Value Added Data systems:

An architecture for end-user informed data preparation

As the myriad of data sources continues to grow, so does the need for cost-effective, scalable and principled techniques for integrating and cleaning big data in order to optimise the data quality and thus increase its value. Dr Norman Paton, Professor of Computer Science and Dr Nikolaos Konstantinou, a Research Fellow with the Department of Computer Science, both from the University of Manchester, are carrying out research into Value Added Data systems (VADA) and exploring various techniques for automating the creation of data preparation processes in their development of a cost-effective automated end-to-end data wrangling process.

Data science involves the analysis and synthesis of large amounts of data, exploring and solving complex problems in order to obtain insights from the data. Surveys reveal, however, that data scientists can spend as much as 80% of their time preparing data for analysis. This data preparation, or data wrangling, is expensive and labour intensive because it consists of a number of steps such as web extraction, source selection, data integration and data cleaning. Intense manual involvement in each of these steps makes data preparation as a whole a process that requires both significant skill and time.

DATA PREPARATION

A number of data preparation approaches are currently in widespread use. These tend to fall into three groups: those that involve programming the solution; those that develop workflows that extract, transform and load the data into analysis platforms; and those that develop transformations using tabular representations of the data. These data preparation tools usually offer components that support similar tasks, such as combining data sets and reformatting columns, but they differ in how these can be expressed by data scientists. Even with the support of these tools, data scientists are required to retain fine-grained control over each aspect of the process. While this is appropriate in some circumstances, the costs are high and can be prohibitive.

AUTOMATING THE CREATION OF DATA PREPARATION PROCESSES

As the myriad of data sources within organisations and in the public domain continues to grow, so does the need for cost-effective, scalable and principled techniques for integrating (addressing the variety) and cleaning (addressing the veracity) of big data to create data sets suitable for downstream analysis, thereby obtaining value from the data. This prompts the question: to what extent can the creation of data preparation processes be automated? Dr Norman Paton, Professor of Computer Science at Manchester University, and Dr Nikolaos Konstantinou, a Research Fellow at the University of Manchester, are striving to answer this question. They are carrying out research into Value Added Data systems (VADA) and exploring various techniques for automating data preparation. In the automated approach, the data scientists describe what they need, and the VADA software develops a plan for producing that data from the available sources.

VALUE ADDED DATA SYSTEMS (VADA)

The research team acknowledge that in the big data era, data wrangling must overcome the challenges of the four Vs (volume, velocity, variety and veracity) if the fifth V (value), the reward, is to be achieved. In order to automate the creation of a data wrangling process, some evidence is required to inform a search for suitable approaches to prepare the data. Dr Konstantinou explains that "a key feature of VADA is that the automation takes account of the data context and the user context. The data context is supplementary data about the expected result of the data wrangling process. The user context is information about what is important to the user, as likely there are trade-offs between different features of the result, such as the consistency and the completeness".

As the myriad of data sources continues to grow, so does the need for cost-effective, scalable and principled techniques for integrating and cleaning big data in order to optimise the data quality and thus increase its value.

VADA provides an architecture (a set of components, and techniques for sharing data among them and coordinating their evaluation) to automate the data preparation process. The user provides: some data sources; a schema for the data target i.e. a description of the structure of the required data; some example data; and the criteria to be prioritised when populating the target, so that alternative results can be ranked. Armed with this information, the system automatically populates the data target from the sources.

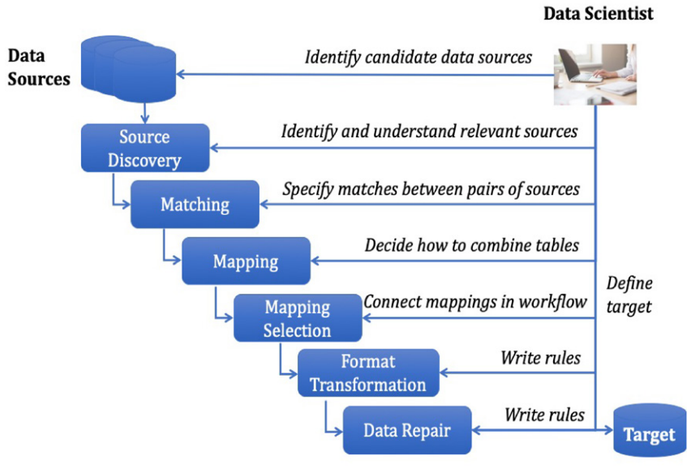
The resulting automated solution may not be suitable or even correct and the user can provide feedback as to the correctness or suitability of the result. In the light of the new evidence, the system can then act on this feedback and automatically generate a revised approach to populating the target. This process is repeated until the result is deemed fit for purpose.

The research team have observed that automating data wrangling in this way involves being able to identify the steps in the data preparation process that can be automated using available evidence and feedback. They are then able to develop components for these steps that take all of the available evidence and feedback into account. This leads to the development of an overall architecture that allows the user to provide the evidence, view the results, and then provide feedback.

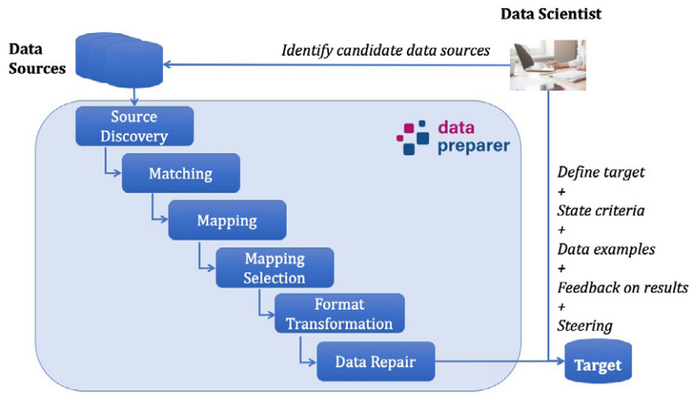
EVALUATION OF VADA

The researchers have carried out an empirical evaluated of VADA using a case study involving real world Web extracted data from real-estate sources

As the myriad of data sources continues to grow, so does the need for cost-effective, scalable and principled techniques for integrating and cleaning big data.



A typical data preparation process has many steps that involve detailed user involvement.



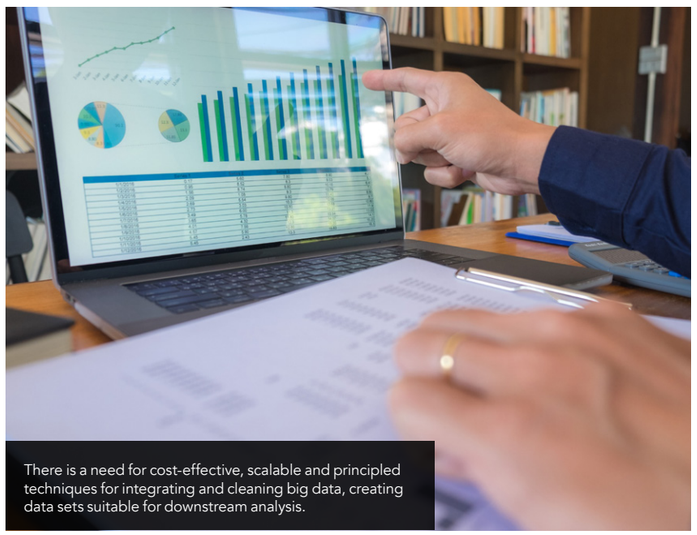
In the VADA data preparation process, users need only specify what they need, not how to produce it.

A key feature of VADA is that the automation takes account of the data context and the user context.

and the UK open government data portal. These data sets are produced by a large number of independent publishers, so inconsistencies need to be resolved in order to maximise their potential value through analysis. The researchers ran a five-step automated wrangling process. They then compared the results with and without data context. Using the f-score to combine precision and recall, examples showed considerable improvements with f-scores rising from around 0.5 to 0.8 when data context was used as evidence throughout the wrangling process.

VADA BENEFITS

These encouraging results show that with the VADA user interface, data scientists can obtain cleaned and integrated data from multiple data sets, having provided only a target schema and associated data context. The process requires only modest effort when compared to the intense manual involvement that is currently required. For example, the automatically produced wrangling process combines data sets, reformats inconsistent attribute values, and resolves certain inconsistencies. Manually crafting queries and rules to carry out such tasks requires significant skill and effort.



There is a need for cost-effective, scalable and principled techniques for integrating and cleaning big data, creating data sets suitable for downstream analysis.

VADA also takes user preferences and user feedback into account. By capturing preferences through the user context, data can be selected in ways that trade-off the accuracy, consistency and relevance of the wrangling result. Furthermore, the automation means that many alternative candidate data products can be produced. The evaluation of the user context demonstrates how feedback can impact on the utility of the results. Moreover, this approach can wrangle hundreds of sources in minutes. This scale of data preparation would likely take many days of manual effort.

SPIN-OUT COMPANIES

There are two spin-out companies that translate different aspects of this work into practice.

Firstly, The Data Value Factory (https://thedatavaluefactory.com) which offers automated data preparation. The Data Preparer system builds on the research team's experience and provides the first declarative data wrangling software product. A free trial version of the Data Preparer wrangling platform is available for download.

Secondly, DeepReason.ai (https://deepreason.ai) enables 'Knowledge-First' AI Solutions with a Knowledge Graph platform that uses state-of-the-art AI technology to deliver end-to-end AI solutions to Fortune 500 companies in finance, logistics, manufacturing, and engineering.

FUTURE WORK

The University of Manchester research team are exploring how they can combine automated data preparation with discovery of relevant data sets in data lakes. In addition, automating data preparation creates opportunities to explore how the introduction of bias during data preparation can detected and reduced.

Behind the Research

Norman Paton

E: norman.paton@manchester.ac.uk **T:** +44 1612756910
W: https://www.research.manchester.ac.uk/portal/norman.paton.html

Nikolaos Konstantinou

E: nikolaos.konstantinou@manchester.ac.uk **T:** +44 1612756183
W: https://personalpages.manchester.ac.uk/staff/nikolaos.konstantinou/

Research Objectives

Research at the Information Management Group at the University of Manchester focuses on distributed information management for challenging environments and applications.

Detail

Department of Computer Science
University of Manchester
Oxford Road
Manchester M13 9PL, UK

Bio

Norman Paton: Norman has been a Professor of Computer Science at Manchester University since 2000, and is now a Founder/Director at The Data Value Factory, working to commercialise techniques on cost-effective data preparation. His research has focused on distributed information management, including applications in the life sciences.

Nikolaos Konstantinou: Nikos has been a Research Fellow at Manchester since 2015, prior to which he held a variety of technical management and research roles in Greece. He is a Founder/Director at The Data Value Factory, working to bring innovative data preparation techniques to market.

Funding

EPSRC, Grant Title - VADA: Value Added Data Systems - Principles and Architecture (EP/M025268/1).

Collaborators

- Edward Abel, Alex Bogatu, Martin Koehler, Lacramioara Mazilu, Alvaro Fernandes, John Keane, School of Computer Science, University of Manchester
- Luigi Bellomarini, Emanuel Sallinger, Georg Gottlob, Department of Computer Science, University of Oxford
- Cristina Civili, Leonid Libkin, School of Informatics, University of Edinburgh

References

Abel, A., Keane, J.A., Paton, N.W., Fernandes, A.A.A., Koehler, M., Konstantinou, N., Ríos, J.C.C., Azuan, N.A., Embury, S.M. (2018) User driven multi-criteria source selection. *Inf. Sci.* 430: 179-199. https://doi.org/10.1016/j.ins.2017.11.019

Koehler, M., Abel, E., Bogatu, A., Civili, C., Mazilu, L., Konstantinou, N., Fernandes, A., Keane, J., Libkin, L. & Paton, N. (2019) Incorporating Data Context to Cost-Effectively Automate End-to-End Data Wrangling, *IEEE Transactions on Big Data*. https://doi.org/10.1109/TBDATA.2019.2907588

Konstantinou N., Abel E., Bellomarini L., Bogatu A., Civili C., Irfanie E., Koehler M., Mazilu L., Sallinger E., Fernandes A.A.A., Gottlob G., Keane J.A., Paton N.W. (2019) VADA: an architecture for end user informed data preparation. *J. Big Data*, 6 p74, https://doi.org/10.1186/s40537-019-0237-9

Personal Response

What initially sparked your interest in automating data wrangling?

"We worked over a considerable period with life science researchers, who often need to combine experimental results with existing information about an organism. In these collaborations, it became clear that the cost of preparing the data for analysis was a significant barrier to progress."

MANCHESTER 1824
The University of Manchester