Codemeta: A Rosetta Stone for Software Metadata

EARly-concept Grants for Exploratory Research (EAGER)

Carl Boettiger, UC Berkeley

Abstract

Research relies heavily on scientific software, and a large and growing fraction of researchers are engaged in developing software as part of their own research (Hannay et al 2009). Despite this, infrastructure to support the preservation, discovery, reuse, and attribution of software lags substantially behind that of other research products such as journal articles and research data. This lag is driven not so much by a lack of technology as it is by a lack of unity: existing mechanisms to archive, document, index, share, discover, and cite software contributions are heterogeneous among both disciplines and archives and rarely meet best-practices (Howison 2015). Fortunately, a rapidly growing movement to improve preservation, discovery, reuse and attribution of academic software is now underway: a recent NIH report, conferences and working groups of Force11, WSSSPE & Software Sustainability Institute, and the rising adoption of repositories like GitHub, Zenodo, figshare & DataONE by academic software developers. Codemeta is a distributed open source project to improve how these resources can talk to each other.

Example JSON-LD

Example JSON-LD

A Metadata CrossWalk

18 contributors met in Portland, OR on April 15-17, 2016 to agree on a draft crosswalk between major software metadata schema.

Schemas in Crosswalk

DataCite, OntoSoft, Zenodo, GitHub, Figshare, Software Ontology, Software Discovery Index, Dublin Core, R Package Description, Debian Package Metadata, Python Distutils (PyPI), Trove Software Map, Perl Module Description (CPAN::Meta), JavaScript package description (npm), Java (Maven), Octave, Ruby Gem, ASCL, Schema.org

A JSON-LD Context for Software Metadata

doi: 10.5063/schema/codemeta-1.0

- A linked-data representation of codemeta concepts
- Basic types from Schema.org, also Dublin Core and XSD
- 26 new terms introduced in codemeta namespace

Project Deliverables

Project Website: codemeta.github.io

JSON-LD Context file: github.com/codemeta/codemeta/codemeta.jsonld

Crosswalk Table: github.com/codemeta/codemeta/crosswalk.csv

codemeta:json examples: github.com/codemeta/codemeta/examples

Coming soon

Utilities for automatically generating codemeta.json for common software package types (e.g. R packages)

Zenodo parsing of codemeta.json into Zenodo metadata and on to DataCite records.

Partner organizations extend metadata representation

Organizations

Workshop Participants

- Carl Boettiger, UC Berkeley
- Matt Jones, NCEAS
- Arfon Smith, GitHub
- Yolanda Gil, USC ISI
- Martin Fenner, DataCite
- Krzysztof Nowak, Zenodo
- Mark Hahnel, figshare
- Abby Mayes, Mozilla Science Lab
- Luke Coy, RIT & MSL
- Kyle Niemeyer, Oregon State
- Alice Allen, ASCL
- Mercé Crosas, Harvard, IQSS
- Ashley Sands, UCLA
- Neil Chuong Hong SSI
- Peter Slaughter, NCEAS
- Patricia Cruse, DataCite
- Dan Katz, NCSA
- Carole Goble, University of Manchester

See more contributors at github.com/codemeta/codemeta/graphs/contributors and github.com/codemeta/codemeta/network/members
Computational Infrastructure for Geodynamics: Accelerating Discovery in the Solid Earth through Sustainable Scientific Software

Louise H. Kellogg, University of California, Davis
Algorithm:
1. Project high-D → low-D
2. LSH low-D vectors
3. Count-min sketch
4. Top k counters are centroids

<table>
<thead>
<tr>
<th>HAR Data</th>
<th>ARI</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kmeans</td>
<td>0.461</td>
<td>24.746</td>
</tr>
<tr>
<td>RPHash</td>
<td>0.363</td>
<td>0.4838</td>
</tr>
<tr>
<td>RPHash-Dis</td>
<td>0.48</td>
<td>8.116</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dim.</th>
<th>KMeans</th>
<th>RPHash</th>
<th>RPHash-Dis</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>3.9656</td>
<td>0.4890</td>
<td>7.54</td>
</tr>
<tr>
<td>500</td>
<td>32.92</td>
<td>0.5594</td>
<td>7.87</td>
</tr>
<tr>
<td>1000</td>
<td>142.2341</td>
<td>0.7206</td>
<td>8.51</td>
</tr>
<tr>
<td>1500</td>
<td>237.0007</td>
<td>0.8233</td>
<td>9.48</td>
</tr>
<tr>
<td>2000</td>
<td>366.0743</td>
<td>0.8796</td>
<td>10.73</td>
</tr>
<tr>
<td>2500</td>
<td>431.5876</td>
<td>0.9997</td>
<td>11.43</td>
</tr>
<tr>
<td>3000</td>
<td>542.0223</td>
<td>1.1122</td>
<td>13.39</td>
</tr>
<tr>
<td>3500</td>
<td>631.8423</td>
<td>1.2421</td>
<td>12.97</td>
</tr>
<tr>
<td>4000</td>
<td>741.915</td>
<td>1.3157</td>
<td>13.78</td>
</tr>
<tr>
<td>4500</td>
<td>811.3911</td>
<td>1.3986</td>
<td>14.40</td>
</tr>
<tr>
<td>5000</td>
<td>909.223</td>
<td>1.5095</td>
<td>15.14</td>
</tr>
<tr>
<td>5500</td>
<td>975.2703</td>
<td>1.5977</td>
<td>15.98</td>
</tr>
<tr>
<td>6000</td>
<td>1076.882</td>
<td>1.6805</td>
<td>17.24</td>
</tr>
<tr>
<td>6500</td>
<td>1187.6062</td>
<td>1.7933</td>
<td>18.02</td>
</tr>
</tbody>
</table>
Automated Statistical Mechanics for the First-Principles Prediction of Finite Temperature Properties of Crystals

Jonathon Bechtel, Julija Vinckeviciute, Sanjeev Kolli, S. N. Harsha Gunda, Anton Van der Ven
Materials Department, University of California Santa Barbara

CASM
A Clusters Approach to Statistical Mechanics

Electronic Structure Methods
\[ -\nabla_i^2 + V_{\text{ion}}(r_i) + \int \frac{\rho(\mathbf{r})}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + V_{\text{xc}}(\rho(\mathbf{r})) \phi_i(\mathbf{r}_i) = \varepsilon \phi_i(\mathbf{r}_i) \]

Effective Hamiltonians

(kinetic) Monte Carlo

Thermodynamics

Kinetics

Ti-C

Ti-N

Ti-O
SI2-SSE: GraphPack: Unified Graph Processing with Parallel Boost Graph Library, GraphBLAS, and High-Level Generic Algorithm Interfaces

Andrew Lumsdaine
ELSI - Infrastructure for Scalable Electronic Structure Theory

Victor Wen-Zhe Yu, Fabiano Corsetti, Alberto García, Stefano de Gironcoli, William Paul Huhn, Mathias Jacquelin, Weile Jia, Murat Keçeli, Raul Laasner, Lin Lin, Jianfeng Lu, Yingzhou Li, Álvaro Vázquez-Mayagoita, Chao Yang, Haizhao Yang, and Volker Blum

Yu et al., Comp. Phys. Commun. 222, 267 (2018)
http://elsi-interchange.org

Part of upcoming CECAM “ESL (Electronic Structure Library) bundle”
Parallel Parsing of Application Binaries

A New Dynamic Data Race Detector for OpenMP

- Instrument data accesses in binaries
- Track task orderings, mutual exclusion, and data environments during execution
- Assess races involving each access by reasoning about locks and orderings

Performance Analysis of GPU Accelerated Applications
Background on ASSISTments

ASSISTments is a free, university-based platform. Each day, teachers assign problems to thousands of students (currently 50,000 students) through problem sets aligned to the Common Core State Standards. Randomized controlled experiments are often embedded in these problem sets to evaluate the efficacy of learning interventions. Heffernan has been funded by the NSF to conduct certain types of research using ASSISTments (i.e., spacing studies). Seventeen peer-reviewed publications have resulted from controlled experiments conducted within ASSISTments.

Purpose of the SI2 Grant

This grant proposed the evolution of ASSISTments into a shared scientific instrument that researchers could use to experiment with learning at scale. Software infrastructure modifications were necessary to aid in this transformation. The result is the ASSISTments TestBed (www.assistmentsTestbed.org). The community of potential users include educational psychologists and mathematics education researchers. Developing relationships with schools has traditionally been costly for researchers. The ASSISTments TestBed reduces these costs by bridging relationships with teachers and researchers to conduct noninvasive classroom experiments that improve education.

WPIs Contribution

- Created the infrastructure to allow researchers to design and implement their own RCTs.
- Created a way for teachers to access materials with embedded RCTs but not be distracted by them.
- Ran multiple trainings (AERA, as well as a webinar) to recruit pilot researchers.
- Created a workflow for idea submission. WPI mentors researchers toward study designs compliant with our IRB via ‘normal instructional strategies.’
- Created the Assessment of Learning Infrastructure (ALI) to provide easy access to study data. This tool eases the data processing and analysis required of researchers as raw files can be overwhelming.

ASSISTments TestBed Research Progression

Universal Reporting

Raw Data Files

Raw data files contain the logged information for each student that has participated in your study. We provide this data in a variety of formats, as shown below, to assist in your analytic efforts.

- Student Corrective Dataset
- Action Level Dataset
- Problem Level Dataset
- Student Level + Problem Level Dataset

For a glossary and dataset tutorials, please visit our Glossary Page.

Knowledge Gained

Good Experiences...

- Recruiting was easy. This service is in high demand and we filled our yearly researcher quota in just a single day.
- Researchers love our IRB terms. By separating recruitment from actual data collection it was much easier to get external IRB approval.

Difficulties...

- We spent time negotiating with researchers about what constitutes ‘normal instructional practice’ (researchers want extensive pre/post tests).
- We have pivoted our focus to assignments where completion goals are well known to users, rather than letting researchers design new content that we then had to try to entice teachers to use.
- We have noticed a lack of clear, conceptual assessment items aligned to the Common Core State Standards in our content.

Improving Science by Promoting Replication...

- Pre-registration of studies and hypotheses
- Open data and open materials
- Reductions to the ‘File Drawer’ problem

Success To Date

- Participation from a dozen researchers, representing: Boston College, Freiburg University, Harvard University, Indiana University, Northwestern University, Southern Methodist University, Texas A&M, University of Colorado - Colorado Springs, University of California–Berkeley, University of Maine, University of Wisconsin, and Vanderbilt.
- Participation from an educational company hoping to evaluate their product.
- Publications:
  - The first manuscript resulting from this pool of research is currently in press.
  - Three researchers have published at international conferences.
  - Two manuscripts are in press to promote this ‘research evolution’ in similar learning platforms.

About the Author

Neil T. Heffernan (neilh@wpi.edu) is a Professor of Computer Science, the Director of the Learning Science & Technologies Program, and the creator of ASSISTments.

SI2 - Adding Research Accounts to the ASSISTments Platform: Helping Researchers Do Randomized Controlled Studies with Thousands of Students

Worcester Polytechnic Institute
Artifact Execution Curation for Repeatability in Artifact Evaluation

Copy artifacts from the ephemeral internet

Use the artifacts in workflows

Execute the artifacts in repeatable environments

The ephemeral Internet
CDS&E: Numerical Investigation of Two-Particle Response Functions of Correlated Materials

Emanuel Gull
CIF21 DIBBs: PD: Building High-Availability Data Capabilities in Data-Centric Cyberinfrastructure

Haiying Shen, Associate Professor
Computer Science, University of Virginia

• HPC and HDFS storage architectures are not scalable enough to adapt to exascale systems

Motivation
Achieving exascale data-centric cyberinfrastructure capabilities

Solution
Highly scalable and efficient data storage system (i.e., file system)

Goal
Building Scalable High-Availability Data Capabilities in Data-Centric Cyberinfrastructure

Proposed project
Distributed data indexing and load balancing
Energy-efficient and popularity-adaptive data replication
Adaptive low-overhead data replication
Distributed low-overhead consistency maintenance

BROADER IMPACTS
• Curriculum development activities
• Student recruiting and mentoring
• Impacts on the research community

INTELLIGENTIAL MERIT
• Distributed data indexing and load balancing
• Energy-efficient and popularity-adaptive data replication
• Adaptive low-overhead data replication
• Distributed low-overhead consistency maintenance

U.S. NSF-OAC-1724845
CitSci.org: Advancing and Mobilizing Citizen Science Data through Integrated Sustainable Cyber-Infrastructure

Greg Newman*, Stacy Lynn*, Melinda Laituri*, Russell Scarpino†, Louis Leibenberg†, Sarah Newman*, Justin Steventon†

We power citizen science projects all over the world

4,958 people
572 projects
796,158 data points
66,775 locations
2,099 protocols

120+ customer discovery interviews
know your audience
build relevant software

design & analysis tools

goals

- Broaden inclusivity of citizen science (CS)
- Improve CS software
- Mobilize CS data for meta-analyses
- Elevate the value and rigor of CS software and data

future directions

- Integrate real time precipitation data
- Integrate field based with online CS
- Offer design & implementation support services

know your audience

partner for sustainability

integrations & interoperability

why?

- science use cases & needs
- research
- diversity
- sharing
- quality

Colorado State University
SI2-SSE: Reusable Image Analytics Pipelines

Dino Bektasevic (dinob@uw.edu), Magda Balazinska, Alvin Cheung, Andrew Connolly (PI), Mario Juric

University of Washington

Scaling the image processing and analytics pipelines for the Large Synoptic Survey Telescope (LSST) to run efficiently across a broad range of architectures and within the cloud
SI2-SSE: Development of Cassandra, A General, Efficient and Parallel Monte Carlo Multiscale Modeling Software Platform for Materials Research

Jindal K. Shah
**GOAL**  Offer support for **software-defined events (SDE)** to extend PAPI's role as a standardizing layer for monitoring performance counters.

**VISION**  Enable HPC software layers to expose SDEs that performance analysts can use to form a **complete** picture of the entire application performance.

**BENEFIT**  HPC application scientists will be able to better understand the interaction of the different application layers, and the interaction with external libraries and runtimes.

All new PAPI SDE API functions are available in C and FORTRAN, e.g.:

```c
void *papi_sde_init(char *lib_name);
void papi_sde_register_counter( ..., char *event_name, ..., void *counter);
void papi_sde_register_fp_counter( ..., char *event_name, ..., func_ptr_t fp_counter);
void papi_sde_describe_counter( ..., char *event_name, char *event_description);
```
1) **Self-aware** Information Extraction (SELFIE) workflows

- IE alternatives are organized in cost-incremental order. They are able to **evaluate** each generated value and **decide** if it is correct or should be extracted by another IE process.
- Tasks **actively improve** from humans (Active Learning - Humans in The Loop)

2) **Smart use of the crowdsourcing interfaces.**

We have shown that hybrid approaches can yield **human-like quality** results and **machine-like** execution speed.

---

**Problem:** Scientific data digitization is a slow process, commonly performed using humans (crowdsourcing).

**HuMaIN’s Objective:** Efficient digitization of scientific data.

**HuMaIN:** Information services framework for the integration of human & machine Information Extraction (IE) tasks.

**Use case:** Digitization of biocollections’ labels.
A sustainable Software Environment for Integrated Seismic Modeling

Pt: John E. Vidale (USC), Award ID: ACI-1450451
Co-PIs: YiFeng Cui (SCEC), Kim B. Olsen (SCEC), Ricardo Taboada (University of Memphis)

SEISM

SCDC and SCEC Project Objectives

The Southern California Earthquake Center (SCDC) coordinates fundamental and applied research on earthquake processes using Southern California as a natural laboratory. This research program is designed to support research and education in seismology, structural geology, earthquake, and computational science. The SCEC community advances earthquake system science through three basic activities: (i) gathering information from seismological and geodetic systems, geophysical field observations, and laboratory experiments; (ii) synthesizing knowledge of earthquake phenomena through physics-based modeling, including system-level hazard modeling, analysis and communication of understanding to relevant communities; and (iii) preparing and executing new large-scale simulation experiments.

The SCEC/SCDC project integrates scientific software elements (SSEs) into a software ecosystem for physics-based seismic hazard analysis (SHA) capable of using current and future extreme-scale computing systems.

1. Incorporate into community models the physics necessary to extend deterministic simulations to seismic frequencies of engineering interest (>1 Hz).
2. Incorporate into community models time elapse before they can be confidently used for designing structural systems.
3. Provide comprehensive seismograms as input to soil and structure analyses.

Challenge: Stability is inherently difficult to obtain in overlap between seismic wave speeds, such as basin models.

Open-source Software

Cybershake: A simulation-based platform for probabilistic seismic hazard analysis (PSHA). CyberShake simulates ground motions by combining finite-fault rupture descriptions of earthquakes and near-source propagation in 1-D structural models of the Earth. CyberShake ground motions account for complex effects such as the coupling of directivity and basin response that cannot easily be captured with empirical modeling of ground motions. For the latest CyberShake implementation (v3.5) and cannot (CA) an updated Blue Waters and Titan to compare hazard results from over 400,000 scenarios at 64 locations, using alternative velocity models, producing 285 million reproducible waveforms, and computing over 43 trillion simulation instances (responsive spectral acceleration and duration matrices)

• Averaged 300 million - 1% 1% (vca, vssa) (CA/fix) 1% (2.5%)
• 800,000,000 cores (CA/fix)
• FORTRAN, C++, Python, MATLAB, and CMake.
• Transferred 1.6 TB of interactivity data between the two systems
• Workload managed using Pegasus (7,000 jobs).
• Transfered CyberShake to the new region requiring the development of software to conduct simulations.

Cybershake Maps: Toward a CA-wide model

The map below shows the 1% in 50 probability of exceedance of 10 + do, pseudo-spectral acceleration (PSA) for the combined region created following Study 117.7 (shock box for local, mag-11 central CA, stress drop for future comparison of NGA/W).

Ground Motion Simulations

Envisioning "Virtual Topography" Using Hierarchies

Topography affects ground motions differently depending on the topographic features. This effect is frequency-dependent.

• We partition the domain using a non-conforming mesh technique.
• We develop an efficient virtual topography generation algorithm.

We solve the weak form of the elastodynamic equation in space by an octree-based finite-element method and a second-order central difference step-by-step marching forward scheme in time.

The HPCC environment (HPC) enables running different suites of scientific modules that models come, path and frequency (PSA). We have validated the PEER NGA-West2 project and CyberShake. The new PEER tool is to be released on May 29, 2016.

High-P Performance

Open-source Software

HP-ODC on Nvidia GPU

• First 6-to-10 simulation of M7.7 on the southern San Andreas Fault conducted using 4,200 Blue Waters GPUs.
• 40% of parallel efficiency achieved for both linear and nonlinear versions of HP-ODC on GPGPUs.
• Accelerated time-to-solution from original Deconv to 12.65 on Nvidia Titan X (64 cores).
• Blue Waters (MDS-1) project ported to CyberShake. The new MDS tool is to be released on May 29, 2016.

UMd Committee and Code Application

Software Scaling

Cybershake in a distributed fashion on different suites of scientific modules that models come, path, and frequency (PSA). We have validated the PEER NGA-West2 project and CyberShake. The new PEER tool is to be released on May 29, 2016.

Using simulation-based and traditional empirical-based approaches, the SCEC Committee for Utilization of Ground Motion Simulations (USUMS) is working to develop improved response spectral acceleration maps for the Los Angeles region for possible future inclusion in the NIST and SCEC - California Building Code. The USUMS is developing recommendations for the specification of the inelastic Elemental-Constrained Earthquake (ECEE) design bases for spectrums larger than 0.29sec per iteration (6.5speedup). The challenge has developed a criterion based on the combination of values from ground motion prediction equations (GMPEs) from the PEER NGA-West2 project and CyberShake. The new MDS tool is to be released on May 29, 2016.

The HPCC environment (HPC) enables running different suites of scientific modules that models come, path, and frequency (PSA). We have validated the PEER NGA-West2 project and CyberShake. The new PEER tool is to be released on May 29, 2016.

Open-source Software

CyberShake: A simulation-based platform for probabilistic seismic hazard analysis (PSHA). CyberShake simulates ground motions by combining finite-fault rupture descriptions of earthquakes and near-source propagation in 1-D structural models of the Earth. CyberShake ground motions account for complex effects such as the coupling of directivity and basin response that cannot easily be captured with empirical modeling of ground motions. For the latest CyberShake implementation (v3.5) and cannot (CA) an updated Blue Waters and Titan to compare hazard results from over 400,000 scenarios at 64 locations, using alternative velocity models, producing 285 million reproducible waveforms, and computing over 43 trillion simulation instances (responsive spectral acceleration and duration matrices)

• Averaged 300 million - 1% 1% (vca, vssa) (CA/fix) 1% (2.5%)
• 800,000,000 cores (CA/fix)
• FORTRAN, C++, Python, MATLAB, and CMake.
• Transferred 1.6 TB of interactivity data between the two systems
• Workload managed using Pegasus (7,000 jobs).
• Transfered CyberShake to the new region requiring the development of software to conduct simulations.

Cybershake Maps: Toward a CA-wide model

The map below shows the 1% in 50 probability of exceedance of 10 + do, pseudo-spectral acceleration (PSA) for the combined region created following Study 117.7 (shock box for local, mag-11 central CA, stress drop for future comparison of NGA/W).

Ground Motion Simulations

Envisioning "Virtual Topography" Using Hierarchies

Topography affects ground motions differently depending on the topographic features. This effect is frequency-dependent.

• We partition the domain using a non-conforming mesh technique.
• We develop an efficient virtual topography generation algorithm.

We solve the weak form of the elastodynamic equation in space by an octree-based finite-element method and a second-order central difference step-by-step marching forward scheme in time.

The HPCC environment (HPC) enables running different suites of scientific modules that models come, path and frequency (PSA). We have validated the PEER NGA-West2 project and CyberShake. The new PEER tool is to be released on May 29, 2016.

High-P Performance

Open-source Software

HP-ODC on Nvidia GPU

• First 6-to-10 simulation of M7.7 on the southern San Andreas Fault conducted using 4,200 Blue Waters GPUs.
• 40% of parallel efficiency achieved for both linear and nonlinear versions of HP-ODC on GPGPUs.
• Accelerated time-to-solution from original Deconv to 12.65 on Nvidia Titan X (64 cores).
• Blue Waters (MDS-1) project ported to CyberShake. The new MDS tool is to be released on May 29, 2016.

Software Scaling

CyberShake in a distributed fashion on different suites of scientific modules that models come, path, and frequency (PSA). We have validated the PEER NGA-West2 project and CyberShake. The new PEER tool is to be released on May 29, 2016.

Using simulation-based and traditional empirical-based approaches, the SCEC Committee for Utilization of Ground Motion Simulations (USUMS) is working to develop improved response spectral acceleration maps for the Los Angeles region for possible future inclusion in the NIST and SCEC - California Building Code. The USUMS is developing recommendations for the specification of the inelastic Elemental-Constrained Earthquake (ECEE) design bases for spectrums larger than 0.29sec per iteration (6.5speedup). The challenge has developed a criterion based on the combination of values from ground motion prediction equations (GMPEs) from the PEER NGA-West2 project and CyberShake. The new MDS tool is to be released on May 29, 2016.

The HPCC environment (HPC) enables running different suites of scientific modules that models come, path, and frequency (PSA). We have validated the PEER NGA-West2 project and CyberShake. The new PEER tool is to be released on May 29, 2016.
A parallel computing framework for large-scale real-space and real-time TDDFT excited-states calculations
E. Polizzi, UMass Amherst
Project Motivation

Image registration and segmentation are vital enabling technologies for addressing many complex, data-driven problems. Examples include individualized medical treatment where disease progression is monitored by analyzing MRI, CT, or ultrasound images over time; identifying anatomical structures in medical images; recognizing objects and people in video footage; and extracting imageable biometrics such as fingerprints, faces, and the iris. Images and videos can now be easily acquired at a rate that far surpasses our capacity to perform advanced image analysis. For this reason, advanced registration and segmentation algorithms are not routinely used for many large-scale and time-sensitive applications because they require more processing time than is available.

Real-time Image Registration

Cine-mode MR is an ideal imaging method for real-time therapeutic control. Our deformable image registration method can be computed in less than 200 ms, to enable real-time tracking of multiple soft-tissue targets. The intended application is automatic optimization of radiotherapy in the thorax and abdomen, to compensate for respiratory motion. (Submitted, RSNA 2018)

Multi-grid Image Registration

B-spline deformable image registration is well established for mapping 3D medical imaging volumes. B-spline coefficients can be optimized simultaneously at multiple resolutions using a multi-grid approach. Algorithm performance is boosted by detecting grid sparsity and removing unnecessary parameters from the coefficient grid. (Accepted, CVPR 2017)

Work in Progress

Automatic image segmentation methods use distance maps during training and classification. Drexel PhD students Shihao Song and Michael Spanier are developing high-performance methods to accelerate this important algorithm.

Contact and Visit

NSF Award #1642380

http://www.libkaze.com
http://www.platimatch.org

James Shackleford
shack@drexel.edu

Naga Kandasamy
kandasamy@drexel.edu

Greg Sharp
gcsharp@mgh.harvard.edu

Cine-mode MR is an ideal imaging method for real-time therapeutic control. Our deformable image registration method can be computed in less than 200 ms, to enable real-time tracking of multiple soft-tissue targets. The intended application is automatic optimization of radiotherapy in the thorax and abdomen, to compensate for respiratory motion. (Submitted, RSNA 2018)
SI2-SSE: Abaco - Flexible, scalable, and usable
Functions-as-a-service via the Actor Model

PRIMARY CAPABILITIES

- **Reactors** provide functionality for event-driven programming
- **Asynchronous Executors** enable executing functions on the Abaco cluster from directly within a running application.
- **Data adapters** enable users to create high-quality API access to data from disparate external sources

EARLY ADOPTERS

- ETL pipelines in the SD2E platform
- Scheduling containers on an elastic cloud in the IPT on the Web gateway
- Automatically creating Singularity images from new Docker images in the Biocontainers project.
- Scaling out pleasantly parallel Opensees function calls within an engineering Jupyter notebook
The free energy is the single-most important quantity describing biomolecular systems at equilibrium.

Biomolecular free energy surfaces are rugged and complex requiring very long time scales to explore.

To address issue, we developed the Adaptively Biased Molecular Dynamics (ABMD) method with belongs to the general category of umbrella sampling with methods with time-dependent potential [Babin et al, JCP 2008]

- ABMD calculates Landau free energies as a function of suitably chosen collective variables
- Implemented with multiple walker and replica exchange enhancements
- Implemented along with Steered Molecular Dynamics (SMD) in AMBER software suite

Our SEE program aims to develop software tools around this program

Current extensions:
1. Port ABMD suite from SANDER to PMEMD in AMBER v.16 +
2. GPU compatible
3. Introduce interacting multiple walker algorithms for enhanced sampling
4. Well-tempered ABMD
5. Driven-ABMD with combines SMD and ABMD
6. Introduce swarms-of-trajectories string method (STSM) for calculating minimum free energy path (MFEP)

Our group has applied methodology to variety of biomolecular systems: small peptides, sugar puckering, polyproline systems, guest-host systems, polyglutamine systems, DNA/RNA

E.g., investigation of hairpin loops and stems for CAG and GAC DNA/RNA systems as implicated in trinucleotide repeat expansion diseases

- RNA free energy landscapes as a function of two collective variables – see Pan et al, Biophysical J., 2017 for details
What is OpenAtom

NSF SI2-SSI: Scalable, Extensible, and Open Framework for Ground and Excited State Properties of Complex Systems

NSF ACI 1339804 & 1339715

- Quantum simulation of materials and molecules (DFT)
- **OpenAtom** software package: molecular dynamics now, excited electrons in progress (GW)
- Plane waves and pseudopotentials
- charm++ parallel infrastructure
Mining Seismic Wavefields (EAR-1551462)

PI: Gregory C. Beroza (Stanford); Co-PIs and Contributors: Clara Yoon, Karianne Bergen, Kexin Rong, Hashem Elezabi, Peter Bailis, Philip Levis (Stanford), Yehuda Ben-Zion, Haoran Meng, Philip Maechling, John E. Vidale (USC), Egill Hauksson, Zachary Ross (Caltech), Zhigang Peng, Zefeng Li (Georgia Tech)

Scientific Challenge: Develop new methods and software to detect weak and unusual seismic events that currently go unreported.

Solution Approach: We have developed multiple techniques that use waveform similarity to detect and identify previously undetected signals.

The “large-T” approach: Use waveform similarity of multiple events over long periods of time (T).

The “large-N” approach: Use waveform similarity of single events as recorded on dense seismic arrays featuring a large number (N) of stations.

Fig. 1. Using a template matching procedure (Large-T method), we reprocessed the SCSN continuous waveform archive using the seismograms of about 300,000 previously recorded earthquakes as templates which took about 1 million GPU hours. This method identified 2.4 million earthquakes for the period 2008-2017, which is a 13 times increase over the standard SCSN regional catalog.
MEDDLY: Multi-terminal and Edge-valued Decision Diagram Library

What is MEDDLY?

- Library for decision diagrams (BDDs, MDDs, EVMDDs)
- Graph data structures
- Represent functions over discrete variables
- Can be very compact
- State-of-the-art algorithms

Why use MEDDLY?

- Digital logic verification (BDDs)
- Model checking (BDDs, MDDs)
- Counterexamples (EVMDDs)
- Integer constraint problems
- Applications needing large sets of vectors of integers

http://meddly.sourceforge.io  Andrew Miner (PI)  Gianfranco Ciardo (Co-PI)  Iowa State University
Larch: Integrating Synchrotron X-Ray Analysis Methods

Larch provides an open-source set of libraries and applications for synchrotron X-ray methods.

Synchrotron beamlines like X-ray microprobes produce *multi-modal data* including X-ray absorption spectroscopy, fluorescence imaging, X-ray diffraction, diffraction imaging, and tomography.

**Larch** provides a coherent, tested, documented, and extensible library for treating these datasets.

It uses the *scientific Python* stack, and provides GUI apps for visualization and analysis of X-ray spectroscopies and related methods. This gives a shallow entry for non-expert scientists users, and also full scripting capabilities for more experienced users.

Synchrotron users run thousands of experiments per year, producing complex and heterogeneous datasets in a wide variety of fields.

Matthew Newville, U Chicago
NSF SI2 (SSI) Project #1450468
[http://cars.uchicago.edu/xraylarch](http://cars.uchicago.edu/xraylarch)
Distributed MultiThreaded Checkpointing (DMTCP)

NSCI: SI2-SSE: Extensible Model to Support Scalable Checkpoint-Restart... Save running computation and restart (possibly on different computers)

Several Use Cases:

1. **Checkpoint Many Times / Restart Once:** Save and restart long-running computation (in case of computer crash); EXAMPLE: MPI for HPC

2. **Checkpoint Once / Restart Many Times:**
   - Execute long-running initialization once and checkpoint — Then restart many times
   - EXAMPLE (Emulation of new CPU chips by Intel):
     - Start an operating system using hardware emulator, checkpoint, and restart many times to check common applications (e.g., office suite)

3. **Checkpoint Once / Restart Many Places:**
   - EXAMPLE (formal verification): Explore many paths in parallel

DMTCP transparently supports a wide range of environments:
- MPI, InfiniBand, CUDA/GPU, Distributed Software, ...
Empirical Methods for Computational Science

Tim Menzies
Petascale astrophysics, software infrastructure development, and community engagement

6. Next phase

Exascale Enzo-E
strong scaling heterogeneous

5. Scaling

256K BW cores
hydro+tracer
≈ 98% || effic.
fully distributed
>> ENZO

1. ENZO
community developed structured AMR
astrophysics & cosmology
powerful but scaling issues

2. Enzo-P
“Petascale Enzo”
ENZO physics
redesigned AMR
Cello framework

3. Cello
highly scalable reusable framework
array of octrees
uses Charm++

4. Charm++
parallel programming system
data-driven / asynchronous
targeting Exascale apps.
**Problem:** Reproducibility in scientific computing depends on the ability to run the exact same analysis. Any change in the environment, makes consistency a challenge.  

**Wholly!:** Tool to build reproducible, portable, minimal, and self-contained software packages. Specify environment to minimize external dependencies.