Data as objects & architectures

Readings for today

- Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1-23.
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., ... & Handwerker, D. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific data, 3(1), 1-9.

Supplemental reading: Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".

Topics

1. Data tables

2. Tidy data

3. Standardized data architectures

Data tables

What is data?

Data: Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored & used in a computer. - Cambridge Dictionary

Data Types:

1. **Quantitative** - a direct and continuous mapping between variation in an observable phenomenon and numeric value.

2. **Qualitative** - a discrete mapping between an static environmental state or category and an alphanumerical symbol.

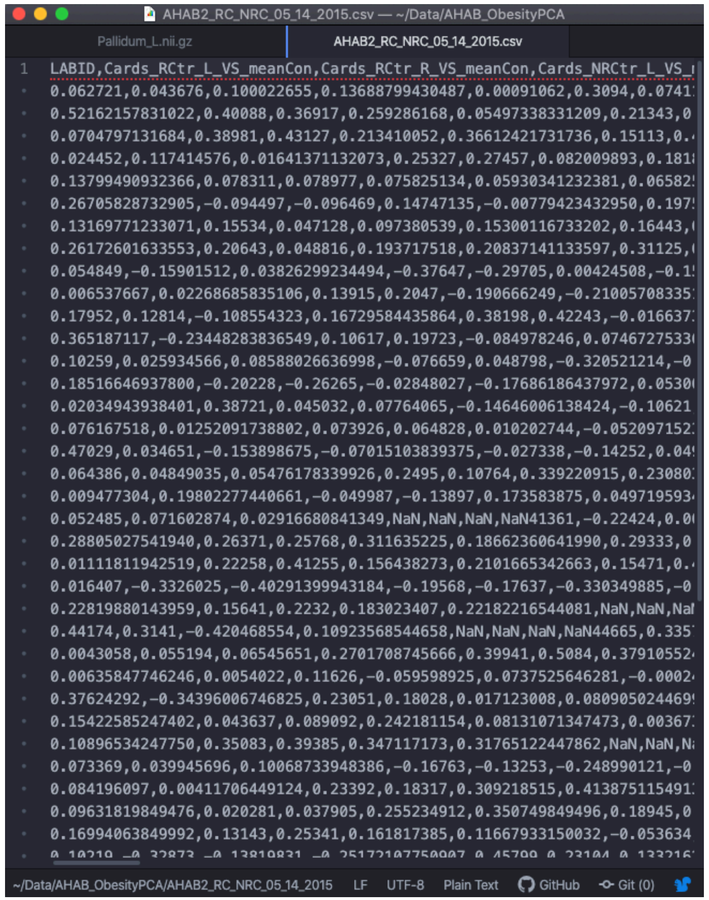
How is data stored?

Data files: How your data are stored in a physical medium.

1. **Human readable** - files that can be read in text editors by humans (e.g., .csv, .txt)

2. **Binary** - files that are compressed for storage or security (e.g., .mat, .R, .sav)

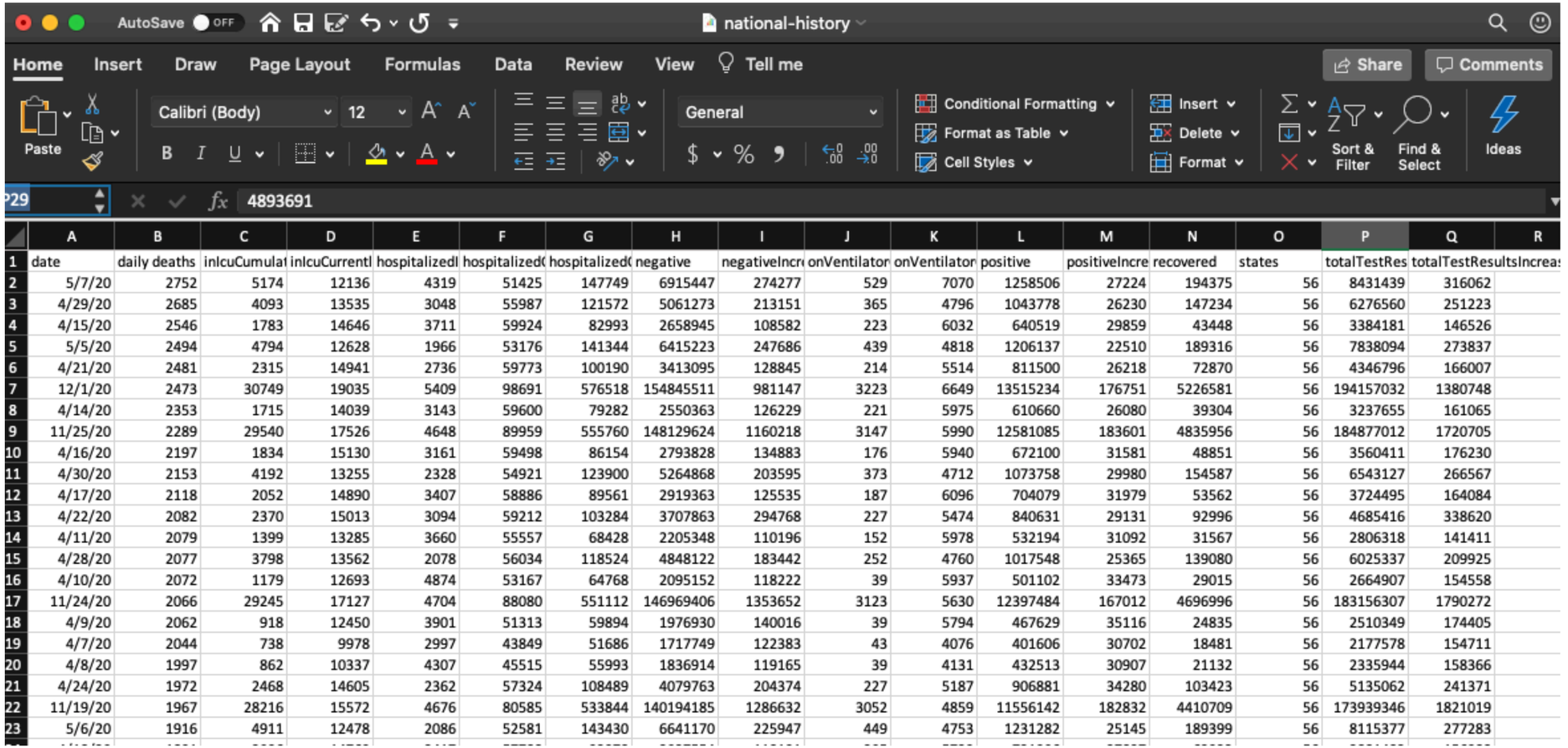
Human readable (.csv)



Binary (.mat)

How is data organized?

Data Tables: Data that is loaded into memory and organized in a way that allows for it to be analyzed.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	date	daily deaths	inIcuCumulat	inIcuCurrentl	hospitalizedI	hospitalizedC	hospitalizedC	negative	negativeIncr	onVentilator	onVentilator	positive	positiveIncre	recovered	states	totalTestRes	totalTestResultsIncreas	
2	5/7/20	2752	5174	12136	4319	51425	147749	6915447	274277	529	7070	1258506	27224	194375	56	8431439	316062	
3	4/29/20	2685	4093	13535	3048	55987	121572	5061273	213151	365	4796	1043778	26230	147234	56	6276560	251223	
4	4/15/20	2546	1783	14646	3711	59924	82993	2658945	108582	223	6032	640519	29859	43448	56	3384181	146526	
5	5/5/20	2494	4794	12628	1966	53176	141344	6415223	247686	439	4818	1206137	22510	189316	56	7838094	273837	
6	4/21/20	2481	2315	14941	2736	59773	100190	3413095	128845	214	5514	811500	26218	72870	56	4346796	166007	
7	12/1/20	2473	30749	19035	5409	98691	576518	154845511	981147	3223	6649	13515234	176751	5226581	56	194157032	1380748	
8	4/14/20	2353	1715	14039	3143	59600	79282	2550363	126229	221	5975	610660	26080	39304	56	3237655	161065	
9	11/25/20	2289	29540	17526	4648	89959	555760	148129624	1160218	3147	5990	12581085	183601	4835956	56	184877012	1720705	
10	4/16/20	2197	1834	15130	3161	59498	86154	2793828	134883	176	5940	672100	31581	48851	56	3560411	176230	
11	4/30/20	2153	4192	13255	2328	54921	123900	5264868	203595	373	4712	1073758	29980	154587	56	6543127	266567	
12	4/17/20	2118	2052	14890	3407	58886	89561	2919363	125535	187	6096	704079	31979	53562	56	3724495	164084	
13	4/22/20	2082	2370	15013	3094	59212	103284	3707863	294768	227	5474	840631	29131	92996	56	4685416	338620	
14	4/11/20	2079	1399	13285	3660	55557	68428	2205348	110196	152	5978	532194	31092	31567	56	2806318	141411	
15	4/28/20	2077	3798	13562	2078	56034	118524	4848122	183442	252	4760	1017548	25365	139080	56	6025337	209925	
16	4/10/20	2072	1179	12693	4874	53167	64768	2095152	118222	39	5937	501102	33473	29015	56	2664907	154558	
17	11/24/20	2066	29245	17127	4704	88080	551112	146969406	1353652	3123	5630	12397484	167012	4696996	56	183156307	1790272	
18	4/9/20	2062	918	12450	3901	51313	59894	1976930	140016	39	5794	467629	35116	24835	56	2510349	174405	
19	4/7/20	2044	738	9978	2997	43849	51686	1717749	122383	43	4076	401606	30702	18481	56	2177578	154711	
20	4/8/20	1997	862	10337	4307	45515	55993	1836914	119165	39	4131	432513	30907	21132	56	2335944	158366	
21	4/24/20	1972	2468	14605	2362	57324	108489	4079763	204374	227	5187	906881	34280	103423	56	5135062	241371	
22	11/19/20	1967	28216	15572	4676	80585	533844	140194185	1286632	3052	4859	11556142	182832	4410709	56	173939346	1821019	
23	5/6/20	1916	4911	12478	2086	52581	143430	6641170	225947	449	4753	1231282	25145	189399	56	8115377	277283	

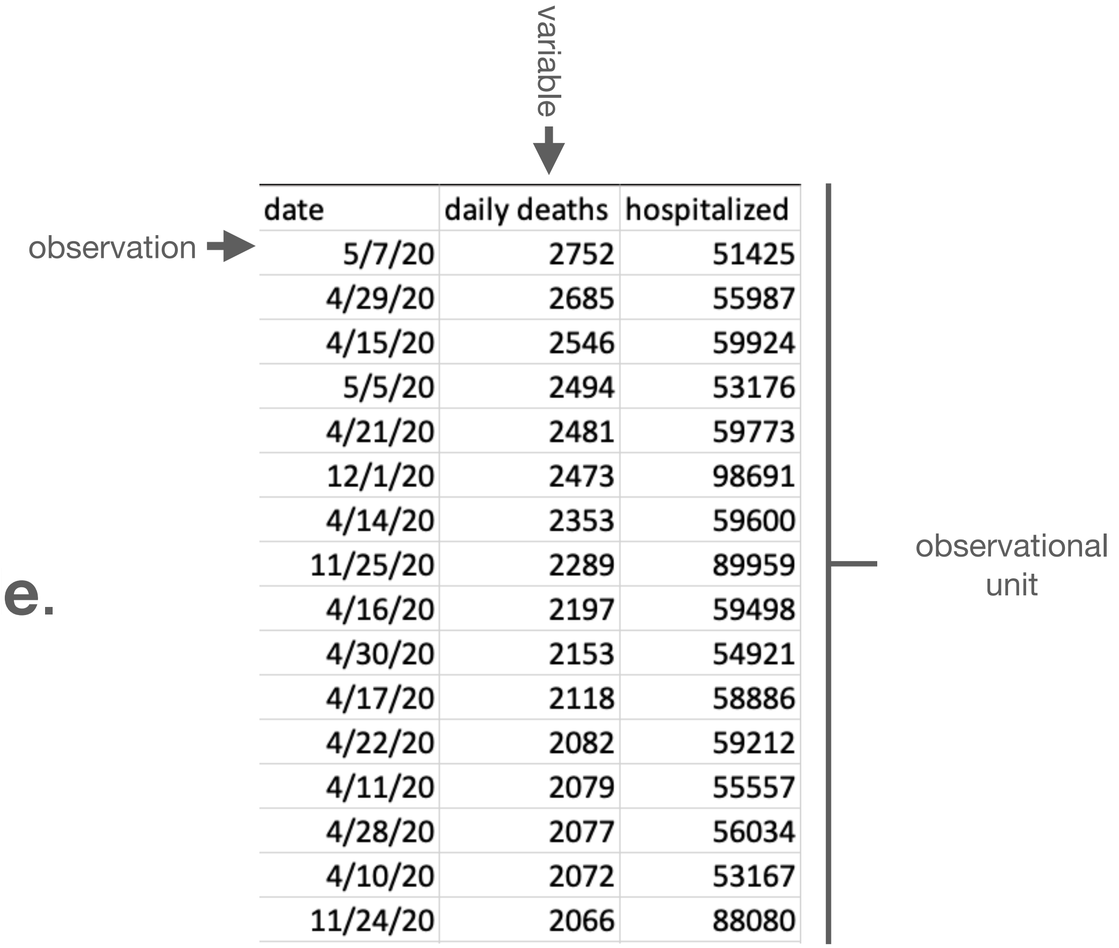
Tidy data

Tidy data

Definition: A standard way of mapping the meaning of a dataset to its structure.

Properties:

1. Each **variable** forms a **column**.
2. Each **observation** forms a **row**.
3. Each **observational** unit forms a **table**.



date	daily deaths	hospitalized
5/7/20	2752	51425
4/29/20	2685	55987
4/15/20	2546	59924
5/5/20	2494	53176
4/21/20	2481	59773
12/1/20	2473	98691
4/14/20	2353	59600
11/25/20	2289	89959
4/16/20	2197	59498
4/30/20	2153	54921
4/17/20	2118	58886
4/22/20	2082	59212
4/11/20	2079	55557
4/28/20	2077	56034
4/10/20	2072	53167
11/24/20	2066	88080

(Wikham et al. 2014)

Key terms

Dataset: collection of values.

Value: analytical unit (number/string).

Variable: values that measure the same attribute across units.

Observation: all values measured on the same unit.

Table: collection of variables and observations organized as a two-dimensional array.

(Wikham et al. 2014)

Dirty vs. tidy tables

Dirty

Person	Treatment	Result
Joe	a	-
June	a	16
Mary	a	3
Joe	b	2
June	b	11
Mary	b	1

Tidy

	a	b
Joe	-	2
June	16	11
Mary	3	1

1. Each **variable** forms a **column**.
2. Each **observation** forms a **row**.
3. Each **observational** unit forms a **table**.

(Wikham et al. 2014)

Tidying approaches

Problem	Solution
1. Column headers are values, not variable names.	Melting

(Wikham et al. 2014)

Melting

Dirty

row	a	b	c
X	1	4	7
Y	2	5	8
Z	3	6	9

Tidy

label	condition	Value
X	a	1
Y	a	2
Z	a	3
X	b	4
Y	b	5
Z	b	6
X	c	7
Y	c	8
Z	c	9

Melting: Unify data across columns that are subordinate to a common variable.

(Wikham et al. 2014)

Tidying approaches

Problem	Solution
1. Column headers are values, not variable names.	Melting
2. Multiple variables are stored in one column.	Splitting

(Wikham et al. 2014)

Splitting

Dirty

Country	Group	Cases
US	M014	0
US	M1524	0
US	F014	1
US	F1524	0
UK	M014	2
UK	M1524	1
UK	F014	0
UK	F1524	3

Tidy

Country	Gender	Age	Cases
US	M	0-14	0
US	M	15-24	0
US	F	0-14	1
US	F	15-24	0
UK	M	0-14	2
UK	M	15-24	1
UK	F	0-14	0
UK	F	15-24	3

Splitting: Separating one column with multiple variables in to multiple columns with one variable.

(Wikham et al. 2014)

Tidying approaches

Problem	Solution
1. Column headers are values, not variable names.	Melting
2. Multiple variables are stored in one column.	Splitting
3. Variables are stored in both rows and columns.	Casting

(Wikham et al. 2014)

Casting

Dirty

Date	Measure	Value
1/20	mean	23
1/20	variance	10
2/20	mean	35
2/20	variance	7
3/20	mean	29
3/20	variance	15

Tidy

Date	Mean	Variance
1/20	23	10
2/20	35	7
3/20	29	15

Casting: Values in a single column reflecting multiple types of variables are rotated into separate columns.

(Wikham et al. 2014)

Tidying approaches

Problem	Solution
1. Column headers are values, not variable names.	Melting
2. Multiple variables are stored in one column.	Splitting
3. Variables are stored in both rows and columns.	Casting
4. Multiple types of observations are stored in the same table.	Parsing

(Wikham et al. 2014)

Parsing

Raw

DOI	Author	Title	Year	Citations
0.001	Verstynen	“Data rules”	2017	2
0.001	Verstynen	“Data rules”	2018	10
0.001	Verstynen	“Data rules”	2019	50
0.001	Verstynen	“Data rules”	2020	101
0.002	Holt	“Theory rules!”	2017	10
0.002	Holt	“Theory rules!”	2018	211
0.002	Holt	“Theory rules!”	2019	561
0.002	Holt	“Theory rules!”	2020	1014

Types of observations:

1. Paper identity
2. Citations across time

(Wikham et al. 2014)

Parsing

Article Identity (Table 1)

DOI	Author	Title
0.001	Verstynen	“Data rules”
0.002	Holt	“Theory rules!”

Citations (Table 2)

DOI	Year	Citations
0.001	2017	2
0.001	2018	10
0.001	2019	50
0.001	2020	101
0.002	2017	10
0.002	2018	211
0.002	2019	561
0.002	2020	1014

Parsing: Take 1 table, with multiple observational units, and separate it into multiple tables with unique observational units.

(Wikham et al. 2014)

Tidying approaches

Problem	Solution
1. Column headers are values, not variable names.	Melting
2. Multiple variables are stored in one column.	Splitting
3. Variables are stored in both rows and columns.	Casting
4. Multiple types of observations are stored in the same table.	Parsing
5. Single observational unit stored in multiple tables.	Concatenation

(Wikham et al. 2014)

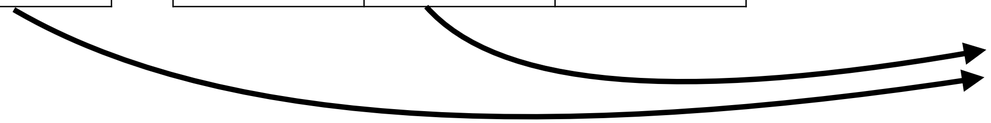
Concatenation

Raw (Subject 1)

Trial	Cond	RT
1	A	380
2	B	599
3	A	240

Raw (Subject 2)

Trial	Cond	RT
1	A	692
2	B	476
3	A	301



Tidy

Subject ID	Trial	Cond	RT
S001	1	A	380
S001	2	B	599
S001	3	A	240
S002	1	A	692
S002	2	B	476
S002	3	A	301

Concatenate: Merge multiple data tables, with the same observational units, into a single table.

(Wikham et al. 2014)

Tidying approaches

Problem	Solution
1. Column headers are values, not variable names.	Melting
2. Multiple variables are stored in one column.	Splitting
3. Variables are stored in both rows and columns.	Casting
4. Multiple types of observations are stored in the same table.	Parsing
5. Single observational unit stored in multiple tables.	Concatenation

(Wikham et al. 2014)

Standardized data architectures

Beyond data tables

Data Standardization: Systematic file formats, directory organization, & naming logic that allows for effective collaboration & sharing.

Advantages:

1. **Minimal curation**: Makes it possible for those not involved in the research to understand the data with minimal instructions.

2. **Error reduction**: Reduces errors attributed to misunderstanding the meaning of a given datum.

3. **Optimal Data Use**: Facilitates aggregation across studies to promote reuse & re-analysis.

4. **Automation**: Fosters the development of tools that work across data sets.

(Gorgolewski et al. 2016)

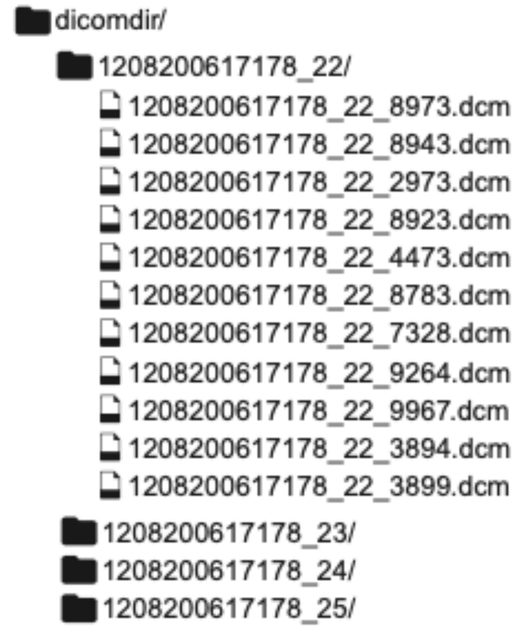
Data architectures

A formal logic of data & directory organization.

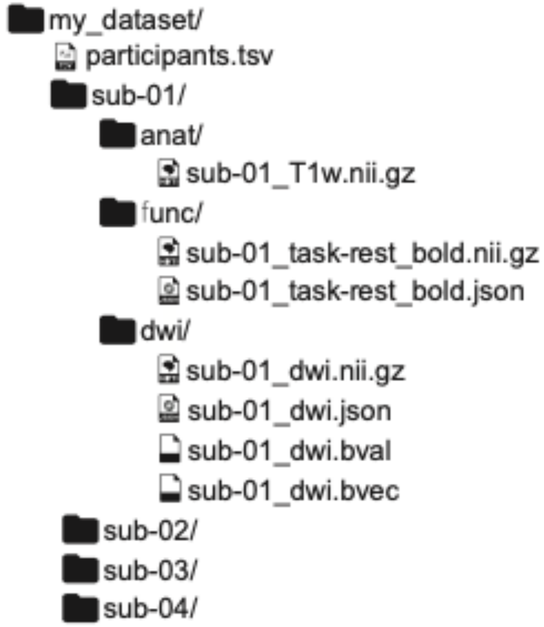
Pay attention to:

- File types
- Naming logic
- Directory hierarchy
- Documentation & meta-data

Raw architecture



BIDS



(Gorgolewski et al. 2016)

Take home message

Having a standard logic for your data tables (e.g., tidy) and files (e.g., BIDS) reduces many sources of errors and facilitates data & code sharing.