An Iterative Approach to Synthesize Data Transformation Programs

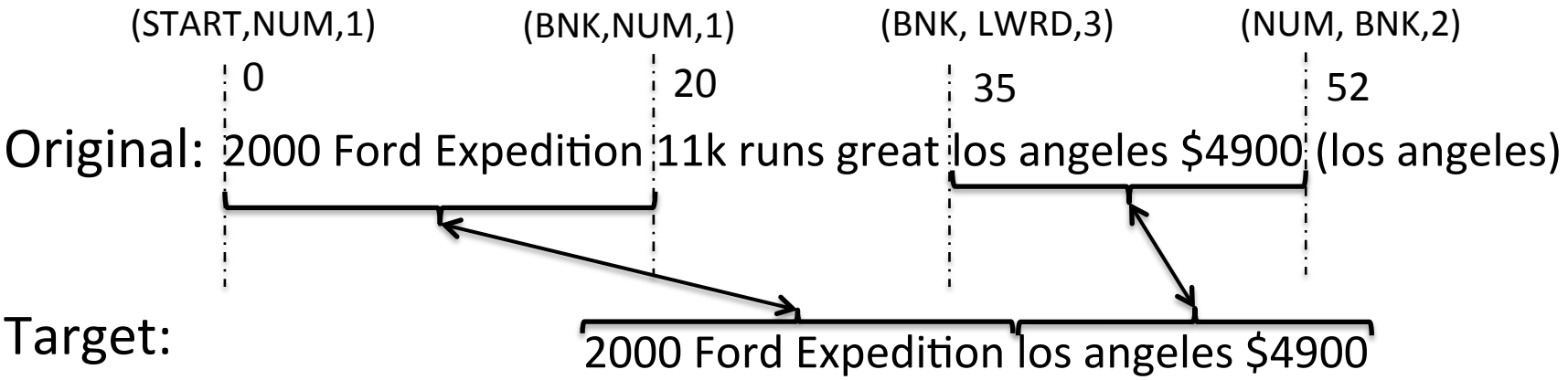
Bo Wu and Craig Knoblock
University of Southern California

Learning Transformation Programs by Example

Input Data	Target Data
2000 Ford Expedition 11k runs great los angeles $4900 (los angeles)	2000 Ford Expedition los angeles $4900
1998 Honda Civic 12k miles s. Auto. - $3800 (Arcadia)	2008 Mitsubishi Galant Sylmar CA $7500
2008 Mitsubishi Galant ES $7500 (Sylmar CA) pic	1998 Honda Civic Arcadia $3800
1996 Isuzu Trooper 14k clean title west covina $999 (west covina) pic	1996 Isuzu Trooper west covina $999
…	…

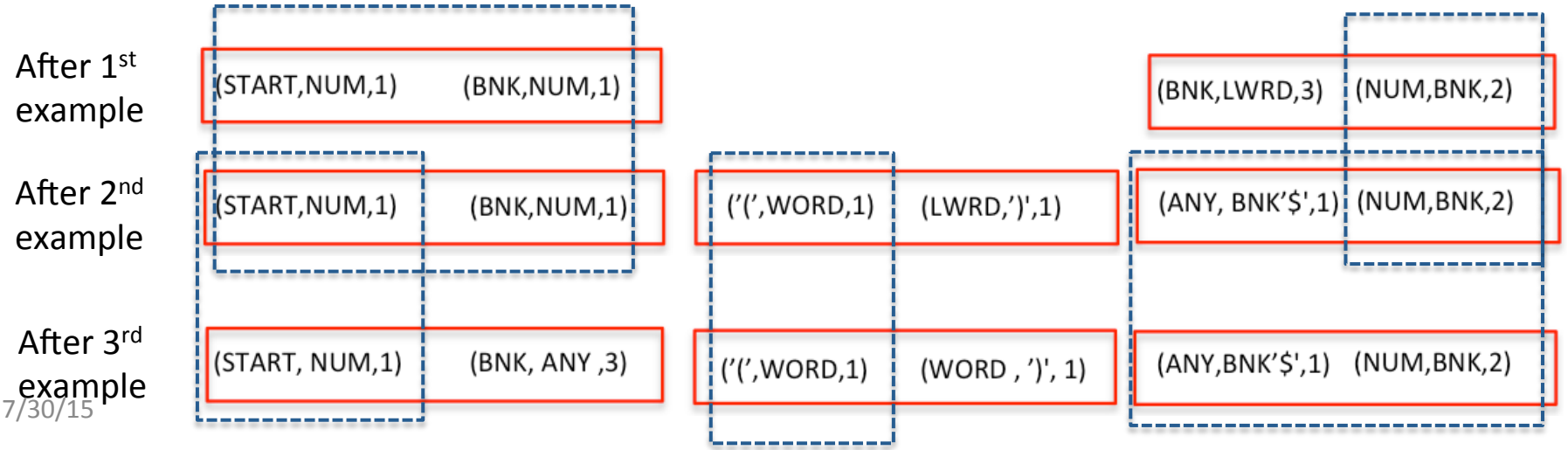
Time complexity is **exponential** in the **number** and a **high polynomial** in the **length** of examples

Reuse subprograms



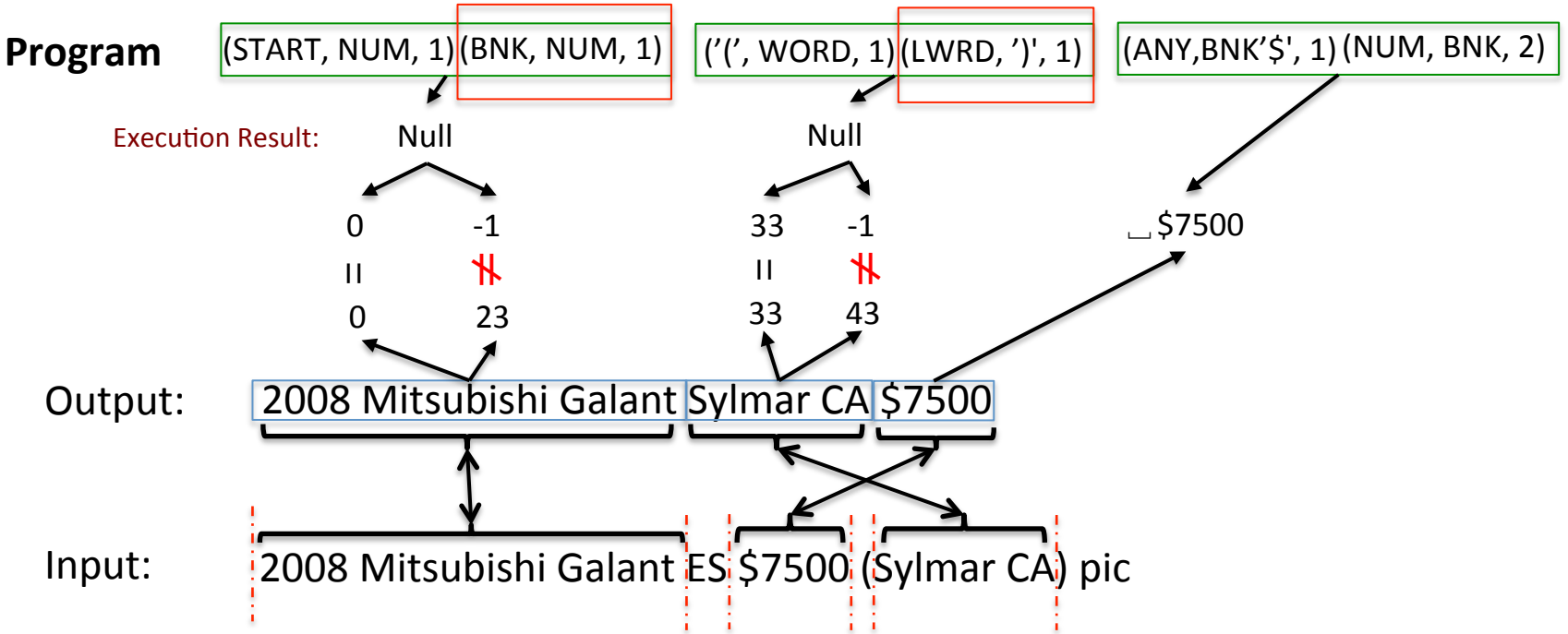
Position program= (left context, right context, occurance)

Learned Programs



Identify incorrect subprograms

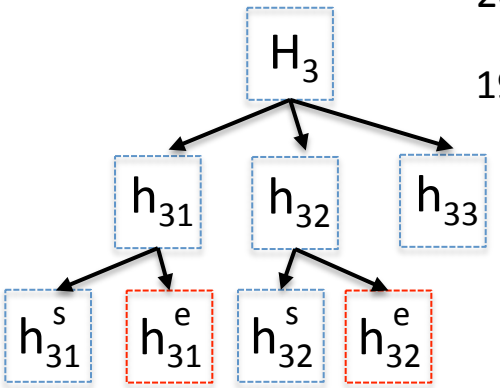
Input	Output
2000 Ford Expedition 11k runs great los angeles $4900 (los angeles)	2000 Ford Expedition los angeles $4900
1998 Honda Civic 12k miles s. Auto. - $3800 (Arcadia)	2008 Mitsubishi Galant Sylmar CA $7500



Update hypothesis spaces

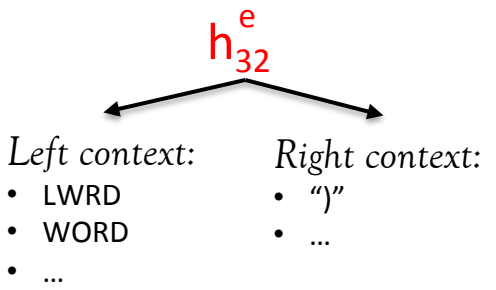
Program (START, NUM, 1) (BNK, NUM, 1) ('(', WORD, 1) (LWRD, ')', 1) (ANY,BNK'$', 1)

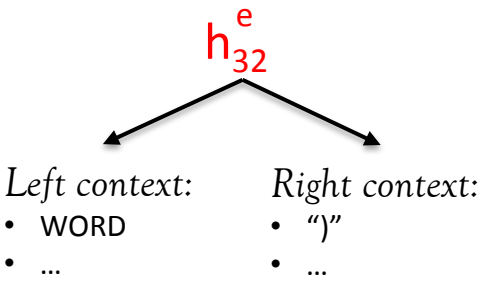
Hypothesis H_3



2000 Ford Expedition 11k runs great los angeles $4900 (los angeles)

1998 Honda Civic 12k miles s. Auto. - $3800 (Arcadia)





2008 Mitsubishi Galant ES $7500 (Sylmar CA) pic

Evaluation

- Dataset
  - **D1**: 17 scenarios used in (Lin et al., 2014)
    - 5 records per scenario
  - **D2**: 30 scenarios collected from student data integration projects
    - about 350 records per scenario
  - **D3**: synthetic dataset
    - designed to evaluate scale-up
- Alternative approaches
  - **Our implementation of Gulwani's approach**: (Gulwani, 2011)
  - **Metagol:** (Lin et al., 2014)
- Metric
  - Time (in **seconds**) to generate a transformation program

Program generation time comparisons

Table: time (in seconds) to generate programs on D1 and D2 datasets

		Min	Max	Avg	Median
D1	IPBE	0	5	0.34	0
	Gulwani's approach	0	8	0.59	0
	Metagol	0	213.93	55.1	0.14
D2	IPBE	0	1.28	0.20	0
	Gulwani's approach	0	17.95	4.02	0.33
	Metagol	~	~	~	~

D3

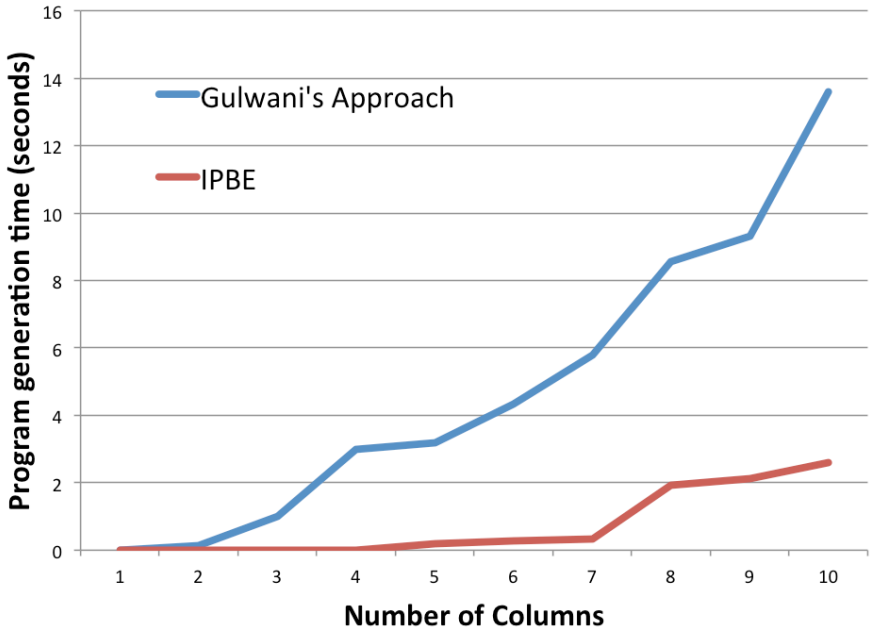


Figure: scalability test on D3

Discussion

- Our iterative PBE approach significantly reduces time in synthesizing programs

  Future work

- Extend to domains with only partial traces
- Help user to determine when to stop transforming on large datasets.

Thanks

Please come to my poster #23 for more details

Bo Wu

bowu@isi.edu

References

[Lin et al., 2014] Dianhuan Lin, Eyal Dechter, Kevin Ellis, Joshua Tenenbaum, and Stephen Muggleton. Bias reformulation for one-shot function induction. In ECAI, 2014.

[Gulwani, 2011] Sumit Gulwani. Automating string processing in spreadsheets using input-output examples. In POPL, 2011.

Different number of segments