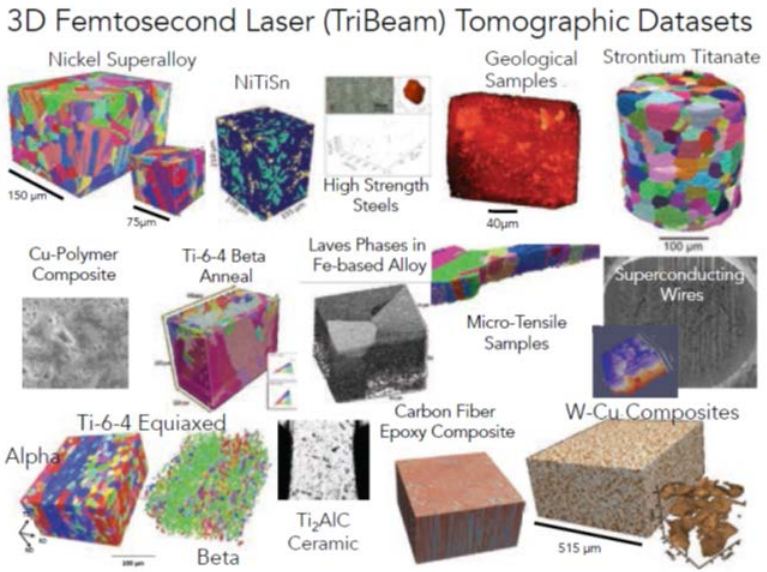
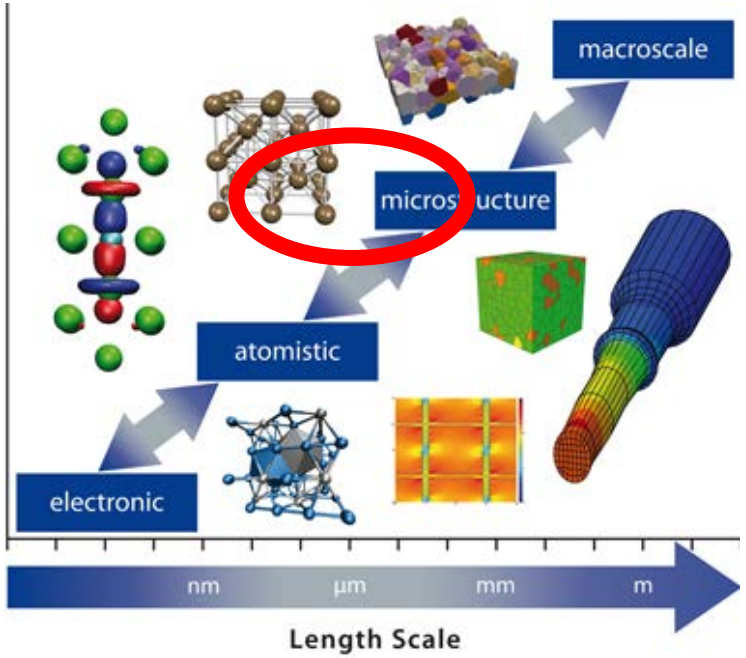
HARNESSING THE DATA REVOLUTION

*"Engage NSF's research community in the pursuit of **fundamental research in data science and engineering**, the development of a cohesive, federated, national-scale approach to **research data infrastructure**, and the development of **a 21st-century data-capable workforce**."*

MATERIALS LENGTH SCALES





- "ARRANGEMENT OF PHASES AND DEFECTS" – MAY BE COMPLEX
- RICH IN DATA FROM IMAGING PROBES
- DATA ANALYTICS WITH EXPERIMENT AND COMPUTATION TO SPAN THE MICROSTRUCTURE ⇔ PROPERTY CHASM

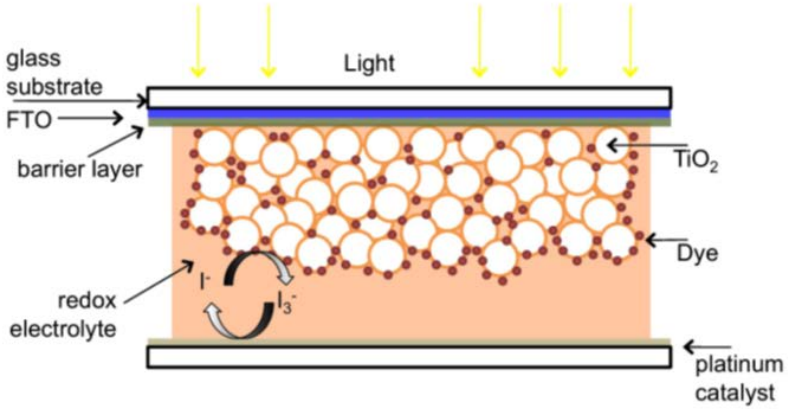
MATERIALS PROPERTIES ARE CONTROLLED BY STRUCTURE AT DIFFERENT SCALES

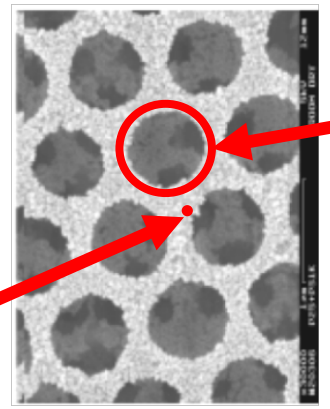
CHALLENGE:

- **DISCOVER MICROSTRUCTURE ⇔ PROPERTY RELATIONSHIPS**
- **CONTROL MICROSTRUCTURE ➔ DESIRED PROPERTY**

Dye Sensitized Solar Cells

Can Fossil Fuel Electricity Generation Price/Performance be achieved?





Nanopores

High surface area for dye

How can microstructure be optimized for maximum efficiency?

Apply concepts

- Algebraic topology and geometry
- Applied statistics
- Algorithms
- Graph theory

Understand topological interconnections, shapes, and dynamics

Investigate topological concepts

- Finite data,
- Approximations
- Noise
- Constraints

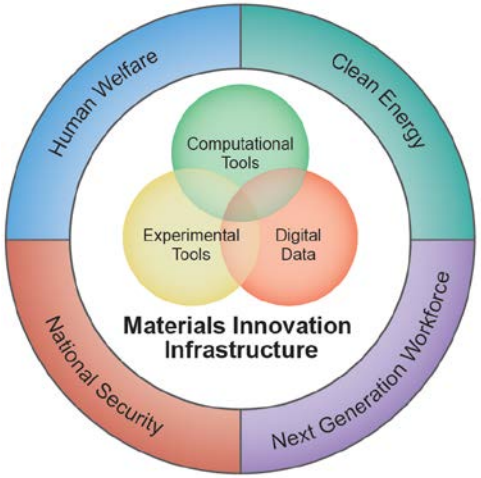
These are always encountered in real materials characterization!

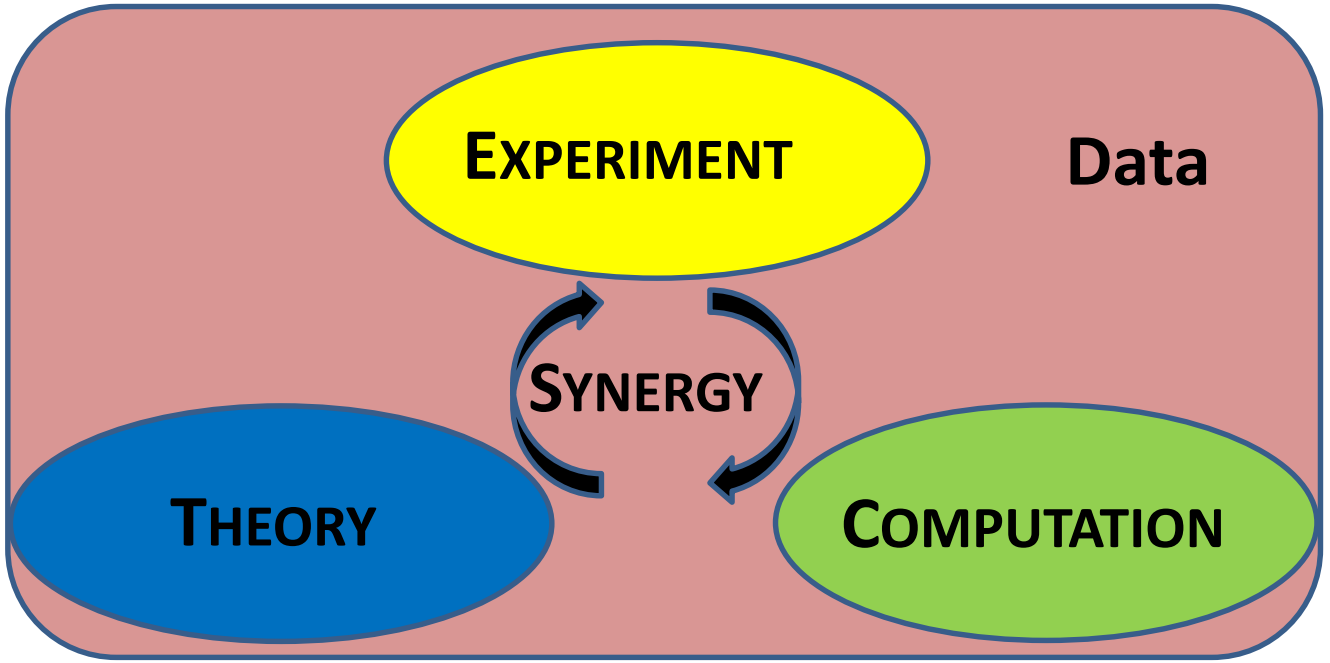
Data Science ⇔ Materials Science

synergy enables materials to device design

THE MATERIALS GENOME INITIATIVE

DISCOVERY-TO-MARKET IN LESS THAN HALF THE TIME AT HALF THE COST





A NEW PARADIGM FOR DISCOVERY:

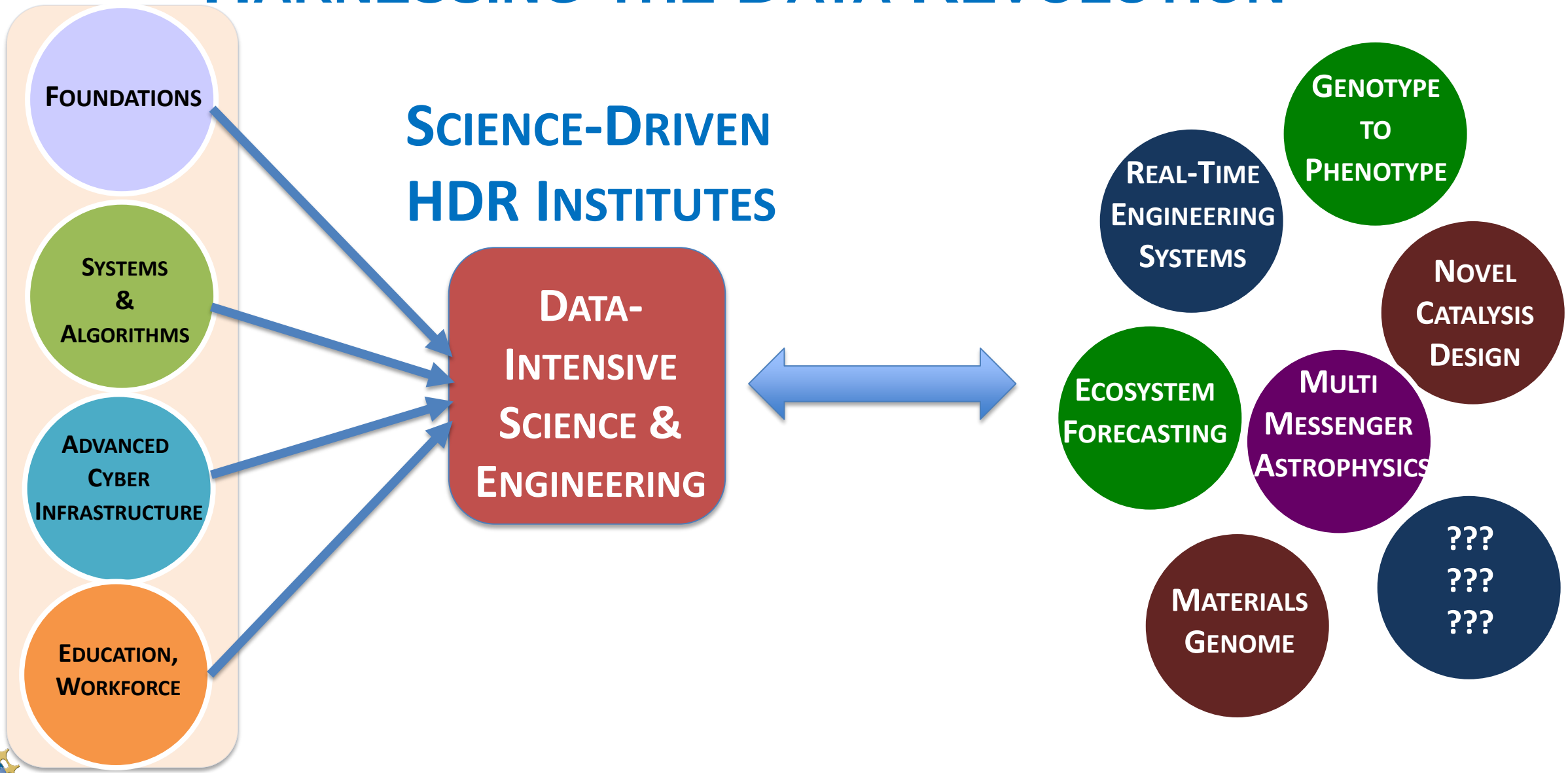
THE SYNERGISTIC INTERACTION AMONG

COMPUTATION, *DATA*, EXPERIMENT, AND THEORY

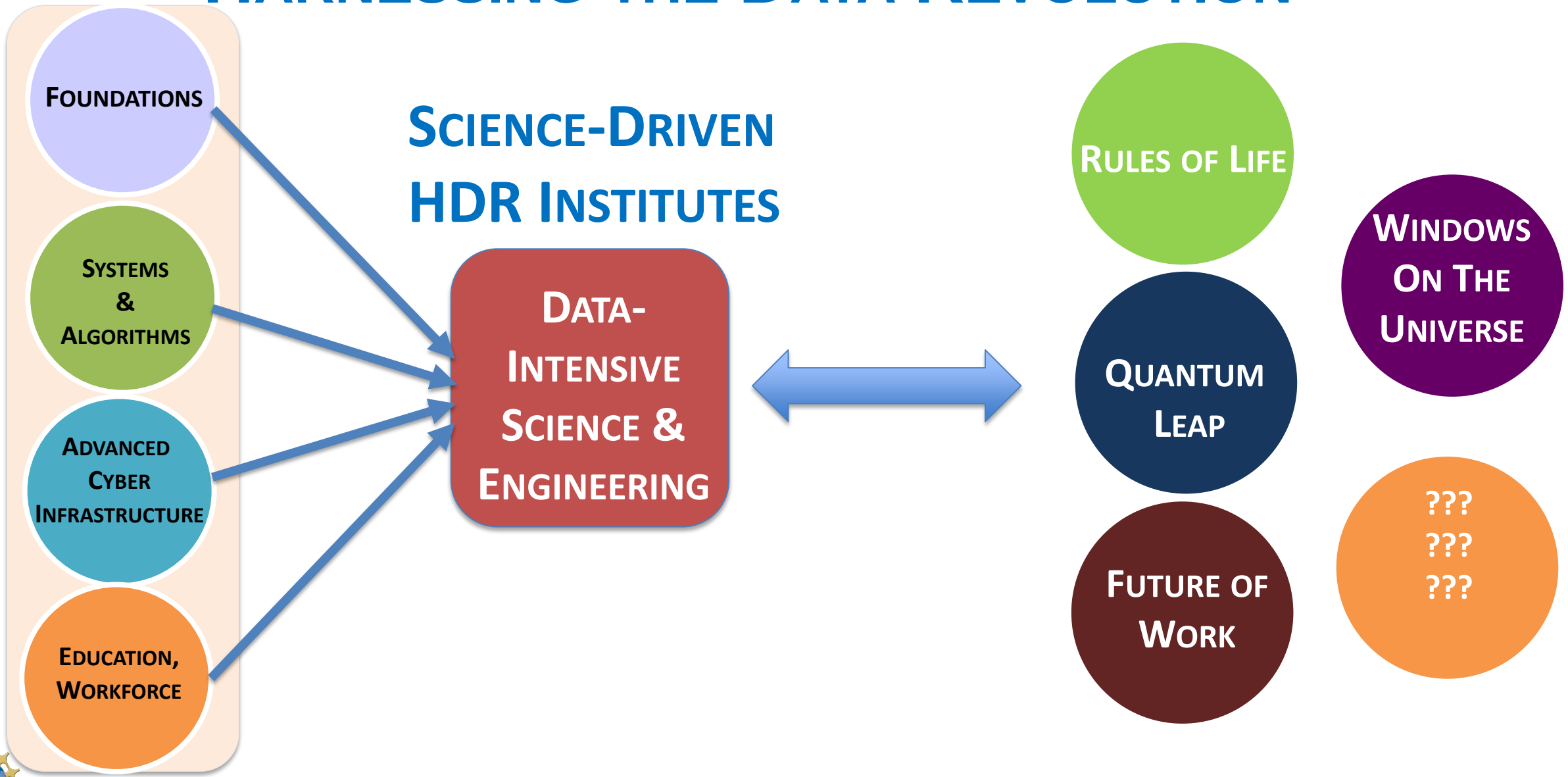
HDR Roadmap Has 5 Major Components

- Theoretical Foundations
- Systems Foundations
- Data-Intensive Research Across All S&E
- Data Cyberinfrastructure
- Education & Workforce Development

Harnessing the Data Revolution



Harnessing the Data Revolution



HDR Foundations

- TRIPODS: Transdisciplinary Research in Principles of Data Science
  - Collaboration among computer and computational scientists, statisticians and mathematicians to develop the principles of data science
- TRIPODS+X
  - Collaboration among domain research and TRIPODS projects, so that foundational approaches are informed by real Science & Engineering Problems

Foundations

HDR Systems and Algorithms

- Open Knowledge Network (OKN)
  - An open web-scale knowledge network of semantically-linked concepts and data
  - To foster research on a new generation of applications leveraging data, context, and inferences from data
- ModelCommons
  - Sharing and reuse of Machine Learning and other data-intensive models
  - Support for reproducibility and reuse (transfer learning…)

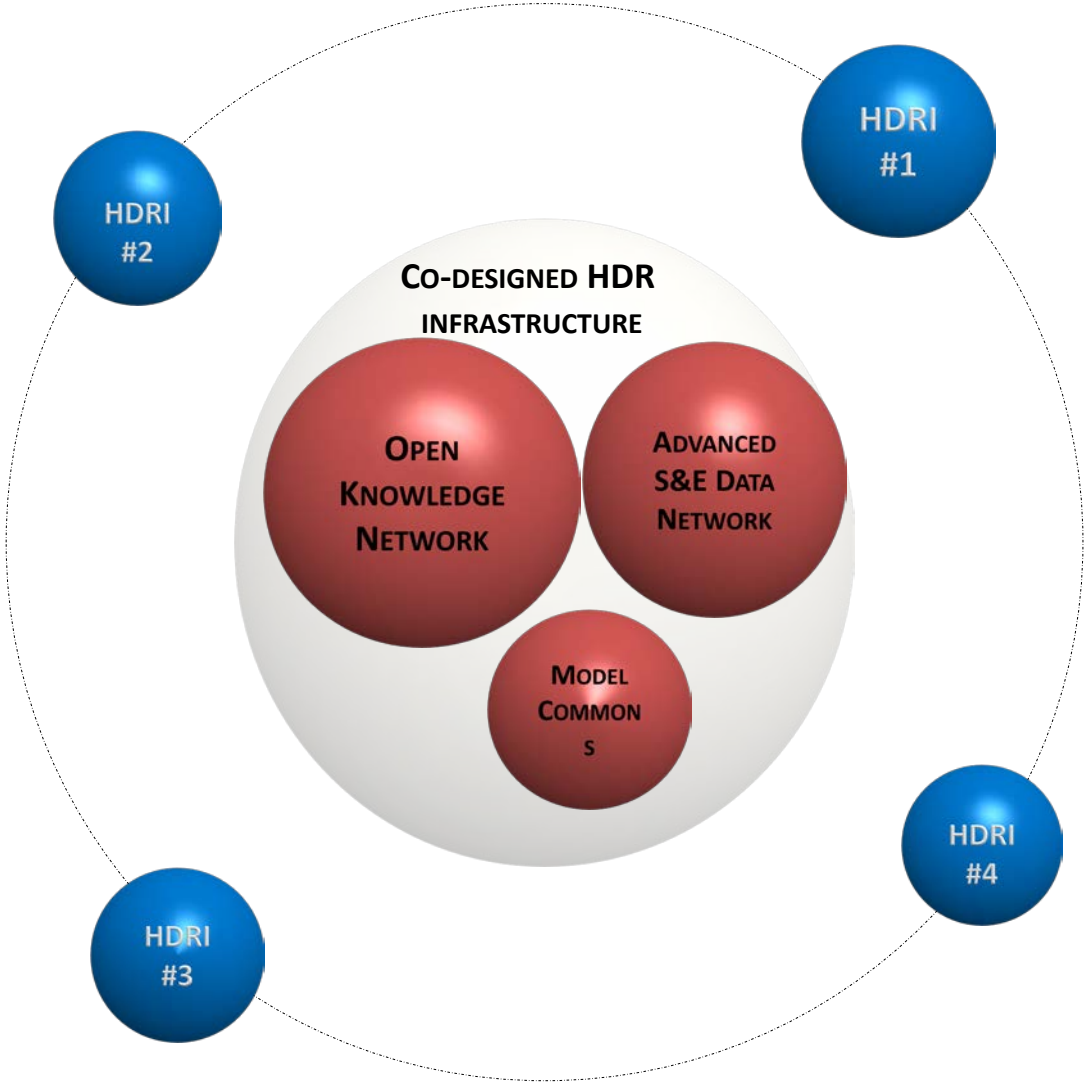
Systems, Algorithms

HDR Institutes: Coupling Science and Engineering Challenges with Data Challenges

Data-Intensive Science & Engineering

Data Challenges:

- Machine learning
- Data provenance
- Data heterogeneity
- Data security
- Data ethics
- Data storage & access



Science and Engineering Challenges:

- Near-term ecological forecasting
- Real-time sensing, learning, and decision making
- Climate, weather, hydrological, and hazard forecasting
- Novel materials and chemical designs
- Multi-messenger astrophysics

HDR – Building on Existing Programs

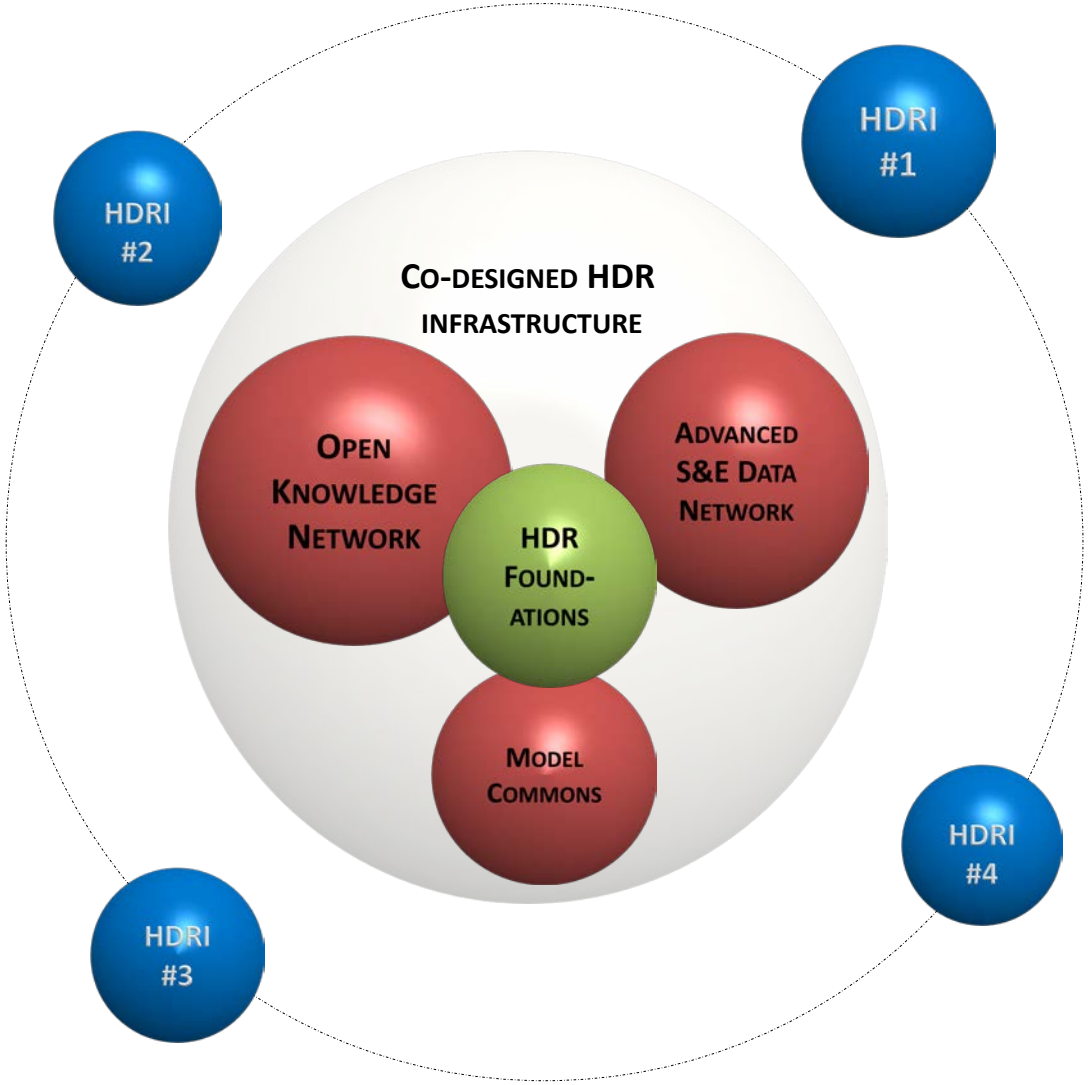
Data-Intensive Science & Engineering

HDR Themes:

- Theoretical and Systems Foundations
- Data Intensive Research
- Cyberinfrastructure
- Learning and Workforce Development

HDR Foundations:

- TRIPODS
- TRIPODS+X



Leverage CI Investments:

- CSSI – Cyberinfrastructure for Sustained Scientific Innovation
- SI2 – Software Infrastructure for Sustained Innovation
- DIBBs – Data Infrastructure Building Blocks
- EarthCube
- BIGDATA
- Big Data Hubs and Spokes
- ….

Education & Workforce Development and Evaluation

- **HDR Academy**
  - Catalog, collect, create education/training materials
  - HDR Postdocs, HDR Bootcamps
- **Data Science Corp connecting data scientists/science students to data science projects**
  - Special focus on data science programs at community colleges, 4-year colleges, MSIs, etc.
  - Data Science Corps Workshop, Dec 7-8, 2017, Georgetown University
- **Program evaluation**
  - Metrics for success, Evaluating convergence
- **Sociotechnical study**
  - Similar to work being conducted for the NSF Big Data Hubs

Current Status

- **TRIPODS**
  - 12 Phase I projects Funded
  - TRIPODS + X proposals in review
- **HDR Advanced Cyberinfrastructure**
  - Open Storage Network Award (June 2018)
- **Steering Committee and Working Group In Discussions**
  - Open Knowledge Network
  - Data Science Corps
  - ModelCommons
  - HDR Institutes

HDR Structure

- **Co-Chairs:** Chaitan Baru, CISE, Juan Meza, MPS
- **Steering Group :** Jim Deshler, BIO; Robin Wright, EHR; Fil Bartoli, ENG; Anjuli Bamzai, GEO; Manish Parashar, OAC; Daniel Sui, SBE
- **Working Group:** Peter McCartney, BIO; John Cherniavsky, EHR; Tony Kuh, Akbar Sayeed, ENG; Eva Zanzerkia, GEO; Daryl Hess, Lin He, Nandini Kannan, Slava Lukin, Angela Wilson, MPS; Amy Walton, OAC; Paul Morris, OIA; Charles Estabrook, OISE; Cheryl Eavey, Cassidy Sugimoto, SBE
- **Exec Secretary:** Vandana Janeja

THANK YOU

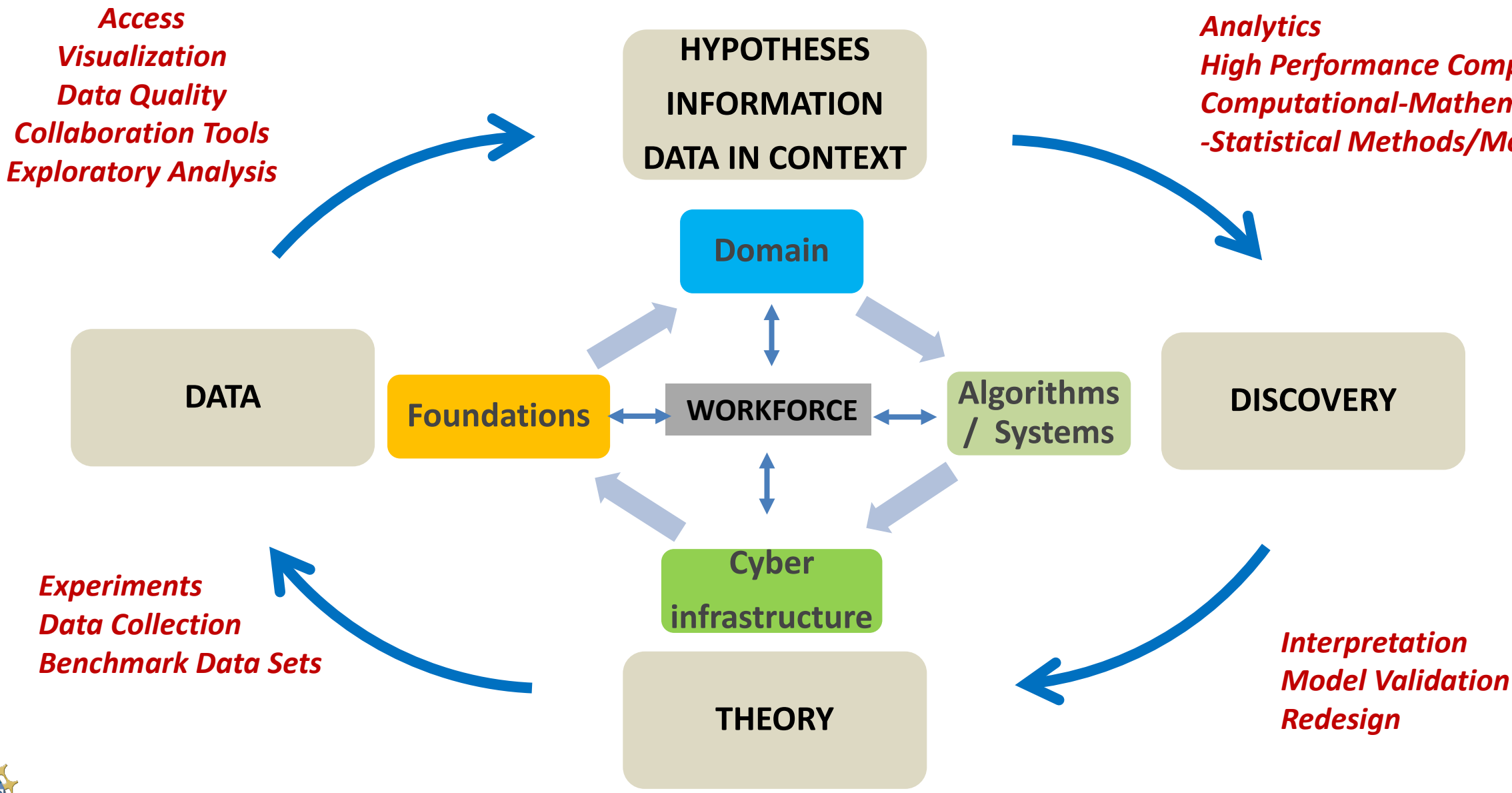
Extra Slides

Science-Driven HDR Institutes

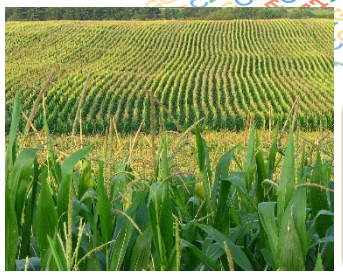
Data-Intensive Science & Engineering

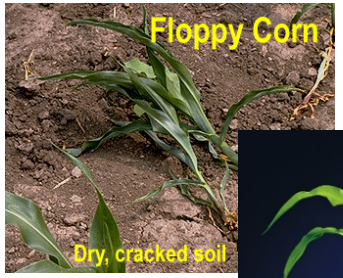
- Clear and compelling science- and engineering-driven goals
  - Enable significant progress within a 3-5 Year Time Period
- Convergent
  - Teams of domain scientists and computer scientists, mathematicians and statisticians
- Co-design of HDR "infrastructure"
  - Coordination with Other HDR Components: Foundations, Open Knowledge Network, ModelCommons, etc.
- Leverage other NSF investments:
  - Cyberinfrastructure: CSSI, SI2, DIBBs, EarthCube, etc.
  - BIGDATA, Big Data Hubs
- Enhance education, diversity, and public outreach

ENABLING AND ACCELERATING DISCOVERY: CONVERGENCE & CO-DESIGN



FROM GENOTYPES TO PHENOTYPES



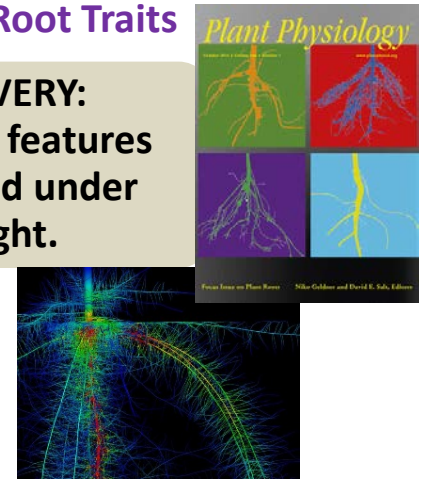


Access
Visualization
Data Quality
Collaboration Tools
Exploratory Analysis

HYPOTHESIS:
Bigger root systems => better water use and grain yield

Analytics
High Performance Computing
Models/Methods

Digital Imaging of Root Traits



DATA:
Genome Sequences
Trait Measurements
Environmental Data

Domain

Foundations

WORKFORCE

Algorithms/ Systems

Cyberinfrastructure

DISCOVERY:
Some root features affect yield under drought.

Experiments
Data Collection
Benchmark Data Sets

THEORY:
Root variables influence yield, but … How…? What if…?

Interpretation
Model Validation
Redesign