NSF PREDICTIVE INTELLIGENCE FOR PANDEMIC PREVENTION (PIPP)

National Symposium on PRedicting Emergence of Virulent Entities by Novel Technologies (PREVENT)

Workshop Summary Report

June 2021

NSF Georgia Tech NYU Michigan Wisconsin

Acknowledgements

Many thanks to everyone who contributed and participated in the PREVENT Symposium[1] held virtually on February 22 – 23, 2021. This report was prepared with the input from over 60 scientists, engineers, and other stakeholders, and represents the synthesis of results and discussion from the February 2021 virtual Symposium on PRedicting Emergence of Virulent Entities by Novel Technologies (PREVENT). Any opinions, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the United States Government.

Special thanks are extended to those listed below and to active discussants who participated in the breakout sessions (a complete list is provided in *Appendix B: Workshop Participants*).

Plenary Speakers:
Rita Colwell (University of Maryland), Bryan Grenfell (Princeton University), Ruian Ke (Los Alamos National Laboratory), Denise Kirschner (University of Michigan), Vipin Kumar (University of Minnesota), Madhav Marathe (University of Virginia), Jordan Peccia (Yale University), Debra Peters (United States Department of Agriculture), Marc Riedel (University of Minnesota), Paul Turner (Yale University), David Van Valen (California Institute of Technology), Bin Yu (University of California – Berkeley)

Prepared by the Workshop Planning Committee

B. Aditya Prakash, PhD	Georgia Institute of Technology
Paul Torrens, PhD	New York University
Krista Wigginton, PhD	University of Michigan
John Yin, PhD	University of Wisconsin – Madison
Kenta Shimizu	Energetics Incorporated
Emmanuel Taylor, PhD	Energetics Incorporated
Ridah Sabouni	Energetics Incorporated

This workshop was supported by NSF award # 2115126

[1] Symposium homepage: http://prevent-symposium.org/

From the Organizers

Since December 2019, the COVID-19 pandemic has caused the loss of nearly 4 million lives across the globe at the time of this report. Despite the record-time production and distribution of multiple highly effective vaccines, the pandemic was bad and it is not yet over. Much remains uncertain about where the virus came from and how/when it will ultimately be contained. COVID-19 prompted the scientific community to look at what we know about pathogen emergence and made us quickly aware of the many things we do not know. From surveillance science to decontamination strategies, scientists worldwide mobilized quickly to start to fill in some of these critical research gaps.

When we were contacted by the NSF PIPP Working Group, it was a call to duty, and we were grateful to have an opportunity to contribute our own training, experience and perspectives toward the effort. We represented a broad range of disciplines and research expertise; Aditya is a computer scientist with a research focus in data science, machine learning, and Artificial Intelligence (AI). Paul has a computer science and engineering background with a research focus in using modeling and simulation to study complex urban systems. Krista is an environmental engineer studying the detection and fate of viruses outside of their host organisms. John's background is in chemical engineering and has a research focus in systems biology. Coming together from these backgrounds to organize the workshop was challenging due to the different languages (e.g., jargon) used by our fields, not to mention our fundamental understandings of what prediction and preparedness even meant. Our experience underlines the critical importance of researchers with different backgrounds and perspectives to join together to solve the complex issues around pandemics. After several months working together, we had a much better appreciation of the research our disparate fields were contributing to the pandemic and our scientific language barriers were greatly reduced. Indeed, by the end of our time together, we were brainstorming numerous ways our research could come together to address aspects of pandemic preparedness and prediction.

We hope that the workshop attendees and readers of this report have a similar experience. That is, whatever specific discipline or research area they inhabit, the attendees and readers will learn about new tools, approaches, and hypotheses related to predicting pathogen emergence and pandemic preparedness. Many of these are likely far from their own field. In doing so, we hope to spark creative approaches to solving the critical knowledge gaps that are identified in the report.

We named our symposium 'PREVENT' as we wanted to understand what it will take for us to go beyond reacting and managing an outbreak to preventing future ones. In addition, some form of the word, "to predict" also appears in the title of the NSF Working Group, in the name of our symposium, and different forms appear throughout this report. This word is defined by *Merriam-Webster*: to declare or indicate in advance; and to foretell on the basis of observation, experience, or scientific reason. We aim "to predict" in an aspirational sense; we recognize that the prevention of future pandemics will be hard to achieve, even when based on our most updated observations, by integrating our collective experiences, and by applying our best reasoning. A more appropriate term might be "to forecast," as we do for the daily weather, where atmospheric measurements over space and time are used to guide physics-based computational models. However, pandemic forecasting will present still greater challenges: to account not only for virus-human, virus-environment, and human-human interactions, but also how such forecasts will impact human behaviors.

It was a daunting task to organize an international meeting on 3-months' notice, especially during a pandemic and when we have large research efforts related to the pandemic. We nonetheless thank NSF for this opportunity. We are grateful for the many who helped, including (but not limited to) colleagues at NSF: Goli Yamini, Mitra Basu, Mamadou Diallo, Elebeoba May, and Wendy Nilsen. Co-workers at Georgia Tech also helped in many ways: Carly Ralston, Pulak Agarwal, Christina Camejo, Javen Ho, Harshavardhan Kamarthi and Alexander Rodriguez. We also thank the speakers and attendees who helped to make the workshop a success.

June 10, 2021

B. Aditya Prakash *Atlanta, GA*
Paul Torrens *New York, NY*
Krista Wigginton *Ann Arbor, MI*
John Yin *Madison, WI*

List of Acronyms

AI	Artificial intelligence
BIO	National Science Foundation Research Directorate for Biology
CISE	National Science Foundation Research Directorate for Computer Information Science and Engineering
CLEP	Combined linear and exponential predictor
CRISPR	Clustered regularly interspaced short palindromic repeat
ENG	National Science Foundation Research Directorate for Engineering
HPC	High performance computing
MHC	major histocompatibility complex
ML	Machine learning
MM	Molecular mechanics
NGO	Non-governmental organization
NSF	National Science Foundation
OISE	Office of International Science and Engineering
PCR	Polymerase chain reaction
PIPP	Predictive Intelligence for Pandemic Prevention
PPE	Personal protective equipment
PREVENT	Predicting Emergence of Virulent Entities by Novel Technologies
RNA	Ribonucleic acid
SBE	National Science Foundation Research Directorate for Social, Behavior, and Economic Sciences
TB	Tuberculosis
TGDS	Theory-guided data science
TGML	Theory-guided machine learning
VS	Vesicular stomatitis

Executive Summary

The National Science Foundation (NSF) held a virtual Symposium on PRedicting Emergence of Virulent Entities by Novel Technologies (PREVENT), on February 22 – 23, 2021 as part of its series on Predictive Intelligence for Pandemic Prevention (PIPP). The workshop brought together more than 60 leading experts, representing NSF research directorates for Biological Sciences (BIO), Computer Information Science and Engineering (CISE), Engineering (ENG), Social, Behavioral and Economic Sciences (SBE), and the Office of International Science and Engineering (OISE), to discuss how the global behavior of an infectious entity can emerge from the interactions that begin occurring between components at the molecular level and expand to physiological, environmental, and population scales.

The workshop was divided into four sessions, each focusing on one of four different scales: 1) end-to-end (or multi-scale) 2) molecular, 3) physiological and environmental, and 4) population and epidemiological. Particular focus was given to identifying challenges and opportunities in each of these domains.

The workshop aimed to:

- Identify interdisciplinary advances in science, technology, and human behavior to enable prediction and prevention of future pandemics
- Begin to build the necessary convergence to be optimally prepared to prevent future pandemics
- Establish convergent data commons and cyberinfrastructure for PIPP

This workshop report summarizes the plenary presentations, panel discussions, and breakout group sessions that took place at this event. The results presented here are drawn from the viewpoints expressed by the participants and do not necessarily reflect those of the broader pandemic research community.

Priority Challenges

Summarized below, and in graphical form in Figure ES-1, are the major takeaways from each of the four scales discussed during the workshop and additional priority research areas that emerged.

End-to-End: Pandemics are complex problems requiring expertise across multiple scales, from the behavior of molecules in living cells to humans traveling across the planet. Equally important is the proper integration of this knowledge to form a cohesive scientific framework to tackle multi-scale problems. Theories or conceptual frameworks, based on mathematical and computable models, integrate vast data and principles across many scientific disciplines. The primary challenge identified in the End-to-End session was effectively integrating models at different scales to provide a holistic, accurate predictive pandemic model. To collect sufficient data for these predictive models, a surveillance program, whether through active or passive methods, may be required and balancing data collection with privacy and security will be paramount for widespread acceptance of such a program. Other priority challenges include:

- Developing a generalized, theoretical framework to address multi-scale, multi-dimensional problems
- Developing models and ML techniques that can project reliably from sub-cellular level to population level

- Inclusion of viral screening during testing of other diseases or routine check-ups to evaluate whether new variants or emerging pathogens are identified
- Building a team with expertise at all scales

Molecular: Molecular-scale interactions can often determine macro-scale phenomena. The cellular immune response is the first line of defense for any foreign pathogen and often determines disease trajectory and outcome. The primary research challenge identified at this scale was bridging the gap between viral genomics and pathogen transmission rate, disease severity, and patient outcomes. A deeper understanding of the connection between viral genotype and phenotype will help determine transmission mechanism and ultimately help determine effective treatments and interventions. Other priority challenges include:

- Understanding pathogen prevalence on surfaces and the effectiveness of common disinfectant protocols
- The lack of reproducibility in experimental results across different laboratories
- Quantifying the strength of the virus and how quickly it can infect a host
- Integrating biological processes with chemical drivers

Population and Epidemiological: As the current COVID-19 pandemic has shown, even with mounting scientific evidence for effective intervention protocols, humans may not be inclined to follow them. Human behavior is inherently unpredictable and current epidemiological models that do not account for this may be too simplistic to accurately predict pandemic progression. The primary challenge identified in the population and epidemiological topic was the integration of human behavioral data into pandemic prediction models. Incorporation of this data will provide a more holistic model, combining physical and biological models with social theories. To accomplish this, socio-behavioral sciences must be integrated into interdisciplinary pandemic research teams from the beginning. Other priority challenges include:

- Incorporating individualistic behaviors into specific agents in agent-based models to provide a population-level digital twin
- Addressing inherent unpredictability of human behavior through social science theory and probability theory
- Enhancing data collection methods on human interactions and behaviors
- Understanding asymptomatic transmission on a population-level

Environmental and Physiological: Development of a personalized, physiological digital twin that incorporates molecular, tissue, and organ level phenomena will greatly advance predictive capabilities for disease outcome and intervention efficacy of individuals. However, in order to realize this, greater understanding of the physiological immune response and subsequent mapping of individual immune responses is needed, representing the primary challenge identified at this scale. In the realm of the environment, an equally important challenge is integrating climate data with predictive models. As global warming continuously shifts climate patterns, understanding how these changes affect disease drivers, transmission mechanisms, and individual susceptibility will be critical in developing accurate predictive models. Other priority challenges include:

- Identifying the relationship between wastewater data and true viral prevalence in the population

- Collecting a more comprehensive, multi-dimensional environmental dataset for different areas around the world. Including data on biodiversity and resilience
- Developing a computationally intensive model that incorporates patient data with known scientific data and models
- Improving methods for working with animals in the wild

Data Science, AI/ML, and Computing: Data is the cornerstone of any predictive model. Data access and quality are paramount in developing accurate models that can undergo robust model validation and uncertainty quantification. Outside of the lab environment, data streams are often noisy and inconsistent leading to difficulty in subsequent processing and analysis, including AI/ML techniques. Determining the most effective method to using limited and noisy data and overcoming gaps while quantifying uncertainty was identified as one of the primary challenges on this topic.

In addition, integrating mechanistic and data-driven models can provide an alternative approach in situations where theory and data alone are not sufficient. This third type of modeling capability harnesses the advantages of both and can be quite useful in identifying areas where improvements in theory are needed. However, whether the integrated model can outperform purely mechanistic or purely data driven models is still unknown. To process and analyze large data streams necessary for building pandemic prediction models, large-scale cyberinfrastructure will be required to help scale the necessary analytics, modeling, and simulation efforts. The best methods to develop and deploy such a cyberinfrastructure must be addressed.

Other priority challenges include:

- Establishing robust data sources for: social media, mobility, compliance, human interactions, and viral/protein genomes
- Developing well-defined ML tasks which go beyond prediction to help with modeling and decision making
- Moving scientific apparatus to the edge (mobile devices) which may help alleviate handling of large, dynamic data streams.
- Developing a sandbox for testing models

Interdisciplinary Collaboration: Pandemics are multi-scale, highly complex phenomena and interdisciplinary collaboration will be required to effectively develop a holistic pandemic prediction model. Throughout the workshop, participants identified challenges in interdisciplinary collaboration. These are not technical research questions, but rather are barriers researchers face when trying to collaborate with others. Communication was identified as the primary barrier to interdisciplinary collaboration. The same words often don't mean the same thing across disciplines and a consistent methodology to establish a common language must be developed for effective and efficient collaboration. Other priority challenges include:

- Models rely on data from experiments, so model builders should help in the design of experiments
- Tighter collaboration is needed between data/computational researchers with modelers and experimentalists
- High quality outlets for disseminating interdisciplinary research need to be established
- There is a need to coordinate data and results across fields to foster interdisciplinary work

Translation of Research to Action: The current COVID-19 pandemic has highlighted the difficulty in translating research findings and data into actionable information for policymakers and public health officials. The two primary challenges in accomplishing this are balancing the timescales of researchers and policymakers and the communication barrier between researchers, policymakers, and the medical community. The research community desires longer timescales to more fully develop theories and models, while policymakers are under short timeframes to communicate actionable information to the public. Furthermore, researchers must effectively communicate their results to the medical community to provide guidance on the best interventions. Clinicians neither have the time nor the energy to sift through data troves or muddled results to identify information relevant to them. Other priority challenges include:

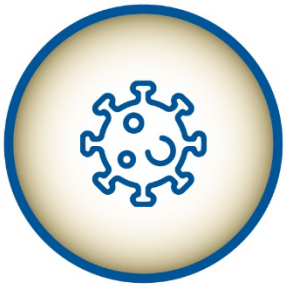
- Improving methods of interacting with practitioners, overcoming privacy and "busyness" constraints
- Connecting policy directly to the data analysis taking place
- Reducing misinformation and providing various stakeholders with clear, concise, and consistent information

Funding and Academic Institutional Structure: It was stated numerous times in the workshop that the current funding and academic structure greatly hinders interdisciplinary research. The primary challenge is that funding and academic institutions are structured such that there is little motivation to work on interdisciplinary topics. Hiring and promotion at academic institutions prioritize deepening knowledge in a specific field, de-incentivizing research faculty, to explore interdisciplinary research questions. Funding agencies similarly focus on deepening the breadth of knowledge in specific fields and do not often fund highly interdisciplinary work. Both of these aspects contribute to the silo-ing of the research community, making interdisciplinary research particularly hard to execute.

PRIORITY RESEARCH AREAS AND TECHNICAL CHALLENGES

End-to-End: Pandemics are complex problems requiring expertise across multiple scales, from the behavior of molecules in living cells to humans traveling across the planet.

- Effectively integrating models at different scales to provide a holistic, accurate predictive pandemic model
- Balancing privacy and security of a surveillance program while collecting enough data to feed into predictive models
- Developing models and ML techniques that can project reliably from sub-cellular level to population level
- Developing a generalized, theoretical framework to address multi-scale, multi-dimensional problems



Molecular: Molecular-scale interactions can often determine macro-scale phenomena.

- Bridging the gap between viral genomics and pathogen transmission rate, disease severity, and patient outcomes
- Understanding pathogen prevalence on surfaces and the effectiveness of common disinfectant protocols
- The lack of reproducibility in experimental results across different laboratories
- Quantifying the strength of the virus and how quickly it can infect a host

Population and Epidemiological: Human behavior is inherently unpredictable and current epidemiological models that do not account for this may be too simplistic to accurately predict pandemic progression.

- Integrating human behavioral data into pandemic prediction models
- Incorporation of individualistic behaviors into specific agents in agent-based models to provide a population-level digital twin
- Addressing inherent unpredictability of human behavior through social science theory and probability theory
- Enhancing data collection methods on human interactions and behaviors

Environmental and Physiological: Global warming is rapidly changing climate patterns that may fundamentally alter disease drivers and physiological susceptibility.

- Greater understanding of the physiological immune response
- Understanding how global warming affects local climate patterns and how these changes affect disease drivers, transmission mechanisms, and individual susceptibility
- Identifying the relationship between wastewater data and true viral prevalence in the population
- Collecting a more comprehensive, multi-dimensional environmental dataset for different areas around the world. Including data on biodiversity and resilience

Data Science, AI/ML, and Computing: Data is the cornerstone of any predictive model. Data access and quality are paramount in developing accurate, predictive models.

- Determining the most effective method to using limited and noisy data while overcoming gaps and quantifying uncertainty
- Integrating mechanistic and data-driven models
- Establishing large-scale cyberinfrastructure to help scale analytics, modeling, and simulation efforts
- Establishing robust data sources for social media, mobility, compliance, human interactions, and viral/protein genomes

Interdisciplinary Collaboration: Pandemics are multi-scale, highly complex phenomena and interdisciplinary collaboration will be required to effectively develop a holistic pandemic prediction model.

- Developing a common language between disparate research fields to foster better communication
- Integrating model builders in the design of experiments
- Improving collaboration between data/computational researchers with modelers and experimentalists
- Establishing high quality outlets for disseminating interdisciplinary research

Translation of Research to Action: The COVID-19 pandemic has highlighted the difficulty in translating research findings and data into actionable information for policymakers and public health officials.

- Balancing the timescales of researchers and policymakers
- Establishing channels of robust communication and common language between researchers, policymakers, and clinicians
- Connecting policy directly to the data analysis taking place
- Reducing misinformation and providing various stakeholders with clear, concise, and consistent information

Funding and Academic Institutional Structure: Funding and academic institutions prioritize research to deepen knowledge in specific fields, de-incentivizing interdisciplinary research and collaboration.

- Re-structuring funding and academic institutions to make it easier to collaborate on interdisciplinary research
- Re-thinking the hiring and promotion process to encourage interdisciplinary research

Figure ES-1: Priority research areas and associated technical challenges

Conclusions and Actionable Recommendations

Addressing the challenges outlined above will not only provide deeper foundational knowledge of the multi-scale drivers of pandemics but will also provide the modeling tools and data streams necessary to develop a robust pandemic prediction model and provide a framework for communication with policymakers, clinicians, and the public.

The traditional structure of siloed research will not be sufficient to predict or prevent future pandemics. A more collaborative research environment, including shifts in academic structure, funding interdisciplinary research, and robust lines of communication with various stakeholders, must be nurtured to establish a public health infrastructure, and ultimately a society that is more resilient and prepared for future pandemics.

High impact recommendations to tackle technical challenges associated with predicting and preventing future pandemics include:

- Improve foundational understanding of pandemic drivers at all scales, including:
    - Viral processes and transmission mechanisms
    - Bridging the gap between genomics and function for viruses and their animal and human hosts
    - Environmental factors and how they affect pandemic occurrence and trajectory
    - Physiological immune response
    - Asymptomatic transmission within a population
    - Incorporating feedback loops, like compliance behaviors, into frameworks
- Develop and deploy a viral surveillance program that collects appropriate levels of data for predictive models while maintaining security and privacy
- Improve techniques to incorporate uncertainty and variation to create more robust pandemic predictive models
- Incorporate social, behavioral, and economic sciences into pandemic research teams from the onset of research projects
- Reduce misinformation surrounding pandemic progression, disease outcome, and intervention efficacy by providing stakeholders with clear, concise, and consistent information
- Better understand the timeliness of information required by policymakers and public health officials to provide critical scientific findings that can inform public policy
- Improve collaboration between data producers (i.e., modelers) with data consumers (AI/ML researchers) to create high impact pandemic modeling capabilities
- Establish high quality outlets for disseminating interdisciplinary results to motivate interdisciplinary research
- Develop a holistic framework that bridges disparate research communities to tackle complex, multi-dimensional problems
- Develop well defined ML/AI testbeds including and also beyond prediction, to help develop techniques while also directly helping modelers and decision makers

Table of Contents

1. Background and Workshop Proceedings

Workshop Motivation and Purpose

In the last year, the ongoing COVID-19 pandemic has severely disrupted the livelihoods of our planet's human inhabitants, infecting over 126 million individuals, and causing roughly 4 million deaths at the time of this report. Many actions could have played key roles in minimizing the severity of this pandemic including: environmental monitoring for potential animal-to-human infection spillovers; establishment of pipelines for rapid vaccine development and optimal deployment and distribution; designing data science tools to accurately forecast disease trajectories; fast and adaptive syndromic surveillance and behavior tracking of humans; designing and timing effective interventions; and training susceptible and infected individuals for measures needed to inhibit the spread of infectious agents. Gaps in knowledge, methodologies, technologies, and policies must be addressed to begin developing a holistic solution that will prevent or minimize the effects of future pandemics.

The National Science Foundation (NSF), in coordination with the Directorates for Biological Sciences (BIO); Computer Information Science and Engineering (CISE); Engineering (ENG); Social, Behavioral, and Economic Sciences (SBE); and the Office of International Science and Engineering (OISE), organized a series of four workshops on the topic of Predictive Intelligence for Pandemic Prevention (PIPP) to bring together a diverse research community to start conversations and catalyze ideas on how to advance scientific understanding beyond the state-of-the-art in pre-emergence and emergence forecasting, real-time monitoring, and detection of inflection point events in order to prevent and mitigate the occurrence of future pandemics. The four workshops within the series looked at different aspects of PIPP to provide NSF and the greater research community a holistic vision of the challenges and opportunities needed in this space.

For several significant reasons, the topic of PIPP stands to benefit by drawing upon convergent science and engineering alongside traditional disciplinary reservoirs of expertise. First, the broad-reaching drivers and collateral effects of pandemics (between health, social, technological, economic, and environmental systems, for example) may outstrip the expertise of a single or a few disciplines to respond with solutions that factor the evolving scope of pandemics as they unfold. Second, the diverse multi-scale nature of pandemics (from molecular therapeutics to societal policies) often yields a knowledge base that is uneven across fields. Third, PIPP can involve a great degree of novelty, identifying and anticipating possible future scenarios will be critical in addressing what needs to be done now to be better prepared. There is an urgent need to develop sound theoretical principles and transformative experimental and computational approaches that will address the escalating threat of current and future pandemics.

Workshop Overview

Over 60 experts from academia, industry, and government gathered for the February 22nd-23rd 2021 Workshop on PRedicting Emergence of Viral Entities by Novel Technologies (PREVENT). The workshop focused on better understanding how the global behavior of an infectious agent emerges from the interactions that occur between components at the molecular, physiological, environmental, and population scales. In addition, the workshop aimed to initiate the development of a convergent interdisciplinary research community to identify and tackle the most pressing challenges in pandemic

prevention, as well as develop robust modeling capabilities through data commons and cyberinfrastructure.

The convergence of computational, biological, environmental, and social science communities of scholars provided each community with new perspectives on their roles for predicting and preventing future pandemics. The PREVENT workshop provided a valuable opportunity for the community to begin building the necessary convergence to be optimally prepared to prevent future pandemics.

The workshop (see *Appendix A: Agenda*) included four plenary sessions that each focused on one of four different scales or levels: 1) molecular, 2) physiological and environmental, 3) population and epidemiological, and 4) end-to-end. Each plenary session consisted of a keynote address, two presentations, and a panel discussion/combined Q&A. Following each plenary session was an extended breakout session, providing active participants a chance to further discuss, in small groups, the ideas presented during the plenary session. The workshop opened with the end-to-end scale, aiming to provide examples for predicting and controlling phenomena across multiple scales.

Plenary Presentations

Presentations from leading experts at each scale discussed their work, including intersections to the current pandemic, and set the stage for subsequent discussions in the workshop breakout sessions. Speaker presentation slides and plenary session recordings can be found on the PREVENT website.

End-to-End Level

Real-Time Pandemic Planning, Prediction and Response: *Madhav Marathe, Distinguished Professor, University of Virginia*

Dr. Madhav Marathe discussed real-time pandemic planning, prediction, and response in three parts: roles of models in decision making; challenges associated with planning, prediction, and response; and his own work in the on-going COVID-19 pandemic. For models to go beyond prediction, they must have the ability to synthesize data as it becomes available, provide a range of interpretations, evaluate a range of responses, monitor the effects of interventions, and coordinate understanding such that it can provide actionable information for decision-makers. These holistic models must integrate local models from a wide range of stakeholders including individual citizens; US, State, and local authorities; and on the ground responders and have the ability to incorporate each models' differences while maintaining privacy and security.

The challenges associated with end-to-end planning, prediction, and response are multi-faceted. Each of ecology, biology, epidemiology, and sociology, has its own fundamental questions that must be addressed. An additional layer of difficulty is that processes and interactions at varying spatial, temporal, and social scales all affect the course and outcome of a pandemic – hence multi-scale, multi-level network representations are needed. Having multiple sources for datasets and consistently performing extensive validation and uncertainty quantification is critical in building accurate and explainable models.

By leveraging a data-driven networked epidemiology approach using artificial intelligence (AI) and high-performance computing (HPC), Dr. Marathe and his team successfully built a prediction and decision informatics framework to provide critical, informational briefings for federal, state, and local authorities and deployed a dashboard for the general public. Dr. Marathe closed by highlighting the challenges on

the horizon, including climate change, expectation of timely information, anti-microbial resistance, and increasing urbanization.

Big Data-Model Integration as a Multi-scale Approach to Predicting the Spread of Vector-Borne Diseases: an End-to-End Vision and Operational Framework: *Debra Peters, Research Ecologist, USDA*

Dr. Debra Peters discussed developing a strategy and operational framework for complex ecological problems using as an example vesicular stomatitis (VS) disease, which is caused by infections of a rabies-like virus. VS is the most commonly reported vesicular disease of livestock in the Americas, with outbreaks in the United States roughly every 6-10 years. Dr. Peters outlined a multi-step workflow that she and her team followed to investigate the spatial variability in VS occurrence observed throughout cyclic outbreaks. The workflow is as follows: create a trans-disciplinary team; develop a conceptual model; develop a hypothesized relationship between processes and variables with each driver; identify the datasets associated with the variables in the eco-transfer function; standardize and harmonize the data in time and space; conduct analyses and interpret results; conduct experiments to test new hypotheses; develop early warning strategies based on variables and processes related to patterns; apply the approach to additional vector-borne diseases; and predict future dynamics and spread of disease based on past outbreaks.

This multi-step workflow provides a flexible framework to address many multi-scale, multi-dimensional problems including pandemic prediction. She highlighted the importance of an iterative process to build a meaningful dataset and being adaptable for future outbreaks. In conclusion, Dr. Peters cited three primary challenges while investigating this problem: limited data availability, different ecological responses between virus serotypes, and the wide array of skilled personnel needed in developing the necessary computational models.

Estimating Key Parameters for Novel Infectious Disease Outbreaks and Implications for Control – SARS-CoV-2 as an Example: *Ruian Ke, Staff Scientist, Los Alamos National Laboratory*

Dr. Ruian Ke discussed two key parameters to evaluate during early outbreaks and their implications on controlling pandemics. The parameters, early epidemic exponential growth rate, r, and basic reproductive number, R_0, can be used during the pre-pandemic interval to assess epidemic potential of novel outbreaks, evaluate the most effective intervention and control strategies, and determine herd immunity thresholds. Early in the COVID-19 outbreak, estimates of r were vastly different leading to significant differences in predicting COVID-19 severity and leading some to believe that SARS-CoV-2 was comparable to SARS-CoV-1. Furthermore, based on the inconsistent estimates of r, estimates of R_0 were similarly skewed. This led to R_0 values significantly higher in the US than widely reported values.

This illuminates the importance of model validation, especially in information poor environments. Early in the pandemic, inconsistent data collection methods, low surveillance intensity, and social, economic, and political factors all created an information poor environment leading to significant bias in reported data and uncertainty in model estimation. Dr. Ke stressed cross validation with multiple datasets and evaluation of predictability can be done to mitigate these challenges.

Combined Q&A/Panel Discussion: *Madhav Marathe, Debra Peters, and Ruian Ke*

The Q&A / Panel Discussion focused on the following key questions and discussion points:

- How do you measure success at the end of the day? In other fields, having clear success metrics has led to explosive growth in research/advancements. Are the success measures clear?
    - The ultimate success metric is lives saved. But it is hard to measure because pandemics can't be replicated to quantify how many lives were saved. In specific areas, it might be possible to identify early measures to use as metrics, but these must be determined carefully. An example metric is measuring how well a society functions. However, counterfactuals will always be hard to measure. In all cases, if we can assign values to important actions that helped an event not take place, this would be very helpful, but it is difficult. In terms of model inference, cross-validation and predictability can be used as success measures.

      With respect to VS, there were multiple outbreaks such that some data could be held back to validate the model. Significant amounts of data are required to build accurate predictive models, but with climate change, the target is always moving and historic data has limited use. We found that even with a validated model, our model did not accurately predict the most recent outbreak.

- With respect to drivers of outbreaks, how might climate change impact outbreaks in future?
    - For model development, historic data and pattern-process relationships are traditionally used. Instead of looking at correlations of output data, evaluating relationships between drivers is a more powerful way of predicting the effects of climate change. However, using the right climate model is paramount in how it will affect our predictive models.

- How do we ensure the needs of global health care workers are considered in research efforts?
    - Medical workers, as a demographic group, have been incorporated into a digital twin of cities. The models that have been built accommodate this demographic quite well.

- With public awareness, as high as it is now, what can we do now that we couldn't even a year before?
    - The momentum gained from the pandemic could be utilized to coordinate national centers to foster multi-disciplinary research. Synergizing different perspectives and backgrounds and feeding these inputs into a system model will be critical in developing a robust, multi-scale predictive pandemic model. Traditional silo-based science is not going to solve the problem.

      In addition, establishing a surveillance program to illuminate which viruses are near the animal/human interface will be very important. This can provide information on zoonotic transmission mechanism and how interactions between animal and humans affect transmissibility.

      Finally, as a community, we must provide information to the public that is both understandable and trustworthy.

- How does social network information across the US consider "on-the-ground" contexts, especially in minority communities?
    - The minority community gets disproportionately affected in any large-scale disasters. From a modeling perspective, a digital twin has been built to account for these

demographics. The data is captured structurally and the predictive models do show that they get affected disproportionately. To refine and improve these models, deeper data for these communities must be collected, while considering privacy concerns.

Molecular Level

Predicting Evolution of Virus Emergence: *Paul Turner, Distinguished Professor, Yale University*

Dr. Paul Turner discussed using experimental evolution as a method to study viral emergence. The goal of this method is to more accurately predict emergence potential on new hosts, understand why pathogens are successful at infecting new hosts, and what rules govern pathogen evolution, adaptation constraint, and extinction.

Novel host encounters and its role in emergence were of particular interest and whether dynamic environments, including wet markets, animal environments, and highly disturbed environments, led to greater emergence potential of novel pathogens. Using experimental evolution, Dr. Turner and his colleagues found that the rate of exposure to novel host species affects viral emergence potential. When gradually exposed to a host, lesser phenotypic and genetic variation and a greater adaptation to the host were found, showing that viral populations achieve higher and similar fitness in a gradual way. Further, there was a much greater reliance of genetic mutations working together. These findings highlight the importance of genetics when building predictive model for biological processes.

Dr. Turner concluded by raising other approaches for studying emergence including:

- investigation of other model and non-model systems, including true pathogens and microbes in the lab and tissue culture systems, to more accurately describe what is happening inside a macroorganism host;
- high throughput phenotyping to rapidly characterize properties of emerging pathogens and elucidate interactions of pathogens with microbiomes, viromes, host cells, and other pathogens;
- computer and data science techniques to measure phenotypic and molecular “rules” of viral interactions and fitness;
- machine learning to accurately predict infection potential of emerging pathogens based on their genetics and estimate microbial “background extinction” rates; and
- utilization of diverse approaches and workforce to incorporate a wide array of technical backgrounds and viewpoints to tackle difficult problems.

Computationally Predicting and Characterizing the Immune Response to Viral Infections: *Marc Riedel, Associate Professor, University of Minnesota*

Dr. Marc Riedel discussed using molecular simulations to predict and characterize the immune response to viral infections. The cellular immune response is the first line of defense against any pathogen. Generally, the cellular immune response involves cleaving foreign proteins, transporting them to the cell surface, binding them to major histocompatibility complex (MHC) Class I receptor molecules, and destroying the infected cell. Understanding this process is critical in predicting disease severity, developing vaccines, and identifying the impacts of viral mutations. The core problem lies in predicting whether peptides associated with the virus will bind to an individual’s MHC I receptor. Efficacious binding will result in mild symptoms while non-binding will result in a full infection. However, the

primary challenges are the sheer number of peptides to evaluate and elucidating MHC I receptor structures, the bulk of which are unknown.

Dr. Riedel and his team are developing parallel algorithms and cloud-computing infrastructure to quickly calculate binding energy using correctly aligned peptides and evaluating the torsional space to find the optimal configuration. If successful, this generalized approach can be used to predict disease severity of novel pathogens and variants as well as the effectiveness of various vaccines.

Uncovering Host-Virus Interactions with Imaging-Based Reverse Genetics: *David Van Valen, Assistant Professor, California Institute of Technology*

Dr. David Van Valen discussed developing an image-based platform to study host-virus interactions through reverse-genetics. Leveraging large scale data annotation and advances in deep learning, quantitative information can now be captured from images and provide an image-based link between phenotype and genotype. This significantly reduces the cost of genomic studies by reducing sequencing burden and may potentially make images the universal biological data type.

Host-virus interactions that ultimately govern the viral life cycle can be studied using fluorescent reporter viruses, high throughput imaging, and reverse genetics. Imaging-based reverse genetics can reveal the role of the host within the viral lifecycle, which is mostly unknown. Advances in CRISPR technology and deep learning driven analysis algorithms have made imaged-based, viral studies on mammalian cells more practical and sustainable. However, large-scale perturbation studies are complex, expensive, and un-scalable.

Combined Q&A/Panel Discussion: *Paul Turner, Marc Riedel, David Van Valen*

The Q&A / Panel Discussion focused on the following key questions and discussion points:

- Virus variants of SARS-CoV2 are creating broad concern over more efficient transmission and/or more severe disease. How might experimental evolution be harnessed to give insight into these challenges?
    - The role of experimental evolution is to evaluate evolution and adaptability of variants. However, in practice, this is very difficult and dangerous to do because of the potential pathogenicity of the virus and the necessary infrastructure to handle such pathogenic viruses. Studying close relatives of pathogens that are less dangerous is one way. An alternative method is to use safer viruses that have the same exterior protein modalities to mimic viral variants; however, the genome is still different making the transmission mechanism different. Ultimately, having the proper infrastructure to study highly pathogenic organisms is the best way to study variants.
- Are you developing your own simple torsion-only molecular mechanics (MM) force field or using existing MM force fields?
    - Existing force field calculations are used. The innovation of the approach is starting with the best confirmation and evaluating the torsional moves to find the best binding strength, instead of expending a lot of resources on finding the best binding confirmation.

- Is there scarcity of labeled data? How would you describe the main computational challenge in your problem - scale, noise, scarcity, dimensionality? Do you utilize any simulations in conjunction with data as well?
    - The lack of annotated training data for images is a significant challenge. The most useful annotation method for biologists is the most difficult and time consuming. Annotating images and ensuring data quality are two of the largest barriers in using machine learning techniques. Even though it is essential, the work is often unattractive. Methods are being developed to achieve human level accuracy in annotating images to mitigate this burden.

      With respect to modeling the cellular immune response, the scarcity of data in peptide structure is the primary challenge in using computational tools to predict immune response. Furthermore, the scale of computation requires rewriting of existing software to run more efficiently; however, as noted above, this work is often unattractive and not viewed as rewarding.

Population and Epidemiological Level

What Cross-Scale Research can tell us about Predicting, Understanding and Mitigating Future Pandemics: *Bryan Grenfell, Professor, Princeton University*

Dr. Bryan Grenfell discussed epidemic complexity, pandemic dynamics across scales and how best to address immunological and evolutionary complexity, and his perspectives on research and global health priorities. Like most pathogen researchers, Dr. Grenfell completely shifted his work to study COVID-19. In particular, he focused on modeling local dynamics, immunological uncertainties, and seasonality of COVID. Before the pandemic, he studied measles and applied his insights to COVID-19. Unlike COVID-19 and the flu, infected individuals with measles don't shed the virus enabling a simple mass action model with an extra term for seasonality and stochasticity to accurately capture the complex phenomena occurring at many scales, including network dynamics of populations, immunological dynamics of individuals, and molecular responses of cells.

A multitude of complicating factors, including heterogeneity of hosts and interventions, immunity, pathogen-host evolution, and pathogen community dynamics, necessitates more complex models for COVID-19. However, early in the pandemic, researchers, using simple SIR models, were able to determine that seasonality due to the large number of susceptible individuals and variations in immune response and NPIs.

The low cross-immunity of influenza and COVID-19 causes re-infection of susceptible individuals and enables strong transmission of antigenic novelty and evolutionary immune escape (i.e., variants). If secondary infections are used as a metric, it is clear that global vaccinations are necessary to minimize the threat of viral variants. Dr. Grenfell stressed that it is critical to think about many scales when tackling these problems and the primary gaps are in modeling and quantifying transmission. Dr. Grenfell concluded with a discussion on research priorities. These include:

- gathering data via pathogen surveillance including genomic, OneHealth, and syndromic surveillance and cross-scale interactions to get a better understanding of pandemic course and trajectory;

- immune surveillance and tools for mitigation to measure serology of both susceptible and recovered individuals to provide more informed predictions;
- human and environmental drivers including population, movement, networks, and climate to better understand how these drivers affect transmission; and
- foundational science including:
    - systems immunology including population immunity;
    - transmission biology to better understand drivers such as seasonality and aligning fluid dynamics and droplet chemistry with human movement;
    - host-pathogen evolutionary biology investigating cross-scale dynamics;
    - social dynamics including mask wearing and vaccine hesitancy and how it affects pandemic trajectory;
    - modeling/computation leveraging model hierarchies, inference, and averaging and establishing dynamical "case law"; and
    - polymicrobial interactions evaluating how interventions for a specific disease can cause outbreaks of other diseases due to the buildup of susceptible individuals.

COVID-19 Data Repository and Country-level Death Count Prediction in the US: *Bin Yu, Distinguished Professor, University of California – Berkeley*

Dr. Bin Yu discussed her experiences with creating a robust data repository and developing a county-level COVID-19 dashboard. The primary motivation of the project was to support non-government organizations (NGO) prioritization of personal protective equipment (PPE) distribution to help ease the challenges associated with the shortage. There was a dearth of data when they started building their repository and had to rely on personal networks for access to data. Dr. Yu and her team, ultimately, developed one of the first county-level dashboards, COVIDseverity.com. Data quality was the primary issue when collecting data; they found chronic undercounting of deaths, mismatch in data presentation even when it came from the same source, large differences between weekdays and weekends counts, and revisions to past data. A combined linear and exponential predictor (CLEP) was developed and took a weighted average between the linear and exponential models to produce the most accurate predictions. Conformal prediction was used to address uncertainty. More specifically, the maximum error of the previous five days was used as the uncertainty.

Future work will focus on hospital-level predictions and causal investigation to supply actionable information to policymakers. Dr. Yu concluded by highlighting the future challenges in developing a robust pandemic forecasting model. These include deploying a nimble surveillance and intervention network; building a completely integrated supply chain; improving data quality; developing a responsible, trustworthy, and reproducible AI system; and coordinating a trans-disciplinary framework to tackle large, complex problems.

Tracking Epidemics at the Population Level through Wastewater-based Epidemiology: *Jordan Peccia, Distinguished Professor, Yale University*

Dr. Jordan Peccia discussed the advantages of wastewater-based epidemiology and its future outlook for pandemic prediction. Using deep sequencing and publicly available pathogen genome databases, a long number of pathogens was found in sewage streams. Wastewater testing can be more efficient, faster, and potentially more accurate than traditional testing protocols. Using polymerase chain reaction (PCR)

on sewage sludge, researchers were able to detect a peak in COVID-19 prevalence in wastewater that correlated with the first outbreak. In Connecticut, where much of the work was completed, six daily samples tracked more than 1,000,000 residents. Where traditional testing is prompted by the onset of symptoms, wastewater testing can detect pathogens up to seven days before cases are reported. Using simple regression, case rates can be predicted from wastewater data alone. In this way, wastewater testing can provide valuable information for local, state, and federal authorities.

Further analysis of wastewater can determine changes in human behavior by evaluating chemical signatures, such as compounds found in motor vehicles, to evaluate whether the population is adhering to lockdown guidelines. Dr. Peccia closed by highlighting future directions for wastewater testing including moving towards full automation, surveillance of other infectious diseases, tracking genetic strains and evolution, and deploying the technology in the developing world.

Combined Q&A/Panel Discussion: *Bryan Grenfell, Bin Yu, Jordan Peccia*

The Q&A / Panel Discussion focused on the following key questions and discussion points:

- If major restrictions will cause a larger breakout of COVID-19 later, how can we predict the magnitude and time of the next breakout wave? And will such restrictions cause more or less total infections?
    - It ultimately depends on how the host immune dynamics operate. Using wastewater might be an elegant way of getting a handle on this.

- Recovery from measles meant permanent immunity to measles! Why it isn't this the case with COVID-19?
    - There is no such thing as permanent immunity; however, because measles is incredibly invariant, vaccines are effective and long-lasting. COVID-19 has low cross-immunity making immunity more difficult to achieve.

- Looking at other pathogens, can we develop an early warning system for pathogens that aren't as prevalent due to non-pharmaceutical interventions?
    - It is a great idea but not all pathogens are gut-trophic so many may not be detected in wastewater. For example, influenza is not found in wastewater and coronavirus is more prevalent than rhinovirus. Understanding why this is the case would help refine this methodology.

- With respect to surveillance programs, how often do you have to deal with concept drift? For example, relationship of surveillance data with questions, changes in vocabulary, sudden novel changes in how you ask questions.
    - As technology, supply chains, and methodologies advance, it is very hard to monitor the same way. Wastewater testing is valuable and can serve as an early warning system so not letting the concept drift is critically important.

      Furthermore, by developing an agile surveillance program, dynamic situations can be better adapted to. Quality control by humans is one way to ensure data quality has not drifted. To align terminology for better communication across disciplines, the research community must be come together and decide.

- Do you extract RNA from virus particles? Does anyone look at infectious activity of virus particles in the sewage?
    - Evaluating infectious activity of SARS-CoV-2 particles in sewage is not currently done. More generally, evaluation of this kind is not traditionally performed for waste streams. Moreover, for those who have the capabilities to measure infectious activity, measuring waste streams is not a high priority.

- How do we integrate social and behavioral sciences into these models? At the institutional level, how do we get these people involved?
    - Incorporation of social and behavioral sciences is critical for these models to better predict disease trajectories. Understanding the determinants of social norms and how it affects acceptance of an intervention (e.g., mask wearing) has huge implications on outcome. The pandemic is a good time to think about how to better incorporate social and behavioral sciences.

      There are economists that are thinking about very interesting questions related to human behavior and the pandemic. For example, how particular interventions can modulate transmission and whether some are risk-averse or risk-seeking. However, data quality is paramount in developing accurate, predictive models. Social and behavioral sciences can greatly help by developing more sophisticated ways of collecting high quality socio-economic data to feed into predictive models.

      In addition, the dangers and severity of respiratory illnesses are well known; however, the quantities of pathogens found in indoor environments are worrisome and would never be allowed in drinking water. In order to address this gap, interdisciplinary barriers need to be broken. More specifically, social and behavioral sciences need to be part of the process.

Physiological and Environmental Level

A Multi-Scale Systems Biology Approach towards Tuberculosis Infection Interventions: *Denise Kirschner, Professor, University of Michigan*

Dr. Denise Kirschner discussed using a systems biology approach to understand infection interventions. This approach, which integrates data from multiple model systems, including animal models, human models, immunological models, and mathematical models, can be used to address multi-scale, multi-dimensional biological research problems. These models can be integrated to build a holistic, virtual human model that incorporates all scales and can be used for many applications.

Dr. Kirschner used her work on tuberculosis (TB), as an example, to illustrate the development of a multi-scale model. Despite TB being the leading cause of the death in the world due to infectious disease, prior to COVID-19, there is still limited understanding of infection trajectory. TB is a multi-scale infection in nature with interactions at the cellular, tissue, organ, host, and population scales all affecting the disease outcome.

To best address these scales, Dr. Kirschner and her team developed GranSim, an agent-based cellular/tissue scale model that incorporates and addresses interactions at higher and lower scales.

Furthermore, GranSim integrates stochastic and discrete cellular dynamics to capture and track mechanistic interactions. The model provides a deeper understanding of the evolution of granulomas and evaluates how granulomas will react to certain perturbation. The team further scaled GranSim to develop HostSim to understand how granulomas form within the lung as a whole.

These models provide a testbed to evaluate the evolution of granulomas, reactions to specific perturbations such as knock-outs and interventions, and which interventions will translate between human and non-human primates. They are particularly useful in providing the best targets for potential vaccines and the efficacies of them through virtual clinical trials, enabling significant cost savings in vaccine development. Finally, Dr. Kirschner highlighted the barriers she faced during her work including,

- collaboration between experimental, clinical, and computational/mathematical scientists;
- integration of co-morbidity data into models; and
- access to high throughput computing.

Physics Guided Machine Learning: A New Framework for Accelerating Scientific Discovery: *Vipin Kumar, Distinguished Professor, University of Minnesota*

Dr. Vipin Kumar discussed combining machine learning and scientific modeling to build predictive models for complex systems. In physics-based models, relationships between input and output are governed by partial differential equations with parameters and equations based on first principles. Limitations of this class of models include missing or incomplete physics resulting in model bias and calibration of unknown parameters resulting in computationally expensive models.

Machine learning based models do not require fundamental physical understanding of the system and have been highly successful in commercial applications; however, the requirement of high-quality data for accurate predictions and its inability to be generalized to unseen scenarios limit its usefulness in many scientific applications. To put it simply, physics-based models are limited by what is known and data science models are limited by the availability and quality of the data.

To overcome these limitations, he elaborated his experience in developing theory-guided data science models for climate and environmental problems. He talked about how such models can be developed to harness the advantages of both. However, much more development will be required to evaluate whether this class of models are capable of outperforming pure physics or data science models, dynamically assimilate new data, and model multi-dimensional, multi-scale processes.

Climate, Oceans, and Human Health: What Cholera Can Teach Us About COVID-19: *Rita Colwell, Distinguished Professor, University of Maryland College Park*

Dr. Rita Colwell discussed the prevalence of Cholera epidemics and how the lessons learned from mitigating Cholera can be applied to COVID-19. Cholera is a bacterial disease caused by *vibrio cholerae* and is naturally exists in copepod plankton. Previous research identified poor sanitary conditions as the primary driver of infection, transmission between individuals, and a strong correlation between cholera outbreaks and sea surface temperature. A theoretical framework incorporating environmental data with socio-economic factors to predict cholera outbreaks was developed to create risk maps for epidemic prone regions. Further, this theoretical framework was adapted to create COVID-19 risk maps. Finally, Dr. Colwell highlighted that this theoretical framework for COVID-19 disease predictions can integrate

interactions at multiple scales by taking advantage of sophisticated tools not available when cholera was first being studied, enabling more accurate and refined predictions.

Combined Q&A/Panel Discussion: *Denise Kirschner, Vipin Kumar, Rita Colwell*

The Q&A / Panel Discussion focused on the following key questions and discussion points:

- Could you elaborate a bit more on how you merge synthetic and real data for doing machine learning (ML)? Do you just treat them as equivalent?
    - In order to integrate synthetic and real data, we simulate the exact same experiments in the virtual environment, such that the same type of data is generated during each experiment. These results are then integrated with experimental data to increase the machine learning data set.

- Do you think theory-guided data science (TGDS)/theory-guided machine learning (TGML) takes away from the "surprise findings" side of data science?
    - No, it does not take away from “surprise findings”. When machine learning is constrained by physical theories, as they are in TGML, it helps uncover missing physics or theories. Because all models are approximations of the natural world, if a TGML model does a better job at predicting phenomena than a simple mechanistic model, the mechanistic model is not accurately depicting real world phenomena, identifying areas where an improvement in theory is needed.

- Can you expand on what your findings were with respect to sequencing SARs-CoV-2 in wastewater?
    - DNA sequencing in wastewater testing identified the presence of SARS-CoV-2 seven to 10 days before it was reported, showing its usefulness as an early warning system. Additionally, this testing effort was able to identify variants. Wastewater streams from dormitories and assisted living facilities can be tested to target specific populations to establish early interventions in affected communities to reduce spread.

- What are the most promising alternative approaches to allometric scaling to scale animal data to better predict behaviors in humans?
    - Scaling data between small and large scales is not a one-step, linear process. There is a big black box in between and modeling is a useful tool in connecting these two domains. As an example, conducting dose-response experiments in small animal models and feeding the resulting data into a virtual dose-response experiment for large, animal models (non-human primate and human) can provide a more accurate prediction of outcomes in clinical trials.

- Can fluctuations in animal biodiversity data and their habitat be integrated with climate change data?
    - Many different data sources can be integrated together. For example, biodiversity data can be integrated with existing water systems models to create a more refined and realistic model. Many of these models that lend themselves to TGML can produce global scale results.

Breakout Session Overview

Over the course of two days, the invited speakers and discussants participated in breakout sessions focusing on each scale:

- Breakout Session 1: End-to-End
- Breakout Session 2: Molecular
- Breakout Session 3: Population and Epidemiological
- Breakout Session 4: Physiological and Environmental

Within each breakout session, breakout groups discussed the most relevant contribution from their domain of expertise that would be most valuable to pandemic prediction; data and computing needs; and interdisciplinary research challenges.

In this summary report, the ideas discussed during the breakout sessions, plenary presentations, and panel discussions have been combined and organized topically into major themes, summarized in *Chapter 2. Summary of Research Challenges*. A comprehensive summary of participant input during the breakout session is found in *Appendix C: Breakout Session Participant Input*.

2. Summary of Research Challenges

Numerous technical interdisciplinary research challenges emerged throughout the workshop discussions. The following section is organized into the four scales, as well as any additional topics that were not scale-specific. Challenges that were given additional focus in conversation are detailed below. A more comprehensive list of research challenges identified during the workshop are summarized in Table 1.

End-to-End Research Challenges

Pandemics are multi-scale, multi-dimensional problems requiring expertise across all scales. The primary challenge raised in the end-to-end scale was effectively integrating models at different scales. Models, frameworks, and theories that are effective at a certain scale may not be successful in others. For example, many unknown problems may be lurking when connecting network and agent-based models, especially at the interface. To mitigate these challenges, developing a holistic framework that incorporates multi-scale phenomena would greatly help in developing models for pandemic prediction.

In order to collect sufficient data to feed into these holistic models and establish accurate baselines, a surveillance program may be needed. Methods and frequency of sample collection and data storage need to be considered properly to balance security and privacy concerns with acquiring enough data. However, it was noted that with the current pandemic, the public may be more accepting of a viral surveillance program.

Molecular Research Challenges

Interactions at the molecular scale underpin interactions at higher levels. As Dr. Riedel noted in his plenary presentation, the cellular immune response is the first line of defense for any invading pathogen, often determining disease severity and outcome. Research challenges primarily focused on foundational understanding at the molecular scale. More specifically, bridging the gap between viral genomics and pathogen transmission rate, disease severity, and patient outcomes.

It was further noted that there is a lack of knowledge in disease prevalence. How long viruses last on surfaces, how effective common disinfectant protocols are, and whether natural antimicrobial materials can be developed are all fundamental research questions that emerged. Finally, there was discussion on lack of reproducibility in experimental results. Researchers often find that they cannot reproduce their own results or results from a published work, even though they follow the same experimental protocols. Challenges in accessing data and gaps in experimental methods in published results may be cited as justifications but on a more fundamental level, there may be gaps in knowledge that lead to such effects.

Population and Epidemiological Research Challenges

The integration of human behavior into epidemiological and population-level models is paramount for accurate pandemic prediction models. Due to the inherent unpredictability of human behavior, current epidemiological models may be too simplistic to accurately predict pandemic progression. Even with the most the sophisticated models for viral transmission and host interactions, if human behavior is not incorporated into pandemic forecasting models, they will be unable to provide accurate predictions. To amend this, social, behavior, and economic sciences must be fully integrated into pandemic research from the beginning.

Incorporation of individualistic behaviors into specific agents in agent-based models can provide a population-level digital twin with accurate predictive capabilities for community-level interactions. This more realistic and "personalized" agent-based model can alleviate the numerous security and privacy concerns associated with surveillance programs meant to collect data for pandemic forecasting models by running scenarios *in silico*.

Environmental and Physiological Research Challenges

Developing a physiological digital twin that integrates molecular, tissue, and organ level interactions will greatly advance predictive capabilities for disease outcome and intervention efficacy. As Dr. Kirschner highlighted in her keynote presentation, significant knowledge gaps exist in the physiological immune response to yield accurate development of digital twins. With deeper understanding of the immune response, individual immune responses can be mapped and integrated to create robust, fully personalized digital twins.

Incorporation of climatological data will be critical as global warming continuously changes our climate patterns. As Dr. Colwell discussed in her presentation, a strong correlation was found between environmental conditions and disease outbreaks. As our world warms and weather patterns change, the environmental drivers of pandemic emergence may be rapidly evolving and viral prevalence and disease risk may be especially heightened in regions that may be least expecting it. To combat this, the integration of more environmental data into predictive models will be critical.

In addition, with climate change, new methods for model validation may need to be developed. As Dr. Peters stressed, the traditional paradigm of using historic data and holding data back may be obsolete as climate change continuously shifts disease outbreak potential.

Data Science, AI/ML, and Computing Research Challenges

A primary focus of the workshop was to identify data and computing needs to create an effective pandemic prediction capability and is thus, the largest collection of research challenges. Limited data

access and poor data quality are two of the most common challenges researchers face across disciplines. Open access to databases, data collection techniques, and data standardizations will only encourage robust scientific discovery. Conversely, messy, inconsistent datasets, which is often the norm, limit insights and modeling capabilities. Further, integration of data at multiple scales, such as microbiome, physiological, environmental, and population data, can lead to refined insights in disease outcome and intervention efficacy. However, collecting appropriate data with the proper coverage at the right scale to feed into models for the best outcome is a still an open challenge.

An additional point of topic involved challenges associated with model validation and uncertainty quantification in a highly dynamic world. Data collection methods and model development can infringe implicit biases in the model. Using external data sources and holding back data were identified as two methods for model validation; however, with the unpredictability of human behavior, these methods may not be sufficient to validate models in dynamic environments. Further complications arise when sufficient data cannot be collected for AI/ML models. The most effective method of using limited and noisy data and overcoming gaps while quantifying uncertainty for these models will need to be addressed.

Open-source modeling was raised as an idea to democratize the development of a holistic pandemic prediction model. Coordinating and integrating different models from disparate research fields is a major obstacle in and of itself and an open-source framework could help alleviate this burden by allowing everyone to individually address their portion of the model. However, it was strongly stressed that researchers neither have the time nor the money to both develop robust modeling capabilities and make it user-friendly such that it can be incorporated into a larger open-source model.

Bridging mechanistic and data driven models was discussed as a potential path forward in developing an accurate pandemic prediction model. Combining the two approaches allows for theory-driven, data-backed models that exploit the advantages of both modeling methods. However, the proper integration of these methods and whether they are capable of outperforming pure physics or data models is still unknown.

With the collection of large data streams, a commensurately large cyber infrastructure will be required to help scale the necessary analytics, modeling, and simulation efforts. HPC will be particularly useful in this domain. With such a large repository of data, stringent security and privacy measures to keep the data secure was identified as a critical challenge.

Finally, there is a need to create a common testbed or sandbox to validate and improve population-level models. Sandboxes can accelerate development of modeling capabilities by providing a low-pressure environment for researchers to test and improve their models. It was noted that there are limited capabilities for doing this, often leading to incomplete models being deployed.

Challenges to Interdisciplinary Collaboration

Throughout the workshop, participants noted many challenges in interdisciplinary collaboration. These are not technical research questions, but rather are barriers researchers face when trying to collaborate with others. Differences in terminology present significant challenges between researchers in different fields. The same words often don't mean the same thing across disciplines. For effective interdisciplinary collaboration, a common language must first be established to mitigate any miscommunication and

confusion. There is a need for high quality outlets for disseminating this interdisciplinary research which may not fit neatly in traditionally defined venues.

Specific examples to foster interdisciplinary research include incorporating modelers when designing experiments to greatly aid in getting the most impactful data for predictive models, tighter collaboration between data/computational researchers with modelers/experimentalists, and incorporating social sciences from the beginning to integrate human behavioral aspects to develop comprehensive pandemic prediction models.

Translation of Research to Action

The scientific community must effectively communicate to provide timely, consistent, and trustworthy information to various stakeholders. During the current pandemic, it has become abundantly clear that the timescales between scientists and policymakers are drastically different. Scientists desire longer timescales to fully develop theories and models to make the most accurate predictions, while policymakers must make decisions quickly, often with incomplete information, to try to mitigate the pandemic's effects and save lives.

Similarly, researchers must effectively communicate their results to the medical community to provide guidance on the best interventions. Currently, translation of research results and data to application in the medical domain is lacking and must be clarified and improved. It was noted that clinicians often find research data unusable when determining prognosis and treatment courses for sick patients. Further, clinicians don't have the time or energy to sift through data troves or muddled results. The scientific community must provide concise, actionable information that is relevant to the medical community.

More generally, the research community, as a whole, needs to better understand how data is understood and used within each stakeholder community (scientific, medical, policymaker, and general public).

Funding and Academic Institutional Structure

Institutional structures of academia and research organizations limit motivation to work on interdisciplinary research topics. Hiring and promotion traditionally values advancing a specific discipline, de-incentivizing research faculty, especially young faculty, to explore interdisciplinary research questions. Similarly, funding agencies focus on projects that deepen the breadth of the knowledge in specific, disciplined fields and do not typically fund highly interdisciplinary work. These aspects contribute to the silo-ing of research fields making interdisciplinary research exceptionally hard to execute.

Table 1: Summary of Research Challenges
End-to-End
• Connecting models at different scales • Developing a generalized, theoretical framework to address multi-scale, multi-dimensional problems • Deploying a viral surveillance program that collects appropriate levels of data to build better baselines, while maintaining security and privacy • Inclusion of viral screening during testing of other diseases or routine check-ups to evaluate whether new variants are identified

- Chunking domains in the PIPP process
- Developing small scale networks that project reliably
- Developing models that can project from sub-cellular level to population level
- Building a team with expertise at all scales to build a holistic, accurate model
- Connecting microbiome data, metadata, and the macro data needed for prediction

Molecular

- Bridging the gap between genomics and function
- Studying pathogen prevalence with respect to surfaces and common disinfectant protocols
- Developing robust, naturally anti-microbial materials
- Developing more reproducible experiments
- Mapping of DNA to disease features. Models for severity and transmissibility exist but do not translate genomic data into disease features
- Quantifying the strength of the virus and how quickly it can infect a host
- Building a comprehensive viral genome and viral protein database
- Integrating biological processes with chemical drivers

Population and Epidemiological

- Incorporating human behavior into population/epidemiological models
- Developing agents based on personalized behavior for agent-based models, mitigating security and privacy concerns
- Enhancing data collection methods on human interactions
- Addressing inherent unpredictability of human behavior (social science theory and probability theory)
- Understanding asymptomatic transmission on a population-level
- Developing more robust experimental techniques in epidemiology to mitigate data reproducibility problems
- Collecting more types of population level data that go beyond case data
- Integrating electronic health records with determinants of health to better understand interplay of factors

Physiological and Environmental

- Improving the understanding of the immunological response of an individual and groups of individuals
- Developing a comprehensive digital twin to evaluate disease severity, disease outcome, and vaccine efficacy
- Developing a computationally intensive model that incorporates patient data with known scientific data and models
- Identifying the relationship between wastewater data and true viral prevalence in the population
- Developing HVAC sensors for pathogen detection
- Developing accurate models for physiological organs
- Collecting a more comprehensive, multi-dimensional environmental dataset for different areas around the world. Including data on biodiversity and resilience
- Understanding individual responses to pathogens. Why are some individuals asymptomatic?

- Improving methods for working with animals in the wild

Data Science, AI/ML, and Computing

- Establishing robust data sources for: social media, mobility, compliance, human interactions, and viral/protein genomes
- Deploying devices and ubiquitous computing to monitor microbes
- Establishing data format standardizations and open access of databases
- Building an open-source model that incorporates all levels and disciplines
- Improving techniques to incorporate uncertainty and variation into models and analysis
- Developing a sandbox for testing models
- Effectively addressing gaps in data for AI/ML models
- Developing innovations that advance computing capability for simulations of pandemic processes or pathways (molecular, physiological, population) while also advancing the complexity of the simulations
- Finding sources of data variability
- Establishing large-scale cyberinfrastructure to help scale data analytics, simulation, and modeling efforts
- Improving parameterization of pandemic models
- Bridging mechanistic and data driven models
- Establishing a network of better databases and annotated biobanks
- Improving model validation and uncertainty quantification
- Improving availability of reliable data to help with reproducibility
- Moving scientific apparatus to the edge (mobile devices) which may help alleviate handling of large, dynamic data streams.
- Improving understanding of biases involved in surveillance and other methods of data collection
- Improving understanding of data fusion patterns

Interdisciplinary Collaboration

- Developing a common language between research fields
- Incorporating social sciences in all stages of pandemic science
- Incorporating modelists to inform designed experiment from the start – rather than mining of data from clinical records or other post-hoc strategies.
- Overcoming the social challenges of getting the community to work together more efficiently
- Breaking down barriers between modelers and academia
- Increasing collaboration between data producers/modelers with data consumers/AI/ML researchers.
- Establishing high-quality research outlets for disseminating work in the overlap
- Coordinating data and results across fields to foster interdisciplinary work

Translation of Research to Action

- Balancing timescales (research vs policy) for predictions
- Communicating more effectively with policymakers
- Connecting policy directly to the data analysis taking place
- Improving methods of interacting with practitioners, overcoming privacy and “busyiness” constraints

- Reducing misinformation and providing various stakeholders with clear, concise, and consistent information
- Shortening the length of time between data acquisition and it being public availability
- Overcoming privacy implications

Funding and Academic Institutional Structure

- Developing a more agile funding system. Funding was not ready when it came time to support pandemic solutions (8 months!) - need action now - only well-funded organizations could act quickly.
- Providing strong funding support for very basic science/engineering and highly interdisciplinary work

Appendix A: Agenda

Day 1 – Monday, February 22

Time (ET)	Segment	Speaker
10:00 – 10:10 AM	Opening Remarks	Sethuraman Panchanathan, NSF
10:10 – 10:15 AM	Welcome Statement	Mitra Basu, NSF
10:15 – 10:35 AM	Technical Background	Symposium Chairs
10:35 – 10:45 AM	Workshop Agenda, Structure, and Processes	Emmanuel Taylor, Energetics
Session 1: End-to-End Theme Session Chairs: B. Aditya Prakash and Krista Wigginton		
10:45 – 11:05 AM	Keynote	Madhav Marathe, UVA
11:05 – 11:15 AM	Presentation	Debra Peters, USDA
11:15 – 11:25 AM	Presentation	Ruian Ke, Los Alamos National Lab
11:25 – 11:45 AM	Panel and Q&A	• Madhav Marathe • Debra Peters • Ruian Ke • Session Chairs
11:45 – 11:55 AM	Break	
11:55 AM – 12:55 PM	Breakout Sessions	Active Participants
12:55 – 1:05 PM	Break	
1:05 – 1:25 PM	Report Outcomes	Breakout Session Volunteer
1:25 – 2:25 PM	Lunch Break	
Session 2: Molecular Level Theme Session Chairs: John Yin and Paul Torrens		
2:25 – 2:45 PM	Keynote	Paul Turner, Yale
2:45 – 2:55 PM	Presentation	Marc Riedel, UMN
2:55 – 3:05 PM	Presentation	David Van Valen, Caltech
3:05 – 3:25 PM	Panel and Q&A	• Paul Turner • Marc Riedel • David Van Valen • Session Chairs
3:25 – 3:35 PM	Break	
3:35 – 4:35 PM	Breakout Sessions	Active Participants
4:35 – 4:45 PM	Break	
4:45 – 5:05 PM	Report Outcomes	Breakout Session Volunteer
5:05 – 5:35 PM	Closing Discussion	Symposium Chairs

Day 2 – Tuesday, February 23

Time (ET)	Segment	Speaker
10:00 – 10:30 AM	Opening Remarks	Emmanuel, Symposium Chairs
Session 3: Population and Epidemiological Level Theme Session Chairs: Paul Torrens and B. Aditya Prakash		
10:30 – 10:50 AM	Keynote	Bryan Grenfell, Princeton
10:50 – 11:00 AM	Presentation	Bin Yu, UC – Berkley
11:00 – 11:10 AM	Presentation	Jordan Peccia, Yale
11:10 – 11:30 AM	Panel and Q&A	• Bryan Grenfell • Bin Yu • Jordan Peccia • Session Chairs
11:30 – 11:40 AM	Break	
11:40 AM – 12:40 PM	Breakout Sessions	Active Participants
12:40 – 12:50 PM	Break	
12:50 – 1:10 PM	Report Outcomes	Breakout Session Volunteer
1:10 – 2:10 PM	Lunch Break	
Session 4: Physiological and Environmental Level Theme Chairs: Krista Wigginton and John Yin		
2:10 – 2:30 PM	Keynote	Denise Kirschner, UM
2:30 – 2:40 PM	Presentation	Vipin Kumar, UMN
2:40 – 2:50 PM	Presentation	Rita Colwell, UMD
2:50 – 3:10 PM	Panel and Q&A	• Denise Kirschner • Vipin Kumar • Rita Colwell • Session Chairs
3:10 – 3:20 PM	Break	
3:20 – 4:20 PM	Breakout Sessions	Active Participants
4:20 – 4:30 PM	Break	
4:30 – 4:50 PM	Report Outcomes	Breakout Session Volunteer
4:50 – 5:20 PM	Closing Discussion	Symposium Chairs
5:20 – 5:25 PM	Closing Remarks	Elebeoba May, NSF

Appendix B: Workshop Participants

Name	Institution
Pulak Agarwal	Georgia Tech
Les Atlas	University of Washington
Arindam Banerjee	University of Illinois Urbana-Champaign
Mitra Basu	NSF
Catherine Beauchemin	Ryerson University / iTHEMS at RIKEN
Stephen Beckett	Georgia Institute of Technology
Michal Ben-Nun	Predictive Science Inc
Phoebe Brown	Energetics
Sharada Buddha	Saint Xavier University/Chemistry
Rita Colwell	University of Maryland College Park
Angelique Corthals	CUNY - John Jay College
Jeseth Delgado Vela	Howard University
Kossi Edoh	NC A&T State University
Rebecca Ferrell	NSF
Lauren Giles	Energetics
James Glazier	Indiana University
Theresa Good	NSF
Bryan Grenfell	Princeton University
Alison Hill	Johns Hopkins University
Javen Ho	Georgia Tech
Vandana Janeja	University of Maryland-Baltimore County
Ananth Kalyanaraman	Washington State University
Harshavardhan Kamarthi	Georgia Institute of Technology
Ruian Ke	Los Alamos National Laboratory
Denise Kirschner	University of Michigan Medical School
Vipin Kumar	University of Minnesota
Astrid Lewis	U.S. Department of State
Heather Liddell	Energetics
Cynthia Lord	University of Florida
Tommi Makila	Energetics
Madhav Marathe	University of Virginia
Elebeoba May	NSF
Dana Pasquale	Duke University
Jordan Peccia	Yale University
Sen Pei	Columbia University
Avital Percher	NSF
Debra Peters	USDA ARS
Elsje Pienaar	Purdue University
Aditya Prakash	Georgia Institute of Technology
Marc Riedel	University of Minnesota
Alexander Rodriguez	Georgia Institute of Technology

Name	Institution
Ridah Sabouni	Energetics
Siqian Shen	University of Michigan
Kenta Shimizu	Energetics
Andrea Silverman	New York University
Sylvia Spengler	NSF
Ashok Srinivasan	University of West Florida
Lauren Steimle	Georgia Institute of Technology
Alex Szczuka	University of Michigan
Sindy Tang	Stanford University
Emmanuel Taylor	Energetics
Paul Torrens	New York University
Paul Turner	Yale University
David Van Valen	California Institute of Technology
Srini Venkentramanan	University of Virginia
Peter Vikesland	Virginia Tech
Krista Wigginton	University of Michigan
John Yin	University of Wisconsin-Madison
Bin Yu	University of California-Berkeley
Walt Zalis	Energetics
Preeti Zanwar	Texas A&M University

Appendix C: Breakout Session Participant Input

Greatest Contribution to Pandemic Science

The tables below show a more comprehensive summary of the breakout session input and are included for completeness. The input summarized in this section has been minimally edited, clarified, and relocated when necessary. Insights have been drawn from this data and have been highlighted in the main body.

Greatest Contribution to Pandemic Science
End-to-End
• An app for providing relevant data to end users, similar to weather forecasting. Gives an end user a comprehensive view of relevant data, with tools for aiding decision making, and incorporating feedback and learning principles. • Making sure that teams come together that involve SMEs and domain experts collaborative work • Figuring out better ways to understand the interface between people and populations • Measured viral load, how infectious the individual is at various times, and population infection rates • Systems approach to integrating biological processes with chemical drivers • Bringing the correct people together who have the right understanding • Chunking domains in the PIPP process; how each domain interacts with the end goal in a real-world simulation • Geographic influence of pandemics, or are there related factors in underlying biology? • Impacts of weather on virus spread? What cross-domain relationships are causal? • Rapid testing and DNA sequencing of pathogens for civilians • Intervention: Intersection of high-performance computing (HPC) and algorithms • Coordination between different reporting sources • Comparative understanding of viral response in animals and humans • Construct models with dynamics and controls for infectious disease analysis • Method/systems for assimilating observations from different scales into a mathematical model • Multiple data sets coming from different organizations but limits research - people are picking/choosing data to be favorable for research - need cross validation in different fields (wastewater, etc.). Could incorporate this into data for forecasting • Multiple data sets coming from different organizations - missing interoperability standards for multi-domain analysis - need those standards • Different data sets had different temporal compilations - different time scales - data set 1 and 2 may be on a different temporal scale, and without a match, it is hard to harmonize data • Determine optimal time and place to collect data in order to properly conduct surveillance • Improved methods for better surveillance and linking specific data sets to incidents in a community • Meta-analysis of existing research - so much research around COVID other viruses - take research and an AI meta-analysis - could find threads in different studies. This would require standardization for this analysis. • Addressing data analysis problems; developing a suite of data tools to address unique data issues in this field

- Economic sciences can be helpful -- need to be able to communicate limited resources -- opportunity costs and tradeoffs
- Daily information -- enables forecasting that is much more accurate (much better than what the CDC does)
- Spatial detail of the information is missing to enable -- like weather forecast. Can develop software for the clinics, what is going to happen in XX County in the next four weeks? How to ensure the data is free?
- Rapid Agile information (with surveillance) to policy makers based on research, socio-economic, epidemiology and. Interdisciplinary effort. Need evaluation to see how long-term predictions are working (impact of decisions). Change approach based on evaluation as needed.

Molecular

- Bridging techniques used in models to experimental research
- Greater ability to look at phenotypic variations; better bridge between systems in the lab
- Better understanding of the variants, and the way they differ from one another
- Tracking variants over time, and how they will evolve in the future
- Difference between intra- and inter host variation as a pathogen moves through a network
- Evolution of variants within host, and between transmissions; variability within a host, and within the population"
- From the mathematical modeling perspective; process based models to track pathogen evolution;
- Antigenic drift, due to evolution of the virus; develop models to predict how virus will evolve in the near future and relation to immunity; can be used to track emerging viral entities with antiviral resistance
- Better job at sequencing the various variants and identifying potential variants that we haven't seen yet.
- Easier ability to obtain genetic data and look at variations within a population
- Using ML models on peptides and connect to RNA based vaccines
- Develop computationally intensive models to handle various variants
- Innovations that advance computing capability for simulations and advancing complexity of simulations of pandemic processes from the organism and up
- Greater understanding on how different diseases mutate
- Further understanding uncertainty in the viral process. Understanding stochastic process and distilling it into a modeling framework
- Using synthetic biology both as a tool and model system; need ML/AI models to interact and facilitate mechanistic models to understand molecular mechanisms of host-viral interaction
- Develop a systems level approach (problem agnostic) or virus agnostic. Generalize the process.
- Cell culture testing; investing in full sequencing of what is cultures from patients, continuous monitoring
- Machine learning critical at the molecular scale
- ML at molecular scale could that be tied in at different scales
- Utility of edge computing. Shift computing closer to the phenomena at hand. This will enable information distributed more broadly and kickstart innovation. Develop this into IoT and Smart Cities. Build this into communities. Build it into local community dashboards
- How molecular characteristics impact persistence

- Scalable methods for quantifying interactions
- Likely transmission roots
- Understanding of base of virus at molecular level - the transfer of animals to humans
- Molecular interactions with objects
- Surrogate virus' (that can be safely studied)
- Having the tools to know which viruses to study

Environmental and Physiological

- How host factors and pathogen factors interact
- Predicting the physiological impacts on pathogens
- Use ML / deep learning for prediction of epidemiology parameters
- How disease severity impacts different populations
- Connecting environmental and physiological dynamics
- Spatial temporal risk model that considers environmental factors
- Predictive and uncertainty quantified; regions where we need to collect more data to improve model
- Perfecting sewage testing systems; adding automation
- Comprehensive and interoperable digital twins to predict immune response in individuals
- Accessibility of tissue samples; few efforts that have tried to create a global network of biobanks and the data they contain
- Offering access to scientists and researchers; breaking silos between biobanks
- Increasing knowledge on immunological responses -- individually and across groups
- Sampling of waste-water -- is there information we can gather there? How representative are those markers? Need franchise-able data.
- Identifying the origin of disease in a population
- intertwining between humans and environment
- Environmental factors and their influence on health of human communities
- Future understanding, COVID transmission over long distances, >6 ft
- Incorporate environmental factors into predictive models
- Reconcile impact at different scales
- Leveraging mechanistic models in ML
- Parameterizing pandemic models for specific scenarios, localities, etc., environmental conditions required to produce maps by county, etc. that may not reflect data;
- Opportunity to use ML/AI, to support image super-resolution
- Integrating immune response to in-vivo modeling: cross-model validation
- Building virtual host models
- Ability to have whole body in-host (open-source) model where you can select level of detail and then customize the model to better understand the target organ system (Demonstration models using a modular architecture to allow other researchers to build components)
- Create simulations based on future consequences of climate change
- Location aware technologies that can embed diverse data to common platforms (GIS - new type that is design to work across maps, networks, graphs, chemical signatures, IoT events)
- Modeling and describing medical interventions: what are the things I can do the system? Can I make this change?

- Within ecology - species distribution modeling - correlate against environmental habitat - early-on approach in analyzing the pandemic. Statistical top-down model; what are the traits we can evaluate for determining how pathogens come into a specific environment.
- Global model - environmental and physiological – spatio-temporal model that takes in many factors to characterize the region.
- Describing transmission through the environment - for example, can look at decay of virus as it travels through a room/through water - can provide data to build various models in this space.
- Overuse of disinfectant - providing research into resistance issues.
- Environmental radar: Accurate, more specific radar type technology for virus/bacteria detection.

Population and Epidemiological

- Using social, economic, and behavioral data
- Build very high-resolution models / simulation capabilities to describe human behavior
- Models for fine scaled movement of people in crowded locations
- Not necessary to deploy tests that are 100% sensitive, lower can still be useful
- Improved sensitivity analyses to identify mechanisms driving behaviors
- Social networks as substrates for contact networks
- Gain a sense of contact patterns associated with infections
- How to streamline who is sampled to identify when a pathogen enters a population;"
- Nailing down the relative strength of the various transmission routes for respiratory diseases
- airborne, surface contact, environmental, etc.
- Right-sizing the surveillance and model structure. Too many data sets -- modeling and data needs advancing. Understanding at what resolution the data sets are needed.
- Being able to predict when and where humans will be able to come into contact infections vectors. Better understanding the vector human contact. Can things be generalized across systems?
- Providing the specific right information to policymakers to lead to actionable policies (public at large and to specific communities like nursing homes) including the relevant stakeholders.
- Framework to deal with data analysis from multiple sources. Better understanding of data fusion patterns.
- Move beyond proxy models toward a truly behavioral agent library. Agents are adaptive -- proactive, reactive, and interactive. For everyday human phenomena and across inclusive range of demographics - including key or underrepresented groups.
- Construct models with dynamics and controls for infectious disease analysis.
- Multiple data sets coming from different organizations but limits research - people are picking/choosing data to be favorable for research - need cross validation in different fields (wastewater, etc.). Could incorporate this into data for forecasting.
- Multiple data sets coming from different organizations - missing interoperability standards for multi-domain analysis - need those standards.
- Determine optimal time and place to collect data in order to properly conduct surveillance.
- Improved methods for better surveillance and linking specific data sets (like WW data) to incidents in a community.

- Meta-analysis of existing research - so much research around COVID other viruses - take research and an AI meta-analysis - could find threads in different studies. This would require standardization for this analysis.

Information and Computing Needs

Information and Computing Needs
End-to-End
• Not enough data sets. Not enough connections between the data • Simulated data sets around populations • Building a good social network with information about properties of the nodes • Spatial and temporal data (inside organisms) • Proxy data now from social media records. Limited case study data. • Location based services data from cell phone apps. And network streams from video cameras • Local data from hospitals. Also national data • Very poor data availability on this. • Need more granularity • Need more reliable temporal networks that change over time. More diversity in the networks that we get • Data uniformity - so much of the data we have is at different local levels - not uniform. Need uniform structure. • Reliant on volunteer efforts - somewhat surprising there isn't more for unifying this data. • Standards are needed for interoperability and connectivity of the models • Need well-organized ongoing observatories • Need social media data • Better management and data science. Cross-validation needed. • Colonies of the animals are needed to do the experiments. Need different bat colonies • Surveillance, at the human level; new concept, global surveillance using new field diagnostic technology; combination of surveillance tech that uses centralized sequencing facilities and decentralized, people in the field (distributed sequencing); better, more user-friendly equipment for field collection of data; give attention to all potential pathogens, and not just the one that drives our attention at the time • Require integration between data scientists and disciplinary scientists
Molecular
• Utility of edge computing. Shift computing closer to the phenomena at hand. This will enable information distributed more broadly and kickstart innovation • Develop this into IoT and Smart Cities. Build this into communities. Build it into local community dashboards • Making data available to different fields and openly accessible • Data coordination across fields can provide more venues for research • Daily information -- enables forecasting that is much more accurate (much better than what the CDC does) • If new virus emerges, need to quickly identify and understand interactions with host cells and objects - that data would help • Surrogate virus data

- Cytokines panels (immune reports), lacking quality quantitative experimental datasets (chicken/egg)

Environmental and Physiological

- Physiological measurements
- Fine grained hierarchical biodiversity data; interactions with the food cycles, humans, animals, etc.; overlayed with public health data
- Case data; timely public health data
- Integration of meteorological and geoscience, atmospheric science, public health; currently disconnects between all
- Passively collected data; even when there is no emerging pandemic; for establishing baselines
- Abundant data for model validation; long time scales for some environmental factors; require large data sets
- Transparency around compliance, globally; influences diffusion rate
- Human behavior data, response, compliance
- Easy to access datasets required. Poor availability of rich data sets. Need better experimental techniques to improve accuracy and things we measure (and more affordable/accurate)
- Computing infrastructure
- Need access to high throughput computers
- Need access to high throughput computers.
- Models that are parallel - running replicas. Don't need fancy HPC. Need lots of processors. Optimization is an issue though
- Integration with climate science and data
- Requires persistence across very hard barriers -- outdoor and indoor positions, time scales (annual to second by second), ontologies, sensor modalities.
- Standard language to describe software tools and data is critical. Model tools as services?
- Lots of data in the wrong format - different scales (temporal, etc.). Remove interoperability from the equation and look at this as a knowledge issue. Tackling scaling issues - new information may take time to appear.
- Data is valuable and important - but there are limits on how its processed together. How do we align data to find meaningful - need to work with partners to ensure right data is generated/collected
- Full hydrological model would need to be developed

Population and Epidemiological

- Surveys, social media data, globally available
- Number of people wearing masks
- Quantitative social science
- Population data with a certain level of sensitivity
- Full genome testing for more samples, using less money;"
- Adaptation of human behavior, compared to the resolution of models
- Understanding of how human activity changes when informed about risks? Does that impact strategy and effectiveness?
- "where people go and why?
- Placing someone in space when assigning their risk of infection
- Where is home-base for infection; boundaries of social and spatial autocorrelation

- Cell phone data, seasonality; variations
- Passive data collection in social environments
- need better controls on individual privacy, sharing;
- Mobility data, county hospital level data, vaccination data with highest spatial granularity, each zip code etc. Coordination with other relevant agencies like FEMA (state or federal)
- Need to have a survey of academic teams to better understand the most useful data. In peacetime, run scenarios to test models and data needs. Platform needed to do that
- Fine-scaled human mobility data. Interactions between the different communities
- Behavioral data (e.g., intrinsic behaviors interplay with things like mask wearing)
- Need inputs from policy and public health officials on how data processing is done
- Real-time sensing and data exchange that can capture and contextualize human social phenomena in situ as they emerge (spying on everyone??). Ethics and cybersecurity.
- We lack granularity in important ways - computing needs to compare models with more complex data/models
- We lack detailed data to validate these methods in real world. We have the case data and genetic data - but need from one location to evaluate and incorporated into one model - major challenge
- Data sources are just from one specific method of coordination - maybe data sources form different fields will help validate
- Data example: temperature checks; so many devices are in play for measurement - across same type of sensing problem, there are different ways to acquire/harmonize the data. Other medical data points - how do we bring them all together? Need to fill this need
- Open-knowledge network to facilitate creation of knowledge depository - drop data in own format without parameters around to allow others to work with the data
- Missing data - bad surveillance of several viruses. Poorly characterized - don't trust clinical data

Research Challenges

Research Challenges
End-to-End

- Availability of reliable data to help with reproducibility
- Social challenge - how to get the community to work together efficiently?
- Theory doesn't move well across different fields
- Lack of reliable data
- Different methods of recording are needed
- Technology infrastructure is missing
- Interacting with practitioners -- they are very busy and privacy
- Constructing models for these tests
- How to develop small scale networks to project reliably?
- How to do model validation and uncertainty quantification?
- How to get around privacy implications
- Humans behave in unpredictable manner (social science theory and probability theory)
- Different definitions of parameters are needed

- Don't have interdisciplinary community model to start from. Unlike climate change for example
- Improved ways to communicate the uncertainty to the public and policymakers.
- Reduce disconnect between people collecting data and the people making models
- Coordinating efforts between different research groups; likely some overlap and resources could be allocated more efficiently to target specific problems
- Breaking down barriers between disciplines - funding agencies complicate this with how they distribute funds - needs to be addressed
- Individualistic - people will behave in uncertain ways in reaction to policy - need to bring in human-centered research
- Figuring out what data to collect, balancing quantity and usability
- Post-model development stage - how can we communicate with policy makers? Behavioral science support needed
- Connection of the microbiome, metadata, and the macro data we need for prediction
- Data gaps: massive sensing systems - privacy preserved but knowing enough about behavior to come to helpful conclusions. Compliance data is missing in current analysis.
- How time and energy is supported. Individual vs team science (NIH model requires you to be a PI - not great motivation for team support)
- How do we implement surveillance methods to gather data for modeling frameworks?
- How to design deliverables to engage policy makers - science community has tools, but how do we encourage policy makers and individuals to use/trust tools? Missing self-reported feedback in models - how we improve while protecting privacy? Uniformity in data packaging
- Data fusion; grand algorithm; surrogates for variables that we cannot measure directly
- Decolonizing infrastructure - in country, giving the power to countries to build infrastructure

Molecular

- Samples need to be stored around the world, requires innovations in chemistry, and related disciplines
- Samples cannot all be stored subzero in all locations
- Gain more in-situ visibility within experiments
- High-throughput phenotyping
- Can theories in network analysis be used to identify key players in the transmission process?
- Acquiring the necessary data to fill gaps between models
- Map sequencing of DNA to the features of disease; severity or transmissibility; we have models, but they do not translate to the features of disease.
- Cell biology - creating more realistic environments in the lab
- Collecting the right amount of data; the right coverage, with enough depth, at the right scale
- Bridging the gaps between genomics and function
- Quantify the strength of the virus - how quickly it infects a host - can sequence quickly, but knowing how it will behave is a challenge
- Genotype/phenotype - very unclear - there is fundamental research that still needs to be done here
- Viral genome databases. Are they curated well for quickly locating? No. And trust issues exist. This sequencing is not trivial. Details matter. There are now some helpful databases for influenza, etc. but still pieces are missing.
- Protein side: curated databases of viral protein database

Environmental and Physiological
• Working with animals in the wild is extremely challenging • High-level and low-level measurements in a single person • Diversity in their microbiome vs population level impacts • Lack of environmental data in different areas of the world; biodiversity, resilience; challenging to create these multidimensional data sets • Dealing with biophysical / natural systems, that have multiple significant factors • ML algorithms can highly overtrain on one variable • Capturing the physiological organs (very complex) - don't know the best model to capture • Better understanding of the immune system at basic science level is lacking. • Enhanced replicability and model sharing to make the science progressive and accumulative • General Research: How can we further stimulate immune system response? • Understanding the role of the environment in transmission is still important and less is known still - have found assumptions on past viruses aren't always true. Ability of organism to persist vs. amount of virus released • Integrate the physiological questions: How were some people asymptomatic? So there is a threshold variable viral load individuals are able to handle- is this based on genetics or hygiene or ethnicity? • Opportunity for sensors in HVAC systems for detection (though surface detection is still used most often) • Processing power for massive environmental data sets • Difficult to measure viruses in the air - samples are very dilute in the air - lack data in this space
Population and Epidemiological
• Deciding between developed contract tracing technologies for an application? navigating the available option space • Ethicists will be needed at all levels of research, and in interactions with vulnerable populations • Measuring contact patterns and modeling / measuring person-to-person interaction • Matching the right type of technology to the tier / population level we're investigating • On greater automation in data collection and analysis: how to do so more equitably around the world, and not just in developing countries • Can data scientists and MI be used to bridge that gap? work regularly with 'low quality' data • Data privacy concerns influence analysis and understanding of contact patterns associated with infections • Collecting data on human interactions • Quantifying exposure of people to each other (or animals, surfaces, etc.); understanding its influence on transmission • Sandbox for testing models is lacking along with its computing architecture • Lacking models and data analysis at multiple scales that can interact with different connections -- how to connect a model at one scale to another. • Integration of electronic health records into social determinants of health. Where people work/live/play integrating with mobility data (e.g., like nursing homes). Hierarchical modeling to better understand interplay of factors

- Need open shared community models for socio-behavioral sciences. Existing in climate science - possible to develop similar open models for the SB sciences?
- Improved method to integrate cognitive science. Better understanding on how decisions are being made. Community decision making needs to be integrated into computational models.
- Providing agile data to provide the appropriate intervention alerts to the public -- e.g., apps, phones
- How to Move scientific apparatus to mobile devices. Solving for connected problem of handling massively big and dynamically streaming data.
- Understanding of - Difference between causation and association
- Building synthetic sensors (connecting to synthetic models).
- Better understanding biases involved in surveillance and other methods of data collection
- Lacking evaluation metrics for forecasting and other things like measuring the robustness of signal properties
- Need to develop synthetic characters like computer games -- will that allow for unsolvable issues of privacy
- Make cross-country comparisons. Including things like seasonal patterns. How to best learn from others.
- More types of population level data beyond case data - social data like mask-wearing is very informative/social media data can reflect the attitudes of a community - how to use at the population level is missing. We don't understand how this data interacts with surveillance data. What data is informative for linking at the population level?
- Missing data - how can researchers who have expertise in social science be brought in on teams of researchers at the beginning? Social psychologists, others could help with asking the right questions. Geographers for example, anthropologists who have a deeper understanding of impacts of a pandemic in a community.
- Missing social science at the start - some issues concerning ownership. For example - ethics and data science; sense of ownership around ethics. need an ethics discussion up front, which can be difficult across disciplines. This creates issues with ownership. If done right - big payoff. If we can measure mask wearing - connect to policy is difficult.

Appendix D: Speaker Biographies

Madhav Marathe, Distinguished Professor and Division Director, Biocomplexity Institute and Department of Computer Science, University of Virginia

Dr. Madhav Marathe is an endowed Distinguished Professor in Biocomplexity, Director of the Network Systems Science and Advanced Computing (NSSAC) Division, Biocomplexity Institute and Initiative, and a tenured Professor of Computer Science at the University of Virginia. Dr. Marathe is a passionate advocate and practitioner of transdisciplinary team science. During his 25-year professional career, he has established and led a number of large transdisciplinary projects and groups. His areas of expertise are network science, artificial intelligence, high performance computing, computational epidemiology, biological and socially coupled systems, and data analytics.

He obtained his Bachelor of Technology degree in 1989 in Computer Science and Engineering from the Indian Institute of Technology, Madras, and his Ph.D. in 1994 in Computer Science from the University at Albany -SUNY, under the supervision of Professors Harry B. Hunt III and Richard E. Stearns. Before coming to Virginia Tech in 2005, he worked in the Basic and Applied Simulation Science group (CCS-5) in the Computer and Computational Sciences Division at Los Alamos National Laboratory where he was team leader in a theory-based, advanced simulation program to represent, design, and analyze extremely large socio-technical and critical infrastructure systems. He holds adjunct appointments at Chalmers University and the Indian Institute of Public Health.

Debra Peters, Research Ecologist, Agricultural Research Service, U.S. Department of Agriculture

Dr. Debra Peters, an ecologist for the United States Department of Agriculture Agricultural Research Service's (USDA-ARS) Jornada Experimental Range and lead principal investigator for the Jornada Basin Long Term Ecological Research program in Las Cruces, New Mexico, has served on the editorial boards of Ecological Society of America's journals Ecological Applications, Ecology, and Ecological Monographs. She chaired the Society's Rangeland Ecology Section, was a founding member and chair of the Southwest Chapter, and has served as member-at-large on the Governing Board. As program chair for the 98th Annual Meeting of the Society, she inaugurated the wildly popular Ignite talks, which give speakers the opportunity to present conceptual talks that do not fit into the standard research presentation format.

Dr. Peters has greatly contributed to the broader research enterprise as senior advisor to the chief scientist at the USDA, and as a member of the National Ecological Observatory Network's (NEON) Board of Directors. She has provided this amazing array of services in support of the Society and her profession while maintaining an outstanding level of research productivity and scientific leadership in landscape-level, cross-scale ecosystem ecology. Many of her more than 100 research publications have been cited more than 100 times. Her fine record of research led to her election as a Fellow of ESA and the American Association for the Advancement of Science.

Ruian Ke, Staff Scientist, Los Alamos National Laboratory

Dr. Ruian Ke is currently a staff scientist at Los Alamos National Laboratory (LANL). His research group focuses on developing mathematical/quantitative theories and tools to understand the spread of viruses, viral-immune interactions and viral evolutionary dynamics across multiple scales of biological organization, i.e. at intracellular, cellular and population scales. Since January 2020, he has been working on modeling the transmission dynamics of SARS-CoV-2 across the globe. More recently, his work

focused on characterizing within-host dynamics and immune responses to SARS-CoV-2 infection. Before joining LANL, he was a tenure-track assistant professor of mathematics at North Carolina State University between 2015 and 2018. He did his Ph.D. at Imperial College London followed by post-docs at University of California, Los Angeles and LANL.

Paul Turner, Distinguished Professor and Faculty Member, Department of Ecology and Evolutionary Biology and Microbiology, Yale University and Yale School of Medicine

Dr. Paul Turner is the Elihu Professor of Ecology and Evolutionary Biology at Yale University, and faculty member in Microbiology at Yale School of Medicine. He studies the evolutionary genetics of viruses, particularly bacteriophages that specifically infect bacterial pathogens, and RNA viruses that are vector-transmitted by mosquitoes. Dr. Turner received a Biology degree (1988) from University of Rochester, and Ph.D. (1995) in Zoology from Michigan State University. He did postdoctoral training at National Institutes of Health, University of Valencia in Spain, and University of Maryland-College Park, before joining Yale's Ecology and Evolutionary Biology Department in 2001.

Dr. Turner's service to the profession includes Chair of the American Society for Microbiology (ASM) Division on Evolutionary and Genomic Microbiology, as well as membership on the National Science Foundation's Biological Sciences Advisory Committee, ASM Committee on Minority Education, and multiple National Research Council advisory committees. Dr. Turner was elected Member of the National Academy of Sciences, Fellow of the American Academy of Arts & Sciences, Fellow of the American Academy of Microbiology, Councilor of the American Genetic Association, Chair of the Gordon Research Conference on Microbial Population Biology, and Chair of the CNRS Jacques Monod Conference on Viral Emergence. He chaired the Watkins Graduate Research Fellowship award committee for ASM, and received the E.E. Just Endowed Research Fellowship and William Townsend Porter Award from Marine Biological Laboratory, and fellowships from Woodrow Wilson Foundation, NSF, NIH and HHMI. Dr. Turner has served as Director of Graduate Studies and as Chair of the Ecology and Evolutionary Biology Department at Yale, as well as Yale's Dean of Science and Chair of the Biological Sciences Advisory and Tenure Promotion Committees.

Marc Riedel, Associate Professor, Department of Electrical and Computer Engineering, University of Minnesota

Dr. Marc Riedel is Associate Professor of Electrical and Computer Engineering at the University of Minnesota. From 2006 to 2011 he was Assistant Professor. He is also a member of the Graduate Faculty in Biomedical Informatics and Computational Biology. From 2004 to 2005, he was a lecturer in Computation and Neural Systems at Caltech. He has held positions at Marconi Canada, CAE Electronics, Toshiba, and Fujitsu Research Labs. He received his Ph.D. and his M.Sc. in Electrical Engineering at Caltech and his B.Eng. in Electrical Engineering with a Minor in Mathematics at McGill University. His Ph.D. dissertation titled "Cyclic Combinational Circuits" received the Charles H. Wilts Prize for the best doctoral research in Electrical Engineering at Caltech. His paper "The Synthesis of Cyclic Combinational Circuits" received the Best Paper Award at the Design Automation Conference. He is a recipient of the NSF CAREER Award.

David Van Valen, Assistant Professor, Division of Biology and Biological Engineering, California Institute of Technology

Dr. David Van Valen is an Assistant Professor in the Division of Biology and Bioengineering at the California Institute of Technology. His research group's long-term interest is to develop a quantitative

understanding of how living systems process, store, and transfer information, and to unravel how this information processing is perturbed in human disease states. To that end, his group leverages—and pioneers—the latest advances in imaging, genomics, and machine learning to produce quantitative measurements with single-cell resolution as well as predictive models of living systems. Prior to joining Caltech, he studied mathematics (B.S. 2003) and physics (B.S. 2003) at the Massachusetts Institute of Technology, applied physics (Ph.D. 2011) at the California Institute of Technology, and medicine at the David Geffen School of Medicine at UCLA (M.D. 2013)

Bryan Grenfell, Professor, Department of Ecology and Evolutionary Biology and School of Public and International Affairs, Princeton University

Dr. Bryan Grenfell is a population biologist, distinguished for his investigation into the spatiotemporal dynamics of pathogens and other populations. Dr. Grenfell studies processes that occur in populations at different scales and how infections move through such groups of organisms. His work is crucial in helping to control disease in humans and animals.

His research is theoretical as well as based on large datasets, demonstrating how the density of a population and randomness interact to change the size and composition of populations. Alongside colleagues from the National University of Singapore, he studied measles in developed countries and is now extending his investigations to whooping cough and other infectious diseases.

Dr. Grenfell is currently Professor of Ecology and Evolutionary Biology and Public Affairs at Princeton University in New Jersey. He was awarded the T. H. Huxley Medal from Imperial College London in 1991, and the Scientific Medal of the Zoological Society of London in 1995.

Bin Yu, Distinguished Professor, Department of Statistics and Department of Electrical Engineering and Computer Sciences, University of California – Berkeley

Dr. Bin Yu is the Chancellor's Distinguished Professor and Class of 1936 Second Chair in the Departments of Statistics and of Electrical Engineering & Computer Sciences at the University of California at Berkeley and a former chair of Statistics at UC Berkeley.

Dr. Yu's research focuses on practice, algorithm, and theory of statistical machine learning and causal inference. Her group is engaged in interdisciplinary research with scientists from genomics, neuroscience, and precision medicine. In order to augment empirical evidence for decision-making, they are investigating methods/algorithms (and associated statistical inference problems) such as dictionary learning, non-negative matrix factorization (NMF), EM and deep learning (CNNs and LSTMs), and heterogeneous effect estimation in randomized experiments (X-learner). Their recent algorithms include staNMF for unsupervised learning, iterative Random Forests (iRF) and signed iRF (s-iRF) for discovering predictive and stable high-order interactions in supervised learning, contextual decomposition (CD) and aggregated contextual decomposition (ACD) for interpretation of Deep Neural Networks (DNNs).

Dr. Yu is a member of the U.S. National Academy of Sciences and a fellow of the American Academy of Arts and Sciences. She was a Guggenheim Fellow in 2006, and the Tukey Memorial Lecturer of the Bernoulli Society in 2012. She was President of IMS (Institute of Mathematical Statistics) in 2013-2014 and the Rietz Lecturer of IMS in 2016. She received the E. L. Scott Award from COPSS (Committee of Presidents of Statistical Societies) in 2018. Moreover, Yu was a founding co-director of the Microsoft Research Asia (MSR) Lab at Peking University and is a member of the scientific advisory board at the UK Alan Turing Institute for data science and AI.

Jordan Peccia, Distinguished Professor, Department of Chemical and Environmental Engineering, Yale University

Dr. Jordan Peccia is the Thomas E. Golden Jr. Professor of environmental engineering at Yale University. His research mixes genetics with engineering to study childhood exposure to bacteria, fungi and viruses in buildings. Dr. Peccia is a member of Connecticut Academy of Science and Engineering and associate editor for the journal Indoor Air. He earned his Ph.D. in environmental engineering from the University of Colorado.

Denise Kirschner, Professor, Department of Microbiology and Immunology, University of Michigan

Dr. Denise Kirschner received her Ph.D. in dynamical systems from Tulane University in 1991, which included training at Los Alamos National Laboratories as part of her studies. She did a Postdoctoral Fellowship at Vanderbilt University with joint appointments in both Mathematics and Infectious Diseases. She joined the University of Michigan in 1996 where she is now a Full Professor in the department of Microbiology and Immunology.

At UM, she is involved with the Center for Complex Systems as well as Biomedical Engineering, Bioinformatics and at the School of Public Health. Her research for the past 20 years has focused on applying mathematical and computational techniques to study questions related to host-pathogen interactions. Her main focus has been to study persistent infections with pathogens that have evolved strategies to evade or circumvent the host-immune responses. Her goal is to understand the complex dynamics involved, together with how perturbations (e.g. treatment) can lead to health. Her work is well funded by the National Institutes of Health and she has published over 100 research papers. In addition, she serves as editor for a number of journals in both immunology and mathematics, including serving as Editor-in-Chief of the Journal of Theoretical Biology. Dr. Kirschner's original interests were in medicine, and she has now come full circle to realizing her strength in mathematics with her love of biology.

Vipin Kumar, Distinguished Professor and Endowed Chair, Department of Computer Science and Engineering, University of Minnesota

Dr. Vipin Kumar is a Regents Professor and holds William Norris Chair in the department of Computer Science and Engineering at the University of Minnesota. His research interests include data mining, high-performance computing, and their applications in Climate/Ecosystems and health care. He also served as the Director of Army High Performance Computing Research Center (AHPCRC) from 1998 to 2005.

He has authored over 300 research articles, and co-edited or coauthored 10 books including the widely used text book ``Introduction to Parallel Computing", and "Introduction to Data Mining". Dr. Kumar's current major research focus is on bringing the power of big data and machine learning to understand the impact of human induced changes on the Earth and its environment. Dr. Kumar’s research on this topic is funded by NSF’s BIGDATA, INFEWS, and HDR programs, as well as DARPA and USGS He has recently finished serving as the Lead PI of a 5-year, $10 Million project, "Understanding Climate Change - A Data Driven Approach", funded by the NSF's Expeditions in Computing program that is aimed at pushing the boundaries of computer science research.

Dr. Kumar is a Fellow of the ACM, IEEE, AAAS, and SIAM. Kumar's foundational research in data mining and high performance computing has been honored by the ACM SIGKDD 2012 Innovation Award, which

is the highest award for technical excellence in the field of Knowledge Discovery and Data Mining (KDD), and the 2016 IEEE Computer Society Sidney Fernbach Award, one of IEEE Computer Society's highest awards in high performance computing.

Rita Colwell, Distinguished University Professor, University of Maryland Institute for Advanced Computer Studies and Bloomberg School of Public Health, University of Maryland and Johns Hopkins University

Dr. Rita Colwell is a Distinguished University Professor at the University of Maryland at College Park and at Johns Hopkins University Bloomberg School of Public Health, senior advisor and chairman emeritus at Canon US Life Sciences, Inc., and president and chairman of CosmosID, Inc. Dr. Rita Colwell's interests are focused on global infectious diseases, water, and health. Dr. Colwell developed an international network to address emerging infectious diseases and water issues, including safe drinking water for both the developed and developing world, in collaboration with Safe Water Network, headquartered in New York City.

She served as the eleventh director of the National Science Foundation (NSF) from 1998 to 2004. In her capacity as NSF director, she served as co-chair of the Committee on Science of the National Science and Technology Council. Before joining NSF, Dr. Colwell was president of the University of Maryland Biotechnology Institute and a professor of microbiology and biotechnology. She was also a member of the National Science Board from 1984 to 1990.

One of Dr. Colwell's major interests is K-12 science and mathematics education, graduate science and engineering education, and the increased participation of women and minorities in science and engineering.

She has held many advisory positions in the U.S. government, nonprofit science policy organizations, and private foundations, as well as in the international scientific research community. Colwell is a nationally-respected scientist and educator, and has authored or co-authored 19 books and more than 800 scientific publications. She produced the award-winning film, "Invisible Seas," and has served on editorial boards of numerous scientific journals, including GeoHealth, which she founded at AGU in 2015.

Dr. Colwell has previously served as Chairman of the Board of Governors of the American Academy of Microbiology and also as President of the American Association for the Advancement of Science, the Washington Academy of Sciences, the American Society for Microbiology, the Sigma Xi National Science Honorary Society, the International Union of Microbiological Societies, and the American Institute of Biological Sciences (AIBS).

Dr. Colwell is a member of the U.S. National Academy of Sciences, the Royal Swedish Academy of Sciences, Stockholm, the Royal Society of Canada, the Royal Irish Academy, the Bangladesh Academy of Science, the Indian Academy of Sciences, the American Academy of Arts and Sciences, and the American Philosophical Society. Colwell is an honorary member of the microbiological societies of the UK, Australia, France, Israel, Bangladesh, Czechoslovakia, Royal Irish Academy and the U.S. She has held several honorary professorships, including the University of Queensland, Australia.

Dr. Colwell has been awarded 62 honorary degrees from institutions of higher education, including her alma mater, Purdue University.

A geological site in Antarctica, called Colwell Massif, has been named in recognition of her work in the Polar Regions.

Dr. Colwell has published a new book, "A Lab of One's Own: One Woman's Personal Journey Through Sexism in Science"