Data Structures and Formats

A few best practices

Rosemary Hartman

DWR

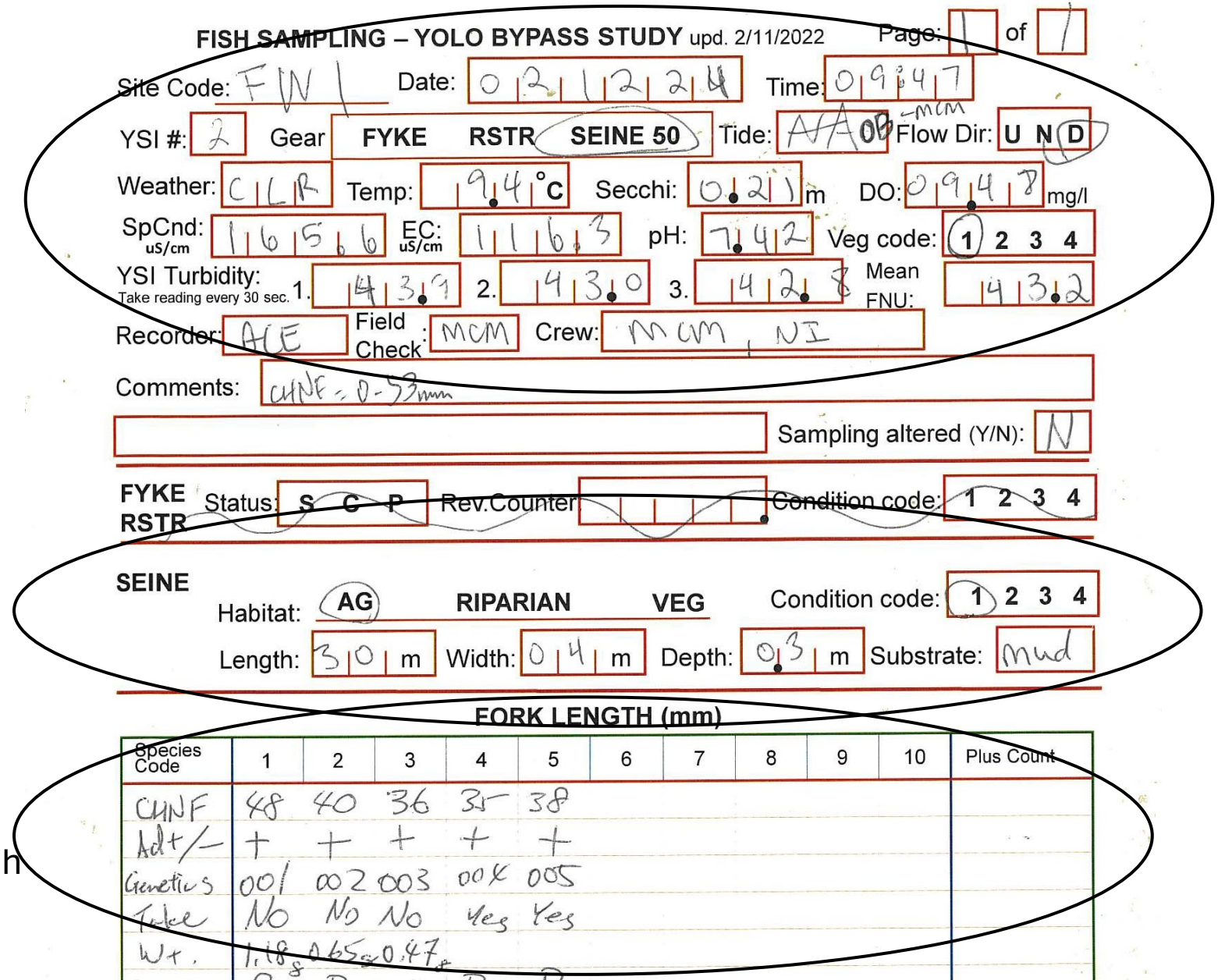

DISCLAIMER
YOUR MILAGE MAY VARY
imgflip.com

So you have some data.

Information about the surrounding environment

Information about sampling effort

Information about each fish



FISH SAMPLING – YOLO BYPASS STUDY upd. 2/11/2022 Page: 1 of 1

Site Code: FWL Date: 02/12/24 Time: 0947

YSI #: 2 Gear: FYKE RSTR SEINE 50 Tide: AA OB -mcm Flow Dir: U N D

Weather: CLR Temp: 9.4 °C Secchi: 0.21 m DO: 09.48 mg/l

SpCnd: 1656 uS/cm EC: 1163 uS/cm pH: 7.42 Veg code: 1 2 3 4

YSI Turbidity: Take reading every 30 sec. 1. 431 2. 430 3. 428 Mean FNU: 431.2

Recorder: ACE Field Check: MCM Crew: MCM, NI

Comments: CHNF = 0-53mm

Sampling altered (Y/N): N

FYKE RSTR Status: S C P Rev.Counter: Condition code: 1 2 3 4

SEINE

Habitat: AG RIPARIAN VEG Condition code: 1 2 3 4

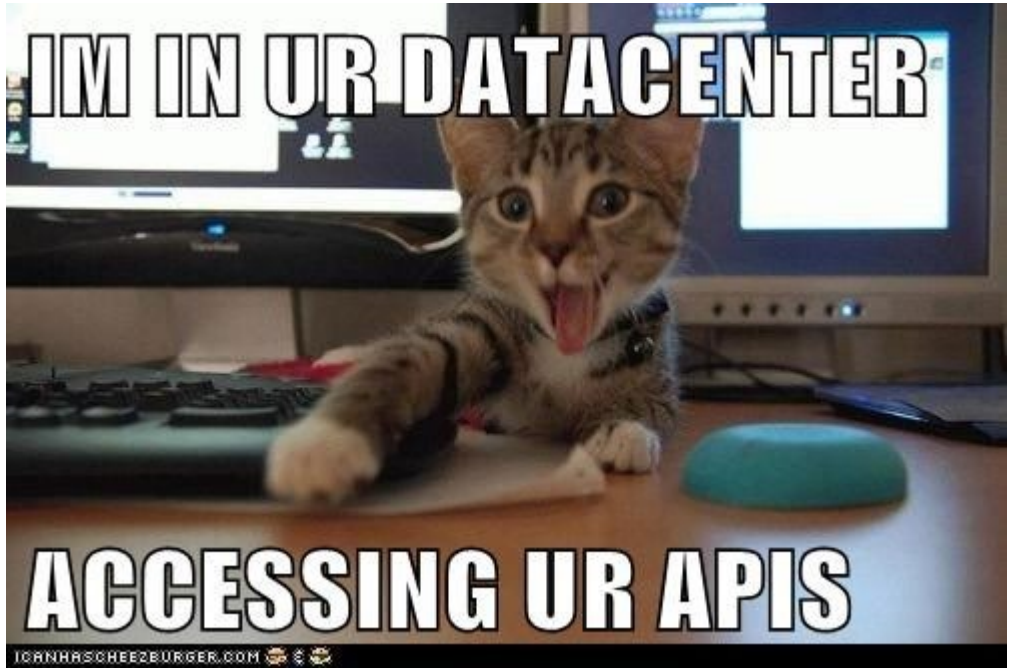
Length: 30 m Width: 04 m Depth: 03 m Substrate: mud

FORK LENGTH (mm)

Species Code	1	2	3	4	5	6	7	8	9	10	Plus Count
CHNF	48	40	36	35	38						
Ad+/-	+	+	+	+	+						
Genetics	001	002	003	004	005						
Take	No	No	No	Yes	Yes						
Wt.	1.18g	0.65g	0.47g								

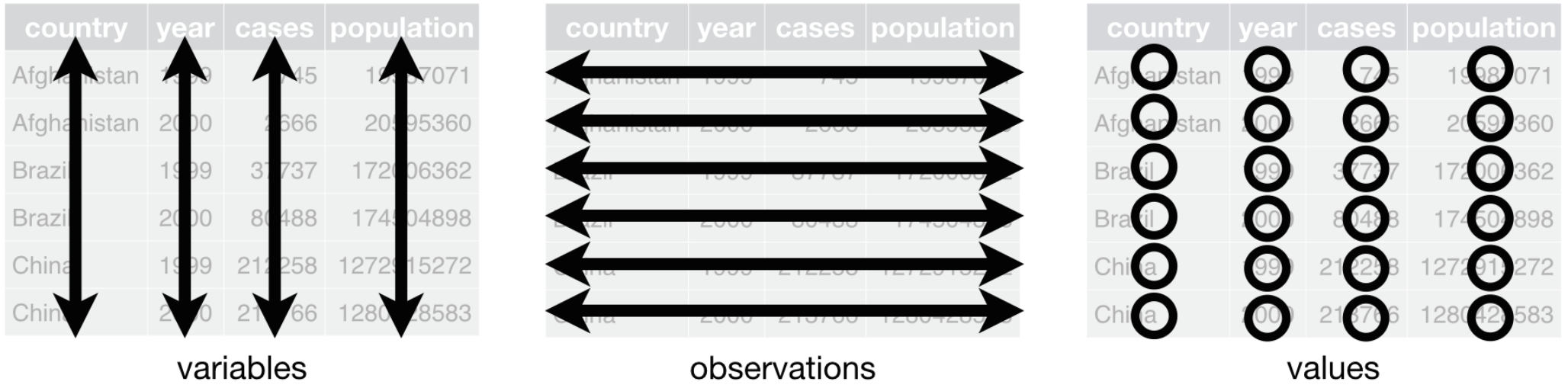
Your database should be…

- Consistent (but flexible)
- Efficient
- Easily queryable
- As simple as possible (and no simpler)
- Sustainable



Basic rules

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.



Credit – R for Data Science, by Hadley Wickham

What not to do

Multiple columns with same name

Multiple columns with same type of info

Repeated information

Site Code	Date	Time	Tide	Temperature	Secchi	DO	Sample Type	Gear ID	Length (m)	Width	Vol	Time	Species 1	Count	Lengths	Species 2	Count	Lengths	Comments
A1	2024-01-01	6:34	Ebb	18	74	12.2	Seine	Seine1	30	15	225	NA	CHISAL	16	23, 45, 56	WHICAT	2	76, 23	
A1	2024-01-01	6:34	Ebb	18	74	12.2	Seine	Seine1	23	6	69	NA	WHICAT	5	Didn't measure	MISSIL	135	123, 24,56	Geneitc tag 001
A1	2024-01-01	6:34	Ebb	18	74	12.2	Zooplankton	Zoop3	18	0.2	1.8	5							
BL5	2024-01-04	8:04	Low Slack	17.8	Forgot secchi	9.9	Fyke	Fyke1	none	none	none	354	REDEAR	1	110	BLABAS	1	345	Geneitc tag 002
C3	2024-01-03	10:43	High Slack	12.6	110	10.2	Seine	Seine2	34	12	345	NA	WHICAT	5	Didn't measure	MISSIL	135	123, 24,56	

Comments in numeric columns

Multiple values stored in one cell

Multiple indicators of missing values

Valuable info in com

So what should we do?

Information about the surrounding environment

TableSiteVist

VisitID	SiteCode	Date	Time	Tide	Weather	Temperature	Secchi	DO	Comments
A145292	A1	2024-01-01	6:34	Ebb	Clear	18	74	12.2	
B245293	B2	2024-01-02	9:34	Flood	Cloudy	25.2	15	2.5	Nearby ag field draining, water was nasty looking
C345294	C3	2024-01-03	10:43	High Slack	Rainy	12.6	110	10.2	
BL545295	BL5	2024-01-04	8:04	Low Slack	Clear	17.8	NA	9.9	Forgot secchi
Al245296	Al2	2024-01-05	13:45	Ebb	Clear	20.1	60	11.5	
FW45297	FW	2024-01-06	15:13	Ebb	Rainy	15.4	45	10.4	
H2245298	H22	2024-01-07	7:34	High Slack	Clear	13.4	67	12.1	

Information about sampling effort

TableSampleInfo

SampleID	VisitID	SampleType	GearID	Length	Width	Volume	Time	CoditionCode	Comments
1	A145292	Seine	Seine1	30	15	225	NA	1	
2	A145292	Seine	Seine1	23	6	69	NA	1	
3	A145292	Zooplankton	Zoop3	18	0.2	1.8	5	1	
4	B245293	Seine	Seine2	30	8	120	NA	2	Net got snagged
5	B245293	Zooplankton	Zoop1	20	0.2	2	5	1	
6	B245293	Fyke	Fyke1	NA	NA	NA	3600	1	

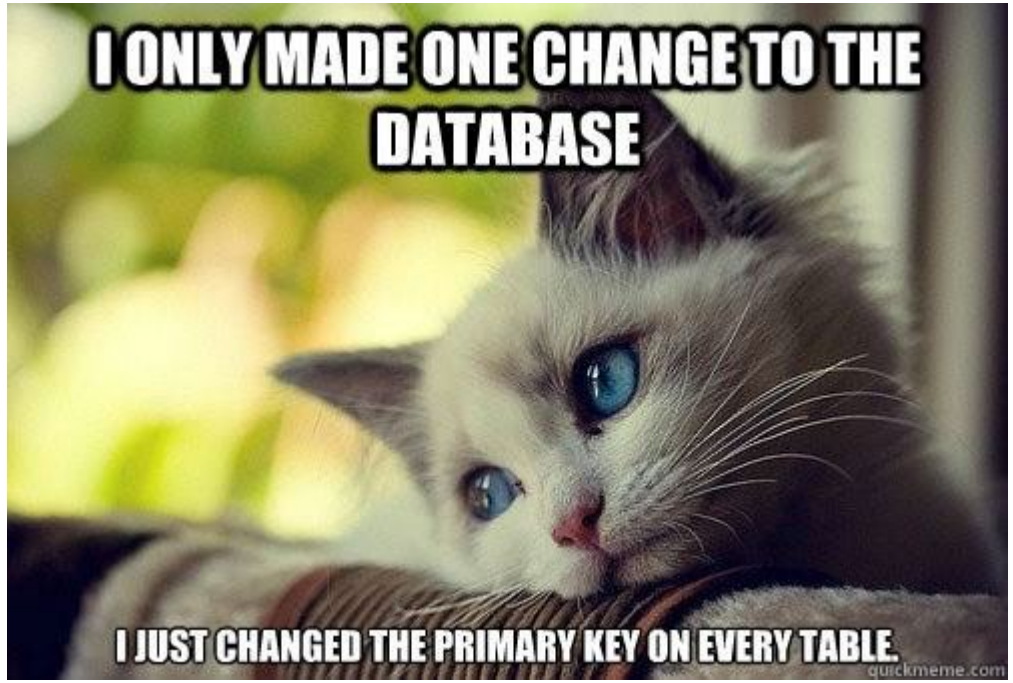
Information about each fish

TableFishInfo

FishID	SampleID	SpeciesCode	Count	ForkLength	Weigth	Genetics	FishComments
123	2	CHISAL	1	68	2.21	SR001	
124	2	CHISAL	1	57	2.56	SR002	
125	2	CHISAL	1	72	3.48	SR003	
126	2	CHISAL	1	48	1.04	SR004	
127	2	WHICAT	1	75	NA		Weird paracite on tail
128	2	WHICAT	1	112	NA		
129	2	WHICAT	1	135	NA		
130	2	WHICAT	34	NA	NA		

Keys

- Each table should have a field indicating unique values
- Auto-numbered versus informative
- Keys are used to link tables



Keys

Primary key

Foreign key

TableSiteVist

VisitID	SiteCode	Date	Time	Tide	Weather	Temperature	Secchi	DO	Comments
A145292	A1	2024-01-01	6:34	Ebb	Clear	18	74	12.2	
B245293	B2	2024-01-02	9:34	Flood	Cloudy	25.2	15	2.5	Nearby ag field draining, water was nasty looking
C345294	C3	2024-01-03	10:43	High Slack	Rainy	12.6	110	10.2	
BL545295	BL5	2024-01-04	8:04	Low Slack	Clear	17.8	NA	9.9	Forgot secchi
Al245296	Al2	2024-01-05	13:45	Ebb	Clear	20.1	60	11.5	
FW45297	FW	2024-01-06	15:13	Ebb	Rainy	15.4	45	10.4	
H2245298	H22	2024-01-07	7:34	High Slack	Clear	13.4	67	12.1	

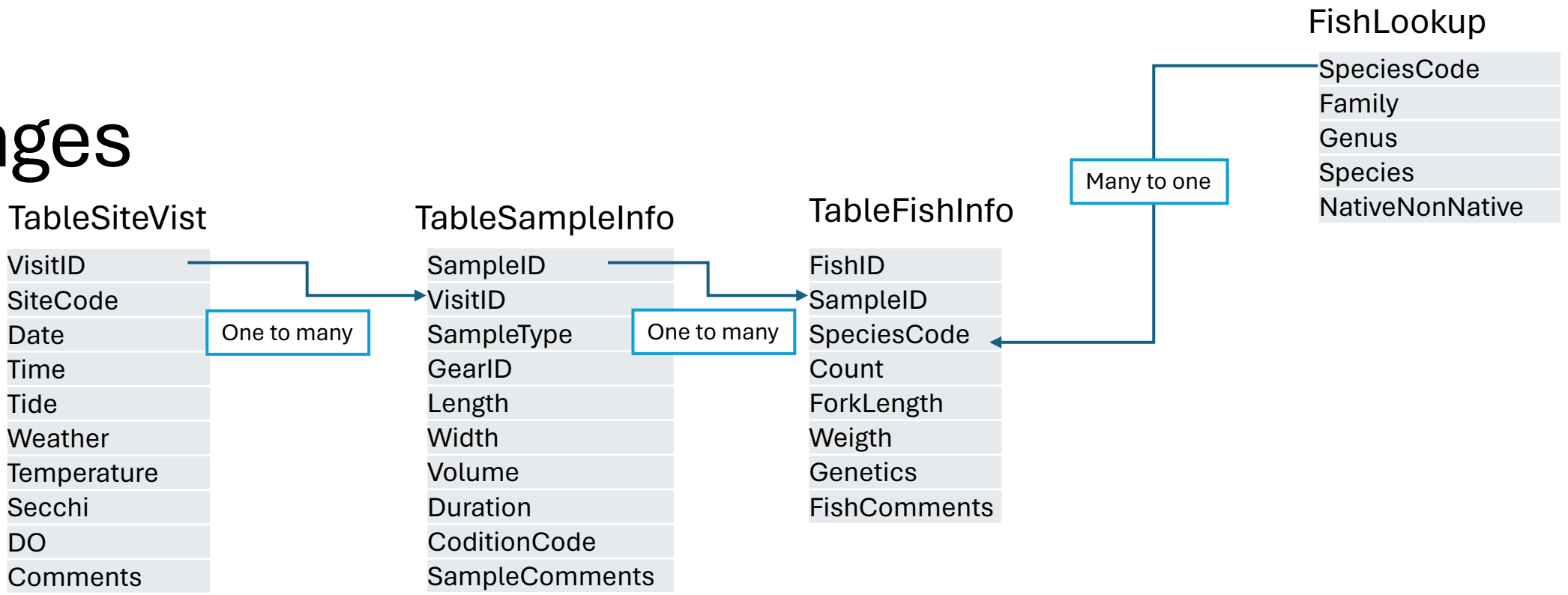
TableSampleInfo

SampleID	VisitID	SampleType	GearID	Length	Width	Volume	Time	CoditionCode	Comments
1	A145292	Seine	Seine1	30	15	225	NA	1	
2	A145292	Seine	Seine1	23	6	69	NA	1	
3	A145292	Zooplankton	Zoop3	18	0.2	1.8	5	1	
4	B245293	Seine	Seine2	30	8	120	NA	2	Net got snagged
5	B245293	Zooplankton	Zoop1	20	0.2	2	5	1	
6	B245293	Fyke	Fyke1	NA	NA	NA	3600	1	

TableFishInfo

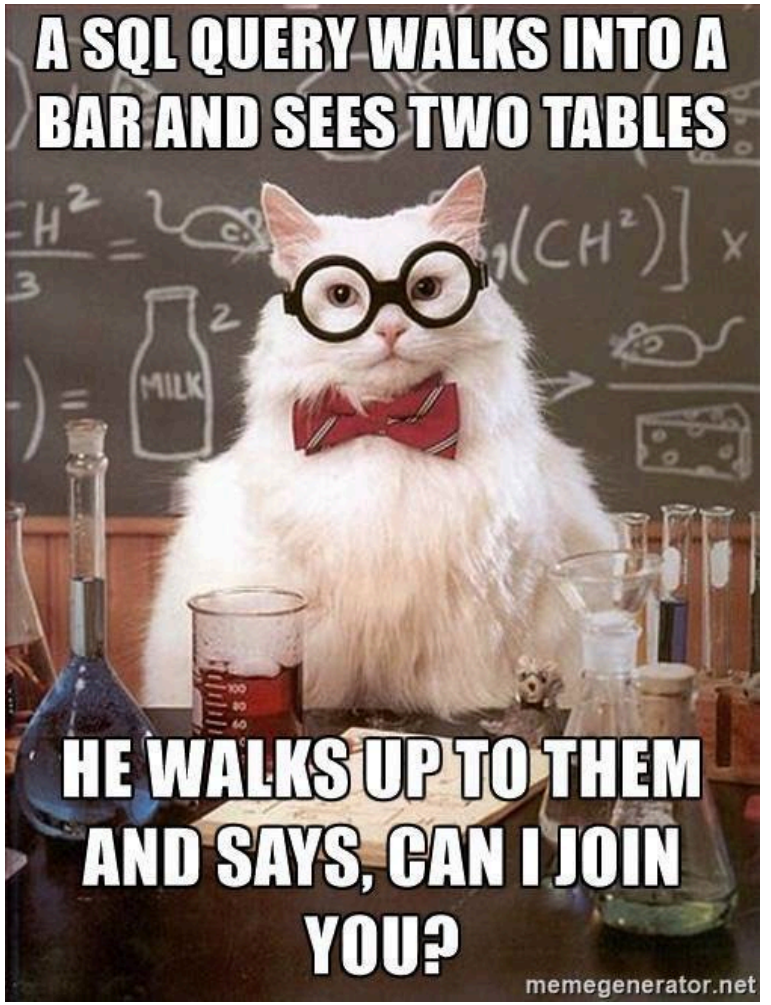
FishID	SampleID	SpeciesCode	Count	ForkLength	Weigth	Genetics	FishComments
123	2	CHISAL	1	68	2.21	SR001	
124	2	CHISAL	1	57	2.56	SR002	
125	2	CHISAL	1	72	3.48	SR003	
126	2	CHISAL	1	48	1.04	SR004	
127	2	WHICAT	1	75	NA		Weird paracite on tail
128	2	WHICAT	1	112	NA		
129	2	WHICAT	1	135	NA		
130	2	WHICAT	34	NA	NA		
131	6	CHISAL	1	46	1.12	FR002	

Linkages

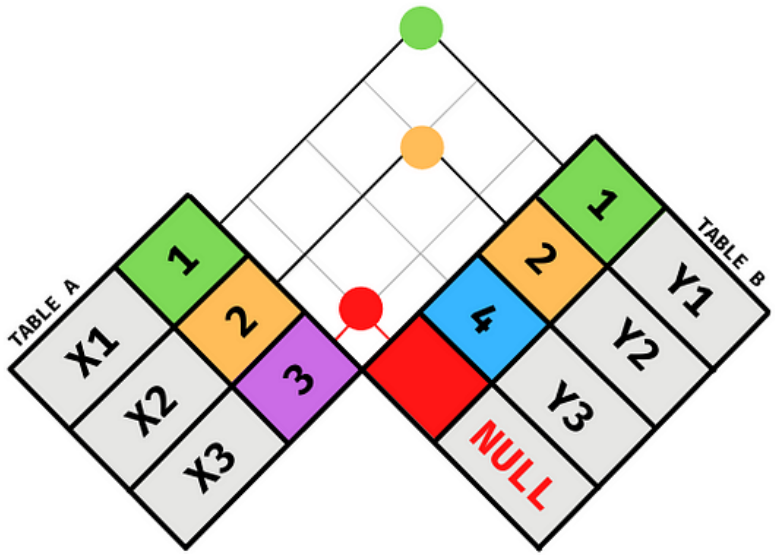

TableSiteVist
VisitID
SiteCode
Date
Time
Tide
Weather
Temperature
Secchi
DO
Comments
One to many
TableSampleInfo
SampleID
VisitID
SampleType
GearID
Length
Width
Volume
Duration
CoditionCode
SampleComments
One to many
TableFishInfo
FishID
SampleID
SpeciesCode
Count
ForkLength
Weigth
Genetics
FishComments
Many to one
FishLookup
SpeciesCode
Family
Genus
Species
NativeNonNative

Queries

- JOIN tables you want
- SELECT columns you want
- FILTER rows you want



Joins



LEFT JOIN

```
SELECT
    <SELECT LIST>
FROM      TABLE_A A
LEFT JOIN TABLE_B B
    ON A.KEY = B.KEY
```

KEY	VAL_X	VAL_Y
1	X1	Y1
2	X2	Y2
3	X3	NULL

Left Join – TableSiteVisit, TableSampleInfo

TableSiteVist

VisitID	SiteCode	Date	Time
A145292	A1	2024-01-01	6:34
B245293	B2	2024-01-02	9:34
C345294	C3	2024-01-03	10:43
BL545295	BL5	2024-01-04	8:04
Al245296	Al2	2024-01-05	13:45
FW45297	FW	2024-01-06	15:13
H2245298	H22	2024-01-07	7:34

TableSampleInfo

SampleID	VisitID	SampleType	GearID	Length	Width
1	A145292	Seine	Seine1	30	15
2	A145292	Seine	Seine1	23	6
3	A145292	Zooplankton	Zoop3	18	0.2
4	B245293	Seine	Seine2	30	8
5	B245293	Zooplankton	Zoop1	20	0.2
6	B245293	Fyke	Fyke1	NA	NA

VisitID	SiteCode	Date	Time	SampleID	SampleType	GearID	Length	Width
A145292	A1	1/1/2024	6:34	1	Seine	Seine1	30	15
A145292	A1	1/1/2024	6:34	2	Seine	Seine1	23	6
A145292	A1	1/1/2024	6:34	3	Zooplankton	Zoop3	18	0.2
B245293	B2	1/2/2024	9:34	4	Seine	Seine2	30	8
B245293	B2	1/2/2024	9:34	5	Zooplankton	Zoop1	20	0.2
B245293	B2	1/2/2024	9:34	6	Fyke	Fyke1	NA	NA
C345294	C3	1/3/2024	10:43	NA	NA	NA	NA	NA
BL545295	BL5	1/4/2024	8:04	NA	NA	NA	NA	NA
Al245296	Al2	1/5/2024	13:45	NA	NA	NA	NA	NA
FW45297	FW	1/6/2024	15:13	NA	NA	NA	NA	NA
H2245298	H22	1/7/2024	7:34	NA	NA	NA	NA	NA

Left Join – Fish Info

VisitID	SiteCode	Date	Time	SampleID	SampleType
A145292	A1	1/1/2024	6:34	1	Seine
A145292	A1	1/1/2024	6:34	2	Seine
A145292	A1	1/1/2024	6:34	3	Zooplankton
B245293	B2	1/2/2024	9:34	4	Seine
B245293	B2	1/2/2024	9:34	5	Zooplankton
B245293	B2	1/2/2024	9:34	6	Fyke
C345294	C3	1/3/2024	10:43	NA	NA
BL545295	BL5	1/4/2024	8:04	NA	NA
Al245296	Al2	1/5/2024	13:45	NA	NA
FW45297	FW	1/6/2024	15:13	NA	NA
H2245298	H22	1/7/2024	7:34	NA	NA

FishID	SampleID	SpeciesCode	Count	ForkLength
123	2	CHISAL	1	68
124	2	CHISAL	1	57
125	2	CHISAL	1	72
126	2	CHISAL	1	48
127	2	WHICAT	1	75
128	2	WHICAT	1	112
129	2	WHICAT	1	135
130	2	WHICAT	34	NA
131	6	CHISAL	1	46
132	6	MISSIL	1	35
133	6	MISSIL	1	40
134	6	MISSIL	1	42
135	6	MISSIL	123	NA
136	6	REDEAR	1	36

VisitID	SiteCode	Date	SampleID	SampleType	FishID	SpeciesCode	Count	ForkLength
A145292	A1	1/1/2024	1	Seine	NA	NA	NA	NA
A145292	A1	1/1/2024	2	Seine	123	CHISAL	1	68
A145292	A1	1/1/2024	2	Seine	124	CHISAL	1	57
A145292	A1	1/1/2024	2	Seine	125	CHISAL	1	72
A145292	A1	1/1/2024	2	Seine	126	CHISAL	1	48
A145292	A1	1/1/2024	2	Seine	127	WHICAT	1	75
A145292	A1	1/1/2024	2	Seine	128	WHICAT	1	112
A145292	A1	1/1/2024	2	Seine	129	WHICAT	1	135
A145292	A1	1/1/2024	2	Seine	130	WHICAT	34	NA
A145292	A1	1/1/2024	3	Zooplankton	NA	NA	NA	NA
B245293	B2	1/2/2024	4	Seine	NA	NA	NA	NA
B245293	B2	1/2/2024	5	Zooplankton	NA	NA	NA	NA
B245293	B2	1/2/2024	6	Fyke	131	CHISAL	1	46
B245293	B2	1/2/2024	6	Fyke	132	MISSIL	1	35
B245293	B2	1/2/2024	6	Fyke	133	MISSIL	1	40
B245293	B2	1/2/2024	6	Fyke	134	MISSIL	1	42
B245293	B2	1/2/2024	6	Fyke	135	MISSIL	123	NA
B245293	B2	1/2/2024	6	Fyke	136	REDEAR	1	36
C345294	C3	1/3/2024	NA	NA	NA	NA	NA	NA
BL545295	BL5	1/4/2024	NA	NA	NA	NA	NA	NA
Al245296	Al2	1/5/2024	NA	NA	NA	NA	NA	NA
FW45297	FW	1/6/2024	NA	NA	NA	NA	NA	NA
H2245298	H22	1/7/2024	NA	NA	NA	NA	NA	NA

Select

- Date, SampleType
  Species Code, Count

VisitID	SiteCode	Date	SampleID	SampleType	FishID	SpeciesCode	Count	ForkLength
A145292	A1	1/1/2024	1	Seine	NA	NA	NA	NA
A145292	A1	1/1/2024	2	Seine	123	CHISAL	1	68
A145292	A1	1/1/2024	2	Seine	124	CHISAL	1	57
A145292	A1	1/1/2024	2	Seine	125	CHISAL	1	72
A145292	A1	1/1/2024	2	Seine	126	CHISAL	1	48
A145292	A1	1/1/2024	2	Seine	127	WHICAT	1	75
A145292	A1	1/1/2024	2	Seine	128	WHICAT	1	112
A145292	A1	1/1/2024	2	Seine	129	WHICAT	1	135
A145292	A1	1/1/2024	2	Seine	130	WHICAT	34	NA
A145292	A1	1/1/2024	3	Zooplankton	NA	NA	NA	NA
B245293	B2	1/2/2024	4	Seine	NA	NA	NA	NA
B245293	B2	1/2/2024	5	Zooplankton	NA	NA	NA	NA
B245293	B2	1/2/2024	6	Fyke	131	CHISAL	1	46
B245293	B2	1/2/2024	6	Fyke	132	MISSIL	1	35
B245293	B2	1/2/2024	6	Fyke	133	MISSIL	1	40
B245293	B2	1/2/2024	6	Fyke	134	MISSIL	1	42
B245293	B2	1/2/2024	6	Fyke	135	MISSIL	123	NA
B245293	B2	1/2/2024	6	Fyke	136	REDEAR	1	36
C345294	C3	1/3/2024	NA	NA	NA	NA	NA	NA
BL545295	BL5	1/4/2024	NA	NA	NA	NA	NA	NA

SiteCode	Date	SampleType	SpeciesCode	Count
A1	1/1/2024	Seine	NA	NA
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	WHICAT	1
A1	1/1/2024	Seine	WHICAT	1
A1	1/1/2024	Seine	WHICAT	1
A1	1/1/2024	Seine	WHICAT	34
A1	1/1/2024	Zooplankton	NA	NA
B2	1/2/2024	Seine	NA	NA
B2	1/2/2024	Zooplankton	NA	NA
B2	1/2/2024	Fyke	CHISAL	1
B2	1/2/2024	Fyke	MISSIL	1
B2	1/2/2024	Fyke	MISSIL	1
B2	1/2/2024	Fyke	MISSIL	1
B2	1/2/2024	Fyke	MISSIL	123
B2	1/2/2024	Fyke	REDEAR	1
C3	1/3/2024	NA	NA	NA
BL5	1/4/2024	NA	NA	NA
Al2	1/5/2024	NA	NA	NA

Filter

- Filter
  - SpeciesCode = “CHISAL”

SiteCode	Date	SampleType	SpeciesCode	Count
A1	1/1/2024	Seine	NA	NA
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	WHICAT	1
A1	1/1/2024	Seine	WHICAT	1
A1	1/1/2024	Seine	WHICAT	1
A1	1/1/2024	Seine	WHICAT	34
A1	1/1/2024	Zooplankton	NA	NA
B2	1/2/2024	Seine	NA	NA
B2	1/2/2024	Zooplankton	NA	NA
B2	1/2/2024	Fyke	CHISAL	1
B2	1/2/2024	Fyke	MISSIL	1
B2	1/2/2024	Fyke	MISSIL	1
B2	1/2/2024	Fyke	MISSIL	1
B2	1/2/2024	Fyke	MISSIL	123
B2	1/2/2024	Fyke	REDEAR	1
C3	1/3/2024	NA	NA	NA
BL5	1/4/2024	NA	NA	NA

SiteCode	Date	SampleType	SpeciesCode	Count
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
A1	1/1/2024	Seine	CHISAL	1
B2	1/2/2024	Fyke	CHISAL	1

Naming conventions

- Keep column names short but informative
- Avoid spaces or special characters
- Avoid having the same column names in multiple tables if they mean different things
- Definitely don’t have multiple columns in the same table with the same name!

Naming conventions

- Good names
  - SampleDate
  - Distance_m
  - Biovolume
  - LabComments

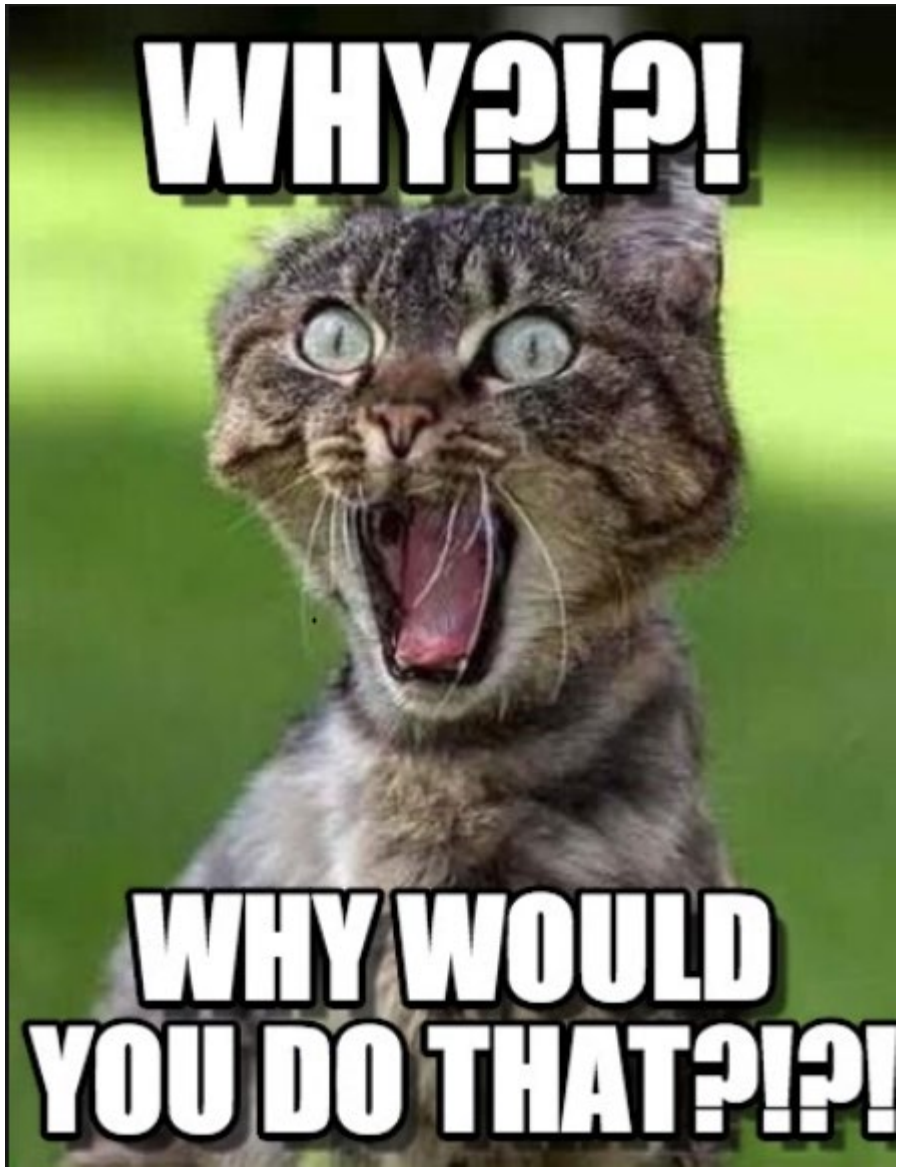
  IEP naming conventions

  https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=197029

- Bad Names
  - Date
  - Distance (m^2)
  - Bio_Volume_msquared_per_Liter_from_zooplankton
  - com

Missing Values

- Use comment field to say why value is missing
- Use a single, consistent indicator for missing values
  -
  - NA
  - -9999



Comments

- The comment field is for WHY, not WHAT.
- If you find yourself making the same comment frequently, put it in it’s on column.

- Good Comments
  - Secchi was broken
  - Fish was missing its tail
  - Lots of cow poop in the water may have been causing the low DO
- Bad Comments
  - Dead fish
  - Genetic sample ID 0001
  - Dang, it’s freezing out today.

But How?

- Microsoft Access
- SQL Server
- ESRI products
- Custom-built databases
- ~~Lots of .csv's~~

Further reading

- https://r4ds.had.co.nz/tidy-data.html
- https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=203207&inline

Questions?