Artificial Intelligence for Data Analytics – Final Project Report

PI: Christopher K I Williams
School of Informatics, University of Edinburgh, UK
The Alan Turing Institute, London, UK

March 11, 2022

The Artificial Intelligence for Data Analytics (AIDA) project at the Turing was concerned with data engineering (aka data wrangling). This is a rather under-researched area, despite often being laborious and time-consuming, and accounts for up to 80% of the effort in a typical data science project.

The AIDA project was led by Prof Chris Williams (PI, Edinburgh), with co-investigators Dr James Geddes (Turing), Prof Zoubin Ghahramani (Cambridge), Prof Ian Horrocks (Oxford), Dr Tomas Petricek (Kent), Dr Charles Sutton (Edinburgh). The main research activity was been carried out by the three project RAs, Drs Ernesto Jiménez-Ruiz, Alfredo Nazabal and Gerrit van den Burg, along with Chris Williams' PhD student Taha Ceritli. We also carried out a lot of work with the Turing Research Engineering Group (REG), with both data scientists and software engineers.

It is important to realize that predictive models (such as deep neural networks) are just one part of the data analytics process. This process starts with a problem specification, asks what data are available to address this question, then prepares the data before carrying out predictive modelling. After this the model needs to be evaluated and then deployed. This is not a linear process, there can be many feedback loops that require revisiting earlier steps.

We find that human insights can be critical for the process of data engineering. Thus we do not aim to fully automate data engineering but instead focus on 'semi-automated' tools that keep the human in the loop, and let the analyst guide the overall process and provide key insights.

The project has been successful against all of its objectives, which were to:

- build AI assistants for individual tasks,
- to build an open-source platform (Wrattler) and integrate the assistants into the platform, and
- to provide exemplar use cases of real-world data wrangling.

More than 20 papers have been published, along with associated code and datasets. The individual assistants include topics such as as probabilistic type inference (ptype), unioning of pairs of related datasets (datadiff), outlier detection and repair (RVAE), reading messy CSV files (CleverCSV), and semantic annotation of data columns (ColNet).

The scope of AIDA is best summarized by our review paper "Data Engineering for Data Analytics: A Classification of the Issues, and Case Studies" by Nazabal et al (arXiv, 2020). This paper provides a classification of data engineering problems appearing in messy datasets when a data scientist faces an analytical task. We have identified three high-level groups of problems: Data Organization issues (DO), related to obtaining the best data representation for the task to be solved, Data Quality issues (DQ), related to cleaning corrupted entries in the data, and Feature Engineering (FE) issues, related to the

creation of derived features for the analytical task at hand. Additionally, we have further divided the DO and DQ groups according to the nature of data wrangling problem they face. Under Data Organization we include data parsing (DP), data dictionary (DD), data integration (DI) and data transformation (DT). Under Data Quality we include canonicalization (CA), missing data (MD), anomalies (AN) and non-stationarity (NS). The AIDA project has made contributions in almost all of these areas.

Members of the AIDA team have given many talks at academic conferences and workshops, and to industry, to disseminate the work of the project. For example, Chris Williams gave a talk in March 2021 to the AIUK conference, the UK's national showcase of artificial intelligence (AI) and data science research and collaboration.

AIDA benefited from various funding sources: starter funding from Lloyds Register Foundation (Nov 2016-Apr 2017), funding from the UK Government's Defence & Security Programme in support of the Alan Turing Institute (Apr 2017-Mar 2019), and from the Alan Turing Institute (Apr 2019-May 2021) under EPSRC grant EP/N510129/1.

Papers

- Sutton, C, Hobson, T, Geddes, J, & Caruana, R., Data Diff: Interpretable, Executable Summaries of Changes in Distributions for Data Wrangling. Knowledge Discovery and Data Mining Conference 2018, London, United Kingdom.

- Petricek, T, Geddes, J & Sutton, C. Wrattler: Reproducible, live and polyglot notebooks. 10th USENIX Theory and Practice of Provenance, London, United Kingdom, 2018

- Christopher K. I. Williams, Charlie Nash, Alfredo Nazabal. Autoencoders and Probabilistic Inference with Missing Data: An Exact Solution for The Factor Analysis Case. Posted on arXiv 11 Jan 2018, `https://arxiv.org/pdf/1801.03851.pdf`.

- Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks and Charles Sutton (2019). ColNet: Embedding the Semantics of Web Tables for Column Type Prediction. The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19).

- Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks and Charles Sutton (2019). Learning Semantic Annotations for Tabular Data. Proc. 28th International Joint Conference on Artificial Intelligence (IJCAI 2019).

- Jiaoyan Chen, Ernesto Jiménez-Ruiz and Ian Horrocks. Canonicalizing Knowledge Base Literals. International Semantic Web Conference (ISWC-19), 2019

- Van den Burg, G. J. J., Nazabal, A., and Sutton, C. (2019). Wrangling messy CSV files by detecting row and type patterns. *Data Mining and Knowledge Discovery*, 33(6), pp. 1799-1820.

- Ceritli, T., Williams, C.K.I. & Geddes, J. (2020) ptype: probabilistic type inference. *Data Mining and Knowledge Discovery*, 34(3), pp. 870–904.

- Handling incomplete heterogeneous data using VAEs. Alfredo Nazabal, Pablo M. Olmos, Zoubin Ghahramani, Isabel Valera. *Pattern Recognition*, 107, 107501, November 2020.

- Simao Eduardo, Alfredo Nazabal, Christopher K. I. Williams, Charles Sutton. Robust Variational Autoencoders for Outlier Detection in Mixed-Type Data. In 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020. PMLR Vol 108.

- Mark Collier, Alfredo Nazabal, Christopher K.I. Williams. VAEs in the Presence of Missing Data. Published on arXiv 13 July 2020. Presented at the first Workshop on the Art of Learning with Missing Values (Artemiss) hosted by the 37th International Conference on Machine Learning (ICML 2020).

- van den Burg, G.J.J. and Williams, C.K.I., 2020. An Evaluation of Change Point Detection Algorithms. Submitted for publication, arXiv preprint arXiv:2003.06222.

- Alfredo Nazabal, Christopher K.I. Williams, Giovanni Colavizza, Camila Rangel Smith, Angus Williams. Data Engineering for Data Analytics: A Classification of the Issues, and Case Studies. Published on arXiv 27 April 2020, `https://arxiv.org/pdf/2004.12929.pdf`.

- Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, Kavitha Srinivas. SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. ESWC 2020: 514-530.

- Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, Kavitha Srinivas, Vincenzo Cutrona: Results of SemTab 2020. SemTab@ISWC 2020.

- Probabilistic Sequential Matrix Factorization. Omer Deniz Akyildiz, Gerrit van den Burg, Theodoros Damoulas, Mark Steel. Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, PMLR 130:3484-3492, 2021.

- ptype-cat: Inferring the Type and Values of Categorical Variables. Taha Ceritli and Christopher K. I. Williams. Presented at the ECML-PKDD Workshop on Automating Data Science, 17 Sept 2021. arXiv:2111.11956.

- Identifying the Units of Measurement in Tabular Data. Taha Ceritli and Christopher K. I. Williams. Presented at the ECML-PKDD Workshop on Automating Data Science, 17 Sept 2021. arXiv:2111.11959.

- Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks, Denvar Antonyrajah, Ali Hadian, Jaehun Lee: Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. Extended Semantic Web Conference (ESWC) 2021: 392-408

- Jiaoyan Chen, Pan Hu, Ernesto Jiménez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, Ian Horrocks: OWL2Vec*: embedding of OWL ontologies. *Machine Learning* 110(7): 1813-1845 (2021).

- Erik B. Myklebust, Ernesto Jiménez-Ruiz, Jiaoyan Chen, Raoul Wolf, Knut Erik Tollefsen. Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings. Semantic Web (2021).

- On Memorization in Probabilistic Deep Generative Models. Gerrit J. J. van den Burg, Christopher K. I. Williams. In Proc. NeurIPS 2021.

- INDIGO: GNN-Based Inductive Knowledge Graph Completion Using Pair-Wise Encoding. Shuwen Liu, Egor Kostylev, Bernardo Cuenca Grau and Ian Horrocks. In Proc. NeurIPS 2021.

- Automating Data Science, T. De Bie, L. De Raedt, J. Hernández-Orallo, H. Hoos, P. Smyth, C. K. I. Williams. *Communications of the ACM* 65(3) 76-87, March 2022.
- AI Assistants: A framework for semi-automated, accountable, tooling-rich data wrangling. Tomas Petricek, Gerrit J.J. van den Burg, Alfredo Nazabal, Taha Ceritli, Ernesto Jiménez- Ruiz, Christopher K. I. Williams. Submitted for publication.

Software

Much of the work carried out has produced code. See:

- Wrattler, a polyglot notebook environment for exploratory data science: `https://github.com/wrattler/wrattler`
- ptype: Probabilistic type inference `https://github.com/alan-turing-institute/ptype`
- Robust Variational Autoencoders for Outlier Detection and Repair of Mixed-Type Data `https://github.com/sfme/RVAE_MixedTypes`
- CleverCSV, a Python package for handling messy CSV files: `https://github.com/alan-turing-institute/CleverCSV`
- SemAIDA: Semantic Technologies for the AIDA project: `https://github.com/alan-turing-institute/SemAIDA/`
- OWL2Vec* code: `https://github.com/KRR-Oxford/OWL2Vec-Star`
- Datadiff: `https://github.com/alan-turing-institute/datadiff`
- Probabilistic Unit Canonicalizer: `https://github.com/tahaceritli/puc`
- Data Science in ecotoxicology: `https://github.com/NIVA-Knowledge-Graph`
- On Memorization in Probabilistic Deep Generative Models: `https://github.com/alan-turing-institute/memorization`

Datasets

- A repository with the data wrangling challenges addressed in four case studies in the review paper: `https://github.com/alan-turing-institute/aida-data-engineering-issues`
- The Turing Change Point Dataset - A collection of time series for the evaluation and development of change point detection algorithms: `https://github.com/alan-turing-institute/TCPD`
- SemTab: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching: `https://www.cs.ox.ac.uk/isg/challenges/sem-tab/`