

Relational Databases

Data Management for Biologists

Lind and Cariveau

February 19, 2018

Database

- An organized system for storing data

Databases

- Database: An organized system for storing data
- Database Management System (DBMS)
- NoSQL/NewSQL
 - MongoDB; JSON-based
 - Good for unstructured data



Databases

- Relational Databases (RDMS)
 - MySQL (Oracle), PostgreSQL, SQLite, many many others



PostgreSQL

Databases

- Relational Databases (RDMS)
 - Data are kept separate
 - Data are accessed via queries and data are not changed
 - Fast
 - It improves quality control of data entry

Databases

- Relational Databases (RDMS)
 - Table/Relation-based

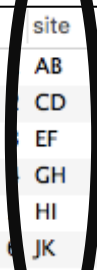
id	site	treatment	latitude	longitude	soil
1	AB	T	44.9442	-93.0936	clay
2	CD	C	44.426	-93.565	silt
3	EF	T	44.642	-92.0936	clay
4	GH	C	45.345	-93.77	sand
5	HI	T	43.999	-91.655	sand
6	JK	C	42.99	-92.0936	silt

uniqueID	genus	species	plant	site	fdate
1	Nomada	illinoensis	RUHI2	AB	2014-07-07
2	Nomada	illinoensis	BICO	AB	2014-07-07
3	Sphecodes	aroniae	BICO	AB	2014-07-07
4	Bombus	auricomus	BICO	AB	2014-07-07
5	Bombus	impatiens	LEVU	AB	2014-07-07
6	Bombus	impatiens	LEVU	AB	2014-07-07
7	Bombus	impatiens	MEOF	AB	2014-07-07
8	Bombus	impatiens	MEOF	AB	2014-07-07
9	Bombus	fervidus	RUHI2	AB	2014-08-07
10	Bombus	fervidus	RUHI2	AB	2014-08-07
11	Bombus	impatiens	RUHI2	AB	2014-08-07
12	Andrena	wilkella	BICO	AB	2014-08-07
13	Lasioglossum	admirandum	BICO	AB	2014-08-07
14	Bombus	impatiens	PEGR7	AB	2014-08-07

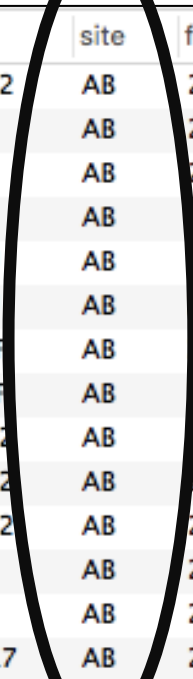
Tables (Relations)

Databases

- Relational Databases (RDMS)
 - Table/Relation-based



id	site	treatment	latitude	longitude	soil
	AB	T	44.9442	-93.0936	clay
	CD	C	44.426	-93.565	silt
	EF	T	44.642	-92.0936	clay
	GH	C	45.345	-93.77	sand
	HI	T	43.999	-91.655	sand
	JK	C	42.99	-92.0936	silt



uniqueID	genus	species	plant	site	fdate
1	Nomada	illinoensis	RUHI2	AB	2014-07-07
2	Nomada	illinoensis	BICO	AB	2014-07-07
3	Sphecodes	aroniae	BICO	AB	2014-07-07
4	Bombus	auricomus	BICO	AB	2014-07-07
5	Bombus	impatiens	LEVU	AB	2014-07-07
6	Bombus	impatiens	LEVU	AB	2014-07-07
7	Bombus	impatiens	MEOP	AB	2014-07-07
8	Bombus	impatiens	MEOP	AB	2014-07-07
9	Bombus	fervidus	RUHI2	AB	2014-08-07
10	Bombus	fervidus	RUHI2	AB	2014-08-07
11	Bombus	impatiens	RUHI2	AB	2014-08-07
12	Andrena	wilkella	BICO	AB	2014-08-07
13	Lasioglossum	admirandum	BICO	AB	2014-08-07
14	Bombus	impatiens	PEGR7	AB	2014-08-07

Variables (Attributes)

Databases

- Relational Databases (RDMS)
 - Table/Relation-based

id	site	treatment	latitude	longitude	soil
1	AB	T	44.9442	-93.0936	clay
2	CB	C	44.426	-93.558	silt
3	EF	T	44.642	-92.0936	clay
4	GH	C	45.345	-93.77	sand
5	HI	T	43.999	-91.655	sand
6	JK	C	42.99	-92.0936	silt

uniqueID	genus	species	plant	site	fdate
1	Nomada	illinoensis	RUHI2	AB	2014-07-07
2	Nomada	illinoensis	BICO	AB	2014-07-07
3	Sphecodes	aroniae	BICO	AB	2014-07-07
4	Bombus	auricomus	BICO	AB	2014-07-07
5	Bombus	impatiens	LEVU	AB	2014-07-07
6	Bombus	impatiens	LEVU	AB	2014-07-07
7	Bombus	impatiens	MEOF	AB	2014-07-07
8	Bombus	impatiens	MEOF	AB	2014-07-07
9	Bombus	fervidus	RUHI2	AB	2014-08-07
10	Bombus	fervidus	RUHI2	AB	2014-08-07
11	Bombus	impatiens	RUHI2	AB	2014-08-07
12	Andrena	wilkella	BICO	AB	2014-08-07
13	Lasioglossum	admirandum	BICO	AB	2014-08-07
14	Bombus	impatiens	PEGR7	AB	2014-08-07

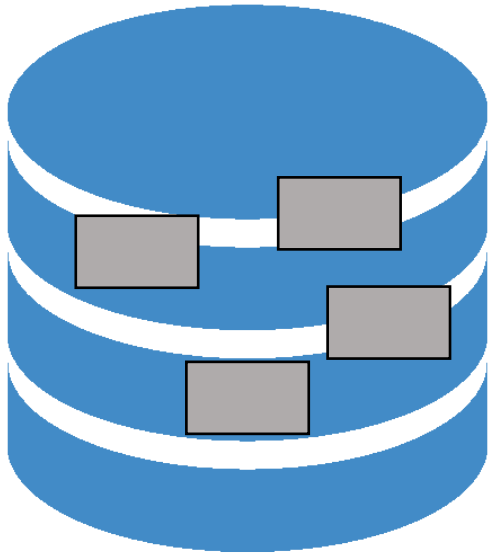
Rows (Tuples)

Many relations vs. one big table

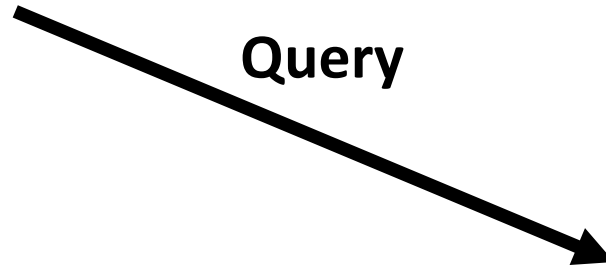
uniqueID	fdate	round	site	treatment	genus	species	detby	plant_genus	plant_species	collect
NAT.CIG2011_1000	2011-07-07	2	MO	T	Halictus	confusus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1003	2011-07-07	2	MO	T	Halictus	ligatus	Ascher	Leucanthemum	vulgare	CLW
NAT.CIG2011_1005	2011-07-07	2	MO	T	Augochlora	pura	Ascher	Asclepias	tuberosa	CLW
NAT.CIG2011_1006	2011-07-07	2	MO	T	Halictus	ligatus	Ascher	Rudbeckia	hirta	CLW
NAT.CIG2011_1007	2011-07-07	2	MO	T	Ceratina	calcarata_dupla_mikmaqi	Ascher	Rudbeckia	hirta	CLW
NAT.CIG2011_1012	2011-07-07	2	MO	T	Halictus	ligatus	Ascher	Rudbeckia	hirta	CLW
NAT.CIG2011_1013	2011-07-07	2	MO	C	Halictus	ligatus	Ascher	Leucanthemum	vulgare	CLW
NAT.CIG2011_1014	2011-07-07	2	MO	C	Halictus	confusus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1015	2011-07-07	2	MO	C	Halictus	ligatus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1016	2011-07-07	2	MO	C	Halictus	ligatus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1021	2011-07-07	2	MO	C	Halictus	ligatus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1022	2011-07-07	2	MO	C	Halictus	ligatus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1023	2011-07-07	2	MO	C	Halictus	confusus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1024	2011-07-07	2	MO	C	Halictus	ligatus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1025	2011-07-07	2	MO	C	Halictus	confusus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1027	2011-07-07	2	MO	C	Halictus	ligatus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1028	2011-07-07	2	MO	C	Halictus	ligatus	Ascher	Achillea	millefolium	CLW
NAT.CIG2011_1032	2011-07-11	2	MO	T	Halictus	ligatus	Ascher	Erigeron	philadelphicus	CLW
NAT.CIG2011_1033	2011-07-11	2	MO	T	Halictus	ligatus	Ascher	Erigeron	philadelphicus	CLW
NAT.CIG2011_1035	2011-07-11	2	MO	T	Agapostemon	virescens	Ascher	Digitalis	purpurea	CLW
NAT.CIG2011_1036	2011-07-11	2	MO	T	Halictus	ligatus	Ascher	Coreopsis	lanceolata	CLW
NAT.CIG2011_1037	2011-07-11	2	MO	T	Halictus	ligatus	Ascher	Erigeron	philadelphicus	CLW

SQL

- Structured Query Language



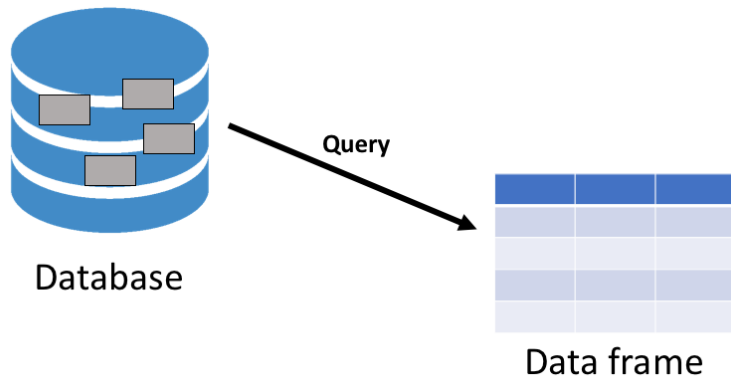
Database



Data frame

SQL

- Structured Query Language



Queries:

**Select, join, group,
aggregate, calculate**

Databases

- Relational Databases (RDMS)
 - Table/Relation-based
 - Based on relational algebra (Codd 1977)

id	site	treatment	latitude	longitude	soil
1	AB	T	44.9442	-93.0936	clay
2	CD	T	44.426	-93.565	silt
3	EF	T	44.642	-92.0936	clay
4	GH	C	45.345	-93.77	sand
5	HI	T	43.999	-91.655	sand
6	JK	C	42.99	-92.0936	silt

uniqueID	genus	species	plant	site	fdate
1	Nomada	illinoensis	RUH1	AB	2014-07-07
2	Nomada	illinoensis	BICO	AB	2014-07-07
3	Sphecodes	aroniae	BICO	AB	2014-07-07
4	Bombus	auricomus	BICO	AB	2014-07-07
5	Bombus	impatiens	LEVJ	AB	2014-07-07
6	Bombus	impatiens	LEVJ	AB	2014-07-07
7	Bombus	impatiens	MEBF	AB	2014-07-07
8	Bombus	impatiens	MEBF	AB	2014-07-07
9	Bombus	fervidus	RUH12	AB	2014-08-07
10	Bombus	fervidus	RUH12	AB	2014-08-07
11	Bombus	impatiens	RUH12	AB	2014-08-07
12	Andrena	wilkella	BICO	AB	2014-08-07
13	Lasioglossum	admirandum	BICO	AB	2014-08-07
14	Bombus	impatiens	PEGR7	AB	2014-08-07

Attributes

Databases

- Relational Databases (RDMS)
 - Table/Relation-based
 - Based on relational algebra (Codd 1977)

id	site	treatment	latitude	longitude	soil
1	AB	T	44.9442	-93.0936	clay
2	CD	T	44.426	-93.565	silt
3	EF	T	44.642	-92.0936	clay
4	GH	C	45.345	-93.77	sand
5	HI	T	43.999	-91.655	sand
6	JK	C	42.99	-92.0936	silt

uniqueID	genus	species	plant	site	date
1	Nomada	illinoensis	RUH1	AB	2014-07-07
2	Nomada	illinoensis	BICO	AB	2014-07-07
3	Sphecodes	aroniae	BICO	AB	2014-07-07
4	Bombus	auricomus	BICO	AB	2014-07-07
5	Bombus	impatiens	LEVJ	AB	2014-07-07
6	Bombus	impatiens	LEVJ	AB	2014-07-07
7	Bombus	impatiens	MEBF	AB	2014-07-07
8	Bombus	impatiens	MEBF	AB	2014-07-07
9	Bombus	fervidus	RUH12	AB	2014-08-07
10	Bombus	fervidus	RUH12	AB	2014-08-07
11	Bombus	impatiens	RUH12	AB	2014-08-07
12	Andrena	wilkella	BICO	AB	2014-08-07
13	Lasioglossum	admirandum	BICO	AB	2014-08-07
14	Bombus	impatiens	PEGR7	AB	2014-08-07

Attributes

Databases

- Relational Databases (RDMS)
 - Table/Relation-based
 - Based on relational algebra (Codd 1977)

1:n

id	site	treatment	latitidue	longititude	soil
1	AB	T	44.9442	-93.0936	clay
2	CD	T	44.426	-93.565	silt
3	EF	T	44.642	-92.0936	clay
4	GH	C	45.345	-93.77	sand
5	HI	T	43.999	-91.655	sand
6	JK	C	42.99	-92.0936	silt

uniqueID	genus	species	plant	site	fdate
1	Nomada	illinoensis	RUH1	AB	2014-07-07
2	Nomada	illinoensis	BICO	AB	2014-07-07
3	Sphecodes	aroniae	BICO	AB	2014-07-07
4	Bombus	auricomus	BICO	AB	2014-07-07
5	Bombus	impatiens	LEVJ	AB	2014-07-07
6	Bombus	impatiens	LEVJ	AB	2014-07-07
7	Bombus	impatiens	MEBF	AB	2014-07-07
8	Bombus	impatiens	MEBF	AB	2014-07-07
9	Bombus	fervidus	RUH12	AB	2014-08-07
10	Bombus	fervidus	RUH12	AB	2014-08-07
11	Bombus	impatiens	RUH12	AB	2014-08-07
12	Andrena	wilkella	BICO	AB	2014-08-07
13	Lasioglossum	admirandum	BICO	AB	2014-08-07
14	Bombus	impatiens	PEGR7	AB	2014-08-07

Attributes

Databases

site_id	site	treatment	latitude	longitude	soil
1	AB	T	44.9442	-93.0936	clay
2	CD	C	44.426	-93.565	silt
3	EF	T	44.642	-92.0936	clay
4	GH	C	45.345	-93.77	sand
5	HI	T	43.999	-91.655	sand
6	JK	T	42.99	-92.0936	silt

	species	plant	site	fdate	siteID	
1	Andrena	illinoensis	RUHI2	AB	2014-07-07	1
2	Andrena	illinoensis	BICO	AB	2014-07-07	1
3	Sphecodes	aroniae	BICO	AB	2014-07-07	1
4	Bombus	auricomus	BICO	AB	2014-07-07	1
5	Bombus	impatiens	LEVU	AB	2014-07-07	1
6	Bombus	impatiens	LEVU	AB	2014-07-07	1
7	Bombus	impatiens	MEOF	AB	2014-07-07	1
8	Bombus	impatiens	MEOF	AB	2014-07-07	1
9	Bombus	fervidus	RUHI2	AB	2014-08-07	1
10	Bombus	fervidus	RUHI2	AB	2014-08-07	1
11	Bombus	impatiens	RUHI2	AB	2014-08-07	1
12	Andrena	wilkella	BICO	AB	2014-08-07	1
13	Lasioglossum	admirandum	BICO	AB	2014-08-07	1
14	Bombus	impatiens	PEGR7	AB	2014-08-07	1
15	Bombus	impatiens	PEGR7	AB	2014-08-07	1
16	Nomada	illinoensis	LEVU	CD	2014-07-08	2
17	Andrena	wilkella	LEVU	CD	2014-07-08	2
18	Nomada	illinoensis	RUHI2	CD	2014-07-08	2
19	Andrena	wilkella	RUHI2	CD	2014-07-08	2
20	Bombus	impatiens	RUHI2	CD	2014-07-08	2

Primary Key
Unique (no repeats)

Databases

site_id	site	treatment	latitude	longitude	soil
1	AB	T	44.9442	-93.0936	clay
2	CD	C	44.426	-93.565	silt
3	EF	T	44.642	-92.0936	clay
4	GH	C	45.345	-93.77	sand
5	HI	T	43.999	-91.655	sand
6	JK	C	42.99	-92.0936	silt

	species	plant	site	date	siteID
1	illinoensis	RUHI2	AB	2014-07-07	1
2	Nomada illinoensis	BICO	AB	2014-07-07	1
3	Sphecodes	BICO	AB	2014-07-07	1
4	Bombus auricomus	BICO	AB	2014-07-07	1
5	Bombus impatiens	LEVU	AB	2014-07-07	1
6	Bombus impatiens	LEVU	AB	2014-07-07	1
7	Bombus impatiens	MEOF	AB	2014-07-07	1
8	Bombus impatiens	MEOF	AB	2014-07-07	1
9	Bombus fervidus	RUHI2	AB	2014-08-07	1
10	Bombus fervidus	RUHI2	AB	2014-08-07	1
11	Bombus impatiens	RUHI2	AB	2014-08-07	1
12	Andrena wilkella	BICO	AB	2014-08-07	1
13	Lasioglossum admirandum	BICO	AB	2014-08-07	1
14	Bombus impatiens	PEGR7	AB	2014-08-07	1
15	Bombus impatiens	PEGR7	AB	2014-08-07	1
16	Nomada illinoensis	LEVU	CD	2014-07-08	2
17	Andrena wilkella	LEVU	CD	2014-07-08	2
18	Nomada illinoensis	RUHI2	CD	2014-07-08	2
19	Andrena wilkella	RUHI2	CD	2014-07-08	2
20	Bombus impatiens	RUHI2	CD	2014-07-08	2

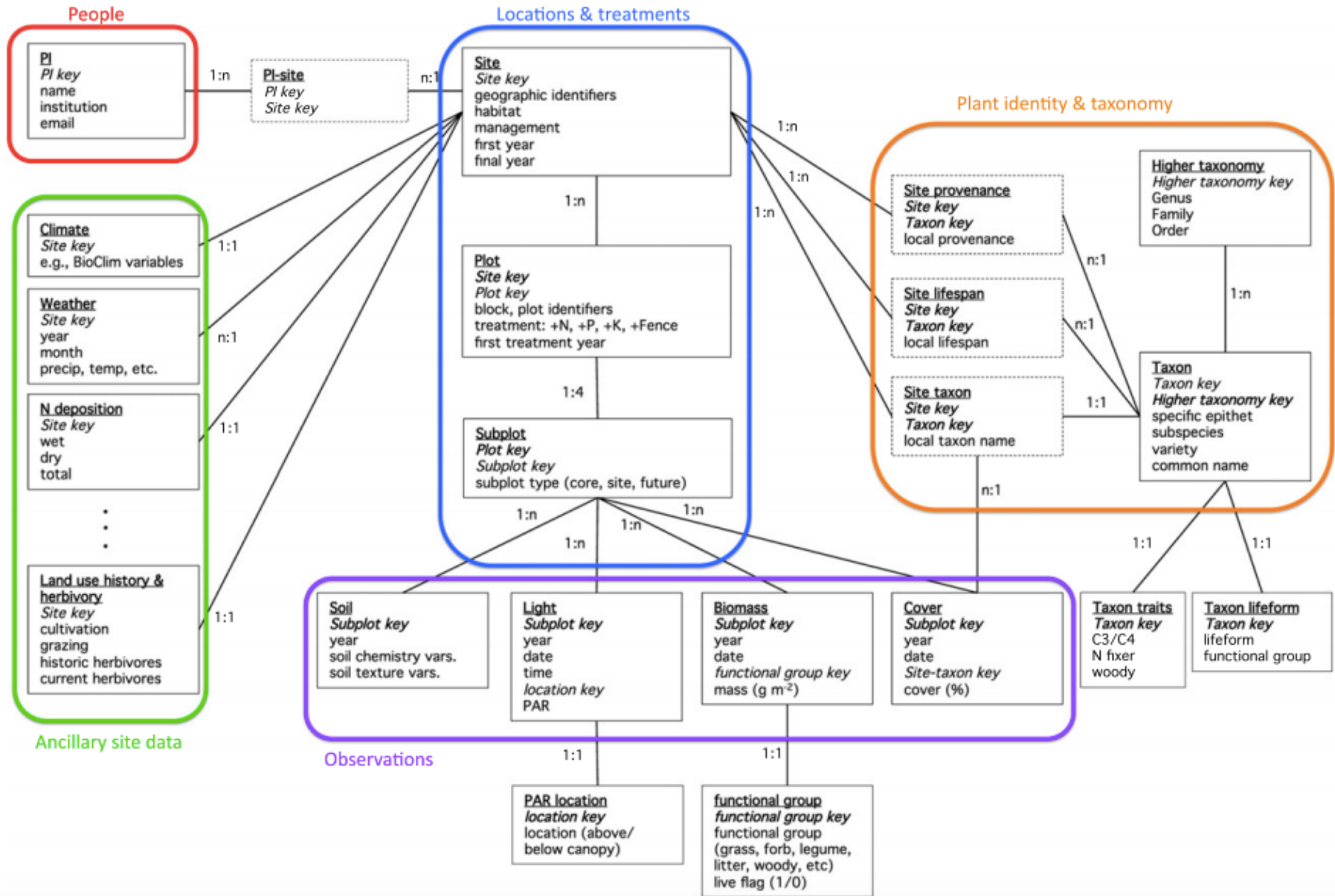
Foreign Key
Can have repeats



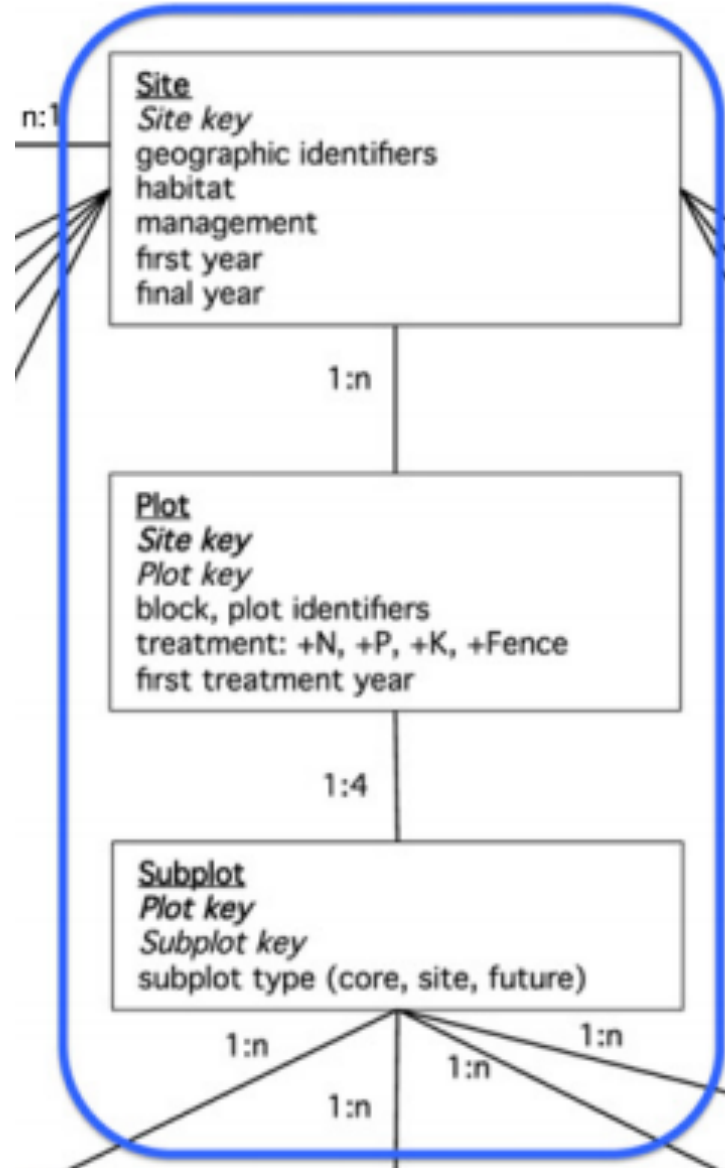
Databases

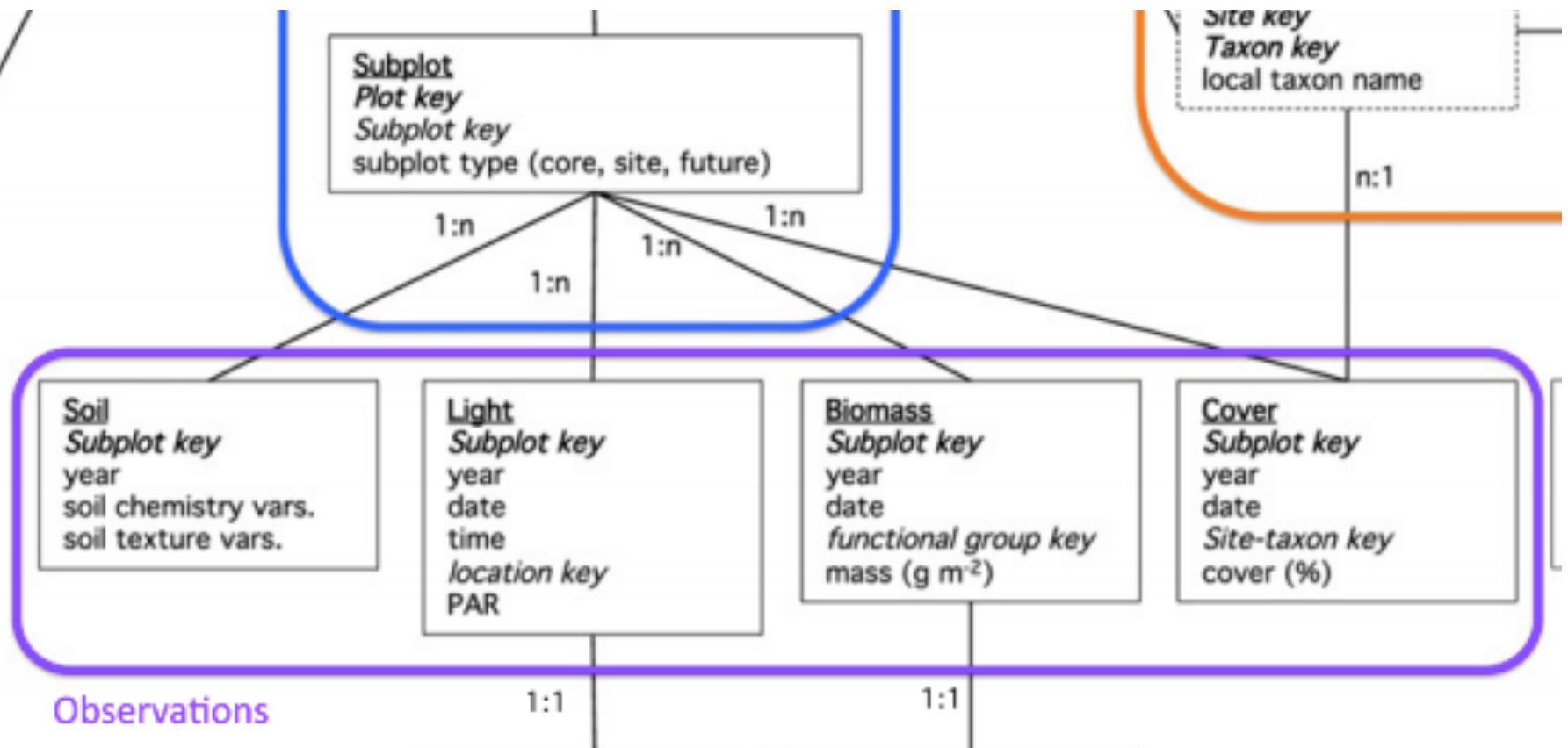
site_ID	site	treatment	latitude	longitude	soil	uniqueID	genus	species	plant	site	fdate
1	AB	T	44.9442	-93.0936	clay	1	Nomada	illinoensis	RUHI2	AB	2014-07-07
1	AB	T	44.9442	-93.0936	clay	2	Nomada	illinoensis	BICO	AB	2014-07-07
1	AB	T	44.9442	-93.0936	clay	3	Sphecodes	aroniae	BICO	AB	2014-07-07
1	AB	T	44.9442	-93.0936	clay	4	Bombus	auricomus	BICO	AB	2014-07-07
1	AB	T	44.9442	-93.0936	clay	5	Bombus	impatiens	LEVU	AB	2014-07-07
1	AB	T	44.9442	-93.0936	clay	6	Bombus	impatiens	LEVU	AB	2014-07-07
1	AB	T	44.9442	-93.0936	clay	7	Bombus	impatiens	MEOF	AB	2014-07-07
1	AB	T	44.9442	-93.0936	clay	8	Bombus	impatiens	MEOF	AB	2014-07-07
1	AB	T	44.9442	-93.0936	clay	9	Bombus	fervidus	RUHI2	AB	2014-08-07
1	AB	T	44.9442	-93.0936	clay	10	Bombus	fervidus	RUHI2	AB	2014-08-07
1	AB	T	44.9442	-93.0936	clay	11	Bombus	impatiens	RUHI2	AB	2014-08-07
1	AB	T	44.9442	-93.0936	clay	12	Andrena	wilkella	BICO	AB	2014-08-07
1	AB	T	44.9442	-93.0936	clay	13	Lasioglossum	admirandum	BICO	AB	2014-08-07
1	AB	T	44.9442	-93.0936	clay	14	Bombus	impatiens	PEGR7	AB	2014-08-07
1	AB	T	44.9442	-93.0936	clay	15	Bombus	impatiens	PEGR7	AB	2014-08-07
2	CD	C	44.426	-93.565	silt	16	Nomada	illinoensis	LEVU	CD	2014-07-08
2	CD	C	44.426	-93.565	silt	17	Andrena	wilkella	LEVU	CD	2014-07-08

Join in MySQL
Merge in R



Locations & treatments





Site table

siteID
site
treatment
latitude
longitude
soil

Schema

Specimen table

uniqueID
genus
species
plant
side
fdate
siteID

Collection event table

collevent_id
site
fdate
weather
temp
wind
siteID

Schema

As a group:

1. Circle the primary keys in each table
2. Circle the foreign keys in each table (not all tables have one)
3. Draw relationships indicating one to many among tables
4. Describe 2-3 tables from your own data (or to be collected data) and discuss relationships

Tidy Data

- Store data as tidy data – messy data maybe need for analyses (but do this using code!)
- Example: community-level data in vegan

Site	Species1	Species2	Species3
A	0	1	10
B	10	3	10
C	2	3	5
D	3	0	16
E	0	0	20

Tidy Data

- Columns are values and not variable names

religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Tidy Data

- Columns are values and not variable names

religion	income	freq
Agnostic	<\$10k	27
Agnostic	\$10-20k	34
Agnostic	\$20-30k	60
Agnostic	\$30-40k	81
Agnostic	\$40-50k	76
Agnostic	\$50-75k	137
Agnostic	\$75-100k	122
Agnostic	\$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Tidy Data

- Multiple variables in one column

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Tidy Data

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Tidy Data

- Multiple variables in one column

uniqueID	year	round	fdate	site	genus	species
NAT.CIG2011	2011	1	5/31/11	HP	Andrena	brevipalpis
NAT.CIG2011	2011	2	7/7/11	MO	Augochlora	pura
NAT.CIG2011	2011	2	7/7/11	MO	Ceratina	calcarata
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	confusus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus

Tidy Data

- Multiple variables in one column

uniqueID	year	round	fdate	site	genus	species
NAT.CIG2011	2011	1	5/31/11	HP	Andrena	brevipalpis
NAT.CIG2011	2011	2	7/7/11	MO	Augochlora	pura
NAT.CIG2011	2011	2	7/7/11	MO	Ceratina	calcarata
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	confusus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus

Easy to do in R: `data$gen_sp <-
paste(data$genus, data$species, sep = "_")`

Tidy Data

- Multiple variables in one column

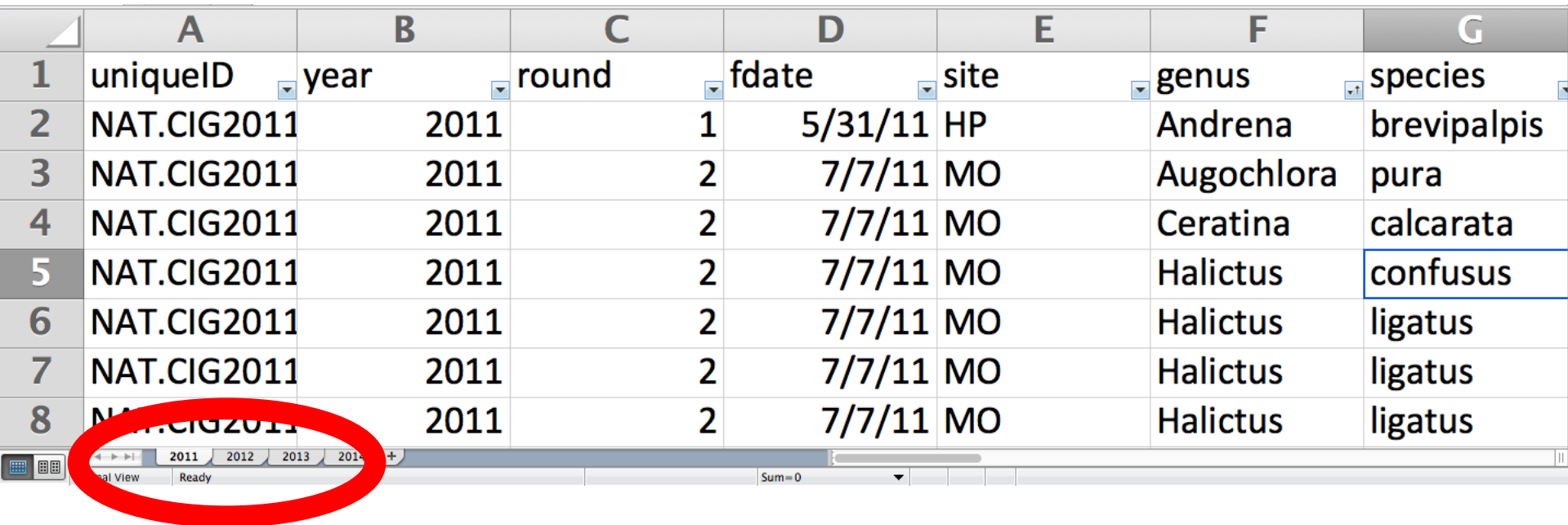
uniqueID	year	round	fdate	site	genus	species
NAT.CIG2011	2011	1	5/31/11	HP	Andrena	brevipalpis
NAT.CIG2011	2011	2	7/7/11	MO	Augochlora	pura
NAT.CIG2011	2011	2	7/7/11	MO	Ceratina	calcarata
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	confusus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus
NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus

Easy to do in MySQL:

```
select variable 1, variable2, concat(genus, "_",  
species) from data
```

Tidy Data

- Many tables when there should be one



	A	B	C	D	E	F	G
1	uniqueID	year	round	fdate	site	genus	species
2	NAT.CIG2011	2011	1	5/31/11	HP	Andrena	brevipalpis
3	NAT.CIG2011	2011	2	7/7/11	MO	Augochlora	pura
4	NAT.CIG2011	2011	2	7/7/11	MO	Ceratina	calcarata
5	NAT.CIG2011	2011	2	7/7/11	MO	Halictus	confusus
6	NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus
7	NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus
8	NAT.CIG2011	2011	2	7/7/11	MO	Halictus	ligatus