INNOVATION IN SCIENTIFIC MODELLING AND SIMULATION AS INFRASTRUCTURE ASSEMBLY

Mark Hartswood (1), Alex Voss (2)

(1) School of Informatics, University of Edinburgh, mjh@inf.ed.ac.uk; (2) School of Computer Science, University of St Andrews, alex.voss@st-andrews.ac.uk

Abstract

We propose a generalised framework to describe the development of scientific models and simulations as infrastructure assembly. *We take our inspiration from Paul Edwards' classic account of the emergence of Climate Science as the creation layer by layer of supportive socio-technical infrastructures over timescales measured in decades leading to a mature simulation based-discipline able to make authoritative predictions with a significant impact upon policy formation. Our framework isolates important infrastructural components that appear to be common in modelling and simulation based science across a number of domains, the sophistication of which we use as indices to describe the maturity of a given modelling approach. This enables us to explain otherwise confusing reports from the literature that alternately describe modelling and simulation based science to be in crisis, or else in a healthy state based upon the degree that projects studied manage to implement robust software engineering and testing practices. Equating innovation in scientific models with long cycles of infrastructure assembly enables us to distinguish between those modelling approaches that have run into real difficulties, and those that are merely less mature, enabling better targeting of remedial interventions.*

Keywords: Information Infrastructure, Model and Simulation based science, Software Engineering, Validation and Verification practices.

1 INTRODUCTION

Use of computational models and simulations to underpin scientific research is now commonplace in a broad range of disciplinary areas including the physical sciences, engineering, economics and the social sciences. The ubiquity of computational models and their strong contributions to policy formation and scientific advance, often in economically important or politically contentious arenas such as climate change, has led to an increasing focus on the quality of the engineering practices used in their production. A number of scholars have examined the practices of code production and use within science to understand how scientists can take greater advantage of software engineering 'best practice' to improve the reliability of their codes. Taken at face value, their findings reflect diverse and often contradictory accounts of the quality of software engineering and testing practice across a range of scientific communities. Some papers pessimistically report 'gridlock' from 'divergent values' of software engineers and computational scientist (e.g. Faulk et al, 2009; Kelly and Saunders, 2008), while others present a rosier picture of mature testing practice, and of sophisticated software engineering process (e.g. Easterbrook and Johns, 2009; Holzworth et al, 2011). Those presenting a negative perspective point to lack of data against which to test models and simulations, poor use of software process and tools, and immature testing procedures that rely too heavily on professional judgement. An important concern of these papers is that errors may go undetected undermining trust in modelling approaches and thereby degrading their worth. To make sense of this seemingly disparate array of findings, and informed by interview data from a pilot study, we developed an analytical framework based on the idea that creation of models and simulations is inseparable from the creation of infrastructure.

2 INFORMATION INFRASTRUCTURE AND COMPUTATIONAL SCIENCE

The inspiration for our approach came from Paul Edward's book 'A Vast Machine', which charts the history of climate science from its amateur origins to its present day status as a highly organised and coordinated set of scientific processes depending on a series of infrastructures that bind together a

heterogeneous array of current, historical and simulated data, technicians and scientists and models, simulations and software processes (Edwards, 2010). This summary of Edward's account forms the central idea for our analytical framework, which aims to produce a generalised description of how model and simulation based science advance by the incremental development of supportive infrastructural components over lengthy timescales – decades in the case of climate science. One example from Edwards that illustrates this process is the increasing diffusion across wider geographical areas of standardised referents for the measurement of time, driven initially as a requirement for smooth operation of developing rail transport infrastructures, and then appropriated as a means of coordinating geographically dispersed meteorological readings so that they could be meaningfully combined. The increasing sophistication and predictive power of climate science is shown to be punctuated by a series of infrastructural innovations that create the necessary supportive elements for 'next steps' in the progression to be made. These infrastructural elements are socio-technical in character, where technical and non-technical components (such as standards, agreements, processes and collaborations) are mutually supportive. Another key feature of modelling science revealed by Edwards' account is the relationship between the suppliers and consumers of data obtained from the real world. This data is needed by climate modellers both to inform model parameters (so the models have a basis in reality) and as a source of validation data (so that the models are tested against reality). Initially, meteorological data collection networks were created for the purpose of weather forecasting, as opposed to climate modelling, and the emergence of climate science is also partly the story of the emerging alignment between forecasting and modelling communities. We elaborate this point below when we discuss our interview data, but emphasise here that it seems typical for modelling science to be highly dependent on, and have evolving relationships with, 'primary' experimental science communities. The final point to take from Edward's work concerns his term 'Knowledge Infrastructures' used to convey how infrastructure underwrites the knowledge claims by providing a foundation of trusted and taken for granted components, processes and theories upon which new claims can rest. Taken together, these ideas suggest an interpretation of our case study data whereby the sophistication of a modelling approach and its predictive power varies with the sophistication and quality of its supportive infrastructures.

3 CASE STUDIES

In a pilot project we have interviewed eight scientists and technicians engaged in model- or simulation based science to inform the development of a research proposal aimed to assist computational scientists to adopt software engineering techniques and approaches. Our initial sample was selected by contacting scientists known to Voss via his existing research into barriers to uptake of advanced technology support for science. A 'snowball' approach was used to identify additional interviewees. We conducted semi-structured interviews lasting about an hour in which participants were asked to explain: their modelling approach and its scientific context; their software engineering and testing practices; and what additional support would help most to advance their work. Our interviewees fell into three broad groups: those creating agent-based models of social phenomena such as crime, population movements and responsibility (n=4); those working within particle physics creating and interpreting experiments conducted on the Large Hadron Collider at CERN (n=3) and a data driven scientist involved biomedical research (n=1). The work of each of these groups is briefly illustrated in the following sketches:

Sketch 1: Particle physics: *Experimental work is supported by a large and complex technical infrastructure created to handle and process the massive volumes of data generated by particle collisions within the accelerator. A modelling process is used to show what phenomena described by the theory should look like in the data from actual experiments. A sophisticated series of software processes has been implemented, including: a tagging system coupled with version control to enable software configurations used in a given processing chain to be identified and replicated; and a sophisticated series of nightly software and validation tests to confirm that updates to the software function correctly and are consistent with scientific understanding. Publications are produced that advance the field of theoretical physics.*

Sketch 2: Data driven science: *Data culled from existing biomedical publications assembled through exhaustive literature searches is used to populate models of human biological processes, for example, hormonal responses. This data is assembled from tables of data within the papers, or from software that can reverse-engineer data points from published graphs. The modeller, a computer scientist, works closely with clinical scientists to establish the quality of the input data and the interpretation and validity*

of the model results. The numerical modelling approaches used are all long established and available as libraries bundled with routinely used scientific software packages. The aim of the exercise is not to make scientific discoveries per se, but by pooling a range of historical data to generate hypotheses that can then be tested by experimental science. The work is however published in the primary literature in the relevant clinical domains.

Sketch 3: Agent based modelling: *By modelling the actions of individuals within a population and enabling them to interact with each other and their environment it is possible to explore a range of emergent behaviours by adjusting the 'rules' of those interactions. Most of the agent based modellers we spoke to use existing modelling frameworks to provide basic configurable agent behaviour, and wrote additional modules to constrain the behaviour of the framework in relevant ways. Those working in population-based modelling had a series of links to a series of other communities that were able to provide validation data, but often not at the frequency or granularity that was ideal. None of the agent-based modellers produced publications that directly advanced the primary scientific field, but instead published improved modelling approaches.*

4 AN ANALYTICAL FRAMEWORK FOR MODELLING PRACTICE

Comparison between the practices of the scientists interviewed in our pilot also revealed differences in access to validation data, sophistication of software process and the maturity of the modelling approach similar to those reported in the literature briefly reviewed above. However, our interviews also gave the sense that those deficits were 'appropriate', or at least expectable, given the developmental stage of the science being undertaken. Moreover, less mature aspects of a model's development did not risk an erroneous contribution to science because scientists were acutely sensitive to the model's competence, and were very careful only to make knowledge claims that were proportional to the model's capabilities. Figure 1 attempts to capture the idea that the maturity of modelling disciplines falls along a spectrum.

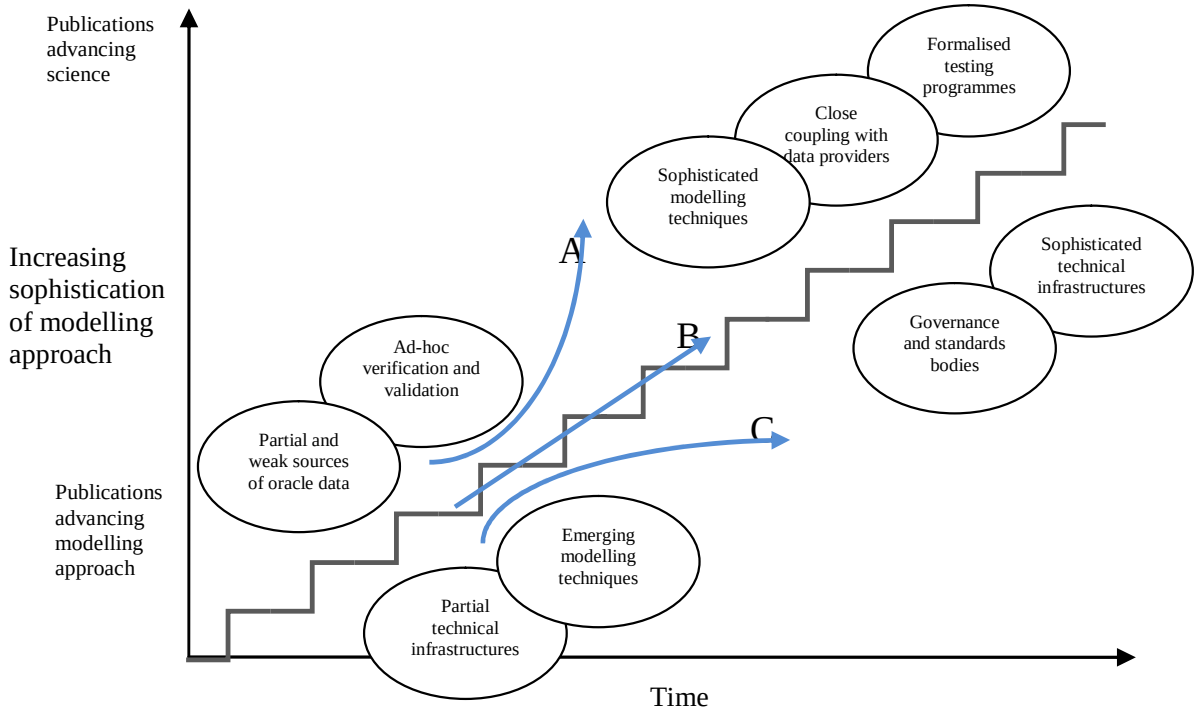

Figure 1: How modelling approaches mature with time in terms of their infrastructural components.

Models in an early stage of their career can be found in the bottom left hand corner of the diagram, and those that are more mature in the top right hand quadrant. From the literature and from our own data we identify models as increasingly mature with:

- the increasing coupling between data providers and modellers, where experimentation becomes geared to the modeller's hunger for data,

- the emergence of 'rounded' and 'complete' technical infrastructures – providing access to mature modelling frameworks, appropriate compute and storage resources, automation, data visualisation etc.,
- the emergence of governance and standards bodies coordinating activity across a discipline,
- increasing standardisation and automation of testing regimes, and
- increased acceptability of the contribution made by the model to the primary scientific discipline.

If we map our interview sketches onto the diagram we can locate the agent based modellers in the bottom left hand quadrant because their approach has a number of novel aspects. It relies on partial infrastructural arrangements which it is seeking to solidify, for example, through strengthening its relationships with data providers, accessing better compute resources and improved storage, advancing the core modelling approach, and engaging in community building work to establish standards of acceptability for results reporting. The particle physicists, with mature practices and infrastructure, can be seen as occupying the top right hand quadrant of the diagram. The stepped line represents the idea that at any one moment existing infrastructural arrangements constrain the modelling approach, and that shifts in those arrangements, whether this be in the form of improved compute resources, better validation data, improved software process (and so on) enables progress to be made more rapidly. The diagram suggests a normative picture of even progression along the X=Y line, indicated by the arrow marked (B). In real life the trajectory of a given modelling approach will more than likely deviate from this line, perhaps to different degrees and in differing directions over the course of its career. For example, the work of the data-based scientist has followed a trajectory indicated by the arrow marked (A). They have rapidly reached the point where they are able to publish results in the primary literature and have achieved this by taking advantage of pre-existing mature infrastructural components from which they assembled their solution.

Using this analytical framework we can understand the puzzle box of the apparent variations in effective delivery of scientific computing reported in the literature. For example:

"*In addition to the challenge posed by obtaining a reliable oracle, it is challenging to determine what constitutes sufficient validation. Sometimes the oracle used is too simplistic or not in the range in which the scientific software will actually be used. S4, who created her own oracle data, describes her dataset as "simple" and not representative of the data that it will eventually be used to process. Whether this simple simulated data validates the software adequately enough to justifiably increase her confidence in the model's applicability on complex inputs is questionable.*" (Kelly and Saunders, 2008)

If taking a normative view of how models are supposed to be validated, then it seems reasonable to be concerned that S4 is working with (what looks like) deficient validation data. However, in the context of our analytic framework this should not be seen as so unusual for a nascent modelling approach. Many modellers start out by using data that imperfectly matches the model's requirements, chiefly because there are no data suppliers creating data tailored to the their needs. Our findings suggest those working on 'early career' models appropriate data opportunistically from a range of sources (all of our agent based modellers worked in this way), none of which are yet geared to deliver data in the format, frequency and granularity that are ideal from the modellers' perspective. Both Edward's work and our interviewees highlight the importance of establishing *over time* strong sources of empirically based data, in a process that often takes many years to complete. A general lesson is that the modeller must not only advance their model, but also the infrastructural components that underpin their modelling work. This can be seen in the way our interviewees nurture relationships with data providers, and how they develop community practice and technologies. There are, however, also significant interdependencies between infrastructural components, meaning that any underdeveloped component can be rate-limiting and impeded the progress of the entire project. Interdependencies are demonstrated by this quote from one agent based modeller:

"Dynamic data assimilation - which is principally what I'm going to spend the next year looking at with any luck ... we're not really at that stage in terms, in terms of the data coming in but nonetheless that's kind-of what I'm interesting in."

Developing a "*dynamic data assimilation"* approach by itself will not improve the model's performance if access to improved empirical data is not also forthcoming. This implies that development of separate infrastructural components needs to keep pace with each other in order for the modelling science to progress effectively. This phenomenon resembles Hughes' notion of a 'reverse salient' whereby lag in the

development of a single system component can jeopardise its viability creating pressure for more fundamental reconfiguration (Hughes, 1987).

5 MODEL MATURITY SPECTRUM

The model we have outlined and the notion of maturity it encompasses is meant to be interpreted as a loosely fitting framework to help us think about modelling activities from an infrastructural perspective. In particular, we do not mean for our account to be a teleological one, whereby each modelling enterprise strives for some ideal 'mature state'. Our definition of maturity is a pragmatic and reflects simply the ability of the modelling approach to advance its primary scientific discipline implying that a mature infrastructure for one modelling approach might differ significantly in scope and ambition than for another. Putting it another way, not all modelling approaches grow up to be a climate science or particle physics equivalent. However, caveats aside, there are a number of important aspects of our case studies help support and elaborate our framework. Firstly, sophisticated science can be carried out with a minimum of innovation when important pre-existing infrastructural components can be easily appropriated, as for our data driven scientists. Secondly, regardless of scale, requirements for *similar* sorts of infrastructural components emerged as important for each of our case studies. Thirdly, mature and sophisticated modelling approaches such as those found in climate and ecology can act as a beacon and source of inspiration for emerging modelling approaches in other areas. Finally, but not least, that 'maturity' is not bounded. For example, those with 'mature' modelling approaches (in data driven science and particle physics) found aspects of their infrastructural arrangements unsatisfactory, and amenable to improvement.

6 CONCLUSIONS

We argue that advancing a modelling and simulation enabled science depends not only on local iterative refinements to the approach, but also on the ability to build strong interdisciplinary relationships and community spanning infrastructures that enable the flow of expertise and resources. Inspired by Edwards, we have developed an analytical framework that explains the maturation of scientific models and simulations in terms of the development of their supportive infrastructures, and predicts different trajectories for projects depending upon their ability to access or create appropriate infrastructural elements. Our framework allows us to harmonise findings from prior studies by explaining some instances of apparently worryingly immature practice as being a common feature of modelling approaches early in their career. Also, the presence of rate-limiting interdependencies suggests that development of a modelling approach can be hindered by the slower development or neglect of a single infrastructural component, represented by the trajectory labelled C in figure 1. Some examples of troubled projects described in the literature can be categorised as being of this type, where the development of testing regimes and software engineering practices has failed to keep pace with the development of other components of the modelling infrastructure. The distinction between projects that are merely immature and those that have a real deficiency is an important one, because helps us both to anchor our judgement about what is really problematic, and to guide how we might best intervene. Our pilot project was concerned to understand how scientific modellers can be better supported. The step-like character of advances in modelling science suggests that assistance should be tailored by identifying those infrastructural elements currently constraining progress, and by being sensitive to how the project's maturity will limit the sorts of assistance it is able to absorb.

References

Edwards, P. (2010) A Vast Machine. Computer models, Climate data, and the Politics of Global Warming. MIT Press.

Easterbrook, S. M. and Johns, T. C. (2009) Engineering the software for understanding climate change. Computing in Science and Engineering, 11(6) 65-74.

Faulk S., Loh, E. and Squires, S. (2009) Scientific Computing's Productivity Gridlock: How can Software Engineering help Computing in Science and Engineering? Computing in Science and Engineering, 11(6) 30-39.

Holzworth, D. P., Huth, N. I. and deVoil, P. G. (2011) Simple software processes and tests improve the reliability and usefulness of a model. Journal of Environmental Modelling and Software 24(4) 510-516.

Hughes, T. P. (1987) “The Evolution of Large Technological Systems,” In Wiebe E. Bijker, Hughes and Trevor J. Pinch (Eds.) The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology, Cambridge, MA, MIT Press.

Kelly, D. and Sauders, R. (2008) The challenge of testing scientific software. Conference of the Association of Software Testers, Toronto, Canada.