Welcome everyone to the webinar today. My name is Fen Zhao, Staff Associate, Strategic Innovation at NSF in the Directorate for Computer and Information Sciences and Engineering, Office of the Assistant Director. Today I'll be giving you some background on the New BD Spokes program.

Before we start, I want to make sure to acknowledge the whole team here at NSF that has been working on the BDHubs program, especially Al Suarez, Chaitan Baru, and Sylvia Spengler here in the NSF CISE Directorate. Chaitan is joining us today and will be able to answer your questions about the broader Big Data NSF CISE portfolio of programs at the Q&A section of this webinar.

I also want to thank my co-program officers for the BD Spokes solicitation: Heng Xu, from our Social, Behavioral, and Economics Directorate, and Earnestine Psalmonds, from our Education & Human Resources Directorate. In addition, I'd like to thank the other members of the BD Spokes working group, in particular Ken Whang, Martin Wiener, Anita Nickolich, Seta Bogosyan, and Lin He.

The agenda for our webinar today is three fold. I'll run start by running through some the historical timeline and background surrounding the BDHubs and BDSpokes program, setting some context on where this program sits within broader NSF and CISE efforts, as well as the interaction between the established BDHubs and the new BDSpokes to be funded.

Second, I'll try to explain some of the overall strategy and motivation behind this BD Spokes program and the major themes and topic areas of interest to NSF. The thrust behind the BD Spokes solicitation is accelerating Big Data applications, which topically has a very large scope. The thematic areas articulated in the solicitation help guide the proposer to understand the dimension where NSF feels innovation in Big Data could make the most impact.

Finally, I will discuss some of the logistical details related to the proposal and review process for the BD Spokes solicitation.

The BDHubs and Spokes comprise just one component of a broader Big Data portfolio of programs that NSF funds. Here at NSF we have a four-part framework of how we think about Big Data. One is the foundational research that's needed, which at NSF is funded primarily by the Critical Techniques and Technologies for Big Data program (often just called BIGDATA, all caps). This program currently has an open call and I encourage everyone to take a look. We also think about supporting infrastructure design and development, which is primarily supported by the Data Infrastructure Building Blocks or DIBBS program. We also fund education in Big Data through multiple programs, including, in the past the National Research Traineeship. Finally, on the partnerships and engagement quadrant of the portfolio, we have the BDHubs and BDSpokes program.

So as you listen through this webinar, you may find that the unique nature of the BD Spokes isn't a good fit for the project you have in mind. I encourage you to take a look at the other programs we mention today.

The BD Spokes solicitation is part of the Big Data Regional Innovation Hubs Program (also known as BD Hubs). The BD Hubs program was launched in March of 2015 with a call to develop 4 Big Data hubs around the country. NSF held 4 regional charrettes gathering stakeholders from each region, who collaborated to submit a single BD Hub proposal to NSF. The awards were made in Sept 2015.

Now the BD Spokes solicitation kicks off the second phase of the BD Hubs program. A 5th DC charrette was organized in early November gathering leadership from all 4 BD Hubs, where preliminary ideas for moving forward on the Hubs and in preparation for the BD Spokes was discussed. You will find links to artifacts from this meeting at https://bdhub.info.

A little more about the BDHubs. In September, NSF announced four awards totaling more than $5 million to establish regional hubs for data science innovation.

The consortia are coordinated by top data scientists at Columbia University (Northeast Hub), Georgia Institute of Technology and the University of North Carolina (South Hub), the University of Illinois at Urbana-Champaign (Midwest Hub) and the University of California, San Diego, the University of California, Berkeley, and the University of Washington (West Hub).

Covering all 50 states, they include commitments from more than 250 organizations--from universities and cities to foundations and Fortune 500 corporations--with the ability to expand further over time. The organizational structure for each BD Hub varies by region. Some are organized into standing sub committees and task forces, others into flexible working groups. You see on this map not only the Hub coordinators, but also the organizations leading the subcommittees and sub divisions of each Hub.

To help regions and stakeholders find each other to form these partnerships, the NSF funded BDHubs and BDSpokes aim to support the development of a coordination network. The top-level of this network contains the four geographically based hubs– the basis of what this webinar is about. One can think of each hub having a series of “spokes” which are targeted activities in local priority areas that the Hub wants to engage in– for example: transportation, manufacturing, or land use–effectively the Big Data application areas. These are the “spokes” of the hub. Each spoke could lead to multiple activities or partnerships. Not to belabor the terminology, but let’s call these nodes. These are defined as partnerships between two or more organizations that are geared towards driving some specific end goal in that spoke. We would hope that the hubs would be able to drive 20-30 of these partnership nodes during their first award period of 3 years as well as hr though this solicitation and future BD Spokes solicitations.

Each Spoke should be coordinating with a Big Data Hub to make sure that they align with the broader interests in the region and by its Big Data stakeholders. The regional BD Hub Steering Committee will provide general guidance to BD Spokes and assist with coordinating with the national BD Hub network, other BD Spokes, and the broader innovation ecosystem. We leave it to the spoke and the hub to articulate the proposal and the corresponding letter of collaboration, the details of how the Spoke and Hub will interact on a logistical, day-to-day level, and the details of how that collaboration will move forward.

As such, all proposals to this solicitation must include a letter of collaboration from a BD Hub coordinating institution. **Any proposals not including a letter of collaboration from a BD Hub coordinating institution will be returned without review. No exceptions will be made.**

Next I want to talk a little about the nature of the kind of proposals we would like to see submitted for the Big Data Spokes solicitation. BD Spokes are not typical R&D projects because of their focus on impact and partnerships across sectors and different types of organizations (academic, non profit, industry, etc). Nor are Spokes mini-Hubs. While Hubs are designed to be more exploratory and focused on ideation, full Spoke proposals are meant to have specific end goals in mind. (I will caveat this by saying there are 1 year 100K planning grants as well as full Spoke grants available through this call. Planning grants can be more exploratory. They are meant for collaborations that are of areas of interest to the region, but where it may be too early in the process for partners to articulate clear goals and timeline for a potential Spoke proposal. The planning grant offers resources for those partners to host workshops, feasibility studies, and other preparations for a more developed project. One goal for a planning grant may be to help design a submission to a later BD Spokes call, though that is not required; it may be that at the end of the planning grant, partners may find other NSF or federal agency program or other sources of support moving forward.)

The activities of a BD Spoke should address one or more of the following Big Data Innovation themes:

- *Accelerating progress towards societal grand challenges relevant to regional and national priority areas.*
- *Helping to automate the Big Data lifecycle.*
- *Enabling access to and increasing the use of important and valuable available data assets, also including international data sets, where relevant.*

Note that this list is not MECE, mutually exclusive or collectively exhaustive. These are simply a few critical dimension we find to be some priorities in accelerating innovation in Big Data. Any proposal may touch on more than one of these themes. There may also be Big Data proposals that make an important impact on data innovation in a way that this framework does not capture, however, for this competition, those projects would not be a good fit.

The first theme is "accelerating progress towards societal grand challenges relevant to regional and national priority areas". Due to the pervasiveness of Big Data in virtually all national priority areas, the BD Spokes have the opportunity to bring rapid change in application areas, by facilitating the creation of interdisciplinary and multidisciplinary data-intensive teams. Such areas broadly include healthcare, climate change, urban sciences and many more. These grand challenges may have widespread global implications, but be of specific interest to the region. For example, the South Hub has articulated an interest in their region on coastal hazards, in particular of the gulf coast. While projects on coastal hazards have global implications, the immediate output is also of interest to stakeholders in that region.

The second theme is "helping to automate the Big Data lifecycle." Managing the end-to-end lifecycle of Big Data assets can be a tedious and manual task. Steps in the data lifecycle include: ingestion, validation, curation, quality assessment, anonymization, publication, active data management, and analysis (including information extraction, visualization, and annotation). Automated (or, semi-automated) techniques are needed in order to keep up with the rapid data rates, large volumes, and immense heterogeneity of Big Data. Automation may also aid the reproducibility of data processing and analysis workflows. The data challenges and lessons learned by a BD Spoke on such automation efforts are expected to be shared with the BD Spoke's stakeholders as well as more broadly across the network of BD Hubs and Spokes.

The third theme is "enabling access to and increasing the use of important and valuable available data assets, also including international data sets, where relevant." Many valuable data sets are underutilized, and results from the analysis of such data are not shared, due to a variety of actual or perceived costs, including cost of curation, cost of data reuse, attribution and intellectual property considerations, etc. One of the desirable roles for a BD Spoke is as a catalyst for organizing and sharing datasets and related data services among a larger set of stakeholders, across disciplinary areas, within the geographic region, or across the national community. BD Spokes are expected to play an important role in supporting and promulgating open data and open source software policies within their projects—at the Hub-level, and across Spokes—to further facilitate the sharing of data and outcomes of analyses.

Broadly, for projects addressing any of the three themes, NSF recognizes that Big Data are global due to the way they are collected and analyzed and, hence, encourages international collaborations that will enhance the capacity and capabilities of the BD Hubs and Spokes.

BD Spokes proposals must articulate a clear focus within a specific Big Data topic or application area, while highlighting their Big Data Innovation theme. All BD Spokes must have clearly defined mission statements with goals and corresponding metrics of success. Some templates illustrating the specificity and level of detail for missions include:

- Use a specific set of analytical tools to improve the lead time for predictions of certain critical regional indicators by a given percentage.
- Given a specific set of high value data sets that were previously siloed and, therefore, usable only within a single research group or institution, make them available to a broader set of groups, or to the public at large, along with appropriate privacy and access control mechanisms.
- Adapt specified Big Data technologies to automate previously tedious and manual data collection and curation processes for specific types of data in a given field of science.
- For a specific genre of data, introduce new types of (automated) analytics—which were previously tedious to perform and manual in nature—that can be performed with minimal human intervention.

BD Spokes can initiate many different kinds of activities in support of their mission goals. The BD Spokes role is meant to convene stakeholders to augment and spawn new research efforts as opposed to directly carrying out traditional research. Potential activities for BD Spokes include, for example:

- Accelerating the ideation and development of Big Data solutions relevant to its mission by convening stakeholders across sectors (e.g., academic, industry, non-profits, etc.) to partner in results-driven programs and projects;
- Driving successful pilot programs by acting as a matchmaker between the various academic, industry, and community stakeholders;
- Engaging stakeholders across the region—including solution providers and end users—to enable dialogue, share best practices, and/or set standards for data access, data formats, metadata, etc.; and
- Connecting critical data resources to stakeholders that can best utilize them to fulfill the BD Spoke mission.

Note that BD Spokes funding from this solicitation is not intended to primarily support research activities. Rather, the goal of the program is to enhance and amplify collaborative efforts focused on achieving specific mission-based goals. For example, BD Spokes funding could support staff efforts in maintenance and/or improvement of existing data assets; integration of siloed datasets; analytics using existing high-value data assets, or of datasets made available via the efforts of a given BD Spoke; curation efforts; and workshops, travel, and other activities to support the collaborative and community-building nature of the BD Spokes. For specific questions about appropriate activities for BD spokes proposals, please reach out to me.

Proposed BD Spoke projects are expected to focus on their articulated regional challenges and opportunities. In addition, NSF has a working group of program officers from a number of different directorates and cross agency programs that are interested in solicitation additional proposals from the Big Data community that addresses some of their interests and priority areas. This includes:

- *Neuroscience*: Engage questions and opportunities in neuroscience that leverage BD Hub resources, such as enabling large scale, integrative modeling, sharing of diverse data and resources, and other neuroscience and neurotechnology approaches that require very large-scale, complex, or diverse data. Connections to other NSF programs on neuroscience research (www.nsf.gov/brain/) are welcomed.
- *Replicability and Reproducibility in Data Science*: Facilitate robust and reliable science by improving the replicability and reproducibility of research instruments, procedures, codes and results.
- *Smart and Connected Communities*: Stimulate innovative applications and services to enable more livable, workable, sustainable, and connected communities (www.nsf.gov/pubs/2015/nsf15120/nsf15120.jsp).
- *Data Privacy*: Ensure transparency by helping to identify when and how the data collected are being used; as sensor technologies collect more information, data is often reused and combined with other data in ways that go beyond the intent of the original collection.
- *Data Intensive Research in the Social, Behavioral, and Economic Sciences*: Accelerate research infrastructure and frameworks that integrate and operate on data from multiple sources including administrative data; scientific instruments from large-scale surveys, brain research, large-scale simulations, etc.; digitally-authored media, including text, images, audio, and emails; and streaming data from weblogs, videos, and financial/commercial transactions.
- *Education:* Support innovations in software infrastructure and use of data sets, and training, that facilitate research on STEM learning and learning environments, STEM workforce development, and broadening participation in STEM.

And on that note, I'd like to move onto some specifics of the solicitation that was posted as NSF document 16-510.

As I mentioned there are 2 types of awards, planning grants and Spoke projects. The planning grants are 1 year awards of $100K, and spokes grants are $1M (total) grants over 3 years. We anticipate a total funding amount of $10M for the program. This would come to approximately 9 spokes and 10 Planning Grants, though distribution would depend on quality and volume of proposal in these two categories.

The letter of intent is due on Jan 12, 2016. In the letter of intent, the proposer must have communicated with their collaborating hub and gotten preliminary approval to move forward. Each Hub has a distinct process for communicating with parties interested in designing spokes, which are detailed in each Hubs webpage. You'll find a link to each Hub's webpage on https://bdhubs.info. Please take a look as soon as possible; given the level of interest and timeline, each Hub has a tight set of deadlines for this process that will enable them to move forward efficiently.

The Full proposal deadline is Feb 25, 2016, 5pm proposers local time).

Proposals can be submitted by a wide number of organizations such as universities, colleges, non-profits, state and local governments. NSF welcomes collaborative proposals from for-profit organizations and FFRDCs as well, but those organization can only be subawardees and not the lead awardee. An individual can only serve as PI or Co-PI in at most one submission, but may participate as senior personnel in multiple proposals.

The review process for the solicitation will be by the standard NSF merit review process with a panel and additional ad-hoc reviews. As this is a unique solicitation, we've added a four additional review criteria to the standard intellectual merit and broader impacts criteria.

- How the Spoke will impact Big Data innovation and/or the application area of the BD Spoke's mission;
- Does the Spoke needs the support and infrastructure of the BD Hub network to carry out its mission and the effectiveness of its plans to coordinate with the host BD Hub;
- Does the Spoke have a feasible mission (given the Spoke's participants, planned activities, and management plan) that aligns with the relevant BD Hub's priorities; and
- Will the Spoke aid education and training of the Big Data workforce as well as related external groups such as end users, students or managers, and contribute to the education and training plan of the regional BD Hub.

To best help a proposal address these criteria, there are a series of mandated sections in the proposal that are outline in the solicitation. I won't go into detail for them on screen because they are different for planning grants and full proposals, but please take a look.

So both NSF and the Big Data Hubs are excited to see what your ideas are for Big Data Spokes.

Here is my contact information (fzhao@nsf.gov, 703-292-7344), and feel free to reach out to me if you have any questions. We'll open up the line to take some questions. I want to add that joining us to answer some questions are Chaitan Baru, the lead of the NSF BIGDATA program.

As we wait for questions to come in, here's a first one that I know is on many of your minds. This is one Chaitan Baru can best answer– "How do you see the Big Data Spokes plugging into the R&D NSF programs, such as Big Data?"