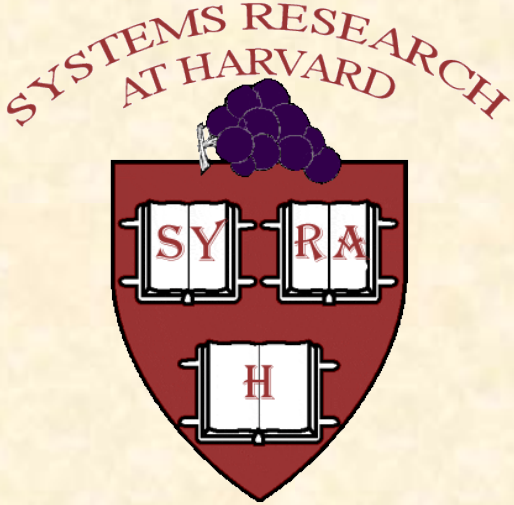
World Domination Through Provenance



Margo Seltzer
Harvard School of Engineering and Applied Sciences

April 3, 2013

What Should World Domination Look Like?



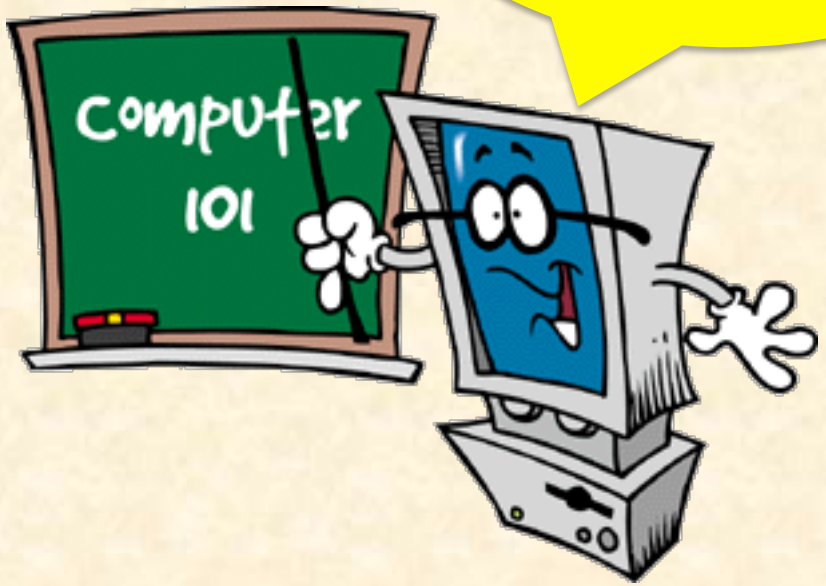


What Should World Domination Look Like?

Where did this file come from?



It came to you yesterday from your dept. chair marked urgent.

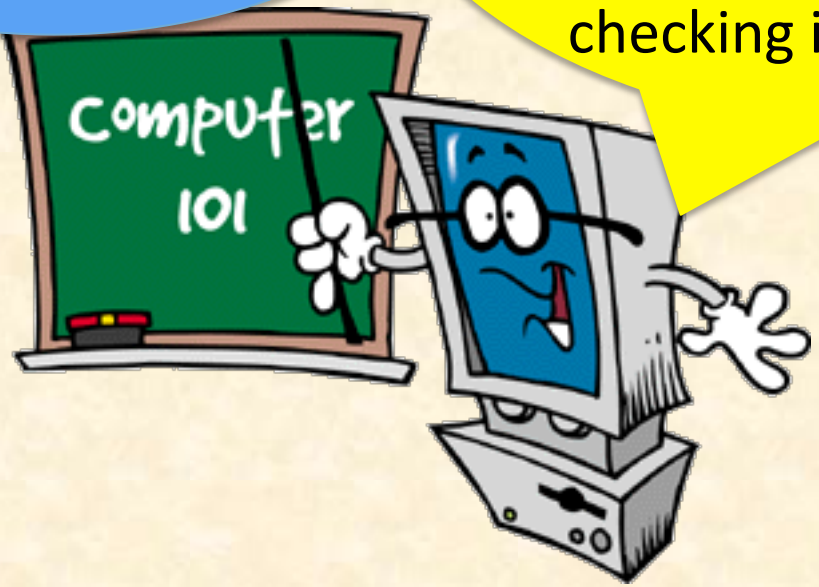


What Should World Domination Look Like?

Oh @#$%^. Someone changed something in our data and now our graphs are corrupt. What happened???

Relax! Amos changed the format of our data files; I've put the old data files back and sent him an email asking him to fix the plotting scripts before checking in new data files.





What Should World Domination Look Like?

What is the risk profile of this mortgage backed instrument?





What Should World Domination Look Like?

I work on Provenance!

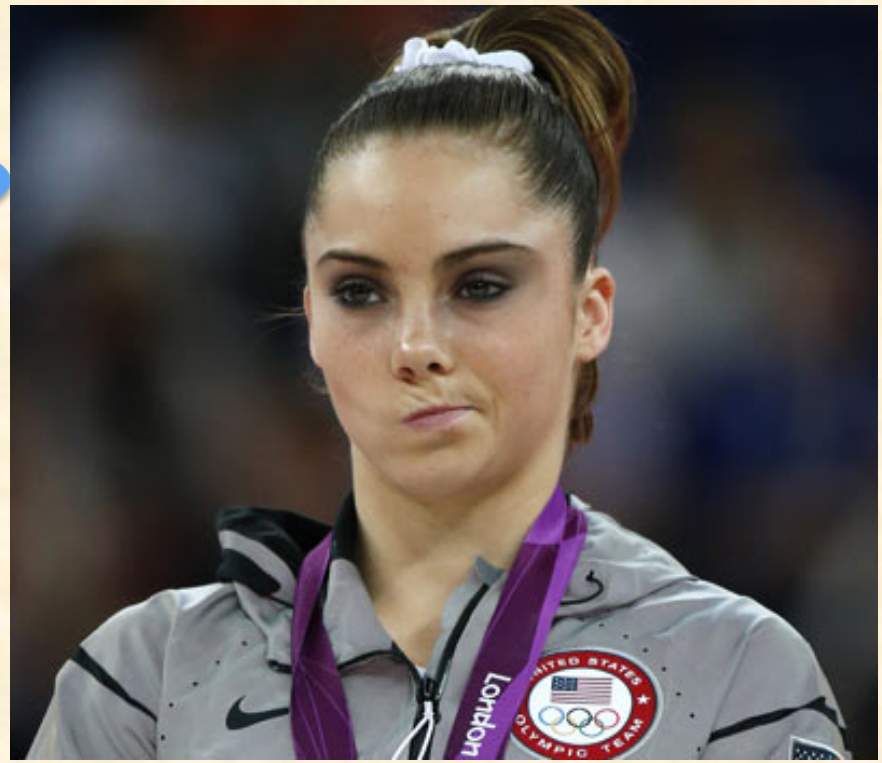


You're our hero!



What Does the World Look Like Today?



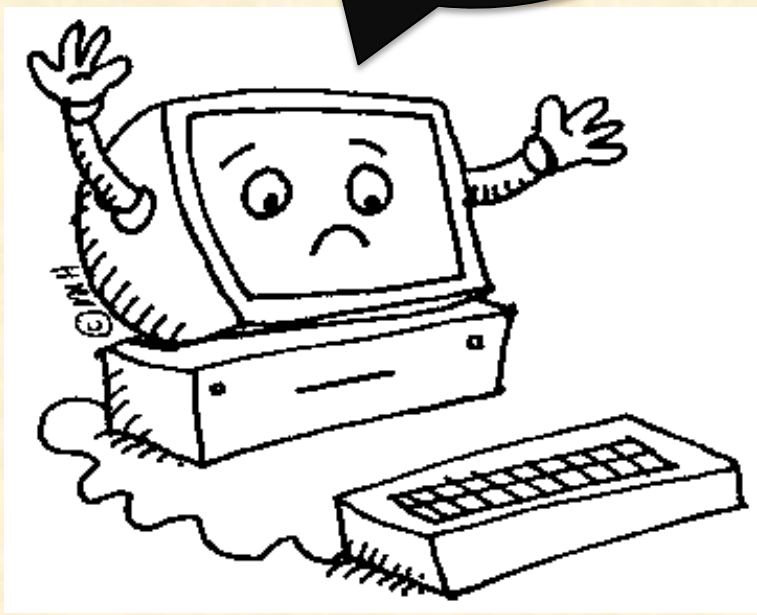


What Does the World Look Like Today?

Where did this file come from?



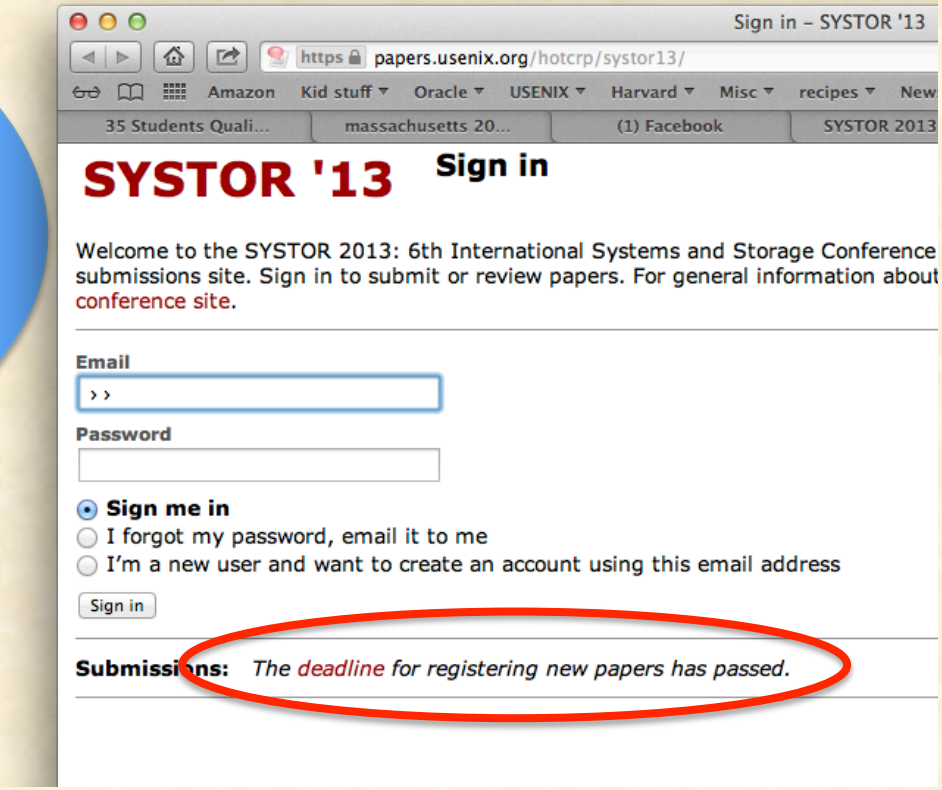
Um, um, it was created yesterday. Is that helpful?



What Does the World Look Like Today?

Oh @#$%^. Someone changed something in our data and now our graphs are corrupt. What happened???





What Does the World Look Like Today?

What is the risk profile of this mortgage backed instrument?



No Problem!

BEAR STEARN

LEHMAN BROTHERS 150 1850-2000 years A TRADITION OF INNOVATION

Merrill Lynch

Goldman Sachs

Morgan Stanley

What Does the World Look Like Today?

I work on Provenance!





World Domination?

NOT!

What Impact Have We Had?

- Papers Published: Perhaps a few thousand?
- Where and Why Provenance : 729 citations
- Standards Specifications: 13
- NYT Headlines: 0
- Products: 1?
- Turing Awards: 0

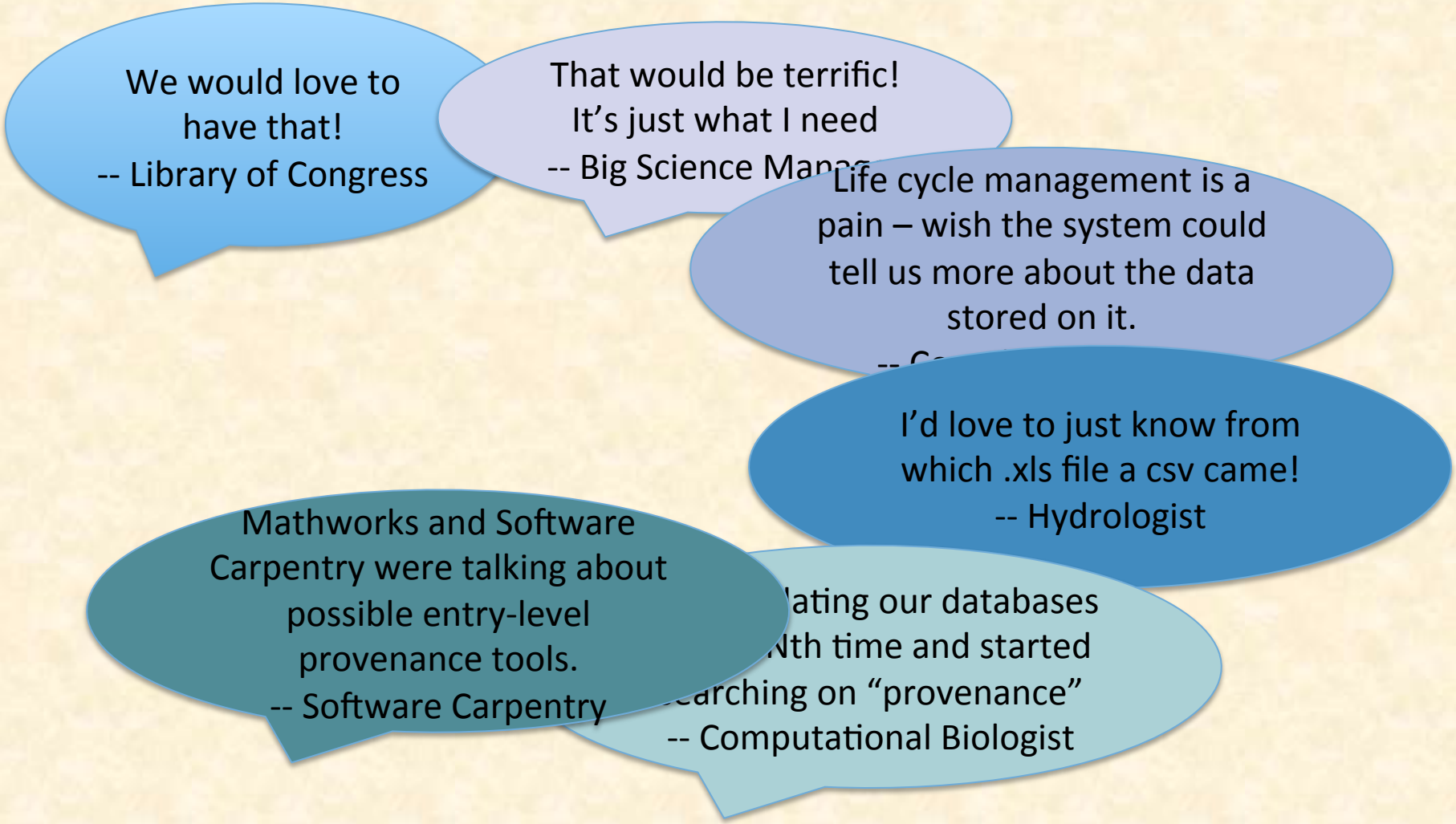
Where Should We Be?



So, What's the Problem?

- People don't really care about provenance?
- It's like security: people want it but won't pay for it?
- General provenance doesn't help: people want provenance specific to their problem?
- Provenance is like insurance: you never really want it until it's too late.
- We're giving them the wrong thing?

People Don't Care?



People Want it but Won't Pay for It?

- Maybe, but …

- There really aren't any products to buy, so we don't even know.
- (So we can't use this as an excuse.)

General Provenance is Not Useful?

- Maybe general purpose provenance solutions aren't useful?
- Perhaps, in order to be useful, the provenance solution must be completely embedded in the application.
- Do we see evidence of this?
  - Some application areas are pretty heavily vertically integrated (e.g., health care).
  - Users are tied to their own tool sets.
    - Those who use R, use R.
    - Those who use Python, use Python.
    - Those who use MySQL, don't use Postgres

Provenance as Insurance?

- When you gather it, provenance is pure overhead.
- You don't know that you want provenance until you need it, and then it's too late.
- Why do people buy insurance?
  - Required [car insurance, mortgage insurance].
  - Down side is huge [fire insurance].
- But in some cases, people don't think of it: most people do not carry renter's insurance.
- Maybe provenance is the renter's insurance of the digital age? Why???

We're Giving Them the Wrong Thing?

- What have we given them?
  - A fragile kernel module
  - A library
  - Workflow engines
  - Relational databases
- But …
  - No one likes either fragility or kernel modules.
  - Libraries aren't bad if I'm looking for them.
  - If I want a workflow engine, I get a workflow engine, but if I want provenance, I may not want a workflow engine.
  - Science seems to be a common target use case for provenance, but most scientists don't use databases.

So, What's the Problem?

- People don't really care about provenance? **Definitely Not**
- It's like security: people want it but won't pay for it? **Maybe. Too early to tell.**
- General provenance doesn't help: people want provenance specific to their problem? **Maybe**
- Provenance is like insurance: you never really want it until it's too late. **Maybe**
- We're giving them the wrong thing? **Probably**

What Should we Give People?

- Something we have used ourselves
- Solutions
- Friendly interfaces
- A long term plan
- Insurance
- Support
- An open mind

Eat our own Dog Food

- You cannot expect other people to use the things you build if you do not use them yourself!
- We are all scientists
  - We manage data
  - We sometimes need to understand where our data came from
  - We should be using our own tools

Dog Food Implications



Solutions That Work For Users

"One of the main problems in software design is that the users and the computer scientists do not understand one another."

" However, even more to the point is that researchers work the way they do for a reason, and adding extra obstacles to the procedures that they carry out would be a terrible design decision, and probably more of a hindrance to their research than a help."

Solutions Not Tools

- We give people tools
  - No one likes to learn new tools
  - People who are getting real work done aren't going to change their environment just to get provenance.
- Work closely with real users.
- Figure out what they need, what they want, and what they'll tolerate.
- Then just do it!

Friendly Interfaces

- For some users that means GUIs; for others it means command line.
- Meet the user where s/he works.
- We had one group who really wanted some command line options to `find`.
- Others wanted a GUI
- Others wanted a browser
- None had really thought about truly large collections.

A Long-Term Plan

The first is that experiments
[in Bioanthropology] may span
years, even decades.

- We (researchers) think small and short term.
- We don't necessarily think of the provenance that collects over the long term, but that's what our users are thinking.

Insurance

Researchers work with primary data: the fact that metadata might become useful in the future will not provide them with enough of an incentive to collect it, if collecting it will take away time and resources that they would spend on their primary work.

Support

- If people use our stuff, we have to be responsive!
- What would outreach look like?
- Should we have a provenance community presence? A blog?

An Open Mind

- We know a lot about provenance.
- We've thought a lot about provenance.
- BUT – our users know what they need.
- We have to believe them.
- We may not know best …

The Challenge

- Let's put together an event with:
  - ~10 provenance weenies
  - ~10 users who say they want provenance
- Outcome:
  - Commitments from provenance weenies for real live deliverables
  - Commitments from users that they will actually use what we build

Here is Where We Want to Be

What we did: Technology

- Systems: we deployed real systems
  - We used them
  - We made others use them.
- Formalism: we developed precise solutions
  - We applied theory to real systems; real systems relied upon formalism to define correctness.
- Security: we addressed the tension
  - Provenance provides accountability and must be available
  - Provenance is powerful information and must be carefully protected

TaPP: February 2019 43

What we did: Standardization

- Patience
- We let the practice evolve so we knew what needed standardization.
- We focused on interoperability.
- Standardized:
  - Protocols
  - Query methods

TaPP: February 2019 44

Where We Are

- Papers Published: Perhaps a few thousand?
- Where and Why Provenance : 729 citations
- Standards Specifications: 13
- NYT Headlines: 0
- Products: 1?
- Turing Awards: 0

A Call to Action

- Let's hold the challenge suggested
- Let's start a community site with:
  - Information
  - Tools
  - A Blog
  - Success stories (that means we have to have successes)
- Let's get ourselves into the NYT