SCITECH +friends

RESEARCH SYMPOSIUM

May 12 and 13, 2021

Dialectical Tensions Surrounding CI
Fair Sharing of Network Resources Among Workflow Ensembles
Galaxy Morphology Classification
Crisis Computing Workflow
Lung Image Segmentation Using U-net Workflow
Best Practices In External Communication for Cyberinfrastructure Technology
Women In CI
and more...

https://scitech.group/research-symposium

C. Hayes, C. Kulkarni,
E. D. Milman, S. Okunloye,
A. Olshansky, R. Oruche,
K. Kee, P. C. S. Moreira,
C. Vardeman, T. Coleman,
T. M. A. Do, A. Jain,
P. Krawczuk, K. Lam,
S. Nagarkar, G. Papadimitriou,
S. Subramanya, R. White,
W. Whitcup, R. Ferreira da Silva,
E. Deelman

USC Viterbi School of Engineering
Information Sciences Institute

TEXAS TECH UNIVERSITY

UNIVERSITY OF NOTRE DAME

Preferred citation

C. Hayes, C. Kulkarni, E. D. Milman, S. Okunloye, A. Olshansky, R. Oruche, K. Kee, P. C. S. Moreira, C. Vardeman, T. Coleman, T. M. A. Do, A. Jain, P. Krawczuk, K. Lam, S. Nagarkar, G. Papadimitriou, S. Subramanya, R. White, W. Whitcup, R. Ferreira da Silva, E. Deelman, "SciTech and Friends Research Symposium 2021", Technical Report, May 2021, DOI: 10.5281/zenodo.4847543.

```
@misc{scitech2021symposium,
  author    = {Hayes, Cassandra and Kulkarni, Chaitra and Milman, Eric D. and
               Okunloye, Sayo and Olshansky, Alex and Oruche, Roland and
               Kee, Kerk and Moreira, Priscila C. S. and Vardeman, Charles and
               Coleman, Tain\~a and Do, Tu Mai Anh and Jain, Aditi and
               Krawczuk, Patrycja and Lam, Kelsie and Nagarkar, Shubham and
               Papadimitriou, George and Subramanya, Srujana and White,
               Rebecca and Whitcup, Wendy and Ferreira da Silva, Rafael
               and Deelman, Ewa},
  title     = {{SciTech and Friends Research Symposium 2021}},
  month     = {May},
  year      = {2021},
  publisher = {Zenodo},
  doi       = {10.5281/zenodo.4847543}
}
```

Table of Contents

Introduction

The Science Automation Technologies Research Group SciTech [1] at the USC Information Sciences Institute [2] aims to empower the scientific community by conducting research and software development in the area of automation of scientific computing, providing tools such as workflow management systems like Pegasus [3,4]. As a result, scientists can focus on their research questions, while our open-source tools provide the computational foundations to seamlessly run their experiments and analyses in local and distributed resources. In addition to workflow management, Scitech conducts research in resource scheduling and provisioning, cyberinfrastructure management and deployment, applied machine learning, and modeling and simulation of distributed computing systems. Scitech research is funded by the National Science Foundation, the U.S. Department of Energy, and the National Institutes of Health.

In the past academic year, the SciTech Group has included a number of Master's and undergraduate students. This "SciTech and Friends Research Symposium" aimed to provide a forum for students to publicize their work. SciTech is lead by Research Professor and Research Director Ewa Deelman. Her group collaborates with a number of diverse researchers and for this Symposium she partnered with her colleagues Prof. Kerk Kee [5] from Texas Tech University and Prof. Charles Vardeman [6] from the University of Notre Dame. The "SciTech and Friends Research Symposium" showcases the scholarly work of students from the three groups.

The Symposium was organized by the SciTech's Project Manager, Wendy Whitcup, with the support from Research Assistant Professor Rafael Ferreira da Silva (Research Lead at SciTech). The event was held virtually on May 12 and 13, 2021, and included 36 participants: undergraduate and graduate students from

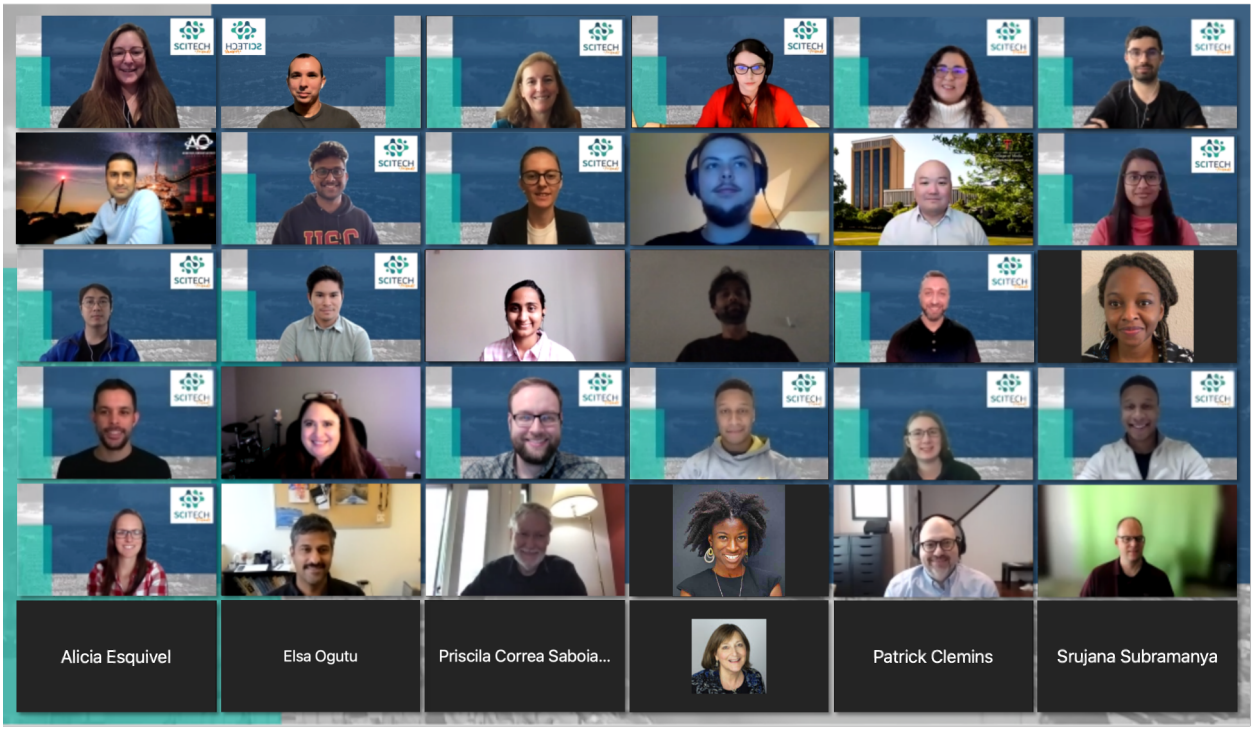


Figure 1: Screenshot of the SciTech & Friends Research Symposium participants. (The event was held virtually via Zoom on May 12 and 13, 2021.)

the three universities as well as researchers and professors from the community (Figure 1). A total of 16 students (1 high-school, 1 undergraduate, 3 MS, and 11 PhD) from the three universities presented their work in the form of a 20 minutes talk (15 minutes presentation and 5 minutes for questions), and provided a research abstract that comprises the remaining of this report.

The presentation slides, videos, and additional materials are all freely available on the symposium website: https://scitech.group/research-symposium.

Research Abstracts

Winning friends and influencing people: Characteristics of cyberinfrastructure opinion leaders

Cassie Hayes, Chaitra Kulkarni, and Alex Olshansky (Texas Tech University)
casshayes1024@gmail.com, {chaitra.kulkarni,alex.olshansky}@ttu.edu

Research Problem. The nature of cyberinfrastructure (CI) is complicated and abstract, yet the innovation is essential to support and advance modern research practices [7]. Thus, how and why professionals adopt CI as an innovation is a necessary area of examination to expand the community and technologies in the future. Effective opinion leadership has a large influence on whether an individual or organization decides to adopt complex innovations. Thus, in this study we focus on providing an answer to the following research question:

> **RQ:** What qualities of opinion leadership do CI professionals emphasize when describing their adoption of CI technologies and the CI community?

Challenges. Within this study, we contribute to an expanding body of literature regarding the role of influencers and opinion leaders within the diffusion of innovations. Diffusion describes how new ideas, behaviors, and technologies spread through social systems, a process driven and facilitated, in part, through the influence of opinion leadership [8]. Though only a feature of the diffusion process, scholars have noted the essential nature of opinion leadership to successful innovation diffusion [9, 10]. While such studies note the importance of opinion leadership, the particular qualities of opinion leaders remain understudied. With this study, we hope to fill in this gap in research.

Approach & Preliminary Results. We examined the qualities of CI opinion leaders as described by individuals within the CI community through grounded theory analysis of semi-structured interview data from a continuing project on organizational capacity for CI. During preliminary analysis, we found that CI influencers:

- Lead by example—spread excitement about CI by showing successes and enthusiasm.
- Write "love letters" about CI—advocate for the innovation.
- Bridge the gap between users, the public, and the community of CI—are accessible.
- Share information openly—are trainers, professors, and mentors.
- Navigate CI advancements toward the needs the innovation can meet for users.
- Envision and articulate a sustainable future for the CI discipline—spark organic interest.

Through identifying qualities of CI opinion leadership, we hope such features could be learned by individuals hoping to further diffuse CI technologies and expand to the surrounding community. Such a diffusion process could revolutionize scientific practice in the U.S. [7], as well as facilitating responsible diffusion practices in other areas.

Acknowledgements. This work is funded by NSF contracts #1453864 and #1939067.

Women in Cyberinfrastructure: Using Positive Deviance Strategies for Motivating New Generation of CI Women

Chaitra Kulkarni, Alex Olshansky, Cassandra Hayes, and Oluwabusayo Okunloye
(Texas Tech University)
{chaitra.kulkarni,alex.olshansky,Oluwabusayo.Okunloye}@ttu.edu, casshayes1024@gmail.com

Research Problem. In a report by the Commission on the Advancement of Women and Minorities in Science, it was found that women represent about 46% of the workforce, but comprise only 19% of the science, engineering, and technology workforce. A similar representation of women is also seen in cyberinfrastructure, a technologically inclined field that provides resources for advanced research capabilities not otherwise possible [7, 11], Young women usually find it difficult to foresee a career in technological fields due to limited women role models in advanced computing or by extension in CI [12]. Singhal introduced the "positive deviance/outliers" approach as an extension of diffusion of innovations theory [8] which posits that it is possible to motivate others by providing examples of those who took the less-taken path and succeeded [13, 14].

Challenges. Attracting women to science, engineering, and technology fields can be a challenge because of the stereotype that this field is more suitable for men [12]. To combat this, efforts have been taken by academic institutions, funding agencies, and private organizations to encourage more women in STEM and high-performance computing (HPC). However, there is limited literature on what brought these women to CI in the first place. Taking from Singhal's [13, 14] concept of positive deviance, diffusing the stories of these women to encourage more women to join this field is a way to tackle unequal representation of women in CI.

Approach & Preliminary Results. In-depth interviews of these women were conducted as a part of continued NSF-funded project. We analyzed the data using grounded theory approach [15, 16] of open, axial, and selective coding and implementing specific thematic analysis [17] of reoccurrence, repetition, and forcefulness.

Some themes found in this study are:

- First exposure to cyberinfrastructure—motivators to enter this field, influencers.
- Current work profile—what position, project, a goal they are working on.
- Motivators of new CI generation—train new students, collaborations with others.
- Career growth—possibility of climbing up the career ladder.

Using the positive deviance approach positive success stories can operate as motivations for encouraging a new generation of women in cyberinfrastructure. Strategies generated from this study can be implemented to combat the problem on unequal women representation in CI.

Acknowledgements. This work is funded by NSF contracts #1453864 and #1939067.

Opposites Attract: Dialectical Tensions Surrounding Cyberinfrastructure

Oluwabusayo Okunloye and Cassie Hayes (Texas Tech University)
Oluwabusayo.Okunloye@ttu.edu, casshayes1024@gmail.com

Research Problem. Dialectic tensions is derived from dialectical theory, which describes "all human relationships as grounded in contradictions: relational partners experience the pull of opposing forces that are desirable but mutually negate one another" [18]. Dialectical tensions represent two opposing ideas that are, in practice, connected to each other—opposite poles of the same underlying force. Although tension implies destructive disorder and dysfunction, they form the unifying negotiation of needs and fears that provide the foundation for organizations and relationships themselves [19]. This study examines the dialectical tensions that shape conceptualization of research within the cyberinfrastructure (CI) community. The research question guiding the study includes:

> **RQ:** What dialectical tensions do CI professionals attempt to navigate when discussing research practices relating to CI technologies?

Challenges. The term cyberinfrastructure describes a community of 'hardwares' 'softwares', 'virtual organizations', and people that make highly complex scientific computation accessible and possible [11]. CI is complex in terms of technologies as well as a surrounding community of diverse backgrounds. Thus, understanding the holistic culture and concept of CI is difficult and requires negotiation of the inherent tensions within the community. Kee and Browning [20] revealed five dialectic tensions relating to funding that affect the successful operation of CI and occur at the various overlapping levels of the institutional, organizational, and individual. However, there are other dialectical tensions necessary for CI as a concept, community, and set of functioning technologies have yet to be identified and examined. We attempt to provide some illumination on this understudied area of the community.

Approach/Preliminary Findings. Employing the grounded theory approach [15] in analyzing a series of semi-structured interviews from an ongoing project on organizational capacity for CI, we attempt to identify dialectical tensions surrounding CI, focusing on how dialectics negotiate opposing views on research practice and purpose within the community. Our examination of a subset of 33 randomly-selected transcripts from the original dataset produced the following preliminary findings:

- *Both* working with multidisciplinary fields—with different ontological perspectives and research paradigms—*and* relying on specialized technical computer science knowledge.
- Within this community, professionals work with individuals and organizations who are at different levels of the diffusion bell curve (early adopters, early majority,etc). Thus, they emphasize *either* innovation and being ahead of the curve or safety, doing what has been proven to work.
- *Both* the desire to share data openly *and* ensure security of data by CI users. Although the CI community runs an 'open data' policy, it must protect data from corruption and improper handling.
- Because of limitations in staffing or shifts in responsibilities, CI professionals must *either* focus on their own research projects *or* maintaining CI technologies.
- Research with CI must be *both* human *and* superhuman. It must be accessible and compatible to the work of users while providing adequate advancement that goes beyond what would be possible on the researcher's own.

Overall, findings revealed that in addition to identified tensions surrounding funding for CI [20], tensions surround the research practices that involve CI technologies. Researchers using CI must navigate multidisciplinary fields with different values and perspectives as well as the ways in which securing CI resources, the availability of adequate personnel, and the nature of the complex technologies themselves impact research. We believe our findings have possible implications for establishing policies that will make both interdisciplinary research practices and diffusion of CI more effective.

Acknowledgements. This work is funded by NSF contracts #1453864 and #1939067.

Best Practices in External Communication for Cyberinfrastructure Technology

Alex Olshansky, Chaitra Kulkarni, and Cassie Hayes (Texas Tech University)
{alex.olshansky,chaitra.kulkarni}@ttu.edu, casshayes1024@gmail.com

The burgeoning field of big data and advanced high-performance computational research, commonly referred to as cyberinfrastructure (CI) or e-science, is revolutionizing STEM research and education, and promises to drive economic growth through discovery and technological development [21, 22]. CI encompasses a multidimensional sociotechnical structure merging computational systems, data management and storage systems, scientific instruments and sensors, visualization tools, advanced networks, and the people that build, use, and maintain them [11]. Taken together, CI technology enables scientists to make discoveries from big data and further advance research, innovation, and education in the U.S., helping to maintain its position as a global leader in science, technology, and engineering [7].

The goal, then, as acknowledged by the National Science Foundation would be to facilitate the adoption and diffusion of CI technology within and among increasingly more STEM fields. However, communicating CI's importance, accessibility, and functionality, a necessary antecedent to its adoption and diffusion, is under-studied and not well understood. Furthermore, the complexity, scope, and requisite knowledge to operate advanced high-performance computers represents a major hurdle for communicating the accessibility of CI and significantly narrows the potential for widespread adoption [23–25].

Diffusion of innovations theory [8] has traditionally centered around the adoption and diffusion of a new technology or innovation. However, CI's sociotechnical structure represents a departure from traditional diffusion theory and an opportunity to expand its theoretical reach, laying the foundation for a new multidimensional approach to technology that involves both material objects and the social dimensions of multidisciplinary collaborations and support networks that enable its adoption. This study, therefore, seeks to bridge this gap and help advance the adoption and diffusion of CI technology by offering a collection of best practices in communication from CI experts at supercomputing centers across the country. Interviews were conducted with 132 CI practitioners, including domain scientists, computational technologists, and supercomputing center administrators, among others.

Preliminary thematic analysis from interview transcripts suggests several key aspects of external communication and outreach. First, adoption and diffusion depend primarily on raising awareness of the technology, which can involve aspects of public relations, marketing, and social media outreach. Next, establishing education and outreach training programs where users experience hands-on training and gain practical knowledge of CI tools is key for CI adoption. Building and establishing relationships with new users also facilitates further message dissemination and CI adoption. When new users establish relationships and build trust with CI support staff, they are more likely to encourage others to adopt. Further, understanding the subtleties in communicating to different stakeholders and avoiding a "one size fits many" approach is vital, and reinforces all of the above elements. Finally, establishing quantitative and qualitative feedback mechanisms can help encapsulate and inform diffusion efforts and improve them going forward.

Acknowledgements. This work is funded by NSF contracts #1453864 and #1939067.

Developing a COVID-19 Science Gateway with Chatbot Support for Clinicians

Roland Oruche (University of Missouri-Columbia) and Eric D. Milman (Texas Tech University)
rro2q2@mail.missouri.edu, Eric.d.milman@ttu.edu

In recent years, the development of cyberinfrastructure (CI) as a means of collecting, analyzing, and sharing large amounts of data (also known as "big data") has been rapidly emerging. In order to make CI more accessible to users, systems called "science gateways" have been developed to function in the same way as graphic user interfaces and internet browsers, in that they allow for the navigation of computed data in a more intuitive way. As such, they may appear similar to any other website or program. The difference lies in the computing power behind them, which, through CI, allows for the analysis of very large amounts of data. While existing science gateways have manifested its advanced capabilities for consumers, a science gateway that has dynamic ability to provide chatbot assisted guidance and a monitoring coupled with feedback process to improve functionality for gateway data users has yet to be developed.

In this study, we are focusing on the development of science gateways that enables data providers (e.g., CI developers, CI administrators) and data consumers (e.g., data professionals, researchers) to execute a set of science gateway tasks with advanced CI tools powered by a conversation agent viz., *Vidura Advisor*. The Vidura Advisor functions as a guide to the science gateway platforms, as it is capable of answering a variety of relevant questions in order to facilitate the successful navigation of the science gateway—particularly with more difficult tasks. The approach of using chatbots creates a "perspective guidance" to users, i.e., a conversational agent can be designed to provide personalized guidance for various steps of a user task in a science gateway that requires the use of advanced CI tools or in the aid of knowledge discovery from disparate data sources. In our development of the Vidura chatbot, we use Google Dialogflow to train over the set of potential questions a user might have.

In our example use case, we will examine KnowCOVID-19–a science gateway application in the context of the COVID-19 pandemic–used for data providers (e.g, admins, developers) and consumers (e.g., healthcare professionals, clinical researchers) that automatically filter high-quality publications for identifying tools topics and other important criteria for feasible, pandemic-related solutions. We investigate how chatbots can be developed in order to facilitate the use of science gateways through multiple phases. In our preliminary experiments, we conducted a qualitative assessment by interviewing clinicians and medical students to better understand the importance of a COVID-19 driven science gateway like KnowCOVID-19 and conversational agents as a starting point. As such, this phase will examine the following: *What functions do researchers and clinicians want to see in a COVID-19-focused science gateway with chatbot support?*

Our results demonstrate the critical need to incorporate conversational agents to foster the next generation of science gateways. While this has indicated optimism in using Vidura Advisor, there is uncertainty regarding its ability to yield effective responses for users. In the next phase of this study, we plan to focus on a usability assessment to measure the impact of the Vidura chatbot using a set of trivial sets for data consumers and data providers. In this future work, we plan to assess the effectiveness of the chatbot by integrating a systematic framework that explores and analyzes the efficiency of the Vidura chatbot as it relates to aiding users in using the KnowCOVID-19 science gateway. Inspired by conversation analyses, we plan to adopt various metrics such as time between responses and whether a user was able to carry out a feasible outcome based on the chatbot's suggestions and use these objective data points to systematically improve the performance of the Vidura chatbot.

Acknowledgements. This work is supported by the National Science Foundation under awards: OAC-1730655, OAC-2006816 and OAC-2007100. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Generative Adversarial Networks for Paper Analytical Devices

Priscila Correa Saboia Moreira (University of Notre Dame)
pmoreira@nd.edu

Paper Analytical Devices. Paper Analytical Devices (PADs) [26] are paper cards with imprinted lanes of reagents that aim at quickly and cheaply testing the presence of substances of interest, such as medicine and drugs. PADs are useful in scenarios where one has suspicions about the quality or even the falsification of pharmaceuticals, an unfortunate reality in some developing-world markets.

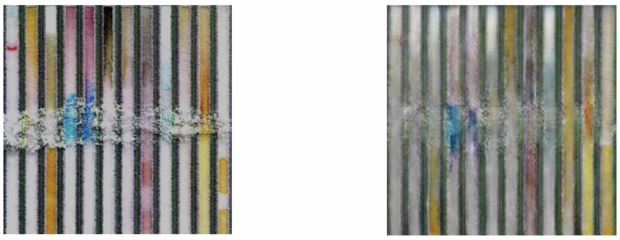
The usage of PADs comprises two main stages, namely (i) the presentation and the chemical reaction of the questioned substance with the card reagents, and (ii) the reading of the results of the reaction through the interpretation of the modification of the visual appearance of the paper card. The first stage was designed to be quick and easy, via a simple friction of the pharmaceutical on the card. The second stage, in turn, demands the inspection of a trained expert (either human or computer program), to detect the presence and quality of the questioned substance. Human experts, in particular, can become very skilled in reading the test results, with an average accuracy of 97% of correct detection. However, the amount of cards needing evaluation can quickly surpass the number of available experts, especially in the case of remote locations.

Computational Reading of PAD Results. Computer programs embedded in handheld camera devices can be used to automatically process the picture of a test card after reaction and detect the presence of a target substance [27]. Such programs are developed with techniques from the computer science subfields of image processing, computer vision, and machine learning. In particular, our team of computational scientists was able to train a deep learning [28] solution that presents an average accuracy of 96% of correct substance detection.

Training accurate deep learning solutions, though, requires a large amount (preferably in the order of hundreds of thousands) of carefully curated, labeled, and scanned paper card samples after reaction, for each substance of interest. This requirement may be rendered impossible to be achieved in the case of time sensitive scenarios or of expensive substances.

Generative Adversarial Networks. Generative Adversarial Networks (GANs) [29] can be used to generate a large set of inexpensive and realistic synthetic examples of paper test cards after reaction, which might be useful as additional training data. New synthetically generated cards might help to train additional deep learning solutions to detect new substances that might become of interest in future scenarios.

We have verified the feasibility of generating synthetic samples, which are indeed hard to be spotted as synthetic by human experts. Figure 2 brings a pair of real and synthetic paper cards, for illustration sake. Next steps include using these cards to train a new solution.



(a) Real card (b) Synthetic card

Figure 2: Penicillin procaine paper card reaction samples. In (a), a real card after reaction and proper scanning. In (b), a synthetic sample generated by a GAN.

Results. In the conduction of this research, we were able to:

- Train an automatic solution that detects substances on cards with an accuracy of 96%.
- Generate synthetic and realistic pictures of cards after reaction.

Acknowledgements. This project was supported by Management Sciences for Health (MSH) under award number A18-0197-001.

WfChef: Automated Generation of Accurate Scientific Workflow Generators

Tainã Coleman (University of Southern California)
tcoleman@isi.edu

Scientific workflow applications have become mainstream and their automated and efficient execution on large-scale compute platforms is the object of extensive research and development. For these efforts to be successful, a solid experimental methodology is needed to evaluate workflow algorithms and systems. A foundation for this methodology is the availability of realistic workflow instances. Dozens of workflow instances for a few scientific applications are available in public repositories. While these are invaluable, they are limited: workflow instances are not available for all application scales of interest. To address this limitation, previous work has developed generators of synthetic, but representative, workflow instances of arbitrary scales. These generators are popular, but implementing them is a manual, labor-intensive process that requires expert application knowledge. As a result, these generators only target a handful of applications, even though hundreds of applications use workflows in production.

WfCommons [30] is a framework for enabling scientific workflow research and development. It provides foundational tools for analyzing workflow execution instances, generating workflow recipes, and generating synthetic, yet realistic, workflow instances (Figure 3). WfCommons importance lies in the use of the synthetic generated workflow instances to develop new techniques, algorithms, and systems to overcome the challenges of efficient and robust execution of ever larger workflows on increasingly complex distributed infrastructures.

WfInstances. Collection and curation of open access production workflow executions from various scientific applications shared in a common instance format (WfFormat). The WfCommons Python package provides a set of tools for analyzing instances, which can be used to develop workflow recipes for the WfGen.

WfGen. WfGen targets the generation of realistic synthetic workflow traces with a variety of characteristics. The workflow generator uses recipes of workflows for creating different synthetic instances based on distributions of workflow job runtime, and input and output file sizes.

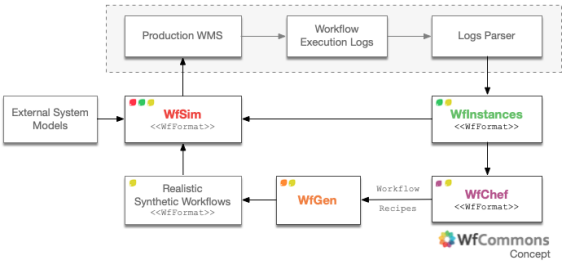


Figure 3: WfCommons framework concept. Overview of the workflow research life cycle process.

WfSim. WfSim fosters the use of simulation for the development, evaluation, and verification of scheduling and resource provisioning algorithms (e.g., multi-objective function optimization, etc.), evaluation of current and emerging computing platforms – i.e. clouds, IoT, extreme scale, etc.

WfChef. In this work, we present the newest component of WfCommons, WfChef [31], a tool that fully automates the process of constructing a synthetic workflow generator for any scientific application. Based on an input set of workflow instances, WfChef automatically produces a synthetic workflow generator. We define and evaluate several metrics for quantifying the realism of the generated workflows. Using these metrics, we compare the realism of the workflows generated by WfChef generators to that of the workflows generated by the previously available, hand-crafted generators. We find that the WfChef generators not only require zero development effort (because it is automatically produced), but also generate workflows that are more realistic than those generated by hand-crafted generators.

Acknowledgements. This work is funded by NSF contracts #1923539 and #1923621; and partly funded by NSF contracts #2016610, and #2016619.

Efficiency and Resource Evaluation of In Situ Workflows

Tu Mai Anh Do (University of Southern California)
tudo@isi.edu

Advances in high-performance computing (HPC) allow scientific simulations to run at an ever-increasing scale, generating a large amount of data that needs to be analyzed over time. Conventionally, the simulation outputs the entire simulated data set to the file system for later post-processing. Unfortunately, the slow growth of I/O technologies compared to the computing capability of present-day processors causes an I/O bottleneck of post-processing as saving data to storage is not as fast as data is generated. Several new high-performance memory systems that reside closer to the computation units have been developed (e.g., burst buffers, non-volatile memory, high-bandwidth memory, etc.). This trend leads to the transition from compute-centric models, in which data are moved to where the computation takes place, to data-centric models, which require locating computing components close to where data are placed.

According to data-centric models, a new processing paradigm has been recently emerged, called *in situ*, where simulation data is analyzed on-the-fly to reduce the expensive I/O cost of saving massive data for post-processing. In *in situ* analytics, the main simulation periodically sends data to analysis kernels running simultaneously with simulation component as the run progresses. Since the *in situ* analysis interleaves its execution with the simulation, running the analysis *in situ* helps to reduce the time to solution of the analysis pipeline.

An *in situ* workflow describes a scientific workflow with multiple components (simulations and analyses) running concurrently, potentially coordinating their executions using the same allocated resources to minimize the cost of data movement. However, co-locating the simulation and the analysis requires them to be effectively managed to deal with performance interference between the *in situ* components running on the same resource. The ability to address the performance issue depends on the in-depth understanding of the *in situ* design space, which includes component placements, data coupling schemes, and synchronization methods. Since *in situ* workflows are comprised of co-located tasks running concurrently in an iterative manner, the execution yields complicated behaviors that create challenges in evaluating the efficiency of an *in situ* run.

To enable efficient execution of *in situ* workflows, our work introduces an approach to characterize *in situ* execution in an accurate and efficient way. Based on the insights gained from this characterization, we propose a theoretical framework that models the efficiency of *in situ* execution for evaluating the performance of *in situ* workflows. By applying the proposed performance model to a real molecular dynamic simulation use case, we have shown the advantages of data locality when co-locating the simulation with the corresponding analyses in an *in situ* workflows. More importantly, the performance model is lightweight to run concurrently with the *in situ* workflow and enable its adaptation, in which the performance assessment can drive and adjust the execution at runtime.

Acknowledgements. This work is funded by NSF contracts #1741057, #1740990, and #1741040, and DOE contract DE-SC0012636.

Lung Image Segmentation Using U-Net

Aditi Jain (University of Southern California)
amjain@usc.edu

Motivation. Lung segmentation constitutes a critical procedure for any clinical-decision sup- porting system aimed to improve the early diagnosis and treatment of lung diseases. For decades, acute lower respiratory tract infections have been among the top three causes of death and disability among both children and adults. Segmentation of lung images help in easier and more efficient medical diagnosis.

Deep learning in lung segmentation. Deep learning (DL) has become a conventional method for constructing net- works capable of successfully modeling higher-order systems to achieve human- like performance. Tumors have been direct targets for DL-assisted segmentation of medical images. A lung cancer screening tool was implemented using DL structures aiming to lower the false positive rate in lung cancer screening with low-dose CT scans. Also, researchers attempted to segment brain tumors from MRI images with a hybrid network of U-NET and SegNet, reaching an accuracy of 0.99. The accuracy achieved by using deep learning techniques for lung segmentation has been really high.

Unet. Unet architecture consists of a contracting path to capture context and asymmetric expanding path that enables precise localization.

Workflow. The complete workflow is divided into the following jobs:

- **Preprocessing:** This job performs image augmentation and splitting up of data into train, test and validation set. This job has been parallelized due to independent and non-overlapping execution.
- **hpo:** This job performs the optimization of hyperparameters like learning rate using Optuna.
- **Train Model:** This job performs the actual training of the UNet. We are using a pre-trained model and tuning over our small, augmented dataset.
- **Prediction:** This job is responsible for predicting the masks of the test dataset.
- **Evaluate:** This job evaluates the model performance and generates a pdf showing the results.

Dataset. The dataset is collected from NCBI (National Center for Biotechnology Information) that consists of a total of around 800 images and masks [32].

Results. Using Unet for the task of lung image segmentation has been highly useful in achieving a good IOU score (Intersection over union). The IOU score obtained was around 0.8.

Acknowledgements. This work is funded by NSF contract #1664162.

Using Gantt Charts and CNNs to Study end-to-end Scientific Workflows and their Anomalies through Visual Analysis

Patrycja Krawczuk (University of Southern California)

krawczuk@isi.edu

Modern scientific experiments are conducted on complex, large-scale, distributed high-performance infrastructure like DOE Leadership Computing Facilities (e.g., NERSC, OLCF). Even though these systems are designed with reliability in mind [33], they can experience anomalies ranging from subtle (e.g. network performance degradation) to critical (e.g., file system integrity errors) [34], affecting the performance of the applications leveraging their resources and increasing the chances of failures.

Workflow management systems, such as Pegasus [4], have emerged as very important tools for managing the execution of science applications. Automation, reproducibility, resiliency, and monitoring are some of the key advantages they can provide. However, anomalies remain a significant barrier to the reliable execution of scientific workflows at scale. It is still difficult to understand subtle anomalies that impact performance and provide useful feedback to users and system operators.

We suggest a novel approach to model, detect and classify anomalies in scientific workflow traces, that leverages advances in computer vision. First, we introduce a methodology to calculate and visualize high-resolution end-to-end workflow execution timelines (Gantt charts). These Gantt charts provide us with a complete overview of the workflow's execution and break down the timeline of each individual task into 9 phases: ready time, WMS prepare time, queue time, pre-script and postscript time, stage in and stage out time, runtime, finished. Then, we explore and evaluate different Convolutional Neural Network (CNN) architectures when applied to the Gantt chart for anomaly detection and classification. We employ CNN models trained from scratch and we investigate whether transfer learning from pre-trained models (AlexNet [35], VGG-16 [36], and ResNet-18 [37]) on ImageNet [38] leads to better accuracy.

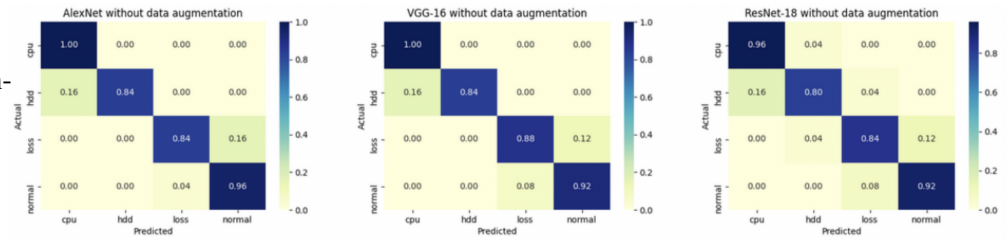


Figure 4: Confusion matrices for the pre-trained models.

This work is the first to apply computer vision methods and identify anomalies in workflow execution traces. Our initial results have been produced using a dataset containing 1000 execution traces of the Pegasus 1000Genome workflow [39], executed on the ExoGENI testbed [40] under normal and anomalous conditions (we introduced CPU, HDD, and network-related anomalies). oth trained from scratch and pre-trained models show promising results with accuracy of over 90% at anomaly detection and classification. The pre-trained models perform slightly better than the models trained from scratch, but in our analysis, there is only a 2% difference between the provided accuracies. Our simple CNN, AlexNet, and VGG-16 achieve 100% accuracy when detecting CPU-related anomalies (see Figure 4). All of the models can identify the normal runs (without any anomalies) with an accuracy of at least 92%. The misclassification errors are most common between normal runs and packet loss injected anomalies. This interference influences the characteristics of the traces subtly and it is hard to identify by our model. More data is needed to improve the generalization and the performance of the models.

Acknowledgements. This work is funded by DOE contract DE-SC0012636, and NSF contract #1664162.

Face Mask Detection Workflow

Kelsie Lam (University of Southern California)
kelsielam17@gmail.com

The rapid spread of Covid-19 throughout the world has caused the Centers for Disease Control and Prevention to set up a mask mandate within the country. However, this guideline is not always followed by everyone so we need to figure out what percentage of a population is truly following this rule. Thus, we have developed a face mask detection workflow so that if we set up a camera to take pictures of people at public places periodically, we can run those images into the workflow to determine how many people are actually wearing their masks.

This workflow has a total of eight different steps: data acquisition, data exploration, data split, data preprocessing, hyperparameter optimization, model training, model evaluation, and inference. We first gathered all the data we will be using which are a variety of images of people either wearing a mask, not wearing a mask, or wearing a mask incorrectly. With these images, we learned about the distribution of the dataset so we can account for any imbalances and then split them into three groups: training, testing, and validation. Additionally, we also standardized our data to ensure that it is all uniform and consistent as well as added gaussian noise to the images, making it more robust to allow us to increase the number of training images we can use. Next, we used an automatic hyperparameter optimization software framework called Optuna to help us find the best hyperparameters to use for our model. We also used a pre-trained model called the FastRCNNPredictor in the training step since it is a lot more convenient for our project. With this pre-trained model, we fine-tuned it to better detect peoples' faces and classify them each into one of the three different categories. This model simply extracted region proposals of the image we provided it with, classified the people, and added bounding boxes to them. Once our model had been trained, we evaluated its performance. We evaluated the model by looking at the loss curves and validation loss plots to see if the model was overfitting. In our case, since our data was unbalanced, we also used a confusion matrix since it is the most beneficial for imbalanced data and can give us the exact cause of errors we may encounter. Lastly, our final step was to look at our results and determine what is working well and what needs more work.

Our results show that the model does a tremendous job detecting people wearing their face masks correctly. On the other hand, it needs more work detecting those wearing their masks incorrectly or not wearing them at all. Another issue that the model ran into is detecting head coverings. As shown in the diagram, it has a difficult time differentiating between a hat and a person's face. Thus, a solution we have come up with to help resolve these issues is to balance out and diversify our dataset. By adding more images of people not wearing a mask, wearing a mask incorrectly, or wearing head coverings, we can help the model learn and better understand the differences between these categories.

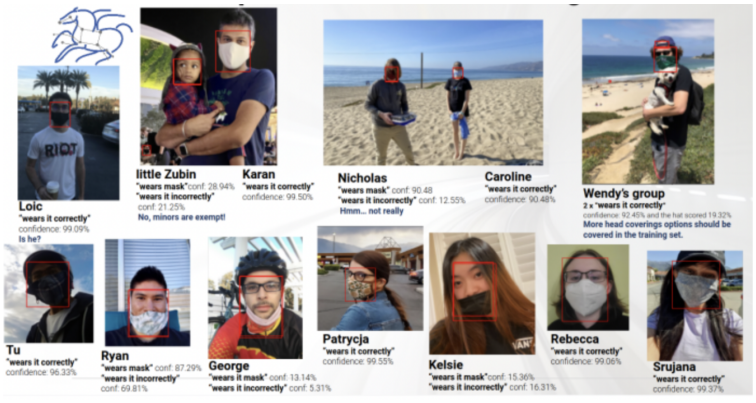


Figure 5: Some examples from members of the Pegasus Team.

Acknowledgements. This work is funded by NSF contract #1664162.

Crisis Computing Workflow

Shubham Nagarkar (University of Southern California)
slnagark@isi.edu

Introduction and Motivation. An increasing number of people use Social Media (SM) platforms like Twitter and Instagram to report critical emergencies or disaster events. Multimodal data shared on these platforms provide information about the scale of the event, victims, and infrastructure damage. The data can provide local authorities and humanitarian organizations with a big-picture understanding of the emergency. Moreover, it can be used to effectively and timely plan relief responses. Addressing this problem requires more research and has been impeded by the lack of large-scale annotated datasets. In our work, we aim to address the challenge of finding relevant information among the vast amount of published SM posts. Specifically, we aim to improve state-of-the-art performances on classification tasks on the CrisisMMD dataset using novel methods like Contrastive Learning, Transfer Learning and Multi-Modal Late and Early Fusion techniques.

Challenges. One of the biggest challenges is handling social media information overload. To extract relevant information a computational system needs to process massive amounts of data and identify which data is *informative* in the context of disaster response. Another issue is data scarcity. To develop effective applications that could assist in crisis response, researchers need access to the large-scale annotated dataset.

Approach. We attempted to solve this problem by using a two step approach as follows:

1. **Improved state-of-the-art performance on the classification task i.e. *informativeness* on the Crisis-MMD dataset.** Implemented Transfer Learning approach by using a ResNet-50 model pre-trained on ImageNet dataset for the classification task at hand which served as a baseline for the Image pipeline. Used Bi-LSTM networks to classify the tweets corresponding to the images which served as a baseline for the Text pipeline. Employed a novel method called the Supervised Contrastive Learning (SupCon) to create effective and robust feature level representations of the crisis-related images. Used attention-based model Distil BERT to create domain-specific contextual crisis embeddings.
2. **Fused both the modalities to further improve the classification accuracy.** Implemented Late Fusion techniques like Mean Probability Concatenation, Custom Decision Policy and Logistic Regression on the results obtained from our baseline models i.e ResNet-50 for images and Bi-LSTM for text. Implemented Early Fusion by merging the robust intermediate feature representations obtained from SupCon for Images and Distil BERT for text.

Preliminary Results. **Late Fusion:** In this strategy, we merged the output probabilities from Bi-LSTM and ResNet-50 using techniques like Mean Probability Concatenation, Custom Decision Policy (MLP) and Logistic Regression. We saw improved performances in all these techniques due to well-trained base models, thereby beating existing baselines by a clear margin.

Model	Mean Probability	Custom Decision	Logistic Regression
Our Model	**92.0**	**91.8**	**91.5**
Gautam et al.	79.2	80.2	80.2

Early Fusion: We explored the early fusion paradigm, where the multimodal data representations are fused at the level of hidden layers. We combined our SupCon's image embeddings of size 1x128 and fine-tuned DistilBERT's sentence embeddings of size 1x 64 and fine-tuned the last layers. We successfully outperformed the baseline architectures with our proposed model.

Modality	Olfi et al.	Ours
Text	0.808	**0.84**
Image	0.833	**0.89**
Image + Text	0.844	**0.91**

Acknowledgements. This work is funded by NSF contract #1664162.

Cross-Cloud Resource Management in Scientific Workflows

George Papadimitriou (University of Southern California)
georgpap@isi.edu

Resource provisioning is getting more and more complicated with the introduction of new hardware accelerators, highly distributed compute and storage resources, and complicated network topologies. Additionally the advent of cloud and serverless computing has created the trend of leasing (acquiring resources on demand) rather than maintaining a static infrastructure. Having this resource heterogeneity at our disposal and the power to acquire them on demand, raises the question of what resources are more suitable for a scientific application's execution. Moreover, new clouds and testbeds oriented to scientific research and computing are funded by national agencies, and offer bleeding edge capabilities. It is important to understand how to best leverage these new resources and integrate them into the scientists' infrastructure and applications designed with older paradigms in mind.

What am I trying to do. Introduce a methodology to model the resource requirements of scientific applications encapsulated into workflows. Study how the application requirements evolve under load or anomalous conditions, and enable the application to drive allocation of the infrastructure available to it, spanning across multiple clouds.

How it is done today. Ad-hoc resource acquisition is a common approach in the scientific community. If an application has a predictable workload, load balancing techniques can be utilized to scale up or scale down the resources. These techniques take into consideration QoS requirements, costs, scalability but usually focus on a single cloud provider (usually commercial) and rely on platform specific tools.

What is new. My methodology will be applicable to a large number of scientific applications, by automating the process of profiling on compute, storage and network. To achieve this a requirement of modeling the scientific application into a familiar structure (a scientific workflow) will be required. Using static code analysis techniques (e.g., libraries used), and collecting and analysing execution traces containing statistics of the infrastructure's subcomponents, we can provide insight into their utilization and the level of support by the application. During the analysis of the execution traces, correlating them with the scientific application will be of high importance, in order to achieve reliable conclusions. Following this approach we could provide meaningful feedback to the user and generate an algorithm that can automatically identify the more suitable resources from a pool of available resource specifications (e.g., cloud resources), alleviating scientists from actively managing their scientific workflow resources.

Who cares. Scientific community, system operators of computational infrastructures

Acknowledgements. This work is funded by NSF contracts #1664162, and #1826997.

Event Horizon Telescope Data and Open Source Methods

Rebecca White (University of Southern California)
beckswhi@usc.edu

Introduction. The Event Horizon Telescope (EHT) array aims to image the event horizons of supermassive blackholes with high resolution so we can better understand these astrophysical phenomena. The EHT collaboration released some of their data publicly and the methods used are outlined in papers. Unfortunately, not everything is easy to find and run. Our goal is to report the experience in replicating their work as scientists outside of the collaboration. This reproducibility study will determine if there is enough open-source methods to get the same or similar images to those done by scientists inside the collaboration.

Reproducibility and having data in the public domain is important because the more scientists working on this data, the more we could potentially find out about black holes as well as improve upon the methods of data cleansing to get sharper images. Also, the results could be verified by people outside of the collaboration, bringing more validity to the results.

Challenges. The biggest challenge so far has been finding the methods and parameters to run the tests. In the papers we have looked at to gather the methods used, we have come across some dead links. The papers also have not been the clearest on what parameters are needed to run the methods.

Approach. Our approach leading up to recently has been scouring EHT papers for methods, equations, and parameters pertaining to data analysis and compiling them. Recently and in the near future, our approach will consist of: synthesizing the open source data to put into algorithms; plotting the data from all the days by recreating figures in the papers; and compiling all our notes in a github repo.

Preliminary Results.

(a) Preliminary results from April 5th 2017 data.

(b) Paper results from April 5th 2017 data.

Figure 6: Preliminary Results vs. Paper Results - The preliminary results look similar, however are using a different scale and are missing the circles representing the baseline lengths.

Acknowledgements. This work is funded by NSF contracts #1664162, and #2041901.

References

[1] "SciTech: Science Automation Technologies," http://scitech.group, 2021.

[2] "USC Information Sciences Institute," http://isi.edu, 2021.

[3] "Pegasus workflow mangement system," http://pegasus.isi.edu, 2021.

[4] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. F. Da Silva, M. Livny *et al.*, "Pegasus, a workflow management system for science automation," *Future Generation Computer Systems*, vol. 46, pp. 17–35, 2015.

[5] "Prof. Kerk F. Kee," http://www.ekerk.com, 2021.

[6] "Prof. Charles Vardeman," https://crc.nd.edu/about/people/charles-vardeman, 2021.

[7] D. E. Atkins, *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*. National Science Foundation, 2003.

[8] G. Orr, "Diffusion of innovations, by everett rogers (1995)," *Retrieved January*, vol. 21, p. 2005, 2003.

[9] T. W. Valente and R. L. Davis, "Accelerating the diffusion of innovations using opinion leaders," *The Annals of the American Academy of Political and Social Science*, vol. 566, no. 1, pp. 55–67, 1999.

[10] P. S. Van Eck, W. Jager, and P. S. Leeflang, "Opinion leaders' role in innovation diffusion: A simulation study," *Journal of Product Innovation Management*, vol. 28, no. 2, pp. 187–203, 2011.

[11] C. A. Stewart, S. Simms, B. Plale, M. Link, D. Y. Hancock, and G. C. Fox, "What is cyberinfrastructure," in *Proceedings of the 38th annual ACM SIGUCCS fall conference: navigation and discovery*, 2010, pp. 37–44.

[12] K. W. English, K. F. Hulme, and K. E. Lewis, "Engaging high school women in engineering design using cyberinfrastructure," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 43291, 2008, pp. 543–554.

[13] A. Singhal, J. Sternin, and L. Dura, "Combating malnutrition: Positive deviance grows roots in vietnam in the land of a thousand rice fields," *British Columbia, Canada: The Communication Initiative Network*, 2009.

[14] A. Singhal, "Communicating what works! applying the positive deviance approach in health communication," *Health communication*, vol. 25, no. 6-7, pp. 605–606, 2010.

[15] J. M. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative sociology*, vol. 13, no. 1, pp. 3–21, 1990.

[16] A. Strauss and J. Corbin, "Grounded theory methodology: An overview." 1994.

[17] W. F. Owen, "Interpretive themes in relational communication," *Quarterly journal of Speech*, vol. 70, no. 3, pp. 274–287, 1984.

[18] G. J. Galanes, "Dialectical tensions of small group leadership," *Communication Studies*, vol. 60, no. 5, pp. 409–425, 2009.

[19] G. e. A. Trethewey and K. L. Ashcraft, "Special issue introduction: Practicing disorganization: The development of applied perspectives on living with tension," *Journal of Applied Communication Research*, vol. 32, no. 2, pp. 81–88, 2004.

[20] K. F. Kee and L. D. Browning, "The dialectical tensions in the funding infrastructure of cyberinfrastructure," *Computer Supported Cooperative Work (CSCW)*, vol. 19, no. 3-4, pp. 283–308, 2010.

[21] C. L. Borgman, H. Abelson, L. Dirks, R. Johnson, K. R. Koedinger, M. C. Linn, C. A. Lynch, D. G. Oblinger, R. D. Pea, K. Salen *et al.*, "Fostering learning in the networked world: The cyberlearning opportunity and challenge. a 21st century agenda for the national science foundation," 2008.

[22] T. Hey and A. E. Trefethen, "Cyberinfrastructure for e-science," *Science*, vol. 308, no. 5723, pp. 817–821, 2005.

[23] M. J. Bietz, T. Ferro, and C. P. Lee, "Sustaining the development of cyberinfrastructure: an organization adapting to change," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 2012, pp. 901–910.

[24] K. F. Kee and L. D. Browning, "Challenges of scientist-developers and adopters of existing cyberinfrastructure tools for data-intensive collaboration, computational simulation, and interdisciplinary projects in early e-science in the us," in *Data-Intensive Collaboration in Science and Engineering workshop, Computer Supported Cooperative Work (CSCW 12), Seattle, WA*, 2012.

[25] D. Spencer, A. Zimmerman, and D. Abramson, "Special theme: project management in e-science: challenges and opportunities," *Computer Supported Cooperative Work (CSCW)*, vol. 20, no. 3, pp. 155–163, 2011.

[26] "Paper analytical devices," https://padproject.nd.edu, 2021.

[27] S. Banerjee, J. Sweet, C. Sweet, and M. Lieberman, "Visual recognition of paper analytical device images for detection of falsified pharmaceuticals," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[29] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[30] R. Ferreira da Silva, L. Pottier, T. Coleman, E. Deelman, and H. Casanova, "Workflowhub: Community framework for enabling scientific workflow research and development," in *2020 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*. IEEE, 2020, pp. 49–56.

[31] T. Coleman, H. Casanova, and R. Ferreira da Silva, "Wfchef: Automated generation of accurate scientific workflow generators," *arXiv preprint arXiv:2105.00129*, 2021.

[32] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[33] M. Snir, R. W. Wisniewski, J. A. Abraham, S. V. Adve, S. Bagchi, P. Balaji, J. Belak, P. Bose, F. Cappello, B. Carlson *et al.*, "Addressing failures in exascale computing," *The International Journal of High Performance Computing Applications*, vol. 28, no. 2, pp. 129–173, 2014.

[34] M. Rynge, K. Vahi, E. Deelman, A. Mandal, I. Baldin, O. Bhide, R. Heiland, V. Welch, R. Hill, W. L. Poehlman *et al.*, "Integrity protection for scientific workflow data: Motivation and initial experiences," in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*, 2019, pp. 1–8.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[39] R. Ferreira da Silva, R. Filgueira, E. Deelman, E. Pairo-Castineira, I. M. Overton, and M. P. Atkinson, "Using simple pid-inspired controllers for online resilient resource management of distributed scientific workflows," *Future Generation Computer Systems*, vol. 95, pp. 615–628, 2019.

[40] I. Baldin, J. Chase, Y. Xin, A. Mandal, P. Ruth, C. Castillo, V. Orlikowski, C. Heermann, and J. Mills, "Exogeni: A multi-domain infrastructure-as-a-service testbed," in *The GENI Book*. Springer, 2016, pp. 279–315.

Appendix A: Agenda

Day 1 - May 12, 2021

Time	Topic
9:00-9:10am PDT	Day 1 Welcome
9:10-9:30am PDT	**Dialectical Tensions Surrounding CI** Sayo Okunloye *(Texas Tech University)*
9:30am-9:50am PDT	**Efficiency and Resource Evaluation of In Situ Workflows** Tu Mai Anh Do *(University of Southern California)*
9:50-10:10am PDT	**Developing a COVID-19 Science Gateway with Chatbot Support for Clinicians** Roland Oruche *(University of Missouri)* and Eric Milman *(Texas Tech University)*
10:10-10:30am PDT	**Fair Sharing of Network Resources Among Workflow Ensembles** George Papadimitriou *(University of Southern California)*
10:30-10:50am PDT	**Galaxy Morphology Classification** Srujana Subramanya *(University of Southern California)*
10:50-11:10am PDT	**Crisis Computing Workflow Multimodal Social Media Content for Improved Emergency Response** Shubham Nagarkar *(University of Southern California)*
11:10-11:30am PDT	**Using Gantt Charts and CNNs to Study end-to-end Scientific Workflows and their Anomalies through Visual Analysis** Patrycja Krawczuk *(University of Southern California)*
11:30-11:50am PDT	**Generative Adversarial Networks for Paper Analytical Devices** Priscila C. Saboia Moreira *(University of Notre Dame)*
11:50-12:00pm PDT	Day 1 Closing Remarks

Day 2 - May 13, 2021

Time	Topic
9:00-9:10am PDT	Day 2 Welcome
9:10-9:30am PDT	**Face Mask Detection Workflow** Kelsie Lam *(University of Southern California)*
9:30am-9:50am PDT	**Lung Image Segmentation Using U-Net Workflow** Aditi Jain *(University of Southern California)*
9:50-10:10am PDT	**Best Practices in External Communication for Cyberinfrastructure Technology** Alex Olshansky *(Texas Tech University)*
10:10-10:30am PDT	**Winning Friends and Influencing People: Characteristics of Cyberinfrastructure Opinion Leaders** Cassie Hayes *(Texas Tech University)*
10:30-10:50am PDT	**Women in Cyberinfrastructure: Using Positive Deviance Strategies for Motivating New Generation of CI Women** Chaitra Kulkarni *(Texas Tech University)*
10:50-11:10am PDT	**Event Horizon Telescope data with Open Source methods** Rebecca White *(University of Southern California)*
11:10-11:30am PDT	**WfChef: Automated Generation of Accurate Scientific Workflow Generators** Tainã Coleman *(University of Southern California)*
11:30-11:40am PDT	Closing Remarks

Appendix B: List of Participants

Name	Affiliation
Aditi Jain	University of Southern California
Alex Olshansky	Texas Tech University
Alicia Esquivel	University of Missouri
Angela Murillo	Indiana University
Cassandra Hayes	Texas Tech University
Chaitra Kulkarni	Texas Tech University
Charles Vardeman	Notre Dame University
Ciji Davis	University of Southern California
Elsa Ogutu	University of Southern California
Eric Milman	Texas Tech University
Ewa Deelman	University of Southern California
George Papadimitriou	University of Southern California
Julio Alvarado	University of Central Florida (Arecibo Observatory)
Karan Vahi	University of Southern California
Kerk Kee	Texas Tech University
Loic Pottier	University of Southern California
Maciej Krawczuk	University of Southern California
Mary Gohsman	Notre Dame University
Mats Rynge	University of Southern California
Oluwabusayo Okunloye	Texas Tech University
Patrick Clemins	University of Vermont
Patrycja Krawczuk	University of Southern California
Priscila Correa Saboia Moreira	Notre Dame University
Rafael Ferreira da Silva	University of Southern California
Rajiv Mayani	University of Southern California
Rebecca White	University of Southern California
Roland Oruche	University of Missouri
Ryan Tanaka	University of Southern California
Shubham Nagarkar	University of Southern California
Srujana Subramanya	University of Southern California
Tainã Coleman	University of Southern California
Tu Mai Anh Do	University of Southern California
Wendy Whitcup	University of Southern California