

Open Access in the sciences - the author and research perspective

Cambridge OA conference, 21 Oct 2009

Tim Hubbard, Head of Informatics

Wellcome Trust Sanger Institute

th@sanger.ac; @timjph

Topics

- Background
 - Post-publication data release
 - Pre-publication data release (e.g. human genome)
- Managing OA at WTSI
- Challenges of data and privacy

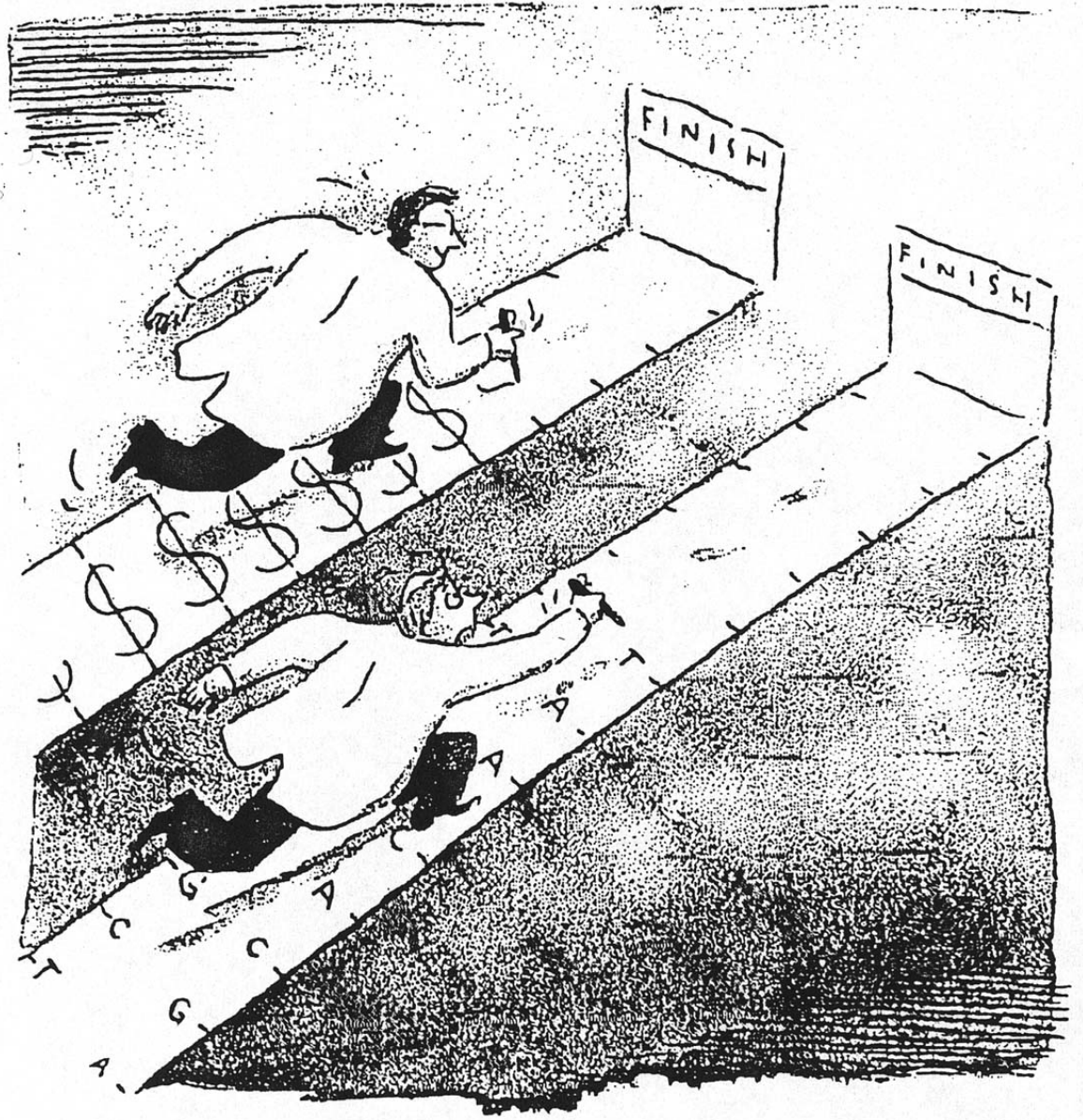
OA Principles

- Easier to share and reuse data due to internet/computers
- Disciplines such as biology increasingly data driven
 - No one can imagine all types of analysis that could be carried out
- Much of research paid by public through taxation
 - Increasing interest in reading original articles (particularly medical)
 - Tendency of researchers to be monopolistic, in interest of their careers, however have obligation to share data to advance research

Human
genome
race

won by
public
project

open
access for
all



International agreement on data release

“All human genomic sequence information should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society.”

The Bermuda Statement, February 1996

Assemblies of 1-2 kb are deposited in public database (GenBank, EBI) every 24 hours

No patents are filed

Benefits of early data release

- Several years ago, WTSI started sequencing two organisms
 - Streptococcus equi - strangles in horses
 - Streptococcus zooepidemicus - inflammatory disease in horses.
- Sequences were very repetitive and tricky
 - 5 years to get from shotgun to finished sequence.
- Collaborators at Animal Health Trust used early draft sequence
 - designed a novel diagnostic test for Strangles, now in commercial production
 - a vaccine, which is currently undergoing trials
- Advances came earlier due to early data release.

Ft Lauderdale, Jan 2003

- Bermuda principles reaffirmed*
- led to new NIH/WT policies to divide funding into two classes:
 - R01 projects:
 - Competitive
 - Release data on publication
 - “Community Projects”
 - Non-competitive
 - Managed
 - Release data in real time

*Nature 421 , 875 (2003)

Broader adoption

- Funders grant forms include data sharing section
- Consortia data sharing / publishing agreements
- Increasingly publishing requirements
 - WT PMC 6 months rule
 - NIH PMC 12 months rule
- 11/2008: Toronto follow up meeting on early data sharing, writeup published in Nature online today.

Problems to solve

- Cultural attitudes towards data sharing
 - New ways of allocating credit
 - Adjustment to more competitive environment
- Practical issues of data sharing
 - Standardization of data sets
 - Engineering to allow distributed data access
 - Stable infrastructure funding to support data archives
 - Secure analysis of private data

WTSI implementation of OA

- Previously
 - Genome sequence (human consented + model organisms)
- Today
 - Genome sequence, genotypes (some disease related), phenotype data (models), high throughput assay data (transcriptomics), WT publishing policy
- Data sharing committee
- Data sharing policy
- Tracking of compliance

OA publications policy

- WT policy
 - All original articles must be deposited in UK pubmed central (ukPMC) within 6 months
- WTSI implementation
 - Library tracks compliance internally (currently 76% for all publications since 10/2006)
 - OA publication encouraged

WTSI data sharing policy

- **Access:** The Institute aims to provide rapid access to data sets of use to the research community and will place these in publicly accessible repositories when possible. The Institute will support data and interoperability standards to maximise access and ensure ease of integration with other global resources
- **Ethical considerations:** Conducting genetic and genomic research carries responsibilities to protect confidentiality and the privacy of research participants. Access to certain data sets will therefore be carefully managed and granted in a transparent manner to all appropriately qualified researchers
- **Rights of data providers:** The Institute recognises the need for researchers to be appropriately credited for their scientific contribution and investment in data generation. It is therefore expected that all researchers both honour agreements in line with Fort Lauderdale's data sharing principles and appropriately acknowledge the contributions of others
- **Optimising Translation:** The Institute recognises that, in specific instances, the use of intellectual property protection and attendant potential delays to data sharing may be necessary to prevent inappropriately exclusive claims by others and to ensure health benefits occur

OA issues for institutes

- Tracking compliance needs to be proactive
- Pragmatic assessment of value of intermediate data
- WTSI approach to pre-publication data release
 - Raw data automatically, immediately deposited to repositories
 - Intermediate analysis provided via institute websites
 - Final analysis outputs (e.g. linked with publication) submitted to appropriate database, repository

Repositories must keep up

- 1000 genomes project generated more raw sequence data in a few months than entire existing Genbank/EMBL/DDBJ archive
- Technology improvements in sequencing increasing output ~10x each 2-3 years. First genome cost ~£1billion (2000); currently £40,000; projected £1,000 ~2013
- ELIXIR project to improve long term support for biological data repositories at EBI

Openness

Privacy

- Can you have both?
- Biology getting closer to medicine

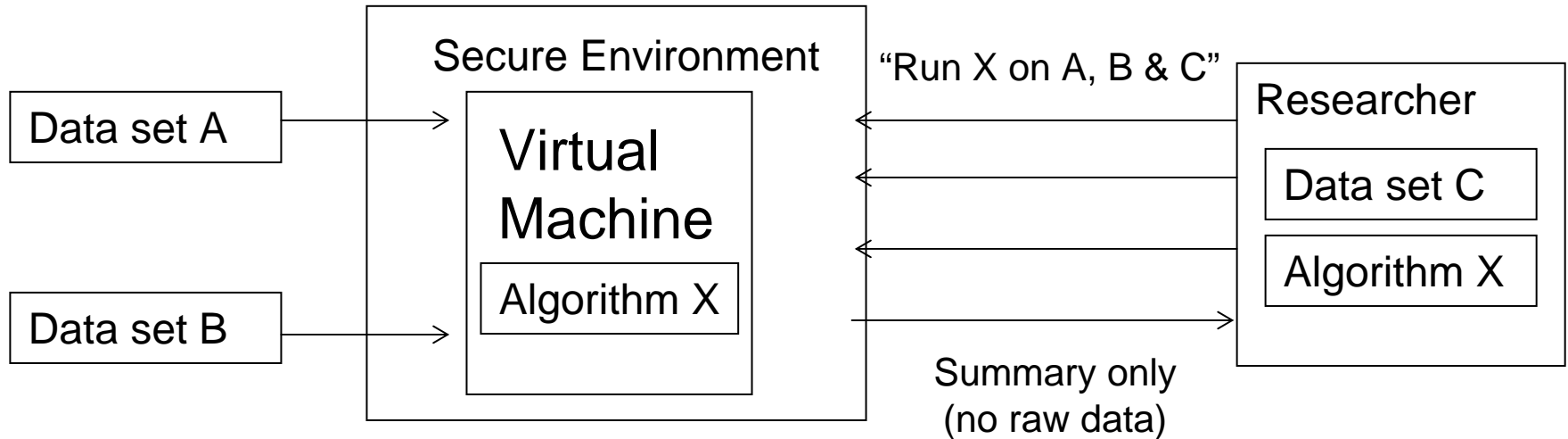
Hard to be anonymous and still useful

- Example:
- Looking for long term effects of environmental pollution
- Require location based data (postcodes) to link to patients
 - How to link without disclosing postcodes?
- Algorithm for integration could be quite complex, linked to weather, wind patterns etc.
- Any effect might be correlated with genotype

Secure analysis of private data

- Privacy is an issue
 - Public happy to contribute to health research
 - Public not happy to discover personal details have been lost from laptops / DVDs etc.
- 3 potential solutions
 - “Fuzzify” data accessible for research
 - Social revolution (personal genetic openness)
 - Technical solution

Honest Broker



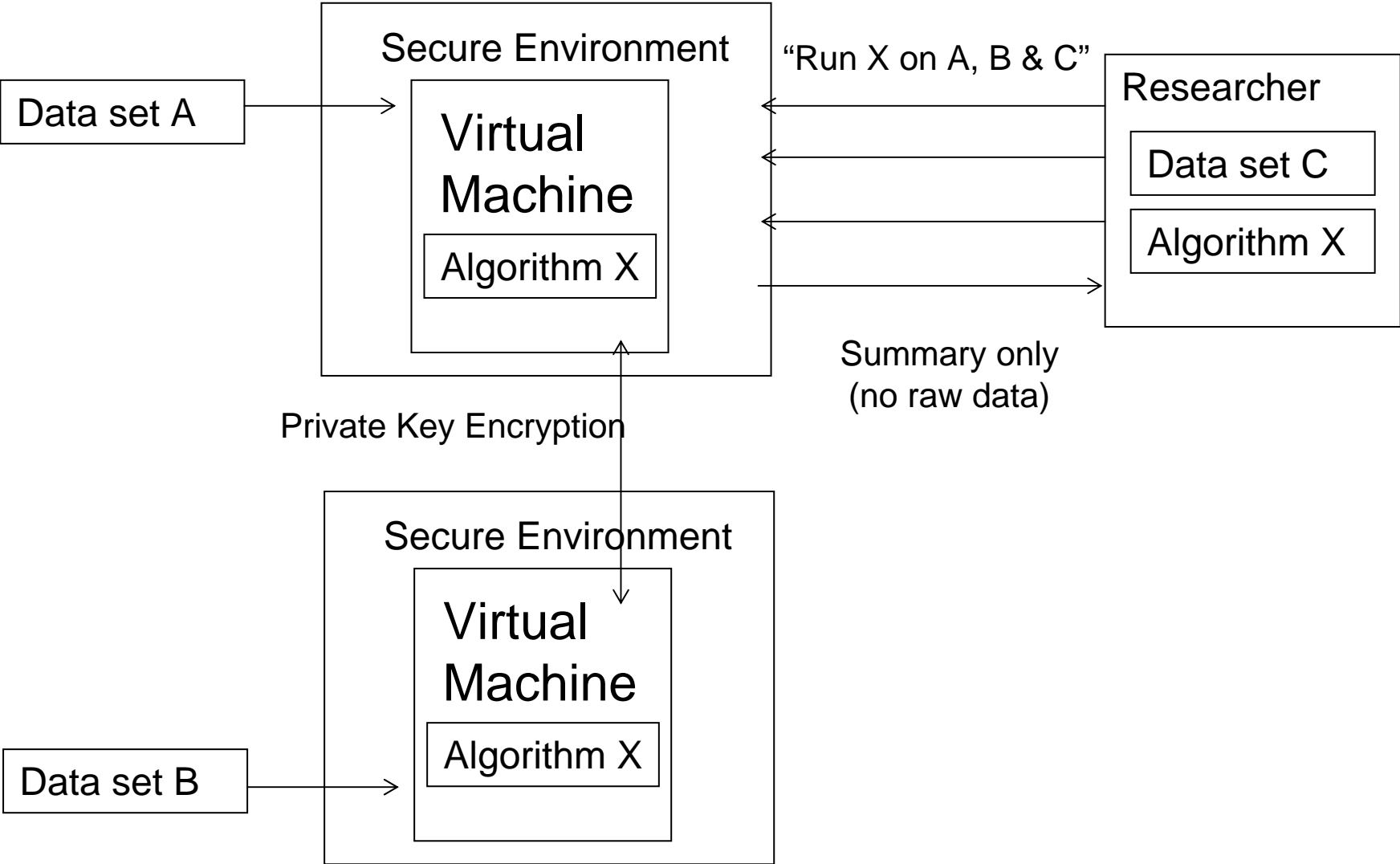
Virtual machine (VM):

- VM has sole access to raw data.
- Algorithms implement analysis within VM.
- VM guarantees that only summary data can be exported

Existing examples:

- cloud computing: Amazon ec2
- iphone SDK (all software is developed against SDK, with controlled access)

Honest Broker



Technical solution hard, but could be good enough

- Objective is to avoid leakage of raw identifiable or potentially re-identifiable data
- Make it easy enough to do practical research
- Make it hard enough and illegal to bypass the system

A generic problem of our time

- Age of pervasive surveillance
- DNA database
- Police National Computer
- Exchange of government data

