1G04

Novel applications of data analytics to higher education.

Lily Yu[1], Mikiko Shimaoka[2], Yasuyuki Koga[2], Azusa Tanaka[3], Naoto Kobayashi[2], John Higginson[1].
Affiliations: 1. Dart Partners, 2. Waseda University, 3. University of Bath

Abstract

Data analytics is revolutionizing various sectors from healthcare and energy to financial services and retail. In line with this trend, higher education institutions around the world are starting to integrate analytics to support decision-making and management in a range of areas including teaching assessment, staff performance, partnership development, finance and resources. In this paper, we introduce ways in which text mining based approaches can be used to understand the global reach of UK research, with particular emphasis on Japan.

Introduction

The production of data is growing at an unprecedented rate. Data analytics is becoming integral to diagnosing and solving business issues in the commercial sector. However, data analytics is also making a difference across many aspects of society. Recent reports from Japan indicate that data analytics is contributing to various areas of public service, from rare cancer diagnosis and healthcare management, to match making and marriage services (The Japan Times, 2016) (The Japan Times, 2015).

Data analytics is also starting to have an impact on the management of higher education institutions. Analytics is aided by the fact that a wide variety of data is already being collected at universities. This ranges from data on students (e.g. admissions, course enrolment, study performance and completion statistics, alumni) and staff (e.g. teaching evaluation, demographics, grants and funding) to research output metrics (e.g. papers, publications and other bibliometric measures). Moreover, data analytics can be used to complement existing decision making processes in core strategic areas including research management, student management and institutional strategy (see Fig. 1).

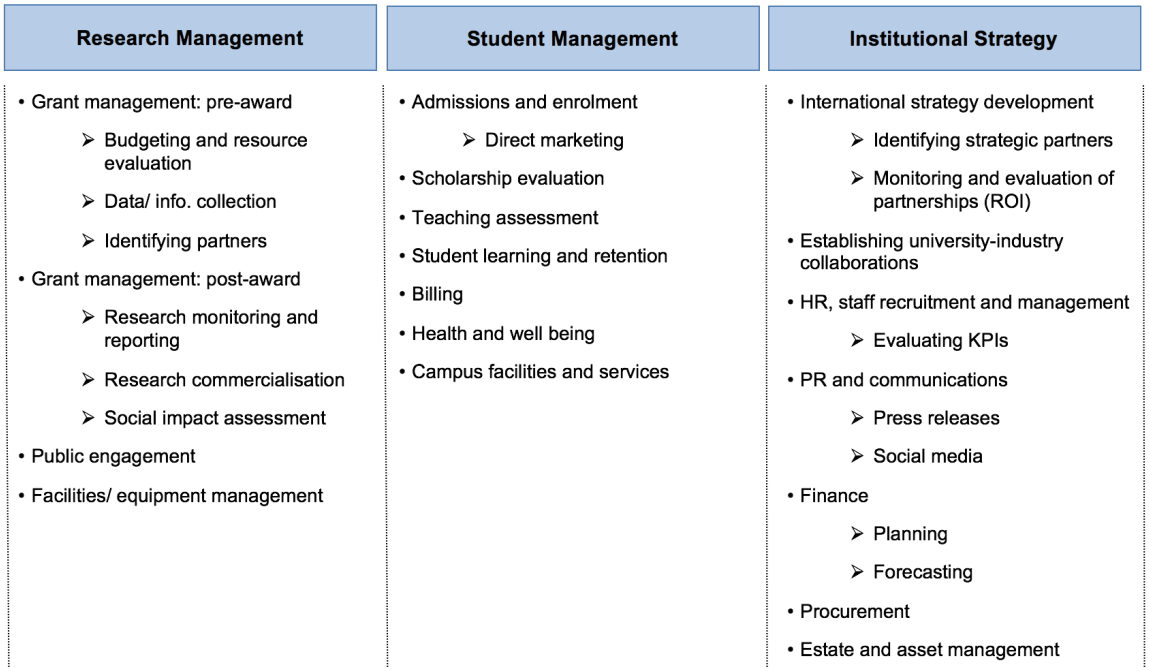


Fig 1. Areas of Higher Education Management that are benefitting from data analytics.

In this paper, we introduce ways in which text mining based approaches can be used to understand the characteristics of UK research and its global reach, with particular emphasis on Japan.

Using text mining approaches to understand the characteristics of UK research.

The national evaluation of university research is a mechanism to understand how research is contributing to both academic progress and the impact on wider society. One well-documented example of university research assessment is the UK Research Excellence Framework (REF). REF (and its predecessor, the Research Assessment Exercise (RAE)) is a national assessment exercise that has taken place every 5-7 years for the past three decades. During REF 2014, the research of 154 universities, accounting for 52,061 academic staff and

around 191,150 examples of research output, was assessed and ranked according to overall quality, to determine the distribution of £1.58 billion budget for university based research.

Additionally, REF impact case studies (RICS), a new component of assessment, required universities to submit case studies describing how the underlying research led to the creation of impact in the areas of politics, health, technology, economics, law, culture, society and the environment. RICS accounted for 20% of the total evaluation score and can be viewed in a publically accessible and searchable database of 6,637 case studies (HEFCE, 2014). Due to the scale (over 6 million data points) and structure (largely free form text) of the content, the database lends itself well to meta-analysis using a combination of natural language processing, text mining and machine learning approaches to investigate deeper relationships between the case studies.

A collaborative project commissioned by the Higher Education Funding Council for England (HEFCE) and directed by Kings College London and Digital Science, used a multifaceted text mining approach in combination with qualitative methods to examine the types of impact, pathways to measuring impact and beneficiaries of university research. Fig 2 outlines the general methodology used for automated text mining of REF impact case studies.

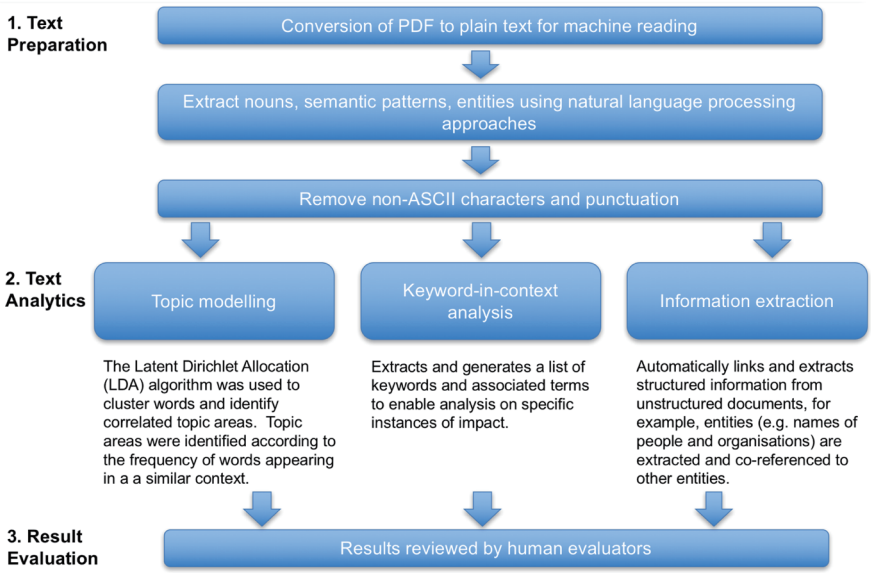
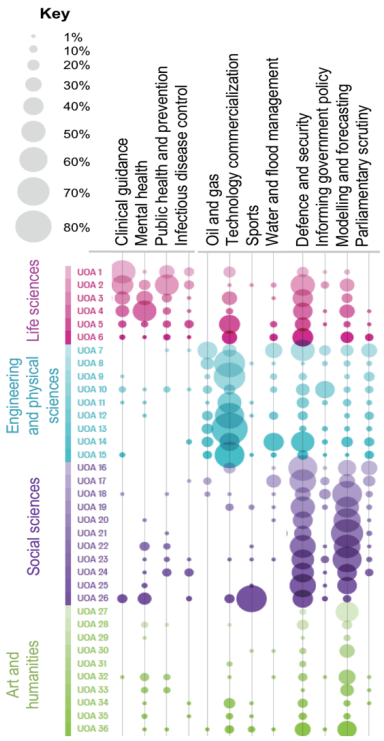


Fig 2. General methodology for automated text mining of impact case studies.
Source: modified from KCL/ Digital Science (King's College London and Digital Science, 2015).



Through topic modelling, KCL and Digital Science revealed that the societal impact of UK university research was diverse and wide-ranging. Fig 3 shows a heat map of selected topics by units of assessment (UOA, see appendix i for complete list), identified by text mining of case studies (see KCL/ Digital Science report for full map). The circles represent the proportion of case studies that have been allocated to a topic for a given UOA, and each UOA is grouped into a discipline category (e.g. life science: UOA 1-6 (pink), engineering and physical science: UOA 7-15 (cyan), social science: UOA 16-26 (purple), art and humanities: UOA 27-36 (green)). The figure shows that some topic areas were highly discipline specific e.g. under "clinical guidance" the majority of text references (indicated as the percentage of case studies per UOA) were associated with UOAs within life sciences, while "oil and gas" mainly contained references in the engineering and physical sciences category. However, other topic areas contained text that spanned several UOAs, e.g. "technology commercialization" contained topics that primarily spanned life sciences and engineering and physical sciences, and "informing government policy" was highly disciplinary with text spanning all UOA fields.

Fig 3. Heat map of selected topics. Circles indicate the percentage of case studies per UOA. Modified from KCL/ Digital Science 2015.

Using RICS to understand the global reach of UK university research.

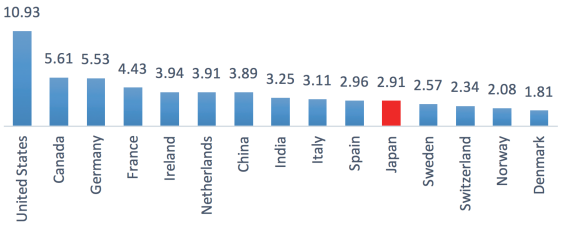


Fig 4. Top 15 countries in RICS classified according to global location. Data shown as a percentage of all tagged global locations.

Automated text indexing was used to classify RICS according to global location, based on references (tags) to geographical place names in the text, which were then grouped by country (see (HEFCE, 2016) and appendix ii for data classification methods). Classified case studies could be viewed in the REF database, according to country. By evaluating the number of references by country, it was possible to calculate the proportion of case studies with references to different foreign countries, and by proxy estimate the international reach of UK research.

The United States was the most highly referenced country, accounting for 10.93% of references, followed by Canada with 5.61% and Germany with 5.53%. China and Japan were the top countries in Asia with 3.89% and 2.91% of references respectively (Fig 4). The remaining sections of this paper will largely focus on analysis of Japan tagged RICS to evaluate the potential impact of UK research on Japan.

Medical and health sciences is the most highly represented research area in Japan tagged RICS.
Further analysis of RICS according to global location revealed heterogeneity in the representation of different research areas. Quantification revealed that the most highly referenced research area in Japan tagged RICS was medical and health sciences accounting for 20.33% of references (Fig 5). Language, communication and culture (8.37%); engineering (8.11%); and information and computing sciences (7.72%) were the next most commonly referenced research areas.

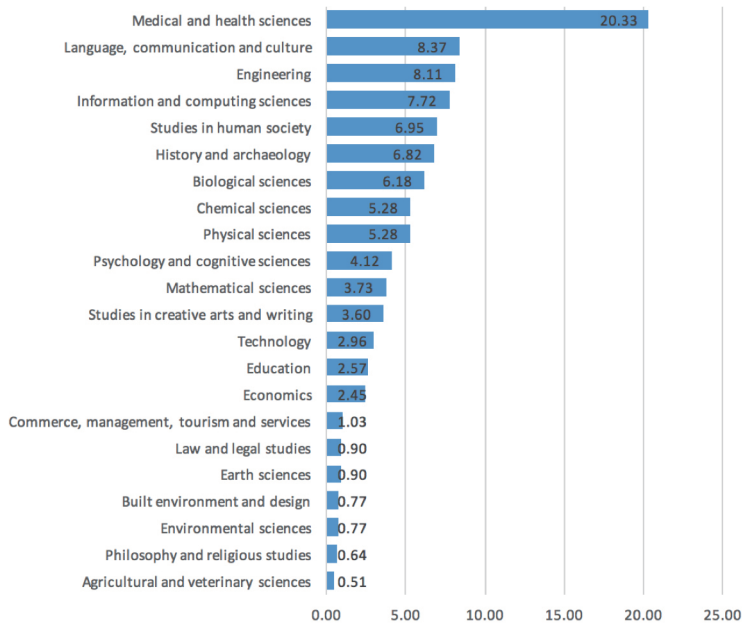


Fig 5. Research areas represented in Japan tagged RICS.
Values correspond to the percentage of research area references from all Japan tagged RICS.

RICS provide insights into the type of impact associated with UK research.
An aim of the RICS was to gain a better understanding of the type of impact UK research creates. Based on text mining approaches, all RICS were assigned to one of eight impact types: technological, health, cultural, societal, economic, political, environmental and legal.

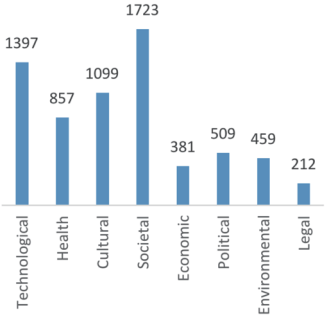


Fig 6. The number of RICS per impact category.

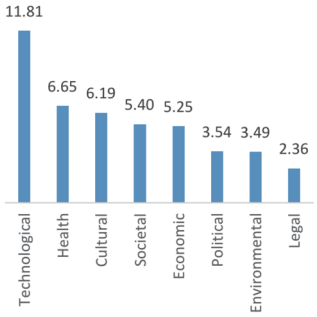
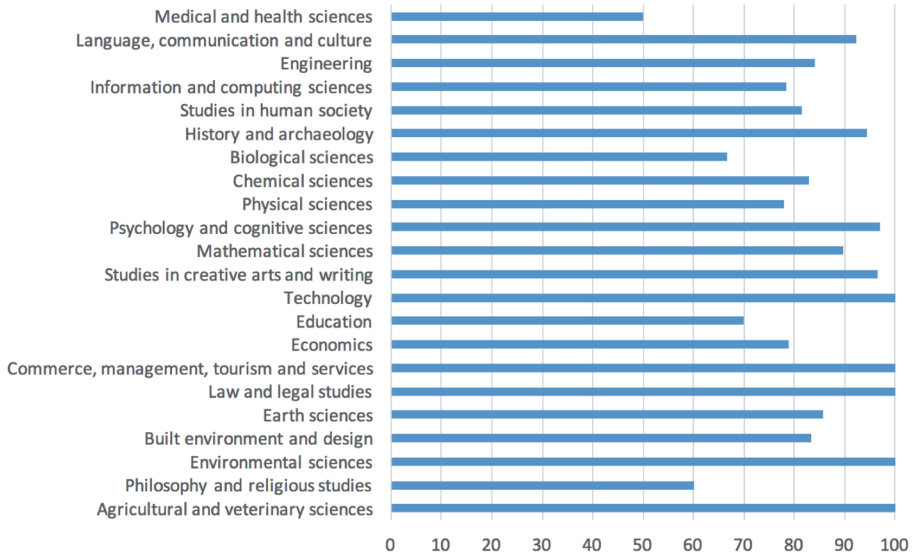


Fig 7. The percentage of Japan tagged RICS according to impact category

Analysis of all RICS revealed that the majority of case studies created societal, technological and cultural impact (Fig 6). However, analysis of Japan tagged RICS showed that technological, health and cultural impact accounted for the largest share (Fig 7).

UK research associated with Japan is highly interdisciplinary.
Reports indicate that interdisciplinary research is necessary to tackle some of the global challenges facing society (Ledford, 2015) (Van Noorden, 2015). To better understand the interdisciplinarity of UK research, text mining was used to classify RICS associated with multiple research subject areas as interdisciplinary. Analysis of Japan tagged RICS showed that the majority of UK research associated with Japan was highly interdisciplinary (Fig 8).



Interestingly, 100% of Japan tagged RICS in the research areas of technology; commerce, management, tourism and services; law and legal studies; environmental sciences; and agricultural and veterinary sciences were interdisciplinary (Fig 8).

Fig 8. The percentage of Japan tagged RICS that are interdisciplinary, according to research subject area.

Discussion and future directions
The REF 2014 impact case studies (RICS) database is a rich source of information on UK university research and lends itself well to natural language processing and text mining techniques. As demonstrated by the KCL/ Digital Science collaboration, topic modelling can reveal insights into the disciplinary make-up of different research topics and taken together showed that UK research is highly diverse and multidisciplinary (Fig 3). Further analysis showed that frequent references to foreign countries were made in RICS, indicating that UK research has significant global reach (Fig 4). By focusing on Japan tagged RICS, it was possible to deduce that medical and health sciences was a key area for UK research associated to Japan (Fig 5). The database also enabled RICS to be classified based on technological, health, cultural, societal, economic, political, environmental and legal impact. Analysis of all RICS showed that the largest number of references related to societal, technological and cultural impact (Fig 6). However, analysis of Japan tagged RICS revealed slightly different trends with technological, health and cultural impact being the most highly referenced areas, suggesting that Japan associated UK research has a greater focus on these areas (Fig 7). Finally, by examining Japan tagged RICS it was shown that Japan associated UK research is highly interdisciplinary (Fig 8).

REF 2014 was the first time that UK research was evaluated based on impact. The recent independent review of REF 2014 lead by Lord Stern found that the introduction of impact case studies provided valuable insights into the wider achievements of UK research and could be used to foster institutional strategy to support greater societal engagement by researchers (Stern, 2016). It would be of interest to determine how impact changes over time by tracking the progress of case studies from one exercise to the next.

This paper aims to introduce some of the ways text processing can be used to understand research impact. However, it is valuable to note that whilst text mining allows tentative associations to be made, human review is important to validate associations. Additionally, it should be noted that whilst large data sets contain much valuable information, pre-sorting and structuring of data is often required prior to automated processing (Fig 2). Text mining techniques can also complement other research evaluation approaches, such as bibliometrics, to gain a deeper understanding of research impact and interdisciplinarity (Elsevier, 2015).

Using the REF database to identify new research collaborations.
Research collaborations are often established as a result of common research interests. It is often challenging for researchers to identify suitable partners based on common impact objectives, especially amongst researchers in different disciplines. As the REF database is uniquely positioned to facilitate searches based on impact and interdisciplinarity, it could be a valuable starting point for researchers and companies to explore new areas for UK-Japan research collaboration and university-industry partnerships.

Assessment of research conducted by universities in Japan.
One of the main challenges to conducting data analytics is the availability of data. The REF database is unique in terms of the volume and variety of research information. To our knowledge, there is no publically accessible data of a similar scale on research conducted by Japanese universities. However, there are several smaller scale sources of data on Japanese research. For example, recipients of the JSPS Grant in aid for scientific research (Kakenhi) are required to submit annual progress reports, which would allow analysis to be performed over the duration of a grant. Of particular relevance to understanding the social contribution of research is 各法人の評価結果 data published by MEXT, which features research conducted by national universities in Japan, and could potentially be used to explore how university research is contributing to regional growth, university-industry partnerships and other impact metrics.

Other applications of data analytics in higher education
Learning analytics
Besides evaluating research impact, data analytics is also being applied to many other areas of university management. Predictive analytics is being used by universities to support student learning and development, a term known as learning analytics. Student dashboards, which help students visualize their learning progress are one example of learning analytics. For example, at Nottingham Trent University (NTU) UK, virtual dashboards have been developed to help students visualize their level of engagement by integrating data including access to the university's virtual learning environment, class attendance, library and e-book usage and door swipes (Nottingham Trent University , 2016) (JISC, 2016). By analyzing aggregate data, it was discovered that the level of engagement was positively correlated with learning. Hence, the engagement score is seen as an indirect predictor of whether a student is at risk of underperforming or dropping out of the course. However, learning analytics is still very much in its infancy, with a recent study showing that less than 2% of UK universities have fully implemented learning analytics, whilst a further 51% have either partially implemented or are working towards implementation (Higher Education Commission, 2016).

Strategic planning
Data analytics has also been used to develop institutional strategic plans. As a way to illustrate the university's strategic plan, Texas A & M University, US, is using business intelligence and visualization software to make key indicators and aggregated data publically accessible (Texas A & M University., 2016). A comprehensive collection of metrics can be viewed through interactive graphs and maps, including student metrics (e.g. demographic information, enrollment, retention and graduation rates), faculty metrics (e.g. funding and grants, headcount, demographics) and statistics relating to specific aspects of the institutional strategy (e.g. research expenditures, faculty accomplishments, endowed assets and alumni giving rates, retention and graduation rates).

As data at universities often exists across different departments and schools, the value of analytics will inevitably come from cross-departmental integration of academic (e.g. publication output, grants) and non-academic (e.g. finance, HR, facilities) information. Business intelligence and data visualization software will become increasingly important to support data-driven decision making. For example, as part of the Top Global University Project, the office of institutional research at Sophia University in Tokyo has used Tableau software to analyze both curriculum and administrative data to support student admission and the university's globalization strategy (Tableau, 2016). In light of demographic change, universities around the world are facing increasing pressure to attract the best students and staff in an increasingly competitive business. The use of data analytics could help universities to better evaluate and understand their strengths and achievements, which not only supports reporting, but more significantly, the ongoing development and monitoring of institutional strategy, goals and partnerships.

Appendix i: Unit of Assessments (UOA).

Modified from (King's College London and Digital Science, 2015)

	Unit of Assessment	# of case studies	% of total
	Life Sciences		
UOA 1	Clinical medicine	383	6%
UOA 2	Public health, health services and primary care	163	2%
UOA 3	Allied health professions, dentistry, nursing, pharmacy	347	5%
UOA 4	Psychology, psychiatry, neuroscience	318	5%
UOA 5	Biological sciences	257	4%
UOA 6	Agriculture, veterinary and food science	126	2%
	Physical Sciences and Engineering		
UOA 7	Earth systems and environmental sciences	171	3%
UOA 8	Chemistry	125	2%
UOA 9	Physics	182	3%
UOA 10	Mathematical sciences	210	3%
UOA 11	Computer science and informatics	253	4%
UOA 12	Aeronautical, mechanical, chemical and manufacturing engineering	124	2%
UOA 13	Electrical and electronic engineering, metallurgy and materials	127	2%
UOA 14	Civil and construction engineering	51	1%
UOA 15	General engineering	242	4%
	Social Sciences		
UOA 16	Architecture, built environment and planning	141	2%
UOA 17	Geography, environmental studies and archeology	235	4%
UOA 18	Economics and econometrics	98	1%
UOA 19	Business and management studies	413	6%
UOA 20	Law	217	3%
UOA 21	Politics and international studies	167	3%
UOA 22	Social work and social policy	187	3%
UOA 23	Sociology	97	1%
UOA 24	Anthropology and development studies	80	1%
UOA 25	Education	214	3%
UOA 26	Sport and exercise sciences, leisure and tourism	124	2%
	Art and Humanities		
UOA 27	Area studies	69	1%
UOA 28	Modern languages and linguistics	190	3%
UOA 29	English language and literature	283	4%
UOA 30	History	263	4%
UOA 31	Classics	59	1%
UOA 32	Philosophy	98	1%
UOA 33	Theology and religious studies	75	4%
UOA 34	Art and design: history, practice and theory	236	4%
UOA 35	Music, drama, dance and performing arts	196	3%
UOA 36	Communication, cultural and media studies, library and information management	158	2%

Appendix ii: Information on data sources

Data for figures 4-8 was taken from the REF 2014 database (HEFCE, 2014) based on the following:
Fig 4: data points for all countries under the category of Impact Global Location were used.
Fig 5: data points for all subjects under the category of Research Subject Area, filtered to Japan under the category of Impact Global Location were used.
Fig 6: data points for all impact types under the category of Summary Impact Type were used.
Fig 7: data points for all subjects under the category of Research Subject Area, filtered to Japan under the category of Impact Global Location and interdisciplinary Case Studies were used.
Fig 8: data points for all subjects under the category of Research Subject Area, filtered to Japan under the category of Impact Global Location were used.

Bibliography

Elsevier. (2015). *A review of the UK's interdisciplinary research using a citation-based approach.* Elsevier.

HEFCE. (2014). *REF 2014 impact case studies*. Retrieved from http://impact.ref.ac.uk/CaseStudies/

HEFCE. (2016, 09 05). *REF2014 Impact Case Studies*. Retrieved from REF2014 Impact Case Studies: http://impact.ref.ac.uk/CaseStudies/FAQ.aspx#globalloc

Higher Education Commission. (2016). *From bricks to clicks: the potential of data and analytics in higher education.* London: Policy Connect.

JISC. (2016). *Case Study I: Predictive analytics at Nottingham Trent University.* JISC.

King's College London and Digital Science. (2015). *The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies.* Bristol: HEFCE.

Ledford, H. (2015, 09 21). *How to solve the world's biggest problems* . Retrieved from Nature: http://www.nature.com/news/how-to-solve-the-world-s-biggest-problems-1.18367

Newland, B., Martin, L., & Ringan, N. (2015). *Learning Analytics in UK HE 2015, a HeLF Survey report.*

Nottingham Trent University . (2016, 09 08). *NTU Student Dashboard*. Retrieved from Nottingham Trent University: https://www4.ntu.ac.uk/adq/running_a_course/student-success/ntu-student-dashboard/index.html

Stern, L. (2016). *Building on Success and Learning from Experience. An Independent Review of the Research Excellence Framework.* UK GOV. UK Department for Business, Energy & Industrial Strategy.

Tableau. (2016, 09 05). *Sophia University: Tableau broadens perspectives, promotes IR activities*. Retrieved from https://www.tableau.com/stories/customer/tableau-broadens-perspective-university-staff-member-and-promotes-ir-activities

Texas A & M University. (2016, 09 09). *Accountability*. Retrieved from Measuring the pursuit of excellence: https://accountability.tamu.edu/All-Metrics

The Japan Times. (2015, 12 31). *Big data used to support marriage hunting in Japan*. Retrieved from The Japan Times: http://www.japantimes.co.jp/news/2015/12/31/national/big-data-used-support-marriage-hunting-japan/#.V9Uof5N96Rs

The Japan Times. (2016, 08 11). *IBM big data used for rapid diagnosis of rare leukemia case in Japan*. Retrieved from The Japan Times: http://www.japantimes.co.jp/news/2016/08/11/national/science-health/ibm-big-data-used-for-rapid-diagnosis-of-rare-leukemia-case-in-japan/#.V9Uof5N96Rs

Van Noorden, R. (2015, 09 16). *Interdisciplinary research by the numbers* . Retrieved from Nature: http://www.nature.com/news/interdisciplinary-research-by-the-numbers-1.18349