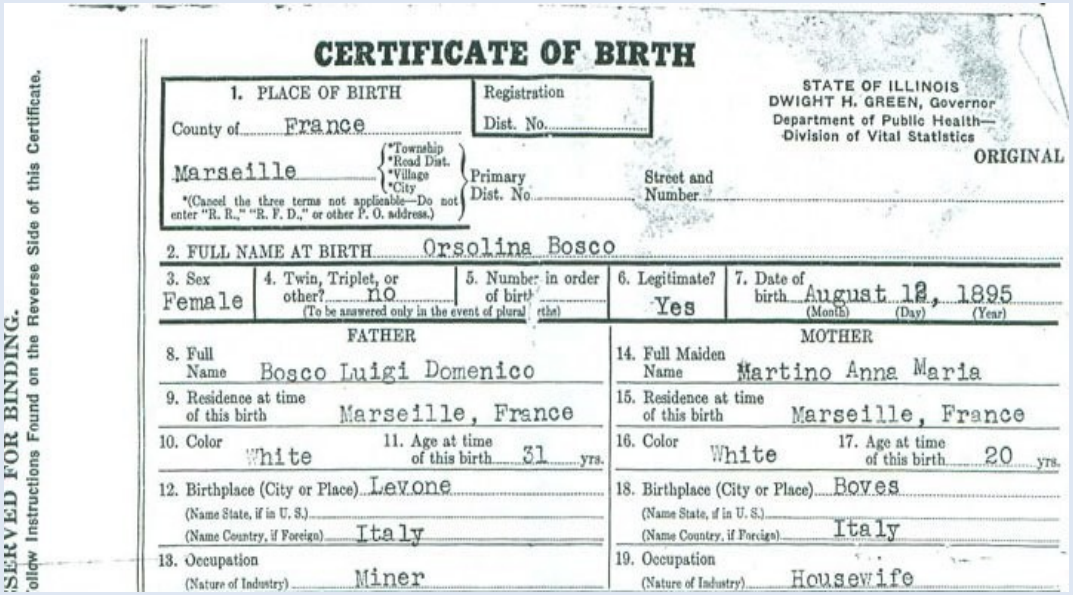
Principles of Provenance

eScience Institute Theme Opening Lecture

James Cheney
April 15, 2008

Provenance is...

- **Records** of origin, modification, influences
- Where something "comes from"



CERTIFICATE OF BIRTH

STATE OF ILLINOIS
DWIGHT H. GREEN, Governor
Department of Public Health—
Division of Vital Statistics

ORIGINAL

1. PLACE OF BIRTH — Registration Dist. No.
County of France
Marseille (*Township, *Road Dist., *Village, *City) Primary Dist. No. Street and Number
*(Cancel the three terms not applicable—Do not enter "R. R.," "R. F. D.," or other P. O. address.)

2. FULL NAME AT BIRTH Orsolina Bosco

3. Sex	4. Twin, Triplet, or other? (To be answered only in the event of plural birth)	5. Number in order of birth	6. Legitimate?	7. Date of birth (Month) (Day) (Year)
Female	no		Yes	August 12, 1895

FATHER	MOTHER
8. Full Name: Bosco Luigi Domenico	14. Full Maiden Name: Martino Anna Maria
9. Residence at time of this birth: Marseille, France	15. Residence at time of this birth: Marseille, France
10. Color: White — 11. Age at time of this birth: 31 yrs.	16. Color: White — 17. Age at time of this birth: 20 yrs.
12. Birthplace (City or Place): Levone; (Name State, if in U. S.); (Name Country, if Foreign): Italy	18. Birthplace (City or Place): Boves; (Name State, if in U. S.); (Name Country, if Foreign): Italy
13. Occupation (Nature of Industry): Miner	19. Occupation (Nature of Industry): Housewife

SERVED FOR BINDING.
Follow Instructions Found on the Reverse Side of this Certificate.

Provenance is...

- **Evidence** of authenticity, integrity, quality
- Certifies products of good “process”

Provenance is...

- **Valuable** because hard to collect, verify
- **Necessary** to assign credit

Provenance is...

- **Valuable** because hard to collect, verify
- **Necessary** to assign credit and blame

Why is it important for data?

- For traditional (paper) information:
  - Creation process leaves “paper trail”
  - Easier to detect modification, copying, forgery
  - Can *usually* judge a book by its cover
- For *electronic* information:
  - Often no such thing as a “bit trail”
  - Easy to forge, plagiarize, alter data undetected
  - Can't judge a database by its cover - *there isn't one*
- Provenance essential for judging quality of data

This talk

- Areas where provenance is *needed*
  - *bioinformatics & other curated scientific databases*
  - *workflow/grid/distributed computation*
- Why this is a hard problem
  - *& why principles/foundations need development*
- Areas of CS that can help
- Overview of **Principles of Provenance** theme

Relevance to eScience?

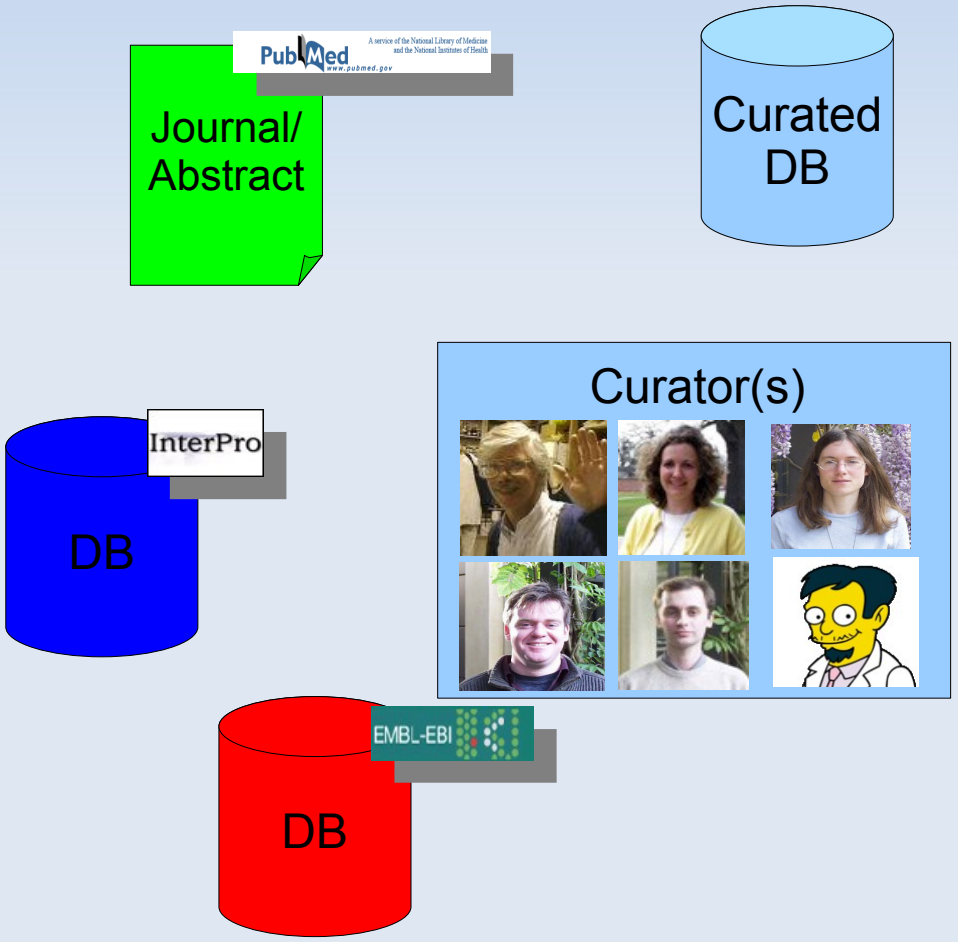
- Curated scientific (biological) databases
  - Manual curation process
  - Need provenance for quality control, accountability
  - Currently maintained by hand
- Scientific workflows. grid computation
  - Hides a complex execution process
  - Need provenance for reproducibility, efficiency
  - Currently provided by various customized systems

Biological databases

- *1000s* of specialized biological DBs
  - Independent
  - Heterogeneous
  - Change frequently
- Many *curated*
  - Expensive!

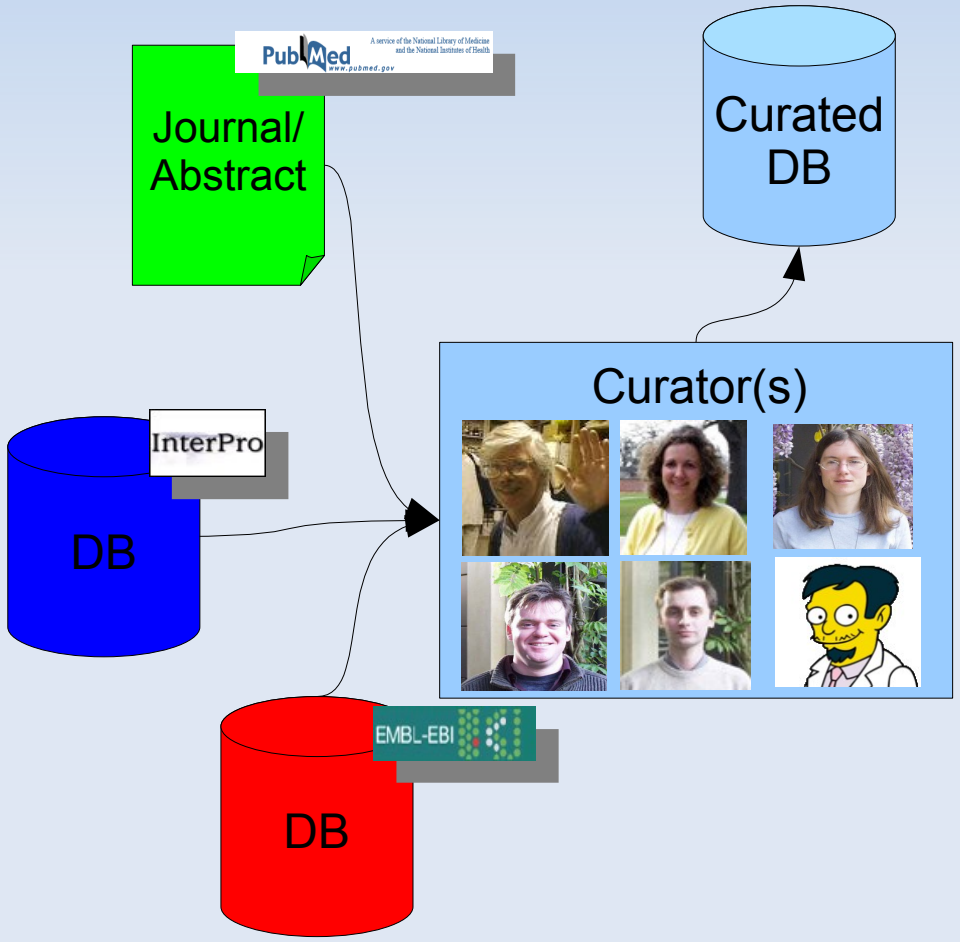
Curated databases

- Created by manual effort of scientists



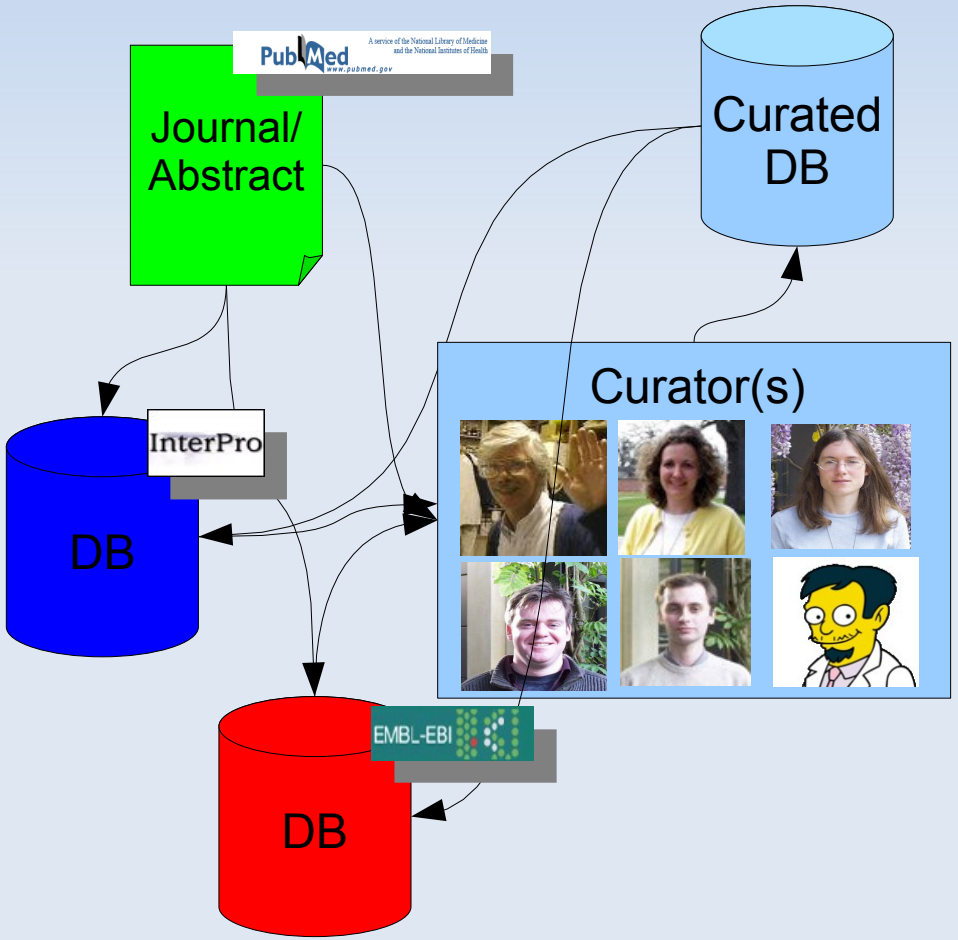
Curated databases

- Created by manual effort of scientists
- Curators copy from papers, other DBs



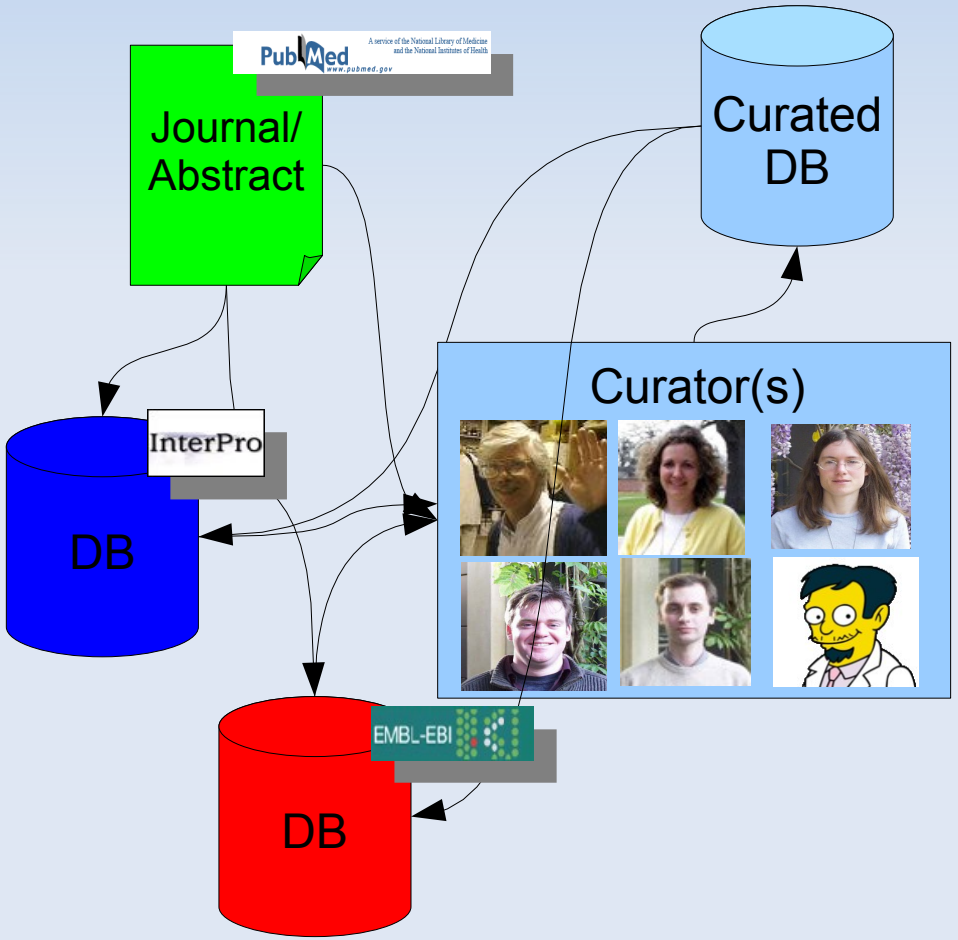
Curated databases

- Created by manual effort of scientists
- Curators copy from papers, other DBs
  - Which often copy from each other...



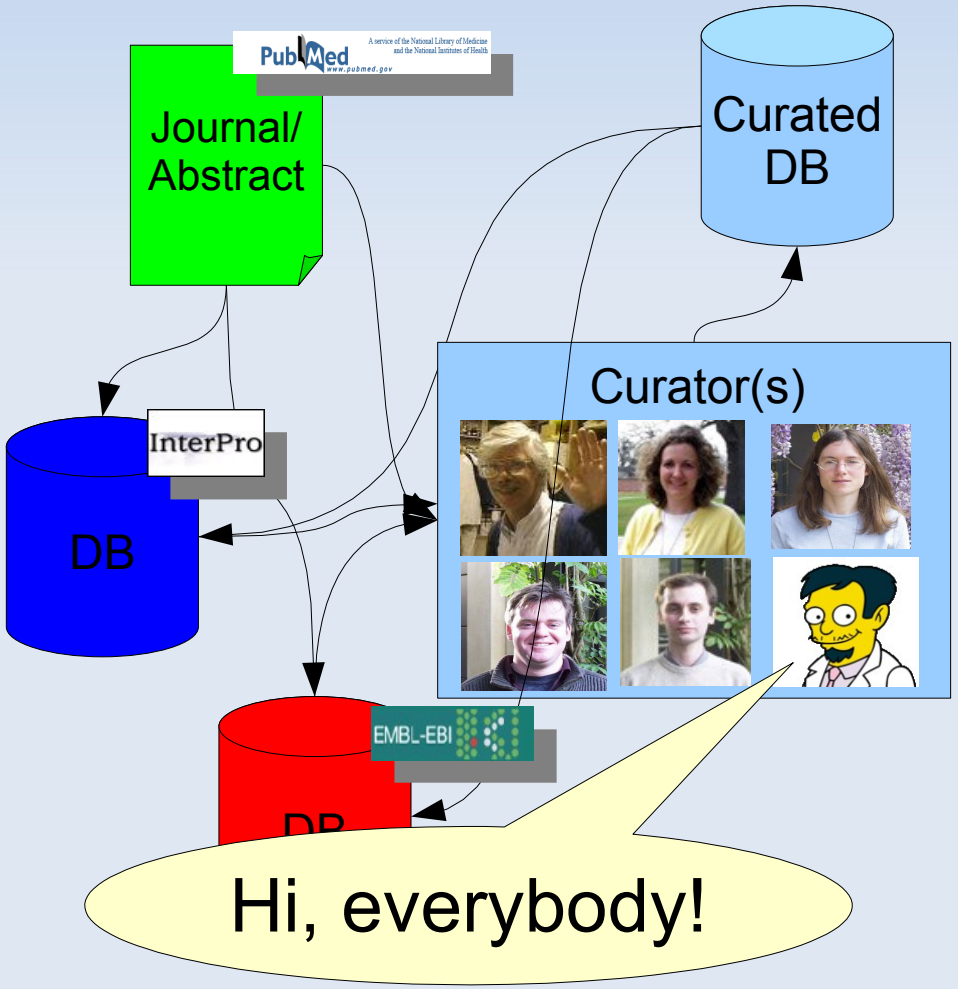
Curated databases

- Created by manual effort of scientists
- Curators copy from papers, other DBs
  - Which often copy from each other...
- Some sources unreliable



Curated databases

- Created by manual effort of scientists
- Curators copy from papers, other DBs
  - Which often copy from each other...
- Some sources unreliable
  - and some curators too

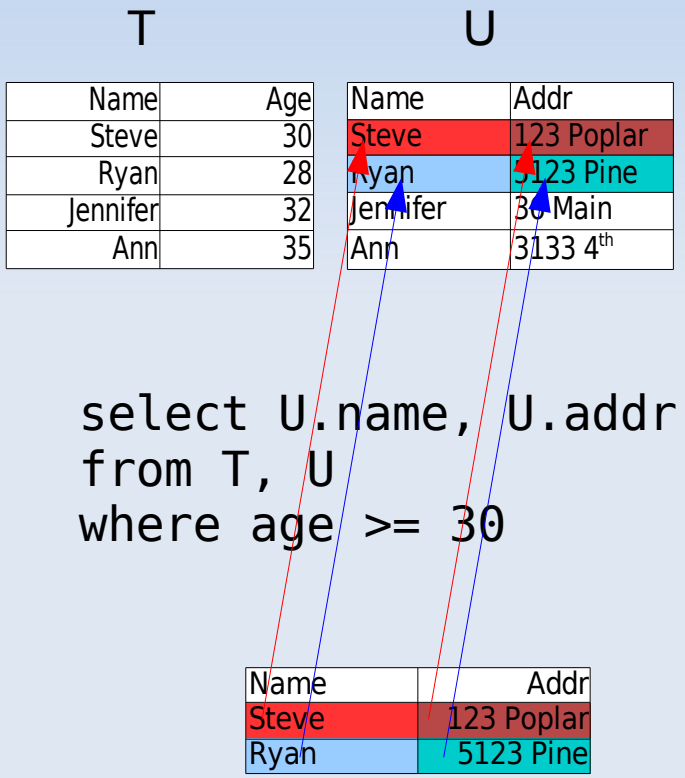


State of practice

- Scientists believe provenance essential for curated DBs
- Existing systems do not track provenance well
- Instead, curators currently do this *manually*
  - boring; waste of curators' valuable time (= $$)
- or using ad hoc, custom systems
  - few guarantees; lots of wheel reinvention
- Want to *automatically* record provenance

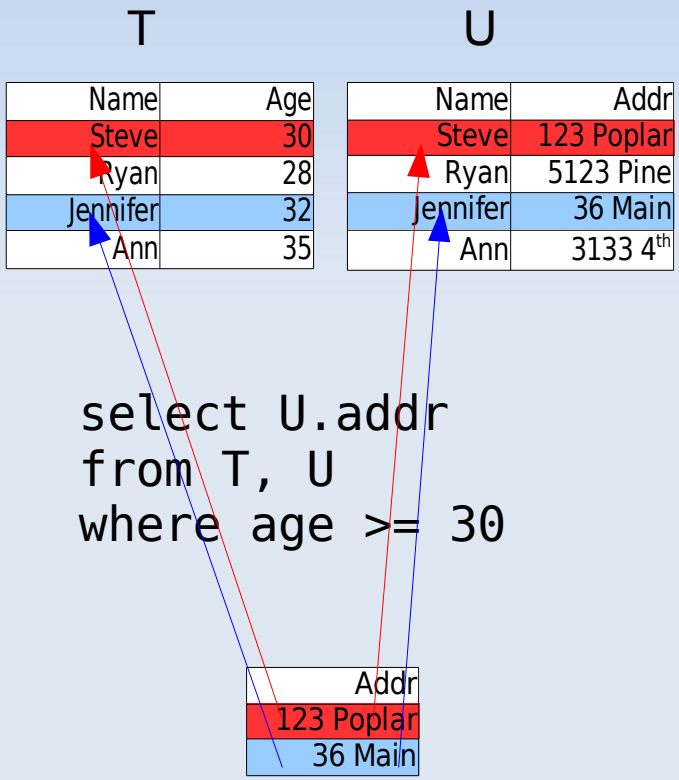
Where-provenance

- Where-provenance
  - Shows where data in each tuple was “copied from”
  - [Buneman, Khanna, Tan 2001]



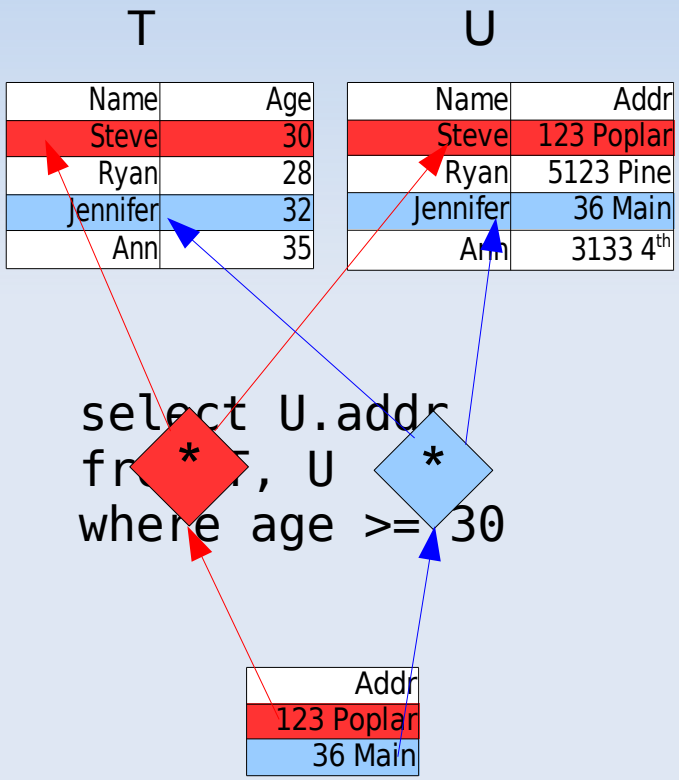
Why-provenance

- Why-provenance:
  - Shows sets of tuples “contributing” to result tuple
  - [Buneman, Khanna, Tan 2001]
- See also: “lineage”
  - [Cui, Widom, Wiener 2000]

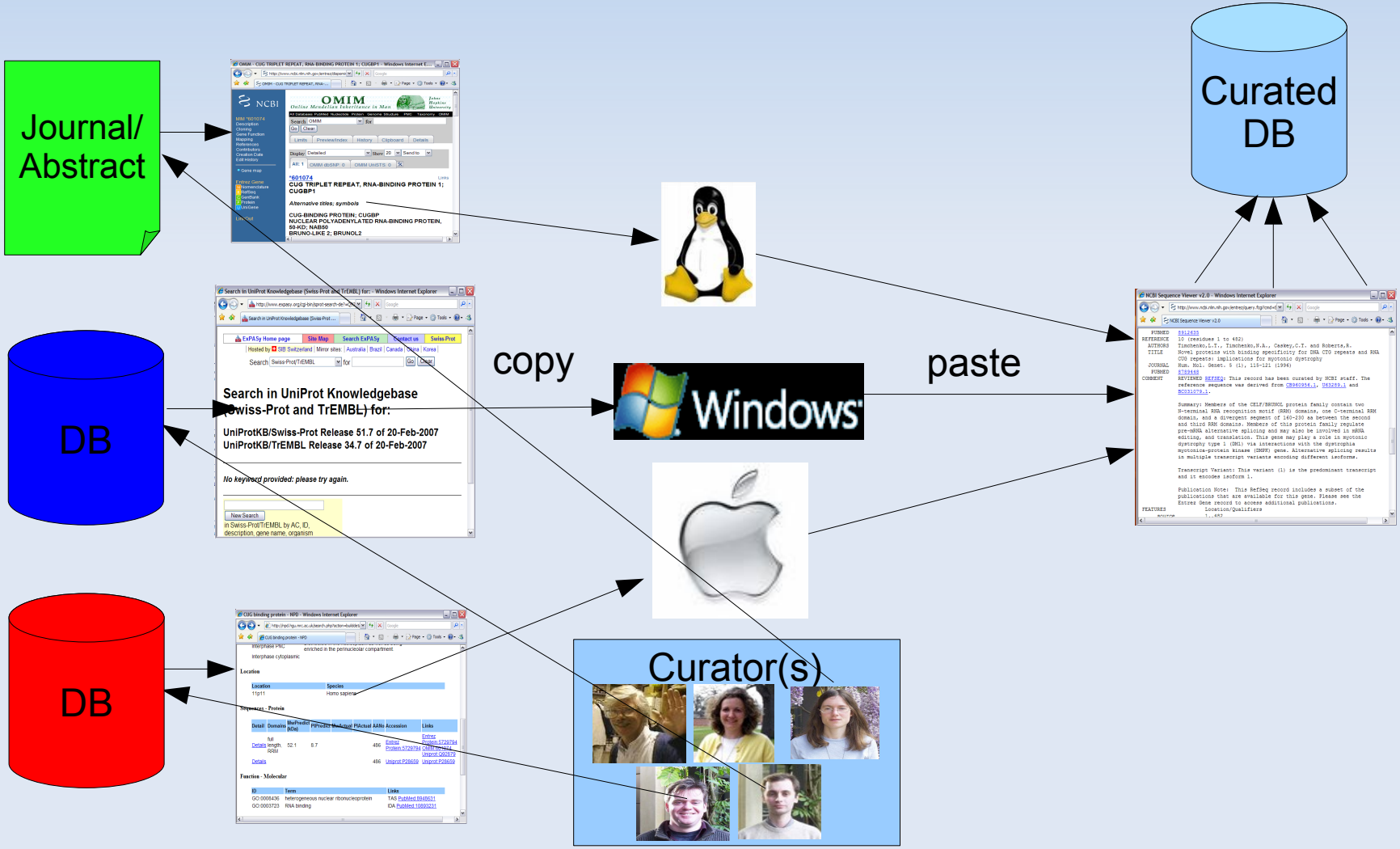


How-provenance

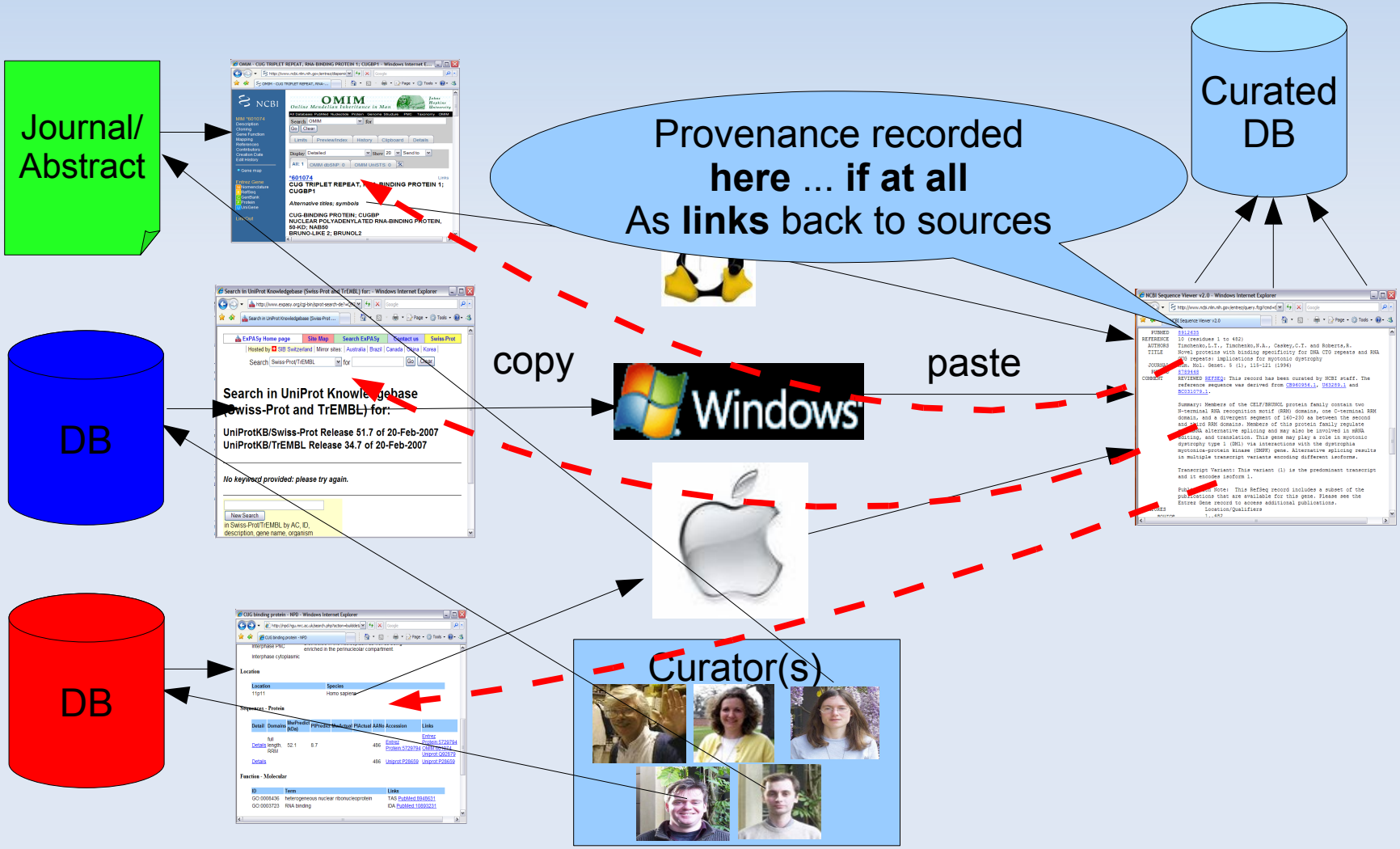
- How-provenance:
  - Gives “expression” showing “how” tuple was obtained from input
  - [Green, Karvounarakis, Tannen 2007]



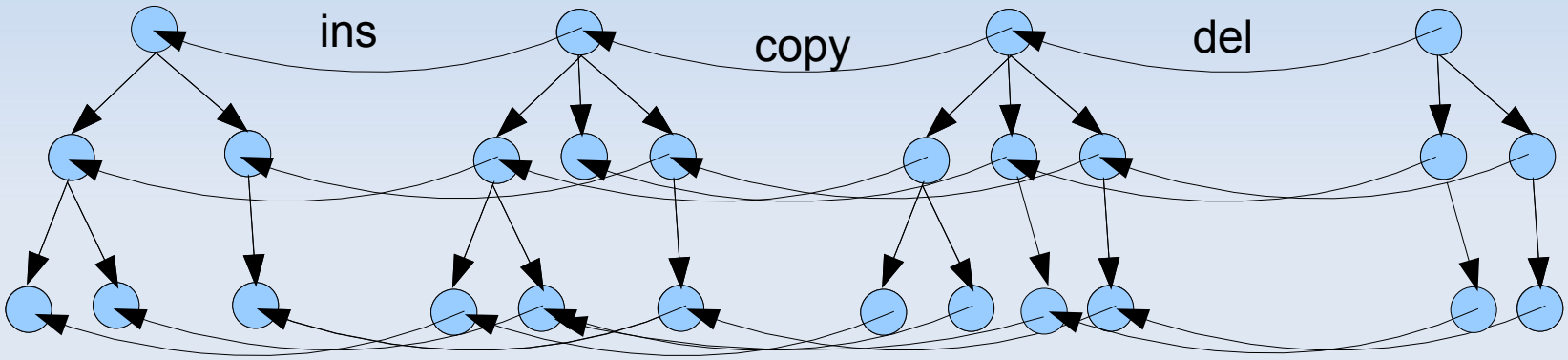
What curators do now



What curators do now



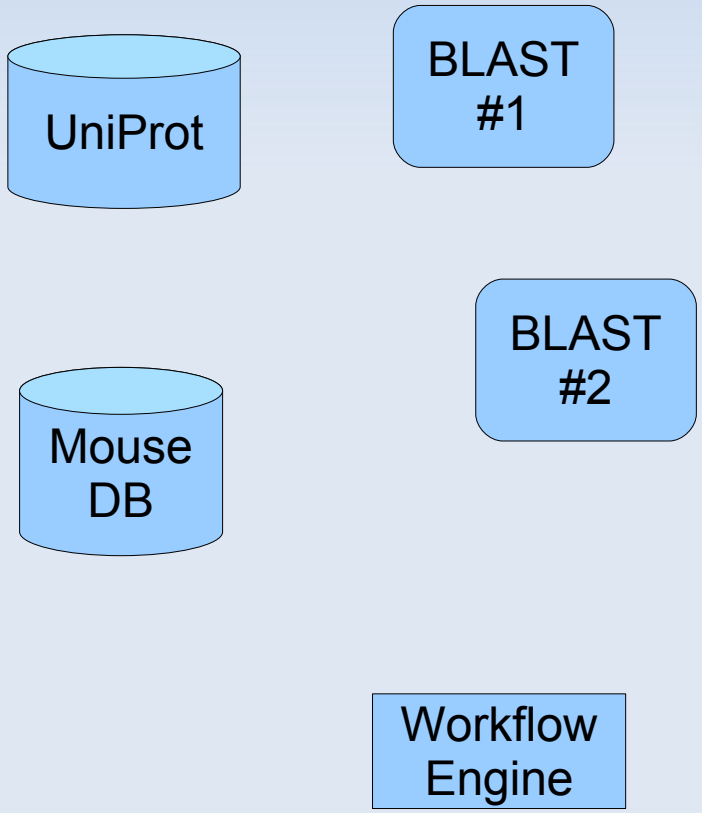
Copy-paste provenance



- First approach to provenance for *manually curated databases*
  - Sequence of inserts, deletes, copies
  - [Buneman, Chapman, Cheney 2006]

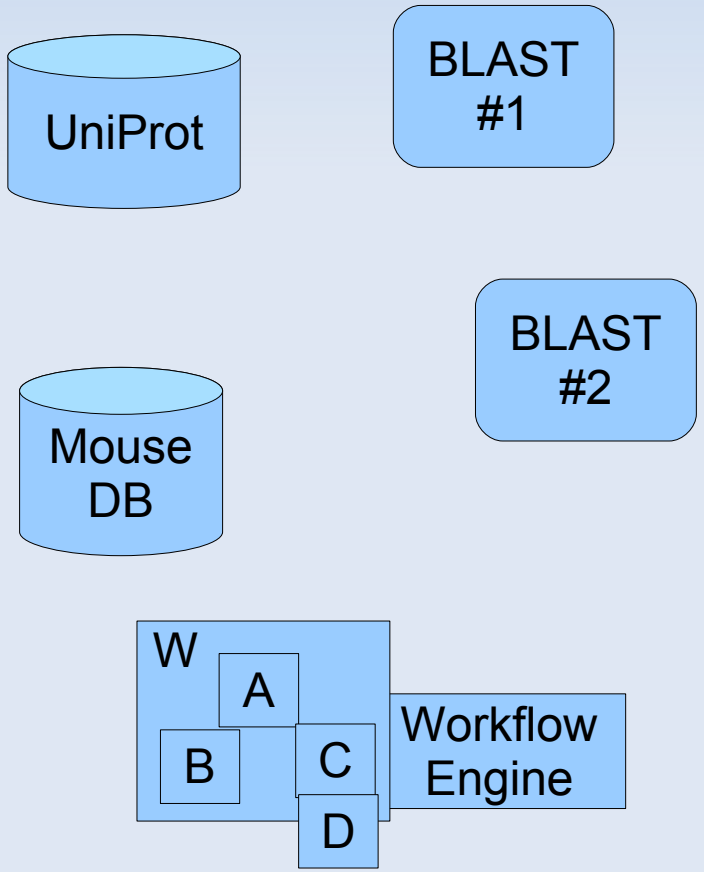
Workflows

- Computations packaged into *workflows*



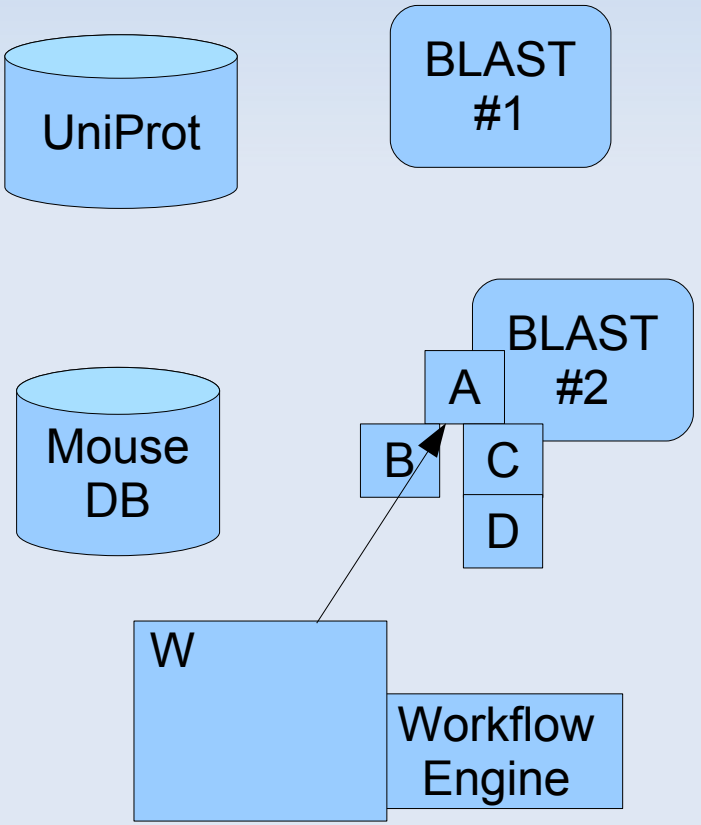
Workflows

- Computations packaged into *workflows*
- Workflow engine executes program



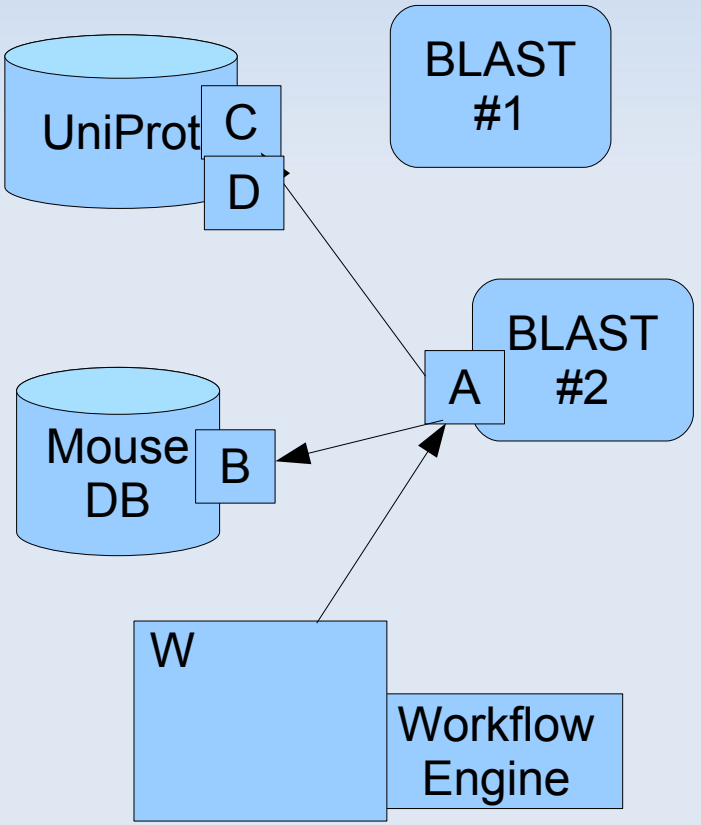
Workflows

- Computations packaged into *workflows*
- Workflow engine executes program



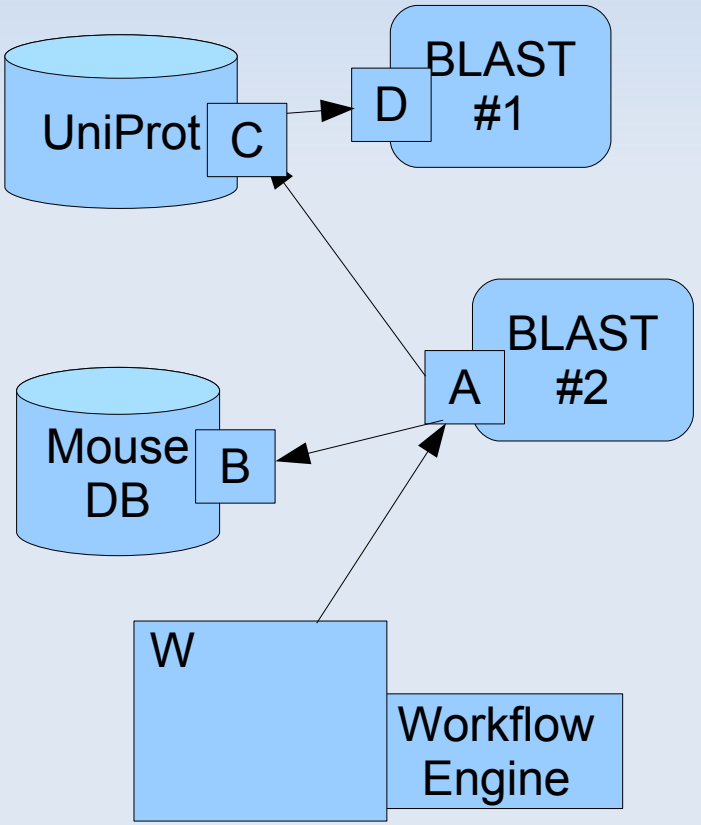
Workflows

- Computations packaged into *workflows*
- Workflow engine executes program



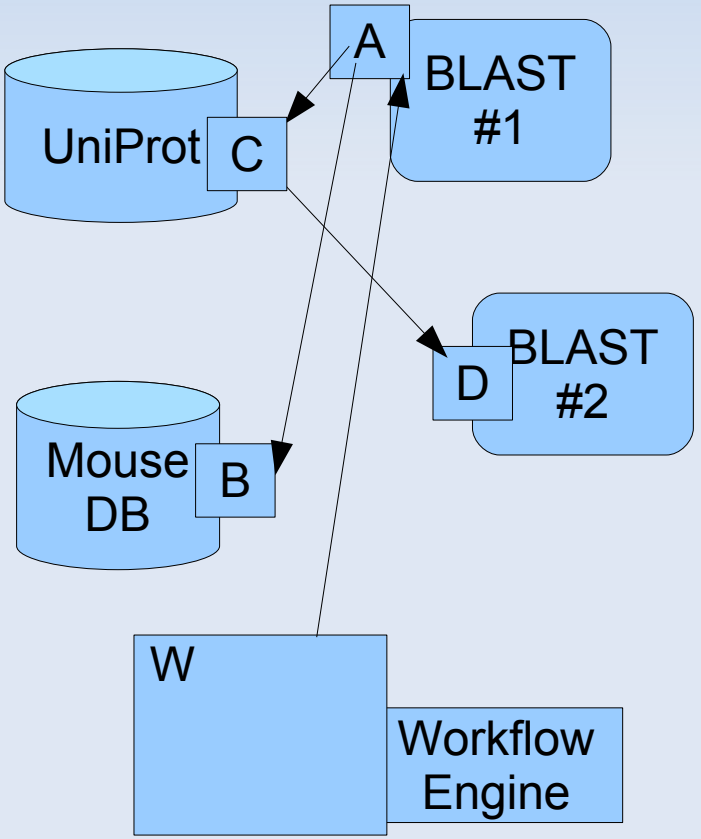
Workflows

- Computations packaged into *workflows*
- Workflow engine executes program



Workflows

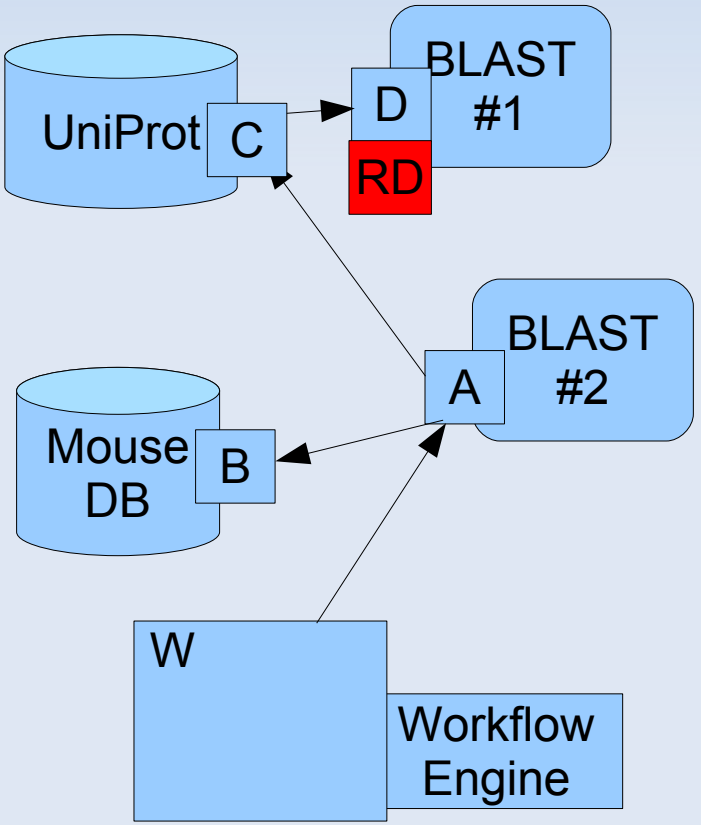
- Computations packaged into *workflows*
- Workflow engine executes program
- A workflow can be executed in different ways



Workflow provenance

- Provenance should show *what actually happened*
  - so that we can repeat computation
  - or track down bugs

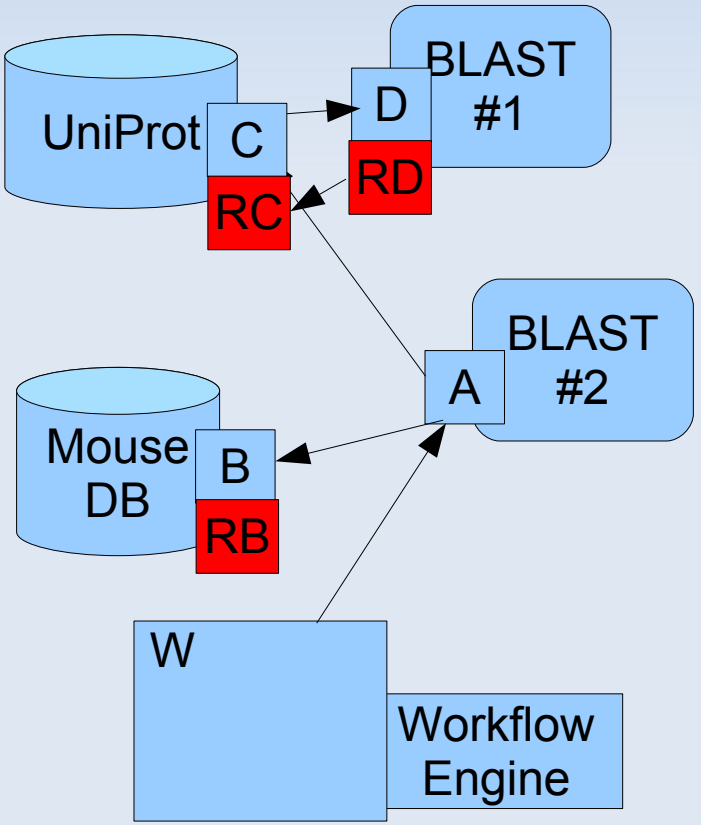
```
RD := Blast_1(QD)
```



Workflow provenance

- Provenance should show *what actually happened*
  - so that we can repeat computation
  - or track down bugs

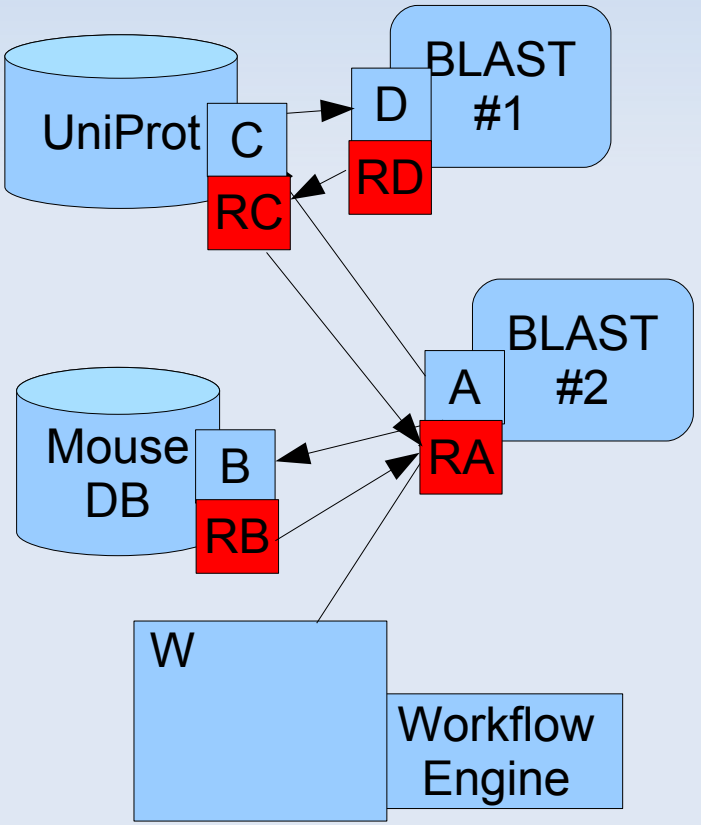
```
RD := Blast_1(QD)
RC := UniProt(QC,RD)
RB := MouseDB(QB)
```



Workflow provenance

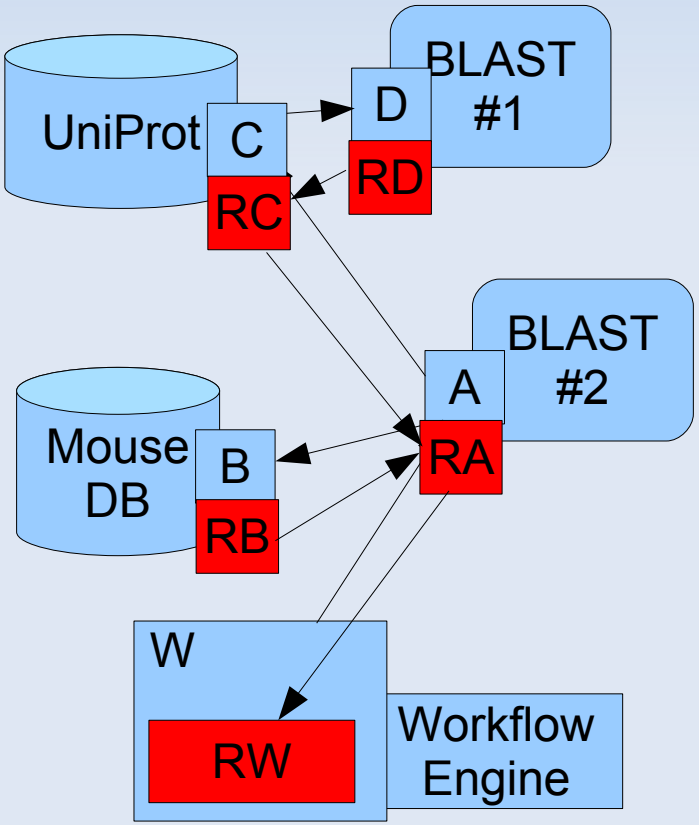
- Provenance should show *what actually happened*
  - so that we can repeat computation
  - or track down bugs

```
RD := Blast_1(QD)
RC := UniProt(QC,RD)
RB := MouseDB(QB)
RA := Blast_2(QA,RB,RC)
```



Workflow provenance

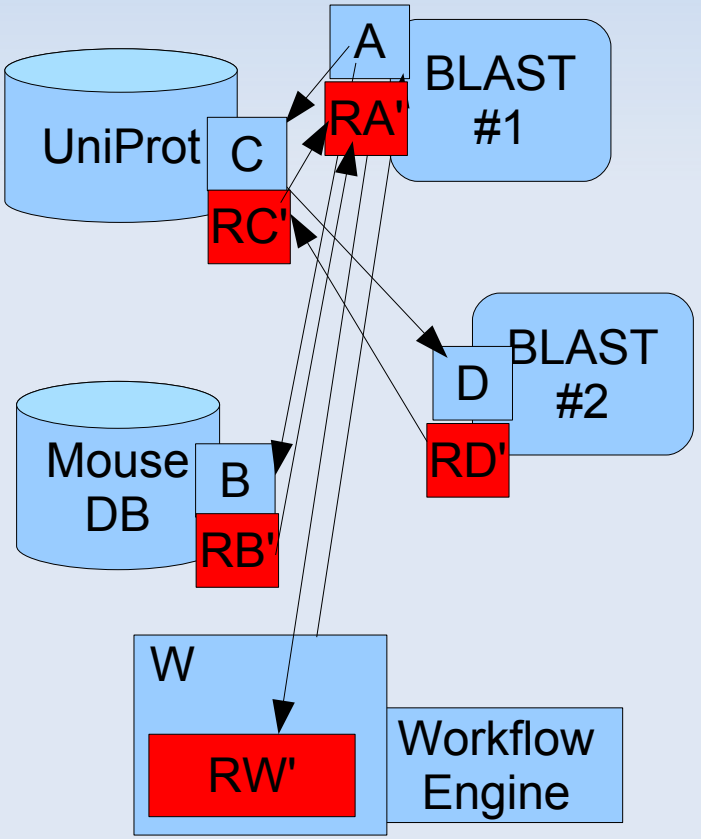
- Provenance should show *what actually happened*
  - so that we can repeat computation
  - or track down bugs




```
RD := Blast_1(QD)
RC := UniProt(QC,RD)
RB := MouseDB(QB)
RA := Blast_2(QA,RB,RC)
RW := Format_Results(RA)
```

Workflows

- Provenance should show *what actually happened*
  - so that we can repeat computation
  - or compare two runs




```
RD' := Blast_2(QD)
RC' := UniProt(QC,RD')
RB' := MouseDB(QB)
RA' := Blast_1(QA,RB',RC')
RW' := Format_Results(RA')
```

State of practice

- Useful for efficient recomputation
  - May also reassure users
- *Coarse-grained* model
  - Computation steps are *"black boxes"*
  - Doesn't deal well with structured data (DBs, XML)
  - Can't track errors at data level
- Does not deal with *updates* to data
  - What if DB gets updated in the middle of a workflow?

The *real* problem

- Many *systems*, few *specifications*
  - what makes a technique correct?
  - how will we know when we've solved the problem?
  - how to generalize existing techniques?
- Little understanding of
  - “correctness”, “completeness”
  - suitability for given application
  - relationships among techniques

Why is this hard? (1)

- Users can't always explain
  - what they *want*
  - what they *need*
  - these are **not** necessarily identical!
- **This is not a criticism of users**
  - This is our (computer scientists') job
  - Not enough **interaction** between CS research and potential users

Want vs. need

- "Users usually ask for what they *want;* our job is to figure out what they really *need*."

  --Jim Gosling, paraphrase

- If scientists **want** provenance information...
- What **needs** drive this want?
  - Data integrity/quality filtering?
  - Error correction/propagation?
  - Credit/citation?
  - Efficiency?

Can needs be met?

- Most provenance information currently is
  - added manually
  - or maintained by customized systems
- Reasonable to automate/generalize?
  - Is “intelligence” needed?
  - maybe this is an AI problem...
- Automatic, general purpose systems must:
  - **provide clear guarantees**
  - **elucidate responsibilities of users**

Why is this hard? (2)

- In practice, several **working definitions** used
  - **DBs**: polygen, lineage , why, where, how
  - **Workflows**: variants of "directed acyclic graphs"
- But little agreement about **definitions, goals**
  - what ***is*** foo-provenance?
  - what is foo-prov ***for***?
- # definitions ≈ # systems/papers
  - Not a good thing! Tower of Babel effect

What *should* we be storing?

- Store “everything”
  - But this is impossible
  - Unbounded amount of information we could track
- Store whatever the users (say they) want
  - Difficult to please everyone
  - Hard to justify adding to general purpose systems
  - May not address (unstated) needs
- Store “enough”
  - What is “enough”?

How much is “enough”?

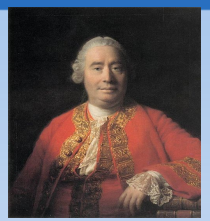
- This depends on what we want to do with the information
  - *But we don't always know this in advance!*
- Thus, provenance should be as general as possible
  - “Suitable” for many applications
  - While still being “manageable”
- Not sure what this means yet
- **This** is why foundational study is needed!

Why is this hard? (3)

- **Claim:** Work on provenance treads in deep philosophical waters
  - Seldom recognized!
- **Causality, influence, dependence**
- **Explanation, justification**
- **Knowledge, belief**
- are **all** nontrivial concepts
  - many with relatively little **formal understanding**

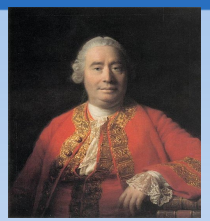
Some Pseudo-Definitions

- Papers on provenance are littered with **pseudo-definitions** such as
  - *"the **process** which led to a result"*
  - *"a summary of the **history and context** of the data"*
  - *"the parts of the input that **influenced** (or that **explains**) a part of the output"*
  - "*the part of the input that shows where a part of the output **came from***"
  - *"a **causal** graph that shows how a result was computed"*



Hume on causality

- What does it mean to say "A **causes** B"?
  - e.g. Rain **causes** the sidewalk to get wet
  - e.g. Flame **causes** heat
- WWHD? (What would Hume do?)
- Hume said:
  - Thus we remember to have seen that species of object we call **FLAME**, and to have felt that species of sensation we call **HEAT**. We likewise call to mind their constant conjunction in all past instances. Without any further ceremony, *we call the one **CAUSE** and the other **EFFECT**, and infer the existence of the one from that of the other*.

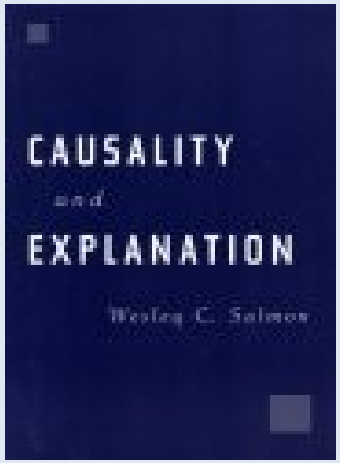


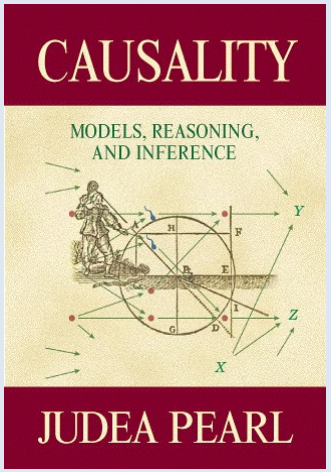
Hume on causality

- Usual interpretation of Hume's views:
  - **causal relationships** are **not** real things that we can directly perceive
  - Instead, causality is (only) **psychological!**
  - A **causes** B means *in our experience whenever we observe A, we also observe B*
- Since then most scientific disciplines have **eschewed causality** in favor of **correlation**
- *Is (causality-oriented) provenance legitimate?*

Modern view

- Causality = correlation view is changing.
- Causality is neither settled nor trivial.
- Much recent activity in philosophy, AI:





Relevance to Computer Science?

- Databases
  - Data models, query languages, expressiveness
- Concurrency/distributed computing
  - Workflow models, causality, explanation
- Programming Languages/Software Engineering
  - Bidirectional, adaptive programming, dependency analysis, program slicing
- Systems/Security
  - Integrity, (information flow) security, trust

Provenance in databases

- Provenance in database **queries/views**
  - **Polygen,** source tagging [WM 1990]
  - **Lineage** [CWW 2000], Trio [Widom et al. 2005/6],
  - "**why**" and "**where**" [BKT 2001]
  - Semiring-valued relations and "**how**" [GKT07]
- Provenance for database **updates**
  - Simple "**copy-paste**" curation model [BCC06]
  - Extended to **SQL-like updates** [BCV07]

Provenance in workflows

- Focus on **system development**
  - based on a variety of ad hoc provenance models
- Many similar efforts in
  - **Geospatial inf. syst.** [Bose, Frew 2005, ...]
  - **Kepler** system [Ludaescher et al.]
  - **PASOA** project [Groth, Moreau, Miles]
  - **Taverna**, myGrid [Goble et al.]
  - **Provenance Challenge(s)**, **Open Provenance Model** [Moreau, Freire, et al.]

Provenance in PL

- Provenance-like concepts widespread in PL
  - **line-number maintenance** for error messages
  - **debugging symbol** propagation
  - **dependency analysis** and **program slicing**
    - [CAA07] used to define a new form of DB provenance
  - **adaptive functional programming** [ABH02,...]
    - similar to **view maintenance** in DBs
  - **bidirectional computation** [Foster et al. 05,...]
    - similar to **view update** in DBs

Provenance in systems/security

- **Provenance aware storage systems (PASS)**
  - Records "provenance" at level of file system/OS calls [Muniswamy-Reddy et al. 2006]
- **Information flow security** [e.g. Myers/Liskov 97]
  - Ensure that **low-security outputs** cannot depend on **high-security inputs**
  - Connections to **why-prov, dependency analysis**
- **Trust and Security in Virtual Communities**
  - concurrent eSI Theme, Jan-Dec 2008

Problems

- Provenance-like ideas occur in (or may benefit from) many areas of CS
- But there is little **contact** between these areas
  - and little **recognition** of provenance as a deep and interesting **cross-disciplinary** problem
- And often little contact between CS and working scientists
  - hence, research efforts may **miss the mark**

What we already did

- Last November, held a 1.5-day **Workshop on Principles of Provenance** (in Edinburgh)
  - 12 invited/contributed abstracts & talks
  - 25-30 participants
- My observations:
  - Many interesting discussions; not enough follow-up.
  - Would have been **great** to keep the participants together longer...
  - and **focus on** certain **key areas**

Principles of Provenance Theme

- Running **April 2008 - March 2009**
- **Goals:**
  - Support **focused research** into **foundations of provenance in computer science**
  - **Bridge CS/eScience gap**
  - Identify **key problems** and set **research agenda**
  - **Disseminate results** and **incubate further research programs/funding proposals**
- Led by James Cheney, Peter Buneman, Bertram Ludaescher (U. C. Davis)

What we will do

- We plan to host four **small symposia** focusing on these areas
  - 3-5 **CS researchers** working on provenance
  - 3-5 **working scientists** with provenance needs
  - One day of **public lectures and discussion**
  - Follow-up **research/collaboration time**
  - Nominal goal of drafting **white paper**
    - summarize state of art and future directions for each area
  - But hope for more (collaboration/papers/proposals)

What we will do (2)

- We also plan to organize another **Principles of Provenance Workshop**
  - Present results, white papers
  - Solicit contributions from community
  - Hope to have (tentative)
    - peer-review
    - formal publication
    - collocation with external conference
  - Target: Q1-2 2009

Current plan

- May 19-23, 2008: Provenance in **Databases**
  - public talks/discussion on **May 21**
- late July 2008: Provenance in **Workflow, Grid, and Distributed Computation**
- Q4 2008: Provenance in **Programming Languages and Software Engineering**
- Q1 2009: Provenance in **Operating Systems and Security**
- Q2 2009: Follow-up **workshop**

Current plan

- **Additional speakers** as opportunities arise
  - Stuart Madnick (MIT)
    - TBA, April/May
  - Umut Acar (Toyota Technological Institute, Chicago)
    - “Adaptive functional programming”, late May
- Suggestions welcome!

Conclusions

- Provenance
  - Important for scientific record(s)
  - few clear definitions/problem statements
- Principles of Provenance theme will...
  - develop **foundations** of subject
  - **build awareness** of problems
    - & connect provenance in different CS disciplines
  - **engage with** scientists/DB curators
    - to make sure we're solving the right problems

More information

http://wiki.esi.ac.uk/Principles_of_Provenance