# Data and Data Science Challenges in Understanding & Innovating for F-E-W Nexus

April 7th, 2016

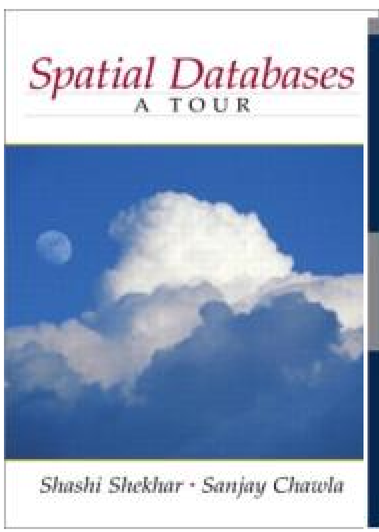
Fate of the Earth 2016 Workshop, Michigan State University

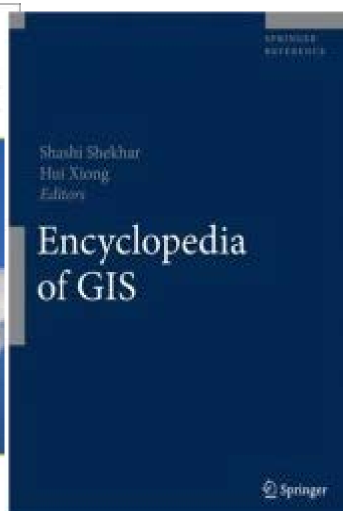

Shashi Shekhar

McKnight Distinguished University Professor

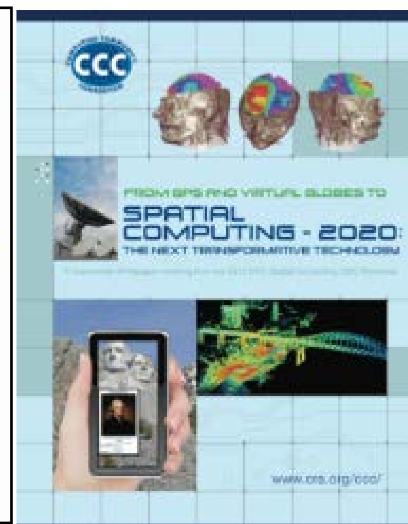
Computer Sc. & Eng., University of Minnesota

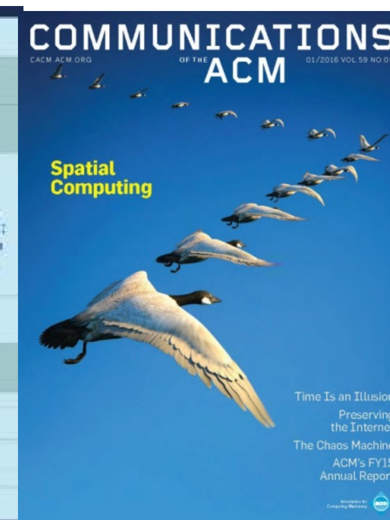
www.cs.umn.edu/~shekhar





# Outline

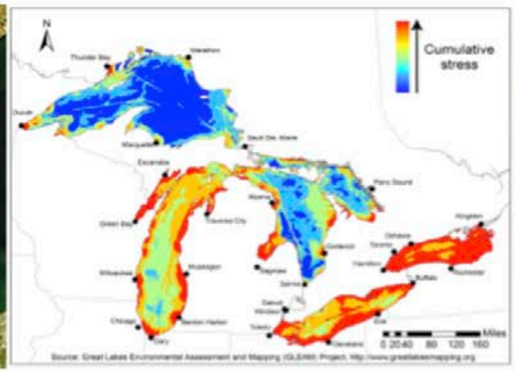
- F-E-W Nexus
  - Context
  - History
- Role of Data & Data Science
- Data & Data Science Gaps
- Next

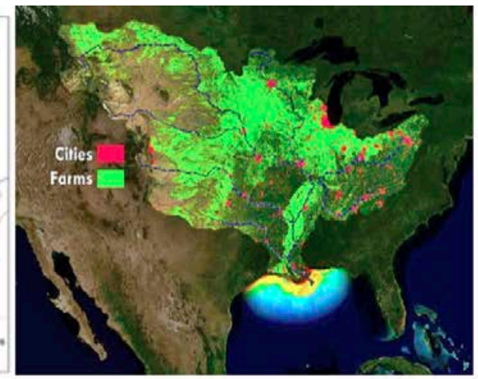
# U.N. Sustainable Development Goals 2030

includes Food (2), Energy (7), Water (6), Climate Action (13), …

# Downside of Piece-meal Approach

- Piece-meal policies => unanticipated problems
  - Ex. Fertilizers affect Water quality (e.g., Great Lakes, Mississippi River)
  - Ex. Bio-fuel subsidy => Rise in food prices (2008)


Courtesy: Wikipedia

- Crucial to understand interactions across Water, Food, Energy Systems
  - Not just for mid-west
  - National priority with initiatives from USDOD/NIC, NSF, USDA USDOE, USGS, …
  - Global priority with initiatives from U.N. University and many countries

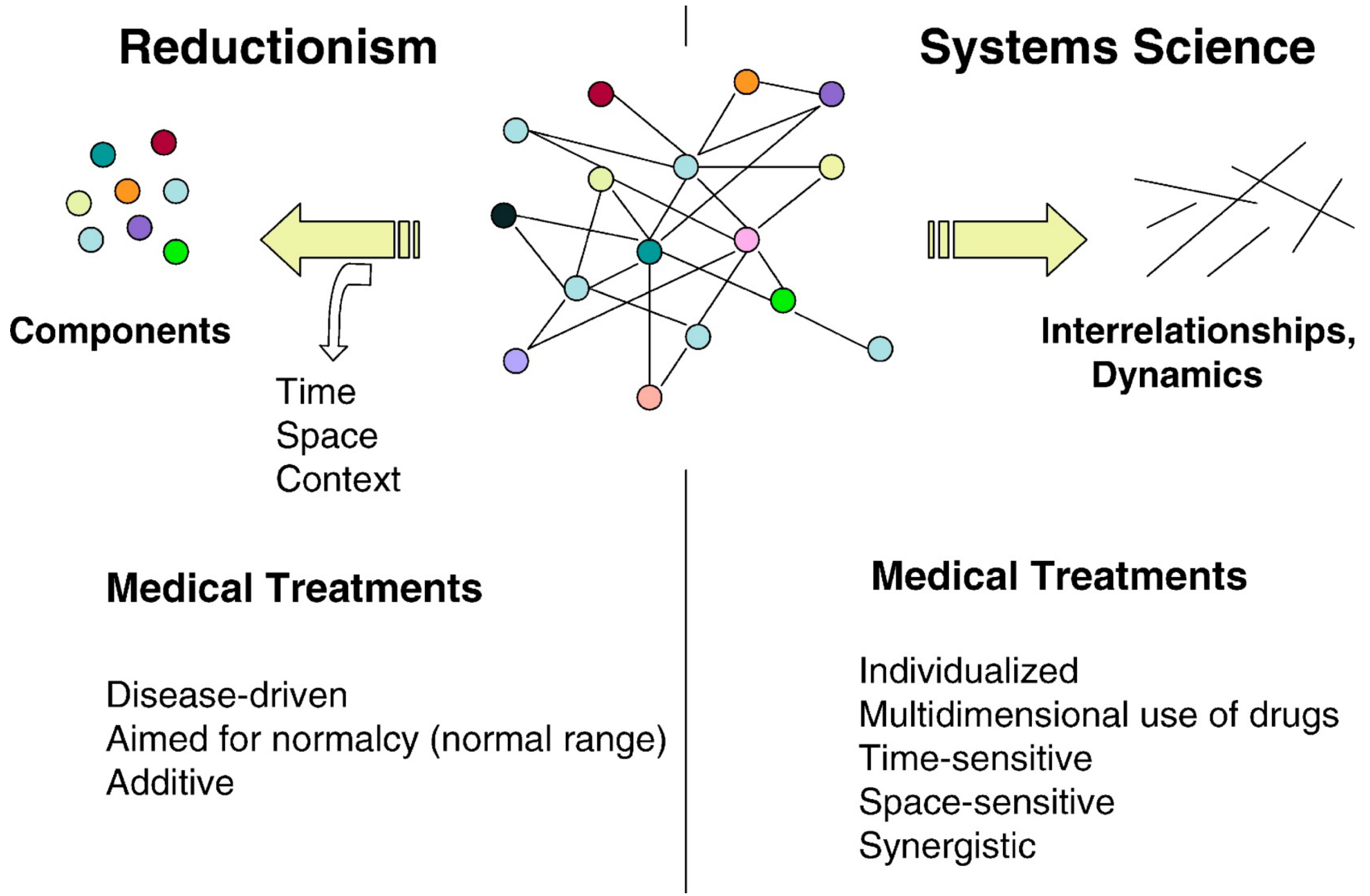
# Alternative to Piecemeal Approach

- Context-aware
  - Systems, Holistic, Integrated, Networks, Exogeneous, …
  - Ecological, Environmental, Geo, Spatial, …

- Ex.: Ecology – Study of interactions among organisms and environment
  - Barry Commoner's four laws
    - Everything is connected to everything else.
    - Everything must go somewhere.
    - Nature knows best.
    - There is no such thing as a free lunch.
- Ex.: Geography
  - Tobbler: Everything is connected to everything else but
  - but nearby things are more related than distant things

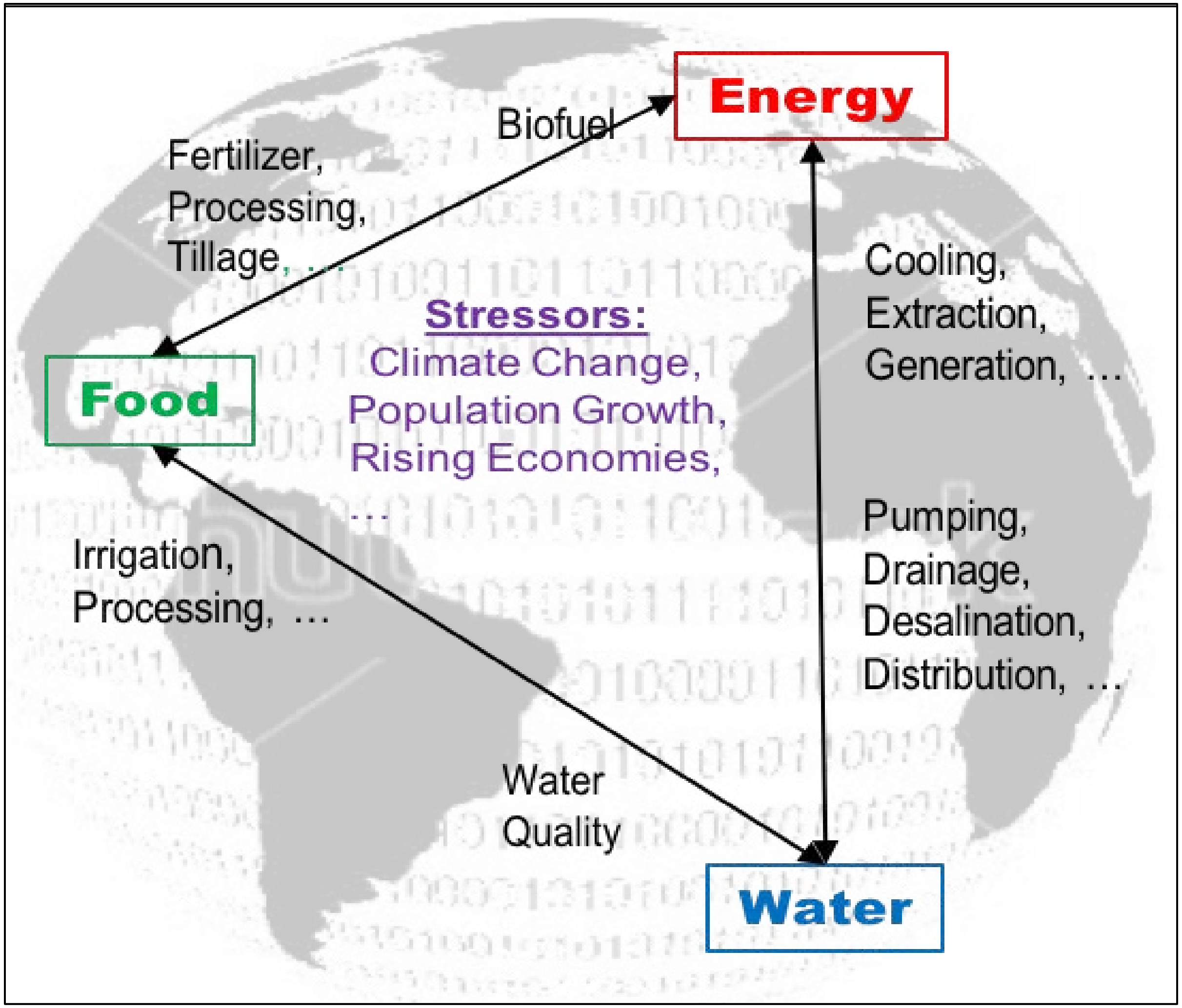
# Alternatives to Piecemeal Approach

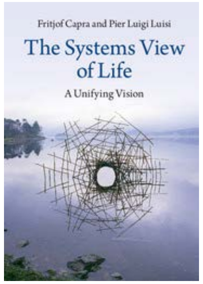
- Systems, Holistic, Integrated, Ecological, environmental traditions
  - Source: "Ahn AC et al, PLOS Medicine Open Access, July 2006"



# Interactions among Food, Energy Water Systems

- Piecemeal decisions in one affect the other
- Efficiency/ abundance in one reduces scarcity in other
- Stressed by Climate, Population, etc.
- Broader context





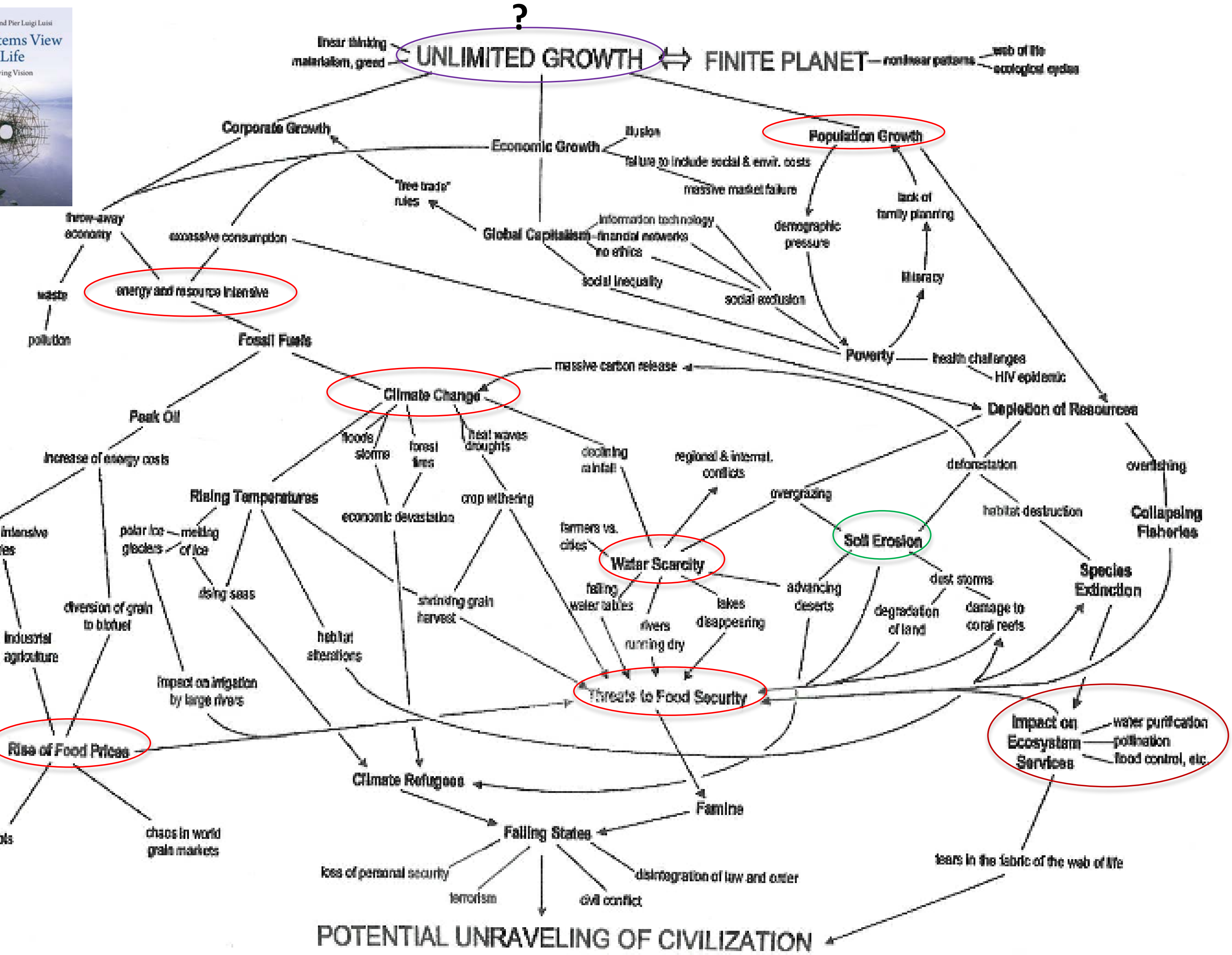



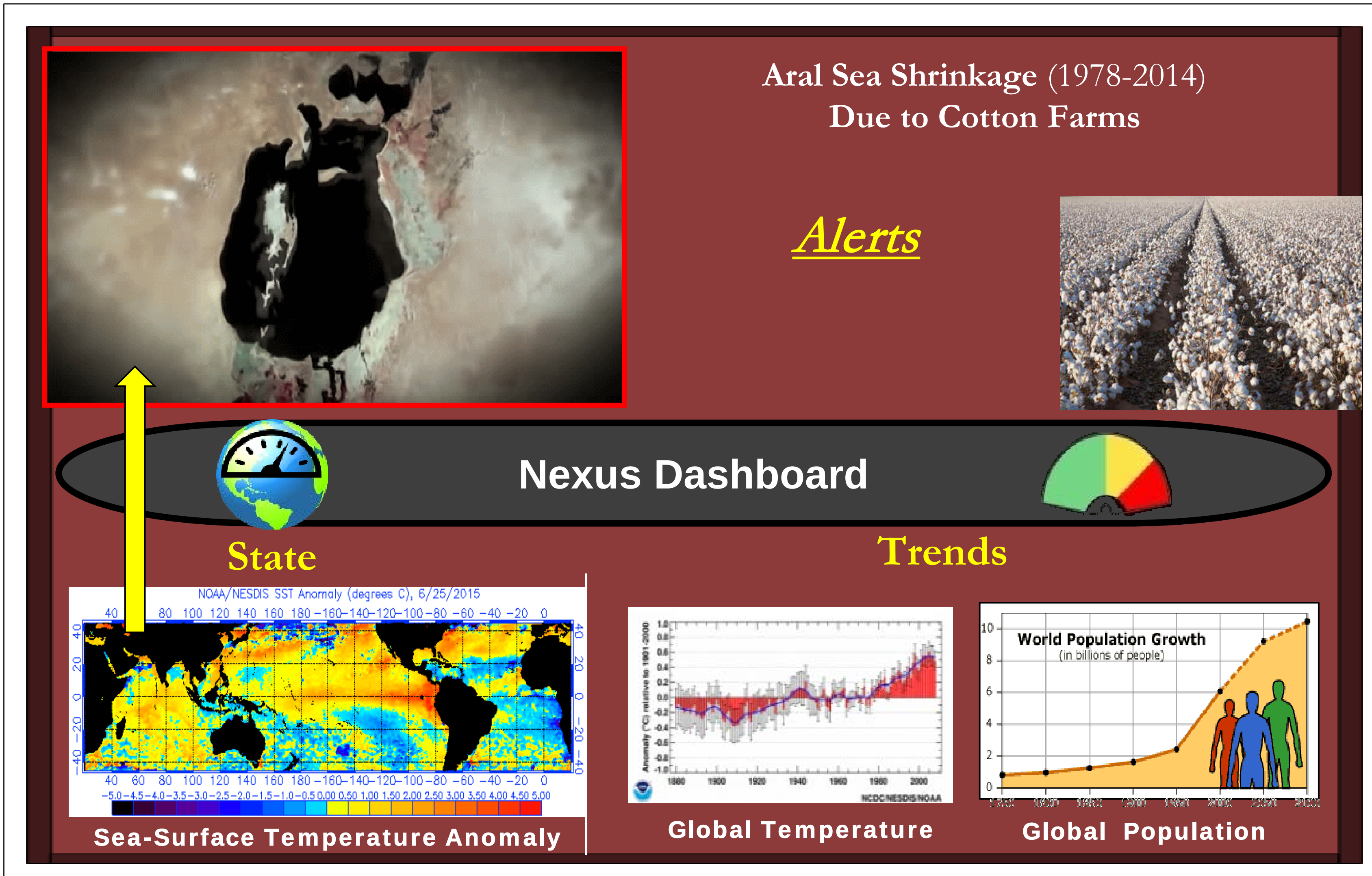

Figure 17.1 Interconnectedness of world problems (based on Brown, 2008).

# Outline

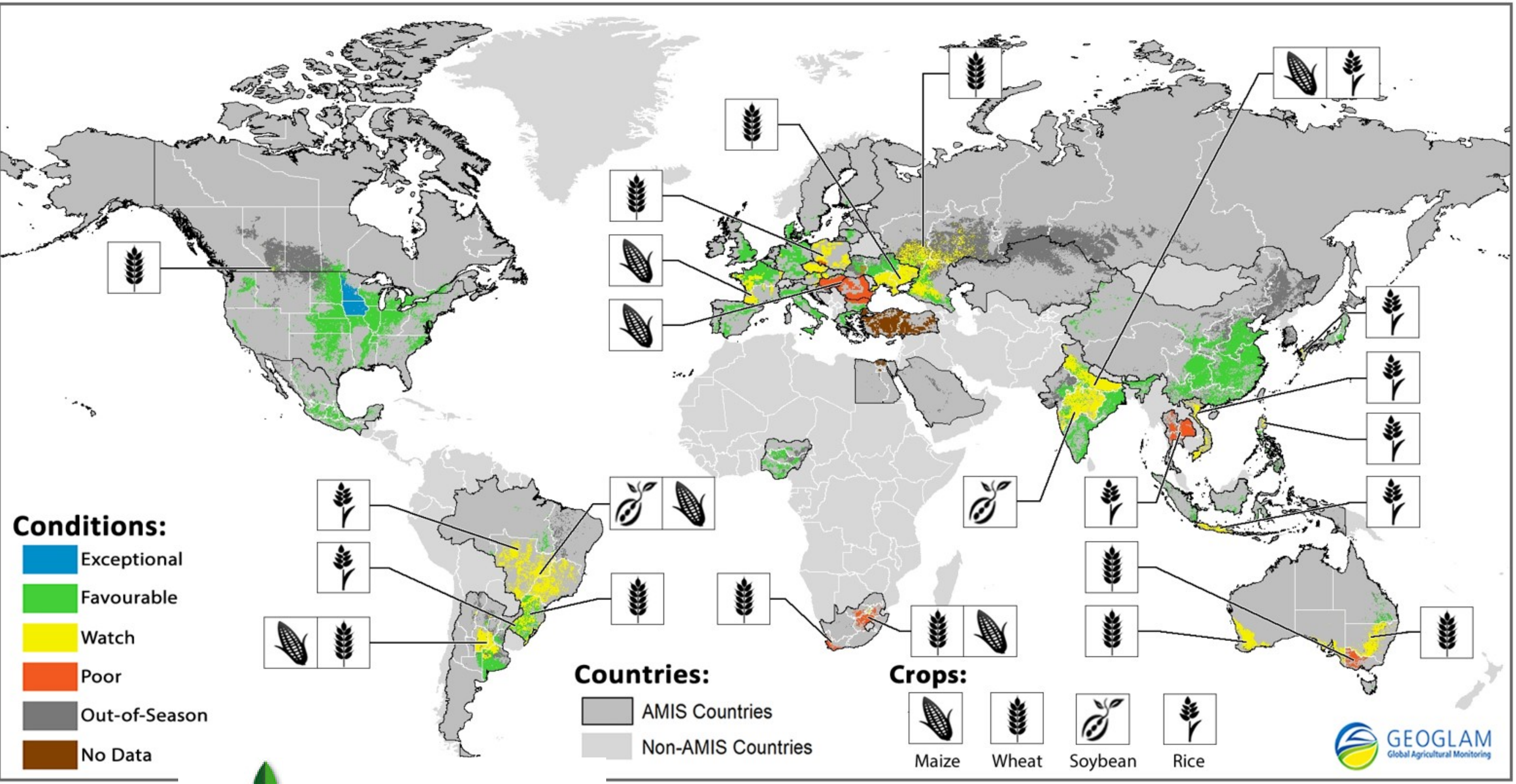
- F-E-W Nexus
- Role of Data & Data Science
  - Q? Which track of INFEWS Solicitation?
  - Monitor FEW resources, and trends to detect risks
  - Support Decisions & Policy making
  - Innovate to address nexus
  - Understand problems, connections, dynamics
- Impact of FEW Nexus on Data Science
- Next

# Monitor resources & trends to detect risks

Communicate with public and stakeholders



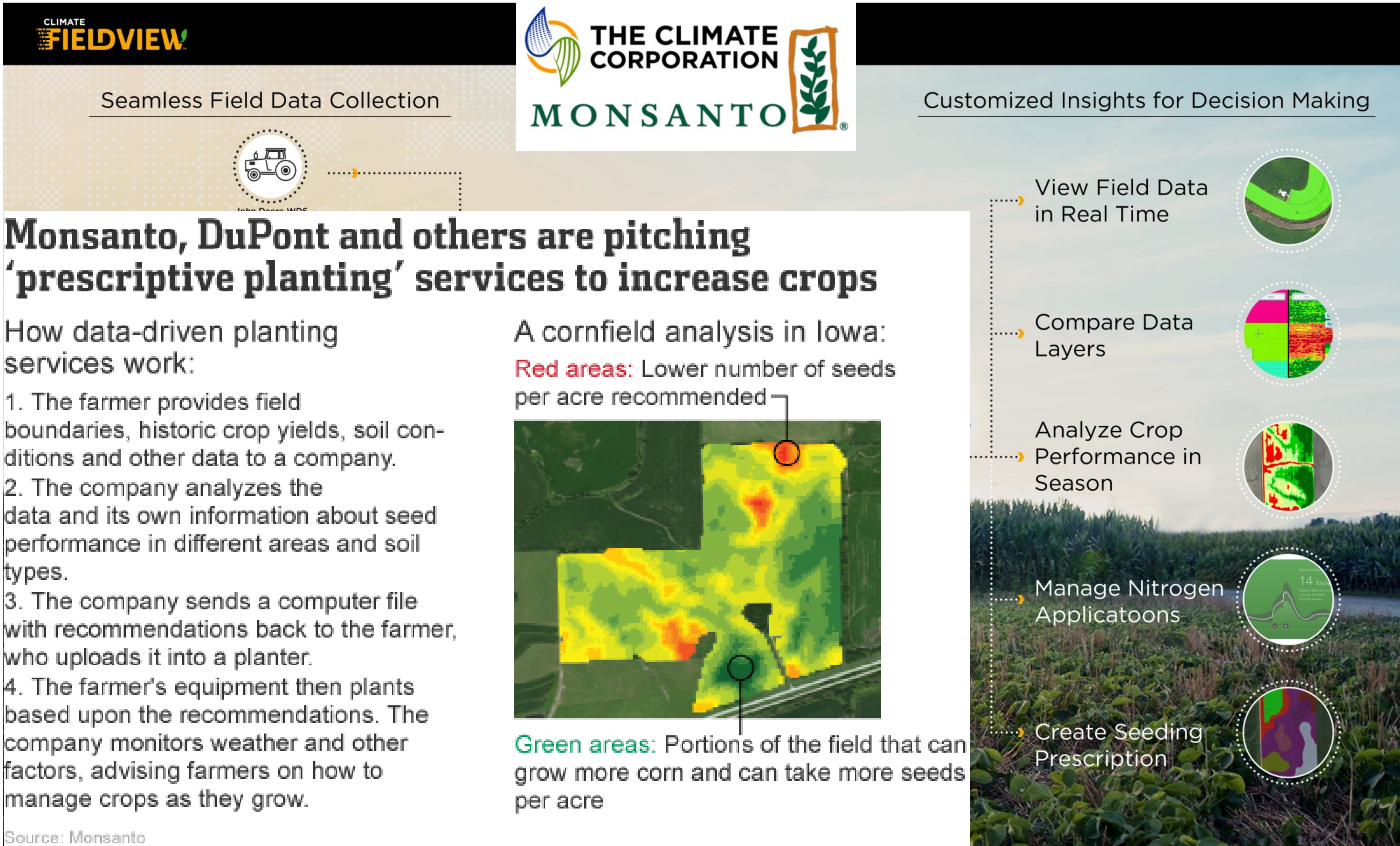
# Support (Global) Decisions and Policy Making




AMIS Agricultural Market Information System

GEOGLAM Global Agricultural Monitoring

# Support (Farm-level) Decisions and (Insurance) Policy




## Monsanto, DuPont and others are pitching 'prescriptive planting' services to increase crops

### How data-driven planting services work:

1. The farmer provides field boundaries, historic crop yields, soil conditions and other data to a company.
2. The company analyzes the data and its own information about seed performance in different areas and soil types.
3. The company sends a computer file with recommendations back to the farmer, who uploads it into a planter.
4. The farmer's equipment then plants based upon the recommendations. The company monitors weather and other factors, advising farmers on how to manage crops as they grow.

### A cornfield analysis in Iowa:

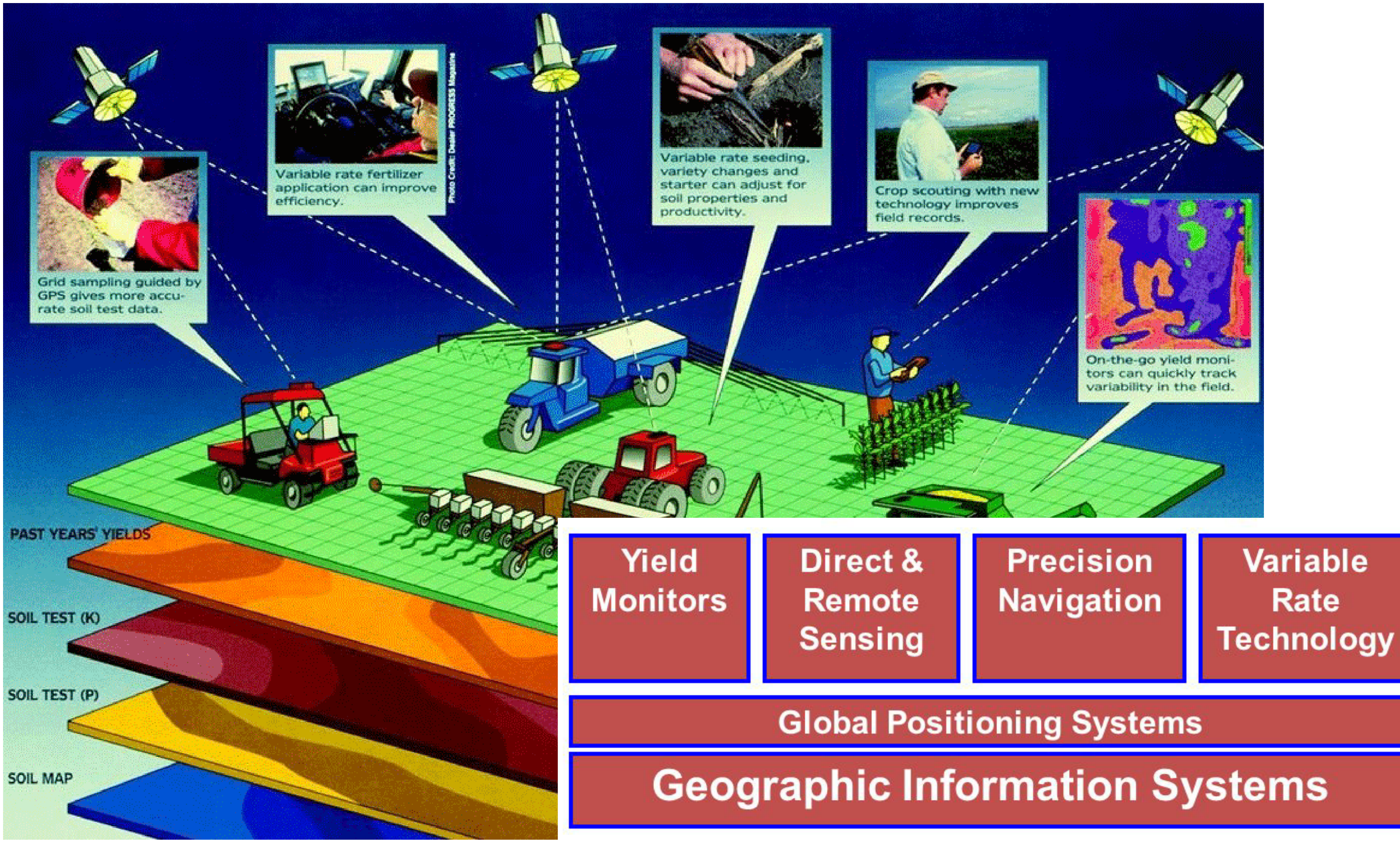
Red areas: Lower number of seeds per acre recommended

Green areas: Portions of the field that can grow more corn and can take more seeds per acre

Source: Monsanto

# Innovate to address nexus

- **Example**: **Precision Agriculture** to reduce run-off



# Deconstructing Precision Agriculture

## #AgInnovates2015

Wednesday, March 4, 2015
Reception | 5:00 to 7:00 pm

House Agriculture Committee Room,
1300 Longworth House Office Building,
Washington, DC

Think Moon landing.

Think Internet.

Think iPhone and Google.

**Think bigger.**

Come hear U.S. farmers, leading agriculture technology companies, and scientists tell how they work together to fuel U.S. innovation and the economy to solve this global challenge.

The event will exhibit three essential technologies of precision agriculture that originated from a broad spectrum of federally funded science: Guidance Systems and GPS, Data & Mapping with GIS, and Sensors & Robotics.

### Moderator

**Raj Khosla**, Professor of Precision Agriculture at Colorado State Univ.

### Farmers

**David Hula**, of Renwood Farms in Jamestown, Virginia

**Rod Weimer**, of Fagerberg Produce in Eaton, Colorado

**Del Unger**, of Del Unger Farms near Carlisle, Indiana

### Speakers

**Mark Harrington**, Vice President of Trimble

**Carl J. Williams**, Chief of the Quantum Measurement Division at NIST

**Bill Raun**, Professor at Oklahoma State Univ.

**Marvin Stone**, Emeritus Professor at Oklahoma State Univ.

**J. Alex Thomasson**, Professor at Texas A&M Univ.

**Dave Gebhardt**, Director of Data and Technology at Land O'Lakes/WinField

**Shashi Shekhar**, Professor at the Univ. of Minnesota

### RSVP

**http://bit.ly/1CoOYoa**

Hosted by
**the Congressional Soils Caucus**

**In partnership with**

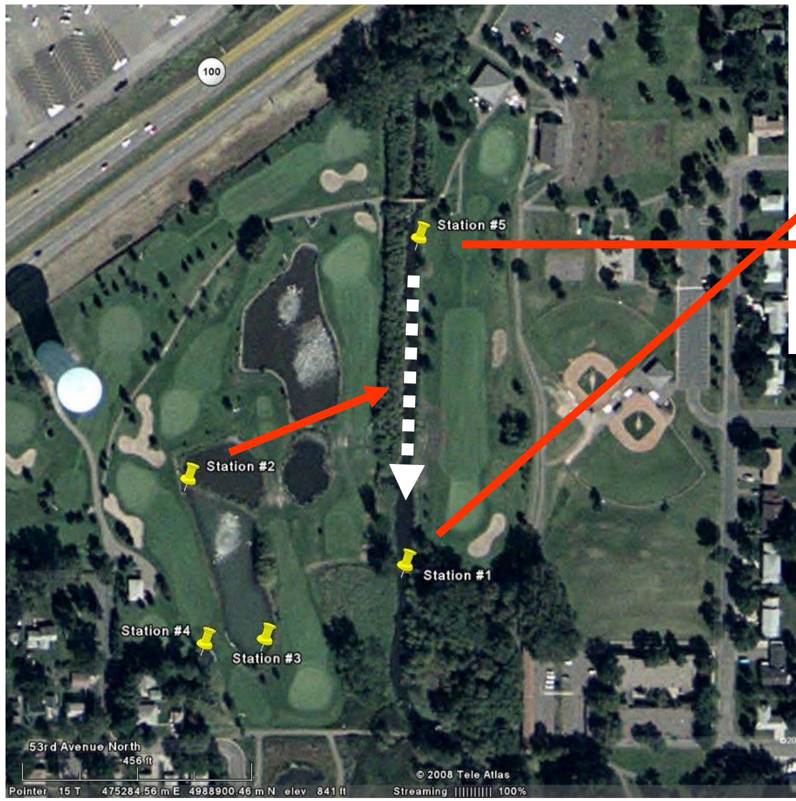
Agricultural Retailers Association
American Society of Plant Biologists
American Physical Society
American Society of Agronomy
Association of Equipment Manufacturers
Coalition for the Advancement of Precision Agriculture
Computing Research Association
CropLife America
Crop Science Society of America
PrecisionAg Institute
Soil Science Society of America
Task Force on American Innovation
Texas A&M AgriLife
Trimble
WinField

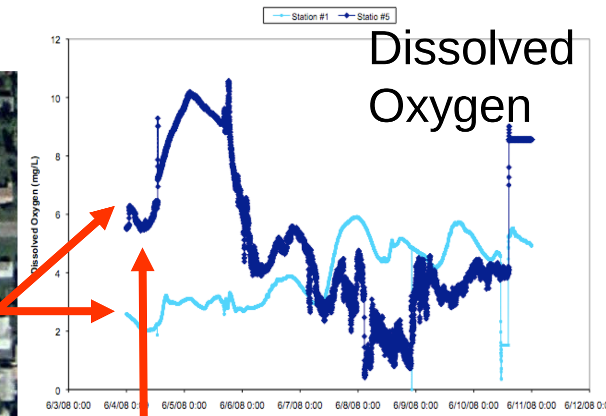
**This is about feeding the world.**

# **Understand** problems, **connections**, **impacts**

**Q? When and where do contaminants enter Shingle Creek, MN?**

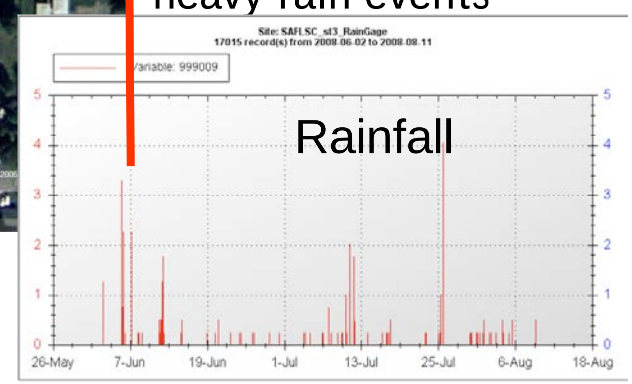
Ex. Spill, low O2


**Flow anomaly**

After consecutive heavy rain events




Details: J. M. Kang, S. Shekhar, C. Wennen, and P. Novak,
Discovering Flow Anomalies: A SWEET Approach, IEEE Intl. Conf. on Data Mining, 2008.

# Understanding Nexus: Data-Intensive Science

- (Causal) Theory
- (Controlled) Experiments
- Computational Simulations
- Data-Intensive Science


- What causes Algal bloom in lake Erie?
- What is fate of P in surface water?
- Does precision agriculture increase yield and reduce run-offs?
- Forecast climate, populations, …
- Q? What are side-effects of GMOs? Hydraulic fracturing?

# Outline

- F-E-W Nexus
- Role of Data & Data Science
- Data & Data Science Gaps
  - NSF INFEWS Data Science Workshop (Oct. 2015)
  - Data Gaps
  - Data Science Gaps
- Next

# NSF INFEWS Data Science Workshop

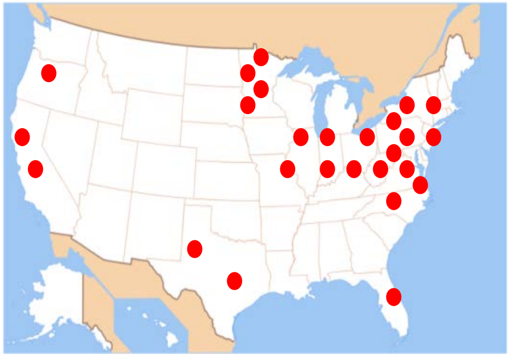
- **Goals**
  - Design compelling visions, Identify gaps
  - Develop a research agenda

- **Location:** USDA NIFA
- **Dates:** Oct. 5th-6th, 2015
- **Co-organizers:** Shekhar, Mulla, Schmoldt
- **URL:** www.spatial.cs.umn.edu/few

- **55 Participants** (Data-driven FEW & Data Sciences)

| Gov. | Aca. | Industry |
|---|---|---|
| 26 | 24 | 5 |

| Food | Energy | Water | DataSc. |
|---|---|---|---|
| 14 | 10 | 11 | 20 |



# Participants

| Data Science | Names |
|---|---|
| Data Collection | Thomas G. Dietterich<br>Tom Shapland |
| Data Exploration, Management | Chandra Krintz<br>Dieter Pfoser<br>Hanan Samet<br>Goce Trajcevski |
| Data Extrapolation | Peggy Agouris<br>Chid Apte<br>Paul Gader<br>Vasant Honavar<br>Zico Kolter<br>Inna Kouper<br>Vipin Kumar<br>Sanjay Ranka<br>Raju Vatsavai |

| FEW | Names |
|---|---|
| Food | Parag Chitnis<br>Jason Hill<br>Rattan Lal<br>Lakshmi K. Matukumalli<br>Rachel Melnick<br>Rabi Mohtar<br>Sonny Ramaswamy<br>Susan Jean Riha<br>Paul Tanger<br>Moira Zellner |
| Energy | Noel M.Bakhtian<br>Robie Lewis<br>Bob Vallario<br>Tamara Zelikova |
| Water | Richard Alexander<br>Brad Doorn<br>Alan Hecht |
| Cross-cutting | Zachary Hayden<br>Luis Tupas,<br>Ariela Zycherman |

# Day 1: Agenda

- **Pull Panel**:
  - Data-Driven FEW Nexus Science and Application Innovations
- **Pull Break Out**:
  - Challenges & Opportunities in Data-Driven FEW Nexus Science & Application Innovation
- **Push Panel**:
  - Science Research Needs to Understand and Innovate for FEW Nexus
- **Push Break Out**:
  - Discussion for Data Science Research Needs

# Panel Presentations

- **Pull Panel:** Data-Driven FEW Nexus Science and Application Innovations
    - FEW Nexus Overview (with life-cycle analysis): Rabi Mohtar (TAMU)
    - **Energy - Water** Nexus: Bob Vallario (USDoE)
    - FEW: A NIFA Perspective: Sonny Ramaswamy (NIFA)
    - **Water - Food** Nexus: Rich Alexander (USGS)
    - **Energy - Food** Nexus: Louis Tupas (NIFA)
    - Drivers of FEW Nexus: Rattan Lal (OSU)

- **Push Panel:** Data Science Research Needs to Understand & Innovate for FEW Nexus
    - Data Science Challenges in **Sustainable Energy**: Zico Kolter (CMU)
    - Open-Source **Precision Agriculture** and Analytics Driven Decision Support: Chandra Krintz (UCSB)
    - **Machine Learning** Challenges: Thomas Dietterich (Oregon U)
    - **Trustworthiness** and Sustainability: Data Science for FEW Nexus in the Developing Regions: Inna Kouper (Indiana U)
    - **Informatics** Challenges: Vasant Honavar (Penn State)
    - **Remote Sensing** and Water: Brad Doorn (NASA)

# Break Out Sessions

**Pull Break Out:** Data-Driven FEW Nexus Science and Application Innovations: Challenges & Opportunities

- Food-Water Nexus,
- Water-Energy Nexus,
- Energy-Food Nexus,
- Cross-cutting (FEW Nexus)

**Questions**:

- Q1. What are most pressing societal needs?
- Q2. What are the data needs?
- Q3. What are promising quantitative methods?
- Q4. What innovations may arise from using the quantitative methods identified in Question 3?

**Push Break Out:** Discussion for Data Science Research Needs

- FEW Nexus Data Collection
- FEW Nexus Data Exploration
- FEW Nexus Data Extrapolation
- Cross-cutting

**Questions**:

- Q5. What are data-science needs?
- Q6. What are limitations of current data science methods?
- Q7. List promising approaches to address the limitations identified in
- Q8. Recommend a couple of data science research topics.

# Day 2: Workshop Agenda

- **Synthesis Panel**:
  - Summarize Day 1 Discussions
  - into a Small Number of Potentially-Transformative Data Science Research Topics Needed to Understand and Innovate for FEW Nexus
  - Develop Consensus
- **Synthesis Break Out 1**:
  - **Quad-Chart** Potentially-Transformative Data-Science Research Topics
- **Synthesis Break Out 2**:
  - **Detail** Potentially-Transformative Data-Science Research Topics

# Sample Domain Context Goals

- Increase efficiency and sustainability of farming
- How can consumer behavior be changed to create more sustainable FEW systems?
- Sustainability and productivity of soils
  - Restoration of degraded soils and ecosystems
- Impacts of climate change on FEW systems
- FEW strategies for mitigation and adaptation to climate change

# Outcomes: F-E-W Nexus Data Gaps

- Water – No global water quantity & quality census, e.g., aquife
- NewYork Time, 3/16/2017: Charles Fishman: Water is broken, Data can fix it.
- Energy – No global energy census
- Food - consumption data is sparse

- Interactions – aquifer recharge from surface water
- **Need**: a nexus data community (BD FEW Spoke)
- **Action**: Midwest Big Data Hub launching a Registry for nexus datasets

- Varied data collection (e.g., aquifer withdrawal meter in TX & CA)
- Heterogeneous data format (e.g., raster climate data, vector population)
- Lack of integrated datasets

# Outcomes: Data Science Gaps

1. Methods to help stakeholders reach consensus on FEW issues
   - Social science methods: scenario-based discussion, design exercises, etc.
   - Computational tools: visualization, explainable/interpretable models, interactive simulation and optimization

2. Spatio-temporal modeling
   - Dealing with data collected multiple spatial, temporal scales,
   - missing values

3. Fusion of multiple model types
   - Data-driven, process-driven, economic, etc.

4. Lifecycle thinking for the FEW Nexus
   - modeling human behavior, understanding indirect effects of perturbations, supply chains, opportunity costs, agent-based modeling

5. Data uncertainty, incompleteness, bias
   - provenance, conflict of interest, capturing and visualizing uncertainty

# Methodology for Guiding Stakeholders to Reach Consensus on FEW Issues

David Mulla, Moira Zellner, Ariela Zycherman, Zachary Hayden, Tamara Zelikova, Inna Kouper

## Innovative Data Science Research Questions

How can data/computational science integrate existing platforms for visualization of tradeoffs and fuller implications associated with various scenarios in complex FEW systems.

Better methods are needed to represent uncertainty in meaningful ways so stakeholders can incorporate this into the decision making process.

## Meaningful Graphic

## Intellectual Challenges and Hurdles

Connecting small scale process modeling with coarse scale data driven models that can be used for stakeholder decision making efforts.

Representing temporal evolution of complex landscapes is challenging.

How to identify data surrogates in the absence of data?

## Transformative Potential

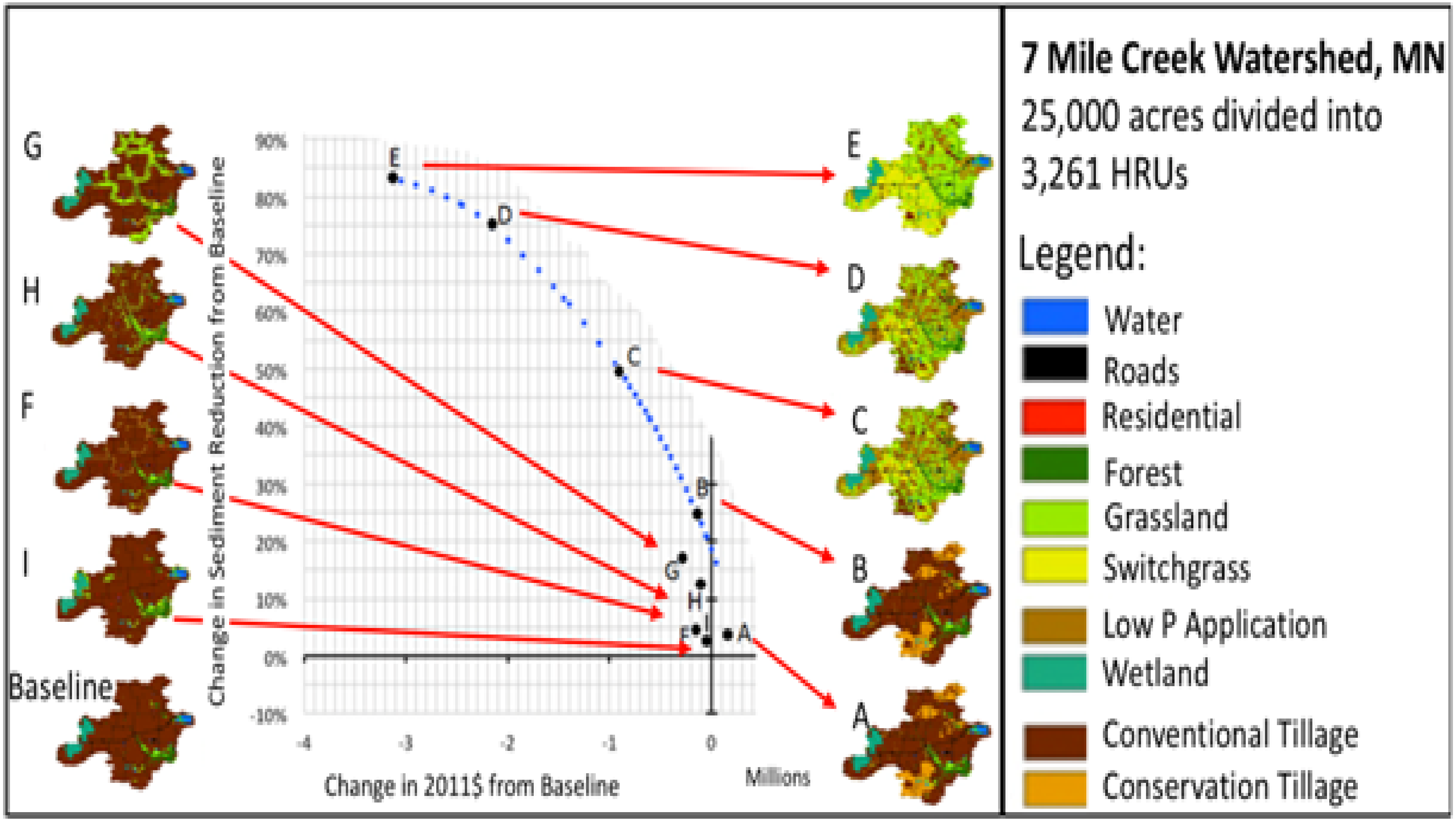
Enabling data informed stakeholder consensus and compromise. Stakeholders make better decisions based on understanding complex FEW systems from different stakeholder perspectives.

## Impact on Food, Energy, and Water Nexus

Scenario analysis and visualization could lead to significant shifts in stakeholder behavior and a more sustainable FEW system.

# Reducing Spatial Fragmentation in Optimization

Landscape geodesign via stakeholder collaboration (designs F, G, H, I) vs linear programming optimization approach (designs A, B, C, D, and E)



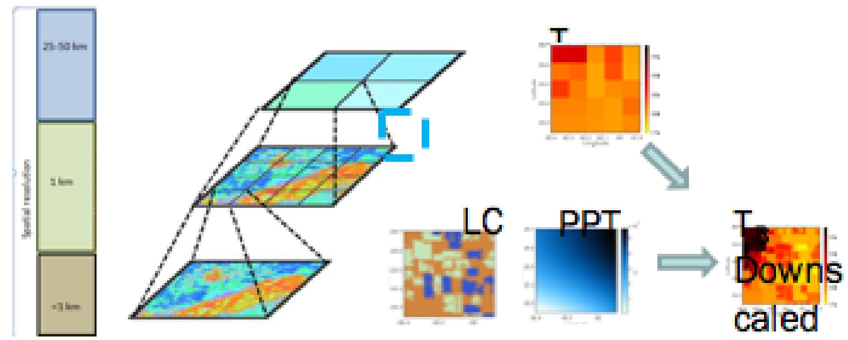
# Potentially-Transformative Data-Science Challenge/Opportunity

T. Archer (NOAA), H. Samet (Maryland), S. Riha (Cornell), T. Shapland (Tule Technologies),
D.Pfoser (GMU), P. Agouris (GMU), C. Apte (IBM-Watson), S. Ranka (UFL), Brad Doorn (NASA), R. Vatsavai (NCSU)

## Innovative Data Science Research Questions

1. What is the spatiotemporal grid that can effectively model multi-sensor, multi-resolution, and multimodal data?
2. How do you collect and fuse data and its provenance?
3. How do you leverage the open software ecosystem for processing big spatiotemporal data?
4. How do you find answers to complex queries such as spatiotemporal causes, events and changes?
5. How do you allow interaction of experts and stakeholders in the data collection and modeling process?

## Meaningful Graphic




- Spatial Correlations
- Hierarchical Constraints
- Parallel Computing

## Intellectual Challenges and Hurdles

1. Develop techniques for leveraging spatiotemporal correlations and causations
2. Develop techniques for merging multi-resolution, multi-model data
3. Develop methods which deal with uncertainty
4. Develop big data techniques for scalable parallel and distributed processing.
5. Develop ML techniques for solving complex queries and allow expert interaction

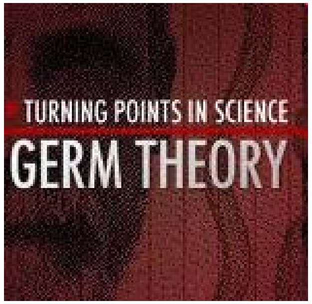
## Transformative Potential

1. Allow scientists to understand the relationships between FEW systems
2. Better understanding of impact of climate change on FEW applications
3. Participation of a wider community of stakeholders in data collection and information dissemination

## Impact on Food, Energy, and Water Nexus

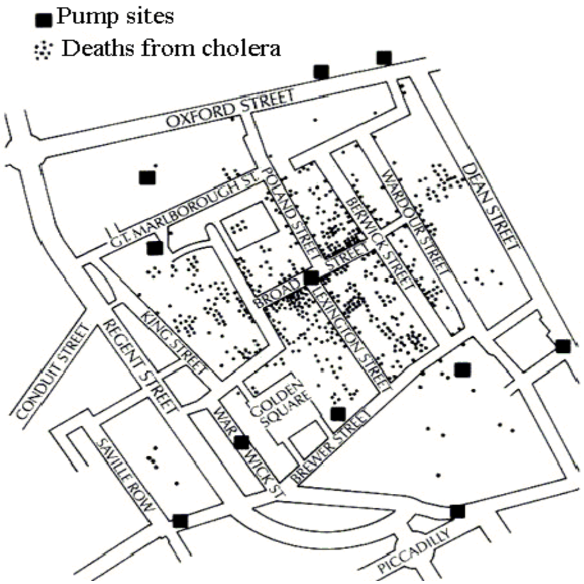
1. More environmentally rationalized use of our resources
2. Sustainability and productivity of soils
3. Empower stakeholders in the decision making processes

# Spatio-temporal (ST) Data Mining




- ST Data Mining?
  - Identifying interesting, useful, non-trivial patterns
  - in large ST datasets despite
    - Costly false positives, sparse data, auto-correlation, physics,
    - Non-isotropic non-Euclidean space

- Pattern Families
  - Hotspots, e.g., 1854 London Cholera
  - ST discontinuities, e.g., change
  - Co-locations, Tele-connections
  - Predicting location, trajectories, …



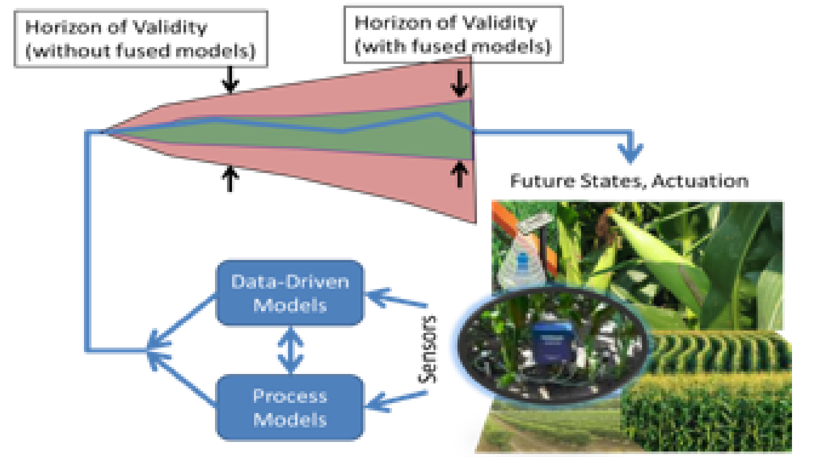

**Details:** (a) Spatiotemporal Data Mining: A Computational Perspective, ISPRS Intl. Journal on Geo-Information, 4(4):2306-2338, 2015.
(b) Identifying patterns in spatial information: a survey of methods, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery , 193-214, 1(3), May/June 2011.

# Fusion of Data-Driven and Process Models

Richard Alexander, Tom Dietterich, Vasant Honavar, Chandra Krintz, Paul Gador, Goce Trajcevski

**Innovative Data Science Research Questions**

- How can we combine data-driven and process models to
  - improve the accuracy of measurements
  - improve state estimation for process models
  - quantify and reduce uncertainty in forecasts
  - lengthen the validity horizon of forecasts
- How can we use data to characterize the relationships between different models at different scales
  - to optimize model ensembles, assemblies, couplings
  - match the spatiotemporal resolution of the models to the resolution of the FEW questions




**Intellectual Challenges and Hurdles**

- Lack of formalisms for
  - modeling complex measurement processes including privacy-preserving transformations
  - interfacing process models to data-driven models
- Gap between the scale of measurement and the scale of models
- Computational and approximation tradeoffs in solving coupled models
  - frequency and scale of model coupling and intercommunication
  - validity of rescaling for coupling
- Interpreting the results of coupled models
- End-to-end validation of coupled models

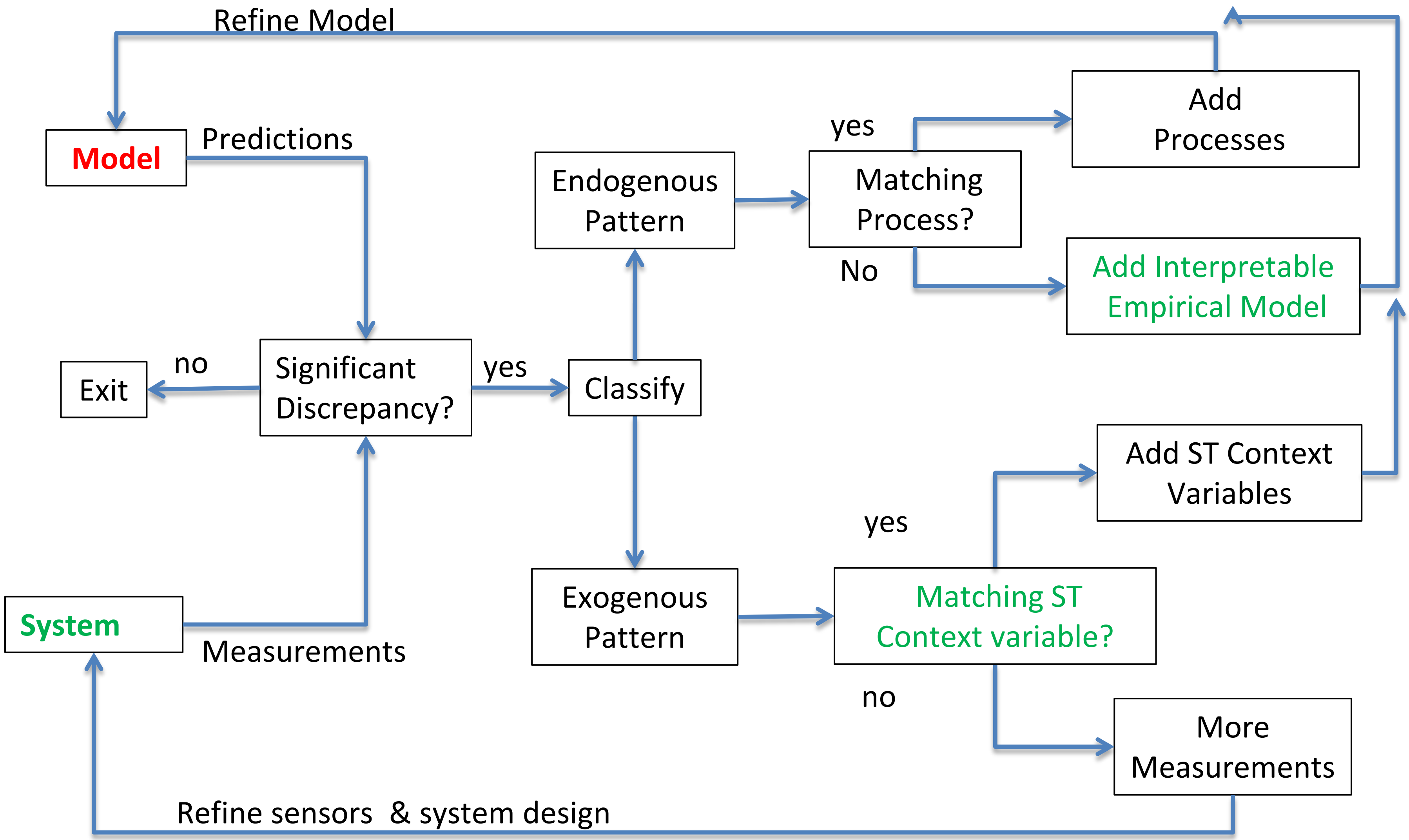
**Transformative Potential**

- Huge gains to be obtained by integration of multiple modeling paradigms
- Increased scope, accuracy, and robustness of coupled model systems

**Impact on Food, Energy, and Water Nexus**

- Increase efficiency and sustainability of farming
- Improve productivity of soils
- Improved risk models for insurance providers

# Understanding by Nexus of System Models



# Discussion

- Discussions during the workshop brought few topics.
  - Optimization and trustworthiness under uncertainty,
    - Develop optimal decisions under uncertainty.
    - Experiment optimization in support of the FEWS-related research in chemistry (materials, fuel cells) and biology.
  - Visualization of spatial and spatiotemporal data combined with other dimensions (e.g., time, flows, etc.)
  - Develop scalable cyber infrastructure to support spatiotemporal data and scalable algorithms for various data mining tasks (pattern mining, change detection, etc.)
    - Parallelization and support of spatial data as well as streaming data.
  - Data collection, curation, and sustained support
    - Data on supply chain, trade, socio-economic state,
    - Data on food production (yield, etc.), use, and waste,
    - Data on energy production, use, and waste
    - Data on water production, use, and waste.

# Outline

- F-E-W Nexus
- **Role** of Data & Data Science
- Data & Data Science **Gaps**
- Next
  - CCC Workshop on Computing and National Priorities
  - ACM SIGKDD Workshop on FEW
  - AGU session proposal

# Activities

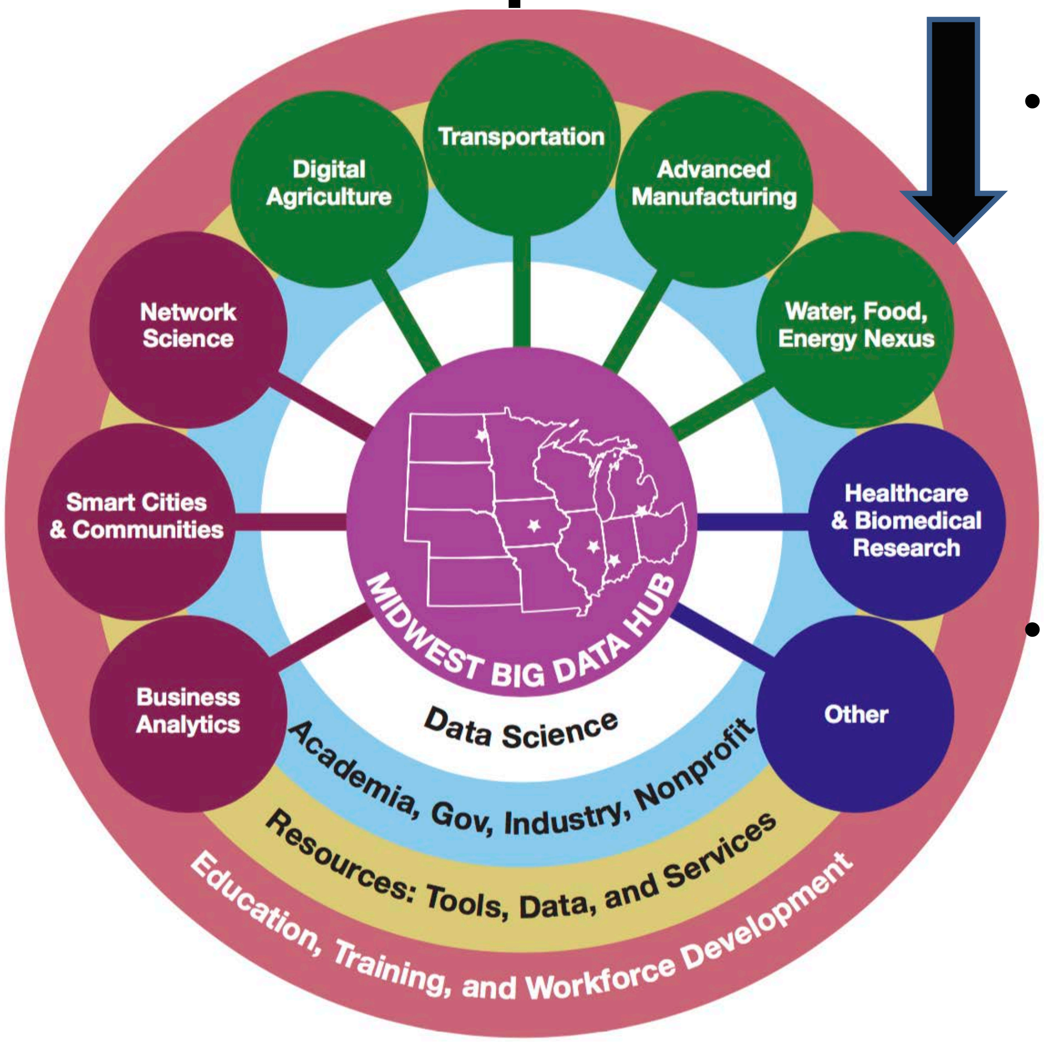
- **Recent**
  - Capitol Hill Presentation, House Ag Committee Reception on “Deconstructing Precision Agriculture” (3/15)
  - NSF Workshop to Identify Interdisciplinary Data Science Approaches and Challenges to Enhance Understanding of Interactions of Food Systems with Energy and Water Systems (10/15)
  - Symposium S-E2 (Towards a Food-Energy-Water Nexus Data and Data Science Community”, NCSE 2016 Conference (1/16)
  - NSF BDSpokes proposal Big Data Community for the Nexus of Food, Energy, and Water Systems (2/16)
- **Upcoming**
  - CCC/CRA Symposium on Computing Research: Addressing National Priorities and Societal Needs (5/16)
  - ACM SIGKDD Workshop on Data Science for Food-Energy-Water (8/16)
  - AGU session proposal (12/16)

# FEW Big Data Community Challenges

- Challenge:
  - Spatial nature of FEW datasets.
  - Popular Big Data tools (e.g. MapReduce, Spark) inadequate for Spatial Data

- State of Spatial Big Data Tools
  - Siloed with small user communities
  - Ex. Minnesota Population Center: TerraPop
  - Ex.: GABBS for Agriculture Model Inter-comparison(AgMIP)
  - Ex. CyberGIS, SpatialHadoop

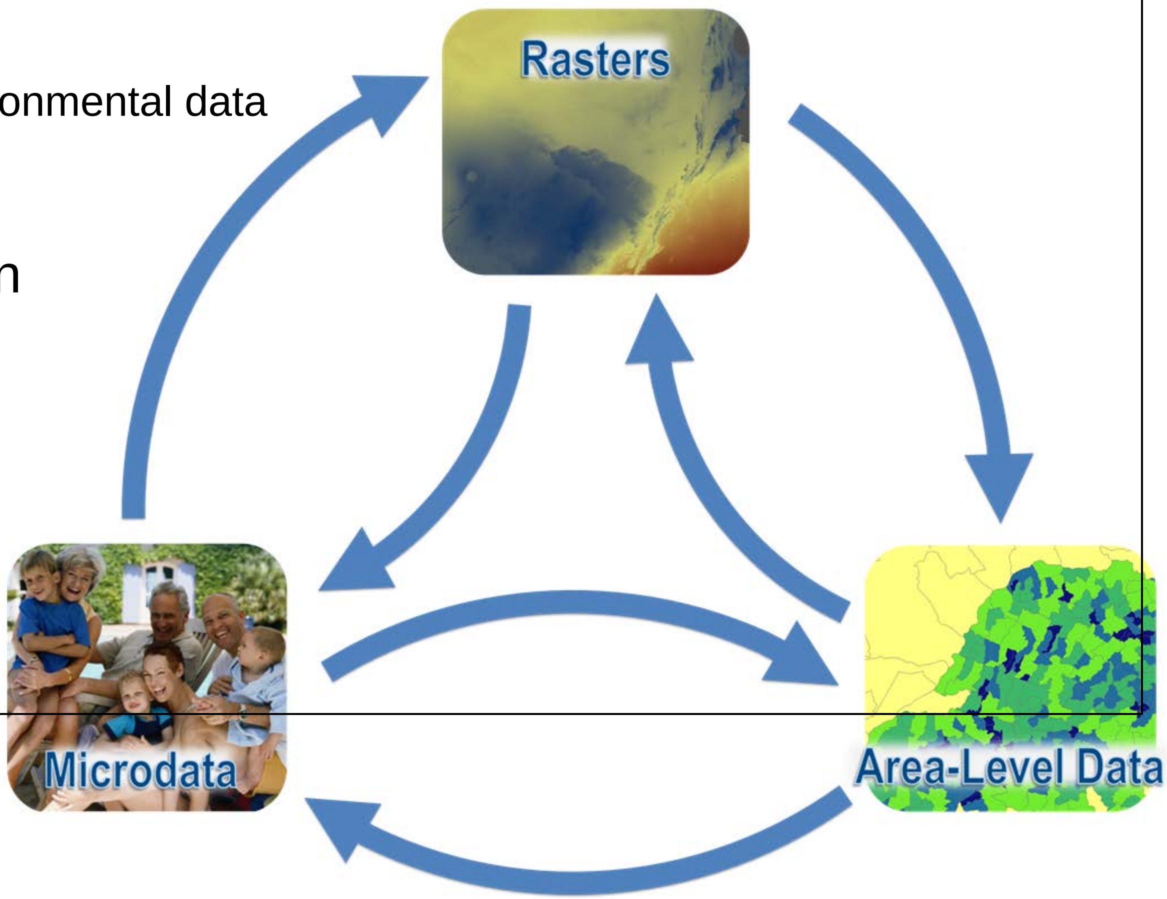
# FEW Spoke in Midwest Big Data Hub




- Midwest
  - Water: Largest freshwater reserves, e.g., Great Lakes.
  - Food: Leader in agricultural production, processing, transportation, distribution
  - Energy: Dominant Biofuel Supplier
- NSF Cross-Directorate Initiative
  - Research: Innovations for Food, Energy, Water Nexus (INFEWS)
  - Education: NRT solicitation listed INFEWS as a priority
  - Infrastructure & Community Building: BD Hub, BD Spoke

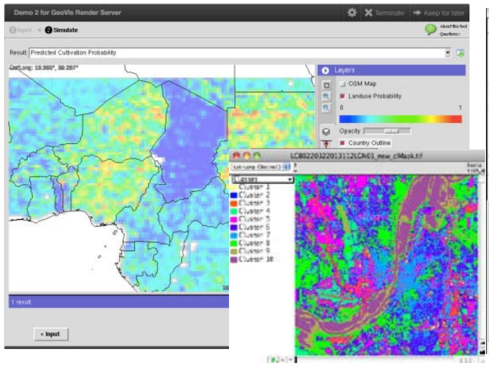
# Integrated Data: Terra Populus (A NSF Datanet)

- Goal: Enable investigation of human impacts and vulnerability
- Integrated data related to
  - agriculture, land cover, climate, and population
- Curated collection
  - global population and environmental data
  - linked to locations
- Location-based integration
  - Raster
  - Vector
  - microdata

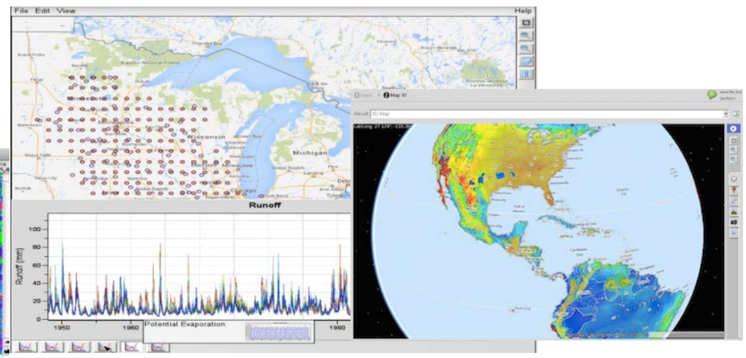


# G·A·B·B·S

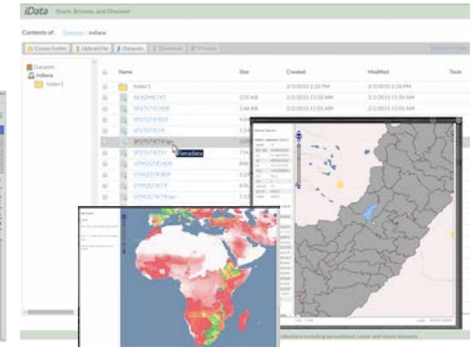
geospatial data analysis building blocks



Modeling & Analysis



Explore and visualize data

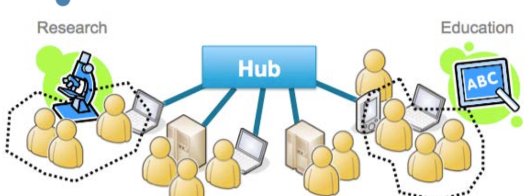


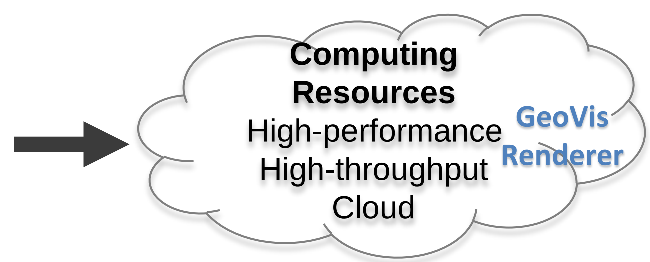
Share & Publish

## Geospatial data analysis building blocks for Agricultural Model Intercomparison (AgMIP)

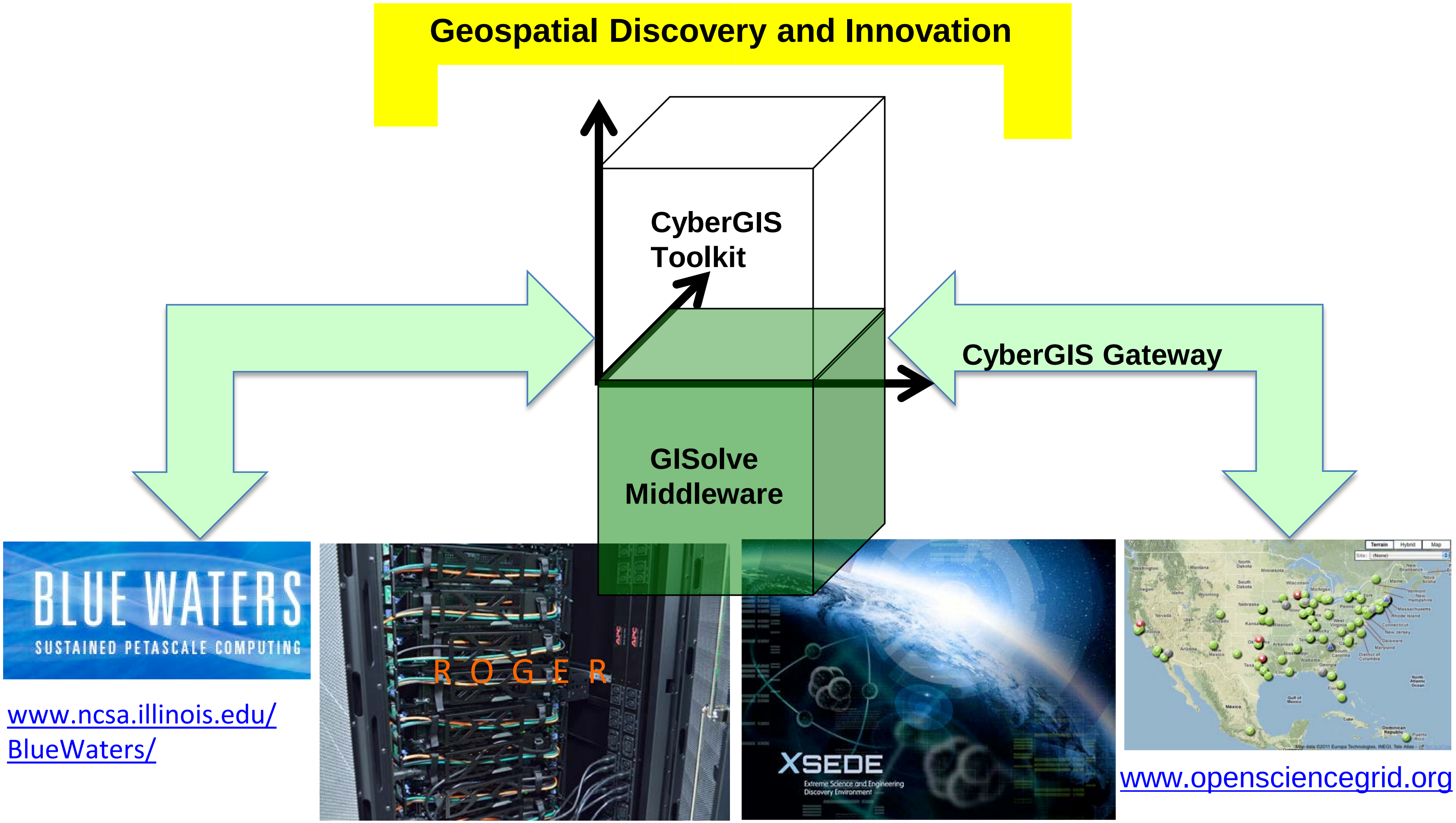
- Geospatial data processing, analysis and visualization support inside HUBzero
- Map library, Rapid Tool Development API (Rappture) with geospatial extension for developing online applications without web programming
- Online data management system linked to user tools
- DIY online interactive tool and data publishing (with DOI), publications linked to viewers and interactive tools

hubzero **Platform for Scientific Collaboration**



# Cyberinfrastructure: cybergis.org




www.ncsa.illinois.edu/BlueWaters/

cybergis.illinois.edu

www.xsede.org

www.opensciencegrid.org

# Spoke Mission & Vision

- Grow and Connect Communities
  - Producers and Consumers of FEW Nexus Data, Tool, Services