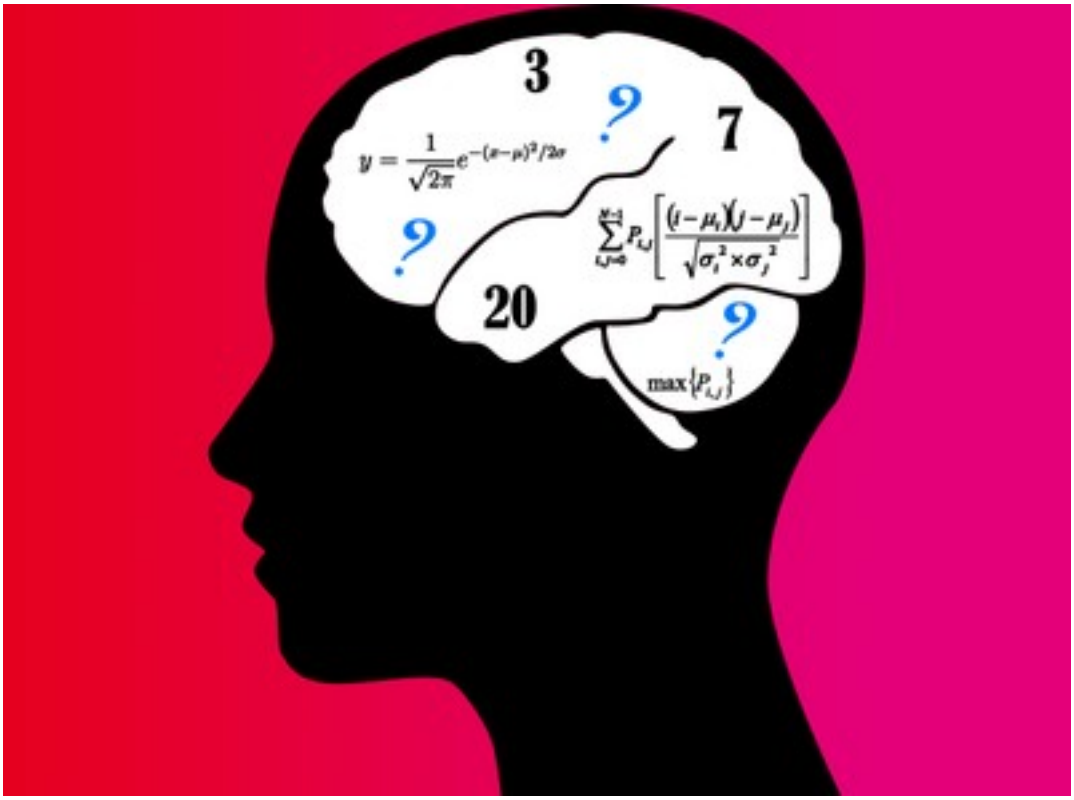

iNeuro Conference Report

- **Preparing a workforce to meet the challenges of large-scale neuroscience data**
- **Producing curricula and resources for large-scale neuroscience data analysis**

Arlington, VA • November 13 & 14, 2014

Table of Contents

A. Summary

At the iNeuro Conference in November 2014 35+ neuroscientists, library/information scientists, computer scientists, bioinformatics scientists, administrators, and educators came together to discuss how to structure training programs that will allow scientists to use large-scale data (a.k.a big data) to help advance understandings of the brain. This approach, called information neuroscience (iNeuro; a.k.a. neuroinformatics) is a rapidly emerging field that teams experimental neuroscience data with computational power and calls for training a new generation of talented scientists who can navigate both neuroscience and data science. As one iNeuro conference attendance remarked, "As technology makes it easier and easier to collect and store large amounts of data about the brain, there will be an increased need for researchers capable of exploring and analyzing these data."

High-profile initiatives such as the White House's Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative [www.whitehouse.gov/brain] and the European Commission's Human Brain Project [www.humanbrainproject.eu] recognize that with the ever-increasing production of data regarding nervous system structure and function, science has newfound and enormous potential to integrate previously discrete understandings of brain function together in profound new ways. By combining information gathered about the nervous system at different scales (molecular, cellular, structural, behavioral, etc.) through different techniques (genetic, anatomical, physiological, behavioral, etc.) neuroscience is embarking on an exciting frontier with bold new opportunities to understand the nervous system. To use an analogy, in the past scientists worked on individual puzzle pieces to understand specific aspects of brain form and function in discrete and often unconnected ways. Large-scale data approaches now give new abilities to begin to put some of these complex and numerous puzzle pieces together. By connecting previously separate information, we will gain new perspectives that will lead to more comprehensive understanding of the brain. Moreover, information neuroscience will not only allow scientists to make new links between brain form and function, but strategic integration of these rich and diverse data sets will allow scientists to make sophisticated predictive models about the brain's function that were never before imaginable.

Yet, at this important and exciting moment in time with powerful experimental, computational, and analytical tools, strong interest in uncovering the mysteries of the brain, and prominent public initiatives, very few training programs in neuroinformatics exist to prepare the next generation of scientists to harness the potential of big data to unlock some of the mysteries of the nervous system. Indeed, few scientists have the necessary fluency in both neuroscience and computer/data science to link diverse datasets in neuroscience in powerful new ways. Although interest and potential are both in place for information neuroscience to generate new knowledge, the need for talented scientists in neuroinformatics is an immediate concern. If students are not trained to use big data in

neuroscience, we will lose an important opportunity to transform modern neuroscience with the use of important new computational and modeling tools.

Although all large data sets face significant organizational challenges, neuroscience data is unusually complex because it is collected in particularly diverse and unconnected ways. By transcending scales from genetics to anatomy to physiology to behavior and examining model systems from single molecules to groups of organisms in complex environments, the datasets that describe the brain are enormously heterogeneous, ranging from genetic sequences to brain images to physiological activity patterns. This complexity and diversity makes neuroinformatics both unusually challenging as well as unusually exciting in its potential. Data describing brain form and function have never been more plentiful and will become increasingly so. Consequently, the need to stimulate strong training programs is urgent in order to create the next generation of scientists who can harness the power of large-scale data to transform it into new knowledge and ultimately wisdom. There are tremendous opportunities (and returns on investment) by using large-scale data in both education and research. An iNeuro participant remarked, "We have new computational tools that are helping to reveal the brain's mysteries, the better we can use them, the more we will learn."

Participants at the iNeuro Conference readily agreed on the magnificent potential and promise of neuroinformatics to reveal novel insights into brain function as well as its inherent challenges. Conference participants concurred that educational institutions must delineate multiple new pathways to train a new generation of scientists to contribute to information neuroscience initiatives from a variety of perspectives because few current training programs exist to meet this need. Moreover, participants recognized that life science education is transitioning to student-centered pedagogies. Articulated as a call to action in *Vision and Change* [visionandchange.org], future students will become scientists not by memorizing lists of facts, but by immersion in novel scientific problems. Neuroinformatics, as an emerging field, is particularly well positioned to borrow from best practices in neuroscience and informatics education to train future scientists encouraging them learn via team-based activities that immerse students into exciting and real problems that use data sets that describe the nervous system in various aspects. Conference attendees recommended that examples of collecting, sharing, and analyzing data can (and should) be infused early and often throughout science curricula. Both data mining and modeling should be emphasized in a variety of undergraduate science curricula where the power of large-scale data and the exciting challenges of understanding the human brain should be introduced.

Discussions at the iNeuro Conference ultimately suggested that information neuroscience curricula at the graduate level are urgently needed. Participants envisioned that graduate education in neuroinformatics can be implemented in a wide variety of ways from non-degree training sessions to certificate programs to tracks within existing degree programs to degrees in information neuroscience. Although there was limited resolution on the specific components of such curricula because of the many disciplines that come together

under a neuroinformatics umbrella, it was clear that neuroinformatics research questions are so large and diverse that they will continue to be pursued most effectively by transdisciplinary teams rather than solitary scientists. Although no single curricular formula will suffice to train information neuroscientists, existing and future curricular frameworks were suggested that included coursework in neuroscience, library and information science, and computer science along with experiences to foster strong communication and teams science skills.

A wide variety of graduate curricula that provide students with rich opportunities to gain experience and expertise in contemporary information neuroscience can be created de novo and/or from existing personnel and programs by intentional collaborations between neuroscientists and computer scientists, data scientists, bioinformaticians, and/or library scientists. Such curricula should be proposed and piloted as soon as possible. In addition some iNeuro participants envisioned the need to train data curators, a relatively new category of professional scientists with graduate training that prepares them for important roles in transdisciplinary teams of scientists with responsibilities to ensure integrity and interoperability of large data sets. In order to achieve the desired variety of training programs the scientific community needs leadership and resources to articulate explicit curricular design options, arrange strategic partnerships and build infrastructure then pilot and assess a variety of training programs in neuroinformatics.

Large-scale data will rapidly become the norm for cutting edge research. Being able to ask research questions on this scale and capitalize on this technology hinges on having a workforce that is able to work with large-scale data.

iNeuro participant (2014)

B. Introduction & Overview

For centuries scholars and artists have been fascinated with the brain at every level from molecular to philosophical. The more we learn about the nervous system, the more we realize we do not understand. Even in the simplest organism, the nervous system's abilities are beyond remarkable, its mysteries are infinitely intriguing, and its sophisticated organizational structures beyond complex. Scientifically, the brain can be studied at many levels from the ions that cross its membranes in coordinated fashions to the intricate synaptic connections that cells make with each other to the sophisticated circuits and firing patterns that underlie sensations, thoughts, and interactions. Neuroscientists are generating important new data on the brain at these dramatically different scales in vastly different model systems. As examples, biochemistry in single-celled organisms (without a nervous system) can inform our understanding of protein interactions used by signaling neurons, physiological recordings from slices of rodent brains can help reveal how drugs act at synapses, and fMRI data from humans doing sophisticated mental tasks can reveal which portions of the brain are most active during behaviors. Both the quantity and diversity of neuroscience data being produced are rising exponentially. As neuroscience generates even more large-scale data, the field desperately needs talented scientists to organize and interrogate that data to produce predictive models and transformative knowledge and wisdom that will enhance our understanding of brain function in health and disease.

C. What are Large-scale Data?

1. Large-scale Data

The term "large-scale data" (a.k.a. "big data") describes information sets that are so complex in scale and/or structure that they require sophisticated, specialized, and/or distributed computational resources in order to extract meaningful information, make predictions/hypotheses, and/or test models. An informal definition asserts that if data fits on a hard drive or can be handled by a single computer, it is not big data. Large-scale data requires parallel processing by multiple platforms. Currently, data of all types are being collected and analyzed to extract important knowledge that simply cannot be resolved by examining small or uncorrelated data sets. Large-scale data comprise a rapidly expanding array of metrics being collected in increasingly automated fashions. Moreover, nearly every sector (science, technology, business, finance, government, health, etc.) has keen interests in using large data sets to improve their enterprises. These large data sets can be relatively easy to obtain, yet are far more challenging to manage and analyze in meaningful ways.

2. Large-scale Data in Neuroscience

Neuroscientists share a common goal of understanding and explaining the form and function of the nervous system. They go about that quest using a remarkably broad array of tools, model systems, levels of analysis, and approaches. Given the complexity of event the simplest nervous system and the diversity of ways to study the brain, it is not surprising that neuroscientists work at vastly different levels of biological organization from molecules to societies using many different tools pull in very diverse animal species. As examples, current large-scale data collections created by and of interest to neuroscientists include genetic sequences, epigenetic modifications, expression patterns, signaling cascades, neuronal morphologies, synaptic connections, neuronal physiology, network activity patterns, fMRI scans, behavioral responses, disease conditions, and demographics. Bringing previously disparate data together to build models and determine emergent properties will provide countless new insights into a better understanding of brain function.

D. How Will Large-scale Data Explain Mysteries of the Brain?

1. Research

The potential to create new knowledge and understanding of the brain is by far the most important benefit of applying big data strategies to neuroscience. Large-scale data will undoubtedly facilitate our abilities to uncover previously unknown links that transcend traditional levels of analysis. Current science is producing information principally at a single level of analysis (e.g., molecular, cellular, physiological, anatomical, behavioral). Many neuroscientists are ultimately interested in linking understandings at these traditional levels in new ways to understand emergent properties of the brain. As examples, neuroscientists might use big data to ask how specific behaviors correlate with specific epigenetic markers or how physiological activity patterns in the brain predict specific actions that animals make. Thus, large-scale data provides new opportunities to identify and fill gaps in our fundamental understandings of the brain; it has the exciting potential to help merge previously disparate bodies of knowledge. Moreover, large-scale data allows increasingly sophisticated computational approaches to the brain, allowing the creation and testing of new models that can be compared to experimental data. Modeling has the important advantage of being able to substitute for experiments that are important but impossible to conduct because of ethical or resource constraints. Ultimately, by combining both theoretical and experimental data, scientists will be able to achieve a richer understanding of how the brain works.

2. Education

In addition to the obvious implications for enhancing research on the brain, large-scale data approaches also offer unprecedented opportunities for enhancing math and science education at multiple levels. The nervous system is an inherently intriguing topic with many unanswered questions and great potential to engage young minds. With publically accessible databases accompanied by calls for enhancing research transparency and access by providing raw data as publication supplements, students have exciting opportunities to query exiting data sets to learn techniques, confirm previous findings, and ask new questions from existing data. Students can learn to become scientists by actively doing science; they can grapple with real data to address unique questions. An iNeuro Conference participant commented, "current educational structures need to change to better adapt to a changing world of data availability." Thus, in both research and educational settings, the use of large data sets has the powerful ability to transform existing data into both new knowledge and new scientific talent in neuroscience and many other areas of study.

E. What are the Challenges of Large-scale Data in Neuroscience?

Although the challenges of large-scale data are numerous and substantial, these challenges need to be addressed in the immediate term because large-scale data holds remarkable potential for making important advances in understanding the brain that are simply not achievable by current means. Most critically, science needs a workforce excited by and capable of addressing these challenges. Every step of gathering, organizing, relating, maintaining, protecting, and interrogating large-scale data in neuroscience includes challenges. Conference conversations acknowledged the pressing need to address the practical challenges that will limit ability to harness the power of large-scale data effectively. Data are not useful if they cannot be understood. Simply creating data sets is not enough; data sets need to be organized and curated in smart ways that the information they contain can be shared, related, and interrogated by multiple groups. Moreover, knowing what data types are useful (and which are not) is important as summarized in the maxim by William Cameron (1963), "not everything that can be counted counts; and not everything that counts can be counted." Specifically, the challenges of using large-scale data in neuroscience fall into three broad categories:

1. Standards

An absence of standards, minimum requirements, and/or best practices for collecting and curating data create open questions for the field such as: how should data be defined, prioritized, configured, stored, blinded, annotated (workflow and metadata), verified, reviewed, replicated, authenticated, combined, owned, shared, attributed, maintained, sustained, supported, scaled, and protected? These many unanswered questions reveal that the scientific community has a large, urgent, and important task at its doorstep to create and sustain necessary standards for data as soon as possible so that large data sets can be shared in efficient and coherent ways. This challenge can be met by an infusion of talented neuroscientists with fluency in working with large-scale data who will provide leadership to establish standards and best practices for future work. In developing and implementing standards in neuroinformatics, there is much to be learned from communities such as bioinformatics and information science that are addressing similar challenges. Ultimately, a broad community of stakeholders must collaborate to determine best practices. Funding agencies such as the National Science Foundation (NSF) and National Institutes of Health (NIH) and collaborative organizations such as the International Neuroinformatics Coordinating Facility (INCF) and the Neuroscience Information Framework (NIF) are the most likely centralizing factors to generate, support, and sustain standards. NSF's 2011 requirement for data management plans (DMPs) in all grant proposals is important (though small) first step in developing practices that encourage and facilitate mindsets for protecting, organizing, and sharing data throughout the scientific community. Similarly, INCF's task forces on standards for data sharing that concentrate on electrophysiology and neuroimaging were viewed as critical community initiatives to reach these goals. Future data standards, support mechanisms, and sharing platforms must be developed from both the perspectives of funding agencies (top-down) as well the scientists who generate and use data (bottom-up).

2. Integration

Second, how can large, hierarchical, heterogeneous, and/or incomplete data sets be integrated coherently into interoperable repositories? The particularly vast differences in experimental methods, levels of analysis, and model systems make connecting individual databases a particular challenges in neuroscience. For example, how can neurophysiological data from a specific neuron or network be considered in the context of relevant genomic and anatomical data in ways that will allow useful new insights into brain function? The challenge of integrating existing data sets is a substantial task that demands a talented workforce capable of integrating existing data and anticipating ways to integrate future data sets that have yet to be imagined or collected.

3. Training

In addition to the development of best practices, infrastructure, and standards that will make it achievable to collect data sets in ways that will facilitate meaningful insights into brain function, it is critical that all scientists be trained with the skills necessary to navigate these large data sets. Urgent needs for the development of best practices and data standards repeatedly emerged throughout iNeuro conversations because such standards and practices are necessary foundations of curricula to educate future scientists to become adept at working with large-scale data. At the same time, training programs simply cannot afford to wait to design curricula until common standards are in place. Strong programs will prepare their trainees to contribute directly to the conversations that build and revise such standards. Moreover, a recent call to reform life science education [visionandchange.org] recommends undergraduate curricula be transformed from a focus on content or unsustainable lists of facts to be memorized learned by passive means to a focus on conceptual frameworks and the process of scientific inquiry via student-centered, active courses.

F. How Can Training Facilitate Advances in Information Neuroscience?

1. Foundational Questions

To address how current and future training programs can facilitate the use of large data sets to reveal new information about brain function, iNeuro participants considered several organizing questions such as:

- What **skill sets** does a scientist/curator of large-scale neuroscience data need?
- What **disciplinary training** does a scientist/curator of large-scale neuroscience data need (neuroscience, computer science, information science, mathematics, etc.)?
- What **degree** level(s) should these individuals hold (AA, BA/BS, MA/MS, MD, PhD, etc.)?
- Are **new and/or existing training programs** sufficient to generate individuals with the desired skills?
- What is the desired **curriculum** in programs that train individuals to use large-scale neuroscience data?
- How can the ***Vision and Change*** **recommendations** for transforming undergraduate life sciences education inform the curricula training individuals using large-scale neuroscience data?

2. Skill Sets & Disciplinary Training

It is unlikely that scientists using big data to decode the mysteries of the brain will be able to do so from strictly computational approaches, without at least some foundational knowledge of neuroscience concepts, dimensions, and experimental methodologies used to generate the data they are analyzing. At the same time, it is also unlikely that a robust understanding of experimental neuroscience by itself without knowledge of computation or informatics will be sufficient to navigate big data in ways that deepen our understanding of the brain. Thus, neuroinfomatics requires scientists with experience in both the "wet" bench sciences and the "dry" computational and data sciences. Very few individuals will be able to invest the time necessary to develop fluency in both areas. Instead, there is a great need for transdisciplinary training programs that help students become proficient in both the neuroscience and the computation (or perhaps fluent in one and conversant in the other). It is expected that transdisciplinary teams of individuals will make advances in neuroinformatics, each member bringing strong expertise in one aspect and sufficient knowledge in complementary aspects to engage productively with professionals who bring different skills and expertise to the table. Moreover, individuals with the interests and skills to advance neuroinformatics will need to navigate dynamic and evolving conditions over the long term because new methods of acquiring, organizing, and analyzing data sets will undoubtedly continue to be developed. Accordingly, information neuroscientists must expect to continue learning and expand their skills throughout their careers. Neither single dimensional skills in one area, nor static skill sets will be sufficient for an individual to contribute successfully to questions that rely on neuroinformatics approaches.

Consequently, a consensus repeatedly emerged among iNeuro participants that effective information neuroscience training programs will feature active, team-based transdisciplinary experiences to prepare a new generation of scientists with skills to collaborate effectively with peers who have different yet complementary expertise. Such hands-on challenges with real data sets are ideally integrated throughout a curriculum as case studies, assignments, examples, capstones, and/or theses. Importantly, such training programs also need to emphasize skills, traits, and environments where continuous development and learning is expected to extend far beyond the boundaries of the formal training period. A training program needs to provide its students with opportunities to develop an understanding of contemporary neuroscience research, gain experience with transdisciplinary approaches to interesting and applied problems, understand both data curating and data sharing through first-hand experience working in large-scale and centralized databases, be able to work across scales and modalities (genetic, molecular, cellular, physiological, anatomical, behavioral, etc.), and understand experimental designs and workflows. The inherent challenges of multidisciplinary training in any field of study (scientific and beyond) were noted. These challenges are both theoretical and practical and in no way unique to neuroinformatics. Such challenges include differences in vocabularies, cultures, criteria, protocols, priorities, and organizational structures. Much can be learned from successful interdisciplinary training STEM programs that have been established. Moreover, neuroscience itself is an inherently multidisciplinary area of study and most existing neuroscience training programs effectively navigated such challenges in bringing together multiple disciplines in an educational framework.

3. Necessary Skills for Information Neuroscientists

Participants at the iNeuro Conference outlined numerous dimensions of skills and knowledge that will make a student most likely to make new insights into the brain via large-scale data. These competencies fit into four general categories as described below that emphasize research, computation, strategy, and relational skills.

a. Research Skills

Conversations at iNeuro spent relatively little time discussing wet lab or bench research skills because these skills were relatively easy to identify as fundamental principles and experimental methodologies currently forming the basis of nearly all existing undergraduate and graduate neuroscience training programs. Such wet neuroscience skills are firmly based in the life sciences overlapping considerably with fields such as biochemistry, genetics, cell biology, physiology, anatomy, medicine, and behavior. It was also acknowledged that neuroscience training intersects meaningfully and abundantly with statistics, engineering, math, physical sciences, social sciences, computer sciences, and health sciences. Although no individual neuroscientist will be trained in all the interconnecting fields mentioned, whatever suite of experimental techniques that a student learns as part of her/his training, emphasis on strong experimental design and analysis was expected to be foundational to all neuroscience programs.

b. Computational Skills

In comparison to the wet neuroscience research skills and knowledge, computational and modeling skills received far more attention in conversations, drawing directly from skills emphasized in existing quantitative training programs including: computing principles, high performance computing techniques, data visualization, programming, database design, web technologies, and data transfer methods. Additional skills from library science, data science, and/or informatics programs were also expected and included examples such as: understanding existing resources, data formats, standards, vocabularies, lexicons, ontologies, semantics, lifecycles, workflows, annotation, metadata, and interoperability. Finally, necessary skills from the quantitative sciences included: data analysis, machine learning, programming, coding, scripting, probability, statistics, signal processing, imaging, and standardization of workflows.

c. Strategic Skills

Conversations at iNeuro describing how scientists interact with large data sets frequently generated lively terms that went far beyond "managing" data toward more active verbs that included: hacking, curating, translating, wrangling, stewarding, and advocating. Although each individual term has important and distinct nuances, taken together this collection suggests current and future neuroinformatics practitioners need to be particularly imaginative, nimble, collaborative, and strategic if they are to engage effectively with large and diverse data sets as well as with other scientists who create and interrogate the data. No matter how well organized, collections of data are fashioned, profound new insights into how the brain is organized and operates will require savvy and creative minds with perseverance and creativity to overcome new challenges at multiple stages of transforming data into powerful new knowledge.

d. Relational Skills

In addition to disciplinary and attitudinal skills described above, iNeuro participants acknowledged that scientists best positioned to make advances in neuroinformatics also need skills and experiences in communication, collaboration, and ethics. Strong interpersonal communication skills are critical to collaborations that transcend multiple boundaries (disciplinary, structural, institutional, international) to build successful teams that communicate effectively and efficiently. Moreover, robust written, oral, and visual communication skills are needed to communicate research outcomes with a wide variety of audiences from scientists to administrators to policy makers to the general public. Finally, future scientists should balance responsible stewardship of shared and open data for the scientific community's use with the relevant ethical and legal understandings of sensitive privacy, legal, licensing, and attribution responsibilities for various data types.

4. Degrees – Associates vs. Bachelors vs. Masters vs. Doctoral

iNeuro conversations focused on undergraduate foundations and developing graduate degrees as pathways to encourage and train young scientists to use large-scale data to understand the brain. Few, if any, participants suggested that an undergraduate degree alone could provide sufficient training, given the depth and range of skills expected for information neuroscience, though the necessity of strong undergraduate training was

repeatedly acknowledged. Similarly, there was little support for a single, stand-alone neuroinformatics course at any level as sufficient training. Instead, conversations assumed that excitement for and skill in neuroinformatics would be best achieved by infusing neuroscience and informatics throughout curricula so that students have repeated and varied exposures to and experiences in information neuroscience in multiple contexts.

Despite the emphasis on graduate training, the need to introduce the principles and excitement of using big data in neuroscience as well as cultivating an ethos of generating and sharing data as part of undergraduate scientific training, were widely and repeatedly endorsed as fundamental throughout a wide variety of existing undergraduate majors most likely to produce neuroinformaticians (biology, computer science, engineering, informatics, math, neuroscience, psychology, etc.). Simply put, graduate school is far too late for a scientist to have a first encounter with the power and utility of big data, to work with a computational model, to practice good data sharing habits, or appreciate the mysteries of the nervous system; these elements must also appear throughout undergraduate science curricula. Most undergraduates will not likely have deep understandings or skill sets in neuroscience, information science, and computation, but they should emerge from college with strong disciplinary skills and experiences in one area that have primed them to see the potential of large-scale data applied to the brain. If undergraduate programs prepare STEM students with the skills and experiences necessary to see how neuroscience and computation collaborate to make a powerful insights into neural function and point them toward strong, transdisciplinary graduate programs, then a new generation of scientists can develop the skills and knowledge to use big data to understand the brain in exciting new ways.

5. Data Professionals in Neuroinformatics (Wranglers & Curators)

Scientific teams addressing novel questions in neuroinformatics will benefit tremendously by strategically employing data specialists who may not directly create or analyze data, but can assume the important responsibilities of organizing, annotating, and/or making data accessible. Data specialists were envisioned as a new professional position with training at a M.S. level or beyond who work closely with neuroscientists and computational experts. At the acquisition stage a data professional in something of a data wrangler role makes critical contributions by ensuring all data are collected and organized in appropriate ways to facilitate its efficient use in hypothesis-driven scientific inquiry. At post-acquisition stages, a data professional in a curator role transports and maintains data appropriately within repositories so data may be integrated meaningfully with additional data sets and accessed by others. Incorporating professional data positions such as these wranglers and curators acknowledges the significant and necessary (often overwhelming) tasks of collecting, developing, and organizing interoperable data sets with appropriate metadata so data can be shared broadly and mined deeply. Data professionals might be employed in a variety of ways: by individual labs, by transdisciplinary teams, as consultants, and/or by institutional core service providers. While the utility of professional data specialists was acknowledged, it was challenging for iNeuro participants to delineate a specific training curriculum for this critical role. Data professionals were viewed as

essential members of transdisciplinary teams who likely possessed strong training in computer science, data science, and/or library science as well as interest or experience in neuroscience. Moreover, data professionals play important roles in developing and upholding much-needed standards and best practices for ensuring data consistency, quality, and interoperability. In these roles, data professionals are also important hubs, connecting members of transdisciplinary teams with distinct expertise and facilitating new insights. An iNeuro attendee commented, "Neuroscience data cannot be used to their fullest extent without dedicated personnel concerned with their curation."

6. Identifying Existing Training Programs & Building New Training Programs

a. Existing Training Programs

Although it is very easy to identify numerous examples of strong undergraduate and graduate training programs in disciplines such as neuroscience, bioinformatics, computer science, and data science, it is more difficult to identify existing training programs that intentionally coordinate these disciplines to train students in ways that specifically advance neuroscience via big data. Some students within these traditional graduate programs are unquestionably doing research in neuroinformatics as part of their training. Yet, only one graduate training program, the Computational Neuroscience and Neuroinformatics graduate program at the University of Edinburgh, was cited by iNeuro participants as an example where a specific graduate curriculum was constructed to train students to use large-scale data in neuroscience [http://www.anc.ed.ac.uk/dtc/]. Similarly, a single specific neuroinformatics undergraduate degree at the University of Warsaw was identified [http://neuroinformatyka.pl]. The Warsaw curriculum, based in biomedical physics, intentionally integrates traditional undergraduate coursework in biology, mathematics, and physics along with research apprenticeships in neuroscience labs to prepare undergraduates to understand and analyze neurophysiological data and advance to graduate programs. The Edinburgh PhD program begins with an emphasis on the range of inquiry and methods in neuroscience, then stresses computational and informational expertise to prepare its graduate students to relate theory to experimental during their thesis research. Additional examples of undergraduate and graduate programs in computational neuroscience and neuroinformatics later came to light at institutions in the US, Canada, Europe, Israel, and Japan [www.incf.org/resources/training]. Not surprisingly, the curricular structures of these programs feature strong quantitative skills and neuroscience context via a wide variety of curricular models that take advantage of local expertise and resources.

b. Creating Training Programs from Existing Resources

Most scientists currently engaging in neuroinformatics developed their skills through ad hoc training fueled by a combination of curiosity, necessity, personal motivation, and accessible resources. Very few are products of a coherent training program or intentional institutional structure in neuroinformatics. Instead, they built the professional networks they needed. iNeuro participants commented that scientists at many institutions likely have access to existing personnel and resources that might allow a meaningful assemblage

of essential constituencies (neuroscience, data science, bioinformatics, etc.) necessary to navigate neuroscience databases to generate new knowledge.

It is easy to identify institutions where many of the essential curricular elements of neuroinformatics training are currently in place, yet it is unusual to identify institutions where those elements are coordinated or encouraged to come together to address questions of brain function at new levels. Many iNeuro participants indicated that numerous public and private educational institutions have significant and diverse expertise in house, but lack incentives or structure to coalesce into research teams and curricula addressing information neuroscience. If individuals currently housed in existing departments or programs could be catalyzed around common neuroinformatics goals and provided with appropriate resources, then new educational programs could be developed relatively rapidly and easily. With the proper catalysts, new degree programs in neuroinformatics could be created largely from existing substrates at many institutions. Moreover, many universities have considerable experience building and sustaining transdisciplinary and/or interdepartmental graduate programs in related areas such as life sciences, neuroscience, bioinformatics, applied computation, etc. The lessons learned in creating and sustaining other interdepartmental graduate programs translate readily to launching neuroinformatics training programs.

Institutions not interested or able to construct discrete neuroinformatics degree programs could also train bright young scientists for neuroinformatics futures by creating curricular emphases (or tracks or certificates) that promote "cross-training" within existing graduate programs such as an informatics track within a neuroscience graduate program or a neuroscience track within a computer science graduate program. Similarly, an informatics certificate pathway as an add-on open to graduate students, postdocs, and other professionals in a variety of programs such as physiology, neuroscience, cell biology, cognitive psychology, etc. may be a structure better suited to some institutions. In addition to creating tracks or certificate programs, iNeuro participants suggested that existing neuroscience graduate programs looking to prepare their students for a big data future should include a quantitative informatics component as part of the training for *all* neuroscience graduate students. Even those students who may not use big data in their graduate thesis research face a future in which large-scale data will be part of many scientific discourses. Moreover these students training in "small data" labs can benefit tremendously by learning quantitative skills and best practices for data sharing and experiencing transdisciplinary collaboration. Regardless of how institutions build neuroinformatics programs, the necessary catalysts will depend on inherently unique institutional cultures and resources. Appointing personnel, reconfiguring research spaces, creating and supporting infrastructure, articulating priorities, strategically recruiting talent, and/or targeting funding to priorities are a few examples of the many stimuli needed to create and sustain neuroinformatics training programs.

7. Neuroinformatics Curricula

Preliminary curricular design drafts that emerged from conversations at the iNeuro conference were, not surprisingly, both ambitious and diverse. Regardless of emphasis or level, however, a singularly important and essential priority of future neuroinformatics training emerged: active, multidisciplinary, team-based learning on genuine and compelling challenges in neuroinformatics using real data sets must be key features of any neuroinformatics training. Learning science by doing authentic scientific inquiry was strongly endorsed by iNeuro conference participants as a necessary approach to train emerging scientists. Because neuroinformatics is such a contemporary, multidisciplinary, and rapidly evolving field, for its training programs to prepare scientists effectively for large and complicated challenges of understanding the brain through large-scale data, these training programs must be ambitious, forward-looking, and focused on developing skills and attitudes. Insisting on designing training programs that immerse students in hand-on, authentic, research experiences with unknown outcomes aligns remarkably well with recommendations articulated in *Vision and Change* [visionandchange.org]. Though *Vision and Change* addressed undergraduate biology curricula specifically, its recommendations that curricula emphasize core concepts and competencies, data fluency, student-centered learning, community engagement, and strategic partnerships all echo curricular elements emphasized by iNeuro curricular conversations.

a. Undergraduate Foundations

In creating graduate programs or tracks in neuroinformatics, undergraduate curricular preparations for such graduate work are essential considerations. The next generation of successful life scientists will undoubtedly need a computational foundation; wet bench skills are necessary, but insufficient to navigate in the "-omics" age of life science where data can be mined in genomes, epigenomes, proteomes, metabolomes, connectomes, interactomes, and many more dimensions. Conference participants suggested that competitive applicants to MS or PhD neuroinformatics degrees would likely enter with some undergraduate training in several (but rarely all) of the following diverse disciplinary foundations:

Computing Theory
Database Design
Web Programming
Data Structures
Script Writing
Statistics
Research Methodology and Design
Ethics
Intellectual property
Neuroscience
Biology
Physical Science
Engineering
Psychology.

Regardless of major, both quantitative literacy and hands-on research experience with at least one novel scientific question were essential and expected components of an undergraduate degree. In addition to disciplinary content knowledge, iNeuro participants emphasized that undergraduate science curricula should cultivate a mindset of good data habits where students both learn the value of collecting strong and reproducible data and develop an ethos that strongly encourages sharing that data with others. Finally, encouraging undergraduates to develop creative and hacking mindsets that allow them to view challenges as exciting open frontiers to be navigated (rather than obstacles) will prepare them for success in information neuroscience. Such emphases on experiential learning, original research questions, and thoughtful integration of quantitative skills with the life science are cornerstones of the *Vision and Change* call to action. These *Vision and Change* recommendations aim to transform undergraduate biology education by creating curricula that emphasize foundational concepts (not facts or details that can be looked up), and provide novel educational experiences with scientific problems that allow students to develop scientific skills and cultivate inquisitive and flexible mindsets.

b. Masters Curricula

Although conversations at iNeuro produced recommendations for undergraduate preparation in traditional, existing disciplines, there was less agreement regarding how graduate curricula should be structured. A comprehensive list of knowledge and skills desired far exceeds what training might prudently and sustainably fit into a maters degree trajectory of two to four years of full-time study. Necessarily, a MS degree in informatics would need to address both breadth across transdisciplinary fields and depth within an area of expertise. Hands-on experiences using large datasets with transdisciplinary teams to investigate original questions in neuroscience were expected foundations of any graduate degree. Such work would provide experiences with not only large-scale data, but with team science and open-ended research challenges. Additional elements of masters degree programs in neuroinformatics included curricular elements addressing:

- Neuroscience
  - methodologies
  - research techniques
  - data collection
  - analysis
- Library and Information Science
  - metadata
  - annotation
  - data management
- Computer Science
  - machine learning
  - data mining
  - coding
- Communication
  - data visualization

writing
speaking

c. Doctoral Curricula

As described above for MS degrees, the list of desirable knowledge and skills for a PhD in neuroinformatics goes well beyond what can be accomplished by most students during four to six years of full-time doctoral work. A PhD in neuroinformatics expects both breadth across transdisciplinary fields and depth within an area of expertise and indicates more substantial original research experiences using large-scale datasets in collaborative, transdisciplinary teams environments to investigate original questions in neuroscience. In addition to the curricular elements for MS programs, desirable elements of doctoral program in neuroinformatics included: math (probability, statistics, linear algebra), machine learning, information technology, systems, and networks. PhD training in neuroinformatics would include both the wet and dry aspects of contemporary information neuroscience, expecting students to produce PhD theses that directly linked laboratory experimentation (and/or validation) with modeling or informatics using large-scale data sets.

d. Non-traditional and Non-degree Training

Neuroinformatics training should not be limited to graduate degree programs, nor can a graduate degree provide a scientist will all the skills necessary for success over a career trajectory given how rapidly neuroscience, computation, and data technologies evolve. All scientists engaged in neuroinformatics will need, at multiple points, to get up to speed, keep up with changes, and/or learn about new resources, knowledge, tools, and strategies. iNeuro participants recognized that a holistic educational strategy goes well beyond graduate degrees, acknowledging the dynamic nature of neuroinformatics as well as the diverse and understandably incomplete expertise of individuals who engage in information neuroscience. Therefore, it is necessary to make information neuroscience training widely accessible for all those who aim to use large-scale data to comprehend the brain.

Collectively, non-degree neuroinformatics training methods must serve many different constituents with goals of both increasing audience and broadening participation through a variety of formats. Individually, each training experience will be most effective if it has a clear topic, sharp focus, and well-defined target audience. Examples of such training units might include tutorials, seminars, bootcamps, MOOCs, workshops, short courses, jamborees, on-going training plans, trainer training, and hack-a-thons. Individuals, teams, universities, professional societies, government agencies, private foundations, and businesses are all appropriate sponsors of such non-degree training opportunities. As discussed for graduate curricula above, because neuroinformatics assumes a team approach to scientific discovery, many of these continuing education formats will necessarily emphasize and organize different teams with different backgrounds collaborating, potentially at multiple levels of work.

e. Applying Neuroinformatics Training Beyond Neuroscience

Although by name neuroinformatics may sound like a highly specialized program of study, iNeuro participants envisioned the skills trainees acquire in working with large-scale data will be broadly applicable well beyond the research sciences to extend to public and private pursuits of many types. Because nearly all sectors are actively exploring frontiers and opportunities in large-scale data, iNeuro participants envisioned that most neuroinformatics training programs have strong potential to produce graduates that will be able to apply their coding, quantitative, and analytical skills in domains well beyond neuroscience if desired. Challenges of curating data and shortages of talented people (a.k.a. the "big data gap") are in no way limited to scientific research; businesses and industries well beyond the sciences are also very concerned with developing sufficient talent in this area and have developed informatics training programs often within business schools. Consequently, training programs in neuroinformatics are likely to produce graduates with skills that will be valuable across myriad sectors using big data. In fact, some training programs might consider partnering with industry in order to provide students with case studies, data sets, exercises, or internships directly that reveal the links between information neuroscience and other endeavors. Thus, talented students with a passion for scientific questions, but reluctant to commit to a future in scientific research perhaps because of gloomy prospects for funding and/or academic job opportunities, may consider neuroinformatics training because the quantitative and transdisciplinary skills cultivated in neuroinformatics are both marketable and readily transferrable to other sectors in ways that traditional scientific training may not be. Current conversations in the academy are acutely aware that most PhDs in the life sciences will not become tenured faculty members and thereby investigating best practices for training for "alternative" or non-academic careers. Thus, graduate programs in neuroinformatics will be uniquely positioned to allow graduates a variety of academic and non-academic career paths that may be less dependent on federal funding.

f. Scientific Rigor, Data Sharing, & Reproducibility

Although the topics of scientific rigor, reproducibility, and data sharing were not explicitly listed on the iNeuro Conference agenda, these topics emerged in interwoven ways in many discussions. All scientific training programs must emphasize rigorous experimental design, analysis, and ethics to ensure the contemporary scientific record presents the best possible understanding of the natural world. Contemporary media reports suggest a crisis of confidence in peer-reviewed scientific results that subsequently do not pass tests of reproducibility due to inappropriate reagents, bias, design, and/or analysis with neuroinformatics research included in these concerns. Efforts to enhance transparency, access, rigor, and replication include statements of best practices in research design and analysis, symposia, courses, editorials, manuscript checklists, data management plans, data sharing expectations, and data repositories. Training programs as early as the undergraduate level have strong potential to emphasize these important principles in active, hands-on ways advocated by *Vision and Change* by incorporating reproduction or reanalysis in ways that allow students to make contributions by verifying or updating the scientific record as they simultaneously learn to create, share, and analyze large data sets.

Consequently, information neuroscience with its emphases on design, interoperability, modeling, statistics, and careful analysis is well positioned to lead initiatives that encourage thoughtful data annotation, repositories, electronic lab notebooks, open source code and other best practices. In doing so neuroinformatics students can take active roles in confirming and/or correcting the scientific record as they learn. When datasets and code are made widely accessible students and scientists in resource-limited situations such as undergraduate programs, small institutions, and/or poorly funded labs can advance science by confirming, correcting, or annotating previous analyses. Changes in both incentives and infrastructures are necessary to encourage scientists to share their data and invest time investigating questions of reanalysis and reproducibility. The scientific community must acknowledge the considerable time and talent required to create, curate, share, integrate, and interrogate large data sets. With the current emphasis on novel discovery by largely independent research labs seen as key to obtaining jobs and funding, incentives that make sharing interoperable datasets through affordable and accessbile repository structures are critical both to enhancing scientific rigor and to providing opportunities to all skilled scientists to contribute to the advancement of scientific knowledge. As one iNeuro attendee stated, "Neuroscience data of many types are rapidly growing. They cannot be effectively used nor fully appreciated without thoughtful and consistent curation." A new category of trained data professionals who can shift such daily burdens of data automation, annotation, and/or analysis away from the experimental scientists has strong potential to advance both scientific rigor and access to data sets.

g. Assessing, Sustaining, and Improving Neuroinformatics Training Programs

It is important to note that in creating neuroinformatics training programs of all types, the designers and providers need to delineate clear and measurable educational goals at the outset. A careful assessment of any curriculum necessarily measures how well students develop the concepts and skills deemed most important and foundational to the program's design. Metrics such as student applications to, satisfaction with, and completion of the degree program are important indicators. As well, program outcomes such as research catalyzed, methodologies devised or improved, data sets created or analyzed, presentations given, papers published, and placement of graduates all provide valuable information to assess a program's ability to meet its goals by making adjustments and improvements. As neuroinformatics programs are deployed in various forms these goals should be regularly assessed to make smart changes that allow this new and rapidly evolving transdisciplinary training to as effective as possible in producing much-needed smart and nimble scientists who can harness the potential of large scale data to advance understandings of the nervous system.

G. Building Solutions for Training in Information Neuroscience

1. The Problem

An urgent problem facing neuroscience is how to train scientists who will transform the breathtaking power of large-scale data at multiple scales into novel insights explaining how the nervous system forms and functions. Long gone are the days when scientists worked in relative isolation, built most of their instruments, purified their reagents, wrote their own code, and reasonably mastered all the relevant literature of their field. Many contemporary scientists use kits, equipment, and software without fully understanding the components of the tools they rely on to do their research. Additionally, today teams of scientists report the vast majority of contemporary scientific discoveries. Such changes have undoubtedly accelerated the pace of scientific discovery and created scientists who are more narrowly specialized, more broadly aware, and more collaborative in practice than scientists of previous generations. As we embark on the exciting frontiers of information neuroscience, fueled by high-profile initiatives and impressive achievements in imaging, sequencing, physiology, and computation, talented and trained minds that can nimbly and bravely navigate these new territories are urgently needed. Without an appropriately trained workforce in information neuroscience, tremendous potential will be wasted and our understanding of the brain unnecessarily limited.

2. Solutions

Participants at the iNeuro Conference articulated long lists of disciplinary knowledge, experimental skills, and quantitative proficiencies needed in neuroinformatics research. These lists far exceed what one talented individual could reasonably achieve in graduate and postdoctoral training. Consequently, most neuroinformatics questions will require teams of scientists with distinct but overlapping skill sets. To contribute to teams using large-scale data to understand the brain, individual scientists need to develop quantitative literacy and expertise in at least one area as well as sufficient familiarity in one or more related areas. Despite iNeuro's ambitious goal of articulating curricular frameworks to train information neuroscientists, no single curricular formula emerged from the Conference to train a multi-dimensional scientific workforce. Instead iNeuro conversations focused on active engagement strategies within a wide variety of new and existing curricula, emphasizing hands-on, team-based experiences using big data in training programs. iNeuro conversations also focused on common skills, traits, and collaborative qualities needed by all team members to make advances in neuroinformatics. This emphasis on quantitative literacy and developing skills through team-based learning experiences echoed recommendations in *Vision and Change* [visionandchange.org], which aims to convert undergraduate biology away from timeworn curricula of canned lab exercises and list of facts to be memorized toward dynamic curricula where doing biology research is an integral part of biology training with an intentional focus on teaching foundational knowledge and building quantitative skills that will help students continue to learn and innovate long after their formal education ends. Much like *Vision and Change*, iNeuro Conference participants advocated that universities develop curricula that train information neuroscientists by direct immersion in real scientific data and

research experiences rather than by transferring content knowledge outside the context by which that knowledge was acquired. Ultimately, when any training program articulates a specific course of study in neuroinformatics within its institutional framework, those information neuroscience programs must propose an educational pathway that develops scientists with solid quantitative and experimental skills and understandings who are creative, rigorous, collaborative, and resilient.

As an example of the difficulty in specifying a discrete curriculum, participants discussed basic software expectations for coding and statistical analysis. It was clear that quantitative proficiency beyond basic spreadsheet programs was a baseline expectation. Several contemporary software programs were suggested as gateways to programming, yet there was no one program or scripting language that emerged as necessary. Instead it was the experiences of coding, quantitative literacy, and hands-on experiences with large-scale data sets that were viewed as necessary. Ultimately, the ability to work nimbly with quantitative information, experience with programming logic and languages, experience working in transdisciplinary teams, and competence in statistics were articulated as expected competencies for information neuroscientists (as well highly desirable outcomes for scientists who use "small" data). It is unlikely that these elements can be conveyed appropriately in a single course, but instead should be intentionally interwoven throughout a curriculum in neuroinformatics in ways that take advantage of each institution's unique sets of strengths and resources.

While there was no question that existing undergraduate or graduate neuroscience program would be enhanced by emphasis on quantitative skills that will allow brain scientists to use big data, there were no suggestions for what aspects of current training could or should be eliminated to make room for additional training in large-scale data approaches. Similarly, there is little room in computational degrees such as data science or computer science to infuse experimental life sciences experiences. Recognizing that there will be continued needs for focused scientific training and maximum capacity on what any undergraduate or graduate degree program can reasonably accommodate, the potential emerged that a new type of scientific data specialists be developed. These specialists would serve crucial roles on research teams to facilitate appropriate data standards are developed and maintained for proper data acquisition, analysis, and integration and could presumably be developed in exiting and future neuroinformatics training programs.

H. iNeuro Action Plan

1. Short Term

- Propose a symposium on neuroinformatics training at an upcoming Society for Neuroscience (SfN) meeting

- Share this iNeuro Conference report with SfN (cNDP), NSF, White House (OSTP), and iNeuro participants, and any other interested parties

2. Intermediate Term

- Articulate specific graduate curricular models for certificate programs, neuroinformatics tracks within existing graduate programs, and de novo degree programs in neuroinformatics

- Encourage the scientific community to articulate standards and best practices in data sharing

- Host information neuroscience workshops and bootcamps for scientists (modeled after INCF and/or SfN short courses and MBL/CSH summer courses)

3. Long Term

- Develop incentives and infrastructure that allow scientists to follow best practices in sharing data sets

- Pilot and assess a variety of graduate program models in neuroinformatics

I. Acknowledgements

This report summarizes a 1.5-day conference held in November 2014 in Arlington, VA. The iNeuro Conference was generously supported by funds from the National Science Foundation. It brought together a group of approximately 35 professionals from neuroscience, education, bioinformatics, engineering, mathematics/statistics, computer science, data science, and library/information science to discuss training strategies for developing a workforce that will catalyze transdisciplinary advances in understanding the brain via approaches that make use of large-scale data. Thanks are extended to each participant who attended and engaged in the Conference, to Lisa McCauley for surveying Conference attendees. Special thanks also go to iNeuro organizer Bill Grisham (UCLA) for his leadership and vision that brought together a diverse team and stimulated a unique and focused conversation on an important topic.

J. APPENDIX of iNEURO CONFERENCE PARTICIPANTS

Katherine Akers (Wayne State University)
Brian Athey (University of Michigan)
Diane Baxter (San Diego Supercomputer Center)
Reed Beaman (National Science Foundation
Lukas Buehler (Southwestern College)
Melissa Cragin (University of Illinois)
Chiquito Crasto (University of Alabama at Birmingham)
Chinh Dang (Allen Institute for Brain Science)
Heather Dean (National Science Foundation)
Ying Ding (University of Indiana)
Pauline Fujita (University of California, Santa Cruz)
Daniel Gardner (Cornell University)
William Grisham (University of California, Los Angeles)
Amy Hodge (Stanford University)
Lisa Johnston (University of Minnesota)
Mark Kramer (Boston University)
Aric LaBarr (North Carolina State University)
Linda Lanyon (International Neuroinformatics Coordinating Facility)
Mahria Lebow (University of Washington)
Mike Levine (University of California, Los Angeles)
Monica Linden (Brown University)
Barbara Lom (Davidson College)
Amitava Majumdar (University of California, San Diego)
Maryann Martone (National Center for Microscopy and Imaging Research)
Lisa McCauley (Immaculata University)
Tom Morse (Yale University)
David Patterson (University of Sydney)
Russ Poldrack (Stanford University)
Raddy Ramos (New York Institute of Technology)
Gary Reiness (Lewis & Clark College)
David Sheinberg (Brown University)
Friedrich Sommer (University of California, Berkeley)
Cathy Strasser (California Digital Library)
Chuck Sullivan (National Science Foundation)
Laura Symonds (Michigan State University)
Brian Westra (University of Oregon)
Martin Wiener (National Science Foundation)
Rob Williams (University of Tennessee)
Diane Witt (National Science Foundation)
Terry Woodin (National Science Foundation)
Lisa Zilinski (Purdue University)