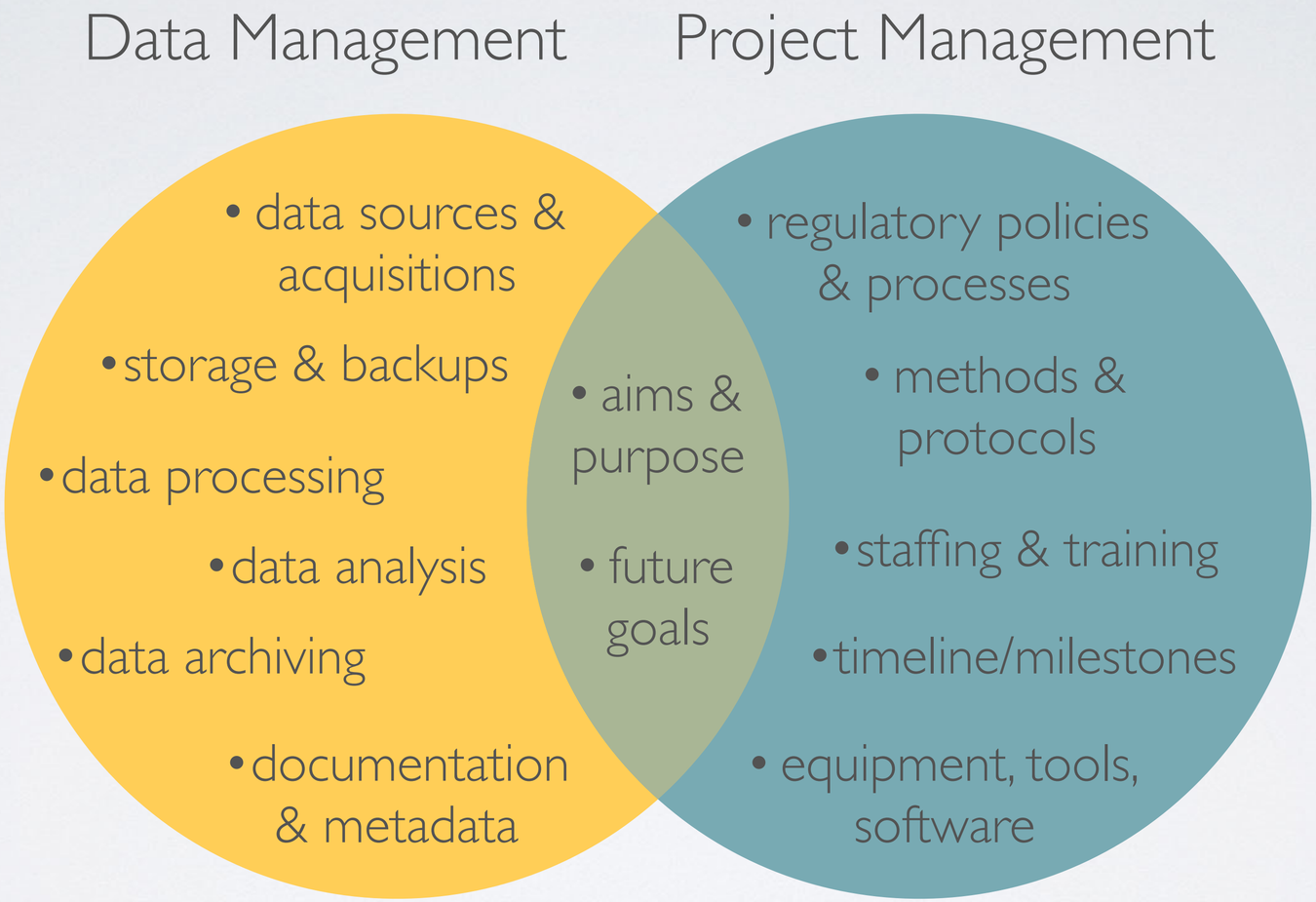
Good Enough Practices for Data and Project Management

Alaina Pearce

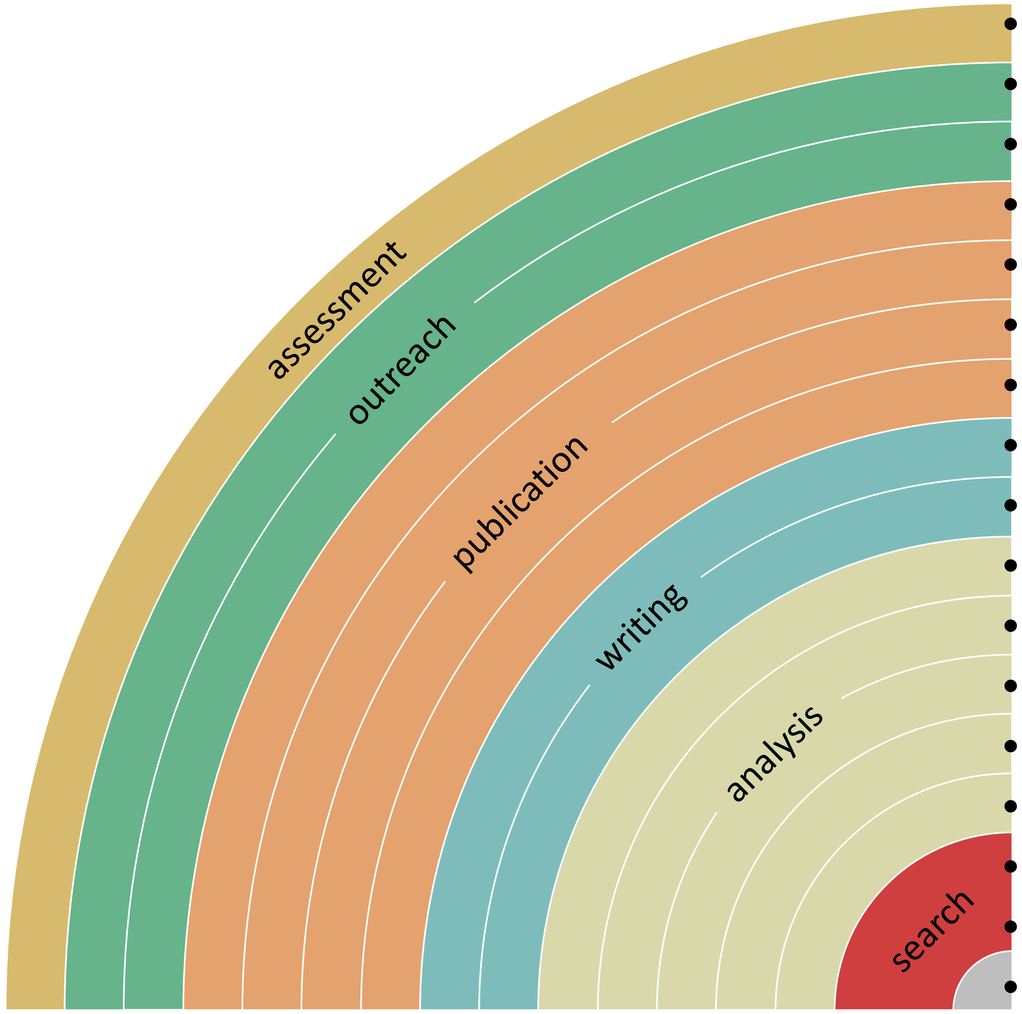
Open Science Bootcamp 2023

Project vs Data Management



Goal: extract meaningful insight and information

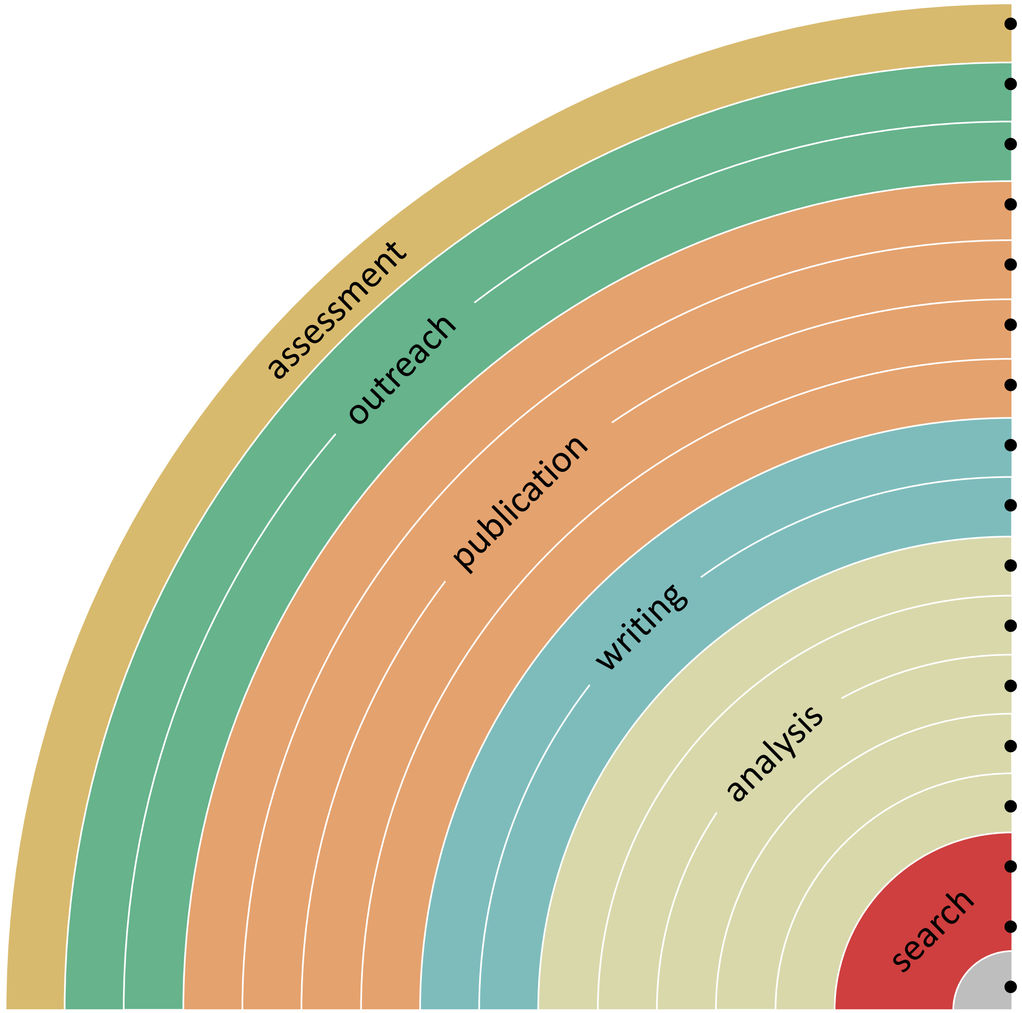
Goal: meet project goals within set timelines



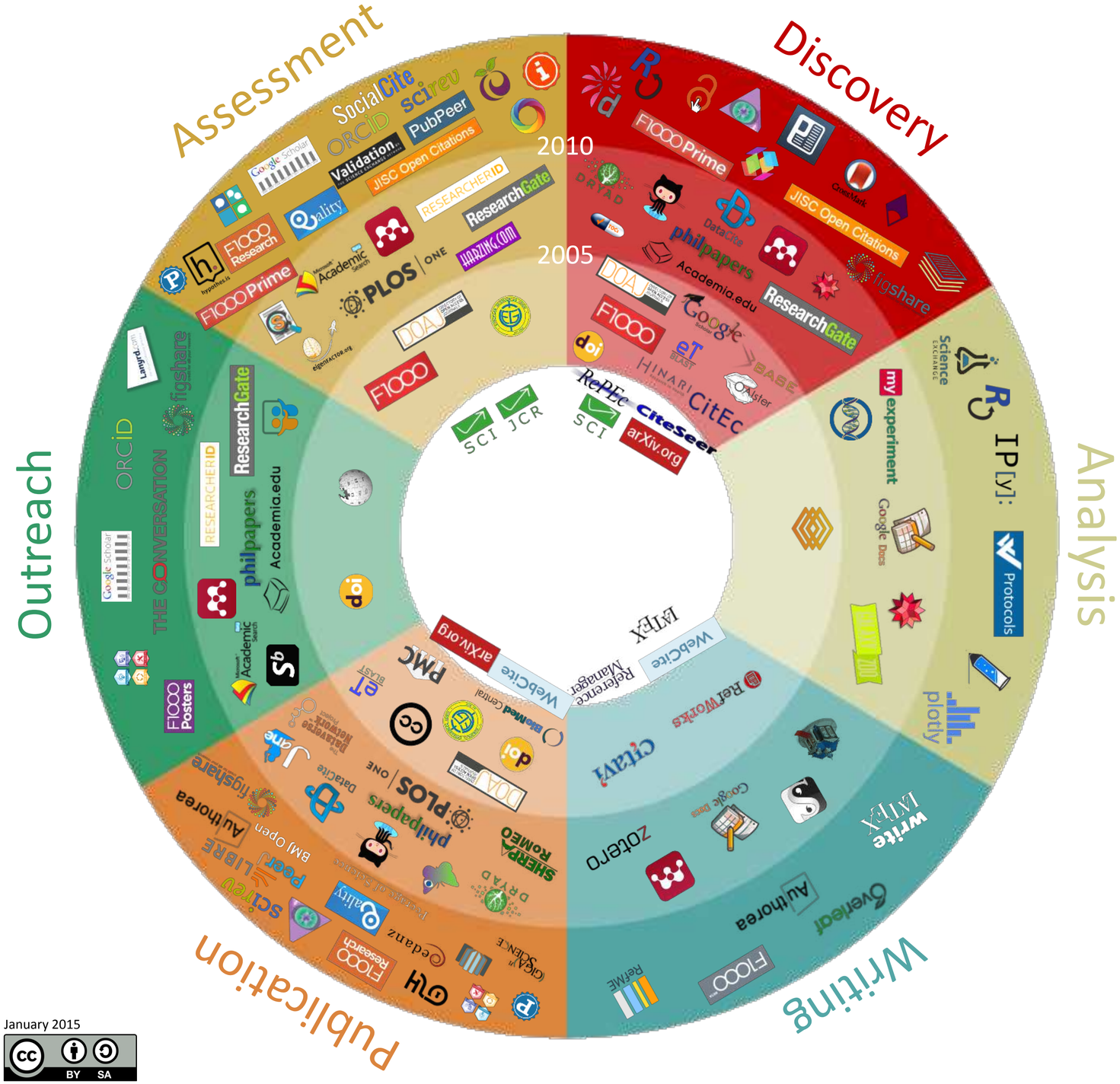
- adding alternative evaluation, e.g. with altmetrics
- communicating through social media, e.g. Twitter
- sharing posters & presentations, e.g. at FigShare
- using open licenses, e.g. CC0 or CC-BY
- publishing open access, ‘green’ or ‘gold’
- using open peer review, e.g. at journals or PubPeer
- sharing preprints, e.g. at OSF, arXiv or bioRxiv
- using actionable formats, e.g. with Jupyter or CoCalc
- open XML-drafting, e.g. at Overleaf or Authorea
- sharing protocols & workfl., e.g. at Protocols.io
- sharing notebooks, e.g. at OpenNotebookScience
- sharing code, e.g. at GitHub with GNU/MIT license
- sharing data, e.g. at Dryad, Zenodo or Dataverse
- pre-registering, e.g. at OSF or AsPredicted
- commenting openly, e.g. with Hypothes.is
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. at RIO

Bianca Kramer & Jeroen Bosman https://101innovations.wordpress.com

[DOI: 10.5281/zenodo.1147025](https://doi.org/10.5281/zenodo.1147025)



Bianca Kramer & Jeroen Bosman https://101innovations.wordpress.com



January 2015

CC BY SA

all logos excluded

	Discovery	Analysis	Writing	Publication	Outreach	Assessment
Traditional	Web of Science	SPSS	Word → EndNote	nature	ResearcherID	Journal Citation Reports
Modern	Google Scholar	R	RefWorks	arXiv.org / PLOS ONE	Institutional Repository	eigenFACTOR.org / PLOS Article-Level Metrics
Innovative	Mendeley	Zooniverse / Labspace	Authorea / Mekentosj	figshare / PeerJ	ResearchGate / ORCID	Peerage of Science / Altmetric
Experimental	sparrho / d	RunSciOpen / IP[y]: Nb	Overleaf / Docear	the WINNOWER	Labtrove	PubPeer
Google	Google Scholar / Google Books	Google Drive	Google Research	?	Google Scholar	Google Scholar / Google Scholar Metrics
NPG/Macmillan	Nature Publishing Group	labguru	Overleaf	figshare / nature	Nature Blogs	nature INDEX Beta / Altmetric

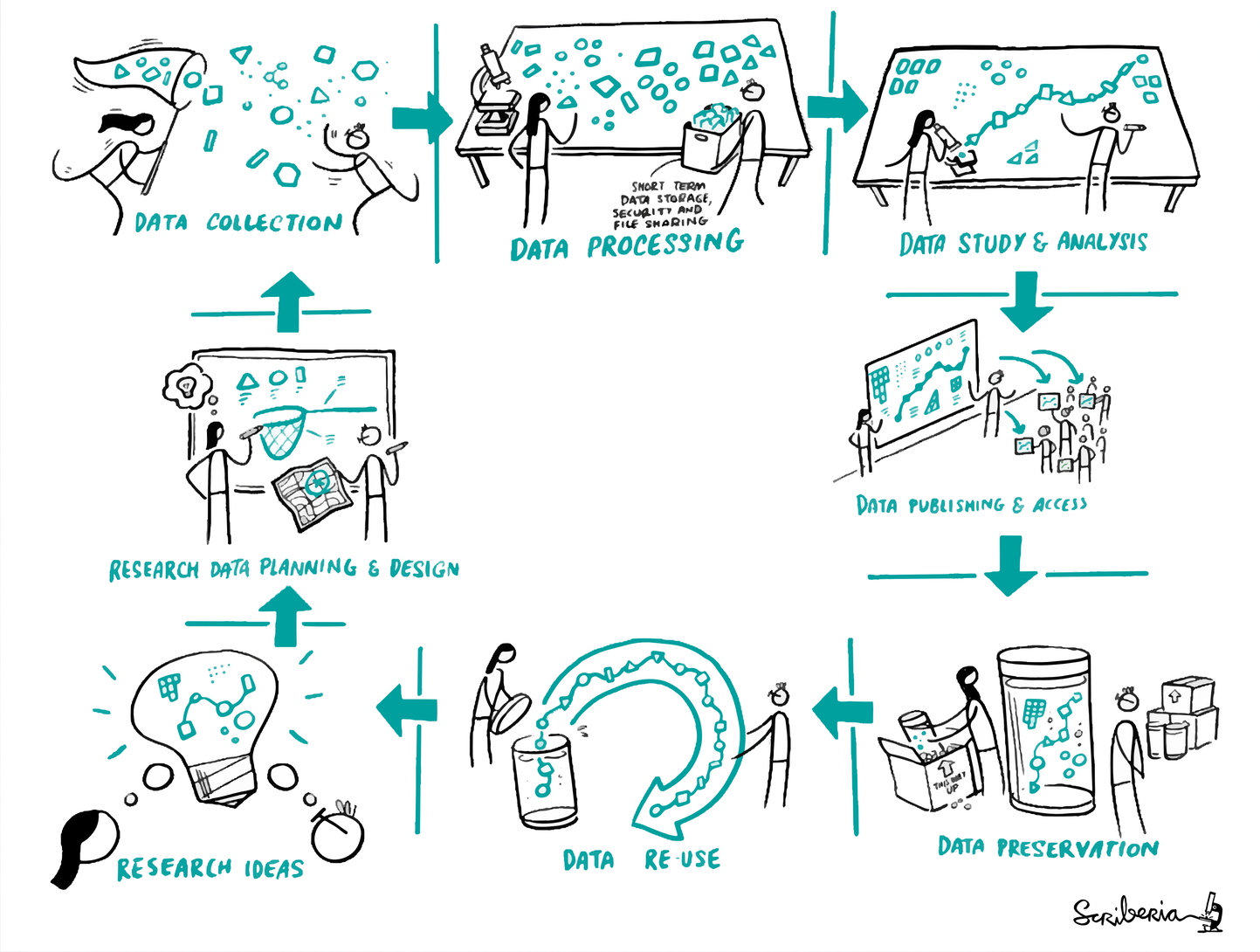
Bianca Krame
all logos excluded

With What Time???

‘Good Enough’

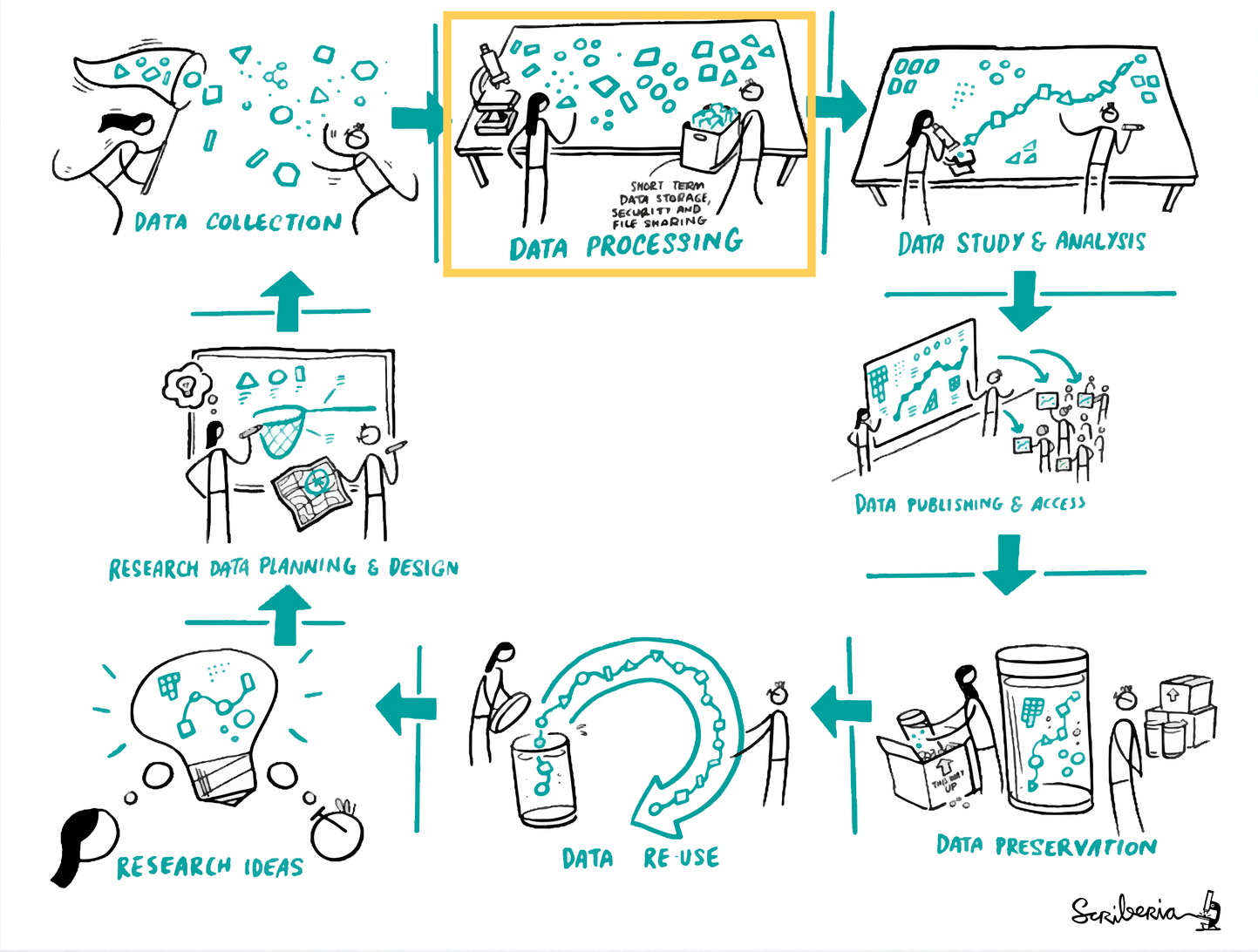
- (relatively) low effort
- shallow learning curve
- beneficial to current and future you
- increases ‘openness’ of research

Project Lifecycle



The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 license. DOI: 10.5281/zenodo.3332807.

Project Lifecycle



The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 license. DOI: 10.5281/zenodo.3332807.

Data Processing Pipelines

The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 license. DOI: 10.5281/zenodo.3332807.

1. Preserve Raw Data

Raw Data: data as it was originally collected

Save in data in its original form and DO NOT alter or ‘improve’ it

1. Preserve Raw Data

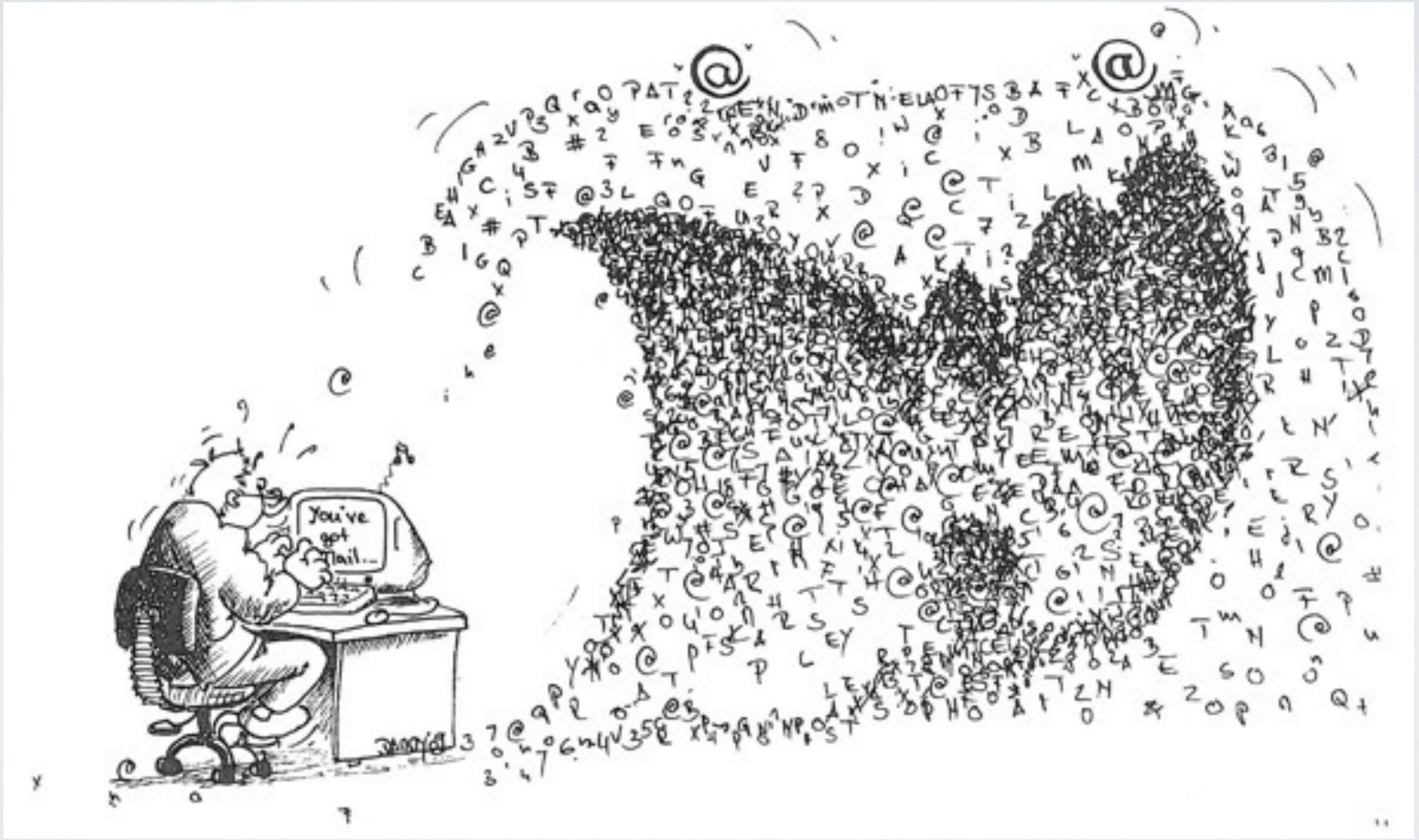
Raw Data: data as it was originally collected

Save in data in its original form and DO NOT alter or 'improve' it

What makes this 'Open'?

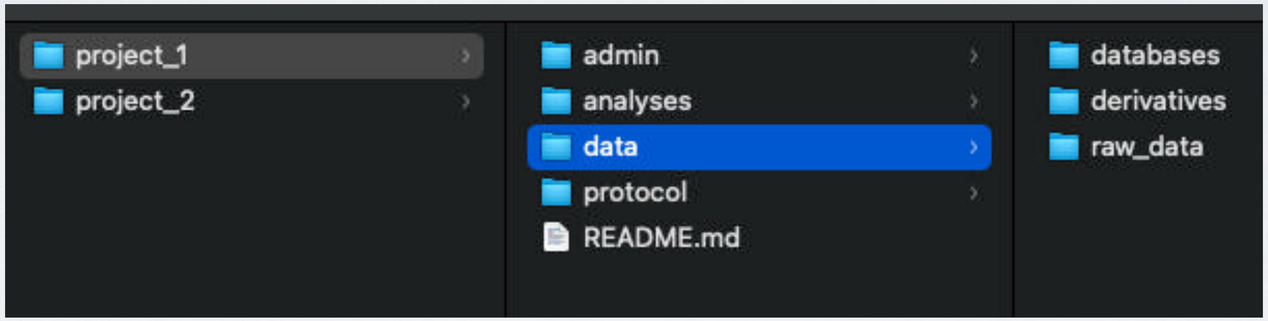
- Stable starting point
- Test reproducibility of pipeline
- Recover from mishaps
- Experiment without fear

Data Tsunami



2. Create a Central Hub

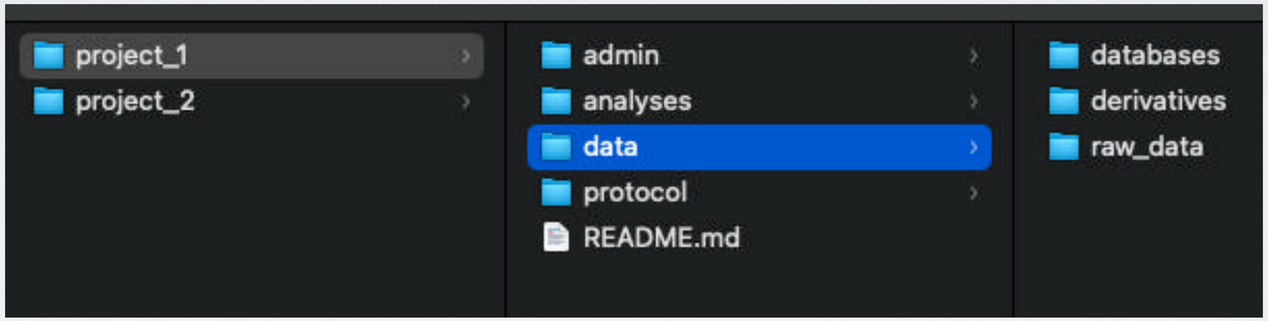
Directory Structures: organization of files into a hierarchical structure



- Create a directory for each project
- Use a consistent structure
- Separate data management from project management

2. Create a Central Hub

Directory Structures: organization of files into a hierarchical structure



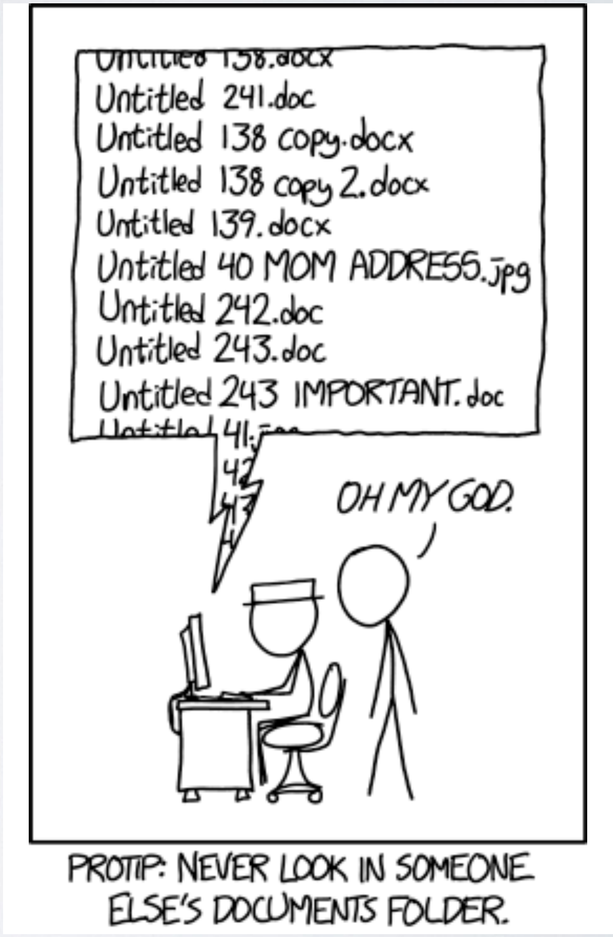
- Create a directory for each project
- Use a consistent structure
- Separate data management from project management

What makes this 'Open'?

- Easy to find data, code, protocol
- Consistent (at least within lab)
- Bigger Lift: match field standards (e.g. , BIDS, MIxS)

3. Use Meaningful Names

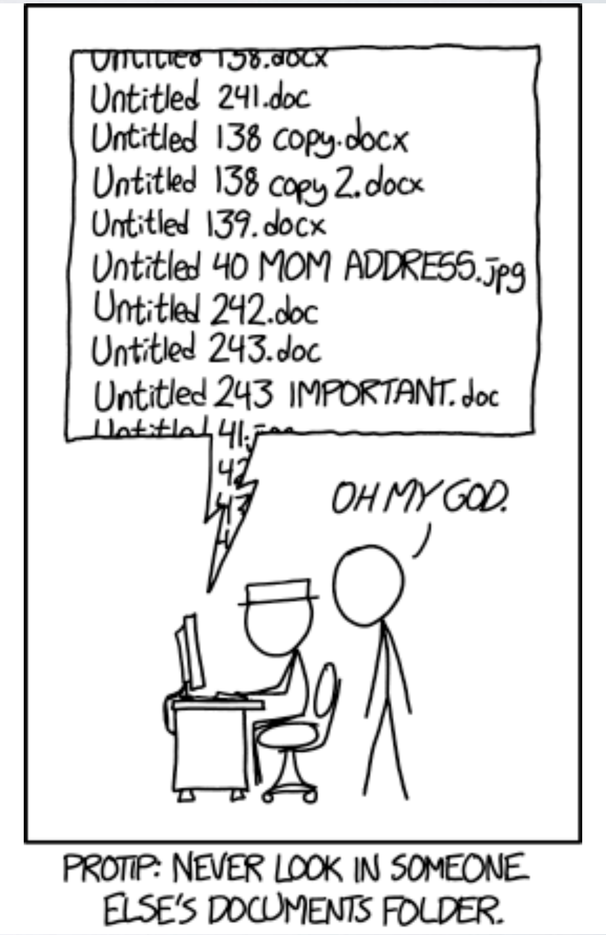
Leverage filenames to help you manage complex projects



3. Use Meaningful Names

Leverage filenames to help you manage complex project

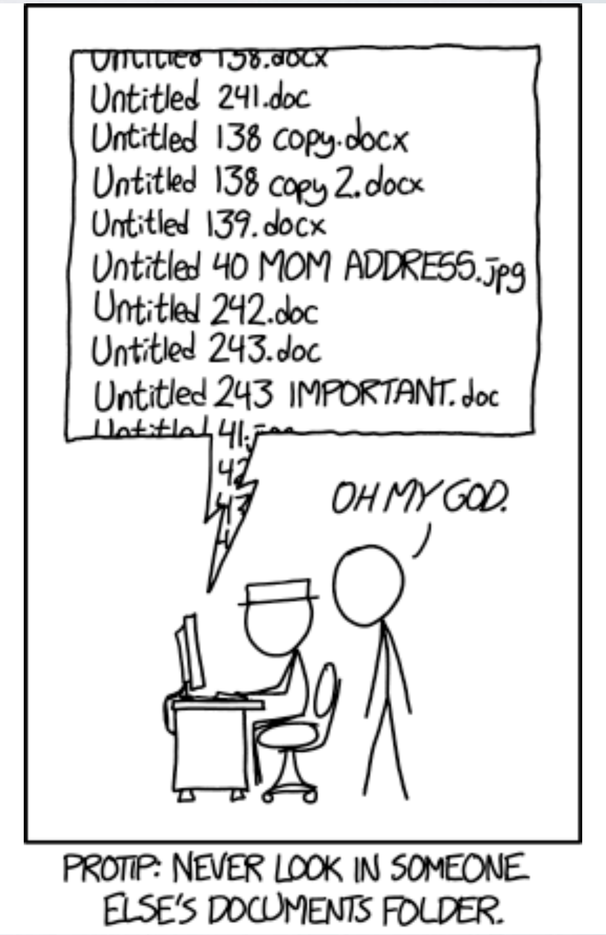
- Human Readable: names should clearly describe content in the simplest way possible (e.g., ‘code’, ‘data’)



3. Use Meaningful Names

Leverage filenames to help you manage complex project

- Human Readable: names should clearly describe content in the simplest way possible (e.g., 'code', 'data')
- Computer Readable: ability of a computer to parse a name
  - Use '-' or '_' in place of spaces
  - No special characters (e.g, '&', '#', '^', etc)



3. Use Meaningful Names

Leverage filenames to help you manage complex project

- Human Readable: names should clearly describe content in the simplest way possible (e.g., 'code', 'data')
- Computer Readable: ability of a computer to parse a name
  - Use '-' or '_' in place of spaces
  - No special characters (e.g, '&', '#', '^', etc)
- Sortable: help you find what you need in the future
  - Dates: YYYY-MM-DD
  - Pad with zeros (subject IDs, versions, etc)

3. Use Meaningful Names

Leverage filenames to help you manage complex project

- Human Readable: names should clearly describe content in the simplest way possible (e.g., 'code', 'data')
- Computer Readable: ability of a computer to parse a name
  - Use '-' or '_' in place of spaces
  - No special characters (e.g, '&', '#', '^', etc)
- Sortable: help you find what you need in the future
  - Dates: YYYY-MM-DD
  - Study IDs: Pad with zeros

What makes this 'Open'?

- Makes data more findable
- Can be a form of metadata
- Bigger Lift: adopt field standards

4. Preserve the Journey

Version control: tracking and managing changes to documents or code

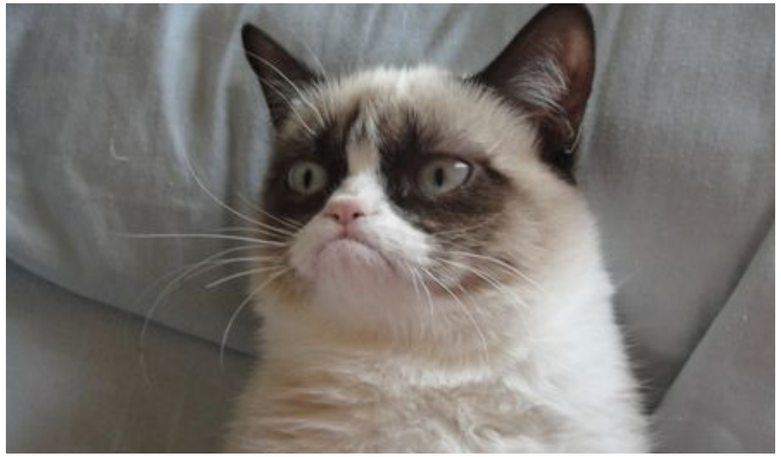
4. Preserve the Journey

Version control: tracking and managing changes to documents or code

- Manual: use file naming to document drafts (e.g., dates, version numbers)
- Software: git, GitHub, subversion
- Allows you to trace your steps

4. Preserve the Journey

Version control: tracking and managing changes to documents or code



99 little bugs in the code
99 little bugs
Take one down and compile it
117 little bugs in the code...

- Manual: use file naming to document drafts (e.g., dates, version numbers)
- Software: git, GitHub, subversion
- Allows you to trace your steps

4. Preserve the Journey

Version control: tracking and managing changes to documents or code

What makes this 'Open'?

- Documents project and data history
- Can reproduce process if needed
- Bigger Lift: use a version control software (e.g., git)

- Manual: use file naming to document drafts (e.g., dates, version numbers)
- Software: git, GitHub, subversion
- Allows you to trace your steps

5. Avoid Manual Manipulations

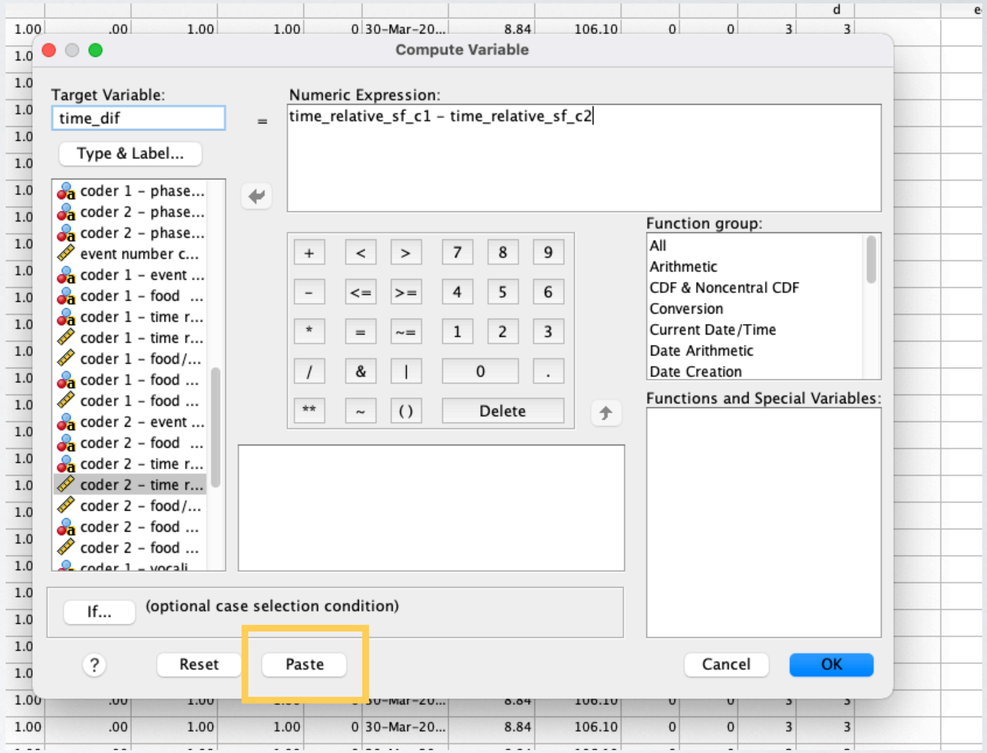
- Manual data manipulations leave no trace
  - Hard to reproduce
  - Error prone

- Alternatives:
  - Save Syntax in SPSS
  - Include calculations in variable descriptions
  - Script data cleaning

5. Avoid Manual Manipulations

- Manual data manipulations leave no trace
  - Hard to reproduce
  - Error prone

- Alternatives:
  - Save Syntax in SPSS
  - Include calculations in variable descriptions
  - Script data cleaning



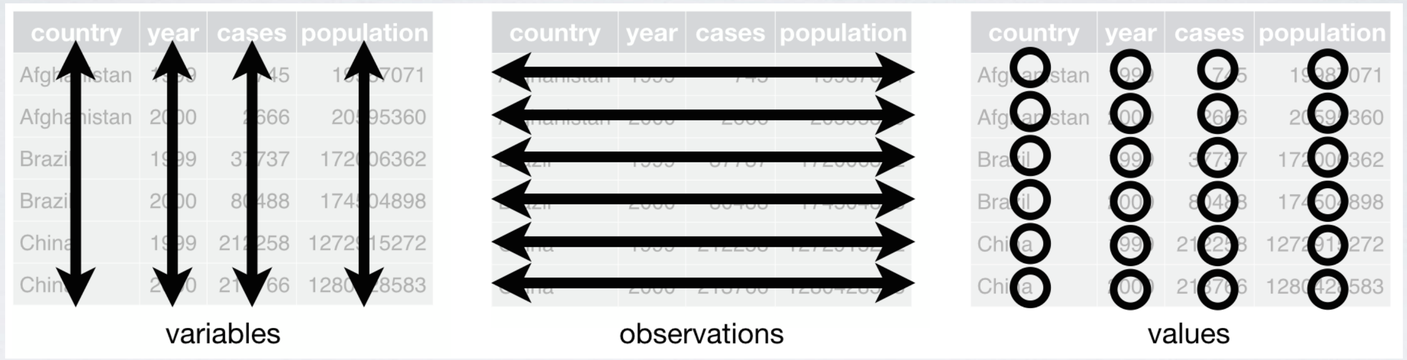
5. Avoid Manual Manipulations

- Manual data manipulations leave no trace
  - Hard to reproduce
  - Error prone
- Alternatives:
  - Save Syntax in SPSS
  - Include calculations in variable descriptions
  - Script data cleaning

What makes this 'Open'?

- Data processing will be reproducible
- Can reverse to original data if needed
- Bigger Lift: move away from GUI-based analysis software to open code/syntax based programs (e.g., R, python)

6. ‘Tidy’ Your Data



- Every variable is in its own column
- Each participant/sample is in its own row
- Each value is in its own cell

6. ‘Tidy’ Your Data

- Use open file formats — csv, html, txt, jpeg

6. 'Tidy' Your Data

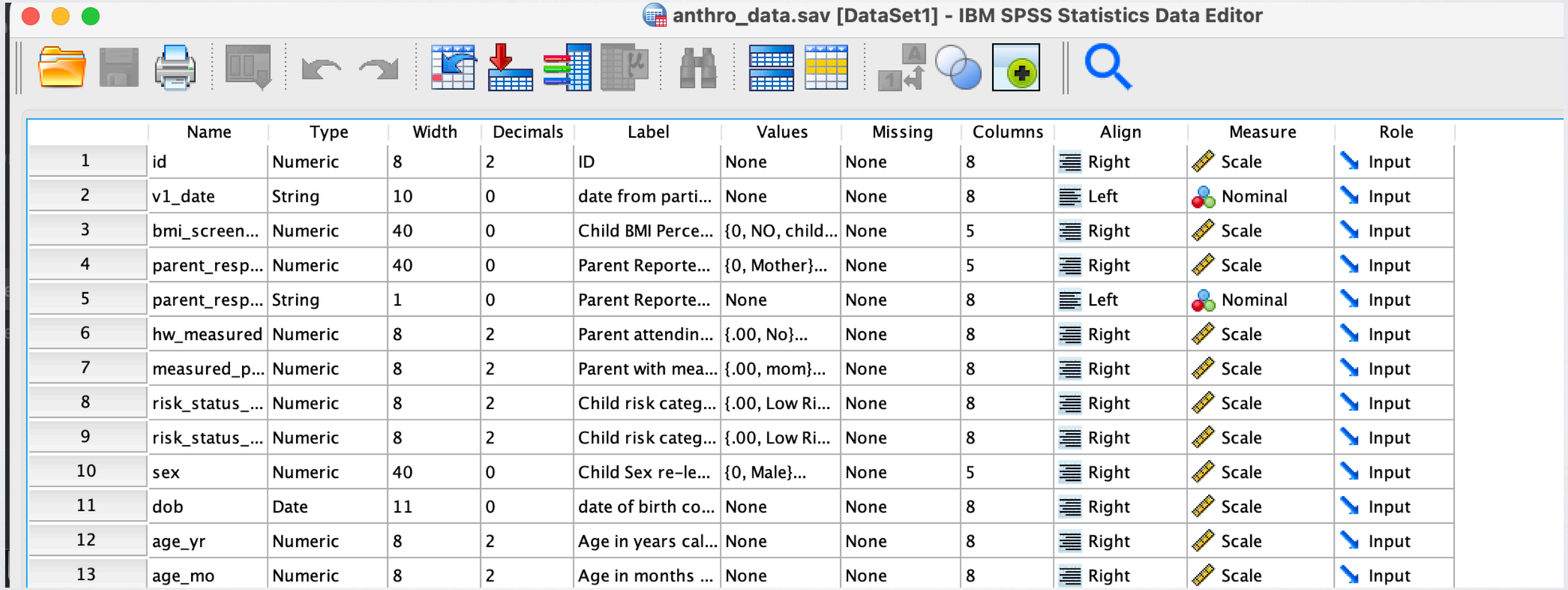
- Use open file formats — csv, html, txt, jpeg
- Create a data dictionary

variable	explanation	examples	prev work w this variable	Why?
Frame of reference	**Who is the referent compared to?**			
	A group of ppl = **class**	"tu eres muy ñoño"	BCT, 2012	
	Itself in another time = **individual**	"estoy negra"	diaz-campos Geeslin, 2011	individual favor estar
Susceptibility to change	**Is it changeable?**			
	Yes	hair	diaz-campos Geeslin, 2011	changeable favored estar This variable has been proven to be a predictor of estar use across varieties of Spanish (Geeslin 2003; Geeslin & Guijarro-Fuentes 2008; Silva-Corvalán 1994).
	No	building		
Experience with the referent	**How much the speaker knows bout the referent**			
	speaker has some previous knowledge = **ongoing**		BCT, 2012	immediate, use of estar is more likely than ser
	speaker just encountered the referent = **immediate**	"la casa esta grande"		
	speaker has previous experience = **direct**			
Resultant state	**Is it a participial adjective/perfective?**			
	Yes = [+ resultant]	el piso esta mojado	Diaz-Campos Geeslin, 2011	[+ resultant] adj favor estar
	No = [- resultant]	el piso es muy bonito		
Animacy	**Is the referent animate?**			
	Yes	el cabello esta largo		
	No	el departamento es muy bonito		

Example provided by Dr. Blanchette, Penn State Center for Language Science, Department of Psychology

6. ‘Tidy’ Your Data

- Use open file formats — csv, html, txt, jpeg
- Create a data dictionary



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	8	2	ID	None	None	8	Right	Scale	Input
2	v1_date	String	10	0	date from parti...	None	None	8	Left	Nominal	Input
3	bmi_screen...	Numeric	40	0	Child BMI Perce...	{0, NO, child...	None	5	Right	Scale	Input
4	parent_resp...	Numeric	40	0	Parent Reporte...	{0, Mother}...	None	5	Right	Scale	Input
5	parent_resp...	String	1	0	Parent Reporte...	None	None	8	Left	Nominal	Input
6	hw_measured	Numeric	8	2	Parent attendin...	{.00, No}...	None	8	Right	Scale	Input
7	measured_p...	Numeric	8	2	Parent with mea...	{.00, mom}...	None	8	Right	Scale	Input
8	risk_status_...	Numeric	8	2	Child risk categ...	{.00, Low Ri...	None	8	Right	Scale	Input
9	risk_status_...	Numeric	8	2	Child risk categ...	{.00, Low Ri...	None	8	Right	Scale	Input
10	sex	Numeric	40	0	Child Sex re-le...	{0, Male}...	None	5	Right	Scale	Input
11	dob	Date	11	0	date of birth co...	None	None	8	Right	Scale	Input
12	age_yr	Numeric	8	2	Age in years cal...	None	None	8	Right	Scale	Input
13	age_mo	Numeric	8	2	Age in months ...	None	None	8	Right	Scale	Input

6. 'Tidy' Your Data

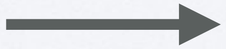
- Use open file formats — csv, html, txt, jpeg
- Create a data dictionary

column	variable	label	value_labels	type	n_na	range
1	id	ID	NULL	double	0	c(1, 133)
2	v1_date	date from participant contacts databases ('verified_visit_da	NULL	character	0	c("2018-01-31", "2022-05-07")
3	bmi_screenout	Child BMI Percentile Screen Out	c(`YES, child is overweight, sc	double	0	c(0, 1)
4	parent_respondent	Parent Reported: Parent relationship to child re-leveled in R	c(Mother = 0, Father = 1, Oth	double	0	c(0, 1)
5	parent_respondent_o	Parent Reported: Parent specify relationship to child if other	NULL	character	0	c("", "")
6	hw_measured	Parent attending Visit 1 had measured height and weight	c(No = 0, Yes = 1)	double	0	c(1, 1)
7	measured_parent	Parent with measured BMI at Visit 1	c(mom = 0, dad = 1)	double	0	c(0, 1)
8	risk_status_mom	Child risk categor: Low risk: Mom BMI < 26, High Risk: Mom	c(`Low Risk` = 0, `High Risk` =	double	0	c(0, 1)
9	risk_status_both	Child risk category: Low Risk: Mom and Dad BMI < 25, High	c(`Low Risk` = 0, `High Risk` =	double	0	c(0, 2)
10	sex	Child Sex re-leveled in R to start with 0	c(Male = 0, Female = 1)	double	0	c(0, 1)
11	dob	date of birth converted to format yyyy-mm-dd in R	NULL	double	0	c(14333, 16391)
12	age_yr	Age in years calculated from dob and start_date	NULL	double	0	c(7, 8.99)
13	age_mo	Age in months calculated from dob and start_date	NULL	double	0	c(84, 107.9)
14	ethnicity	Parent Reported: Child ethnicity	c(`NOT Hispanic or Latino` = 0	double	0	c(0, 0)
15	race	Parent Reported: Child race -- Note: prefer not to answer (p	c(`White/Caucasian` = 0, `Am	double	0	c(0, 2)
16	income	Parent Reported: Yearly household income -- Note: prefer n	c(`Less than $20,000` = 0, `$2(	double	3	c(0, 5)
17	parent_ed	Parent Reported: Parent education re-leveled in R to start w	c(`High School or GED (12 yea	double	0	c(0, 5)

6. ‘Tidy’ Your Data

- Use open file formats — csv, html, txt, jpeg
- Create a data dictionary
- One piece of information per cell

height
5 ft 6 in
5 ft 2 in
7 ft
5 ft 11 in

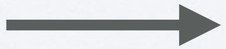


height_ft	height_in
5	6
5	2
7	0
5	11

6. 'Tidy' Your Data

- Use open file formats — csv, html, txt, jpeg
- Create a data dictionary
- One piece of information per cell
- Do not use highlighting/font color as data

height
5 ft 6 in
5 ft 2 in
7 ft
5 ft 11 in



height_ft	height_in	check_height
5	6	0
5	2	0
7	0	1
5	11	0

6. ‘Tidy’ Your Data

- Use open file formats — csv, html, txt, jpeg
- Create a data dictionary
- One piece of information per cell
- Do not use highlighting/font color as data

What makes this ‘Open’?

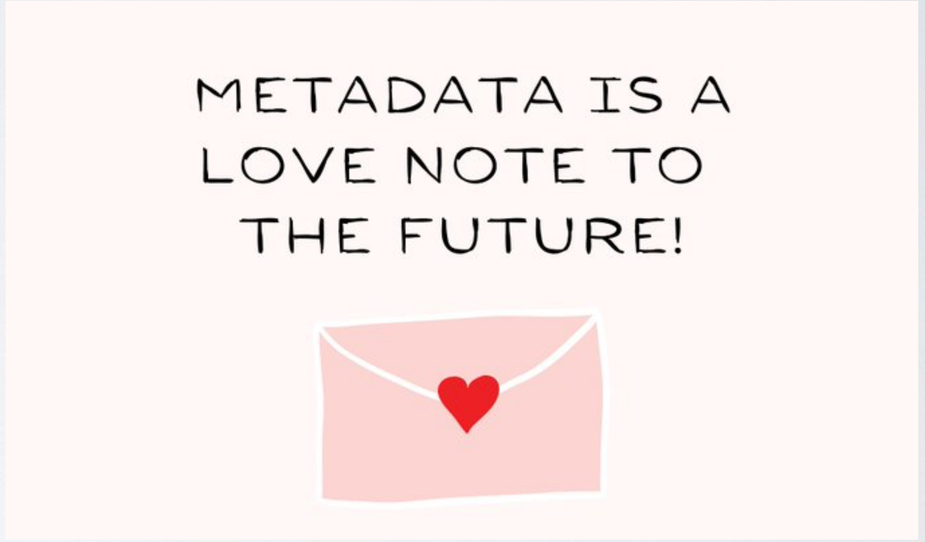
- Open formats are accessible
- All data are computer readable
- Data are documented
- Makes data re-use and sharing easier

7. Metadata Magic

Metadata: the who, what, when, where, and why of your data

What makes this 'Open'?

- Makes data more findable
- Helps others (and future you) understand the data
- Shared vocabularies help to harmonize data within a field

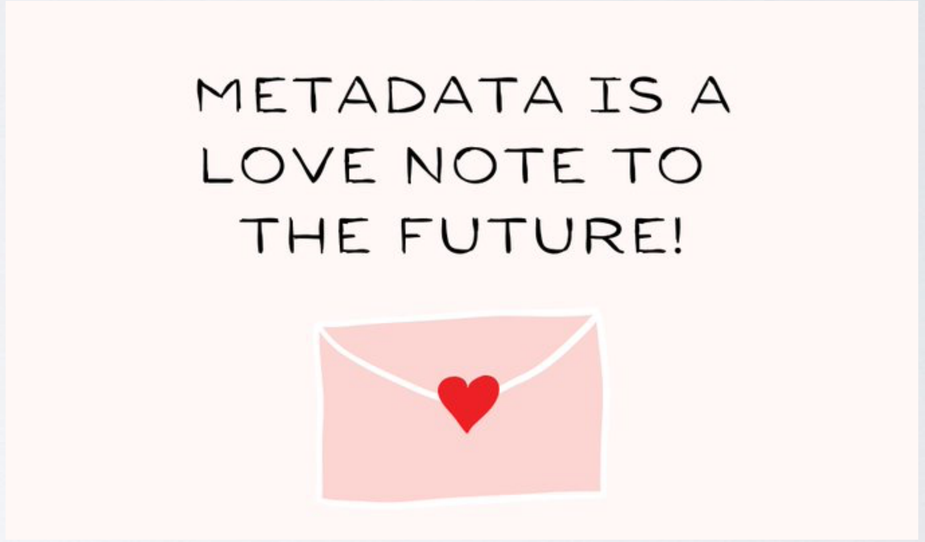


7. Metadata Magic

Metadata: the who, what, when, where, and why of your data

Easiest: when in doubt, document

- Data dictionaries
- Standard operating procedures manuals
- Lab notebooks
- changelog file (document versions)
- README
  - Description of folders/files
  - Can provide instructions on use of code/ data
  - License information



7. Metadata Magic

Metadata: the who, what, when, where, and why of your data

Easiest: when in doubt, document

- README

```
FACTORS (same for both PA and NE data frames)

Participant: Each experiment participant was assigned a unique ID

Item: Each individual sentence presented to participants was also associated with a unique ID

SentType: Sentence Type (2-word-order-errors, be-drop, causative, factive-say, like, need, negative-concord,
non-factive-believe, raising-seem, singular-dont, want)

Rating: Sentences were rated on a scale of 1-7 based on how "natural" participants thought they sounded

TrialOrder: Sequence in which a given item was presented during the experiment (items were presented in randomized order
so no 2 participants saw the items in the exact same order)
```

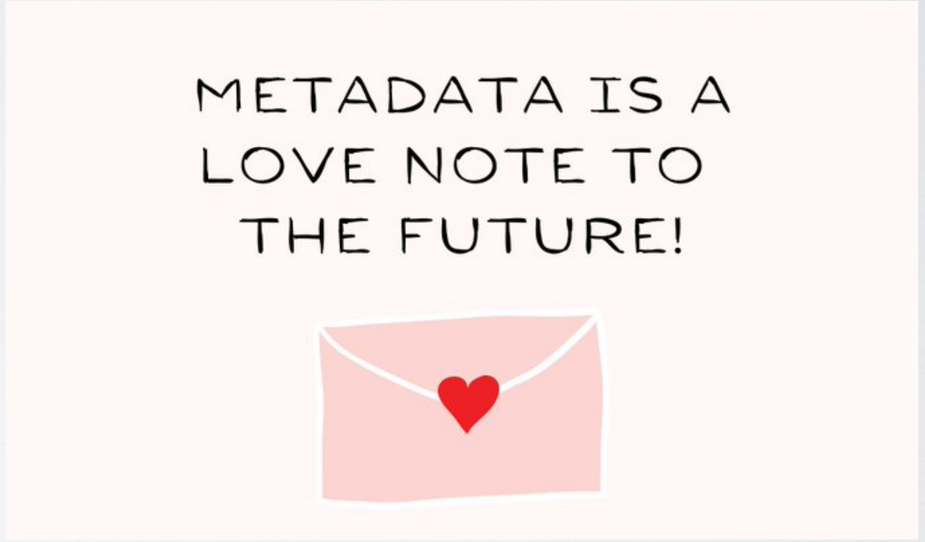
Example provided by Dr. Blanchette, Penn State Center for Language Science, Department of Psychology

7. Metadata Magic

Metadata: the who, what, when, where, and why of your data

Medium Effort: Data Manual

- Larger
- More verbose and detailed
- Can include science/rational/citations
- Like a user manual for data



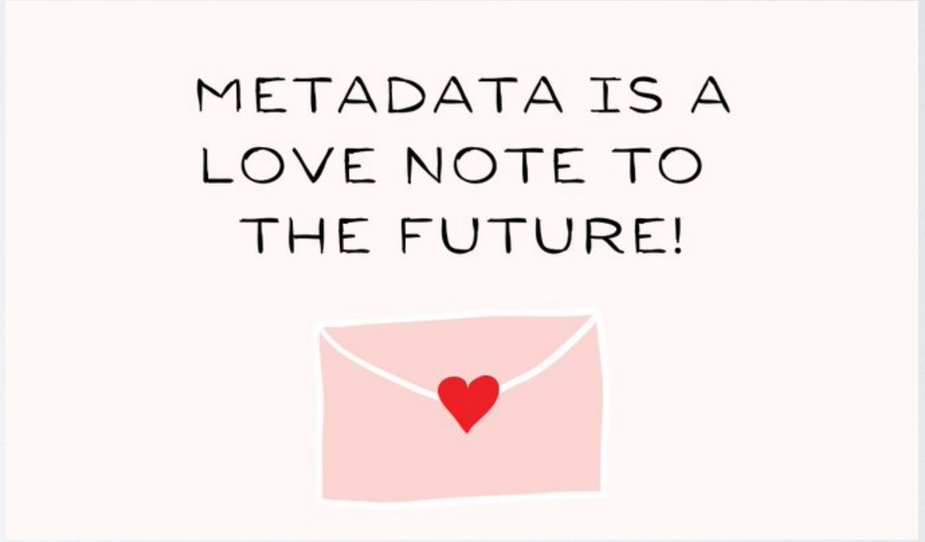
7. Metadata Magic

Metadata: the who, what, when, where, and why of your data

Medium Effort: Data Manual

Data Manual - R01: Food and Brain Study

1. Data Manual - R01: Food and Brain Study
2. Introduction
    1. Overview of Study
    2. Inclusion/Exclusion Criteria
3. Study Design
    1. Visit 1
    2. Visit 2
    3. Visit 3
    4. Visit 4
    5. Visit 5
    6. Visit 6
    7. Visit 7
4. Outcome Measures: Descriptions and Protocols
    1. Laboratory Eating Paradigms
        1. Standard Laboratory Meal
        2. Eating in the Absence of Hunger
        3. Portion Size Meals

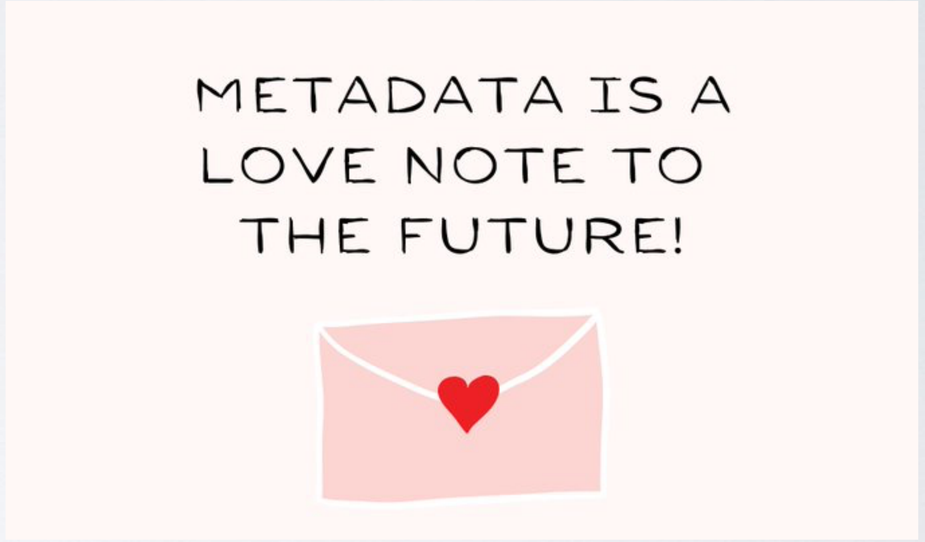


7. Metadata Magic

Metadata: the who, what, when, where, and why of your data

Bigger Lift: Structured Metadata

- Often laid out in fields
- Can require use of shared vocabularies
- Often field/data type specific

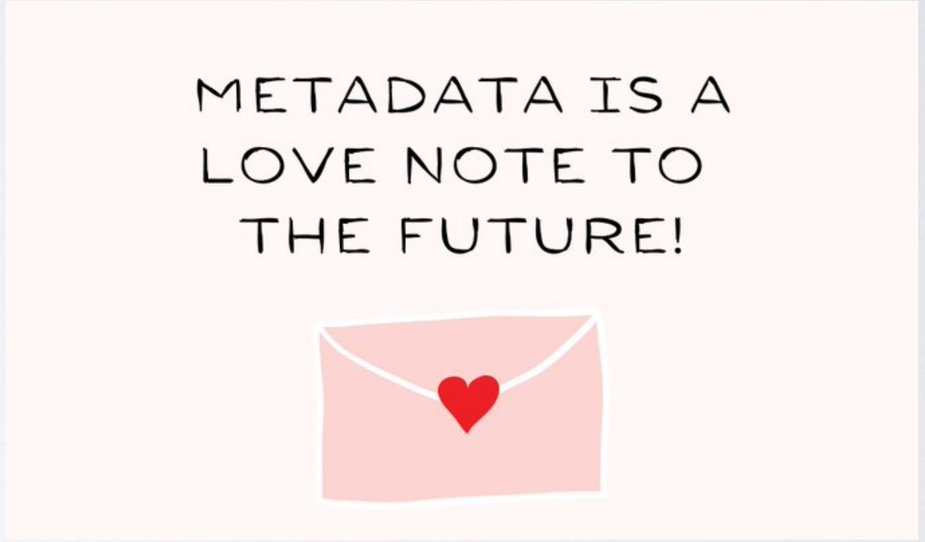


7. Metadata Magic

Metadata: the who, what, when, where, and why of your data

Bigger Lift: Structured Metadata

```
"age": {
    "Description": "age of the participant",
    "Units": "years"
},
"sex": {
    "Description": "sex of the participant as reported by the participant",
    "Levels": {
        "M": "male",
        "F": "female"
    }
},
"handedness": {
    "Description": "handedness of the participant as reported by the
        participant",
    "Levels": {
        "left": "left",
        "right": "right"
    }
},
```



‘Good Enough’ Practices

1. Preserve Raw Data
2. Create a Central Hub
3. Use Meaningful Names
4. Preserve the Journey
5. Avoid Manual Manipulations
6. ‘Tidy’ Your Data
7. Metadata Magic