



The National Consortium for Data Science

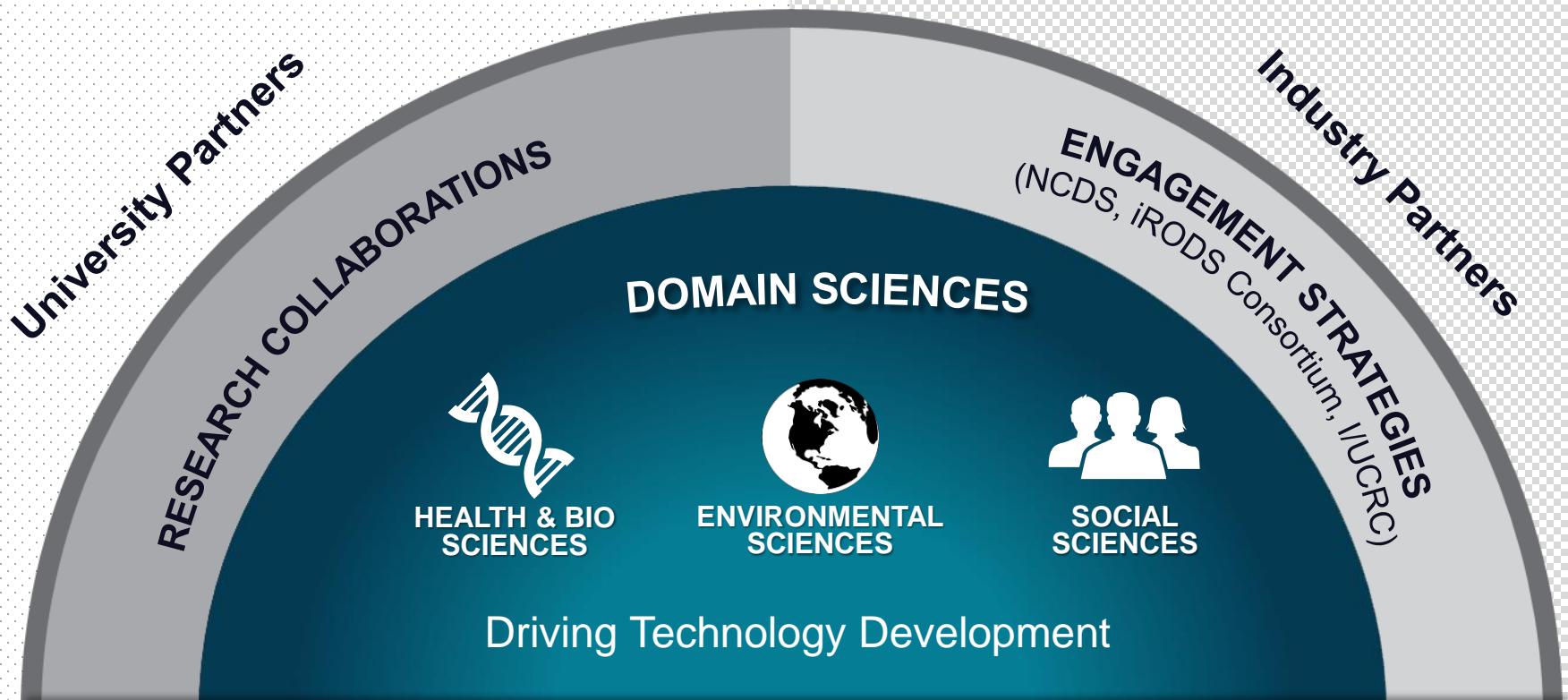
A public-private collaboration to advance data science

Renaissance Computing Institute (RENCI)
University of North Carolina, Chapel Hill

ncds

THE NATIONAL CONSORTIUM
for DATA SCIENCE

RENCI FOCUS: DATA SCIENCE TO ENABLE RESEARCH, INNOVATION, AND ECONOMIC DEVELOPMENT



FOUNDATIONAL TECHNOLOGIES



DATA MANAGEMENT

- Irods
- Data Federation



COMPUTING

- HPC
- Performance Optimization



NETWORKING

- Software Defined Networks
- NlaaS



SOFTWARE

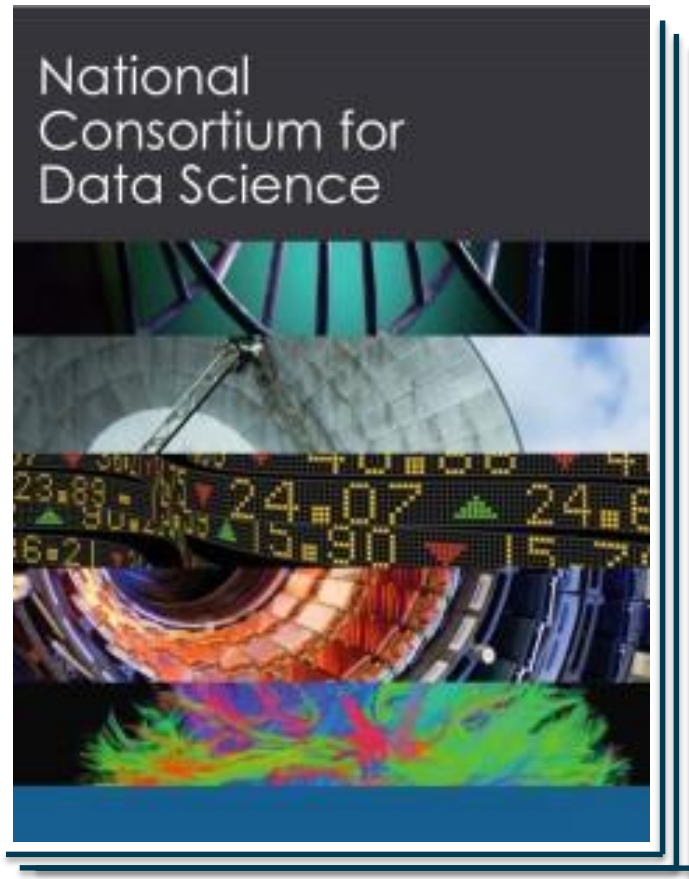
- Open Source Development
- Hydroshare
- SRS



VISUALIZATION

- InfoViz
- SciViz
- GeoViz

Foundational Concepts



- Data science needs to emerge as a complete domain science on par with other domains
- Public-private partnerships are necessary to access the requisite talent and perspectives
- Long-term thinking is essential
 - focus on research, education, and commerce

The National Consortium for Data Science



Mission

Help secure US role as leaders in data science research & education; position US industry to use the power of data to drive economic growth



Goals

- **Engage** broad communities of data experts
- **Coordinate** data science research priorities that span disciplines and industries
- **Facilitate** development education & training programs
- **Apply** NCDS expertise to data challenges in science, business and government



Vision

Focused multi-sector, multidisciplinary data science community to solve big data challenges and drive the field forward

ncds is a **strategic** approach to data science and big data opportunities.

NCDS Members



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

NC STATE UNIVERSITY



Deloitte
Consulting



THE UNIVERSITY of NORTH CAROLINA
GREENSBORO



Consortium Highlights



Research

- Data Fellows program
- Partnerships with NCBSTI & the DSA Initiative
- Focus on three overarching research themes



Education

- *Data Matters* Summer Short Course Series
- Data Science Career Panel Events



Engagement

- *DataBytes* Lunchtime Webinar Series
- Support for multiple data science conferences, hackathons
- NSF Big Data Hub

Current Research Themes



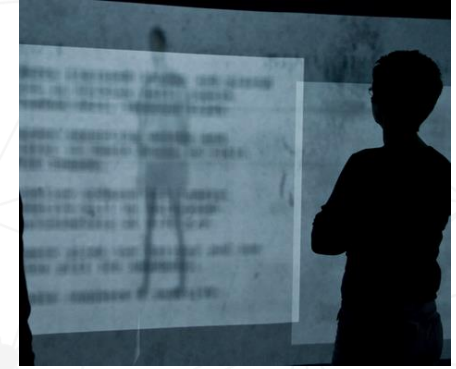
Innovative Infrastructure & the IOT

Purpose: Focus NCDS activities on data challenges of interest to members in order allow members to have a voice, increase the value of NCDS membership, produce products, and establish thought leadership.

Outcomes: White papers, position papers, best practice documents, lectures, panels, special events, etc.



Workforce Development



Anonymizing Data

Data Fellows Program

Fosters private-public relationships; **engages** future data scientists; **bridges gaps** between research and practice; **creates** NCDS-sponsored scholarship

\$50,000 seed grants for early career faculty



Seed grant approach to fund initial cadre of Fellows from NCDS academic member campuses



Teaming with an NCDS member encouraged; industry-led selection process



Funds used for **course buy-outs, summer salary, graduate student support, conference travel** and modest infrastructure costs



Target: **3 awards, \$50K each, focus on early career faculty**

Additional support provided by UNC General Administration to **offer fellowships to UNC System campuses.**

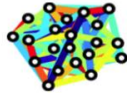
NCSU Data Fellows Highlight

- Dr. Blair Sullivan selected for \$1.5 million Moore Investigator Award as part of its Data-Driven Discovery Initiative
- Sullivan's work is based on a field of study called parameterized complexity. These algorithms leverage a graph's structure to solve time-consuming problems much more quickly.

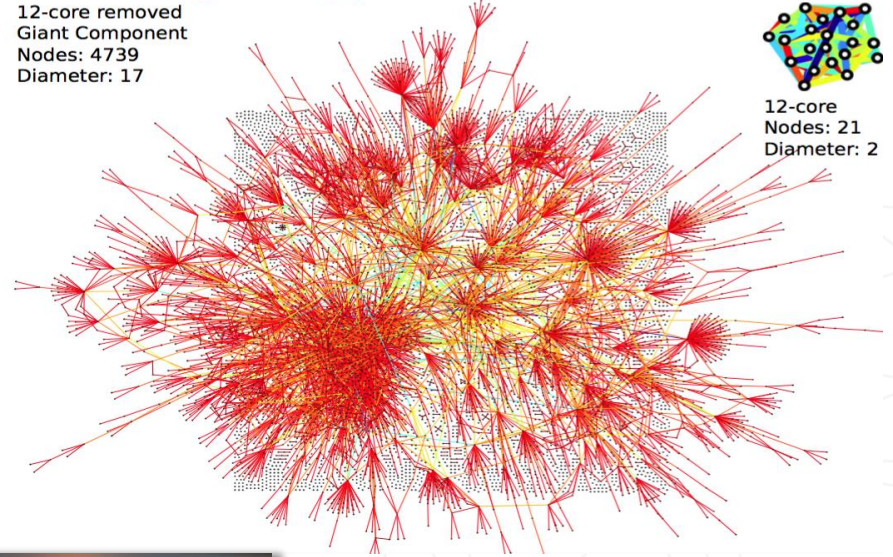
NC STATE UNIVERSITY

AS: removing 21 of 6474 nodes

12-core removed
Giant Component
Nodes: 4739
Diameter: 17



12-core
Nodes: 21
Diameter: 2



“The NCDS congratulates Dr. Sullivan on this great opportunity to expand her research and turn theory into practice. This is exactly what we hope NCDS Data Fellowships will enable for talented faculty—new opportunities to move their research forward and answer important data science questions.”

- Stan Ahalt, director of the NCDS Data Fellows Steering Committee

NC Data Science and Analytics Initiative

- Shared, multi-institutional data environment for collaboration & discovery
- Collaborative data store and discovery - iRods V4, DataVerse Network
- On-demand cloud computing for analytics workflows – NCSU’s Virtual Computing Lab (VCL)
- Includes proposal process and financial support for promising data-related research across campuses
- Web interface & connectors to “conventional” HPC and Hadoop
- Distributed authentication managed by each campus
- 3 year & \$2.2M investment by the State of North Carolina



Partnership with NC Dept. of Commerce

NCBSTI identified data science as one of six “Grand Opportunities” for the NC economy

Project underway to survey NC data science leaders on identification of the most valuable strategy:

- Survey/research of data science assets in the state
- Develop “Data Sciences Index” and other methodology to assess the state’s relative standing nationally
- Survey employer skill needs
- Engagement with KFBS “STAR” Program student team

Key question: *How important is data science to North Carolina’s economy now, and how important could it be?*



Workforce Training

DATA MATTERS

DATA SCIENCE SHORT COURSE SERIES



2016 JUNE 20 - 24,
CHAPEL HILL, NC



THE NATIONAL CONSORTIUM
for DATA SCIENCE



PREVIOUS COURSES

A week-long series of two-day and one-day courses aimed at professionals in business, research, and government. If your organization struggles to stay afloat in the data deluge, if you grapple daily with large, complex data, if you want to capitalize on the opportunities of big data, **Data Matters is for you.**

DATA MATTERS 2016 UPDATES

- ▶ Choose courses to meet your needs and interests; come for a day or stay all week.
- ▶ All courses held at the Friday Center for Continuing Education in Chapel Hill, NC.
- ▶ Registration includes lunch each day and an evening cocktail reception during the week.

DATA MATTERS

DATA SCIENCE SHORT COURSE SERIES

Monday & Tuesday June 20 & 21	Wednesday June 22	Thursday & Friday June 23 & 24
Intro to Data Science	Conceptual Diagrams in Information Visualization	Intro to Data Mining & Machine Learning
Intro to Information Visualization	Programming in R	Health Informatics in the Age of Interoperability
Intro to Data Science Using R	Open(ing) Data	Collecting, Classifying, & Analyzing Textual Data
Data Curation: Managing Data throughout the Research Lifecycle	Creating Surveys in Qualtrics	Simulation Strategies in Data Science: System Dynamics & Agent-based Modeling
Writing Questions for Surveys	Intro to Big Data & Machine Learning for Survey Researchers & Social Scientists	Conducting & Analyzing Cognitive Interviews
Intro to Survey Sampling	Geospatial Data Science	Analysis with Complex Sample Survey Data

Career Networking Events

Student-focused, cross campus events designed to connect industry with future data science talent




DataBytes Webinar Series


Objective: Present novel and compelling data science concepts and projects


Member and non-member presenters


First Wednesday of the month


2015

 *Geospatially Explicit Synthetic Populations*
Bill Wheaton


 *The Internet of Everything: The Next Big Wave*
Russ Gyurek

 *Leveraging AI for Big Industrial Data Science*
Steve Gustafson


 *Big Data, Small Data, & Behavioral Insights*
James Guszcza


 *Analytics for Large-Scale Temporal Event Data*
David Gotz


 *Toward Machine Oblivious Graph Analysis*
Erik Saule


 *Assessing the Impact of Data & Software on Science*
Erjia Yan


2016

JAN *Semantics: Revolutionary Breakthrough or Same Old Thing* 
Andy Crapo

FEB *Simulating Storm Surge for Coastal Hazards and Risk Assessment* 
Brian Blanton

MAR *The Computing Universe* 
Tony Hey

APR *NOAA's Collaborative Big Data Project* 
Jeff de La Beaujardiere

MAY *Title TBD* 
Stephen W. Edwards

JUN *TBD*

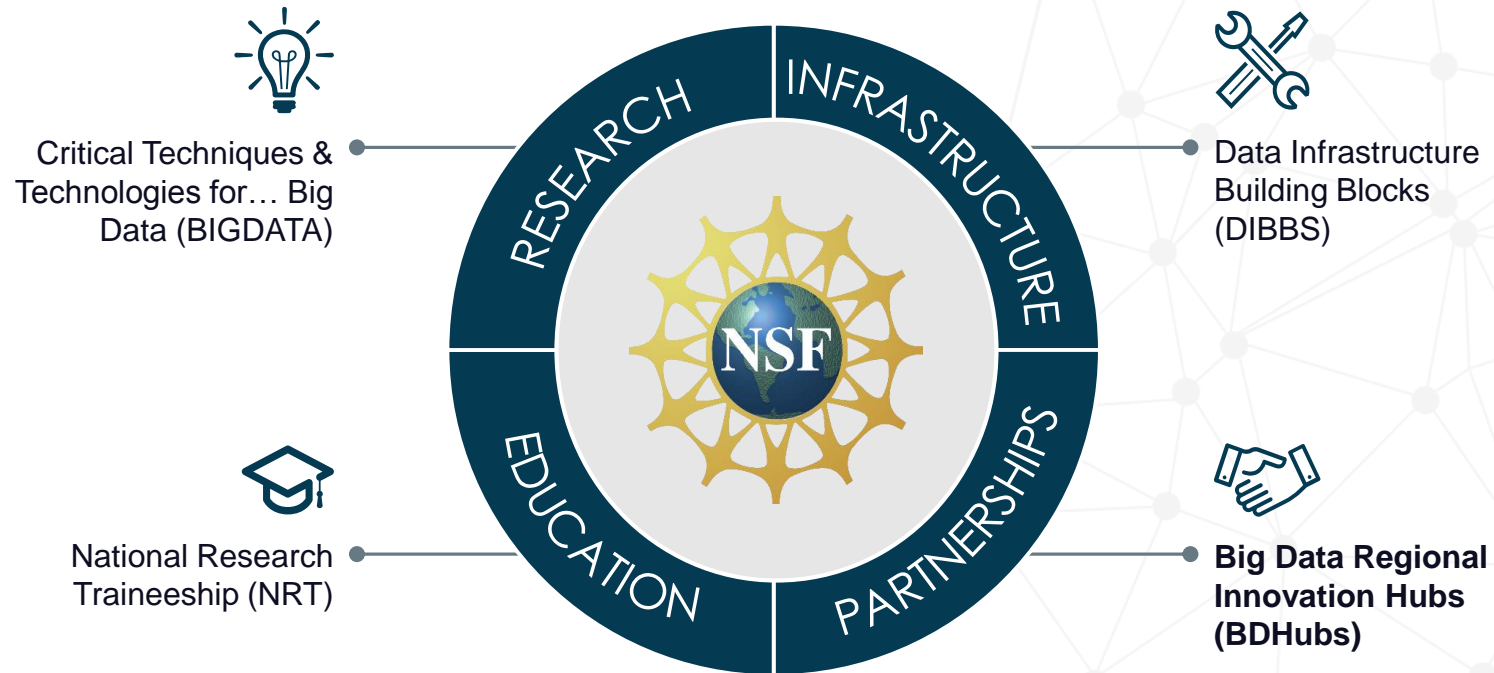
AUG *TBD*

SEP *TBD*

OCT *TBD*

NOV *TBD*

NSF Big Data Portfolio of Programs



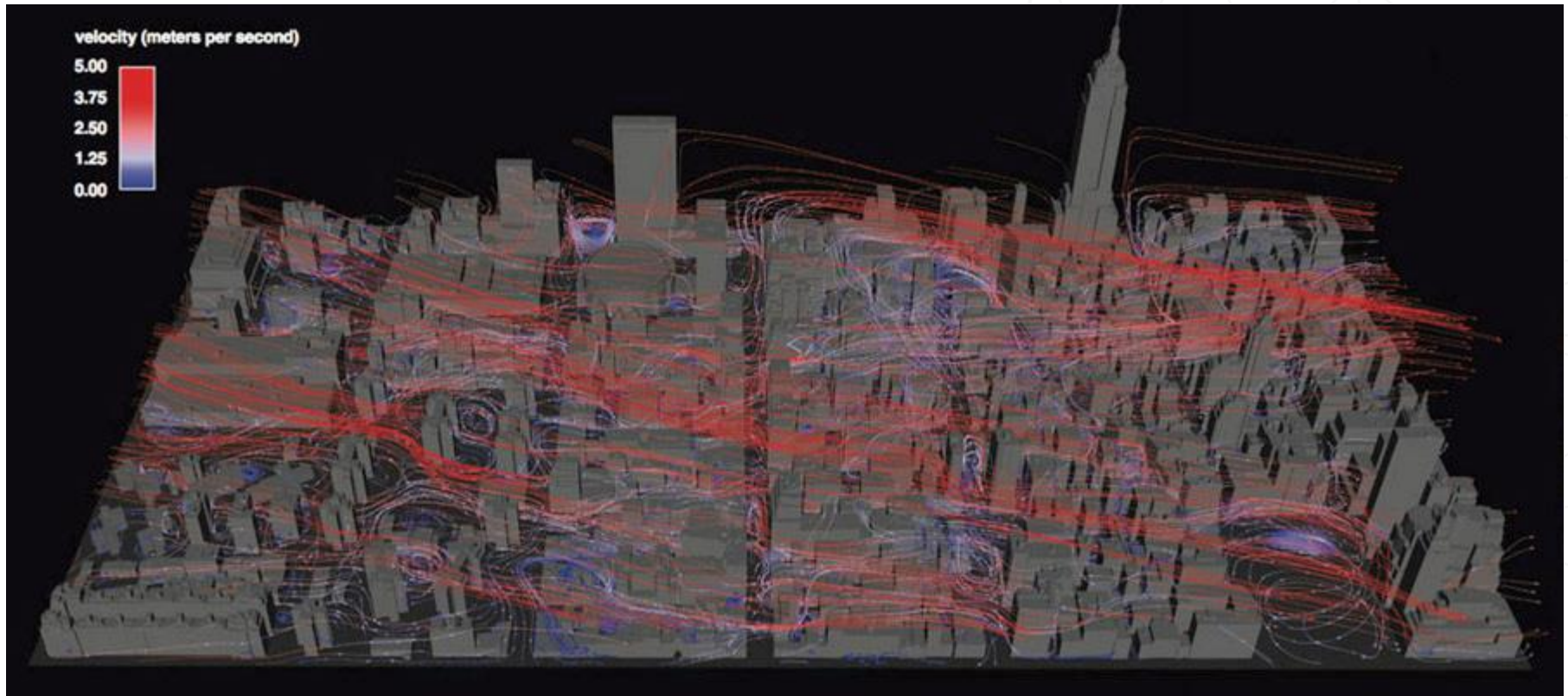
GOAL: To stimulate regional and grassroots partnerships involving public and private organizations of all types in order to establish a national Big Data innovation ecosystem

RENCI Engagement Center at NCSU

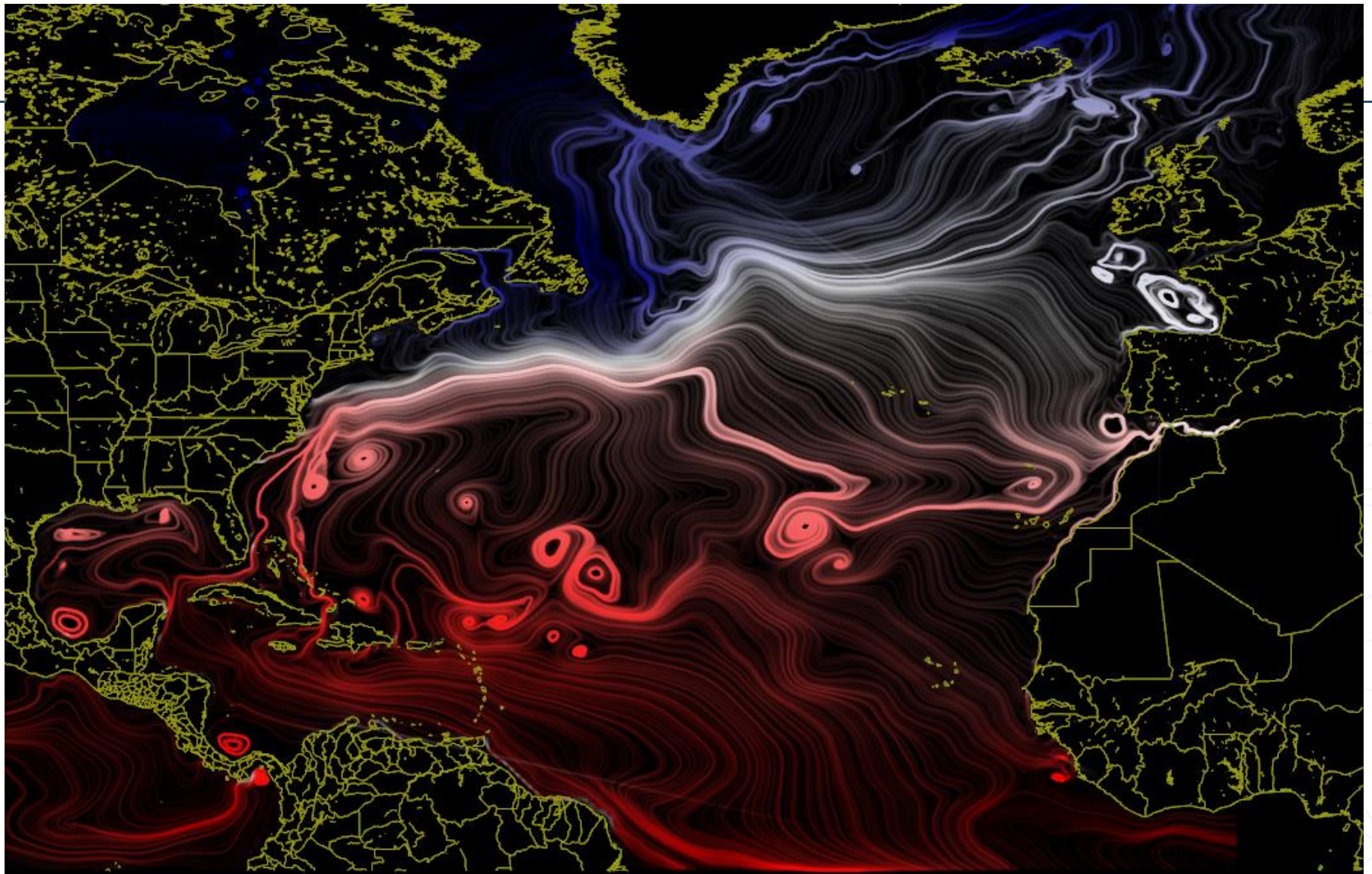
- Established in 2007 at Centennial Campus to foster collaborations with NCSU research community
- Visualization-focused; Two full-time resident visualization researchers (PhD)
- Project focus spanning scientific visualization and information visualization
- Establishment of **Visualization Studio** (RENCI's Social Computing Room) at DH Hill Library
- RENCI-sponsored Vis Call for Projects focusing on Applied Visualization Research
 - \$12,000 (each) for two projects
 - 50% dedicated personnel time
 - Running two years in a row (2012-13)



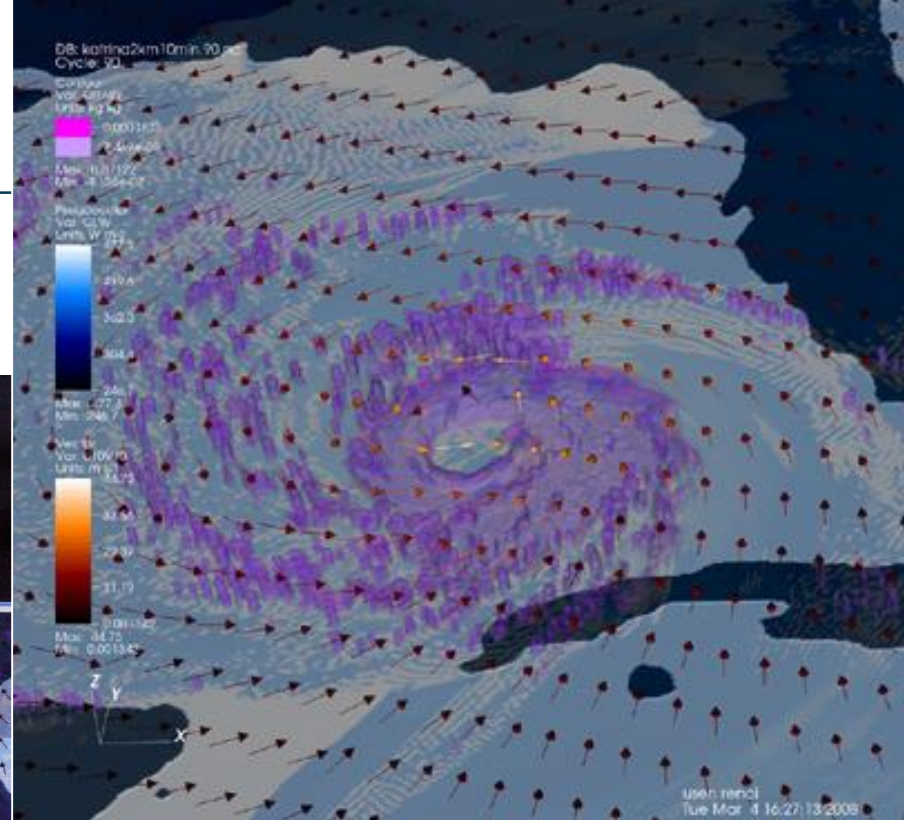
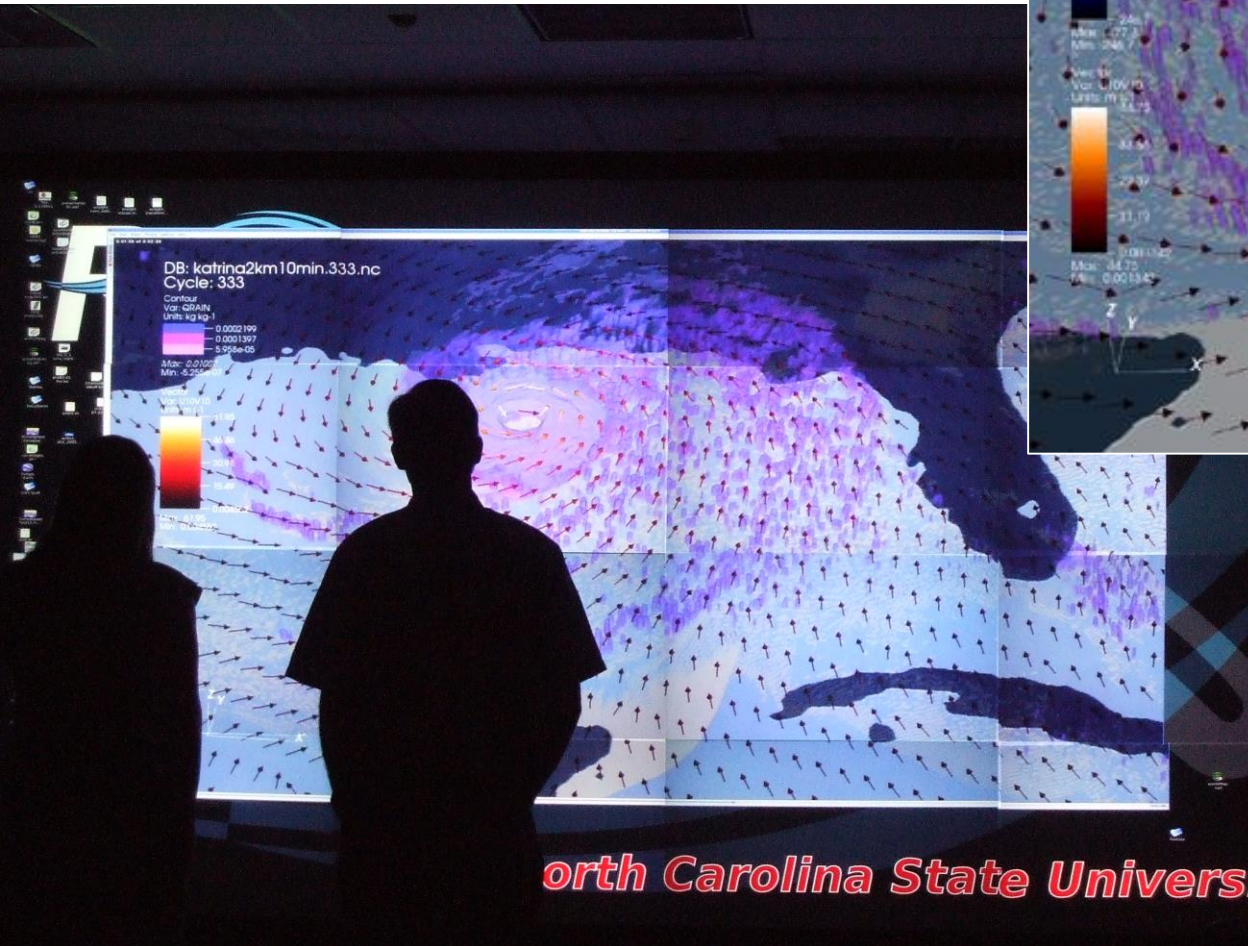
RENCI Visualization Examples



Visualization: Dr. David Borland, RENCI

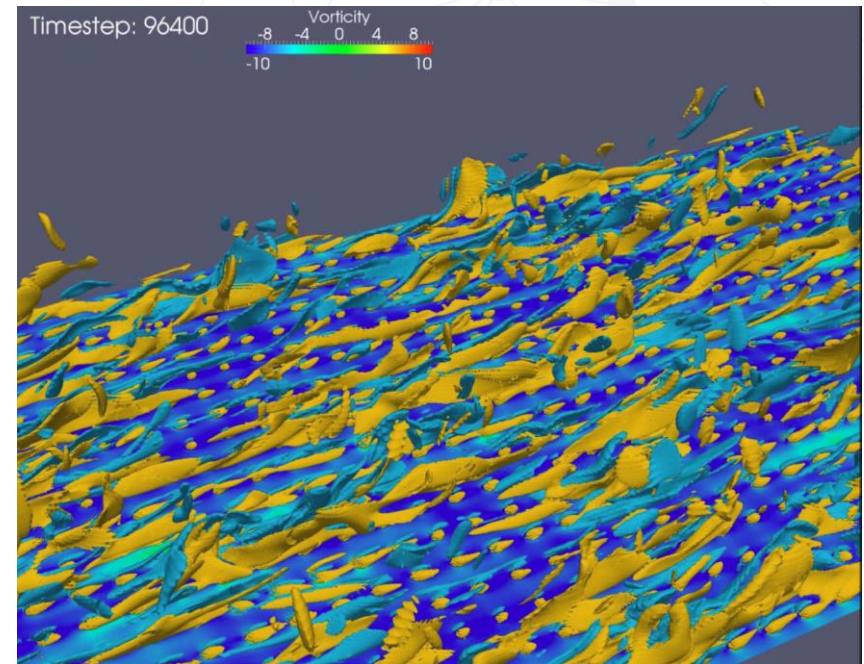
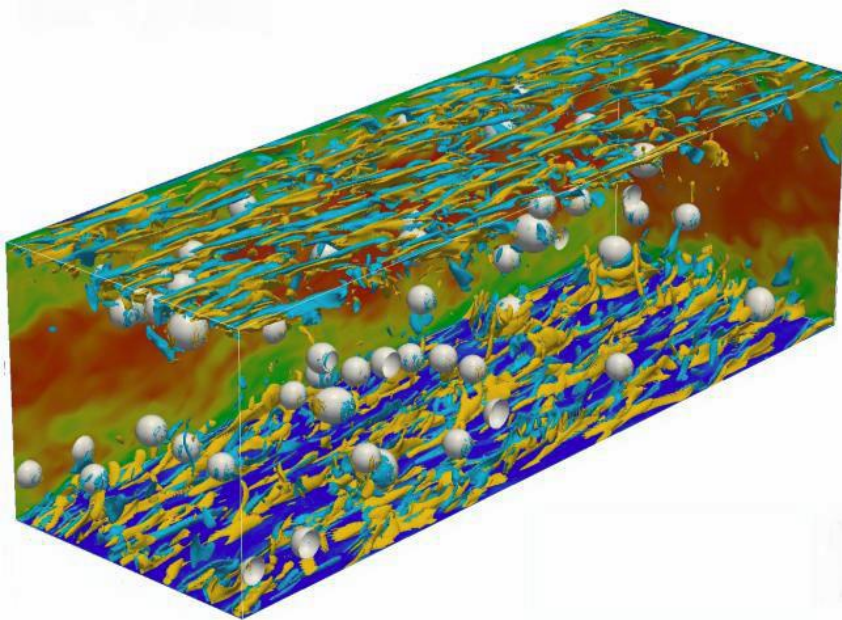


Visualization: Dr. David Borland, RENCI



Visualizations: Steve Chall,
Theresa-Marie Rhyne

North Carolina State University



Visualizations: Dr. Igor Bolotnov's group,
Dr. Hong Yi (RENCI)

X3DBio1: 3D Biomolecular Structure Visual Analytical Tool

File

Drawing method: Ball-and-stick Stick

Atom coloring: Atom type Residue type

Selection: All Backbone Secondary structure

Alpha helix Beta sheet Turn Coil Show cartoon representation

Display density map

Cube size: 1 Clipping Range filter Minimum: 1.0 Maximum: 10.0

Search by sequences: 4-20 | qfpk | 35

Distance range filter
Minimum: 0.00 Maximum: 20.00

Greyscale color lookup table ramp: Linear S-Curve Binning 256

Non-bonded interactions only
 Apply to highlighted cells only

Info of Loaded molecules

Name	Residue Range
Mol-0	1-506

Residue Sequence: QFPKWSPINRETYIDRLSARFEREGEQSQL
NALVAKAQKTPPEGWTMQDGTSPGNNT
AMRNLEFRDUMMBGLDGLGCRGQVCTM

Selected Location: (11, 7, 47)
Density Value: (6.08928)
Selected Residues (Atoms): R474(H, C, C, H, H, C, H, H, C, O), R475(N, R477(N, C, H, C), R478(N, H, H)

Location: (11, 7, 47)
Value: (6.08928)

47
Slice Z

6.90
4.60
2.30
0.000

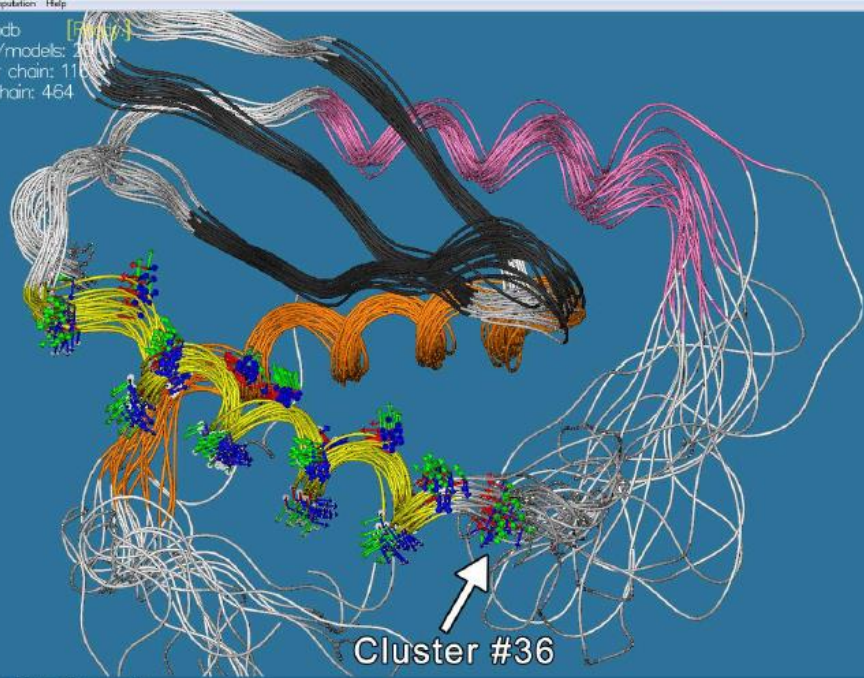
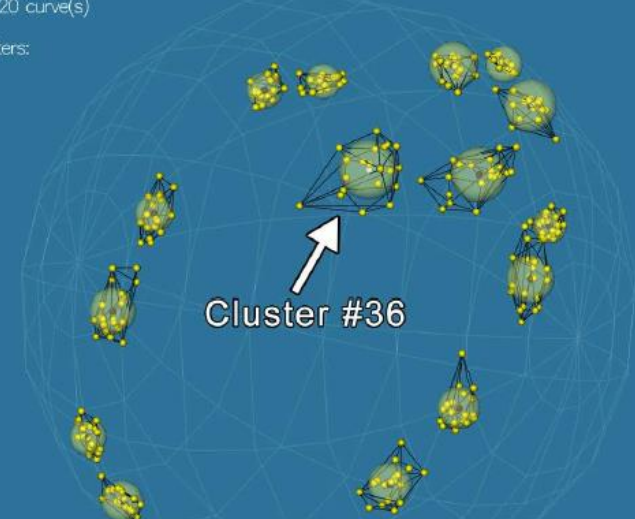
Visualization: Dr. Hong Yi, RENCI

PDB: 2HC5.pdb [R] [V]
 Num chains/models: 1
 Residues per chain: 110
 Atoms per chain: 464

Quaternion Map: [Discrete frames] C(alpha)
 Projection: X Y Z (change: mouse middle button / control+left button)
 Quaternions shown for all 20 curve(s)

Standard deviation of clusters:
 (transparent spheres)
 > Max: 0.1541
 > Min: 0.0157

Average dot products
 of quaternion clusters:
 > Max: 0.9482
 > Min: 0.8513



Secondary structures: (color indicates helices and sheets)

Residue Sequence

Residue index

ALA	LEU	LEU	MET	PHE	TRP	TRP	VAL
ARG	THR	ASN	GLN	CYS	CYS	GLY	PRO
ASP	GLU	LYS	ARG	HE			

Visualizations: Dr. Sid Thakur, RENCI

Protein structure visualization

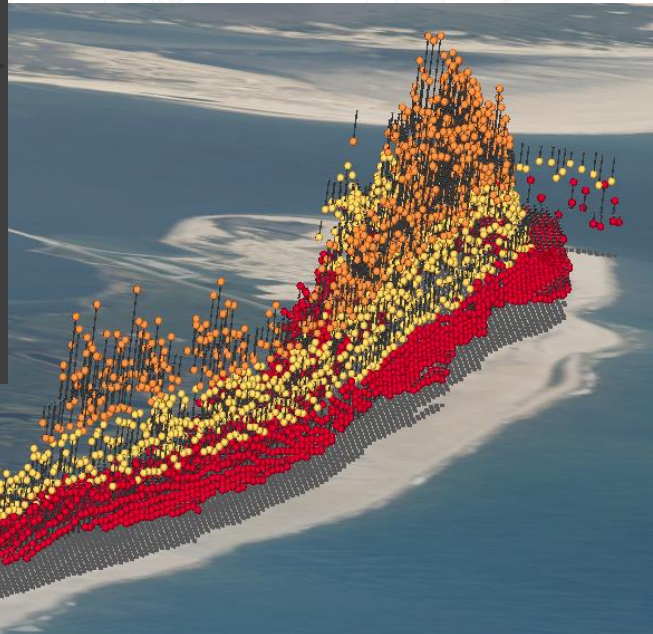
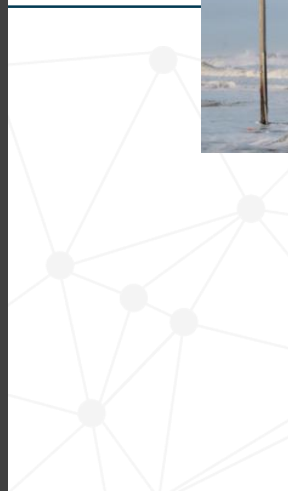
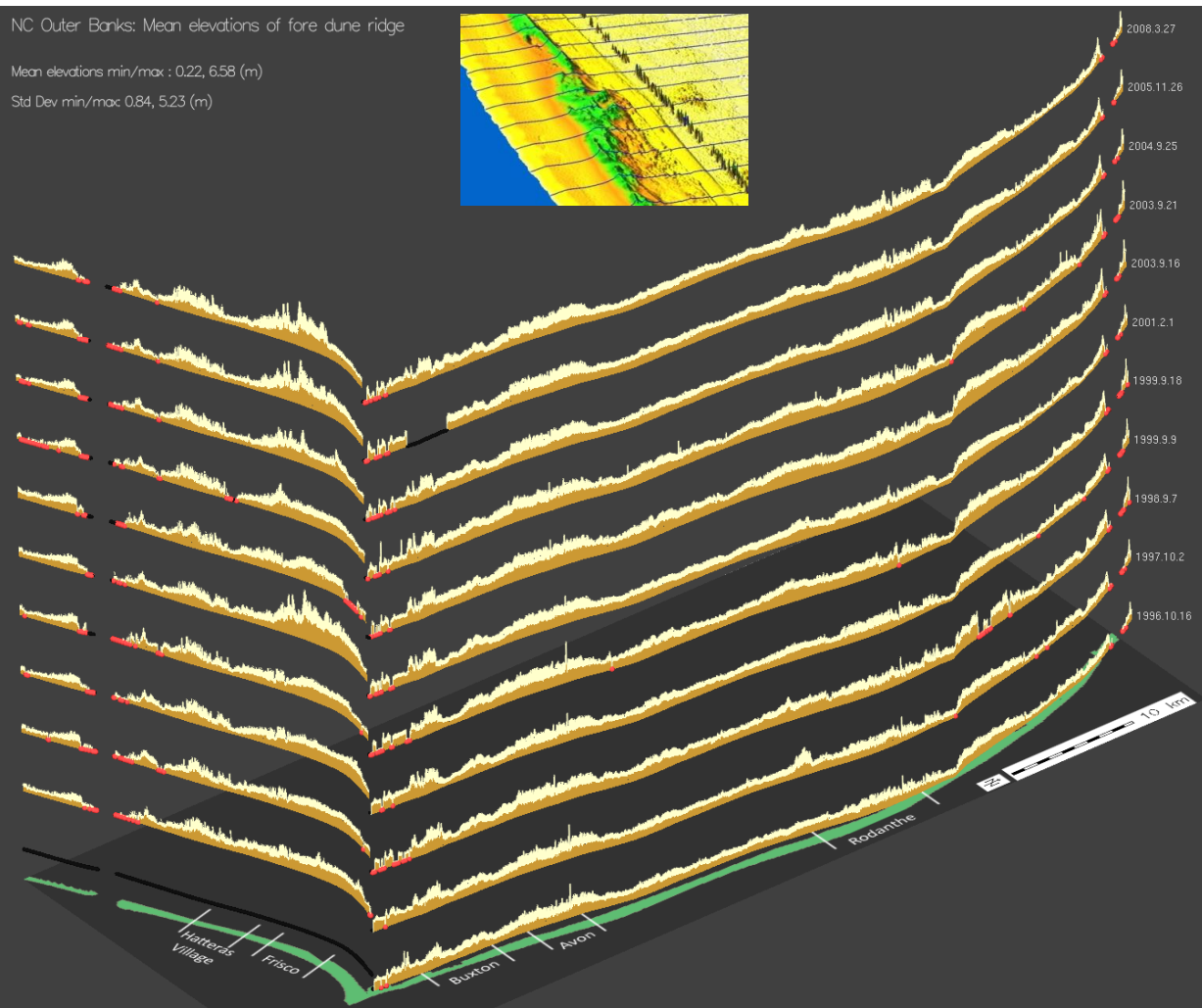
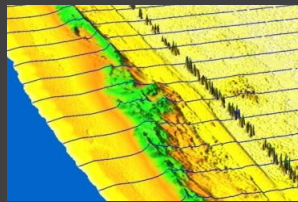
Solution NMR Structure of Protein yyvC from Bacillus subtilis.
 Source: PDB.org

Visualization by:

NC Outer Banks: Mean elevations of fore dune ridge

Mean elevations min/max : 0.22, 6.58 (m)

Std Dev min/max: 0.84, 5.23 (m)



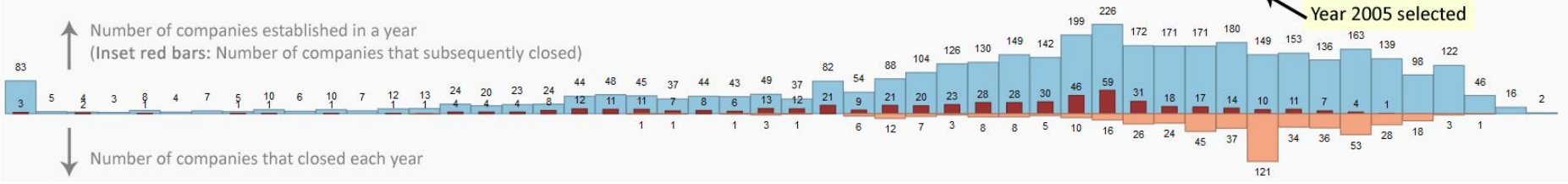
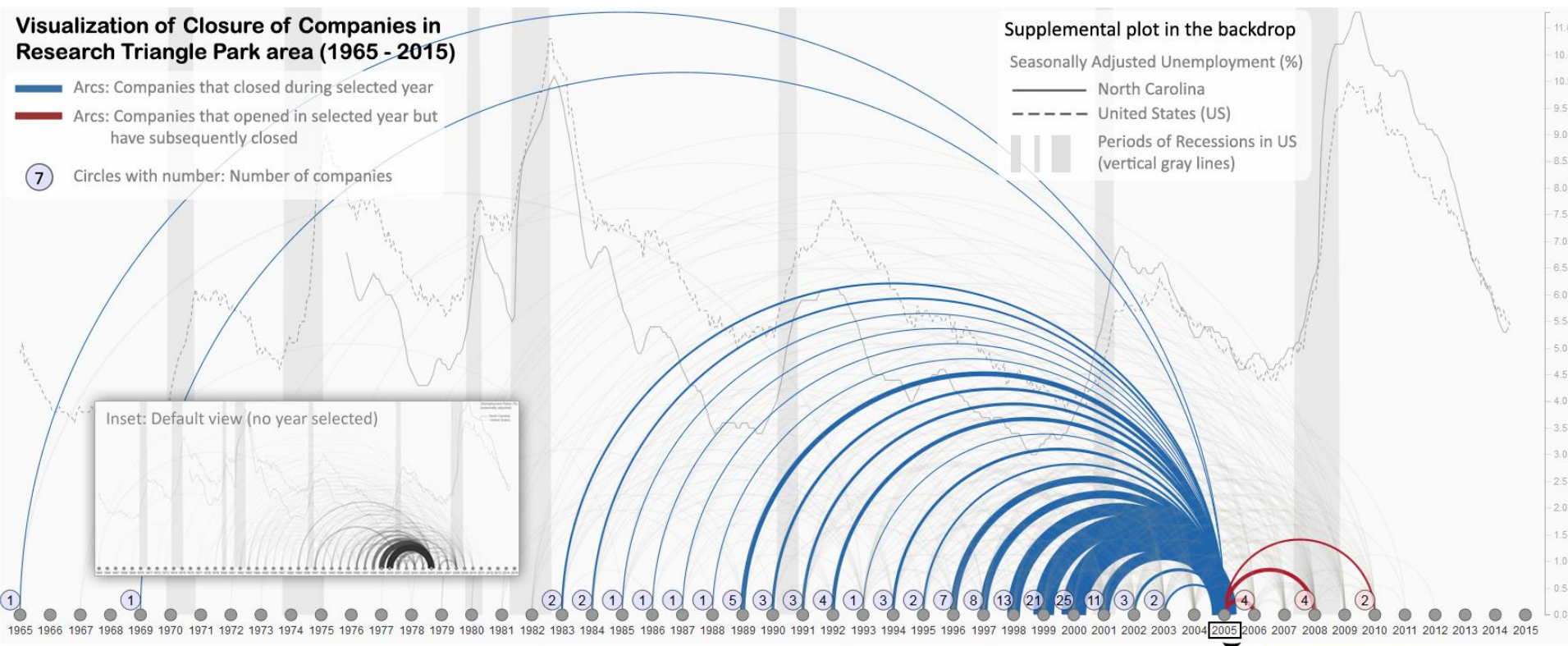
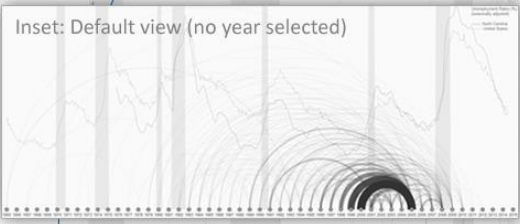
Visualizations: Dr. Sid Thakur, RENCI

Visualization of Closure of Companies in Research Triangle Park area (1965 - 2015)

- █ Arcs: Companies that closed during selected year
- █ Arcs: Companies that opened in selected year but have subsequently closed
- ⑦ Circles with number: Number of companies

Supplemental plot in the backdrop

- Seasonally Adjusted Unemployment (%)
- North Carolina
- - - United States (US)
- █ Periods of Recessions in US (vertical gray lines)

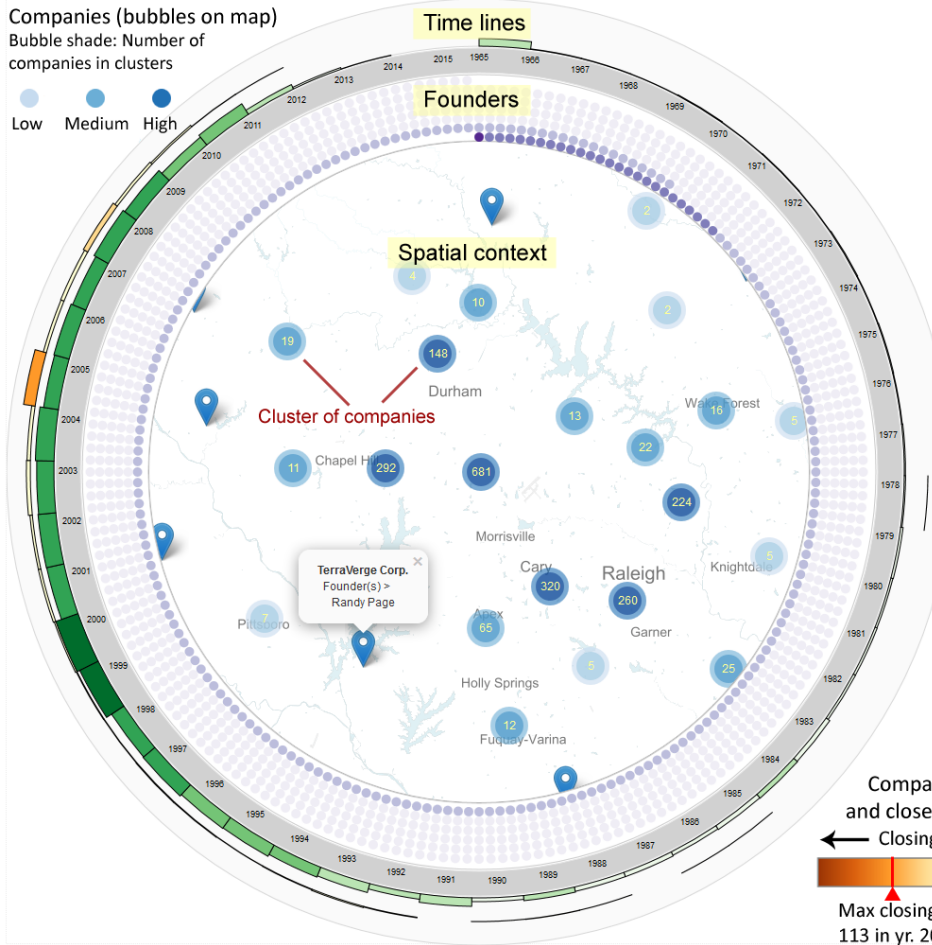


Visualization: Dr. Sid Thakur, RENCI

Time, People, Geography

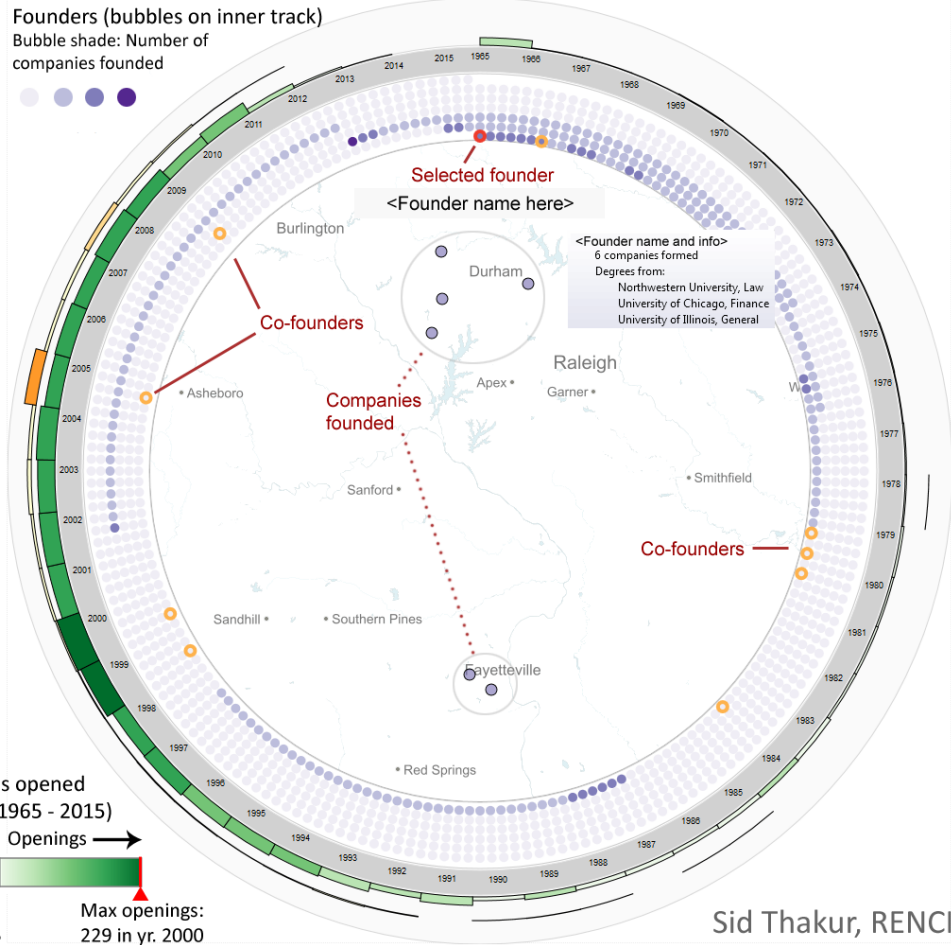
Companies (bubbles on map)
Bubble shade: Number of companies in clusters

Low Medium High



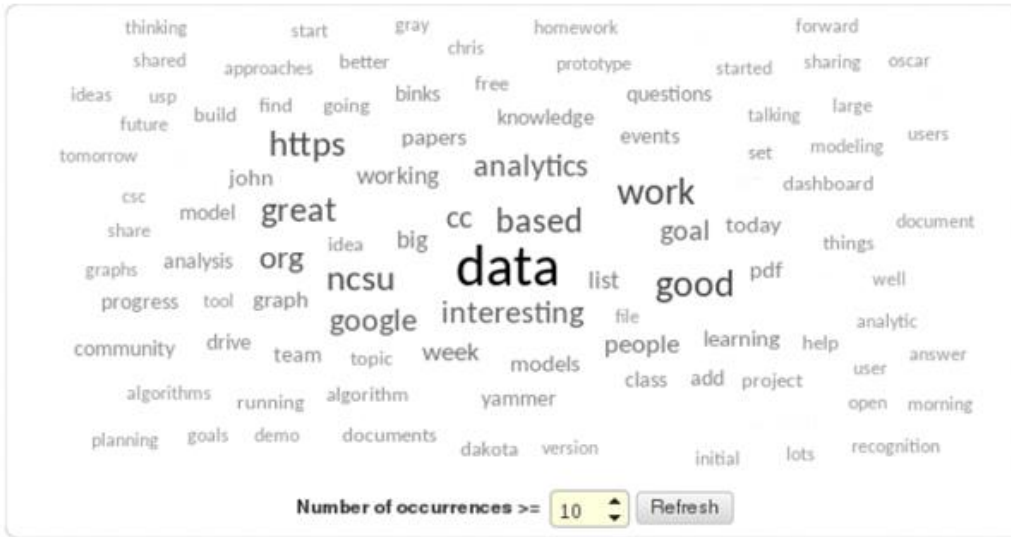
Founders (bubbles on inner track)
Bubble shade: Number of companies founded

Low Medium High

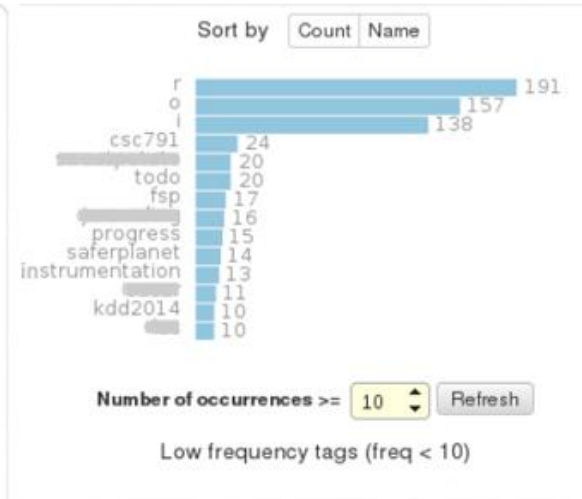


Sid Thakur, RENCI

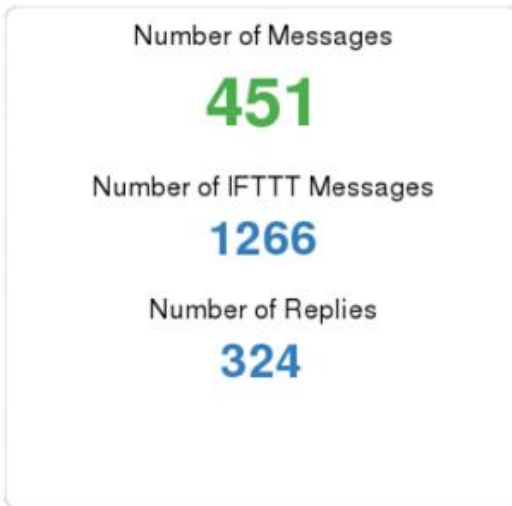
Word Cloud



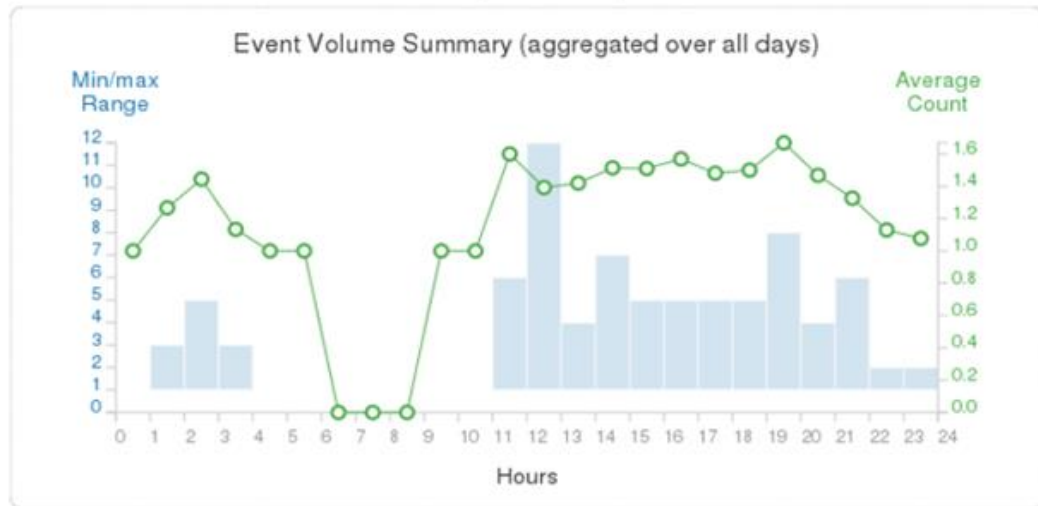
Tag Frequency



Yam Stats



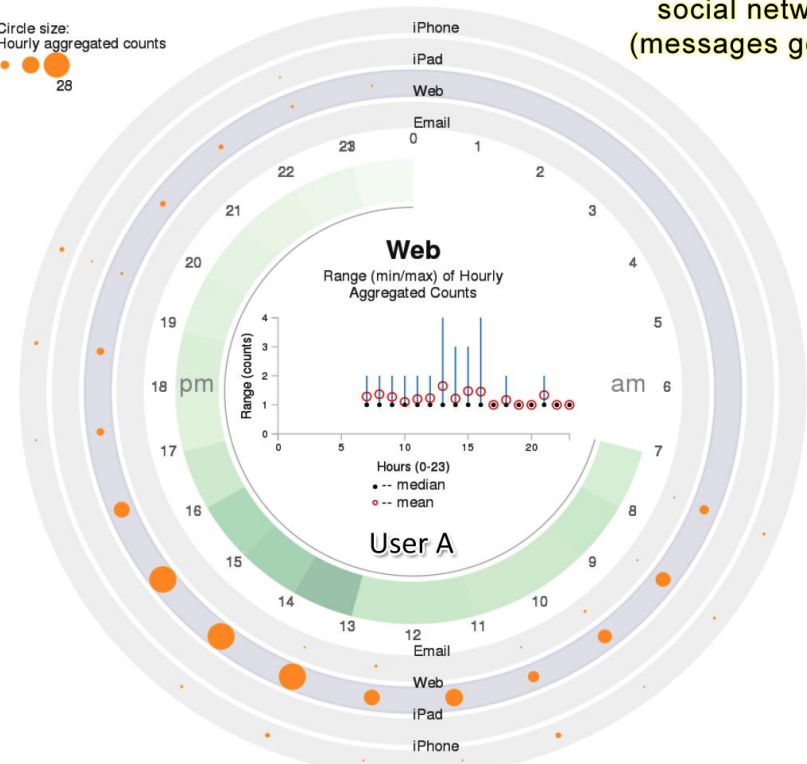
Daily Summary



Visualization: Dr. Sid Thakur, RENCI

Activity of a user on a social networking site in 2015 (messages generated on Yammer)

Plot #1



Detailed hourly view

Plot #2

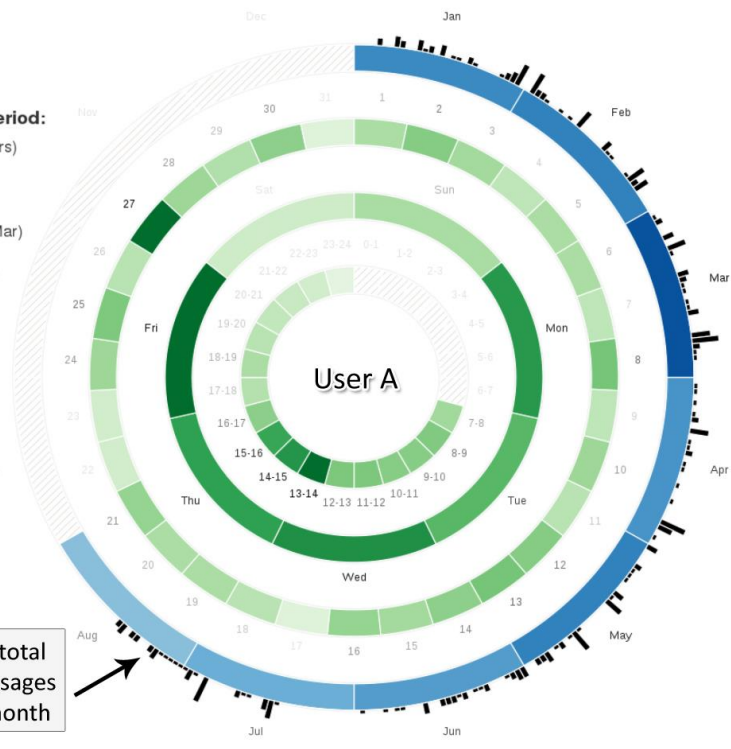
Maximum total events per time period:
 Hour: 33 max events (during 13-12 hrs)
 Week day: 51 max events (on Fri)
 Day: 23 max events (on 27 day)
 Month: 40 max events (in month of Mar)

Maximum events per app:
 Web: 211
 iPhone: 26
 Email: 8
 iPad: 2

No data

Bars represent total number of messages each day in a month

Arc color: Aggregated number of events
 Low (light) ---> High (dark)



Summary view (multiple time periods)

Visualization: Dr. Sid Thakur, RENCI