

Summary

Researchers in diverse domains from astronomy and atmospheric sciences, to earth sciences and geonomics are generating massive datasets at an unprecedented scale. The availability of rapidly evolving computational and data technologies for harnessing these datasets are enabling a rich environment for establishing robust collaborations and building communities of data scientists and domain scientists and engineers that are central for transforming these datasets into information and knowledge. With this opportunity on the forefront, NSF will soon fund multiple projects that require data science and domain science collaborations. TRIPODS+X presents an outstanding opportunity for assembling transdisciplinary teams that will harness vast amounts of data, developing novel methods and techniques in the process. However, it is very likely that the fate of many of these new collaborations will be similar to prior efforts, encountering many technical and non-technical barriers that hamper productivity when highly productive teams (with diverse expertise and computational backgrounds) work on common problems. *We need to instill a paradigm shift that improves productivity for transdisciplinary collaborations in today's open science, open data and open innovation age.* TRIPODS+X provides the essential *avenue* for a TRIPODS Commons where participating data scientists and domain scientists, while readily exchanging their ideas, data, code and best practices, can guide their transdisciplinary communities towards more productive paths. This proposal aims to initiate this paradigm shift through a series of *innovation lab* (IL) events that 1) facilitate and empower TRIPODS+X projects to collaborate and share their expertise beyond their teams and institutions, and 2) outline pragmatic pathways of data science for domain science and engineering that will lead to rapid, efficient and more productive interactions. Among tangible outcomes of the innovation lab project will be 1) TRIPODS Commons (TC), a nascent Data and Information commons for the TRIPODS+X community that leverages existing computational infrastructure, 2) on-line pedagogical resources that document exemplar data science and field X collaborations; and 3) standard operating procedures and strategic imperatives for a data-driven research infrastructure.

Intellectual Merit: The proposed two part IL will bring together participants that represent thought leaders and practitioners in data-driven science collaborations, including TRIPODS+X project teams. Topics of emphasis will intersect with and learn from Astronomy and Earth Sciences communities, both prolific data generation disciplines, with mature computational tools, data repositories, and savvy community of modelers and users eager to collaborate with data scientists for developing their next generation of analysis capabilities. This diverse group will provide unique and pragmatic perspectives for guiding outcomes. Disruptive technologies and sensors producing new data types pose significant challenges for transdisciplinary teams developing analysis methods for a rapidly evolving technology landscape. For contextualizing these challenges we will engage lidar (Light Detection and Ranging) users, a disruptive technology that has seen rapid adoption in Earth Sciences, vehicular guidance systems and beyond.

Broader Impacts: The ILs will provide a rare opportunity for TRIPODS+X participants to establish productive foundational best practices during early phase of their collaboration, learning by example from other projects and communities. The National Optical Astronomy Observatory (NOAO) and Earth Science Information Partners (ESIP) will benefit from receiving focused input and key capabilities essential for fostering data science driven projects for their communities. We will introduce participants to the comprehensive national cyberinfrastructure ecosystem available at no cost through NSF-funded centers, allowing participants to share, scale, and extend their ideas and projects through these venues. Prototype TC will be a cohesive platform to showcase and share research products and outcomes, enabling other communities to emulate and extend these capabilities; eventually becoming an avenue that provides visibility to the vibrancy and productivity for projects at all TRIPODS institutes. Through ILs, we will provide approaches and pathways for establishing successful transdisciplinary collaborations that enable teams to work across domains and institutional boundaries, and at scales essential for addressing the research, education, and advanced cyberinfrastructure needs outlined in NSF's 10 big ideas, especially for harnessing the data revolution.

Contents

1	Overview and Objectives	D-1
2	Innovation Labs Mission and Plan	D-2
2.1	The <i>Lemon</i> Innovation Lab	D-2
2.1.1	Target Participants	D-2
2.1.2	Activities	D-3
2.1.3	Expected Outcomes	D-3
2.2	The <i>Lemonade</i> Innovation Lab	D-4
2.2.1	Target Participants	D-4
2.2.2	Activities	D-4
2.2.3	Expected Outcomes	D-4
3	Envisaging the TRIPODS Commons	D-5
4	Science Areas of Emphasis	D-5
4.1	Astronomy	D-5
4.2	Earth Sciences	D-6
5	Broader Impacts	D-6
6	Prior Experience Organizing Workshops, Seminars, and Hackathons	D-7
7	Results from Prior NSF Support	D-8
8	Results from current TRIPODS project	D-8

1 Overview and Objectives

Central to establishing productive data-driven, transdisciplinary collaborations is the ability for teams of domain scientists (+X) to articulate the problem space and provide representative datasets and associated metadata to groups of data scientists (TRIPODS). These prototype datasets are used to refine novel methods and techniques for solving various collaborative and information-sharing challenges. Team members share outcomes with collaborators and develop a common understanding of the problem space, while creating solutions that can often benefit a much larger community. It is fairly common for domain scientists to independently explore the expansive options of open source tools, platforms and publications in search of appropriate methods and techniques; similarly, data scientists look for well-curated datasets for experimenting with their algorithms and methods. In fortunate circumstances, these two independent quests lead to crossing of paths which may lead to productive collaborations. More often, these journeys go down paths that are unproductive and fraught with frustration and friction at the boundaries of technology, communication and infrastructure, preventing the collaboration from realizing its full potential.

Significant barriers exist (both technical and non-technical), which hamper productivity for these transdisciplinary collaborations. Such barriers are often associated with obtaining or providing access to appropriate datasets in forms which can be readily consumed by tools and platforms that are familiar to the data scientists, as well as the ability of domain scientists to articulate the core computational challenges needing to be addressed for a given dataset.

These challenges in collaborations are further exacerbated when new data types and disruptive technologies are introduced, such as lidar (Light Detection and Ranging) or IoT (Internet of Things). Rapid adoption across disciplines, despite the lack of standardized formats and limited availability of canonical data repositories, also limits how transdisciplinary teams can collaborate and share their outcomes and products when working with these data types.

Datasets provided by domain scientists can be extremely large and can have inherent complexities that are not obvious when represented as smaller subsets. Using these datasets requires important cognitive overhead in terms of learning the domain specific vocabulary and problem space necessary for establishing meaningful collaborations. Improved access to training, awareness of technical computing, adoption of software engineering (and best practices of data management) have reduced computational infrastructure related friction encountered when domain and computational scientists collaborate.

Limitations associated with size and scale continue to pose significant barriers to adoption, use and exploration of datasets by the broader data science communities. These issues are further confounded when solutions that are collaboratively developed need to be scaled for real world (production) use. Many of these well-known issues are being resolved over time by advances in technologies such as cloud platforms (public, private) and containers (e.g., Docker, Singularity etc.) that are changing the way complex software is shared, deployed and scaled. This has allowed researchers to reproducibly share software they develop. Container based technologies facilitate packaging of the desired operating system and analysis programs developed in any language. Software dependencies are also packaged into portable containers that can scale from a researcher's own laptop to the cloud and HPC (High Performance Computing) systems with ease. Similarly, the evolution of web-based computational notebooks (e.g., Jupyter, Zepplin etc.) are making it easier to provide documentation narratives, original code, corresponding analysis steps and embedded visualizations. These advances allow collaborators to rapidly share, reproduce, and experiment with analyses by modifying parameters. Such capabilities are part of the essential technology stack for supporting contemporary analysis pipelines in industry and disciplines such as bioinformatics [23]. Unfortunately, these solutions have not seen wide-spread adoption in many academic disciplines.

Similar barriers and limitations of adoption will likely be encountered by the participants of the (soon to be funded) NSF TRIPODS+X projects. *Envisioning and prototyping of a Commons* that specifically targets the essential interactions between data science and domain sciences can greatly facilitate these collaborations. "Commons" is a term often used in reference to cultural and natural resources accessible to all members of a society. These resources are held in common (not owned privately) by communities that man-

age them for individual and collective benefit. Today, Data, Information and Knowledge commons include agency based resources that are domain specific such as the US Department of Agriculture [2] Ag Data Commons, National Institute of Health (NIH) Data Commons [7], NSF CyVerse Data Commons [4] and re-imagined libraries at many academic institutions. The TRIPODS+X initiative presents a unique opportunity to apply lessons learned from prior collaboration experiences between data science and X fields, explore the current state of the art for data science tools and platforms, and develop guidelines for best practices for current and future collaborations.

We propose a series of innovation-lab inspired events that will lay the foundation for a TRIPODS Commons (TC) that supports multiple communities and not constrained by domain specificity. Innovation Labs (IL) aim to engage diverse participants on a sustained open collaboration for the purpose of creating, elaborating, and prototyping radical solutions to pre-identified systemic challenges [27]. Specifically, we plan to bring the TRIPODS+X awardees community together with domain scientists from Astronomy and Earth Sciences with emphasis on lidar. Collectively these will represent the “X” for our TRIPODS+X proposal. The IL will have the following **Objectives**:

1. *Be an Enabling Space for the Success of TRIPODS+X Projects* through: 1) Augmenting the traditional habits of data science collaboration with domain science and engineering and incubating a new culture that facilitates communication and accelerates mutual growth. 2) Laying the foundation for a support infrastructure for these projects that is domain independent. 3) Encouraging new collaborations.
2. *Build Strategic Imperatives for Data-Driven Science* by going beyond the TRIPODS+X projects scope and attracting other funding agencies efforts on data-driven science that collectively seek to reshape how science is done in the 21st-century.

Our project objectives will be achieved mainly through two 4-day long events. Following the proverbial phrase “when life gives you lemons (collaboration difficulties), make lemonade (collaboration how-tos),” we will create a *Lemon* Innovation Lab and a *Lemonade* Innovation Lab and lay the foundation for a *Lemonade Stand* (TRIPODS Commons) for the community to coalesce around.

The **outcomes** of our proposed project are:

1. *TRIPODS Commons*. Prototype of a Data and Information commons for the TRIPODS+X community that leverages existing infrastructure.
2. *On-line resources for pragmatic utilization of data science in domain science and engineering*. Document the analysis steps for successful TRIPODS+X projects in forms of narratives and notebooks that are clear and easy to follow, much like how YouTube is utilized for user-contributed videos, providing step-by-step instructions for diagnosing problems and performing repairs.
3. *Data-driven research how-tos*. Standard Operating Procedure (SoP) checklists for domain scientists to prepare data and associated questions for data scientists, conversely checklists for sharing computational methods and results with domain scientists
4. *Strategic imperatives of scientific cyberinfrastructures*. A starting point for pursuing funding opportunities through NSF CSSI (Cyberinfrastructure for Sustained Scientific Innovation) planning grant as well as non traditional avenues such Earth Science Information Partners (ESIP) Idea Farm.

2 Innovation Labs Mission and Plan

2.1 The *Lemon* Innovation Lab

Collaborations between data scientists (computer scientists, statisticians, mathematicians) and domain scientists are not new. In the first innovation lab, we wish to examine the past carefully for both failures and successes in order to guide the new TRIPODS+X collaboration efforts. With this in mind, the primary **mission** of this innovation lab is to incubate a vibrant and collaborative culture in the TRIPODS+X community, and equip the participants with a set of procedures, tools, and common vocabulary to facilitate faster, smarter, and more efficient data-driven research.

2.1.1 Target Participants

This innovation lab will include representatives from:

- TRIPODS+X awardees from all three tracks, including representatives from domain and data sciences.

- Data scientists from Computer Information Science and Engineering (CISE) disciplines (*see Rajasekar letter of collaboration*).
- Domain experts and practitioners from Astronomy (*see Turk letter of collaboration*) and Earth Sciences (*see Russell letter of collaboration*) will complement the X domains from TRIPODS+X awardees.
- Data and computational infrastructure providers (e.g., NOAO, ESIP, XSEDE) and representatives from NIH and NIST Data Commons efforts. (*see Bolton, Robinson letter of collaboration*).
- Data scientists involved in developing methods for analyzing lidar data and practitioners that acquire and manage lidar data (*see Robinson letter of collaboration*).

2.1.2 Activities

Prior to the event, a series of webinars conducted using the Zoom conferencing system will introduce the participants to IL concept and help align expectations. The four day on-site event will be comprised of activities focused around:

1. *Claiming previous failures.* Invited guest speakers will discuss challenges encountered when establishing inter and transdisciplinary collaborations that prevented their projects from achieving their full potential (e.g., caBIG: Cancer Bioinformatics Grid, VAO: Virtual Astronomical Observatory). Participants will share with each other their versions of past failures.
2. *Learning by example.* Overview of how successful data driven projects have implemented their innovation, software and data life-cycle management, including use of methods such as idea sketching, mind maps and the role of contemporary software tools for improving productivity and communication.
3. *Bite-sized basics.* With the purpose of equipping participants with basics of data science and key science domains topics, a series of short (20–60 min) tutorials will be offered throughout the 4 days of the gathering. These tutorials will be followed with hands-on experimentations by the participants (e.g., turn a GitHub code repository into a collection of interactive notebooks using binder).
4. *Surveying the status quo.* Participants will survey the landscape of how various communities share ideas, tools and data and how they synthesize concepts, including overview of successful NSF Synthesis centers (NIMBioS, NCEAS, SESYNC, etc.) that manage transdisciplinary projects [16, 22]. Survey technologies and platforms that facilitate collaboration, sharing of data, tools and how communities are adopting them.
5. *Dream Platform Attributes.* Participants will be engaged to inquire and outline the key capabilities and features for a Commons that would support the research needs of the participants, define at least two use cases in detail for how the domain and data science groups would collaborate, identify two datasets that are challenging to manage and curate, identify two analysis pipelines that utilize these datasets.
6. *Premortem discussions.* Participants will be asked to imagine that their TRIPODS+X project has failed, and then work backward to determine what potentially could have lead to that failure. Define expected challenges and barriers to productivity and perform gap analysis.
7. *Emulating exemplars.* Participants will identify candidate procedures, tasks and projects (use cases) that utilize currently available resources, platforms, tools and technologies to define exemplar approaches for managing transdisciplinary data science collaborations.

Both ILs will have facilitator-led activities to ensure full engagement of the participants. The main facilitators will be Sondra LoRe, NIMBioS Evaluation Associate at the National Institute for STEM Evaluation and Research, and Kimberly Eck, Director of the Research Development team at the Office of Research and Engagement at the University of Tennessee.

2.1.3 Expected Outcomes

1. Standard Operating Procedure (SOP) checklist for domain scientists to prepare data and associated questions for data scientists, conversely a checklist for sharing computational methods and results with domain scientists.
2. Techniques for developing a minimal framework of shared vocabulary, terminologies and key con-

cepts for accelerating the initial learning process.

3. A paper following the “PLOS 10 simple rules” guidelines [1] to create “10 simple rules for working productively with transdisciplinary data science teams.”

2.2 The *Lemonade* Innovation Lab

The primary **mission** of the *Lemonade* IL is to transform the guidelines and ideas from the *Lemon* IL into usable and tangible deliverables. This IL, while intended to be pragmatic, will also push the participants to explore concepts for the next generation of scientific research infrastructures (for Data Science projects), as well as the technical underpinnings for creating a *Science Analytics* platform.

2.2.1 Target Participants

A subgroup of 15 participants from the *Lemon* IL representing data science, engineering, and domain sciences will form the core of this second IL. Additionally, local participants (from University of Tennessee and Oak Ridge National Laboratory) will be included in the lab, to enhance the representation of X domains. We will also extend invitations to NSF program officers outside of the TRIPODS program, as well as from other agencies such as DoD, DoE, DHS, NIH with interest in this space.

2.2.2 Activities

This 4-day gathering will be centered on:

1. *Case study review*: The collaborations and case studies developed after the *Lemon* IL will be discussed to identify main breakthroughs and breakpoints. Participants then create a showcase of exemplar projects.
2. *Testing the SOP and 10 simple rules*: The participants will be assigned to subgroups (each with at least one representative of data science, engineering, and domain sciences) and will work on new transdisciplinary research questions, using the SOP and the “10 simple rules” produced by the *Lemon* IL.
3. *Prototyping the TRIPODS Commons*: Participants will perform a gap analysis for the Dream Platform and current available technologies and resources, to examine use cases from prior IL. Participants will propose the architecture for TC, which utilizes existing infrastructure resources available through CyVerse, Jetstream, other NSF XSEDE centers and commercial cloud. Furthermore, as part of a hackathon, participants will prototype key features of TC.
4. *Usability*: In this hands-on activity, participants will curate nominated datasets using prototypes, and build computational notebooks and analysis, and assess the usability and value of the platform. External users will also be surveyed.
5. *Proposal development and competition to envision Sustained Cyberinfrastructure*: Participants in teams develop 1-page proposals (5 minute pitch) and get immediate feedback from the attending program officers. Finally, participants document the outcomes and prepare material for multiple NSF CSSI (Cyberinfrastructure for Sustained Scientific Innovation) planning grants with domain emphasis in Data Science, Astronomy and Earth Sciences.
6. *Envisioning a Science Analytics platform*. Data science has made its way through private sectors. Many business analytics platforms exist (e.g., Google data studio, Domo) where companies can readily analyze data of interest and get insights into patterns and trends, *without the user having to know the nuts-and-bolts of the underlying platform*. The scientific research community lacks such platforms and it is worth exploring what the next generation of Science Analytics would look like and how do we get there.

2.2.3 Expected Outcomes

1. A visioning paper that summarizes the efficiency of SoP and “10 simple rules” and outlines the desired capabilities for a platform (TRIPODS Commons) that fosters productive transdisciplinary collaboration and can serve as the starting point for NSF CSSI for the TRIPODS community.
2. A series of papers specific to domain sciences (X) explored during the innovation labs that inform the researchers about best practices from the data science perspective, as well as the analytic tools and workflows that have broader applications and can standardize methodology.
3. Curated datasets from the Earth Science and Astronomy and appropriate TRIPODS+X projects that

do not have a canonical repository and are large or complex to manage. Making these datasets meet the basic FAIR (Findable, Accessible, Interoperable and Re-usable) principles [44], create well-documented notebooks that support exploratory analysis and visualization.

4. A collection of proposal ‘pitches’ that will be shared with funding agencies program officers.

3 Envisaging the TRIPODS Commons

Open data mandates by funding agencies have paved the way for domain scientists to deposit their final research products into canonical repositories, such as NCBI. For many science domains there are no canonical data repositories and this leads to lost and dark data [29]. The problem is becoming more prominent as data science (DS) and machine learning (ML) become more and more important in all areas of science. Repositories do not usually support work-in-progress data as they are designed for end products, thus forcing collaborators to rely on free or institutionally-provided services (e.g., Box, Dropbox) which primarily cater to document and photo sharing and do not lend themselves to scientific datasets that need size, scale, metadata or streaming capabilities. There are limited avenues to create or retrieve well-curated datasets associated with ML/DS friendly analysis platforms that can operate at scale, to which researchers from any domain (+X) can readily contribute datasets. For example, the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/>) provides access to nominally curated datasets geared towards the data science community, but with a 200Mb limit and without analysis capabilities [8]. Platforms such as Kaggle (<https://www.kaggle.com/datasets>) allow computational notebooks to be associated with the datasets, although also limiting the datasets to 10Gb or 50 files [6]. Current domain science datasets are significantly larger and have inherent complexity making them unsuitable for these platforms.

The aim of the TRIPODS Commons is an amalgamation of best practices and capabilities that overcome current limitations and support the agile usage patterns necessitated by transdisciplinary projects operating at scale. This includes issues such as securely sharing data and tools among collaborators at any stage of the analysis life cycle and making outcomes publicly available regardless of size and format. Other activities of the TRIPODS Commons include creating a series of online tutorials and how-tos from the IL prototyping efforts. These also would be disseminated as part of hands-on sessions at the annual TRIPODS workshops conducted by each TRIPODS institute. This content would be developed and delivered by team of TRIPODS postdocs and graduate students that are well-versed in the use of these capabilities.

4 Science Areas of Emphasis

To frame and guide the innovation lab activities, we will rely on two mature data-driven disciplines – Astronomy and Earth Sciences – and focus on specific topics to highlight the strengths, weakness, opportunities and threats (SWOT) for transdisciplinary teams working in these disciplines.

4.1 Astronomy

Recent advancements in digital electronics and computing hardware allow astronomers to gather data at a very large scale. By way of example, in order to take the first pictures of black holes, the Event Horizon Telescope (EHT) coordinates many radio telescopes around the globe and collects up to 10 petabytes (PB) of raw data during each of its observation runs [43]. Similarly, the Large Synoptic Survey Telescope (LSST) will soon collect 15 terabytes (TB) of raw data every night, survey the full sky every 3 nights, and accumulate 500 PB of processed data during its 10 years lifespan [31]. As these PB-scale datasets become the norm, astronomy enters the Fourth Paradigm of data-intensive research and data-enabled discovery, with growing reliance on ML/DS tools and analysis.

In the absence of the proposed Commons, the EHT is sharing TBs of its numerical simulations through the cloud infrastructure of CyVerse in order to train machine learning algorithms to automate and speed up part of its data process pipeline. As another example, although astronomers have been sharing and interoperating observational data through virtual observatories (VO) for decades, they were designed for and are limited to small datasets. The DataLab at the National Optical Astronomy Observatory (NOAO) is an example of a nascent Commons bringing data and computational resources together. As a final example, both CyVerse and the NSF-funded Partnerships for International Research and Education (PIRE) program started to host hackathons, webinars, and workshops to train the next generation astronomers to use cloud

computing tools and data science techniques. With this in mind, we will ensure that IL participants from EHT, LSST and NOAO will share their experience and knowledge and help the proposed Commons achieve its goals.

4.2 Earth Sciences

Remote sensing is currently experiencing a renaissance. As sensor technology, e.g. lidar and hyperspectral imaging, as well as techniques like structure from motion and multi-view stereo (SfM-MVS) mature, ever larger quantities of data are generated. For example, the ArcticDEM Project [3] used satellite image stereo pairs and SfM-MVS to generate high resolution multi-temporal digital elevation models over Earth's polar regions, requiring over a billion CPU hours on HPC to process, creating over 500 TB output data (to date). Privately owned cubesats (<https://planet.com>) image the entire terrestrial Earth surface daily at high spatial resolution (+7 TB/day). Hyperspectral cameras along with lidar have been miniaturized to fly on small unmanned aerial systems (sUAS) ([42], [41]), producing terabytes of data over areas less than a hectare with daily repetition.

Fine-tuning of technology and broad adoption in engineering have made lidar and hyperspectral instruments more affordable and appealing to earth scientists. However, the fast pace of technological advancement and its adoption by various domain sciences has not been paralleled by robust development of analytical tools to deal with the volume and veracity of the data. An array of tools and platforms are scattered across domains, creating a difficult and time-consuming maze to be navigated by domain scientists that sense the environment. Data have variable spatial densities, e.g. low density spaceborne and manned aircraft collections to extremely dense scans with multi-temporal repetition from terrestrial, fixed locations or SLAM (Simultaneous Localization And Mapping). Programs like the National Ecological Observatory Network (NEON), National Center for Airborne Laser Mapping (NCALM), and USGS 3D Elevation Program (3DEP) are producing lidar and hyperspectral datasets too large to move over conventional distributed networks, including Internet2. These multi-petabyte archives reside on public services like NSF OpenTopography, or Amazon Web Services, where users can query subsetted extents, and run jobs on HPC or cloud. Importantly, in the coming years new producers of data, i.e. sUAS, self-driving vehicles, and a cadre of increasingly affordable portable lidar instruments, portend an explosion of data which will dwarf what is currently being collected and shared by national research groups. Making long tail data findable, accessible, interoperable, and reusable, the so-called FAIR data principles [44], will require educating the scientific lidar community toward adopting integrated data science practices from the onset of their research until its terminus.

5 Broader Impacts

IL based approach provides a rare opportunity and an avenue for students and participants at TRIPODS centers along with their domain collaborators to establish a strong foundation, ensuring productive collaboration, by learning from each others efforts and sharing best practices that can be readily adopted by other communities. Through IL we will also introduce participants to national cyberinfrastructure resources that are available at no cost through NSF-funded centers, allowing them to scale and extend their projects.

The prototype TC will be a cohesive avenue for interested participants to showcase and share their approaches and outcomes, allowing other communities to extend and utilize these resources; eventually becoming a platform that can demonstrate the vibrancy and productivity of all TRIPODS institutes. IL will provide approaches for establishing successful transdisciplinary collaboration and the ability to work across domains and institutional boundaries, especially at scales demanded by multiple disciplines, which is absolutely essential and required for addressing the research, education and advanced cyber-infrastructure needs outlined in multiple NSF 10 big ideas, especially for harnessing the data revolution. The National Optical Astronomy Observatory (NOAO) and Earth Science Information Partners (ESIP) representatives will share their experiences with using distributed and shared infrastructure and their vision for a Commons, they will also benefit from receiving focused input and key capabilities essential for fostering data science driven projects for their communities. The proposed IL at a greater view will be among rigorous efforts to transform how science is conducted in the 21st century. As a result, this project has the potential to enhance

and facilitate domain sciences progress, and each and everyone will serve the nation's progress.

Outreach Efforts: The second year of the project will emphasize outreach efforts to bring the outcomes of the two IL and the TRIPODS Commons to the attention of other universities by IL participants, and to present them at TRIPODS summer workshops by co-PI Sahneh. With NOAO and ESIP partnerships we will share the outcomes utilizing their working groups and events. Participation of underrepresented minorities are valued and encouraged and such issues will be taken under consideration when assembling participant lists. The South Dakota (SD) NSF EPSCoR (Established Program to Stimulate Competitive Research) program will assist to identify suitable participants for IL, and project PIs will present IL and TC outcomes at annual SD EPSoR symposiums and webinars (*see Lushbough letter of collaboration*)

6 Prior Experience Organizing Workshops, Seminars, and Hackathons

Merchant as part of CyVerse (formerly iPlant Collaborative) and as the founding member of the Software Carpentry Foundation (SCF) has designed multiple bootcamps and workshops for teaching technical computing fundamentals to scientists. He has developed hands-on workshops for teaching cloud computing concepts for scalable analysis, content that is central to many advanced domain specific tutorials for CyVerse and Jetstream. Container Camp for reproducible science was co-developed with senior personnel Swetnam and Upendra Devisetty and has been extended and customized for multiple disciplines, including Astronomy and for data scientists as part of the UA-TRIPODS summer workshop (May 21-24, 2018) conducted along with co-PI Sahneh. Merchant has extensive experience conducting hackathons and code sprints along with technical staff and scientists from projects such as iRODS, HTCCondor, Openstack, DataOne etc. for rapidly prototyping and integrating complex software tools into CyVerse and developing new community standards. With senior personnel Walls he is responsible for the CyVerse Data Commons. He has coordinated global hackathon events (17 countries) such as the The #Great Antarctic Climate Hack along with Joellen Russell.

Papes has taught an NSF NIMBioS funded, 2.5 days tutorial on spatial data applications, specifically for ecological niche modeling; was a member of a multinational team funded by the Global Biodiversity Information Facility to organize and teach four one-week workshops for researchers from 28 countries around the world interested in biodiversity informatics and ecological niche modeling techniques; co-led a similar course in Manaus, Brazil, for 22 Brazilian students; led a niche modeling course focused on disease ecology applications for the agriculture ministry in Lima, Peru; was instructor for Conservation Implementation, a two-week course in Ethiopia funded by JRS Biodiversity Foundation, with participants from seven African countries. Several events in the past 12 months at University of Tennessee have initiated contacts among domain scientists and to a lesser degree data scientists and engineers. These events include a monthly UAS brown bag lunch series, a monthly lidar discussion group, both organized by the Spatial Analysis Lab, and a SPARKS (Seeking Partnerships for Research and Knowledge) event organized by Office of Research Engagement.

Kobourov has organized six Dagstuhl Seminars, including several multi-disciplinary events where the main goal was to bring together researchers from different areas. "Beyond-Planar Graphs: Algorithmics and Combinatorics (Dagstuhl Seminar 16452) included mathematicians, theoretical computer scientists, and information visualization experts. "Drawing Graphs and Maps with Curves (Dagstuhl Seminar 13151) included mathematicians, geographers, and computer scientists, which lead to a Dagstuhl exhibition "Bending Reality: Where Arc and Science Meet." Such seminars enable the transfer of knowledge between theory and applications, tapping into new fields of application, and fostering the next generation of researchers by including them in the research dialogue. They are great for establishing new connections, initiating new collaborations, and broadening one's academic social network.

Sahneh organized the 1st TRIPODS Southwest Conference (BioSphere 2, Tucson, AZ, May 21-24, 2018). He also served as a workshop organizer, Frontiers in Multiscale Computational Modeling for Zoonotic Epidemics (Kansas City, MO, Oct. 10-12, 2011), where two groups of mathematical/computational modelers and epidemiologists/public health experts gathered together.

The Team has a history of successful collaborations by the PIs, such as theoretical work on multi-level representation of networks [9] and practical work on visualizing global trends in research [20].

7 Results from Prior NSF Support

Kobourov has been PI or co-PI on several NSF projects, including being PI on a project that originated in one of the earliest NSF Innovation Labs. More details are provided about the project **Algorithms for Visualizing Data with Contact Graphs (Award #1115971, 2011–15, \$296,001)**. **Intellectual Merit:** This project led to several publications of theoretical nature [12–15, 21, 24, 26, 32, 33] and with practical applications [10, 11, 17–19, 25, 30]. **Broader Impacts:** The design and implementation of algorithms for contact representations has also led to software systems, including the Graphs-to-Maps (GMap) framework [30], Maps of Computer Science [25], and WordCloud [18], all available as functioning online tools and as open-source software. The PhD theses of Jawaherul Alam and Sankar Veeramoni were based on this work. Several undergraduates involved in this project won awards, including Daniel Fried (Churchill Scholarship, IBM Watson Scholarship, Outstanding University of Arizona College of Science) and Katie Cunningham (Outstanding Senior UA College of Science).

Merchant has been co-PI on multiple NSF projects which involve community-driven platform development for large scale analytics and data management, as well as establishing consortia and federations to share these capabilities. **LIMPID: Large-Scale Image Processing Infrastructure Development (Award # 1664172, 2017–22, \$3,400,000)**; **High Performance Computing System Acquisition: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment (Award #1445604, 2014–18, \$11,836,981)**; **DataNet Full Proposal: DataNet Federation Consortium (Award #0940841, 2011–16, \$8,300,992.00)**. The project most relevant to this effort is: **The iPlant Collaborative: Cyberinfrastructure for the Life Sciences (Award #1265383, 2013–18, \$50,300,000)**. **Intellectual Merit:** Design, implement and maintain an extensible infrastructure for supporting data-driven collaborations for diverse communities. **Broader Impact:** CyVerse (formerly iPlant Collaborative) is a cross-disciplinary project initially designed to build scalable computational infrastructure for the plant sciences. It was given an expanded mandate to support the broader life science community by providing access to high performance computational resources, large-scale data management systems and cross-disciplinary collaborations. Results include community extended platforms for conducting scalable data analysis using cloud and distributed computing infrastructure (MAKER as a Service: Moving HPC applications to Jetstream Cloud), extending into disciplines beyond life sciences such as Astronomy [28] and Geosciences [40]

Sahneh has been a co-PI on the NSF project titled **CIF: Small: Spreading Processes over Multilayer and Interconnected Networks (Award #1423411, 2014–18, \$522,042)**. **Intellectual merit:** This project advances the boundaries of network theory by analyzing multilayer and interconnected networks. Results include characterization of structural transition phenomenon in interconnected networks [35] for diffusion process; optimal interconnection [39]; and several other publications [36–38]. **Broader impacts:** developed and publicly posted GEMFsim [5], a simulation toolbox for spreading processes on multilayer networks based on [34, 37], also incorporated into the curriculum for Network Theory class ECE841 at Kansas State University.

Papes is a new NSF investigator and has not received prior NSF funds.

8 Results from current TRIPODS project

Kobourov is a co-PI on **UA-TRIPODS Building Theoretical Foundations for Data Sciences, NSF-DMS-1740858, 2017-2020, \$1,360,000**. He leads a research group with a focus on large scale graphs. The first peer-reviewed publication with co-authors in Computer Science, Mathematics, and Statistics appeared within the first 9 months of the start of this project [9]. The lead two authors are a PhD student, R. Spence, and a postdoc, F. Sahneh, both supported by the TRIPODS project. UA-TRIPODS organizes a regular seminar series inviting researchers from the UA campus, other universities and industry with topics relevant for UA-TRIPODS member.

References

- [1] 10 Simple Rules Collections, Public Library of Science. <http://collections.plos.org/ten-simple-rules>.
- [2] Ag Data Commons, United States Department of Agriculture. <https://data.nal.usda.gov/>.
- [3] ArcticDEM Project. <https://www.pgc.umn.edu/data/arcticdem/>.
- [4] CyVerse Data Commons. <http://www.cyverse.org/data-commons>.
- [5] GEMFsim. <http://www.ece.k-state.edu/netse/software/index.html>.
- [6] Kaggle. <https://www.kaggle.com/about/datasets/publish>.
- [7] NIH Data Commons, National Institute of Health. <https://commonfund.nih.gov/commons>.
- [8] UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets.html>.
- [9] A. R. Ahmed, P. Angelini, F. D. Sahneh, A. Efrat, D. Glickenstein, M. Gronemann, N. Heinsohn, S. G. Kobourov, R. Spence, J. Watkins, and A. Wolff. Multi-level Steiner trees. In *17th Symposium on Experimental Algorithms (SEA)*, 2018.
- [10] J. Alam, T. Biedl, S. Felsner, M. Kaufmann, S. G. Kobourov, and T. Ueckerdt. Computing cartograms with optimal complexity. In *28th ACM Symposium on Computational Geometry (SoCG)*, pages 21–30, 2012.
- [11] J. Alam, T. C. Biedl, S. Felsner, A. Gerasch, M. Kaufmann, and S. G. Kobourov. Linear-time algorithms for hole-free rectilinear proportional contact graph representations. *Algorithmica*, 67(1):3–22, 2013.
- [12] J. Alam, T. C. Biedl, S. Felsner, M. Kaufmann, and S. G. Kobourov. Proportional contact representations of planar graphs. *J. Graph Algorithms Appl.*, 16(3):701–728, 2012.
- [13] J. Alam, D. Eppstein, M. Kaufmann, S. G. Kobourov, S. Pupyrev, A. Schulz, and T. Ueckerdt. Contact graphs of circular arcs. In *Algorithms and Data Structures Symposium (WADS)*, pages 1–12, 2015.
- [14] J. Alam, W. Evans, D. Eppstein, S. G. Kobourov, S. Pupyrev, J. Toeniskoetter, and T. Ueckerdt. Contact representations of non-planar graphs. In *Algorithms and Data Structures Symposium (WADS)*, pages 14–27, 2015.
- [15] J. Alam and S. G. Kobourov. Proportional contact representations of 4-connected planar graphs. In *20th Symposium on Graph Drawing (GD)*, pages 211–223, 2012.
- [16] J. S. Baron, A. Specht, E. Garnier, P. Bishop, C. A. Campbell, F. W. Davis, B. Fady, D. Field, L. J. Gross, S. M. Guru, et al. Synthesis centers as critical research infrastructure. *BioScience*, 67(8):750–759, 2017.
- [17] L. Barth, S. Fabrikant, S. G. Kobourov, A. Lubiw, M. Nöllenburg, Y. Okamoto, S. Pupyrev, C. Squarcella, T. Ueckerdt, and A. Wolff. Semantic word cloud representations: Hardness and approximation algorithms. In *11th Latin American Theoret. Inform. Symp. (LATIN)*, pages 514–525, 2014.
- [18] L. Barth, S. G. Kobourov, and S. Pupyrev. Experimental comparison of semantic word clouds. In *13th Symposium on Experimental Algorithms (SEA)*, pages 247–258, 2014.
- [19] M. Bekos, T. van Dijk, P. Kindermann, S. G. Kobourov, S. Pupyrev, J. Spoerhase, and A. Wolff. Improved approximation algorithms for box contact representations. In *22nd European Symposium on Algorithms (ESA)*, pages 87–99, 2014.

- [20] R. Burd, K. A. Espy, M. I. Hossain, S. G. Kobourov, N. Merchant, and H. C. Purchase. GRAM: global research activity map. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI 2018, Castiglione della Pescaia, Italy, May 29 - June 01, 2018*, pages 31:1–31:9, 2018.
- [21] S. Chaplick, S. G. Kobourov, and T. Ueckerdt. Equilateral L-contact graphs. In *39th Workshop on Graph-Theoretic Concepts in Computer Science (WG)*, pages 139–151, 2013.
- [22] K. Crowston, A. Specht, C. Hoover, K. M. Chudoba, and M. B. Watson-Manheim. Perceived discontinuities and continuities in transdisciplinary scientific working groups. *Science of The Total Environment*, 534:159–172, 2015.
- [23] F. da Veiga Leprevost, B. A. Grüning, S. Alves Aflitos, H. L. Röst, J. Uszkoreit, H. Barsnes, M. Vaudel, P. Moreno, L. Gatto, J. Weber, et al. Biocontainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, 33(16):2580–2582, 2017.
- [24] C. Duncan, E. Gansner, Y. Hu, M. Kaufmann, and S. G. Kobourov. Optimal polygonal representation of planar graphs. *Algorithmica*, 63(3):672–691, 2012.
- [25] D. Fried and S. G. Kobourov. Maps of computer science. In *7th IEEE Pacific Visualization Symposium (PACIFICVIS)*, pages 113–120, 2014.
- [26] E. Gansner, Y. Hu, and S. G. Kobourov. On touching triangle graphs. In *18th Symposium on Graph Drawing (GD)*, pages 250–261, 2010.
- [27] L. Gryszkiewicz, I. Lykourantzou, and T. Toivonen. Innovation labs: leveraging openness for radical innovation? 2016.
- [28] A. Haug-Baltzell, J. R. Males, K. M. Morzinski, Y.-L. Wu, N. Merchant, E. Lyons, and L. M. Close. High-contrast imaging in the cloud with klipreduce and findr. In *Software and Cyberinfrastructure for Astronomy IV*, volume 9913, page 99130F. International Society for Optics and Photonics, 2016.
- [29] P. B. Heidorn. Shedding light on the dark data in the long tail of science. *Library trends*, 57(2):280–299, 2008.
- [30] Y. Hu, E. Gansner, and S. G. Kobourov. Visualizing graphs and clusters as maps. *IEEE Computer Graphics and Applications*, 30(6):54–66, 2010.
- [31] M. Jurić, J. Kantor, K. Lim, R. H. Lupton, G. Dubois-Felsmann, T. Jenness, T. S. Axelrod, J. Aleksić, R. A. Allsman, Y. AlSayyad, J. Alt, R. Armstrong, J. Basney, A. C. Becker, J. Becla, S. J. Bickerton, R. Biswas, J. Bosch, D. Boutigny, M. Carrasco Kind, D. R. Ciardi, A. J. Connolly, S. F. Daniel, G. E. Daues, F. Economou, H.-F. Chiang, A. Fausti, M. Fisher-Levine, D. M. Freemon, P. Gee, P. Gris, F. Hernandez, J. Hoblitt, Ž. Ivezić, F. Jammes, D. Jevremović, R. L. Jones, J. Bryce Kalmbach, V. P. Kasliwal, K. S. Krughoff, D. Lang, J. Lurie, N. B. Lust, F. Mullally, L. A. MacArthur, P. Melchior, J. Moeyens, D. L. Nidever, R. Owen, J. K. Parejko, J. M. Peterson, D. Petravick, S. R. Pietrowicz, P. A. Price, D. J. Reiss, R. A. Shaw, J. Sick, C. T. Slater, M. A. Strauss, I. S. Sullivan, J. D. Swinbank, S. Van Dyk, V. Vujčić, A. Withers, P. Yoachim, and f. t. LSST Project. The LSST Data Management System. *ArXiv e-prints*, Dec. 2015.
- [32] S. G. Kobourov, D. Mondal, and R. I. Nishat. Touching triangle representations for 3-connected planar graphs. In *20th Symposium on Graph Drawing (GD)*, pages 199–210, 2012.
- [33] S. G. Kobourov, T. Ueckerdt, and K. Verbeek. Combinatorial and geometric properties of planar Laman graphs. In *24th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1668–1678, 2013.

- [34] F. D. Sahneh, C. Scoglio, and P. Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Transactions on Networking*, 21(5):1609–1620, 2013.
- [35] F. D. Sahneh, C. Scoglio, and P. Van Mieghem. Exact coupling threshold for structural transition reveals diversified behaviors in interconnected networks. *Physical Review E*, 92(4):040801, 2015.
- [36] F. D. Sahneh, A. Vajdi, and C. Scoglio. Delocalized epidemics on graphs: A maximum entropy approach. In *2016 American Control Conference (ACC)*, pages 7346–7351, July 2016.
- [37] F. D. Sahneh, A. Vajdi, H. Shakeri, F. Fan, and C. Scoglio. Gemfsim: A stochastic simulator for the generalized epidemic modeling framework. *arXiv preprint arXiv:1604.02175*, 2016.
- [38] C. M. Scoglio, C. Bosca, M. H. Riad, F. D. Sahneh, S. C. Britch, L. W. Cohnstaedt, and K. J. Linthicum. Biologically informed individual-based network model for rift valley fever in the us and evaluation of mitigation strategies. *PloS one*, 11(9):e0162759, 2016.
- [39] H. Shakeri, N. Albin, F. D. Sahneh, P. Poggi-Corradini, and C. Scoglio. Maximizing algebraic connectivity in interconnected networks. *Physical Review E*, 93(3):030301, 2016.
- [40] T. Swetnam, J. Pelletier, C. Rasmussen, N. Callahan, N. Merchant, E. Lyons, M. Rynge, Y. Liu, V. Nandigam, and C. Crosby. Scaling gis analysis tasks from the desktop to the cloud utilizing contemporary distributed computing and data management approaches: A case study of project-based learning and cyberinfrastructure concepts. In *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*, page 21. ACM, 2016.
- [41] T. L. Swetnam, J. K. Gillan, T. T. Sankey, M. P. McClaran, M. H. Nichols, P. Heilman, and J. McVay. Considerations for achieving cross-platform point cloud data fusion across different dryland ecosystem structural states. volume 8, page 2144, 2018.
- [42] S. T. T., M. Jason, S. T. L., M. M. P., H. Philip, and N. Mary. Uav hyperspectral and lidar data and their fusion for arid and semiarid land vegetation monitoring. volume 4, pages 20–33.
- [43] L. Vertatschitsch, R. Primiani, A. Young, J. Weintroub, G. B. Crew, S. R. McWhirter, C. Beaudoin, S. Doeleman, and L. Blackburn. R2DBE: A Wideband Digital Backend for the Event Horizon Telescope. *PASP*, 127:1226, Dec. 2015.
- [44] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.