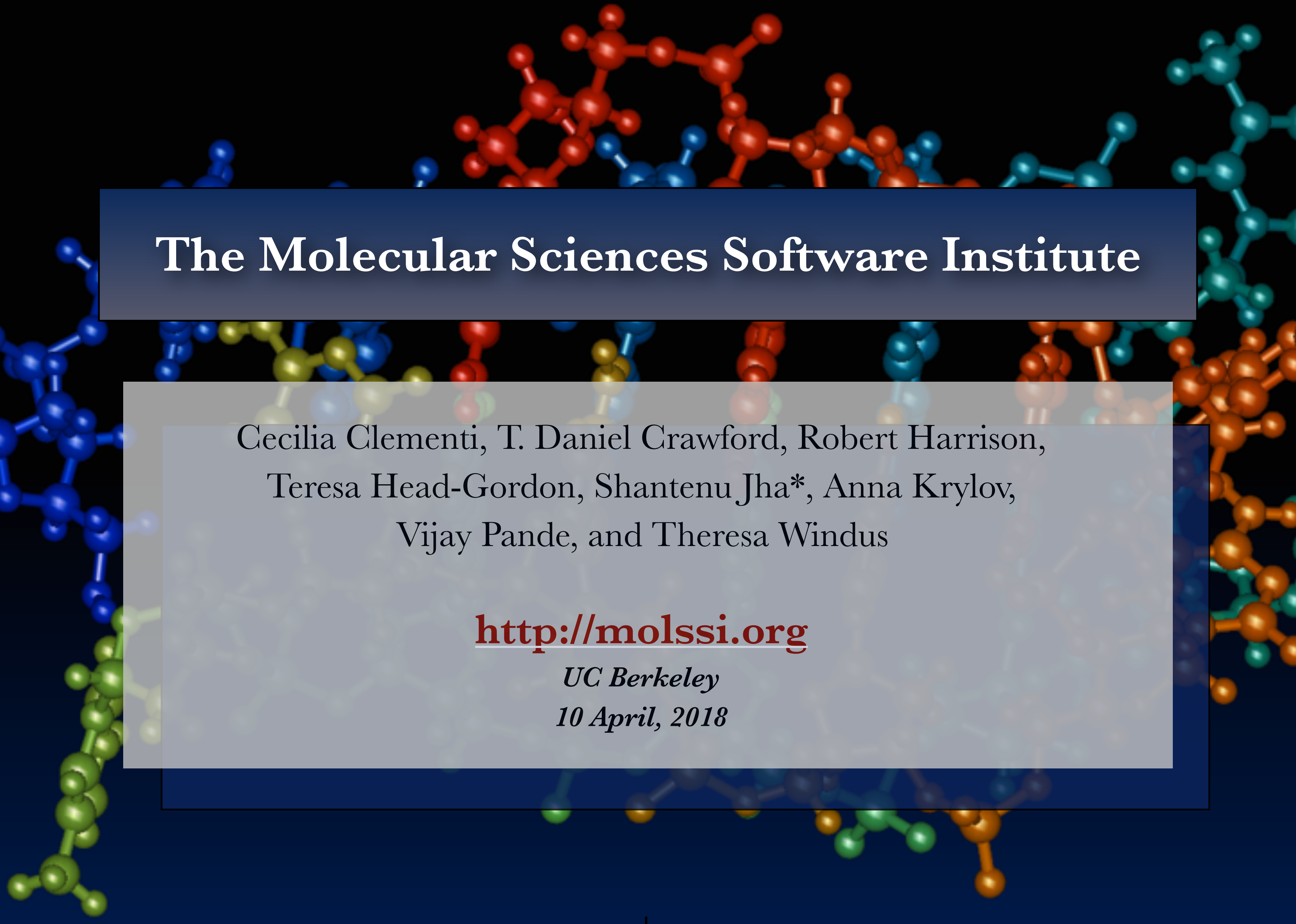


The Molecular Sciences Software Institute

Cecilia Clementi, T. Daniel Crawford, Robert Harrison,
Teresa Head-Gordon, Shantenu Jha*, Anna Krylov,
Vijay Pande, and Theresa Windus

http://molssi.org

UC Berkeley

10 April, 2018

The Molecular Sciences Software Institute (MolSSI)

- Project (start date of August 1st, 2016) funded by the National Science Foundation.
- Collaborative effort by Virginia Tech, Rice U., Stony Brook U., U.C. Berkeley, Stanford U., Rutgers U., U. Southern California, and Iowa State U.
- Total budget of $19.42M for five years, potentially renewable to ten years.
- Joint support from numerous NSF divisions: Advanced Cyberinfrastructure (ACI), Chemistry (CHE), Division of Materials Research (DMR), Office of Multidisciplinary Activities (OMA)
- Designed to **serve** and **enhance** the software development efforts of the broad field of computational molecular science.

Computational Molecular Sciences (CMS)

- The history of CMS – the sub-fields of **quantum chemistry**, **computational materials science**, and **biomolecular simulation** – reaches back decades to the genesis of computational science.
- CMS is now a **"full partner with experiment"**.
- For an impressive array of **chemical**, **biochemical**, and **materials** challenges, our community has developed simulations and models that directly impact:
  - Development of new chiral drugs;
  - Elucidation of the functionalities of biological macromolecules;
  - Development of more advanced materials for solar-energy storage, technology for CO_2 sequestration, etc.

CMS Codes Are Developed and Used Globally

CMS Codes Are Developed and Used Globally

These codes represent decades of development by thousands of programmers, and are used by tens of thousands of scientists worldwide.

Code Complexity and Historical Legacy

- CMS programs contain millions of lines of hand-written code and require hundreds of programmers to develop and maintain.
- Incredible language diversity: F77, F90, F95, HPF, C, C++, C++11, Python, perl, etc.
- Incredible algorithmic diversity: structured and unstructured grids, dense and sparse linear algebra, graph traversal, fast Fourier transforms, MapReduce, and more.
- The packages have evolved in an *ad hoc* manner over decades because of the intricacy of the scientific problems they are designed to solve
- …. against a backdrop of rapidly evolving hardware and computing platforms!

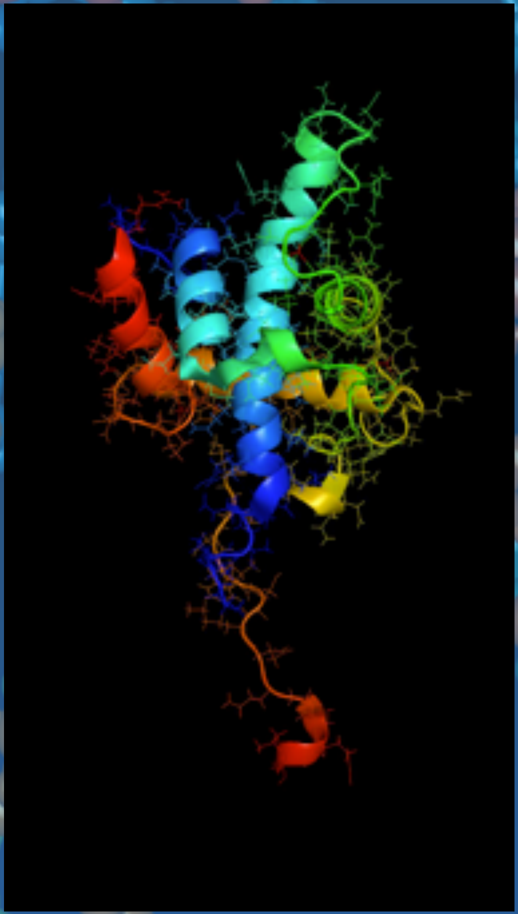
Inertia in the Scientific Education Culture

- Undergraduate programs in chemistry and physics typically require no training in software development or programming.
- Graduate programs in these areas require minimal coursework between the bachelor and Ph.D.
- Most computer science students lack the underlying knowledge of the scientific domains to help develop creative software solutions.
- Due credit for software development is elusive due to a culture that judges productivity based on citations of peer-reviewed papers.
- Thus, a “just get the physics working” approach pervades much of CMS software development.

What Scientific Areas *Could* We Enable?

Structure-Function of Intrinsically Disordered Proteins

- Required to understand biochemical function and disease: cellular regulation and signaling; associated with cancer, diabetes, and Alzheimers.
- New area of structural biochemistry: ~25% of proteome consists of proteins that are fundamentally dynamic in nature, with no intrinsic order.
- Computational models must fill experimental gap: IDPs confound spectroscopic characterization such that structure-function is highly underdetermined
- Importance of adequate sampling and workflows: accurate potential energy surfaces, aggressive sampling methods, probabilistic models, developed in state-of-the-art codes and analysis tools.



The IDP TAZ1-domain–CITED2 complex (PDB: 1R8U)

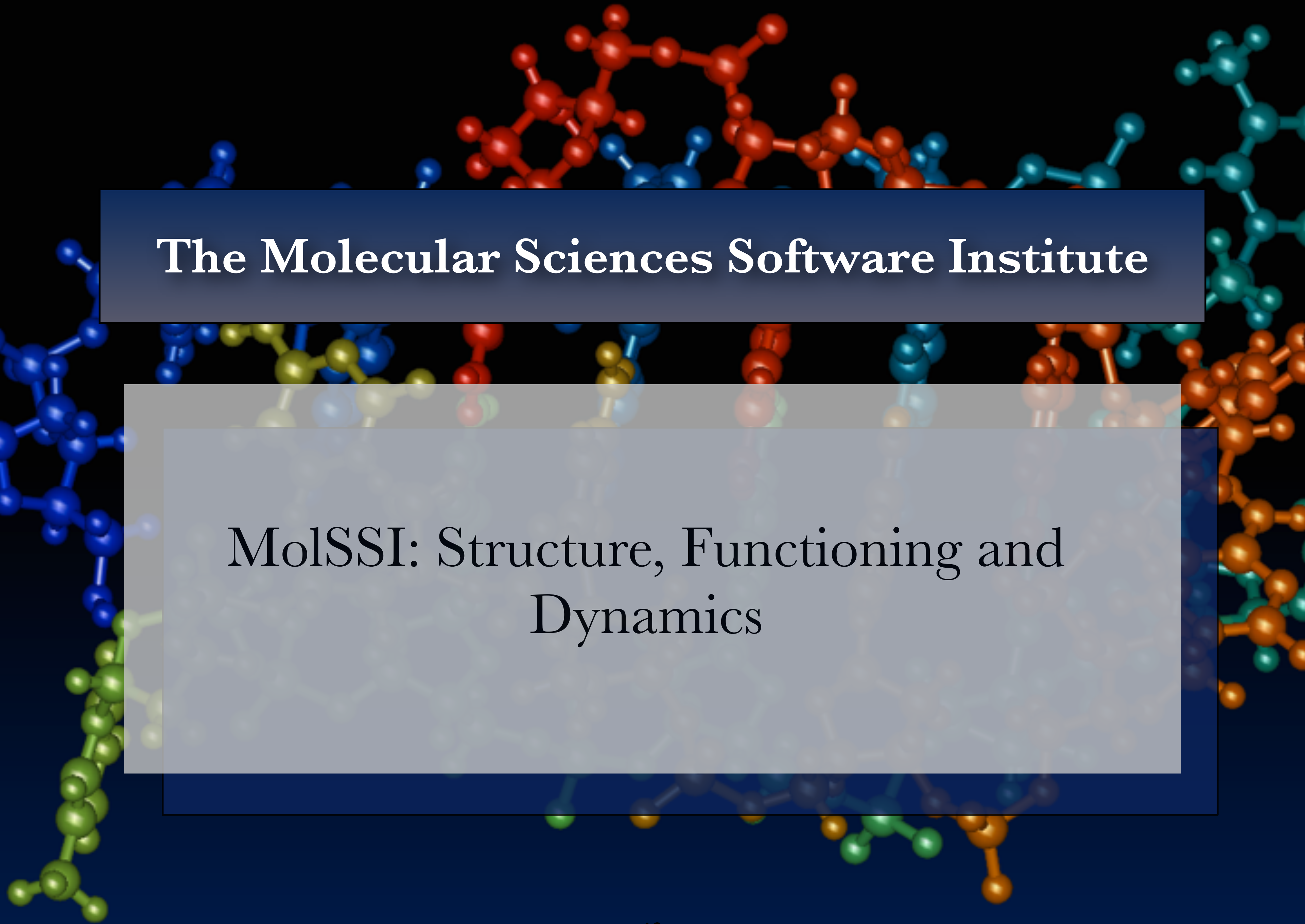
Conceptualization Phase and Activities

BMS: S2I2 Conceptualization Activities

- Started September 2014
  - Cecilia Clementi (Rice), Teresa Head-Gordon (Berkeley), Shantenu Jha (Rutgers) and Vijay Pande (Stanford)
  - https://sites.google.com/site/s2i2biomolecular/
- Two workshops at Berkeley (November 2014) and Houston (January 2015) focused on BMS engagement, requirements and the development of a community-wide vision for the Institute.
  - Overlap in workshop scope but different parts of the community
  - Houston Workshop had a focussed session on Cyberinfrastructure Problems in Molecular Simulation
- Informal "brain storming" and community engagement at other meetings
- CI Workshop (Rutgers) postponed due to solicitation!

QM: S2I2 Conceptualization Activities

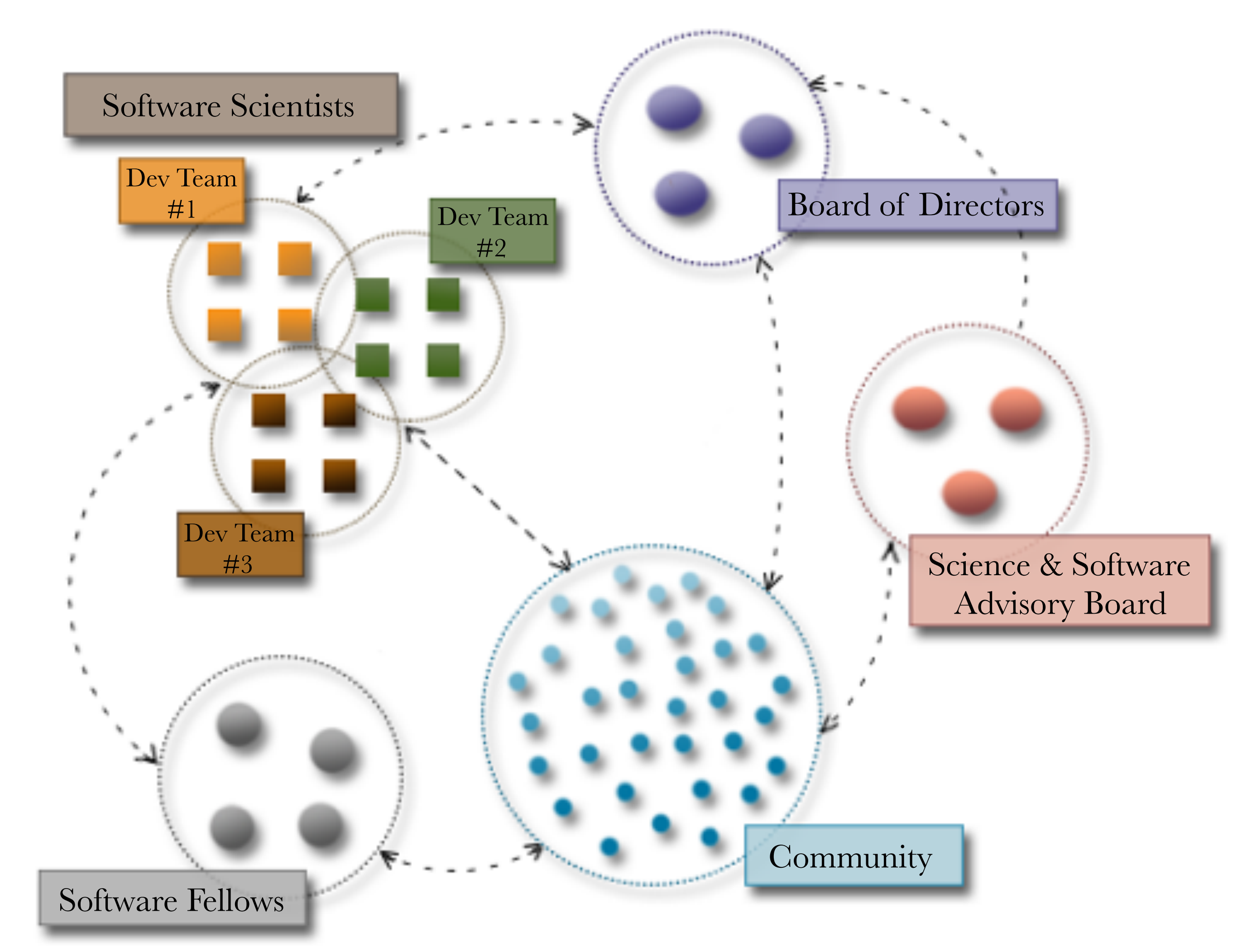
- Kick-off meeting in 2013 to develop initial vision of Institute
- Three workshops in 2013 and 2014 focused on potential software framework targets and community needs:
    - Portable parallel infrastructure (Manhattan, NY)
    - Code and data interoperability (Blacksburg, VA)
    - Tensor representations and algebras (Laguna Beach, CA)
- 2015 Summer Training Workshop in Biomolecular Simulations (Pasadena, CA);
- Three software summer schools (2013-15) for more than 100 students;
- Symposium at the ACS National Meeting in San Francisco (2014);
- Dozens of presentations at national and international conferences to encourage community engagement.
- Conceptualization activity separate to BMS conceptualization!
    - We're all molecular scientists (now)!



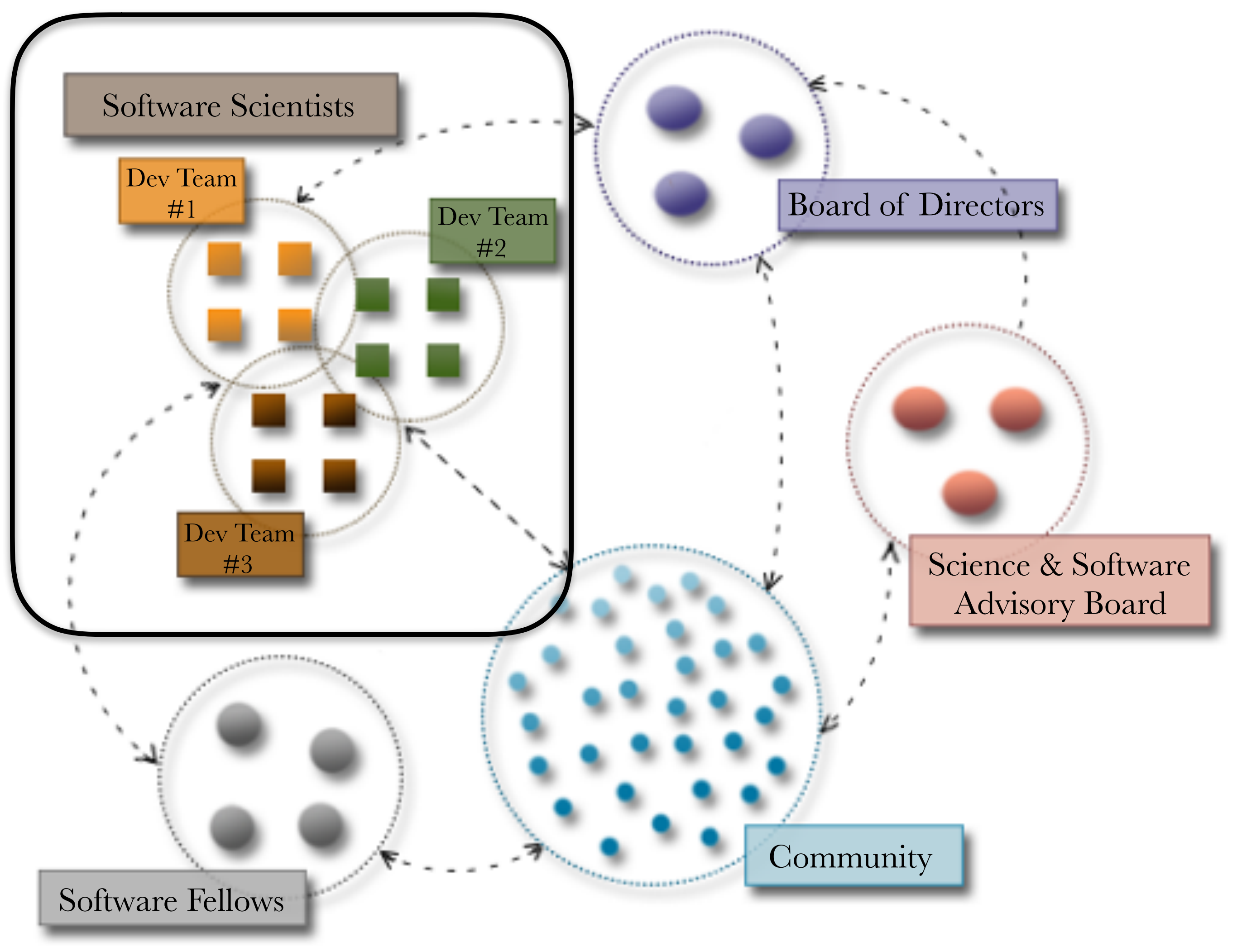
The Molecular Sciences Software Institute

MolSSI: Structure, Functioning and Dynamics

The Molecular Sciences Software Institute (MolSSI)



The Molecular Sciences Software Institute (MolSSI)

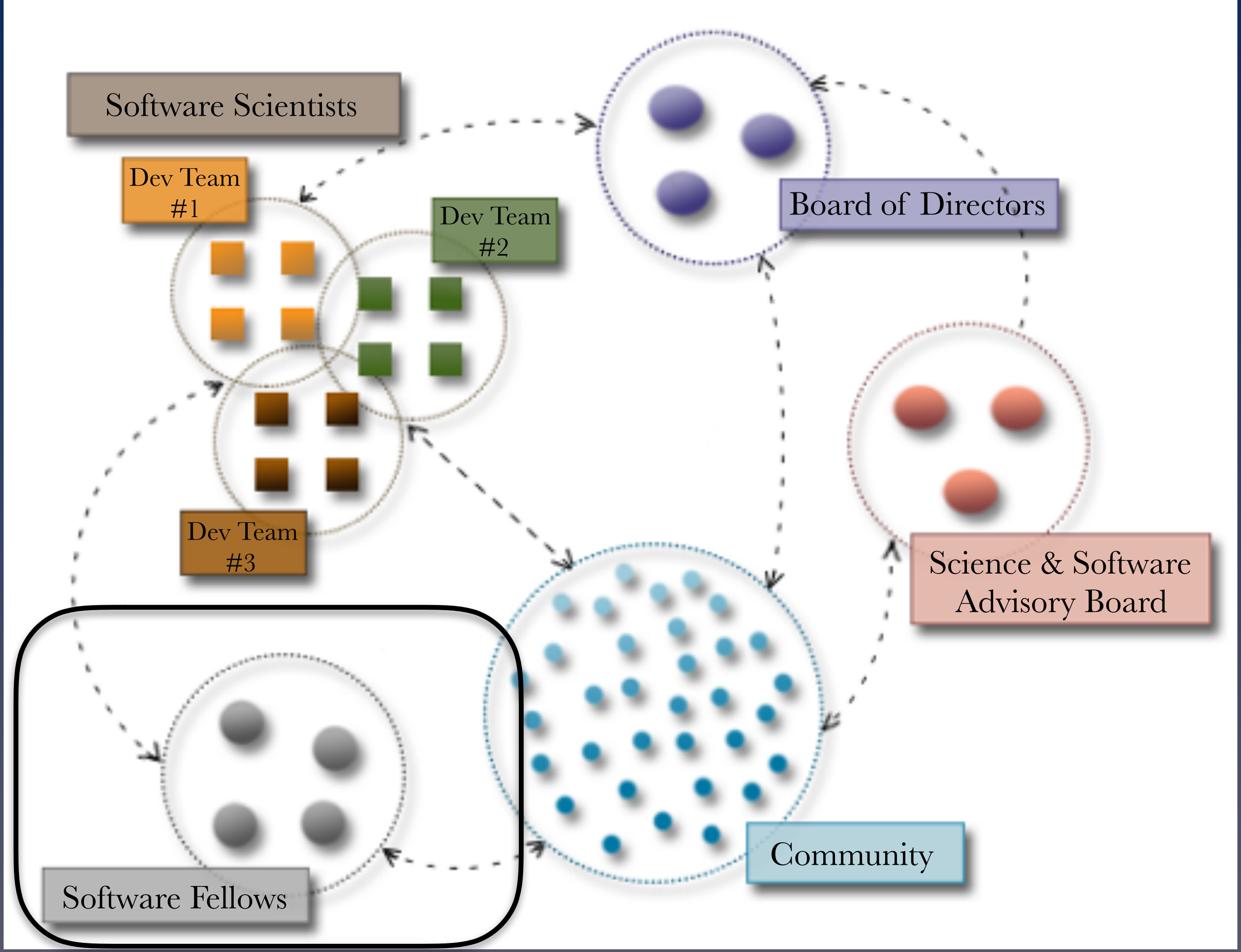


The MolSSI Software Scientists (MSSs)

- A team of ~12 software engineering experts, drawn both from newly minted Ph.D.s and established researchers in molecular sciences, computer science, and applied mathematics.
- Dedicated to multiple responsibilities:
  - Developing software infrastructure and frameworks;
  - Interacting with CMS research groups and community code developers;
  - Providing forums for standards development and resource curation;
  - Serving as mentors to MolSSI Software Fellows;
  - Working with industrial, national laboratory, and international partners;

Approximately 50% of the Institute's budget will directly support the MolSSI Software Scientists.

The Molecular Sciences Software Institute (MolSSI)

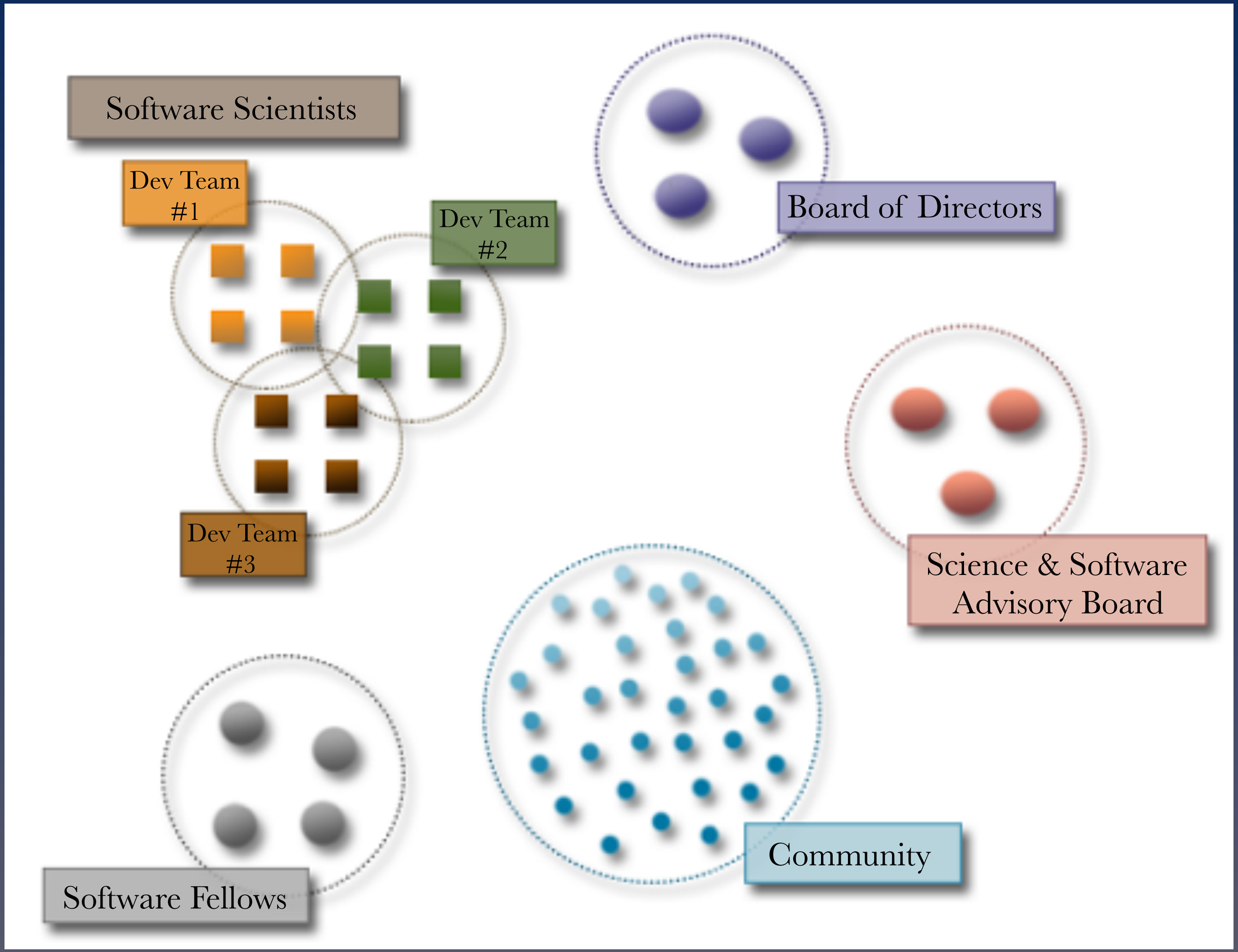


The MolSSI Software Fellows (MSFs)

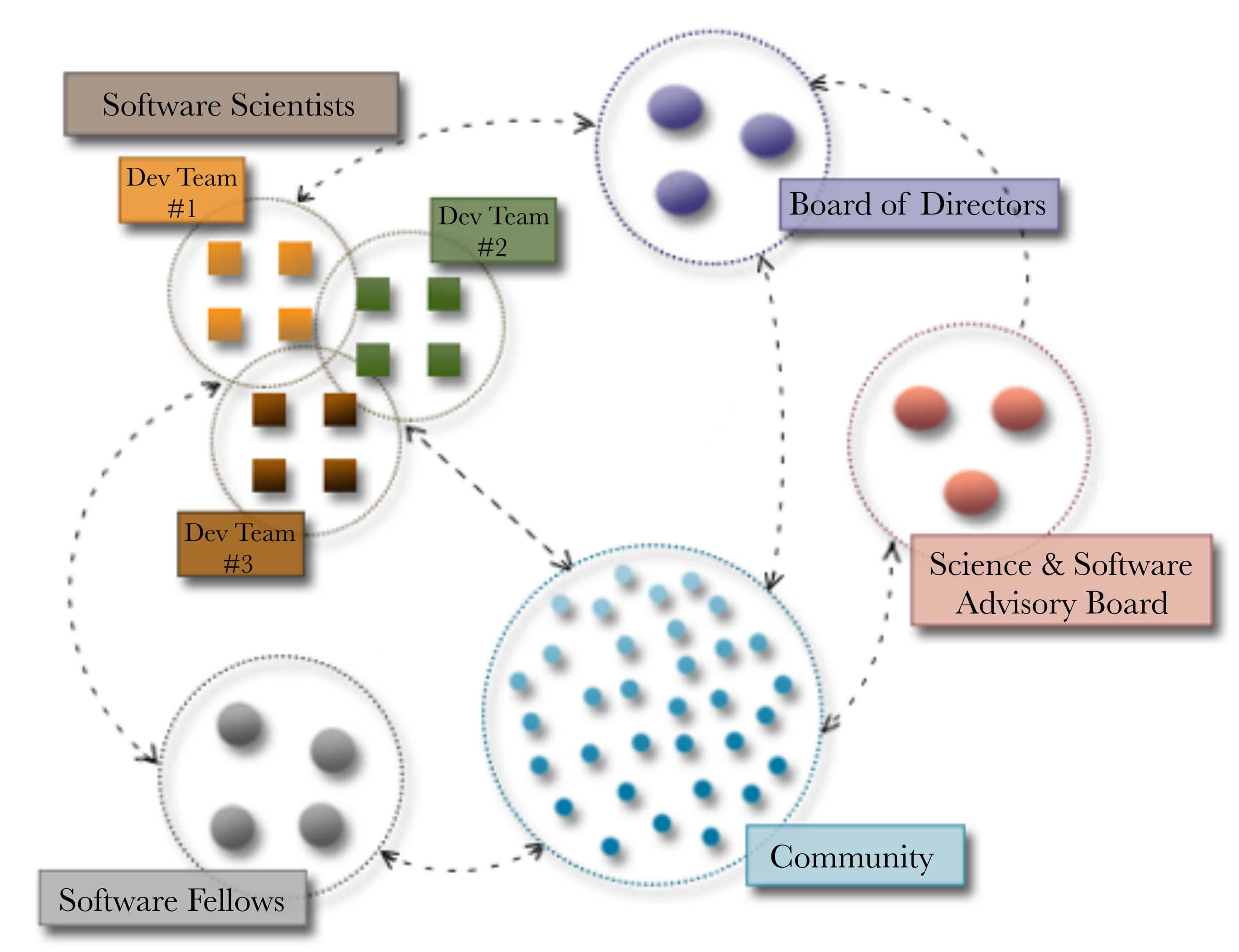
- A cohort of ~16 Fellows supported simultaneously – graduate students and postdocs selected by the Science and Software Advisory Board from research groups across the U.S.
- Fellows will work directly with both the Software Scientists and the MolSSI Directors, thus providing a conduit between the Institute and the CMS community itself.
- Fellows will work on their own projects, as well as contribute to the MolSSI development efforts, and they will engage in outreach and education activities under the Institute guidance.
- Funding for MolSSI Software Fellows will follow a flexible, two-phase structure, providing up to two years of support.

Approximately 25% of the Institute's budget will directly support the MolSSI Software Fellows.

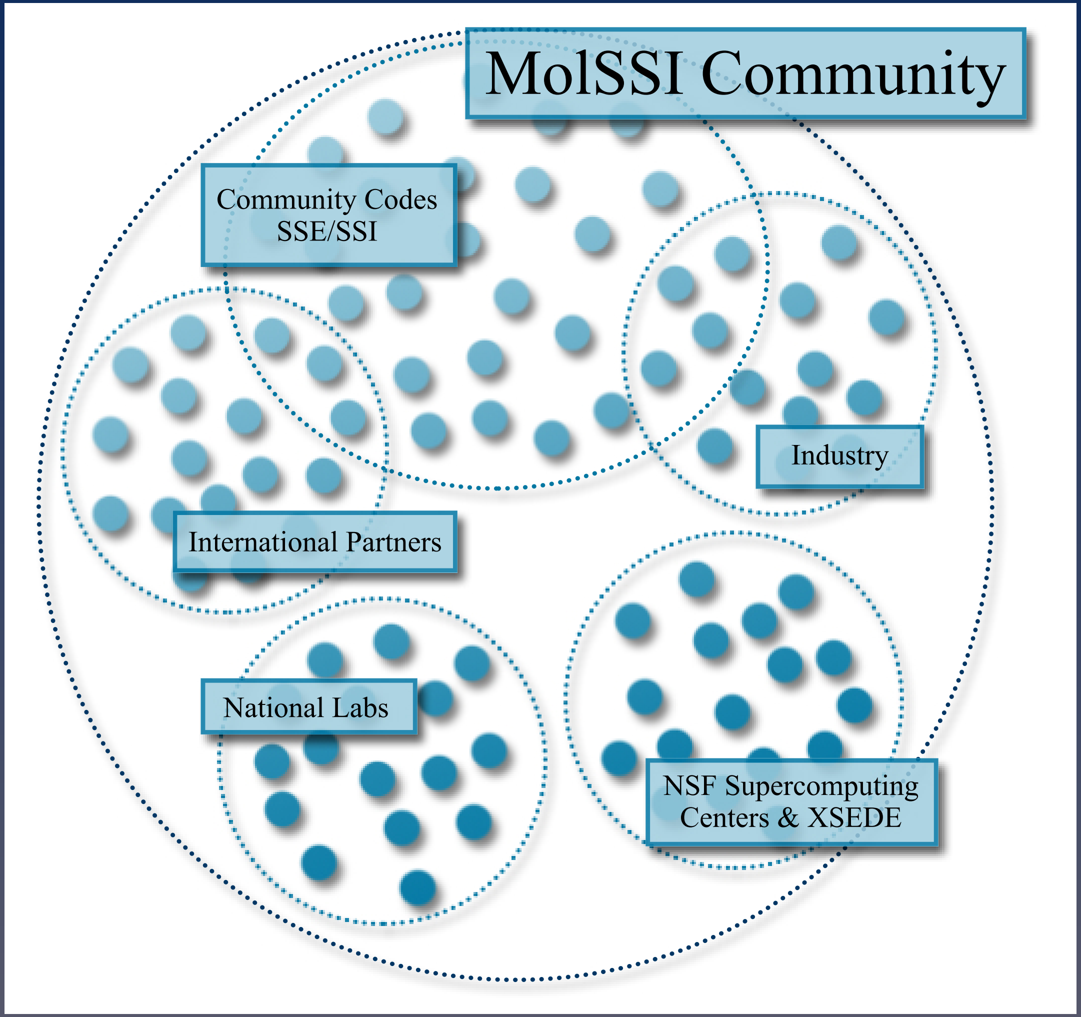
The Molecular Sciences Software Institute (MolSSI)



The Molecular Sciences Software Institute (MolSSI)



The MolSSI Community



Engaging Community Codes & SSE/SSI

- A principal representative from each of the following community codes has committed to collaboration between their development team and the MolSSI Software Scientists and leadership:
  - Gaussian
  - GAMESS
  - Molpro
  - Q-Chem
  - PSI4
  - ACESIII
  - CFOUR
  - Molcas
  - Orca
  - SISSA
  - CHARMM
  - Amber
  - BOSS
  - Gromacs
  - OpenMM
  - LAMMPS
  - Plumed
  - Turbomole
  - NWChem
  - ONETEP
  - NAMD
  - Dalton
  - Columbus
  - Dirac
  - DL_POLY
  - Tiger-CI
  - Schrödinger
  - Quantum ESPRESSO
  - PARSEC
  - APBS
- MolSSI will coordinate with all relevant SSE/SSI projects to bring their software products to the community.

Engaging the International Community

- MolSSI’s Board of Directors and SSAB have established numerous community code partners worldwide.
- EPSRC: ARCHER eCSE
- EU Computational Materials Centers
- EU Center of Excellence on Biomolecular Simulation (BioExcel)
- Our S2I2 Conceptualization workshops prompted the UK’s EPSRC to report on how the British CMS community could interface to MolSSI.
- The SSAB will maintain an international representative.

NOMAD

NOVEL MATERIALS DISCOVERY

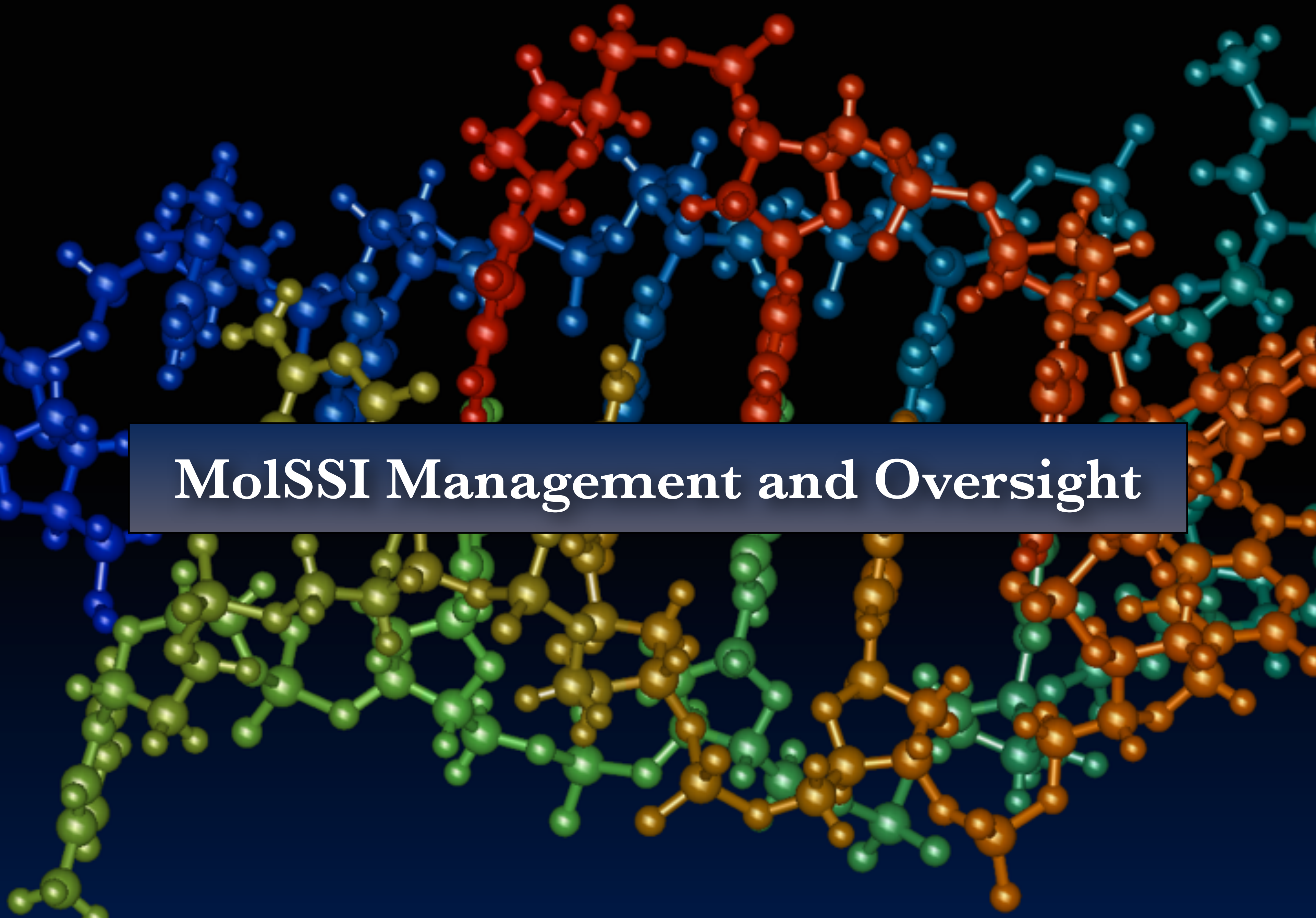
archer

Horizon 2020

Education and Training

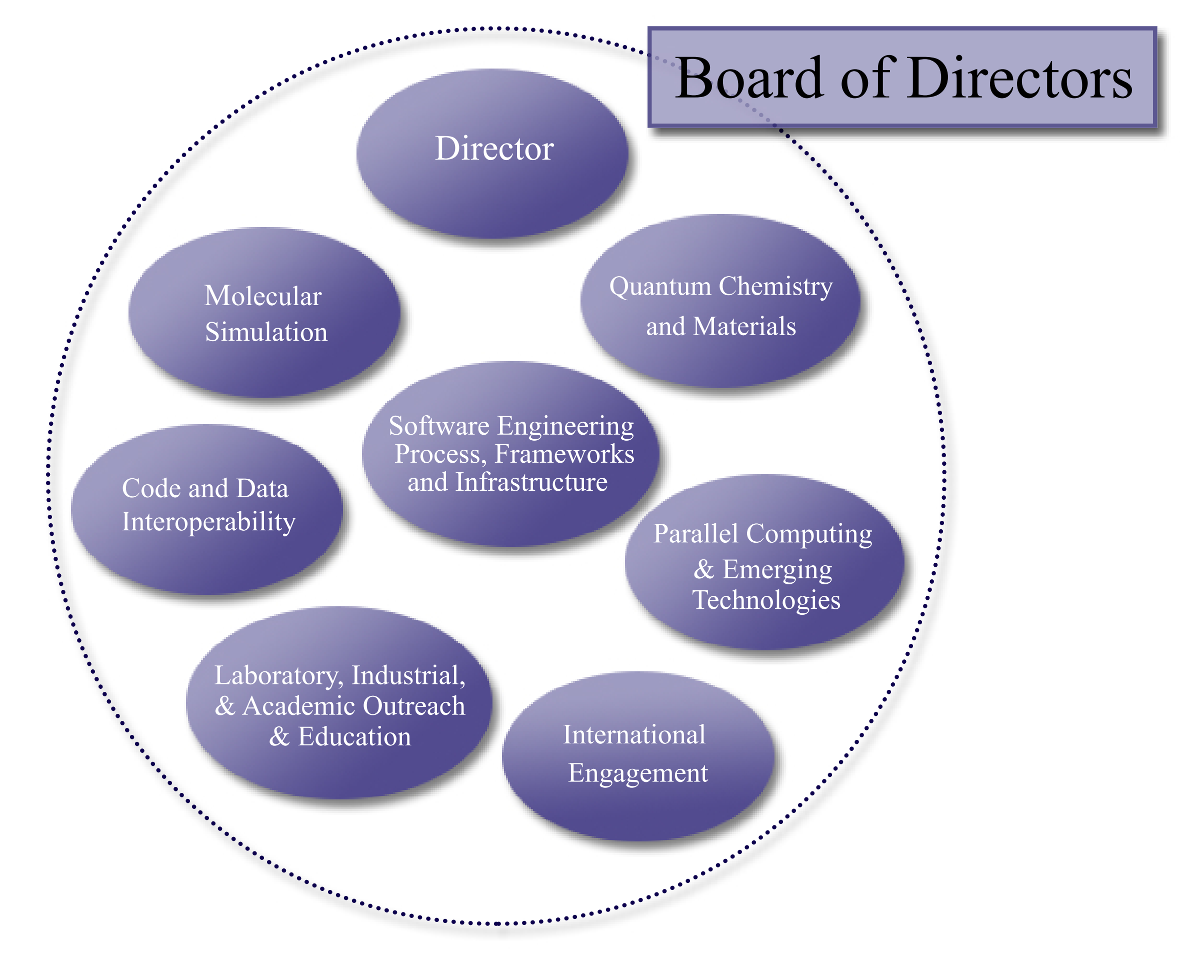
Professional Master's in Molecular Simulation and Software Engineering (MSSE)

- A two-year, self-supporting, part-time Master's program comprised of 26 units including:
  - Computational chemistry
  - Materials science
  - CS294: Software Engineering for Scientific Computing (P. Colella)
  - CS267: Applications of Parallel Computers (J. Demmel)
  - Leadership, management, and communication (Fung Institute):
    - E271/272: Engineering Leadership I & II
    - E273: Ethics and Capstone Project
- MolSSI will engage with industry and government labs for capstone projects, help with outreach for admissions, and provide a career fair at the Virginia Tech Arlington Center that will include remote access.



MolSSI Management and Oversight

The MolSSI Board of Directors



Board of Directors

Director

Molecular Simulation

Quantum Chemistry and Materials

Software Engineering Process, Frameworks and Infrastructure

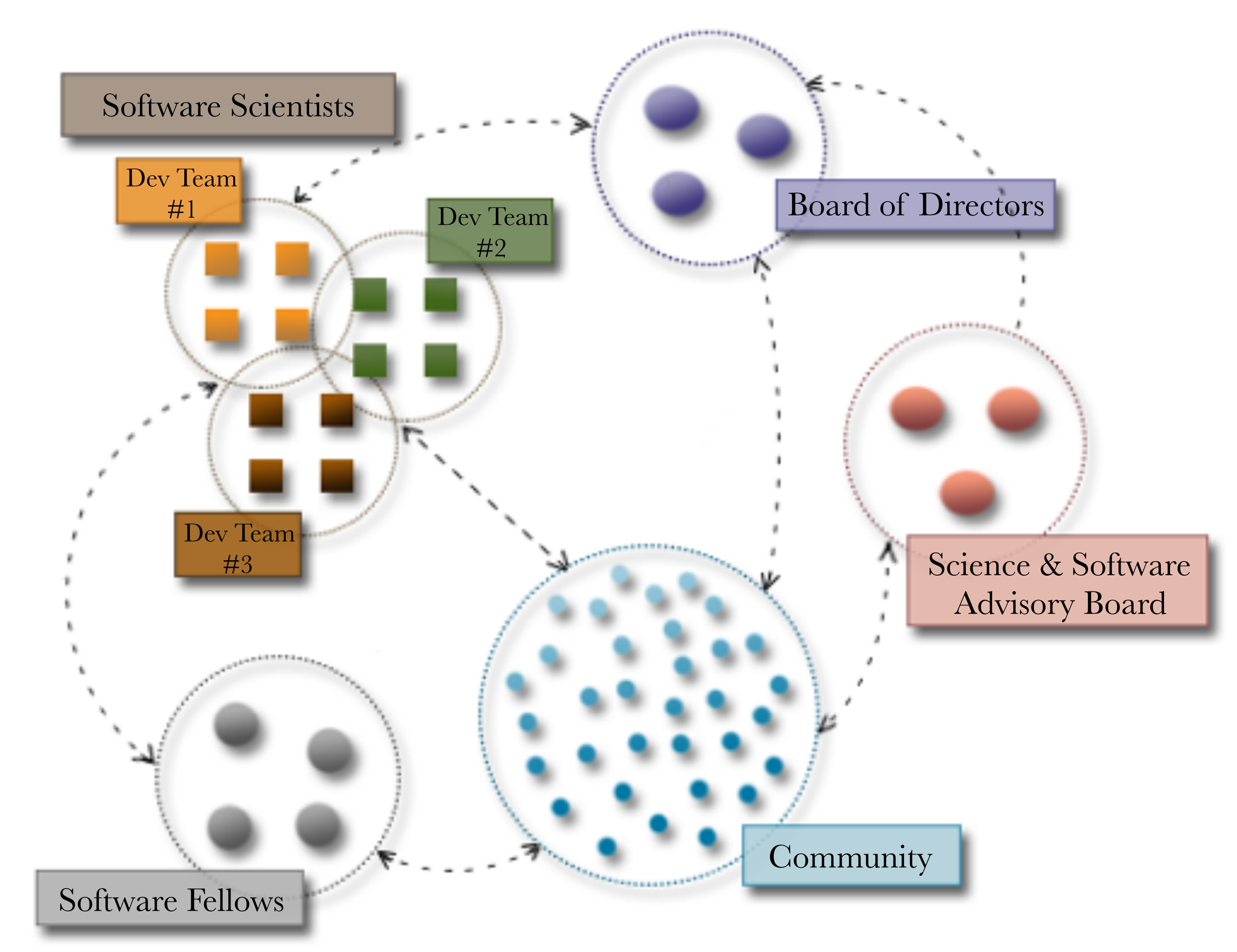
Code and Data Interoperability

Parallel Computing & Emerging Technologies

Laboratory, Industrial, & Academic Outreach & Education

International Engagement

The Molecular Sciences Software Institute (MolSSI)



Potential Initial Software Frameworks and Use Cases

- Interoperability frameworks between QM/MM codes:
    - Code interoperability – set APIs that allow algorithms to be easily migrated from code to code;
    - Data interoperability – data structures and mathematical definitions of key quantities for easier sharing;
- Parallel task managers and DSLs targeted toward many-body methods;
- Load-balancing infrastructure for advanced sampling methods;
- Use cases outlining interaction schemes between multiple QM and MD codes;
- Use cases derived from current and future SSE/SSI projects;
- DSLs that hide multi-model and multi-code tasks from the user enabling new science.

Acknowledgements

- Daniel Crawford, Cecilia Clementi, Robert Harrison, Teresa Head-Gordon, , Anna Krylov, Vijay Pande, Theresa Windus
- The dozens of members of the CMS community who helped to develop the vision for the Institute over the last five years
- NSF ACI-1547580

Watch molssi.org for the latest information!