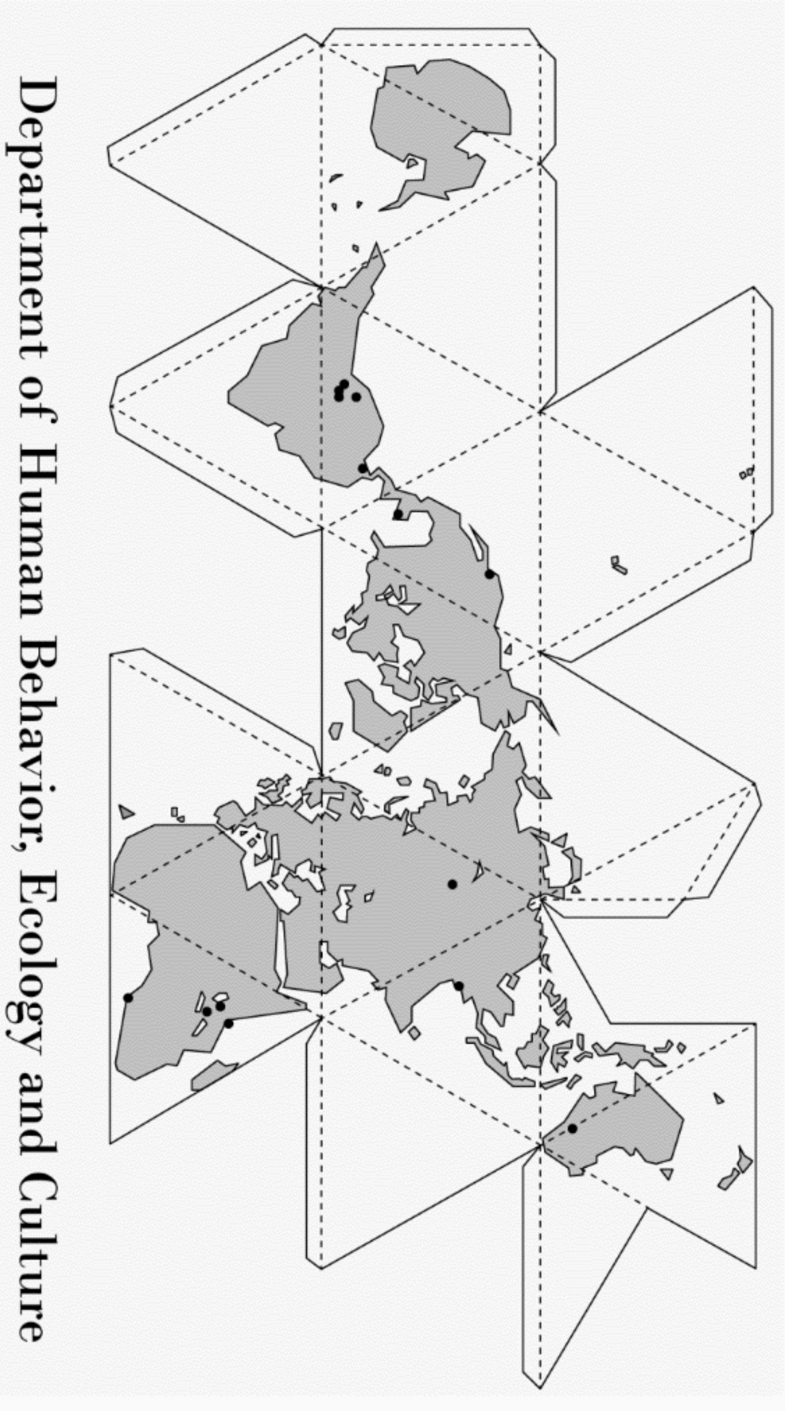
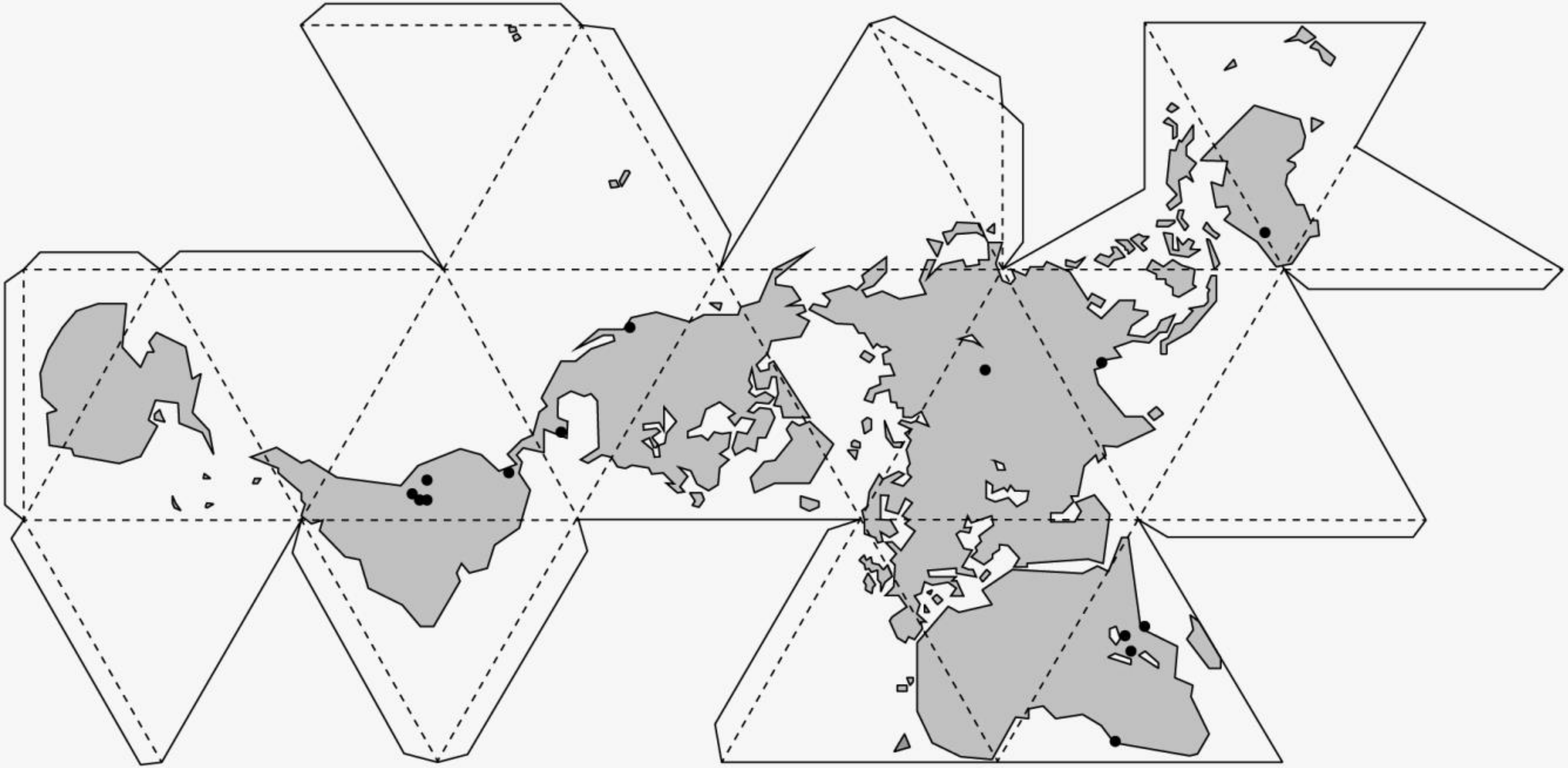


Open Science in the Field

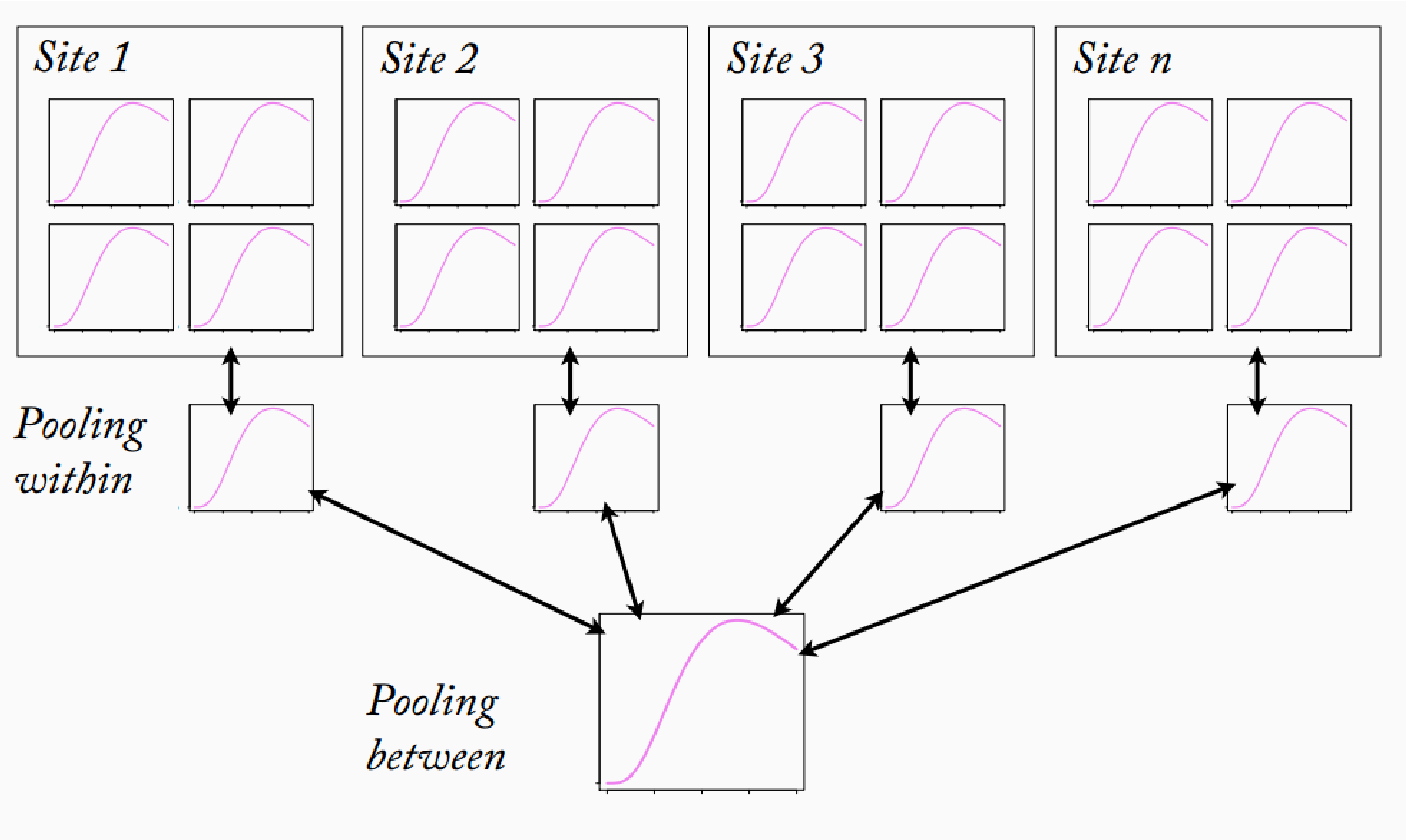
Bret Beheim

🐦 @babeheim

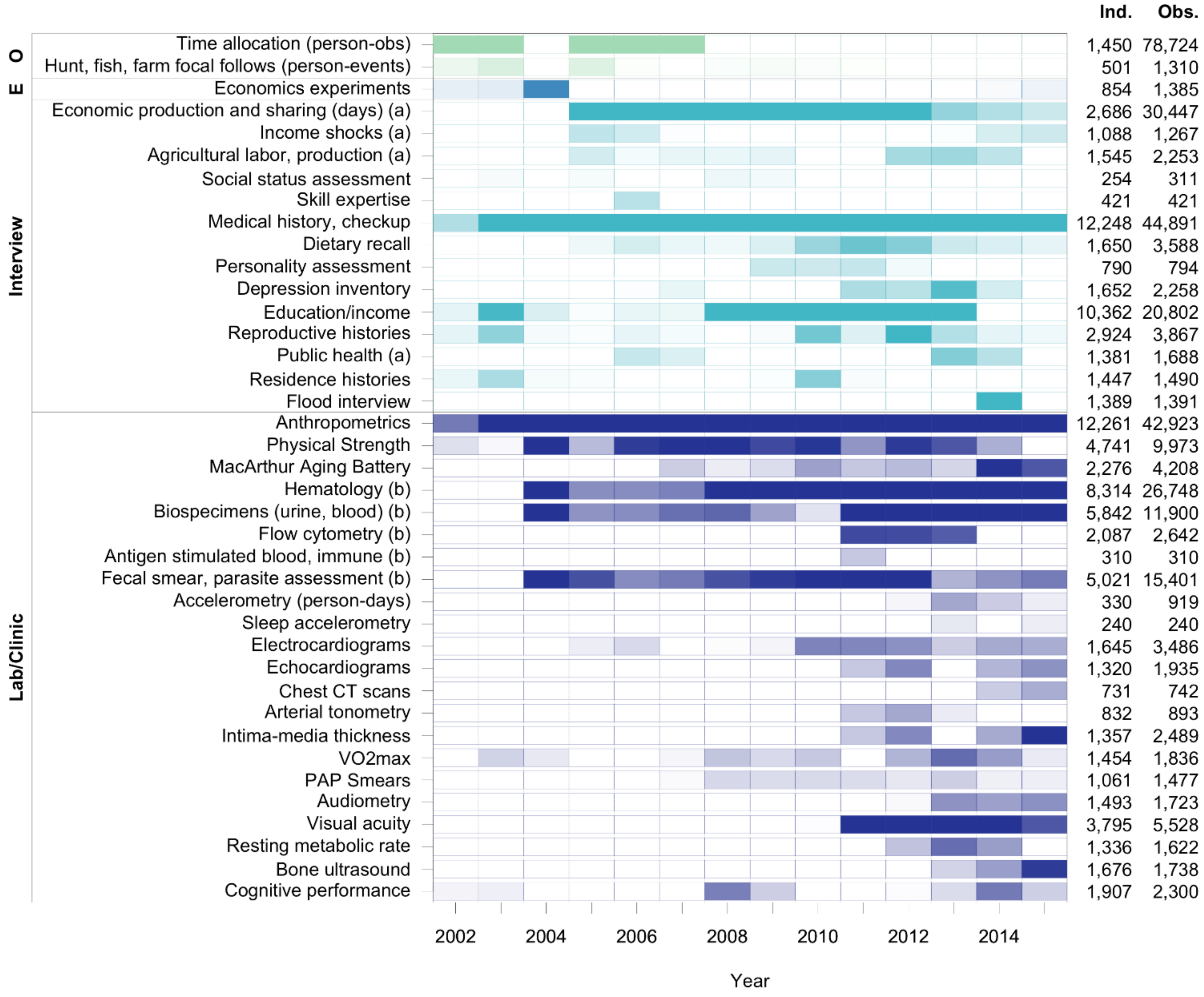
🏠 Max Planck Institute for Evolutionary Anthropology

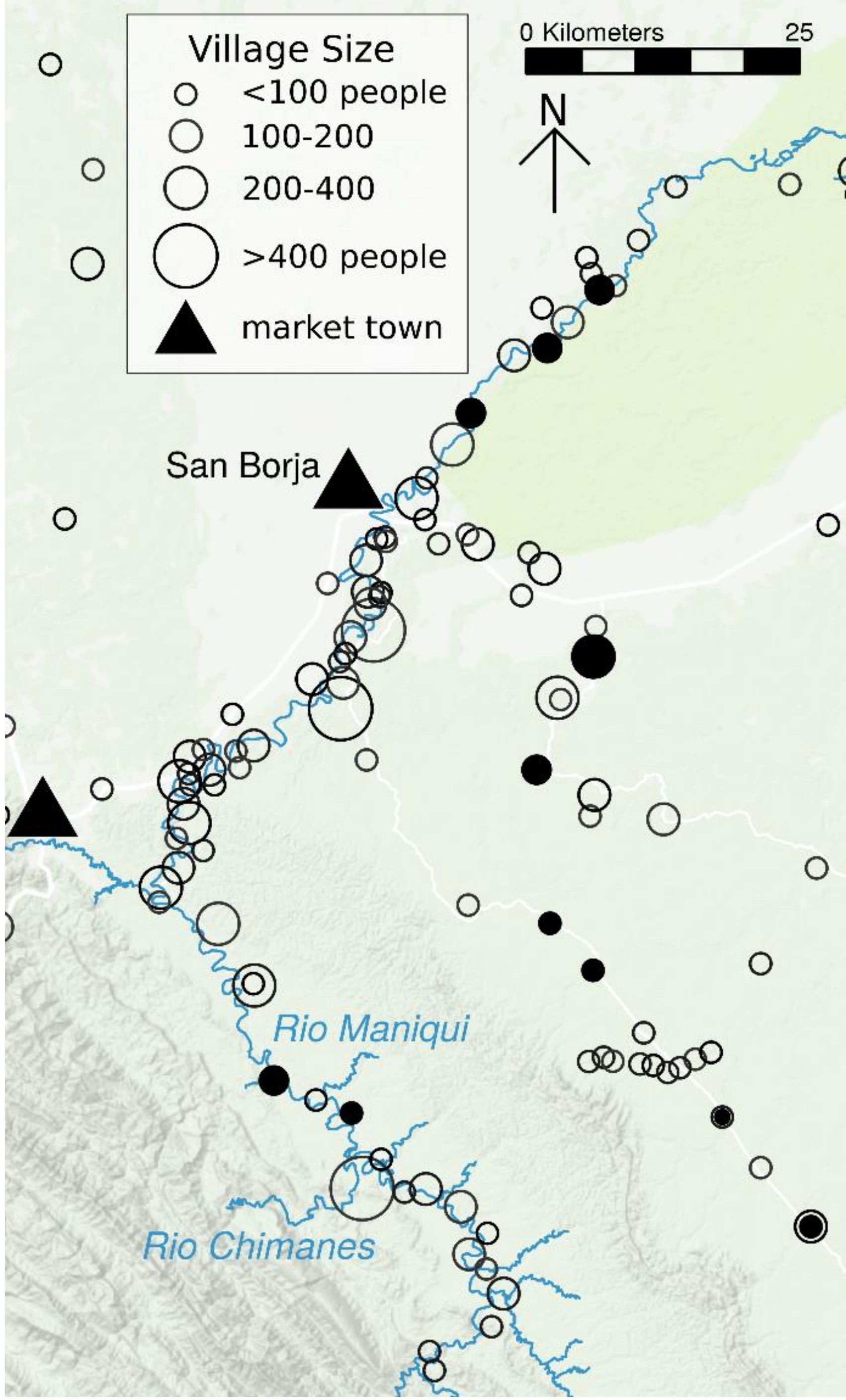


Department of Human Behavior, Ecology and Culture



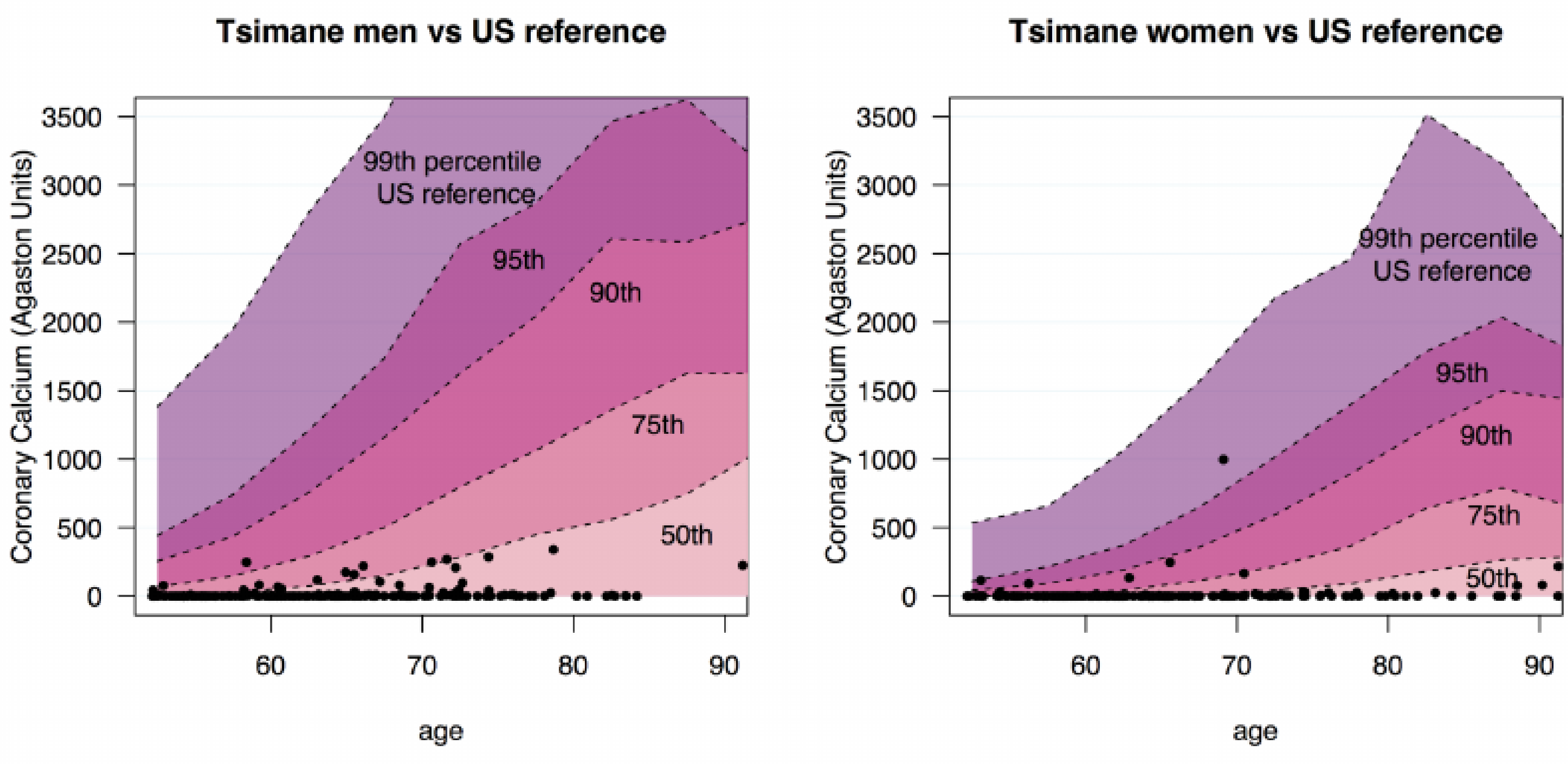
Koster, McElreath, et al. (2019) The Life History of Human Foraging





Gurven, et al. (2016) The Tsimane Health and Life History Project

Tsimane hearts are some of the healthiest in the world

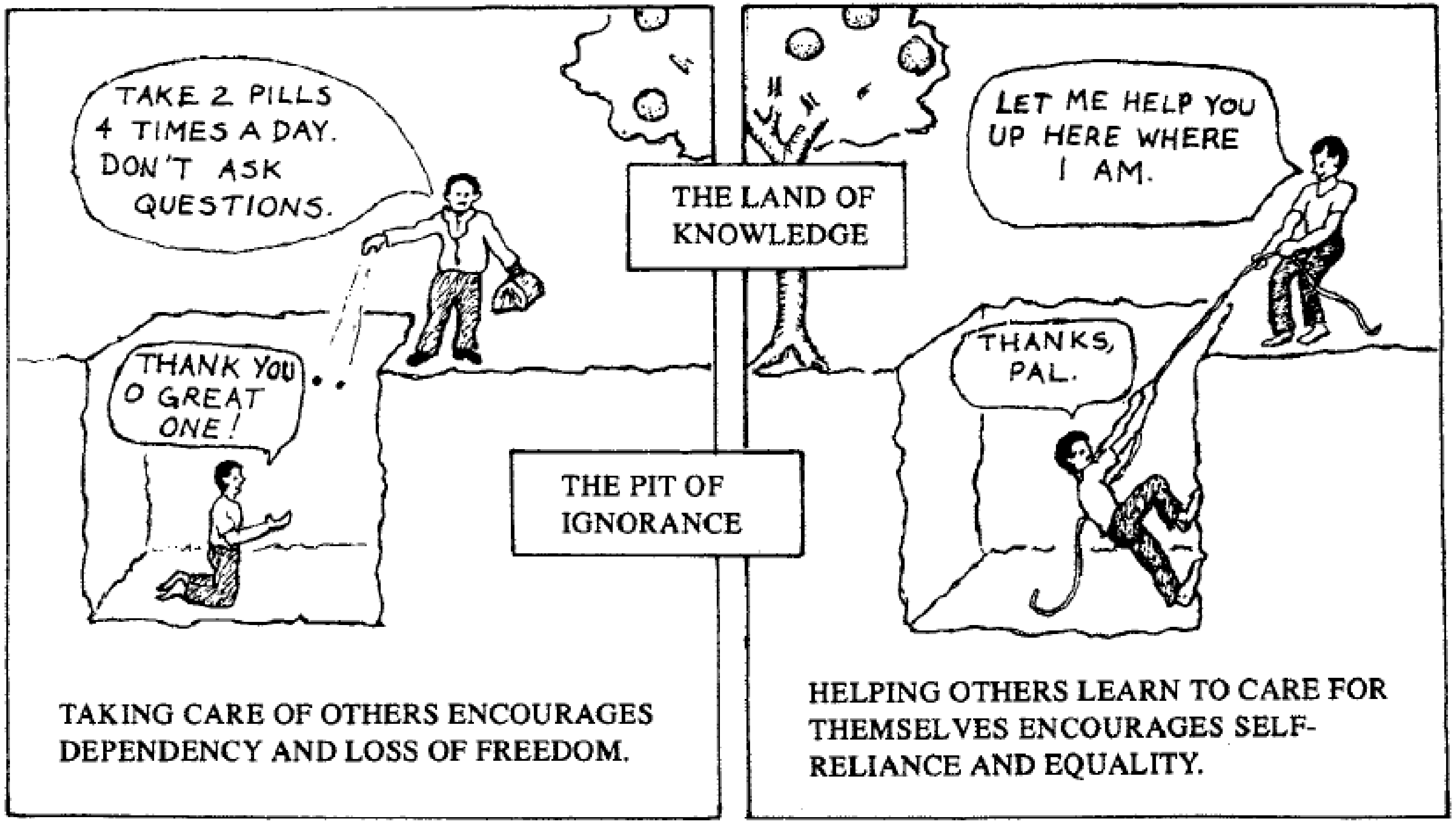


Kaplan, et al. (2017) "Coronary atherosclerosis in indigenous South American Tsimane: a cross-sectional cohort study" **DOI: 10.1016/S0140-6736(17)30752-3**

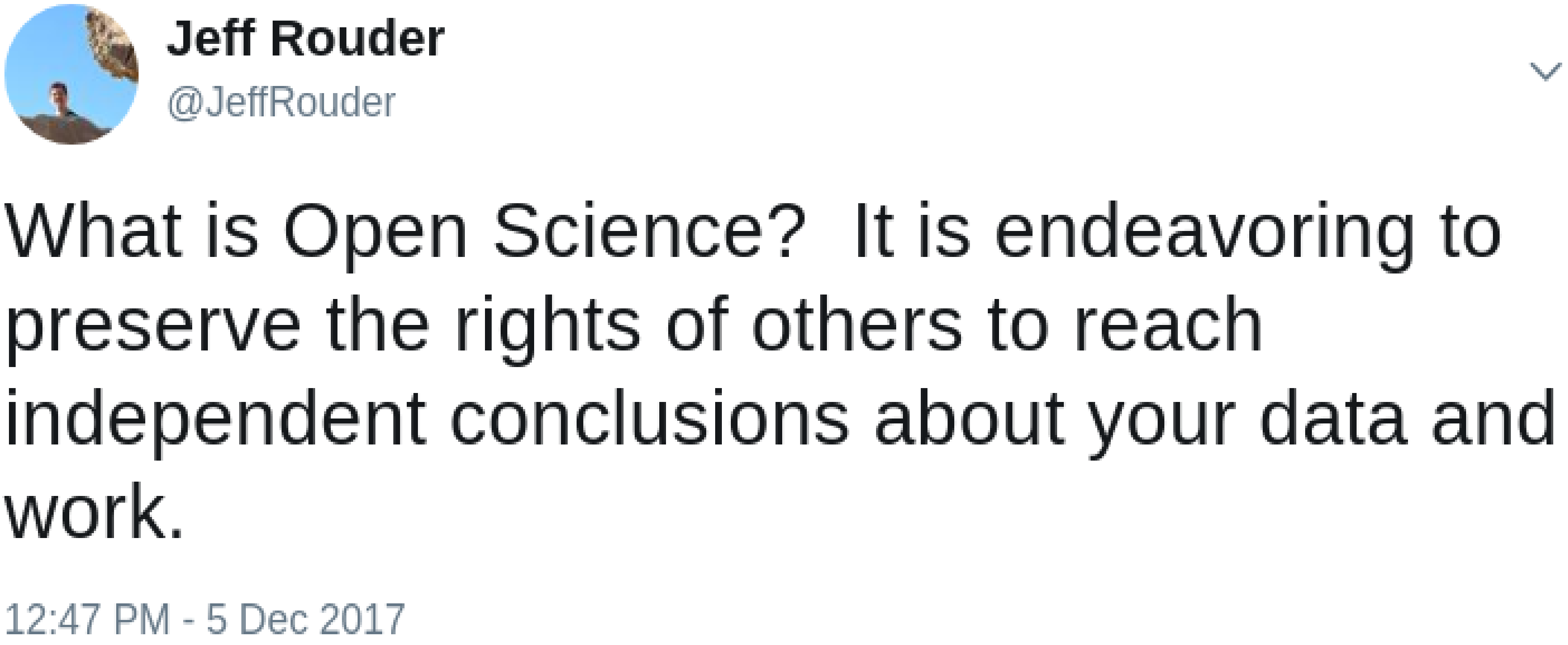
Typical Field Datasets in Evolutionary Anthropology

- Quantitative
- Small-scale
- By/from humans
- Noisy
- Complicated
- Longitudinal

TWO APPROACHES TO HEALTH CARE


TAKE 2 PILLS 4 TIMES A DAY. DON'T ASK QUESTIONS.
THANK YOU O GREAT ONE!
THE LAND OF KNOWLEDGE
THE PIT OF IGNORANCE
LET ME HELP YOU UP HERE WHERE I AM.
THANKS, PAL.
TAKING CARE OF OTHERS ENCOURAGES DEPENDENCY AND LOSS OF FREEDOM.
HELPING OTHERS LEARN TO CARE FOR THEMSELVES ENCOURAGES SELF-RELIANCE AND EQUALITY.

From *Where There Is No Doctor: A Village Health Care Handbook,*
http://hesperian.org/books-and-resources/



Jeff Rouder
@JeffRouder

What is Open Science? It is endeavoring to preserve the rights of others to reach independent conclusions about your data and work.

12:47 PM - 5 Dec 2017

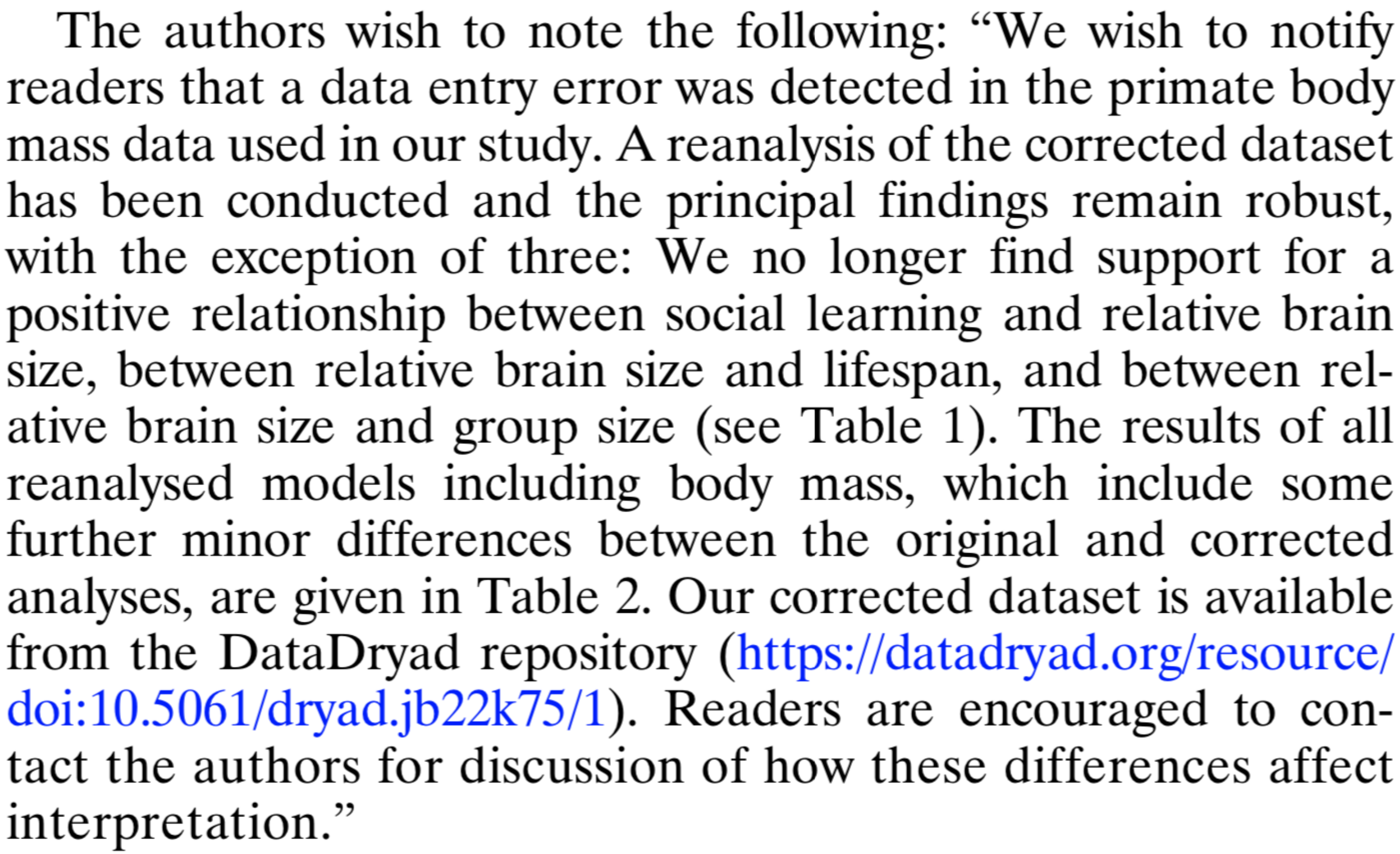
How to produce reliable knowledge about the world?

replication: similar results from
different researchers,
different data

reproduction: similar results from
different researchers,
same data

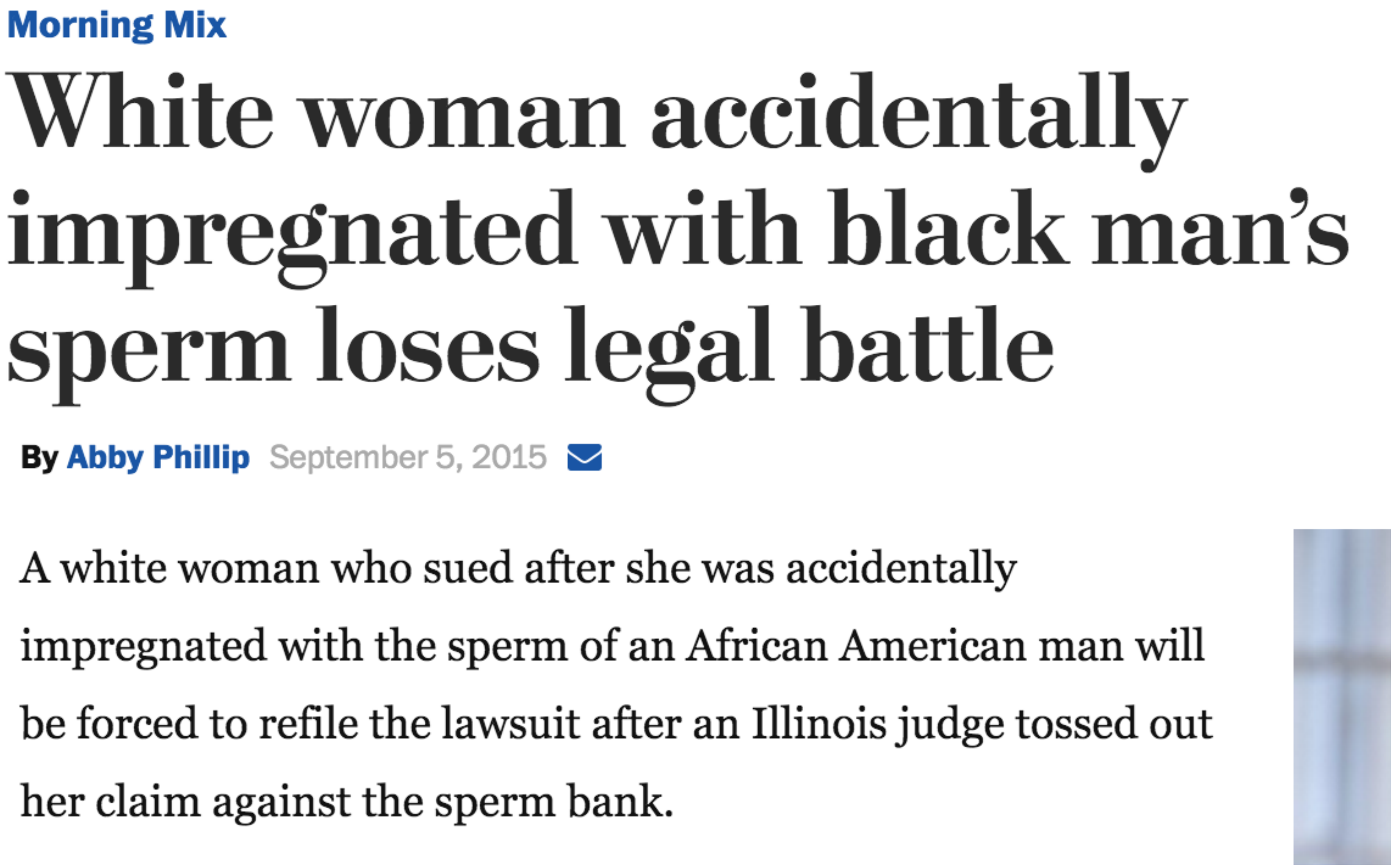
1. *Patil, Peng, Leek (2016) A statistical definition for reproducibility and replicability.*
2. *Ihle, et al. (2017) Striving for transparent and credible research: practical guidelines for behavioral ecologists.*

Bad data can sink results



The authors wish to note the following: “We wish to notify readers that a data entry error was detected in the primate body mass data used in our study. A reanalysis of the corrected dataset has been conducted and the principal findings remain robust, with the exception of three: We no longer find support for a positive relationship between social learning and relative brain size, between relative brain size and lifespan, and between relative brain size and group size (see Table 1). The results of all reanalysed models including body mass, which include some further minor differences between the original and corrected analyses, are given in Table 2. Our corrected dataset is available from the DataDryad repository (https://datadryad.org/resource/doi:10.5061/dryad.jb22k75/1). Readers are encouraged to contact the authors for discussion of how these differences affect interpretation.”

Typos at the Sperm Bank



Morning Mix

White woman accidentally impregnated with black man's sperm loses legal battle

By Abby Phillip September 5, 2015

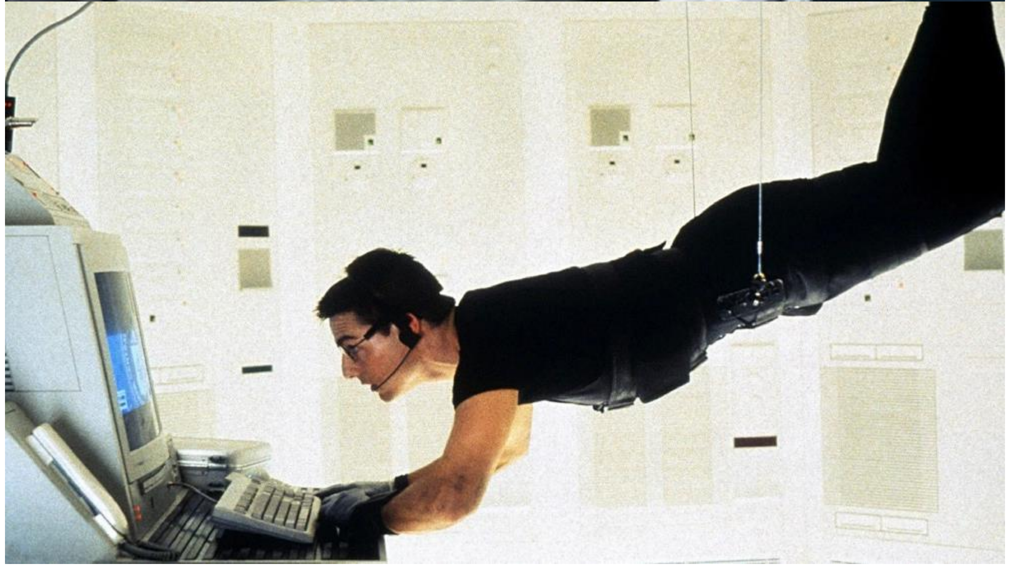
A white woman who sued after she was accidentally impregnated with the sperm of an African American man will be forced to refile the lawsuit after an Illinois judge tossed out her claim against the sperm bank.

'According to the suit, the couple chose sperm from donor No. 380...instead, they were given sperm from donor No. 330...They blame a paper records system that allegedly caused an employee to misread the numbers.'

Barriers to data availability

Reasonable:
- Access to computer / internet
- Read through supplementary info
- Detailed data request
- Respecting authors’ restrictions

Unreasonable:
- Privileged insiders only
- Promises of payment, etc.
- Decoding undocumented data
- Requires travel great distances



The Reasonable Researcher Test

For a given empirical analysis, can a *typical* early-career researcher reproduce the essential results with a *reasonable* amount of effort?

i. Can they acquire materials for the published analysis?

ii. Given materials, can they recover the reported results?

Literature: Empirical Social Learning

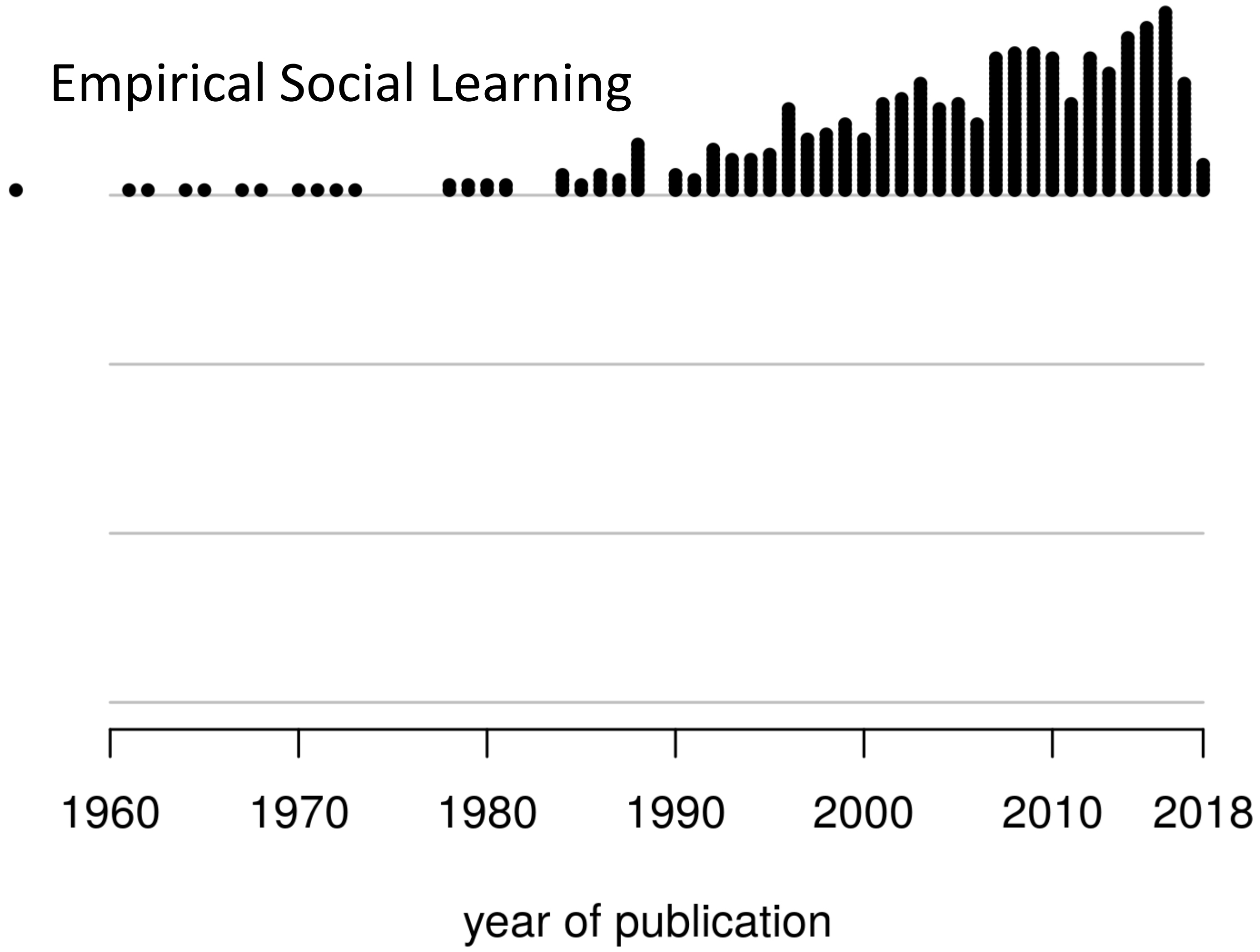
- papers & chapters through citation backtracking

- three criteria for inclusion in study:
    - empirical
    - quantitative
    - “social learning” (or similar)

- requests sent from reproducibility@eva.mpg.de

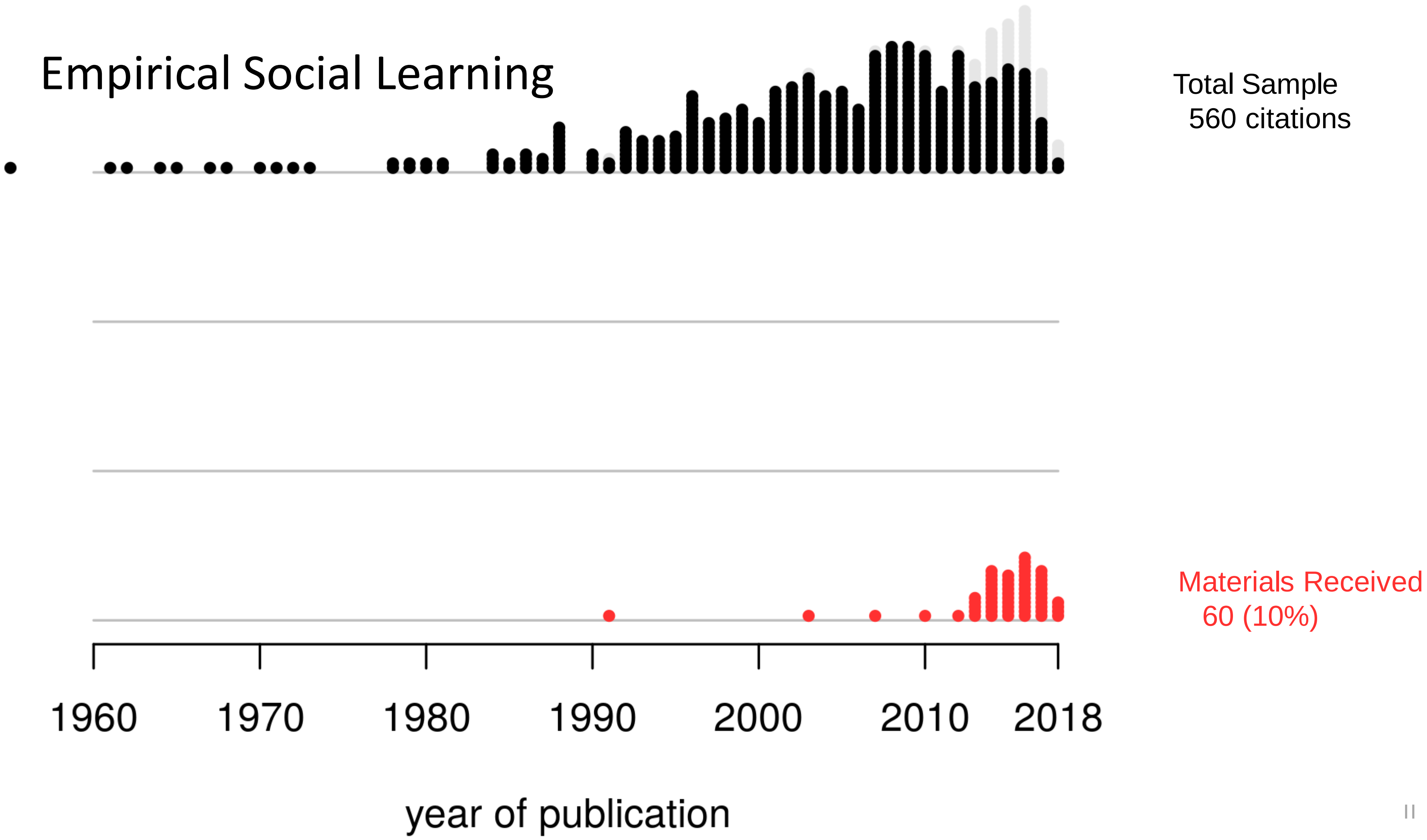
(1270 total emails last count)

Empirical Social Learning

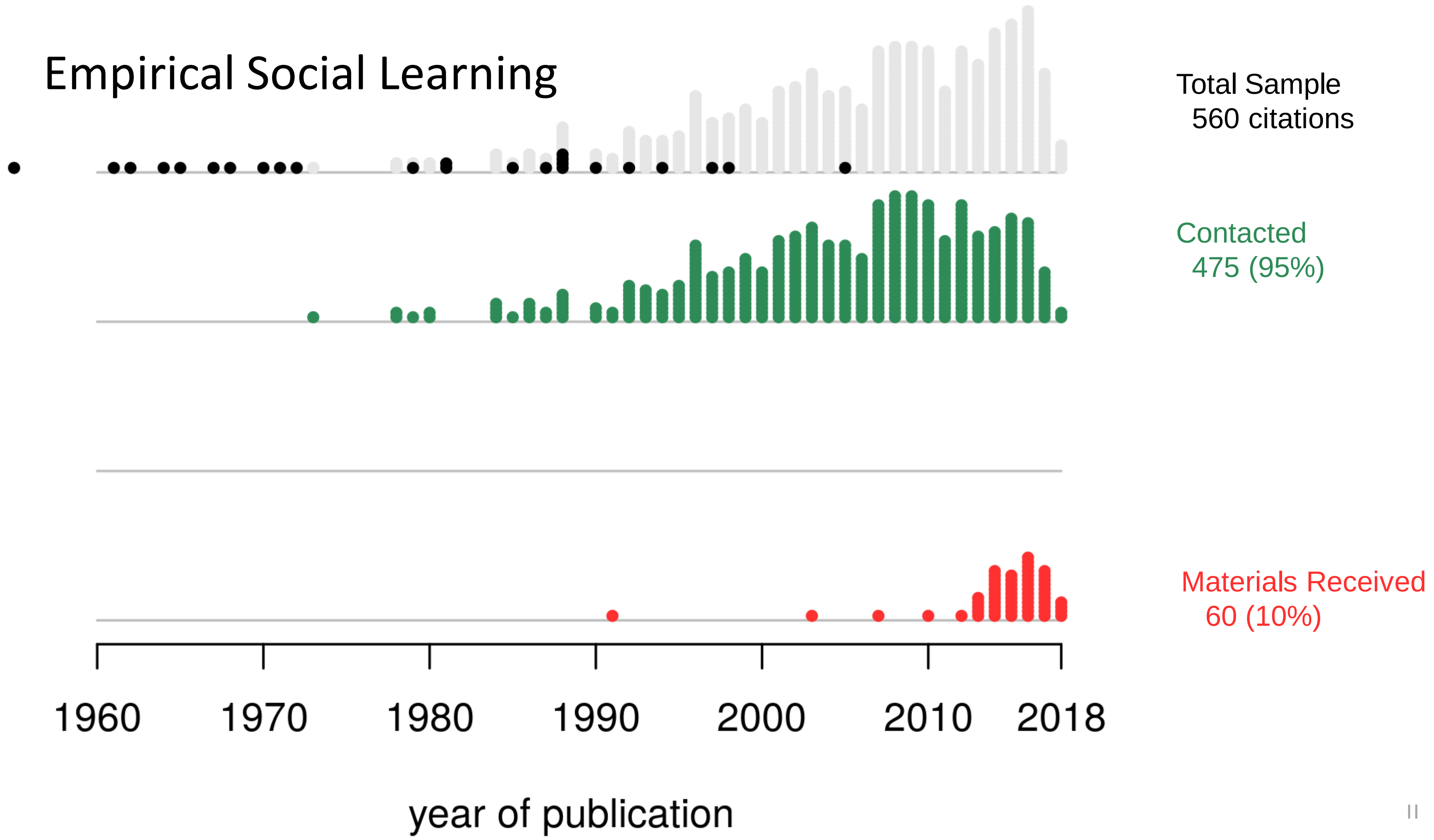


Total Sample
560 citations

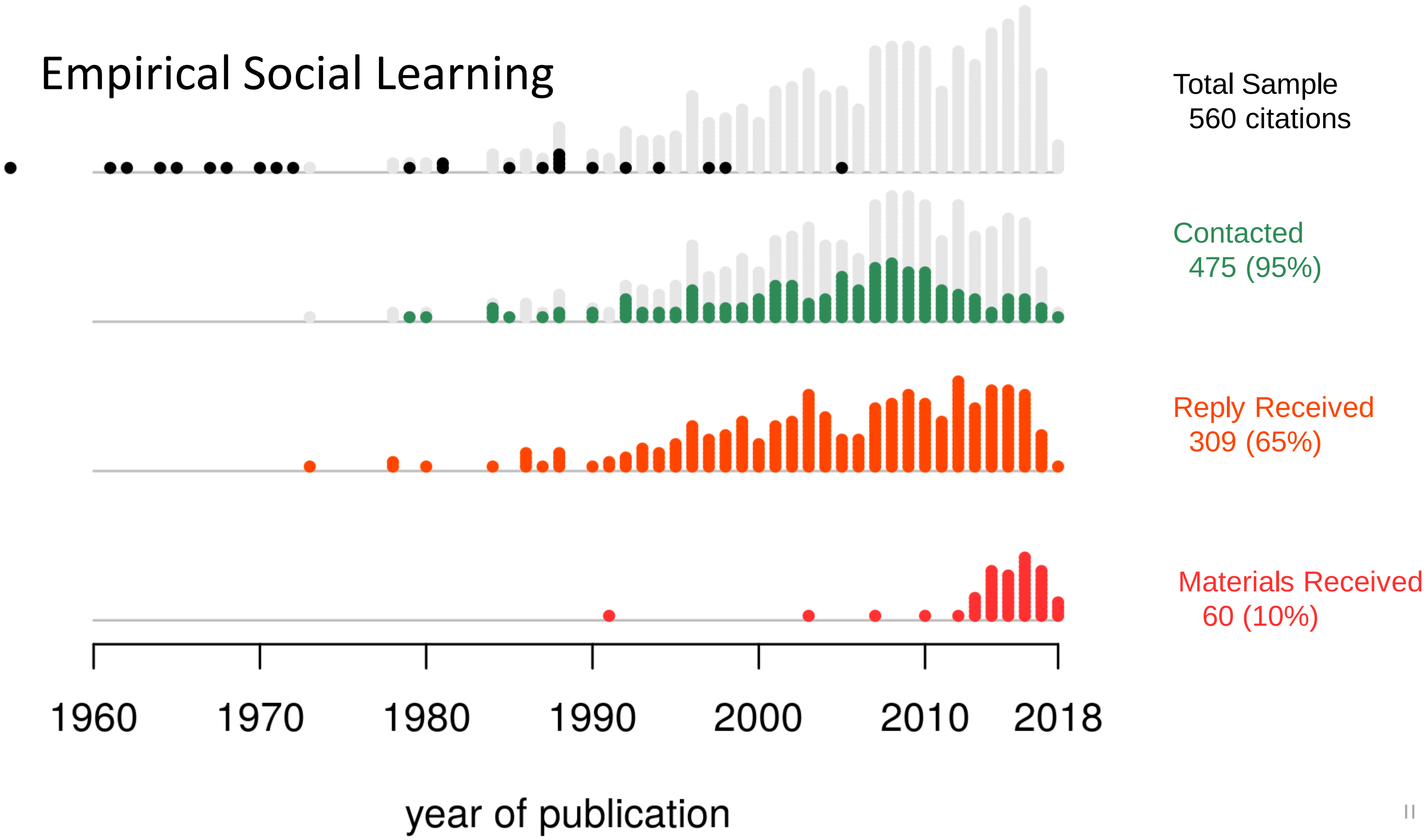
Empirical Social Learning



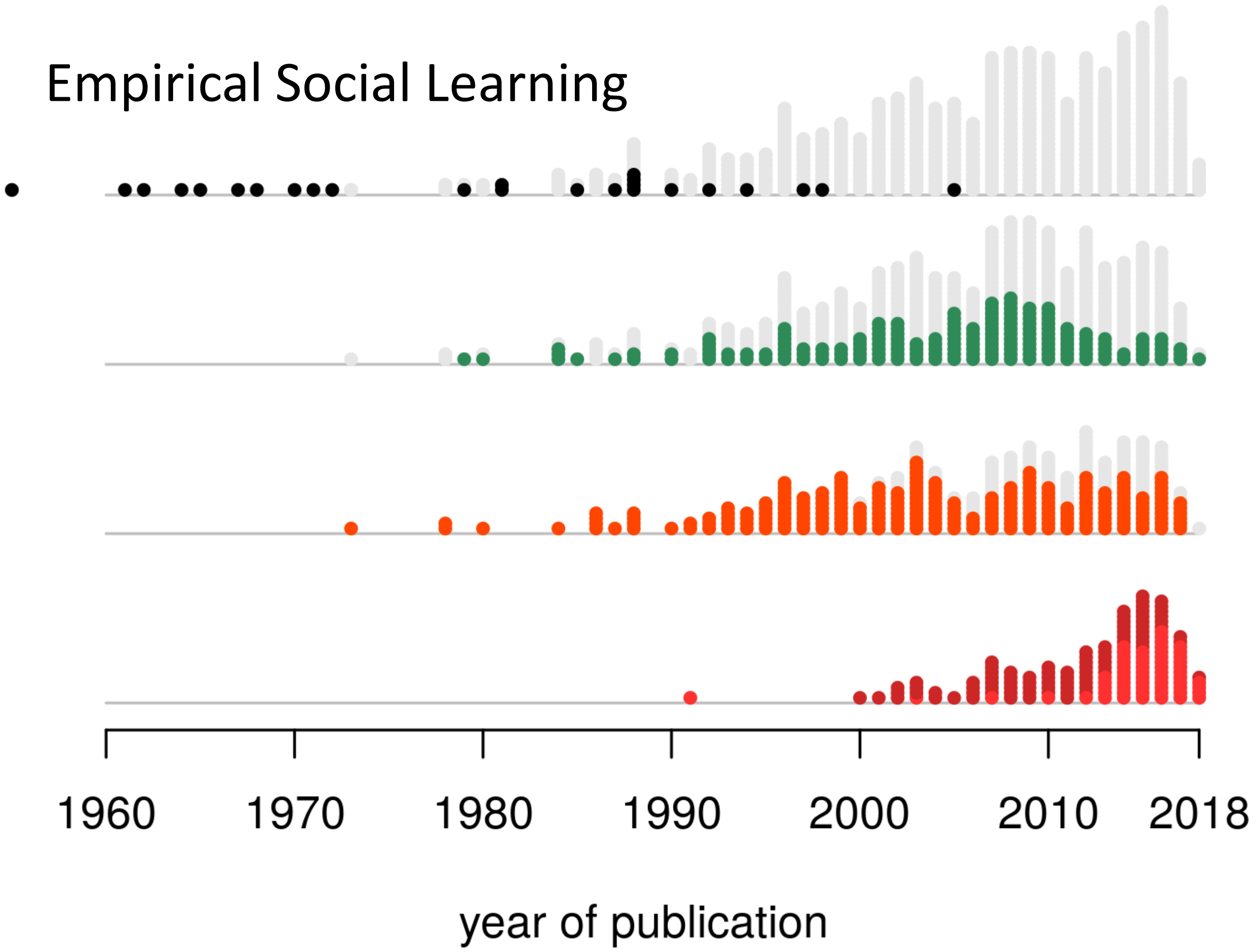
Empirical Social Learning



Empirical Social Learning



Empirical Social Learning



Total Sample
560 citations

Contacted
475 (95%)

Reply Received
309 (65%)

Materials Received
60 (10%)
+ 87 (28%)

147 (26%)

Ethnographic context on causes of data decay

"there is no code, scripting or paper data still in existence...this does NOT mean that the results are in any way invalid. I do not agree with the focus of your project “

“looking at studies done even just a few years ago (i.e. published 2015 or earlier) is completely counter-productive”

"I don't see the point here really.“

Ethnographic context on causes of data decay

"Your project will be a good opportunity for me to put my folders in order and to make this material available.‟

"My initial thought was to say that this has been far too long – this study was conducted almost 20 years ago. But I had a quick look and amazingly I managed to find the data.‟

"something I’ve been meaning to do myself for a while for these older studies‟

Ethnographic context on causes of data decay

"My god, that was 17 years ago. I have no idea where that data is."

"I wish you had asked us for the data 3 years ago because I too would like to have the data."

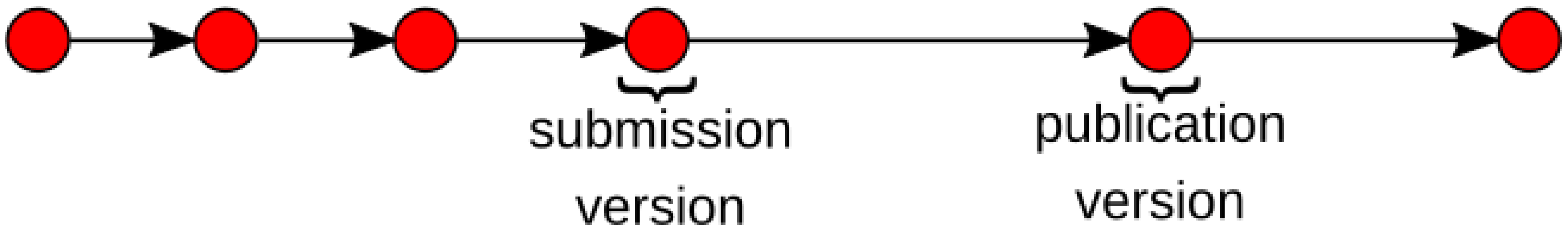
"…such studies often uses expertise from several people, and make multiple intermediate versions of the datasets. This is often done without really knowing what will end up in a paper and what will not, or even what the paper will be about"

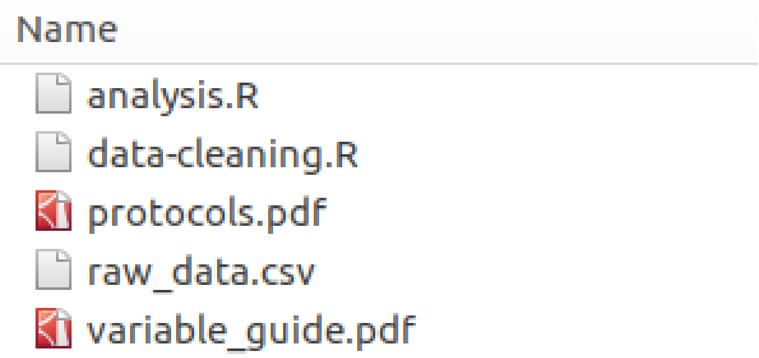
Version control for reasonable researchers

happygitwithr.com

atlassian.com/git

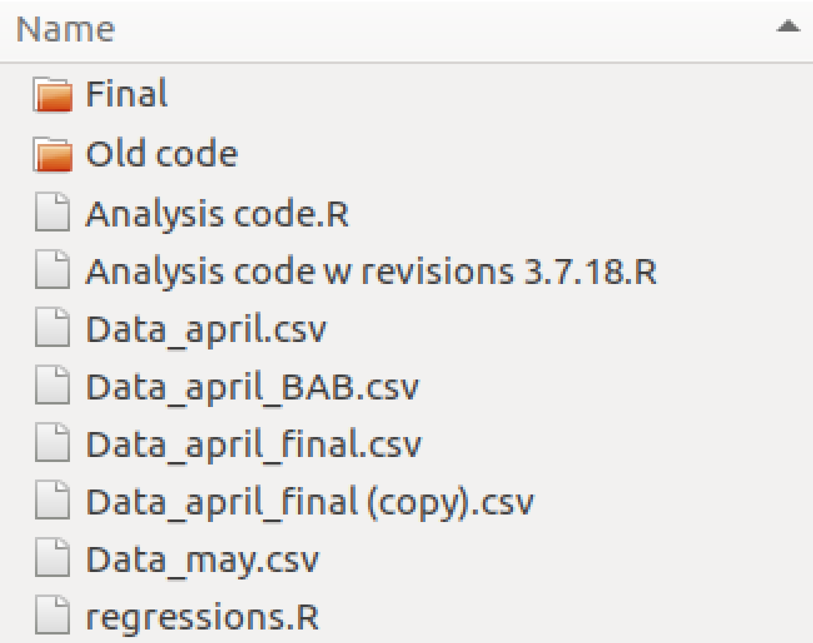
swcarpentry.github.io/git-novice/





Name

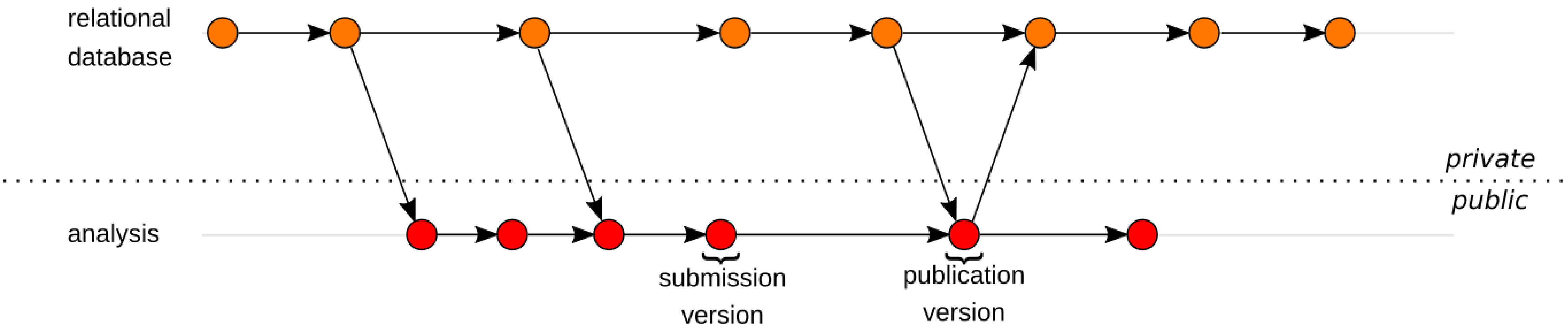
analysis.R
data-cleaning.R
protocols.pdf
raw_data.csv
variable_guide.pdf



Name

Final
Old code
Analysis code.R
Analysis code w revisions 3.7.18.R
Data_april.csv
Data_april_BAB.csv
Data_april_final.csv
Data_april_final (copy).csv
Data_may.csv
regressions.R

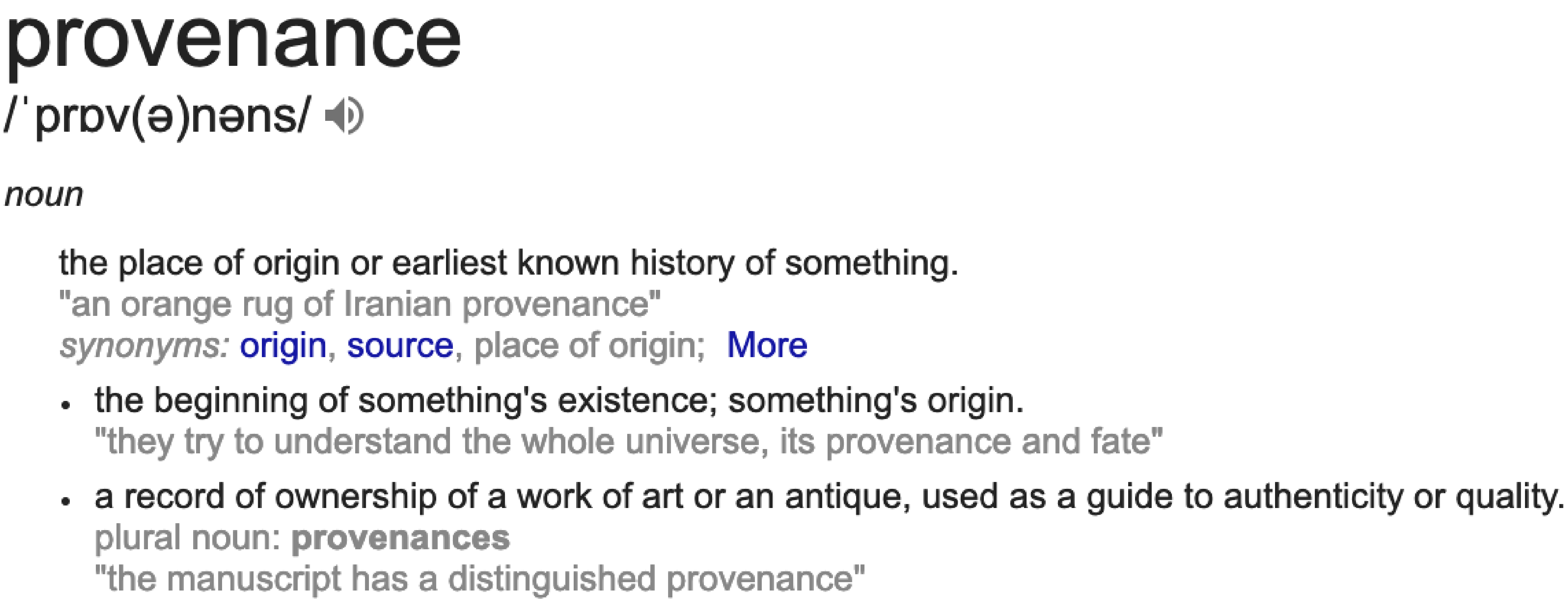
Idea: versioning the data too



Two barriers to reliable field analyses

Data Decay

Provenance Problems



provenance

/ˈprɒv(ə)nəns/ 🔊

noun

the place of origin or earliest known history of something.
"an orange rug of Iranian provenance"
synonyms: origin, source, place of origin; More

- the beginning of something's existence; something's origin.
  "they try to understand the whole universe, its provenance and fate"
- a record of ownership of a work of art or an antique, used as a guide to authenticity or quality.
  plural noun: **provenances**
  "the manuscript has a distinguished provenance"

Stamp’s Law

“The individual source of the statistics may easily be the weakest link. Harold Cox tells a story of his life as a young man in India. He quoted some statistics to a Judge, an Englishman, and a very good fellow. His friend said, "Cox, when you are a bit older, you will not quote Indian statistics with that assurance. The Government are very keen on amassing statistics — they collect them, add them, raise them to the *n*th power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the *chowty dar* (village watchman), who just puts down what he damn pleases.“

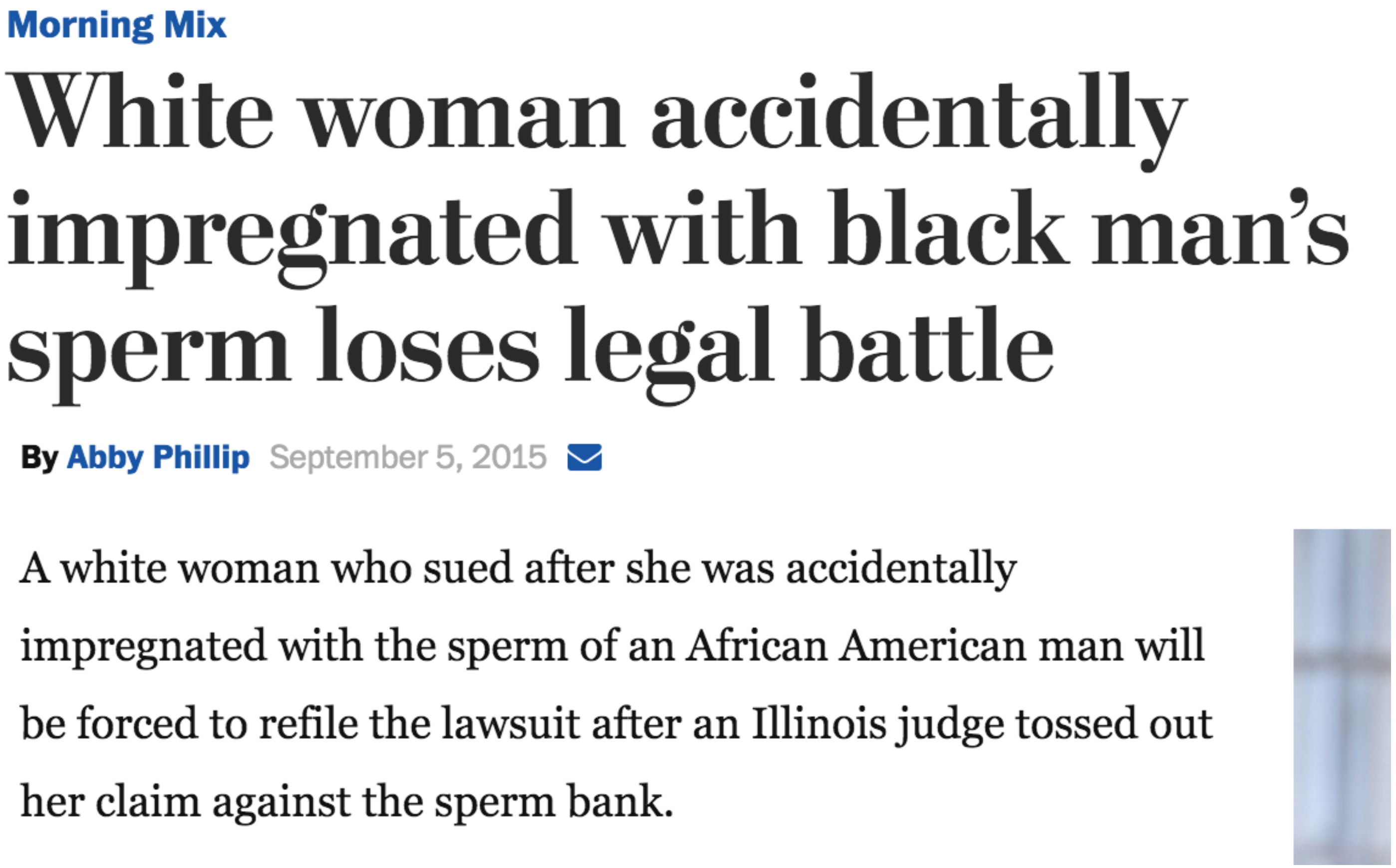
- J. Stamp, *Some Economic Factors in Modern Life* (1929), p. 258

Josiah Stamp,
1st Baron of Stamp
(1880 – 1941)

Assessing the provenance of the data

- Who collected the data, from what sources, when?
- What did the data look like at point-of-collection?
- Is there documentation, metadata, a dictionary, etc.?
- After collection, what was changed, when, and by whom?
- Is there a clear chain of custody in the dataset?
- What quality checks have been run on the data?
- What versions are available?
- Have the collection protocols changed over the course of time?

What went wrong here?



Morning Mix

White woman accidentally impregnated with black man's sperm loses legal battle

By Abby Phillip September 5, 2015

A white woman who sued after she was accidentally impregnated with the sperm of an African American man will be forced to refile the lawsuit after an Illinois judge tossed out her claim against the sperm bank.

'According to the suit, the couple chose sperm from donor No. 380...instead, they were given sperm from donor No. 330...They blame a paper records system that allegedly caused an employee to misread the numbers.'

Databasing Lingo

record collision: two individuals were given the same ID
record duplication: one individual given two IDs

If typos are possible, using serial numbers as IDs will create collisions. Alternatives:

- Cryptographic IDs
- Random IDs

Build Redundancies into Data Entry

Enter critical data multiple times

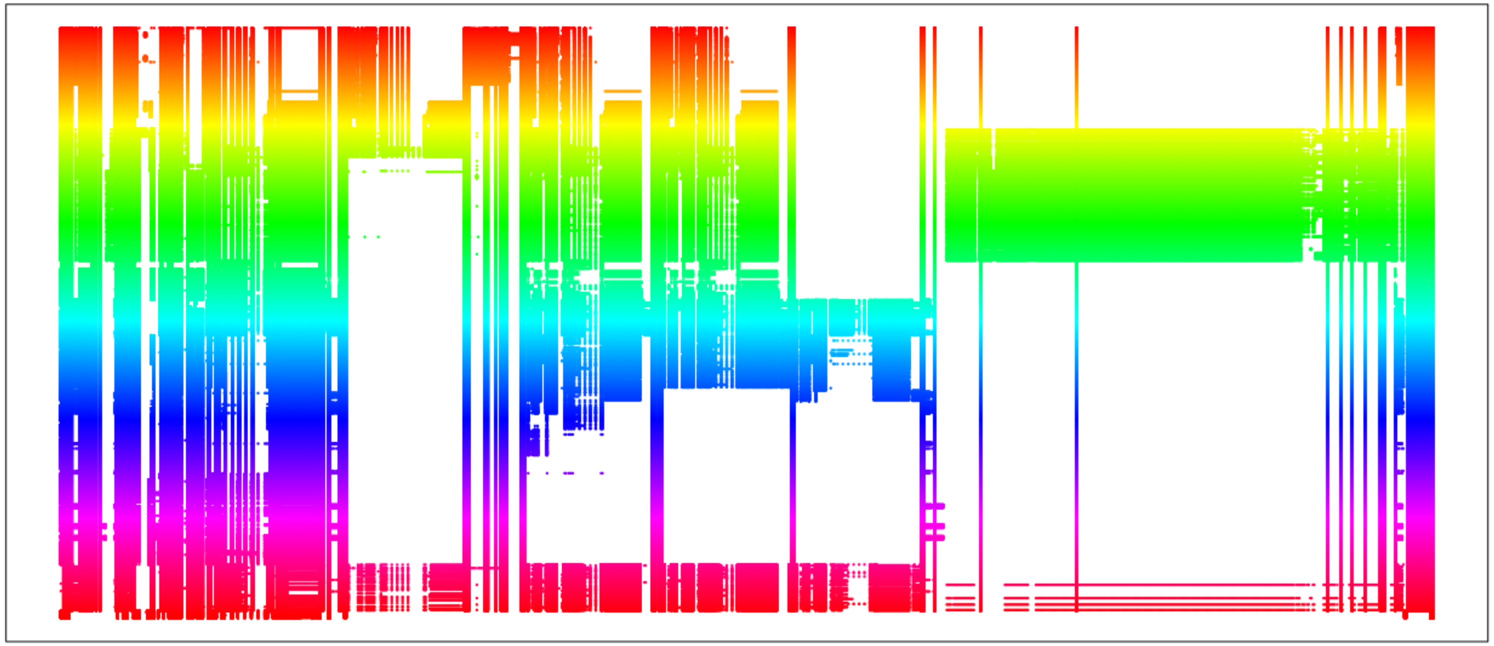
Include redundant information identifiers

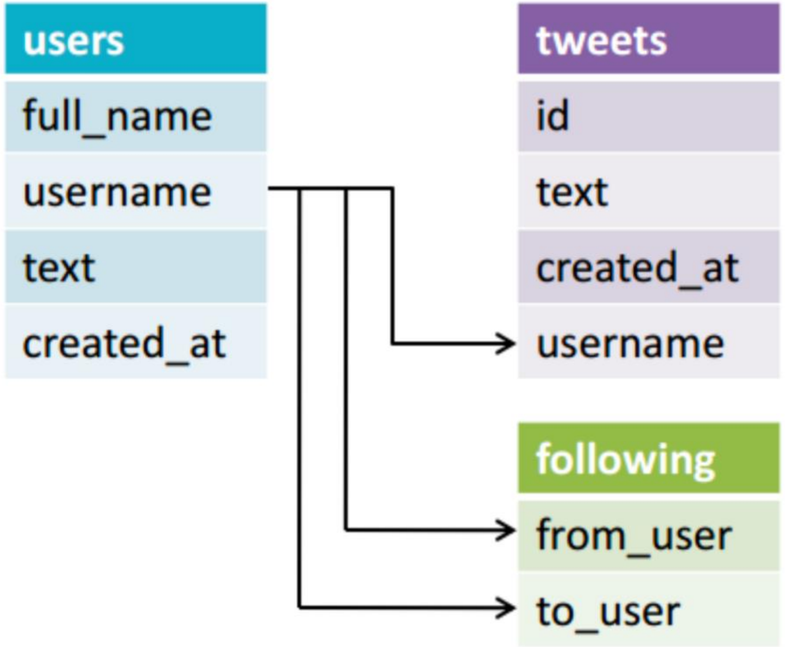
Remove Redundancies from Data Structures

Each value exists in one location

Relational subtables, not more columns

Flat Files vs Relational Tables





users

full_name

username

text

created_at

tweets

id

text

created_at

username

following

from_user

to_user

A common scenario

Researcher A takes raw field data and cleans and codes it for an analysis. He drops some bad cases, creates new variables from existing ones, and imputes missing values to fill in a column.

Some time later, Researcher B starts to do an analysis. Researcher A suggests she start with the “clean” version A produced, rather than go through all the steps of cleaning it again. But she wants to be careful about data provenance, starting as raw as possible.

What should B do?

Distinguishing “Germ-Line” Data Modifications

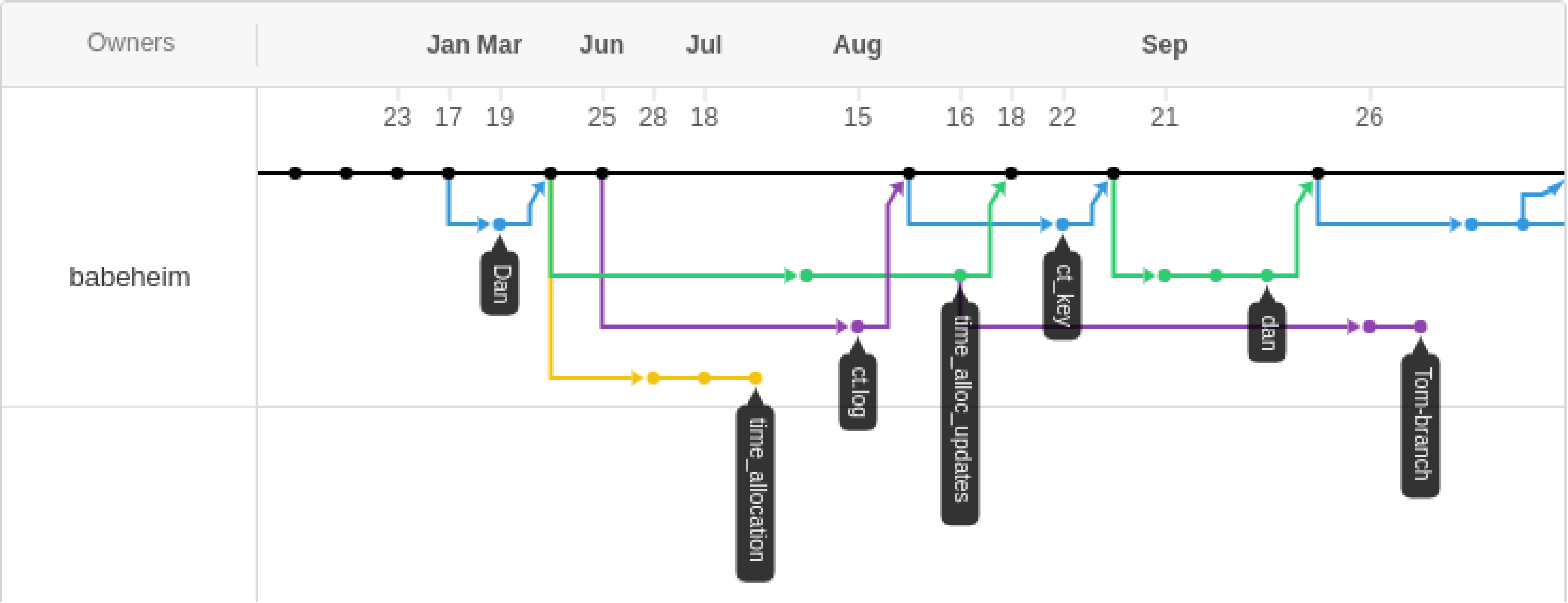
Incidental changes made for an analysis:

- imputed values
- dropping cases
- transformed variables
- merges between relational tables

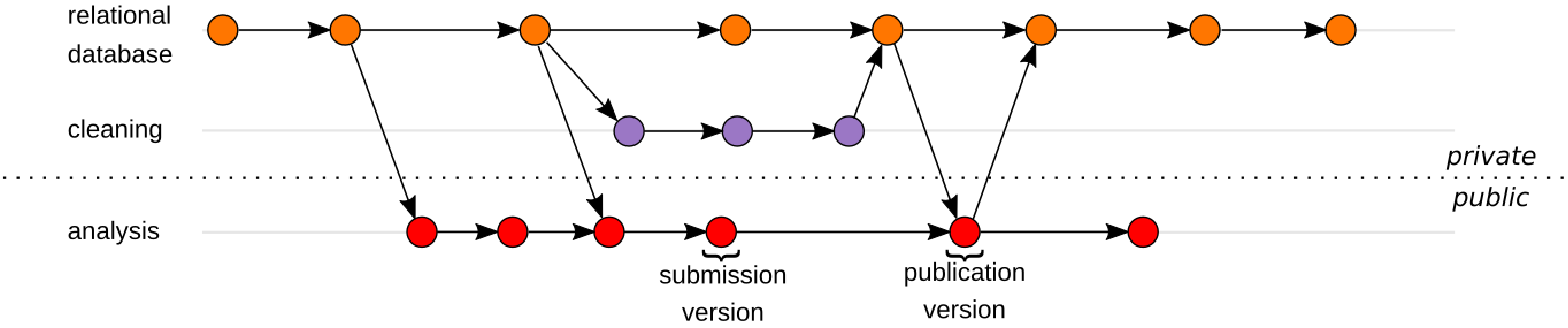
Permanent modifications to standing database:

- new data added
- record linkage corrections
- fixing transcription errors
- Cleaning/standardizing variables

Tsimane' Database on Github

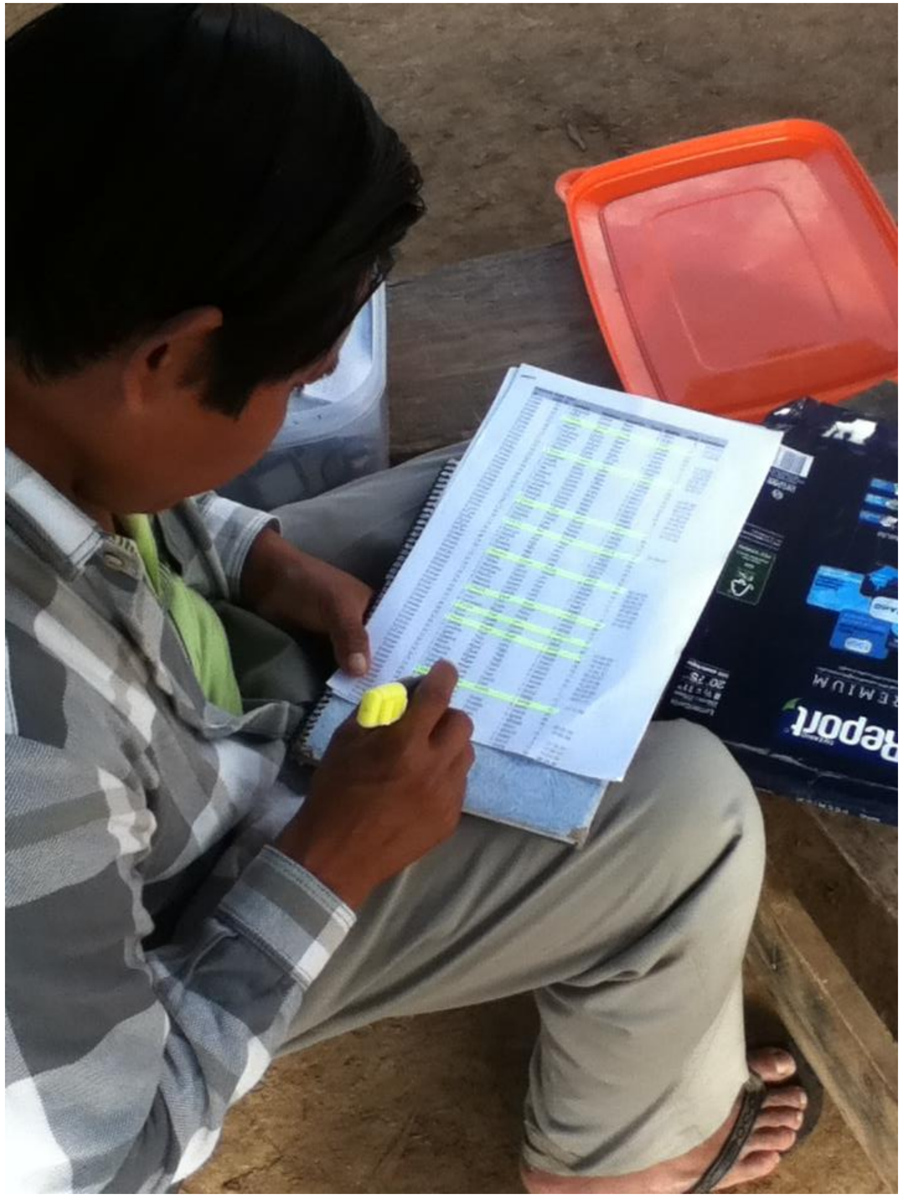


Multiple branches processed simultaneously



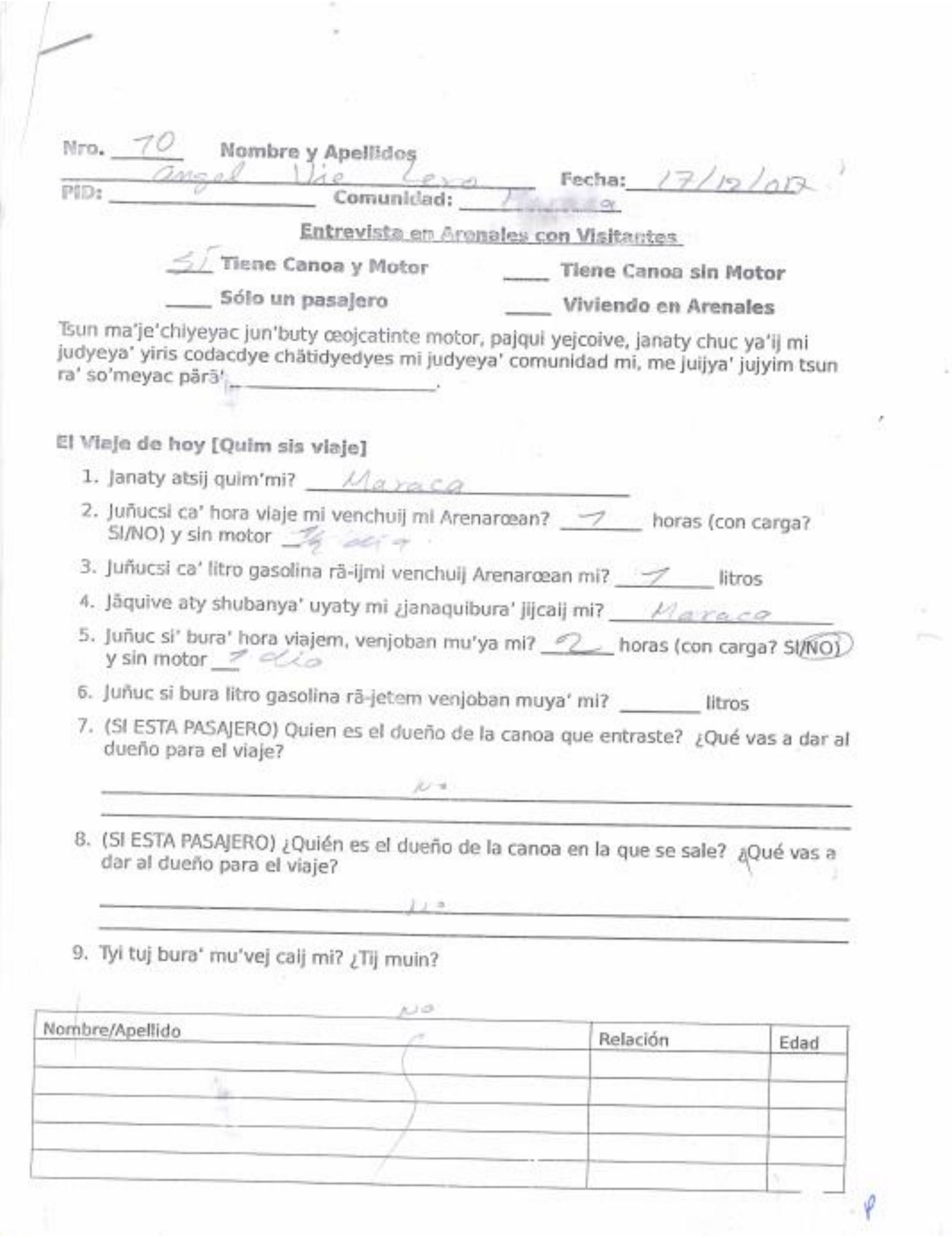
A Data Interchange Pipeline

```
household-interview/
├── metadata/
│    ├── template.yaml
│    └── data dictionary.xlsx
├── code/
│    └── check-yamls.r
├── primary sources/
│    └── audiovideo
└── project-overview.Rmd
```



A Data Interchange Pipeline

```
household-interview/
├── metadata/
|     ├── template.yaml
|     └── data dictionary.xlsx
├── code/
|     └── check-yamls.r
├── primary sources/
|     ├── audiovideo
|     └── pdf
└── project-overview.Rmd
```



Nro. 70 Nombre y Apellidos
Angel Vie Lera Fecha: 17/12/017
PID: ______ Comunidad: [redacted]

Entrevista en Arenales con Visitantes

SI Tiene Canoa y Motor ____ Tiene Canoa sin Motor
____ Sólo un pasajero ____ Viviendo en Arenales

Tsun ma'je'chiyeyac jun'buty œojcatinte motor, pajqui yejcoive, janaty chuc ya'ij mi judyeya' yiris codacdye chätidyedyes mi judyeya' comunidad mi, me juijya' jujyim tsun ra' so'meyac pärä' ________.

El Viaje de hoy [Quim sis viaje]

1. Janaty atsij quim'mi? Maraca
2. Juñucsi ca' hora viaje mi venchuij mi Arenarœan? 7 horas (con carga? SI/NO) y sin motor ½ día
3. Juñucsi ca' litro gasolina rã-ijmi venchuij Arenarœan mi? 7 litros
4. Jãquive aty shubanya' uyaty mi ¿janaquibura' jijcaij mi? Maraca
5. Juñuc si' bura' hora viajem, venjoban mu'ya mi? 2 horas (con carga? SI/[✓ NO]) y sin motor 1 día
6. Juñuc si bura litro gasolina rã-jetem venjoban muya' mi? ____ litros
7. (SI ESTA PASAJERO) Quien es el dueño de la canoa que entraste? ¿Qué vas a dar al dueño para el viaje?
   no
8. (SI ESTA PASAJERO) ¿Quién es el dueño de la canoa en la que se sale? ¿Qué vas a dar al dueño para el viaje?
   no
9. Tyi tuj bura' mu'vej caij mi? ¿Tij muin?
   no

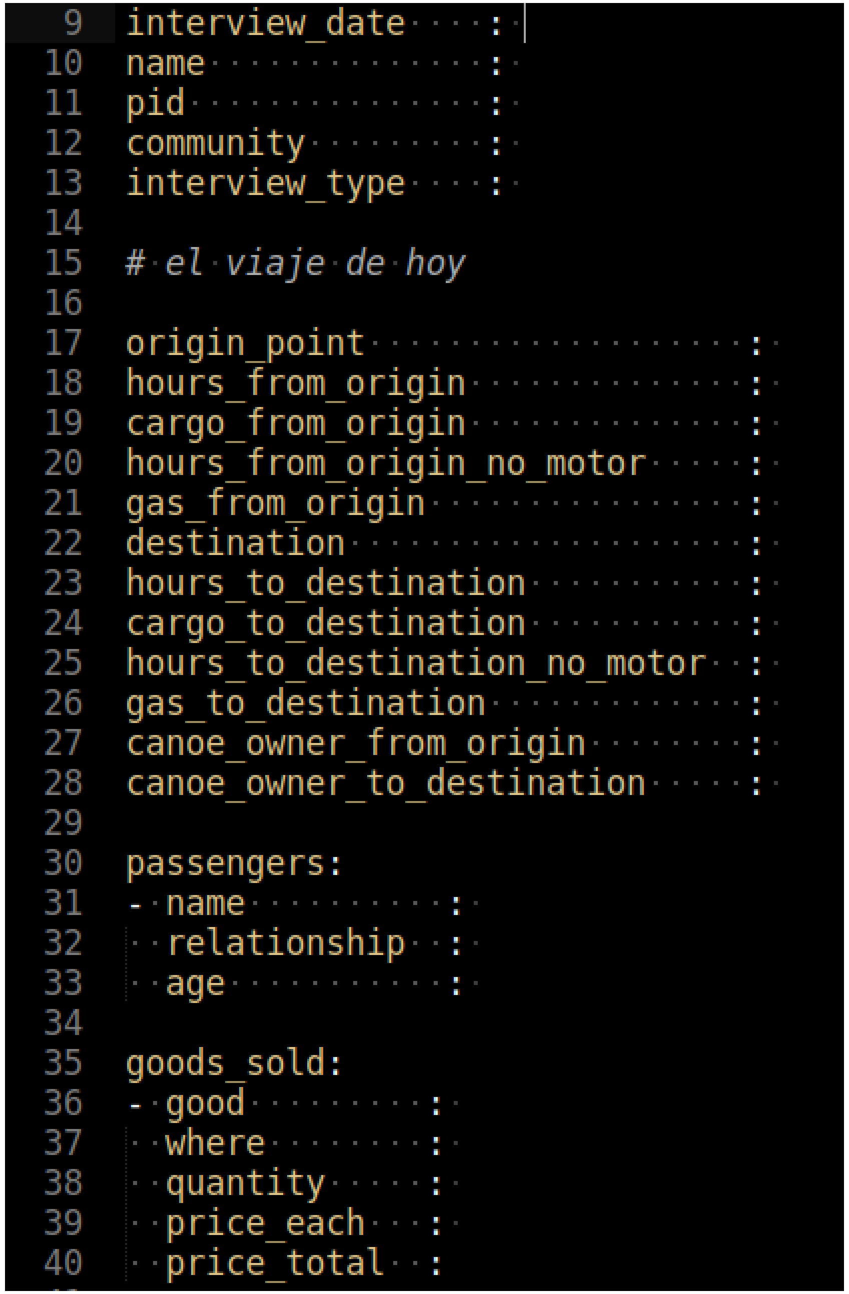
Nombre/Apellido	Relación	Edad

A Data Interchange Pipeline

```
household-interview/
├── metadata/
|     ├── template.yaml
|     └── data dictionary.xlsx
├── code/
|     └── check-yamls.r
├── primary sources/
|     ├── audiovideo
|     |     └── completed
|     └── pdf
└── project-overview.Rmd
```

A Data Interchange Pipeline

```
household-interview/
├── metadata/
│     ├── template.yaml
│     └── data dictionary.xlsx
├── code/
│     └── check-yamls.r
├── primary sources/
│     ├── audiovideo
│     │     └── completed
│     ├── pdf
│     └── yaml
└── project-overview.Rmd
```



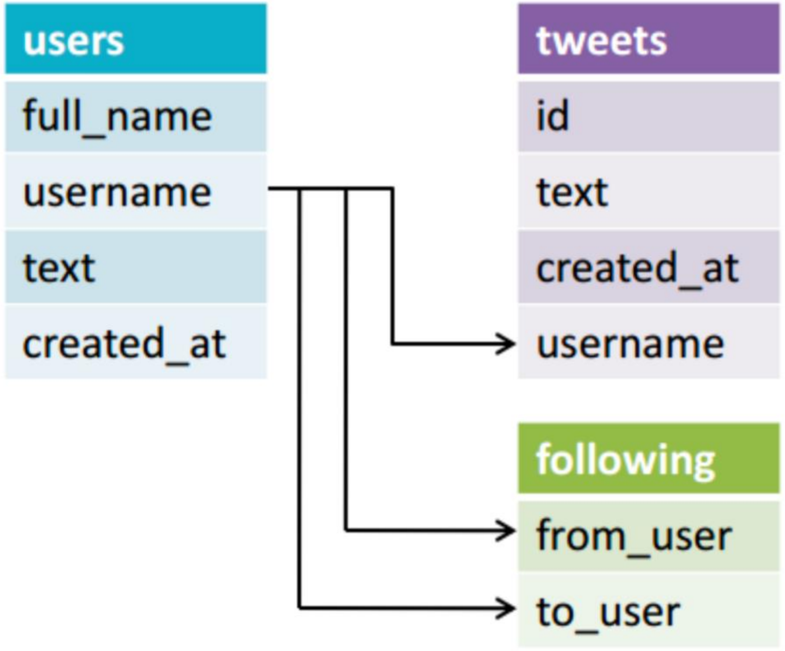
```
interview_date    : 
name              : 
pid               : 
community         : 
interview_type    : 

# el viaje de hoy

origin_point                   : 
hours_from_origin              : 
cargo_from_origin              : 
hours_from_origin_no_motor     : 
gas_from_origin                : 
destination                    : 
hours_to_destination           : 
cargo_to_destination           : 
hours_to_destination_no_motor  : 
gas_to_destination             : 
canoe_owner_from_origin        : 
canoe_owner_to_destination     : 

passengers:
- name          : 
  relationship  : 
  age           : 

goods_sold:
- good         : 
  where        : 
  quantity     : 
  price_each   : 
  price_total  : 
```

A Data Interchange Pipeline

```
household-interview/
├── metadata/
|    ├── template.yaml
|    └── data dictionary.xlsx
├── code/
|    └── check-yamls.r
├── primary sources/
|    ├── audiovideo
|    |    └── completed
|    ├── pdf
|    |    └── completed
|    └── yaml
└── project-overview.Rmd
```

A Data Interchange Pipeline

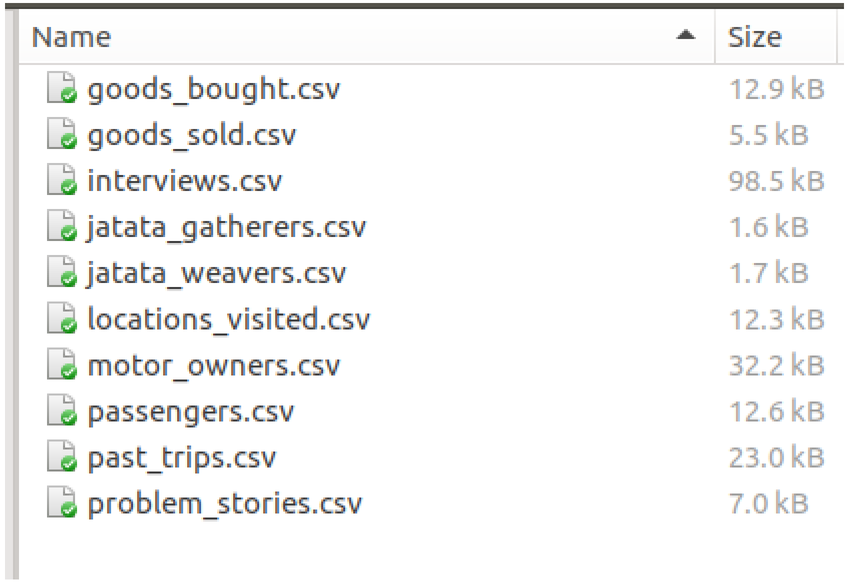
```
household-interview/
├── metadata/
|     ├── template.yaml
|     └── data dictionary.xlsx
├── code/
|     ├── check-yamls.r
|     └── scrape-yamls.r
├── primary sources/
|     ├── audiovideo
|     |     └── completed
|     ├── pdf
|     |     └── completed
|     └── yaml
├── data tables/
|     ├── interviews.csv
|     ├── household-wealth.csv
|     └── family-members.csv
└── project-overview.Rmd
```



users
full_name
username
text
created_at

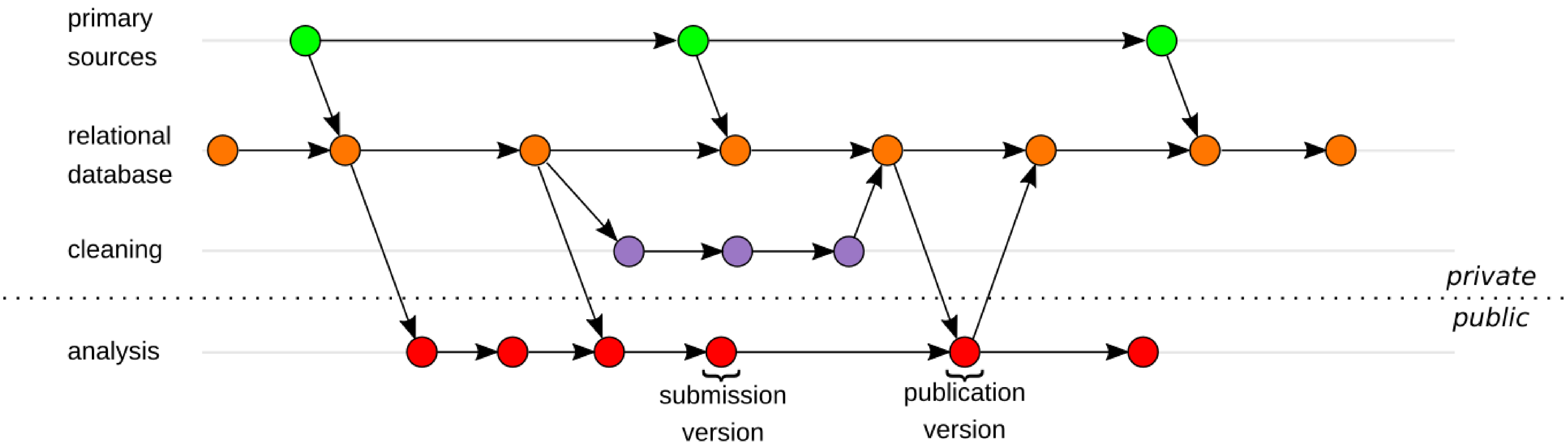
tweets
id
text
created_at
username

following
from_user
to_user



Name	Size
goods_bought.csv	12.9 kB
goods_sold.csv	5.5 kB
interviews.csv	98.5 kB
jatata_gatherers.csv	1.6 kB
jatata_weavers.csv	1.7 kB
locations_visited.csv	12.3 kB
motor_owners.csv	32.2 kB
passengers.csv	12.6 kB
past_trips.csv	23.0 kB
problem_stories.csv	7.0 kB

Versioning the primary sources



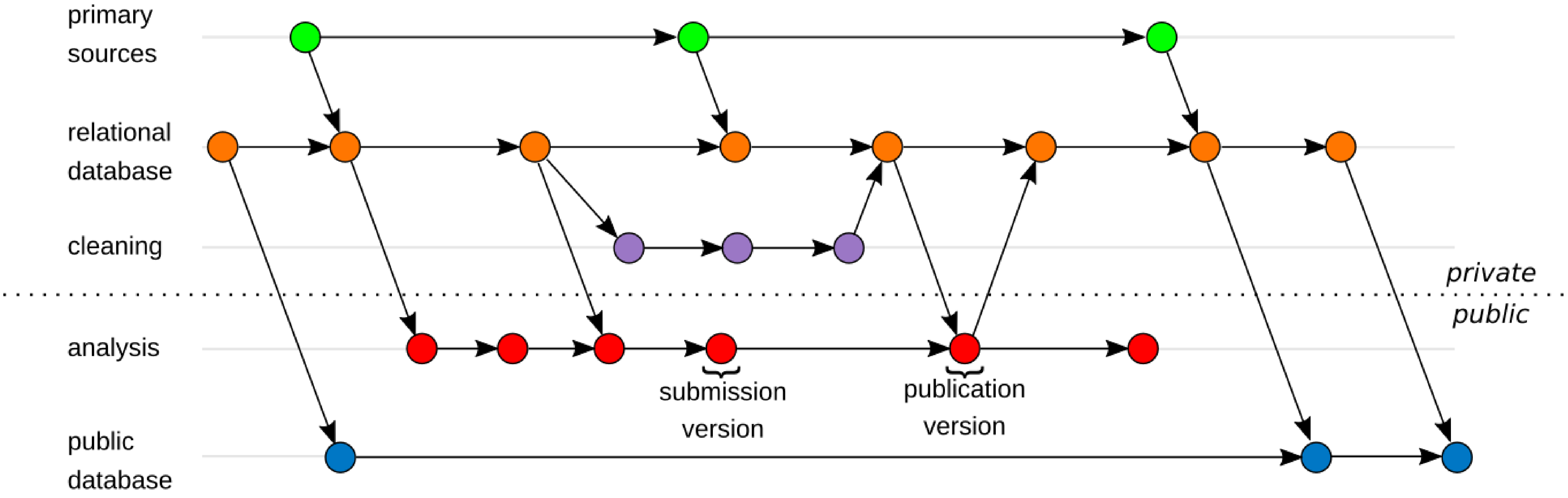
Improving reliability of field datasets

Data decay: version control on data

Provenance problems:

anticipate points of failure,
use redundancy and fail-safes in data entry,
minimize redundancy by “tidy” relational tables,
germ-line vs analysis-only data changes,
script as much as possible

Adding a public database


primary sources
relational database
cleaning
private
public
analysis
submission version
publication version
public database