

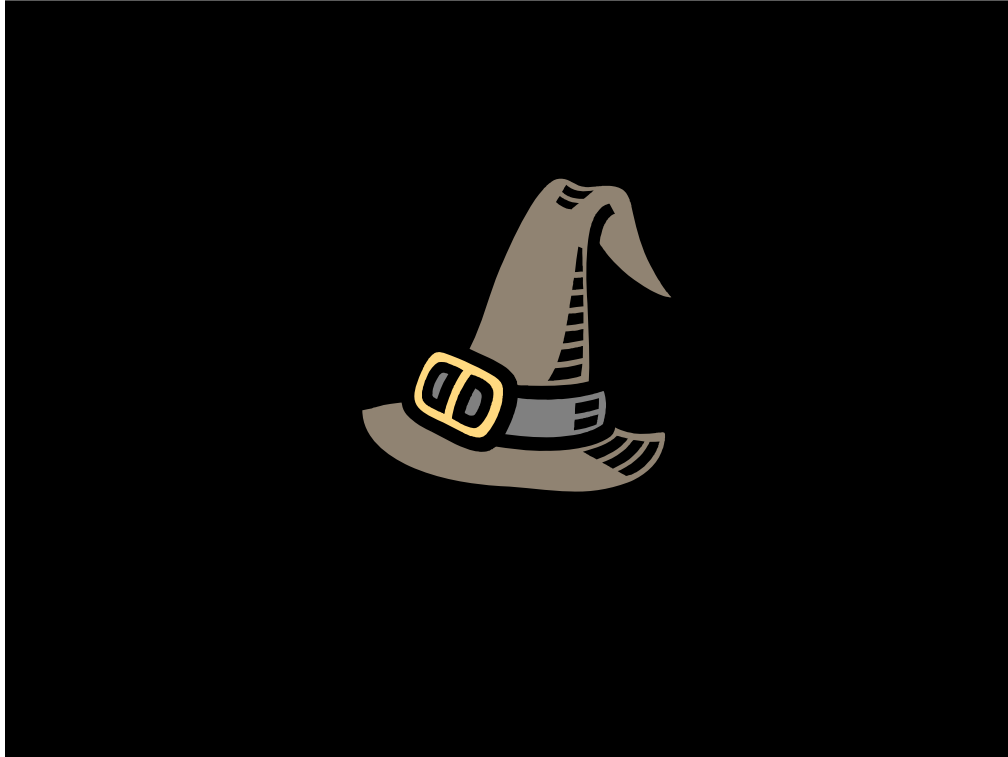
Why study Data Sharing?

(+ why share your data)

Heather Piwowar
DBMI Colloquium, March 28 2008
University of Pittsburgh



My name is Heather, and as many of you know my area of research is the evaluation of biomedical Data Sharing and Reuse. Today I'm going to be talking about what this means, why it is important, some of what we've learned, and what we still don't know.

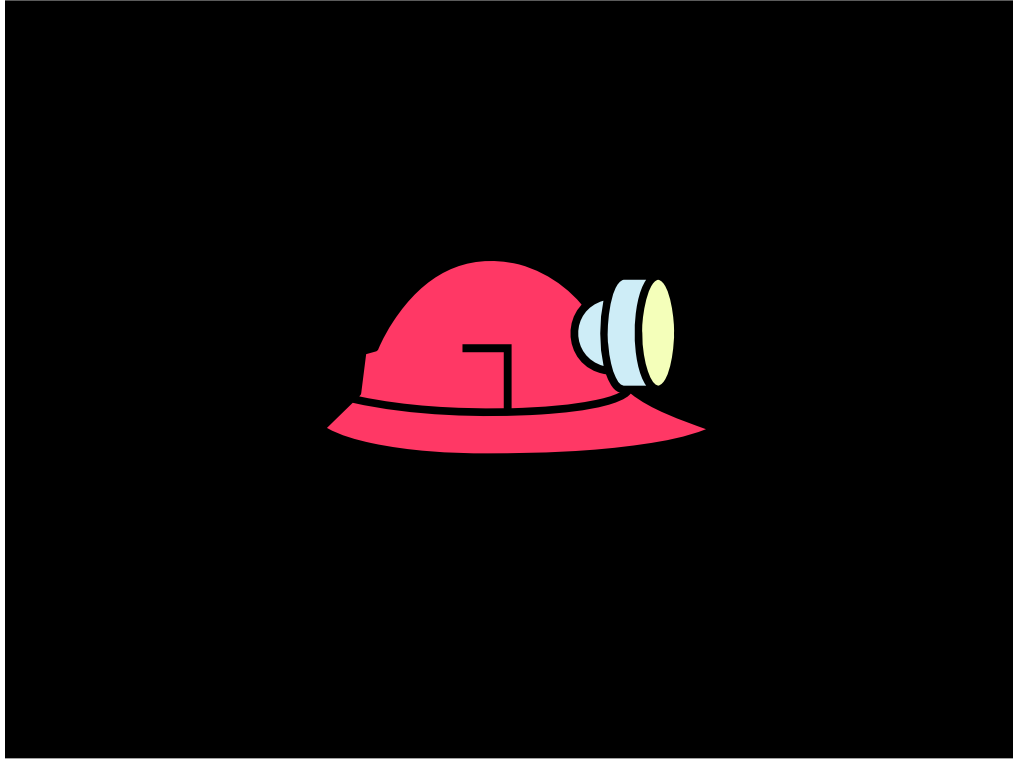


Let's start by putting on a few hats.

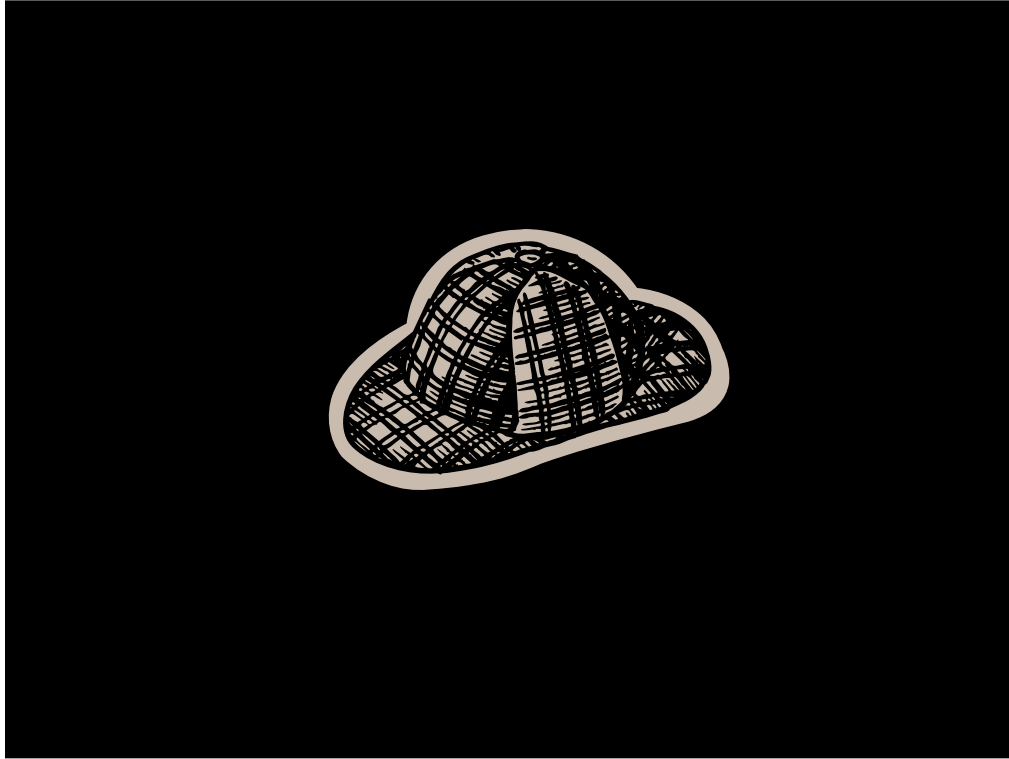
You've just finished analyzing the results of your clinical trial. You suspect your results are generalizable. It would sure help the impact of your publication if you could include an analysis on an independent population.



You believe your research might help to improve prognosis in a rare disease in a rare population. Because of the rarity, it isn't possible to recruit enough people in any one trial. Perhaps the effect could be studied by re-examining patient outcomes across a number of previous trials?



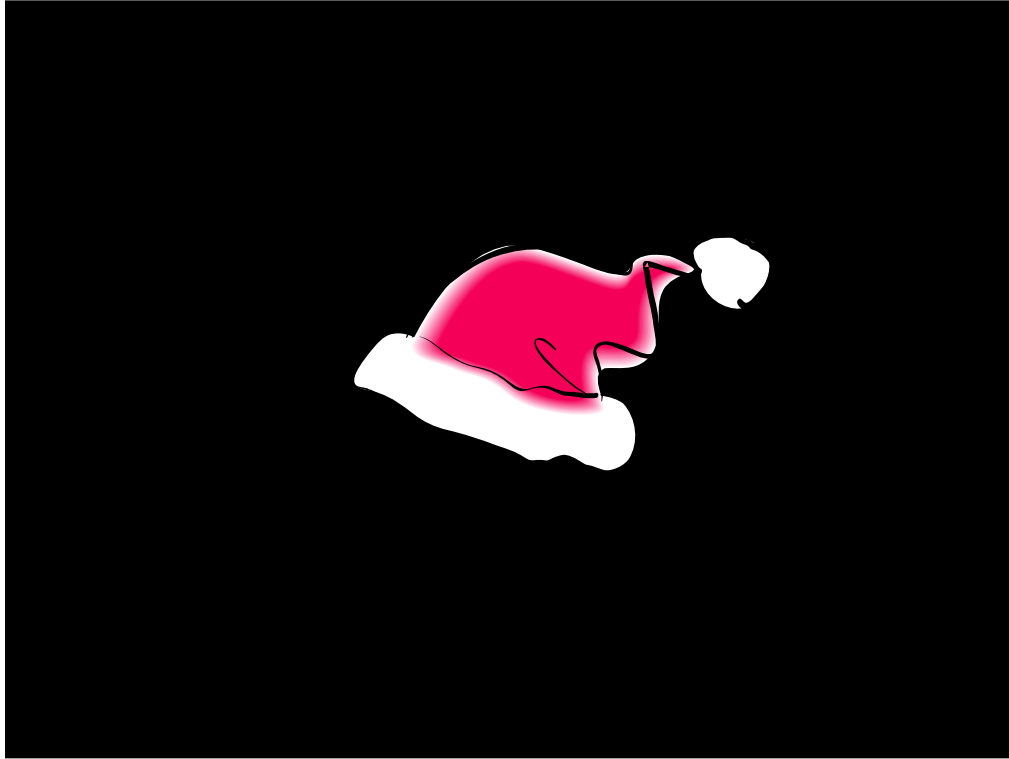
You have a new data mining technique you'd like to try.



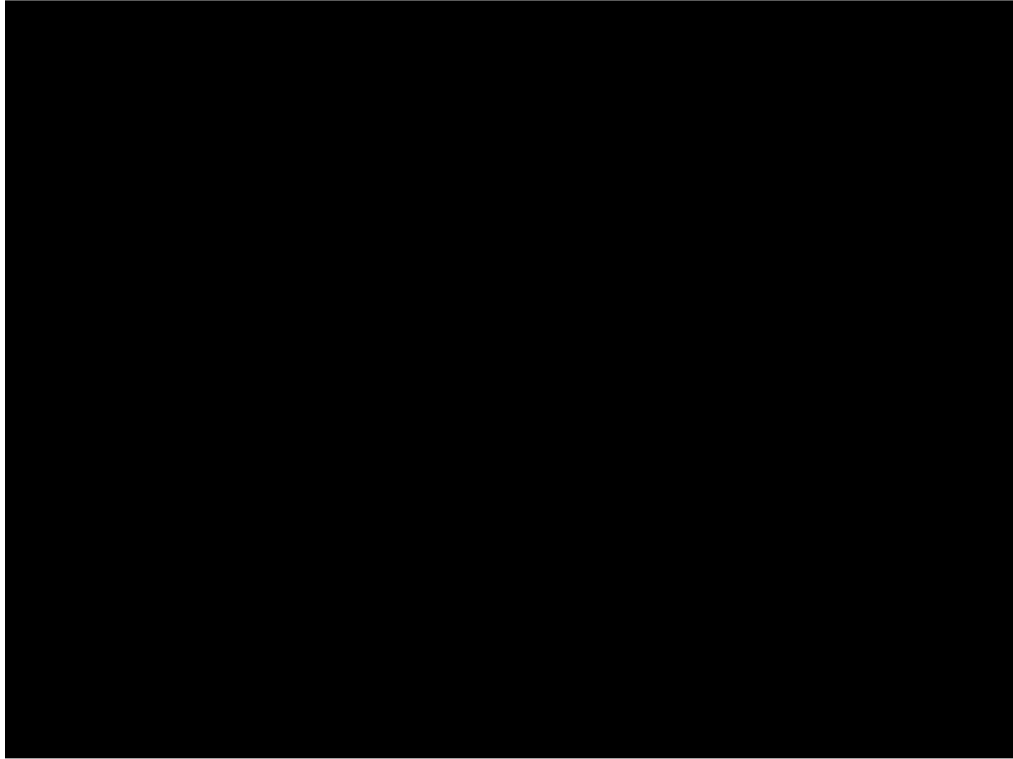
You just read the latest issue of your favorite journal, and one of the results contradicts the research you've been doing. Is your research off base? Perhaps the authors made a mistake? You'd like to take a closer look.



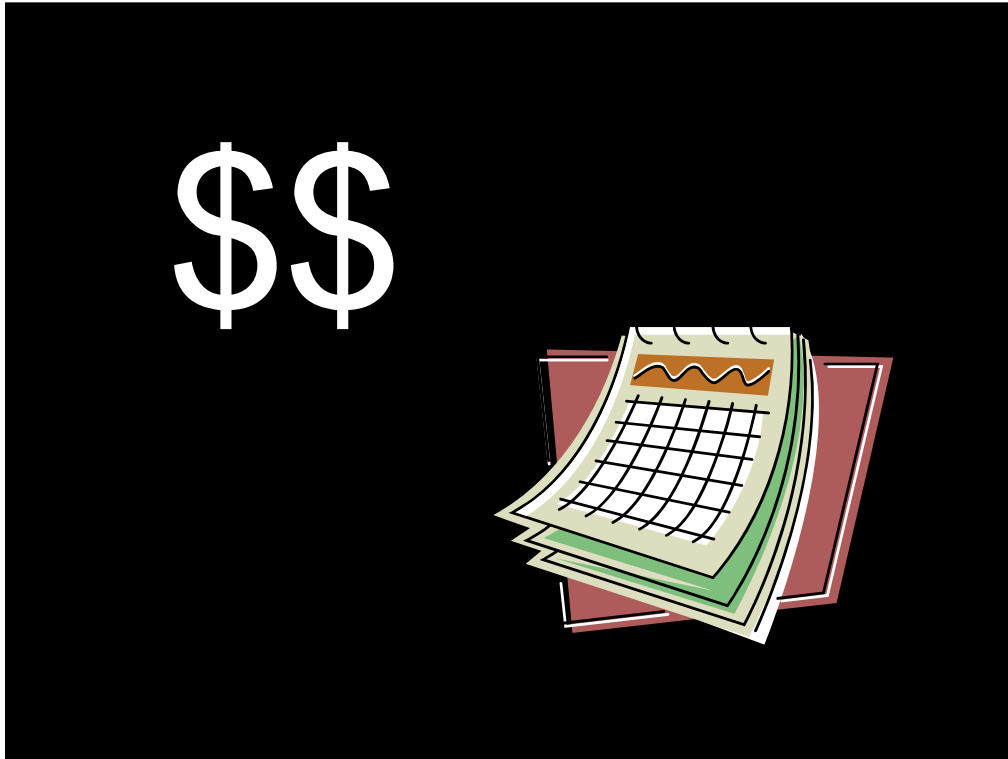
You are a student, and you relish the rich learning experience of applying the new concepts to real data with all of its inherent complexities.



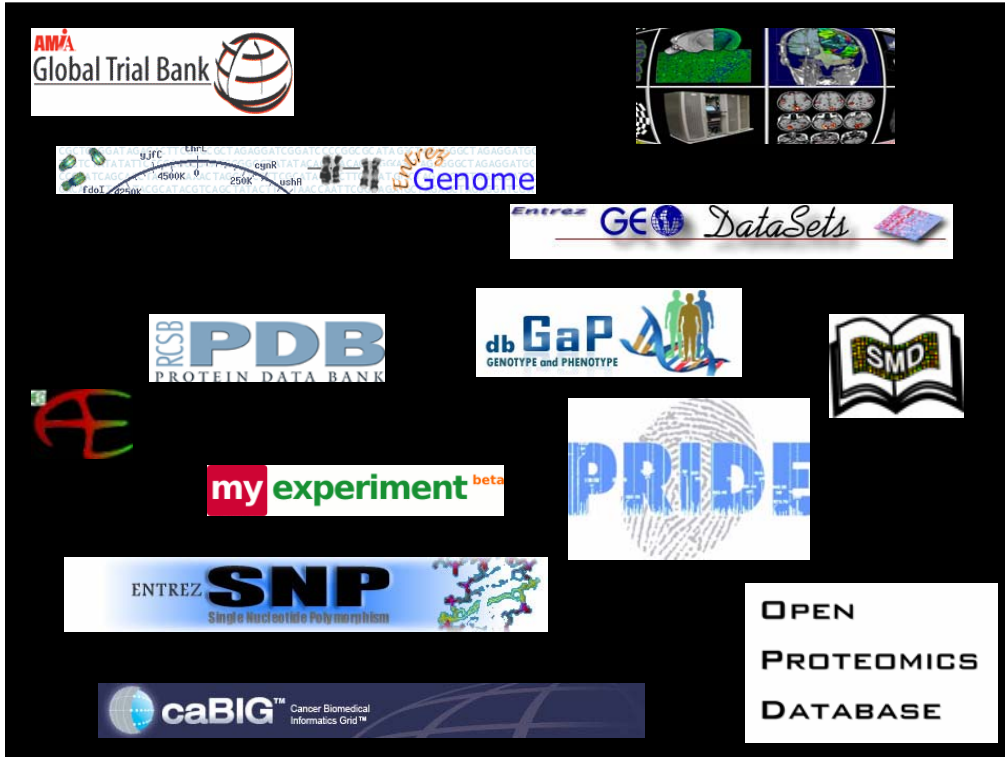
You are a smart cookie with unique perspective and skills, and you'd like to help advance the research frontier. Unfortunately, you live in a rural area that doesn't have access to hospitals or wet labs to generate your own data.



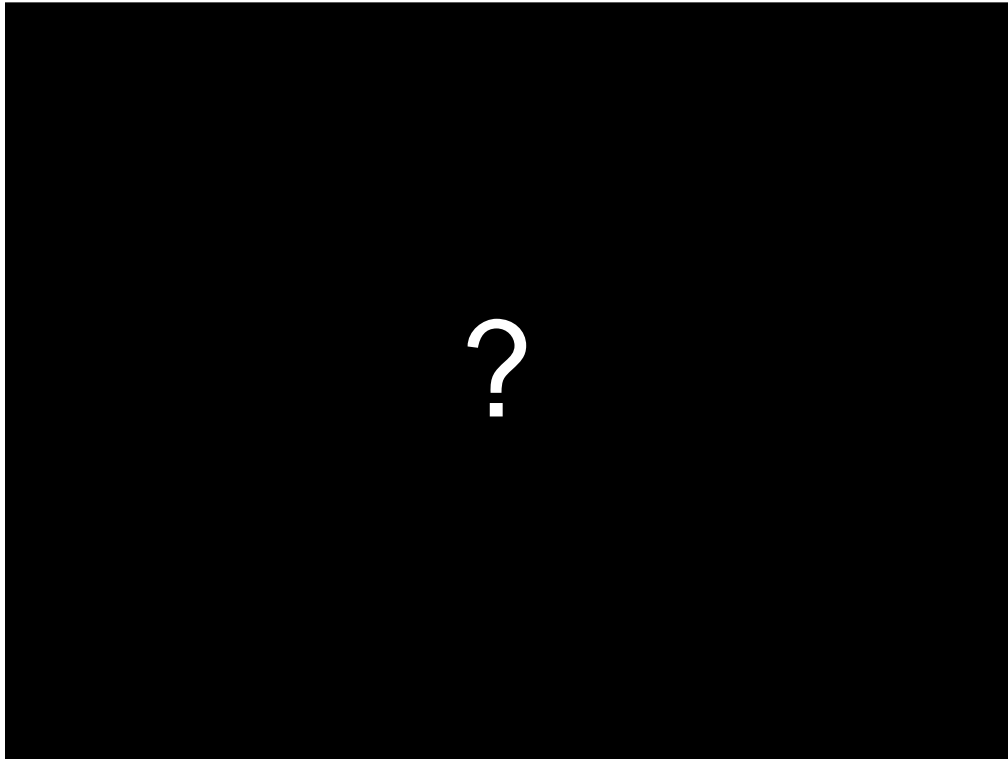
And we could imagine many more....



To help achieve these benefits, the science community is spending lots of time and money on databases, initiatives, and policies that encourage people to share their data.




Tip of the iceberg



But is it worth it? Are we realizing the benefits? Could we be doing better?



To answer that, we need to evaluate.



We cannot manage
what we do not measure

We cannot manage what we do not measure. If we believe that data sharing and reuse has value, or even if we think it might, we need to study our current practices and behavior to choose the most effective path for the future.

Long-term research hypothesis:

There exist many underutilized data resources.

Evaluating data sharing and reuse
policies and behaviors

will eventually lead to

improved scientific progress

I believe there are a whole lot of underutilized data resources out there. By evaluating data sharing policies and behaviours we can learn how to put these to better use for the progress of science.

Since I don't have much time today, I'm going to give you a really quick rundown of the many open questions in this area, and a bit about what we've learned.

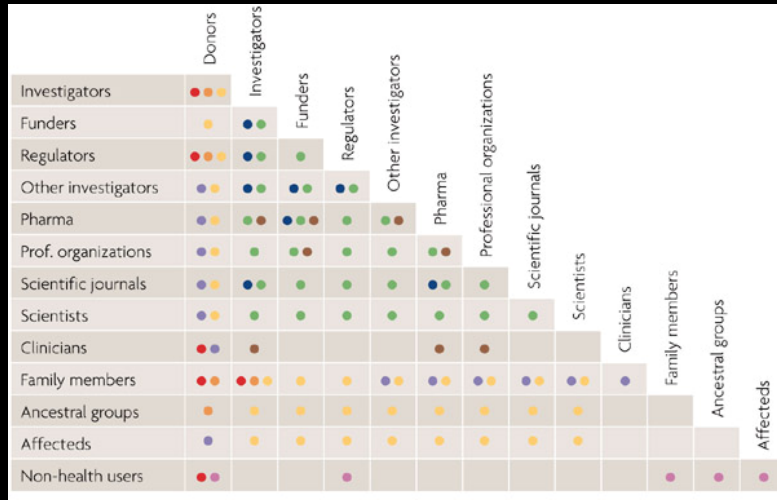
5 W's + H

- Who
- What
- When
- Where
- Why
- How

... out of order

We're going to go over the 5 W's, but we're going to do them out of order. No fear, you can handle it.

Why share data?



Foster et al. Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. Nature Reviews Genetics 8, 633-639

Many stakeholders involved in decision to share data, all with different reasons for and against sharing and reusing datasets.

Why share: funder requirements

- NIH
 - Sharing plan for grants > \$500k
 - Sharing in dbGaP for GWAS studies
- NSF, Wellcome Trust, ...

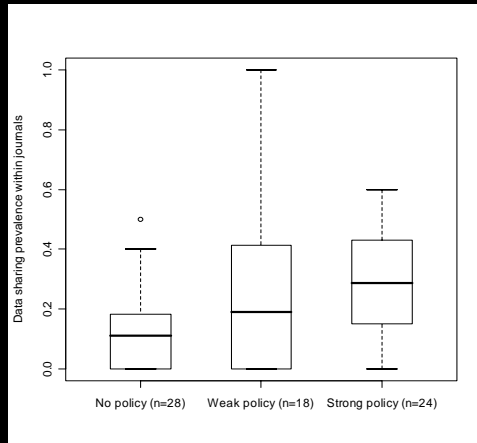
*Funded by NIH: 2.0x more likely to share
More funding sources -> More sharing*

Piwowar, Chapman. Prevalence and Patterns of Microarray Data Sharing. Poster at PSB 2008.

Why share: journal requirements

Percent journal studies with links from microarray databases, by strength of journal data-sharing policy

Note:
Studies are about microarray data but they may not have all generated microarray data. Calculated percentages may be <100% even when universal sharing.



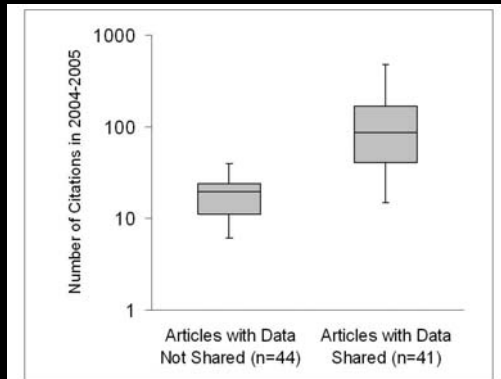
Piwovar, Chapman. A review of journal policies for sharing research data. Accepted to ELPUB2008

Note: There are interactions between journal policy strength and impact factor.

Why share: investigator benefit

Number of citations received by clinical trial publications, by whether or not they publicly shared their microarray data

Note the logarithmic scale



Piowar, Day, and Fridsma DB (2007). PLoS ONE 2(3): e308.

A look into the individual benefits for those investigators who share their data: The publications which did not make their data available had a median of 20 citations, whereas on the right, the publications which did make their data available had a median of about 100 citations. The Y axis is a log scale, in case you can't see that in the back. Note, again there are interactions with impact factor.

Why withhold: investigator cost

- Among geneticists who said they had intentionally withheld data regarding their published work:
 - 80% too much effort
 - 64% protecting the ability of a graduate student, postdoctoral fellow, or junior faculty member to publish
 - 53% protecting their own ability to publish

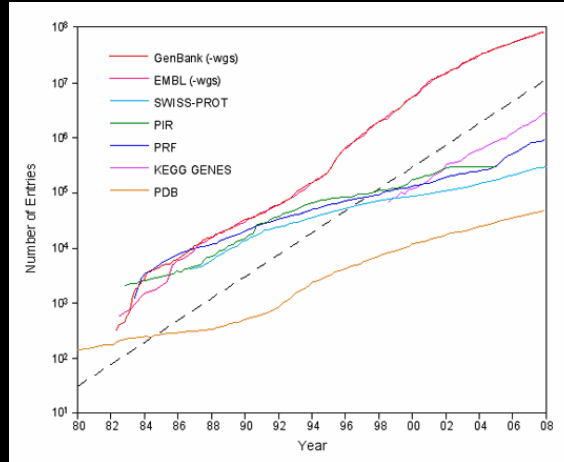
Campbell et al. JAMA 2002. Jan 23039;287(4):473-80

In other parts of this presentation, I'm defining "data sharing" to imply public, open, free sharing on the internet. There are of course other kinds of sharing. A series of studies by Campbell and collaborators looked at data withholding. They defined this as an investigator not supplying research related information to other investigators upon request.

What types of data?

Growth of Sequence and 3D Structure Databases over time

Note the logarithmic scale



Much less sharing of other datatypes...

http://www.genome.jp/en/db_growth.html

These are the most commonly shared datatypes

What data elements?

- ArrayExpress lists MIAME score:

The screenshot shows the ArrayExpress website interface. At the top, there are navigation tabs: Databases, Tools, EBI Groups, Training, Industry, About Us, and Help. Below these are search filters: 'Text search...' with a search button, 'Filter on species...' dropdown, 'Filter on array...' dropdown, and a 'reset' button. There are also options for 'detailed view' and 'show 50 experiments'. Below the filters is a table with columns: ID, Title, Hybs, Species, Date, and Process. The first row is: E-MEXP-727, Oliver - S cerevisiae - Respiration related deletion mutants, 20, Saccharomyces cerevisiae, 2008-03-20, and a small icon. Below the table, there is a detailed view for the selected experiment. A red circle highlights the 'MIAME score' field, which contains the text: '4 (array: 1, protocols: 0, factors: 1, raw data: 1, processed data: 1)'. Other fields include 'Title: Oliver - S cerevisiae - Respiration related deletion mutants', 'Sample annotation: > Tab-delimited spreadsheet', and 'Array: Affymetrix GeneChip Yeast Genome S98 [YG_S98] (> A-AFFY-27)'. At the bottom, it shows 'Experiments: 3409' and 'Hybridizations: 101017'. The URL at the bottom is 'http://www.ebi.ac.uk/microarray-as/aer/?#ae-browse[2]'.

ID	Title	Hybs	Species	Date	Process
E-MEXP-727	Oliver - S cerevisiae - Respiration related deletion mutants	20	Saccharomyces cerevisiae	2008-03-20	

MIAME score: 4 (array: 1, protocols: 0, factors: 1, raw data: 1, processed data: 1)

Where is it stored?

- Personal or lab websites
- Journal supplementary information
- Federated data stores
- Centralized databases



- Ease of location, retrieval
- Long term availability
- Computational access
- Control

When is it shared?

- As you are doing it?
 - Open Notebook Science
- Within 24 hours of collection?
 - Human Genome Project
- After QA, with publication embargo?
 - NIH GWAS studies
- On publication?
 - Journal microarray policies

How is it shared?

- Semantic and syntactic standards
- How (+ What + Where + ...) aspects are critical for sharing health care data:

Journal of the American Medical Informatics Association Volume 14 Number 1 Jan / Feb 2007

1

Perspectives on **Informatics**

JAMIA

White Paper ■

Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper

CHARLES SAFRAN, MD, MS, MERYL BLOOMROSEN, MBA, W. EDWARD HAMMOND, PHD,
STEVEN LABKOFF, MD, SUZANNE MARKEL-FOX, PHD, PAUL C. TANG, MD, DON E. DETMER, MD, MA,
WITH INPUT FROM THE EXPERT PANEL (SEE APPENDIX A)

Who shares data?

Don't really know the characteristics of people who most openly share.

Who withholds data?

From one survey, in multivariate analyses:

- participation in relationships with industry,
- mentors' discouraging data sharing,
- receipt of formal instruction in data sharing,
- negative past experience with sharing,
- male gender

Blumenthal et al. Acad Med. 2006 Feb; 81(2):137-45

The authors note that the receipt of formal instruction was not widespread, and may have been given only in situations which were predisposed to high rates of withholding.

Who: Trainee experience

- Survey of 1,077 2nd year doc+postdocs at 50 US universities
- **23.0% been denied access** to information, data, materials, or programming associated with published research
- **7.9% reported they had denied access** to another academic scientist's request(s) related to their own published research.
- **28-50%** reported withholding caused **negative effects** on these aspects of **their training**:
 - progress of their research,
 - rate of discovery in their lab/research group,
 - quality of their relationships with academic scientists,
 - quality of their education,
 - level of communication in their lab/research group.

Vogeli et al. Acad Med. 2006 Feb; 81(2):128-36

Who

- You?

You

- I think data sharing is relevant for all of us.
- IRB? Privacy? Animal studies?
Collaborators? Esoteric data sets?

Some is better than none!
Share what you can.

Does anyone want your data?

That's hard to predict, but **the easier** it becomes to request data and to receive credit for sharing it, **the more likely people are to ask**. After all, no one ever knocked on your door asking to buy those figurines collecting dust in your cabinet before you listed them on eBay. Your data, too, **may simply be awaiting an effective matchmaker.**

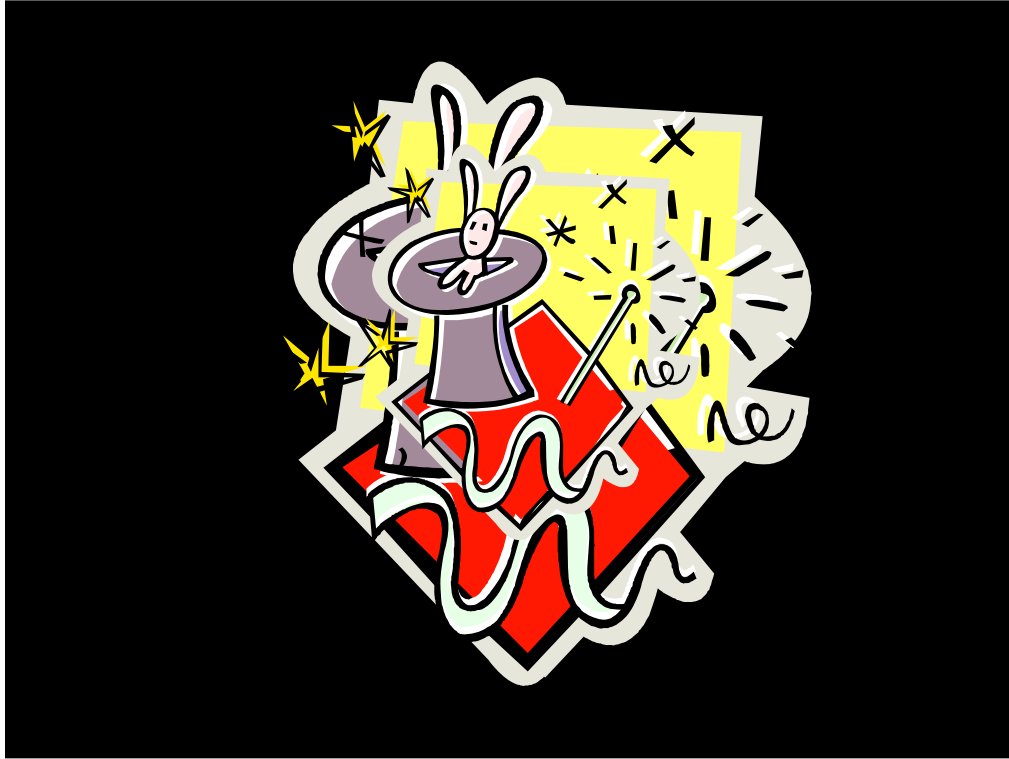
Got data? Nature Neuroscience 10, 931 (2007)

Got data? Nature Neuroscience 10, 931 (2007)

Not sure what your data is good for? A Nature editorial this summer. I'll let you read it.



Of course you're afraid some joker is going to find a mistake, and you might get in trouble.



But you never know unless you try it. I'm guessing you'll be pleasantly surprised.

Mine is here

www.dbmi.pitt.edu/piwowar

You can ask for a DBMI homepage too.

Why is this? Who does share and who doesn't? What are effective ways to get people to share? Are the benefits of sharing actually realized? How much time and effort is it worth?

In summary

I hope I've given you an overview of why data sharing is important, why it would benefit from evaluation, a taste of the various open questions and current answers, and finally prompted each of you to think a bit about making your own detailed research data available to more people.

Thank you

- Peter Suber's blog: "Open Access News"
- Wikipedia: "Open Data"
- Got data? *Nature Neuroscience* 10, 931 (2007)

Questions?



I'd like to thank to do thanks times three. To the NLM for my generous funding support, my advisors for their insightful feedback, and last but not least, thanks to each and every one of you who have made your detailed research data available in the past, or will do so the future. It matters.