Data sharing:
lessons from genome sequencing

Tim Hubbard

CASIMIR Sharing Data and Resources
for Functional Genomics

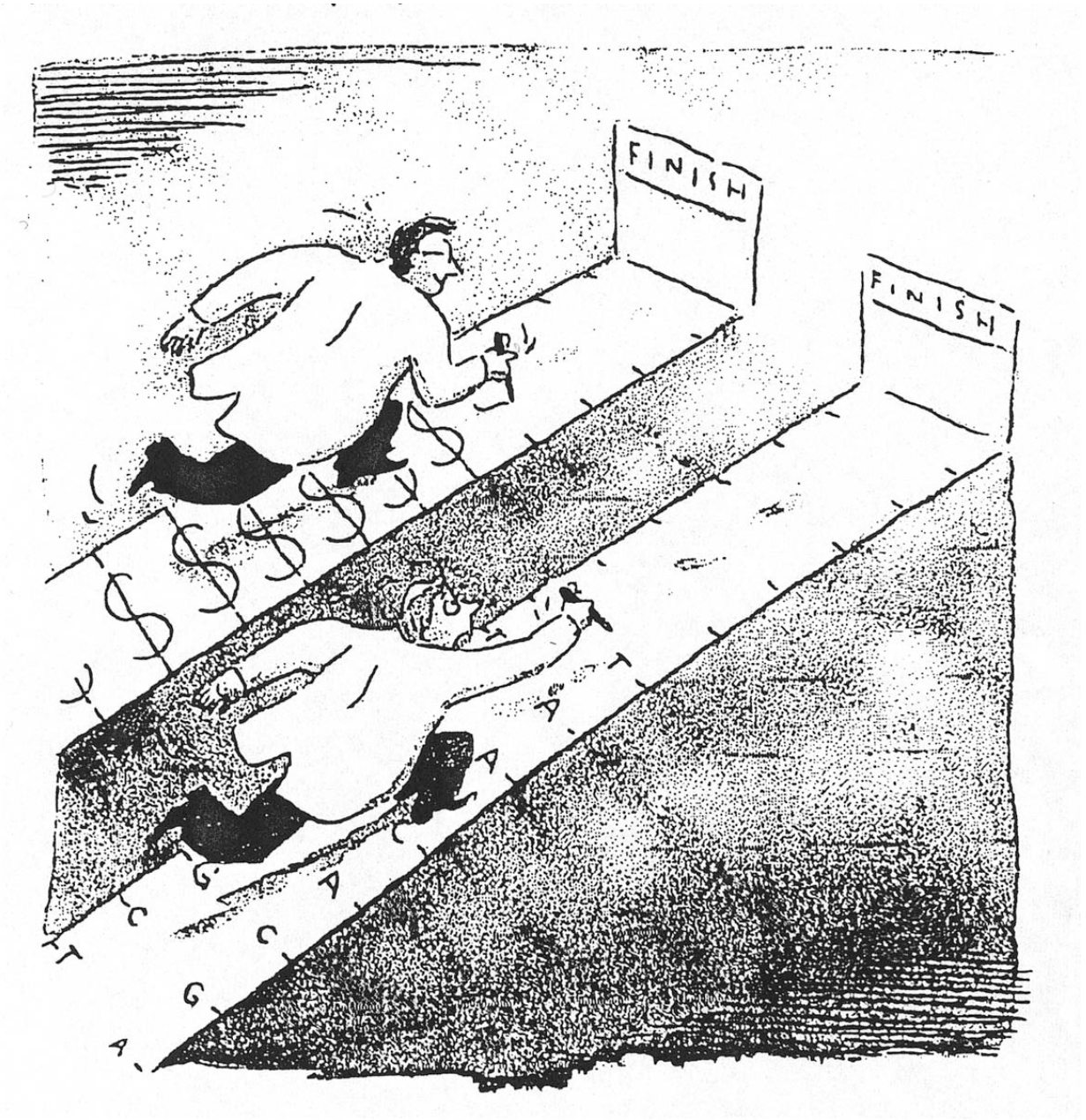
21st May 2009

History

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

Positive v Negative

- Genome centres were forced into this
  - but they recognise they got a lot of credit
  - even when someone else took their data and published a first analysis, official publication got more citations

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)

Tripartite Sharing of Responsibility

- *Funding agencies should*:

- Designate appropriate efforts as Community Resource Projects;
- Require, as a condition of funding, free and unrestricted data release from community resource projects to appropriate central & searchable public databases, & vigorously ensure that this occurs;
- Ensure sufficient support for curation, maintenance & distribution of the data to the community, as well as resources to perform initial analyses using the resources that they have generated;
- Support central databases that will house & distribute the data in a way that prevents fragmentation of the data.

Tripartite Sharing of Responsibility

- *Resource producers should:*

- When feasible, publish a Project Description, to inform the scientific community about the resource project & to provide a citation to reference the source of the data;
- Produce data of consistently high quality;
- Make the data generated by the resource immediately & freely available without restriction;
- Recognize that even if the resource is occasionally used in ways that violate normal standards of scientific etiquette, this is a necessary risk set against the considerable benefits of immediate data release.

Tripartite Sharing of Responsibility

- *Resource users should:*

- Appropriately cite the source of the data & acknowledge the resource producers;
- Recognize that the resource producers have a legitimate interest in publishing prominent peer-reviewed reports describing & analyzing the resource that they have produced;
- Respect the producer's legitimate interests. In some cases, this might best be done by discussion or coordination with the resource producers;
- Assist journals & funding agencies to play their proper roles in ensuring, through the peer review system, that the system works fairly for all constituents.

Examples: Community Resource Projects

- The International Human Genome Sequencing Consortium
- Mouse Genome Sequencing Consortium
- Mammalian Gene Collection
- The SNP Consortium
- The International HapMap Project
- The Cancer Genome Atlas

Broader adoption

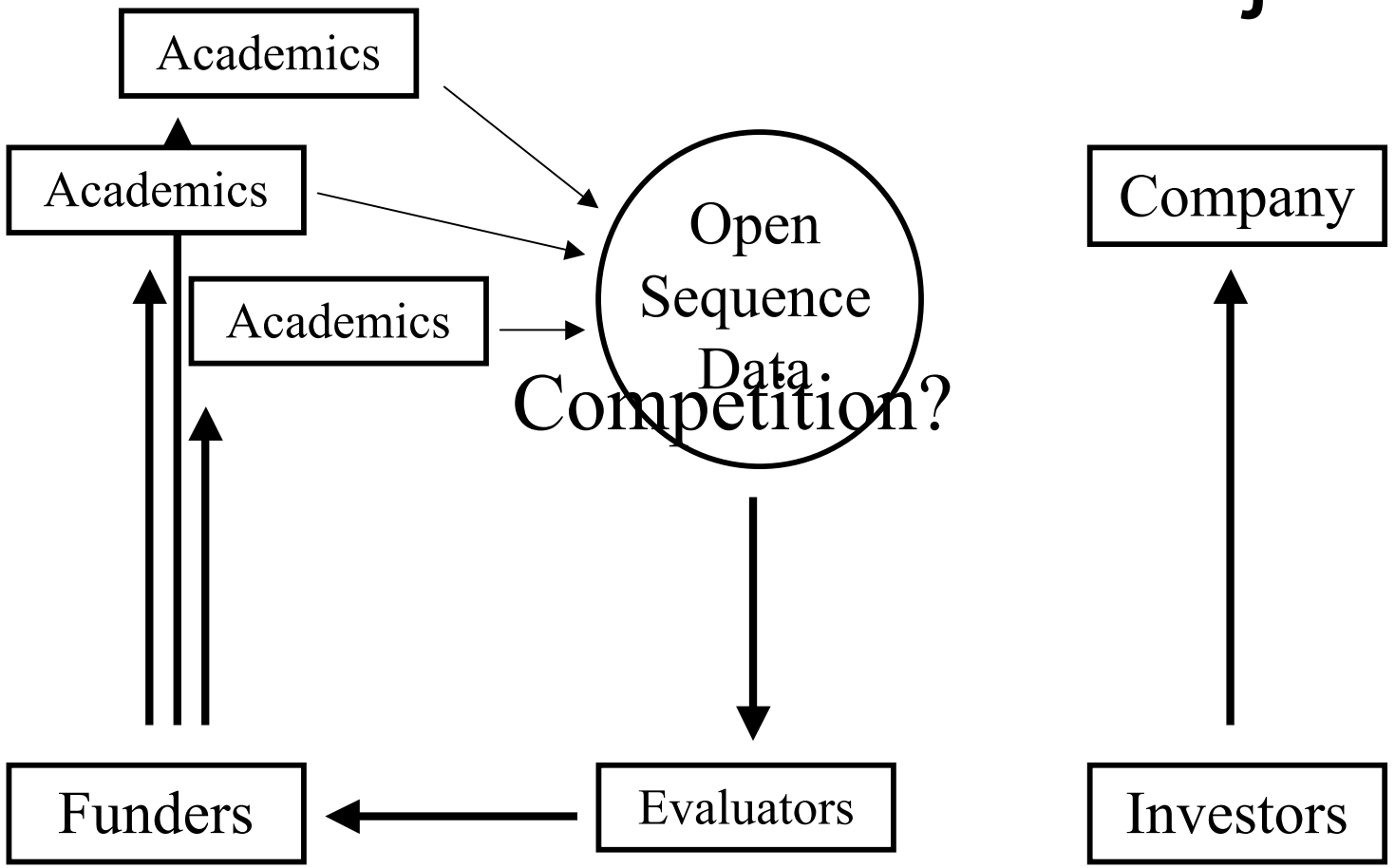
- Funders grant forms include data sharing section
- Consortia data sharing / publishing agreements

- Increasingly publishing requirements
  - WT PMC 6 months rule
  - NIH PMC 12 months rule

Why widespread adoption?

- Value for money
  - Government recognition that if data is kept closed, they don't maximise their investment
- Competition
  - In a sea of data, keeping your data/publications private is counter productive: lose credit; inhibits collaborations
- Harness others
  - Companies chose to release data rather than keep secret: the more people analysing a block of data, the more valuable it is.
- Capacity building

The Human Genome Project



Weekly conference calls
Competitive Collaboration

Toronto: issues

- Credit
  - Marker publications with DOIs: could be easier now in open access/web journal world
  - Microcitation
  - Clarity on publishing restrictions
    - Ft Lauderdale: vague
    - ENCODE: sharp and clear
- Timeline issues
  - Release policies need to be tied to data creation
  - Avoid delays due to ‘Q/C’
- No repository
  - Generic location for unformated data
- Medical
  - Applies to all datasets (pre and post publication)
    - Consent; Regulating access;

Toronto: conclusions

- Narrow focus - issues around pre-publication data release only
  - Noted that everyone should release on publication, but many still don't.
- Conclusions
  - All or some?
    - Mandatory for community projects
    - 'Encouraged' for rest

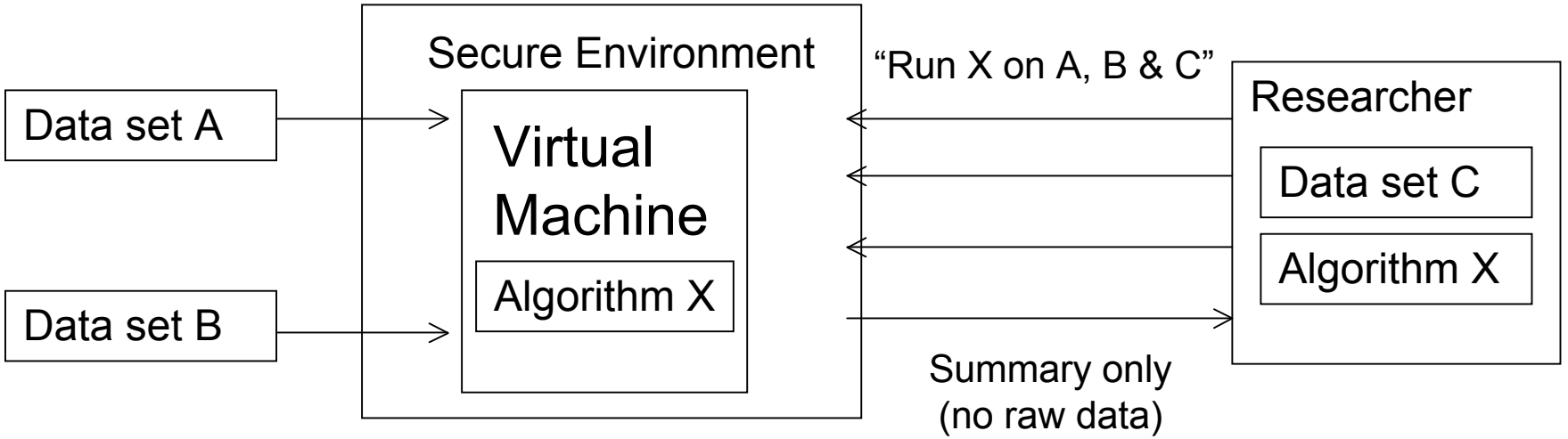
Problems to solve

- 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

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



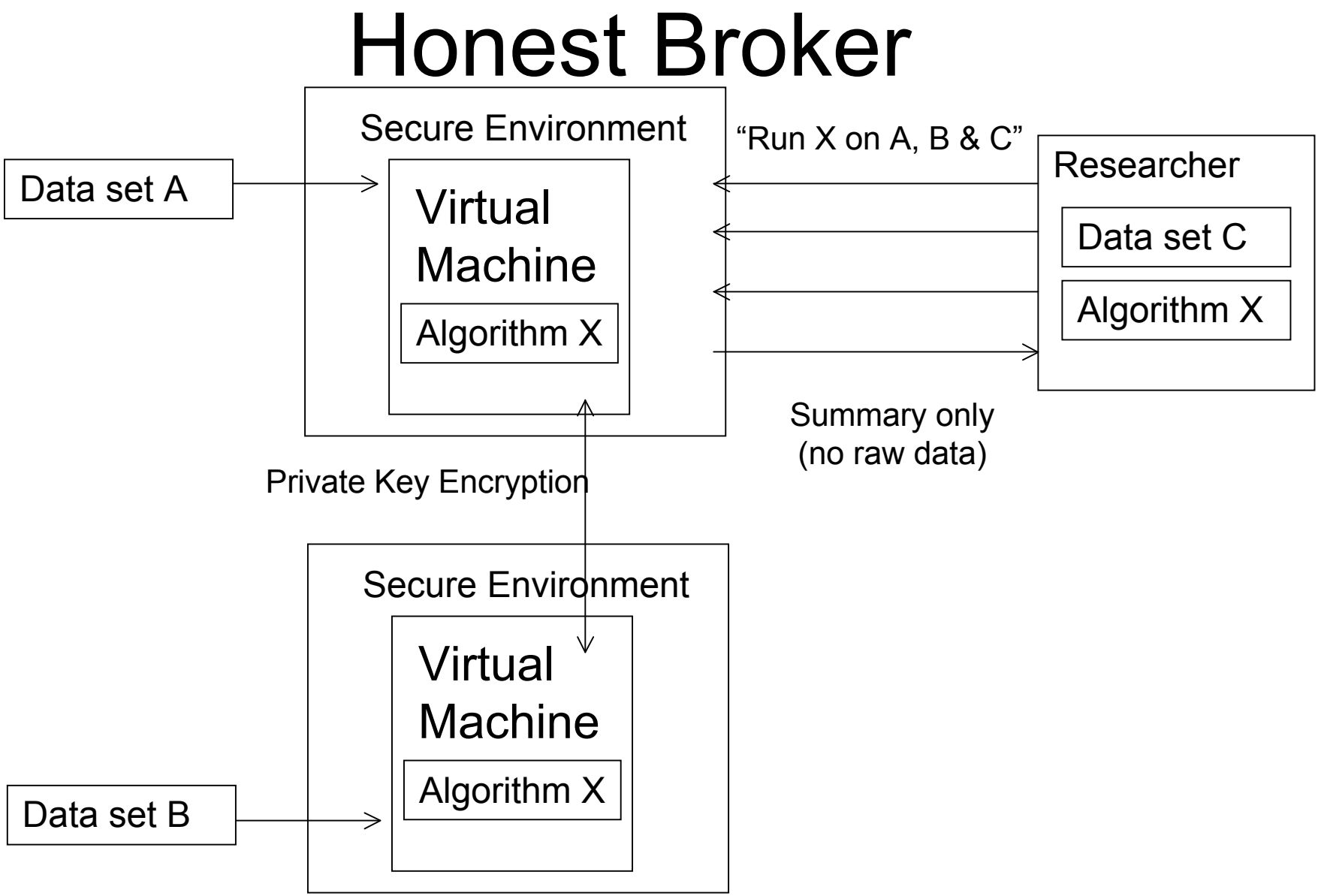
Virtual machine (VM):

- 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

Enforcement?

- We are really talking about benefits, hazards of ‘free trade’ between scientists
- Free trade between countries
  - difficult bilateral agreements, big transaction costs
  - global norms coupled to enforcement: WTO
- Free trade between scientists
  - scientists worried about loosing credit
  - global association of funders and journals could provide confidence of enforcement: WDSO
  - more of professional than legal association