

# Science Gateways – Leveraging Modeling and Simulations in HPC Infrastructures via Increased Usability

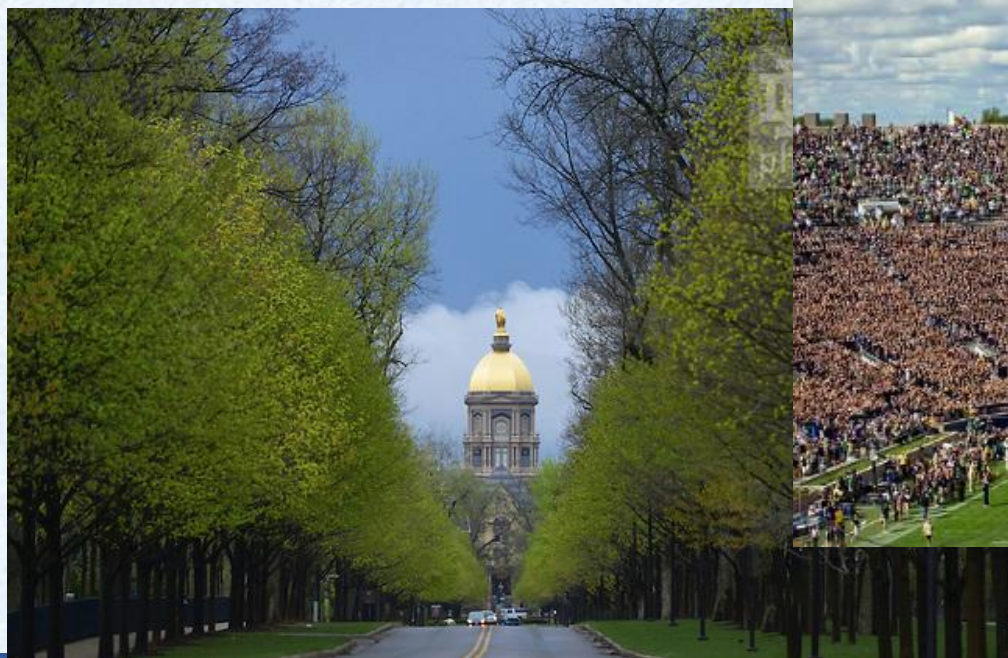
Sandra Gesing

sandra.gesing@nd.edu

cHiPSet Training School 2016

22 September 2016

- In the middle of nowhere of northern Indiana (1.5 h from Chicago)
- 4 undergraduate colleges
- ~35 research institutes and centers
- ~12,000 students





- Genomics
- Proteomics
- Metabolomics
- Immunomics
- System biology
- Molecular simulations
- Docking
- Epidemiology
- ...



Black Swallowtail –  
larvae and butterfly



February 16, 2001  
biotech company Celera



February 15, 2001  
The Human Genome Project



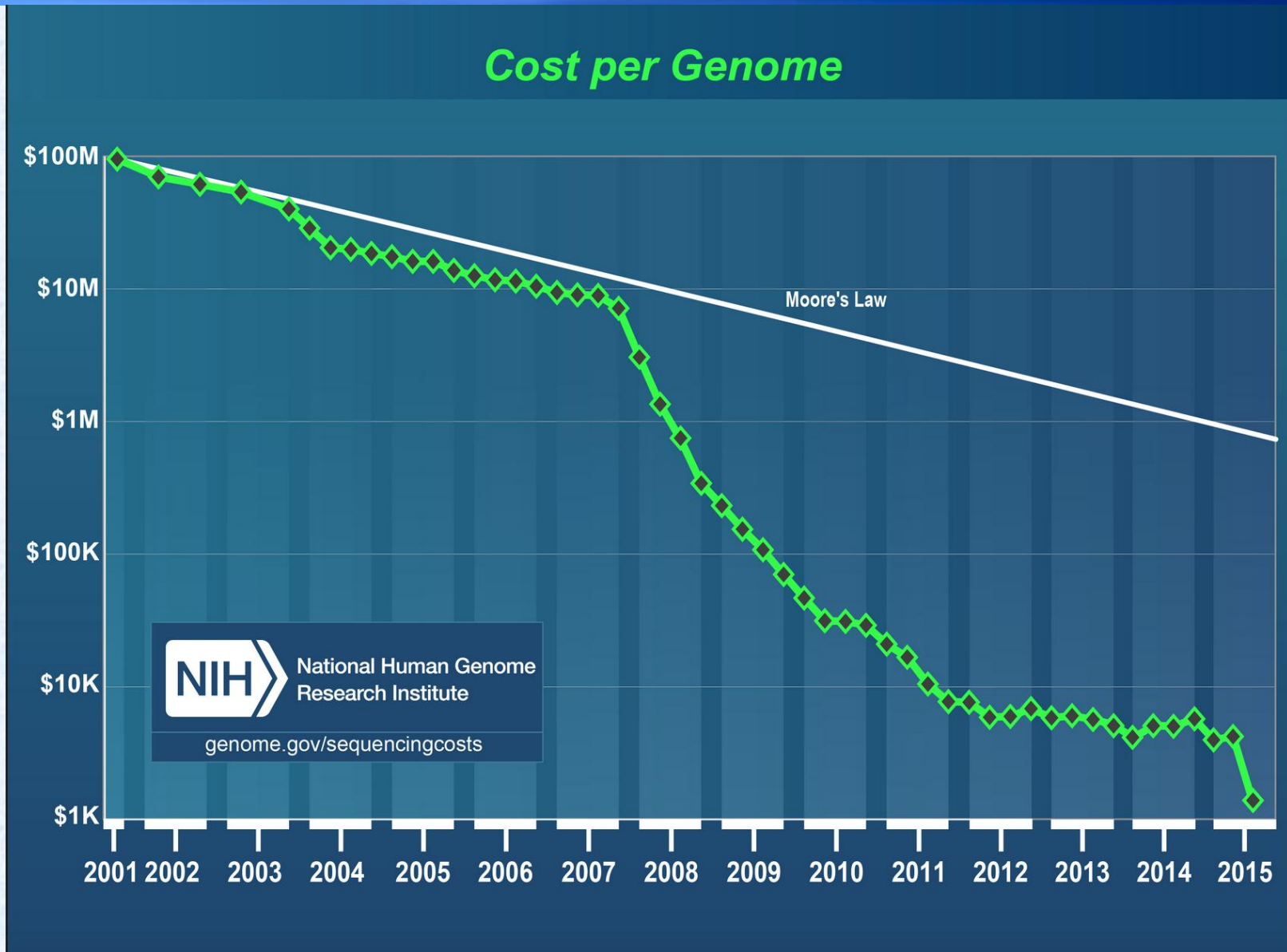


Craig Venter (left) and Francis Collins (right)

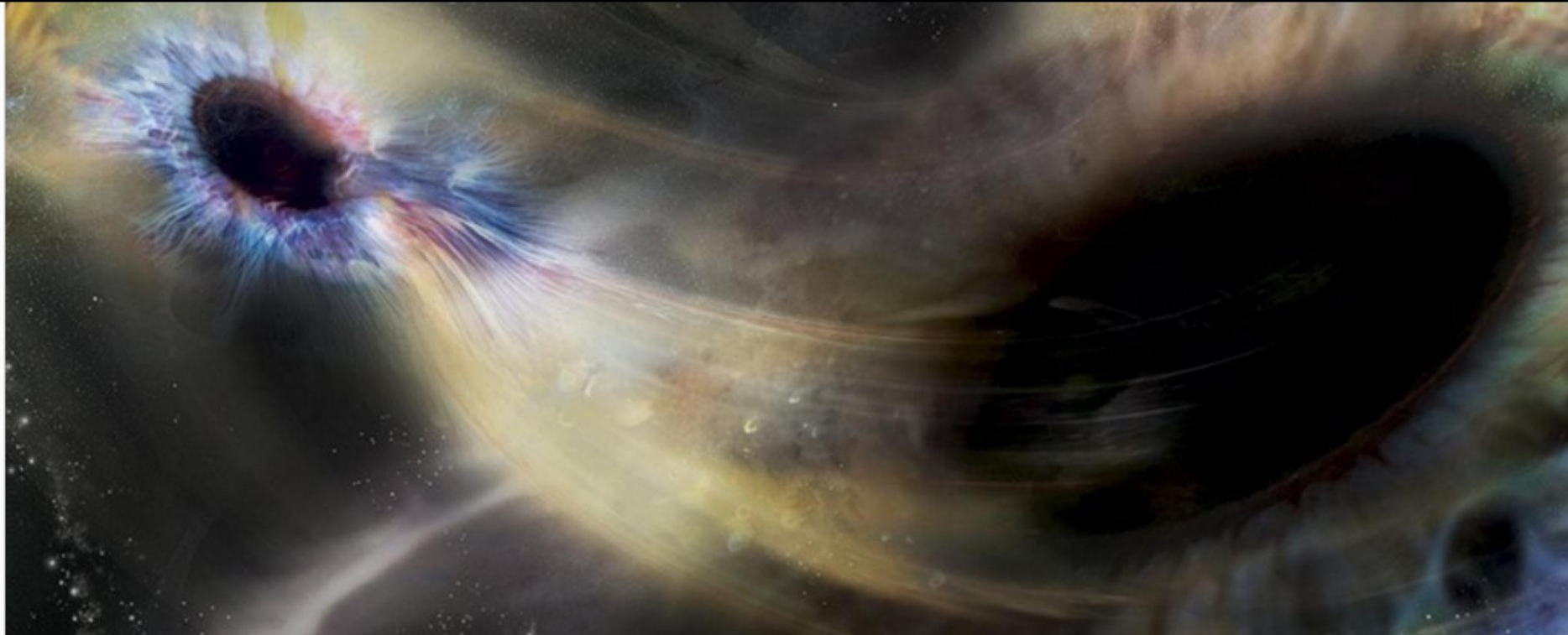
- Explosion in the quantity, variety and complexity of data
- Questions can be answered impossible to even ask about 10 years ago
- Costs far reduced (e.g., Human Genome project, 15 years, ~\$2 billion; today ~3 days, \$1000)







[http://www.genome.gov/images/content/cost\\_per\\_genome\\_oct2015.jpg](http://www.genome.gov/images/content/cost_per_genome_oct2015.jpg)



LIGO/Aurore Simonnet/Sonoma State University

## IT'S OFFICIAL: Gravitational waves have been detected, Einstein was right

"Ladies and gentlemen, we have detected gravitational waves. We did it!"

FIONA MACDONALD 11 FEB 2016



[illegible]

Slide copied from: Stuart Owen „Workflows with Taverna“

- Different workflow concepts
- Different workflow languages
- Different workflow constructs

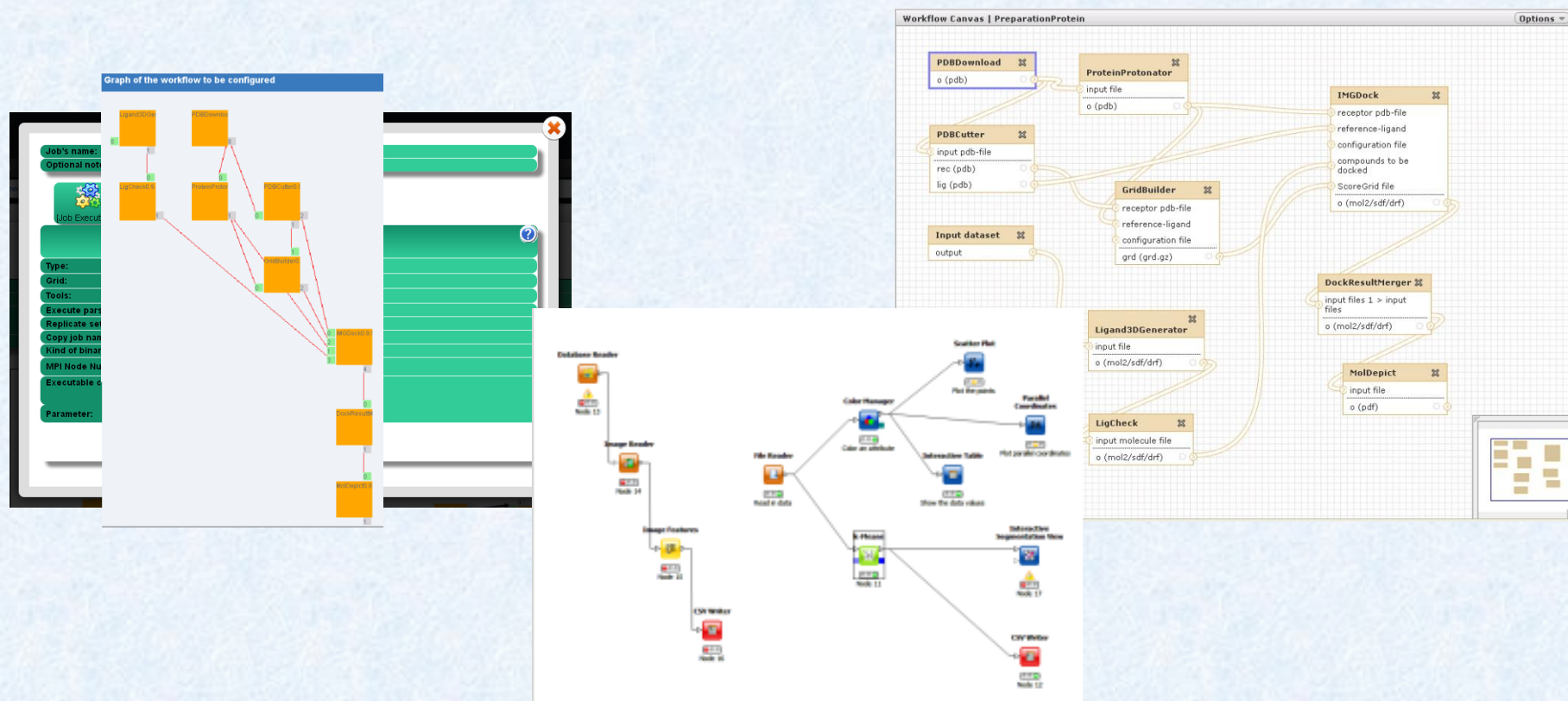


Taverna





- Different technologies (workbenches, web-based)
- Different look-and-feel



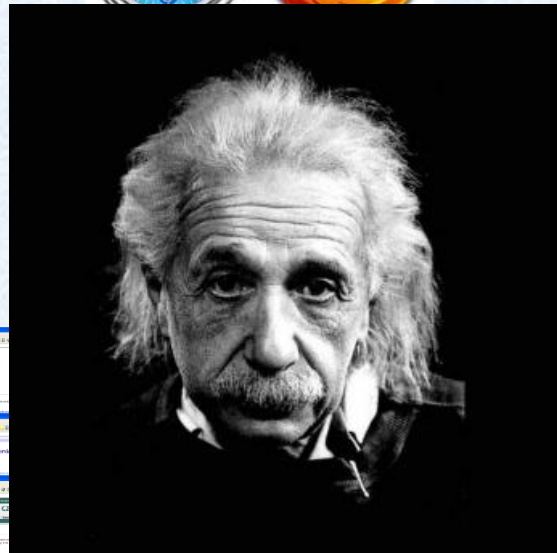
Data and compute-intensive problems



Web-based agile frameworks



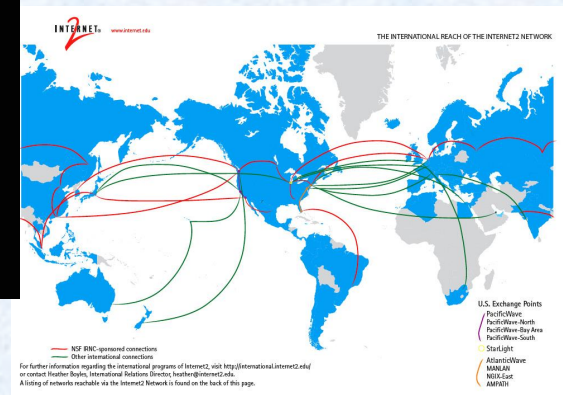
Distributed data and computing infrastructures



Users generally not IT specialists



Tools and workflow engines



High-speed networks



Data and compute-intensive problems



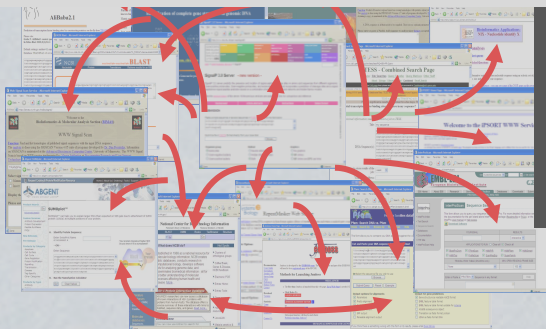
Web-based agile frameworks



Distributed data and computing infrastructures



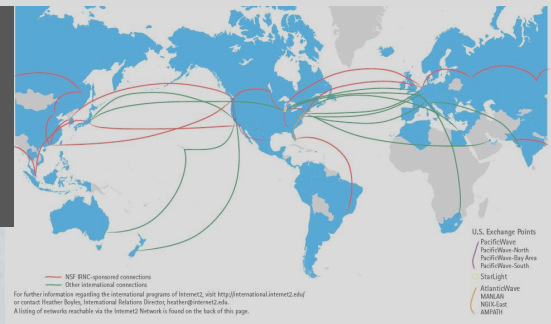
**Need for intuitive and self-explanatory user interfaces!**



Tools and workflow engines



Users generally not IT specialists



High-speed networks

# Challenge for Developers

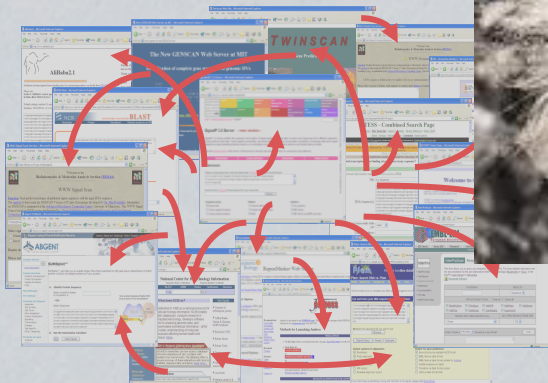
Data and compute-intensive problems



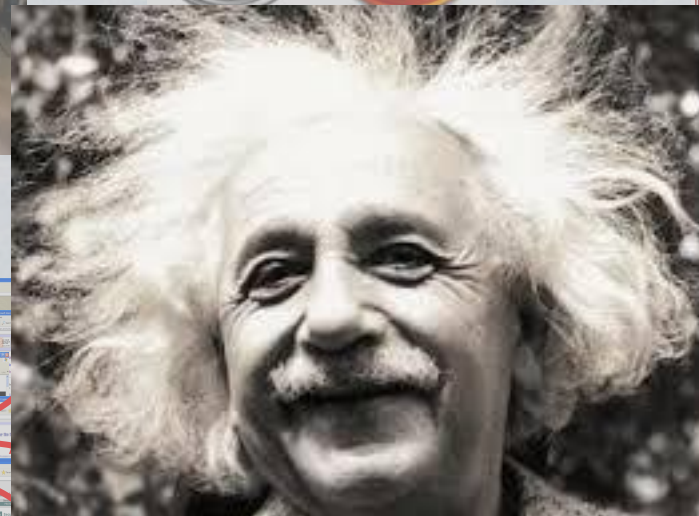
Web-based agile frameworks



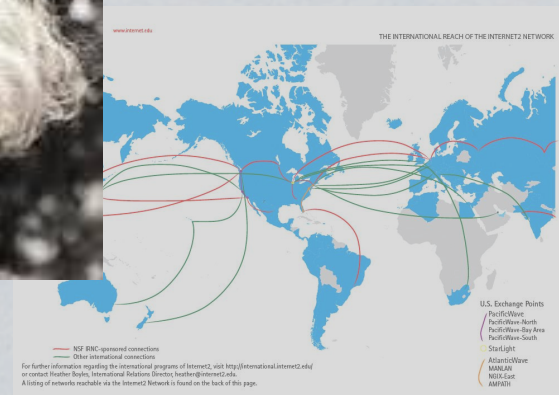
Distributed data and computing infrastructures



Tools and workflow engines



Users generally not IT specialists



High-speed networks



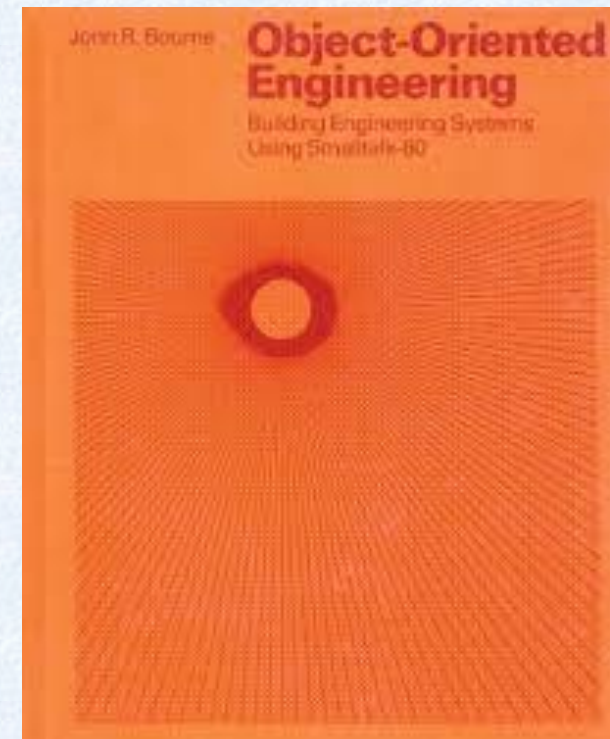
“After all, usability really just means that making sure that something works well: that a person ... can use the thing - whether it's a Web site, a fighter jet, or a revolving door - for its intended purpose without getting hopelessly frustrated.”

(Steve Krug in “Don't make me think!: A Common Sense Approach to Web Usability”, 2005)



“The key to productivity is reusability. The easiest way to produce code is obviously to have it already!”

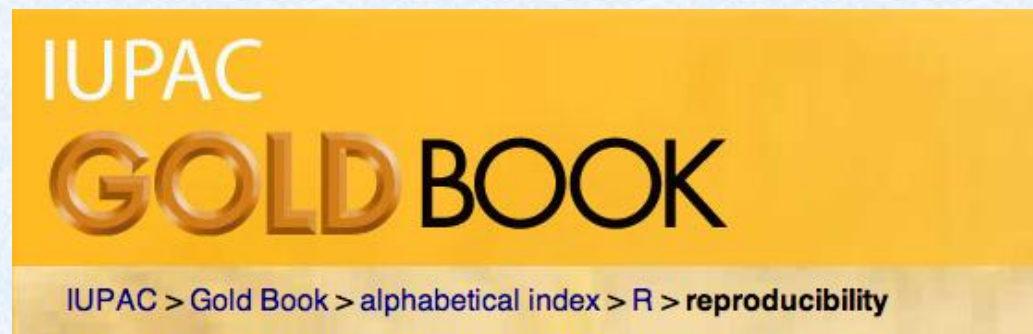
(John R. Bourne in “Object-oriented Engineering: Building Engineering Systems Using Smalltalk-80”, 1992)





“The closeness of agreement between independent results obtained with the same method on identical test material but under different conditions (different operators, different apparatus, different laboratories and/or after different intervals of time)...”

(IUPAC (International Union of Pure and Applied Chemistry [iupac.org](http://iupac.org)) GoldBook)



## “The closeness of agreement between independent

### BioMed Central pilots projects around enhancing reproducibility

#### Reproducibility: what are we going to do about it?



With an increasing number of studies revealing much of science is not able to be reproduced or replicated, the question is now being asked, *Can we do science better?*

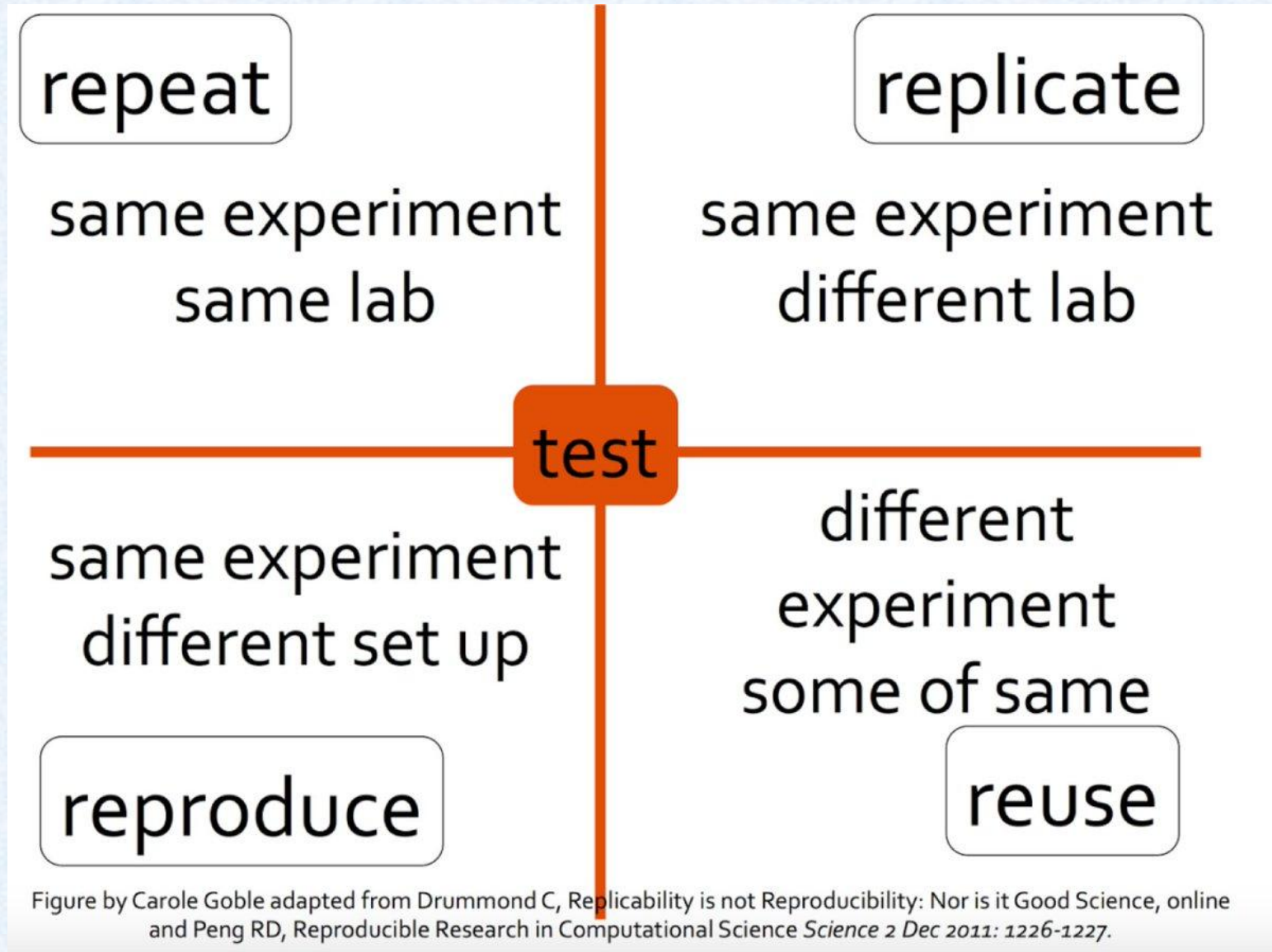
In an effort to address this we are pleased to announce that we are piloting a new **Minimum Standards of Reporting Checklist** for authors and reviewers.

The checklist addresses three areas of reporting: experimental design and statistics, resources, and availability of data and materials. Authors will have to confirm they have complied with the checklist and reviewers are asked to confirm the author's answers. For more information read our [launch editorial](#).

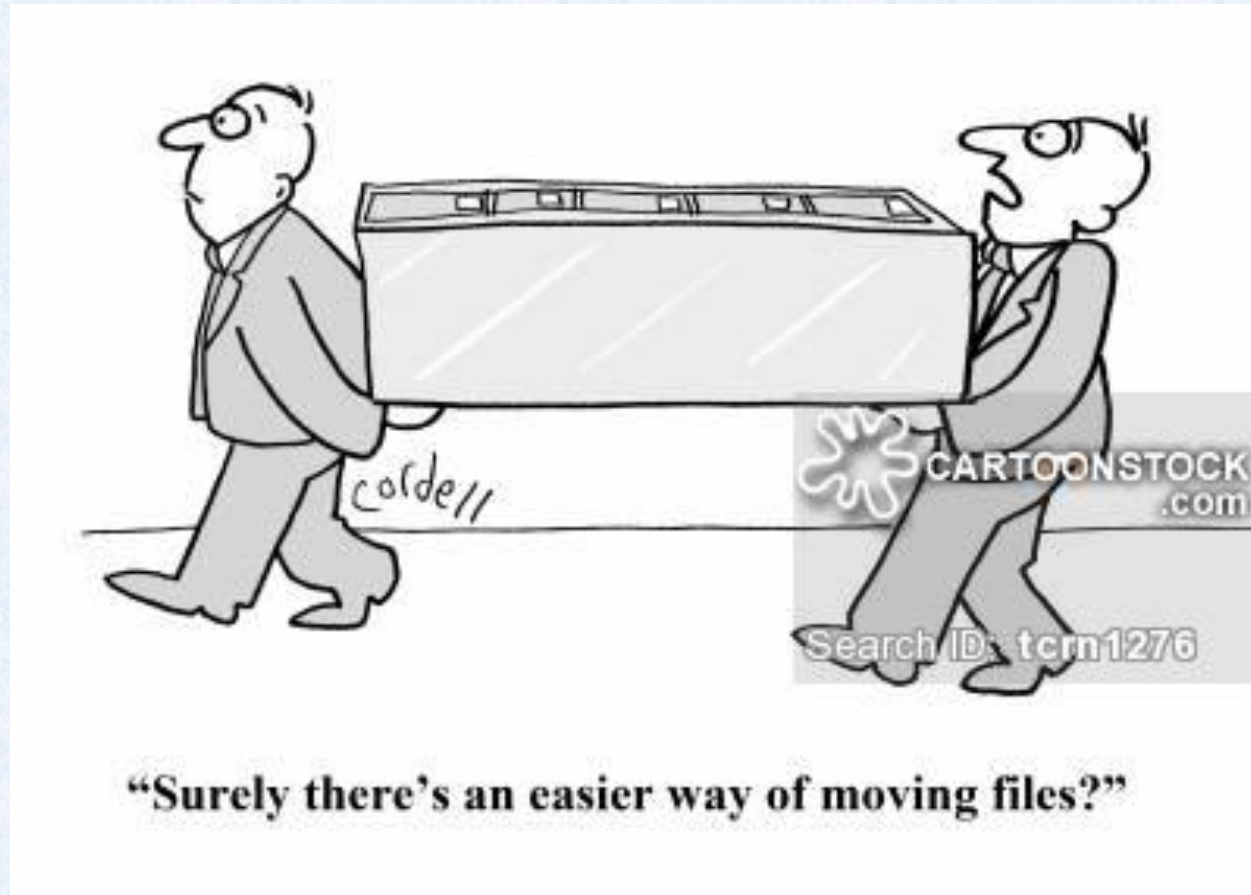
# GOLD BOOK

IUPAC > Gold Book > alphabetical index > R > reproducibility





- Time
- Computational resources
- Money





**science gateway** /sī' əns gāt' wā/ n.

1. an online community space for science and engineering research and education.
2. a Web-based resource for accessing data, software, computing services, and equipment specific to the needs of a science or engineering discipline.

The collage displays five distinct science gateway interfaces:

- VecNet**: A red-themed site titled "Vector-Borne Disease Network" with a navigation bar and sections for "Explore the VecNet", "Information Sharing", "Collaborations", and "News".
- VectorBase**: A green-themed site titled "Bioinformatics Resource for Invertebrate Vectors of Human Pathogens" featuring a mosquito icon, a "Welcome to VectorBase!" message, and sections for "DATA" (Genomes, mRNA, Proteins, Mitochondrial Sequences) and "TOOLS & RESOURCES".
- SPACES**: A site titled "Spatial Portal for Analysis of Climatic Effects on Species" with a map background, a "Welcome to SPACES!" message, and a "RECENT ADDITIONS" section.
- CyberEye**: A site titled "Research Computing" with a blue header, a "CyberEye" logo, and a sidebar with links to "About", "People", "Portal", "Tutorials", and "Publications".
- NOTRE DAME CyberEye**: A site titled "OFFICE of the VICE PRESIDENT for RESEARCH" with a "Portal" section, a "Search" bar, and a "WORKFLOWS" section showing "Rapid Risk Assessment (RRA)" and "Data Intake & Discovery (DID)".

- Increased complexity of
  - today's research questions
  - hardware and software
  - skills required
- Greater need for openness and reproducibility
  - Science increasingly driving policy questions
- Opportunity to integrate research with teaching
  - Better workforce preparation

*We need interfaces  
that provide  
broad access to  
advanced resources  
and  
allow **all** to tackle  
today's challenging  
science questions.*



Untitled form

File Edit View Insert Responses (0) Tools Add-ons Help

◀ ▶ Edit questions Change theme View responses View live form

Form Settings

- ☒ Require University of Notre Dame login to view this form
- ☐ Automatically collect respondent's University of Notre Dame username
- ☐ Show progress bar at the bottom of form pages
- ☐ Only allow one response per person (requires login) ?
- ☐ Shuffle question order ?

Page 1 of 1

**Untitled form**

Form Description

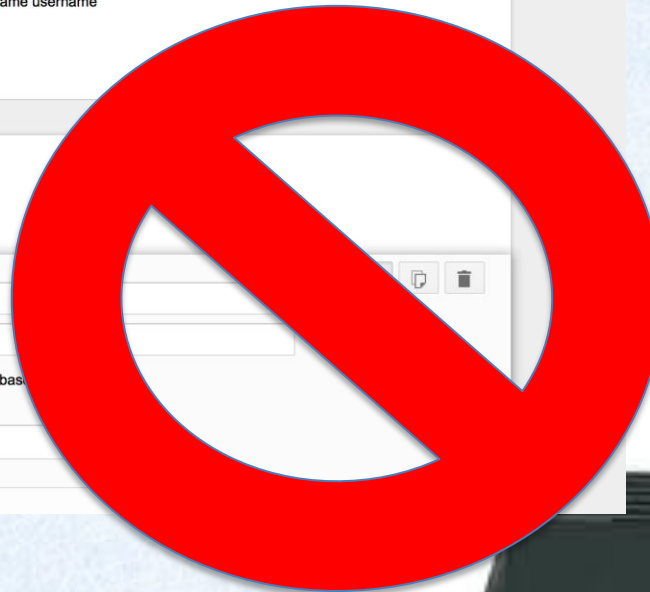
Question Title: Untitled Question

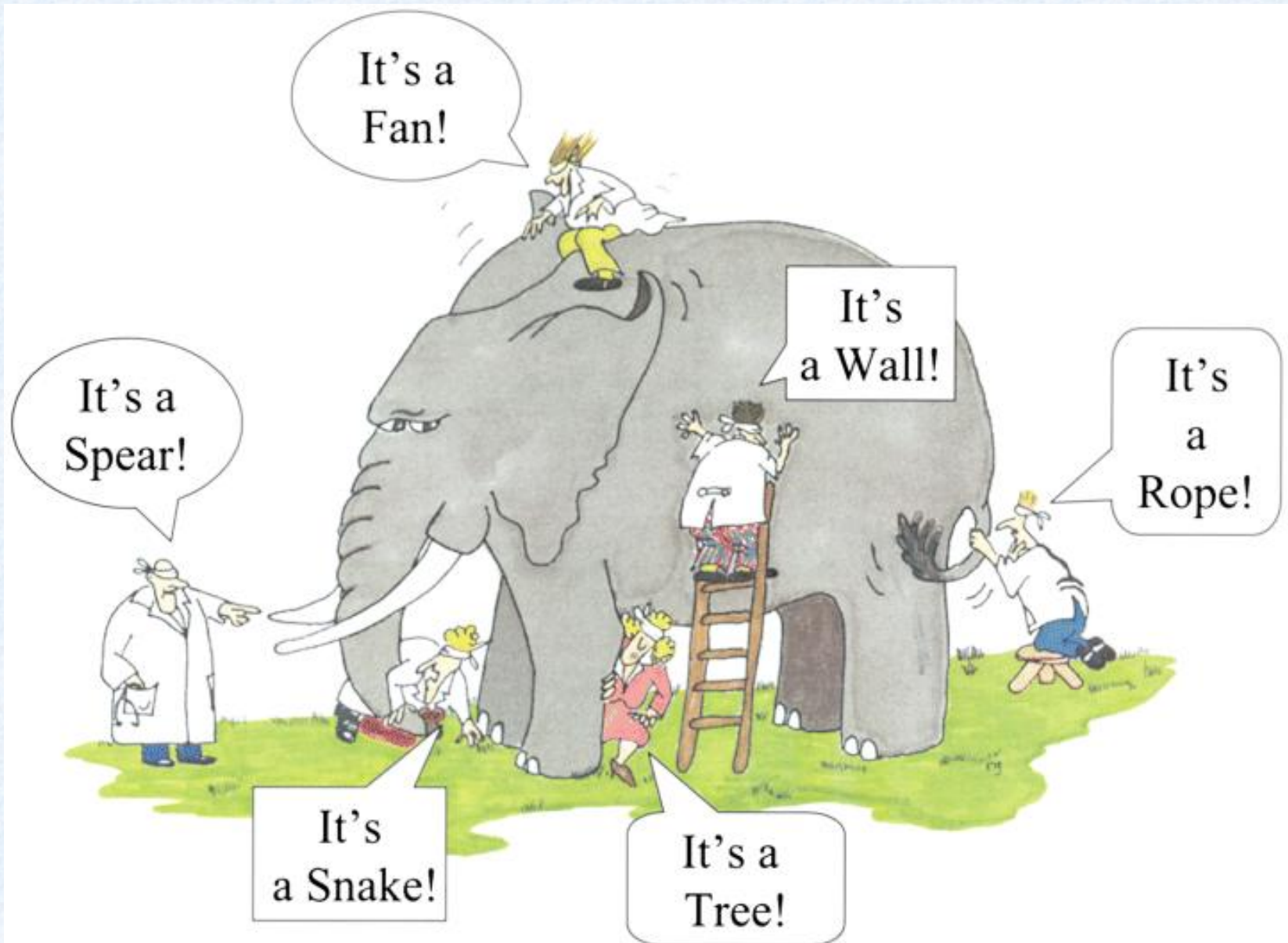
Help Text:

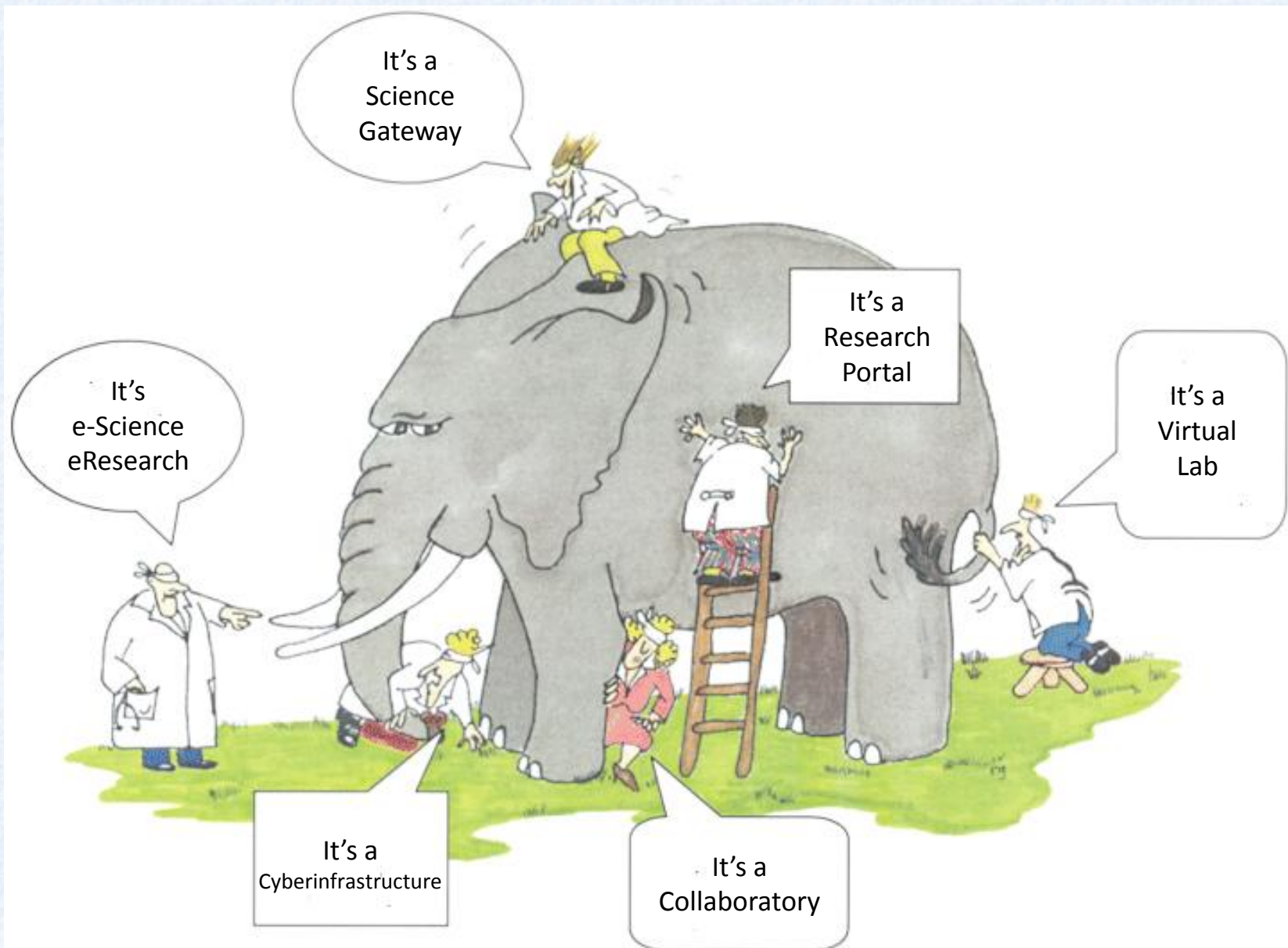
Question Type: Multiple choice ▾ ☐ Go to page base

☐ Option 1

☐ Click to add option









Re-inventing is not always necessary..



... and users should get more features easily...



... but the model should fit to the demands of the community





Questions around frustration and limitations of using

- Bioinformatic software
- Bioinformatic resources
- HPC and Cloud infrastructures

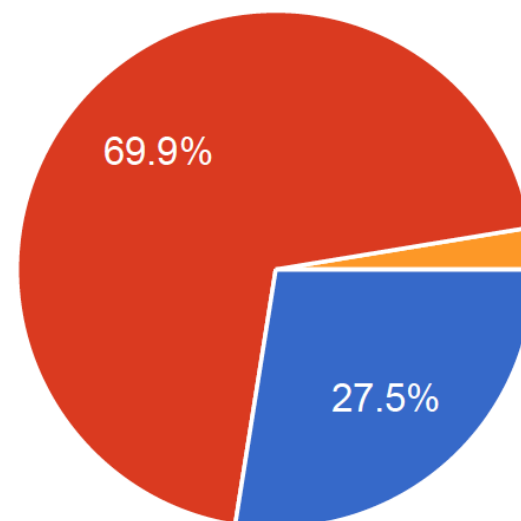
and about challenges to train students in bioinformatics

Answers often address

- Hurdles to use bioinformatic resources because of commandline access or not available software
- Quality of documentation of software
- Need for parsers and converters for diverse data formats
- Long waiting time for support or even lack of support

- Nick Loman  
(Birmingham, UK)
- Thomas Connor  
(Cardiff, UK)
- October 2015
- 272 answers

## Are you located in the UK?



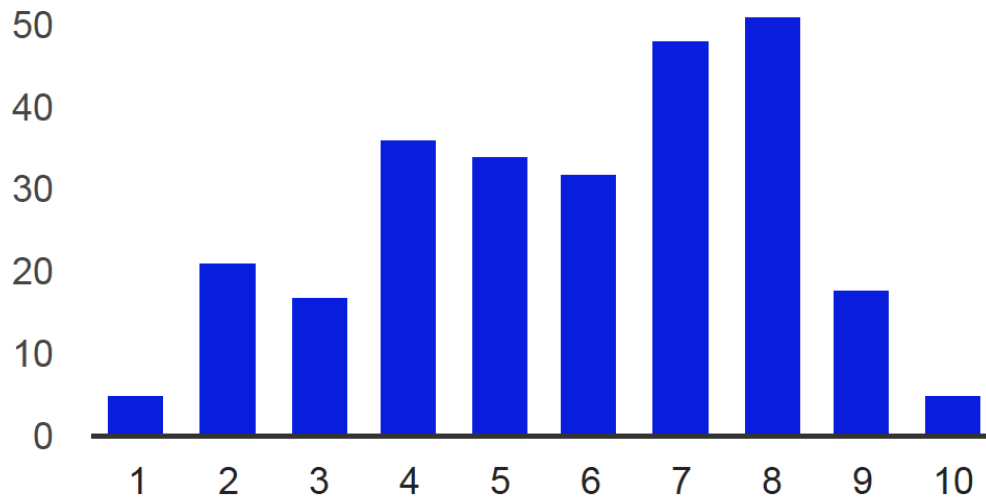
Yes      **74**      27.5%

No      **188**      69.9%

Is Scotland still in the UK?      **7**      2.6%

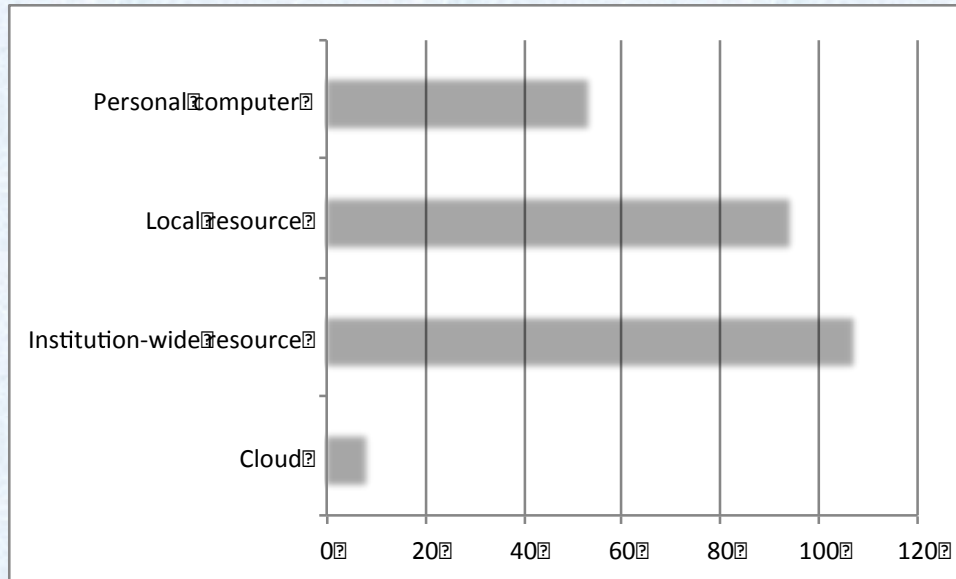
<https://drive.google.com/drive/folders/0B7KZv1TRi06fLUJCU1BYM3JScjg>

**How would you rate your level of bioinformatics expertise (0 = total n00b, 10 = Heng Li) ?**

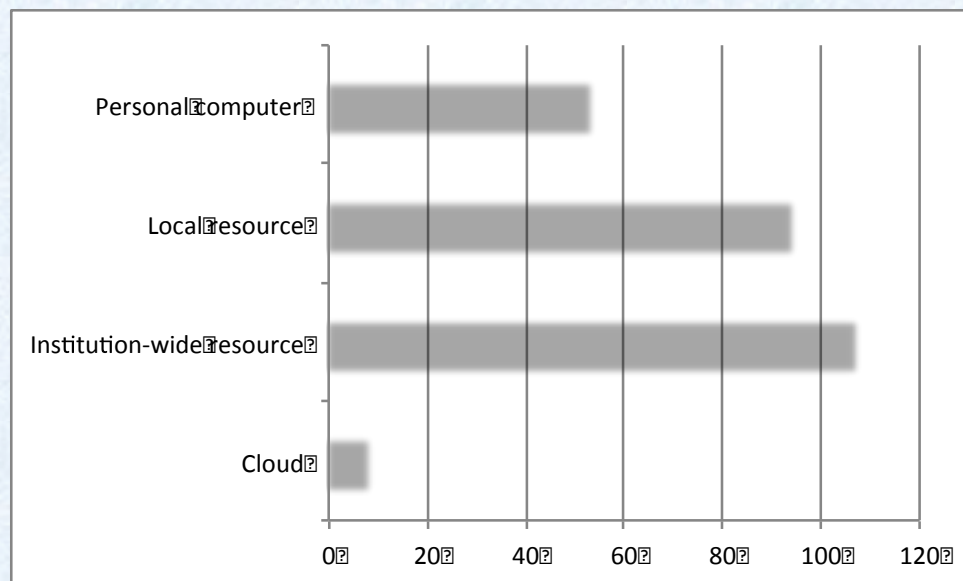




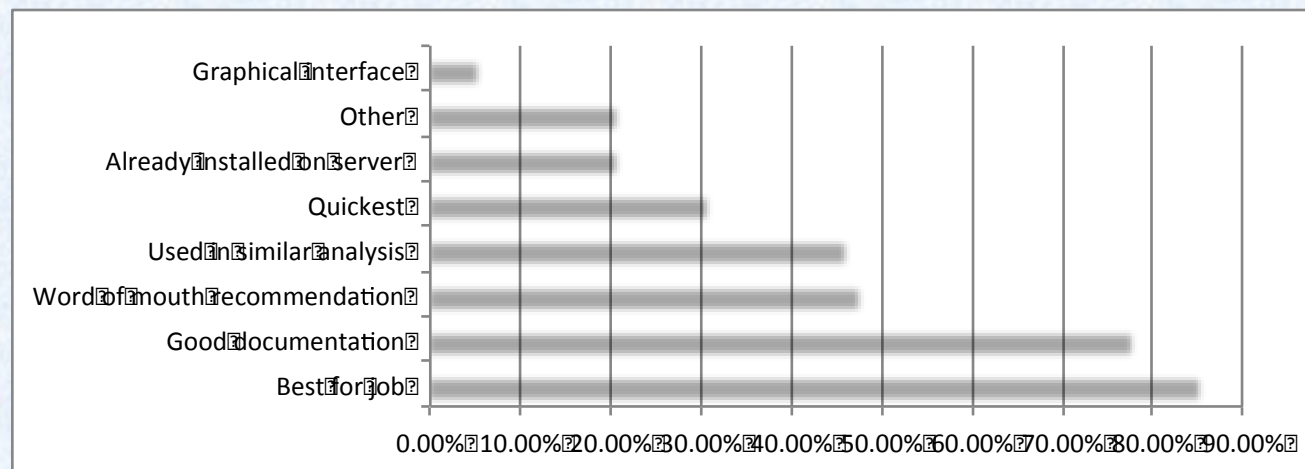
## Where do bioinformaticians do most of their work



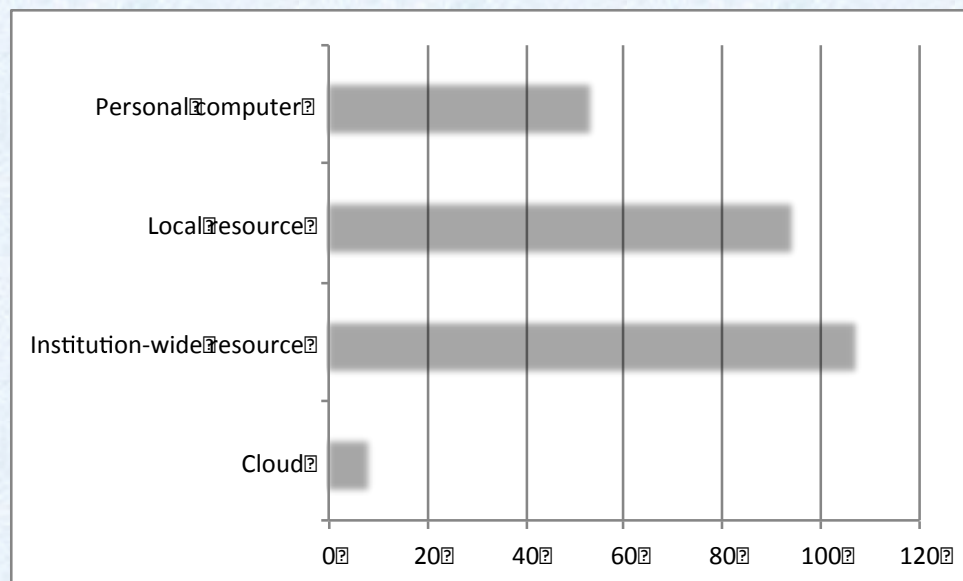
## Where do bioinformaticians do most of their work



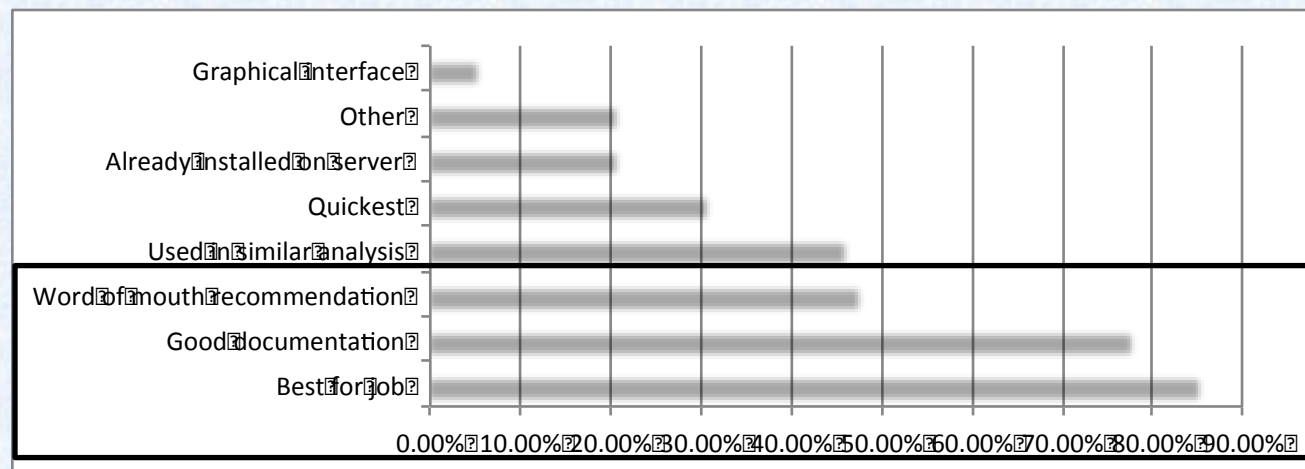
## Why do bioinformaticians use the software they use



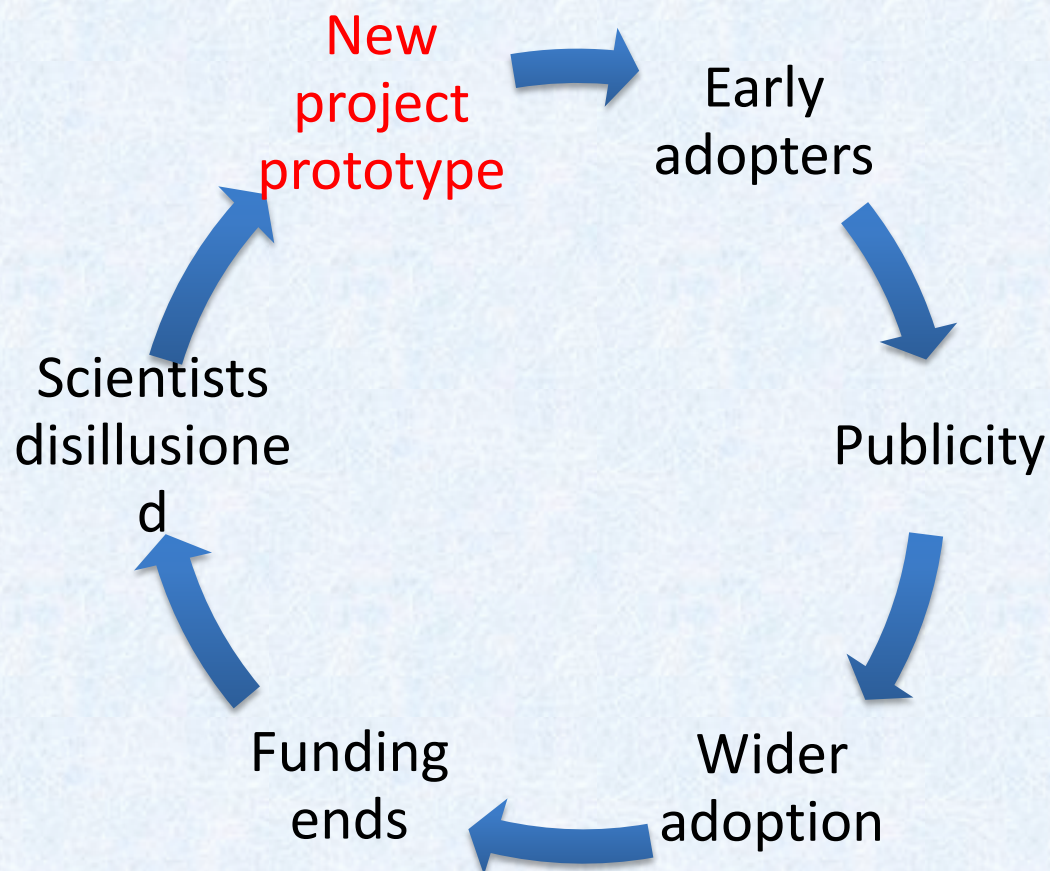
## Where do bioinformaticians do most of their work



## Why do bioinformaticians use the software they use







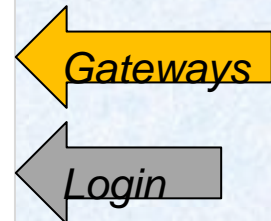
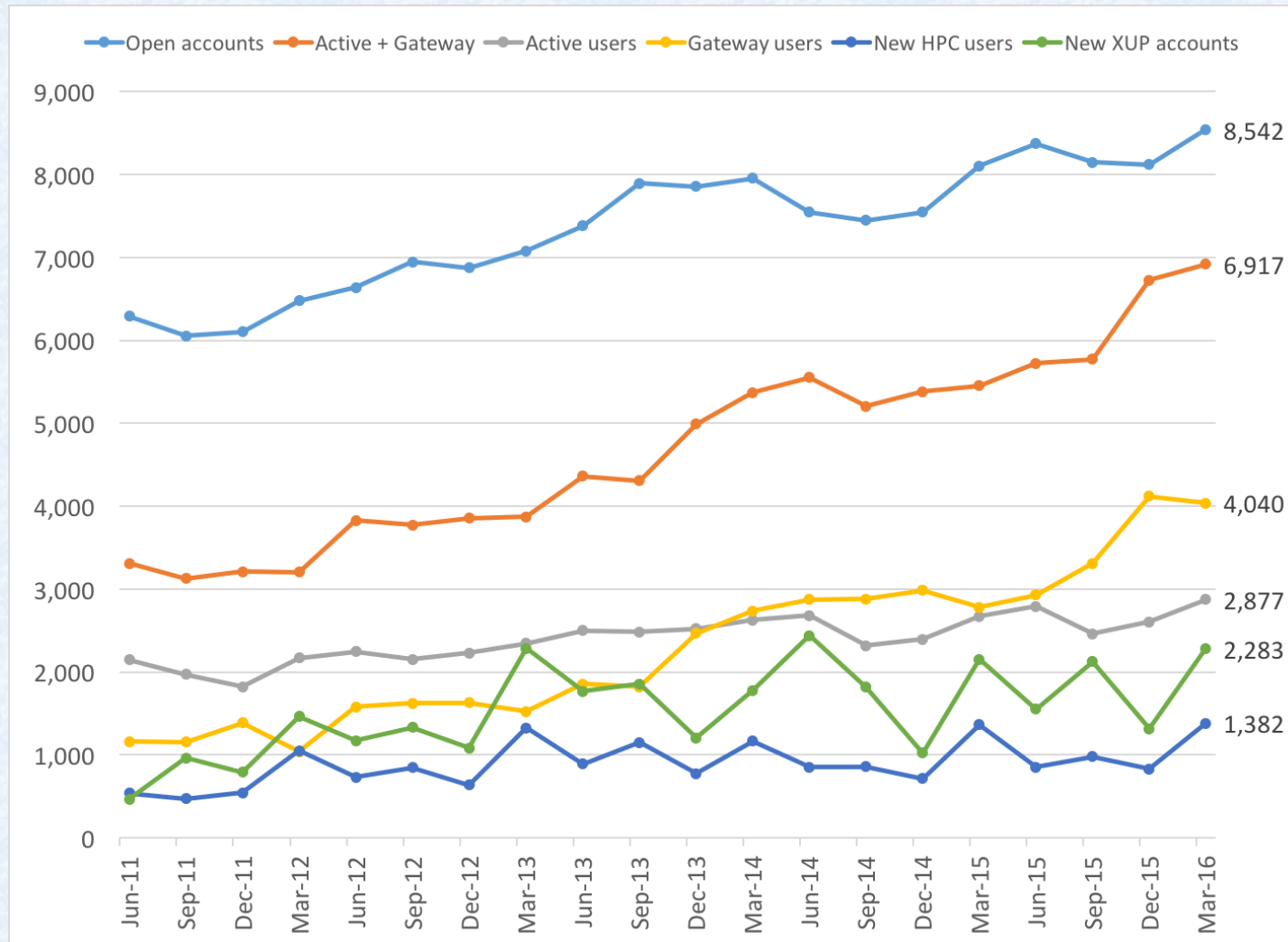
**Gateways enable research, but are not research projects themselves...**

**Sustainability is a problem...**

A new era...

- Novel developments of web-based agile frameworks
- Infrastructure providers report that science gateways are more used than commandlines

## A new era...



<https://www.xsede.org/>

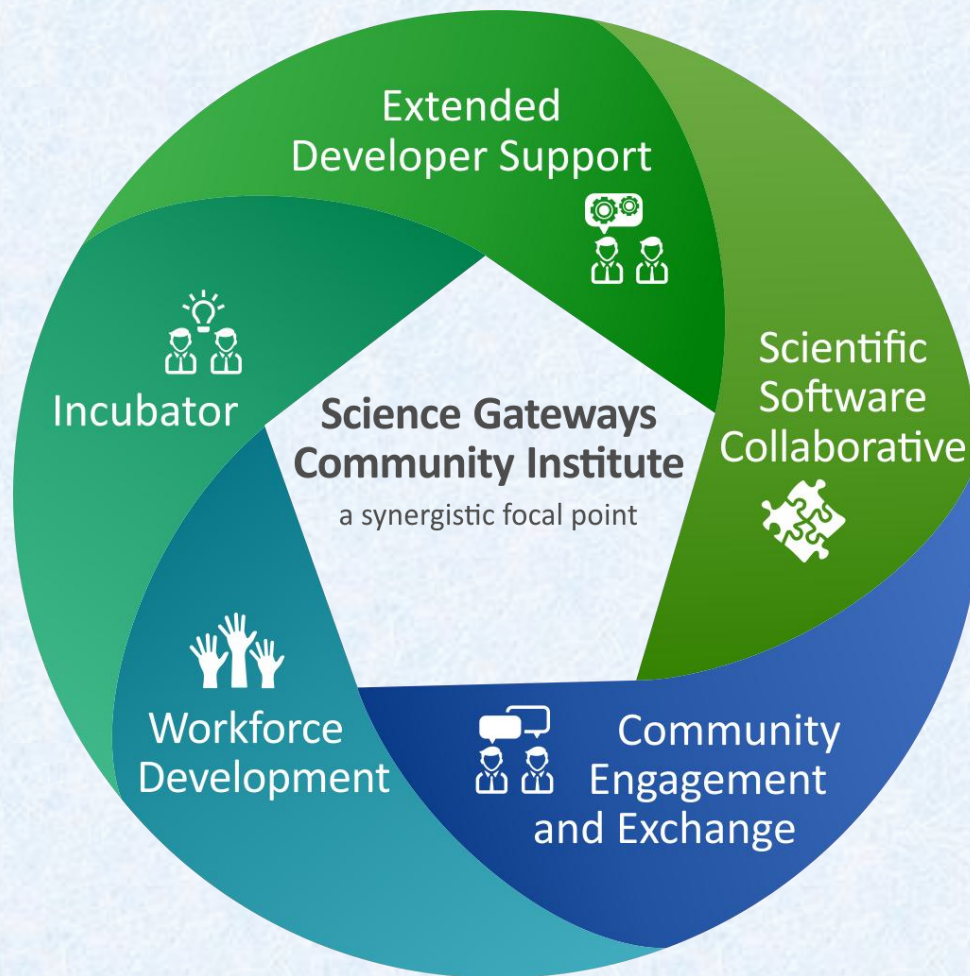


A new era...

- Novel developments of web-based agile frameworks
- Infrastructure providers report that science gateways are more used than commandlines

But also always new challenges...

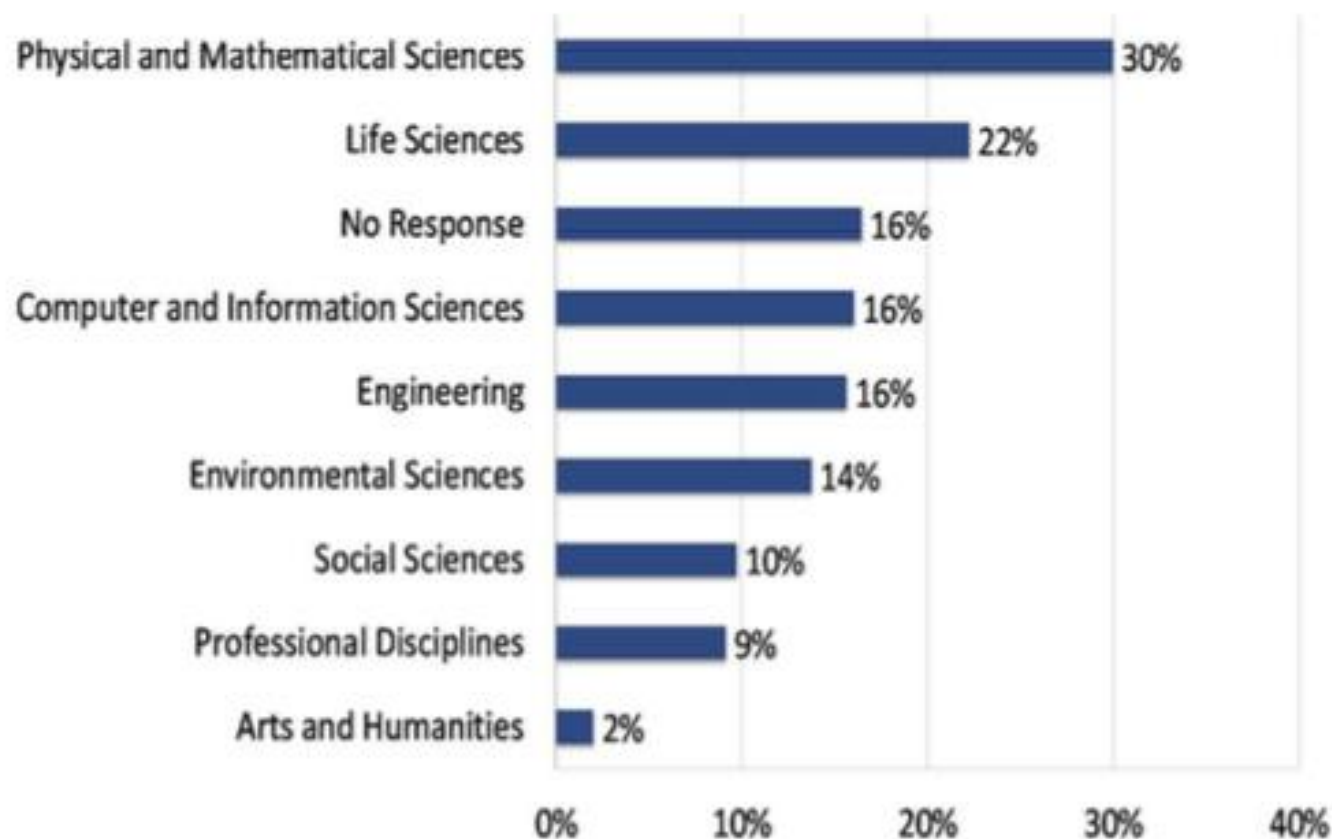
- Novel infrastructures
  - Novel data sources like NGS sequencing machines, telescopes such as the Square Kilometre Array (SKA) (will create data rates in exa-scale size)
- ➔ Support of developers necessary



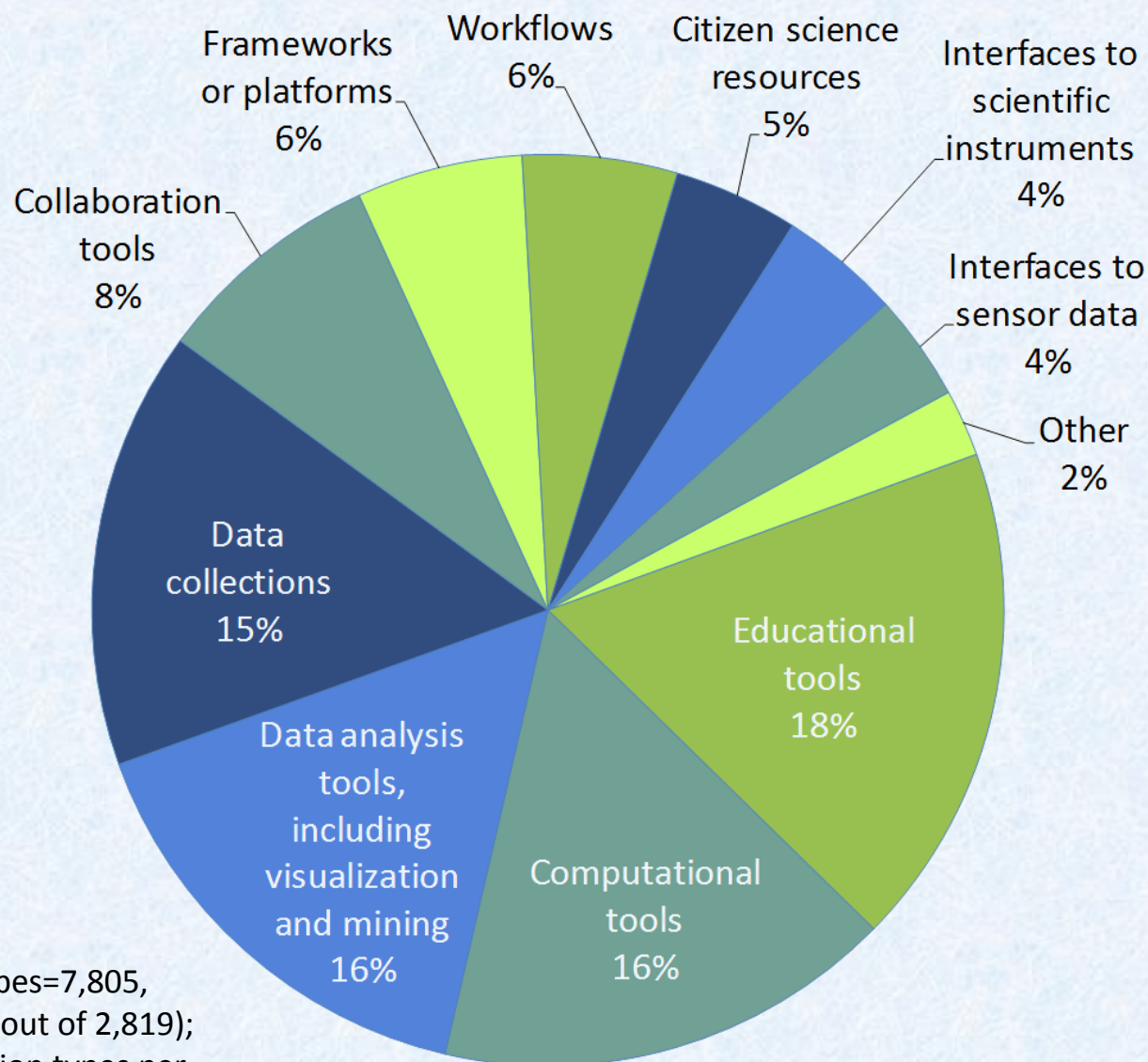
- Diverse expertise on demand
- Longer term support engagements
- Software and visibility for gateways
- Information exchange in a community environment
- Student opportunities and more stable career paths

<http://sciencegateways.org>

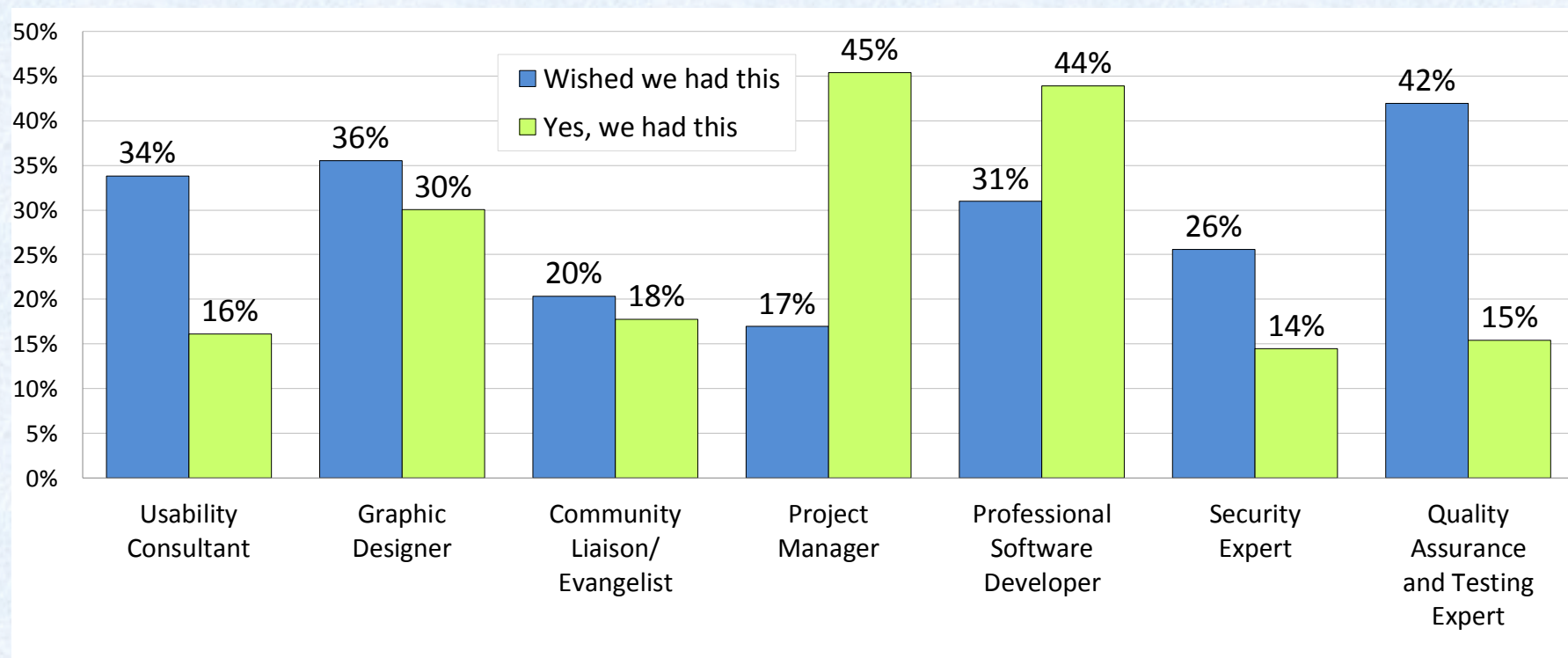
- 29,000-person survey
- 4957 responses from across domains





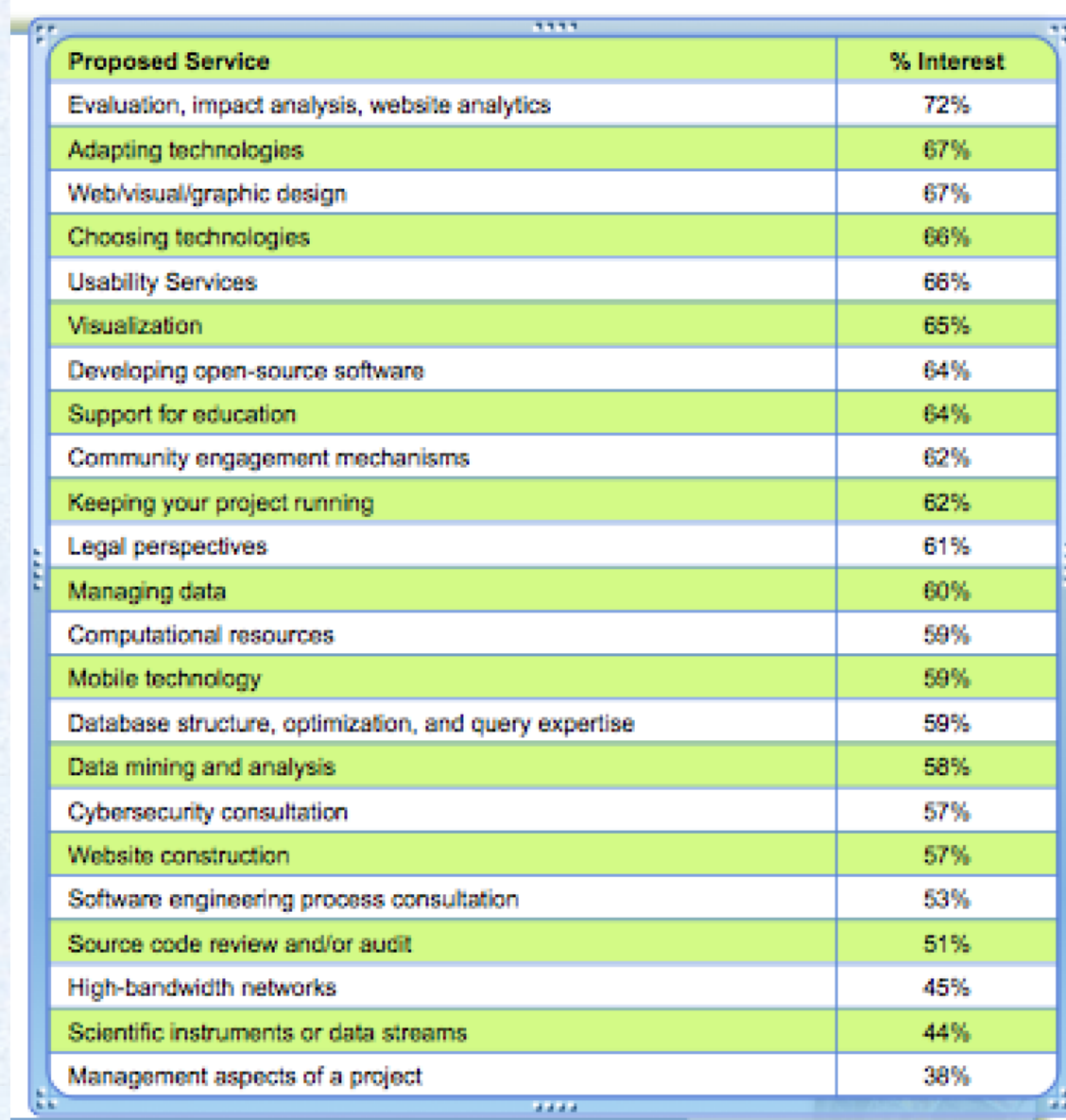


n of application types=7,805,  
by 2,756 creators (out of 2,819);  
mean=2.8 application types per  
application creator



n=2,756 respondents or 98% of application creators

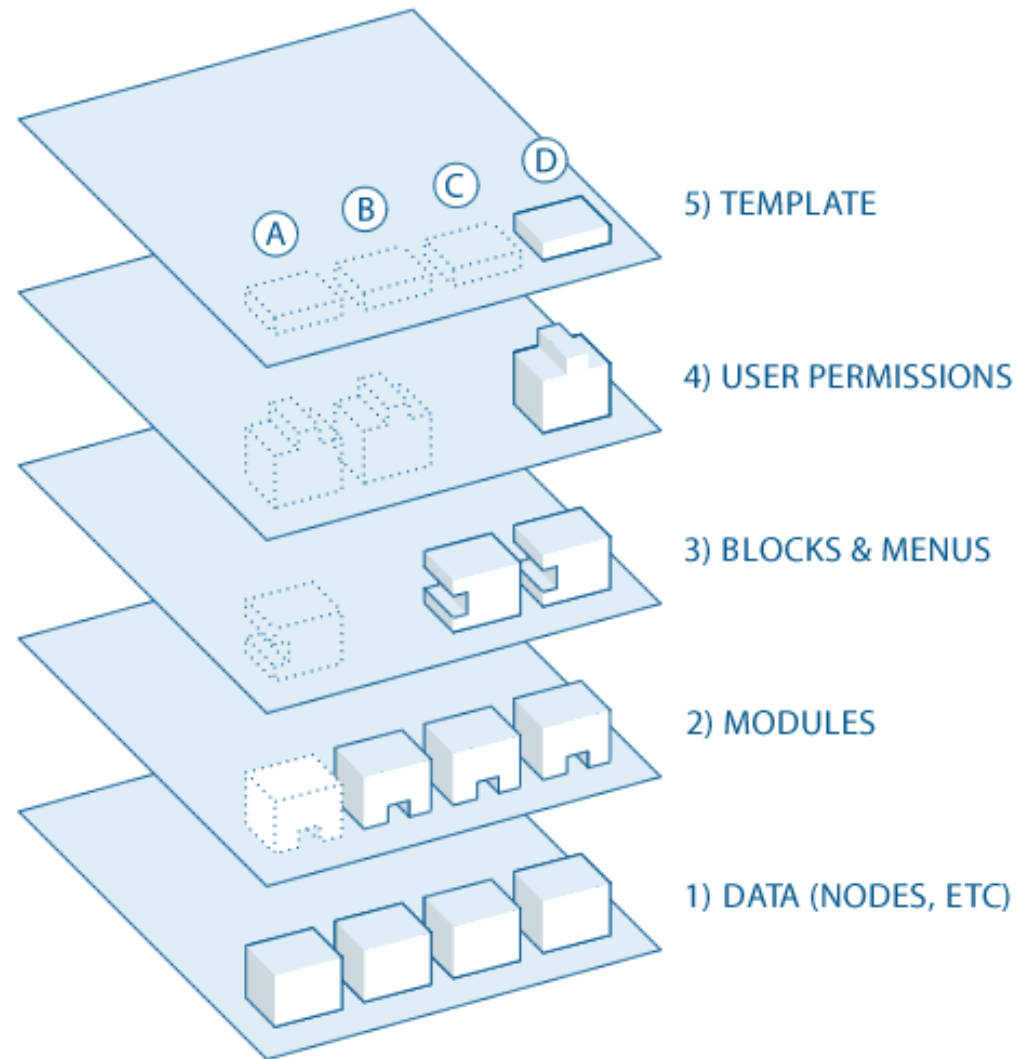
What services  
would be helpful?



Proposed Service	% Interest
Evaluation, impact analysis, website analytics	72%
Adapting technologies	67%
Web/visual/graphic design	67%
Choosing technologies	66%
Usability Services	66%
Visualization	65%
Developing open-source software	64%
Support for education	64%
Community engagement mechanisms	62%
Keeping your project running	62%
Legal perspectives	61%
Managing data	60%
Computational resources	59%
Mobile technology	59%
Database structure, optimization, and query expertise	59%
Data mining and analysis	58%
Cybersecurity consultation	57%
Website construction	57%
Software engineering process consultation	53%
Source code review and/or audit	51%
High-bandwidth networks	45%
Scientific instruments or data streams	44%
Management aspects of a project	38%



- Content management systems (Drupal)
- Libraries for implementation (Django)
- Portal frameworks (Liferay)
- Science gateway frameworks (WS-PGRADE, Galaxy)
  - Static layout
  - Layout extendable
  - Workflow-enabled
- APIs for implementation (Apache Airavata, Agave)





## VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

GO

LOGIN

ABOUT

ORGANISMS

DOWNLOADS

TOOLS

DATA

HELP

COMMUNITY

CONTACT US

## Welcome to VectorBase!

VectorBase is an NIAID Bioinformatics Resource Center dedicated to providing data to the scientific community for Invertebrate Vectors of Human Pathogens. We aim to provide a forum for the discussion and distribution of news and information relevant to invertebrate vectors, as well as access to tools to facilitate the querying and analysis of the data sets presented on this site.

DATA



GENOMES



TRANSCRIPTS &  
TRANSCRIPTOMES



PROTEINS &  
PROTEOMES

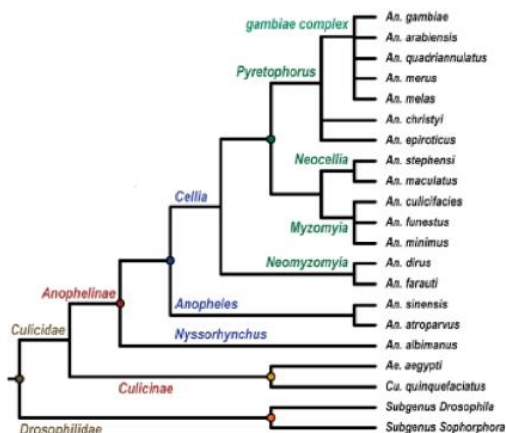


MITOCHONDRIAL  
SEQUENCES



POPULATION  
BIOLOGY

## TOOLS & RESOURCES



### First pass annotation for nine Anopheline species available

VectorBase and The Anopheles Genomes Cluster announce the first pass annotation of nine Anopheline genomes. The predictions were generated using *ab initio* and similarity approaches utilising transcriptome data and taxonomically informative proteomes. Gene models were aggregated using the MAKER system. These gene sets are available for browsing, searching via BLAST and download.

*An. albimanus*

*An. christyi*

*An. epiroticus*

*An. minimus*

*An. stephensi*

*An. arabiensis*

*An. dirus*

*An. funestus*

*An. quadriannulatus*

## Want to see your BLAST, ClustalW and HMMer jobs?

Login above or Register here.

## POPULAR ORGANISMS



*Anopheles gambiae*



*Aedes aegypti*



*Culex quinquefasciatus*

## RECENT ADDITIONS



*Anopheles funestus*



*Phlebotomus papatasi*



*Biomphalaria glabrata*

All organisms

## LATEST NEWS

August 14, 2013

VectorBase Release VB-2013-08

June 28, 2013

VectorBase Release VB-2013-06

[More news](#)

## DID YOU KNOW?

### New search engine at VectorBase

Searching via the box at the top of all pages now lets you find more than just genes! Most site content is now searchable. ... From Newsletter 13 (Sep





## VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

GO

LOGIN

ABOUT

ORGANISMS

DOWNLOADS

**TOOLS**

DATA

HELP

COMMUNITY

CONTACT US

Home » Tools

### BLAST

Due to browser compatibility problems in Safari and Firefox, we recommend using **Google Chrome** when using blast. We are working on these issues and apologize for any inconvenience this may cause you.

Paste your sequences here

#### Upload FASTA File

Browse...

No file selected.

#### Program

☒ blastn

☐ tblastn

☐ tblastx

☐ blastp

☐ blastx

blastn - Nucleotide vs. Nucleotide

#### Job Control

Load results

Job ID

RESET

SUBMIT

#### Datasets

☐ All Datasets

☐ *Aedes aegypti*

☐ *Anopheles albimanus*

☐ *Anopheles arabiensis*

☐ *Anopheles christyi*

☐ *Anopheles darlingi*

☐ *Anopheles dirus A*

☐ *Anopheles epiroticus*

#### Options

Maximum E-Value

1

Word Size

11

Complexity Masking

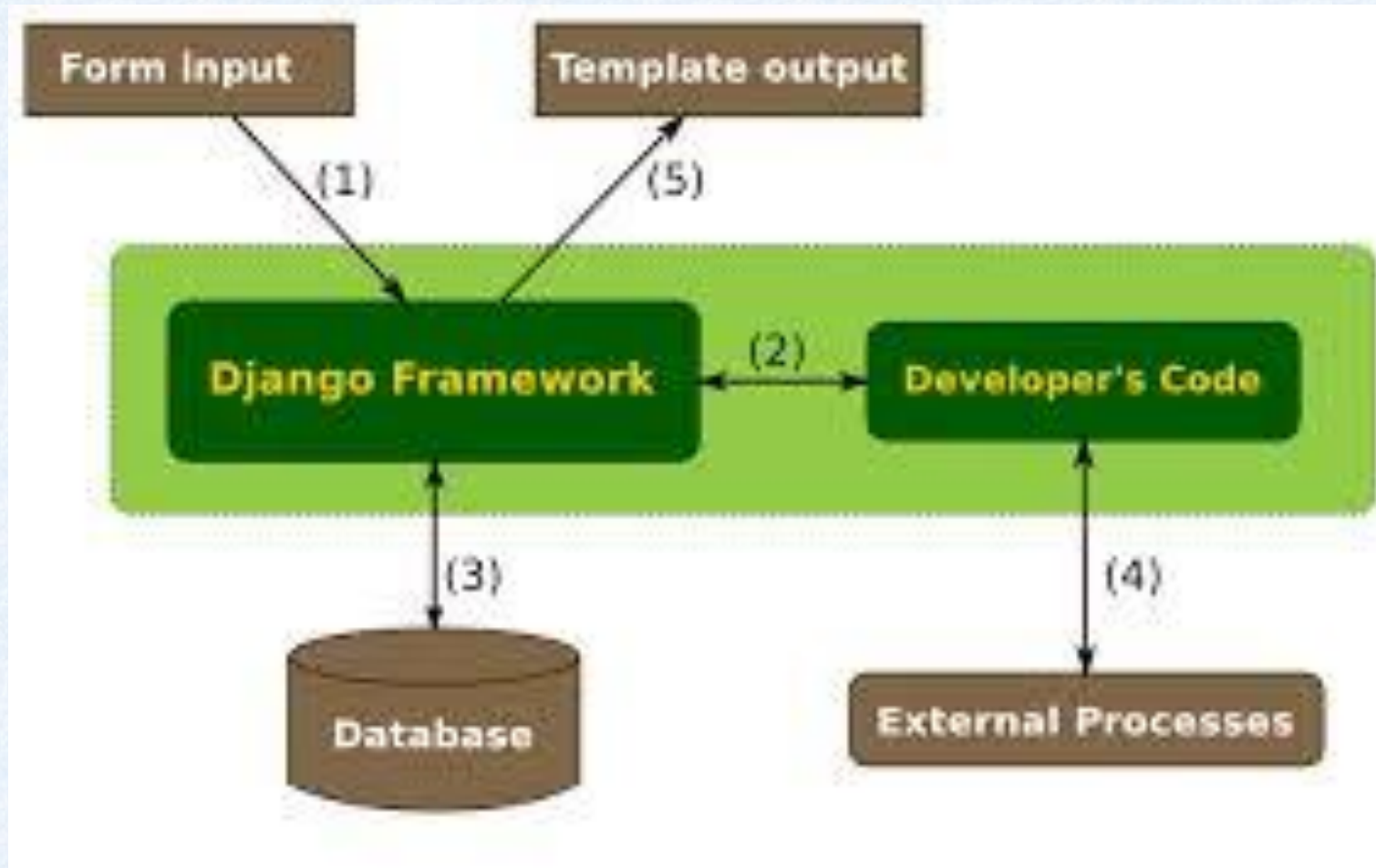
Default

Results per Query per Database

50

Tool

The screenshot displays the VectorBase Bioinformatics Resource website. At the top, there is a search bar with the text "Enter search terms" and a "GO" button. Below the search bar is a navigation menu with tabs for "DOWNLOADS", "TOOLS", "DATA", "HELP", and "COMMUNITY". The "TOOLS" tab is currently selected, and a dropdown menu is open, listing the following tools: **BLAST**, **ClustalW**, **HMMer**, **BioMart**, **Genome browser**, **Ontology browser**, **Expression browser**, **Insecticide resistance**, and **Population biology browser**. The main content area on the left is titled "BLAST" and includes a text input field for "Paste your sequences here", an "Upload FASTA File" section with a "Browse..." button, and a "Program" dropdown menu set to "blastn". Below this is a "Datasets" section with a list of mosquito species, each with a checkbox: *Aedes aegypti*, *Anopheles albimanus*, *Anopheles arabiensis*, *Anopheles christyi*, *Anopheles darlingi*, *Anopheles dirus A*, and *Anopheles epiroticus*. On the right side of the main content area, there are additional settings: "Maximum E-Value" (set to 1), "Word Size" (set to 11), "Complexity Masking" (set to Default), and "Results per Query per Database" (set to 50). Buttons for "RESET" and "SUBMIT" are visible at the bottom right of the form.





## VECNet

### Vector Ecology and Control Network

#### Our Work

Though malaria remains both treatable and preventable, 350-500 million people worldwide are infected with the disease every year, with up to one million cases ending in death. Nearly 85 percent of the victims who die are younger than five years old.

Recent global efforts have contributed to declines in malaria-related sickness and death, but while the present available tools for controlling malaria are effective, they will not by themselves eliminate the disease. There is a need for new strategies to eliminate malaria.



VECNet

VECNet is a [consortium of institutions](#) assembled to address the need for new strategies to eliminate malaria, which requires an understanding of how interventions affect the transmission of the disease across different geographic areas where the mosquitoes that transmit malaria differ in their behavior.



Follow @VECNetNews

#### VECNet Alpha Release August 14, 2013

The VECNet website is currently in "Alpha Release" while the VECNet team tests its design and functionality. If you received an invitation to be an Alpha tester, please [request an account here](#). If you are interested in helping to test the "Beta Release" expected later this year, please [register](#) and indicate that you would like to be a Beta tester.

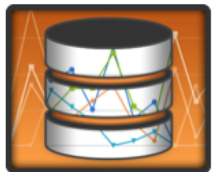
VECNet enables national malaria control managers, researchers, product developers, funding bodies and policy makers to ask questions such as: *‘What is the intensity and type of intervention/s required to achieve one malaria death per 100,000 in this particular population?’* and *‘What is the impact on malaria transmission of a new tool that potentially kills a mosquito any time it takes a sugar meal, seeks a blood meal or lays eggs?’*

To find answers to these questions, VECNet is developing three resources: the **Digital Library**, the **Data Warehouse Browser** and a **Modeling Platform**. These tools provide users with access both to data on malaria transmission and to modeling software to create simulations of various scenarios. The simulations use geospatially specific data to analyze the potential of different combinations of control interventions to reduce malaria transmissions in a given area.



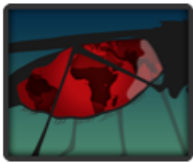
## Digital Library

The Digital Library assembles all published and unpublished data on malaria vectors. This extensive database enables the analysis of transmission risk as a function of vector ecology and behavior via the Modeling Platform.



## Data Warehouse Browser

The Data Warehouse Browser is an incentive-based platform for data sharing, and enables easy-to-use, secure access to data for use with any of the Modeling Platform tools. It presents research data in ways that allow model simulations with specific data in geographically defined areas. Users can access all existing data, including their own model input and output data.



## Transmission Simulator

Researchers use their data as input to model the sensitivity of transmission to changes in the behaviors of vectors resulting from responses to interventions or changes in the environment/ecology.



## Risk Mapper

Risk Mapper analyzes the impact of particular interventions on malaria. National malaria control programs can use it to compare the distribution of vector control interventions against the distribution of specific vector species to determine the appropriateness of the intervention.



## Product Impact Evaluator (PIE)

Investors use PIE to estimate the value of new control tools. With PIE, product developers estimate the effect of candidate tools on malaria transmission and can then develop and refine Target Product Profiles to achieve a desired level of impact.



## Computational Intervention portFolio EvaluatoR (CIFER)

CIFER is an amalgamation of the outputs of Transmission Simulator, Risk Mapper, and PIE, combining the estimates of contributions from individual vector species, individual geographies, and individual interventions to overall malaria transmission. Policymakers can refine the suite of tools needed to achieve malaria eradication by analyzing how interventions affect the transmission of the disease and, as importantly, where interventions will not be able to achieve effective control and elimination.



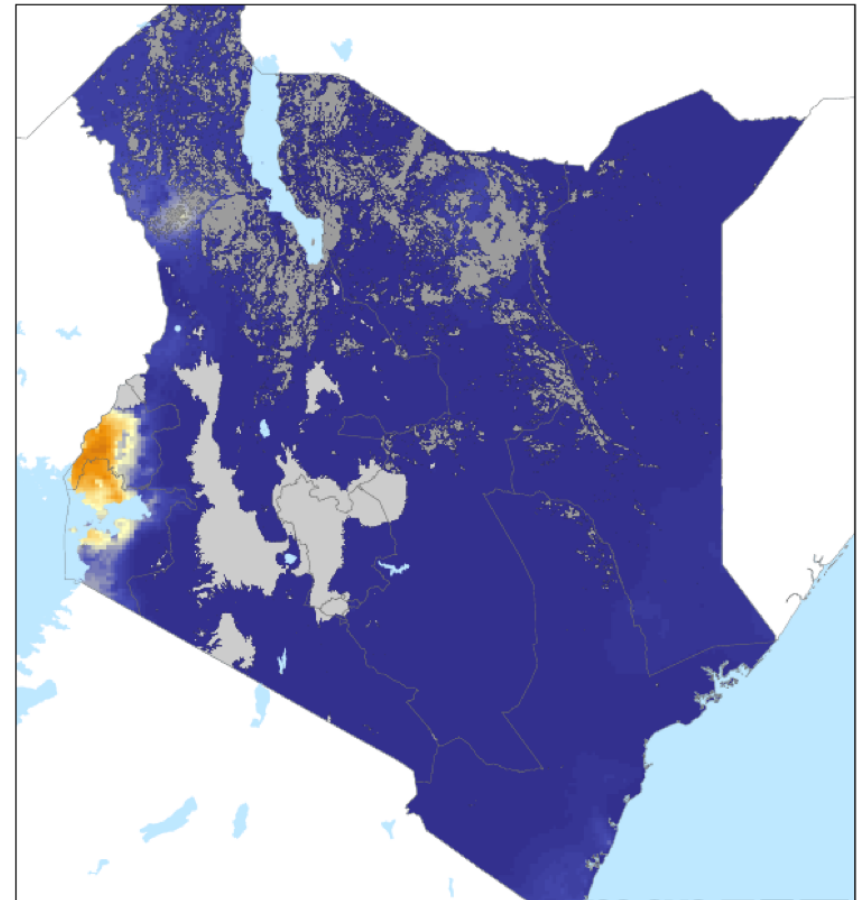
[Home page](#)

## BUILD SCENARIO

[Start page](#) >[Location](#) >[Model Parameters](#) >[Dominant vectors](#) >[Behavior/Habitat](#) >[Baseline Transmission](#) >[Interventions](#) >[Summary](#) >[Risk Mapper](#) / [Build Scenario \(Kenya\)](#) / [Baseline Transmission](#)

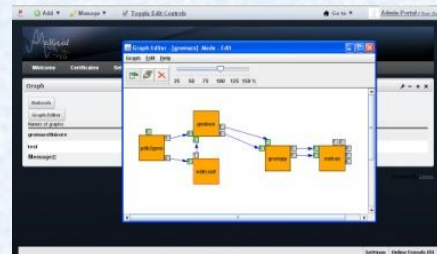
## Step 6 of 8 : Baseline Transmission

Annual EIR data, Courtesy: Malaria Atlas Project



## Portal framework

- Authentication (e.g., OpenSSO, CAS)
- Authorization
- Standards compliant
  - JSR168/286
  - Web services
  - Web 2.0 websites
- Web Publishing and Shared Workspaces
- Collaboration
- Social Networking



**User Interface**  
WS-PGRADE  
Liferay

**Workflow  
storage**

**Application  
repository**

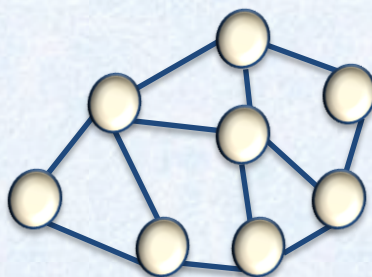
**Information  
system**

**High-Level  
Middleware  
Service Layer**  
gUSE

**Workflow  
engine**

**Submitters**

**Logging**

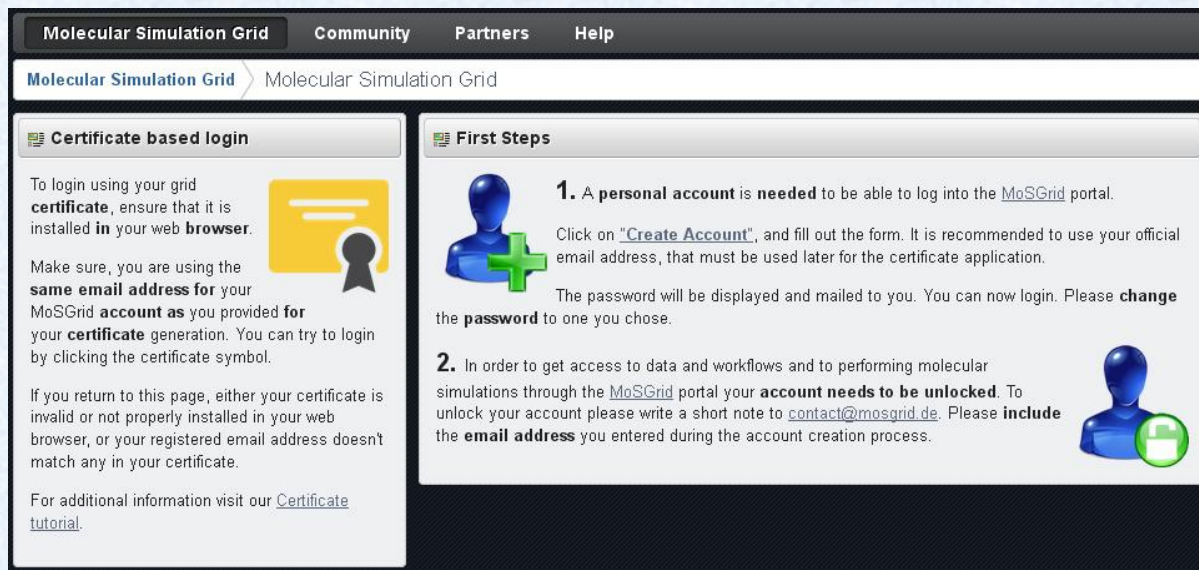


**DCI Resources**  
**Middleware Layer**



## Molecular Simulation Grid

- Science gateway integrated with underlying compute and data management infrastructure
- Distributed workflow management
- Data repository
- Metadata management



The screenshot shows the Molecular Simulation Grid portal interface. At the top is a navigation bar with links: Molecular Simulation Grid, Community, Partners, and Help. Below this is a breadcrumb trail: Molecular Simulation Grid > Molecular Simulation Grid. The main content area is divided into two panels. The left panel, titled 'Certificate based login', provides instructions on how to log in using a grid certificate, emphasizing the need for a valid certificate and the correct email address. It includes a yellow icon of a certificate and a key. The right panel, titled 'First Steps', contains two numbered instructions. Step 1 explains that a personal account is needed to log in, with a blue person icon and a green plus sign. Step 2 explains that the account needs to be unlocked, with a blue person icon and a green lock icon. Both panels include links to 'Create Account' and 'contact@mosgrid.de'.

**Molecular Simulation Grid** Community Partners Help

Molecular Simulation Grid > Molecular Simulation Grid

**Certificate based login**

To login using your grid **certificate**, ensure that it is installed in your web **browser**.

Make sure, you are using the **same email address** for your MoSGrid **account** as you provided for your **certificate** generation. You can try to login by clicking the certificate symbol.

If you return to this page, either your certificate is invalid or not properly installed in your web browser, or your registered email address doesn't match any in your certificate.

For additional information visit our [Certificate tutorial](#).

**First Steps**

**1. A personal account is needed** to be able to log into the [MoSGrid](#) portal.

Click on "[Create Account](#)", and fill out the form. It is recommended to use your official email address, that must be used later for the certificate application.

The password will be displayed and mailed to you. You can now login. Please **change** the **password** to one you chose.

**2. In order to get access to data and workflows and to performing molecular simulations through the [MoSGrid](#) portal your account needs to be unlocked.** To unlock your account please write a short note to [contact@mosgrid.de](mailto:contact@mosgrid.de). Please **include** the **email address** you entered during the account creation process.

[Add](#) [Manage](#) [Edit Controls](#)

[Go to](#) [Sandra Gesing](#) (Sign Out)



[Welcome](#) [Workflow](#) [Storage](#) [Settings](#) **[Security](#)** [Statistics](#) [Information](#) [Data Avenue](#) [Publications](#) [Help](#) [End User](#)

Coming Soon

[WS-PGRADE gUSE](#) [Security](#) [Assertion](#)

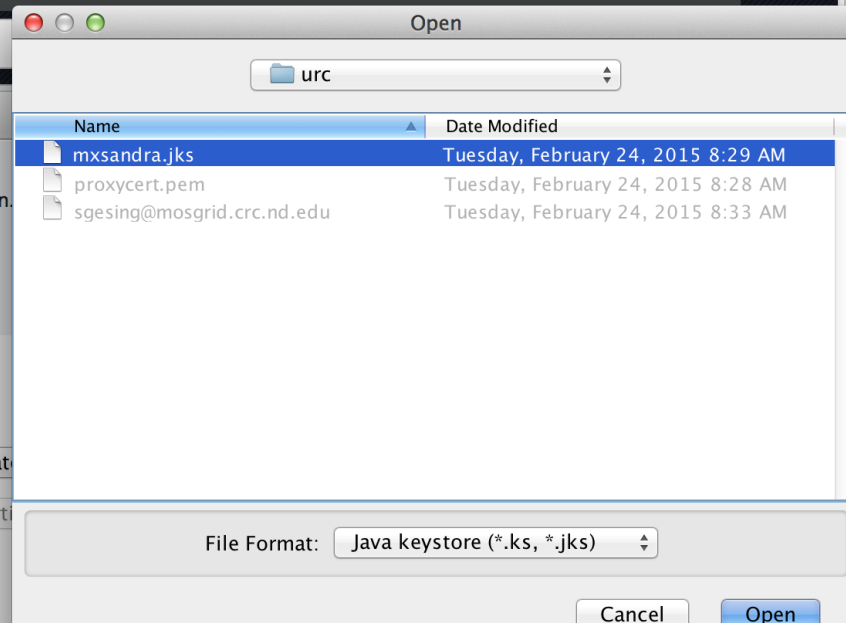
## Assertion

You have to generate an assertion.

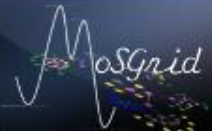
### Generation of new assertion:

User Certificate (.p12):  [Select Certificate](#)

Validity (in days):  [Generate assertion](#)



[Add](#) | [Manage](#) | ☒ [Toggle Edit Controls](#) | [Go to](#) | [Admin Portal](#) (Sign Out)



[Welcome](#) | [Certificates](#) | [Set](#)

**Graph**

Refresh

Graph Editor

Names of graphs

**gromacsUnicore**

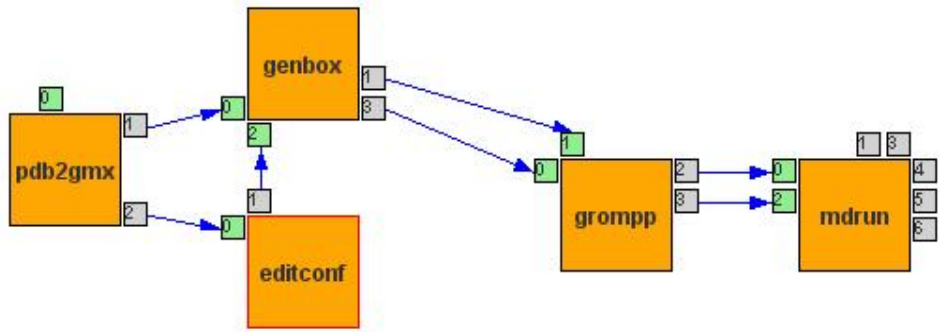
test

Message:[]

**Graph Editor - [gromacs] Mode - Edit**

Graph Edit Help

25 50 75 100 125 150 %



```

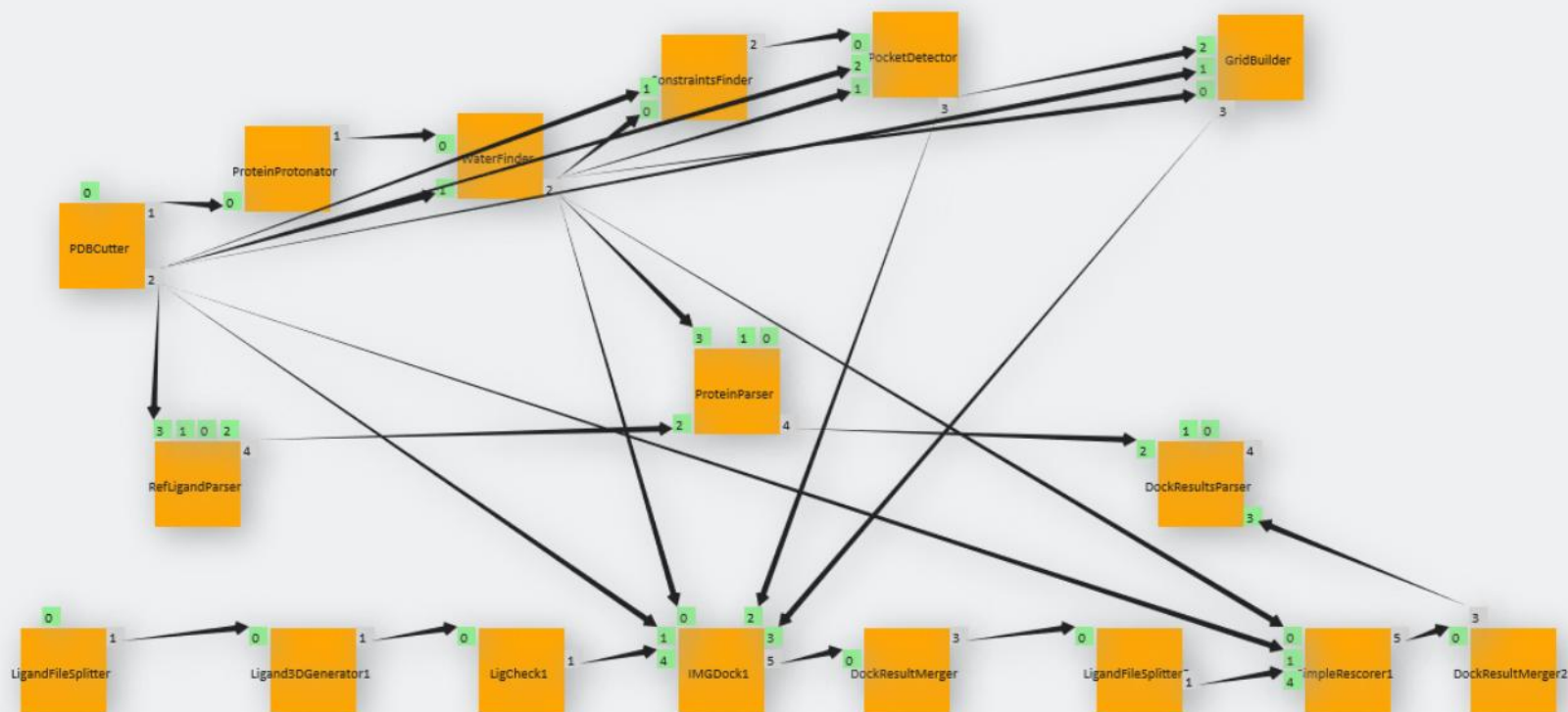
graph LR
    pdb2gmx[pdb2gmx] -- 1 --> genbox[genbox]
    pdb2gmx -- 2 --> editconf[editconf]
    genbox -- 1 --> grompp[grompp]
    genbox -- 8 --> grompp
    editconf -- 1 --> grompp
    grompp -- 1 --> mdrun[mdrun]
    grompp -- 8 --> mdrun
    
```

The diagram illustrates a workflow for GROMACS simulation setup. It starts with **pdb2gmx**, which outputs 1 unit to **genbox** and 2 units to **editconf**. **genbox** then outputs 1 unit to **grompp** and 8 units to **grompp**. **editconf** outputs 1 unit to **grompp**. Finally, **grompp** outputs 1 unit to **mdrun** and 8 units to **mdrun**.

Powered By [Liferay](#)

Settings | Online Friends (0)






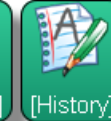




MoSGrid / Generic Workflows / Concrete


ph:   
 plat


**Job's name:** PDBCutter


**Optional note:** Description of Job


 [Job Executable]
  [Job I/O]
  [JDL/RSI]
  [History]

**WorkflowService Binary**  

**Type:** uncore 

**Grid:** flavus.informatik.uni-tuebingen.de:8090 

**Tools:** PDBCutter 1.0.0 

**Execute parser:** ModelCreator 1.0.0 

**Replicate settings in all Jobs:** MolCombine 1.0.0

**Copy job names to tools:** MolDepict 1.0.0

**Kind of binary:** MolFilter 1.0.0

**MPI Node Number:** MolPredictor 1.0.0

**Executable code of binary:** nwchem 6.1

**Parameter:** obabel (OpenBabel) 2.3.1

PartialChargesCopy 1.0.0

pdb2gmx 4.5.5

**PDBCutter 1.0.0**

PDBDownload 1.0.0

Perl 5.8.8

PocketDetector 1.0.0

POVRay 3.5

Predictor 1.0.0

PropertyModifier 1.0.0

PropertyPlotter 1.0.0

ProteinCheck 1.0.0



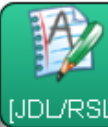
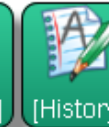
ProteinProtonator 1.0.0

Python Script 2.4.2

2012-11-21



✕

**Job's name:** ParserProtein  
**Optional note:** Description of Job

 [Job Executable]
  [Job I/O]
  [JDL/RSL]
  [History]





?



Port Number:0   Port Name: genparser   Description of Port

**Input Port's Internal File Name:** genparser.jar  

**Port dependent condition allowing the run of the job:** ☐ View ☒ Hide

**Source of input directed to this port:**

 xtreemfs://test/genparser.jar  
☒ Copy to WN:  

**Parametric Input details:** ☐ View ☒ Hide

Port Number:1   Port Name: startscript   Description of Port



WS-PGRADE gUSE Workflow Concrete



Refresh

Workflow name	Running	Finished	Error	Suspended	Actions					
<b>NWChem-specWF_2015-03-05-092900</b> 2014-4-9	0	3	0	0	Configure	Info	Details	Submit	Delete	Export
<b>NWChem_Mull-part_2015-03-05-092900</b> 2014-4-9	0	5	0	0	Configure	Info	Details	Submit	Delete	Export
<b>NWChem_TD-part_2015-03-05-092900</b> 2014-4-9	0	5	0	0	Configure	Info	Details	Submit	Delete	Export
<b>NWChem_basic-part_2015-03-05-092900</b> 2014-4-9	0	8	3	0	Configure	Info	Details	Submit	Delete	Export
<b>NWChem_freq-part_2015-03-05-092900</b> 2014-4-9	0	5	0	0	Configure	Info	Details	Submit	Delete	Export
<b>NWChem_solv-partc_2015-03-05-092900</b> 2014-4-9	0	3	0	0	Configure	Info	Details	Submit	Delete	Export
<b>hello_2015-02-07-114739</b> 2015-2-7	0	2	0	0	Configure	Info	Details	Submit	Delete	Export

MoSGrid Portal

Workflows


Concrete

2011-1-31 14:24	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-1-13 14:53	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-1-17 12:0	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-2-9 9:34	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-1-18 9:40	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-2-1 14:44	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-2-7 18:55	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-2-15 9:21	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-1-14 10:38	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-1-18 10:13	finished		<a href="#">Details</a>	<a href="#">Delete</a>
2011-2-10 12:56	finished		<a href="#">Details</a>	<a href="#">Delete</a>

Selected WF Instance:

2011-2-10 12:56

Job	Status	Instances	[ Actions ]
cell	finished	1	<a href="#">View finished</a> <a href="#">Hide</a>



Sorting method: 

Method 1

 Range: 

20

 From: [Refresh](#)

PID	Resource	Status	View info
0	<a href="https://unicore6-bisgrid.uni-paderborn.de:8080/bisgrid/services/JobManagement?res=0106fb75-d006-488b-a97a-6f8c60c25daa">https://unicore6-bisgrid.uni-paderborn.de:8080/bisgrid/services/JobManagement?res=0106fb75-d006-488b-a97a-6f8c60c25daa</a>	finished	<a href="#">Logbook</a> <a href="#">std. Output</a> <a href="#">std. Error</a> <a href="#">Download file output</a>

Powered By [Liferay](#)

## Molecular Dynamics

- Study and simulation of molecular motion

## Quantum Chemistry

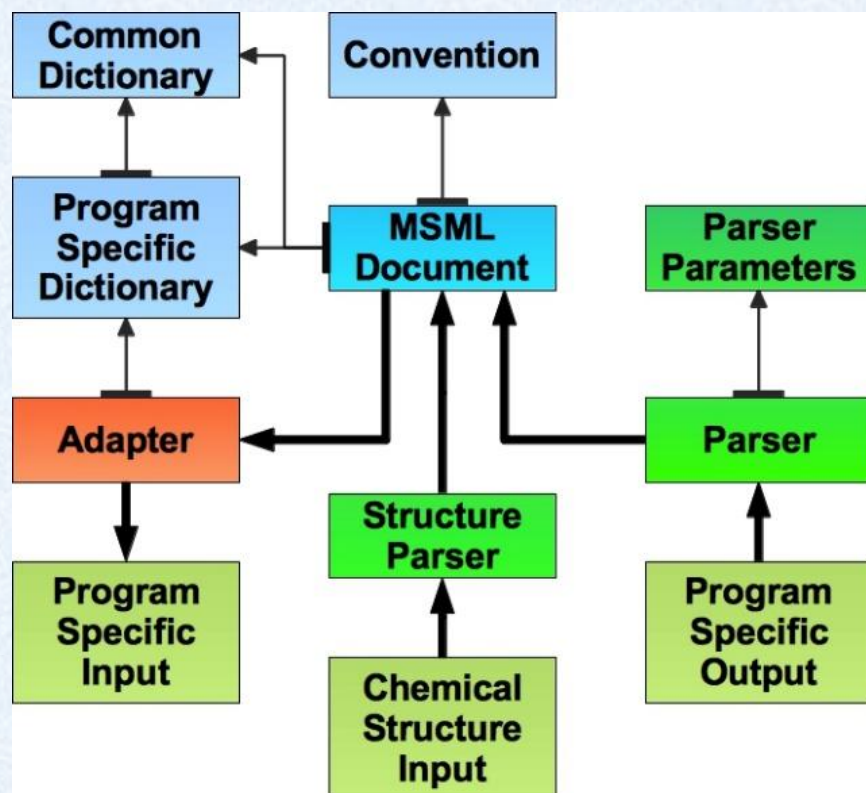
- Study and simulation of molecular electronic behavior relative to their chemical reactivity

## Docking

- Main focus on evaluation of ligand-receptor interactions (e.g., for drug design)



- Molecular Simulation Markup Language (MSML)
- CML compliant
- Template for each and every workflow
  - Molecular input
  - Domain specific tools
  - Job configuration
  - Optimized structures, trajectories, energies, ...
- Semantic search (Apache Lucene)



```

<?xml version="1.0" encoding="UTF-8"?>
<cml convention="convention:compchem">
  <module dictRef="compchem:jobList">
    <cmlx:parserConfiguration/>
    <module dictRef="compchem:job" id="Job1">
      <module dictRef="compchem:environment">
        <propertyList/>
      </module>
      <module dictRef="compchem:initialization">
        <parameterList/>
        <cmlx:adapterConfiguration/>
        <cmlx:parserConfiguration/>
      </module>
      <module dictRef="compchem:finalization">
        <propertylist/>
      </module>
    </module>
    <module dictRef="compchem:job" id="Job2"/>
    ...
  </module>
</cml>
  
```

```
<module dictRef="compchem:initialization">
  <parameterList>
    <parameter dictRef="g09:loglevel">
      <scalar dataType="xsd:string" units="si:none">p</scalar>
    </parameter>
    <parameter dictRef="g09:jobtype">
      <scalar dataType="xsd:string" units="si:none">opt</scalar>
    </parameter>
    <parameter dictRef="g09:hf.theory" cmlx:editable="true">
      <scalar dataType="xsd:string" units="si:none">hf</scalar>
    </parameter>
    <parameter dictRef="g09:basisset" cmlx:editable="true">
      <scalar dataType="xsd:string" units="si:none">6-31G</scalar>
    </parameter>
    <parameter dictRef="g09:formal.charge" cmlx:editable="true">
      <scalar dataType="xsd:integer" units="si:none">0</scalar>
    </parameter>
    <parameter dictRef="g09:spin" cmlx:editable="true">
      <scalar dataType="xsd:integer" units="si:none">1</scalar>
    </parameter>
    <parameter dictRef="g09:checkpointfile">
      <scalar dataType="xsd:string" units="si:none">job.chk</scalar>
    </parameter>
  </parameterList>
  <cmlx:adapterConfiguration adapterID="g09adap"
    fileExtension="com" portName="job.com"/>
</module>
```

Quantum Chemistry Portlet

Import

Submission

Monitoring

Welcome

Welcome to the Quantum Chemistry portlet.

Import a workflow

Toolsuite

Gaussian 09

Workflow \*

Optimization with DFT methods

Submission of prepared job-files. (Yields formatted checkpoint file)

Optimization with DFT methods

Optimization with HF methods

Optimization + frequency calculations with DFT methods

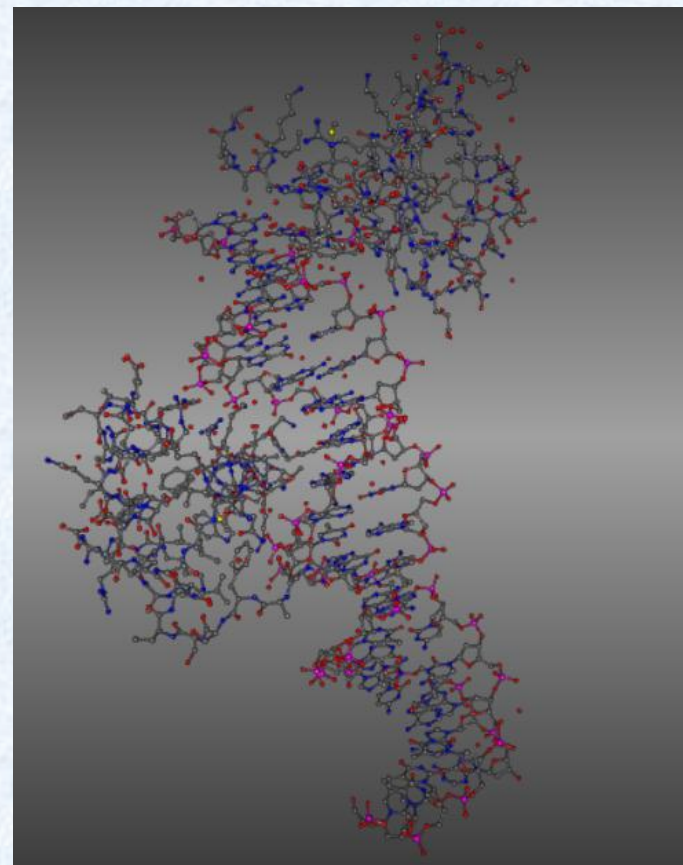
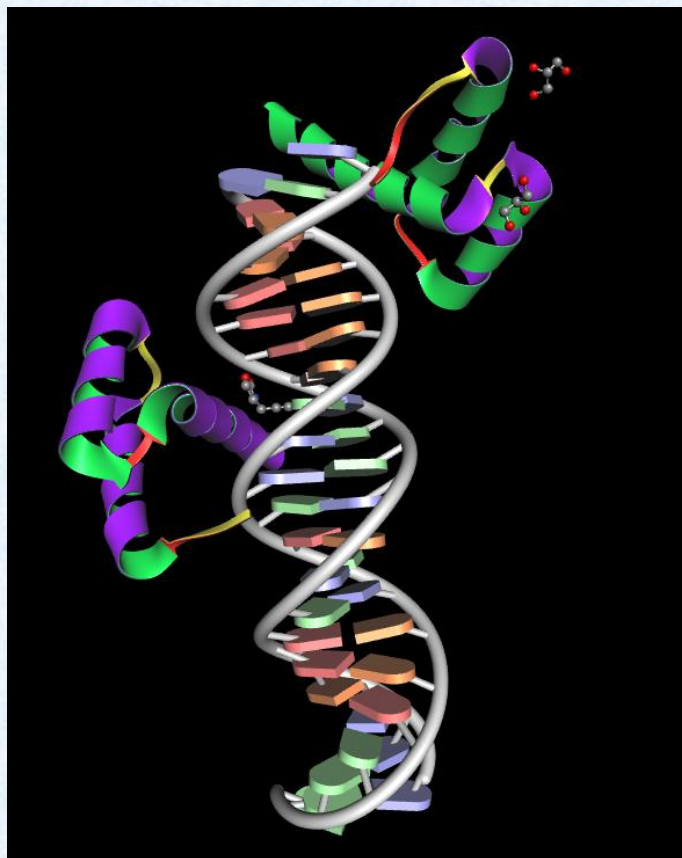
Submission of prepared job-files. (No postprocessing)

Generate cube files for the visualization of HOMO and LUMO

Submission of prepared job-files. (No postprocessing)

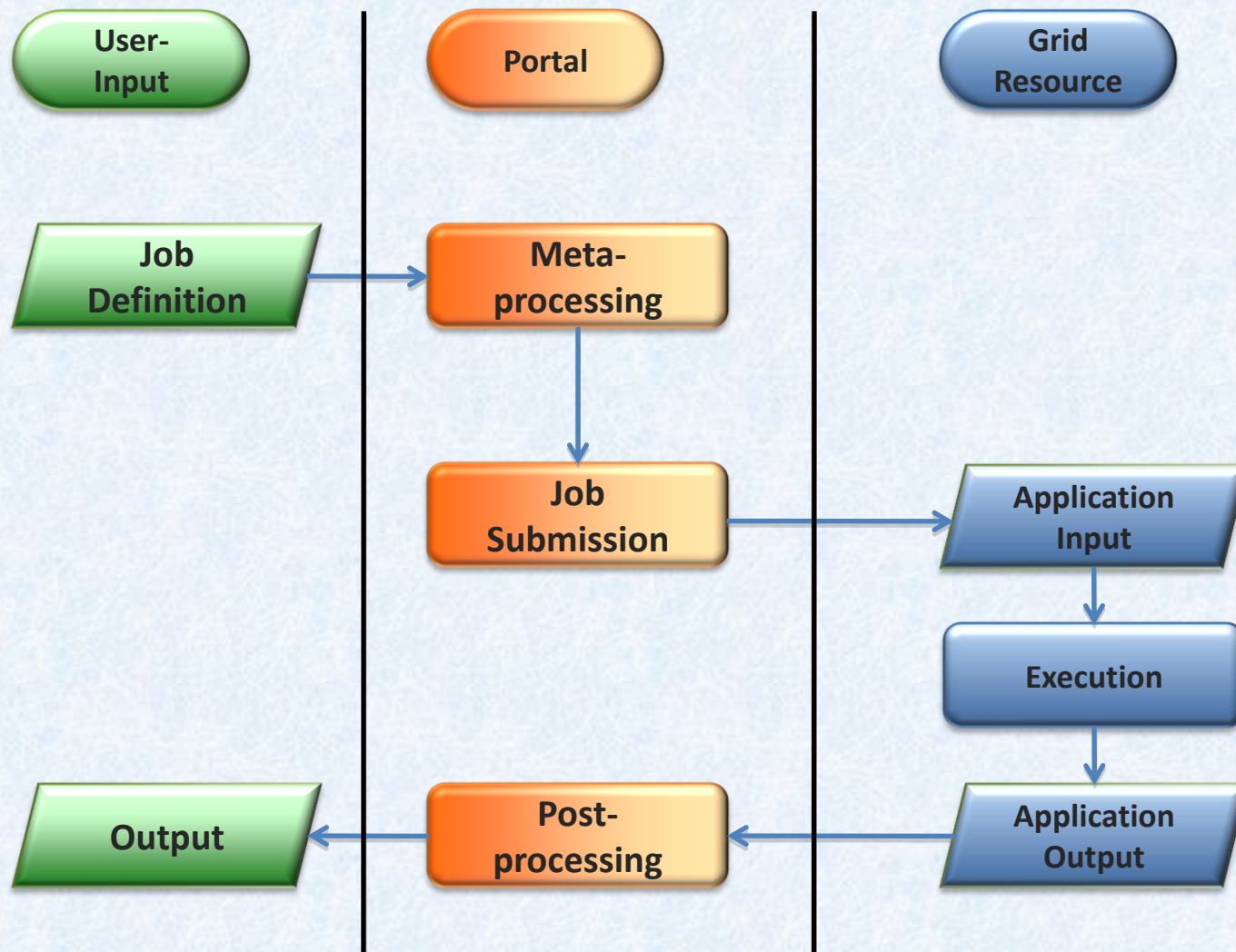


## Testing of ChemDoodle and MolCAD



[web.chemdoodle.com](http://web.chemdoodle.com)

[molcad.de](http://molcad.de)



**QCPortletVAPI**

Import Submission Monitoring About

**Welcome**


Welcome to the Quantum Chemistry portlet.

**Import a workflow**

Please select a toolsuite  
Gaussian 09

Please select a workflow  
G09Minimal

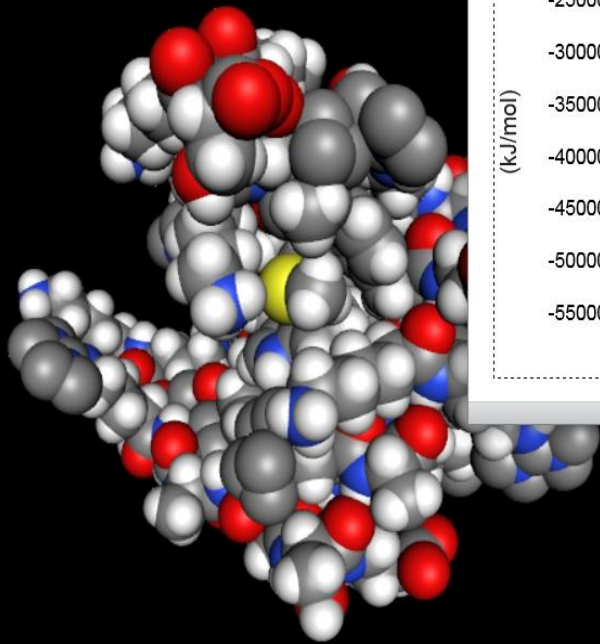
Please enter a name  
TEGqu\_26

 Import

Gaussian optimisation and invocation of parsertools.

- Specialised interface for quantum chemistry software (Gaussian, NWChem, ORCA)
- Basic workflows
- Easy Generation or Uploading of Input Files
- Parsing of result files





## Docking Portlet

Import Submission Monitoring About ?

### Select an imported instance

Import

StandardDockingWorkflow\_2012-03-30-125439\_29.0

### Please fill the input mask to submit your workflow

#### PDBCutter

Filename \*

1DX6.pdb

Upload PDB

PDB Model \*

Model 0

☒ Chain A

☐ Chain S

Chain name of ligand \*

A

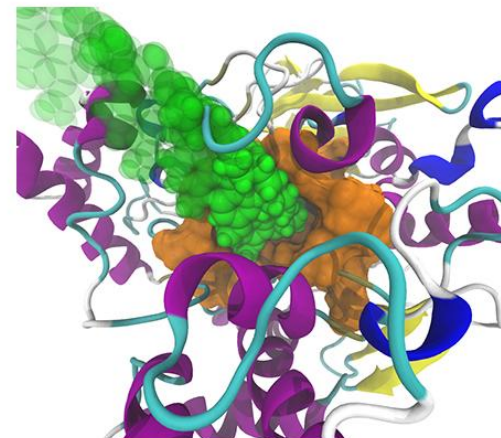
Name of ligand as stated in pdb file \*

GNT

Protein Chains that are to be deleted

Select a protein chain from your PDB input file to act as receptor (secondary structure) including the binding pocket (orange).

Specify a reference ligand (green) by it's three letter code including the corresponding chain. It might be necessary to open the input PDB file with an editor. This information is required for the identification of the binding site and the calculation of RMSD values.



Docking Portlet

Import

Standard

Select an i

Import

Standard

Please fill

PDB

File name

1DX6.j

PDB Mo

Model

☒ Chain

☐ Chain

Chain n

A

Name c

GNT

Protein

Docking-Portlet

Welcome

Monitoring

Debug

Workflows:

1EVE

TEST20111025

TEST20111025

1EVE

Docking

recH.pdb

results.sorted.sdf

1EVE/ STATUS: FINISHED / selected

ATOM	1	N	SER	A
ATOM	2	CA	SER	A
ATOM	3	C	SER	A
ATOM	4	O	SER	A
ATOM	5	CB	SER	A
ATOM	6	OG	SER	A
ATOM	7	HN1	SER	A
ATOM	8	HN2	SER	A
ATOM	9	HN3	SER	A
ATOM	10	HA	SER	A
ATOM	11	HB1	SER	A
ATOM	12	HB2	SER	A
ATOM	13	HG	SER	A
ATOM	14	N	GLU	A
ATOM	15	CA	GLU	A
ATOM	16	C	GLU	A
ATOM	17	O	GLU	A
ATOM	18	CB	GLU	A
ATOM	19	CG	GLU	A
ATOM	20	CD	GLU	A
ATOM	21	OE1	GLU	A
ATOM	22	OE2	GLU	A
ATOM	23	HN	GLU	A
ATOM	24	HA	GLU	A
ATOM	25	HB1	GLU	A
ATOM	26	HB2	GLU	A
ATOM	27	HG1	GLU	A
ATOM	28	HG2	GLU	A

Update

Delete

Download

View in Jmol

Jmol

1320400625949recH.pdb

Jmol

-7.010 86.637 39.078 1.00 0.00 H



## Tools

[Get Data](#)  
[Send Data](#)  
[ENCODE Tools](#)  
[Lift-Over](#)  
[Text Manipulation](#)  
[Filter and Sort](#)  
[Join, Subtract and Group](#)  
[Convert Formats](#)  
[Extract Features](#)  
[Fetch Sequences](#)  
[Fetch Alignments](#)  
[Get Genomic Scores](#)  
[Operate on Genomic Intervals](#)  
[Statistics](#)  
[Wavelet Analysis](#)  
[Graph/Display Data](#)  
[Regional Variation](#)  
[Multiple regression](#)  
[Multivariate Analysis](#)  
[Evolution](#)  
[Motif Tools](#)  
[Multiple Alignments](#)  
[Metagenomic analyses](#)  
[FASTA manipulation](#)



Hello world! It's running...

To customize this page edit [static/welcome.html](#)



Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

## History

### test history

8.8 MB

**44: Main output for creating files (Test3.txt)**

1 line  
format: text, database: ?

Test3.txt

**43: Main output for creating files (Test2.fasta)**

1 line  
format: text, database: ?

Test2.fasta

**42: Main output for creating files (Test1.txt)**

1 line  
format: text, database: ?

Test1.txt

**41: Main output for**

## Compute sequence length (version 1.0.0)

Compute length for these sequences:

2:  ▾

How many title characters to keep?:

'0' = keep the whole thing

**Execute**

## What it does

This tool counts the length of each fasta sequence in the file. The output file has two columns per line (separated by tab): fasta titles and lengths of the sequences. The option *How many characters to keep?* allows to select a specified number of letters from the beginning of each FASTA entry.

## Example

Suppose you have the following FASTA formatted sequences from a Roche (454) FLX sequencing run:

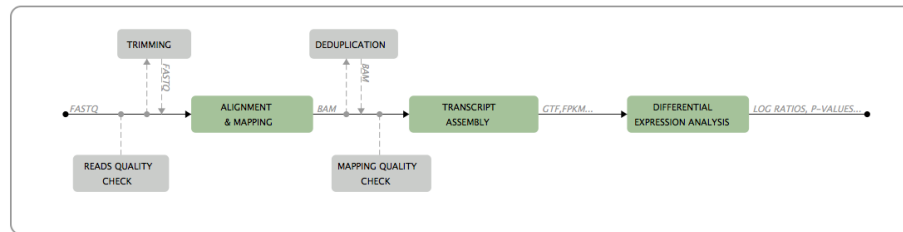
```
>EYKX4VC02EQLO5 length=108 xy=1826_0455 region=2 run=R_2007_11_07_16_15_57_
TCCGCGCCGAGCATGCCCATCTTGGATTCCGGCGCGATGACCATCGCCCGCTCCACCACG
TTCGGCCGGCCCTTCTCGTCGAGGAATGACACCAGCGCTTCGCCCACG
>EYKX4VC02D4GS2 length=60 xy=1573_3972 region=2 run=R_2007_11_07_16_15_57_
AATAAACTAAATCAGCAAAGACTGGCAAATACTCACAGGCTTATACAATACAAATGTAAfa
```

Running this tool while setting **How many characters to keep?** to 14 will produce this:

```
EYKX4VC02EQLO5 108
EYKX4VC02D4GS2 60
```

View a [list of supported genomes](#) from [EuPathDB](#), [PATRIC](#), and [VectorBase](#).

Have a question? Contact the [Pathogen Portal Team](#)



Choose an activity below



**Uploads** **Login to get started**

From my computer or a URL

From ENA/SRA



**Quality Control**

How good are my base calls?

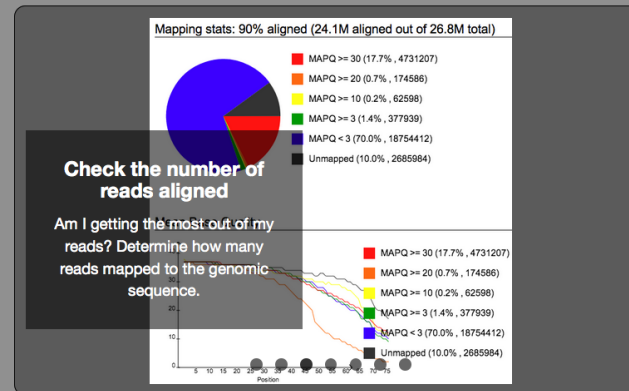
Trim low quality sequence

Are all my reads mapped?



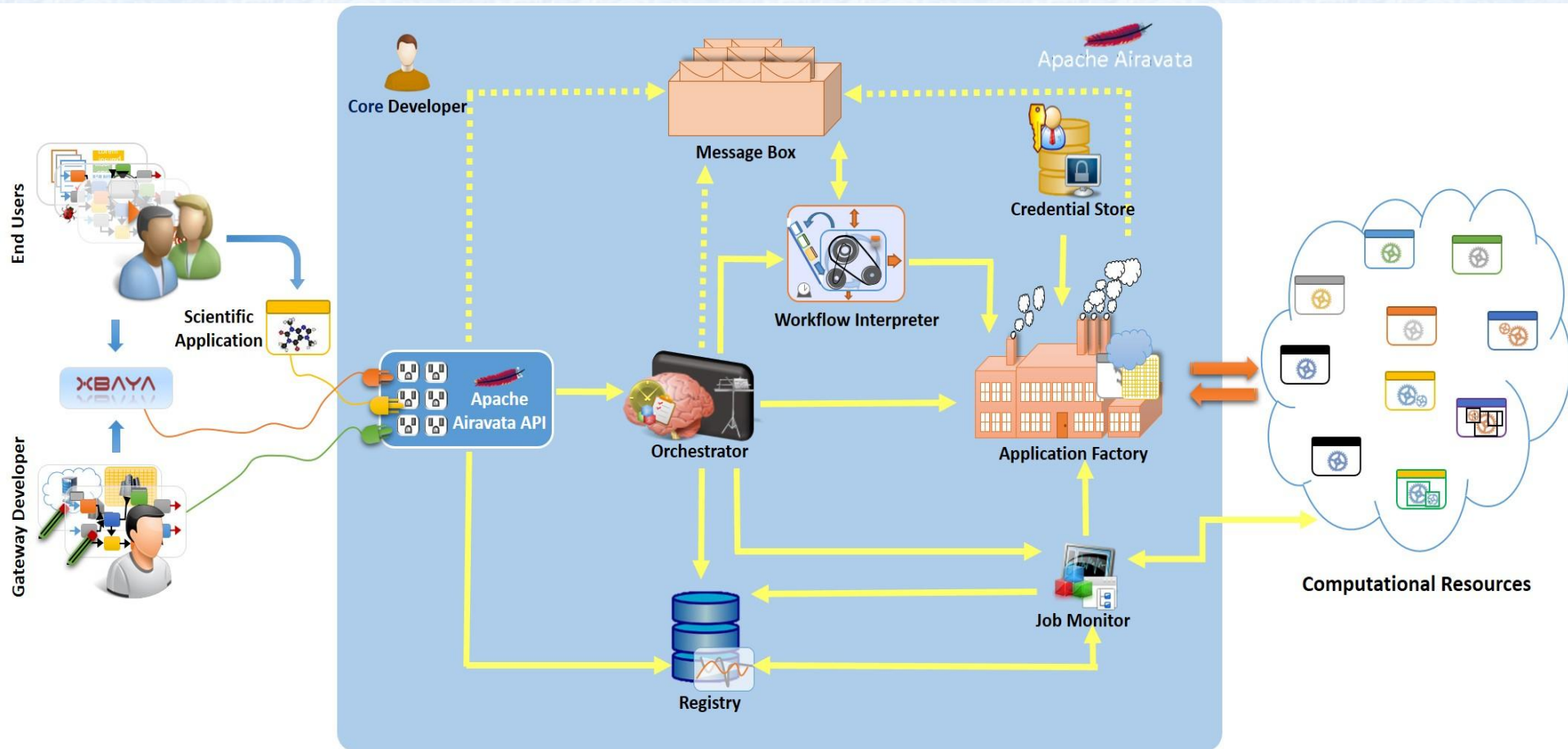
**RNA-Seq Analysis**

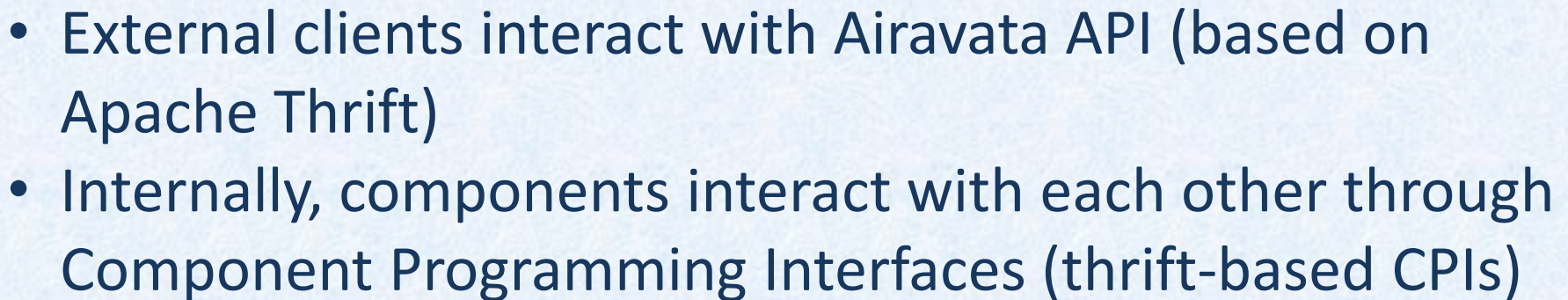
Map Reads & Assemble Transcripts





- Airavata is a general purpose distributed system software framework build on micro-service and component based architecture principles
- Airavata provides capabilities to compose, manage, execute and monitor large scale applications and workflows on distributed computing resources
- Airavata supports executions on local clusters, national grids, academic and commercial clouds
- Airavata is inherently multi-tenanted





- Experiment data model is a **complex data model**
- Data structures : string, type-defs, integers, lists, sets
- Can refer other structs, enums

```
struct Experiment {  
  1: required string experimentID = DEFAULT_ID,  
  2: required string projectID = DEFAULT_PROJECT_NAME  
  3: optional i64 creationTime,  
  4: required string userName,  
  5: required string name,  
  6: optional string description,  
  7: optional string applicationId,  
  8: optional string applicationVersion,  
  9: optional string workflowTemplateId,  
 10: optional string workflowTemplateVersion,  
 11: optional UserConfigurationData userConfigurationData,  
 12: optional string workflowExecutionInstanceId,  
 13: optional list<DataObjectType> experimentInputs,  
 14: optional list<DataObjectType> experimentOutputs,  
 15: optional ExperimentStatus experimentStatus,  
 16: optional list<WorkflowNodeStatus> stateChangeList,  
 17: optional list<WorkflowNodeDetails> workflowNodeDetailsList,  
 18: optional list<ErrorDetails> errors  
}
```

Defining a struct

- Can send such **complex data model** over the wire
- Also note that Thrift support **exception handling** too.

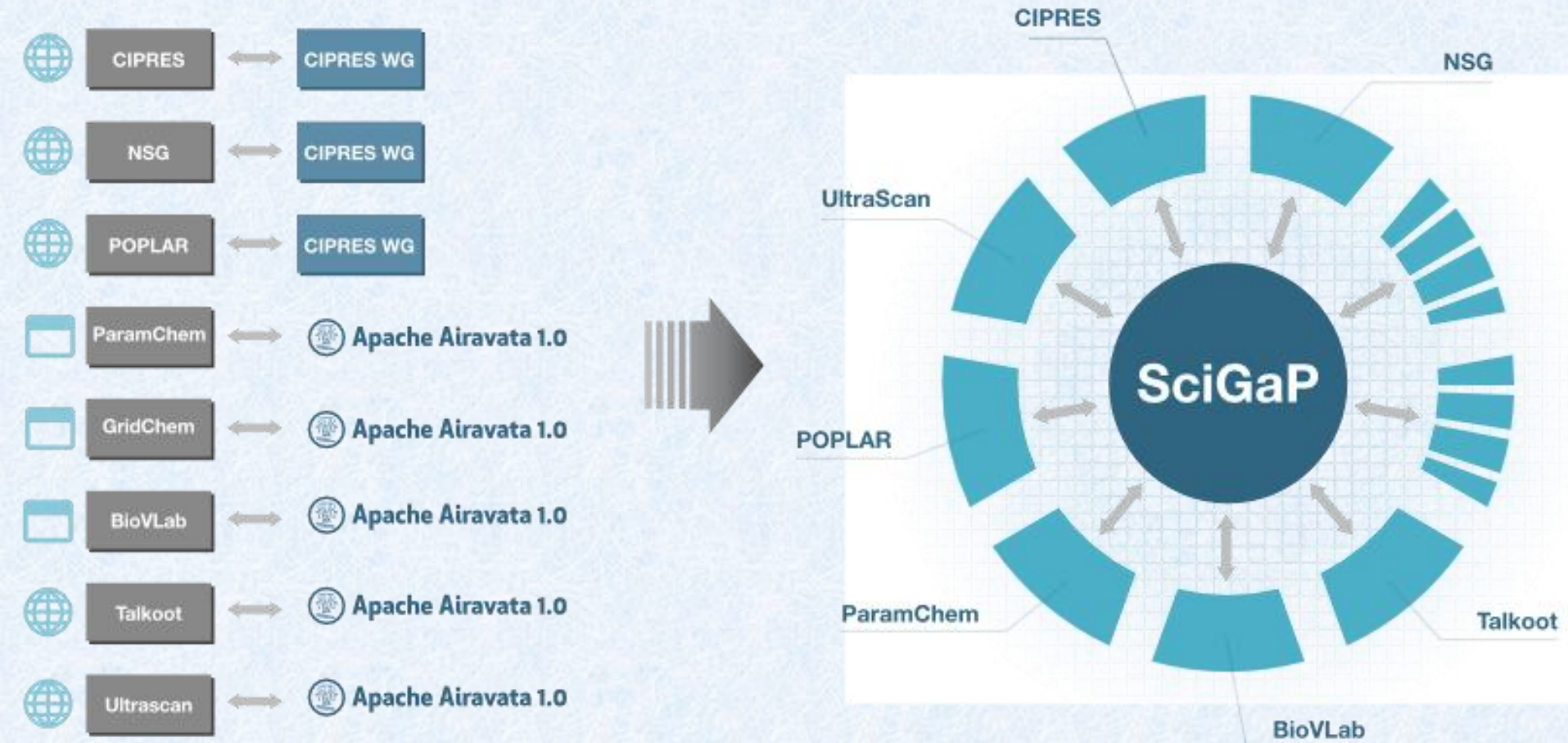
```
string createExperiment(1: required experimentModel.Experiment experiment)  
  throws (1: airavataErrors.InvalidRequestException ire,  
         2: airavataErrors.AiravataClientException ace,  
         3: airavataErrors.AiravataSystemException ase)
```

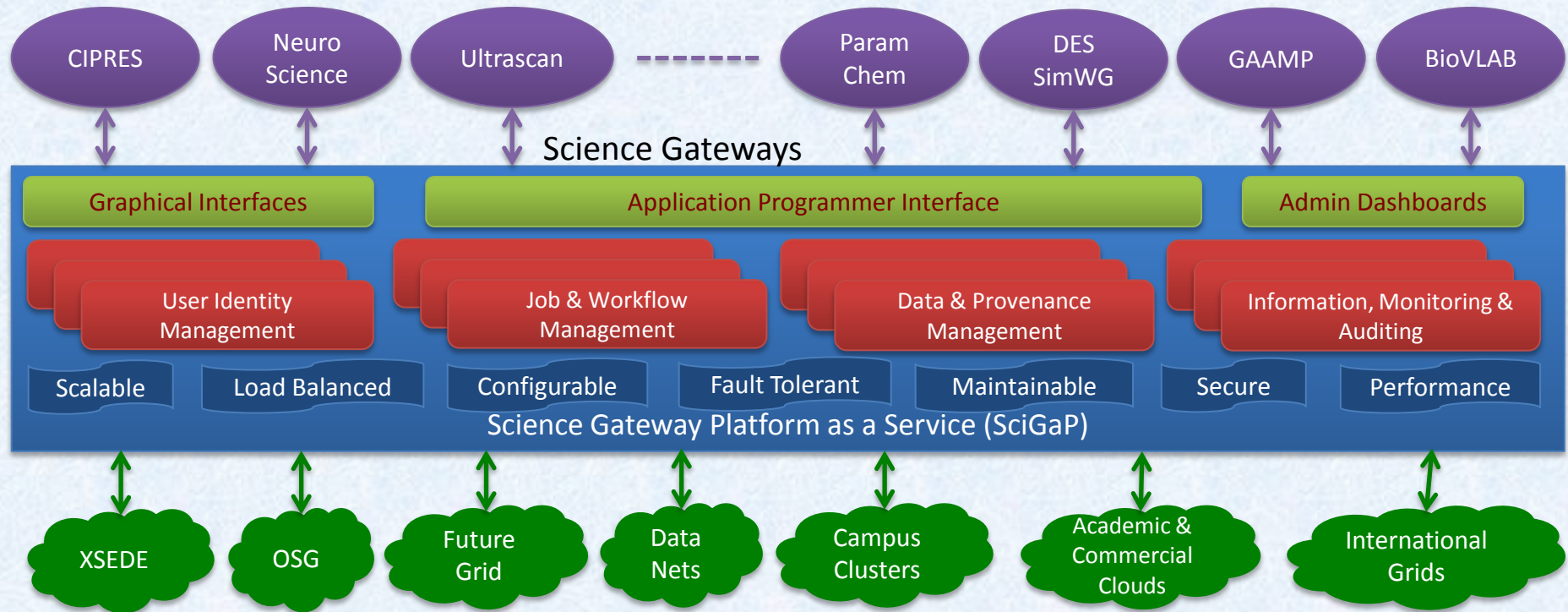
Defining a service method

Clean way to define IDLs with  
richer data structures



## Science Gateway Platform as a Service





## Community Hangout

Mailing lists:

[architecture@airavata.apache.org](mailto:architecture@airavata.apache.org)

[dev@airavata.apache.org](mailto:dev@airavata.apache.org)

[users@airavata.apache.org](mailto:users@airavata.apache.org)



Extend Airavata from your project or extend your project from Airavata

Agave is a *Science-as-a-Service* web API platform

## **Run scientific codes**

- your own or community provided codes

## **...on HPC, HTC, or cloud resources**

- your own, shared, or commercial systems

## **...and manage your data**

- reliable, multi-protocol, async data movement

## **...from the web**

- webhooks, rest, json, cors, oauth2

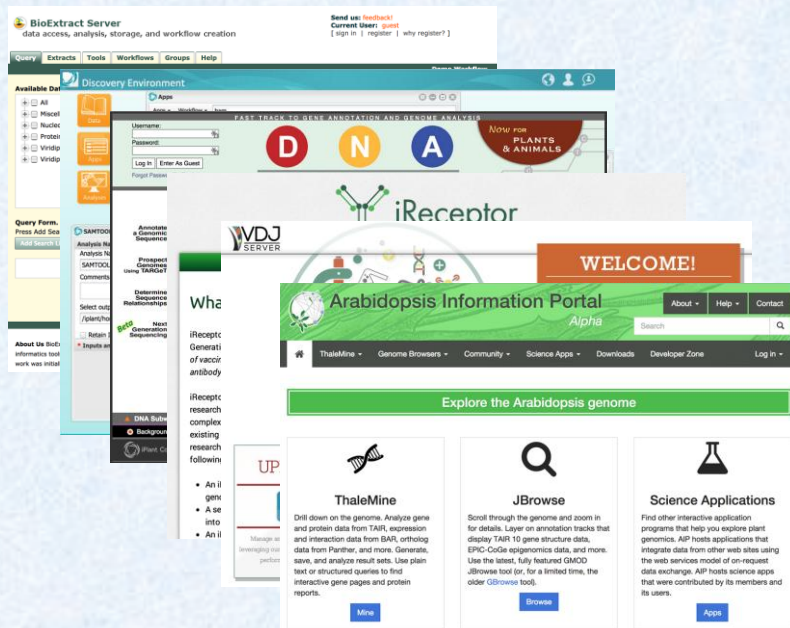
## **...and remember how you did it**

- deep provenance, history, and reproducibility built in



- Multitenant
- Hosted identity management
- Supports multiple IdP
- OAuth2/OIDC server
- API Management
- Hosted or on premise
- Vertical SSO
- Analytics and reporting
- Developer resources
- Multiple SDK & CLI
- Reference gateway
- White labeled
- 100% open source

## Used to power web & mobile applications



## Used to extend existing processes

### What is the Agave CLI

The Agave CLI is a collection of Bash shell scripts allowing you to interact with the Agave Platform. The CLI allows you to streamline common interactions with the API and automating repetitive and/or background tasks.

### Installation from source

The following technologies are required to use the Agave API cli tools.

- \* bash
- \* curl
- \* Perl
- \* Python (including json.tool)

Just clone the repository from Bitbucket and add the bin directory to your classpath and you're ready to go.

```
git clone https://bitbucket.org/taccaci/foundation-cli.git agave-cli
export PATH=$PATH:$PWD/agave-cli/bin
```

### Getting started

From here on, we assume you have the CLI installed and your environment configured properly. We also assume you either set or will replace the following environment variables:

- `AGAVE_USERNAME`: The username you use to login to Agave for your organization.
- `AGAVE_PASSWORD`: The password you use to login to Agave for your organization.

926eb83 Removed quotes from null  
Rion Dooley · 2014-10-16

1 commit  
Pushed to taccaci/foundation/cli  
c8675d4 Adding script to automagic  
Rion Dooley · 2014-10-16

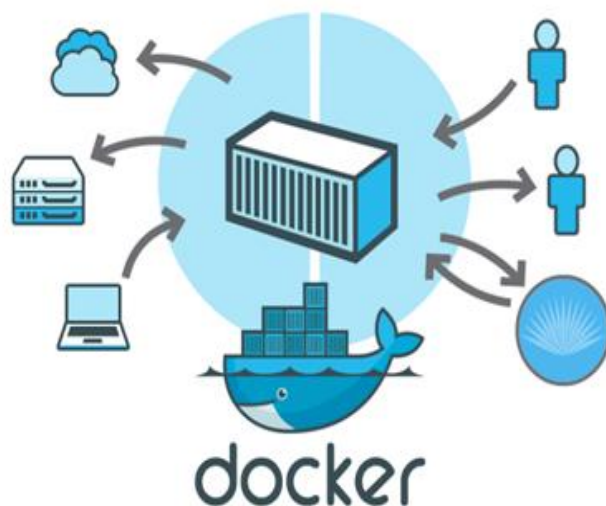
1 commit  
Pushed to taccaci/foundation/cli  
7bf1c3a Fixing verbosity in recursive  
Rion Dooley · 2014-10-13

1 commit  
Pushed to taccaci/foundation/cli  
d105495 Cleaning up output messages  
Rion Dooley · 2014-10-02

1 commit  
Pushed to taccaci/foundation/cli  
b221009 Working around bug in curl  
Rion Dooley · 2014-10-02

```
dooley$ systems-list
lonestar4.tacc.teragrid.org
systest-rodeo-storage
rodeo.storage.demo
osg-dooley
storage.example.com
dooley-stampede-gsi-teragrid
data.iplantcollaborative.org
condor.opensciencegrid.org
data.vdjservice.org
condor-dooley.execute.example.com
dooley-lonestar-gsi-teragrid
dooley-rodeo-storage1
docker.iplantcollaborative.org
stampede.tacc.utexas.edu
execute.example.com
stampede-dooley
dooley-rodeo-docker1
dooley-lonestar-gsi
dooley-docker
dooley-stampede-gsi
dooley-ranch
agave-demo
meetings-104-172:~ dooley$ files-list -S docker.iplantcollaborative.org .
dooley
lmjohnstone
nirav
sjmiller
systest
welge
meetings-104-172:~ dooley$ files-list -S osg-dooley .
dooley
.bash_history
.bash_logout
.bash_profile
.bashrc
.gnome2
.m2
.mozilla
.mysql_history
```

## (Re)Introducing the Micro App Paradigm

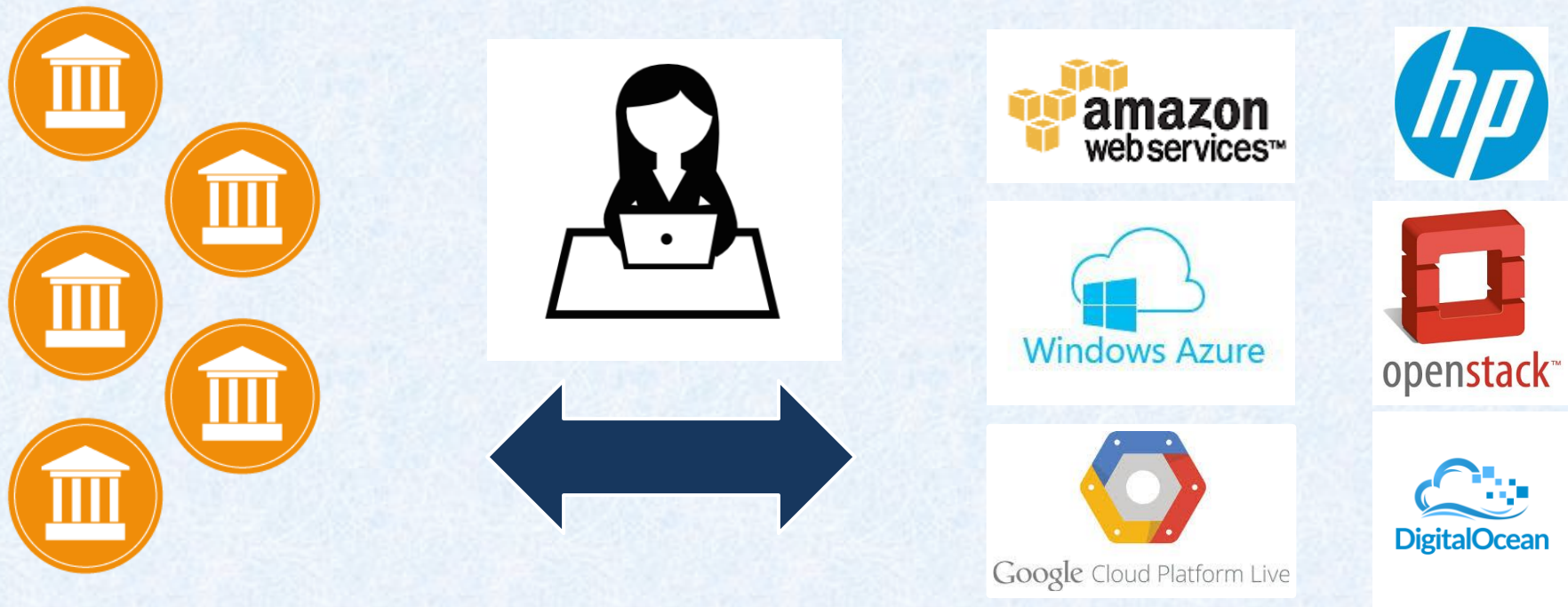


### AGAVE + DOCKER

Agave uses Docker container technology to safely and securely run your code on HPC, HTC, Cloud and your local resources.



## Agave Delivers Process-as-a-Service



## PRODUCTS



## FOUNDATIONAL SERVICES



## LOW-LEVEL SERVICES, SECURITY, ASSETS, etc



## HARDWARE RESOURCES





Bisque WebApp



Services



Upload



Download



Images

Find images using tags



Martha Narro



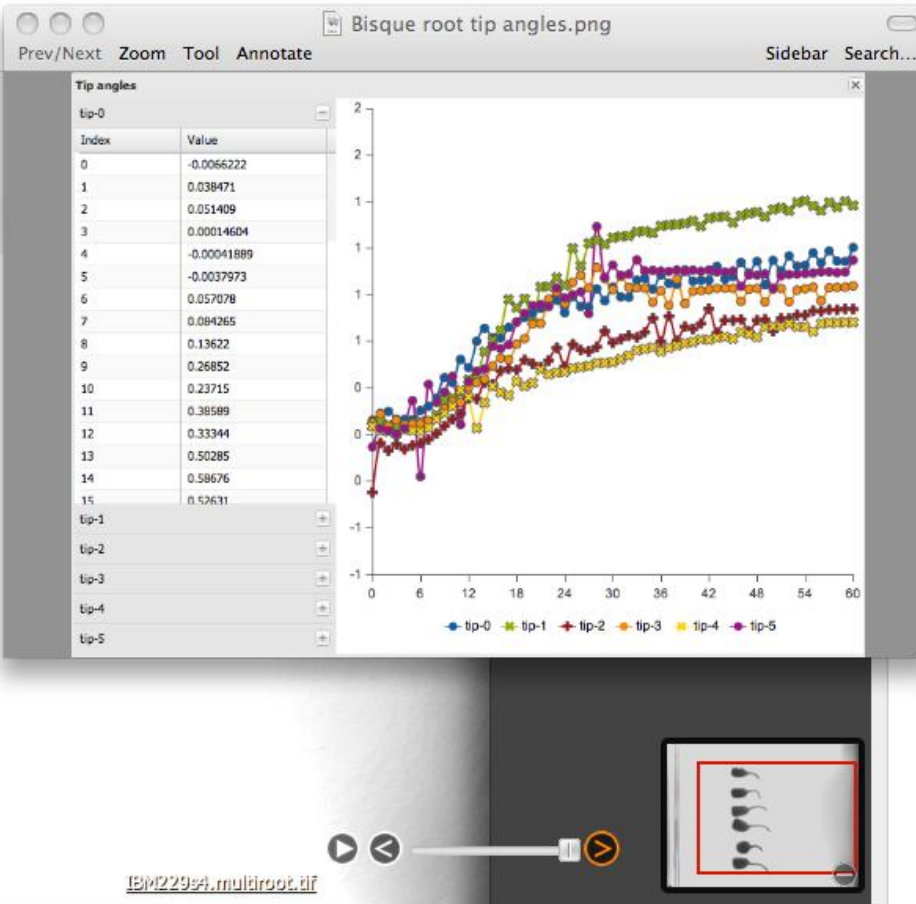
?

## 3. Results:

The module ran in 57 seconds

Tracked root tips

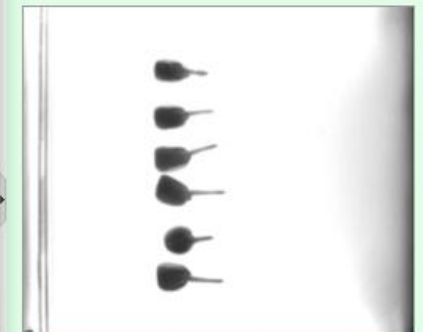
Plot Export



## Help and workflow

### 1. Select input image

Select an image by double clicking the image of interest, it will be loaded into the viewer for the step 2. The input time series first image should look something like this:



The 100% magnification on our images looks like this:



### 2. Select initial tip positions

[Home](#)[News](#)[About](#)[Documentation](#)[User Guide](#)[Beginner's Guides](#)[Authorization Guide](#)[Advanced Tutorials](#)[- Apps](#)[- Client Registration](#)[- Data](#)[- Jobs](#)[- Metadata](#)[- Notifications](#)[- Systems](#)[- Users](#)[Interactive REST Docs](#)[Event Reference](#)[Changelog](#)[Roadmap](#)[Best Practices](#)[Use Cases](#)[Tools](#)[Terms of Service](#)

## HANDS-ON TUTORIALS

### DEEP-DIVE INTO THE AGAVE REST APIS

## Advanced Tutorials

Dive deeper into the Agave REST APIs with these advanced tutorials on the individual APIs.



### Client Registration

Learn how to register your client applications and obtain API keys.



### Authorization

Learn how about authentication and authorization in Agave.



### App Management

Learn how to wrap your existing scientific applications and expose them for execution through the API.



### Agave + Docker (TODO)

Learn how to use Docker and Agave to conduct portable, reproducible science.





## System Management

Learn how to access your own HPC, HTC, Cloud, and Big Data resources with Agave.



## System Monitoring (TODO)

Learn how to monitor system uptime and availability with Agave.



## Job Management (TODO)

Learn how to run applications, monitor jobs, and archive data in Agave.



## Data Management

Learn how to manage, move, and share your data with others in this tutorial.



## User Management

Learn how the Agave profile service can help you locate and interact with other users in your organization.



## Using Postlts (TODO)

Learn how to create disposable, pre-authenticated URLs that you can share with anyone.



## Metadata Management

Learn how to view, validate, and manage metadata in Agave.



## Notifications and Events

Learn about Agave's event system and how to get real time notifications about any event, any time.

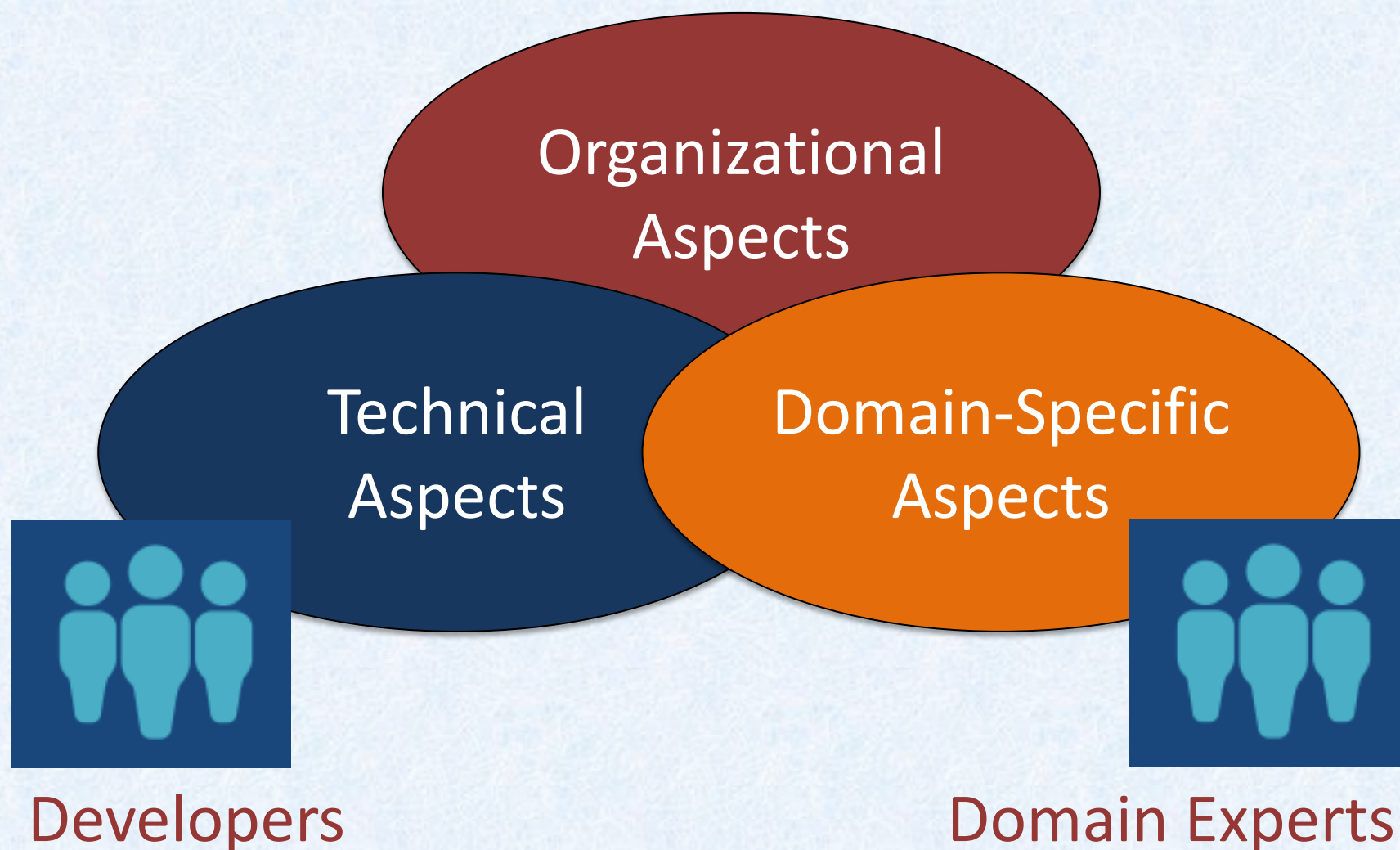
## Crucial Topics

- Close collaboration with user communities
- Knowledge about available technical solutions

## Sounds easy but...

- Requirements of user communities often not so clear
- Technologies sometimes still under development for certain building blocks
- ➔ Slow uptake of solutions
- ➔ Larger effort for creating science gateways

## DISCUSSION



## Domain-specific aspects:

- Goal, target area and target users
- Visions/demands on the layout
- Priorities of features and options, e.g., a list from must-have to great-to-have options
- Integration of existing applications or development of applications
- Technologies of the applications
- Visualization
- Security demands
- Workflows



## Organizational aspects:

- Time constraints for the development, agreement on a (maybe even rough) project plan with milestones
- Agreement on alpha- or beta-tester
- Regular meetings

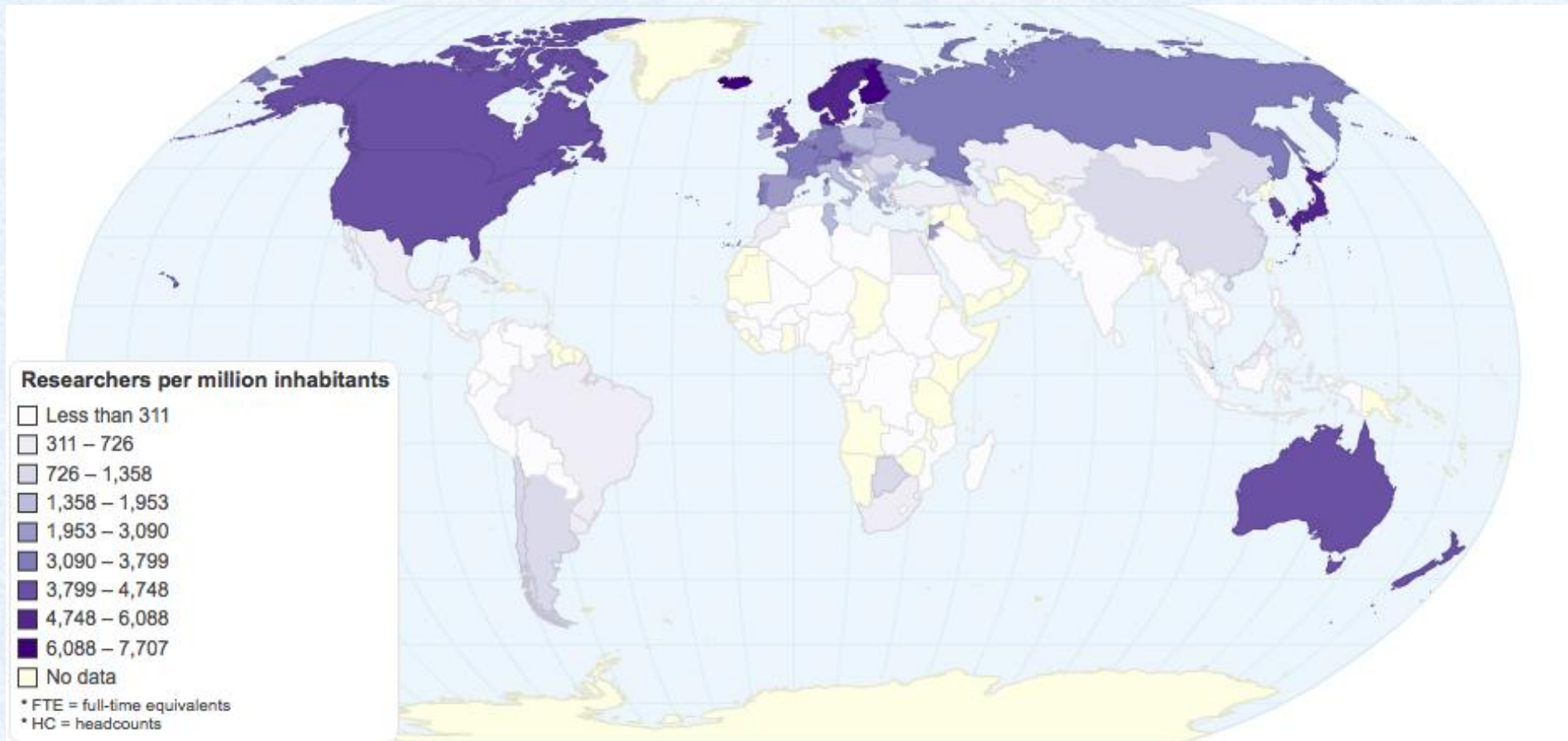
## Technical aspects:

- Experience with existing frameworks and programming languages
- Available infrastructure including security infrastructure and resources
- Available support of suitable technologies
- Scalability of suitable technologies
- Effort for extending existing technologies compared to novel developments
- Synergy effects with other science gateway projects

## A world-wide research computing infrastructure

- Transparent service selection
  - e.g., Docker could be part of the solution
- Access to data irrespective of location
- Options to share data efficiently
- Appropriate privacy and security measures
- Optimized usage of resources
  - e.g., optimized usage of cloud computing and their business models

~7 million researchers world wide

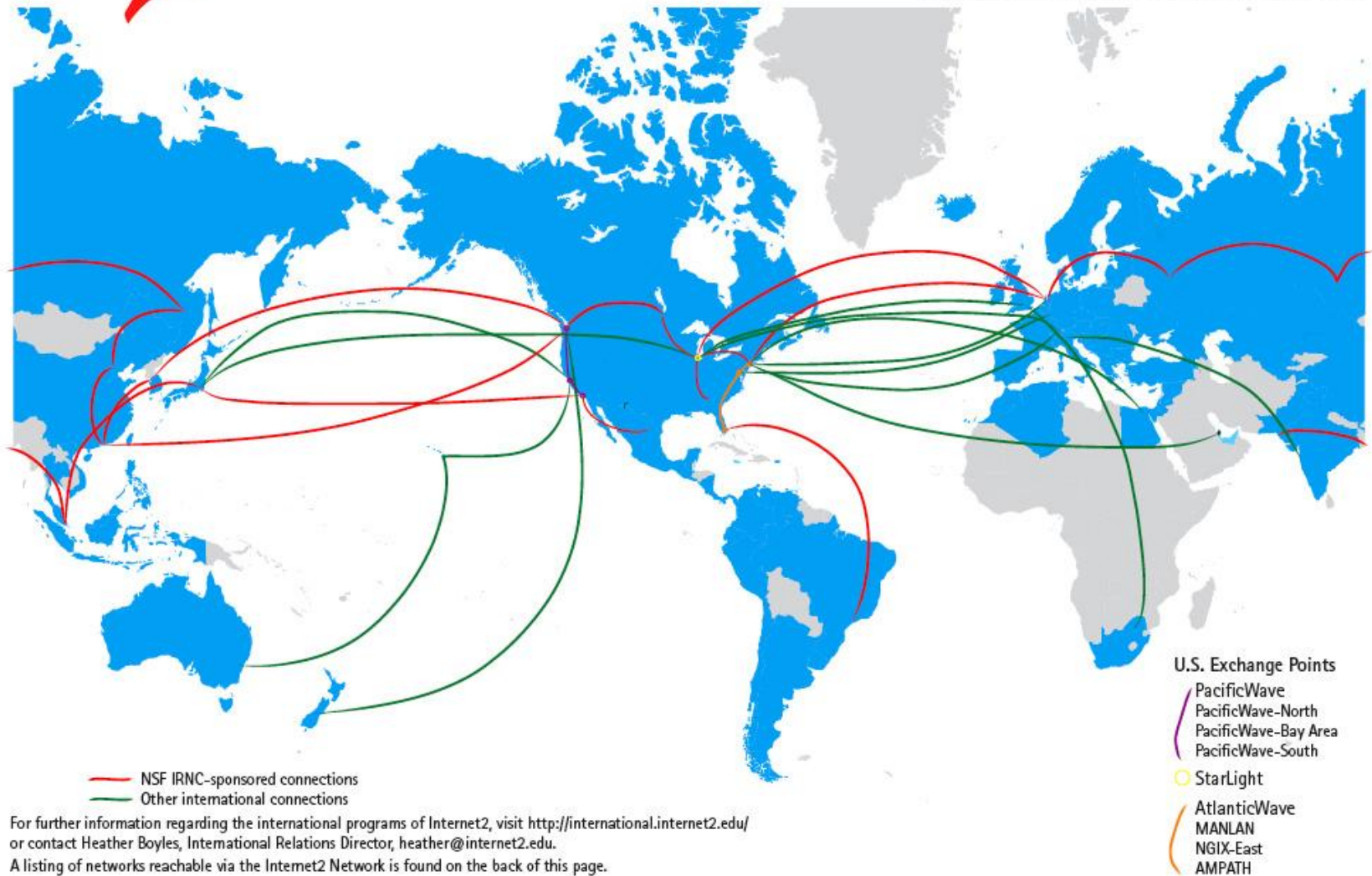


<http://chartsbin.com/view/1124>



INTERNET<sup>2</sup>® [www.internet2.edu](http://www.internet2.edu)

THE INTERNATIONAL REACH OF THE INTERNET2 NETWORK



## Integration of data sources and instruments

- Different data formats
- Different interfaces
- Different hardwares and technologies

... from small ones to the big ones...





## Software searchability, reproducibility and reusability

- Science gateways step in the right direction but ... much more work necessary on searchability... Not only finding any data for a research area but finding the right data
- Metadata approaches
- Dictionaries
- More involvement of librarians



## Software searchability, reproducibility and reusability

- Science gateways step in the right direction but ... much more work necessary on reproducibility and reusability...
- studies in medicine and pharmacology: 11% or 6% of the analysed research was reproducible
- myExperiment: only 20% of workflows reusable because of dependencies on hardware, local or distributed data, software versions



## Software searchability, reproducibility and reusability

- Science gateways and workflow systems step in the right direction but ...

much more work necessary on reproducibility and reusability...

- Containerization approaches
- Migration approaches
- Combination of both

... require novel solutions!



## Open Science Framework

Cloud-based management for your projects.



### Structured projects

Keep all your files, data, and protocols in **one centralized location**. No more trawling emails to find files or scrambling to recover from lost data. [SECURE CLOUD](#)



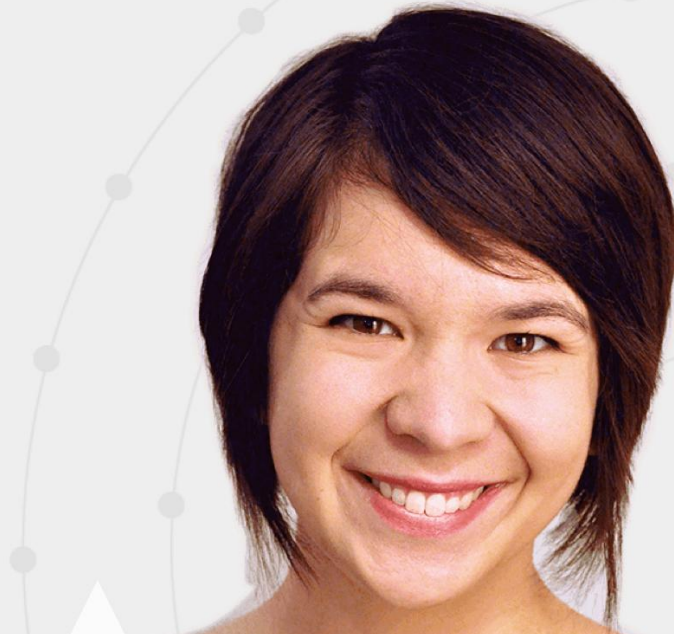
### Control access

**You control which parts of your project are public or private** making it easy to collaborate with the worldwide community or just your team. [PROJECT-LEVEL PERMISSIONS](#)



### Respect for your workflow

**Connect your favorite third party services** directly to the Open Science Framework. [3RD PARTY INTEGRATIONS](#)



"The OSF is a great way to collaborate and stay organized while still using your favorite external services."



- **Logical level: Meta-workflows**

Herres-Pawlis, S., Hoffmann, A., Rösener, T., Krüger, J., Grunzke, R., and Gesing, S. “Multi-layer Meta-metaworkflows for the Evaluation of Solvent and Dispersion Effects in Transition Metal Systems Using the MoSGrid Science Gateways” Science Gateways (IWSG), 2015 7th International Workshop on, pp.47-52, 3-5 June 2015, IEEE Xplore, doi: 10.1109/IWSG.2015.13

- **System level: Combination of strengths of workflow systems**

Hazekamp, N., Sarro, J., Choudhury, O., Gesing, S., Scott Emrich and Thain, D. “Scaling Up Bioinformatics Workflows with Dynamic Job Expansion: A Case Study Using Galaxy and Makeflow”, e-Science (e-Science), 2015 IEEE 11th International Conference on, pp.332-341, Aug. 31 2015-Sept. 4 2015

- **Prediction: Model for optimization of tasks and threads**

Choudhury, O., Rajan, D., Hazekamp, N., Gesing, S., Thain, D., and Emrich, S. “Balancing Thread-level and Task-level Parallelism for Data-Intensive Workloads on Clusters and Clouds”, Cluster Computing (CLUSTER), 2015 IEEE International Conference on, pp.390-393, 8-11 Sept. 2015, doi:10.1109/CLUSTER.2015.60

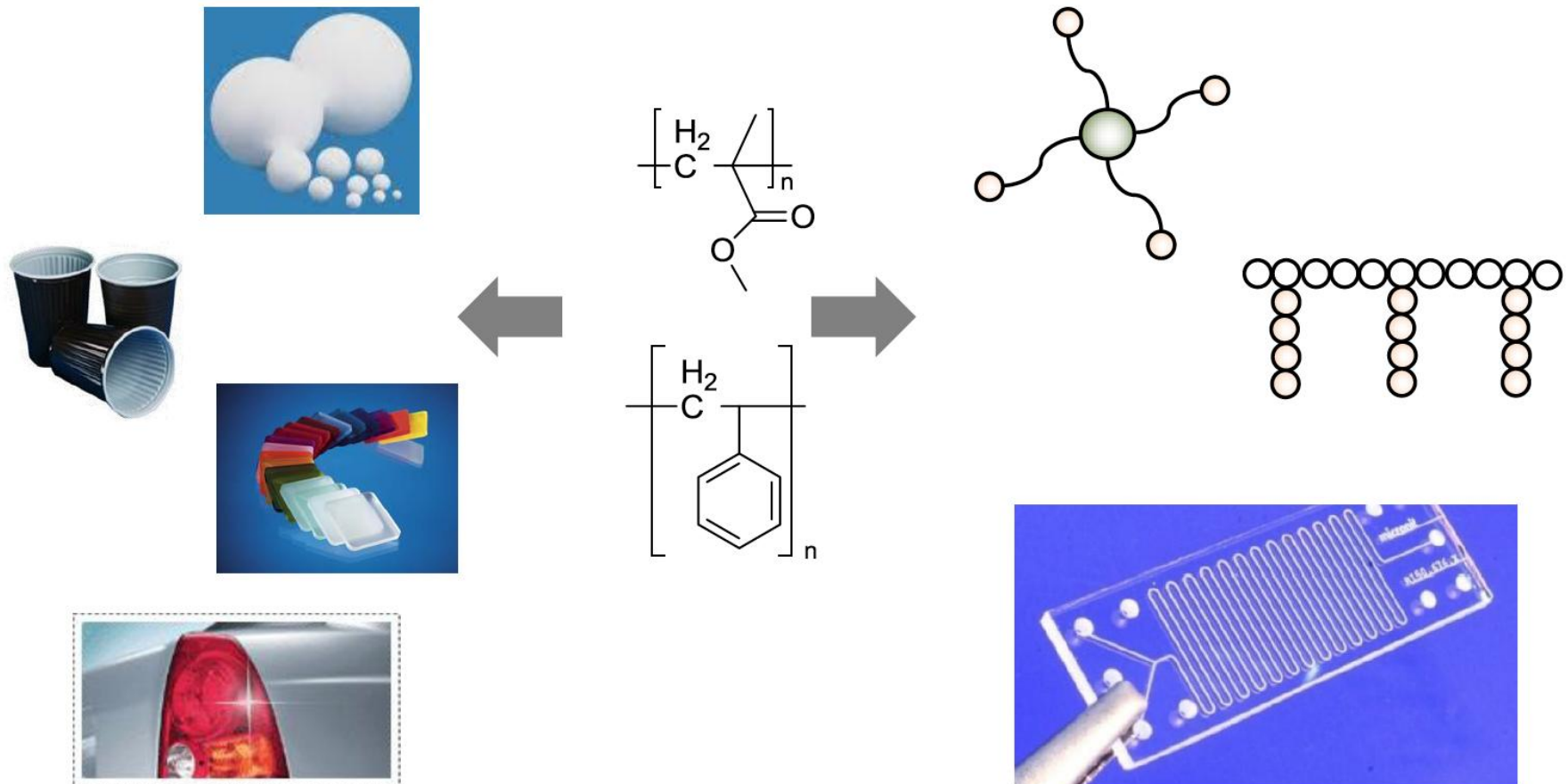


Free radical polymerisation (FRP)

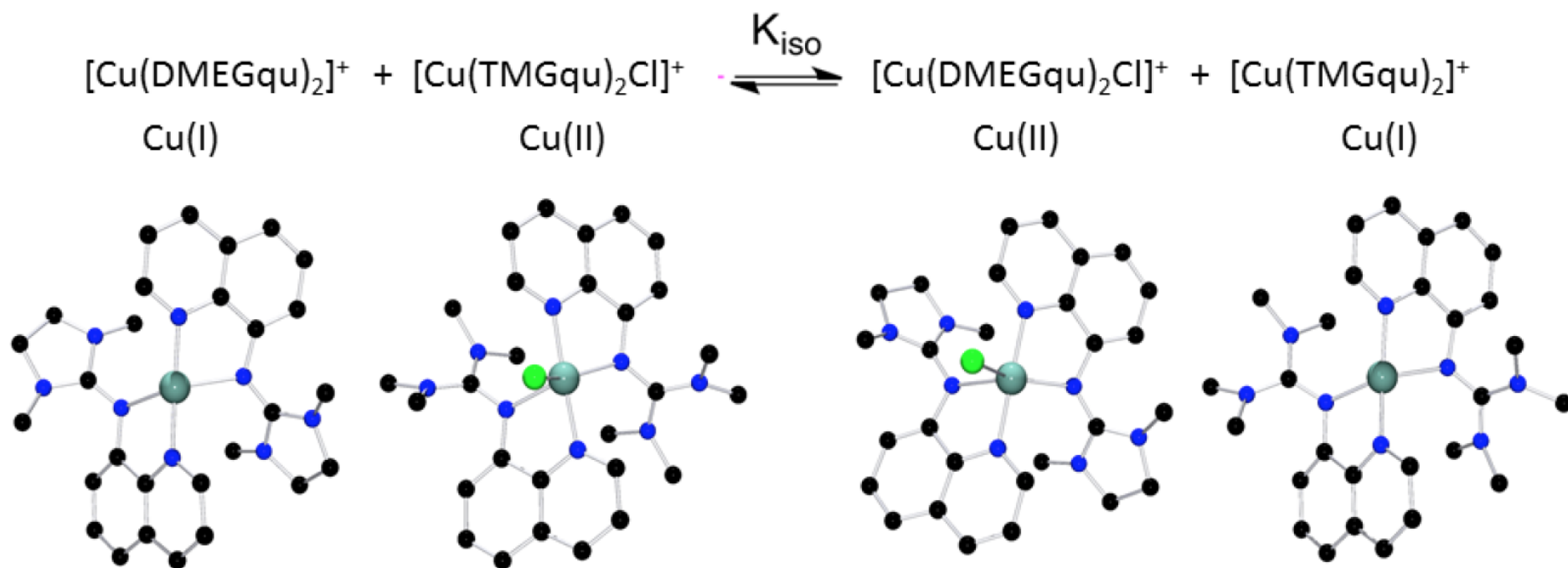
→ Commodity polymers

Controlled radical polymerisation (CRP)

→ Tailored and/or intelligent polymers



Handbook of Radical Polymerization, K. Matyjaszewski 2004, Wiley, New York.



→ Correct structural description needed by quantum chemistry

- Evaluation of structural description with suited functionals and basis sets
- **Evaluation of structural description with dispersion and solvent models at suited temperatures needed for an accurate description!**



# Translation into workflows

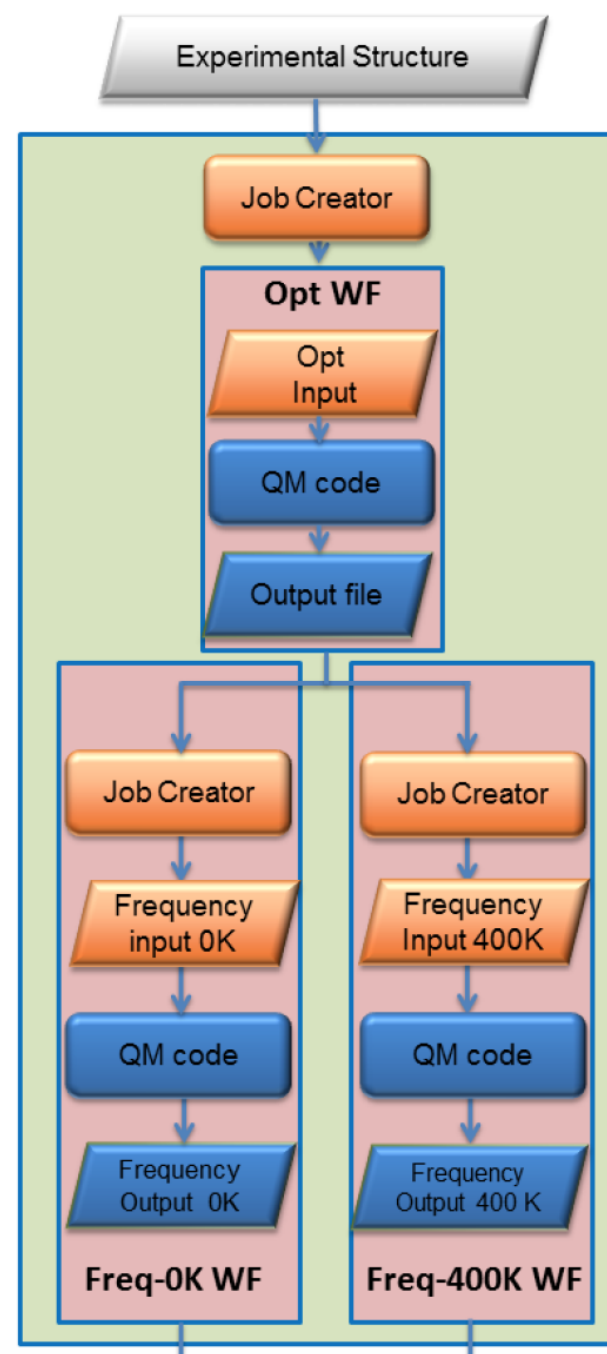
Fundamental step: optimisation

Frequency calculation at 0 K and at 400 K

(polymerisation temperature) for the same molecule

→ Small workflow with 3 atomic workflows (opt and freq)

S. Herres-Pawlis, A. Hoffmann, A. Balasko, P. Kacsuk, G. Birkenheuer, A. Brinkmann, L. de la Garza, J. Krüger, S. Gesing, R. Grunzke, G. Terstyansky, N. Weingarten, *Concurrency Computat.: Pract. Exper.* 2015, 27, 344-357.

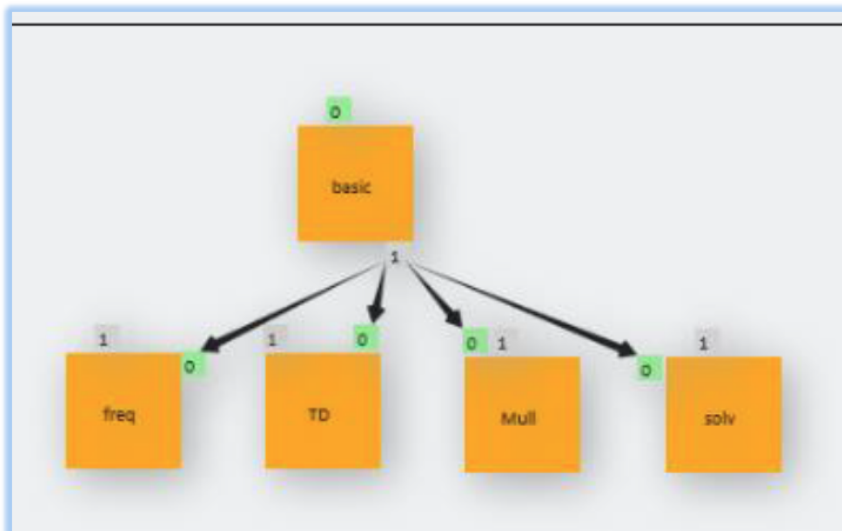


# Meta-Workflows

*Implementation in WS-PGRADE:*

## **White-box approach**

- Definition of subworkflows
- Creation of templates
- Creation of concrete WFs out of templates
- Definition of meta-workflows by using these sub-concrete WFs



## *Advantages*

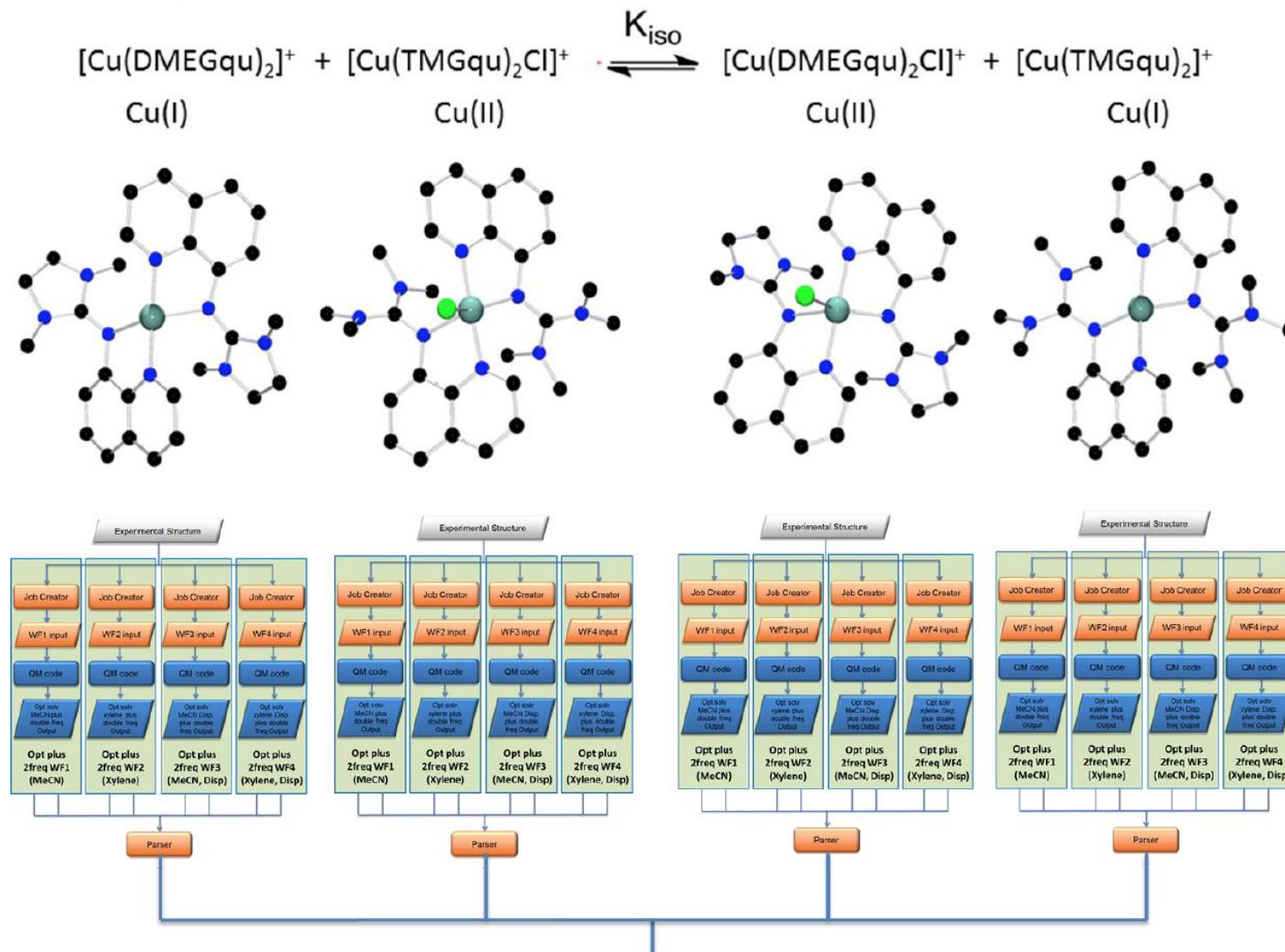
- High re-usability of subworkflows
- High-reusability of final workflows
- High flexibility in combination

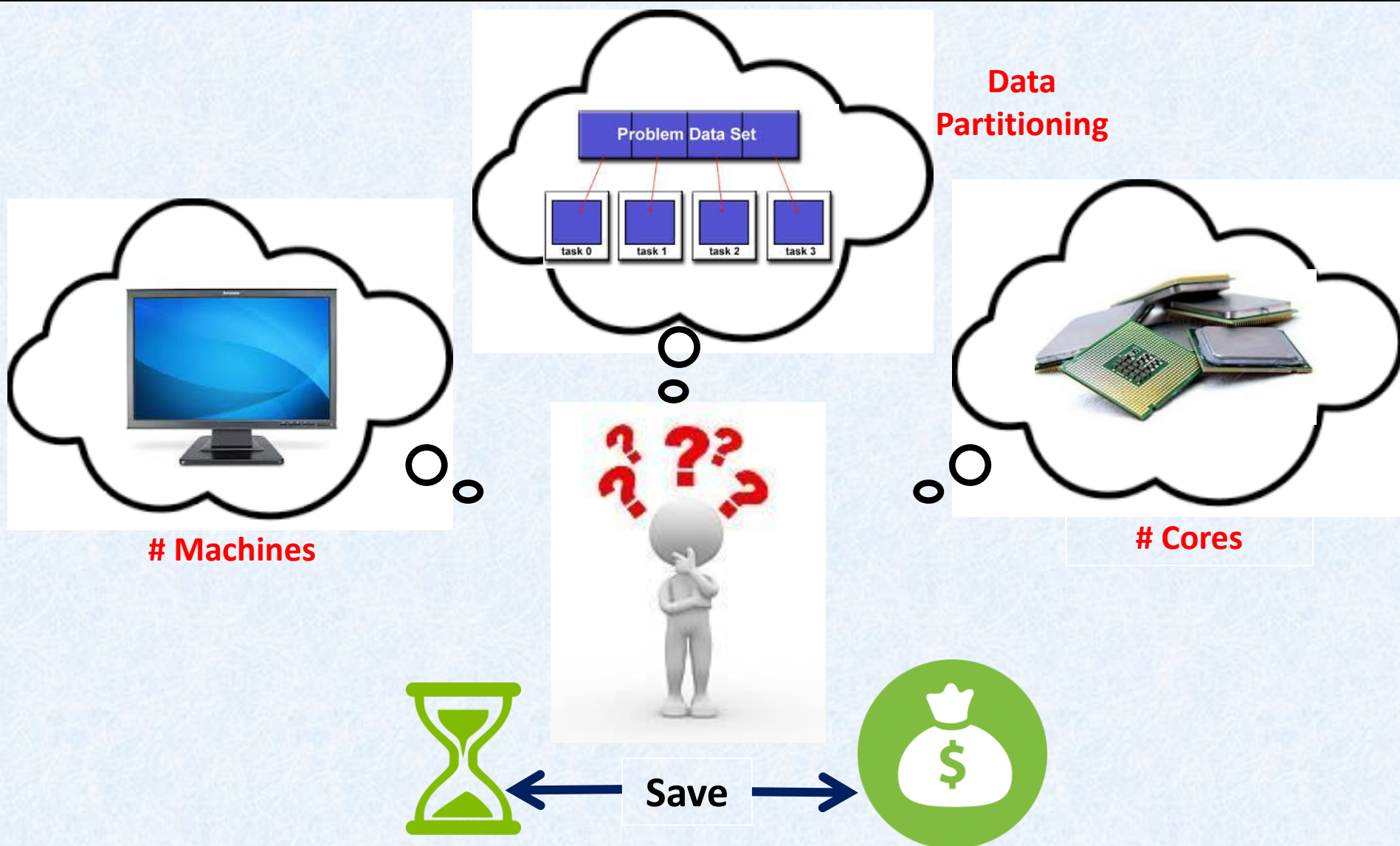
## *Disadvantages*

- Very complex for chemists
- Long-learning curve
- Error-prone in details



Doing it for 4 complexes ....



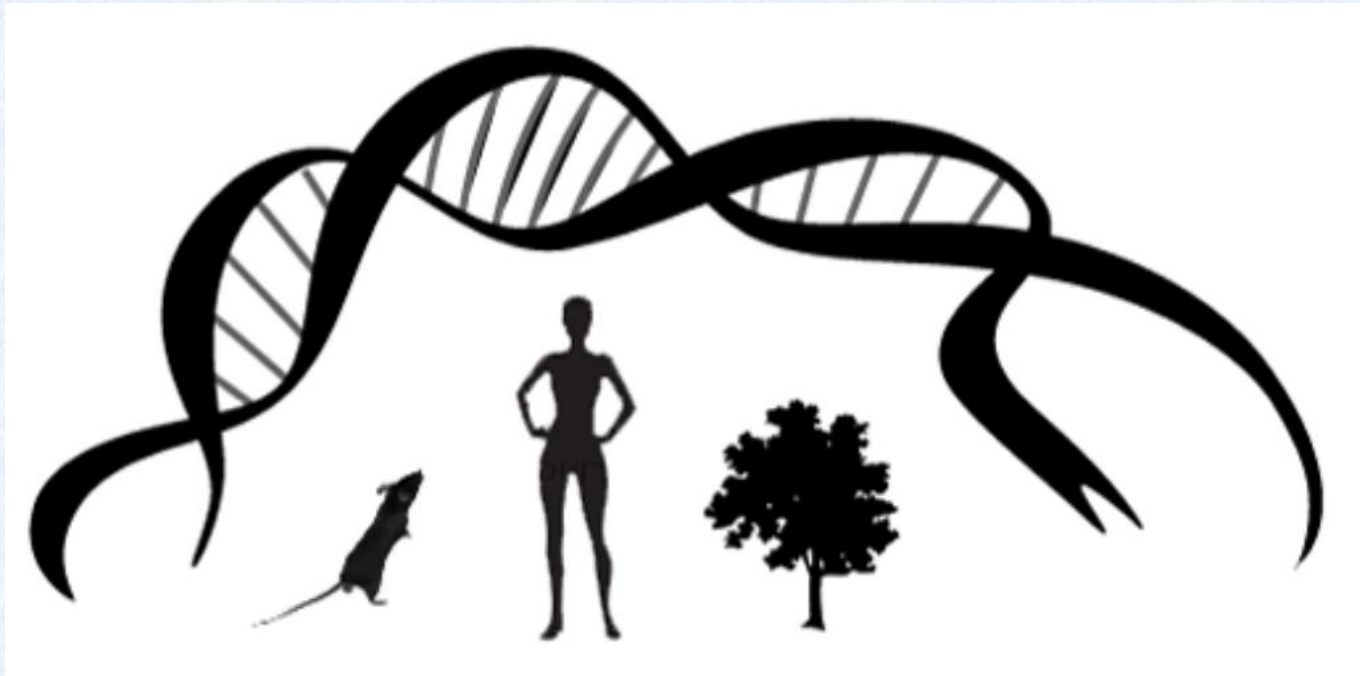


## Galaxy

The screenshot displays the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Cloud, Help, and User. The left sidebar lists various tools categorized by function, such as Get Data, Text Manipulation, and NGS tools. The main workspace is divided into several panels:

- Galaxy / VectorBase**: A central panel showing a workflow canvas with a grid of tools. The workflow includes steps like 'Input dataset', 'BWA-Makeflow', 'SAM-to-BAM', 'Add or Replace Read Groups', and 'Haplo-type-Makeflow'. The workflow is titled 'Workflow Canvas | BWA-GATK'.
- History**: A panel on the right showing a search for datasets and a message indicating that the history is empty.
- Tools**: A panel on the left showing a search for tools and a list of available tools, including 'BWA-Makeflow', 'SAM-to-BAM', and 'Haplo-type-Makeflow'.
- Workflow Parameters**: A panel on the right showing the parameters for the selected tool, 'SAM-to-BAM', including version, source, and input file.
- Public Galaxy Servers**: A banner for Penn State and Johns Hopkins University.
- Tweets**: A panel showing tweets from the Galaxy Project.

- Finding precise order of nucleotides within a DNA molecule
- A (adenine), G (guanine), C (cytosine), and T (thymine)  
(Human genome over 3 billion of nucleotides)





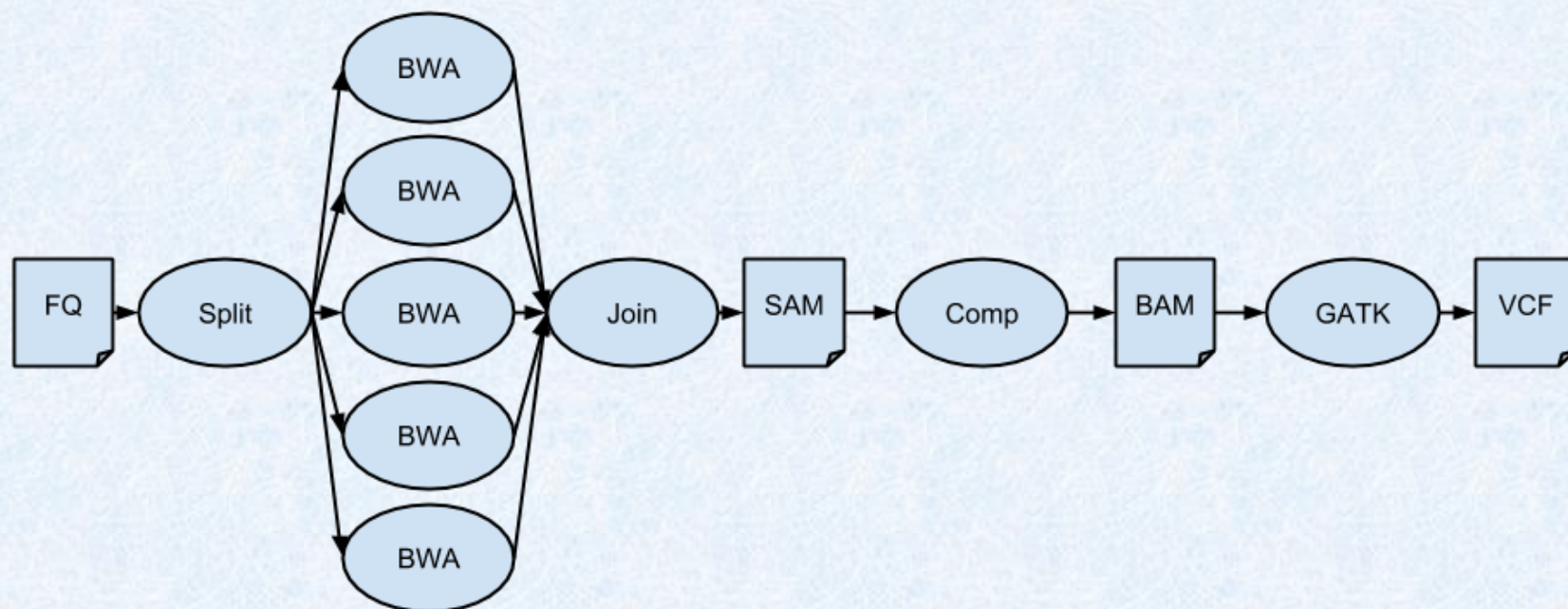
Let's imagine a party game. The game is a guessing game. Here is how it is played:  
You are thinking of a number and the group has to guess it. The tricky part is that the number is 200-digits in length. You are reading the digits of the number in your head without making a sound. Every so often a person interrupts you, and you tell them the single digit you were just thinking and where it is in the sequence of 200. Each time you are interrupted, you have to start again. You leave after a few hours and the group has to figure out the 200-digit number. They have to piece together the information you gave them, for example the 25<sup>th</sup> number was 5, the 40<sup>th</sup> number was 0, and so on. Using the information from their interruptions, they can repeat the number they gave you.

## Simple Workflow in Galaxy



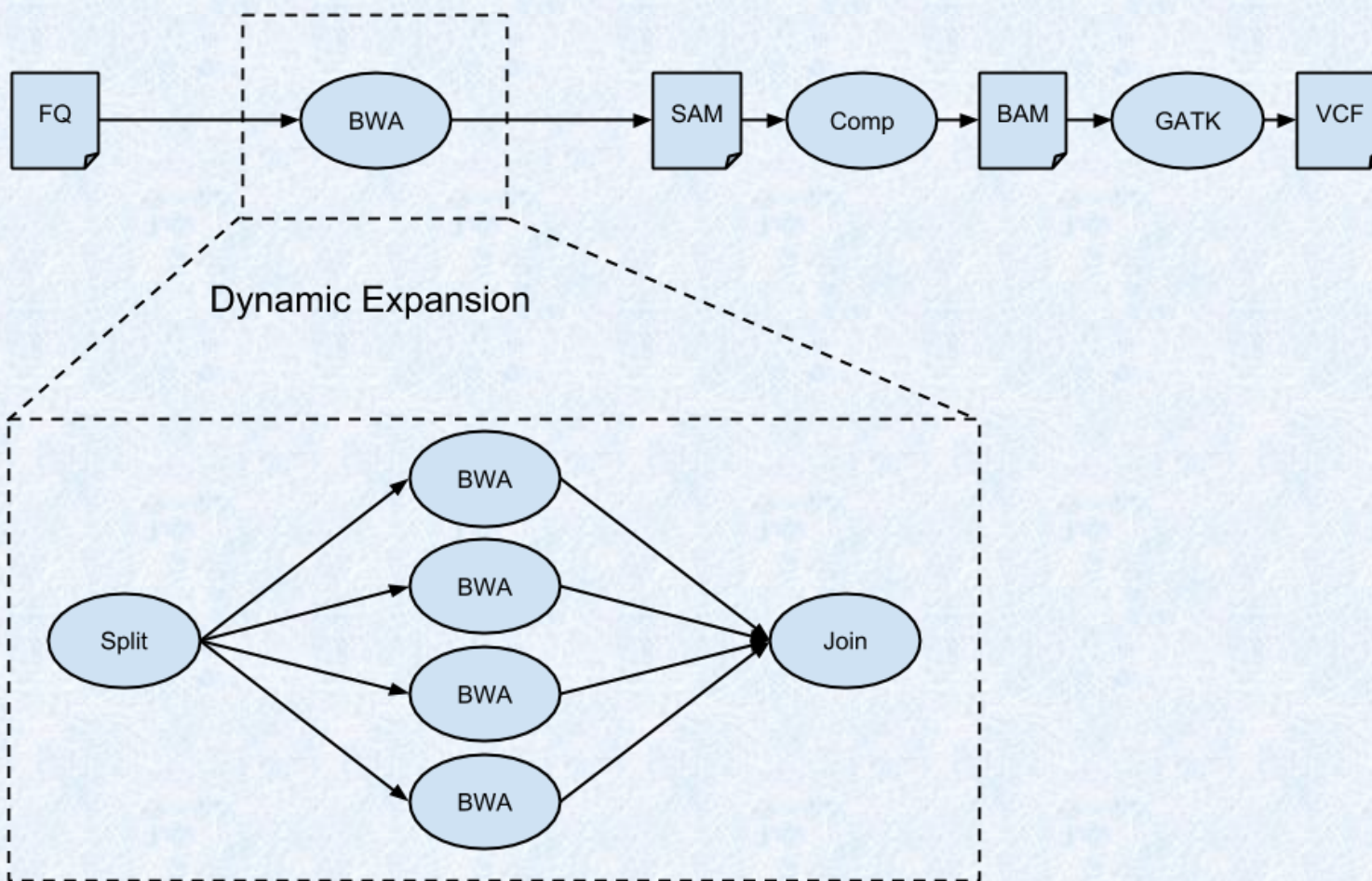
Problem: As Size increases so does Time

## Workflow with Parallelism added in Galaxy

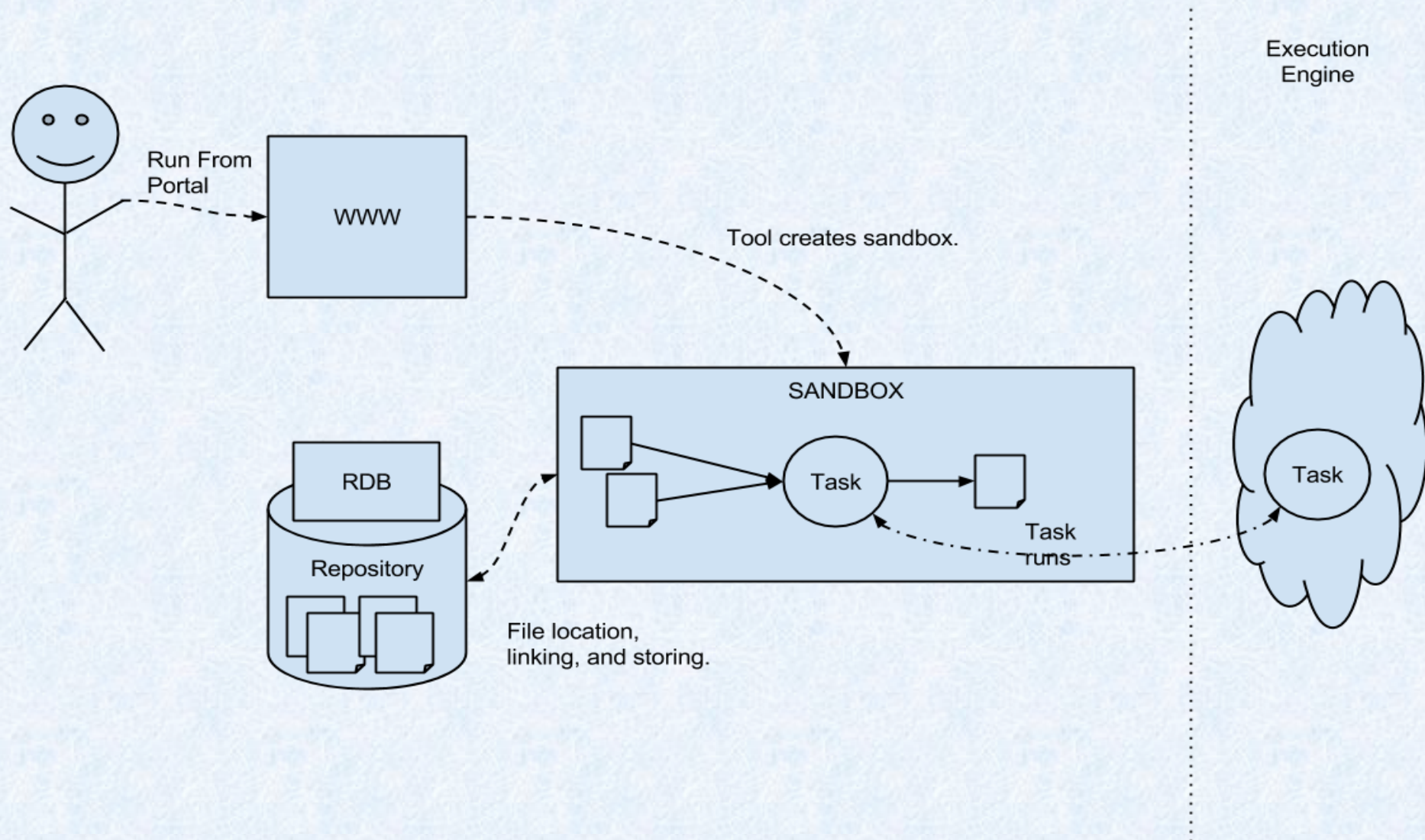


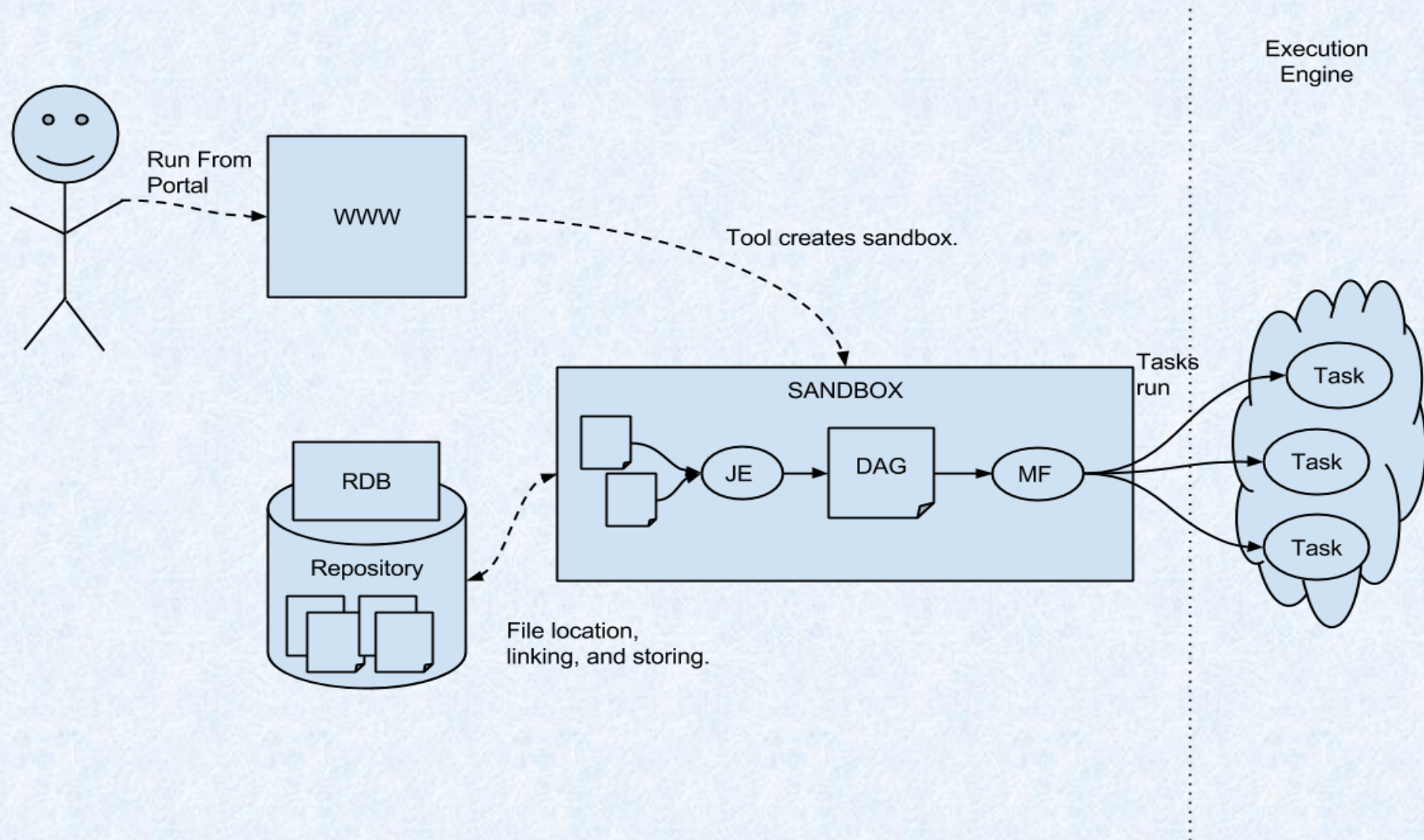
Problem: Tools must be updated every change in Parallelism/Relies on Scientist

## Workflow Dynamically Expanded behind Galaxy









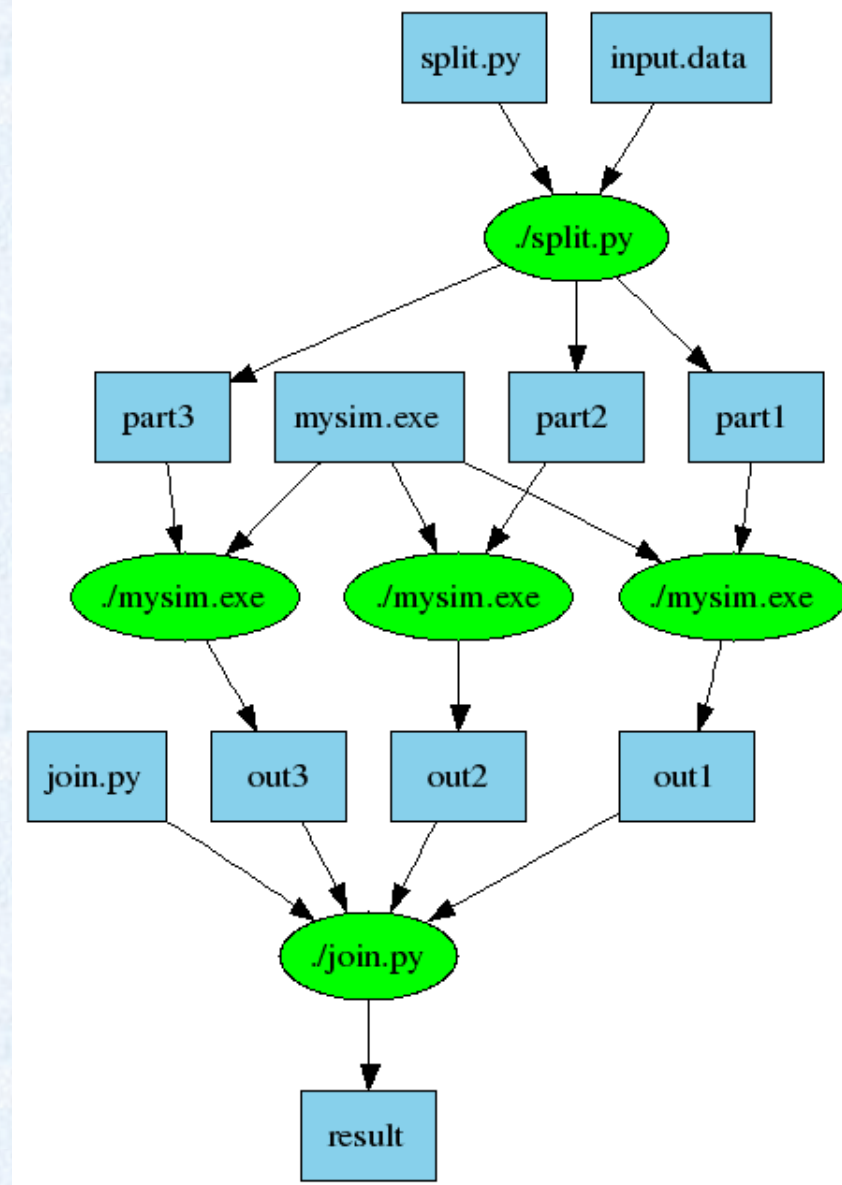
## Makeflow

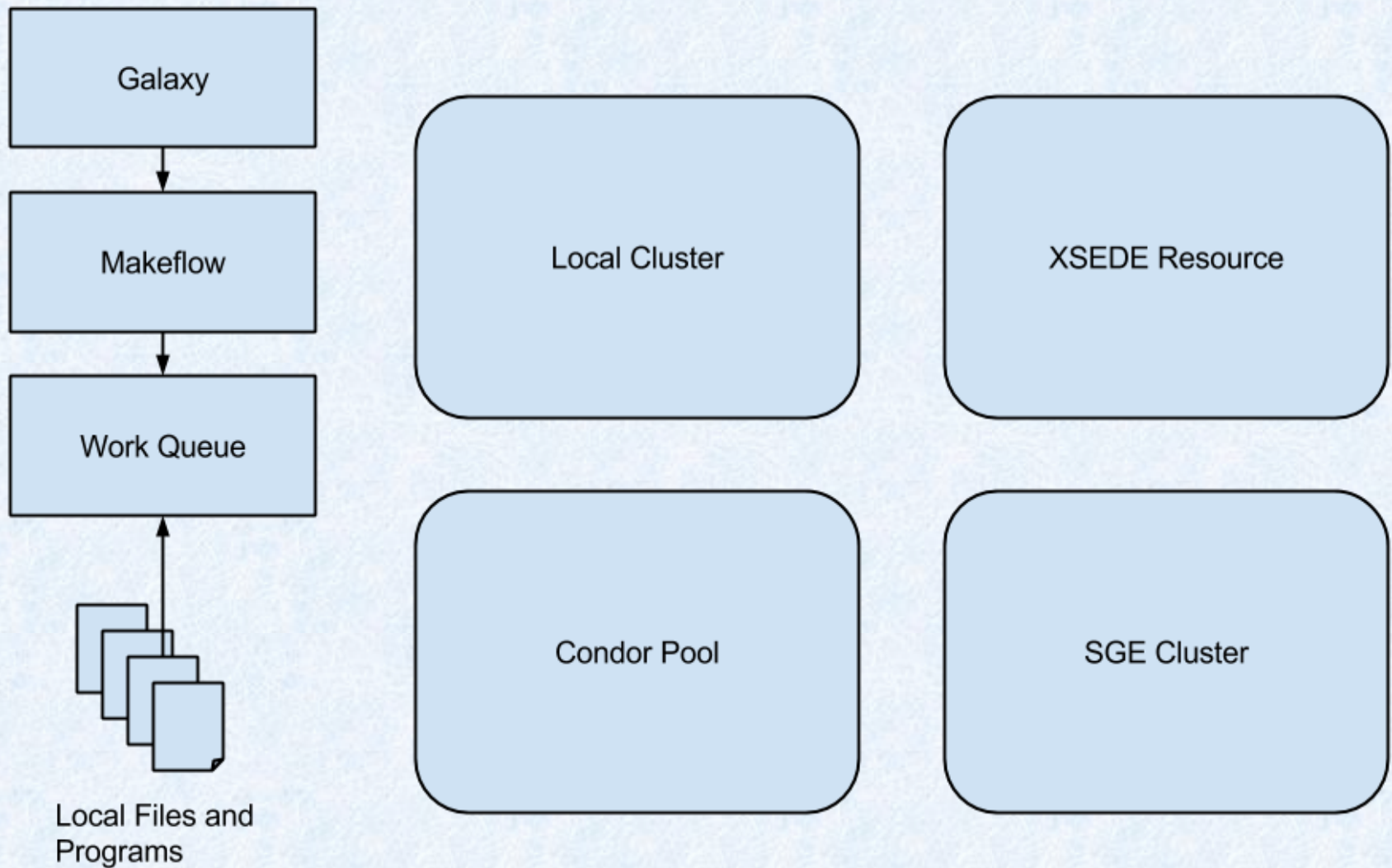
- Task Structure

INPUTS : OUTPUTS

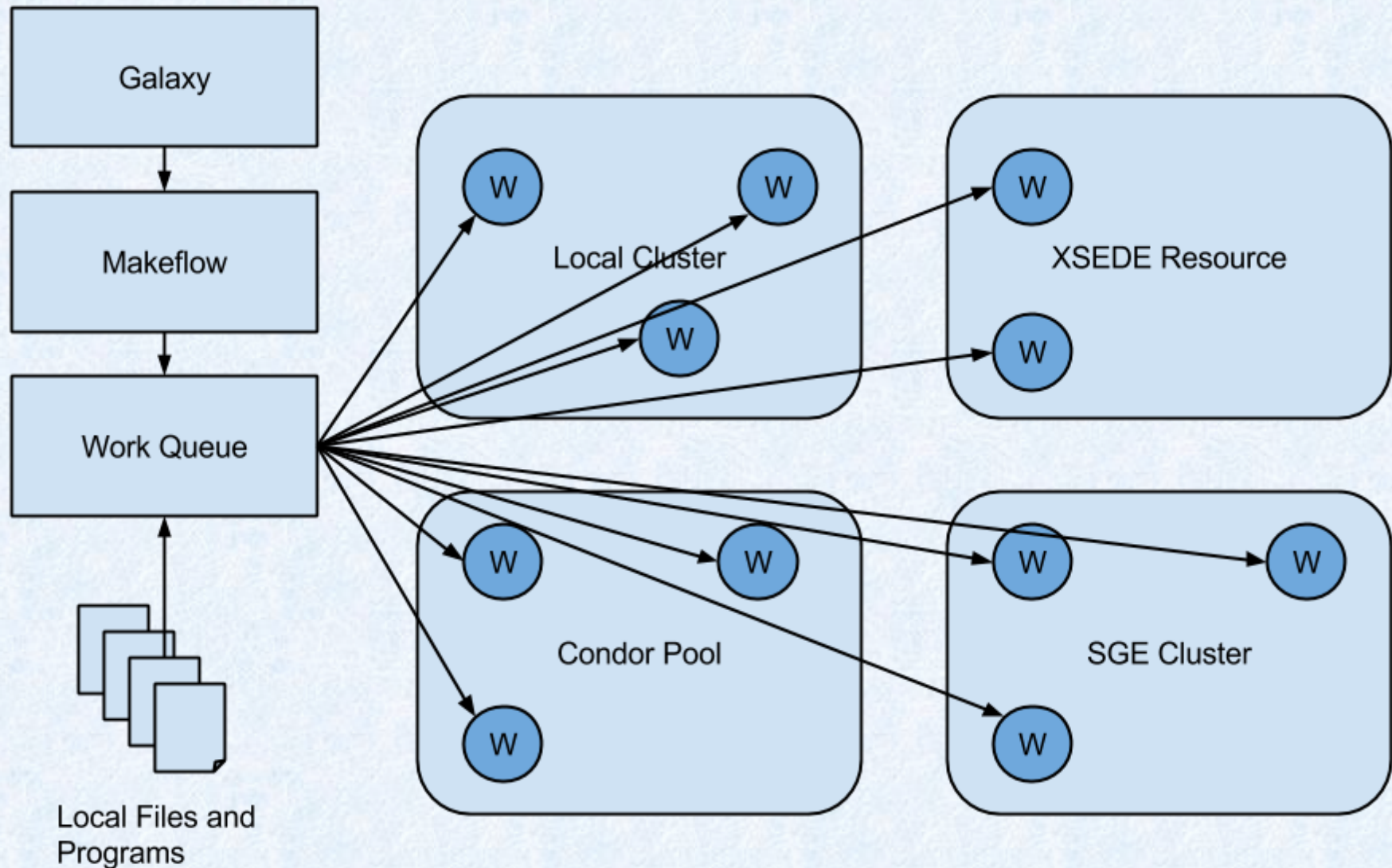
COMMAND

- Directed Acyclic Graph (DAG)
- Programmatically Generated

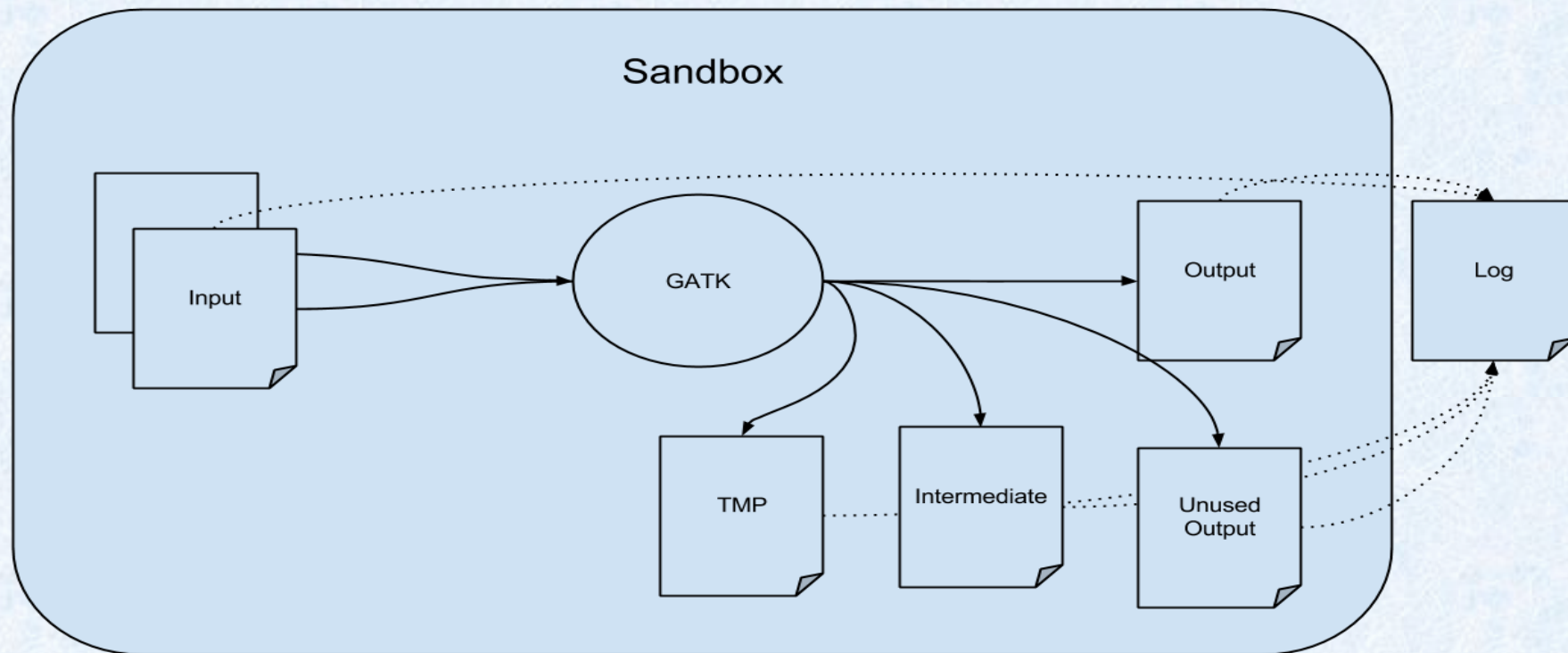








## Job Sandbox – Log file creation for cleanup



## Dynamic Job Expansion

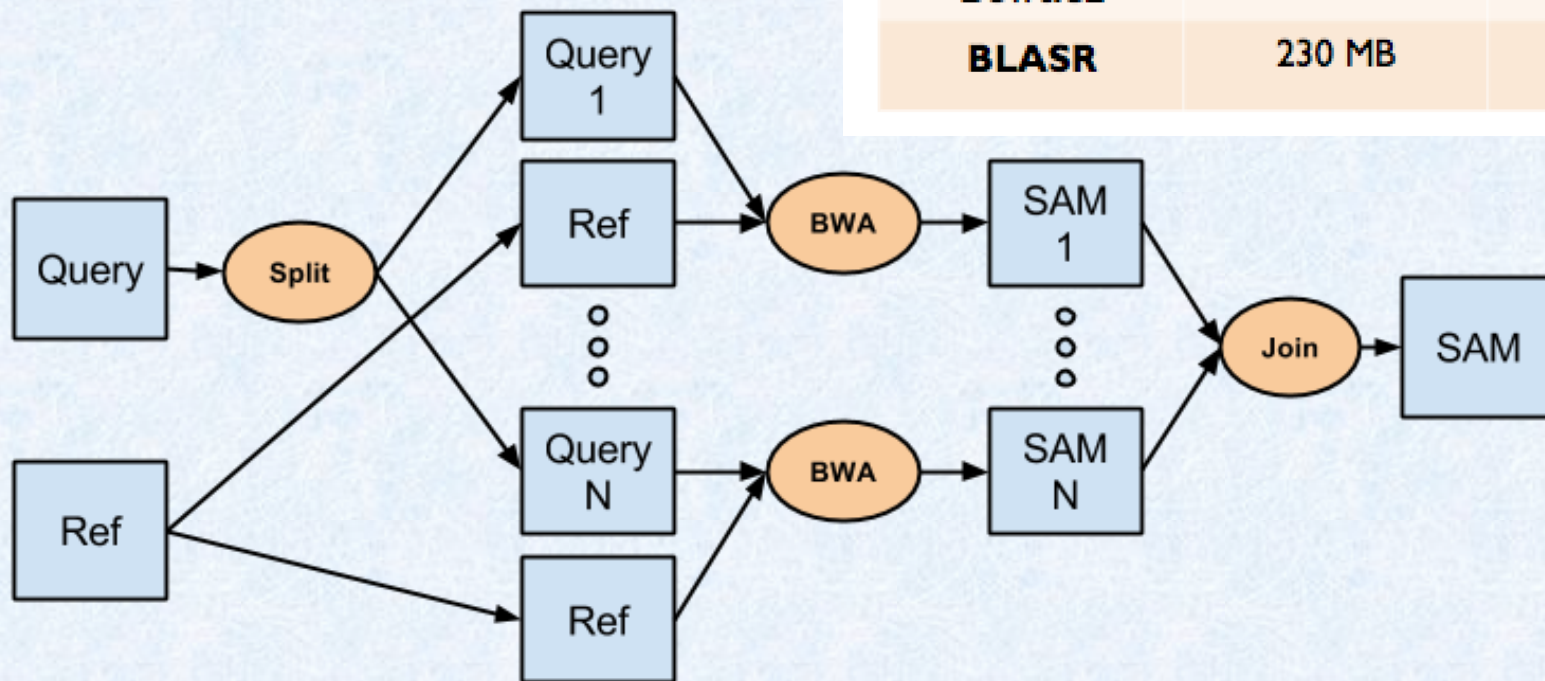
- Work Queue: we utilized 100s of cores from a Condor Pool
- Cleaning Sandbox using knowledge of intermediates and logging
- Explored methods to transmit needed environments such as executables and Java

61.5X speed-up on 32 GB dataset utilizing these methods

- Develop predictive performance models for an application domain
- Achieve acceptable performance the first time
- Optimize resource utilization
  - Execution time
  - Memory usage



Tools	Data	
	Reference	Query
<b>BWA, Bowtie2</b>	562 MB	45 GB
<b>BLASR</b>	230 MB	1 GB



- WorkQueue master-worker framework
- Sun Grid Engine (SGE) batch system

## 1. Application-level model for time:

$$T(R, Q, N) = \beta_1 RQ/N + \beta_2$$

## 2. Application-level model for memory:

$$M(R, N) = \gamma_1 R + \gamma_2 N$$

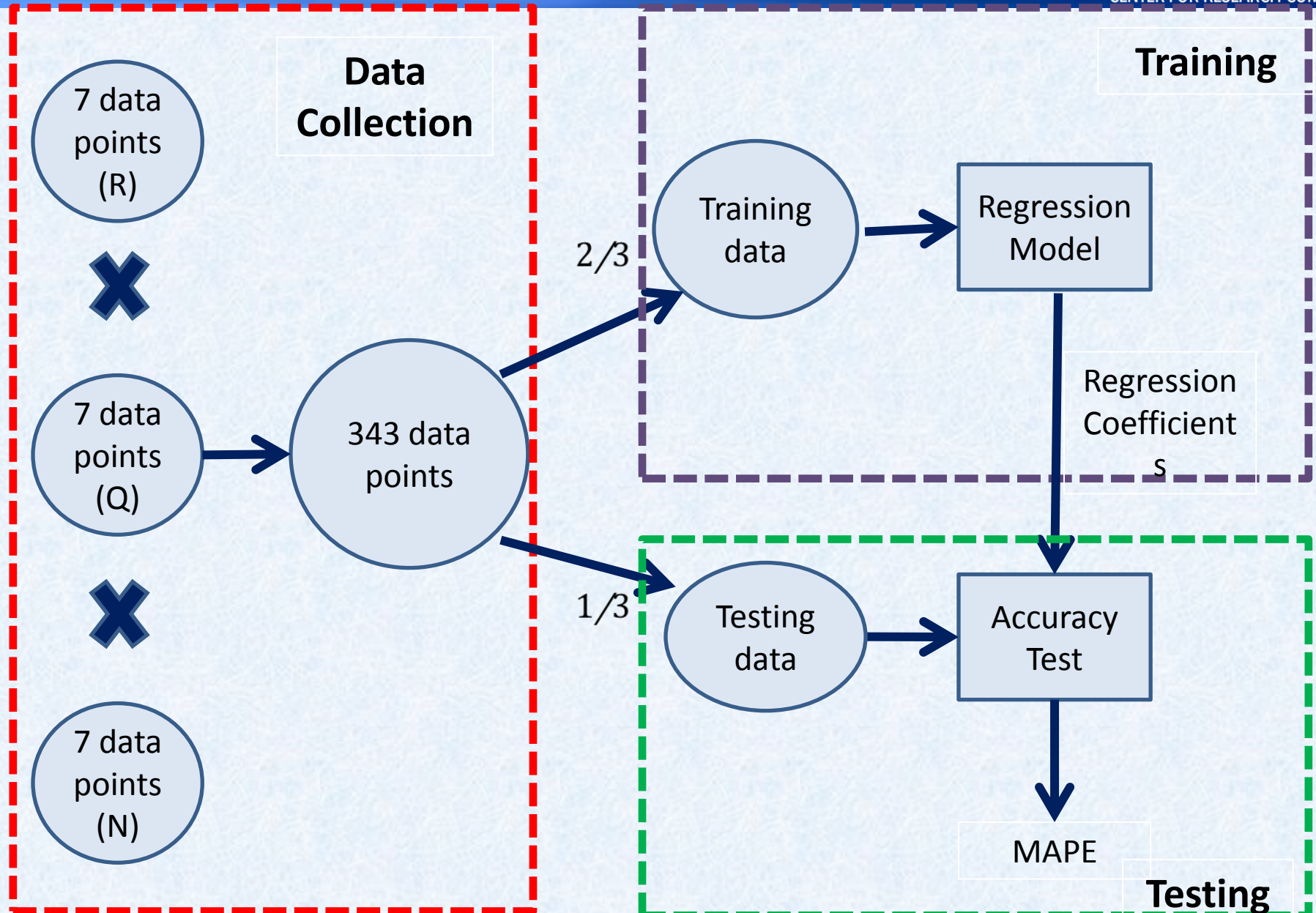
## 3. System-level model for time:

$$T_{Total} = \eta_1 QK/D + \eta_2 (Q/B + RKN/BC) + \eta_3 T(R, Q/K, N) * KN/MC + \eta_4 O/B + \eta_5 OK/D$$

## 4. System-level model for memory:

$$M_{Master}(R, Q) = \phi_1 R + \phi_2 Q$$

Terms	Meaning
R	Reference size
Q	Query size
N	No. of threads
K	No. of tasks
D	Disk speed
B	Network bandwidth
M	No. of available machines
C	No. of cores/machine

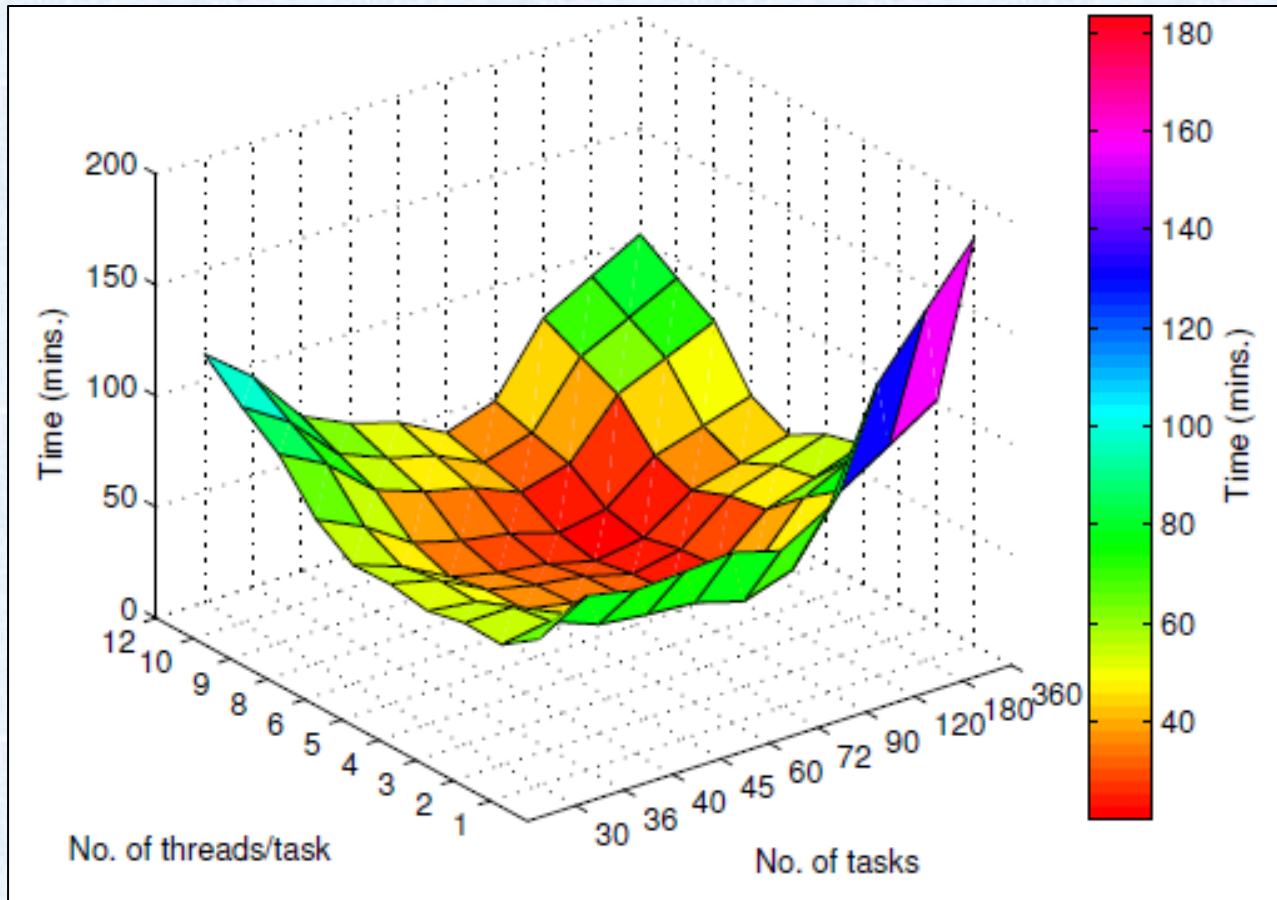


Model	Application	Configuration	MAPE(%)
Application-level Model for Time	BWA	Vary R, Fix Q,N	3.4
		Vary Q, Fix R,N	3.8
		Vary N, Fix R,Q	2.6
	Bowtie2	Vary R, Fix Q,N	1.6
		Vary Q, Fix R,N	2.2
		Vary N, Fix R,Q	1.3
	BLASR	Vary R, Fix Q,N	4.3
		Vary Q, Fix R,N	5.1
		Vary N, Fix R,Q	3.6
Application-level Model for Memory	BWA	Vary R, Fix N	3.9
		Vary N, Fix R	3.3
	Bowtie2	Vary R, Fix N	2.6
		Vary N, Fix R	1.9
	BLASR	Vary R, Fix N	4.7
		Vary N, Fix R	4.2
System-level Model for Time		Vary K, Fix R,Q,P	2.1
		Vary N, Fix R,Q,P	2.7
System-level Model for Memory		Vary R, Fix Q	2.5
		Vary Q, Fix R	3.3

Avg. MAPE  
= 3.1

MAPE = Mean Absolute Percentage Error





For the given dataset,  $K^* = 90$ ,  $N^* = 4$

# Cores/ Task	# Tasks	Predicted Time (min)	Speedup	Estimated EC2 Cost (\$)	Estimated Azure Cost (\$)
1	360	70	6.6	50.4	64.8
2	180	38	12.3	25.2	32.4
4	90	24	19.5	18.9	32.4
8	45	27	17.3	18.9	32.4

- Science Gateway Institute  
<http://sciencegateways.org>
- Science Gateway Workshops  
Europe: IWSG - <http://iwsg.info>  
USA: GCE - <http://sciencegateways.org>  
Australasia: IWSG-A - <http://iwsg.info>
- IEEE Technical Area on Science Gateways  
<http://ieeesciencegateways.org>
- XSEDE Science Gateways  
<https://www.xsede.org/gateways-overview>
- CRC Science Gateways  
<https://crc.nd.edu/index.php/research/gateways>

Questions and exercises at  
<http://bit.ly/2dlkySW>

Data at  
<http://bit.ly/2cTwKaN>





[sandra.gesing@nd.edu](mailto:sandra.gesing@nd.edu)