6- Cleaning Real Data

Questions

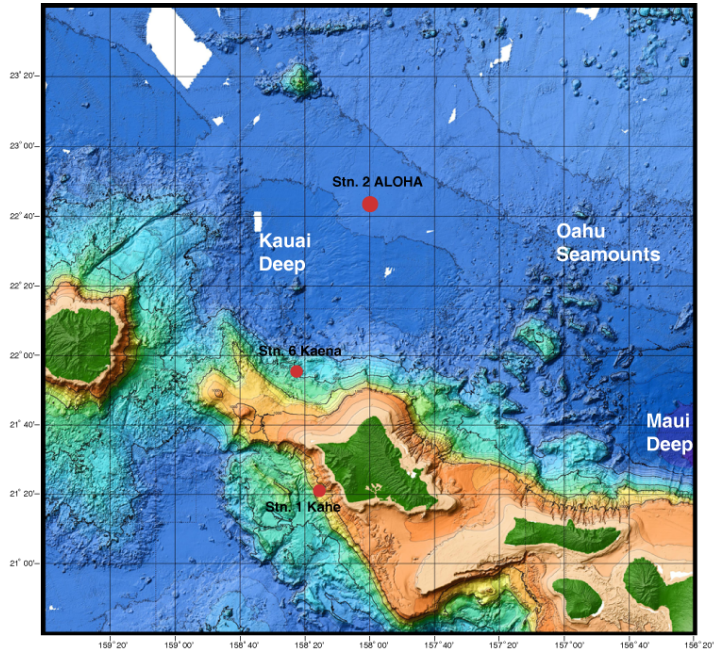
- How do you clean an example dataset?
- How do you deal with missing data?
- How do you fix column type mismatches?

Objectives

- Clean an example dataset using previously described concepts and some new ones.

Cleaning Real Data: HOT Data

- Data from the Hawaiian Ocean Time-Series (HOT).
    - Data collected since 1988 from station ALOHA located just North of Oahu.



Cleaning Real Data: HOT Data

- Some of the environmental variables recorded are:

Column name	Environmental Variable It Represents
botid #	Bottle ID
date mmddyy	Date
press dbar	Pressure
....	...

- The dataset contains over 20,000 samples.
- Dataset needs to be cleaned up.
    - We will focus on simple yet useful operations to illustrate the process.

In [5]:
```
pd.read_csv("./data/hot_dogs_data.csv", nrows=5)
```

Out[5]:

	botid #	date mmddyy	press dbar	temp ITS-90	csal PSS-78	coxy umol/kg	ph	phos umol/kg	umo
0	2190200124	30910	5.5	23.0629	35.2514	214.1	-9	0.10	
1	2190200123	30910	59.6	23.0670	35.2506	214.6	-9	0.11	
2	2190200122	30910	90.7	21.7697	35.1897	213.4	-9	0.12	
3	2190200121	30910	119.4	20.7957	35.1666	208.5	-9	0.15	
4	2190200120	30910	153.6	20.1517	35.2192	204.6	-9	0.15	

Data Issues

- There are a lot of -9 values
- Empty column
- Date is incorrectly inferred (number instead of mmddyy)
- Index should be "botid #" instead of the automatically generated integer value.
- Column order inconvenient: Should start with `botid #` `date`, `ph`, etc...
- How can we fix these issues?

Exercises

- Complete the 4 exercises in Notebook 6.
    - We can answer individual question if any.