The crumbling wall: data archiving and reproducibility in published science

Tim Vines, University of British Columbia

The crumbling wall: data archiving and reproducibility in published science

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

OPEN ACCESS Freely available online PLOS ONE

Priming Intelligent Behavior: An Elusive Phenomenon

David R. Shanks[1]*, Ben R. Newell[2], Eun Hee Lee[1], Divya Balakrishnan[1], Lisa Ekelund[1], Zarus Cenac[1], Fragkiski Kavvadia[1], Christopher Moore[2]

1 Division of Psychology and Language Sciences, University College London, London, United Kingdom, **2** School of Psychology, University of New South Wales, Sydney, Australia

Abstract

Can behavior be unconsciously primed via the activation of attitudes, stereotypes, or other concepts? A number of studies have suggested that such priming effects can occur, and a prominent illustration is the claim that individuals' accuracy in answering general knowledge questions can be influenced by activating intelligence-related concepts such as *professor* or *soccer hooligan*. In 9 experiments with 475 participants we employed the procedures used in these studies, as well as a number of variants of those procedures, in an attempt to obtain this intelligence priming effect. None of the experiments obtained the effect, although financial incentives did boost performance. A Bayesian analysis reveals considerable evidential support for the null hypothesis. The results conform to the pattern typically obtained in word priming experiments in which priming is very narrow in its generalization and unconscious (subliminal) influences, if they occur at all, are extremely short-lived. We encourage others to explore the circumstances in which this phenomenon might be obtained.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

Reproducibility

- Science is the search for general ‘rules’

Reproducibility

- Science is the search for general ‘rules’
- Replication tests different circumstances

Reproducibility

- Science is the search for general 'rules'
- Replication tests different circumstances
- Reproducibility checks existing results

Reproducibility

- We hope bad papers will be discarded

Reproducibility

- We hope bad papers will be discarded

- But maybe many papers are ‘wrong’?
  - We need to quantify this problem…

Reproducibility

- Reproducibility needs the original data
- Then we need to repeat the analyses

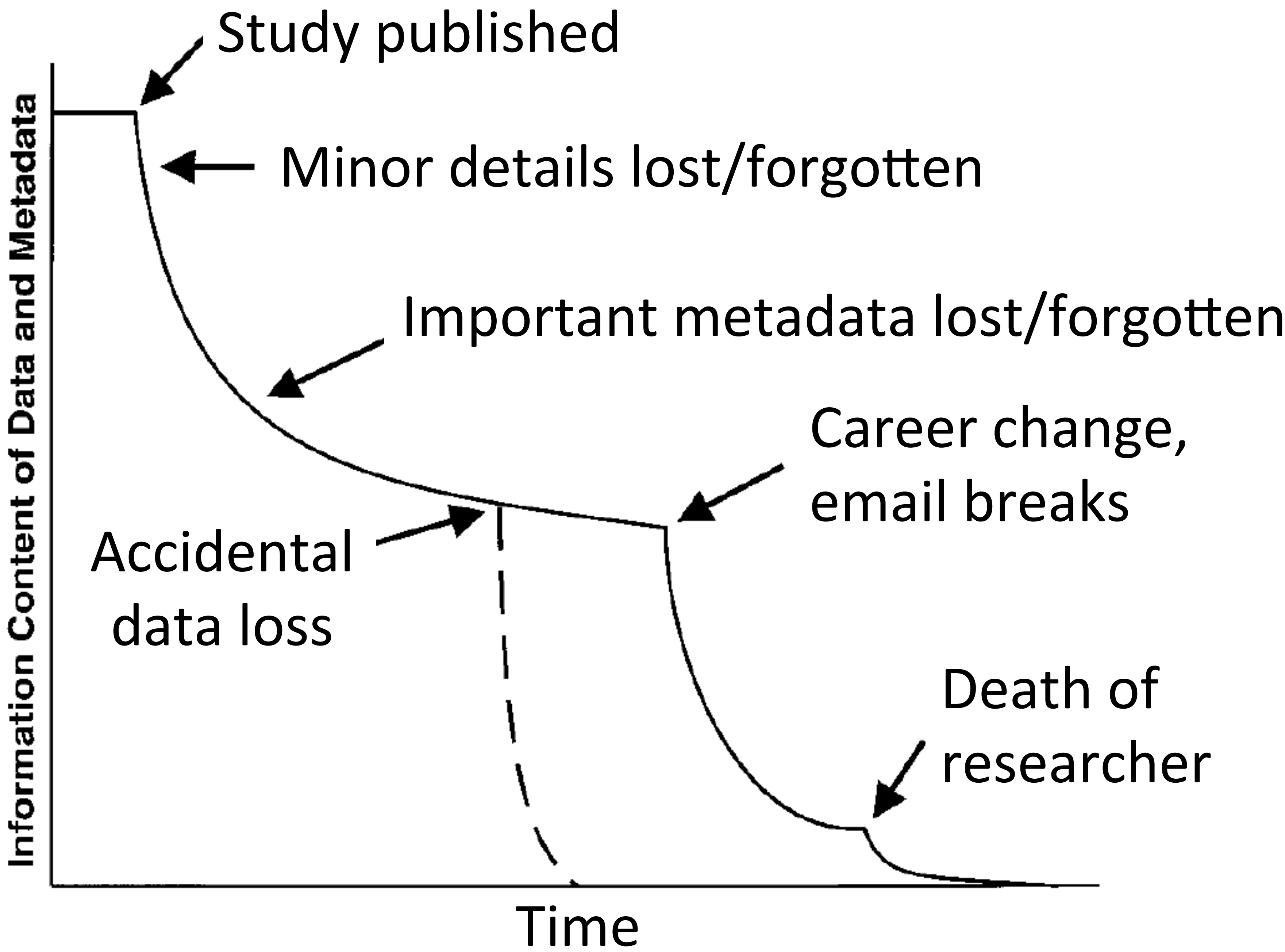
Reproducibility

- Reproducibility needs the original data
- Then we need to repeat the analyses
- Here are two iterations of this process…

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

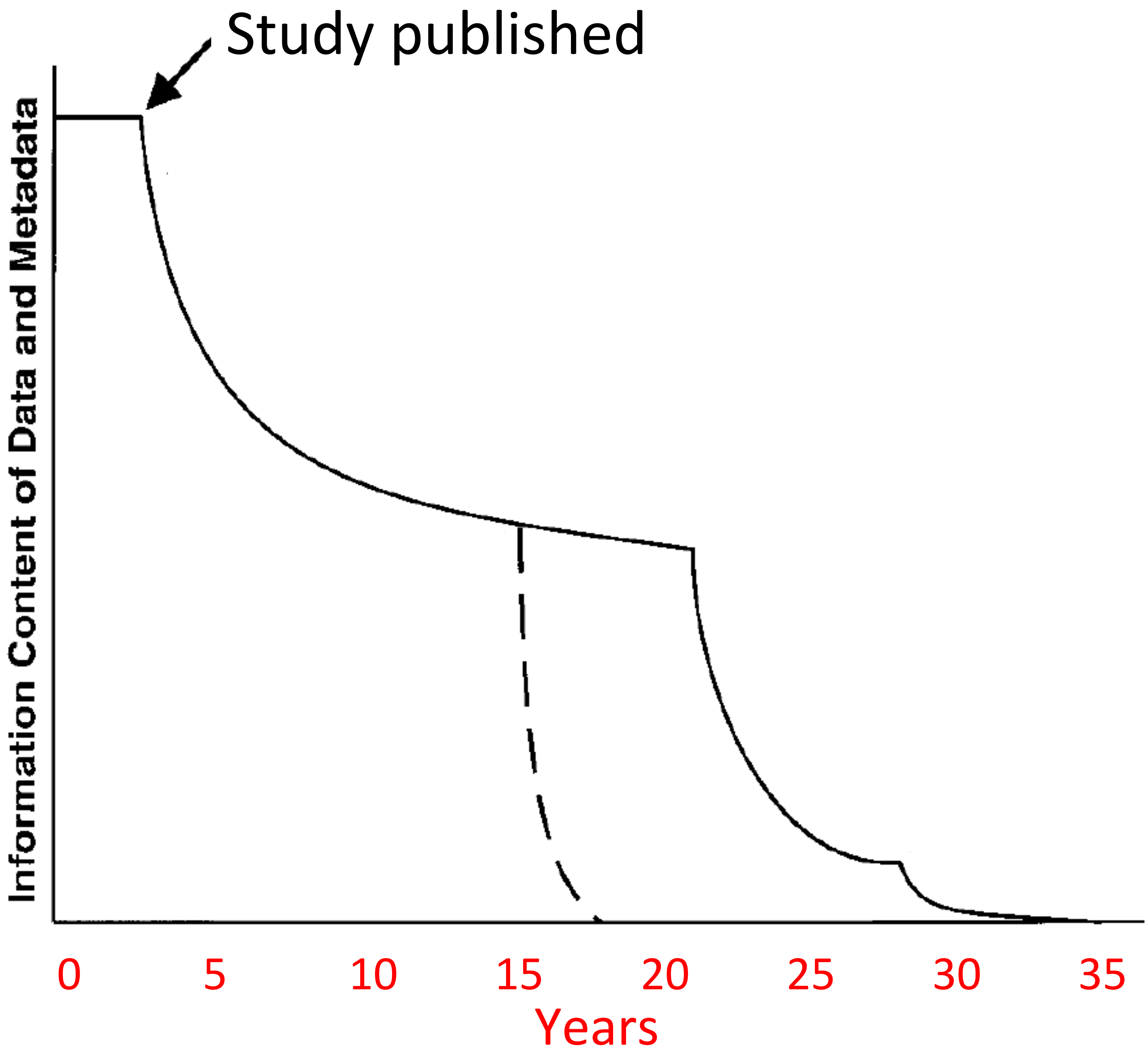
Vines *et al*. Current Biology 2014

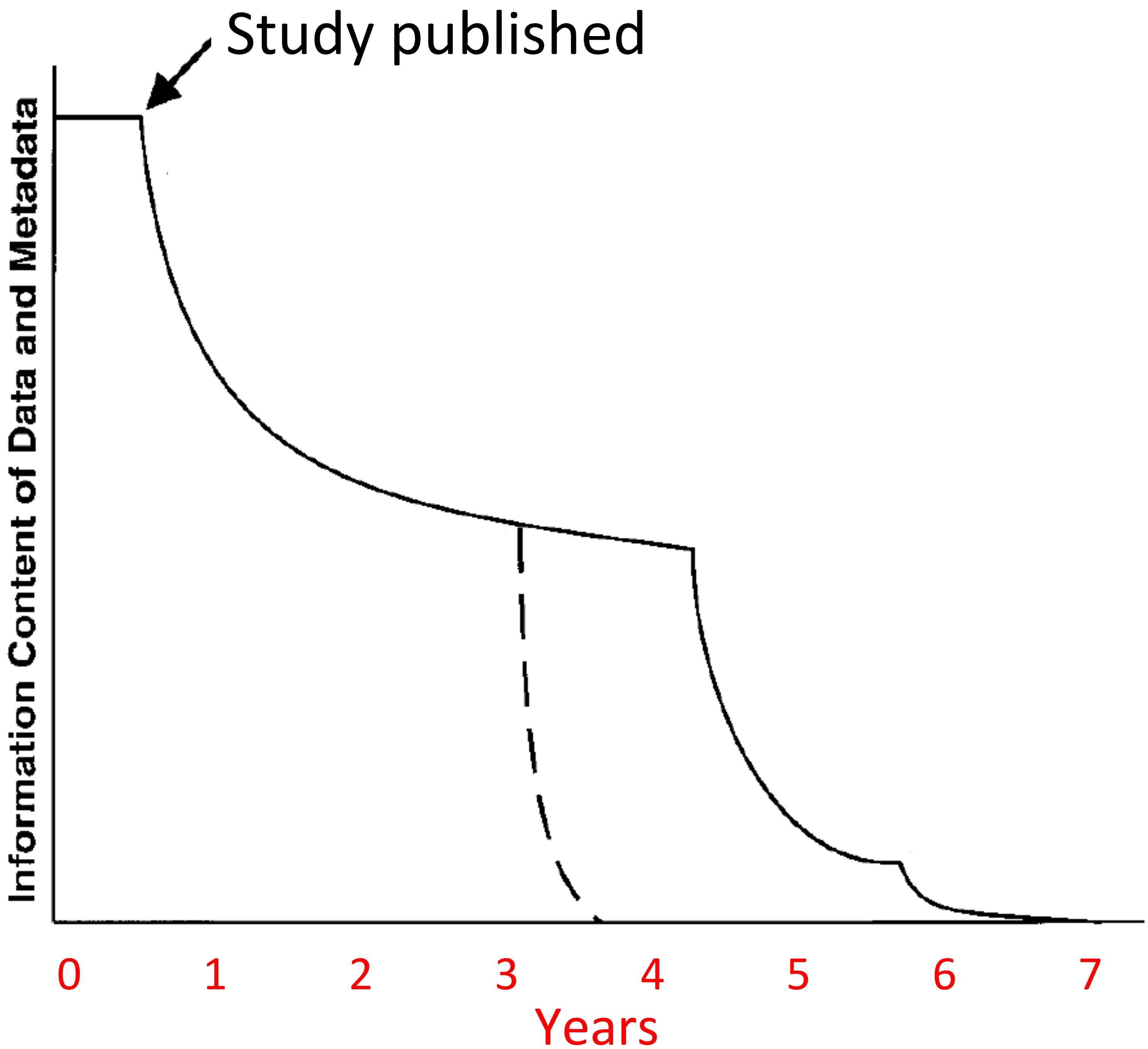
Introduction



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

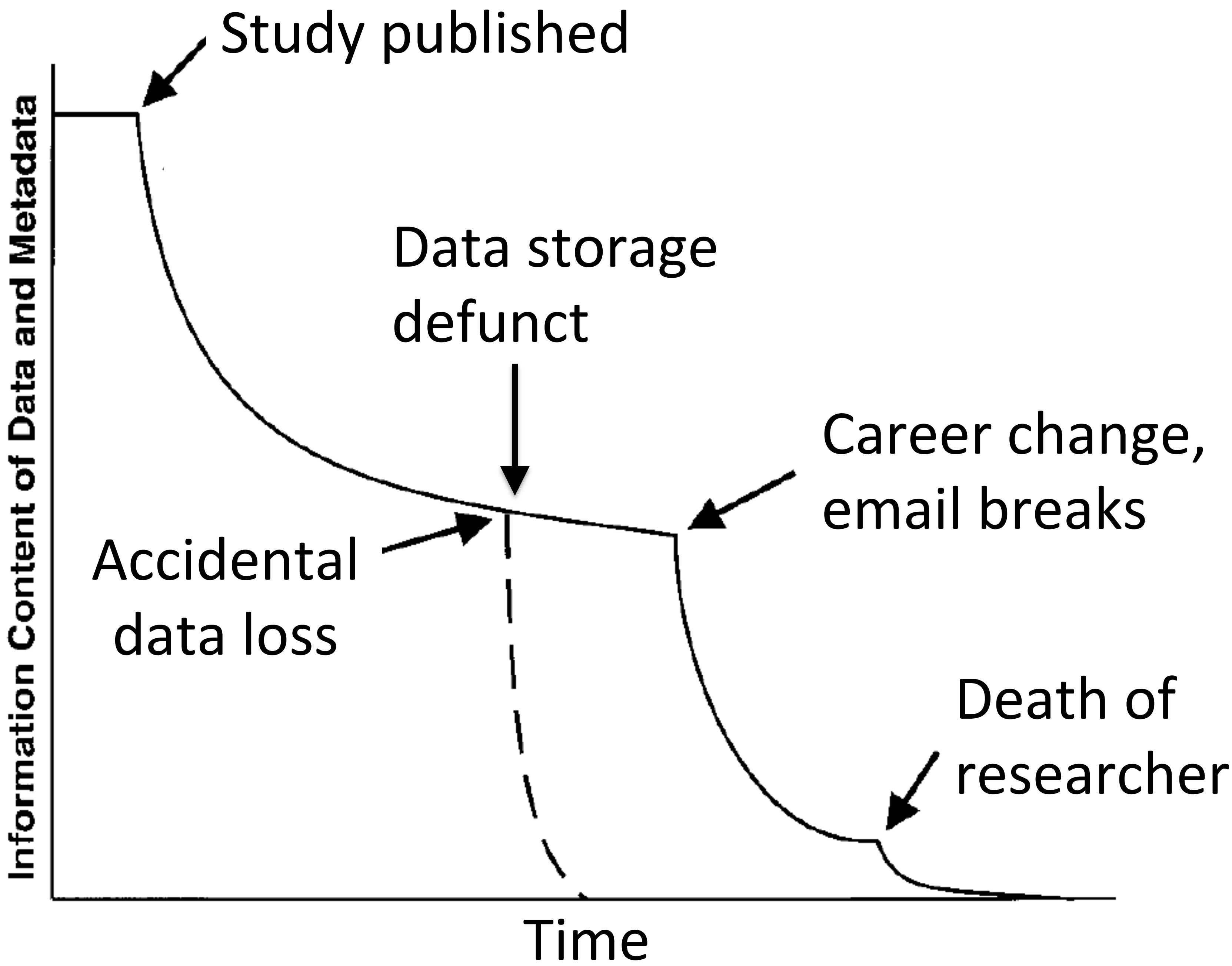
- How fast does this happen?


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



Study published
Information Content of Data and Metadata
0
1
2
3
4
5
6
7
Years

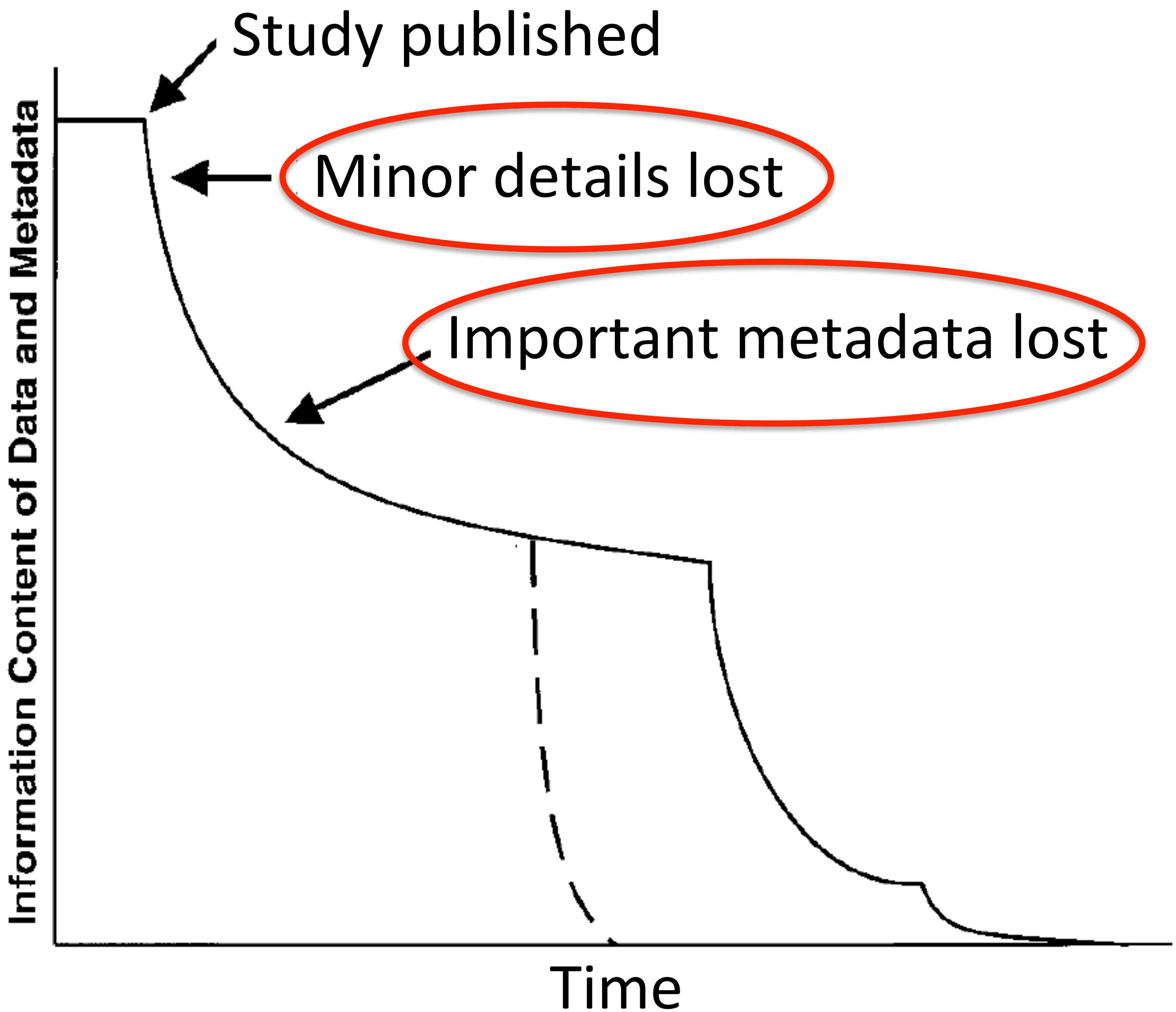
- How fast does this happen?
- What are the main causes of data loss?


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


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

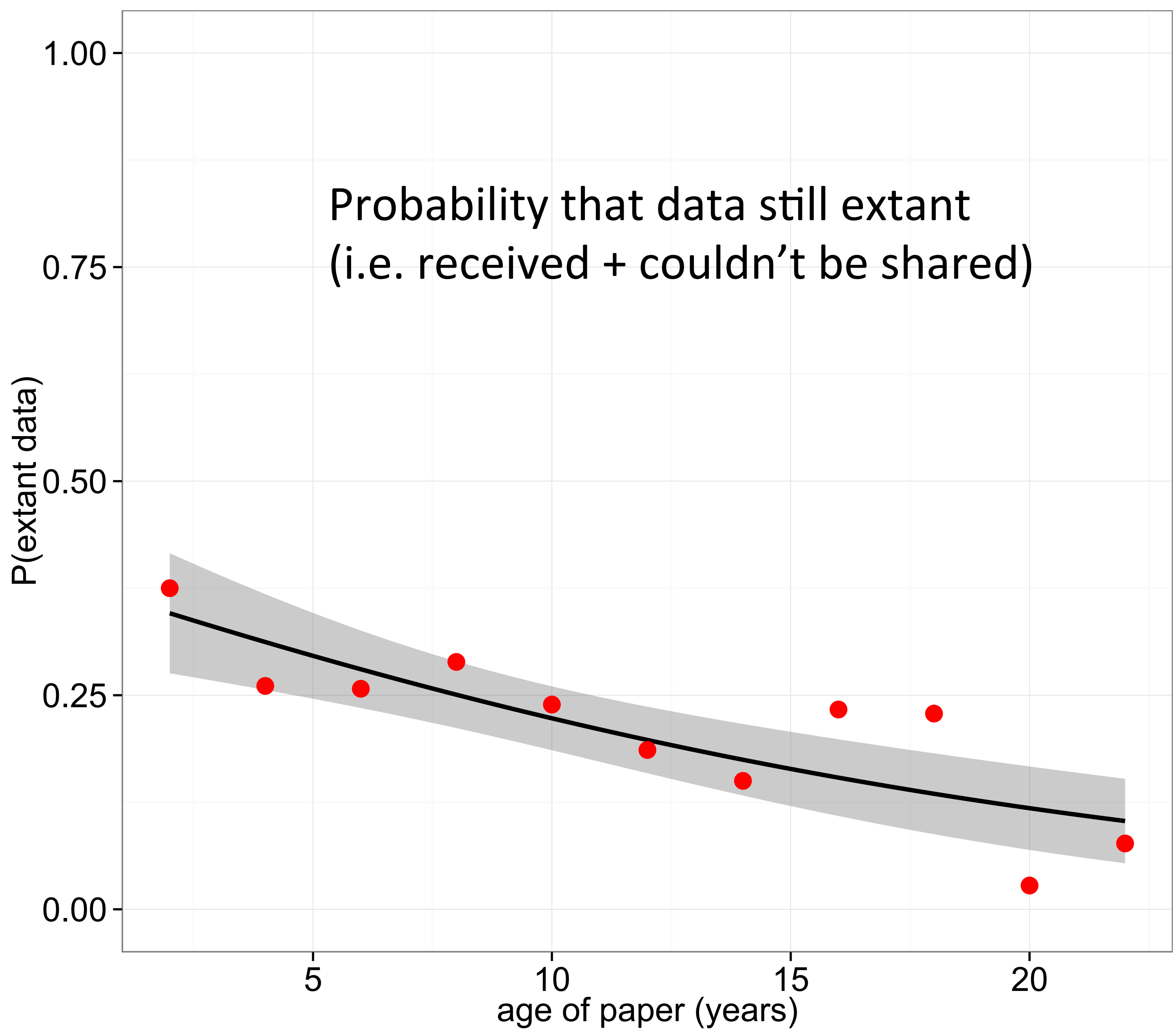
- Motivation sets total % of data we receive

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

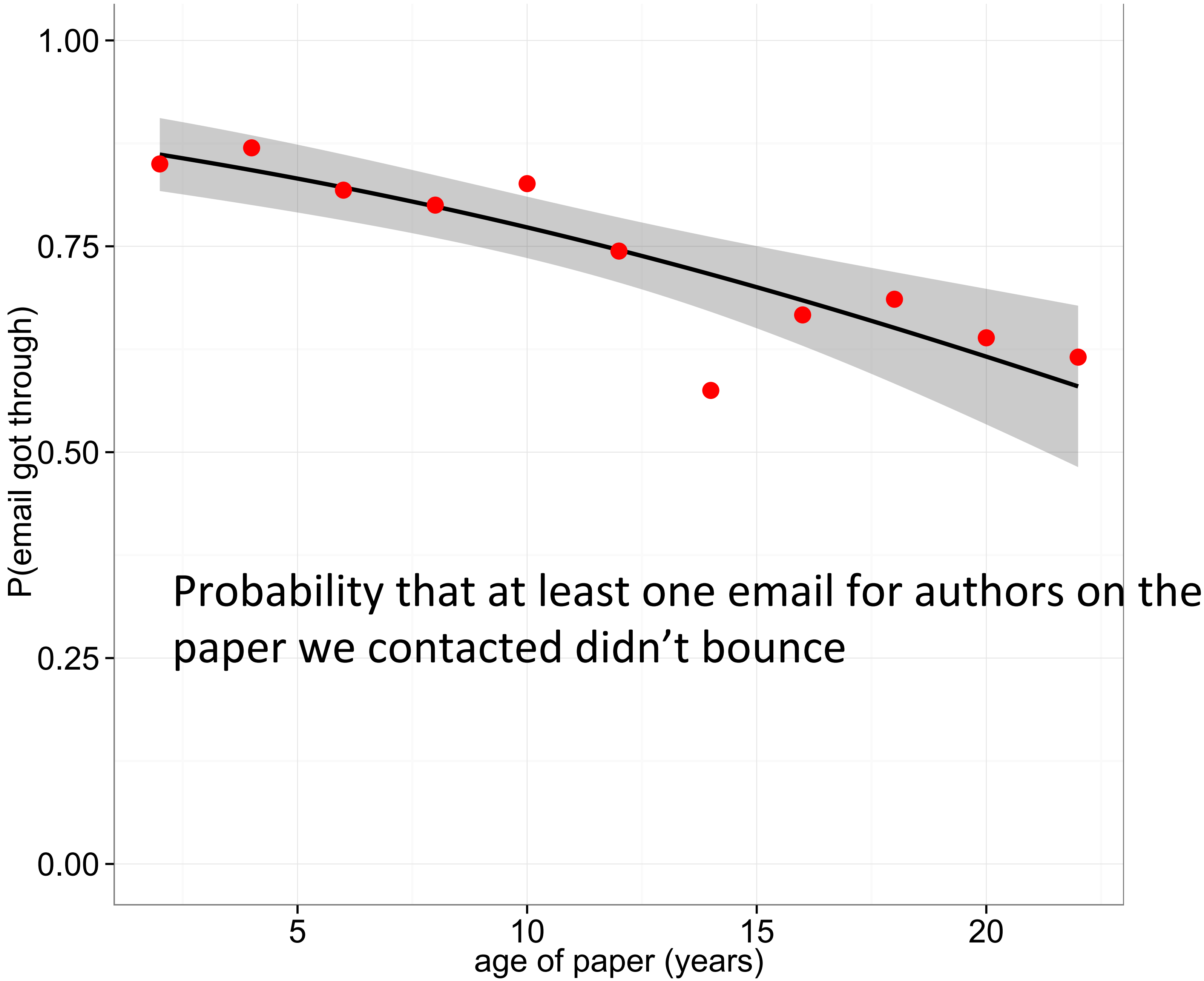
- If data were gone, we asked for the reason

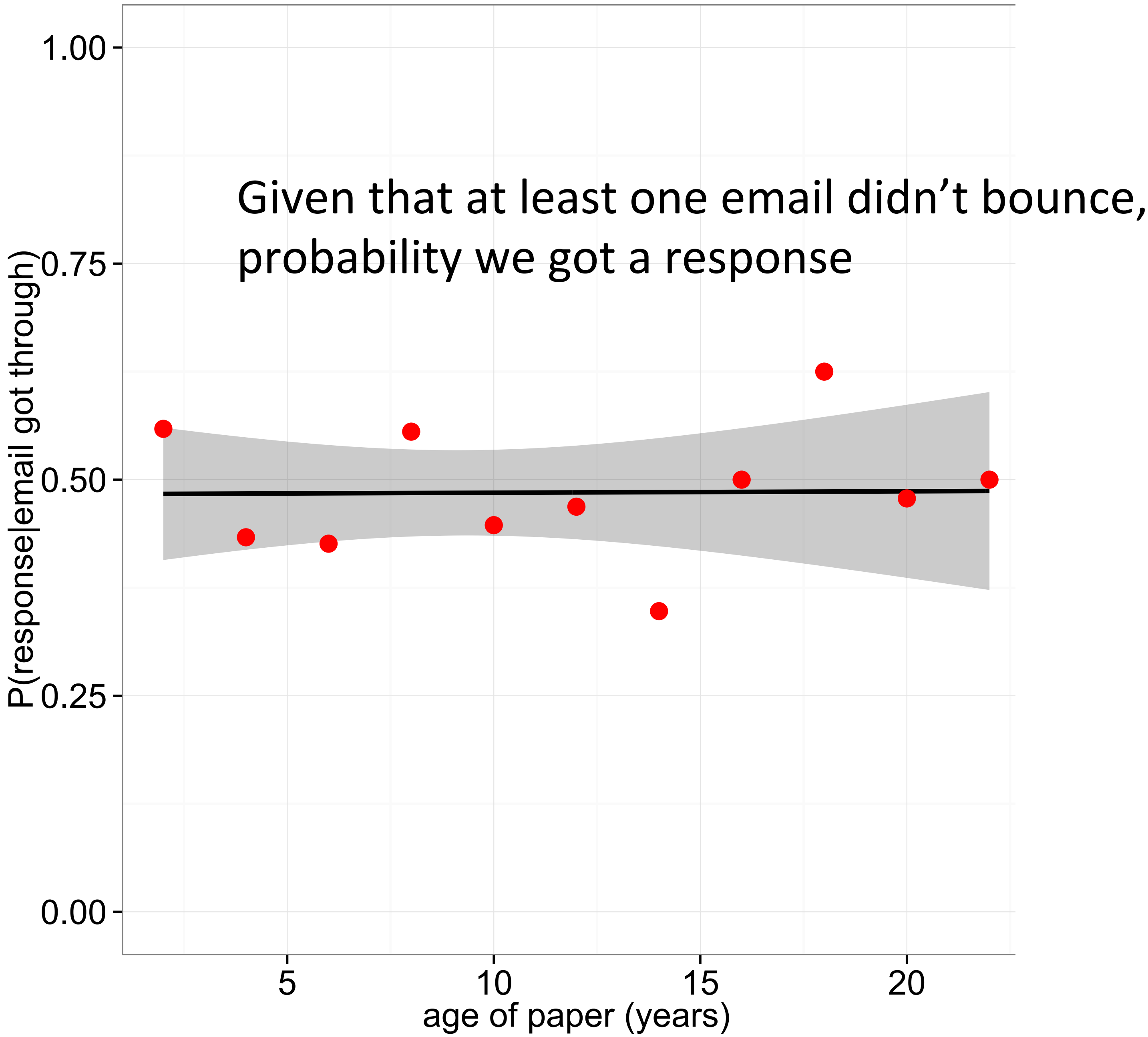
Results

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

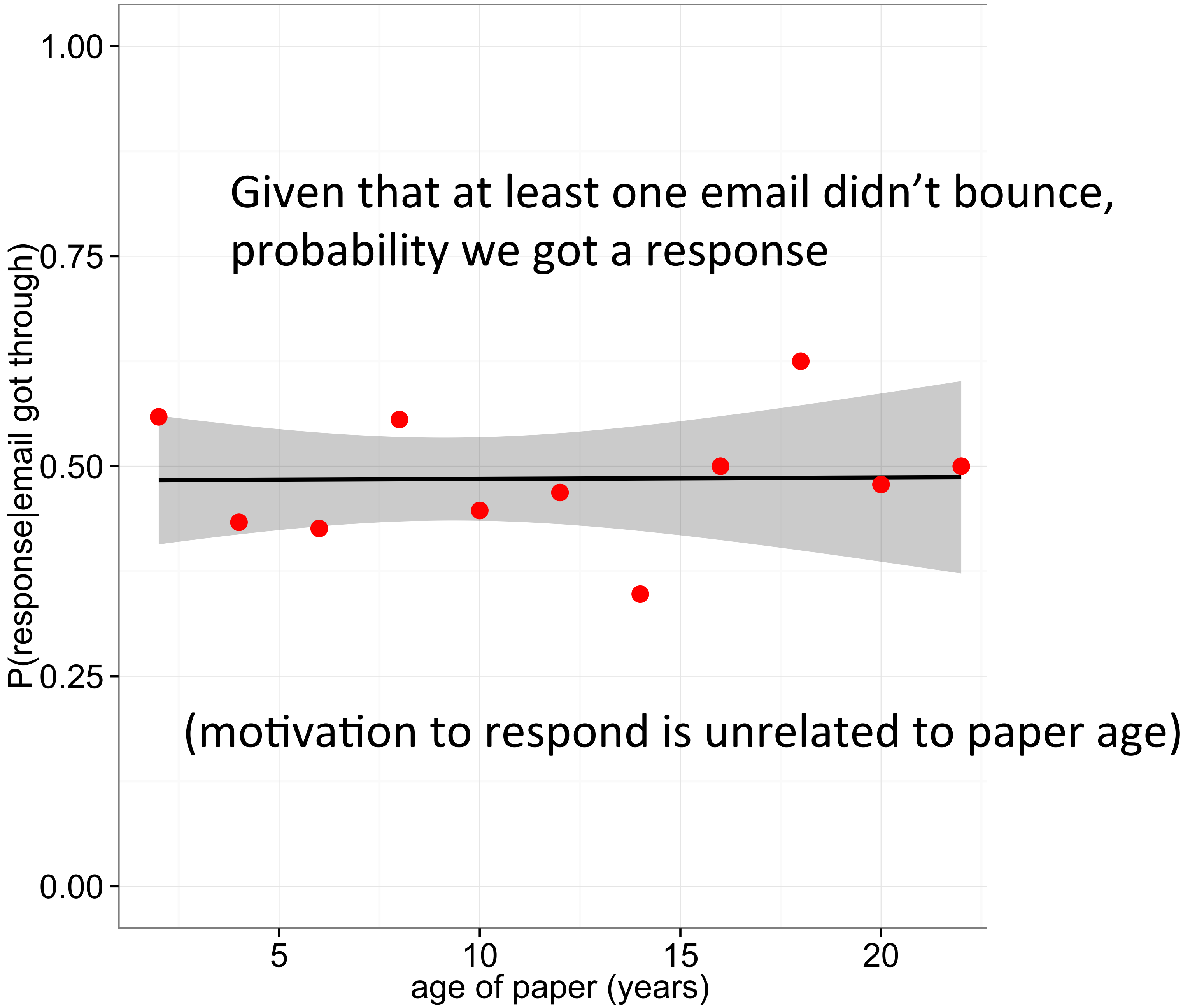

Probability that data still extant
(i.e. received + couldn't be shared)
P(extant data)
1.00
0.75
0.50
0.25
0.00
5
10
15
20
age of paper (years)

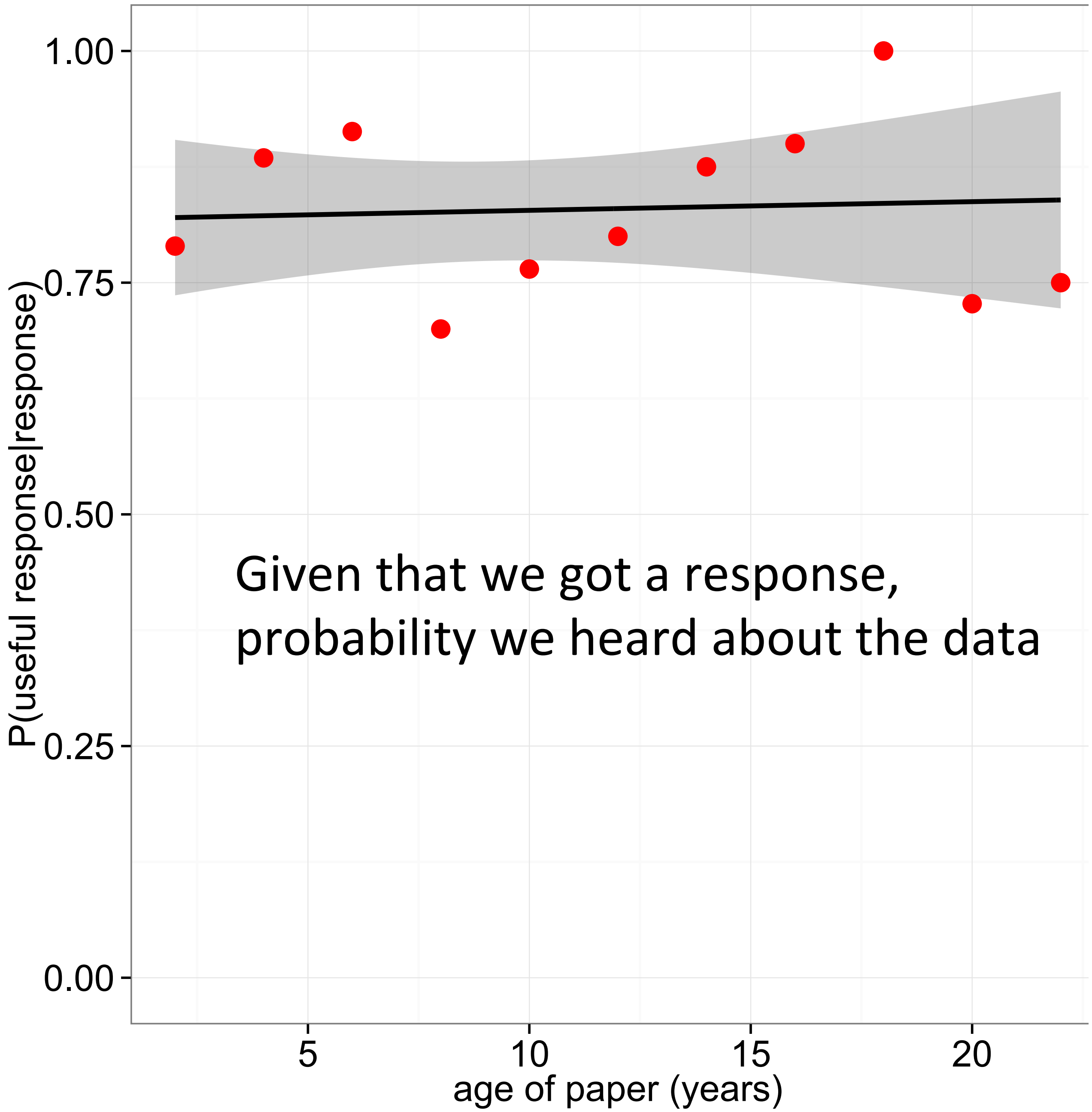
- Odds of data being extant fall by 8% per yr
- Almost all gone after 20 years
  - just 3 of 61 datasets extant for 1991 and 1993
- Why were we unable to get the data?
  - which reasons are related to paper age?

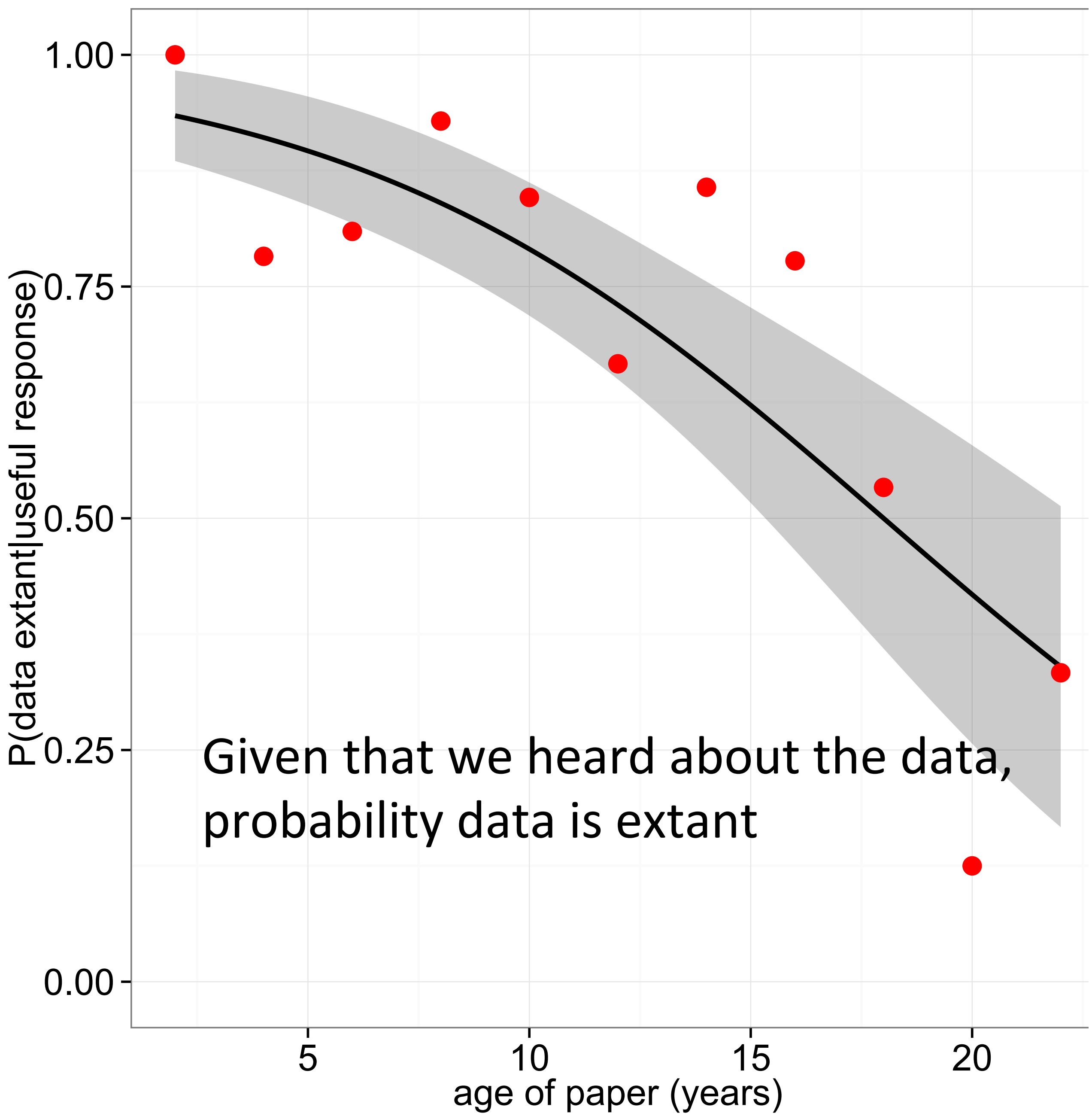

P(email got through)
1.00
0.75
0.50
0.25
0.00
5
10
15
20
age of paper (years)
Probability that at least one email for authors on the paper we contacted didn't bounce



Given that at least one email didn't bounce,
probability we got a response
1.00
0.75
0.50
0.25
0.00
P(response|email got through)
5
10
15
20
age of paper (years)


Given that at least one email didn't bounce, probability we got a response
(motivation to respond is unrelated to paper age)
P(response|email got through)
1.00
0.75
0.50
0.25
0.00
5
10
15
20
age of paper (years)


1.00
0.75
0.50
0.25
0.00
P(useful response|response)
5
10
15
20
age of paper (years)
Given that we got a response,
probability we heard about the data



1.00
0.75
0.50
0.25
0.00
P(data extant|useful response)
5
10
15
20
age of paper (years)
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

Reproducibility Part I: Discriminant Functions

Reproducibility Part I

- We received 101 files from authors
  - these are only the first step
- Are these the actual data from the paper?
- We tried to repeat their DFA

- What's a Discriminant Function Analysis?

  - you have 2 or more groups of something
  - you want be be able to tell the groups apart
  - the groups differ in e.g. size & shape
  - you measure a few things

  - the DF says what aspect of size/shape is best for distinguishing the groups

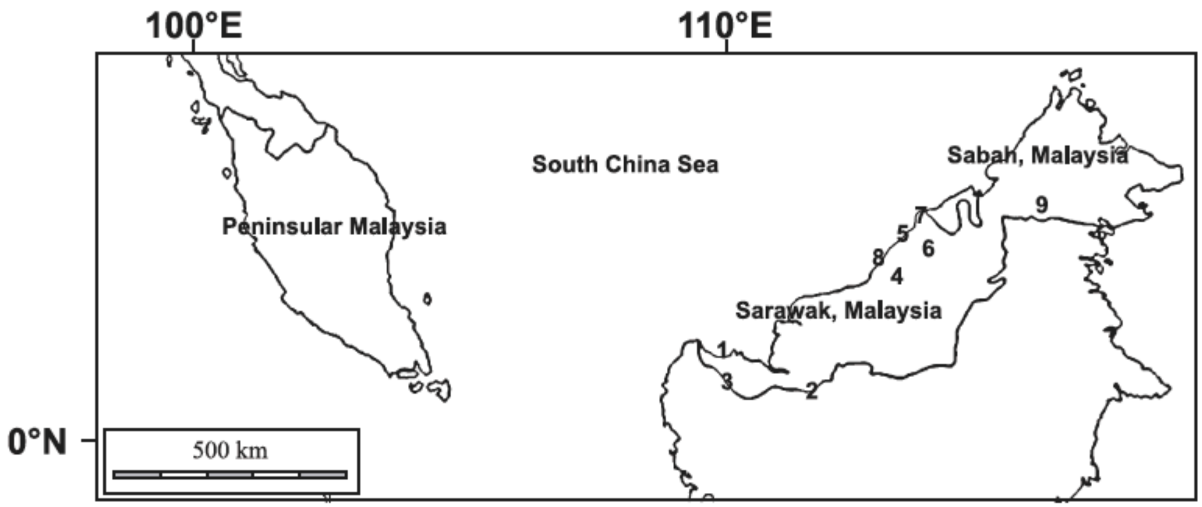


Fig. 1. Localities of the specimens used in this study. Localities were as follows: 1, Kubah National Park, Sarawak; 2, Batang Ai National Park, Sarawak; 3, Bau, Sarawak; 4, Bintulu, Sarawak; 5, Niah National Park, Sarawak; 6, Loagan Bunut National Park, Sarawak; 7, Lambir Hills National Park, Sarawak; 8, Similajau National Park, Sarawak; 9, Gua Madai, Sabah.

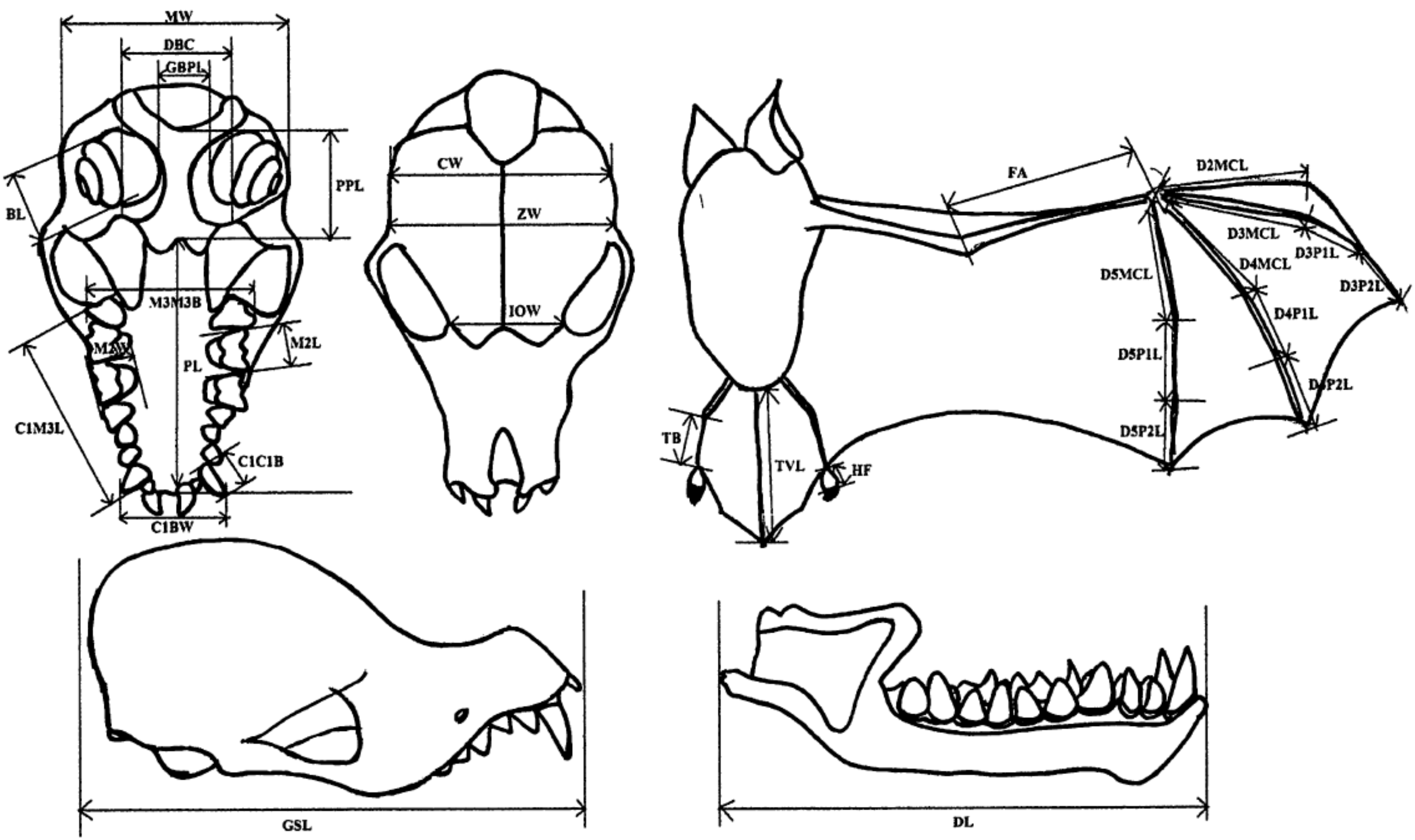


Fig. 2. Thirty-one characters used for measurements and morphological analysis in *Kerivoula.*

Table 6. Standardized canonical discriminant function coefficients of selected *Kerivoula*

	Functions				
	1	2	3	4	5
TB	–.612	–1.716*	–.294	.106	–.305
HF	.720	.937	.343	–.066	.700*
TVL	.676	–.155	.744	–.156	–.085
D3P1L	–.501	–.156	–.297	1.159*	–.409
D3P2L	1.201	.371	.848*	–.483	.645
D4MCL	1.202*	1.227	–.210	–.148	–.088
D4P1L	.184	–1.124	.599	.010	–.146
D4P2L	.222	1.184	–.080	–.257	–.052
CW	.387	–.986	–.543	–.250	.420
BL	.908	1.263	.101	.283	.588
DL	.963	.814	–.173	.641	.088
M2W	.905	–.636	.159	–.726	–.654

Highest character loadings for each function are indicated with asterisks (*)

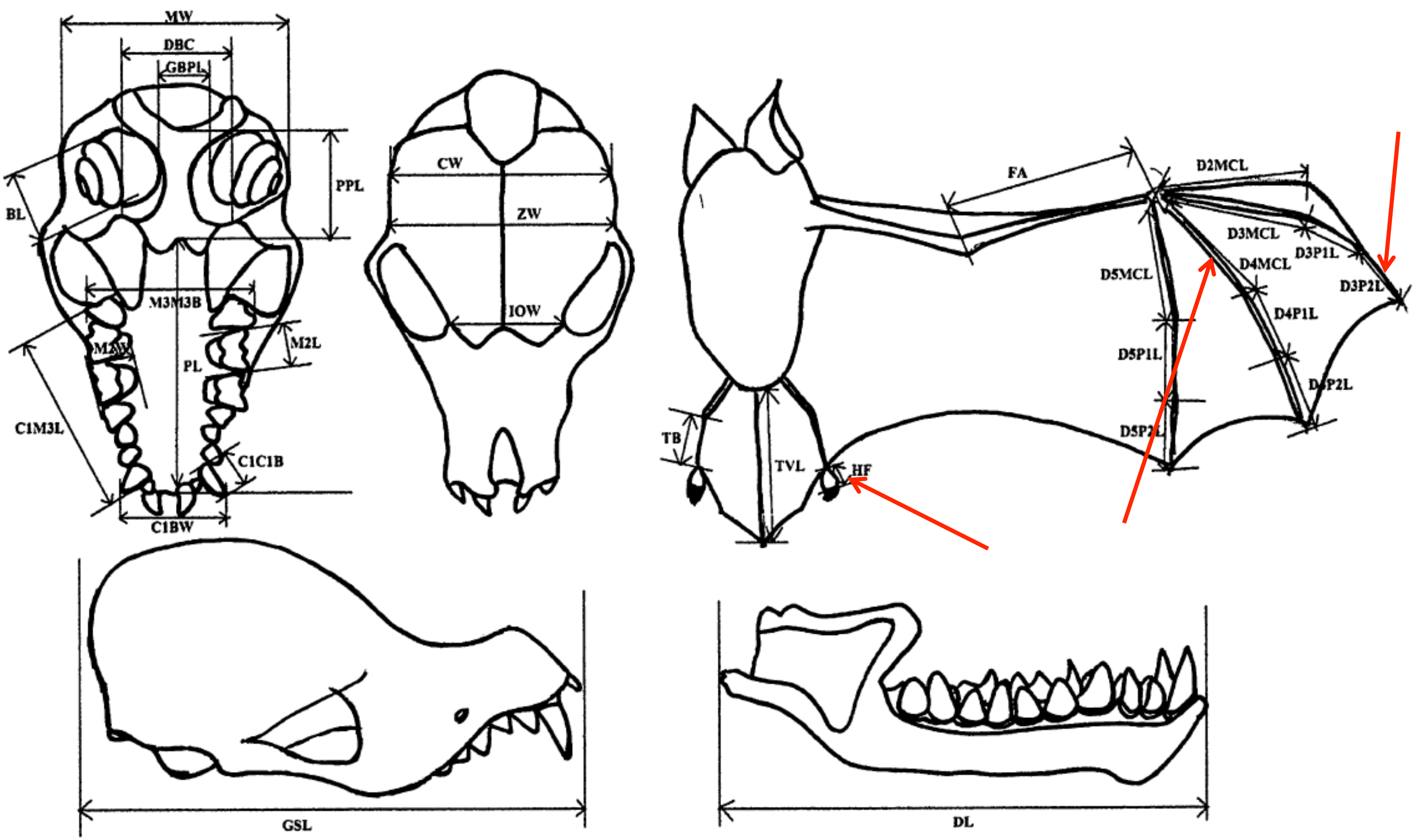
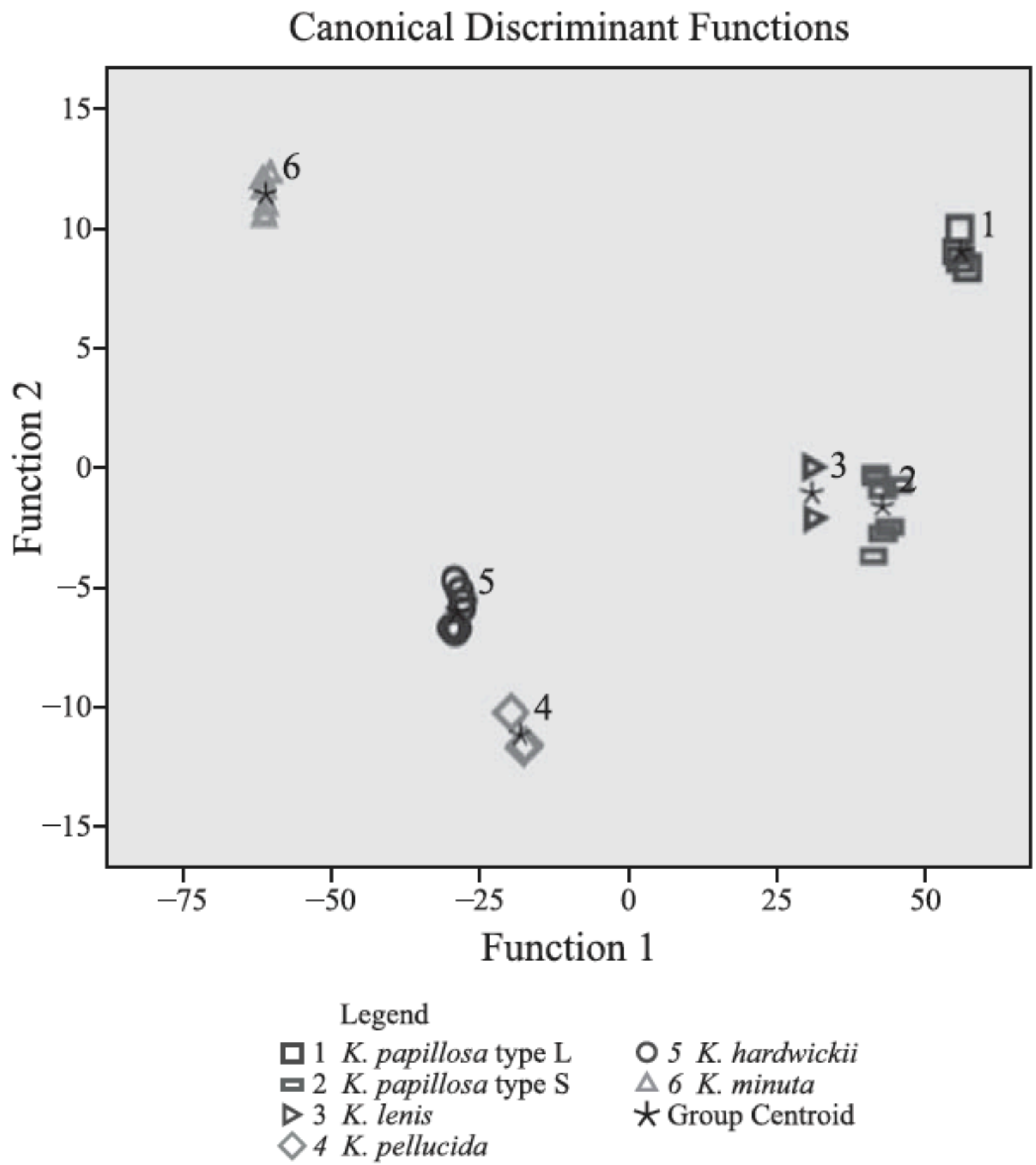


Fig. 2. Thirty-one characters used for measurements and morphological analysis in *Kerivoula*.

Table 6. Standardized canonical discriminant function coefficients of selected *Kerivoula*

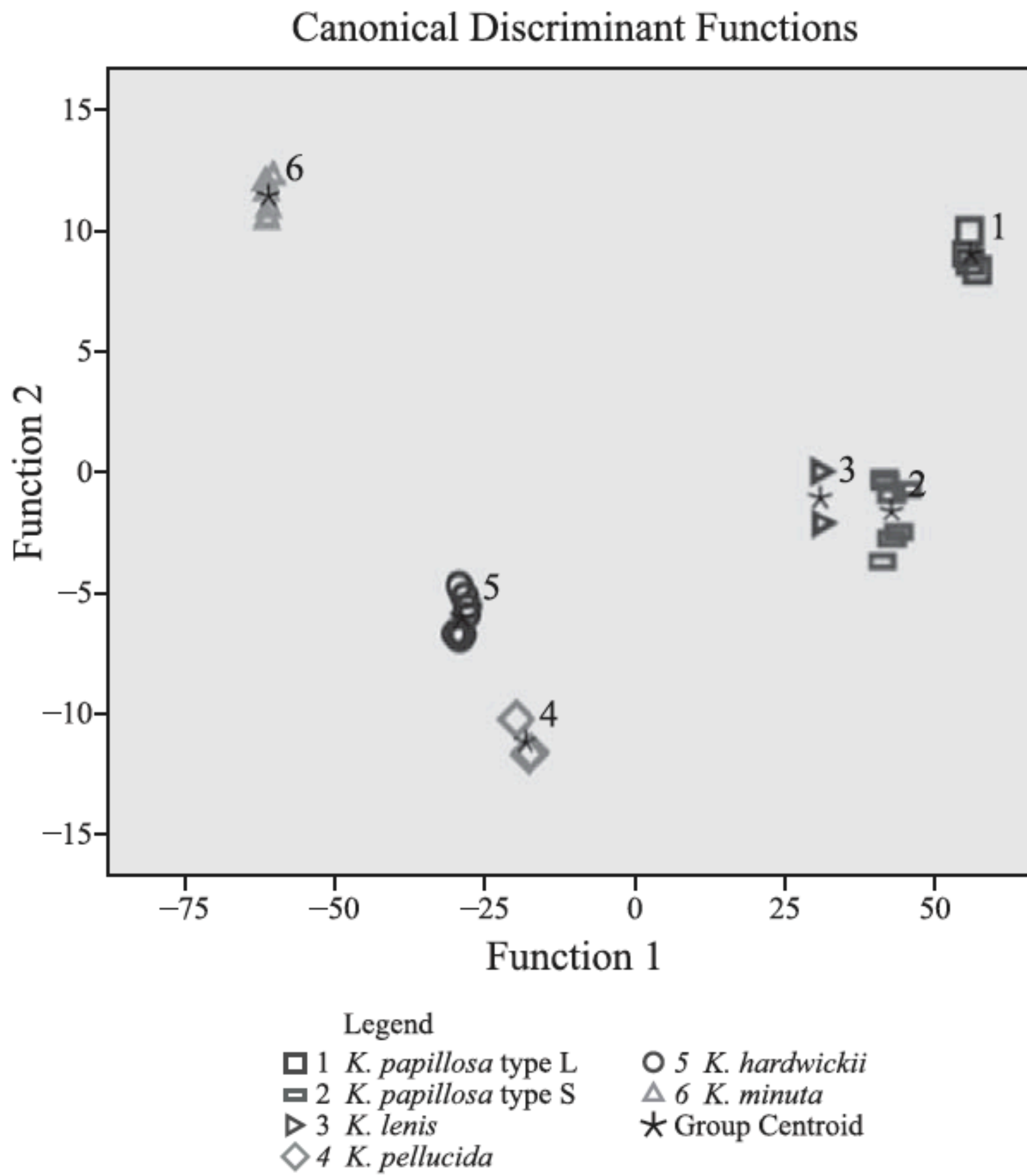
	Functions				
	1	2	3	4	5
TB	–.612	–1.716*	–.294	.106	–.305
HF	.720	.937	.343	–.066	.700*
TVL	.676	–.155	.744	–.156	–.085
D3P1L	–.501	–.156	–.297	1.159*	–.409
D3P2L	1.201	.371	.848*	–.483	.645
D4MCL	1.202*	1.227	–.210	–.148	–.088
D4P1L	.184	–1.124	.599	.010	–.146
D4P2L	.222	1.184	–.080	–.257	–.052
CW	.387	–.986	–.543	–.250	.420
BL	.908	1.263	.101	.283	.588
DL	.963	.814	–.173	.641	.088
M2W	.905	–.636	.159	–.726	–.654

Highest character loadings for each function are indicated with asterisks (*)



Canonical Discriminant Functions
Function 2
Function 1
15
10
5
0
−5
−10
−15
−75
−50
−25
0
25
50
1
2
3
4
5
6
Legend
1 K. papillosa type L
2 K. papillosa type S
3 K. lenis
4 K. pellucida
5 K. hardwickii
6 K. minuta
Group Centroid

- What's a Discriminant Function Analysis?
  - the DFA produces three useful metrics:
    1. the percent variance explained by the 1st axis



Canonical Discriminant Functions
Function 2
15
10
5
0
−5
−10
−15
Function 1
−75
−50
−25
0
25
50
6
1
3
2
5
4
Legend
1 K. papillosa type L
2 K. papillosa type S
3 K. lenis
4 K. pellucida
5 K. hardwickii
6 K. minuta
Group Centroid

- What's a Discriminant Function Analysis?

  – the DFA produces three useful numbers:

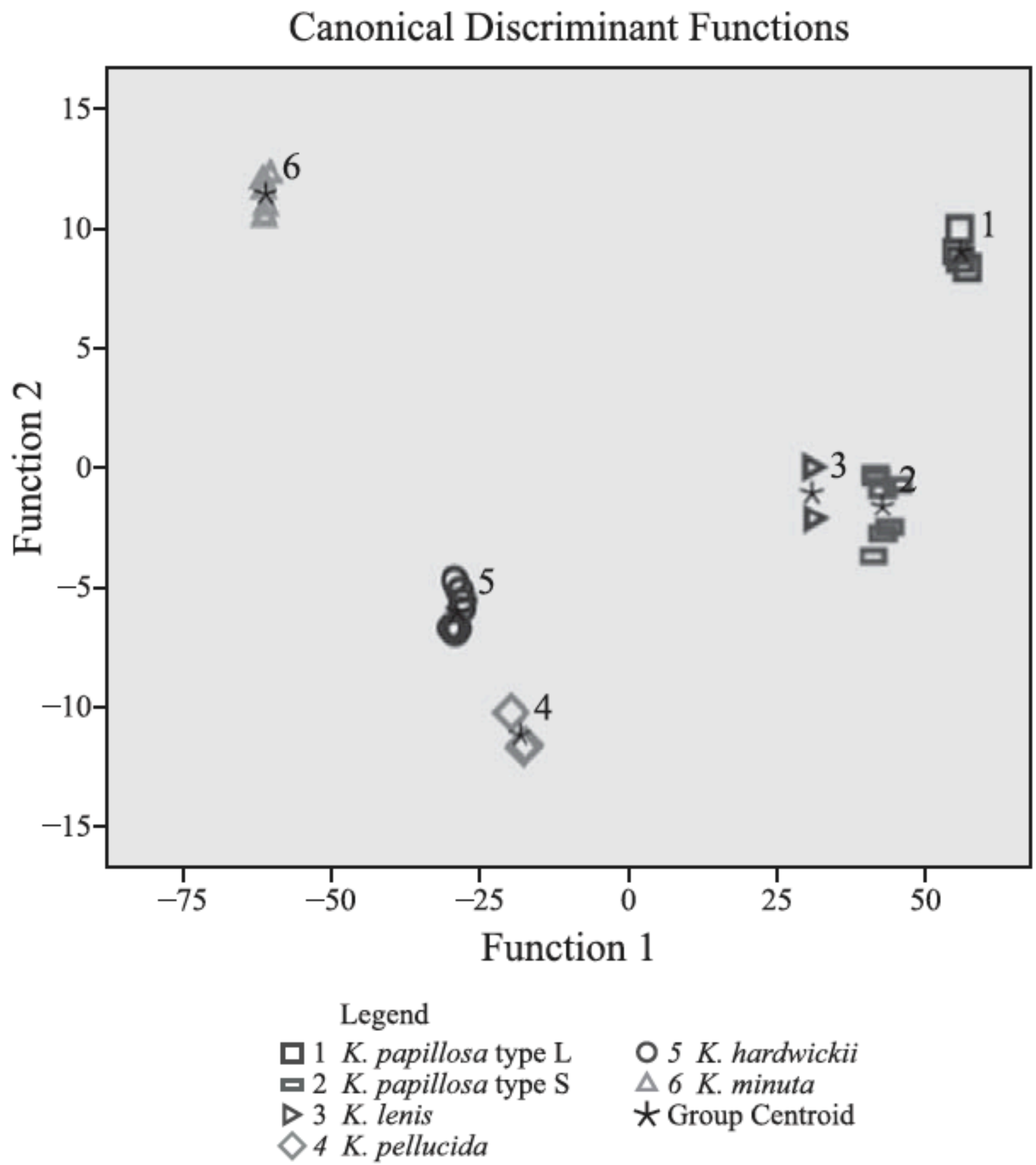
    1. the percent variance explained by the 1st axis
    2. the loading coefficient

Table 6. Standardized canonical discriminant function coefficients of selected *Kerivoula*

	Functions				
	1	2	3	4	5
TB	–.612	–1.716*	–.294	.106	–.305
HF	.720	.937	.343	–.066	.700*
TVL	.676	–.155	.744	–.156	–.085
D3P1L	–.501	–.156	–.297	1.159*	–.409
D3P2L	1.201	.371	.848*	–.483	.645
D4MCL	1.202*	1.227	–.210	–.148	–.088
D4P1L	.184	–1.124	.599	.010	–.146
D4P2L	.222	1.184	–.080	–.257	–.052
CW	.387	–.986	–.543	–.250	.420
BL	.908	1.263	.101	.283	.588
DL	.963	.814	–.173	.641	.088
M2W	.905	–.636	.159	–.726	–.654

Highest character loadings for each function are indicated with asterisks (*)

- What's a Discriminant Function Analysis?
  - the DFA produces three useful numbers:
    1. the percent variance explained by the 1st axis
    2. the loading coefficient
    3. the percentage of individuals correctly assigned



Canonical Discriminant Functions
Function 2
Function 1
15
10
5
0
–5
–10
–15
–75
–50
–25
0
25
50
1
2
3
4
5
6
Legend
1 K. papillosa type L
2 K. papillosa type S
3 K. lenis
4 K. pellucida
5 K. hardwickii
6 K. minuta
Group Centroid

- What's a Discriminant Function Analysis?

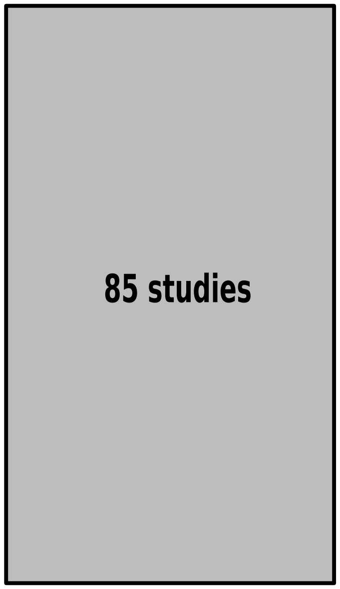
  - the DFA produces three useful numbers:
    1. the percent variance explained by the 1^{st} axis
    2. the loading coefficient
    3. the percentage of individuals correctly assigned

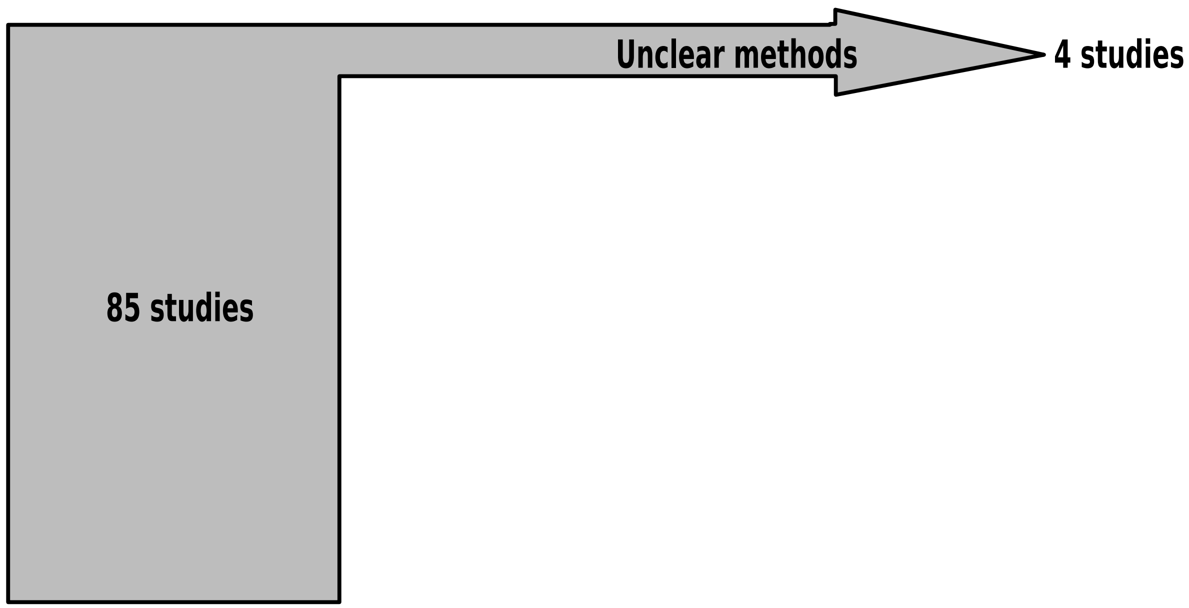
- We tried to reproduce these metrics

Reproducibility Part I

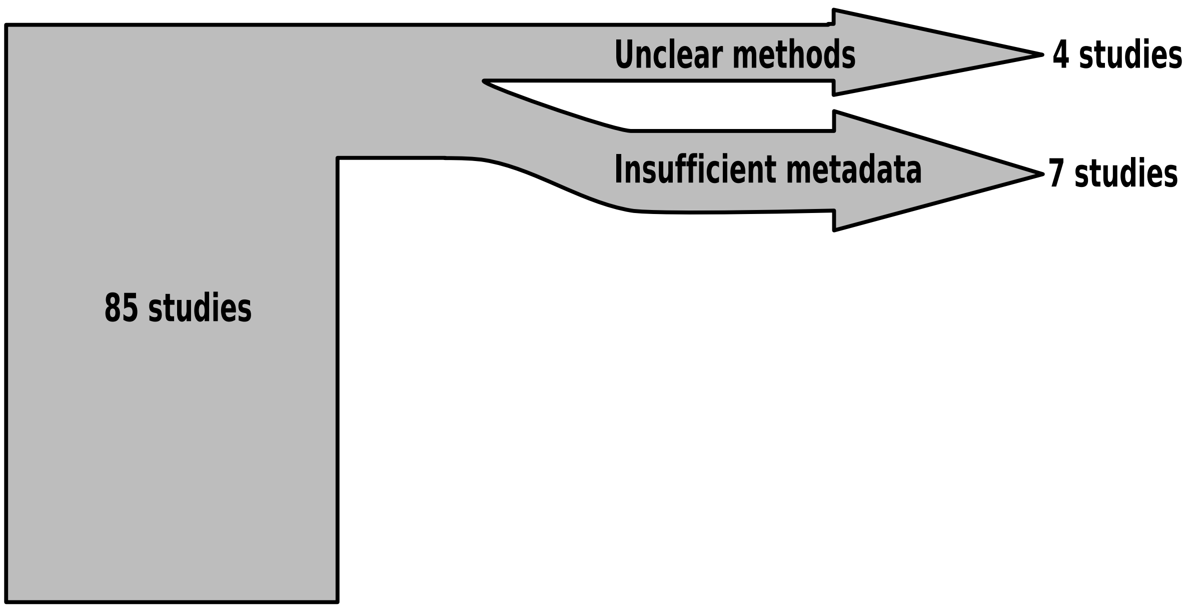
- We started with 101 studies
  - 16 didn't contain any of our three metrics
  - these were excluded

- What happened with the rest?

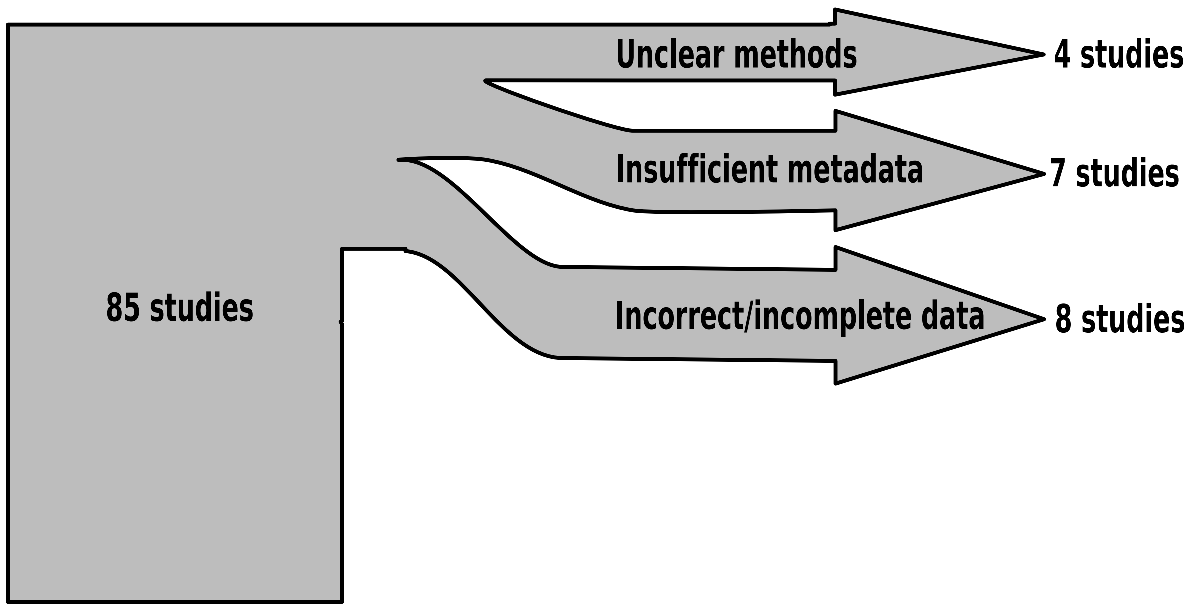

85 studies



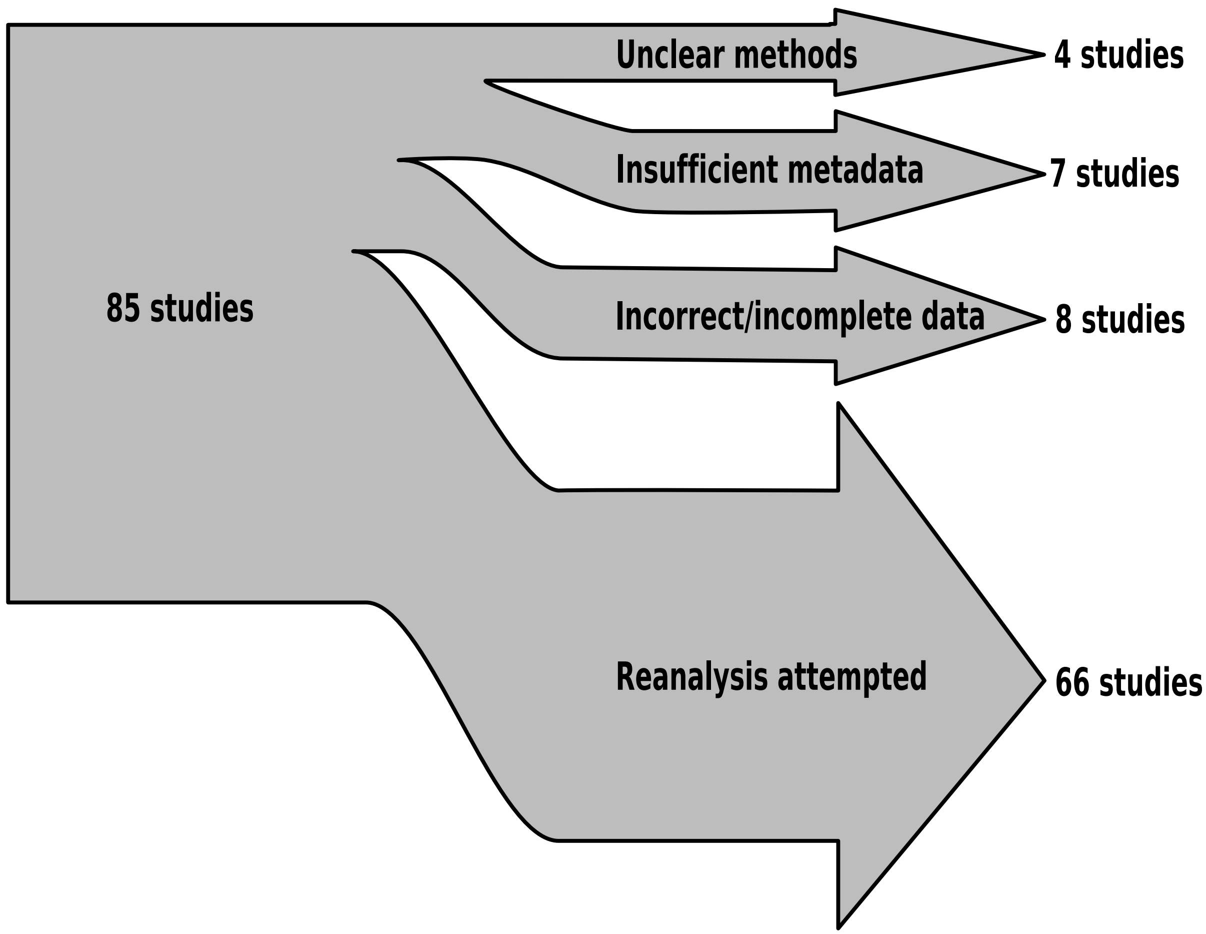
85 studies
Unclear methods
4 studies

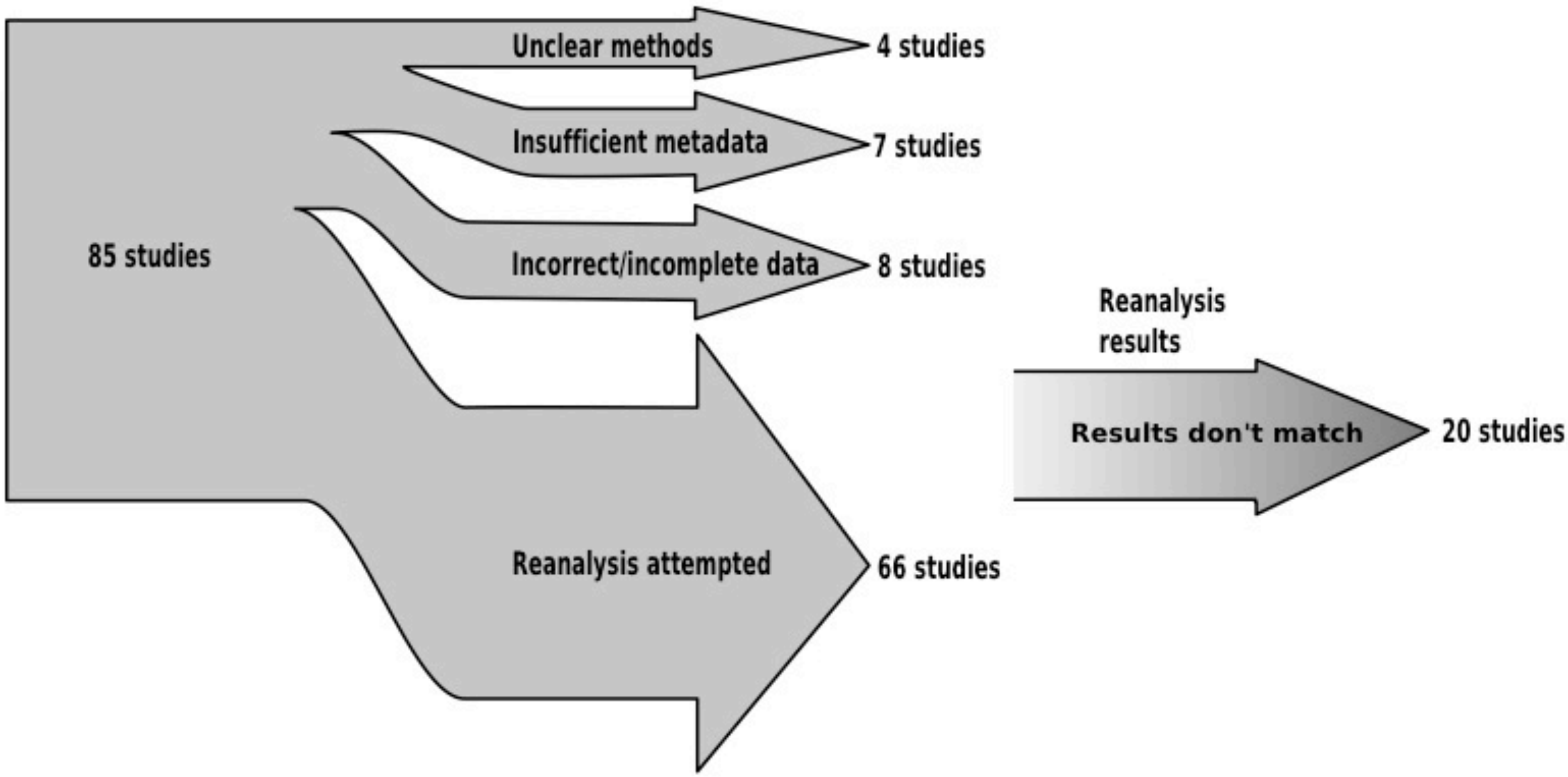


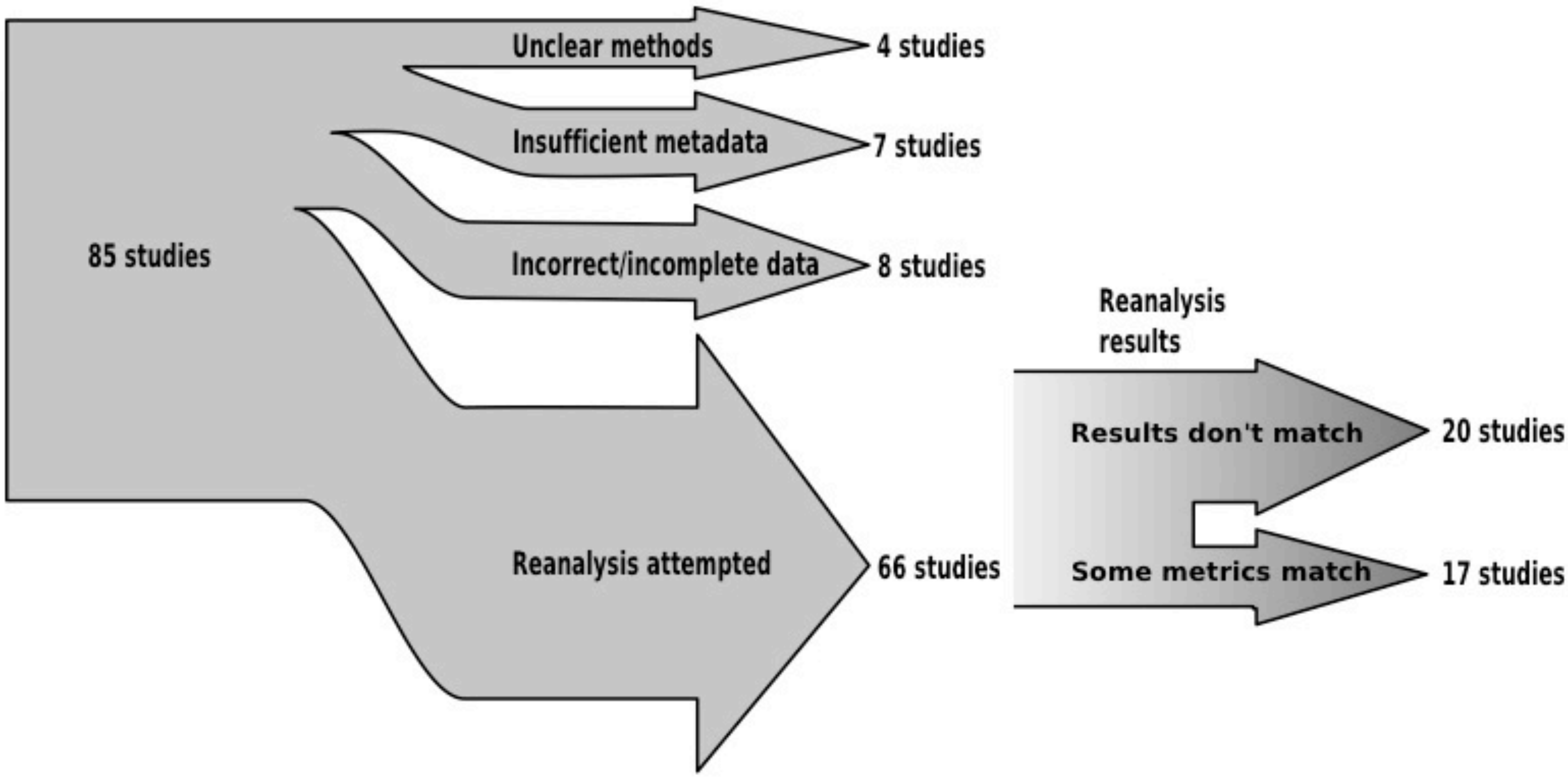
Unclear methods
4 studies
Insufficient metadata
7 studies
85 studies



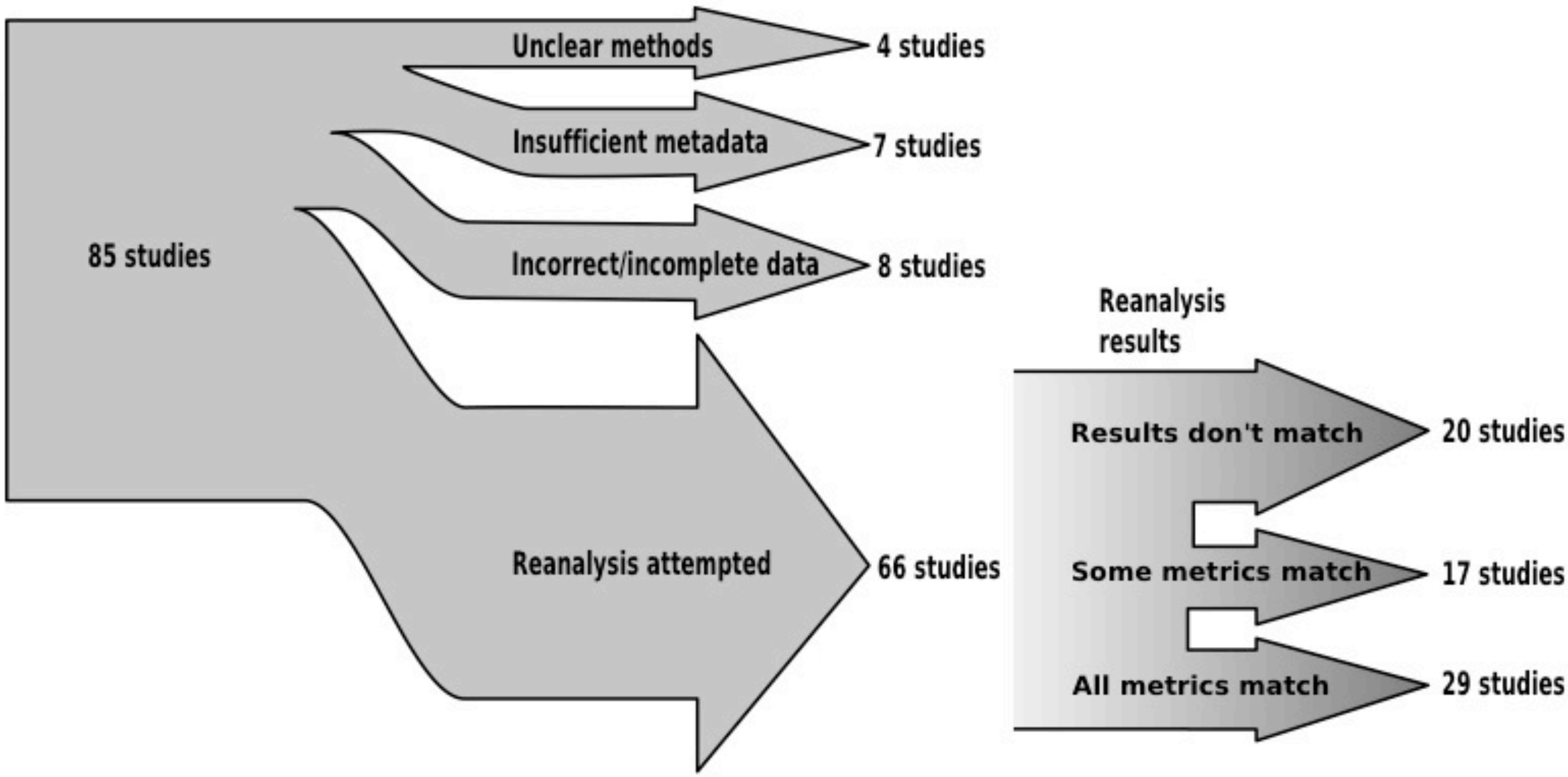
Unclear methods
4 studies
Insufficient metadata
7 studies
85 studies
Incorrect/incomplete data
8 studies


Unclear methods
4 studies
Insufficient metadata
7 studies
85 studies
Incorrect/incomplete data
8 studies
Reanalysis attempted
66 studies


Unclear methods
4 studies
Insufficient metadata
7 studies
85 studies
Incorrect/incomplete data
8 studies
Reanalysis
results
Results don't match
20 studies
Reanalysis attempted
66 studies



Unclear methods
4 studies
Insufficient metadata
7 studies
Incorrect/incomplete data
8 studies
85 studies
Reanalysis results
Results don't match
20 studies
Some metrics match
17 studies
Reanalysis attempted
66 studies


85 studies
Unclear methods
4 studies
Insufficient metadata
7 studies
Incorrect/incomplete data
8 studies
Reanalysis attempted
66 studies
Reanalysis results
Results don't match
20 studies
Some metrics match
17 studies
All metrics match
29 studies

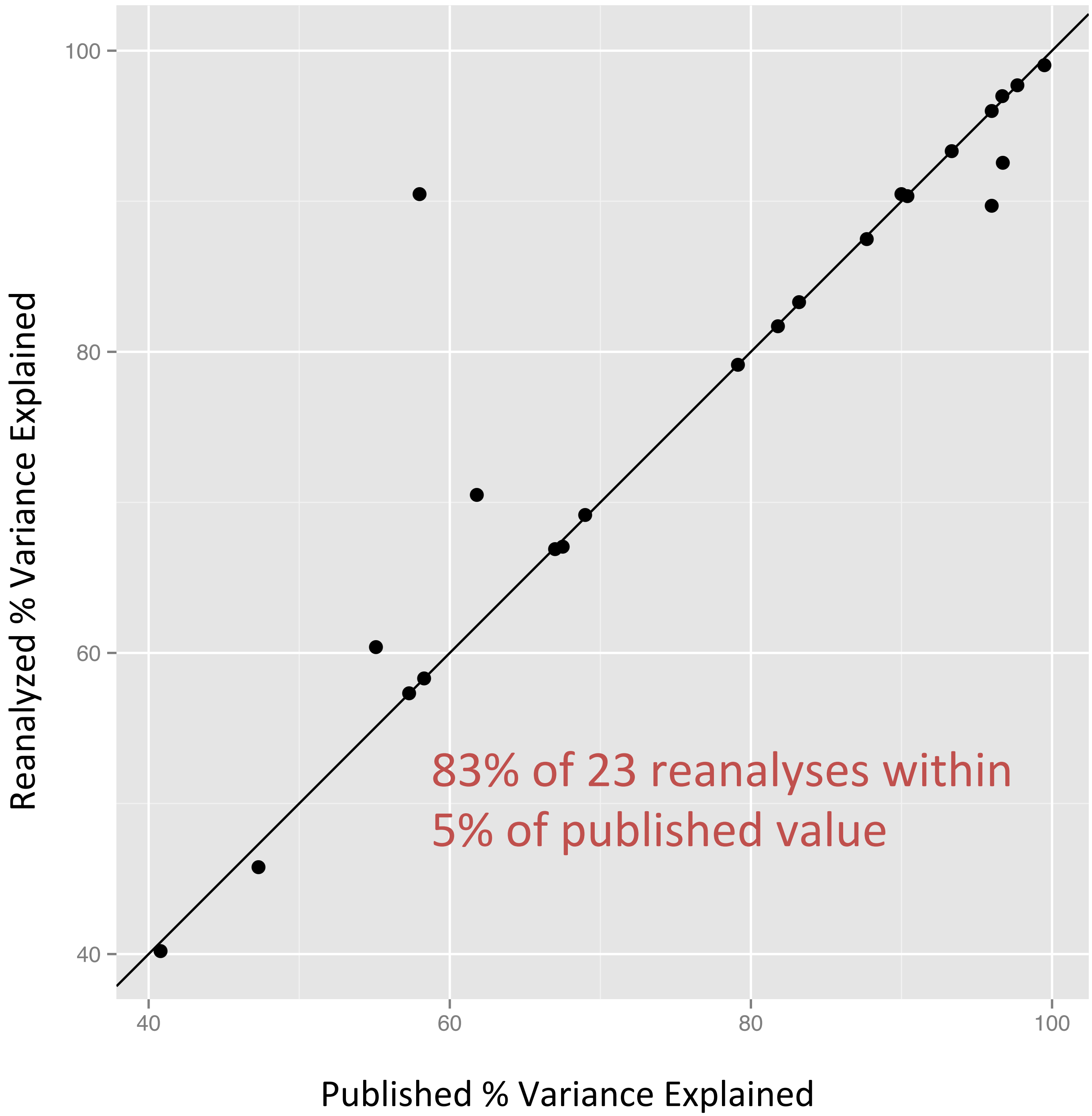
Outcome	**Percent**
Unclear methods	4
Insufficient metadata	7
Incorrect/incomplete data	9
[Subtotal]	[20]
Reanalysis attempted:	
Results don't match	21
Some metrics match	18
All metrics match	31
Overall Total	100

Reproducibility Part I

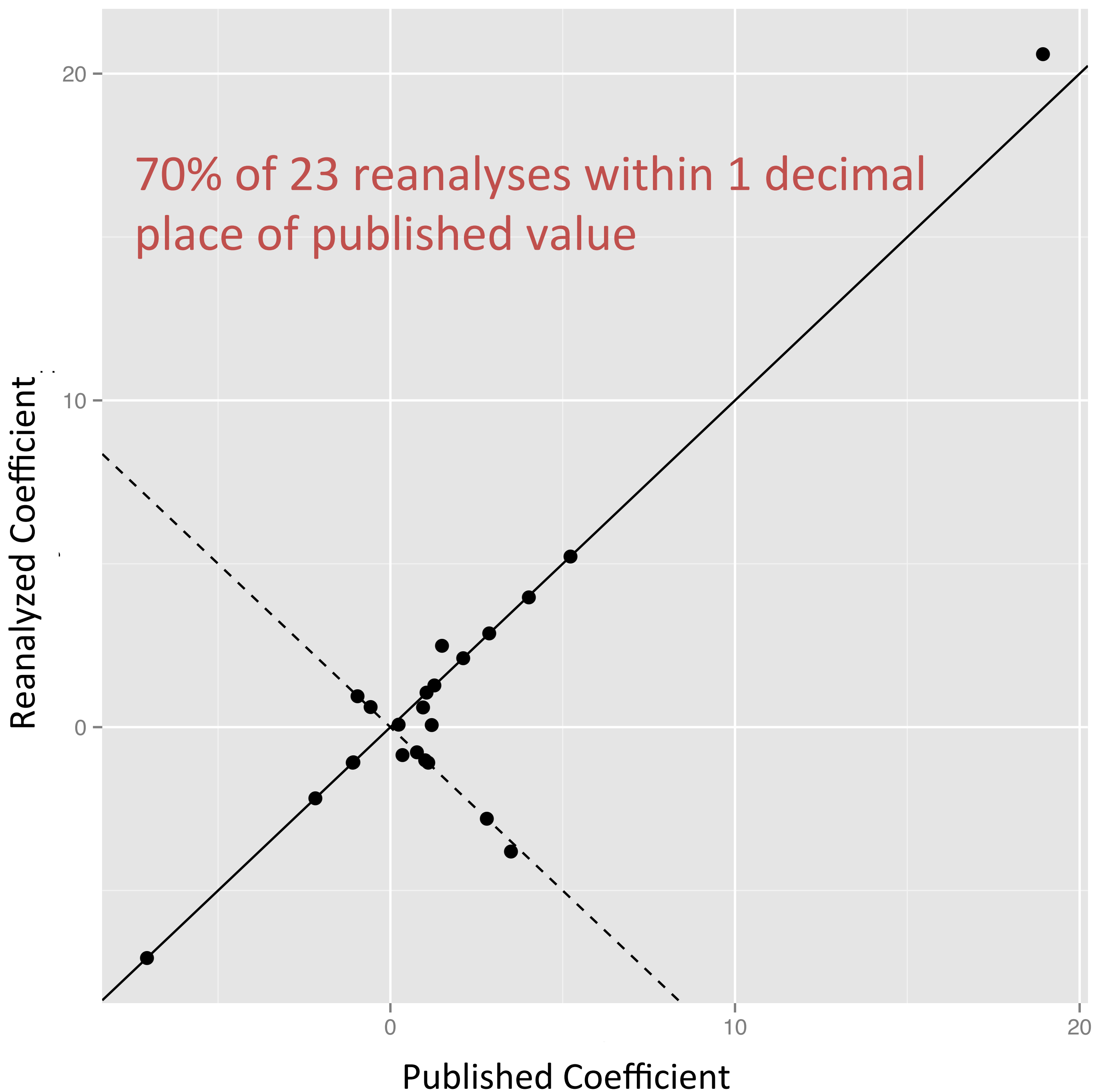
- We started with 101 studies
  - 16 didn't contain any of our three metrics
  - these were excluded
- Only 52 could be reproduced
  - 10% of the 516 datasets requested

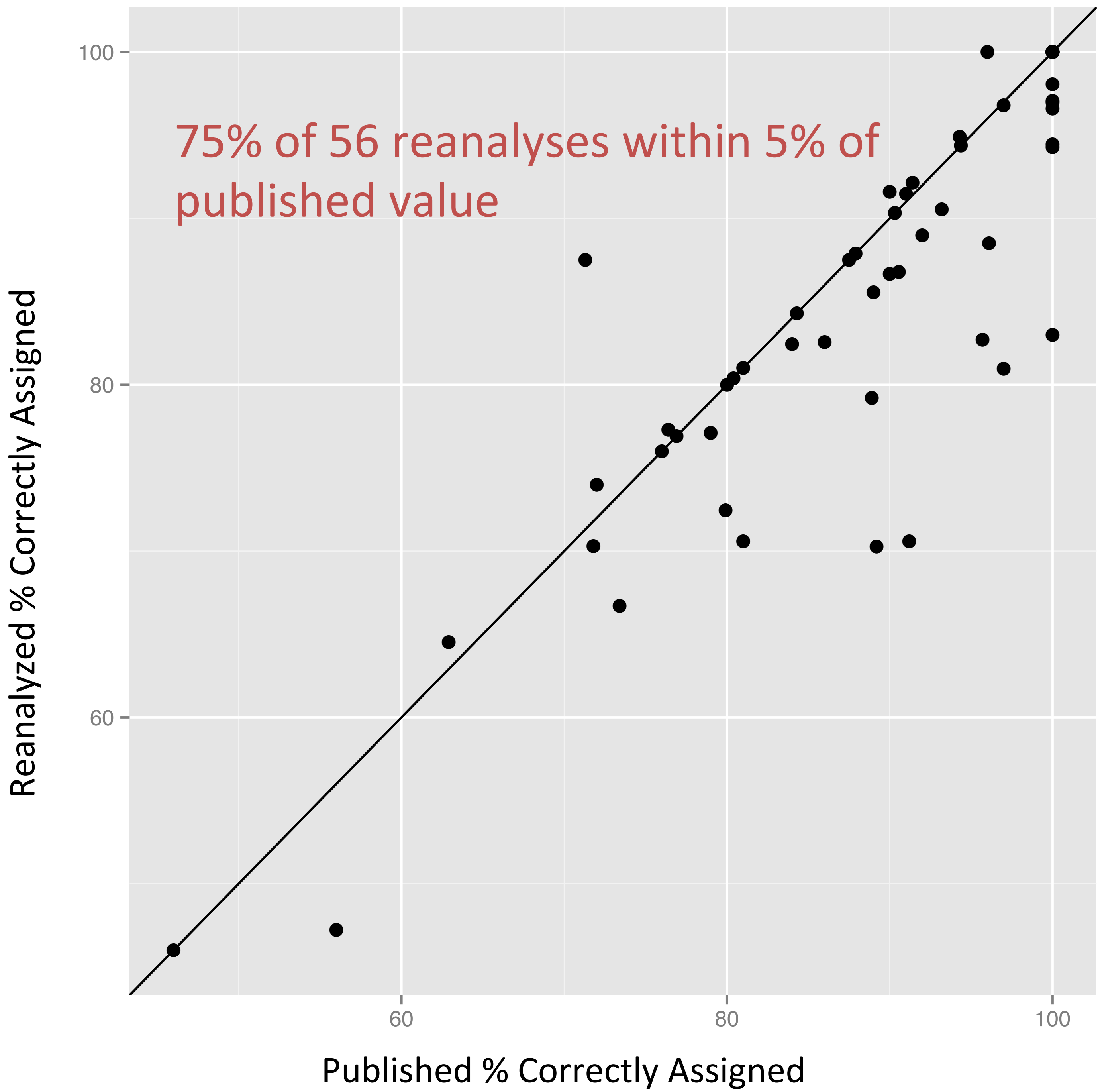
Reproducibility Part I

- We started with 101 studies
  - 16 didn't contain any of our three metrics
  - these were excluded
- Only 52 could be reproduced
  - 10% of the 516 datasets requested
- How far off were we?



83% of 23 reanalyses within
5% of published value
Reanalyzed % Variance Explained
Published % Variance Explained
100
80
60
40
40
60
80
100


70% of 23 reanalyses within 1 decimal place of published value
Reanalyzed Coefficient
Published Coefficient
20
10
0
0
10
20


75% of 56 reanalyses within 5% of published value
100
80
60
60
80
100
Reanalyzed % Correctly Assigned
Published % Correctly Assigned

Reproducibility Part I

- We started with 101 studies
  – 16 didn’t contain any of our three metrics
  – these were excluded

- Only 52 could be reproduced
  – 10% of the 516 datasets requested

- Strong differences between metrics

Conclusions

Reproducibility Part I

- Getting the data is the biggest obstacle
  - accounts for 80% of total

- Poor curation takes out only 4%
  - 22% of received datasets

- For DFA, reproducibility is quite good
  - but depends a lot on the metric used

Do data archiving policies work?

- journals now have data archiving policies

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

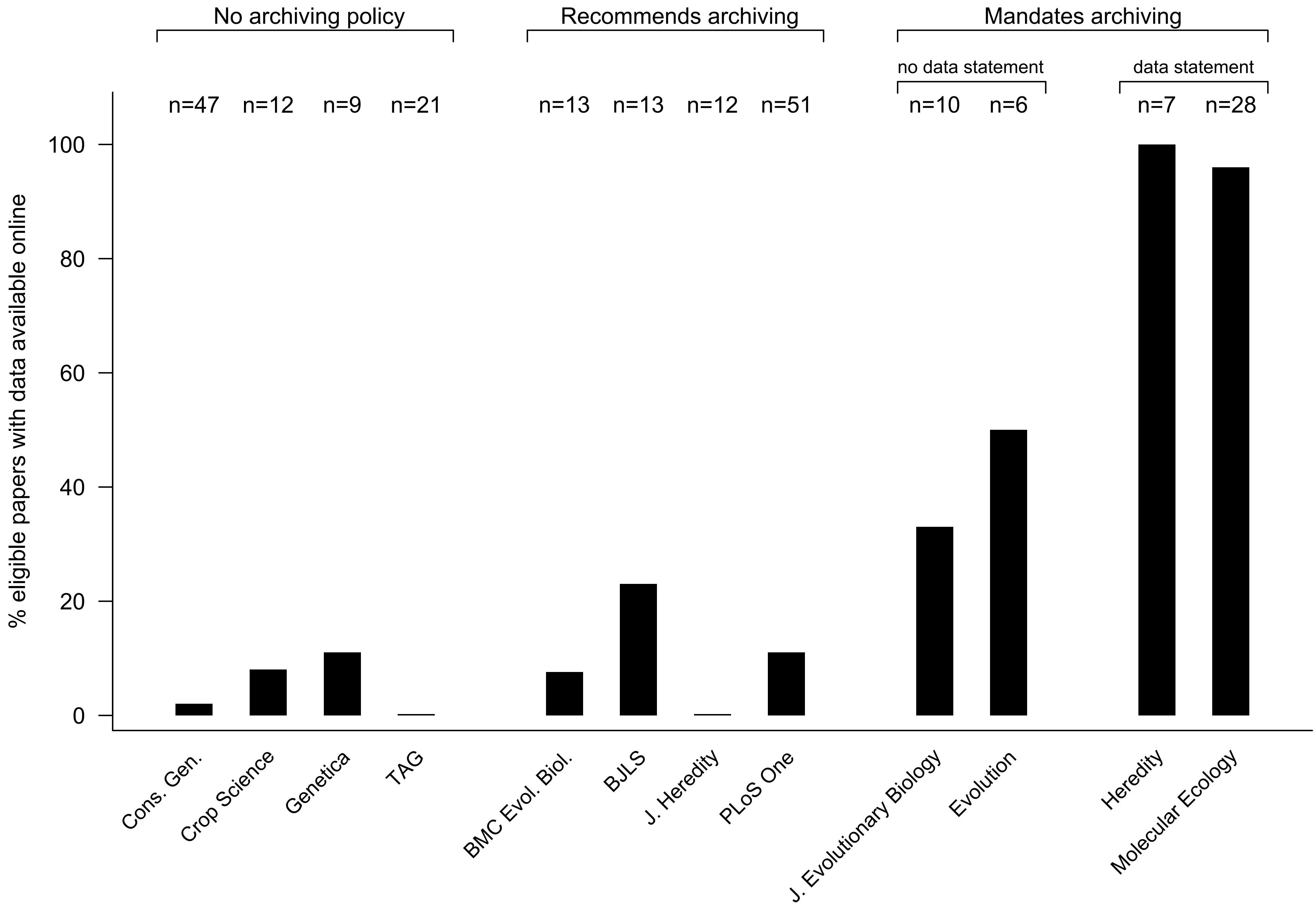
Vines *et al*. (2013) FASEBJ

- journals now have data archiving 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
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

Conclusions

- journals need to get tough
- give priority to papers with good archiving?
- have reviewers assess data statement

"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"

How journals can boost data sharing

The journal ecosystem is a powerful filter of scientific literature, promoting the best work into the best journals. Why not use a similar mechanism to encourage more comprehensive data sharing?

Several journals have introduced policies mandating that data be shared on a public archive at publication. However, these have met with limited success, perhaps because of authors' fear of losing control, being scooped in subsequent papers or having errors exposed. Moreover, compliance with data sharing policies is typically only checked after the paper is accepted.

To spur excellence in data sharing, journals must recognise that better sharing leads to stronger papers, and judge their submissions accordingly. Articles with feeble sharing efforts should either improve or be rejected.

A focus on publishing verifiable research correspondingly boosts journal reputation, and signals to the author community that withholding data restricts them to publication in less prestigious journals.

Timothy H. Vines
University of British Columbia

Reproducibility Part II: genetic data

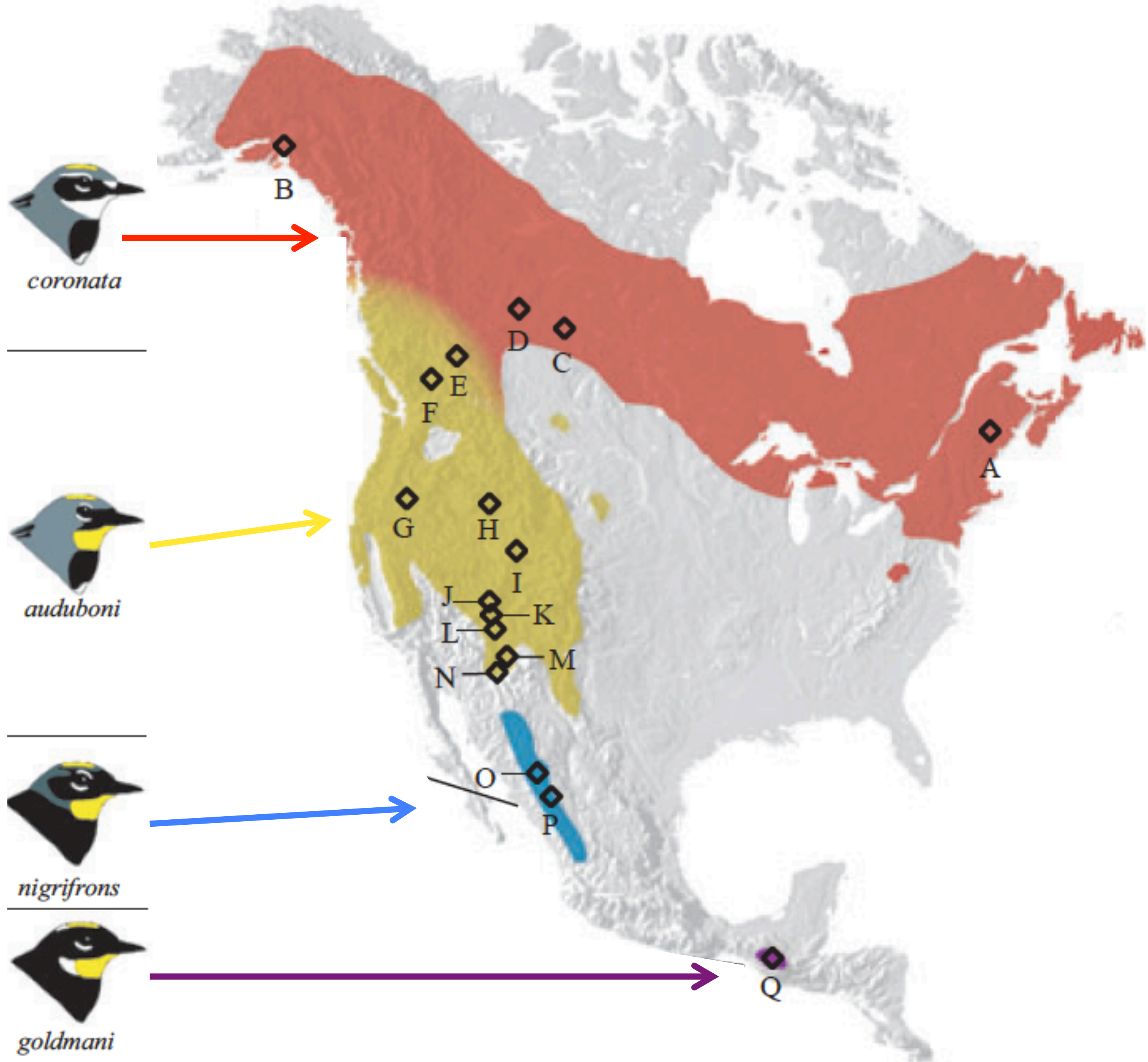
Gilbert *et al*. (2012) Molecular Ecology

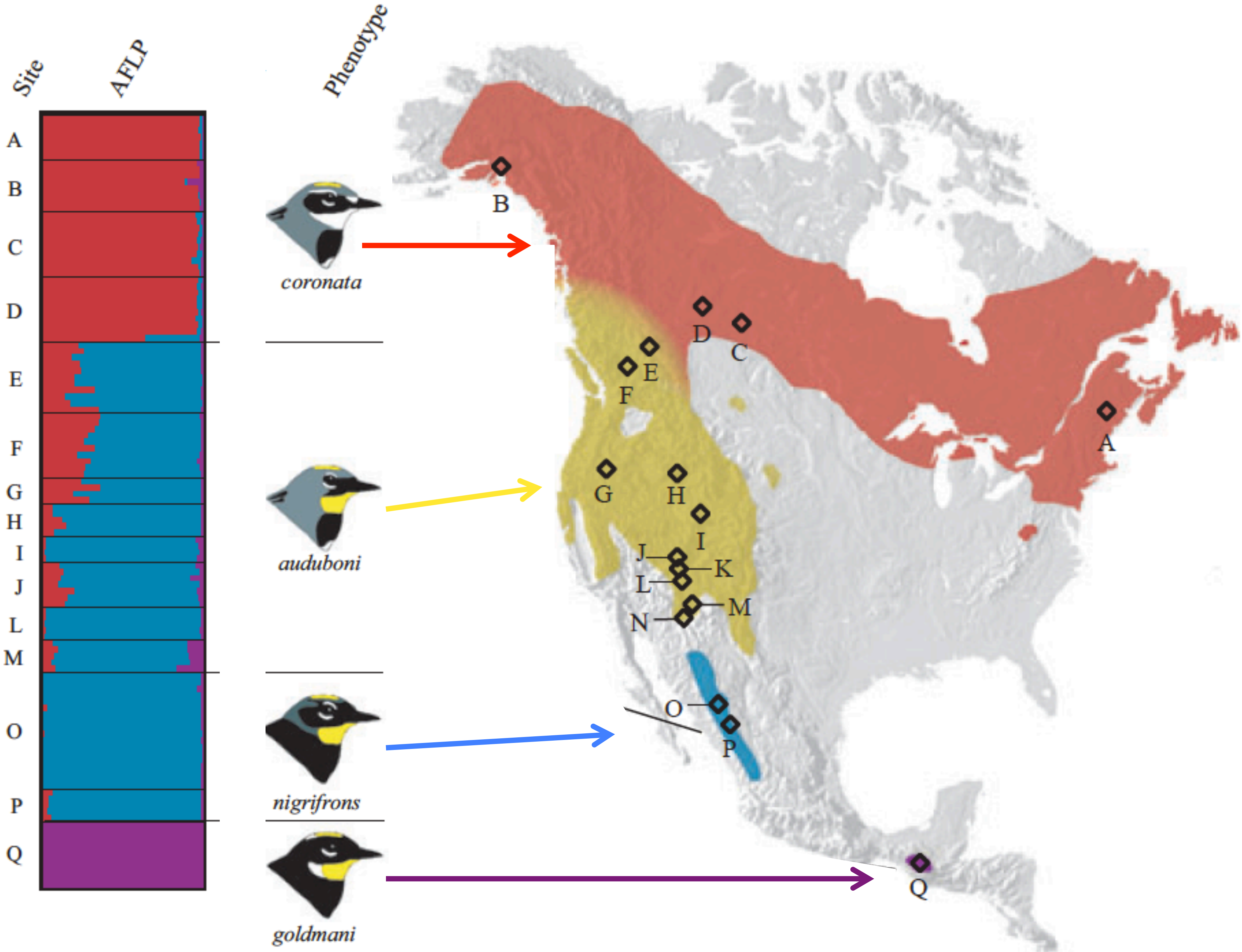
Reproducibility Part II

- Reproducing simple stats (a DFA) was OK
- modern stats are more sophisticated
- most involve numerical optimization
  - can get a different answer each time

Reproducibility Part II

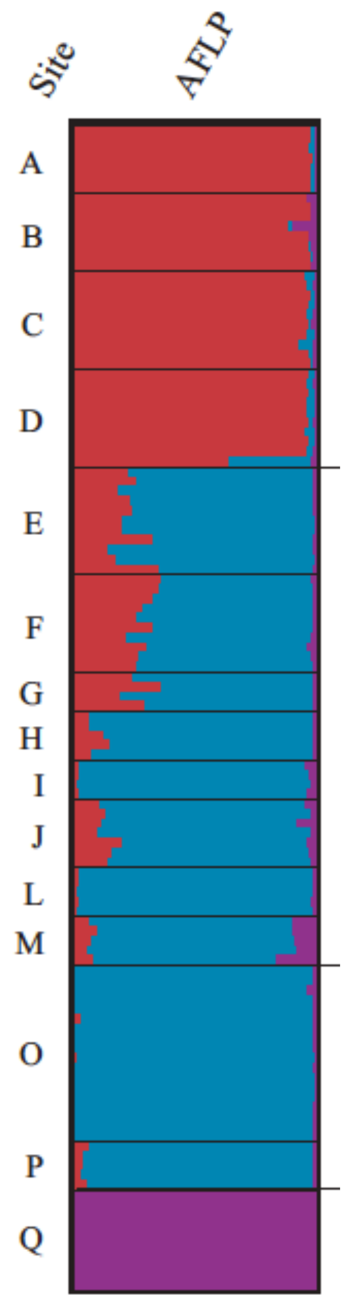
- 34 datasets from the previous study
- all have a STRUCTURE analysis


coronata
auduboni
nigrifrons
goldmani
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q


Site
AFLP
Phenotype
A
B
C
D
E
F
G
H
I
J
L
M
O
P
Q
coronata
auduboni
nigrifrons
goldmani
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q

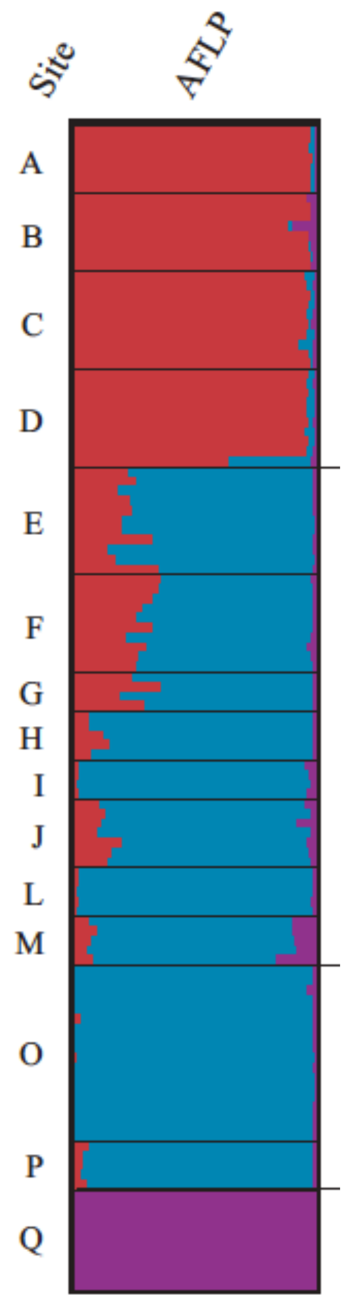
Reproducibility Part II

- 34 datasets from the previous study
- all have a STRUCTURE analysis
- this uses extensive numerical optimization


Site
AFLP
A
B
C
D
E
F
G
H
I
J
L
M
O
P
Q

Reproducibility Part II

- 34 datasets from the previous study
- all have a STRUCTURE analysis
- this uses extensive numerical optimization
- output is K, the number of distinct clusters



Site
AFLP
A
B
C
D
E
F
G
H
I
J
L
M
O
P
Q

Reproducibility Part II

- Can we reproduce their value of *K*?

- 4 studies were excluded
  - no data, irregular use of STRUCTURE

- Reanalyzed remaining 30 datasets

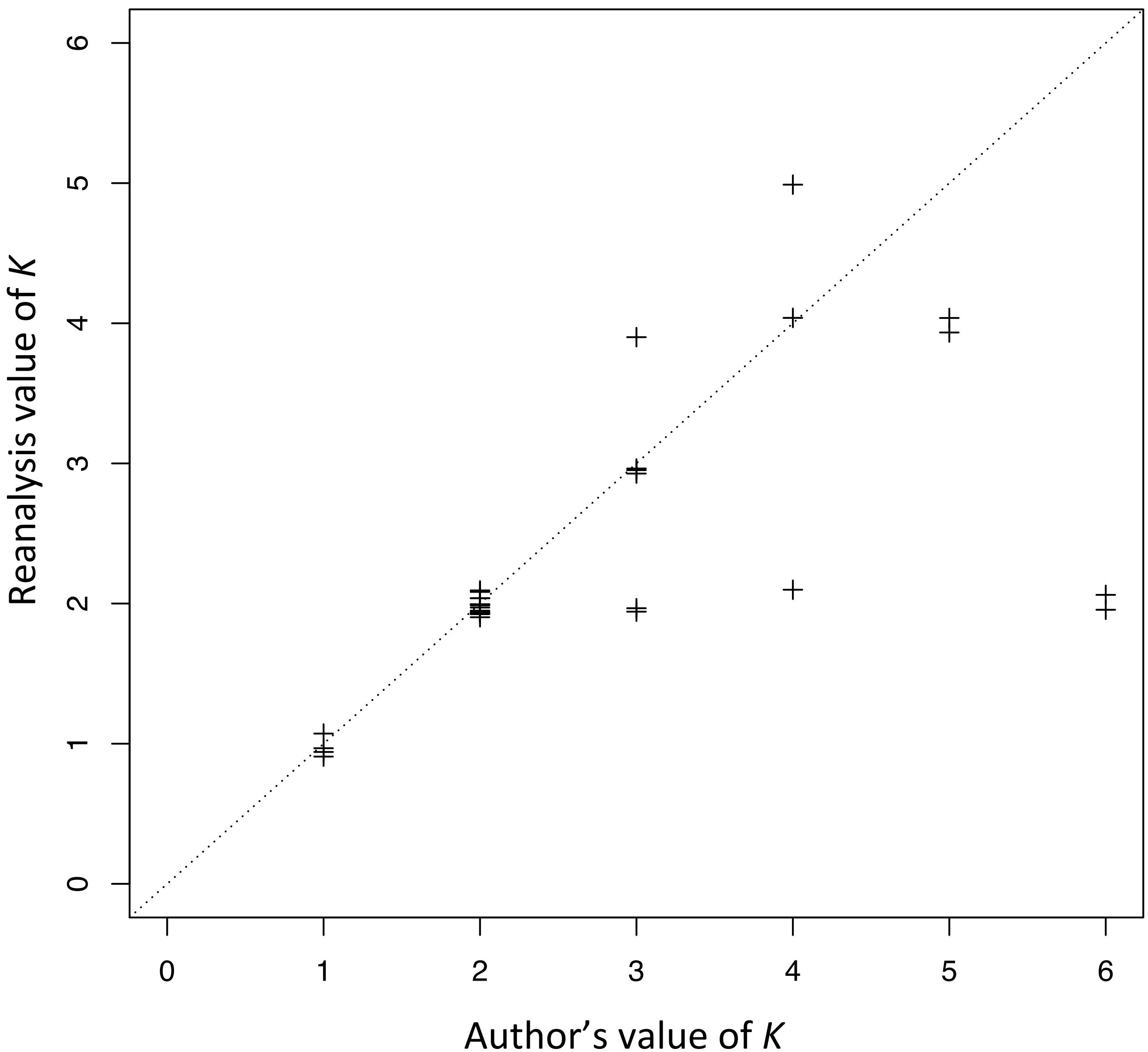
Outcome	**No. datasets**	**Percent**
Strange use of STRUCTURE	2	6
Missing data	2	6
Incorrect/incomplete data	3	9
Reanalysis attempted:		
K didn't match	6	18
K matched	21	62
Overall Total	34	100

Reproducibility Part II

- Can we reproduce their value of *K*?

- 4 studies were excluded
  - no data, irregular use of STRUCTURE

- How close did we get?



Reanalysis value of K
Author's value of K
0
1
2
3
4
5
6

Reproducibility Part II

- Most mismatches from poor software use
  - stochastic methods need many iterations
  - too few and the answer is unreliable

Reproducibility Part II

- Most mismatches from poor software use
  - stochastic methods need many iterations
  - too few and the answer is unreliable

- Poor curation was less of a problem

Grand Conclusions

- STRUCTURE reproducibility > DFA
  - 65% vs 50%
- Is under 100% reproducibility unacceptable?
- Maybe replication is more important

- Data availability is the biggest problem
  - without it, reproducibility = 0
- We need stronger data archiving policies
- May mean better science as well
  - someone will check your data…

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