Original Article

Computational Social Science Methodology, Anyone?

Joop J. Hox

Department of Methodology and Statistics, Utrecht University, The Netherlands

Abstract: This article reviews computational social science methods and their relation to conventional methodology and statistics. Computational social science has three important features. Firstly, it often involves big data; data sets so large that conventional database and analysis techniques cannot handle them with ease. Secondly, dealing with these big data sets has given rise to analysis techniques that are specially developed for big data. Given the size of the data, resampling and cross-validation approaches become feasible that allow both data-driven exploration and checks on overfitting the data. A third important feature is simulation, especially agent-based simulation. Here size also matters. Agent-based simulation is well known in social science, but modern computer equipment and software allows simulations of unprecedented scale. Many of these techniques, especially the resampling and cross-validation approaches, are potentially very useful for social scientists. Given the relatively small size of social science "big data" is useful to explore how well these techniques perform with smaller data sets. Social science methodology can contribute to this field by exploring if well-known methodological distinctions between external validity, internal validity, and construct validity can help clear up discussions on data quality (veracity) in computational social science.

Keywords: Big Data, computational social science, analytics, data science

It seems that almost any scientific field can now be preceded with the adjective "computational." This is certainly true for the social and behavioral sciences, where several departments and centers for computational social science exist, and several handbooks describing methods and applications exist (e.g., Attewell & Monaghan, 2015; Foster, Ghani, Jarmin, Kreuter, & Lane, 2016).

Nevertheless, computational social science is an emerging field. A search in Google Scholar for "computational social science" returns about 5,700 results. "Computational psychology" returns 3,180 results, and "computational economics" returns about 7,530 results. In comparison, "computational physics" returns about 475,000 results, and "computational biology" a whopping 1.7 million. Compared to some other scientific fields, computational social science is a small affair.

In this article, the term "computational social science" is used to refer also to related fields such as computational psychology and computational economy. Computational social science refers to the application of computational methods to explore and test scientific (social, psychological, economic) theories. It is often interpreted as equivalent to the use of big data in social science. However, although computational social science relies strongly on big data and the ability to analyze these, it also includes computer simulation and text analysis.

Computational social science is an interdisciplinary field, that includes mathematics, statistics, data science, and of course social science. Collecting and analyzing big data sets, including large bodies of texts, was developed mostly by computer scientists and mathematicians, often in a commercial context. Amazon's "Frequently Bought Together" recommendations are based on predicting search and purchase behavior in its enormous database of transactions. This origin in a nonstatistical environment shows up in a distinctive terminology, which is a barrier to understanding the techniques and tools for big data sets. This article introduces computational terminology where needed, and the appendix lists the most important terms with their social science methodological counterparts.

This article reviews three important elements of computational social science: big data, analytics, and simulation. The fourth section looks at the question: What Can We Learn? Are there tools that social science methodology should be using more often? The fifth section looks at the converse question: What can we contribute? Are there methodological issues that are well known to social science methodologists, but are unknown to the computer scientists involved in computational social science?

Big Data

Big Data in Social Science

How big is big data? According to Donoho (2015), John Tukey defined "Big Data" as anything that does not fit on one device. In the early days of computing, around

DOI: 10.1027/1614-2241/a000127

1955, the device would be a magnetic tape holding about 6 megabytes (MB). Nowadays, around 2016, a USB stick easily holds 256 gigabytes (GB), and a hard disk several terabytes (TB). To get a feeling for "How big is big?," Table 1 shows a list of prefixes indicating amount of information, together with examples.

Table 1. Decimal (SI) prefixes indicating amount of data

Prefix	10^n	Example (approximate)
Byte	10^0	One single character
Kilobyte	10^3	Very short story
Megabyte	10^6	Small novel
Gigabyte	10^9	TV movie
Terabyte	10^{12}	One day data of NASA Earth Observing System
Petabyte	10^{15}	Five years data of NASA Earth Observing System
Exabyte	10^{18}	5,000,000,000 CD-ROMs
Zettabyte	10^{21}	All Internet traffic in 2016

Note. Source: http://simplyted.blogspot.nl/2005/12/how-to-visualize-data.html

Most social science data are not in the realm of Big Data. For example, the cumulative SPSS datafile from the European Social Survey (http://www.europeansocialsurvey.org) is about 50 GB. Daas, Tennekes, de Jonge, Priem, and van Pelt (2012) presented a paper that gave some examples of Big Data in official statistics. Statistics Netherlands creates every month a linkage between tax, insurance, and social security records, producing 20 million records per month. Assuming that each record is 1 KB, 12 months of financial statistics is 240 GB. That will just fit on a big USB stick. Traffic loop tracing data, also collected by Statistics Netherlands, produces 80 million records per day. One year of data would be about 3 TB and just fit on a large hard disk.

The promise of big data for social science is that people's behavior increasingly leaves digital traces, which can be collected and analyzed to make inferences about these behaviors. Expensive data collections are replaced by inexpensive "found" data, and sampling can be replaced by analyzing N = *All*. Digital traces can be intentional, such as messages on Facebook or Twitter, or discussions on Internet discussion lists. Many traces are unintentional: mobile phones provide information about the location; phone calls in general are monitored by the providers, who store location, who calls whom, and length of call. All these traces of behavior are available to be collected and analyzed.

Collecting or "scraping" data from the Internet is not always easy. Big data have been characterized by the Three V's: Volume, Velocity, and Variety (Laney, 2001). Volume means that big data refers to very large volumes of data. Velocity means that this data may come in in large volumes in real time and must be analyzed very quickly. Variety means that much of this data may be weakly structured or consist of texts and multimedia files that must be coded before they can be analyzed. Various authors have added a fourth V: Veracity. Veracity refers to the uncertainty in the data (IBM, 2016): quality issues due to inconsistencies in the data, ambiguity in coding, dependencies, and measurement issues.

There are certainly quality issues in Big Data. Firstly, there is often no notion of sampling from a well-defined population. For example, a popular source for Internet-based big data is Twitter, because there is a streaming API (Application Programming Interface) that can be used to sample tweets for free. It is also possible to obtain *all* tweets generated, using a service called firehose. Morstatter, Pfeffer, Liu, and Carley (2013) compared data from the firehose to data sampled from the free streaming API and conclude that the API data are *not* a random sample from the firehose. Since there is no definition from which population the complete set of tweets derives, this may seem unimportant. But Morstatter et al. (2013) also point out that it is unknown which method is used to subsample the tweets, and that the coverage of the API seems to depend on the search terms. As a result, a trend found on Twitter may be real or simply the result of changes in the sampling algorithm. Much more problematic is the finding by Kollanyi, Howard, and Woolley (2016) that after the second presidential debate between Clinton and Trump about one third of the pro-Trump tweets and about one quarter of the pro-Clinton tweets were generated by bots. These tweets were dubbed as "Computational Propaganda." Another example of dubious data quality was found by Madden et al. (2013), who report that in a study about privacy management by teens on social media sites 26% of the respondents admitted to posting false information.

An example of computational social science and big data is the study by Onnela et al. (2007) on the importance of weak ties in social networks. Granovetter (1973) has argued that information in social networks is carried by weak ties, by weak relationships between members of the network. Onnela et al. (2007) used data from 4.6 million mobile phone users to investigate this hypothesis. Two users were connected with a unidirectional link if there had been at least one reciprocal call (A called B, and B called A). The strength of the tie was defined as the total duration of all calls between A and B. The resulting network had 4.6 million nodes (the mobile phone users) and 7.0 million links.

Figure 1 shows the network around a randomly selected individual. It shows a structure of small clusters with strong ties within clusters and weak ties between clusters, consistent with Granovetter's (1973) hypothesis. More formal analyses (Onnela et al., 2007) show that the stronger the tie between two individuals, the more their friends overlap.

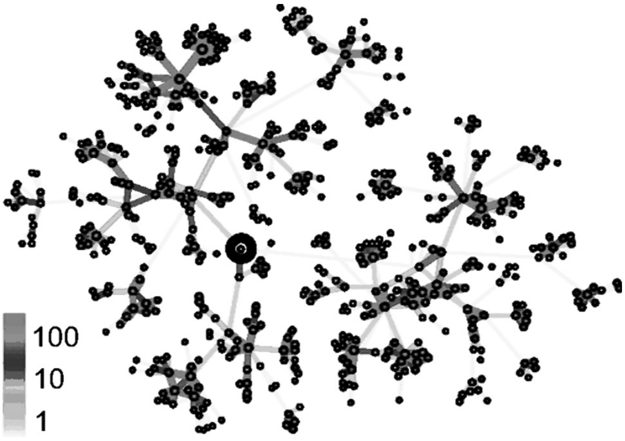

Figure 1. Network structure around a randomly chosen individual. Tie strength indicated by color. Reprinted with permission from Onnela et al. © (2007) National Academy of Sciences.

Finally, Onnela et al. (2007) find that removal of strong ties makes the network smaller, but if the weak ties are removed at some point the network structure falls apart. All these findings corroborate the strength of weak ties hypothesis.

Analytics

Big Data Analytics

Big data analytics is the analysis of data so big and complicated that it is difficult to analyze them with traditional database and analysis tools. The mobile phone network example contains 4.6 million nodes and 7.0 million ties, and does not fit the traditional rectangular data matrix. Analyzing this structure is computationally demanding. The technology of big data analysis was developed as a combination of statistics, machine learning, and computer science, with emphasis on developing software that works with very large data sets. The main aim is not modeling and inference, but rather exploring data and prediction. This goes back to Tukey (1962, 1977) who emphasized exploration of data, graphical techniques, and determining the real indeterminacy by using prediction, subsampling, and resampling of the data, instead of relying on formal statistical inference. For an overview of “50 years of data science” I refer to Donoho (2015).

Big data analytics needs algorithms that are efficient for large data sets, and software that can run on clusters of computers. The software must be robust; if one of the computers in the cluster malfunctions, it must not crash. A good example is Hadoop. Hadoop is a MapReduce procedure that maps features of data (in social science parlance: variable values) onto a reduced set of values (e.g., a frequency distribution). Hadoop is designed to run on a large cluster of machines. It splits large data sets into smaller blocks that are handled by different machines. Each block is replicated three times across different machines, to make sure that the data are still complete after two concurrent machine failures. Other program features ensure the efficiency and integrity of the algorithm in an environment of massive parallel processing. I refer to Vo and Silva (2016) for a nontechnical description of the inner workings of Hadoop and similar systems such as Apache Spark.

The origin of big data analytics is applied statistics in the real world, often developed by computer scientists, not statisticians. The result is a strong emphasis on prediction instead of modeling and theoretical understanding. *Deep Learning* is associated with neural networks, which mimic neural connections. It consists of layers, where each layer connects input nodes to output nodes, and weights, summarizing and transforming the input values into output values. The more layers there are, the better they perform (Ghani & Schierholz, 2016). Neural networks are especially effective in pattern recognition, such as recognizing patterns in images or text. IBM’s Watson is an example of deep learning (Presenti, 2015). Since deep learning is not part of the social science toolbox, it is not treated here, see Attewell and Monaghan (2015, Chapter 11) for an introduction. In contrast, *machine learning* has its origin in computer science and uses many statistical techniques well known to social scientists (Ghani & Schierholz, 2016).

In machine learning, two distinctions are important. First, there is supervised and unsupervised learning. In supervised learning, there is a target variable or a set of target variables (called *labels*), and the goal is to predict (if the target is continuous) or to classify (if the target is categorical nominal). The prediction uses *features* of the data, in social science called independent variables. Familiar techniques here are (logistic) regression models, often including nonlinear effects and interactions, and tree methods. In unsupervised learning, there are no target variables. The goal in unsupervised learning is to distinguish groups of objects or features that are similar. Familiar techniques are cluster and principal component analysis. Actually, the distinction between supervised and unsupervised learning is a dimension rather than a dichotomy; the degree of supervision can vary. For an introduction to semi-supervised learning, see Zhu and Goldberg (2009).

Supervised Learning

Supervised learning involves prediction or classification. One key difference between machine learning analytics and the usual social science statistics is that there is a strong emphasis on predicting new data. In social science statistics, the emphasis is on modeling the data at hand as well as possible, using a model that has an explicit and reasonably simple structure, for example a structural equation model. Machine learning models are not necessary regression models, they may also be formulated as a set of rules that transform the independent variables in a specific way to produce predicted outcomes. Since these transformation rules are explicit, they can still be given a substantive interpretation, unlike the situation in neural networks where the prediction system is a black box because the rules are unknown.

The techniques used to make predictions were developed in the context of large data sets with many variables. Most models explore nonlinear effects and interactions, either explicitly or implicitly, and big data analytics tends to use predictor variable selection as well. Even with large data sets, there is a serious risk of overfitting. Overfitting occurs when a model is overly complex and predicts mostly variation that is specific to the data at hand. Overfitted models do not replicate well.

To assess the fit of the model, big data analytics uses a combination of resampling and cross-validation methods. In social science, cross-validation generally means dividing the data at random in an exploration and a validation sample. Model exploration is carried out on the exploration sample. When a well-fitting model is found, it is next estimated in the validation sample. If a model fits well in the validation sample, it is retained. Variations are exploring the model twice in both subsamples and validating it twice on the other subsample. Camstra and Boomsma (1992) review such cross-validation techniques.

Big data analytics improve on these methods in several ways. A common variant is the *k*-fold. The data are partitioned into *k* parts called *folds*, with *k* typically set to ten. Next, the procedure iterates *k* times. Each time, fold *k* is held out, and the model is trained (estimated) on the combined $k - 1$ remaining folds. Subsequently, the generalizability of the model is tested by assessing the accuracy of the predictions made for the holdout fold *k*. Finally the *k* results are aggregated to a single model fit value. Typically, several different models are assessed this way, and the model that generalizes best is retained. If variable selection is involved, this entire process is also iterated *v* times, each time starting with a different random subset of the available features (predictors).

An example of supervised learning and feature selection is Google's attempt to predict influenza prevalence in the US. Influenza-like illness (ILI) is monitored in the US by the Centers for Disease Control (CDC) by monitoring physician visits and virologic data. The data are published weekly with a 1–2 week time lag. The system proposed by Google (Ginsberg et al., 2009) monitors influenza-related search terms, and uses these to make predictions about Influenza prevalence, almost without time lag. Google collected 50 million common search queries in the US, aggregated to weekly counts. Next, these were used to predict the weekly CDC influenza figures. Logistic regression was used to predict ILI-related physician visits. Each of the 50 million search terms was tested individually. To find the best set of predictors, different sets were combined of the *N* top query terms. *N* was chosen using a holdout sample procedure, the best *N* turned out to be $N = 45$. In an appendix Ginsberg et al. report fitting 450 million models to test each of the candidate queries, dividing the analysis task among a cluster of hundreds of machines.

For validation, the prediction was correlated with the CDC figures across the nine regions that CDC uses, with correlations ranging from 0.80 to 0.96. However, when the prediction system was used on new data, it predicted less well, until in February 2013 Google Flu Trends (GFT) predicted more than double the actual CDC figures. A careful analysis (Lazer, Kennedy, King, & Vespignani, 2014) showed that GFT had had a tendency to overpredict in other years as well. They conclude that the prediction errors are likely the result of overfitting: 50 million queries were tested on 1,152 available data points (weekly ILI figures). All the *k*-folding and other methods to combat overfitting were simply not enough.

Unsupervised Learning

Unsupervised learning, meaning there are no *labels* (dependent variables), is often based on familiar techniques such as cluster or principal component analysis. Unsupervised learning also refers to clustering qualitative data, such as bodies of text, email records, or tweets. Many clustering algorithms used are familiar to social scientists; *k*-means clustering, hierarchical clustering, and latent class analysis.

An example of unsupervised learning is a study by Roberts, Stewart, and Tingley (2016). They examined political event data, records of interactions between political actors. Their data contains large bodies of text. In addition to "found" data, such as online news sites, there are large collections of text data made available by scientific consortiums. Roberts et al. examine a collection of 13,246 blogs written during the 2008 US elections. They use a latent class statistical topic model (STM), which assigns each document to several topics. There are in total 100 topics. To examine the stability of the results, several runs of the algorithm are carried out with different random starts, and the similarity between different runs is evaluated by comparing the distribution of words within the same topic. Examples of topics are Supreme Court, global warming, and Iran/North Korea nuclear arms. Typical for big data analytics is comparing different models, different preset numbers of topics, different convergence criteria, and even an entirely different model like *k*-means clustering.

Simulation

In social science methodology and statistics, simulation generally refers to statistical simulation. In statistical simulation, data are generated according to a specific model, with some violation of the assumptions of the estimation method. A large number of data sets are generated, and the parameter estimates are summarized and compared to the known population parameters, to find out to what extent the assumption violations lead to misleading results. In computational social science, simulation generally refers to model-based simulation or agent-based simulation. In model-based simulation, a computer model is specified to represent a complex system using a large set of equations and rules. Such models are understood as simpler than the system they simulate, yet complex enough to be useful for prediction and understanding. Agent-based simulation places virtual agents in a computer generated environment and allows these agents to interact according to certain rules. Agent-based simulation is often used to study how individual rules of behavior and interaction between individual agents give rise to macro-level regularities. Model-based simulation is problem-specific and is treated here only briefly. Agent-based simulation is more general and is treated here in more detail.

Model-Based Simulation

Model-based simulation builds a computer model of a complex system and then uses this model to study that system. O'Reilly and Munakata (2000) give an example involving deep dyslexia. In deep dyslexia, patients make errors based on semantics, for example when the stimulus word is *rose* they may respond *tulip*. They also make the more common misreading errors. This suggests two separate damages to the brain. O'Reilly and Munakata (2000, Chapter 10), using a neural network model, show that it is possible to explain these two different types of errors assuming only one locus of damage, which is a more parsimonious model.

In model-based simulation, a model must be specified that encompasses both substantive knowledge about the system studied, and the equations and rules that represent the theoretical perspectives. Thus it requires understanding the subject matter and building a computer model for each research problem. One cannot imagine user-friendly software that can be used to model both the human brain and a nuclear power plant. The study by O'Reilly and Munakata (2000, Chapter 10) is a case in point. They explore language in general, investigating the consequences of assuming that words are stored in the brain in a distributed lexicon, rather than one single location. This has consequences for writing (orthography), speaking (phonology), and understanding (semantics), including different forms of dyslexia. They use several different (sometimes competing) models to assess how different inputs or differences in model parameters influence the model's behavior. Since model-based simulation is heavily dependent on the subject matter, it is not treated further here.

Agent-Based Simulation

Agent-based simulation models social systems using multiagent computer simulations. In these simulations, individuals or groups are represented by agents, who interact according to certain rules, in a computer generated environment. Agent-based simulation is often used to study how individual behavior at the micro-level can produce macrolevel structures and regularities. A simple goal of agent-based simulation is prediction, for example when the behavior of a large number of cars is simulated to predict when and where traffic jams occur. A goal more central to social science is using agent-based simulation to understand social systems.

A very early example of agent-based simulation is Loehlin's ALDOUS program. Aldous reacted to stimuli that differed in their values on the attributes attraction, anger, and fear. Loehlin (1965) created a program in which two copies of Aldous were interacting in a single computer environment. The values of the two Aldouses on the three attributes were experimentally manipulated. According to Loehlin (1965), the interaction sequences were surprisingly varied and had some similarity to human behavior. This early example has all the elements of an agent-based simulation. There are multiple agents, agents have different attributes, and react to each other in a series of encounters. Even this very simple simulation created results that were not entirely predictable by a rule-based system.

A more complex example of agent-based simulation is the work by Axelrod. Axelrod (1984) studies how cooperation can evolve if there is competition for resources and no central authority enforces cooperation. The central tool in Axelrod's research is a prisoner's dilemma game. There are two players, and each has a choice between cooperation and defection. If both players cooperate, they both gain. If both players defect, they both lose. If one player cooperates and the other defects, the cooperating player loses, and the defecting player gains. In this game, defecting is always better than cooperation. The dilemma is that if both players defect, they both lose, and do worse than if both had cooperated. Axelrod assumes that if both players interact only once, they will both defect. However, in a continuing interaction, an iterated prisoner's dilemma, there is an opportunity for cooperation to evolve, because in the long run both players will gain by this strategy. To investigate this, Axelrod set up a computer tournament to assess effective strategies. Game theorists were invited to send in strategies for the iterated prisoner's dilemma. Each entry faced all other entries for 200 iterated encounters, plus a copy of itself and a randomly responding entry. The entire tournament was repeated five times to get more stable estimates of each entry's success rate. The tournaments involved 14 players, 120,000 moves, and 240,000 separate choices (Axelrod, 1984, p. 31).

In the end, the clear winner was an entry called Tit-For-Tat. Tit-For-Tat cooperates on the first move and subsequently reciprocates whatever the other player did on the last move. Against a cooperating opponent, Tit-For-Tat will always cooperate, and it will defect against an opponent who defects. After the first tournament, the results were published and a second tournament was held. Tit-For-Tat won again. When analyzed in more detail, it became clear that "nice" strategies performed well by eliciting cooperation from other "nice" strategies.

Axelrod's agent-based simulations set off a whole string of studies looking into the conditions that encourage the evolution of cooperation, some based on further simulations, others investigating how realistic these simulations are. An example of the latter is Phelps' (2012) study that applies the reciprocity principles behind "nice" strategies in a simulation modeling the emergence of social networks in the real world. An interesting result from the 3.6×10^5 simulations is that at the individual (agent) level there is change, while at the network level there is stability, which mimics empirical observations of large-scale real networks.

What Can We Learn?

There are lessons to be learned for social scientists and methodologists. Social scientists are more conservative than investigators who work in market and applied research. Important research techniques were first developed in applied research and later incorporated in academic research. Examples are telephone survey methodology, pioneered in the 1970s in market and opinion research, and later used in academic research, computer-assisted psychological testing, and the use of web panels for data collection. The same pattern is visible in big data analytics, where large commercial organizations and market researchers have pioneered collecting data from the web and analyzing large and weakly structured data sets. Official statistics and academic researchers are just starting to get interested in these methods.

Collecting Big Data

Collecting data from the web or from other automated sources is potentially very useful. The promise of big data collection is that expensive data collections are replaced by less costly "found" data, and that sampling is replaced by analyzing all existing data: $N = All$, avoiding sampling error.

"Found" data have a methodological advantage; they are similar to the traces of behavior described by Webb, Campbell, Schwarz, and Sechrest (1966) as unobtrusive measures. If subjects are aware of being research subjects, this influences their behavior, including answer behavior in surveys. Such research is reactive; subjects react to being researched. Unobtrusive measures collect data already generated and cannot influence that behavior. Webb et al. mention as examples of unobtrusive measures physical traces, archives, organizational and personal records, and observation. The difference with data collection on the web or using data generated for other purposes, such as cash register scan data, is a matter of scale, not of kind. So, actually, the kind of data used in big data is well known to social scientists.

Size does matter. In 2014, Twitter handled 400 million tweets per day (Grishenko, 2014). Collecting even 1% of that stream in real time is nontrivial, and so is the ensuing analysis. But this is an only problem of hardware and software. Found data can answer interesting social science questions, too. Gonçalves-Sá, Varela, Wood, Bollen, and Rocha (2015) investigated whether human sexual cycles are driven by culture or environment, such as the yearly climate cycle. Using surveys is problematic here, because this is a sensitive topic, and the answers can be biased both by underreporting and by boasting. Gonçalves-Sa et al. collected Google sex-related searches over a 10-year period. The data were coded into searches originating either in the Northern or the Southern Hemisphere, and whether the region of origin can be described as Christian or Islamic. If the search pattern is seasonal, the pattern in the Southern Hemisphere should be the opposite of the pattern in the Northern Hemisphere. If the pattern is cultural, both hemispheres should exhibit the same pattern. Moreover, if the pattern is cultural, it should be related to important dates, such as Christmas in Christian regions, and Eid-al-Fitr ("sugar feast") in Islamic countries. The results clearly indicate that human sexual cycles are primarily driven by culture (Gonçalves-Sá et al., 2015).

Big Data Analytics

With big data, size matters as well. Analyzing really big data sets is challenging. But really big data sets have methodological advantages as well. In social science, the statistical approach is largely model driven. We have a theory, translate that into a model, for instance a structural equation model, and examine if the model describes the data well enough to be useful. In big data analytics, the attitude is to let the data speak, the analysis is largely data driven.

Most of the models used are well known to social science methodologists. The large size of the data makes it possible to use the models as tools. The data may be explored by several models, including different kinds of models. For example, in unsupervised learning one could use both *k*-means clustering and a tree-based clustering method. Data exploration includes model tuning, such as choosing different values for *k* in the clustering model, or different criteria for splitting branches in the tree-based clustering.

All this exploration carries the risk of overfitting by capitalizing on chance. The large size of the data set makes it possible to rely on cross-validation techniques to decide where to stop. The *k*-fold described earlier is a nice example of this. Categorizing subjects on the basis of their characteristics (features) is also dependent on the actual choice of variables. Here, the size of the data set helps again. We can run several models, using different sets of variables each time, and again rely on cross-validation techniques to investigate how well the results replicate across different sets of variables. This is exploited to the full in regression and classification trees. In a classification tree, a nominal target variable is predicted by variables (features) that are split at an optimal point to distinguish between the categories of the target variable. Figure 2 is an example based on the Titanic survival data (Dawson, 1995).

The tree shows that survival is predicted by being female, and for males by being younger than 9.5 years, and by traveling with more than two sibs. Note that the classification tree naturally includes nonlinear trends and interactions. Often several trees are generated. A technique called *bagging* relies on bootstrap sampling to generate several sets of predictions and averages these to produce final predictions. A different technique is known as *random forest*: many trees are generated with different sets of predictor variables. In the end, the predictions from the forest are averaged to get a final set of predictions. The random forest is an ensemble method: fitting many models to the same data with small variations and combining the results.

An empirical example of the strength of big data analytics is "Crowdsourcing Analytics" by Silberzahn et al. (2015). A data set was published online of all soccer players (N = 2,053) playing in the first divisions of four European countries in the 2012–2013 season. All interactions with all referees (N = 3,147) in their career were coded. The target variable was the number of red cards given to each player by each referee. There were 146,028 player-referee dyads. The primary research question was if player skin tone (ranging from 1 = *very light* to 5 = *very dark*) predicts receiving red cards. Seventeen potentially confounding predictor variables were included, describing the player, the referee, or the dyad. Twenty-nine teams of in total 61 analysts analyzed these to answer the primary research question. After several rounds of analysis and feedback, the 29 teams came up with highly varying results. Twenty teams (69%) reported significant results.

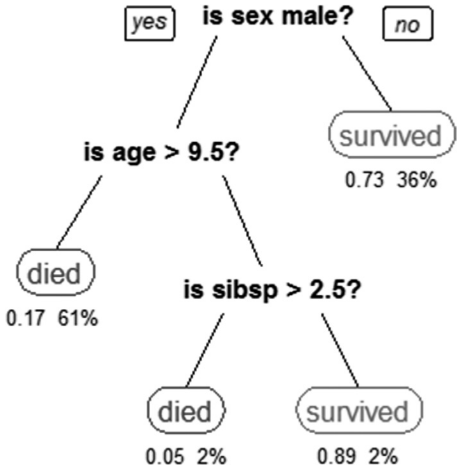

Figure 2. Decision tree on Titanic survival data. Source: https://en.wikipedia.org/wiki/Decision_tree_learning.

The authors (Silberzahn et al., 2015) conclude that their approach highlights the effects of reasonable but different decisions in the analysis on the results. This conclusion is actually based on weak evidence; almost all teams chose for a form of regression analysis (only three teams did something else). The differences are in the choice of distribution (normal, logistic, or Poisson) or control variables. Whether skin tone has an effect is mostly based on the significance of the regression coefficient of skin tone, which is probably inaccurate because there are obvious dependencies in the data and the target variable is highly skewed.

Given these difficulties, predictive analytics are attractive. The procedure was a 10-fold; for each fold a logistic model was run with all covariates included, once with and once without the skin tone predictor. For each of these two models, the proportion of correct predictions in the holdout sample was calculated. Next, the mean proportion over the 10 holdout samples was calculated. To obtain stable results, this entire procedure was repeated 100 times.

Table 2 shows the results from the first 10-fold. In all 10 folds, adding skin tone to the predictor set results in predictions for the holdout sample that are less accurate. So, after taking the covariates into account, skin tone has no effect. In ten out of ten folds the results are less accurate, using a sign test for the null hypothesis that there is no difference between models with and without skin tone yields a p-value of .002.

What Can We Contribute?

There are some contributions that social science methodology can bring to computational social science. In brief,

Table 2. Proportion of correctly predicted red cards across 100 ten-folds

Fold	Model *with* skin tone	Model *without* skin tone	Difference
Fold 1	0.0353	0.0382	−0.0029
Fold 2	0.0324	0.0353	−0.0029
Fold 3	0.0324	0.0353	−0.0029
Fold 4	0.0294	0.0324	−0.0029
Fold 5	0.0382	0.0441	−0.0059
Fold 6	0.0206	0.0265	−0.0059
Fold 7	0.0382	0.0412	−0.0029
Fold 8	0.0265	0.0353	−0.0088
Fold 9	0.0353	0.0412	−0.0059
Fold 10	0.0440	0.0499	−0.0059
Mean	0.0332	0.0379	−0.0047
Mean 100 repetitions			−0.0060

there is the problem of (lack of) transparency, the issue that the size of the data by itself is not a quality indicator, the meaning of veracity, and the question how well big data analytics works on smaller data. The last section of this paper looks into these questions.

Transparency

There is an increasing interest in "Science 2.0" or "Open Science," meaning sharing and collaborating by researchers, using web resources. Ideas, data, methods, and results are shared, with author copyright protected by various licenses such as copyright, creative commons, or public domain. The goal is that the entire research process is transparent because it is accessible to all interested parties, scientists, or otherwise. Some of the practices in computational social science do not fit well in this open science movement. For example, Google searches lack transparency, because the search algorithm is not accessible. In fact, Google constantly adapts search algorithms to improve them further, so a specific Google search is not even replicable. This is clearly problematic.

Big is Good, Bigger is Better

A central idea behind big data is collecting relatively inexpensive data, aiming at $N = All$. The problem is that such data are not just "found," they have been created for a purpose, and that purpose is rarely scientific research. And, provided sampling is random, the error introduced by sampling is well understood. A case in point is the study by Bond et al. (2012). Bond et al. carried out a randomized experiment sending different political messages to 61 million Facebook users. They report that the messages had an effect, not only on the target, but also on their friends. All effects were smaller than 0.5%. A power analysis requiring a power of 95% with an α of 0.05 shows that a sample size of 4.4 thousand suffices. For an effect size of 0.01%, we need 108.6 thousand. Surely, if we take a 10% random sample of 6.1 million, that should be enough?

Veracity

The name "veracity" contains many data quality issues. IBM (2012) describes veracity as uncertainty of data. Rose Technologies (IBM, 2016; http://www.rosebt.com/blog/data-veracity) describes Veracity as "Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, model approximations." This list is rather diverse. Social science methodologists would prefer to discuss these data quality issues under separate terms such as external validity, internal validity, and construct validity. If we are unsure whether we measure what we intend to measure, the issue is construct validity, and social science methodologists have developed procedures to look into that. It would be an interesting endeavor to develop similar procedures that work for big data.

How Much Room is There at the Bottom?

The analysis techniques have been developed for use on very big data sets. This is what allows techniques such as *k*-fold and ensemble methods. Social science methodologists encounter far smaller samples. The question is: how well do these techniques work with "small Big Data"? Comparative survey research routinely deals with samples of several thousand subjects in dozens of countries. At these, compared with Big Data really small sample sizes, formal statistical testing of structural equation models becomes already problematic, because a small degree of misfit already rejects the model. This has led to an array of goodness-of-fit indices. However, a cross-validation or *k*-fold approach that analyzes how well the structural equation model predicts actual scores in the holdout samples provides a much more direct test. If such procedures can be developed, they can also be applied on Big Data, where explicit modeling may help to turn black box results into white boxes.

References

Attewell, P., & Monaghan, D. (2015). *Data mining for the social sciences*. Oakland, CA: University of California Press.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*, 295–298. doi: 10.1038/nature11421

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods and Research, 21*, 89–115. doi: 10.1177/0049124192021001004

Cioffi-Revilla, C. (2014). *Introduction to computational social science*. Berlin/New York: Springer.

Daas, P., Tennekes, M., de Jonge, E., Priem, A., & van Pelt, M. (2012, September 27). *Statistiek en Big Data: De kracht van datavisualisaties* [Statistics and Big Data: The power of data visualisations]. Paper presented at the Nyenrode Symposium on Big Data. Retrieved from http://www.slideshare.net/marketingfactsnl/big-data-piet-daas-cbs

Dawson, R. J. M. (1995). The "unusual episode" data revisited. *Journal of Statistics Education, 3*. Retrieved from https://ww2.amstat.org/publications/jse/v3n3/datasets.dawson.html

Donoho, D. (2015). *50 years of data science*. Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). (2016). *Big data and social science: A practical guide to methods and tools*. London, UK/New York, NY: CRC Press.

Ghani, R., & Schierholz, M. (2016). Machine learning. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (pp. 147–186). New York, NY/London, UK: Chapman & Hall/CRC Press.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012–1014. doi: 10.1038/nature07634

Gonçalves-Sá, J., Varela, P. L., Wood, I. B., Bollen, J., & Rocha, L. M. (2015, December 2–3). *Human sexual cycles are driven by culture and collective moods*. Poster presented at the 2015 Winter Symposium of Computational Social Science, Cologne, Germany. Retrieved from http://www.gesis.org/css-wintersymposium/home/past-events/css-winter-symposium-2015/slides-videos/

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*, 1360–1380. doi: 10.1086/225469

Grishenko, A. (2014). *Twitter architecture analysis (part 1)*. Retrieved from https://0x0fff.com/twitter-architecture-analysis-part-1/

IBM. (2016). *The four V's of big data*. Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Kollanyi, B., Howard, P. N., & Woolley, S. C. (2016, October). Bots and automation over Twitter during the second U.S. Presidential Debate. *COMPROP Data Memo 2016.2/19 October 2016*. Retrieved from http://politicalbots.org/?p=711

Laney, D. (2001, February 6). 3-D data management: Controlling data volume, velocity and variety. *META Group Research Note*. Retrieved from http://gtnr.it/1bKflKH

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in Big Data analysis. *Science, 343*, 1203–1205. doi: 10.1126/science.1248506

Loehlin, J. C. (1965). "Interpersonal" experiments with a computer model of personality. *Journal of Personality and Social Psychology, 2*, 580–584. doi: 10.1037/h0022457

Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., & Beaton, M. (2013). *Teens, social media, and privacy*. Retrieved from http://www.pewinternet.org/2013/05/21/teens-social-media-and-privacy/

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). *Is the sample good enough? Comparing data from Twitter's streaming API with Twitter Firehose*. Retrieved from https://arxiv.org/abs/1306.5204

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabo, G., Lazer, D., Kaski, K., . . . Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences USA, 104*, 7332–7336. Retrieved from http://www.pnas.org/content/104/18

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Phelps, S. (2012). *The emergence of social networks via direct and indirect reciprocity*. doi: 10.2139/ssrn.2109553

Presenti, J. (2015). *Five new services expand IBM Watson capabilities to images, speech, and more*. Retrieved from https://developer.ibm.com/watson/blog/2015/02/04/new-watson-services-available/

Roberts, M. E., Stewart, B. M., & Tingley, R. (2016). Navigating the local nodes of big data: The case of topic models. In R. M. Alvarez (Ed.), *Computational social science. Discovery and prediction* (pp. 51–97). New York, NY: Cambridge University Press.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., . . . Nosek, B. A. (2015). *Crowdsourcing data analysis: Do soccer referees give more red cards to dark skin toned players*. Retrieved from https://osf.io/qix4g/

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33*, 1–67. doi: 10.1214/aoms/1177704711

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Vo, H., & Silva, C. (2016). Programming with big data. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (pp. 125–143). New York, NY/London, UK: Chapman & Hall/CRC Press.

Webb, E. J., Campbell, D. T., Schwarz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago, IL: Rand McNally College.

Zhu, X., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. San Rafael, CA: Morgan & Claypool.

Published online June 1, 2017

Joop J. Hox
Department of Methodology and Statistics
Utrecht University
POB 80140
3508 TC Utrecht
The Netherlands
j.hox@uu.nl

Appendix

Glossary

Examples. Analysis units, cases, subjects.
Features. The *independent* or predictor variables.
Label. The *dependent* or target variable.
***k*-fold**. A procedure that divides the data in *k subsamples* or *folds*. The analysis is then iterated *k* times. Each time fold *k* is left out, the model is trained on the combined $k - 1$ folds, and tested on the holdout fold *k*.
Machine learning. Fitting a prediction or classification model.
Tree. A regression or classification tree predicts a continuous or categorical dependent variable using splits on the predictors. A tree naturally includes nonlinear effects and interactions.
Random forest. A large number of trees grown for the same data, but including different sets of predictor variables in each tree.
Ensemble method. Fitting many models to the same data, often using different subsets of subjects and/or variables in each model. The random forest is an ensemble method.
Training. Used in machine learning for *fitting*, as in "training a model."
Tuning. Choosing values for parameters that are not estimated but specified by the analyst, for example the number of clusters for a cluster analysis. The optimal value for a tuning parameter is usually found by using cross-validation or *k*-fold techniques.