

ConvertGrid

Keith Cole
Pascal Ekin
Linda Mason
Jon Maclaren

Presentation Overview

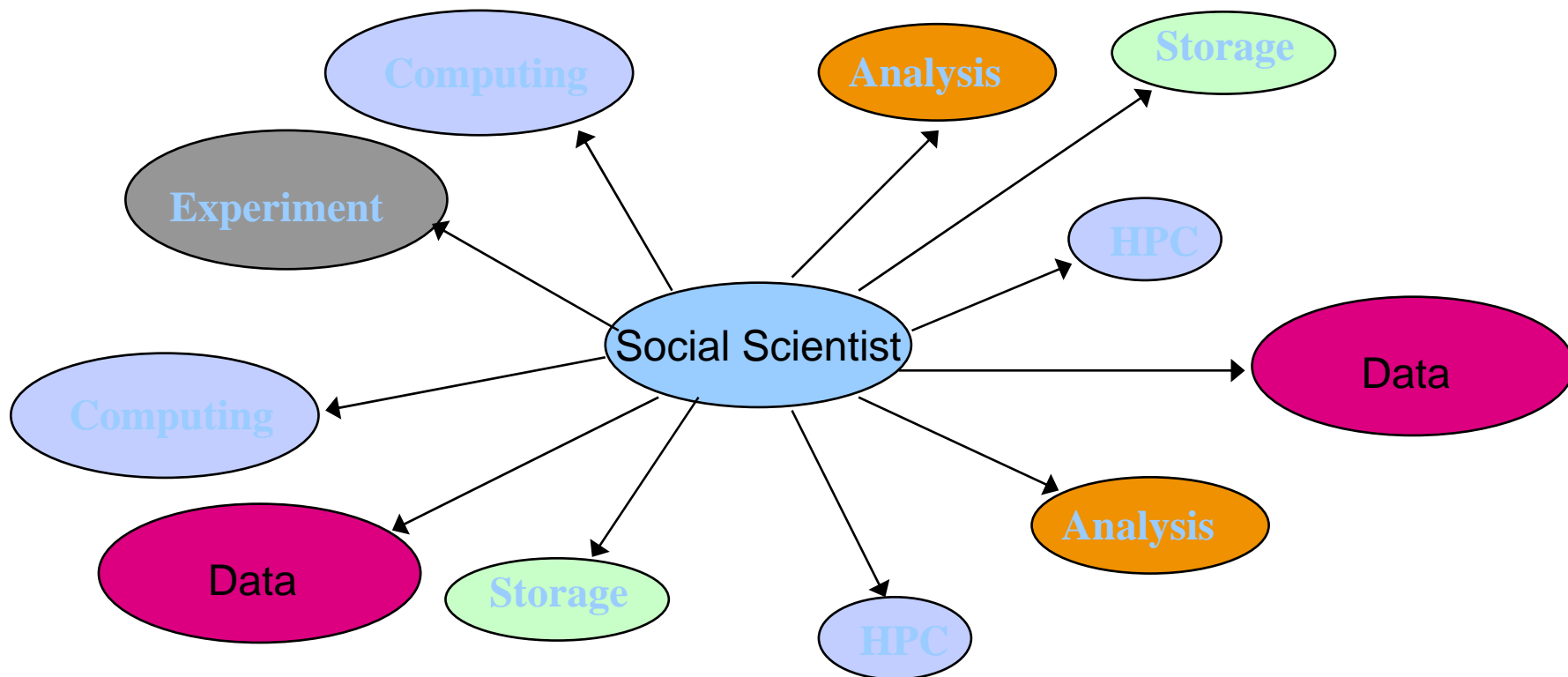
- Data Grids and the Social Sciences
- ESRC Pilot Demonstrator Programme
- ConvertGrid Objectives
- Research Context
- The ConvertGrid Demonstrator
 - Grid enabling population datasets & an existing web based service
 - A worked example
- Issues and Challenges
- Building the Social Science Data Grid
 - lessons learned and the next steps

What are the benefits of Data Grids for Social Science?

- Data Grids facilitate unimpeded use of distributed, heterogeneous, autonomous data resources.
 - Integrated view of the data resources that allow users to interact with them as if they constituted a single, global, integrated data resource.
- Grid enabling a dataset creates new opportunities for its use.
 - enables users to **integrate** it with other datasets
 - makes it possible to **analyse** the dataset using techniques that require the kind of computational power that it is only feasible using the Grid (e.g. more complex models, more data points).
 - standardisation of procedures and mechanisms used to access and update the dataset, increase its **shareability**
 - analyses can be re-run automatically when databases are updated.

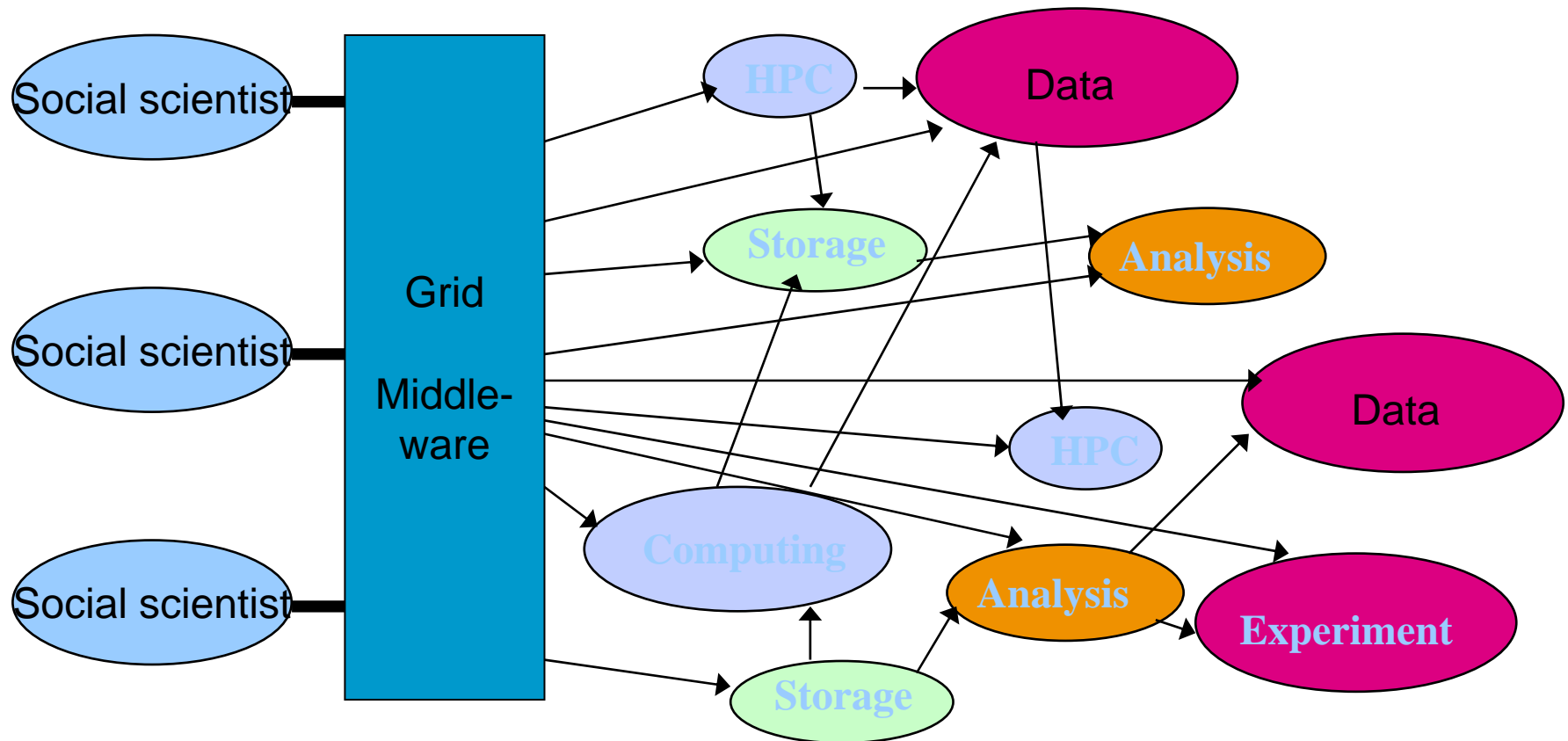
Research Infrastructure Today

- Many separate accesses, multiple architectures



Future Research Infrastructure –The Vision

- Grid middleware provided seamless integration of data, analytic tools and compute resources



ESRC Pilot Demonstrator Programme

- An early strand of ESRC's e-social Science Strategy
- Key objectives:
 - To demonstrate how Grid technologies can be used to enhance the capacity to address substantive social science research questions;
 - To produce a substantive research output based on the application and development of Grid technologies;
 - To produce a training output that can form part of a broader strategy aimed at demonstrating the application of Grid technologies to the wider social science community;
 - To work collaboratively with computational scientists to build an interdisciplinary community of scientists who can carry forward and develop e-science within the social sciences.
- 11 projects funded (2003-2005)

ConvertGrid – Key Objectives

- Provide a practical demonstration of how the Grid can be used to facilitate data integration and overcome a major barrier to research use of multiple datasets;
- Demonstrate how to build a social science Data Grid by grid enabling a number of key geo-referenced socio-economic data sources;
- Use Grid technologies to extend the functionality of an existing web based data service (i.e. Convert) to exploit the existence of a Data Grid;
- Demonstrate how Grid technologies can automate complex workflows and facilitate new forms of research
- Build a user interface to a Grid based service which is suitable for student/teaching use

ConvertGrid – The Research Context

- Many research questions require the combination of a data from multiple geo-referenced datasets
- The conversion of data relating to different geographies to a common target geography is a complex time consuming task requiring a range of data handling/processing skills. A major barrier to use!
- The data conversion process will require users to perform the following generic tasks:
 - Extract and download data in different formats from a number of databases using different interfaces
 - Convert each dataset to the desired target geography using geographical conversion tables
 - Combine the converted sets into a single dataset for analysis
- These generic tasks can be automated!

Different Source Geographies

- 1991 Wards
- 1991 Postcode Sectors



Source: Office for National Statistics

Existing Convert Service

- Developed as part of an ESRC funded project and transferred into service
- Provides access to 225 UK-wide geography conversion tables between census, electoral, administrative, postal, health and statistical geographies derived from the All Fields Postcode Directory.
- Facility to convert a researcher's data from one set of geographical units to another (e.g. from postcode geography to health geography)
- The data conversion is improved by proportional allocation for those source units that do not fit within a single target unit.
- Extensible system - further conversion tables from any source can be incorporated

ConvertGrid - Data Sources Used

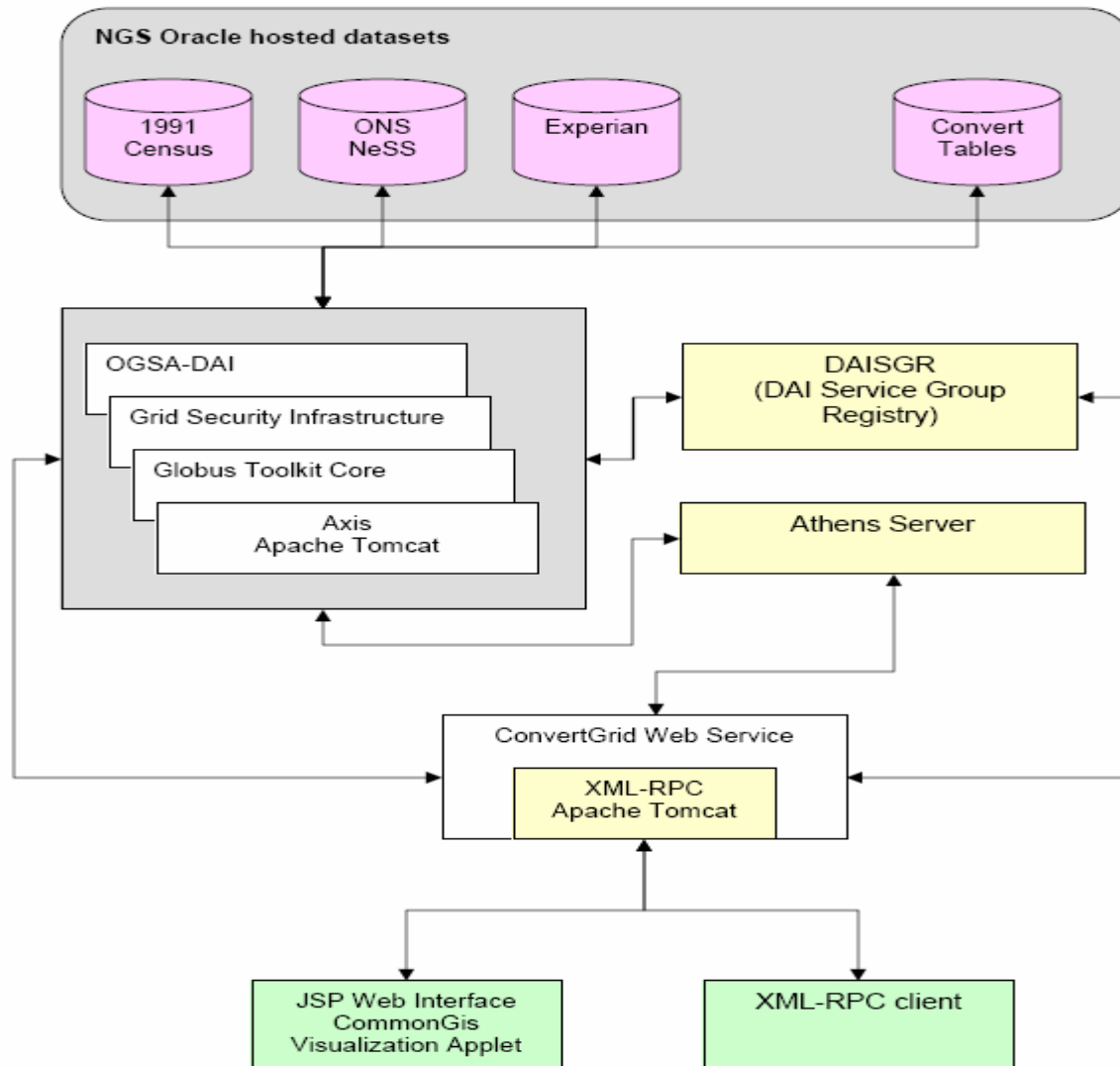
■ Data Sources

- 1991 Census Aggregate Statistics (1991 Census geographies)
- ONS Neighbourhood Statistics (1998 Ward & Districts)
- Experian (2000 Postcode Sectors)
- All Fields Postcode Directory (AFPD) (1999b)

■ Selection criteria

- Data on a range of themes to support Health, Education and Crime use cases.
- Different geographies and time points
- AFPD derived conversion tables available for geographies via Convert system

ConvertGrid Architecture



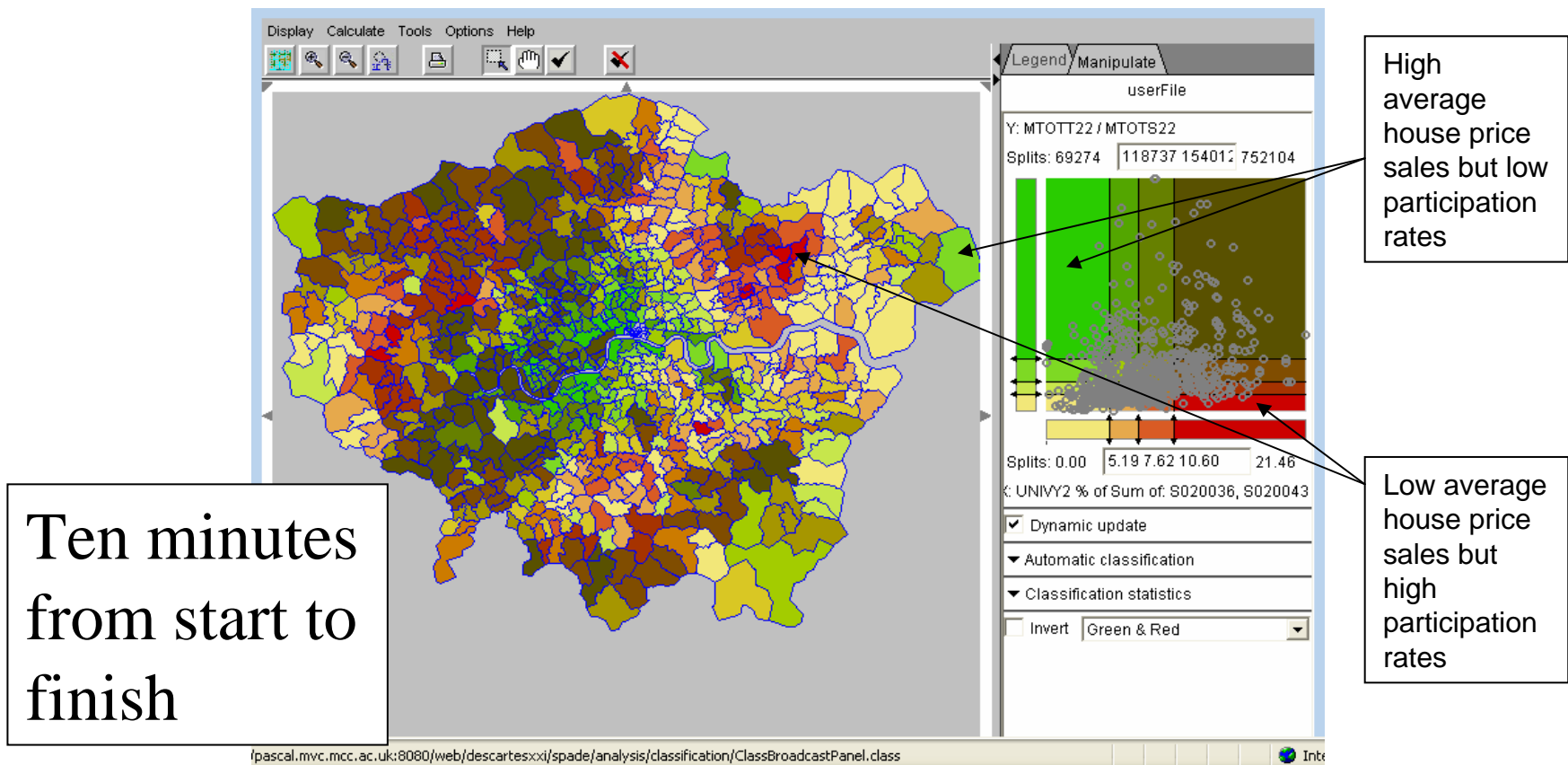
ConvertGrid – Services Provided

- Converts data sources with different native geographies to a common Target Geography and outputs combined data as:
 - A data stream in CSV or XML format or
 - Transferred to a web based visualisation system
- Grid-enabled datasets (incl. AFPD)
 - Available to other Grid services via NGS
- Accessible to users via a ‘classic’ web based interface
 - Essential for demonstration purposes
 - Step by step guide developed
- Extensible system
 - Available to other applications via a web services interface
 - Easy to add other Grid-enabled datasets to the system

ConvertGrid – A Worked Example

- What factors explain spatial variations in participation rates in higher education
- Study target geography –1991 Census Ward
- Data required:
 - 1991 Census
 - Total persons aged 16-17 & 18-19 (1991 Census Ward)
 - Neighbourhood Statistics
 - Number of applicants aged under 20 entering university (1998 Electoral Ward)
 - Experian
 - Average house price sales Quarter 2 2000 to Quarter 1 2001 (1999 Postcode Sectors)

ConvertGrid – Data Visualisation Interface



- Relationship between average house price sales (Experian) and percentage of 16-19 year olds entering university (Neighbourhood Statistics & Census aggregate statistics)

ConvertGrid – Issues and Challenges

- Establishment of a Grid infrastructure
 - Early adopter of the National Grid Service
 - Key Grid middleware immature and under rapid development
- Database migration problems
 - SQLServer to Oracle on the National Grid Service
 - Maintaining multiple databases resource intensive
- Data comparability issues a problem
 - Postcode formats
- Developing metadata registries
 - For resource discovery, data access and interpretation
- System performance, scalability and security
 - OGSA-DAI still relatively inefficient
 - Implementation of Grid security non-trivial

Building the e-Social Science Data Grid - The Next Steps (1)

- Establishing a production social science Data Grid is a key component of the wider e-Social Science strategy.
- Current social science data infrastructure (academic and non-academic) needs to be Grid enabled in a standards compliant and sustainable way.
- Data service infrastructures need to be able to support multiple forms of access (i.e. single database approach) to minimise duplication of effort.
- Still technical problems to resolve
 - Mapping UK e-Science certificates to current and future access management protocols
 - Linking data and metadata

Building the e-Social Science Data Grid - The Next Steps (2)

- Managing user expectations is very important.
 - Data Grids do not solve all research problems
 - Many demonstrators but few production services
- Complex, distributed workflows makes the development and deployment of services challenging and may require substantial software engineering!
- Grid enabling the underlying databases may turn out to be the easy bit! Methodologies and intermediary applications/interfaces to facilitate data integration/analysis is much harder.
- Finally, MIMAS is being funded by JISC to Grid enable the 2001 Census aggregate statistics (GEMS project) as part of building a production Data Grid via the NGS.
 - Connecting the MS SQLServer databases holding the 2001 Census aggregate data directly to the Grid via the NGS
 - Grid enabling the current data access system (Casweb)

Acknowledgements

- ConvertGrid & GEMS Teams @ Manchester
 - Pascal Ekin
 - Linda Mason
 - Stephen Pickles
 - Jon McLaren
 - Justin Hayes
 - Mat Ford (NGS)

- NCeSS
 - Laura Bond (NCeSS)
 - Alvaro Fernandes (Computer Science)