Authors:
Petricek, Tomas; Heravi, Bahareh R; Stark, Jennifer A; Diakopoulos, Nicholas; Chorley, Martin; Mottershead, Glyn; Broersma, Marcel; Esteve del Valle, Marc;

Copyright 2017 The Authors

University College Dublin, July 2017
ISBN: 9781910963159
Welcome to the European Data and Computational Journalism Conference

The European Data and Computational Journalism Conference aims to bring together industry, practitioners and academics in the fields of journalism and news production and information, data, social and computer sciences, facilitating a multidisciplinary discussion on these topics in order to advance research and practice in the broad area of Data and Computational Journalism.

Held in Dublin, Ireland, the first edition of this conference presents a mix of academic talks and keynotes from industry leaders. It is followed by a half-day 'Introduction to Data Journalism' workshop and the 'Computational and Data Journalism Unconference'.

Submissions of both academic research-focused and industry-focused talks were invited for the conference, on the subjects of journalism, data journalism, and information, data, social and computer sciences.

Topics of interest include, but are not limited to:

- Application of data and computational journalism within newsrooms
- Data driven investigations
- Data storytelling
- Open data for journalism, storytelling, transparency and accountability
- Algorithms, transparency and accountability
- Automated, robot and chatbot journalism
- Newsroom software and tools
- 'Post-fact' journalism and the impact of data
- User experience and interactivity
- Data and Computational Journalism education
- Post-desktop news provision/interaction
- Data mining news sources
- Visualisation and presentation
- Bias, ethics, transparency and truth in Data Journalism
- Newsroom challenges with respect to data journalism, best practices, success and failure stories

Collected within these proceedings are the academic abstracts presented at the conference.

We would like to take this opportunity to thank the programme committee for their hard work reviewing submissions and helping us to come up with the fantastic line-up of talks for this year.

Welcome to Dublin, and welcome to DataJConf 2017!

Bahareh R Heravi, Martin J Chorley & Glyn Mottershead
DataJConf 2017 Organising Committee
# European Data and Computational Journalism Conference

## Schedule

<table>
<thead>
<tr>
<th>Invited Talk</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conor Ryan, RTÉ Investigations, Ireland</td>
<td></td>
</tr>
<tr>
<td>International Data Journalism year in review</td>
<td></td>
</tr>
<tr>
<td>Marianne Bouchart, HEI-DA, France</td>
<td></td>
</tr>
<tr>
<td><strong>The Gamma: Programming tools for open data-driven storytelling</strong></td>
<td>3</td>
</tr>
<tr>
<td>Tomas Petricek, Alan Turing Institute, UK</td>
<td></td>
</tr>
<tr>
<td><strong>Detail Data: Data journalism with impact</strong></td>
<td></td>
</tr>
<tr>
<td>Kathryn Torney, The Detail Data, Northern Ireland, UK</td>
<td></td>
</tr>
<tr>
<td><strong>The state of Data Journalism globally</strong></td>
<td>5</td>
</tr>
<tr>
<td>Bahareh Heravi, University College Dublin, Ireland</td>
<td></td>
</tr>
<tr>
<td><strong>Launch of UCD Data Journalism programme</strong></td>
<td></td>
</tr>
<tr>
<td>Professor Colin Scott, University College Dublin, Ireland</td>
<td></td>
</tr>
<tr>
<td><strong>Leveraging Search data to bring Elections to life</strong></td>
<td>8</td>
</tr>
<tr>
<td>Clara Guibourg, Google News Lab, UK</td>
<td></td>
</tr>
<tr>
<td><strong>Using baselines for algorithm audits</strong></td>
<td></td>
</tr>
<tr>
<td>Jennifer Stark and Nicholas Diakopoulos, University of Maryland, USA</td>
<td>8</td>
</tr>
<tr>
<td><strong>Data Journalism as a civic duty: Lessons from Data Journalism in Developing Countries</strong></td>
<td>11</td>
</tr>
<tr>
<td>Eva Constantaras, Internews, USA</td>
<td></td>
</tr>
<tr>
<td><strong>Now we are three: A perspective on Computational and Data Journalism Education</strong></td>
<td>11</td>
</tr>
<tr>
<td>Martin Chorley and Glyn Mottershead, Cardiff University, Wales, UK.</td>
<td></td>
</tr>
<tr>
<td><strong>Automated Analysis of Journalists' and Politicians' Online Behavior on Social Media</strong></td>
<td>14</td>
</tr>
<tr>
<td>Marcel Broersma and Marc Esteve Del Valle, University of Groningen, The Netherlands.</td>
<td></td>
</tr>
<tr>
<td><strong>Invited Talk</strong></td>
<td></td>
</tr>
<tr>
<td>Megan Lucero, Bureau Local, The Bureau of Investigative Journalism, UK</td>
<td></td>
</tr>
</tbody>
</table>

## Closing
The Gamma: Programming tools for open data-driven storytelling

Tomas Petricek
Alan Turing Institute
tomas@tomasp.net

Abstract: The rise of open data initiatives means that there is an increasing amount of raw data available. At the same time, the general public increasingly distrusts statistics and "post-truth" has been chosen as the word of 2016. The objective of The Gamma project is to help reverse this development. We argue that an important part of the problem is the fact that data science is often opaque, non-experts find results difficult to interpret and verify, and creating data-driven reports is limited to a small number of specialists.

We aim to address this problem by building novel programming tools that make it easy for journalists to build open, transparent data-driven reports and encourage interested citizens to understand how the presented claims are justified, further explore data and make their own factual claims. We show that recent ideas from programming language research can dramatically simplify the complexity of code to perform data access and aggregation and make it possible to automatically provide user interfaces that enable the readers to modify parameters of data visualizations.

Keywords: Data-driven storytelling, Programming languages and tools, Data visualization, Transparency

Introduction: Why programming tools matter

There has never been a greater need for increasing data literacy and building tools that let anyone explore data and use it transparently to tell stories backed by data. Spreadsheets made data exploration broadly accessible, but operations performed on spreadsheets cannot be easily reproduced or replicated with different inputs. This manual mode of interaction breaks the link with the original data source and makes spreadsheets error-prone (Krugman, 2013; Panko, 1998). In contrast, when exploring data programmatically, the resulting scripts can be run repeatedly and their parameters can be modified, but this typically requires non-trivial programming knowledge.

In this paper, we discuss a new programming environment that makes it easy to access data from external sources, aggregate and summarize it and create transparent, reproducible visualizations. To do this, we combine two ideas from programming language research. First, we use type providers (Syme et al., 2013; Petricek et al., 2016) for integrating external data into the programming environment and for providing easy access to data aggregation. Second, we build a direct manipulation user-interface (Shneiderman 1997, Shneiderman et al. 1992) that allows the user to change scripts without directly editing the source code, while preserving the reproducibility. In this paper, we also report our experience with building a case study data visualizations using the presented tooling.

Methods: Language-integrated data access using type providers

In editors for typed object-oriented languages such a Java, typing “.” opens an auto-complete list with operations that can be invoked on a given object. Type providers extend this mechanism such that the available operations are offered not just based on class members, but can also be generated on-the-fly based on an external data source. For example, consider the following example, which accessed CO2 emissions of the UK from the World Bank:

```python
worldbank.byCountry.'United Kingdom'.Climate Change.'CO2 emissions (kt)'
```

Here, worldbank represents a type provider that exposes data from the World Bank. We choose byCountry view of the data, choose the UK and then select the required indicator. This simplifies writing of data-driven scripts in two ways. First, the language is extremely simple and most operations are captured as member access via the “.” operator. Second, the tooling offers auto-complete with available members and the user only needs to choose. For example, after typing byCountry, a list with all available countries appears.

The type provider mechanism can be used for navigating through a data source, but it can also be used for constructing more complex queries. The following example takes a table of individual Olympic medals, counts the number of medals per athlete and sorts the data:

```python
olympics.'group data'.'by Athlete'.sum Gold'.sum Silver'.sum Bronze'.then
  .sort data'.'by Gold descending'.'and by Silver descending'
```
As before, the whole program can be constructed by repeatedly typing "," and choosing one of the offered members. The type provider understands the constructed query — after grouping data by athlete, it offers aggregations such as 'sum Gold' to count the total number of gold medals. The aggregated columns are then offered for sorting. The type provider also guarantees that all scripts constructed in this way will be formally correct (Petricek, 2017).

**Findings and argument: Enabling new kind of data-driven storytelling**

The key finding from our work is that treating data visualizations as reproducible programs rather than as embedded images has overarching consequences. It makes the use of data transparent, allowing the readers to find the original source and check how data is used. Furthermore, it also enables development of other tools for data exploration.

The user interface in Figure 1 is available as part of The Gamma package (http://thegamma.net). It allows readers to change add or remove transformations and change their parameters. Each operation makes a corresponding change to the source code, so the script remains a full representation of the data transformation. In another case study, which presents a number of visualizations around Olympic medal data (http://rio2016.thegamma.net), the user interface is even simpler. It generates drop-down controls based on member names and so, e.g. 'Rio (2016)' can be easily substituted for 'London (2012)' or 'Beijing (2008)'.

Most of the Olympic medal visualizations are around 20 lines with the core logic often shorter than 10 lines. Treating data visualizations as scripts also offers new experiences for the readers that encourage data literacy. We are actively investigating how to simplify the import of external data sources — developing type providers using services such as the World Bank currently requires programming skills and uploading a CSV file does not guarantee full transparency.

**Conclusions**

We believe that the increasing distrust in statistics and facts is reinforced by the fact that creating data-driven reports requires significant expert skills. In this paper, we present The Gamma project, which aims to make data-driven storytelling simpler and more transparent. Data visualizations created using The Gamma inspire the reader to actively engage with data, verify sources and adapt visualizations to focus on other facts. We believe that active approach to data is essential for increasing data literacy and that treating data visualizations as scripts, rather than an embedded graphics, is the key component of transparent, open and reproducible data-driven storytelling.

**References**

The State of Data Journalism Globally

Bahareh R. Heravi
School of Information and Communication Studies
University College Dublin
Bahareh.Heravi@ucd.ie

Abstract: This paper presents the initial results of the first Global Data Journalism survey, which is a study into data journalism practices in newsrooms internationally. This study reveals various aspects of data journalism practice and characteristics of data journalists and data teams. These include demographics, education level, proficiency in various topics, size and formation of data teams, and common workflows. This study further looks into the values associated with journalistic work and analyses the ways in which the community believes data journalism has improved or undermined these values.

Keywords: Data journalism, data driven journalism, journalism, computational journalism

Introduction
Data driven journalism or data journalism is an emerging discipline that brings together knowledge from several disciplines, including journalism, social sciences, information science, data and computer sciences, data analytics, information design, and storytelling. Despite the growing interest in data driven journalism in newsrooms and its emergence as an academic discipline in its own right, there is a lack of systematic research in this domain, and a divide between academic and industry practices. This paper presents the initial results of the first Global Data Journalism survey, which studies the current state of data journalism in newsrooms across the globe, with the aim of providing theoretical, practical and pedagogical guidelines for the future of data driven journalism.

Methods
The survey was launched on the 3rd December 2017 and closed on the 10th May 2017 and was open to all data journalists and journalists globally. The survey was limited to those who identify as having worked as a journalist or a data journalist in the past year. In addition, as a qualifying but not mandatory criteria, we asked if journalism or data journalism formed a significant part of participants’ income.

The survey was carried out using the online Google Forms, and was circulated and promoted as broadly as possible through various platforms and channels. A link to the survey was distributed widely through social media channels and relevant listservs, two Slack groups - News Nerdy and DJA 2017 - and a number of articles about the survey featured in the media (Lorenz, 2016; Burke, 2016, Plaum, 2017). This survey was conducted following an ethical approval from the University College Dublin’s Research Ethics Committee.

The survey consisted of 48 questions in 7 sections.

Findings
Two hundred and six participants from 43 countries participated in this survey between 3rd December 2016 and 10th May 2017, with 181 respondents filling it out to completion. Considering the small community of Data Journalism we believe this is a considerable sample size. In terms of gender balance 57.5% of our participants identified as male and 42.5% as female. Of all participants 64% were in full time employment, 18% freelance, 12% part time and 4% casual/retainer. 32% of participants worked in large organisations of 500+ employees, 22% in organisations of size 10-49, 17% in organisations with 100 to 499 employees, 15% of these in small organisations of 2 to 9 employees and only 8% in mid-sized organisations of 50-99 employees. 42% of participants work in national organisations, 20% in local, 18% in International and the rest in a combination of these types, or other types of organisations.

1 This survey was conducted by Bahareh Heravi and Mirko Lorenz and was open between 3rd December 2016 and 10th May 2017.
Out of all participants 43% produce content for online platform of broadcast or print media outlet and 34% produce content for online only publications. This makes a total 77% of all participants producing content for online publications. This figure is followed by print newspaper (8%), Radio (4%), TV (4%), print magazines (3%), personal blog (2%) and producing content for news agency makes only 1% of the total. In terms of experience as a journalist a majority of our respondents (78%) were individuals with 1 to 10 years experience as a journalist with breakdown of 2% having less than a year experience, 41% having 1 to 4 years experience and 26% 5 to 9 years. 19% of our participants have 10 to 19 years experience and only 11% have over 20 years experience as a journalist.

We asked our participants about the status of data journalism in their organisations. Forty-six per cent claimed that they have a dedicated data desk/team/unit/blog/section. This figure is followed by 29% who expressed that they do not have a dedicated data desk/team/unit/blog/section, but publish data driven projects on a regular basis. 7% of participants noted that they plan to work with data in the next six months and 7% expressed that they have no immediate plan to start working with data. Of those who do have dedicated data desk/team/unit/blog/section, 40% have a data team consisting of 3 to 5 people and 30% have a team of 1 to 2 people. This means a vast majority (70%) of organisations with data teams operate with small teams of 1 to 5. On the other side of the spectrum 22% of participating organisations have data teams of 6 to 10 people and 3% have a team of 11 to 15 people and 5% have large data teams of more than 15 people.

While 86% of our participants consider themselves to be data journalists, in terms of data journalism proficiency only 18% rate themselves as experts in data journalism, while 44% of respondents identify as having a better than average knowledge in data journalism and 26% identify as having average knowledge in the field. 13% of participants identified as novice or below average level of expertise in the field. Half of our participants (50%) had formal training in data journalism and the other half did not. In terms of a wider understanding of formal training in knowledge areas used in data journalism, most of our participants demonstrate a high degree of formal training in journalism, while a lower and varying degrees of formal training in the more data oriented and technical aspects such as data analysis, statistics, coding, data science, machine learning and data visualisation. Figure 2 depicts the breakdown of formal training in various related fields between our participants.
In terms of general education level, 96% of our respondents had a university degree, with a breakdown of 40% at undergraduate (bachelor) level, 53% postgraduate level and 3% with a doctorate or above degree. This shows that data journalism community is a highly educated community composed of 96% university graduates, 50% of whom have a postgraduate university degree. Looking into the degrees obtained by these participants 62% are formally educated in Journalism at the university level. While Journalism is by far the most prevalent obtained higher education degree between our participants, it is followed by a combination of other degrees Politics (15%), Computer/Information/Data Science/Engineering (12%) and Communication and Language/Literature each 10.5%, with 26% listing a combination of other degrees. This shows that while most participating journalists have formal higher education training in Journalism, Communication, Politics and related degrees such as Literature, only 12% have higher education training in the more data related and technical topics. This further reflects on the basic underlying reasons behind the level of training demonstrated in Figure 2. It further denotes that formal training between the participants seems to have been mainly obtained through higher education and university degrees, and highlights the importance of including data related courses and modules in relevant higher education Journalism and Communication programmes.

In terms of values associated with journalism we asked our participants a series of questions, a number of which are briefly discussed in this section. Sixty-five per cent of the respondents somehow agree or strongly agree that data journalism allows them or their organisation to produce more stories. On the end of the spectrum 13% somewhat disagree (10%) or strongly disagree (3%) with this statement. Moving on from quantity to quality, 90% of respondents agree somewhat (21%) or strongly agree (69%) that data driven journalism adds rigour to journalism, with only 5% expressing the opposite. Similarly 91% agree or strongly agree that data journalism improves the quality of journalistic work in their organisation, with only 4% believing the opposite. Tapping into traditional journalistic values, while leaving the definition of these values to the participants, 83% of participating journalists disagree somewhat or strongly disagree that data journalism undermines traditional journalistic values, while only 11% agree somewhat or strongly agree that data journalism is undermining these values. On a final note, 70% of participants expressed that they will not be able to carry out their work without data as a source.

**Conclusion**

Data journalism has evolved tremendously in the past few years, and is rapidly becoming an integral part of many newsrooms. This paper presented the initial results of the first Global Data Journalism survey, studying and analysing the current state of data journalism in newsrooms internationally. 206 participants from 43 countries participated in this study. The results show that the data journalism community is a highly educated community, while it has its roots mostly in Journalism and Communication degrees, and less so in data/information and computer related disciplines. Additionally journalists engaged in data journalism form a younger cohort of journalists, with less than 10 years experience as a journalist. While technical, data analytics and statistical skills do not seem to be the strength of participating journalists put next to their journalism background, it appears that many newsrooms already have dedicated data team and produce data driven stories on a regular basis. This study further reveals that despite debates in the use of data for producing journalistic work, both in terms of quantity and quality, a vast majority of journalists believe that data journalism allows them to create more stories in terms of quantity, which also are of higher rigorous and of higher quality.

This paper presented a small fraction of the data collected in the Global Data Journalism survey in a descriptive manner. A further, more detailed, and multivariate analysis of the results is yet to take place in the future.

**References**


Lorenz, M., (2016), What is the current status of data journalism? Participate in our global survey to find out, Data Driven Journalism, [online]. Available at http://datadrivenjournalism.net/news_and_analysis/2017_global_data_journalism_survey

Using Baselines for Algorithm Audits

Jennifer A. Stark and Nicholas Diakopoulos
University of Maryland, College Park
{starkja, nad}@umd.edu

Abstract: Algorithm audits that assess bias often call for a suitable baseline to be defined. Bias becomes salient and newsworthy when put into contrast with a comparison, often an expectation derived by defining a baseline. This paper considers a baseline and three bias-rating methods for a study investigating bias in Google’s image selection for presidential candidates’ main search results page. Images of U.S. presidential candidates Hillary Clinton and Donald Trump were scraped from Google’s image box on the main results page, and also from all results in Google Image search for the same time period, which served as the baseline. Using Google as its own baseline was sufficient to demonstrate that images on the main results page exhibit different distributions of news sources and sentiment, indicating that bias may have been introduced by the Google image selection algorithm.

Keywords: Algorithm Audits, Algorithmic Accountability, Baseline, Search Engines

Introduction
Reporting on bias in search engines is relatively new, and effective methodologies are still being established (Diakopoulos et al. 2018). One challenge is defining and obtaining baseline data appropriate to the specific study: assessments of bias must be put in context. The news value of an audit may depend on the deviation of some distribution with respect to that baseline. To date, studies have mostly not used baselines, opting instead to perform within-sample comparisons. For example, an exploratory study compared the emotions and age of women in news photographs with those of men, along with ratios of photographs of women and men across different news organisations (Kwak & An 2016) using data collected from the global-level news database GDELT (Global Data on Events, Location, and Tone), for a given time period. Another study compared images from specific queries between search engines Bing and Google, and between countries (Magno et al. 2016). Conversely, a study exploring stereotypical images of occupations used real-life counts from the Bureau of Labor Statistics as its baseline (Kay et al. 2015), enabling it to draw conclusions regarding stereotypes. While such comparisons are interesting and meaningful, isolating the role of a particular search algorithm in introducing bias into the most visible results requires that the comparison, or baseline, represent the universe of potential results surfaced by the search query studied.

We use this definition of baseline to investigate the visual framing of 2016 US presidential candidates Hillary Clinton and Donald Trump. Candidate images were collected from Google’s main (non-personalized) search results page. Images were also collected from Google’s image search that served as our baseline. Search queries were “hillary clinton” and “donald trump”.

We show that, compared with baseline images, the proportion of images of Clinton displayed on the main results page showing happiness was greater, while the proportion with a neutral expression was reduced. For Trump, percentages of happiness and anger were both increased compared with the baseline. Regarding political ideology, the cumulative proportion of left-leaning image sources was augmented in the image box for Clinton compared with the baseline, while the proportion from centrist sources was reduced. The proportion of Trump images from the centrist sources was reduced compared with baseline, while that from right-leaning sources was enhanced in the image box. The baselines of both candidates privileged left-leaning sources, overall. Together, this suggests that Google’s main results page images are not curated in an ideologically balanced way.

Methods
We collected images from the main search result image box once per day from September 3rd to October 28th 2016 for the search queries “hillary clinton” and “donald trump”, resulting in nine unique images for Clinton, and 11 for Trump. Baseline images were collected in early 2017 using Google Image search with advanced options specifying the same time period as for the image box collection, resulting in 353 images of Clinton, and 298 of Trump. Baseline images that were not of the queried candidate or that contained more than one face were removed. We
selected this baseline because Google was the universe we wanted to investigate; we assumed that images presented in the image box on the main results page would be gathered from Google’s own indexed images, and therefore all potential images that could be presented in the image box would be found in Image Search. We acknowledge that images available in Google Images may themselves be biased in terms of selection, but evaluating the bias of Google Images itself (with respect to some other baseline) was beyond the scope of this study. All Google searches were non-personalized so that search history would not affect search results.

Emotion was determined for each unique image using the Microsoft emotion API set to a confidence threshold of ≥ 0.55 (considered ‘likely’) out of a total of 1.0 for an emotion to be declared as present. Sources were counted for each time the image occurred in a data-collection of the image box (i.e. once per day). All image sources were tagged with a bias rating based on a dataset aggregated, calculated, and provided by Allsides.com, combined with bias ratings from a study of news sharing on Facebook (Bakshy et al. 2015). Sources that remained unrated, or whose ratings from Allsides and the Facebook study disagreed, were assessed for bias by the first author according to the source’s About pages, content of several articles not including that which the image linked to, and ratings from a third site MondoTimes.com. Code for all collection and analysis can be found on GitHub (https://github.com/comp-journalism/GoogleScraper, https://github.com/comp-journalism/Baseline_Problem_for_Algorithm_Audits).

Findings and Argument

We found that Clinton expressed emotions congruent with gender stereotypes for women in news photographs (Kwak & An 2016): that women smile more. However, this was not reflected in the baseline in which neutral expressions dominated instead (Figure 1 left). Trump had higher proportions of images showing happiness, anger, and slightly fewer showing no emotion compared with the baseline (Figure 1 right).

We found that sources of baseline images were mostly left-leaning with lower proportions from centrist and right-leaning sources for both candidates (Figure 2). Further, a chi-square test showed that the distribution of political ideology for sources of images in the image box is significantly different from those in the baseline (Clinton: χ² = 68.1, p < 0.001; Trump: χ² = 36.4, p < 0.001). Specifically, the proportion of left-learning sources of Clinton image box images was augmented at the expense of centrist sources. For Trump, left-learning sources remained highly represented in the image box, while right-leaning representation was increased at the expense of centrist. Of note, right-leaning sources in the image box were highly partisan, including pro-Trump subreddit r/The_Donald, pro-Russia web news site Russia Insider, and an obscure pro-Trump YouTube channel. Finally, not all sources of image box images were represented in the baseline. Some of these sources were identified in Google Image search results after extending the Google Image search time-frame further back in time, while others – including r/The_Donald and Russia Insider remained absent, indicating that curation of images for the image box may have different temporal dependencies than we assumed when initially collecting Google Image search images. Another possibility is that images in Google Image search do not represent all Google’s indexed images.

---

Figure 1 – Emotion disparities: Percentage of Clinton and Trump images from Image box, or Google Images (Baseline) that show listed emotions.

Figure 2 – Image Source Ideological Representation: Percentage of Clinton and Trump images from image box or Google Images baseline identified from each listed ideology.

---

1 https://www.allsides.com/bias/about-bias
2 https://www.youtube.com/watch?v=dY77j6uBHI
Conclusions
We found that representation of candidates’ emotions diverged from that seen in baseline images, whereby Clinton was predominantly happy, and Trump was happier and more angry compared with the baseline. We also found that the baseline images from Google Image search for Clinton and Trump were biased toward liberal sources. Sources of images on Google’s main results page had a different political ideological distribution that suggests some degree of bias may have been introduced by the Google image selection algorithm. Finally, by incorporating a baseline into this study, we ensured that errors incurred by the emotion API resulting from its own training data are mitigated in part, since errors applied to the images of interest are also applied to the baseline images. Limitations include: a failure to disambiguate the influence of photographers’ aesthetics and editors’ selection criteria that may introduce biases into Google’s image universe; we were not ultimately able to assign a bias rating to all sources, such as the stock image site Getty Images, foreign news outlets, and magazines with little to no news or political content (41 missing for Clinton; 23 for Trump), meaning that, were all sources rated, the final pattern of results may be different; and finally, that the chosen time-frame for the baseline to match that of the image-box image collection was insufficient to capture all the sources, and therefore did not represent the universe of potential images Google could select from to populate the main results page image box. We recommend that studies investigating bias in search engines consider and define baselines carefully and iteratively so as to ensure a reliable expectation to contrast with another sample.

Acknowledgements
This work has been supported by grants from the Tow Center for Digital Journalism and from the VisMedia project at the University of Bergen.

References
Now we are three: A perspective on Computational and Data Journalism education

Martin Chorley
School of Computer Science & Informatics
Cardiff University
chorleymj@cardiff.ac.uk

Glyn Mottershead
School of Journalism, Media and Cultural Studies
Cardiff University
mottersheadgg@cardiff.ac.uk

Abstract: In this talk we present the key findings from the process of designing a new joint-honours Master’s degree combining Computer Science and Journalism, and discuss the lessons learnt from the first three years of delivery. We discuss the design and evolution of the course in terms of its learning outcomes and the practical skills taught, the changes made during the first three years, and reflect upon the experiences and outcomes from delivering the course.

Keywords: Data Journalism Education.

Introduction

The MSc in Computational and Data Journalism is a one-year level 7 postgraduate course taught at Cardiff University, designed to take applicants from a broad spectrum of backgrounds\(^1\) and teach them a set of Journalism and Computing skills needed to succeed in the modern newsroom. First introduced in 2014/2015, it is now in its third year and has produced graduates who have gone on to work for journalism organisations around the world in roles such as data-journalist and newsroom developer. We present here a reflection on the first three years of delivering the course, the key principles underlying the design of the course, and lessons learnt that may be useful to those interested in computational and data journalism education.

A unique proposition of the course is that it is truly a joint-honours course, with invested and accessible course-directors from both academic schools involved. Both schools have made significant commitments in terms of new modules and teaching resources to develop and deliver the course.

Course Design

One of the key struggles for designing a cutting-edge joint-honours course is a lack of relevant subject benchmarks and people specifications, and few examples of existing UK courses. Individual subject benchmarks and specifications are not entirely relevant to the course\(^2\). Some previous data was available for guidance (e.g. (Yarnall, Johnson, Rinne, & Ranney, 2008)) but these are limited, being focused on past versions of the tools and techniques used in the news industry. The relatively recent introduction of freedom of information to the UK compared to the US results in the UK industry playing catch up. A lack of data journalism tradition results in less guidance on which to build a practice focused course. A reliance on generic frameworks (Higher Education Funding Council For Wales, 2015; QAA, 2008) was backed up by consultation with several industry sources to gather input and feedback on the plans. It was deemed more relevant to consult practitioners on the required skills, knowledge and qualities of potential graduates, rather than relying upon what could be perhaps a more academic assessment of the discipline. This industry input helped guide our thinking about how an ‘educated person’ (Boyer, 1995) from this course should look. No single curriculum design model was used in the design process, but rather a ‘best effort’ approach was used combining relevant modules from both schools, with gaps in curriculum filled with new modules where necessary. Elements of Tyler’s model (Tyler, 1949) can be found in our process, as are aspects of Alexander’s model (Alexander, 1997), although with a focus on what should be learnt, how it should be learned,

\(^1\) No prior knowledge of either computing or journalism is required, just a ‘good’ first degree or relevant work experience.

\(^2\) No QAA subject benchmark exists for Journalism at the postgraduate level - other postgraduate courses rely on benchmarks from professional organisations such as the NCTJ, however, these bodies are not necessarily relevant for a course in Computational and Data Journalism. Meanwhile, the QAA subject benchmark for Computing (QAA, 2011), while a useful guide, is not entirely relevant.
taught and assessed, and what an educated person should be, with not as justification as to why learners should be educated this way. ILOs are written using a Constructive Alignment model (Biggs, 1996).

Course Structure
Resource pressures meant that we had to fit as much as possible within existing courses and modules. Gaps were filled by new modules: two new Computer Science modules: one on web development and one on data analysis and visualisation, backed up in the third year with a new Journalism module: Data Journalism. Inclusion of the core foundational data journalism and coding knowledge taught in the first semester is critical, a view reinforced by (Berret & Phillips, 2016) in their recent report, which highlights the same key skills and knowledge requirements identified when designing our course. Comparing the structure of our MSc to their suggested programme structures shows a commonality of modules and topics which validates our approach. One key point raised by our guiding industry and academic panel was that students needed space to fail that was free from assessment load and supported by the teaching team. This resulted in a series of harmonisation workshops known as The Lab. This space allows staff from computer science and journalism to support students with their learning – inevitably their coding in the early part of the course – but also allows the team to discuss application of techniques in the newsroom and run mini projects. These often involve Freedom of Information applications for data as well as domain-specific implementations of their computer science classes.

Does everyone need to learn to code?
The coding aspect of the course is a particular concern for those students entering the course from a non-technical background, as ‘programming’ itself can be considered a threshold concept (Meyer & Land, 2003). In our experience so far this has not proven to be an insurmountable barrier. The benefits of including technical coding modules have been clear to see in the outputs from the students and the career paths they have taken. The coding modules within the programme are shared with other ‘conversion’ MSc courses which have a long tradition of successfully educating students from non-technical backgrounds.

Digital Investigation
A key innovative module on the course is ‘Digital Investigation’. In this module, students work in teams as close to a real-world ‘investigation’ team as possible. They are responsible for managing their own projects using an Agile methodology (Fowler & Highsmith, 2001; Maher, 2009). They are also responsible for self-managing the learning and skills acquisition required to successfully complete their project. A Connectivist learning model (Siemens, 2005) assumes that the student teams will work as social networks, sharing and gaining knowledge together from a diverse set of sources in a form of peer learning (Havnes, 2008), which has been found to work quite successfully in our experience. In this module a diverse set of technological tools are used, such as Slack\(^3\) for communication, Trello\(^4\) for project management, and Github\(^5\) for code and resource sharing. This module provides students with a safe space to fail, before heading out to work on real-world projects.

Reflection and Review
The course launched (as the MSc in Computational Journalism) in 2014/2015 with our first cohort of students. Based on personal reflection and feedback and evaluation from the first ‘alpha-test’ year, we made initial immediate changes for the second (‘beta-test’) year in 2015/2016, while planning for larger changes for the 2016/2017 year. Initial changes were fairly small such as adjusting the group assessments in Digital Investigation in order to take individual contributions into account to mitigate issues such as social loafing (Latané, Williams, & Harkins, 1979)\(^6\). We have kept many ‘unique’ aspects, such as the harmonisation ‘Lab’ seminars in the autumn semester. Data Journalism in the Autumn semester replaces the original Digital Journalism.

---

\(^3\) http://www.slack.com
\(^4\) http://www.trello.com
\(^5\) http://www.github.com
\(^6\) Although a small change this is a minefield of potential pitfalls (Gibbs, 1997)
Intake
Initial assumptions held that the course would skew mostly towards home students, with international students as a minority. This situation was reversed in the first two years, with more international students than Home/EU. A further assumption was that as a technical course intake would be skewed towards male students, as in many other STEM subjects. However, the first year was 75% female, and both the second and third years are an equal split male/female. The issue of gender equality in STEM subjects is well considered (Baker, 2013; Cheryan, Plaut, Davies, & Steele, 2009), but so far we have found no issues with gender participation in the new course.

Conclusions
Key to the success of this course has been a truly joint operation between both involved schools, with time and effort allocated for course-directors to ensure that the course is not a ‘data journalism + some coding’ or a ‘computing + some journalism’ programme. Providing ample practice focused time for the students and allowing them the space and support to be able to try new things, and fail if necessary has been vital.

References
QAA. (2011). Subject benchmark statement Master’s degrees in computing.
Automated Analysis of Online Behavior on Social Media

Marcel Broersma
Professor and director, Centre for Media and Journalism Studies, University of Groningen (The Netherlands)
m.j.broersma@rug.nl

Marc Esteve Del Valle
Assistant Professor, Centre for Media and Journalism Studies, University of Groningen (The Netherlands)
m.esteve.del.valle@rug.nl

Abstract: This research studies a) which discursive practices politicians and journalists use on Twitter and how these have changed, and b) to what extent institutional difference between agents still matter, or even exist, now they all have the power to publish on social media. To analyze this machine learning is applied to do large-scale automated content analysis of discursive practices. Our work suggests that although journalists and politicians have different roles and goals, their behavior on social media is surprisingly similar. This hypothesized redistribution of power in the so-called “triangle of political communication” calls for a revision of classical theoretical insights that are key to both political communication and journalism studies.

Keywords: Social media, Politics, Journalism, Discursive Practices, Automated Content Analysis.

Introduction
In the past decades, the use of Twitter and Facebook has become fully integrated in the professional repertoires of politicians and (political) journalists (Broersma and Graham, 2016; Cision, 2013; Graham et al., 2016; Esteve Del Valle, 2015). They are both very aware of the opportunities and challenges networked communication offers them in both relating to each other and reaching out to the public. A hybrid media system is taking shape where a mass media logic and a networked logic interact, and the various agents in the so-called triangle of political communication “create, tap, or steer information flows in ways that suit their goals and in ways that modify, enable, or disable others’ agency, across and between a range of older a newer media settings” (Chadwick, 2013, p.4; Klinger and Svensson, 2015). The hybridization that occurs can create significant changes to established working practices. For politicians, impression management is thus increasingly important to win voters (Bruns et al., 2016). At the same time, journalists cope with the issue of how to attract the attention of this fragmented audience and how to profile their brand in a hybrid communication paradigm.

The possibility of direct and open communication with and to citizens, even when it is unidirectional, has changed the power structures in both political communication and the media ecology for better and for worse. The power relations in the triangle of political communication are redistributed. However, to what extent and how the communicative spaces of politics and journalism are and will be transformed is still up for debate. The use of various social media platforms and their functions in an evolving hybrid media system are still very much in flux while scholarship especially on the levels of non-manifest content and theory is still in its infancies.

This research explores automated computational analysis of the (latent) discursive practices of politicians and journalists on Twitter, using machine learning. It asks:

• RQ1: Which discursive practices do politicians and journalists use on Twitter and how do these change?

• RQ2: To what extent do institutional differences between agents still matter, or even exist, now they all have the power to publish on social media?

The main hypothesis underlying the work is that networked communication is blurring the distinctive but interdependent roles of journalists and politicians. Now that they can both broadcast information, the online behavior of politicians and journalists converges. They might have different norms, aims and functions in society, but their discursive practices such as broadcasting, promoting, criticizing, branding or requesting input overlap. This changes the power relations in the triangle of political communication.

Methods
We will take three manually coded (annotated) datasets as a starting point to develop algorithms that allow for automated content analysis:

1) Tweets of Dutch political candidates during the 2012 election campaigns in the Netherlands (N= 55,992)
2) Tweets of British political candidates during the 2015 election campaigns in the UK (N= 31,796)
3) Tweets of Dutch and Belgian journalists in two random weeks (N=5,978)

Tweets have been gathered via Twitter’s API and were presented in an online interface that allowed coders to study the tweet in the context of the thread in which it was posed (see for the code of this tool: https://github.com/valeriobasile/twittercrawler). The coding unit for each dataset was the individual tweet. These have been manually coded for a range of manifest and latent variables, based on an extensive coding scheme. The three training datasets are big enough to provide the information to train the supervised learning algorithms. Based on the manual codes of latent categories (which needs much human interpretation) the computer will be “trained” through supervised machine learning to code tweets automatically (Petchler and González-Bailón, 2015). Latent codes are, most importantly, the function of the tweet/online behavior which consists out of 12 categories in the case of politicians and 18 categories in the case of journalists.

Findings and Argument
Research conducted by the authors has shown that, amongst other functions, journalists are using social media to monitor information, source stories, build and maintain professional networks and interact with citizens (Broersma and Graham, 2013). Reporters indicate that monitoring is an important reason for using social media; between 81 (Australia) and 66 percent (Germany) of the surveyed journalists in nine Western countries use Twitter for this reason (Cision, 2013). Political journalists, for instance, use Twitter to take the citizens’ pulse of certain topics or to have immediate access to the information delivered by politicians (Broersma and Graham, 2016). Journalists also use social media to build and maintain professional networks. They use social networking sites to directly contact politicians. The direct messaging function of Twitter is in this regard a key affordance for journalists who want to establish individual contact with politicians. Furthermore, journalists use social media to network with other peers. To do so, Twitter offer the possibility to create lists of accounts, a well-known function by journalists. Last, social media offer journalists the opportunity to directly interact with citizens. On Twitter, journalists can use their retweets to enjoy a brokerage position between citizen’s information and their audiences, or they can engage in a more dialogical conversation by mentioning other users.

These roles adopted by journalists in the social media age open room for new techno-mediated relations between journalists, politicians and citizens. The bulk of research studying journalists’ use of social media has focused, however, on the study of their behavior in this new medium without considering the role that they play as interconnected agents in the triangle of political communication. We aim to solve this lacuna by analyzing the relations between politicians and journalists on Twitter.

Conclusions
In the classic triangle of political communication, politicians, journalists and citizens are sharply distinguished categories with different aims and behavioral patterns (Broersma and Graham, 2016). Conversely, in the new networked communication paradigm these agents are adopting similar roles, i.e. becoming hubs in the networks. While some research on how social media have changed journalism and politics argues that social networking sites are normalizing established politicians and journalists’ professional practices and routines, some of our previous findings show that social media are redefining their behavior. Politicians and journalists’ use of social media is facilitating the emergence of new discursive practices which shape their functions and capabilities in modern democracies.

By using automated content analysis and machine learning methods, this research will allow us to automatically analyze the discursive practices (or: online behavior) of agents, also beyond politicians and journalists, in big social media data sets. This does not only provide empirical insights about the use of Twitter by politicians and journalists, but will also show how power is redistributed among these agents. This will contribute to the revisiting of classical theoretical insights (Coleman, 2009) that are key to both political communication and journalism studies, and will move research from a mass communication paradigm to one of networked communication.

References
European Data and Computational Journalism Conference
Programme Committee 2017

- James Hamilton (Stanford University, U.S.)
- Nicholas Diakopoulos (University of Maryland, U.S.)
- Meredith Broussard (New York University, U.S.)
- Cheryl Phillips (Stanford University, U.S.)
- Eddy Borges Rey (University of Stirling, Scotland, U.K.)
- Marc Esteve del Valle (University of Groningen, The Netherlands)
- Miranda McLachlan (Goldsmiths, University of London, U.K.)
- Stefano Cecccon (The Times and The Sunday Times, U.K.)
- Jonathan Gray (University of Bath, U.K.)
- Paul Bradshaw (Birmingham City University, U.K.)
- Bahareh Heravi (University College Dublin, Ireland)
- Glyn Mottershead (Cardiff University, Wales, U.K.)
- Martin Chorley (Cardiff University, Wales, U.K.)

Thanks to all the committee for their timely and informative reviews!