AAAI Fall Symposium on "Discovery Informatics: AI Takes a Science-Centered View on Big Data," Arlington, VA, November 15–17, 2013.

Capturing Data Analytics Expertise with Visualizations in Workflows

David Kale, Yan Liu
Department of Computer Science
University of Southern California
941 Bloom Walk
Los Angeles, CA 90089
dkale@usc.edu, yanliu.cs@usc.edu

Samuel Di
Department of Computer Science
University of Mississippi
University Cir
University, MS 38677
sldi@go.olemiss.edu

Yolanda Gil
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
gil@isi.edu

Abstract

In the age of big data, data analytics expertise is increasingly valuable. This expertise includes not only formal knowledge, such as algorithms and statistics, but also practical skills that are learned through practice and are difficult to teach in classroom settings: management and preparation of data sets, feature design, and iterative exploratory analysis. Semantic workflows are a valuable tool for empowering non-expert users to carry out systematic analytics on large datasets using sophisticated machine learning methods captured in the workflows and their semantic constraints. In this paper we motivate and illustrate the role of visualizations in the usability of workflows by non-experts as well as their role in learning practical data analytics skills to gain interesting insights into data and methods. This capability is particularly important when confronting large datasets, where the selection of appropriate methods and their configuration with the best parameter and algorithm selections can be crucial in obtaining useful results.

Introduction

Data is increasingly and widely available to study complex phenomena in science, industry, health, and many other areas of societal importance. A multitude of data sources await analysis and mining, but a number of barriers stand in the way. Consider the survey results shown in **Figure 1**: most scientists do not have the expertise they know is required, and typically resort to collaborators for developing joint approaches to analyze the data they already have [Science 2011]. These are the results from a survey of Science reviewers, who represent the elite of the scientific data analysis community. Limited access to data analytics expertise is hampering even their ability to extract useful knowledge from the data that is already available and that will only grow in coming years.

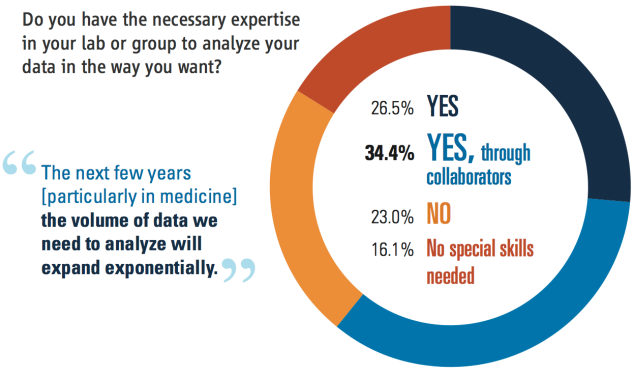


Figure 1. In 2011, *Science* reported that the majority of research labs do not have the expertise required to analyze their data [Science 2011]. The actual numbers are likely much worse, as the survey was done among its exceptionally qualified reviewers.

Oftentimes, data represents complex interrelated phenomena, requiring sophisticated knowledge about intricacies of data analytics algorithms and methods as well as their scalability. Therefore, learning advanced data analytics skills cannot easily be done from books or in a classroom setting. Many courses on different levels of statistics, machine learning and data mining are offered in most universities and train users on the relative merits of different algorithms and statistical techniques. Machine learning is one of the most popular topics in on-line courses. But a course on machine learning or data mining is not sufficient to prepare students to perform data analytics in practice, which often involves a significant amount of intuition and skill that can be developed only

with experience. This experiential learning requires understanding complex multi-step methods that often include key data preparation steps that must be customized to the data and task at hand. Visualizations can play a key role in this understanding process, but can be just as inaccessible to end users as the methods themselves. End users might not know what visualization software is appropriate to use to understand the effects of different aspects of complex methods. It is no wonder, then, that the use of sophisticated data analytics and the learning of such skills remain largely inaccessible to the people who have expertise on particular kinds of data and want to analyze it.

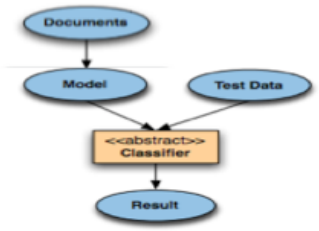
In this paper, we motivate the need to capture data analytics expertise particularly in this era of big data and the challenges many people face when trying to acquire it. We then discuss how to encapsulate such expertise in semantic workflow systems and then walk through a concrete example of such a system, combined with visualization tools, can facilitate both task-driven and exploratory analysis.

Acquiring Practical Data Analytic Skills

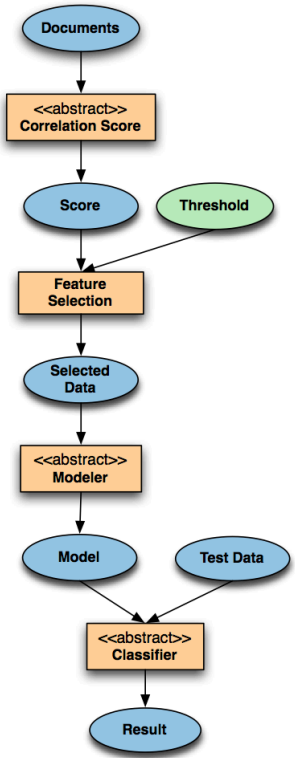
The era of "big data" poses significant new challenges for data analytics. Techniques that work for small data sets may not transfer well to data sets with increased volume, variety, and velocity of collection. Practitioners must master high performance computing techniques such as parallelization. In addition, data pre-processing steps (e.g., quality control, outlier detection, etc.) have increased importance, and so effective data scientists require comprehensive training in not only statistics and algorithms but also end-to-end analytic methods. For example, a commonly held piece of wisdom among machine learning practitioners is that the choice of good features plays a far more influential role than the choice of classifiers or statistical models (a fact that is often downplayed or ignored in both education and research). In both text and image classification, there are many cases in which the difference in error rates between two classifiers (e.g., a support vector machine and Naïve Bayes) may depend almost entirely on the choice of data representation (binary vs. word count vs. TF-IDF, unit normalized, unit variance, PCA whitened, etc.). Thus, designing an appropriate end-to-end data preparation process is critical to successful data analytics. Another example is the use by experts of visualization tools to understand a data set and to diagnose whether a given method is working and refine it. This process, which is both interactive and iterative, often provides as much insight into a given problem as does the final analysis itself. Additionally, state-of-the-art data analytics often involves multi-step methods with sophisticated statistical techniques such as cross-validation and combinations of algorithms such as ensemble methods. These techniques are challenging to set up and run and few users will have requisite experience and infrastructure to experiment with them. Finally, large data sets often introduce a number of related but independent engineering challenges. Naïve implementations of learning algorithms often prove prohibitively slow, and so analysis of large data sets necessitates significant optimization. And the largest data sets may reside in distributed shared-nothing storage architectures, requiring algorithms that can run in parallel. Analytics in these scenarios often requires sophisticated knowledge of algorithms, computational optimizations, and even parallel execution. Thus, modern data analytics may require specialized knowledge spanning fields as diverse as statistics, psychology, algorithms, and hardware, as well as substantial practical experience and know-how.

In both research projects and industrial practice, advanced data analytic skills are usually learned through both experience working on real-world projects and being mentored by data analytics experts. This creates a "chicken-and-egg" problem for end users who are interested in acquiring these skills to analyze their own data but do not have access to such settings. Assembling an appropriate framework for solving any real problem (e.g., email prioritization) requires a solid understanding of the state-of-the art analytics required (e.g., text analytics), creating a barrier for many end users who do not have easy access to that requisite expertise. Such frameworks often involve setting up significant infrastructure even when reusing widely available open source software, requiring significant investment in terms of time and resources and deterring many potential users. In addition, setting up this infrastructure often requires programming skills, making it infeasible for users without the right background. Developing an appropriate infrastructure for any real domain may take as long as months or years, making it impractical for users who want to rapidly acquire these skills for one or more domains.

Currently there are no courses or training programs offered that systematically teach students key data analytics skills, especially those informal ones concerned with pre-analytic processes, such as collecting high-quality data sets for training statistical models, processing data so that noisy information can be removed and effective representations can be extracted, and selecting appropriate statistical models for real data sets. Most researchers and practitioners acquire the necessary skills through years of experience. Many students do not recognize the importance of a systematic data analysis process until they graduate and face a real-world analytic task. Naïve approaches often do not lead to.



(a) Naïve document classification workflow



(b) Expert document classification workflow

Figure 2. Contrasting naïve and expert approaches to data analytics: (a) a naïve approach to classifying documents, (b) an expert approach to classify documents includes important data pre-processing steps.

Figure 2 contrasts naïve and expert approaches to data analytics using document classification as an example. **Figure 2(a)** shows a simple approach where a modeler is used on the training data as is to produce a classifier that is then applied to the test data. **Figure 2(b)** shows a more realistic approach to document classification that an expert would follow, in which a number of preprocessing (e.g., term weighting) and feature selection (based on correlation scoring) are done to create a more effective initial dataset for learning. Finally a model is trained to classify test data.

Semantic Workflows to Capture Data Analytics Expertise

Workflows represent complex applications as a dependency network of individual computations linked through control or data flow. The use of workflows for representing and managing complex scientific data analysis processes has been described in [Taylor et al 2006; Gil et al 2007ab; Gil 2009ab]. Workflow repositories have been created to enable sharing and reuse of workflows [De Roure et al 2009], but sharing is effective only when potential users understand the shared workflows and target problems. Without meaningful annotation, inexperienced users will have difficulty utilizing them.

To address this, we use *semantic workflows* [Gil 2009ab]. Semantic workflows capture end-to-end expert-grade analytic methods as computational workflows together with use constraints about the data and the algorithms used. The workflow system uses these constraints to ensure that the workflow is used correctly by non-experts. Semantic workflows can then be disseminated and reused by end users with no data analytics expertise.

Figure 3 shows a semantic workflow, designed to preprocess documents for document classification following the expert method shown in **Figure 2(b)**. This workflow is part of a library designed for text analytics [Hauder et al 2011ab] that we implemented in the WINGS semantic workflow system [Gil et al 2011a; Gil et al 2011b]. This workflow includes a number of data preprocessing and feature engineering steps before training the classifier. The first step removes *stop words*, extremely common words present in many documents. The two inputs to the *Stop Words* filter are the raw data and a list of stop words to be removed. This output is then passed to the *Small Words* filter, which removes words shorter than some minimum length (controlled by an input parameter). During the *Stemmer* step, a stemmer is applied in order to decrease superfluous variation in the vocabulary by removing plural endings, tenses, etc. [Porter 1980]. The *TermWeighting* step converts each document into a *vector space* representation, for example based on term frequency-inverse document frequency (TF-IDF) which counts the per-document occurrences of each word and then multiply that times the inverse of the fraction of documents that word occurs in, emphasizing rare words. After this step, each document is now represented as a list

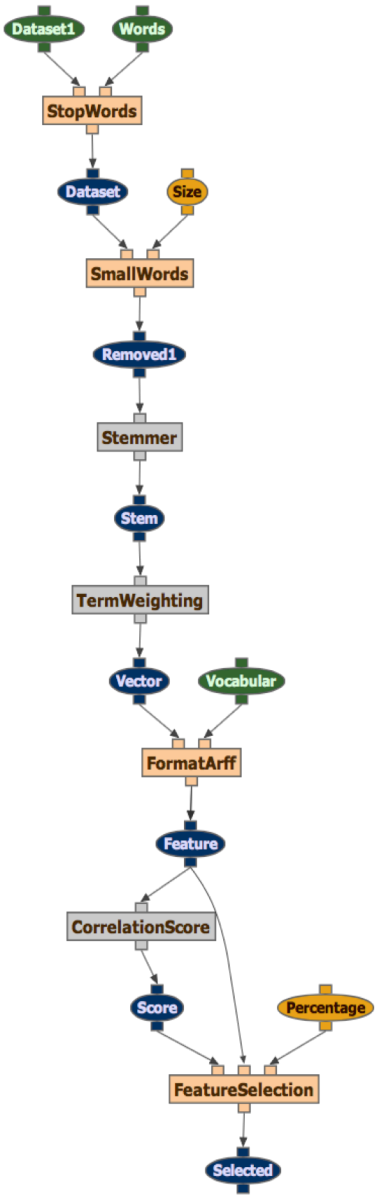


Figure 3. A portion of a WINGS workflow for document classification, focused on the data pre-processing and preparation steps.

of unique words and respective scores. The final preprocessing step, *FormatArff*, does not alter the data but merely re-formats it into a compact representation. It requires a fixed vocabulary as an additional input. Finally, the *FeatureSelection* component reduces the data dimensionality by keeping only the features with the highest correlation scores, calculated in a prior step.

Semantic workflows exhibit several important characteristics. First, they can capture expertise expressed as constraints. For example, an expert would know that decision tree modelers can only be trained with discrete data sets, so that is a constraint on their input training data. Second, semantic workflows can include steps that represent classes of components, and can be specialized to specific component implementations. For example, in the workflow of shown in Figure 3, there is a step labeled *CorrelationScore*. This represents a class of components including *ChiSquared*, *InformationGain*, and *MutualInformation* as specific techniques for calculating a "correlation score" for features. The user, or the system automatically upon request, can select an algorithm appropriate for the data. Third, semantic workflows can include rules that use data set constraints to assist non-experts with parameter selection. For example, a rule of the *FeatureSelection* step could be to use the top 60% of the features with the highest correlation scores. Finally, workflow templates can also abstract away computational details such as parallel processing of data collections, so end users do not have to worry about scalability [Gil et al 2009]. Properly designed workflows can capture and systematize valuable data analytics expertise and make it accessible to non-experts. High-school students and other non-experts have used the WINGS semantic workflow system successfully to perform a variety of analytics tasks [Hauder et al 2011b].

Visualization can often reveal patterns in data that defy more traditional statistical methods. One of the most famous examples is "Anscombe's quartet," a set of four data sets whose standard statistical properties (mean, variance, regression line, etc.) are virtually indistinguishable but which are clearly different when graphed [Anscombe 1973]. Experts routinely use visualizations throughout the different pre-processing steps and core analytic algorithms to understand how the method is working with the dataset at hand.

The use of semantic workflows can help disseminate expertise to end users, particularly with respect to visualizations:

- Workflows capture multi-step methods that include data preparation as well as visualization steps. Experts can create workflows based on their knowledge about how to prepare data. They know how to use visualizations effectively to understand whether a complex method is working for the data at hand, so these visualization steps can be included in the workflows. End users can simply reuse these workflows for their own data.
- Workflows allow end users that are not data analytics experts to experiment with different variations of those methods and adapt them to their own data, for example to try different algorithms and parameters. The expert use of visualizations greatly facilitates this exploration.

- Workflows encapsulate heterogeneous code into a simple dataflow paradigm that is more accessible for end users with no computer science or programming background. They do now have to download and install software packages, or figure out how to set up complex visualization tools to understand results that typically are generated in complex formats.
- Workflows make high-end computing techniques such as parallelization and distributed computation accessible to non-experts. This enables scientists to analyze data at scale while being isolated from system-level concerns. Visualization tools often require significant computing resources, which can make them even more inaccessible to end users.

Visualizations can be included in a workflow to show final results of the analysis, which means that visualization steps may be included as the last steps in a workflow. More importantly, visualization steps can be included to show intermediate data as well. **Figure 4** illustrates the use of visualization steps throughout a workflow for document clustering, generating visualizations for both intermediate and final results. These visualizations enable end users to understand how the data is being pre-processed and prepared for the core analytic algorithms, as we will see in the next section.

Exploring a Data Analytics Method through Workflow Visualizations

To illustrate how visualization expertise can be conveyed through workflows, we walk through the text classification workflow from **Figure 3** and discuss visualizations of the output from several intermediate steps. This allows us to gain insight into the behavior and impact of each data preparation or feature-processing step in the workflow, as well as demonstrate some of the informal and visual analysis that data scientists often perform. The impacts of different workflow steps on classification performance are discussed in detail in [Hauder et al 2011ab].

We use the *Termite* visualization system [Chuang et al 2012] to examine the *topics* discovered in intermediate data products by *latent Dirichlet allocation* (LDA) [Blei et al 2003]. LDA is a generative mixture model in which each topic places a different probability distribution over words and can be thought of as an unsupervised method for discovering groups of words that frequently co-occur in clusters of documents. Thus, it can provide useful insights into the structure of text data, which in turn helps us to understand how each step of the workflow impacts our

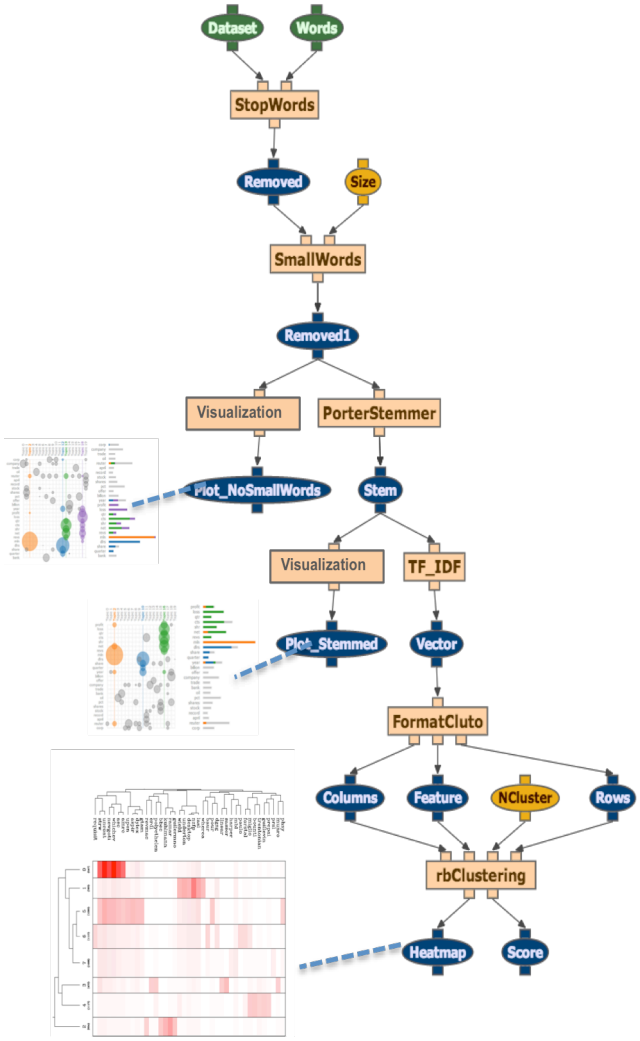


Figure 4. A WINGS workflow for document clustering that includes visualization steps for intermediate data as well as for the results of the analysis.

data and task. Only an expert would know of the existence of Termite, or think of using this kind of visualization for this particular kind of set of pre-processing steps.

Termite's main visual component is the term-topic matrix, which enables efficient visual comparison topics by encoding per-topic word probabilities using circle area. A configurable subset (between 10 and 250 words) of the vocabulary are included in the matrix, usually based on probability or *salience*, a novel measure of the word's distinctiveness similar to the term frequency-inverse document frequency (TF-IDF) score. The default setting, which compares topics based on probabilities assigned to the top 25 most "salient" terms, suits our purposes.

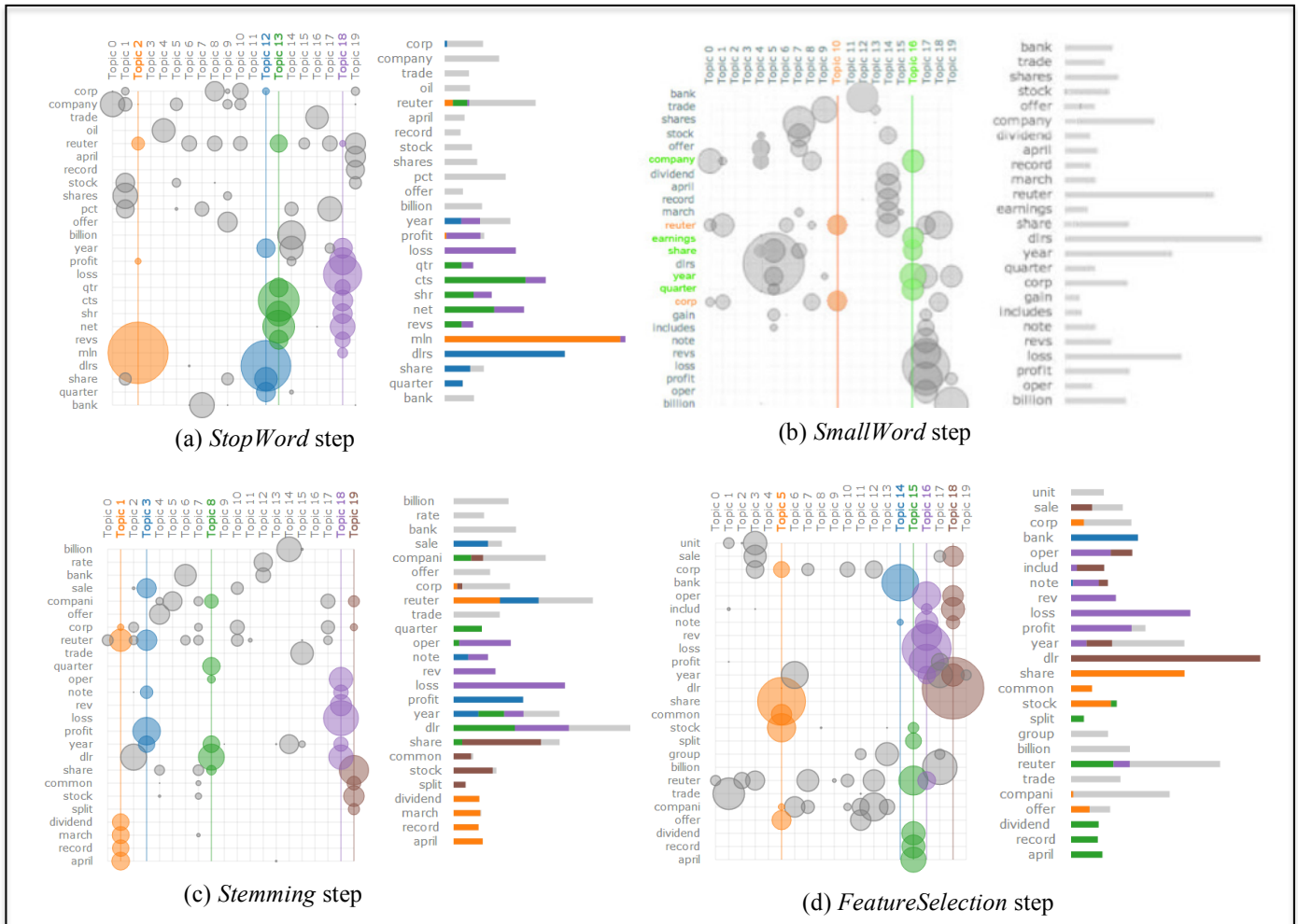


Figure 5. Termite visualizations of topic models learned from intermediate data products after (a) stop word removal, (b) removal of short words, (c) Porter stemming, and (d) elimination of words with the bottom 50% correlation scores.

For this work, we use the *Reuters R8* data set [UCI 1999], which includes 5,845 documents and a vocabulary of 19,982 distinct terms. Removing stop words and then "small" words (length less than four characters) reduces the vocabulary to 19,868 and 17,934 words respectively, while stemming reduces it to 13,221 distinct words. Finally, selecting only the features with the top 50% correlation scores further reduces the vocabulary to 6,579 terms. We used Termite to construct and visualize an LDA topic model after the *StopWord*, *SmallWord*, *Stemming*, and *FeatureSelection* steps. We then compare the lists of most salient words, as well as per-topic word probabilities, to determine the significance and impact of each stage.

Figure 5(a)-(d) show these visualizations. Comparing the output of independent runs of Termite can be tricky since the learned topics are not guaranteed to be the same or even related in any meaningful way. Nonetheless, we can still derive some useful insights by examining these visualizations.

Figure 5(a) shows the salient topics in the dataset after removing stop words. Compared to the topics of the original dataset (not shown here), removing stop words does not drastically alter the results for this particular dataset. The top 25 most salient terms remain the same as with the original text, and while the topics are not identical, their structure is quite similar. In both cases, *mln* features prominently in Topic 2 along with *profit* and *reuter* to a lesser extent. After the *StopWord* step, however, this topic also includes *net* and *year*. In both cases, there is a *dlrs* + *share* + *quarter* + *year* topic: Topic 12 in the raw data, which also includes *corp*, and Topic 10 in the *StopWord* output. In the *StopWord* output, there is an interesting *profit* + *loss* + *qtr* + *cts* + *shr* + *net* + *revs* topic (18); the raw data has two topics that overlap with it. This suggests that both LDA and the salience measure already disregarded stop words.

Figure 5(b) shows a visualization after removal of all words of length less than four (the *SmallWord* component).

This naturally eliminates seven of the top 25 most salient words, replacing them with longer words. Topic 2 of 5(a) disappears, though there is a similar *profit* + *reuter* topic (17) that is also most strongly associated with *loss* + *revs* + *oper*, etc. (Topic 18in 5(a)). There is once again a *dlrs* + *share* + *quarter* + *year* topic (5) that now also includes *gain* and *includes*. Topic 16 from 5(a) does not reappear; instead we see in 5(b) two new topics: *dividend* + *april* + *record* + *march* + *reuter* + *stock* (Topic 14) and *shares* + *stock* + *offer* + *share* + *march* (Topic 7).

By comparing the visualizations in 5(a) and 5(b), an end user could see the effect of removing short words. Quantitatively speaking, that step reduced the vocabulary by 2,000 words, some highly salient, so it is not surprising that this made a large impact in cleaning up the data to highlight useful features.

Figure 5(c) shows the visualization after stemming. We see a modest change (five words) in the top salient words. One obvious thing to notice is that *shares* and *share* are now combined, and many plural words have been replaced by their singular version (e.g., *dlr*). More interesting, however, is the relatively radical change in topics. Although Topics 1 (*dividend* + *april* + *record* + *march* + *reuter*, etc.) and 18 (*loss* + *revs* + *oper*, etc.) are quite similar to Topics 17 and 14 from before (respectively), the rest are different. There is also greater overlap between *Stemming* topics: *dlr* is equally likely in three topics, and *reuter*, *compani* (a weird stemming artifact), and *year* are also show up in several topics. It is possible that stemming has somehow merged the distinct roles previously played by *dlr* and *dlrs*. On the other hand, *share* demonstrates the opposite effect: the new Topic 19 assigns *share* the same amount of likelihood that the similar looking *StopWord* Topic 7 in 5(a) assigned to *share* and *shares* combined.

Figure 5(d) visualizes the data following the *FeatureSelection* step. Most interestingly, we see minimal change in both our top salient words list and topics. Only four words (including *march* and *quarter*) disappear from our salient list (replaced by things like *unit* and *rate*). The *Stemming* Topic 19 (*share* + *common* + *stock* + *split*, etc.) is very similar to the new Topic 5 (*share* + *common* + *stock* + *offer*, etc.). Topic 18 (loss + *rev* + *note* + *oper*, etc.) and Topic 16 (*loss* + *rev* + *note* + *oper* + *profit*, etc.) in 5(c) share many terms. Topic 1 in 5(c) (*reuter* + *dividend* + *april* + *record* + *march*, etc.) is very similar to new Topic 15 (same words minus *march*). However, several more minor topics (such as Topic 3 in 5(c)) disappear altogether. Also, the "popular word" phenomenon is reduced somewhat, particularly for *dlr*. All in all, the feature selection step appears to have minor impact on the topical structure showed by the visualizations, suggesting that the correlation scores do reasonably good job of capturing the topic structure and saliency discovered by LDA and Termite. If these features are retained for constructing a classification model downstream, it appears that they should prove discriminative, so long as the topic structure in the data is associated with the actual classes.

These visualizations illustrate the kinds of understanding that is difficult to teach in formal settings and often must be learned through experimentation and consultation with experts. Workflows can automatically provide such visualizations, and enable efficient and frictionless experimentation for end users to learn whether a method is useful for their data and how to customize its parameters and algorithms appropriately

We note that the visualization steps shown here are significantly more expensive computationally than calculating correlation scores through correlation scoring (e.g., Chi Squared), and they seem to capture some of the same structure in the data. They are extremely useful for understanding and exploring the method, but at performance time they should be eliminated or else they could become a bottleneck for delivering results in time.

Target Users

Our target end users include not just students, but also researchers and practitioners who intend to use data analytics in industry or scientific research. Anyone can easily find courses and training materials to be familiar with basic machine learning and statistical data analysis techniques. The goal of our work is to supplement those materials with practical learning experiences.

Semantic workflows could significantly reduce the learning curve for novices in data analytics. Existing workflows could be used to start running experiments within minutes, while the usual cycle of implementing a process from raw input to final results can take anywhere from days to months, depending on the complexity of the problem and the skill of the scientist. In many computational biology applications, for example, it may take an inexperienced user several months to implement a basic protein secondary structure prediction framework that consists of sequence analysis, feature extraction, classification, and post-processing. With workflows, a user will be able to achieve this in several minutes. Our approach provides an effective solution to lower the barriers to learning advanced skills for data analytics.

Semantic workflows could also enable access to data analytics training experiences for students who have no computer science or programming background. For example, many students in statistics or bioinformatics end up being limited to painstakingly reformatting and preparing data by hand or using only what is available in end-user environments such as MATLAB. A workflow system can be delivered together with real-world data sets and an extensive list of already-packaged state-of-art data

analysis components, such as feature extraction, feature selection, classifiers, unsupervised learning algorithms, and visualization tools. This combination would enable non-programmers to experiment with this rich set of components by easily assembling them into end-to-end data analytic processes represented as workflows.

Semantic workflows could also be used by researchers that have developed initial data mining applications and are seeking to improve the performance of their application. A good example here is compiler optimization, where data mining is used to rapidly customize optimizations to new computer architectures that are released every few months. We found that most of the work in this area focuses on decade old methods, far from the state-of-the-art [Hall et al 08]. Lowering the cost of learning data analytics skills would enable compiler researchers to achieve new levels of performance. Similarly, sophisticated analytic skills are required to analyze the reams of data in mobile devices and other human-computer interfaces.

Finally, expert-level data analytics practitioners could also use semantic workflows to learn new techniques. Experts can read and be aware of the newest algorithms but may lack a practical means to obtain hands-on experience with them because of limited time and resources. Moreover, experts often reach a comfort zone with algorithms and techniques that they have experience with, and are reluctant to invest the effort to learn novel state-of-the-art methods. We envision sustaining learning as a long-term activity throughout a professional career to keep up with research innovations.

Conclusions

We have illustrated the role of visualization steps in workflows as a means to enable end users to understand not just analytic results but also intermediate results that give insight on whether and how a multi-step analytic method is working for a given dataset. Semantic workflows can effectively lower the barriers to end users of a variety of experience levels, allowing them to experiment and learn in practical settings. They also systematize common pre-analysis tasks (e.g., data processing) that are critical to successful analysis. Workflow visualization steps convey to novices the importance of these pre-processing steps to the success of the overall method. This encourages greater understanding of the key role of data preparation steps when analyzing real data.

Acknowledgements. We gratefully acknowledge support from the National Science Foundation with award ACI-1355475, and from the Defense Advanced Research Projects Agency (DARPA) with award FA8750-13-C-0016.

References

Anscombe, F. J. Graphs in Statistical Analysis. *American Statistician* 27 (1): 17–21, 1973.

Blei, D., Ng, A., and Jordan, M. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993-1022, 2003.

Chuang, J., Manning, C., and Heer, J. Termite: Visualization Techniques for Assessing Textual Topic Models. *Advanced Visual Interfaces*, 2012.

De Roure, D., Goble, C., and Stevens, R. The design and realizations of the myExperiment Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, 25, 2009.

Garijo, D. and Gil, Y. A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data. In *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science* (WORKS-11), 2011.

Gil, Y., Deelman, E., Ellisman, M. H., Fahringer, T., Fox, G., Gannon, D., Goble, C. A., Livny, M., Moreau, L., and Myers, J. Examining the Challenges of Scientific Workflows. *IEEE Computer* 40 (12): 24-32, 2007.

Gil, Y., Groth, P., Ratnakar, V., and Fritz, C. Expressive Reusable Workflow Templates. *Proceedings of the Fifth IEEE International Conference on e-Science*, 2009.

Gil, Y., Gonzalez-Calero, P., Kim, J., Moody, J., and V. Ratnakar. A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs. *Journal of Experimental and Theoretical Artificial Intelligence*, 23 (4), 2011.

Gil, Y., Ratnakar, V., Kim, J., Gonzalez-Calero, P. A., Groth, P., Moody, J., and E. Deelman. Wings: Intelligent Workflow-Based Design of Computational Experiments. *IEEE Intelligent Systems* 26 (1), 2011.

Hall, M., Gil, Y., and Lucas, R. Self-Configuring Applications for Heterogeneous Systems: Program Composition and Optimization Using Cognitive Techniques. *Proceedings of the IEEE*, 96, 2008.

Hauder, M., Gil, Y., and Liu, Y. A Framework for Efficient Text Analytics through Automatic Configuration and Customization of Scientific Workflows. *Proceedings of the Seventh IEEE International Conference on e-Science*, 2011.

Hauder, M., Gil, Y., Sethi, R., Liu, Y., and Jo, H. Making Data Analysis Expertise Broadly Accessible through Workflows. *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science (WORKS-11)*, 2011.

McCallum, A. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

Porter, M. An Algorithm for Suffix Stripping. *Program* 14 (3): 130–137, 1980.

Science Magazine Editorial: "Dealing with Data Science: Challenges and Opportunities". *Science* 331, 2011.

Taylor, I., Deelman, E., Gannon, D., and Shields, M. (Eds). *Workflows for e-Science*, Springer Verlag, 2007.

UCI, Reuters-21578 Text Categorization Collection, 1999. Available from http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html