Karen Spärck Jones Lecture

2012 Lecture
Work with Scientists: Fun, Profit & the e in e-Science
Professor Carole Goble, University of Manchester
Working with Scientists for Fun and Profit:

the e in e-Science

Prof Carole Goble FREng FBCS CITP
carole.goble@manchester.ac.uk

Karen Spärck Jones Lecture, BCS Academy Conference, York, UK, 11 April 2012
Talk Plan

• e-Science Context
• Automated computation
  – Provenance
• Semantic computation
  – Acquisition of semantics; New Publishing
• Human computation
  – Credit tracking
• The Social and The Technical are Inseparable
  – Reciprocity, co-shaping and disintermediation
• Lessons
What is e-Science

Using computing infrastructure, techniques and approaches to:

• Systematically support global scientific **collaboration**, enable large scale resource, tools and results sharing, assist scientific processing, avoid unnecessary repeated work.
• **Accelerate** scientific discovery, improving scientific **productivity** and stimulate technological **innovation**.
• **Cope** with the scales and speed of scientific innovation and data.
• **enabling**-Science and **empowering**-Scientists
"To be a proper professional you need to think about the context and motivation and justifications of what you're doing... once you see how important computing is for life you can't just leave it as a blank box and assume that somebody reasonably competent and relatively benign will do something right with it."

Karen Spärck Jones

What is e-Science

Apply & test
- Computational thinking
- Computational practices
- Computational innovations
- Software engineering

Stimulate & validate
- Computer Science Research
- Fundamental and Theoretical
- Applied and empirical
- Systems

Application  Technology  Techniques

Fundamental

Applied
"The computer industry knows that scientists can come up with strange ideas and requirements that may well, in time, have broader commercial application elsewhere."

X Science and Y CS Workshop

• NIPS conference 2011
  – Cosmology meets Machine Learning
  – Machine Learning and Interpretation in Neuroimaging
  – Computational Social Science and the Wisdom of Crowds
  – Machine Learning for Sustainability
  – Machine Learning in Computational Biology
  – Machine Learning in Computational Photography
  – Philosophy and Machine Learning
  – Music and Machine Learning

• ICML conference 2012
  – Machine Learning in Genetics and Genomics
  – Statistics, ML and Neuroscience
  – Music and Machine Learning
  – Machine Learning for Clinical Data Analysis

[Gavin Brown]
Agile Data exploration unifying theoretical, empirical, & simulation science

Jim Gray

- Turing Award winner, researcher and manager of Microsoft Research's eScience Group
- Astronomy, geography, hydrology, oceanography, biology, and health care. TerraService, SkyServer databases
• Bio and Medical Informatics
• 1989 – Pen&Pad, GALEN, TAMBIS
• 2001 – UK e-Science Programme
• Research project
• Research, production and support
• Alliance of projects and partners
• Open Source Software
• Services
• Free
Observing Systems Simulation Experiments
JPL, NASA

GWAS, Pharmacogenomics Association study of Nevirapine-induced skin rash in Thai Population

Trypanosomiasis (sleeping sickness parasite) in African Cattle

Astronomy & HelioPhysics

Library Doc Preservation

Systems Biology of Micro-Organisms

BioDiversity Invasive Species Modelling
Software Design for Empowering Scientists

David De Roure, University of Southampton
Gusinde Gibble, University of Manchester

Science is becoming increasingly digital. Scientists' tools are not just the experimental apparatus of the laboratory but are also the software apparatus they use to conduct their research, analyze data, search databases, run simulations, and record their scientific process. New scientific techniques—from DNA microarrays to sensor networks in the environment—are generating volumes of data that
The Current Manchester myGrid Team
Why Science?
Why Science?
Tech Revolution, Social Change

Affordable
- Commodity platforms, instruments and technologies
- Remote access to equipment
- Cheap computers
- Cloud computing

Available and Accessible
- Online open datasets, Open access
- Publication bloat, Service Oriented
  - Open science & Cloud services

Capable
- Shared tools and infrastructure
- Social networks and virtual collaboration frameworks
- Collectivism
Academic spring: how an angry maths blog sparked a scientific revolution

Alok Jha reports on how a Cambridge mathematician’s protest has led to demands for open access to scientific knowledge

The Guardian, April 2012
Scientific Instrument Technology

Data-oriented
Service-oriented
The Cloud

Scientific Information Technology

The Web
The Cloud
Social Media

Scientific Community
Social
The Crowd

[Adapted from Shapin 1984]
Accelerating In the Wild Science
Disintermediation and Co-Shaping

Automated Computation
- Scientific workflows
- Distributed web/grid/cloud services
- Data Integration
- Science as a Service

Semantic Computation
- Ontologies & Knowledge Management
- Linked Data
- Data Spaces
- Research Objects
- Executable papers

Human Computation
- e-Laboratories
- Collaboration & Sharing environments

Instrument
Info
Social

Volunteerist
Crowd-sourcing
Credit/Attribution
Siloed, Distributed
Integrative, Fragmented
Ecosystem of Resources

[Janet Thornton]
In The Wild
Data

- Creation Scales
- Distributed
- Heterogeneity
- Turn cycles
- Integration, Multi-Scale
- Cumulative, Propagation
- Analytics
- Quality and Trust
- Preservation & Conservation
- Available Programmatically

Nature Methods 7, 495 - 499 (2010)
Automated Computation
Scientific Workflows

Dataflow: multi-step coordinated execution of computational processes.
Execution platform, design suite, development kit.
81,000 downloads
Underpinning Science infrastructure
> 400 organisations
Commercial exploitation
>2000 web services catalogued

Principles:
• Just enough Just in Time not Just in Case. Jam Today.
• Be Useful. KISS.
Automated Computation
Scientific Workflows

Trypanosomiasis resistance in African Cattle

Software as a Service / (Cloud) Appliance

Analytic bottleneck

Repetitive, unbiased, accurate record, taming data, transparency, avoiding shortcuts.

Fisher, NAR, 2007, 35(16) 5625-5633
Noyes, PNAS 2011 108(22) 9304-9309

BioDiversity Invasive Species Modelling
American Horseshow Crabs in the Baltic

Dev. Years->Weeks
Runs. Weeks -> Hours
Generalised ENM data mapping and and overlaying pipelines.
In: Computational Lambda Calculus


Out: Systems and Semantics Research


– **Interaction**: Visual programming, workflow reusability, workflow quality, workflow discovery
– **Service** oriented computing, cloud computing, grid computing, optimisation, parallelism, adaptation, security, monitoring and fault correction
– **AI**: re-run analysis, auto-planning, auto-repair, auto-composition, auto-annotation, service matching, auto-substitution
– **Data** integration, data mapping, service integration, provenance tracking, credit propagation, data spaces
– **Data quality**: Missier, CPHC/BCS Distinguished Dissertation *Modelling and Computing the Quality of Information in e-Science*, 2008
I ♥ Mess

Data format mismatches, Identity mismatches
Auto-generated afterthoughts
Contract-last, not Contract-first

- API volatility
- Operation <-> Task Mismatch, Implicit Interaction

1 of 134 BLAST services. 1 search
Simple operation: 5 meanings
Reproducible Science?

1. Provenance
   - Service method obscurity
   - Poor provenance and version reporting
   - Absent input-output correlations
   - Poor credit tracking

2. Quality
   - No peer review
   - Mis-comparison and poor interpretation

3. Preservation
   - Mixed hosting and stewardship
   - Service decay
   - Workflow frailty

D. De Roure Replacing the Paper: The Twelve Rs of the e-Research
Provenance
Where, How, Why, Who, When
Derivation, History, Justification
Reuse, Reproducibility

Semantic Models OPM, VoID, PROV-DM, PAV…
-> Web (W3C) Standards
Provenance
Practice: Models, Methods, Tools.

• >400 publications, 200 in past few years

Workflow Models
OPM, Janus

Database Models
SemiRings and Polynomials

Missier, Sahoo, Zhao, Sheth, Goble, LNCS 6378 2010

Buneman, Cheney, Vansummeren ICDT07/TODS08
Provenance
Practice: Models, Methods, Tools
Foundations: Principles

Pioneered by e-Science, feeding into Web (W3C) Standards

- Models: Janus, OPM, PROV-DM, VoID….

Workflow: Goble, Moreau, Groth, Zhao, Miles, Missier, Ludäscher, Freire…

- Open Provenance Graphs, round-trip. FGCS, 27(6), 2011
- Extending Semantic Provenance into the Web of Data, IEEE Internet Computing, 15(1) 2011
- Mining Taverna’s Semantic Web of Provenance. CCPE 20(5) 2008
- Understanding Collaborative Studies Through Interoperable Workflow Provenance, & Janus: from Workflows to Semantic Provenance and Linked Open Data, LNCS 6378 2010
- Using Semantic Web Technologies for Representing e-Science Provenance LNCS 3298, 2004

DB: Buneman, Cheney, Davidson, Green…SemiRings and Polynomials
Software Engineering, Security, Prog. Languages….

- Components: black boxes, fragile
- Provenance inference and repair Disclosure
- Polysemes and identity management.
Dealing with the mess: Components: black boxes, fragile provenance inference and repair, polysemes and identity management, disclosure.
Putting Lipstick on Pig: Enabling Database-style Workflow Provenance
Amsterdamer, Davidson, Deutch, Milo, Stoyanovich, Tannen, VLDB 2012

[Cheney, 2012]
Reproducible Science?

• Data Quality & Trust
  – Gamble and Goble Quality vs. Trust of Scientific Data on the Web: Towards a Joint Model. 3rd ACM Intl Conf on Web Science 2011

• Decay, Repair & Recovery
  – Belhajjame, Goble, Soiland-Reyes, De Roure: Fostering Scientific Workflow Preservation through Discovery of Substitute Services. 7th IEEE Intl Conf on eScience 2011

• Calculating Credit, Tracking Contribution
  – Buneman, Silvello: A Rule-Based Citation System for Structured and Evolving Datasets. IEEE Data Eng. Bull. 33(3), 2010
  – Buneman, Cheney, Lindley, Müller: DBWiki: a structured wiki for curated data and collaborative data management. SIGMOD Conference 2011: 1335-1338
In The Wild Metadata

- Unstructured/
  Semi-Structured
- Scruffy
- Inconsistent/Variable
- Imprecise
- Open to interpretation
- Mismatching
- Emergent/Changeable
- Premature formalisation
- Retrospective
- Multiple names

74% / 26%
31% / 8%

Scale
Long lived
Social undertaking
Semantics
• WSL-1 protein
• Apoptosis-mediating receptor DR3
• Apoptosis-mediating receptor TRAMP
• Death domain receptor 3
• WSL protein
• Apoptosis-inducing receptor AIR
• Apo-3
• Lymphocyte-associated receptor of death
• LARD
• GENE: Name=TNFRSF25

Q93038 = Tumor necrosis factor receptor superfamily member 25 precursor

Annotation history:

<table>
<thead>
<tr>
<th>ID</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q92983</td>
<td>P78515</td>
</tr>
<tr>
<td>O00275</td>
<td>Q93036</td>
</tr>
<tr>
<td>O00276</td>
<td>Q93037</td>
</tr>
<tr>
<td>O00277</td>
<td>Q99722</td>
</tr>
<tr>
<td>O00278</td>
<td>Q99830</td>
</tr>
<tr>
<td>O00279</td>
<td>Q99831</td>
</tr>
<tr>
<td>O00280</td>
<td>Q9BY86</td>
</tr>
<tr>
<td>O14865</td>
<td>Q9UME0</td>
</tr>
<tr>
<td>O14866</td>
<td>Q9UME1</td>
</tr>
<tr>
<td>P78507</td>
<td>Q9UME5</td>
</tr>
</tbody>
</table>

http://www.expasy.org/uniprot/Q93038
Semantic Web in a Nutshell

e-Science Incubator

- Identify it.
- Self-Describe it.
  - Annotate. Context.
- Publish it.
- Index it.
- (Inter)Link it.
- Map it.
- Query it.
- Reason over it.
- Validate it.

- Ontologies
  - Controlled Vocabularies.
- (Meta)Data publishing
  - Schema-late

Linked Knowledge

Linked Data

Open Data
Semantic Web in a Nutshell

e-Science Incubator

- Identify it.
- Self-Describe it.
  - Annotate. Context.
- Publish it.
- Index it.
- (Inter)Link it.
- Map it.
- Query it.
- Reason over it.
- Validate it.

- Ontologies (OWL, SKOS)
  - Controlled Vocabularies.
- (Meta)Data publishing (RDF)
  - Schema-late

Bio2RDF, Chem2Bio2RDF
LinkedLifeData, LinkedScience.org

Open PHACTS
Open Pharmacological Space
http://www.openphacts.org
Semantic Computation

Encoding Knowledge

• Logic based Knowledge Rep: OWL
  Inverse properties, Property chains
• Simplified KR: SKOS

• Building ontologies: automated reasoning, explanation, debugging, modularity, impact analysis, integration with other formalisms…
  – A suite of DAML+ OIL ontologies to describe bioinformatics web services and data IJCIS 12(2)

• Using ontologies: exploiting expressivity, annotating resources; querying data; verification
  – Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation Bioinformatics 19(10)
  – TAMBIS: transparent access to multiple bioinformatics information sources Bioinformatics 16(2)

Software:
Reasoners, Editors:
Pellet, Fact+, Protégé-OWL
Semantic Computation

- Knowledge Capture (Rightfield)
- Formalization and Verification (SNOMED-CT)
- Consistency Checking (SBML Harvester)
- Classification (Phosphatases, Compounds)
- Semantic Annotation (Array Express/ Gene Expression Atlas, Semantic Assistant)
- Query Formulation (Array Express/ Gene Expression Atlas)
- Query Answering (KUPD)
- Search & co-occurrence (gopubmed)
- Semantic Assistant
- Hypothesis Testing (HyQue)
- Disease Similarity and Model Organism prediction (phenomeBLAST)
- Function Prediction (genemania)

[Michel Dumontier]
Knowledge Acquisition & Publishing Ramps

- Annotation by Stealth
- Automated annotation
- Knowledge enabled routine.
- Open Science
- Tools. Embedding.
- Simplicity.
- A little goes a long way.

http://www.rightfield.org.uk
Kidney and Urinary Pathway Ontology
~1800 classes (~40,000 after imports closure)

Experimental data
195 KUP experiments/databases integrated. Text mining.

Bio2RDF
Linked data

Excel 2 RDF/OWL

Sesame + OWLIM ~50M triples

RDF triple store

Experimental factors

What has been observed, where and when?

E.g. Experiment X showed gene TGFB1 over-expressed in location Kidney under condition model of diabetic nephropathy

Ontologies provide the schema

disease ontology
animal ontology
mouse anatomy ontology

tart publication
[groth et al 2010]
Can calreticulin be associated to the development of human kidney disease?

KUPKB *in silico* result confirmed in wet lab.

Buried in a PDF supplementary file.

1st prize Ontologies Come of Age Challenge (*Jupp et al*)

10th *International Semantic Web Conference 2012: 02/09*
In the Wild Social

Identify Problem
Investigate Prior Knowledge
Make Observations
Generate Hypothesis
Make Prediction
Test Hypothesis
Perform Experiments
Devise New Experiments
Organise Data
Analyse Data
Validate & Compare Results
Draw Conclusions
Log
Communicate
Validate & Compare Results
Pool Results

Hypothesis

Results
Methods

Fast cycle of hypothesis testing

Post hoc Rationalization
Filtering through results
Re-running experiments
Justifying hypotheses
Checking parameters

Formalization

Workflow Description
Conclusions
In the Wild Social EgoSystem

- Open Science....
- Commons based production
- Community curation
- Data Hugging and Suspicion
- Centralisation & Decentralisation
- Long Tail Science
- Provider/Consumer connection
- Coordinated / emergent / evolving team-based, pan-institution, global collaboration
- (Mis)Trust
- Personal reward & risk
- Multi-disciplinary
- Mono-discipline credit
Identify biological pathways implicated in resistance to Trypanosomiasis in cattle using mouse as a model organism.


Identify biological pathways involved in sex dependence in the mouse model, believed to be involved in the ability of mice to expel the whipworm parasite.

Levison S.E., et al Inflammatory Bowel Diseases (2010)
Human Computation: Social Media

voluntary social collaboration, contribution & curation

Lone Scholars
Private Groups
Trusted Collaborators
Scientific Community
Citizens

reproducibility, reusability, transparency
Combining People and Computing
Crowd-Sourced Collective Intelligence
Social Machines

Online game **Foldit**
- University of Washington
- 57,000-plus players
- Deciphered the structure of the retroviral protease protein: an enzyme implicated in HIV
- 10 days. Decade long problem.

**Galaxy Zoo**
- Oxford University
- XXX-plus participants

**Old Weather**
- 27000+ participants
- 950,000+ pages done
Empirical
Social network analysis
Socio-Technical behaviours
  (e.g.) Workflow/Service Reuse+
Instrumentation

Social Machinery
Data entity recognition & mapping
Semantic annotation
Filtering, Learning

Cross-Disciplinary
Social Science
Information Science

Interventional
HCI, CSCW
Socio-Technical behaviours
Web Science

Machinery for Social
Credit and Responsibility
Quality and Trust
Security and Privacy
Filtering, Learning
Semantic Social Data
Disintermediation

+ Zhang et al. IEEE SCC 2011; Tan et al. IEEE Computer 43(9); Starlinger et al. SSDBM 2012
Research Objects
Reproducibility, Integrated Publishing

• Workflow
  – Provenance
  – Conservation & Preservation
  – Executable Publication

• Human
  – Credit Tracking
  – Unit of Scholarship
  – Crowd management

• Semantics
  – Acquisition & Publishing
  – Encoding, Encapsulation
  & Annotation: OAI-ORE, AO…
Accelerating \textit{In the Wild} Science

Disintermediation and Co-Shaping
Some Data Intensive e-Science Directions

- Big Data, sure….
- Text Mining
- Linked Data Publication
- Provenance
- Quality and Trust
- Open Science
- Credit & Attribution
- Reproducibility
- Executable papers
- Science as a Service
- Disintermediation
Grand Challenges in CS Research, 2008

In Vivo - In Silico - Bangham
Ubiquitous Computing - Tom Rodden
Memories for Life - Nigel Shadbolt
Dependable Systems Evolution - Jim Woodcock

Research on the Edges

Semantics for service discovery
Social Semantics
Semantics for discovery

Automated semantic annotation
Auto-annotation of Workflow components
Social computing for crowd curating
The Social and The Technical are Inseparable

Solving the Scientific Problem
Collaborating with Scientists


www.paulgraham.com
Mars and Venus

- Problem rewrite: Solve another problem which isn’t yours but is more researchy.
- Over-complication: Solve this harder problem rather than offering the straightforward solution.
- Over-simplification: Consider the body is a cylinder.
- Proof of concept: Well I’m done.
- Be like me: It would be easy if you were like me.
- Paternalism: Repeating past mistakes despite our experience & protests.
- Short term instant gratification: Just about holds together to get the results for my paper.
- Hackery. Simplifications, and monoliths store up trouble down the road.
- Service provider. I’m now maintaining legacy code, managing releases and manning a help desk.
Reciprocity Balance
Economics of Reputation

**Reward**
- Competitive advantage.
- Be the first with the Nature paper.
- Credit, fame, acclaim, peer respect.
- More funding.
- Get my X adopted

**Risk**
- Beaten by Prof Y.
- Looking stupid.
- Distraction from my Science
- Being left behind.
- Being out of fashion.
- Retaining my special sauce.
- Sustaining software and services
- Being scooped.
- Being scrutinised.
- Being misinterpreted or misrepresented.
- Being blamed..
- Releasing results too early.
Reciprocity Dating

e-Science
- Long term relationship
- Impact and adoption
- Mutual Trust, Mutual Benefit

Science Case Studies
- Flirtatious speed dating
- One night stand
- At a distance crush
- Self-centric-ness
A NLP People Speak
Jock McNaught, Goran Nenadic

Frustration: scientists becoming aware of what research conducted over several prior decades could actually offer, and consequently becoming more demanding so I don’t have to dumb-down.

Inspiration:

• Connecting up text mining results with how a scientist wants to interact with such knowledge,
  – Instead of information need templates, match one word queries facts and rewrite into questions, classify and show to the user to pick from
• Handle integration of results. The Significance of context
  – Information Retrieval -> Question Answering
  – Not just find a fact or an association between a gene and a disease. What is the causality? What is the temp for an interaction?
• Integration of text mining results from many different huge data sets
  – Text mining as a Service
• Non-trivial error-rates in manually curated databases.
  – What is the true state of our science base that we've plugged into our processing actually is and can we rely on what we're inferring from it.
Dating Gone Bad

How can I get Scientists™ to use my software/technique, which is clearly superior to what they already have and unaccountably like? How do I make them work with me

Why don’t these computer people give me something I can use. Make my favourite desktop application faster. Make my dataset bigger. Get me their data set. Don’t let them see my results until I say so. Give me something I couldn’t get before.
Some Salt that works Today, Better Salt Tomorrow
Just Enough, Just in Time
Some for me, Some for you
Customised specific preferred to Unconfigurable Generic
- Rarely a direct link between Science and any kind of Computer Science Research
- Rarely a direct link between fundamental Computer Science Research and Computational Science
- Computational Science + Applied CS
- Tools and Techniques
• Rarely a direct link between Science and any kind of Computer Science Research
• Rarely a direct link between fundamental Computer Science Research and Computational Science
• Computational Science + Applied CS
• Tools and Techniques
Reality Checks

- User Experience beats Smart.
- Sequins & Sparkle matter.
- Generic solutions have to be configurable and specialised.
- Scientists are Naughty.
- Eat Your Own Dog Food.
- Innovation is not always smarter CS.
- Stability outweighs latest FooBar Thing.
- The Bigger Picture.
- There are No Green Fields.
- The scientists’ tech-ecosystem.
- Embed staff.

If you are successful, you will be used – have a plan!

Multi-disciplinary research lip service.
Overarching challenge to CS

• How to combine an emerging formal understanding of a problem with some kind of evaluation to determine which techniques actually provide value to the scientists.

• And thus worth the huge effort to actually deploy in a widespread and usable way.

James Cheney on Provenance and Databases
Empirical Research is OK

- "... the fundamental difference between a theoretician and an engineer is that an engineer is away and doing something once he's grasped that there's a way of doing something, and the theoretician is strangely enough most of the time lagging behind, cleaning-up -- that's the best word I can use ..."

Tom Kilburn 1992
Take Home Messages

• Working with Scientists is Fun and Profitable.

• Work hard and take care to make it fun and profitable for all parties.

• Mess is fun. And profitable.
Talk Acknowledgements

- James Cheney (Edinburgh)
- David De Roure (Oxford)
- Paul Groth (Vrije, Amsterdam)
- Michel Dumontier (Carleton)
- Simon Jupp (EMBL-EBI)

- The many members of the myGrid projects and family.
Further Information

- myGrid
  - http://www.mygrid.org.uk
- Taverna
  - http://www.taverna.org.uk
- myExperiment
  - http://www.myexperiment.org
- BioCatalogue
  - http://www.biocatalogue.org
- SysMO-SEEK
  - http://www.sysmo-db.org
- MethodBox
  - http://www.methodbox.org.uk
- OMII-UK
  - http://www.omii.ac.uk
- Wf4ever
  - http://www wf4ever-project.org
- Software Sustainability Institute
  - http://www.software.ac.uk