Explorable, Self-Explaining Data Visualisations

What is your work package title?

See above

Lead Investigator:

Roly Perera (The Alan Turing Institute)

Co-Investigators:

Tomas Petricek (University of Kent, Turing Fellow), James Cheney (University of Edinburgh, Turing Fellow), Benjamin Bach (University of Edinburgh), Tanu Malik (DePaul University), Levi Wolf (University of Bristol, Turing Fellow)

Name of PI/Co-I present at scoping workshop (or Turing Director):

Roly Perera, Tomas Petricek, James Cheney

Earliest possible start date:

1 June 2019

Latest possible end date:

31 December 2020

Duration (in Months):

12

Abstract (500 words):

Data visualisation is essential to data science and science communication, but is open to both misinterpretation and misuse: patterns in raw data can be obscured, statistical assumptions hidden, and effect sizes misrepresented [16]. These concerns can be addressed in part through improved statistical practices and better plot and chart designs [1], but also by making visualisations themselves more open and explorable [6].

Understanding a visualisation requires grasping how it relates to the underlying data and other visualisations. For example, geoscientists often work with multiple layered views. To show how these are related, spatial analytics applications like GeoDa [2] can automatically select the relevant part of one view as the user changes the selection in a related view, say a choropleth map. However, this feature is available only if it was specifically anticipated by the application or library developer; if the geoscientist uses custom libraries or wants other views linked that the developer did not consider, they are out of luck.

Our project will deliver a framework for authoring visualisations where support for linking, between data, code, and visualisations is built in, making this powerful comprehension feature automatic. To allow us to focus on the needs of SPF projects working within the Urban Analytics theme, we have assembled a team with expertise in geocomputation and spatial analytics (Wolf, Malik). This is complemented with expertise in data visualisation (Bach, Wolf), data provenance (Cheney, Perera, Malik), and programming languages (Cheney, Perera, Petricek). The following three tracks will run in parallel, building on a proof-of-concept also developed within the TPS programme; see https://www.turing.ac.uk/research/research-projects/data-science-toolkit-explorable-data-visualisations.

Track 1: Domain use cases

[Wolf, Malik, Bach, Perera, 0.5 FTE RA] We will develop a public repository of substantive geospatial visualisation use cases that demonstrate the explorability features of our approach to prospective users. This will ensure that the work in track 1 stays focused on real needs of researchers in the urban analytics and other geospatial domains.

Track 2: Infrastructure for explorable data visualisations

[Perera, Petricek, Cheney, Malik, Bach, 0.5 FTE RA] This will involve user interface and library design, and new programming language infrastructure. Driven by our domain use cases (track 1), we will further develop our toolkit, building on a dynamic program analysis technique developed by Perera and Cheney [9, 13]. We will combine the linking feature with the ability to explore the pipeline of data and visual transformations that yielded the final artefact, for example by tabbing through the various intermediate artefacts. We call the resulting explorable visualisations "self-explaining" because they can be explored in situ (say in an online paper) to reveal how the visualisation was created.

Track 3: Wrattler integration

[Perera, Petricek, 0.5 FTE RSE] We will integrate our toolkit with the Wrattler notebook [11]; this strand of work will also be supported by the "Evolving Wrattler into the Turing tools platform" work package. This is central to our impact strategy and how we connect to other Turing SPF projects and is discussed in detail later.

Route to Impact (300 words):

Workshop on explorable data visualisation. To reach out to prospective users in the Urban Analytics and Data Science for Science SPF programmes within the Turing, we will hold a hands-on workshop/tutorial, using Wrattler, on explorable visualisations. Bach is giving a workshop on data visualisation at the Turing on 31 May; this will serve as a useful networking exercise and precursor to our event.

Online gallery of examples/interactive essays. To generate excitement beyond the Turing network, we will publish our domain use cases as an online suite of examples showcasing our approach. We will encourage contributions from the community, both within and without the Turing. In collaboration with other relevant TPS projects, we will also aim to publish an interactive research paper in Google's Distill [7] journal, a high-profile venue for non-traditional research artefacts, describing the analysis techniques and design principles underlying our framework.

Standalone self-explaining visualisations. We will make it possible to deploy our visualisations as standalone charts or infographics that can be embedded into web pages or online papers, following the approach taken by Petricek for The Gamma [10] and leveraging Malik's expertise in reusable research objects [18], substantially increasing the potential reach of our approach beyond practising data scientists to the wider community.

Naturally, we will also disseminate our research through the usual academic channels. We will aim for a significant publication at a programming languages, data visualisation or human-computer interaction conference (perhaps POPL, VIS or CHI), and one at a data science or quantitative geography/geovisualisation venue (perhaps SciPy or NACIS). The software will be released under an open source licence through the Turing's public GitHub repository.

Research challenges (300 words):

The project is an unusual blend of programming languages, data visualisation and HCI research, and faces technical, usability, and interoperability challenges. We have tried to anticipate these with an eclectic but synergistic team.

The basic technical challenge of our project is implementing infrastructure for "linking", or multiple-coordinated views [15], in an application-independent way. We are taking a principled approach based on our own prior work on dynamic dependency analysis and provenance; in

contrast to prior work on provenance in data visualisation [5], our approach is much more fine-grained and is able to associate specific bits of data or code with parts of a visualisation in a precise way. While the theoretical technique is proven, this approach has never been applied to data visualisation before.

A central usability challenge is visualising these complex relationships between the various parts of a visualisation and the relevant data and/or visualisation code. This is essentially a higher-order visualisation problem: visualising information about the provenance of visualisations. We will draw on Bach's expertise in temporal data visualisation [3] and data-driven storytelling [4], and Cheney, Malik and Perera's background in data provenance. We may also build on recent work on "literate" visualisation [17].

Finally, we will need to solve challenging interoperability problems in order to coexist usefully with popular visualisation libraries such as Bokeh [8], visual specification languages like Vega [14], spatial analytics libraries such as GeoPandas and PySAL [12], and notebooks like Wrattler and Jupyter. To use our analysis techniques with such third-party frameworks we may need to propose new metadata formats or find efficient ways to instrument existing libraries; Wolf's experience with developing open source spatial analytics software and Petricek's Wrattler expertise will be useful for exploring appropriate techniques.

Community benefit (300 words):

The project involves a new collaboration between Kent (Petricek), Edinburgh (Cheney, Bach) and Bristol (Wolf), with Turing Fellows from all three partner institutions contributing, as well as DePaul University in the US (Manik) and the London Turing office (Perera, Petricek). A wide range of career stages and disciplines are represented in our investigator team. Our unique selling point is bringing techniques and perspectives from programming languages to bear on problems in data visualisation and data science; we see this as a fertile area for future interdisciplinary research and community-building.

Our work package will bring together Turing researchers interested in explorability and explainability in data science, and methodological and ethical issues surrounding data visualisation in particular, such as trustworthiness, transparency, pedagogy and digital literacy. Our initial efforts will focus on urban analytics researchers from the Digital Twins SPF programme, but we will also seek engagement from the Data Science for Science programme, and other SPF projects dependent on visualisation.

Beyond the Turing network, we envisage our framework being used as part of the various external projects and collaborations we are involved with. Wolf holds a fellowship at the Center for Spatial Data Science at University of Chicago and is co-maintainer of the PySAL spatial analytics library and lead author of Python libraries for geovisualisation. Bach has collaborated with Glasgow's Urban Big Data Centre, and historians and archaeologists in the UK, France and Luxembourg. Cheney has been involved in digital curation and is a research leader in data provenance. Malik's SciUnit tool is integrated with Hydroshare, a social platform for sharing code and data serving more than 3500 geoscientists. Finally, Perera and Petricek are involved

in conference and workshop organisation in live programming, a related (and hot) topic in programming languages. These will provide a wide variety of opportunities for community outreach and involvement.

Connections to other activities (300 words):

We see important synergies with at least the following 3 TPS project proposals:

1. Evolving Wratter into the Turing Tools Platform (Tomas Petricek)

Integrating our framework into Wrattler as a new cell type will significantly lower the barriers to entry for a data scientist wishing to experiment with our visualisations, since they will be able to add them easily to an existing a Wrattler-based notebook written in R or Python. The synergy should flow in the other direction too: while Wrattler already has excellent support for coarse-grained dependency-tracking and provenance, our toolkit will show the advantages of fine-grained dependency tracking, as well as fully transparent visualisations. This should also make Wrattler more appealing to potential users.

2. The Turing Way (Kirstie Whitaker)

Our project is motivated by concerns surrounding trust and traceability in data visualisation, important methodological issues for science communication. While developing new visualisation methodology is not an explicit goal of the project, we are keen to be informed by current methodological thinking. Moreover, in the longer term, we hope to influence developing methodologies, not only by delivering tools that enable more transparency out of the box, but also by developing a suite of visual design patterns, libraries and programming practices that take advantage of these new tools.

3. Explainability and Interpretability in AI (Emmanouil Benetos)

Benetos' project aims to consolidate and develop explainability techniques for machine learning, such as guided backpropagation, and is highly complementary to ours. They too are concerned with tracking input-output relationships in a fine-grained way, but for neural networks rather than symbolic programs. By composing their approach with ours, it might be possible to provide end-to-end explainability for pipelines with a mixture of computational styles (say conventional visualisation libraries mixed with CNNs). We believe the two projects would benefit from a certain amount of temporal overlap.

References and URLs:

Project pilot: https://www.turing.ac.uk/research/research-projects/data-science-toolkit-explorable-data-visualisations

Put a screenshot online?

Investigator home pages

Perera:	http://www.dcs.gla.ac.uk/~roly
Petricek:	http://tomasp.net
Cheney:	http://homepages.inf.ed.ac.uk/jcheney
Bach:	http://benjbach.me
Malik:	http://facsrv.cs.depaul.edu/~tmalik1
Wolf:	http://www.bris.ac.uk/geography/people/levi-j-wolf

Costings

The 0.5 FTE Research Associate is Stuart Presnell, currently University of Bristol, available to start 1 June.

- Perera (1 FTE): £50,000.00
- Petricek (0.1 FTE): £6,500.00
- Cheney (0.05 FTE): £3,939.03
- Bach (0.05 FTE): £3,376.00
- Malik (0.1 FTE): £8,244.36
- Wolf (0.1 FTE): £6,275.00
- Research Associate (0.5 FTE): £25,000.00
- Research Software Engineer (0.5 FTE: £33,802.50
- Travel: 2 conferences for Perera; 1 for 0.5 FTE RA (US/Europe): £7,500.00
- Visualisation workshop + 2 project meetings: £2,500.00
- Laptop for RA: £1,200.00

References

[1] Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Rogier A. Kievit. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Research*, 4(63), 2019.

[2] Luc Anselin, Ibnu Syabri, and Youngihn Kho. GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, 38(1):5–22, 2006.

[3] B. Bach, P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale. A descriptive framework for temporal data visualizations based on generalized space-time cubes. *Computer Graphics Forum*, pages 1–26, 2016.

[4] Benjamin Bach, Zezhong Wang, Matteo Farinella, Dave Murray-Rust, and Nathalie Henry Riche. Design patterns for data comics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 38:1–38:12, New York, NY, USA, 2018. ACM.

[5] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos Eduardo Scheidegger, Cláudio T. Silva, and Huy T. Vo. Vistrails: Visualization meets data management. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 745–747, 2006.

[6] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. Increasing the transparency of research papers with explorable multiverse analyses. In *CHI 2019 - The ACM CHI Conference on Human Factors in Computing Systems*, Glasgow, United Kingdom, May 2019.

[7] Jochen Görtler, Rebecca Kehlbeck, and Oliver Deussen. A visual exploration of Gaussian processes. *Distill*, 2019. https://distill.pub/2019/visual-exploration-gaussian-processes.

[8] Kevin Jolly. *Hands-On Data Visualization with Bokeh: Interactive Web Plotting for Python Using Bokeh.* Packt Publishing Ltd, Birmingham, UK, 2018.

[9] Roly Perera, Umut A. Acar, James Cheney, and Paul Blain Levy. Functional programs that explain their work. In *Proceedings of the 17th ACM SIGPLAN International Conference on Functional Programming*, ICFP '12, pages 365–376, New York, NY, USA, 2012. ACM.

[10] Tomas Petricek. The gamma: Programming tools for open data-driven storytelling. In *Proceedings of European Data and Computational Journalism Conference (EDCJC 2017)*, 2017.

[11] Tomas Petricek, James Geddes, and Charles Sutton. Wrattler: Reproducible, live and polyglot notebooks. In *Proceedings of 10th USENIX Workshop on The Theory and Practice of Provenance (TaPP 2018)*, 2018.

[12] Sergio J. Rey and Luc Anselin. PySAL: A Python library of spatial analytical methods. *Review of Regional Studies*, 37(1):5–27, 2007.

[13] Wilmer Ricciotti, Jan Stolarek, Roly Perera, and James Cheney. Imperative functional programs that explain their work. *Proc. ACM Program. Lang.*, 1(ICFP):14:1–14:28, 2017.

[14] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-Lite: A grammar of interactive graphics. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2017.

[15] M. Tobiasz, P. Isenberg, and S. Carpendale. Lark: Coordinating co-located collaboration with information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1065–1072, Nov 2009.

[16] Tracey L. Weissgerber, Natasa M. Milic, Stacey J. Winham, and Vesna D. Garovic. Beyond bar and line graphs: Time for a new data presentation paradigm. *PLOS Biology*, 2015.

[17] J. Wood, A. Kachkaev, and J. Dykes. Design exposition with literate visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):759–768, Jan 2019.

[18] Zhihao Yuan, Dai Hai Ton That, Siddhant Kothari, Gabriel Fils, and Tanu Malik. Utilizing provenance in reusable research objects. *Informatics*, 5(1), 2018.