Institutionalizing Big Data methods in social and political research

Ahonen, Pertti Pellervo

2015-06-30


http://hdl.handle.net/10138/178745
https://doi.org/10.1177/2053951715591224

Downloaded from Helda, University of Helsinki institutional repository.
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
Institutionalizing Big Data methods in social and political research

Pertti Ahonen

Abstract
We expect Big Data methods to contribute to research with results that are not inferior to those attained in other ways but possibly better, or hard or impossible to generate in other ways. Those who apply these methods may also aspire to augment the arsenal of research methods, offer surrogates for existing research designs, and re-orient research. Moreover, we can critically examine the institutional, societal and political effects of the Big Data methods and the conditions for the solid institutionalization of these methods in social and political research. To reach its primary objective, this article elaborates conclusions on how Big Data methods, not only by means of their ‘social life’ but also by their ‘political life’, may influence the institutionalization of social and political research. To reach its secondary objective, the article re-examines a study of budgetary legislation in 13 countries carried out by means of Big Data methods to draw conclusions concerning the augmentation of the arsenal of research methods, the surrogation of existing research designs, and the re-orientation of research.

Keywords
Digital social research, institutional analysis, neo-institutionalism, institutional change, latent trait scaling, topic modeling

Introduction
We expect Big Data methods to contribute to research with results that are not inferior to those attained in other ways but are possibly better, or hard or impossible to generate in other ways. Those who apply these methods may also aspire to use them to augment the available arsenal of research methods, offer surrogates for existing research designs, and re-orient research (Edwards et al., 2013). Moreover, we can critically examine the direct and indirect societal and political effects and conditions of the institutionalization of Big Data methods (Law and Ruppert, 2013).

The primary objective of this article is to elaborate conclusions on how Big Data methods, not only by means of their ‘social life’ (Law and Ruppert, 2013) but also their ‘political life’ (a notion that will be explained in the next section), influence the institutionalization of social research with special reference to political research. To support its pursuit of its primary objective, the article also pursues a secondary objective, comprised of the re-examination of a study of budgetary legislation in 13 countries carried out by means of Big Data methods. The purpose of this re-examination is to elaborate conclusions on the augmentation of the arsenal of research methods, the surrogation of existing research designs, and the re-orientation of research.

The following two sections elaborate the theoretical rationale of this article and reproduce the research hypotheses of the study that the article re-examines. The subsequent section and Appendix 1 introduce the Big Data methods and the research material of the re-examined study. Utilizing the re-examined study as a context, this article examines in its next two sections two methods of Big Data analysis, comprising a method of unsupervised latent trait scaling and a method of topic modeling. The purpose of the last section is to summarize the contributions of this article.
The theoretical rationale of this article

In important respects this article leans on neo-institutional analysis with no fewer than seven to twelve present-day orientations (Lowndes and Roberts, 2013; Peters, 2010). Making its pick from among this multitude, the article follows authors starting from Berger and Luckmann but principally comprising John W Meyer and his colleagues and followers (see Powell and Colyvas, 2008). Berger and Luckmann (1991) understood institutionalization to take place by means of habituation and resulting taken-for-grantedness (p. 72): ‘Institutionalization occurs whenever there is reciprocal typification of habitualized actions by types of actors. Put differently, any such typification is an institution.’

This article shares the assumption that not only human beings but also artifacts and therefore also Big Data methods and Big Data itself play active roles in research and other institutionalized action (generally see, for instance, Cecez-Kecmanovic et al., 2014; D’Adderio, 2011; Latour, 2005). The article fixes its attention to institutionalization related to Big Data methods in an analogous although not identical way to Law and Ruppert (2013), who elaborate the examination of the ‘social life’ of methods by means of characterizing what they call the relevant ‘patterned teleological arrangements’. Moreover, this article agrees with boyd and Crawford (2012) that mythologies of institutionalization in Big Data contexts need study, and therefore pays attention to the examination of ‘rationalized myths’ of Big Data, which rather than fulfilling common explicit promises to contribute to rationality support institutional legitimation (Meyer and Rowan, 1977).

This article not only seeks to examine the habitually institutionalized social life of Big Data methods but also the ‘political life’ of disruptions and radical transformations related to these methods. The article shares the understanding of politics of Pocock (1975: 156) as the ‘art of dealing with the contingent event... with pure, uncontrolled, and unlegitimated contingency’, and also fixes its attention to the role of performativity (Austin, 1975) that catalyzes the disruptions and transformations indicated. We find analogous arguments on performativity in political theory (Skinner, 2009), science and technology studies (D’Adderio, 2008, 2011; Latour, 2005), and organization research (D’Adderio, 2014; Deroy and Clegg, 2015; Maguire and Hardy, 2009). The authors indicated locate disruptive, radical institutional change in situations in which contentious actors, finding that unprecedented opportunities have opened up, mobilize performatives by means of which these actors may successfully de-legitimate the incumbent actors and those performatives that have helped entitle the latter actors to their positions.

The rationale of this article also pushes it to examine chances to augment the arsenal of existing methods, offer surrogates for known research designs, and more generally re-orient research (see Edwards et al., 2013). A common motivation to use Big Data methods is to try to reap economies of scale in managing large datasets, data dredging, and implementing research designs. Moreover, Big Data methods may lead researchers to transcend boundaries between research fields and re-orient research in other ways (Ruppert, 2013).

The hypotheses proposed in the study that this article re-examines

Big Data methods can be used either in exploratory analysis (O’Neil and Schutt, 2013) or explanatory research, the latter being the case in the study that this article re-examines. The empirical subject matter of the study indicated comprised a ubiquitous trite practice, namely government budgeting. However, the objectives of the study that this article re-examines had pushed it away from entrenched research on budgetary governance (de Haan et al., 2013; Hallerberg et al., 2007) or adaptations of generic research on public sector reform adapted to examine government budgeting (see, for instance, Hyndman et al., 2013). Instead, the study indicated examined historical legal traditions as possible influences on government budgeting.

The study that this article re-examines started with observations that despite indications that legal traditions are dead (Lindahl and Schadewitz, 2013; Pargendler, 2012), research acknowledging these traditions continues (Ma, 2012; Painter and Peters, 2010). Moreover, heterogeneity between national systems of government budgeting despite decades of global harmonization called for explanation (International Monetary Fund (IMF), 2013; Jones et al., 2013; Lienert, 2013; OECD, 2005; Wanna et al., 2010). The indicated study proposed hypotheses that legal system traditions indeed explain differences in government budgetary legislation according to entrenched historical, social, cultural, and political divisions in common law with British origins, Napoleonic civil law first codified in France, civil law of the German type, and Nordic law resembling German law in many respects (Glenn, 2010; Zweigert and Kötz, 1998). The study that this article re-examines approached differences between the legal systems taking the less common inroad of examining differences in legal language (Brake and Katzenstein, 2013; Kischel, 2009).

The first three hypotheses proposed in the study that this article re-examines were tested by means of a
Big Data method of unsupervised latent trait scaling. The first two of these three hypotheses were:

- Hypothesis 1: A grand division into common law legal systems and civil law legal systems explains differences in government budgetary legislation.
- Hypothesis 2: Within countries of civil law traditions, finer divisions into Napoleonic, German and Nordic legal systems explain differences in government budgetary legislation.

The longer-term harmonization efforts of government budgeting indicate countries with common law traditions as the foremost present-day global ideal models (IMF, 2013; Wanna et al., 2010). A third hypothesis was proposed:

- Hypothesis 3: The older the budgetary legislation is in a civil law country, the wider its differences from budgetary legislation in common law countries.

The last two hypotheses of the study that this article re-examines were tested by means of a Big Data method of topic modeling. These hypotheses were:

- Hypothesis 4: Divisions into legal system traditions explain divisions into topics in government budgetary legislation.
- Hypothesis 5: Divisions within government budgetary legislation according to different topics explain differences in the vocabularies of these topics in the legislation indicated.

**Research methods and the research material in the study that this article re-examines**

**Methods**

The foremost methods of the study that this article re-examines comprised a method of unsupervised latent trait scaling and a method of topic modeling. A brief background characterization follows, and Appendix 1 gives more details.

Latent trait scaling first appeared in social research of the political science variety in the shape of supervised latent scaling (Laver et al., 2002), which requires the researcher first to feed a ‘seed text’ and next utilize machine learning algorithms to do the scaling. Later, unsupervised scaling evolved (Proksch and Slapin, 2008), using machine learning algorithms throughout the research process. First and foremost, both two scaling methods have been applied to examine ideological polarization in politics by means of using such Big Data textual materials as parliamentary speeches, political statements of government ministers, or political party programs (Lowe and Benoit, 2013).

Topic modeling found its way into social research at about the same time as unsupervised latent trait scaling evolved within the political science variety of this research (Grimmer and Stewart, 2013). Rather than within political science research (Clark and Lauderdale, 2010), this method has found applications in sociological research (see, for instance, DiMaggio et al., 2013; Fligstein et al., 2014; Levy and Franklin, 2013).

**Research material**

The research material in the study that this article re-examines comprised legislation on federal or national government budgeting in the stable and highly developed Western democracies of Australia, Austria, Canada, Estonia, Finland, France, Iceland, Italy, the Netherlands, New Zealand, Sweden, Spain, and the United Kingdom (Table 1). It would have been both relevant and interesting to examine the United States, Germany, Switzerland, and Belgium, but these countries were omitted because of their complex federal structures. The corpus examined had 13 documents with a total of 270,000 words.

Most of the materials were derived from a World Bank (2013) website with legal texts in English or English translation. Two texts derived were from governmental websites, in English translation (Austria, 2013) or in Italian (Italy, 2013), of which the latter was translated into English. Applying a common and allowable practice of Big Data research for the UK and Finland, two separate legal documents were merged into a single text.

Examining augmentation, surrogation, and re-orientation in research by means of Big Data methods

**Examining unsupervised latent trait scaling**

The method of unsupervised latent trait scaling used in the study that this article re-examines (for technical details see Appendix 1) enables delivering estimating and testing results on what the developers of this method have named the fixed effects related to individual words (\(\theta\), theta), the fixed effects related to each different document included in the data (\(\alpha\), alpha), the word weights (\(\beta\), beta), and the positions taken in the documents (\(\omega\), omega). The notion of ‘fixed effect’ is generic to many varieties of quantitative analysis, indicating that a variable is treated for the technical purposes of quantitative analysis as if it were non-random. All that is specific to estimating fixed effects by means
of latent trait scaling derives from the character of this method as a Big Data method of textual analysis. The results obtained by means of using this method are typically uninteresting as concerns the fixed effects, the calculation of which therefore plays only a subordinate technical role to enable the calculation of the word weights and the positions taken in the documents. Moreover, the calculation of the fixed effects may essentially enhance the possibilities graphically to display the estimation of the word weights, the positions taken in the documents, or both the weights and the positions.

Words best discriminating between the documents examined and therefore between the countries from which these documents derive situated themselves at the extreme ends of $\beta$ values, as is generally expected in applications of the unsupervised latent trait scaling method. This pinpoints an asset of this method: it emphasizes rare as opposed to frequent words discriminating between text documents. This very characteristic also eliminates the confounding effects of the different lengths of the texts examined. However, the assets of the unsupervised latent trait scaling examined must be weighed against its less outstanding characteristics, which included a graphical output plot with no fewer than 2992 different words according to their weights ($\beta$), or far too many either to present meaningfully or to make subject to a sensible substantive interpretation. In their turn, the $\alpha$ (alpha) fixed effects, related to documents and the countries behind these documents, are of no interest for the model interpretation, but the dimension they represent usefully takes visually apart the positions taken in the documents, or the important $\omega$ (omega) values (Figure 1).

As proposed in Hypothesis 1 of the study that this article re-examines, a grand division indeed prevails between countries with common law traditions and countries with civil law traditions. Hypothesis 2 could also be sustained, although the countries examined do not arrange themselves quite neatly into the Napoleonic, German, and Nordic variants of civil law. Certain countries sharing legal system traditions received resembling $\omega$ (omega) values, such as Spain and Italy, or Sweden and Finland, but there were exceptions such as Iceland more resembling Spain and Italy than the two other Nordic countries of Sweden and Finland. Moreover, Austria, one of the two representatives of the German tradition in the analysis, received a resembling $\omega$ (omega) value with Sweden and Finland. The Netherlands, with its Roman–Napoleonic–German legal characteristics (Glenn, 2010), comprised a unique case as could be expected. The fact that budgetary legislation was old at the time of the investigation both in Estonia and Iceland suggested the acceptance of Hypothesis 3 on differences between civil law countries with older as opposed to more recently reformed budgetary legislation.

In the study that this article re-examines, statistical credibility intervals were calculated for the $\omega$ (omega)

---

### Table 1. The text material of the study that this article re-examines.

<table>
<thead>
<tr>
<th>Country</th>
<th>Legal tradition</th>
<th>Titles of legal texts examined, in English</th>
<th>Latest upgrade</th>
<th>Words in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>Common law</td>
<td>Financial Management and Accountability Act 1997</td>
<td>2013</td>
<td>12,503</td>
</tr>
<tr>
<td>Austria</td>
<td>German</td>
<td>Federal Organic Budget Act 2013</td>
<td>2013</td>
<td>40,151</td>
</tr>
<tr>
<td>Canada</td>
<td>Common law</td>
<td>Financial Administration Act 1985</td>
<td>2013</td>
<td>37,994</td>
</tr>
<tr>
<td>Estonia</td>
<td>German</td>
<td>Law on State Budget 1999</td>
<td>NA</td>
<td>4,659</td>
</tr>
<tr>
<td>Finland</td>
<td>Nordic</td>
<td>1988 Budget Act, 1992 Budget Decree</td>
<td>2007</td>
<td>21,494</td>
</tr>
<tr>
<td>France</td>
<td>Napoleonic</td>
<td>The Organic Law on Budget Laws 2001</td>
<td>NA</td>
<td>8,944</td>
</tr>
<tr>
<td>Iceland</td>
<td>Nordic</td>
<td>The Government Financial Reporting Act 1997</td>
<td>NA</td>
<td>4,690</td>
</tr>
<tr>
<td>Italy</td>
<td>Napoleonic</td>
<td>Law of Accounting and Public Finance 2010</td>
<td>2013</td>
<td>23,306</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>Napoleonic, German</td>
<td>Government Accounts Act 2001</td>
<td>2005</td>
<td>15,332</td>
</tr>
<tr>
<td>New Zealand</td>
<td>Common law</td>
<td>Public Finance Act 1989</td>
<td>2013</td>
<td>47,378</td>
</tr>
<tr>
<td>Spain</td>
<td>Napoleonic</td>
<td>General Budgetary Law 2001</td>
<td>NA</td>
<td>43,711</td>
</tr>
<tr>
<td>Sweden</td>
<td>Nordic</td>
<td>Budget Act 2011</td>
<td>2011</td>
<td>4,290</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>Common law</td>
<td>Exchequer and Audit Departments Act 1866, Charter of Budget Responsibility 2011</td>
<td>2013</td>
<td>3,847</td>
</tr>
</tbody>
</table>

Notes: Column 4 gives the year of the latest legislative upgrade available in the research material of this article, and column 5 indicates the number of words in the legislation examined before the text pruning preceding the analysis proper in the study this article re-examines.
values of each document and, indirectly, for the country which each document represents (Table 2). According to the results, certain documents and countries were similar enough to receive overlapping credibility intervals with the intervals for other documents and countries. Australia and New Zealand did not differ statistically from each other, nor did Austria, Finland and Iceland, or Estonia and Spain. The other countries, Canada, Italy, France, the Netherlands, Sweden, and the United Kingdom, were empirically more unique.

**Examining topic modeling**

The topic modeling program library used in the study that this article re-examines would have set no technical limits to estimating high numbers of topics, but Hypotheses 4 and 5 and their background argumentation indicated a maximum of four topics that could be given a substantive interpretation. However, a German topic and a Nordic topic did not evolve separately in the four-topic estimation, although this is understandable given the closeness of the two civil law varieties. Only a three-topic solution is presented (Table 3).

The topic modeling results (Table 3) allowed Hypothesis 4 to be sustained, revealing three topics: a joint topic of ‘German and Nordic law’, a topic of ‘common law’, and a topic of ‘Napoleonic law’. As is commonplace in topic modeling, many words appear in two or all three topics. However, some of these words only illustrate that the corpus comprised legal texts (for instance, ‘law’, ‘must’, ‘shall’, and ‘may’), or indicated references and cross-references common in legal texts (for instance, ‘section’, ‘article’, or ‘paragraph’). Looking at words of substance that characterize certain topics rather than others, Hypothesis 5 on differences between the legal vocabularies of different legal systems could be sustained (Table 3). Words dealing with the contents of budgeting can be observed in the first topic of ‘German and Nordic law’, such as ‘budget’, ‘finance’, ‘financial’, ‘expenditure’ (both in the singular and the plural), ‘cash’, ‘assets’, and ‘provisions’. In the ‘common law’, topic words were found to
characterize the institutionalization of budgetary governance with special reference to the key actors: ‘minister’, ‘crown’, ‘corporation’, ‘treasury’, ‘board’, ‘parliament’, ‘council’, and ‘governor’. Finally, words characterizing accountability received emphasis on the third, ‘Napoleonic’ topic: ‘accounts’, ‘accounting’, ‘report’, and ‘auditing’. Unfortunately, the relationships between different regimes of budgetary governance and different textual frames of budgetary legislation were too under-researched to receive an elaboration proper in the study that this article re-examines, and have to await future study.

In the study that this article re-examines, the distribution of the three topics in the documents examined and, respectively, in the countries that these documents represent gave further support to Hypothesis 5 (Table 4). In the Austrian document, only the German-Nordic topic was present. This was almost the case in the Swedish and Finnish documents, whereas the weight of this same topic in the Icelandic document was about two-thirds. In Estonia, despite its predominantly German legal heritage (Glenn, 2010), more than two-thirds of the vocabulary examined represented the ‘Napoleonic law’ topic. Only future studies can examine the reasons for this aberration.

The countries in which only the second, ‘common law’ topic was present comprised Canada, New Zealand, and Australia, and this topic also predominated in the United Kingdom with a weight over half of the total. According to the estimation results, in Spain and Italy only the third, ‘Napoleonic law’ topic was present. France, in its turn, represented itself rather as a hybrid between the Napoleonic topic and the German-Nordic topic, possibly reflecting the historical influence of German legal traditions in northern France and the characteristics of the Code

### Table 2. Test results of unsupervised latent trait scaling in the study that this article re-examines.

<table>
<thead>
<tr>
<th>Countries</th>
<th>Indicator of position taken in documents representing the country, ( \omega )</th>
<th>95% Lower credibility interval of ( \omega )</th>
<th>95% Upper credibility interval of ( \omega )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>1.4401</td>
<td>1.4002</td>
<td>1.4600</td>
</tr>
<tr>
<td>Austria</td>
<td>-0.4418</td>
<td>-0.4493</td>
<td>-0.4043</td>
</tr>
<tr>
<td>Canada</td>
<td>1.8086</td>
<td>1.7689</td>
<td>1.8227</td>
</tr>
<tr>
<td>Estonia</td>
<td>-0.9689</td>
<td>-1.0135</td>
<td>-0.9469</td>
</tr>
<tr>
<td>Finland</td>
<td>-0.4544</td>
<td>-0.4684</td>
<td>-0.4097</td>
</tr>
<tr>
<td>France</td>
<td>-0.7192</td>
<td>-0.7476</td>
<td>-0.6848</td>
</tr>
<tr>
<td>Iceland</td>
<td>-0.5077</td>
<td>-0.5420</td>
<td>-0.4436</td>
</tr>
<tr>
<td>Italy</td>
<td>-1.0599</td>
<td>-1.1131</td>
<td>-1.0666</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>0.0644</td>
<td>0.0412</td>
<td>0.1218</td>
</tr>
<tr>
<td>New Zealand</td>
<td>1.4748</td>
<td>1.4423</td>
<td>1.4895</td>
</tr>
<tr>
<td>Sweden</td>
<td>-0.1680</td>
<td>-0.0422</td>
<td>-0.1220</td>
</tr>
<tr>
<td>Spain</td>
<td>-0.9953</td>
<td>-1.0317</td>
<td>-0.9922</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>0.4814</td>
<td>0.4300</td>
<td>0.5579</td>
</tr>
</tbody>
</table>

### Table 3. Characteristics of a three-topic model by words in the study that this article re-examines.

<table>
<thead>
<tr>
<th>Words</th>
<th>Topic 1, German-Nordic law</th>
<th>Topic 2, Common law</th>
<th>Topic 3, Napoleonic law</th>
</tr>
</thead>
<tbody>
<tr>
<td>'shall'</td>
<td>'minister'</td>
<td>'article'</td>
<td></td>
</tr>
<tr>
<td>'budget'</td>
<td>'may'</td>
<td>'state'</td>
<td></td>
</tr>
<tr>
<td>'government'</td>
<td>'crown'</td>
<td>'shall'</td>
<td></td>
</tr>
<tr>
<td>'federal'</td>
<td>'act'</td>
<td>'budget'</td>
<td></td>
</tr>
<tr>
<td>'section'</td>
<td>'corporation'</td>
<td>'public'</td>
<td></td>
</tr>
<tr>
<td>'finance'</td>
<td>'section'</td>
<td>'law'</td>
<td></td>
</tr>
<tr>
<td>'management'</td>
<td>'financial'</td>
<td>'general'</td>
<td></td>
</tr>
<tr>
<td>'act'</td>
<td>'public'</td>
<td>'finance'</td>
<td></td>
</tr>
<tr>
<td>'minister'</td>
<td>'must'</td>
<td>'year'</td>
<td></td>
</tr>
<tr>
<td>'year'</td>
<td>'money'</td>
<td>'accounts'</td>
<td></td>
</tr>
<tr>
<td>'accounting'</td>
<td>'person'</td>
<td>'expenditure'</td>
<td></td>
</tr>
<tr>
<td>'may'</td>
<td>'treasury'</td>
<td>'may'</td>
<td></td>
</tr>
<tr>
<td>'financial'</td>
<td>'report'</td>
<td>'financial'</td>
<td></td>
</tr>
<tr>
<td>'statement'</td>
<td>'shall'</td>
<td>'paragraph'</td>
<td></td>
</tr>
<tr>
<td>'accounts'</td>
<td>'department'</td>
<td>'entities'</td>
<td></td>
</tr>
<tr>
<td>'ministry'</td>
<td>'appropriation'</td>
<td>'referred'</td>
<td></td>
</tr>
<tr>
<td>'expenditures'</td>
<td>'year'</td>
<td>'accounting'</td>
<td></td>
</tr>
<tr>
<td>'expenditure'</td>
<td>'board'</td>
<td>'ministry'</td>
<td></td>
</tr>
<tr>
<td>'fiscal'</td>
<td>'parliament'</td>
<td>'following'</td>
<td></td>
</tr>
<tr>
<td>'state'</td>
<td>'subsection'</td>
<td>'treasury'</td>
<td></td>
</tr>
<tr>
<td>'audit'</td>
<td>'council'</td>
<td>'provisions'</td>
<td></td>
</tr>
<tr>
<td>'central'</td>
<td>'fiscal'</td>
<td>'revenue'</td>
<td></td>
</tr>
<tr>
<td>'referred'</td>
<td>'information'</td>
<td>'said'</td>
<td></td>
</tr>
<tr>
<td>'cash'</td>
<td>'respect'</td>
<td>'provided'</td>
<td></td>
</tr>
<tr>
<td>'information'</td>
<td>'regulations'</td>
<td>'credits'</td>
<td></td>
</tr>
<tr>
<td>'account'</td>
<td>'amount'</td>
<td>'auditing'</td>
<td></td>
</tr>
<tr>
<td>'assets'</td>
<td>'governor'</td>
<td>'account'</td>
<td></td>
</tr>
<tr>
<td>'line'</td>
<td>'means'</td>
<td>'report'</td>
<td></td>
</tr>
<tr>
<td>'report'</td>
<td>'made'</td>
<td>'sector'</td>
<td></td>
</tr>
<tr>
<td>'provisions'</td>
<td>'general'</td>
<td>'social'</td>
<td></td>
</tr>
</tbody>
</table>

Note: The table indicates words in a descending order of probability that these words belong to a given topic. Only 30 words receiving the highest posterior probabilities in each topic are included in this table. Referring to the article text related to this table, the italicized words comprise examples of content words of budgetary governance in the first topic, words referring to institutionalization in the second topic, and accountability words in the third topic.
Napoleon as a hybrid between the law of southern France and northern France (Glenn, 2010).

In documents representing six countries (Australia, Austria, Canada, Italy, New Zealand, and Spain) only one topic evolved, which was almost the case in Finland and Sweden. In Estonia, France, and Iceland a dominant topic and a minor topic could be found, and in the United Kingdom and the Netherlands the study that this article re-examines revealed a dominant topic and two minor topics.

Baseline comparisons concerning the Big Data methods that this article examines

Baseline comparisons have been carried out before between traditional methods and the unsupervised scaling method that the study that this article re-examines utilized (see, for instance, Grün and Hornik, 2013; Lowe and Benoit, 2013). The performatve struggles between researchers using traditional methods (see, for instance, Biernacki, 2014; Budge, 2013) on the one hand, and researchers using Big Data methods (DiMaggio et al., 2013; Laver et al., 2002; Proksch and Slapin, 2008) on the other, also offer lessons for baseline comparisons.

The baseline comparisons in this article receive support from long-time researcher familiarity with both quantitative research and semiotic and rhetorical textual analysis. During the preparation of an international refereed article that went into press in 2014, extending a three-country study (Hyndman et al., 2013) to cover a fourth country, the laborious material collection, the tedious and error-prone manual coding and dissatisfaction with the analysis of word frequencies made methods of Big Data analysis attractive (for the methods actually chosen, see Appendix 1). Given this modest baseline, the study that this article examines, once completed, represented a substantial improvement.

While preparing the study indicated above, the path led to the best-evolved Big Data methods to examine ideological polarization, comprised of methods of latent trait scaling. An additional interest to learn what unsupervised rather than supervised scaling would deliver made the Wordfish program (Proksch and Slapin, 2009) the choice. The study that this article re-examines was driven towards topic modeling by two additional forces. The former of these comprised researcher familiarity with the classical rhetorical examination of (Aristotle, 2006) and the resulting curiosity to learn about the performance of a Big Data method that steps forward as an inheritor of the classical tradition. The latter driving force was a technical interest in using two methods of Big Data analysis as baselines in respect to each other in examining the same research material.

Conclusions and discussion

Augmentation, surrogation, and re-orientation of research

From among the two types of Big Data methods that this article has examined, latent trait scaling has evolved with the explicit purpose of augmenting (Edwards et al., 2013) earlier methods. The authors of both supervised latent trait scaling (Laver et al., 2002) and unsupervised latent scaling (Proksch and Slapin, 2008) have been critical towards the conventional methods applied in a minor international research project started in the mid-1980s to examine political preferences of parties since 1945 finally in more than 50 countries (MARPOR 2015). Although sensitizing MARPOR representatives towards vulnerability in their methods, the alterations have ended in stalemate (Budge, 2013; Volkens et al., 2013), and comparative examinations of the merits of the traditional and Big Data methods (see, for instance, Lowe and Benoit, 2013) have failed to resolve the fundamental disputes.

Big Data methods of topic modeling originate from within hybrids of computer science and statistics rather than from within social research. Topic modeling explicitly augments traditional latent variable modeling.
initiated with Karl R Pearson’s principal component analysis at the beginning of the 20th century and continued with such methods as LL Thurstone’s factor analysis in the 1930s (Blei, 2014; Grimmer and Stewart, 2013). However, researchers with classical humanist inclinations may want to critically consider how much topic modeling actually augments the 2500-year rhetorical tradition of examining topics or other procedures of classical humanist textual interpretation (see, for instance, Biernacki, 2014).

The results of this article suggest agreement with Edwards et al. (2013) that Big Data methods may not easily provide surrogates for conventional research designs of social research despite the fact that such advances are not ruled out (see, for instance, Hale et al., 2014; Nickerson and Rogers, 2014). In the study that this article has re-examined, the research design was cross-sectional and comparative by modest default rather than by explicit design. A possible future step forward would lead to longitudinal Big Data research of texts of budgetary legislation in a number of countries (for analogous examples of longitudinal Big Data analysis, see, for instance, DiMaggio et al., 2013; Grimmer, 2010; Proksch and Slapin, 2008).

Edwards et al. (2013) indicate that the re-orientation of research by means of Big Data methods may start from where their mere augmentation of conventional methods stops. The empirical results on government budgetary legislation in the study that this article has re-examined suggest that latent trait scaling has potential in empirical research over and above the study of the ideological dimensions of politics (Laver et al., 2002; Lowe and Benoit, 2013; Proksch and Slapin, 2008). More specifically, in the study indicated, a specific method of unsupervised latent trait scaling was used to discern latent traits of legal traditions in government budgetary legislation in 13 countries by means of examining the texts of this legislation. Big Data research has been evolving within legal research proper (see, for instance, Hildebrandt, 2012; Surden, 2014), but the study that this article has examined indicates re-orientation within political science research on government budgeting and legal policy-making rather than legal studies.

The foremost contribution of the topic modeling application in the study that this article has re-examined was confined to social research of the political science variety. This is possibly a contribution in itself, as topic modeling has been applied within political science research comparatively rarely thus far (but, see, for instance, Clark and Lauderdale, 2010), whereas it has been substantially more common within sociological research (DiMaggio et al., 2013; Fliigstein et al., 2014; Levy and Franklin, 2013; more generally, see also Bail, 2014).

The social and political life of Big Data methods influencing the institutionalization of research

As indicated at the beginning, as its foremost objective this article examines the ‘social life’ and the ‘political life’ of Big Data methods as influences on the institutionalization of research. The social structures, the organization and the identity of Big Data research, and its researchers within the social and political sciences were still weak rather than entrenched in most countries and most research fields in the mid-2010s. At the same time, mainstreaming of Big Data methods was still advancing within statistics proper rather than having been fully accomplished (see, for instance, EMC, 2015). Genuine innovations in social and political research concerning Big Data methods have been relatively few thus far although not nonexistent (see, however, Grimmer and Stewart, 2013; Laver et al., 2002; Proksch and Slapin, 2008). However, the Big Data activism of distinguished social and political scientists (see, for instance, DiMaggio et al., 2013; Fliigstein et al., 2014; Laver et al., 2002) may be changing the situation.

According to one scenario, social and political research utilizing Big Data methods will institutionalize itself, that is, it will become habitual and taken-for-granted in social and political research. However, should this institutionalization take place, it would also generate its ‘rationalized myths’ (Meyer and Rowan, 1977) with the exaggeration of the merits and contributions of these methods and formal rather than substantive commitment to them. We would also witness not only the enhancement of the rationality of the core analytic processes of research by means of the Big Data methods, but also the strengthening of the external legitimation of institutions of research by the same means.

For a social researcher in political science, particularly interesting characteristics of the political life of the Big Data methods include the performative roles that these methods and their applications may play in disrupting the achieved institutionalization of research and, possibly, in catalyzing institutional transformations. This article has indicated two specific performative struggles. The first one of these struggles has revolved around a long-term international project examining ideological political polarization (MARPOR, 2015), and has been waged between those who defend traditional methods (Budge, 2013; Volkens et al., 2013) on the one hand, and those who apply Big Data methods (Laver et al., 2002; Proksch and Slapin, 2008) on the other. As indicated above, proposals of appeasement (Lowe and Benoit, 2013) have reaped little success so far. The other performative struggle, which to this date has been milder, has been...
waged between those who apply Big Data methods including topic modeling (DiMaggio et al., 2013; Fligstein et al., 2014) on the one hand, and those who prioritize methods derived from classical humanistic traditions (Biernacki, 2014) on the other.

In the mid-2010s, time has generally been working to the advantage of those social and political researchers who apply or promote Big Data methods, as observations accumulate that in many countries public sector and private sector funding authorities have placed heavy emphasis upon Big Data research projects. Exhibiting a ‘vivid awareness of the relationship between personal experience and the wider society’ better known as ‘sociological imagination’ (Mills, 1959: 3), these developments have lately become keenly acute not only for numerous global colleagues but also for the research team from among whose products this article is one. Observations of opportunities intermingle with experiences of smaller or larger moments of success in attracting external research funding and having refereed article manuscripts accepted in international journals, winning variable success in mainstreaming Big Data research and related teaching in academia, and disappointments insofar as funding applications or article manuscripts receive rejections. What else do the current characteristics of the ‘political life’ of Big Data methods and its researchers re-present than politics in the very sense of Pocock (1975: 156) as the ‘art of dealing with the contingent event... with pure, uncontrolled, and unlegitimated contingency’?

Declaration of conflicting interests
The author declares that there is no conflict of interest.

Funding
This study was supported by Koneen Säätiö/Kone Foundation, Helsinki, Finland, project grant.

References
Fligstein N, Brundage JS and Schulz M (2014) Why the Federal Reserve failed to see the financial crisis of 2008: The role of ‘macroeconomics’ as sense-making and
cultural frame. Available at: irle.berkeley.edu (accessed 31 January 2015).


Unsupervised latent trait scaling. The unsupervised latent trait scaling applied a version of the Wordfish program included in the R library system (Feinerer, 2013), such as eliminating numbers, punctuation marks, stopwords, and all words present in only one or two of the 13 texts of the corpus, leaving 2992 of the original 5351 different words.

In equation (1), \( y_{ij} \) is the count of word \( j \) in document \( i \), and in equation (2), \( \alpha \) is the set of fixed effects related to a document, \( \psi \) is the set of fixed effects related to words, \( \beta \) is the set of estimates of word-specific weights to capture the importance of each individual word in discriminating between positions that the texts represent, and \( \omega \) is the set of estimates for the position that each document represents (for explanations of these notions see the text proper of this article).

In the Wordfish method of unsupervised latent trait scaling, the estimation and testing of the fixed effects, \( \alpha \) and \( \psi_j \), technically control for the common possibility that different documents may include some words rather than others, and that some documents have more words than others. Moreover, the estimation results of the fixed effects are instrumental in visually displaying the results of Wordfish unsupervised latent trait scaling (see the text proper of this article and, for an example, see Figure 1). Unsupervised latent trait scaling first and foremost estimates the \( \beta_j \), indicating the words that differentiate between the positions taken in the documents examined, and the \( \omega \), indicating the positions of each document derived from the positions that its constituent words express.

The right-hand side of equation (2) is estimated by means of an expectation maximization algorithm to compute maximum likelihood estimates for the latent variables (Proksch and Slapin, 2009). First, starting values for \( \alpha \), \( \psi \), \( \beta \), and \( \omega \) are obtained. Second, the document parameters \( \omega \) and \( \alpha \) are estimated conditional on the expectation for the word parameters \( \psi \) and \( \beta \) by maximizing a log-likelihood function for each of the documents examined. Third, the word parameters \( \psi \) and \( \beta \) are estimated conditional on the expectation of the document parameters \( \omega \) and \( \alpha \), obtained in step 2. For each word, a log-likelihood function is maximized. Fourth, the log-likelihood of the model is calculated as the sum of the individual word log-likelihoods obtained in step 3. Fifth, steps 2–4 are repeated until the model converges. Last, credibility intervals are calculated for the scaling estimates of the \( \beta_j \) and \( \omega \) by means of a parametric bootstrap with a Wordfish default of 500 iterations. As is commonplace with Big Data methods, the Wordfish latent trait scaling method in some respects utilizes analogous elements as older statistical methods on the one hand, but adds certain new elements and procedures on the other let alone represents a novel combination of its constituent elements (more generally, see EMC, 2015).

Topic modeling. ‘Topics’, first evolved in classical rhetoric (Aristotle, 2006), comprise devices that language users invent or innovatively transfer from other uses to those of their own to influence the cognitions, emotions, and perceptions of their audience, including their opponents (Greer, 2011; Walter and Uhr, 2013). The research methods of topic modeling support the detection and interpretation of topics in texts. The study that this article re-examines applied program scripts of probabilistic topic modeling called ‘latent Dirichlet allocation’ (LDA; Blei et al., 2003).

Applications of LDA discern latent topics that support the persuasive capacity of texts by means of inclusions and emphases and exclusions and omissions in these texts. LDA builds on the understanding that the documents examined comprise mixtures of topics that regulate the probability that different words find their way into these documents. The topics and their probability distributions in documents are understood to
comprise hidden random variables, which are built into a hierarchical probabilistic model, and estimated by means of approximating the conditional distribution of these variables in the documents. In the topic modeling program library used in the study that this article re-examines (Grün and Hornik, 2013), this was done by means of Gibbs sampling and constructing a Markov chain Monte Carlo algorithm (Blei, 2012). These characteristics comprise another example on how Big Data methods combine elements of known earlier methods with newer elements (EMC, 2015).

The examination advanced as follows in the study that this article re-examines. First, the probability distribution of words, $\beta$, was determined for each topic, assuming a Dirichlet probability distribution, $\delta$:

$$\beta \sim \text{Dirichlet}(\delta) \quad (3)$$

Second, the topic distribution, $\theta$, was determined for each document $w$, also assuming a Dirichlet distribution, $\alpha$:

$$\theta \sim \text{Dirichlet}(\alpha) \quad (4)$$

Third, for $w_i$ or each of the $N$ words, each topic, $z_i$, was chosen on the assumption of the multinomial distribution of this topic:

$$z_i \sim \text{Multinomial}(\theta) \quad (5)$$

Fourth, $w_i$, each of the $N$ words, was chosen on the assumption of a multinomial prior distribution of these words, each of which is conditioned on the topic, $z_i$:

$$z_i : p(w_i|z_i, \beta) \quad (6)$$

After the four steps, the log-likelihood of the research material, calculated as the sum over the log-likelihoods of all documents included, was maximized with respect to the model parameters $\alpha$ and $\beta$. However, no mathematical criteria are available to decide how many topics, $k$, the researcher should distinguish, but he or she has to decide the number of topics based on the study objectives, the background theory, and the hypotheses proposed. In the study that this article re-examines, the model was fitted using the function LDA included in the program library utilized in the analysis. The fitting required assigning an initial value to the Dirichlet prior $\alpha$, defining the number of iterations to reach as stable estimates as possible, and introducing a random ‘seed word’ with which to start the estimation to ensure the reproducibility of the results (Grün and Hornik, 2013). Last, a topic model function called ‘posterior’ was used to obtain the distribution of words on each topic and the distribution of topics in each document.