RENCI Collaborations and Consortia to Advance Data Science

Stan Ahalt

Professor, Department of Computer Science

Director, Renaissance Computing Institute (RENCI)

Director, Bioinformatics Core, NC TraCS, UNC-CH School of Medicine

ncds | THE NATIONAL CONSORTIUM for DATA SCIENCE

renci

RESEARCH \ ENGAGEMENT \ INNOVATION

RENCI's Mission

- **Be a leader in cyber-infrastructure (CI) research and development**
- **Be an essential CI partner for:**
  - Triangle university research teams
  - Research Triangle area industries
  - State of NC and federal agencies

IN ORDER TO address complicated multidisciplinary problems and research.

- **Data is central to all we do. Underlying theme: *Data to Decisions***

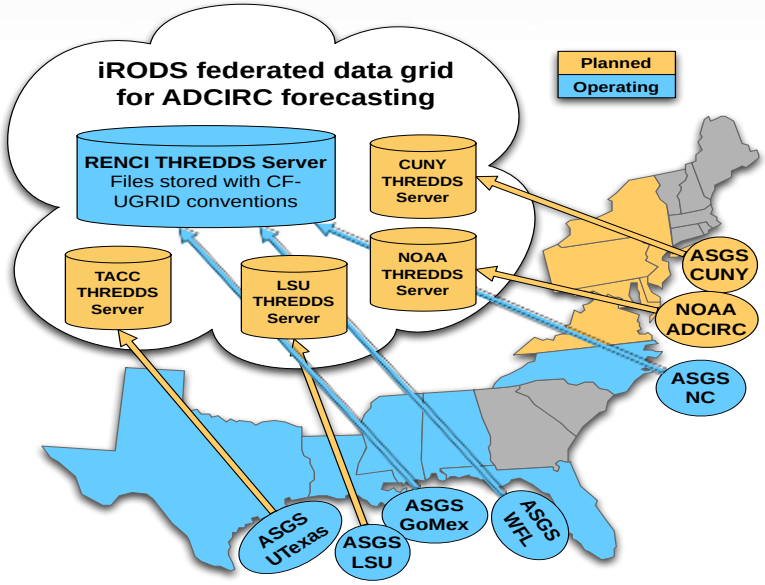
Collaboration is key to advance data science

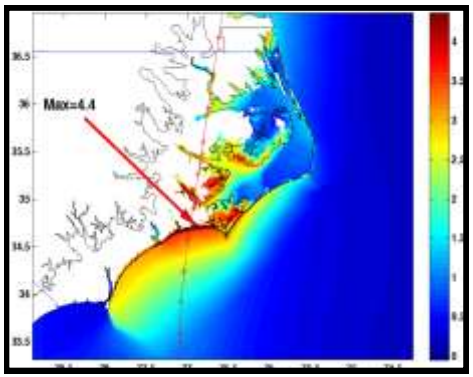
- Almost all large-scale data projects:
  - Include multiple producers of data
  - Data experts to curate and model the data
  - CI experts to store, manage, analyze and serve the data
  - Many end users who will use the information from the data for action

Selected RENCI Collaborative projects in data science

- **Collaboration can be:**
  - ***Science Driven***
    - ***Storm surge prediction (ADCIRC)***
    - ***Clinical genomics (NCGENES)***
  - ***Technology Driven***
    - ***Integrated Rule-Oriented Data-management System (E-iRODS)***
    - ***Network infrastructure for data (NIAAS)***
  - ***Data Driven***
    - ***National Consortium for Data Science (NCDS)***
    - ***Datanet Federation Consortium (DFC)***

Storm Surge Forecasting (ADCIRC)





Sandy (2012) and Irene (2011) flooding forecasts used by

- **National Hurricane Center in Miami**
- **US Coast Guard Atlantic Command**
- **Regional National Weather Service Offices**
- **State and local emergency managers**

- System uses **NSF/NARA** funded iRODS**, NOAA NOS** gauge data , **USGS** data, **DHS/FEMA** collected high-water mark, meteorological forecasts from **NOAA's NCEP** and **NHC**
- Very large pre-existing datasets; provides early guidance information, available about 10 minutes after official NHC forecast storm advisory
- **DHS-funded research activity** through the DHS Coastal Hazards Center of Excellence at the University of North Carolina at Chapel Hill
- **Winner, DHS Science & Technology Impact Award, 2012**

NCGENES - Clinical Genomics

Today:

- NIH prototype to evaluate the *ethical* and *social* challenges of genomic sequencing in clinical care
- Big Data to clinically-relevant knowledge ('Clinical bins')
- Over 100 patients in the system today…

Tomorrow:

- 100M+ genomes scattered throughout the health care system
- We face a multitude of data challenges before we realize the potential of genomics in healthcare…

North Carolina Genetic Evaluation by Next-generation Exome Sequencing

Criteria:	*Loci with Clinical Utility*	*Loci with Clinical Validity*			*Loci with Unknown Clinical Implications*	Loci with important reproductive implications
Bins:	**Bin 1** Genes, which when mutated, result in high risk of clinically actionable condition	**Bin 2A** Low risk incidental information	**Bin 2B** Medium risk incidental information	**Bin 2C** High risk incidental information	**Bin 3** All other loci	**Bin R** Carrier status for severe AR disease

E-iRODS: Enterprise iRODS data grid technology

Tailored E-iRODS distribution

- Broaden adoption in federal agencies and industry
- Broaden developer-base
  - Pluggable framework = data building blocks

E-iRODS Consortium membership model

- Membership funding model
- Building an ecosystem of invested partners (Max Plank Society, NASA, NOAA, others)

At a glance:
National Consortium for Data Science

www.data2discovery.org

- **Mission:** Secure US role as leaders in data science research & education, position US industry to use the power of data to drive economic growth
- **Vision:** Focused multi-sector, multidisciplinary data science community to solve big data challenges and drive the field forward
- **Goals:**
  - **Engage** broad communities of data experts
  - **Coordinate** data science research priorities that span disciplines and industries
  - **Facilitate** development education & training programs
  - **Support** development of technical, ethical & policy standards
  - **Apply** NCDS expertise to data challenges in science, business and government

National Consortium for Data Science

Founding Members

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci

cisco

GE

NC STATE UNIVERSITY

IBM

sas

Duke UNIVERSITY

Drexel UNIVERSITY

RTI INTERNATIONAL

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

NetApp

UNCCHARLOTTE

NIEHS National Institute of Environmental Health Sciences

THE HAMNER INSTITUTES FOR HEALTH SCIENCES

MCNC Connecting North Carolina's Future Today

NCDS components

- **Data Observatory**
  - Shared, distributed infrastructure housing large organized research data sets to enable fundamental advances in data science
- **Data Laboratory**
  - R&D into critical tools and techniques for data science
- **Data Fellows program**
  - Educate and Train data science workforce and leaders
- **Data Science curriculum**

NCDS: A Public – Private Partnership

Shared Benefits

- Access to organizations with complimentary agendas
- Glimpse into future trends, leads to competitive advantages
- Positive exposure and visibility
- Opportunities for joint educational/workforce materials
- Data Laboratory/Observatory (access to shared data platform)
- NCDS helps to fill a "concierge" role facilitating such things as:
  - Identifying ideas for collaboration, revenue generation
  - Identifying opportunities for cross-marketing, public relations and communications

Industry		Academic		Nonprofit and agency	
Benefits	Through	Benefits	Through	Benefits	Through
Access to: • Data science research on the horizon • Potential future employees, lower-risk vetting/recruiting • Opportunities for pre-competitive collaboration • Place industry scientists in academe	• Hosting student interns • Sponsoring research fellows • Working directly with academic researchers on joint-projects • Preferred access to and/or customized training and education for industry staff	• Funding for faculty and students • Opportunities to participate in collaborative research with NCDS partners • Access to industry • New curriculum, new programs • Attract best students and faculty	• Faculty course 'buy-outs' to fund selected research projects • Funding for graduate students to work in partnership with industry • Access to industry resources such as reduced cost software and hardware	• Access to: • Leading edge research • Access to industry • Applied problem solving • Regional economic development • Policy enhancements	• Hosting research fellows • Working with industry and academe • Increased understanding of issues and opportunities • Coalitions to provide end-to-end solutions for business development

NCDS Data-centric research framework

Grand Data Challenges

1. How do we translate genomic data into better healthcare?
2. How will data help us understand climate change and manage natural resources?
3. How can data help us understand human behavior and social trends?
4. How can an understanding of materials—from atoms to structures—lead to better products?
5. How will we manage the exploding Internet of Things?

Data Users

- Policy Makers
- Scientists
- Clinicians
- First Responders
- Industry
- Engineers
- Educators
- Public
- …

Kickoff: First NCDS Leadership Summit

Data to Discovery: Genomes to Health, held April 23 – 24, 2013

- Keynote address: **Dr. Eric Green,** Director, National Human Genome Research Institute, **NIH Interim Director for Data Science**
- First in annual **Data to Discovery Leadership Summits**: environmental science, homeland security, etc.
- Purpose: Focused discussion among data science leaders to elicit key data problems and opportunities
- Final Product: Leadership Summit Report on data challenges and opportunities in genomic science.