

Grid technologies for Social Science: the Seamless Access to Multiple Datasets (SAMD) project

Authors: Celia Russell, Keith Cole, M. A.S. Jones, S.M. Pickles, M. Riding, K. Roy, M. Sensier

NCeSS All Hands Meeting
5-6 July 2004, Hulme Hall, Manchester



Seamless Access to Multiple Datasets

- A project to demonstrate the benefits of applying e-Science grid technologies to an ordinary social science query
- Use a grid approach to solve a genuine problem from the UK academic social science community - a multivariate analysis using a complex mathematical algorithm
- Based on a major social science databank, the UK Office for National Statistics Time Series Data, hosted at MIMAS

The problem

- Published as **Sensier, M., Osborn D.R. and Öcal N.** (2002) 'Asymmetric Interest Rate Effects for the UK Real Economy' , Oxford Bulletin of Economics and Statistics, Volume 64, September 2002, n°4
- The research query looks at the effect interest rate changes had on Gross Domestic Product in the UK over the period 1960 – 2000

The Model

$$y_t = \phi_{0,t} + \sum_{i=1}^k \phi_{i,t} y_{t-i} + \sum_{i=1}^k \delta_{i,t} z_{t-i}$$
$$F(r_{t,t}) [\phi_{0,t} + \sum_{i=1}^k \phi_{i,t} y_{t-i} + \sum_{i=1}^k \delta_{i,t} z_{t-i}] = \varepsilon$$

$$F(r_{t,t}) = \frac{1}{1 + \exp(-\gamma(r_{t,t} - \mu)) \sigma(r_t)}$$

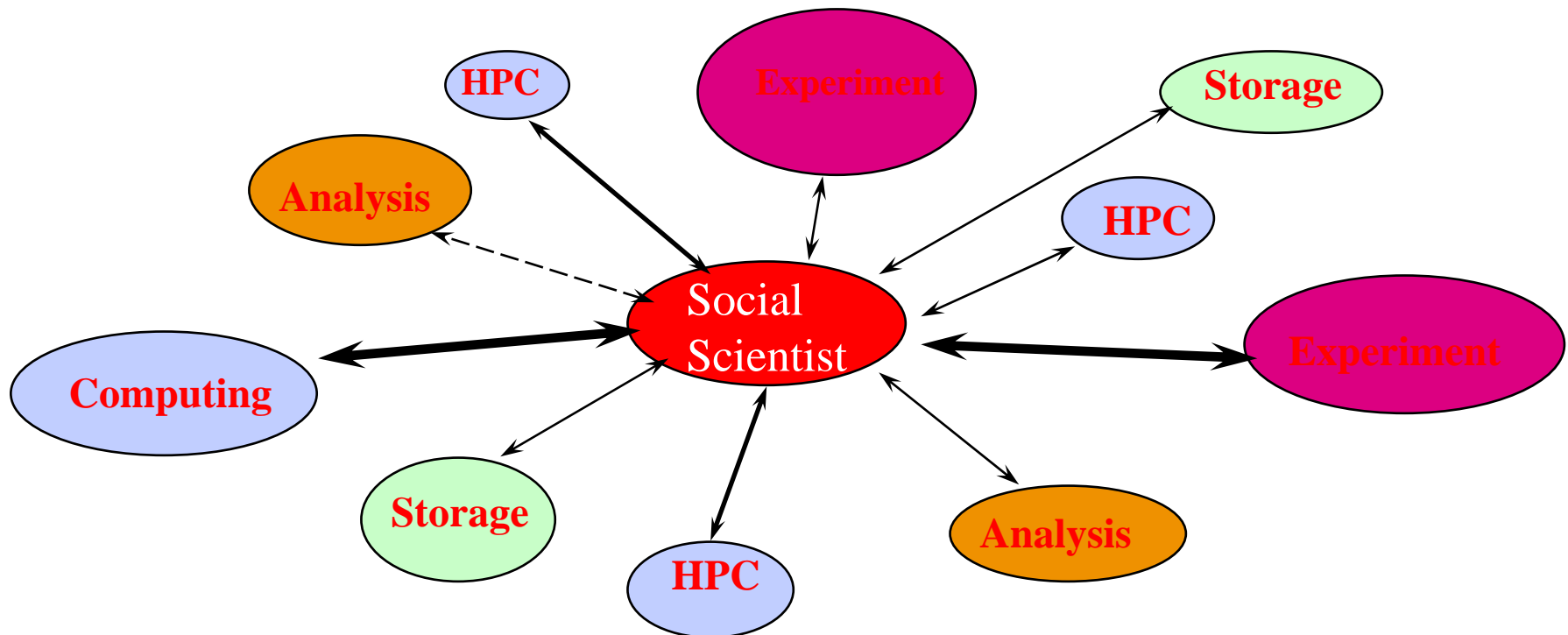
Where y is the quarterly change in GDP and z is the quarterly change in interest rates

Before SAMD

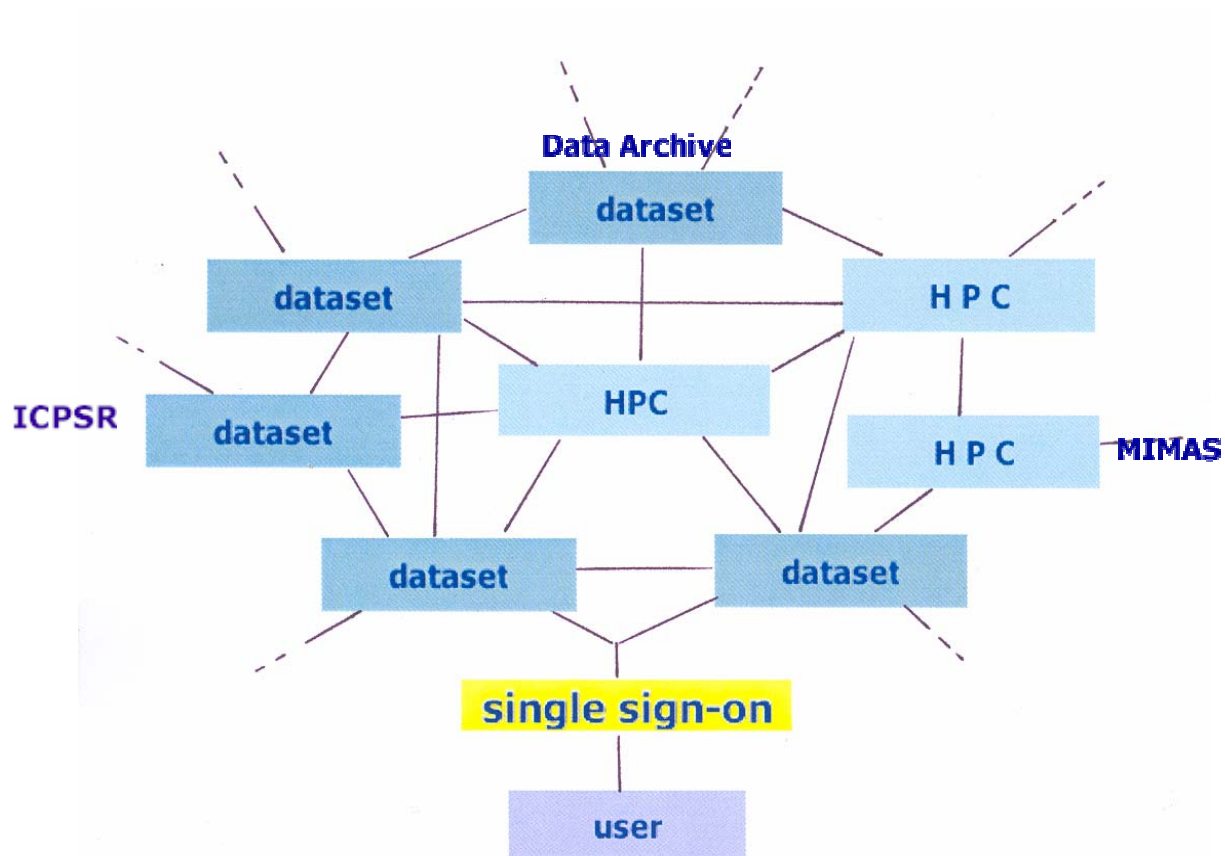


Current web model

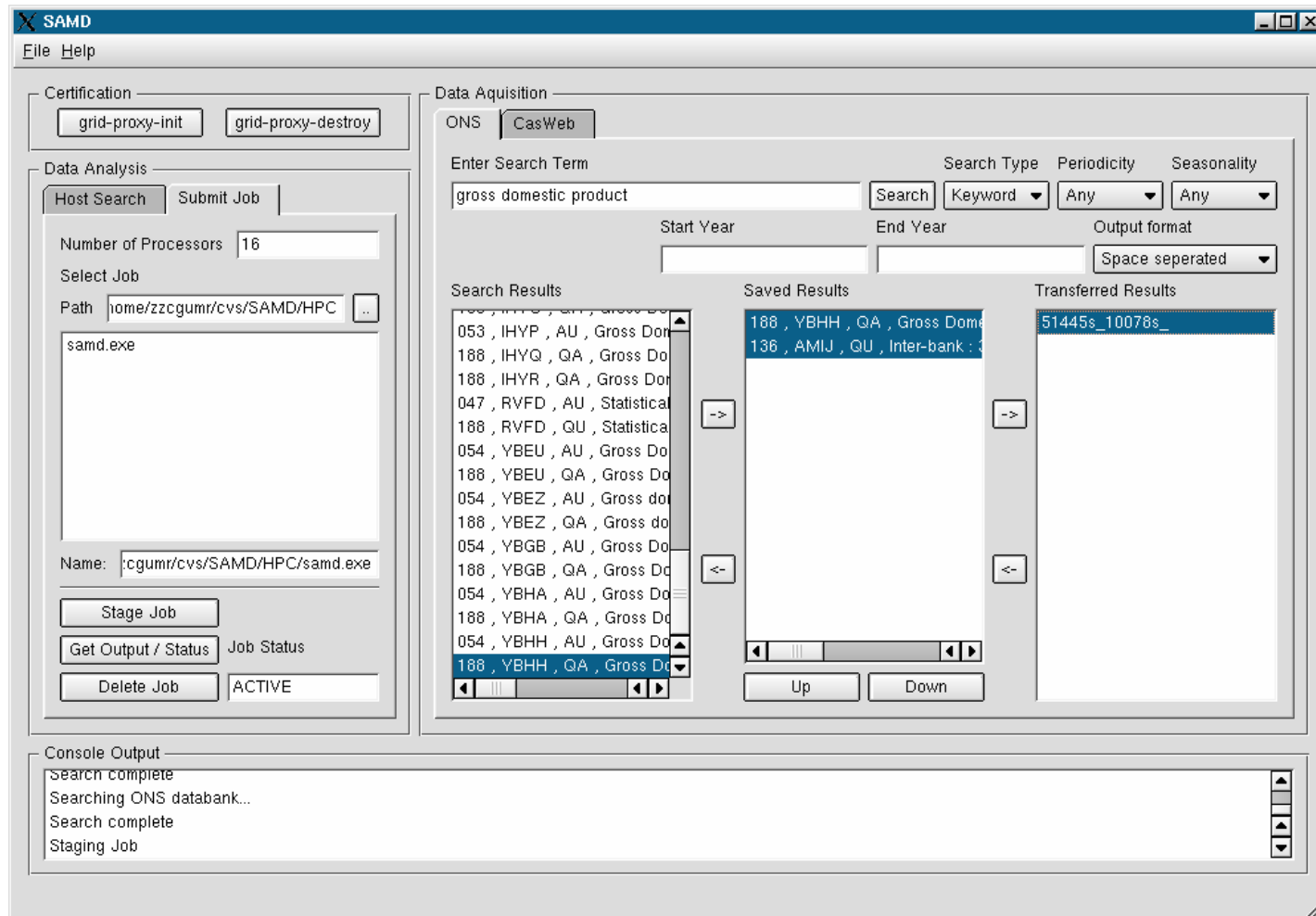
- Today – many separate accesses, ad hoc Client-Server



Grid Model Used



SAMD user interfaces



SAMD Methodology

We built a mini demonstrator grid for SAMD by:

- Grid-enabling the NS Time Series Databank
- Parallelising the code to represent the HPC facilities
- Using Grid protocols for data transfer
- Creating a graphical user interface that included a single sign-on
- It all worked, and cut the data collection and analysis time down to around 8 minutes.

The SAMD solution

- Use Grid Security Infrastructure for "single sign-on" authentication everywhere
 - Modified standard Apache web server to accept proxy credentials
 - Permits re-use of existing CGI code
- Use third party file transfers (grid-ftp) to move data directly to where it's needed
- Use standard globus mechanisms to
 - Locate HPC facility for analysis
 - Stage analysis binary from local repository and run analysis job on HPC facility
 - Retrieve results

Extending SAMD

- The approach and methods of SAMD are applicable to more general social science applications involving data collection and analysis
- Some of the SAMD resources reused in other Grid applications. These are available on the SAMD website:

<http://www.sve.man.ac.uk/Research/AtoZ/SAMD>

What's new with SAMD?

- More efficient handling of datasets – data is moved to where it's needed, not just to web browser
- The single sign-on for all databanks means users can cross search datasets and perform cross analyses of multiple datasets from different providers
- Grants access to high performance computing facilities without the user having to learn how to use them
- Can automate routine enquiries
- Cuts the time taken to run computing intensive problems by a factor of around 100

Scaling up with e-Social Science

A Grid approach allows the social scientist to scale up their quantitative research by:

- Including many more data points in their analysis
- Developing more complex models incorporating more variables
- Dropping assumptions
- Exploring new types of analyses

SAMD Acknowledgments

Keith Cole

Mark Riding

Geoff Lane

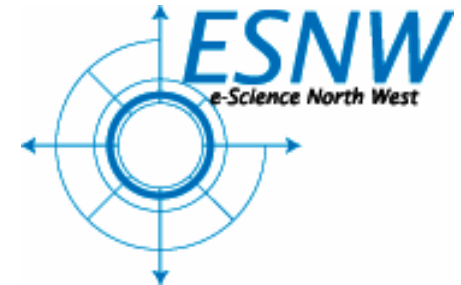
Celia Russell

Kevin Roy

Tim Hateley

Marianne Sensier

Stephen Pickles



Funded by the



and the

