Running Large Workflows in the Cloud

Paul Watson
School of Computing Science & Digital Institute
Newcastle University, UK
Paul.Watson@ncl.ac.uk

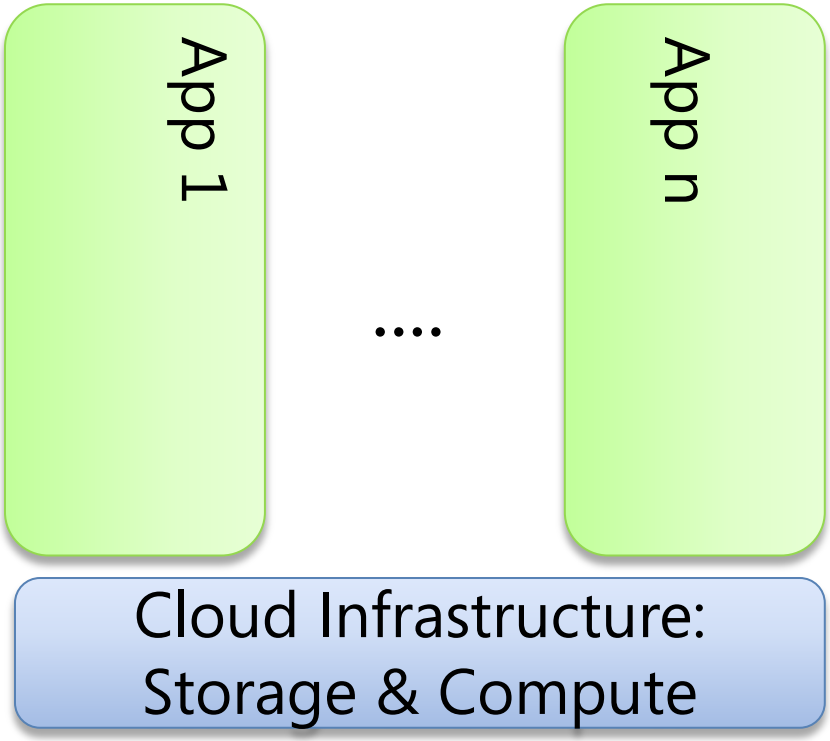
The team: Jacek Cala, Hugo Hiden, Simon Woodman, David Leahy
With thanks to:
-Vlad Sykora, Martyn Taylor, Christophe Poulain, Savas Parastatidis
-Microsoft External Research & the EU (Venus-C)
for their financial support

Venus-C
Virtual multidisciplinary EnviroNments USing Cloud infrastructures

The Promise of Clouds

- Reduced capital expenditure
- Reduced Energy
- Scalability
- “On-demand resources”

The Problem with Clouds

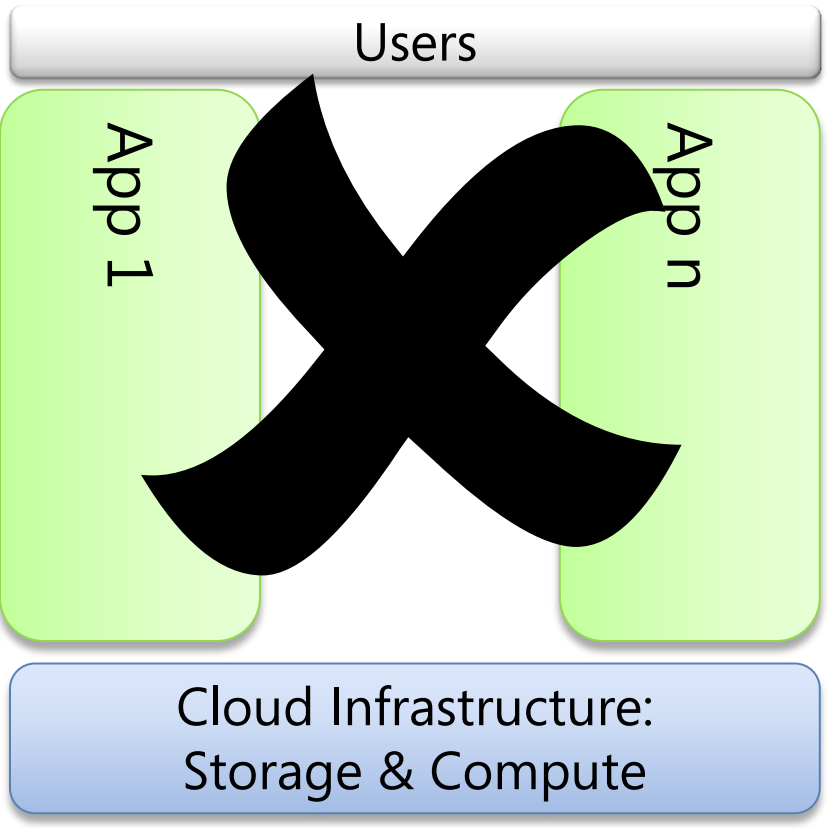


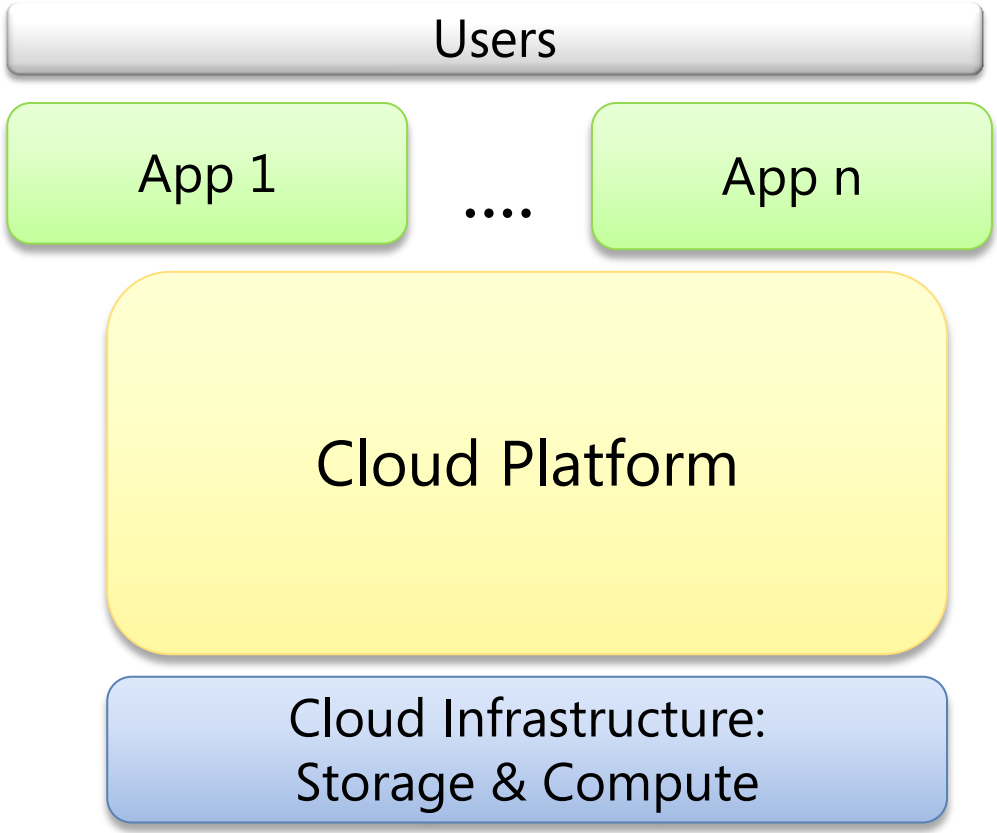
Building scalable, reliable, secure systems on cloud infrastructure is still hard

- *deep IT skills*
- *bespoke*
- *on-going management costs*
- *lock-in*

Realising cloud advantages is beyond most who could benefit

Cloud Options



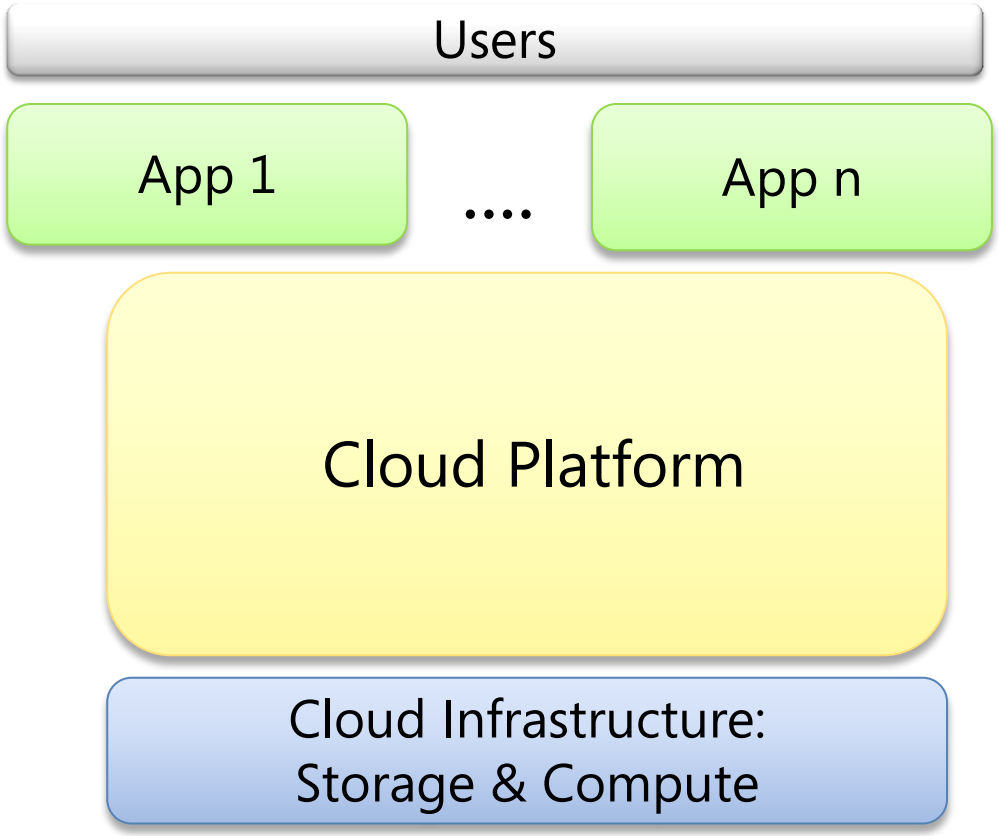


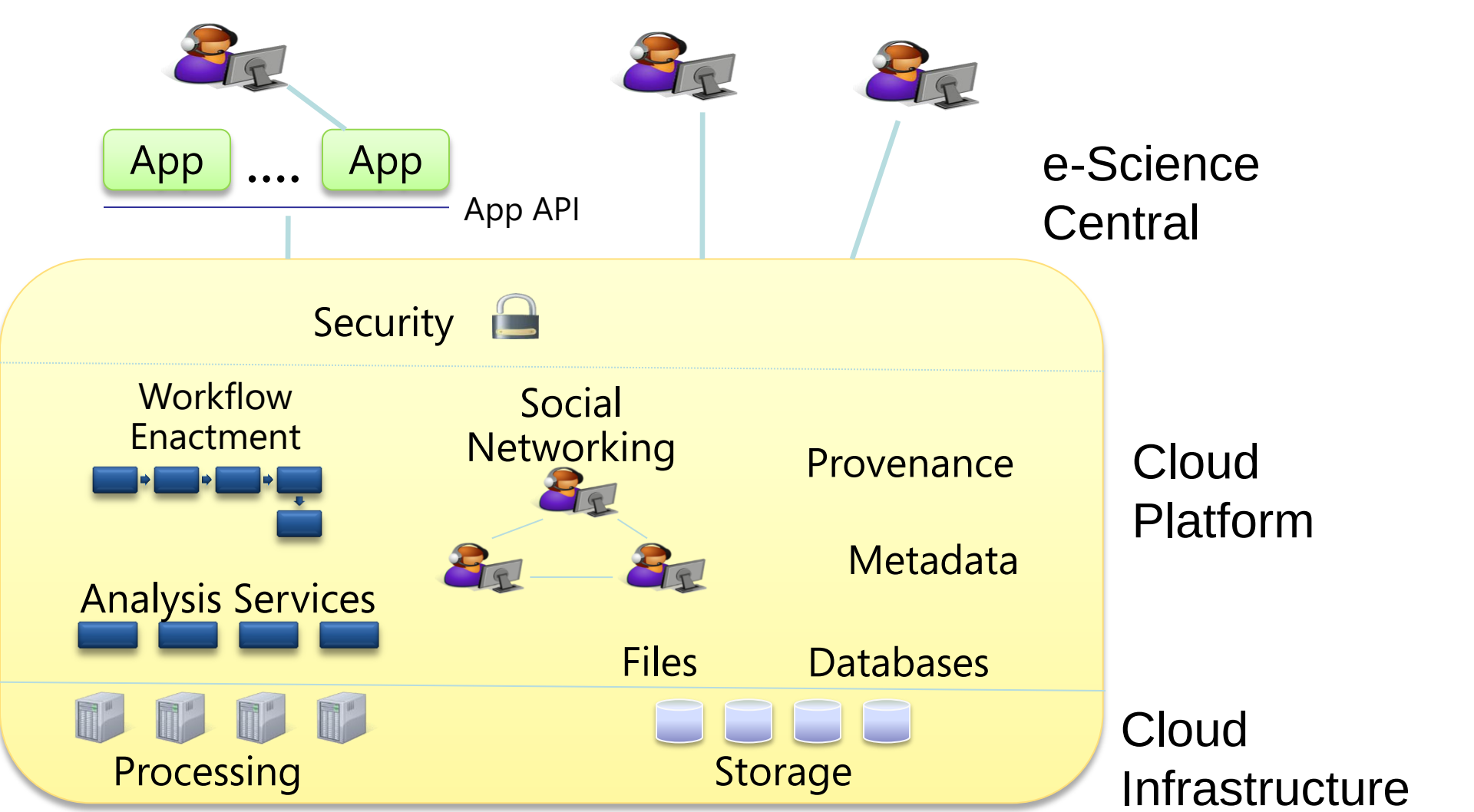
e-Science Central

Science as a Service
for users & programs

Cloud Platform
for developers

Portable: run on
- Azure
- Amazon
- Private Cloud




App
....
App
App API
e-Science
Central
Security
Workflow
Enactment
Social
Networking
Provenance
Metadata
Analysis Services
Files
Databases
Cloud
Platform
Processing
Storage
Cloud
Infrastructure

e-science central

Search...

Home Workflows Data Blogs Notes Apps Hel

Timeline

Show Filters

ed Three in a row... Hanani Abang-Ibrahim published Oct2010 John Colquhoun published Bug fix

quhoun published No teaching, but still busy John Colquhoun published Moving over

-accredited :-) Gavin Wood updated Gavin Wood - SiDE Mobile Developer

eparation Gavin Wood updated Gavin Wood - SiDE Mobile Developer

John Colquhoun published Today's fascinating fact

John Colquhoun published A busy day of teaching and research

Gavin Wood published Week beginning 18th October 2010

Gavin Wood published Week beginning 18th October 2010

Hanani Abang-Ibrahim updated My PhD Research Diary

Hugo Hiden has made eScience 2010.pptx public

John Colquhoun published I am blogging today though :-)

John Colquhoun published I didn't blog yesterday...

d published Week beginning 11th October 2010

d published Week beginning 11th October 2010

published More teaching and demonstrating

DLL

Hugo Hiden has made commons-math-2.1 public

Hugo Hiden has made XPSChart public

Hugo Hiden has made XPSImport public

Hugo Hiden has made XPSReader public

Gavin Wood updated Gavin Wood - SiDE Mobile Developer

Gavin Wood updated Gavin Wood - SiDE Mobile Developer

Gavin Wood updated Gavin Wood - SiDE Mobile Developer

Gavin Wood updated Gavin Wood - SiDE Mobile Developer

Gavin Wood published 25th to 29th October 2010

Gavin Wood published 25th to 29th October 2010

17 Oct 24 Oct 31 Nov 7

eline © SIMILE

Inbox - Mai... Calendar - ... Untitled - ... iGoogle - ... KTN Webin... e-Science ... Humming... e-S

FoodDataPW.csv

Owned By: Paul Watson

Open

Delete

Save

Description:

B I U ABC

Tags

Food [x]
Model [x]
Biscuits [x]

Add Tags Add

Attached Files

Attach File

Versions

Version Comments

Version 2: 08 Mar 2010 11:26

Version 1 : 08 Mar 2010 10:17

Access Control

Document is not Public

Make Public

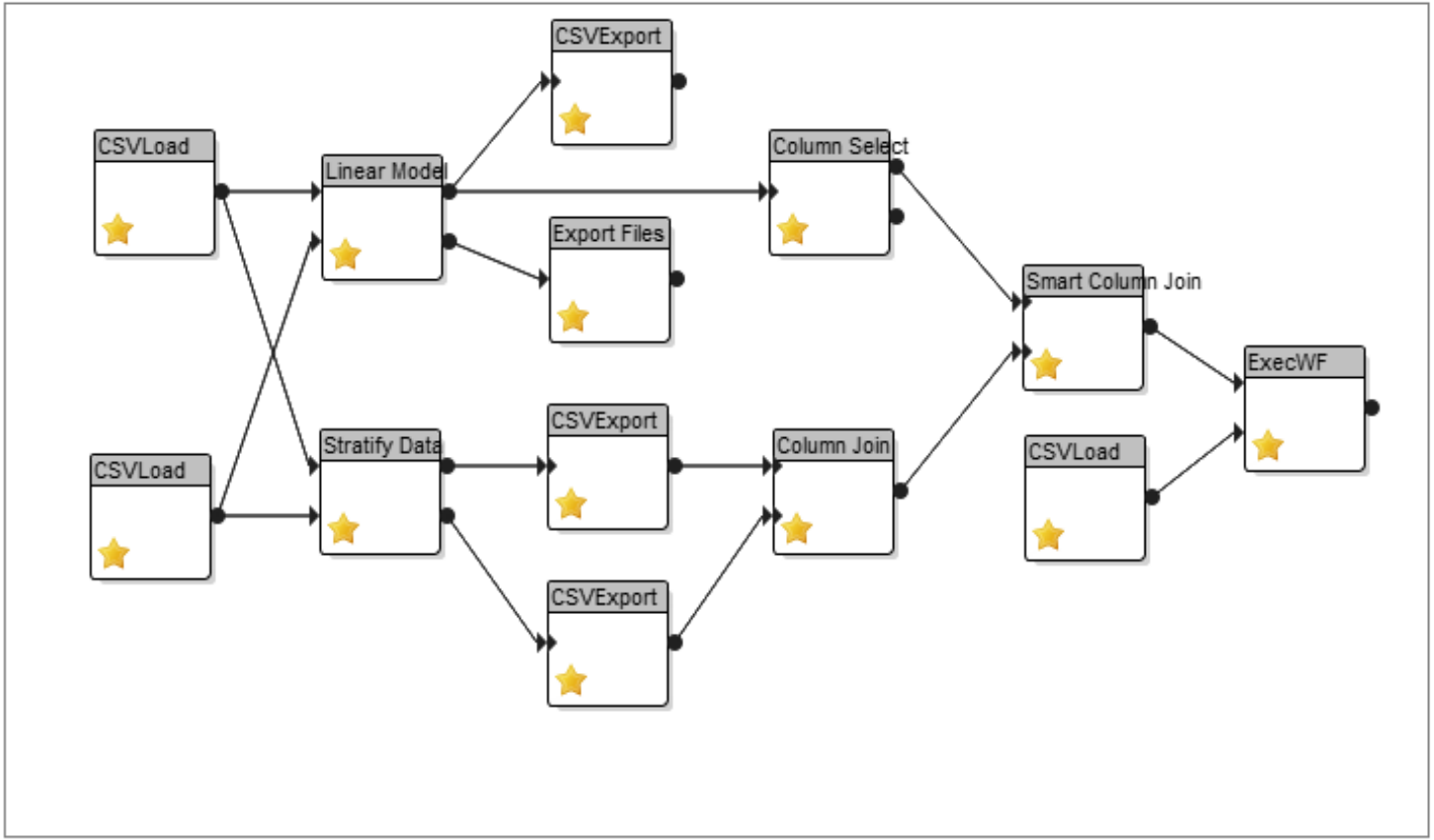
Add Permissions

Name	Access
Stephen Andrews	read
Simon Woodman	read
David Leahy	read
Hugo Hiden	read
Joanna Berry	read,write
Stephen McGough	read

Workflow Editor

1: May 26, 2011 4:04:55 PM | Help | Properties | Run | Delete Selected | Save

- **Discovery Bus**
  - RPart Model
  - Neural Net Model
  - xValidateNNM
  - xValidatePLSM
  - Linear Model
  - xValidateRPartM
  - xValidateLM
  - StratifyData
  - PLS Model
  - Filter Features
- **File Management**
- **Import**
- **Workflow**
- **Windows Azure**
- **Manipulation**
- **Numerical**
- **Chemistry**
- **Export**



Selected block:

e-Science Central. Science. On Demand.. - Windows Internet Explorer

http://esc.selfip.com/esc/secure/dataproperties.jsp?id=ff8081811db9952c011de37b29f20044

nokia mail for exchange e71

Favorites | iGoogle | e-Science Central. Scie... | Nokia Europe - Device soft...

Page | Safety | Tools

Profile | People | Groups | Settings | Inbox

Logged in as Paul Watson **Logout**

Search...

e-science central

Home | Workflows | Data | Blogs | Notes | Apps | Help

trend.jpg

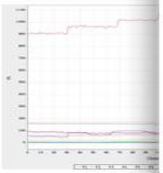
Owned By: Paul Wa

Open

Delete

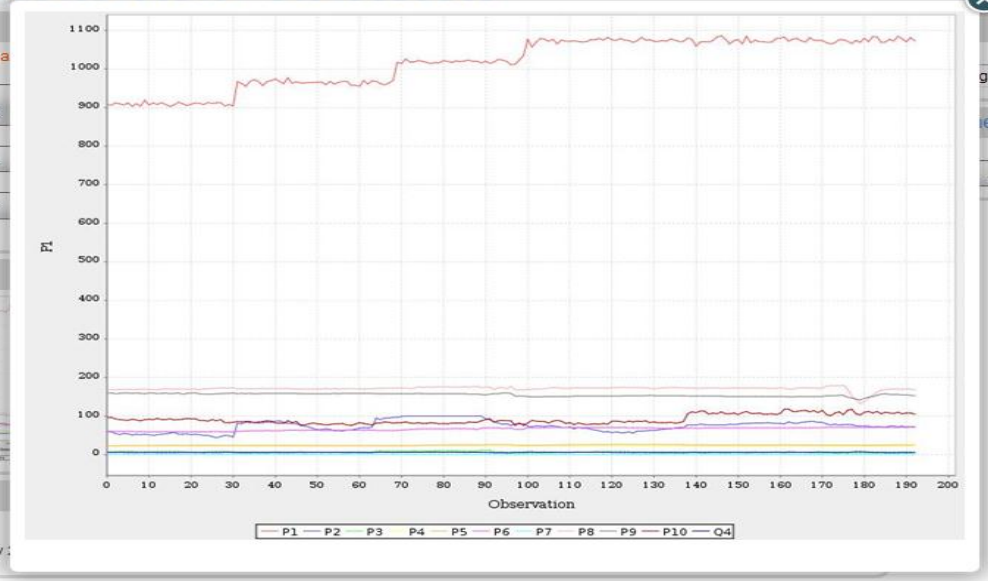
Save

Preview



Versions

Version 0: 28 Nov



gs Add

ed Files

Attach File

History

28 November 2008: Created by Paul Watson (1)

Access Control

Document is not Public

Make Public

Add Permissions

Name	Access

Meta Data

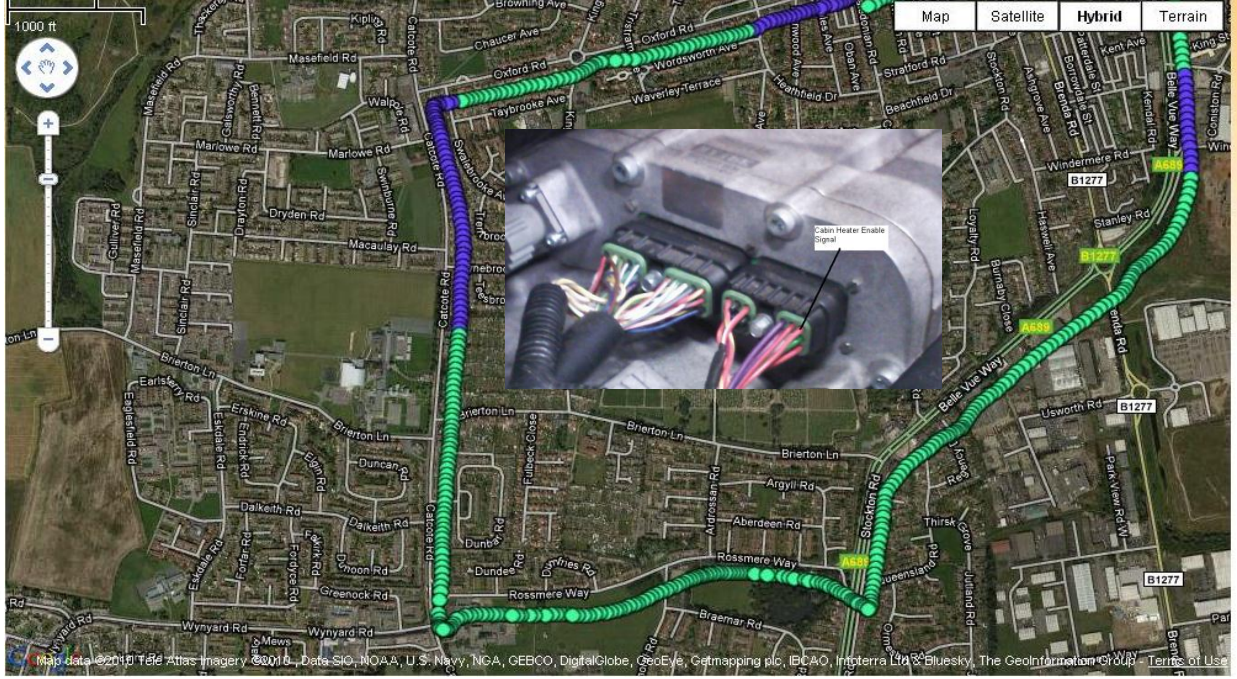
Done

Internet | Protected Mode: On

100%

Used for variety of research:

- activity recognition
- spectroscopy
- driving analysis
- chemistry…

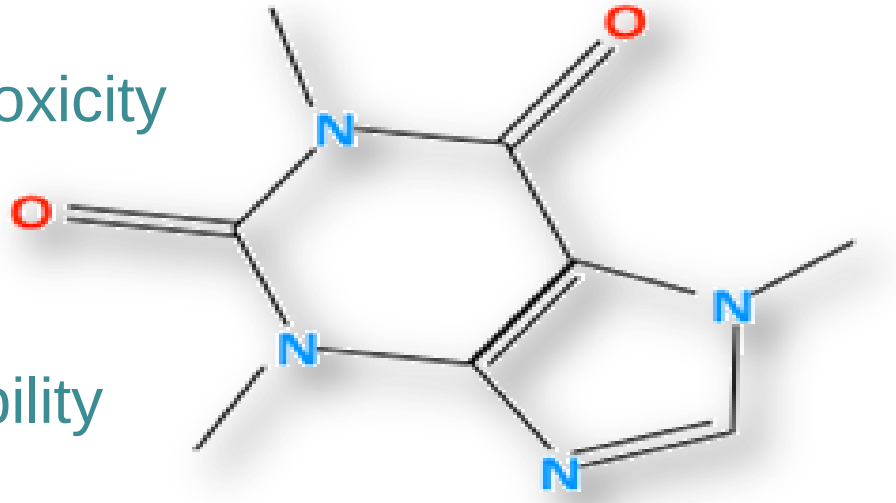


Chemists want to know:

Q1. What are the properties of this molecule?

Toxicity

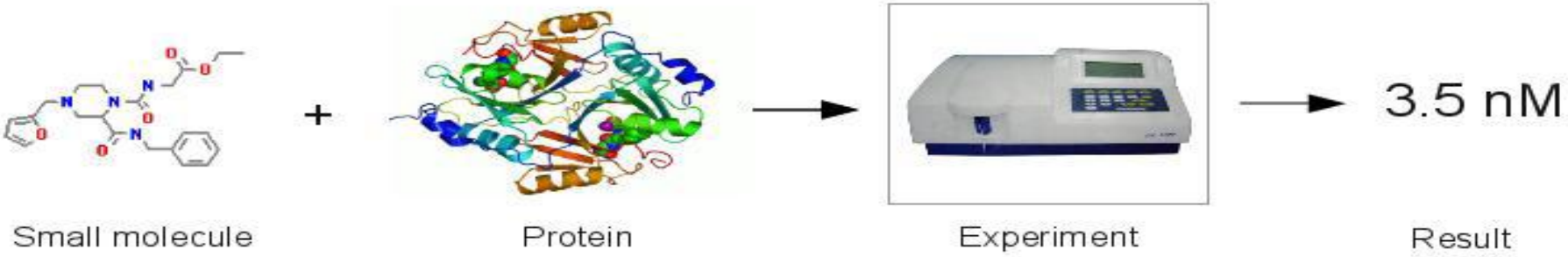
Biological Activity



Solubility

Q2. What molecule would have aqueous solubility of 0.1 µg/mL*?*

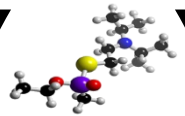
Answering the Question by performing experiments



..... time consuming, expensive, ethical Issues

An alternative to experimentation: QSAR

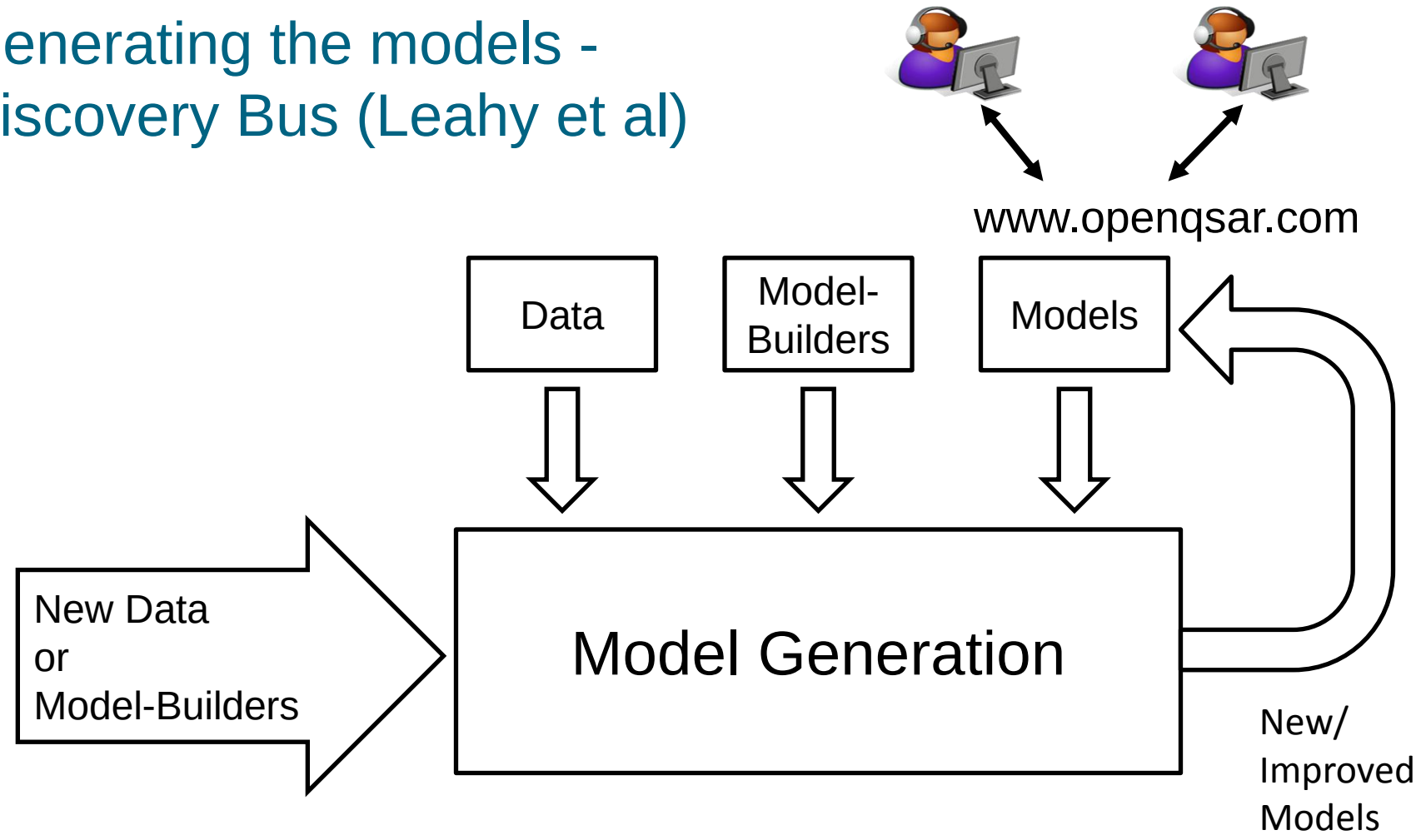
Quantitative Structure Activity Relationship
- predict properties based on similar molecules

$$Activity \approx f(\text{  })$$

quantifiable structural attributes, e.g.

#atoms
logp
shape
.....

Generating the models - Discovery Bus (Leahy et al)



Increasing amounts of data for model building...

CHEMBL : data on 622,824 compounds,
collected from 33,956 publications

WOMBAT : data on 251,560 structures,
for over 1,966 targets

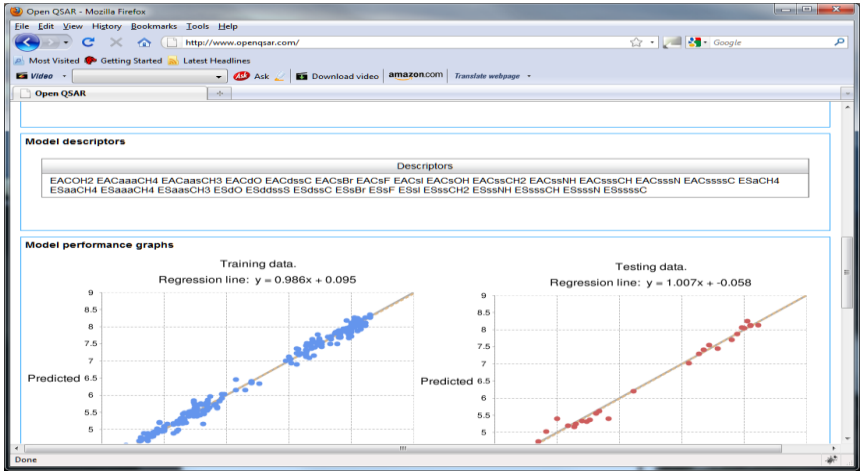
WOMBAT-PK: data on 1230 compounds,
for over 13,000 clinical measurements

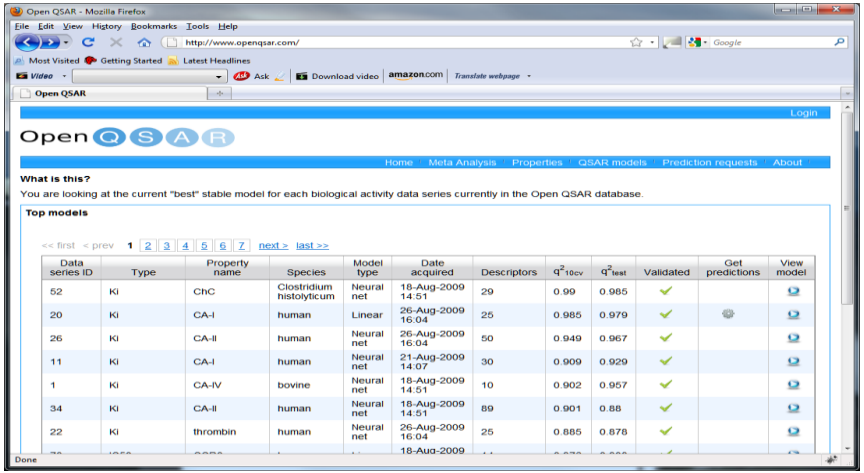
✓ More models
✓ Better models

✗ est. 5 years to process new datasets on existing server

JUNIOR Project Aim

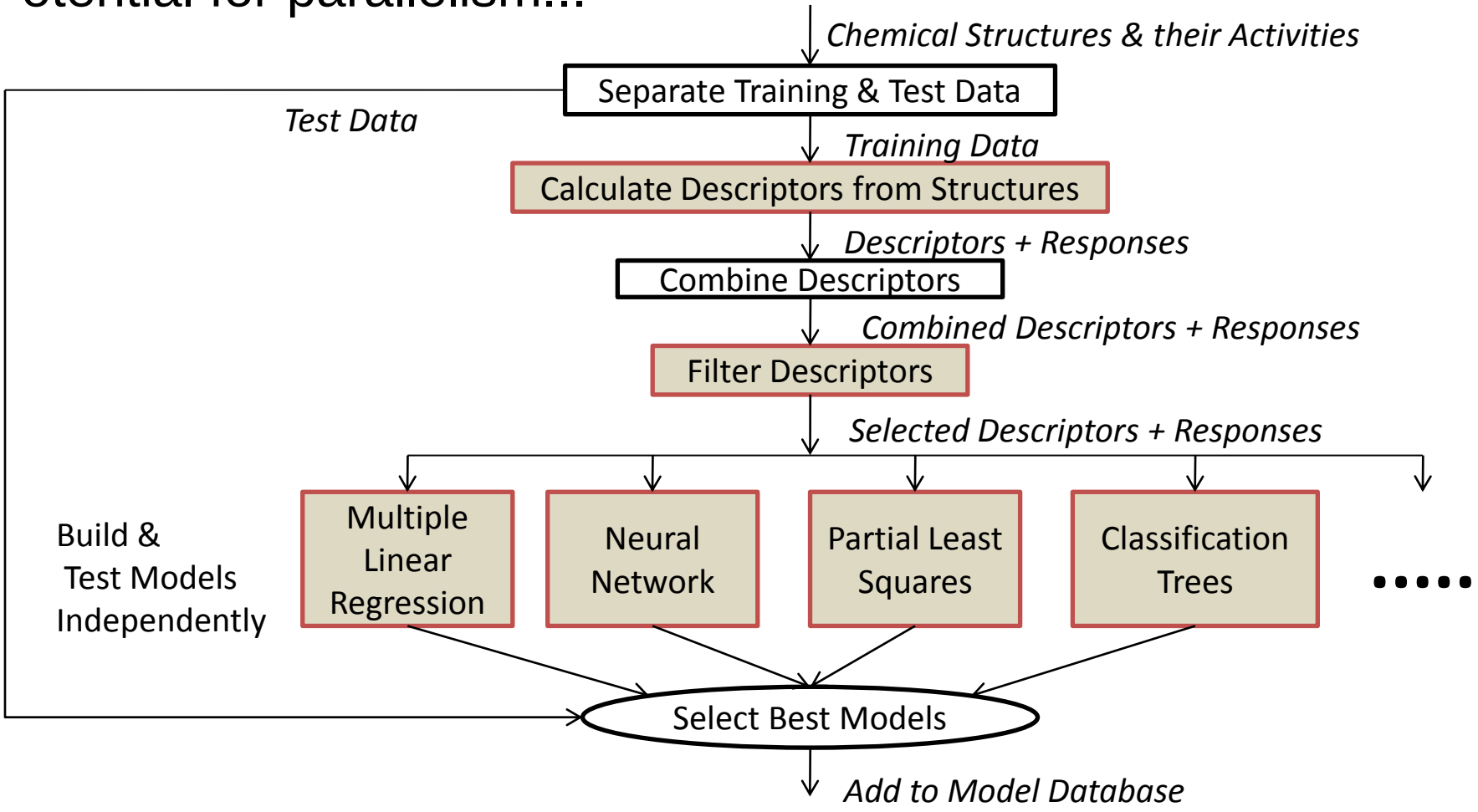
Use Azure to generate models in weeks not years





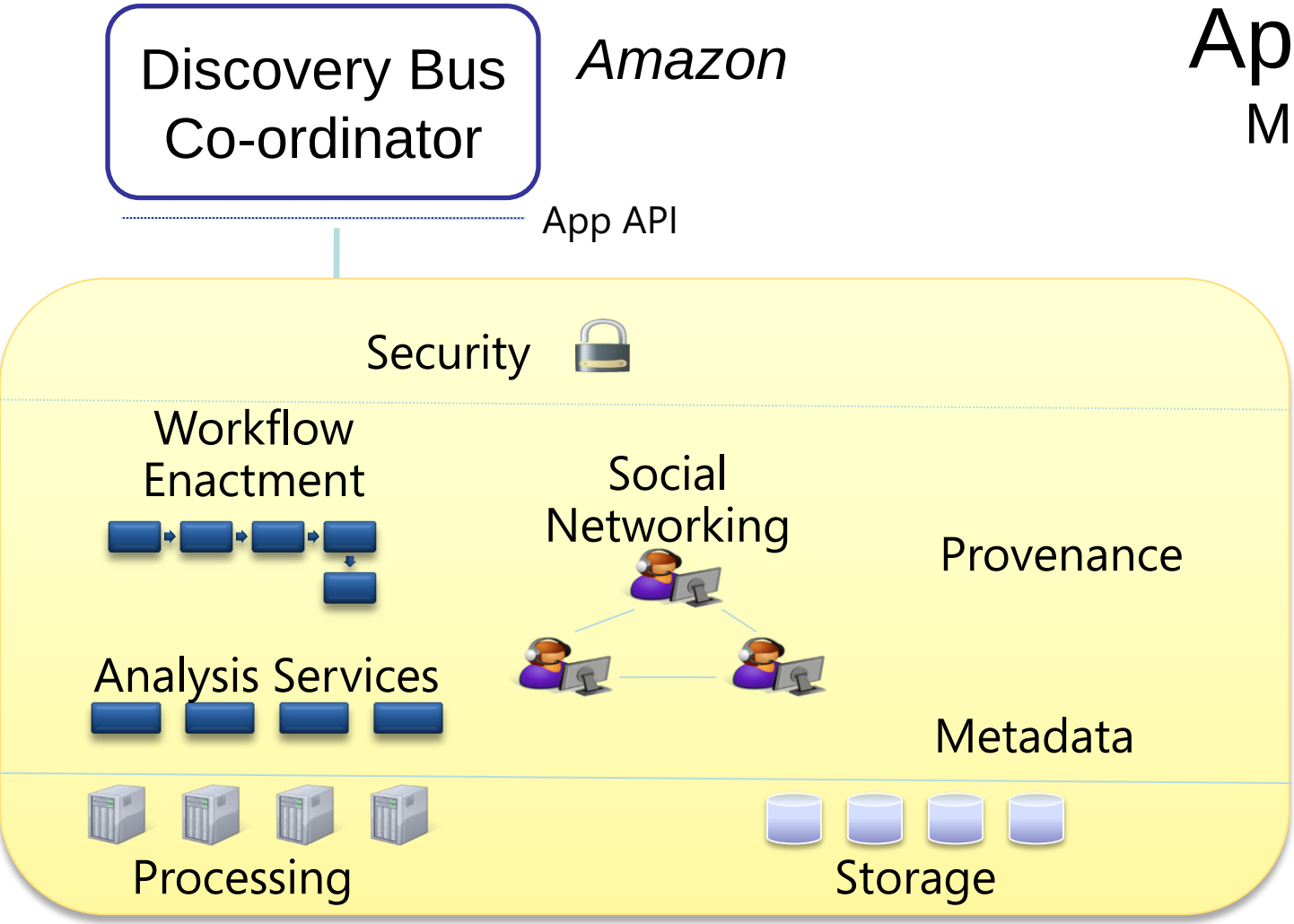
Data series ID	Type	Property name	Species	Model type	Date acquired	Descriptors	q^2_{10cv}	q^2_{test}	Validated	Get predictions	View model
52	Ki	ChC	Clostridium histolyticum	Neural net	18-Aug-2009 14:51	29	0.99	0.985	✓		
20	Ki	CA-I	human	Linear	26-Aug-2009 16:04	25	0.985	0.979	✓		
26	Ki	CA-II	human	Neural net	26-Aug-2009 16:04	50	0.949	0.967	✓		
11	Ki	CA-I	human	Neural net	21-Aug-2009 14:07	30	0.909	0.929	✓		
1	Ki	CA-IV	bovine	Neural net	18-Aug-2009 14:51	10	0.902	0.957	✓		
34	Ki	CA-II	human	Neural net	18-Aug-2009 14:51	89	0.901	0.88	✓		
22	Ki	thrombin	human	Neural net	26-Aug-2009 16:04	25	0.885	0.878	✓		

Potential for parallelism...

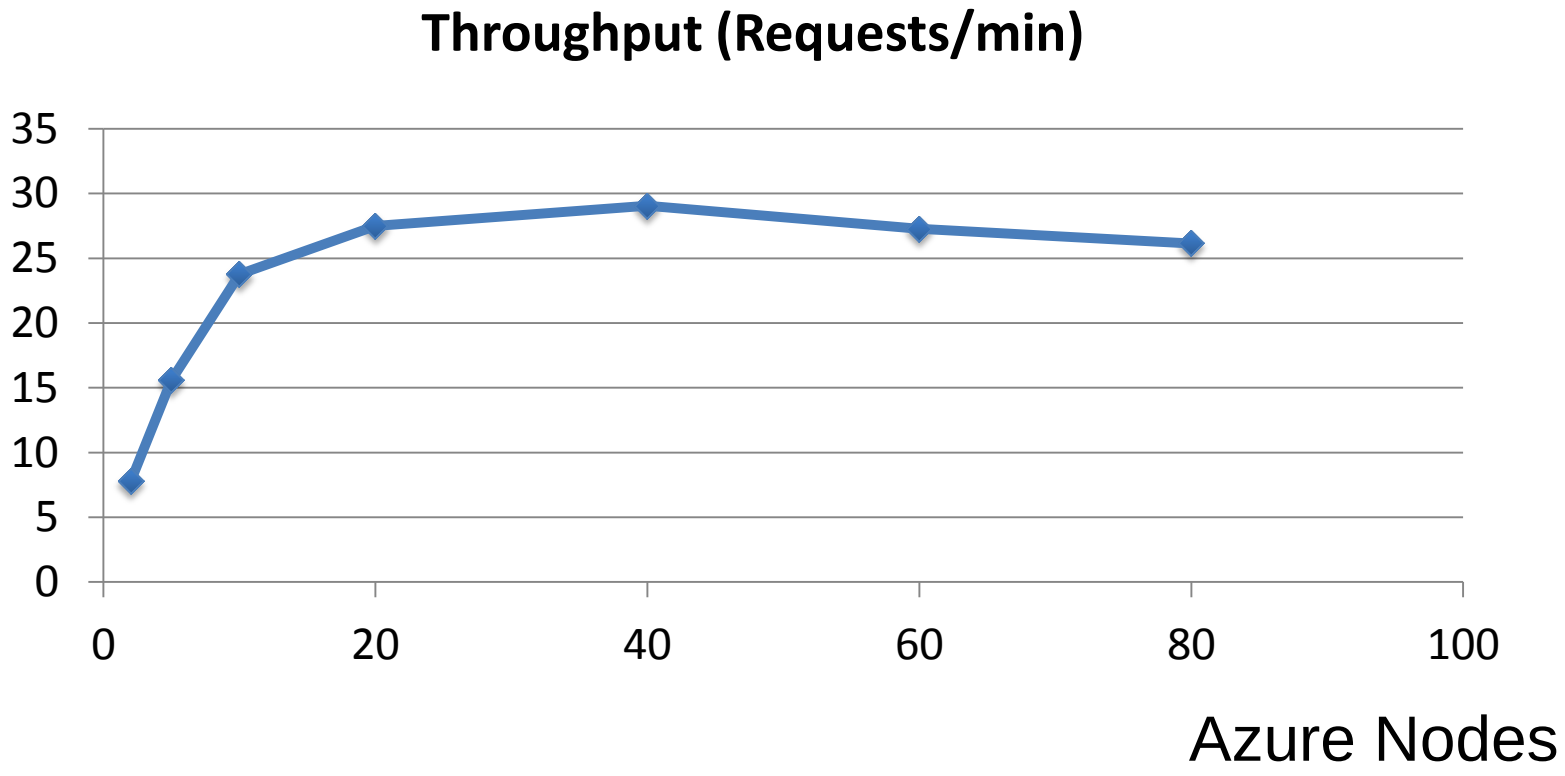


Approach #1

Minimal change to Discovery Bus

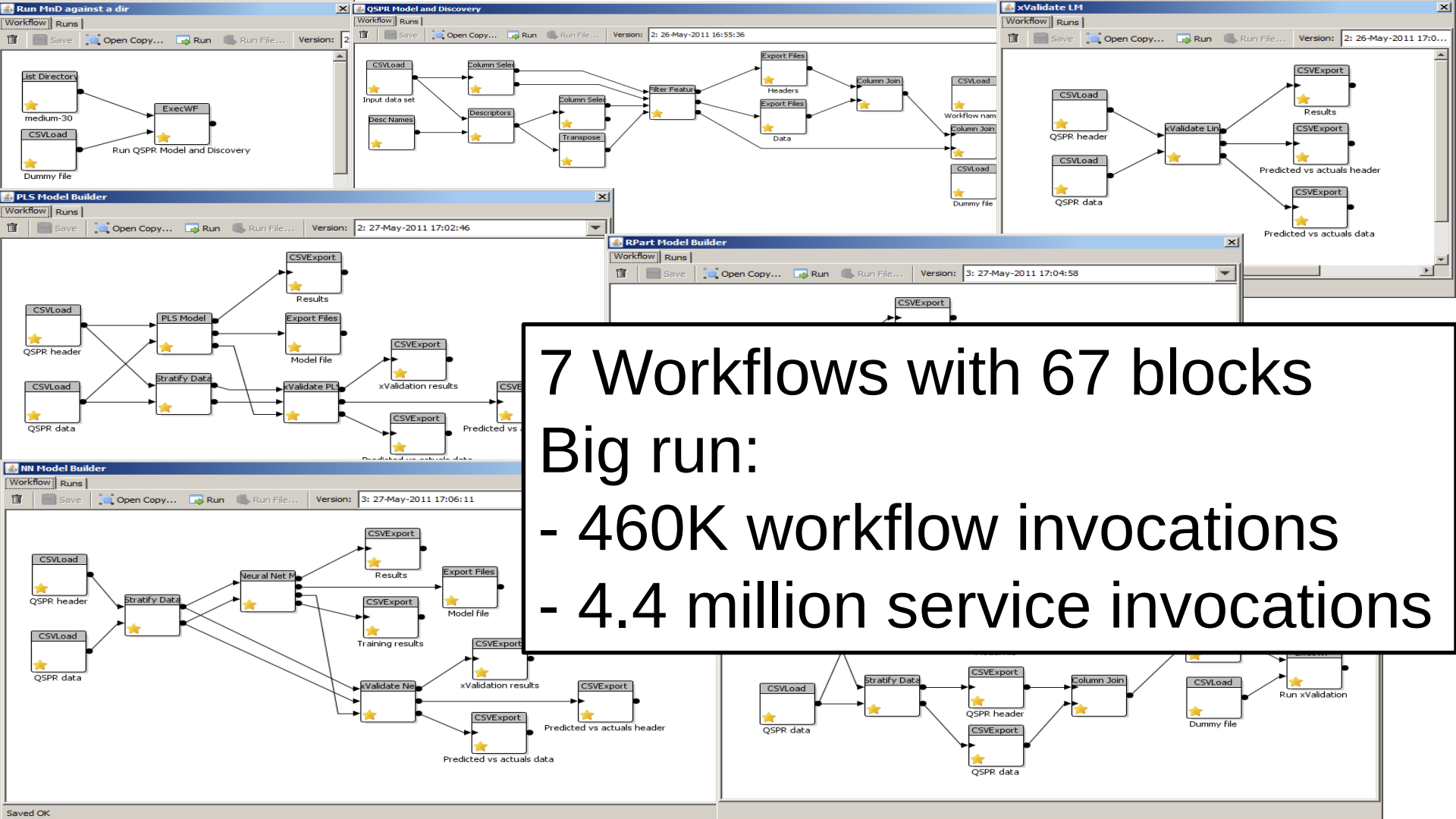


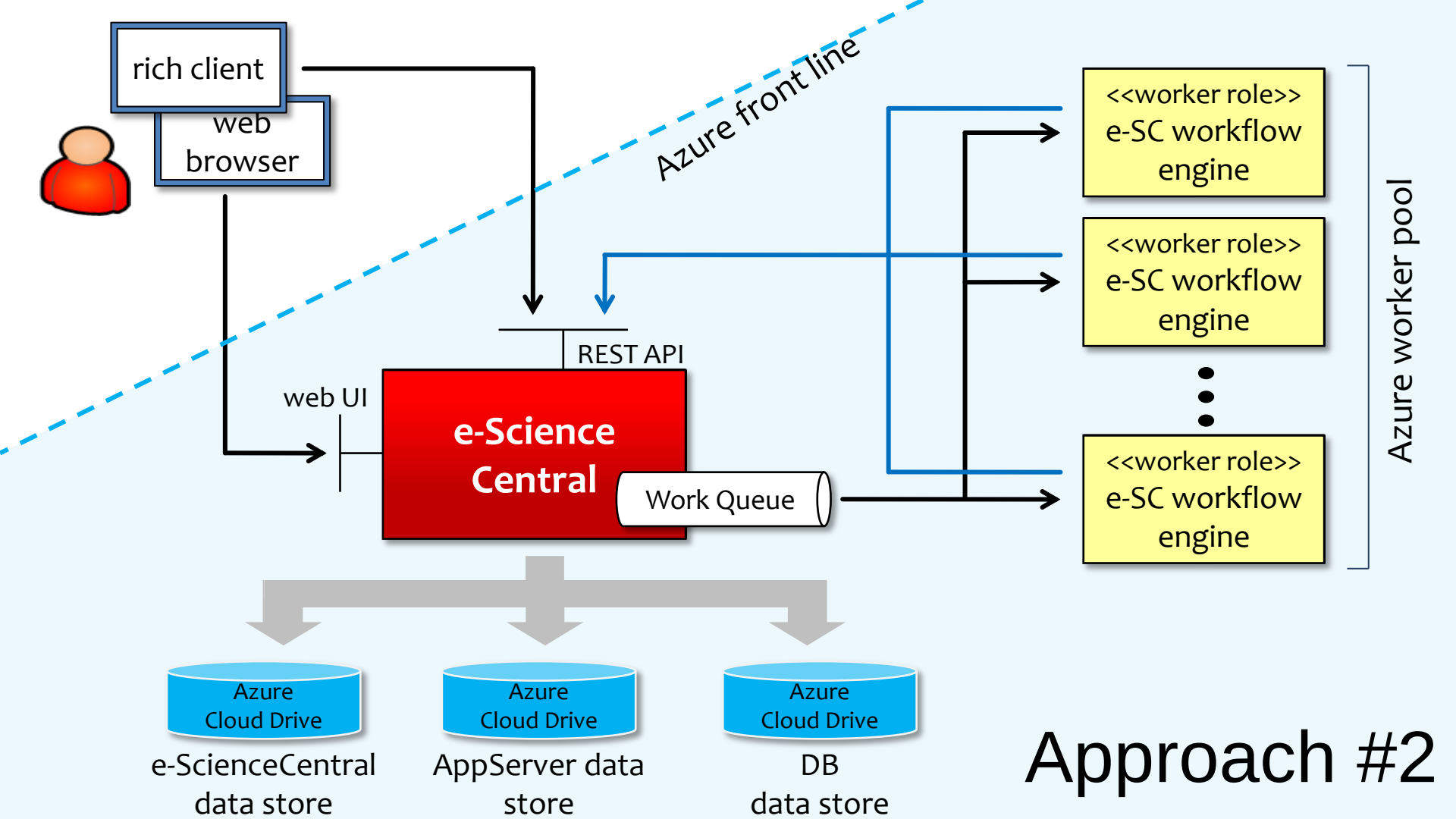
Succeeded in reducing time from (est.) 5 years to 3 weeks, but…



Approach #2

- run entirely within Azure
  - through e-Science Central on Azure

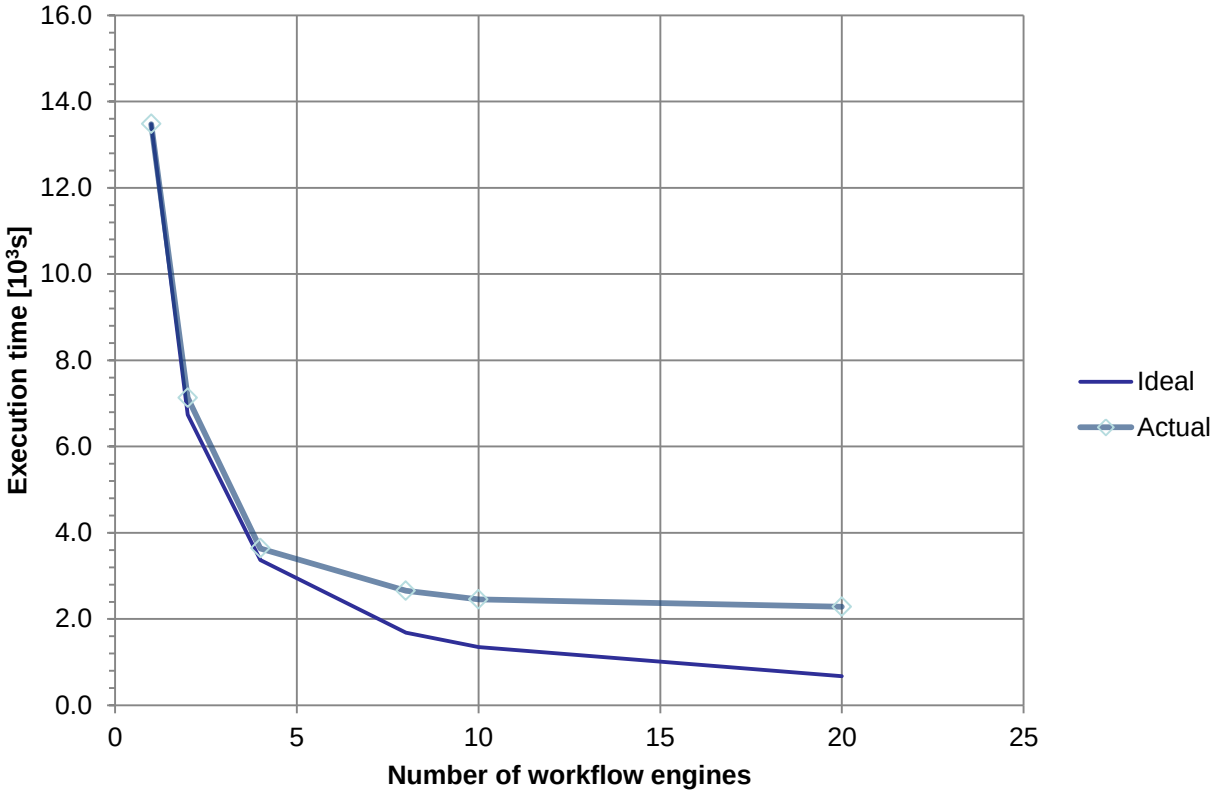

Run MnD against a dir
QSPR Model and Discovery
xValidate LM
PLS Model Builder
RPart Model Builder
NN Model Builder
7 Workflows with 67 blocks
Big run:
- 460K workflow invocations
- 4.4 million service invocations


rich client
web browser
Azure front line
<<worker role>>
e-SC workflow engine
<<worker role>>
e-SC workflow engine
<<worker role>>
e-SC workflow engine
Azure worker pool
REST API
web UI
e-Science Central
Work Queue
Azure Cloud Drive
Azure Cloud Drive
Azure Cloud Drive
e-ScienceCentral data store
AppServer data store
DB data store
Approach #2

Early Results

- Reduces time to 9 days
  - 5yrs $\rightarrow$ 22 days $\rightarrow$ 9 days
  - but room for improvement….

Scalability



Summary

- *Discovery Bus* exemplifies a good Cloud pattern
  - large, variable, bursty requirements
- e-Science Central is a scalable, secure, portable cloud platform for Azure (and Amazon, and Private Clouds)
- next steps
  - optimize large workflow scheduling
  - automatically adapt #workers