

Common provenance questions across e-science experiments

Simon Miles

King's College London

Outline

- Gathering Provenance Use Cases
- Our View of Provenance
- Sample Case Studies
- Generalised Questions about Provenance
- General Issues in Creating an Infrastructure for Provenance
- Further Details

Projects and Colleagues

- Started gathering use cases 5 years ago
 - Provenance-Aware Service-Oriented Architecture (PASOA) project
 - EU Provenance project
 - Case studies used as requirements for general infrastructure, a subset implemented
- Many collaborators, and particularly
 - Luc Moreau, University of Southampton
 - Paul Groth, Information Sciences Institute, University of Southern California

Gathering Use Cases about Provenance

Methodology

- Describe idea and model of provenance
- Give others' use cases as examples
- Ask what information (about the past) is being determined and/or used in each case

Use Case Expression

- Describe preceding actions by scientist(s)
- State what the scientist determines

A bioinformatician, B, downloads sequence data of a human chromosome from GenBank and performs an experiment. B later performs the same experiment on data of the same chromosome, again downloaded from GenBank. B compares the two experiment results and notices a difference. **B determines whether the difference was caused by the experimental process or configuration having been changed, or by the chromosome data being different (or both).**

Provenance

What Provenance Is

- Oxford English Dictionary:
 - the fact of coming from some particular **source** or quarter; **origin**, derivation
 - the **history** or pedigree of a work of art, manuscript, rare book, etc.;
 - concretely, **a record of the passage** of an item through its various owners.
- Provenance is important for:
 - Interpretation
 - Judging value

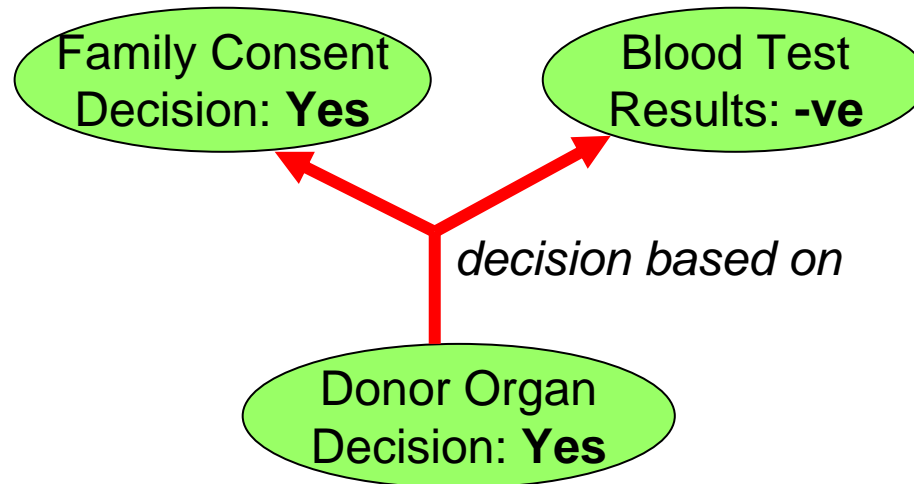
Causation

- Everything that is part of the provenance of an item is a **cause** of that item being as it is
- For example, provenance of a bottle of wine includes:
 - Grapes from which it is made
 - Where those grapes grew
 - Steps in the wine's preparation
 - How the wine was stored
 - Between which parties the wine was transported, e.g. producer to distributor to retailer

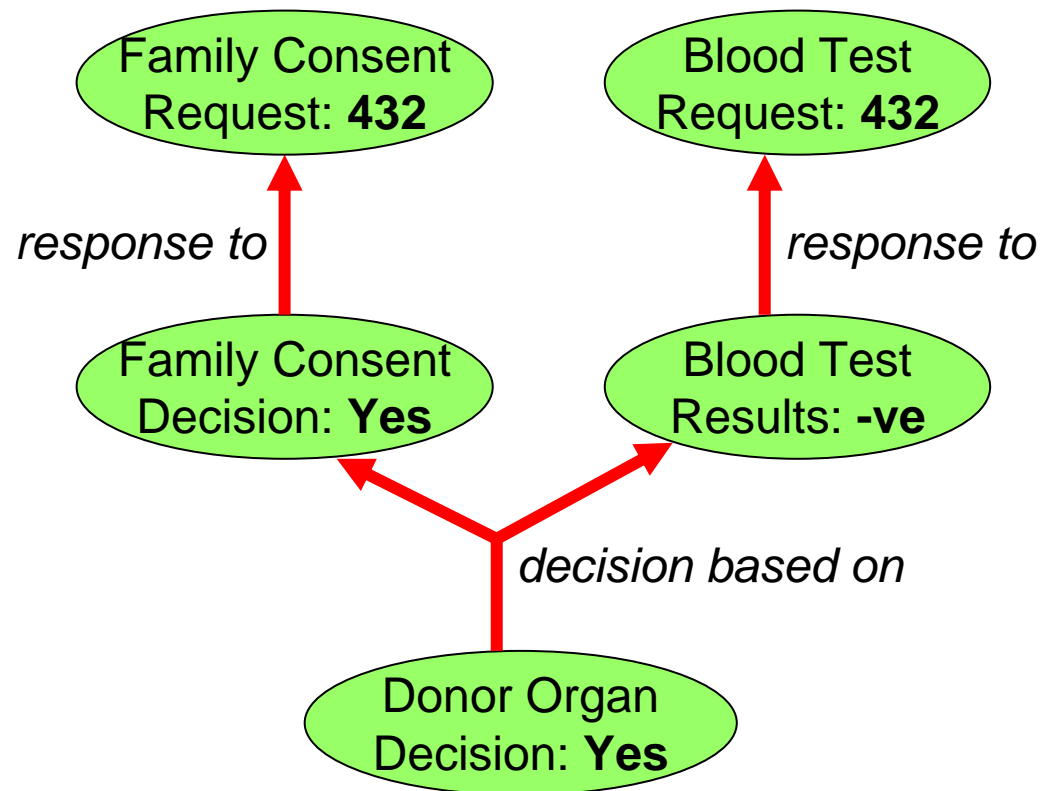
Causal graphs

Donor Organ
Decision: **Yes**

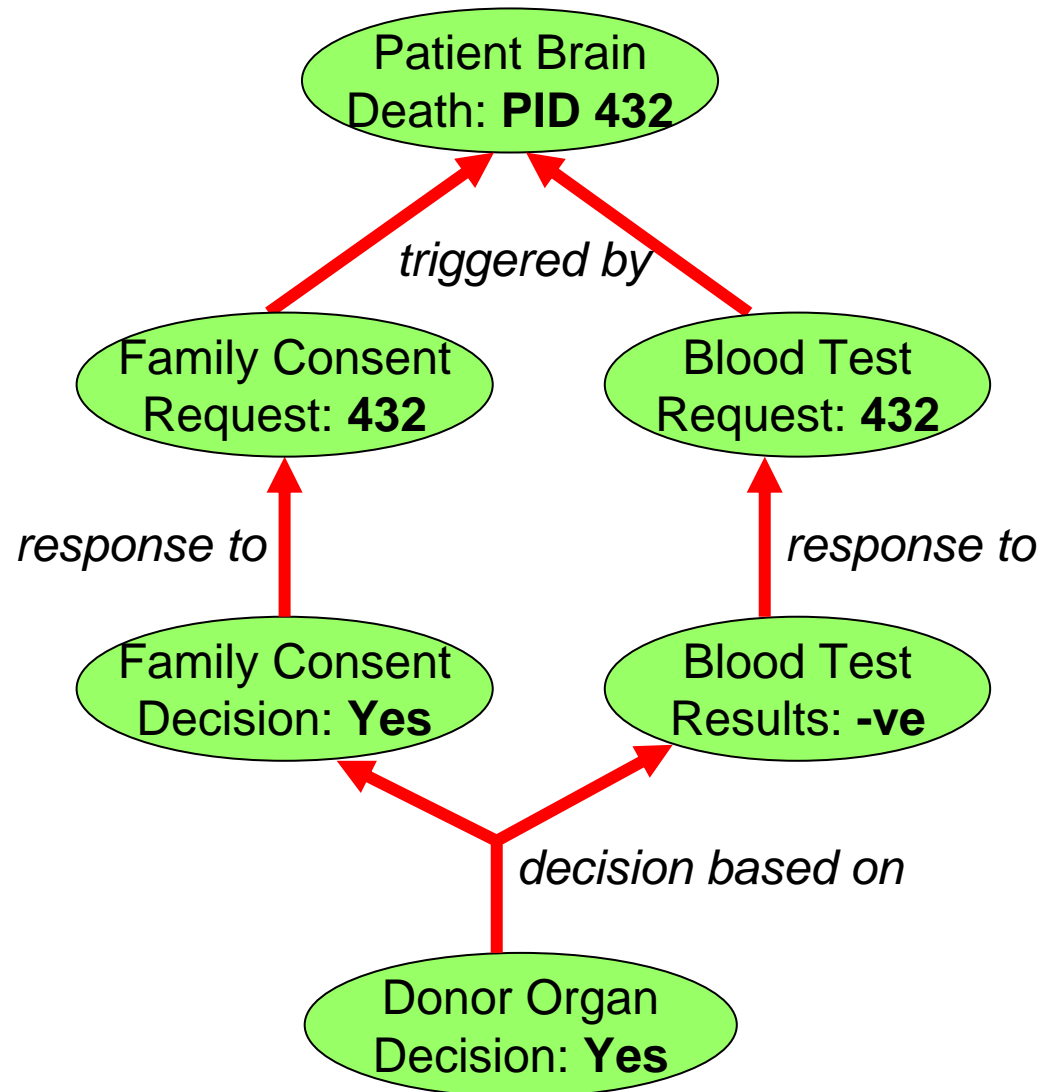
Causal graphs



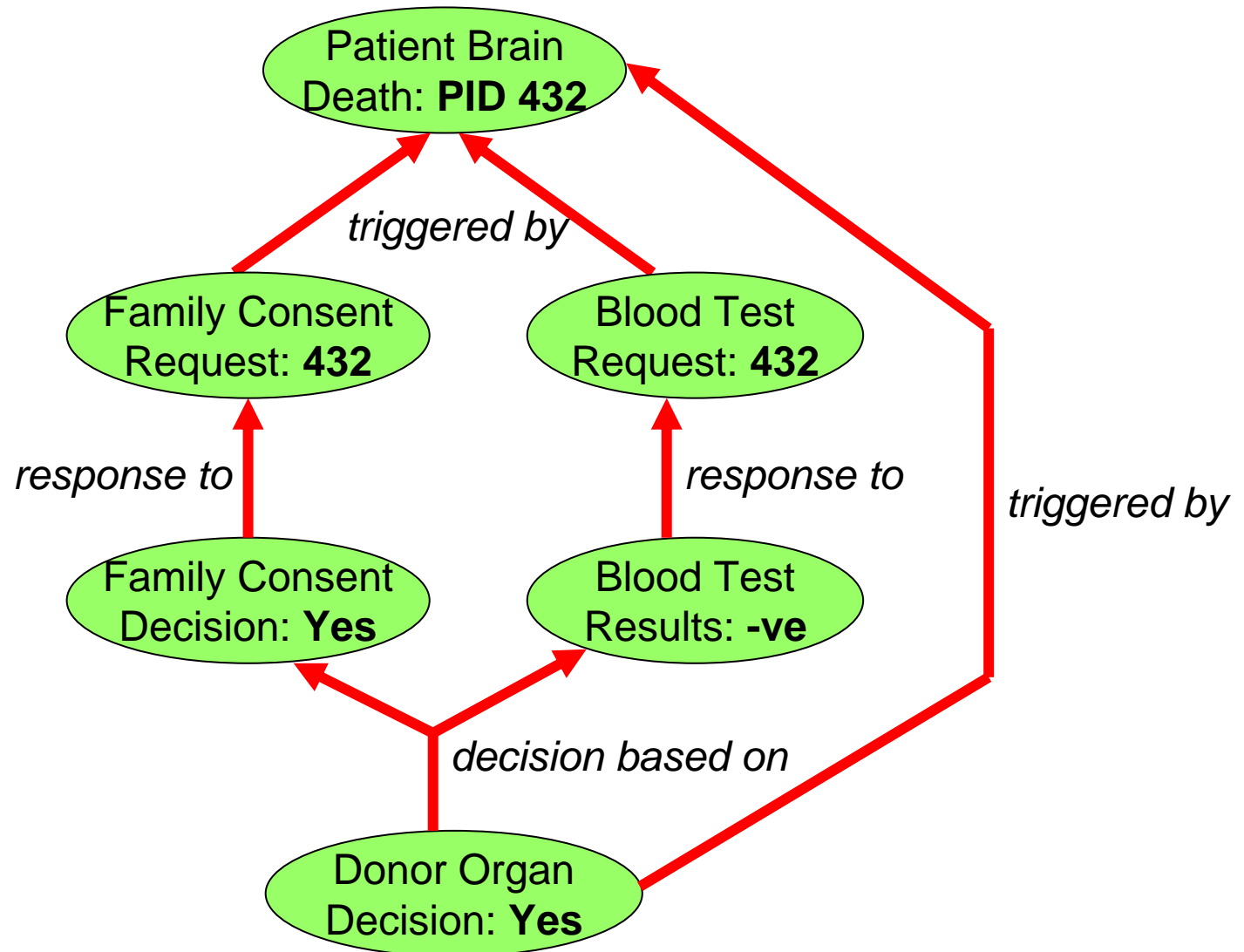
Causal graphs



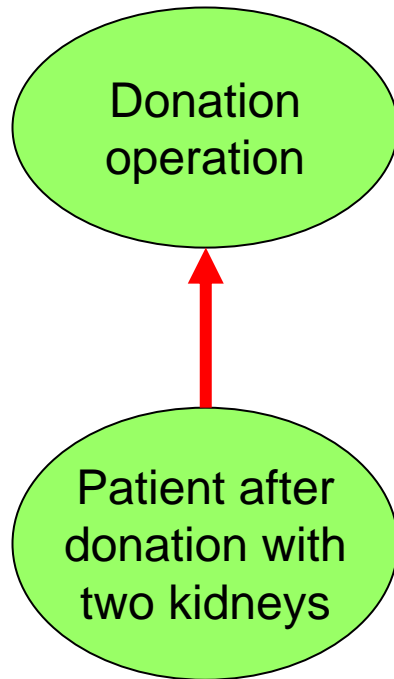
Causal graphs



Causal graphs



Causal Connections



- Causes and effects are **occurrences**
 - Occurrence of a process or event, or
 - Occurrence of a data item or physical artefact being in a particular state
- Counter-factual definition:
 - Effect would not have occurred if cause had not occurred

Sample Case Studies

Bioinformatics

- Klaus-Peter Zauner at the University of Southampton
- Analysing the complexity (information content) of gene and protein sequences
- Purely electronic experiment implemented as UNIX shell scripts calling local executables
- Inputs downloaded from RefSeq, GenBank
- Output data is a graph plot graphics file

Provenance Questions

- Questions included:
 - **What sequences led to the production of this output graph?**
 - **I ran what I thought was the same experiment (same configuration, same input data) on multiple occasions, but the output looks different - what was different?**

Proteomics

- Centre for Proteomic Research at the University of Southampton
- Identifying proteins within biological samples
- A lab based experiment to extract data used as evidence for identification
- Followed by search of public and local databases for proteins matching this evidence

Provenance Questions

- Questions included:
 - **What machine settings did I use in obtaining this successful identification (so I can try similar settings in a later experiment)?**
 - **What was the perceived reliability of the pieces of evidence and database entries used to identify this protein?**

Particle Physics

- ATLAS experiment at the Large Hadron Collider, CERN
- Identifying traces of particles produced by the collision of particles at high energies
- Much data processing, first at CERN, then by physicists around the world
- A lot of processing in terms of large sets of data, of which only subsets may be used in any one experiment

Provenance Questions

- Questions included:
 - **Has the data set from which the subset of data I am experimenting on is extracted, been updated?**
 - **Were these results produced by processing involving a version of a library now known to have bugs?**

Organ Transplant Management

- Inter-hospital organ transplant management with software support
- Governed by the Catalan Health Authority
- Patients build up healthcare records through check-ups, tests, surgery
- When a donor dies, standardised procedures guide transplant process involving tests of donor organ, recipient, and making use of healthcare records

Provenance Questions

- Questions included:
 - **Who made the critical decisions which led to this donor organ being accepted/denied for transplantation?**
 - **Where were the time lags in getting from donation to transplant?**

And more...

- Genetic diseases
 - Aircraft simulation
 - Police databases
 - Social planning
 - Chemicals and lasers
 - Grid service reliability
 - Brain image analysis
- Healthcare records
 - Ecological simulation
 - Medical images
 - Aerospace aftersales
 - Chemical prediction
 - Galaxy formation
 - Near-earth objects

Generalised Common Questions

Generalised Questions

- How did I (or someone else) come by this result?
(genetic diseases, aerospace examples)
- What was common and relevant in the history of this set of successful outcomes?
(proteomics, social planning examples)
- Was the process claimed to be performed the one which was actually performed?
(organ transplant, chemistry examples)

Generalised Questions

- What inputs were used to derive this output? (bioinformatics, particle physics examples)
- What software produced this data? (particle physics, genetic diseases examples)
- Can I generalise from the process by which this result was produced to a reusable plan? (chemistry example)

Generalised Questions

- Were these regulations followed in producing this result? (proteomics, transplant examples)
- Are these two independent conclusions actually based on the same faulty assumption/input? (grid reliability, policing examples)
- What differed between the way these two results were produced? (social planning, bioinformatics examples)

Generalised Questions

- Were tools or services used in a meaningful way? (bioinformatics examples)
- What effect do the tools used have on my rights to patent or publish? (bioinformatics examples)
- Which inputs have a pronounced effect on the output? (social planning, galaxy formation examples)

Generalised Questions

- Were the inputs to this experiment of reliable quality? (chemical prediction, biodiversity examples)
- Who was the source of this decision or input fact? (organ transplant examples)

General Issues for Provenance Infrastructures

Infrastructure Issues

- Record or infer connections between data, processes, events (and plan in advance)
- Naming no longer existent data, processes, events, states of artefacts
- Scalability of storage for large data sets
- Privacy infringement by ability to infer
- Requirements to delete old data
- Querying vast causal graphs
- Post-processing for most appropriate answers

Extra Resources

More Detail, More Use Cases

- The Requirements of Using Provenance in e-Science Experiments
 - by Miles, Groth, Branco and Moreau
 - Journal of Grid Computing
- <http://twiki.pasoa.ecs.soton.ac.uk>
 - See Use Cases section
- <http://www.gridprovenance.org>
 - See Applications section

More Detail, More Use Cases

- The Provenance Challenge
 - First and Second used brain image analysis case study
 - Third (current) uses near-earth object detection (astronomy) case study
- Workflow-oriented but trying to make connections with database provenance
- <http://twiki.ipaw.info/>

Credits

- Thanks to the many who were interviewed and supplied the use cases (see papers and websites for all the credits)