRING(@PhoneNumber,PATINDEX('%[^0-9]%',@PhoneNumber),1),'')

    -- skip foreign phones
    IF (SUBSTRING(@PhoneNumber,1,1) = '1'
        OR SUBSTRING(@PhoneNumber,1,1) = '+'
        OR SUBSTRING(@PhoneNumber,1,1) = '0')
        AND LEN(@PhoneNumber) > 11
        RETURN @Phone

    -- build US standard phone number
    SET @Phone = @PhoneNumber

    SET @PhoneNumber = '(' + SUBSTRING(@PhoneNumber,1,3) + ') ' +
        SUBSTRING(@PhoneNumber,4,3) + '-' + SUBSTRING(@PhoneNumber,7,4)

    IF LEN(@Phone) - 10 > 1
        SET @PhoneNumber = @PhoneNumber + ' X' + SUBSTRING(@Phone,11,LEN(@Phone) - 10)

    RETURN @PhoneNumber
END

```

(taken from stackoverflow)



What if you could just tell it what you want it too look like?

'949-727-6490'

+1 949 727 6490

In PBE, rather than specifying the program directly, the user specifies input/output **examples**, and the machine figures out the program the user would like to craft

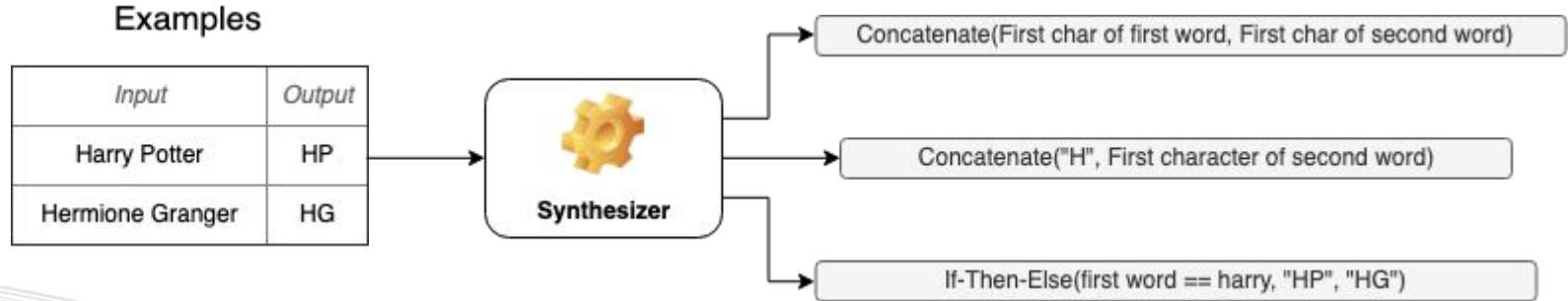


Building a PBE Algorithm



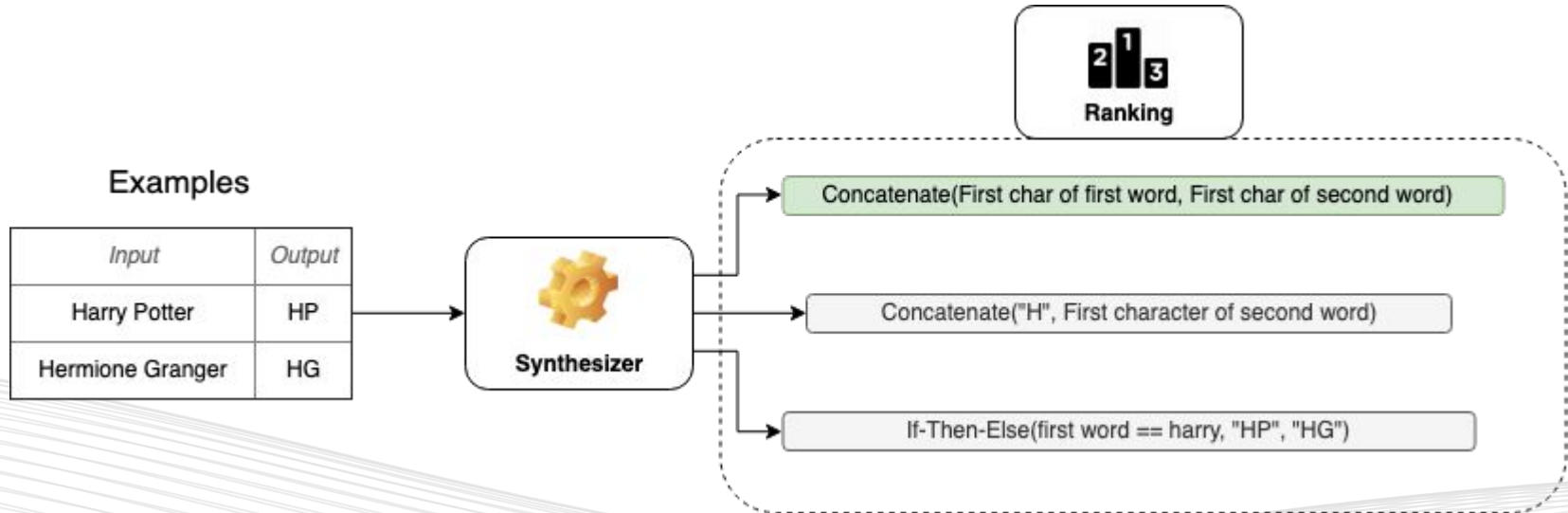
How it works

- General Idea: Given a set of input and output examples,
 - ◆ **synthesize** a set of programs that could represent that state



How it works

- General Idea: Given a set of input and output examples,
 - ◆ **synthesize** a set of programs that could represent that state
 - ◆ then **rank** them to pick the best one



Synthesis

- Domain specific languages (the language programs are written in, e.g. SQL) are usually too big to synthesize over
 - ◆ Large numbers of functions
 - ◆ Nesting
 - ◆ Multi-step programs
 - ◆ Numeric + String parameters
- Most PBE systems therefore restrict the DSL to something smaller, more task oriented
 - ◆ String Formatting DSL
 - ◆ Supports operations like Substring(), Concat(), Upper/Lowercasing



FlashFill (Gulwani 2011)

- First real software application of PBE (shipped in Microsoft Excel 2013)

<i>Input v_1</i>	<i>Output</i>
<i>(6/7)(4/5)(14/1)</i>	<i>6/7 # 4/5 # 14/1 #</i>
<i>49(28/11)(14/1)</i>	<i>28/11 # 14/1 #</i>
<i>() (28/11)(14/1)</i>	<i>28/11 # 14/1 #</i>

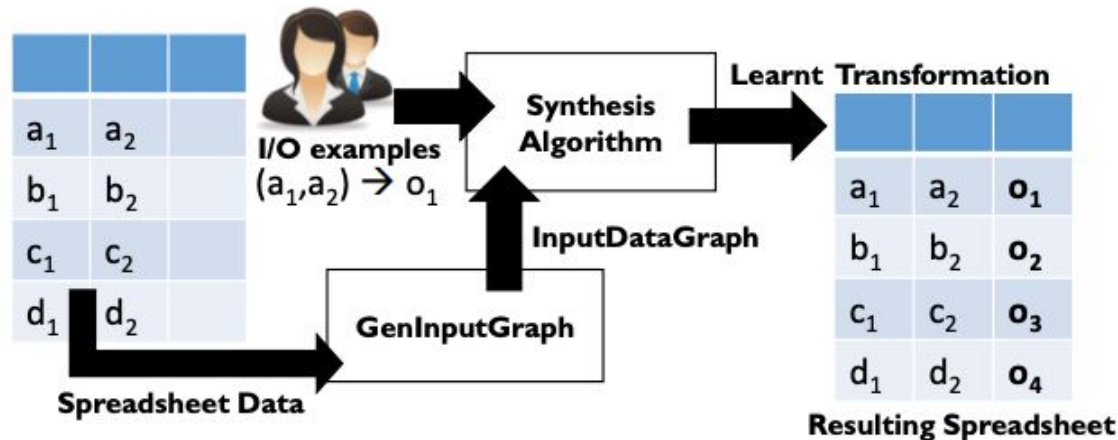
String Program:

Loop($\lambda w : \text{Concatenate}(\text{SubStr}(v_1, p_1, p_2), \text{ConstStr}(\text{"# "}))$)
where $p_1 \equiv \text{Pos}(\text{LeftParenTok}, \text{TokenSeq}(\text{NumTok}, \text{SlashTok}), w)$
and $p_2 \equiv \text{Pos}(\text{TokenSeq}(\text{SlashTok}, \text{NumTok}), \text{RightParenTok}, w)$.



BlinkFill (Singh 2016)

- Idea: Programs should be semantically valid for the *whole column*, not just for input examples provided
- Space of such programs is also dramatically smaller, leading to increased performance (up to 40x as fast as FlashFill, according to authors)



Ranking: Heuristics

- Simplest: Occam's Razor (prefer simpler, shorter programs)

```
MERGE([SUBSTRING(column1, 0, 1), SUBSTRING(column1,  
1, 2)], '')
```

```
SUBSTRING(column1, 0, 2)
```

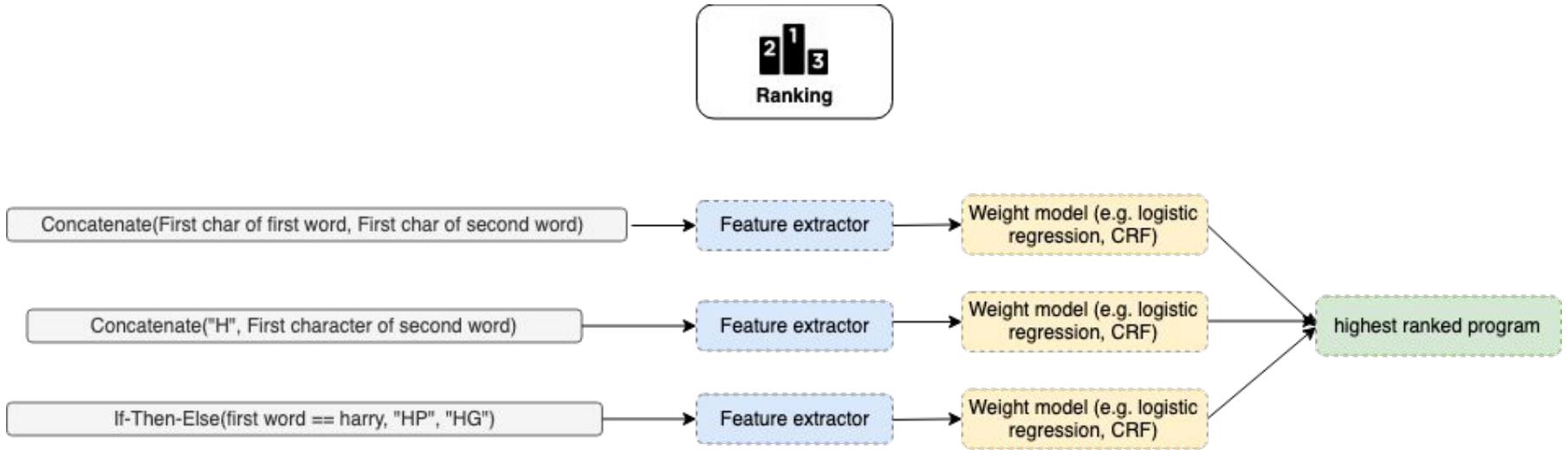


Ranking

- More sophisticated:
 - Prefer certain functions (e.g. Propercase over UPPER + lower)
 - Prefer substring boundaries that end at delimiters
 - Use *metadata* about the column (e.g., use date formatting functions in a date column)
- Can we improve these heuristics by looking at user data?



Ranking with ML



mixture of hand tuned **heuristics** (feature extractor) and **ml** (weight models are trained on data)



Ranking with ML: Challenges in Production

- Training Data: simply look at hand crafted transformations!
- I.E. - save data before a transformation, data afterwards as a set of *input examples*, save the transformation itself as the *output program*
- Operations that people are doing on your product are a great source of training data
- Personalization potentially possible through transfer learning



Ranking with ML: Challenges in Production

- How do you train models on user data while respecting data privacy?
 - Ideal is online trained models, but those may be hard to deploy
 - Another strategy: Mask sensitive fields in analytics pipeline
 - Fields like SSN, credit card numbers, email addresses should be "masked" before saving

original: 123-45-6789 -> 123 45 6789

masked: 999-99-9999 -> 999 99 9999

- Model still has access to the informational content of the pattern transformation

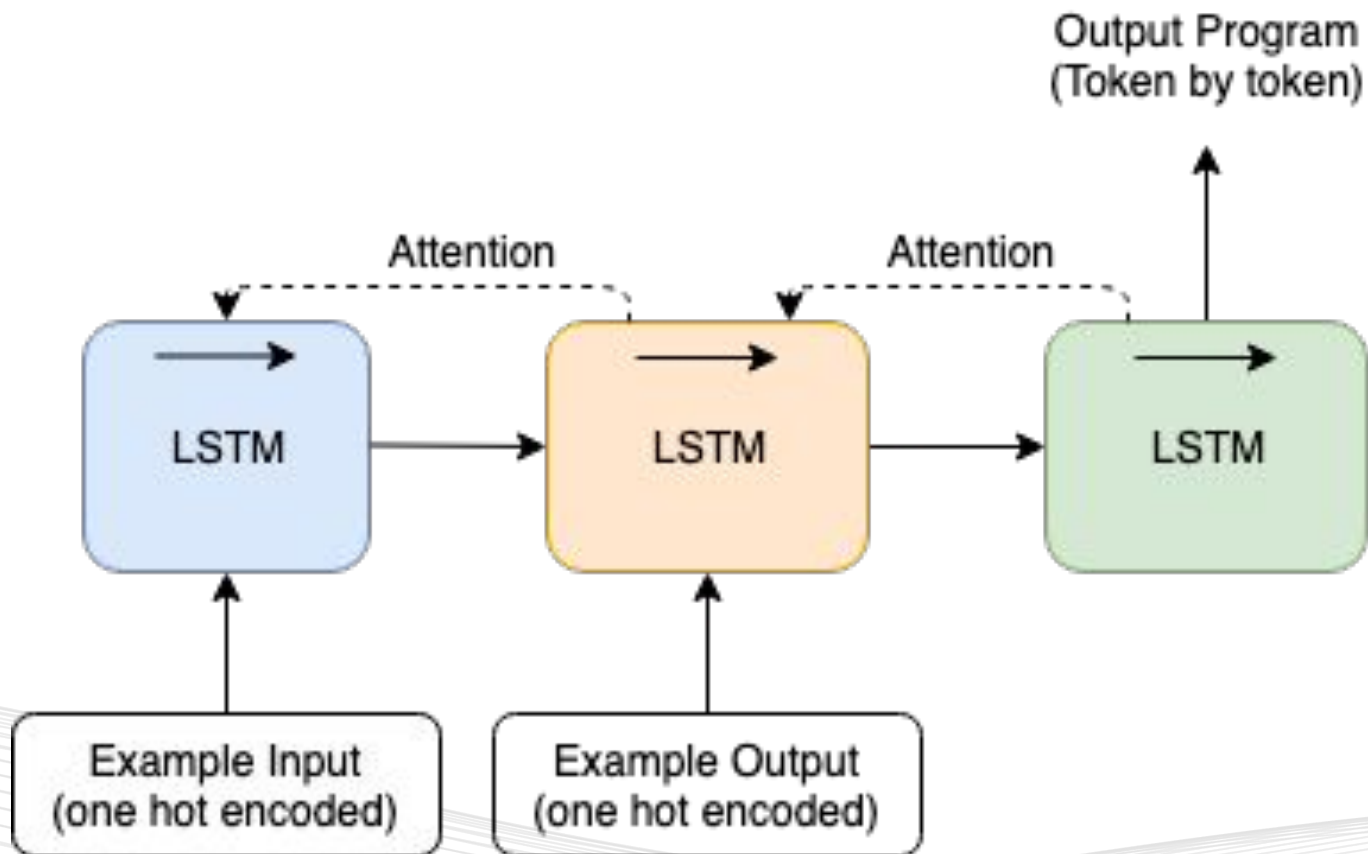


Neural Programming by Example

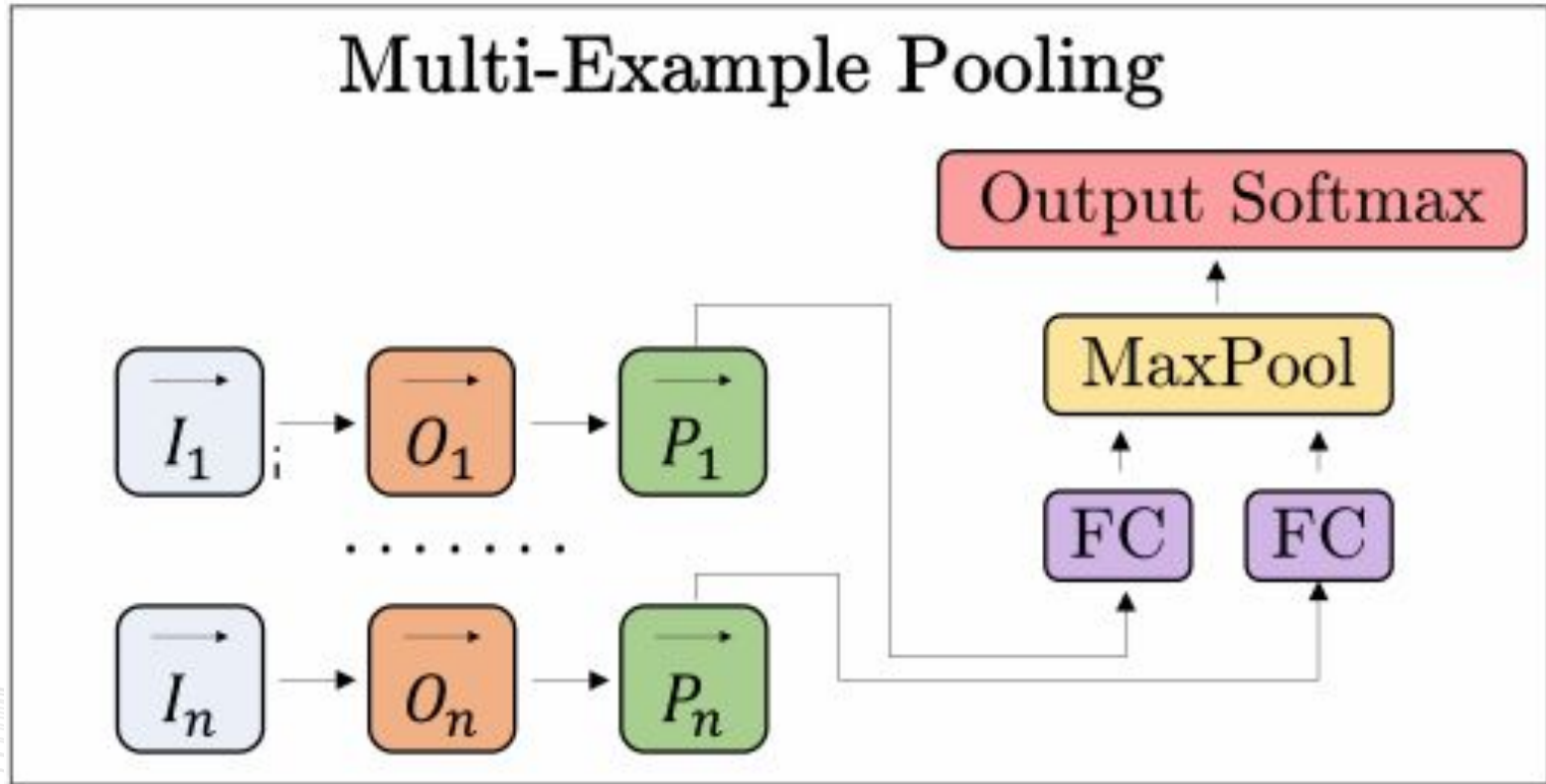
- Idea - Train a neural network directly to output a program given some encoding of input/output examples
- "Output a program" can mean a bunch of things:
 - Selecting a program from a preset list (a classification problem)
 - Hard to predict on such a large space - maybe prefilter to a threshold amount using heuristics, and then predict
 - Write out a program token by token (e.g. with an RNN)
 - Output a vector in some embedding space, and then find the closest valid program that satisfies the validity constraint
- Program Synthesis \neq Program Induction



RobustFill (Devlin, Uesato et al. 2017)



RobustFill (Devlin, Uesato et al. 2017)



RobustFill (Devlin, Uesato et al. 2017)

- How do you make sure the generated program actually works?
 - Uses a **modified beam search** when outputting program tokens to make sure the program result is as consistent with the examples as possible.
 - Relies on nature of the DSL (String concatenation based DSL similar to FlashFill/BlinkFill)
- Pros
 - Continuous space, so tolerant to noise in examples (e.g. typos)
 - Could be trained on data directly, no need for custom heuristics
- Cons
 - Potentially hard to interpret results
 - Hard to verify determinism



Neural Programming by Example: Challenges in Production

- Deployment
 - How do you make sure the prediction step happens in a scalable way?
 - Where do you store the neural network's weights, which can be quite large?
- Testing
 - How do you make guarantees on an inherently probabilistic operation?
 - Can you make guarantees about the **number of examples** it takes to output a correct program?
- Usability
 - How would users provide feedback to the operation of the network?



Building a User Interface for PBE



Started with a prototype

PBE Demo

Upload file

Phone Numbers (example)

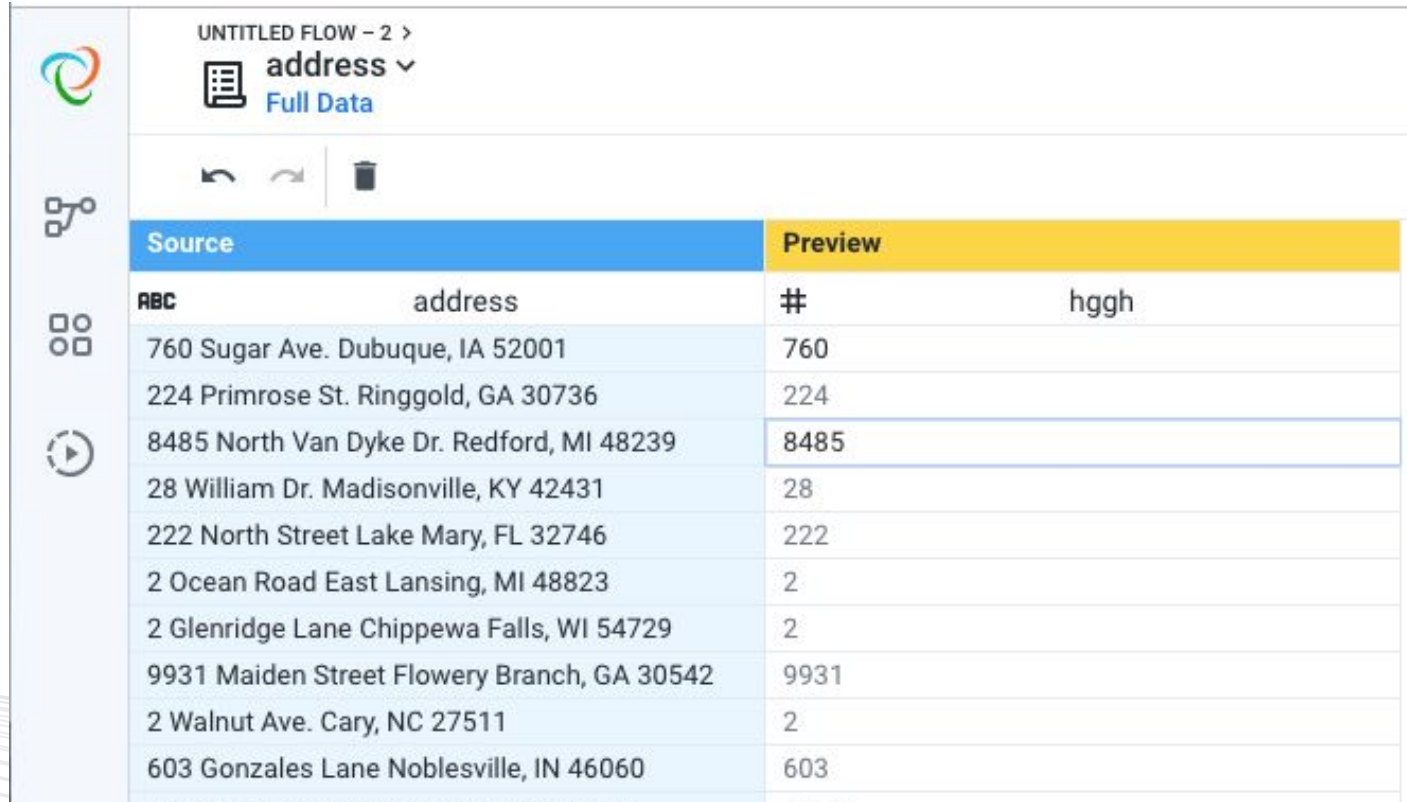
Cluster

Inputs	Outputs
(908) 902-5338	I
(302) 012-3489	
(124) 533-1207	
(128) 001-0123	
(963) 456-2039	

Interactivity and **Previewing** are important



Same basic idea applied in our main application...



The screenshot shows a software interface for a data flow. At the top, it says 'UNTITLED FLOW - 2 >' and 'address' with a dropdown arrow. Below that is a 'Full Data' link. There are navigation icons (undo, redo, delete) and a sidebar with icons for flow, data, and play. The main area contains a table with two columns: 'Source' (blue header) and 'Preview' (yellow header). The table has 11 rows of data, with the third row highlighted.

Source	Preview
ABC address	# hggh
760 Sugar Ave. Dubuque, IA 52001	760
224 Primrose St. Ringgold, GA 30736	224
8485 North Van Dyke Dr. Redford, MI 48239	8485
28 William Dr. Madisonville, KY 42431	28
222 North Street Lake Mary, FL 32746	222
2 Ocean Road East Lansing, MI 48823	2
2 Glenridge Lane Chippewa Falls, WI 54729	2
9931 Maiden Street Flowery Branch, GA 30542	9931
2 Walnut Ave. Cary, NC 27511	2
603 Gonzales Lane Noblesville, IN 46060	603



...but that raised a lot more questions

The screenshot shows a data tool interface with a table of data and a 'Select Column' dialog. The table has columns 'RBC', 'address', '#', and 'hggh'. The 'Preview' header is highlighted in yellow. A blue arrow points from the 'Preview' header to the 'Select Column' dialog. The dialog has a dropdown menu with 'address' selected and a 'New Column name' field with 'hggh' entered. The dialog also has 'Cancel' and 'Add to Recipe' buttons.

RBC	address	#	hggh
760 Sugar Ave. Dubuque, IA 52001		760	
224 Primrose St. Ringgold, GA 30736		224	
8485 North Van Dyke Dr. Redford, MI 48239		8485	
28 William Dr. Madisonville, KY 42431		28	
222 North Street Lake Mary, FL 32746		222	
2 Ocean Road East Lansing, MI 48823		2	
2 Glenridge Lane Chippewa Falls, WI 54729		2	
9931 Maiden Street Flowery Branch, GA 30542		9931	
2 Walnut Ave. Cary, NC 27511		2	
603 Gonzales Lane Noblesville, IN 46060		603	
7739 Hilldale Court Nazareth, PA 18064		7739	
921 Van Dyke Court Eastlake, OH 44095		921	
37 1st Street Roselle, IL 60172		37	
698 Nichols Lane Carrollton, GA 30117		698	
794 San Carlos Lane Lake Worth, FL 33460		794	
10 Carriage Ave. Key West, FL 33040		10	
31 St Margarets Court Manahawkin, NJ 08050		31	
7474 High Ridge Dr. Easton, PA 18042		7474	
460 Bayport Ave. Goldsboro, NC 27530		460	
98 Rock Creek Street Cambridge, MA 02138		98	
8908 Ridgeview Street Whitestone, NY 11357		8908	
8492 West Garfield Ave. Fort Lauderdale, FL 3330		8492	
9719 North Mechanic Dr. Sarasota, FL 34231		9719	
773 Railroad Lane Shelton, CT 06484		773	
54 Brookside Drive Bowling Green, KY 42101		54	
7 Swanson Street Brainerd, MN 56401		7	
1 Young Court Mahwah, NJ 07430		1	
6 S. Woodside Court Mount Prospect, IL 60056		6	
188 Court St. Niagara Falls, NY 14304		188	
14 Chestnut Road Sewell, NJ 08080		14	
414 Lakeview Street Defiance, OH 43512		414	
22 Tower St. Pueblo, CO 81001		22	
4 Church St. Hinesville, GA 31313		4	
477 Ravherry Court Toledo, OH 43612		477	

Can we allow users to interact, filter, sort their data from a **toolbar**?

If we know where the user should be entering examples, can we **prompt** them to do that somehow?



...but that raised a lot more questions

The screenshot shows a data tool interface with a table of addresses and a 'Select Column' dialog box. The table has columns for 'Source' and 'Preview'. The 'Preview' column shows a '#' and a program name 'hggh'. The 'Select Column' dialog box has a dropdown menu for 'Example Column' with 'address' selected, and a text input for 'New Column name' with 'hggh' entered. A callout box with a blue arrow points to the dialog, containing two questions:

Should users be allowed to **pick** between the top k ranked programs?

Should they be able to **edit** the generated program directly, in addition to providing examples?

Buttons at the bottom of the dialog include 'Cancel' and 'Add to Recipe'.

Source	Preview
RBC address	# hggh
760 Sugar Ave. Dubuque, IA 52001	760
224 Primrose St. Ringgold, GA 30736	224
8485 North Van Dyke Dr. Redford, MI 48239	8485
28 William Dr. Madisonville, KY 42431	28
222 North Street Lake Mary, FL 32746	222
2 Ocean Road East Lansing, MI 48823	2
2 Glenridge Lane Chippewa Falls, WI 54729	2
9931 Maiden Street Flowery Branch, GA 30542	9931
2 Walnut Ave. Cary, NC 27511	2
603 Gonzales Lane Noblesville, IN 46060	603
7739 Hilldale Court Nazareth, PA 18064	7739
921 Van Dyke Court Eastlake, OH 44095	921
37 1st Street Roselle, IL 60172	37
698 Nichols Lane Carrollton, GA 30117	698
794 San Carlos Lane Lake Worth, FL 33460	794
10 Carriage Ave. Key West, FL 33040	10
31 St Margarets Court Manahawkin, NJ 08050	31
7474 High Ridge Dr. Easton, PA 18042	7474
460 Bayport Ave. Goldsboro, NC 27530	460
98 Rock Creek Street Cambridge, MA 02138	98
8908 Ridgeview Street Whitestone, NY 11357	8908
8492 West Garfield Ave. Fort Lauderdale, FL 3330	8492
9719 North Mechanic Dr. Sarasota, FL 34231	9719
773 Railroad Lane Shelton, CT 06484	773
54 Brookside Drive Bowling Green, KY 42101	54
7 Swanson Street Brainerd, MN 56401	7
1 Young Court Mahwah, NJ 07430	1
6 S. Woodside Court Mount Prospect, IL 60056	6
188 Court St. Niagara Falls, NY 14304	188
14 Chestnut Road Sewell, NJ 08080	14
414 Lakeview Street Defiance, OH 43512	414
22 Tower St. Pueblo, CO 81001	22
4 Church St. Hinesville, GA 31313	4
477 Ravenna Court Toledo, OH 43612	477



...but that raised a lot more questions

The screenshot shows a data processing application interface. On the left, a table displays a list of addresses with columns for 'RBC', 'address', and 'Preview'. The 'Preview' column contains a '#' and a value 'hggh'. A blue arrow points from a text box on the right to the row containing '8485' in the 'Preview' column.

Source	Preview
RBC address	# hggh
760 Sugar Ave. Dubuque, IA 52001	760
224 Primrose St. Ringgold, GA 30736	224
8485 North Van Dyke Dr. Redford, MI 48239	8485
28 William Dr. Madisonville, KY 42431	28
222 North Street Lake Mary, FL 32746	222
2 Ocean Road East Lansing, MI 48823	2
2 Glenridge Lane Chippewa Falls, WI 54729	2
9931 Maiden Street Flowery Branch, GA 30542	9931
2 Walnut Ave. Cary, NC 27511	2
603 Gonzales Lane Noblesville, IN 46060	603
7739 Hilldale Court Nazareth, PA 18064	7739
921 Van Dyke Court Eastlake, OH 44095	921
37 1st Street Roselle, IL 60172	37
698 Nichols Lane Carrollton, GA 30117	698
794 San Carlos Lane Lake Worth, FL 33460	794
10 Carriage Ave. Key West, FL 33040	10
31 St Margarets Court Manahawkin, NJ 08050	31
7474 High Ridge Dr. Easton, PA 18042	7474
460 Bayport Ave. Goldsboro, NC 27530	460
98 Rock Creek Street Cambridge, MA 02138	98
8908 Ridgeview Street Whitestone, NY 11357	8908
8492 West Garfield Ave. Fort Lauderdale, FL 3330	8492
9719 North Mechanic Dr. Sarasota, FL 34231	9719
773 Railroad Lane Shelton, CT 06484	773
54 Brookside Drive Bowling Green, KY 42101	54
7 Swanson Street Brainerd, MN 56401	7
1 Young Court Mahwah, NJ 07430	1
6 S. Woodside Court Mount Prospect, IL 60056	6
188 Court St. Niagara Falls, NY 14304	188
14 Chestnut Road Sewell, NJ 08080	14
414 Lakeview Street Defiance, OH 43512	414
22 Tower St. Pueblo, CO 81001	22
4 Church St. Hinesville, GA 31313	4
477 Ravenna Court Toledo, OH 43612	477

On the right, a 'Select Column' dialog box is open, showing 'Example Column' set to 'address' and 'New Column name' set to 'hggh'. At the bottom right, there are 'Cancel' and 'Add to Recipe' buttons.

How do we handle **failure** states?

How does the user get a **guarantee** about what will happen to the rest of their data?



Key Takeaways

- Programming by Example is a methodology for users to interact with data in new way
- Tradeoffs between ML and heuristics, in expressibility and determinism
- Building it requires full stack, cross-disciplinary thought





Darin Friedrichs

@crushspread

Follow



AI is going to take over the world... and this is what Excel auto-populated today.

DEC	December
NOV	November
OCT	October
APR	Aprembur
AUG	Augember
FEB	Febember
JAN	Janember
JUL	Julembur
JUN	Junember
MAR	Marembur
MAY	Mayember
SEP	Sepember

5:00 AM - 23 Oct 2018



Questions + Thanks!

www.trifacta.com

adoshi@trifacta.com

