# 

# 

# 

# 

# **Volume I: Technical Proposal**

# 

# **I. Abstract**

EcoHealth Alliance (EHA) proposes a Global Rapid Identification Tool System (GRITS) to deliver disease diagnostics, decision support, and data processing to the BSVE. We aim to enhance our current GRITS platform, developed with DTRA support, by scaling this system to handle large data volumes enhancing diagnostic capabilities through network and cluster analysis. GRITS will also utilize the benefits and explore the integration of crowdsourcing, collective intelligence, and expert review. This tool will rely on automation to ingest media, extract key disease characteristics, and recommend resources, increasing the specificity of the data feed to an analyst’s workflow. We propose building an app with the BSVE SDK that allows BSVE users to submit resources to diagnose and to connect with experts from ProMED, HealthMap and EcoHealth, thereby increasing our network of experts from digital surveillance and open source communities. The diagnostic results generated by GRITS have immediate applications for risk mapping, outbreak prediction, and early warning of microbial threats, advancing our readiness to combat the broad class of chemical and biological threats posed by EIDs.

**Keywords**

disease ecology, digital diagnostics, natural language, machine learning, emergent media, network, data science

# **II. Scope**

## A. Objective

EcoHealth Alliance (EHA) proposes a Global Rapid Identification Tool System (GRITS) for diagnosing infectious disease events. Our current GRITS media diagnostic tool (GRITS.md) delivers ranked differential diagnoses to pinpoint and identify disease threats more rapidly than current public health systems. Our specific objective with this proposal is to leverage the GRITS platform, developed with DTRA support since January 2013, and expand it to deliver disease diagnostics, decision support, and data processing to the BSVE. These enhanced capabilities will be powered by GRITS analytics, visualizations, and data; all connected to our network of experts from digital surveillance and open source communities.

## B. Background

The discovery of HIV/AIDS in the 1980s marked the transition from declaring victory over infectious diseases, to global increases in disease emergence and re-emergence1. EcoHealth Alliance (EHA) is at the forefront of organizations working to ‘get ahead of the epidemic curve’ by identifying these threats before the next pandemic or extinction event. Our researchers pioneered the identification of the origins of pathogens, such as Nipah virus2, SARS3, and MERS4. We constantly seek innovative approaches to target our field surveillance efforts on emerging threats5. To this end, the global increase in the volume of data, from the growth of the web and instrumented systems, presents both a challenge to traditional surveillance approaches and a tremendous opportunity for novel discoveries.

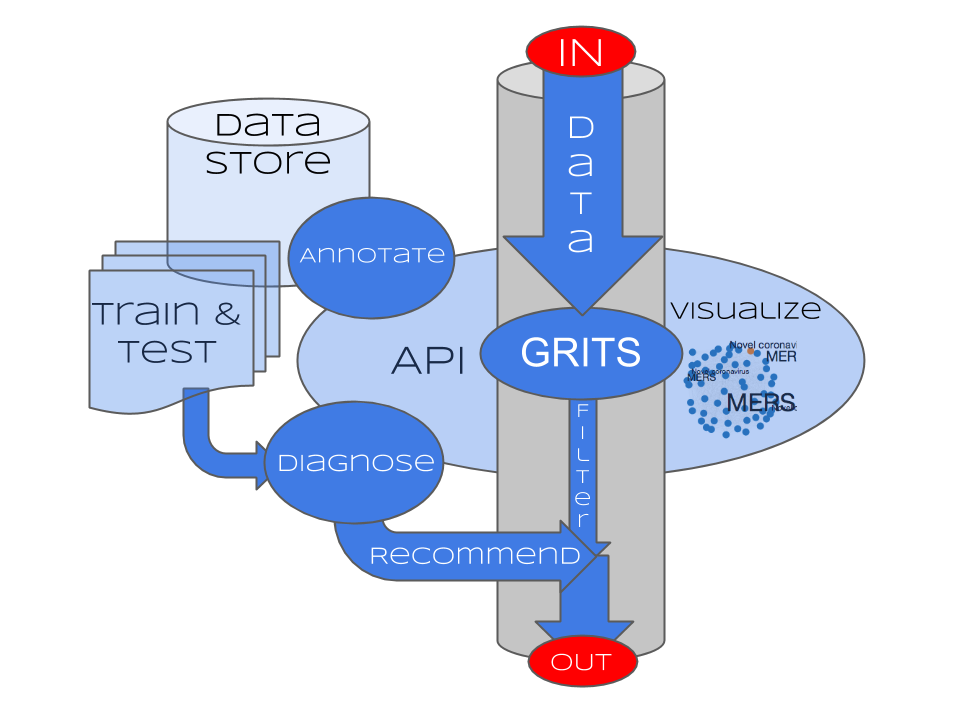
Various biosurveillance technologies have been developed to monitor Internet data sources, for instance, syndromic surveillance initiatives target surrogate indicators of a disease outbreak6. As with other sectors, such as commerce, marketing, and finance, the challenge is to manage the expanding volume of data while identifying signals of interest7. In the case of EIDs, access to emergent media, including participatory, personal, and interactive media, characterized by decentralized content generation (e.g., the blogosphere, internationalization, and social media), presents an opportunity to detect early instances of disease characteristics and anomalies of interest.

The GRITS partners are among the leading organizations in this domain. ProMED-mail (International Society of Infectious Diseases) manages a global email network of clinicians who are often among the first to identify and report disease threats8. Epidemico (HealthMap)9 actively curates an expanding catalog of relevant news and social media assets. Both organizations are unique among their peers in leveraging broad networks of experts to curate and reduce the volume of media to a high-quality feed. Kitware Inc., our technical partner, has engineered sustainable communities around successful, high-impact open source scientific software.

EcoHealth assembled the GRITS team to advance the state-of-the-art in the detection and diagnosis of disease threats. GRITS combines new technologies with expert networks to apply the sciences of disease ecology and epidemiology to diagnosing big data for near-real-time disease situation awareness. We built GRITS upon pre-existing communities of expertise, rather than *de novo* technical solutions, recognizing that human experts must be integrated into the platform to ensure its intelligence and growth. In addition, we sought novel mechanisms for decision support by organizing, prioritizing, contextualizing, and linking information to relevant current and historic resources. Overall, we deploy science at the core of our systems to assemble near-real-time information in a manner that supports decision makers with diagnostic tools built by, and for, the digital disease ecology community. The diagnostic results generated by GRITS have immediate applications for risk mapping, outbreak prediction, and early warning of microbial threats.

|  |
| --- |
| **Description of GRITS Capabilities**  We take a diagnostic approach to identifying disease threats, setting research priorities, and strategically deploying our field teams. GRITS ingests and processes data feeds to provide decision support to analysts, with the following capabilities:  **Diagnose** - Identify the disease(s) described in a resource. Return a ranked list of diseases with quantitative metrics of certainty (differential diagnoses).  **Mine** - Extract the key components of document via automated analysis, collective annotation, and expert curation. Return a set of relevant information.  **Recommend** - Expand the materials available to the analyst. Return a collection of recommended resources that provide historic and contemporary context.  **Filter** - Reduce the complexity of the data feed by selecting those documents that meet diagnostic criteria of interest to the analyst. Return a filtered subset of data.  **Connect** - Relate the material to underlying ontologies to provide a decision framework for the analyst. Return connections to other bioevents based on underlying network structures (e.g., geography, ecology, time, host, pathogen, and environmental drivers). |

### 



### Sources of GRITS Biointelligence

|  |  |
| --- | --- |
| **A. Machine** | Data mining, machine learning, and natural language processing |
|  |  |
| **C. Collective** | Crowdsourcing annotation and human intelligence tasks |

### A. Machine

GRITS leverages automation to ingest, process, and return an initial diagnosis of digital media, reducing the extraneous sources of information through which experts and reviewers must sort. The GRITS text mining system extracts key disease characteristics, such as locations, case counts, and dates, for our metadata and diagnostic models. These features, which may be combined into composite features by high-level rules, are based on sentence patterns and keywords chosen from third-party sources such as WordNet, Geonames, Disease Ontology, Symptom Ontology, and Biocaster Ontology. The categories we extract, such as hosts, pathogens, diseases, signs and symptoms, drivers, and transmission types, are also informed by historic disease event data curated via our GRID project, and prioritized through consultation with our experts. The GRITS diagnostic models, that use machine-learning algorithms based on extracted keywords to classify articles, are trained on test articles labeled with diseases by our partners. A key function of this machine intelligence is to provide the materials and platforms to crowd source the training data.

### C. Collective

The machine-learning component of GRITS requires training data from portfolios of articles with disease labels and disease characteristic metadata, information collected from HealthMap and ProMED editors to highlight and categorize important features and relationships. These are then used to train classifiers to identify specific features of importance. To extend the range of articles we can classify, we propose using crowdsourcing methods via Amazon’s Mechanical Turk and Zooniverse to generate labels and annotations for articles that do not require domain expertise.

To further enhance the accuracy of the GRITS classification algorithms, we propose crowdsourcing a diagnostic challenge, whereby we host and share our training data on an existing challenge platform (e.g., Kaggle), where programmers compete to build classifiers to improve accuracy. These approaches would compliment the models developed by our internal users and promote global citizenship in combating the scourge of emerging infectious disease. Finally, we propose integrating EHA’s Global Repository for Infectious Diseases (GRID) to solicit collective intelligence and peer editing to develop a canonical collection of event data portfolios for BSVE users and the broader scientific community. The BSVE will have access to the GRID API throughout the contract period with an option to continue access through continued funding. EHA will also provide a copy of the GRID data to DTRA at the end of the contract period.

Our two-pronged approach (machine and collective) is a unique strength of the GRITS platform. The automated tools enable experts to make the best use of their time by farming out less specialized tasks to the crowd, identifying errors in the data sources, and notifying editors or creating tasks to crowd source solutions. In the case that there isn’t enough data to deliver an accurate diagnosis, GRITS will identify the most useful information in distinguishing disease candidates. Combined, this presents a unique conceptual model and workflow for supporting outbreak investigation.

## Proposed Development

GRITS enhances traditional search with adaptive diagnostic models, trained on a broad spectrum of resources from GRITS partners and curated by our communities of experts. This diagnostic function reduces the complexity and increases the specificity of the data feed to an analyst’s workflow. We propose scaling this system to handle large data volumes, along with enhancing diagnostic capabilities, thereby allowing users of GRITS and the BSVE to monitor data feeds for diseases or disease characteristics of interest. To ensure rapid diagnosis when a disease first emerges, GRITS blends social media, news media, and scientific literature. This approach may be further leveraged to identify unusual events and diagnostic gaps that may herald an emerging infectious disease of unknown etiology.

### BSVE Integration and Benefits

We propose building an application with the BSVE SDK that allows BSVE users to submit resources to diagnose. The app will display a basic visualization of a diagnosis, and provide links to diagnostic dashboards for decision support via additional visualizations and diagnostic tools. The data processing stack will provide the capabilities identified above (diagnose, mine, recommend, filter, and connect) to data being ingested or monitored by the BSVE. Our diagnostic service is designed to continually evolve and improve via mechanisms for input from collective and machine intelligence sources. Our novel approach to diagnostics builds upon traditional search with the infusion of disease ecology into the method.

### Technical Challenges

Due to incomplete reporting, errant diagnoses, delayed onset of symptoms, and delays in laboratory results, it is difficult diagnose diseases in the early phase of an outbreak10. Dialect varies among data sources and there is high variability in accessibility of news sources and scientific literature. Software has a language bias where non-Latin characters are often the source of bugs, and foreign language articles require translation to fit into a single NLP pipeline.

Additionally, media attention is biased toward OECD countries, with blind spots in some of the locations where diseases more frequently emerge, particularly those with limited clinical infrastructure and access to health professionals11. However, as a shared and accessible platform that can maximize the utility of both expert networks and machine intelligence, our tool is of particular value in resource-poor areas without laboratory diagnostic facilities or, with past events, where clinical samples are no longer available. To reduce the time lag in detecting EIDs, we are interested in identifying prompt data sources and recruiting members of the GRITS network and BSVE community in disease prone environments to provide additional reports of emerging diseases.

When attempting to respond to user submitted corrections and feedback, some classifiers take considerable time to train (e.g., neural nets), and completely retaining any classifier on big data can be prohibitive. We will need to investigate methods of incrementally retraining our classifier in a timely manner as new data is submitted.

**Technical merit**

Overall, we propose to expand our automated data extraction to support reporting initiatives and event notification systems with near-real-time automated extraction of disease-relevant information from an arbitrary, unstructured data feed. Further development of the GRITS text mining system will allow us to introduce ecological and epidemiological concepts to extract more complex information such as host-pathogen relationships and quantitative data (e.g., case counts) for modeling and analysis12. We want to improve the precision of the features we extract by developing more sophisticated natural language processing pipelines. Additionally, we plan to reduce error rates by taking advantage of the open source NLP software ecosystem to perform word sense disambiguation and coreference resolution. Identifying the features GRITS and BSVE users value and providing channels for them to provide us with feedback will also be important to improving GRITS. We propose dynamically retraining the classification algorithm GRITS uses to diagnose diseases based on user submitted corrections to its diagnosis. Finally, we plan on open sourcing reusable components of our system so that other groups can benefit from our work.

Our technology stack consists of modern web technologies like Meteor that allow us to seamlessly update the content of the UI. These frameworks are compatible with the current technologies in the BSVE, such as AngularJS. Furthermore, we employ implementations of machine learning algorithms from scikit-learn, and are using natural language processing algorithms (e.g., tokenizing, part-of-speech-tagging, lemmatization) from nltk, and CLIPS pattern. We have started with a Python ML/NLP stack because of the rich software ecosystem available, and the ability to prototype rapidly with IPython notebooks.

**Scientific merit**

Automated categorization of EID reports by the GRITS diagnostic classifier will enable analysts and to search and filter them, increasing the ratio of relevant information they review. Furthermore, GRITS will provide decision support by suggesting potential diseases in reports of unknown diseases, and by recommending relevant data to review. Established networks of experts from EcoHealth and One Health will provide input on our diagnostic engine enabling us to continually improve upon it. Infectious disease emergence is a global-scale challenge requiring an extended, engaged community to monitor, track and respond to new threats; it is also intrinsically interdisciplinary given the complex life histories of many disease agents.

Ontologies are invaluable tools in artificial intelligence systems, decision support systems, data exploration and research where they can be used to make complex inferences and generate rich datasets. We seek to expand the scope of our ecological ontologies by considering taxonomic, distribution, ecological niche, and networks for pathogens and hosts. This will help us provide state of the art diagnostics for the BSVE by providing structured data from which we can make inferences with GRITS diagnostic algorithms. We plan to use the biointelligence extracted from news media and submitted by GRITS users to assemble a corpus of new information to form the basis of new relationships and entities within biological ontologies to help explain patterns of disease emergence. We are interested in incorporating these relationships and entities into public ontologies they derive from or relate to so that other projects benefit from our work. Our goal, with the work we propose, is to push our system to provide recommendations of to users for additional data-sources, work with citizen scientists to expand the data inputs to diagnostic tools, and integrate various GRITS components into the BSVE community.

# **C. Programmatics**

This effort will support DoD CBDP, DTRA, and the Joint Science and Technology Office for Chemical and Biological Defense (JSTO-CBD), HDTRA1-14-CHEM-BIO-BAA, Chemical Biological Technologies Department and is submitted in response to Topic: CBA-03: Next Generation Analytic Capabilities for BSV. The work will support DTRA's mission to safeguard America and its allies from biological WMDs by providing diagnostic capabilities to reduce, eliminate, and counter microbial threats. Specifically this "Global Rapid Identification Tool System" is aimed at developing capacity for the detection of EID threats, to support protection efforts and mitigation of the threats posed by disease agents through the New Initiatives in Science and Technology program. This technology is intended for eventual transition through the DTRA R&D Enterprise.

## Management plan

Our management plan blends strong scientific expertise in global EID surveillance, with agile software engineering as well as iterative and incremental rapid application development. The project will be managed by our team of data scientists and software developers at EcoHealth Alliance, in consultation with thought leaders in the field of biosurveillance (Epidemico & ProMED-mail), and infused with innovative technologies Kitware Inc., developers of leading edge, high quality software.

## Key personnel (roles/responsibilities)

* EcoHealth Alliance (prime): management, delivery, software integration, computing, diagnostic analysis, data science, data mining, disease ecology
  + Andrew Huff, Ph.D., M.S. (Principal Investigator - PI)
* Kitware Inc.: data management, visualization
  + Jeff Baumes, Ph.D. (Technical Sub-contractor)
* Epidemico: data curation, digital surveillance
  + John Brownstein, Ph.D. (Scientific Consultant)
* ProMED-mail: data curation, disease outbreak reporting
  + Larry Madoff, Ph.D. (Scientific Consultant)

## Current data providers and collaborating centers:

* **ProMED-mail** - the Program for Monitoring Emerging Diseases - is an open source Internet-based reporting system dedicated to rapid global dissemination of information on outbreaks of infectious diseases and acute exposures to toxins that affect human health, including those in animals and in plants. Electronic communications enable ProMED-mail to provide up-to-date and reliable news seven days a week. Sources of information include media reports, official reports, online summaries, local observers, and others. A team of expert moderators screen, review, and investigate reports before posting to the network and distributing by email. ProMED-mail currently reaches over 40,000 subscribers in at least 185 countries.
* **HealthMap** - The freely available Web site 'healthmap.org' and mobile app 'Outbreaks Near Me' deliver real-time intelligence on a broad range of emerging infectious diseases. HealthMap brings together disparate data sources, including online news aggregators, eyewitness reports, expert-curated discussions and validated official reports, to achieve a unified and comprehensive view of the current global state of infectious diseases and their effect on human and animal health.
* **Global Repository of Infectious Disease (GRID) project** - an EHA project describing the initial emergence of global infectious disease bioevents since 1940. We collected direct language from the primary literature describing the agent, time, place, impact, transmission, host, driver, EID category, and economics of the event to discover patterns and trends among these variables across time and space. This is used to train the models with historic context to support both our Hotspot predictive modeling efforts and diagnostic capabilities.
* **GIDEON** - Global Infectious Disease and Epidemiology Network - is the world's premier global infectious diseases knowledge management database. It contains a diagnostic module that employs information on symptoms, country, incubation period, and laboratory tests to construct a ranked differential diagnosis. The Infectious Diseases module encompasses over 340 infectious diseases, 231 countries, over 500 anti-infective drugs and vaccines.
* **PubMed, Google Scholar, and Web of Science** - will be used to generate records of confirmed diagnoses and historical outbreaks. Additionally, we hope to explore archival resources such as CDC disease reports.

# **D. Relevance**

Our goal with this project is to develop a tool system of high-relevance to DTRA’s Goals and Objectives. With the support of DTRA, we could advance our GRITS to full development and extend timely operational capability to all sectors affected by the threat of EIDs. This state-of-the-art technology will help advance our readiness to combat the broad class of biological threats posed by EIDs, including the capability to identify agents with the potential to be used as WMDs. We have the expertise and capacity to ensure useable capability of the GRITS application within the timeline of tasks we propose. The tools will provide near-real-time decision support to end-users of GRITS and the BSVE. By using open source and transparent methods, we ensure that our results are reproducible and that the technology is portable, reliable, agile, and flexible in confronting emerging threats. The tools we propose are state-of-the-art, both scientifically and technically. GRITS leverage the latest reactive web technologies (e.g., Meteor), visualization environments (e.g., WebGL and Tangelo), and scientific databases (e.g., Girder). The machine learning and clustering algorithms are drawn from Scikit-learn, an open and accessible, community-supported Python library.

A key strategy in mitigating EID risk is to build situational awareness as far forward from our shores as possible, by using advanced digital biosurveillance to detect early signals that portend the emergence of high-risk, priority diseases and pathogens, including bioterrorism agents. This biosurveillance technology is adaptable to low resource settings, among them those most vulnerable to EIDs. Ultimately we hope to empower our warfighters and allies with the tools necessary to adapt and shape the dynamic Global Security Environment, as it pertains to the acute threat of infectious diseases.

**Responses to DTRA’s questions from the White Paper**

**How will the proposed system be sustained? Is a fee for service model envisioned? If so, details should be provided.**

We propose providing to DTRA all of the data hosted in GRITS.db and all of the code developed for this proposed system under permissive open source licenses (e.g., MIT and Apache2). Furthermore, EHA has established the Data Science and Research Technology (DART) lab with the express goal of supporting the services outlined in this proposal. We envision sustaining the service under a ‘fee for service’ model to be negotiated with DTRA upon delivery of the work. We would do so in coordination from our technical subcontractor, Kitware Inc., who has extensive experience supporting open source development for federal agencies, including DoD. Our data subcontractors have agreed to make the full data available to DTRA spanning the duration of the contract (beginning Jan 18, 2013). Where copyright restrictions limit our ability to distribute the full text of the media, we will provide programmatic tools to retrieve the text in compliance with the terms and conditions of the source. Additionally, we propose a ‘freemium’ mechanism to sustain data hosting and processing costs, whereby DTRA-approved users or organizations could access the service on a pay-as-you-go basis (e.g., paying for characters processed or volume stored).

**What will actually be delivered to the Government/BSVE? Will this simply be an API to the EHA system, or will underlying tools and data be delivered? If yes, please clearly describe all deliverables.**

We are prepared to deliver all materials developed through the contract under permissive open source licenses. Given the complexity of the system, and the maintenance burden, we recommend that that EHA continue to maintain the materials on a mutually agreed upon third party service (e.g., AWS) and provide the access to the data and diagnostic capabilities through our API. The source code will be made available through a private repository. We will provide full documentation, as we have done for the data service we are currently providing to the BSVE developed by Digital Infuzion. This applies to all deliverables, including GRITS.app, GRITS.db, and GRITS.md. The GRID platform has been developed with support from other agencies, foundations, and universities; however, we will provide unlimited access in perpetuity to DTRA via the API and web interface. The source code for Tangelo visualizations Girder are publicly available on Github and permissively licensed.

# **III. Credentials**

## A. Summary of Credentials

## EcoHealth Alliance (EHA):

## Building on over 40 years of groundbreaking science, EHA is a global nonprofit organization dedicated to protecting wildlife and safeguarding human health from the emergence of disease. The organization develops ways to combat the effects of damaged ecosystems on human and wildlife health. Using environmental and health data covering the past 60 years, EHA’s scientists created the first ever global disease hotspots map that identified at-risk regions, to help predict and prevent the next pandemic crisis. That work is the foundation of EHA’s rigorous, science-based approach, focused at the intersection of the environment, health and capacity building. Working in the U.S. and more than 20 countries worldwide, EHA’s strength is founded on innovations in research, training, global partnerships, and policy initiatives.

EHA is a partner of the USAID Emerging Pandemic Threats PREDICT program, a $75 million effort focused on predicting and preventing pandemic diseases. PREDICT is building a global early warning system to detect and reduce the impacts of emerging diseases that move between wildlife and people (zoonotic diseases). PREDICT has developed a SMART surveillance method (Strategic, Measurable, Adaptive, Responsive, and Targeted) that accounts for the fact that zoonotic pathogens, like influenza and SARS, are responsible for the majority of emerging infectious diseases in people, and that more than three quarters of these emerging zoonoses are of wildlife origin. The SMART surveillance approach is designed to detect novel diseases with pandemic potential early, giving health professionals the best opportunity to prevent emergence and spread. It also targets sentinel animal species at active human interfaces in hotspot regions to improve surveillance efficiency.

The PREDICT team builds on a broad coalition of partners to develop the global capacity to monitor diseases at the animal-human interface and develop a risk-based approach to concentrate these efforts in surveillance, prevention, and response at the most critical points for disease emergence from wildlife.

PREDICT project objectives:

* Assess local surveillance capacity;
* Implement targeted and adaptive wildlife disease surveillance systems;
* Develop and deliver new technologies to improve efforts close to the source;
* Use cutting-edge information management and communication tools to bring the world closer to realizing an integrated, global approach to emerging zoonotic diseases.

**Partners:**

**HealthMap/Epidemico**

HealthMap is a team of researchers, epidemiologists and software developers based out of the Children's Hospital, Boston. Founded in 2006, HealthMap is an established global leader in utilizing online informal sources for disease outbreak monitoring and real-time surveillance of emerging public health threats. The freely available Web site 'healthmap.org' and mobile app 'Outbreaks Near Me' deliver real-time intelligence on a broad range of emerging infectious diseases for a diverse audience including libraries, local health departments, governments, and international travelers. HealthMap brings together disparate data sources, including online news aggregators, eyewitness reports, expert-curated discussions and validated official reports, to achieve a unified and comprehensive view of the current global state of infectious diseases and their effect on human and animal health. Through an automated process, updating 24/7/365, the system monitors, organizes, integrates, filters, visualizes and disseminates online information about emerging diseases in nine languages, facilitating early detection of global public health threats.

**ProMED**

ProMED-mail was established in 1994 with the support of the Federation of American Scientists and SatelLife. Since October 1999, ProMED-mail has operated as an official program of the International Society for Infectious Diseases, a nonprofit professional organization with 20,000 members worldwide.

**Kitware, Inc.**

Kitware, Inc. creates and supports leading edge, high quality software in the fields of computer vision, medical imaging, visualization, 3D data publishing, and technical software development. Kitware employs an open source development model to foster extended, collaborative communities, and an open source business model to provide flexible, low-cost technical solutions. The Company's services and products include technology integration, software support, consulting, custom application development, and training and productivity tools that leverage our open-source software systems.

**B.** **Summary of Qualifications for PI and Key Personnel**

**Dr. Andrew Huff (PI),** is a senior research scientist at EcoHealth Alliance. Over the past 4 years, he patented a novel technology at the University of Minnesota, to collect and combine disparate spatial data sources to rapidly identify biologically or chemically contaminated sources of food. To accomplish this difficult task, he has engineered spatial algorithms and has created novel primary functional keys for data fusion. While at Sandia National Laboratories, he developed multiple computational models for infectious disease consequence prediction and developed unique analytical methods for infectious disease epidemiology and emergence. He has worked on projects funded by DOE, DHS, DoS, DTRA, DVA FDA, HHS, and USDA.

**Dr. John Brownstein** is an Associate Professor at Harvard Medical School and directs the Computational Epidemiology Group at Children's Hospital Informatics Program in Boston. His group is supported by a multi-million dollar budget with support from NIH (NLM and NIAID), USAID, Centers for Disease Control and Prevention, and Google.org. He has pioneered efforts in participatory epidemiology, using statistical and informatics approaches aimed at improving public health surveillance and practice. He recently was awarded the Presidential Early Career Award for Scientists and Engineers, the highest honor bestowed by the United States government to outstanding scientists and engineers.

**Dr. Lawrence Madoff** is an infectious disease physician whose career has been devoted to disease surveillance. Dr. Madoff is the Editor of ProMED-mail, which uses Internet-based communication and social media to detect and report emerging infectious diseases globally. He is currently Director of the Division of Epidemiology and Immunization at the Massachusetts Department of Public Health, which oversees infectious disease surveillance and immunization activities in the state. He is a fellow of the American College of Physicians and a Fellow of the Infectious Disease Society of America.

**Dr. Jeff Baumes** is a technical lead at Kitware Inc. He has significant expertise in information analysis and presentation. His contributions include novel graph clustering algorithms that allow cluster overlap, and algorithms for discovering subsets of individuals persistently connected over time. Over the last six years, he has been a major technical contributor to the open source Titan scalable analysis and visualization toolkit, which is an extension of the Visualization Toolkit to include informatics and information visualization capabilities. He has worked on projects in several fields surrounding Titan such as text analysis, bioinformatics, and social network analysis including funding from NSF, DoD, DOE, and NIH.

**C.** **Summary of Facilities to Perform the Proposed Work**

**Facilities at EcoHealth Alliance (EHA)**

EcoHealth Alliance (EHA) is a 501(c)(3) nonprofit organization that specializes in scientific research on the causes, origins, and spread of zoonotic emerging diseases. EHA scientists have been working on spatial modeling for over 15 years, and on modeling of infectious disease emergence and spread for over a decade. EHA is based in New York City with 10,000 square feet of office space including a meeting room and basic laboratory. A core administrative staff of 11 employees support EHA’s scientific team (15 core scientists, 100+ field staff) and are available for work on this project through foundation support. EHA is equipped with 25 networked PCs including ARRA funded International Live Meeting Video Conferencing facilities. EHA has access to multiple servers, server support, and all necessary software on Mac, Linux, and Windows operating systems. Additional computing power is acquired from commercial cloud providers to meet project needs.

EHA has an active program of staff development and this is reviewed and adjusted annually as part of each employee’s evaluation process. Specific provisions are made for internal training and external training resources, tuition support programs via a partnership with Columbia University, and active support of staff to spend time in collaborators organizations. All early stage investigators are mentored to provide guidance in research practices, grant management, administration and project management. Financial support from EHA core funds is available to support external tuition, travel to conferences and to conduct joint research in collaborator’s institutions. There is no obligation for teaching time at EHA and all research staff are funded for 100% research time; however, there is a provision, through partnership with Columbia University, to enable staff to teach at the undergraduate and graduate level, with monetary support provided by Columbia University. Administration and other staff are supported in their efforts to enhance their careers by the provision of tuition fees for external courses, travel funds for conferences, and time off their core activities.

## Facilities at Kitware Inc.

Kitware Inc. is headquartered just north of Albany, New York in a Clifton Park office complex. Kitware rents approximately 27,000 square feet of office space at this location. Kitware also has an office in Chapel Hill, North Carolina approximately 6,200 square feet in size. Both offices are linked via a common virtual private network and a shared phone system, and share financial and administrative personnel. They also have on-site office managers, lunchrooms, private meeting rooms, and advanced conference facilities including large screen projection systems and whole-room Polycom video conferencing systems. The proposed work will be performed at the Clifton Park site.

Kitware has a mixed environment of personal and shared computing platforms. Employees average three computers per person (desktop, laptop, and/or home system), with each computer typically equipped with multiple multi-core processors, a high-performance graphics card, dual monitors, and 8GB or more of main memory. These personal systems run a mix of Windows, Mac OS X, and Linux operating systems. Shared resources include compilation and testing farms as well as workstations running a variety of alternative operating systems for testing purposes, e.g., Windows XP or Vista. Kitware also maintains several servers to provide public access to the open source VTK, ITK, TubeTK, Titan, Slicer, CMake, and ParaView systems; to host web pages and web services for open source communities such as NA-MIC and Visomics; to operate open-access journals such as the Insight Journal and the Midas journal which has hosted workshop papers for nearly ten years; and to provide access to massive collections of public data for computer vision and medical imaging algorithm evaluation. Access to these systems is provided by a fiber connection to the Internet yielding a total of 100 Mbit/second data rate.

Kitware hosts several special-purpose, high-end workstations, GPU systems, haptic systems, and magnetic and optical trackers. One such workstation is a multi-GPU computer featuring 6 NVidia GPU boards: 5 C2050 Tesla and 1 Quadro 5000 Fermi, as well as a 6-Core X5680 3.33GHz processor. A noteworthy haptic system is a MBP Freedom 7S haptic device, configurable for 6 or 7 degrees of freedom. To address large-scale distributed computing, the company maintains several clusters for development and testing. The clusters include: a twelve node testing cluster running mixed Linux and Windows operating systems; a four node Windows cluster with a gigabit Ethernet network; and a seven node Linux cluster with dual 64-bit Xenon processors, high-end graphics accelerators, and an Infiniband network driving a 3x2 PowerWall display. Kitware also has access to several external computer systems (e.g., HP, IBM, and Intel) through various vendor partnership programs.

# **IV. WORK TO BE PERFORMED.**

## A. General

Our goal over the 3-year time period is to expand our DTRA-funded GRITS platform to deliver disease diagnostics, decision support, and data processing. These enhanced capabilities will be powered by GRITS analytics, and data.

We are currently building a robust and scalable software infrastructure to provide a diagnostic decision support system for analysts. The final deliverable will include our user interface (GRITS.app), application interface (GRITS.api), media diagnostics (GRITS.md), and database (GRITS.db). Overall, GRITS will be deployable and generalized application that will output probabilities and lists of pathogens likely responsible for an outbreak on the basis of user-provided data. Accordingly, the resource will be adaptable to a specific organization or agency’s needs or emerging threats, and geographic areas of high concern. The source code for this algorithm will be made available to interested parties for further development and adaptation.

The original Rapid Identification Tool (RIT) prototype was developed by manually extracting symptoms from encephalitides reports in ProMED-mail to train a diagnostic model. Through rigorous testing, we identified modeling approaches that improved performance by combining natural language processing and machine learning algorithms. We recognized automated data collection and crowd sourced data curation would be needed to scale to disease coverage and diagnose additional diseases with greater precision. The GRITS diagnostic dashboard provides decision support to experts at our partner organizations by automatically extracting and visualizing information from media. We leverage crowdsourcing techniques by providing experts with tools for curating disease portfolios and annotating articles. We have developed an application programming interface (API) for access to our data and diagnostics by third party developers and users. Additionally, we integrated the project with an ongoing EHA initiative to collect historical disease outbreak data (Global Repository for Infectious Diseases - GRID). Finally, we integrated work from our colleagues at Kitware to support the storage and visualization of the large, complex datasets being generated.

During the proposed contract period, we will improve the accuracy and robustness of the GRITS media diagnostics. We will experiment with crowdsourcing as an additional source of training data and knowledge to further improve GRITS.md and expand the capabilities of the system to provide decision support and recommendations to analysts.

By the end of the contract period GRITS will provide:

1. Robust architecture fordiagnostics and data processing;
2. Interface from BSVE to GRITS with SDK;
3. Healthmap, ProMED-mail, and EcoHealth data to the BSVE; and,
4. Machine and community intelligence.

**Diagnostic modeling**

We plan to use the biointelligence extracted from news media and submitted by GRITS users to form the basis of new relationships and entities within biological ontologies that link EID events, diseases, symptoms, hosts, and other entities and attributes related to infectious diseases together in a rich network structure. This ontology will be a critical data source for making domain specific inferences in diagnostic algorithms.

**Data Storage and Management**

Girder is a data management platform built to meet the needs of distributed, data-centric web applications. Girder is a modular framework that allows developers to build systems that use any or all of the components necessary to create a system tailored to their needs. All data sharing web applications need the same core functionality: upload, download, large data storage, supplemental metadata storage and indexing, authentication/authorization, a RESTful API, and extensible plugin architecture. Girder provides these components and is currently being used for GRITS as well as several DOE projects.

## B. Summary

#### **Year 1 (2015)**

1. Connect GRITS Girder database to the BSVE
2. Develop recommendation and decision support capabilities
3. Connect GRITS diagnostic and text-mining APIs to the BSVE
4. Build BSVE interface to GRITS with the SDK
5. Build mechanisms to crowd source annotations
6. Incorporate disease network graphs to assist diagnostics
7. Support diagnostic algorithm development with dashboard

Year 2 (2016)

1. Expand diagnostic capability to arbitrary data feeds
2. Connect GRITS to GRID
3. Update diagnostic model in near-real-time
4. Use text mining to extend network graphs/ontologies.
5. Connect GRID’s collective intelligence editor to the BSVE
6. Connect GRITS diagnostic data filtering to the BSVE
7. Generate disease summary reports from diagnostics
8. Forecast disease emergence

## C. Detailed Tasks

**Task 1: Connect GRITS Girder database to the BSVE**

**Description:** Coordinate with BSVE developers to provide API access to GRITS database (HealthMap, ProMED and EcoHealth data with diagnostic metadata).

**Resources:** EcoHealth & Kitware (API, BSVE support), HealthMap & ProMED (data).

**Metric(s) of success:** BSVE access to GRITS data via API.

**Deliverable:** API key and access for the BSVE team, communication between BSVE/GRITS.

**Subtasks:**

1. Solicit feedback on GRITS data API from BSVE team
2. Create API key generation and assignment infrastructure for GRITS
3. Connect Girder data storage to API key infrastructure
4. Incorporate feedback from the BSVE team into GRITS API
5. Generate documentation for API access

**Task 2: Develop recommendation and decision support capabilities**

**Description:** Use our GRITS.md to extract features of incoming articles to inform media recommendations for analysts based on areas or keywords of interest or a collection of documents being evaluated. Identify targets for data collection that would most enhance diagnostic capabilities for a particular event.

**Resources:** EcoHealth (algorithms and infrastructure), Kitware (storage & visualization), HealthMap & ProMED (testing).

**Metrics of success:** Recommendation system returns relevant articles, people and organizations. Adding recommended information improves diagnosis.

**Deliverable:** API and diagnostic dashboard interface to recommendations.

**Subtasks:**

1. Set up job to extract and store features of incoming articles
2. Provide recommendations based on keywords
3. Recommend similar articles based on an article or portfolio
4. Create mechanism for people and organizations to opt in to be recommended
5. Recommend people or organizations based on keywords, articles, or portfolios
6. Identify missing features that differentiate between top diagnoses for an article
7. Connect recommendation system to API

**Task 3: Connect GRITS diagnostic and text-mining APIs to the BSVE**

**Description:** Provide BSVE developers with API access to GRITS media diagnostics and text mining tools and support them as they integrate these features into the BSVE interface.

**Resources:** EcoHealth (API development, BSVE support).

**Metrics of success:** BSVE team satisfied with API structure. GRITS features and diagnoses accessible through BSVE.

**Deliverable:** API access for the BSVE team, communication between BSVE/GRITS.

**Subtasks:**

1. Connect diagnosis and text-mining to API infrastructure
2. Write documentation on diagnosis and text mining APIs for BSVE team
3. Incorporate feedback from the BSVE team into GRITS API
4. Support BSVE team as they integrate GRITS extracted features and diagnostics

**Task 4: Build BSVE interface to GRITS with the SDK**

**Description:** Build an app on the BSVE SDK that allows BSVE users to submit text to diagnose. The app will display a basic visualization of a diagnosis, and provide links to diagnostic dashboards where users can see additional visualizations and interact further with GRITS diagnostic tools.

**Resources:** EcoHealth (app development & testing).

**Metrics of success:** Users able to access GRITS through app deployed to BSVE.

**Deliverable:** Deployed app, BSVE users able to access diagnostic dashboards.

**Subtasks:**

1. Obtain and review SDK documentation from BSVE
2. Develop backend component to access GRITS API
3. Set up mechanism for BSVE users to authenticate and/or register with GRITS
4. Develop frontend component to allow submission
5. Develop frontend component to provide diagnosis and dashboard links
6. Coordinate with BSVE to deploy application

**Task 5: Build mechanisms to crowd source annotations**

**Description:** Identify human intelligence annotation tasks for crowdsourcing by citizen scientists and Amazon’s Mechanical Turk. Extend the annotation interface so that users can correct auto-generated annotations.

**Resources:** EcoHealth (annotation interface), Mechanical Turk (pay for annotations), Citizen Scientists (volunteers).

**Metrics of success:** Annotations crowd sourced. Improves diagnostics from annotations.

**Deliverable:** Crowd sourced annotations incorporated into GRITS training data.

**Subtasks:**

1. Prepare annotation interface to allow correcting annotations
2. Write documentation of annotation interface
3. Test annotation interface and incorporate feedback
4. Develop infrastructure to identify and assign annotation tasks
5. Integrate annotation interface into Mechanical Turk or similar platform
6. Integrate and deploy citizen science platform
7. Evaluate obtained annotations and report on quality (e.g., inter-annotator agreement)

**Task 6: Incorporate disease network graphs to assist diagnostics**

**Description:** Model the geographic, ecological, and information structures of infectious disease networks and connect them to diagnostic API. Develop visualizations for diagnostic dashboard. (NEEDS UPDATE)

**Resources:** EcoHealth & Kitware (network modeling, visualization).

**Metrics of success:** Improved diagnostics from network reasoning.

**Deliverable:** Visualizations of the network model in the diagnostic dashboard.

**Subtasks:**

1. Develop features to extract from texts and map to ontologies
2. Develop geographic, ecological, and transportation network graphs
3. Develop information network graph
4. Build graph visualizations with Tangelo
5. Develop tool to identifies ‘blind spots’ in the information network

**Task 7: Support diagnostic algorithm development with dashboard**

**Description:** Support multiple models for diagnosis, and continually reevaluate their effectiveness. Different algorithms may perform differently with time, origin, or diseases. Allow users to run and compare the results of different models via the diagnostic dashboard. Run automated jobs to compare the performance of models over time.

**Resources:** EcoHealth & Kitware (algorithm and interface development).

**Metrics of success:** Run and compare models from the dashboard.

**Deliverable:** Capacity to compare model performance in the dashboard.

**Subtasks:**

1. Develop infrastructure for training multiple machine learning algorithms
2. Create interface to select algorithms and compare results
3. Create interface for selecting algorithm parameters
4. Automate algorithms and record performance

**Task 8: Expand diagnostic capability to arbitrary data feeds**

**Description:** Develop robust scraping algorithms and provide an interface for users to connect data sources to GRITS.

**Resources:** EcoHealth (algorithms), Language translation service.

**Metrics of success:** Users can submit relevant feeds to the GRITS system.

**Deliverable:** Interface for submitting data feeds.

**Subtasks:**

1. Build a submission interface for users to submit arbitrary feed
2. Integrate article-processing pipeline with a translation service to process non-English articles in near real time.

**Task 9: Connect GRITS to GRID**

**Description:**Incorporate GRID dataset into GRITS to give the platform a comprehensive historical perspective on EIDs and enhance analytic capabilities around drivers of disease behavior. Extend the recommendation system to include historic context from GRID (e.g., media and data). Use historic event data to expand to match targets for a disease or keyword.

**Resources:** EcoHealth (recommendation system, GRID).

**Metrics of success:** Diagnostic decision support enriched by historic event data.

**Deliverable:** Improved recommendation system, match new reports with historic events.

**Subtasks:**

1. Evaluate existing GRID API against needs of recommendation system
2. Develop capacity of GRID API to deliver historic event matches
3. Recommend GRID events (e.g., current event is similar to past outbreak)
4. Use GRID media to improve diagnostics and recommendation quality
5. Match GRITS data to events in GRID (expanding from EIDs to all outbreaks)

**Task 10: Update diagnostic model in near real time**

**Description:** Retrain the classifier with new reports from HealthMap/ProMED or users submissions. Classify labeled training data with GRITS, for example correcting misclassification. Investigate classifiers capable of distributed training or incremental retraining.

**Resources:** EcoHealth (algorithm and architecture development).

**Metrics of success:** Model updates improve performance. Classifier updates.

**Deliverable:** Enhanced classification infrastructure, retraining interface.

**Subtasks:**

1. Create a service for retraining the classifier with new labeled data
2. Research distributed training and incremental retraining algorithms
3. Add features to the UI that allow users to confirm or correct classifications

**Task 11: Use text mining to extend network graphs/ontologies**

**Description:** Use features extracted from disease reports to add entities and relationships to disease network ontology. For example, linking case counts to locations.

**Resources:** EcoHealth (feature extraction and ontologies).

**Metrics of success:** Accuracy of entities and relationships extracted from our data sources.

**Deliverable:** An extended ontology generated by text-mining algorithms.

**Subtasks:**

1. Infer set of subjects (e.g., EID events) and predicates (hasCase, reportedBy)
2. Create feature extractors for interesting subject-predicate-object groups
3. Use named entity classification or related technologies to identify diseases and disease attributes that are unknown to our system in advance (e.g., viral strains that have mutated or developed antibiotic resistance).

**Task 12: Connect GRID collective intelligence editor to the BSVE**

**Description:** Incorporate BSVE users as experts for the review and editing GRID events.

**Resources:** EcoHealth (GRID).

**Metrics of success:** BSVE users contribute to GRID. Diagnostic models improve from GRID data. Additional features are extracted based on GRID data.

**Deliverable:** BSVE/GRITSaccess to GRID, improved diagnostics and feature extraction. **Subtasks:**

1. Evaluate and implement changes to GRID API for of GRITS diagnostic needs
2. Generate keywords and rules from GRID data to incorporate into GRITS text mining
3. Incorporate GRID data into diagnostic model training
4. Design user interface for expert review and editing of GRID events
5. Implement expert review and editing interface
6. Set up mechanism for registering and authenticating BSVE users with GRID

**Task 13: Connect GRITS diagnostic data filtering to the BSVE**

**Description:** Use diagnostics to reduce data volume to relevant reports. Users should be able to list diseases or regions of interest.

**Resources:** EcoHealth & Kitware (GRITS integration and data filtering).

**Metrics of success:**  GRITS will return a subset of related assets.

**Subtasks:**

1. Generate diagnostic rules to match regions and diseases of interest
2. Test filtered results with experts to determine accuracy of filtering
3. Add near-real-time diagnostic filtering to GRITS API

**Task 14: Generate disease summary reports from diagnostics**

**Description:** Aggregate the data collected by text-mining and diagnostic algorithms to give an meta overview of a collection of reports or disease outbreak.

**Resources:** EcoHealth (algorithms and reporting).

**Metrics of success:** Number of visits per unique user to the summary report website

**Deliverable:** API returns summary reports to diagnostic dashboard and BSVE.

**Subtasks:**

1. Create algorithms for generating statistics (e.g., case counts) and visualizations (e.g., epidemic curves) to include in the summary report.
2. Integrate diagnostic filtering to identify data sources for the summary report

**Task 15: Forecast disease emergence**

**Description: EHA are experts in disease hotspot mapping.** Use GRITS.db and GRITS.md to identify diseases that pose the greatest threat. Use case-counts and data from historical epidemic curves to model probable epidemic curves for new diseases.

**Resources:** EcoHealth (hotspots modeling).

**Metrics of success:** The cumulative difference between the actual epidemic curve and predicted curve. Identify hotspots for disease emergence based on diagnostics.

**Deliverable:** Return geocoded hotspots via API for visualization on BSVE.

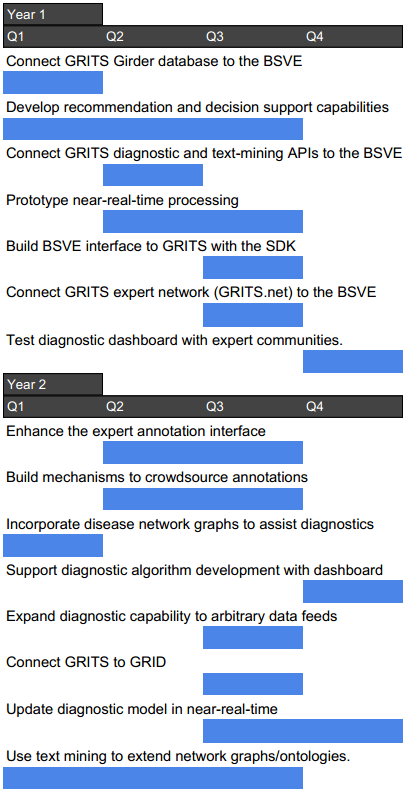
**Subtasks:**

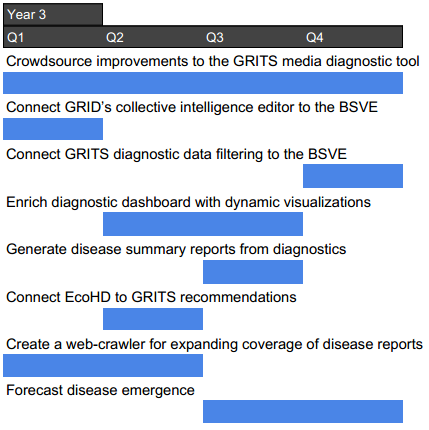
1. Build a mathematical model of disease emergence that incorporates GRITS networks
2. Create a geo visualization for GRITS dashboard
3. Add geocoded hotspot data to the API

## Maintenance Plan

* Support the Global Rapid Identification Tool System (GRITS) on the cloud
* Develop a software process infrastructure for the GRITS developer community
* Build a robust server network to ensure uptime of the GRITS platform
* Develop a test suite to identify issues and ensure compatibility with the BSVE
* Develop user and developer documentation for the platform
* Maintain the software via a ticketing system.

**V. Performance of Work (NEEDS UPDATE when tasks are finalized)**





**VI. References**

1. Jones, Kate E., et al. "Global trends in emerging infectious diseases." *Nature* 451.7181 (2008): 990-993.
2. Epstein, Jonathan H., et al. "Nipah virus: impact, origins, and causes of emergence." *Current Infectious Disease Reports* 8.1 (2006): 59-65.
3. Li, Wendong, et al. "Bats are natural reservoirs of SARS-like coronaviruses."*Science* 310.5748 (2005): 676-679.
4. Memish, Ziad A., et al. "Middle East respiratory syndrome coronavirus in bats, Saudi Arabia." *Emerging infectious diseases* 19.11 (2013): 1819.
5. Preston, Nicholas D., Peter Daszak, and Rita R. Colwell. "The Human Environment Interface: Applying Ecosystem Concepts to Health." *One Health: The Human-Animal-Environment Interfaces in Emerging Infectious Diseases*. Springer Berlin Heidelberg, 2013. 83-100.
6. Pavlin, Julie A. "Investigation of disease outbreaks detected by “syndromic” surveillance systems." *Journal of Urban Health* 80.1 (2003): i107-i114.
7. Balazinska, Magdalena, et al. "Data management in the worldwide sensor web."*IEEE Pervasive Computing* 6.2 (2007): 30-40.
8. Madoff, L. C. ProMED-mail: An early warning system for emerging diseases. *Clinical Infectious Diseases* 39, 227-232 (2004).
9. Brownstein, J. S., Freifeld, C. C., Reis, B. Y. & Mandl, K. D. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLos Med.* 5, 1019-1024 (2008).
10. Ames, G. M. *et al.* Using network properties to predict disease dynamics on human contact networks. *Proc. R. Soc. B* 278, 3544-3550, doi:10.1098/rspb.2011.0290 (2011).
11. Chan, Emily H., et al. "Global capacity for emerging infectious disease detection." *Proceedings of the National Academy of Sciences* 107.50 (2010): 21701-21706.
12. Holt, Robert D., and Andrew P. Dobson. "Extending the principles of community ecology to address the epidemiology of host-pathogen systems." *Disease ecology: community structure and pathogen dynamics* 1 (2006): 6.