Re-Using Workflow Fragments Across Multiple Data Domains

Ricky J. Sethi
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
Email: rsethi@isi.edu

Hyunjoon Jo
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
Email: hyunjoon@usc.edu

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

***Abstract*—In this paper, we demonstrate the ability to re-use workflow fragments in different data domains: from text analytics to image anlysis to video activity recognition. We highlight how the re-use of workflows allows scientists to link across disciplines and avail themselves of the benefits of inter-disciplinary research beyond their normal area of expertise. In addition, we present an in-depth study of a Social Media Analysis (SMA) task, wherein we show how the re-use of workflow fragments can extend a pre-existing, rudimentary analysis; we also examine how workflow fragments save time and effort in SMA while bringing together multiple areas of machine learning and computer vision.**

I. Introduction

Scientific workflows help computational research in many different disciplines by consolidating heterogenous codebases and programs written in many different languages [1], [2], [3], [4], [5], [6]. Such workflows, designed by domain experts in their own fields, may also be of great utility to scientists in other disciplines; in fact, sites like http://www.myexperiment.org spotlight the need for re-using and re-purposing scientific workflows [7], [8]. Although such sites help reproduce and re-use entire workflows, the adaptation of complete workflows can be daunting for experts in other disciplines who might need to re-purpose specific components of the workflows for new research purposes. Therefore, the ability to share components of workflows would allow scientists in different disciplines to compose applications that utilize the same functionality across very different domains of data.

An elegant solution to this research development problem is to utilize and share workflow fragments [6]. Workflow fragments are a coherent sub-workflow designed by a domain specialist. They have the potential to reduce workflow authoring time and improve quality of the final workflows by allowing re-use of established, validated workflows. Each workflow fragment, in fact, is a useful resource in its own right and allows for cross-fertilization across scientific domains [6].

In this paper, we utilize workflow fragments to demonstrate the ability to re-use workflows as a way to facilitate development and bridge expertise across disciplines. We introduce several workflow fragments for Text Analytics, Image Analysis, and Video Activity Recognition. We also export these workflow fragments and make them available to the research community. In addition, we examine case studies that highlight the re-usability of workflow fragments across multiple data domains, from video analysis to multimedia analysis, which involves both text and image analysis. In particular, we show how a pre-existing but incomplete Social Media Analysis task can be developed more rapidly and extended by simply re-using the workflow fragments we have already developed; this complete workflow can subsequently be deployed as a system on the web, accessible by end-users or researchers to conduct further analysis or reproduce results, as needed.

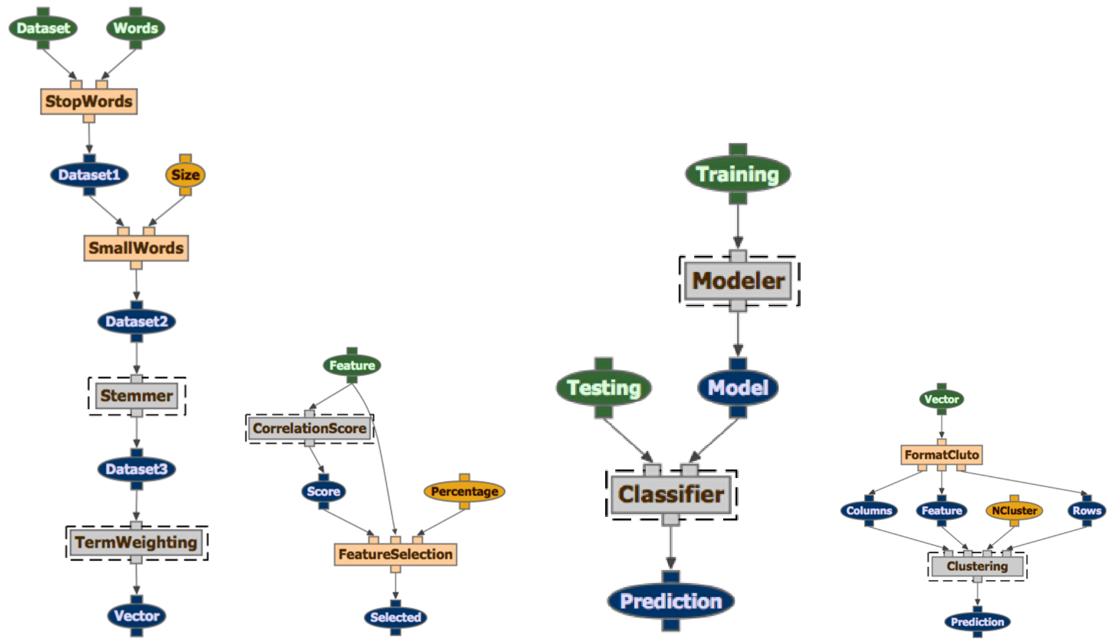
To facilitate this export of workflows and workflow fragments, we utilize the Wings workflow system, which was developed to assist scientists in managing complex computations [3] and has been used in several large-scale distributed scientific applications [5]. Wings uses semantic workflow representations that capture the requirements and semantic constraints of individual steps and datasets explicitly, as well as workflow reasoning algorithms to generate and validate possible combinations of workflow components systematically.

A. Contributions

Our main contributions in this paper are:

- Creation of various workflow fragments for Text Analytics, Image Analysis, and Video Activity Recognition which are exported and made available to the research community.
- Case studies that show the re-usability of workflow fragments across multiple data domains, including computer vision and machine learning applications for multimedia analysis.
- Analysis of development time and effort to both extend a nascent, rudimentary analysis of a Social Media Analysis project using the provided workflow fragments and to port its pre-existing code as new workflow fragments.

a) Overview of Article: The rest of this article is organized as follows: in Section II, we discuss the Wings workflow system. Then, in Section III, we show several workflow fragments we created for Text Analytics, Image Analysis, and Video Activity Recognition which incorporate vastly hetergenous codebases. In Section IV, we provide an in-depth case study of the re-use of workflow fragments to



(a) Feature Generation (b) Feature Selection (c) Classification. (d) Clustering.

Figure 1. Workflow Fragments previously developed [9] for Text Analytics. Here we see workflow fragments for a) Feature Generation; b) Feature Selection; c) Training and Classification; and d) Clustering.

extend a rudimentary Social Media Analysis task. We then demonstrate how the workflow fragments we created can be re-used to enable rapid development and deployment of several research projects in Section V. Finally, in Section VI, we highlight how the re-use of workflows allows scientists to link across disciplines, followed by a discussion of future directions in Section VII.

II. Wings Semantic Workflows

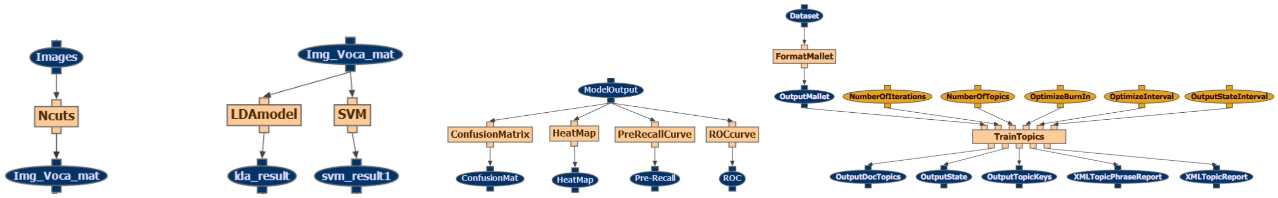
Our approach uses the Wings workflow system [10], which has three key features that make workflows accessible to users: a simple dataflow structure, an easy-to-use web interface, and an ability to export workflows and workflow fragments as web objects. This framework allows us to structure computer vision and machine learning tasks as computational workflow fragments described in high-level declarative notations and capable of processing large quantities of data that comes from multiple sources or files [3], [11]. Wings is open source, built upon open web standards from the World Wide Web Consortium (W3C), and is available at http://www.wings-workflows.org/.

Using a semantic workflow system like Wings to assist with the design of such computational experiments allows for creating structured, systematic experiments that can be automated, thus allowing anyone to re-purpose entire workflows or workflow fragments. In addition, the Wings workflow system has an open modular design and can be easily integrated with other existing workflow systems and execution frameworks to extend them with semantic reasoning capabilities.

In fact, the Wings workflow system is pre-equipped with several expert-quality workflows that represent a powerful set of analytic methods [12]. It includes workflow fragments for general machine learning packages like Weka [13], document clustering packages like CLUTO, etc. We extend these repositories by creating workflow fragments based on popular computer vision and machine learning packages like OpenCV [14], a standard computer vision library, and MALLET [15], a standard package for statistical processing and information extraction, as well as adding custom implementations of some state-of-the-art computer vision/machine learning models.

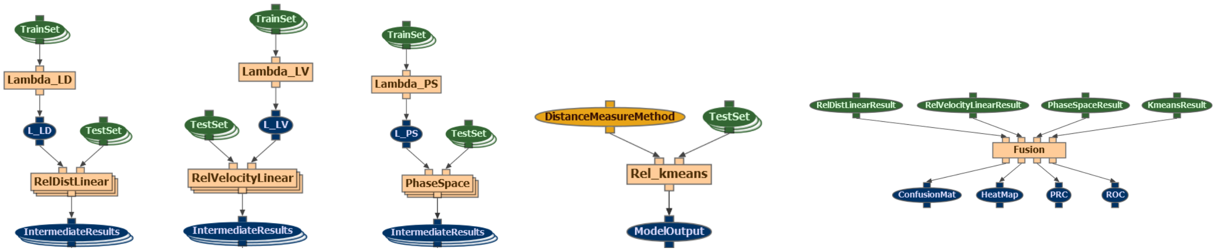
These packages have vastly heterogeneous implementations but the workflow fragments encapsulate the software with interfaces described by data types in the workflow system to make them reusable in different workflows. Wings ensures that only the right components are used in workflows by checking the semantic constraints of the input and output types for every component. The system ensures that only workflows with valid combinations of components are executed. The framework also includes several widely used datasets used for comparison purposes in the text analytics and computer vision community.

In addition, these workflow fragments can be exported in Wings by publishing them as web objects using Linked Data principles [16] and can be made available as part of a workflow library. These web objects, represented in RDF, allow direct access via unique URIs to workflow fragments or workflows, their components, and their associated datasets. Such web objects can then be imported into any workflow system that is compatible with the standard Open Provenance Model for workflow publication [16] so that other researchers can directly re-use or re-purpose any single workflow fragment or entire workflows.



(a) N-Cuts Image Segmentation (b) LDA and libSVM (c) Statistics Evaluation (d) Topic Models

Figure 2. Workflow Fragments for Image Analysis. Here we see workflow fragments for a) N-Cuts Image Segmentation; b) Latent Dirichlet Allocation (mallet) and Support Vector Machines (libSVM); c) Statistics Evaluation (Confusion Matrices, Heatmaps, Precision Recall Curves, and Receiver Operating Characteristic Curves); and d) Topic Modelling (mallet).



(a) RDL (b) RVL (c) PS (d) K-Means (e) FUSION: weighted average of all other methods

Figure 3. Workflow Fragments for five Video Activity Recognition (VAR) Models. Here we see workflow fragments for a) RDL: Relative Distance with Linear Fit VAR Model; b) RVL: Relative Velocity with Linear Fit VAR Model; c) PS: Phase Space (MOPA) as Relative Distance with Exponential Fit VAR Model; d) K-Means: Relative Distance with k-means clustering VAR Model; and e) FUSION: a weighted average of all other methods.

III. Workflow Fragments for Text Analytics, Image Analysis, and Video Activity Recognition

Workflows are usually composed of workflow fragments that are reused across workflows. Such predefined workflow fragments make complex analytics expertise readily available to new users. The components that make up workflow fragments can be written in heterogeneous languages: e.g., some components are in Java, others in matlab, and still others in C++ but the language of choice is irrelevant as the components are integrated into the workflows without reliance upon their individual implementation idiosyncrasies. This is possible because each individual program is converted into a workflow component via a short wrapper shell script (usually 3-5 lines of code) thus allowing any pre-existing program to be incorporated as a new component in a workflow or workflow fragment.

These previously defined workflow fragments can be executed independently from each other. This is helpful as some researchers might choose to focus on particular parts in order to optimize or improve their understanding of the behaviour in the individual steps. A good starting point for researchers in other disciplines, however, is to create end-to-end workflows that are formed by re-using and re-purposng workflow fragments. These end-to-end workflows would then incorporate and represent advanced expertise in that they would capture complex combinations of components that are known to work well in practice. Such re-usable workflow fragments are pre-defined by domain experts and available as part of workflow libraries. They can be executed with available datasets or adapted by adding or changing components.

A. *Workflow Fragments for Text Analytics*

Here, we detail some of the workflow fragments we have previously developed [9] for Text Analytics as seen in Figure 1.

1) Text Pre-Processing and Feature Generation: Analytic tasks usually begin with some preprocessing steps to generate the features of a document. The workflow fragment for feature generation is shown in Figure 1(a). Morphological variations are removed from the dataset with a stemmer component. The Wings workflow system provides several choices, including a Porter Stemmer and a Lovins Stemmer. It further provides term weighting components that is used to transform the dataset into the vector space model format. Among them are term frequency-inverse document frequency, corpus frequency or document frequency for instance. The generated outcome can now be used with different other workflows and is independent of a particular implementation at this stage in the workflows.

2) Feature Selection: A very common step for many classification problems is feature selection, as shown in Figure 1(b), whose main purpose is to reduce the training set by only using the most valuable features. This will reduce the necessary time for training the model and can improve the results of the classifier in some cases. The goodness of a feature in the dataset is measured with the correlation score. Typical implementations for this step are Chi Squared, Mutual Information or Information Gain that can be found in [17] and are all implemented in the framework. The resulting score is used in a feature selection step to retain the most valuable features in the training set. The percentage of selected features is typically changing for every dataset respectively classifier used in the computational experiment.

Another characteristic for this workflow fragment is that it uses heterogeneous implementations for the components. While the components for the computation of the correlation score take advantage of the capabilities of MATLAB to handle large matrices very elegant, the component for the feature selection uses an implementation written in Java.

3) Classification and Clustering: The resulting training set after the feature selection can be used for the training of a model with the workflow fragment shown in in Figure 1(c). Both components in the workflow use the Weka machine learning framework. Thus, many different machine learning algorithms can be used to perform experiments with the dataset. Among them are very popular algorithms from the text analytic community like Support Vector Machines, Naive Bayes or k-Nearest Neighbor. The computed model can be stored in the data catalog and reused for later classifications. Since the training is usually a very time demanding task in the workflows, it is very desirable to reuse previously created models. Existing models are also easier to compare against each other, because the metadata information of the model carries provenance information from the components used and their configuration during the workflow execution. In the second step a classifier uses the trained model with the testing set to compute the predictions.

In Figure 1(d), we see the workflow fragment for clustering. The Vector that results from the Feature Generation workflow fragment in Figure 1(d) can be used as input for clustering. It needs to be formatted into the suitable format for the clustering software. The result of this step is the Feature output with the transformed Vector. Next to this output there are additional intermediate files called Rows and Columns that contain the label names that are used to annotate the final result with the right names for the features and labels. The parameter for this component is used to specify the number of clusters to be applied on the data set.

B. *Workflow Fragments for Image Analysis*

Here, we detail some of the Workflow Fragments we have developed for Image Analysis as seen in Figure 2. In particular, we created workflow fragments for a) Normalized Cuts Image Segmentation [18] which views image segmentation as the optimal partitioning of a graph by minimizing the cut with a modified cost function; b) Latent Dirichlet Allocation (MALLET) [15] for visual-word clustering and Support Vector Machines (libSVM) for visual-word classification [19]; c) Statistics Evaluation (Confusion Matrices, Heatmaps, Precision Recall Curves, and Receiver Operating Characteristic Curves) [20], [21], [22]; and d) Topic Modelling (MALLET) [15] for video-word clustering. In particular, the statistics evaluation workflow fragment allows for easy visualization of diverse summary and graphical statistical measures which are the outputs of that component (i.e., summary measures like Equal Error Rate, Mean Average Precision, etc., as well as the graphical outputs of Confusion Matrices, Heatmaps, Precision Recall Curves, and Receiver Operating Characteristic Curves). Visual-words and video-words are the image and video equivalent of text words used in textual bag-of-words models; in computer vision, they are created by partitioning an image or video into interest point cuboids or segments and then computing some features (for which it is possible to calculate a distance metric) for each interest point cuboid. The centers of each of these clusters are the visual words (codewords) in the visual vocabulary (codebook).

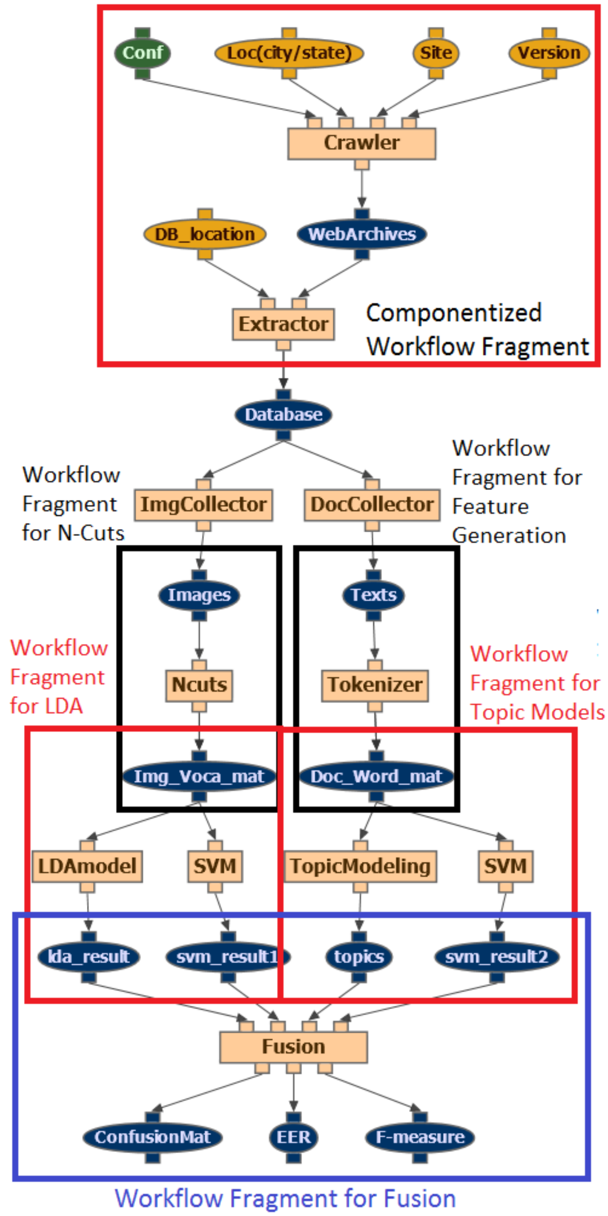


Figure 4. Fusion of Image and Text Analyses in the Social Media Analysis workflow.

C. Workflow Fragments for Video Activity Recognition

Here, we detail some of the Workflow Fragments we have developed for Video Activity Recognition (VAR) as seen in Figure 3. We implemented the most appropriate models for the ISI Atomic Actions datasets (which consists of 190 videos as explained in Section V), derived from [23], [24], [25], [26], [27], [28], [29], [30], as workflow fragments in the Wings workflows system; the workflow fragments corresponding to these five models for VAR are shown in detail in Figure 3 and are: a) RDL: Relative Distance with Linear Fit VAR Model; b) RVL: Relative Velocity with Linear Fit VAR Model; c) PS: Phase Space (MOPA) as Relative Distance with Exponential Fit VAR Model; d) K-Means: Relative Distance with k-means clustering VAR Model; and e) FUSION: a weighted average of all other methods, which outputs the graphical statistical measures Confusion Matrices, Heatmaps, Precision Recall Curves, and Receiver Operating Characteristic Curves, as well as summary statistical measures of Equal Error Rate, Mean Average Precision, etc.

IV. Case Study: Social Media Analysis

We demonstrate the utility of the Image Analysis and Text Analytics workflow fragments by extending a pre-existing text analysis for a Social Media Analysis (SMA) task that tries to detect human trafficking. This project analyzes posts on various sites on the internet in order to determine if the subject of that post is a victim of human trafficking. The ultimate goal of this project is to create intelligence which may be used by law enforcement to detect and combat trafficking by making a determination of whether or not the subject of a post was trafficked or not.

The initial development of the project had progressed to creating a crawler, which downloads posts from various posting sites, and an extractor, which extracts the text and images and stores them in a database. However, there had been no substantial analysis of the posts in this nascent project. We extended the project to examine both the text of the post (using the Text Analytics workflow fragments we had already developed), as well as the associated images (using the Image Analysis workflow fragments we developed); a final determination about trafficking of the subject of the post was made by fusing the results of the Text and Image analysis via the Fusion workflow fragment we developed. The goal of this project is to use both the text and image content of posts to make a stronger determination of whether or not the subject of the post was trafficked. Thus, the re-use and re-purposing of workflow fragments allowed a multimedia analysis spanning data domains of text and image analysis, including the fusion of their results in the final determination.

In particular, we componentized, re-used, and re-purposed the following workflow fragments in the SMA task:

- Componentized: Crawler and Extractor Workflow Fragment
- Re-used: N-Cuts, Feature Generation, LDA, and SVM Workflow Fragments from Figures 1 and 2.
- Re-purposed: Fusion Workflow Fragment from Figure 2.

We first componentized the previously developed crawler and extractor as workflow fragments using the Wings framework and then re-used/re-purposed our workflow fragments to create the final workflow for human trafficking detection. The resulting workflow is shown in Figure 4 where the top black box labelled "Componentized Workflow Fragment" shows the original crawler and extractor incorporated as components in Wings. This is followed by:

- Re-Use: The next two black boxes labelled "Workflow Fragment for N-Cuts" and "Workflow Fragment for Feature Generation" show re-use of the Image Analysis workflow fragments from Figure 2 as well as the re-use of the Text Analytics workflow fragments from Figure 1, respectively. Here, the "Tokenizer" component represents the entire workflow fragment in Figure 1(a).
- Re-Use: The next two red boxes labelled "Workflow Fragment for LDA" and "Workflow Fragment for Topic Models" show the re-use of workflow fragments for unsupervised analysis using MALLET and supervised analysis using SVM in a bag-of-words model from both the Image Analysis workflow fragments in Figure 2 and the Text Analytics workflow fragments in Figure 1. Here, the "TopicModeling" component represents the entire workflow fragment in Figure 2(d).
- Re-Purpose: The final blue box labelled "Workflow Fragment for Fusion" shows the re-purposed Fusion workflow fragment from Figure 3 for fusing the results of the Text and Image Analysis and visualizing those results.

Summary results from the Fusion module showed an Equal Error Rate of 0.37 and F-Measure of 0.47.

A. Analysis of Time/Work Savings

In this implementation, we incorporated the original crawler and extracter into Wings and then added on various Text Analysis and Image Analysis workflow fragments, including fusing their results and adding components to help visualize the results. This involved writing simple component wrapper scripts for both of the existing python scripts and setting up the mySQL database interface. The original development of the python version of the crawler/extracter had taken several months; this was quite involved as appropriate algorithms had to be researched, in addition to developing the code. The original crawler and extracter were componentized into Wings components, as shown in Figure 4.

This process took roughly two days as the original programs had to be made independent of the original development environment, account for supporting libraries, and had to interface with the external Database system that was distributed on the Web. Once this was done, the extension of the other components via workflow fragments for Image Analysis, Text Analysis, and their Fusion and visualization, took approximately one day, saving effort estimated to be on the order of 300 man-hours of work. This was estimated by the original developers using one postdoc and one graduate

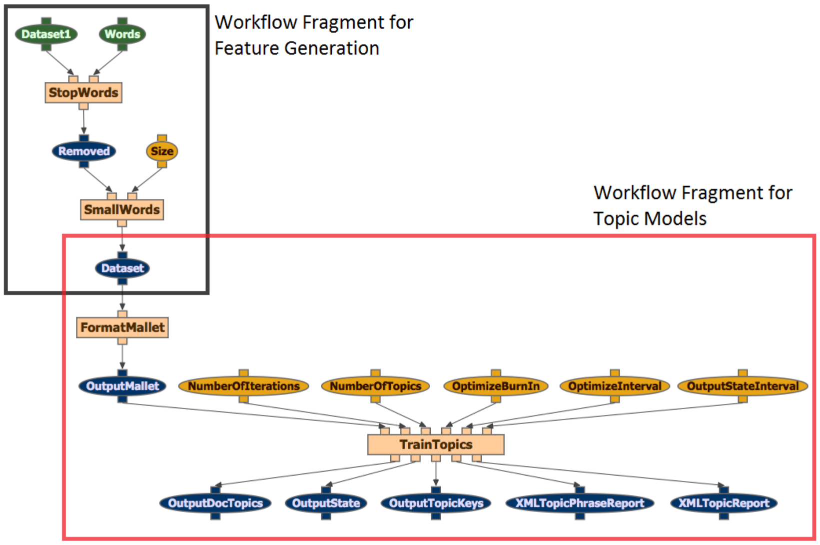


Figure 5. Topic Modeling Workflow.

student working at a similar pace as in the development of the original prototype as they worked to identify appropriate algorithms for image and text analysis, implement them and incorporate them into their nascent crawler/extracter prototype, and then investigated a fusion methodology as well as the tools to visualize and analyze the results.

V. Overview of Workflow Fragments in Text Analytics and Video Activity Recognition

In this section, we demonstrate how the workflow fragments we created can be re-used and re-purposed to enable rapid development and deployment of several research projects spanning such diverse data domains as analysis of the text questions and answers on the The Madsci Network to video activity recognition in the Atomic Actions dataset analysis.

The Madsci Network is an Ask-A-Scientist website [31] and the **Madsci Topic Modelling Workflow**, shown in Figure 5, re-uses the Feature Generation Workflow Fragment from Figure 1 and the Topic Model Workflow Fragment from Figure 2. The various parameters associated with MALLET, the topic modeling framework utilized here, as well as the various outputs, can all be easily specified, customized, and used in subsequent processing.

For example, we can easily take one of the MALLET outputs, the OutputDocTopics in Figure 5, which shows the distributions over topics for each document, and insert a Weka component to visualize it. This visualization is shown in Figure 6. This is the plot of a single question, and its distribution over topics, which clearly shows the dominance of a single topic in the distribution. Such plots intuitively reveal insights about the individual questions and about the overall dataset.

It is also relatively simple to analyze how questions and answers cluster together, using the clustering workflow fragments from Figure 1. The results are shown in Figure 7. We can use this workflow to show how documents and topics

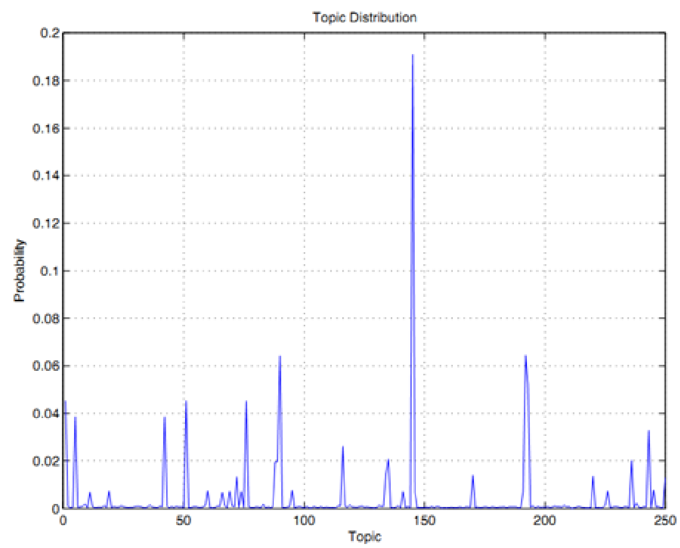


Figure 6. Topic distribution for a sample question on The Madsci Network.

cluster; this can be used by both the users and the moderator. When a new question is submitted, a new clustering diagram will be produced in which topics would be on the y-axis and documents on the x-axis; this would clearly show which questions/answers cluster with the correct answer (on the x-axis) for the user and which topics cluster together on the y-axis for the moderator to see which topics are most relevant for the new question.

The **Atomic Pair Actions Workflow** re-uses the Video Activity Recognition Workflow Fragments and Statistics Workflow Fragments from Figures 2 and 3. Similarly, the **Atomic Group Actions Workflow** re-purposes the Statistics Modules from Figure 2, in addition to implementing custom workflow

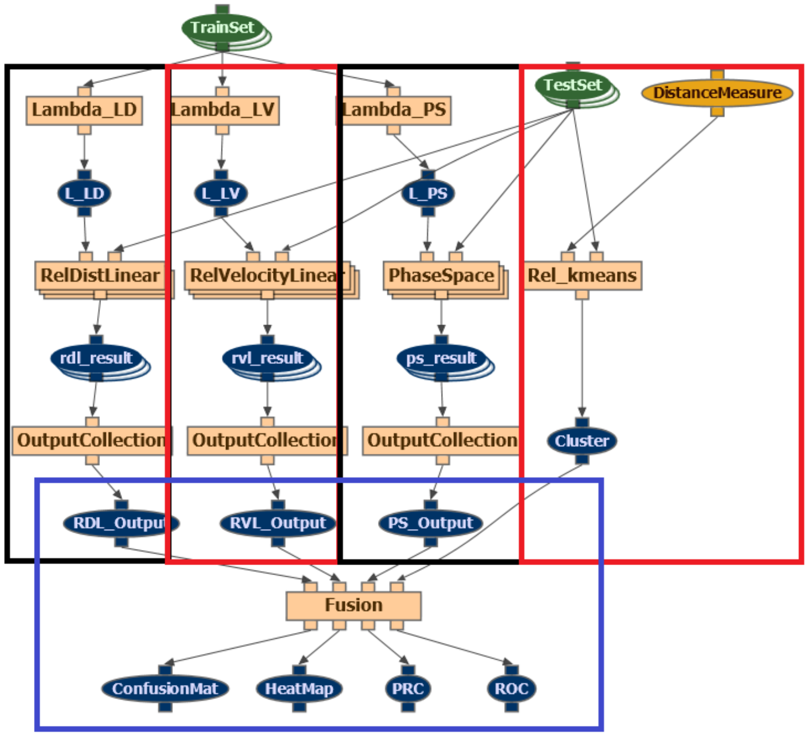


Figure 8. Overall Workflow for the Atomic Pair Actions Dataset Analysis. Some of the individual component workflow fragments, like the Video Activity Recognition models and Fusion module, that are re-used in this overall workflow are shown in Figure 3.

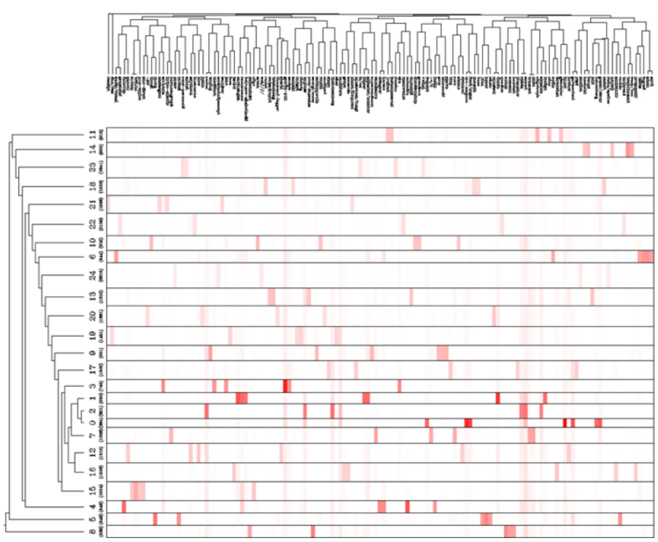


Figure 7. Clustering output for The Madsci Network dataset.

fragments for the Group Transition Ratio (G_{tr}) model [25], [26].

The ISI Atomic Pair and Group Actions Dataset contains videos obtained from a combination of YouTube or public domain datasets. It examines Atomic Pair Actions (converging, parallel, diverging) as well as Atomic Group Action (group formation, dispersal, and movement) [25], [26]. The entire dataset contains 190 videos and some sample frames are shown in Figure 10.

In Figure 8, we see the overall workflow for the Atomic Pair Actions Dataset analysis. Some of the individual component workflow fragments, like the Video Activity Recognition models and Fusion module, that are re-used in this overall workflow are shown in Figure 3. Then, in Figure 9, we see the workflow for calculating the Group Transition Ratio (G_{tr}) model, used for the Atomic Group Actions Dataset, which re-purposes the Statistics Evaluation workflow fragment from Figure 2.

In both of these workflows, we re-use or re-purpose the Statistics Evaluation workflow fragment from Figure 2 to calculate Heatmaps, PRC curves, ROC curves, Confusion Matrices, etc., as shown in Figure 11, in addition to various summary statistical measures like Equal Error Rate, F-Measure, Mean Average Precision, etc., as shown in Table I.

Here we see a variety of different levels of workflow fragment componentization, re-use, and re-purposing: in Figure 4, we componentize two components, re-use four different workflow fragments, and re-purpose one workflow fragment, all analyzing *image* and *text* data; finally, in Figure 8, we directly re-use five workflow fragments and, in Figure 9, we componentized four new components and re-purposed one workflow fragment, all for the analysis of *video* data.

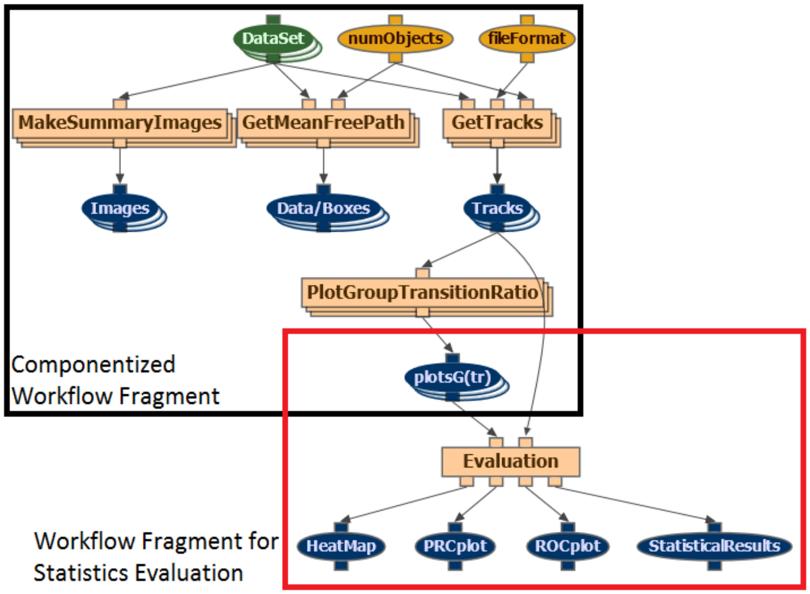


Figure 9. Workflow for calculating the Group Transition Ratio (G_{tr}) model, used for the Atomic Group Actions Dataset, which re-purposes the Statistics Evaluation workflow fragment from Figure 2.

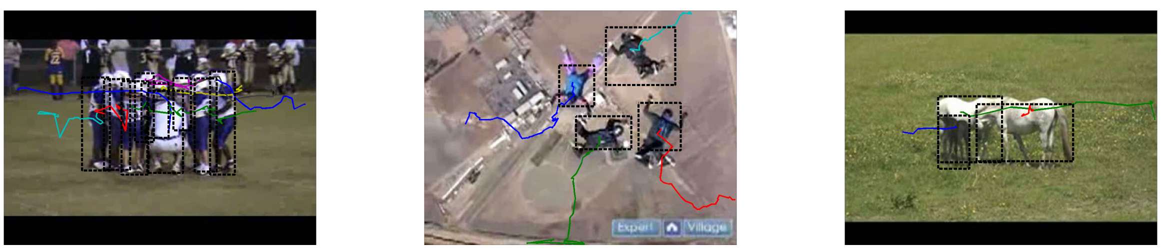
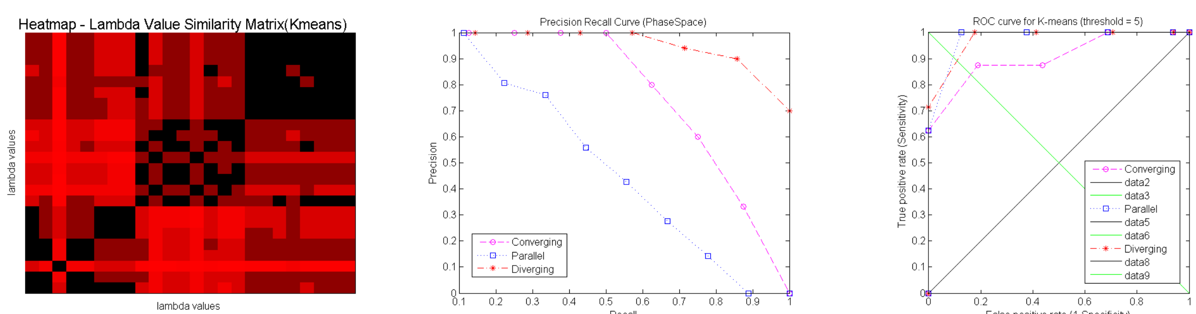


Figure 10. Sample frames from video clips from the ISI Atomic Actions datasets.



(a) K-Means Heatmap (b) PS PRC curve (c) K-Means ROC Curve

Figure 11. Results showing Heatmaps, PRC, ROC curves, and Confusion Matrices of the best performing Video Activity Recognition models on the ISI Atomic Actions datasets.

	F-Measure			Equal Error Rate			Average Precision			MAP
	Converging	Diverging	Parallel	Converging	Diverging	Parallel	Converging	Diverging	Parallel	
RDL	0.928571	0.823529	0.814815	0.235300	0.176500	0.235300	0.875000	0.961700	0.642500	0.826400
RVL	0.235294	0.320000	0.800000	0.956500	0.250000	0.600000	0.000000	0.150900	0.672500	0.274500
PS	0.846154	0.736842	0.666667	0.277800	0.315800	0.176500	0.750000	0.957100	0.422000	0.709700
K-Means	1.000000	0.933333	0.941176	0.187500	0.125000	0.176500	0.930600	0.861100	1.000000	0.930600
FUSION	0.222222	0.777778	0.476190	0.000000	0.235000	0.467000	0.638900	0.640300	0.738500	0.672600

Table I

CUMULATIVE STATISTICAL COMPARISON OF ALL FIVE VAR MODELS FROM SECTION III-C FOR THE ATOMIC PAIR ACTIONS. HERE WE SEE THE K-MEANS OUTPERFORMS IN THE F-MEASURE, EQUAL ERROR RATE (EER), AND AVERAGE PRECISION/MEAN AVERAGE PRECISION (AP/MAP). RVL IS THE WORST PERFORMER IN ALL CATEGORIES.

VI. LINKING ACROSS DISCIPLINES BY RE-USING WORKFLOWS

This kind of re-use and re-purposing of workflow fragments across different data domains can be generalized to other scientific fields and allows scientists to link across different disciplines [2], [5]. For example, in geosciences, researchers observe the surface of the Earth at criticial points and examine moisture levels from above and below. This, in turn, depends closely on weather models, models of soil, rain, etc. Thus, they also need to use approaches from many different disciplines to analyze data from multiple domains. Examples such as this also abound in Biology [5], particularly in Proteomics and Genomics.

VII. DISCUSSION

In this work, we only illustrate re-use of our workflow fragments by ourselves and our collaborators, not by third parties [6]. However, this paper does illustrate the potential for re-use of workflow fragments and, if they are shared with other researchers, more scientists can use such workflow fragments in their own applications instead of having to re-implement them or, even worse, forego such an analysis. One of the issues involved with sharing workflow fragments is the open question of how to describe them so they are re-usable by others. We intend to examine this in future work where one promising approach is to use a Component Ontology by function as an aspect [32]: i.e., being able to find workflow fragments according to a user query to search for a specific kind of component that is retrieved for the user; we can also find workflow fragments by typing about them in English [33].

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation under Grant #1019343 to the Computing Research Association for the CIFellows Project.

REFERENCES

[1] Y. Gil, V. Ratnakar, J. Kim, P. A. Gonzalez-Calero, P. Groth, J. Moody, and E. Deelman, "Wings: Intelligent workflow-based design of computational experiments," *IEEE Intelligent Systems*, vol. 26, no. 1, 2011.

[2] Y. Gil, P. Szekely, S. Villamizar, T. Harmon, V. Ratnakar, S. Gupta, M. Muslea, F. Silva, and C. Knoblock, "Mind your metadata: Exploiting semantics for configuration, adaptation, and provenance in scientific workflows," in *Proceedings of the Tenth International Semantic Web Conference (ISWC)*, (Bonn, Germany), 2011.

[3] Y. Gil, V. Ratnakar, E. Deelman, G. Mehta, and J. Kim, "Wings for pegasus: Creating large-scale scientific applications using semantic representations of computational workflows," in *AAAI*, (Vancouver, British Columbia, Canada), pp. 1767–1774, 2007.

[4] Y. Gil, P. Groth, V. Ratnakar, and C. Fritz, "Expressive reusable workflow templates," in *Proceedings of the Fifth IEEE International Conference on e-Science (e-Science)*, (Oxford, UK), 2009.

[5] T. M. Kurc, S. Hastings, V. S. Kumar, S. Langella, A. Sharma, T. Pan, S. Oster, D. Ervin, J. Permar, S. Narayanan, Y. Gil, E. Deelman, M. W. Hall, and J. H. Saltz, "High performance computing and grid computing for integrative biomedical research," *Journal of High Performance Computing Applications*, vol. 23, no. 3, pp. 252–264, 2009.

[6] A. Goderis, U. Sattler, P. Lord, and C. Goble, "Seven bottlenecks to workflow reuse and repurposing," in *The Semantic Web* (Y. Gil, E. Motta, V. Benjamins, and M. Musen, eds.), vol. 3729 of *Lecture Notes in Computer Science*, pp. 323–337, Springer Berlin / Heidelberg, 2005. In ISWC 2005.

[7] D. D. Roure, C. Goble, S. Aleksejevs, S. Bechhofer, J. Bhagat, D. Cruickshank, D. Michaelides, and D. Newman, "The myexperiment open repository for scientific workflows," in *Open Repositories 2009*, January 2009. Event Dates: May 2009.

[8] K. Wolstencroft, P. Fisher, D. D. Roure, and C. Goble., *Research in a Connected World*, ch. Scientific Workflows. 2009.

[9] M. Hauder, Y. Gil, R. Sethi, Y. Liu, and H. Jo, "Making data analysis expertise broadly accessible through workflows," in *In Proceedings of the Sixth Workshop on Workflows in Support of Large-Scale Science (WORKS'11), held in conjunction with SC 2011*, (Seattle, Washington), 2011.

[10] J. Kim, Y. Gil, and M. Spraragen, "Principles for interactive acquisition and validation of workflows," *Journal of Experimental and Theoretical Artificial Intelligence*, 2009.

[11] I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, *Workflows for e-Science: Scientific Workflows for Grids*. Springer-Verlag, 2006.

[12] M. Hauder, Y. Gil, and Y. Liu, "A framework for efficient text analytics through automatic configuration and customization of scientific workflows," in *In Proceedings of the Seventh IEEE International Conference on e-Science*, (Stockholm, Sweden), 2011.

[13] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with java implementations," 1999.

[14] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[15] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002.

[16] D. Garijo and Y. Gil, "A new approach for publishing workflows: Abstractions, standards, and linked data," in *SC WORKS*, (Seattle, Washington), 2011.

[17] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," pp. 412–420, Morgan Kaufmann Publishers, 1997.

[18] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *CSVT*, 2008.

[19] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, 2012.

[20] B. Song, R. J. Sethi, and A. K. Roy-Chowdhury, "Robust Wide Area Tracking in Single and Multiple Views," in *Visual Analysis of Humans* (T. B. Moeslund, L. Sigal, V. Krüger, and A. Hilton, eds.), pp. 1–18, Springer-Verlag, 2011.

[21] N. M. Nayak, R. J. Sethi, B. Song, and A. K. Roy-Chowdhury, "Motion Pattern Analysis for Event and Behavior Recognition," in *Visual Analysis*

of Humans (T. B. Moeslund, L. Sigal, V. Krüger, and A. Hilton, eds.), pp. 289–309, Springer-Verlag, 2011.

[22] J. Aggarwal and Q. Cai, “Human motion analysis: A review,” *CVIU*, 1999.

[23] Y. Wang, Y. Tian, L. Duan, Z. Hu, and G. Jia, “ESUR: A system for Events detection in SURveillance video,” in *ICIP*, no. 60973055, pp. 2317–2320, IEEE, 2010.

[24] D. Zhang, D. Gatica-Perez, and S. Bengio, “Modeling individual and group actions in meetings with layered HMMs,” *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 509–520, 2006.

[25] R. J. Sethi and A. K. Roy-Chowdhury, “Modeling and Recognition of Complex Multi-Person Interactions in Video,” in *ACM MM MPVA*, pp. 0–3, 2010.

[26] R. J. Sethi and A. K. Roy-Chowdhury, “Physics-based Activity Modelling in Phase Space,” in *ICVGIP*, 2010.

[27] N. M. Oliver, B. Rosario, and A. P. Pentland, “A Bayesian computer vision system for modeling human interactions,” *PAMI*, vol. 22, no. 8, pp. 831–843, 2000.

[28] N. Oliver, B. Rosario, and A. P. Pentland, “A Bayesian computer vision system for modeling human interactions,” in *ICVS*, Springer-Verlag, 1999.

[29] B. Ni, S. Yan, and A. Kassim, “Recognizing Human Group Activities with Localized Causalities,” in *CVPR*, pp. 1470–1477, 2009.

[30] Y. Zhou, B. Ni, S. Yan, and T. S. Huang, “Recognizing pair-activities by causality analysis,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 1, pp. 1–20, 2011.

[31] R. J. Sethi and L. Bry, “The Madsci Network,” tech. rep., madsci network research, 2011.

[32] R. Bergmann and Y. Gil, “Retrieval of semantic workflows with knowledge intensive similarity metrics,” in *Proceedings of the Nineteenth International Conference on Case Based Reasoning (ICCBR)*, (Greenwich, London), 2011.

[33] Y. Gil, V. Ratnakar, and C. Fritz, “Tellme: Learning procedures from tutorial instruction,” in *Proceedings of the ACM International Conference on Intelligent User Interfaces*, (Palo Alto, CA), 2011.