Measuring the adoption of open science

Heather Piwowar

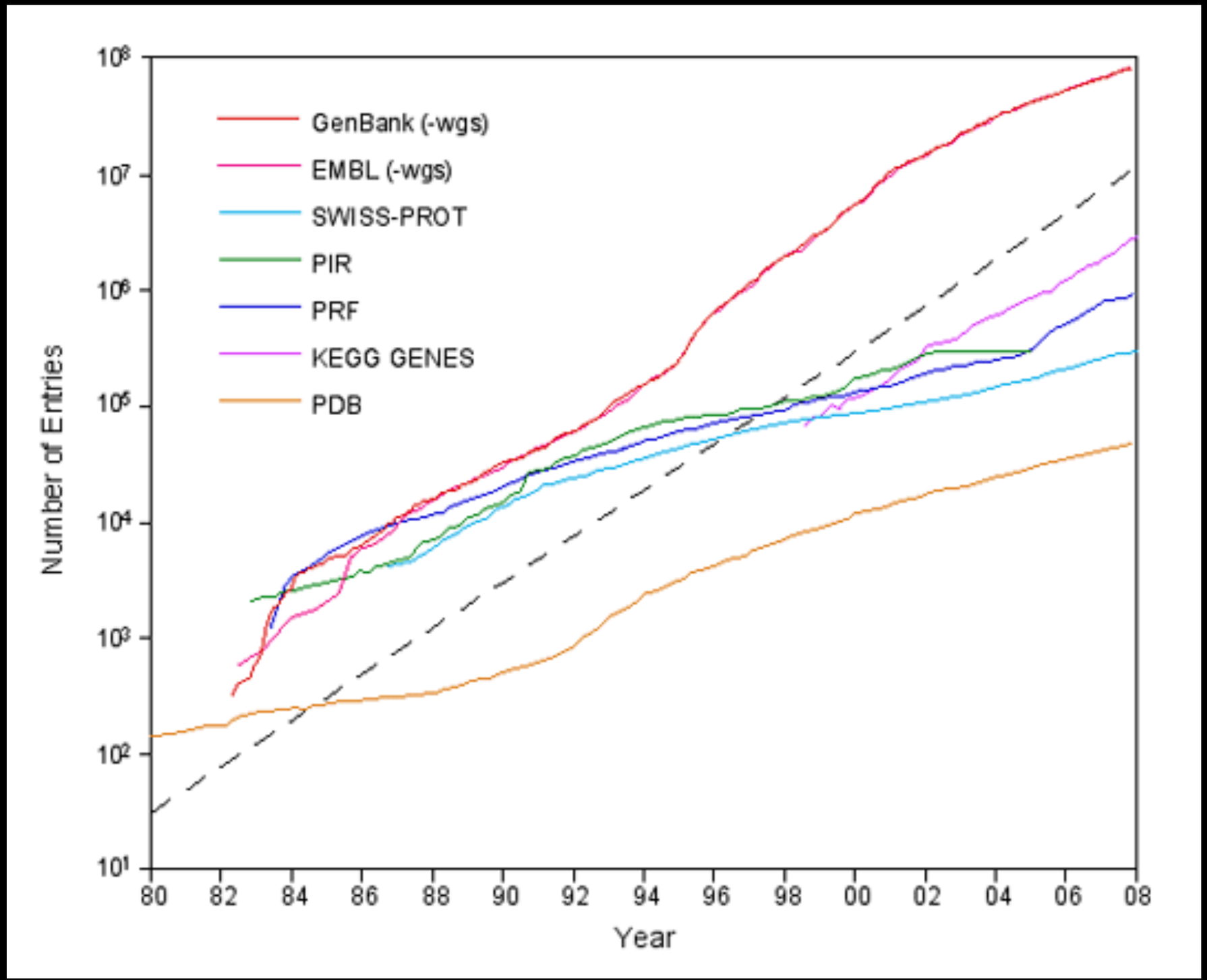
Department of Biomedical Informatics
University of Pittsburgh

PSB workshop on Open Science, January 2008

you can not manage
what you do not measure

Measuring the adoption of ~~open science~~ sharing data

lots of data sharing!



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

but how much isn't shared?

what isn't shared?

who isn't sharing it?

why not?

how much does it matter?

what can we do about it?

I'll be highlighting the results of a number of studies:

surveys
manual reviews
citation analyses

Preview

Although some scientists voluntarily share their research data, many don't.

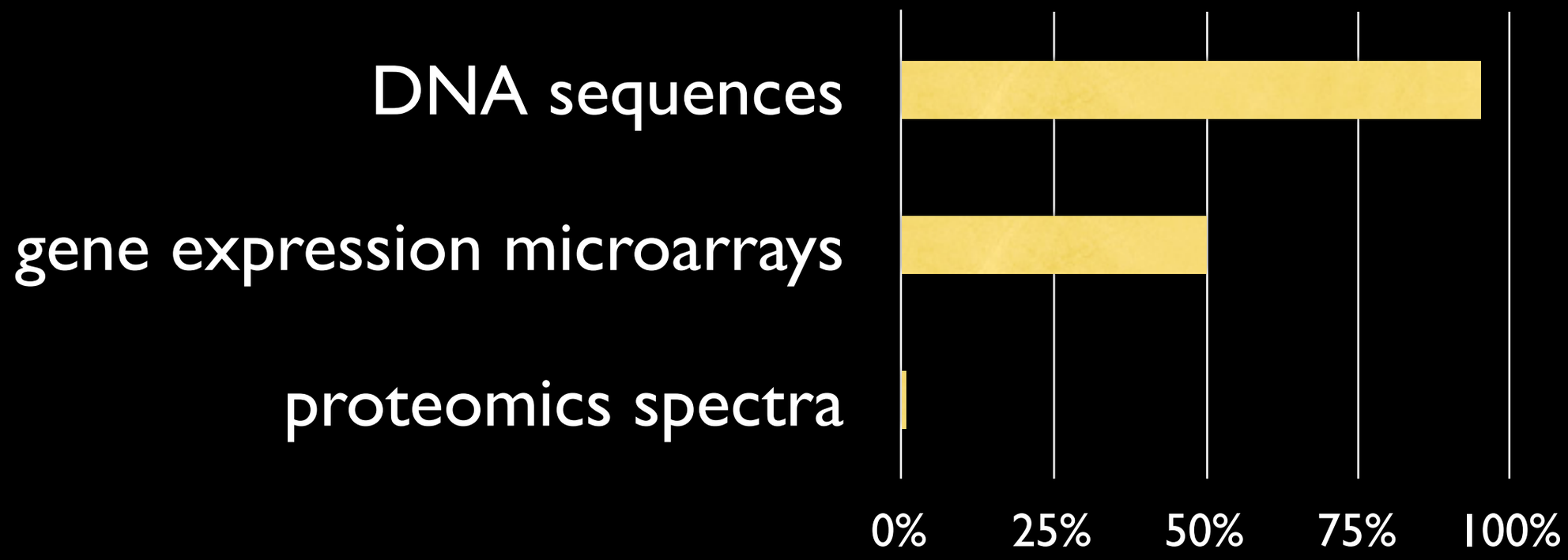
Data withholding correlates with the usual suspects.

Feedback on incentives may surprise you.

Much room for continued research, including several ways that you can help.

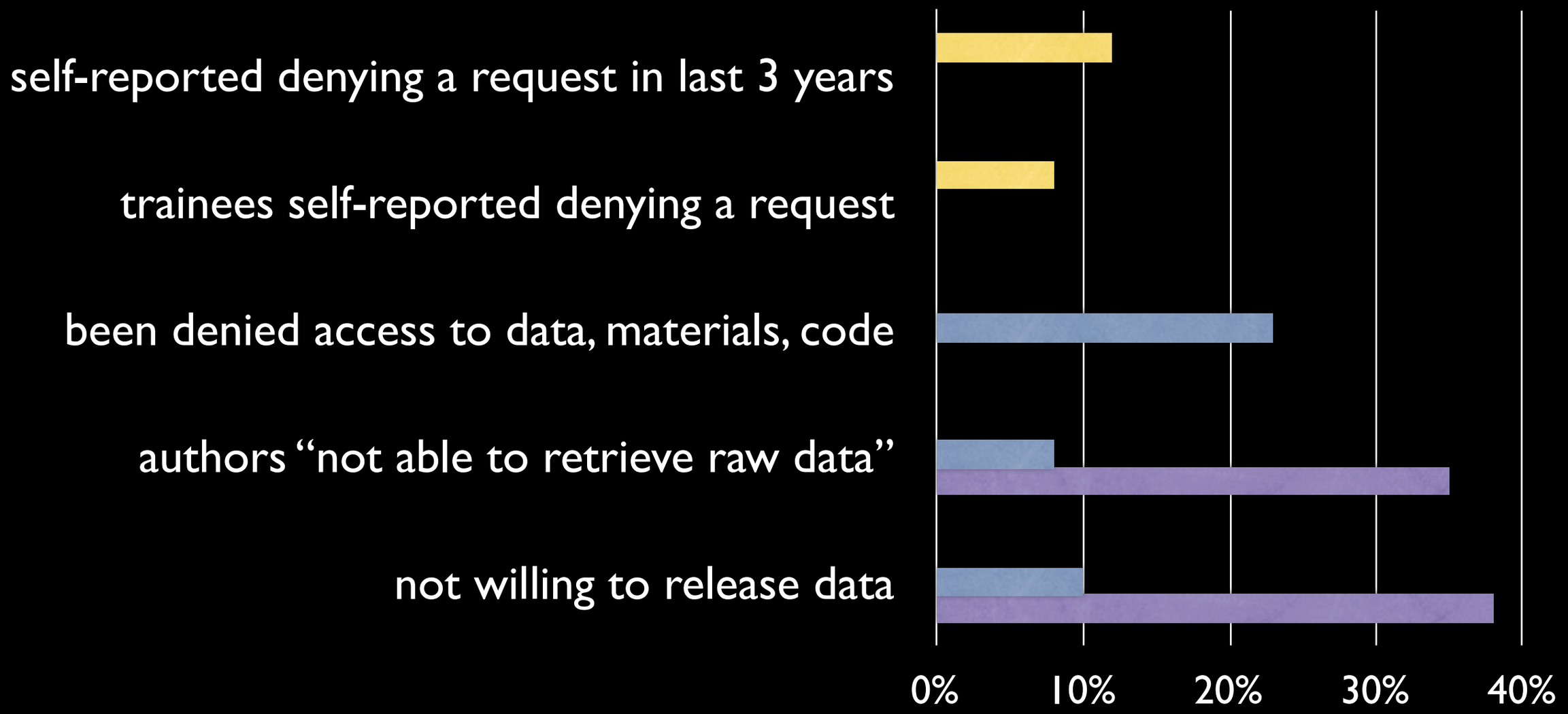
How much data gets shared?

Data sharing frequency depends on ***datatype***



Noor et al. PLoS Biology 2006.
Ochsner et al. Nature Methods 2008.
Piwowar et al. PLoS ONE 2007.
Editorial. Nature Biotech 2007.

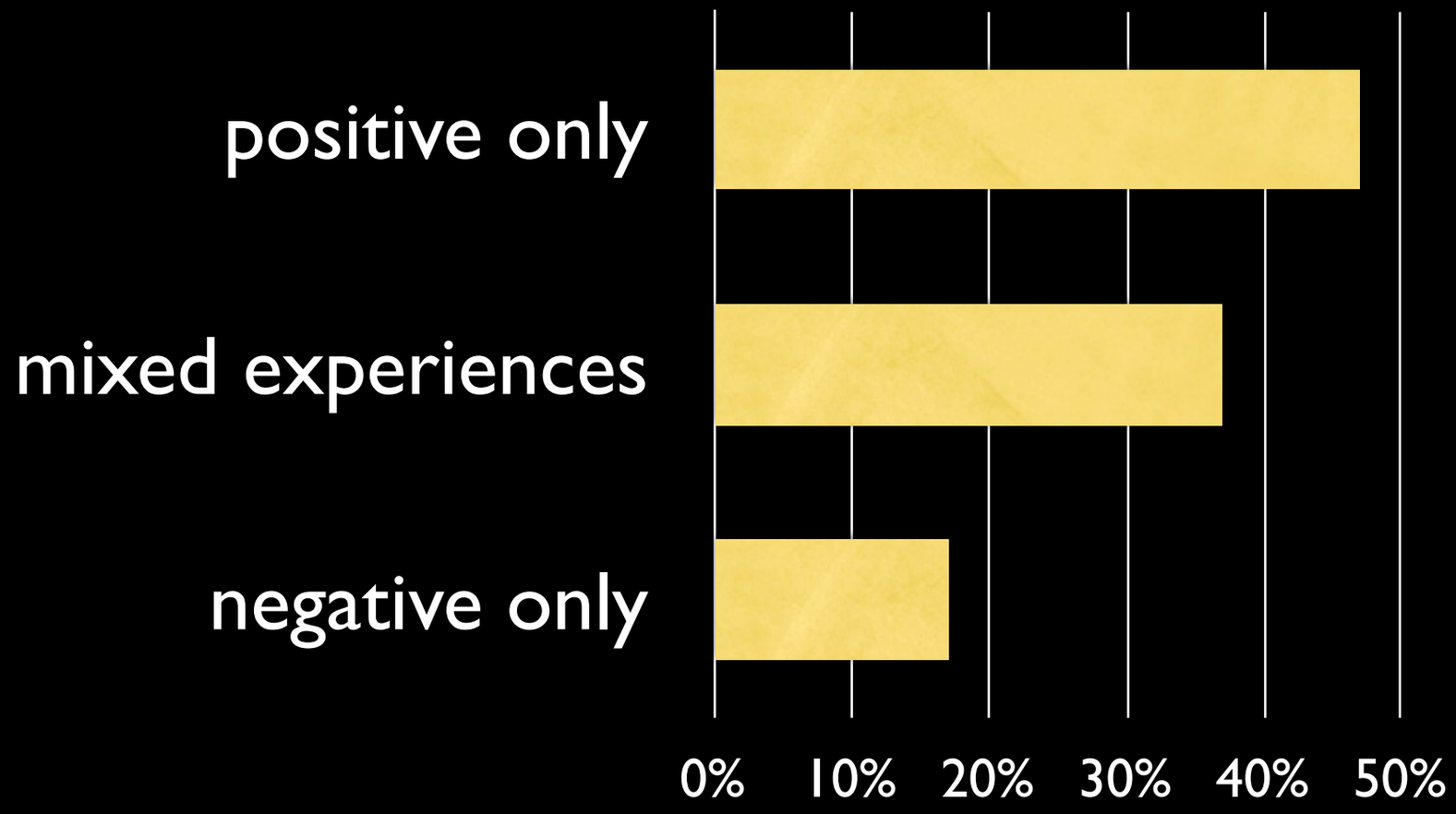
Data sharing frequency depends on *who you ask*



Campbell et al. JAMA. 2002.
Kyzas et al. J Natl Cancer Inst. 2005.
Vogeli et al. Acad Med. 2006.
Reidpath et al. Bioethics 2001.

Are the outcomes of data sharing positive or negative?

80% of scientists report positive experiences from data sharing

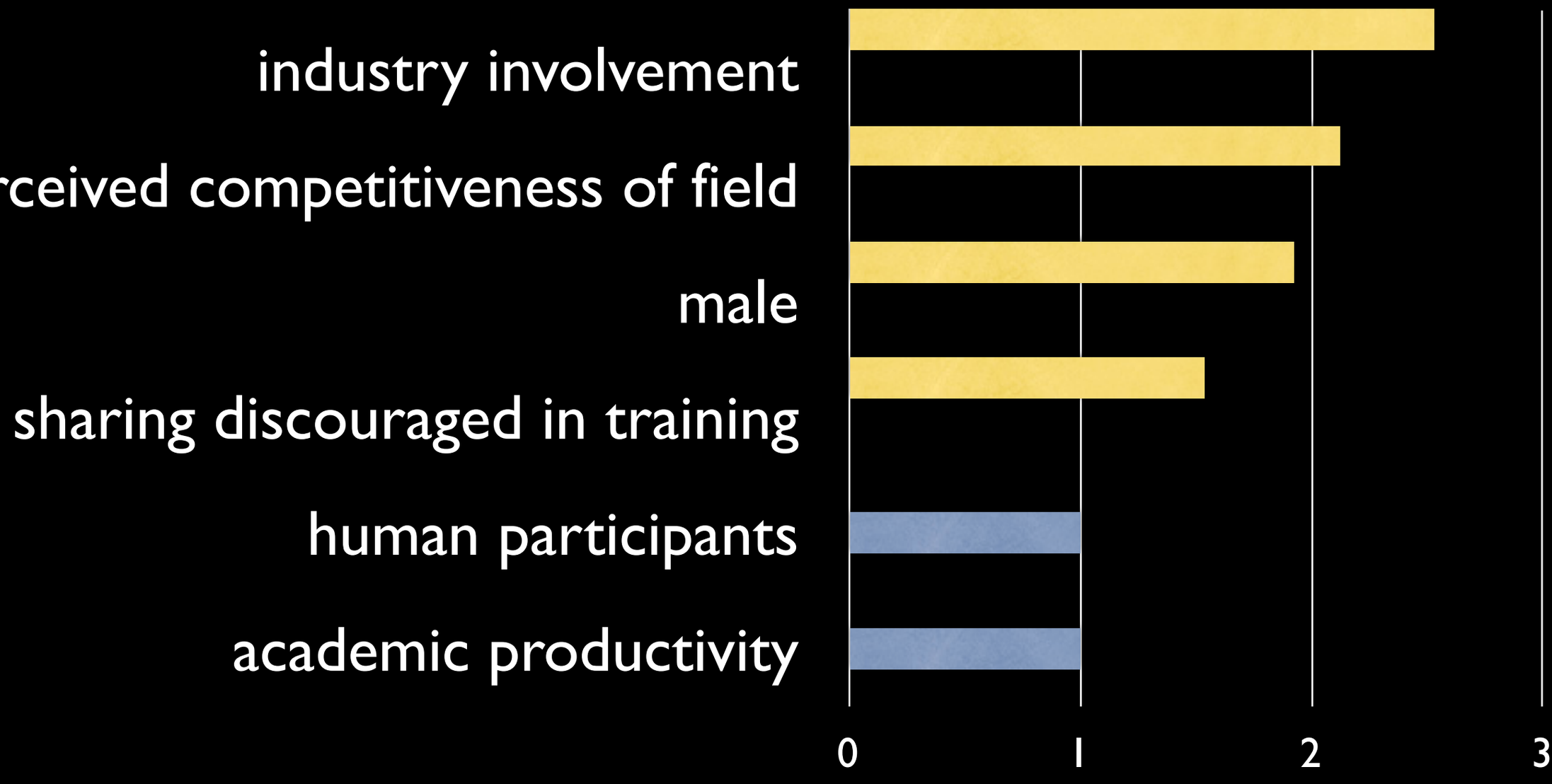


Positive experiences: collaboration, new research, etc.
Negative: scooping, preventing publishing, IP, or $$ benefit, etc.

Blumenthal et al. Acad Med. 2006

Why is data withheld?

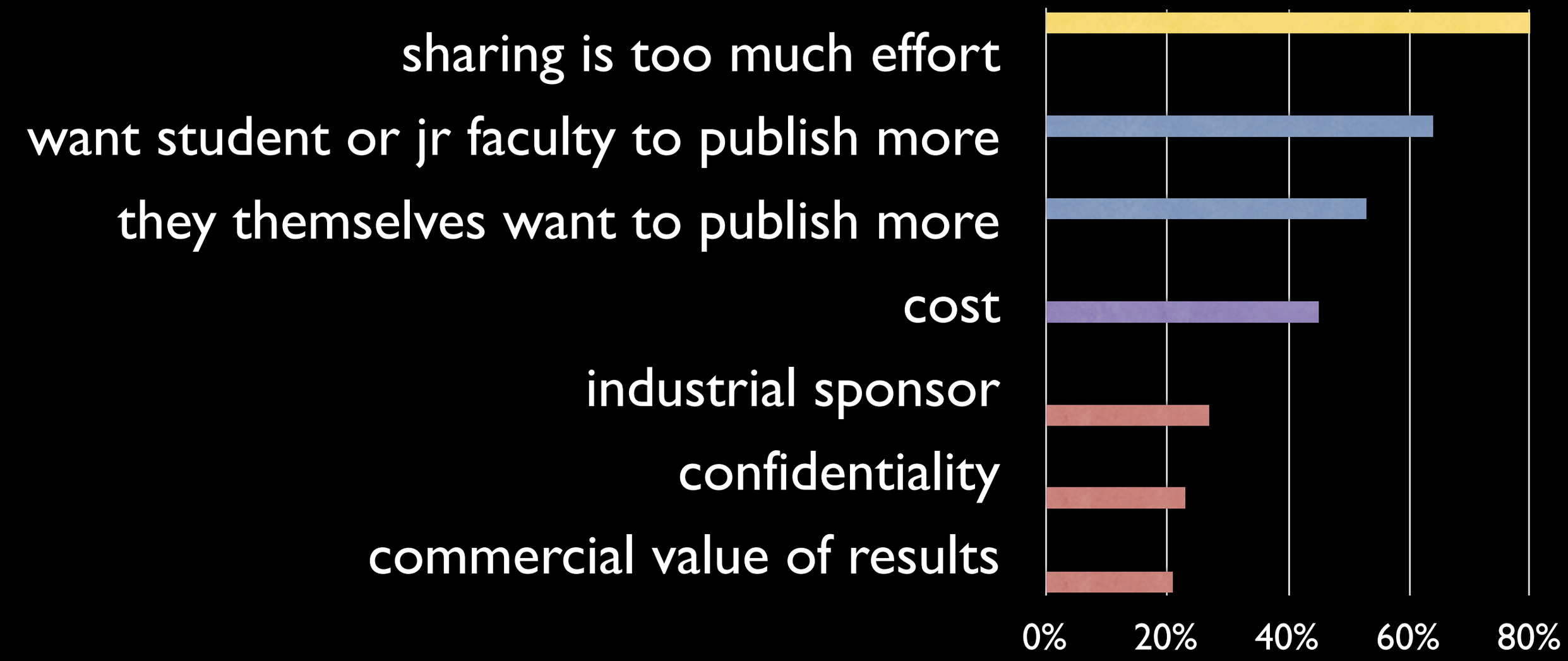
Withholding is associated with *industry links, competiveness*



40% of surveyed scientists said data sharing was discouraged during their training!

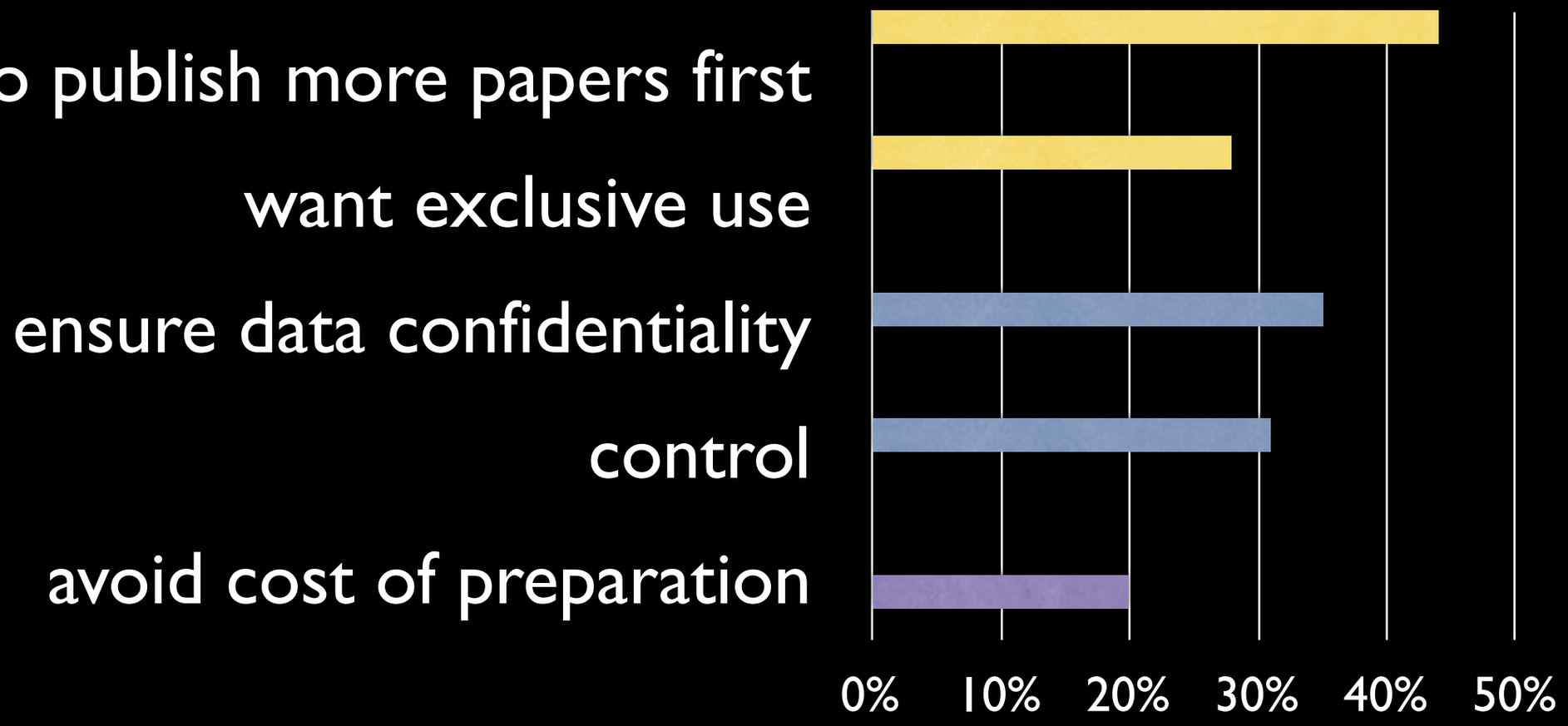
Blumenthal et al. Acad Med. 2006

Withhold because *too much effort, desire for continued publishing*



Campbell et al. JAMA 2002.

Obstacles for sharing: *publishing, control, cost*



Hedstrom. Society of Am Archivists Ann Meeting. 2008.

Comments show desire for ***control***

`Before I send you the data could I ask what you want it for?'

`Can you be more explicit, please, about the analyses you have in mind and what you plan to do with them?'

`We'll have to discuss your request with the other coauthors. Before we do that, I'd like to know your proposed analysis plan.'

`We are not finished using the data, but when we are finished with it, we would be open to requests for the data.'

`Any use of the data other than for the specific purpose laid down in the contract of collaboration is effectively ruled out.'

Reidpath et al. Bioethics 2001

What are the perceived and measured benefits?

Benefits both *societal and personal*

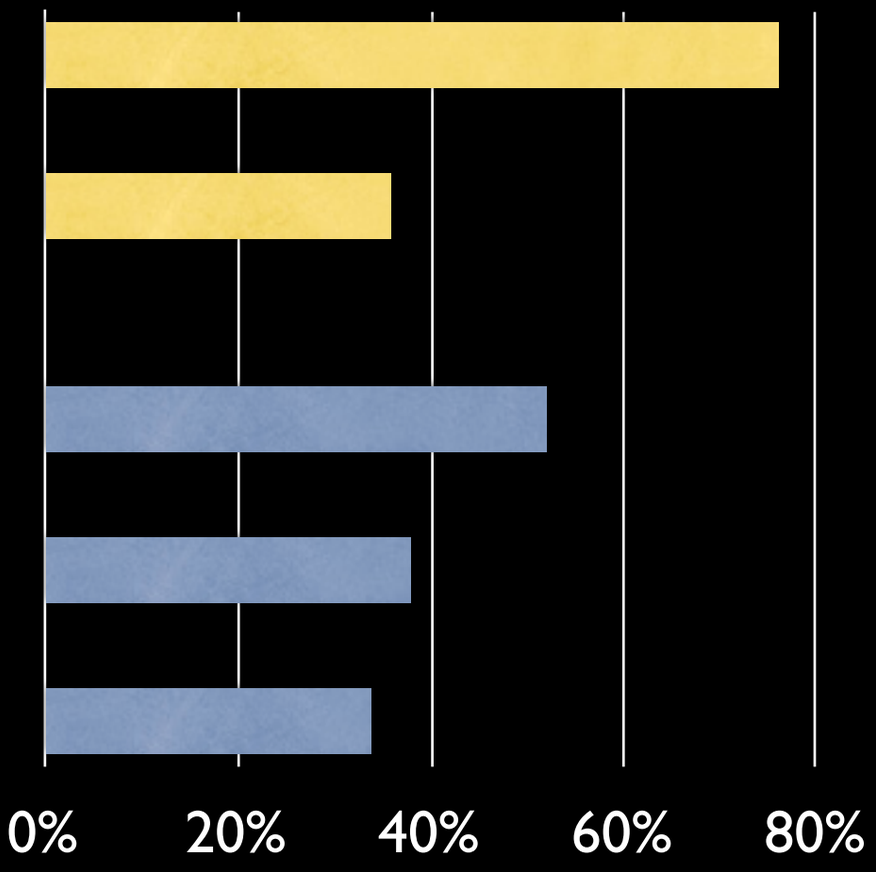
saves other people effort

for the public good

will be cited and enhance my reputation

saves me effort in answering questions

saves me effort in managing my data



Hedstrom et al. IASSIST 2006.

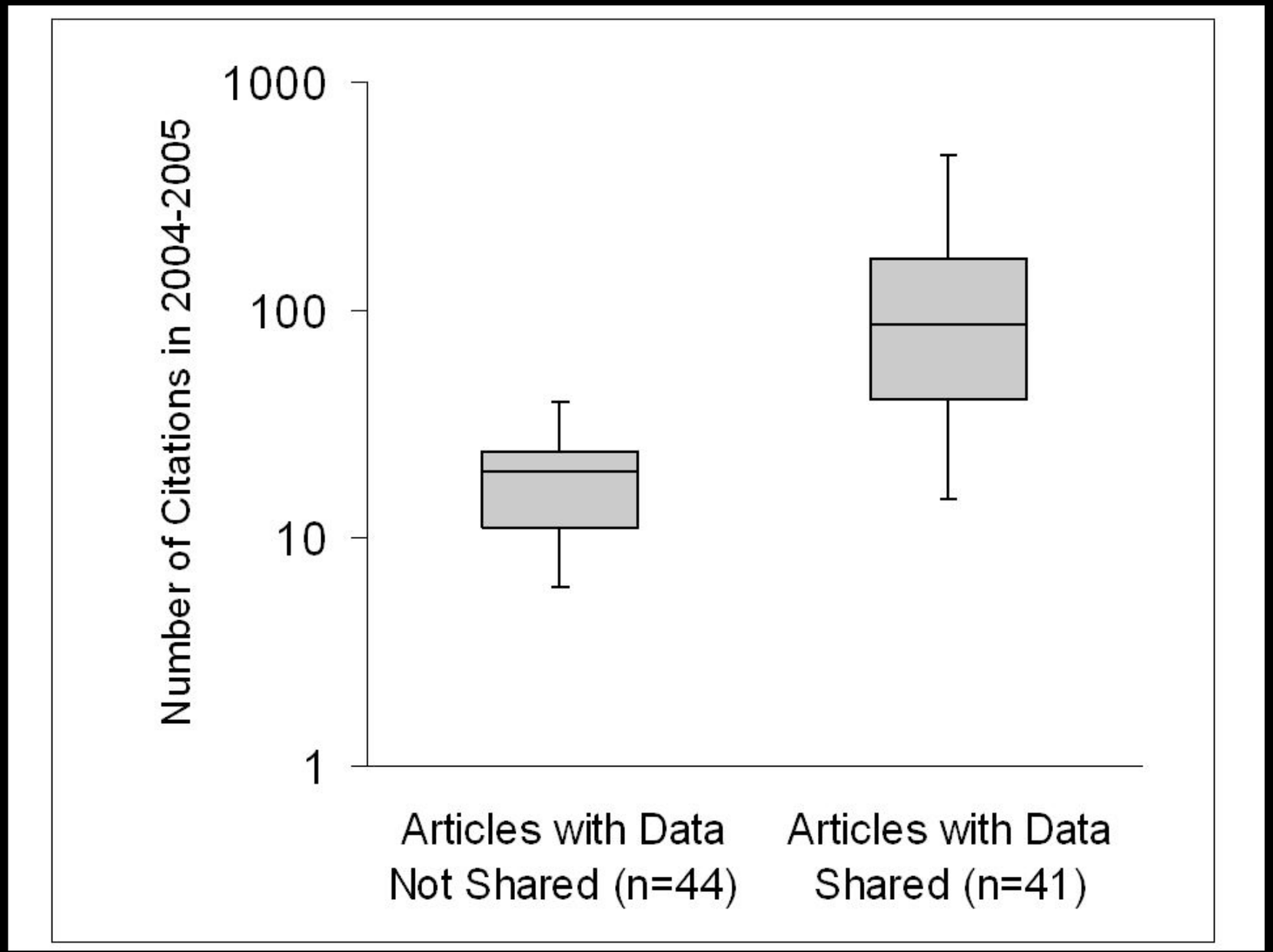
Measuring societal benefit

- assume each database hit saves $0.10, or a fraction of data collection costs
- assume the value is approximated by the (idealized) funding target for data maintenance:
  20-25% the cost of generating the data

Remembering, moreover, the indirect benefits are much higher than the direct ones.

Ball et al. Nature Biotechol. 2004.

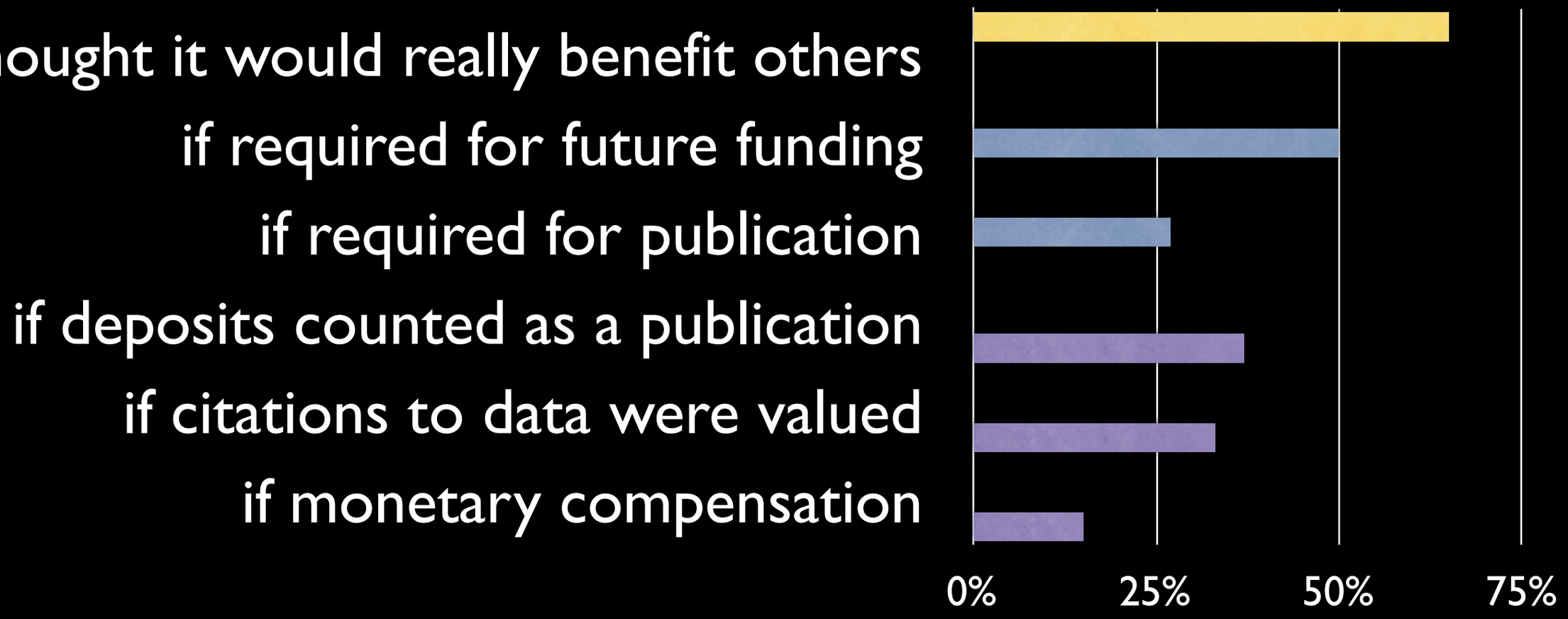
Measuring personal benefit: ***increased citations***



Gleditsch et al. Int Studies Perspectives. 2003.
Piwowar et al. PLoS ONE. 2007.

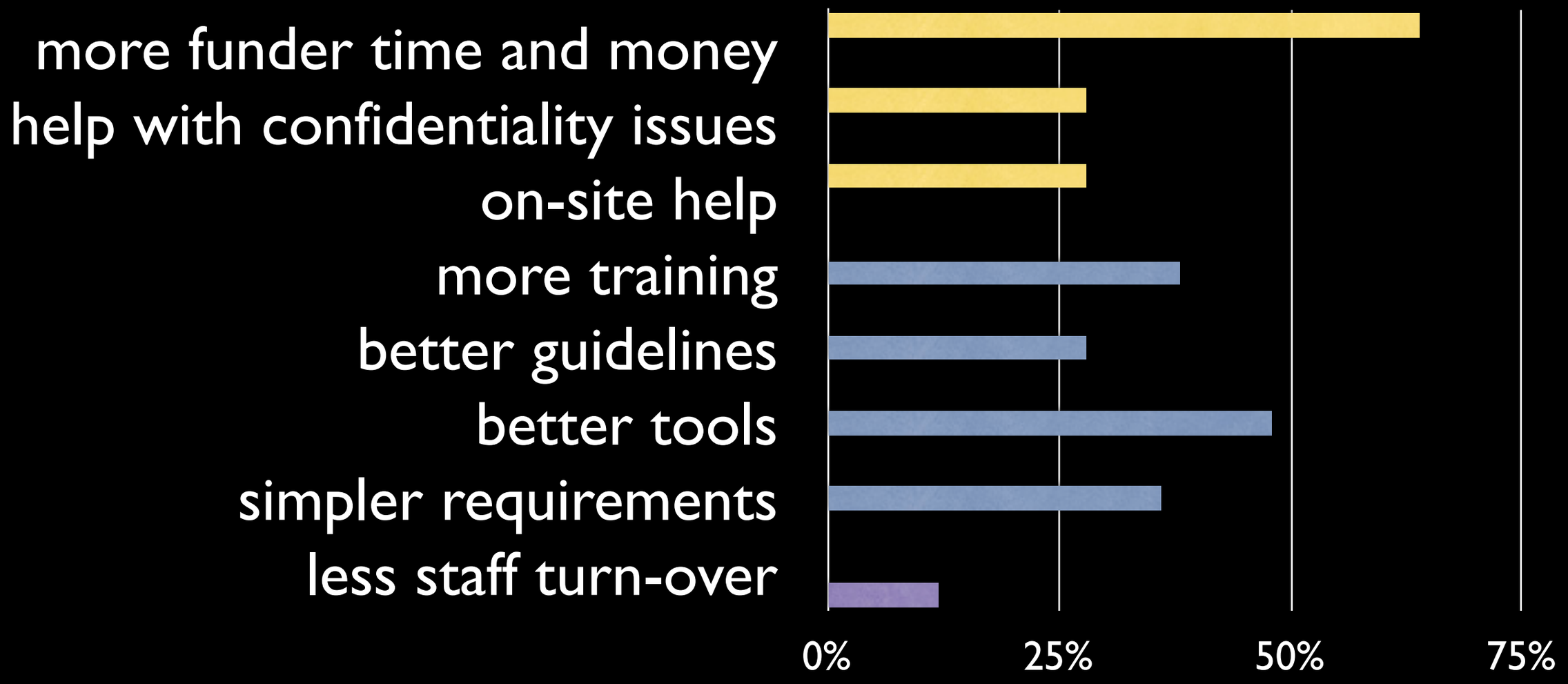
What incentives are valued?

Incentives to share: *perceived value, mandates, recognition as publication*



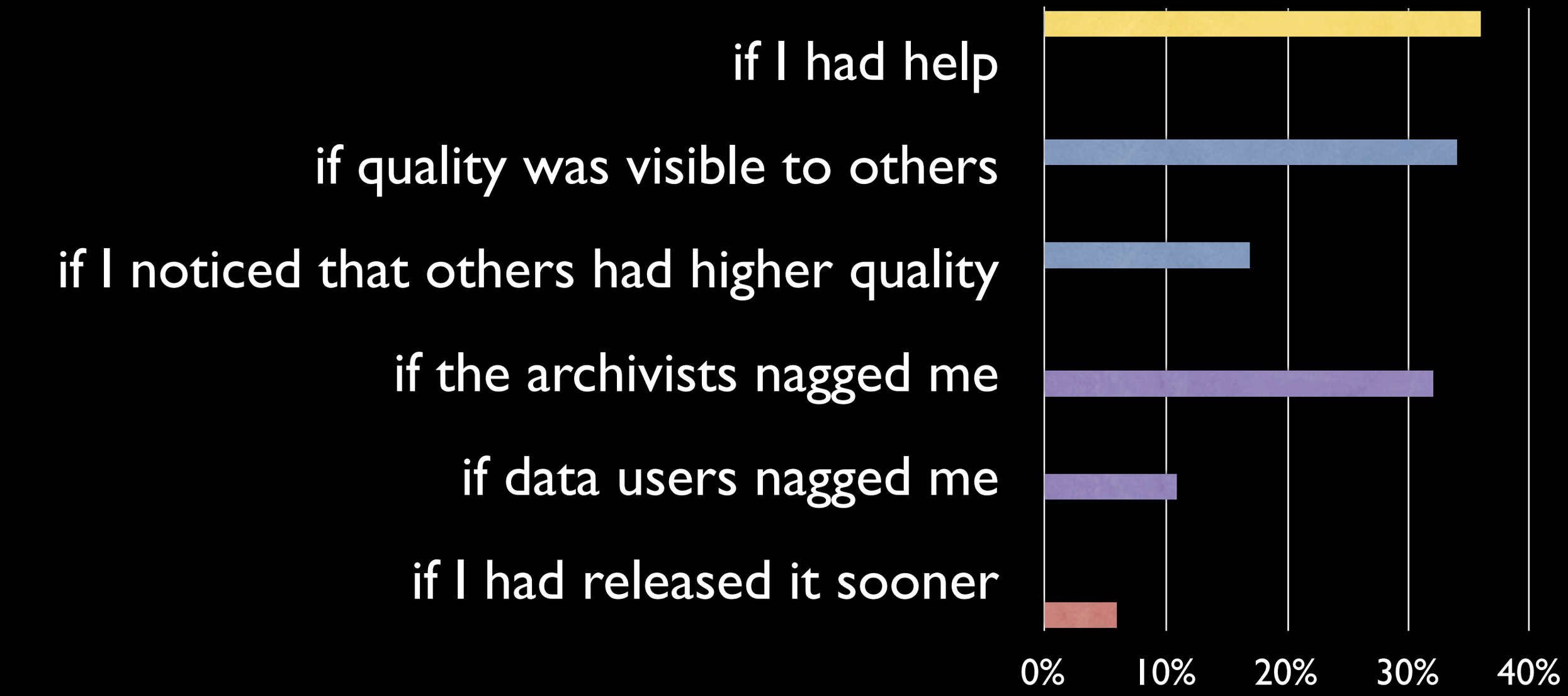
Hedstrom. Society of Am Archivists Ann Meeting. 2008.

What would make it easier? *help and straightforward guidelines*



Hedstrom et al. IASSIST 2006.

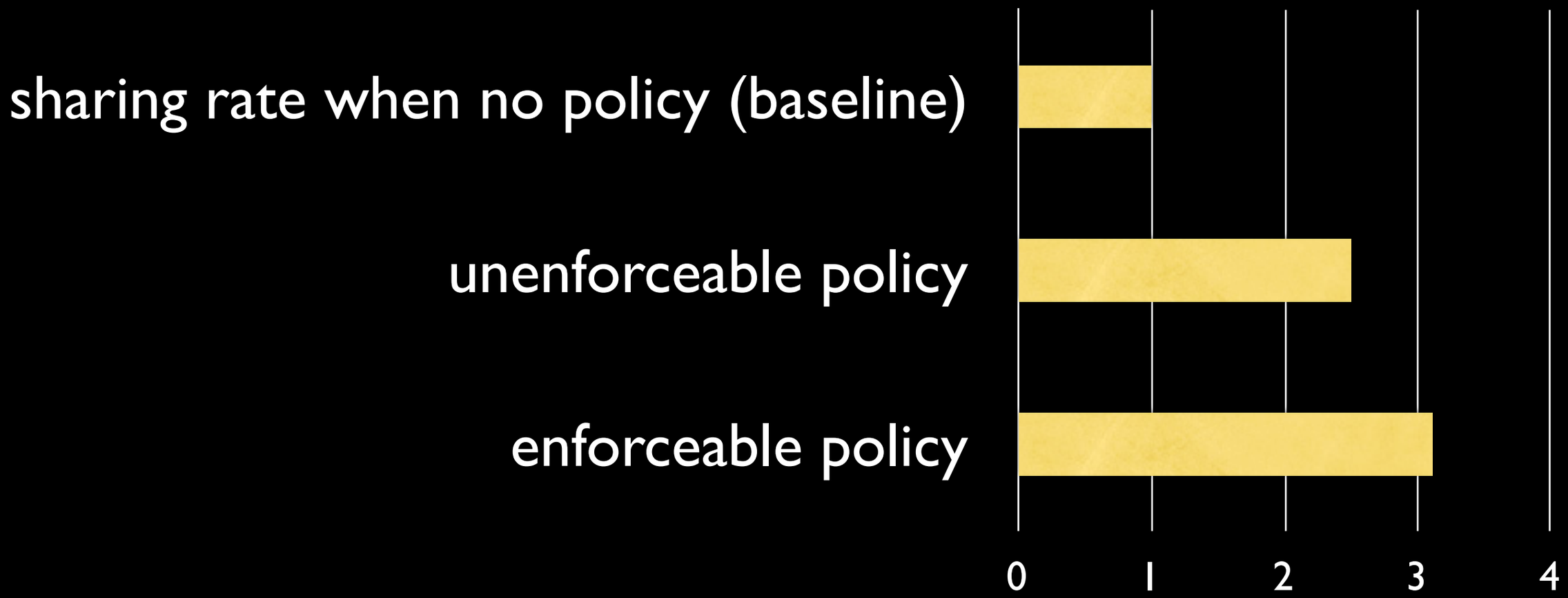
Incentives for quality and docs: *help, visibility, and nagging*



Hedstrom et al. IASSIST 2006.

Do journal mandates work?

Journals with *enforceable policies* have more shared datasets

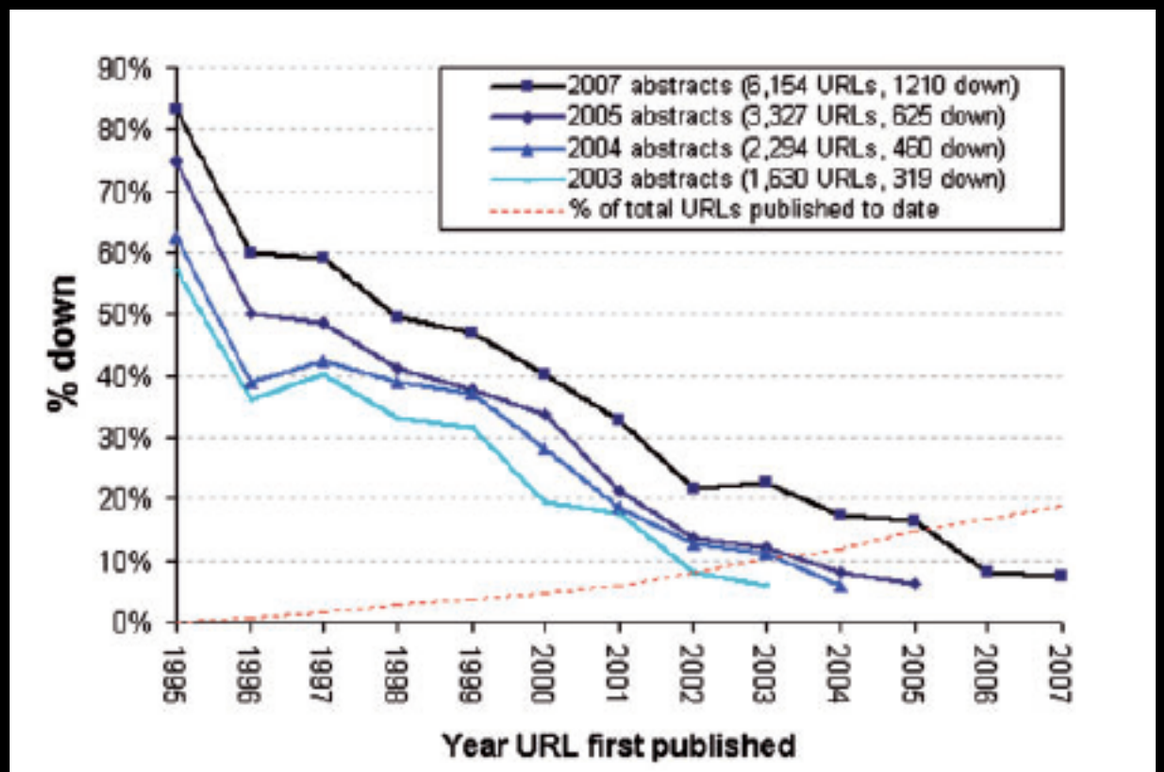


Piwowar, Chapman. A review of journal policies for sharing research data. ELPUB 2008.

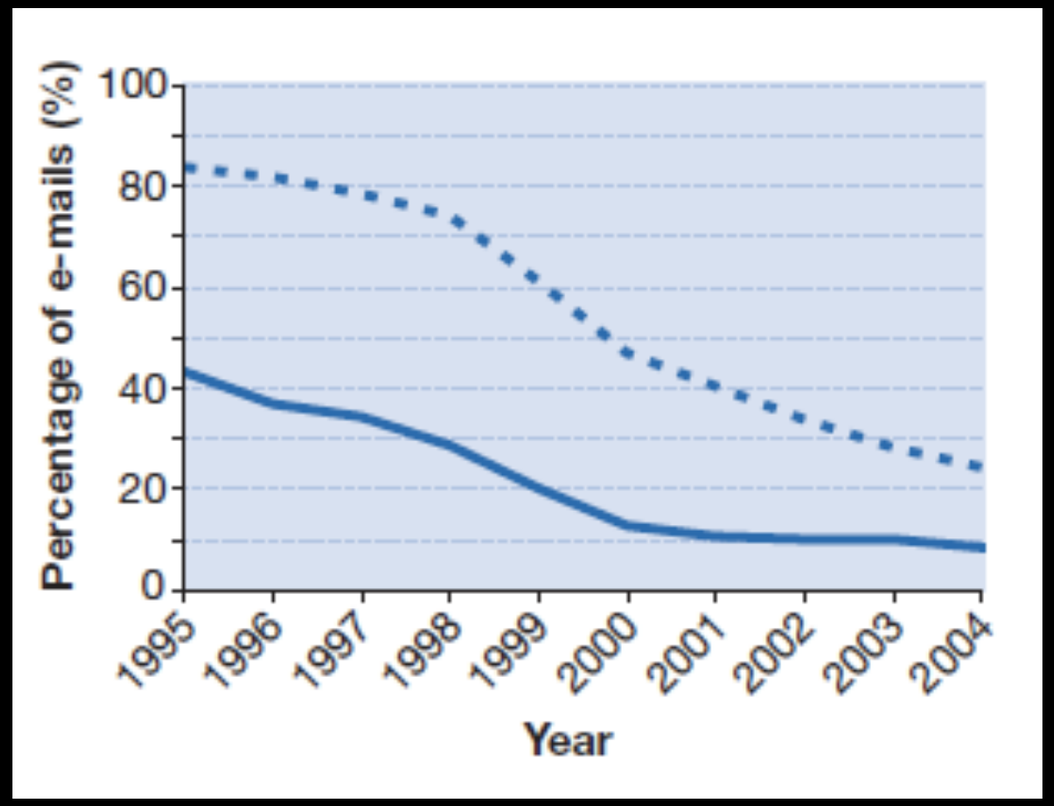
Once shared, always there?

Data contacts and storage *decay with time*

URL decay:



email decay:



Supplementary information: in 6 top journals:
5% unavailable after 2 years, 10% unavail after 5 years

Evangelou et al. FASEB J. 2006.
Wren. Bioinformatics 2008.
Wren et al. EMBO Rep 2006.

Anything else?

data completeness?
replicability?
theoretical models of info behaviour?

Good questions.
Out of time.
Ask or see online bibliography for more info.

Do funder mandates work?

Which subdisciplines have best practices?
particular weaknesses?
why?

Good questions.
Research underway….

NIH: Haga, S.
Exploring Attitudes About Data Disclosure and Data-Sharing in Genomics Research.

NSF: Hedstrom, M.
Incentives for Data Producers to Create Archive-Ready Data Sets.

National Inst of Nursing Research: Pienta, A.
Barriers and Opportunities for Sharing Research Data.

NLM training grant: Piwowar, H.
Impact, prevalence, and patterns of shared biomedical data.

+others

In some cases do the
costs outweigh the benefits?

Do mandates **decrease quality** of shared data?

What is the prevalence of **data reuse**?

What would **facilitate** reuse?

Good questions.
We don't know.
Future research!

Conclusions

Take home #1

Although some researchers voluntarily share data, many don't.

the frequency of sharing depends on
data type,
who you ask,
how you ask,
what you plan to do with the data,
what journal it is published in....

Take home #2

Withholding is correlated with the usual suspects:
desire to publish more, avoid effort, maintain control, industry relationships.

Relative value of incentives is surprising:
demonstrated value, visibility, help, straightforward guidelines, effective mandates, and nagging :)

Each of us can make a difference here:
Write letters to the editor about journal policies,
blog a how-to guide in plain English,
get involved in data standards,
offer help to colleagues,
communicate instances of value.

Take home #3

Much room for future research:
costs and benefits, data quality, reuse ....

Opportunities for traditional large-scale grants across a range of disciplines and agencies

But also opportunity for impact in less formal channels:
You can help communicate
anecdotes, evaluations, and visualizations

via blogs, published research notes, perspectives, letters to the editor, and water-cooler conversations.

you can not manage
what you do not measure

->

If we measure current behaviour,
we'll learn how to **facilitate** the adoption
of open science, and

We'll know what and when to celebrate!

Thanks to

Wendy Chapman + the Dept of Biomedical Informatics at U of Pittsburgh

NLM for training grant funding: 5 T15 LM007059-22 (U of Pitt DBMI)

NIH for research and travel funding: 1R01LM009427-01 (Wendy Chapman)

PSB, Shirley, and Cameron for organizing this workshop

Study references available at http://www.citeulike.org/user/hpiwowar/tag/psb-talk

Contact me for more info at hpiwowar@gmail.com

My shared data: www.dbmi.pitt.edu/piwowar

Share your research data too!