Future Generation Computer Systems 72 (2017) 11–22

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Why good data analysts need to be critical synthesists. Determining the role of semantics in data analysis

Simon Scheider [a,*], Frank O. Ostermann [b], Benjamin Adams [c]

[a] Human Geography and Spatial Planning, Utrecht, Netherlands
[b] Faculty of Geo-Information Science and Earth Observation (ITC), Enschede, Netherlands
[c] Department of Geography, University of Canterbury, Christchurch, New Zealand

HIGHLIGHTS

- We explain why learning from data is more than just analyzing data, including synthetic tasks.
- We provide arguments from statistical learning, workflow reproducibility, and philosophy.
- We propose a learning model that highlights the roles of semantic technology in data analysis.
- Based on this model, we review current analysis and workflow tools and Semantic Web research.
- We propose a roadmap of 8 challenging research problems which currently seem largely unaddressed.

ARTICLE INFO

Article history:
Received 13 May 2016
Received in revised form
7 February 2017
Accepted 24 February 2017
Available online 2 March 2017

Keywords:
Data driven analysis
Learning
Semantic Web
e-Science
Data science

ABSTRACT

In this article, we critically examine the role of semantic technology in data driven analysis. We explain why learning from data is more than just analyzing data, including also a number of essential synthetic parts that suggest a revision of George Box's model of data analysis in statistics. We review arguments from statistical learning under uncertainty, workflow reproducibility, as well as from philosophy of science, and propose an alternative, synthetic learning model that takes into account semantic conflicts, observation, biased model and data selection, as well as interpretation into background knowledge. The model highlights and clarifies the different roles that semantic technology may have in fostering reproduction and reuse of data analysis across communities of practice under the conditions of informational uncertainty. We also investigate the role of semantic technology in current analysis and workflow tools, compare it with the requirements of our model, and conclude with a roadmap of 8 challenging research problems which currently seem largely unaddressed.

1. What's missing from modern data-driven analysis?

Modern data science is predicated on the assumption that data is synonymous with facts, and thus the more data that can be gathered for analysis, the better. Today's increased availability of data has thus not only paved the way for data-driven analysis, it has even led some researchers to proclaim a radical *end of theory* [1], as well as the rise of a new, *fourth paradigm* of science as such [2]. According to this latter idea, science in the future will exclusively rely on efficient strategies to gather and analyze data, and less and less on theory building itself, a task which is largely taken over by powerful machine learning algorithms. Pushing this to its ultimate consequence, it is an argument for taking theory out of the hands of humans (because they are not only incapable of handling huge amounts of data, they are also error prone) and for reducing science to a mere computational problem.

However, if we take a closer look at the current practice of data science, we see that analysts struggle with issues very different than simply improving their computational efficiency [3,4]. Today, we have more data and computational power than ever before, and we have more software tools implementing a vast array of statistical methods than ever before. And yet, in more and more cases, this results in spurious discoveries, statistically insignificant findings, and ridiculous correlations [5,6].

* Correspondence to: Human Geography and Spatial Planning, Universiteit Utrecht, Heidelberglaan 2, 3584 CS UTRECHT, Netherlands.
E-mail address: simonscheider@web.de (S. Scheider).

http://dx.doi.org/10.1016/j.future.2017.02.046
0167-739X/

A parallel problem is the missing reproducibility of scientific results [7–9]. We often *do* reuse models, methods, algorithms and code from other researchers. However, this can be dangerous and problematic even under optimal conditions [9]. The problem is that effective reuse goes considerably beyond just sharing code or data [10]. Very often, scientists need to know more about foreign data than can be discovered from its surface, causing what one might call *science friction* [11]. This becomes even more important if we want to exploit the advantages of citizen science, i.e. local and traditional knowledge from the citizens, and increased participation for the citizens [12,13].

These issues, along with the increasing publicity of big data and data science more broadly, has led in recent years to social critiques of the big data phenomenon [14]. There is little doubt that much of the massive amount of data that we have now has the potential to open up new discoveries in the sciences and to aid decision making in certain domains. The question is: how can we leverage that opportunity in a way that builds off of the knowledge we have learned about the process of science, an endeavor which still requires humans in the loop performing interpretation in context [15]?

The technology to support data analysts is in principle available, and can be based on Semantic Web standards, ontologies, sharable linked data repositories and other e-Science technology [15,16]. There is also an increasing awareness that the Semantic Web may serve as a way to invest knowledge into the discovery process itself [17]. The knowledge to "do it better" is also there, but it is scattered in many papers and textbooks about data analysis, thus difficult to obtain and internalize, and even more difficult to apply to a specific problem [18]. We argue that *good data analysts need to be critical synthesists*. Under the conditions of uncertain information gathered from diverse sources, there is no meaningful knowledge discovery without information about semantic domains of origin and the context in which data is generated. This synthetic context consists of observation, data and model selection, data derivation and interpretation, and it serves to clarify the roles that semantic technology could take in the learning process. Consequently, there remains a series of unsolved research challenges to develop tools that help analysts in becoming critical synthesists.

In this review article, we present a number of arguments shedding light on the reasons why synthesis should be considered an integral part of data analysis (Section 2), requiring a revision of Box's model of data analysis and learning. We then critically examine current research in analysis support technology in this respect. In Section 3, we investigate the principle role that synthetic and semantic tools could play in the process of data analysis, before assessing in Section 4 how they are currently used in practice. For this purpose, we review recent research on semantics in statistics and data mining and assess in how far existing workflow tools actually cover semantic methods, using a geospatial analysis scenario. Based on these theoretical and practical insights we suggest a research road map in Section 5 that is capable of filling in some of the detected gaps. We formulate eight open problems that should be addressed by future research.

2. The synthetic parts of learning

In 1976, George Box [19] published a pragmatic critique of contemporary statistics, which he argued had by his time become a playing field for specialized mathematicians with little relevance for scientific practice. Methods that were once invented with a particular context and practical meaning in mind, such as Fisher's ideas of distribution-free tests or combinatorics in experiment design, came to live a life on their own, totally disconnected from practice. Box warned us about the consequences of both, mathematicians who do not understand the practical context naively applying mathematical models, as well as practitioners who naively apply mathematical theories to their domain to produce questionable results. He concluded that mathematical knowledge about statistics, which prevents mere *cookbookery*, needs to be backed up by knowledge of scientific practice, avoiding *mathematistry*. This results in a closed loop that links induction with deduction, as well as experimental design with a hypothesis (see Fig. 1).

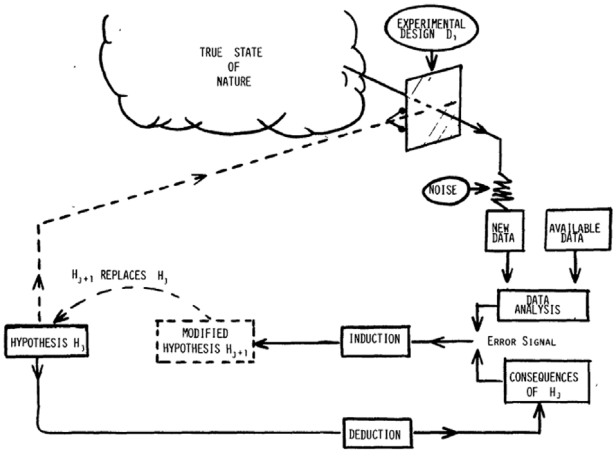

Fig. 1. Box's model of "data analysis in scientific investigations" includes an experimental design "window", but misses other synthetic parts of learning.

Box's model has become sort of a standard in statistical methodology. It has recently been used to underpin modern Bayesian and so called "latent variable" approaches [20]. However, in trying to give it a "modern" interpretation, Blei [20] seems to reduce Box's model to a mere inductive revision loop of statistical models based on data, excluding any considerations of experimental design or background knowledge.[1] In fact, we think that Box's critique concerning the acknowledgment of practice and the underestimation of experimental design perfectly applies also to today's data driven practice, for which Blei's framework is a good example. The "modern" approach seems to reduce every single step of scientific practice into a question about modeling that needs to be answered based on data.[2] We hardly believe that this is what Box intended, let alone what scientific practice consists of. One could say, therefore, that data driven science today has fallen behind Box in neglecting an essential part of science.

However, we think that we need to go into the exactly opposite direction: Box's model is not only in danger of being oversimplified, it is also incomplete itself. Box largely excluded what we call the *synthetic parts of learning* [22], which consist of available knowledge, interpretation, observation, data selection and derivation, terms that we will explain below in more detail. When Box speaks about a theory, then he only means an isolated "hypothesis" that needs to be accepted or refuted, just like in test statistics. This neglects the dependency of acceptable statements on other kinds of knowledge and leaves open how a hypothesis alone (without any other kinds of knowledge) could ever serve to design a statistical experiment to test it. Furthermore, when Box speaks about an "error signal" that causes a model revision, he only means a contradiction between predicted and observed facts which leads to an inductive correction step of the model. This, however, neglects that in scientific practice, counter-induction and non-analytic methods are often used which ignore the *rule of facts*, as explained in the following.

[1] This is very much in the tradition of Bayesian model averaging, where decisions for models are not based on knowledge but on how well they explain data or how much faith we put into them regardless on which grounds [21].

[2] Note that, paradigmatically, the word "deduction" completely disappeared from Blei's model, and that what he calls "critique" is merely model testing based on new data.

2.1. Against the rule of facts: arguments from philosophy of science

Very often, new problems can be recast in older terms to show the futility of alleged solutions. Philosophers of science have been trying for centuries to unravel the method of science, and in doing so, created forceful ideologies, such as Popper's idea that scientific progress is a linear process of approaching truth (the *correspondence* theory of truth), the idea that science can be clearly separated from pseudo-science (the *demarcation* problem), and that the latter is a matter of exposing theories to the scrutiny of empirical facts [23]. In the history of science of the 20th century, this led to the widespread opinion that facts (in the form of data) are what makes science scientific, and that therefore every method that induces theories from as many facts as possible will advance science.

This view, however, has been confronted with substantial critique from other scholars, such as W.V.O. Quine [24], W.W. Bartley [25] and P. Feyerabend [26]. For instance, Feyerabend [27] gathered a large list of evidences from the history of physics to show that scientific progress in fact often relies on:

- *Counter-induction*, i.e., on choosing theories that seem unnatural or even contradict given facts. Reasons for counter-induction are threefold: the primacy of alternatives, the dependency of facts on theories, and the possibility of re-interpretation of facts. We shortly explain the latter two.
- *The dependency of facts on theories*. Which facts are available is influenced by the theory one adopts. This is not only because under a given theory, some facts may appear trivial or remain unnoticed, but also because alternative theories make certain facts observable in the first place, as they inform us on *how* to observe matters of interest. An example are Galilei's observations of the moon through a telescope, which were influenced by novel theories of vision, unraveling some observed properties of the moon as being mere artifacts of the telescopic process of perception (cf. [27], Chapter 10).
- *The re-interpretation of facts*. The reason why facts contradict theories is not necessarily because theories are wrong, but because facts may need to be reinterpreted in order to resolve the contradiction. Thus, in the light of a contradiction, one always has a choice as to whether one changes a theory or whether one rather changes the interpretation of facts that contradict it (cf. [24]). An example is Galilei's re-interpretation of facts about the motion of a falling stone in his refutation of the Aristotelian *tower argument* against the motion of the Earth (cf. [27], Chapter 6).

Our point is that these insights can serve to protect us against uncritical views on futile promises and alleged benefits of a data driven science. We illustrate this with a couple of examples in the following.

For example, the dependency of facts on theories has practical consequences for those people who want to base data-driven science (solely) on knowledge discovery. Suppose you obtain data about passenger statistics in public transportation in different cities [28]. Suppose the best predictor of traffic (with the least squares error under cross-validation) over all cities is a non-linearly increasing monotonic function of tour frequency, since the passenger distribution spreads a lot between larger and smaller cities. Note this pattern will not change regardless of how many cities one adds to the data. In later tests it turns out, however, that predictions of this model catastrophically overestimate higher frequencies in smaller and middle-sized cities. With a bit of (economic) background theory, one could have guessed that passenger numbers can never increase with tour frequency in an unlimited way, since there must be a saturation of mobility needs for any given city. The joint distribution of passenger frequency over all cities is, however, a mixture of the distributions of very different cities. Thus, it would have been advisable to prefer a conservatively increasing model, and, in effect, to "distrust" or "disregard" some facts about tour frequency dependencies in incompatible cities in the joint data set based on a background theory.

Furthermore, re-interpretation of facts has similarly devastating influence on the effectiveness of a data driven science. To illustrate, we use an example first suggested by Goodman [29]: suppose there is a color term called "grue". Grue applies to all observations of *blue* things before a given point in time t. After time t, it applies to all things that are *green*. Up to time t, based on gathering observable facts about utterances of "grue" in the presence of blue things, and based on the conservative strategy of choosing only the most simple theory that explains the given facts, one would end up with the wrong conclusion about the extension of "grue" ("grue" means blue), no matter how many facts were gathered before time t. We suggest that Goodman's example perfectly illustrates how a data-driven science will suffer from semantic re-interpretation. Grue is a color term that is re-interpreted from green to blue. In the light of contradictions of our learned color theory (predicting "this is grue" for blue things) with observations after time t ("this is not grue" for blue things), we should better learn how to re-interpret "grue" as green, rather than adapting our color theory to contradicting facts.[3] If this example seems artificial, consider that this is exactly the situation faced by biodiversity scientists who use syntactically equivalent taxonomical terms that have quite different semantics as the taxonomic classification is updated and changed over time [30]. The same situation occurs in the Earth and environmental sciences, where over time different methods of land use categorization result in very different semantics for regions mapped to syntactically equivalent land-use labels [31]. In summary, what is needed are methods that allow researchers to effectively disregard or preselect facts.

2.2. Why bias is needed: arguments from (statistical) learning under uncertainty

In order to learn a curve from observations, it is always possible to choose among a set of linear and non-linear regression models. However, this choice is often not clearly decidable from empirical evidence: it needs a *theoretic bias* [32,33], an a-priori decision that is based on a theory or on general experience of a domain, and not on data.

Why is this so? The reason lies in *uncertainty*, but not in stochastic uncertainty, rather in the *lack of knowledge about what information is relevant*, which leaves us no way of determining optimal behavior [35]. This contradicts some fundamental assumption of statistical learning, namely that optimal predictors *minimize bias error*[4] by *maximizing effort*, i.e., by being maximally flexible in adapting to the imagined "true" function that needs to be predicted using more data, more computation, and more time.

The program of minimizing error by minimizing bias and maximizing computational effort, however, has very limited chances of success under the conditions of informational uncertainty. An optimal estimator is only optimal under the assumption of "optimal" knowledge. For statistical learning [36], this essentially means samples are *representative* (have low noise), (2) large (relative to the number of variables), and (3) phenomena are predictable. If any

[3] The reason is that the latter can only result in an impoverished color theory, where color terms like "grue" can mean blue *or* green, and thus the distinctive power of color terms is lost altogether.

[4] The *bias error* is the difference between the "true" function to be predicted and the function induced from samples by an algorithm, averaged over these samples.

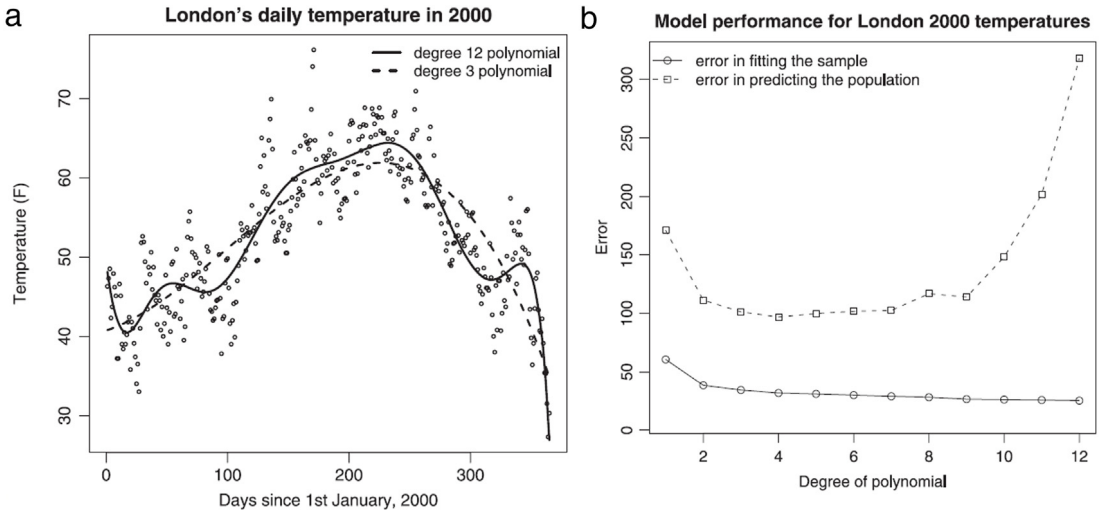

Fig. 2. The bias–variance trade off illustrated by daily temperatures of the year 2000 in London. Polynomials of different degree can be fitted into the scatter plot (a), but even though the error of fitting decreases, the prediction error increases with the degree (b). Furthermore, even a balanced model fails to learn that temperature is a periodic phenomenon.
Source: Example taken from [34], by kind permission of Gerd Gigerenzer and Henry Brighton.

of these criteria are not met, then the most sophisticated estimators can perform worse than simple *heuristics* (compare the evidences given in [34]). Heuristics, in contrast, typically have a large bias error but a *minimal variance error*[5] with *minimal effort*. That is, under informational uncertainty conditions, heuristics do not trade accuracy for efficiency, but rather bias for variance error, and thus can be both more efficient *as well as* more accurate [34].

In contrast to heuristics, textbook measures against variance error, such as cross validation and balancing of model complexity, [36] (1) do not take into account background knowledge and (2) can easily fall victim to biased samples, as illustrated by the following. Consider the case of choosing among a set of polynomials of increasing degree to predict London's mean daily temperature over the year 2000 (see Fig. 2(a)). One can fit a curve with increasing perfection to this data, as depicted in the lower curve in Fig. 2(b), but this usually increases the error with the polynomial degree (upper curve in Fig. 2(b)). This is normally called "overfitting", and it means that the flexibility of a predictor is too high and models noise instead of the temperature function in question. If the sample size is large and representative, then we can penalize overfitting based on random re-sampling (cross validation) or by balancing model complexity with the minimum description length principle.[6] In this way, we might be safely guided in choosing, e.g., the degree-3 polynomial in Fig. 2(b), a moderately complex and balanced model. However, in this specific case, even a balanced model is actually not adequate, because it lacks a certain kind of bias: temperature variation is strongly periodic, and therefore the ends of the temperature curve should meet at the beginning and end of a one-year period. Even the degree-3 polynomial fit in Fig. 2(a) does not take into account this basic kind of background knowledge. It suggests that Dec. 31 is a singular temperature minimum. A semantically adequate model would consist of a periodic function representing the average temperature variation plus a polynomial for the particular deviations over the year 2000. Regarding the second claim, if the data sample is biased itself, then the only chance we have is to rely on a heuristic. Only background knowledge can then usefully guide our choice of a plausible model. To illustrate, consider the case of a large sample of road traffic measured at the locations of billboard posters to assess their value for advertising [37], which is naturally biased towards streets with high traffic frequency. Estimating traffic on side streets with this sample will inevitably lead to overestimation [38], regardless of any measures taken against overfitting. Preventing overestimation of road traffic in small streets requires a-priori knowledge of traffic conditions, which may, for example, enter a KNN predictor of road traffic in the form of synthetic ("low traffic") data exemplars for side streets [37]. Further examples were found by Gigerenzer, who has shown in many studies how "less can be more", i.e., how heuristic models can outperform optimal estimation models under informational uncertainty [34].

The robustness of biased methods, however, comes at a price: whereas machine learning methods adapt to any possible domain, biased models depend on a domain or a certain type of information [34]. We argue therefore that the only way to fight the detrimental effects of informational uncertainty in data analysis lies in knowing your domain, not in building more powerful general-purpose learning models or using more data. And the only way to support this technically over distant communities of practice is to use *semantic metadata*. In principle, there are two ways how metadata can be used in countering uncertainty:

1. Either, one can use *biased re-sampling*, i.e., one can ignore or weight data items based on knowing what items are relevant and which are noise depending on what is modeled. A standard example for this is outlier removal, but also sample stratification and data prototyping [37].
2. Or, one can use *biased models*, i.e., heuristics for prediction. This also depends on knowing your domain, because whether a heuristic or biased model works well depends on an information environment [34] that needs to be described with metadata.

2.3. The curse of modularity: arguments from re-usability of analysis

How could we support analysts in biased model/data selection? A crucial role in this respect is played by the way we help researchers reuse workflows. Current approaches to re-usability of analytic methods often rely on the idea of re-computation [39];

[5] The *variance error* is the difference between the mean function and the function induced from each sample. It exactly captures that part of the error which comes from overfitting, i.e., from uncertainty about what is noise and what is representative for the process to be modeled.

[6] https://arxiv.org/abs/math/0406077.

we hope to be able to reproduce and reuse analyses by sharing software code and data. Reusing analysis methods across application domains, however, goes considerably beyond replicating results [10]. For this purpose, the mere sharing of executable code and source data is not sufficient. It needs to be amended in some manner to share workflows that can be adapted with respect to data sources and models.

Why is this so? Notice that all arguments discussed above, starting from counter-induction, through the dependency of facts on theories, to the necessity of biased re-sampling and biased model selection, basically require a way to react to a data analysis problem by *adapting* parts of the processes involved in a purposeful way, either by adapting observation techniques, data, tools or models, or even by adapting data or model interpretation. This adaption needs to be based on *a-priori knowledge*, and so is difficult to manage simply based on sharing data or software scripts. Also, particular data and software may be unavailable for various reasons (e.g. license restrictions or simply research culture).

What is needed is rather a *modular* way of sharing workflows, namely by distinguishing tools from methods and data from their semantic types. This requires abstracting from particular tools and data sets and describing data analysis on a conceptual level [40], which would allow partial adaption of workflows across computing environments. It would also allow analysts to focus on the questions they want to answer and the methods to answer them, instead of the software and data formats needed for computation [41,42]. A particular challenge is to decide when such adaptions can be considered *meaningful* [18]. Building truly modular analysis workflows first requires a more comprehensive picture of the synthetic operations involved in data analysis and learning.

3. An alternative, "synthetic" model of learning

In this section, we argue that the crucial role of observation, selection, derivation, interpretation and background knowledge leads to a different, "synthetic" model of learning (depicted in Fig. 3), in which the import of semantic operations and semantic metadata stands out clearly from the background of classical analysis procedures.

3.1. Model revision: semantic conflicts vs. data conflicts

If we analyze data or models, then we compare them against each other as well as against our theories. In Fig. 3, a conflict (detected by an error term or a logical contradiction) can occur on the level of data (where derived data may contradict observed data), as well as on the level of statements that data is interpreted into (e.g., when a statement extracted from data contradicts a known theory). In Box's model (Fig. 1), a comparison is only possible on the data level, and only to a single purpose, namely in order to detect data conflicts that may lead to a revised selection of a model. However, in reality, in such a case, we may also decide to keep our model, and instead either re-observe, re-sample or generate new data, all of which might resolve the conflict. Furthermore, we might also re-interpret data into different conceptual statements, which might resolve the conflict on another, *semantic* level (Fig. 3). Even if data and model might not be in conflict, then the interpreted statements it produces might still be in conflict with our theories, and thus cause us to revise a model just because we put more trust in theories.

Semantic conflicts are conflicts between interpreted statements and conceptual theories, and they exist independently from data conflicts. We argue that both kinds of conflicts can trigger adjustments on all levels of the following analytic and synthetic operations (Fig. 3). These operations are also summarized in Table 1.

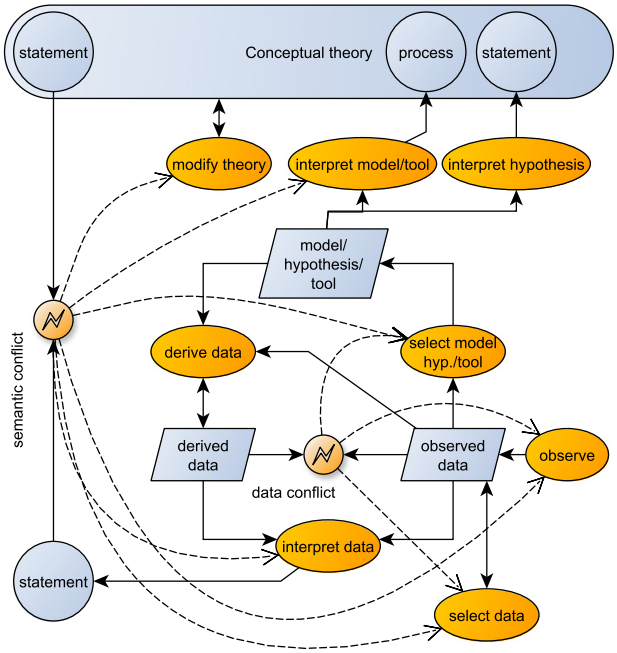

Fig. 3. A model of the operations involved in data analysis and learning. Orange ellipses denote operations and normal arrows their in- and outputs. Boxes denote participating data and tools and big circles denote semantic concepts. Dashed arrows denote influences between chosen operations. One might "enter" this model at any point in the process, and conflicts may be resolved by modifying any kind of operation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Observation

Observation processes not only generate data, they also influence our interpretation of data [43]. For example, knowing that a temperature sensor was calibrated in a certain way and measured one meter above the ground surface allows us to give this temperature quality a *spatio-temporal support* as well as associating it with a *phenomenon* in a domain theory [18], such as ground surface temperature. Making this distinction is crucial for analysis. If we confuse ground surface temperature with the temperature of the ground, we mix up two different domains of interpretation of temperature measurements, and thus two different stochastic processes with different probability distributions and means. Furthermore, the way we interpret data can influence the way we observe. For example, we may plan a certain sensor measurement to gather new data based on interpreting existing data as ground surface temperature. Observation also influences the way we interpret a model. For example, we can interpret a KNN model as a model of the traffic process only if it was trained on traffic data. Semantic technology could therefore help analysts in sharing information about *observational origins* of data, which helps them interpreting data as well as models.

The need for semantic information on observations was formulated a decade ago for ecological data, through development and use of a formal ontology [44]. This was followed by a proposal for a Semantic Sensor Web [45] and its application using the OGC's Sensor Web Enablement's Sensor Observation Service (SOS) [46]. However, it remains to be demonstrated that such an approach can cope with the increasing volume of sensor data being generated. The emergence of several ontologies related to data collection (e.g. [47]) has led to recent research activity on semantic mediation of observation ontologies for the SOS [48] and Observation & Measurement standards [49].

Table 1
A summary of analysis operations for which semantic technology is crucial.

Operation type	Subtype	Example	Semantic information
Observation	Sensor measurement	A smartphone sensor	Sensor calibration, sensed phenomenon and measurement error
	Human observation	User-generated web content	Expertise and trustworthiness of observer
Tool selection/data derivation	Data fusion	Table join	Entity (web) identifiers and space–time support
	Data summarization	Box plot, estimating any statistic parameter	Scale levels and meaningful statistics
	Data transformation	Spatial reprojection	Standardized (coordinate) reference frames with web identifiers
	Data construction	Building a ratio from two numbers	Scale levels and meaningful data derivation
Data selection	Retrieval	A statistical data portal	Question based analysis support: what constitutes relevant data for an analytic question?
	Biased resampling	Stratified resampling	What are relevant strata/important data points?
Data interpretation	Cleaning	Outlier detection	
	Data annotation	Provenance and workflow models	Semantic workflow support
Model/hypothesis selection	Biased model selection	Linear regression	Best practice model repository for problem domains
	Formulation of hypothesis	0-hypothesis selection	Question based analysis support
Model/hypothesis interpretation	Model documentation	annotation of a regression model	Model repository for problem domains
	Hypothesis documentation	Linking hypotheses to research question	Model/question repository for problem domains

3.3. Selection of data

As important as observation "from scratch" is the ability to select from existing data. Under this heading, we subsume *data retrieval* (the selection and access of data from external sources), but also *resampling* (the selection of data subsets), as well as *data cleaning* and *preparation*. Retrieval is clearly knowledge-based, but also resampling and data cleaning sometimes need to be done in a *biased* and thus *domain specific* way (compare Section 2.2). Semantic tools have been proposed to support analysts not only in retrieval of data [50,51], which is an on-going major research effort, but also in *biased sampling and data cleaning* [52–54], i.e., in deciding which sub-sample is representative and which is adding only noise.

3.4. Selection of data derivation tools, models and hypotheses

We often need data in a different form in order to answer questions not answerable based on the available data sets. One way to obtain such data is *data derivation*, i.e., fusing, summarizing, transforming, or constructing data. For example, suppose we want to compare the age structure across statistical regions in space, but the only kind of data available is absolute population counts for different groups of people. Then, we can construct a new data set of age class ratios by combining absolute numbers for each region. This new variable has a different meaning, and thus gives rise to different kinds of models. It also requires that underlying counts have comparable spatial and temporal references in order to prevent semantic conflicts and meaningless ratios. Semantic technology could support analysts in deciding whether such derivation tools are *meaningfully applicable* to a data set [55]. Also they could help them in *formulating an appropriate hypothesis* that corresponds to a research question [42,56]. As another example, in the geosciences, georeferencing information (assigning coordinates to information including a textual reference to a place) is an important and ubiquitous step. However, the process itself is often not disclosed completely, which can have important ramifications for future analysis. For example, in a previous study [57], Tweets were geocoded based on lexical matching with LAU2 placenames, and the centroid of a matching municipality was assigned as a coordinate. This was sufficient for a study of the spread of bushfires with social media data, but others using the data set for more precise kinds of analysis may require a different geocoding method.

While *selection of models* in data science is mainly data-driven, it used to be (and in many applied science disciplines still is) a matter of a domain theory. Data driven models can be more accurate and successful than domain driven models, but this is not necessarily the case under uncertainty (Section 2.2). The selection of a model can also be driven by its (or its underlying data's) interpretation into a theory, and may even be done by neglecting data. Semantic technology could support scientists in choosing models based on a library of successfully tested models for a given semantic domain, allowing computational scientists to *share and document models* and their results in a way that goes beyond sharing software and code [41]. Some of the workflow tools we present in the following section allow this.

3.5. Interpretation into background knowledge

As shown in Fig. 3, hypotheses need to be interpreted into theory statements and linked to research questions, models need interpretation into (stochastic) processes, and data need interpretation into theory statements. Note that on a conceptual level, statements can also be produced in a theory, not only through a data driven environment, and interpretations can be adjusted in case of a conflict. In fact, all operations discussed so far influence and are influenced by interpretation processes. Background knowledge has played a large role in ontology engineering, however, its integration into analysis tools is largely missing. For example, the background assumptions usually taken in statistics (about forms of distributions and covariance) are usually not made explicit in terms of meta-data [18]. One proposed solution is to aggregate data and metadata that enriches the specifications of workflows into research objects [58]. Another suggested approach is to generate workflow description summaries and aggregate workflow elements using reduction to primitives [59].

4. The current role of semantic technology in analysis

Once it is clear in what ways semantics can help in the analysis process, we can identify corresponding support technology. In this section, we investigate the state of the art in data analysis support technology and the role that the Semantic Web currently plays in it. We first review related research efforts in semantic based statistics and data mining, before we take a closer look at existing analysis workflow management systems. Our specific view on this work is driven by the question in how far it supports the synthetic operations identified in Table 1.

4.1. Semantic support for statistics and data mining

Biased theory building and deductive elements of learning used to be a topic in early machine learning books, namely in the form of *analytical learning* (learning from scarce data with a biased domain theory) and explanation-based generalization [60,61], and in artificial intelligence based endeavors to model scientific and commonsense reasoning [62]. However, since the triumphal success of data mining in the nineties and 2000s, such kind of research was largely given up.[7] The current state of the art in statistical learning theory [36] seems to have excluded semantic issues from its agenda. Nowadays the dominant way how machine learning and Semantic Web technology contribute to each other is by helping the latter scale across unstructured texts or noisy linked data [64], e.g., in terms of Semantic Web Mining [65].

There is, however, a growing number of researchers who investigate potential linkages between statistics/knowledge discovery and semantics.[8] For example, there is recent research on bringing statistical meta-data, such as data cubes [66], on the level of the Web for increased accessibility. Also, the value of ontologies for data mining has been investigated by [67]. In a revised knowledge discovery (KDD) process on graph data, Ristoski and Paulheim recently suggested that data resources as well as mining results all should be linked to the Linked Open Data (LOD) Web [17]. In fact, a lot of research has demonstrated how ontologies and linked data can support the whole KDD process, including enrichment of tabular data on the Web [68], data cleaning (outliers and redundancy reduction) [52,53], feature vector generation (propositionalization of linked data) and feature selection (based on relations between features) [54], valid data mining process design (as in Rapid Miner [69]), as well as generalization of learned hypotheses [70] (compare the complete survey in [17]). However, the authors conclude that reasoning on expressive schemas and ontologies about the analysis workflow itself (e.g., to infer result types or to advise meaningful analysis steps for given data) is rarely done in data mining [17].

In the information ontology community, there is research on measurement scale ontologies [71] and observation concepts [72,73] that can be used for choosing relevant services and data, but it does not focus on data analysis. There are many approaches to integrating ontologies into workflow management systems [58], but these are mainly used for purposes of documentation and replication, not for analysis support or reproducibility across data or tool environments [10]. The work in [74] is a notable exception, in that the authors propose a reasoning and query service for interpreted observations gained from biomedical experiments. A semantic approach to meaningful spatial data analysis is presented in [18].

In summary, while many semantic methods for sharing data and workflows and for supporting the KDD process in various ways have been proposed, current research hardly addresses the problem how Semantic Web tools could support researchers in designing the process of data analysis in a modular way, including synthetic operations such as observation, biased data/model selection or modular and meaningful workflow design. It seems that data analysis is still largely considered a process that is independent from semantics and thus supported mainly with analysis methods, not with synthetic methods. In the following, we will take a closer look at existing workflow management systems.

[7] Observe, e.g., how the idea of incorporating *prior knowledge* still ranges among future work in Mitchell's later publications [63].

[8] Compare the workshop series on Semantic Statistics (SemStats 2015) https://semstats.wordpress.com/.

4.2. Semantic support for workflow management

In this section, we investigate workflow management systems for their semantic support of the analysis process. We do this using an example analysis workflow from the geographic information science domain. There are two reasons for this choice. First, the authors' backgrounds in the Geosciences provide them with a background in a domain that relatively early on began to embrace semantic technologies to improve discovery and interoperability of heterogeneous data sets [75]. However, at the same time a typical Geoscience research team is comparatively small, thus they often have little incentive or resources to develop or adopt advanced scientific workflow management systems. Second, geographic information science is a highly integrative science, potentially interfacing with environmental science, social sciences, and planning sciences. Therefore, there is a strong potential benefit from workflow management systems that allow interdisciplinary sharing of analysis workflows.

Our example workflow combines relatively straightforward, typical objectives and design choices in geographic information science. An analyst wants to compare the frequency of georeferenced social media posting with the official population density records. The objective is to find out at which geographic scale population records can serve as a normalization factor for detecting unusual clusters of social media posts. The motivation is that we can assume that clusters of social media activity are dependent on population density (more people, more posts), but authoritative population records (i.e., census) captures mostly residential (nightly) population, which may not be suited to act as baseline for social media activity normalization. The required analysis steps can be summarized as follows. First, we need to obtain two data sets, one for georeferenced social media, one for population density. Then, we need to count the number of unique geosocial media contributors per census district (point-in-polygon analysis). Lastly, we need to compare this geosocial media distribution with the population density and identify the correlation between them (and possible clusters of high or low correlation).

There are several potential pitfalls in this analysis, which need to be captured in semantic annotations in order to make the whole workflow shareable. They include varying geographic coordinate systems (often, geographic data come in different coordinate systems, requiring harmonization or re-projection before analysis), the modifiable areal unit problem (aggregating over areas is influenced by their geometric shape, e.g. changing census districts), edge effects of study area (influence from outside factors on study objects at the periphery, or analysis methods that rely on neighborhoods), uncertainty and imprecision of geolocation (e.g. lineage of georeferencing can be different for two data points from the same source), scale effects (at which scale do we measure?—the finest scale is not always the most suitable), and parameterization of quantitative methods (in this case spatial clustering). The gold standard would be a fully semantically annotated workflow which allows modular reproduction of the analysis by researchers from different domains, e.g. a geographic information scientist or a sociologist or a urban planner, and also addresses the most common reasons for workflow decay over time [76]: insufficient documentation, missing example data, volatile third-party resources, and incompatible execution environment.

For a structured investigation of workflow management systems, we developed (partly based on [76]) several criteria that need to be fulfilled for a gold standard workflow:

- Annotation granularity. At which level or granularity of the workflow are annotations possible? Possible values are Node (every single atomic analysis module can be annotated, required for the gold standard), Model (the combination of several Nodes into a coherent workflow, or the combination of

several atomic analysis steps into an aggregated Node), Project (the project file, which can contain several Models and many Nodes).
- Annotation abstraction level. At which abstraction level are annotations supported? Possible values are Natural Language, Formal (any formalized language), or both (gold standard).
- Annotation languages. Which are supported: formal and structured languages for storing and sharing the annotations or the whole workflow? This is a list of supported languages, e.g. XML and dialects. A gold standard would support as many as possible.
- Annotation generation. How are the annotations generated? Possible values are Manual (the analyst has to enter every annotation manually), Automated (the workflow management system creates annotations based on the meta-data of input data sources and analysis nodes), or both (gold standard).
- Semantic richness. How rich are the semantics of the analysis workflow? Possible values are High (a gold standard workflow management system checks the meaningfulness of an analysis step, e.g. combining different geographic coordinate systems, different temporal resolutions or periods, or parameterization of clustering), Low (the workflow management system only enforces data type consistency), or None.
- Recommender system. Does the workflow management system recommend analysis steps based on input and expected output, e.g. the operation types of data or tool or model selection as presented in Table 1? Possible values are Data, Tool, Model, combinations thereof, or None.
- External resources. Which external resources does the workflow management system support? This is a list of resources, the more the higher the support for extensibility and interdisciplinarity.

4.2.1. Review of workflow management systems

As a starting point for the choice of workflow management systems, we reviewed Talia's recent work [76]. Several of the examined workflow management systems do not seem to have strong development support anymore (exemplified in lack of an accessible website, time elapsed since last update, low forum activity, etc.). After excluding those, seven workflow management systems remain: Kepler [77], Knime [78], Orange [79], Taverna [80], VisTrails [81], Pegasus [82], and WINGS [83]. Although the scope of this paper does not allow an in-depth evaluation of these complex systems, the ability to test them first hand was deemed a necessary requirement. However, given that our test case is from the geospatial domain, we added ESRI's ArcGIS model builder[9] to our set as a widely used workflow management system in the geospatial domain. A brief description of each tested system follows:

Kepler is a very mature scientific workflow management system. It builds on the Ptolemy II engine that users Modeling Markup Language (MoML). Kepler's origins from 2002 are in the ecological and environmental sciences, but it has found wide adoption in many other scientific disciplines.

Knime is another mature scientific workflow management system, based on Eclipse, having a strong support and user base (commercial versions exists), and offering many extensions with wide functionality. It was originally developed to support pharmaceutical research, but has found adoption in many other business areas as well.

Orange is a relatively recent addition to the pool of scientific workflow management systems. Its focus lies on ease of use, with an intuitive GUI providing access to SciPy functionality. It is developed by a bioinformatics institute.

Apache Taverna is another 'established' scientific workflow management system and widely in use. It strongly supports service-oriented architecture through the functionality of plugging in network processing services with a URL. It was originally developed within the bio and life sciences.

VisTrails supports a very low-level assembly of workflows, with the components providing atomic functionality. Its domain background are computer sciences.

Similar to the Kepler and Taverna scientific workflow systems, **Pegasus** allows for the construction of models of data flow with very complex topologies that can be mapped to high-performance computing systems with extensive error handling. The **WINGS** system has been designed to build off of Pegasus and similar workflow systems.

ArcGIS model builder is a module of the ArcGIS software suite that allows researchers to build geoprocessing workflows using a visual language including a huge variety of geoprocessing tools and models.

4.2.2. Applying the example workflow

Our test is based on trying to implement the straightforward geospatial analysis workflow in all available systems. An initial observation regarding their utility for geospatial analysis is that most of the systems do not support basic geospatial analysis methods. Kepler has geospatial actors (through GRASS[10] and GDAL[11]) that allow coordinate system transformation and point-in-polygon analysis (among others), and allows to include R[12] scripts, which provides (spatial) clustering algorithms. However, support for geospatial analysis seems to have stopped (the website referenced in the point-in-polygon module has not been updated since 2007). Knime has some geospatial extensions (OpenStreetMap[13] and ESRI shapefile[14] support), but their functionality is basic. None of the required case study analysis steps is currently supported, although there are plans to bring Open Spatial Analytics[15] to Knime. Apache Taverna allows in principle any geospatial operation to be included through the OpenGeospatialConsortium[16] WebProcessingService[17], but apart from the pioneering work of de Jesus et al. [84] there appears little progress. VisTrails and Orange are both built on Python, and therefore facilitate plugging-in of Python scripts for all the required analysis steps. However, there is little innate geospatial analysis support. It should be noted that the available tools are very powerful, very complex, and possess a strong community and corresponding diverse ecosystem in other domains. Therefore, our analysis is possibly not complete. The single domain-specific system ArcGIS model builder offers all of the required analysis functionality, but this is to be expected given its background in the Geosciences. Although there is no direct module to import geosocial media from APIs, these can be included through

9 http://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/modelbuilder/what-is-modelbuilder-.htm.

10 Popular and established free and open source geographic informations system (https://grass.osgeo.org/).

11 Widely used library for transformations of geospatial data (http://www.gdal.org/).

12 Free and open source project for statistical computing (https://www.r-project.org/).

13 https://openstreetmap.org.

14 Industry de-facto standard file format for geospatial data.

15 http://www.crcsi.com.au/research/2-rapid-spatial-analytics/2-23-open-spatial-analytics/.

16 International not-for-profit organization committed to develop open geospatial standards (http://www.opengeospatial.org/).

17 Inferface standard providing rules for networked geospatial analysis (http://www.opengeospatial.org/standards/wps).

Table 2
Semantic support for (geospatial) analysis in current analysis workflow tools.

System	Annotation granularity	Annotation languages	Annotation generation	Semantic richness	Recommender system	Annotation abstraction level	External resources
Kepler	Nodes	XML	Manual	Low	None	Natural language	Ontologies, Web Services
Knime	Nodes	XML	Manual	High	None	Natural language	
Orange	Project		Manual	Low	None	Natural language	
Taverna	Model		Manual	Low	None	Formal and natural	MyExperiment, Web services
VisTrails	Model	XML	Manual	Low	None	Natural language	Crowdlabs, Web services
Pegasus	Model	SQL	Both	Low	None	Natural language	
WINGS	Model	OWL	Automated	High	Data	Formal and natural	Ontologies, Web services
ArcGIS model builder	Nodes	Python	Manual	Low	None	Natural language	

Python scripts with little effort. As a next step, we investigated how each of the tools supports semantic annotations to make the workflow reproducible.

Table 2 summarizes the results of the test regarding the criteria for semantic annotation. In summary, the tested workflow management systems offer little support for an analyst to enrich the workflow with structured semantic annotations.

Orange is the most accessible system, but is geared towards single users with little support (or encouragement) for semantics and sharing. Workflow semantics are limited to free-form annotations within the project or some components. Knime has several extensions aimed at incorporating semantics and elements of the semantic web, but a search of the extensions and forums seems to indicate that the only way to use ontologies is to model them as network graphs (for which ample support exists). Apache Taverna supports the myGrid ontology and semantics at the level of component families, i.e. groups of components that the creator of a workflow can define. Vistrails supports detailed logging and versioning for collaborative analysis, but there is no explicit support for semantics or ontologies.

At the other end of the spectrum, Kepler and WINGS offer the most support for analysis semantics. Kepler offers ontology-based semantic annotations. Its MoML does not define any semantics for the connections between workflow components. However, it does represent hierarchical relationships between workflow components that have properties, and the components' interaction ports and the connections between those. The Ptolemy II engines uses the concept of director to describe the overall type of workflow and the connections, hence they inherit any properties from the directors, which is another class for MoML. The other important concept is actor, which describes the components of workflow, i.e. the actual processes. These can have semantic type annotations that describe the actor and its data schema. These annotations are part of the component's metadata and can serve to ontology-based discovery, integration and validation. However, at the same time, Kepler's functionality comes at the price of high complexity and a very steep learning curve.

In Pegasus, there is no semantic annotation of the nodes and edges in the workflow. However, the WINGS add-on to Pegasus provides the most advanced system in terms of reasoning capability over semantic annotations. It provides a semantic layer used to generate valid workflow models. The key purpose of WINGS is to use semantics to define *constraints* on both the data that can be used in the workflow as well as how the components of the workflow are connected. These constraints are defined using the OWL standard and build off of existing ontologies, such as the PROV-O provenance ontology [85]. The semantic constraints defined in WINGS can be used to suggest parameter settings for operations based on metadata properties in the data, and they can also be used to suggest data sets that can be used in a workflow. In order to facilitate re-use of workflows, WINGS provides the ability to define abstract components that can be specialized with different operations depending on the type of data that is input into the component. Finally, WINGS can produce metadata for the output data from components of the workflow. However, similar to Kepler, there exists a high practical barrier to making this research software usable and accessible to non-specialist users who are not working in high performance computing environments. Although the ArcGIS model builder supports the example case study in a rather intuitive way, there is no support for encapsulating the semantics and addressing the domain-specific problems outlined above, except manually added annotation boxes. Adding those annotations is tedious work even for a small analysis workflow, and still offers no structured way to annotate. The resulting model can be exported as Python script, which is, however, of little use without the proprietary ArcGIS software libraries. Based on this assessment of the state-of-the-art, we propose in the following section a research and development agenda.

5. A research road map

How exactly should semantic technology support data analysts? The following is a list of research problems that take up the argumentation thread from Section 3 and that currently seem largely unaddressed (compare Section 4). Possible solutions are shortly discussed. For the most part, these problems progressively build off of each other, i.e., each problem will require some kind of solution of its predecessors in order to be solved. We hope that by formulating these problems in an explicit way, they can be better acknowledged in future Semantic Web research.

1. The problem of (semi-automatic) *documentation of analysis workflows*. People usually do not document what they do and for which purpose. Yet, semantic documentation is a prerequisite for any kind of semantic analysis support and for reproducibility of research. This is why machines have to support people in annotating what they do. While automatic documentation of workflows for particular tools is available, the main challenge lies in finding general procedures for annotation of data analysis operations and data types that could be used across data and tool environments [86]. One possibility is to use *information concepts* to describe operations on an abstract level [42], based on their input and output types [87], and to reuse them across the Web of data [58].
2. The problem of *biased data selection*. Under the conditions of informational uncertainty, random sampling is not an option to avoid overfitting. How could semantic technology help analysts in biased sampling, in trusting or distrusting data items [88], and in purpose oriented data retrieval based on knowing about observational origins and representativity for the phenomenon in question? It seems that ontologies of observation and sampling techniques are needed for this purpose, but also a repository of typical distribution parameters found in a domain is needed. While semantically-enabled data retrieval [50] and observation portals [51] have been previously proposed, to the best of our knowledge, the semantic support of biased sample selection seems a largely new exercise (compare [18]).

3. The problem of *biased model selection*. Under the conditions of informational uncertainty, biased models can be more successful than unbiased ones [34]. What would a semantic model repository look like that allows researchers to share and link their models to information environments and modeling problems, allowing them to reuse models that have proven robust in a certain domain and for a certain problem? Data scientists have pointed to the necessity of software-independent model databases [41], however, it remains unclear how such models could be described together with the problems they solve in a way that is independent from scripting languages [10].
4. The problem of *typing and meaningful application* of models and tools. Since data scientists are overwhelmed by the amount of data and tools available, it is crucial to support them in choosing (types of) tools meaningfully applicable to (types of) data in an automated way, very similar to statically typed functional programming [87]. This requires extending the notion of meaningfulness from measurement theory to design *reusable semantic types* that go beyond ordinary data types based on specific information concepts. For example, in the Geosciences, types for data could be spatio-temporal fields or trajectories, and types for tools could be interpolation, and meaningful interpolation can only be performed on data that represents fields [18].
5. The problem of *modular workflow design and analysis design patterns*. Workflow tools nowadays quite successfully document, share, and replicate what has been done, but they hardly help scientists in designing workflows in a *modular* way, and thus, in reproducing research across contexts [10]. So far, the closest instantiation of this idea is the ability to create abstract steps in WINGS workflows [89]. Fostering reproducibility across contexts would require adapting both data inputs and tools. A prerequisite for doing so is that nodes and links in a workflow are semantically typed, so that tools and data of equivalent, similar or even different type can be substituted [40]. In analogous fashion, ontology design patterns were proposed as a cure to limited re-usability of modeling solutions, as they can be adapted to particular applications by filling in open slots [90].
6. The problem of *formulation and linking of analysis questions*. Analysis questions are those questions that motivate an analysis because they need to be answered, and thus determine its purpose. Questions need to be formulated on a formal level that allows machines to share them and to compare and link them with tools and analysis workflows [42]. The question how this can be done remains largely open, even though the problem has gained attention in the e-science community [74].
7. The problem of *recommendation/exploration of analysis solutions*. This problem can be paraphrased as follows: given a data set and an analysis question, which steps would need to be taken to obtain a solution, i.e., an answer to this analysis question? Solving this problem, at least to a limited extent, is a prerequisite for any kind of analysis support system that truly deserves its name. Recently, we proposed an algebraic model of data generation that is capable of computing possible data derivations based on a functional type system [87] and that may be used to construct and search through a data derivation graph.
8. The problem of *indirect question answering* in data analysis. We cannot start our analysis unless we know that an available data set is in principle suitable for our task. A semantically informed retrieval portal should therefore expand a data query in such a way to encompass also those data sets that do not directly answer a given query, but which can be made to do so via some analysis steps. For example, if an analysis of environmental air pollution in Germany requires a fully covering raster data set of PM10, then a data query could also search for point measurements based on knowing that with a suitable interpolation technique, the point set can be turned into a PM10 data set covering the whole of Germany [18]. In order to expand queries in such a way, one needs to be able to construct possible analysis solutions, and then reformulate the data query to search for possible data inputs [87].

6. Conclusion

In this article, we have examined a number of arguments for regarding data analysis as a synthetic enterprise, where knowledge is not only extracted from data but needs also to be invested into the analysis process in order to prevent non-reproducible or meaningless results. From a theoretical viewpoint, the reasons lie in the need for counter-induction and biased learning under informational uncertainty. From a practical point of view, this essentially requires tool support not only regarding model building and inference and the computational replication of workflow processing steps, but also regarding the critical adaption of workflows across data, tools and research environments. This will help analysts to take into account semantic conflicts, perform conceptual (re-)interpretations and biased selections of models and tools, as well as give them information about a data set's observational origin.

While current research on the interface between semantics and analysis focuses largely on using machine learning to help the Semantic Web do its job (of classifying unstructured data), or on linking and publishing data and results in analysis portals, there seems a void of research that specifically addresses how the Semantic Web could support the data analysis process itself. Current workflow tools offer a very limited level of semantic annotation. They focus on tool documentation and computational replication of workflows, while genuine reproducibility would require semantic abstraction from particular data and tool environments, allowing analysts to do modular adaption. Furthermore, semantic support of the workflow design, i.e., of data, tool, or model selection and the meaningful applicability of methods, as well as the linking of workflows with research questions, seems largely missing. In this article, we have suggested a research roadmap that shows a way to incrementally fill this gap, starting from Semantic Web based automatic documentation and sharing of workflows and moving toward functional typing of analysis operations and the linking of workflows to questions, which may lead to a new generation of analysis recommendation technology.

Acknowledgments

We would like to thank Gerd Gigerenzer and Henry Brighton for their inspiring work and the permission to reprint Fig. 2.

References

[1] C. Anderson, The end of theory: The data deluge makes the scientific method obsolete, Wired mag. 16 (2008).
[2] A. Hey, S. Tansley, K. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009.
[3] A. Jacobs, The pathologies of big data, Commun. ACM 52 (8) (2009) 36–44.
[4] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J.M. Patel, R. Ramakrishnan, C. Shahabi, Big data and its technical challenges, Commun. ACM 57 (7) (2014) 86–94.
[5] D. Boyd, K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, Inf. Commun. Soc. 15 (5) (2012) 662–679.
[6] D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of google flu: traps in big data analysis, Science 343 (14 March) (2014).
[7] F.O. Ostermann, C. Granell, Advancing science with VGI: Reproducibility and replicability of recent studies using VGI, Trans. GIS (2015).
[8] M. McNutt, Journals unite for reproducibility, Science 346 (6210) (2014) http://dx.doi.org/10.1126/science.aaa1724, 679–679.
[9] A. González-Beltrán, P. Li, J. Zhao, M.S. Avila-Garcia, M. Roos, M. Thompson, E. van der Horst, R. Kaliyaperumal, R. Luo, T.-L. Lee, et al., From peer-reviewed to peer-reproduced in scholarly publishing: The complementary roles of data models and workflows in bioinformatics, PLoS ONE 10 (7) (2015) e0127612.

[10] C. Drummond, Replicability is not reproducibility: Nor is it good science, in: Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, National Research Council of Canada, 2009.
[11] P.N. Edwards, M.S. Mayernik, A.L. Batcheller, G.C. Bowker, C.L. Borgman, Science friction: Data, metadata, and collaboration, Soc. Stud. Sci. 41 (5) (2011) 667–690.
[12] C. Lagoze, Big data, data integrity, and the fracturing of the control zone, Big Data Soc. 1 (2) (2014) 1–11.
[13] A. Voinov, N. Kolagani, M.K. McCall, P.D. Glynn, M.E. Kragt, F.O. Ostermann, S.A. Pierce, P. Ramu, Modelling with stakeholders Next generation, Environ. Modell. Softw. 77 (2016) 196–220.
[14] K. Crawford, M. Kranzberg, G. Bowker, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, Inf. Commun. Soc. 15 (5) (2012) 662–679.
[15] K. Janowicz, F. Van Harmelen, J.A. Hendler, P. Hitzler, Why the data train needs semantic rails, AI Mag. 36 (1) (2014).
[16] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, D. Ayers, Scovo: Using statistics on the web of data, in: The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009 Heraklion, Crete, Greece, May 31–June 4, 2009, Proceedings, Springer, 2009, pp. 708–722.
[17] P. Ristoski, H. Paulheim, Semantic web in data mining and knowledge discovery: A comprehensive survey, Web Semant. Sci. Serv. Agents World Wide Web 36 (2016) 1–22.
[18] C. Stasch, S. Scheider, E. Pebesma, W. Kuhn, Meaningful spatial prediction and aggregation, Environ. Modell. Softw. 51 (2014) 149–165.
[19] G.E. Box, Science and statistics, J. Amer. Statist. Assoc. 71 (356) (1976) 791–799.
[20] D.M. Blei, Build, compute, critique, repeat: Data analysis with latent variable models, Ann. Rev. Stat. Appl. 1 (2014) 203–232.
[21] S. Senn, et al., You may believe you are a Bayesian but you are probably wrong, Ration. Market. Morals 2 (42) (2011).
[22] S. Scheider, W. Kuhn, How to talk to each other via computers: Semantic interoperability as conceptual imitation, in: Applications of Conceptual Spaces, Springer, 2015, pp. 97–122.
[23] K.R. Popper, Logik der Forschung, Vol. 4, JCB Mohr (Paul Siebeck), 1982.
[24] W. Quine, Two dogmas of empiricism, Phil. Rev. 60 (1951) 20–43.
[25] W.W. Bartley, Theories of demarcation between science and metaphysics, in: Problems in the Philosophy of Science, in: Studies in Logic and the Foundations of Mathematics, vol. 3, North-Holland, 1968, pp. 40–119.
[26] P.K. Feyerabend, How to be a Good Empiricist-A Plea for Tolerance in Matters Epistemological, Interscience Press, New York, 1963, pp. 3–39.
[27] P. Feyerabend, Against Method: Outline of an Anarchist Theory of Knowledge, Verso, 1993.
[28] S. Scheider, M. May, A method for inductive estimation of public transport traffic using spatial network characteristics, in: Proceedings of 10th AGILE International Conference on Geographic Information Science, 2007.
[29] N. Goodman, Fact, Fiction, and Forecast, Harvard University Press, 1983.
[30] D. Remsen, The use and limits of scientific names in biological informatics, ZooKeys 550 (2016) 207.
[31] P. Gupta, M. Gahegan, G. Dobbie, Adventures of categories: Modelling the evolution of categories during scientific investigation, in: D. Bailo, K.G. Jeffery, A. Spinuso, G. Fiameni (Eds.), 2015 IEEE 11th International Conference on e-Science (e-Science), IEEE, 2015, pp. 1–11.
[32] T. Mitchell, The need for biases in learning generalizations, Cbm-tr 5-110, Rutgers University, 1980.
[33] E. Sober, What is the problem of simplicity? in: A. Zellner, H. Keuzenkamp, M. McAleer (Eds.), Simplicity, Inference, and Modelling, Cambridge University Press, Cambridge, Massachusetts, 2002, pp. 13–32.
[34] G. Gigerenzer, H. Brighton, Homo heuristicus: Why biased minds make better inferences, Top. Cogn. Sci. 1 (1) (2009) 107–143.
[35] G. Gigerenzer, On the supposed evidence for libertarian paternalism, Rev. Phil. Psychol. (2015) 1–23.
[36] T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The elements of statistical learning: data mining, inference and prediction, Math. Intelligencer 27 (2) (2005) 83–85.
[37] M. May, D. Hecker, C. Korner, S. Scheider, D. Schulz, A vector-geometry based spatial knn-algorithm for traffic frequency predictions, in: IEEE International Conference on Data Mining Workshops, 2008, ICDMW'08, IEEE, 2008, pp. 442–447.
[38] M. May, S. Scheider, R. Rösler, D. Schulz, D. Hecker, Pedestrian flow prediction in extensive road networks using biased observational data, in: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2008, p. 67.
[39] F. Leisch, M. Eugster, T. Hothorn, Executable papers for the r community: The r2 platform for reproducible research, Procedia Comput. Sci. 4 (2011) 618–626.
[40] S. Scheider, A. Ballatore, Semantic typing of linked geoprocessing workflows, Int. J. Digit. Earth (2017) forthcoming.
[41] K. Hinsen, Computational science: shifting the focus from tools to models, F1000Research 3 (101). http://dx.doi.org/10.12688/f1000research.3978.2.
[42] W. Kuhn, A. Ballatore, Designing a language for spatial computing, in: F. Bacao, Y.M. Santos, M. Painho (Eds.), AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities, Springer International Publishing, 2015.
[43] S. Scheider, Grounding Geographic Information in Perceptual Operations, Vol. 244, IOS Press, 2012.
[44] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, F. Villa, An ontology for describing and synthesizing ecological observation data, Ecol. Inform. 2 (3) (2007) 279–296.
[45] A. Sheth, C. Henson, S.S. Sahoo, Semantic sensor web, IEEE Internet Comput. 12 (4) (2008) 78–83.
[46] C.A. Henson, J.K. Pschorr, A.P. Sheth, K. Thirunarayan, SemSOS: Semantic sensor observation service, in: Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems, IEEE Computer Society, 2009, pp. 44–53.
[47] J.S. Horsburgh, D.G. Tarboton, M. Piasecki, D.R. Maidment, I. Zaslavsky, D. Valentine, T. Whitenack, An integrated system for publishing environmental observations data, Environ. Model. Softw. 24 (8) (2009) 879–888.
[48] M.A. Regueiro, J.R. Viqueira, C. Stasch, J.A. Taboada, Semantic mediation of observation datasets through sensor observation services, Future Gener. Comput. Syst. 67 (2017) 47–56.
[49] S.J.D. Cox, Ontology for observations and sampling features, with alignments to existing models, Semant. Web 8 (3) (2016) 453–470.
[50] K. Janowicz, S. Schade, A. Bröring, C. Keßler, P. Maué, C. Stasch, Semantic enablement for spatial data infrastructures, Trans. GIS 14 (2) (2010) 111–129.
[51] O. Corcho, R. García-Castro, Five challenges for the semantic sensor web, Semant. Web 1 (1, 2) (2010) 121–125.
[52] X. Wang, H.J. Hamilton, Y. Bither, An Ontology-Based Approach to Data Cleaning, Regina: Department of Computer Science, University of Regina, 2005.
[53] C. Fürber, M. Hepp, Using semantic web resources for data quality management, in: P. Cimiano, H.S. Pinto (Eds.), Knowledge Engineering and Management by the Masses, 17th International Conference, EKAW 2010, Lisbon, Portugal, October 11–15, 2010, Proceedings, Springer, 2010, pp. 211–225.
[54] H. Paulheim, Exploiting linked open data as background knowledge in data mining, in: J. Völker, et al. (Eds.), Workshop on Data Mining on Linked Open Data, Vol. 1082, 2013.
[55] S. Scheider, M. Tomko, Knowing whether spatio-temporal analysis procedures are applicable to datasets, in: R. Ferrario, W. Kuhn (Eds.), 9th International Conference on Formal Ontology in Information Systems (FOIS 2016), IOS Press, 2016.
[56] W. Kuhn, Core concepts of spatial information for transdisciplinary research, Int. J. Geogr. Inf. Sci. 26 (12) (2012) 2267–2276.
[57] L. Spinsanti, F.O. Ostermann, Automated geographic context analysis for volunteered information, Appl. Geogr. 43 (2013) 36–44.
[58] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J.M. Gómez-Pérez, S. Bechhofer, G. Klyne, C. Goble, Using a suite of ontologies for preserving workflow-centric research objects, Web Semant. Sci. Serv. Agents World Wide Web 32 (2015) 16–42.
[59] P. Alper, K. Belhajjame, C. Goble, P. Karagoz, Small is beautiful: Summarizing scientific workflows using semantic annotations, in: 2013 IEEE International Congress on Big Data (BigData Congress), IEEE, 2013, pp. 318–325.
[60] T.M. Mitchell, Machine Learning, first ed., McGraw-Hill, Inc., New York, NY, USA, 1997.
[61] T.M. Mitchell, R.M. Keller, S.T. Kedar-Cabelli, Explanation-based generalization: A unifying view, Mach. Learn. 1 (1) (1986) 47–80.
[62] E. Davis, G. Marcus, Commonsense reasoning and commonsense knowledge in artificial intelligence, Commun. ACM 58 (9) (2015) 92–103.
[63] T.M. Mitchell, Does machine learning really work? AI Mag. 18 (3) (1997) 11.
[64] C. d'Amato, N. Fanizzi, M. Grobelnik, A. Lawrynowicz, V. Svatek, Inductive reasoning and machine learning for the semantic web, Semant. Web 5 (1) (2014) 3–4.
[65] G. Stumme, A. Hotho, B. Berendt, Semantic web mining: State of the art and future directions, Web Sem.: Sci. Serv. Agents World Wide Web 4 (2) (2006) 124–143.
[66] S. Capadisli, S. Auer, A.-C. Ngonga Ngomo, Linked sdmx data, Semant. Web 6 (2) (2015) 105–112.
[67] H.O. Nigro, Data Mining with Ontologies: Implementations, Findings, and Frameworks, IGI Global, 2007.
[68] V. Mulwad, T. Finin, Z. Syed, A. Joshi, Using linked data to interpret tables, in: O. Hartig, A. Harth, J. Sequeda (Eds.), Proceedings of the First International Conference on Consuming Linked Data, CEUR-WS. org, 2010, pp. 109–120.
[69] J.-U. Kietz, F. Serban, S. Fischer, A. Bernstein, "Semantics Inside!" but let's not tell the data miners: Intelligent support for data mining, in: V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, A. Tordai (Eds.), The Semantic Web: Trends and Challenges: 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25–29, 2014, Proceedings, Springer International Publishing, 2014, pp. 706–720.
[70] H. Paulheim, Generating possible interpretations for statistics from linked open data, in: E. Simperl, P. Cimiano, A. Polleres, O. Corcho, V. Presutti (Eds.), The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27–31, 2012, Proceedings, Springer, 2012, pp. 560–574.
[71] H. Rijgersberg, M. van Assem, J. Top, Ontology of units of measure and related concepts, Semant. Web 4 (1) (2013) 3–13.

[72] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, et al., The SSN ontology of the W3C semantic sensor network incubator group, Web Semant. Sci. Serv. Agents World Wide Web 17 (2012) 25–32.
[73] F. Probst, Ontological analysis of observations and measurements, in: M. Raubal, H.J. Miller, A.U. Frank, M.F. Goodchild (Eds.), Geographic Information Science, 4th International Conference, GIScience 2006, Münster, Germany, September 20–23, 2006, Proceedings, Springer, 2006, pp. 304–320.
[74] T.A. Russ, C. Ramakrishnan, E.H. Hovy, M. Bota, G.A. Burns, Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case, BMC Bioinformatics 12 (1) (2011) 1–15.
[75] K. Janowicz, S. Scheider, T. Pehle, G. Hart, Geospatial semantics and linked spatiotemporal data–past, present, and future, Semant. Web 3 (4) (2012) 321–332.
[76] D. Talia, Workflow systems for science: Concepts and tools, ISRN Softw. Eng. 2013 (2013) 1–15. http://dx.doi.org/10.1155/2013/404525.
[77] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E.A. Lee, J. Tao, Y. Zhao, Scientific workflow management and the kepler system, Concurr. Comput.: Pract. Exper. 18 (10) (2006) 1039–1065.
[78] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, Knime-the konstanz information miner: version 2.0 and beyond, AcM SIGKDD Explor. Newslett. 11 (1) (2009) 26–31.
[79] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, et al., Orange: data mining toolbox in python, J. Mach. Learn. Res. 14 (1) (2013) 2349–2353.
[80] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M.R. Pocock, A. Wipat, et al., Taverna: a tool for the composition and enactment of bioinformatics workflows, Bioinformatics 20 (17) (2004) 3045–3054.
[81] J. Freire, C.T. Silva, S.P. Callahan, E. Santos, C.E. Scheidegger, H.T. Vo, Managing rapidly-evolving scientific workflows, in: Provenance and Annotation of Data, Springer, 2006, pp. 10–18.
[82] E. Deelman, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, S. Patil, M.-H. Su, K. Vahi, M. Livny, Pegasus: Mapping scientific workflows onto the grid, in: M.D. Dikaiakos (Ed.), Grid Computing: Second European AcrossGrids Conference (AxGrids 2004), Springer, 2004, pp. 11–26.
[83] Y. Gil, V. Ratnakar, E. Deelman, G. Mehta, J. Kim, Wings for pegasus: Creating large-scale scientific applications using semantic representations of computational workflows, in: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, Vol. 22, 2007, p. 1767.
[84] J. de Jesus, P. Walker, M. Grant, S. Groom, WPS orchestration using the taverna workbench: The eScience approach, Comput. Geosci. 47 (2012) 75–86.
[85] P. Missier, K. Belhajjame, J. Cheney, The W3C PROV family of specifications for modelling provenance metadata, in: N.W. Paton (Ed.), Proceedings of the 16th International Conference on Extending Database Technology, ACM, 2013, pp. 773–776.
[86] M. Gahegan, B. Adams, Re-envisioning data description using Peirce's pragmatics, in: M. Duckham, E. Pebesma, K. Stewart, A.U. Frank (Eds.), Geographic Information Science, Springer, 2014, pp. 142–158.
[87] S. Scheider, B. Gräler, E. Pebesma, C. Stasch, Modeling spatiotemporal information generation, Int. J. Geogr. Inf. Sci. 30 (10) (2016) 1980–2008.
[88] M. Bishr, W. Kuhn, Trust and reputation models for quality assessment of human sensor observations, in: S.I. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, S. Bell (Eds.), Spatial Information Theory, 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12–16, 2015, Proceedings, Springer, 2013, pp. 53–73.
[89] E. Deelman, J. Moody, J. Kim, V. Ratnakar, Y. Gil, P.A. González-Calero, P. Groth, Wings: Intelligent workflow-based design of computational experiments, IEEE Intell. Syst. 1 (2011) 62–72.
[90] A. Gangemi, Ontology design patterns for semantic web content, in: Y. Gil, E. Motta, V.R. Benjamins, M.A. Musen (Eds.), The Semantic Web–ISWC 2005, Springer, 2005, pp. 262–276.

Simon Scheider is an assistant professor in geographic information science at the Department of Human Geography and Spatial Planning, university of Utrecht. His research lies at the interface between conceptual modeling, geographic data analysis and knowledge extraction. Before coming to Utrecht University, he was a post-doc at the Institute of Cartography and Geoinformation of ETH Zürich, Switzerland, a post-doc at the Institute of Geoinformatics in Münster, Germany, and a visiting scholar and research fellow at the Geography department of the University of California Santa Barbara, USA. He received his Ph.D. in Geoinformatics from the University of Münster in 2011, and worked for 6 years at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) in Sankt Augustin, Germany.

Frank O. Ostermann is an assistant professor for cloud and crowd geo-information processing at the Faculty for Geo-Information Science and Earth Observation, University of Twente. His main research interests are the opportunities and challenges of distributed collection, storage, processing and sharing of crowd-sourced and volunteered geographic information. Since 2009, he holds a Ph.D. (Dr. sc. nat.) in Geographic Information Science from the University of Zurich. Previous work includes three years as a post-doctoral researcher at the Joint Research Center of the European Commission, as well as several years as a research assistant at the University of Zurich and Hamburg on EU-funded projects on user-generated geographic content, and spatio-temporal data analysis and visualization in urban contexts.

Benjamin Adams is a research fellow at the Centre for eResearch at the University of Auckland. His research interests include knowledge representation, the semantic web, and computational models of place. He worked as a Scientific Observations Network postdoc on the semantic modeling of scientific observations to aid data integration and interoperability. He is motivated by the desire to develop semantic technology solutions that help domain scientists better share and discover data relevant to their research interests. In addition to traditional knowledge modeling, he uses bottom-up approaches, including natural language processing, to populate ontologies with terms and concepts that are used by working scientists.