REPRODUCIBLE WORKFLOWS: THE WAY FORWARD

John D. Chodera
MSKCC Computational and Systems Biology Program
http://www.choderalab.org

DISCLOSURES:
Scientific Advisory Board, Schrödinger
All opinions/views are my own.

SAMPL/D3R Workshop - 23 Feb 2018 - La Jolla, CA

COMPUTATIONAL CHEMISTRY IS FACING SIGNIFICANT CHALLENGES

INTEROPERABILITY

Current software communities are **balkanized**

Poor (or no) standards for moving data between codes/packages

If there *was* a good standard, developers would adhere to it

(where **good** = it made our lives **easier**, not harder)

EVALUATION

Comparison of predictive modeling on retrospective data hindered by **lack of standard datasets** and **absence of common benchmark framework**

Predictive challenges (e.g., SAMPL, D3R) often end up **testing unrelated choices** (such as biomolecular setup pipeline), not the scientific core code

BIOMOLECULAR SYSTEM PREPARATION REQUIRES MANY CHOICES

Before beginning, we have to make many decisions about structural data:

* Which structure(s) do we want to use? Often multiple

* What do we do about missing loops, termini, and residues?

* How do we treat modified residues? (phosphates, unnatural amino acids, PTMs)

* What do we do with cofactors? Keep or discard?

* What about crystallographic waters?

* How do we treat non-biological crystal contacts or domain swaps?

WHAT ARE WE EVALUATING IN BLIND COMPETITIONS?

evaluating the driver

evaluating the technology

Need to separate capabilities of technology from skill of driver

ENABLING FOCUS ON KEY SCIENCE

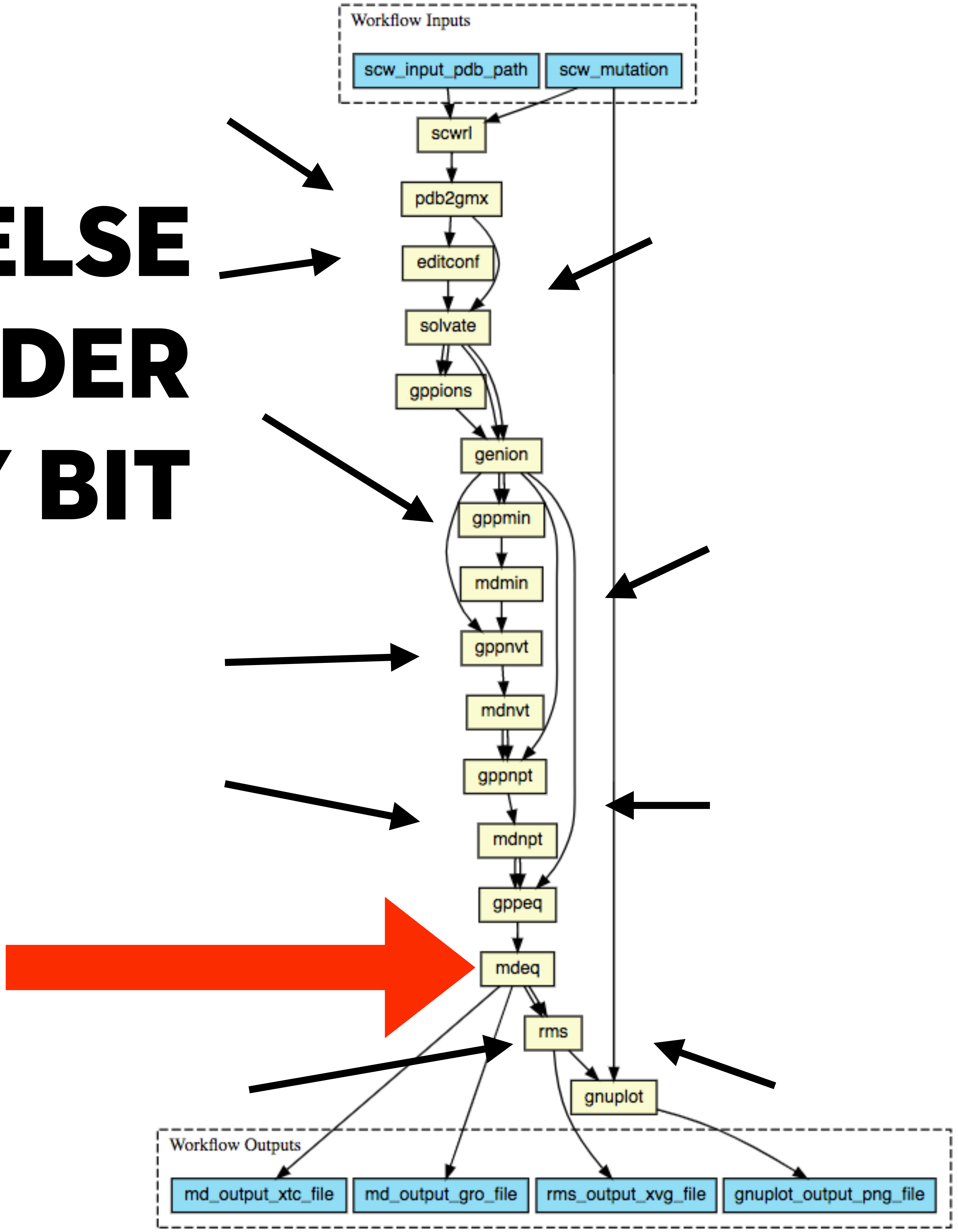
Academic scientists want to **focus creative efforts on a specific part of the process**, but are often forced to build everything from scratch to have a working framework in which they can carry out productive research

Industry wants to **combine best practices** from academia into useful pipelines for discovery, but has to hack everything together if they want to make this work

EXAMPLE: SETTING UP A FREE ENERGY CALCULATION IN GROMACS

EVERYTHING ELSE
I NEED IN ORDER
TO RUN MY BIT

THE SCIENCE
I'M INTERESTED IN



http://bioexcel.eu/

REPRODUCIBILITY

Reproducing work from a published computational chemistry paper is currently **nearly impossible**, which **minimizes opportunities for learning and improvement**

Translating best performers from D3R/SAMPL blind challenges into production pipelines is **nearly impossible** for the same reason

Example: SAMPL pKa methods

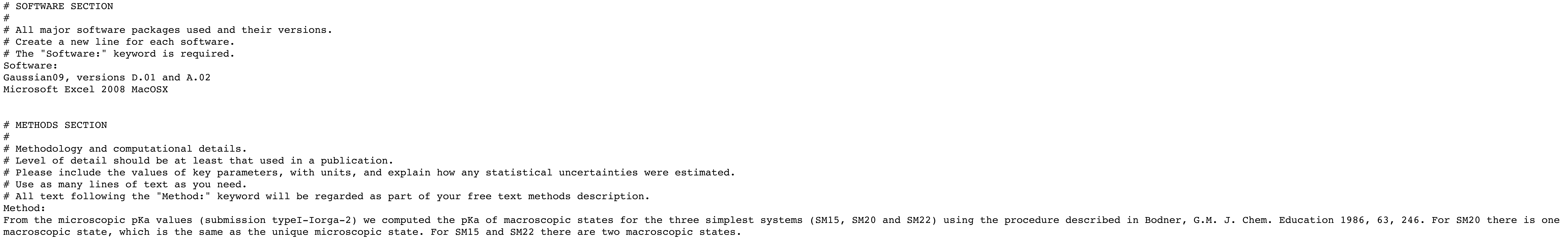
some are detailed:

```
# SOFTWARE SECTION
Software:
COSMOtherm C30_1701
Turbomole 7.2
COSMOconf 4.2
COSMOquick version 1.6
COSMOpy (version2017) & Python 2.7

# METHODS SECTION
#
Method:
The pKa dataset consists of 24 small to medium sized drug-like molecules which combine several functional groups whereas most of them have at least one basic functional group. Molecules SM01, SM08, SM15, SM20 and SM22 possess an additional (significant) acidic
functional group
Possible deprotonated and protonated species (anions, cations, zwitterions) have been generated automatically via the COSMOquick software package. A few further potential ions and tautomers were determined from visual inspection of the neutral forms as provided
for the challenge. In all cases, only single protonation or deprotonation turned out to be relevant at the experimental region from pH=2 to pH=12.
For all compounds, including the ionic and tautomeric forms, independent sets of relevant conformations were computed with the COSMOconf 4.2 workflow. Additional neutral conformers which are thermodynamically relevant in water according to COSMOtherm computations
have been found only for compound SM18 (tautomeric) and SM22 (zwitterionic) and have been included into the respective conformer sets used later on for the COSMOtherm pKa calculations.
The quantum chemistry calculations of COSMO sigma-surfaces were done at the BP//TZVPD//FINE single point level based upon BP//TZVP//COSMO optimized geometries to match the parameterization (BP-TZVPD-FINE-C30-1701) used in the 2017 COSMOtherm-release. All quantum
chemical calculations were carried out with the TURBOMOLE 7.2 quantum chemistry software.
The COSMOtherm pka-module uses a simple linear free energy relationship (LFER) in order to correct the free energy differences of the neutral and protonated (deprotonated) forms. ( Klamt, A. et al. J. Phys. Chem. A 107, 9380�9386 (2003). & Eckert et al. J Comp
Chem 27, 11�19 (2006).):
pKa = c0 + c1*(DG_neutral-DG_ionic)
with
c0=-131.7422 and c1=0.4910 mol/kcal (for acids in water)
c0=-171.1748 and c1=0.6227 mol/kcal (for bases in water)

pKa values were computed for all identified single protonated and deprotonated sampl6 molecules and the respective zwitterions using the COSMO-RS method as implemented in the COSMOtherm software. The workflow for the batch computation about 80 pKa reactions has
been automated via an in-house script based on Python 2.7 (COSMOpy).
For the final submission, only relevant pKa-values were included. For bases all protonation reactions with predicted pKa>0 and for acids all pKa values <14 were selected.
The pKa value of basic molecule SM14 containing 2 equivalent basic groups according to our calculations was corrected by the addition of log10(2).
The accuracy of the pKa prediction with the current COSMOtherm parameterization is about 0.65 log units root mean squared deviation (RMSD). The RMSD was evaluated on a validation set of about 160 basic and acidic compounds having a fairly simple molecular
structure. However, due to the somewhat more complex structure of the sampl6 molecules the mean of the expected error may be somewhat higher.
```

some are brief:



```
# SOFTWARE SECTION
#
# All major software packages used and their versions.
# Create a new line for each software.
# The "Software:" keyword is required.
Software:
Gaussian09, versions D.01 and A.02
Microsoft Excel 2008 MacOSX


# METHODS SECTION
#
# Methodology and computational details.
# Level of detail should be at least that used in a publication.
# Please include the values of key parameters, with units, and explain how any statistical uncertainties were estimated.
# Use as many lines of text as you need.
# All text following the "Method:" keyword will be regarded as part of your free text methods description.
Method:
From the microscopic pKa values (submission typeI-Iorga-2) we computed the pKa of macroscopic states for the three simplest systems (SM15, SM20 and SM22) using the procedure described in Bodner, G.M. J. Chem. Education 1986, 63, 246. For SM20 there is one
macroscopic state, which is the same as the unique microscopic state. For SM15 and SM22 there are two macroscopic states.
```

DEPLOYMENT

Translating academic research software into a tool that can be employed within industry is **extremely difficult** if not impossible for reasons of code quality, robustness, interoperability, and user-friendliness

Example from my own group: Merck KGaA pays us to fly a postdoc out once a quarter to do software updates and ensure code remains fully interoperable with their batch queue system, even though we try hard to make code conda-installable, use continuous integration, etc.

TRAINING

Pharma and comp chem are facing an exodus of talent due to wave of retirements

Need better tools to train the next generation of computational chemists
(which we're in also danger of losing to machine learning and data science)

FUNDING

Industry and federal funding agencies (NSF, NIH) tired of investing $ in software or research that is not useful to them or others

Easier to justify small investments in funding to deliver new features if they can be rapidly deployed and utilized/combined

VALIDATION AND ANALYSIS

For blind challenge participants, it's difficult to **validate** the output of your scripts to make sure it's in the right format, and to test on known datasets with the same analysis pipeline that will be used for assessment.

For blind challenge assessors, it's almost impossible to guarantee everyone will submit the data in the right format. (Sorry, Pat!)

WORKFLOWS TO THE RESCUE

Workflows (and the machinery to support them) can address many of these issues:
* Training
* Interoperability
* Reproducibility
* Evaluation
* Deployment
* Funding
* Enabling focus on key science
* Producivity

MolSSI

The Molecular Sciences Software Institute

… a nexus for science, education, and cooperation for the global computational molecular sciences community.

WHAT IS THE MOLSSI?

- New project (as 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.
- Part of the NSF's commitment to the White House's National Strategic Computing Initiative (NSCI).
- Total budget of $19.42M for five years, potentially renewable to ten years.
- Joint support from numerous NSF divisions: Advanced Cyberinfrastructure (ACI), Chemistry (CHE), and Division of Materials Research (DMR)
- Designed to serve and **enhance** the software development efforts of the broad field of computational molecular science.

Prof. T. Daniel Crawford

Director

crawdad@vt.edu

Prof. Robert J. Harrison

Co-Director for Parallel Computing and Emerging Technologies

Robert.Harrison@stonybrook.edu

Prof. Robert J. Harrison will oversee the Institute's

Prof. Shantenu Jha

Co-Director for Software Engineering Process, Middleware, and Infrastructure

shantenu.jha@rutgers.edu

Prof. Shantenu Jha will be responsible for the (i)

Prof. Vijay Pande

Co-Director for Molecular Simulation

pande@stanford.edu

Prof. Vijay Pande will be the primary liaisc
MolSSI and the biomolecular simulation/r

Prof. Cecilia Clementi

Co-Director for Molecular Simulation, and International Engagement

cecilia@rice.edu

Prof. Teresa Head-Gordon

Co-Director for Laboratory, Industrial, and Academic Outreach and Education

thg@berkeley.edu

Prof. Teresa Head-Gordon will lead MolSSI outreach

Prof. Anna Krylov

Co-Director for Quantum Chemistry and Materials

krylov@usc.edu

Prof. Anna Krylov will be the primary liaison to the quantum chemistry and soft materials community.

Prof. Theresa Windus

Co-Director for Code and Data Interoperability

twindus@iastate.edu

Prof. Theresa Windus will oversee the Institute's interoperability projects, an area in which she has long been a community leader. She has participate

MolSSI

MOLSSI SOFTWARE SCIENTISTS

- 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.

MOLSSI NEEDS BIOMOLECULAR SOFTWARE SCIENTISTS

MOLSSI IS SEEKING SOFTWARE SCIENTISTS IN THE BIOPHYSICAL DOMAIN

Qualified applicants must have a PhD in biophysics, chemistry, biology, materials science, applied mathematics, or related areas and experience in theoretical and computational methods for biophysical sciences.

Preferred Qualifications

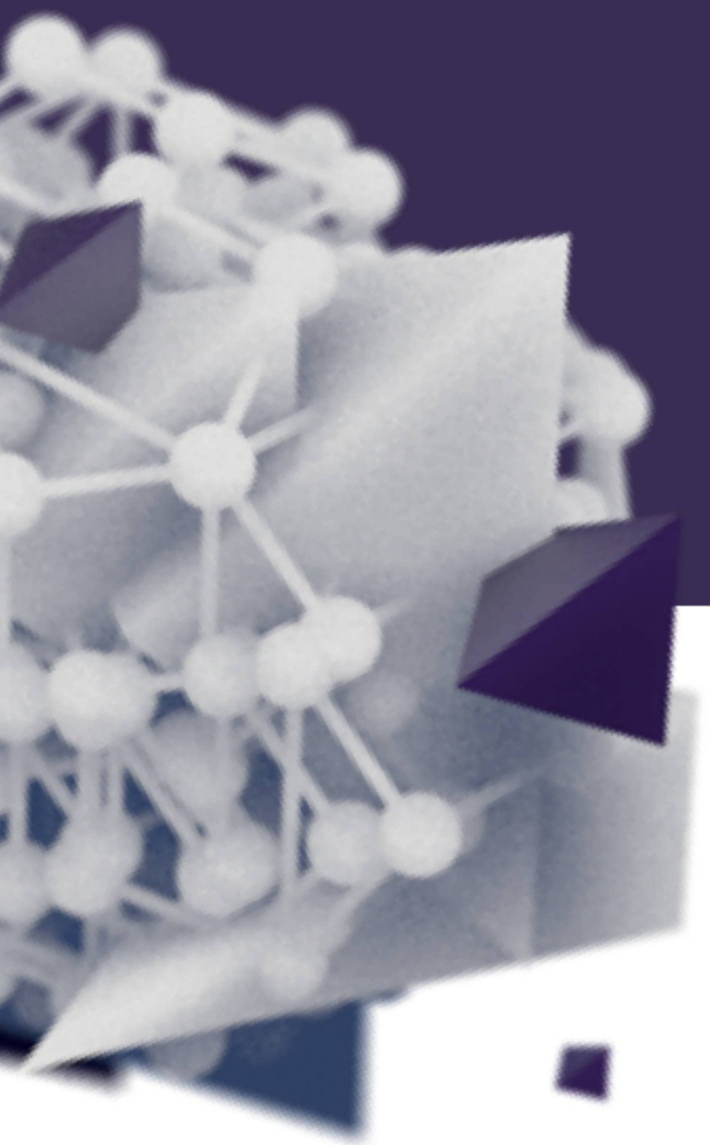
- Experience in successful software development activities such as bioinformatics, molecular dynamics and simulation, coarse graining, statistical mechanicsExperience in modern computational software development cycle methods;
- Experience with high performance computers and associated centers
- Ability to meet intermediate objectives towards the accomplishment of milestones for the advancement of concurrent projects
- Excellent publication record
- Excellent written and oral communication skills

Duties and Responsibilities of the Software Scientist Team as a Whole

- Develop software infrastructure and frameworks for community use and development
- Collaborate with scientists both within and without MolSSI to address the priorities of the community and MolSSI
- Provide expertise in design, optimization, verification, and documentation of software
- Provide forums for standards development and resource curation to the community
- Serve as mentors to Software Fellows by training them in software engineering best practices, API development, unit-testing, documentation, version control, performance profiling and other issues essential to community software development.
- Interact with partners in industry, NSF supercomputing centers, national laboratories, and international facilities to identify emerging hardware trends, software priorities and future career paths
- Lead and participate in outreach and educational activities, as well as developing instructional materials
- Author/co-author articles for publication and presentation in scientific journals Present MolSSI activities and research at professional and project meetings
- Ensure all relevant safety policies and procedures are followed and appropriate training is acquired and maintained
- Personal professional development activities

Applicants must submit their applications online at http://www.jobs.vt.edu and locate the posting for Staff Software Scientists (Posting SR0180022) under the Department of Chemistry. Applicants will submit a curriculum vita, a cover letter, and provide three references. The Search Committee Coordinator is available to address any specific questions related to the position: Professor Theresa Windus, Iowa State University, Department of Chemistry, 125 Spedding Hall Ames, IA 50011; twindus@iastate.edu.

http://molssi.org/2018/02/21/molssi-is-seeking-software-scientists-biophysics/



A MolSSI Workshop

DISTRIBUTED WORKFLOWS FOR BIOMOLECULAR SIMULATION

September 12-13, 2017 | Autodesk Gallery, 1 Market Street, San Francisco, CA

Distributed Workflows for Biomolecular Simulations is an invite-only, innovation-driven workshop hosted by MolSSI and Autodesk Life Sciences for academic and industry experts on how workflow technologies will vastly accelerate pipelines from academic research to industrial discovery.

PLEASE SAVE THE DATE, REGISTRATION LINK TO FOLLOW

AUTODESK LIFE SCIENCES

MolSSI THE MOLECULAR SCIENCES SOFTWARE INSTITUTE

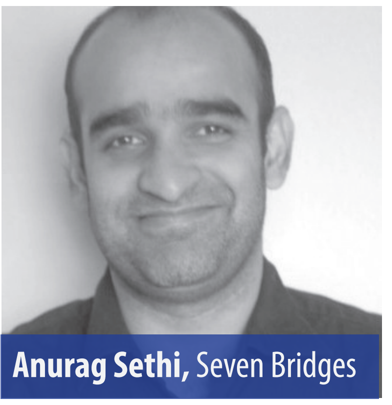
BACKGROUND

Workflow technologies simplify the processes of developing reliable computational methods, deploying reproducible and reliable software, exploiting scalable computing, and sharing standardized best practices. With increasing interest in such systems from academic, industrial, and computing groups, this two-day workshop will bring together a diverse group of experts to catalyze and develop modern workflow

WORKFLOWS TO THE RESCUE

Success stories from industries transformed by workflows

Anubhav Jain, LBNL



Anurag Sethi, Seven Bridges

Pharma industry needs for workflow engines

Brad Sherborne, Merck & Co.

Jeff Blaney, Genetech

Great workflow engines for computational chemistry are emerging now:

* OpenEye **Orion**

* Autodesk **Molecular Design Toolkit (MDT)**

* Schrödinger **LiveDesign**

Bob Tolbert, OpenEye

Aaron Virshup, Autodesk

Cecilia Cheng, Schrodinger

Cloud computing technologies that are eliminating computing constraints

* Google Life Sciences / Verily

* Amazon Web Services

Joe Corkery, Google

Jamie Kinney, Amazon

HOW CAN WE MAKE THE FUTURE BETTER THAN THE PAST?

What could **computational chemistry in 2020** look like?

Computational chemistry publications include a DOI-indexed workflow that can be pulled from a **common workflow registry** to reproduce the calculations in the paper.

Publications require virtual screening or affinity prediction tools to report performance on **standard benchmark datasets.**

Academics can focus their efforts on improving the science underlying **specific components** of versioned best practices workflows, and share them in a common **app store**.

Industry can easily **evaluate academic tools or workflows** on internal datasets without having to embark on a multi-year effort to reimplement, hack together, or harden the software.

Vendors could flexibly charge for use of their tools, potentially by **pay for privacy/ownership** so tools could be evaluated freely but funded by use for IP generation.

Don't miss **Alpha Shock**: Murcko and Walters, JCAMD 26:97, 2012.
http://doi.org/10.1007/s10822-011-9532-z

WHAT HAPPENS IF WE DO NOTHING?

We pay an enormous opportunity cost.

Stage 1: PROLIFERATION.
Many competing non-interoperable workflow engines emerge, remain balkanized.
Toolmakers must wrap their tools separately for each engine, wasting time.
Workflows must be tediously re-implemented in each engine.

Stage 2: METASTASIS.
One workflow engine dominates, leading to monoculture,
which is also not good for innovation.

MolSSI is here to catalyze change that would be otherwise difficult

OPPORTUNITIES

Workflow component interoperability:

- Components could be portable between workflow engines
  - Academics could **wrap tools once** to make them available to many systems
  - Software vendors could make components available via licensing models
  - Workflow engines could benefit from large **ecosystem** of components
- Common component format could be supported alongside specialized formats
- Enable a common "app store" or registry of components?
- We would need to define:
  - How components are **encapsulated**
  - What **information must be exchanged**
  - How components **expose their functionality**
  - Different **licensing models** that enable research, use, and fair compensation
  - How toolmakers can get **feedback** (especially regarding failures)

OPPORTUNITIES

Workflow definition interoperability:

- Workflows could be portable between workflow engines
  - Different workflow engines may be ideal for different hardware environments
- Common workflow format could be supported alongside specialized formats
- Workflows could implement **versioned best practices** (LiveCoMS)
- Enable a **common registry of workflows**?
  - Computational chemistry papers could contain workflow references to reproduce calculations performed in paper
  - Workflows could be evaluated retrospectively on common benchmark datasets or prospectively on blinded datasets
- Would also require interoperable workflow components

FOCUS WORKFLOW GROUPS

Free energy calculations: Michael Shirts

Michael Shirts, UC Boulder

Molecular dynamics simulation: Pek Leong & Paul Saxe

Pek Leong, UCSD / NBCR

Paul Saxe, MolSSI

Biomolecular complex setup pipeline: David Mobley

David Mobley, UC Irvine

Docking, scoring, and quantitative affinity prediction blind assessment: Jeffrey Wagner & Ajay Jain

Daniel Smith, D3R/UCSD

Ajay Jain, UCSF

WHAT ARE THE INCENTIVES?

- To **workflow engine developers?**
  - Access to many more components / workflows without needing to wrap tools
  - Continual supply of updated versions of components
- To **tool developers?**
  - Large user base (via multiple workflow engines)
  - Don't need to directly support users
  - Academics can focus on science, software vendors on their strengths
- To **industry?**
  - Rapid translation of new science from academia or vendors to pharma
  - Facile benchmarking of new technologies
- To **infrastructure providers**
  - Better scalability of tools; greater utilization of resources
- Makes lives of all stakeholders better

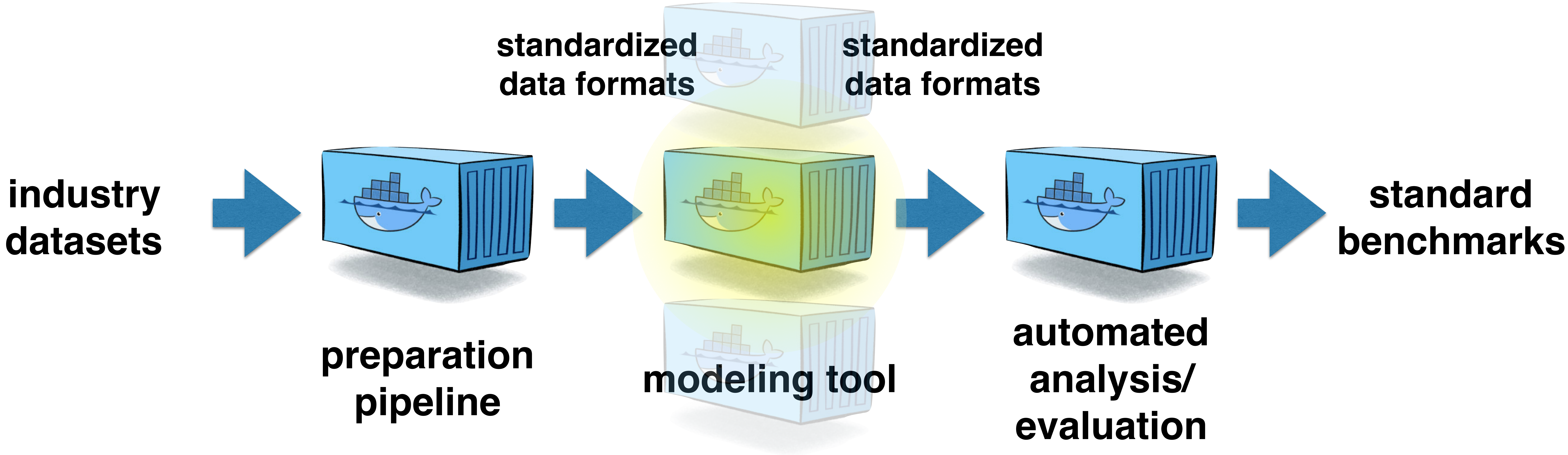
WHAT ARE WE EVALUATING IN BLIND COMPETITIONS?

evaluating the driver

evaluating the technology

Need to separate capabilities of technology from skill of driver

WORKFLOWS USING BEST PRACTICES WOULD ALLOW US TO EVALUATE THE TECHNOLOGY


industry datasets

preparation pipeline

standardized data formats

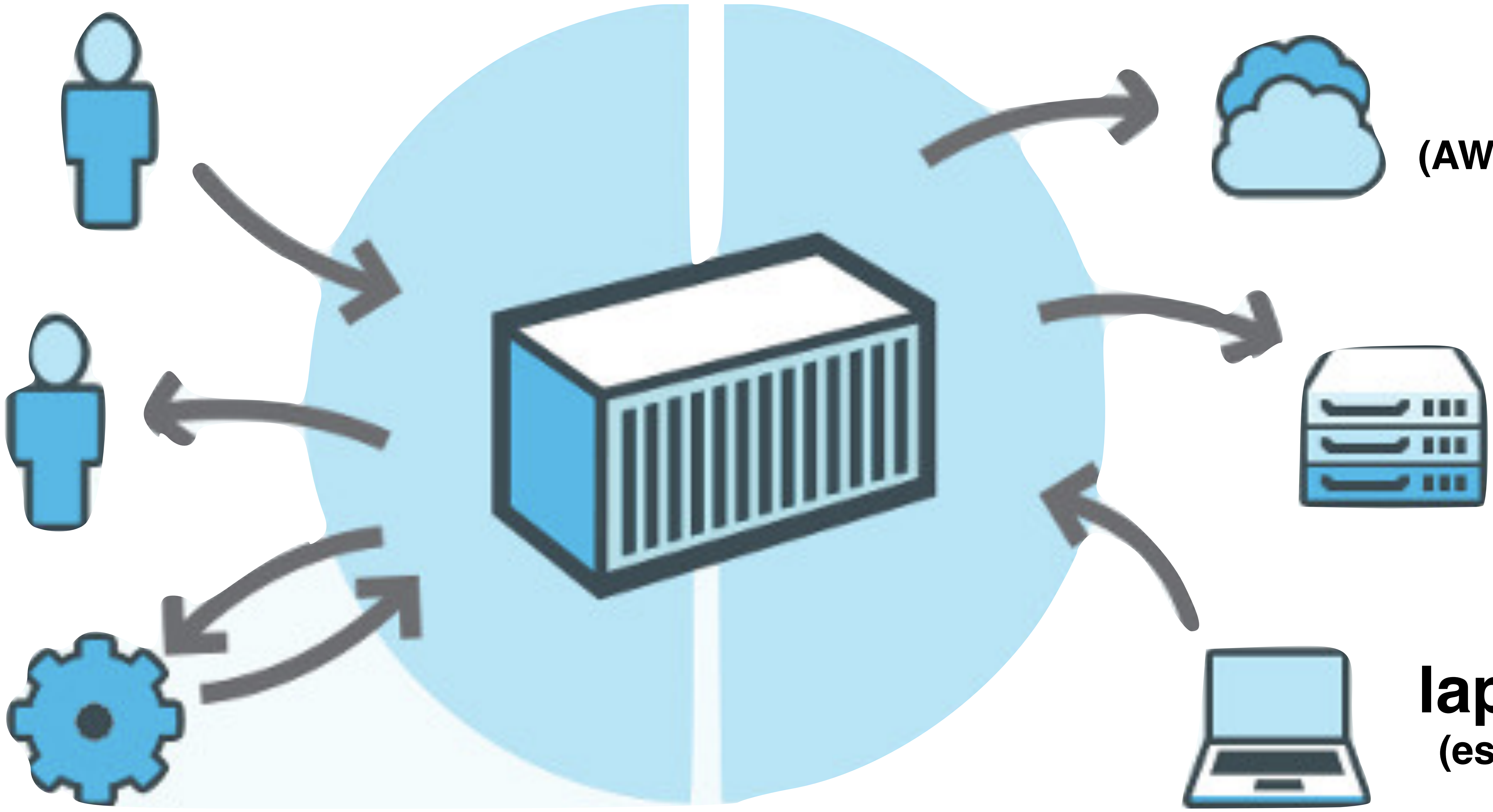
modeling tool

standardized data formats

automated analysis/ evaluation

standard benchmarks

CONTAINERS SOLVE THE PORTABILITY PROBLEM



interactive terminal/GUI sessions

standardized programmatic interfaces

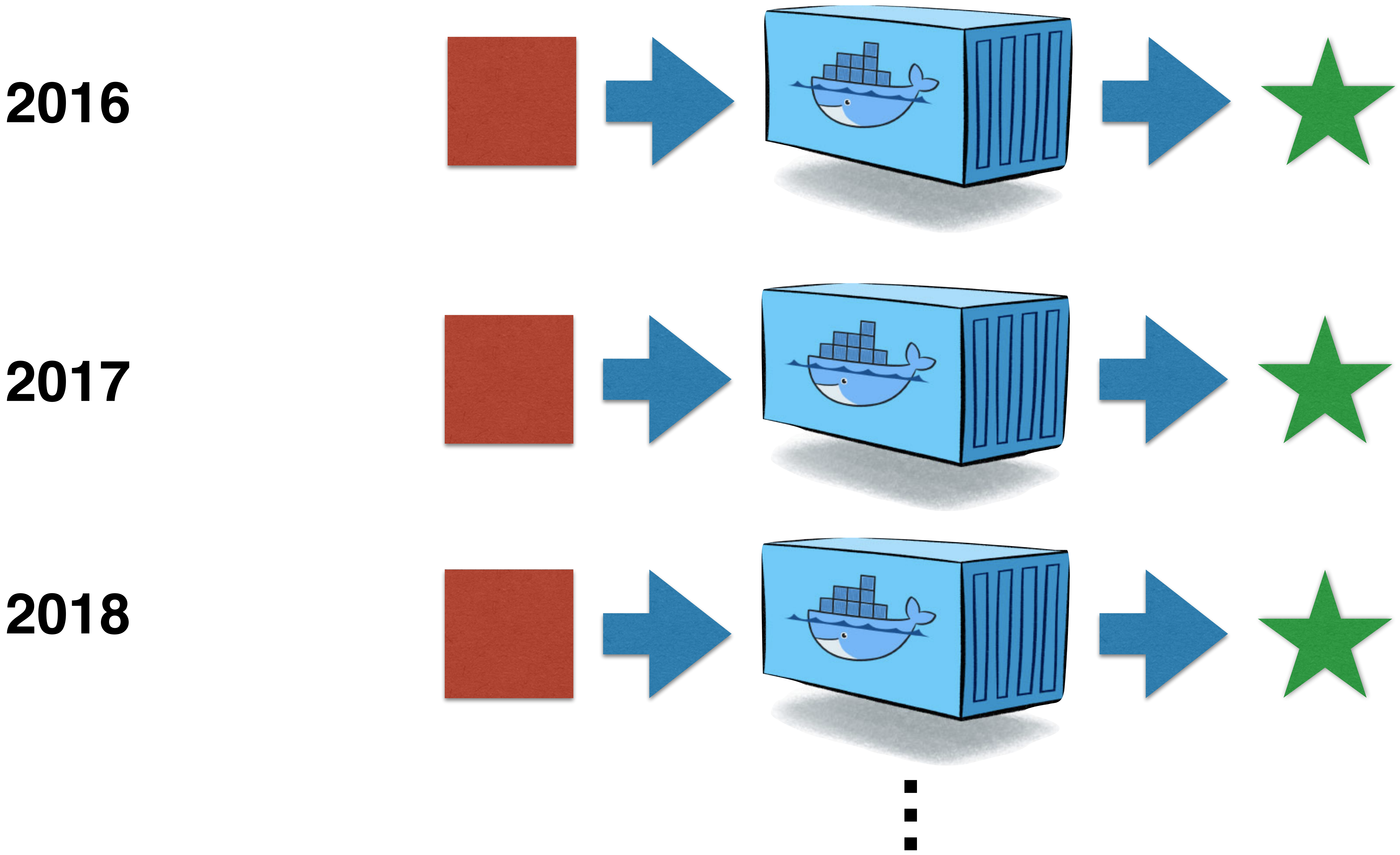
cloud
(AWS, Google Compute)

local resources
(OS independent)

laptop/desktop
(essential for training)

docker

CONTAINERS SOLVE THE REPRODUCIBILITY PROBLEM

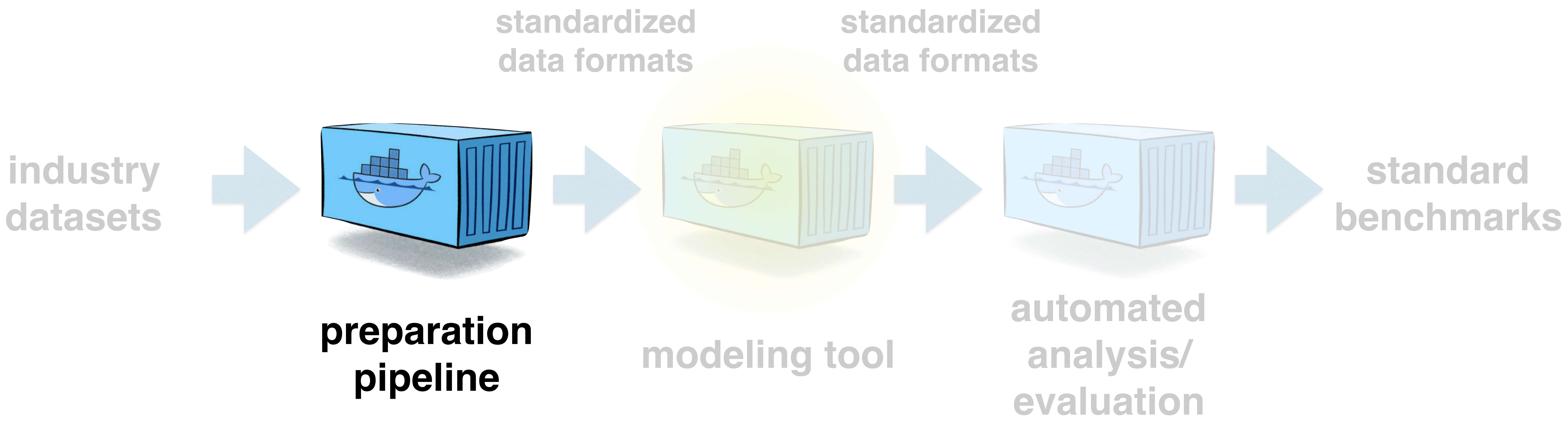


2016

2017

2018

OPEN PREPARATION PIPELINES COULD CAPTURE COMMUNITY-DRIVEN BEST PRACTICES



industry datasets

standardized data formats

standardized data formats

standard benchmarks

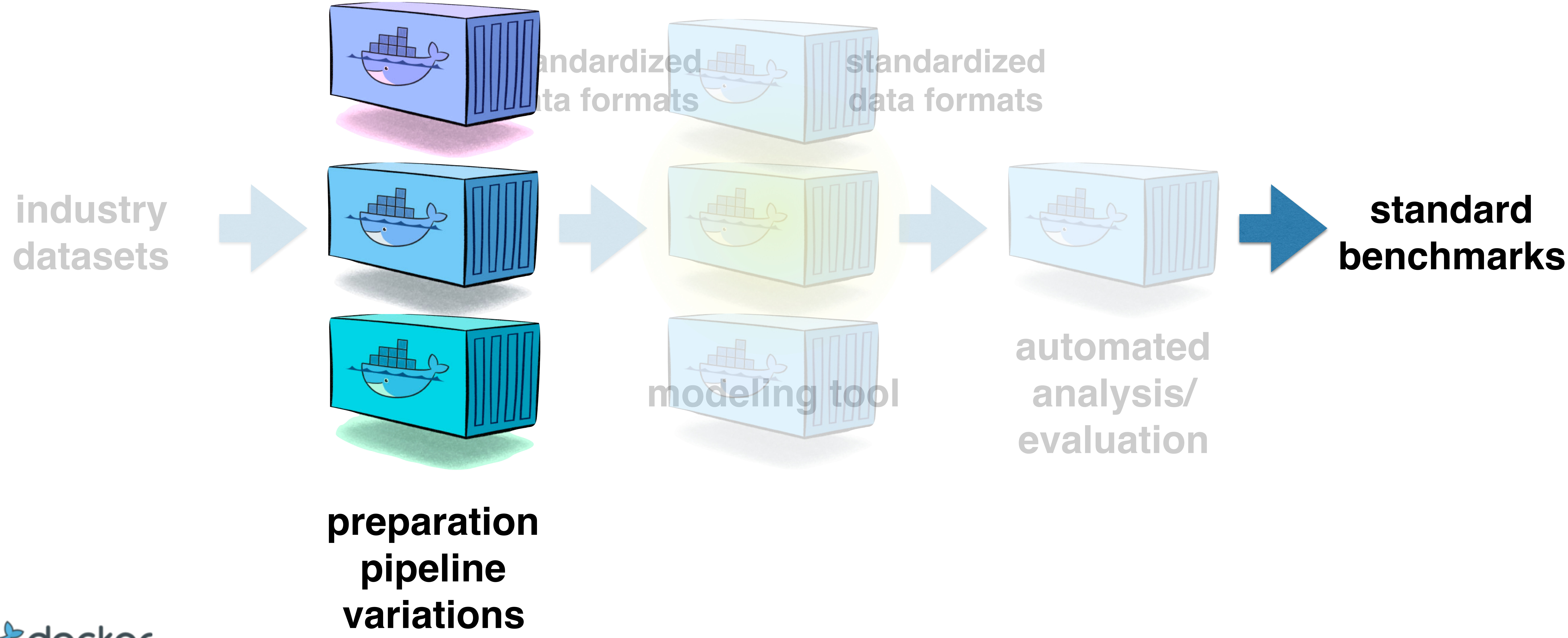
preparation pipeline

modeling tool

automated analysis/ evaluation

docker

BEST PRACTICES CAN BE EVALUATED BY TESTING VARIATIONS ON A VARIETY OF MODELING TOOLS



industry datasets

standardized data formats

standardized data formats

modeling tool

automated analysis/ evaluation

standard benchmarks

preparation pipeline variations

THIS REQUIRES STANDARDIZED DATA INTERCHANGE FORMATS

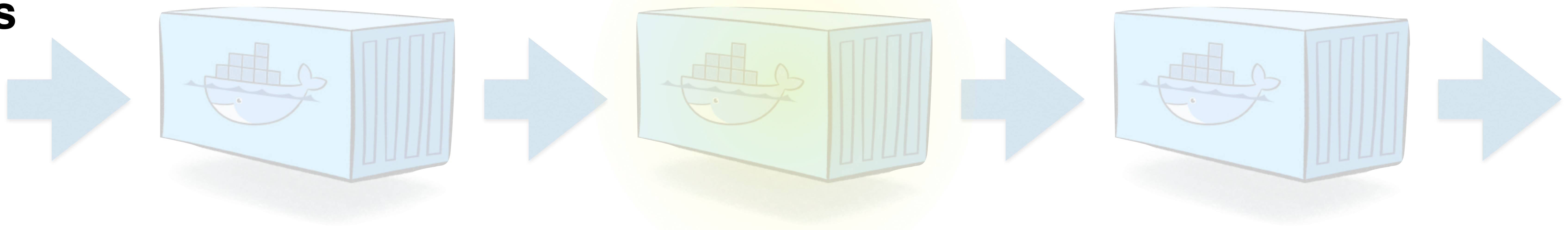
standardized data formats

standardized data formats

protein constructs

assay conditions

molecules



industry datasets

preparation pipeline

modeling tool

automated analysis/ evaluation

standard benchmarks

biomolecular target
replace aging PDB format
handle charges, parameters, etc.
robust open source readers/writers

parameterized small molecules
make up for shortcomings in mol2, SDF
suitable for the internet age (e.g. JSON)

prediction formats
binding poses
predicted affinity/assay data
predict confidence/uncertainties
exception logging

assessment formats
standard representations
standard assessments
standardized uncertainty analysis

docker

WHAT WOULD DO WE NEED TO DO?

Articulate workflows, workflow components, and tools of interest

Determine what kinds of data they consume/emit

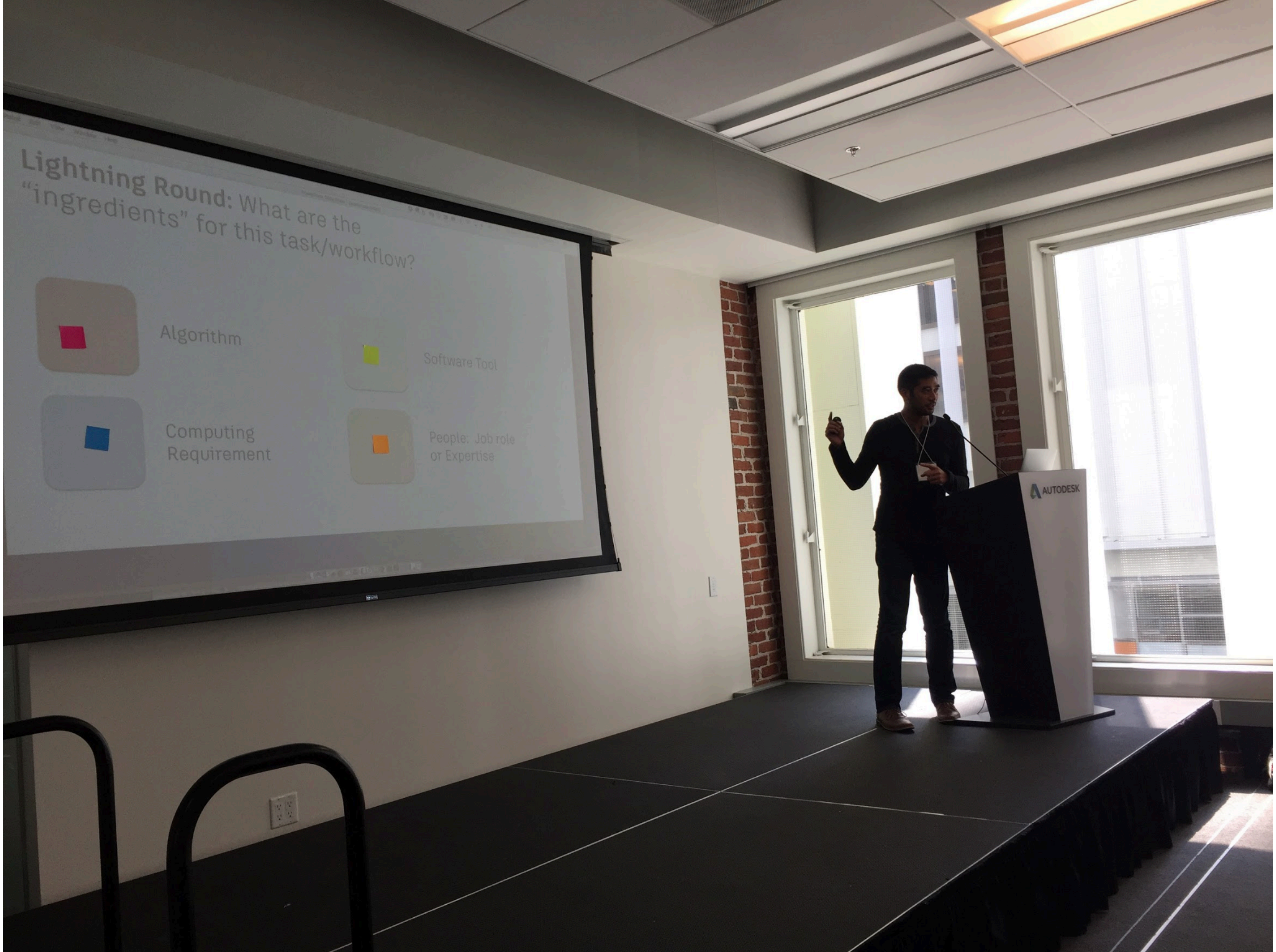
Identify what, if any, new standards, formats, or APIs are needed

Create working groups to establish standards for building interoperable components/workflows

urrent Medical and Pharmaceutical Practices are Insufficient and Costly
40-60% of medicines do not work as described
$610B/YR in pharmacy spending
350M people have a rare disease (globally)
7 YRS to diagnosis $300B for diagnosis
Cancer is a genetic disease
Oncogenes
$150B/YR for cancer care
SevenBridges

Our System

The most practical, flexible and versatile approach to innovation in the world, that anyone can learn and apply.



Lightning Round: What are the "ingredients" for this task/workflow?

- Algorithm
- Software Tool
- Computing Requirement
- People: Job role or Expertise

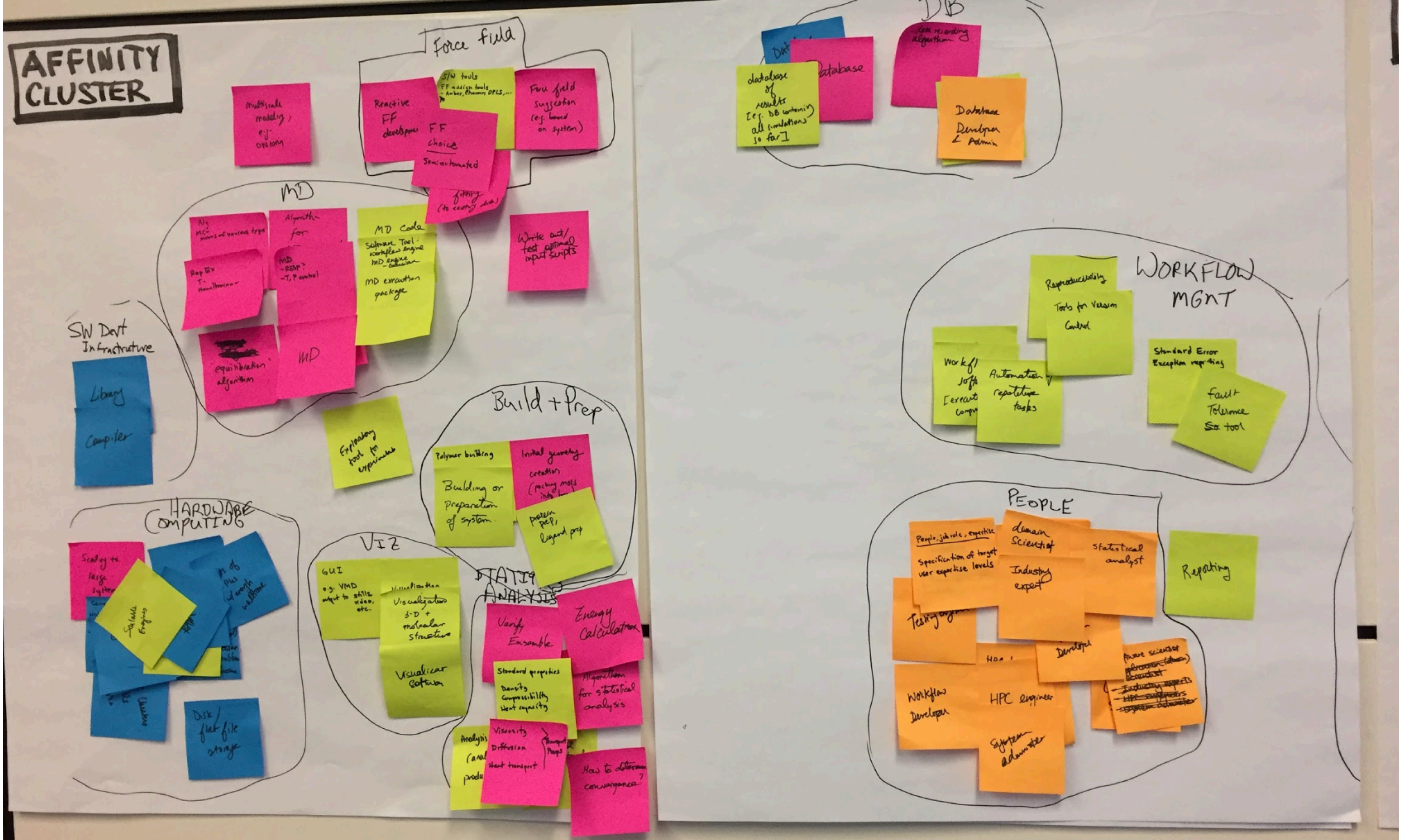
AFFINITY CLUSTER

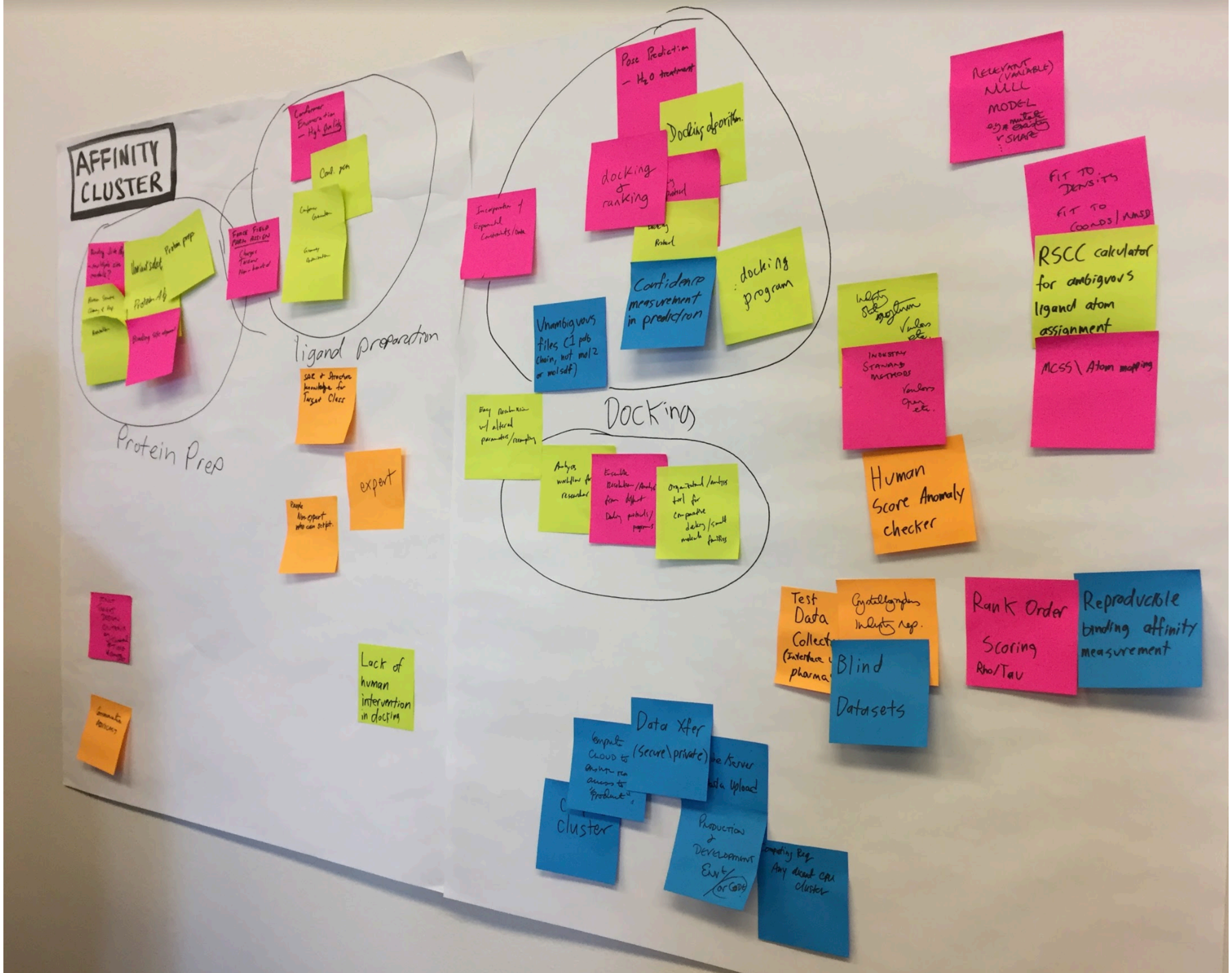
AFFINITY CLUSTER
WET STANDPI
AFFINITY CLUSTER

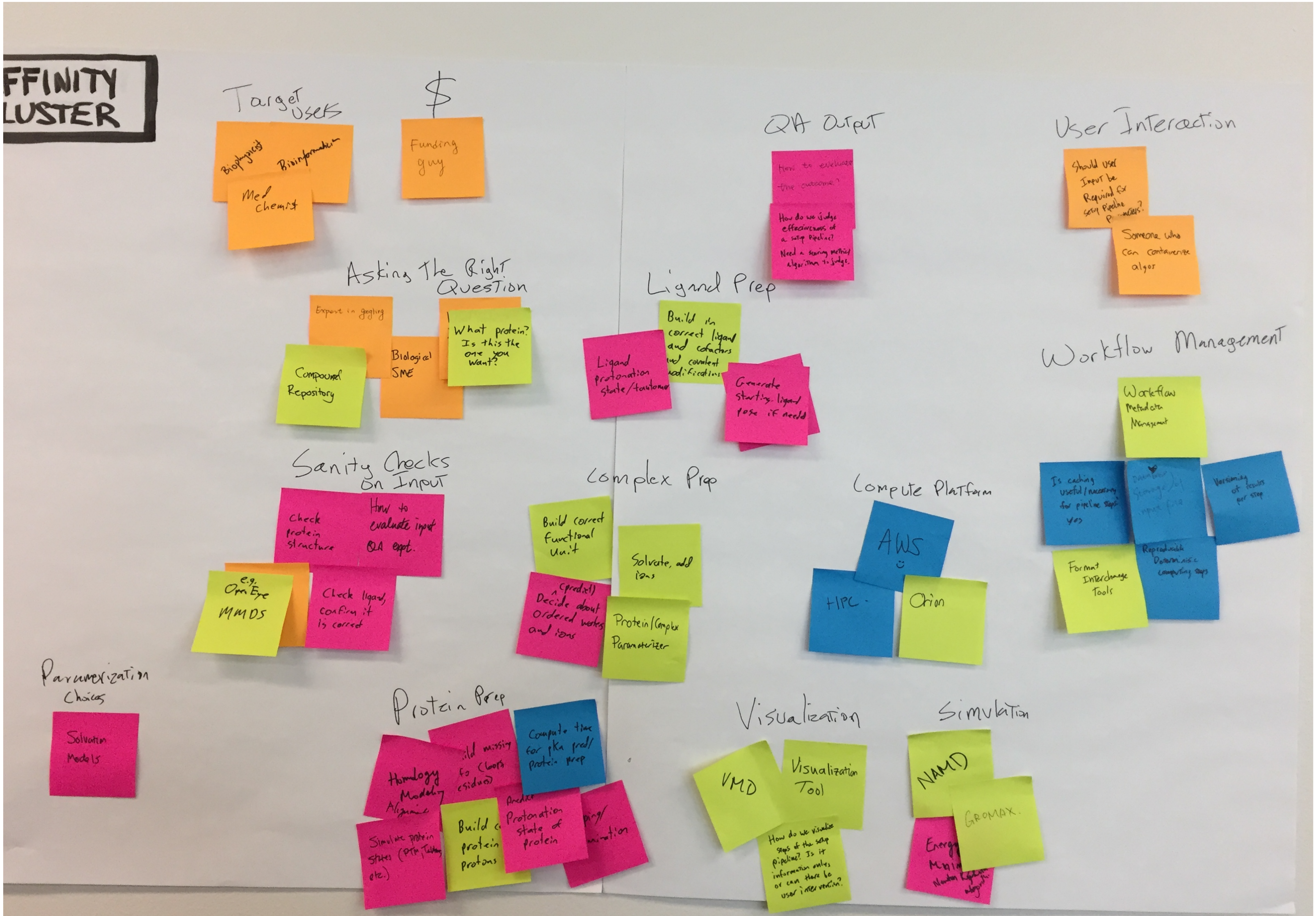
SCHEMATIC DIAGRAM
AFFINITY CLUSTER
SCHEMATIC DIAGRAM
FREE

AFFINITY CLUSTER
SCHEMATIC DIAGRAM

AFFINITY CLUSTER


AFFINITY CLUSTER
Multiscale modeling, e.g. ONIOM
Force field
Reactive FF development
S/W tools
FF assign tools
- Amber, Charmm, OPLS, ...
Force field suggestion (e.g. based on system)
FF choice
Semiautomated
fitting (to existing data)
MD
Alg
MC-
moves of various types
Algorithm for
Rep Ex
T-
Hamiltonian
MD
-RESP?
-T, P control
equilibration algorithm
MD
MD Code
Software Tool:
workflow engine
MD engine
- Calculation
MD execution package
Write out/test optimal input scripts
SW Devt Infrastructure
Library
Compiler
Exploratory tool for experimentals
Build + Prep
Polymer building
Building or Preparation of system
Initial geometry creation (packing mols into box)
protein prep, ligand prep
HARDWARE COMPUTING
Scaling to large systems
Scalable Engines
Clusters
Disk/flat file storage
VIZ
GUI
e.g. VMD
output to stills, video, etc.
Visualization
Visualization 3-D + molecular structure
Visualizer Software
STATISTICAL ANALYSIS
Verify Ensemble
Energy Calculations
Standard properties
Density
Compressibility
Heat capacity
Algorithm for statistical analysis
Viscosity
Diffusion
Heat transport
Transport Props
How to determine convergence?
DB
Database
database of results [e.g. DB containing all simulations so far]
data recording algorithm
Database Developer & Admin
WORKFLOW MGMT
Reproducibility
Tools for Version Control
Automation of repetitive tasks
Standard Error Exception reporting
Fault Tolerance ~~Sw~~ tool
PEOPLE
People, job role, expertise
Specification of target user expertise levels
domain Scientist
statistical analyst
Industry expert
Testing engineer
Developer
Workflow Developer
HPC engineer
System administrator
software scientist
~~Industry experts~~
~~HPC engineers~~
~~System administrator~~
Reporting


AFFINITY CLUSTER
Protein prep
Variant select
Protein Alf
Protein Prep
Force Field Param Assign
Charges
Torsion
Non-bonded
Conformer Enumeration – High Quality
Conf. gen
Conformer Generation
Geometry Optimization
ligand preparation
SAR & Structure knowledge for Target Class
expert
People
Non-expert who can script.
Lack of human intervention in docking
Pose Prediction – H_2O treatment
Docking algorithm.
docking & ranking
Docking Protocol
Confidence measurement in prediction
docking program
Incorporation of Experimental Constraints/data
Unambiguous files (1 pdb chain, not mol2 or molsdf)
Docking
Easy Resubmission w/ altered parameters/resampling
Analysis workflow for researcher
Ensemble Resolution/Analysis from different Docking protocols/programs
Organizational/analysis tool for comparative docking/small molecule families
RELEVANT (VARIABLE) NULL MODEL
FIT TO DENSITY
FIT TO COORDS/RMSD
RSCC calculator for ambiguous ligand atom assignment
MCSS\Atom mapping
INDUSTRY STANDARD METHODS
Vendors
Open
etc.
Human Score Anomaly checker
Test Data Collect
(Interface w/ pharma
Crystallographers
Industry rep.
Blind Datasets
Rank Order Scoring
Rho/Tau
Reproducible binding affinity measurement
Data Xfer (Secure\private)
Compute CLOUD to ensure real access to "Product".
cluster
Data Upload
PRODUCTION & DEVELOPMENT Env't
or CODE
Computing Req
Any decent CPU cluster



FFINITY
LUSTER

Target users

Biophysicist

Bioinformatician

Med chemist

$

Funding guy

Asking The Right Question

Expert in googling

Compound Repository

Biological SME

What protein? Is this the one you want?

Sanity Checks on Input

Check protein structure

How to evaluate input QA expt.

e.g. OpenEye MMDS

Check ligand, confirm it is correct

Parameterization Choices

Solvation Models

Protein Prep

Homology Modeling Alignment

ild missing fo (loops esidues)

Compute time for pKa pred/ protein prep

Simulate protein states (PTM, Tautomers etc.)

Build c protein protons

Predict Protonation state of protein

ping/ ination

Ligand Prep

Ligand protonation state/tautomer

Build in correct ligand and cofactors and covalent modifications

Generate starting ligand pose if needed

Complex Prep

Build correct functional unit

Solvate, add ions

(predict) Decide about ordered waters and ions

Protein/Complex Parameterizer

QA Output

How to evaluate the outcome?

How do we judge effectiveness of a setup pipeline? Need a scoring metric/ algorithm to judge.

User Interaction

Should user Input be Required for setup pipeline parameters?

Someone who can containerize algos

Workflow Management

Workflow Metadata Management

Is caching useful/necessary for pipeline steps? yes

Database Storage of input file

Versioning of results per step

Format Interchange Tools

Reproducible Deterministic computing steps

Compute Platform

AWS ☺

HPC.

Orion

Visualization

VMD

Visualization Tool

How do we visualize steps of the setup pipeline? Is it information only, or can there be user intervention?

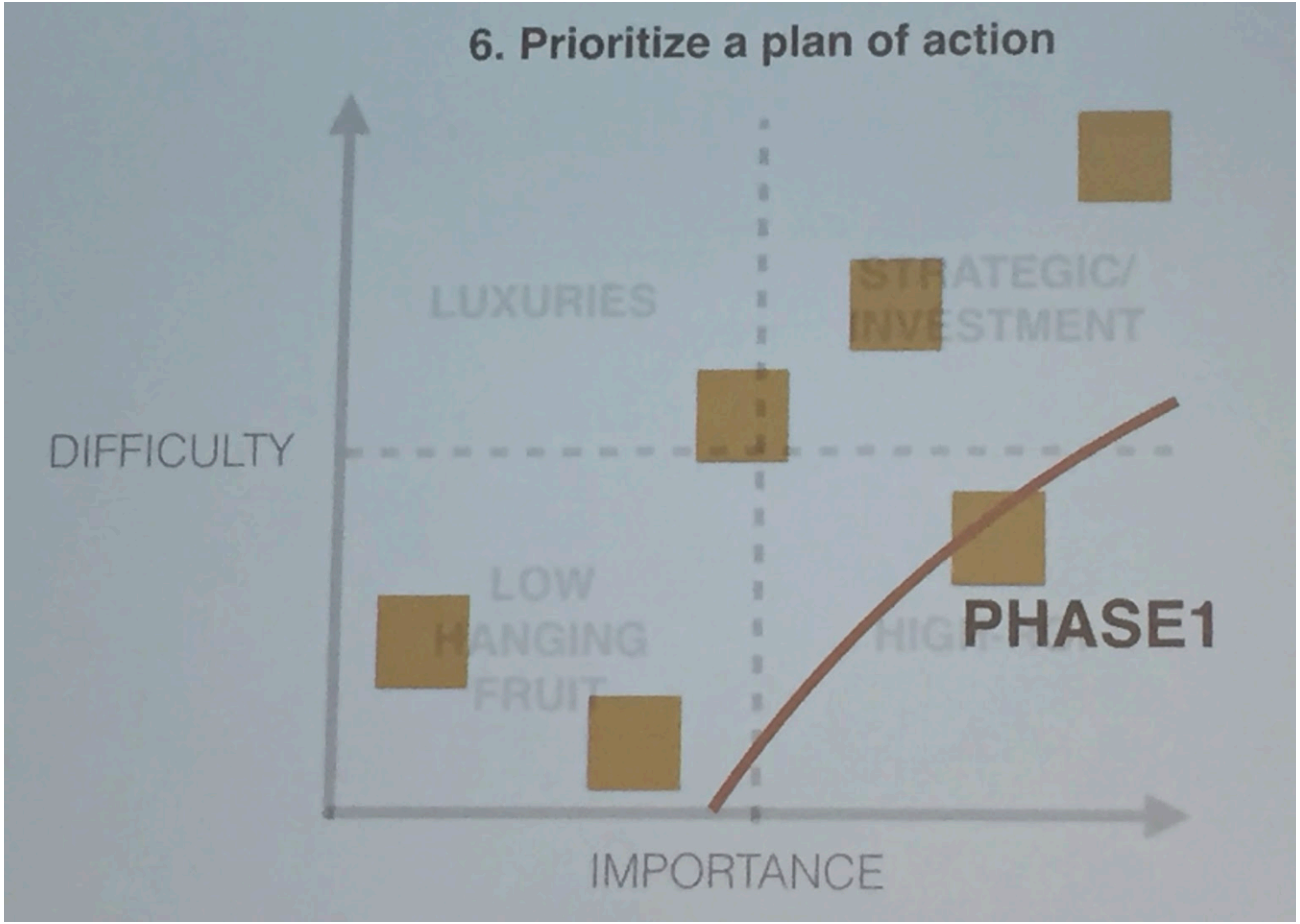
Simulation

NAMD

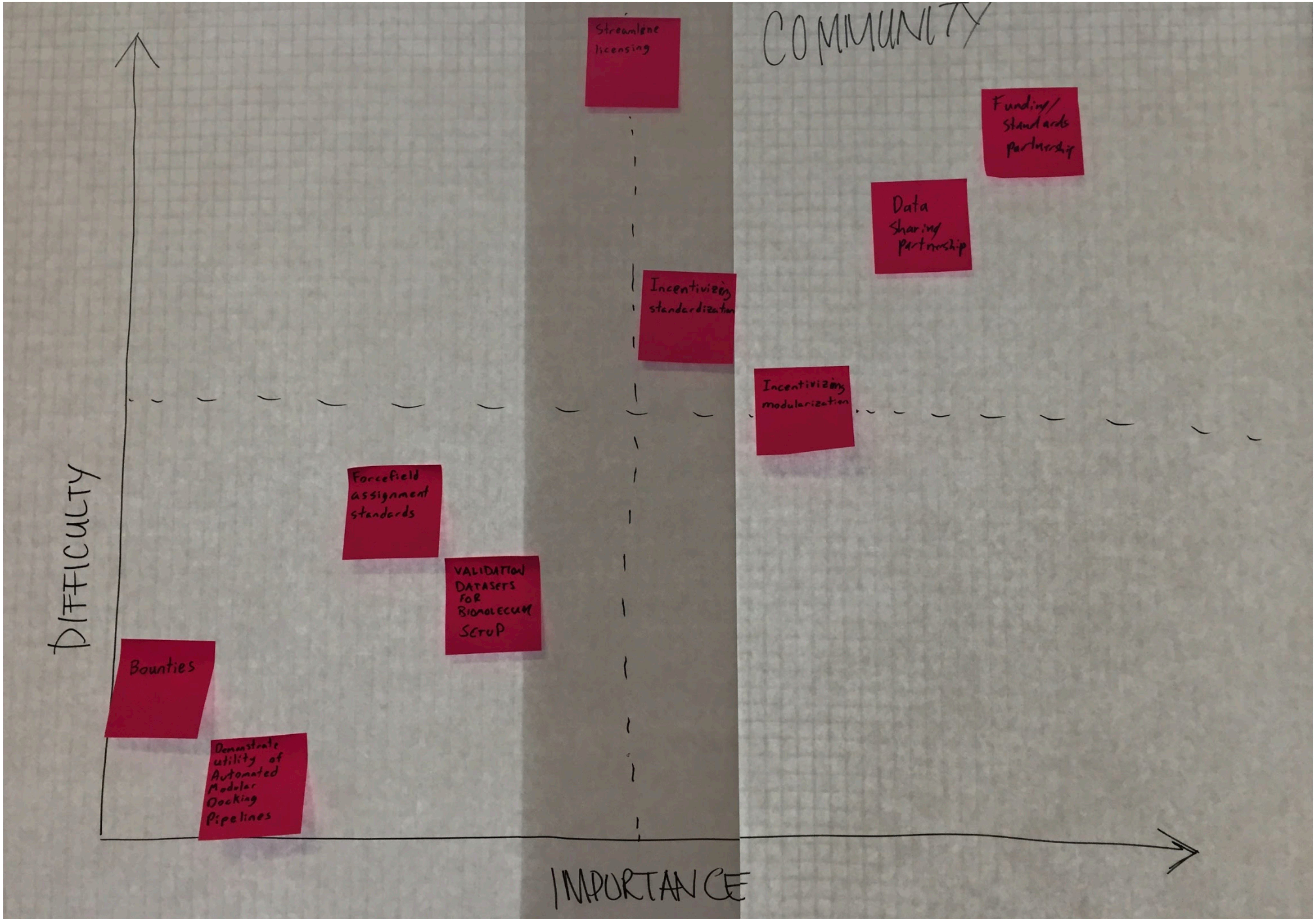
GROMAX.

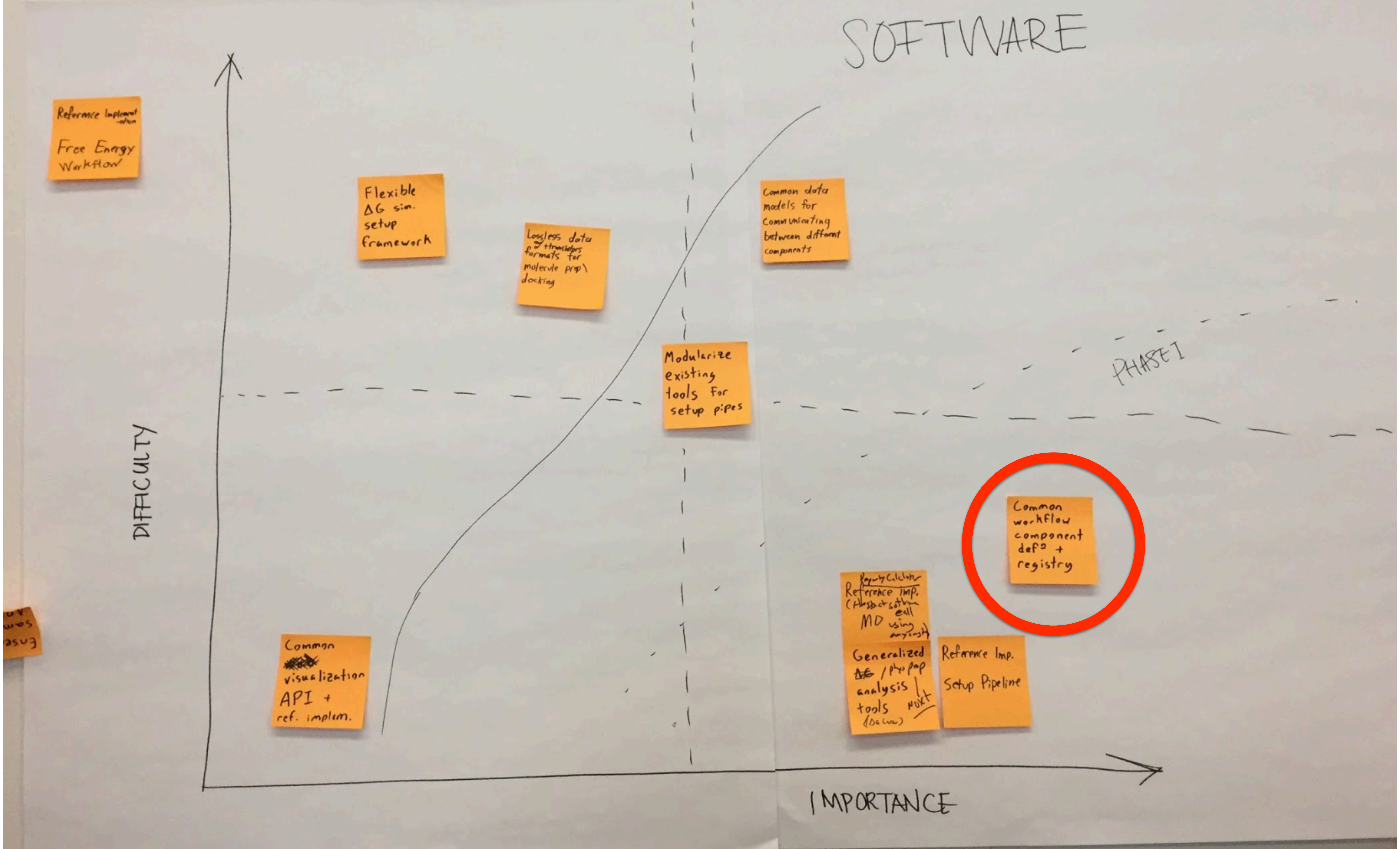
Energy Minim Newton Raphson algorith.

FREE ENERGY CALCULATIONS
AFFINITY CLUSTER
Experimentalist
Biomolecular Setup
Experienced User (comp chemist)
System setup (placing coordinates) of molecules
Desktop
PREPARING THE MODEL
FF
Construct map between endstates
Free Energy
ALGORITHMS MD
Software that can implement chosen map
COSMO
APPROXIMATING
Software Architect
computational scientist
Data Archivist
ANALYZE RESULTS
Statistician Stat Mech
Free Energy Analysis
Convergence Analysis
MANAGING INSTANCES
Efficient Sampling Algorithms
Coordinating the execution & managing communication b/w ensembles
RESOURCES
Application Engineer
cloud
Supercomputer
SCHEMATIC DIAGRAM
Targets
Pdbs
New molecules (sdf?)
Error Trace
Biomolecular Simulation Setup
Topologies Coordinates
Configure for ΔG calc
MD/MC Simulation inputs
Sample the equilibrium Distribution (MD Simulation)
or do proper NEQ
Coordinates Energies
Take many MD Runs → ΔGs, δ(ΔGs) (Analysis)
ΔG & δGs
Interpret
Simulation Driving Engine
Hpc
All Exist
None Modular
All flawed
All limited in scope



6. Prioritize a plan of action


COMMUNITY
Streamline licensing
Funding/ Standards partnership
Data Sharing partnership
Incentivizing standardization
Incentivizing modularization
Forcefield assignment standards
VALIDATION DATASETS FOR BIOMOLECULAR SETUP
Bounties
Demonstrate utility of Automated Modular Docking Pipelines
DIFFICULTY
IMPORTANCE


SOFTWARE
DIFFICULTY
IMPORTANCE
PHASE I
Reference Implementation Free Energy Workflow
Flexible ΔG sim. setup framework
Lossless data + translators formats for molecule prop\ docking
Common data models for communicating between different components
Modularize existing tools for setup pipes
Common ~~[illegible]~~ visualization API + ref. implem.
Common workflow component defⁿ + registry
Property Calculator Reference Imp. (Abstract software call MD using anything)
Generalized ~~ΔG~~ / phys prop analysis tools (DG Lunar) next
Reference Imp. Setup Pipeline

NEXT STEPS:
COMMON COMPONENT WORKING GROUP

standardized data formats

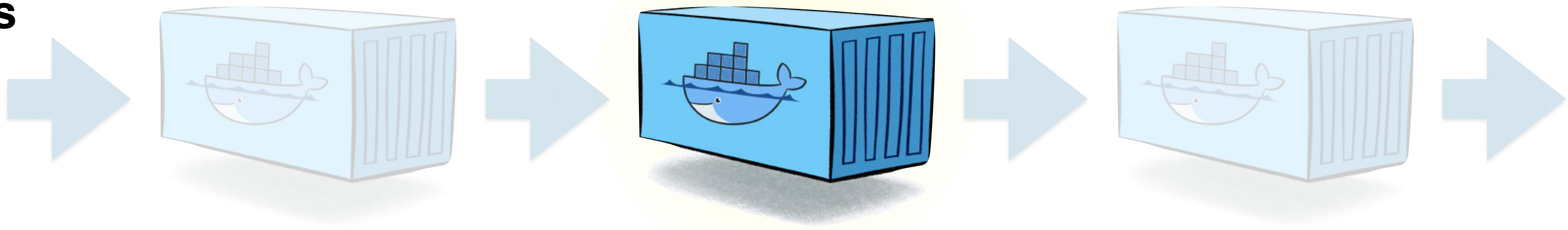
standardized data formats

protein constructs

assay conditions

molecules

industry datasets



standard benchmarks

preparation pipeline

modeling tool

automated analysis/ evaluation

biomolecular target
replace aging PDB format
handle charges, parameters, etc.
robust open source readers/writers

parameterized small molecules
make up for shortcomings in mol2, SDF
suitable for the internet age (e.g. JSON)

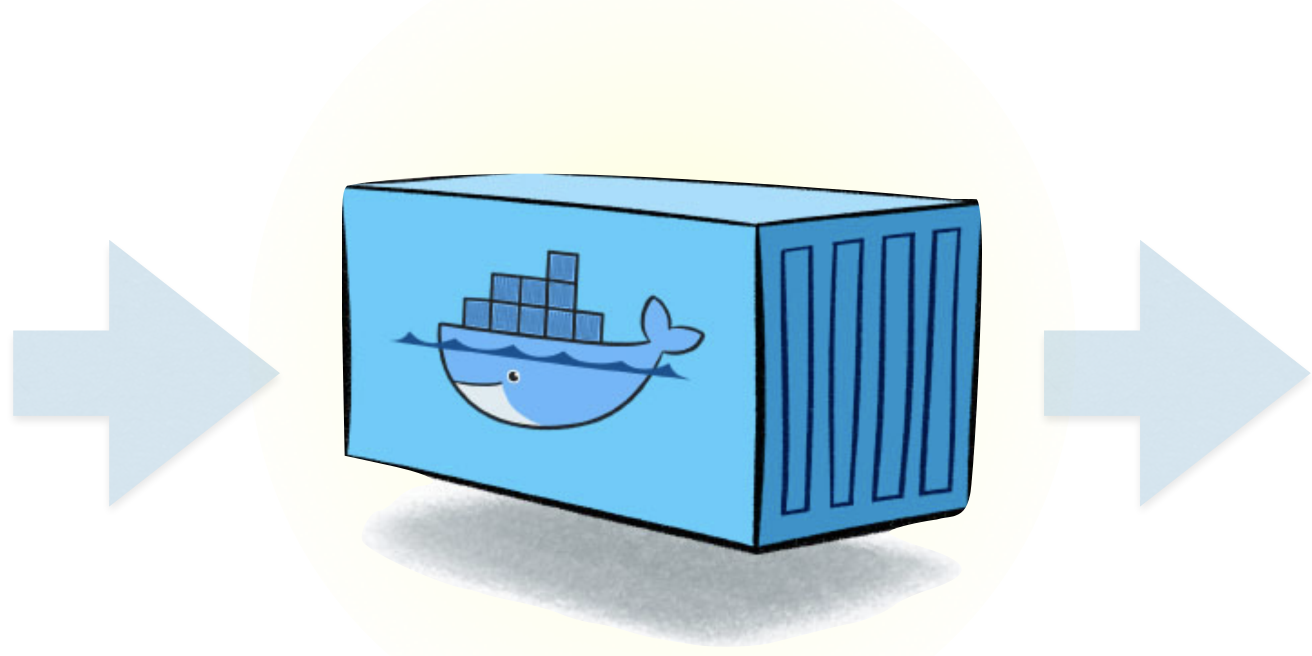
prediction formats
binding poses
predicted affinity/assay data
predict confidence/uncertainties
exception logging

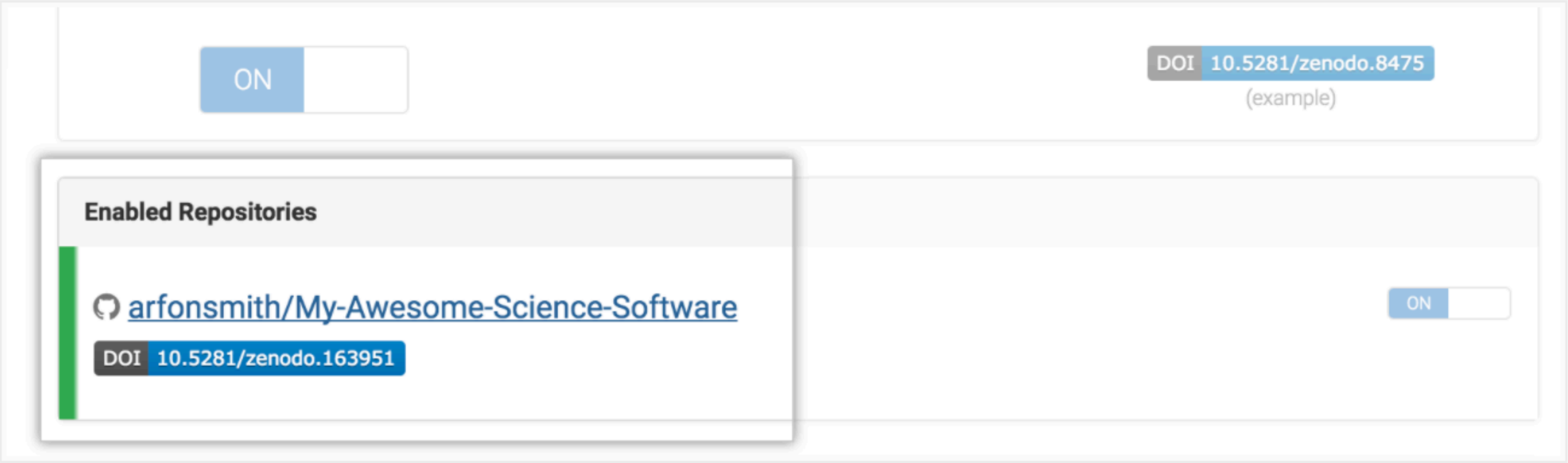
assessment formats
standard representations
standard assessments
standardized uncertainty analysis

docker

DEFINE COMMON COMPONENT FORMAT, I/O, API, AND REGISTRY

What if every modeling tool paper came with a DOI that let you pull the exact tool used in that paper from a common component registry and evaluate it yourself?





ON

DOI 10.5281/zenodo.8475
(example)

Enabled Repositories

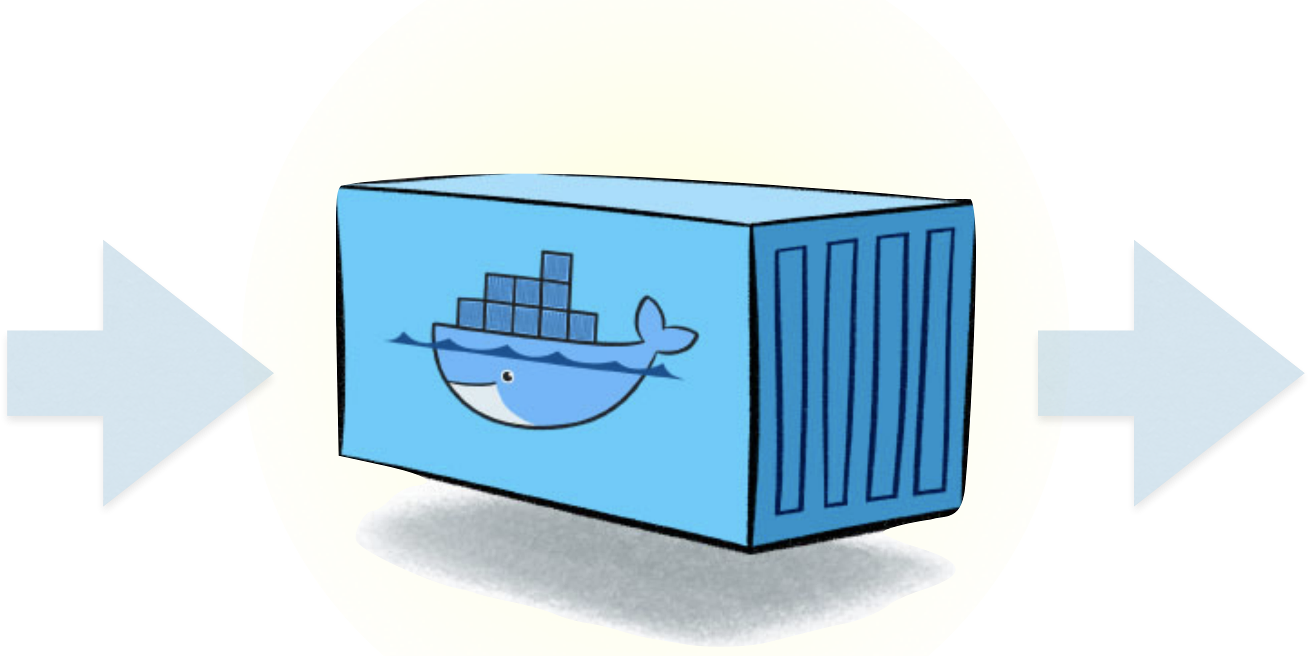
arfonsmith/My-Awesome-Science-Software

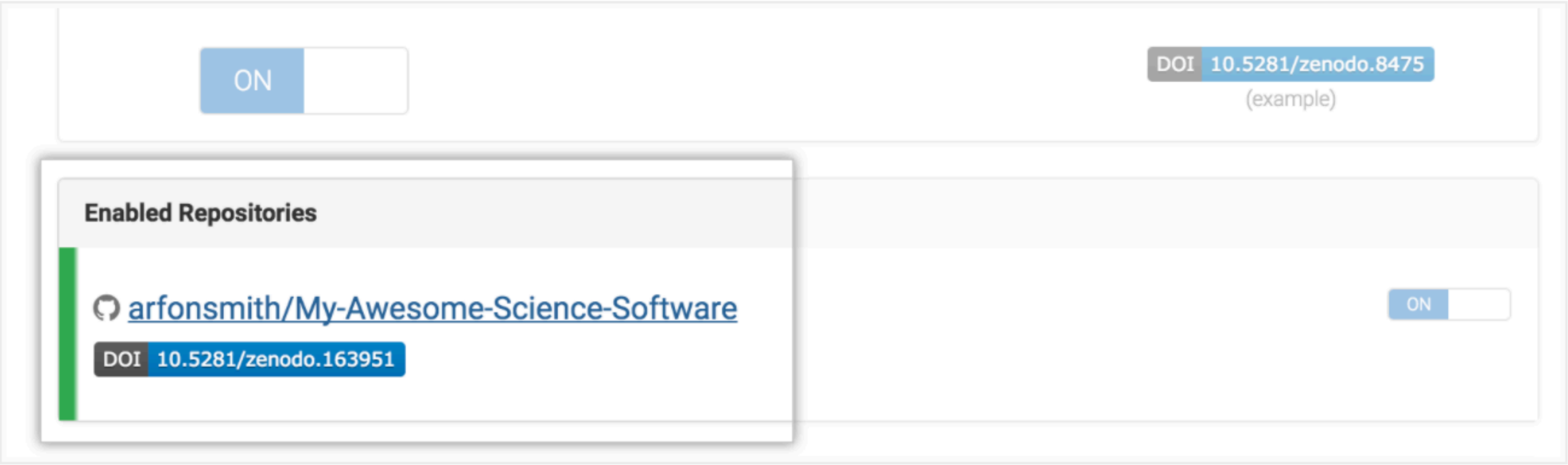
DOI 10.5281/zenodo.163951

ON

DEFINE COMMON COMPONENT FORMAT, I/O, API, AND REGISTRY

What if every modeling tool paper came with a DOI that let you pull the exact tool used in that paper from a common component registry and evaluate it yourself?





ON

DOI 10.5281/zenodo.8475

(example)

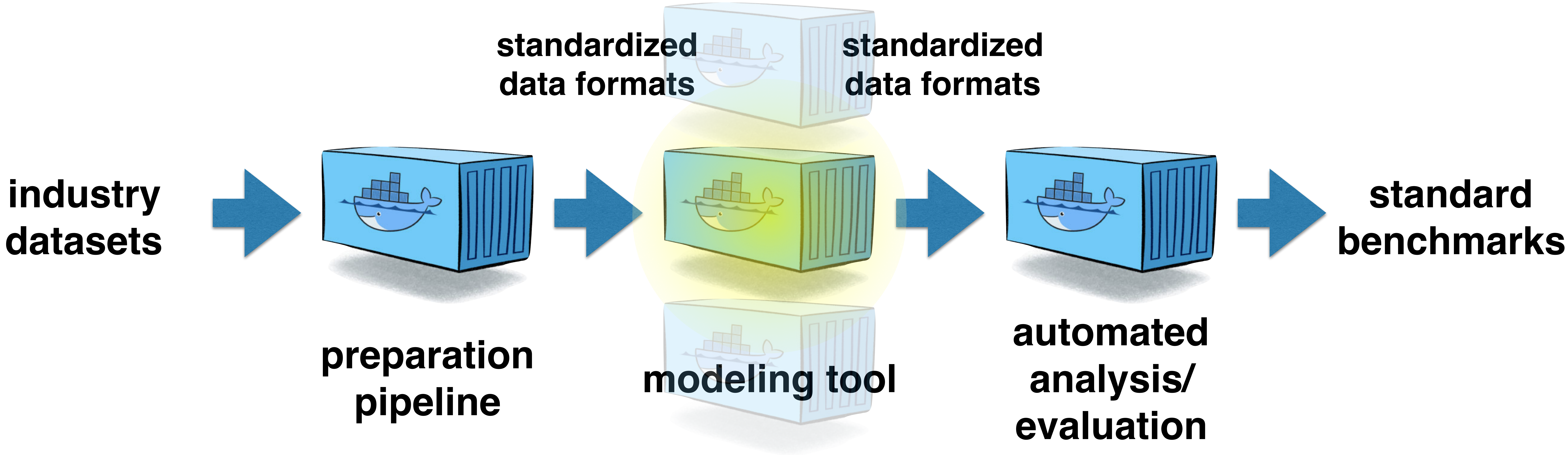
Enabled Repositories

arfonsmith/My-Awesome-Science-Software

DOI 10.5281/zenodo.163951

ON

AUTOMATED SAMPL/D3R?



industry datasets

preparation pipeline

standardized data formats

modeling tool

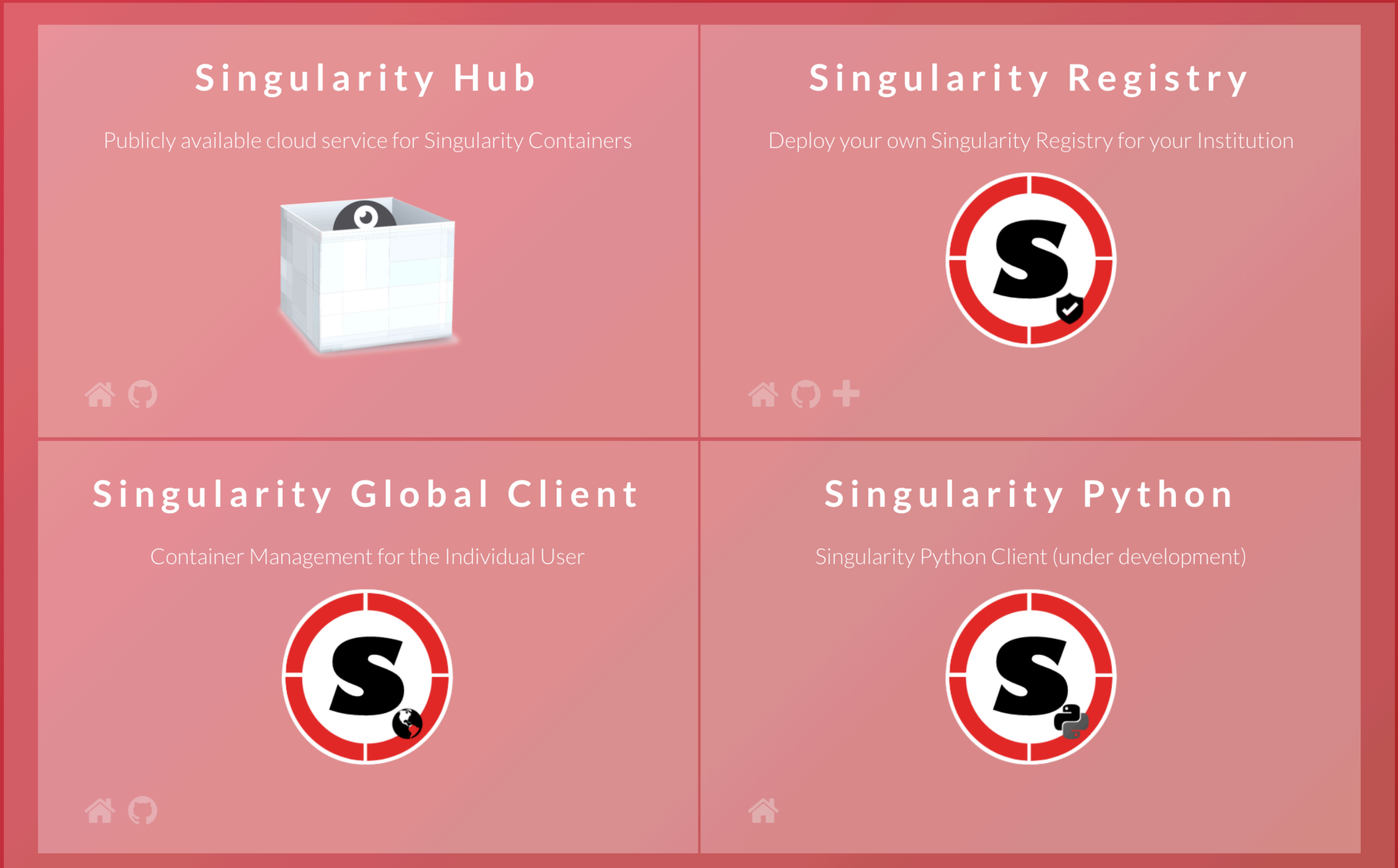
standardized data formats

automated analysis/ evaluation

standard benchmarks

We can likely find a way to raise funds for AWS / GCE time to run tools retrospectively and prospectively for modeling evaluation.

SOME NEAT TECHNOLOGY IS HELPING MAKE THIS EASY



Singularity Hub

Publicly available cloud service for Singularity Containers

Singularity Registry

Deploy your own Singularity Registry for your Institution

Singularity Global Client

Container Management for the Individual User

Singularity Python

Singularity Python Client (under development)