

International Workshop  
**Data-driven research in the biological and biomedical sciences**

**Reed Hall,  
University of Exeter**

Sponsored by the British Academy and the Economic and Social Research Council

**THURSDAY 15 APRIL 2010**

10:00 - 11:30 **Opening lecture:** 'Data-driven science: why and how'  
Professor Douglas Kell and Professor Tony Hey  
Chair: John Dupré  
*(Lecture hall Streatham Court C)*

*12:00 – 13:00 Lunch at Reed Hall*

13:00 – 13:15 Introduction to the workshop (Sabina Leonelli)

13:15 – 15:30 First session: **Experimentation**. Chair: Staffan Mueller-Wille

- 13:15 Ulrich Krohs: 'Convenience experimentation and the quest for scientific hypotheses'
- 13:45 Maureen O'Malley: 'Iterativity in the practice and philosophy of biological research'
- 14:15 Dick Burian: 'A Philosophical Perspective on Some Methodological and Epistemological Issues Raised by Recent Developments in Data-Driven Research'
- 14:45 *break*
- 15:00 Comments by Jane Calvert
- 15:15 General discussion

*15:30 – 16:00 Coffee Break*

16:00 – 18:00 Second session: **Cyberinfrastructure**. Chair: Annamaria Carusi

- 16:00 Sabina Leonelli: 'On the role of theory in data-driven research: the case of bio-ontologies'
- 16:30 Edna Suárez-Díaz: 'The place where similarity counts: sequence analysis meets taxonomy and evolution'
- 17:00 Comments by Peter Keating
- 17:15 General discussion

*19:30 Workshop Dinner at Strada*

## FRIDAY 16 APRIL 2010

9:00 – 9:30 *Coffee and Tea*

9:30 – 12:30 Third session: **Instruments and materials.** Chair: Mathias Grote

- 9:30 Bruno Strasser: ‘Data-driven research: Natural history in the age of experimentation?’
- 10:00 Staffan Mueller-Wille and Isabelle Charmantier: ‘Natural history and information overload: the case of Linnaeus’
- 10:30 Rachel Ankeny and Sabina Leonelli: ‘Re-thinking organisms: the impact of databases on model organism biology’
- 11:00 *Coffee break*
- 11:30 Comments by Dick Burian
- 11:45 General discussion

12:30 - 13:30 *Lunch at Reed Hall*

13:30 – 15:30 Fourth session: **Translation and clinical applications.** Chair: Jane Calvert

- 13:30 Alberto Cambrosio and Peter Keating: ‘Too many numbers: microarrays in clinical cancer research’
- 14:00 Susan Kelly: ‘Application driven research in biomedicine: dynamics of context, theory, evidence and interpretation’
- 14:30 General discussion

15:30 – 16:00 *Coffee and Tea*

16:00 – 17:00 **Final roundtable** introduced by John Dupré and Werner Callebaut

17:00 – 17:30 Discussion of future plans

## DIRECTIONS

The workshop will take place in Reed Hall at the University of Exeter. The only exception is the introductory lecture on Thursday morning, which will take place in the Streatham Court C lecture hall. The Streatham building is a ten minute walk to Reed Hall. Directions to Reed Hall, Streatham Court C and a map of the campus can be found here: <http://www.genomicsnetwork.ac.uk/egenis/aboutus/contactus/>

The workshop dinner will take place on the upper floor of the Italian restaurant ‘Strada’, located north of Exeter Cathedral in Princesshay Square (about 20 minutes walk downhill from Reed Hall).

## ABSTRACTS

**Rachel Ankeny (University of Adelaide) and Sabina Leonelli (University of Exeter)**

***Re-thinking organisms: the impact of databases on model organism biology***

Community databases have become crucial to the collection, ordering and retrieval of data gathered on model organisms. It is not yet clear, however, precisely how their use will shape the production of knowledge about organisms. This paper offers a qualitative analysis of the impact of community databases on research practices in model organism biology. The study is based on a comparison of the history and current use of four community databases, each of whom was set up to study a different organism: FlyBase, for *Drosophila melanogaster*; Mouse Genome Informatics, for *Mus musculus*; WormBase, for *Caenorhabditis elegans* and The Arabidopsis Information Resource [TAIR], for the plant *Arabidopsis thaliana*. For community databases to function efficiently, their curators need to set standards for what counts as reliable evidence, acceptable terminology, appropriate experimental set-ups, adequate materials (e.g. specimens) and research ethos. These choices affect the skills, practices and background knowledge required of the database users. We conclude that the increasing reliance on databases as vehicles to circulate data is having a major impact on the ways in which data are used as evidence, and consequently on how researchers understand the biology of model organisms and its relation to the biology of other species.

**Richard Burian (Virginia Tech)**

***A philosophical perspective on some methodological and epistemological issues raised by recent developments in data-driven research***

The post-genomic development of major databases and tools for handling them has led to surprising developments of considerable interest in a number of disciplines and interdisciplinary research programs that fall under the general umbrella of molecular biology. These developments are of considerable philosophical interest, especially for methodology, epistemology, and issues concerning the centrality of what I sometimes call high-level theory. Roughly speaking, data-driven research (DDR) proceeds in part by using methods for hypothesis generation ill-understood by philosophers that are also capable of testing some of the hypotheses thus generated. Examples of this sort are provided by the 'generation' or 'forcing' of phenomena exemplified nicely by data-driven research using microRNAs and RNA interference. The main point, however, is that DDR forces the development of an integrative methodology based on a plurality of methods, a plurality of means for generating and testing hypotheses, and indirect interactions with the underlying theories prevalent in, or employed by, molecular biology. The result is a methodology based on iterative procedures for generating hypotheses, relating them to available theoretical and empirical knowledge, testing and revising them, generating new hypotheses, etc. These methodological procedures require theoretical pluralism, a position with strong implications for epistemology and for our understanding of the relations of theory to hypothesis generation and testing. While there cannot be full-fledged autonomy of data-driven research from high theory, there is, I shall argue, autonomy from any specific theory bearing directly on the subject matter and the objects under investigation.

**Alberto Cambrosio (McGill University) and Peter Keating (Université du Québec à Montréal)**

***Too many numbers: microarrays in clinical cancer research***

When DNA microarray technology first appeared at the beginning of the 1990s, the possibilities for data generation for biomedical and clinical purposes seemed endless. Gene expression profiling in particular offered the possibility of comparing normal and

pathological genomes and thus to isolate the bad actants in, for example, cancer. As it is often the case with new biomedical technologies, however, the introduction of microarrays into the clinic as diagnostic and prognostic devices has not proceeded as smoothly as expected. Three main obstacles have confronted microarray developers all of which revolve around the fact that microarrays produce too much data and thus threaten to overwhelm traditional practices. Clinicians complain that, while a “nice idea”, the amount and kind of raw data produced by a single microarray precludes their routine use; companies should be forced to supply a statistician with every microarray they produce. Statisticians quickly realized that the amount of data generated by a single patient swamped common notions of significance and undercut a number of statistical assumptions that underlay standard data management techniques. In recognition of the fact that microarrays did not conform to common statistical practices, in 1997 the NCI set up a “Molecular Statistics” subunit to deal with the problem. In the decade that followed, together with clinicians and manufacturers, researchers at the subunit created a number of standards for the field including a common statistical package. Finally, despite a few exceptions (most notably the *Mammaprint* gene expression profiling signature for breast cancer, which has been cleared by the FDA), microarray technology has so far not attained the status of a routine diagnostic tool within the field of cancer. This is partly the result of the third and final obstacle: all those numbers have to displace the visual signs used by pathologists in diagnosis and prognosis. Microarrays may be as good as the visual techniques used by histopathologists, but are they better? Current clinical trials seek to answer this question. While reaching an answer might be more difficult than first suspected, the use of microarrays as tools of prediction rather than prognostication or diagnosis has, for the moment, breathed new life into the microarray enterprise. This paper will examine the constitution of the obstacles above and how they have been circumvented, denied and confronted.

**Tony Hey (Microsoft)**

***Data-intensive scientific discovery: The fourth paradigm***

There is a sea change happening in academic research -- a transformation caused by a data deluge that is affecting all disciplines. Modern science increasingly relies on integrated information technologies and computation to collect, process, and analyze complex data. It was Ken Wilson, Nobel Prize winner in physics, who coined the phrase “Third Paradigm” to refer to computational science and the need for computational researchers to know about algorithms, numerical methods, and parallel architectures.

However, the skills needed for manipulating, visualizing, managing, and, finally, conserving and archiving scientific data are very different. “The Fourth Paradigm” is all about data and the computational systems needed to manipulate, visualize, and manage large amounts of scientific data. A wide variety of scientists—biologists, chemists, physicists, astronomers, engineers – require tools, technologies, and platforms that seamlessly integrate into standard scientific methodologies and processes.

This talk will illustrate the far-reaching changes that this new paradigm will have on scientific discovery.

**Douglas Kell (The University of Manchester and BBSRC)**

***Data-driven science: Why and how***

Much of science, and especially the biology of the 20th Century, has been based on the primacy of hypothetico-deductive reasoning, in which one starts with a hypothesis or idea and performs an experiment whose results (data) are seen as either consistent or otherwise with the predictions of the hypothesis. This kind of view was emphasised by Karl Popper (and more readable commentators such as Peter Medawar). (Where the hypotheses were

supposed to come from was not discussed.) By contrast, inductive modes of reasoning begin with the data and find the hypotheses that best fit those data. These different approaches bear similarities to the distinction to be made between (Neyman- Pearson) statistics and machine learning. In reality, these inductive and deductive phases form an iterative 'cycle of knowledge' (Kell DB, Oliver SG: Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* 2004; 26:99-105; Westerhoff HV, Kell DB: The methodologies of systems biology. In Boogerd FC, Bruggeman FJ, Hofmeyr J-HS, Westerhoff HV (eds.): *Systems biology: philosophical foundations*. Amsterdam: Elsevier, 2007:23-70), and there is an increasing epistemological trend from hypothesis-dependent to data-driven science. As the scientific literature (Hull D, Pettifer SR, Kell DB: Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput Biol* 2008; 4:e1000204) and necessary datasets become ever larger, and available to all online, we need new kinds of approaches - including for data-intensive science - to optimise our analyses of the two arcs of the cycle of knowledge.

**Susan Kelly (University of Exeter)**

***Application driven research in biomedicine: dynamics of context, theory, evidence and interpretation***

While there is much discussion about both data and hypothesis driven biological research, much research within biomedicine is driven by specific application needs or contexts. Application contexts are existing or envisioned complex socio-technical arenas that include standards of evidence, reliability, analytic equipment and skills, information practices, and ethical and normative frameworks. In this talk I focus on research in which the production of knowledge about biological objects is shaped not by hypotheses but by the 'work' those objects are intended to do within specified application contexts - efforts to exploit findings of fetal cells in maternal circulation for developing techniques of non-invasive prenatal diagnosis of Down Syndrome. The research has proceeded along two intersecting lines – the production of evidence of physiological phenomena upon which technological applications can be developed, and the simultaneous production of evidence of an optimal configuration of methods around which technological applications can be developed. I take these dynamics to raise questions about the nature of evidence, data, theory and interpretation. Further, for data to translate into clinical applications, experimental contexts must reflect features of application contexts, in terms of what counts as 'reliable' with regard to evidence, materials, clinical practices and equipment.

**Ulrich Krohs (University of Bielefeld)**

***Convenience experimentation and the quest for scientific hypotheses***

Systems biology aims at explaining life processes by means of detailed models of molecular networks, mainly on the whole-cell scale. The whole cell perspective distinguishes the new field of systems biology from earlier approaches within molecular cell biology. The shift was made possible by the high throughput methods that were developed for the collection of 'omic' (genomic, proteomic, etc.) data. These new techniques, however, have also induced a change with regard to the modeling strategies that are applied to gain biological insights out of experiments. My paper analyzes the altered roles that experiment and hypotheses play in systems biology with respect to model building and draws some conclusions about epistemic consequences of the change in modeling strategies.

In times when experimentation was automated only to a lesser degree, modeling and experimentation used to be driven by hypotheses. While experimental feasibility was a serious constraint to data collection, there were always enough interesting hypotheses to test.

In systems biology, the situation is dramatically different. While data collection is easy now (though not cheap), it seems difficult to find biologically relevant hypotheses that could guide modeling. Model building, consequently, is driven to a high degree by the vast amount of data that is produced by convenience experimentation, i.e., by commercially available semi-automatic analytic equipment. Data-drivenness has several epistemologically relevant effects. I will focus on the consequences of the following two effects: First, modeling is less theory laden than it used to be. Second, data sets and therefore the models are in part determined by the available equipment (DNA probe arrays etc.). I will demonstrate that the resulting models shift the explanatory focus of biology from a concept of the living entity as an organized whole to a concept of the cell as a dynamic system. While models following this new focus do have their value in unraveling the mechanisms that underlie cellular dynamics, it turns out that they alone cannot answer genuine biological questions about molecular and cellular physiology.

**Sabina Leonelli (University of Exeter)**

***On the Role of Theory in Data-Driven Research: The Case of Bio-Ontologies***

The availability of data on an unprecedented scale, the global increase in research projects and funding, the insistence on systemic approaches and the growing reliance on bioinformatics are having an impact on knowledge-making practices in biology. As a first step towards clarifying the nature of that impact, I focus on the case of bio-ontologies, classification tools that have recently become crucial to the organisation and sharing of results across research contexts, by enabling biologists to store and retrieve data on the basis of their own research interests. I argue that bio-ontologies play the role of theories in data-centred research modes. They express the knowledge that is relied upon when analysing data for the purpose of discovery. At the same time, they can be challenged and modified depending on shifts in research contexts. They thus represent the theoretical framework underpinning the extraction of inferences from high-throughput data: a set of hypotheses about biological phenomena, which is (1) understood in reference to the research context(s) in which phenomena are experimentally examined, (2) constantly tested against new evidence and (3) adapted to scientific developments.

**Staffan Mueller-Wille and Isabelle Charmantier (University of Exeter)**

***Natural History and Information Overload: The Case of Linnaeus***

Natural History can be seen as the paradigmatic discipline engaged in “data-driven research”. Historians of early modern science have begun to emphasize its crucial role in the Scientific Revolution (Harold Cook), and some historians of present day genomics see it engaged in a return to natural history practices (Bruno Strasser). A key concept developed by some historians to understand the dynamics of natural history is that of an “information overload” (Brian Ogilvie, Ann Blair). Taxonomic systems, rules of nomenclature, technical terminologies and even theories of evolution were developed in botany and zoology to catch up with the ever increasing amount of information on hitherto unknown plant and animal species. In our contribution, we want to problematize this notion. After all, the same people who were driven by, *also drove* the production of information. In order to understand this complex relationship, we will turn to Linnaeus, who used his publications to create a veritable data production machinery.

**Maureen O'Malley (University of Exeter)**

***Iterativity in the practice and philosophy of biological research***

A growing number of philosophers have recognized the need to account for the complex and highly variable nature of scientific inquiry. This paper will suggest that the concept of

iterativity could play a fruitful role in that endeavour, especially in the era of burgeoning genomic sequence data and attempts to integrate it into systems-based modelling. Iterativity in the context of scientific practice refers to a repeated series of activities – a series that comes back repeatedly but not necessarily repetitiously to a particular sphere of inquiry. One of the few explicit treatments of iterativity in the history and philosophy of science is in the form of ‘epistemic iteration’, a term devised by philosopher of science Hasok Chang in his 2004 book, *Inventing Temperature*. The most notable scientific area in which such discussions are occurring is today’s systems biology, where iterativity has become a virtue, an aim, and even a manifesto of inquiry. It is epistemic, in Chang’s sense, but also methodological, in that it specifies a cycle or even barrage of methods that lead to improved epistemic outcomes as the researcher moves from one phase of inquiry to the next. Within this context, hypothesis testing can be regarded as a refined form of questioning that occurs in appropriately narrowed (and transient) nexuses of inquiry that are created by the iterative interplay of multiple modes of investigation. These investigative practices include not only the proposal and testing of hypotheses, but also exploratory, technology-oriented, and question-driven modes of research. Also relevant to any discussion of iterativity is the notion of bootstrapping, which is a metaphor to describe the way in which successful investigations can be founded on incorrect assumptions and inadequate models. The iterativity of investigative approaches plays a key role in such ‘corrective evolution’ as research is pushed onwards through the interplay of different strategies of inquiry. I will outline two general research programmes that demonstrate iterative movement between multiple modes of investigation. These cases are drawn from metagenomics, the study of molecules of uncultured communities of organisms, and synthetic biology, the effort to engineer living organisms at the molecular level.

**Bruno Strasser (Yale)**

***Data-Driven Research: Natural History in the Age of Experimentation?***

The perception that data-driven research represents a new episode in the history of science tells us more about our current assumptions regarding the nature of scientific research than it does about the actual development of science. If data-driven research is understood as a way of knowing based on the collection, comparison, and computation of data which has not been produced to test a particular theoretical conjecture, then its historical origins predates even the scientific revolution. Indeed, from the Renaissance to the present day, one science has been fundamentally “data-driven”: natural history. The collection, comparison, and computation (not necessarily with computers) of fossil, plants, and animals by naturalists who participated in systematic surveys was just as much “data-driven” as systems biology, for example, is today. It is only because the experimental sciences have taken the upper hand over natural history in the late nineteenth century and come to dominate the public perception of science that data-driven research is perceived as a novel feature of twenty-first century science.

Yet, there is perhaps something new after all about current data-driven research. To be sure, computers and the internet, so important in contemporary data-driven research, were not in wide use in early modern natural history. But that difference is rather superficial, as the comparison and distribution of natural facts was simply practiced through other means. What is more fundamental, however, is the fact that contemporary “data-driven research” constitutes a natural historical way of knowing applied to *experimental* data. The experimental and the natural historical traditions which had grown increasingly separate have recently become part of a new hybrid culture, of which data-driven research is an illustration. Understanding the dual historical origins and actual hybrid constitution of data-driven research illuminates many of the current debates around the value of this way of knowing

which result from tensions between the epistemic, social, and cultural norms and values of these two traditions.

To explore these themes, this paper will focus on the history of the Protein Data Bank (PDB), created at Brookhaven National Laboratory in 1973. This electronic collection of three-dimensional coordinates of protein structures has since then become a major tool for data-driven research. Initially, the PDB contained only coordinates of structures determined experimentally, by x-ray crystallography. By the late 1970s, however, it began to include structures which had been elucidated through “comparative modelling”, i.e. the fitting of an amino acid sequence on the structure of a homologous protein. The PDB thus contained empirically determined structures along with theoretically derived structures prompting debates about the value of empirical and data-driven methods. A different, but related, set of issues were at stake, when the organisers of the PDB attempted to bring crystallographers to make their data public, revealing profound disagreement as to what exactly constituted data, results, and interpretations. This issue was not only epistemological, but also practical, as it determined what research outcomes should be made public upon publication and what could remain private. Finally, when researchers began to establish taxonomies of protein structures based on the comparison of the PDB’s entries, they were torn between the experimental sciences’ reliance on objectivity and quantification and natural history’s emphasis on subjective and visual methods. These debates are much less surprising, and make much more sense, if we consider that data-driven research, such as that carried out with the PDB, is not a new way of knowing, but a new hybrid between experimentalism and natural history.

**Edna Suárez-Díaz (Universidad Autónoma México)**

***The place where similarity counts: sequence analysis meets taxonomy and evolution***

At the beginning of the 1980s, as molecular sequences started to accumulate, the need for data bases and sequence-analysis tools became evident at many places. While a US-national nucleic acid sequence data-base was created in 1982 at Los Alamos National Laboratory at New Mexico (GenBank), the funds for the creation of a center for sequence analysis were not granted. Sequence analysis tools, however, were created by members of Walter Goad’s group at Los Alamos (Smith and Waterman 1981) following the steps of molecular evolutionists that had attempted quantitative-computational analysis since the mid-1960s (Fitch and Margoliash 1967, Needleman and Wusch 1970).

Measures of similarity, and their counterpart, distance, have played a crucial role in the development of these tools since the early times of molecular evolution. Tools for sequence analysis may search for local and global similarities, for fast- but rough *patterns* of similarity, and for similarity or distance that may reveal ancestry –and functional- relations among organisms. Although the statistical analysis needed for searches in huge data bases has been driving the developments in mathematics, genomics and computational biology in the last three decades, theoretical (evolutionary) assumptions have not been abandoned. On the contrary, one of the visible trends –at least in some areas of bioinformatics and very clearly in molecular phylogenetics- has been to incorporate theoretical assumptions and explicit criteria and models of molecular evolution in sequence analysis tools (for instance, as in the ‘weighting of gaps’ and in models to calculate ‘multiple-hits’, see Suárez and Naya 2008).