NEUROSCIENCE AND DATA SHARING SYMPOSIUM REPORT

A ONE-DAY SYMPOSIUM
March 21, 2014

AAAS
1200 New York Ave
Washington, DC, 20005

Sponsored by:
AAAS & the Potomac Institute for Policy Studies

SCIENCE & TECHNOLOGY
POLICY FELLOWSHIPS
AAAS

Hosted by **The NeuroPolicy Affinity Group**

POTOMAC INSTITUTE FOR POLICY STUDIES

The sponsors and publisher gratefully acknowledge participants' contributions included in this event report. Please note that the transcript has been edited for publication.

The sponsors and publisher cannot assume responsibility for the validity of all materials or the consequences of their use; the view and opinions expressed do not necessarily reflect those of the Potomac Institute for Policy Studies, Potomac Institute Press, or AAAS.

The video recording of the event and the panelists' Powerpoint presentations are available on the Potomac Institute website at:
http://www.potomacinstitute.org/past-events/2673-neuroscience-and-data-sharing.html

Potomac Institute for Policy Studies
901 N. Stuart St, Suite 1200
Arlington, VA, 22203
www.potomacinstitute.org
Telephone: 703.525.0770; Fax: 703.525.0299
Email: webmaster@potomacinstitute.org

TABLE OF CONTENTS

NEUROSCIENCE AND DATA SHARING SYMPOSIUM REPORT

SYMPOSIUM AGENDA: MARCH 21, 2014

8:00 am – 9:00 am	**Coffee and check in** *Sponsored by: The Society for Neuroscience*
9:00 am – 9:10 am	**Welcome** *Alan Leshner, Chief Executive Officer of AAAS and Executive Publisher of the journal, "Science"*
9:10 am - 9:20 am	**Perspectives and Opening Remarks** *Michael Swetnam* *President, CEO, Potomac Institute for Policy Studies*
9:20 am – 11:30 am	**Panel I – Challenges and Opportunities in Data Sharing** *Moderator:Jerry Sheehan* *Panelists: Marcia McNutt, Yuan Liu, Nina Preuss, Paul Albert*
11:30 am – 1:00 pm	**Lunch**
1:00 pm – 2:30 pm	**Neuromorpho.org – A Working Example in Data Sharing** *Giorgio Ascoli*
2:30 pm – 2:45 pm	**Coffee Break** *Sponsored by: SfN DC Metro Area Chapter*
2:45 pm – 4:45 pm	**Panel II: Building the Road Forward** *Moderator: Jennifer Buss* *Panelists: Rita Colwell, Michael Huerta, Justin Sanchez, Kristin Branson*
4:45 pm – 5:00 pm	**Closing Comments** *Philip Rubin, Principal Assistant Director for Science in the Office of Science and Technology Policy*
5:00 pm – 6:30 pm	**Networking**

EXECUTIVE SUMMARY

Neuroscience is a rapidly-advancing, interdisciplinary field which has seen growth in both research and public interest since the announcement of the BRAIN Initiative. With this growth comes an increased opportunity for collaboration and sharing of data. Technological advances make sharing data easier and less expensive than ever before. Despite the many advantages to data sharing, there remain many hurdles to overcome. These include insufficient resources, non-standardized data and procedures, and a lack of incentives to share data. In order to create a culture of sharing, these and other concerns need to be addressed. On March 21, 2014, the Potomac Institute for Policy Studies, in conjunction with the American Association for the Advancement of Science (AAAS) Science and Technology Policy Fellowships NeuroPolicy Affinity Group, convened a day long symposium to allow the neuroscience community to discuss data sharing, including the obstacles facing the community as well as how cultural changes can lead to improvements in data sharing.

Members of the scientific community came together for this one day symposium to discuss data sharing in neuroscience. Attendees included those who are interested in improving the current culture in neuroscience, in which data is withheld and sharing is not encouraged. Speakers included personnel from government, industry, journals, and academia. The symposium addressed the many obstacles to data sharing that face the neuroscience community and raised policy recommendations for moving forward.

Opening remarks were given by Dr. Alan Leshner, Chief Executive Officer of AAAS and Executive Publisher of the Journal Science, and Michael Swetnam, President and CEO of the Potomac Institute for Policy Studies. Following opening remarks, the symposium was organized into two panels and a keynote address. The first panel discussed the current challenges and opportunities in neuroscience data sharing. Keynote speaker Giorgio Ascoli then shared a working example of neuroscience data sharing: his website and database, NeuroMorpho.org. The final panel discussed building the road of data sharing forward. Philip Rubin, Principal Assistant Director for Science in the White House Office of Science and Technology Policy, delivered closing remarks.

The report that follows has been prepared by the Potomac Institute for Policy Studies and the Neuropolicy Affinity Group and is intended to contain a factual summary of the events and discussions that occurred at the symposium. The views contained in the report are those of the individual symposium participants and do not necessarily represent the views of all symposium participants, the Potomac Institute for Policy Studies, AAAS S&T Policy Fellows at large, or the AAAS.

FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

THE STANDARDIZATION OF DATA SHARING

FINDINGS

- **There are no regulations for neuroscience data sharing.**

  In neuroscience, data and code from publications are not universally available to readers, and the timing of any release is not standardized. Other disciplines like meteorology, biochemistry, and astronomy have robust standardized data sharing processes. Because neuroscience is embedded in the medical industry, there is a greater incentive to avoid publishing data and ensure proprietary claims to findings. Even if this incentive is not removed, there is still a basic need for standardization of data sharing in terms of the waiting period for data release, information on methods and code, and compatibility of data elements. Journals like Science adhere to specific policies in this regard and are able to reap benefits through increased collaboration, novel interdisciplinary investigations, and ensured reproducibility of publications.

- **Unstandardized data formats pose a technological challenge.**

  Neuroscientists are not using the same standard data formats. Individual laboratories produce data sets that cannot be transmitted to others for analysis because of inconsistencies in formatting. When it comes to standardizing this data collection, today's available technology can easily meet the field's needs. Computer scientists and companies are able to develop software that transforms proprietary formats into a standardized one.

CONCLUSIONS

- **There needs to be an organizational overhaul for neuroscience data sharing.**

  Data and code that is unavailable to other scientists puts limits on scientific progress, novel investigations into neurological processes, and computational analysis tools. Scientists cannot extract information and knowledge from data that is inaccessible. There is a need for standardized data sharing practices to improve access to the most recent, innovative science research. Encouraging computer and data scientists from both academia and private companies to design a system for sharing data will solve these problems. The government can mediate these interactions and mandate standardization processes.

RECOMMENDATIONS

- Encourage the publication of data sets, either in dedicated data journals or in existing ones. Researchers can generate hypotheses from these data sets and can avoid duplicating research if data sets with negative results are published.

- Journals should have standardized requirements for making data and code available to readers and for setting a data sharing time frame.

- Require each science field to establish a set of guidelines for reproducibility and accountability. There is not necessarily a need to dictate which repository each field chooses to place their data, as long as it is standardized and open.

- Provide researchers access to databases across all neuroscience fields for cross-disciplinary secondary research and analyses.

INCENTIVIZING DATA SHARING TO TRANSFORM NEUROSCIENCE RESEARCH

FINDINGS

- **Primary investigators are not incentivized to share their data.**

  The academic environment spurs researchers to avoid sharing data in order to ensure their own success. In a research economy where funding limitations create strong competition, scientists need to feel secure in their livelihood. From grant applications to journal submissions to evaluation for tenure and promotion, researchers feel that they need to prioritize being the first to publish a finding. Those who do share their data are often unacknowledged and unrewarded by the researchers who then perform secondary analyses. These situations and environments breed distrust and discourage potentially willing open data contributors.

- **Students are not trained in data sharing and its positive effects on research.**

  Students are not always taught how data sharing can benefit the research process as well as their own career. They do not receive an interdisciplinary higher education that promotes data sharing between neuroscientists as well as between disparate fields. A neuroscience student may be afforded the opportunity to learn about computer science, statistics, or physics but these lessons alone do not necessarily translate into a science career that involves collaboration with other fields on research projects. Students also require exposure to cooperative research experiences, avenues for interacting with other students and researchers, and the technology that allows this kind of research. They do not always see that their mentors and primary investigators value data sharing and collaborative research. Again, the current academic environment focuses on teaching students how to complete an independent thesis project and measures of personal achievement. It should incorporate a greater focus on the overall growth of the field of neuroscience through data sharing.

CONCLUSIONS

- **A powerful data sharing environment changes neuroscience research for the better.**

  Lack of incentives and understanding of the benefits of data sharing inhibit the widespread adoption of the practice. A more collaborative approach to neuroscience is unavailable until researchers actually want to share their data. Incentives are needed to produce standards in data sharing procedures. A powerful data sharing system and a research culture that fosters collaboration can completely alter the face of neuroscience research. Increasing the number and scope of interdisciplinary and cross-institutional research projects can lead to more thorough data collection methodologies and higher resolution data. A parallel example is the Large Hadron Collider at CERN. This facility produces petabytes of data every year, which is then analyzed by 140 different research centers in 35 countries. The LHC's research goals are only feasible because of this immense data sharing venture. Neuroscience research can be accomplished on a similar scale. It is not an issue that data would be coming in from hundreds or thousands of laboratories instead of from a single research center. Solving the challenges in sending out data to many locations (as seen in the LHC) additionally solves any issues in collecting data inputs from the same locations. Researchers can participate in large-scale projects where individual labs are contributing data to discover meaningful findings at all levels of neuroscience.

- **The data sharing environment of the future requires current students' participation.**

  If students are taught about interdisciplinary science and methodologies, then they will have a framework for developing investigations that draw on statisticians, computer scientists, and neuroscientists alike. This knowledge, in combination with salient evidence that collaboration is a skillset valued by universities and thesis panelists, will provide students with the necessary motivation to become data sharing-focused researchers. Data itself does not reveal significant information. It is only through data analysis that we retrieve information, conclusions, and knowledge. Scientists' value in society is their ability to conduct analyses and to extract meaningful information from data sets. Focusing on proprietary aspects of data is a waste of resources and time that detracts from scientists' capability to achieve progress in neuroscience research. Teaching students to value collaboration and data sharing will enable permanent improvements in the neuroscience research community.

RECOMMENDATIONS

- Funding institutions should require data sharing plans within grant applications and take the quality of these plans into account when evaluating them. They should also follow up on the actual execution of these data sharing plans. They should ensure that original authors are acknowledged in any secondary analysis.

- Research institutions should use article-level metrics (from both primary and secondary analyses) to evaluate scientists for promotion and tenure. Transforming researcher evaluations to have more emphasis on collaboration and broader implications on a scientific field, rather than proprietary claims, will help to resize motivations for data sharing. They should also train students to be proficient in both their own field of research as well as data sharing practices and methods. Targeting contributors at all levels with short and long term training initiatives to teach data sharing and interdisciplinary approaches to science research. Existing databases can be used by students as a tool to learn about the data sharing process.

- Journals should mandate that authors provide an explicit data sharing statement. This could motivate authors to share their data because it becomes public knowledge if they do not share. Journals can additionally encourage secondary data usage by requiring the inclusion of all information needed to reproduce and reuse primary data in publications.

TECHNOLOGY ENABLES NEUROSCIENTISTS TO PARTICIPATE IN DATA SHARING

FINDINGS

- **Automation and computational power are lacking in data sharing.**

  Large data sets are difficult to collect, curate, and analyze. Many researchers who are adding to large data sets are untrained in their successful management. They are unaware of the technologies that are available to them and are often unable to understand how to use these technologies. They are already spending an inordinate amount of time programming data collection and analysis tools. Scientific fields and research institutions are not sharing resources and knowledge with each other to improve data sharing.

- **The right technology for data sharing can be made available to neuroscientists.**

  The Internet, information technologies, cloud computing, and advances in automation are all completely capable of achieving neuroscience's data sharing goals. Cloud computing provides immediate, convenient access to large, federated data sets and makes sharing a simple process. Automation and raw computational power remove the need for human resources spent on transforming data elements into standard forms, synthesizing datasets, and curating data repositories.

- **Data sharing has multiple privacy complications.**

  In medical and clinical research, it is necessary to protect sensitive participant data sets. Participants should have access to their own raw data sets and the neuroscientists who can successfully analyze the raw data should have access as well, but it is important to limit access beyond that. Again, information technology and computer science provide ample resources for accomplishing this goal and neuroscientists should not hesitate to set up data sharing platforms for medical and clinical research.

CONCLUSIONS

- **Data sharing systems need to be created and implemented.**

  Modeling and simulating the brain will improve our understanding of neuroscience, but this task requires a well-implemented large data repository and the proper tools to analyze the data. There is a need to expose neuroscientists to the best available data sharing and collection technologies. Improving the computing and informatics technologies that neuroscientists use would vastly increase the efficiency of neuroscience research. These technologies can become available to neuroscientists but they will not be used without greatly improved guidance and translation. There is a definite need to involve computer scientists and private companies who work in data and information technologies to guide neuroscientists through successful implementation of a data sharing environment.

- **Enabling neuroscience data sharing changes the role of neuroscientists.**

  Once this data sharing system is in place, neuroscientists will be able to alter the tasks that they complete and the resources that they implement. Instead of spending their time collecting and managing data, neuroscientists can focus on analysis of their own data, other labs' data, and large scale data sets. They could begin developing novel imaging tools, cellular and circuital models, and other research modalities. The research economy would be less focused on insular, individual data sets and more focused on scientists' abilities to synthesize data and information and extract the sort of findings that revolutionize the field. Extracting meaningful information from large data sets is difficult, but scientists are the best suited to focusing on this task. Physicists do not fight with each other over the ownership of black hole imaging data or supercollider datasets; they focus on achieving monumental steps forward in their field.

  Through data sharing and automation, neuroscientists will be able to reconstruct the brain from multiple temporal and spatial perspectives and elucidate complex networks and systems. Neuroscience can quickly become a field with a diverse set of models from the genetic and molecular levels up to the cognitive and neural network levels. Data sharing is a potent tool for changing the way that research works.

RECOMMENDATIONS

- Researchers and database curators should set up programs to crowdsource analysis of large data sets. This could help to scale down current barriers to large, data-intensive research projects.
- Use challenges and competitions to fuel cost-effective solutions for research and analysis problems.
- Researchers and database developers should draw lessons from existing databases in other fields and share methods of database creations and optimizations.
- The research community and software developers need to collaborate on platforms that allow for collection and analysis of data in a reproducible and standardized format. Developers should maximize the usability of their databases by making it easier to search for data and to transfer data.
- Involve data industry companies and businesses in the development and maintaining of research databases. This will allow for increased computational power, shared knowledge and wisdom, and ease of curation.

EVENT TRANSCRIPT

The first panel brought together researchers with backgrounds in publishing, federal research funding, industry, and statistics to discuss what they saw as obstacles to data sharing in neuroscience. Data availability is essential for ensuring reproducibility and successful science. Decisions on how to publish data sets should focus on the repository location, the timing of the data release, and the privacy concerns inherent to biological data and user information. Data sets will often contain more useful information than one lab can analyze by itself. By encouraging researchers to perform secondary analyses on data sets, they will generate hypotheses, results, collaborations, and publications. Getting to this point requires a plan to overcome technological, financial, and cultural challenges. Platforms for sharing data need to include common data elements, standardized regulations, and restructured incentives for researchers.

NITRC is an example of a successful data sharing endeavor. It is a neuroimaging resource repository, image repository, and a computational environment. It provides avenues for sharing data sets, a vibrant community to discuss technical challenges, and the potential to be used as a teaching tool for good data sharing practices. Data repositories can improve their chances of success if they incorporate public-private partnerships, market their product successfully, highlight the potential for secondary analysis to bolster researchers' careers, and encourage open dialogue and communication between users.

It is increasingly common for neuroscience research to include complex statistical analysis. Data, analytic code, and methodological planning should all be shared to improve the reproducibility of a study. When neuroscience data is longitudinal and multi-dimensional, it becomes more important to ensure that it is simple for researchers to access the data itself as well as the robust methods that can handle the data properly.

By providing specific examples of current data sharing policies, the speakers were able to establish the framework by which different actors contribute to a successful data sharing environment. The process for making data sharing an engrained part of neuroscience research must involve researchers with fluency in multiple fields, user-friendly tools, and changes to researcher evaluation for tenure and promotion.

HEATHER DEAN

Introductions

Good morning and welcome to the 2014 Data Sharing in Neuroscience Symposium. We are very excited to bring this to you. I am a AAAS Science and Technology Policy Fellow and the founder and co-leader of the Neuroscience Policy Affinity Group. I am also the chair of the policy and advocacy committee in the Society for Neuroscience DC Metro Area Chapter. I wanted to note that this symposium is one of a series. As you came in you should have picked up some of the fliers that list the other seminars hosted by the AAAS and the PIPS. We hope that you will join us for many of these. We will host another symposium on May 14th dealing with neuroscience and education. I also encourage you to pick up a flier listing the NeuroPolicy group events. We have an evening speaker series that is held at AAAS and all are welcome to join. Just to give you a quick overview of the day, we will start out with a panel discussion highlighting challenges and opportunities of data sharing in neuroscience. At 1PM, our keynote talk will be given by Giorgio Ascoli. This will be followed by our second panel discussion: building the road forward. How do we overcome the challenges discussed in the first panel? I would like to thank all our sponsors: AAAS, the Big Data and NeuroPolicy Affinity Groups, the Society for Neuroscience, and our series partner the Potomac Institute for Policy Studies. I now want to introduce Jen Yttri of the Big Data Affinity Group.

JEN YTTRI

My name is Jen Yttri and I am a first year science and technology fellow at AAAS. I am representing the Big Data Affinity Group. The Big Data Affinity Group was created within the last year in response to growing interest in big data especially as it relates to ethics, security, health, medical sciences, energy. We are open to fellows, alumni, and anyone with an interest in data mining. We have a flyer out in the lobby with more information and a list of upcoming events. If you are interested in further updates we do have a Google group at aaasbigdata@googlegroups.com.

ALAN LESHNER

Welcome

It is a pleasure to be here, greetings and good morning. I am a neuroscientist myself, so I naturally love the NeuroPolicy Affinity Group and the Big Data Affinity Group. I am very intrigued by this symposium today. When I was the director of the NIMH, we embarked on something called the Human Brain Project (which has no relation to the European initiative). It was an attempt to centralize neuroscience data and ensure that all data about the brain would be collected. It had a tremendous influence because it started a large number of grants that in fact started the process of collecting data. There were technological challenges to sharing the data, but the data was still being stored. Now, we are living in a time where we can see the formation of an intersection between neuroscience and big data. These topics are receiving serious support from policy makers all over the world. In fact, Representative Chaka Fattah called me the other day to remind me that he is brokering a coordination of initiatives between Europe, the US, and Israel. All have made large investments in neuroscience, particularly in neuroscience data.

This morning I comprised a list of words that denote just how complex this issue is. Data sharing can sound like a straightforward, easy task. You can hold that any data should be made available in a shared way once it is published. When you start thinking about some of the complexities to generating and sharing new data, storing the data and moving the data between repositories, you begin to realize that it is not so straightforward. All of the issues of proprietary rights and researchers' phenomenal desire to ensure that they own their data exclusively also contribute to the thorniness of implementing a solution. Neuroscience is particularly hard for data sharing because of the potentially large amounts of data that can be generated, collected, and curated.

I'm really delighted that you all are going to solve the complexity of these issues today and figure out what we can do. I would make one request: have an interesting and provocative conversation that also provides insight into the policy framework for data sharing. This really is a moment in neuroscience history. There is tremendous enthusiasm from policy makers of all sorts and we actually have some money dedicated to neuroscience research that implements collection of sharable data. Having an interesting and provocative conversation would be good, but what would be great would be if you could try to articulate some of the major recommendations you have to make this opportunity one that we can seize. With that, congratulations to the affinity groups and our partner, the Potomac Institute for Policy Studies, for developing this symposium. Thank you.

MIKE SWETNAM

Opening Remarks

Thank you all, and thank you to Alan and everyone at the AAAS for giving us the honor of partnering with you in this series and the many things we have done together. I echo Alan's comments about the importance of not only data sharing, but also the importance of this time and opportunity. I hope we will not only find ways to discuss good ideas and good thoughts, but also make recommendations on how our society can move forward.

The currency behind science and technology is information and knowledge. Research is all about accumulating data and transforming that data into knowledge. Slowly but surely, the nations and industries of the world are coming to realize that knowledge creates and enables powerful tools. Information and knowledge move mountains, empower societies, and improve our lives. As people have come to see the value in information, their problematic tendency has been to seek sole ownership of information and milk as much value out of it as possible. Doing so is illogical, as information, particularly science and technology information and knowledge, is only valuable when it is used. If you lock that information away in a safe and never share it with anyone, it has no value and it does nothing for you.

A little over a decade ago at PIPS, we began to analyze neurotechnology research into understanding brain function. We learned about the projects and data collection that were being performed at research centers, which were often funded by private donations. The data formats were unique and the data was hard to share, but the main concern was that these privately-funded research centers were incentivized to keep data to themselves and exploit all of the available information in a dataset before it was shared across the country. That slowed down the progress of one of the potentially fastest-moving fields of research today: neurotechnology. We have to find a way to keep money, influence, and resources from delaying the movement of data, because once again, data that is locked up does us no good. Sharing it is the key to the advancement of science and is a basic doctrine on which we should base our policy. I encourage you today to discuss how can we actually articulate to law makers and policy makers laws, procedures, and statements that make it clear and easy for people to share data in a way that keeps the technology and the science moving as fast as possible. I believe that sharing data and knowledge to further our understanding of the world around us is one of the highest callings for all scientists. We have to find a way to share our data. Once again, thank you for coming and for your support. If there is anything we can do to support you in this effort, please let us know.

PANEL 1: CHALLENGES AND OPPORTUNITIES IN DATA SHARING

JERRY SHEEHAN

Moderator

I'd like to welcome you to the symposium today and to panel one, which is going to look at neuroscience data sharing–opportunities and challenges. My name is Jerry Sheehan, and I am the assistant director for policy development at the National Library of Medicine (NLM), which is part of the NIH. I'm delighted to have the opportunity to introduce and moderate this panel.

You've heard a lot already about the motivations for this panel and for today's event. From my perspective at the NIH and NLM, I believe there are a number of opportunities and challenges related to data sharing in the neurosciences. On the opportunities side, we've heard about – and we'll hear more throughout the day about – some of the great advances that have been made in the technologies that can allow us to look inside the human brain and understand neurological function. These are going to generate large volumes of data. We see increasing effort and opportunity related to the BRAIN initiative that will help us unlock even more data that we can convert into knowledge that will help us make progress in understanding neuroscience and neurological diseases and lead us towards treatments and cures. In terms of challenges, we've heard some of them mentioned already. They include the sheer volume of data that could be created by neuroscience research and the difficulties in managing, organizing, and providing access to that data. At the same time, we need to keep in mind the challenges created by the heterogeneity of the data that can be generated in the neurosciences. We talk a lot about the large volume of data sets from neuroimaging and fMRI and other technologies, but across the spectrum of neuroscience research, there are other types of studies like electrophysiological studies, animal model and clinical studies trying to understand neurological disease, observational studies trying to understand the progression of those diseases, and genome-phenome studies. Part of the challenge of data sharing is trying to make sense of that data, not just as independent data sets, but combining data across disciplinary or sub-disciplinary boundaries to advance our understanding of neuroscience. I'd say from my perspective at NIH that neuroscience is among the leaders in the biosciences in making progress in data sharing. The genomics community gets all the attention in this area, but from what I've seen at the NIH, there are great advances in neuroscience, bringing together and developing repositories of data to facilitate data sharing, including the National Database for Autism Research and FITBIR, the database for traumatic brain injury research. We are going to hear the panelists speak about tools like NITRC, which make tools and resources for manipulating neuroscience data more accessible. We will hear about other tools like the Neuroscience Information Framework that tries to pull together and provide access to a lot of different neuroscience data and standardization efforts as well. We will hear about standards like NIfTI (Neuroimaging Functional Tool Kit) that make the collection of fMRI data more consistent. Additionally, there are efforts at NIH now within NINDS to advance clinical studies and make that data more accessible through the use of common data elements.

I would now like to move on to our panel. I will give a brief introduction of the panelists, and each of the panelists will have fifteen minutes to make their presentations. Then, we are going to open the floor to a discussion. I will start the discussion with a few questions and then open it up to the audience. The first panelist is Marsha McNutt, a geophysicist who is now the editor in chief of Science, which is a journal I expect most of you have read. Prior to coming to Science, Dr. McNutt was Director of the US Geological Survey. Dr. McNutt is not a neuroscientist but is going to try to advise the neuroscience community today with her remarks. The second speaker will be Dr. Yuan Liu who is chief of the office of international activities at the NIH in the NINDS. She is involved in a range of trans-NIH and international activities related to neuroscience and is a card-carrying neuroscientist. Nina Preuss, who is the senior IT and scientific program manager for Turner Consulting Group, will follow her. She is another non-neuroscientist on the panel but has been involved from a computing and a project management point of view in the development of tools like NITRC for making scientific data more accessible. Finally, we will hear from Dr. Paul Albert, another NIH colleague of mine who is from the NICHD. He is a biostatistician who works in the intramural research program. I expect we'll see very different views on data sharing across the panel. We are going to take Dr. Leshner's advice to heart and try to both have a very good discussion during our time and see if we can identify some areas in which we can make progress toward improving data sharing in the field. With that, I will give the floor to Dr. Marcia McNutt.

MARCIA MCNUTT

Speaker

Good morning everyone. Rather than try to channel my inner neuroscientist, which I probably wouldn't be able to do effectively, I will try to set the tone this morning by talking a little more broadly about the challenge of data sharing from the standpoint of Science as a journal and the initiatives we have. I hope that will stimulate your thinking as to what some other communities are doing and give you an idea of where you might go and what communities you might look to for inspiration. First let me describe to you Science's own policy for data and materials. Our policy states that "all data necessary to understand assess and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must be available to the reader of Science. After publication, all reasonable requests for data and materials must be fulfilled." The reason we have this policy is that we want all papers published in Science to be reproducible. Reproducibility of research published in Science is a hot topic these days. Not just research published in Science magazine but in all published research.

The availability of data is a cornerstone to reproducibility. There are two standards of reproducibility. One is to take authors data, reprocess that data, and get the same result. The other is can you start the study from square one to the end by collecting your own data and get the same answer.

That is the gold standard and of course one would like to be able to see that happen. Simply being able to get the same answer from the same data is the first step and having access to that data is essential to be able to reproduce the answers that the author got. Science's policy doesn't require that Science hold the data.

Yesterday I was at a meeting at the National Academy of Sciences' bi-annual journal summit where a number of editors of journals from across the nation and some international editors attended. The topic about who should actually hold the data was discussed. One editor was quite insistent that journals should create their own repositories. Her journal had actually gone to great lengths to create their own repository of data and required that anyone who published in that journal deposit data that was used in that paper in that journal's own repository. Science does not ask that. It says that all data necessary to assess and extend the conclusions of the manuscript must be available. It doesn't say where it is available. Science's policy, for reasons that are both practical and philosophical, is that we urge authors to use whatever data repositories are standard for that community. If there is a public repository for your community we urge you to use that repository. We urge authors to do that rather than Science create its own repository and have authors populate our repository. We don't want to impoverish the public repositories by having authors send us their data rather than sending it to the public repository and leaving holes in the public repository that many federal agencies and private groups have put a lot of funding into trying to make more complete and standardized. Science is a relatively small journal when you look at the number of papers it publishes from any single field and for us to duplicate what people have done across many fields doesn't make much sense. There are also some very good practical reasons why we wouldn't do that, and I will give you one example. Recently we published a paper that had authors from Google, the University of Maryland, and the USGS that reprocessed 40 years of LANDSAT data from the LANDSAT archive to look at loss of forest across the entire globe to identify areas that were winning and losing. The bottom line was that all the gains in forest that had come in Brazil from the government being vigilant in stopping deforestation were being offset by Indonesia being very lax in its forest policy. For us to have insisted that the entire LANDSAT archive be deposited with Science magazine in order for that paper to be published would have been ludicrous because the entire LANDSAT archive is already deposited and backed up many times over. In fact, most of the electric bill for the state of South Dakota is devoted to the LANDSAT archives. There are many cases where for the underlying information there is no practical archive, including tissue samples and biological samples. There are lots of cases where proper archives have not been established and we require that the PI's hold those data. When other PI's come to us and say they have approached the authors we do put pressure on the original authors to be sure that they do make those samples available, and there is very good compliance.

The timing of the hold on the data is another question. We know there are some communities where they actually say, "If you have collected data and you publish a paper, the data should be made available within months." One good example of it is Dryad, a data repository used by ecologists. In Dryad, they have an official embargo policy that allows them to hold their data for a year. They collect their data, deposit it in Dryad then the PI can put a one-year embargo on it. Even if someone published a paper using it, the PI gets exclusive use on it for a year. If they were to try to publish that paper in Science, we say "no". You have to remove the embargo. We do not allow them to retain the hold on the data. There are Dryad-affiliated journals that allow the PI to continue their hold on the data but Science does not allow that. We have heard recently that there are some initiatives afloat, in particular in Africa, that have their basis (a noble aim), to give African scientists an advantage by arguing that they need more time to get the full scientific value of genomic and other scientific data that they have collected. They are now arguing that when they put their data in gene banks and other data repositories like that, that they want the African scientists to have proprietary hold over their data, even if they have published it, for a year when the norms in the community are 3-6 months. We have not yet had a case come to Science where an African scientist, who was publishing a paper in Science, wanted to extend a hold on data that they have put in a data bank. That is an issue we will have to address when we come to it.

Privacy is another issue that we frequently encounter at Science magazine with the free and open release of data. It is frequently encountered with medical data, but not exclusively with medical data. We had a case where we published a paper by Wang, et. al. in 2009 that looked at the way that mobile viruses can spread on through cell phones. We had billing records that had been provided to us and we found a way for the authors to be able to share those records with others who wanted to replicate or extend that study as long as they agreed to observe the same privacy, technology security, and legal limitations about the rights of those records. The records could not be published in some freely available site, but the records could be released on request as long as the requester agreed to the same restrictions that the original authors had. The data can be shared, but it has to be shared with restrictions that the new user has to agree to and that are the same restrictions that the original authors had.

Finally, let me leave you with a challenge. My challenge to you is that *Science* magazine is embarking on a reproducibility initiative and so far we have implemented reproducibility guidelines taken from the NINDS report that came out of a workshop that set four main guidelines for best practices in preclinical studies for transparency and reproducibility. Those four guidelines for transparency and reproducibility are that studies should have a pre-study plan that looks at things like how will outliers be treated, and how is the study going to end. In other words don't just decide on the fly how you're going to treat your data. Do a power analysis so that you have a good idea about how many samples you need. Do you need 50 mice or 500 mice in order to get a good signal to noise? Make sure the samples are randomly assigned to your treatment group and your control group. Determine whether the experimenter is blind to which are controls and which are treatment groups. These are questions that we are asking for our authors now for pre-clinical studies.

I was at a meeting yesterday at the NIH with Story Landis, who was the lead author on that report, and she was telling us that she was reading a paper in Nature that they were very interested in at the NIH and that they were thinking about putting out a press release. Nature has also adopted these guidelines. They were thinking about making a big deal out of this paper until they read in the paper that the four guidelines had not been met. The experimenter had not been blind and that one of the other guidelines had not been met. They decided they weren't going to put out a press release about it because they were not confident in the conclusions of the paper. Whether or not someone follows these guidelines can help readers and reviewers with their confidence in conclusions of the paper. We believe that these guidelines are important to pre-clinical studies but not widely applicable to other fields. Science has now obtained funding from the Arnold foundation to have three additional workshops in other areas of study to produce NINDS-type reports to help guide other fields in best practices. We have already decided that the second workshop is going to be in the social and behavioral sciences because we believe that community is poised to go forward with defining their best practices because we see lots of action in that community to come together, define standards and best practices, and work toward transparency. The other two workshops have not been defined. What I would like to see coming out of this meeting today is strong movement forward on data transparency that could help the neurosciences be a good candidate for one of the remaining two workshops. So that is my challenge to you. Thank you very much.

YUAN LIU

Speaker

I am so glad that Dr. Marcia McNutt mentioned the work of science rigor and reproducibility that the National Institute of Neurological Disorders and Stroke is doing. Today, I would like to share with you some of my thoughts about data sharing, including the challenges and opportunities associated with it, as well as possible solutions. We heard from Dr. Leshner and Mr. Swetnam about why we need to share. In short, it is not good to lock your data in the closet. A second question is what to share. This is a big question because if you only provide raw data, then no one is going to be able to use it. The third is how to share. I think a policy perspective is especially important for this question. The fourth is when to share; whether it is the day after the publication or a one-year hold as in the case with the African scientists. Fifth is with whom to share. Should the data have controlled access or be freely shared on a public database?

There are many challenges in data sharing. The first barrier is sociological. Some investigators may say: "If I spent many years collecting data on monkeys that are very difficult to obtain, why should I share it with others? A person without knowledge of how I collected the data may misinterpret, misuse, or abuse it." There are also technical challenges. I was talking with Jerry Sheehan the other day and he used the term "enabling", which means not only providing technologies, but also enabling an easier way to share data. Finally, there are practical and financial challenges because sharing data costs money and effort. I am going to use a few examples on how we can address issues in data sharing. To address reluctance, we can use a metaphorical "stick" or a "carrot", or both. Some of the existing "sticks" are quite weak in my opinion.

Currently, principal investigators with NIH grants over $500,000 in any year are required to develop a data sharing plan. The plan is reviewed by the study section but is not counted when calculating scores. The progress is monitored by program directors, some of them are very rigorous, while some others have hundreds of grants and may not have the time to address data sharing issues. The good news is that we now have a better data sharing policy, particularly with the Genomic Data Sharing Policy. The spirit of this policy is that all NIH-funded research that generates large-scale human and non-human genomic data should be made available through any widely used data repositories. The NIH cannot afford to make a centralized repository because it is impossible to share every piece of data that is produced from an NIH funded grant. This policy will apply to all funding mechanisms and we removed the $500,000 threshold.

Why do we share and reuse data? Many data sets contain far more information than any single lab has the time to analyze. It would take not only your life but your postdocs' and students' lives to analyze every single piece of the data you collected. Another point is that a single data set may answer many more questions than the initial question. Reusing data is cost-effective. I understand that Heather Dean is a monkey physiologist, so she knows how expensive and time consuming it is to collect monkey data. A couple of days ago Jerry Sheehan organized a trans-NIH forum on policy. We learned a case from the National Libraries of Medicine where the intramural group analyzed Kaiser's emergency room data, and they learned about the relationship between survival rate and obesity. Their investigation of re-used data prompted many good results, thus highlighting advantages to sharing and reusing data.

I would like to share some other successful cases with you. A few years ago, Dr. Ascoli and I sat down together and discussed how we could provide some "chocolate" or "carrots" to encourage data sharing. We organized a satellite meeting at the Society for Neuroscience meeting where we showcased five success stories on how good data sharing practices can add value to your own research. It can generate new hypotheses, new results, new publications and new collaborations. The details of that symposium are available online (http://www.nitrc.org/forum/forum.php?forum_id=225). We talked about reanalyzing existing data or reanalyzing gene expression data from multiple papers to reach new conclusions where the original author was wrong. Another case is fMRI data for anonymous users outside neuroscience to use. There is an economist using this data to publish papers, and the original author is not upset by it. We can draw an evolutionary tree about how many papers and publications grow out of this shared data. We also want to emphasize on how to share data between experimentalists and computational modelers because they don't generate the data themselves.

Today, you are going to hear examples from Giorgio Ascoli himself. We talk about policy and the needs for policy. I think there are still some gaps and needs in respect to policy, one being how we develop policy in regards to acknowledgements. I think journals, funding agencies, stake holders could all be involved. One example is the Database of Genotypes and Phenotypes (dbGAP) at the NIH. We require that approved users agree to acknowledge contributing investigators. However, we did a study analyzing dbGaP secondary users' publications and the results were not that good. Not everyone complied with this policy, and many of them did not acknowledge the original contributors. This raises the question: once we have a policy in place, how do we enforce that policy? A couple years ago, a representative from the OSTP came to a meeting where he asked what could we do to encourage data sharing? So, I raised my hand and asked if it was possible for the White House to do something. If this message comes from the

President and the OSTP people will listen. To our surprise, they took us seriously and launched a program promoting open data. President Obama, OSTP, or the NIH can give more "chocolate" or "carrots" to the people who are willing to share.

There are still a lot of gaps to be filled. Take the way journals or PubMed acknowledge citations. I think Giorgio Ascoli suggested that we add a button to every publication that says how much and how many times data has been shared from this particular article. If the NLM can do something like that, it would be very helpful. Another concern is that you spend a lot of time analyzing your data to clean it up and put it on the web, which costs time and money that doesn't lead anywhere. Should we encourage or demand institutions include this as criteria for promotion or tenure? As funding agencies like ourselves, should we include a data sharing track record as criteria for merit awards?

I'm going to switch gears to technology challenges. I mentioned the five questions we are interested in addressing today, include what to share. We know some data is easier to share than others. For example, genetics or anatomy data is probably easier to share than physiological data. Being a physiologist myself, I know how hard it is to analyze and share multi-electrode recording time series data. Another example is animal vs. human data. Working with human data is much more complicated because of the privacy issues. If you do EEG or fMRI how do you consolidate, integrate, and make the format shareable? For meaningful sharing, the data collector needs to provide a lot of things including cleaning the data, making sure there is clear annotation about under what conditions you collected this data, and sharing the meta-data format.

This brings us to the question how to share. We need to establish a sharing platform and environment. I would like to mention a few examples: one is the International Neuroinformatics Coordinating Facility. Mike Huerta and I participated in the establishment of this international organization, and they are establishing global data sharing policies, platforms, and technologies. Another example is the Neuroscience Information Framework, which is a blueprint contract under the NIH to support a clearinghouse for all neuroscience related data. Our colleague, Nina Preuss, is going to showcase the Neuroimaging Informatics Tools and Resources Clearinghouse in more detail. For data sharing, we also need to develop common data elements. Yesterday, we heard at the NLM that there are twenty different terms for "stroke". We really need common data elements to consolidate these different types of data and make them interoperable. The key to all of these is community buy-in. It does not matter how interoperable your format is and how sharable your platform or environment are if no one is using them.

There are some practical challenges regarding when to share. Dr. McNutt already mentioned this, so I am not going to expand too much. Our draft of the Genomic Data Sharing Policy states that we would like to share all non-human data on the date of the publication. For human data, we would like to share the data six months after publication. Current policies are all over the place; there are six months, twelve months, and the day after publication. We need some more consistent policy, and all of us need to work together to determine the best data sharing policy.

Another practical challenge is with whom to share. With collaborators? Of course. With potential future collaborators or with known users? Once you deposit your data on a controlled database, you will be able to know who downloads your data because the moderator of the database will be contacted by an investigator requesting your data. He or she will contact you about the request, making you feel more comfortable about what groups are using your data. However, if you put it in a public database, there are hundreds of thousands of unknown users, and this raises the concern of misuse. Therefore, we need some data sharing policy that focuses on how we can control the traffic and acknowledge contributors who post their data to public databases.

The lack of training regarding the management of data sharing also presents an issue. Training is necessary for the data users, as well as the data contributors. What type of meaningful data should you deposit? More importantly is training the next generation of scientists to establish sharing as the norm, and grow up as scientists in this environment. The good news is that we recently released training opportunities called Big Data to Knowledge (BD2K). If you are interested, there is more information available on our website (http://bd2k.nih.gov/funding_opportunities.html#sthash.bmSLa10a.dpbs). There is also a lack of community awareness. Funding agencies, journals and societies put out information about data sharing, new clearinghouses are established, and there are booths, posters and satellite events at meetings. We have already done a lot, but there is still a lack of community awareness. What else can we do to reach out to the community?

There are also financial challenges. Good data sharing requires appropriate time and money. We have something in place where you can budget in your grant application and apply for special supplements. For example, the NSF-NIH Collaborative Research in Computational Neuroscience Program has a data sharing website (http://crcns.org/), and also has a special data sharing proposal to which you can apply (http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5147.)

The majority of the big databases that have been developed need to be sustained and further developed, but how do we do that? Through user fees or through private-public partnership? You can hear a little bit more from Nina Preuss about this. I am going to end with a slogan, Data Sharing and Data Mining Lead to New Discoveries! Jerry Sheehan asked me what my biggest question is and I believe it is this: how can we make data sharing and data mining lead to new discoveries?

NINA PREUSS

Speaker

I would like to thank the AAAS for inviting me to talk about NITRC as a case study. I have structured my talk after Dr. Liu's talk to address the three challenges that she had mentioned. Along the way, I will sidebar the NIH-funded Blueprint for Neuroscience Research program in addition to NITRC that are playing in this "sandbox". The reason I want to showcase NITRC is because the NIH Blueprint funded it to address the sociological, technological, and practical financial challenges that Dr. Liu had addressed. I am the program manager for healthcare at TCG. We are a local small business focused on making the magic happen for government clientele between scientists and technologists for cutting edge solutions whether the domain is neuroscience, bioinformatics, or even infectious disease. So what is NITRC? NITRC is a combination of services. We started out as a NITRC Resource Repository (NITRC-R) and have grown over the past seven years to include not just software but test data sets, workflows, and other things that are openly shared by the research community. We have 650 different software packages that are shared on NITRC-R. Most of them are open source but we try to be agnostic so that someone new to the community can come in and try all of the commercial products that are out there. You can search by licenses and say that you only want software results, or open source programs, or all data in a certain domain.

The second service that we rolled out was the Image Repository. Dr. McNutt was talking about a public repository and that there are certain ones that people should send their data to. The NITRC Image Repository (NITRC-IR), is focused on neuroimaging related data. Jerry was talking about NIFTY as a sexy term for data format. NITRC started out with NIFTY format and we've also started to add in some DICON data as well. We have over 20,000 images of NIH funded studies of children with autism, ADHD, schizophrenia, depression, and anxiety. If you want to access some of the data, you have to actually register on the NITRC resource project that is associated with it because they do want to have access and reach back to other researchers who are going to use their data for secondary analysis. Other data sets on NITRC, like Candi Share, are just shared openly. It is available for download and the contributors are just hoping people will cite it properly. On NITRC, you can tell people how you want your data to be cited and hope that secondary users will follow.

The third service that we just rolled out is the NITRC Computational Environment (NITRC-CE). It addresses the very real challenge of computing against big data sets that are typical of neuroimaging. How did we figure out what tools to put on the NITRC Computational Environment? We looked at the NITRC Resources Repository and we looked for the most popular software tools. For the tools that were most popular by download numbers, we then reached out to the neuroscience community and asked how they combined their usage of tools and organized data. Once we learned that information, we started pre-loading computational environments with a series of very popular data sets and software tools for people to compute against. The first thing I did when I was invited to talk was to go on to the NITRC mini-forum, which is viewable by the public (we've got 3,700 subscribers now as well), and I just typed in the search strings AAAS and data. I pulled up results that touched on many of the challenges that were discussed so far. There are technical challenges in making raw data and processing consistent. People were chatting about how to document algorithms and workflows to enable reproducibility. The search string also returned information about financial challenges such as deciding who bears the cost of storing the data, how long data should be stored, and who pays for the manual work of preparing the data for distribution.

We have accomplished the initial mandate, to create a repository and clearinghouse with a vibrant online community. In the beginning, I think everyone was wondering if the community would embrace it and play nicely. We have been pleasantly surprised over the years that they were willing to do that. There are a lot of new collaborations that have cropped up as a result of NITRC, where people can just contact someone directly and start collaborating with them. Dr. Liu mentioned training as a practical challenge. Twice now, MIT faculty members have used NITRC Computational Environment to create a standard environment for a classroom of 50 students. They pre-populate the computer lab with an exact replica of dozens of open source neuroimaging software. Then, they upload either the NITRC-IR data or MIT's own data. Once there, they walk the students through the process of bringing data to the cloud and downloading that process data for their research. Those students are learning by example that they benefit from others sharing software and data. They learn that they can conduct secondary analyses on those shared resources to devise their own novel hypotheses and that does lead to good things. For example, three students won a competition using shared data that directly led to Washington University winning their $40 million Human Connectome Project grant, which many of you are familiar with. Where best to teach good data sharing behavior but the classroom? These classes can lead to the culture change that Dr. Liu mentioned by the open source community and the White House open data policies.

The best neuroscience minds are using the NITRC Computational Environment, as evidenced by its use at the Organization for Human Brain Mapping Conference last year. At the conference, there was a competition to create the best visual or data mash-ups. We hope to prove that data sharing does lead to citations (and grants) as another tactic in the community awareness that is very essential to change that culture. To do that, metrics are a key to measuring success. These graphics show data sharing pays dividends over time. More and more downloads lead to more and more citations. One success story is the Nathan Kline Institute sharing the 1,000 Functional Connectomes data via NITRC. They have used their high download numbers and citations to then go to the NIH and receive roughly $9 million in funding and three new research grants to enhance processes and share that data further with the community. The NIMH Director included that initiative in his blog which also helps reinforce the message that those who share are recognized in ways that matter.

So what are some other ways NITRC ensures that the community is aware of our services? What are the ways to influence a cultural change of sharing? Even though marketing might be a dirty word in science, we have proven that we need to get the word out. We staff a booth at three conferences a year and researchers who know of NITRC regularly swing by the booth to help administer shared resources. It is a challenge to catch the eye of an introverted scientist who is walking down the aisle communicating "don't talk to me, don't talk to me" with their body language, but once we can get them in the booth they say, "Wow, this is great! This data is really free to download? That's so great!" While they are in the booth, we figure out what their pinpoints are and try to document that information so that we can improve our services. We then follow up on suggestions that they give us to add other data sets or other software to NITRC. We're using the community to tell us what to add to the software and then we reach out to those tool developers, software developers, and data holders to say 'please share'. Finally, we ask those people in the booth to give ratings and reviews on the tools and data that they use so that others can benefit from their knowledge. Other marketing tactics include ensuring that NIH institutes carry our YouTube NITRC training videos on their channels. We ship NITRC data sheets to labs that are conducting internal training seminars, society satellite meetings and other international data sharing working groups. We print posters so other people can spread the word at conferences like the New Horizons in Human Brain Imaging Conference that took place in Hawaii. We also cross-list searches on other sites. Dr. Liu mentioned the European counterpart to NITRC, the NIH sites, and other initiatives like NIF. Technological challenges include the very real problem of searching and finding data. The Proper Data Federation has a set of agreements and standards so you can find data across sites. It allows you to use metadata to find data that is at site A and collect it from site A or store it only at site B only because it was collected at site B. It is working in many places now, including the National Database for Autism Research, where they share their data across five different sites. NITRC is fettering its data with its European counterpart, INCF Data Space. Once you have the data you want, you have the technological challenge of getting the data and the computer together in one place. Currently, neuroimaging data is somewhat big. Not as big as genomics, but somewhat big. Option 1 is to copy the data over the Internet or get a hard drive. Now you have the data near your computing environment and your research institution. That works if you need 20 out of the 20,000 images that you want. Option 2 is that you bring your computing power to your data. That is, you do not incur the time and cost of transferring the data. Moving large amounts of data can be time consuming so the Human Connectome Project created a service called Connectome-in-a-Box. They ship four terabyte hard drives to people so they can use the data instead of trying to download the sets. Microsoft is quite interested in solving the technical and financial challenges for scientists. They recently rewarded NITRC a grant to mirror image repository and merge data so researchers will have the computing power on their account. If you want to use NITRC on Amazon Now it will take you 20 minutes to compute against this data when it would have normally taken 6 hours. I started this presentation thinking it cost $4, but then I saw recently that they dropped their price and now it only costs $1.

For those who want more, we offer two clusters on the NITRC computational environment as well. Imagine taking six different software packages and running one set of data through each, one at a time, for an optimal result. That is what the NITRC Computational Environment does. You can plug your data in and chug without having to spend hours configuring a bunch of software packages that may or may not be on the same operating system. A researcher at a nearby well-respected university came to us at a conference and said since her institution didn't have a lot of computing power, she and her colleagues actually went office by office, collected the computers, and strung them together to make their own homegrown cluster. They uploaded and configured all that software on those computers, processed their data, and returned the computers once the work was complete. That is unbelievable. Researchers should not have to do that. These are researchers who get NIH grants. The NITRC Computational Environment helps labs like hers who do not have the big budget. It also helps new researchers get their work done when established PI's already have talked to the institution and reserved x amount of time to compute for their research. Here you are, waiting a month and a half to just do your small study. The new generation trained on cloud computing will say 'Forget it. I'm going to use my Amazon account or I'm going to use my Microsoft account and I'm going to compute my own data.' These pictures are samples of a few of the neuroimaging tools that are shared on the NITRC Computational Environment and on NITRC itself. Most of them are open source. In addition to the YouTube videos, we have a very detailed user guide and we update the list of software we have available on the Computational Environment regularly. There are technological practical issues associated with licensing of software tools. We created a pop-up that identifies licenses and agreements in white that don't require codes and in red where you have to put your license code in. Dr. Liu mentioned supplement grants. NIDA and NIMH issued at least $2 million in supplement grants to help with the financial burden of ushering data, such as preparing your data to be distributed openly.

The Human Connectome Project and Nathan Kline Institute received administrative supplements to address these practical and financial challenges after sharing metrics with NIH about how valuable their data had become. It is a known challenge for new investigators to break into the established world of receiving NIH grants. Secondary analysis success stories include two grad students who use the 1,000 Functional Connectomes data for their dissertations. Will one of them discover the cure for Alzheimer's? A bioengineering student who is now first author on a publication that mashed up two data sets from the NITRC Image Repository and a third set from the Human Connectome Project. Finally, in doing its share to enable the kind of reproducible research that Dr. McNutt had spoken about, NITRC makes it easy now for researchers to replicate other researchers' processing approach down to the operating system, software package, version, and orders, which is that gold standard that we are shooting for.

Some of the entities that I mentioned have both public and private funding, along with federal funding research going into NIF and NDAR have NSF funds. Alzheimer's Disease Neuroimaging Initiative has pharmaceutical funds. NKI's 1,000 Functional Connectomes project has non-profit funds. These are all examples of projects that are moving forward with public-private partnerships. Researchers using NITRC-CE can pay Amazon directly for their compute time. NITRC does not markup that time but we could be positioned at some point to mark that up if that was required. As those MIT students conduct their own research, they will be starting with their Amazon and Azure accounts and will be trained to expect their data to be processed immediately and will be more willing to pay to play. Indeed, because of the size of autism data, NITRC-CE is becoming an essential service for NDAR researchers. They will be able to build the Amazon usage fees into their grant requests and then NIH is happier knowing that the fees are directly related to specific research study. I do not know how many of you play Candy Crush, but it teaches us that 70% finish without paying while 30% are willing to pay a fee. I will not ask for a show of hands but think of how many times you've been willing to pay that $1 fee for that app. It really needs to be worth the cost, so the jury is still out on whether scientists will pay to download other researchers' data. The NITRC team felt strongly that we are still in that phase of changing the culture so we are not charging and are making sure that people do not charge for the data downloading because we do want to encourage secondary analysis and usage. The challenge will be getting past the current business model where universities apply prorated facilities and administrative costs of their ever-expanding IT infrastructures to grants. PI's are rewarded for bringing in the big grants. Thank you all for your interest in this fascinating and important subject. Indeed, there are many challenges for us to address and you all will help us address them.

PAUL ALBERT

Speaker

Thank you for the invitation to speak today. I feel a little bit like a fish out of water because I worked in neuroscience about 20 years ago. I worked as a staff scientist at NINDS but it is remarkable how much neuroscience has changed over the past decades, especially from what I've heard so far this morning. My talk is on data sharing reproducibility and analytic strategies. I am a research biostatistician and I lead a group at the National Institute of Child Health and Human Development. Our focus is primarily on developing new analytic strategies, so that is the main perspective that I come with.

So first of all, why share data? What is the purpose of doing that? I think there are three major groupings or reasons for sharing data. One is to reproduce the findings of others and this falls into the category of reproducible research, which as was pointed out earlier, is its own important topic that I will touch on briefly but could be the subject of many workshops. Reproducing the work of others is important. That is, taking a data set and then validating another manuscript's findings is valuable research. Another reason to share is individualized meta-analysis. For example, in neuroscience as well as other areas of science, researchers conduct small studies which are sometimes underpowered. Or maybe they have enough power for the objective of interest but for secondary analysis they are underpowered. We want to combine information from many studies and that is a challenge. It is a lot of work to ensure that the data sets have the same platform and that there is agreement to get the data. At my institute, there are some really good examples of that in looking at pre-term birth. We have large cohorts and we want to predict pre-term birth from genomics data. Pre-term births are relatively rare so our cohorts do not always have the power to do that, but if we put the cohorts together, then we can achieve this goal (especially when we also incorporate international studies). Another reason to share data is to perform new investigations. In a large cohort study, you may have a series of questions that are the primary objectives but there is a lot of very interesting, new scientific exploration that can be conducted with the same data.

So what are the goals of reproducible research? One is to reconstruct data, either raw data or normalized data. Another goal is to analyze the methods used in publications, which typically involves code. These should be easily accessible to the outside community. It is also important in this modern era to reconstruct complex statistical analyses in a clear and concise way. With the increased focus on "-omics" data that we see, the statistical analyses are more complicated and more subtle. Often, peer review does not catch errors or issues. I have seen instances in top journals where statistical analyses were later shown to be invalid and this made a difference in terms of the results. We need a mechanism so that they can be clearly provided for any published paper. Increasingly, journals and specifically statistical journals are including analytic code as supplementary material in papers. Leading statistical journals now require that all code is put in the supplementary material. One journal has an editor who goes through the code and checks to see that it works and actually gives the correct results, which actually takes a lot of time. This process is optional but if the paper goes through it, it gets an official stamp as a reproduced paper. This acts as sort of a feather in the author's cap and can increase the likelihood that the process will be used. We should also describe all steps of an analysis and really show the analysis plan and strategy. Increasingly, we need to document the final analysis that authors do as well as the discovery process. What was the analysis plan? Did you follow the analysis plan? Sometimes, in a perfectly reasonable study, you start with one plan but then adapt it because you learn and want to have an adaptive procedure. This needs to be described so when someone reads the results, they understand the context of the whole analysis procedure.

Using proprietary software for analysis should be avoided. In statistical analyses, it has become increasingly important to use one software package for analysis (usually R now) which is available to everybody. It does not cost anything and it is easy reproduce procedures for others' usage. If you use proprietary software, then there is a chance that someone else does not have it or cannot afford it and then this becomes a problem. R and other software packages have ways to perform reproducible research within the programs themselves. The idea is that if you are writing a paper, then you can implement the analysis within the paper so readers can check the code associated with each result. This is a way for these complex procedures to easily be reproduced by other people. R-sweave has been developed recently and is indicative of a trend of an increased reproducibility. There is also the Comprehensive R Archive Network (CRANS), where you can publish software. If a researcher has a new statistical procedure or new imaging method for data, they can put software on the CRANS system for access by anyone using R.

Some examples of data sharing and platforms that I have seen include data sharing platforms designed by people in my division. My division performs mostly large-scale, large cohort studies. These studies are almost always longitudinal and have between 2,000 and 10,000 individuals. There are some real big challenges in that data because we have so many subjects and "-omics" data as well. We put data on the web after the primary and secondary objectives have been completed, or 5 years after the data has been completed. It does not generally take 5 years because objectives are completed fairly quickly but we do have a policy for those situations. Investigators in our programs might spend 10 years on these cohorts and might want the opportunity to publish those findings first. Other examples of data sharing platforms include the database of genotypes and phenotypes (DbGAP) and then there is the NHBI BioLINCC. These platforms are nice for accessing study data with large cohorts.

There are challenges involved in sharing data images. My experience is that there are not a lot of issues involving neuroimaging but rather with examples like the NICHD fetal growth study that incorporates ultrasound data. This might be an instance where the ultrasound community is a little behind the neuroimaging community in some of the issues in data sharing and using proprietary software. The goal of the NICHD intermural study on fetal growth is designed to estimate race-specific standards for fetal growth in singleton pregnancies. Another component of the study is to analyze fetal growth for twins. There are 3000 singleton pregnancies and 150 twin pregnancies monitored throughout gestation. 2D and 3D ultrasounds are being collected at a high rate. We are still struggling with developing a strategy for sharing these images. There are tens of thousands of these images being collected and they are being brought to us in hard drives. How are we going to deal with this data flow and deal with the proprietary software that these images are stored in? The images are stored with Volusom, which is used in GE-based machines. We are still not certain of the best way to get the images from the GE machine.

Some neuroscience data-sharing issues come up in terms of analysis and complex data that are longitudinal and spatially coordinated. That is, data may be high dimensional and span in many directions (time, number of pixels, and even the differences in modalities). Many new statistical methods for analyzing experimental and population-based imaging data are being developed now. There should be simple procedures for accessing these methods.

PANEL 1 QUESTION & ANSWER

Jerry Sheehan

Thank you for that assessment of what is happening in the statistical community and the activities that statistical journals are taking in this regard. I have a few questions for the panel and then we will open up the floor for questions as well.

This may be jumping ahead a little bit in terms of taking next steps to answer this great challenge, but I'd like each of you to think about what the next big opportunity in neuroscience data sharing could be. Are there particular types of data that are ripe for study or is it something more broad-based for the general community to participate in?

Yuan Liu

I've been thinking about this question for many years. I think a lot of this depends on the readiness of the community and the community buy-in. If we are not ready to share data and use the tools, then no matter how hard you use your stick or chocolate, it's not going to work. I would like to use NIFTI as an example. NIFTI is the Neuroimaging Informatics Technology Initiative and we started with workshops to discussing sharing data in the neuroimaging field. First, we identified the needs and came up with a long laundry list of what we need to do. The first item on the list was metadata format, which is integral to sharing. We invited people who developed software and the users to identify their needs. Neuroimaging researchers use only a few software packages and we get the software developers to come into town a couple of times a year. They developed NIFTI 1 and then we hosted a large community meeting to push for researchers to buy into the program. You have to identify what the community's needs are and then provide visibility. Even hardware companies like Phillips and GE need to buy in. If you are going to use a specific software system, you have to put your data in that particular format. We made this available to the whole world so roughly 90 to 95% of neuroimaging researchers are on the same page now. These programs are also expanding with cute names like GIFTI or SIFTI for surface mapping. I think if we want to share more, then we should have more conferences to identify the need and readiness.

Nina Preuss

We always look at the money side of things and how that makes an impact. The BRAIN Initiative is being funded by $110 million in taxpayer dollars. Additionally, private foundations are adding another $100 million. Data funded from that initiative should be made available to the public in all forms (raw, processed, etc.). There should be specific timelines for release and data should be put into existing repositories as appropriate. Dr. Liu gave the example of metadata in NIFTI where neuroimaging data was previously stored in DICOM but the community realized that this was not sufficient for federating and combining data. NIFTI enabled the collection of disparate research activities and studies and now researchers can search for left-handed males with a certain genetic profile. It needs to come from the community but as taxpayers, we can demand that data gets shared within a reasonable amount of time.

Paul Albert

One important goal would be to develop ways for flexible analyses to be easily reproducible. In the statistical imaging field, there is an explosion of new methods for fMRI and functional data methods to flexibly model the longitudinal spatial correlation. The methodology is complex and the software is in hard-to-reach places. If we want to bridge innovative analyses and data collection, we should look to developing a new platform.

Marcia McNutt

My best contribution might be what I can see from the perspective of an editor of a journal that sees what is new and hot in a field or what is a wave on the horizon for the community to get out ahead of before that wave becomes a tsunami. The wave I see is optogenetics. Optogenetics is a very exciting technique with high spatial and high temporal resolution. Any time I see those words, I'm thinking lots of data and lots of information. The fact that it is probing individual neurons with the opportunity to look at the function of neural networks in a fundamentally new way is exciting. Lots of people are going to want to use this data in ways that the original person that gathered the data are maybe not using it. This is the kind of data that should be freely shared and this is the kind of information that there should be freely available data archives.

Jerry Sheehan

Is there any thought on how we would capture this kind of data and make it more available for future study?

Yuan Liu

My colleagues at NINDS and I discussed the opportunity that optogenetics presents a couple years ago. Optogenetic data is a tool; you may or may not want to establish a database of all the optogenetic research in a single repository. The future may be federated databases that are not centrally controlled. This is also important in context of the BRAIN initiative and the European brain project, where we are collecting thousands of data points. The Human Connectome Project acts as a control, or gold standard, for imaging data. Whether we should establish an optogenetic data repository or work with Society for Neuroscience to make data formats elements standard and interoperable, we can improve data sharing.

Marcia McNutt

AAAS Science magazine gave the Newcome Cleveland Prize to the nematode connectome project. When I look at the technology that was used for that, it was quite primitive compared to what we have now. I'm wondering if these optogenetic databases would be optimized for the kind of data that is coming out or if we would have to look back at these databases from the past, as if we were building them with optogenetics in mind, and structure them differently.

Yuan Liu
Another thing we can discuss is contacting NIF, which is still contracted under the Blueprint. They are supposed to be the central clearinghouse for all neuroscience data and we can ask them what the common element that we need to establish with this optogenetic research might be. We could even ask the INCF if there is a global standard to enable sharing. I agree that optogenetics is a hot area and we can get a lot out of reanalyzing and reusing existing databases.

Jen Buss
The database I am familiar with as a biochemist is NCBI but there are also databases for crystallography and genomic data. With neuroscience, there are ways through Blue Brain or the Connectome to use structured data. We will have operable databases eventually, but it is just a matter of getting there.

Jerry Sheehan
One of the issues is getting from operable databases to interoperable databases, which requires thinking about standards and representations of data. I wonder if some of the panelists want to comment on that topic of interoperability and standardization.

Nina Preuss
I wanted to differentiate between aggregating data and federating data. To find different data in different databases, or to know about a database and pull information from it, require aggregation. NIF knows what is out there and has automated scripts to find data from NITRC or NDAR. Then, when you are trying to get to the data itself, you have a federated approach. You look at the metadata, maybe the left handed male that has a diagnosis of autism. Something like NIFTI allows you to look in the different databases for where those particular images are, whether they are on NITRIC's image repository or INCF's database. To reiterate, there are two different technological approaches that differentiate how to find data across the world and then how to delve down and find the data you want.

Jerry Sheehan
From the library side, we talked about discovery and then there is the metadata that enables aggregation of datasets.

Dave Clifford
I'm Dave Clifford, here with Dr. Sanchez from DARPA. If I want to share a news article from my smartphone or browser, I can just find a button and share it through email, my blog, or Facebook. I have talked to a few neuroscientists about this and the culture of using a LIMS system (laboratory information management) isn't in place so there is not a simple protocol for publishing a dataset to a repository. Has anyone given any thought to writing a MATLAB plugin or some other software to make these datasets easily publishable?

Yuan Liu
In the past years, the NIH Neuroscience Blueprint worked together to develop tools like NITRC and NIF. These programs have several different levels of registration, where you provide different levels of technical help for the users.

Heather Dean
If I was understanding the question right, it was how do you get your workflow to become easier and just share what you want when you want?

Dave Clifford
As a user, I'm referring to barriers to contribution. We were talking to some of our performers and asked them why they don't share data now. For them, it is an afterthought, but if you were able to push a button and sent it to a repository with associated metadata all at once and quickly, that would seem effective. Discoverability and download tools are obvious needs, a motivated user will go out there and get data, but we also need upload tools.

Heather Dean
You have to make data easy to share, part of the workflow. There has to be a benefit to the researcher. My understanding of Open Science Framework, Brian Nosek's project, is that their goal is to make documentation on data collection easier. You share what you want and direct data and information to the groups you want. I think there is certainly thought to this process. Data sharing is a burden on the PI, but if you make it so that it has benefits to him or her at no extra cost, then it is easy.

Marcia McNutt
I would say that this has to do with the issue of standards. I am an oceanographer and there are several standards that are widely used. MATLAB and GMT (generic mapping tools) are used widely. When you collect data on oceanographic ships, your data is provided in several formats that are sent to the repositories. Your data is automatically downloadable to MATLAB and GMT. GMT is also compatible with the standard format that goes into the data repositories which have been around since the 1970's. Once you have "standard" standards, this happens.

Jerry Sheehan
It goes to the difference between tools and standards. Yuan, you had mentioned this in your remarks. We can use carrots and sticks to provide recognition to people for having shared data, perhaps for shared data in a way that makes it easily usable. We might have policies that require people to do data sharing. We look at funding agencies as the people that carry sticks because you want to come in to your next funding increment with evidence of compliance with data sharing policies. This notion of the enablers is that if you can make it easier for researchers to collect data in standardized formats and have the data annotated in a way that makes it easily and readily shared, my sense is that you don't need as big of a stick. Even if the rewards aren't quite as great or aren't quite there yet, you've lowered the burden in a way that makes it a more natural part of the process. I've seen NITRC as one mechanism of doing that, but I'd be interested in thoughts about other tools. I am often envious of the oceanographic, geographic, and space science communities because there have been systems established to collect data

and a lot of people can reuse that data. Therefore, it is all collected in a standardized way and a centralized place. In contrast, biomedical data are collected by individual PI's in individual labs using the local standard, which makes it inherently more difficult to share. I am curious to your reactions; whether you see that as well and what ways can we address the challenge.

Nina Preuss
Following up on the question out there, the first thing I would have done would be to go onto NITRC and search for MATLAB scripts and then search the forums for others' solutions to these issues. Each investigator does their research differently and MATLAB is creative enough that they are bending it to their specific research methods. I would encourage others to go out and research these solutions on forums and post forum questions as well.

Yuan Liu
I think a good thing about NITRC or NIF as a clearinghouse is that they really serve as a middleman with community feedback to the data contributors. I think it still boils down to several fundamental challenges like data format. Like Dr. McNutt mentioned, if everyone in oceanography uses two data formats, this community buy-in is very beneficial. Translational tools would also be very useful for moving between data formats and making contributions easier for users of secondary analysis. It is easier in some fields than others because genomic data is linear for example. But if you want to integrate different types of formats, that will be much more challenging.

Judy Kosovich
I was speaking to someone who worked at a coroner's office and they told me that every cause of death goes into a public record. I don't know how accessible that is but it seems like a valuable source of information and perhaps some kind of interaction between the science community and public records could create a valuable database. I was researching radio frequency pollution and came across a young lady who had three friends die of aneurysms. I had not come across this in the literature but if people are paying attention to coroner's reports, then maybe there would be more research on that outcome.

Nina Preuss
There was an article in the past two days on the subject of new startups that specialize in publicly available data and aggregating it along with data that is not as public and finding mash-ups. That way, researchers can try to find correlations. I think we'll see a lot more of that in the future. Banks are starting to rely on those startups to find information about small businesses that do not submit as much financial information but still provide clues about their status in other data forms.

Greg Hale
I'm a graduate student at MIT. I think that we can be optimistic about this goal of sharing data because the problem is completely solved in software engineering, without even meeting the goal of standardizing data formats. There are platforms that abstract away from data formats. For example, there is no common data format for the Web or word documents. Github and Travis CI are great programs that show that these issues are tractable. I agree that community buy-in is really important. 98% of our scientists' thought cycles go towards making sure that we have a path for getting tenure in the future. If data sharing isn't on that critical path, then I'm

not sure how much buy-in there is going to be. My question is: when will I be able to publish a science paper that is not a publishable unit with a conclusion, but is rather a dataset that will be analyzed by others?

Nina Preuss

I brought up the 1,000 Functional Connectomes example because that there was the data without a hypothesis and shared for free on NITRC. It wasn't structured in any way. Because they did that, people downloaded it and played with it and came up with hypotheses. They published and got those three R01 grants. The more people hear about situations like this, the more they'll connect why there is reason to just share the data and how you can benefit.

Marcia McNutt

Nature has a new journal devoted to datasets. Science has not yet launched a journal that is devoted to descriptions of data sets. The readers of Science would complain if we took a slot away from a paper that had a result and gave it to a paper without results, but right now I would urge you to look into that new Nature journal.

Yuan Liu

I would also like to add a little bit more to that. I think that some journals are now willing to provide short comments and publish some data for people to access. There is also an issue of negative results. Scientists collect a lot of unpublishable data, some of which is a negative result. Having a place for this data will prevent duplication of failed research.

Jerry Sheehan

There's another consideration here, can we have data publications that are recognized as valuable counting towards tenure and promotion? If not, what can we do to advance data curation and maintenance as appropriate career track?

Marcia McNutt

There are article-level metrics where it is possible to go to your paper and see the pageviews, citations, downloads, tweets, etc. For data papers, this is a measure of the impact of a paper that might go way beyond how many times it is cited. Especially for data papers, it would certainly be an early indication of the long term impact of that paper.

Jerry Sheehan

In your question, you mentioned solutions that come out computer science or informatics and how we often say we can fix things if only we get access to these tools. There are many ways of trying to do that: we might look long-term and say eventually we will train people who are trained in both science and information technology. The alternative is bringing those groups together to work on projects jointly. We've tried to do that at NIH and NLM and give small supplements to grants at other institutes that help hire an informatician or a data manager. We saw greater enthusiasm than we anticipated and other institutes were also interested in putting money into these programs. I thought that was an interesting model for bridging these two communities. I am interested to hear if others on the panel have heard of similar models for bridging this gap?

Paul Albert
Statistics has a long history of that and statisticians collaborate with other scientists and physicians to publish papers and perform innovative methodology. This is motivated by the science but really mostly about the analytic techniques, where a separate literature appears. For tenure, you are evaluated by your contributions to the statistics literature, and not so much your collaboration. More and more in the training for biostatistics, there is a data science orientation. Students are learning a scientific area very well and then are also learning much more about informatics. I see more and more that there is this sort of interdisciplinary culture.

Nina Preuss
Data scientists graduate and make $117,000 a year. They know there is a great market for people who can merge data, statistics, and science. Universities usually train people in math, physics, etc. or in the scientific domains like neuroscience, but they don't usually cross between the two. Students don't learn how to best structure the data. Moving forward, that's a wonderful way for people who are graduating to cross over from statistics and applied math to science so they can help be on grants and help universities share data in a structure that makes sense.

Jerry Sheehan
We have a lot of AAAS fellows here today. How many of you feel you are adequately trained in data sharing and informatics? I don't see a lot of hands going up so I think there is quite a lot of opportunity there.

Question
I am a AAAS fellow and I am the representative at the Big Data affinity group here but I was not trained formally in informatics, which I think is something that is lacking. I think that negative results in science will become a big issue, which is to say that the pressure of publication is associated with it. How are we going to be able to apply these same policies to that information? In terms of the data quality and management side, I agree that we are not trained as scientists to handle those kinds of tasks. In this world of now providing open data, whose responsibility is it going to be to ensure quality? We all know that secondary analysis is nothing if you don't have quality data.

Marcia McNutt
I can give an example from the oceanographic community. In my view, there is nothing more important than the quality of information. If you have quality information, then everything flows down from that. Good research projects, good results, good policies, and good translational medicine all result from good science at the beginning. It is very hard for someone who is in charge of a data repository to be able to police the data that comes into it. Peer reviewers have a responsibility to evaluate quality when they go over results, but as you know, there are journals on the scene with a variety of standards. Some are high-quality, some are minimal quality, but it is still possible for people to collect data and deposit it in a repository even if it is not connected

to any published material. Thus, the data may not have been put through any quality control before it was uploaded. Putting the onus on publications to say that data is quality controlled in this situation is not correct.

For example, an investigator looked at quality of data as a function of the PI who collected it for all of the bathometric data in the National Geophysical Data Center. Ship tracks are random through the ocean so the investigator looked at crossover error, or whenever two ship tracks crossed each other. The investigator looked at the size of the mismatch between the ship tracks as a function of the scientist conducting the research. It is hard to pick out between A and B whose fault this is, but when you have the final analysis, you can see that a couple people consistently come out with crossover errors. This research translated into funding in the future. Some researchers were rewarded for consistently collecting good data and some were punished for not being watchful in terms of collecting high-quality data with their ship time. The investigator was also able to account for the few incorrect datasets and improve the NGDC records because of this analysis of data quality.

Yuan Liu
By sharing data itself, there is inherent policing. At a recent symposium, someone reanalyzed another researcher's public data and discovered the previous conclusions were wrong. By putting data out there, peers can try to reproduce and reanalyze it. We were talking about lack of training and opportunities for training. If you are being trained in bioinformatics, can you be hired as a professor in academic science? We use bioinformatics as a partner and a tool, and we need to provide faculty positions for these newly generated scientists who are bilingual. We should recognize them as part of the workforce for the biology and neuroscience community. I think that is something we lack in policy and strategy: how to really embrace these new researchers who are biologists and statisticians at the same time.

Jerry Sheehan
The optimist in me likes to think that as we increase the access to data and the steps that we've taken already through repositories and tools, the value and contribution of data science will become more visible. That might be self-reinforcing to some degree but we need to push it. People who are in this room and people in the Big Data affinity group can really play a leading role in reminding the broader community about issues of data quality, management, and stewardship.

Paul Allen
We need to recognize bioinformaticians as key players and PI's, which has happened in the biostatistics community over the past 40 years. These people shouldn't be seen as service on a grant or a co-PI, but rather an independent grantee with specific methodologies. Money speaks, so you get a professorship if you bring in a lot of money.

Nina Preuss
As someone without a PhD, I did have the opportunity to be a PI on an NIH grant. That trend is happening and it is being recognized that non-PhD contributors have a role in bioinformatics research.

Yuan Liu
We have been making effort to call for people who are not just neuroscientists but people with strengths in informatics and computational applications. For example, Dr. Ascoli has grants with us and he is bilingual in neuroscience and computational science. At NSF, we established collaborative research in computational science and require 2 PI's. The computational science doesn't collect data at the bench but their contribution is equally important to neuroscience. We need academic institutions to realize that these people are very important.

Jerry Sheehan
I have one last question. Cognizant of the fact that we have a lot of different stakeholders represented here, what is the one step that your group can take to advance data sharing?

Paul Allen
In the biostatistics field, which has a lot of applications in neuroimaging and neuroscience, the new statistical methods for complex analysis need to be accessible and reproducible. The data is the obvious point to share but the methods themselves need to be validated by other statisticians, peer-reviewed and scrutinized, and made usable by the community as a whole.

Nina Preuss
Industry and academia need to be part of the culture change and provide the tools and effort to support researchers. We need to make it easy for these people to share data and put software and technology out there in the open. What we consider to be big data now will be small data in five to ten years, so costs will become less and less of an issue moving forward. It is mostly a cultural change that is required.

Yuan Liu
From the funding agency point of view, we can work together with developers by funding them (NITRC and NIF) as well as through communication. Journals don't need our money to fund but we work together with them. We can address issues like optogenetics data sharing through workshops, seminars, and communication.

Marcia McNutt
From the journal perspective, we do not view our role as either luring the community with a carrot or beating them with a stick. We see it as a dance, where we are partners. The dance works best if we decide what we are doing together first. We have to agree on a common goal. I think it is important that the community is a willing partner in all of this. We have seen that there are communities that are motivated to be transparent, to share data, and to get behind this cause. We are ready and available to be a convener to this community. We can be a convening body for journals as well as meetings between societies, journals, academia, and funding agencies.

.

KEYNOTE: NEUROMORPHO.ORG

The keynote address was delivered by Dr. Giorgio Ascoli, creator of NeuroMorpho.org, a successful neuroscience database. NeuroMorpho.org contains morphological data from thousands of neurons, which were collected from publications by many different researchers. This database has had great success in standardizing the way morphological data is represented. Researchers can use the database to classify their neuron datasets and to model characteristics such as function, development, and network connectivity. Curating a large dataset like Neuromorpho.org allows for researchers to investigate fundamental principles of neuroscience that require immense, diverse datasets. Neuromorpho.org is interoperable with other similar databases and tries to mirror any data that is uploaded to another database and freely available. Data sharing will occur more frequently if the neurocience field takes lessons from other research fields, incentivizes searches for solutions to computational problems, and publicizes the information on which researchers are publishing their data.

GIORGIO ASCOLI

I will talk about the one example of data sharing that is nearest and dearest to my heart, NeuroMorpho.org. I will give a very brief history and then I will spend most of the next hour showing you through the website and showing what data sharing is accomplishing and can accomplish when it works. The project started due to needs that came through my lab and those in parallel from many other labs. I was one of the original awardees of the human brain project when it was a US endeavor in the 1990s and I needed a large amount of data to constrain the computational models I was interested in. I realized that only some of the data was out there, but not all of it, and it was very hard to find and to collect the data that was available.

I have just a few introductory slides, and then I will go live online. The data that NeuroMorpho.org is concerned with is the structures of neurons. So the typical laboratory pipeline is that you first fill the neuron or stain the neuron by making it visible somehow. There are many diverse ways to do that, such as bulk intracellularly, genetic labeling, and immunolabeling. Once the neuron is contrasted with respect to its background, it can be visualized microscopically. The two leading ways to do this at the level of whole neuron morphology are bright field microscopy and confocal microscopy. There is a higher resolution technique called electron microscopy, however it does not have the span of the field to capture entire neuronal arbors. Here is where digitization comes in. People can just look through microscopes, they can also acquire image stacks, but then there are ways to combine hardware and software, now mostly imaging and software, to render the neuron in digital formats, which are essentially linked X, Y, Z coordinates that contain the full information of what this neuron does in space and where it is. The stunning diversity that fascinated all neuroscientists since the early days of Cajal and Golgi, continues to be a draw for all of us, and this is just a sample of all the variety of neurons that are being reconstructed (pictures of neurons on the screen) all around the world. These are just some of the neurons on NeuroMorpho.org and each one is a labor of love because just the tracing takes somewhere on the order of several weeks to several months, so we are talking about massive amounts of data to the finest level of detail.

So, what are the data good for? In most cases, neurons are traced just to establish their identities. The shape of the neuron is one of the most quintessential signatures of what that neuron is in terms of how we understand neuroscience today. So even if you are just interested in the molecular phenotype characterization or electrophysiology characterization of a neuron, you still typically want to look at its morphology to make sure it is the neuron you think it is. In addition, or after establishing neuronal identity, people have been using reconstructions in digital form to model the function as well as the structure and the development of the neurons; to establish network connectivity, which is becoming in the connectomic area more and more important of an application; to do morphometric analysis and comparative analysis between species, between experimental conditions, and so forth; and increasingly, applications that have to do with the usage of the data when it is actually stored in databases. Any one lab only has the interest and the bandwidth to peer into a very small amount of the complexity of the nervous system. But if all the data is shared and curated in repositories and databases, the diversity and the complexity of the data becomes available both for mining and scientific purposes, but also for just understanding at all levels, not just at the highest levels. Not just for the best and brightest among the scientists who want to understand how the brain works, but also for young students and trainees, all the way down to kids who might want to start their own fascination as to what is in our skulls.

Neuromorpho.org was launched under the US Human Brain Project in 2006 after years of testing. Since 2006, hundreds of articles have been published based on data from the database. These are scientific discoveries and analysis, in some cases quite prominent discoveries, which simply could not have been possible from data from individual labs, even if the investigators who made these discoveries were the ones who collected the data. And the reason for that is that most of these data are data that come from a variety of different studies. So if you want to establish, for example, something like principles governing the operation of synaptic inhibition in dendrites, you cannot just do that in one neuronal type, because that would not be a principle, that would just be how it works in that one neuronal type. But if you find the same pattern of activity over and over in many species, in many developmental stages, in many kinds of preparations, then you can claim that maybe it is a principle. And you can look at the theme here, if you are really seeking a theory of neuroscience laws or universal properties, you really need massive amounts of data, well beyond what individual labs can do.

It is not just papers. The power of the archive itself extended way beyond the reach of what we researchers imagined ourselves, and to surprising extents. One of the most surprising ones was when we were contacted in August 2010 by the director of the Applied Math Olympics in China with a very interesting email, saying that they were going to select neuronal classification and the data from NeuroMorpho.org for that weekend's competition across all of the high schools of China. So they said please make sure that your database stays up, and they actually nearly crashed our servers because of the huge numbers of visitors to the site. I also mentioned that this can serve as an inspiration for even children. One of the achievements that I am proudest of is that NeuroMorpho.org was named site of the month in "Neuroscience for Kids", which is a resource that I myself continue to look at and encourage my kids to look at. It was also highlighted in Scientific American in a piece called "Know your Neurons." And generally, unsolicited positive reviews from journals and books. Last but not least, last year, we organized a conference at George Mason University after the Wellcome Trust contacted us asking what they could do for neuroscience and data sharing. I told them that it would be useful to foster a culture of data sharing and convince people to share more. Later, the Burroughs Fund was gracious in allowing us to invite 50 of the most prominent players in the field of digital reconstructions to come and share their success stories for a full 3 days.

The one example that I am going to highlight is the model we put together in 3D of my favorite structure in the brain, called the hippocampus. This is just an example of a few of the neurons on NeuroMorpho.org. It is only about 30 neurons, it shows you the complexity of the data and you can actually see the shape of the hippocampus. The reason that I wanted to close with this one example is that there is a sculpture that an artist at George Mason and I put together with pilot funds for the collaboration between art and science. I figured that we really had to embody the 3-dimensional structure of neurons and circuitry in 3D. As an example of a more scientific example of this, I am just going to show an application of data from my own lab that we could not have done without NeuroMorpho.org. We created a 3D rendering of the hippocampus and embedded all the neurons we could get from the archive in their proper locations. Just based on this we were able to statistically compute the overlap probability between the axons of neurons and the dendrites of other neurons and therefore to establish the statistics of the potential circuitry which is as of yet unknown in quantitative terms. We are still working on Cajal's qualitative description. We were able to show some of the potential synapses one neuron could make in the hippocampus. This example goes to show that this kind of complexity is impossible to tackle without digitizing the data and without sharing the data thoroughly.

Before I switch into demo mode, I really want to give credit to the people who do the work and did the work in the past. There are a lot of programmers; there is a lot of information technology infrastructure and it is always the case when you try efforts like this that intuitively we think that other people did the work and we are just putting it in one place, but there is more to it than that. There are a lot of data curators. These data come from different labs and these labs are asking different questions which means that the format, the meta-data, of these data are all different and of course the strengths and weaknesses in each data set are different. Some data might have very diameter resolution for each branch and not very good shrinkage correction for the tissue, and other data may be just the other way around, because of course every lab tries to maximize the aspect of the data that is relevant to their scientific questions. There are a lot of lab members in general that worked on this project. Interestingly, NeuroMorpho.org was picked up by a lot of different resources so there are now a lot of external developers both in companies such as the Mitre Corporation here in DC, but also in places like the UK where Robert Cannon and Padraig Gleeson are working on code that interact and port with NeuroMorpho.org. The same is going on in NIF (Neuroscience Information Framework) in San Diego. And of course support. The R01 grant that started this and still supports much of it is going through for the nineteenth year and is now funded under BISTI (Biomedical Information Science and Technology

Initiative) because the Human Brain Project is no longer here in the US. As the project grew, we were very fortunate to secure funding from other sources including Department of Defense, National Science Foundation, and non-profit foundations. Finally the people who actually put all this together in the lab: Maryam Halavi started this and Ruchi Parekh is now in the lead.

We track and keep a list of papers that used NeuroMorpho.org, which are searchable on Google Scholar.

Now I am going to start navigating around NeuroMorpho.org. We currently have more than 10,000 reconstructions of neurons and we have been working for a year on a major release that will bring the content to more than 25,000 neurons. There are quick facts and stats on what is on NeuroMorpho.org. We are getting data from all sorts of species and brain regions, and we are trying to translate and estimate these data into things such as how many person-hours of lab work this takes, the total length of dendrites and axons on the site, how many branches we have, how many countries are visiting, and so forth.

You can also see what changed from one release to another. Typically at each release we add data sets and we add meta-data such as species and brain region. We also continually update functionalities such as statistics functionality and specific search functionalities.

The terms of use are quite simple. The data is free and available for everyone to download and use. There is no restriction, no registration; we just keep track of the IP address so we can do geographical distributions of where the data is downloaded from. The only thing we ask is, please cite the papers of the authors that you are using so credit can be given to the people who actually collected the data and traced the data. There have been some instances where people did not cite the original investigators, and we caught that and brought it to the attention of the original authors. We added a terms of use tab so that it is easily seen, in order to avoid this type of problem. The detailed statistics give you an idea of hit over the years by country, by cell type, by species; we can see what is more popular and less popular. The increased views in 2010 are due to the Chinese study where, given the meta-data and the morphological data, the students attempted, using machine learning and applied math tricks, to determine what the meta-data is just by using the X, Y, and Z coordinates of the points of each neuron. This would be impossible to do with only 100 neurons, but when you have 10,000 neurons, then you can start producing smart algorithms. They found that it was pretty easy to classify the animal species in this way, but unfortunately it was a trivial effect because they simply found that the axon length of humans tend to be bigger than rats. So this has to be taken out in order to really see the "gems". We acknowledge all the data owners and we give links to where they post their own data or if they have a website they want us to link to.

Let me start going through the data themselves. We have the ability to browse data by a variety of dimensions such as species, brain region, and cell type. Everything on this site is clickable. If you click on a name or graphic, it will actually give you the data.

The rat used to be the dominant species for a while but the mouse is quickly catching up so we think that it will be reversed. In fact, many of the neurons we are working on now for the next release are from *Drosophila*, so you can really see where the trends of neuroscience are coming through. Interestingly, it is not just animal models; there are many post-mortem data from clinical studies in humans (>2000 reconstructions).

As an example, when you go to salamander, there is a list of cells types with a preview mode where if you just pass your cursor through, you can see the morphology of neurons. For any neuron you select, you can choose to download all of the data, the data with the meta-data, the images, and so forth.

All the pages for browsing modes are similarly organized into pie charts. You can also search for random neurons. One of the functionalities that we introduced most recently is the neuron atlas which allows you to search directly by 3D exploration of the anatomy. For now we have only implemented this from the rodent brain which we did in collaboration with the Allen Brain Atlas. You can rotate the image and click on the dots which are individual neurons, which can give meta-data for each neuron. You can choose different parameters to look for and if you click on an individual neuron, you will be given the actual reconstructed image of it.

In search mode, you can search by meta-data which are uniformly assigned, such as species classification from NCBI taxonomy. You can narrow down search parameters such as by sub-species or by transgenic strain. Search can be narrowed by experimental parameters as well, such as staining or experimental condition.

Last, you can search by keywords.

This is the easiest; it is like a search bar in Google. So if you know what you are looking for, you can simply type it in and get those neurons. When you get to the neuron page, you can download the file itself, which has some meta-data and a long list of X, Y, Z coordinates, diameter, and intracellular connectivity information. On the neuron page there are the main kinds of meta-data as well as the reference articles. We are very careful to refer to both the original articles from which the data were reconstructed as well as the secondary articles if the data were imaged or reconstructed. The articles have links to PubMed. You can also go directly from the literature to links to neurons on NeuroMorpho.org.

I want to show you the kinds of data available. You have the original file that was contributed by the lab. We have the standardized file, because all data needs to be put in the same format. We keep a log of every change that is made to the data so that there is transparency for the data. Finally, you can visualize and interact with the cell image using very simple clicks. This is in Java so can be used on any platform.

We do have some minimum standards for what types of meta-data need to be included when data is reported. In the section "how to contribute data to NeuroMorpho.org", we say to fill in as much as possible in the meta-data form. The idea is that if you have that information in your lab notebooks, we would like to know it. In some cases the gender is not reported simply because that information is not known or available. So if there is ever meta-data missing from the list, it is because it is not reported in the literature. But there are some minimum standards for reporting data.

The very last functionality is the search by morphometry which gives morphological data so you have X, Y, Z coordinates and a lot of geometry information. We have a tool that we created in the lab to do our own analysis and then created a Java-compatible version of this tool so you can essentially search the neurons not just by meta-data but by morphometry. For example, all the neurons that have at least 1000 branches, and it will instantaneously load the distribution of values for that statistic from the database so you know what to aim for with the data that is in the database. It tells you how many neurons fit that parameter and you can organize them by animal species, brain regions, etc. or by summary. The reconstructions with that morphometry of >1000 branches will load and you can make powerful searches with this feature.

Let me get to the core issue of how we get the data. We have a constant literature search to monitor the literature and we look for when the data gets published. In the absence of the culture of data sharing where people would want to share their data as soon as it's out, we have to wait until the publication is out. We find out that neural reconstructions have been published and we simply contact the authors and ask if they would like to deposit this data in NeuroMorpho.org. We tell them that they just need to send a .zip file and we will do most of the leg work. We allow authors to review the data before it is put up on the site so they can make edits and approvals before it is posted. You can actually search the literature by PubMed ID to see if a published reconstruction is in the database. It is now NIH policy that all publications have PubMed IDs. This makes it much easier for use to find publications. If the data is not in the database, an email is sent to us suggesting that we check out that paper because it might have reconstructions. If the PubMed ID search comes back positive, it will say that the data is available, and provide a link; or that the data has been sent and is being processed; or that the data has been requested and was denied.

We actually have a database of all of our communications where you can see all the papers that have data in the repository, the ones we are currently communicating with, and ones which are not available.

As a journal editor, in *Neuroinformatics* we have implemented a policy that we are not mandating data sharing; we are simply mandating that authors of each article end the article with an explicit data sharing statement. That way, authors do not have to share the data, but they have to explicitly state that they are not sharing. This encourages researchers to share the data because it is public knowledge if they do not.

For the data that is not available, the typical responses we get are that they lost the data, or the contact information is incorrect, or some simply decline to share.

I wrote an editorial of problems with data sharing. Typical reasons people decline to share data when asked: creative lies (e.g. my hard disk crashed); personal commitments (e.g. exclusive agreement with X); "I want exclusive rights to my own data"; matter of trust: others will misuse/over interpret the data; it is a tough world ("why should I give it away to my competitors?"); time commitment (too much effort, no time); "I will do this really soon…"

I want to end my demo by showing how you can search for data availability. I can show you from these numbers that as of the last release, less than half of the data that is out there is in the repository. If the data is available for data sharing we are committed to putting it out there, so this is dense coverage of whatever is available. But what is available is far from 100%, but we are getting close to 50%, up from 30% in 2006. About 2-3% of the data is from authors who contact us wanting to contribute data but the vast majority of data is from us asking researchers to contribute.

The availability status is just one way to search the literature; you can also search it by the information of the publication such as author, journal, reconstruction information, etc. It is all color-coded by whether it is available or not.

Sometimes when we contact researchers to try to get them to contribute data, they respond that they already had plans to share the data elsewhere. We actually consider those to be positive responses. If they say that they are in the process of setting up their own database or are depositing in another database, we consider those as positive responses because our policy is that if the data gets deposited elsewhere, we simply mirror them. We put up links and there is code in our server that makes it completely seamless for the user. It is as if the data were on our server, except it is not. It is transparent as to where the data is coming from. This does not happen often. But if you click on one of the neurons that are from another database, you will see the link to the original archive and the link to the original neuron. The link to the original neuron will cull the individual neuron from that archive entry. It will link to the archive so you can see their set up, and so forth. This used to be the case earlier on and we have actually seen this trend going down, because now people who want to post the data but do not want to do the work involved, send their data to us and we do the work for them. This is a small fraction of the data that is in the repository.

There are some examples of similar databases in other subfields. I wish there were many, many more and with many different models of how to make it come alive. The one example that I am thinking of is the ModelDB repository in SenseLab out of Yale University. It was actually originally funded through the very same US Human Brain Project in the late 1990s. If you go to SenseLab and click on ModelDB, it strives to achieve the same goal of covering all the scripts for neuronal simulations of biophysical activity and electrophysiology. It started as a neuron-specific model, but it expanded and now has models created with all sorts of software. In fact we made ModelDB and NeuroMorpho.org directly interoperable with each other, and with several others, through NIF so that now there is a very easy pipeline if people want to do compartmental modeling which is one of the main applications of these data. They can take the morphology from NeuroMorpho.org and the model from ModelDB and run their simulations within hours, whereas this was the kind of thing that in the mid-1990s would easily take months to get to the point where you could start this. So, there are a few cases, not very many, I think there is a lot going on in the brain imaging community which is quite separate from the level of individual neurons. Obviously there is a lot in the genetic community, and the application of the genetic community to neuroscience, but I wish there were a lot more.

There are currently a few ways in which databases are linked across various scales in neuroscience. In fact the 3D rat brain I showed that came from the Allen Brain Atlas came from an attempt to link genetic and neuron morphology data. We have done it in a few cases with Brain Info, for example, which is a brain atlas out of University of Washington. They have a browse-able knowledge base of brain regions and they have links saying which neurons are in this region and if a neuron can come from that region, they will open NeuroMorpho.org in a sub-window. We are developing capabilities to do the same, when people search for region we give the option, if you want to know more about this region, go to BrainInfo.org. The idea of doing this comprehensively was the original idea behind the Neuroscience Information Framework (NIF). There is a lot that the NIF has achieved already. All of the so-called deeply registered resources within the NIF, including NeuroMorpho.org, allow for cross talk between resources and the PubMed link out is enabled through this. But it is still sparse and there is a lot more, even in the database and the electronic tools world of neuroscience that is not yet at that level.

We are able to use the advances in artificial intelligence and deep learning to adopt a crowdsourcing mode for the database. A lot of the searches that right now are done locally at the Krasnow Institute were not sustainable when they were done by me originally and then I started distributing them through the lab and that would not have been sustainable in the long term from the lab itself. I created a course which is now in the 6th or 7th year and it is a dry-lab course called Neuroinformatics for neuroscience undergrads where students come in, we show them how to mine the data, and then they actually have to do the work for the rest of the semester. Of course there is the issue of quality control because these students are students who took introductory neuroscience courses and now they have to figure out whether a reconstruction is an axon or a dendrite, whether agouti is a rodent or not, and it is very challenging from that point of view. So what we did is, some of it we automate so when we have enough students, we can put more than one student onto the same data set and look at the distribution of their answers to see if there are conflicts. We thought eventually that this has to become broader and I am looking at a time where there are not going to be 1 billion people on the internet right now, but 7 billion people on the internet. Right now we are mining at a very small proportion of the gist of the human soul that can actually extract that knowledge that machines cannot quite do yet, which is to have a little bit of expertise, enough to say something about the meta-data. We actually had a vision that we brought up to the National Academy of Science when they organized the future initiatives program and received some funding from Keck to pilot this, and we have a collaboration with Michigan State University to see if we are able to get some good annotators from Michigan State without ever meeting in person. In the meantime we are collaborating with the director of the neuroscience undergraduate program there and they are getting credit for whatever they are able to do from their institution. But the idea is there that you can open it up, crowdsourcing, to the rest of the world. The question is: at what point will we reach a threshold of data and annotation such that machines will take over? As I was saying before, if you have lots of data, eventually machines can figure it. So, eventually this will have to become crowdsourced, and there is going to be a threshold for this that is going to be passed.

Not only should there be openness and sharing of data, but also for the code of our website. When it comes to this, I think that the coding community is farther ahead than the data community, in the sense that there is a very mature open source movement that we are certainly leveraging as we look for code and adapt code. The code for NeuroMorpho.org is available, and I am sorry to say that when some people look at the code they say that it can be done better, faster, etc. To that I say "please, be my guest." On the budget of an R01, if I hire a professional programmer, I have to fire two post-docs, and I simply will not do that. I know there are programs for particularly pushing small business research. If a small business can reproduce this, we should run with it. I wrote a paper in Nature Reviews Neuroscience called "Mobilizing the base of neuroscience data" in which I took each and every example of the negative answers that we got and essentially rebutted them. Although some of the concerns were legitimate, the reality is that the negatives are minimal. Even from the personal interest point of view, we saw citations of given papers going up very dramatically, sometimes doubling, after they deposited the data in databases. In fact it might increase a researcher's chances of getting their grant renewed if they can tell the reviewers that not only did they publish their data in a journal, but the data was also mined by other labs, leading to even further publications. We keep track of the data people deposit so that we can show them statistics on their data such as number of downloads, location of downloads, papers using the data. This way we can send this information to the authors of the original data and show how their data was used, and this information can be used in grant renewals. Study sections are still very much publication-driven, but we should give more attention to the secondary data cascade when reviewing grants. So, yes there are negatives to data sharing, but the benefits far outweigh the negatives.

One point I would like to make is that we are not alone. Neuroscience data is complex, but there are other fields that we can learn from that have a much better mentality about sharing data such as in physics, astronomy, and geography. They are also dealing with very complex data, so I think in the next forum or meeting, it would be helpful to invite a representative from another field who can give insight into the data sharing practices of their own field so that we can learn from them not only sociologically but also technologically. Another point I have is that people say that in order to do this, you have to have money. This is partly true, as the databases like NeuroMorpho.org are funded by grants; however, there are other creative ways to do this. One thing that we did was the Diadem contest, which did not cost very much, about $75,000 altogether, calling for the automatic solution for tracing a single neuron down to every branch, every detail. A challenge is whether we can make better algorithms to do the annotation of data more automatically. Maybe we can have another challenge calling the world to come up with good solutions. From the Diadem contest we did, we did not have anyone who reached all the criteria, but the most hopeful solution was provided by a group of computer science graduate students in Switzerland which provided some solutions that could have the most potential to solve some of the automated problems. So I think there are other ways besides just providing grant funding that are more innovative and less expensive. In fact, the director of NINDS Dr. Story Landis came to the final awards ceremony for the Diadem challenge and she looked at the work that had been done and was surprised that only $75,000 bought all of that work, and thought that this mechanism should be used more. The CEO of the Allen Institute was also there and had similar remarks. So I think that this would go a long way if more of these types of challenges are done. Only 1 year after the Diadem challenge, someone downloaded the algorithms and codes that were posted and was able to make modifications that would have, in fact, won the prize. So again, that type of development was not possible without open data and coding. I think that the same model could be used for a challenge for the infrastructure as opposed to the reconstruction.

I would say there are hurdles and really the remaining challenge is that more than half of the data is not shared. I think that is really where we need to push right now. Yes, there are technological issues, but I think there really needs to be a cultural change. I think that in some cases there can be policies and laws to enact that and I think that the funding agencies and journals can do a lot to get us to the critical mass but in some cases, my experience with this data is that if you make the information as to who is sharing public, people will come. A lot of it is just social constraints. Again, the biggest challenge to data sharing is getting people to share more data.

PANEL 2: BUILDING THE ROAD FORWARD

The speakers at the second panel provided their perspectives as stakeholders in academia and federal agencies by discussing values inherent to data sharing that should be communicated clearly. In an era of interdisciplinary science, collaborations between disparate research fields can be very useful. We can take research on microorganisms within the human gut flora and begin to investigate their relationship with chronic infections and neurological diseases. This kind of cross-disciplinary research is enabled and bolstered by the data sharing process. Institutions like DARPA buy into open science and data sharing because they believe that it is their governmental obligation to share innovations and discoveries with the public and the research community. Data sharing improves society by enhancing science research and reducing issues in reproducibility and discoverability of findings.

Currently, the neuroscience community is not data-centric, but advances in standardization and increased communication between data, tools, and literature will make a data-centric environment possible. Programs like the Big Data to Knowledge initiative will help us get to that point by advancing the technology behind big data, developing a data science workforce, and facilitating the broad use of research data. Data sharing becomes complicated when large data sets are involved. Research into methods of analyzing raw data and transforming it into manageable information is a key to improving the data sharing process. Biologists are not always trained in computer science, so there is a need for user-friendly tools and for support from experts in coding and data manipulation.

There must be community buy-in from publishers, funders, institutes, industry, and researchers to establish data sharing as a standard practice. There is a need for an all-hands-on deck approach, with industry technology adoption, government oversight and funding, research institution cultural changes, and people in all fields and domains who are willing to collaborate.

JENNIFER BUSS

Moderator

I'd like to introduce myself; I am Jennifer Buss, the Director for the Center for Neurotechnology Studies at the Potomac Institute for Policy Studies. Thank you all for coming today. The Potomac Institute is a non-for-profit think tank in DC. We do science for policy and policy for science. We shepherd discussions on key science and technology issues that are facing our society to develop meaningful science and technology policy options to ensure their implementation at the intersection of government and business. The topics of data sharing and open access have been analyzed in science policy for a long time. It is therefore crucial to discuss them in context of new programs like the BRAIN Initiative and to present the tools that make research more effective. Today, we bring together the policymakers from academia, industry, publishing, and government who play an integral role in defining and implementing a well-designed data sharing environment. Without further delay, I would like to begin our panel titled "Building the Road Forward."

RITA COLWELL

Speaker

I would like to give a brief commentary for the discussion. This entire issue concerning data is not new. In fact, the work I did years ago, in what was then called numerical taxonomy was simply coding data that could be used for identifying bacteria and viruses using computers. This was 30 years ago now. With respect to building a neuroscience database, the microbial database that we established and many of the issues that were discussed by the previous speaker were resonant then and continue to be now. My colleagues at the National Institute of Dental Research, NIH, and I published a book on how to code data and enter it into databases. As it turns out, microbiologists have Bergey's Manual to identify microorganisms and that manual also has been uploaded and the data made available, complete with illustrations of microorganisms. I would like to emphasize that we are in an era of interdisciplinary science and interdisciplinary collaborations. With that said, I will be iconoclastic and present some data on metagenomics of the human gut.

You might ask how this is applicable to a neuroscience research group. This is because we are finding that what happens to microorganisms in the gut and what they produce has a great deal of effect on behavior and well-being and on neurological functions. In addition, some of these products of microorganisms of the gut cross the blood brain barrier and influence pathways of some of the neurons that were described so beautifully earlier today by Dr. Ascoli.

I will describe briefly a study on the metagenomics of the gut of hospital patients in Calcutta, India. The data I am presenting is part of a long-term study that I have been involved with over the last 35 years. This particular analysis, with the National Institute of Cholera and Enteric Diseases (NICED), was one of three phases where patients coming in to the hospital and their clinical stool samples were analyzed for 26 different pathogens using standard techniques. That is, culture was used to determine the presence or absence of enteric viruses, rotavirus, norovirus, parasites, giardia, endamoeba, salmonella, and a variety of enteropathogenic coli, etc. At first, they did the analyses and then sent us DNA extracted from the stool samples which we then sequenced. They sent us samples from patients for whom they had concluded what the pathogen was. We had developed a mathematical approach so that we could take the sequence from the computer and from the sequence, we are able to deduce which pathogens were present, the quantity present, and which specific genes were present. The second lot was composed of 30 samples from both patients where doctors could not identify a pathogen as well as controls. The controls were samples from individuals who were perfectly healthy, without manifestation of disease and from individuals from the same family or same community as the patients. In the third phase, they sent a set of 39 samples from patients for which they could identify the pathogen in 19, could not in 10, and another set of controls. With the analyses we obtained some very interesting results. By downloading all the data from the human microbiome project, which is available on the web through GenBank, we compared healthy individuals from the United States with the healthy individuals (controls) from India. One of the major findings is that the healthy Indian individuals carry low levels of pathogens. In other words, they are able to tolerate a small number of pathogens and a larger variety of pathogens in their gut flora. Another interesting finding was that healthy individuals in India had higher levels of protective bacteria in their gut. In fact, if you have some sort of enteric disturbance, often the suggestion by your doctor is to eat yogurt. We suspect that since most Indian diets incorporate yogurt, that may be how that custom evolved. Now the other aspect I think is important to point out is that through the DNA approach, we were able to identify the same pathogens as identified by traditional methods present in patients with known etiology as well as the presence of pathogens like *Shigella* and *E. coli* that could not be picked up with traditional methods in the Indian patients with no known etiology. One finding is that the standard laboratory techniques are rather opaque and with the more sophisticated approach of sequencing and mathematical algorithms we can much more accurately identify species, strains, and sub-strains of bacteria.

The message is that many of these pathogens produce toxins and it is very likely that the non-pathogens are producing metabolites that may be protective. We are now beginning to understand that the gut flora plays a role in a variety of what was previously believed to be chronic infections rather than infectious agent-mediated. We may find that for some diseases with difficult-to-identify origins such as Parkinson's disease and other neurological diseases, we need to look to the gut.

The gut flora of healthy volunteers indicates that the microbiome of healthy humans in India is markedly different from that of Western Europeans with respect to mix of pathogens and that the population in Calcutta tolerates a small number of pathogenic microorganisms that would comprise a disease state for Westerners. Multiple pathogens were identified from the disease patients. This was observed both by standard tests and our DNA approach. We have learned that a patient coming into the hospital with an enteric infection usually carries 4 to 10 pathogens, not a single pathogen. Therefore it is time to reevaluate Koch's postulate which has served us well for over 100 years, in which you isolate an agent and introduce it to a test animal and if you reproduce the disease then you have the agent identified. It is not so simple now that we have new techniques, but I would continue to emphasize that we need to look at the output of these organisms in the gut with respect to metabolites that may have an influence on the central nervous system and the brain. By enterotyping both healthy Western subjects and healthy Indian subjects by mathematical techniques we can show quite significant differences based on diet.

I close with a quote from John Muir, the founder of the Sierra Club who said, "When one tugs at a single thing in nature, he (or she) finds it hitched to the rest of the universe." With the elegant presentations of databases that have been described, we now need to intercalate databases from many disciplines because, in the era of big data, we will find very interesting new ways of thinking. A point was made about collaboration. It has been pointed out that when there is a very diverse community, from many countries, there are diverse ideas, and you get more powerful output from scientific work done collaboratively. That has certainly been true with our team from India, Bangladesh, many countries in Europe and from many universities within the US. Science today is genuinely both international and interdisciplinary, and the presentation today has been very exciting. We are poised for a huge saltative leap in science and engineering today.

MICHAEL HUERTA

Speaker

Good afternoon. Today's biomedical research enterprise has a lot of data; you could say the NIH generates about $30 billion worth of data every year. These data are largely held in individual labs. There are notable exceptions like genomics but, by and large, few data are broadly available to the biomedical research community. In fact, the major public products of the enterprise are not data, but are instead concepts and ideas as described in scientific papers. If you think about the scientific process, the findings, interpretations and conclusions are the most fragile parts. Today, biomedical research is concept-centric, not data-centric.

In the biomedical research enterprise of tomorrow, I see increased data sharing that will make data broadly available. I see the use of standards that will make those data usable. I see data being brought into the research ecosystem by becoming discoverable and citable, with data that communicates with other data, software tools, and the scientific literature. Finally, I see advances in data science and science tools that will enable scientific innovation. In my view, tomorrow's enterprise will be data-centric.

How do we get from where we are today to where we need to be tomorrow? NIH has an initiative called Big Data to Knowledge, or BD2K. This will be a major factor in getting us there. It began with a working group on data and informatics of the advisory committee to the director of NIH. They issued a report in June 2012. Since then, many of us at NIH have been busy translating that report into this initiative.

BD2K is a significant, unique, and transformative initiative. I have been at NIH for 23 years and I have been involved in other transformative initiatives, some of which you have heard about earlier today. I think this is going to take the cake, though. It is significant because it's going to fund research, development and training in big data. But, that is what NIH often does when we identify a scientific priority. It is unique in that each and every institute and center is contributing funds, so everyone has an investment in this. When Dr. Collins, the director of NIH, brought this to the institute and center directors, there was a clear recognition of this area as important across all of NIH. Finally, BD2K will be transformative because, as you will see, this initiative will not just fund grants and support training but it will ultimately make biomedical research more data-centric.

I think of this initiative as having 3 major thrusts: the first is to advance the science and technology of biomedical big data, the second is to enhance and develop the work force in this area, and the third is to facilitate the broad use of biomedical research data. The intent of this last thrust is meant to apply across the board, for big data and small, alike.

In terms of advancing science and technology, there are initiatives underway already and there are others in planning stages as well. We will have centers of excellence to support data science research, tool development, engagement of the scientific community, and training opportunities. Research project awards will support the development of software tools and methods for big data, with an initial focus on data compression, visualization, provenance, and data wrangling. We are also looking at ways to expedite the wide use of large scale computing. We heard Nina

Preuss speak this morning about using the cloud with NITRC and imaging data. How can NIH, with thousands of investigators, make it easier for everyone to use these valuable resources? We have ongoing discussions with leading players in this area to figure this out.

In terms of the second thrust (to enhance and develop the workforce in biomedical big data), the idea is to aim this at undergraduate through senior investigators, to offer a variety of short and long-term training initiative and resources, and to emphasize interdisciplinary and team approaches. We are offering mentored career development awards that will train big data scientist in biomedical research and vice versa. We will be developing courses, again from undergraduate to senior investigator levels, to develop skills to use and analyze biomedical big data. We are going to support open educational resources so, for example, people who may be in a great computer science department not connected to biomedical research might have access to biomedical big data. In the next couple of months, we are also expecting to offer solicitations for institution training awards.

The third thrust does not have a big budget because it does not need one. This is the effort to facilitate the broad use of data. Specifically, we need to make data broadly available in the research ecosystem. One effort is to try to change policies, practices and the culture at NIH to increase data availability. When I say at NIH, I don't mean just the staff on campus, I mean all researchers supported by NIH. We developed recommendations that data management and sharing plans should be part of all requests fo research funding, whether it is for a contract or grant, big or small, extramural or intramural. If you are going to receive money from NIH, we would like you to have a data management and sharing plan. These plans would describe the standards you are going to use, the kind of data you will collect, how you plan to share it, what data repository would be used, etc. We also recommended that data management and sharing plans be peer reviewed and that the merit of the plans be reflected in the overall assessment of merit of the proposed project. In addition, we recommended that investigators provide information on the data sets that they are collecting and that they use existing standards and repositories when possible. These recommendations are well aligned with a subsequently issued (February 22, 2013) memo from the Office of Science and Technology Policy that asked for plans from major research funding agencies, including NIH, to provide plans to make results from federally funded research more widely accessible.

At NIH, making data more broadly available includes sensitive data from clinical research. More sensitive data – and perhaps more useful – are those not generated by research, but contained in electronic health records. With the advent and imminent ubiquity of electronic health records, such data represents the foundation for a whole new biomedical research paradigm. BD2K is now exploring how to make the most of this trove of important data which demands attention to ethical, legal and social issues.

We have heard a lot about standards today. You can make data available but if your data is, for example, in a format that does not match anyone else's, like the old days of PC vs. Mac and Betamax vs. VHS. Data must be able to work with other data, tools and resources, or it is not really usable. We talked about NIfTI-1 this morning. When we got that started, a workshop was held and we asked investigators about their thoughts on barriers to progress in neuroimaging studies. These brain researchers were performing research with functional magnetic resonance imaging (fMRI) but little new was coming out of it. After extensive discussion, it became clear that many diverse data formats were used in different labs. Since the data format determined the subsequent processing pipelines that could be used, and since the processing was so complex, relying on many assumptions that differed across pipelines, the use of different formats meant that the same study done in two different labs might lead to different conclusions (which could not practically be checked in other labs using other data formats) This led to a community-based standard for fMRI data.

Standards are crucial to data being broadly usable. Under BD2k, data-related standards will be promoted and encouraged. One initiative will be to make key information about widely used standards available to investigators via a standards information resource. This information will allow them to choose standards that will best serve their scientific needs (e.g., allowing them to use particular software tools, combine or compare their data with other data sets, or to deposit their data in particular repositories). The standards information resource will help investigators understand and discover which standards they should use, encourage the adoption of existing standards, and discourage people from re-inventing the wheel

We heard a lot this morning about the importance of community buy-in. Dr. McNutt talked about bringing people together to talk about social or behavioral science standards and it turns out that, at the NIH, while we support standardization efforts but there is no routine way to do it. If someone comes up to us and says “I have an experiment I want to do. My graduate student and I are going to work on these three hypotheses and it will take us about four years to do this research”, we say, “That sounds like an R01”. It is a mechanism, a set of policies, etc. We know how it is going to be reviewed, and we can give them the form to fill out. We have a number of different frameworks for supporting different kinds of research projects.

However, if someone comes to us, as they did many years ago, and says, “We need a common data format for fMRI data”, we would say “Let’s see what we can do”. This request could be handled in many different, mostly ad hoc and idiosyncratic ways, but we have no routine framework to support such efforts. BD2K will develop frameworks to provide catalytic support for particularly opportune community-based standards efforts. And, it will use those frameworks to support such activities that are broadly relevant to the NIH mission.

Finally, we need to bring data into the research ecosystem. This is crucial. The idea here is to catalog information about data sets. If someone publishes a paper, it would be helpful to have a minimal set of information about the data set serving as the basis of that paper. What would that be? It might include a descriptive title of the data set, a list of authors of the data set (who may or may not be the same authors of the paper) allowing those data authors to be cited and credited for their work. Other things you might want to include would be whether, when, where, and how the data will be available for data sharing. That might be a link to a database or it might be something simpler than that. Importantly, you will want to have a description of the data set. This might include the name of the organism from which data were collected, the type of data, the modality of the data, and further data elements for refined definitions. Information about the data set would then be cataloged, indexed, or a registered to allow people to search and discover data sets of interest to them Over time, you would have a compendium of information about data sets, perhaps ncluding what software was used to generate the data (not analyze, that would presumably be in the paper), what instruments were used to generate the data, and what standards were used that are related to the data. All of those kinds of important points could be in that compendium. Having that information there would allow you to index and search for whatever it is that you are interested in. And those capabilities could connect with resources such as Pubmed so as to relate the data to literature.

In terms of the impact of BD2K, the initiatives I have highlighted will advance the science and technology of biomedical big data, and bolster the expertise of biomedical researchers to use these approaches for scientific innovation. BD2K will make data broadly available, broadly usable, and will bring it into the ecosystem of research and scholarship, making the biomedical research enterprise more robust by making it more data-centric.

JUSTIN SANCHEZ

Speaker

I'd like to start off by saying thank you to the organizers for the invitation to be here because it really has been an exciting and fascinating meeting so far. One of the great benefits of meetings like this are the sidebar discussions that really addressing some of the tough issues we are talking about here today. I'd like to share with you a few remarks about this really challenging area. In 1945, Vannevar Bush was the director of the Office of Scientific Research and Development and he wrote a very important memo to President Truman tackling four extremely important points. He was addressing issues along the lines of what can be done to transition scientific knowledge gained during the war back to the public. He was also asking how we can better organize the fight against disease. He was asking what government can do to aid future research and how we can improve science, technology, engineering, and math education for the public benefit of the country. The recommendations Bush made resulted in a memo titled 'Science: The Endless Frontier'. This memo resounds with all of us. These recommendations connect with all of us who try to champion science and its importance in the public interest. Bush said, "Health, well-being, and security are proper concerns of government." Scientific process is and must be a vital interest to the government. And without scientific progress, the national health would deteriorate. Without scientific progress, we could not hope improvements in our standard of living and an increased number of jobs for our citizens. Without scientific progress, we could not have maintained our liberties against tyranny." What Bush is really talking about is a social contract for science. This is not just about what government should do for science, but how science benefits society. It's our obligations to deliver back to society. Bush also championed the concept of scientific publications resulting from these investments stating, "We should get scientific material to scientists everywhere with great promptness and at as low a price as is consistent with suitable format."

Nearly 70 years later, DARPA has taken up this banner with a renewed interest in administration and support. On top of broad agency announcements, we have added specific criteria directly addressing policies around data sharing. These policies reflect a belief that science conducted with the public dollar should be maximized for the public interest, especially around the most pressing challenges of our age. I manage a set of programs related to the brain and have had the great fortune of arriving at DARPA while we were gearing up for President Obama's great BRAIN Initiative. Currently, we have the expectation that scientists funded through these new DARPA awards also rapidly return their knowledge back to the community. We have direct conversations with investigators saying that it is their obligation to deliver this back to the community. This includes publishing in peer-reviewed articles with open access articles to expand the dissemination of knowledge. These are the same concepts that Bush was talking about. However, we want to go well beyond just scientific publication.

Let's talk about some of the issues, including the reproducibility issue in science. Many of us are familiar with the work of John P. A. Ioannidis, who published a very interesting paper titled, 'Why most research findings are false'. There are external sources of bias, small sample sizes, and poor standards of publishing. These are just some of the hurdles to overcome. We also know that science has a discoverability problem. Over a million and a half journal articles will be published this year across scientific disciplines. I read voraciously and even then it is extremely difficult to capture all of that knowledge. While methods for staying current on the literature are becoming more available, it's still easy to overlook transformative findings, the needle in a haystack. Science also has a transparency problem. Though many professional societies require authors to make their data available for publication of results in literature, many researchers do not comply with these requests.

Thinking toward the future, we are moving into an era of science where the challenges have expanded beyond and outside of the scope of the individual investigator. DARPA has continuously and consistently held the belief that teams from very different background can bring perspectives and fresh ideas to problems. Open science and data sharing offers a means for groups to systematically collaborate and pioneer new discoveries in how the brain operates, how we can harness new forms of energy, and how we can even explore the stars. At DARPA, for a small set of programs, we're hoping to address these problems in this very collaborative spirit. We anticipate creating a repository for scientific data collected in programs funded through this President's BRAIN Initiative. While many aspects are still under negotiations with our performers, and these are tough negotiations but we are up for the challenge, we would anticipate that the data resulting from publicly funded experiments might be made available through such repositories back to the community of interest within a year of publishing those experiments. I want to emphasize this very important point that this repository won't just be a collection of .csv files and MATLAB scripts, but will truly be a 21st century set of tools and techniques derived from the open access community. There is a huge opportunity here to really take the reigns and do something different. In a great spirit, we would like to work together with the publishing community and the scientific community to build a better awareness in this resource through special journal articles derived for secondary data users. We don't want to spend several years arguing over what metadata standards to adopt, but rather believe that annotation and data capture systems can lower this burden for the investigator to translate knowledge in their heads into bits and bytes able to be queried and called out by.

We don't anticipate doing all this alone. In a recent phone call Tom Kalil, the Deputy Director for the Office of Science and Technology Policy, referred to the initiative as an all-hands-on-deck moment. We all need to come together to make this happen. In the coming weeks and months we welcome and encourage all of your participation in this endeavor. Thinking back to 1945 and Bush's interesting, compelling, and transformative ideas he also noted that "Advances in science will also bring higher standards of living, will lead to the prevention or cure of diseases, will promote conservation of our limited natural resources and will assure means of defense against aggression. To achieve all of these objectives, to secure high level of employment, and to maintain a position of world leadership the flow of new scientific knowledge must be both continuous and substantial." This statement remains as true today as it does then. Thank you very much.

KRISTIN BRANSON

Speaker

My name is Kristin Branson and I'm a lab head at the Janelia Farms Research Campus. Janelia Farms is a Howard Hughes Research Institute that opened ten years ago with a primary focus in neuroscience. We work to develop imaging analysis technologies for novel neuroscience data collection. As a scientist, I focus on the research side more so than the policy side, so I am going to tell you about the type of data, the methods we are trying to share, and issues that we've been thinking about.

One of the large projects that I've been working on is called the Fly Olympiad. The goal is to understand the relationships between neural circuits and behavior. We are using genetic tools to target specific sets of neurons and activate those neurons using different transgenic lines. The goal of my research is to figure out how activating these neurons affect behavior. The flies in this video have specific subsets of neurons activated in the visual system and are trying to avoid each other.

Continuing what Dr. Liu was talking about, this research brings up the question, "What types of data should we share?" There is the raw data and then there is the processed data. The raw data in our experiments is video data. We put 10 male and 10 female flies in a bowl and we videotaped them. We did this for 2,000 types of genotypes of flies, each with different subsets of neurons activated. We collected about 20,000 videos that, in their raw form, take up about half a petabyte of data. We don't really want to share that raw data because it is something other people wouldn't be able to use. The raw data is very big, it's difficult to store, and it's difficult to transfer. We never actually store the uncompressed format; we do online compression of the data. This dataset is particularly difficult to visualize and interpret. There is nothing that a scientist can do with this data.

We've been working on ways to compress this data and highlight the important statistics. We want to collapse one video, or several videos describing 15 minutes in the life of these flies, into just a few statistics. For instance, this graph is showing how much of the time the flies spend chasing each other. We are comparing different flies with different genotypes to normal, wild type flies and we see that the flies in the boxed area court and chase each other more. This is the type of analyzed data we are trying to produce. One of the problems that we have is that we don't know what format to put the data into. We don't know what other behaviors flies perform, we don't know what other behaviors are interesting to scientists, and we don't have technology to automatically pull this information out of the video. One of the advantages of this dataset is that you now have a small number of statistics about the behavior of the flies that you can compare over 20,000 different genotypes of flies that you are looking at. We can start looking at what is common about these genotypes like whether they share certain neurons that were activated.

How do we get from this raw data to this analyzed data? There are analysis methods and, being a computer scientist, I mainly focus on these. The types of methods that we use to look at neuroscience data often require computer vision. We use tracking algorithms where take videos and estimate the position of the animals in every frame. We use machine learning methods in combination with human annotation to define behaviors like chasing. We use this machine learning to create classifiers that can predict whether the animal is chasing or not in each frame. We spend a lot of our time trying to figure out how to share methods like this with biologists. In computer science, you typically give someone your code and they can figure out how to run it. There is no graphic user interface associated with it. If you want biologists to adopt these tools, then you have to make it really easy to do this. A lot of biologists do not have a background in computer science and do not know how to program. We have to spend a lot of time trying to make our tools really usable. One of the things we focus on is how to make machine learning, which is cutting edge in the academic community, usable by someone without a PhD in computer science. There is more to behavior data than just the raw data. You start with the biological specimen, the sample preparation, and the data capture method. Behavior is a very fickle thing. If you change anything about how you prepare the animals, you will get a different result. For instance, we changed the incubator temperature by one degree and our flies became one and half times bigger in size and walked half as much. Small perturbations in the system can result in significant changes.

The types of information that we want to share are pretty common. There are several large data sets that are being collected at Janelia. One of these data sets is the ongoing fly connectome effort. Researchers are trying to take an electron microscopic (EM) image stack of the entire fly brain and trace all of the neurons and their synapses. There are also whole brain functional imaging projects where researchers are using calcium imaging to measure activity in all neurons in the zebrafish brain. One of the struggles inherent to this research is that there are somewhat limited incentives for enabling others to use ones' data methods. You have to be really devoted to wanting your work to be used by other people. You have to spend a lot of time developing GUIs and developing programs. One thing that HHMI allows us to do is to fund consultants who will help us support users who implement our tools. I've hired someone who answers emails about our tracking tools and does software management and updates. There are limited benefits for one's career in normal academic institutes. In my experience, it is the scientist's responsibility to develop the logistics of trying to prepare your code and your data to be shared as well as maintaining the dataset once it is shared. It would be really useful to have support for this. We think a lot about data storage, web servers for presenting your software, and data exploration and visualization tools. We have so much data now and when we want to describe a complex process, it is important to think about the best way to manipulate data that we are presenting to other people. I am excited about this because neuroscience is becoming a bigger science now. You need to have expertise in a lot of different fields to make progress. There is not just one field that you are a specialist in anymore; you have to combine fields. You have to know electrophysiology, you have to know how to perform genetic manipulations, and you have to know computer science to make a lot of progress.

PANEL 2 QUESTION & ANSWER

Jen Buss

Thank you Dr. Branson, for not being in public policy you did a really good job at framing policy questions out of your research. The first question I want to discuss is related to a lot of things that each of you touched on. Dr. Caldwell, one of the first things I noted was the cause and effect or the causation vs. correlation, and Dr. Sanchez then mentioned social responsibility and false publishing which begs the question, with whom should data be shared? If we are making this public then that means that the general public has access to this research as well. What are they going to do with it and how might it be framed in a way that we don't expect? What is that going to do to us as scientists? How is that going to hurt us potentially? Where do we draw the line and how do we determine who has access to this information?

Rita Colwell

Let me point out first that one component of society that we have not discussed is industry. Many industry players are beginning to recognize that these databases are valuable. There is a lot of work done to curate databases and eliminate errors. We have proprietary databases for the work that we do in Metagenomics, simply because there is a huge amount of work completed on eliminating errors in them. This is not necessarily criticism as to what data is stored in Genbank, but we have found that even the best laboratories that study X, Y, and Z bacteria, the DNA sequence for X has a little bit of Y and Z in it. A way around this problem is that a fee to use these databases could provide a source of income to people who curate these databases and ensure their quality. I am beginning to see a kind of shift: 25 or 30 years ago, maintaining databases in microbiology was a tedious job that was funded by NIH and NSF. The job was to keep cultures, data, and metadata, which was a labor of love. The databases got very large and cumbersome and a lot of data has been lost. The university holding the dataset without the individual curating it could not keep the data and turned it over to places like the American Type Culture collection, but even that has become a biotech company rather than a data collection service. It has been an interesting path over the past 25 or 30 years, but now it is well-understood that it is important to have good, reliable, publishable data.

Justin Sanchez
You asked who data should be shared with. The pursuit is to enable the path to discovery. This is about getting data to the people who can innovate. We need to get data to the people who can see the new discoveries and have the potential to advance the field forward. I think that is the fundamental charge here. Without a doubt, all of the annotation and data quality points are agreed requirements. The back end of your question is about the danger of open data. As with all things, there are secondary consequences to all of this. In light of our scientific mission and moving things forward in the best way possible, that may outweigh some of those other situations.

Michael Huerta
In terms of asking with whom we should share data, the notion that someone is going to misuse or misinterpret data is not an issue. You can bring up the same issue with scientific papers. People do this all the time. Obviously, with NIH, we have to consider patient privacy and so forth, which can be considered an issue of the type of data being shared. However, there is still a policy aspect to who can access this data. A non-scientist could apply to access NDAR data but they won't get it. I think an important question is what data should be shared. We do not often have a conversation about the realization that not all data are equal. It is true that we do not know how useful data might be in the future but we can't keep everything so we have to start making choices. These choices should be driven by assessing the likely purpose of a dataset. If data are going to be used primarily to understand a particular paper, you do not need a lot of standards and you can afford to be more idiosyncratic with proper justification. If you are going to use data to re-aggregate with other labs's data, standards are crucial. Experimental reproducibility will also require standards, as the lot number of the chemical used (not just the manufacturer) makes a huge difference. At the National Academy Journal summit, researchers brought up examples like this as reasons why entire studies could not be reproduced. We need to talk about this and make decisions on funding and otherwise: what are we going to use the data for?

Kristin Branson
In terms of data sharing as a researcher, I think there are two stages. One is when you are finished with a first publication on a dataset and another is when you are not done with it. I am hesitant to share my data before I have published something on it unless I know what they are going to do with it. After I am done my publication, I am happy to share it with anyone. If you are talking about a year between data collection and publication, it takes a year to analyze one's data.

Justin Sanchez
Let me connect back with your question. In academia, there really is not an incentive for sharing data because there is this need to get your work out and be the sole researcher to get a proprietary claim on something. How can we change that academic environment and create a set of incentives so that a researcher can establish his or her career but also think about the broader implications of the science that they are doing. I have spoken with others who have said that at multiple points in their career, they have intended to perform collaborative science and administrators think that they are really just trying to further their own reputation. The science gets clouded by these things. I think now is the time to start renovating some of those ideas.

Rita Colwell

One of the things that help is to be a little older and a little wiser. I can remember being a young investigator who was very concerned with the data we collected in fear of being "scooped". I learned that when you collaborate you get a whole lot more out of the data and the publications that come out of it. You will still want to be the first author but these aspects are negotiable. As a result, you get a string of publications that are more knowledge-based because you have a multi-dimensional, kaleidoscopic view of the ideas you are trying to put into the literature. I want to come back to Dr. Ascoli's talk: at the end, he left out the point that there is a genuine cost to maintaining data and curating it. If you are an investigator with a couple of post-docs, the time spent with this work is intensive, and maybe you have to hire a technician to carry out the work. I think there has to be a cost to accessing the data because a lot of work has gone into its placement in the database.

Heather Dean

I want to return to the question of what data should be shared. I don't know much about fly behavior but there could be a lot of data that you could never imagine to search for at first. One of the reasons Paul Allen was invited to speak at the first panel is that he has written about the fact that when data is shared amongst other fields, questions appear that were never imagined by the original investigators. In monkey electrophysiology research, I tied behavior to what I saw in local field potential and spiking activity. It could be that someone would find interesting data during the resting periods where the monkey was staring at a screen, grooming itself, etc. There could have been valuable information in there about activity in different layers of cortex. So I get a little bit nervous when we discuss having to ascertain a purpose to our data sets because we might not know as investigators.

Michael Huerta

You can't keep it all. It is not just something that just sits around like a piece of paper. If you have recordings on a tape from 30 years ago, the tape might be getting brittle and you might not have the equipment to play it back. We are talking about lots and lots of stuff because you did not do just that one experiment but if you are going to keep everything, you might spend the rest of your career collecting data from that one experiment.

Yuan Liu

I think it is a balance between Michael's comment and Heather's comment. I am a nature photographer and film is expensive, so you are careful about what you are photographing. With digital photography, you can use the camera like a machine gun and take thousands of pictures. It is not just the storage base; it is also the time you spend looking at the database. You can imagine two photographers: one deletes everything extra while the other keeps everything. I think there is a balance, where you need to be selective for your own research purposes but we showcase data like monkey's facial expressions and an economist can get something out of it.

Kristin Branson

One way would be to store the semi-raw data in a compressed format, where I've already thrown away what I am confident is not needed, just to have it. What I serve up on a web server is the fully processed form of the data.

Question:
Justin was getting at the question of tenure. As the only dean in the room, I agree with you. The way we reward faculty has to change over time. What we've tried to do at the Krasnow Institute for Advanced Study is actually change that equation so that the data sharing paradigm can thrive. It is thriving on the basis of faculty rewards.

Justin Sanchez
That is refreshing to hear. If we can use that as a model for other institutions, I would say we should try to promote these things.

Jen Buss
I had a question here regarding changing the culture of the faculty position to make data sharing as a problem obsolete. Faculty are so scared to share their data that it makes them more insular, which makes data sharing a cultural problem more so than a technological one. If we can change the attitudes, then we can alleviate some of the problems there.

Judy Kosovich
I do some work in the area of medical devices and so my question deals with abuse of science and data. When you get a device approved, you have to show substantial equivalence to an existing device. Increasingly, the FDA wants controlled studies. I worked on a device that simply measures resistance. Can we go to Radioshack and buy some resistors to show that the device is about as accurate as those? You need clinical trials and double-blind studies, which can be a waste of resources in this case. Anecdotal evidence also doesn't count, but you can gain ideas from anecdotal evidence. The legal implication is that your claims based on anecdotes are considered false. There needs to be a balance of what is considered convincing evidence and when you are dealing with individual health, there is a whole lot of variability.

Michael Huerta
You referred to a cookie-cutter approach in regulatory agencies. That is an important consideration and I tried to cover the notion of peer review of data management and sharing plans. The idea is not only to have these plans affect the score but also to bring community norms and expertise in the area to bear in deciding whether a data management or sharing plan is appropriate on a project-by-project basis. That gets to the earlier conversation about what data should be shared. If an investigator decides that he or she is going to share a specific data set, the reviewers might look at it and also suggest additional areas to share. There will be feedback on this point and in the context of the study. If you are doing a small, tech development project for the brain of zebrafish, you might not need to focus so much on the animal data because what you are doing is tweaking technology. There might be no use in sharing this data. The peer review aspect is important. Hopefully, NIH and other agencies will not take a cookie-cutter approach.

Giorgio Ascoli
I would like to follow up on Rita's point that there is a cost to data sharing. I would dare to disagree that there should be a cost to getting data from the databases. The issue is that would work for hypothesis-driven research. With a hypothesis, you buy some equipment and buy some data if you can. However, there is a whole lot of research where you don't know what you're looking for

and you are mining the data or are just attracted to the beauty of the dataset itself. You would not spend money to download the data just because you like it. As you mine the dataset, you might start to find interesting qualities. In journal publications, we have transitioned from a pay-per-read model to a pay-per-publish model. Right now the only users of data are humans, but increasingly, we are getting machines and artificial agents to mine data.

Rita Colwell
The collaborative use of shared data has been emphasized. If you are collaborating, you are using a currency, namely shared data and the data go back and forth. One of the biggest problems at NSF was trying to maintain funding for databases. Congress does not really understand why data storage would need funding. At least half a dozen meetings and conferences have been focusing on the topic of how to maintain funding for databases. Historical data often become hugely valuable. For example, a plankton repository in Southampton, holding a collection of plankton samples collected over the last 50 years is available to researchers internationally. When trying to understand processes like climate change, these collections and the databases are worth their weight in gold. This involves different methods for paying for the use of and access to such data. Industry is becoming much more involved in data management. Companies like Lockheed Martin and Northrup Grumman are involved in maintaining databases for the National Institutes of Health simply because the kinds of computations required are so complicated and there are lessons to be learned in computational ability from industry.

Giorgio Ascoli
I know that there is a program to maintain simulation and programming tools that has not extended to databases. Perhaps a way to let Congress know that there should be movement in this direction is to put them before the cost of not maintaining databases. If these databases went down, it would be a significant loss for the community and a monetary loss as well. If instead of just downloading data, we also had to start performing old experiments from scratch, we would be talking about orders of magnitude of cost increases.

Justin Sanchez
Instead of presenting that as here is what could happen as a warning, you can think about how to illustrate how data aggregation and secondary analysis enables you to see deeper to expand your horizons. Putting the positive perspective on this problem could be very effective. There are multiple examples that we have heard today in this vein and there are many more examples to come. The challenge to the community is to illustrate how these examples are useful and to create a roadmap for doing this process correctly.

Jen Buss
The positive spin is great, but progress is really only made when there is a dire need. If we told Congress that the entire VA database was down, chances are that changes would be made. One of the things we mentioned at lunch and Dr. Huerta did a good job with this in his presentation was that biomed/health data from private hospitals are not being aggregated well both in-house and for external sharing. The government has it pretty well figured out between Medicare and the VA.

Michael Huerta
We were talking about this in context of the ability of the same vendor of electronic health records having to customize their product significantly from hospital to hospital so much that there cannot be inter-hospital communication and data sharing. We talked about how the VA at least has a system where there can be these processes between hospitals internally.

Dave Clifford
We're talking about Vista, an open-source platform that the VA uses. The VA, Vista, and the Army records don't intercommunicate. They are secondarily not interoperable with any sort of research framework and they are not interoperable with Medicare claims. If you're looking at Vista, coming up with a clinical data warehouse is incredibly difficult. The standards adopted in electronic health records don't map to any standards of common data elements and interoperability at the NINDS. We talked a lot today about propagating standards and every organization that you give the opportunity to use a standard will create their own standard. Standards also produce technological lock-in: oceanographic standards used today were created in the 1970s. I don't necessarily view that as a good thing. Computers have computational elegance in certain tasks (e.g. using the minimum number of descriptors to describe a set of operating conditions). They are good at self-describing what they have. It is essentially all they can do. It is only when humans start ascribing labels to data that computers start to get confused. Every .csv file is interoperable with all other .csv files, but the data labels that we put on the files are not interoperable. If we think about systems that can self-describe their data (e.g. Javascript object notation, or DOI). This doesn't decide what is and what isn't recorded, but it provides an interoperable language that says a data element has certain properties and stores these properties. These concepts are something that computer scientists understand very well as compared to other scientists.

Michael Huerta
I was not saying earlier that the VA data sharing system is a great system as a whole, but rather I was talking about the success of inter-hospital communication at a basic level, which is a low bar to reach. At NIH and NLM, we are trying to increase coordination across all of the institutes, with common data elements for clinical research data. We have also put together shared data repositories that are fairly robust and can accept data from outside investigators. We have 55 biomedical databases at BMIC.

Question
Speaking from the perspective of a citizen, I want you all to tell the story that has been discussed today so that it gets written up in *Bloomberg Businessweek*, the *New York Times*, etc. The problems here are incredibly complicated and the general public does not have a clue. They probably have a set of expectations about research that do not pertain to reality whatsoever. I come from the standpoint that everyone who is educated in science has the responsibility to communicate their research at an 11-year-old level. Otherwise, people outside of the stovepipe won't be able to understand it.

Michael Huerta
The major, public product of science is not data. Maybe an 11-year-old wouldn't catch that point, but a 15-year-old would because they just assume this to be the case.

Greg Hale
I wanted to share my experiences with a couple pieces of software that get at the issues that we've described today. Github solves the issue originally raised about preventing people from messing with data. It also has a business model that addresses Rita's issue with charging users for databases. Github never took a government grant or VC funding. They provide a platform for controlling versions, branching, and experimentation in code. Linux was developed across multiple continents thanks to the power of this version control software. The platform that Github put together was so useful that companies that wanted to use it but not share their code were able to pay for all of the operating expenses and open source users could access the program for free. Travis CI is also a good program for making sure that for code that depends on other libraries, changes are made simultaneously at both ends. Travis is a system for simulating your code on evolving platforms and it automatically tells you when your code stops working. Travis solves this problem through virtualized machines. This addresses the issue Michael brought up of saving all data. These are some nice building blocks.

Michael Huerta
The issue to keep in mind is that biologists are not going to follow all of that and they are not interested in computer science because they went into biology. There are certainly solutions out there, and it is a cost-benefit analysis.

Question
When we started a few years ago with NITRC, a lot of the big labs had software that they developed and they did not want to share it because they said that other labs would use it wrong. Over the years, we convinced them that they do need to share the tools. The same thing happens with data, where researchers are concerned that their data will be mishandled and used to make incorrect conclusions. PubMed comments allow people to give feedback about these results. The culture might be changing: feel free to share, learn from sharing, etc. There are some really cool collaborations where institutions are not just holding hack-a-thons but are including crowdsourcing to take data from a variety of industry, science, and academia to come up with solutions. Share the data, share the software, and see what happens.

Rita Colwell
My team has been working on a set of algorithms for identifying microorganisms to species and strain. The team comprises microbiologists, but also software engineers, computer scientists, statisticians, and a cryptomathematician. The latter member, who does not know a lot about biology, often comes up with really snazzy programs, but what comes out may not have biological relevance. It is the interaction among all members of the team, with so many different perspectives, that has led to progress in finding microbiological solutions.

Justin Sanchez
DARPA often has challenges. There are a variety of opportunities and situations for new ideas from disparate areas of expertise.

Yuan Liu
Sharing and mining of data lead to new discoveries. The challenge over the next ten years is integration of data in two senses. We look at the brain at the molecular level, genetic level,

subcellular level, network level, and system level. In the past, we worked in silos. I worked in electrophysiology and Dr. Ascoli works in cellular data. How do we integrate all of that data obtained at different levels?

Justin Sanchez
All new neuroscience DARPA projects, by design, require fusing of hierarchical perspectives in the brain from multiple systems and connect all that information back to behavior. We do research for a purpose: restoring function in the brain. We put that end goal in mind and put the pieces together in a natural way to get at all of the research.

Yuan Liu
I think DARPA is a little bit ahead of us, and that's great! The second point I wanted to make was the example of inflammation. Inflammation could be a risk factor for cancer as well as a lot of neurodegenerative disorders, but there is little crosstalk between fields of research. Those kinds of integrations of data could really help science to advance. There are a lot of avenues to explore and we need guidance and encouragement to support them.

Rita Colwell
Dr. Branson's data provide a very useful example of what can be done by codifying complex behaviors and linking them to genomes. That's beautiful and where science is going, in a very big way.

Heather Dean
So, this is a panel on building the road forward and one of the big challenges that I've been thinking of is that science is increasingly international. Here we are thinking about the role of US tenure and promotion practices, US funding agencies, etc. but really there are more and more international collaborations. This morning, Dr. McNutt brought up the differences in data sharing policies in Africa, and the different needs there, where they have different sets of resources and different needs as they build their research programs. How do you deal with an international science and different policies and practices around the world.

Rita Colwell
One thing that is really important, in my experience working in many countries, mainly Bangladesh and India, is that NIH human subjects guidelines are adopted and followed in all of those countries. There have been instances where human studies were done in other countries and when results were submitted for publication (without following NIH guidelines) the papers were rejected. An international code of behavior and procedures for science is being created. It is happening slowly, but it is happening.

Mike Huerta
My office oversees international activities of the NLM (National Library of Medicine) and we have a lot of work going on in sub-Saharan Africa. There, our approach is to disseminate our best practices and to convey what is going on in the developed world, and that is a start. In all of these things that we do, we have to keep in mind how important it is to keep the silos down between disciplines and keep in mind how what we do is going to influence the geographically

dispersed community. In general, I don't think this is at the front of the minds of people at NIH but I think it is good to have that awareness raised.

Question – Yuan Liu

I actually happened to develop several bilateral initiatives in the past year. We established collaborations with Japan, China, and India. In any of these initiatives we make sure the data can be shared. Some sharing is challenging because some of the countries did not want to let the DNA out. But in neuroscience we have organizations like the INCF which is stationed in Sweden but we have member countries around the world, including the US. This organization's goal is to coordinate the policy of how to share neuroscience data. Several times a year, they have working groups on how to share data and providing enabling tools and they run workshops around the world to help us to do so.

CLOSING

PHILIP RUBIN

Closing Remarks

Good afternoon, everybody. I am going to be talking about the White House Neuroscience Initiative, the BRAIN initiative, and data sharing. The Obama administration has placed a strong emphasis on both ongoing and novel neuroscience, as well as related research efforts, under the auspices of the White House Neuroscience Initiative. This initiative encompasses neuroscience and mental health-related activities directed by the White House or supported by the White House Office of Science and Technology Policy (OSTP). I am the principal assistant director for science at OSTP and also lead the White House Neuroscience Initiative. This initiative seeks opportunities to build upon and coordinate across established efforts within the federal agencies by identifying strategic opportunities to work across agencies and promote collaboration between the federal government and the private sector.

The White House Neuroscience Initiative aims to increase the positive impact of federal investments in neuroscience to improve health, learning, and other outcomes of national importance. The White House Neuroscience Initiative includes or supports such activities as the interagency working group on neuroscience (IWGN), the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative, the National Alzheimer's Project, and other programs related to Post-Traumatic Stress Disorder (PTSD), traumatic brain injury (TBI), and mental health. With the encouragement of Congress, the IWGN was chartered by OSTP in June 2012. It was chartered under the National Science and Technology Council (NSTC) committee on science, which I co-chair. IWGN membership comprises more than 20 federal departments and agencies. Its mission is to enhance federal efforts related to improving understanding of learning and cognition, elucidating the causes and impacts of neurological disorders and injuries, and developing appropriate resources, tools, interventions, and therapies to assist in research, treatment, and recovery. Ongoing IWGN efforts coordinated by OSTP involve encouraging and supporting scientific research, sponsoring workshops to set forward-looking research agendas, developing and establishing common standards and guidelines, and sharing data and information. The IWGN recently released a report, available on both the NSTC and White House websites. This report identifies challenges and proposes recommendations in each of five areas of research, policy, and communication. Those areas include understanding and applying the brain's information processing capabilities, understanding and treating brain disorders and trauma, understanding and optimizing interactions between the environment and the brain across the lifespan, translating research to practice, and improving communication and engaging the public.

Let me talk about the BRAIN Initiative, which has gathered a lot of attention. On April 2, 2013, President Obama launched the BRAIN Initiative, which is a grand challenge designed to revolutionize our understanding of the human brain. Under this initiative, federal agencies such as the Defense Advanced Research Projects Agency (DARPA), the National Institutes of Health

(NIH), the National Science Foundation (NSF), and the Food and Drug Administration (FDA) are supporting the development and application of innovative new technologies that can create a dynamic understanding of brain function and its relationship to behavior. These scientific and technological advances could lead to improvements in our ability to diagnose, treat, and even prevent diseases of the brain. The President's 2015 budget proposes to double the federal investment in the BRAIN Initiative from about $100 million in fiscal year 2014 to approximately $200 million in fiscal year 2015. Given the audacious goals of the initiative, the president has called for this to be an "all hands on deck" effort involving not only the federal government, but also companies, health systems, patient advocacy organizations, philanthropists, state governments, research universities, private research institutes, scientific societies, and others. Later this year, the White House will hold an event to feature the role of these organizations in achieving the President's bold vision. At this White House event, we are looking for commitments such as research and shared facilities at universities and private research institutes; efforts by patient advocacy organizations to accelerate the development of diagnostics, treatments, and cures; information technology infrastructure to store, share, visualize, and analyze the huge volumes of data that will be generated; pre-competitive collaborations involving industry; education and training programs; regional clusters to accelerate economic growth; job creation and innovation; commercial neurotechnology domains; and well-designed incentive prizes.

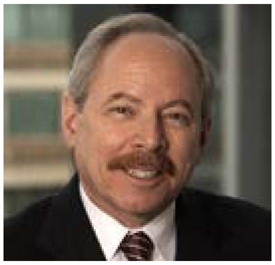
The Obama Administration is committed to the proposition that citizens deserve easy access to the results of the scientific research that their tax dollars have paid for. That is why, in a policy memorandum released on February 22, 2013, OSTP director John Holdren directed federal agencies with more than $100 million in research and development (R&D) investments to develop plans to make the published results of federally-funded research freely available to the public within 1 year of publication and require researchers to better account for and manage the digital data resulting from federally-funded scientific research. The final policy reflects substantial inputs from scientists and scientific organizations; publishers; members of congress; and members of the public, over 65,000 of who signed a We the People petition asking for expanded public access to the results of taxpayer-funded research. Since this announcement, the Obama Administration has expanded this effort to one that more fully embraces open data, open access, and open government.

Technology evolves rapidly, and it can be challenging for policy and its implementation to evolve at the same pace. In May, 2013, Obama launched the administration's new open data policy and released an executive order aimed at ensuring that data released by the government will be as accessible and useful as possible. The executive order of May 9 makes open and machine-readable the new default for all unrestricted government information. To make sure that this tech-focused policy can keep up with the speed of innovation, the administration also created Project Open Data, an online repository intended to foster collaboration and promote the continual improvement of the open data policy. Todd Park, the United States' Chief Technology Officer, and Steve VanRoekel, the United States' Chief Information Officer, indicated that "the administration wants to foster a culture change in government, where we embrace collaboration and where anyone can help us make open data work better." The project is published on GitHub, an open source platform that allows communities of developers to collaboratively share and enhance code. The resources and plug-and-play tools of Project Open Data can help accelerate the adoption of open data practices. The idea is that anyone from federal agencies to state and local governments to private citizens can freely use and adapt these open source tools, and that is exactly what is happening. In a memorandum released March 20, 2014, OSTP director John Holdren directed federal agencies to develop policies that will improve the management of and access to scientific collections that they own or support. According to Michael Stebbins and Erica Lieberman, who authored a White House blog on this announcement, "scientific collections are assemblies of physical objects that are valuable for research and education – including drilling cores from the ocean floor and glaciers, seeds, space rocks, cells, mineral samples, fossils and more. Federal agencies develop and maintain scientific collections as records of our past and investments in our future. These collections are public assets. They play an important role in promoting public health and safety, homeland security, trade, and economic development, medical research, resource management, education, and environmental monitoring."

Another concern is privacy and big data. On January 23, 2013, the President announced that he was appointing John Podesta to be Counselor to the President and head a fast-track 90 day review that is now ongoing examining the policy and privacy implications of big data. Podesta said, "we expect to deliver to the President a report that anticipates future technological trends and frames the key questions that the collection, availability, and use of big data raise, both for our government and the nation as a whole." On March 17, 2014, the Data and Society Research Institute, OSTP, and New York University's Information Law Institute co-hosted a public event entitled "The Social, Cultural, and Ethical Dimensions of Big Data." The purpose of this event was to convene key stakeholders and thought leaders from across academia, government, industry, and civil society to examine the social, cultural, and ethical implications of big data, with an eye to both the challenges and opportunities presented by the phenomena. This is one of a series of activities that are part of an ongoing effort by the Obama Administration to review the implications of collecting, analyzing, and using massive or complex data sets and the implications for privacy, the economy and public policy. The OSTP effort is being led by Nicole Wong, the Deputy Chief Technology Officer for the United States. The last topic I will cover is neuroethics. On February 10 and 11, 2014, the Presidential Commission for the Study of Bioethical Issues held a public meeting in Washington, D.C. on neuroscience and related ethical issues. The commission is an independent panel of experts that advises the president and the administration, and in so doing, educates the nation on bioethical issues. The meeting focused on President Obama's request that the bioethics commission examine the ethical implications of neuroscience research and the application of neuroscience research findings as part of the federal government's new BRAIN Initiative. The meeting was free and open to the public on a first come, first served basis, was live streamed and live blogged on the Bioethics Commission website at www.bioethics.gov.

We have a goal of fostering an integrative science of mind, brain, and behavior. We also face profound health challenges including over 200 neurodegenerative diseases such as Alzheimer's, Parkinson's, epilepsy, ALS, and many others. Other challenges facing our service members, veterans, and other citizens include stroke, paralysis, TBI, concussion, and a host of mental health issues including PTSD, depression, and suicidality. We believe that open data, open access, and open government are tools that help foster progress in many areas, including neuroscience. We have seen, in the case of GPS and weather data, that liberating government data can foster innovation. We also believe that public-private partnerships, while not a replacement for federal funding, can help to accelerate and advance federal endeavors. We believe that the BRAIN Initiative is a good example of that. Sharing of federal resources, including data, code, tools, etc. is an important component of this initiative. We are also exploring international cooperation. Thank you for your time.

PARTICIPANT BIOGRAPHIES



ALAN LESHNER

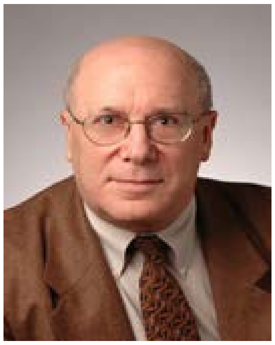
Chief Executive Officer of AAAS and Executive Publisher of *Science*

Dr. Leshner is the Chief Executive Officer of the American Association for the Advancement of Science (AAAS) and Executive Publisher of the journal *Science*. Before this position, Dr. Leshner was Director of the National Institute on Drug Abuse at the National Institutes of Health. He also served as Deputy Director and Acting Director of the National Institute of Mental Health, and in several roles at the National Science Foundation. Before joining the government, Dr. Leshner was Professor of Psychology at Bucknell University. Dr. Leshner is an elected fellow of AAAS, the American Academy of Arts and Sciences, the National Academy of Public Administration, and many other professional societies. He is a member and served on the governing Council of the Institute of Medicine of the National Academies of Science. He was appointed by President Bush to the National Science Board in 2004, and then reappointed by President Obama in 2011. Dr. Leshner received Ph.D. and M.S. degrees in physiological psychology from Rutgers University and an A.B. in psychology from Franklin and Marshall College. He has been awarded six honorary Doctor of Science degrees.

MICHAEL SWETNAM

CEO and Chairman of Potomac Institute for Policy Studies

Michael Swetnam assisted in founding the Potomac Institute for Policy Studies in 1994. The Potomac Institute for Policy Studies focuses on Science and Technology Policy. Since its inception, he has served as Chairman of the Board and currently serves as the Institute's Chief Executive Officer. He has authored and edited several books and articles including: *#CyberDoc, No Borders, No Boundries; Al-Qa'ida: Ten Years After 9/11 and Beyond; Cyber Terrorism and Information Warfare,* a four volume set he co-edited; *Usama bin Laden's al-Qaida: Profile of a Terrorist Network; ETA: Profile of a Terrorist Group;* and *Best Available Science: Its Evolution, Taxonomy, and Application*. Mr. Swetnam is currently a member of the Technical Advisory Group to the United States Senate Select Committee on Intelligence. In this capacity, he provides expert advice to the US Senate on the R&D investment strategy of the US Intelligence Community. He also served on the Defense Science Board (DSB) Task Force on Counterterrorism and the Task Force on Intelligence Support to the War on Terrorism. From 1990 to 1992, Mr. Swetnam served as a Special Consultant to President Bush's Foreign Intelligence Advisory Board (PFIAB) where he provided expert advice on Intelligence Community issues including budget, community architecture, and major programs. He also assisted in authoring the Board's assessment of Intelligence Community support to Desert Storm/Shield. He has served in several public and community positions including Northern United Kingdom Scout Master (1984-85); Chairman, Term limits Referendum Committee (1992-93); President (1993) of the Montgomery County Corporate Volunteer Council, Montgomery County Corporate Partnership for Managerial Excellence (1993); and the Maryland Business Roundtable (1993).



PHILIP RUBIN

Principal Assistant Director for Science, Office of Science and Technology Policy

Dr. Philip Rubin is the Principal Assistant Director for Science at the Office of Science and Technology Policy (OSTP) in the Executive Office of the President of the United States, where he also leads the White House Neuroscience Initiative. His responsibilities also include serving as the Assistant Director for Social, Behavioral, and Economic Sciences and serving as the co-chair of the National Science and Technology Council (NSTC) Committee on Science with Dr. Francis Collins of NIH and Dr. Cora Marrett of NSF. He is on leave as the CEO of Haskins Laboratories in New Haven, Connecticut, where he remains as a Senior Scientist, and is also a Professor Adjunct in the Department of Surgery at Yale School of Medicine and a Fellow of Yale's Trumbull College. Rubin is a cognitive scientist, technologist, and science administrator who for many years has been involved with issues of science advocacy, education, funding, and policy. His research spans a number of disciplines, combining computational, engineering, linguistic, physiological, and psychological approaches to study embodied cognition, most particularly the biological bases of speech and language. He is best known for his work on articulatory synthesis (computational modeling of the physiology and acoustics of speech production), speech perception, sinewave synthesis, signal processing, perceptual organization, and theoretical approaches and modeling of complex temporal events. From 2000-2003 Rubin was the Director of the Division of Behavioral and Cognitive Sciences at the National Science Foundation (NSF), where he helped launch the Cognitive Neuroscience, Human Origins (HOMINID), and other programs and was the first chair of the Human and Social Dynamics priority area. He was the NSF ex officio member of the National Research Protections Advisory Committee (NHRPAC) and the Secretary's Advisory Committee on Human Subjects Protections, both advisory to the Secretary of the Department of Health and Human Services, and was the Chair of the inter-agency NSTC Committee on Human Subjects Research Subcommittee (HSRS). From 2006-2011 he was the chair of the National Academies Board on Behavioral, Cognitive, and Sensory Sciences. He is also the former Chairman of the Board of the Discovery Museum and Planetarium in Bridgeport, Connecticut. Rubin is a Fellow of the American Association for the Advancement of Science, the Acoustical Society of America, the American Psychological Association (APA), the Association for Psychological Science, a Senior Member of the IEEE, and an elected member of the Psychonomic Society and Sigma Xi. In 2010 he received the APA's Meritorious Research Service Commendation "...for his outstanding contributions to psychological science through his service as a leader in research management and policy development at the national level."

JERRY SHEEHAN

Assistant Director for Policy Development, National Library of Medicine

Jerry Sheehan is Assistant Director for Policy Development at the National Library of Medicine where he is responsible for a range of issues related to scientific, technical, and medical data and information. He has played leadership roles in formulation and implementation of policies related to clinical trials registration and results information, genome-wide association studies, NIH data sharing policies, and the NIH Public Access Policy. Mr. Sheehan is active in NIH's Big Data To Knowledge (BD2K) initiative and co-leads its working group on research use of clinical data. He also manages the trans-NIH Biomedical Informatics Coordinating Committee (BMIC) and serves as chair of its Common Data Elements Working Group. Mr. Sheehan is currently the Deputy Chair of CENDI (the organization of federal science, technology, and medical information managers), vice president of the International Council for Scientific & Technical Information, and chair of the OECD Working Group on Innovation and Technology Policy. From 2011-2012, he served as co-chair of the Interagency Working Group on Digital Data, which informed development of the OSTP Memorandum on Increasing Access to the Results of Federally Funded Scientific Research. Prior to joining NLM, Mr. Sheehan was Principal Administrator and Senior Economist in the Science & Technology Policy Division of the Organization for Economic Cooperation and Development (OECD) in Paris. He previously worked in the Congressional Office of Technology Assessment and the Computer Science and Telecommunications Board of the National Research Council, where he directed NIH-sponsored reports on health informatics: *For the Record: Protecting Electronic Health Information* (1997), and *Networking Health: Prescriptions for the Internet* (2000). Mr. Sheehan holds a B.S. degree in Electrical Engineering and an M.S. degree in Technology and Policy, both from the Massachusetts Institute of Technology.

MARCIA MCNUTT

Editor-in-Chief, *Science*

Marcia McNutt is a geophysicist who became the 19th Editor-in-Chief of *Science* in June 2013. From 2009 to 2013, Dr. McNutt was the Director of the U.S. Geological Survey, which responded to a number of major disasters during her tenure, including the Deepwater Horizon oil spill. For her work to help contain that spill, Dr. McNutt was awarded the U.S. Coast Guard's Meritorious Service Medal. She is a fellow of AGU, the Geological Society of America, AAAS and the International Association of Geodesy. Her honors and awards include membership in the National Academy of Sciences, the American Philosophical Society and the American Academy of Arts and Sciences, as well as honorary doctoral degrees from Colorado College, the University of Minnesota, Monmouth University and the Colorado School of Mines. Dr. McNutt was awarded the Macelwane Medal by AGU in 1988 for research accomplishments by a young scientist and the Maurice Ewing Medal in 2007 for her significant contributions to deep-sea exploration.

YUAN LIU

Chief, Office of International Activities NIH/NINDS

Dr. Yuan Liu is the Chief of the Office of International Activities, and the Director of Computational Neuroscience and Neuroinformatics Program at the National Institute of Neurological Disorders and Stroke (NINDS), National Institutes of Health (NIH). Dr. Liu has been serving as the NINDS representative on more than a dozen international, interagency and trans-NIH committees and working groups that develop international and computational/informatics related initiatives and programs. Many of these programs and activities have relevance to bioinformatics and data sharing, such as the NSF-NIH joint program on Collaborative Research in Computational Neuroscience (CRCNS), the Interagency Modeling and Analysis Group (IMAG), the Neuroimaging Informatics Technology Initiative (NIfTI), the trans-NIH Biomedical Information Science and Technology (BISTI), the NSF-NIH BigData initiative, and the trans-NIH Big Data to Knowledge (BD2K). She is also a member of the project teams that oversee the Neuroscience Information Framework (NIF), the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC), and the Human Connectome Project, which are all supported by the NIH Blueprint for Neuroscience. Dr. Liu was constructive in the planning for the establishment of the International Neuroinformatics Coordinating Facility (INCF). Over the past two decades, Dr. Liu organized, chaired and co-chaired, and participated in numerous workshops and meetings in the field of computational biology and bioinformatics, and addressed challenges in data and tool sharing. She also co-authored position papers regarding bioinformatics and computational biology.

Dr. Liu received her bachelors and masters degrees in neurophysiology from Peking University in P.R. China, and her Ph.D.in neuroscience, under the mentorship of Prof. John G. Nicholls, from the Biozentrum, Universität Basel in Switzerland. Following her postdoctoral training at SUNY Stony Brook, she joined the intramural program at NIH. Her research career was focused on the area of neurophysiology at single channel, synaptic and circuit levels. Between 1999 and 2004, Dr. Liu managed a large research portfolio centered on channels, synapses and circuits grants at the NINDS. Prior to joining the NINDS, Dr. Liu was Program Director for Basic Neuroscience Research at the National Institute on Alcohol Abuse and Alcoholism, NIH.

NINA PREUSS

Senior IT & Scientific Program Manager, Turner Consulting Group

Nina Preuss serves as Program Manager for TCG's health care vertical. She is responsible for leading TCG's efforts to promote its competencies in big data, cloud computing, and collaboration environments. Funded by the NIH Blueprint for Neurosciences, Ms.Preuss leads NITRC.org, the "go to" collaboration environment for neuroscience researchers for software, big data, and cloud computing. NITRC comprises an online community, a "big data" federated image repository, and AWS Marketplace on-demand computational environment. To date, NITRC.org has been cited over 1,300 times in Google Scholar. NITRC was voted Overall Best by Excellence.Gov 2009 for transparency, use of technology, acquisition and internal processes.

Ms. Preuss has led a variety of other health care related federal projects including initiatives for the Executive Office of the President, National Institutes of Health's Office of the Director and its various Institutes.

Ms. Preuss holds a Masters of Business Administration from George Washington University and is certified by PMI, Institute as a Project Management Professional.

PAUL ALBERT

Branch Chief, Biostatistics and Bioinformatics at NIH

Paul Albert is senior investigator and Chief of the Biostatistics & Bioinformatics Branch at the NICHD.

His areas of expertise include the analysis of longitudinal data, biomarker data, and diagnostic testing. He is a fellow of the American Statistical Association and has over 25 years of collaborative experience at four different NIH institutes (NINDS, NHLBI, NCI, and now at NICHD). Professional awards include an NIH Merit Award for organization of an interdisciplinary team for methodological development in the design and analysis of biomarker studies in 2010. He has published over 250 papers in statistical, medical, and epidemiologic journals.

He received his A.B. in Mathematics and Psychobiology from Oberlin College and a Ph.D in Biostatistics at The Johns Hopkins University.

GIORGIO ASCOLI

Professor of Neuroscience, George Mason University

Giorgio A. Ascoli is University Professor in the Molecular Neuroscience Department and founding director of the Center for Neural Informatics at the Krasnow Institute for Advanced Study of George Mason University, where he has been since 1997.

Dr. Ascoli was born in Milan, Italy. After an education in the humanities and achieving top national youth ranking in competitive chess, he trained in Chemistry at the Scuola Normale Superiore of Pisa, and received a Ph.D. studying proteins involved in learning and neurodegeneration. Dr. Ascoli won the European Phillips Young Investigator Award in 1989 for the synthesis of a new organic molecule and moved to the National Institutes of Health in Bethesda, MD, in 1994.

Dr. Ascoli is internationally recognized in computational neuroanatomy, and edited the first book on this topic in 2002. He is a leading pioneer in neuroinformatics, and founding editor-in-chief of the premier journal in the field. Dr. Ascoli created and curates NeuroMorpho.Org, the largest collection of three-dimensional digital reconstructions of neurons. This free resource was accessed 50,000 times from hundreds of countries to download 2 million files in five years. Dr. Ascoli is also active in cognitive science and co-edited the book *Consciousness, Mind and Brain* in 2005. The original test he designed to quantify autobiographic memories (cramtest.info) was taken by more than 1,200 subjects who scored over 11,000 memories. Dr. Ascoli's 120 peer-reviewed publications were cited more than 1,000 times, and his work was presented at 350 conferences and invited talks and described in textbooks and national media. Dr. Ascoli serves on review panels for the National Institutes of Health, the National Science Foundation, and Intel Science Talent Search, national and international scientific advisory boards, and editorial boards of numerous biomedical journals. He received $10 million in grants from the US Departments of Health, Education, and Defense. Dr. Ascoli teaches graduate and undergraduate courses; his lab currently includes 5 postdocs, 8 doctoral students, and 10 interns.

Dr. Ascoli is married to Rebecca F. Goldin (a professor of Mathematics, also at George Mason University). They are proud parents of Benjamin (12), Ruben (10), Gabriel (6), and Jonah (4).

JENNIFER BUSS

Research Fellow, Director of Center for Neurotechnology Studies at PIPS

Dr. Jennifer Buss is a Research Fellow at Potomac Institute for Policy Studies. She is a member of the CEO's office and provides the scientific background for the think tank within the Potomac Institute, where she has been for two years. She is the Director of the Center for Neurotechnology Studies (CNS) at the Potomac Institute, having special interests in topics such as music and the brain as well as creativity and cognition. As Director of the CNS, she leads a team studying issues in neuroscience technology and policy and has been instrumental in organizing the Neuroscience Symposia Series 2014. Dr. Buss is a Fellow in the Center for Revolutionary Scientific Thought, a group at Potomac Institute that brings together individuals from a variety of backgrounds to foster discussion on science and technology futures from both an academic and policy perspective. In addition to these efforts, she has supported contracts for DMEA, OSD, and the Office of Corrosion Policy and Oversight. She is the Program Manager for the Rapid Reaction Technology Office contract for OSD in searching for innovative technologies to enhance government systems.

Dr. Jennifer Buss was awarded a doctorate in biochemistry from the University of Maryland Department of Chemistry and Biochemistry in 2012. Her dissertation was on iodide salvage in the thyroid and the evolution of halogen conservation in lower organisms. She performed graduate research in the areas of enzymology, bioinformatics, molecular and structural biology. Dr. Buss received her BS in biochemistry with a minor in mathematics from the University of Delaware. She is a member of the American Chemical Society, the American Association for the Advancement of Science and the American Society for Biochemistry and Molecular Biology.

RITA COLWELL

Distinguished University Professor, University of Maryland College Park and Johns Hopkins University Bloomberg School of Public Health; Senior Advisor and Chairman Emeritus, Canon U. S. Life Sciences; Chairman, CosmosID, Inc.

Dr. Rita Colwell is Distinguished University Professor both at the University of Maryland at College Park and at Johns Hopkins University Bloomberg School of Public Health, Senior Advisor and Chairman Emeritus, Canon US Life Sciences, Inc., and President and Chairman of CosmosID, Inc. Her interests are focused on global infectious diseases, water, and health, and she is currently developing 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. Dr. Colwell served as the 11th Director of the National Science Foundation, 1998-2004. In her capacity as NSF Director, she served as Co-chair of the Committee on Science of the National Science and Technology Council. Dr. Colwell 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. She is a nationally-respected scientist and educator, and has authored or co-authored 17 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.

Before going to NSF, Dr. Colwell was President of the University of Maryland Biotechnology Institute and Professor of Microbiology and Biotechnology at the University Maryland. She was also a member of the National Science Board from 1984 to 1990. 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 multiple scientific societies and has received numerous honorary degrees and prestigious awards.

Born in Beverly, Massachusetts, Dr. Colwell holds a B.S. in Bacteriology and an M.S. in Genetics, from Purdue University, and a Ph.D. in Oceanography from the University of Washington.

MICHAEL HUERTA

Associate Director of the US National Library of Medicine, NIH

Dr. Michael Huerta is Associate Director of the NLM and Director of the NLM's Office of Health Information Programs Development. His office coordinates efforts to make the NLM's considerable resources known to librarians, researchers, healthcare providers, and the general public; it oversees the Library's international efforts as well as NLM's evaluation and strategic planning activities.

Dr. Huerta's research background is in systems neuroscience; his undergraduate and doctoral work was completed at the University of Wisconsin at Madison, he was a postdoctoral fellow at Vanderbilt University and on the faculty of the University of Connecticut Health Center before joining NIH. Since 1991, Dr. Huerta has led several transformational efforts at NIH. These include promoting team & collaborative science through the: NIH Roadmap's Interdisciplinary Research Consortia, NIH Blueprint (http://neuroscienceblueprint.nih.gov/), and the NIH's adoption and mainstreaming of multiple principal investigators on individual projects. He has also led many informatics and data-intensive research initiatives, starting with the Human Brain Project, which helped develop the field of Neuroinformatics. More recently, he led the Human Connectome Project (http://www.humanconnectome.org/), which will provide comprehensive and systematic data about the connectivity of the human brain from some 1,200 healthy adults, and he directed the National Database for Autism Research (http://ndar.nih.gov/), which serves as a collaborative research platform and repository for data from nearly 100,000 subjects.

Today, Dr. Huerta is involved with a number of trans-NIH and trans-government efforts on standards, technologies, practices, and policies to more widely, efficiently, and meaningfully share biomedical research data. He serves on the NIH Steering Group on Public Access to Digital Scientific Data and is helping to lead the NIH Big Data to Knowledge (BD2K) initiative which will support research and development in data science and associated technologies (http://bd2k.nih.gov). Importantly, BD2K will also work to change policies and practices at NIH to raise the prominence of data in the biomedical research enterprise by increasing data sharing, supporting community-based standards efforts, and making data sets discoverable, citable, and linked to the scientific literature.

JUSTIN SANCHEZ

Program Manager, Defense Sciences Office, DARPA

Dr. Justin Sanchez joined DSO as a program manager in 2013. At DARPA, Dr. Sanchez will explore neurotechnology, brain science and systems neurobiology.

Before coming to DARPA, Dr. Sanchez was an Associate Professor of Biomedical Engineering and Neuroscience at the University of Miami, and a faculty member of the Miami Project to Cure Paralysis. He directed the Neuroprosthetics Research Group, where he oversaw development of neural-interface medical treatments and neurotechnology for treating paralysis and stroke, and for deep brain stimulation for movement disorders, Tourette's syndrome and Obsessive-Compulsive Disorder.

Dr. Sanchez has developed new methods for signal analysis and processing techniques for studying the unknown aspects of neural coding and functional neurophysiology. His experience covers *in vivo* electrophysiology for brain-machine interface design in animals and humans where he studied the activity of single neurons, local field potentials and electrocorticogram in the cerebral cortex and from deep brain structures of the motor and limbic system.

He is an elected member of the Administrative Committee of the IEEE Engineering in Medicine and Biology Society.

He has published more than 75 peer-reviewed papers, holds seven patents in neuroprosthetic design and authored a book on the design of brain-machine interfaces. He has served as a reviewer for the NIH Neurotechnology Study Section, DoD's Spinal Cord Injury Research Program and the Wellcome Trust, and as an associate editor of multiple journals of biomedical engineering and neurophysiology.

Dr. Sanchez holds Doctor of Philosophy and Master of Engineering degrees in Biomedical Engineering, and a Bachelor of Science degree in Engineering Science, all from the University of Florida.

KRISTIN BRANSON

Janelia Group Leader, Howard Hughes Medical Institute

Dr. Kristin Branson is a Group Leader at the Howard Hughes Medical Institute's Janelia Farm Research Campus. Her research involves developing machine vision and learning methods for quantitatively understanding animal behavior and its neural substrates. Previously, she was a postdoctoral researcher at the California Institute of Technology. She obtained her Ph.D. in Computer Science from U.C. San Diego, and her B.A. in Computer Science from Harvard University.

HEATHER DEAN

Symposium Organizer

Dr. Heather Dean is currently a AAAS Science and Technology Policy Fellow in the Directorate for Social, Behavioral, and Economic Sciences at the National Science Foundation. At NSF, she is working on big picture issues such as replicability of published scientific findings and broadening participation in science and technology fields. She founded a NeuroPolicy group and speaker series in Washington, DC that is building a neuroscience policy community. Dr. Dean is interested in issues related to cutting-edge interdisciplinary neuroscience, data sharing, science communication, new technologies in science education, and broadening participation.

Dr. Dean started out as an electrical engineering major at Caltech interested in neural networks and was soon exploring the biological side of such networks by studying locust olfaction with Dr. Gilles Laurent. She earned her Master's degree in Computation and Neural Systems along with her Bachelor's degree in Electrical Engineering. This research experience also set her on the path of neuroscience research, and she went on to earn her PhD in Neurobiology at Duke University, where she went into monkey electrophysiology with Dr. Michael Platt. After graduate school, she spent six years at New York University helping to found the lab of Dr. Bijan Pesaran and studying the neural circuitry underlying hand-eye coordination in monkeys.

Dr. Dean currently serves as President of the Caltech Alumni Association and has previously served on the Duke Alumni Association Board and the Duke Board of Trustees.

ABOUT THE SYMPOSIUM SPONSORS

The Potomac Institute for Policy Studies is an independent, 501(c)(3), not-for-profit public policy research institute. The Institute identifies and aggressively shepherds discussion on key science, technology, and national security issues facing our society.

The Institute hosts academic centers to study related policy issues through research, discussions, and forums. From these discussions and forums, we develop meaningful policy options and ensure their implementation at the intersection of business and government.

The Institute remains fiercely objective, owning no special allegiance to any single political party or private concern. With over nearly two decades of work on science and technology policy issues, the Potomac Institute has remained a leader in providing meaningful policy options for science and technology, national security, defense initiatives, and S&T forecasting.

The NeuroPolicy Affinity Group

The NeuroPolicy Affinity Group was established to connect and inform AAAS Science and Technology Policy Fellows who are working in or interested in learning about the intersection of neuroscience with policy, law, ethics, media, and society. The group has since expanded to include others from throughout government, industry, think tanks, and more. It is led by AAAS Policy Fellows Heather Dean, Dorothy Jones-Davis, Laurie Stepanek, and Tom Cheever.

NEUROSCIENCE AND DATA SHARING SYMPOSIUM REPORT

SCIENCE & TECHNOLOGY
POLICY FELLOWSHIPS
AAAS

Hosted by **The NeuroPolicy Affinity Group**

POTOMAC INSTITUTE FOR POLICY STUDIES