Introduction

Dr Katy Wolstencroft

Introduction to the Tutorial

- Using scientific workflows for bioinformatics data analysis, integration and scaling
  - Galaxy – local tools workflows
  - Taverna – distributed computing workflows
  - WS-PGRADE - grid workflows
- RNA-Seq analysis pipeline
- 3 hands-on session
- Discussion
- Workflows surgery session

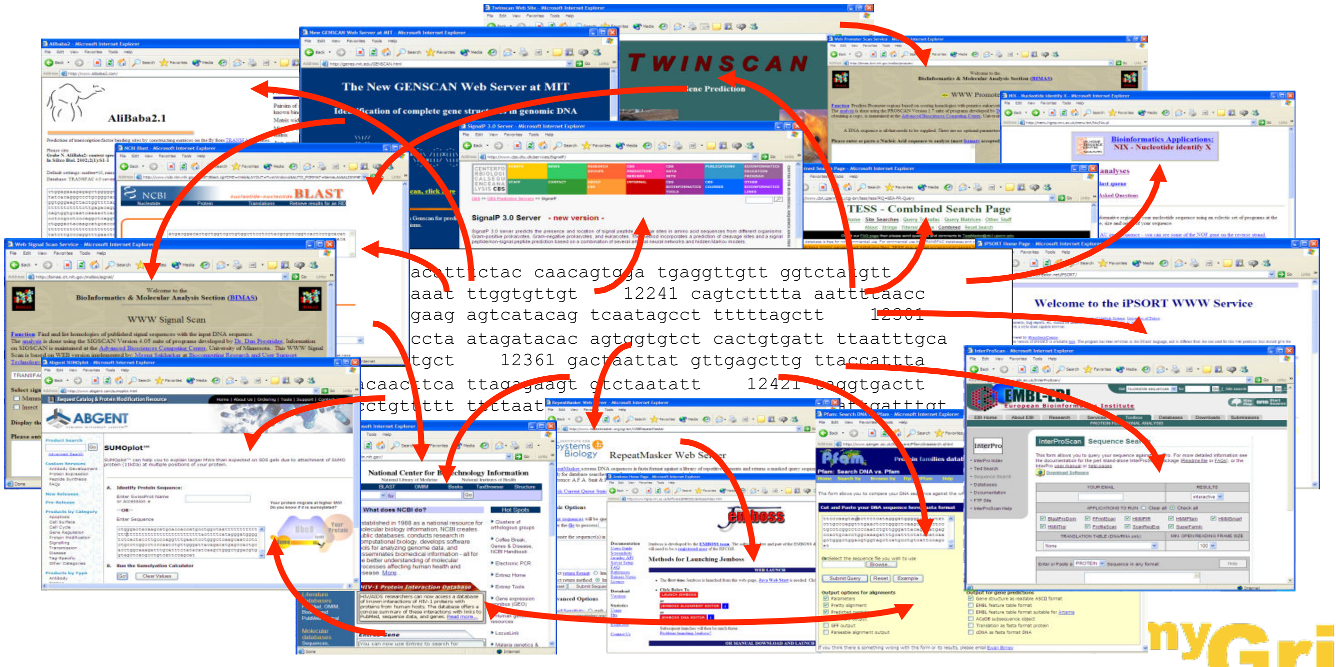
Introduction to Scientific Workflows

Typical Bioinformatics Experiments

1. Collect data / find data in repositories
2. Run analysis algorithm on data
3. Run different analysis algorithm on results of (2) or get more/different data and run (2) again
4. Repeat for any number of sequential algorithms
5. Analyse results
6. Search for supporting information to help explain results and generate new hypotheses
7. Start cycle again with new hypotheses

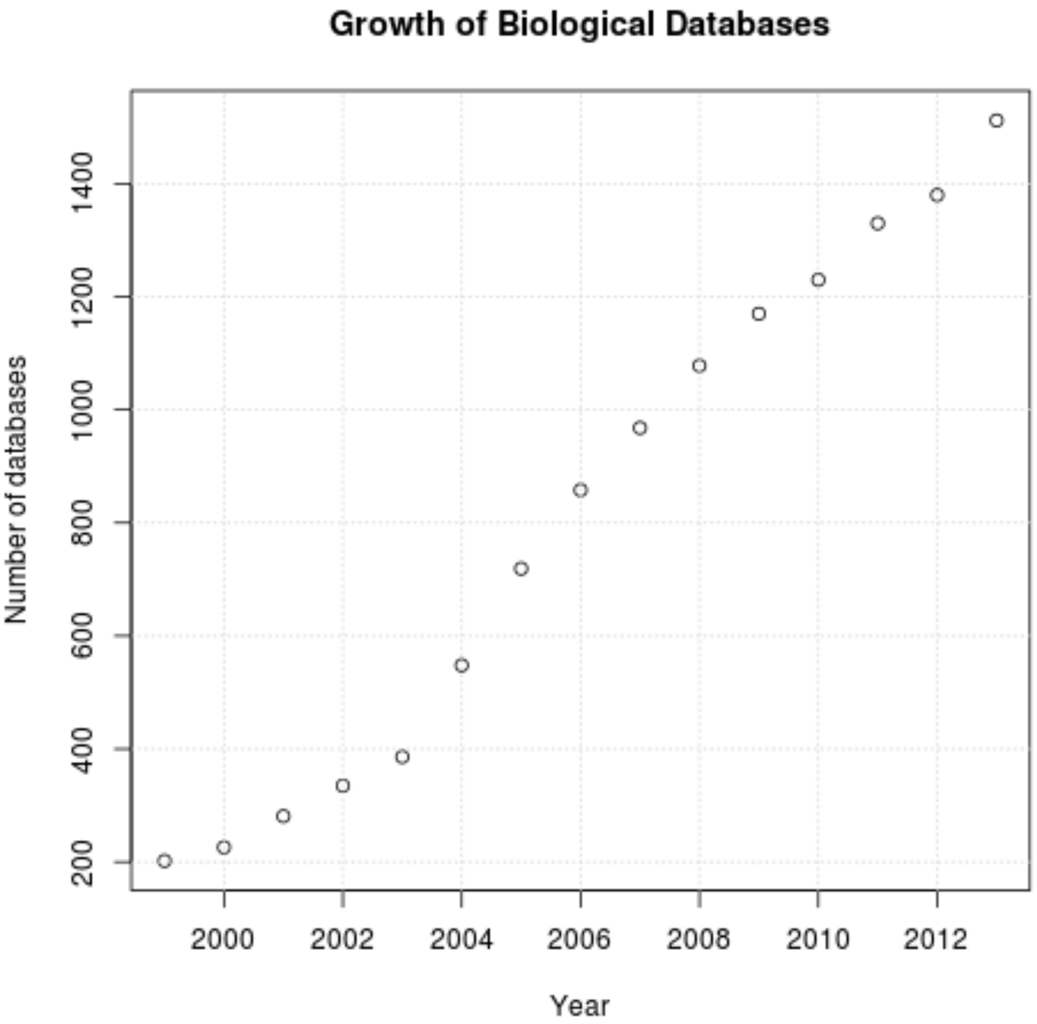
The Practical Reality

- Sequential use of distributed tools
- Incompatible input and output formats
- Analysis of large data sets by multiple researchers
- Difficult to record parameter selections
- Difficult to reproduce analyses


TWINSCAN
Gene Prediction
The New GENSCAN Web Server at MIT
Identification of complete gene structures in genomic DNA
AliBaba2.1
NCBI
BLAST
SignalP 3.0 Server
SignalP 3.0 Server - new version -
Welcome to the BioInformatics & Molecular Analysis Section (BIMAS)
WWW Promoter Scan
WWW Signal Scan
Bioinformatics Applications:
NIX - Nucleotide identify X
TESS - Combined Search Page
Welcome to the iPSORT WWW Service
EMBL-EBI
European Bioinformatics Institute
InterProScan Sequence Search
Pfam
Protein families database
Pfam: Search DNA vs. Pfam
emboss
Methods for Launching Jemboss
RepeatMasker Web Server
Systems Biology
ABGENT
SUMOplot™
National Center for Biotechnology Information
What does NCBI do?
HIV-1 Protein Interaction Database
acttctctac caacagtcga tgaggttgtt ggtctatgtt
aaat ttggtgttgt 12241 cagtctttta aattttaacc
gaag agtcatacag tcaatagcct tttttagctt 12301
ccta atagatacac agtggtgtct cactgtgatt ttaatttgca
tgct 12361 gactaattat gttgagcttg ttaccattta
acaactttca ttagagaagt cctctaatatt 12421 caggtgactt
cctgttttt ttttaat atgatttgt

Lots of Resources

NAR 2014 – 1552 databases



Sources of Data

- In repositories run by major service providers (e.g. NCBI, EBI, DDBJ)
- Group/Institute web sites
- On ftp servers
- In local project stores

- Few defined formats
- Inconsistent metadata

National Center for Biotechnology Information (USA)

Tokyo, Japan

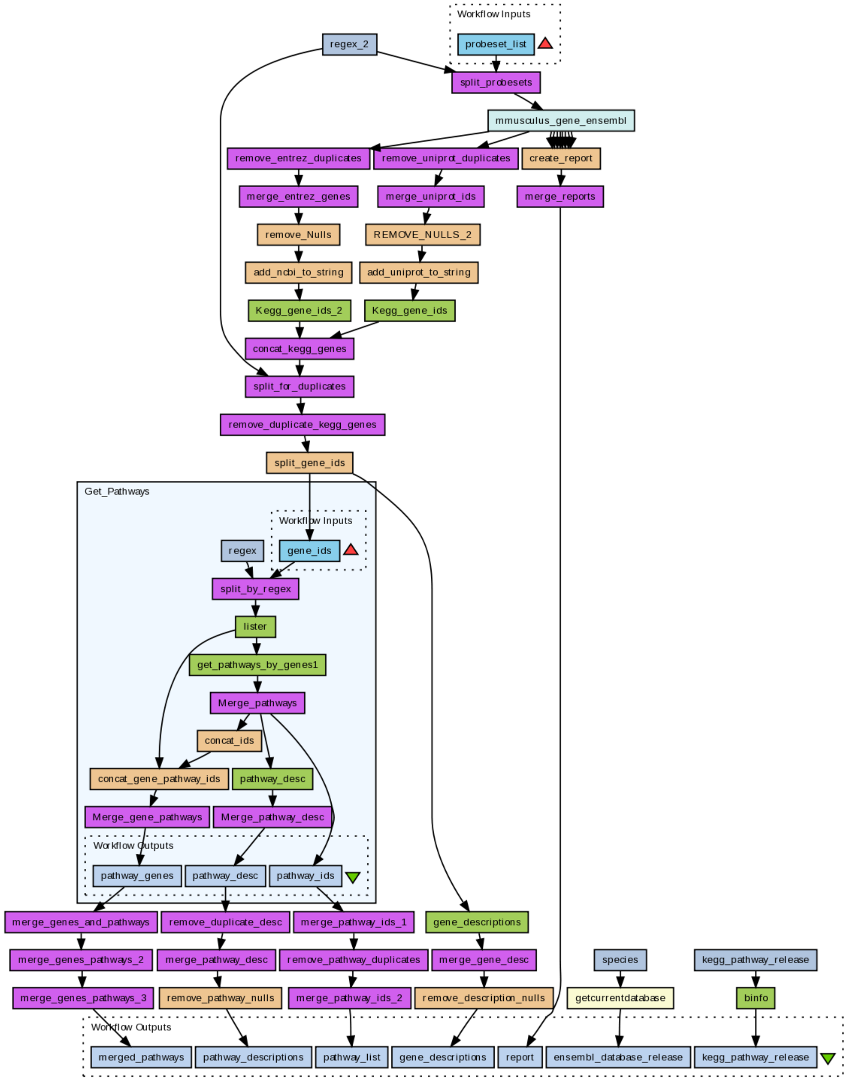
Cambridge, UK

SRS

SeqHound

Workflow as a Solution

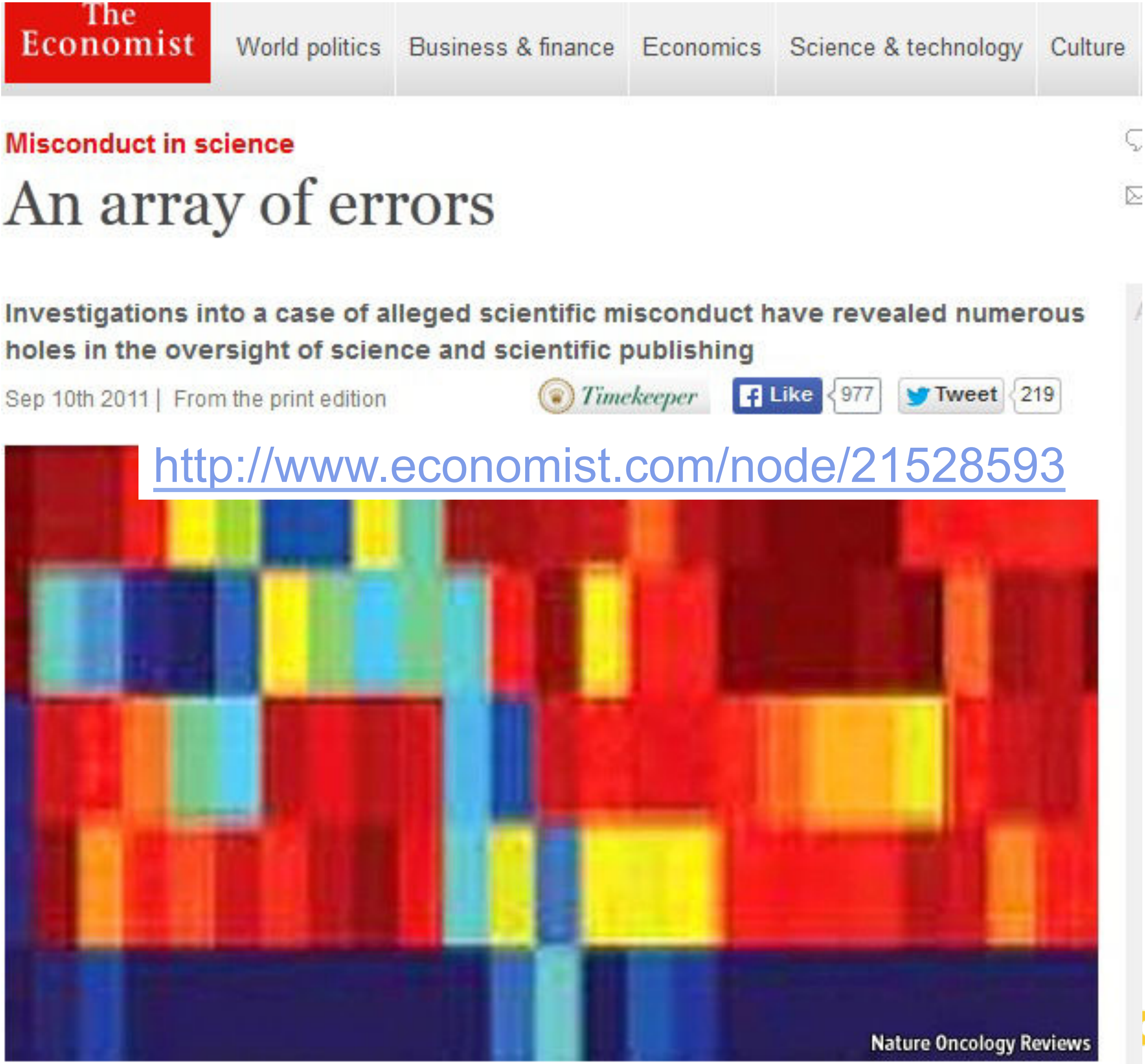
- Automating the process
- Sophisticated analysis pipelines
- A set of **services** to analyse or manage data (either local or remote)
- Data flow through services
- Control of service invocation
- Iteration
- Encapsulates experimental method



Promoting Reproducible Research

Informatics involves

- Complex, multi-step analyses
- Lots of data as inputs
- Lots of data generated
- Workflows encapsulate the methods used and their parameters and the input data
  - Easier to repeat
- Workflows allow you to visualise the methods
  - Easier to assess



Misconduct in science

An array of errors

Investigations into a case of alleged scientific misconduct have revealed numerous holes in the oversight of science and scientific publishing

Sep 10th 2011 | From the print edition

http://www.economist.com/node/21528593

Nature Oncology Reviews

Preventing Irreproducible Research

- Duke University, 2006 -Prediction of the course of a patient’s lung cancer using expression arrays
- *Nature Medicine* - **12**, 1294 - 1300 (2006)
- Recommendations on different chemotherapies from cell cultures
- Clinical trials of personalised treatments
- Major advances in personalised medicine?

Preventing Irreproducible Research

- 3 different groups **could not reproduce** the results and **uncovered mistakes** in the original work
- Requested raw data and more information about all the methods
- Re-created whole analysis methods
- *Nature Medicine* - **12**, 1294 - 1300 (2006) Published online: 22 October 2006; Corrected online: 27 October 2006; Corrected online: 21 July 2008; Retracted: 07 January 2011 | doi: 10.1038/nm1491

If the Analyses were done using Workflows.....

- Reviewers could re-run experiments and see results for themselves
- Methods could be properly examined and criticised
- Mistakes could be pinpointed
  - At what point in the analysis pipeline did it start to go wrong
  - What were the inputs and outputs for each step

Workflows are …

... records and protocols (i.e. your *in silico* experimental method)

... know-how and intellectual property

... hard work to develop and get right

…..re-usable methods (i.e. you can build on the work of others)

So why not share and re-use them

SHIWA

myexperiment

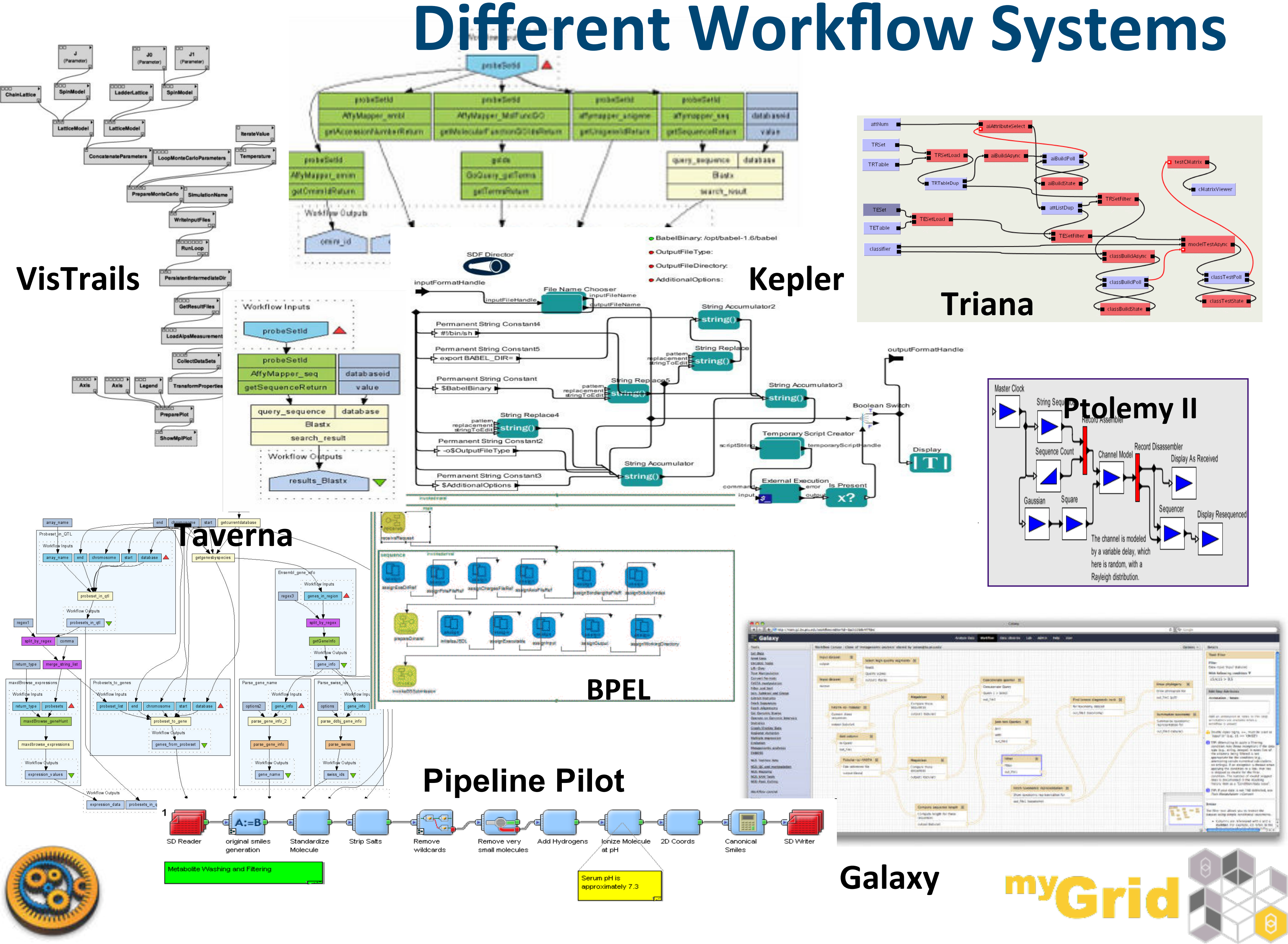
Published workflows

Workflows are ideal for…

- High throughput analysis
  - Transcriptomics, proteomics, next gen sequencing
- Data integration, data interoperation
- Data management
  - Model construction
  - Data format manipulation
  - Database population
  - Semantic integration
  - Visualisation

SCIENTIFIC WORKFLOW MANAGEMENT SYSTEMS

Different Workflow Systems



VisTrails

Kepler

Triana

Ptolemy II

Taverna

BPEL

Pipeline Pilot

Galaxy

myGrid

Example Workflow Systems

Galaxy

https://usegalaxy.org/

Penn State and Emory University James Taylor and Anton Nekrutenko

UK Consortium, led by Carole Goble, University of Manchester

Taverna

http://www.taverna.org.uk/

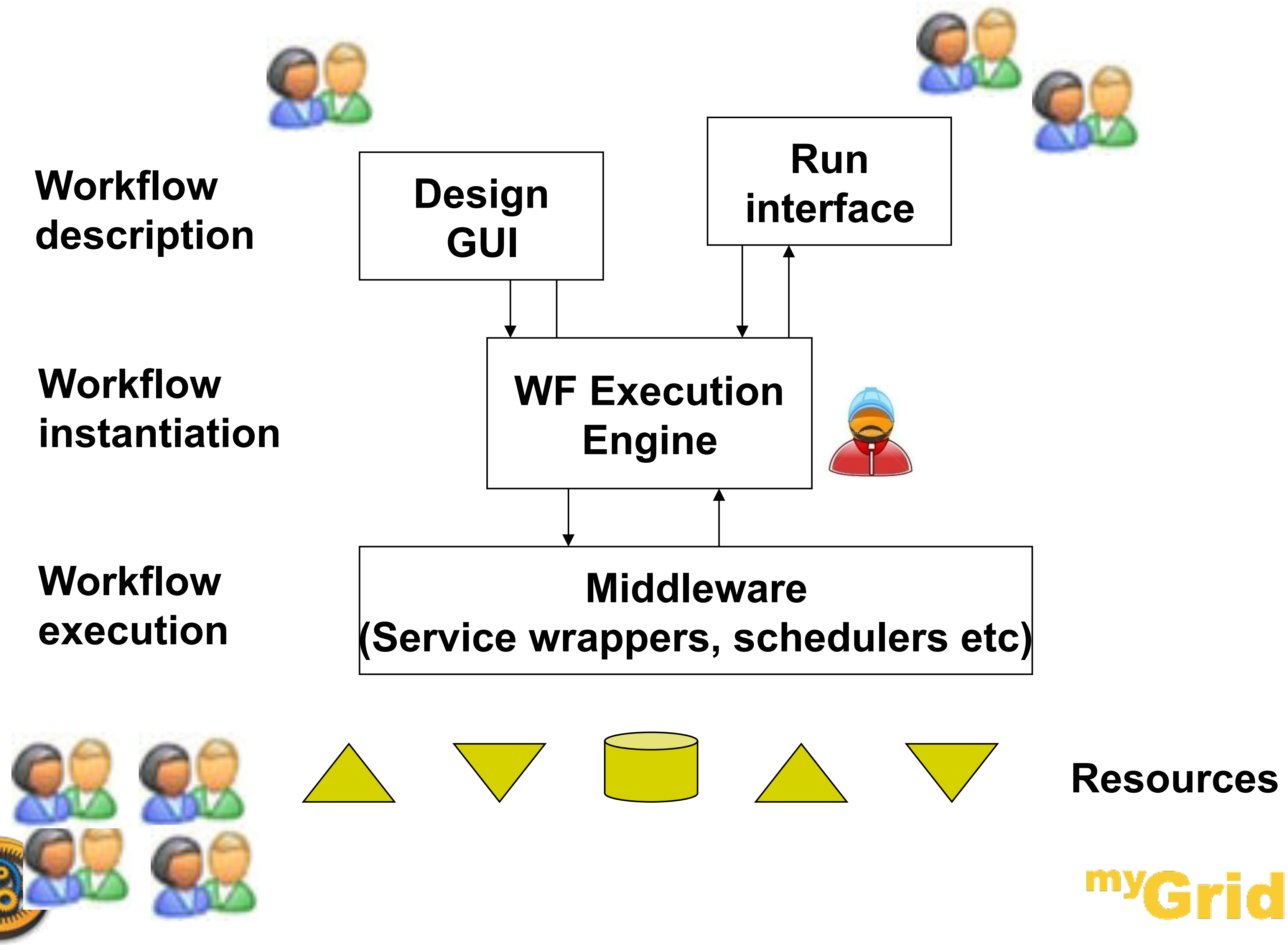
gUSE grid and cloud User Support Environment

http://guse.hu/about/architecture/ws-pgrade

Supported by the ER-Flow project

myGrid

All Workflow Systems at 50,000 feet



Workflow description

Design GUI

Run interface

Workflow instantiation

WF Execution Engine

Workflow execution

Middleware
(Service wrappers, schedulers etc)

Resources

Different Types of Workflows

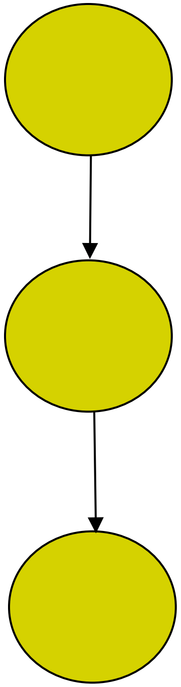
- Two main types of workflows:
  - Data flow workflows

    A task is invoked once its **expected data** has been received. When complete, it passes any resulting data downstream
  - Control flow workflows

    A task is invoked once its **dependant tasks** have been completed
  - Combination

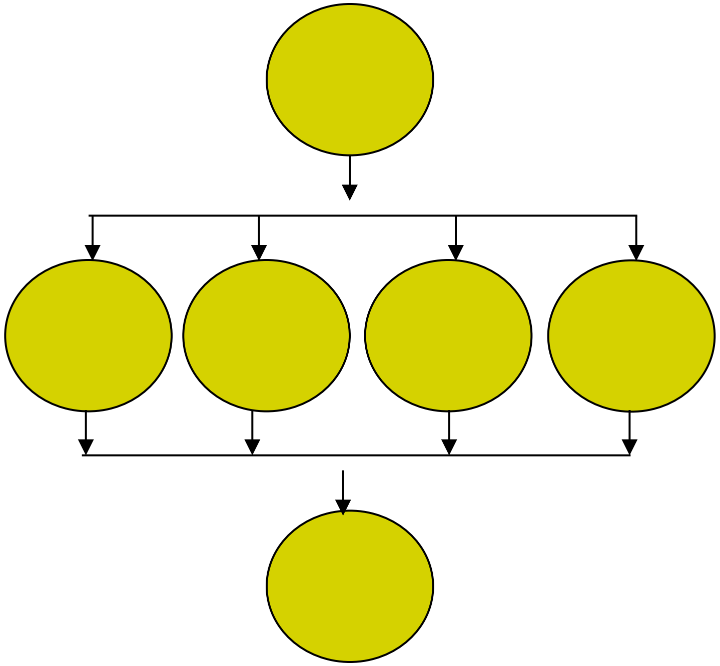
    A task is invoked once its **expected data** has been received AND **dependant tasks** have been completed

Possible Data Flow Structures



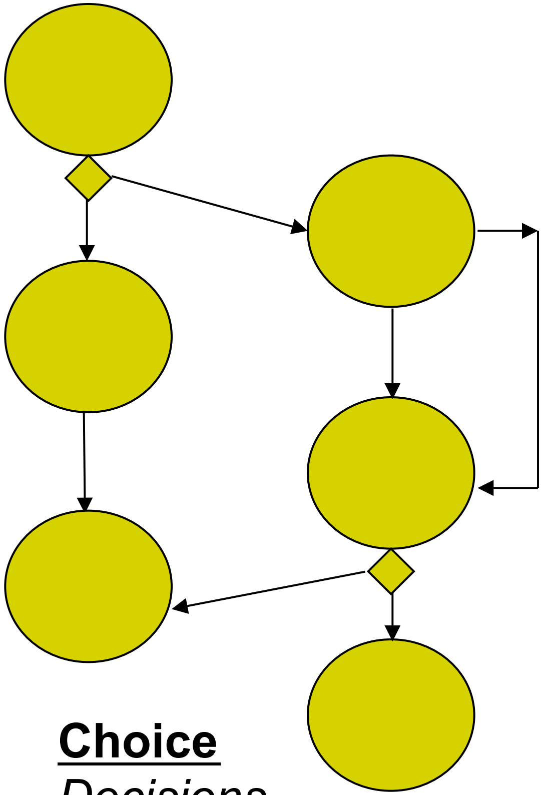
Sequence

Store intermediate results



Parallel

Apply multiple components to a set of data



Choice

Decisions at runtime

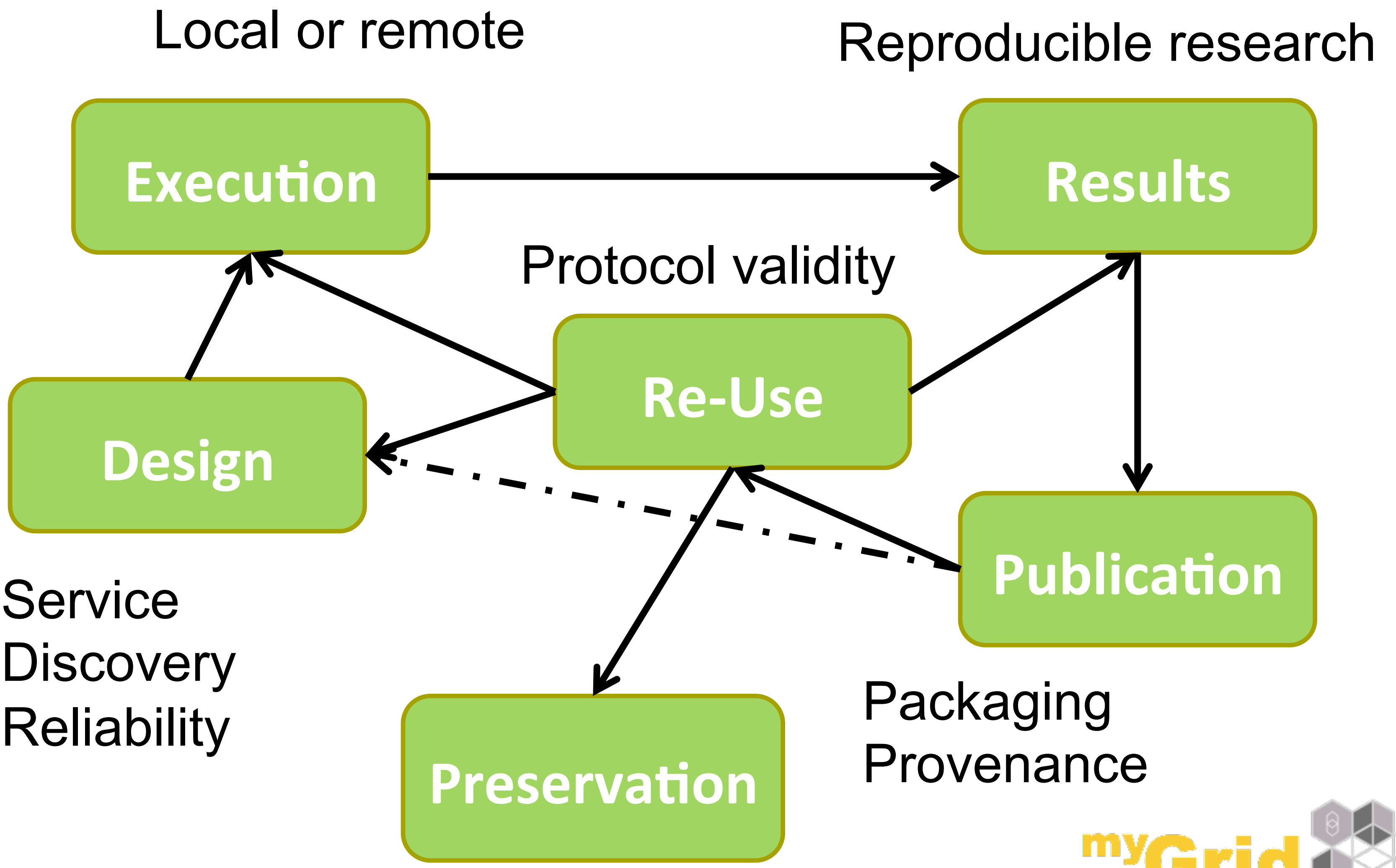
Iteration

Loop through datasets

Workflow engine features

- Implicit iterations
- Parallelisation
  - Run as soon as data is available
- Streaming
  - Process partial iteration results early
- Retries, looping
  - For stability and conditional testing

Workflows Supporting *in silico* Science



SERVICES IN WORKFLOWS

Services in workflows

- Web Services
  - WSDL
  - REST
- Local services
- Grid Services
- Cloud Services
- Workflows
- Secure Services

Different Approaches to Service Connections

- Closed – connect to services installed locally to the server,designed specifically to work together
  - Less heterogeneity, but fewer services
  - Harder to add new services
  - Galaxy server, Knime
- Open – connect to ANY service regardless of type and structure, hosted anywhere in the world
  - More services, but more heterogeneity
  - Easy to add new services
  - Taverna, Kepler

Different Approaches to Service Connections

- On-the-fly service execution
  - Data and legacy code (i.e. Command-line application) are submitted to a workflow engine for execution
  - A language or interactive mechanism to describe the code that needs to be executed, its dependencies, and the arguments. The system in some cases also takes
  - Transports the data to the computing resource where the job is executed

Using and Making Galaxy Services

- Tools installed on the Galaxy server (admin)
- Create Galaxy tool definition file
- Create an entry in the Galaxy tool registry tool_conf.xml
- Finite, but compatible collection of resources
- Focus on genomics and NGS, but expanding
- Includes built-in genome browsers

Galaxy Tool Shed

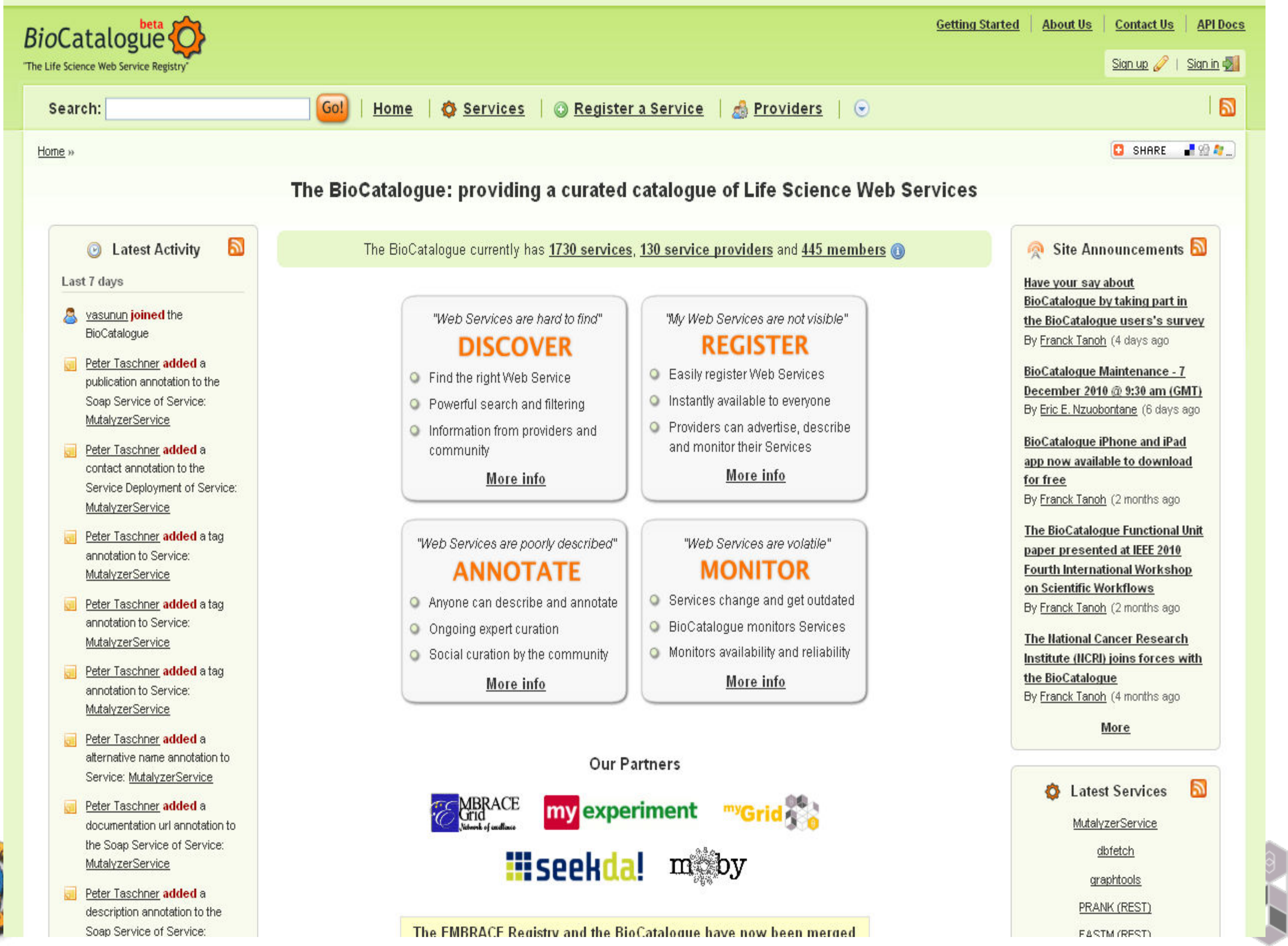
Using Services with Taverna

- Open domain
- Web Services
  - WSDL or REST - 8000 +
- Local services
- Grid Services
- Cloud Services
- Secure Services
- Workflows
- Specialised services
  - BioMart , R

No installation required!

- Third party – we don't own them – we didn't build them
- All the major providers
  - NCBI, DDBJ, EBI …
- Enforce NO common data model

Understanding how services work



BioCatalogue beta
"The Life Science Web Service Registry"

Getting Started | About Us | Contact Us | API Docs

Sign up | Sign in

Search: Go! | Home | Services | Register a Service | Providers |

Home »

SHARE

The BioCatalogue: providing a curated catalogue of Life Science Web Services

Latest Activity

Last 7 days

- vasunun joined the BioCatalogue
- Peter Taschner added a publication annotation to the Soap Service of Service: MutalyzerService
- Peter Taschner added a contact annotation to the Service Deployment of Service: MutalyzerService
- Peter Taschner added a tag annotation to Service: MutalyzerService
- Peter Taschner added a tag annotation to Service: MutalyzerService
- Peter Taschner added a tag annotation to Service: MutalyzerService
- Peter Taschner added a alternative name annotation to Service: MutalyzerService
- Peter Taschner added a documentation url annotation to the Soap Service of Service: MutalyzerService
- Peter Taschner added a description annotation to the Soap Service of Service:

The BioCatalogue currently has 1730 services, 130 service providers and 445 members

"Web Services are hard to find"

DISCOVER

- Find the right Web Service
- Powerful search and filtering
- Information from providers and community

More info

"My Web Services are not visible"

REGISTER

- Easily register Web Services
- Instantly available to everyone
- Providers can advertise, describe and monitor their Services

More info

"Web Services are poorly described"

ANNOTATE

- Anyone can describe and annotate
- Ongoing expert curation
- Social curation by the community

More info

"Web Services are volatile"

MONITOR

- Services change and get outdated
- BioCatalogue monitors Services
- Monitors availability and reliability

More info

Our Partners

EMBRACE Grid Network of excellence | myexperiment | myGrid | seekda! | moby

The EMBRACE Registry and the BioCatalogue have now been merged

Site Announcements

Have your say about BioCatalogue by taking part in the BioCatalogue users's survey
By Franck Tanoh (4 days ago

BioCatalogue Maintenance - 7 December 2010 @ 9:30 am (GMT)
By Eric E. Nzuobontane (6 days ago

BioCatalogue iPhone and iPad app now available to download for free
By Franck Tanoh (2 months ago

The BioCatalogue Functional Unit paper presented at IEEE 2010 Fourth International Workshop on Scientific Workflows
By Franck Tanoh (2 months ago

The National Cancer Research Institute (NCRI) joins forces with the BioCatalogue
By Franck Tanoh (4 months ago

More

Latest Services

MutalyzerService

dbfetch

graphtools

PRANK (REST)

FASTM (REST)

Using and Making WS-Pgrade Services

- On-the-fly service composition
- Execution of web services
- Execution of workflows

Data and Provenance

Data and Provenance

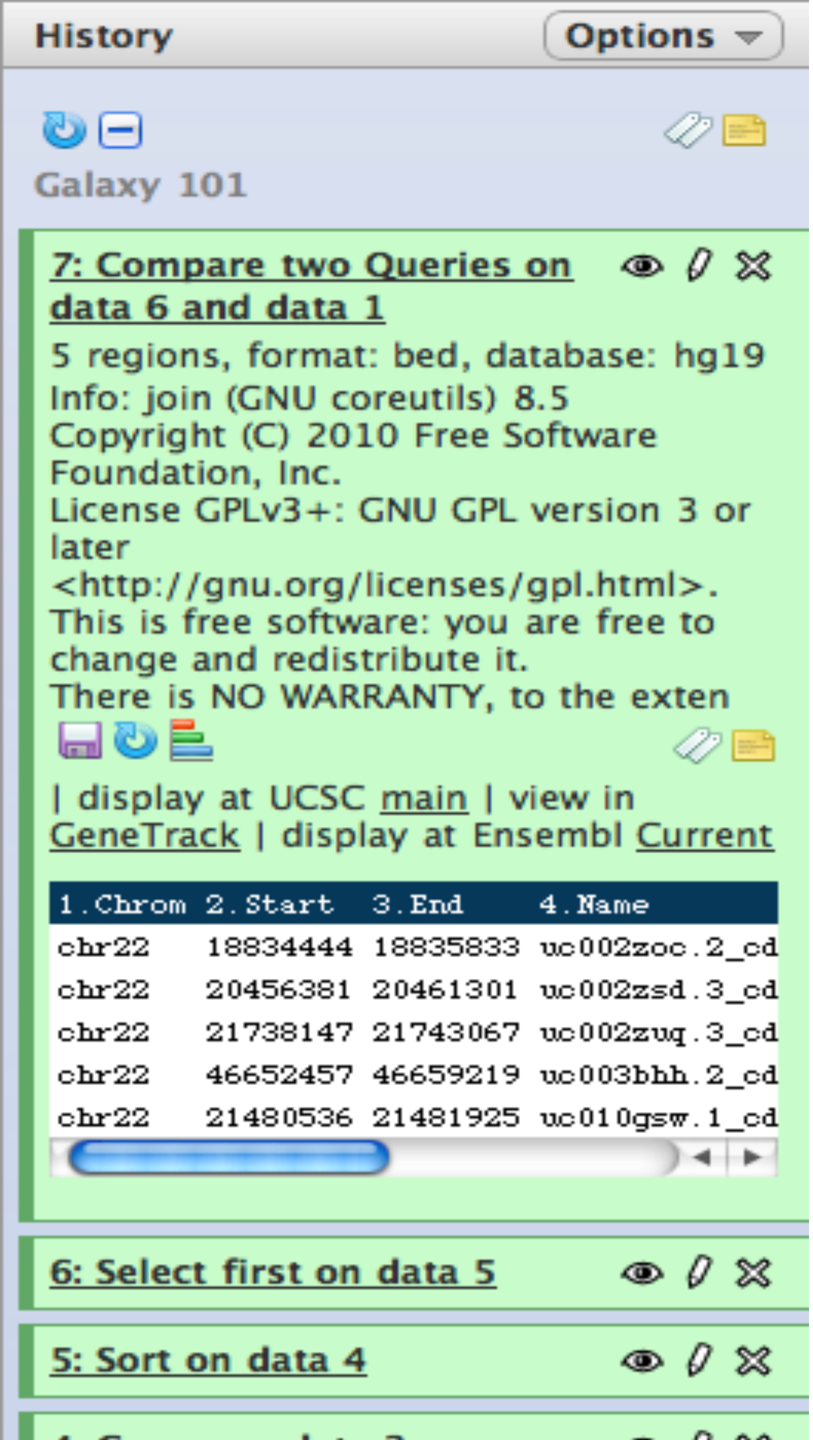
- Workflows can generate vast amount of data - how can we manage and track it?
- We need to manage data AND metadata AND experimental provenance
  - Input data
  - Parameter selection
  - Versions of the analysis tools used
  - Intermediate results
  - Intermediate parameter selection
  - Final results
  - What happened, what went wrong, how long it took

Data and Provenance

Scientists need to:

- Check back over past results, compare workflow runs and share workflow runs with colleagues
- Look at intermediate results when designing and debugging

Histories



History Options

Galaxy 101

7: Compare two Queries on data 6 and data 1

5 regions, format: bed, database: hg19
Info: join (GNU coreutils) 8.5
Copyright (C) 2010 Free Software Foundation, Inc.
License GPLv3+: GNU GPL version 3 or later <http://gnu.org/licenses/gpl.html>.
This is free software: you are free to change and redistribute it.
There is NO WARRANTY, to the exten

| display at UCSC main | view in GeneTrack | display at Ensembl Current

1.Chrom	2.Start	3.End	4.Name
chr22	18834444	18835833	uc002zoc.2_cd
chr22	20456381	20461301	uc002zsd.3_cd
chr22	21738147	21743067	uc002zuq.3_cd
chr22	46652457	46659219	uc003bhh.2_cd
chr22	21480536	21481925	uc010gsw.1_cd

6: Select first on data 5

5: Sort on data 4

- What you did
- With what data
- In what order

- Automatically converts histories to workflows

myGrid

Where Workflows are Executed

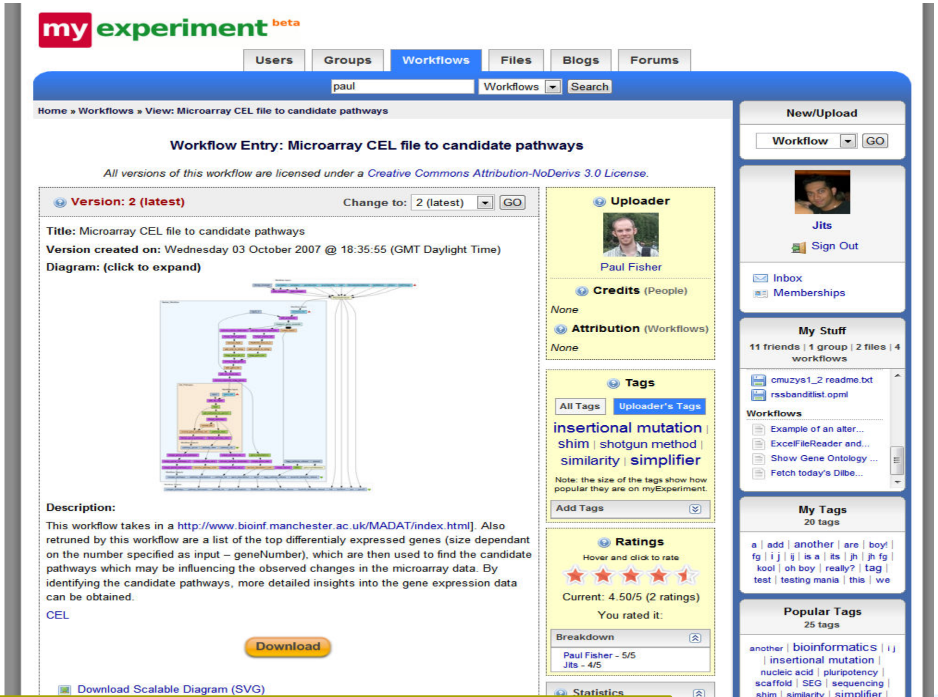
Where workflows are executed

- Local Execution
  - Client supports complex workflow design
  - Download and install client
  - Easy to access and use local data and tools
  - Easy to store results

- Server execution
  - Workflows executed through web interface or other client
  - Workflows run on server, grid or cloud
  - Better for larger, long running workflows
  - Better for scaling-up
  - Serves finished workflows to users

Spectrum of Users

Intermediate users reuse and modify existing workflows



Home » Workflows » View: Microarray CEL file to candidate pathways

Workflow Entry: Microarray CEL file to candidate pathways

All versions of this workflow are licensed under a Creative Commons Attribution-NoDerivs 3.0 License.

Version: 2 (latest) Change to: 2 (latest) GO

Title: Microarray CEL file to candidate pathways

Version created on: Wednesday 03 October 2007 @ 18:35:55 (GMT Daylight Time)

Diagram: (click to expand)

Description:

This workflow takes in a http://www.bioinf.manchester.ac.uk/MADAT/index.html]. Also retruned by this workflow are a list of the top differentialy expressed genes (size dependant on the number specified as input – geneNumber), which are then used to find the candidate pathways which may be influencing the observed changes in the microarray data. By identifying the candidate pathways, more detailed insights into the gene expression data can be obtained.

CEL

Download

Download Scalable Diagram (SVG)

Uploader

Paul Fisher

Credits (People)

None

Attribution (Workflows)

None

Tags

All Tags | Uploader's Tags

insertional mutation | shim | shotgun method | similarity | simplifier

Note: the size of the tags show how popular they are on myExperiment.

Add Tags

Ratings

Hover and click to rate

Current: 4.50/5 (2 ratings)

You rated it:

Breakdown

Paul Fisher - 5/5
Jits - 4/5

Statistics

New/Upload

Workflow GO

Jits

Sign Out

Inbox
Memberships

My Stuff

11 friends | 1 group | 2 files | 4 workflows

cmuzys1_2 readme.txt
rssbanditlist.opml

Workflows

Example of an alter...
ExcelFileReader and...
Show Gene Ontology ...
Fetch today's Dilbe...

My Tags

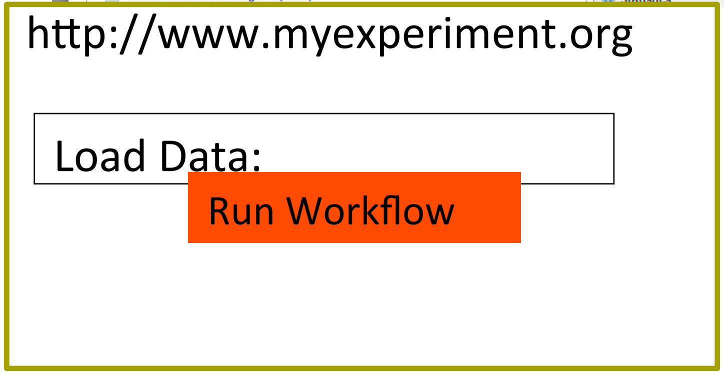
20 tags

a | add | another | are | boy! | fg | i j | ij | is a | its | jh | jh fg | kool | oh boy | really? | tag | test | testing mania | this | we

Popular Tags

25 tags

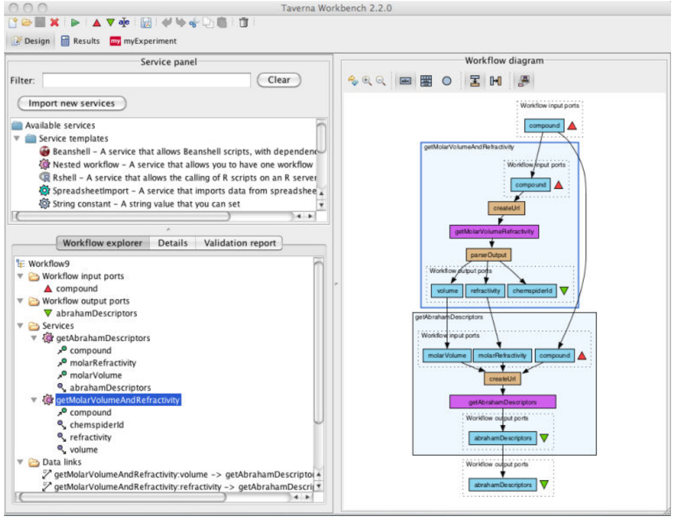
another | bioinformatics | i j | insertional mutation | nucleic acid | pluripotency | scaffold | SEG | sequencing | shim | similarity | simplifier |



http://www.myexperiment.org

Load Data:

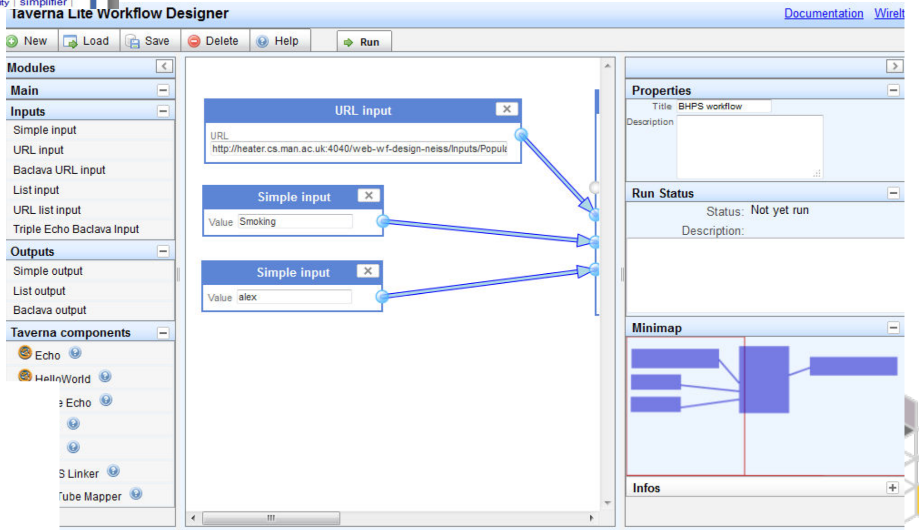
Run Workflow



Available services
- Service templates
  - Beanshell – A service that allows Beanshell scripts, with dependenc
  - Nested workflow – A service that allows you to have one workflow
  - Rshell – A service that allows the calling of R scripts on an R server
  - SpreadsheetImport – A service that imports data from spreadshee
  - String constant – A string value that you can set

Workflow9
- Workflow input ports
  - compound
- Workflow output ports
  - abrahamDescriptors
- Services
  - getAbrahamDescriptors
    - compound
    - molarRefractivity
    - molarVolume
    - abrahamDescriptors
  - getMolarVolumeAndRefractivity
    - compound
    - chemspiderId
    - refractivity
    - volume
- Data links
  - getMolarVolumeAndRefractivity:volume -> getAbrahamDescripto
  - getMolarVolumeAndRefractivity:refractivity -> getAbrahamDescri

Advanced users design and build workflows (informaticians)



URL input
URL http://heater.cs.man.ac.uk:4040/web-wf-design-neiss/Inputs/Popul

Simple input
Value Smoking

Simple input
Value alex

Title BHPS workflow

Description

Status: Not yet run

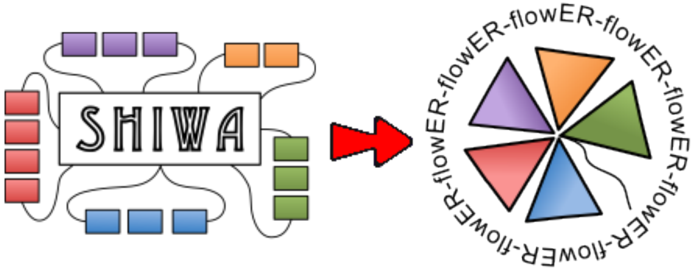
Description:

Others "replay" workflows through a web interface or Taverna Lite

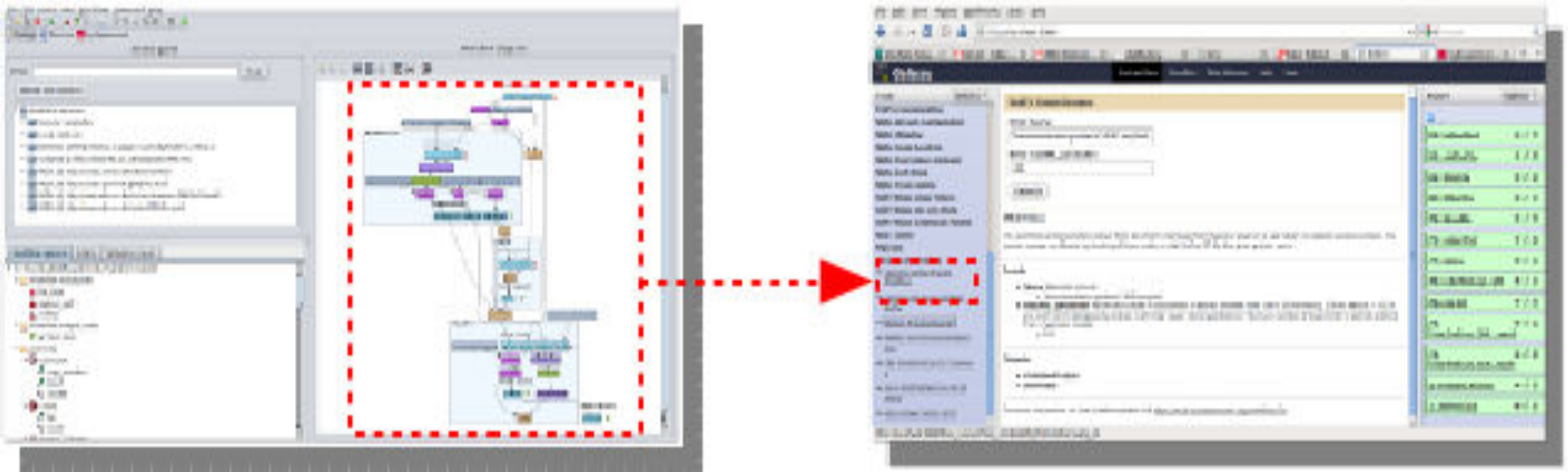
Workflow Interoperability

SHIWA: Sharing Interoperable Workflows

- Development of workflow interoperability technologies
  - Workflow development, testing and validation is a time consuming process and it requires specific expertise
  - Workflows developed for one workflow system not compatible with workflows of other systems



Converting Taverna Workflows into Galaxy Tools



Why?

- Connect the Taverna and Galaxy communities
- Combine the power of Taverna with the simplicity of Galaxy
- Any Taverna service can be made available to Galaxy users

Summary

- Informatics often relies on data integration and large-scale data analysis
- Workflows are a mechanism for linking together resources and analyses
- Automation
- Large data manipulation
- Promote reproducible research
- Easy to find and use successful analysis methods
- Allows scaling-up of computational resources on HPC, cloud and Grid

More Information

Galaxy
  - http://usegalaxy.org

- Taverna
  - http://www.taverna.org.uk

- WS-PGrade
  - http://www.myexperiment.org