



# Maintaining privacy with open data

Ruben Arslan  
Munich, February 23, 2018

[ruben.arslan@gmail.com](mailto:ruben.arslan@gmail.com)

 [@rubenarslan](https://twitter.com/rubenarslan)

blog: <http://the100.ci>

# Time plan

<b>09:00</b>	Short introductory round
<b>09:20</b>	Why share data at all?
<b>10:00</b>	Assessing potential for harm and risk
<b>10:30</b>	Deidentification, anonymisation, Exercise
<b>11:30</b>	Getting the right consent
<b>12:00</b>	Lunch!
<b>13:00</b>	Not collecting personally identifiable information (PII) in the first place, How to manage recontact without PII
<b>13:30</b>	Sharing scientific use files, back ups, encryption
<b>14:00</b>	Your own study, your own problems
<b>15:00</b>	Outlook: synthetic data, differential privacy
<b>16:00</b>	End

# Introductory round

- Name, Department
- Your next planned study
  - Focus on design
  - Cross-sectional/longitudinal/diary/experience sampling/???
  - Social network
  - Randomised experiments
  - Combining with data from lab visits?
  - Rating stimuli?
  - Where do you expect difficulties?
- Past experiences with data sharing

# Why share data at all?

- *Nullius in verba*
  - motto of the Royal SocietyPeople 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

# 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](#)

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

Funder	Data-sharing policy
German Research Foundation	<p>“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: <b>Publicly funded research, including the raw data, belongs to the public.</b> 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).</p>
National Institutes of Health	<p>“The <i>2003 NIH Data Sharing Policy</i> 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</p>

# Why not share data?

# Why not share data publicly?

More papers for you

It's extra work

Ensure analysts know what they're doing

Maintain participant privacy

Prevent overfitting, rat races



# Why not share data publicly?

~~More papers for you~~

Probably *less* papers for you: Benefit of collaboration might outweigh potential loss by scooping; probably more citations (see [osf.io/cdt8y](https://osf.io/cdt8y))

It's extra work

Yes, but it's easier than it ever was

Ensure analysts know what they're doing

Ensure others check you know what you're doing

Maintain participant privacy

Make the best use of participant's hard-won data

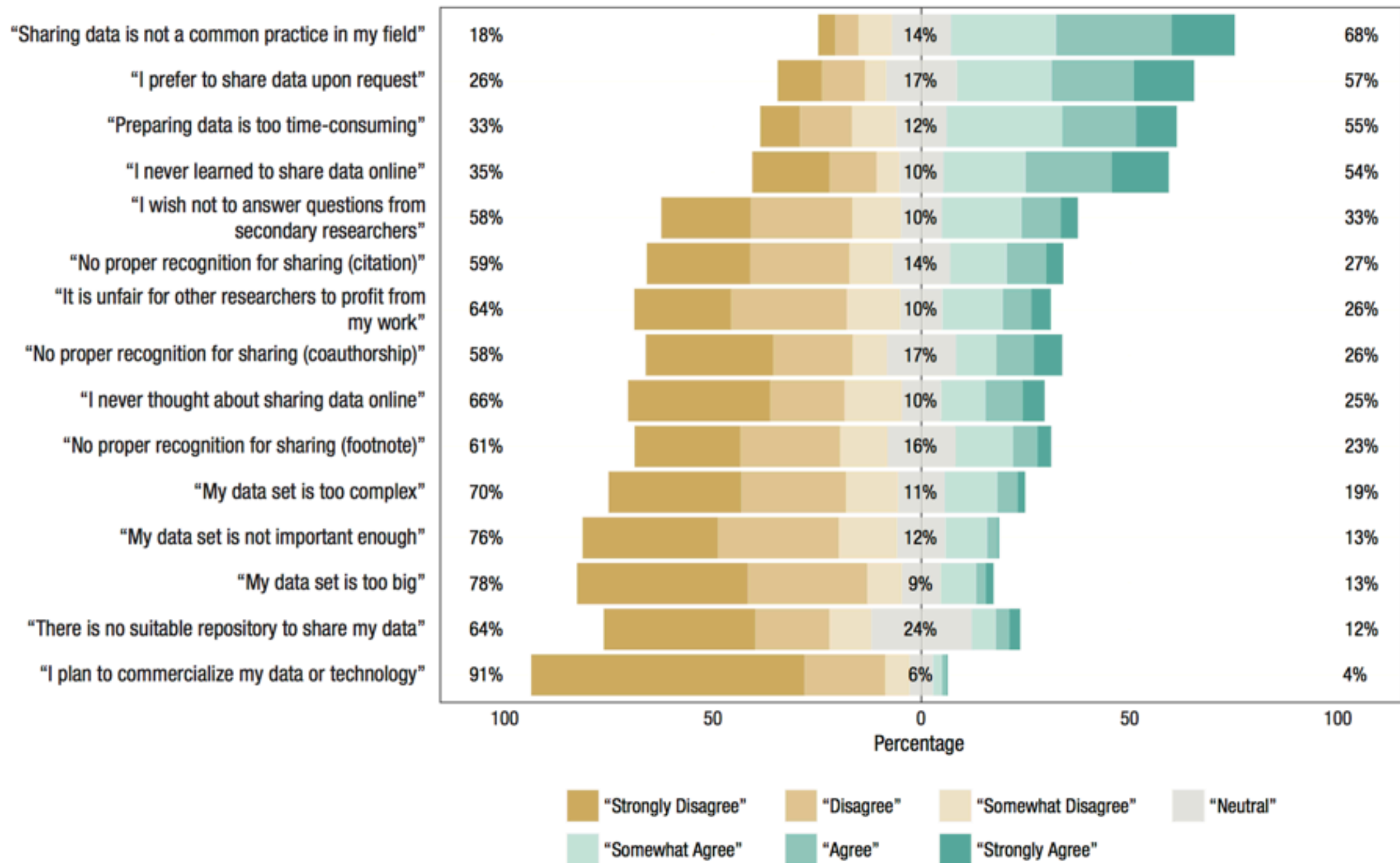
Prevent overfitting, rat races

We overfit plenty on our own

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

Journal or publisher	Data-sharing policy
<i>Nature</i>	“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. . . .” ( <i>Nature</i> , 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. “When submitting a manuscript online, authors must provide a <i>Data Availability Statement</i> 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. “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. “Datasets and code should be deposited in an appropriate, recognised, publicly available repository. . . . “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).
<i>Science</i>	“After publication, all data and materials necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of <i>Science</i> . . . . 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” ( <i>Science</i> , 2017, Data and Materials Availability After Publication).

“To what extent do you agree with the  
following statements about barriers related to data sharing?”

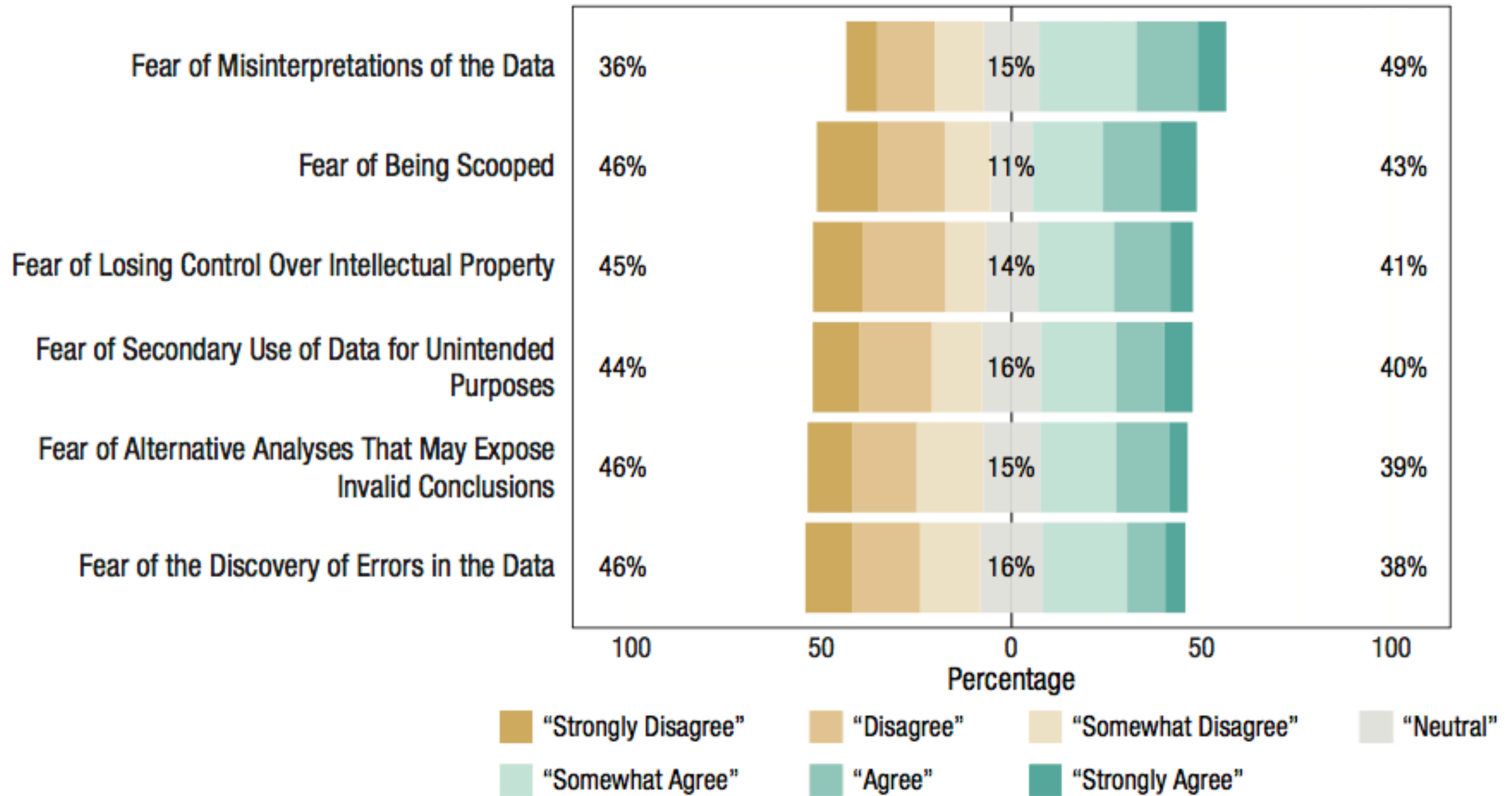


**Fig. 2.** Responses to the survey questions asking respondents to indicate the extent to which the 15 non-fear-related barriers kept them from sharing their research data. For each statement, the number to the left of the data bar indicates the percentage of researchers who responded with “strongly 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 & Speerschnieder, 2015).



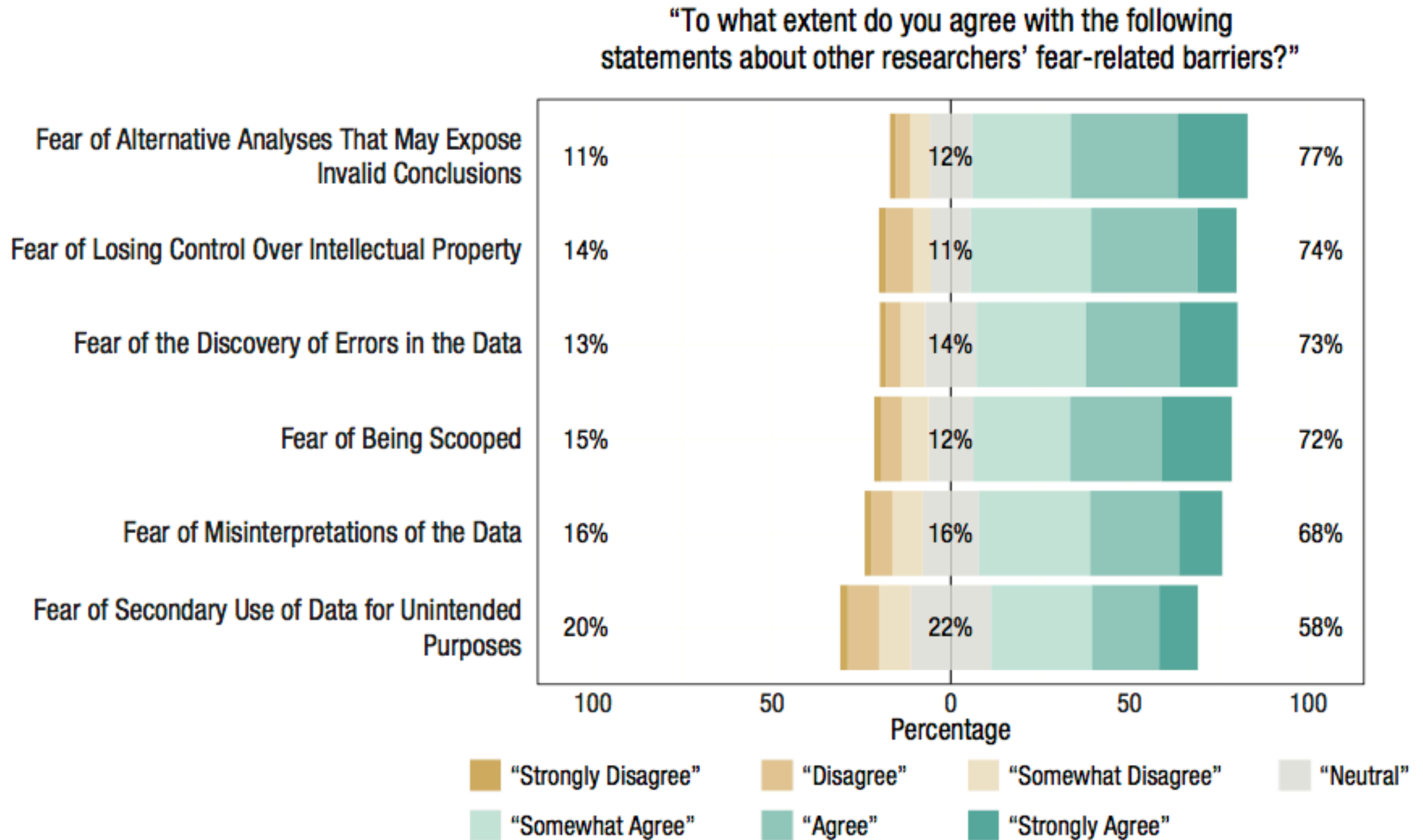
a

"To what extent do you agree with the following statements about fear-related barriers, evaluated for yourself?"

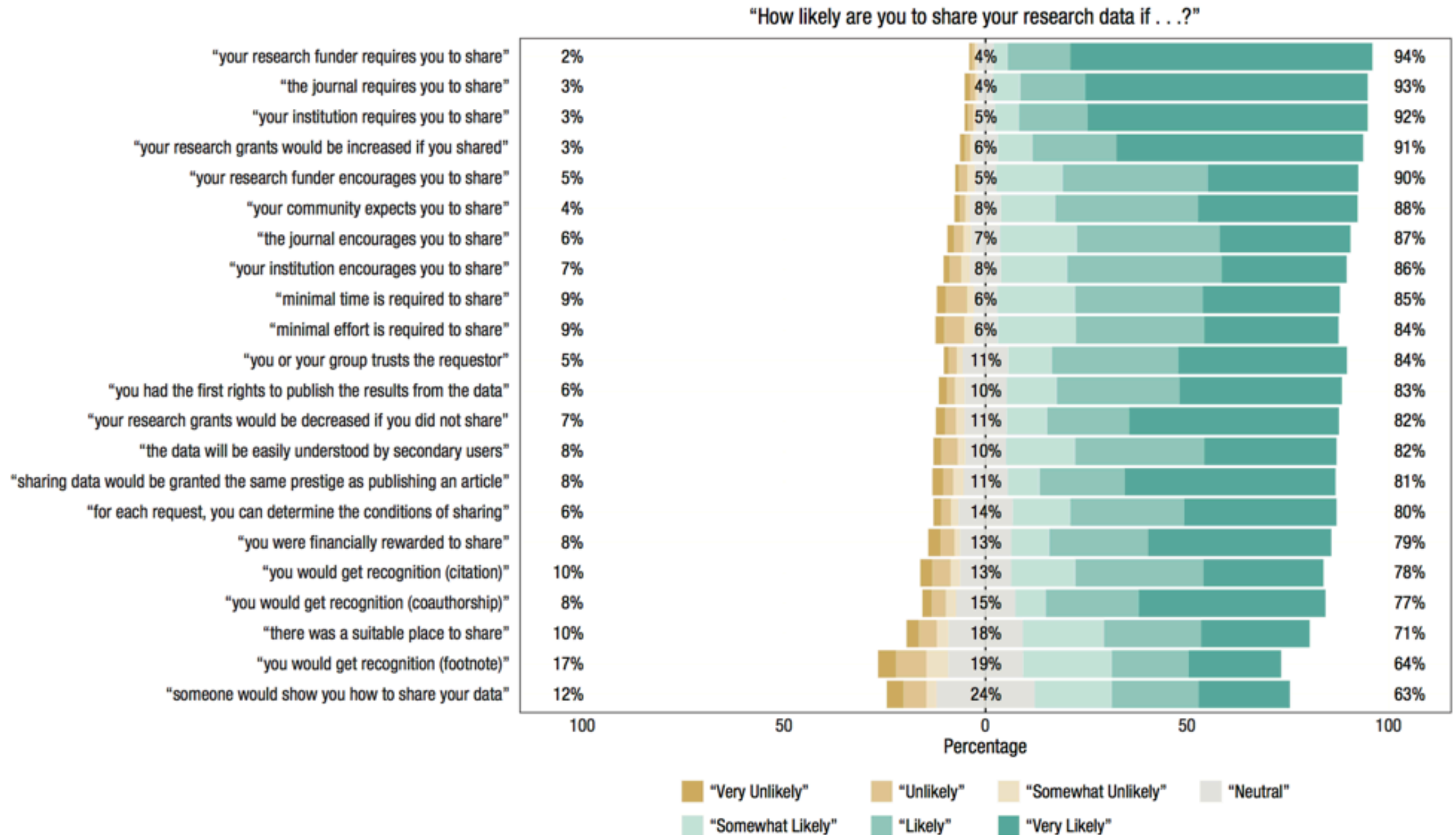


h

D



**Fig. 3.** Responses to the survey questions asking researchers to indicate the extent to which the six fear-related barriers kept (a) themselves or (b) other researchers from sharing their 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 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.” In each panel, 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 & Speerscheider, 2015).



**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 & Speerschnieder, 2015).

# So I just share my data?

- Not so fast
- Conflicting obligation: maintain participant privacy

# Disclaimer

- IANAL: No legal advice.
- If I could provide this, I'd be rich.



# Sensitive data

- Poll: Who collects any of the following?
  - 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?

# What's the worst that could happen? Harm assessment

- mild embarrassment
- worse insurance rates
- jealous partner murder-suicide
- SPAM

# Risk of identification

- Is it clear where and from whom the data was collected (e.g. small subject pool in small university in one year)?
- Are you studying something rare (e.g. trichotillomania)? Are the patient groups socially connected?
- Do you study couples, families, or co-workers?
- Is your data high-dimensional?
  - lots of measures
  - lots of time points
  - including genetic, brain, geographic data

# Risk assessment

- Who are you sharing data with?
  - Student assistants whose friends may be in the sample?
  - Collaborators in/outside EU
  - Scientists that you don't know
  - Random people on the internet

# Risk assessment

- Poll: Who might want to do harm with your data?

# Identifying someone special

- Jealous partner
  - has access to browser, phone, knows age, mother's Maiden name, personality.
  - wants to find out if partner is faithful
- Nosy neighbour, co-ed, co-worker (including Mturk)
  - knows age, demographic info, address, rough idea of personality
  - wants to know if single, psychologically ill, obtain sexual history
- Hacktivists, journalists
  - targets people with a public profile for some reason (politicians, red sweater guy)
  - wants to expose nasty information about them, or get their password question

# Looking for anyone

- Blackmailers, identity thieves
  - target no one specific, have access to third-party data sources (social networks, etc.)
  - want money (by using identity of one individual and embarrassing data or data that allows identity theft)
- Marketers, spammers
  - targets broad groups who might be interested in product (e.g. people with certain preferences, illnesses)
  - wants e.g. their email addresses to send spam/target ads on Facebook
- Computer scientists
  - target hard targets
  - want to show how well their cool re-identification algorithms work
- Insurers
  - target prospective clients
  - want to predict likelihood of claims, to price rates adaptively or reject people

# Lessons

- Harm can come in unforeseen ways
- Users won't think about privacy for you
- Even without *personally identifiable information*, people might be identified or harmed



# What is Personally identifiable information (PII)?

- Traditionally things like:
  - Full name (if not common)
  - Address (post or email)
  - Phone numbers
  - Passport number
  - Face, fingerprints, or handwriting
  - Credit card numbers
  - Date of birth
  - Birthplace
  - Genetic information

# What is *Personal data*?

- European data protection law does not use the concept of PII, and its scope is instead determined by the non-synonymous, wider concept of "personal data".
- The data subject can potentially be identified through additional processing of other attributes—
  - quasi- or pseudo-identifiers.
- In the EU **General Data Protection Regulation**, this has been formalized in Article 4: a "data subject" is one "who can be identified, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person".
- The GDPR becomes enforceable on 25th May 2018.

# What **isn't** Personal data?

- Arguably not much
- Who is the 30-year-old male psychology student in your sample of a Munich participant pool?
- Who is the patient with five unusual co-morbid disorders?
- Who is the 45-year-old patient who enrolled in June and came back for T2 55 days later?
- Brain scans, blood work may all be enough to identify someone

# 人肉搜索

- Human Flesh Search Engine
  - a Chinese term for the phenomenon of distributed researching using Internet media such as blogs and forums. It is similar to the concept of “doxing”
- Exercise

# What is de-identification?

- Removing PII from the data
- Try to reduce the risk that data can be re-identified (linked to individuals) by doing things like
  - binning (age -> age groups, rare categories -> other)
  - fuzzing (e.g.  $\pm 2$  standard deviation of noise)
  - removing free-texts, dates, times
  - removing unique entries
  - and so on

# Risks of de-identification?

- Since it essentially adds noise and systematic missings to the data, it can decrease the usefulness of the data
- *“GDPR applies a standard, considering data anonymous only when it cannot be identified by any means “reasonably likely to be used ... either by the controller or by another person” (Recital 26). Thus, even if a researcher no longer has the ability to re-identify a data set, such data set may still be regulated under the GDPR if it could be re-identified with reasonable effort.”*
- Since you're not the foremost re-identification expert in the world, you might underestimate the ability of others to re-identify the data

# In case of sensitive data

- Don't share what others don't need (e.g. PII)
- If usefulness can be preserved, separate the most sensitive parts from the rest and share those only with a more exclusive set of people
- De-identification techniques
- Don't share with the world's foremost experts in re-identification ;-), or people who know others in your dataset.
- Ask for the right consent – you're not the best judge of what others consider sensitive

# Deidentification exercise

- I shared a dataset with you (I removed truly sensitive parts beforehand).
- We'll work together through an R script and the various ways how to anonymise the data.
- [hirntrichter.de/priv.zip](https://hirntrichter.de/priv.zip)



# Setup

- download & install **R** [cran.r-project.org](https://cran.r-project.org)
- download & install **RStudio** [rstudio.com](https://rstudio.com)
- in RStudio, execute:  

```
install.packages(c("devtools", "synthpop"))  
devtools::install_github("rubenarslan/codebook")
```

# Synthetic data

- Create realistic data mirroring the properties of the real data
- Difficult task
  - Properly done, can preserve privacy much more easily
- R package: synthpop
- [https://en.wikipedia.org/wiki/Synthetic\\_data](https://en.wikipedia.org/wiki/Synthetic_data)

# Documenting data

- You cannot always share all the data
  - No consent
  - Re-identifiability concerns
  - No permission from co-authors, data owners
- This does not mean you cannot share anything useful.
- Field-specific summaries can sometimes be very useful, e.g. GWAS associations per SNP, correlation matrices in psychometrics
- You can almost always share metadata

Which part of the data set is personally identifiable / has sensitive data?

Anonymous part of data set  
(e.g., reaction times,  
questionnaire scales with  
general personality items)

Document and share  
publicly

Personally  
identifiable data

Try to de-identify

Risk assessment:  
Ethical to share?

Yes

Document and  
share in a  
restricted way

Yes

No

Restricted access  
possible (e.g.  
scientific use file)?

Do not share;  
ensure safe  
storage

No

# Informed consent

- Strava had 7 levels of privacy, these soldiers agreed to have their locations revealed
- Most people don't read Terms and Conditions
- You don't even know the risks

# Kinda informed consent?

- Get broad open-ended consent that doesn't preclude sharing data in most ways
- Consider tiered consent
- Dos and Don'ts by Michelle Meyer

# Dos and Dont's (Meyer, 2018)

- DON'T promise to destroy your data
- DON'T promise not to share data
- DON'T promise that research analyses of the collected data will be limited to certain topics
- DO get consent to retain and share data
- DO incorporate data-retention and -sharing clauses into IRB templates
- DO be thoughtful when considering risks of re-identification
- DO consider working with a data repository
- DO be thoughtful when selecting a data repository

# Consent example

The data and samples from this study might be used for other, future research projects in addition to the study you are currently participating in. Those future projects can focus on any topic that might be unrelated to the goals of this study. We will give access to the data we are collecting, including the imaging data, to the general public via the Internet and a fully open database.

The data we share with the general public will not have your name on it, only a code number, so people will not know your name or which data are yours. In addition, we will not share any other information that we think might help people who know you guess which data are yours.

If you change your mind and withdraw your consent to participate in this study (you can call <PI name> at <phone number> to do this), we will not collect any additional data about you. We will delete your data if you withdraw before it was deposited in the database. **However, any data and research results already shared with other investigators or the general public cannot be destroyed, withdrawn or recalled.**

By agreeing to participate, you will be making a free and generous gift for research that might help others. It is possible that some of the research conducted using your information eventually could lead to the development of new methods for studying brain, new diagnostic tests, new drugs or other commercial products. Should this occur, there is no plan to provide you with any part of the profits generated from such products and you will not have any ownership rights in the products.



# Consent example

To the best of our knowledge, the data we release to the general public will not contain information that can directly identify you. The data will not have your name on it, only a code number, so people will not know your name or which data are yours. In addition, the data will not include data that we think might help people who know you guess which data are yours, such as your facial features or the date that you participated. If we write a report or article about this study or share the study data set with others, we will do so in such a way that you cannot be directly identified. However, by using additional data linked to your name (for example brain scans obtained from your medical records) one could potentially associate your imaging or other information in our database back to you. In addition a security breach (break in or cyber attack) might lead to someone being able to link you to your data. This risk is very low because your data are stored in a secure database, and the information about your identity is stored separately from the data themselves, linked only through a code.

We will keep the private portion (name, contact information etc.) of your data in a secure location for at least <x> years. This way if one of the researchers that obtained the data from us will find something in your brain scans that would have a diagnostic value we will be able to contact you. After this period of time we will destroy this information to protect your privacy.

**Letting us use and share your data is voluntary. However, you must be willing to share your data in this way in order to participate in this study. If you are not willing, you cannot participate in this study.**

By signing below, you agree to provide your data for future research. You agree that these may be shared with other investigators at other institutions from around the world. The details, results, and implications of these studies are unknown.

# Controversy about de-identification

- We probably cannot and should not make guarantees, especially when data is high-dimensional.
- Still, I know no examples of people in psychological research data being re-identified using fancy computer science techniques (they prefer bigger datasets than we usually have).
- Maybe we should worry more about student assistants, people who know others in the sample.
- Legal risks still unclear (IANAL)

# Privacy after and during data collection

- If possible, don't collect PII in the first place
- Not only others, also you and your team ideally have minimal access to such data
- Good practice to separate PII from research data (RD) as early as possible
  - e.g. if your account or server is hacked, or laptop is stolen. Damage limited, if data is already anon.

# Anonymity while conducting the study

- many studies require you to record your participant's identity for some reason or other
  - payment
  - feedback (email, mobile phone number)
  - communicating about bugs/problems
  - to delete/ignore data by certain people  
(e.g. somehow who writes you they only filled out nonsense the last few days and now wants to drop out)

# Operational Security

- Don't put your data or passwords in Dropbox or other version control providers (e.g. Github)
- Use a strong password (e.g. four random words of mixed languages)
- Don't email/send your password and account info together.
- Prefer end-to-end encrypted communication (e.g. Whatsapp/Signal) for this stuff

# Connecting research and identifying data

- Hard to avoid, if e.g.
  - your diary study sends emails (PII) until it has been filled out 30 times or even until 2 reports of relationship conflicts have been gathered (RD)
  - you send out feedback via email
  - you pay people on performance
  - someone emails to tell you: your study breaks when I enter 300 sexual relationships in the last 12 months
  - using mTurk worker IDs

# Let the machine do it

- as I understand the GDPR doing this *manually* is no longer seen as better
- you can automate a lot of things that RAs would usually handle
  - re-inviting people
  - calculating how much someone should be paid
  - generating feedback
- a program won't look up its co-students in its database and form a nasty opinion of them based on their sexual history.

# Separate PII from RD



[openhumans.org](https://openhumans.org)



[33mail.com](https://33mail.com)



[formr.org](https://formr.org)



# Dissociate data early

- e.g. in [formr.org](https://formr.org) surveys have an *unlinked* and *hide\_results* option
  - affects only the visible *links* for the researcher, not for formr (formr can still use an email address and link it to progress in the study)
  - whereas the RD is still *linked* by session/user codes, the unlinked ID will be shown
    - without user codes, date times in random order
    - only after at least 10 real entries exist
  - *hide\_results* is also an option, if you know you won't need the data (e.g. to dole out payment by hand)

# Discussion

- At which stage would you consider the data you collect anonymous?
- Who has access to it before then?

# Documenting studies

- There's sharing and there's sharing
- documenting
  - the study structure
  - the survey items, stimuli, programs
  - the collected data

# Documenting studies

- Ideally, you enable others to reproduce your entire study perfectly with minimal effort.
- harder if you use proprietary software
- many software packages don't export the whole study package

# Open Science Framework

- a one-stop shop for
  - preregistration
  - documentation of stimuli, materials, questionnaires
  - data archival
  - preprints



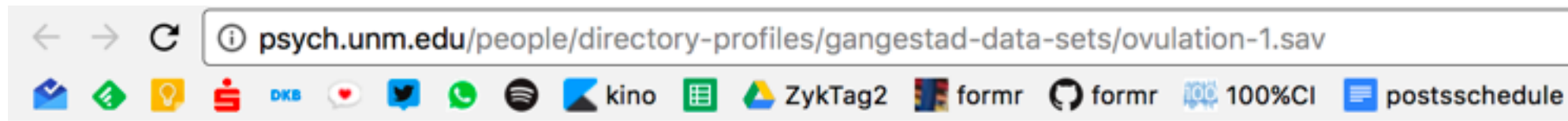
# Other data repositories

- See also Meyer (2018) Table 1
- UK Data Service ReShare: <https://reshare.ukdataservice.ac.uk>
- OpenICPSR: <https://www.openicpsr.org>
- Harvard Dataverse: <https://dataverse.harvard.edu>
- Zenodo: <https://zenodo.org/>
- Figshare: <https://figshare.com>
- Dryad: <https://datadryad.org/>
- Github: <https://github.com>

# Documenting data

- Sharing data is fast becoming the norm in research
- But 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, rdata, or xlsx
  - an upload on your department website vs. a publicly shared file that is discoverable via a search

# The dreaded departmental website



## Not Found

The requested URL /people/directory-profiles/gangestad-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.



# Documenting data

- What do you call data about data?

# Metadata

- Share codebooks ideally containing
  - variable names, item labels, value labels
  - metadata for hungry search engine crawlers
  - information about the data: distributions, means, missings
  - ???

# LengthRep 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



- ein Assistenzsystem für das Management psychologischer Forschungsdaten
- hilft, komplexe Daten in standardisierter Form zu dokumentieren
- umfangreiche Dokumentation
- viel Arbeit, XML-basiert, setzt sich das je durch?

<https://datawiz.leibniz-psychology.org/>

# Codebook

**2\_codebook.Rmd**

---

title: "Codebook formr workshop pre-survey"

output:

html\_document:

toc: true

toc\_depth: 4

toc\_float: true

code\_folding: 'hide'

self\_contained: false

---

```{r}

formr\_workshop = readRDS("formr\_workshop.rds")

codebook(formr\_workshop)

```

[tiny.cc/codebook](http://tiny.cc/codebook)

# Codebook examples

- [https://rubenarslan.github.io/routine\\_and\\_sex/2\\_codebook.html](https://rubenarslan.github.io/routine_and_sex/2_codebook.html)
- [https://rubenarslan.github.io/dating\\_satellites/2\\_codebook.html](https://rubenarslan.github.io/dating_satellites/2_codebook.html)

# Exercise

- Make and customise a codebook for one of your datasets
- You can start from a dta or sav file or do it in R.

[tiny.cc/codebook](https://tiny.cc/codebook)

# What does the future hold?

- I'll venture: more data
- Personality estimated from Facebook likes is a reasonably accurate proxy of self-report (Youyou et al. 2015) – some personalities may be unique enough to allow re-identification
- People will leave ever more data trails, that others will get better re-identifying
- “Moving forward, there is going to be no privacy at all,” Kosinski says. “And the sooner we realize that, the sooner we can start talking about how to make sure that this post-privacy world is still a habitable and safe and nice place to live in.”

<http://www.computerhistory.org/atcm/ai-social-media-data-politics-collide-with-stanford-gsbs-dr-michal-kosinski/>



# What does the future hold?



[openhumans.org](https://openhumans.org)



[midata.coop](https://midata.coop)

Give participant more insight, control over their data?  
Move beyond traditional notions of privacy?

# Open Humans





① SOURCES      ② MIDATA      ③ YOU DECIDE      ④ RESEARCH      ⑤ NEW TREATMENTS




MIDATA enables you to gather all your different health-relevant and other personal data (1) in one secure place (2).

You can decide (3) to share data with friends or physicians or to participate in research by providing access to subsets of your data (4).

In that way you contribute to the development of new treatments for OUR HEALTH (5).

# What does the future hold?

- Differential privacy?
  - Apple, Rapport (Google), PINQ (Microsoft)?
-  aircloak ?
- privacy at the analysis level (how much can be learnt about single individuals from this query)
- bonus: could reduce overfitting (Dwork et al., 2015)

# Differential privacy

- Move privacy from the level of the dataset to the level of the question asked of the dataset/analysis performed
- *Does this analysis reveal private information?*
- More consistent, but with anything but very large datasets, very little utility remains
- I don't predict many psychologists will be using this anytime soon

# Forum

- What are your concerns?
- What are special problems with your study?
- What is keeping you from sharing?
- Where do you need help with anonymising?



# Thank you!

Ruben Arslan  
Munich, February 23, 2018

[ruben.arslan@gmail.com](mailto:ruben.arslan@gmail.com)

 [@rubenarslan](https://twitter.com/rubenarslan)

blog: <http://the100.ci>

# Link collection

## OpenData

- Empfehlungen der DGPs zum Umgang mit Forschungsdaten: <http://econtent.hogrefe.com/doi/pdf/10.1026/0033-3042/a000341>
- Commitment to Research Transparency: <http://www.researchtransparency.org/>
- DS-GVO Datenschutzgrundverordnung: [https://www.ratswd.de/dl/RatSWD\\_WP\\_257.pdf](https://www.ratswd.de/dl/RatSWD_WP_257.pdf)
- Zitieren von Daten: <https://www.force11.org/group/joint-declaration-data-citation-principles-final>
- 21 word solution: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2160588](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2160588)

<http://arx.deidentifier.org/>

<https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative>