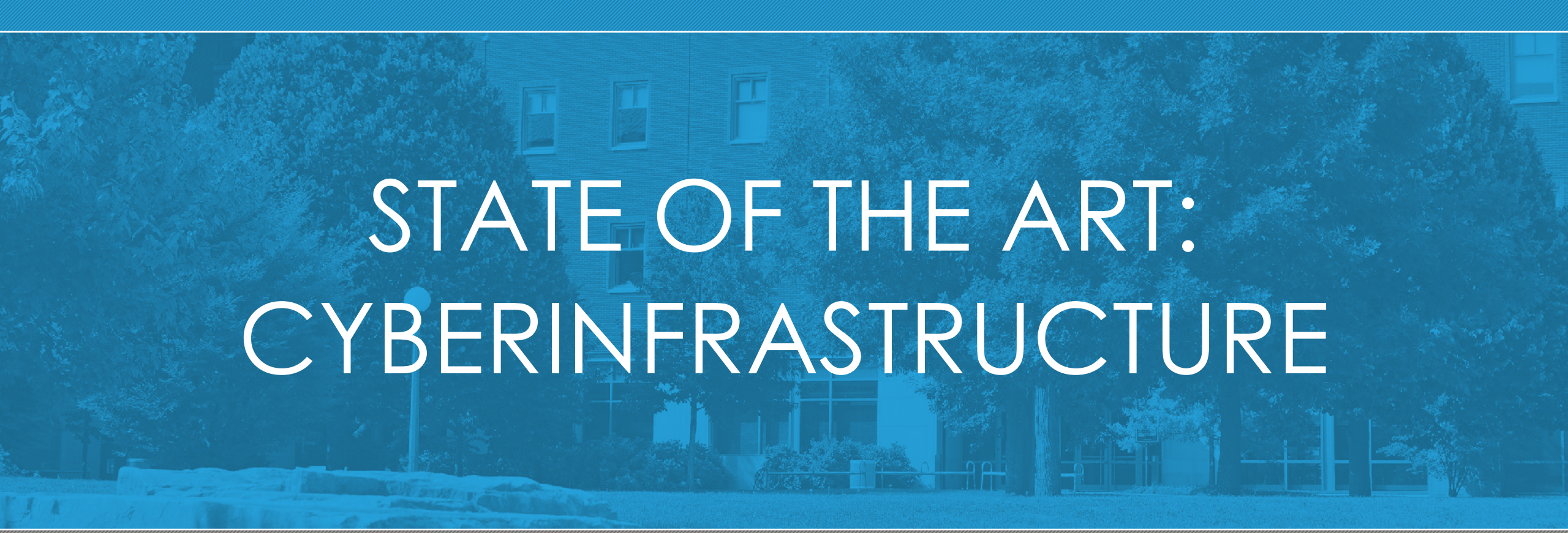


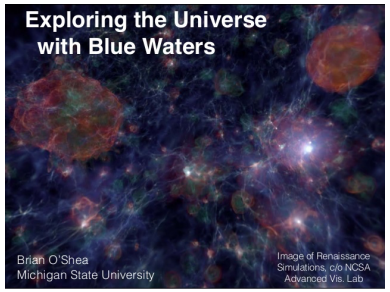
STATE OF THE ART: CYBERINFRASTRUCTURE

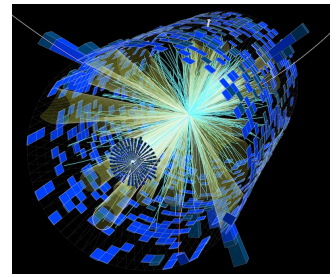
Presented by Ben Galewsky (bengal1@Illinois.edu)

ILLINOIS
NCSA | National Center for Supercomputing Applications

National Center for Supercomputing Applications

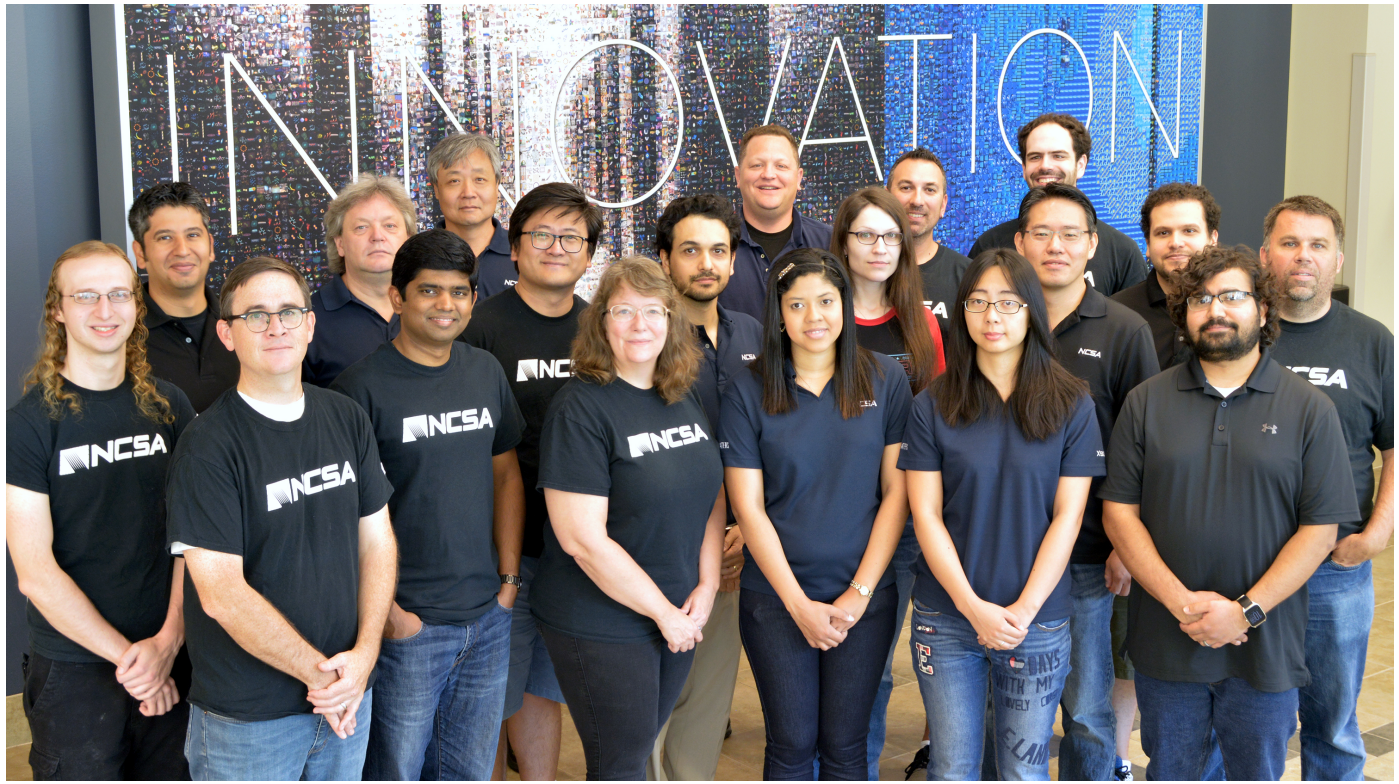
- Technology behind significant scientific research in:
  - Astrophysics
  - Earth Science
  - High Energy Physics
  - Cosmology
  - Materials Science
  - Agriculture
  - Bioinformatics





NCSA
Innovative Software & Data Analysis

- Research & Development
- Reusable software tools & frameworks for data analysis
- Bridging and amplifying efforts across different projects
- New custom software tools & frameworks



NSF Data Infrastructure Building Blocks

National Science Foundation in 2012 Sought Proposals that:

"develop robust, scalable, well-designed cyberinfrastructure contributing to future discovery and innovation across disciplines"

Architectural Vision for Research Cyberinfrastructure

Discipline Specific Environments

Science Portals

Applications & Frameworks

Research Facilities

Integrative Services

Resources

NSF Resources

Commercial Resources

University Resources

International Resources

Community Metadata Improvement #1443062

Computer Aided Discovery in Geo #1442997

Local Spectroscopy Data Infra. #1640899

Mobile Sensor Data #1640813

Archiving U-Series Geochronologic Data #1443037

Middleware & Analytics Libraries #1443054

Ocean Cloud Commons #1640775

Materials Engineering Data Lab #1640867

Modular Eng/Sci Cyber-platform #1443027

Whole Tale #1541450

ClearEarth #1443085

Continuous Capture of Metadata #1640575

LearnSphere #1443068

STORM #1443046

Triple Gateway #1443040

Nanocomposite Resource #1640840

Virtual Data Collaboratory #1640834

Vizier #1640864

Transient Data Acces #1443083

Collaborative Workflow Design #1443069

Spatial Data Synthesis #1443080

Brown Dog #1261582

HUBzero Geospatial #1261727

Virtual Info. Fabric Infrastructure #1640818

Pacific Research Platform #1541349

SciServer #1261715

User Driven Architecture #1443070

Scalable Data Delivery Platform #1541318

Gravitational-Wave Workflow #1443047

DataONE

BODS

XSEDE

Extreme Science and Engineering Discove

Science Grid

Agave

Data Exacell #1261721

4CeeD #1443013

North East Storage Exchange #1640831

Aristotle Cloud Federation #1541215

OSiRIS: Ceph and SDN #1541335

Confidential Social Science Data #1443014

Cyberinfrastructure Process Model

Scientific process described as a series of outcomes

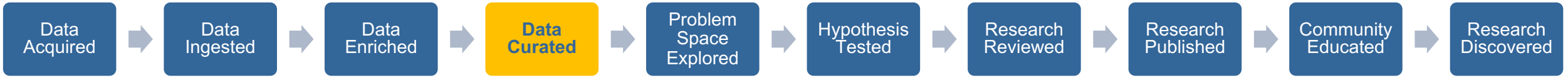
Data Acquired

- Raw data is received from sensors
- Data is safely settled so source data can be purged
- Data loaded into operational system
- Popular Technologies:
  - Globus
  - Kafka
  - Parsl
  - Pegasus

Data Enriched

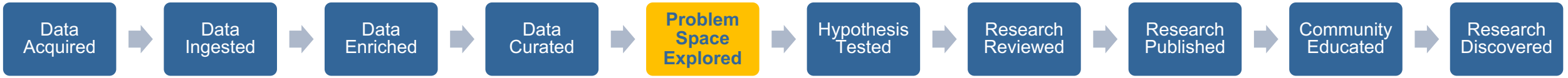
- Derived data is extracted from files
- Data cleaning processes are executed
- Machine learning models assign clusters to data
- Machine learning models detect features
- Popular Technologies:
  - Brown Dog
  - Clowder

Data Curated



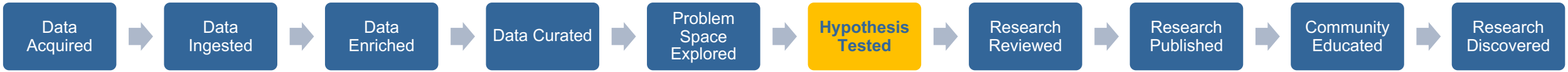
- Data is tagged manually through social curation
- Metadata is extracted
- Datasets organized
- Popular Technologies:
  - Clowder
  - iRODS
  - Dataverse

Problem Space Explored



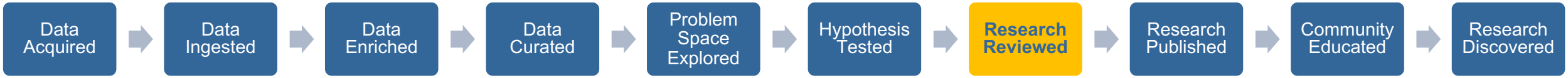
- Researchers perform ad hoc analysis on their data to get a better understanding and to test out models.
- Popular Technologies:
  - JupyterHub
  - NDS Workbench
  - Pangeo
  - Apache Spark

Hypothesis Tested



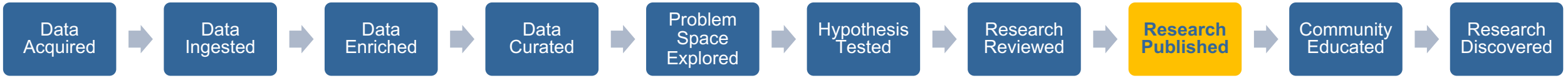
- Batch compute jobs run over dataset
- Popular Technologies:
  - XSEDE
  - Open Science Grid
  - Agave
  - Parsl
  - Pegasus
  - HTCondor

Research Reviewed



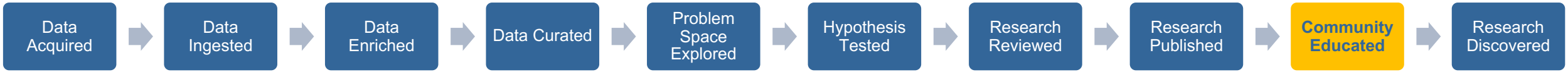
- Data and code packaged up for reproducibility
- Data provenance graphed for traceability
- Data and code shared only with collaborators during embargo
- Popular Technologies:
  - Whole Tale
  - REANA
  - Parsl
  - Pegasus
  - Globus Auth
  - Open CI

Research Published



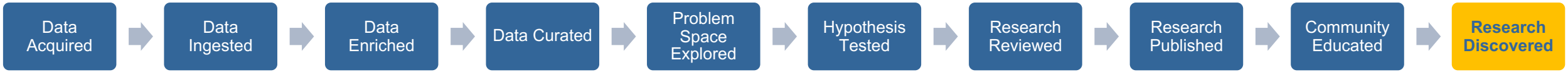
- Mint DOI
- Submit to FAIR repository
- Collocate running code with data
- Popular Technologies:
  - Globus Publish
  - Pangeo

Community Educated



- Members of the scientific community educated in findings, data, and software
- Run workshops and provide interactive notebooks
- Popular Technologies:
  - NDS Workbench
  - JupyterHub
  - Whole Tale

Research Discovered



- Community is able to find code and datasets
- Community members are able to explore and decide if research is useful for their own ends
- Data is safely settled so source data can be purged
- Popular Technologies:
  - Globus Publish
  - Binder
  - Clowder

Borrowing Technology From Other Fields

- High Energy Physics
  - Institute for Research and Innovation in Software in High Energy Physics (IRIS-HEP)
  - DOMA: Data Organization, Management & Access
  - REANA
- Earth Science
  - Pangeo
- Open Source
  - Kubernetes
  - Spark
  - Jupyter

Some Key Technologies

- Clowder
- Brown Dog
- Whole Tale

Clowder: Data Sharing With Active and Social Curation

- **Active curation** involves recording data and metadata as close to the source as practical and driving that acquisition through the deployment of capabilities that help data producers manage their research.
- **Social Curation** drives this economic analysis further, looking at ways that cross group interactions can further motivate best practices.

J. Myers and M. Hedstrom, "Active and Social Curation: Keys to Data Service Sustainability," NDS Consortium Planning Workshop, 2014
http://sead-data.net/sites/default/files/pubs/ActiveandSocialCurationKeystoDataServiceSustainability.pdf

Brown Dog: A Global Data Transformation Service

- Data Transformation Service
  - Low-loss file format conversion
  - Automated metadata extraction
  - File repository independent
  - SDK to ease development of new tools
- Global
  - Public instances for complex tools
  - Local instances near your data
  - Private instances for proprietary tools
  - Centralized tools catalog to find solutions to data transformation needs

Whole Tale: Whole science story for the long and short tails of science

- Provide a living publication, preserving all digital scholarly objects, that can be shared and replayed
  - Input, intermediate, and derived data
  - Software and environment
  - Workflow process
  - Publication narrative
- Capture computational steps and provide compute environment
- Provides unique identifiers to objects (DOI)

bengal1@Illinois.edu

ILLINOIS
NCSA | National Center for
Supercomputing Applications