NSF Support for Advanced Data and Software Research Cyberinfrastructure

Dr. Stefan Robila

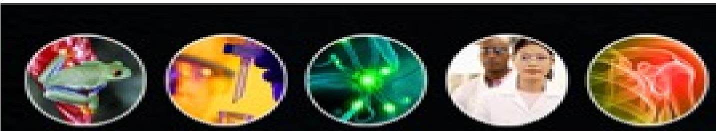
Program Director

Office of Advanced Cyberinfrastructure (OAC)

Directorate for Computer & Information Science & Engineering (CISE)

National Science Foundation (NSF)

Dec 5, 2019 IWCPLMI-Workshop, Gaithersburg, MD

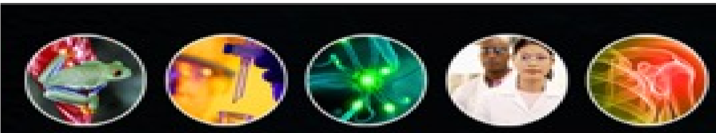


Outline

- NSF Overview
- CISE/OAC
- BIO/DBI
- ENG/DMREF

Outline

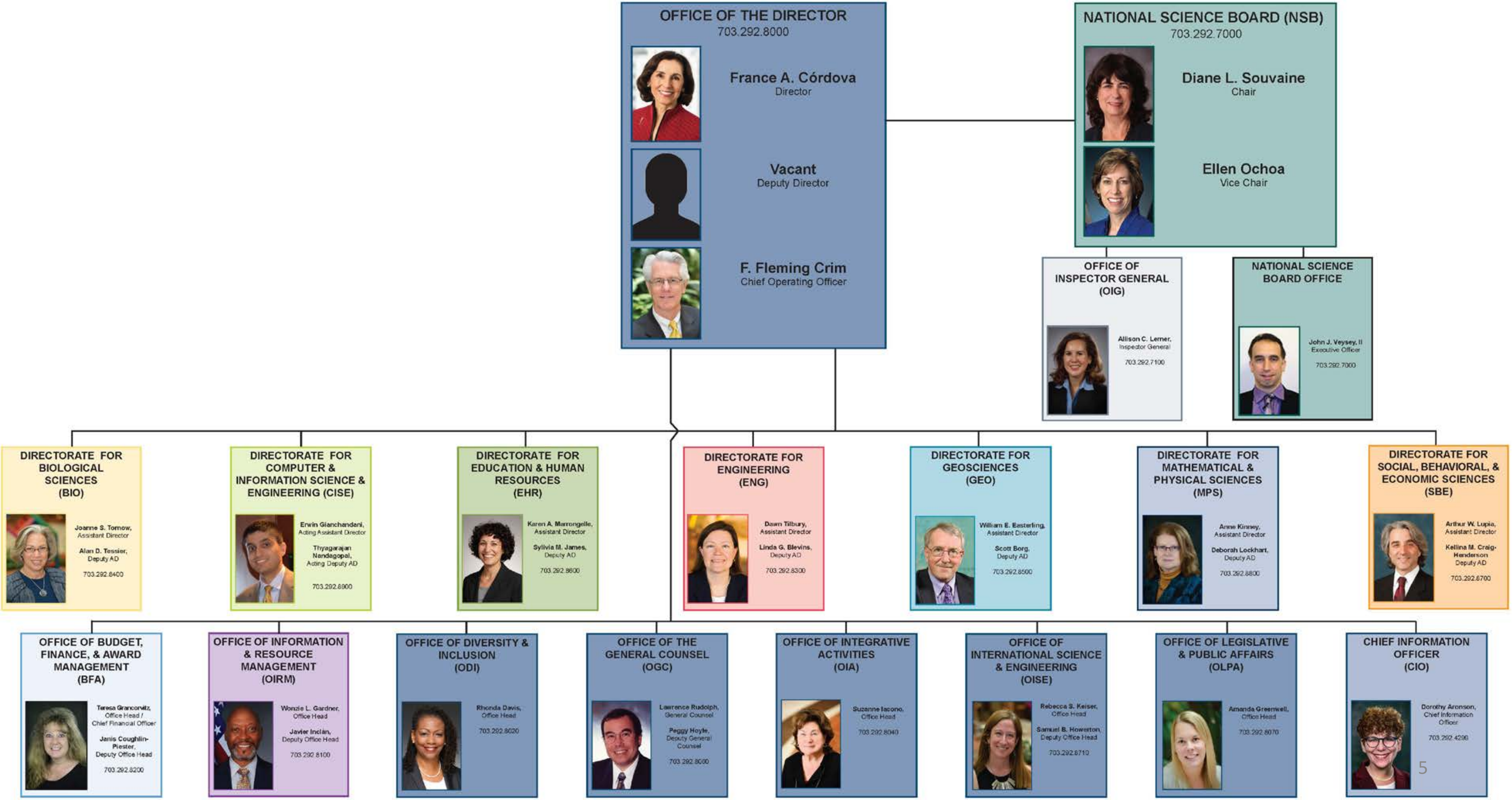
- NSF Overview
- CISE/OAC
- BIO/DBI
- ENG/DMREF



The National Science Foundation's mission

"To promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense..."

NATIONAL SCIENCE FOUNDATION



OFFICE OF THE DIRECTOR
703.292.8000

France A. Córdova
Director

Vacant
Deputy Director

F. Fleming Crim
Chief Operating Officer

NATIONAL SCIENCE BOARD (NSB)
703.292.7000

Diane L. Souvaine
Chair

Ellen Ochoa
Vice Chair

OFFICE OF INSPECTOR GENERAL (OIG)

Allison C. Lerner,
Inspector General

703.292.7100

NATIONAL SCIENCE BOARD OFFICE

John J. Veysey, II
Executive Officer

703.292.7000

DIRECTORATE FOR BIOLOGICAL SCIENCES (BIO)

Joanne S. Tornow,
Assistant Director

Alan D. Tessier,
Deputy AD

703.292.8400

DIRECTORATE FOR COMPUTER & INFORMATION SCIENCE & ENGINEERING (CISE)

Erwin Gianchandani,
Acting Assistant Director

Thyagarajan Nandagopal,
Acting Deputy AD

703.292.8900

DIRECTORATE FOR EDUCATION & HUMAN RESOURCES (EHR)

Karen A. Marrongelle,
Assistant Director

Sylivia M. James,
Deputy AD

703.292.8600

DIRECTORATE FOR ENGINEERING (ENG)

Dawn Tilbury,
Assistant Director

Linda G. Blevins,
Deputy AD

703.292.8300

DIRECTORATE FOR GEOSCIENCES (GEO)

William E. Easterling,
Assistant Director

Scott Borg,
Deputy AD

703.292.8500

DIRECTORATE FOR MATHEMATICAL & PHYSICAL SCIENCES (MPS)

Anne Kinney,
Assistant Director

Deborah Lockhart,
Deputy AD

703.292.8800

DIRECTORATE FOR SOCIAL, BEHAVIORAL, & ECONOMIC SCIENCES (SBE)

Arthur W. Lupia,
Assistant Director

Kellina M. Craig-Henderson
Deputy AD

703.292.8700

OFFICE OF BUDGET, FINANCE, & AWARD MANAGEMENT (BFA)

Teresa Grancorvitz,
Office Head /
Chief Financial Officer

Janis Coughlin-Piester,
Deputy Office Head

703.292.8200

OFFICE OF INFORMATION & RESOURCE MANAGEMENT (OIRM)

Wonzie L. Gardner,
Office Head

Javier Inclán,
Deputy Office Head

703.292.8100

OFFICE OF DIVERSITY & INCLUSION (ODI)

Rhonda Davis,
Office Head

703.292.8020

OFFICE OF THE GENERAL COUNSEL (OGC)

Lawrence Rudolph,
General Counsel

Peggy Hoyle,
Deputy General Counsel

703.292.8060

OFFICE OF INTEGRATIVE ACTIVITIES (OIA)

Suzanne Iacono,
Office Head

703.292.8040

OFFICE OF INTERNATIONAL SCIENCE & ENGINEERING (OISE)

Rebecca S. Keiser,
Office Head

Samuel B. Howerton,
Deputy Office Head

703.292.8710

OFFICE OF LEGISLATIVE & PUBLIC AFFAIRS (OLPA)

Amanda Greenwell,
Office Head

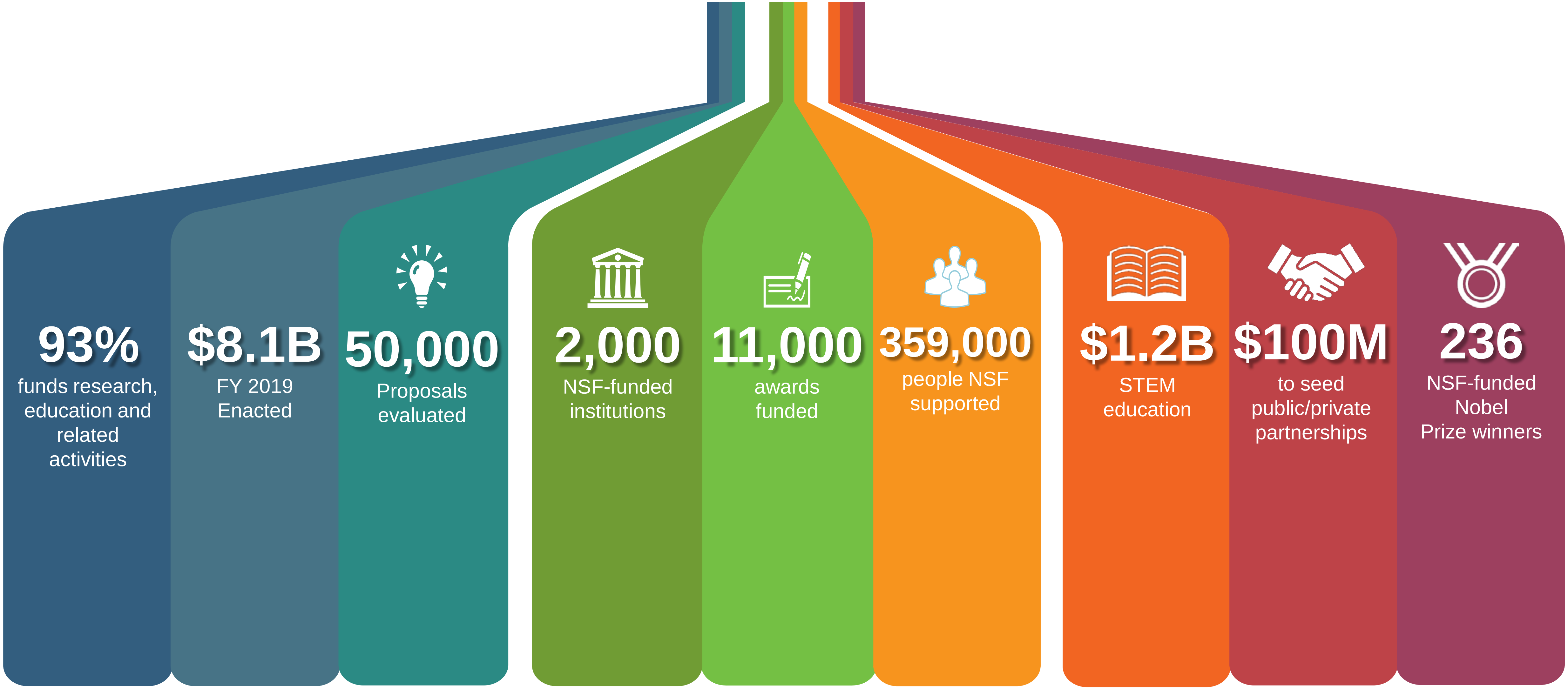
703.292.8070

CHIEF INFORMATION OFFICER (CIO)

Dorothy Aronson,
Chief Information Officer

703.292.4299

NSF by the numbers



93% funds research, education and related activities

$8.1B FY 2019 Enacted

50,000 Proposals evaluated

2,000 NSF-funded institutions

11,000 awards funded

359,000 people NSF supported

$1.2B STEM education

$100M to seed public/private partnerships

236 NSF-funded Nobel Prize winners

Most numbers based on FY 2018 activities.

Outline

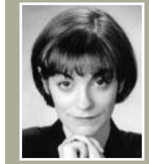
- NSF Overview
- CISE/OAC
- BIO/DBI
- ENG/DMREF

NSF Office of Advanced Cyberinfrastructure (OAC)

Directorate for Computer & Information Science & Engineering (CISE)

Mission: Foster a cyberinfrastructure ecosystem to transform science and engineering research … through Research CI and CI research

Manish Parashar *
Office Director


Amy Friedlander
Deputy Office Director

* IPA Appointment

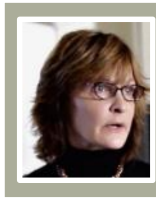
Bob Chadduck

Amy Walton

Vipin Chaudhary *

Kevin Thompson

Bill Miller
Science Advisor


Beth Plale*
Science Advisor
Public Access

Ed Walker

Stefan Robila *

Micah Beck *

Alan Sussman *

Alejandro Suarez
Cooperative Agreements

CISE/OAC Research Infrastructure Investments

- Lead interagency effort to build and support an expansive CI ecosystem driven by research priorities and the scientific process
- Leverage investments by universities, federal agencies, commercial sector
- Establish viable path forward for HPC systems in post-Moore’s Law era; and
- Increase capacity, capability, and sustainability
- Support a diversity of computational resources to meet the growing demands of modern science and engineering
- Align with the National Strategic Computing Initiative (NSCI)

Office of Advanced Cyberinfrastructure (OAC)

Foster a cyberinfrastructure ecosystem to transform science and engineering research…

… through Research CI and CI research

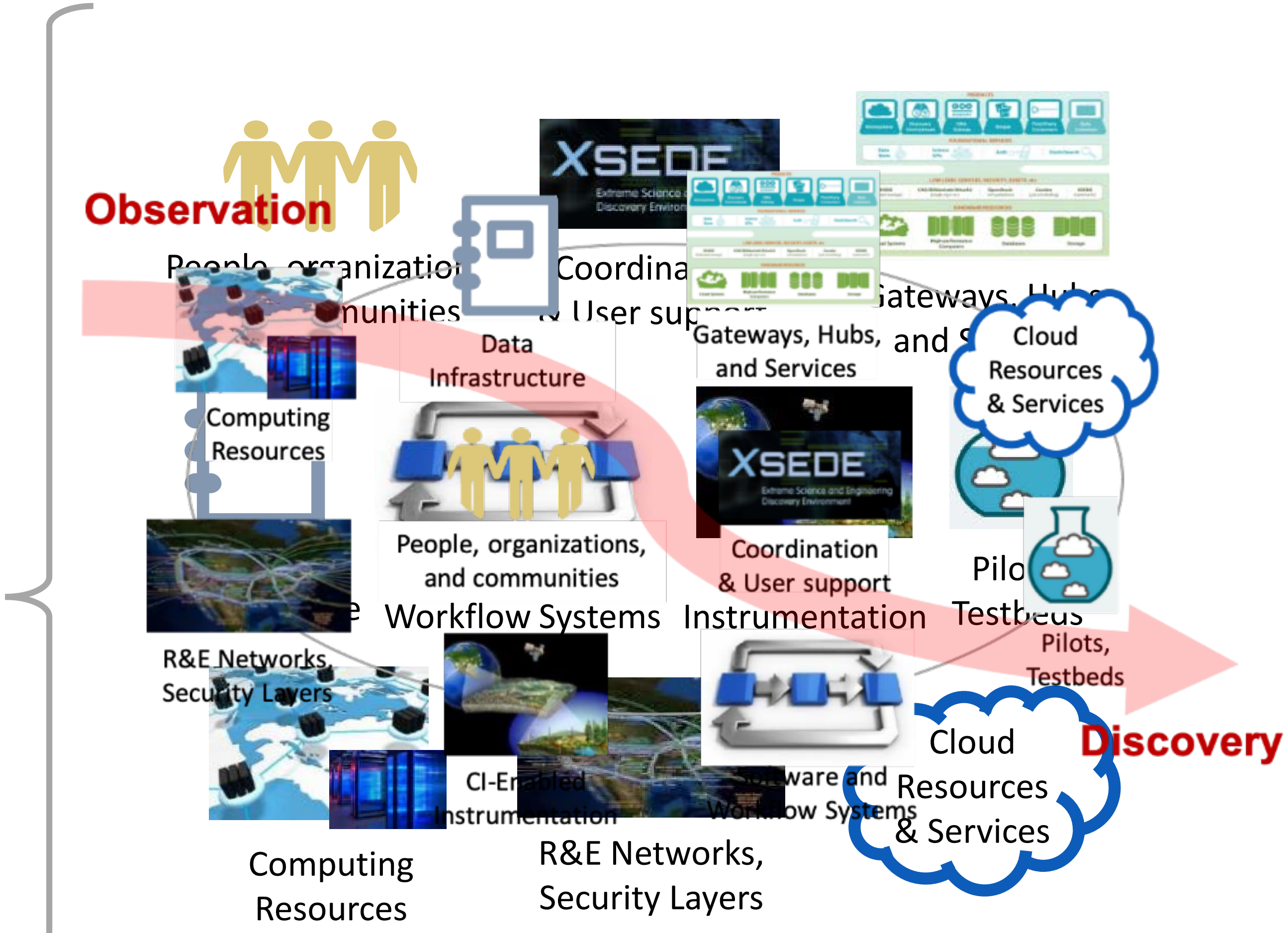
$224M FY 2018 research budget

950 proposals

305 awards

32% Success Rate

Source: https://dellweb.bfa.nsf.gov/starth.asp


Observation
People, organizations, and communities
XSEDE
Coordination & User support
Gateways, Hubs, and Services
Data Infrastructure
Cloud Resources & Services
Computing Resources
People, organizations, and communities
Coordination & User support
Workflow Systems
Instrumentation
Pilots, Testbeds
R&E Networks, Security Layers
CI-Enabled Instrumentation
Software and Workflow Systems
Cloud Resources & Services
Discovery
Computing Resources
R&E Networks, Security Layers

Learning and Workforce Development and Core Research

Communities of Concern

CI Contributors

Cyber Scientists
to develop
new capabilities

CI Professionals

Professional Staff
to deploy & support
new capabilities

CI Users

Area Scientists
to exploit
new capabilities

OAC

CyberTraining

NSF 19-524

OAC-Core Research

NSF 19-589

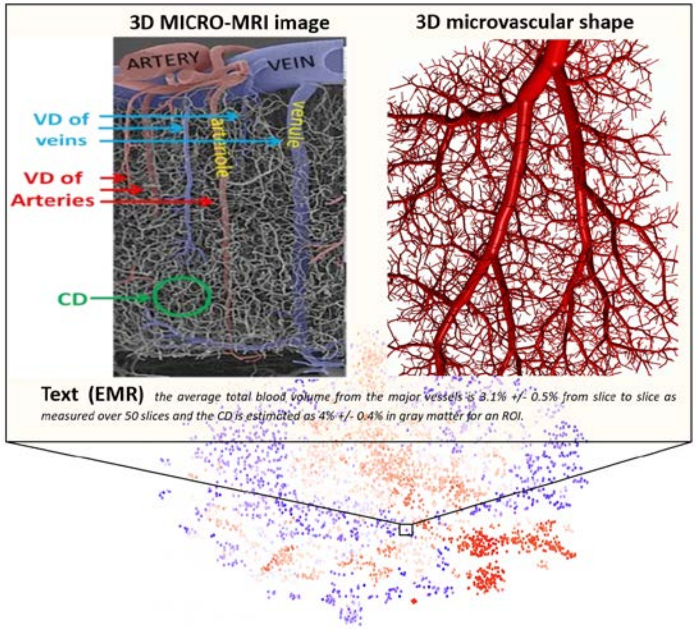
OAC Core: Small: Shape-Image-Text: A Data-Driven Joint Embedding Framework for Representing and Analyzing Large-Scale Brain Microvascular Data

PI: Zichun Zhong, Co-PI: Jing Hua, Wayne State University [Award #OAC-1910469]

(Co-funded by CBET)

A rigorous and scalable computing platform to construct a unified representation and correlation for large sets of multimodal data (i.e., shape, image, and text) through fully data-driven joint embeddings.

- Transform a 3D shape with heterogeneous imaging, textual, and other features from a large dataset to a novel high-dimensional isometric multi-view probability space.
- Allow formal and diverse study of geometry scalability and variability in shape processing and measurement intensively involved in 3D multimodal data, especially the large-scale microvascular networks.
- Significantly increase system's automation, reduce human's interventions, and discover new knowledge in vascular diseases in the joint embedding space, e.g., Microvascular-Multimodal-Embedding.
- Offer an accurate and robust approach for geometric reasoning and quantitative assessment of multi-heterogeneous and multimodal data features across a large number of objects.
- Provide several educational activities for undergraduate, graduate, and local middle school students.



Text (EMR) *the average total blood volume from the major vessels is 3.1% +/- 0.5% from slice to slice as measured over 50 slices and the CD is estimated as 4% +/- 0.4% in gray matter for an ROI.*

Microvascular-Multimodal-Embedding (MME) system

Slide Provided by PI

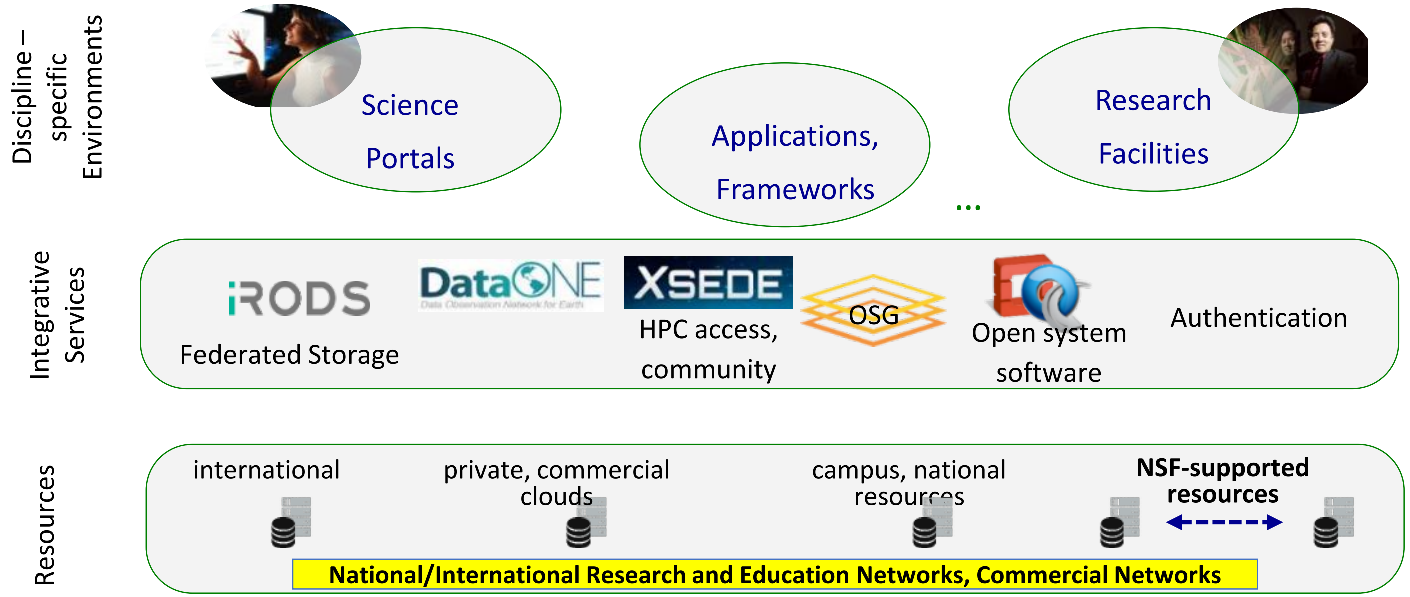
FRONTERA

- New NSF "Flagship" system at TACC/UT-Austin, **available NOW** for large users
- Frontera is the **#5** ranked system in the world – and the fastest at any university in the world.
  - Fastest primarily Intel-based system
  - Highest ranked Dell system ever.
- Dell Servers, Intel Processors, provide 39PF peak (double precision)
- 800+ GPUs for single precision/ML
- 50+ PB of storage, 1.5TB/sec I/O

FRONTERA

Slide Provided by PI

OAC Data and Software Investments



Discipline – specific Environments

Science Portals

Applications, Frameworks

Research Facilities

…

Integrative Services

iRODS Federated Storage

DataONE

XSEDE HPC access, community

OSG

Open system software

Authentication

Resources

international

private, commercial clouds

campus, national resources

NSF-supported resources

National/International Research and Education Networks, Commercial Networks

Guiding Principles

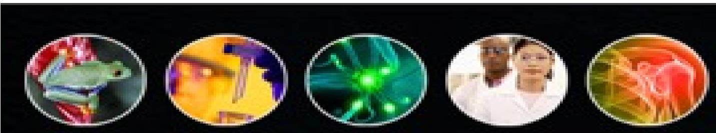
- *Science-driven*
- *Innovation*
- *Close collaborations among stakeholders*
- *Building on existing, recognized capabilities*
- *Project plans, and system and process architecture*
- *Clear deliverables*
- *Measurable outcomes*
- *Sustained and sustainable impacts*

Serves an important and diverse set of research applications and users

- Supports research projects in every NSF science and engineering discipline / Enables new areas of research.
- Cyberinfrastructure for Sustained Scientific Innovation (CSSI). Supports CI/discipline collaborations, cross-disciplinary infrastructure, builds on recognized capabilities, tangible products, Cross directorate investments. **NSF 19-548 Deadline: November 1, 2019**
- CSSI Institutes. Focuses on the establishment of long-term hubs of excellence in software infrastructure and technologies, which will serve a research community of substantial size and disciplinary breadth.
- CC* (Campus Cyberinfrastructure) - Multi--institution collaborations, cloud resources, sharing mechanisms, innovative storage.

Cyberinfrastructure for Sustained Scientific Innovation (CSSI) - Data and Software: Elements and Frameworks - NSF 19-548

- Supports the development and deployment of robust, reliable and sustainable data and software cyberinfrastructure
- Brings innovative capabilities towards sustained scientific innovation and discovery
- Provides a cross-directorate opportunity to advance common approaches to sustain and innovate research cyberinfrastructures.
- Follows accepted data management and software development practices



OAC Data and Software Programs

HYDROSHARE

http://www.hydroshare.org

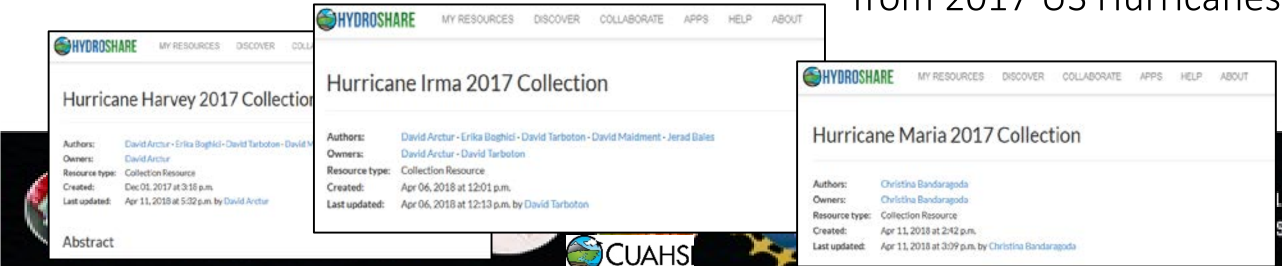
OAC-1664061
OAC-1664018
OAC-1664119
2017-2021

ACI-1148453
ACI-1148090
2012-2017

- Web-based Hydrologic Information System operated by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI)
- Gateway for web based computational analysis and modeling functionality to overcome desktop compatibility, dependency and capacity limitations
- Gives users a way to share datasets, models, and other research products
- Provides permanent publication of data and models with citable digital object identifiers that can link to literature

Better hydrologic forecasting that quantifies effects and consequences of land surface change on hydrologic processes and conditions by enabling access to and organizing data for integrated analysis and modeling

Example: Collection of Data from 2017 US Hurricanes



Hurricane Harvey 2017 Collection

Authors: David Arctur · Erika Boghici · David Tarboton · David M
Owners: David Arctur
Resource type: Collection Resource
Created: Dec 01, 2017 at 3:16 p.m.
Last updated: Apr 11, 2018 at 5:32 p.m. by David Arctur

Abstract

Hurricane Irma 2017 Collection

Authors: David Arctur · Erika Boghici · David Tarboton · David Maidment · Jerad Bales
Owners: David Arctur · David Tarboton
Resource type: Collection Resource
Created: Apr 06, 2018 at 12:01 p.m.
Last updated: Apr 06, 2018 at 12:13 p.m. by David Tarboton

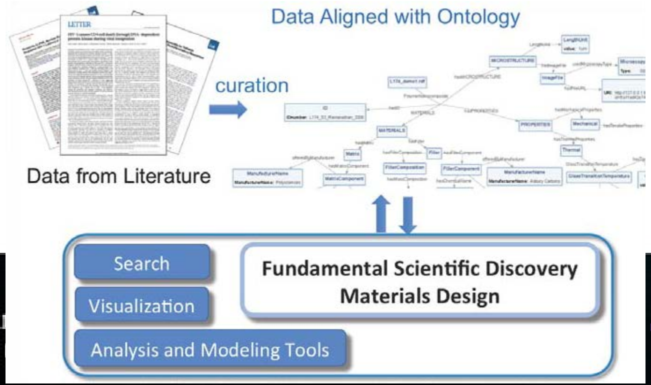
Hurricane Maria 2017 Collection

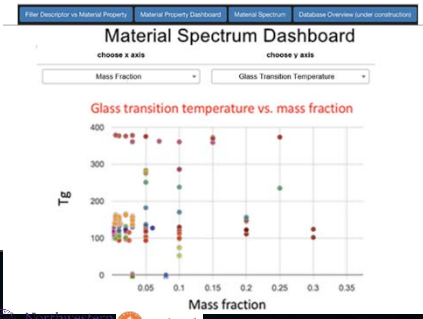
Authors: Christina Bandaragoda
Owners: Christina Bandaragoda
Resource type: Collection Resource
Created: Apr 11, 2018 at 2:42 p.m.
Last updated: Apr 11, 2018 at 3:09 p.m. by Christina Bandaragoda

Nanocomposites to Metamaterials: A Knowledge Graph Framework

OAC – 1835677
2018 - 2023

- Data framework for materials discovery through physics-based modeling and machine learning. Create a Materials Knowledge Graph (MKG) as a materials data framework by developing an extensible semantic infrastructure with ontology-enabled design and analysis tools
- Develop broad use and integration with industry, national labs, and materials research community
- Drastically reduce deployment time of new materials with targeted properties, expand prior work in nanocomposites and metamaterials
- Support an all women research team
- Transdisciplinary Education: mechanics, materials science, engineering design, computer science, and machine learning




Filler Descriptor vs Material Property
Material Property Dashboard
Material Spectrum
Database Overview (under construction)
Material Spectrum Dashboard
choose x axis
choose y axis
Mass Fraction
Glass Transition Temperature
Glass transition temperature vs. mass fraction
Tg
0
100
200
300
400
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Mass fraction

Framework: Software: Next-Generation Cyberinfrastructure for Large-Scale Computer-Based Scientific Analysis and Discovery

Alan Edelman and Juan Pablo Vielma Centeno, MIT (1835443)

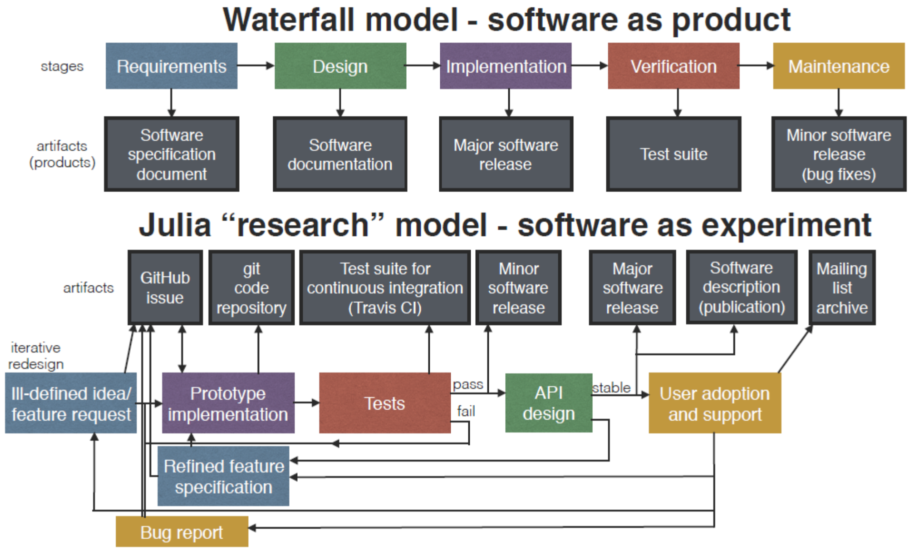
Forge a collaboration between domain and computing experts to build cyberinfrastructure that enables the next generation of computer-based scientific analysis/discovery.

julia Programming

- Over 2532 registered packages
- 7.6M downloads
- 78% annual growth

Contributions:

- Design and implementation of new programming language abstractions to allow close integration of high-level language features with low-level compiler optimizations.
- Build upon the Julia programming language and the Julia-based JuMP modeling language for mathematical optimization.
- Focus on three target scientific applications related to
  - (i) the deep decarbonization of electrical power networks,
  - (ii) image analysis of extreme scale astronomical data, and
  - (iii) pharmacometric models for drug analysis and discovery.



Slide Provided by PI

NSF 20-015, Request for Information (RFI) on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

- **Motivation:** The challenges of growing volumes of scientific data – their availability, transmission, accessibility, management, and utilization – have become urgent and ubiquitous across NSF-supported science, engineering, and education disciplines. Follow on to NSF CI 2030 RFI, towards further development of the OAC CI "vision" (https://www.nsf.gov/cise/oac/vision/blueprint-2019/)

- **Goal:** To update NSF on their data-intensive scientific questions and challenges and associated needs specifically related to data-focused cyberinfrastructure.

- Emphasis on "**cross-disciplinary and domain-agnostic solutions**"

  Q1: Current or emerging data-intensive/data-driven research challenge(s)
  Q2: Data-oriented CI needed to address the research questions/challenge(s)
  Q3: Other considerations.

https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp

Please disseminate to your colleagues!

Outline

- NSF Overview
- CISE/OAC
- BIO/DBI
- ENG/DMREF

Division of Biological Infrastructure (BIO/DBI)

Empowers biological discovery

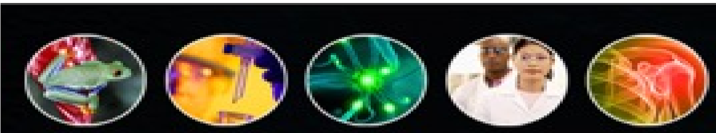
- *Research Resources* (development of informatics and cyberinfrastructure resources; new instrumentation; the curatorial improvement and digitization of research collections, including living stock collections; and the improvements of research facilities at biological field stations and marine laboratories)

- *Human Resources* (development of the biology scientific workforce - support fellowships for postdoctoral research, sites for research experiences for undergraduates, and research coordination networks in undergraduate biology education)

- *Centers and Other Mid-to-Large-Scale Infrastructure* (Addresses targeted biological questions focused toward the needs of a particular research and education community in the biological sciences through managing synthesis centers, science and technology centers, nanoscale science and engineering centers, and other activities)

Peter McCartney (pmccartn@nsf.gov)

Jean Gao (jgao@nsf.gov)



Infrastructure Innovation

NSF 18-595 Infrastructure Innovation for Biological Research (IIBR):

Bioinformatics Research Program

- Support for research on novel methods, approaches, and technologies for capturing, representing, and analyzing biological phenomena in digital form.
- High emphasis on innovation and potential; high tolerance for risk.
- Success measured by the outcomes of the project's research.
- Most closely related to Computational and Data Enabled Science and Engineering (CDS&E) but with a focus on biological applications.

- T.M. Murali & Jean Peccoud: Automated Prioritization and Design of Experiments to Validate and Improve Mathematical Models of Molecular Regulatory System. (https://peccoud.org/automated_design_experiments/)
- Dannie Durand: Domain Architecture simulator. (http://www.cs.cmu.edu/~durand/Lab/research.html)
- Christopher Topp: Algorithms for recovering root architecture from 3D imaging. (https://www.danforthcenter.org/scientists-research/principal-investigators/chris-top)

Infrastructure Capacity

NSF 18-594 Infrastructure Capacity for Biology (ICB):

Cyberinfrastructure for Biological Research Program

- Development of robust cyberinfrastructure such as databases, software, knowledge bases, and other computational resources to enable biological research.
- High emphasis on design, engineering, and community impact; low tolerance for risk.
- Success measured by outcomes of the user community.
- Most closely related to Cyberinfrastructure for Sustained Scientific Innovation (CSSI) but with a focus on biological applications.

Galaxy

Galaxy genome analysis software (www.galaxyproject.org)

PEcAn

Predictive Ecosystem Analyzer (pecanproject.github.io/)

MOVEBANK For Animal Tracking Data

Movebank database for animal tracking data. (www.movebank.org)

Infrastructure Sustainability

NSF 19-569 Sustained Availability of Biological Infrastructure(SABI):

- Limited support for costs of ongoing operations and maintenance of mature, critical cyberinfrastructure.
- Main emphasis on community impact; zero risks because only mature resources with established user communities are considered.
- "Criticality" measured by magnitude and productivity of the user community.
- Most closely related to Operations & Maintenance portion of Major Research Equipment and Facilities Construction projects (MREFC) such as NEON or OOI.

CyVerse computational infrastructure (formerly iPlant) (www.cyverse.org)

DataDryad repository for research data (www.datadryad.org)

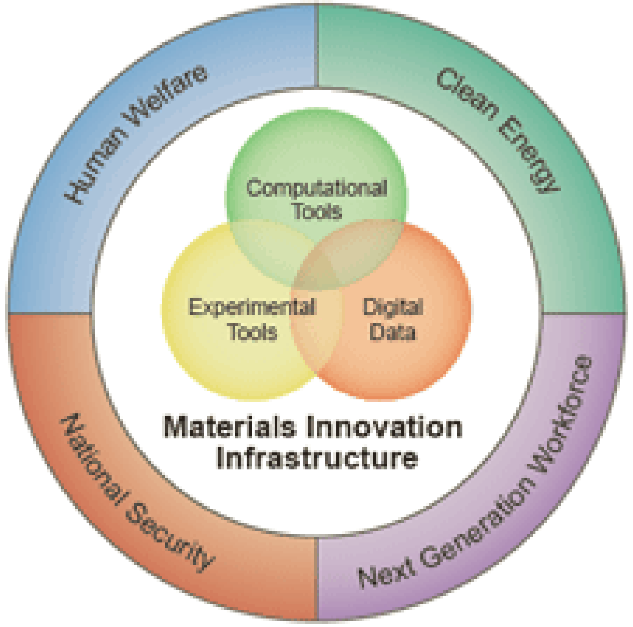
Protein Data Bank repository for 3D structural biology data (www.rcsb.org)

Outline

- NSF Overview
- CISE/OAC
- BIO/DBI
- ENG/DMREF

Designing Materials to Revolutionize and Engineer our Future (DMREF)

- DMREF is NSF's Response to and participation in the Materials Genome Initiative



NSF is interested in activities that accelerate materials discovery and development by building the fundamental knowledge base needed to progress towards designing and making a material with a specific and desired function or property from first principles.

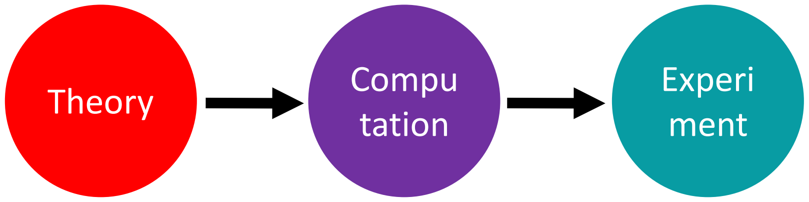
The DMREF goal is to control material properties through design: this is to be accomplished by understanding the interrelationships of composition, processing, structure, properties, performance, and process control.

Materials Design:
Theory, Computation and Experiment

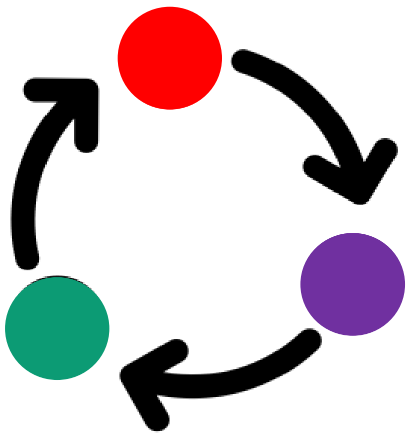
- For DMREF Awards, Theory, Computation and Experiment must guide one another in a closed, iterative feedback loop

- This loop must accelerate materials design, discovery and deployment

- Conventional “Predict – Synthesize – Test” approach is linear; the loop must be “closed” by improving models, theories or algorithms based on experimental measurements



Linear path to Materials Discovery



Iterative feedback loop for Materials Discovery

Open Access to Data and Codes

- In the spirit of the MGI, Open Access to data and codes developed in DMREF proposals is a key criterion

- Proposals must make data accessible (not simply available), including codes, software and algorithms.

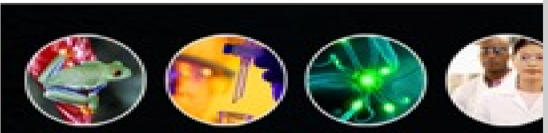
- Numerous Open Access/Open Science activities are underway at NSF-
- Keep an eye out for upcoming opportunities!

DMREF Focus: Workforce Development

- **Community Study: Creating the Next-Generation Materials Genome Initiative Workforce**

- The final report will address the current state of the academic **curriculum and training** approaches of the U.S. workforce to accomplish MGI goals, identify the **key MGI skill requirements** and needs for individuals entering the workforce, and outline curricula development and training guidelines to **improve readiness** of current students and the existing professional workforce.

- Led by the Minerals, Metals and Materials Society (TMS)
- Publication: Fall 2019
- Available for free download at www.tms.org/studies



Example Project: DMREF

- DMREF: Data-Driven Integration of Experiments and Multi-Scale Modeling for Accelerated Development of Aluminum Alloys

CMMI-1921959

Hufnagel et al, Johns Hopkins University

Co-funded by the Division of Civil, Mechanical and Manufacturing Innovation, Division of Materials Research and Office of Advanced Cyberinfrastructure

3D Characterization
EBSD + Serial Sectioning
High Energy Diffraction Microscopy

Image-based Modeling of Deformation

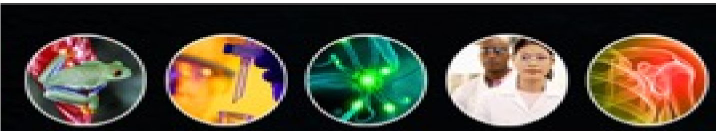
Automated feedback through Materials Semantic Infrastructure (OpenMSI)

Conclusion

- Science and society are being transformed by compute and data – a connected, robust and secure cyberinfrastructure ecosystem is essential

- Rapidly changing application requirements; resource and technology landscapes
  - Our cyberinfrastructure ecosystem must evolve in response

- NSF strives to build a cyberinfrastructure ecosystem aimed at transforming science



Thank You!

Stefan Robila

srobila@nsf.gov