Tools for open, transparent and engaging storytelling

Tomas Petricek
The Alan Turing Institute
tomas@tomasp.net

INTRODUCTION

The rise of Big Data and Open Government Data initiatives means that there is an increasing amount of raw data about the world available. At the same time, "post-truth" has been chosen as the word of 2016 [1] and the general public increasingly distrusts statistics [2]. In other words, data science has more capabilities to help us understand the world than ever before, yet it is becoming less relevant in public discussion.

This should perhaps not be a surprise as data science is often opaque, non-experts find results difficult to interpret and verify, and creating data-driven reports requires advanced skills and is limited to a small number of specialists.

The purpose of the proposed demo is to present The Gamma project (http://thegamma.net) which aims to democratize data science. The Gamma encourages everyone — including journalists and interested citizens — to understand how presented claims are justified, explore data on their own and make their own transparent factual claims. If the society is to benefit from the possibilities available through data science, it is essential to make data-driven storytelling widely accessible, open and engaging.

Open and engaging data-driven storytelling

On one hand, spreadsheets made data exploration accessible to a large number of people, but operations performed on spreadsheets are error-prone and cannot be easily reproduced or replicated with different data source. On the other hand, data analyses written as programs can be modified and run repeatedly, but even with the simplest programming tools available, building an end-to-end analysis that reads data from a government data source, performs analysis and produces an interactive visualization requires expert programming and data science skills.

The Gamma aims to build programming tools that let anyone explore data from a wide range of data sources, including open government data, and publish data-driven reports that are:

- **Transparent and accountable.** Readers can review how data is used and discover misleading uses of data.
- **Reproducible and connected.** Readers can run the analysis themselves using the original data source.
- **Open & engaging.** Readers can modify parameters and share reports on different aspects of the data.

To achieve this, we treat data-driven reports as reproducible programs written in a simple web-based scripting language that is integrated with primary data sources (using type providers [3]) and we develop editor tooling that bridges the gap between programming and spreadsheets.

PROPOSED DEMO

The work that will be presented in the demo session has been focused on building a simple web-based library that could be used by data journalists to present visualizations obtained by aggregating and summarizing tabular data.

Figures 1 and 2 illustrates some of the steps performed during a sample task – given a data table recording individual medals awarded over the entire history of the Olympic games, we want to calculate the number of medals per country.

Innovative aspects of the project

The project is innovative in two ways. It creates a new simple scripting language for working with data and it complements the language with spreadsheet-inspired tooling:

- **Simple data-aware language.** When writing code, the programming language understands the data source and data transformations performed so far and offers all available operations when "." is typed. For example, when "Games is" is typed in Figure 1, the editor understands what values are available and offers the user "Rio (2016)" in the completion list. This means that the user can construct the whole program just by choosing one of the available operations.
- **Spreadsheet-inspired editing.** One of the reasons why spreadsheets are easy to use is that the user can always see the data they are working with and manipulate it directly. We adapt this paradigm to programming — in our live editor, the user can always see preview of the aggregation constructed so far, making data exploration easier. As demonstrated in Figure 2, many transformations can be created using the user interface without writing code directly. Yet, the final result is still an open and reproducible script.

These two innovations make it possible to create web-based data-driven reports that are transparent (anyone can see how they are created), open (readers can modify them and share their results) and engaging (reader can explore other fun aspects of the data).

The Gamma project is the first step of an increasingly important research that aims to democratize data science and encourage every citizen to make factual claims backed by data — be it for fun or to hold the government accountable.

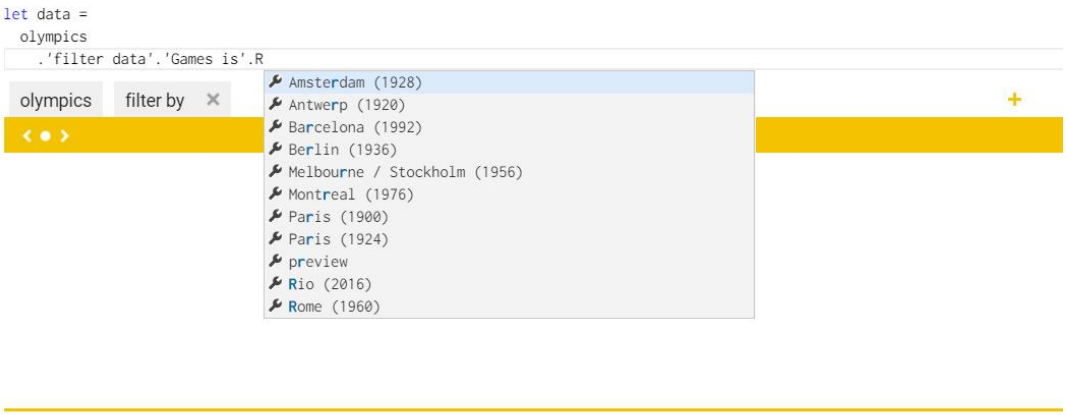

Figure 1. The Gamma project uses a type provider to generate types with members based on the structure of the processed data. When writing script to work with data, the auto-completion offers help based on the types. In the scripting language behind The Gamma, almost all data processing work is done by typing "." and choosing one of the available members, leading to an extremely simple programming model that can be well supported by editor tooling.

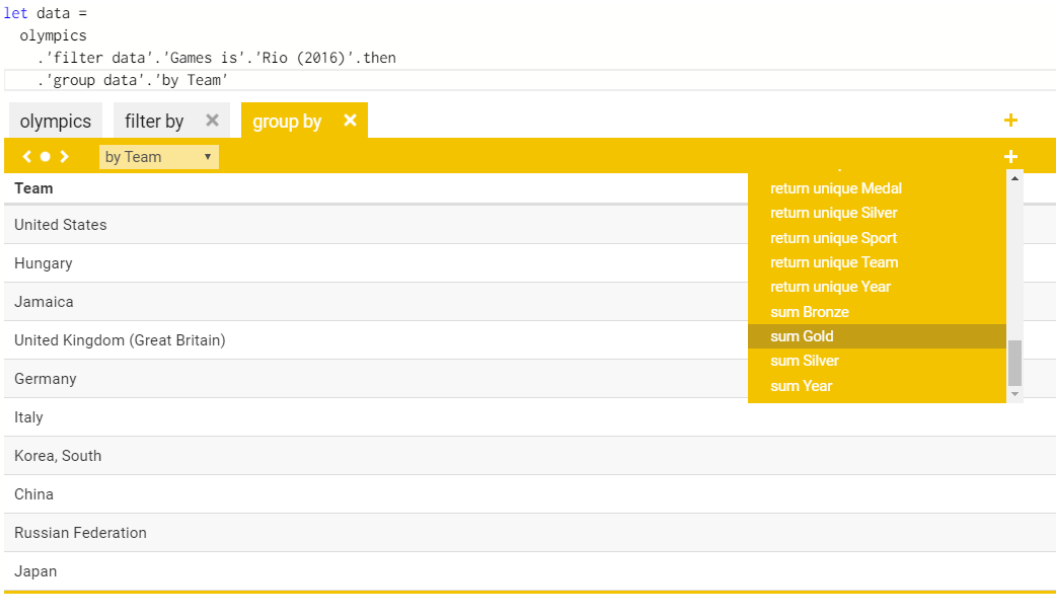

Figure 2. When writing data transformations, users can directly edit code that represents the data transformation, but they can also see the preview of the result of the data transformation written so far and edit the data transformation in a spreadsheet-inspired manner. Here, the user decided to aggregate data by team (country) and is now choosing aggregations to perform over the group.

REFERENCES

[1] BBC News, 'Post-truth' declared word of the year by Oxford Dictionaries, http://www.bbc.co.uk/news/uk-37995600

[2] William Davies, How statistics lost their power – and why we should fear what comes next, The Guardian 2017. https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy

[3] Don Syme, et al. Themes in information-rich functional programming for internet-scale data sources. Proceedings of Workshop on Data driven functional programming, 2013.