Max-Planck-Institut für Bildungsforschung
Max Planck Institute for Human Development

# Automatic Codebooks from Existing Metadata

Ruben C. Arslan
Center for Adaptive Rationality
MPI for Human Development

https://github.com/rubenarslan/codebook

ruben.arslan@gmail.com
@rubenarslan

survey software: formr.org
blog: http://the100.ci

# My data sharing pain points

- I did a lot of work with data that I could not share
  - Minnesota Twin Family Study (always identifiable by the co-twin, not mine to share)
  - Swedish population data (not mine to share)
  - Online sex diary data (~impossible to de-identify)
  - Clue data (not mine to share, too large to be useful for casual users)
- Still want to be as transparent as possible.

# My metadata background

- Before my PhD work, I collaborated on the Archival Project
  an ambitious attempt to collect data on statistical tests in studies through crowdsourcing which failed

- My survey software, formr.org, comes with an R package that exports metadata along with data, but which is underused by our users

- 2017-19 I wrote an R package called *codebook* to entice people to make better use of metadata, by automating tedious tasks for them

- Part of the PsychDS working group

https://github.com/psych-ds/psych-DS

# Why share data at all?

- *Nullius in verba*
  - motto of the Royal Society
  People may no longer trust you unless you share

- Others may derive new insights from your data that you did not think of

- Many have more data than they can ever publish

- Many funders now require it (e.g. NIH, ERC, Wellcome trust, Schweizer Nationalfond, DFG)

- Ensuring the best use of hard-won data is responsible

- Tools may add value to your data in the future, if they can work with it (see psychDS)

# Why share data at all?

**Journals that already require open data (or a justification why it is not possible):**

- Advances in Methods and Practices in Psychological Science (AMPPS)
- Archives of Scientific Psychology
- BMC Psychology
- Collabra: Psychology
- Cognition
- Comprehensive Results in Social Psychology
- European Journal of Personality (EJP)
- European Journal of Social Psychology (EJSP)
- Evolution and Human Behavior
- Experimental Psychology
- Journal of Economic Psychology
- Journal of Open Psychology Data (JOPD)
- Journal of Research in Personality
- Judgment and Decision Making
- Journal of Cognition
- Meta-Psychology
- PLOS ONE
- Royal Society Open Science
- Science

* for a continuously updated list, see https://osf.io/tbkzh/wiki/Psychology%20journals%20with%20mandatory%20open%20data/

**Table 2.** Data-Sharing Policies of Several Funding Organizations

| Funder | Data-sharing policy |
|---|---|
| German Research Foundation | "The German Research Foundation (DFG), the largest public funder of research in Germany, updated their policy on data sharing, which can be summarized in a single sentence: **Publicly funded research, including the raw data, belongs to the public.** Consequently, all research data from a DFG funded project should be made open immediately, or at least a couple of months after finalization of the research project. . . . Furthermore, the DFG asked all scientific disciplines to develop more specific guidelines which implement these principles in their respective discipline" (Schönbrodt, 2017, paragraph 3). |
| National Institutes of Health | "The *2003 NIH Data Sharing Policy* encourages NIH-funded researchers to share their final research data for use by other researchers in a timely way (i.e., no later than the acceptance for publication of the main findings from the final data set). The Policy expects applicants requesting $500,000 or more in direct costs in funding from NIH for research for any one year to include a data sharing plan or state why data sharing is not possible. Supplemental guidance materials suggest that plans should describe |

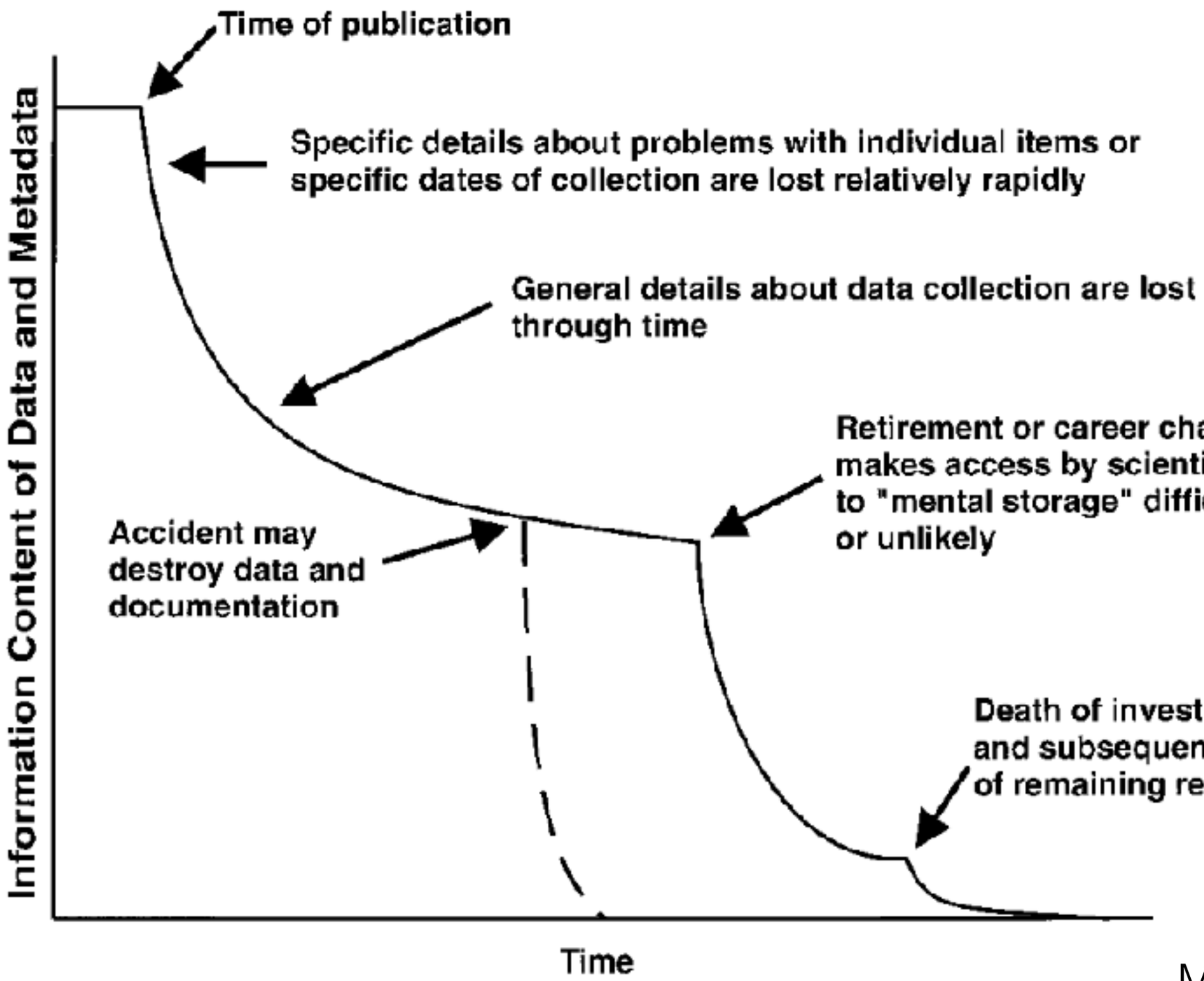
Houtkoop et al. 2018

**Table 1.** Data-Sharing Guidelines of Select Journals With a Clearly Articulated Data-Sharing Policy

| Journal or publisher | Data-sharing policy |
|---|---|
| *Nature* | "Supporting data must be made available to editors and peer reviewers at the time of submission for the purposes of evaluating the manuscript. All manuscripts reporting original research published in Nature Research journals must include a data availability statement. . . ." (*Nature*, 2017, Availability of Data, paragraph 1). |
| PLOS | "PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.<br>"When submitting a manuscript online, authors must provide a *Data Availability Statement* describing compliance with PLOS's policy. If the article is accepted for publication, the data availability statement will be published as part of the final article.<br>"Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection. PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data from articles published in PLOS journals. If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication" (PLOS, n.d., paragraphs 1–3). |
| The Royal Society | "To allow others to verify and build on the work published in Royal Society journals, it is a condition of publication that authors make available the data, code and research materials supporting the results in the article.<br>"Datasets and code should be deposited in an appropriate, recognised, publicly available repository. . . .<br>"Exceptions to the sharing of data, code and materials may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species. Authors must disclose upon submission of the manuscript any restrictions on the availability of data, code and research materials" (The Royal Society, 2017, Open Data Policy). |
| *Science* | "After publication, all data and materials necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*. . . . After publication, all reasonable requests for data or materials must be fulfilled. Any restrictions on the availability of data, codes, or materials, including fees and restrictions on original data obtained from other sources must be disclosed to the editors. . . . Unreasonable restrictions on data or material availability may preclude publication" (*Science*, 2017, Data and Materials Availability After Publication). |

Houtkoop et al. 2018

# Information entropy




Michener et al. 1997

# Why not share data?

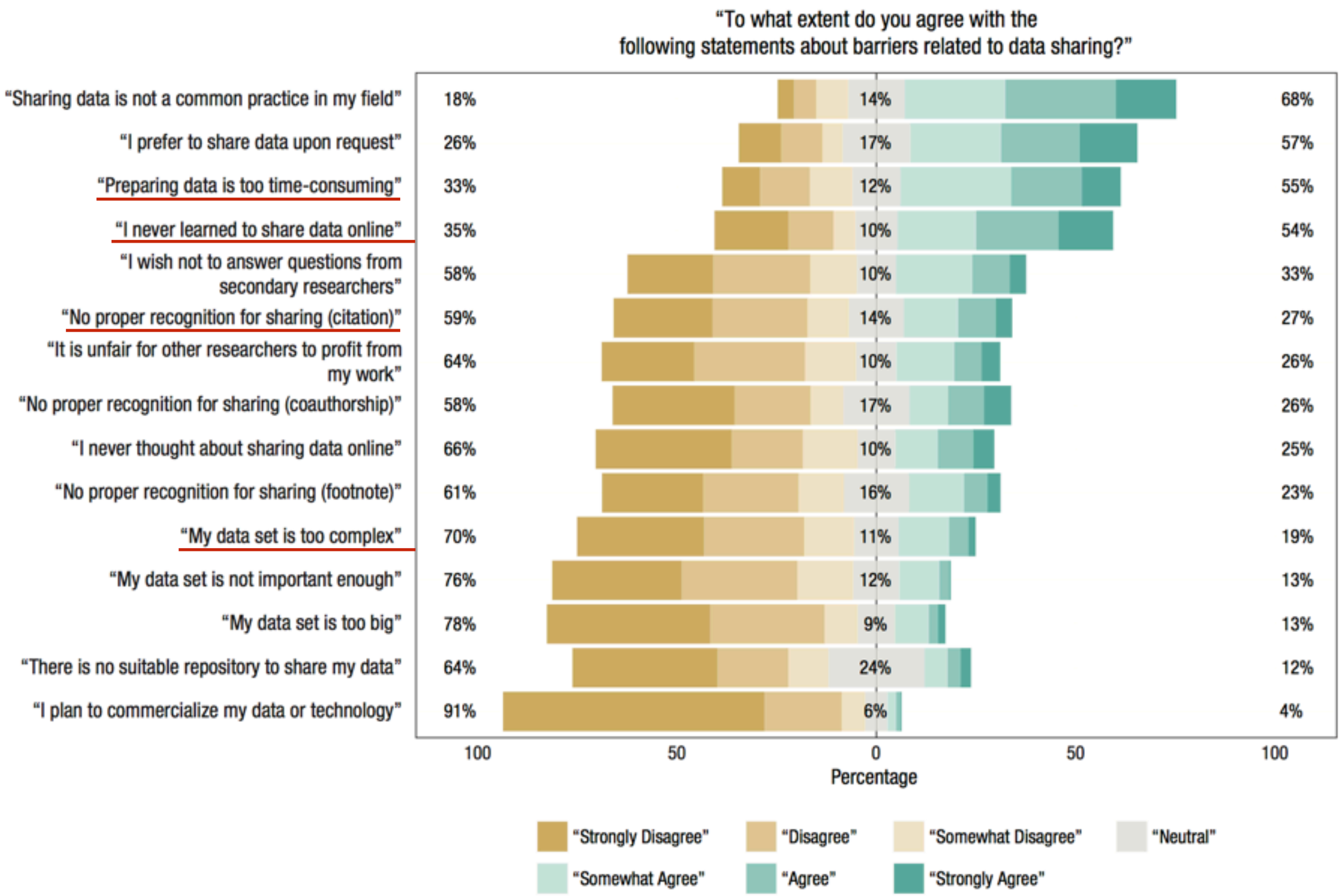
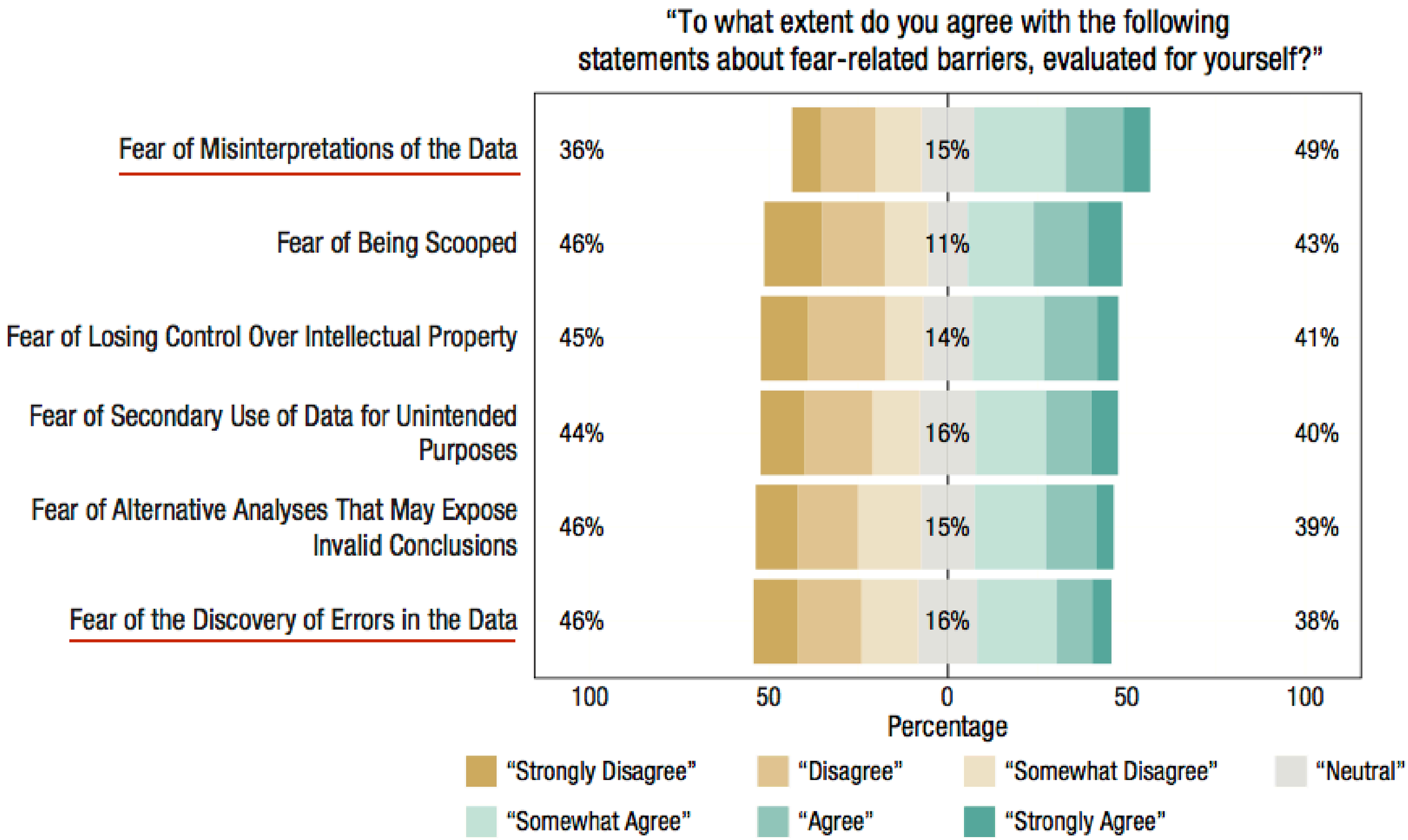



**Fig. 2.** Responses to the survey questions asking respondents to indicate the extent to which the 15 non-fear-related barriers kept them from sharing their research data. For each statement, the number to the left of the data bar indicates the percentage of researchers who responded with "strongly disagree," "disagree," or "somewhat disagree"; the number in the center of the data bar indicates the percentage of researchers who responded with "neutral"; and the number to the right of the data bar indicates the percentage who responded with "somewhat agree," "agree," or "strongly agree." The statements are ordered according to the percentage of agreement (greatest agreement at the top). This figure was created using the *likert* package in R (Bryer & Speerschneider, 2015).

Houtkoop et al. 2018, see also Tenopir, 2011

a




h

Houtkoop et al. 2018

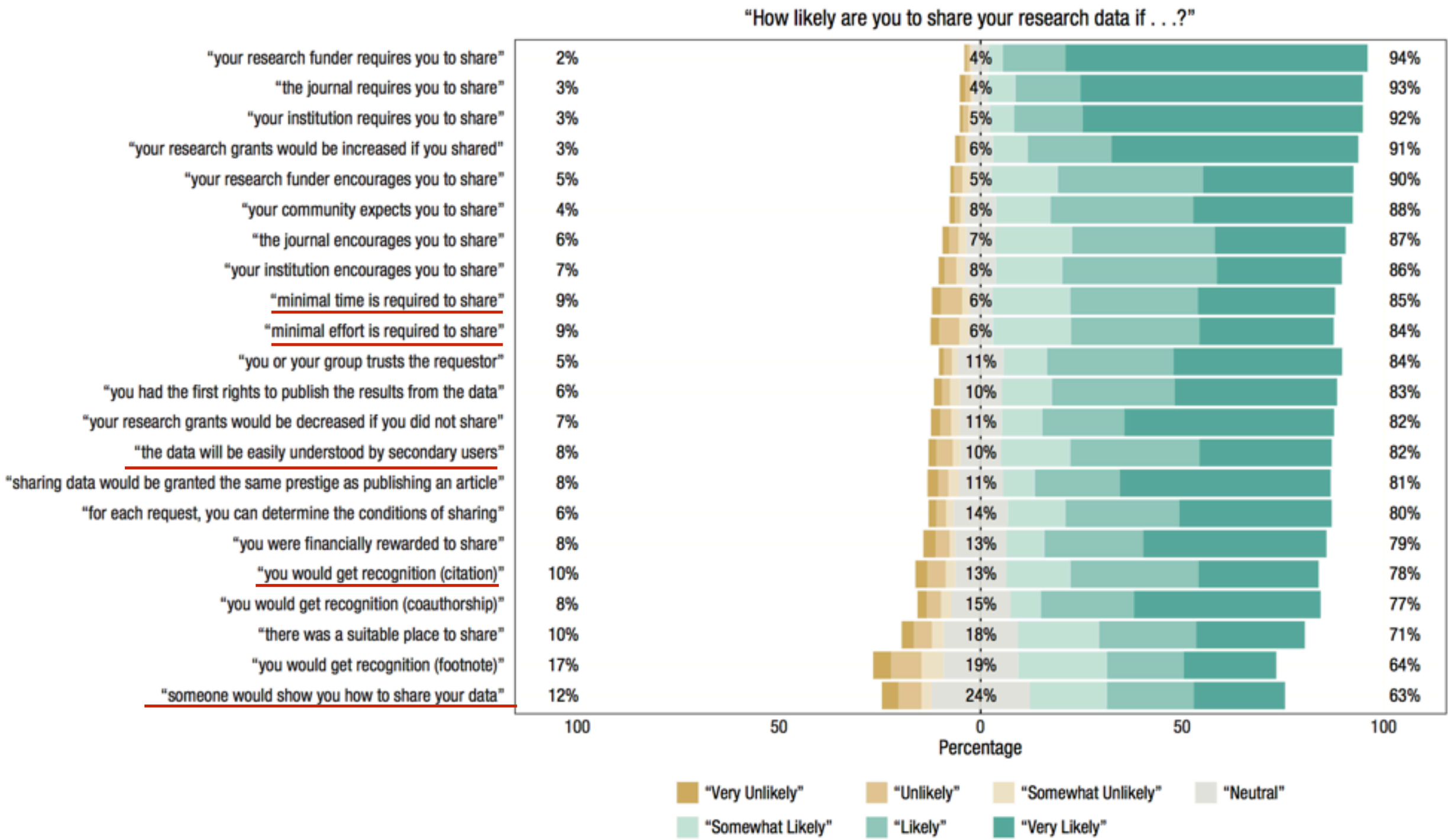
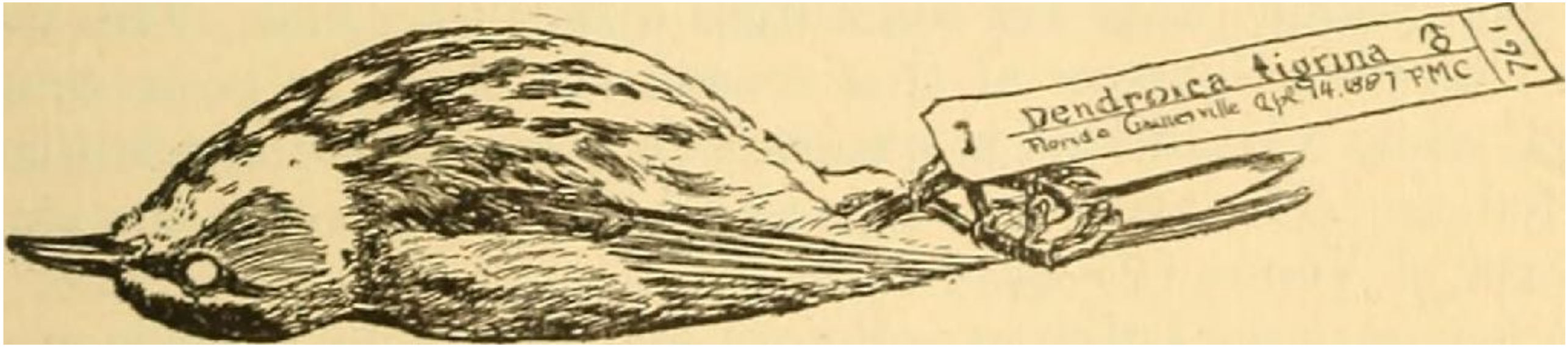



**Fig. 4.** Responses to the survey questions asking researchers to indicate how likely they would be to share their data under several conditions. For each statement, the number to the left of the data bar indicates the percentage of researchers who responded with "very unlikely," "unlikely," or "somewhat unlikely"; the number in the center of the data bar indicates the percentage who responded with "neutral"; and the number to the right of the data bar indicates the percentage who responded with "somewhat likely," "likely," or "very likely." The statements are ordered according to the percentage of agreement (greatest agreement at the top). This figure was created using the *likert* package in R (Bryer & Speerschneider, 2015).

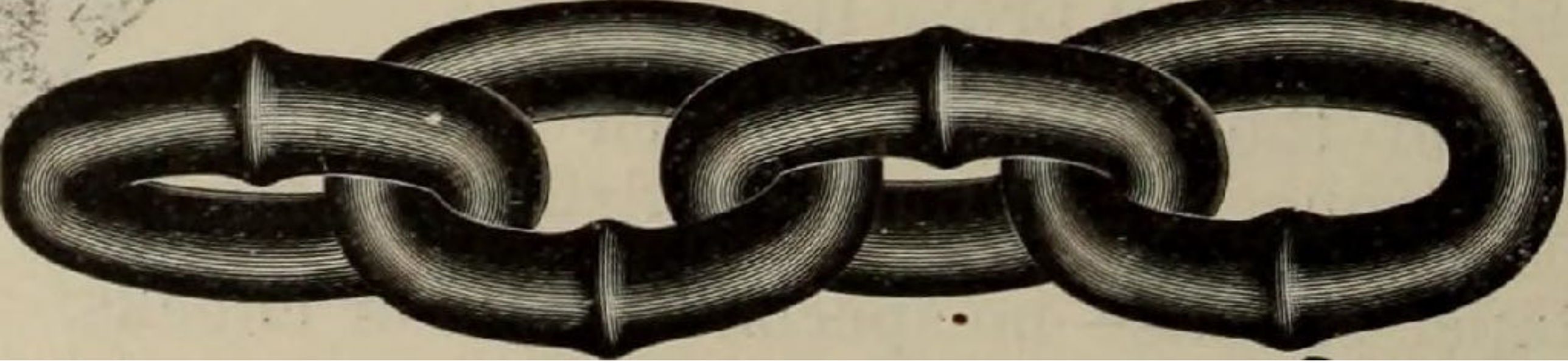
Houtkoop et al. 2018

# So I just share?



# So I just share my data?

- Not so fast



# Sensitive data

- data on sexual life, political preferences, crimes, physical or mental health, racial or ethnic origin, union or party membership

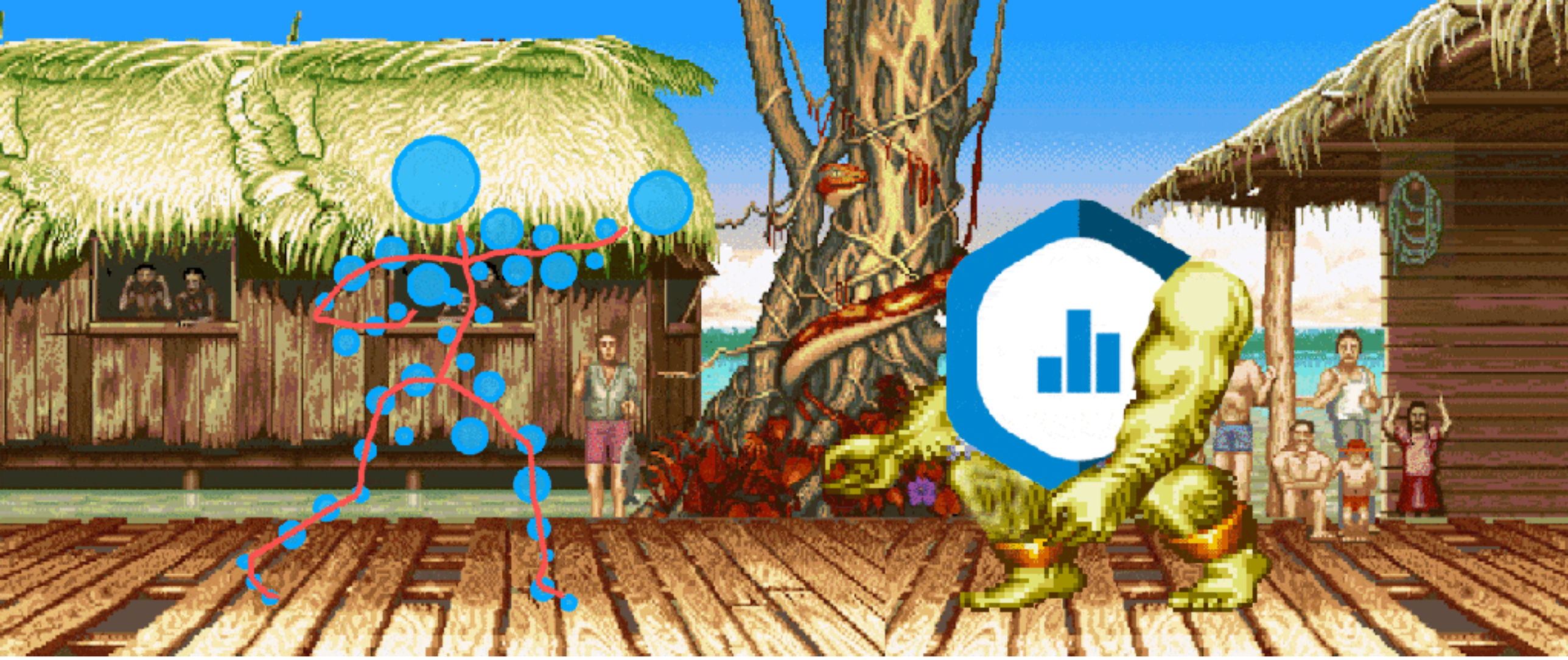
- Anything not on this list that you collect but still consider sensitive?

# Barriers to sharing it all

- You cannot always share all the data
  - No consent
  - Re-identifiability concerns + sensitive data
  - No permission from co-authors, data owners

# Barriers to sharing it all

- Sometimes sharing the raw data is not that useful
  - Format has to be usable
  - Dataset might be too big for most users
  - Specific formats require expertise
  - Many questions that people might have could be easily answered if you went the extra mile



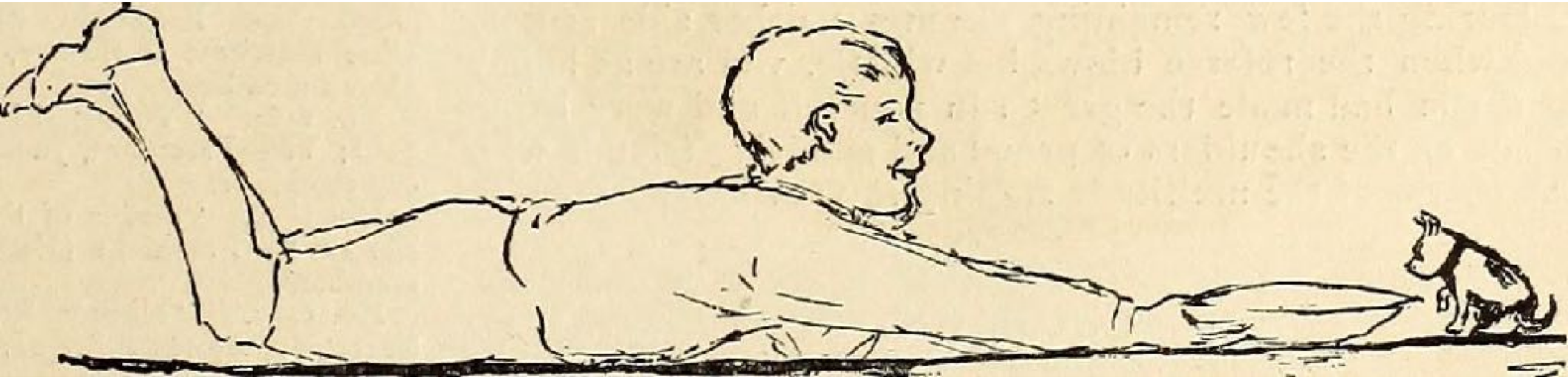
# Barriers to sharing it all

- Sometimes sharing publicly will even make data *less* useful
  - combined with publication bias and the garden of forking paths, rat races may lead to more erroneous results being published (first)

http://www.the100.ci/2017/09/14/overfitting-vs-open-data/

# So, share something useful

- Conversely, being unable to share raw data does not mean you cannot share anything useful.
- Field-specific summaries can sometimes be very useful
  - GWAS associations per SNP (LD Hub)
  - Correlation matrices in psychometrics
  - …
- You can almost always share **metadata**



# So I share metadata?

- There’s different levels of usefulness for metadata too
- documenting
  - the citation metadata (authorship, location, dates)
  - the study structure
  - the survey items, stimuli, programs
  - the collected data

# Documenting data

- Do you share:
  - an SPSS file with variable names like FSC1V2, code2, sex2 **vs.** a properly labelled and documented dataset in an open format like CSV, JSON, or xlsx
  - an upload on your department website **vs.** in a repository?
  - in a way that lets it be indexed and found through search engines **vs.** so that people need to know where too look?

# The dreaded departmental website

**Not Found**

The requested URL /people/directory-profiles/ data-sets/ovulation-1.sav was not found on this server.

Additionally, a 404 Not Found error was encountered while trying to use an ErrorDocument to handle the request.

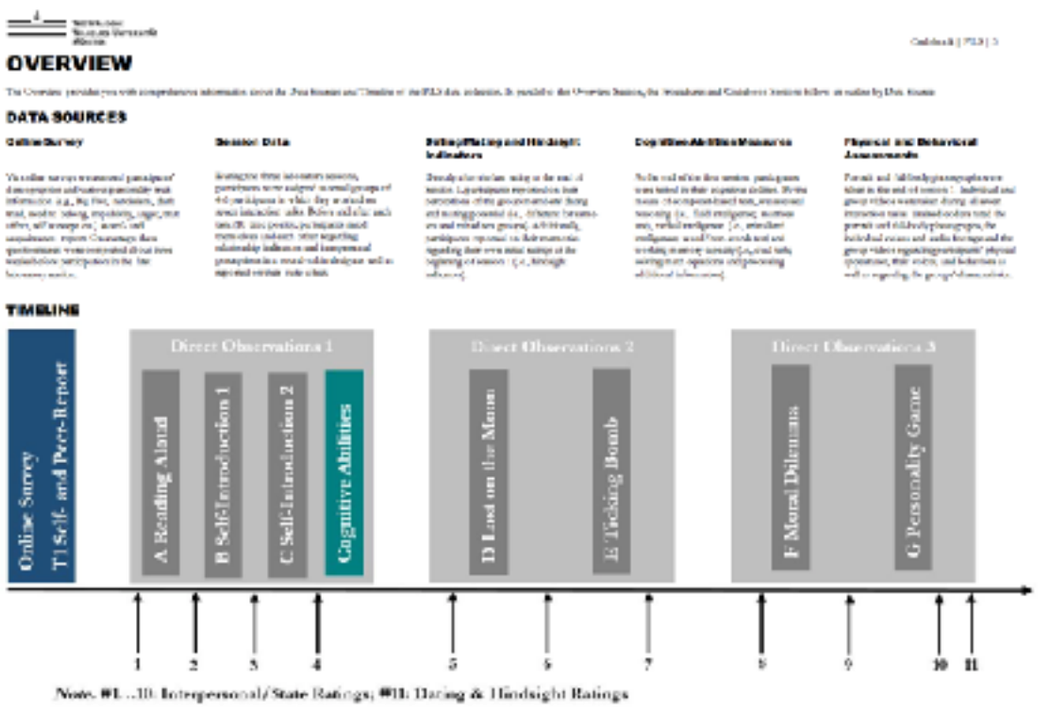
# Respectable README

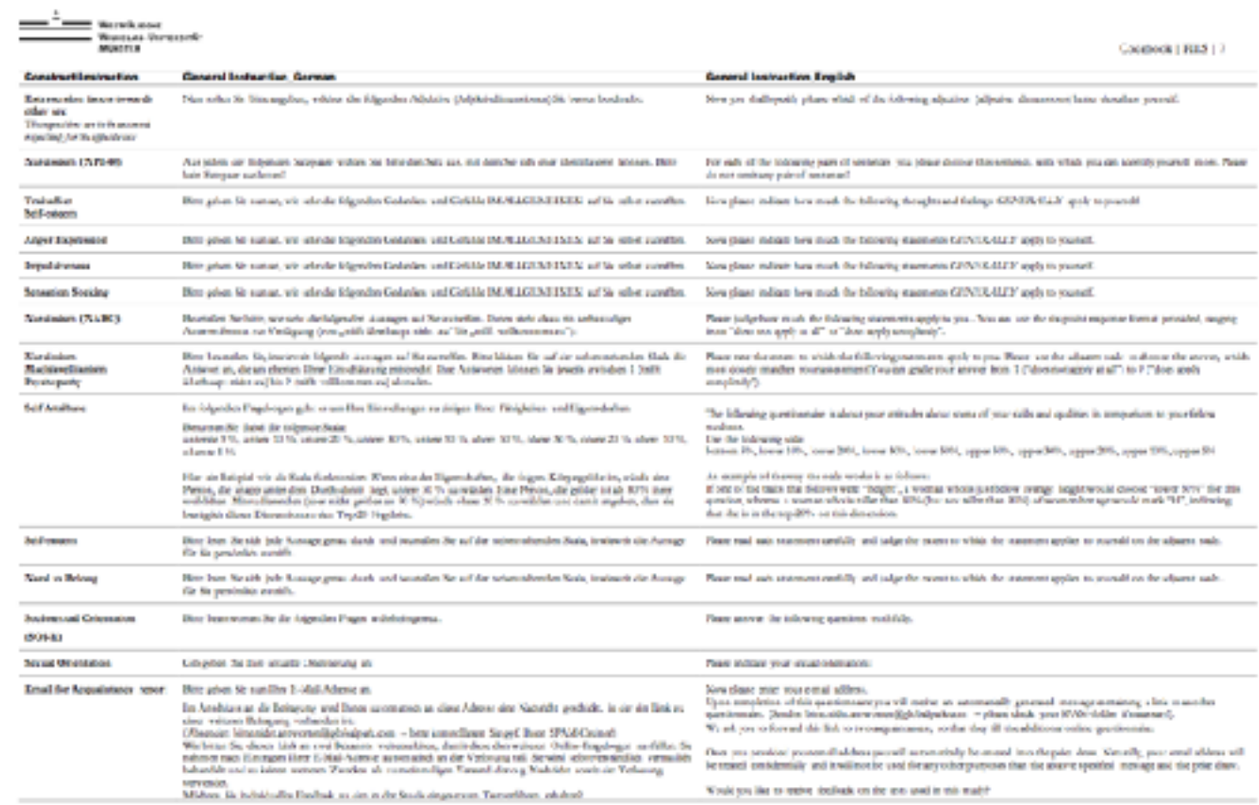
```
Variables are labeled in SPSS. Here is a list of important abbreviations, prefixes and suffixes:

_acq = acquaintance (i.e., variables with this suffix are controlled for prior acquaintance)
_avg = average
_rat = rating variable
_z = z-standardized score
BC = booty call
DG = dating group (three groups in this study)
FIPI = five item personality inventory
FS = friendship
FWB = friends-with-benefits
Int = Intelligence
Like = Likeability
```

- No information on question wording, order of questions, etc.

https://osf.io/pqwm2/

# Pretty PDF




- Useful for humans, but difficult to parse for machines
  -> will not be indexed in search engines

https://osf.io/wtb76/

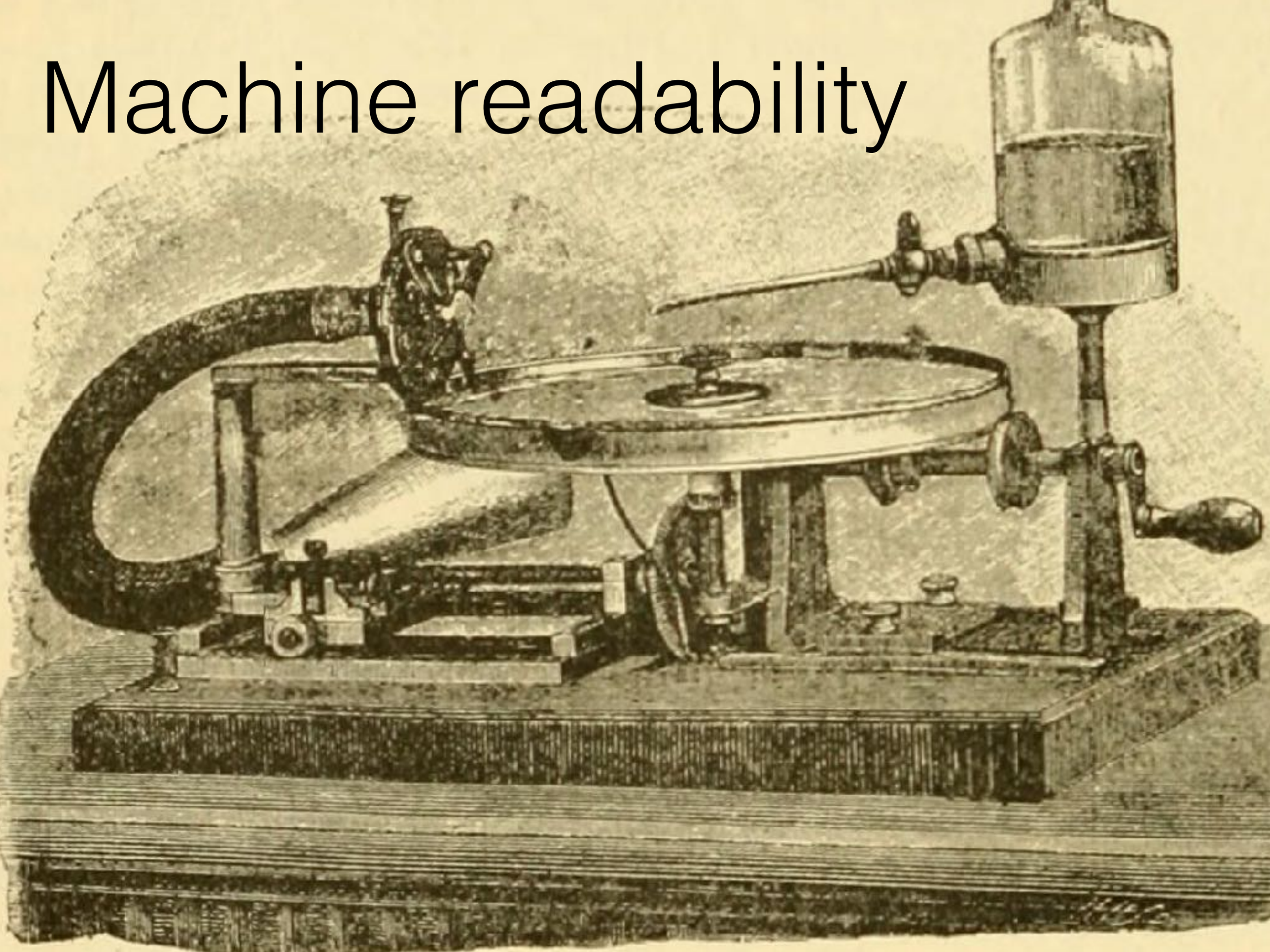
# LengthPriorRepAV indeed

| | | | | | | |
|---|---|---|---|---|---|---|
| NextCycle | Numeric | 8 | 2 | Next cycle onset reported | {.00, No}... | None |
| LHResult | Numeric | 8 | 0 | | {0, Positive}... | None |
| LengthTestingCycle | Numeric | 8 | 2 | | None | None |
| LengthRep | Numeric | 12 | 0 | | None | None |
| LengthPrior | Numeric | 8 | 2 | | None | None |
| LengthPriorRepAV | Numeric | 8 | 2 | | None | None |
| LengthallMonthsRepAV | Numeric | 8 | 2 | | None | None |
| FCDay | Numeric | 8 | 0 | FC surge day | None | None |
| BCActual | Numeric | 8 | 2 | BC surge day (actual) | None | 999.00 |
| BCRep | Numeric | 8 | 0 | BC surge day (reported) | None | None |
| BCPrior | Numeric | 8 | 0 | BC surge day (prior) | None | None |
| BCPriorRepAV | Numeric | 8 | 0 | BC surge day (prior rep av) | None | None |
| BCallMonthsRepAV | Numeric | 8 | 0 | BC surge day (all months) | None | None |
| AccFC | Numeric | 8 | 2 | Accuracy within 2 days (FC) | None | None |
| AccRep | Numeric | 8 | 2 | Accuracy within 2 days (BC rep) | None | None |

# Documenting studies

- Ideally, you enable others to reproduce your entire study with minimal effort.
  - harder if you use proprietary software
  - many software packages don’t export the whole study package
  - description in papers often insufficient as controversies around “direct” replications show

# Machine readability



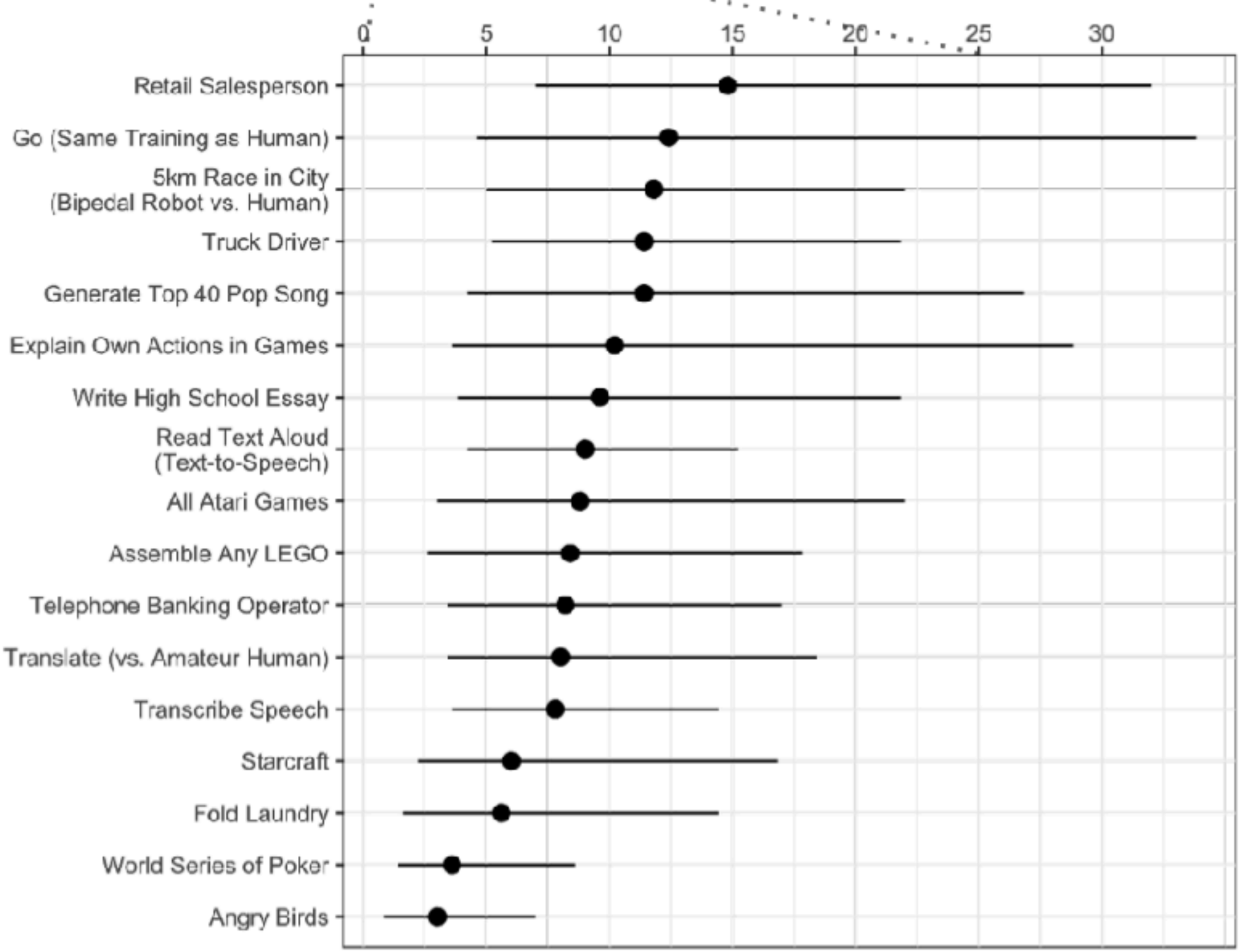
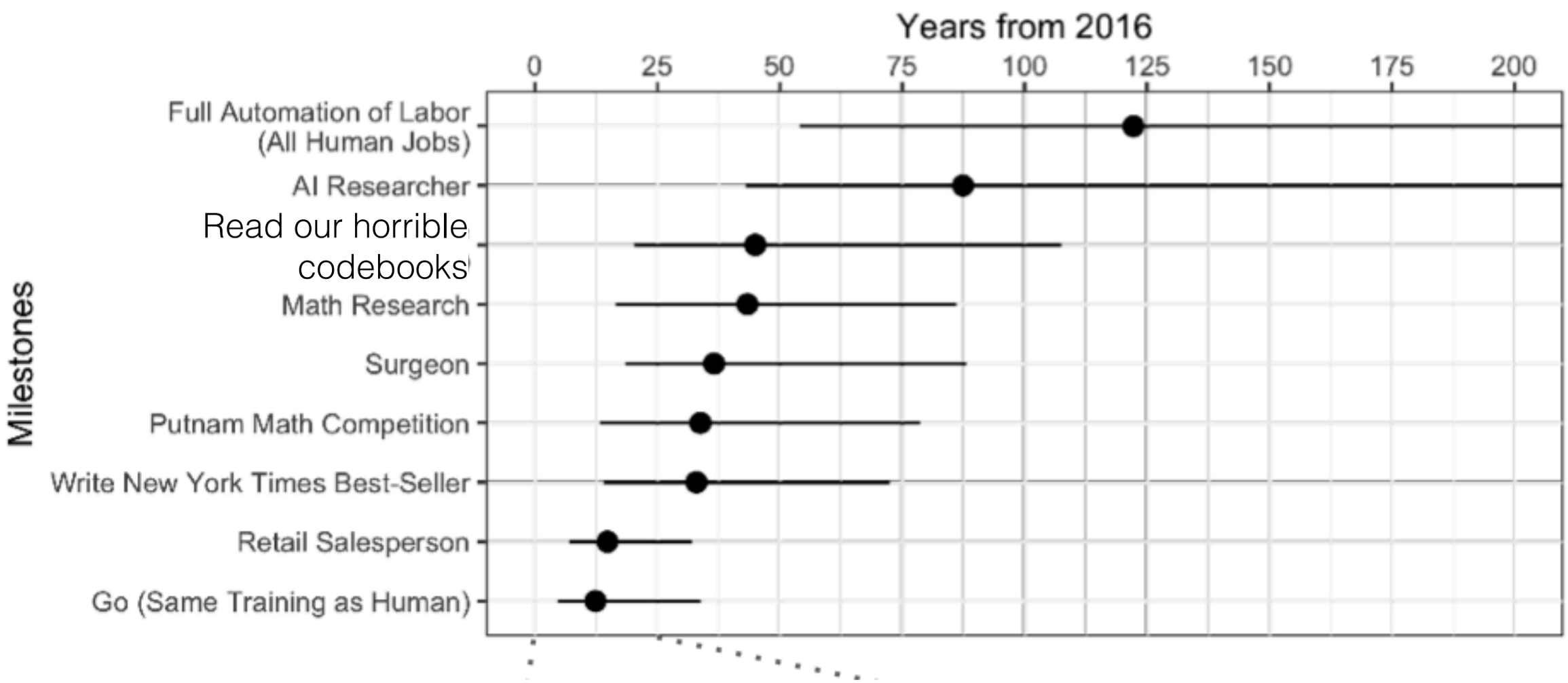


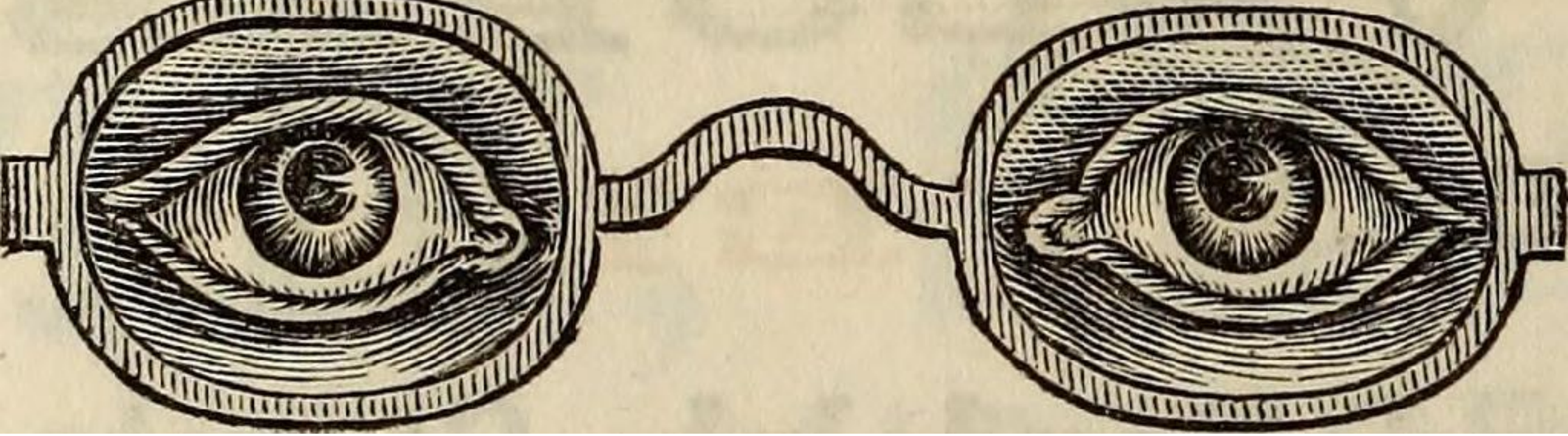

**Figure 2: Timeline of Median Estimates (with 50% intervals) for AI Achieving Human Performance.** Timelines showing 50% probability intervals for achieving selected AI milestones. Specifically, intervals represent the date range from the 25% to 75% probability of the event occurring, calculated from the mean of individual CDFs as in Fig. 1. Circles denote the 50%-probability year. Each milestone is for AI to achieve or surpass human expert/professional performance (full descriptions in Table S5). Note that these intervals represent the uncertainty of survey respondents, not estimation uncertainty.

Grace et al. (2018)

# Machine readability




Grace et al. (2018)

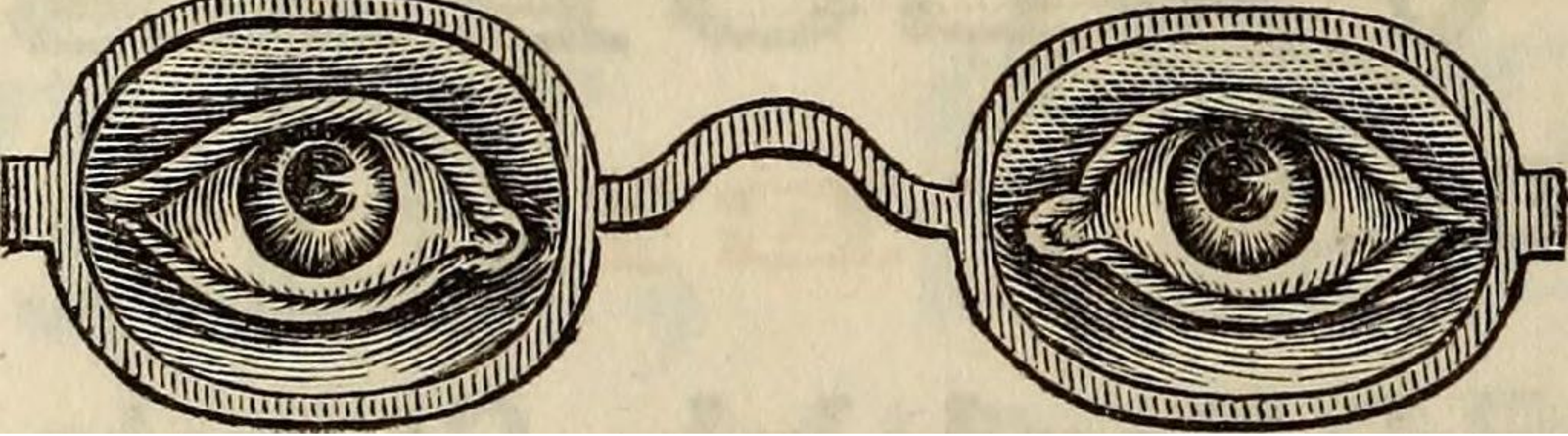


Google Dataset Search Beta

Search for Datasets

Try boston education data or weather site:noaa.gov

https://toolbox.google.com/datasetsearch



- On Google Dataset Search, it's hard to find even datasets that you know exist
- Not to mention datasets that you don't know about, but which might be useful
- The OSF is not indexed
- Variable names and labels are not indexed

# Perfect research metadata

- meaningful variable names, question labels, value labels –> linked to a standard ontology of psychological constructs and instruments
- keywords, descriptions, for hungry search engine crawlers
- information about the data: N, distributions, means, missings
- a format that is standardised, yet fits all purposes
- ???

# FAIR

## Findable

Globally unique, persistent identifier (e.g., DOI), rich metadata, indexed for search

## Accessible

Retrievable using standardised, open protocol (e.g., HTTPS). Metadata stay accessible when data is removed.

## Interoperable

Metadata use a formal, accessible, shared, broadly applicable language/vocabulary, references to other metadata

## Re-usable

Accurate and relevant attributes, provenance and data usage licence is clear, meet domain-relevant community standards.

https://www.force11.org/group/fairgroup/fairprinciples

# Ok, let me just add an ontology then

**Psychology Ontology**
Last uploaded: November 2, 2014

Summary | Classes | Properties | Notes | Mappings | Widgets

Jump to:

**Displaying the path to this class has taken too long. You can browse classes below.**

- Abandonment
- Abdomen
- Abdominal Wall
- Abducens Nerve
- Ability
- Ability Grouping
- Ability Level
- Abnormal Psychology
- Abortion (Attitudes Toward)
- Abortion Laws
- Absorption (Physiological)
- Abstraction
- Abuse of Power
- Abuse Reporting
- Academic Achievement
- Academic Achievement Motivation
- Academic Achievement Prediction
- Academic Aptitude
- Academic Environment
- Academic Failure
- Academic Overachievement
- Academic Self Concept
- Academic Specialization
- Academic Underachievement
- **Acalculia**
- Acamprosate
- Acceleration Effects
- Acceptance and Commitment Therapy
- Accident Prevention
- Accident Proneness
- Accidents

Details | Visualization | Notes ( 0 ) | Class Mappings ( 11 )

| | |
|---|---|
| Preferred Name | Acalculia |
| Definitions | Form of aphasia involving impaired ability to perform simple arithmetic calculations. |
| ID | http://ontology.apa.org/apaonto/termsonly/OUT%20(5).owl#Acalculia |
| comment | Form of aphasia involving impaired ability to perform simple arithmetic calculations. |
| alternative_name | Dyscalculia |
| defaultLanguage | |
| formal_citation | Thesaurus of Psychological Index Terms Edited by Ian Galloway Published by American Psychological Association 2014 |
| label | Acalculia |
| prefLabel | Acalculia |
| subClassOf | http://www.w3.org/2002/07/owl#Thing |

http://bioportal.bioontology.org/ontologies/APAONTO

# Ontologies

- My take is that I would use ontologies if it's easy, but *really* don't want to build them.
- In personality psychology (my main background) many items are almost completely described by the item label. Grouping them by constructs would be less clear.

“Preparing data is too time-consuming”

“I never learned to share data online”

Houtkoop et al. 2018, see also Tenopir, 2011

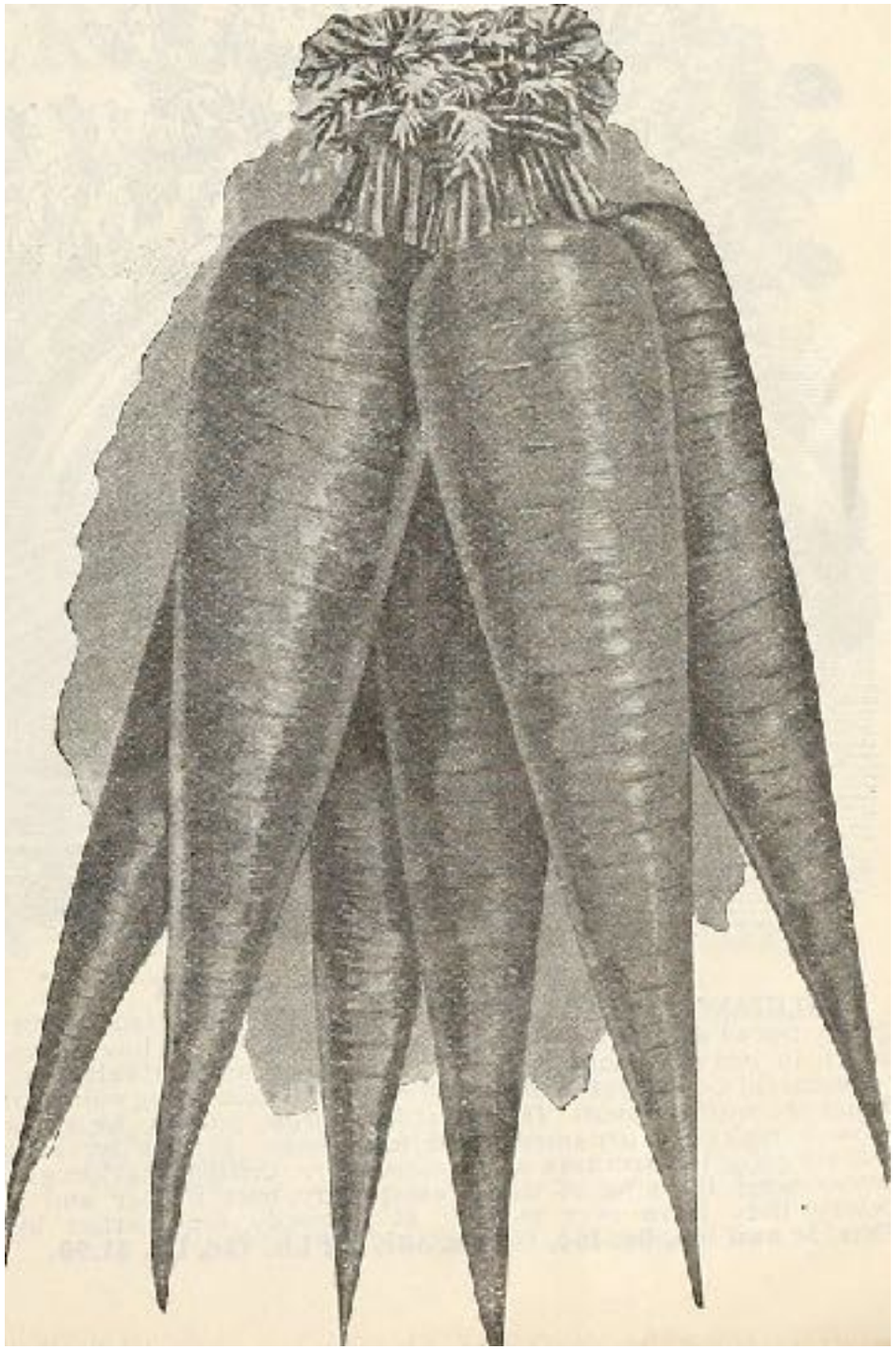
# Why don’t we add metadata?

- Many initiatives seem to have very little buy-in from researchers
- Those that document substantial amounts of data (like social science panels) often rely on specialists (librarians, database administrators, knowledge engineers)
- Not very enticing to enter tons of metadata into a form in the vague hope that some day this will become easier to find for someone else
- So, we usually do not produce the metadata we would like to get ourselves

# Carrots and sticks

- Requiring people to share data is powerful, but you cannot easily enforce the sharing of **good, useful metadata**

- We need some reward for the time and effort spent



# **codebook** package

- Main principles (Arslan, AMPPS, 2019):
  - Minimal effort (especially no **duplicated** effort)
  - Selfish benefits (not just nice for the community)
  - Machine-readable
  - Human-readable

https://github.com/rubenarslan/codebook
https://cran.rstudio.com/web/packages/codebook/index.html

# **codebook** minimal effort

- Reuse metadata that exists already (in formr.org, Qualtrics, Unipark, SoSci, SPSS files)
- Web app (three clicks, 3 minutes: codebook)
- Add metadata locally in R (or SPSS), so *you* have it, but can also *share* it
- No re-entry of existing structured metadata
- Automate tedious tasks with the help of metadata

https://github.com/rubenarslan/codebook
https://cran.rstudio.com/web/packages/codebook/index.html

# Getting variable and value labels into R

- Export data to an SAV file (possible with Qualtrics, SoSci, Unipark) or download metadata using R package (formr.org, Qualtrics)

- Import .sav (or .dta, or .csvy) file
  - `jugendl <- rio::import(“jugendl.dta")`

- If you use formr.org:
  - `s1_demo <- formr_results("s1_demo")`

# **codebook** benefits

- Visualising and aggregating scales of items, computing the appropriate reliability measure depending on whether it's a one-shot, once-repeated, or multiply-repeated study
- Visualising variables
- Making labels for variables and values accessible within RStudio
- Generating a nice study overview to check for errors and share with co-authors

https://github.com/rubenarslan/codebook
https://cran.rstudio.com/web/packages/codebook/index.html

# **codebook** machines

- **JSON-LD**: JSON linked data
  - lightweight metadata markup
  - can be embedded in an HTML webpage
  - indexed by Google Dataset Search
  - human-readable


```
{
  "@context":"http://schema.org/",
  "@type":"Dataset",
  "name":"NCDC Storm Events Database",
  "description":"Storm Data is provided by the National Weather Service (NWS) and contain statistics on
  "url":"https://catalog.data.gov/dataset/ncdc-storm-events-database",
  "sameAs":"https://gis.ncdc.noaa.gov/geoportal/catalog/search/resource/details.page?id=gov.noaa.ncdc:C
  "keywords":[
     "ATMOSPHERE > ATMOSPHERIC PHENOMENA > CYCLONES",
     "ATMOSPHERE > ATMOSPHERIC PHENOMENA > DROUGHT",
     "ATMOSPHERE > ATMOSPHERIC PHENOMENA > FOG",
     "ATMOSPHERE > ATMOSPHERIC PHENOMENA > FREEZE"
  ],
  "creator":{
     "@type":"Organization",
     "url": "https://www.ncei.noaa.gov/",
     "name":"OC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S. D
     "contactPoint":{
        "@type":"ContactPoint",
        "contactType": "customer service",
        "telephone":"+1-828-271-4800",
        "email":"ncei.orders@noaa.gov"
     }
  },
```


https://developers.google.com/search/docs/data-types/dataset#complete-example-dataset-markup-in-json-ld

# Other Metadata standards

- **DDI**: Documenting data Initiative
  - **heavy**weight metadata markup
  - only indexed by custom search engines for datasets they store
  - few open implementations
  - I guess some humans can read this
  - hard to like, hard for small teams to get into


```
<?xml version="1.0" encoding="utf-8"?>
<ddi:FragmentInstance xmlns:ddi="ddi:instance:3_2">
  <ddi:TopLevelReference>
    <Agency xmlns="ddi:reusable:3_2">example.org</Agency>
    <ID xmlns="ddi:reusable:3_2">88d12d54-c6ee-496c-b591-2b52071
    <Version xmlns="ddi:reusable:3_2">1</Version>
    <TypeOfObject xmlns="ddi:reusable:3_2">DDIInstance</TypeOfOb
  </ddi:TopLevelReference>
  <ddi:Fragment>
    <ddi:DDIInstance isUniversallyUnique="true" versionDate="201
      <URN xmlns="ddi:reusable:3_2">urn:ddi:example.org:88d12d54
      <Agency xmlns="ddi:reusable:3_2">example.org</Agency>
      <ID xmlns="ddi:reusable:3_2">88d12d54-c6ee-496c-b591-2b520
      <Version xmlns="ddi:reusable:3_2">1</Version>
      <Citation xmlns="ddi:reusable:3_2">
        <Title>
          <String xml:lang="en-US">CBS News/New York Times Month
        </Title>
        <PublicationDate>
          <SimpleDate>2010-03-31T00:00:00</SimpleDate>
        </PublicationDate>
        <InternationalIdentifier>
          <IdentifierContent>2079</IdentifierContent>
          <ManagingAgency>en-US</ManagingAgency>
        </InternationalIdentifier>
      </Citation>
      <ResourcePackageReference xmlns="ddi:reusable:3_2">
        <Agency>example.org</Agency>
        <ID>265fcdd2-9a9a-4be3-b84c-c433a3233ecd</ID>
        <Version>1</Version>
        <TypeOfObject>ResourcePackage</TypeOfObject>
      </ResourcePackageReference>
      <StudyUnitReference xmlns="ddi:reusable:3_2">
        <Agency>example.org</Agency>
        <ID>bb69605e-50b5-4618-8207-2c4d387b6e48</ID>
        <Version>1</Version>
        <TypeOfObject>StudyUnit</TypeOfObject>
      </StudyUnitReference>
    </ddi:DDIInstance>
  </ddi:Fragment>
  <ddi:Fragment>
    <ResourcePackage isUniversallyUnique="true" versionDate="201
      <URN xmlns="ddi:reusable:3_2">urn:ddi:example.org:265fcdd2
      <Agency xmlns="ddi:reusable:3_2">example.org</Agency>
      <ID xmlns="ddi:reusable:3_2">265fcdd2-9a9a-4be3-b84c-c433a
      <Version xmlns="ddi:reusable:3_2">1</Version>
      <UserAttributePair xmlns="ddi:reusable:3_2">
        <AttributeKey>extension:CodeListReferences</AttributeKey
        <AttributeValue>["urn:ddi:example.org:382be111-50c4-46cc
1","urn:ddi:example.org:8bd6833b-a7f8-4b7e-ad8c-3dff69e64aa0:1",
996f88ce-07c1-403e-9cbb-839228be0c73:1","urn:ddi:example.org:d41
bac2-268d119239a6:1","urn:ddi:example.org:9ed803f3-79ad-4348-abb
e5fd-410d-8315-0f95b6323137:1","urn:ddi:example.org:e2de2245-03e
fd778baf4c5e:1","urn:ddi:example.org:5ea4b8e0-647b-49b8-91e2-997
1","urn:ddi:example.org:3770bd02-75bc-43cf-bb1f-7a133a3c26f7:1",
23b06c96-bdad-4030-8f11-5523693198f4:1","urn:ddi:example.org:d98
85bf84af-3fbd-4786-9b24-71c11af52a97:1","urn:ddi:example.org:c67
fe0b638b23d0:1","urn:ddi:example.org:7290b31b-4860-4aa3-8cb0-d51
ad36-436ec2d298a2:1","urn:ddi:example.org:eae96440-c166-4393-855
1","urn:ddi:example.org:f54a5703-c9b4-4ce5-8788-d301fb43231f:1",
6179df9d-a752-4935-a90c-0ba2a94df40f:1","urn:ddi:example.org:550
b258-4f872e66fef4:1","urn:ddi:example.org:710714d2-b8db-4b9b-bfa
e1d348f5bee9:1","urn:ddi:example.org:5ce7fdeb-a001-4248-9109-081
f52073e46718:1","urn:ddi:example.org:e376c10d-0640-4ea4-9dd8-534
dd463fd0a9b4:1","urn:ddi:example.org:8aa0dd7b-d780-41c3-92df-3a2
e8ee1092dc48:1","urn:ddi:example.org:08bf5adc-cbb2-4388-a839-5b7
c02e1c1bca54:1","urn:ddi:example.org:8b65e1af-b55e-4fc9-931f-dd8
      </UserAttributePair>
      <Citation xmlns="ddi:reusable:3_2" />
      <DataCollectionReference xmlns="ddi:reusable:3_2">
        <Agency>example.org</Agency>
        <ID>feeb5048-da4b-4817-867b-19178761def9</ID>
        <Version>1</Version>
        <TypeOfObject>DataCollection</TypeOfObject>
      </DataCollectionReference>
      <LogicalProductReference xmlns="ddi:reusable:3_2">
        <Agency>example.org</Agency>
```

# Why Use DDI?

DDI encourages **comprehensive description** of data for discovery and analysis and supports **effective data sharing**. Because DDI is a **structured** standard, it facilitates **machine-actionability and interoperability** and it can actually be used to **drive systems.** Another feature of DDI is its focus on **metadata reuse**; "enter once, use often" means you can reuse metadata over the course of the data life cycle to avoid costly duplication of effort.

DDI has advantages for several different audiences:

+ Librarians

+ Managers

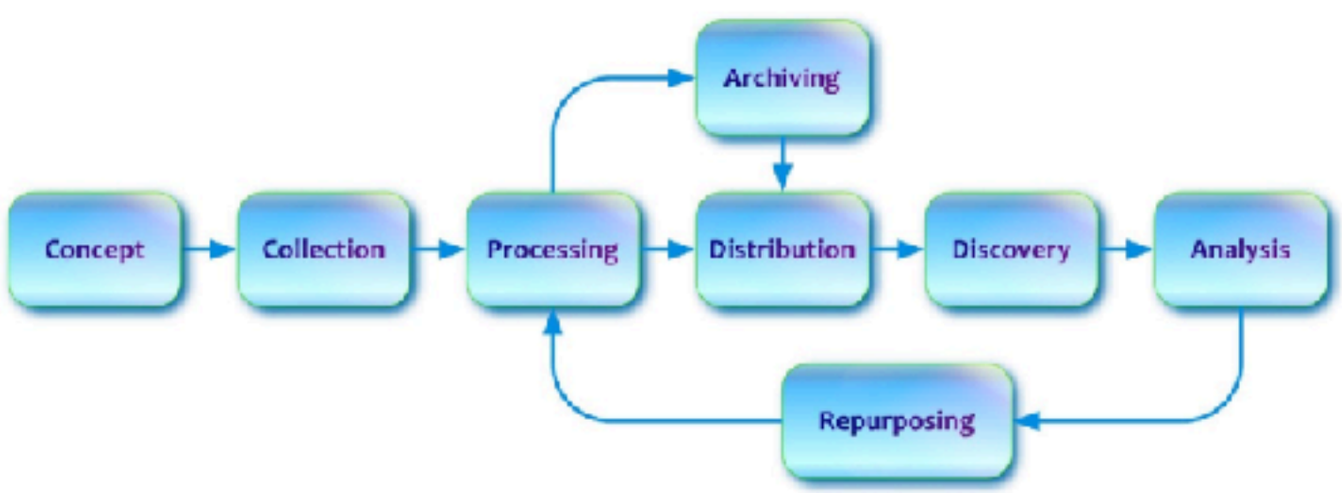
+ Repositories

- Researchers

- Recent open access mandates from funders require that data be shared in order to validate results and to encourage new discoveries. This means that data must be well-documented, which is DDI's strength.
- Complex, longitudinal data projects require additional levels of data management. DDI can support this and can enable creation of reports, displays, and tools that leverage the richness of the data. Some examples are question banks, concordances, and interactive codebooks.
- The structure of DDI can support data comparison and harmonization.
- Interested in learning more about DDI? Contact us.

+ Developers

DDI Data Lifecycle




ddialliance.org

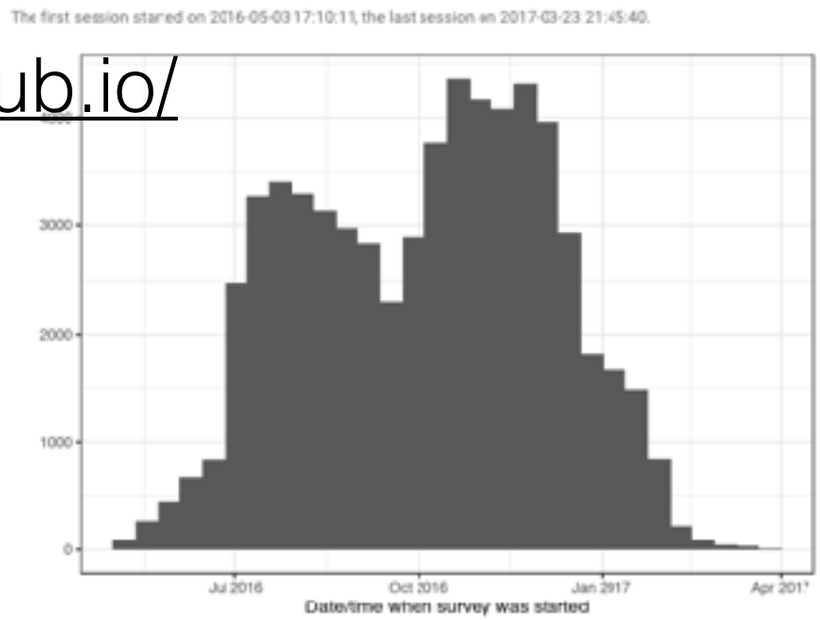
# Document, Discover and Interoperate

- Very complex
- Few resources for individual researchers
- Most tools are proprietary

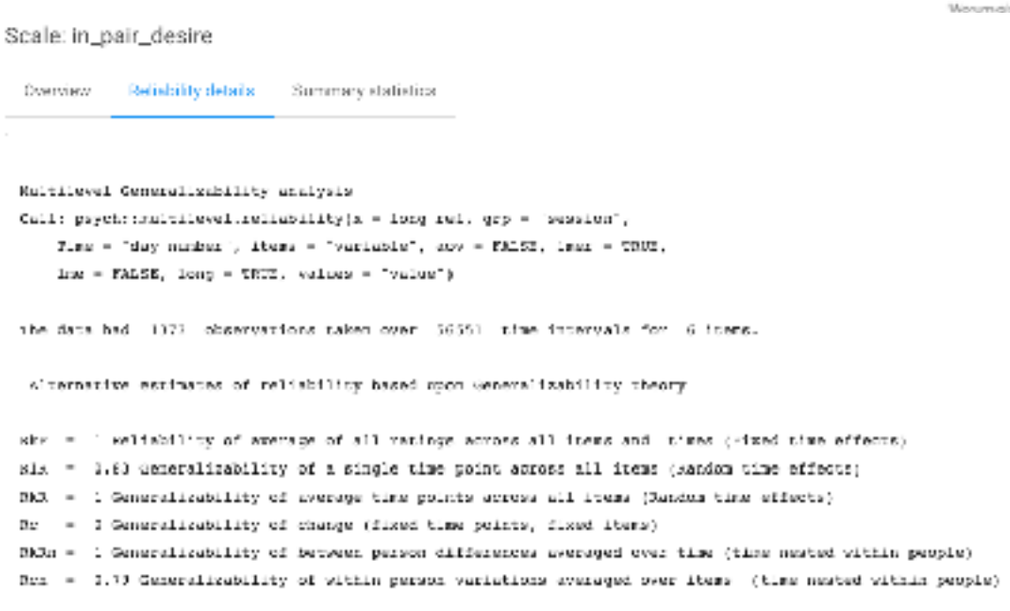

ddialliance.org

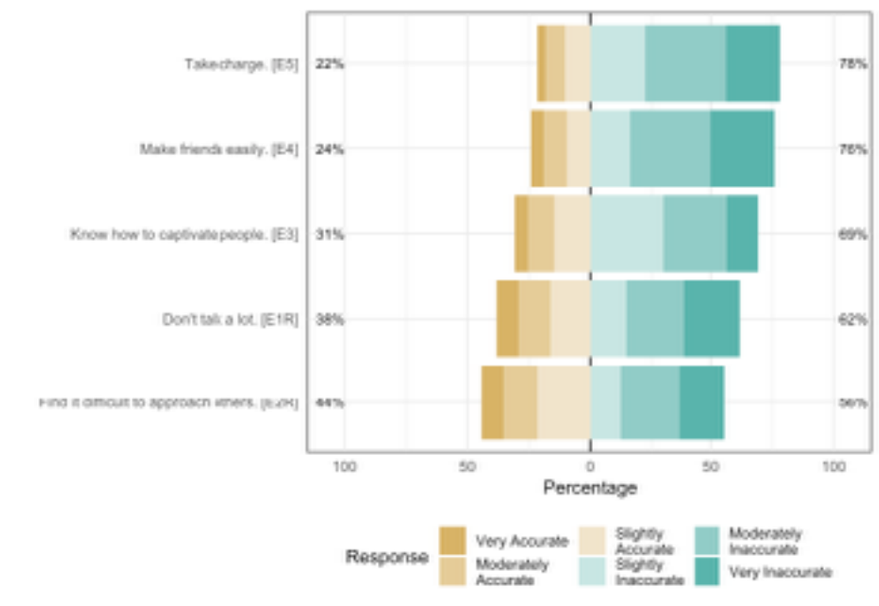
# **codebook** humans

- Example gallery: https://rubenarslan.github.io/codebook_gallery







# **codebook** programming

- 82% test coverage
- Permissive open MIT license
- 13K downloads from Rstudio's CRAN so far
- Archive major versions on Zenodo
- Version control on Github: https://github.com/rubenarslan/codebook

# Online web app

- Go to https://codebook.formr.org
- Do you have a dataset on your computer that fits the following criteria?
  - <2Mb, <50 variables
  - .sav/.dta format with value and variable labels set
  - not sensitive
- If not, use https://osf.io/j4fcb

# Local use

- Make and customise a codebook for one of your datasets in RStudio

- Open RStudio and enter

- `library(codebook)`
  `new_codebook_rmd()`

# Open Science Framework

- a one-stop shop for
  - preregistration
  - documentation of stimuli, materials, questionnaires
  - preprints
  - data archival – easy, but not very useful. doesn't currently do metadata/indexing, plans only for citation metadata.

CENTER FOR OPEN SCIENCE

# Other data repositories

- UK Data Service ReShare: https://reshare.ukdataservice.ac.uk
  Allows for open and limited data access, you have to enter metadata, but sign up sucks for non-UK people and I needed to send an email to learn how to log in.

- OpenICPSR: https://www.openicpsr.org
  Probably the best out there right now, but it requires you to re-enter metadata by hand, even if it's already stored somewhere (e.g. in a JSON file or in attributes), failed in weird ways when I tried it

- IPUMS https://www.ipums.org/
  Fairly specialised on censuses/social surveys, hard to get to raw data on the website AFAICT

- Harvard Dataverse: https://dataverse.harvard.edu
  It seems you would have to enter information on variables in a flat README

- Zenodo: https://zenodo.org/
  Last I checked no support for variable labels etc.

- Figshare: https://figshare.com
  Haven't seen any metadata except citation info and descriptions there

- Dryad: https://datadryad.org/
  Haven't seen any metadata except citation info and descriptions there

- PsychData: https://www.psychdata.de/
  No machine-readable metadata (or at least not indexed in Google?), allows for detailed info, but has to be entered by hand

- Github: https://github.com
  Full flexibility to document your data on Github Pages, but nobody will take you by the hand

See also Meyer (2018) Table 1

# Dilemma

- Use a fully-featured service that’s not user-friendly, gives you little in return
- Use a user-friendly service that isn’t fully featured
- Best of both worlds: Use a user-friendly service and supplement it with a website generated by the codebook package.

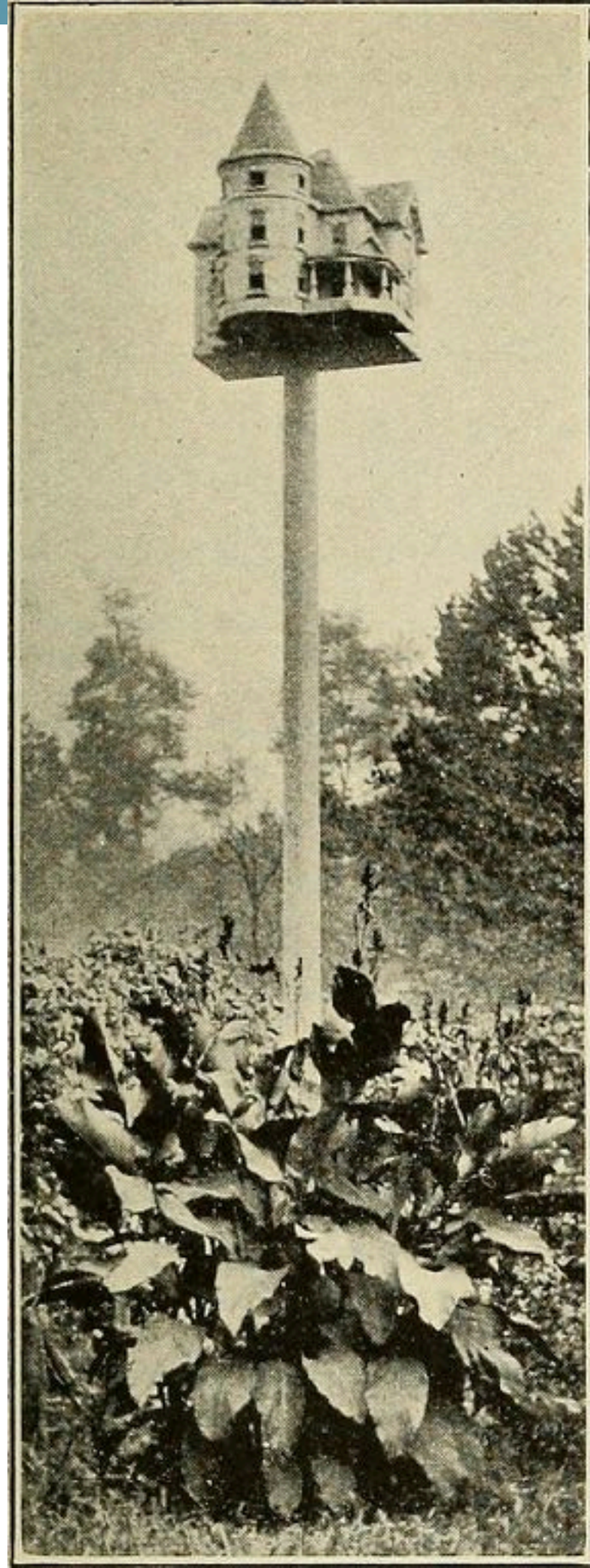
# Publish a codebook

- Sign up on netlify.com (or use Github/Gitlab)
- Rename your codebook.html file to index.html and put it in a folder on its own.
- Drag and drop the folder to your first netlify page
- Give the page a meaningful name (e.g., the name of the dataset)

  (less than 5 minutes)

# Forum

- My main approach is to keep the effort minimal/compensate people for their time
- Psych-DS takes a similar approach
- This leaves some gaps (e.g., we don't enforce proper citation metadata)
- Pro/Contra of a more opinionated approach



Max-Planck-Institut für Bildungsforschung
Max Planck Institute for Human Development


Ruben C. Arslan

ruben.arslan@gmail.com
@rubenarslan

survey software: formr.org
blog: http://the100.ci
also blog: rubenarslan.github.io

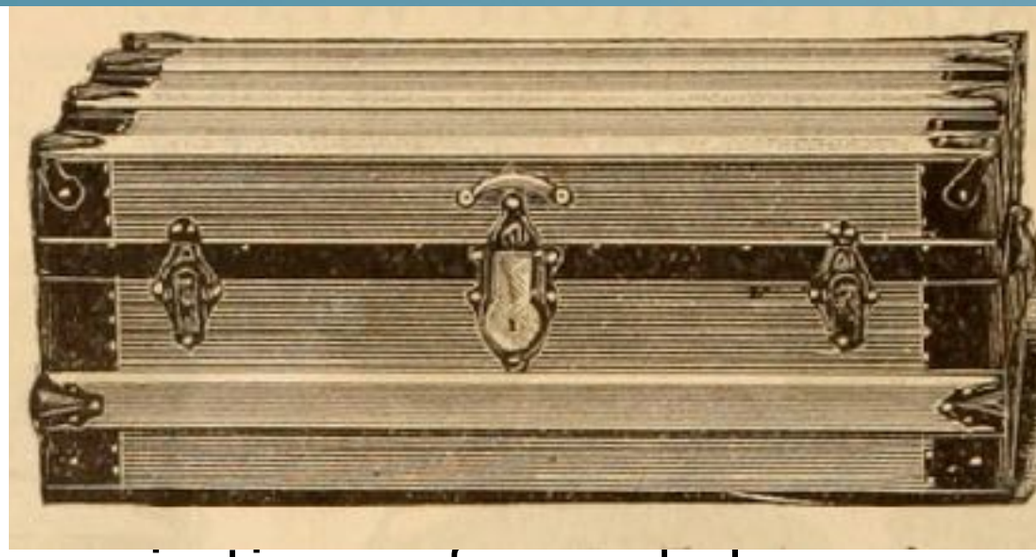
# Weigh in

- We are currently planning PsychDS, a specification for how psychological data should be shared
- Draft
- What are some things the spec should do?

# Codebook examples

- https://rubenarslan.github.io/routine_and_sex/2_codebook.html
- https://rubenarslan.github.io/codebook/
- https://rubenarslan.github.io/codebook_gallery/

# Free idea for an R package



- Come up with a schema.org description of models based on the *tidy* concept in broom.

- Supply nice visual and textual summaries of models using the same approach I used in codebook

- Add machine-readable metadata behind the scenes

  https://scienceverse.github.io/scienceverse/ ?

# Alternative tools

- DDI codebooks
  could not get this to work myself after a few hours of looking into it, will not be indexed by Google (AFAICT)

- dataMaid
  https://cran.r-project.org/web/packages/dataMaid/index.html
  focused on error correction, won't yield machine-readable metadata

- dataspice
  https://github.com/ropenscilabs/dataspice
  very similar, focused on biology/ecology, has a much nicer interactive tool for adding metadata, less visualisation than codebook, does not import existing metadata from other sources