MANCHESTER 1824
The University of Manchester

Feedback Driven Improvement of Data Preparation Pipelines

Nikolaos Konstantinou and Norman Paton

21st International Workshop On Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP 2019)
Co-located with EDBT/ICDT 2019, Lisbon, Portugal, March 26, 2019

Data Preparation

- … or data wrangling , or ETL in data warehouses

  the process of transforming data from its original form into a representation that is more appropriate for analysis

- Similar steps involved in the process
  - *Discovery*
  - *Profiling*
  - *Matching*
  - *Mapping*
  - *Format Transformation*
  - *Entity Resolution*

In this Paper

- How can feedback on the end product be used to revise the result of a multi-component data preparation process?
- Contributions
  - A technique for applying feedback that identifies statistically significant issues and explores the actions that may resolve these issues
  - A realisation of the technique in VADA (http://vada.org.uk)
  - An empirical evaluation of the implementation of the approach

Data Preparation in VADA

- Instead of handcrafting a data preparation workflow, the user focuses on expressing their requirements, and then the system automatically populates the end data product
- In particular, the user provides:
  - *Input Data Sources:* A collection of data sources that can be used to populate the result
  - *Target Schema:* A schema definition for the end data product
  - *User Context:* The desired characteristics of the end product, modelled as a weighted set of criteria
  - *Data Context:* Supplementary instance data associated with the target schema

Example

Source 1 (s_1)

street_name	postcode	city	price	location	type	bathrooms
Burnside Drive	M19 2LZ	Manchester	995	Manchester	Semi-Detached House	2 bathroom(s)
Market Street	M9 8QB	Manchester	500	*Manchester*	Apartment	1 bathroom(s)
Brightman Street	M18 8GN	Manchester	550	Manchester	Terrace House	1 bathroom(s)

Source 2 (s_2)

location	price_asked	postcode	type	bedroom_no	details	street_name
Manchester	£580 pcm	M1 5BY	Apartment	1		*Cambridge Street*
Salford	£830 pcm	M3 7EL	Apartment	3	81.50 sqm approx.	Blackfriars Road
Manchester	£625 pcm	M30 0SW	Apartment	2	50.00 sqm approx.	Devonshire Road
Salford	£720 pcm	M50 1AU	Apartment	2		Pilgrims Way
Salford	£485 pcm	M5 4TD	Apartment	2	36.30 sqm approx.	Ordsall Lane

Source 3 (s_3)

price	location	postcode	property_type	bed_num	city	street	image
£ 1 350 pcm	*Area: Botley*	OX2 9DU	3 X Bed House	3 bed	Botley	*Crabtree Rd*	DSC00195_0.JPG
£470	Cowley - OX4 3EG	OX4 3EG	Room	1 bed			
£ 1 220 pcm	*Area: Cowley*	OX4 2DU	Apartment	3 bed	Cowley	*Oxford Rd*	S1050931_0.JPG
£875	OX1 5PG	OX1 5PG	Flat	2 bed			

Source 4: English Deprivation Indices (s_4)

postcode	incomerank
OX10 1EU	29412
OX1 5PG	29540
OX28 4GE	21324
OX2 9DU	30708
OX4 2DU	9412
OX5 3DH	29567
OX7 6QE	27461
M1 5BY	25794
M18 8GN	3527
M19 2LZ	18597
M3 7EL	26678
M30 0SW	9548
M4 5HU	27939
M50 1AU	8133
M8 4QS	2734
M8 5XJ	2734
M9 8QB	2342

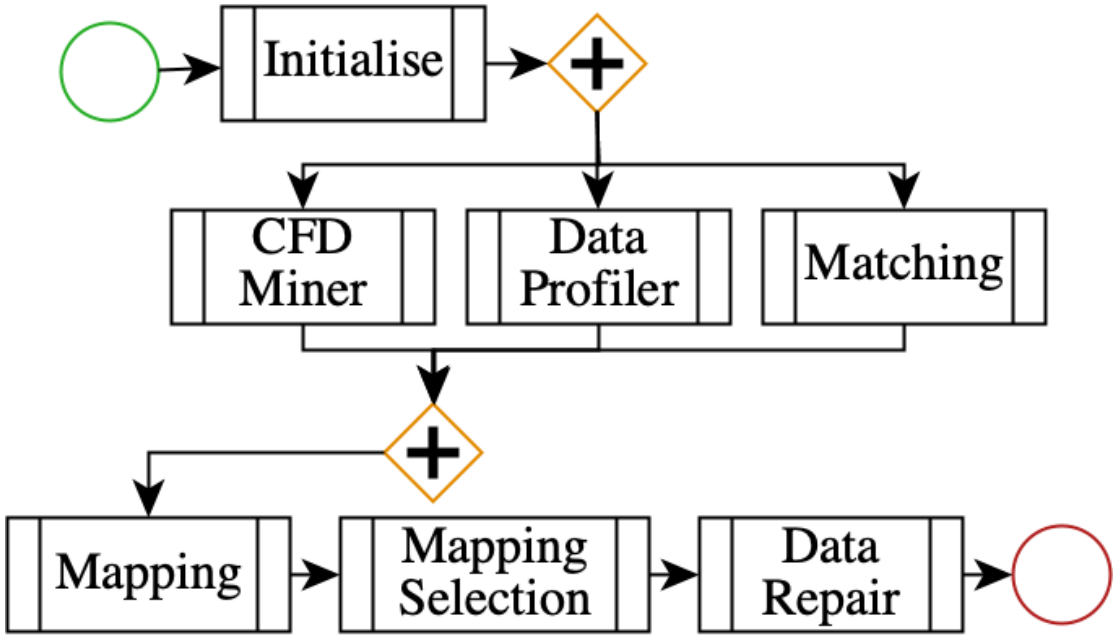
Reference Data

postcode	streetname	locality	pao
M1 5BY	*Cambridge Street*	*Manchester*	3
M1 5BY	*Cambridge Street*	*Manchester*	UNIT B2
M18 8GN	Brightman Street	Manchester	30
M26 3NL	Ashcombe Drive	Manchester	1
M3 7EL	Blackfriars Road	Salford	74
M30 0SW	Devonshire Road	Manchester	41
M50 1AU	Pilgrims Way	Salford	APT 42
M8 4QS	Delaunays Road	Manchester	5
M9 8QB	*Lakeside Rise*	*Manchester*	20
M9 8QB	*Lakeside Rise*	*Manchester*	1
OX2 9DU	*Crabtree Road*	*Oxford*	51
OX2 9DU	*Crabtree Road*	*Oxford*	47
OX28 4GE	*Thorney Leys*	*Witney*	9 PARK
OX28 4GE	*Thorney Leys*	*Witney*	17A PARK
OX4 2DU	*Oxford Road*	*Oxford*	128
OX4 2DU	*Oxford Road*	*Oxford*	132
OX5 3DH	Weston Road	Kidlington	NEW FOLD

- *Target Schema T:*
  property(price, postcode, income, bedroom_no, street_name, location)
- *User Context:* 6 criteria on attribute correctness, each with a weight of 1/6

Basic Flow of Events

- First, *Initialise* using the sources and data context that the user has provided
- Then, run *CFD Miner*, *Data Profiler* and *Matching*
- The *Mapping* component generates a set of candidate mappings, over which *Mapping Selection* evaluates the user criteria to select the most suitable mappings for contributing to the end product
- The *Data Repair* component repairs constraint violations that are detected on the end product



Using Feedback

- Refine the data preparation process
- Revised data product without the problematic values

Initial Repaired End Product

price	postcode	income	bedroom_no	street_name	location
995	M19 2LZ	18597	2 bathroom(s)	Burnside Drive	Manchester
500	M9 8QB	2342	1 bathroom(s)	*Lakeside Rise*	*Manchester*
550	M18 8GN	3527	1 bathroom(s)	Brightman Street	Manchester
£580	M1 5BY	25794	1	Cambridge Street	*Manchester*
£ 1 350 pcm	OX2 9DU	30708	3 bed	*Crabtree Road*	*Oxford*
£ 1 220 pcm	OX4 2DU	9412	3 bed	*Oxford Road*	*Oxford*

Discard match:
$s_1.bathrooms \sim T.bedroom\_no$

End Product after Collecting Feedback

price	postcode	income	bedroom_no	street_name	location
£580	M1 5BY	25794	1	Cambridge Street	Manchester
£830 pcm	M3 7EL	26678	3	Blackfriars Road	Salford
£625 pcm	M30 0SW	9548	2	Devonshire Road	Manchester
£720 pcm	M50 1AU	8133	2	Pilgrims Way	Salford
£ 1 350 pcm	OX2 9DU	30708	3 bed	*Crabtree Road*	*Oxford*
£ 1 220 pcm	OX4 2DU	9412	3 bed	*Oxford Road*	*Oxford*

Problem Statement

- Assume we have a data preparation pipeline *P*, that orchestrates a collection of data preparation steps s_1, ..., s_n, to produce an end data product *E* that consists of a set of tuples
- The problem is, given a set of feedback instances *F* on tuples from *E*, to re-orchestrate some or all of the data preparation steps s_i, revised in the light of the feedback, in a way that produces an improved end data product *E*
- Feedback takes the form of *TP* or *FP* annotations on tuples or attribute values from *E*

- Feedback Propagation:
  - *TP* tuple → all of its attribute values are marked as *TP*
  - *FP* attribute value → all tuples containing any of these attribute values are marked as *FP*

Approach

1. Form a set of hypotheses that could explain the feedback *F*

- Example: Incorrect attribute value. Possible hypotheses:
  - An incorrect match that was used to associate that value in a source with this attribute in the target
  - An incorrect mapping that was used to populate that value in the target (for example joining two tables that should not have been joined)
  - A format transformation has introduced an error into the value

2. Review all evidence to establish confidence in each hypothesis
   - Example hypothesis: incorrect match → consider together all the feedback on data derived from that match, with a view to determining whether the match should be considered problematic
3. Identify actions that could be taken in the pipeline *P*
   - Example hypothesis: Incorrect match → drop the match, or drop all mappings that use the match
4. Explore the space of candidate integrations that implement the different actions

How to Establish Confidence on a Hypothesis?

Statistical technique to test significant difference on the correctness of component products. Given:

Estimated value of criterion $\hat{c}$ on source s

$$\hat{c_s} = \frac{1}{2}(1 + \frac{tp - fp}{|s|}) \quad (1)$$

feedback

source size

…we can evaluate whether an estimated value of criterion $\hat{c}$ is significantly different between sources s_1 and s_2

statistical term measuring the relationship between a value and the mean of a group of values

$$\hat{c}_{s_2} - \hat{c}_{s_1} > z\sqrt{se_{s_2}^2 - se_{s_1}^2} \quad (2)$$

$\hat{c}_{s2}$ significantly better than $\hat{c}_{s1}$

…where se_s is the standard error

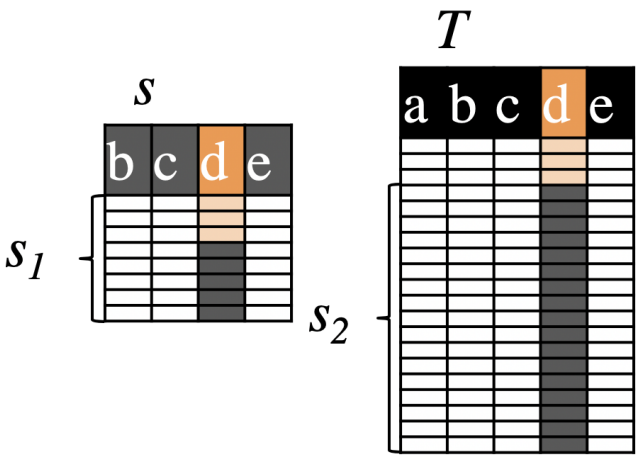
$$se_s = \sqrt{\frac{\hat{c_s}(1 - \hat{c_s})}{L_s}}$$

amount of feedback on s

Testing for Suspicious Component Products

Evaluate significant difference between s_1 and s_2 using Equation (2)

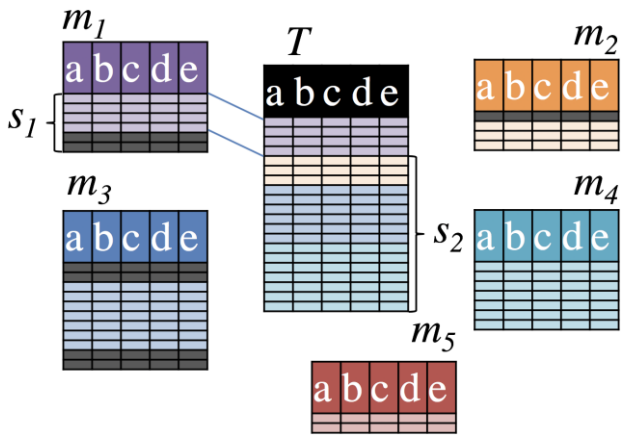
match



match: $s.d \sim T.d$

Test match: use the values from $s.d$ as s_1 and the rest of the values in $T.d$ as s_2

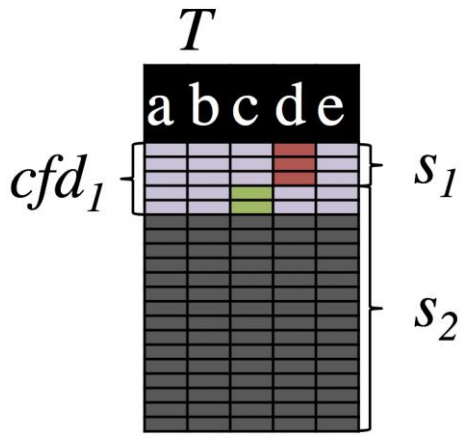
mapping



Candidate mappings m_1 to m_4 contribute to the end product

Test m_1: use the tuples from m_1 participating in the end data product as s_1 and the rest of the tuples in the end data product as s_2

repair rule



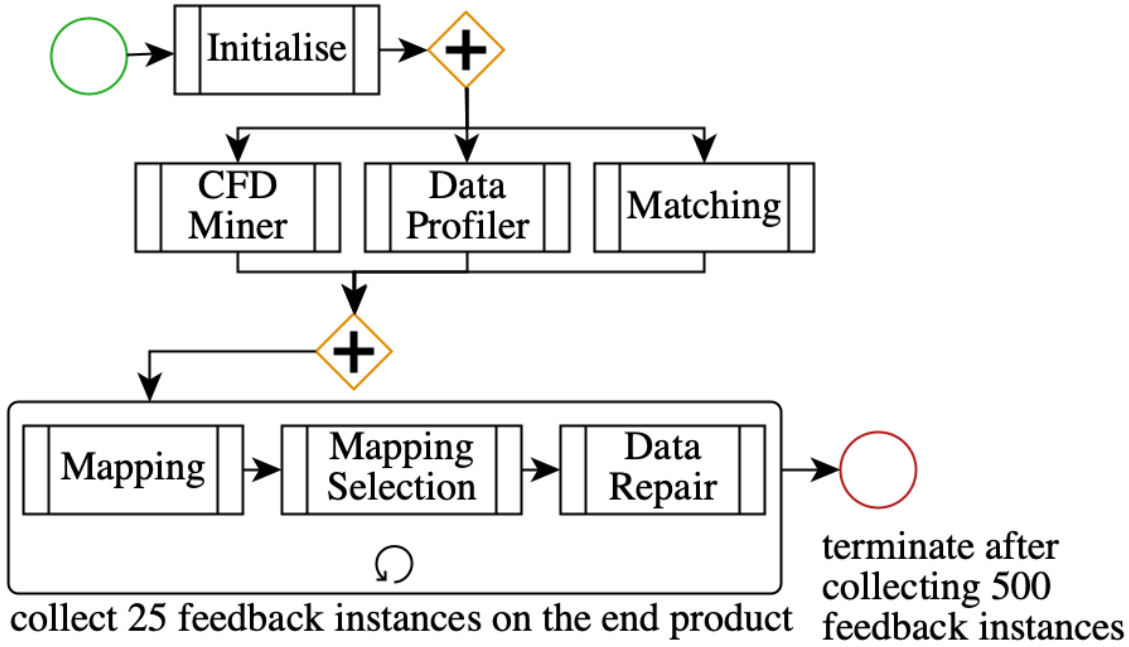
Repair rule cfd_1 has effect on 3 tuples

Test cfd_1: use the repaired tuples as s_1 and the rest of the tuples in the end data product as s_2

Experiments Setup

- Sources:
  - (a) forty datasets with real-estate properties extracted from the web
  - (b) English indices of deprivation data, downloaded from www.gov.uk
- Data context:
  - Open address data from openaddressesuk.org used as reference data
- Ground truth:
  - Manually matched, mapped, deduplicated, and then repaired an end product of approximately 4.5k tuples
- User context and target schema as in the introduction
- Component Parameters
  - Match threshold: 0.6
  - Mapping Selection: select best 1000 tuples from the generated mappings
  - Data Repair: support size set to 5

- Workflow

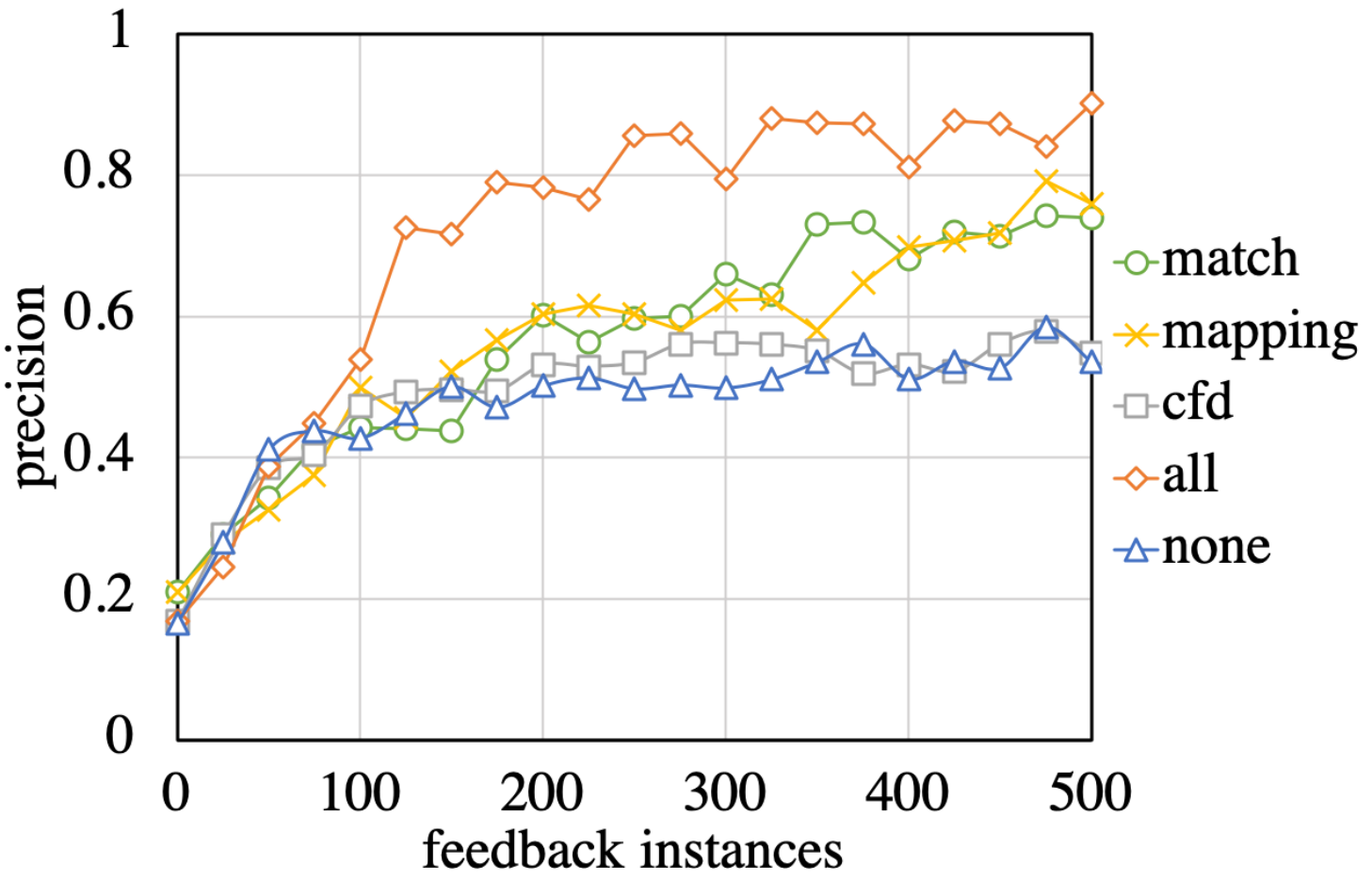


- Random feedback instances, based on the correctness of the respective tuple or attribute value wrt. the ground truth

Results

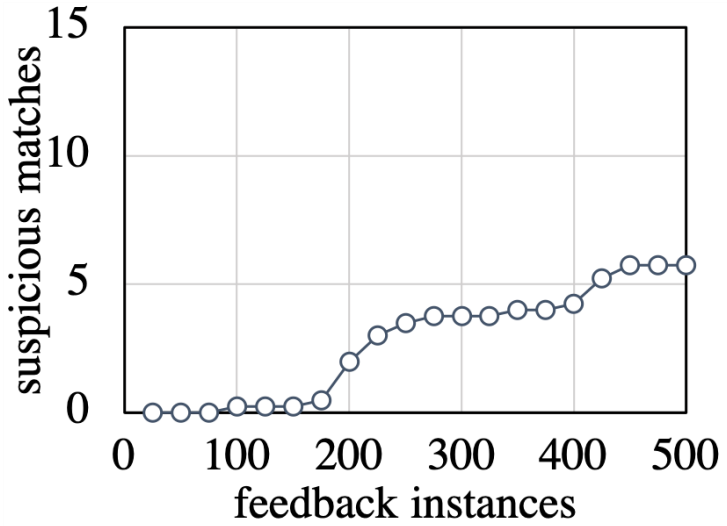
- Precision is 0.2 in the absence of feedback
- Not testing any of the components leads to a slight increase in precision because of the mapping selection component
- Matching and mapping component have approx. similar impact
- CFD component had little impact (numerous rules)
- Discarding suspicious items does not always guarantee an increase in precision

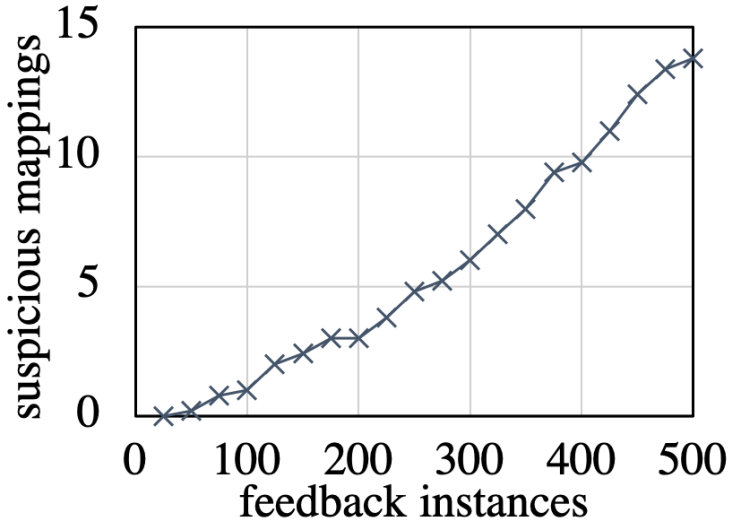
When actions across all components are considered together, the overall benefit is greater, and obtained with smaller amounts of feedback

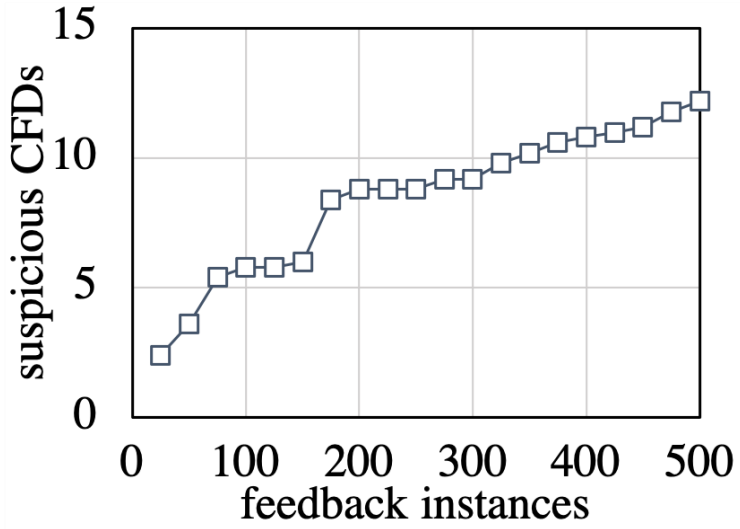


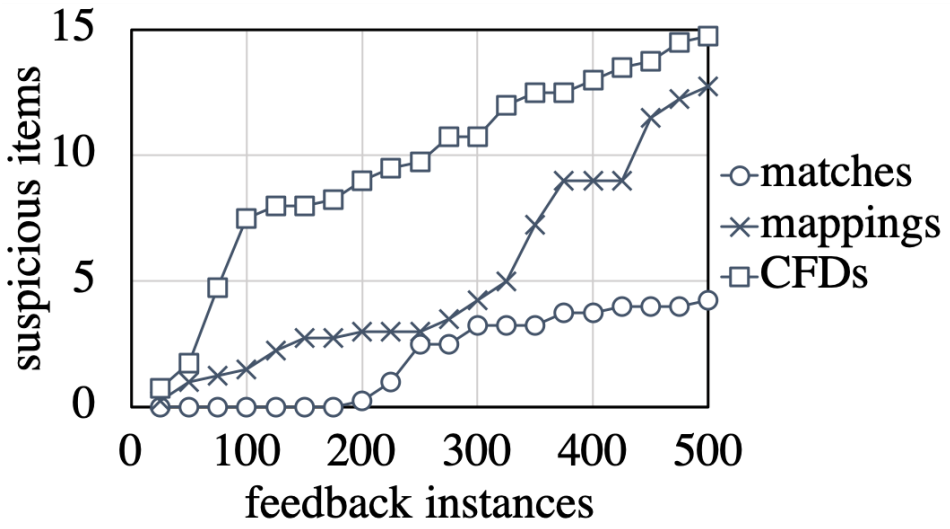
Results Breakdown

- Lines correspond to an average of 5 runs
- Few suspicious matches → substantial benefit obtained from the removal of each such match
- As matches relate to individual columns, obtaining sufficient FP feedback on the data deriving from a match can require quite a lot of feedback
- More suspicious mappings are identified, from early in the process
- Quite a few suspicious CFDs identified, although still a small fraction of the overall number (3526 in total)









Conclusions

- Hypotheses about problems with an integration are tested and acted upon using feedback on the end data product
- Approach potentially applicable to different types of feedback, components, actions
- Applied technique to matching, mapping and repair steps, in VADA
- Experimental evaluation: particularly significant benefits from the combined approach

Thank you!

Acknowledgement:
This work is funded by the UK Engineering and Physical Sciences Research Council, through the VADA Programme.