CRC CENTER for RESEARCH COMPUTING
UNIVERSITY of NOTRE DAME

Data and Software Preservation for Open Science

Jarek Nabrzyski
Director, Center for Research Computing
Department of Computer Science and Engineering
University of Notre Dame
naber@nd.edu

About me

- Born and raised in Poland
- Spent 13 years at Poznan Supercomputing and Networking Center
  - Involved in many EU-funded projects
- 7 years ago moved to the US
  - Center for Computation and Technology, LSU (2008 – mid 2009)
  - Center for Research Computing (CRC), Notre Dame (since mid 2009)

CRC

- Multidisciplinary enterprise
  - 10 faculty, 10 HPC engineers and user support, 20 research programmers, admin staff, plus grad students and undergrad interns
  - HPC/HTC services and research
  - Cyberinfrastructure (eScience) development, Data management
  - Science and CI teams working on projects in science, engineering and humanities

What is this talk really about?

Q: Preservation for what?

A: For reproducibility/reuse/replicability/r…
in computational science

Science and digital age

Science is the mother of the digital age

However, since the moment CERN has created the open internet, science has struggled to go digital and to go open.

What is open science and why is it important?

What is open science?

The term refers to efforts by researchers, governments, research funding agencies and the scientific community itself **to make the primary outputs of publicly funded research results** – publications and the research data (and software if possible) – **publicly accessible in digital format with no or minimal restriction** as a means for accelerating research.

These efforts are in the interest of enhancing transparency and collaboration, and fostering innovation.

Scientific Ideals

Innovative ideas

Reproducibility (the cornerstone of the scientific method)

Accumulation of knowledge

CORRESPONDENCE

Believe it or not: how much can we rely on published data on potential drug targets?

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research ...st appropriately represented ...marized by p-values, but, ...ately, there is a widespread ...at medical research articles

...an be proven that ...t claimed research ...dings are false.

...e interpreted based only on ... Research findings are defined ...ny relationship reaching ...atistical significance, e.g., ... interventions, informative ...s, risk factors, or associations. ...” research is also very useful. ...” is actually a misnomer, and ...terpretation is widespread. ..., here we will target ...hips that investigators claim ...her than null findings. ... been shown previously, the ...ty that a research finding ... true depends on the prior ...ty of it being true (before ...e study), the statistical power ...dy, and the level of statistical ...ice [10,11]. Consider a 2 × 2 ...which research findings are ...d against the gold standard ...lationships in a scientific ...a research field both true and ...otheses can be made about ...nce of relationships. Let R ...tio of the number of “true ...hips” to “no relationships” ...ose tested in the field. R

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets PPV = $(1-\beta)R/(R-\beta R+\alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored

Challenges

Lack of documentation of the workflow

Lack of transparency across the workflow

Lack of discoverability, especially unpublished work

Hard to recover the context of experiments

What do we do about it?

ND's efforts to promote Open Science

- DASPOS – Data and Software Preservation for Open Science
- National Data Service
- Collaboration on Open Science Framework with the Center for Open Science
- Series of Workshops

CRC CENTER for RESEARCH COMPUTING

UNIVERSITY of NOTRE DAME

DASPOS Data and Software Preservation for Open Science

www.daspos.org

ABOUT | PEOPLE | WORKSHOPS | RESEARCH | REPORTS

The massive data sets accumulated by High Energy Physics (HEP) experiments represent the most direct result of the often decades-long process of construction, commissioning and data aquisition that characterize this science. Many of these data are unique and represent an irreplaceable resource for potential future studies. Forward-thinking efforts for preservation are necessary now in order to achieve the relevant parameters, analysis paths and software to preserve the usefulness of these rich and varied data sets.

> "Ten or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able to resuse it. That means we are going to need to save it as open data..."
>
> Rolf-Dieter Heur 2008
> Director General, CERN

Data and Software Preservation for Open Science, DASPOS, represents an initial exploration of the key technical problems that must be solved to provide appropriate data, software and algorithmic preservation for HEP, including the contexts necessary to understand, trust and reuse the data. While the archiving of HEP data may require some HEP-specific technical solutions, DASPOS will create a template for preservation that will be useful across many different disciplines, leading to a broad, coordinated effort.

Discovery and Coordination

Series of highly-structured public workshops to define, discuss and document the details of data and software preservation

Prototyping and Experimentation

Key areas of reserach: data and query models and software sustainability models

The DASPOS Team

Computer science experts, experienced digital librarians, and experts in data-intensive fields, such as physics, astrophysics and bioinformatics

First Workshop Scheduled

The first DASPOS Workshop has been scheduled for Thursday - Friday, March 21-22, 2013, at CERN. More information

Workshop 1

2012-12-17 19:11:04

WORKSHOP 1 Establishment of Use Cases for Archived Data and Software in HEP Date: Thursday-Friday…

Workshop 2

2012-12-17 19:11:04

WORKSHOP 2 Survey of Commonality with other Disciplines Attendees: Broad participation from many…

DASPOS

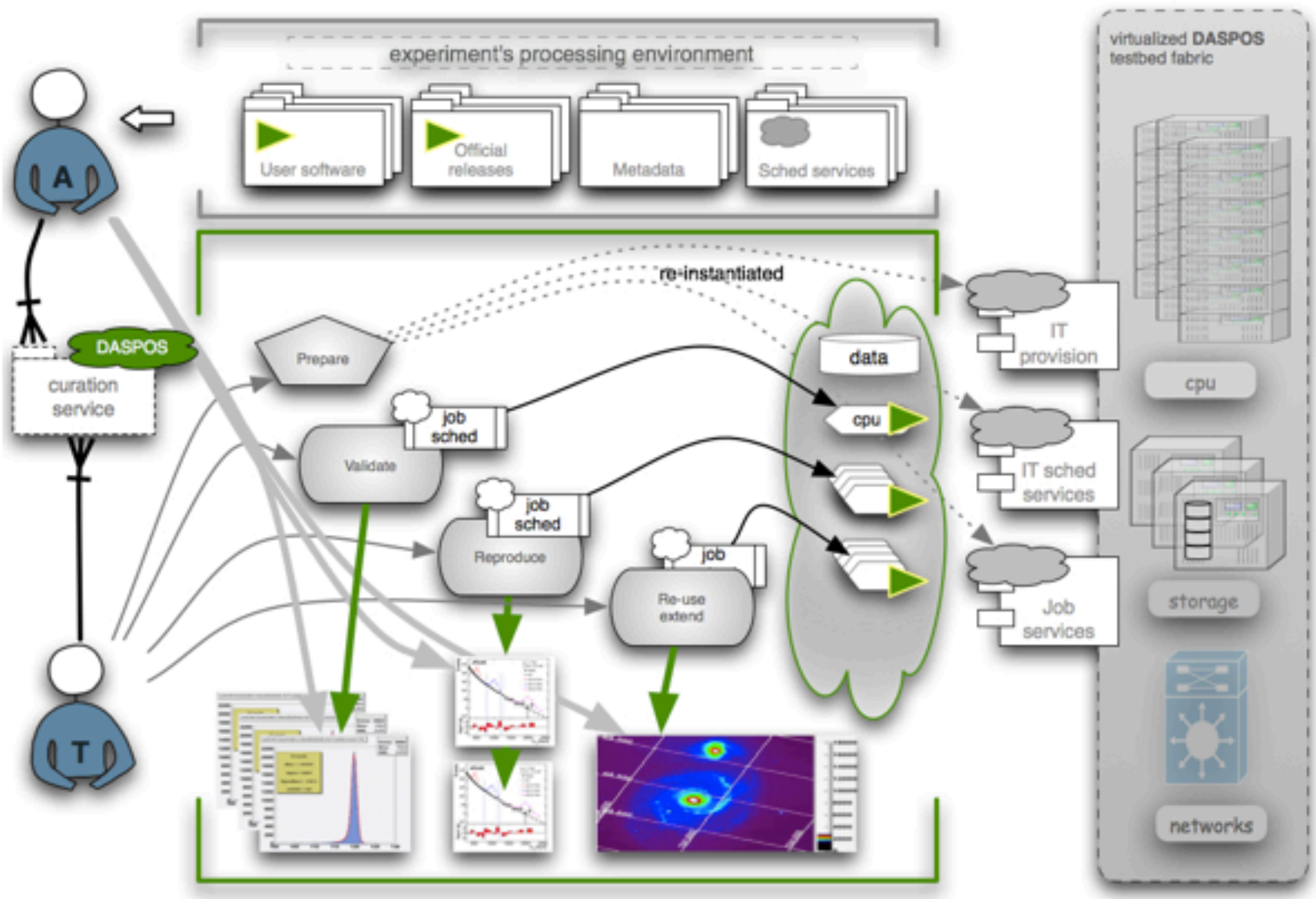
- Data And Software Preservation for Open Science
  - multi-disciplinary effort funded by NSF
    - Notre Dame, Chicago, UIUC, Washington, Nebraska, NYU, (Fermilab, BNL)
- Links HEP effort (DPHEP + experiments) to Biology, Astrophysics, Digital Curation
  - includes physicists, digital librarians, computer scientists
  - aims to achieve some commonality across disciplines in
    - meta-data descriptions of archived data
      - What's in the data, how can it be used?
    - computational description (ontology development)
      - how was the data processed?
      - can computation replication be automated?
    - impact of access policies on preservation infrastructure

Many different Rs…

- Reproduce precisely what someone else did on the same resources, with the same techniques.
- Recreate an equivalent computation on different resources, with similar techniques.
- Repurpose an experiment by running it again with a slight change to the data, software, or environment.
- Reuse the same artifact across many different experiments, for a longitudinal comparison.
- Rely on one party to set up an environment and make it usable for multiple parties. (Think sysadmins)
- Other Rs?

CRC CENTER for RESEARCH COMPUTING
UNIVERSITY of NOTRE DAME

Curation Challenge



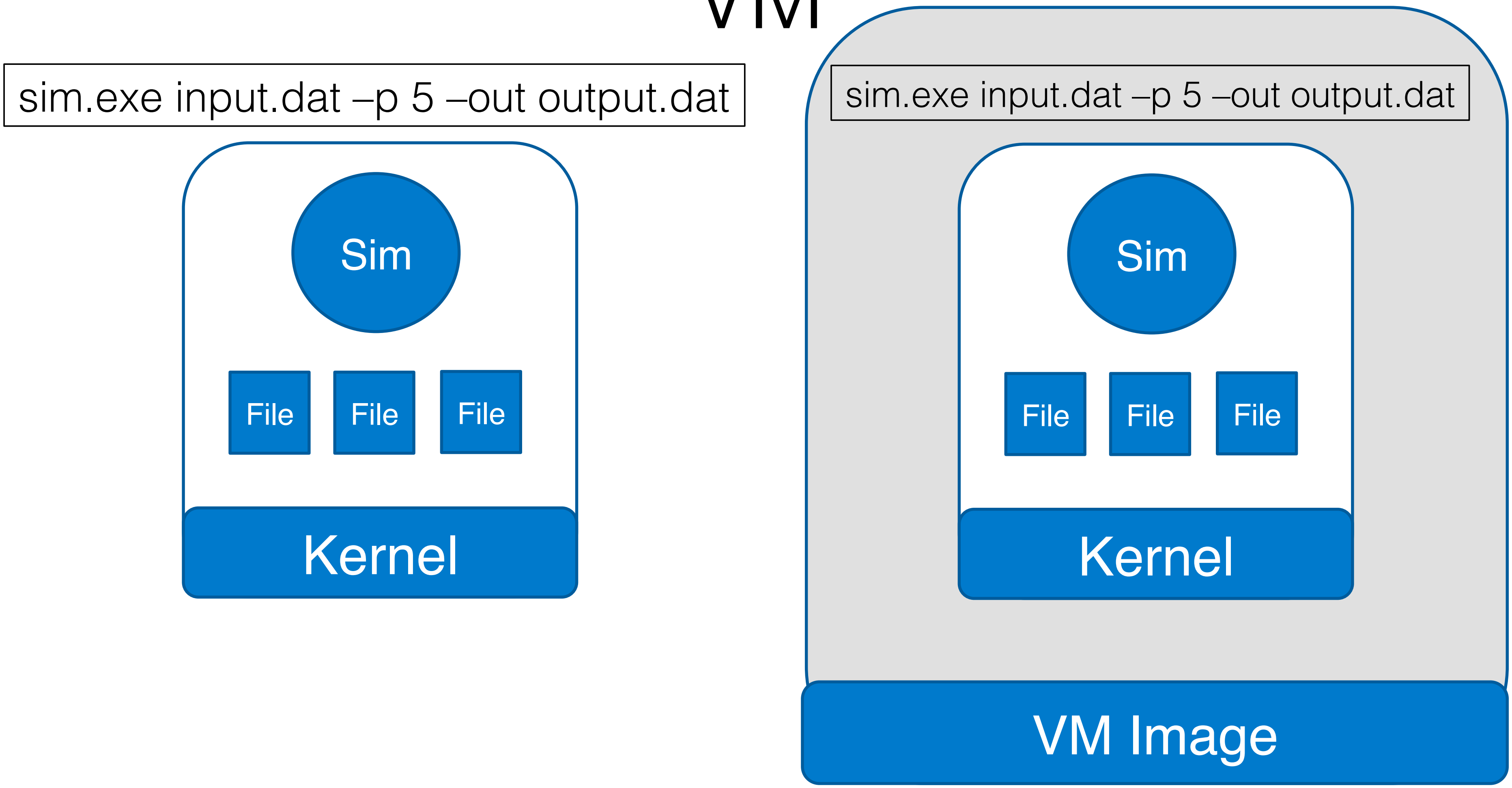
Reproducibility in e-Science is absolutely terrible today!

- Can I re-run a result from a colleague from five years ago successfully, and obtain the same result? How about a student in my lab?
- Today, are we preparing for our current results to be re-used by others five years from now?
- Multiple reasons why not:
  - Rapid technological change.
  - No archival of artifacts.
  - Many implicit dependencies.
  - Lack of backwards compatibility.
  - Lack of social incentives.
  - Lack of transparent tools…

Typical Computational Experiment

- PI gives student some general directions. Student writes some code, does some experiments, saves the outputs, writes the paper.
- Source code is often carefully curated. But what about the operating system, the software dependencies, the experimental configuration, the input data, etc…
- If we did manage to re-run everything, do we have a means of verifying equivalence?
- Concurrency + Floating Point != Bitwise Equality.

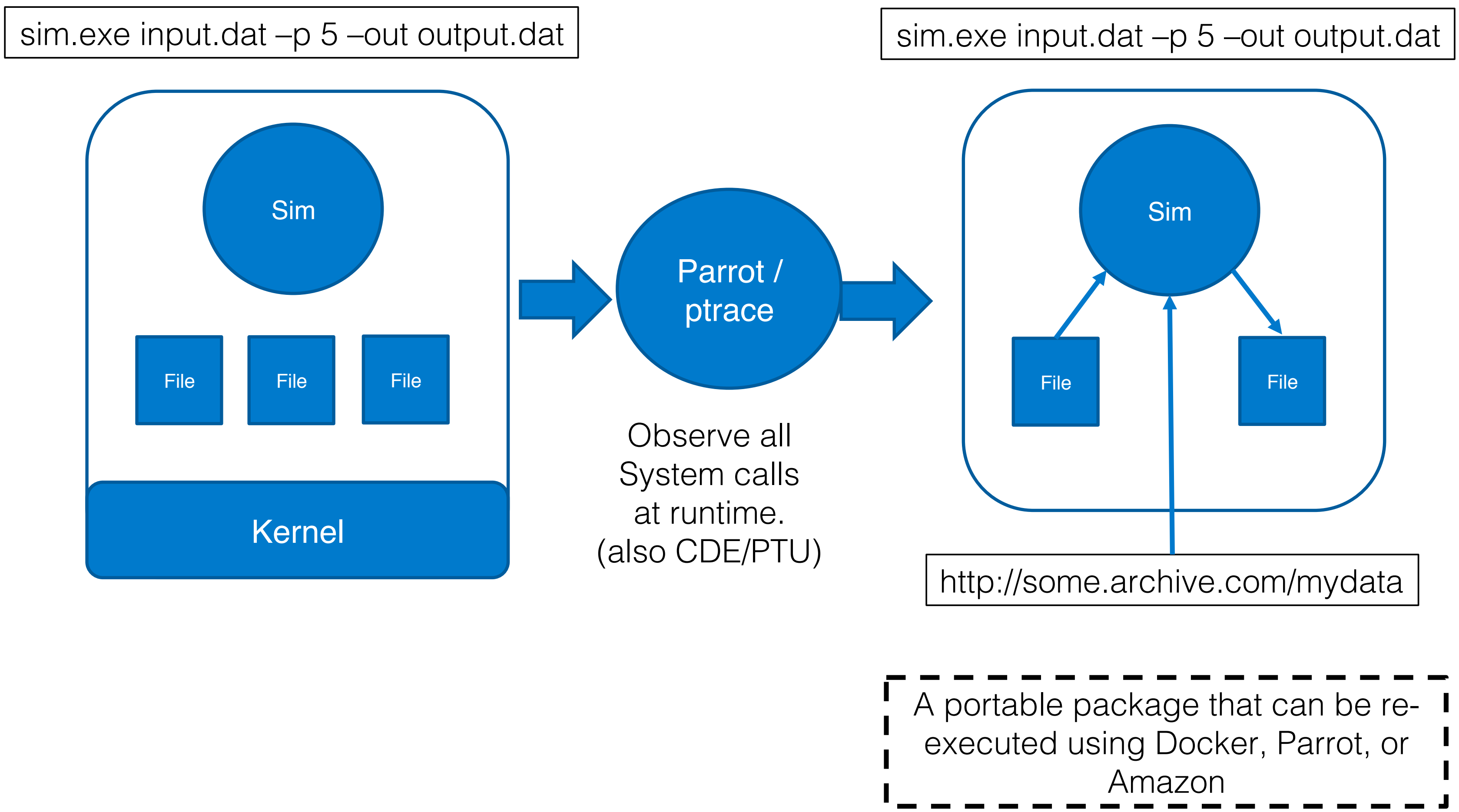
Preserve the Mess: Stick it all into a VM



Preserve the Mess: Stick it all into a VM

- A good place to start, however:
  - Captures more things than necessary.
  - For many experiments will duplicate large amounts of software/data in the VM images.
  - Hard to disentangle things logically – what if you want to run the same experiment with some component of the OS/software/data changed?
  - Doesn't capture network interactions.
  - May be coupled to specific a VM technology.
  - VMs are not the place to archive data.

Preserve the Mess: Trace All Interactions



Preserve the Mess: Trace All Interactions

- Solves some problems:
  - Only captures what is actually used.
  - Once captured, not coupled to a technology.
  - Observes network dependencies.
- But not all of them:
  - For many experiments will duplicate large amounts of software/data in the VM/package images.
  - Hard to disentangle things logically – what if you want to run the same experiment with some component of the OS/ software/data changed.
  - VMs/packages are not the place to archive data.

What we really want:

A *structured* way to compose an application with all of its dependencies.

Enable preservation, but also re-use of data and images for efficiency.

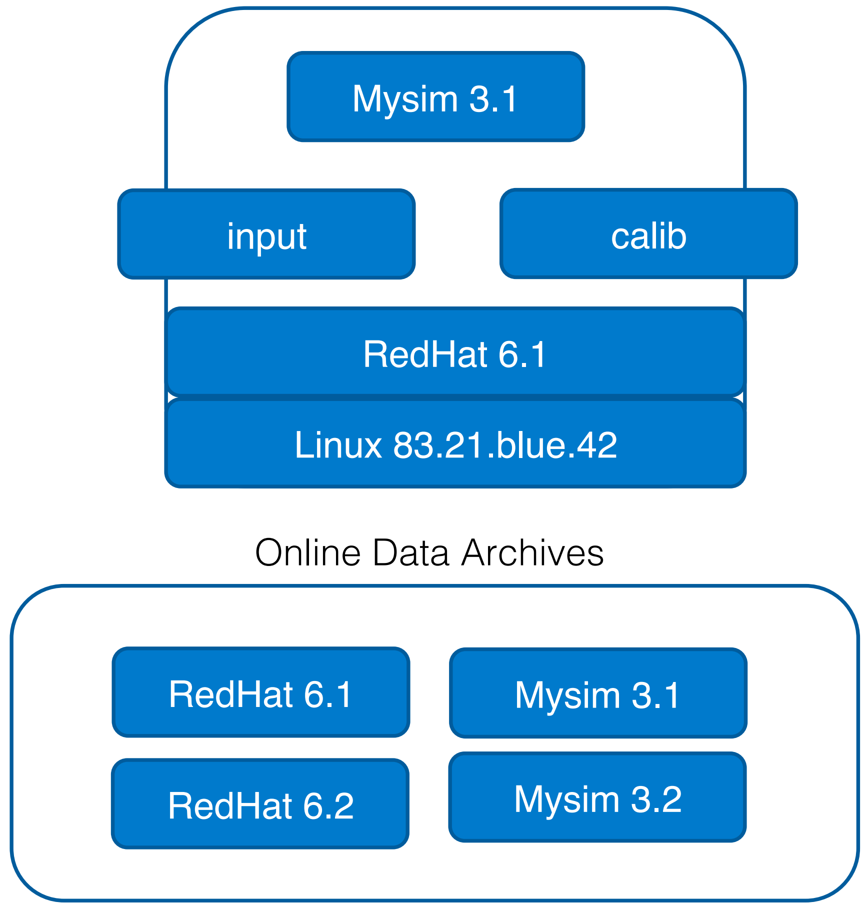
It also would be good to capture the context of the computational experiment.

Umbrella

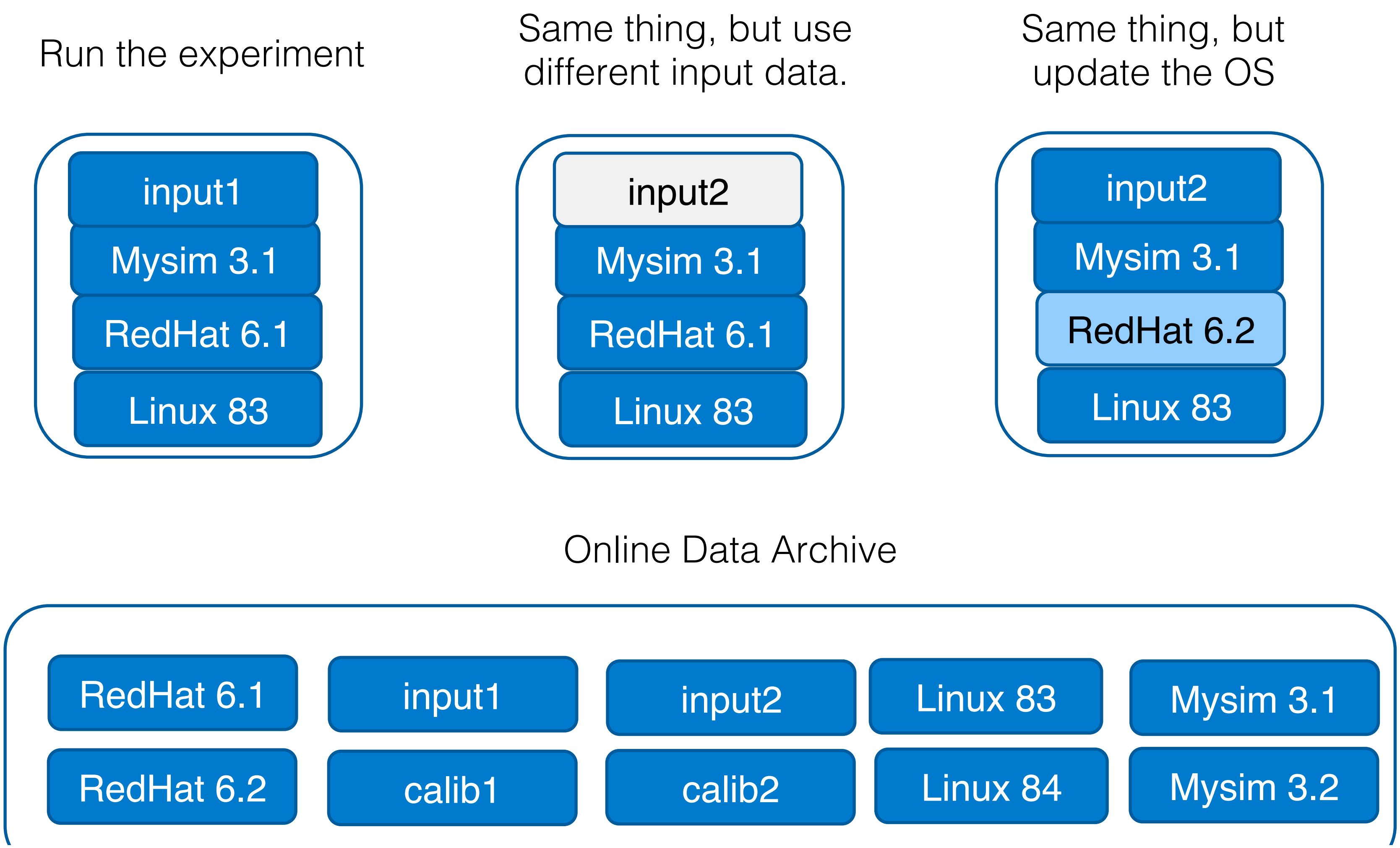
myenv1.json

```
kernel = {
name = "Linux";
version = "83.21.blue.42"
}
opsys = {
name = "RedHat";
version = "6.1"
}
software = {
simulator = {
mount = "/soft/sim";
name = "mysim-3.1";
}
data = {
input = {
mount = "/data/input";
url = "http://some.url"; }
calib = {
mount = "/data/calib";
url = "http://other.url"; }
}
```

umbrella run myenv1.json



Umbrella specifies a reproducible environment while avoiding duplication and enabling precise adjustments.



Specification is More Important than Mechanism

- Current version of Umbrella can work with:
  - Docker – create container, mount volumes.
  - Parrot – Download tarballs, mount at runtime.
  - Amazon – allocate VM, copy and unpack tarballs.
  - Condor – Request compatible machine.
- More ways will be possible in the future as technologies come and go.
- Key requirement: Efficient runtime composition, rather than copying. (Compare to Dockerfile.)

How do we construct complex workflows from these building blocks?

PRUNE – Preservation Run Environment

- Problem: Our user interfaces do not accurately capture the dependencies or the environment of the codes that we run.
- Can we improve upon the standard command-line shell interface to make it reproducible?
- Re-use a good idea: functional representation.

  output = mysim(input, calib) USING ENV myenv.json

- Build on ideas from GridDB, VDL, Swift, Taverna, Galaxy, but focus is on precise reproduction, not on performance (coarse granularity.)

PRUNE – Preservation Run Environment

PUT "/tmp/input1.dat" AS "input1" [gets id 3ba8c2]

PUT "/tmp/input2.dat" AS "input2" [gets id dab209]

PUT "/tmp/calib.dat" AS "calib" [gets id 64c2fa]

PUT "sim.function" AS "sim" [gets id fffda7]

out1 = sim(input1, calib) IN ENV myenv1.json

  [out1 is bab598]

out2 = sim(input1, calib) IN ENV myenv2.json

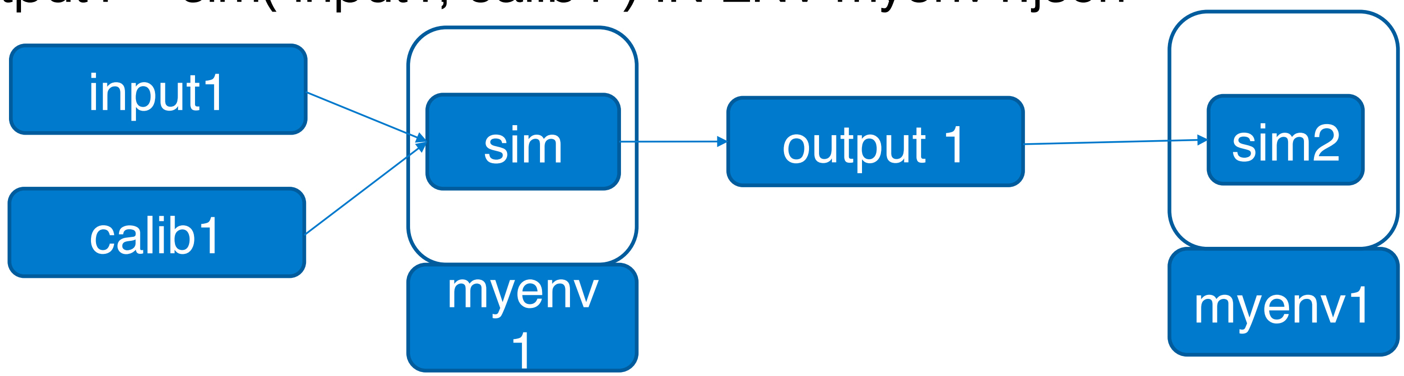
  [out2 is 392caf]

out3 = sim(input2, calib) IN ENV myenv2.json

  [out3 is 232768]

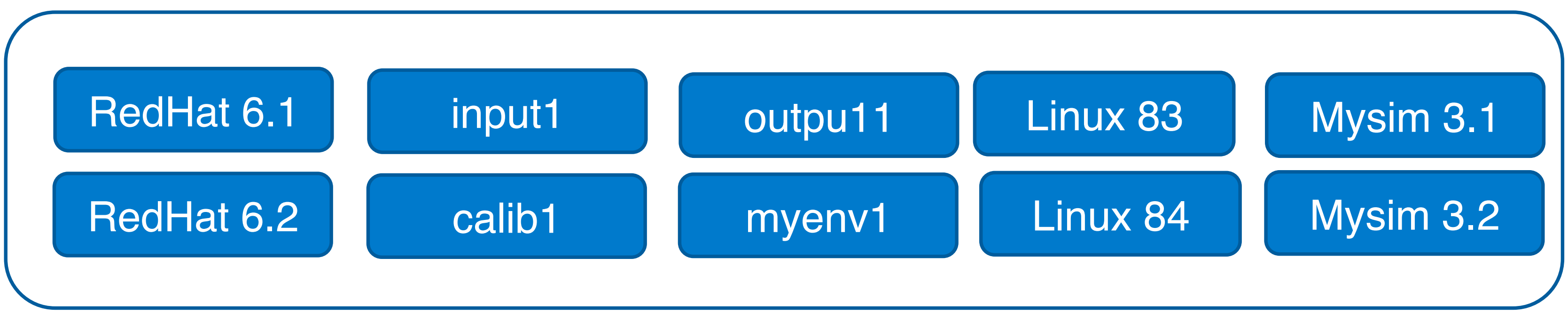
PRUNE connects together precisely reproducible executions and gives each item a unique identifier

output1 = sim(input1, calib1) IN ENV myenv1.json



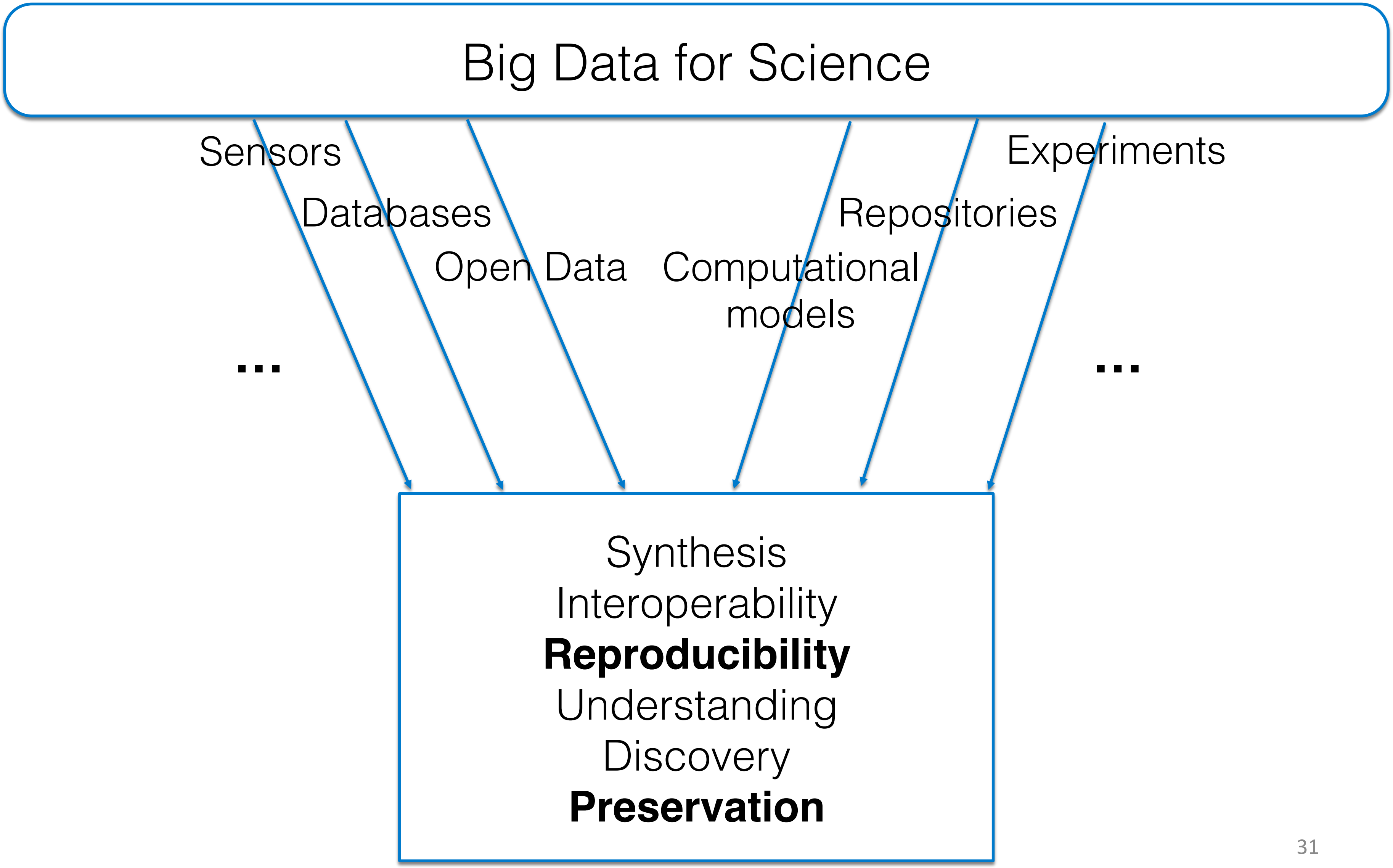
Bab598 = fffda7 (3ba8c2, 64c2fa) IN ENV c8c832

Online Data Archive



All Sorts of Open Problems

- Naming: Tension between usability and durability. At least two levels of naming.
- What is the intersection of version control (doc deltas) and provenance (doc ops) ?
- Usability: Can we accommodate existing work patterns, or do we force new habits?
- Repositories: Who will run them, how many should we have, what will they cost…? Will they be persistent?
- Compatibility: Can we work in existing workflow technologies without starting over?
- Composition: MPI, BoT, Workflows, Map-Reduce, …


Big Data for Science
Sensors
Databases
Open Data
Computational models
Repositories
Experiments
...
...
Synthesis
Interoperability
Reproducibility
Understanding
Discovery
Preservation

"…it does not answer the question how one would **discover** the required data in today's chaotic information universe, how one would **understand** which datasets can be meaningfully **integrated**, and how to **communicate** the results to humans and machines alike."

A sign in New Cuyama, Santa Barbara County, California; original picture by Mike Gogulski (CC BY 2.5).

Krzysztof Janowicz, Frank van Harmelen, James A. Hendler, and Pascal Hitzler. "Why the Data Train Needs Semantic Rails." AI Magazine, 2014

Formal semantic models capture these data dimensions and allow them to be **shared** via linked open data principles.

Krzysztof Janowicz, Frank van Harmelen, James A. Hendler, and Pascal Hitzler. “Why the Data Train Needs Semantic Rails.” AI Magazine, 2014

A “design pattern” based approach allows us to create **reusable** “micro-ontologies” or building blocks that have high **quality** and minimize **unintended** logic consequences.

Aldo Gangemi. “Ontology Design Patterns for Semantic Web Content.” In The Semantic Web–ISWC 2005, 262–76. Springer, 2005. http://link.springer.com/chapter/10.1007/11574620_21.

"Linked Open Data for Computational Science?"

page | discussion | edit | history | Log in

navigation
- Main page
- Recent changes
- Random page

search

Go | Search

tools
- What links here
- Related changes
- Special pages
- Printable version
- Permanent link

GeoVoCampSOCoP2014

Contents [hide]

Where, When and How [edit]

We have now started organizing a SOCoP 2014 workshop and Geo-VoCamp as follow-up to previous SOCoP workshops and VoCamps participated in by SOCoP members.

The GeoVoCamp will take place at the USGS National facility Dec. 1-3 (starting at 12:30 on Dec 1) in Reston VA. The session will start in the **Dallas Peck Auditorium** and be organized into breakouts around 3- 4 Work Groups made up of domain experts, group facilitators and people with semantic/ontological experience.

More details will be available in mid-Nov. Currently 3 workgroup topics have been developed.

We thank the USGS for providing space and support for this event held at the USG facility in Reston VA. As with previous workshops this will be organized around volunteer Work Groups. A focus has been continuation and expansion of work from prior workshops including work on Ontology Design Patterns and the Descartes Core started on in 2012. New topics such as a Settings pattern are started at particular meetings based on the interest of the participants.

Gary Berg-Cross and David Carral

GeoVocamp SOCoP 2014: http://vocamp.org/wiki/GeoVoCampSOCoP2014

CRC CENTER for RESEARCH COMPUTING
UNIVERSITY of NOTRE DAME

kno.e.sis WRIGHT STATE UNIVERSITY

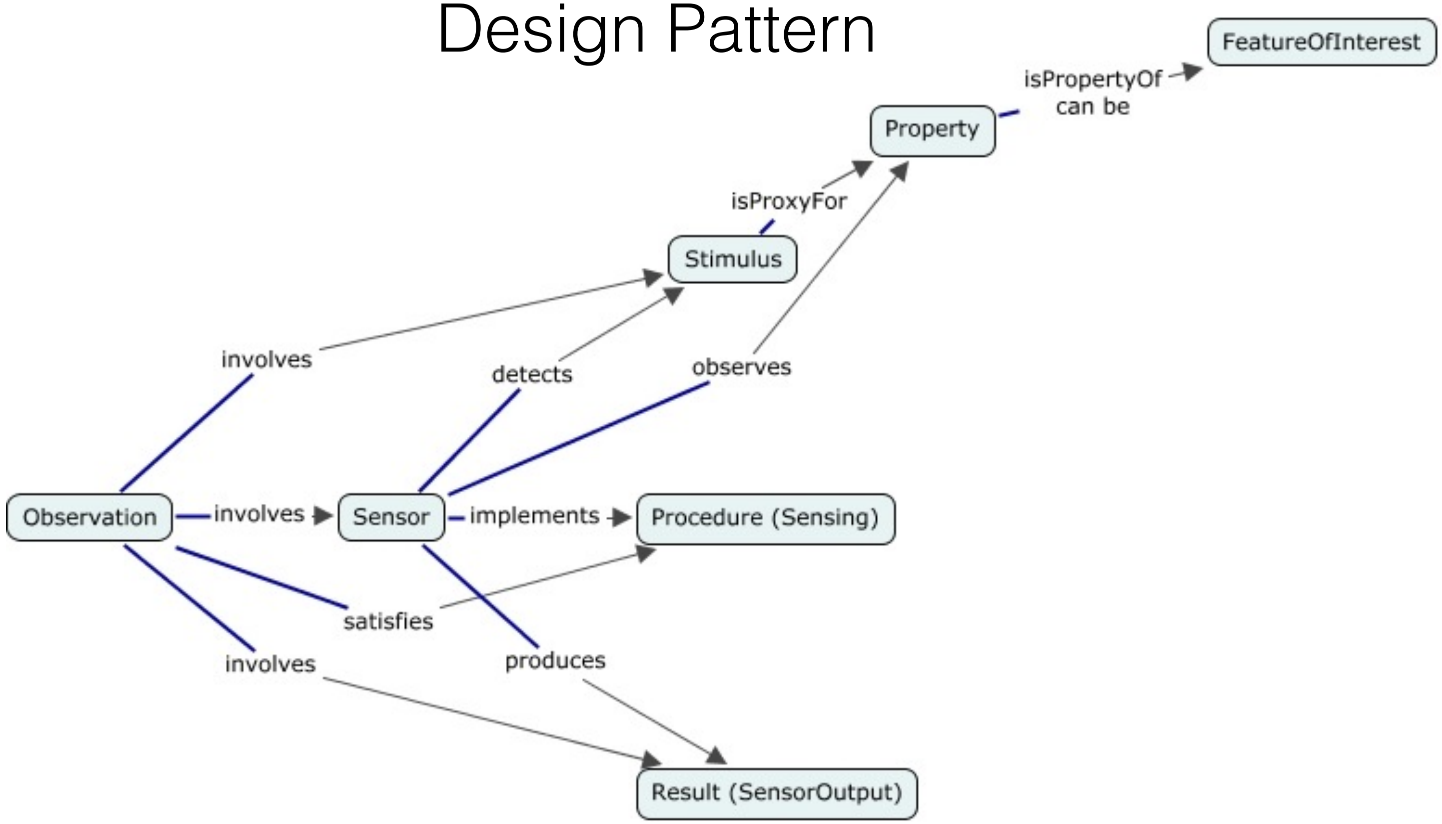
DaSe Lab, Kno.e.sis Center, Wright State University

David Carral, Adila Krisnadhi, Michelle Cheatham, Pascal Hitzler

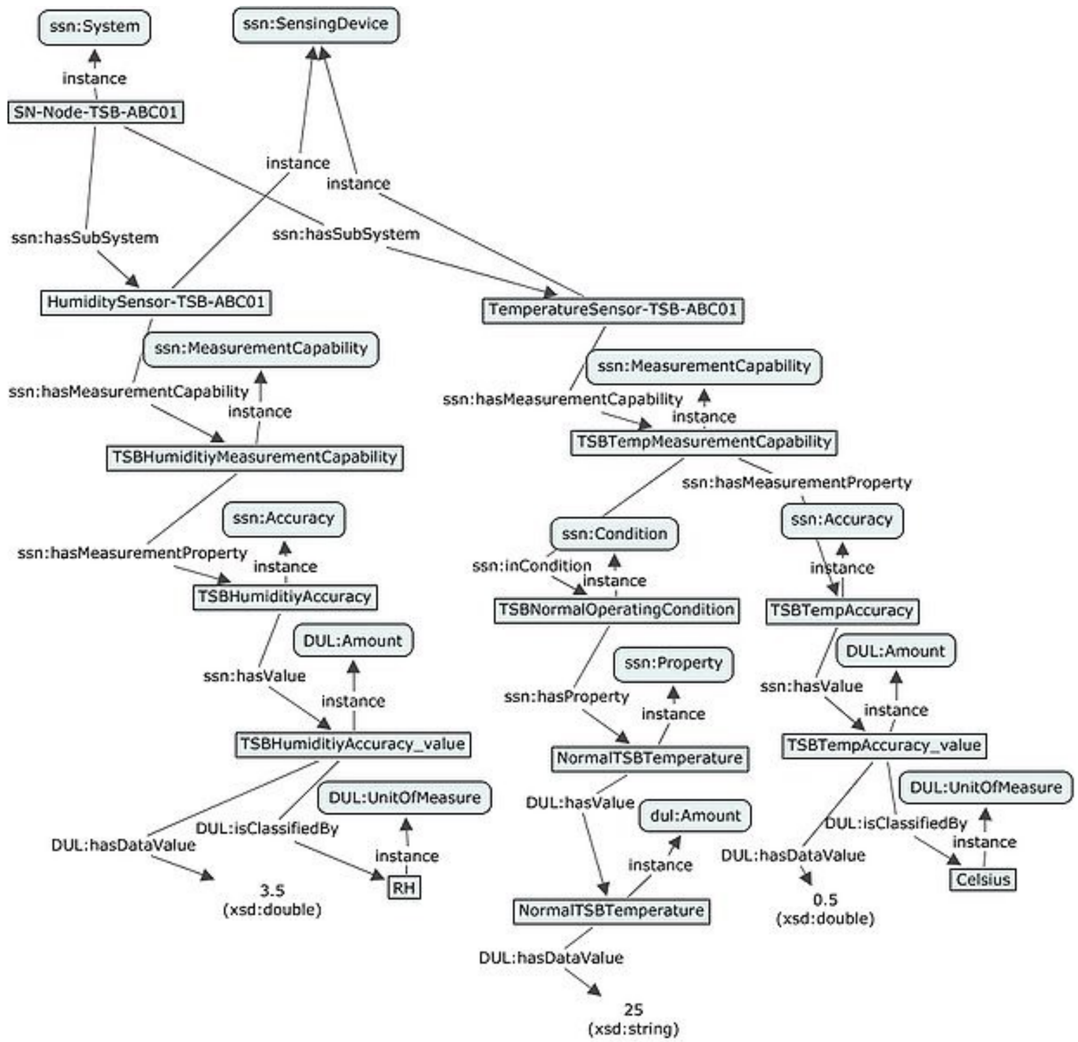
Credit: Beatrice Murch, Creative Commons License, https://www.flickr.com/photos/blmurch/2754681293/sizes/l/

How did you take it’s temperature?

The Stimulus-Sensor-Observation Ontology Design Pattern



SSN: http://www.w3.org/2005/Incubator/ssn/wiki/SSN_Skeleton#The_Stimulus-Sensor-Observation_Ontology_Design_Pattern



SSN: http://www.w3.org/2005/Incubator/ssn/wiki/Report_Work_on_the_SSN_ontology

How might a Computational Scientist take it's temperature?

Temperature

$$T = \frac{2}{3k_B}\left\langle \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathbf{p}_i|}{2m_i}\right\rangle$$

But this definition depends on some **computational** model that captures the molecular behavior of water…

Google

tip4p

+Charles Share

Web Shopping Images Videos News More Search tools

About 685,000 results (0.16 seconds)

Water model - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Water_model Wikipedia
The potential for models such as TIP3P and **TIP4P** is represented by. E_{ab} = \ sum_{i} ^{\text{on. where kC, the electrostatic constant, has a value of 332.1 ...
Simple water models - 2-site - 3-site - 4-site

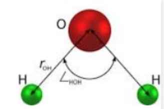
TIP4P model of water page on SklogWiki - a wiki for ...
www.sklogwiki.org/SklogWiki/index.php/**TIP4P**_model_of_water
Jan 20, 2011 - The **TIP4P** model is a rigid planar four-site interaction potential for water, ... The **TIP4P** model consists of a Lennard-Jones site for the oxygen ...
Parameters - Phase diagram - Shear viscosity - Virial coefficients

Water models
www.lsbu.ac.uk/water/models.html London South Bank University
Apr 1, 2014 - Water molecular models including SPC, SPC/E, TIP3P, **TIP4P**, TIP5P, PPC, POL5 , SSD and SWFLEX.

pair_style lj/cut/coul/long - Lammps
lammps.sandia.gov/doc/pair_lj.html Sandia National Laboratories
style = lj/cut or lj/cut/coul/cut or lj/cut/coul/debye or lj/cut/coul/dsf or lj/cut/coul/long or lj/cut/coul/msm or lj/cut/**tip4p**/long; args = list of arguments for a particular ...

[PDF] **TIP4P**-Ew - Stanford University
www.stanford.edu/.../horn_**tip4p**EW_2004jcp.pdf Stanford University
by HW Horn - 2004 - Cited by 557 - Related articles
May 22, 2004 - A re-parameterization of the standard **TIP4P** water model for use with Ewald techniques is introduced, providing an overall global improvement ...

Water model



In computational chemistry, classical water models are used for the simulation of water clusters, liquid water, and aqueous solutions with explicit solvent. These models use the approximations of molecular mechanics. Wikipedia

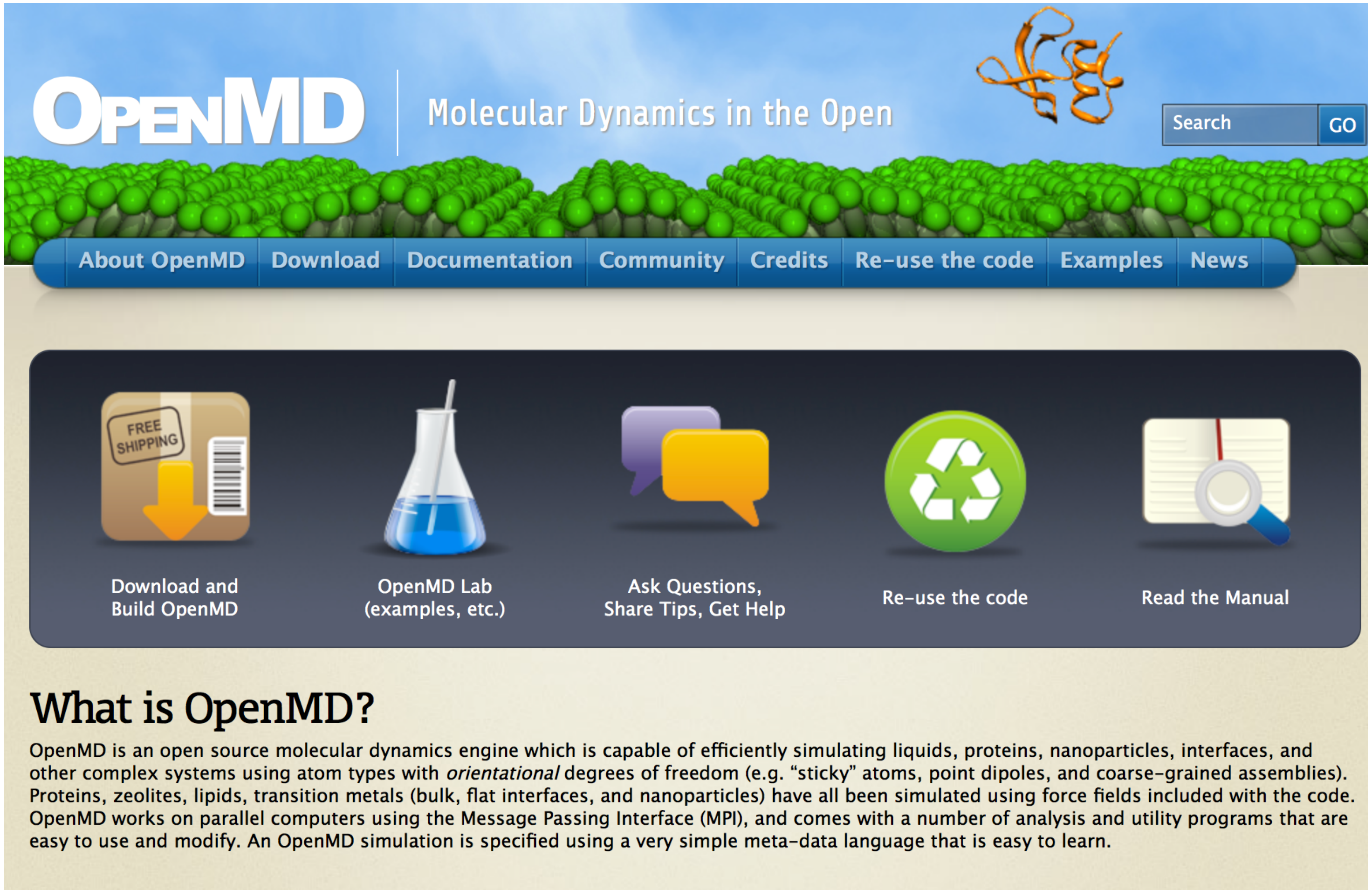
Related topics

In most water models, the **Lennard-Jones** term applies only to the interaction between the oxygen atoms. Wikipedia
Explore: Lennard-Jones potential

In-silico (see: water models), cyclic **water clusters** . . . are found with n = 3 to 60.
Wikipedia
Explore: Water cluster

Feedback

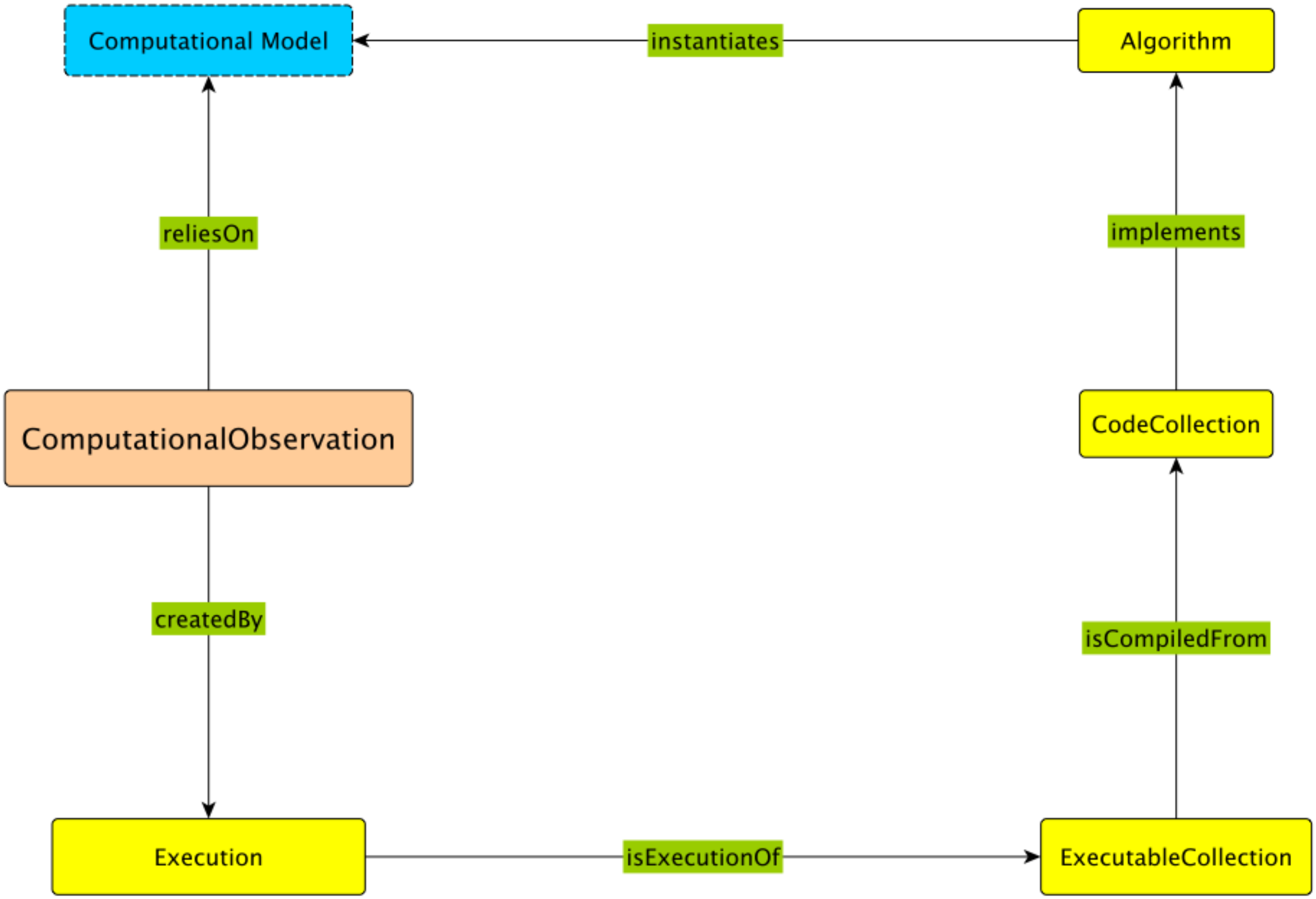
And some **software code** that implements the computational model by **algorithm**…

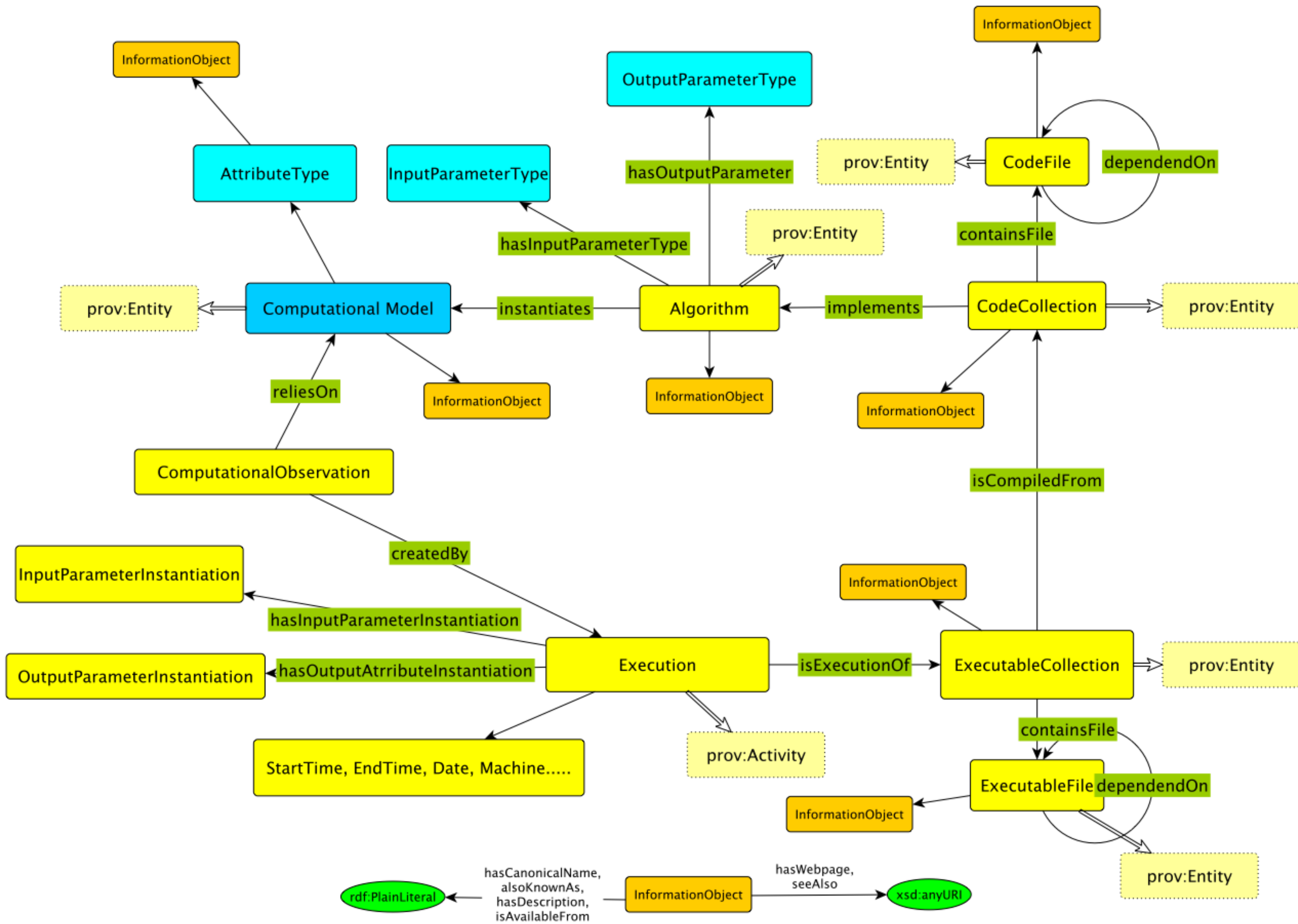


OpenMD: http://www.openmd.org

And some **execution** of the code that produces the data needed for an observation…

How to Connect "Physical Experimental Observation" to "Computational Experimental Observation"?


Computational Model
instantiates
Algorithm
reliesOn
implements
ComputationalObservation
CodeCollection
createdBy
isCompiledFrom
Execution
isExecutionOf
ExecutableCollection


InformationObject
OutputParameterType
InformationObject
AttributeType
InputParameterType
hasOutputParameter
prov:Entity
CodeFile
dependendOn
hasInputParameterType
prov:Entity
containsFile
prov:Entity
Computational Model
instantiates
Algorithm
implements
CodeCollection
prov:Entity
reliesOn
InformationObject
InformationObject
InformationObject
ComputationalObservation
isCompiledFrom
createdBy
InputParameterInstantiation
InformationObject
hasInputParameterInstantiation
OutputParameterInstantiation
hasOutputAtrributeInstantiation
Execution
isExecutionOf
ExecutableCollection
prov:Entity
containsFile
StartTime, EndTime, Date, Machine.....
prov:Activity
ExecutableFile
dependendOn
InformationObject
prov:Entity
rdf:PlainLiteral
hasCanonicalName, alsoKnownAs, hasDescription, isAvailableFrom
InformationObject
hasWebpage, seeAlso
xsd:anyURI

An Ontology Design Pattern towards Preservation of Computational Experiments

- Demand for scientifically reproducible and extensible preservations of computational experiments.
  - Replication of numeric results vs. Context of the calculation
  - Preservation should be described both in machine and human readable fashions
- Smart Container (SC) ontology towards conceptualizing computational experiments from the perspective of computational environments and activities within using Docker[1] LXC as a preservation tool.

Docker

- Docker Linux container as a scaffolder
- Light-weighted virtualization platform
- Versioned file system
- Modular design for distribution of software component
- A sustainable community(industry, CERN[2])

Refer or align existing ontologies and patterns, such as PROV-O[3], CSO[4] and ACT[5] for discoverability, interoperability, queryability and future extensibility.

Goals and "how?"

- Capture existing scientific workflow frameworks and descriptions with a Docker LXC
- Data to be integrated in a consistent manner
- A common description of a computational environment
- How?
    - An automated tool "wraps" the existing Docker command line
    - An infrastructure is transparent to scientists but also captures information necessary to populate the metadata behind the scenes.

Toward the Formalization of "Smart Containers"

- A modular approach by **systematic** alignment of concepts present in Docker as a computational environment.
- Reusing vocabulary terms where possible to contextualize computational activities.
- Assist to answer competency questions:
  1. "What are the requirements for a computational activity?"
  2. "What was the environment in which the activity was performed in terms of software components?"
  3. "What is the order in which provisioning activities must occur?"
  4. "What software agents are responsible for a particular result or outcome".

Toward the Formalization of "Smart Containers"

- PROV is used as a foundational building block to facilitate connection to other vocabularies and preservation efforts.

- The Core Software Ontology (CSO) formalizing concepts of software engineering, such as data, software and executions with data.

- ACT ontology design pattern provides temporal-ordered entities and a planning-related workflow axioms.

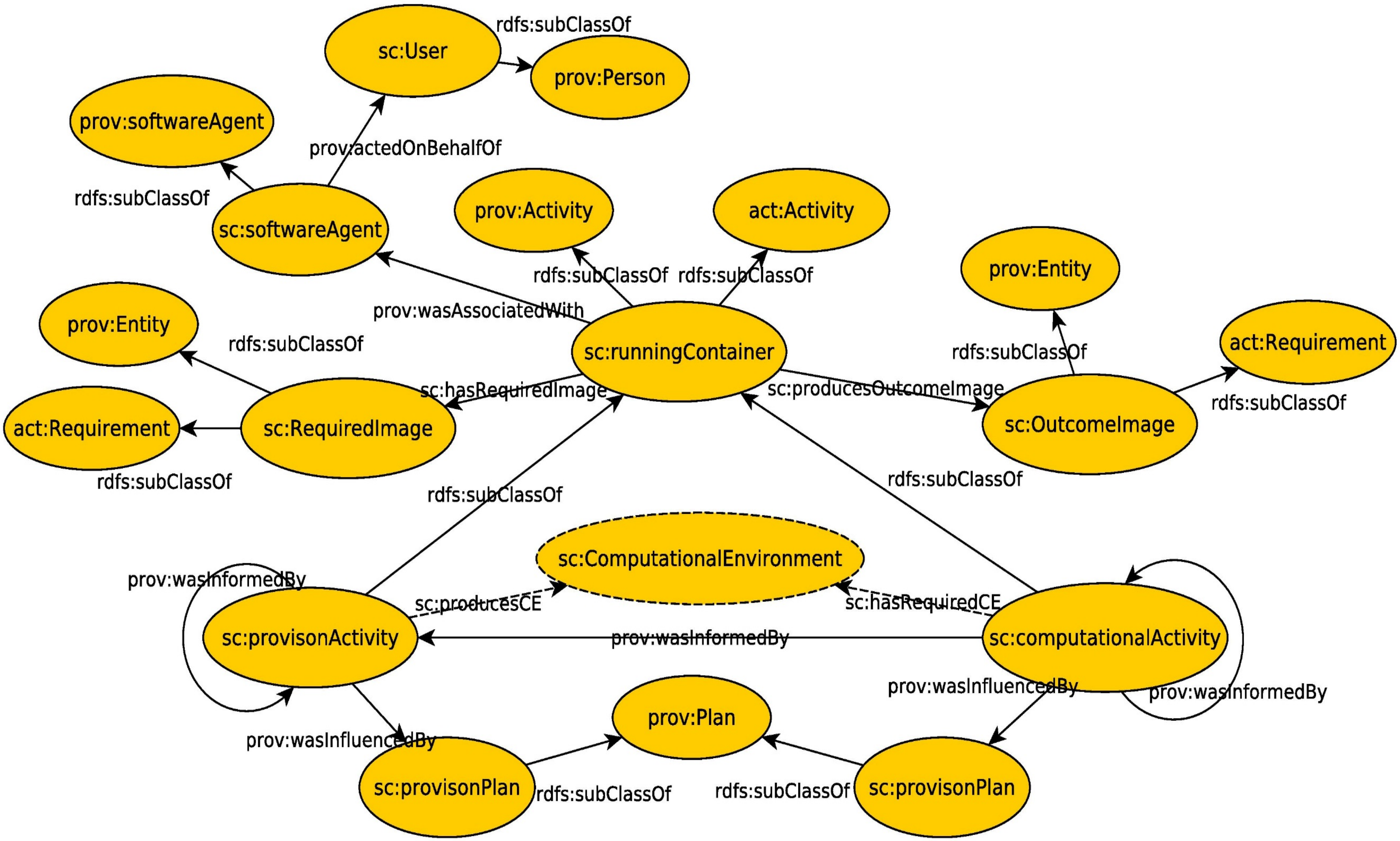
A Proposed Pattern

- **provisioning activities**: create an appropriate environment for computational activities. A sequence of provisioning activities was planned by a **provision plan.**

- **computational activities**: directly produce scientific observations and affected by a **workflow plan**

- **runningcontainer**: a Docker LXC concept that represents an activity **hasRequiredImage** and **producesImage**, also **was Associated With** a **software Agent**, which **actedOnBehalfOf** a **User**

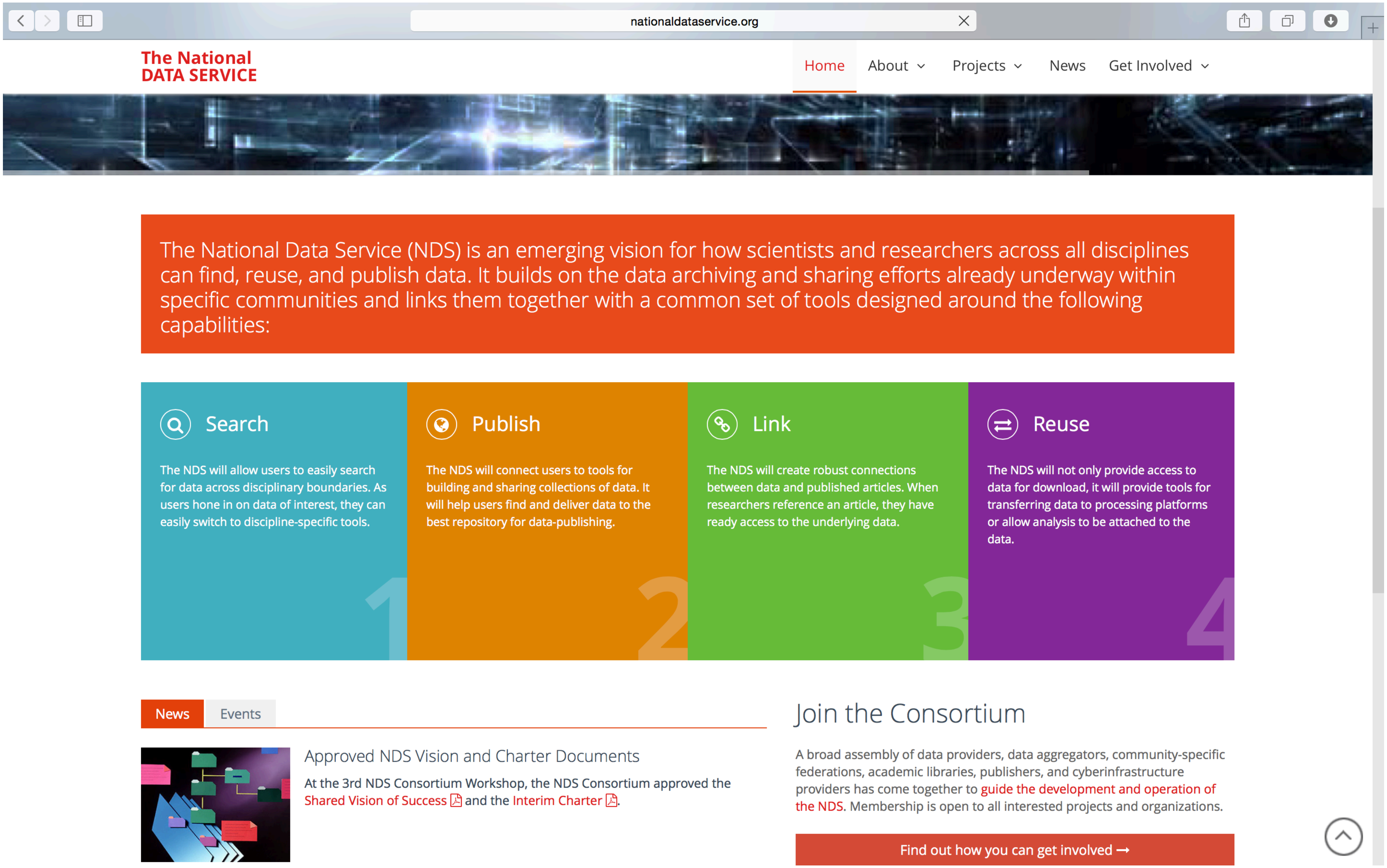
- use **rdf:seeAlso** to reference the human readable encoding of computational experiment as **SoftwareAsCode** which is a kind of **Information Object**


sc:User
rdfs:subClassOf
prov:Person
prov:softwareAgent
prov:actedOnBehalfOf
rdfs:subClassOf
sc:softwareAgent
prov:Activity
act:Activity
rdfs:subClassOf
rdfs:subClassOf
prov:Entity
prov:wasAssociatedWith
prov:Entity
act:Requirement
rdfs:subClassOf
sc:runningContainer
rdfs:subClassOf
rdfs:subClassOf
sc:hasRequiredImage
sc:producesOutcomeImage
rdfs:subClassOf
act:Requirement
sc:RequiredImage
sc:OutcomeImage
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
sc:ComputationalEnvironment
prov:wasInformedBy
sc:producesCE
sc:hasRequiredCE
sc:provisonActivity
prov:wasInformedBy
sc:computationalActivity
prov:wasInfluencedBy
prov:wasInformedBy
prov:Plan
prov:wasInfluencedBy
sc:provisonPlan
rdfs:subClassOf
rdfs:subClassOf
sc:provisonPlan

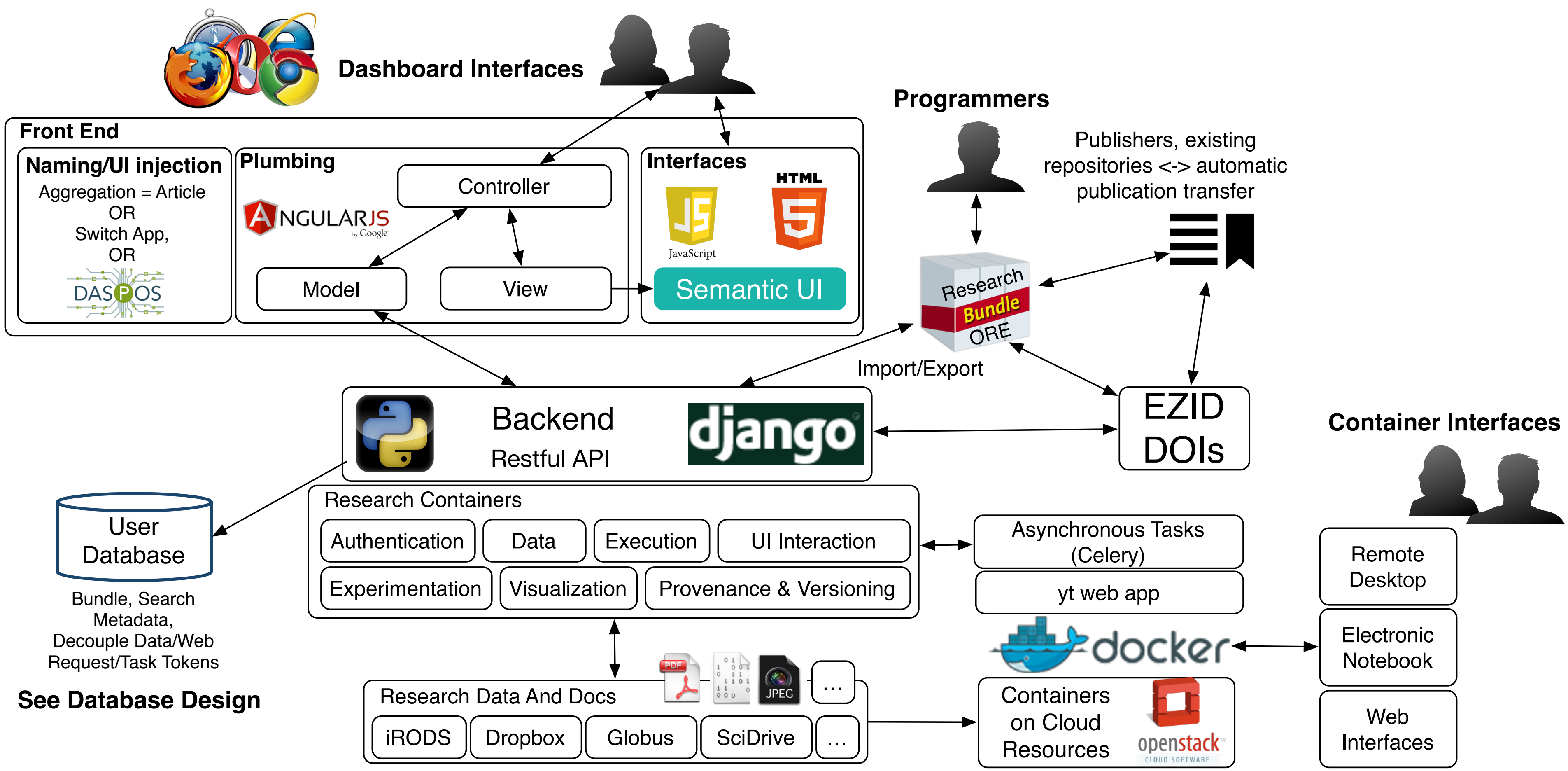
Smart Container

- Capture ENTIRE computational environment by using a modular, reusable, extensible approach
- The Docker effort provides most of this functionality
  - Sustainable Community
  - Versioned file system
  - Repository for components
- Smart Containers extends Docker by providing the ability to capture provenance and other metadata
- Complimentary to Umbrella and Prune approach. Would like to capture workflow but not entire computational environment for each execution of some software
- Ontology design pattern and use of ontologies provides the ability to understand and extend previous work by capturing the CONTEXT behind the scenes

National Data Service



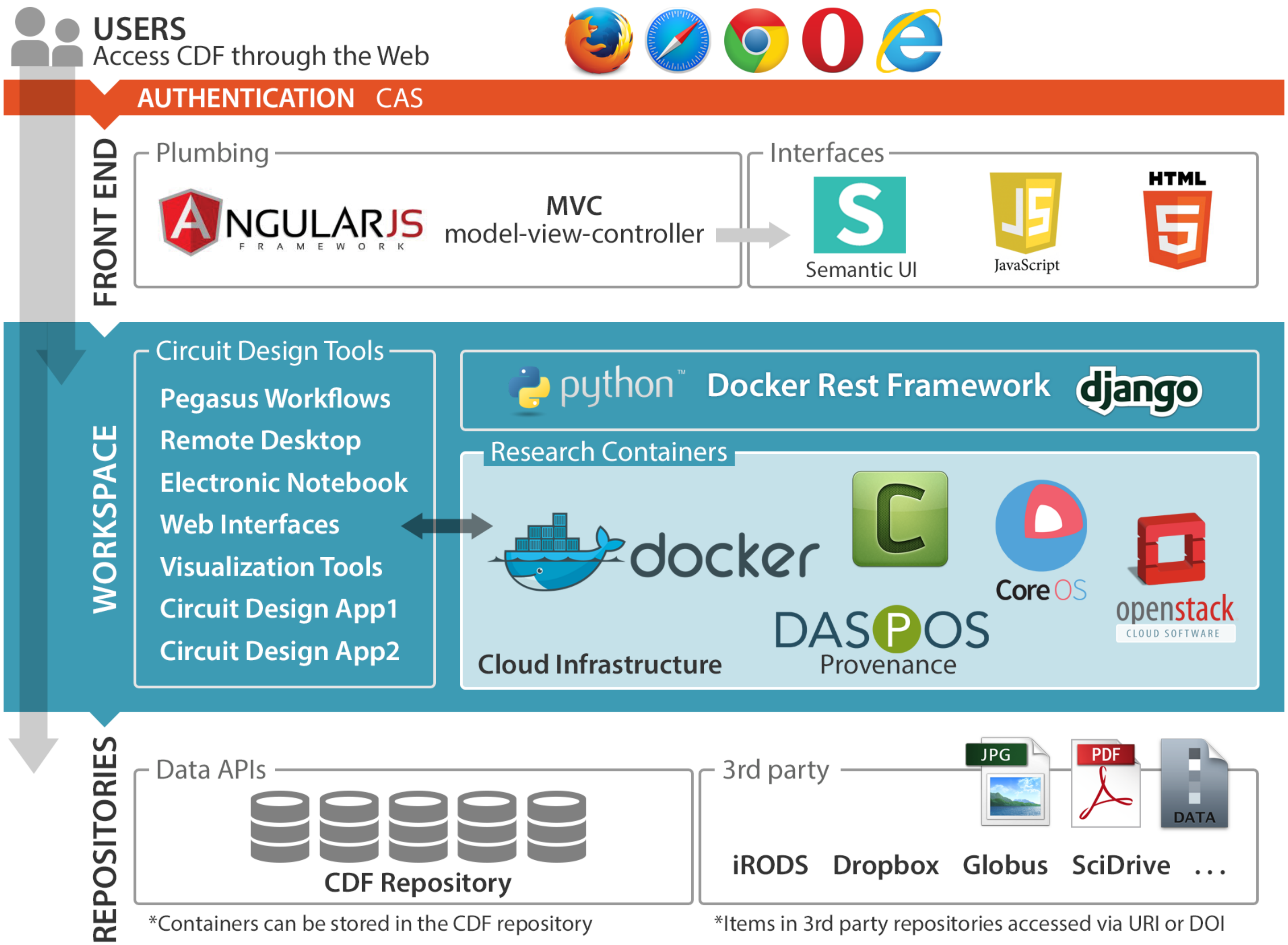
NDS Dashboard



And a video demonstration outlining some of the features can be seen here:

http://ndspilot.com/nds/ndspilot1080p.mp4

Circuit Design Trusted Repository



The Open Science Framework and SHARE

COS

CENTER FOR

OPEN SCIENCE

http://cos.io/top

Citation Standards

Data Transparency

Analytic Methods (Code) Transparency

Research Materials Transparency

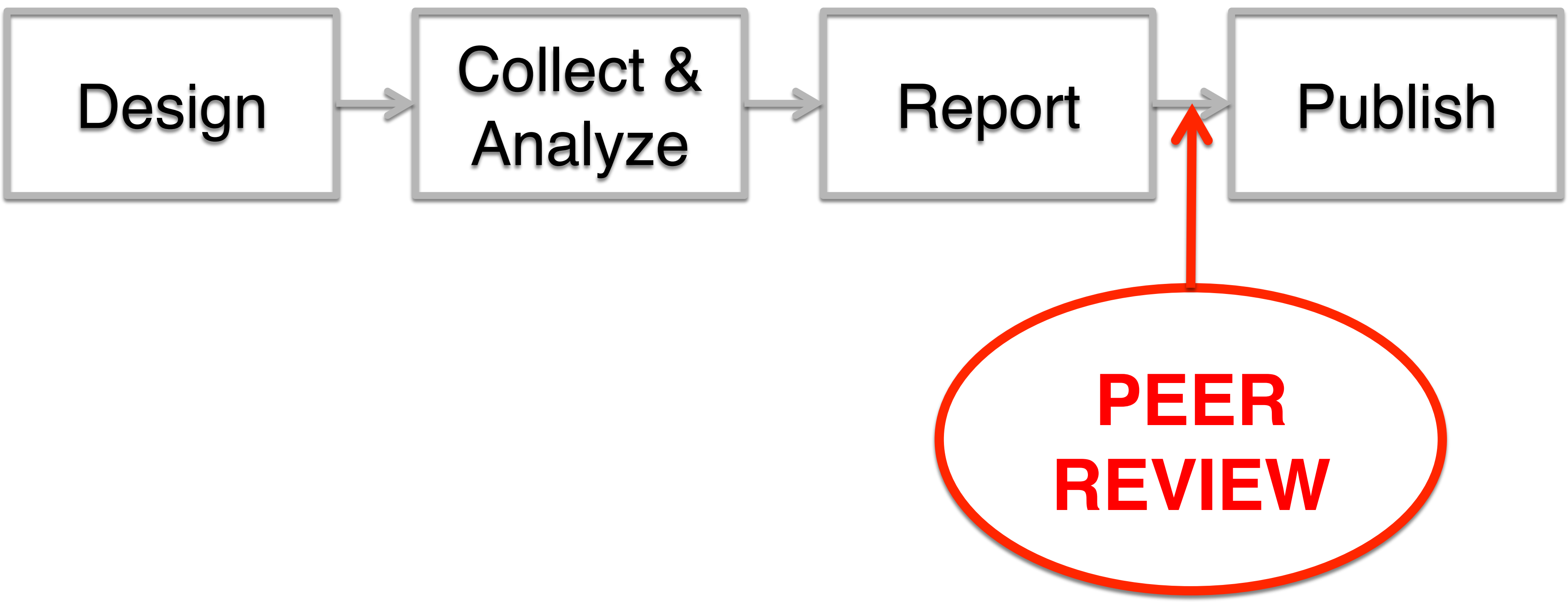
Design and Analysis Transparency

Preregistration of studies

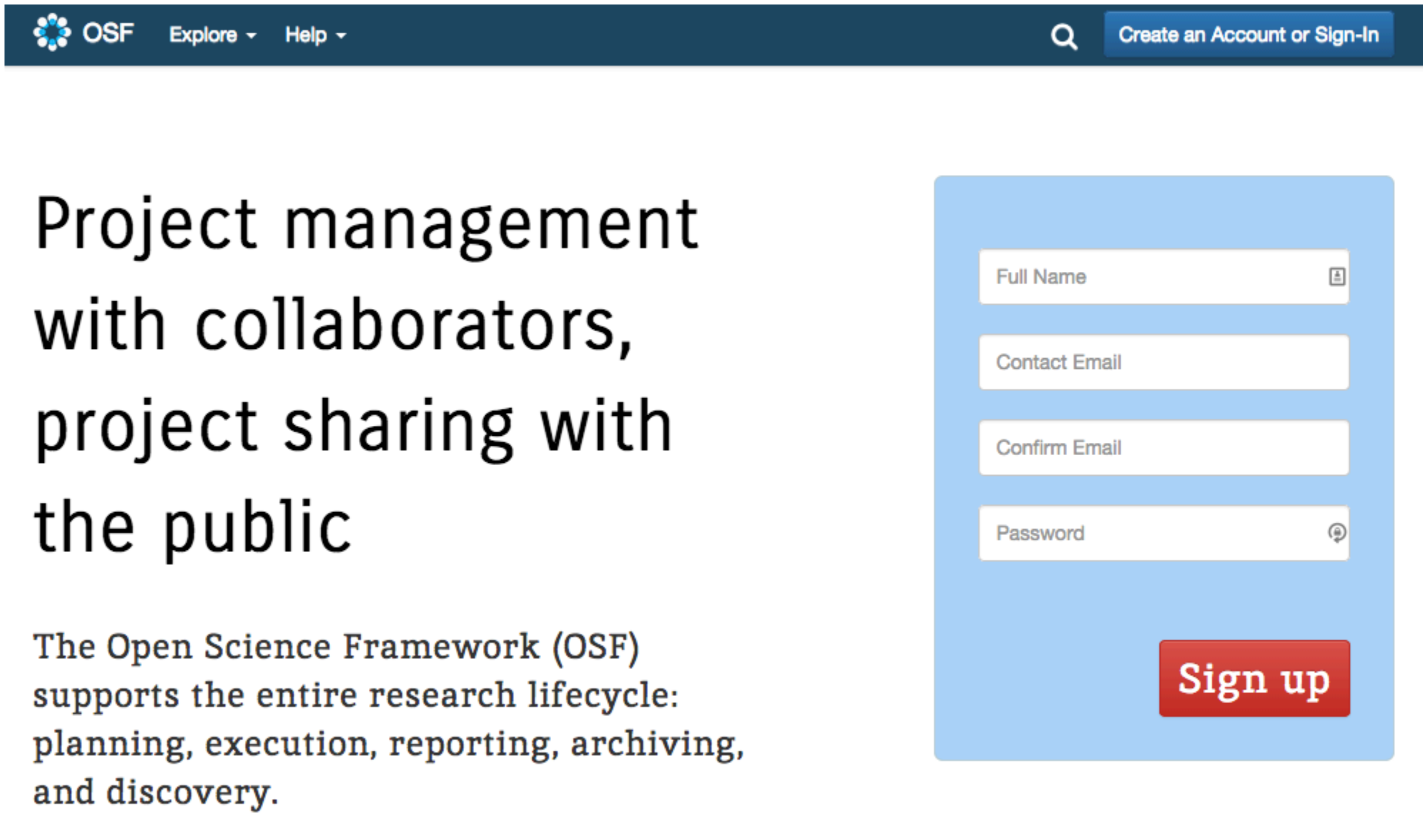
Preregistration of analysis plans

Replication

Registered Reports



Open Science Framework

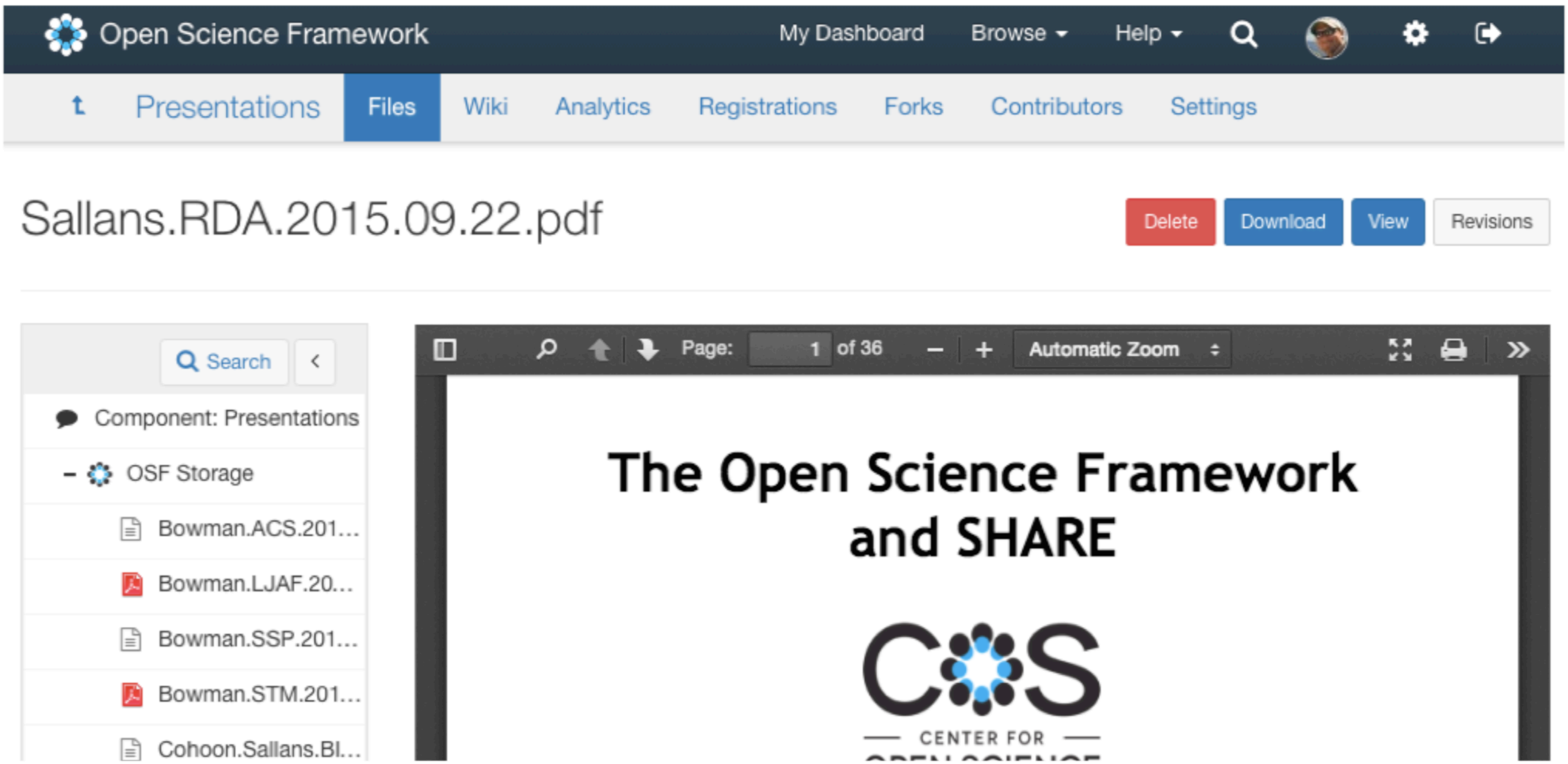


Sign-up now, easy and free!

http://osf.io

Find more on COS/OSF at

https://osf.io/habxc/



Questions: contact@cos.io

Final remarks

- Computer scientists, in collaboration with librarians and domain scientists, can do a lot together to support science integrity and open science efforts.
- Reproducibility is not about technology only.

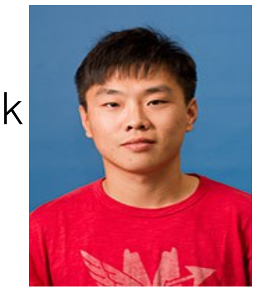
Acknowledgements

CENTER FOR OPEN SCIENCE

Graduate students

Haiyan Meng is leading work on Parrot and Umbrella.

Peter Ivie is leading the work on PRUNE.



Da Huo is leading work on Smart Containers

Collaborators, contributors to this talk

Prof. Doug Thain
CSE @ND
ccl.cse.nd.edu
(provided slides on Umbrella and PRUNE)

Dr. Charles Vardeman
crc.nd.edu
(provided slides on CS Open Linked Data)

Prof. Michael Hildreth
Physics@ND

DASPOS PI

Cooperative Computing Lab Douglas Thain

http://ccl.cse.nd.edu dthain@nd.edu

The Cooperative Computing Lab

Software | Download | Manuals | Papers

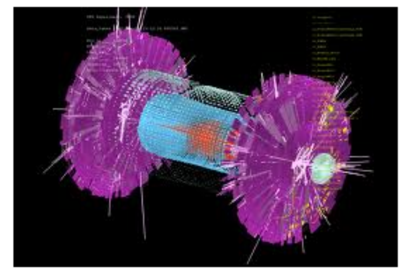
About the CCL

We design software that enables our collaborators to easily harness large scale distributed systems such as clusters, clouds, and grids. We perform fundamental computer science research in that enables new discoveries through computing in fields such as physics, chemistry, bioinformatics, biometrics, and data mining.

CCL News and Blog

- Creating Better Force Fields on Distributed GPUs with Work Queue
- CCTools 4.3 released
- Work Queue Powers Nanoreactor Simulations
- Open Sourcing Civil Engineering with a Virtual Wind Tunnel
- DeltaDB - A Scalable Database Design for Time-Varying Schema-Free Data
- Packaging Applications with Parrot 4.2.0
- CCTools 4.2.0 released
- DeltaDB at IEEE BigData 2014

Community Highlight



Scientists searching for the Higgs boson have profited from Parrot's new support for the CernVM Filesystem (CVMFS), a network filesystem tailored to providing world-wide access to software installations. By using Parrot, CVMFS, and additional components integrated by the Any Data, Anytime, Anywhere project, physicists working in the Compact Muon Solenoid experiment have been able to create a uniform computing environment across the Open Science Grid. Instead of maintaining large software installations at each participating institution, Parrot is used to provide access to a single highly-available CVMFS installation of the software from which files are downloaded as needed and aggressively cached for efficiency. A pilot project at the University of Wisconsin has demonstrated the feasibility of this approach by exporting excess compute jobs to run in the Open Science Grid, opportunistically harnessing 370,000 CPU-hours across 15 sites with seamless access to 400 gigabytes of software in the Wisconsin CVMFS repository.

- Dan Bradley, University of Wisconsin and the Open Science Grid

Research

- Papers
- Projects
- People
- Jobs
- REU

Software

- Download
- Manuals
- Makeflow
- Work Queue
- Parrot
- Chirp
- SAND
- AWE

Community

- Highlights
- Annual Meeting
- Workshops
- Getting Help
- Mailing List
- For Developers

Operations

- Condor Display
- Condor Pool
- Hadoop Cluster
- Biocompute
- BXGrid
- Condor Log Analyzer
- Internal

References

1. Abdalla, A., Hu, Y., Carral, D., Li, N., Janowicz, K.: An ontology design pattern for activity reasoning

2. Compton, M., Corsar, D., Taylor, K.: Sensor data provenance: Ssno and prov-o to- gether at last. In: To appear 7th International Semantic Sensor Networks Workshop (October 2014) (2014)

3. Gangemi, A.: Ontology design patterns for semantic web content. pp. 262–276. Springer (2005), http://link.springer.com/chapter/10.1007/11574620_21

4. Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman II, C.: Five stars of Linked Data vocabulary use. Semantic Web http://iospress.metapress.com/ index/053766UR810L7274.pdf

5. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Gar- ijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. W3C Recommendation 30 (2013)

6. Ma, X., Zheng, J.G., Goldstein, J.C., Zednik, S., Fu, L., Duggan, B., Aulenbach, S.M., West, P., Tilmes, C., Fox, P.: Ontology engineering in provenance enablement for the national climate assessment 61, 191–205, http:// linkinghub.elsevier. com/retrieve/pii/S1364815214002254

7. Oberle, D., Grimm, S., Staab, S.: An ontology for software. In: Handbook on on- tologies, pp. 383–402. Springer (2009)