Grid computing and computational statistics

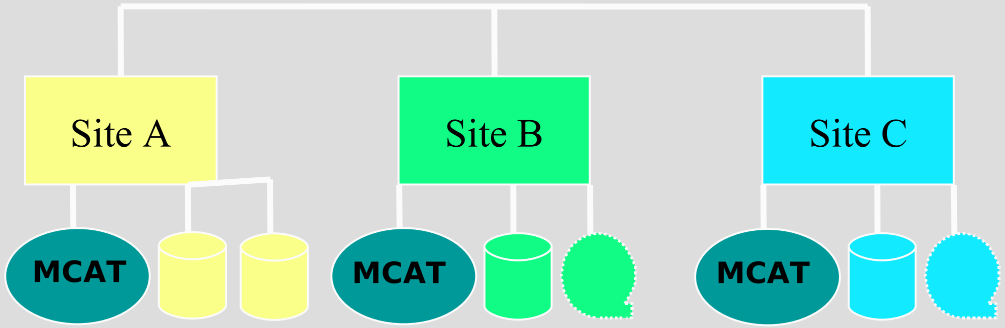
Nigel Sim
VeRG/SIDA

Three focuses

- Data Grid library
  - PGL
- Computational infrastructure
  - Grid Agents
- Data mining:
  - Ensembles
  - Permutation testing

The Data Grid

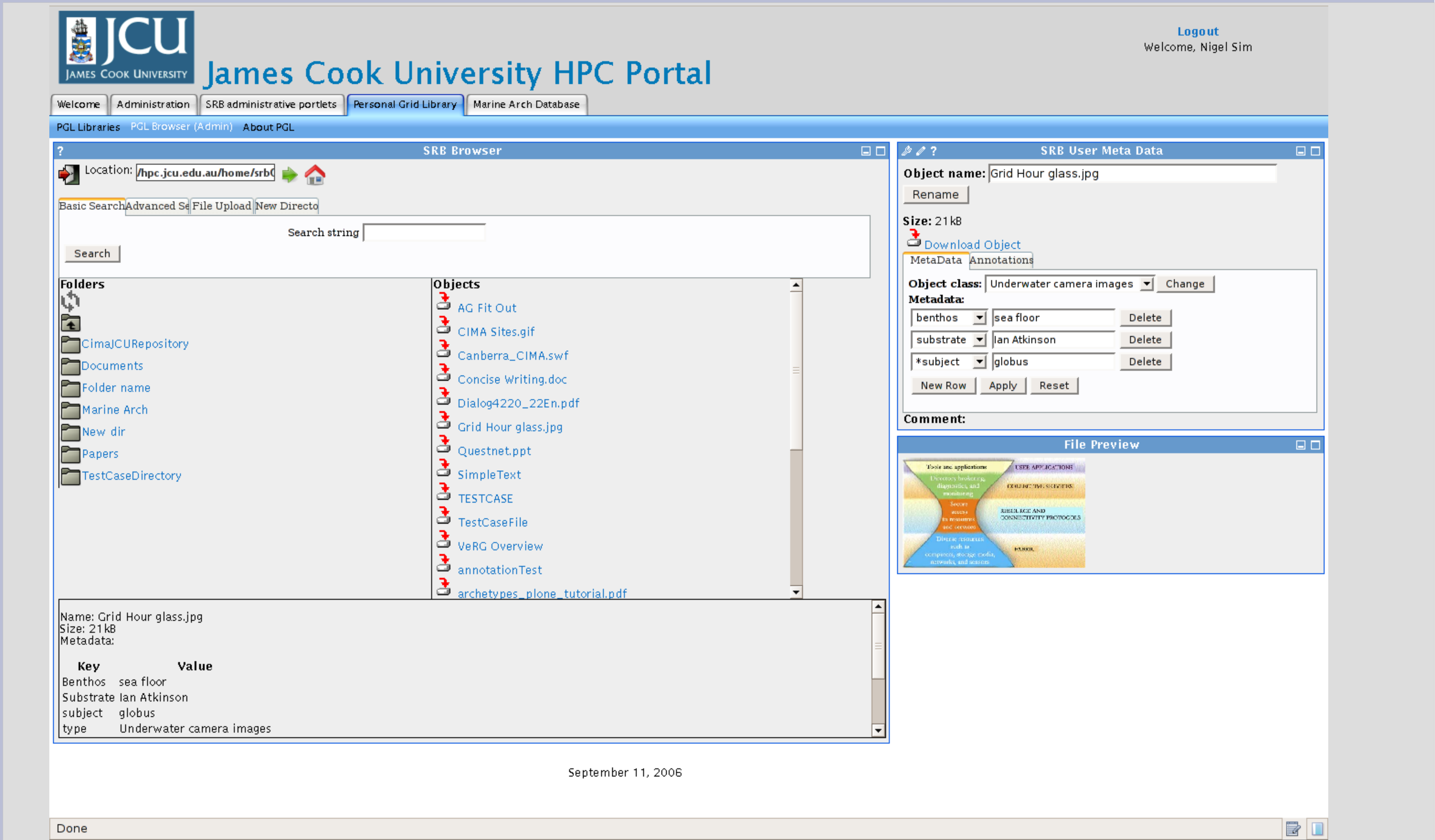
- Federation between sites
- Meta data directly attached to data
- Multiple data locations
- Virtualisation!



Personal Grid Library

- Implementation of a web-based digital library
- Provides an interface to templated metadata manipulation
  - Define object classes (Images, reports, datasets...)
  - Define appropriate metadata tags
  - Define metadata display template
- Exists by placing a metadata template file in a directory – used by interface to render library

Personal Grid Library cont...



JCU JAMES COOK UNIVERSITY James Cook University HPC Portal

Logout
Welcome, Nigel Sim

SRB Browser

Location: /hpc.jcu.edu.au/home/srb(

Search string

Folders

- CimaJCURepository
- Documents
- Folder name
- Marine Arch
- New dir
- Papers
- TestCaseDirectory

Objects

- AG Fit Out
- CIMA Sites.gif
- Canberra_CIMA.swf
- Concise Writing.doc
- Dialog4220_22En.pdf
- Grid Hour glass.jpg
- Questnet.ppt
- SimpleText
- TESTCASE
- TestCaseFile
- VeRG Overview
- annotationTest
- archetypes_plone_tutorial.pdf

Name: Grid Hour glass.jpg
Size: 21 kB
Metadata:

Key	Value
Benthos	sea floor
Substrate	Ian Atkinson
subject	globus
type	Underwater camera images

SRB User Meta Data

Object name: Grid Hour glass.jpg

Size: 21 kB

Download Object

Object class: Underwater camera images ▾ Change

Metadata:

benthos ▾	sea floor	Delete
substrate ▾	Ian Atkinson	Delete
*subject ▾	globus	Delete

Comment:

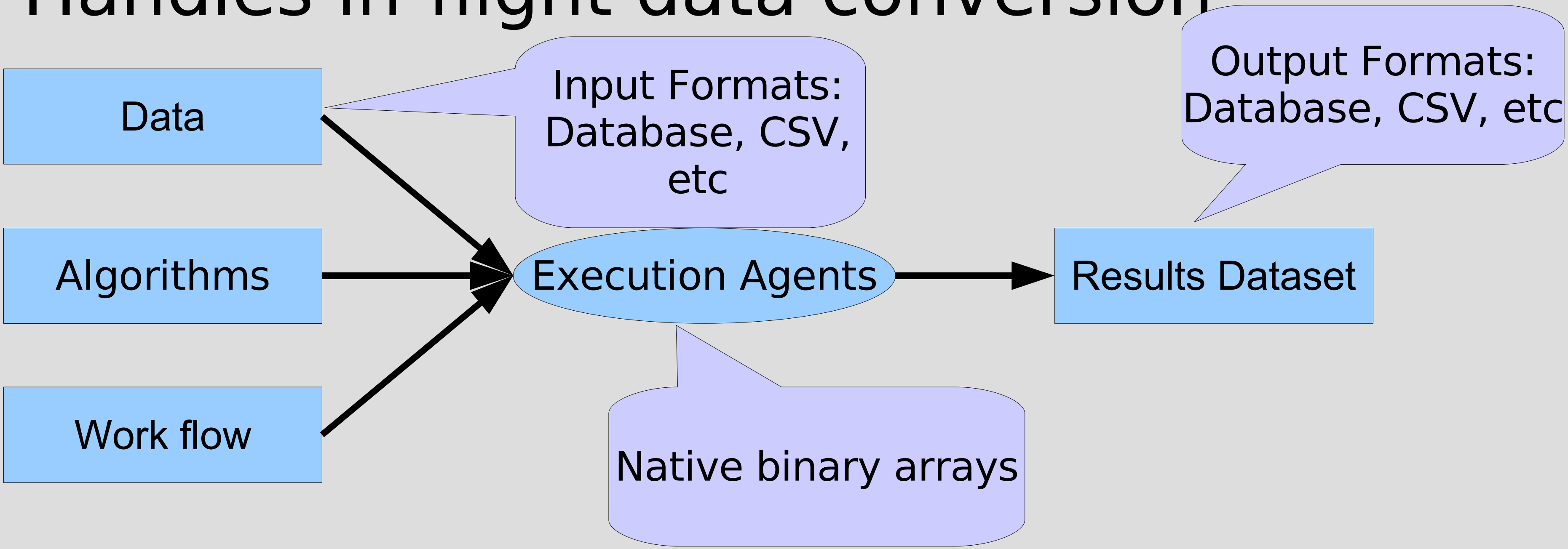
File Preview

September 11, 2006

Done

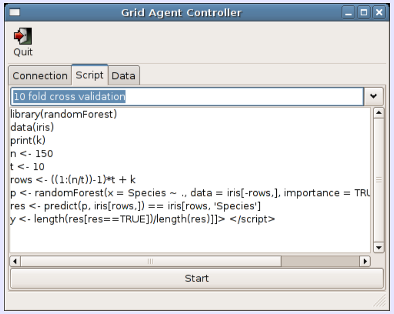
Grid Agents

- Cross platform data flow execution system
- Can utilise R, compiled libraries, legacy apps, etc
- Handles in-flight data conversion



Grid Agents cont...

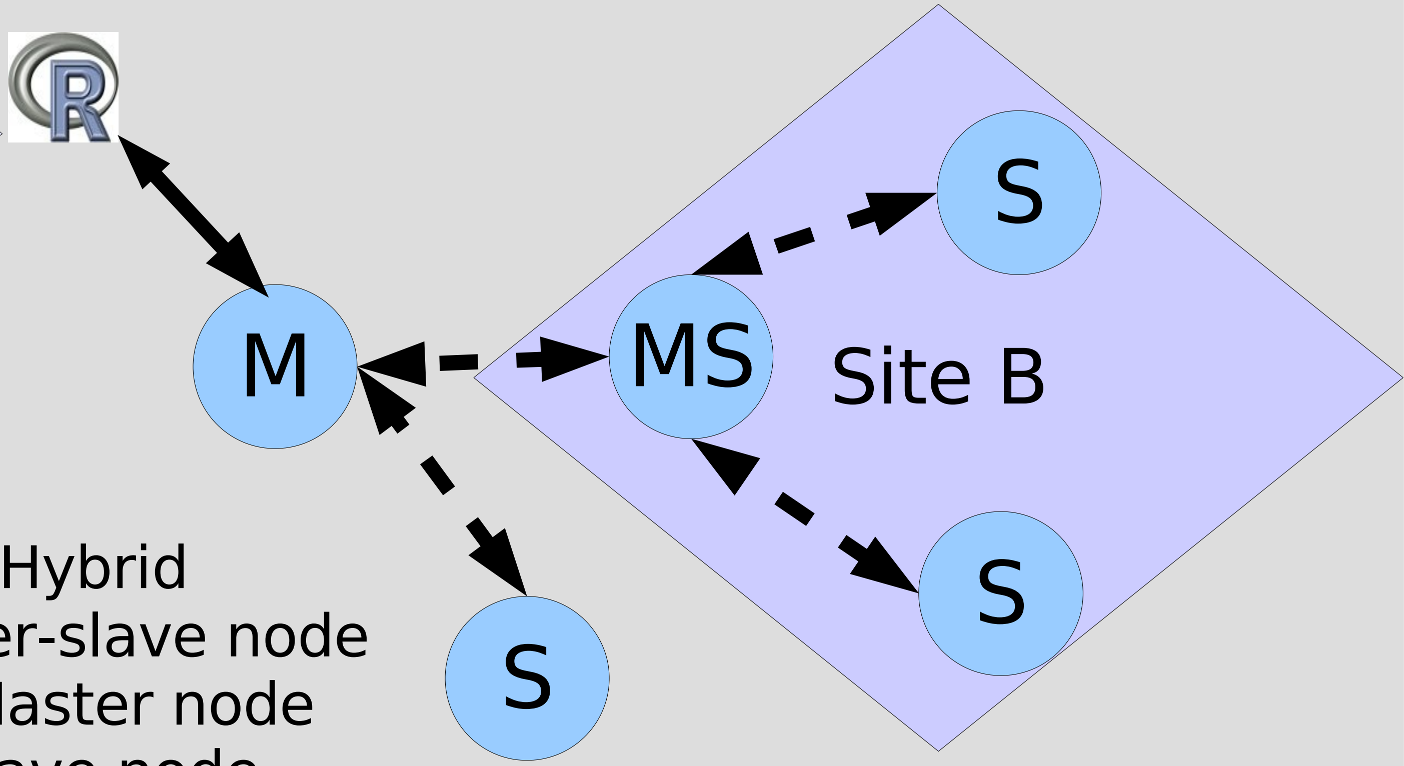
- Scalable work placement
  - Master-slave using workflows
- Utilise multiple CPUs on multiple machines

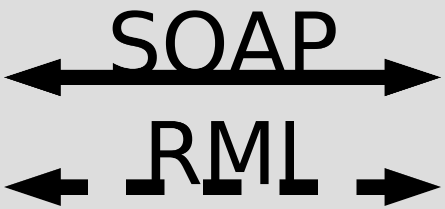


10 fold cross validation

```
library(randomForest)
data(iris)
print(k)
n <- 150
t <- 10
rows <- ((1:(n/t))-1)*t + k
p <- randomForest(x = Species ~ ., data = iris[-rows,], importance = TRU
res <- predict(p, iris[rows,]) == iris[rows, 'Species']
y <- length(res[res==TRUE])/length(res)]]> </script>
```

R monitoring interface

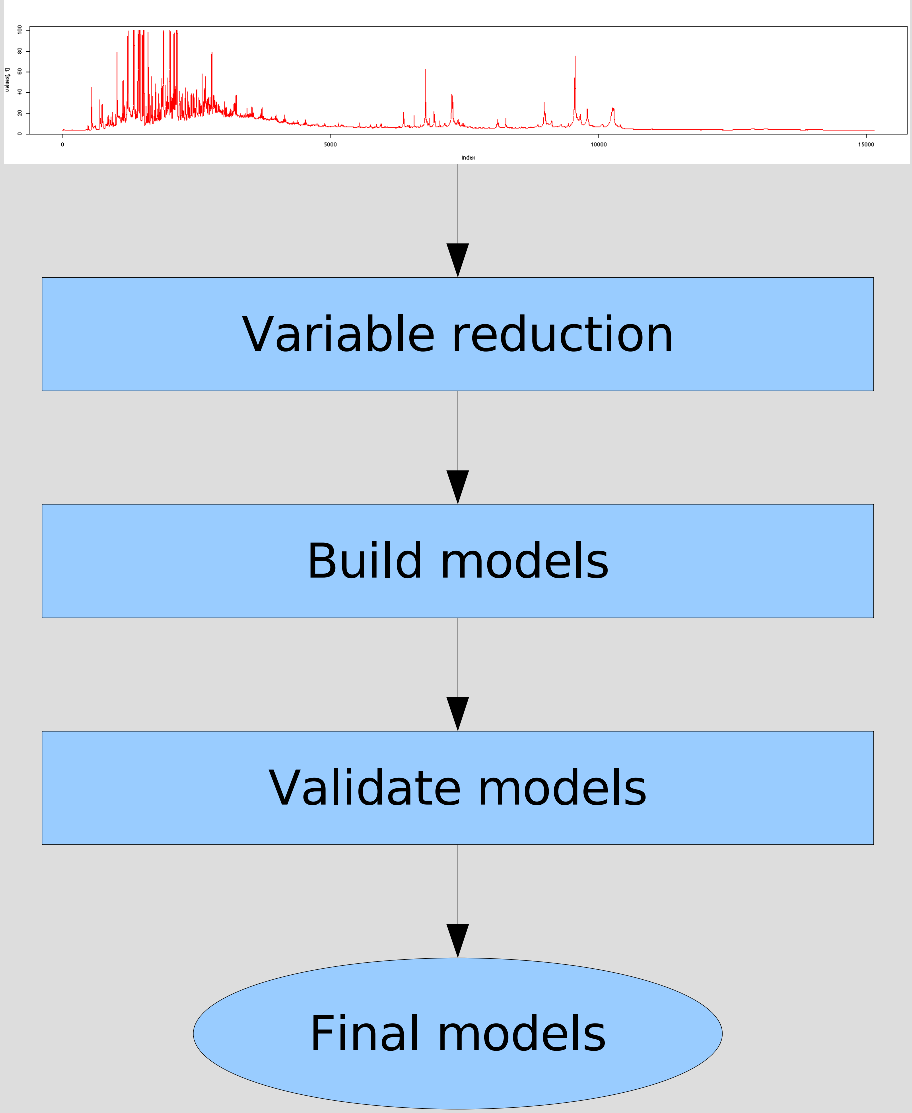




MS – Hybrid master-slave node
M – Master node
S – Slave node

Data mining

- Mainly focusing on spectral data
  - NIR (Quality control)
  - Mass spec (blood proteins)
- Computationally intensive
  - Cross validation
  - Permutation test
  - Optimisation (EA)



Variable reduction

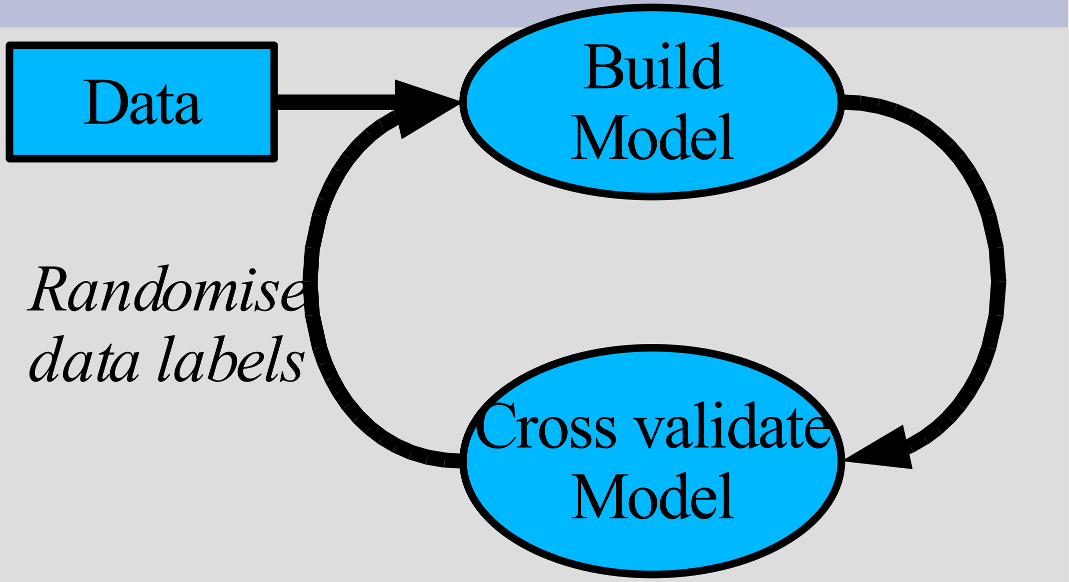
Build models

Validate models

Final models

Permutation tests

- A non-parametric test
- Can be adapted to many situations
  - Model testing
  - Variable testing

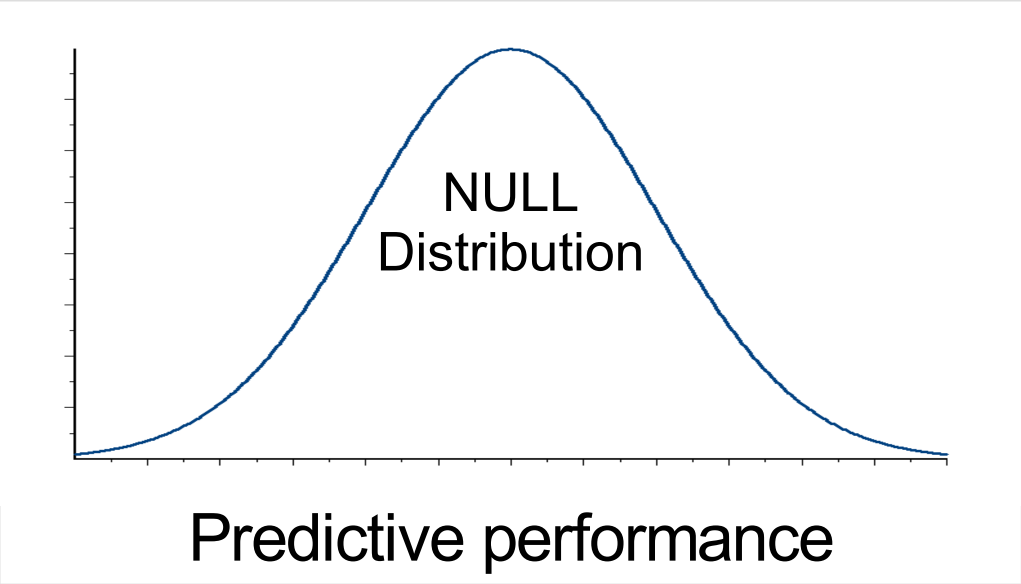


Data

Build Model

Cross validate Model

Randomise data labels



Ensembles

- Linear combination of many models

- Bootstrap using Lasso

- Optimise using evolutionary algorithms

$$F(x)=\sum_{i=1}^{N} w_i f_i(x)$$

$$f(x)=\text{Any model}$$

•Stepwise regression
•Some coefficients may end up zero

	age	sex	bmi	map	tc	ldl	hdl
0	0	0	0	0	0	0	0
1	0	0	60.12	0	0	0	0
2	0	0	361.89	0	0	0	0
3	0	0	434.76	79.24	0	0	0
4	0	0	505.66	191.27	0	0	-114.1
5	0	-74.92	511.35	234.15	0	0	-169.71
		...	...	...	...		
11	-7.01	-237.1	521.08	321.55	-580.44	313.86	0
12	-10.01	-239.82	519.84	324.39	-792.18	476.75	101.04

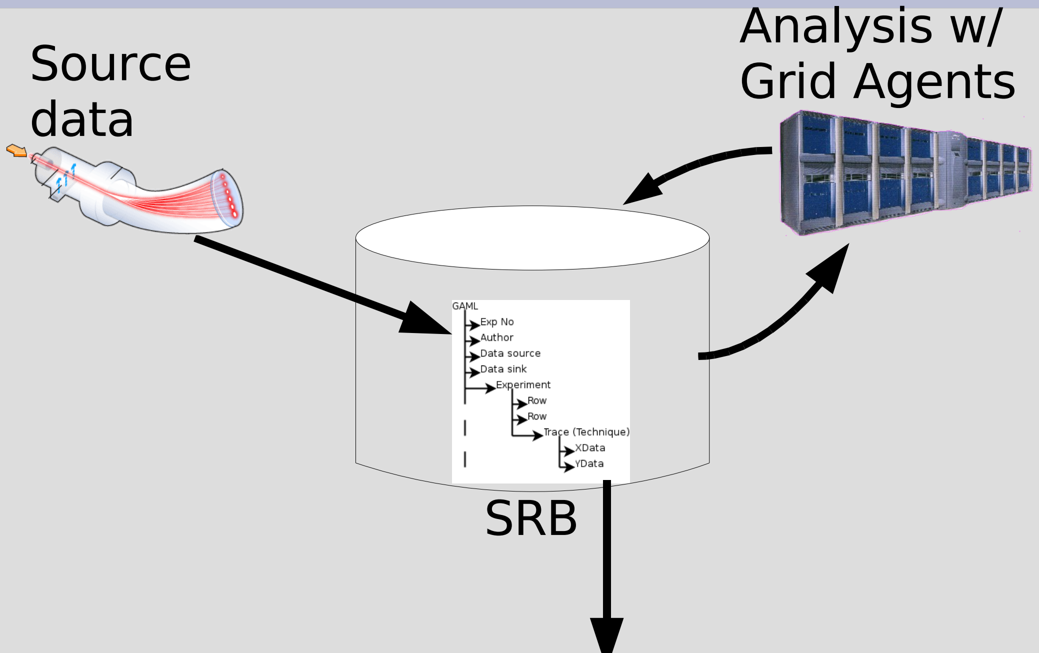
Combined use case

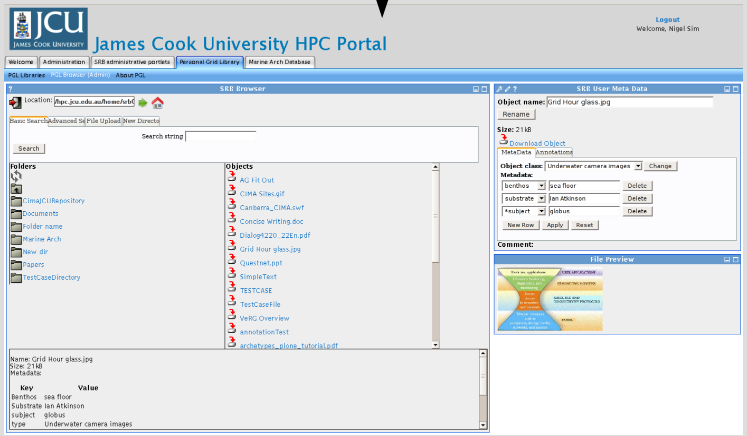
1) Ingest raw data files

2) Start a Grid Agents workflow

3) Results stored as metadata

4) Viewed using PGL





Location: /hpc.jcu.edu.au/home/srb(

Search string

Folders: CimaJCURepository, Documents, Folder name, Marine Arch, New dir, Papers, TestCaseDirectory

Objects: AG Fit Out, CIMA Sites.gif, Canberra_CIMA.swf, Concise Writing.doc, Dialog4220_22En.pdf, Grid Hour glass.jpg, Questnet.ppt, SimpleText, TESTCASE, TestCaseFile, VeRG Overview, annotationTest, archetypes_plone_tutorial.pdf

Name: Grid Hour glass.jpg
Size: 21kB
Metadata:

Key	Value
Benthos	sea floor
Substrate	Ian Atkinson
subject	globus
type	Underwater camera images

Object name: Grid Hour glass.jpg

Size: 21kB

Object class: Underwater camera images

Metadata:

benthos	sea floor
substrate	Ian Atkinson
*subject	globus

Comment:

PGL