Dealing with Data

NeuroHub – case study

Professor Anne E Trefethen

University of Oxford
...the imminent flood of scientific data expected from the next generation of experiments, simulations, sensors and satellites.

Deluge!!!

Courtesy of Christine Borgman
People generating data

• Everyday life has a digital footprint
  – Smart cards, supermarket loyalty cards
  – Sharing photos, videos and ideas...
  – Homes instrumented e.g. smart electricity meters
  – Streets sensing society in motion; e.g. car recognition

By May 2012 more than 48 hours of video uploaded per minute
3 billion views a day

By 2019 50 million electricity meters
THE GEOSOCIAL UNIVERSE

REMEMBER WHEN FRIENDSTER DOMINATED THE U.S. SOCIAL SPACE? LIKE THE UNIVERSE, THE GEOSOCIAL LANDSCAPE IS CONSTANTLY CHANGING. WHILE NEW STARS ARE BEING BORN, BLACK HOLES ARE ALSO DEVELOPING.

WE LOOKED AT THE CURRENT SIZE OF THE MAJOR SOCIAL NETWORKS AND OVERLAI THEIR CURRENT MOBILE USER BASE.

5.3 BILLION MOBILE DEVICES

Sources: TechCrunch | SocialMediaToday | Facebook | Wikipedia | Mashable | SocialSystem | Daily Mail | LinkedIn | Loop | SearchEngineLand | Brightkite | SocialTimes | BadBoy | MobiThinking

MAY 2011
The map shows 16,908,630 entries distributed over 13,116 locations.

November 25th, 2009
The map shows 18,710,024 entries distributed over 62,882 locations.
European Bioinformatics Institute (EBI)
- Hosts European bioinformatics data
- Teams of scientists and curators
- European Nucleotide Archive (ENA) accumulates 5M bases per second
Total number of scans per thousand population = 2259
Medical imaging data

Total amount of data = 68,700,00 Gbytes approx 1M DVDs
Entire E-Science Cycle
Encompassing experimentation, analysis, publication, research, learning

(eBank slides courtesy of Liz Lyon)
The social process of Science 1.0 2.0

Digital Libraries
- Reprints
- Peer-Reviewed Journal & Conference Papers

Local Web
- Technical Reports

Repositories
- Preprints & Metadata

Virtual Learning Environment

Undergraduate Students

Next Generation Researchers

Data, Metadata, Provenance, Scripts, Workflows, Services, Ontologies, Blogs, ...

Graduate Students

scientists

Technical Reports

experimentation

Certified Experimental Results & Analyses
Interacting with data
Data at the heart of understanding

- Datascopes
- Data observatories
- Social observatories
- Social machines
- Knowledge accelerators
Here is the evidence, now what is the hypothesis?
The complementary roles of inductive and hypothesis-driven science in the post-genomic era

Douglas B. Kell¹* and Stephen G. Oliver²

Summary
It is considered in some quarters that hypothesis-driven methods are the only valuable, reliable means of scientific advance. Data-driven advances in scientific knowledge are marginal, irrelevant, insecure or wrong. The development of technology—which ‘hypothesis-led’ (beyond the recognition might be of value)—must be seen as equally the hypothetico-deductive scientific agenda that data- and technology-driven paradigms do not. Hypothesis-led studies of knowledge discovery are complementarities partners with them. Many fields among hypothesis-poor. Here, computational methods of analysis, which may be automated, provide generating novel hypotheses, especially in the post-genomic era. BioEssays 26:99–105, 2004 © 2003 Wiley Periodicals, Inc.
Now we have the data

The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to reorganize it
- How to share with others

- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation
Infrastructure for data?

Repositories

Clouds

Tools

ontologies

compute

standards

Methods

and algorithms

Skilled people

Data Acquisition and Modeling

Collaboration and Visualization

Analysis and Data Mining

Disseminate and Share

Archiving and Preservation
A Research Information Environment for Neuroscientists
NeuroHub Project

• JISC funded project

• Collaboration between Oxford, Southampton and Reading Universities - 4 FTEs, 3 years

• Advisory group, Mark Ellisman UCSD, Jim Austin York, Paul Watson Newcastle

Aim to work with neuroscientists at each site to develop a system to allow them to easily share data throughout lifecycle.
How the project came to be...

And it all came about because of British Rail..........
NeuroHub Team
Professor Mark Baker

Why the young learn more easily

Specific issues such as the study of the molecular basis of synapse formation, plasticity and the regulation of neuronal morphology in the normal and diseased brain. Examining the mechanisms of an activity-dependent form of neural plasticity known as long-term potentiation (LTP). [1]

Studying nerve cell activity - the basis of learning and memory - in rats. Indicated that younger brains may learn things more easily, but older brains may store information more efficiently.

Why the young learn more easily

Highly detailed laser imaging, which looked at images one micron wide - a 100th the width of a human hair, to look at how synapses behave.

A neuron - there are about 10,000 separate synapses on this cell

High resolution cell fluorescent imaging and electrophysiology key technologies to large amounts of raw data from which others might also gain insight

Leading to experimental data, images, electrophysiology data

http://news.bbc.co.uk/1/hi/health/6172048.stm
Science questions...

Focused on integrative analysis of brain function/dysfunction. The group utilizes different invertebrate model systems to understand particular features of the organization of neuronal networks and how different types of constituent interneurones contribute to the processing of sensory signals [2].

How do insects walk?

Phillip Newland

How the nervous system produces limb movements in insects?

Few neurones achieve a high degree of versatility of movement

• how the nervous system performs the complex integrative task of processing different sensory parameters?;

• are there specialist neurones or populations of neurones that process different signals (high frequency versus low frequency inputs, or velocity versus position or acceleration)?

• what do the different branches of a neurone contribute to integration compared to the whole cell?
Unlike us, locusts have taste receptors all over their bodies and legs – they can taste what they stand on! 

Taste receptors on the legs of locusts imaged using a scanning electron microscope

Courtesy of Phil Newland
Intracellular recordings from motor neurones responding to tastants. Physiological data sampled using a Cambridge Electronics Design A/D interface and stored using Spike 2 software.

Courtesy of Phil Newland
Epithelial cells from locusts stain positively for nitric oxide synthase activity (L) and nitric oxide release (R). Images taken using Nikon E-800 digital microscope using 12-bit high resolution camera and Metamorph imaging software.
The effect of acute ethanol on *C. elegans*

Worms with a simple CNS 302 neurons. Locomote in environment by bending. When drunk less bendy and video of worms allow this to be quantified. Neural substrate of behaviour can be quantified.

Courtesy of Vince O’Connor
Research at CINN focuses on physiological and psychological mechanisms underpinning complex cognitive behaviours, targeting typical and atypical development and decline in individuals.

• Facilities
  — 3T MRI Scanner, High-density electroencephalogram (EEG) laboratory, meeting rooms and more ...

• Human data
  — Confidentiality and security issues.

• A strong research group working on issues in signal analysis and computational modelling.
What is required...

Experiments
Analysis
Publication
Dissemination
Experiments

Lasersharp 2000
• 2 photon imaging
• Confocal imaging
• Photolysis

Andor IQ
• TIRF
• QD imaging
• Photolysis

Fortran compiler
JAVA
• Modelling

WinWCP/Strathclyde
• Electrophysiology

Courtesy of Nigel Emptage
Adobe
- Acrobat
- Photoshop
- Illustrator

Publication

Microsoft
- Word
- Excel

Courtesy of Nigel Emptage
More Specifically...  
...Help Managing Data

Initial Experimental Idea

Publication

Experimental Design

Data Collection

Analysis

What are the Data?
Our Approach

• To gain insight into what Neuroscientists do and what they need:
  – Embed developers in the neuroscience labs.
  – Complete structured and unstructured interviews
• Introduce existing standards where they exist
• Use agile development methods
• Investigate further related projects to determine if they can be applied
• Keep the neuroscientists at the centre
Challenges

• Interdisciplinary teams – different expectations, cultures, requirements,
• Agreed standards
  – Data formats
    ➢ Microscopes (Multiphoton or Confocal)
    ➢ Live cell fluorescent imaging
    ➢ Electrophysiology recordings
  – Meta data standards
• Complexity of tools used in community
• Ability to share images, data, analysis
• Network!
User Requirements (1)

NeuroHub needs to
• manage the data throughout the experimental lifecycle, while interfacing with their tools
• be usable by researchers ranging from undergraduate project students to group leaders
• allow users to easily share the research activities and outputs with their collaborators.
• provide effective support for their research tools to annotate images and videos as they are being recorded, while making them available to other group members and collaborators.
• provide a search mechanism for images and videos, along with any associated metadata, from current and past projects is extremely desirable.
User Requirements (2)

• A consistent annotation method for data archiving

• Web based repository for data

• Many file formats to be supported
  • A searchable repository for images
  • A searchable repository for video images
  • A document share tool for ‘live’ manuscript editing

• File space for literature sharing (PDFs)

• Blog interface

• Ability to securely collaborate with international groups
Related Projects

Reuse reuse reuse – sustainability

Integrate integrate integrate integrate – not monolithic

Relatively small project – needs to leverage
• Making better use of computing technology
  – SmartTea investigated how chemists experiment
  – Electronic Lab Notebooks (ELNs)
    • Useful in some scenarios
    • Chemists involved still use their logbooks

• Manages outputs from experiment devices
  • Automatically from electronic devices
  • Otherwise manually (scanned / typed in)

• Creates a visual and easily readable record of experiments
CARMEN: Introduction

- 4 year project funded via the UK EPSRC research council
- Running September 2006 – September 2010
- £4.5 million funding, across consortium of 12 academic institutions, plus industrial partners
- Funded by UK e-Science programme, not medical or biological research councils, as focus is on development of informatics infrastructure to enhance neuroscience collaboration

Courtesy Jim Austin
Electrophysiology Data

- raw voltage signal data collected by patch-clamp and single & multi-electrode array recording
- novel optical recording, particularly the activity dynamics of large networks
- resolving the ‘neural code’ from the timing of action potential activity

Courtesy Jim Austin
• Focussed around Neurophysiology
• Virtual Laboratory Environment through a Portal for Sharing Data and Programs
  – Annotate uploaded files
  – Search across archived datasets
  – Run analysis over datasets using programs on CARMEN as well as their computing resources
• Promotes Standards
  – MINI: Minimum Information about a Neuroscience Investigation
  – NDTF: Neurophysiology Data Translation Format
How does this impact on NeuroHub?

- CombeChem gives insight into what a scientist might find useful
- Blog3 / VRE-CI / Research Objects may provide a basis to help manage data as it moves around the lab
- myExperiment provides facility for social curation of data
- CARMEN provides standards for consistent data annotation
Approach for V1.0

1. Get the network fixed
2. Build on established tools and web 2.0 technologies
3. Establish best practice across labs for data formats
4. Create repository
5. Use myexperiment and Blog3 for capture in lab and sharing
6. Provide suite of translation tools
7. Iterate!
myExperiment for supporting interaction within the community

Built – it didn’t work – we threw it away

Blog3 or similar with additional functionality on the group systems

Community

Meta data generation

Tools on PCs for converting files

Data Capture Device

PC

Data Capture Device

PC

Analysis Server

ORA

Oxford

Southampton

Soton EPrints
NeuroHub 2.0

Moved to a Drupal Commons base

Built against user scenarios
Development practice

• Continual and frequent feedback
  – Over 100 user scenarios created and prioritised over time

• Acceptance test derived from user stories

• Continuous integration and testing
  – Using buildbot
  – 3 builders (quick, standard and cloud

• Release management – NeuroHub-dev
  – Issue Tracking

Configuration details and current build status:
http://neurohub-bot.oerc.ox.ac.uk/buildbot/index.html
NeuroHub Release 2.0 under the hood

- Drupal – Frontend content management. Based on Drupal Commons, Groups
- Alfresco – Backend data management. Modified Alfresco module,
- Apache Solr – Search engine,
- Apache Tika – Metadata extraction toolkit for documents,
- Google services – Docs and Calendar
- Mendeley integration for PDFs
- Cloud-based computation using GPU’s,
- NCBO ontology-based tagging,
- LDAP – Single sign-on
- Digital Pens – Used for recording experiments,
- XML-RPC desktop client – uploading and generating content.
NeuroHub Architectural Overview

Search engine
Metadata extraction toolkit for documents
Frontend content management

Backend data management
National Center for Biomedical Ontology
Welcome to NeuroHub

Welcome to your NeuroHub Commons installation. This is a community version of NeuroHub.

Happening right now in NeuroHub

Neurohub Neuroscience Group

There are no recent items for this group

Our Community

Jumpstarting our community
0 replies
2 days 10 hours ago
admin

Latest featured content

Jumpstarting our community new
Posted: 9 Sep 2012 - 07:50
by admin

Most active groups

Neurohub Neuroscience Group
2 member(s)
0 post(s)

Our Community
1 member(s)
1 post(s)

Latest notices

No notices have been created yet.

Popular tags
NeuroHub - Alfresco

Includes webdav interface to user space
The Oxford e-Research Centre, in collaboration with the Universities of Reading and Southampton was awarded funding by JISC for a project in the A2 Strand - Developing e-infrastructure to support research disciplines. Neurohub project is developing a set of sustainable tools and framework that will allow neuroscientists to efficiently and effectively use existing e-infrastructure and by doing so will enable a more productive research cycle, streamlining the laboratory experience from conception of experiment to publication of the research results. Funded by JISC for a project in the A2 Strand - Developing e-infrastructure to support research disciplines.
Adopted MINI standards from Carmen
- Taxonomy manager built on NCBO tagging tool
- RDF with various ontologies and capability to use SPARQL
Groups

Featured groups

Our Community
Tags: cinn, community
Public information about CINN’s research community

Create a group

Your groups

- Agraphia
- Computational tools and techniques
  1 new post(s)
- EEG / EEG-fMRI
  2 new post(s)
- MRI
  6 new post(s)
- MRS

Most active groups

- Our Community
  37 member(s)
  24 post(s)
- MRI
  22 members
experiences, knowledge and best practices with the aim of reducing those frustrating moments when you don’t quite know how to tell the computer what you want it to do! This is an open group and everyone is invited to participate.

analysis tools, cloud, cluster, code repositories, Condor, CUDA, distributed systems, FSL, GPGPU, network, OpenCL, parallel programming, programming help, recipes, scripting, software libraries, workflow

Featured
Checking this option will make this a Featured group. Featured group's get special placement on the home page. This should only be used for groups with a special purpose.

Shout box

Shouts automatically update every 30 seconds

Latest group activity

knaish joined the group Computational tools and techniques

jason replied on MatLab 2012b:

scratch that the install is

jason has updated Blog entry "MatLab 2012b"

Latest discussions

Preparing brain pictures for participants

Posted: 14 May 2012 - 14:09 by plia
6 comment(s)

Latest blog posts
In 2009, CINN acquired a Siemens TRIO MR scanner suite, which is dedicated to the research activities of groups across all Departments and Schools of the University of Reading.

The suite contains: a whole-body MR scanner, a range of coils (up to 32 channels), including specialist coils for clinical and research use, the Syngo software with a license for developing our own MR sequences, which contains sequences for cardiac, perfusion, BOLD, DTI, angiography and MRS, pre-post testing cubicles, changing and restroom facilities.

Along side these facilities are available two systems to record physiological signals in the scanner, bi-ocular visual stimulation (goggles), and an MR-compatible EEG system (see Neurohub EEG group for more information).

This Neurohub group centralises the information related the MR facility in CINN: information related to safety matters, the details of the procedures in place to getting access to the scanner to run your own studies, links to the booking system, access to the MR helpdesk and the accompanying wiki. If you have any question, or can not find the information you are looking for, do not hesitate to ask!

More on the Siemens TRIO MR scanner suite: http://tinyurl.com/siemens trio

See the installation of the scanner on Youtube: http://www.youtube.com/watch?v=jH6Uf_ZczsQ

*MRI, imaging technologies, MRI*
Run times for analysis using pre revision 5e13d7608d31 code

We encountered some issues with the CINN cluster while processing the sc4 permutation test 10, 500 experiment using the pre revision 5e13d7608d31 (see bitbucket) code. These issues were due to changes to the storage systems used across the cluster that resulted in incorrect configuration for some of the cluster nodes. As a result, we performed a higher number of runs than expected and had to piece together the data list each time to ensure that we did not unduly repeat work. Time was also spent identifying if failed results were due to problems with the cluster, or problems with the data itself. As part of this process, the issues with the cluster were identified and resolved.

The total CPU time for processing the Thousand Brains and Cambridge datasets at 10, 500 (experiment 1 from the sc4 permutation tests) was 7,015 hours (approx 292 days), which includes the CPU hours spent re-computing failed jobs.

thousand brains time_window=10, nshuffles=500
- Job submission: SC4_permutation_test thousand brains time_window=15, nshuffles=500
- Job submission: SC4_permutation_test thousand brains: t_window=5, nshuffles=500
- Job submission: SC4_permutation_test thousand realigned_TAMPERED brains time_window=10, nshuffles=500
- Job submission: SC4_permutation_test thousand realigned_TAMPERED brains time_window=5, nshuffles=500
- Plots from the sc4_permutation tests across thousand brains and cambridge data
- Summary of SC4 permutation test cumulative runtimes

Latest log books

- Thousands Brains Dataset
  Posted: 24 Aug 2012 - 08:46 by garry
  0 comment(s)

- SC4: Permutation test
  Posted: 9 Aug 2012 - 08:55 by garry
Before we could use the thousand brains data set for our experiments, we had to normalise the data. The normalised data set is located on the CINN cluster at:

A Web form was created to elicit expert help in identifying if any of the fMRI data from the thousand brains dataset had not been normalised correctly.

The Web form containing all the brains is located at: http://www.acet.rdg.ac.uk/~gms/thousandbrains

**UPDATE (24/08/12):** In addition, a web form that omits all the brains that have been identified as bad, so far, is located at http://www.acet.rdg.ac.uk/~gms/thousandbrains-with-bad-ones-removed. This is intended to be easier on the eyes as brains that have already been identified as bad are not included in the form output.

The form is split into each of the geographic regions present in the raw thousand brains dataset; with each region-specific part of the form containing an image of the brain alignment, as well as a check box that is used to identify that the brain has not been normalised (aligned) correctly.

An example image for a single brain:
Integrating with NeuroCloud

Workflows using Matlab Engine are run on Amazon cloud and report directly into lab book (using XML-RPC, OAuth)
Forecast: Cloudy with a Chance of Semantics

Future Research Infrastructure will use Client + Cloud resources

- visualization and analysis services
- scholarly communications
- blogs & social networking
- search books citations
- instant messaging
- identity
- notification
- document store
- storage/data services
- knowledge management
- knowledge discovery
- compute services virtualization

Reference management
Project management

Courtesy of Tony Hey
Present Site Usage

• Oxford:
  – Helps Nigel see what researchers are doing....
  – Have tested different input methods including digital pens, iPad and tablet PC.
  – Uploads of their 6000 image files is working fine, with ~250GB of data stored.
• Reading:
  – Adopted by CINN
  – Want to use NeuroHub as a frontend for their whole centre
  – Workflow development has improved scientists working pattern.
• Southampton:
  – Have ~100GB of data stored and have started to organise their files.
  – Interested in digital pen input and want to continue uploading their files in NeuroHub.
Conclusions

• Succeeded at capturing the research processes of the neuroscientists and supporting them.

• Framework used for development has allowed an environment that is adaptable and easily grown to add further features as required – allows personalisation to suit the needs of a given laboratory.

• Platform suitable for other scientific areas and is being reused in other projects
Conclusions (2)

- Southampton scientists collaborate more effectively with groups in Brazil and monitor their researchers at home, whilst in Japan.
- Oxford scientists to access to data on the move and using mobile and tablet devices as well as using digital pens as a way to capture experimental data.
- CINN scientists are also using NeuroHub to communicate with collaborators in Australia.
- Being supported on data analysis by providing mechanisms for using of scientific workflows and cloud services.
- CINN has stopped using mailing lists within the centre and use NeuroHub to collaborate with other departments across the university.
As neuroscientists do we simply require a repository for data that can be accessed via the internet, and that may satisfy a small cohort of scientists? If so then there are many solutions available, but if we want to go further to establish a virtual research environment that can be used globally, to allow us to work within an accessible research environment that integrates across the many fields within the discipline (molecular, genetic, physiological, anatomical, behavioural) then we need to look for something new.

Further information

Welcome To NeuroHub

The Oxford e-Research Centre, in collaboration with the Universities of Reading and Southampton was awarded funding by JISC for a project in the A2 Strand - Developing e-infrastructure to support research disciplines. Neurohub project is developing a set of sustainable tools and framework that will allow neuroscientists to efficiently and effectively use...
More information about NeuroHub and links to our test instances can be found at [http://www.neurohub.ac.uk/](http://www.neurohub.ac.uk/)

Installation designed to take minutes

Fully documented installation and user guide

Final release end of November 2012

Open Source and we would like people to use it!
Thank you!

Questions