Automating Data Preparation: Can We? Should We? Must We?

Norman Paton

University of Manchester

The University of Manchester

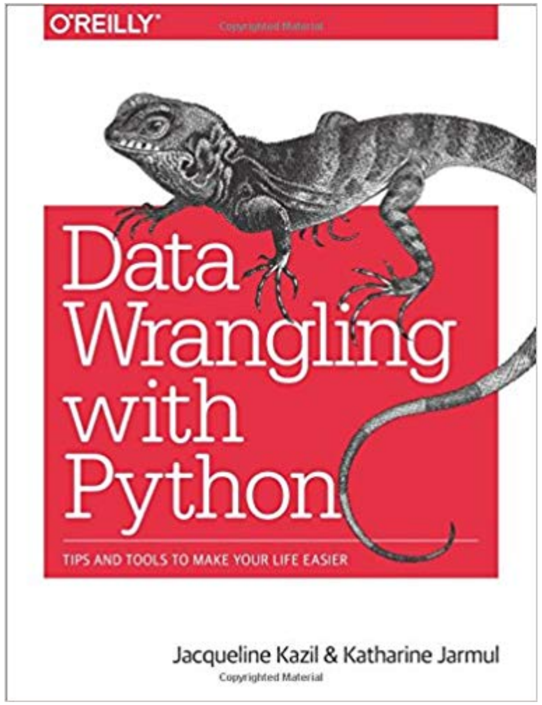
Data Preparation / Data Wrangling

- Definitions:
  - *a process of iterative data exploration and transformation that enables analysis* [1].
- Data preparation often takes 80% of a data scientist's time [2].

[1] S. Kandal, *et al*., Research Directions in Data Wrangling: Vizualizations and Transformations for usable and credible data, *Information Visualization*, 10(4).
[2] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#26ee60816f63

What is Automated?

- The term *automate* is used to refer to:
  1. The repeated execution of a program that carries out a data preparation task.
  2. The authoring of the program that carries out a data preparation task.
- The focus of this presentation is (2).
- With (2), you get (1) as well.



Approaches to Data Preparation

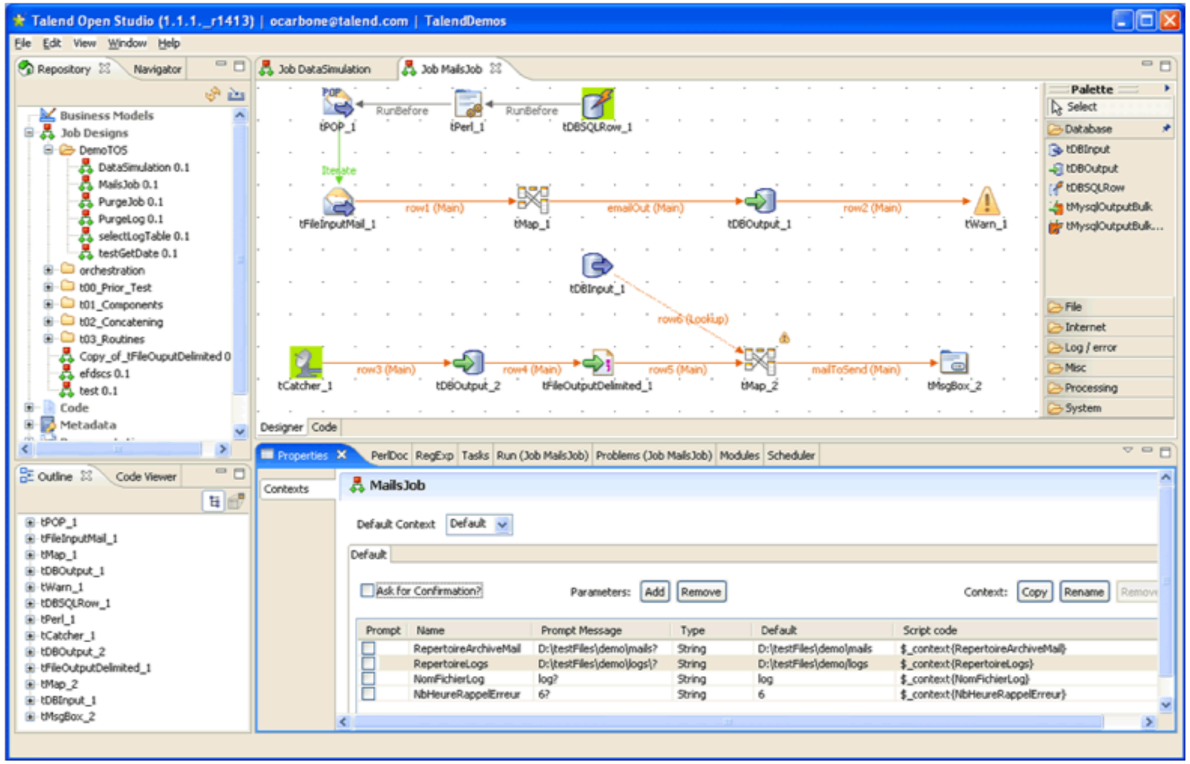
- Products in the $3B data preparation tools market focus on supporting data scientists in writing data preparation programs.
- All approaches carry out similar data preparation tasks, but differ in *how* the user interacts with the system.

Approach	User Interaction	Products
Workflow based	Users manually connect and configure components that combine and clean data sets.	Informatica, Talend, Pentaho
Dataset based	Users interact with spreadsheet like interfaces, applying transformations to individual data sets.	Trifacta, Open Refine, Datawatch
Target based	Users describe what they would like, and the system works out how to produce it.	*Automated proposals are here.*

MANCHESTER
1824
The University of Manchester

Workflow Based

- Extract, Transform and Load (ETL) tools have been around for a significant time.
- ETL tools support source wrapping, warehouse population, data joining, etc.
- ETL vendors also have “big data” offerings.

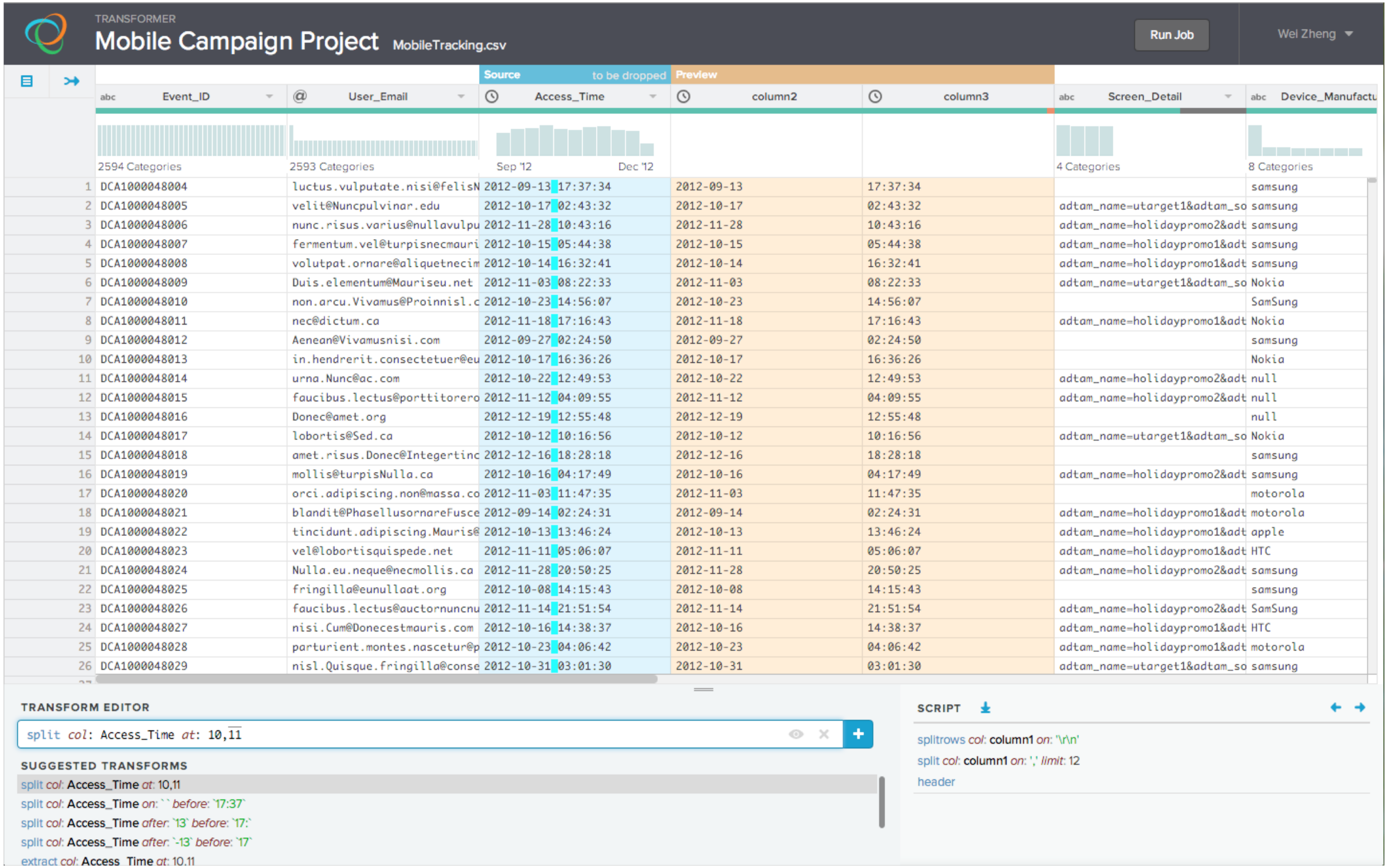


www.talend.com

MANCHESTER 1824

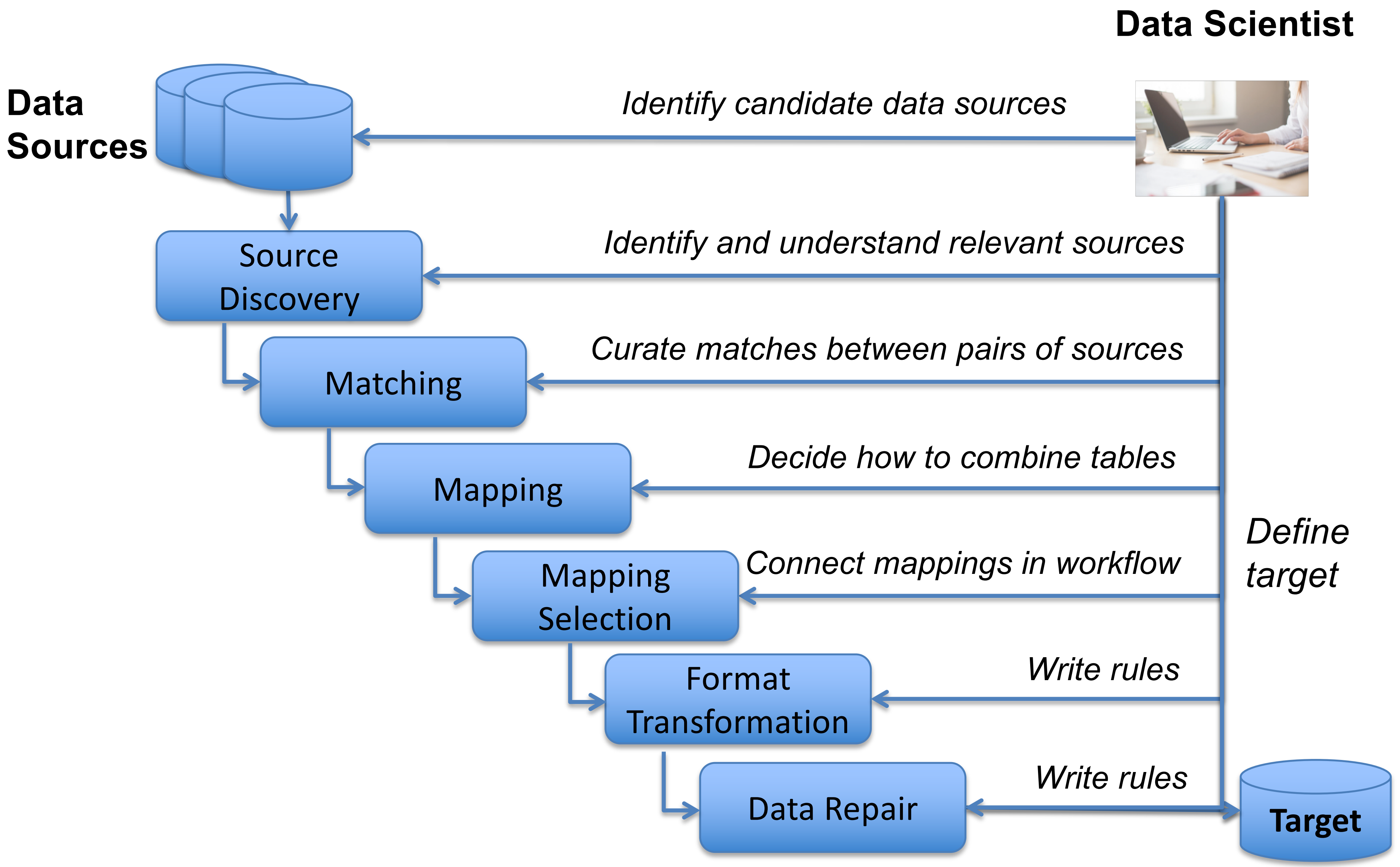
The University of Manchester

Dataset Based: Trifacta



https://www.trifacta.com/

Existing Approaches



Why is this expensive?

- There are many different steps.
- Some of these are technically challenging:
  - Mapping generation, format transformation, …
- Some of these need done for individual sources:
  - Format transformation.
- Some of these need understanding of many sources:
  - Matching, mapping generation, entity resolution.
- The data scientist takes fine-grained control over each of the steps and their combination.

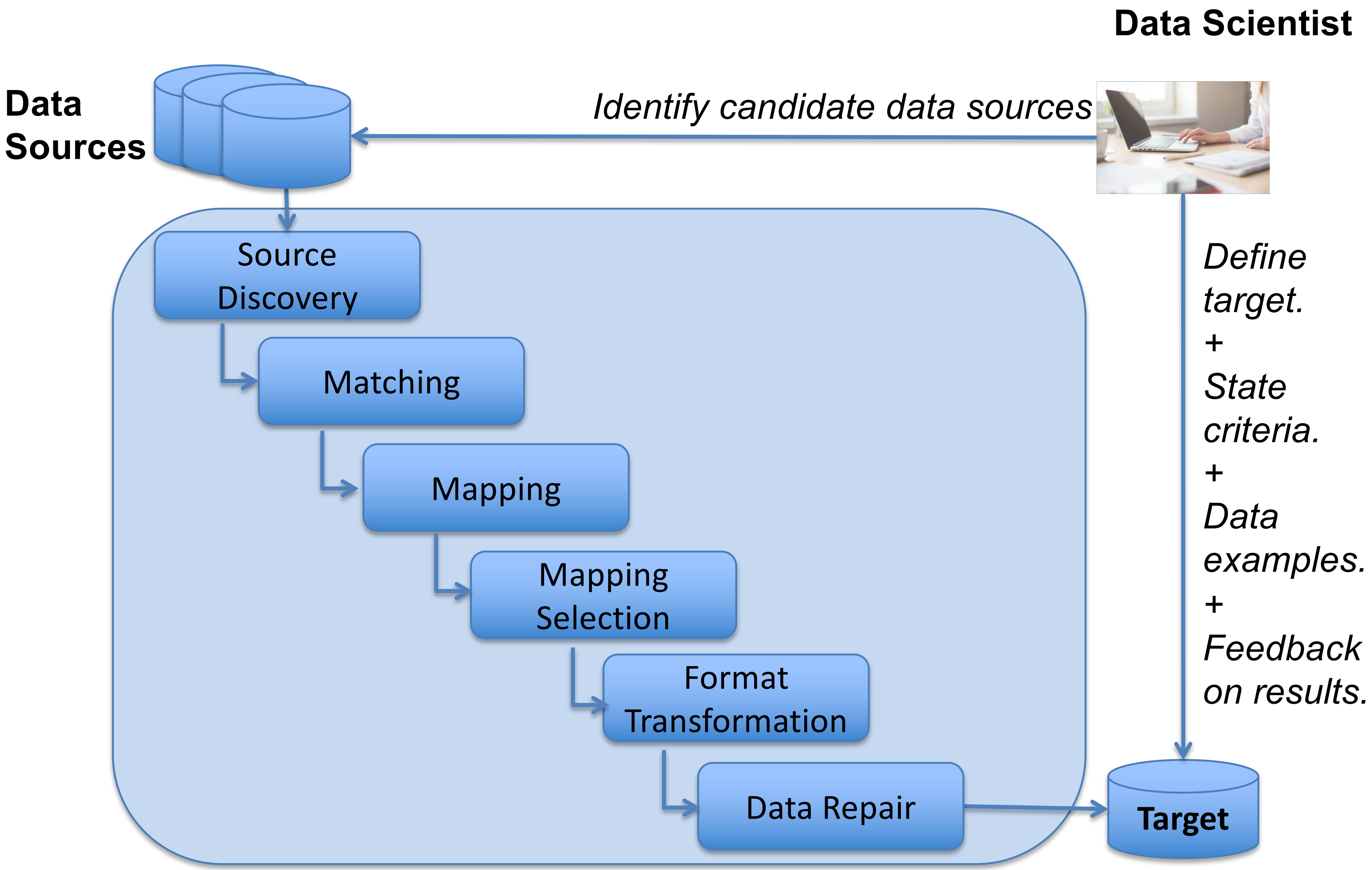
What about automation?

- The hypothesis is that automated approaches should adopt the following principle:

Data preparation systems should involve the description of what is required, and not the specification of how it should be obtained.

MANCHESTER 1824
The University of Manchester

How might this look?



the data value factory

The University of Manchester

Demo

- The demonstration is of DataPreparer, the MVP of our spin-out, The Data Value Factory.
- The research behind this has taken place within the EPSRC VADA project.
  - Nikolaos Konstantinou, Martin Koehler, Edward Abel, Cristina Civili, Bernd Neumayr, Emanuel Sallinger, Alvaro A. A. Fernandes, Georg Gottlob, John A. Keane, Leonid Libkin, Norman W. Paton: The VADA Architecture for Cost-Effective Data Wrangling. SIGMOD Conference 2017: 1599-1602.

So are we done?

- DataPreparer is an early example of end-to-end automation for data preparation. Next:
  - ***Can we?*** To what extent is it understood how data preparation can be automated?
  - ***Should we?*** To what extent can we be confident that automation will be effective?
  - ***Must we?*** In what circumstances is there no option but to automate?

Can we automate?

- There are many steps in data preparation.
- How many of them can be automated?
- What evidence is needed to inform automation?
  - *Bootstrapping*: evidence that can be used to produce an initial result.
  - *Improvement*: evidence that can be used to refine the initial result, using feedback.

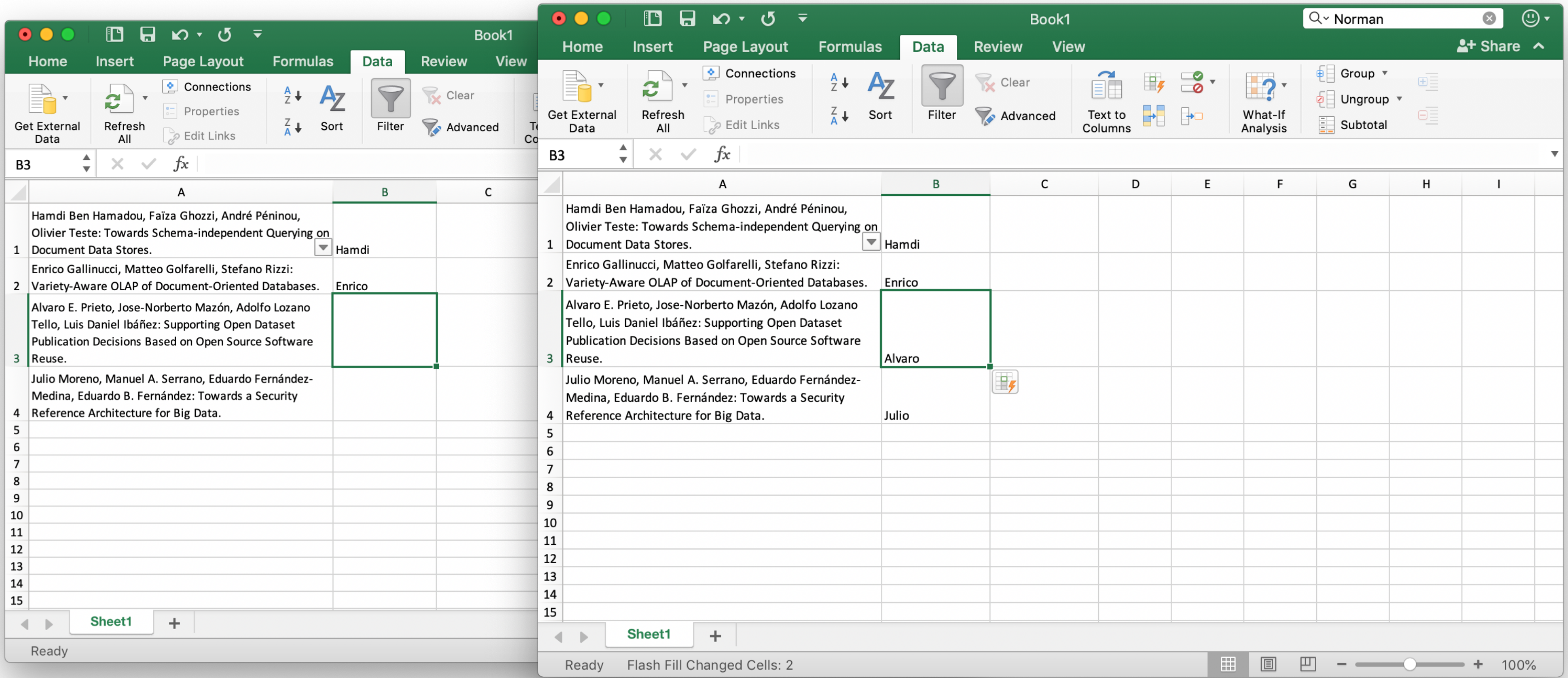
The University of Manchester

Examples of Automation

Stage	What is Automated	Evidence Used	Citation
Data discovery	The search for unionable data sets	Illustrative target examples	[26]
Data extraction	The creation of extraction rules	Training examples, feedback	[11]
Format transformation	The synthesis of transformation rules	Training examples pairs	[15]
Mapping generation	The generation of mappings	Target examples	[28]
Data repair	The generation of repair rules	Master data	[13]
Duplicate Detection	Generation of rules and thresholds	Correctness feedback	[25]

Format Transformation

- Here FlashFill extracts the first names of DOLAP first authors.
- Examples are provided in row B, which is then auto-filled by clicking the FlashFill icon.



How easy was that?

- Pretty easy, but:
  - What if there were a million rows?
  - What if there are a thousand sources?
  - What else can we do with these examples?
  - I typed an example wrongly, with confusing results!
  - My attempt at extracting the surname didn't work.

Sumit Gulwani, William R. Harris, Rishabh Singh: Spreadsheet data manipulation using examples. Commun. ACM 55(8): 97-105 (2012).

The University of Manchester

Questions for Methods

- Where does the evidence come from? Better to discover than ask users.
- How specific to the method is the evidence? Better to apply each piece of evidence several times.
- How does the method scale? FlashFill program synthesis is exponential on number of examples and high quadratic on example size.

- Alex Bogatu, Norman Paton, Alvaro Fernandes, Martin Koehler, Towards Automatic Data Format Transformations: Data Wrangling at Scale, The Computer Journal, https://doi.org/10.1093/comjnl/bxy118, 2018.
- Alex Bogatu, Alvaro Fernandes, Norman Paton and Nikolaos Konstantinou SynthEdit: Format transformations by example using edit operations, EDBT, 2019.

Reuse of Evidence

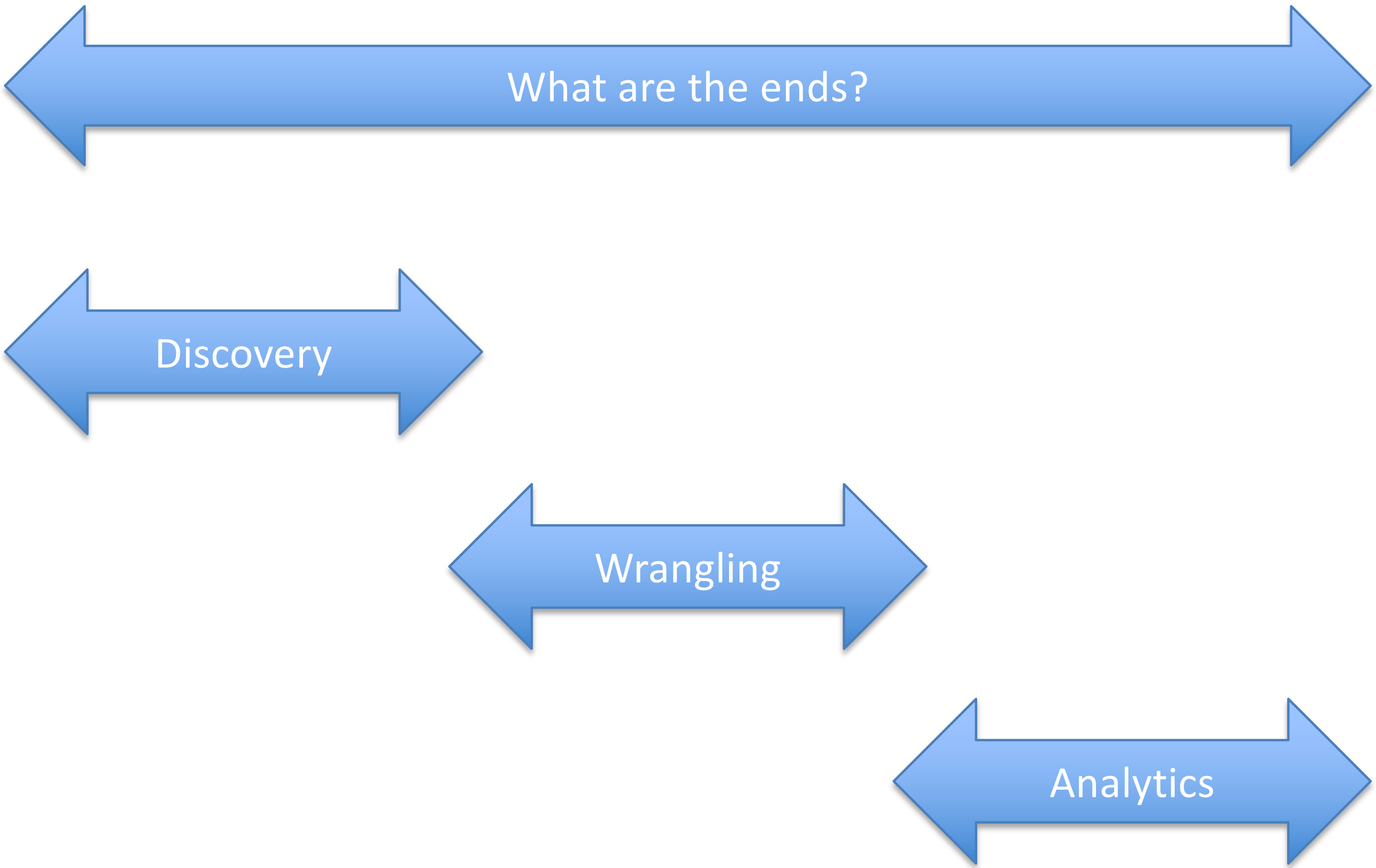
- Some types of evidence can inform several data wrangling steps:
  - Representative result examples.
  - Actual result examples.
  - True/false positive annotations on results.
- Is this true for other types of evidence?

- Martin Koehler, Alex Bogatu, Cristina Civili, Nikolaos Konstantinou, Edward Abel, Alvaro A. A. Fernandes, John A. Keane, Leonid Libkin, Norman W. Paton: Data context informed data wrangling. IEEE BigData 2017: 956-963.
- Nikolaos Konstantinou and Norman W. Paton, Feedback Driven Improvement of Data Preparation Pipelines, DOLAP, 2019.

Feedback

- Feedback has been used to refine the results of many of the earlier tasks, including:
  - Data extraction: correct / incorrect results.
  - Mapping generation: correct / incorrect results.
  - Entity resolution: correct / incorrect pairs.
- Questions for feedback proposals:
  - What else can be done with the collected feedback?
  - How much feedback is needed? Highly variable!
- Automation can generate many alternatives; feedback can be used to choose between them.

End-to-End Proposals



Another Example: Data Tamr

- Tamr aims to bring together key records (parts, customers, suppliers) from across complex enterprises.
- Tamr uses example data plus feedback to categorise attributes, and uses domain experts to refine categories and integration results.
- In Tamr, technical and domain experts contribute to curating data, for example providing examples and feedback.

https://www.tamr.com/

Comparing End-to-End Approaches

- There aren't very many end-to-end approaches that have automation at the core.
- Tamr and VADA/DataPreparer have a similar scope, and follow our earlier principle:

Data preparation systems should involve the description of what is required, and not the specification of how it should be obtained.

- But they are rather different technically, and engage with users differently.
- Likely there are other ways in which end-to-end automation of data preparation can surface.

Should we automate?

- There are now quite a few results on automation.
- Even if you are not targeting end-to-end automation, surely one should automate the steps where automation can do better than an expert.
- Is there evidence as to when this is the case?
  - Not much one way or another …
- Do some tasks look very hard manually?
  - Yes – think about co-optimizing parameters for entity resolution.

Ruhaila Maskat, Norman W. Paton, Suzanne M. Embury: Pay-as-you-go Configuration of Entity Resolution. T. Large-Scale Data- and Knowledge-Centered Systems 29: 40-65 (2016).

Data Extraction

- The problem is to write regular expressions for extracting substrings from text (e.g. URLs, dates, phone numbers).
- Compared a Genetic Algorithm (with 24 training examples) to student users, who self-classified as to their experience writing regular expressions.
- In most cases, the learned extraction rules performed somewhat better than the human users.

Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, Fabiano Tarlao: Can a Machine Replace Humans in Building Regular Expressions? A Case Study. IEEE Intelligent Systems 31(6): 15-21 (2016)

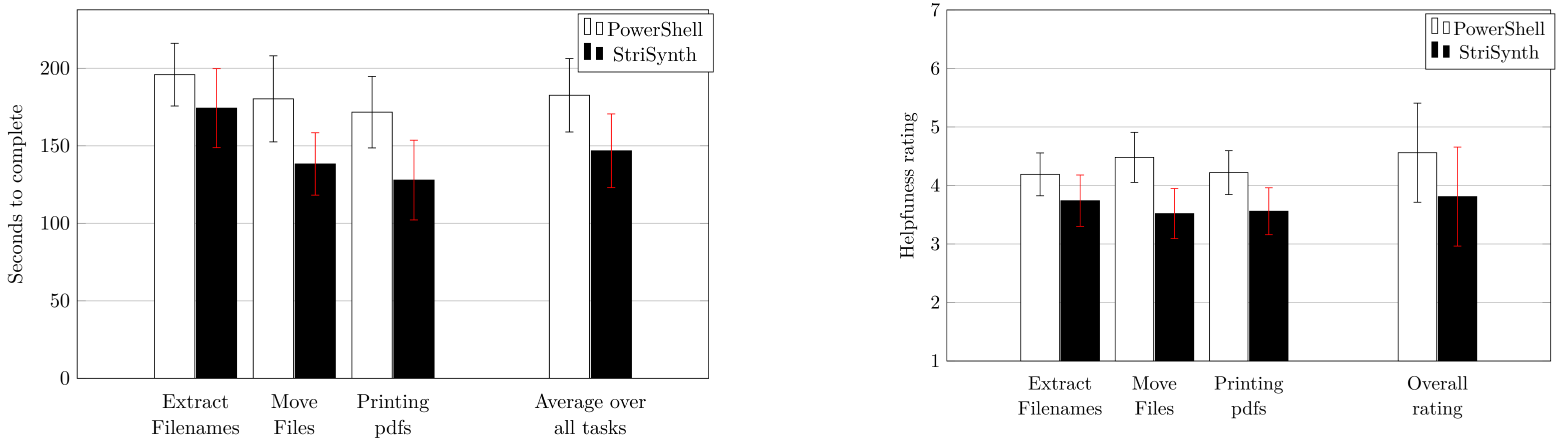
Should we semi-automate?

- Note that semi-automation is typically nothing like automation, in that the user is supported in writing rules, violating our principle.
- An experimental study with proactive suggestions for format transformations yielded mixed results.
  - Proactive suggestions were often ignored.
  - Some proposals were tried and then dropped.
  - The presence of proactive suggestions did not have a significant effect on completion times.

Philip J. Guo, Sean Kandel, Joseph M. Hellerstein, Jeffrey Heer: Proactive wrangling: mixed-initiative end-user programming of data transformation scripts. UIST 2011: 65-74

Should we automate?

- There is a shortage of evidence comparing manual and automated approaches to individual tasks, far less to end-to-end processes.
- This seems like a good topic for further research; in what ways is it most productive to have the human in the loop?



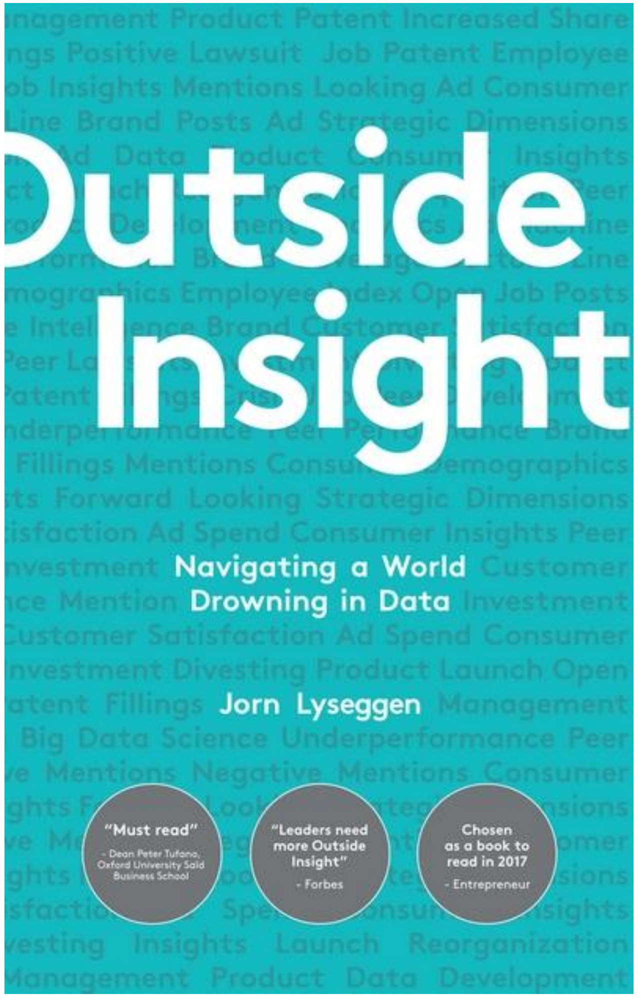
M. Santolucito, D. Goldman, A. Weseley, R. Piskac, Programming by Example: Efficient, but Not "Helpful", 9th Workshop on Evaluation and Usability of Programming Languages and Tools, 2019.

Must we Automate?

- Too much data.
  - The data lakes market is predicted to grow at 28% compound growth rate to $14B by 2023 (www.marketresearchfuture.com/reports/data-lakes-market-1601)
- Not enough resource.
  - 95% of the information economy business in the UK employ fewer than 10 people (www.gov.uk/government/publications/information-economy-strategy).
  - Ability to transform data without programming is an important requirement for end user data preparation (https://www.datawatch.com/wpcontent/uploads/2017/03/2017-End-User-Data-Preparation-Market-Study.pdf)

New Opportunities

- Future data analysis is not sure to be like past data analysis.
- Outside insight is about understanding your business in relation to external data.



Conclusions

- Automating data preparation:
  - Can we?
    - Significant progress has been made, but is typically not joined up.
  - Should we?
    - Automation should be able to compete with experts for a variety of data preparation tasks.
    - Empirical evidence as to when automation is effective, trusted or appreciated is not plentiful.
  - Must we?
    - The current technologies seem not to be up to handling emerging opportunities – ever more data cannot be tackled by labour-intensive techniques.

Acknowledgements

This work is funded by the UK Engineering and Physical Sciences Research Council, through the VADA Programme Grant on Value Added Data Systems: Principles and Architecture: vada.org.uk.