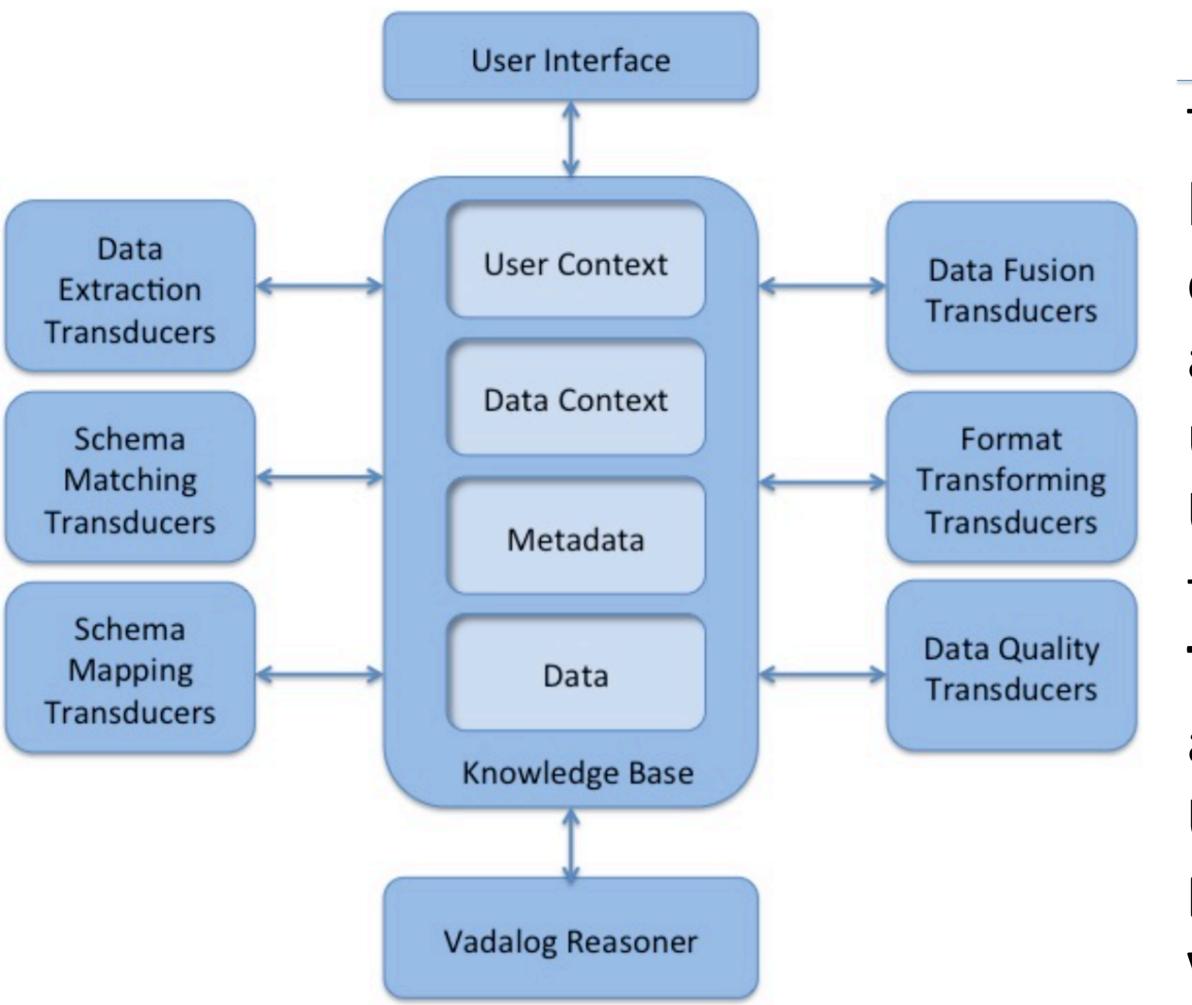
The VADA Architecture for **Cost-Effective Data Wrangling**



End-To-End Data Wrangling

The VADA architecture: (i) combines wrangling components that between them cover the complete data wrangling lifecycle; (ii) builds on automation wherever possible, using whatever information is available; (iii) refines the results of automated processes in the light of user feedback; and (iv) takes into account the user's priorities. User Interface: The data scientist provides information about the data they need, their priorities and feedback.

Transducers: Components with input and output dependencies defined as Datalog[±] rules over the knowledge base.

User Context: Information about the requirements of the user. **Data Context**: Information about the application domain.

erty.bedroom

Vadalog Reasoner: Supports reasoning over the knowledge base using Vadalog, a member of the Datalog[±] family of languages.

Carget Schema ×	Compared to the second se
\leftrightarrow \rightarrow C (i) localhost:8080/vadawebui/target-schema (i)	$\leftarrow \rightarrow \mathbb{C}$ (i) localhost:8080/vadawebui/edit-table?id=2
home target schema sources user context data context end product	home target schema sources user context data context end product
table: property 🖌 🖻	edit property
table properties type description street postcode bedrooms price crimerank	table properties type description street postcode bedrooms price crimerant
table name: property	table name: property
sources < source list create new table view end product 	 upload table sources upload reference data reference.csv upload example values containedvalues.csv upload master data
	add new field apply changes reset

Demonstration

1. Automatic Bootstrapping: The user identifies a collection of sources and defines a target schema. The system automatically orchestrates a collection of transducers that together generate an initial result data set.

2. Data context: The user associates

ome | target schema | sources | user context | data context | end produ

End Product

table:	property	0
--------	----------	---

type 🔺	description \$	street \Leftrightarrow	postcode 🔅	bedrooms 🔅	price 🕴	crimerank 🔅	
3 bedroom detache	48 Totteridge Com	Totteridge Common	OJ9 1AX	3		22809	feedback
3 bedroom detache	New detached luxu	Grange Road, Kenw		3	£119,500		feedback
3 bedroom detache	This immaculate r	Stormont Road, Lo		3	£129,500		feedback
3 bedroom detache	ESCAPE AND THE CI	Newlands Avenue,	ON6 7QT	3	£130,000	22901	feedback
3 bedroom end of	A Grade I listed	Chester Terrace,	OG3 F5J	3	£115,000	22950	feedback
3 bedroom end of	A substantial fam	Queensberry Place		3	£125,000		feedback
3 bedroom farm ho	A beautiful house	Holt End Lane, Be	OL1 F3H	3		19523	feedback
3 bedroom flat fo	Three bedroom lat	Grosvenor Crescen		3	£125,000		feedback
3 bedroom flat fo	A wonderful oppor	Moor Lane, London	OL1 1BL	3	£120,000	21971	feedback
3 bedroom flat fo	An amazing, newly	The Palace View P	OM1	3	£122,500		feedback
Showing 41 to 50 of 125 e	entries			Previous 1	4	5 6 1	13 Next

	alhost:8080/va	adawebui/user	-context#			
l	home targ	et schema	sources	user contex	t data contex	kt end product
criteria compa	arisons					
criterion 1 fi	eld 1		preferer	ice	criterion 2	field 2
completeness pr (v	roperty.crim ery strong in)	accuracy	property.type
	operty trong import	ance)			completenes	s property.bedro
completeness pr	operty.stree	et			completenes	s property.postc (moderate impo
list of criteria						
type		element		weight		
completeness	\$	property.ci	rimerank 🛟	0.223]	
accuracy	\$	property.ty	/pe 🛟	0.116		
consistency	\$	property	ŧ	0.211		
completeness	\$	property.st	treet 🛟	0.134]	
completeness	\$	property.b	edrooms 🛟	0.123]	
completeness	\$	property.p	ostcode 🛟	0.193]	

the target schema with related extents (e.g. master data). Such data allows various of the steps from bootstrapping to be revisited, including matching and mapping validation. Furthermore, it is now also possible to learn quality rules, and thereby to carry out repairs to the mapping results. The result data should now be of better quality.

3. Feedback: The user provides feedback to indicate that some of the results are correct or incorrect. Depending on the feedback provided, this will enable some of the previous steps in the wrangling process to be revisited, giving rise to a revised result.

4. User context: The result is now hopefully of reasonable quality, but the data included in the result may not be especially well-suited to the task at hand. The user specifies the user context by indicating the relative (pairwise) importance of different features in the result. The pairwise comparisons are used to derive weights that inform the selection of mappings based on multidimensional optimization.

A Real Estate Scenario

The demonstration acts on real estate data, bringing together:

Data about properties for sale from web data extraction over

deep web sources.

Open government data that provides information about the areas in which the properties are located.



The University of Manchester



EPSRC VADA: Value Added Data Systems – Principles and Architecture 🞇 www.vada.org.uk

