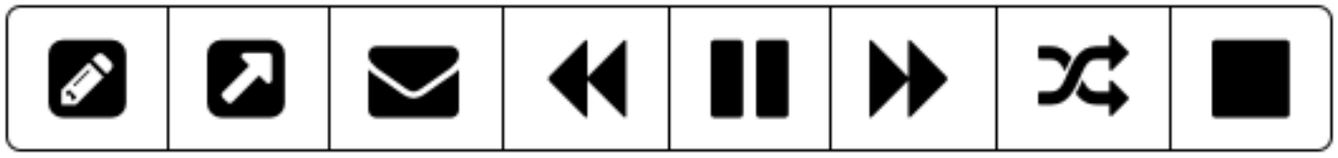
Using formr.org to build complex R-driven online studies with personalised feedback

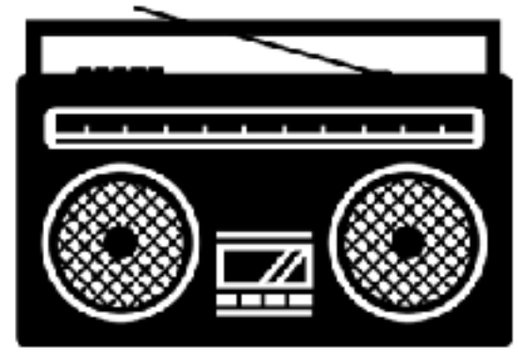


Ruben Arslan
Göttingen, September 24/25, 2018

ruben.arslan@gmail.com
@rubenarslan

blog: http://the100.ci

Time plan

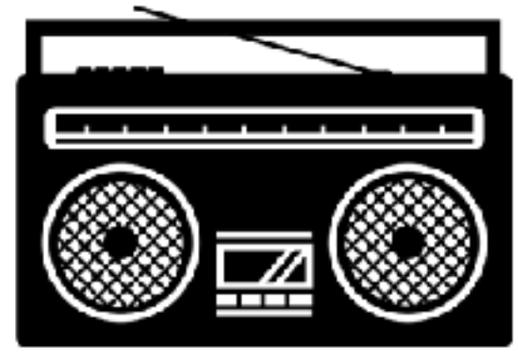


09:00	Short introductory round
09:20	Introduction to formr.org?
10:00	A simple study
10:45	Testing, data entry
11:00	A little complexity
12:00	*Lunch*
13:00	Supervise and monitor an ongoing study
13:30	Maintaining anonymity
14:00	Your own study
17:00	End

Tomorrow

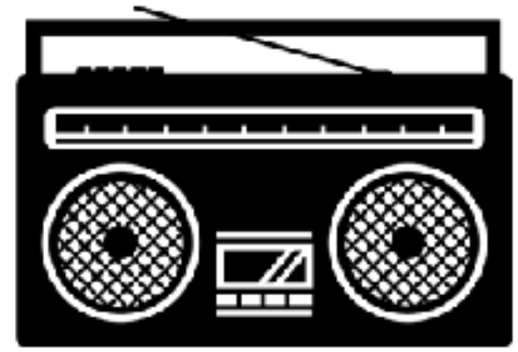
09:00	Intro to (meta)data sharing
11:00	Basics of rmarkdown, Rstudio
11:30	Getting your data out with *formr*
11:00	Making a codebook
12:00	*Lunch*
13:00	Making a codebook
14:00	Publishing the codebook
14:30	Work on your study
17:00	End

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 (frustrating) experiences with survey software

Introduction



- An environment to set up complex studies
- 3 main components
  - surveys (boring questionnaires) – ask questions
  - runs (added value) – define study operation
  - R package – document studies, clean data

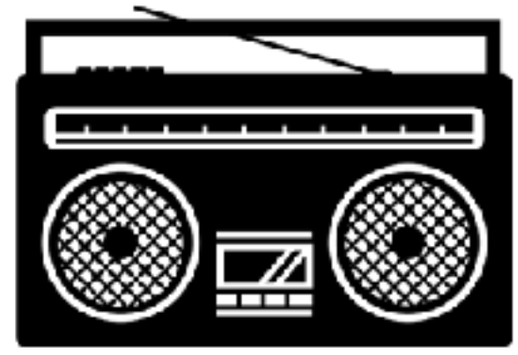
Who? Where?

- Originally developed for FSU Jena, with funding by DFG
- Based on an internal solution we had at HU Berlin
- Ruben Arslan (since 2013),
  Cyril Tata (since 2015),
  Matthias Walther (since 2018)
- Hosted here
- more information:
  formr.org/about

Matthias

Cyril

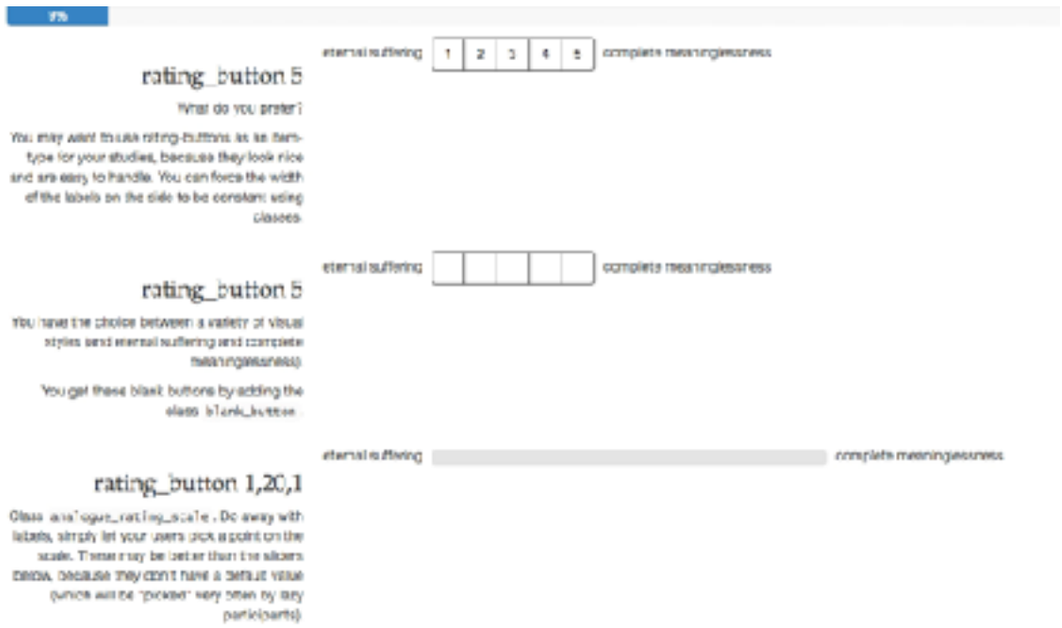
Principles



- Geared towards power users
  - learning curve is steeper (e.g. no interactive way to develop surveys)
  - doing complex things quickly is possible (e.g. surveys are spreadsheets that can be easily combined)

Principles

- Should work on all (modern) devices
  - Smartphones, tablets, computers, (consoles?)
    (i.e. on all operating systems and browser too, macOS, Windows, Linux, Android, etc.)
  - Fluid/responsive layout ->
- web-based, **no** app



y u no app

Pro App

Push Messages (now possible without apps)

Participants will believe that it works offline

Access to phone functionality (Gyrometer, photos, listening in on all activity, GPS-fencing etc.)

Sounds cool in grants?

“Branding”? Actually difficult

Contra App

Fairly difficult to release apps into the Apple AppStore and Google PlayStore on a budget

Study-specific apps -> too costly
vs. formr-App -> do you want that?

we’re a small developer team, don’t want to overstretch

serve more platforms (OSes, devices)

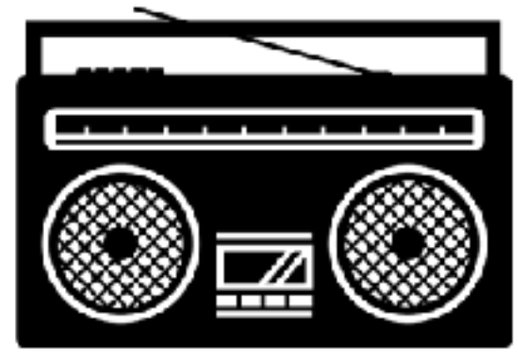
Future websites will seem like apps anyway

Websites can work offline too
(formr unfortunately doesn’t)

open source alternatives exist already

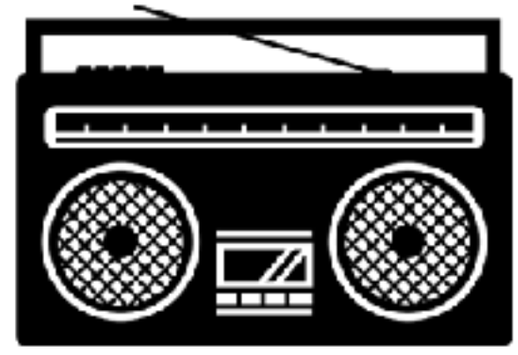
app installation is an extra hurdle for your participants

Principles



- **R** at every step
- Turing-complete (old-school GOTO programming)
- Boombox-metaphor
    - Play/Stop/Rewind/Skip/Record

Important features



- Surveys
- Send Email and SMS
- Automatic participant tracking for e.g. longitudinal studies, diaries
- Complex feedback (with R, knitr), Rmarkdown everywhere

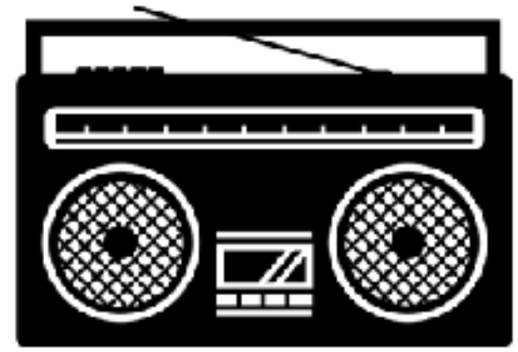
Intro to: **Survey**

- filled out in one sitting/session
- can refer to entered responses in this or previous surveys
- always part of a (sometimes simple) run, never used independently
- based on spreadsheets containing item definitions

Survey spreadsheet

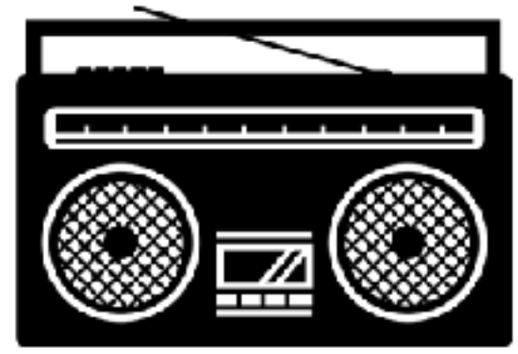
type	name	label	optional	showif
text	name	Please enter your name	*	
number 1,130,1	age	How old are you?		
mc agreement	emotional_stability1R	I worry a lot.		age >= 18
mc agreement	emotional_stability2R	I easily get nervous and unsure of myself.		age >= 18
mc agreement	emotional_stability3	I am relaxed and not easily stressed.		age >= 18

Choices sheet



list_name	name	label
agreement	1	disagree completely
agreement	2	rather disagree
agreement	3	neither agree nor disagree
agreement	4	rather agree
agreement	5	agree completely

(R) Markdown



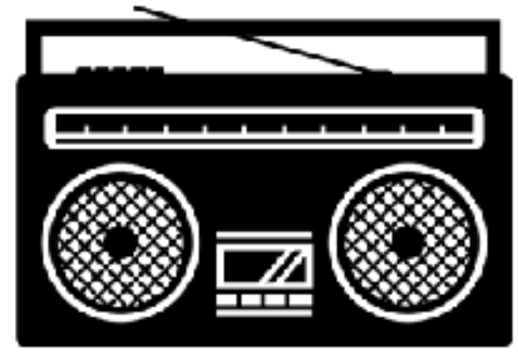
- Markdown: write *plaintext*, so that you have readable “source code”, but can format text

- Rmarkdown (knitr) allows us to blend R code with Markdown (similar to iPython notebooks)

  - established to document statistical reports, some even use it for complete scientific manuscripts

- In formr it can be used anywhere, where the user will see text (item/choice, labels, emails, feedback, …)

(R) Markdown



```
# Markdown

We can make text __fat__ and *italic*.

We can [link stuff](https://formr.org).

We can add images. ![boombox]
(boombox.jpg)

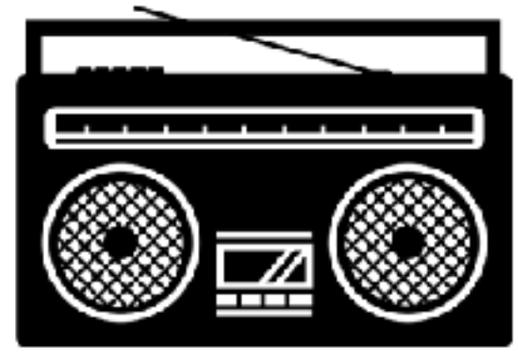
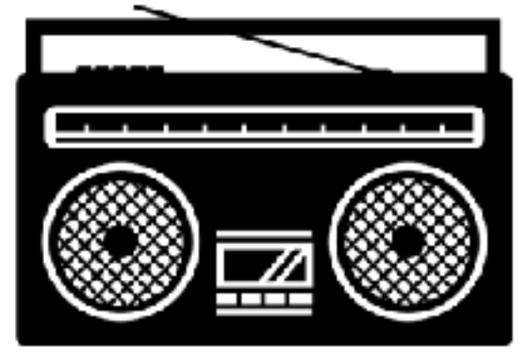
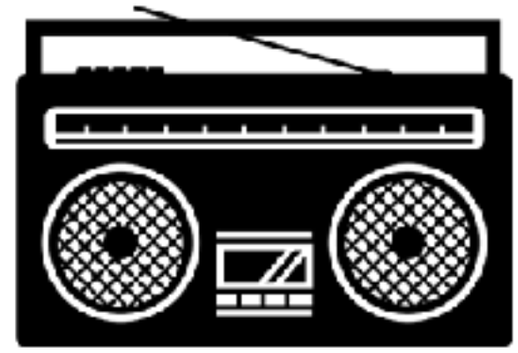
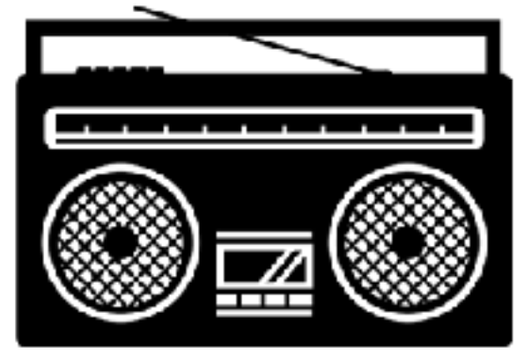
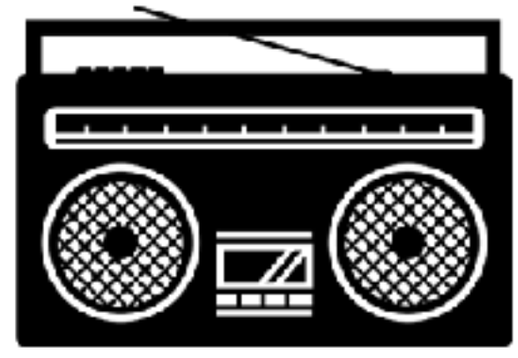
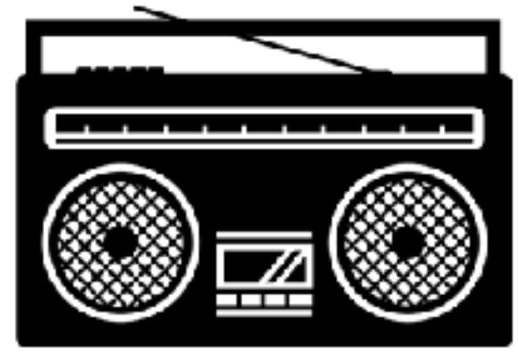
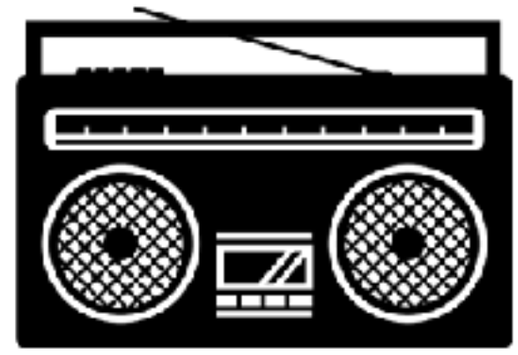
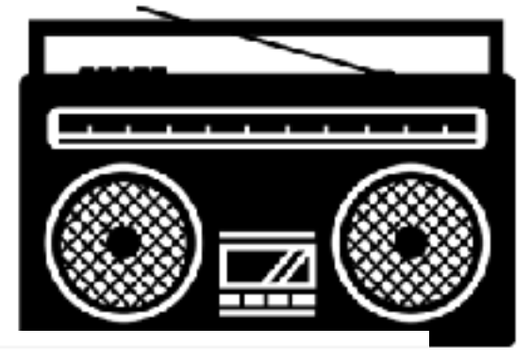
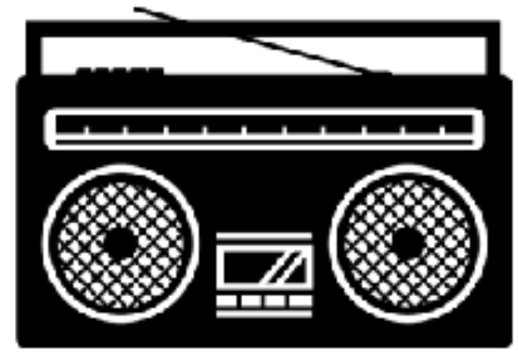
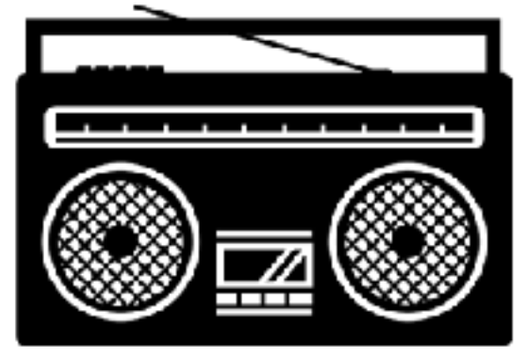
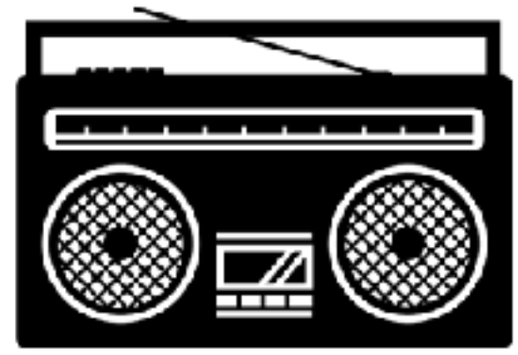
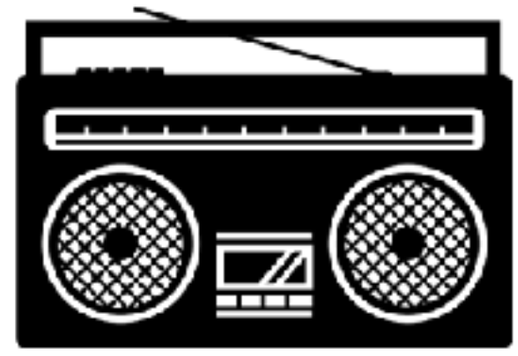
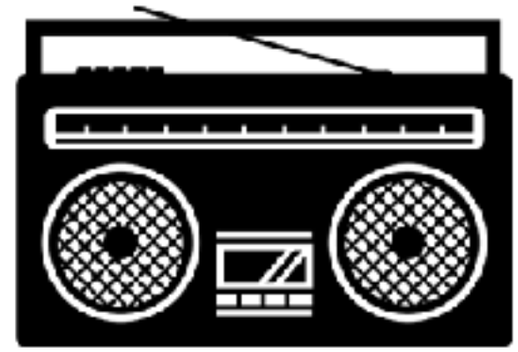
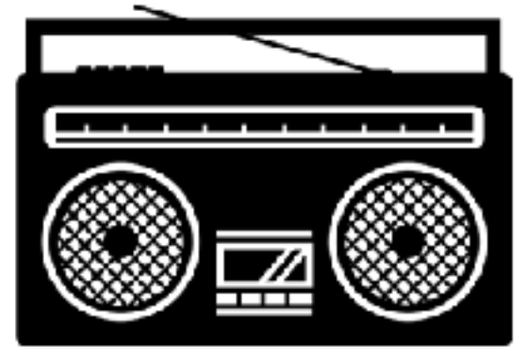
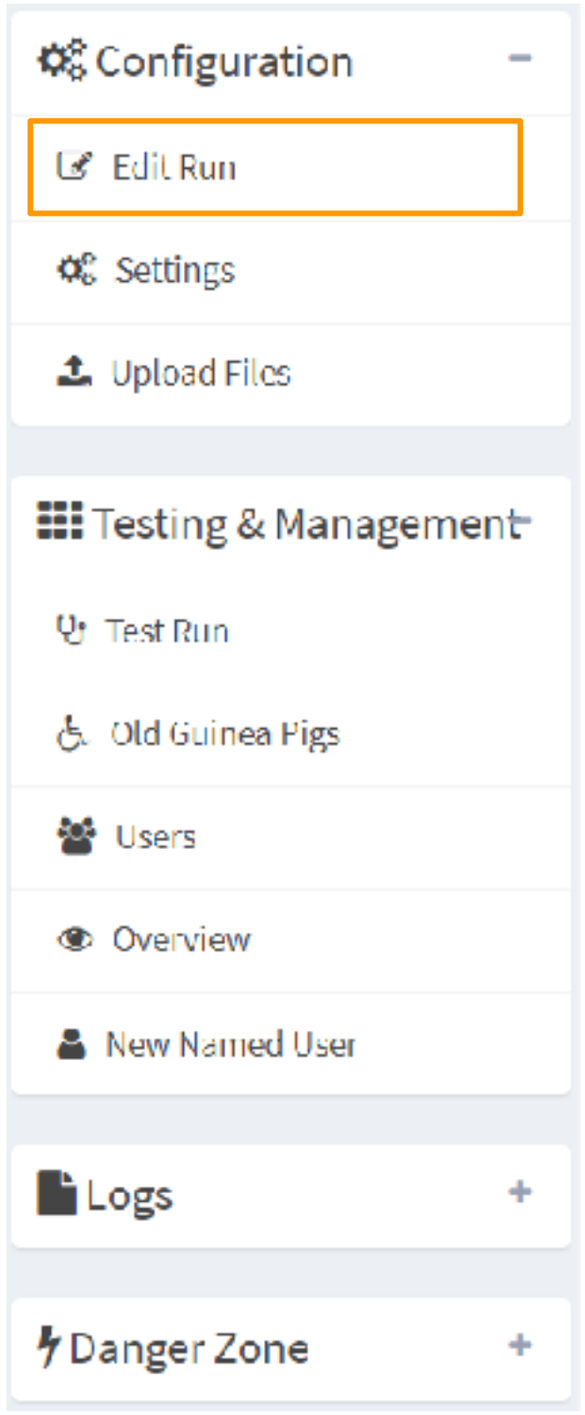
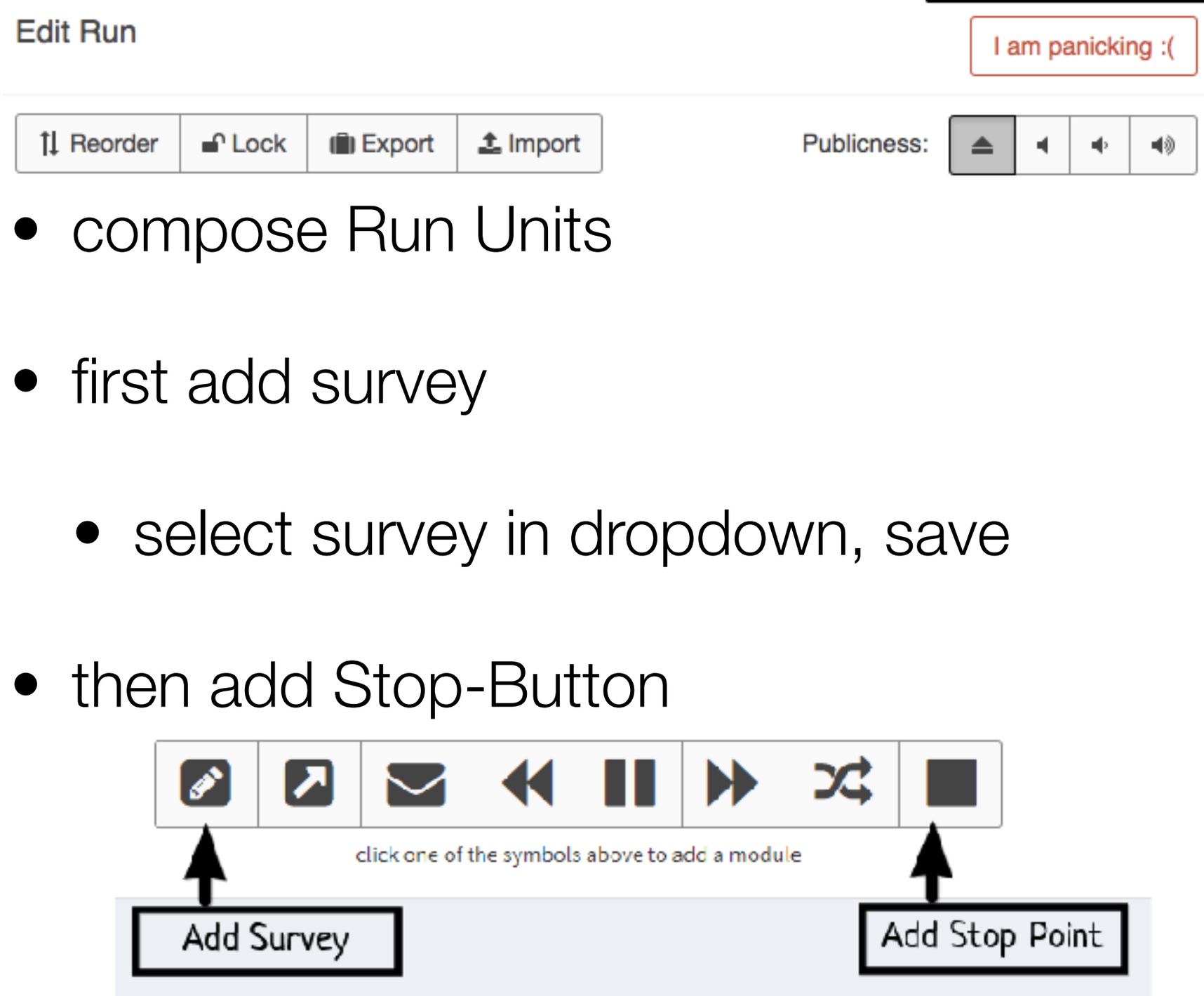
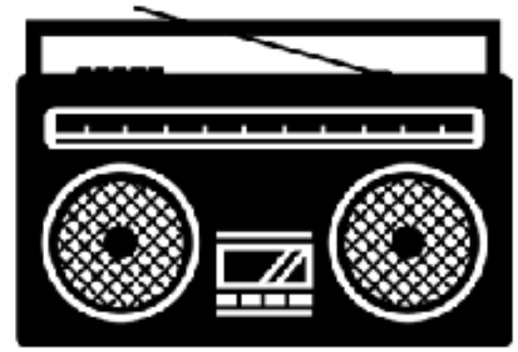
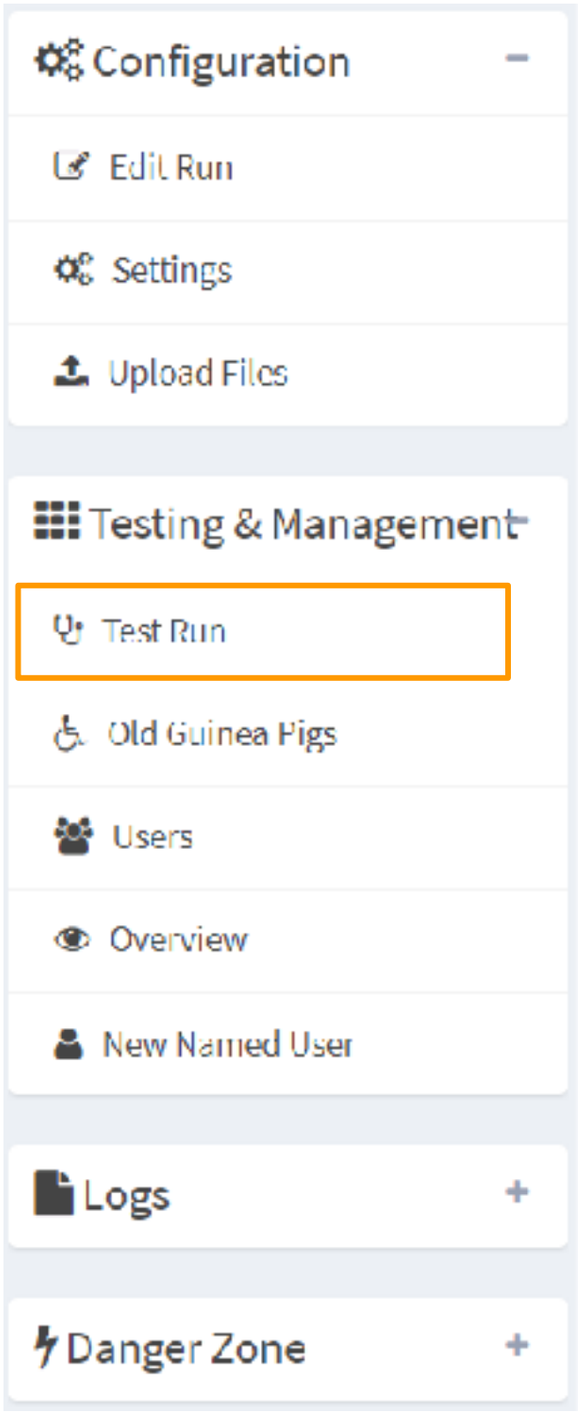
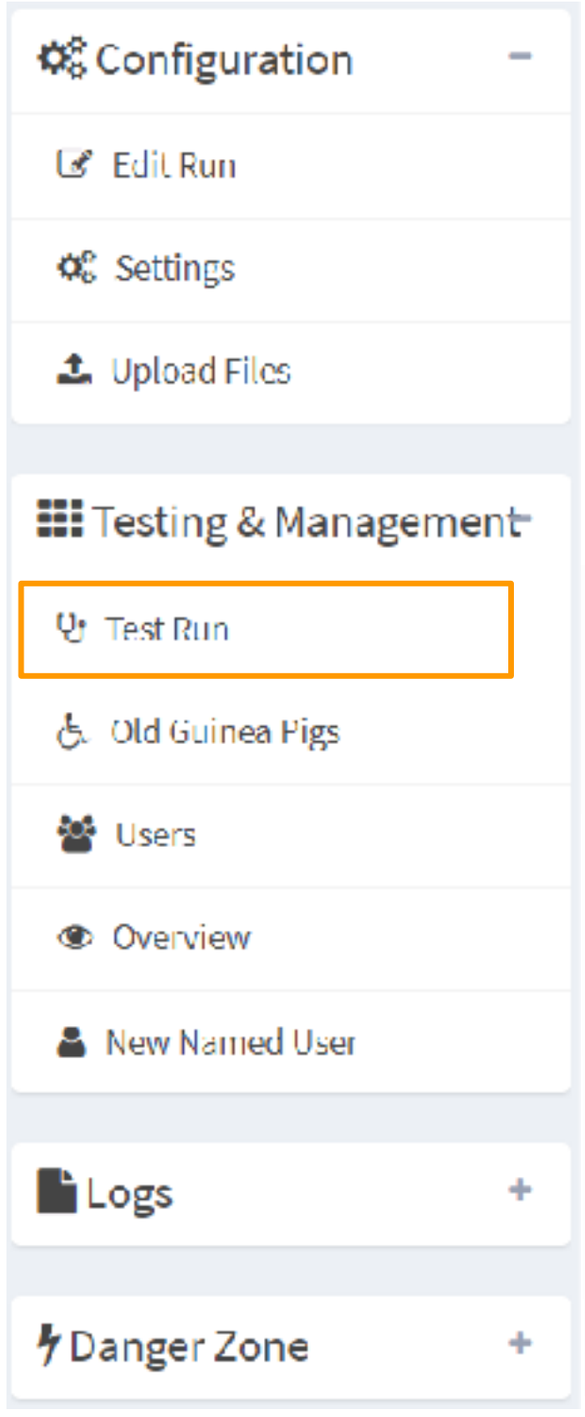
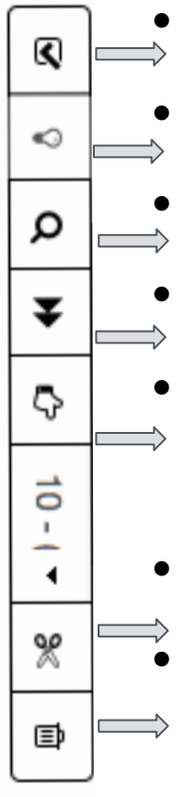
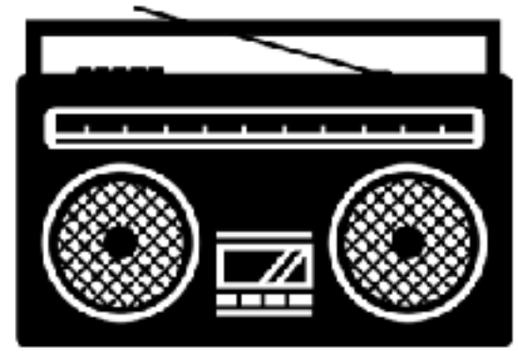
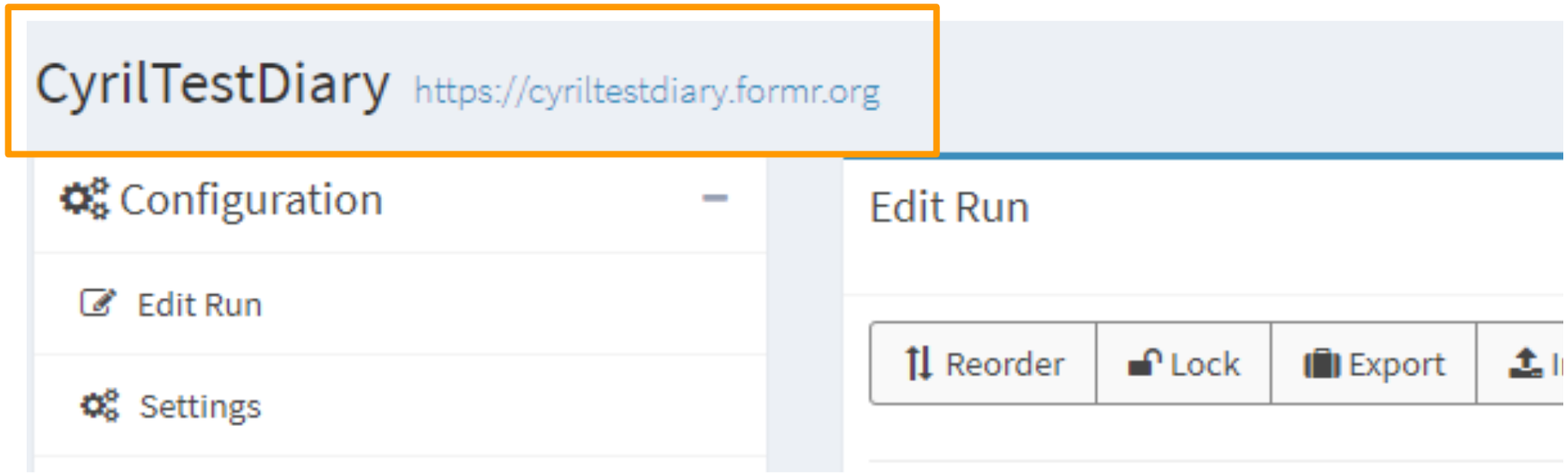
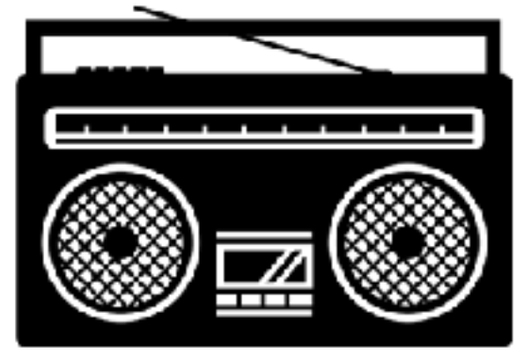
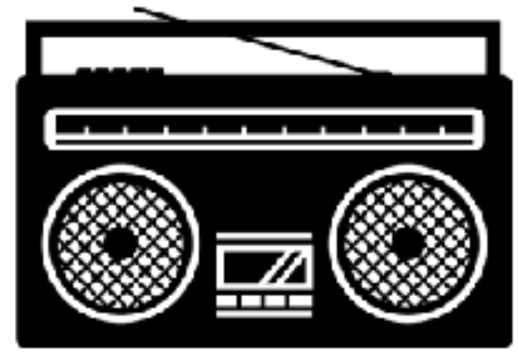
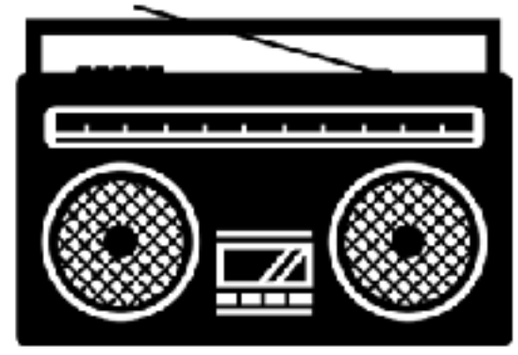
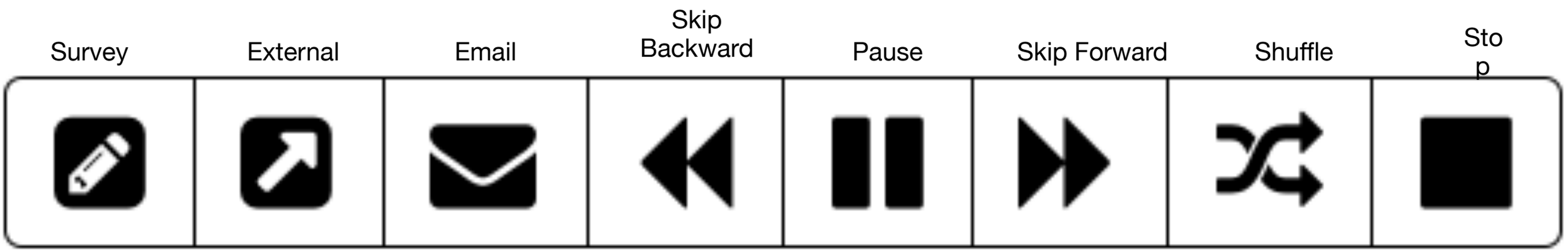
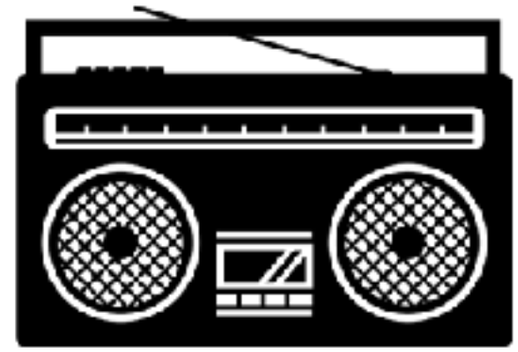
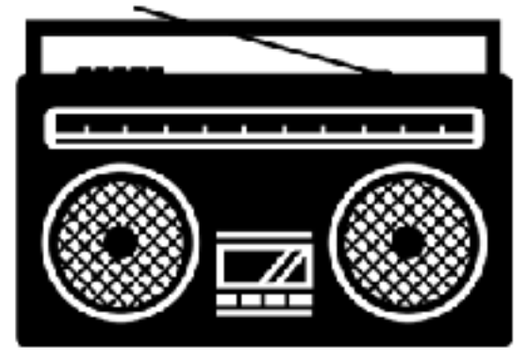
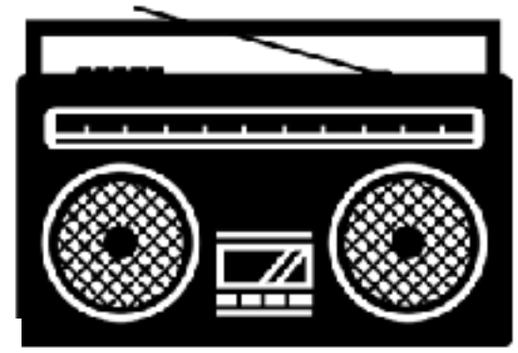
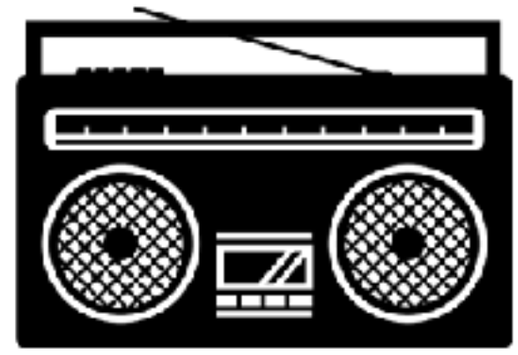
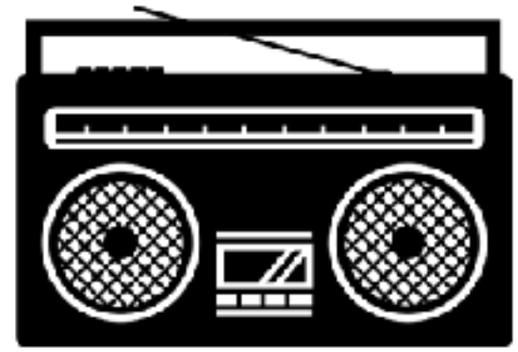
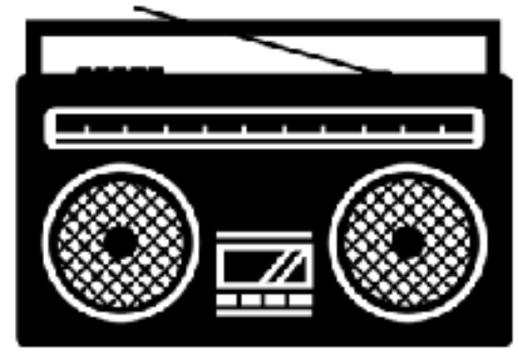
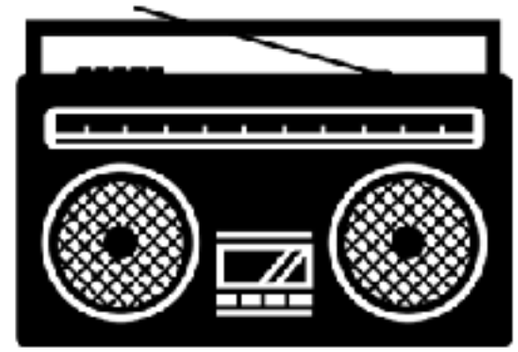
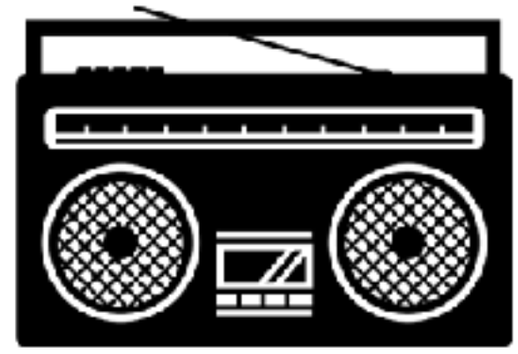
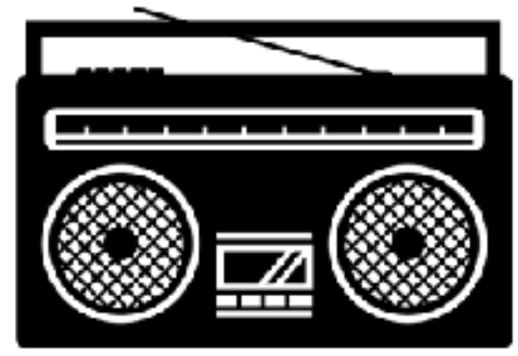
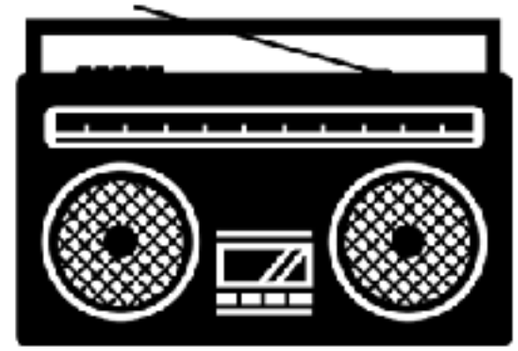
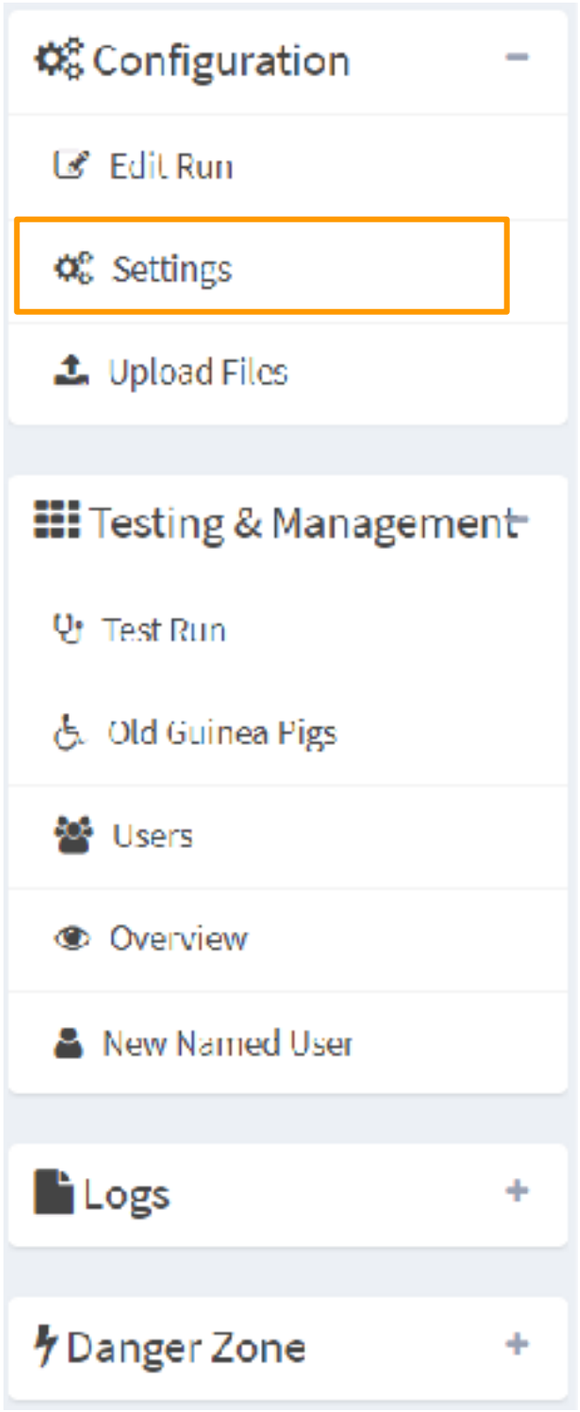
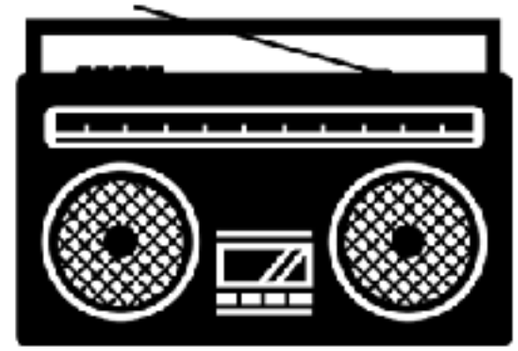
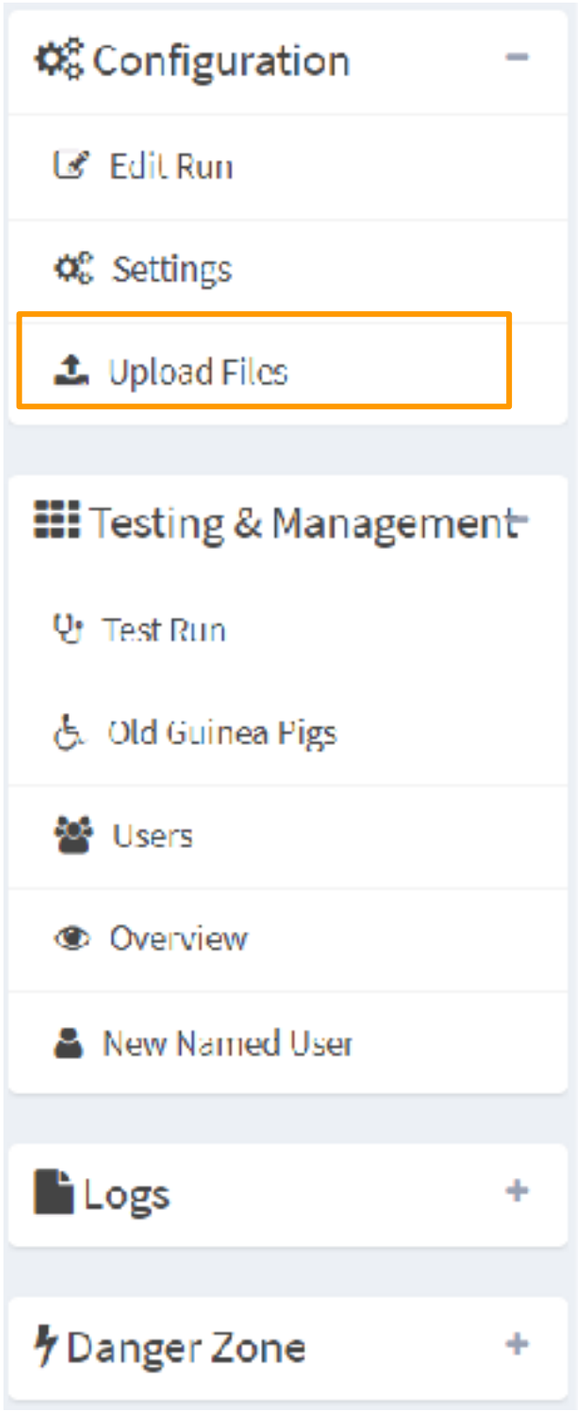
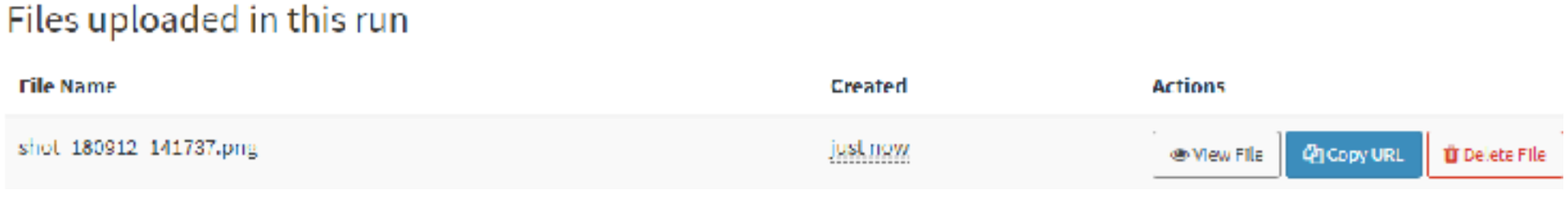
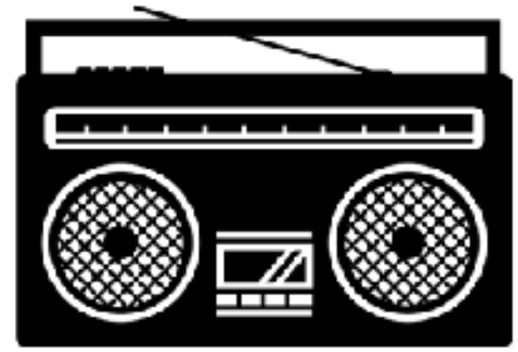
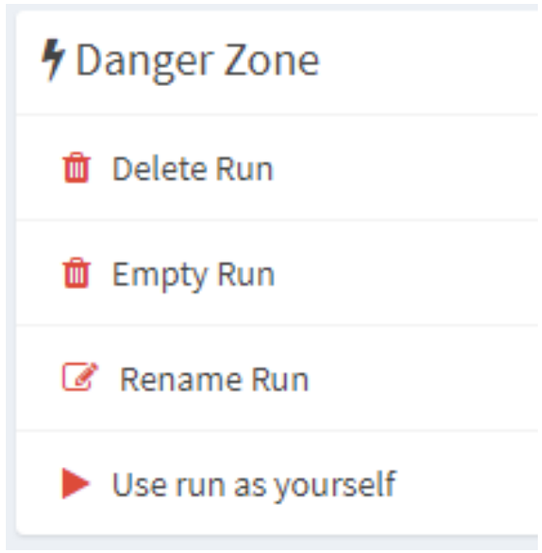
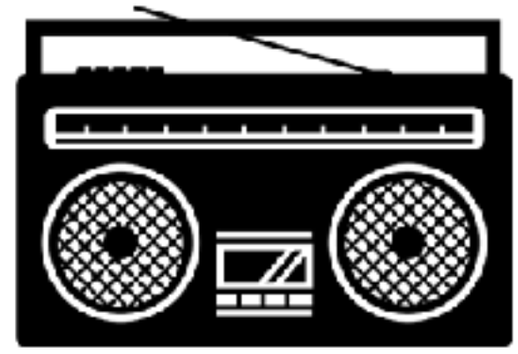
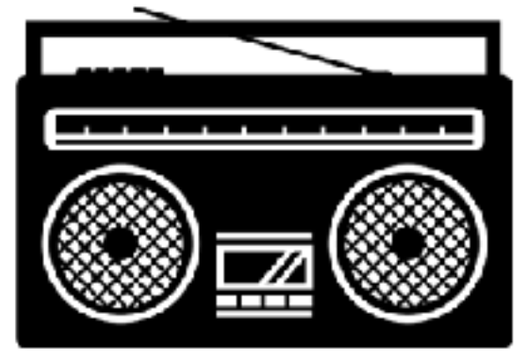
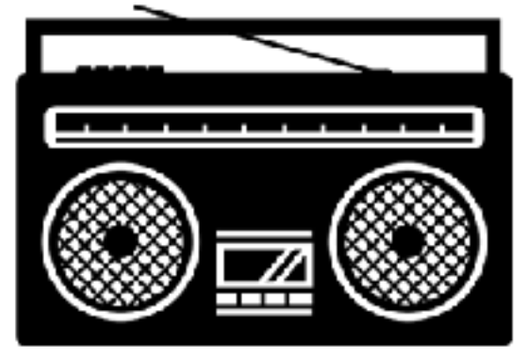
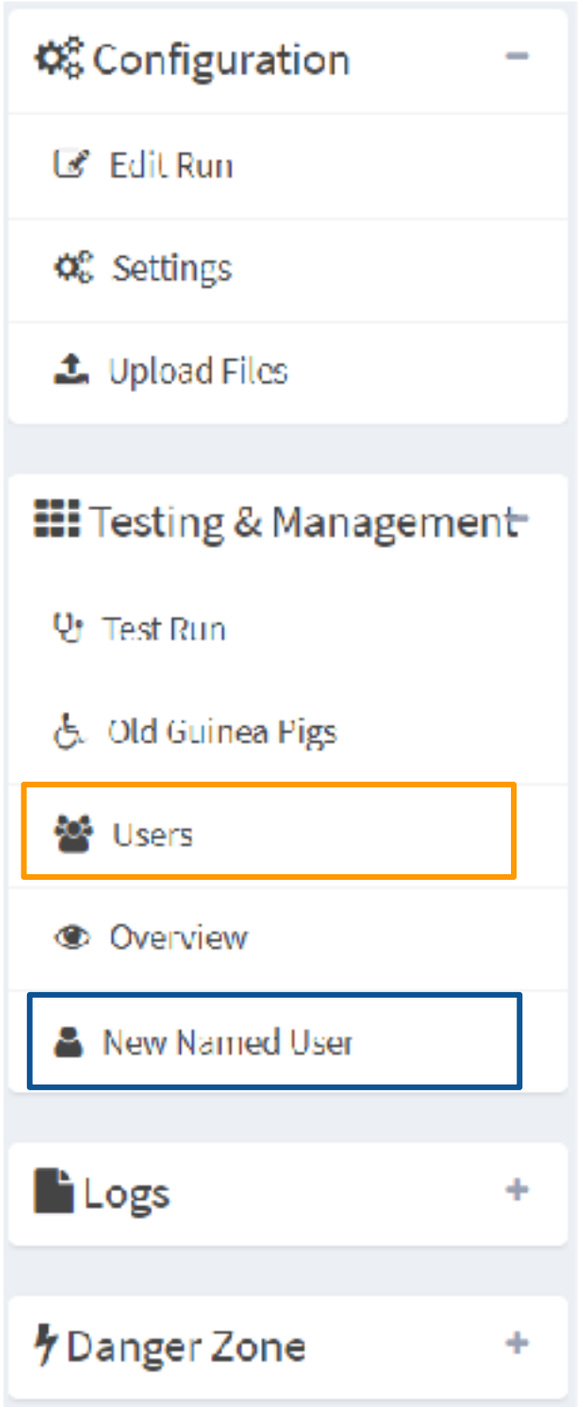
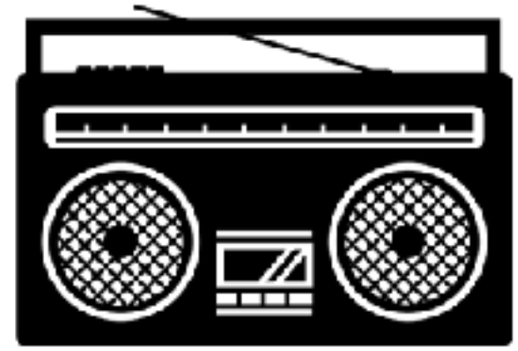
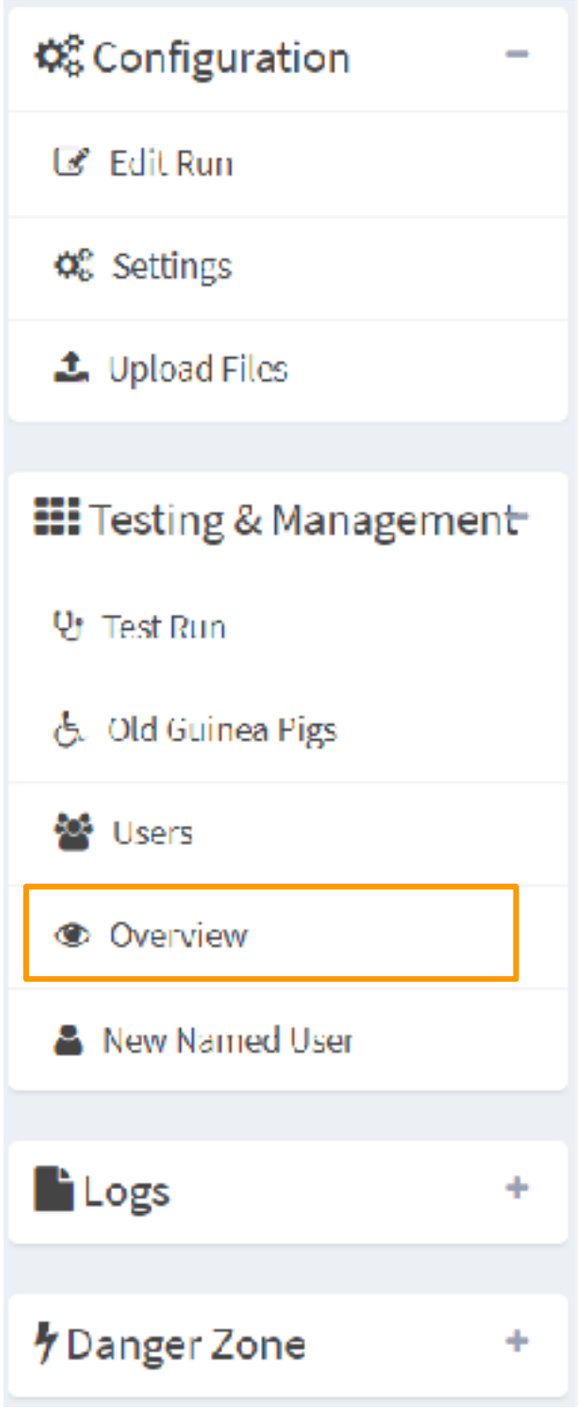
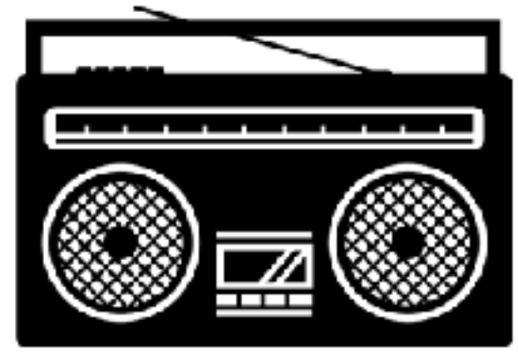
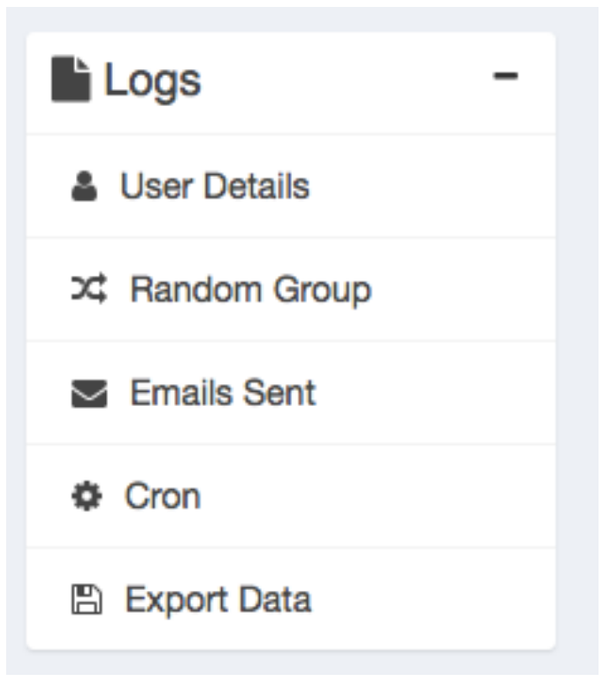
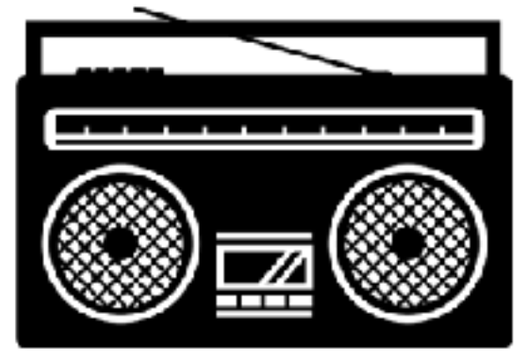
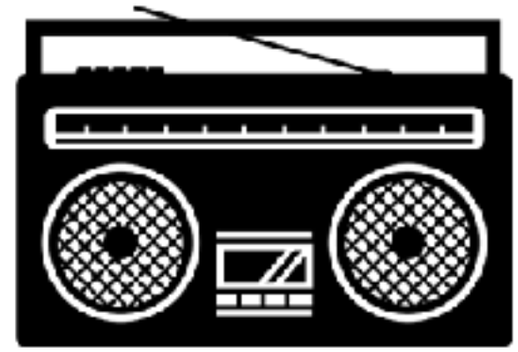
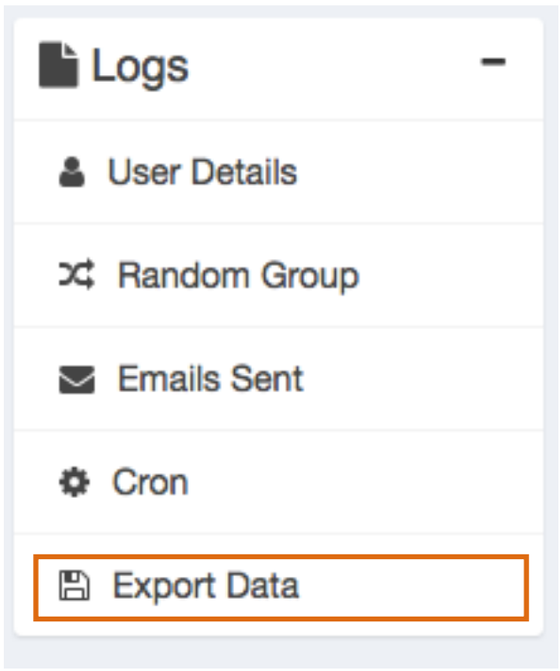
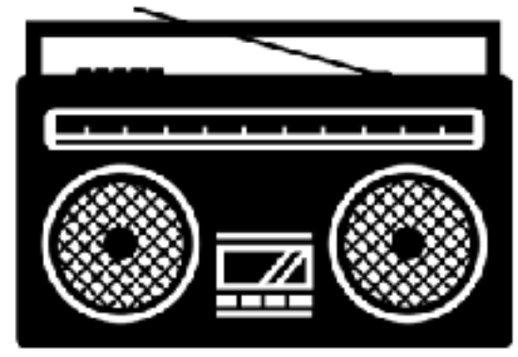
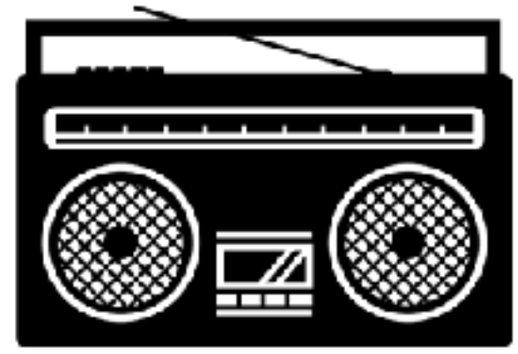
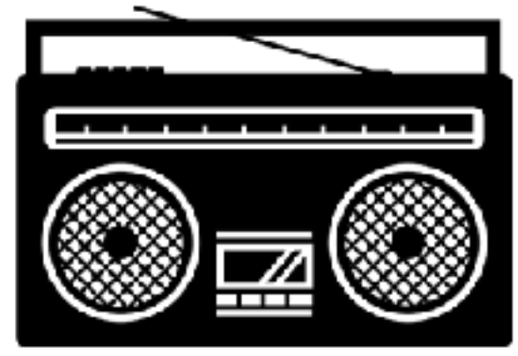
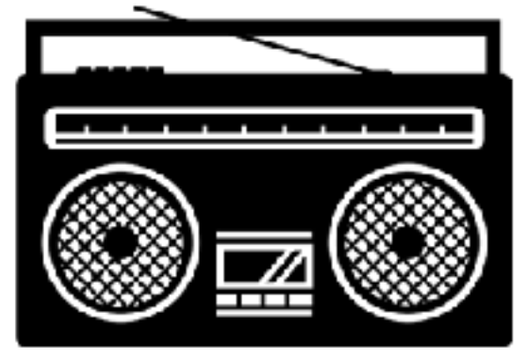
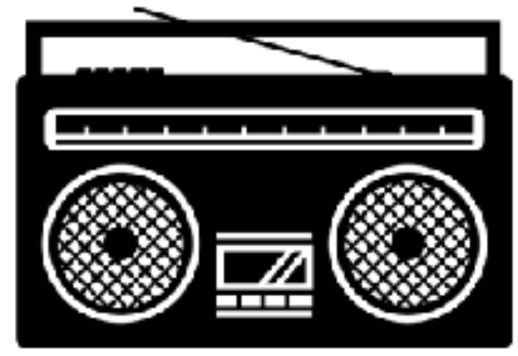
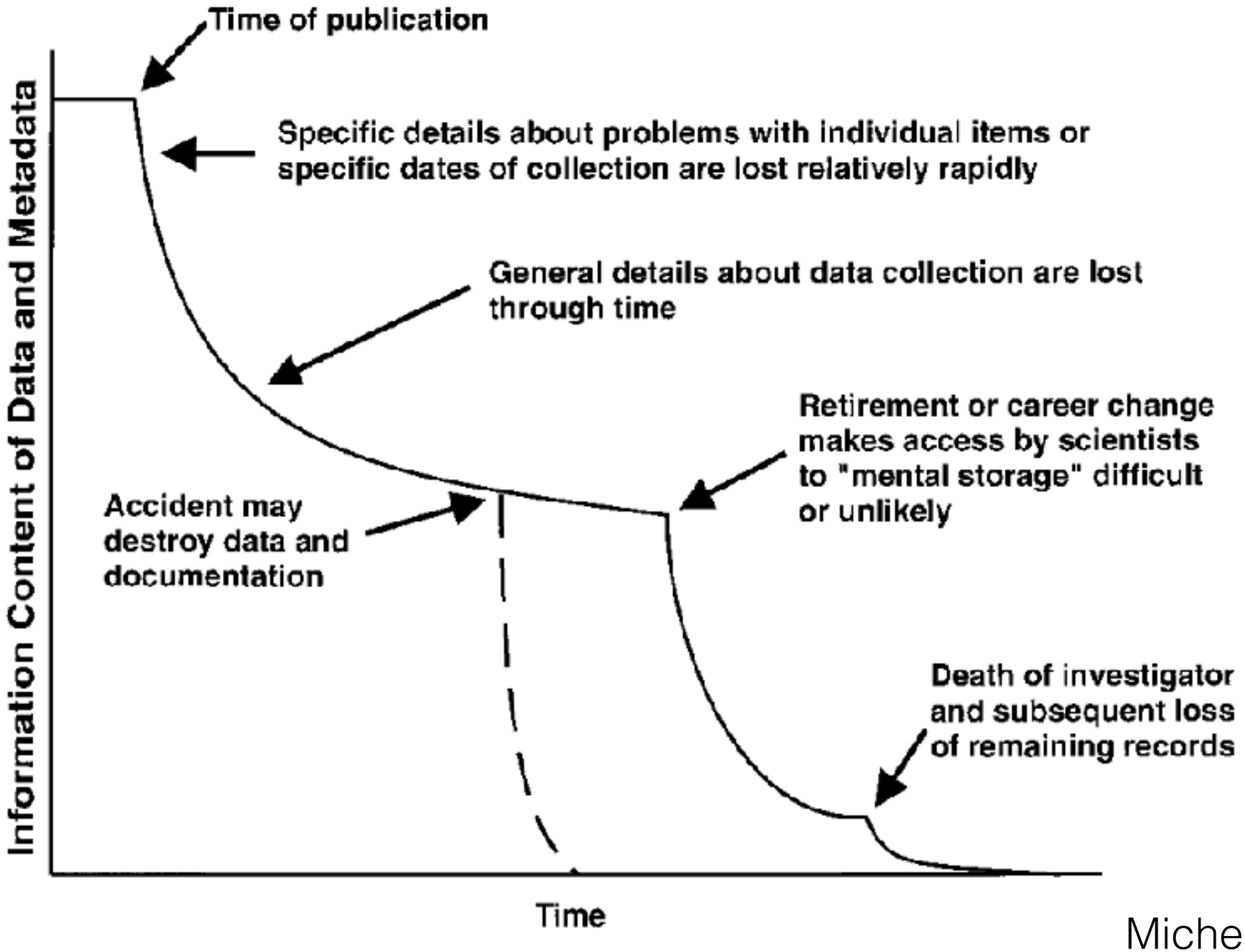
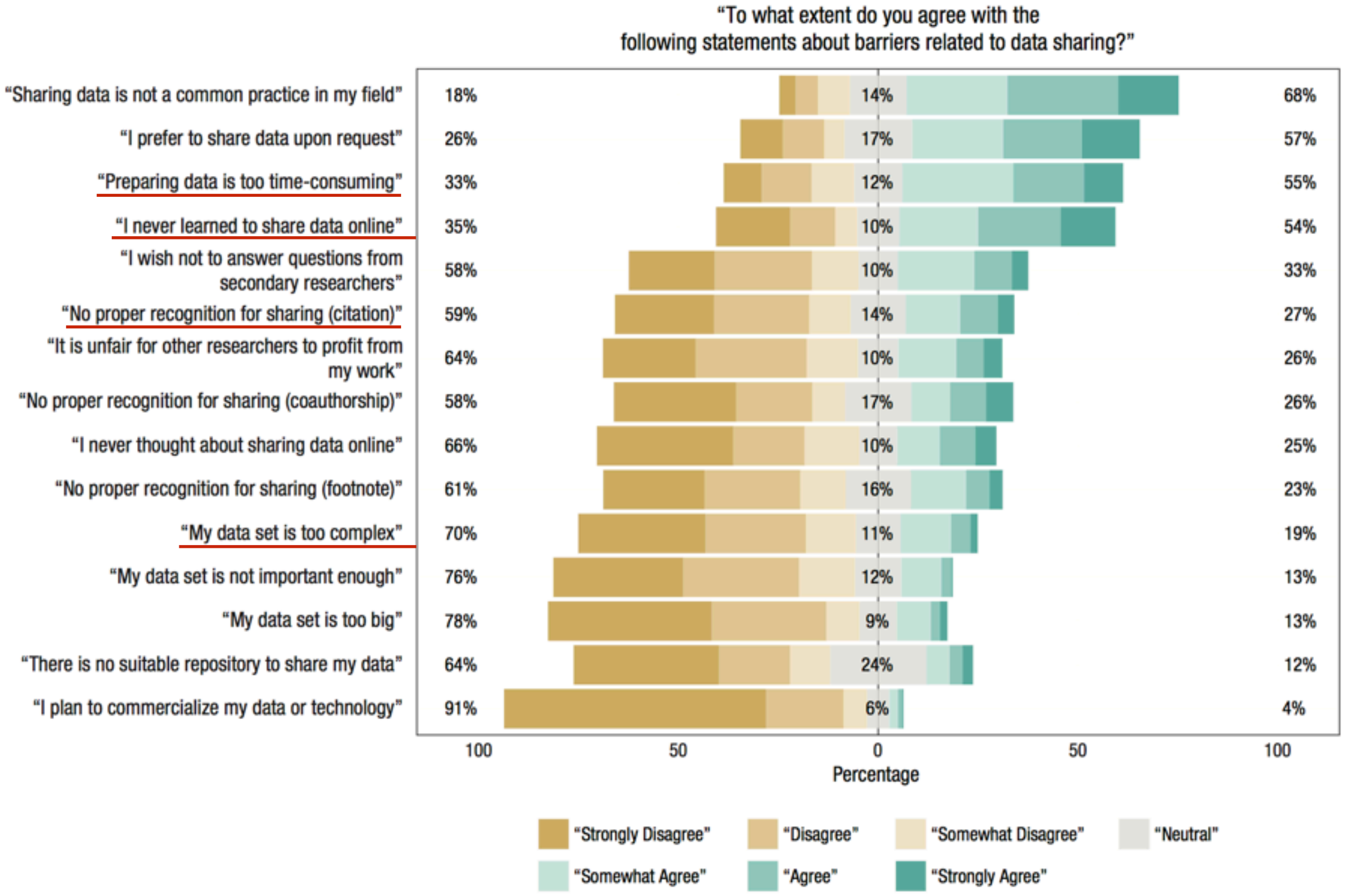
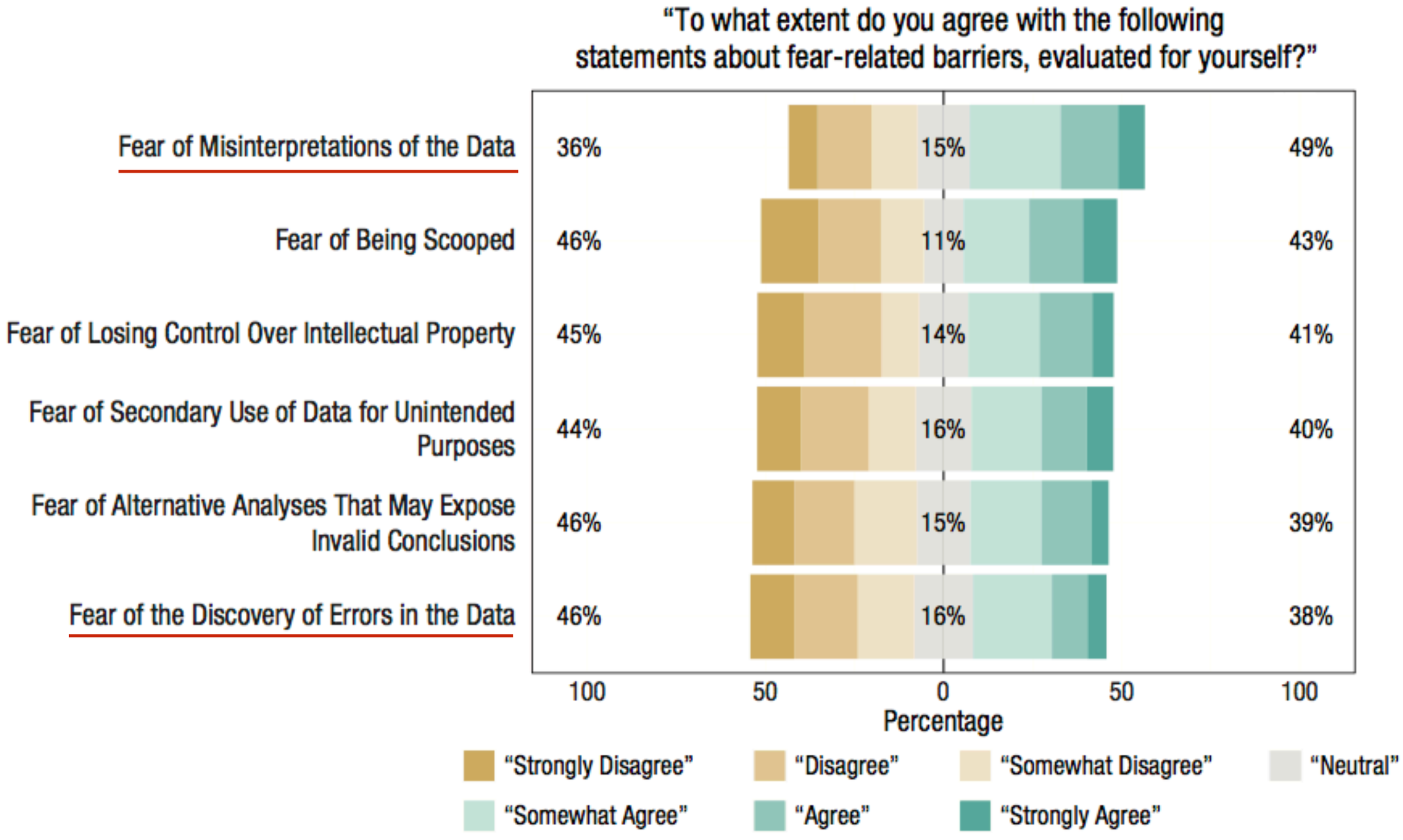
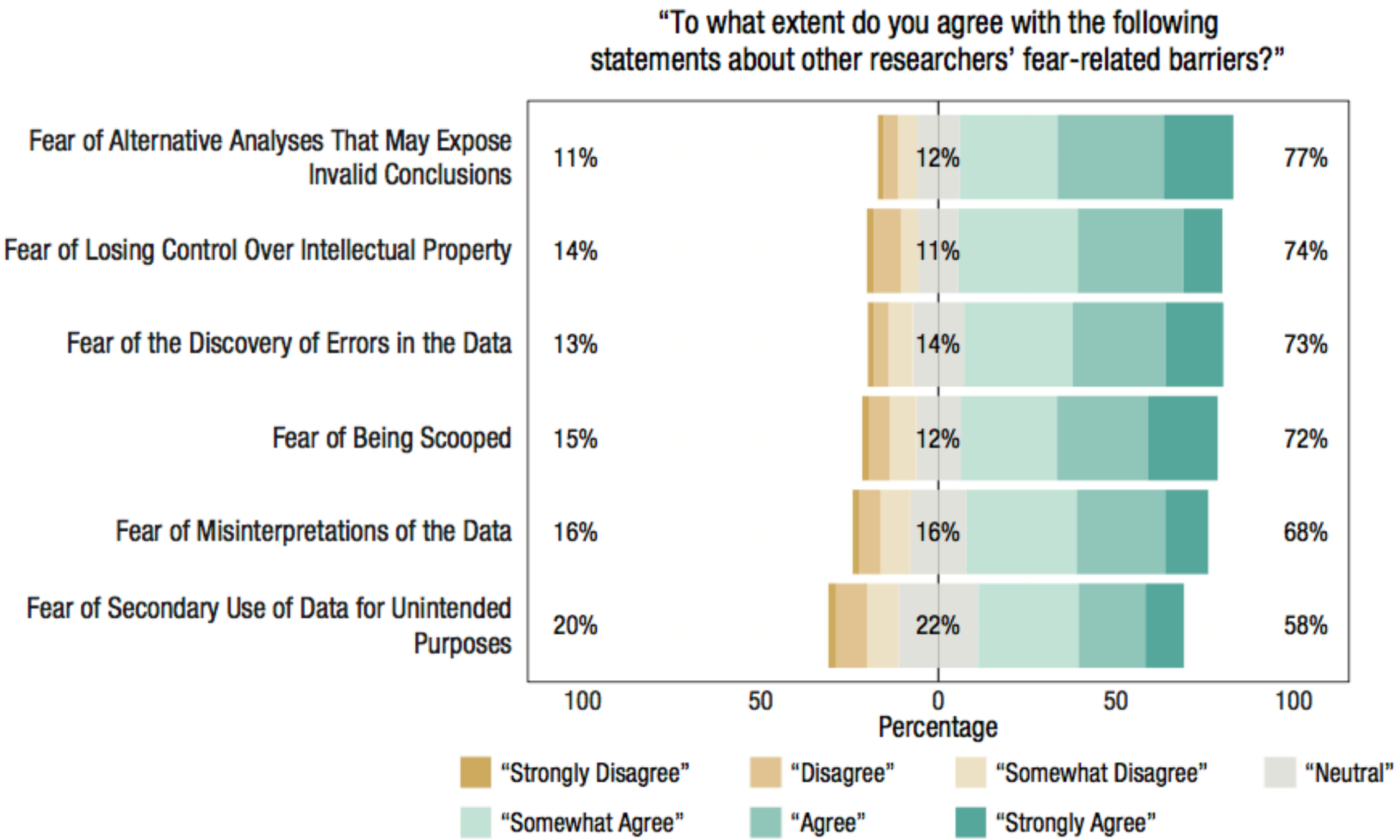
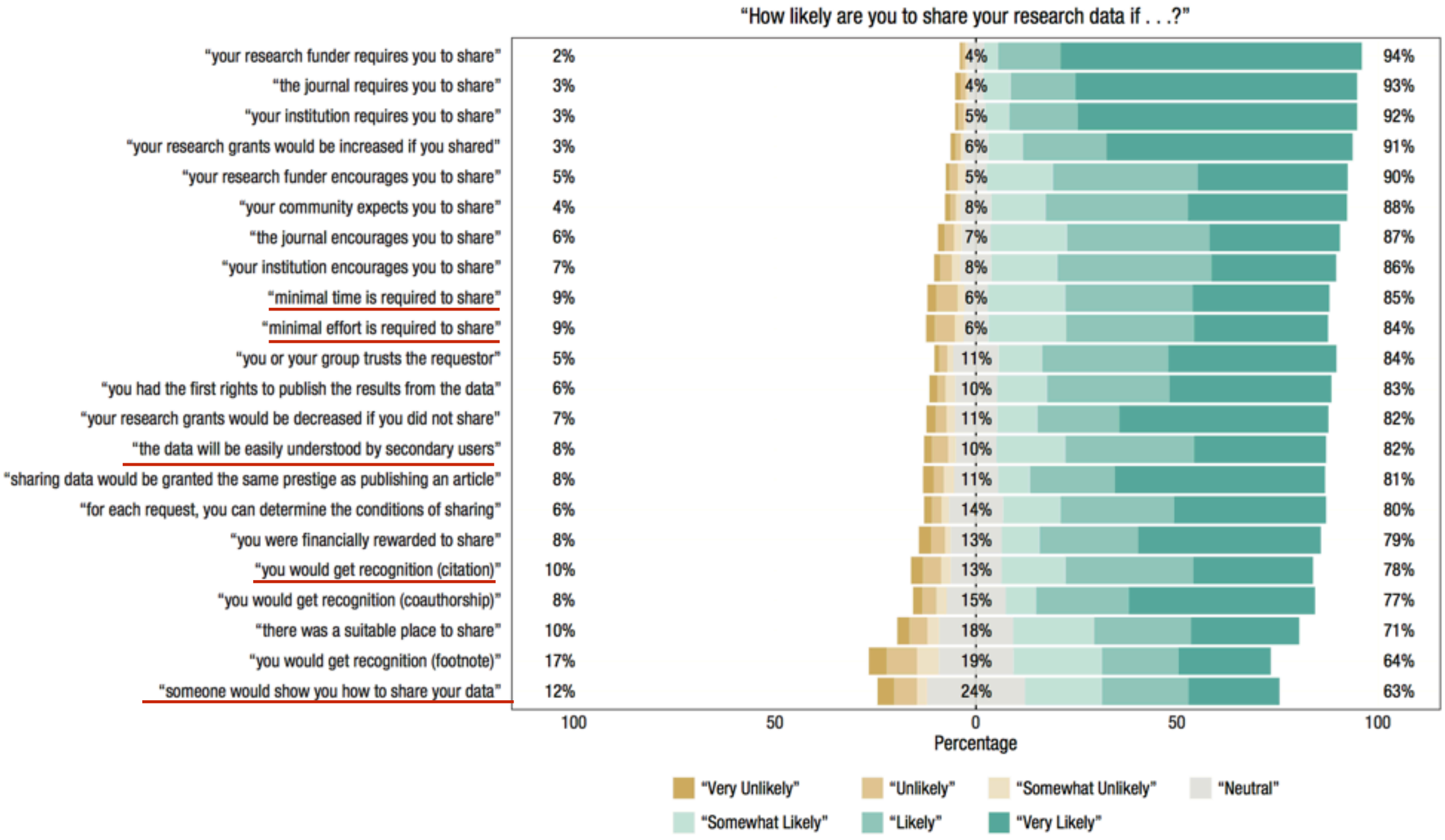
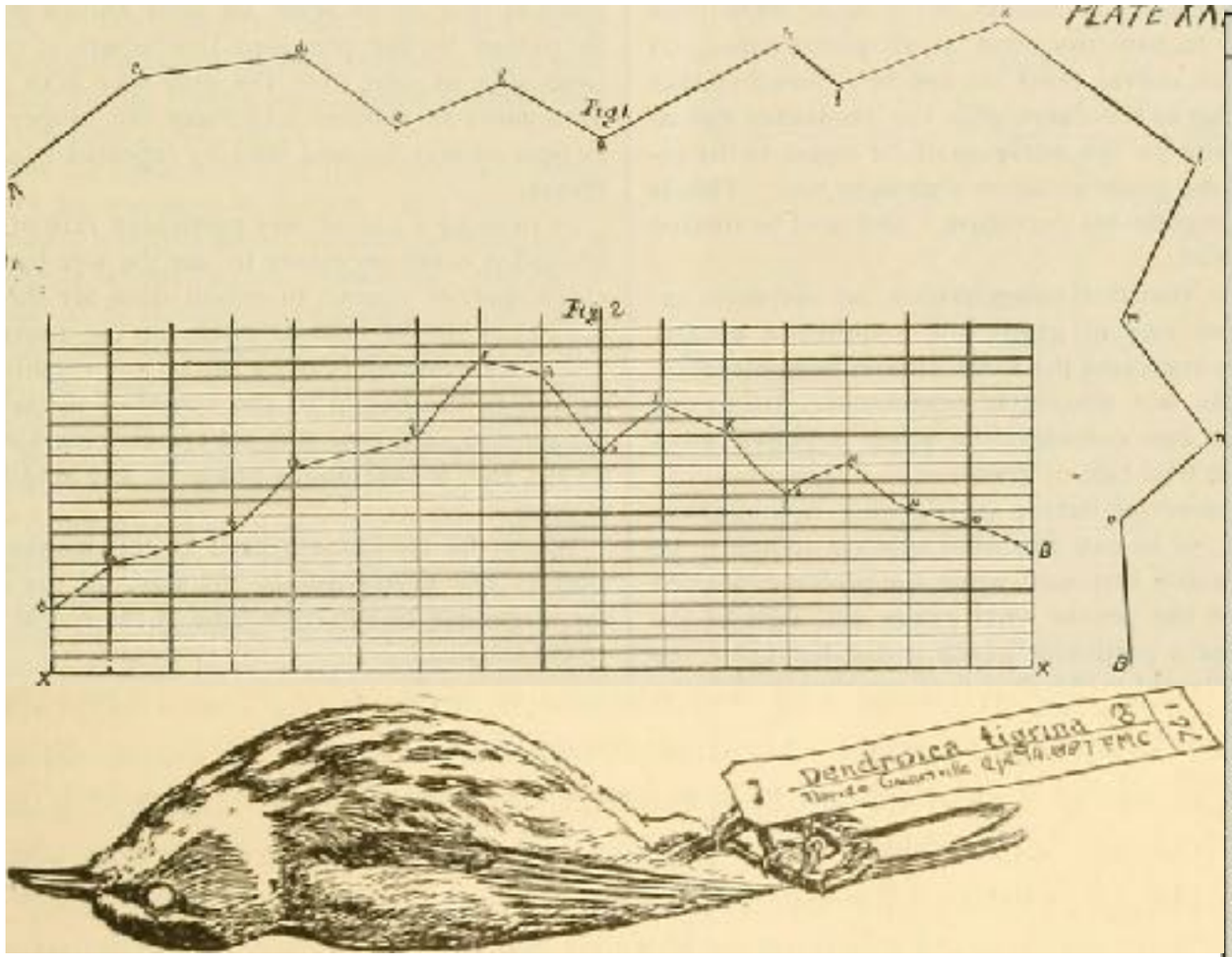
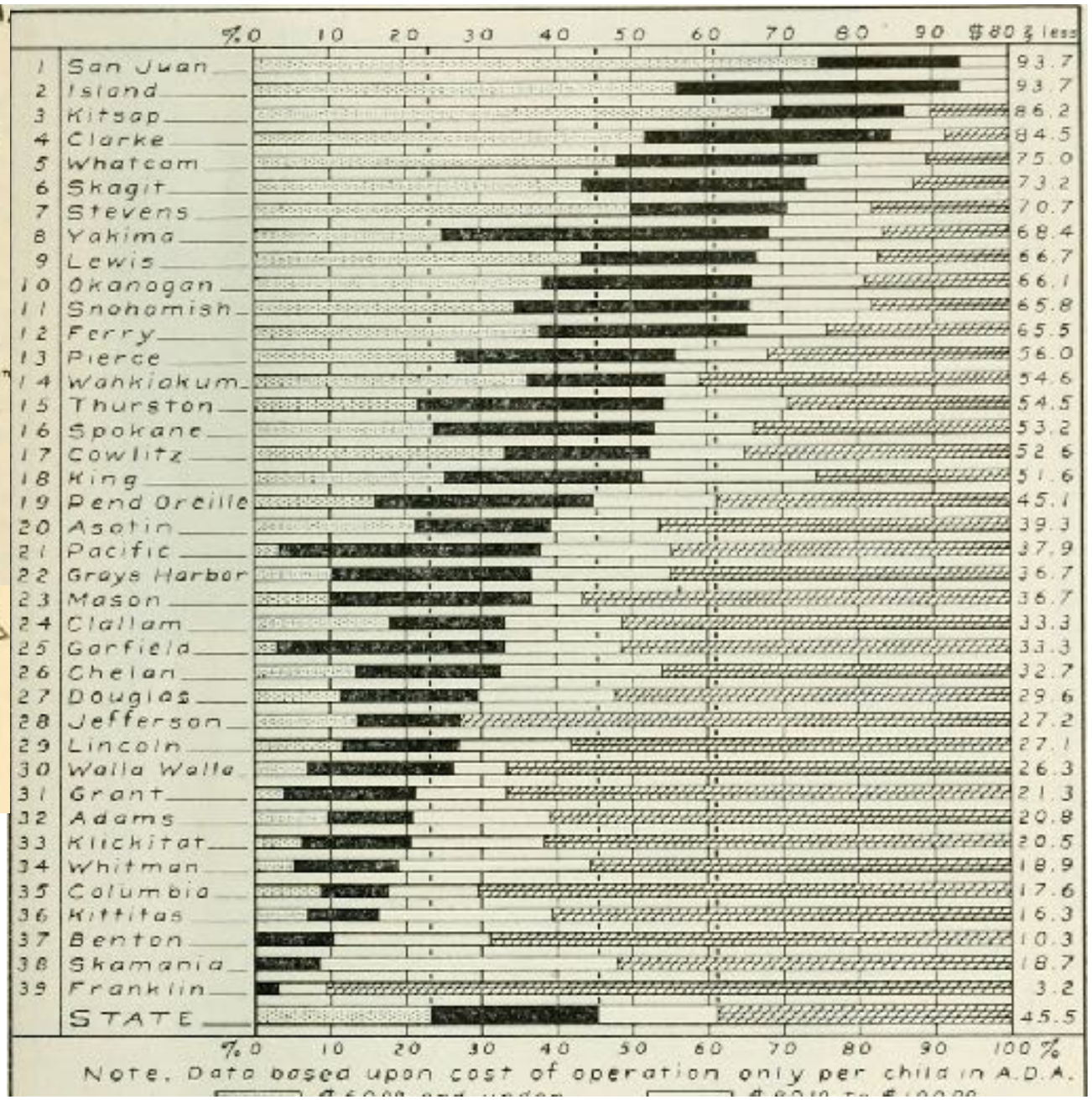
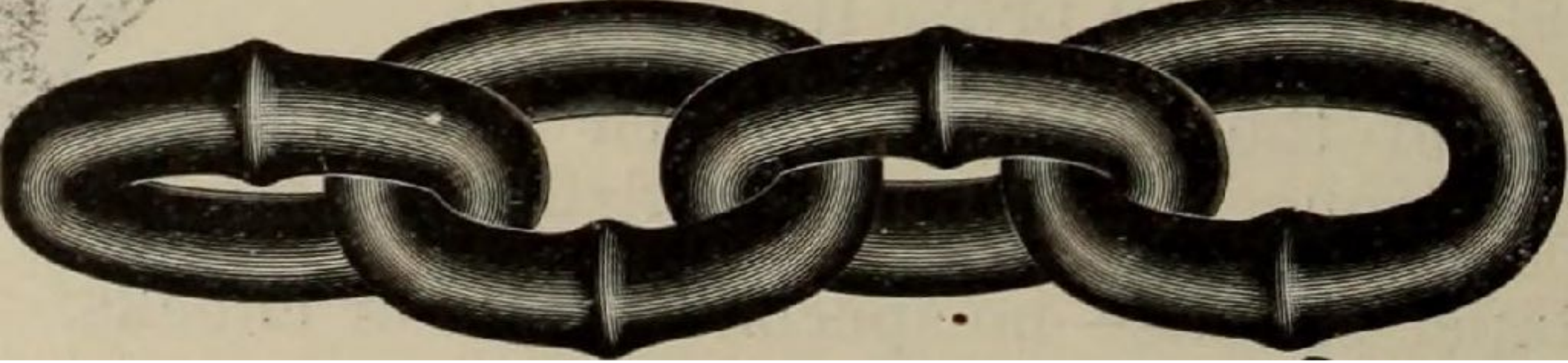
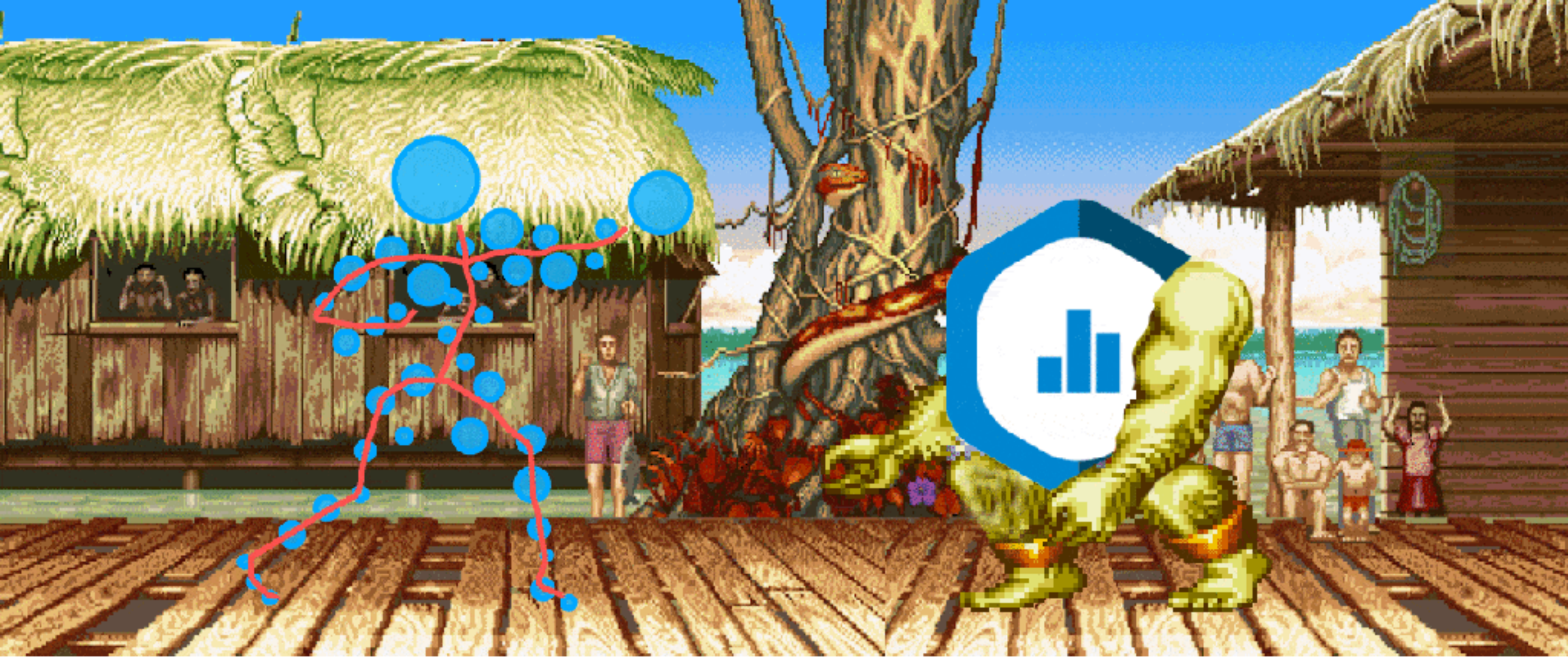
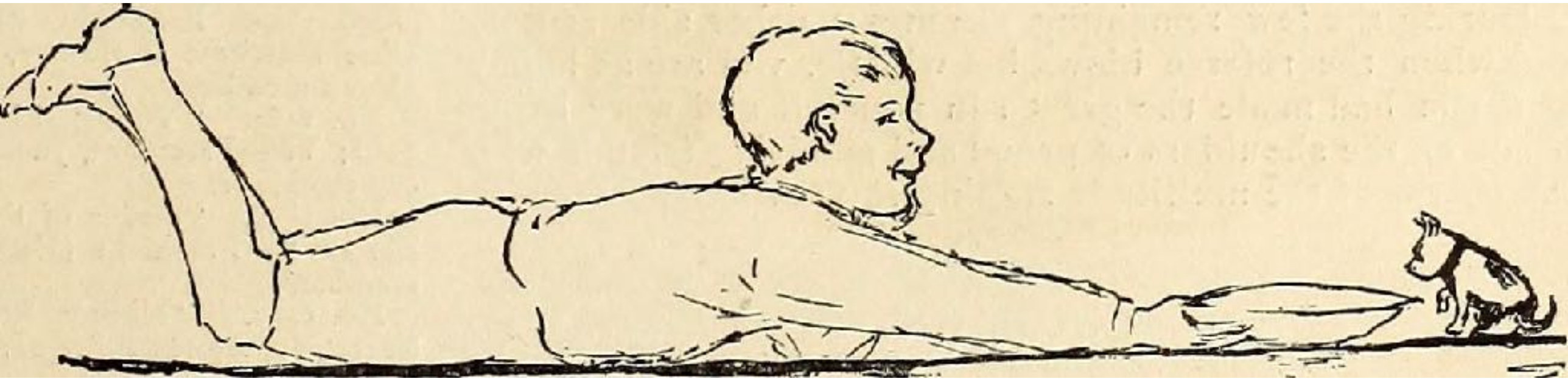
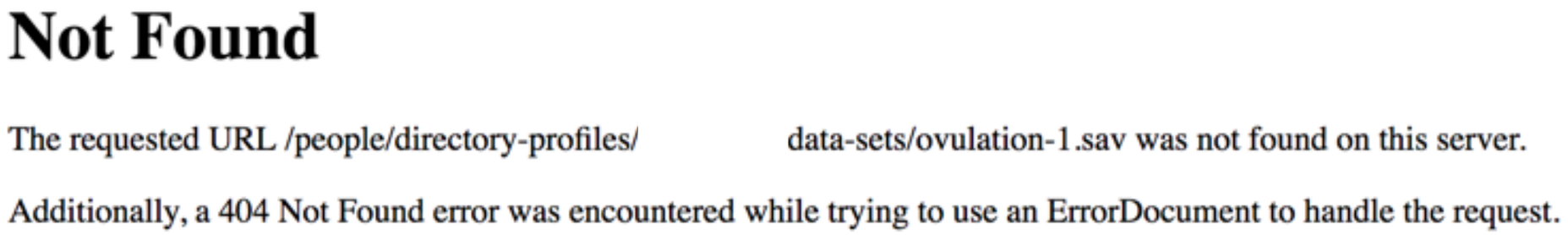
```{r include = F}
library(lubridate)
yesterday = today() - days(1)
```

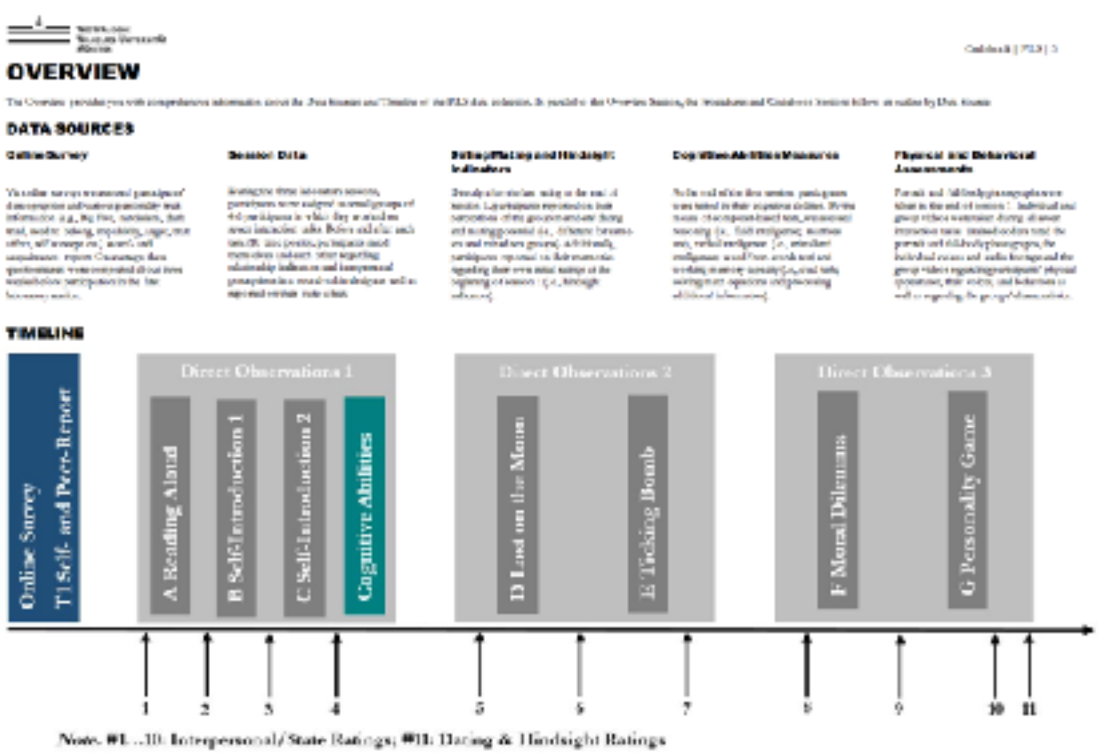
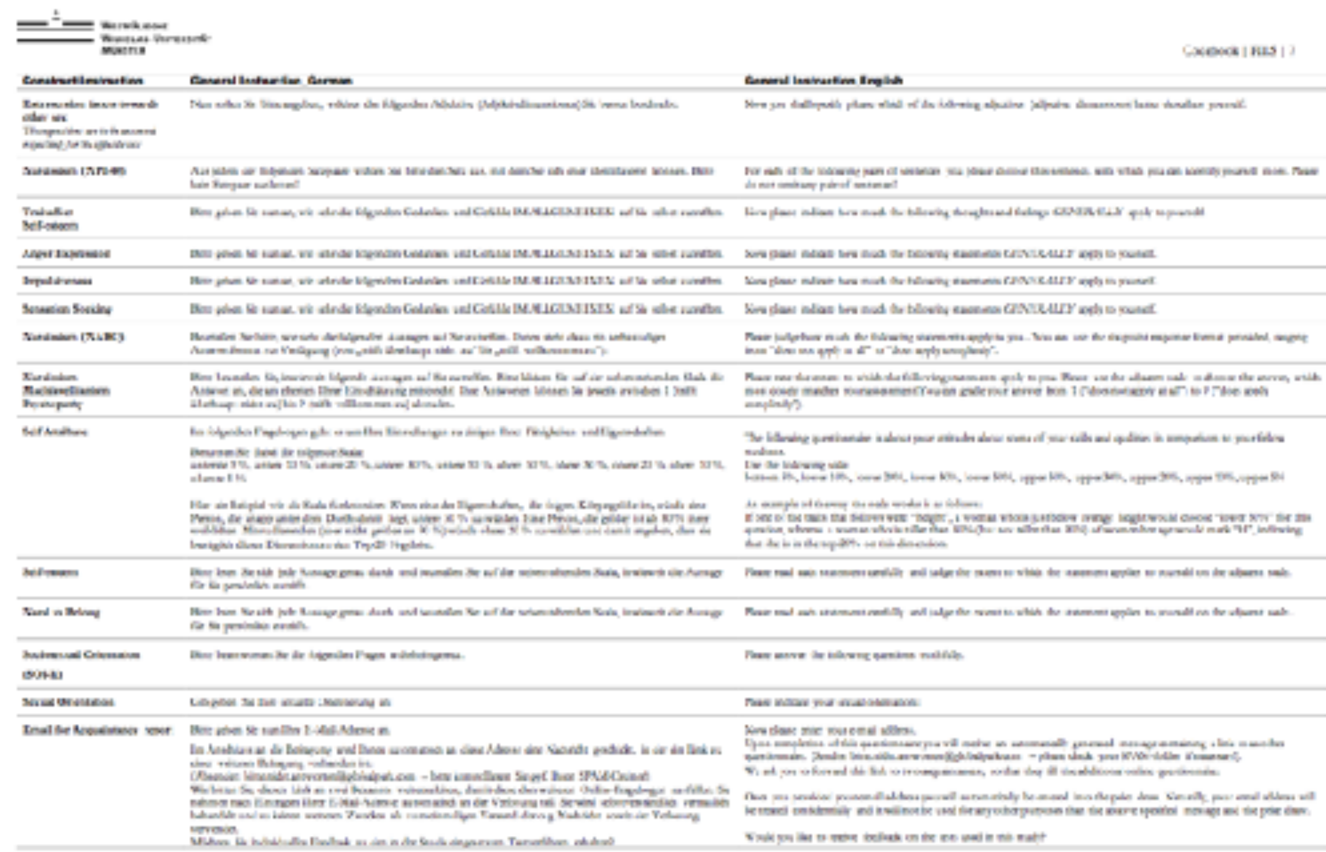
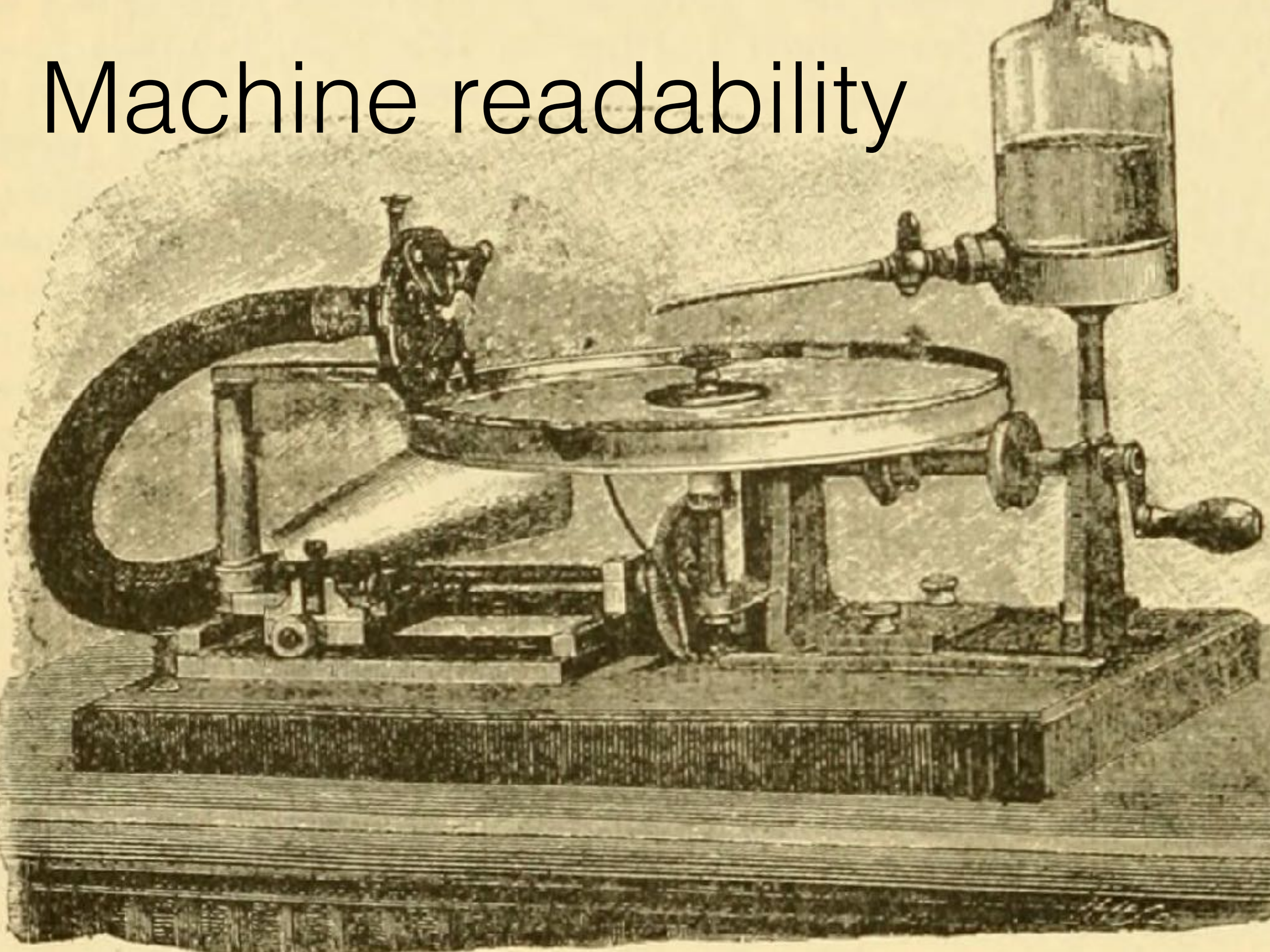
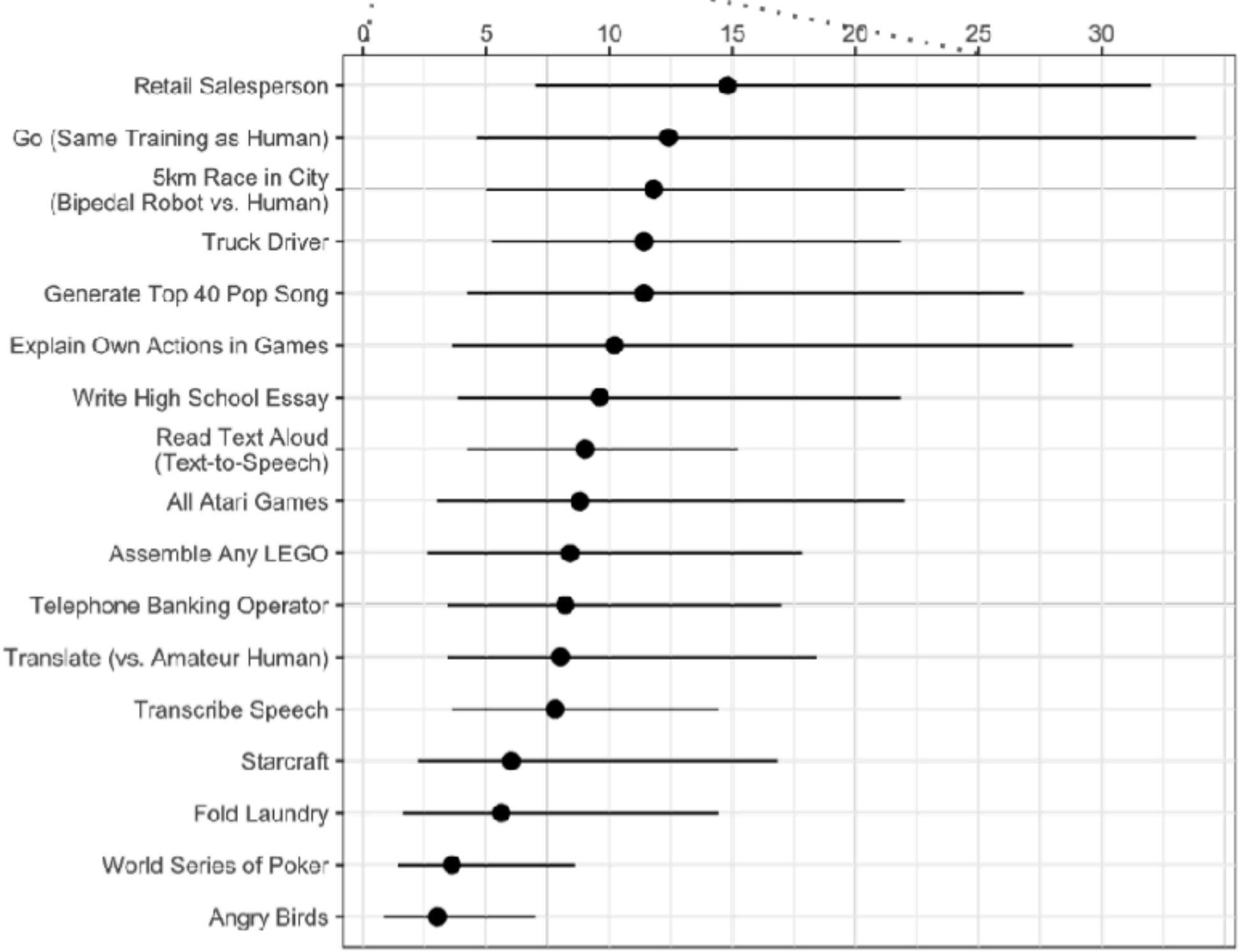
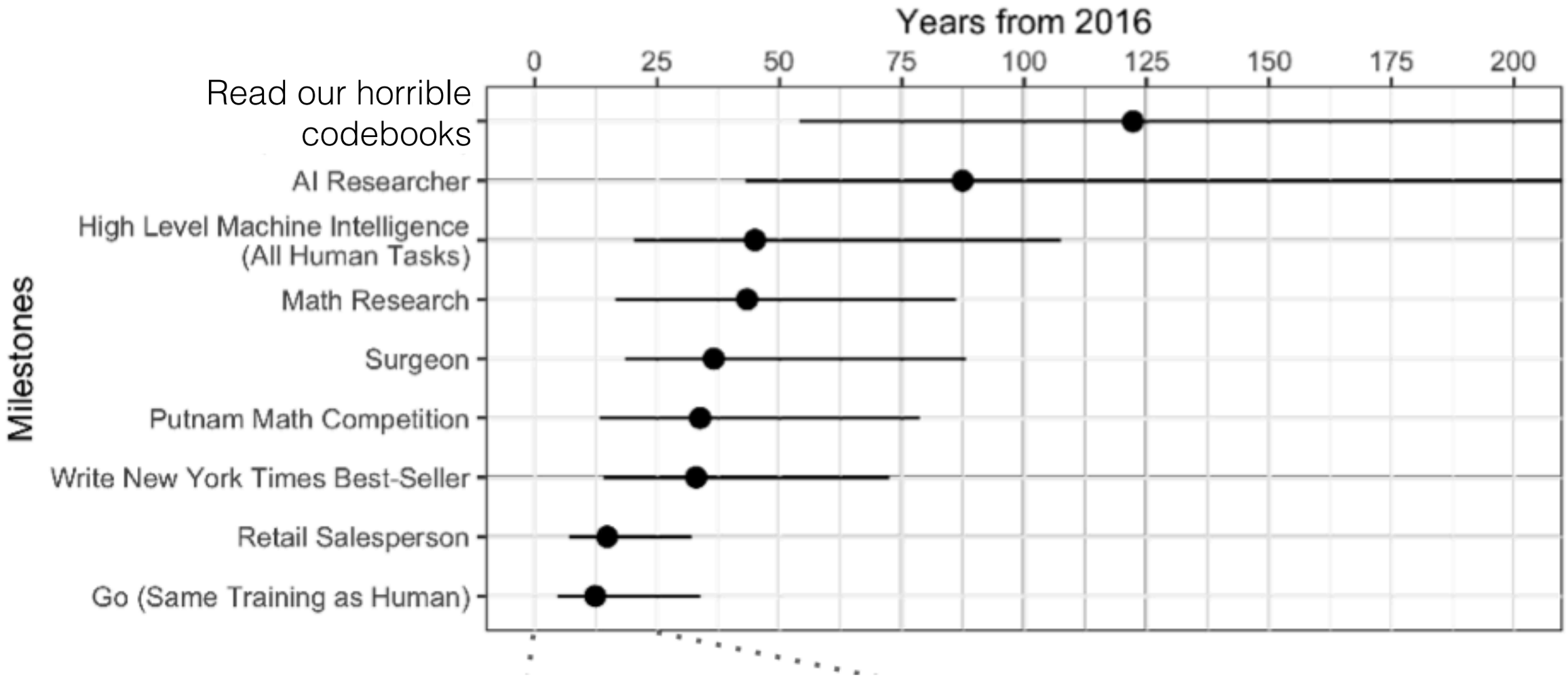
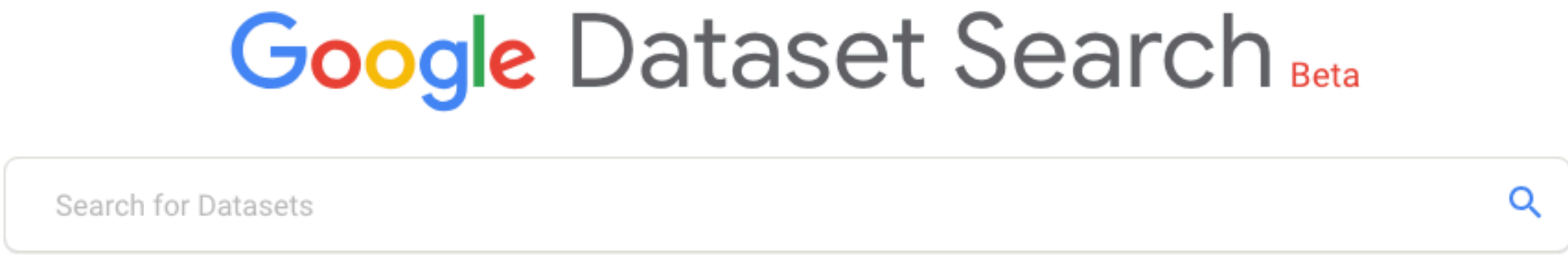
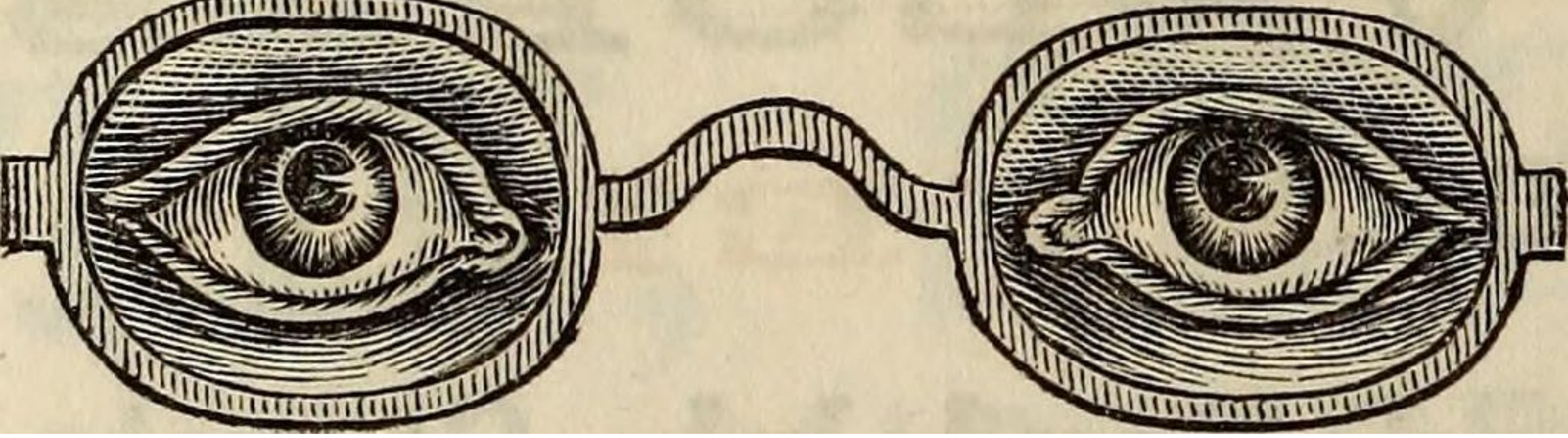
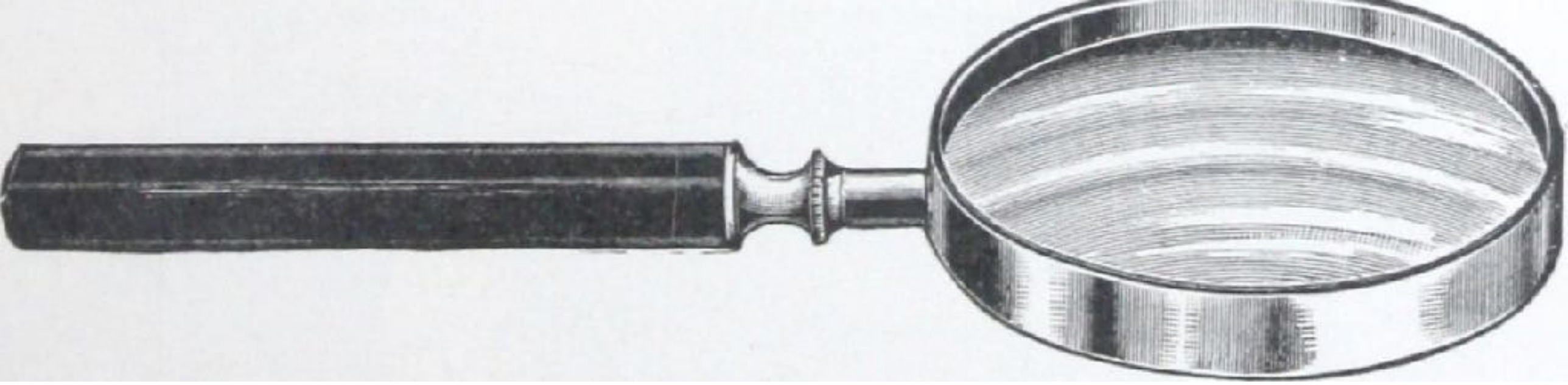
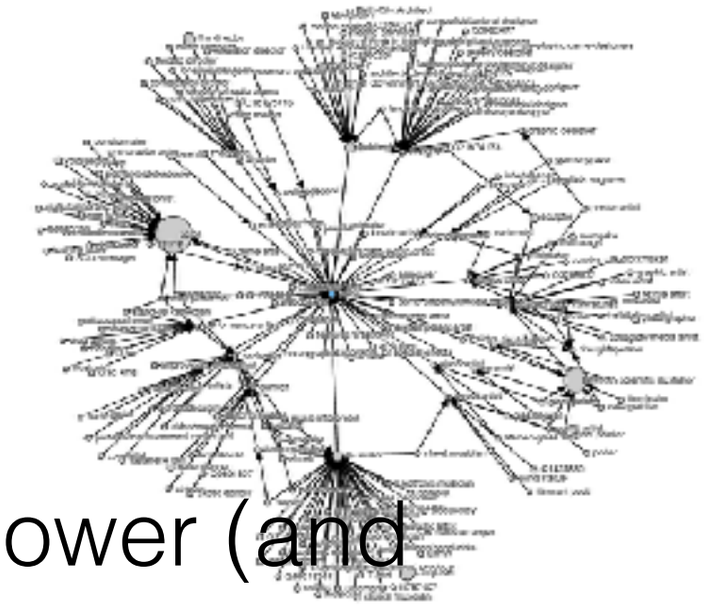
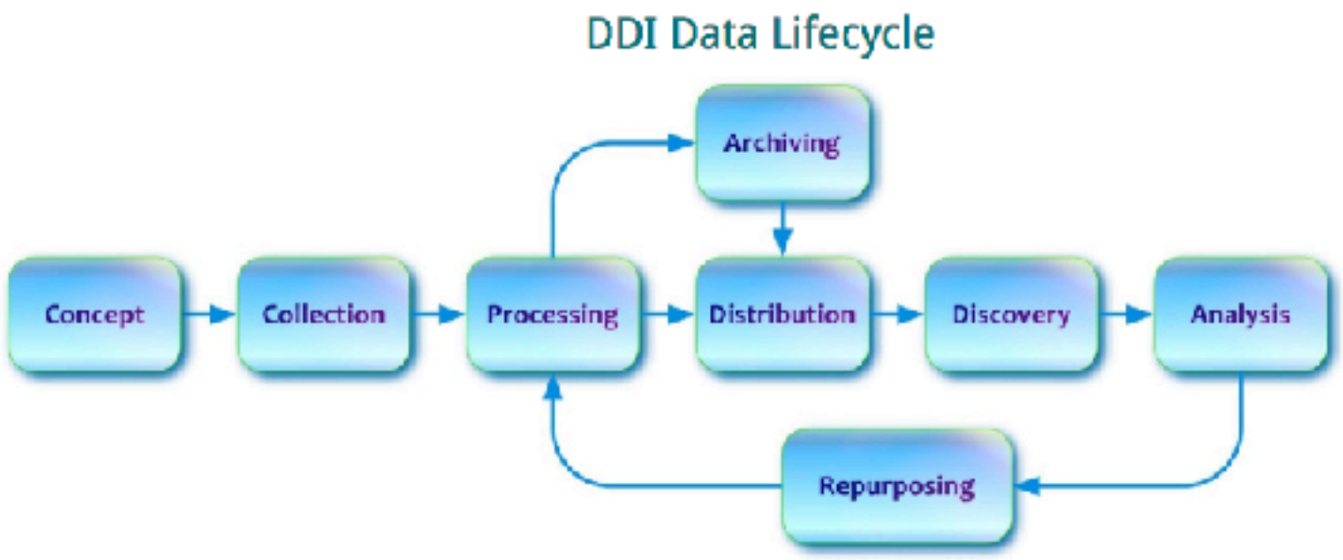
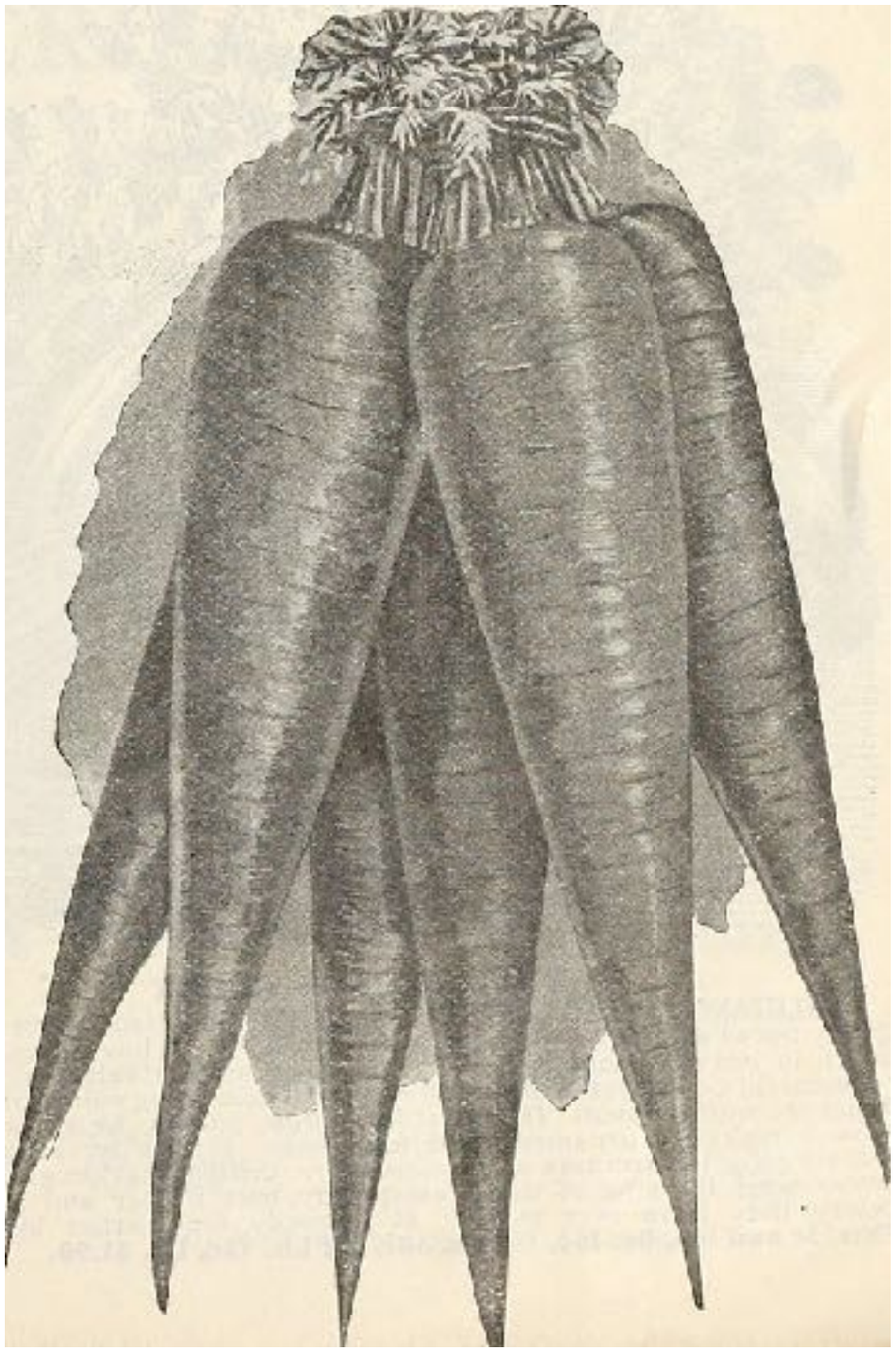
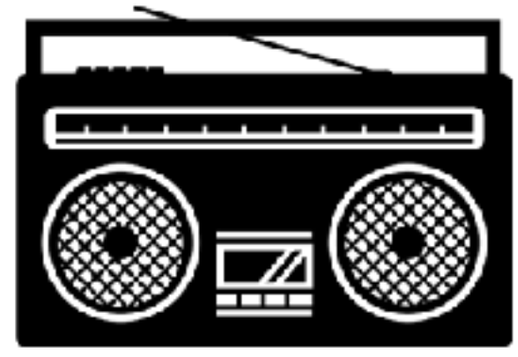
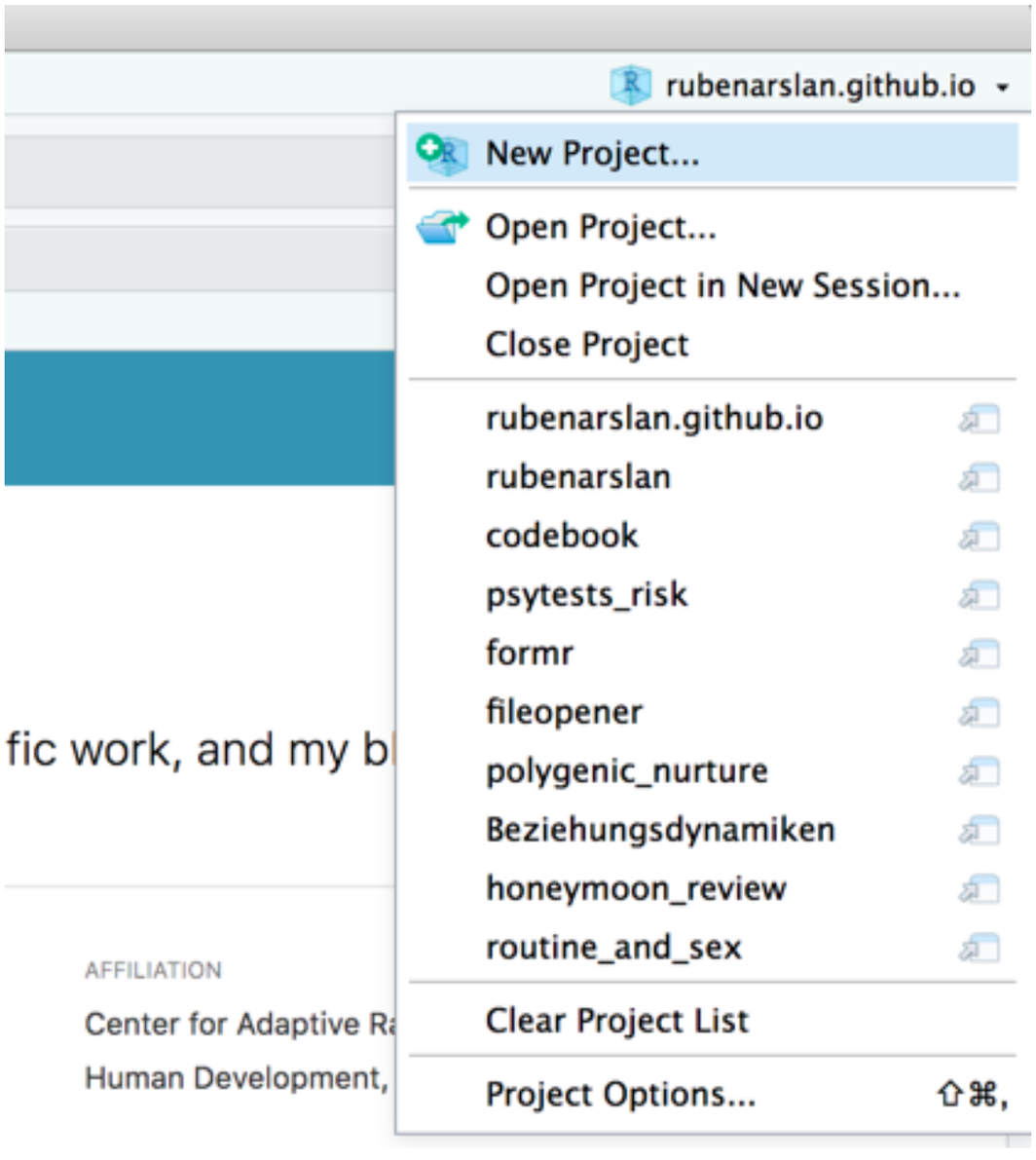
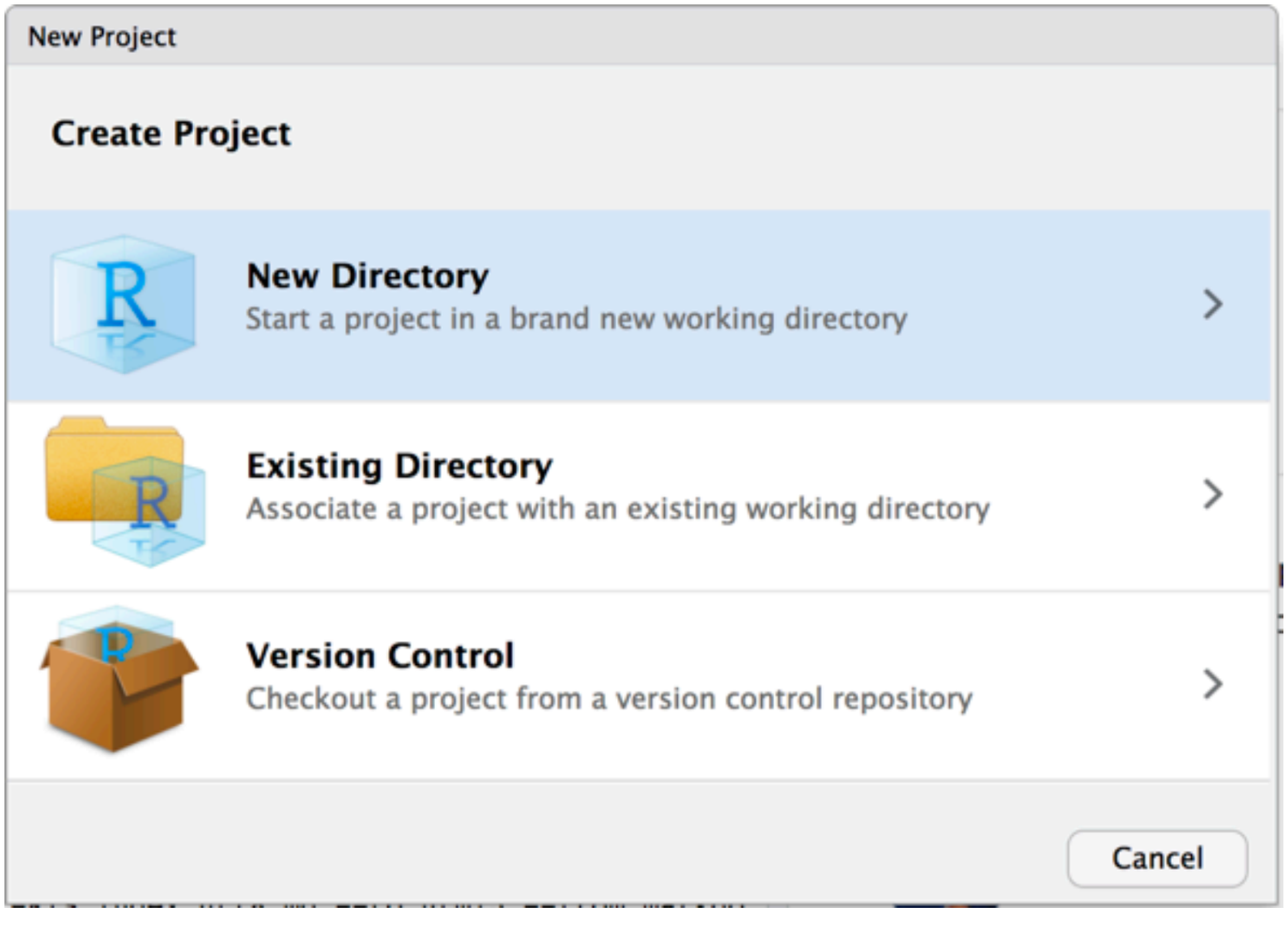
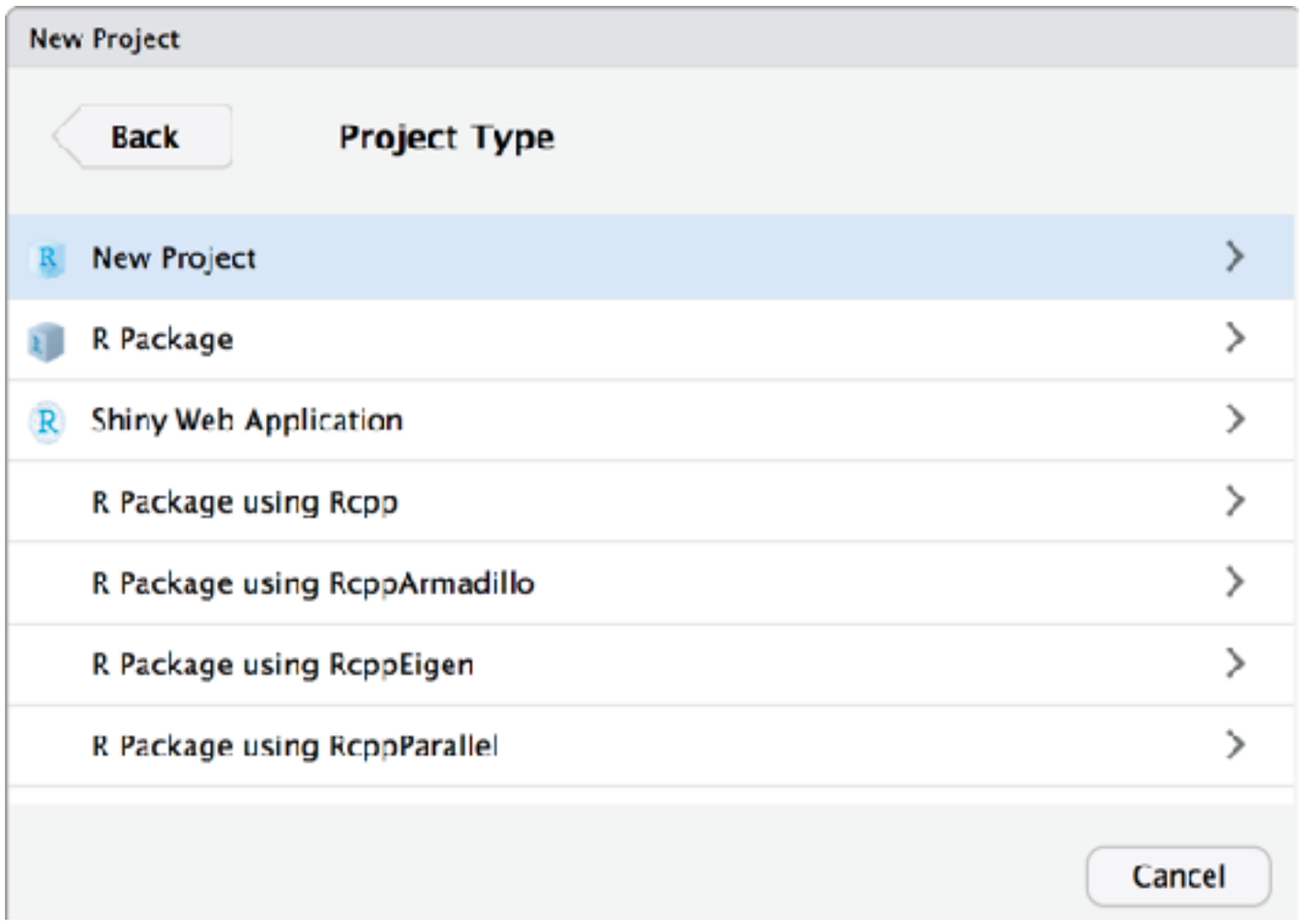
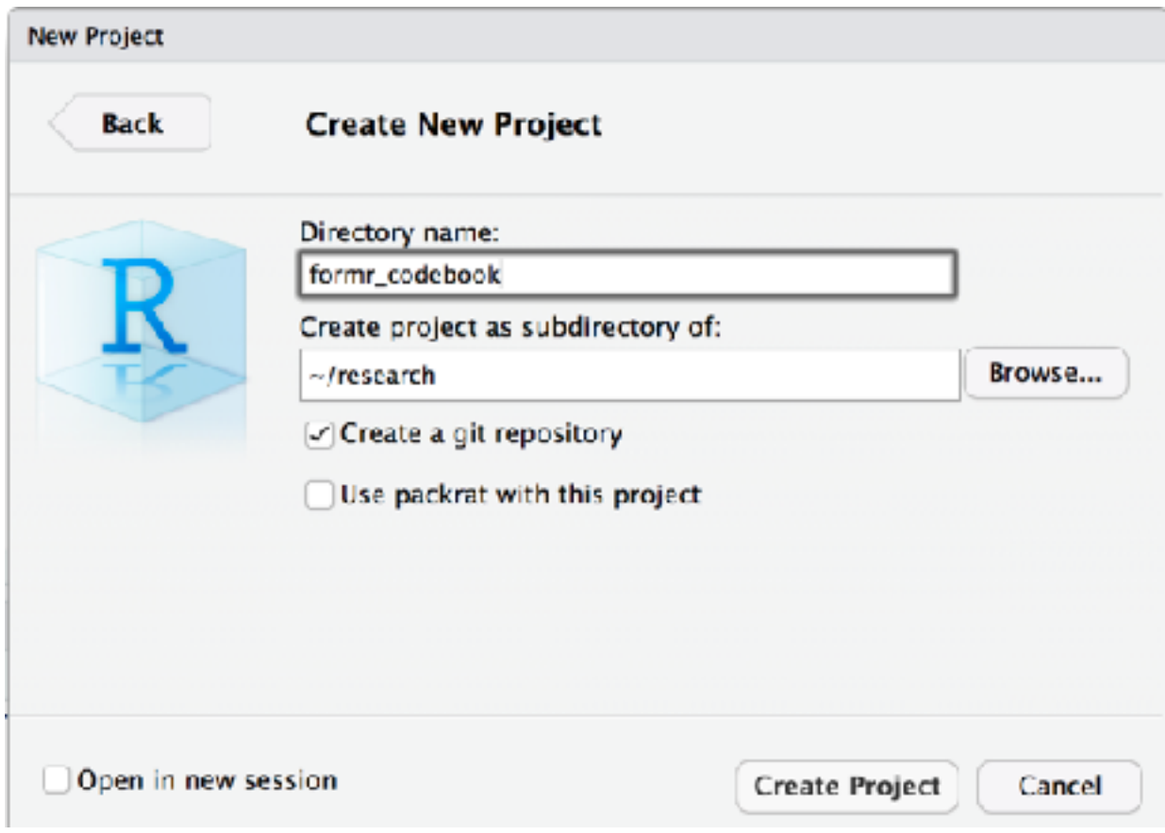
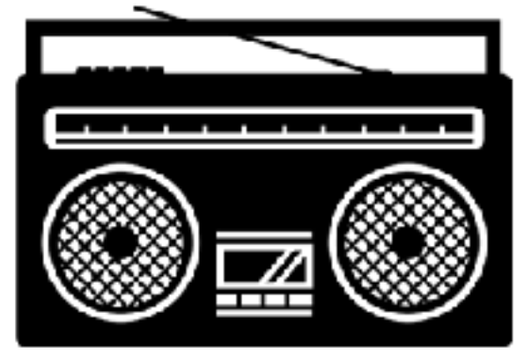
We can blend in R code: Yesterday
was `r yesterday`.
```

```
<h1>HTML</h1>

<p>We can make text <strong>fat</
strong> and <em>italic</em>.</p>

<p>We can <a href="https://
formr.org">link stuff</a>.</p>

<p>We can add images. <img
src="boombox.jpg" alt="boombox"></
p>

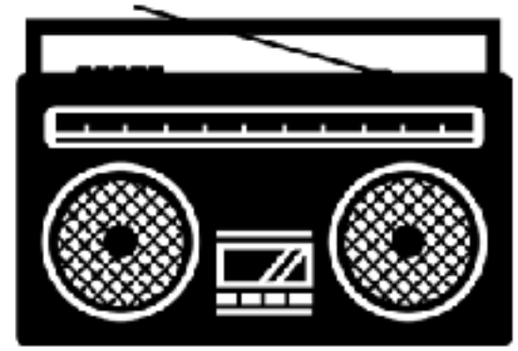
<p>We can blend in R code: Yesterday
was 2016-09-15.</p>
```

Markdown

We can make text **fat** and *italic*.

We can link stuff.

We can add images.

We can blend in R code: Yesterday
was 2016-09-15.

Item types



ℹ Plain display types

note

display text. Notes are only displayed once, you can think of them as being "answered" simple by submitting.

submit *timeout*

display a submit button. No items are displayed after the submit button, until all of the ones preceding it have been answered. This is useful for pagination and to ensure that answers required for `showif` or for dynamically generating item text have been given. If you specify the optional timeout (an integer, milliseconds), the submit button will automatically submit after that time has passed. However, if not all items are answered or optional, the user will end up on the same page. Together with optional items, this is a way to use timed submissions. The data in the item display table can be used to check how long an item was displayed and whether this matches with the server's time for when it sent the item and received the response.

Item types

⌨ Simple input family

text *max_length* allows you to enter a text in a single-line input field. Adding a number `text 100` defines the maximum number of characters that may be entered.

textarea *max_length* displays a multi-line input field

number *min, max, step* for numbers. `step` defaults to `1`, using `any` will allow any decimals.

letters *max_length* like text, allows only letters (`A-Za-züäöß.;,!:`), no numbers.

email for email addresses. They will be validated for syntax, but they won't be verified unless you say so in the run.

Item types

↔ Sliders

range *min,max,step* — these are sliders. The numeric value chosen is not displayed. Text to be shown to the left and right of the slider can be defined using the choice1 and choice2 fields. Defaults are `1,100,1`.

range_ticks *min,max,step* — like range but the individual steps are visually indicated using ticks and the chosen number is shown to the right.

Item types



Datetime family

date for dates

time for times

Item types



Fancy family

geopoint displays a button next to a text field. If you press the button (which has the location icon ➤ on it) and agree to share your location, the GPS coordinates will be saved. If you deny access or if GPS positioning fails, you can enter a location manually.

color allows you to pick a color, using the operating system color picker (or one polyfilled by Webshims)

random min,max generates a random number for later use (e.g. randomisation in experiments). Minimum and maximum default to 0 and 1 respectively. If you specify them, you have to specify both.

Item types



- [x] Multiple choice family

mc *choice_list*	multipe choice (radio buttons), you can choose only one.
mc_button *choices*	like `mc` but instead of the text appearing next to a small button, a big button contains each choice label
mc_multiple *choice_list*	multiple multiple choice (check boxes), you can choose several. Choices defined as above.
mc_multiple_button	like mc_multiple and mc_button
check	a single check box for confirmation of a statement.
check_button	a bigger button to check.
rating_button *min, max, step*	This shows the choice1 label to the left, the choice2 label to the right and a series of numbered buttons as defined by `min,max,step` in between. Defaults to 1,5,1.

Item types



- [x] Multiple choice family

select_one *choice_list*	a dropdown, you can choose only one
select_multiple *choice_list*	a list in which, you can choose several options
select_or_add *choice_list, maxType*	like select_one, but it allows users to choose an option not given. Uses Select2. *maxType* can be used to set an upper limit on the length of the user-added option. Defaults to 255.
select_or_add_multiple *choice_list, maxType, maxChoose*	like select_multiple and select_or_add_one, allows users to add options not given. *maxChoose* can be used to place an upper limit on the number of chooseable options.
mc_heading *choice_list*	This type permits you to show the labels for mc or mc_multiple choices only once. To get the necessary tabular look, assign a constant width to the choices (with classes), give the heading the same choices as the mcs, and give the following mcs (or mc_multiples) the same classes + hide_label.

Item types



Server family

ip	saves your IP address. This should probably not happen covertly but be explicitly announced or made optional.
referrer	saves the last outside referrer (if any), ie. from which website you came to formr
server var	saves the $_SERVER value with the index given by var. Can be used to store one of 'HTTP_USER_AGENT', 'HTTP_ACCEPT', 'HTTP_ACCEPT_CHARSET', 'HTTP_ACCEPT_ENCODING', 'HTTP_ACCEPT_LANGUAGE', 'HTTP_CONNECTION', 'HTTP_HOST', 'QUERY_STRING', 'REQUEST_TIME', 'REQUEST_TIME_FLOAT'. In English: the browser, some stuff about browser language information, some server stuff, and access time.

Particularities



- Item responses are directly checked in the browser
- By default all items are mandatory
- showifs are checked in JavaScript (immediately) or R (after submissions)
- Rmarkdown is usable in all labels
- Saved data only accessible server-side by default (i.e. we never output data without the study author, you, saying so)

Survey settings

Survey Settings

These are some settings for advanced users. You'll mostly need the "Import items" and the "Export results" options to the left.

Items Per Page

Do you want a certain number of items on each page? We prefer speciyfing pages manually (by adding submit buttons items when we want a pagebreaks) because this gives us greater manual control

0

- by default all items until the next submit button are shown

Enable Instant Validation

Instant validation means that users will be alerted if their survey input is invalid right after entering their information. Otherwise, validation messages will only be shown once the user tries to submit.

Enable

Percentage Display

Sometimes, in complex studies where several surveys are linked, you'll want to let the progress bar that the user sees only vary in a given range (e.g. first survey 0-40, second survey 40-100).

from	0	to	100	%

- if you want the progress bar to be consistent across multiple surveys

Survey Unlinking

Unlinking a survey means that the results will only be shown in random order, without session codes and dates and only after a minimum of 10 results are in. This is meant as a way to anonymise personally identifiable data and separate it from the survey data that you will analyze. **You can't change this settings once you select this option.**

Unlink Survey

- to separate research data from identifiable data

Disable Results Display

Selecting this option will disable displaying the data of this survey in forum. However the data will still be available for use. **You can't change this settings once you select this option.**

Disable

- hide results (responses)

Survey settings



Survey access window

Access window

How big should the access window be for your survey? Here, you define the time a user can start the survey (usually after receiving an email invitation). By setting the second value to a value other than zero, you are saying that the user has to finish with the survey x minutes after the access window closed.
The sum of these values is the maximum time someone can spend on this unit, giving you more predictability than the snooze button (see below). To allow a user to keep editing indefinitely, set the finishing time and inactivity expiration to 0. If inactivity expiration is also set, a survey can expire before the end of the finish time. More information.

Start editing within	0	minutes	finishing editing within	0	minutes after the access window closed

- for diaries (access window, after last edit)

Inactivity Expiration (snooze)

If a user is inactive in the survey for x minutes, should the survey expire? Specify 0 if not. If a user inactive for x minutes, the run will automatically move on. If the invitation is still valid (see above), this value doesn't count. Beware: much like with the snooze button on your alarm clock, a user can theoretically snooze indefinitely.

33	Minutes

Survey Paging

- Enable back and forth navigation within a study

Custom Paging

By enabling custom dynamic paging, your survey items will be "grouped" in pages depending on how your *Submit Items* are defined in the Items sheet. That is, each page ends at a defined submit button. Enabling this option nullifies the above "*Items Per Page*" setting, which means the number of items on a page will be determined by where *Submit Items* are placed in your items sheet. **You can't change this settings once you select this option.**

☐ **Enable Paging**

Updating surveys

Only for Google-Sheets & before real data:

Quick-upload items

Otherwise:

Upload an item table

Choose File No file chosen

Did you know, that on many computers you can also drag and drop a file on this box instead of navigating there through the file browser?

Or use this Googlesheet

Sheet link

https://docs.google.com/spreadsheets/d/1gwK03P7bX4-QjZyr0lSVluqx0OifNVDzNTBi3v8aD6I/edit

Make sure this sheet is accessible by anyone with the link

Upload new items, possibly partially delete 0 real results and 1 test sessions.

Updating surveys



If real data might be overwritten:

⚠ Delete Results Confirmation

Do you want to delete the results, if the item table changes were too major?
Enter the survey name below if you're okay with data being potentially deleted.
Leave this field empty if you're fixing typos in a live study.

survey name (see up left)

⚠ Upload new items, possibly partially delete 2 real results and 1 test sessions.

Questions?



Registration



- formr.org -> sign up with the token: **formr friends**
- sign up for Google **Spreadsheets** docs.google.com
- download & install **R** cran.r-project.org
- download & install **RStudio** rstudio.com
- in RStudio, execute:

```
install.packages("devtools")
devtools::install_github("rubenarslan/formr")
```

RStudio Settings



- Options
    - Code
        - Diagnostics
            - Check all boxes

Simple study



- *If you aren't working on something already:*
  - Duplicate tiny.cc/formr_blank into your own account

  - Add some items, that might be in a real study of yours. At least five different item types, max. 15 items. At least one showif

- Share via link (***anyone can view***)

- Upload it in formr via the link.

- Make a *public run* with the survey and one stop button. Name: goe-your-first-name.

Edit Run

I am panicking : (

Reorder | Lock | Export | Import

Publicness:

Survey Short Description

Start_Demo

0 complete results, 1246 begun (in ~-1.55m)

View items | Upload items

1246

Save changes | Test

10

Stop Page Description

Feedback text:

```
# Thank you for participating
```

20

Save changes | Test

click one of the symbols above to add a module

Run: **Edit**







- compose Run Units
- first add survey
  - select survey in dropdown, save
- then add Stop-Button

Run: **Test**





- Creates a new test user
- Participate in run as this new user
- Test users always begin with animal adjectives.
- E.g. **betterFlamingo**XXXUHBtvCNwNfqX09yFzs7db4kSssfTBHqZakQZowM_8nVfjcw

Run: **Testing**



- Testers and admins see this "monkey bar" while testing (bottom right)

- from left to right:



- fill out automatically and submit
- show hidden items (showif = false)
- show debugging info (item names, openCPU report)
- skip a step/end a pause prematurely
- push to position
- Delete this survey entry and start over
- Delete this user and start over

Run: **Participate**





- Name and public link to your run shows up at the top left

- This link can be shared to potential participants

Data entry



- Give your run a name à la *goe-ruben*
- Set it to *public*
- Test your own run with a test code.
- Go to formr.org/studies and fill out 5 surveys by the others (those starting with *goe-*).

Complexity



- Building complex studies with formr:
  - Chain different modules, to steer your participants through the study
  - Basic programming (Loops, if-conditions)

Runs



- Study structure. Allows you to chain
  - Simply surveys
  - Emails/SMS
  - external modules
  - Control modules (Loops, if-conditions, Stop, Pause, Shuffle)



Run: **Shuffle**



randomisation

10

Randomly assign to one of 2 groups counting from one.

You can later read the assigned group using `shuffle$group`.

You can then for example use a SkipForward to send one group to a different arm/path in the run or use a showif in a survey to show certain items/stimuli to one group only.

Saved | Test

Run: **Skip Forward**



control group has to wait first

20

if...

```
shuffle$group == 2
```

automatically skip forward to 50

else automatically **go on**

Saved | Test

Run: **Pause**



wait list control group

wait until time: 15:00 **and**

wait until date: dd/mm/yyyy **and**

43200 convert days

relative to

Text to show while waiting:

```
# Please wait
```

30

Run: **Email**



invitation

40

Account:

rubenarslan@gmail.com

Subject:

Now you may

Recipient-Field:

most recent reported address

Body:

```
{{login_link}}
```

`{{login_link}}` will be replaced by a personalised link to this run, `{{login_code}}` will be replaced with this user's session code.

Saved | Test

Run: **Survey**



survey

Saved

50

R: **Skip Backward**



loop survey twice

60

if...

```
nrow(survey) < 2
```

...skip backward to 50

Saved | Test

Run: **External**



use the info from survey to call magic API

External link:

end using API

Expire after 0 minutes

70

Enter a URL like `http://example.org?code={{login_code}}` and the user will be sent to that URL, replacing `{{login_code}}` with that user's code. Enter R-code to e.g. send more data along: `paste0('http:example.org?code={{login_link}}&age=', demographics$age)`.

Saved | Test

Run: **Stop**

End

Feedback text:

End.

80

Saved | Test

Feedback



- The data for a certain person are automatically available (formr reads your code to see if you mention the names of surveys and items)
- e.g. `demographics$name`
- Can be embedded via Rmarkdown
- bspw. Hej `r demographics$name`.

Feedback



- A few functions to generate feedback
  - qplot_on_normal – Show a value on a normal distribution
  - qplot_on_bar – Show values with CIs in a bar chart
  - feedback_chunk – Text feedback based on a z score

Run: **Settings**





- Set title, description, footnotes, change look-and-feel
- Add custom CSS, JS
- Set “Service Message”, which will be shown to people who don't have access to the study (e.g. because you're fixing problems or because the study ended)
- Pre-formulate Reminder-Emails
- Write an ***overview*** script
- Connect to OSF

Run: **Upload files**





- Upload images, videos, PDFs, other stimuli
- A little annoying: You have to copy the generated long link (works as access protection)
- Embed via Markdown/HTML
- You can also use other sources (YouTube, Imgur, whatever)



Run: **Danger zone**





- Delete: before you can do this, you have to delete all run units one-by-one. (a security measure)
- Empty run: deletes survey data
- Rename run: Change your run's address (previous address will break)
- Test the run using your user account identifier

Complex study



- If you don't have a concrete plan/project: Build a complex diary study
    - 10 questions in the beginning (incl. email address)
    - a daily email invitation for at least 30 days. The email invitation should say how many days they've done already.
    - 10 questions a day, different questions depending on answers in the first survey
    - feedback at the end (plot two diary variables in a scatterplot)
    - Template: Edit Run -> Import or http://tiny.cc/formr_wiki

Study monitoring



- How do you monitor an ongoing study?
    - Get an overview
    - Notice problems early
    - Contact users and fix their problems
    - *live edit* surveys/run structure

Run: **Users**





- Overview over participants
- Who has progressed how far?
- Push people to new position/delete (as in monkey bar)
- Send reminders
- *Use the "**New Named User**" button to create a participant you can later identify using the "name" you specify*

Run: **Overview**





- advanced: write R code to analyse/ monitor a study live

- https://formr.org/admin/run/AlltagUndSex/overview

Run: **Logs**





- Log books
- User detail: Who was where when?
  (you'll primary go here via user overview)
- Random groups: who was assigned to which group? (to be removed)
- Emails sent: which email was sent when (the user detail is the better overview if you're interested in specific users)
- Cron: errors/log messages that occurred when the run progressed automatically (this will change a lot soon, so don't get used to this)

Debugging



- Using **test users** is the royal road to debugging

- Quicker and easier, but won't always help:
  the **Test**-button in the run

  - accordion demo

live edits to surveys

- a certain risk to make whatever you're trying to fix worse

- some precautions are in place

- if there is real data (not by test users) in the survey already, you'll always have to enter the survey name to confirm your okay with potential data deletion

- formr always does a quick internal back-up before deleting anything (even if you confirmed), but you'll need Cyril or me to recover it.
  So, make your own back-ups.

live edits to runs

- Users will always slide to the next position, but there's no "lookahead"

- This means, it's no problem for you to add a second wave to a study later, just append it to the end of your current run (with a pause in front of it)

- If you change R code, feedbacks etc. will be regenerated

- If you had a bug and a user landed somewhere, he shouldn't be, fixing the bug won't move him back, you'll have to do that by hand.

Run: **Export data**





- You can export the whole dataset in your run in the long format
- reshape2/tidyr-fans rejoice?
- I think most use the data export in the survey or via the formr R package

Anonymity



- 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)

Connecting research and identifying data



- It's bad.
- It's hard to avoid, if e.g.
    - your diary study sends emails (ID) 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 1337 sexual relationships in the last 12 months

Let the machine do it



- formr can automate a lot of things that RAs would usually handle
  - re-inviting people
  - calculating how much someone should be paid
- formr won't look up its co-students in its database and form a nasty opinion of them based on their sexual history. RAs might.

dissociate data

- the *unlinked* option in surveys
    - affects only the visible *links* for you, the researcher, not for formr (formr can still use an email address)
    - whereas the RD is still *linked* by session/user codes, the ID will be shown
        - without the codes
        - without date times
        - in random order
        - only after at least 10 real entries exist
- the *hide_results* option does exactly what you would expect

dissociate data



- once enabled, the unlinked option can not be disabled anymore

- this makes it impossible to find people in your run using e.g. their email address

- they can e.g. forward you their diary invitation (it contains their user code), if they need and permit you to look at their RD

- be forward-thinking: e.g. calculate pay in a calculate field in the survey in which you collect payment data

Zukunftsmusik



- Make finer-grained timing and pauses possible while at the same time speeding up formr, making it more efficient

- App after all? Meeting planned with developers to discuss the possibilities of making formr work offline and send push messages

- Revamp the survey validation engine

- A library of formr studies, publishing designs?

- Finish the API

- Your vote?

Wrap up, feedback

- What should I cover tomorrow?
- What should I do differently?
- What did you miss?

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

Day 2: Getting your data out there

Ruben Arslan
Göttingen, September 25, 2018

ruben.arslan@gmail.com
@rubenarslan
blog: http://the100.ci

Plan for this part

Tomorrow

09:00	Intro to (meta)data sharing
10:00	Using metadata tools
11:00	Basics of rmarkdown, Rstudio
11:30	Getting your data out with *formr*
11:00	Making a codebook
12:00	*Lunch*
13:00	Making a codebook
14:00	*Coffee/Tea*
14:20	Publishing the codebook
14:50	Work on your study
17:00	End

Metadata plans

- A dataset you might want to describe in the future/ have described in the past

- max 1 minute

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

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. "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. "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).
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

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 & Speerschneider, 2015).

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 the data?





So I just share my data?

- Not so fast



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?

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 is not enough to make data truly 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

So I just share the data?

Still practicing EMA-compliant data retrieval, in case of #screenonly decision on thursday.
More Tweets: » https://twitter.com/hashtag/@iqwig?f=realtime

Magic: the only data retrieval method not prohibited by EMA's terms of use? #screenonly
More Tweets: https://twitter.com/hashtag/@iqwig?f=realtime

Blimey! Bad lighting conditions today. Can't check the endpoint definitions. #screenonly
More Tweets: » https://twitter.com/hashtag/@iqwig?f=realtime

EMA-compliant #screenonly working mode. The IT department is not amused.
More Tweets: » https://twitter.com/hashtag/@iqwig?f=realtime



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 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:

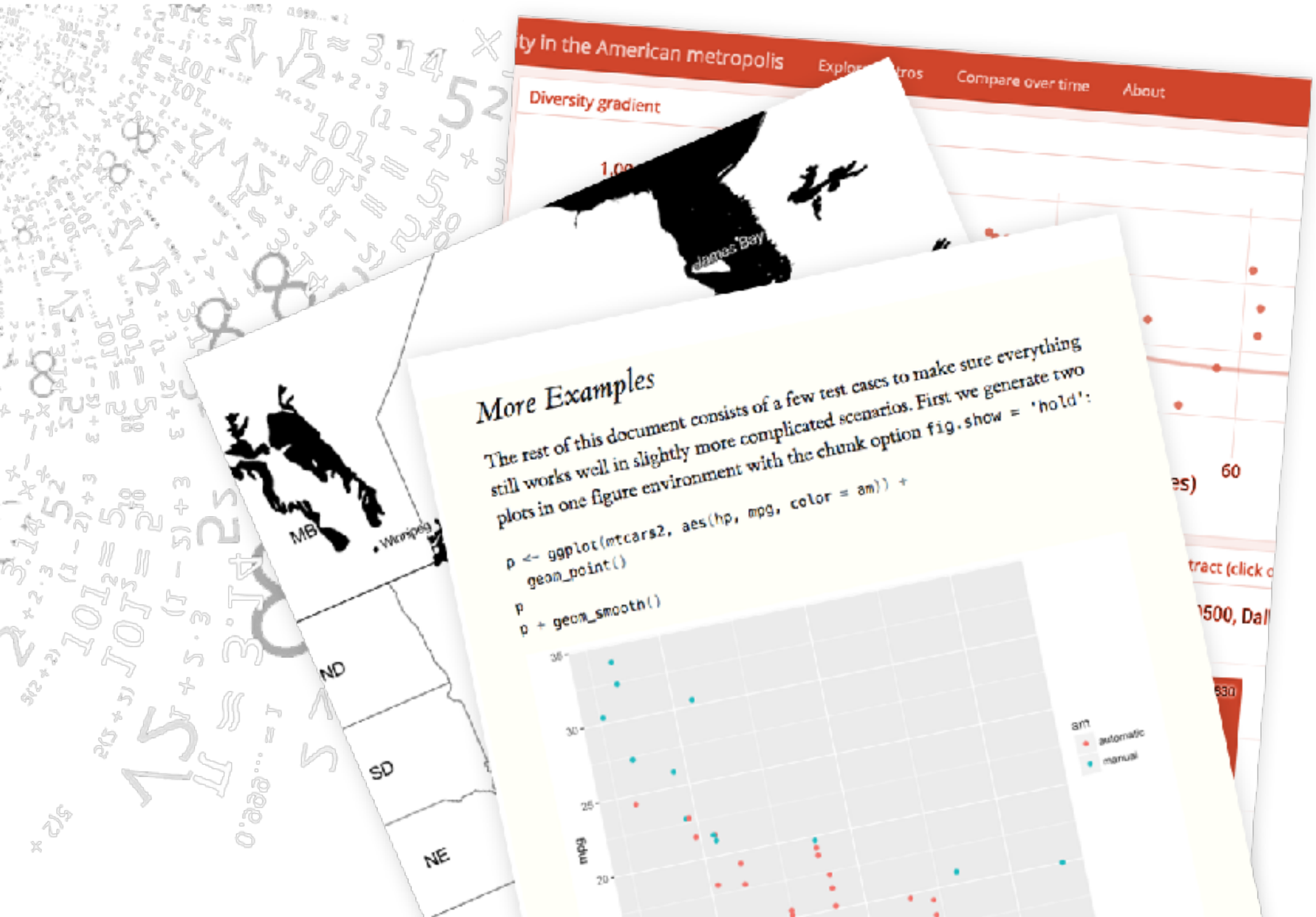
_acq = acquaintance (i.e., variables with this suffix are controlled for prior acquaintance)
_avg = average
_rat = rating variable
_z = z-standardized score
BC = booty call
DG = dating group (three groups in this study)
FIPI = five item personality inventory
FS = friendship
FWB = friends-with-benefits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



Definitions

- **Metadata**
  data about data

- **Linked data**
  published, structured data that can be queried semantically

- **Controlled vocabulary**
  getting our jargon straight

- **Ontology** (quoting Wikipedia)
  encompasses a representation, formal naming, and definition of the categories, properties, and relations between the concepts, data, and entities that substantiate one, many, or all domains.

- **Knowledge Graph**
  Google using ontologies to answer questions to Google Assistant, put infoboxes next to your search

Exercise 1



https://toolbox.google.com/datasetsearch



Exercise 1

- Go to https://toolbox.google.com/datasetsearch
- Two tasks:
  - Try to find a dataset that you know exists online (five minutes max)
  - Try to find a dataset that would be relevant and useful for your research questions & download it
    (10 minutes max)
- Note problems you have while doing so



Exercise 1

- Go to https://www.google.com/publicdata

- Choose some dataset that interests you

Exercise 1



- Try out one of these tools to see the power (and complicatedness) of ontologies
  - https://tools.wmflabs.org/reasonator
  - https://query.wikidata.org/
  - https://www.wikidata.org/wiki/Wikidata:Tools
- Try asking a question like “how many scientists remained childless”

Exercise 1

- Find standards for your fields
  - fairsharing.org
- Do any of them seem useful? Do you get the information you need?

Discussion

- Did you encounter problems?

- What applications for such and similar tools can you imagine in the scientific world you work in?

Perfect research metadata

- meaningful variable names, question labels, value labels –> linked to an 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(DUTF%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 Calloway 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

DataWiz

manage your data

- an assistant for managing psychological research data

- help you document complex data in standardised form

- extensive documentation

- it’s a lot of work, what do you get in return?

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

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



Document, Discover and Interoperate

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

ddialliance.org

Sounds like a ton of work!
As if metadata didn't have a bad enough reputation already!

https://www.nybooks.com/daily/2014/05/10/we-kill-people-based-metadata/

“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

- Add metadata locally, so *you* have it, but can also *share* it
- Automate a few annoying tasks with the help of metadata
    - 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
- Import metadata from places where we have it anyway (formr.org, Qualtrics, SPSS files)

https://github.com/rubenarslan/codebook

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

Exercise 2

- Create a codebook for datasets (ideally collected using formr)

- Publish it and share the link with us

Exercise 2

- Go to https://codebook.formr.org
- Do you have a dataset on your computer that fits the following criteria?
  - <10Mb
  - .sav/.dta format with value and variable labels set
  - not super-sensitive
- If not, use https://osf.io/j4fcb

Make a codebook



- A codebook is a good way to get an overview of your data.

- And it is also a good way to get started with rmarkdown, a powerful tool for making reproducible reports

New RStudio project





New RStudio project





RStudio Settings

- Options
  - Code
    - Diagnostics
      - **Check all boxes**

Exercise

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

- Enter:

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

Getting data out



.passwords.R

```
credentials = list(email =
"youremail@address.com", password =
"yourpassword")
```

codebook.Rmd

```
```{r}
source(".passwords.R")
library(formr)
library(codebook)
formr_connect(credentials$email,
credentials$password)
rm(credentials)
```

```{r}
formr_workshop =
formr_results("pre_formr_english")
```

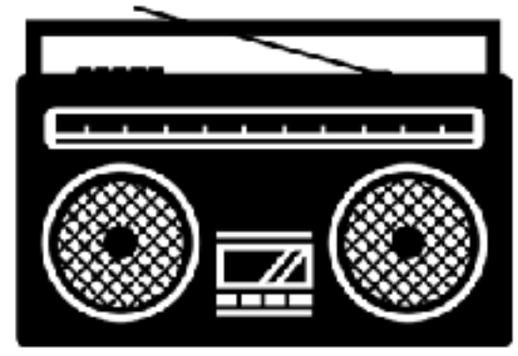
```{r}
codebook(formr_workshop)
```
```

Basics of rmarkdown





Links



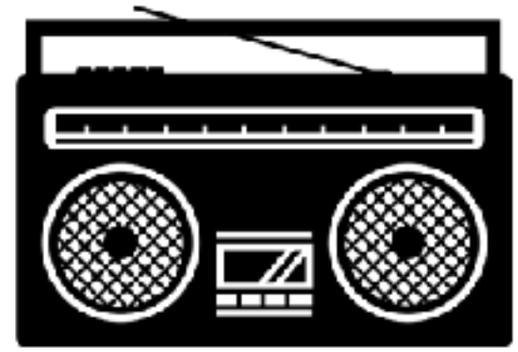
- http://rmarkdown.rstudio.com/lesson-1.html

- https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf

Security

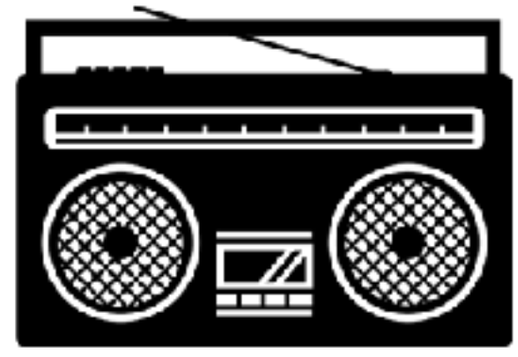
- Don't put your data and **`.passwords.R`** in Dropbox or other version control providers (e.g. Github)
- Use a strong password (e.g. four random words of mixed languages)
- Don't email/send your password and account info together.
- Prefer end-to-end encrypted communication (e.g. Whatsapp/Signal) for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

formr R package



- Most important functions
  - **formr_connect** – login
  - **formr_results** – get results, post-process automatically
  - **formr_raw_results** – for a quick fix/ when formr_results fails to do post-processing to your satisfaction
  - **items(results)$bfi_2R** – how was this item defined
  - **formr_simulate_from_items** – might be useful to test your code or feedback code before getting data

R package



- little helpers useful on formr.org itself
  - first/last/current – *last* takes last non-missing, *current* also takes missings
  - finished(survey) – how many finished?
  - expired(survey) – for who many did the clock run out?
  - survey$var %contains% “rstats”
  - survey$var %contains_word% “rstats”
  - survey$var %begins_with% “rstats”
  - survey$var %ends_with% “rstats”
  - next_day() – has at least one day passed?
  - time_passed(years, days, hours, minutes) – did x time pass?

Data merging

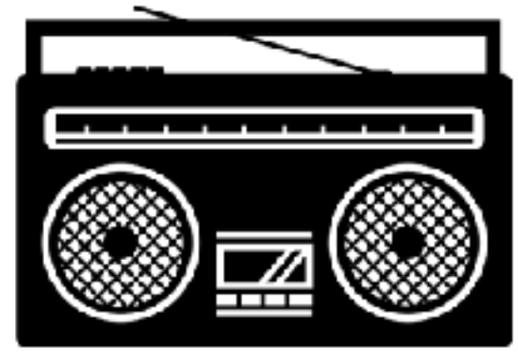
- Merge: from top to bottom in the run: dplyr::left_join
- Merge for complex studies (z.B. diary with repeated social network ratings): ideally prepare your data so there's no merge confusion (e.g. add a column for iterations)
  (if you didn't prepare, use the created date)

- www.the100.ci/2017/02/19/reproducible-websites-for-fun-and-profit/

Documenting studies

- documenting
  - the study structure
  - the survey items
  - the collected data

Documenting studies



- Edit run -> Export -> Include Survey Details

    - this JSON-file (text) contains the entire study structure (and settings) in human- and machine-readable form (no data, no uploads, only links)

- You can share this file (and/or the Excel sheets) with colleagues, upload in your supplement.

- Re-uploading the file in formr should allow others to reproduce your entire study perfectly.

Exercise

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

- Open RStudio and enter

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

Goal

- A codebook with a name and description, variable and value labels (for at least five variables)

- If there are items that form a scale, mark them up too.

- Guidance:
  https://rubenarslan.github.io/codebook/
  https://psyarxiv.com/5qc6h/
  http://tiny.cc/codebooktutorial

Codebook examples

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

- https://rubenarslan.github.io/dating_satellites/2_codebook.html

- https://rubenarslan.github.io/codebook/

Open Science Framework

- a one-stop shop for
  - preregistration
  - documentation of stimuli, materials, questionnaires
  - data archival – easy, but not very useful. doesn't currently do metadata/indexing, you could put a dataset here and link it from netlify. Only open or private, no allowance for gated access.
  - preprints

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/
  Very web 1.0, 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

- Possible way out: Use a user-friendly service and supplement it with a website generated by the codebook package.

Publish a codebook

- Sign up on netlify.com
- Rename your codebook.html file to index.html
- Drag and drop the folder to your first netlify page
- Give the page a meaningful name (e.g., the name of the dataset)
- Send the page name to me via email, so we can see it ruben.arslan@gmail.com

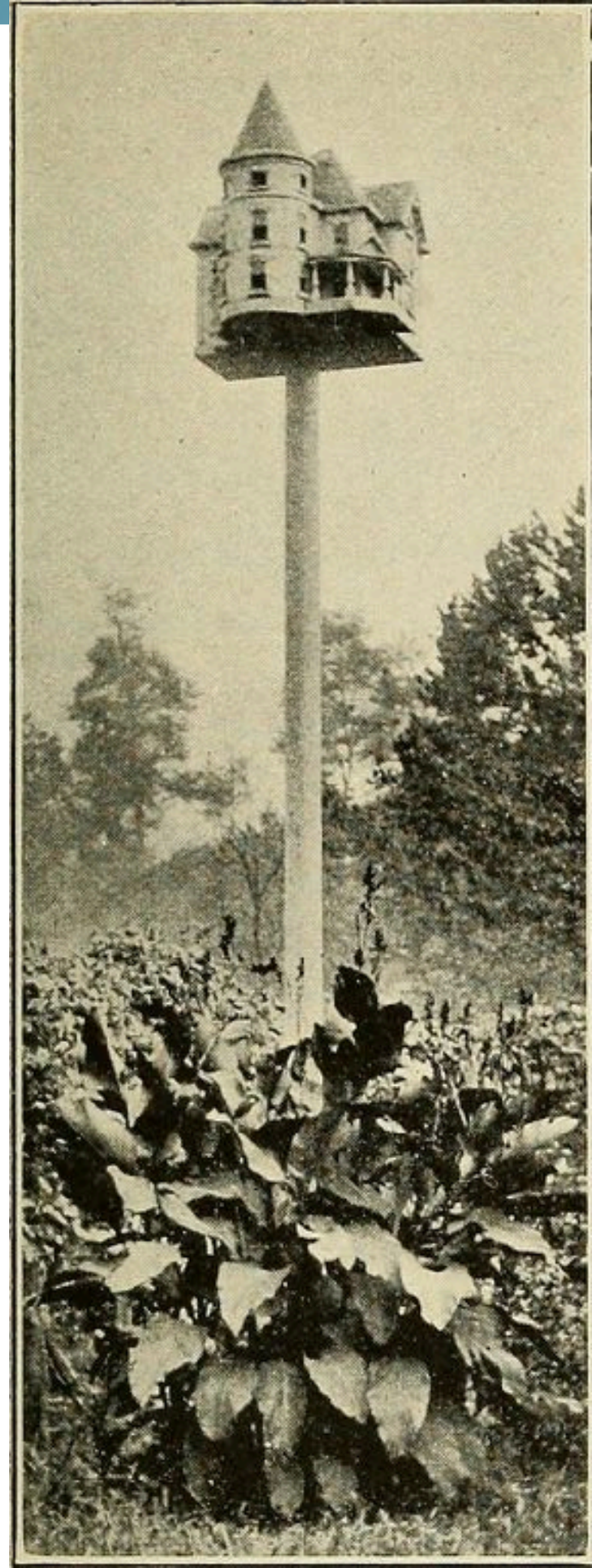
Allowed metadata fields

- For the dataset:
  https://schema.org/Dataset

- For variables in the dataset:
  https://pending.schema.org/PropertyValue

Forum

- Will the metadata you just generated be useful?
- What would you want from other researchers' metadata?
- Can you imagine doing this for future projects?
- What are your concerns? What are special problems with your study?



Weigh in

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

Your own study

- Work on or document your own study and analyses!

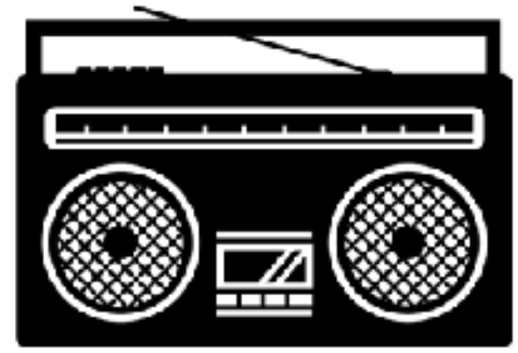
- Discuss with me

- Using formr vs. another software (although I'm not perfectly unbiased, I probably know more options than you)

- Potential hurdles

- Whatever else

Get help



- https://groups.google.com/forum/#!forum/formr - Support queries about formr to mailing list
- https://github.com/rubenarslan/formr.org/issues
  - bugs/feature-wishes, get notifien when/if we get to them (pull requests welcome too)
- https://formr.org/documentation
  - Our documentation
- https://github.com/rubenarslan/formr.org/wiki
  - Your documentation (Wiki), please contribute!
- stackoverflow.com
  - Good resource to ask R questions or find answer to previous R questions (much friendlier than R-Help)

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

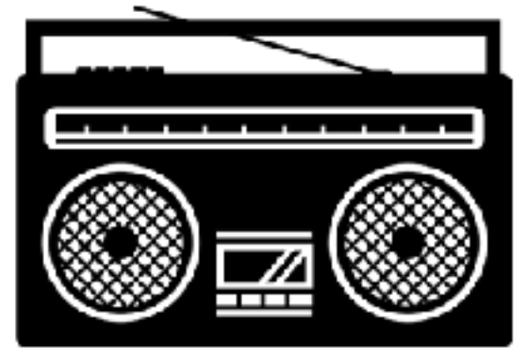
Ruben Arslan
Göttingen, September 25, 2018

ruben.arslan@gmail.com
@rubenarslan
https://tellmeimwrong.formr.org/

pictures: flickr internetarchivebookimages

blog: http://the100.ci

Credits



- Radio by Jon Testa from the Noun Project

Where else could we benefit from structured data?

- Publication-level
  some work on this, mainly focused on citation information afaict, little about content

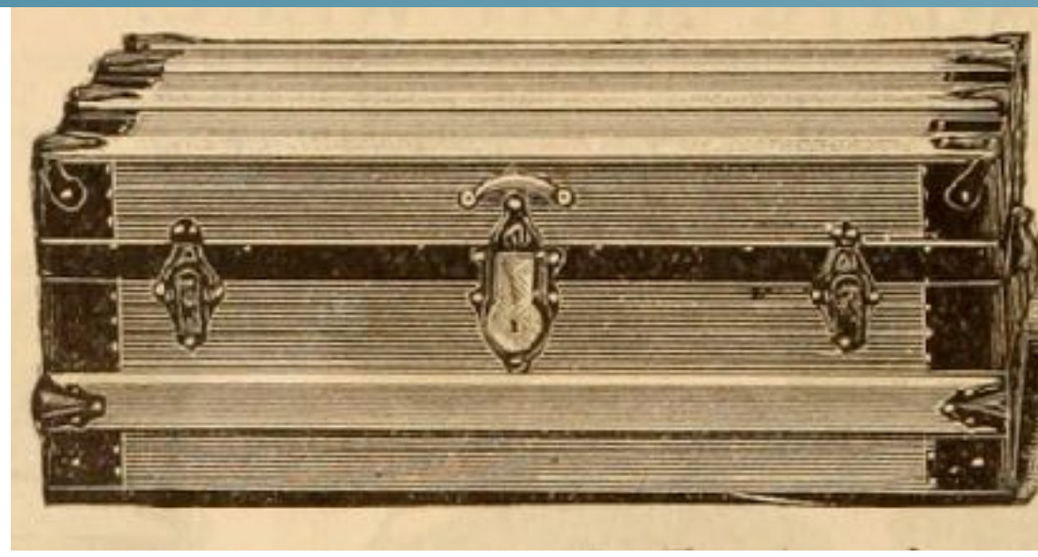
- Study/Investigation-level
  know only about this for funder information, little structured data

- Plots
  seems like a no-brainer: R-generated plots in svg already contain the raw data, axis labels should be meaningful -> don't know if a standard exists, haven't seen one around

- Statistical models and tests
  Some work on boiling this down to a common terminology, but usually just presented as tables, not structured data.

http://www.the100.ci/2017/02/19/reproducible-websites-for-fun-and-profit/

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

Open Data Resources

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

My workshop on maintaining privacy with open data: https://osf.io/n2dsq/