Wading through the swamp: filter systems for geospatial data science

Adrian Tear[*1], Richard Healey[†1]

[1]Department of Geography, University of Portsmouth, United Kingdom

December 1, 2016

Summary

Data Scientists write code, understand statistics and derive insights from data. Commonly used technologies include Relational Database Management Systems and Structured Query Language, 'Big Data' ecosystems such as Hadoop, R, Java, Python or similar code bases, and; Natural Language Processing software used to text-mine large un-/semi-/structured datasets. Data Scientists must also deploy complex software environments built on (or across) various physical, virtual and/or cloud computing infrastructures. This paper outlines the development of one such system designed to comprehensively analyse >8 million Twitter/Facebook social media posts.

KEYWORDS: Data Science, Big Data, Social Media, Text/Data-mining, Analytics

1. Introduction

The role of the Data Scientist is apparently 'The Sexiest Job of the 21st Century' (Davenport & Patil 2012) as Castells' (1996, 2009) *Rise of the network society* creates a data-driven 'information age economy'. Popular online 'Cheat Sheets' (Ohri 2014) for budding Data Scientists encourage their readers to 'write code, understand statistics [and] derive insights from data', detailing a number of skills or technologies which these titans of the new information age must command including 'R, Python, Java, SQL, Hadoop (Pig, HQL, MR) etc.' Data Scientists have been described as 'Engineers of the Future' (van der Aalst 2014); analytical experts (Agarwal & Dhar 2014) offering improved 'data-driven' decision-making (Baesens 2014) and predictive insight into human (or machine) behaviour (Dhar 2013). Here we liken the role of Data Scientist to that of water engineer, 'plumbing' together hardware and software systems (Lin & Ryaboy 2013) to filter the potentially enormous, highly varied, and often messy, 'swamp' of Big Data.

2. Background

Geographic data have always been 'big' (Li et al. 2016) but are usually highly structured (Graham & Shelton 2013). 'Big Data' can be a lot bigger – think in terms of the 'petabytes, exabytes, zettabytes, and yottabytes' of Foley's (2013) *Extreme Big Data* – and are often 'messy', characterised by a mixture of un-/semi-/structured elements (such as text, images or implicit geographic references) which present challenges to conventional computational analysis (Tsou 2015; Kambatla et al. 2014). This research examines:

- 1,718,667 records (up to 146 variables) sampled from an estimated corpus of ~75m+ Twitter tweets and Facebook posts made in the two-month lead-up to the 2012 US Presidential Election (2.46GB in CSV; 3.01GB in JSON).

[*] adrian.tear@port.ac.uk
[†] richard.healey@port.ac.uk

- 6,477,770 records (up to 411 variables) sampled from Twitter tweets and Facebook posts made in the twelve-month lead-up to the 2014 Scottish Independence Referendum (20.1GB in CSV; 19.0GB in JSON).

Social media, or Online Social Network (OSN), data have been widely used in social science and geographical research (see Steiger et al. 2015 for a recent summary) as the explosive upsurge in usage of platforms such as Facebook and Twitter, along with 'public posting' on these sites (Hough 2009), has allowed widespread, sometimes free or otherwise low cost access to very large numbers of user 'interactions', the messages and metadata bundles shared online. 2012 US and 2014 Scottish OSN data were recorded using the DataSift platform (DataSift 2013) which offered, until 5 December 2014, straightforward access to public Facebook posts and the full 'Firehose' of Twitter tweets on a 'pay-as-you-go' basis. The resulting datasets were large and could be downloaded in UTF-8 encoded (Yergeau 2003) Comma Separated Values (CSV) or JavaScript Object Notation (JSON) formats (ECMA International 2013). US and Scottish votes were chosen in this study as elections are data rich events involving a wide range of 'political actors' (Dahlgren 2005; Vergeer 2012) including candidates, political parties, news organisations, journalists, commentators, opinion pollsters and – of course – the electorate who determine outcomes on election day. Geographic data (latitude/longitude geotags, toponymical references and time-zone encoded date/time stamps) are present in OSN content or metadata. Consequently, the research aimed to test several hypotheses:

- H1 – Geographic and non-geographic users exhibit different characteristics
- H2 – Patterns of media link sharing differ between user types
- H3 – Text-mining for geographic references enables greater geographic inference
- H4 – Geographically referenced OSN sentiment offers locally predictive power

Results of this work are forthcoming, depending upon a detailed process of investigation (outlined below) and the development of a highly-functional, multiply-scaled, hardware and software infrastructure designed to filter the messy OSN Big Data 'swamp' using a wide variety of spatio-temporal, data and text-mining processes and technologies.

3. Process of investigation

A comprehensive literature review has resulted in the collection of over 950 academic book, section or journal references. These records have themselves been data-mined.

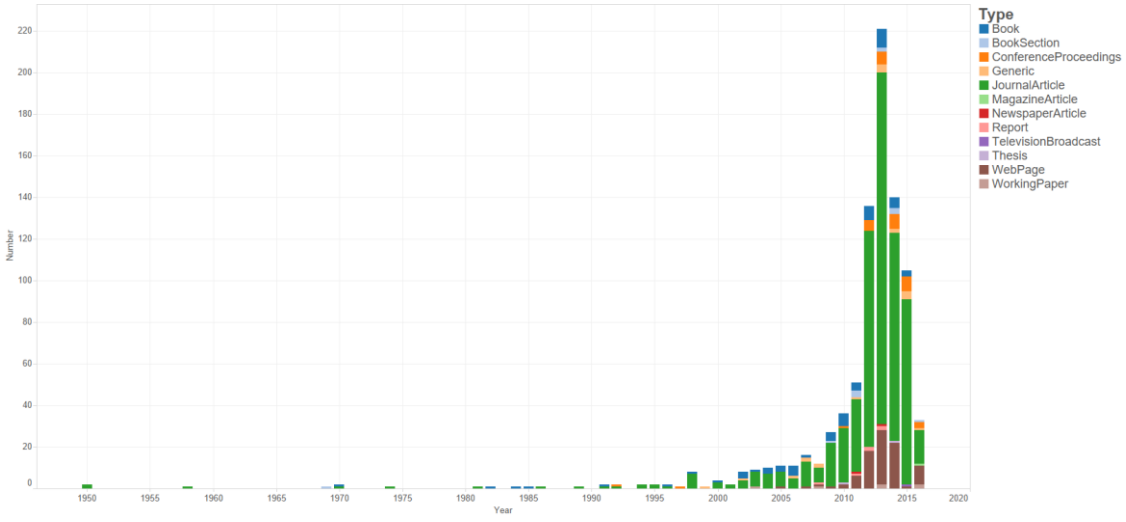

Figure 1 Number of references recorded by Year and Type of publication

Figure 1 shows the temporal spread of references saved in the corpus. Adobe Portable Document

Files (PDFs) saved in Mendeley's reference management software (Mendeley 2016) have been text-mined using Term Frequency-Inverse Document Frequency (TF-IDF) algorithms (Russell 2011) to highlight statistically significant mentions of terms across documents within the corpus. Using the R Project for Statistical Computing (Hornik 2016) and the associated TM (Text-Mining) package (Feinerer et al. 2016) it is possible to effectively process 124MB of text, visualising (McNaught & Lam 2010) the key themes as a Word Cloud (Figure 2).

Figure 2 TF-IDF analysis showing Key Words (Freq>3000) appearing in the academic corpus

Several prominent geo-technological terms are visible, reflecting recent scholarship in Volunteered (Goodchild 2007) and Ambient (Stefanidis et al. 2013) Geographic Information alongside concerns for space, time and network effects which are important components both of OSNs and electoral events.

4. Technological infrastructure

Inherent difficulties in handling Big Data often refer to the 3 or 5 V's of Volume, Velocity and Variety (Laney 2001); Value and Veracity (Demchenko et al. 2013). In this research, Velocity has proved irrelevant as data were analysed *post hoc* following a 'fire-and-forget' data collection exercise. All other V's were present and Veracity, preservation of truth (or prevention of data loss), Volume and Variety have been of critical concern. While the Extract/Transform/Load (ETL) processes to import CSV data are well-established, and may therefore be considered trivial, early experimentation with Microsoft's SQL (Structured Query Language) Server 2012 RDBMS proved anything but. Repeated attempts to load CSV data through the SQL Server 2012 Administrator Graphical User Interface (GUI) or the Bulk Copy (BCP) Command Line Interface (CLI) resulted in data loss either through field truncation or incorrect handling of UTF-8 encoded text, a shortcoming documented on the Web (SQLServerCentral.com 2012) which could not be overcome even with usage of the more advanced SQL Server Integration Services (SSIS, Murray 2013). As the Data Scientist's choice of software stacks and application layers may be strongly influenced by a

combination of familiarity, expediency and inertia (Venkatesh et al. 2007; Agarwal & Prasad 2000) the search for a competing RDBMS solution able to handle large volumes of Web-based UTF-8 encoded data led to experimentation with Oracle and PostgreSQL software.

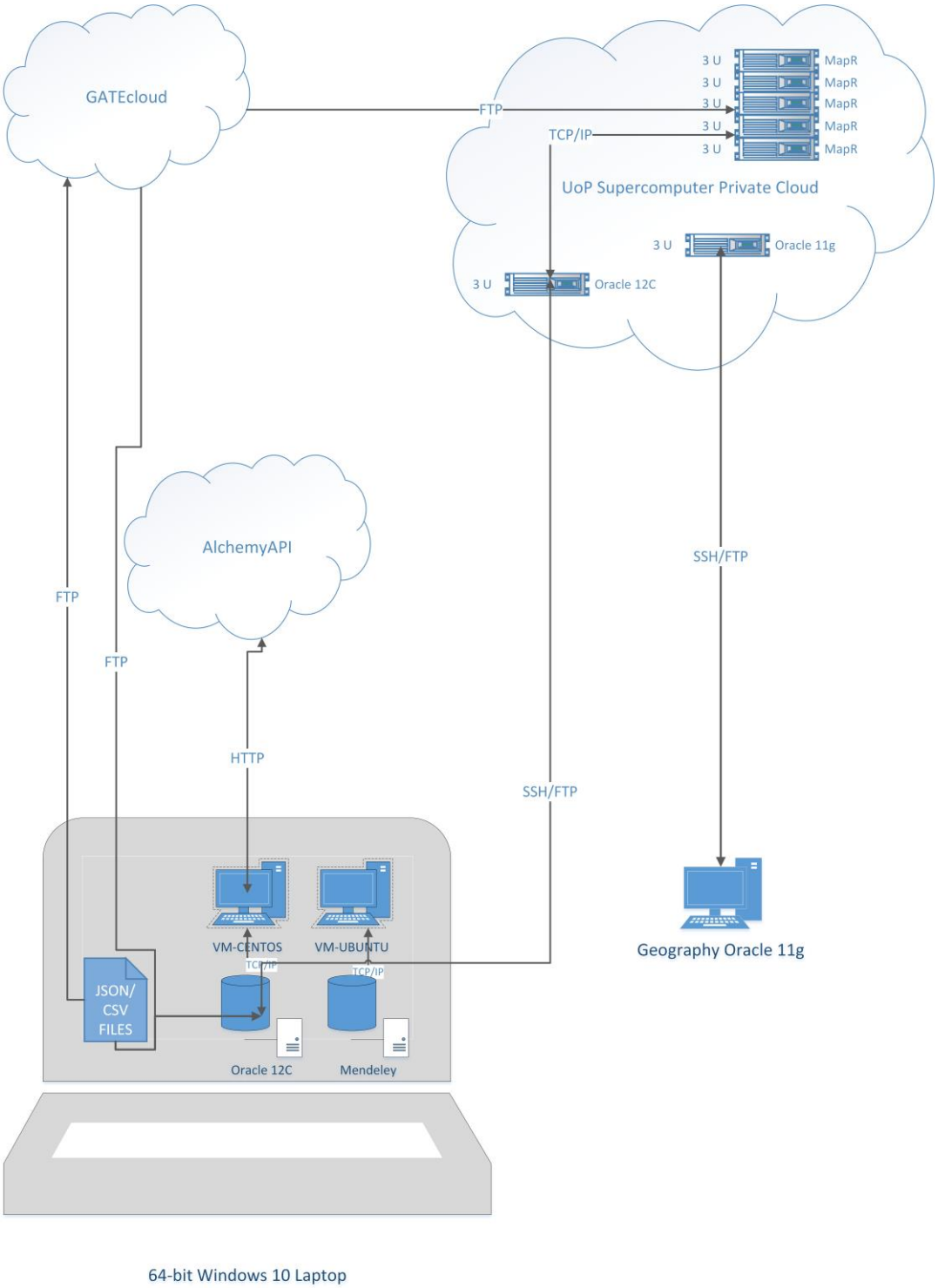

Figure 3 Schematic representation of the system architecture (laptop not to scale)

Only very recently have commercial (Oracle 12.1.0.2.0) and open-source (PostgreSQL 9.3) RDBMSs offered 'native' support for JSON data, the preferred format for OSN data interchange. While PostgreSQL 9.4 Beta (downloaded 11/11/2014), or the operator, proved unable to successfully import DataSift's JSON input files, significant work coding Oracle SQLLDR Control Files eventually

resulted in a verifiably correct import both of CSV and JSON data, the latter stored as a Character Large Object (CLOB) with JSON constraint (Oracle 2014a) populated from an external staging table created by the following SQL statement:

```
CREATE TABLE scot2014_jdump_00001 (json_document CLOB)
ORGANIZATION EXTERNAL
(
TYPE ORACLE_LOADER DEFAULT DIRECTORY scot2014_entry_dir
     ACCESS PARAMETERS
           (RECORDS DELIMITED BY '\n'
           READSIZE 1048576
           CHARACTERSET 'utf8'
           DISABLE_DIRECTORY_LINK_CHECK
           BADFILE scot2014_output_dir: 'JSONDumpFile_00001.bad'
           LOGFILE scot2014_output_dir: 'JSONDumpFile_00001.log'
           FIELDS (json_document CHAR(1048576)))
     LOCATION (scot2014_entry_dir:'part-r-00001.json')
)
PARALLEL
REJECT LIMIT UNLIMITED;
```

Oracle 12C therefore became the *de facto* RDBMS for the project, storing OSN data collected during 2012 US and 2014 Scottish electoral events. Figure 3 shows the system architecture that has been built around this starting point. While Oracle successfully imported JSON data, problems indexing key names greater than 64 bytes forced another search for software capable of efficiently querying JSON in this format; it transpired that many Facebook key names from the 2014 Scottish dataset exceeded Oracle's 64-byte limit (to be increased in future releases of the software) either through representation of complex character sets (e.g. Chinese) or simply lengthy naming in multi-byte/character UTF-8 encoding. Experimentation revealed that Apache Drill (Apache Software Foundation 2014) did not suffer from this limitation, and (very attractively) could query JSON directly from the file using a SQL interface without any ETL import requirements. Unfortunately, in single-user 'embedded' mode (Apache Software Foundation 2016) Drill did not provide the speed of performance required to query large, multi-GB files. This prompted tests (initially on multiple clustered virtual machines under a Windows 2012 host) and the subsequent deployment of Drill within a distributed (MapR 2014) Hadoop 'ecosystem' (Cutting 2013). The architecture, Operating System (OS) and software, tools and flows deployed can be summarised as follows:

- CSV and JSON files flow (by File Transfer Protocol, FTP) from DataSift to Oracle 12C and are imported using SQLLDR or External (staging) Tables respectively.
- Oracle 12C is installed, and works remarkably well, on a commodity Dell Latitude E7440 laptop (Intel i5-4300 CPU @ 1.90GHz with 16GB RAM) running 64-bit Windows 10 equipped with 1*128GB and 1*512GB Solid State Drives.
- The laptop also runs Oracle VirtualBox 5.0.28 which is set up to run:
    - CentOS 7 (1511) with 10GB RAM and 2 CPUs configured to run Ruby:
        - Bespoke Ruby scripts fetch records from Oracle on the host and send/receive data (by Hyper Text Transfer Protocol, HTTP) to/from specialist cloud-hosted AlchemyAPI Natural Language Processing (NLP) software, and;
    - Ubuntu 16.04 LTS with 4GB RAM and 2 CPUs configured to run R and RStudio:
        - Bespoke R scripts read/process/text-mine PDF files stored in the Mendeley reference manager on the host (mapped as a Shared Drive) and are also used to test similar techniques against OSN data re-exported to CSV from Oracle.
- A parallel initiative uses supercomputer resources from the University of Portsmouth's private cloud:
    - Five nodes (each with 12 core CPUs, 24GB of RAM and 2TB of diskspace running Scientific Linux 6) are configured to form a MapR 5.0.0.32987 Hadoop cluster

running Drill and Hive amongst other 'ecosystem' tools. The cluster offers 60 cores, 120GB RAM and 7.5TB storage (available space is lower than total disk space due to replication in the distributed file system). Tests show Hive can successfully (if very slowly) import and query (both in ~24 hours) a ~1.7TB test file with 50 billion rows. Drill is used successfully to query large/complex JSON files directly.
    - Two further supercomputer nodes (specified as above) are configured, one running Oracle 11G, the other Oracle 12C. A large existing Data Warehouse application (Healey 2011) is deployed on 11G with significant speed improvements. The social media OSNDATA database likewise speeds up appreciably when imported to the 12C box.
- JSON data from the original DataSift export flows (by FTP) to the University of Sheffield's GATEcloud servers. TwitIE (Bontcheva et al. 2013) on GATEcloud is used to text-mine OSN tweets and posts using NLP techniques; outputs are downloaded (JSON by FTP) and imported into both Oracle 12C and the MapR cluster for query using Apache Drill. GATEcloud (which shards jobs over multiple servers) is used as the desktop version of GATE is unable to process the number of input records in the files.
- One of the supercomputers (configured with 2*Nvidia GPU cards offering 448 cores and 3GB of RAM/card) is used to run Gephi graph analysis software (Bastian et al. 2009) on large 'Twitter Mentions' networks (>500k nodes, >900k edges) computed from Oracle 12C views. Gephi, like R, runs in-memory; it also runs best with a good graphics cards.

As the project has progressed, the original file-based storage requirements of ~25GB have been supplanted by multiple systems used for storage and analysis. The Oracle 12C database is > 100GB, the NLP output is >25GB and a massive 50 billion row test table occupies >3.5TB on MapR's replicated, distributed file system. Code has been written in SQL, R and Ruby running on machines ranging from laptop through server to supercomputer-cluster. Getting all the computing to work in order to query a large OSN dataset has not been straightforward.

5. Summary

Twitter Data Scientists (Lin & Ryaboy 2013) have described the difficulty in 'plumbing' together complex Big Data software stacks, particularly in real-time/operational environments. In social media analysis, experience suggests that the 'Data Lake' more closely resembles a 'Data Swamp'; it is big and messy, requiring significant skills in Virtual Machine (VM), OS and software setup to create a 'filtering' system capable of revealing meaningful information hidden within otherwise turbid depths. Supercomputers are clearly not accessible to everyone, but offer significant advantages in terms of fast disk Input/Output (I/O), highly parallelised execution and plentiful RAM; sidestepping (by brute force) some of the problems arising from 'hungry' in-memory execution of free, open-source analytical software including R and Gephi. TF-IDF analysis of the entire OSN corpus in R would, for example, require >500GB RAM to hold a large Document Term Matrix in memory. As even the supercomputers in use do not offer this amount of memory, finding solutions to these sorts of problems presents continuing challenges to the analyst. Massively scaling some of the systems already developed on to Public Clouds (e.g. Amazon AWS, Windows Azure) could provide one answer, but costs can rapidly escalate for academic research teams using paid-for services. Instead it seems likely that fast-paced and innovative software development in this domain space (e.g. Oracle Enterprise R to execute R 'in-database') will provide yet another piece of the jigsaw required to analyse the data. Favourable academic licensing terms (Microsoft 2014), open source OS (Ubuntu 2014; CentOS 2014), open source software (Berico-Technologies 2014; GATE 2014; PostgreSQL 2014; The R Project for Statistical Computing 2014) and developer licences for commercial products (MapR 2014; MarkLogic 2014; SAS 2014; Oracle 2014b) now enable rapid test and production builds of effective Data Science stacks on many types of hardware. These new systems and infrastructures enable Data Scientists to more confidently 'wade' through the Big Data swamp, filtering through massive amounts of messy material in search of greater clarity.

6. Acknowledgements

This work arises from Ph.D. research in the Department of Geography, University of Portsmouth. OSN data has been collected from Twitter and Facebook using the DataSift platform. The authors express their gratitude to the ~2.4m unique authors of ~8m OSN posts collected in 2012 and 2014. Gary Burton, High Performance Computing Support Officer in the Institute for Cosmology and Gravitation, has provided invaluable assistance in the setup of a five-node MapR Hadoop cluster on the University of Portsmouth's SCIAMA supercomputer. David Marshall, Principal Database Administrator, University of Portsmouth, has provided invaluable assistance in the setup of Oracle 11g and Oracle 12c RDBMS instances on two further SCIAMA supercomputer nodes. Assistance in the operation of GATEcloud natural language processing software has been provided by Kalina Bontcheva and Ian Roberts, Department of Computer Science, University of Sheffield. Further support has been provided by Richard Pitts and Alastair Fraser of Oracle and Leon Clayton of MapR.

7. Biography

Adrian Tear graduated B.A.(Hons) Geography (Durham) in 1991, gaining his M.Sc. GIS (Edinburgh) in 1992. Following an extremely varied business career, he is currently completing his Ph.D. part time while lecturing at the Universities of Portsmouth and Edinburgh, where he is an Honorary Fellow in the School of GeoSciences.

Richard Healey is Professor of Geography in the University of Portsmouth, before which he taught at the University of Edinburgh, also acting as Co-Director of the ESRC Regional Research Laboratory and Member of the Parallel Computing Centre. Richard's current interests centre on the application of GIS/database techniques to historical railroad and census research.

References

van der Aalst, W.M.P., 2014. Data Scientist: The Engineer of the Future. In K. Mertins et al., eds. *Enterprise Interoperability VI: Interoperability for Agility, Resilience and Plasticity of Collaborations*. Cham: Springer International Publishing, pp. 13–26. Available at: http://dx.doi.org/10.1007/978-3-319-04948-9_2.

Agarwal, R. & Dhar, V., 2014. Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, 25(3), pp.443–448.

Agarwal, R. & Prasad, J., 2000. A field study of the adoption of software process innovations by information systems professionals. *IEEE Transactions on Engineering Management*, 47(3), pp.295–308. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=865899.

Apache Software Foundation, 2014. Apache Drill. Available at: http://incubator.apache.org/drill/ [Accessed October 31, 2014].

Apache Software Foundation, 2016. Drill in 10 Minutes - Apache Drill. *Apache Drill Tutorials*. Available at: https://drill.apache.org/docs/drill-in-10-minutes/ [Accessed November 29, 2016].

Baesens, B., 2014. *Analytics in a big data world: The essential guide to data science and its applications*, John Wiley & Sons.

Bastian, M., Heymann, S. & Jacomy, M., 2009. Gephi: An open source software for exploring and manipulating networks. BT - International AAAI Conference on Weblogs and Social. , pp.361–362.

Berico-Technologies, 2014. Berico-Technologies/CLAVIN · GitHub. Available at: https://github.com/Berico-Technologies/CLAVIN [Accessed October 31, 2014].

Bontcheva, K. et al., 2013. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. *RANLP*. Available at: http://derczynski.com/sheffield/papers/twitie-ranlp2013.pdf [Accessed October 31, 2014].

Castells, M., 2009. *Information Age Series : The Rise of the Network Society, With a New Preface : The Information Age: Economy, Society, and Culture Volume I (2)*, Hoboken, GB: Wiley-Blackwell. Available at: http://site.ebrary.com/lib/portsmouth/docDetail.action?docID=10355273.

Castells, M., 1996. *The Rise of the Network Society, The Information Age: Economy, Society and Culture Vol. I*, Cambridge, MA Oxford, UK: Blackwell.

CentOS, 2014. Download CentOS. Available at: http://www.centos.org/download/ [Accessed October 31, 2014].

Cutting, D., 2013. The Apache Hadoop Ecosystem. Available at: http://assets.en.oreilly.com/1/event/75/The Apache Hadoop Ecosystem Presentation.pdf [Accessed January 30, 2014].

Dahlgren, P., 2005. The Internet, public spheres, and political communication: Dispersion and deliberation. *POLITICAL COMMUNICATION*, 22(2), pp.147–162.

DataSift, 2013. Twitter Data | DataSift Developers. Available at: http://dev.datasift.com/docs/getting-started/data/twitter [Accessed June 20, 2013].

Davenport, T.H. & Patil, D.J., 2012. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, (October 2012), pp.70–77. Available at: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century [Accessed November 29, 2016].

Demchenko, Y., Ngo, C. & Membrey, P., 2013. Architecture Framework and Components for the Big Data Ecosystem. *Journal of System and Network Engineering*, pp.1–31.

Dhar, V., 2013. Data science and prediction. *Communications of the ACM*, 56(12), pp.64–73.

ECMA International, 2013. *ECMA-404 The JSON Data Interchange Format*, Geneva. Available at: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf.

Feinerer, I., Hornik, K. & Artifex Software Inc, 2016. Text Mining Package [R package tm version 0.6-2]. Available at: https://cran.r-project.org/web/packages/tm/index.html [Accessed October 7, 2016].

Foley, J., 2013. OracleVoice: Extreme Big Data: Beyond Zettabytes And Yottabytes - Forbes. Available at: http://www.forbes.com/sites/oracle/2013/10/09/extreme-big-data-beyond-zettabytes-and-yottabytes/ [Accessed January 29, 2014].

GATE, 2014. GATE.ac.uk - download/index.html. Available at: https://gate.ac.uk/download/ [Accessed October 31, 2014].

Goodchild, M.F., 2007. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), pp.211–221.

Graham, M. & Shelton, T., 2013. Geography and the future of big data, big data and the future of

geography. *Dialogues in Human Geography*, 3(3), pp.255–261. Available at: http://dhg.sagepub.com/content/3/3/255.short%5Cnhttp://dhg.sagepub.com/lookup/doi/10.1177/2043820613513121%5Cnhttp://dhg.sagepub.com/lookup/doi/10.1177/2043820613513121.

Healey, R.G., 2011. A Full-Scale Implementation of the NAPP 1880 U.S. Census Data Set Using Dimensional Modeling and Data-Warehousing Technology. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44(2), pp.95–105.

Hornik, K., 2016. {R} {FAQ}. Available at: https://cran.r-project.org/doc/FAQ/R-FAQ.html.

Hough, M.G., 2009. Keeping it to ourselves: Technology, privacy, and the loss of reserve. *Technology in Society*, 31(4), pp.406–413. Available at: http://www.sciencedirect.com/science/article/pii/S0160791X09000815.

Kambatla, K. et al., 2014. Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 74(7), pp.2561–2573. Available at: http://dx.doi.org/10.1016/j.jpdc.2014.01.003.

Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies*, 949(February 2001), p.4.

Li, S. et al., 2016. Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, pp.1–25. Available at: http://arxiv.org/abs/1511.03010.

Lin, J. & Ryaboy, D., 2013. Scaling big data mining infrastructure: the twitter experience. *ACM SIGKDD Explorations Newsletter*, 14(2), pp.6–19. Available at: http://dl.acm.org/citation.cfm?id=2481247 [Accessed February 22, 2014].

MapR, 2014. Quick Installation Guide - Latest Documentation - doc.mapr.com. Available at: http://doc.mapr.com/display/MapR/Quick+Installation+Guide [Accessed October 31, 2014].

MarkLogic, 2014. MarkLogic 7 — MarkLogic Developer Community. Available at: http://developer.marklogic.com/products [Accessed January 30, 2014].

McNaught, C. & Lam, P., 2010. Using wordle as a supplementary research tool. *Qualitative Report*, 15(3), pp.630–643. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-77953058705&partnerID=tZOtx3y1.

Mendeley, 2016. Overview | Mendeley. Available at: https://www.mendeley.com/features/ [Accessed August 31, 2016].

Microsoft, 2014. Microsoft DreamSpark. Available at: https://www.dreamspark.com/ [Accessed October 31, 2014].

Murray, S., 2013. Import UTF-8 Unicode Special Characters with SQL Server Integration Services. Available at: http://www.mssqltips.com/sqlservertip/3119/import-utf8-unicode-special-characters-with-sql-server-integration-services/ [Accessed January 30, 2014].

Ohri, A., 2014. Cheat sheets for data scientists. *SlideShare.net*. Available at: http://www.slideshare.net/ajayohri/cheat-sheets-for-data-scientists/ [Accessed November 29, 2016].

Oracle, 2014a. JSON in Oracle Database. *Oracle Database Online Documentation 12c Release 1 (12.1)*. Available at: http://docs.oracle.com/database/121/ADXDB/json.htm#ADXDB6246.

Oracle, 2014b. Oracle Database Software Downloads | Oracle Technology Network | Oracle. Available at: http://www.oracle.com/technetwork/database/enterprise-edition/downloads/index.html [Accessed October 31, 2014].

PostgreSQL, 2014. PostgreSQL: Downloads. Available at: http://www.postgresql.org/download/ [Accessed October 31, 2014].

Russell, M.A., 2011. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*, O'Reilly Media, Inc.

SAS, 2014. Download SAS University Edition | SAS. Available at: http://www.sas.com/en_us/software/university-edition/download-software.html [Accessed October 31, 2014].

SQLServerCentral.com, 2012. UTF-8 / UTF-16. *SQLServerCentral.com*. Available at: http://www.sqlservercentral.com/Forums/Topic1369404-3077-1.aspx [Accessed November 29, 2016].

Stefanidis, A., Crooks, A. & Radzikowski, J., 2013. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2), pp.319–338.

Steiger, E., de Albuquerque, J.P. & Zipf, A., 2015. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*, 19(6), pp.809–834.

The R Project for Statistical Computing, 2014. CRAN - Mirrors. Available at: http://cran.r-project.org/mirrors.html [Accessed October 31, 2014].

Tsou, M.-H., 2015. Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, 42(sup1), pp.70–74. Available at: http://www.tandfonline.com/doi/full/10.1080/15230406.2015.1059251.

Ubuntu, 2014. Get Ubuntu | Download | Ubuntu. Available at: http://www.ubuntu.com/download [Accessed October 31, 2014].

Venkatesh, V., Davis, F. & Morris, M., 2007. Dead Or Alive? The Development, Trajectory And Future Of Technology Adoption Research. *Journal of the association for …*, 8(4), pp.267–286. Available at: http://aisel.aisnet.org/jais/vol8/iss4/1/ [Accessed October 31, 2014].

Vergeer, M., 2012. Politics, elections and online campaigning: Past, present . . . and a peek into the future. *New Media & Society*, 15(1), pp.9–17. Available at: http://nms.sagepub.com/cgi/doi/10.1177/1461444812457327 [Accessed August 13, 2013].

Yergeau, F., 2003. *UTF-8, a transformation format of ISO 10646*, Available at: http://tools.ietf.org/html/rfc3629 [Accessed October 31, 2014].