BD Hubs: Midwest: "SEEDCorn: Sustainable Enabling Environment for Data Collaboration"

Midwest Big Data Hub

Accelerating the Big Data Innovation Ecosystem

Submitted to the National Science Foundation in response to "Big Data Regional Innovation Hubs (BD Hubs): Accelerating the Big Data Innovation Ecosystem" (NSF 15-562)

Principal Investigator

Dr. Edward Seidel
Founder Professor of Physics, Professor of Astronomy
Director, National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign
Email: h-seidel@illinios.edu

CoPI: Brian Athey
Professor and Chair, Department of Computational Medicine and Bioinformatics
Co-Director, Michigan Institute for Data Science (MIDAS)
University of Michigan

CoPI: Sarah Nusser
Vice President for Research,
Iowa State University

CoPI: Beth Plale
Professor of Informatics and Computing,
Indiana University

CoPI: Joshua Riedy
Vice-Provost and Chief Strategy Officer,
University of North Dakota

A. Project Summary

BD Hubs: Midwest: SEEDCorn: Sustainable Enabling Environment for Data Collaboration, E. Seidel, UIUC. The nation faces increasing challenges in collecting, managing, serving, mining, and analyzing rapidly growing and increasingly complex data and information collections to create actionable knowledge and guide decision-making. All sectors of society are profoundly impacted and need novel solutions that leverage the breadth of expertise in academia, industry, and government. To address this need, a diverse and committed network of partners has created a nimble and flexible regional Midwest Big Data Hub (MBDH), responding to Big Data challenges, capturing special opportunities, interests and resources unique to the Midwest. Within MBDH, our proposed NSF *SEEDCorn* project will leverage partner activities and resources, building a sustainable framework to coordinate existing projects, initiate 20-30 new partnerships, start new pilots, and help acquire funding. It will develop and link collaborations, education, and services around data, involving diverse institutions (universities, non-profits, foundations, national labs, companies, government agencies) in the Midwest region and beyond.

MBDH has formed a distributed hub and governance structure. The hub supports activities that aggregate expertise, projects and resources, enabling communities to assemble and function along multiple *spokes*, including specific themes of importance to the Midwest (across three broad themes of society, natural/built environments, and biomedical sciences). Integrative *rings* connect all spokes and are organized around themes of data sciences, tools and services needed to collect, store, link, serve, and analyze complex data collections, and educational activities to advance the knowledge base and train a new workforce in the practice and use of data science and services. Groups across the region are naturally incentivized to work together as they all realize that the challenges they face are larger than any single group, institution, state, or region can adequately address alone.

SEEDCorn will leverage many existing projects represented by MBDH partners, funded by NSF, NIH, DOE, NIST, DOC, USDA, universities and the private sector. Outcomes will be multifold, including: (a) strengthening, creating and securing funding for 20-30 new public-private partnerships; (b) accelerating technology transfer projects; (c) introducing new Big Data educational activities into universities, industry and government, including data policies, management, social impacts and best practices; (d) starting pilots in a common data environment hosted by the National Data Service; and (e) developing and implementing new sustainability models.

Intellectual Merit. Collecting, harnessing, analyzing, managing, servicing, and sustaining large, complex data sources constitute grand challenges of our age. MBDH will address both the multidisciplinary challenges of creating and supporting collaborations around complex problems and the cyberinfrastructure challenges around creating data services to support them by creating a nimble, efficient and effective organizational and intellectual framework. Energizing MBDH, *SEEDCorn* will build vibrant intellectual partnerships along existing and future themes and will operationalize their complex research problems, exploiting a common platform (NDS Labs) for linking and creating data services. *Spokes* can be created as interest and opportunity evolves. Integration of diverse communities with common interests will aggregate data collections in broad themes, which are cross-connected to facilitate scientific and sharing policy discussions, as well as aligning educational interests. Bottom-up partnership building between academic, industry, nonprofit, and government organizations and individuals will be combined with rapid and responsive top-down big data and knowledge sharing, making MBDH responsive to emerging opportunities and changing conditions.

Broader Impacts. This is an unusually large and diverse consortium of partners, far exceeding usual NSF grants, built from the ground up. In addition to universities of all shapes and sizes across the region, the consortium is built to create and sustain academic-industry-government partnerships, with reach into all sectors of the Midwest. The project includes an experienced diversity coordinator to leverage relationships with national organizations, use social networking tools, and reach into public libraries. Keywords: **data communities and services, public-private partnerships, diverse populations**.

B. Table of Contents

C. Project Description

The nation faces increasing challenges in collecting, managing, serving, mining, storing, and analyzing rapidly growing and increasingly complex data and information collections in order to create actionable knowledge and to guide decision-making. All sectors of society are profoundly impacted and in need of novel solutions that leverage the breadth of expertise in academia, industry, and government at all levels, including setting policy and developing best practices. For example, important relationships between food, water, and energy need to be understood to assess their availability and safety, requiring integration of data across multiple sectors (e,g., agriculture, U.S. Army Corps of Engineers, industry, academia) and with multiple approaches to data science (e.g., genetic sequencing, GIS, simulation, machine learning).

Many data-intensive activities have sprung up around the nation and the world in the last decade, on campuses, in government, and throughout industry, with programs funded by many public and private organizations. While novel approaches are rapidly developing in many isolated settings, they fail to benefit from insights and tools created through other thematic and methodological activities (e.g., physics workflow planning could be applied in multi-sensor environmental data acquisition). In addition, some problems, such as statistical disclosure limitation, are pervasive and immense, requiring integration of numerous approaches and perspectives. A major need has therefore developed to create overarching hubs that provide structures to aggregate and organize activities, and importantly to build and support communities that cross all these sectors to harness the power and realize the promise of Big Data.

To address this need, the Universities of Illinois, Indiana, Michigan, North Dakota, and Iowa State have created a nimble and flexible regional Midwest Big Data Hub [1] (MBDH), with a network of diverse and committed regional supporting partners (including colleges, universities, and libraries; non-profit organizations; industry; city, state and federal government organizations; see *Partnerlist* supplement document) who bring data projects from multiple private, public, and government sources and funding agencies (including NSF, NIH, DOE, NIST, USDA, DOD and others). MBDH is a consortium to better enable the Midwest region and partners across the nation to respond to these Big Data challenges and to capture special opportunities, interests and resources unique to the Midwest. As a framework for developing and deeply linking collaborations, education, and services around data, MBDH was created to facilitate partnerships between diverse institutions. Operating within MBDH, our proposed *SEEDCorn* project will leverage partner activities and resources, coordinating existing projects, initiating new partnerships, sharing best practices and data policies, starting pilots, and helping to acquire funding. It will develop and link collaborations, education, and services around data, involving numerous institutions in the Midwest and beyond.

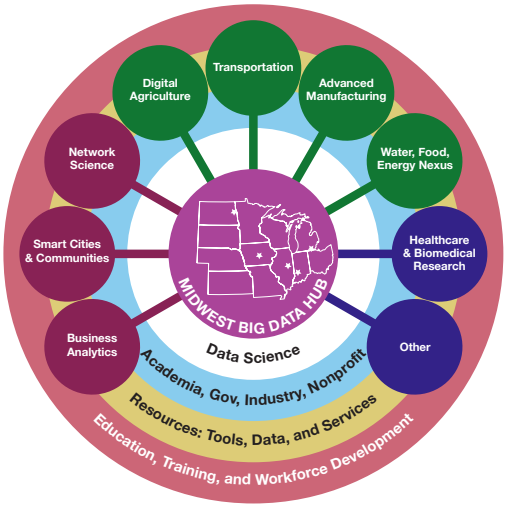

Figure 1. MBDH hub (center) linking thematic spokes of members across the Midwest. Cross-cutting rings provide additional connections between spokes.

As illustrated in Figure 1, MBDH lead institutions and partners in the consortium form a distributed hub that aggregates expertise, projects and resources from their members, enabling and supporting communities to assemble and function along multiple *spokes*, focusing on specific strengths and themes of importance to the Midwest across three sectors: ***Society*** (including smart cities and communities, network science, business analytics), ***Natural & Built World*** (including food, energy, water, digital agriculture, transportation, advanced manufacturing), and ***Healthcare and Biomedical Research*** (which spans patient care to genomics). At the same time, integrative "*rings*" connect all spokes and will be organized around themes of specific MBDH strengths, including (a) *Data Science*, where computational and statistical approaches can be developed and integrated with domain knowledge and societal considerations that support the underlying needs of "data to knowledge," (b) *services, infrastructure, and tools* needed to collect, store, link, serve, and analyze complex data collections, to support pilot projects, and ultimately provide production-level

data services across the hub, and (c) *educational activities* needed to advance the knowledge base and train a new generation of data science-enabled specialists and a more general workforce in the practice and use of data science and services. MBDH has many partners, with working groups led by individuals for each spoke and ring area, as described in the *Partnerlist*.

MBDH will support basic structures (communication, governance, administration, planning) and services (data service hosting, portals, development environments) needed for communities to grow and develop, to communicate with each other and with other hubs nationally. The *SEEDCorn* project will build on these existing structures to catalyze action and create new partnerships. MBDH partners have committed to build and support the hub, assembling many existing projects and funding sources from university, state, federal, and industry sources. The NSF-funded *SEEDCorn* project will leverage these and energize specific activities of the hub, contributing to a full-time executive director (at Illinois), and to part-time personnel at all five institutions (including a project manager and a technical support person at Illinois and local coordinators at all PI sites). A particular strength of MBDH is its existing array of data service projects and cyberinfrastructure offerings, which can be brought together through the National Data Service (NDS) Labs (C.4). NDS Labs will be used for creating data linking and pilot projects across MBDH and beyond, using *SEEDCorn* funds. Very importantly, *SEEDCorn* will fund the costs of travel and a number of workshops, hackathons, training and other events at our member institutions. A number of events are being planned, including a *SEEDCorn*-funded all-hub kickoff workshop in October 2015 at Illinois; each is designed to grow partnerships and catalyze additional actions, including the creation of new funding opportunities (C.8 has a timeline of activities). Hence, *SEEDCorn*-funded activities are front-loaded in year one (Y1) and ramp down as new funded activities are generated. A sustainability plan has been developed to grow funding sources as *SEEDCorn* ramps down, described in C.8. We will mix university, state, industry and agency-funded projects, with community-developed open source and fee-for-service models and will explore new economic models around data partnerships and licensing. By the end of the third year of the project we anticipate that funding from *SEEDCorn* will not be needed to sustain the MBDH.

Another specific strength of MBDH is its focus on education, training and diversity. The region is home to some of the nation's leading computer, informatics and computational science programs and major HPC centers (C.4), and we will combine existing regional activities with specific training programs and hackathons using NDS Labs, creating new data-intensive educational programs and best-practice guidelines that will benefit all hubs (C.3). On diversity, the region is rich in underserved populations, including African-American, Hispanic, and Native American populations. An experienced diversity leader (Franklin, see *Partnerlist*) leads a working group and activities for MBDH.

We have developed a dynamic governance model that allows the organization to support and grow initial partnerships and communities, as well as a process to support the creation of new communities over the lifetime of the project. As described in C.6, we have formed an interim steering council, with working groups around each existing spoke and ring, and developed a process for launching an elected Steering Council (SC) by March 1, 2016.

SEEDCorn will leverage and connect many existing partnerships across the Midwest and beyond, catalyzing many new partnerships. Specific outcomes of *SEEDCorn* will be many, including: (a) strengthening and creating numerous public-private partnerships, built on a stronger funding base, with new projects funded by multiple agencies, government organizations, and industry as a result of workshops and other *SEEDCorn* events; (b) projects will not operate in isolation, but will be better connected through MBDH; (c) new educational activities and best practices will be developed by our collaborations and introduced into the curricula of leading colleges and universities nationally; (d) *SEEDCorn*-supported pilots in a common data service environment through use of NDS Labs will lead to innovation, interlinking and acceleration of data services across dozens or potentially hundreds of projects; and (e) new business models for sustainable data solutions will be developed and implemented.

C.1. Initial Partners, Projects and Overall Collaboration Plan

Led by five universities geographically distributed across the Midwest, initial MBDH partners cover a wide range of institutions, including non-profits, over two dozen industry partners, city and state government organizations, national labs, small colleges, and rural and major state and private research universities. Many letters of collaboration have been collected, detailing some of the institutions' specific interests; additionally, initial leadership roles in key areas of importance have been defined. Operating within MBDH, the NSF *SEEDCorn* project will fund specific activities detailed below that are designed to (a) catalyze and build 20-30 sustainable interdisciplinary public-private data partnerships for research, education, and economic development, and (b) to support pilot projects that link existing data services and explore new ones across the region and with other Hubs.

Highlighting just a few institutions and their roles in leading specific partnerships, the University of Illinois at Urbana-Champaign will take the overall leadership role, hosting the executive director and project manager (both supported in part by *SEEDCorn*) to oversee and manage hub activities, strengthening interdisciplinary partnerships and building new ones, while also leading the National Data Service Labs environment that will support data-sharing projects and pilots (technical support provided in part by *SEEDCorn*). Indiana University will lead data and network science; Michigan will lead healthcare and biomedical research, transportation, and will coordinate business analytics with Wayne State; and Iowa State will lead digital agriculture. Non-profit UILabs in Chicago will lead industry partnerships and advanced manufacturing, while the University of Chicago will lead smart cities and communities. University of North Dakota leadership will be primarily in the area of cross-cutting activities involving Unmanned Aerial Systems (UAS) as they pertain to digital agriculture, transportation, and the food-energy-water nexus. The full list of existing partnerships, with specific leaders, is provided in the *Partnerlist*. Twelve working groups already operate and have produced white papers describing their initial activities, to be made available on the MBDH website (midwestbigdatahub.org).

Collaboration Plan: Building sustainable private-public partnerships is a key goal of MBDH, and key to success is collaboration through both electronic and personal interactions. To foster collaboration in a cost-effective way, we will emphasize teleconferences and web-based collaborative spaces and tools, including monthly meetings of the SC. We will also use web-based sharing of educational materials and programs. The MBDH will allow the Hub to engage with stakeholders, share information about best practices and data strategies, provide access to data tools and repositories, and importantly, provide access to a common, open data services development environment.

Personal engagement is needed for organizational strength and sustainability. The executive director and SC members will be supported by *SEEDCorn* to travel, coordinate and share knowledge among constituencies along several dimensions: connecting existing data projects (e.g. SEAD, DataONE, DataFour) related to spokes and rings that can be leveraged toward the Hub goals; support of regional events and meetings where MBDH initiatives can be presented and discussed; development of potential new spokes and rings as challenges and needs emerge; engagement of organizations like the Research Data Alliance (RDA) and NDS for broader impact and adoption of standards; coordination and communication with other regional Hubs. *SEEDCorn* will support a month of *local coordinators* at all co-PI sites for true engagement across all geographic regions of the MBDH. *SEEDCorn* will also support travel to key MBDH personnel to annual multi-hub meetings organized by NSF.

Very importantly, *SEEDCorn* funding has been designed to frontload workshop and event activities at all five MBDH co-PI sites, with a burst of activity in the first year (Y1) to ignite key hub and spoke collaborations and activities to start strong, with a specific emphasis on developing external funding streams. As external funds are generated for projects over time (from a variety of state and federal agencies, private foundations, and industry), a smaller number of *SEEDCorn*-funded events will occur in years two and three. A *SEEDCorn*-funded all-hands kickoff meeting at Illinois in fall 2015 will include an interim SC meeting, sustainability discussions, and initial spoke-oriented project planning. A second all-hands meeting will be held in spring 2016 with a fully elected SC operating by March 2016. A timeline

for *SEEDCorn*-funded workshops, hackathons, and other events is described in C.8, detailing the locations and approximate dates for each planned event. Additional workshops, funded by NSF during future BD Hubs phases, are anticipated as groups organize and seek additional funding.

MBDH Projects Already Planned with Funding from *SEEDCorn*: MBDH has developed over a dozen partnerships among its initial membership, with working groups already developing concepts for projects. C.5 describes the kinds of projects our current partners (see *Partnerlist*) are developing around eight *spokes* and connecting *rings*, around educational, industry, and sustainability activities. Here we highlight just a few, and discuss mechanisms for growing these activities through *SEEDCorn*.

Project 1 A partnership between MBDH and the National Data Service (NDS) (C.4, **Towns Letter**) will support federation of data services and pilot projects (**Norman, Krishnamurthy Letters)**. MBDH will work with the NDS Labs environment to assess needs of projects ongoing in our partnerships. *SEEDCorn* requests support for senior personnel to facilitate a pilot project that links various data services, providing common tools that may be used by others. Projects involved include three DIBBS projects (Illinois), SEAD (Michigan), Materials Data Service (University of Chicago) and others. NDS Labs will be a common hosting and development environment, accelerating collaboration and development of new projects. The ***Data Tools and Services Ring*** will have a *SEEDCorn*-funded workshop at Illinois (C.8).

Project 2 MBDH Digital Agriculture and related spoke and ring partners (See ***Natural & Built World Spokes***, C.5) will host a workshop at Iowa State (C.8) that brings together diverse stakeholders (producers, ag industry and commodity groups, researchers in agriculture, natural resources, rural sociology, engineering, and data science) to address the most pressing issues and opportunities in improving our capacity to integrate, protect, share, and analyze information that leads to actionable knowledge to support farm productivity, environmental sustainability, and rural well-being. Topics include precision agriculture, ecosystem services, related biosciences, and socio-economic impacts. The workshop will establish and extend partnerships to address specific issues or develop new approaches, tools, and resources that advance the resiliency and sustainability of agricultural and rural life.

Project 3 The ***Health & Biomedical Research (HBR) Spoke*** (C.5) will host a workshop at Michigan (C.8) on developing an open data translational biomedical research repository. The project will use the tranSMART platform, leverage the infrastructure of the Open Cloud Consortium (C.4), combining expertise of Michigan, the University of Chicago, and industry connections (pharma, biotech, nonprofit, academia) of the tranSMART Foundation (SC, **Letter**). The workshop will bring together interested parties from across the MBDH to define, fund, and ultimately implement a sharable HBR open data repository of use to MBDH participants and beyond to other NSF BD Hubs.

Developing New Big Data Partnerships: Within MBDH, *SEEDCorn*-funded activities will support members to build new partnerships, engaging state and local government and industry for support. Successfully engaging new partners requires understanding their needs, which can be many: conducting research and problem solving, developing workforce and talent, building infrastructure, accessing expertise, and sharing of knowledge and information. The expected benefits will depend on the type of partnership—university, industry, government, non-profit—and the goals of the partners. For example, university partners may see value in networking with industry partners through Hub activity, giving them an increased understanding of industry challenges and opportunities. Industry partners will gain access to skilled, experienced data-focused researchers, to human capital for workforce development, to cutting-edge discoveries, and will have influence on new curricula development. Governmental and non-profit partners will benefit from interactions with academia and industry, discussions of big data challenges and possible solutions, and training and internship opportunities. In addition, the Hub will provide shared benefits such as workshops, conferences, publications, and liaisons to other regional Hubs. We will work toward the stated goal of building 20-30 new partnerships catalyzed by *SEEDCorn* within MBDH.

Cross hub collaboration plan: Four proposed hubs (PIs: McKeown, Seidel, Norman, Aluru and Krishnamurthy, **Letters**) have agreed on a collaboration plan that includes development of joint

educational activities, workshops, data sharing and communication initiatives. We propose to share best practices and innovations in education and to coordinate a cross-hub workshop on education, including workforce development. For shared spoke topics, we will ensure that the overlapping spokes complement and support each other. For example, if a spoke from a particular region proposes a workshop at a conference, they will include spoke members from the other three hubs in the workshop design and promotion. Additionally, for regional in-hub meetings (e.g., the Northeast proposes to have these twice per year), each of the other three hubs will send a representative. To aid in cross-regional cooperation and transparent communication among the four hubs, we propose that a representative from each hub form a subcommittee to inform development of a shared, federated environment for data sharing, linking data services, exploring common file formats, and supporting pilot projects that span multiple hubs or may be national in scale; case studies of completed projects will be made available to all hubs. This committee will also inform the adaptation of the hubs to progress in Big Data and the developing interests of hub partners. Finally, the four hub leadership teams will have quarterly phone meetings with one another in addition to discussions at the NSF annual meeting.

C.2. Sustainability

Sustainability has been considered from MBDH's inception, designed as an ecosystem of multiple approaches to business models involving financial and in-kind support of universities and non-profits, fees for private industry, contracts with government organizations, and grants from foundations and traditional agencies (e.g., NSF, NIH). Data services will be supported via mechanisms in which MBDH partners have significant experience, including open-source community development, software as a service (e.g., as Globus operates), and possibly commercialization of services. *SEEDCorn* is designed to catalyze MBDH projects that become self-sustaining (via university, state, non-profit, agency, and/or private industry funds). While funding from sponsoring agencies will be sought (e.g., NSF funding for spokes), MBDH projects are envisioned to operate after *SEEDCorn* funding ends.

Sustainability is a key focus of the SC, which will deliver a detailed sustainability plan by Year 2. A representative from industry leads an SC group on this topic (C.6 and *Partnerlist*). Success will depend in part on MBDH offerings matching partner needs. Members must see value in the partnership, which can be expressed as willingness to pay membership dues and/or to invest in-kind time and energy. Incentives include (1) input and access to workforce development, (2) access to infrastructure and/or expertise, and (3) sharing knowledge, information and data.

MBDH brings significant experience with industry collaborations. Following lessons learned over three decades since the national supercomputing centers were launched and private sector programs were created, we will develop public-private partnerships around data in stages. MBDH has undertaken an initial inventory of numerous partner relationships operating around our region (e.g., UI Labs, NCSA's Private Sector Program, others). In initial discussions, companies have shown great interest in being among the first to join MBDH, and over two dozen companies have signed letters of collaboration outlining areas of interest and detailing "in kind" commitments of staff time to work with MBDH.

We will build on momentum to have other companies join our efforts, further defining interests. Working groups will be formed around common topics, sharing best practices in big data applied to many business sectors, and determining what kinds of pilot projects might be carried out. The pilots may include developing data sharing projects with university research groups, providing advanced data management technologies to companies, and applying advanced business analytics techniques.

The final stage will involve implementing business models for sustaining MBDH activities around specific services and resources, building on successful aspects of existing private sector programs (in many cases with the same companies) that have sustained themselves for decades.

These stages have been designed to scale to partnerships on the state and local level. The MBDH has numerous partnerships previously started by members of the PI team and the SC; we will further develop such partnerships during the life of this award as they contribute to the sustainability of Hub activities.

C.3. Education

Current high demand for big data professionals across industry, academia and government is substantially growing, leading to a critical need for well-trained *data scientists*. McKinsey predicts a shortage of nearly 200,000 professionals in data science and data related jobs in the US [2] and—despite a high median salary [3] of over $117,500—the job market supply will not fill these positions [4] [5].

The MBDH will work to address the challenges of providing effective instruction in big data and satisfying the current and future nation-wide shortage of big data experts across academia, industry, government and non-profits. The goal of the MBDH education ring is to impact Midwest big data sectors through training, outreach and extension, providing coordination, resources and tools to develop a skilled workforce, and to help train the existing workforce in solving real-world big data problems.

While most big data research activities take place at large research-intensive (R1) universities, smaller institutes of higher education carry the responsibility of training the majority of the IT and data science workforce. Therefore, our educational efforts will be geared toward the involvement of small universities (represented in the Steering Council; see *Partnerlist*). The MBDH will provide a unique opportunity for smaller universities to be involved with big data education and research and will expose students in computer science, statistics, mathematics, and in related fields to big data training aimed at developing practical solutions to real-world problems. That unique and critically important experience will be pivotal in the preparation of trainees for becoming big data experts in the present and future job market.

The MBDH will address these needs and will optimize the use of resources by facilitating collaboration between small and large institutes and with industry and government sectors that can offer applications for project-based instruction. Data brought to the attention of the hub, such as transportation or smart city data, will allow students from all institutions, including small colleges, to design and implement solutions using actual problems and real-world data. That will also provide institutions with access to domain scientists who contributed the data or study them, and we fully anticipate these interactions will help MBDH develop continuing education opportunities that will enable the current workforce to expand their knowledge and skills. Because the hub will share computing facilities, underfunded universities and industry will be able to take advantage of available cycles at research universities' computing facilities.

SEEDCorn will coordinate to develop a portal for education, training, outreach and diversity that will integrate information on resources, materials, courses and opportunities related to data sciences. The portal will facilitate partnerships within the hub and across hubs, e.g. linking organizations needing solutions to student engagement to integrate student learning in data science and application fields via a mentored practical experience (e.g., live cases, capstone projects) and advertising tools and services for education (e.g. providing real-world data sets, allowing students to share software solutions). We have identified Steering Council roles to be taken on by Franklin (Senior Personnel (SP)) for diversity and Shamir (SP) for small colleges (See *Partnerlist*).

SEEDCorn will collect information on big data and data science learning options in the region. This will allow sharing of information and options for collaboration as the region's academic institutions build undergraduate, graduate, and continued education curricula, including online and hybrid options such as the specialization in data curation offered by the School of Library and Information Science at Illinois.

In the above activities, we will work with NDS (C.4) to facilitate workshops, training sessions, data and software carpentry, and hackathons and datathons, and to highlight numerous data services and how they can be used in research and education, as well as with numerous other projects ongoing in the Hub.

C.4. Resources

Curated datasets and computational tools to process the data are at the heart of MBDH. Datasets, when properly curated and annotated, serve as a foundation of new economic development and the basis for educational projects and new tools development. The ***Data Sciences Ring*** (Co-PI Plale) will work with the data providers on behalf of the MBDH collective to create agreements and recommend tooling (e.g.,

similar to DataONE federation protocols) that allow the data sets to be as open and available as possible to the MBDH community. ***The Data Sciences Ring*** will strive to harmonize access to MBDH assets.

Numerous data collections have been offered for inclusion in MBDH. The datasets with links to MBDH span from genomics, the earth and atmosphere, materials, texts from great works of science and literature, and social science. We are working with providers to obtain data on transportation and smart cities. Data collections often come with restrictions on their use. The Midwest is host to a number of projects that provide persistent services for interaction with and analysis of sensitive data. Built around an important data set, these services support prior work of data extraction, data cleaning, and data synthesis that takes place before and during analysis itself. Analysis services that provide secure computation on a protected collection are often the only access that is available to the collection. The data collections and service environments already identified are listed here. The Data Science ring will focus its attention on secure computational environments and will identify common services that could benefit MBDH members in the spokes, including certifying repositories as trusted repositories (e.g., Data Seal of Approval). A sampling of MBDH data collections includes:

Genomic Data Genomic data on tumor types from more than 10,000 patients, stored and harmonized by the Genomic Data Commons to advance and transform the study of cancer. *The National Center for Genome Analysis Support (NCGAS)* provides services for analysis of genomic information (transcriptome and genome assembly, phylogenetics, metagenomics/transcriptomics and community genomics).

Materials Research Data *The Materials Data Facility* was established recently to serve as a repository for preservation and sharing of materials research data from both simulations and experiments.

Social Science Data *Inter-University Consortium for Political and Social Research* maintains an archive of more than 500,000 documents in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields. Restricted-use demographic, economic and health microdata are available through the *Central Plains Research Data Center Bureau of the Census*. *Terra Populus* [6] (DataNet) integrates world population and environmental data, including surveys, land cover information from remote sensing, climate records, and land use from statistical agencies. TerraPop data, interoperable across time, space, and scientific domain, inform the dramatic transformation of the earth's inhabitants and their environment. *Iowa State, Michigan, Illinois, Chicago, and Wisconsin survey research programs* offer methodological expertise for household and land-based surveys based on emerging data collection environments and integration of complex survey, administrative and geospatial data resources. Illinois' *Cline Center for Democracy* [7] has an unmatched collection of data, information and millions of documents from media such as The New York Times.

Atmospheric and Earth Surface Data *USGS Earth Resources and Science Center* holds the world's largest civilian collection of images of the Earth's surface, including satellite images, aerial photography, elevation and land cover datasets, and digitized maps. The archive spans from old (1937) aerial photographs to millions of satellite images of the Earth's surface, starting with the original Earth orbits in the 1960's and first Landsat satellite in 1972, to current hourly additions of satellite images. *NOAA observational data*: NOAA gathers 20 terabytes of observational data every day. A real-time copy of this data will soon be accessible at the Open Cloud Consortium (OCC) in Chicago, augmented with curating and management services. *Polar Geospatial Center* [8] (PGC), an NSF-funded research organization supporting polar science and operations, holds an extensive collection of satellite imagery and aerial photography at varying resolutions, including those from the Alaskan High Altitude Aerial Photography (AHAP) Program, Antarctic TMA Aerial Photographs, Landsat and MODIS imagery, among others. In collaboration with the USGS's Antarctic Resource Center, the PGC holds and digitally preserves the entire reconnaissance mapping series and satellite maps to support polar science and operations.

Digital Humanities Data The *HathiTrust Research Center* provides digital text analysis services on copyrighted data. A secure environment for digital humanities analysis, it will soon have 13.4M digitized books (4.7 billion pages) of the HathiTrust Digital Library (62% protected under copyright).

Core Data Services MBDH has among its membership three research funded prototype services that provide data curation and publishing services. MBDH will be able to leverage and use these technologies but also provide to these projects new opportunities to apply those technologies in other domains. *SEAD* (DataNet; Michigan, Indiana, Illinois) is an NSF-sponsored project to create data services designed to meet the needs of sustainability science research, offering a controlled and simple workflow for publishing complex and simple data sets to an array of back-end repositories and storage servers. *Through the Timely and Trusted Curation and Coordination* framework and system, materials-to-devices digital data can be captured, curated, correlated and coordinated in a real-time and trusted manner before fully archiving and publishing the data for wide access and sharing. The NSF-funded *Kurator* project (Illinois and Harvard) is developing tools for automating data curation workflows, focusing on biodiversity data and specimen collection data from natural history museums.

Core data services of MBDH also include identity, profile and group management, third-party data movement, and extraction and conversion tools. *Globus Nexus* provides identity, profile, and group management as a service. It enables users to create a unique identity that can be used across services and allows the creation and management of groups. High-performance, secure, third-party data movement and synchronization between endpoints, provided by Globus Transfer, is critical for moving large amounts of data. Extraction and conversion tools integrated into *Brown Dog*, a framework for plugging in data extraction and conversion tools to facilitate data interoperability, are available. Tools for synthesizing large-scale spatial data are anticipated through two projects: *CyberGIS* [9] and *SpatialHadoop* [10].

Computational, Data, and Visualization Resources The *National Data Service (NDS)* is operated by a consortium of universities, HPC centers, libraries, funded projects, and publishers across all four NSF hub regions that is developing and linking community data sharing services (NDS Share), as well as an open community development environment (NDS Labs) where groups can work together to extend and link existing core services (e.g., those above) and pilot new ones. Primary MBDH partners include Illinois, Chicago, Michigan, and Notre Dame, with other key partners at SDSC (Western Hub), TACC (Southern Hub), and Harvard (Eastern Hub). NDS Labs is a common development platform to support data sharing, linking services, and pilot projects within and across hubs. For example, DataONE (led in the Western Hub) will work with NDS to expose its data discovery services; *iRODS* (led in the Southern Hub; **Letter**) will be available in NDS for data federation. Such collaborations will lead to cross-linking of activities at both scientific and data service levels. The MBDH and the Western Hub (led by NDS partner SDSC; **Letter**) will jointly provide leadership in supporting this environment for all four regional hubs.

The Midwest has numerous large-scale HPC centers: NCSA, Indiana, Minnesota Supercomputing Institute, Iowa State, Nebraska-Lincoln, Argonne, and the Open Science Grid with major anchor points in Wisconsin, Indiana, and Nebraska. Regional/national optical network organizations provide the best optical network footprint in the nation, anchored by NSF StarLight at Northwestern, MREN (Midwest Research and Education Network), Great Plains Network, and many more. Experimental facilities also generate data (FermiLab, the Advanced Photon Source) and carry out visualization (UIC's Electronic Visualization Laboratory, Iowa State's Virtual Reality Applications Center).

C.5. Spokes and Rings

MBDH has already developed activities around eight thematic *spokes*, three cross-connecting *rings* (including education), and industry, shown in Figure 1 (see *Partnerlist*; governance in C.6). Leaders of each activity chair working groups that have developed concept papers with initial plans. Space does not permit a full description of all these activities but the MBDH website has been created with contact information for each activity; as documents are refined they will be made available publicly and updated as they develop. In this section, we have combined these rings into three general overlapping areas, *Society*, *Natural & Built World*, and *Healthcare & Biomedical Research (HBR),* as well as one cross-cutting ring area on *Data Sciences and Services*. We describe the challenges, aspects that are unique to the MBDH, and the initial plans to grow the activities.

Society Spokes Globalization and technological, demographic and environmental changes present enormous challenges in the sustainability, resilience and health of modern societies. Although a vast array of spatiotemporal data are generated by cities, communities, governments, businesses and citizens (e.g., public transportation, traffic, vehicles, cadastral information, utilities, law enforcement, sales, inventories, supply chains, personal smart phones), effectively addressing these challenges will information systems that represent interconnections and interdependencies within the societal system, enabling the shift in focus from objects to interactions. At present, we lack the partnerships, data infrastructure and knowledge required to fully harness the potential that lies with seamlessly integrating available and emerging data resources to respond to critical short- and long-term forces arising in numerous societal contexts. To fully capitalize on these opportunities, people and organizations need (1) access and search capabilities across data within and across entities; (2) standards for defining, organizing, managing, and connecting data with specific contexts; (3) methods for heterogeneous spatial and temporal reference systems; (4) appropriate analytic methods to extract value and generate actionable knowledge from Big Data, as well as methodologies to protect the privacy and confidentiality of individuals and organizations; (5) training opportunities for the current workforce; and (6) a larger supply of data science-savvy graduates interested in solving these challenges in specific contexts that arise in the public and private sectors.

Several MBDH spokes (Network Science, Smart Cities and Communities, Transportation, Business Analytics) will facilitate partnerships among researchers, communities, governments, nonprofits and businesses toward improving the effectiveness, safety, efficiency and effectiveness of how society and its members function. A key characteristic of the Midwest is its heterogeneous spatial distribution, and a corresponding variability in capacity to store, manage and analyze information to address societal challenges. For example, *SEEDCorn* will build on experience and programs associated with partnerships between universities and cities (e.g., the University of Chicago with the City of Chicago (**Letter**), Wayne State University with Detroit (**Letter**), Missouri working with St. Louis) to develop an open-source next-generation data analytics architecture to support city and academic workflows. A prototype of this architecture is the open-source Plenario1 platform, funded by NSF CISE, which serves as a starting point and has been optimized for (1) Chicago's predictive analytics needs, (2) San Francisco's Sustainable Districts evaluation needs, and (3) scientific inquiry from the NSF EHR/SBE-funded Urban Sciences Research Coordination Network2 as well as from the University of Chicago's Data Science for Social Good summer fellowship program. These efforts can be extended to address issues faced by smaller communities that have moderate or limited capacity to develop their own data systems. Extension networks in MBDH land-grant institutions can be leveraged to assist in educating and porting solutions to communities and citizens in rural areas.

Natural & Built World Spokes As a society, and particularly in the Midwest, we face major problems with fresh water, food and agriculture, and energy provisioning that we must address hand-in-hand with transportation and manufacturing demands. Although data science methods have been applied to large and complicated systems such as social networks, data science efforts in complex natural systems (with physical, chemical and biological elements) have been far more limited. MBDH will facilitate new insights into sectors of water, energy, food, agriculture, transportation and manufacturing.

Midwest states border the nation's largest freshwater reservoir and are dominant in agricultural production, transportation and distribution, ranking as the largest supplier of biofuel energy and the agricultural foundation of many local, regional and national economies and populations. Home to major urban centers with multiple modalities of transportation (including railroad and the heart of the automobile industry), the Midwest also includes major advanced and digital manufacturing concerns, including for agricultural and transportation equipment and equipment for the food industry. By enabling unprecedented interdisciplinary scientific exchanges and collaborations among these five sectors, *SEEDCorn* will facilitate partnerships and activities that foster development of necessary analytical methods, big data algorithms, visualization tools, and sensory acquisition methods, as well as access to storage, scalable operational infrastructure, and data management systems.

As an example, food and water quality testing and environmental studies require large volumes of genomic and metagenomics data describing the composition of microbial populations in samples; while next generation sequencing platforms allow for taxonomic profiling, annotating and linking of genetic information, this must subsequently be mined and correlated with other data sources. Sequencing errors, incorrect annotations and translational mistakes complicate this process, which is further impeded by the scale of the data. A second example is that of efficiently deriving relationships between food availability, water quantity/quality, and energy; this requires integration of data typically available only within an individual sector (farmers' use data, Army Corps' report data, energy sector data) as well as sharing of data across regions (i.e., soybean disease spreading from South America into the Midwest via hurricanes).

The spokes of the MBDH will enable sectors and cross-sector partnerships to address distinct research and engineering challenges: (1) understanding the impact of oil fracking on water; (2) understanding cross-sector dependencies and effective sharing of data as related to climate modeling, sustainable and adaptive food systems, and changing climate under economic and demographic conditions; (3) sustainability and ecosystem management using precision agriculture enabled by high-resolution crop yield data; (4) autonomous vehicles and new automotive design and development enabled by ubiquitous sensing and analytics capabilities; (5) quality management, defect tracking and elimination, supply management and shop-floor visibility in manufacturing; (6) intelligent manufacturing, digital manufacturing, design innovation and cyber-physical manufacturing networks and infrastructure.

Healthcare & Biomedical Research (HBR) Spokes There is no sector in American society that is positioned for more dramatic change driven by Big Data than Healthcare and Biomedical Research. Topol [11] points out that the multiplicative effect of innovations of the cell phone, the personal computer, the Internet, numerous digital sensor devices, DNA sequencing and -omics technologies, and social networks has positioned medicine for a "great inflection" before the end of this decade. This transformation is being driven by data, information, and the empowerment and engagement of patients. It is positioning society to move from acute care and over-reliance on emergency room visits to a chronic disease management focus leading to a wellness, prevention, and health focus. It would be hard to imagine a set of more complex big data characteristics than exists in this spoke. Big data challenges in healthcare & biomedical research include: (1) data and information standardization, integration and aggregation of biological and clinical research measurements, patient reported information, and sensor-generated data; (2) data and information privacy and health IT (HIT) security, including policies and regulations; (3) best practices regarding data sharing and use of open-source Big Data analytic applications.

The HBR Spoke will convene a set of hybrid workshops (on-site and virtual), integrated biomedical workforce training programs powered by online learning and on-site opportunities, and community building activities to establish a defined set of sustainable public-private partnerships focused on addressing opportunities and challenges related to Big Data in HBR. The ***HBR Spoke*** will leverage the ***Data Sciences, Education***, and ***Data Tools and Services*** rings and will interact strongly with the ***Network Science*** and ***Business Analytics Spokes*** (leadership detailed in *Partnerlist*).

Annual workshops will be convened at all five MBDH PI sites that will be led by the HBR Spoke Team. These will be attended by a growing set of healthcare and biomedical research partners from academia (Universities of Michigan, Cincinnati, Chicago and the Open Cloud Consortium, Iowa, Illinois, Indiana, Ohio State and Northwestern; and their partners including the Mayo Clinic and NorthShore University Health System; others affiliating), pharmaceutical/biotech industry (Abbvie, Eli Lilly, Transgenomics, Assurex Health, others affiliating), private healthcare systems (Henry Ford, Trinity Health, Regenstrief Institute); and non-profit foundations (tranSMART Foundation, Michael J. Fox Foundation, Open Cloud Consortium). An initial list of potential workshop topics includes:

- Data wrangling, scrubbing, and machine-learning methods to mine, analyze, visualize, and understand biomedical data (from -omics, health records, and mobile platforms) in a temporal fashion, capturing longitudinal trends to alert researchers of discovery opportunities, alert providers of intervention or adverse events, and to alert patients to provide positive feedback;

- Building new communities and partnerships to leverage global open biomedical science data sharing and analytics platforms such tranSMART, i2b2, and i2b2 SMART;
- Geospatial health informatics data analytics/visualization capabilities for HBR communities;
- Understanding how to work with patients in the home using biometric sensors cell phone applications and communications capabilities;
- Workforce development: a cohort of data science-enabled students and trainees at all levels (undergraduate through post-doctoral) to build and use emerging and future data platforms.

Rings Complementing and connecting the thematic *spokes*, we have created *rings* (Figure 1) to integrate spokes in advancing data science; creating and leveraging shared resources for data, tools, services; and as described in C.3 developing and sharing education and training resources and opportunities.

Data Sciences Ring Data science is an emerging field that represents the common core of motivating Big Data applications and includes the *data lifecycle* (data collection, structure, provenance management, curation and digital preservation), *methodologies* for processing and analyzing data (workflow planning, computation, databases, modeling, visualization), and *societal impacts* of big data (privacy and security, policy, ethics, usability). A challenge with data science is its rapid evolution through wide ranging applications, and a lack of integration of knowledge and experiences across these contexts. The Data Sciences ring will aggregate and coordinate expertise, helping to better define data science itself, and through the MBDH will bring this expertise specifically to the spokes, partners, and wider community, including the social and economic impact of data across all spokes. MBDH institutions are especially strong in theoretical, mathematical, statistical, and algorithmic aspects of data analytics; the underlying organization and curation of data; and deep consideration of privacy, confidentiality and benefits to society (with, e.g., extraordinary computer science, bio/statistics, information science departments, as well as top-tier programs that serve the federal statistical system and outreach in helping communities benefit from data). The MBDH and its partners will benefit from shared expertise and ongoing interactions that help define and expand core data science concepts for solving problems, proactively evaluate their impact on society, and educate the workforce in core data science concepts.

Through *SEEDCorn*, the ***Data Sciences Ring*** will work along several thrusts: (1) *community development*, organizing regular interactions and workshops (C.8) among ring and spoke participants to share ideas, opportunities, challenges, and best practices in data science broadly as well as in specific contexts that are meaningful to spokes; (2) *research*, responding to specific applications or common methodological problems arising across contexts, we will serve as a partnership resource for building inter-organizational research teams to develop novel approaches in data science and pursue sponsored funding to support these investigations; (3) *expertise,* providing access to individual/group knowledge on specific topics or services, such as the capacity to conduct audits needed for the Data Seal of Approval [12] to a repository; (4) *educational and workforce development*, developing and supporting educational, internship and hiring opportunities in data science for government, industry, nonprofits and academia.

While the ***Data Sciences Ring*** underpins how all spoke activities extract knowledge from data, it provides a foundation for activities of the ***Data Tools and Services Ring*** for implementation and linking of software and hardware that allow groups to store, retrieve, link, and analyze data.

Data Tools and Service Ring A key role of MBDH is to provide shared structures that support development of community data activities. This includes building on common existing cyberinfrastructure resources (C.4) to support: linking of existing or developing services and repositories, in our region or nationally; creation of new services needed by our growing communities; development of pilot projects that may be undertaken by these communities.

Communities face several key data challenges: how are data to be stored, curated, described, shared, discovered, verified, interpreted and linked, across repositories, with traditional publications, and with richer environments where they may be computed upon for additional investigation? These challenges are not restricted to academic sectors but extend to non-profits, industry, and individuals within the region. A

sophisticated set of services must be developed and made part of the culture of doing research; key challenges identified above relating to industrial and academic interests in the region require these services. Developing mechanisms for linking data repositories is particularly key for those activities that require participation or integration of multiple spokes or rings.

SEEDCorn is not funded to build such environments but will leverage independently funded resources and projects across MBDH, facilitating their integration and interoperation. In the Midwest, there are numerous data-related projects and resources (C.1, C.4), e.g. DIBBS, DataNet, Globus Online, OCC, NDS. Furthermore, the Research Data Alliance (RDA), rapidly emerging as the international organization through which protocols, best practices, and policies can achieve wide adoption, has strong ties within MBDH (e.g. Co-PI Plale). RDA will enable MBDH to gain wider adoption and acceptance for advances in interoperability, access, and curation.

We will actively support information sharing about individual tools and services between spokes and across hubs as well as developing social structures within which tools and services can be connected and developed. To support these (and other) goals, we will create a portal in our website to facilitate the sharing of information and expertise for all hub and spoke activities, including the regionally available tools and services. To support the social and technical aspects of tool sharing, the MBDH and the Western Hub (led by NDS partner UCSD) will jointly provide leadership in supporting this environment for all four regional hubs. *SEEDCorn* will leverage this activity, partially funding a support person to work with the interhub subcommittee on federated data service environments and hub projects to assist in the use of this environment for the MBDH and other hubs. A Data Services hackathon will be hosted in 2016 (C.8).

C.6. Governance

MBDH institutions and partners form a distributed hub that aggregates expertise, projects and resources from their members, enabling communities to assemble and function along distributed thematic spokes. MBDH has been built as a highly dynamic organization: Spokes are created as interest and opportunity evolves. Resources are expected to continually grow, and new opportunities along with them. The MBDH organization was formed as a shared collective, with institutions and partners working together and sharing leadership to make big data in the Midwest region as optimally successful as possible.

The organizational framework of MBDH is a hierarchical structure: At the lowest level of the hierarchy are diverse players engaging in thematic spoke activity around data collections, acquisition, management, etc. At the middle level are cross-spoke communication and cross-sector Big Data and sharing policy discussions as well as aligned educational interests and resources, tools and resources. At the top level the hub has a leadership structure consisting of a Steering Council and an Organizational Partners Board who work through an executive director (ED).

The cornerstone of the organization of MBDH is the Steering Council (SC) and the Organizational Partners Board (OPB). These bodies together provide vision, day-to-day oversight, representation by all stakeholders, and structure to enable cross-hub cooperation. The core organizational entities, their responsibilities and benefits, are described below. An interim SC is in existence. SC membership (see *Partnerlist*) is representative of all early stakeholders in the organization, with working groups around each existing spoke and ring, as well as for topics of education, industry, and diversity. The interim SC will give way to an elected SC by March 1, 2016.

Steering Council (SC) The SC consists of volunteer representatives from organizations who are actively involved in governance, with identified goals for the consortium. The initial membership will consist of the ring and spoke leads, five at large members, and the PI team. At-large membership may be used for consideration of perspectives on big data from underrepresented groups. Decision-making is through by-laws developed by the SC in the first year of the project. The SC chair is an elected position, serving a one-year term with the possibility of a renewal year. The responsibilities of the SC are chosen to reflect foundational aspects that guide the structure of big data partnerships. These include but are not limited to: (1) set strategy and agenda for the Hub; (2) initiate formation of data and tool sharing policies as need

emerges, and work through international bodies like Research Data Alliance (RDA) for broader impact; (3) monitor and evaluate the success of the Hub; (4) invite new members to serve on the OPB; (5) develop procedures and expectations for new partnerships, spokes, or rings; and (5) assess issues surrounding Hub sustainability. SC members are expected to serve as the initial points of contact for potential partners. They will leverage their knowledge of local industry, consortia, nonprofits, governmental agencies, and academic institutions to link existing and new partners with the Hub.

The SC will also hold discussions with the other bodies of governance on further aspects of partnerships, such as (1) public access to data and other products (including permission to publish); (2) expectations for financial support (cost/benefit); (3) educational component (training, outreach and extension); (5) broader impact; (6) ethics and responsible conduct of research; and (7) expectations for productivity (technology, publications, curricula, conferences, workshops, etc.).

Organizational Partners Board (OPB) The OPB is an organization of big data stakeholders (i.e., companies, non-profit organizations, and federal and state agencies) from the Midwest region. OPB members have active and ongoing relationships/partnerships with members of the Hub. Each OPB member institution has a single OPB seat. The roles of the OPB include developing use cases, providing data for research, representing their needs to drive Hub and Spoke strategies, and developing structures to increase the value of data sharing and Hub membership. The OPB nominates and appoints its own chair and develops its own by-laws. The OPB chair holds a seat on the SC.

Affiliated Partners Affiliated partners are individuals who represent non-members institutions from outside the region. Each affiliated partner fills a critical shortage within the region and has an active, short-term relationship with one or more members of the MBDH. Affiliated partners are approved by the SC and will contribute use cases, data sets, and big data resources. In return affiliated partners have limited access to Hub activities as appropriate to their relationship. For instance, an affiliated partner who participates through a Spoke, will have a received benefit derived from that Spoke.

MBDH Staff The executive director (ED) is a full-time paid position, which along with associated staff positions, implements the decisions of the SC and oversees day-to-day operations of the MBDH. Hub operations, however, are distributed; along with the full-time ED, fractionally funded staff working on different aspects of the consortium are at four other PI locations throughout the Midwest. The fractional project coordinator and technical support person report to the ED, who reports to the PI Seidel and SP Nahrstedt at Illinois but is accountable to the SC.

Spoke Governance Spokes are led by spoke leads. It is anticipated that individual spoke groups will require semi-autonomy (i.e. distributed authority; room to optimize terms to maximize value for different spokes). Spokes will develop partnerships and offer services as described above.

Ring Governance Rings are led by a ring lead. Rings will continually refine services that spokes can take up but will offer services of their own. In conjunction with other hubs, the ***Data Tools and Services Ring*** will offer an environment in which services that span all hubs may be developed.

C.7. Goals and metrics

The metrics for evaluating our proposed framework for the Hub and partnerships will be based on growth and economic impact as well as benefits to society and local communities. These metrics will evolve with the Hub and its activities. Specific outcomes of *SEEDCorn* will be many, including:

Partnerships Strengthening and creation of numerous public-private partnerships, built on a stronger funding base, with new projects funded by multiple agencies, government organizations, and private industry as a result of workshops and other *SEEDCorn* events. Progress will be measured by number of public-private partnerships, with tracking of funding received, by number of members (growth, retention), and by the growth of shared data resources.

Joint activity Projects will not operate in isolation, but will be connected through MBDH. Progress will be measured by the increase in external funding for member organizations, funding with demonstrated

links across MBDH members, and by the number of collaborative projects between members and joint publications.

Education and workforce development New educational activities (online lessons, events, curricula) and best practices will be gathered, organized, and delivered through our collaborative efforts. These efforts are measured by the uptake by different organizations within and outside the region and the U.S., graduate and undergraduate courses that include materials developed in MBDH, and individuals interested in retooling their expertise. Also tracked are events and attendance, with focus on outreach to underrepresented groups.

Pilot activity *SEEDCorn* will support pilots in a common technical data service environment through use of NDS Labs, leading to innovation, interlinking and acceleration of data services across potentially hundreds of projects. Progress will be measured by number of pilots and success acquiring funding.

Policies and sustainability New business models for sustainable data solutions will be developed and implemented, as will data policies and standards, including for IP.

These metrics reflect our vision of the Hub as a facilitator of research/education/infrastructure by providing access to shared resources and information from partners about challenges and resources.

C.8. Timelines

We summarize in Table 1 the detailed planning of MBDH events and workshops to be supported by *SEEDCorn*, which is frontloaded so the majority of events are supported in the first year, with a goal of securing funding for subsequent years. A detailed sustainability plan is to be delivered by the SC in Y2.

Table 1: *SEEDCorn* planned milestone/events for MBDH.

Deliverables	**Y1**		**Y2**	**Y3**
	Fall	Spring		
Executive Director identified	X			
Elected Steering Council Operates		X		
Sustainability plan developed by Steering Council			X	
MBDH portal for educational and data service activities			X	
Events				
PI Team coordinating event Project all-hub kick-off workshop at Illinois	X			
PI Team coordinating event Project all-hub kick-off workshop at Illinois; Steering council meeting	X			
All Co-PIs attend NSF BD Hub meeting	X			
Data Spoke workshop on Tools/Services workshop at Illinois	X	X		
UND Big Data Summit	X	X	X	
Food, Water, Energy Workshop at Illinois		X		
Digital Ag workshop at Iowa State				
Industry workshop		X		
Data Science workshop at Iowa State		X		
Tools/services hackathon at Illinois		X		
All hub workshop		X	X	X
BD Hub Consortium Meeting		X	X	X
Workshop Development/Diversity workshop at Illinois			X	
Tools workshop at Illinois			X	
Energy workshop at Illinois			X	X
Healthcare & Biomedical Research and Life Science Big Data Workshop at University of Michigan		X		
Transportation Big Data Workshop at University of Michigan			X	
Business Analytics Big Data Workshop at University of Michigan Wayne State University		X		X

C.9. Broader Impacts of the Proposed Work

We are particularly excited about the broader impact potential for *SEEDCorn.* The project builds on many partners who have come together to create the MBDH, an unusually large and diverse consortium of partners, far exceeding usual NSF grants. In addition to small colleges, large R1 private and state universities, and rural universities across the region, MBDH has dozens of additional organizations, including city and state governments, nonprofits, and companies, among its members. The consortium is built to create and sustain academic-industry-government partnerships, with reach into all sectors of the Midwest. The region is rich in diverse and underrepresented populations, including Hispanic, African-American, and Native American, who have much to gain by participating in the democratization of knowledge through data sharing. The project specifically includes an experienced diversity coordinator (SP Franklin) as an SC member, who will leverage existing and new relationships with national organizations to engage multiple sectors of society.

C.10. Results of Prior NSF Support

Edward Seidel, Brian Athey and Josh Riedy have not had project support from NSF within five years.

Sarah Nusser: SES0822002, $230,189, 10/2008-7/2012, Accommodating Individual Differences in Software Designed for Location-Based Survey Tasks. *Intellectual merit*: Developed a model that describes the variation for low and high spatial persons in how they use maps, work with map software, and perform location-based field tasks for census and survey applications. *Broader impacts*: Contributed to design principles for mobile interfaces for census and survey field work; supported two REU students, one HCI minority and one Computer Science graduate student, five talks and seven publications.

Beth Plale DataNet: Sustainable Environment Actionable Data (SEAD). (Co-PI: NSF ACI 0940824, $8,000,000, 9/27/2011-8/11/2014). SEAD is developing tools to reduce the curation barrier to active data curation and publishing for scientists in the long tail. *Intellectual Merit*: Advanced the understanding of long-tail science [Plale2012]. Prototyped active data curation and publishing services [Plale et al. 2013] that embed a model of minimal provenance in [Plale et al. MIT Press]. *Broader impacts*: SEAD publishes results through DataOne as a member node.

BD Hubs: Midwest: SEEDCorn: Sustainable Enabling Environment for Data Collaboration
PartnerList

Project Personnel and Partner Organizations

1. Ed Seidel; National Center for Supercomputing Applications/University of Illinois Urbana-Champaign; PI and ***Interim Steering Council Chair and Data Tools and Services Lead***[1]
2. Beth Plale; Indiana University; co-PI, Sub-awardee, and ***Interim Steering Council – Data Sciences Spoke Lead***
3. Sara Nusser; Iowa State University; co-PI, Sub-awardee, and ***Interim Steering Council – Digital Agriculture Spoke Lead***
4. Brian Athey; University of Michigan; co-PI, Sub-awardee, and ***Interim Steering Council – Healthcare & Biomedical Spoke Lead***
5. Joshua Riedy; University of North Dakota; co-PI, Sub-awardee and *Interim Steering Council – At Large Member*[2]*/Data Sciences Spoke*
6. Caralynn Nowinski; UI LABS nonprofit; Senior Personnel and ***Interim Steering Council – Digital Manufacturing Spoke Lead***
7. Charlie Catlett; Argonne National Labs and University of Chicago; Senior Personnel and ***Interim Steering Council – Smart Cities and Communities Spoke Lead***
8. Klara Nahrstedt; University of Illinois Urbana-Champaign; Senior Personnel and ***Interim Steering Council – Food, Energy, Water Spoke Lead***
9. R. Babu Chinnam; Wayne State; Senior Personnel and ***Interim Steering Council – Business Analytics Spoke Lead***
10. Wolfgang Kliemann; Iowa State; Senior Personnel and ***Interim Steering Council – Education Ring Lead***
11. H. V. Jagadish; University of Michigan; Senior Personnel and ***Interim Steering Council – Transportation Spoke Lead***
12. Bernice Pescosolido; Indiana University; Senior Personnel and ***Interim Steering Council – Social Networks Spoke Lead***
13. Kevin Franklin[3]; University of Illinois Urbana-Champaign; ***Interim Steering Council – Diversity Lead***
14. Keith Ellison[4]; tranSMART Foundation; ***Interim Steering Council – Industry/Sustainability Lead***
15. Greg Monaco; Great Plains Network; Senior Personnel and *Interim Steering Council – At Large Member/Data Tools and Services Ring*
16. Jennifer Clark; University of Nebraska-Lincoln; Senior Personnel and *Interim Steering Council – At Large Member/Digital Agriculture Spoke*

[1] Each ***Interim Steering Council Lead*** (***listed in bold***) also chairs an MBDH working group on a topic. The interim Steering Council will be superseded by an elected council by March, 2016.
[2] Interim Steering Council *At-Large Members* (*listed in italics*) have specific areas of interest listed, though they do not lead them.
[3] Franklin formally leads a Steering Council working group on diversity.
[4] Ellison formally leads a Steering Council working group on industry interests and sustainability.

17. Jun (Luke) Huan; University of Kansas; Senior Personnel and *Interim Steering Council – At Large Member/Food, Energy, Water Spoke*
18. Lior Shamir; Michigan Technological University; Senior Personnel and *Interim Steering Council – At Large Member/Education Ring*
19. Alex Yahja; National Center for Supercomputing Applications; Senior Personnel
20. Bob Grossman; University of Chicago; Senior Personnel
21. Dan Reed; University of Iowa; Senior Personnel
22. Diego Klabjan; Northwestern University; Senior Personnel
23. Ivo Dinov; University of Michigan; Senior Personnel
24. Joe Colletti; Iowa State; Senior Personnel
25. Kevin Smith; University of Michigan; Senior Personnel
26. Matt Turk; University of Illinois Urbana-Champaign; Senior Personnel
27. Michael Fry; University of Cincinnati; Senior Personnel
28. Placid Ferreira; University of Illinois Urbana-Champaign; Senior Personnel
29. Scott Wilkin; National Center for Supercomputing Applications; Senior Personnel
30. Shaowen Wang; University of Illinois Urbana-Champaign / National Center for Supercomputing Applications; Senior Personnel
31. Shashi Shekhar; University of Minnesota; Senior Personnel
32. Vallabh Sambamurthy; Michigan State; Senior Personnel
33. Brain Athey; University of Michigan; Senior Personnel
34. John Towns; National Center for Supercomputing Applications, National Data Service, XSEDE; Senior Personnel
35. Jun Li; University of Michigan; Senior Personnel
36. Maxine D. Brown; U Chicago; Senior Personnel
37. Sanjay Madria; Missouri University of Science and Technology; Senior Personnel
38. Yinlun Huang; Wanye State; Senior Personnel
39. Ravi Bapna; University of Minnesota; Senior Personnel
40. Gabrielle Allen; University of Illinois Urbana-Champaign; Senior Personnel
41. Allen Renear; University of Illinois Urbana-Champaign; Senior Personnel
42. Bertram Ludaescher; University of Illinois Urbana-Champaign; Senior Personnel
43. Caterina Scoglio; Senior Personnel
44. Hossein Davari; University of Cincinnati; Senior Personnel
45. Dr. Ashok Krishnamurthy; RENCI: Unpaid Collaborator
46. Srinivas Aluru & Dr. Ashok Krishnamurthy; South Big Data Regional Innovation Hub; Unpaid Collaborator
47. Kathleen McKeown; Northeast Big Data Innovation Hub; Unpaid Collaborator
48. Mike Norman; West Big Data Innovation Hub; Unpaid Collaborator
49. Jorge V. José; Indiana University; Unpaid Collaborator
50. David Broecker; Indiana Biosciences Research Institute; Unpaid Collaborator
51. Paul Gunderson; Dakota Precision Ag Center; Unpaid Collaborator
52. Patrick Pope; Nebraska Public Power District; Unpaid Collaborator
53. Kathy Schroeder, HIS Automotive, drive by Polk; Unpaid Collaborator
54. John Ginder; Ford Motor Company; Unpaid Collaborator
55. Daniel Vivian; General Motors Company; Unpaid Collaborator

56. Henry Benedetto; Dominos Pizza LLC; Unpaid Collaborator
57. Josephine Molle; Henry Ford Health System; Unpaid Collaborator
58. Ginny Walls; Macy's; Unpaid Collaborator
59. Mahesh Rajasekharan; Cleo Communications; Unpaid Collaborator
60. Brenna Berman; City of Chicago; Unpaid Collaborator
61. Heather Woodward-Hagg; Department of Veterans Affairs – VA Center for Applied Systems Engineering; Unpaid Collaborator
62. Lisa Phillip; Quicken Loans; Unpaid Collaborator
63. James Buntrock; Mayo Clinic; Unpaid Collaborator
64. John Reid; John Deere; Unpaid Collaborator
65. Nicholas Hatcher; QuesTek Innovations, LLC; Unpaid Collaborator
66. Stuart Aitken; 84.51°; Unpaid Collaborator
67. Kevin Kelley; Great American Insurance Company; Unpaid Collaborator
68. Jude Schramm; GE Aviation Information Technology; Unpaid Collaborator
69. Mitra Dutta; University of Illinois Chicago; Unpaid Collaborator
70. Dayle McDermitt; LI-COR, Inc.; Unpaid Collaborator
71. Beth Niblock; City of Detroit, Information Technology Services Department; Unpaid Collaborator
72. James Anderson; Urban Science Applications, Inc.; Unpaid Collaborator
73. Venkat Gone; Loven Systems, LLC; Unpaid Collaborator
74. P. Brighten Godfrey; Veriflow Systems, Inc.; Unpaid Collaborator
75. Paul Baniewicz; Alcatrel-Lucent; Unpaid Collaborator
76. Christopher Harbourt; Agrible; Unpaid Collaborator
77. Tel Ganesan; Kyyba, Inc.; Unpaid Collaborator
78. Paul Riser; TechTown Detroit; Unpaid Collaborator
79. Tony Brownlee, Kingland Systems; Unpaid Collaborator
80. Matt Spackman, Kum and Go; Unpaid Collaborator
81. Mary Berry; University of South Dakota; Unpaid Collaborator
82. James Tracy; University of Kansas; Unpaid Collaborator
83. Kevin Kephart; South Dakota State University; Unpaid Collaborator
84. Ophir Trigalo; Illinois Institute of Technology; Unpaid Collaborator
85. Chaille Becker; Caterpillar Inc.; Unpaid Collaborator
86. Susan Marquis; Pardee RAND Graduate School; Unpaid Collaborator
87. Nick Lindberg; Milwaukee Institute; Unpaid Collaborator
88. Tony Brownleee; Kingland Systems; Unpaid Collaborator
89. David Dittmann; Proctor & Gamble; Unpaid Collaborator
90. Matthew Gibb; Carle Health System; Unpaid Collaborator
91. Susan Ford; Southern Illinois University Carbondale; Unpaid Collaborator
92. James Garvey; Southern Illinois University Carbondale; Unpaid Collaborator
93. Raymond Goldsteen; University of North Dakota, Center for Comparative Effectiveness Analytics; Unpaid Collaborator
94. Jianglong Zhang; University of North Dakota, John D. Odegard School of Aerospace Sciences; Unpaid Collaborator
95. L. Keith Henry; University of North Dakota, Department of Basic Sciences; Unpaid Collaborator
96. Prem Paul; University of Nebraska, Lincoln; Unpaid Collaborator

97. Vipin Kumar; University of Minnesota; Unpaid Collaborator
98. Claudia Neuhauser; University of Minnesota; Unpaid Collaborator
99. Henry Foley; University of Missouri; Unpaid Collaborator
100. William S. Ball; University of Cincinnati; Unpaid Collaborator
101. S. Jack Hu; University of Michigan; Subawardee

A. References Cited

[1] *MidWest Big Data Hub.* June 22, 2015. http://midwestbigdatahub.org/.

[2] Institute, McKinsey Global. "Big data: The next frontier for innovation, competition, and productivity." 2011.

[3] *Forbes.* "Revealing Data Science's Job Potential." 2014.

[4] *Information Week.* "Wanted: Qualified Data Scientists, People Skills A Plus." 2012.

[5] Gartner. "Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015."

[6] Terrapop. http://www.terrapop.org/.

[7] Center, Cline. http://www.clinecenter.illinois.edu/.

[8] UMN, PGC @. http://www.pgc.umn.edu/.

[9] CyberGIS. http://cybergis.cigi.uiuc.edu/cyberGISwiki/doku.php.

[10] SpatialHadoop. http://spatialhadoop.cs.umn.edu/.

[11] Topol. *The Creative Disruption of Medicine.* Basic Books, 2012.

[12] DataSealofApproval. http://datasealofapproval.org/.