Ch. 8 Intro to Tidy Data

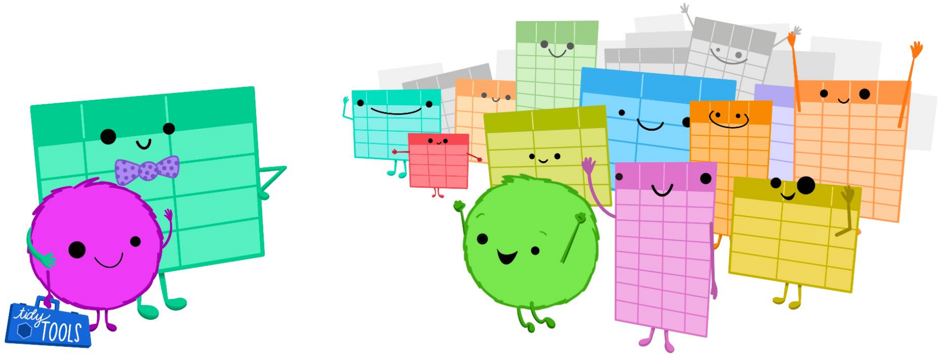
NCEAS Learning Hub
Arctic Data Center

January 2024

Illustrations from the [Openscapes]() blog [Tidy Data for reproducibility, efficiency, and collaboration]() by Julia Lowndes and Allison Horst

Learning Objectives

- Understand basics of relational data models, aka **tidy data**
- Learn how to design and create effective data tables



artwork by @allison_horst

Introduction

book sections here, in case you want to cross reference!

The organizational structure that allows for relating data tables

A type of database that contains data tables that are related to one another

An organized collection of information (i.e. data)

Relational data **models** are what relational **databases** use to organize tables.

However, *you don't need to be using a relational database (e.g mySQL, MariaDB, Oracle, Microsoft Access) to benefit from using a relational data model.*

Benefits:

- Powerful search & filtering
- Handle large, complex datasets
- Enforce data integrity
- Decrease errors from redundant updates

Simple Guidelines for Data Management (Borer et al. 2009)

Use a scripted program

XLSX TXT CSV Nonproprietary file formats

Keep a raw version of the data

```
data
├── raw_data
│   └── my_data.csv
└── processed_data
    └── my_data_clean.csv
```

```
mooredCTD_site1_2020-2023.txt
mooredCTD_site2_2020-2023.txt
mooredCTD_site3_2020-2023.txt
```

Descriptive names

Header line

date	site_name	temp_c
2023-01-01	site1	16.3
2023-02-01	site2	15.9
2023-03-01	site3	16.1

A B C / 1 2 3

Plain ASCII text

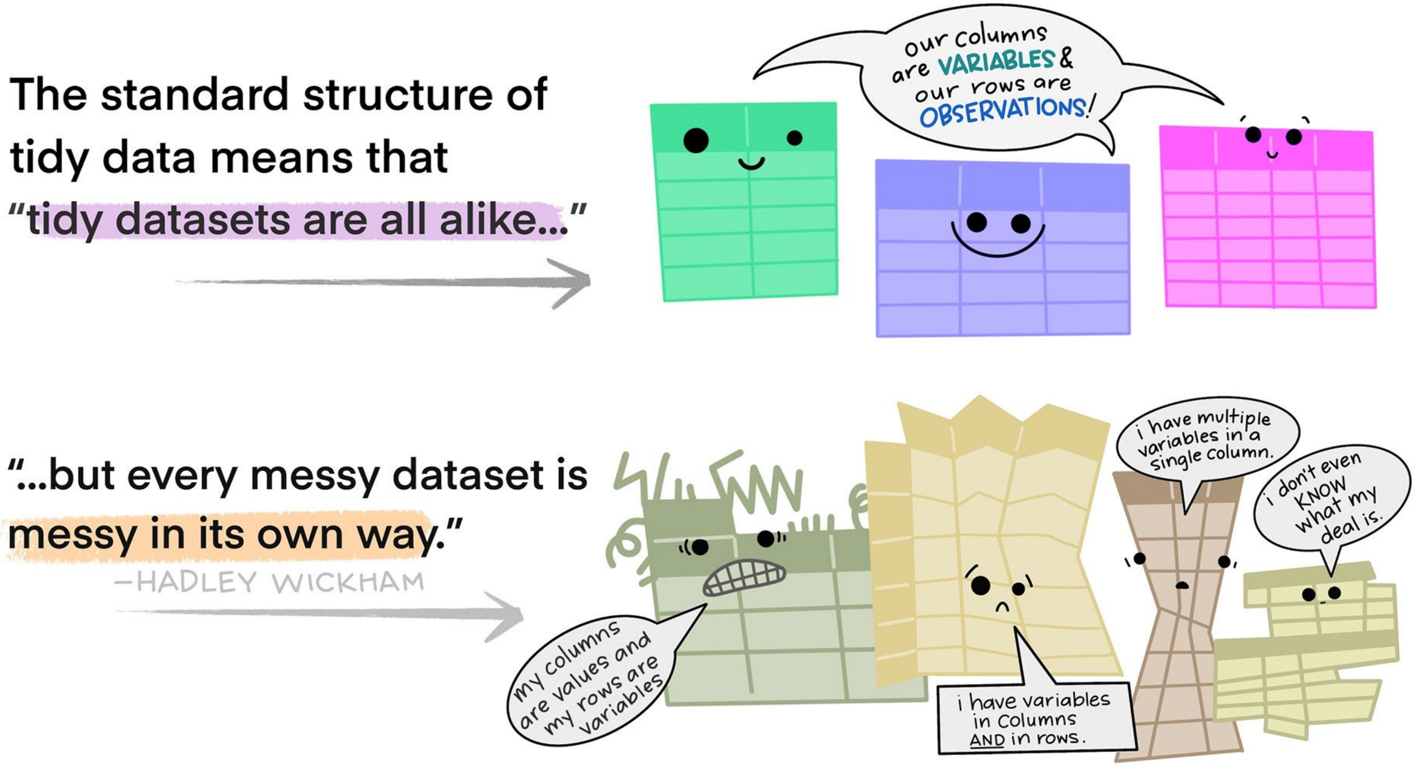
Simple Guidelines for Data Management (Borer et al. 2009)

- Design your tables to add rows, not columns

- Each column should contain only one type of information

- Record a single piece of data only once; separate information collected at different scales into different tables -- in other words, create a *relational database*

Recognizing "untidy" data



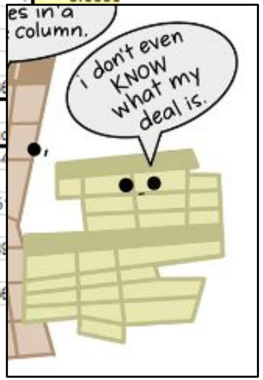
artwork by @allison_horst

Recognizing “untidy” data

A not-so-tidy spreadsheet received by NCEAS....

AtlasGroveCOMPLETE.xls

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		main trunks	reiterated trunks	limbs	branches	leaves						dry masses (kg)				
species	tree	kg	kg	kg	kg	kg		type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2		tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8		tree	PSME	135815	0	0	8338	961	145114	3.3876
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4		tree	THSE	31799	0	0	6343	864	39006	0.9105
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9		tree	ACMA	4444	0	0	925	264	5634	0.1315
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4		tree	UMCA	2921	0	0	937	273	4131	0.0964
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7		shrub	RUSP	0	0	0	1974	686	2660	0.0620
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4		fern	POMU	0	0	0	0	1271	1271	0.0296
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8		shrub	VAOV	0	0	0	526	26	552	0.0129
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3		shrub	COCO	0	0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2		fern	POSC	0	0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7		tree	RHPU	100	0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0		herb	OXOR	0	0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	5524.2	691.2		shrub	VAPA	0	0	0	94	4	99	0.0023
SESE	Zeus	243365.7	2885.5	1620.4	19104.7	954.3		tree	PISI	0	0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	87.6	41.4		tree	CHLA	0	0	0	1	0		
SESE	4	6312.0	356.0	73.5	214.1	43.8		shrub	GASH	0	0	0	0	0		
SESE	5	206.0	0.0	0.0	8.7	2.5		shrub	SACA	0	0	0	0	0		
SESE	6E	18697.4	0.0	0.0	1055.2	66.3				3744390	213247	53714	250519	21767	4283	
SESE	6W	14651.5	7.7	0.0	626.3	49.6										
SESE	11	614.4	0.0	0.0	28.1	17.0				main trunk	reiteration	limb	branch	leaf	to	
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	40844	
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0		
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145	
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0		
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	389	
SESE	23	463.7	0.0	0.0	18.9	4.5			TSHE epi	59	0	0	12	4		
SESE	25	87.7	0.0	0.0	4.1	1.3			ACMA geo	4444	0	0	925	264	56	
SESE	30	512.1	1.8	0.0	18.7	8.7			ACMA epi	0	0	0	0	0		



Recognizing "untidy" data - multiple tables

Easy for humans to interpret (sort of?), hard for computer programs (e.g. R)

AtlasGroveCOMPLETE.xls

Table 1

species	tree	main trunks kg	reiterated trunks kg	limbs kg	branches kg	leaves kg
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0
SESE	Rhea	[illegible]	[illegible]	730.1	5524.2	691.2
SESE	Zeus	[illegible]	[illegible]	1620.4	19104.7	954.3
SESE	3	1761.3	0.0	0.0	87.6	41.4
SESE	4	6312.0	356.0	73.5	214.1	43.8
SESE	5	206.0	0.0	0.0	8.7	2.5
SESE	6E	18697.4	0.0	0.0	1055.2	66.3
SESE	6W	14651.5	7.7	0.0	626.3	49.6
SESE	11	614.4	0.0	0.0	28.1	17.0
SESE	12	232.1	0.0	0.0	11.2	10.3
SESE	18	15632.0	0.0	0.0	946.3	106.8
SESE	19	11805.5	0.0	0.0	770.1	80.3
SESE	20	309.5	0.0	0.0	12.5	5.9
SESE	22	25618.3	0.0	0.0	1504.0	120.2
SESE	23	463.7	0.0	0.0	18.9	4.5
SESE	25	87.7	0.0	0.0	4.1	1.3
SESE	30	512.1	1.8	0.0	18.7	8.7

Table 2

		dry masses (kg)						
type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
tree	PSME	135815	0	0	8338	961	145114	3.3876
tree	THSE	31799	0	0	6343	864	39006	0.9105
tree	ACMA	4444	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	4131	0.0964
shrub	RUSP	0	0	0	1974	686	2660	0.0620
fern	POMU	0	0	0	0	1271	1271	0.0296
shrub	VAOV	0	[illegible]	[illegible]	[illegible]	26	552	0.0129
shrub	COCO	0	[illegible]	[illegible]	[illegible]	6	289	0.0067
fern	POSC	0	0	0	107	89	196	0.0045
tree	RHPU	100	0	0	44	18	162	0.0037
herb	OXOR	0	0	0	0	112	112	0.0026
shrub	VAPA	0	0	0	94	4	99	0.0023
tree	PISI	0	0	0	1	0	1	0.0000
tree	CHLA	0	0	0	1	0	1	0.0000
shrub	GASH	0	0	0	0	0	0	0.0000
shrub	SACA	0	0	0	0	0	0	0.0000
		3744390	213247	53714	250519	21767	4283636	

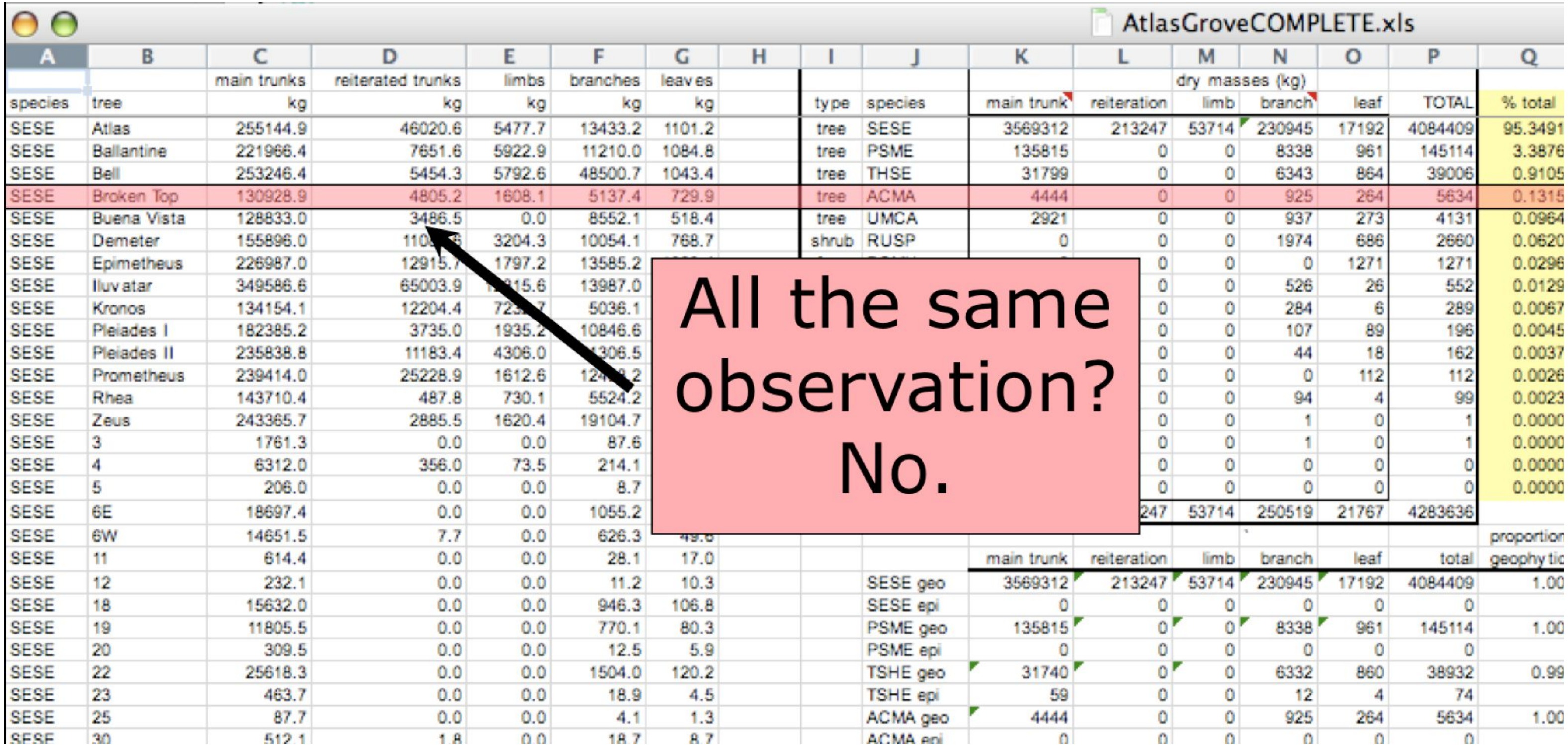
Table 3

	main trunk	reiteration	limb	branch	leaf	total	proportion geophytic
SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE epi	0	0	0	0	0	0	
PSME geo	135815	0	0	8338	961	145114	1.00
PSME epi	0	[illegible]	[illegible]	0	0	0	
TSHE geo	31740	[illegible]	[illegible]	[illegible]	860	38932	0.99
TSHE epi	59	0	0	12	4	74	
ACMA geo	4444	0	0	925	264	5634	1.00
ACMA epi	0	0	0	0	0	0	

INSTEAD: create separate tables/files for each entity measured

Recognizing “untidy” data - inconsistent observations

Each row corresponds to more than one observation



species	tree	main trunks kg	reiterated trunks kg	limbs kg	branches kg	leaves kg
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4
SESE	Demeter	155896.0	[illegible]	3204.3	10054.1	768.7
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	[illegible]
SESE	Iluvatar	349586.6	65003.9	[illegible]	13987.0	[illegible]
SESE	Kronos	134154.1	12204.4	[illegible]	5036.1	[illegible]
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	[illegible]
SESE	Pleiades II	235838.8	11183.4	4306.0	[illegible]	[illegible]
SESE	Prometheus	239414.0	25228.9	1612.6	[illegible]	[illegible]
SESE	Rhea	143710.4	487.8	730.1	5524.2	[illegible]
SESE	Zeus	243365.7	2885.5	1620.4	19104.7	[illegible]
SESE	3	1761.3	0.0	0.0	87.6	[illegible]
SESE	4	6312.0	356.0	73.5	214.1	[illegible]
SESE	5	206.0	0.0	0.0	8.7	[illegible]
SESE	6E	18697.4	0.0	0.0	1055.2	[illegible]
SESE	6W	14651.5	7.7	0.0	626.3	[illegible]
SESE	11	614.4	0.0	0.0	28.1	17.0
SESE	12	232.1	0.0	0.0	11.2	10.3
SESE	18	15632.0	0.0	0.0	946.3	106.8
SESE	19	11805.5	0.0	0.0	770.1	80.3
SESE	20	309.5	0.0	0.0	12.5	5.9
SESE	22	25618.3	0.0	0.0	1504.0	120.2
SESE	23	463.7	0.0	0.0	18.9	4.5
SESE	25	87.7	0.0	0.0	4.1	1.3
SESE	30	512.1	1.8	0.0	18.7	8.7

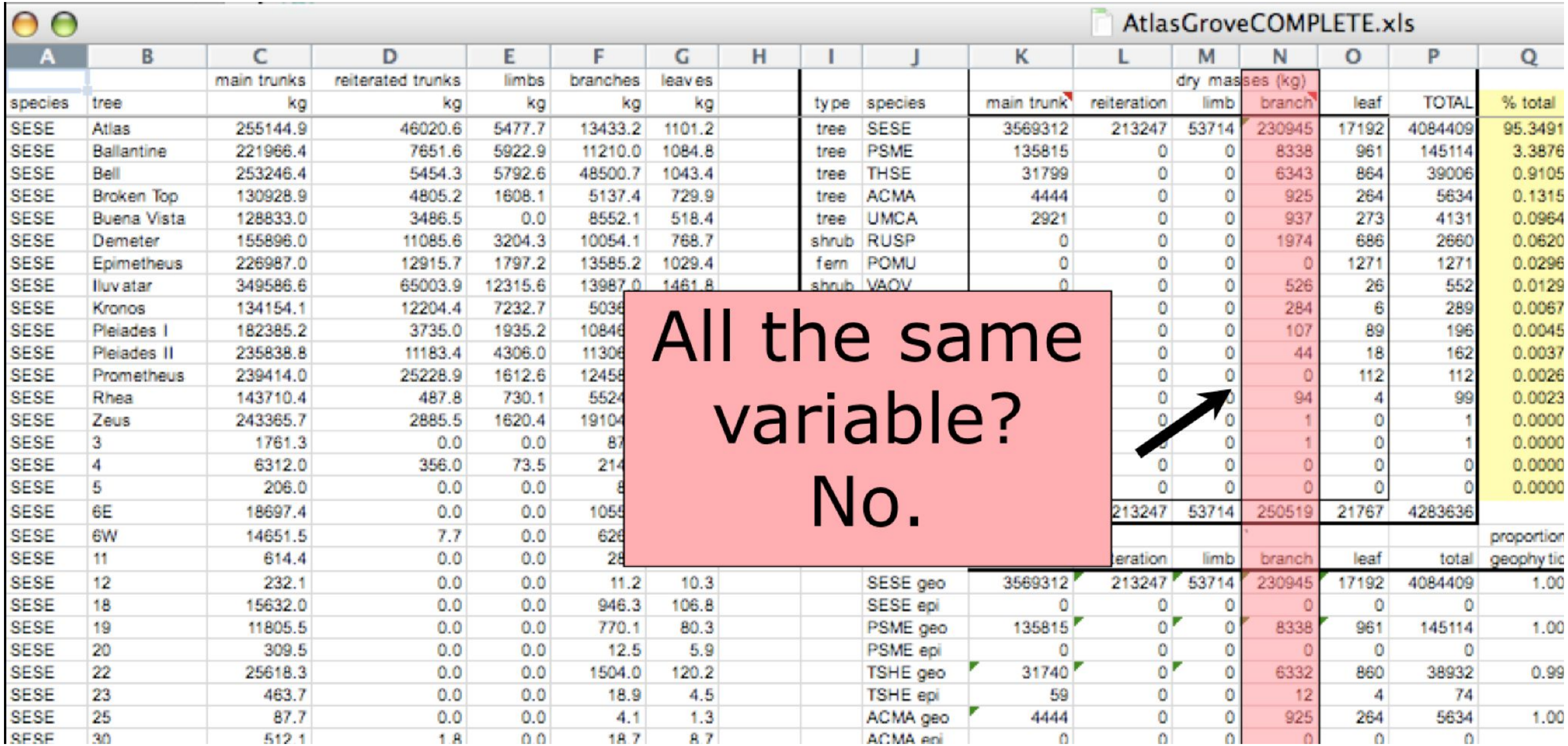
type	species	main trunk	reiteration	limb (dry masses (kg))	branch	leaf	TOTAL	% total
tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
tree	PSME	135815	0	0	8338	961	145114	3.3876
tree	THSE	31799	0	0	6343	864	39006	0.9105
tree	ACMA	4444	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	4131	0.0964
shrub	RUSP	0	0	0	1974	686	2660	0.0620
[illegible]	[illegible]	[illegible]	0	0	0	1271	1271	0.0296
[illegible]	[illegible]	[illegible]	0	0	526	26	552	0.0129
[illegible]	[illegible]	[illegible]	0	0	284	6	289	0.0067
[illegible]	[illegible]	[illegible]	0	0	107	89	196	0.0045
[illegible]	[illegible]	[illegible]	0	0	44	18	162	0.0037
[illegible]	[illegible]	[illegible]	0	0	0	112	112	0.0026
[illegible]	[illegible]	[illegible]	0	0	94	4	99	0.0023
[illegible]	[illegible]	[illegible]	0	0	1	0	1	0.0000
[illegible]	[illegible]	[illegible]	0	0	1	0	1	0.0000
[illegible]	[illegible]	[illegible]	0	0	0	0	0	0.0000
[illegible]	[illegible]	[illegible]	0	0	0	0	0	0.0000
		[illegible]	[illegible]247	53714	250519	21767	4283636	

	main trunk	reiteration	limb	branch	leaf	total	proportion geophytic
SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE epi	0	0	0	0	0	0	
PSME geo	135815	0	0	8338	961	145114	1.00
PSME epi	0	0	0	0	0	0	
TSHE geo	31740	0	0	6332	860	38932	0.99
TSHE epi	59	0	0	12	4	74	
ACMA geo	4444	0	0	925	264	5634	1.00
ACMA epi	0	0	0	0	0	0	

INSTEAD: each row should represent a single observed entity

Recognizing “untidy” data - inconsistent variables

Each column contains more than one variable type



INSTEAD: all values in a column should be of the same type (tip: compare units)

Recognizing “untidy” data - marginal sums & stats

Marginal sums & statistics are combinations of observations

AtlasGroveCOMPLETE.xls

A	B	C	D	E	F	G
		main trunks	reiterated trunks	limbs	branches	leaves
species	tree	kg	kg	kg	kg	kg
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0
SESE	Rhea	143710.4	487.8	730.1	5524.2	691.2
SESE	Zeus	243365.7	2885.5	1620.4	19104.7	954.3
SESE	3	1761.3	0.0	0.0	87.6	41.4
SESE	4	6312.0	356.0	73.5	214.1	43.8
SESE	5	206.0	0.0	0.0	8.7	2.5
SESE	6E	18697.4	0.0	0.0	1055.2	66.3
SESE	6W	14651.5	7.7	0.0	626.3	49.6
SESE	11	614.4	0.0	0.0	28.1	17.0
SESE	12	232.1	0.0	0.0	11.2	10.3
SESE	18	15632.0				
SESE	19	11805.5				
SESE	20	309.5				
SESE	22	25618.3				
SESE	23	463.7				
SESE	25	87.7				
SESE	30	512.1				

I	J	K	L	M	N	O	P	Q
				dry masses (kg)				
type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
tree	PSME	135815	0	0	8338	961	145114	3.3876
tree	THSE	31799	0	0	6343	864	39006	0.9105
tree	ACMA	4444	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	4131	0.0964
shrub	RUSP	0	0	0	1974	686	2660	0.0620
fern	POMU	0	0	0	0	1271	1271	0.0296
shrub	VAOV	0	0	0	526	26	552	0.0129
shrub	COCO	0	0	0	284	6	289	0.0067
fern	POSC	0	0	0	107	89	196	0.0045
tree	RHPU	100	0	0	44	18	162	0.0037
herb	OXOR	0	0	0	0	112	112	0.0026
shrub	VAPA	0	0	0	94	4	99	0.0023
tree	PISI	0	0	0	1	0	1	0.0000
tree	CHLA	0	0	0	1	0	1	0.0000
shrub	GASH	0	0	0	0	0	0	0.0000
shrub	SACA	0	0	0	0	0	0	0.0000
		3744390	213247	53714	250519	21767	4283636	
								proportion
		main trunk	reiteration	limb	branch	leaf	total	geophytic
	SESE	3569312	213247	53714	230945	17192	4084409	1.00
	SE epi	0	0	0	0	0	0	
	ME geo	135815	0	0	8338	961	145114	1.00
	ME epi	0	0	0	0	0	0	
	HE geo	31740	0	0	6332	860	38932	0.99
	HE epi	59	0	0	12	4	74	
	MA geo	4444	0	0	925	264	5634	1.00
	MA epi	0	0	0	0	0	0	

Marginal sums

INSTEAD: only identifying or measured variables should exist here; use a scripted language to analyze data / calculate summary stats

Denormalized (untidy) data

Data are **denormalized** when observations about different entities are combined. For example, each row in the data table below has site characteristics & species observations:

id	date	site	name	temp	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	Taku	23.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	Lituya	19.9	DAMA	3.5	DAPU	3.9

site characteristics (site, name, temp)

species observations (sp1code, sp1height, sp2code, sp2height)

Importantly, a new species observation would require us to add columns (not a row) -- this data table organization is also known as **wide format**

Normalizing (tidying) this data table

To normalize this data table, we want to organize observations about each type of entity in it's own table

id	date	site	name	temp	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	Taku	23.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	Lituya	19.9	DAMA	3.5	DAPU	3.9

Observed entities:

- site characteristics
- plant species

Variables associated with those observations:

- temperature
- height

Normalized (tidy) data

denormalized / untidy / wide format

id	date	site	name	temp	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	Taku	23.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	Lituya	19.9	DAMA	3.5	DAPU	3.9

normalized / tidy / long format

date species height

plants

id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

sites

site	name	temp
1	Taku	23.7
2	Lituya	19.9

name temperature

We now have:

- Separate tables for each type of entity
- Each row represents a single observed entity
- Observations (rows) are all unique

Additionally:

- All values in a column are of the same type
- All columns pertain to the same observed entity
- Each column represents either an identifying variable or a measured variable (no summary stats)

Normalized (tidy) data

Our normalized data now meet the guidelines set by Borer et al. 2009:

- Tables are designed to **add rows,** not columns
- Each **column** contains only **one type of information**
- A single piece of **data is recorded only once** & separated information collected at **different scales** into **different tables**

date species height

plants

id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

sites

site	name	temp
1	Taku	23.7
2	Lituya	19.9

name temperature

Normalized (tidy) data has lots of benefits!

denormalized / untidy / wide format

id	date	site	name	temp	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	Taku	23.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	Lituya	19.9	DAMA	3.5	DAPU	3.9

normalized / tidy / long format

More easily filter rows for observations of interest

```
dplyr::filter(data = plant_data, spcode == "DAPU")
```

Describe columns more precisely

`spcode` is the spp. identifier, but what exactly is `sp1code`, `sp2code`?

Optimize storage

not repeating data (e.g. date) reduces file size

Decrease errors from redundant updates

e.g. only need to update site name in table 2

plants

date species height

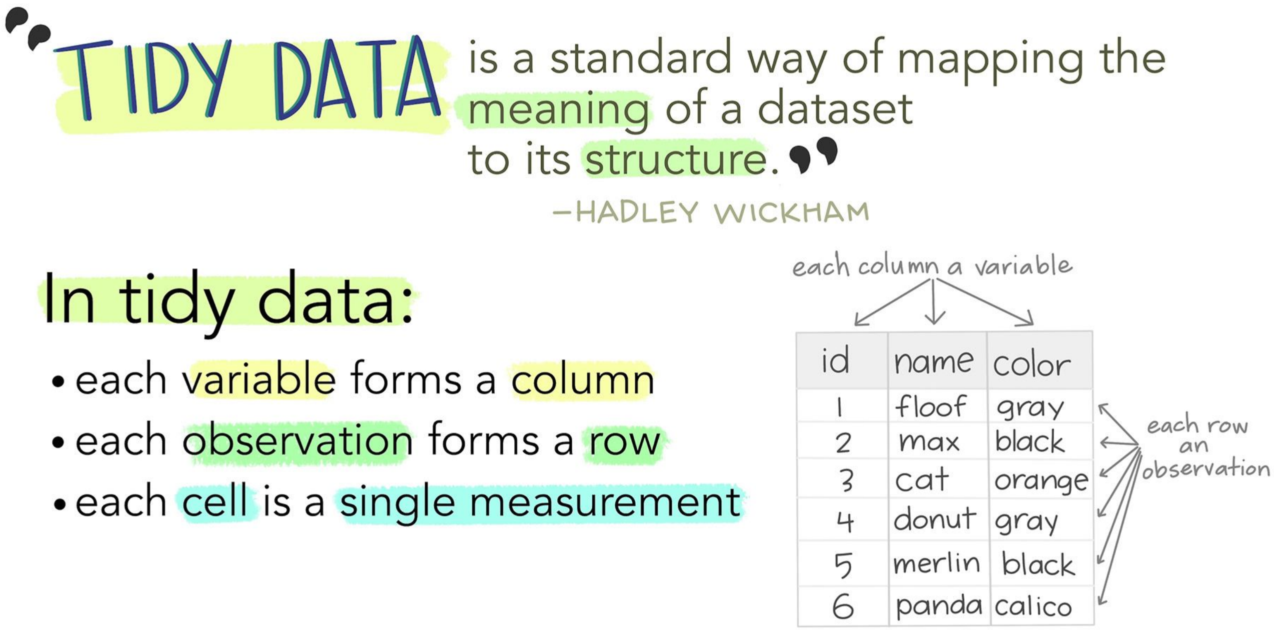
id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

sites

site	name	temp
1	Taku	23.7
2	Lituya	19.9

name temperature

One more look at tidy data



Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

artwork by [@allison_horst]()

Using normalized data

Two tables?!? Don't we want to analyze all these different measurements together??

(e.g. how will we use site temperature as a predictor variable for species composition?)

Keys!

date species height

plants

id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

sites

site	name	temp
1	Taku	23.7
2	Lituya	19.9

name temperature

Keys allow us to link observations across tables

`id` uniquely identifies each row in the *plant* table

`site` references the primary key in the *site* table -- this is our linkage

id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

Primary Key: a unique identifier for each observed entity, one per row

Foreign Key: reference to a primary key in another table (linkage)

site	name	temp
1	Taku	23.7
2	Lituya	19.9

`site` uniquely identifies each row in the *site* table primary key

foreign key

entity: plants

id	date	site	sp_code	sp_height
1	2017-10-10	1	DAPU	4.6
2	2017-10-10	1	DAMA	4.5
3	2017-09-05	2	DAMA	3.5
4	2017-09-05	2	DAPU	3.9

entity: sites

site	name	altitude
1	Taku	944
2	Lituya	525

primary key

surrogate key

natural key

entity: sites

site	name	altitude
1	Taku	944
2	Lituya	525

entity: plants

compound key

id	date	site	sp_code	sp_height
1	2017-10-10	1	DAPU	4.6
2	2017-10-10	1	DAMA	4.5
3	2017-09-05	2	DAMA	3.5
4	2017-09-05	2	DAPU	3.9

Keys allow us to link observations across tables

Joined the tables by **site**

↓

id	date	site	spcode	height	name	temp
1	2017-10-10	1	DAPU	4.6	Taku	23.7
2	2017-09-05	2	DAMA	3.5	Lituya	19.9
3	2017-10-10	1	DAMA	4.5	Taku	23.7
4	2017-09-05	2	DAPU	3.9	Lituya	19.9

Merging data (aka “joins”)

Merging (or joining) two related data tables based on key values is something you’ll probably do often during the data preparation (pre-analysis & visualization) stage. We’ll use these two tables to showcase how different types of joins work:

X

1	x1
2	x2
3	x3

Y

1	y1
2	y2
4	y3

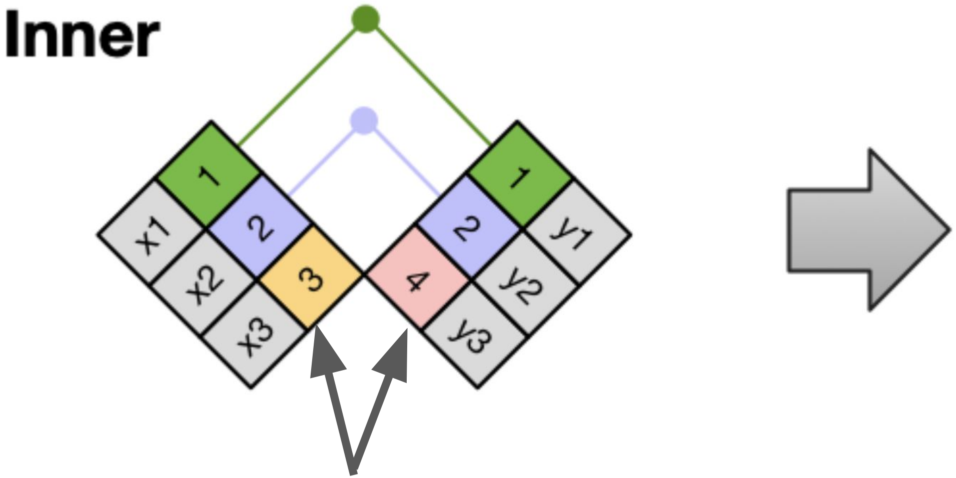
Inner join

X

1	x1
2	x2
3	x3

Y

1	y1
2	y2
4	y3

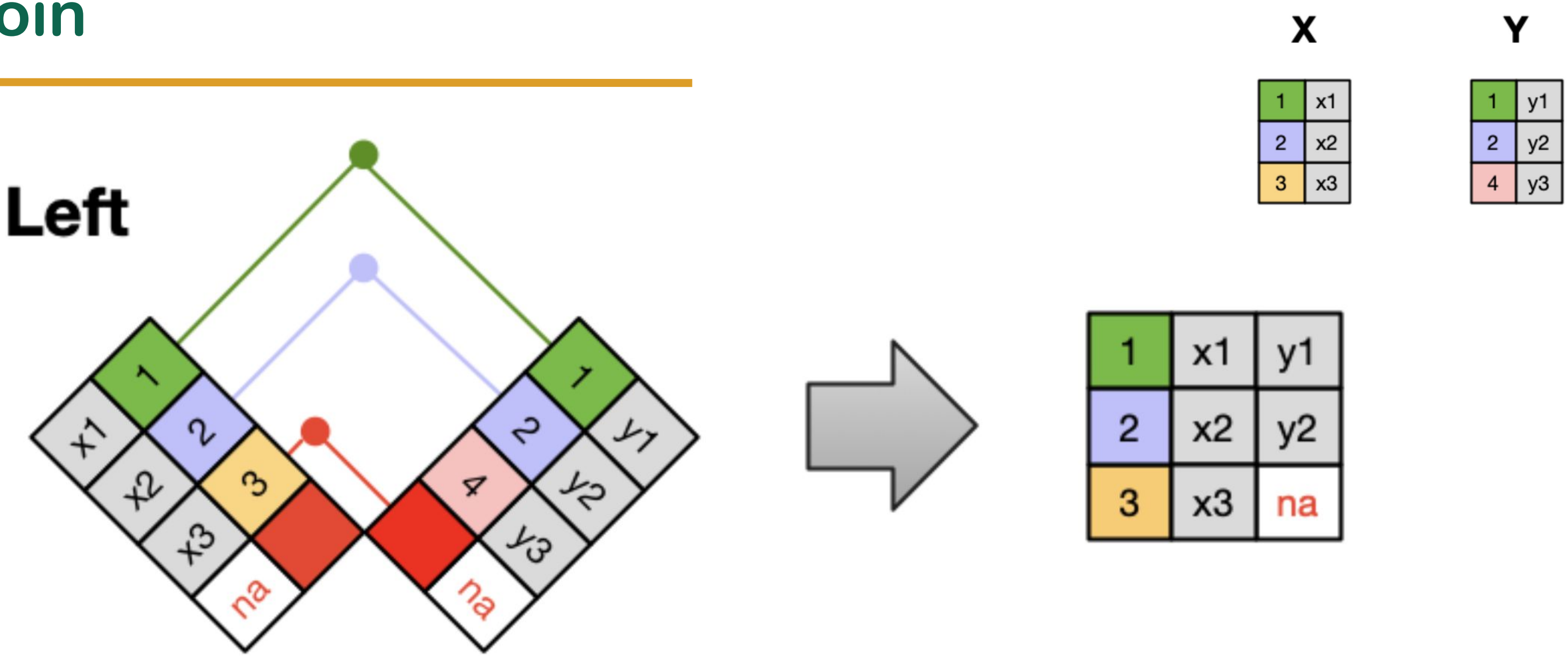


1	x1	y1
2	x2	y2

rows 3 (from left table) & 4 (from right table) are dropped because they have no matches

Merge (i.e. keep) the subset of rows that have matches in both the left and right tables

Left join



rows 1 & 2 (left table) have matches in the right table and are kept;
row 3 (left table) does not have a match in the right table, so it is kept and assigned an NA value

Take all rows from **left** table and merge on data from matching rows in right table

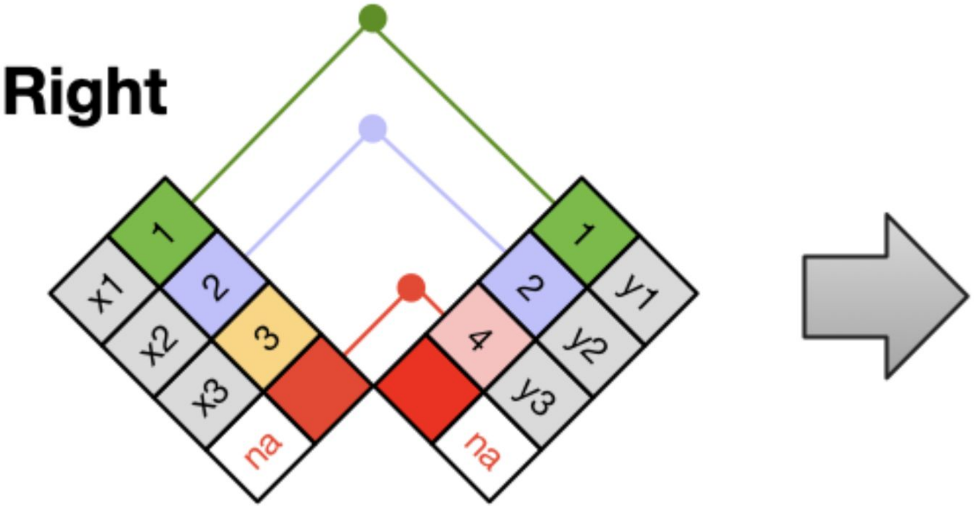
Right join

X

1	x1
2	x2
3	x3

Y

1	y1
2	y2
4	y3

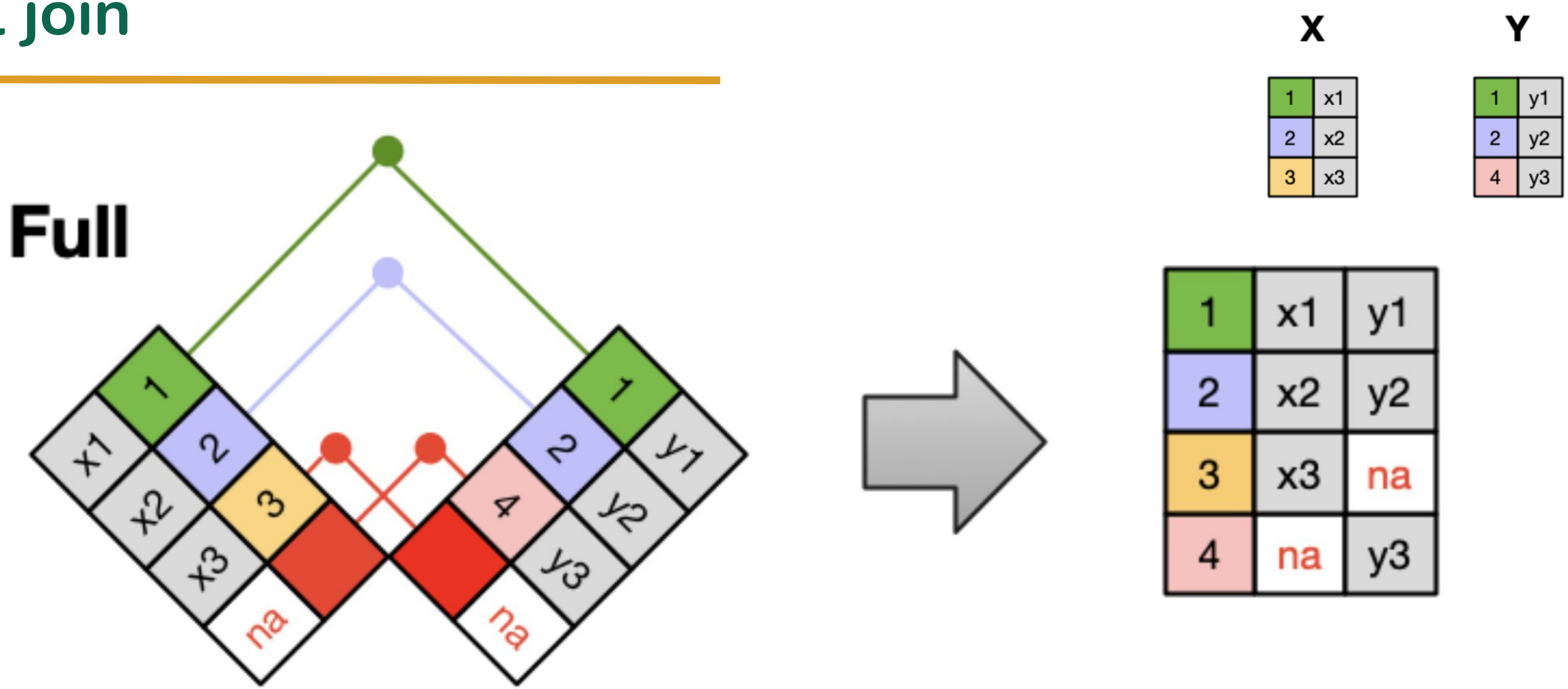


1	x1	y1
2	x2	y2
4	na	y3

rows 1 & 2 (right table) have matches in the left table and are kept;
row 4 (right table) does not have a match in the left table, so it is kept and assigned an NA value

Take all rows from **right** table and merge on data from matching rows in left table

Full join



rows 1 & 2 are matched;
row 3 (left table) and row 4 (right table) are kept despite not having matches (assigned the value, NA)

Includes all rows from both tables and adds missing values (NAs) where necessary

Spoiler: {dplyr} has super helpful functions for joining data

```
inner_join(x, y)
```

```
left_join(x, y)
```

```
right_join(x, y)
```

```
full_join(x, y)
```