EDUCATION AND TRAINING IN DATA HANDLING AND ANALYSIS AT THE INTERFACE BETWEEN E-INFRASTRUCTURE AND RESEARCHERS

David Fergusson

National e-Science Centre, University of Edinburgh, Informatics Forum, 10 Crichton Street, EH8 9AB Edinburgh, United Kingdom
*Email:*dfergusson@dfmac.demon.co.uk

1 STATE OF THE ART

Much effort and concentration has been put into devising training regimes for a number of different technologies in distributed and high-performance computing (Jandric, Artacho, Hopkins, & Fergusson, 2008; Fergusson, D., Romano, van der Meer, & Atkinson, 2008). On the whole, however, these have tended to concentrate on the computational aspects of research tasks rather than the data-related aspects. There have been a number of reasons for this including the immaturity and extra complexity of the data field, the more discipline-specific aspects of data usage compared to computational patterns, and the focus of providers on the "easier" problem of providing distributed computation resources (Fergusson, 2006).

Data are however fundamental to research activities, and nearly all computational tasks (outside pure simulations) involve some form of transformation of a dataset. Having recognized this, we must therefore ask ourselves what type of training support is required by researchers. Do researchers already understand their datasets and how to manipulate them sufficiently?

It may seem impertinent to answer "no" to this latter question but many—to generalize, all—research domains have realized in the last decade or so that the changes in automation and instrumentation have meant that the ability to acquire data has greatly outstripped the ability to process (analyze, manipulate, store, and archive) them. This trend has been particularly recognized within the physics and biology domains, largely due to the advances in data acquisition within these domains.

Thus we can say that there is clearly a need, articulated by the user communities themselves, for both data focused solutions and the concomitant training support.

2 TEN-YEAR VISION

By 2020, e-infrastructure and, more specifically, data management training is included in the academic curricula in Europe and beyond. Elements of data management training are included even in the secondary educational system, and a growing number of skilled ICT pupils and students exist. By 2020 the shortage in the area of skilled people in ICT and e-infrastructures is being reduced.

Curricula Design

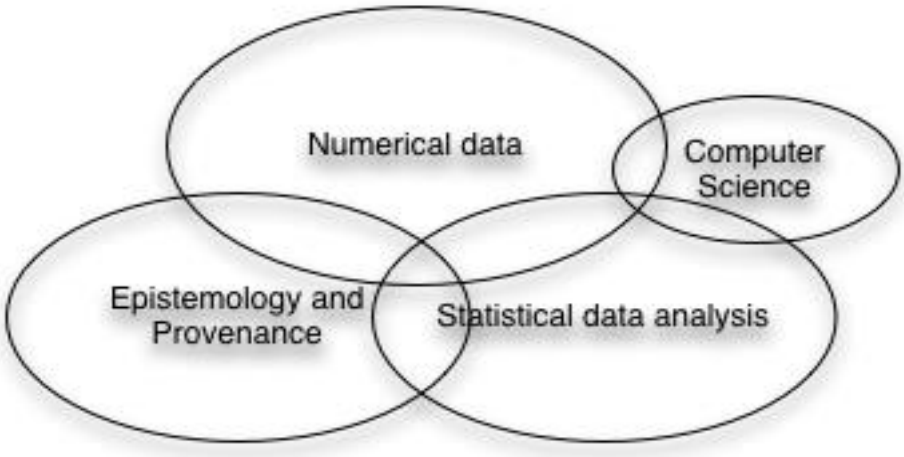
These goals and challenges for e-Science education figured prominently in curricula development discussions at the ICEAGE Curricula Development Workshop held in Brussels from 14–15 February 2008. Professor Malcolm Atkinson (Director, e-Science Institute, University of Edinburgh) and Dr. David Fergusson (Deputy Director of Training, Outreach, and Education, National e-Science Center, University of Edinburgh) co-chaired the workshop (van der Meer, Atkinson, & Fergusson, 2009). e-Science and distributed computing educators from academic institutions in Europe and the US attended the workshop, which answered a call to action voiced by OGF and e-IRG members. Policy reports from the OGF Education and Training Community Group (ET-CG) and e-IRG Education

and Training Task Force (ETTF) (van der Meer, Atkinson, & Fergusson, 2009) documented a profound lack of well-developed e-Science curricula at both the undergraduate and graduate levels.

It is well understood that the data models for different research domains are informed by the research practices within the domain and at the same time the data models must be driven by the nature of the data presented and collected within the scope of the domain (Voss, Asgari-Targhi, Procter, Halfpenny, Dunn, Fragkouli, et al., 2009; Voss, Asgari-Targhi, Procter, van der Meer, & Fergusson, 2010; Voss, van der Meer, & Fergusson, 2010). Thus different domains have different data structures and approaches to analyzing that data. For any training or education activity, the most important and fundamental resource required is some way of simulating the real life activity in a safe and secure environment that supports learning. In the examples under discussion, this means a mechanism for realistically simulating the data and analysis environment that real-life practitioners will find themselves in.

The most important, and consequently often the most difficult to simulate, aspects of this data environment will vary considerably from domain to domain. For example it is clear that in high energy physics the important aspect of the data environment is in the scale of data and rate of acquisition. However, in bioinformatics and biomedical research, it is often the complexity of data and the need to synthesize data from a variety of sources that can be problematic. Further, in both biomedical and social research, constraints are put on the availability and use of data by ethical considerations. In some cases these issues may be addressed in a relatively simple way (although still possibly requiring significant resources), for instance the use of random data derived from Monte Carlo simulations for high energy physics. This may not be the case, however, with biomedical training where data reflecting real life are an absolute requirement (Voss, van der Meer, & Fergusson, 2010).

In order to manage this problem, it is important to try to delineate as much as possible what domains (or aspects of domains) can be treated collectively through broad solutions or approaches and where it will be necessary for custom solutions to be provided.



Numerical Data: Physics, Engineering, Earth Systems, Chemistry, Materials Science
Epistemology and Provenance: Arts, Languages, Humanities
Statistical Data: Biology, Medicine, Social Sciences, Economics

Figure 1: Predominant models for curriculum design

Computer science students and software engineers require different content than students in other domains who use e-Science to further their research (in other words, each discipline needs tailored content). Different models for research prevail in different fields, and educators must consider these models when crafting relevant e-Science curricula. For instance, in physics, engineering, earth systems science, chemistry, and materials science, numerical models predominate. In contrast, biology, medicine, the social sciences, and economics often rely on statistical models while epistemology and provenance (involving conceptual models and narratives) dominate in the arts, humanities, and languages (see Figure 1) (van der Meer, Atkinson, & Fergusson, 2009).

Workshop attendees chose to focus on developing a framework for e-Science education across disciplines rather than on distributed computing education that would target students in computer science. The workshop began with a discussion of undergraduate content and then moved to masters-level content.

The resulting analysis and recommendations have also been accepted by the e-Infrastructures Reflection Group (e-IRG) as part of the output of their Education and Training Taskforce (eTTF) and similarly by the Education and Training Community Group of the Open Grid Forum (OGF).

3 CURRENT CHALLENGES

Models of Data in Training (Disciplinary Approaches to Data and Analysis)

As the different research domains are facing similar problems arising from similar developments, can we find common solutions that will allow for common approaches to training support?

Data in Different Research Domains

Unfortunately, the answer to this question appears to be that the approaches to data, and therefore the requirements of the different fields are not common. (We should not be surprised by this as the divisions between research domains have partly grown-up due to the different nature of their data and the required analysis).

There will therefore have to be significant customisations of training resources to fulfil the requirements of specific research domains (Jandric, Artacho, Hopkins, & Fergusson, 2008).

Some simple models have been used successfully in practice based on the general grouping of domains discussed earlier. For instance a model of a simple data space containing data items, which has to be scanned to locate, identify, and analyse these items. This model allows for the incorporation of noise and the tailoring of the signal-to-noise ratio to allow a more or less realistic model to be presented to the students (Fergusson, Hopkins, Romano, van der Meer, & Atkinson, 2008).

This type of model is useful as an introduction to the use of basic eScience techniques and tools, for instance familiarisation with the EGI infrastructure and its methods for manipulating jobs and data access. However, as these models become more complex and realistic, they also become more and more specialized, requiring more use of specific analysis techniques relating to the particular domain.

The current task for the infrastructure providers is to develop closer ties to their clients (user communities) so that they can work with these communities individually to produce tailored training solutions for data analysis—for instance working closely with ESFRI projects.

Data Standards

Many research domains have been developing successful community approaches to digital data over a number of years. Three example domains spring to mind: biology, geosciences, and physics. Each of these domains has been driven by increasing data acquisition rates, largely driven by automation and improved instrumentation—for

instance, automated sequencing, satellite tracking, and medical imaging in biology; satellite data, geographical information systems and simulations in geosciences; developing colliders, light sources, and simulations in physics. Each domain has developed its own social and informational model to deal with increased data acquisitions rates.

In particular, in the field of bioinformatics, there has been a long-term (approximately 40 years) development of shared community databases of sequences (Genbank, Swissprot, Uniprot, PDB, KEGG). Standardised data formats based on these databases and on the related analytical programs are well established (e.g., FASTA). More recently, sets of standards for biomedical images have emerged (e.g., DICOM).

In contrast, geophysical scientists have generally used commercial analysis programs and datasets, leading to the need to understand various proprietary data formats.

We therefore find that there can often be very advanced understanding of the challenges relating to data within a particular field, but this varies very widely between fields. This results in a situation where standardization of data formats, analysis, and access methods is well developed within fields but there is little or no commonality between different fields. This effect may be seen at many different scales not just between major domain groupings (biology, physics, … etc.) but commonly at increasingly smaller scales, , the stage where health services, for instance, may have difficulties transferring patients records between departments within the same hospital because they take different approaches to data formats and data semantics (Voss, Asgari-Targhi, Procter, Halfpenny, Dunn, Fragkouli, et al., 2009; Voss, Asgari-Targhi, Procter, van der Meer, & Fergusson, 2010; Voss, van der Meer, & Fergusson, 2010) .

We need also to be clear that education alone will not be enough to address these issues; many of them are fundamental and deeply entrenched within working practices. Education will allow users to come to an understanding of the advantages of shared approaches and standardization, but until the tools are available to address the issues directly, this can only be a theoretical appreciation.

These different approaches need to be understood and supported in providing training support for these communities.

Security and Privacy

The needs for security and privacy in dealing with research data varies widely between different domains. Some domains, for example high energy physics (HEP), have effectively no requirements for privacy of data. However, in the biomedical domain, the majority of data may have some privacy implications. This disparity between domains must be supported in the provision of training.

This example also highlights the potential for an apparent tension between security and privacy although it is becoming clearer that these two aspects are facets of the same issues, i.e., security relies on an understanding of what the status of data items is (public/private) in order that the appropriate level of security can be made available.

It is important to understand that training must be tailored to the needs of particular users. It is a mistake to force HEP students to study privacy issues without the underlying knowledge for it making sense to them as much as it would be a mistake to not provide bio-med students with training in the tools to support their needs for data privacy support.

Access to Data

Access to data with the appropriate characteristics—i.e., scale, realistic/real patient data (with appropriate anonymisation), commercial data sources, etc.—is a major problem in presenting training. All of these aspects impose some costs on the provision of the models to be used for training. Commercial data sources, which are essential for instance in much of the geosciences, obviously require direct payment to the data owners; this may be

on the order of 10s to 100s of thousands of Euros. Real patient data are commonly used in teaching in the clinical biomedical sciences, often that derived from research studies. However, this usage requires specific consent and sufficient anonymisation (examples can be seen in the UK Data Archive: http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation). This issue extends also into the social sciences. The alternative of creating realistic but synthetic datasets of sufficient scale and quality is generally more costly than repurposing existing research datasets for educational use. However, students must be given access to data that represent a convincing model of the data sources that they may expect during their normal work.

Analysing and Annotating Data from Multiple Sources

An important and increasing requirement in most domains is the amalgamation of data from a wide variety of sources. These data sources are often a mix of local and public data (increasingly from community databases).

For training purposes this requires access to model local datasets and either the real community datasets or convincing models of them.

Structured and Unstructured Data

The relative use of structured and unstructured data sources is again highly dependent on the domain under discussion. Some domains are mainly concerned with effectively unstructured data directly acquired from instrumentation (e.g., HEP) while others use structured databases as data sources almost exclusively (e.g., geosciences).

Again, the training regime provided must support the correct model for the student's domain.

Instrumentation and Data Capture

Data capture brings a new dimension to data manipulation, that of time. Combined with scale and the likelihood of proprietary data formats from instruments, this brings a special set of problems to providing good models of this process for training.

Data Presentation and Manipulation

The final analysis of data and its presentation is often the realm of proprietorial systems (Spotfire, Power Point, etc.) and is probably beyond the scope of this discussion.

4 RESEARCH DIRECTIONS PROPOSED

To support development of practical skills for e-Science, we recommend emphasizing the production of educational materials, sharing these materials through repositories, and sharing use of specific teaching infrastructures. The e-IRG (www.e-irg.eu), OGF (www.ogf.org/index.php), and the EU FP6 ICEAGE Project identified a lack of suitable e-Science textbooks so it is important for the key stakeholders in e-Science education to support the establishment of specific Web sites and other forums to pool, share, and debate textbook content. For example, see the online SURA Grid Technology Cookbook and the JISC "Research in a Connected World" text book (http://www.lulu.com/product/paperback/research-in-a-connected-world/13525114). It would also be important to develop incentives, such as competitions, in conjunction with editors and publishers, to produce textbooks that follow agreed-upon educational goals and curricula.

Educators and other key stakeholders should then share these materials using repositories and share use of e-infrastructures. We can look to the specific recommendations from the OGF and in other articles for best practice in these areas. Members of the e-Science educational community should support the development of digital libraries

and training infrastructures at local and national levels and provide means to link local and national resources. These initiatives would provide a starting point and standard resources for university programs. Many universities will build on this to tailor education for their students and discipline specialties.

5 RECOMMENDATIONS

Investment in e-Infrastructure Education

Investment in relevant education and training, which aims primarily to equip graduates to use e-infrastructure well, should be comparable with the investment that is going into e-infrastructure provision. Universities should adapt their curricula to prepare graduates. Investment, such as incentives to modify curricula and help with the introduction of new curricula, is necessary to trigger the required rapid and extensive change.

Similarly there should be investment in coordinating the training efforts of large EU research projects (i.e., ESFRI) where these use e-infrastructure and eScience tools in order to promote efficiency through the sharing of resources and knowledge. e-Infrastructure projects should be in an ideal situation to promote this sharing as they are in contact with many of the projects involved.

The Development of Distributed-Computation Knowledge and Skills

Academic institutions, particularly universities, should build on existing "seed" courses and curricula to revise curricula in the majority of disciplines. Relevant professional bodies and individuals proposing to teach this material should undertake further work on curricula. The goal of this alignment should be cross-fertilization, mutual recognition, and increased understanding, not uniformity.

It will be essential in all cases for the education providers to work closely with the local infrastructure providers and also the local research users to provide realistic and appropriate examples for students.

It will be necessary for e-infrastructure projects to work closely with domain researchers in this area in order to ensure that the necessary materials and supporting technologies are in place to allow this type of education to be taken on by existing teachers in the domains.

Education in the Use of e-Infrastructure

Professional bodies, such as the Royal Society of Chemists and the Institute for Engineering and Technology in the UK, should be encouraged to identify target attainments in the exploitation of e-infrastructure for their professions. This will allow harmonization of these attainments across the European research area in accord with the Bologna framework (http://ec.europa.eu/education/policies/educ/bologna/bologna_en.html). The goal of this harmonization is not uniformity of skills and knowledge; rather, it's a common framework to support student and worker mobility and mutual recognition of qualifications, particularly where they influence the appointment of staff to positions in which the use or operation of e-infrastructure is mission critical.

This process has been underway for over a decade within the EU, and it has led to many important developments in the mobility of students and the recognition of qualifications across the EU. In many cases this will mean the incorporation of modules or materials specific to e-infrastructure and their use in the field rather than new qualifications specific to e-infrastructure usage.

Standards

Here we consider standards for identification to enable access to and management of educational infrastructure facilities. A task force, set up by the e-IRG, European Grid Initiative (EGI, http://web.eu-egi.eu), and GÉANT

(www.geant.net), should extend the eduroam protocols (www.eduroam.org) to cover student and teacher use of collaboration facilities and multisite infrastructure.

It will also be important in the data domain for educators to work closely with the domain specific producers of data, for instance with the ESFRI projects in order to understand the needs of the researchers in that domain and so be able to produce training materials appropriate for the students from that domain.

Standards for Sharing Material

Those in the EU developing e-infrastructure courses should build on the Creative Commons for policies governing the sharing of educational material. The EGI, when it becomes operational, should mediate agreement between National Grid Initiatives (NGIs) on sharing infrastructure.

6 REFERENCES

Artacho, M., Jandric, P., & Fergusson, D. (2007) Online learning: the International Winter School on Grid Computing. *International Science Grid This Week*, Geneva: CERN.

Fergusson, D. (2006) Grid education: past experiences (EGEE) and future directions (ICEAGE). *WETICE Conference*, plenary presentation. Manchester, UK.

Fergusson, D. (2006) Grid education: past experiences (EGEE) and future directions (ICEAGE). *EUNIS Conference*, plenary presentation. Tartu, Estonia.

Fergusson, D., Hopkins, R., Romano, D., van der Meer, E., & Atkinson, M. (2008) Distributed Computing Education, Part 2: International Summer Schools. *IEEE Distributed Systems Online 9* (7) art. no. 0807-7002.

Fergusson, D., Romano, D., van der Meer, E., & Atkinson, M. (2008) Distributed Computing Education, Part 1: A Special Case? *IEEE Distributed Systems Online 9*(6)

Jandric, P., Artacho, M., Hopkins, R., & Fergusson, D. (2008) From Distributed Computing to Distributed Learning. Proceedings of the *Seventh IASTED International Conference on Web-based Education WBE 2008*, Innsbruck, WBE.

van der Meer, E., Atkinson, M., & Fergusson, D. (2008) Distributed Computing Education, Part 3: Winter School. *IEEE Distributed Systems Online.*

van der Meer, E., Atkinson, M., & Fergusson, D. (2009) Distributed Computing Education, Part 6: Curriculum Development. *IEEE Distributed Systems Online.*

van der Meer, E., Atkinson, M, & Fergusson, D. (2009) Distributed Computing Education, Part 7: Policy Frameworks. *IEEE Distributed Systems Online*.

Voss, A., Asgari-Targhi, M., Procter, R., Halfpenny, P., Dunn, S., Fragkouli, E., Anderson, S., Hughes, L., Fergusson, D., van der Mere, E., & Atkinson, M. (2009) Paths to Wider Adoption of e-Infrastructure Services. *Information, Communication & Society*.

Voss, A., Asgari-Targhi, M., Procter, R., van der Meer, E., & Fergusson, D. (2010) Adoption of e-Infrastructure Services: findings, issues and opportunities. *Phil Trans R Soc 368*, pp 4161-4176.

Voss, A., van der Meer, E., & Fergusson, D. (2010) Research in a Connected World.

(Article history: Available online 30 July 2013)