IOWA STATE UNIVERSITY

Office of the Vice President for Research

Perspectives on Data Driven Discovery

Sarah Nusser

Vice President for Research

Professor, Department of Statistics

Big Data Summer School, Iowa State University

June 20, 2016

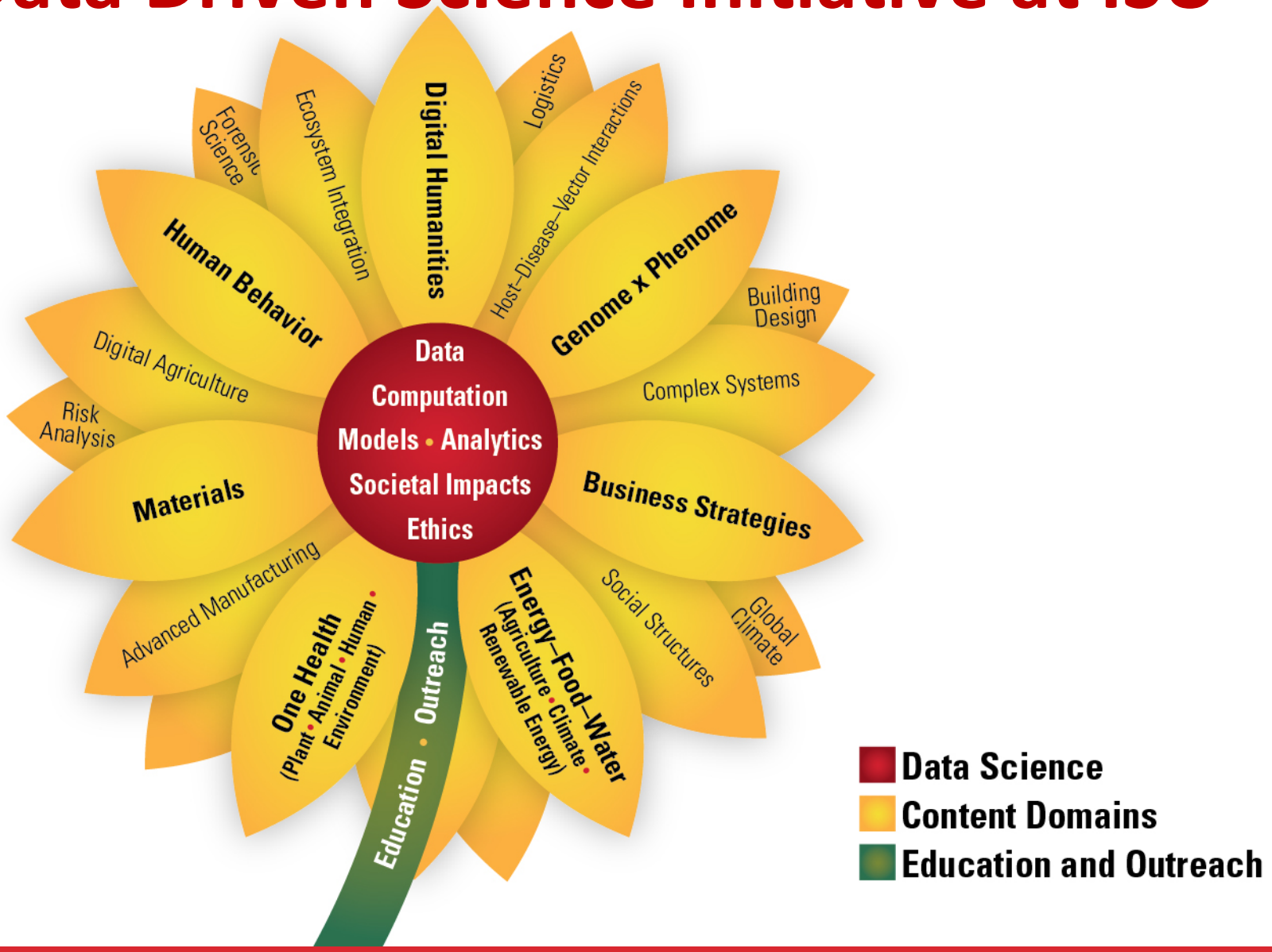
Big Data

- Dozens of Vs?
- Fundamental shift in how we engage with discovery
- Encompasses *not only data*, but *approaches* to research and decision-making that *extract knowledge from complex data sources*
- Our ability to understand data and its underlying structures will soon become as common as regression and microscopy are today

Data Driven Discovery

- *Content domain* or application area – the motivating context
- *Data science* – methodological and broader impacts contexts framed by the application and approach
  1. Methodologies for processing and analyzing data (computing, statistics, curation, …)
  2. Societal impacts (privacy, ethics, …)
- *Education and outreach* – knowledge, tools

Data Driven Science Initiative at ISU



Data Driven Discovery

- Intensely interdisciplinary (applications, methods, societal impacts)
- Funding agencies seek research teams that meaningfully integrate these dimensions
- Pervasive impact also generates huge demand for practical knowledge in the workplace

Challenges

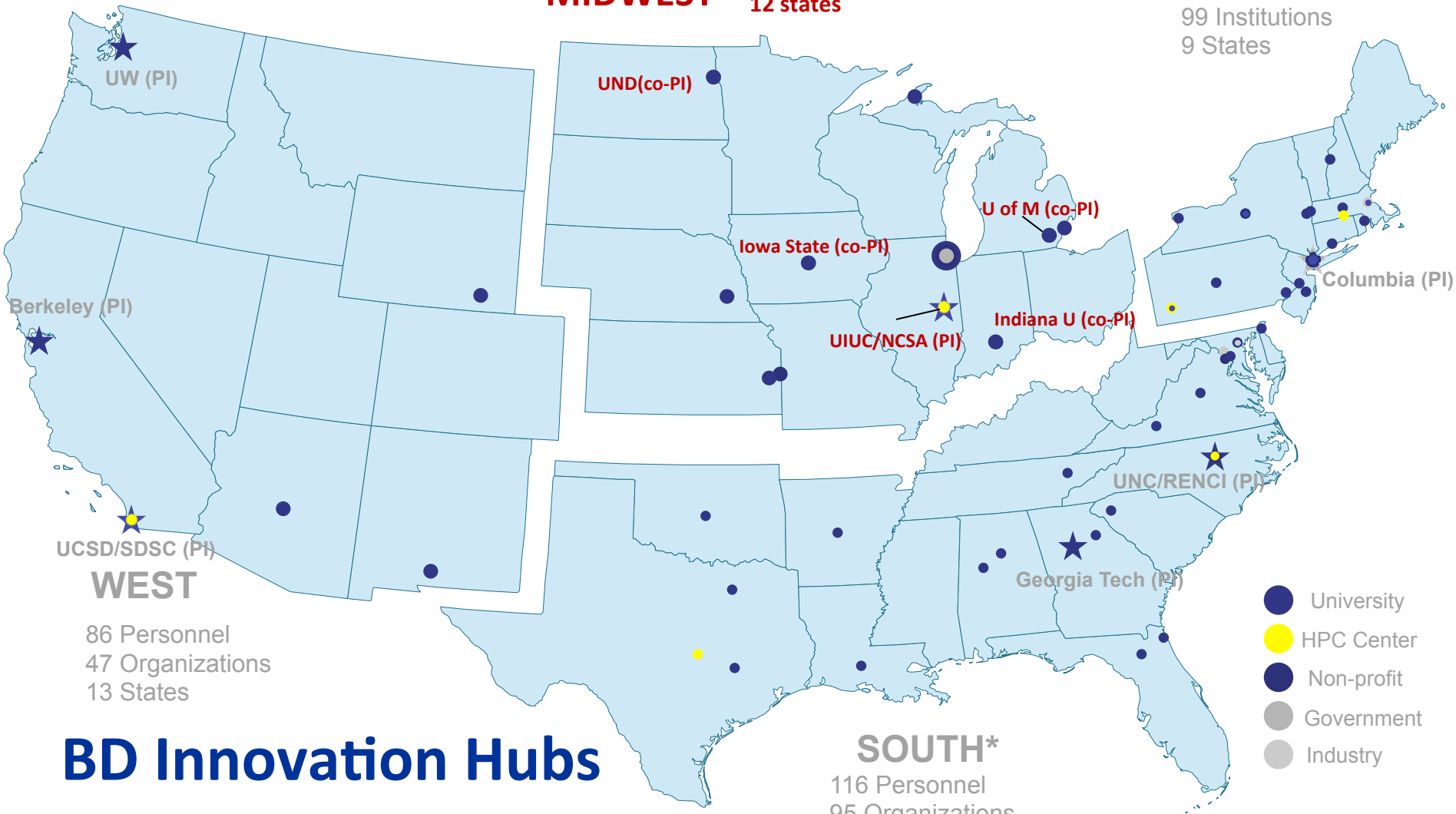
- Inherently interdisciplinary, not really a “field”
- Pervasive, evolving by simultaneously infusing a vast array of fields
- A coherent understanding of big data and data science still emerging
- Insufficient access to knowledge needed to effectively leverage data driven discovery approaches, in all sectors

NSF Support for Addressing Challenges in Big Data

- Research (BIGDATA)
- Computational infrastructure (DIBBS)
- Graduate training (NRT)
- Regional partnership hubs (BD Hubs, Spokes)

An experiment in response to White House mandate to accelerate big data knowledge transfer and innovation across sectors

Alaska & Hawaii are part of the West region
US Territories can participate in any region



BD Innovation Hubs

Points indicate affiliations of individuals named as steering council members and/or task leads.

Midwest Big Data Hub

Accelerating the Big Data Innovation Ecosystem

Ed Seidel
PI (Illinois)

Beth Plale
Co-PI (Indiana)

IOWA STATE UNIVERSITY

Sarah Nusser
Co-PI (Iowa State)

Brian Athey
Co-PI (Michigan)

Josh Riedy
Co-PI (UND)

Melissa Cragin
Executive Director

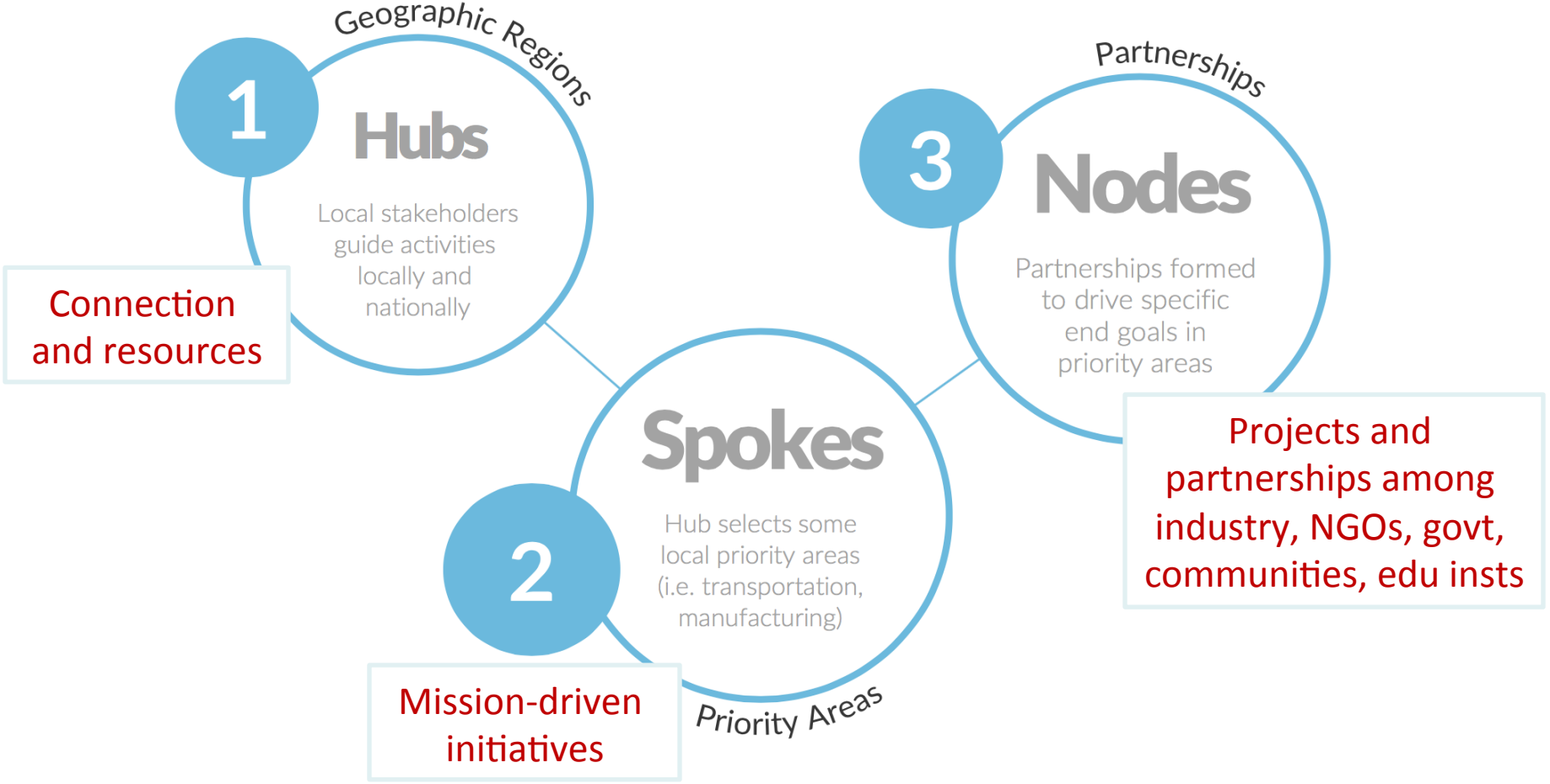
midwestbigdatahub.org

MBDH Mission

Creating effective
cross-sector communities that can
harness the power of data to
address societal problems of relevance
in the Midwest and be a
driver for regional economic development

WHAT IS THE BDHUBS NETWORK?

"Hub and Spoke"– A Nation-Wide Network for Data Innovation



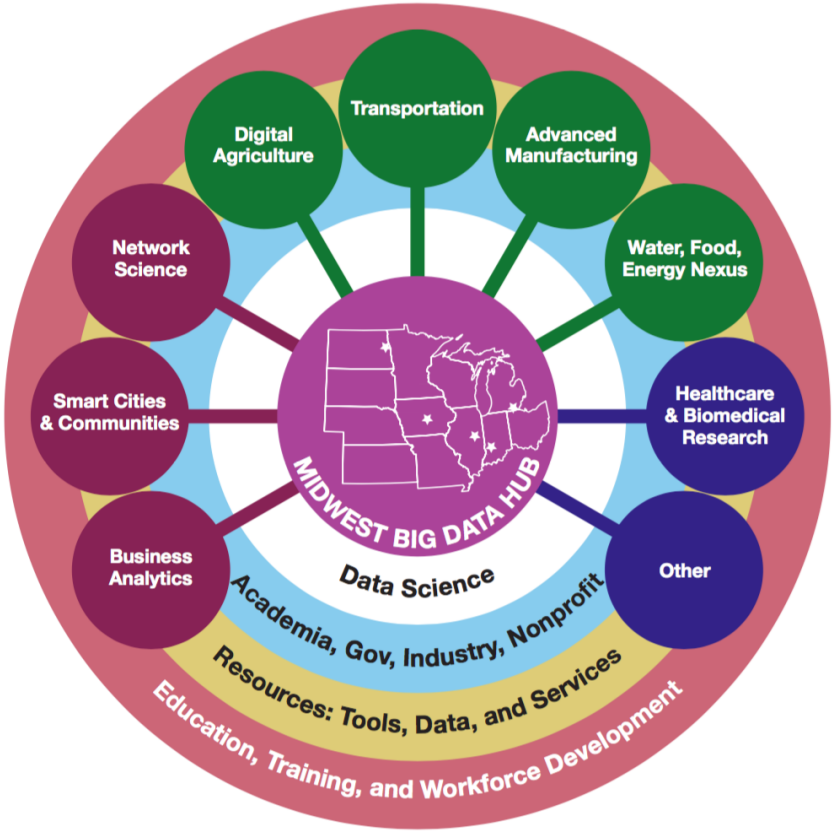
Spokes: Thematic Foci

Cross-sector working groups that facilitate big data innovation in the context of a specific theme

- Contribute to solving grand challenges
- Share assets and resources to address societal issues
- Develop automation opportunities for the data life cycle

MBDH Spokes

- Digital Agriculture
- Food, Energy, Water
- Metropolitan Science
- Transportation
- Healthcare & Biomedical
- Advanced Manufacturing
- Business Analytics
- Network Science
- Others as proposed …



MBDH Cross-cutting Rings

- Data Science
  - Ethical, legal and societal impacts (ELSI)
  - Replicability and reproducibility
- Education
  - New approaches to STEM learning
- Data Tools and Services
  - Data sets, services to share data, and tools to act on them

MBDH Events

midwestbigdatahub.org

- MBDH all-hands meeting (March, Rosemont, IL)
- Digital Agriculture all-hands meeting (May, ISU)
- Midwest Big Data Summer School (June, ISU)
- Data Science for Food, Energy, Water workshop (August, ACM Knowledge Discovery + Data Mining)
- Student Hackathon at U of Iowa
- MBDH all-hands meeting (Oct 10-12, Rosemont, IL)

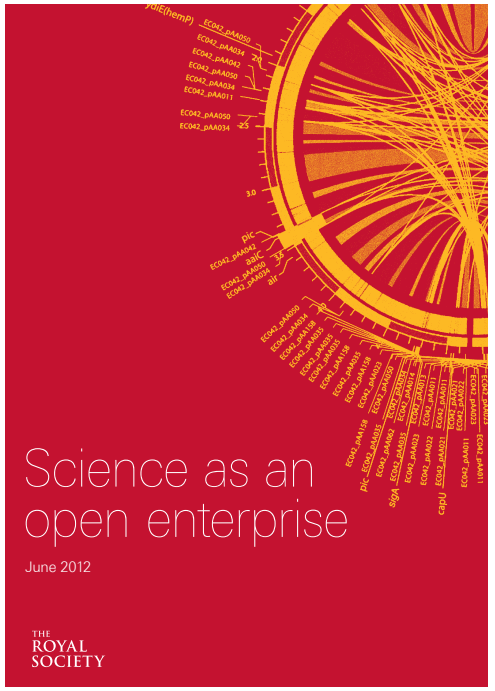
Digital Agriculture All-Hands Mtg

- Participation by academics, foundations, govt, NGOs, industry
- Sally Rockey, Executive Director, Foundation for Food and Agriculture Research
  - Data index – cataloging agricultural data sets and availability
- Industry, start-ups, research communities
  - Sharing new approaches
  - Commercializing tools
  - Creating new business models

Open Science and Inquiry

- Another transformative force
- *Science as an Open Enterprise: Summary Report*

  2012, The Royal Society



royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe-summary.pdf

Premise of inquiry unchanged

"Open inquiry is at the
heart of the scientific enterprise"

***Science as an Open Enterprise Summary Report*, The Royal Society, 2012**

- Historically, this has involved:
  - Publication of theories, experimental and observational data, interpretations
  - Advancement of ideas through exchanges, colloquia and other venues
  - Self-correction through debate and scrutiny

Practice of inquiry is changing

- Centrality of printed page receding with digital technologies
- Increased emphasis on sharing data publicly
- Large-scale data collection and analysis creates challenges for traditional autonomy of individual researchers
- Internet provides a conduit for networks of scientists and public to collaborate and communicate

Public Health Emergency

- Severe gastrointestinal infection outbreak
  - **May 2011** – 50 deaths/4000 cases in EU/US
  - Rare and little known strain of E. coli
- **Collaborations** on 4 continents, open sharing
  - Draft genome (+3d), genome (+1d), virulence and resistance genes, appropriate antibiotics (+2w)
- Open access to **data** and **publications** led to rapid testing and containment of emergency
  - **July 2011** – Papers published

Open Science and Inquiry

Open data

(available, intelligible, assessable & useable)

combined with

Open access to scholarly publications

and

Effective communication of their contents

[among scholars and public]

Open Data requires more than Access

- Accessible publications and data
- **Effective communication** through more **intelligent openness** when **sharing data**
  - **Understandable** to those who wish to scrutinize data
  - **Assessable** for evaluating reliability of data and competence of data producers
  - **Usable** by others (scholars, public) for understanding and new discoveries

Enabling change – Royal Society List

1. **Culture**: Shift away from a research culture where data are viewed as a private preserve
2. **Credit**: Expand criteria for evaluating research to give credit for useful data communication and novel ways of collaborating (tenure, grant proposals)
3. **Standards**: Develop common standards for communicating data

Enabling change – Royal Society List

4. **Pubs**: Mandate intelligent openness for data relevant to published papers
5. **Workforce**: Increase supply of data scientists to manage and support use of digital data
6. **Tools**: Develop and use new software tools to automate and simplify creation and further use of data sets

Open Science and Inquiry in US

- In the US, other dimensions include
  - Accountability for federal agencies
  - Enabling reproducibility studies
  - Public engagement and trust
  - New opportunities with future reuse

Open Access in the US

- 2013 OSTP memo directed agencies to:
  - Ensure pubs, data from research funding are shared (major funders)
  - Enable costs to be covered in grants
  - Protect privacy and confidentiality, proprietary information, security
- Agencies have responded via DMPs, but still in early stages of mature approach

Open Access to Publications

- Many systems already support open access
  - Journals
  - Agency and disciplinary repositories
- Universities developing policies, programs and recommendations to support open access

Open Data

- Data are not the defined final product that publications represent
- Except for data supporting publications and practice in selected fields, most sharing currently occurs on a small scale
- Disciplinary cultures vary widely in their practices for preparing, sharing, curating data
- Repository system is not complete, nor are guidelines all that helpful currently

Open Data

- What is purpose of sharing? Audience?
  - Validate publication vs reuse, reproducibility
  - Scholars vs public
- What data should be shared for that purpose?
- When should data be protected?
  - Evaluating risk of disclosure
- Alternative approaches for secure release of protected data?
  - Disclosure limitation methods (statistical, systems)

Open Data

- How to communicate data?
  - Context of goals, methods, data manipulations
  - Standards for documenting data
  - Curation practices
- Options for storing data?
  - Data commons, federation of repositories
- Tools for proactive planning in research process?
  - Open Science Center, Research Data Alliance, …

IOWA STATE UNIVERSITY

Office of the Vice President for Research

Perspectives on Data Driven Discovery

Sarah Nusser

Vice President for Research

Professor, Department of Statistics

Big Data Summer School, Iowa State University

June 20, 2016