Responsible science at scale: Lessons for everyone from baby research and large collaborations

Melissa Kline Struhl

Research Scientist, MIT

CHS Executive Director

05/11/2026

(1) Sensitive data
(2) Not enough of it

Epic amounts of time (and money)

$20

Baby Elena arrives in the parking lot behind William James Hall with her father. Researcher Lauren greets the family with a parking permit.

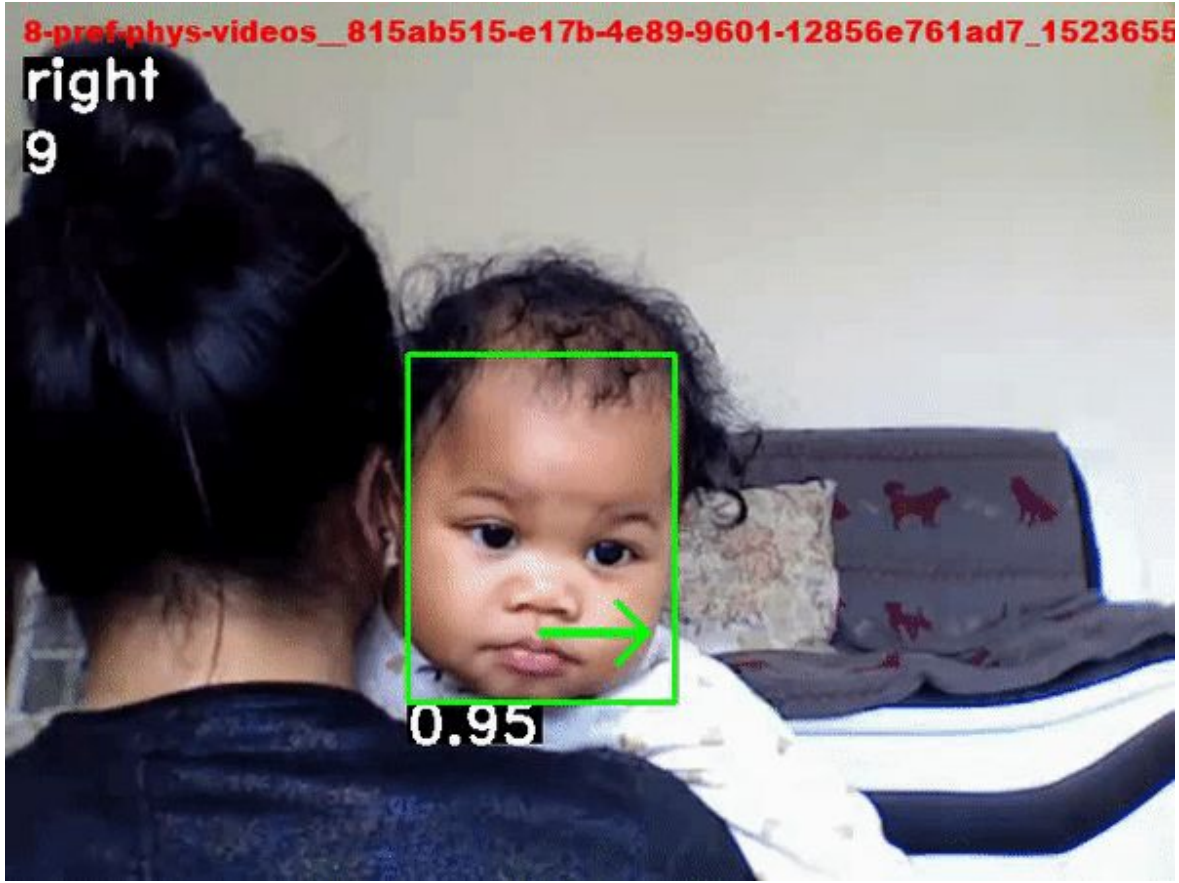
Before she leaves, Elena gets to pick out a prize to take home!

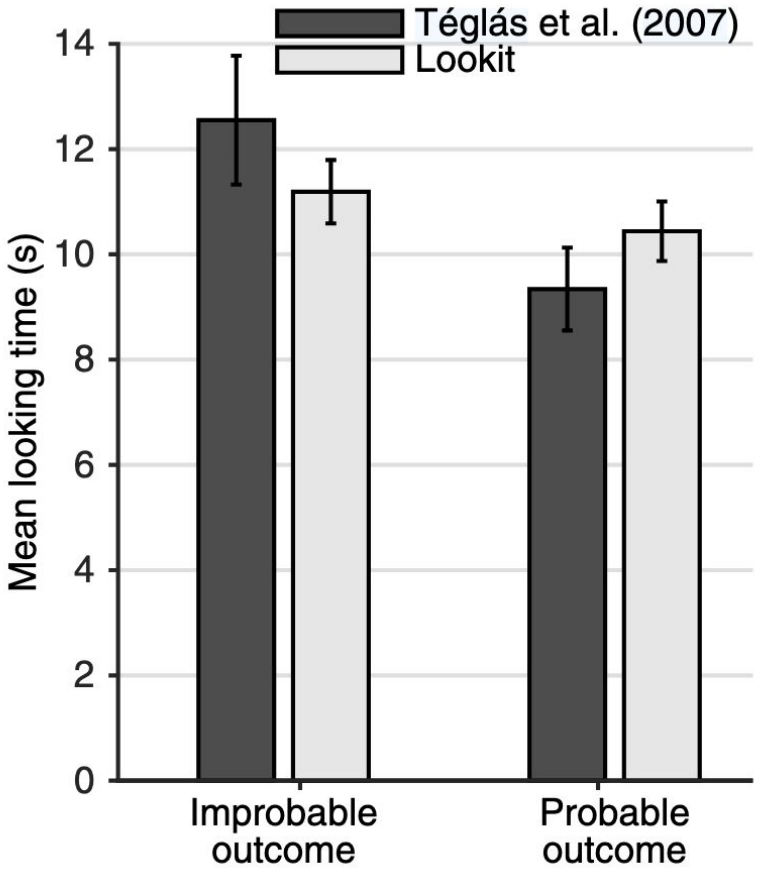
After the study is completed, Elena and Lauren play for a little longer.

Many studies with children involve a toy choice paradigm in which children decide between two objects. Here, Lauren demonstrates a preference for the bunny toy.

Lauren walks Elena and her dad back out to the parking lot. See you next time!

Image sequence: lab tour, Harvard LDS
https://software.rc.fas.harvard.edu/lds/parents/visit-the-lab/


8-pref-phys-videos__815ab515-e17b-4e89-9601-12856e761ad7_1523655
right
9
0.95


Téglás et al. (2007)
Lookit
Mean looking time (s)
0
2
4
6
8
10
12
14
Improbable outcome
Probable outcome

Consequences of the data bottleneck

One study takes **months to years** to collect data

- Large effects only, as few conditions as possible
- Limitations on longitudinal work
- Few replications

Tough to recruiting special populations, kids who come in aren't representative

Behavior changes in the lab (little kids talk less)

“Am I just really bad at experiments?”

Rest of this talk

- ManyBabies
- Psych-DS
- Children Helping Science
- Thinking about data sharing


ManyBabies
A global consortium of developmental researchers
MB

Brianna McMillan [2022 -] [email]

Smith College (Northampton, United States)

Melanie Soderstrom [2016 -] [email]

University of Manitoba (Winnipeg, Canada)

Christina Bergmann [2016 -] [email]

University of Applied Sciences (Osnabrück, Germany) & Max Planck Institute for Psycholinguistics (Nijmegen, Netherlands)

Krista Byers-Heinlein [2016 -] [email]

Concordia University (Montréal, Canada)

Michael Frank [2016 -] [email]

Stanford University (Stanford, United States)

Kiley Hamlin [2016 -] [email]

University of British Columbia (Vancouver, Canada)

Melissa Kline Struhl [2016 -] [email]

Lookit & MIT (Cambridge, United States)

Eon-Suk Ko [2022 -] [email]

Chosun University (Gwangju, South Korea)

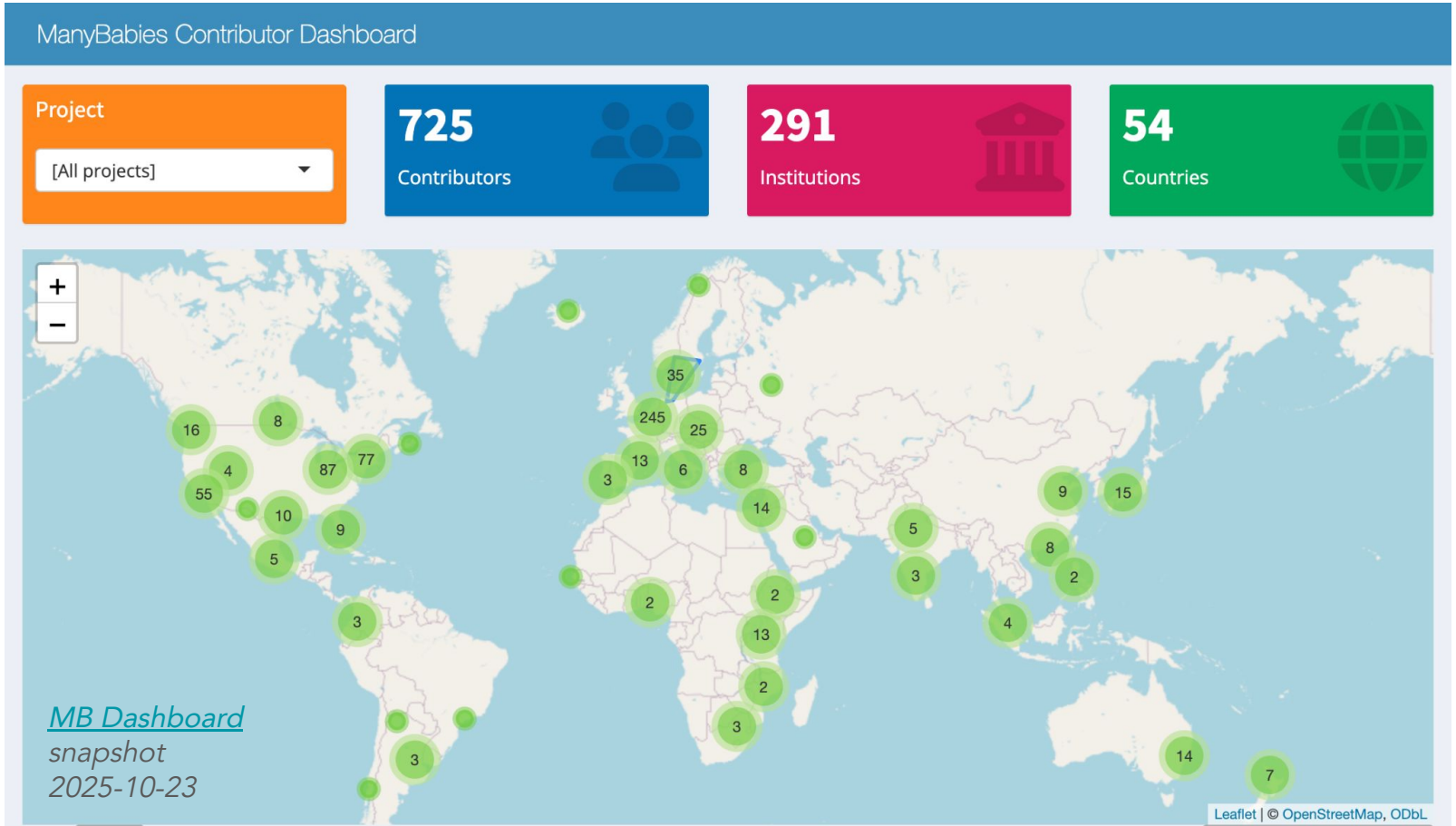
Casey Lew-Williams [2016 -] [email]

Princeton University (Princeton, United States)

Heidi Baumgartner [2021 -] [email]

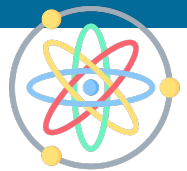
Stanford University (Stanford, United States)

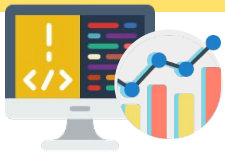
ManyBabies – network



ManyBabies – how it works

identify theoretical claim >> collaboratively design "best test" >> public invitation to collect data >> provide detailed protocols and support >> pool data & analyze via reproducible pipeline >> share FAIR data and encourage re-use





ManyBabies – main projects

manybabies.org/projects

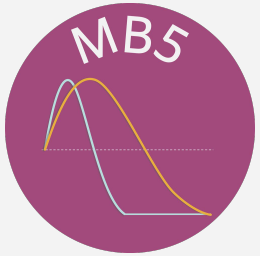
empirical

Preference for Infant-directed speech

Theory of mind

Rule learning

Social evaluation



Hunter & Ames model of infant attention

Neonatal & early imitation

methodological

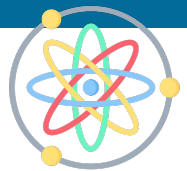
Unified demographic data collection

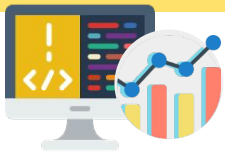
Online infant data collection

Web-based eye tracking with infants

ManyBabies – how it works

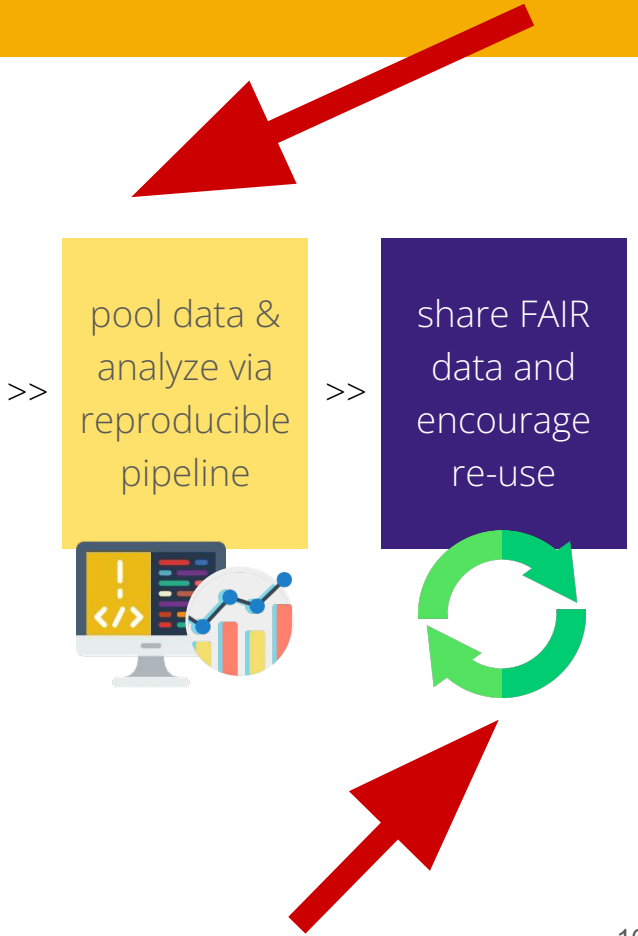
identify theoretical claim >> collaboratively design "best test" >> public invitation to collect data >> provide detailed protocols and support >> pool data & analyze via reproducible pipeline >> share FAIR data and encourage re-use





ManyBabies – how it works

- Two template CSVs shared
- 69 spreadsheets returned!
- Formatting errors, accidental PII, Excel date crimes, missing data, units issues (ms/s), corrupted files…
- Many, many dozens of hours to get to FAIR sharing




SIPS
SOCIETY FOR THE
IMPROVEMENT OF
PSYCHOLOGICAL SCIENCE

We have to start from where we are

- No matter what, any researcher deciding to share data has to start ***from their current data management practices*.**
- Lists of rules and best practices can be daunting, and hard to understand if not in your domain (is your data like my data?)

- Question: What are the **minimal behavior changes** we can ask researchers (= ME) to make to achieve the **maximum payoff** in data quality

- Challenge: **How to get researchers (ME) to actually make these changes?**

Broman & Woo (2018): Data Organization in Spreadsheets

- Be consistent.
- Write dates as YYYY-MM-DD.
- Fill in all of the cells.
- Put just one thing in a cell.
- Make it a rectangle.
- Create a data dictionary.
- No calculations in raw data files.
- Don’t use font color or highlighting as data.
- Choose good names for things.
- Make backups.
- Use data validation to avoid data entry mistakes.
- Save the data in plain text files.

…Excellent rules! (Twelve of them.)

…How do you know if you have you succeeded?

The Psych-DS data specification

- Psychological, behavioral, cognitive science data, including e.g. surveys, self report measures, reaction times, eye tracking, sensor data, ***hand-annotated video, transcribed speech,...***
- …Hand-constructed (or hand-modified) data of many types!
- “Small” (= Fits on your local machine)
- Privacy-sensitive (human data), but not all equally sensitive
  - Proposed 13th rule: Don’t mix sensitive and non-sensitive variables in the same data files
- Method:
  - (1) Copy BIDS where applicable.
  - (2) Use a validator.

The validator is both the final word on specification rules, and a success metric for individual researchers

Psych-DS in One Slide

- example_project/
  - analysis/
  - materials/
  - **data/**
    - **study-1_data.csv**
    - **study-1_sub-w2a1_data.csv**
    - **study-1_sub-dgesi_data.csv**
    - **study-1_sub-aa65_data.csv**
  - **dataset_description.json**
  - README.txt

study-1_data.csv

sub_id	age_years	responded
w2a1	22	True
fds11	34	False
dgesi	21	True
aa65	60	True

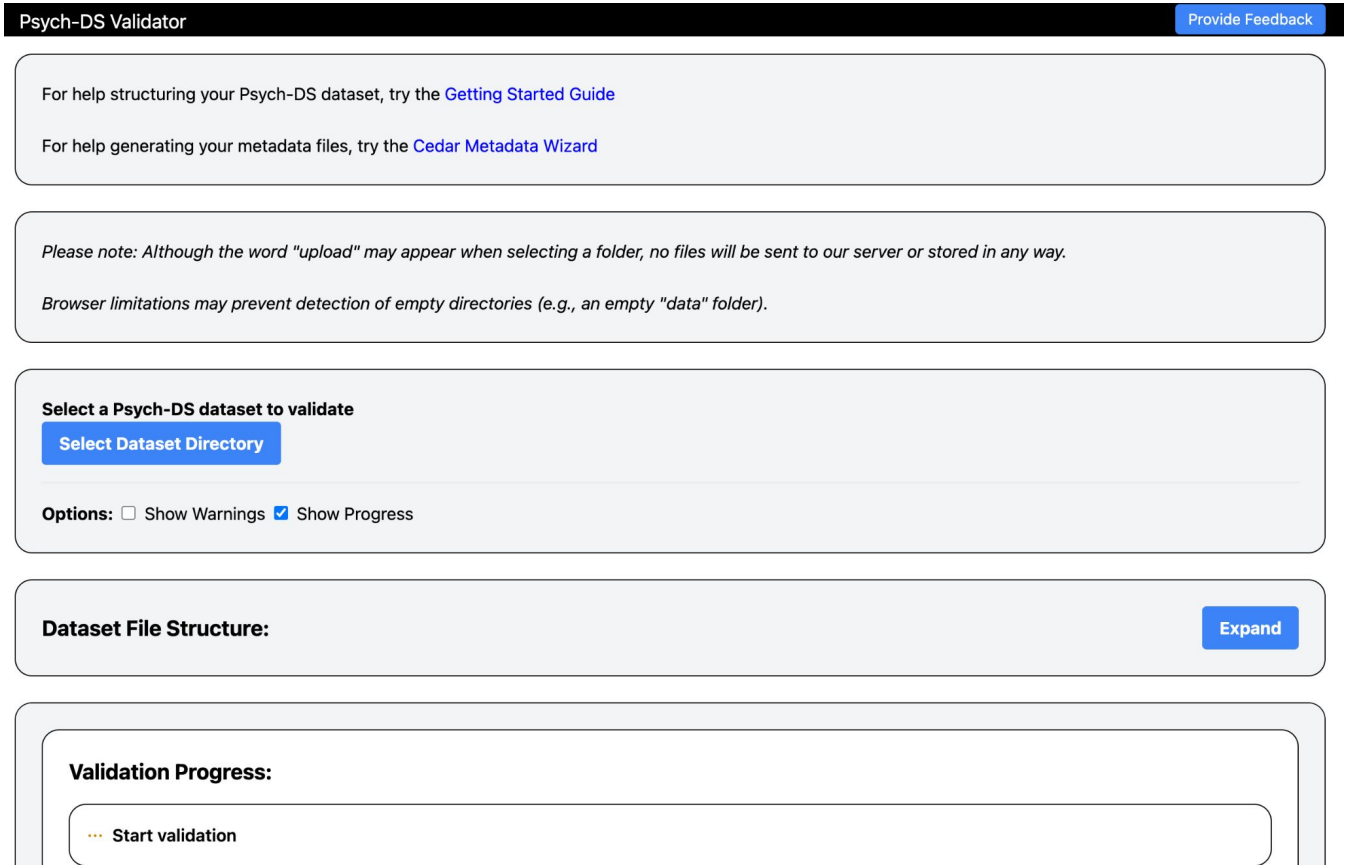
study-1_sub-w2a1_data.csv

sub_id	trial_id	response
w2a1	1	0.12
w2a1	2	0.31
w2a1	3	1.21
w2a1	4	6.41

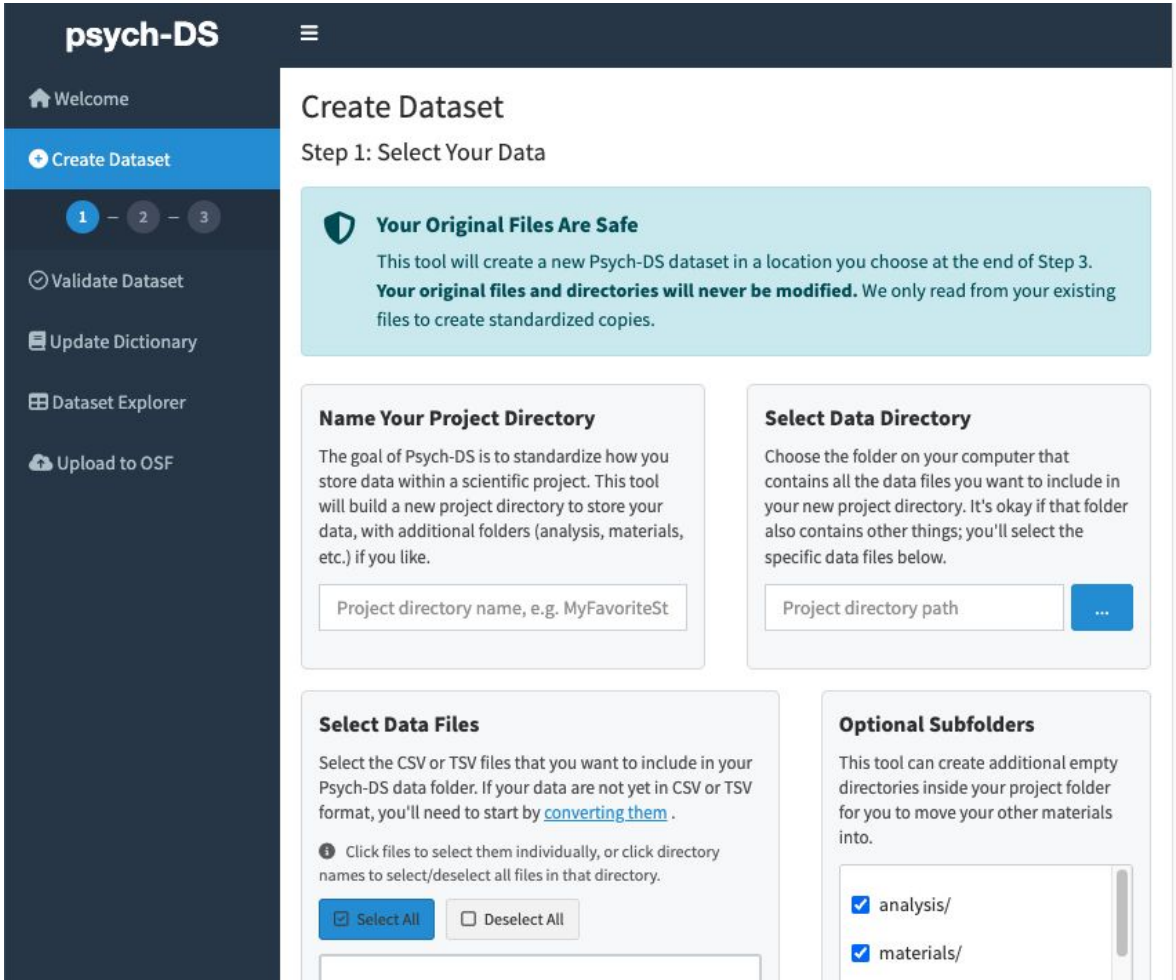
dataset_description.json

```
{ "@context" : "http://schema.org/",
 "@type" : "Dataset",
 "name" : "Psych-DS Example Dataset",
 "description" : "This is a minimal example of a dataset for
Psych-DS",
 "variableMeasured" : [“study_id”,"sub_id", "age_years",
“responded”, “trial_id”, "response"] }
```

https://psych-ds.github.io/validator/



https://github.com/psych-ds/psychds-r



Types of Data

- Terms you may see
  - “Human data” (not always the same as data that is about humans)
  - “Personally identifiable information/PII”
  - “Identifiable data”
  - “Sensitive data”
  - “Shareable data”

Types of Data

- All of these depends on:
  - Your country's laws
  - What your consent form said
  - Your lab or institution's policies
  - The specific information you collect about participants
  - **Your moral intuitions!**

~~"Personally Identifiable Information"~~
"Re-Identification Risk"

How **likely** is this data to re-identify someone?

X

What kind of **harm** could result from being re-identified?

How **likely** is this data to cause unintended exposure

X

What kind of **harm** could result from unintended exposure

ChildrenHelpingScience.com

LOOKIT

An open source online platform for researchers to implement developmental experiments…

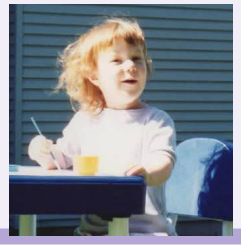
…and for families to participate in them from home

Built for infant looking times, works for **much more…**

…for the whole field (200 institutions and counting)

MelissaAge2.mp4

Name	Birthday	DOT	Age	Total MAB	M1	M2...
Melissa	8/[redacted]	9/18/1988	774	56	0	1...
Fakelina	7/4/1988	12/25/1989	539	89	1	1...



Lesson 1: All data files contain birthdays and faces until proven otherwise

Types of Data Data on Lookit Structuring Data

Everything is “Radioactive”!

My laptop
- My desktop
- My documents
- Downloads
- Random files

Email account
- Random attachments
- Your account & others’ accounts

Google Drive

Lab server

This is a problem for everyone!
Even if you don’t plan to share data...

Please be careful

Don’t email me a combination of some women’s street addresses and also their sexual histories.

Lesson 1: All data files contain birthdays and faces until proven otherwise

Lesson 2: The best way to protect sensitive data is to not store it

Response overview

The response overview file gives high-level information about each response - the account and child IDs, consent approval information, condition assignment, and information about the child such as gender and languages spoken. There is one row per response. This can be used in conjunction with frame data (below) to avoid having to parse JSON in your analysis.

Data (CSV)

Data dictionary (CSV)

Child data

The child data files contain one row per unique child. The data available about each child is the same as is available in the response summary CSV (with the exception of age at time of participation, which depends on the response time). All child data will be included in this file, regardless of selections above; you can use it to store this identifiable data separately from the response data.

Data (CSV)

Data dictionary (CSV)

ID	AGE_ROUND	COND	T1
HS31S	850	Trans	R
KW87	730	Intrans	L

ID	NAME	BIRTHDAY	AGE_DAYS
HS31S	Jane	9/2/2018	847
KW87	Anh	12/3/2017	731

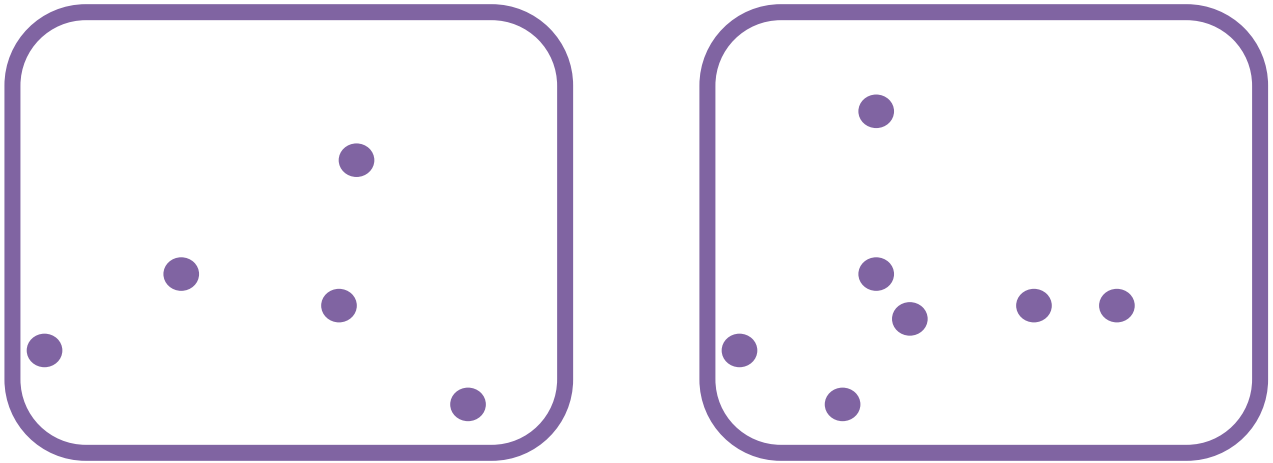
Lesson 3: Isolate your sensitive data, then share it like you mean it

MY SENSITIVE DATA

MY SHAREABLE DATA

A radioactive room, or three

- One folder, in a known location
- A lab video server
- The important thing is that this list is **discrete** and **short**

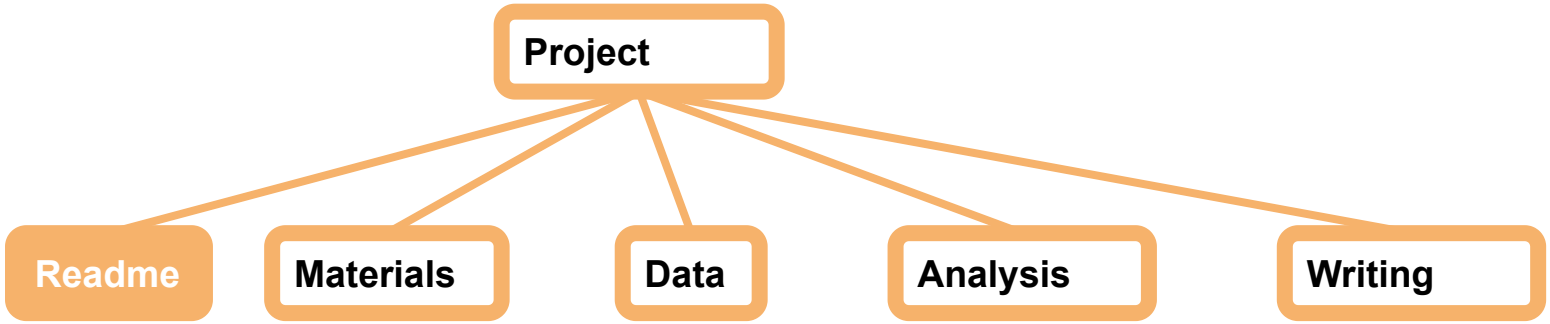


“Melissa. I am in the middle of a project.”

“Melissa. I am a grad student, I don’t choose where my lab stores things.”

It’s all okay!

- Embrace the mess - notice where you are and where you’d like to go
- Method 1: Start on your next project
- Method 2: “Read in” files to a new structure as you go
- Method *: Documentation is ALWAYS better than no documentation. Give others (and yourself) a map



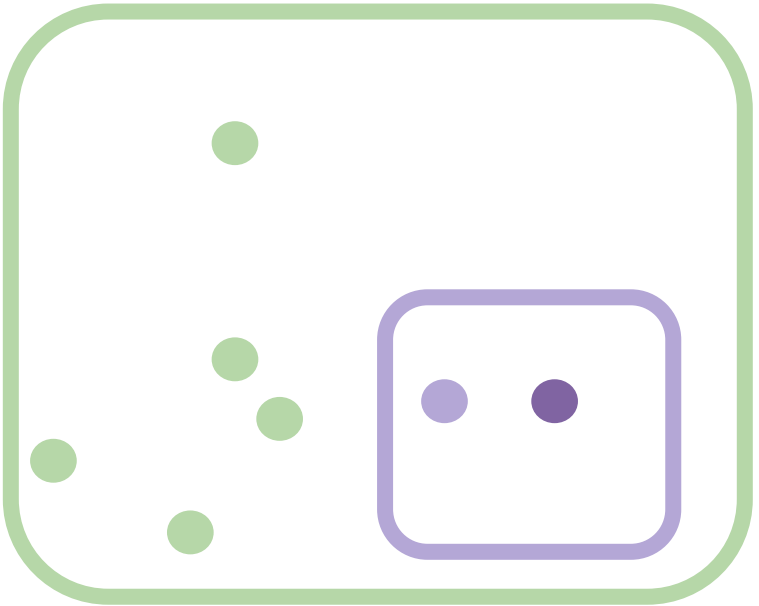
Readme.txt

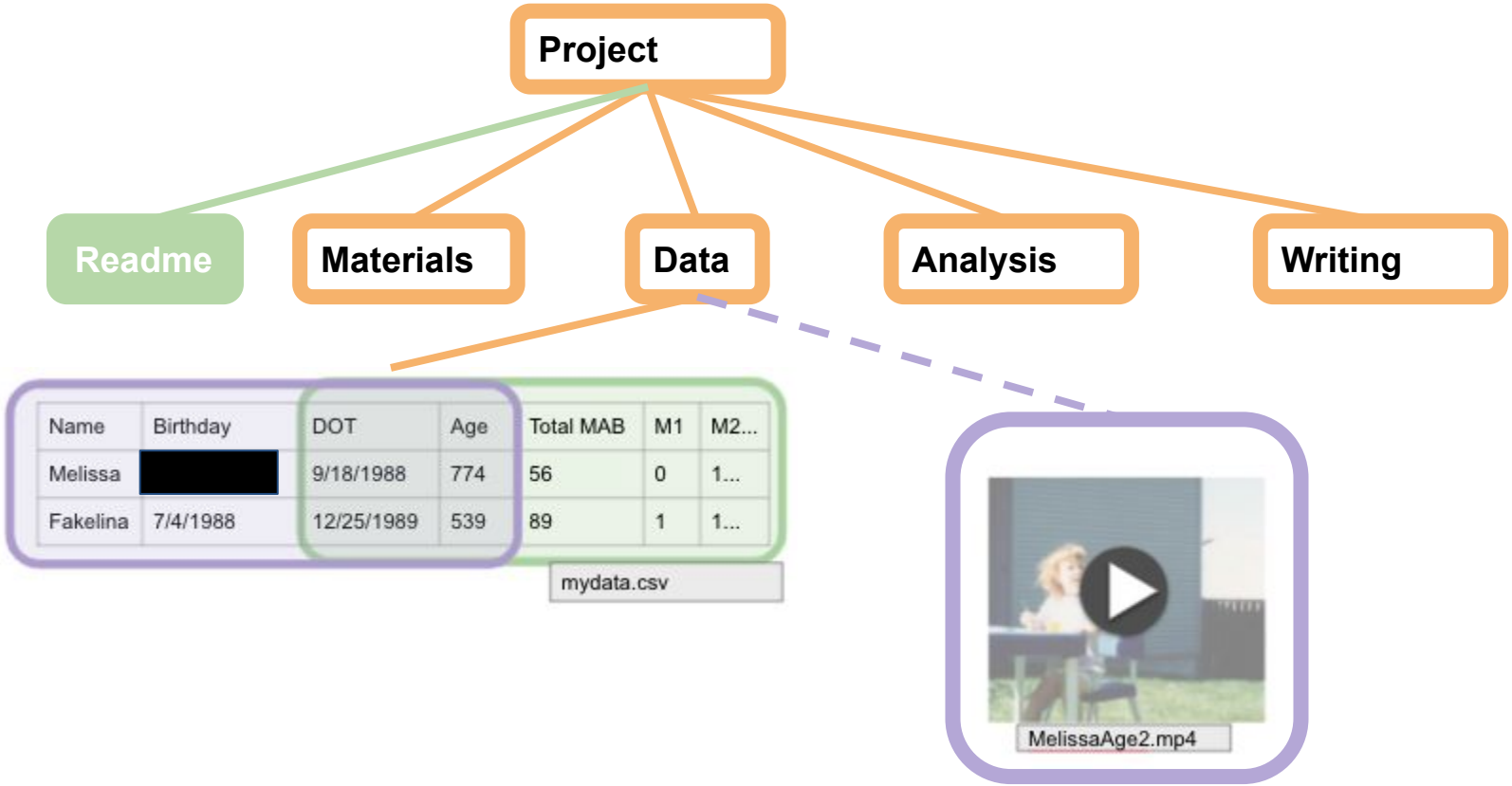
...

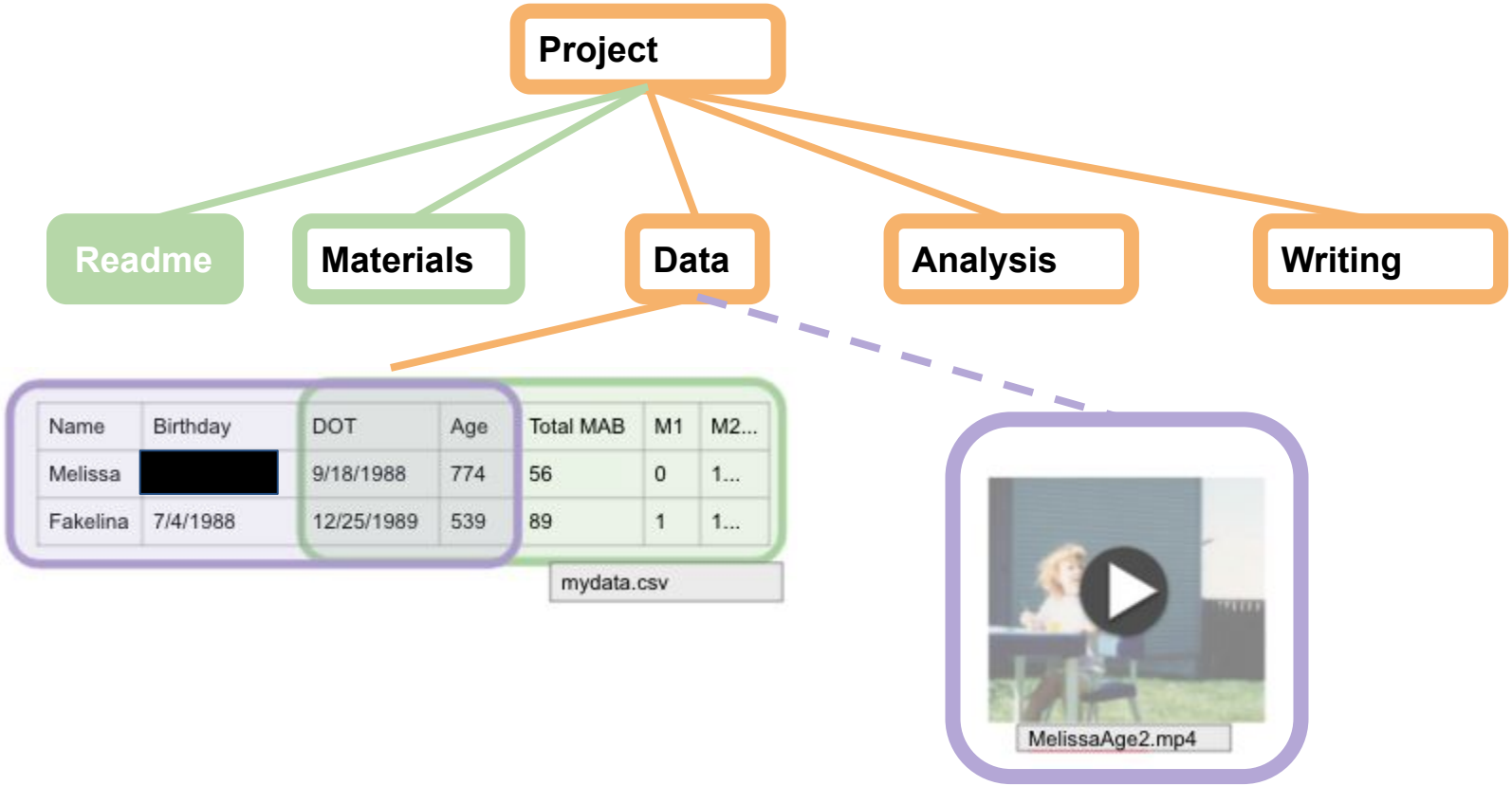
- Data: The data for this project is stored in two places:
  First, the response data files are stored in this folder, one per participant.
  Each participant also has a video with a corresponding filename. These videos are stored on <LAB SERVER> and can be accessed by <HOW TO ACCESS OR REQUEST ACCESS>

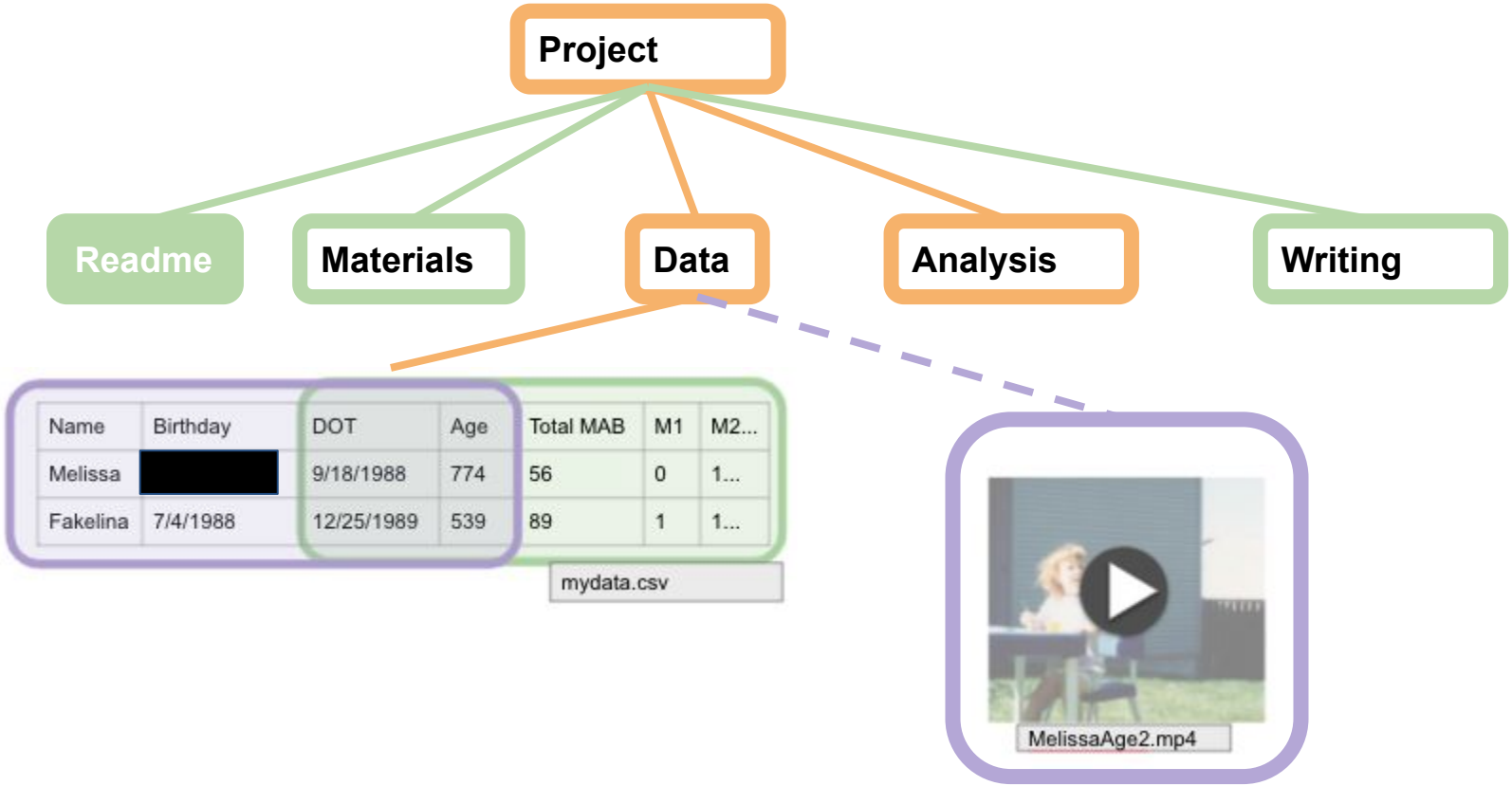
Isolate your radioactive materials!

- If you're still not sure, treat it as sensitive for now




Project
Readme
Materials
Data
Analysis
Writing
Name
Birthday
DOT
Age
Total MAB
M1
M2...
Melissa
9/18/1988
774
56
0
1...
Fakelina
7/4/1988
12/25/1989
539
89
1
1...
mydata.csv
MelissaAge2.mp4


Project
Readme
Materials
Data
Analysis
Writing
Name
Birthday
DOT
Age
Total MAB
M1
M2...
Melissa
9/18/1988
774
56
0
1...
Fakelina
7/4/1988
12/25/1989
539
89
1
1...
mydata.csv
MelissaAge2.mp4

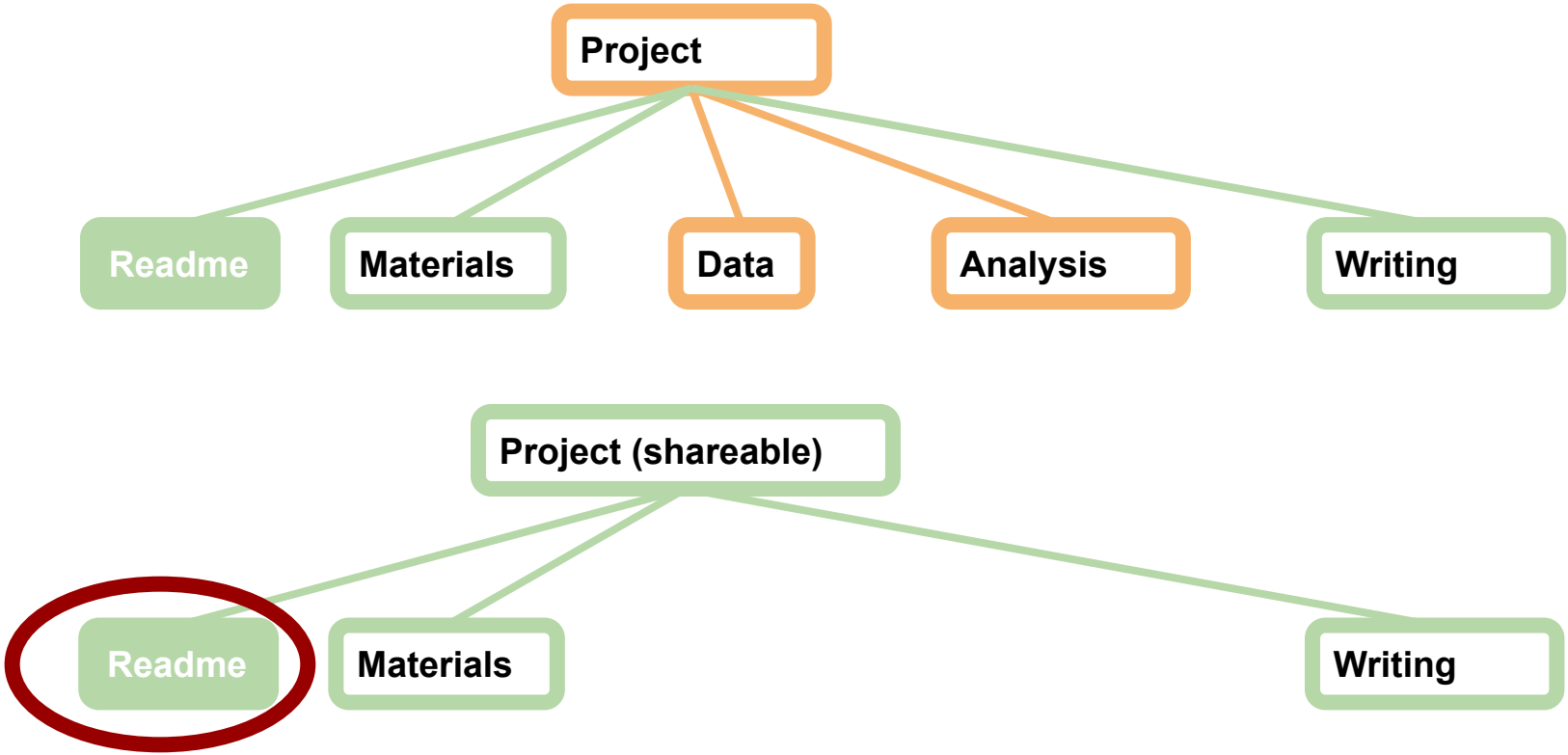

Project
Readme
Materials
Data
Analysis
Writing
Name
Birthday
DOT
Age
Total MAB
M1
M2...
Melissa
9/18/1988
774
56
0
1...
Fakelina
7/4/1988
12/25/1989
539
89
1
1...
mydata.csv
MelissaAge2.mp4

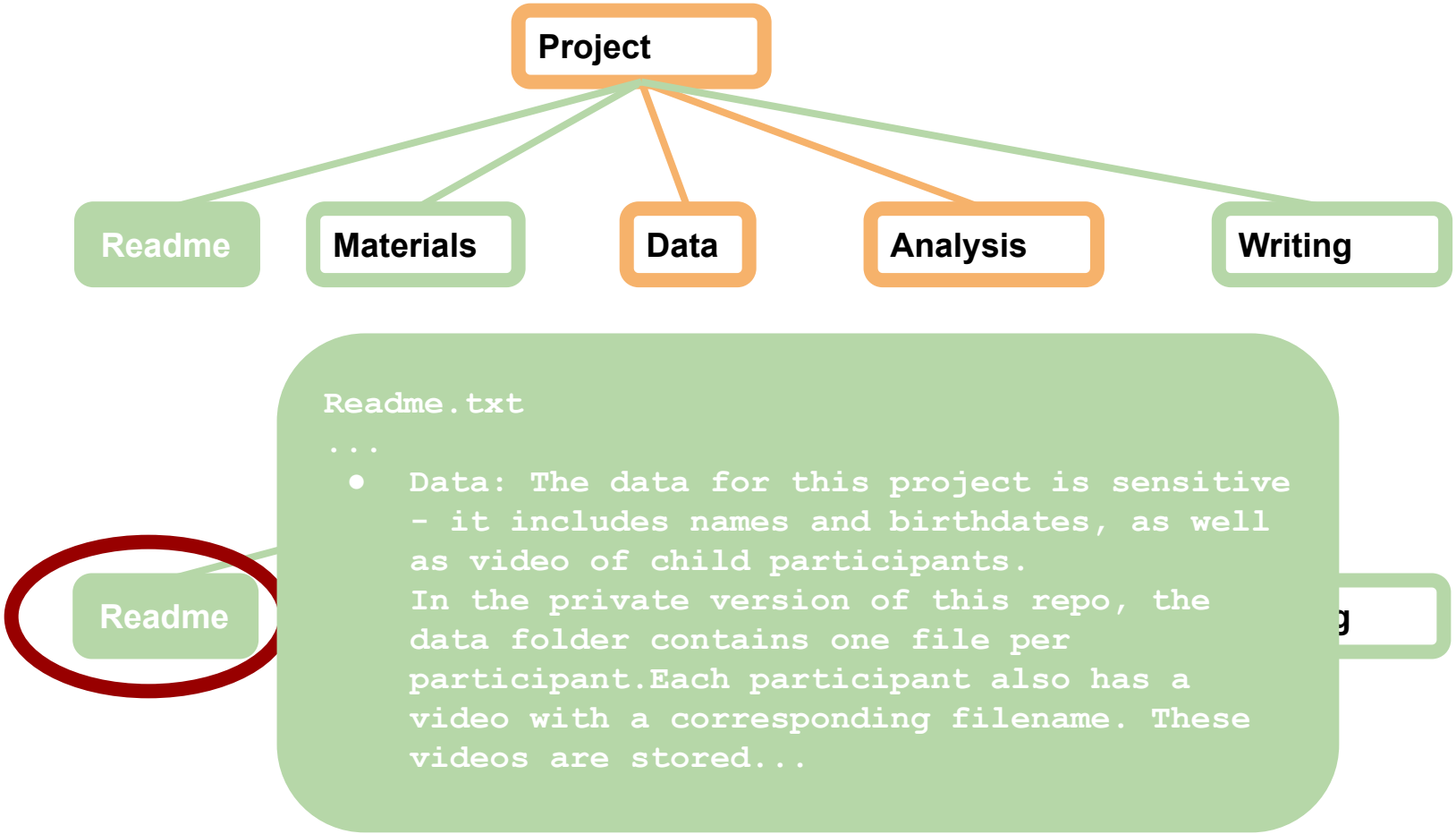
Lesson 1: All data files contain birthdays and faces until proven otherwise

Lesson 2: The best way to protect sensitive data is to not store it

Lesson 3: Isolate your sensitive data, then share it like you mean it

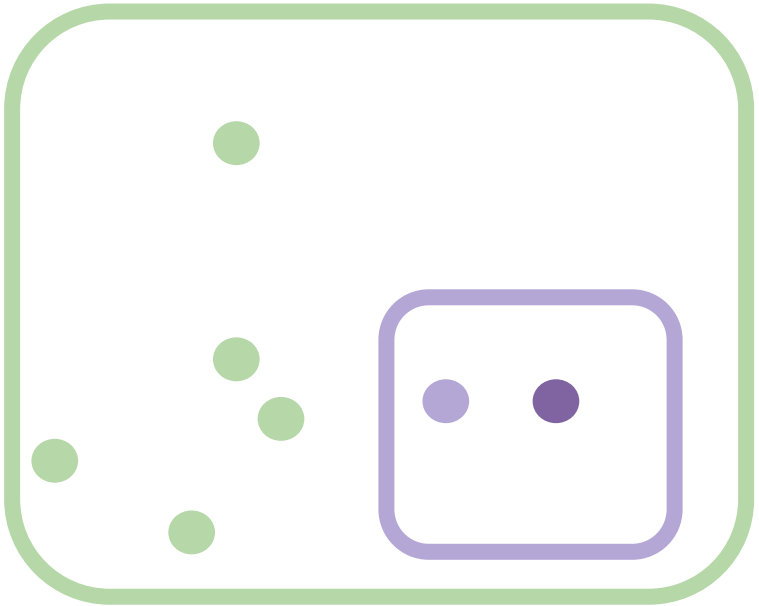
Lesson 4: As open as possible, as closed as necessary


Project
Readme
Materials
Data
Analysis
Writing
Project (shareable)
Readme
Materials
Writing


Project
Readme
Materials
Data
Analysis
Writing
Readme.txt
...
• Data: The data for this project is sensitive - it includes names and birthdates, as well as video of child participants.
In the private version of this repo, the data folder contains one file per participant.Each participant also has a video with a corresponding filename. These videos are stored...
Readme

Share it like you mean it

- Backup and version control
- Know who has access to what
- Remove sensitive data when sharing for any reason
- ...Unless they need it…
- ...In which case, follow a protocol



ID
HE251Q
RT52Z

Name	Birthday	DOT	Age	Total MAB	M1	M2...
Melissa	[redacted]	9/18/1988	774	56	0	1...
Fakelina	7/4/1988	12/25/1989	539	89	1	1...

mydata.csv

- Some properties of good unique IDs
  - Not (only) numerical/sequential
  - Not based on identifiable details
  - Not based on conditions/expectations of your study design!
  - Long enough to be unique

ID
HE251Q
RT52Z

Name	Birthday	DOT	Age	Total MAB	M1	M2...
Melissa	[redacted]	9/18/1988	774	56	0	1...
Fakelina	7/4/1988	12/25/1989	539	89	1	1...

mydata.csv

Age_Round
770
540

- Some good practices for obscuring birthdays
  - Exact age + DOT = Birthday
  - Consider precision of your analyses
  - Rounding vs. jittering
  - Age

ID
HE251Q
RT52Z

Name	Birthday	DOT	Age	Total MAB	M1	M2...
Melissa	[redacted]	9/18/1988	774	56	0	1...
Fakelina	7/4/1988	12/25/1989	539	89	1	1...

mydata.csv

Age_Round
770
540

ID
HE251Q
RT52Z

Data that identifies MKS (Video file, name, birthday, exact age)

mydata_identifiable.csv

Data about MCDI scores (DOT + aprox. age, not traceable to MKS)

mydata_deidentified.csv

ID	Age_Round
HE251Q	770
RT52Z	540

Lesson 1: All data files contain birthdays and faces until proven otherwise

Lesson 2: The best way to protect sensitive data is to not store it

Lesson 3: Isolate your sensitive data, then share it like you mean it

Lesson 4: As open as possible, as closed as necessary

Resources

NYU guide to the Open Science Framework

18 HIPAA Safe Harbor Identifiers

The Turing Way - Working philosophy for ethical sharing & research

Danielle Navarro's slides - Project structure and file naming

Earth Sciences article - The logic/benefits of project structure