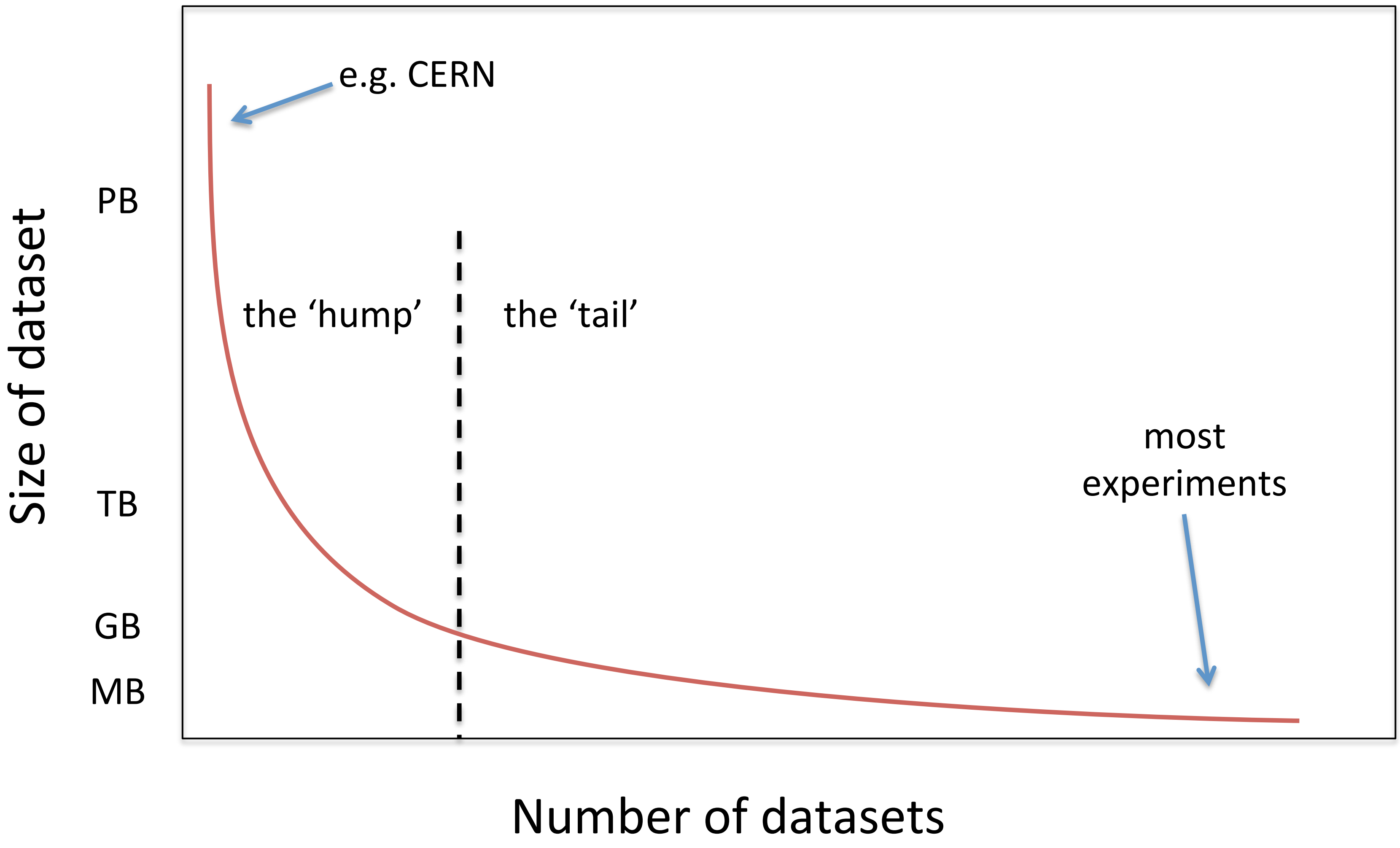
The Long Dark Tail of Research Data

Tim Vines, University of British Columbia

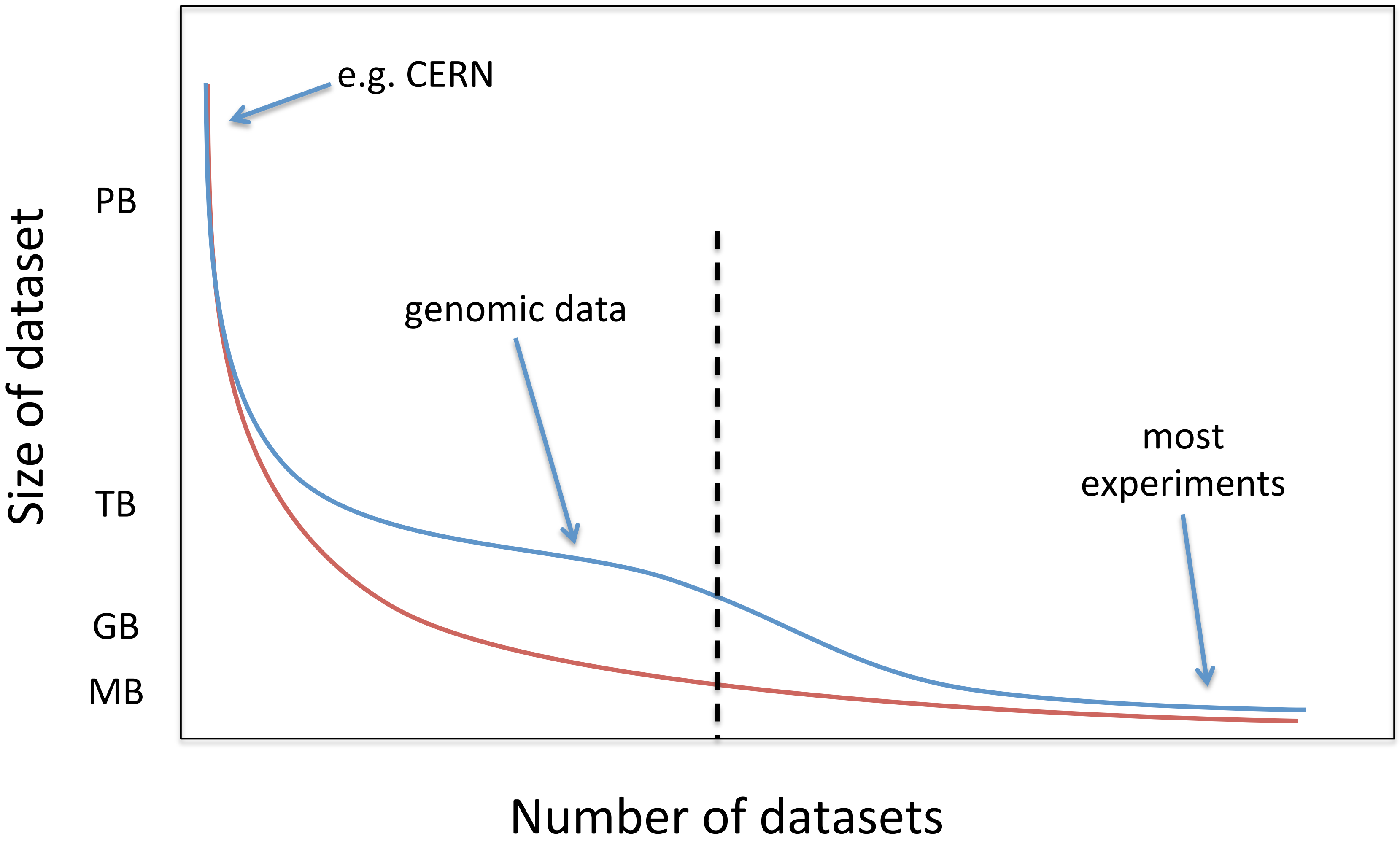
The Long Dark Tail of Research Data

Arianne Albert, Rose Andrew, Florence Débarre, Dan Bock, Michelle Franklin, Kim Gilbert, Nolan Kane, Jean-Sébastien Moore, Brook Moyers, Sébastien Renaut, Diana Rennison, Thor Veen, Tim Vines, and Sam Yeaman

The Long Tail



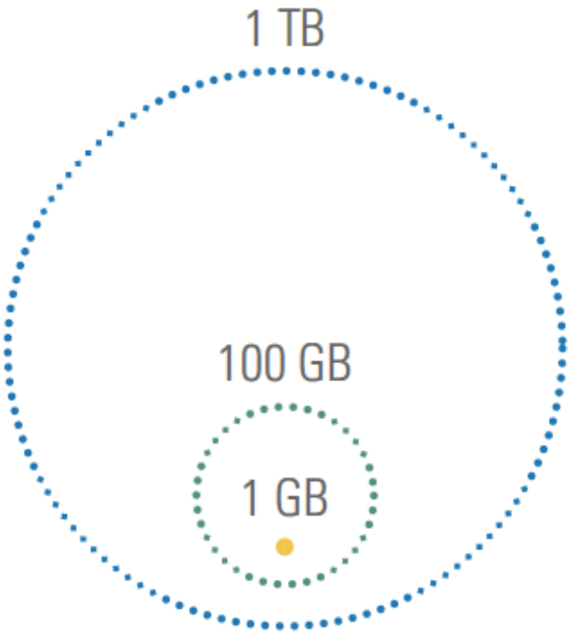
The Long Tail

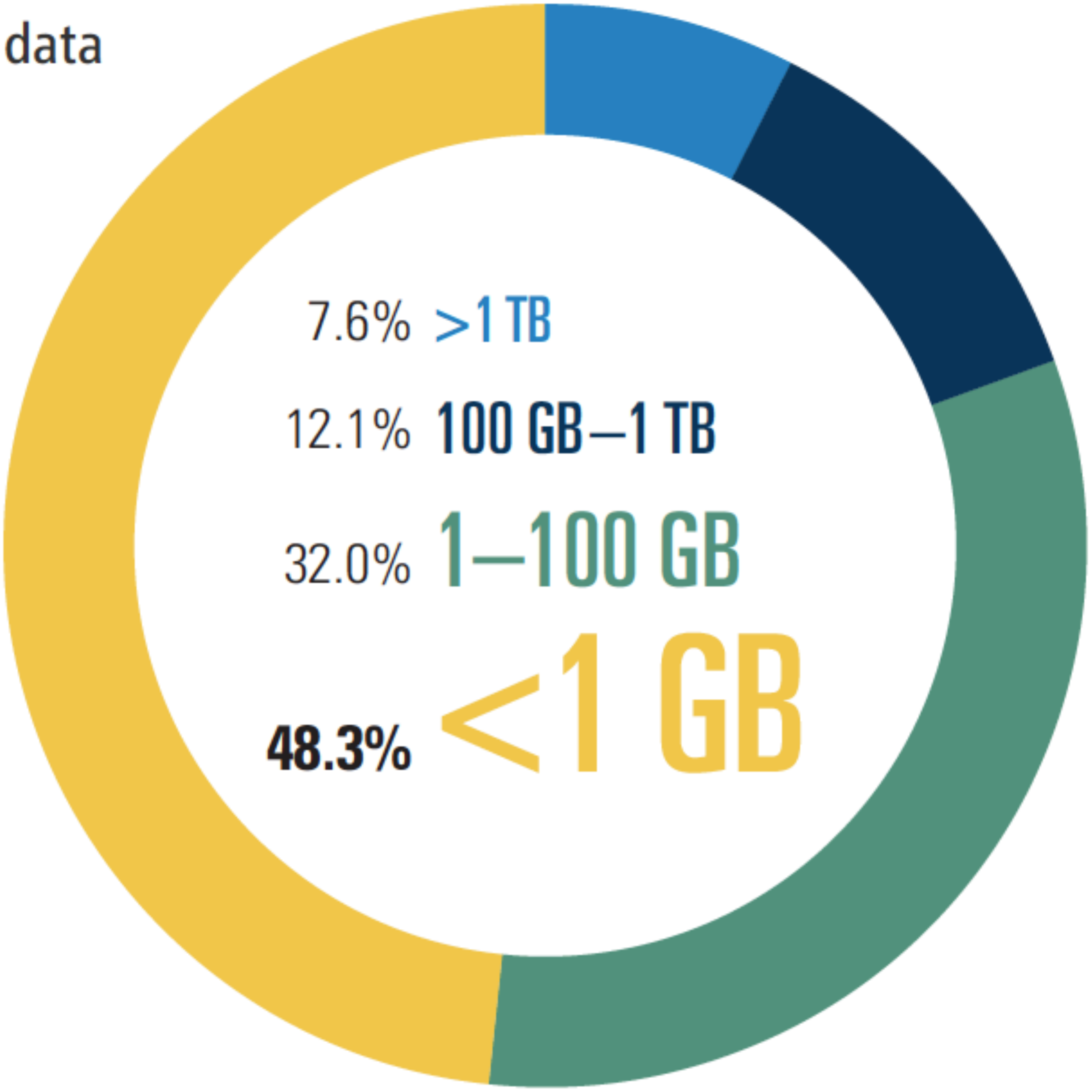


The Long Dark Tail

- The long tail is real

What is the size of the largest data set that you have used or generated in your research?





‘Challenges and Opportunities’ (2011) *Science*

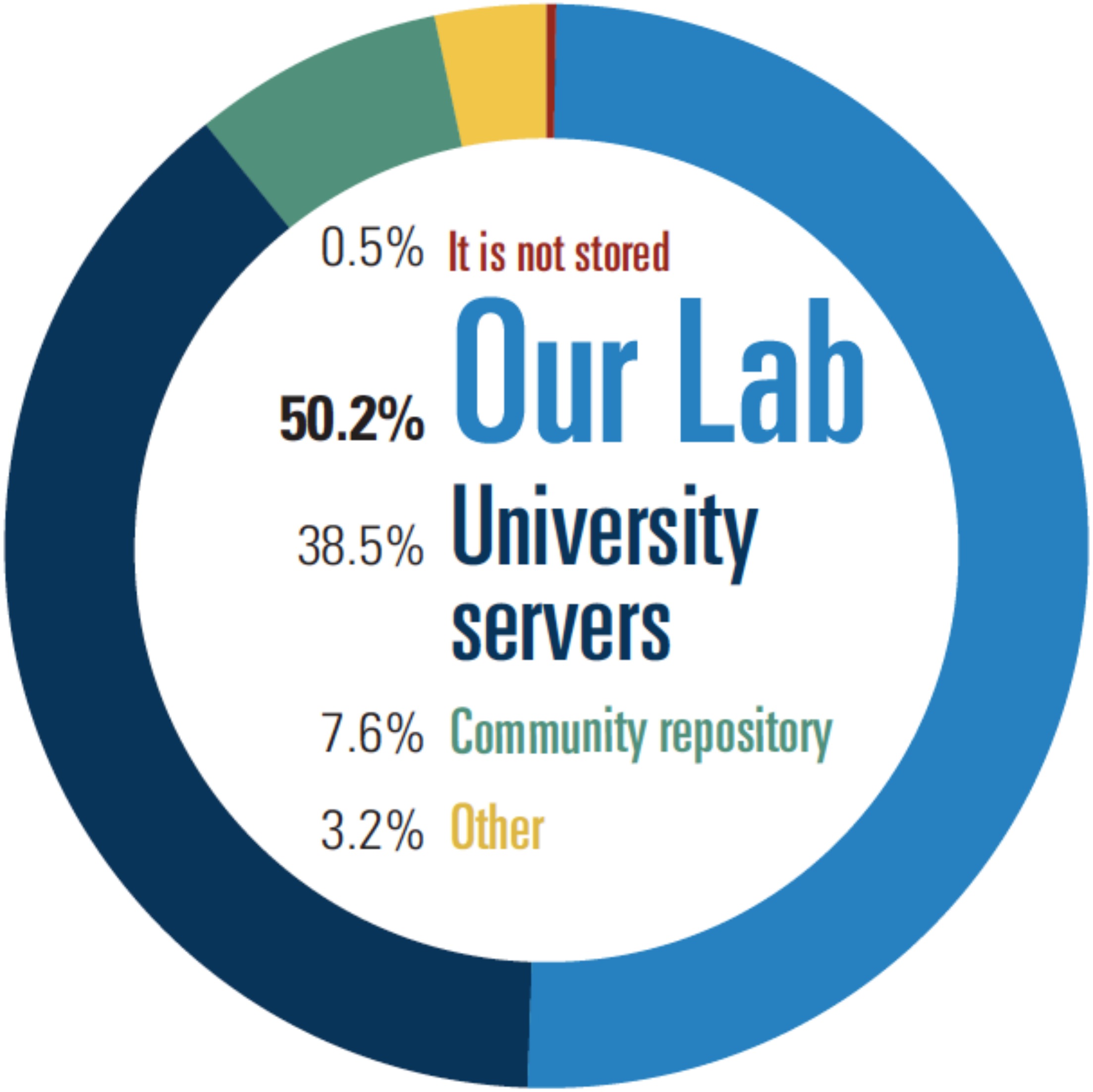
The Long Dark Tail

- The long tail is real

- It’s mostly ‘dark’

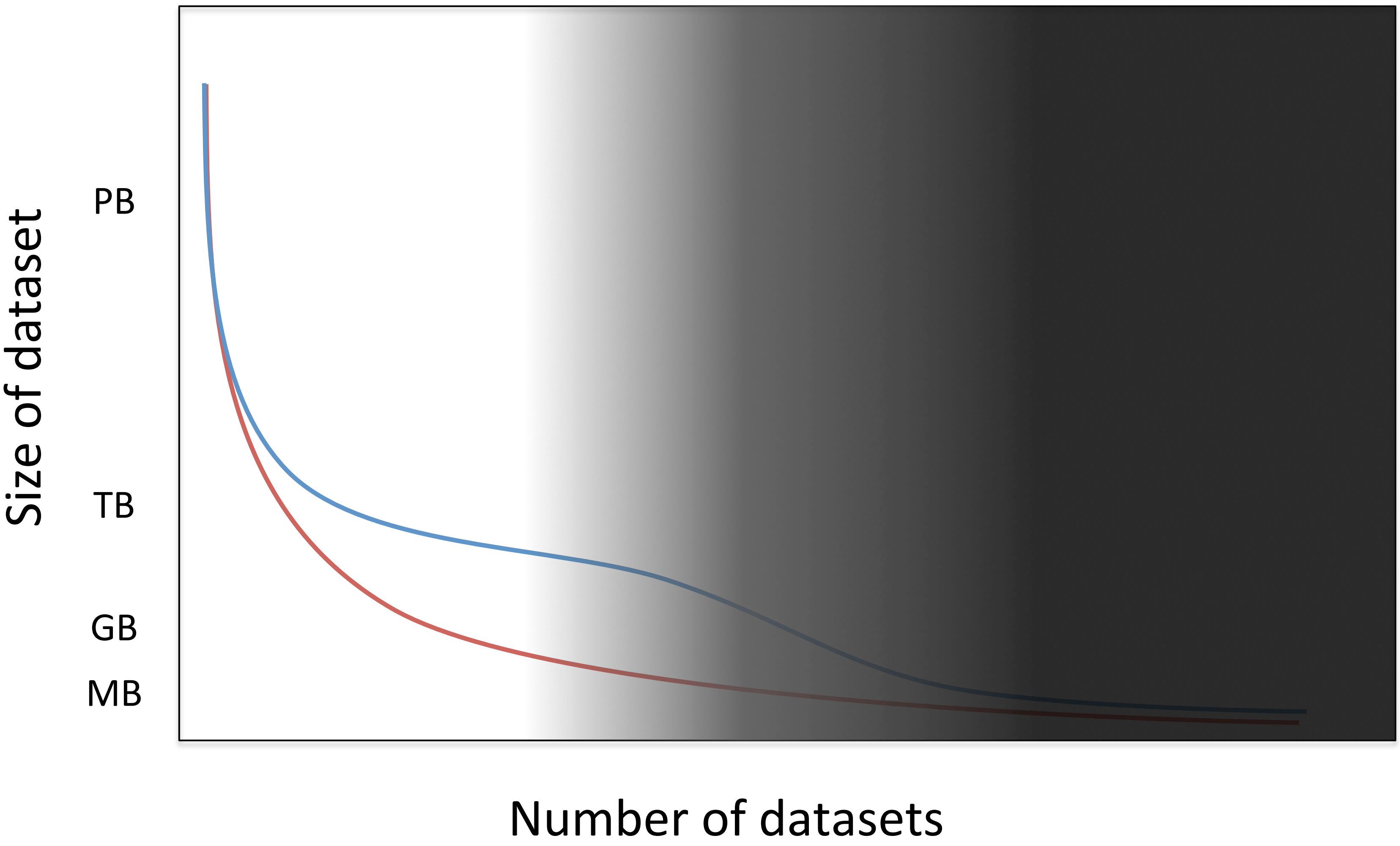
Where do you archive most of the data generated in your lab or for your research?

“Even within a single institution **there are no standards for storing data**, so each lab, or often each fellow, uses ad hoc approaches.”



‘Challenges and Opportunities’ (2011) *Science*

The Long Dark Tail



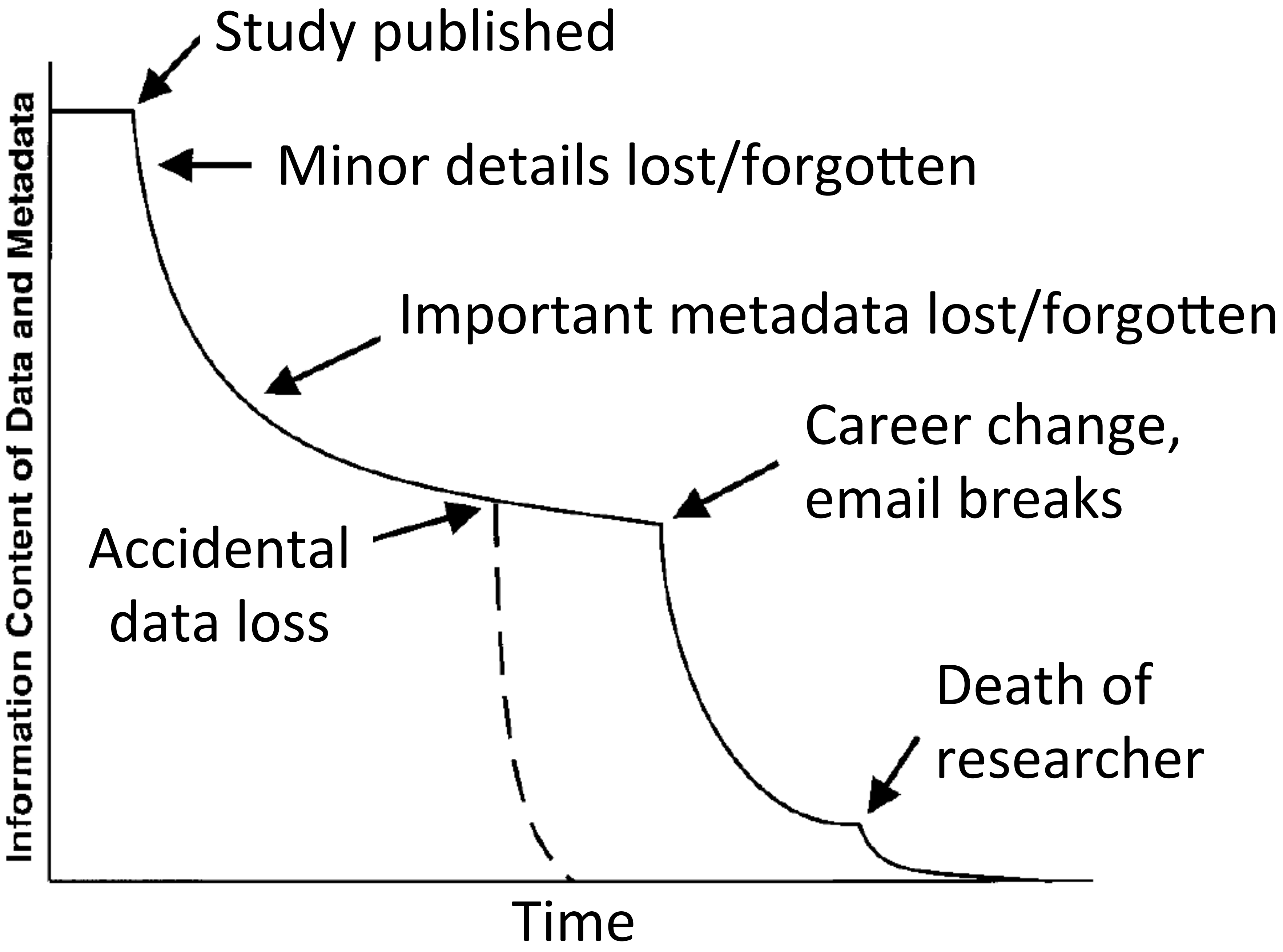
The Long Dark Tail

- What’s happening to these data?

How does the availability of long tail data change with time since publication?

Vines *et al*. Current Biology 2014

Introduction



Study published

Minor details lost/forgotten

Important metadata lost/forgotten

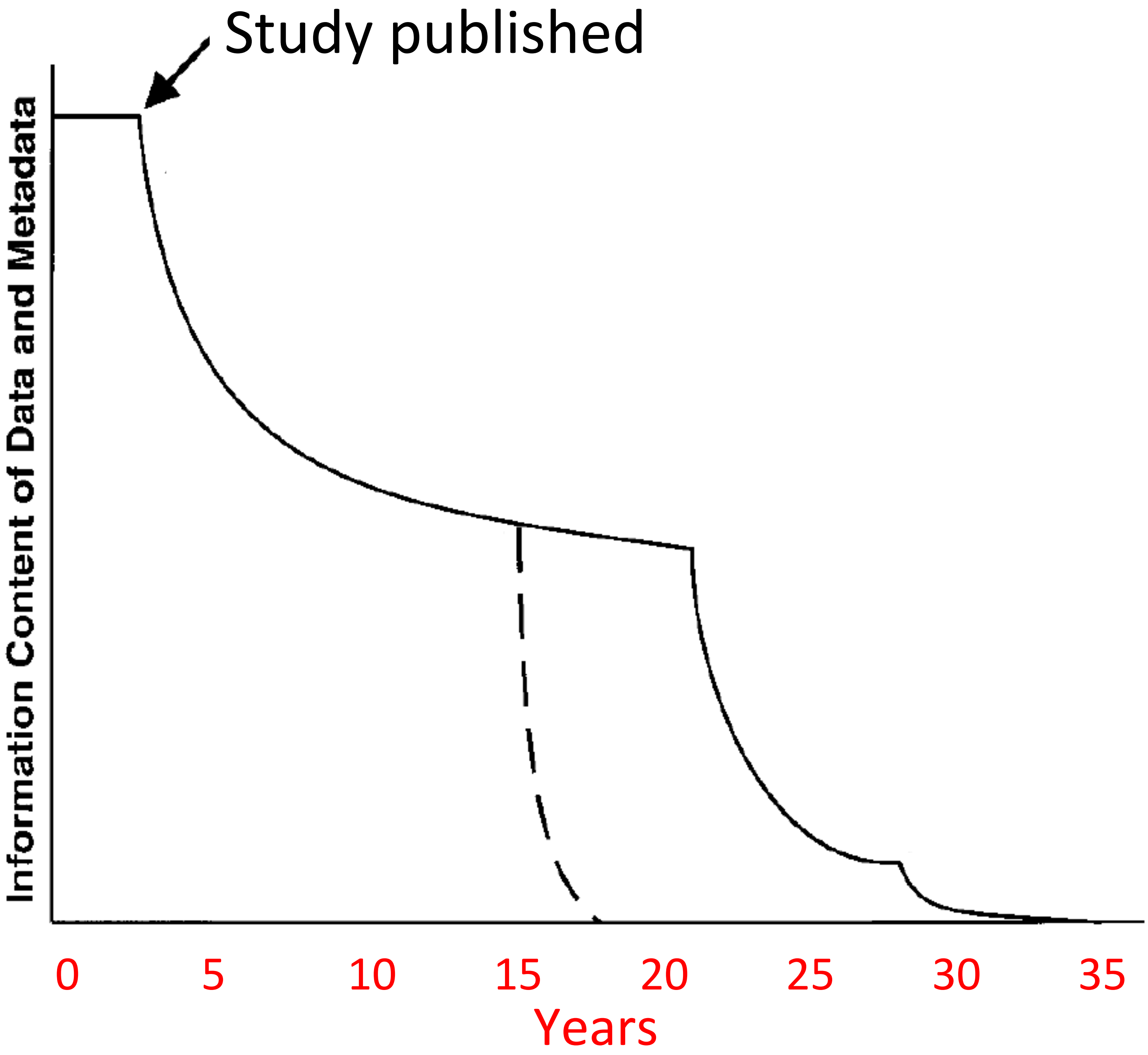
Career change, email breaks

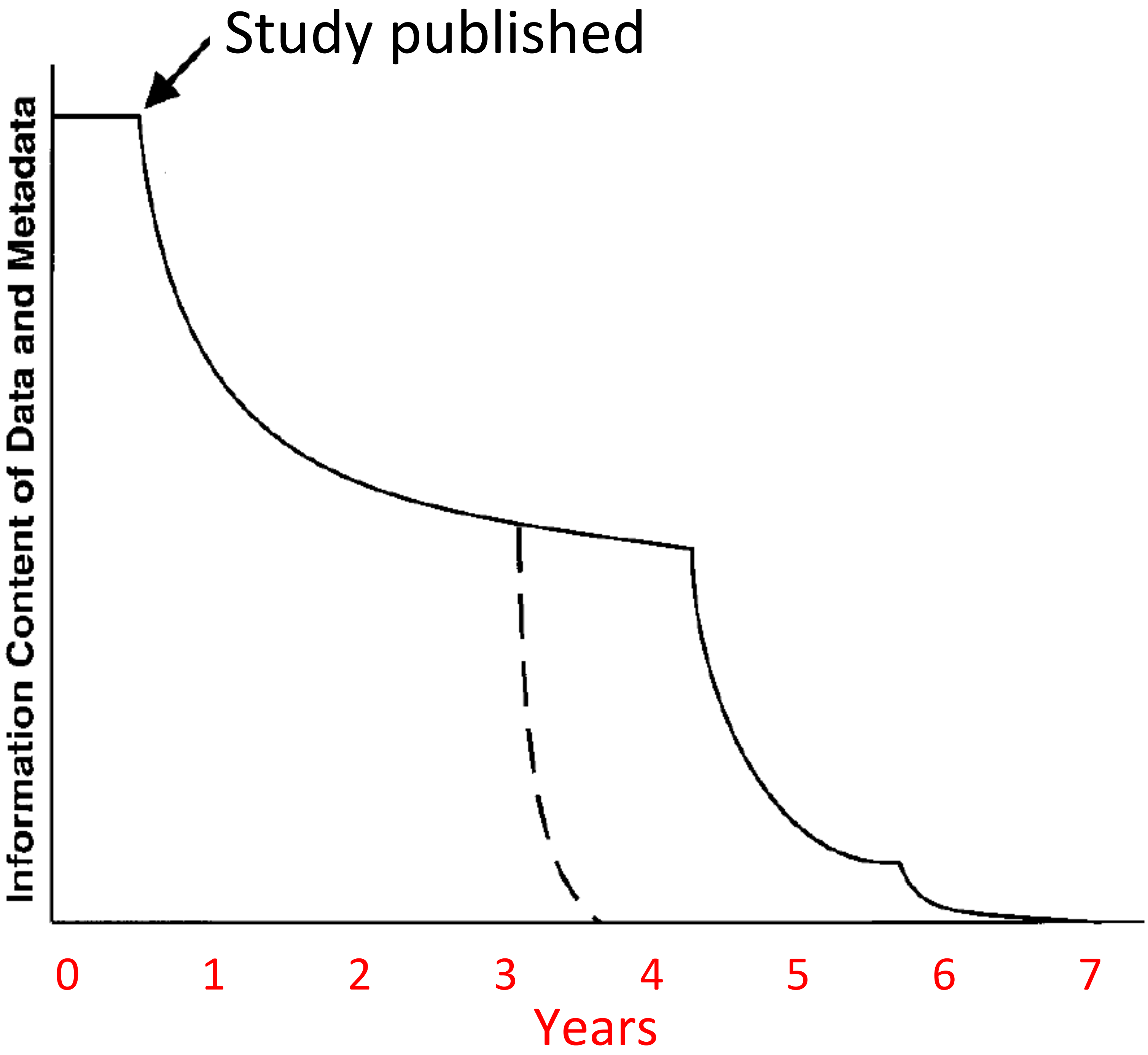
Accidental data loss

Death of researcher

Michener *et al.* (1997) Nongeospatial metadata for the ecological sciences. Ecol. Appl. 7:330

- How fast does this happen?

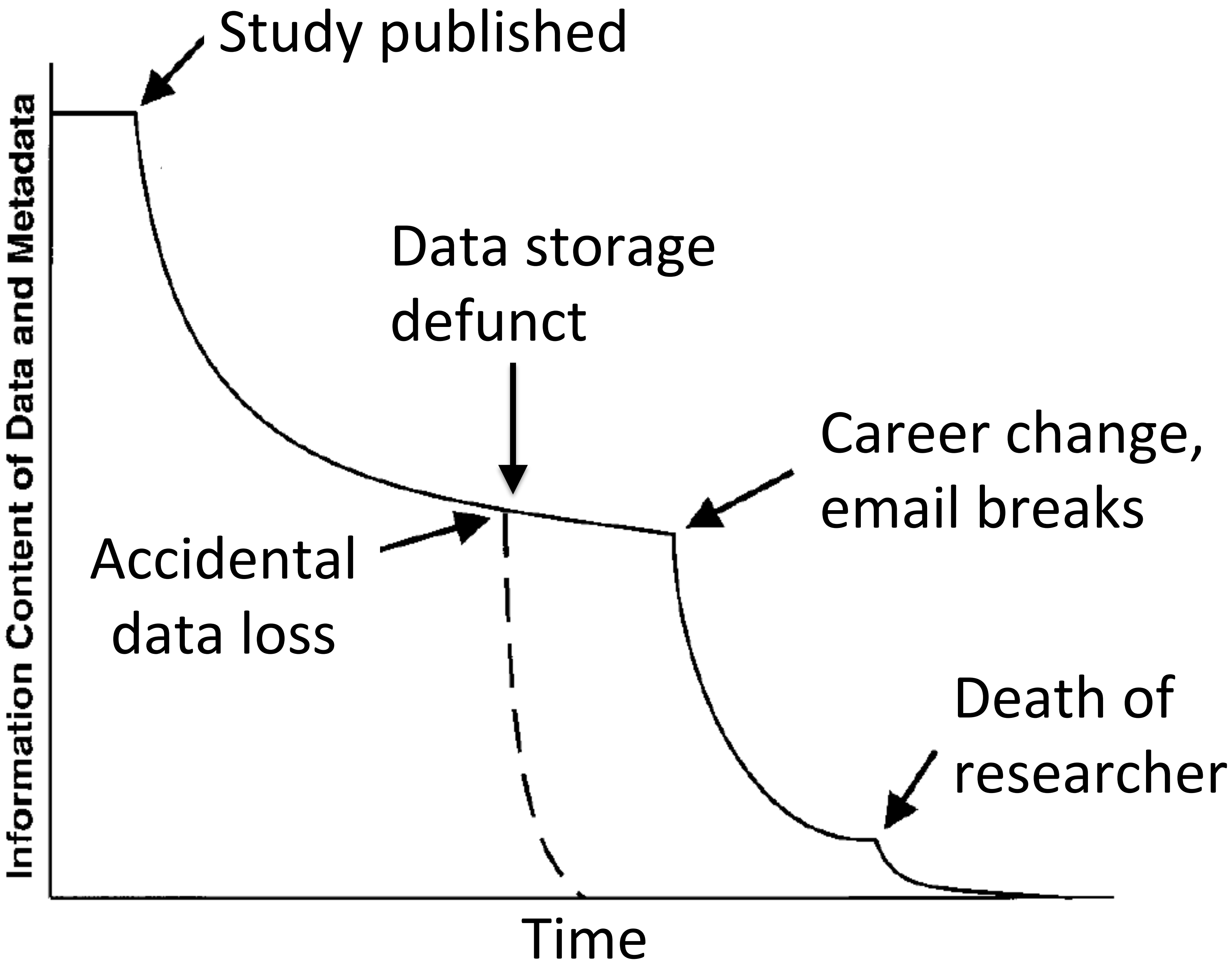

Study published
Information Content of Data and Metadata
0
5
10
15
20
25
30
35
Years



Study published

- How fast does this happen?

- What are the main causes of data loss?


Study published
Data storage defunct
Career change, email breaks
Accidental data loss
Death of researcher
Information Content of Data and Metadata
Time

- How fast does this happen?

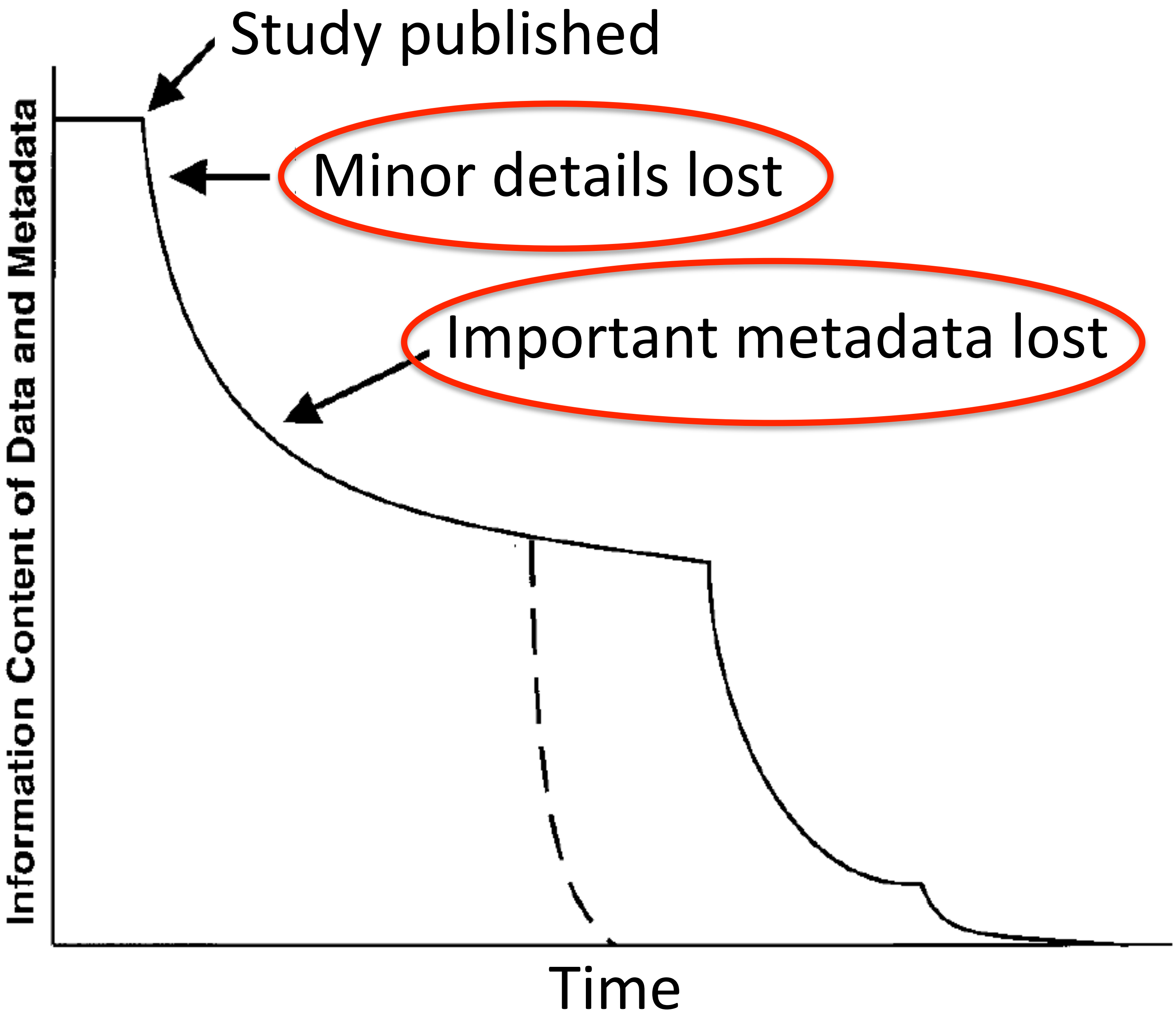
- What are the main causes of data loss?

- Ask for datasets, see how many you get…

Methods

- Need to control for data type
  - morphological data from animals & plants
  - used in a Discriminant Function Analysis

- Reproducing analyses checks the data


Study published
Minor details lost
Important metadata lost
Information Content of Data and Metadata
Time

- 516 studies in odd years 1991 - 2011

- Asked for data by email
  - searched for emails in paper and online
  - contacted first, last & corresponding authors

- “We want to try repeating your DFA”
  - part of study on reproducibility and paper age

- Author motivation :
  - we're trapped in burning building vs
  - we want to print it out for wallpaper

- Our request is fairly common practice
  - expect 20-50% for 2011

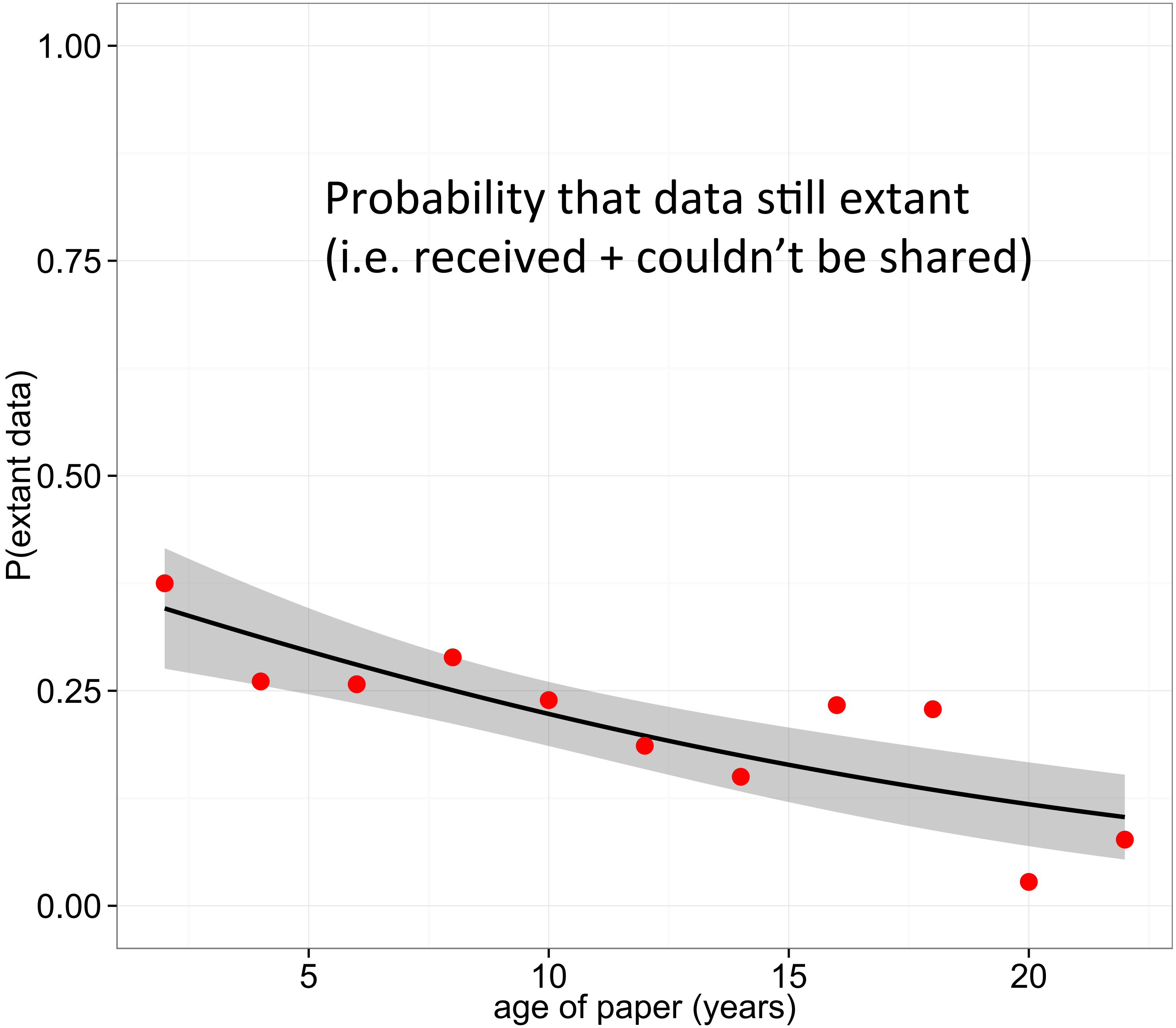
- Motivation sets total % of data we receive

- But our focus is on how % changes with time
  - as long as we get some data we're OK

- If data were gone, we asked for the reason

Results

Probability that data still extant
(i.e. received + couldn’t be shared)

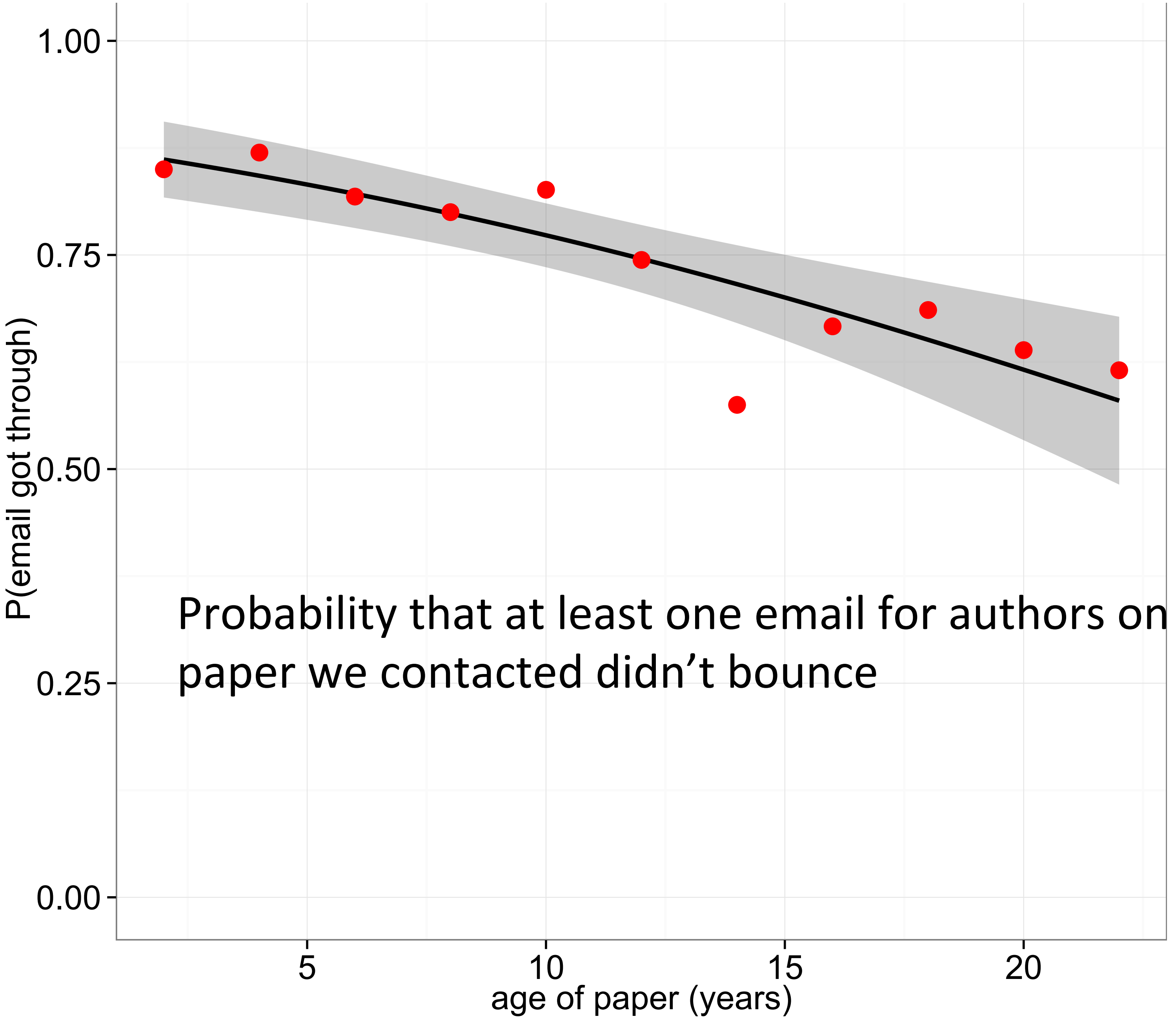


Probability that data still extant
(i.e. received + couldn’t be shared)

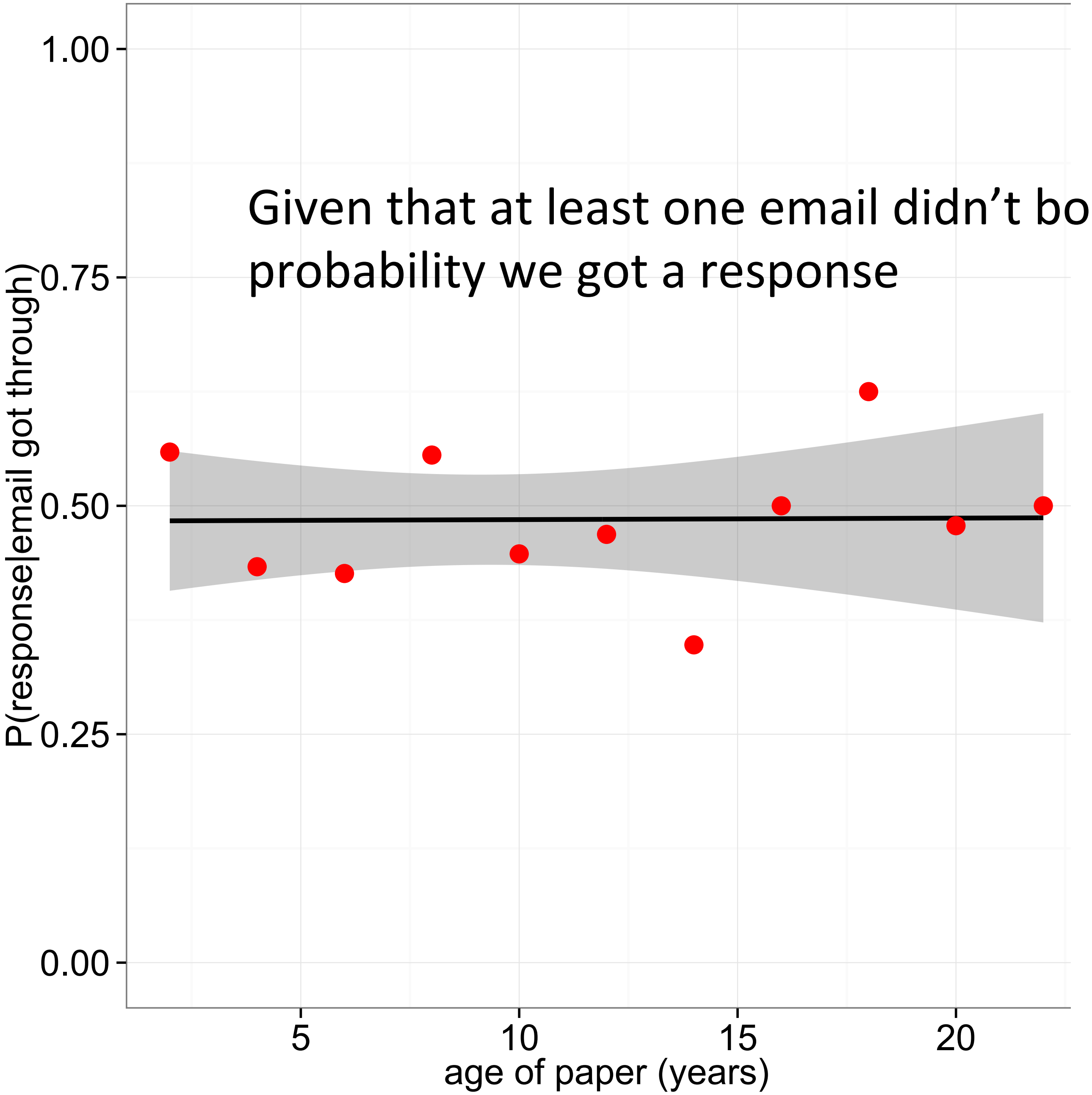
- Odds of data being extant fall by 8% per yr

- Almost all gone after 20 years
  - just 3 of 61 datasets extant for 1991 and 1993

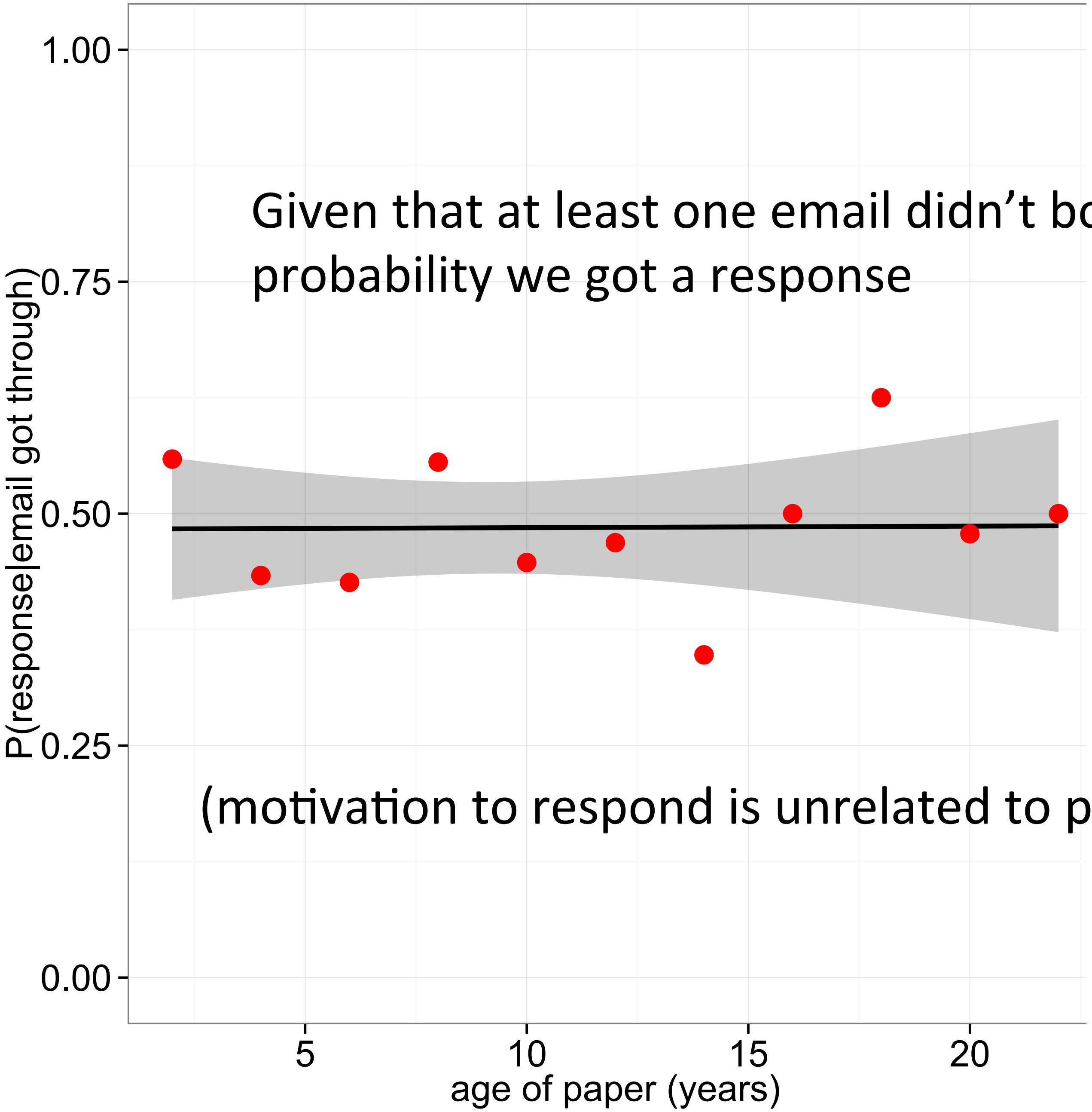
- Why were we unable to get the data?
  - which reasons are related to paper age?



Probability that at least one email for authors on the paper we contacted didn't bounce

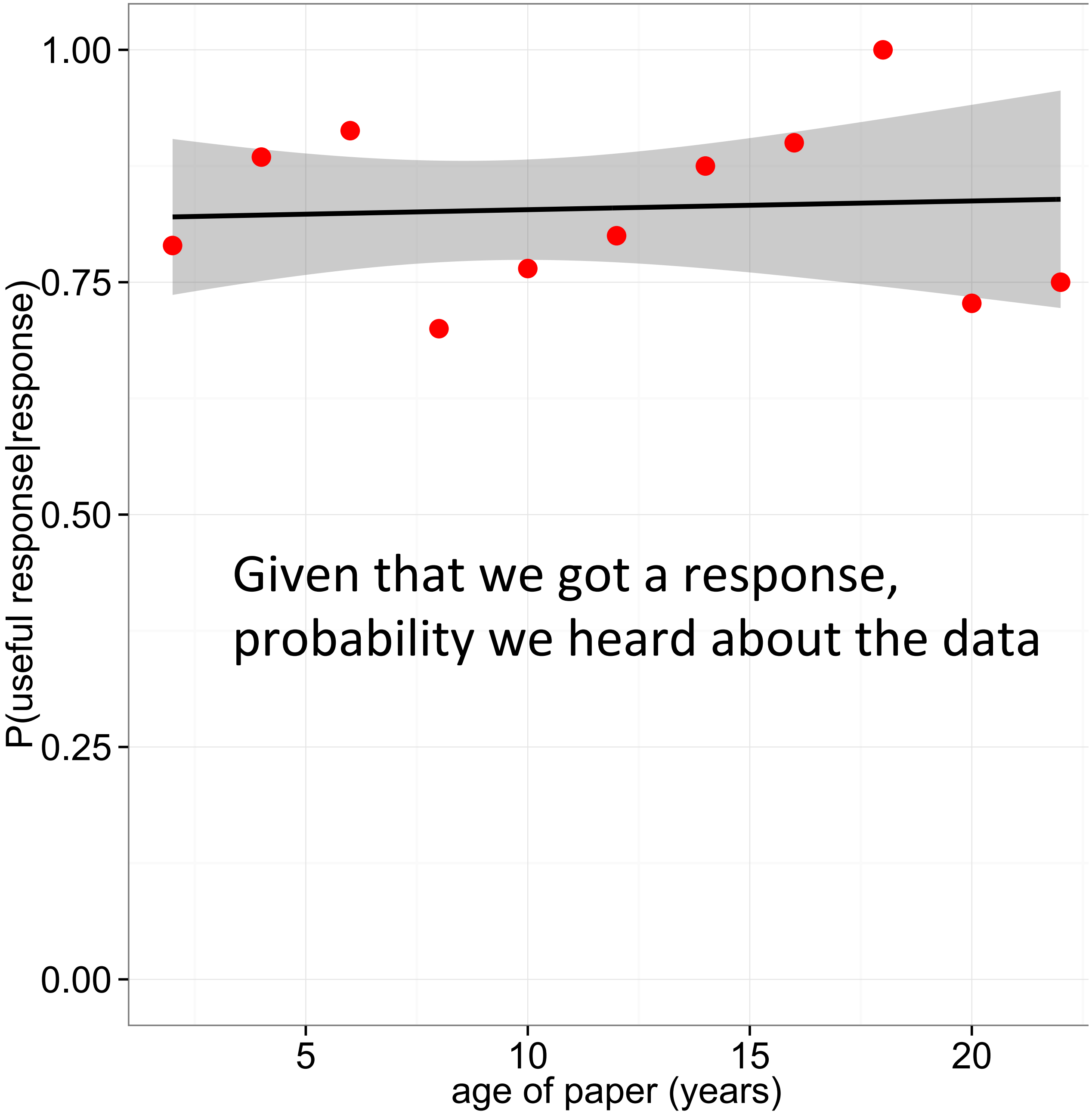


Given that at least one email didn't bounce, probability we got a response

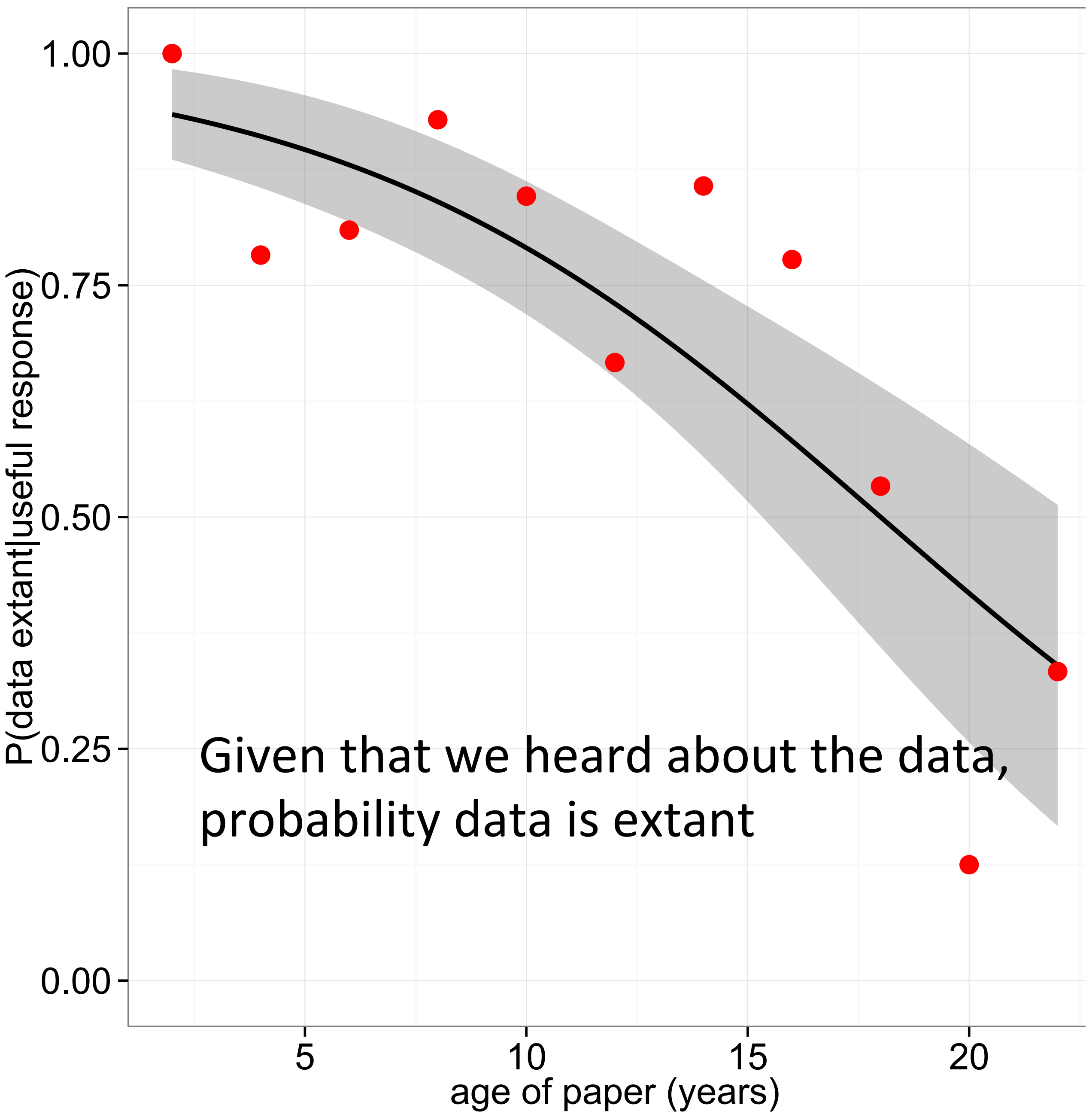


Given that at least one email didn’t bounce, probability we got a response

(motivation to respond is unrelated to paper age)



Given that we got a response, probability we heard about the data



Given that we heard about the data, probability data is extant

Conclusions

- Data held by authors disappears fast

- Almost all gone after 20 years

- Archiving at publication really is crucial

Vines *et al*. Current Biology 2014

The Long Dark Tail

- What’s happening to these data?
  - they’re disappearing!

- How can they be brought into the light?
  - be preserved & made public
  - be re-used in new research

Illuminating the Dark Tail

- First, the bad news…

Many (most?) researchers aren’t interested in

- data curation
- long term data preservation
- data sharing

All they care about is getting publications...

Illuminating the Dark Tail

- First, the bad news…

- Preservating long tail data won't be voluntary

Illuminating the Dark Tail

- Then who's responsible?
  - institutions?
  - funding agencies?
  - journals?

Illuminating the Dark Tail

- Institutions & Funders
  - have influence via $$$
  - but can't monitor data production
  - only hear about data when researchers tell them

Illuminating the Dark Tail

- Journals
  - control access to publication
  - datasets are integral to the paper & hence ‘visible’
  - the data can be delineated
  - can withhold publication until data are shared

Illuminating the Dark Tail

- Who's responsible?

- What can journals do?
  - adopt data sharing policies
  - enforce their data sharing policies

Do data archiving policies work?

- many journals now have data sharing policies

- four flavours:
    1. no policy
    2. recommend
    3. require

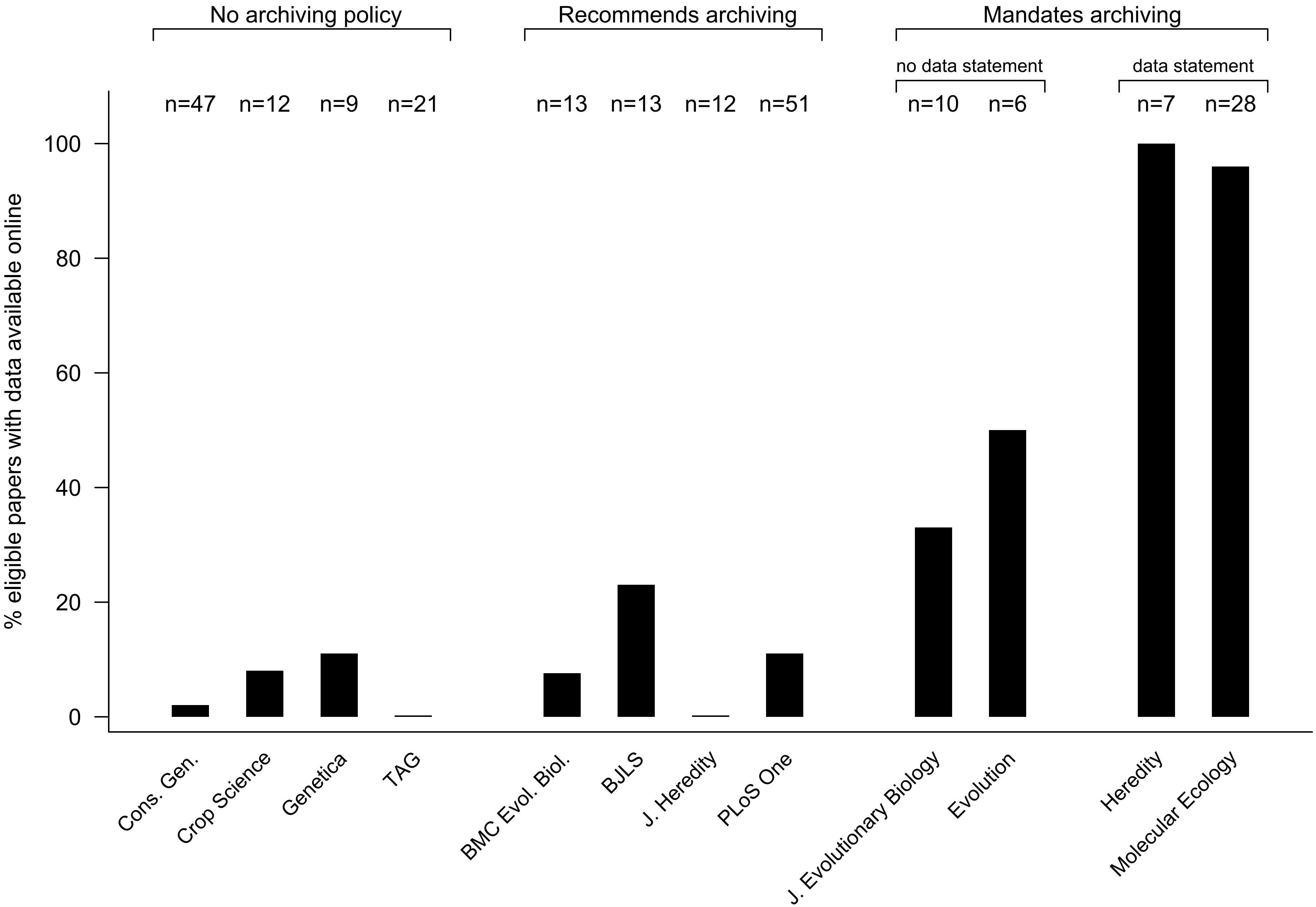
Vines *et al*. (2013) FASEBJ

- many journals now have data sharing policies

- four flavours:
  1. no policy
  2. recommend
  3. require
     a. no ‘data availability’ statement
     b. ‘data availability’ statement

Vines *et al*. (2013) FASEBJ

- focus on single type of data
  - genetic data used in STRUCTURE
- must have established online archive
  - in this case Dryad (or supp. mat.)
- found 229 papers from 2011-12
  - what % had data available?


No archiving policy
Recommends archiving
Mandates archiving
no data statement
data statement
n=47
n=12
n=9
n=21
n=13
n=13
n=12
n=51
n=10
n=6
n=7
n=28
% eligible papers with data available online
100
80
60
40
20
0
Cons. Gen.
Crop Science
Genetica
TAG
BMC Evol. Biol.
BJLS
J. Heredity
PLoS One
J. Evolutionary Biology
Evolution
Heredity
Molecular Ecology

Motivating data sharing

- wide range of approaches available:
  - ask for the data (but don’t follow up)
  - check whether any data is there
  - check whether all data is there
  - check it matches data used in paper
  - reproduce some basic values
  - reproduce all results

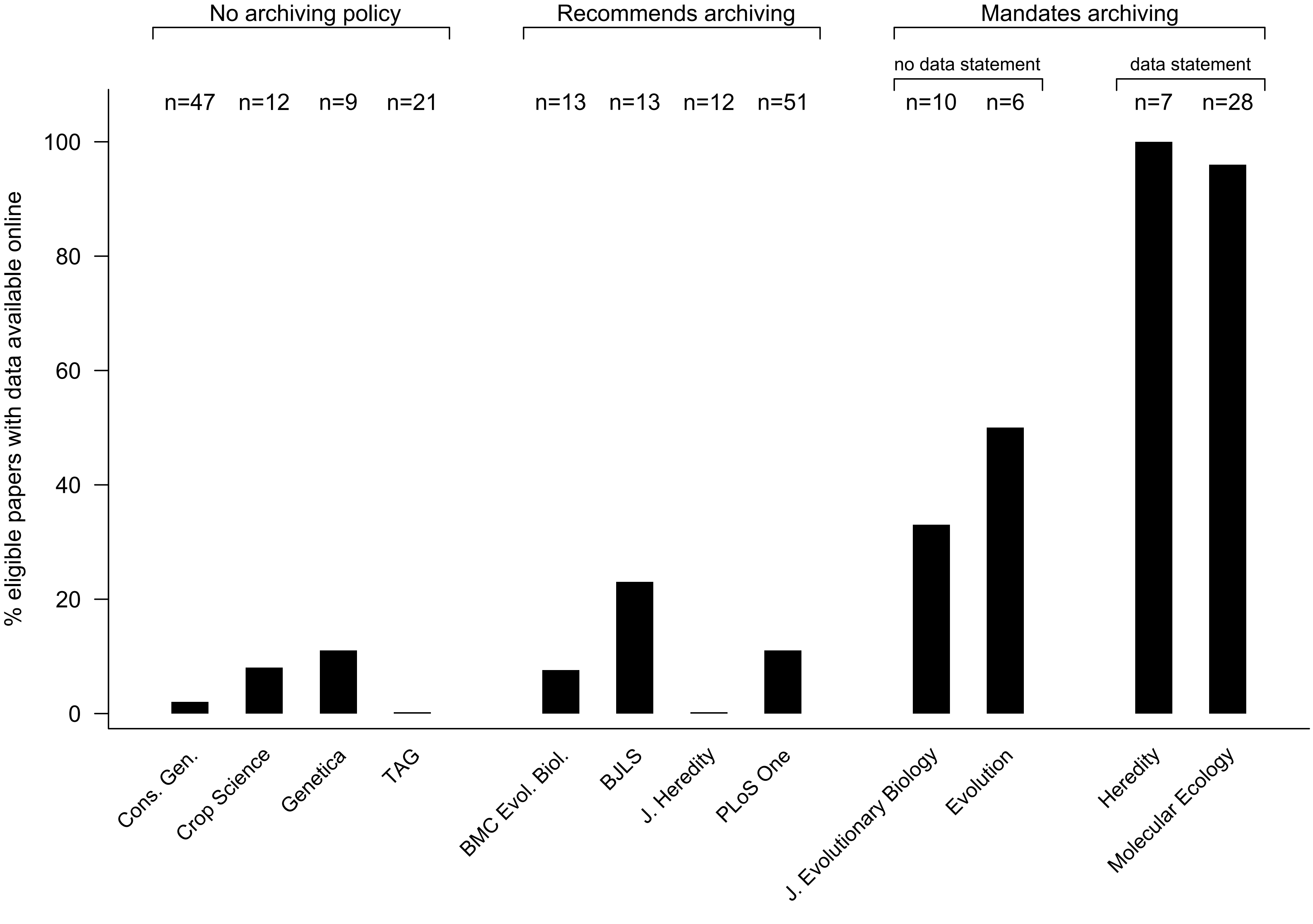
Easy steps

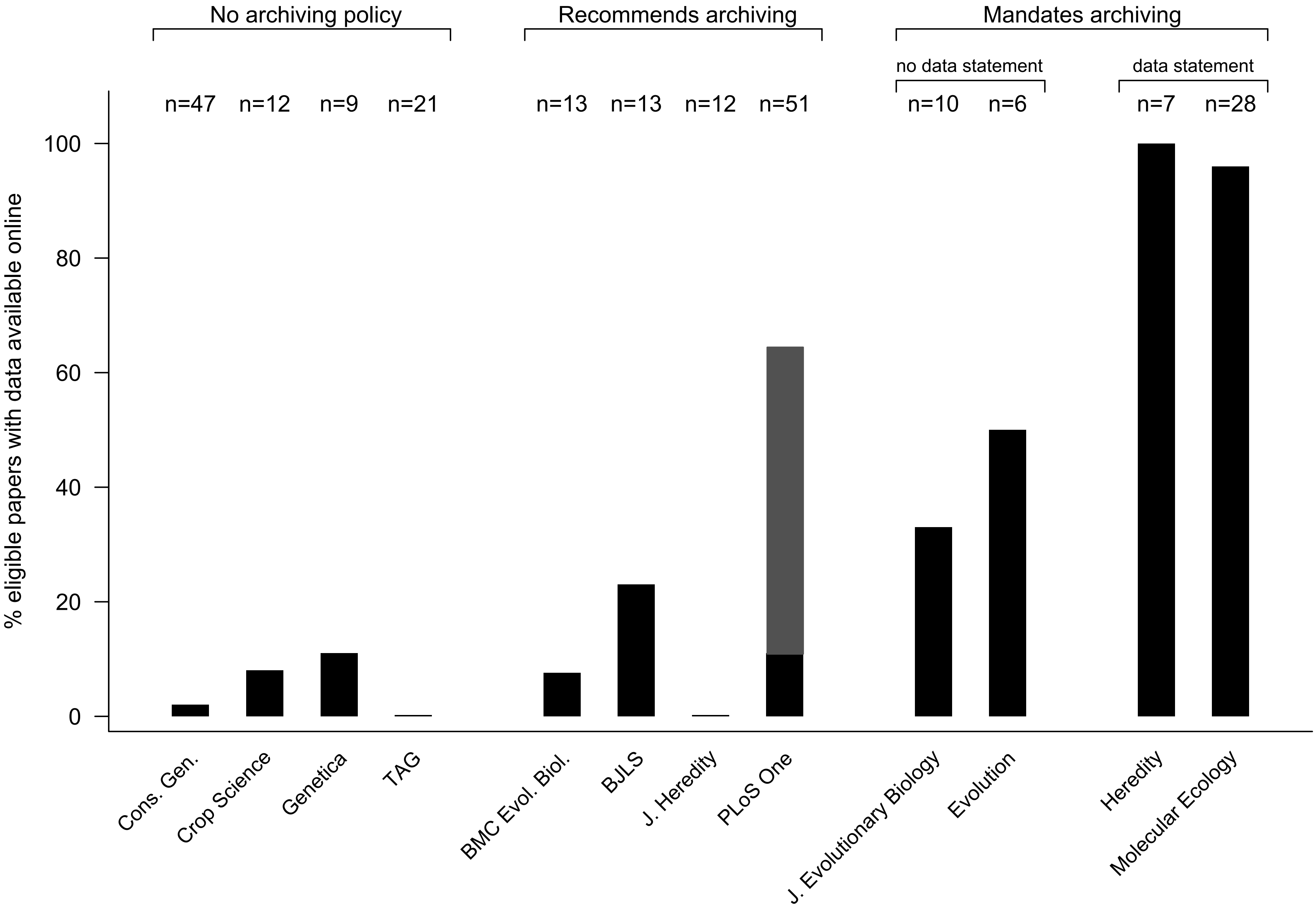
1) Data statements

1) Data statements

These can be effective!

- we repeated analysis for PLoS ONE in 2016
- new data statement rule from March ’14
- 67% of datasets now available


No archiving policy
Recommends archiving
Mandates archiving
no data statement
data statement
n=47
n=12
n=9
n=21
n=13
n=13
n=12
n=51
n=10
n=6
n=7
n=28
% eligible papers with data available online
100
80
60
40
20
0
Cons. Gen.
Crop Science
Genetica
TAG
BMC Evol. Biol.
BJLS
J. Heredity
PLoS One
J. Evolutionary Biology
Evolution
Heredity
Molecular Ecology


No archiving policy
Recommends archiving
Mandates archiving
no data statement
data statement
n=47
n=12
n=9
n=21
n=13
n=13
n=12
n=51
n=10
n=6
n=7
n=28
% eligible papers with data available online
0
20
40
60
80
100
Cons. Gen.
Crop Science
Genetica
TAG
BMC Evol. Biol.
BJLS
J. Heredity
PLoS One
J. Evolutionary Biology
Evolution
Heredity
Molecular Ecology

1) Data statements

- we repeated analysis for PLoS ONE in 2016
- new data statement rule from March ’14
- 67% of datasets now available

- EO checks for statement at initial submission
  - moves archiving earlier in the review process

2) Link data archiving to paper quality

2) Link data archiving to paper quality

- add expectations to author guidelines:

“Papers with exemplary data and code archiving are more valuable for future research, and, all else being equal, these are more likely to get accepted for publication”

Medium steps

3) Ask reviewers to assess data statement:

“Does the Data Accessibility section list all the datasets needed to recreate the results in the manuscript? If ‘No’, please specify which additional data are needed in your comments to the authors.”

[Yes/No/I didn’t check]

- We find only ~15% provide good feedback

4) Ask editors to assess data statement:

- include comments in decision letter
- we didn't explicitly try this at *Molecular Ecology*
    - we would likely run into resistance

Harder steps

5) EO checks data statement

- checks whether datasets listed in paper are also listed in statement (& available)

- have to do all papers, or none
  - can't be inconsistent
  - requires PhD in journal's field

- very effective, but hard work

6) Bring in data reviewers

- keen people to check data statement
- could also do some sanity checks
- only for papers about to be accepted
  - Molecular Ecology hasn’t tried this

Norris steps

7) Full test for reproducibility

- only for fanatics or masochists?

- would certainly motivate/terrify authors

The best lever: peer review

Papers with bad data are bad papers

8) Papers with bad data are bad papers

- authors must upload data & code

- if the data or code are a mess:

  ➔Editorial reject

- if the reviewers find errors in data or code

  ➔Reject, maybe resubmit

8) Papers with bad data are bad papers

- good data management should be a matter of professional pride
  - backed up by reputation damage

- all stakeholders need to emphasize this
  - otherwise we're just wasting resources

Thanks to:

Arianne Albert
Florence Débarre
Michelle Franklin
Nolan Kane
Brook Moyers
Diana Rennison
Thor Veen
Sam Yeaman

Rose Andrew
Dan Bock
Kim Gilbert
Jean-Sébastien Moore
Sébastien Renaut
Loren Rieseberg
Mike Whitlock