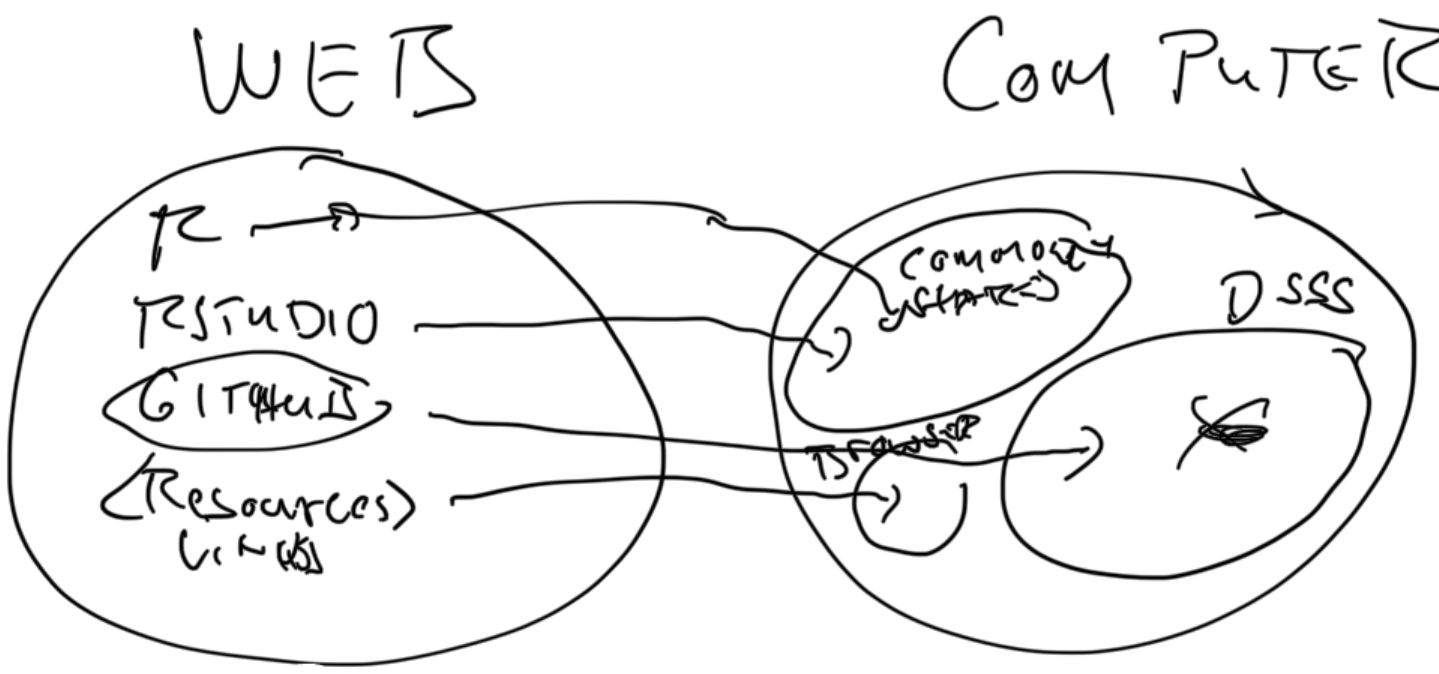
Data science for social sciences



Organization and (File) Names

(1) DSSS.Rproj

DSSS_01.Rmd
data
literature
figures

(2) (File) Names

- machine readable
- human readable
- play well with default ordering

Jenny Bryan
Reproducible
Science Workshop
(see PDF in
literature)

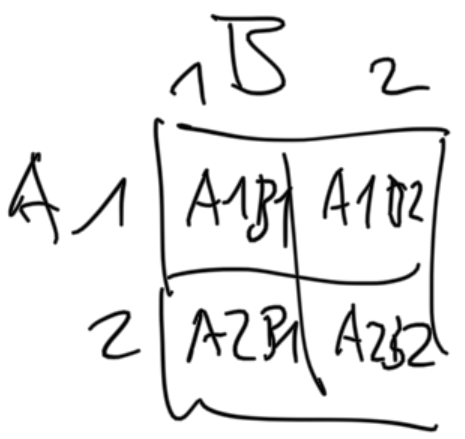
** Machine readable **

- no spaces
- no punctuation (except end)
- no accented or Unicode

- think about case and be consistent

- deliberate use of delimiters

"_" underscore: delimits unit of metadata that are used later



Subj A1_B1 A1_B2 A2_B1 A2_B2

S01 3.4 2.8 Y.8 √.2

S01_K2 Sq1_qA S02_IC4

"-" hyphen delimits words

2022-02-14

J. Bryan: " so my eyes don't bleed "

* human readable

- name contains info about the content

NOT: mydata.csv

data.csv

newdata.csv

graphic.jpg

Poschen.xlsx

Kliegl_01.xlsx

** plays well with default order**

- put numbers first

- use ISO 8601 standard for date

  2022-02-14 not: 14.02.2022

- not: S1, S2, ... S10 ... S100
  But: S001, S002, ... S010 ... S100

! WIDE format:

One row of data for the primary meaningful unit of analysis

- Subjects
- Country
- Family

RANDOM FACTOR

Child_2020.xlsx

Child	School	Birthdate	Testdate	T1...
C0001	S001	2013-03	2020-10-05	

C00001 S001

C20000 S520

School Community Pub/Pr
S001 C400

S520

Community

Questionnaire - WIDE FORMAT

Subj	Item_01	Item_02	...	Item 10	...	Item 33
S01						
S02						
⋮						
S60						

... WITH STRUCTURE

	Biography			Selfrating			Opinion		
Subj	I01	...	I08	I09	...	I18	I19	...	I33
S01									
S02									
⋮									
S60									

SAME QUESTIONNAIRE - LONG FORMAT

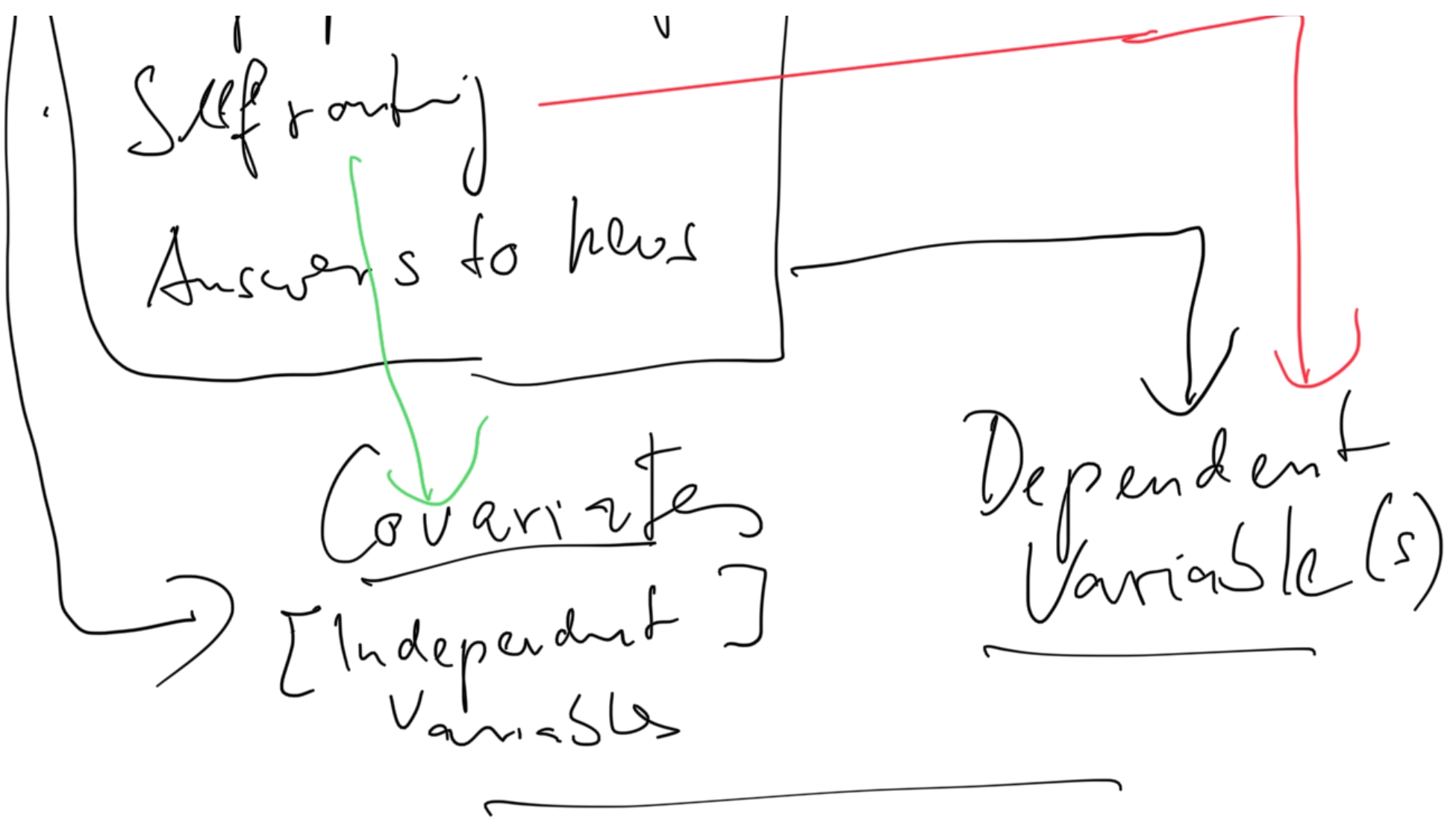
QUESTIONNAIRE

(33 rows / subj)

Subj	Topic	Item	Response
S01	Bio	I01	male ← Gender
S01	Bio	I02	20 ← Age
S01	Bio	I08	2 ← Educ.
S01	SR	I09	3 ← life satisfaction (1-5)
S01	SR	I18	4 ← Health (1-5)
S01	Opn	I19	1 ← Creativity (1-6)
S01	Opn	I33	6
S02	Bio	I01	female ← Gender
S02	Bio	I02	23 ← Age
S02	Opn	I33	5
S03	Bio	I01	male
S60	Opn	I33	4

NOTE: This is usually not the optimal long format! Can you see the problem?

↑ Biographical info



"UPDATED"

- cleaned up data
- Minimal information, nothing that can be computed from the minimal information
- Variable names
- Format store : csv, excel, spss parquet
- Archived for eternity
- unplausible (1) obvious coding error; correct value can be inferred → CORRECT
  (2) obvious measurement, be

error; correct value cannot be
→ NA (missing) inferred

(3) text ones values
→ leave as is

⇓

PREPROCESSING
(Project_preproc.Rmd)

- Read data
- ReadMe file ← describe variables
- Select variables ← usually a subset
- Rename variables ← see below
- Filter values ← remove outliers
- Mutate variables
  - transform variables
  - generate new variables
  - ...
- Relocate variables in colums
- Arrange rows

⇓

CHECKS

- Distributions of variables ?

⇒ boxcox( )

- Table of demographic info
- Table of means, smooths
- Roots

⇓

MULTIVARIATE STATISTICS

- Contrast specification
- Select model supported by data — not overparameterized

Iteration

- Check model residuals →

Iterations must NOT be informed by theoretically relevant statistics (e.g. p-values)

- Fit model

⇓

DOCUMENT RESULTS

- Tables of coefficients
- Visualize effects
- Move into manuscript or supplement

Next: Variable Names

Next: ~~Factors~~ / Covariates

- Random vs fixed
- Between vs. within
- Experimental vs. quasi-experimental