The Alan Turing Institute

Project page template for Turing website

- All fields indicated with an asterisk* are required
- Use third person impersonal, e.g. "this project aims to" not "we aim to" etc
- Send form to website@turing.ac.uk, and cc: research@turing.ac.uk, for review

Example pages

Probabilistic numerics
Capturing complex data streams
Improving cystic fibrosis healthcare
Distributed training for machine translation

Project title*

(Aim for 6 words or fewer. A clear description of the project, can be different to official academic name of the project, please avoid acronyms/initialisms)

Data science toolkit for explorable data visualisations

Project page main contact*

Name: Tomas Petricek

Email address (Optional): T.Petricek@kent.ac.uk

Timeframe*

Start date: 02 / 01 / 19

End date: 31 / 05 / 19

Research page summary*

(1 sentence, present tense, e.g. Using..., Developing..., Investigating...)

Helping scientists understand data analyses by connecting parts of visualisations to the data they depend on

Accessible introduction/summary*

(Clear, concise, ~3 sentences – e.g. 1st sentence: the problem being addressed, 2nd sentence: the potential solution/method, 3rd sentence: applications, output)

Data visualisation is essential to modern science and science communication, but has been criticised for hiding patterns in the raw data, obscuring statistical assumptions and distorting effect size. Using dependency-tracking techniques, we are developing a transparent data visualisation framework which integrates provenance information directly into visualisations, allowing scientists to interactively explore how parts of diagrams are derived from data and/or code. In phase two of the project, we will allow users to explore the impact of changes in a dataset or model on the associated visualisations, so they can assess the implications of modelling decisions by considering alternative scenarios.

Main aims of the project*

(What is the work hoping to achieve? What would define success? Why is this work worth doing?)

The aim of the project is to bring substantially greater transparency to data visualisation. In a notebook system such as Jupyter, this will involve automatically linking diagrams to data and code so that fine-grained dependencies can be explored interactively. For example, moving the mouse over part of a chart will highlight which specific data elements were relevant. A secondary goal to provide this transparency to readers as well as authors, developing technology that makes it is possible to create interactive web content (e.g. online versions of papers) containing charts which provide transparent access to the raw data and underlying code. We believe this is essential to realising the vision of open science, and addresses the growing recognition of the need for more statistically robust, transparent approaches to data visualisation.

In the second phase of the project, we will build on our dependency-tracking approach to allow scientists to explore alternative data or modelling scenarios interactively, obtaining visual feedback on precisely how a chart might change as a consequence of analysing or manipulating data in a different way. Our research hypothesis is that interpreting a model or analysis properly involves understanding counterfactual scenarios – how the model would behave under different assumptions or different inputs.

Applications*

(Where is this work being applied, what area/industry could it benefit?)

We are developing a proof-of-concept implementation which we will integrate into the [Wrattler] ((https://www.turing.ac.uk/research/publications/wrattler-reproducible-live-and-polyglot-notebooks)) notebook being developed here at The Alan Turing Institute. We are looking at application areas in urban analytics, geocomputation, and data-centric engineering, but the techniques and software we are developing should be applicable in any domain that makes use of data visualisation. We also intend to integrate our technique with related approaches to explainability in machine learning, such as guided backpropagation.

Explaining the science*

(Is there theory or methods that would be good to explain to understand the project's work better? Use plain English where possible)

This project adapts techniques from programming languages to the data visualisation domain. In phase 1, we build on prior work in dynamic program analysis. The approach we use, called [Galois slicing](https://dl.acm.org/citation.cfm?id=3110258), formalises a two-way relationship between parts of programs and parts of outputs, identifying the "least" part of the program which is able to account for a given output feature, and conversely the "least" output feature explained by a given part of the program.

Here, we extend this idea to charts and other graphics by defining a simple visualisation language suitable for use by data scientists, and implementing a dependency-tracking interpreter which tracks the fine-grained forward and backward dependencies between parts of images and related code/data. This dependency information will also form the foundation for phase 2 of the project, which will allow the exploration of "counterfactual scenarios" in the form of (hypothetical) model or data changes.

Recent updates

(Achievements/project milestones reached since project started, with month/year)

The team presented the work at a Tools, Practices and Systems workshop in Cambridge in April 2019. You can view our poster from the workshop [here](https://imgur.com/NrXXx6p).

Project leaders*

Title	Name	Organisation/University affiliation	Role on project
Dr	Tomas Petricek	University of Kent; Turing Fellow	PI
Dr	Roly Perera	The Alan Turing Institute	Co-I
Dr	James Cheney	University of Edinburgh; Turing Fellow	Co-I

Researchers

Title	Name	Organisation/University affiliation	Role on project

Collaborating organisations/universities

Name	Role (e.g. funder, collaborator, data supplier etc)
University of Edinburgh	Collaborator

Research areas*

(Please tick the research areas that are **most** applicable, up to approx 10)

- [ ] **Algorithms**
- [ ] Complexity
- [ ] Compression
- [ ] Cryptography
- [ ] Data structures
- [ ] Numerical

- [ ] **Applied mathematics**
- [ ] Dynamical systems & differential equations
- [ ] Information theory
- [ ] Mathematical physics
- [ ] Multi-agent systems
- [ ] Numerical analysis
- [ ] Operations research

- [ ] **Artificial intelligence**
- [ ] Automation
- [ ] Collective behaviour
- [ ] Control theory
- [ ] Evolution & adaptation
- [ ] Game theory
- [ ] Neural networks
- [ ] Neuroscience
- [ ] Nonlinear dynamics
- [ ] Pattern formation
- [ ] Systems theory

- [ ] **Computer systems & architectures**
- [ ] Communications
- [ ] Databases
- [ ] Distributed parallel & cluster computing
- [x] Human computer interface
- [ ] Information retrieval
- [ ] Neural & evolutionary computing
- [ ] Networks
- [ ] Operating systems
- [ ] Real time computing
- [ ] Visualisation

- [ ] **Machine learning**
- [ ] Applications
- [ ] Computer vision
- [ ] Deep learning
- [ ] Natural language processing
- [ ] Pattern recognition
- [ ] Reinforcement learning
- [ ] Supervised learning
- [ ] Semi-supervised learning
- [ ] Unsupervised learning
- [ ] Speech recognition

- [ ] **Mathematical modelling**
- [ ] Automata & algebraic
- [ ] Deterministic/ non-deterministic
- [ ] Dynamic/static
- [ ] Ensemble
- [ ] Graph theory
- [ ] Logics
- [ ] Stochastic

- [ ] **Optimisation**
- [ ] Convex programming
- [ ] Nonlinear programming
- [ ] Stochastic optimisation

- [x] **Privacy & trust**
- [ ] Cryptography
- [ ] Differential privacy
- [ ] Identity management
- [ ] Verification

- [ ] **Programming languages**
- [ ] Hardware optimisation (FPGA/GPU)
- [ ] Probabilistic programming
- [x] Software framework development
- [x] Visualisation

- [ ] **Social data science**
- [ ] Cognitive science
- [ ] Data science of government & politics
- [ ] Developmental psychology
- [ ] Ethics
- [ ] Linguistics
- [ ] Management science
- [ ] Networks
- [ ] Research methods
- [ ] Social media
- [ ] Social psychology

- [ ] **Statistical methodology**
- [ ] Bayesian inference
- [ ] Causality
- [ ] High dimensional inference
- [ ] Monte Carlo methods
- [ ] Non-parametric & semi-parametric methods
- [ ] Simulation
- [ ] Time series

- [ ] **Statistical theory**
- [ ] Asymptotic
- [ ] Estimation theory
- [ ] Information theory
- [ ] Modelling
- [ ] Probability

- [ ] **Theoretical mathematics**
- [ ] Algebra
- [ ] Calculus & analysis
- [ ] Combinatorics
- [ ] Geometry & topology
- [ ] Logic
- [ ] Number theory

Application areas

(Please tick as many as applicable)

- [ ] Culture and media
- [x] Data centric engineering
- [ ] Data science at scale
- [ ] Defence and security
- [x] Economic data science
- [x] Government
- [x] Health and wellbeing
- [x] Urban analytics

Additional content

(If there is any additional content that should be included, or doesn't fit in the previous pages, please add it here. If there any images, videos, or figures (with plain English captions) that would be helpful in communicating the project please include these separately when submitting this form)

Click to add text