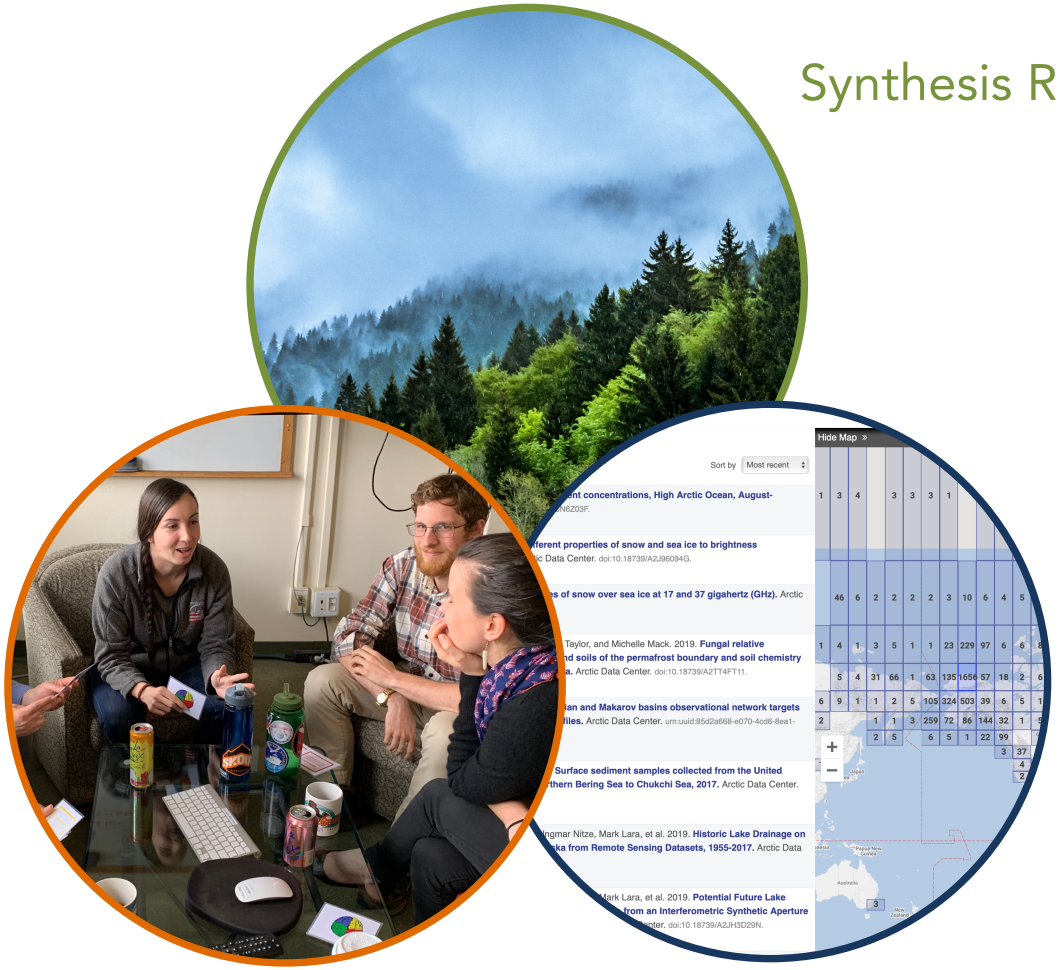
Tidy-ing Your Data: Simple Steps for Reproducible Research

Jeanette Clark
National Center for Ecological Analysis and Synthesis
jclark@nceas.ucsb.edu

NCEAS
National Center for Ecological Analysis and Synthesis

NCEAS

DataONE



Sort by Most recent

ent concentrations, High Arctic Ocean, August-
N6Z03F.

fferent properties of snow and sea ice to brightness
ic Data Center. doi:10.18739/A2J96094G.

es of snow over sea ice at 17 and 37 gigahertz (GHz). Arctic

Taylor, and Michelle Mack. 2019. Fungal relative
nd soils of the permafrost boundary and soil chemistry
a. Arctic Data Center. doi:10.18739/A2TT4FT11.

ian and Makarov basins observational network targets
files. Arctic Data Center. urn:uuid:85d2a668-e070-4cd6-8ea1-

Surface sediment samples collected from the United
rthern Bering Sea to Chukchi Sea, 2017. Arctic Data Center.

Ingmar Nitze, Mark Lara, et al. 2019. Historic Lake Drainage on
ska from Remote Sensing Datasets, 1955-2017. Arctic Data

Mark Lara, et al. 2019. Potential Future Lake
from an Interferometric Synthetic Aperture
enter. doi:10.18739/A2JH3D29N.

Synthesis Research

Training

Data Science Infrastructure

NCEAS

DataONE

Synthesis Research

nt concentrations, High Arctic Ocean, August-

ferent properties of snow and sea ice to brightness
ic Data Center. doi:10.18739/A2J96094Q

s of snow over sea ice at 17 and 37 gigahertz (GHz). Arctic

Taylor, and Michelle Mack. 2019. Fungal relative
nd soils of the permafrost boundary and soil chemistry
s. Arctic Data Center. doi:10.18739/A2TT4FT1

an and Makarov basins observational network targets
files. Arctic Data Center.

Surface sediment samples collected from the United
thern Bering Sea to Chukchi Sea, 2017. Arctic Data Center.

ngmar Nitze, Mark Lara, et al. 2019. Historic Lake Drainage on
ka from Remote Sensing Datasets, 1955-2017. Arctic Data

Mark Lara, et al. 2019. Potential Future Lake
from an Interferometric Synthetic Aperture
enter. doi:10.18739/A2JH3D29N

Training

Data Science Infrastructure

NCEAS
National Center for Ecological Analysis and Synthesis

NCEAS Learning Hub

National Center for Ecological Analysis and Synthesis

```
90) -- "Single Candle"
e R Foundation for Statistical Computing
apple-darwin15.6.0 (64-bit)

e software and comes with ABSOLUTELY NO WARRANTY.
are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

otice the default panes:

- Console (entire left)
- Environment/History (tabbed in upper right)
- Files/Plots/Packages/Help (tabbed in lower right)

YI: you can change the default location of the panes, among many other things: Customizing RStudio.

important first question: **where are we?**

tional Science

SCIENCE
LIGO Tutorial
LIGO Detected Gravitational Waves from Black Holes
On September 14, 2015 at 5:51 a.m. Eastern Daylight Time (09:51 UTC), the twin Laser Interferometer Gravitational-

SCIENCE
Predicting the Properties of Inorga...

A knowledge-sharing community where researchers can learn the latest data skills and technologies to increase efficiency, productivity, transparency, and collaborative capacity.

Courses: Fee-based and grant-supported intensive data science workshops

Mentored Programs: Experiential residential and remote learning programs to build skills in data and open science

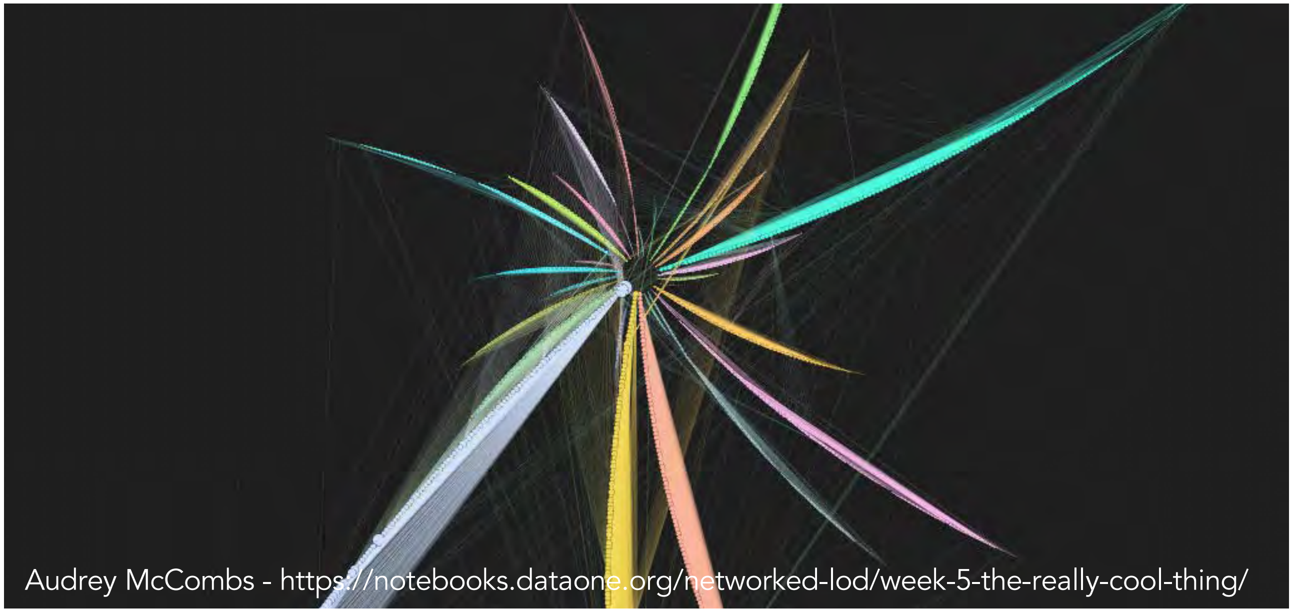
Resources: Extensive online curricula, webinars, training materials and best practices

Partnerships: Customized workshops and collaborative initiatives in data science training

NCEAS
National Center for Ecological Analysis and Synthesis

- Some simple guidelines for effective data management
- How to recognize and tidy untidy data
- Using tidy data in analysis

Data management is for everyone!



Audrey McCombs - https://notebooks.dataone.org/networked-lod/week-5-the-really-cool-thing/

Your data don’t need to be of a particular type, size, or complexity before you start implementing data management practices

NCEAS National Center for Ecological Analysis and Synthesis

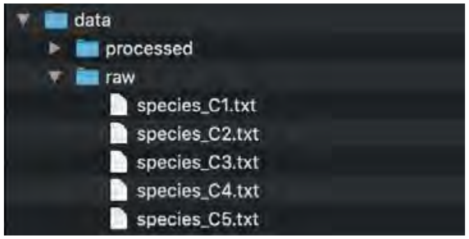
Data management is for everyone!

You don’t have to be using a relational database system to benefit from the concepts of relational data models (aka tidy data)

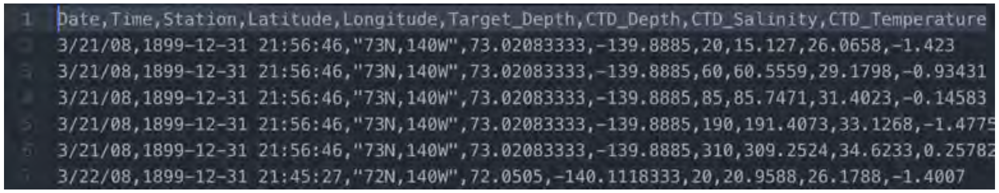
Simple Guidelines for Data Management

- Use a scripted program
- Nonproprietary formats
- Keep a raw version of data
- Descriptive names
- Header line
- Plain ASCII text

.csv, .txt



```
data
  processed
  raw
    species_C1.txt
    species_C2.txt
    species_C3.txt
    species_C4.txt
    species_C5.txt
```



```
Date,Time,Station,Latitude,Longitude,Target_Depth,CTD_Depth,CTD_Salinity,CTD_Temperature
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,20,15.127,26.0658,-1.423
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,60,60.5559,29.1798,-0.93431
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,85,85.7471,31.4023,-0.14583
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,190,191.4073,33.1268,-1.4775
3/21/08,1899-12-31 21:56:46,"73N,140W",73.02083333,-139.8885,310,309.2524,34.6233,0.25782
3/22/08,1899-12-31 21:45:27,"72N,140W",72.0505,-140.1118333,20,20.9588,26.1788,-1.4007
```

Borer, E. T. et al, (2009), Some Simple Guidelines for Effective Data Management. The Bulletin of the Ecological Society of America. doi:10.1890/0012-9623-90.2.205

Simple Guidelines for Data Management

- Design 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, use a relational model

NCEAS
National Center for Ecological Analysis and Synthesis

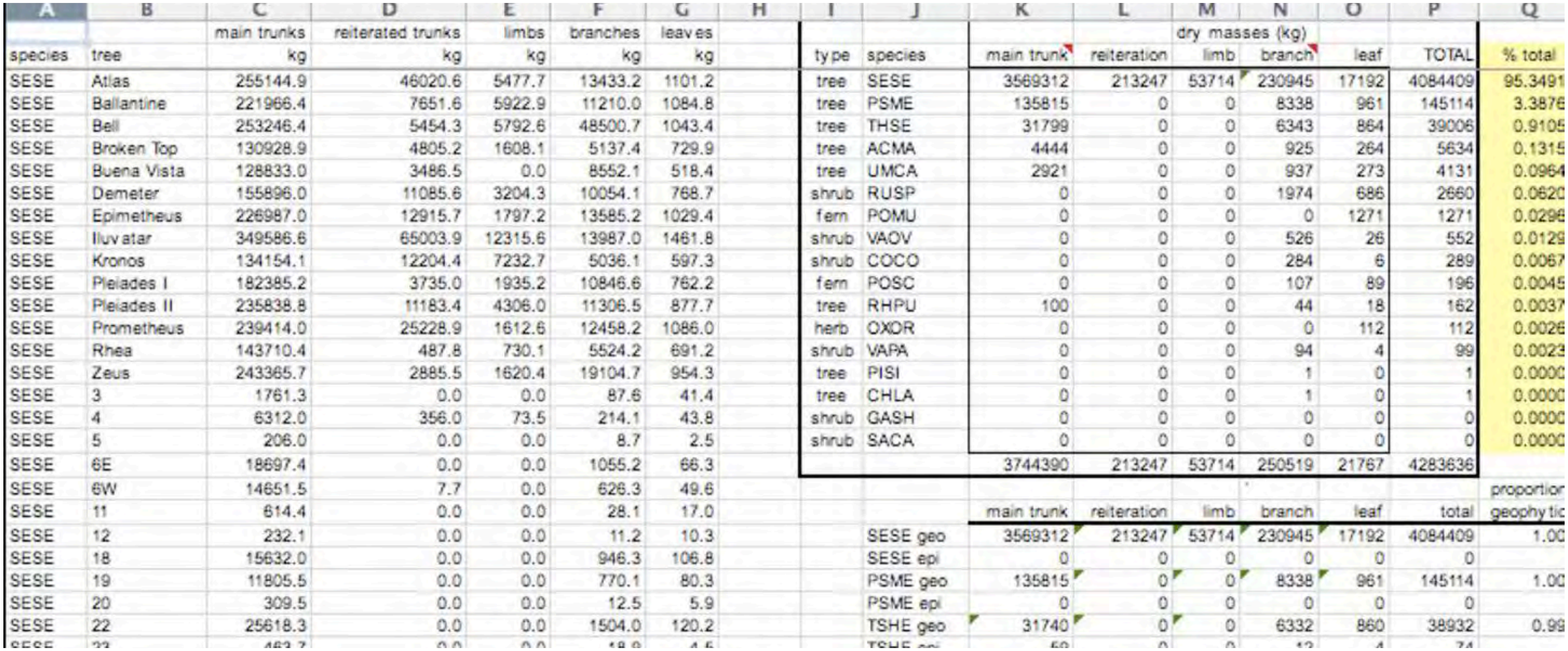
Borer, E. T. et al, (2009), Some Simple Guidelines for Effective Data Management.
The Bulletin of the Ecological Society of America. doi:10.1890/0012-9623-90.2.205

Data model diversity

- There are lots of data models besides tabular data
  - multiband raster
  - matrices
  - spatial vector

NCEAS
National Center for Ecological Analysis and Synthesis

Recognizing untidy data



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		main trunks	reiterated trunks	limbs	branches	leav es						dry masses (kg)				
species	tree	kg	kg	kg	kg	kg		ty pe	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	Iluv atar	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	1	0.0000
SESE	4	6312.0	356.0	73.5	214.1	43.8		shrub	GASH	0	0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0	8.7	2.5		shrub	SACA	0	0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	1055.2	66.3				3744390	213247	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626.3	49.6										proportion
SESE	11	614.4	0.0	0.0	28.1	17.0				main trunk	reiteration	limb	branch	leaf	total	geophy tic
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0	0	
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145114	1.00
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0	0	
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	38932	0.99
SESE	23	463.7	0.0	0.0	18.9	4.5			TSHE epi	59	0	0	12	4	74	

NCEAS
National Center for Ecological Analysis and Synthesis

Characteristics of tidy data

Observations

- Separate tables for each entity measured

NCEAS
National Center for Ecological Analysis and Synthesis

Recognizing untidy data

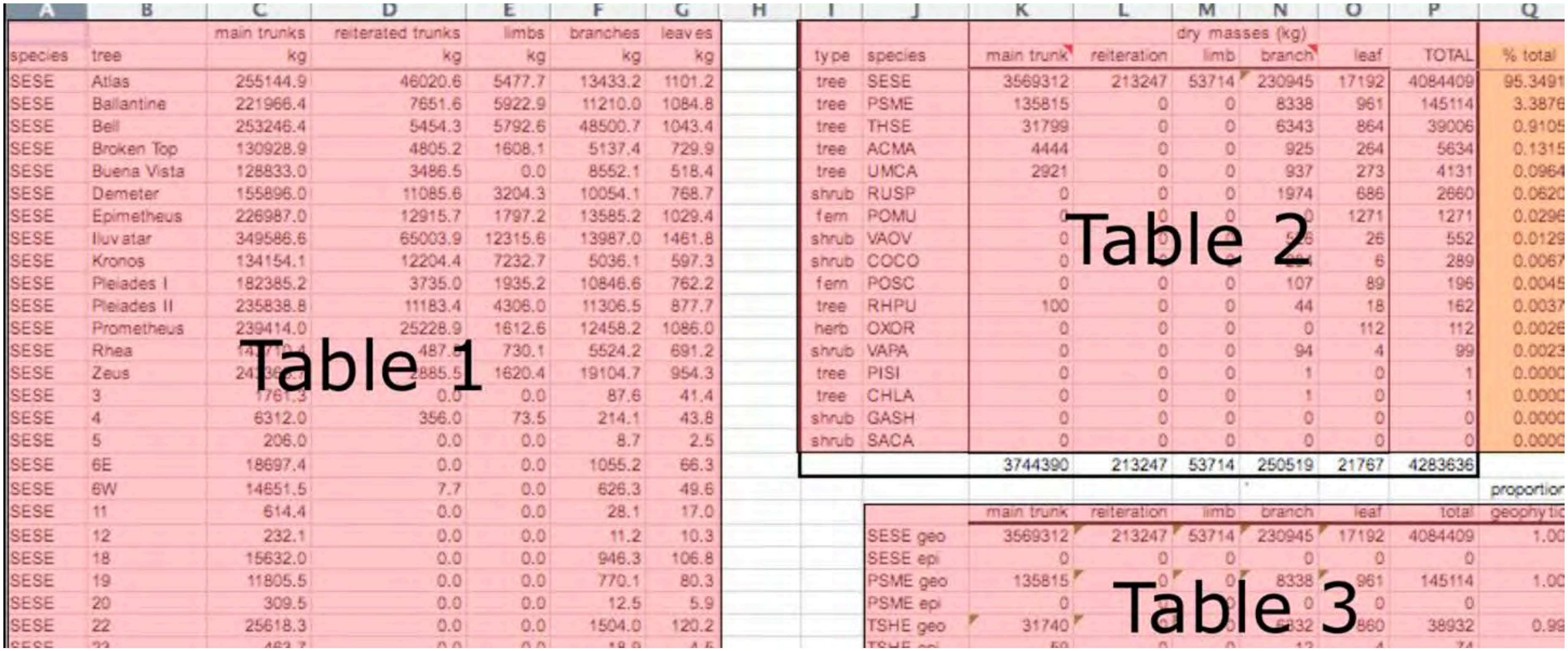


Table 1

		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	[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

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

							proportior
	main trunk	reiteration	limb	branch	leaf	total	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	

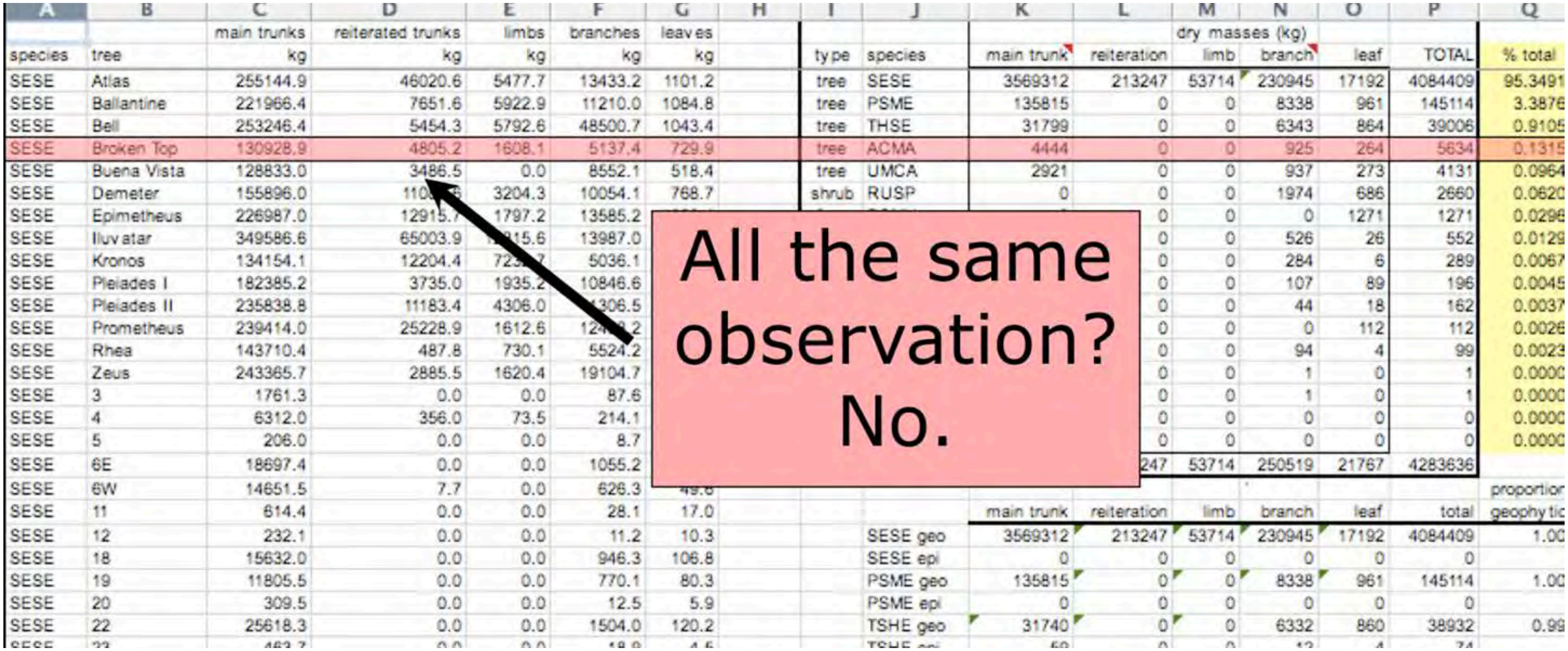
Characteristics of tidy data

Observations

- Separate tables for each entity measured
- Each row represents a single observed entity

NCEAS
National Center for Ecological Analysis and Synthesis

Recognizing untidy data



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		main trunks	reiterated trunks	limbs	branches	leav es						dry masses (kg)				
species	tree	kg	kg	kg	kg	kg		ty pe	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	[illegible]	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						0	0	0	1271	1271	0.0296
SESE	Iluv atar	349586.6	65003.9	[illegible]	13987.0						0	0	526	26	552	0.0129
SESE	Kronos	134154.1	12204.4	[illegible]	5036.1						0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6						0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	[illegible]						0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	[illegible]						0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	5524.2						0	0	94	4	99	0.0023
SESE	Zeus	243365.7	2885.5	1620.4	19104.7						0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	87.6						0	0	1	0	1	0.0000
SESE	4	6312.0	356.0	73.5	214.1						0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0	8.7						0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	1055.2						[illegible]	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626.3	[illegible]										proportior
SESE	11	614.4	0.0	0.0	28.1	17.0				main trunk	reiteration	limb	branch	leaf	total	geophy tic
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0	0	
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145114	1.00
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0	0	
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	38932	0.99

All the same observation?

No.

NCEAS

National Center for Ecological Analysis and Synthesis

Characteristics of tidy data

Observations

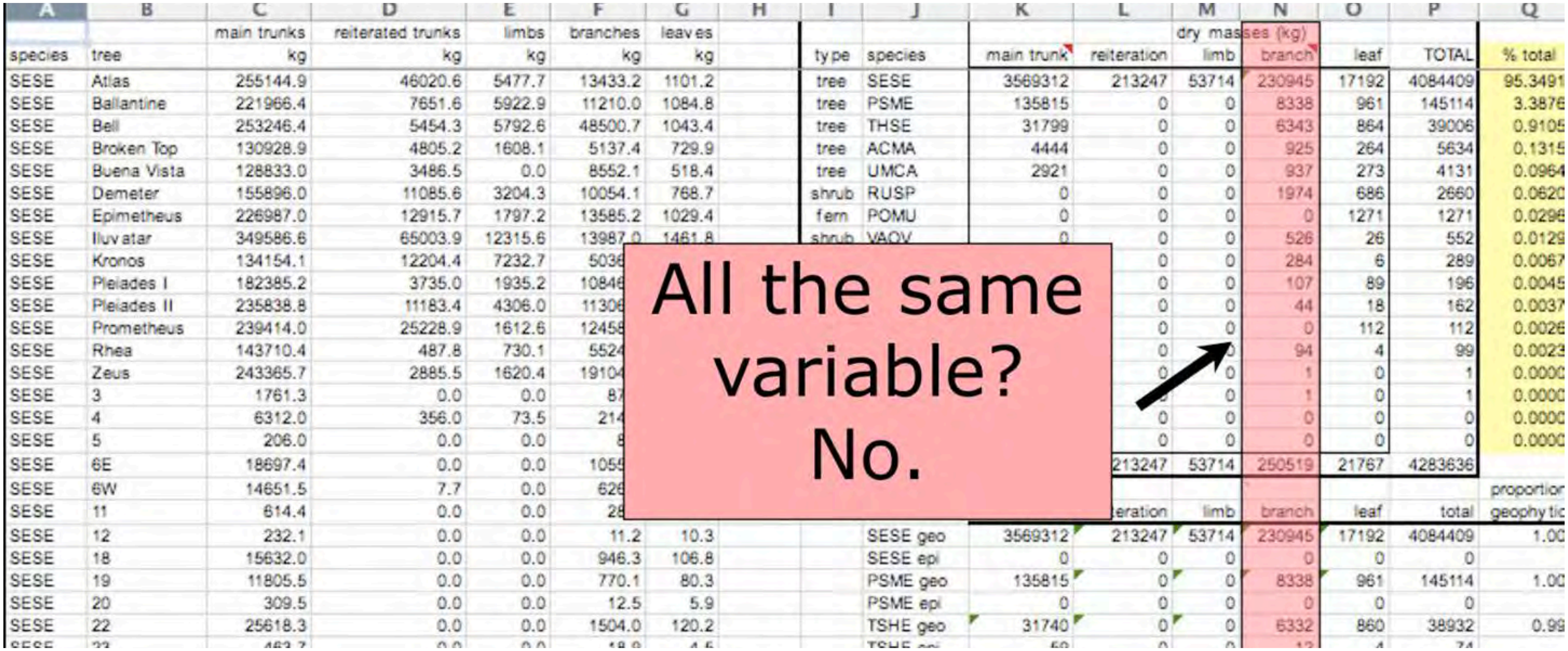
- Separate tables for each entity measured
- Each row represents a single observed entity

Variables

- All values in a column are of the same type

National Center for Ecological Analysis and Synthesis

Recognizing untidy data



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						0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	10846						0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	11306						0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	12458						0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	5524						0	0	94	4	99	0.0023
SESE	Zeus	243365.7	2885.5	1620.4	19104						0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	87						0	0	1	0	1	0.0000
SESE	4	6312.0	356.0	73.5	214						0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0	8						0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	1055						213247	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626											proportio
SESE	11	614.4	0.0	0.0	28						eration	limb	branch	leaf	total	geophytic
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0	0	
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145114	1.00
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0	0	
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	38932	0.99
SESE	23	463.7	0.0	0.0	18.9	4.5			TSHE epi	59	0	0	12	4	74	

All the same variable?

No.

NCEAS
National Center for Ecological Analysis and Synthesis

Characteristics of tidy data

Observations

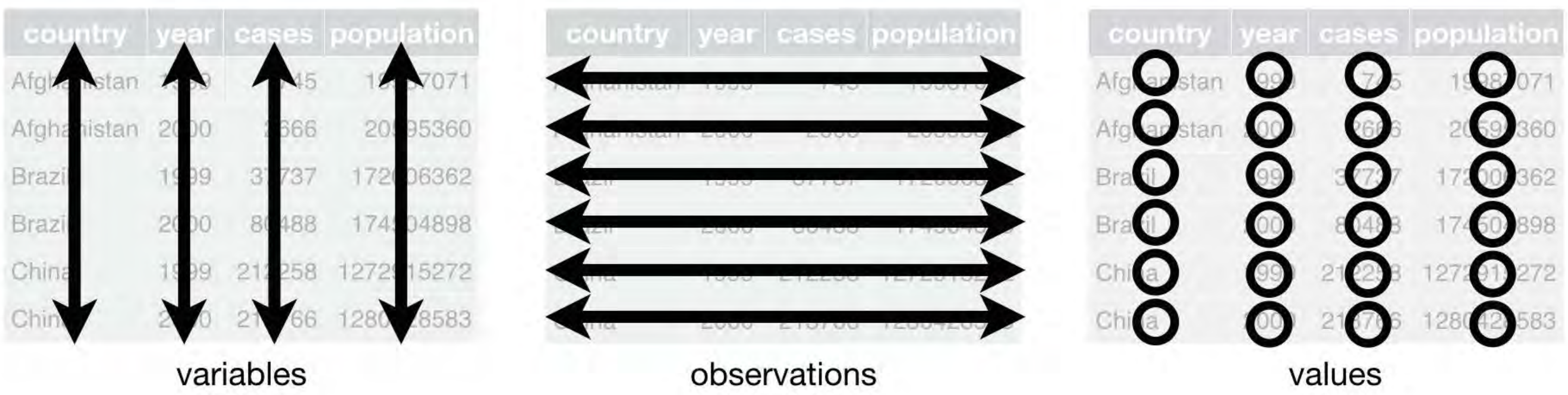
- Separate tables for each entity measured
- Each row represents a single observed entity
- Observations (rows) are all unique

Variables

- All values in a column are of the same type
- All columns pertain to the same observation (row)
- Each column represents either an identifying or measured variable

NCEAS
National Center for Ecological Analysis and Synthesis

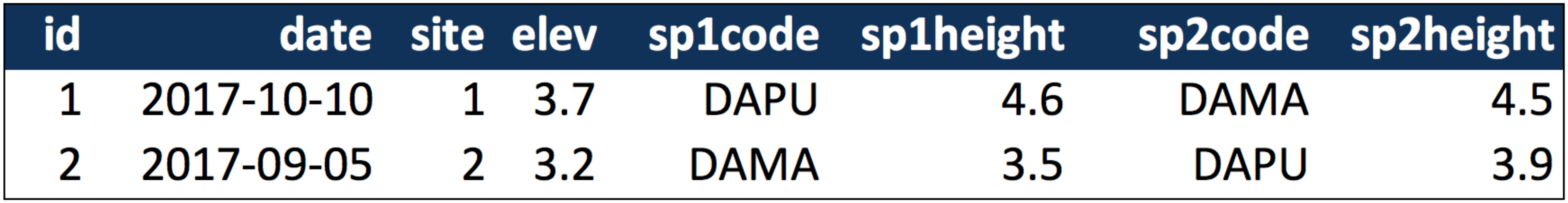
Characteristics of tidy data



NCEAS
National Center for Ecological Analysis and Synthesis

R for data science: import, tidy, transform, visualize, and model data. H Wickham, G Grolemund – 2016. https://r4ds.had.co.nz/

Recognizing untidy data

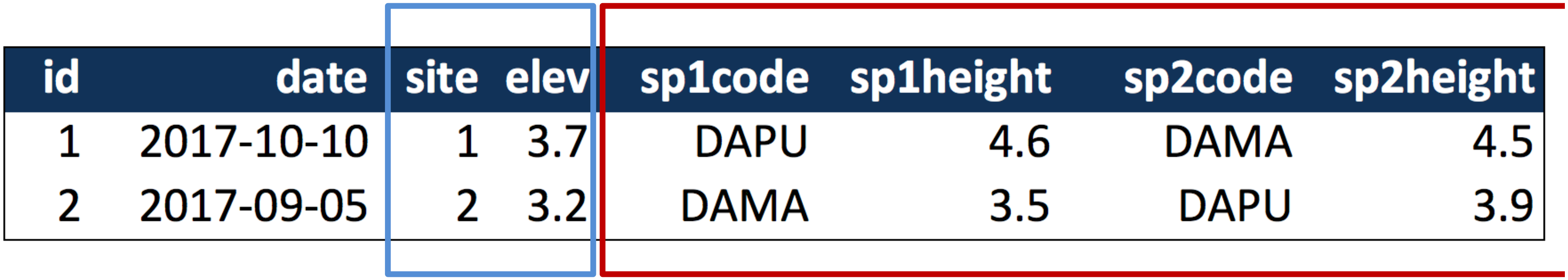


id	date	site	elev	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	3.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	3.2	DAMA	3.5	DAPU	3.9

- Each row contains observations about multiple entities (site characteristics and species observations)
- A new species observation would add a column (wide format)

NCEAS
National Center for Ecological Analysis and Synthesis

Tidying our data

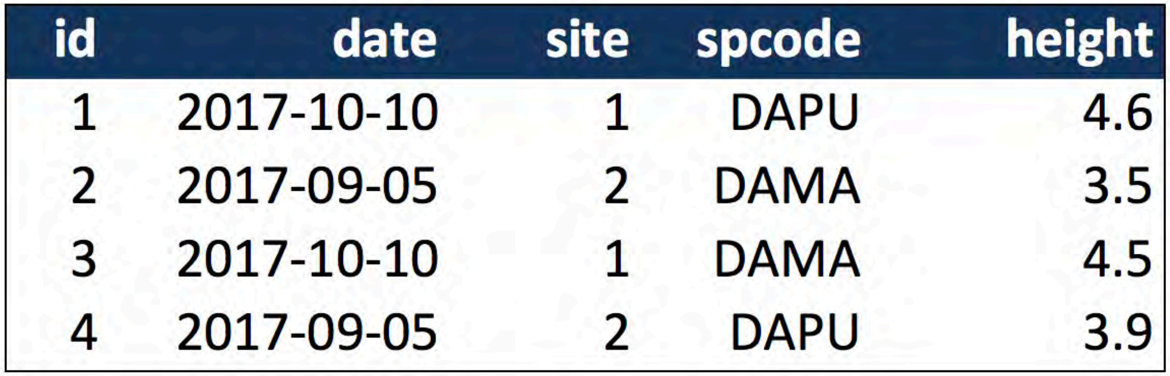


id	date	site	elev	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	3.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	3.2	DAMA	3.5	DAPU	3.9

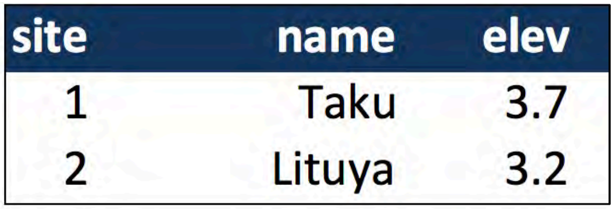
- What are the observed entities?
  - plant species
  - site characteristics
- What are the variables associated with those observations?
  - height
  - elevation

NCEAS
National Center for Ecological Analysis and Synthesis

Tidying our data



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



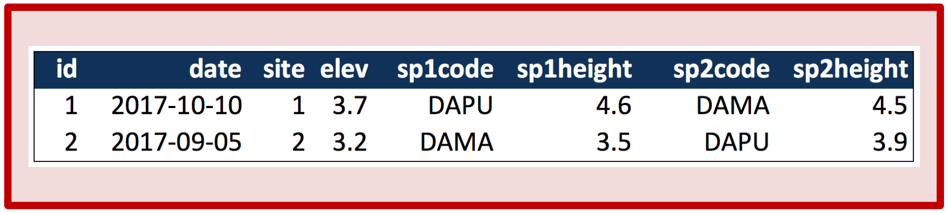
site	name	elev
1	Taku	3.7
2	Lituya	3.2

- Individual species observations
  - identifying variables: id, date, site, spcode
  - measured variables: height

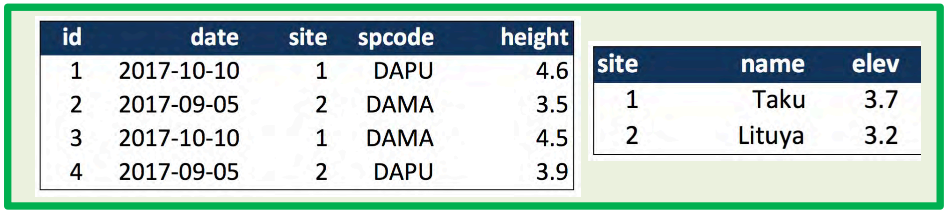
- Site observations where species occurred
  - identifying variables: site, name
  - measured variables: elev

NCEAS National Center for Ecological Analysis and Synthesis

Tidying our data



id	date	site	elev	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	3.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	3.2	DAMA	3.5	DAPU	3.9



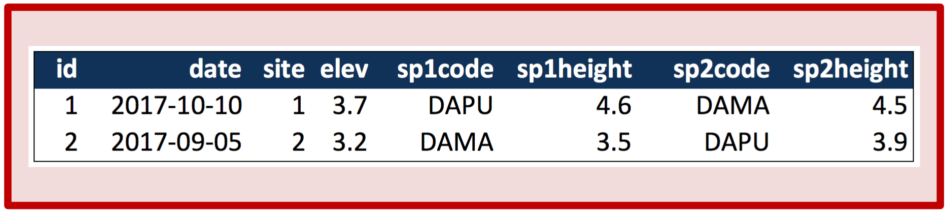
id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

site	name	elev
1	Taku	3.7
2	Lituya	3.2

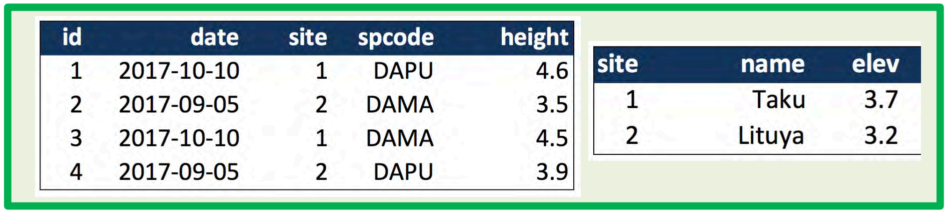
- Add rows not columns
- Separate information collected at different scales into different tables
- Record a single piece of data only once

NCEAS
National Center for Ecological Analysis and Synthesis

Benefits of normalized data



id	date	site	elev	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	3.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	3.2	DAMA	3.5	DAPU	3.9



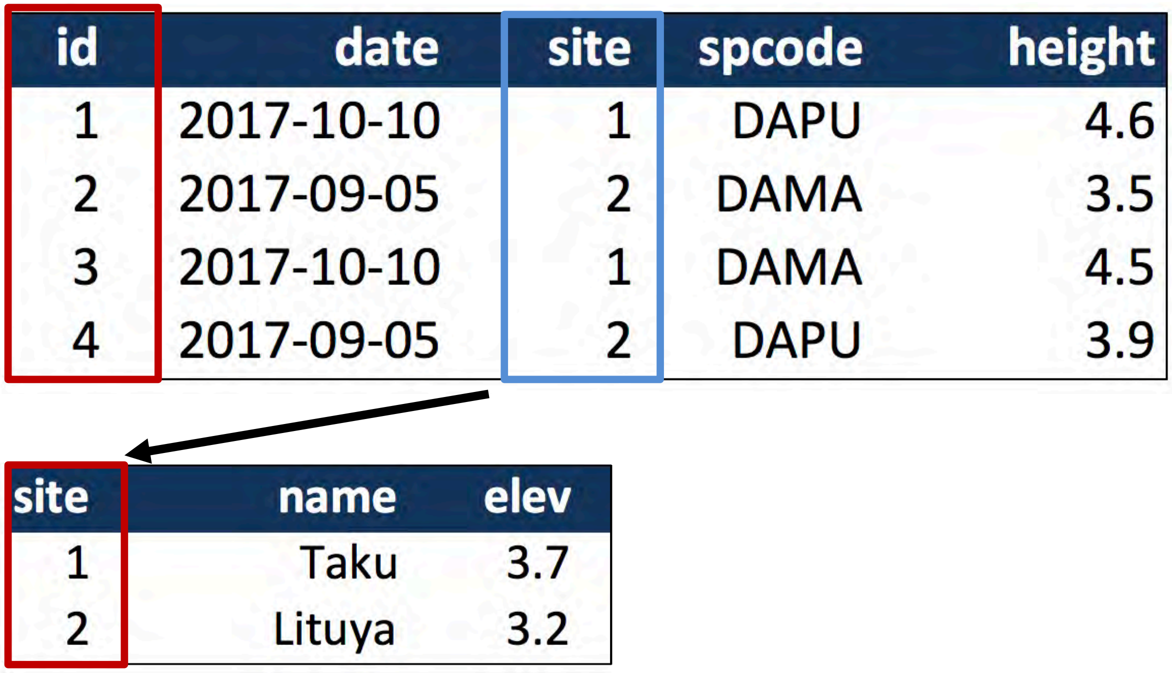
id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

site	name	elev
1	Taku	3.7
2	Lituya	3.2

- Search and filter rows
- Describe columns more precisely
- Optimize storage
- Enforce data integrity

NCEAS
National Center for Ecological Analysis and Synthesis

Using normalized data



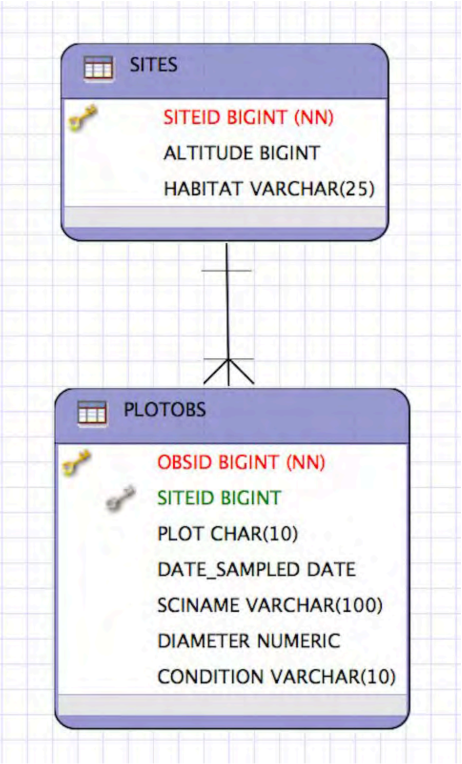
id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

site	name	elev
1	Taku	3.7
2	Lituya	3.2

- **Primary key**
  - unique identifier for each observation within an entity
- **Foreign Key**
  - reference to a primary key in another table

NCEAS
National Center for Ecological Analysis and Synthesis

Entity-relationship diagrams



SITES

SITEID BIGINT (NN)
ALTITUDE BIGINT
HABITAT VARCHAR(25)

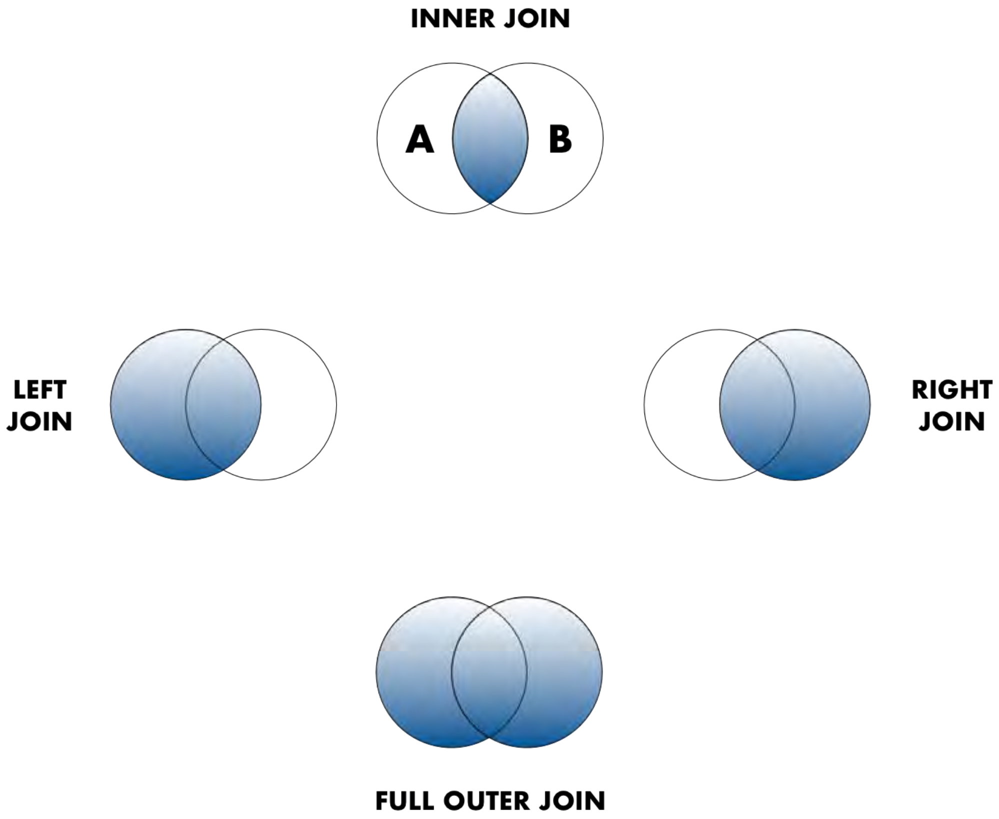
PLOTOBS

OBSID BIGINT (NN)
SITEID BIGINT
PLOT CHAR(10)
DATE_SAMPLED DATE
SCINAME VARCHAR(100)
DIAMETER NUMERIC
CONDITION VARCHAR(10)

- Draw relationships between tables concisely
- Used in database management systems

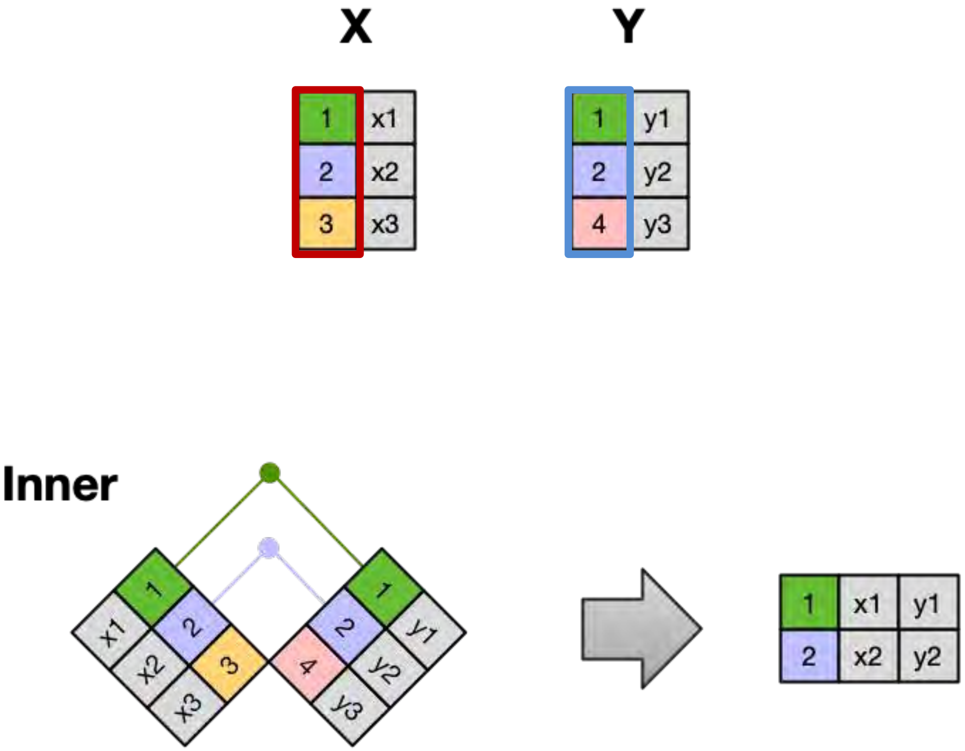
NCEAS
National Center for Ecological Analysis and Synthesis

Merging normalized data



NCEAS
National Center for Ecological Analysis and Synthesis

Merging normalized data



1	x1
2	x2
3	x3

1	y1
2	y2
4	y3

1	x1	y1
2	x2	y2

NCEAS
National Center for Ecological Analysis and Synthesis

R for data science: import, tidy, transform, visualize, and model data. H Wickham, G Grolemund – 2016. https://r4ds.had.co.nz/

Merging normalized data

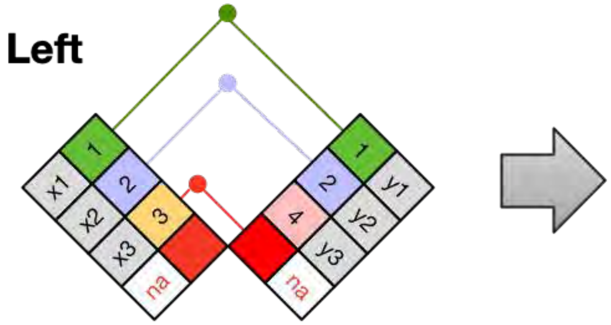
X

🟩 1	x1
2	x2
3	x3

Y

🟩 1	y1
2	y2
4	y3

Left



🟩 1	x1	y1
2	x2	y2
3	x3	na

NCEAS
National Center for Ecological Analysis and Synthesis

Merging normalized data

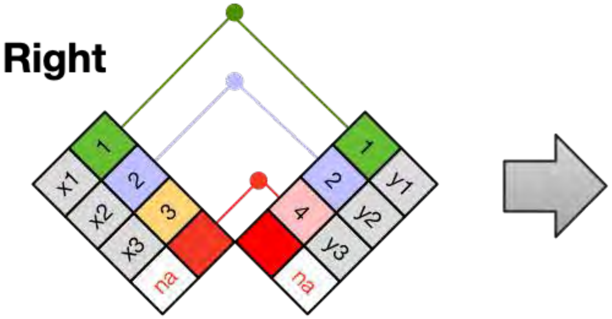
X

1	x1
2	x2
3	x3

Y

1	y1
2	y2
4	y3

Right



1	x1	y1
2	x2	y2
4	na	y3

Merging normalized data

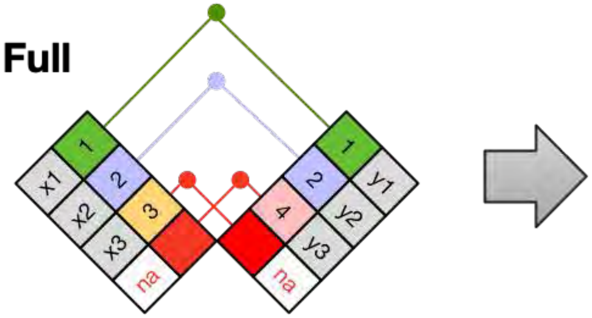
X

1	x1
2	x2
3	x3

Y

1	y1
2	y2
4	y3

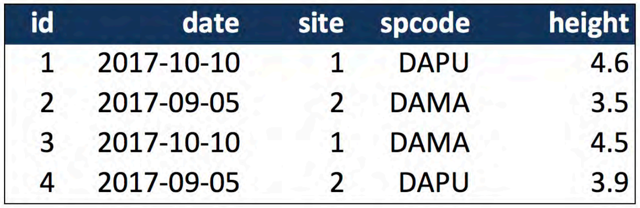
Full



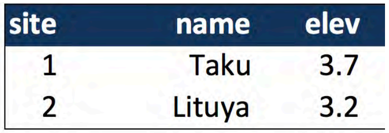
1	x1	y1
2	x2	y2
3	x3	na
4	na	y3

Merging normalized data

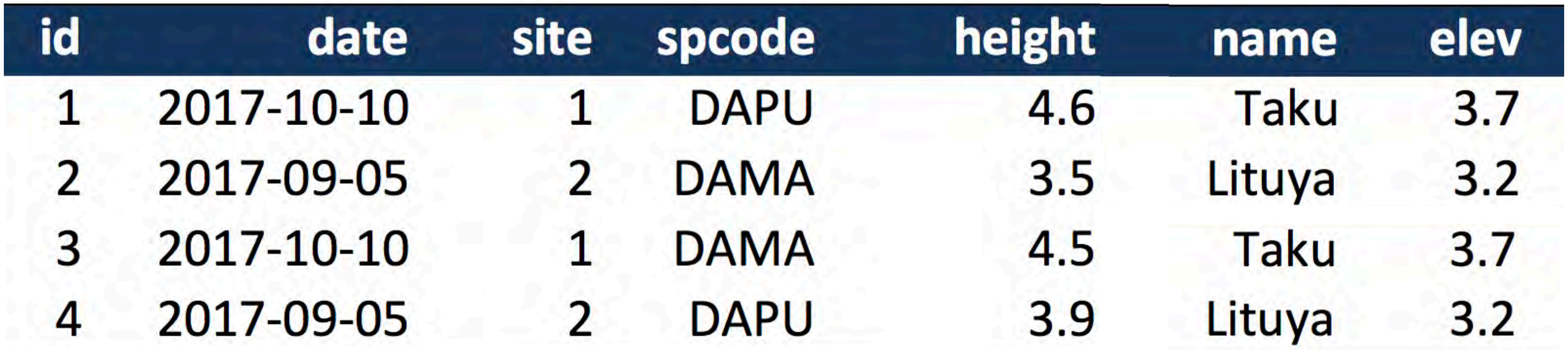
Left join



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



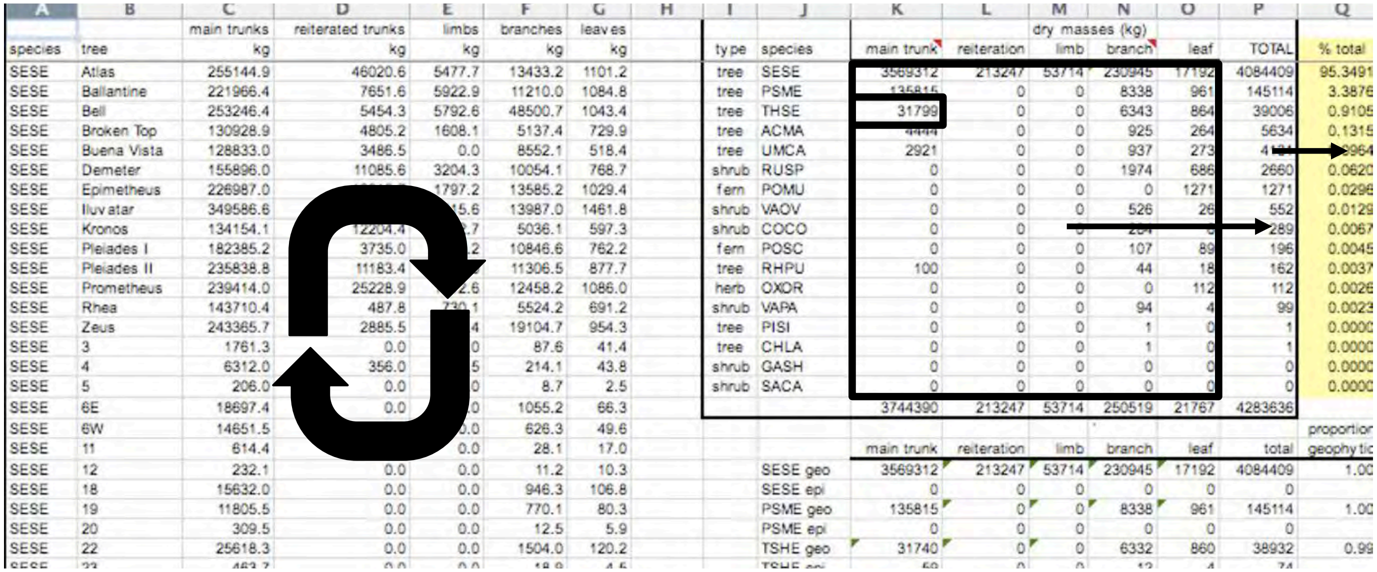
site	name	elev
1	Taku	3.7
2	Lituya	3.2



id	date	site	spcode	height	name	elev
1	2017-10-10	1	DAPU	4.6	Taku	3.7
2	2017-09-05	2	DAMA	3.5	Lituya	3.2
3	2017-10-10	1	DAMA	4.5	Taku	3.7
4	2017-09-05	2	DAPU	3.9	Lituya	3.2

NCEAS
National Center for Ecological Analysis and Synthesis

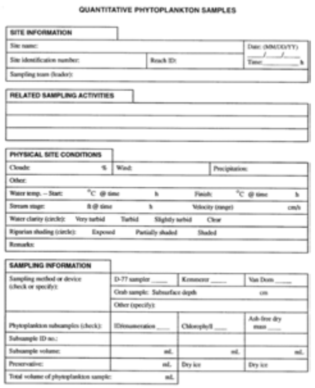
A not-so-reproducible workflow



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	[illegible]	1797.2	13585.2	1029.4
SESE	Iluvatar	349586.6	[illegible]	[illegible]	13987.0	1461.8
SESE	Kronos	134154.1	12204.4	[illegible]	5036.1	597.3
SESE	Pleiades I	182385.2	3735.0	[illegible]	10846.6	762.2
SESE	Pleiades II	235838.8	11183.4	[illegible]	11306.5	877.7
SESE	Prometheus	239414.0	25228.9	[illegible]	12458.2	1086.0
SESE	Rhea	143710.4	487.8	730.1	5524.2	691.2
SESE	Zeus	243365.7	2885.5	[illegible]	19104.7	954.3
SESE	3	1761.3	0.0	[illegible]	87.6	41.4
SESE	4	6312.0	356.0	[illegible]	214.1	43.8
SESE	5	206.0	0.0	[illegible]	8.7	2.5
SESE	6E	18697.4	0.0	[illegible]	1055.2	66.3
SESE	6W	14651.5	[illegible]	[illegible]	626.3	49.6
SESE	11	614.4	[illegible]	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

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	[illegible]	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	[illegible]	[illegible]
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	[illegible]	[illegible]	[illegible]	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 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	

Building a reproducible workflow



QUANTITATIVE PHYTOPLANKTON SAMPLES

SITE INFORMATION

Site name:

Date (MM/DD/YY) ___/___/___

Site identification number: Reach ID:

Time:______ h

Sampling team (leader):

RELATED SAMPLING ACTIVITIES

PHYSICAL SITE CONDITIONS

Clouds: % Wind: Precipitation:

Other:

Water temp. – Start: °C @ time h Finish: °C @ time h

Stream stage: ft @ time h Velocity (range) cm/s

Water clarity (circle): Very turbid Turbid Slightly turbid Clear

Riparian shading (circle): Exposed Partially shaded Shaded

Remarks:

SAMPLING INFORMATION

Sampling method or device (check or specify): D-77 sampler ____ Kemmerer ____ Van Dorn ____

Grab sample: Subsurface depth cm

Other (specify):

Phytoplankton subsamples (check): ID/enumeration ____ Chlorophyll ____ Ash-free dry mass ____

Subsample ID no.:

Subsample volume: mL mL mL

Preservative: mL Dry ice Dry ice

Total volume of phytoplankton sample: mL

Field data sheets

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

site	name	elev
1	Taku	3.7
2	Lituya	3.2

Tidy, raw data

V3

id	date	site	spcode	height	name	elev

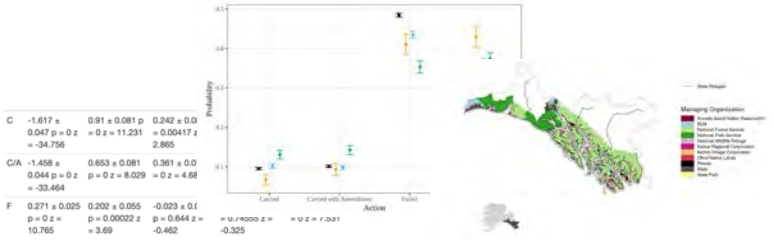
V2

id	date	site	spcode	height	name	elev

V1

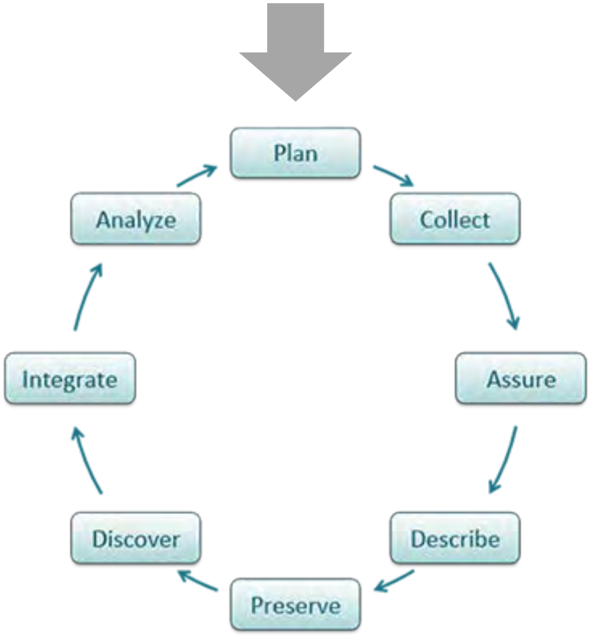
id	date	site	spcode	height	name	elev
1	2017-10-10	1	DAPU	4.6	Taku	3.7
2	2017-09-05	2	DAMA	3.5	Lituya	3.2
3	2017-10-10	1	DAMA	4.5	Taku	3.7
4	2017-09-05	2	DAPU	3.9	Lituya	3.2

Quality controlled, derived data, often merged or summarized



Figures, tables, maps

When to start?



- Thinking about your data model **early** helps you be more efficient at every stage of the data lifecycle

- Its never too late to tidy things up!

NCEAS
National Center for Ecological Analysis and Synthesis

Reproducible Research Techniques for Synthesis

A five day immersion into widely adopted R-based tools for open science

Details

Dates:
February 3-7, 2020
May 11-15, 2020

Location:
NCEAS
Santa Barbara, CA

Cost:
$2100
Includes: 5 days of instruction, refreshments and lunch.

NCEAS
National Center for Ecological Analysis and Synthesis

www.nceas.ucsb.edu/learning-hub/short-course

Questions? **www.nceas.ucsb.edu/learning-hub/short-course**

Jeanette Clark
National Center for Ecological Analysis and Synthesis
jclark@nceas.ucsb.edu

NCEAS
National Center for Ecological Analysis and Synthesis