Data Cleaning

January 29, 2026

Today’s plan

1 **Lab 1: Data Cleaning**

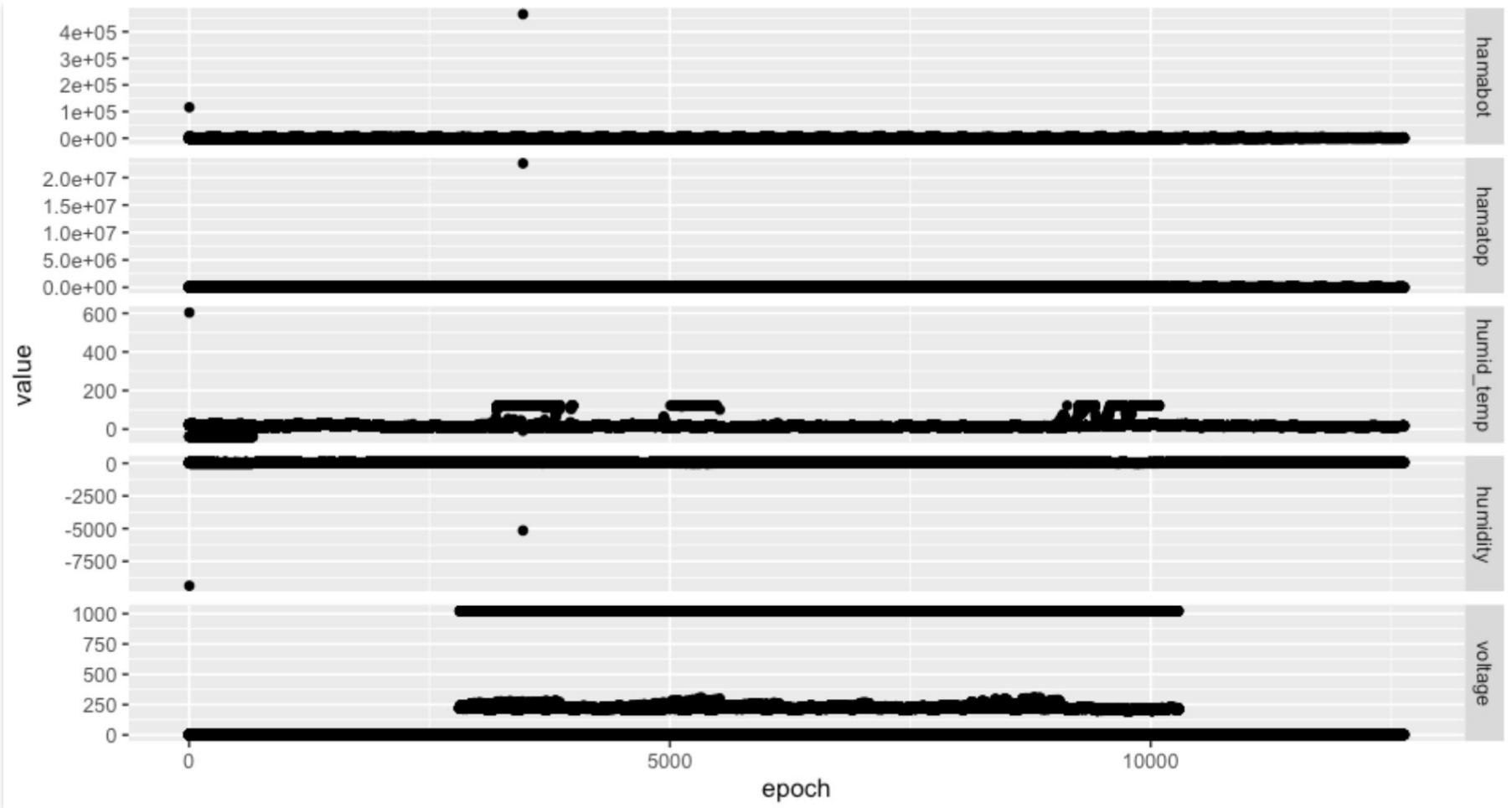
Data Cleaning

Continuation of Last time

1. **Load in the data**
   a. Epoch/dates and redwood datasets have already been filled out for you.
   b. You need to fill out the `load_mote_location_data()` in the load.R/load.py file.
      - The output of this function should be a data frame (or tibble) with 80 rows and 5 columns (column names: "ID", "Height", "Direc", "Dist", "Tree")

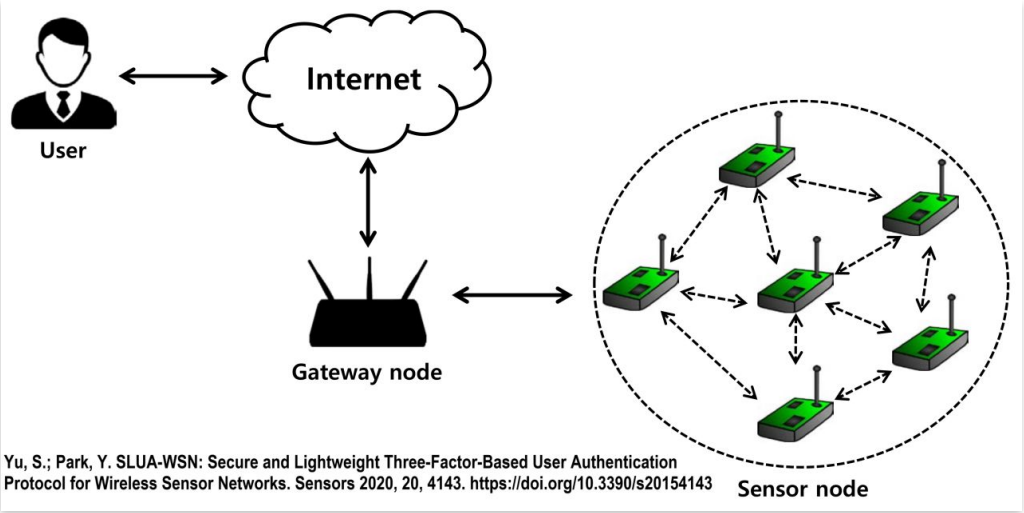
Before proceeding, read all available documentation!

2. Look and "play" around with the data in order to:
   a. Try to **identify as many issues or oddities** with the data as you can. *Hint:* there are many!!
   b. Also **think** about how you might address these issues and clean the data. Jot down these ideas, but no need to take the time to implement it *yet*.
      - *Time permitting*: start implementing your ideas, but prioritize identifying the issues over fixing them.


hamabot
hamatop
humid_temp
humidity
voltage
value
epoch
4e+05
3e+05
2e+05
1e+05
0e+00
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
600
400
200
0
-2500
-5000
-7500
1000
750
500
250
0
5000
10000

What to do about the redwood "all", "log", and "net" datasets?

- Initial reaction: look at the "all" dataset
- **Q: What is the difference between the "all", "log", and "net" datasets?**
  - *Hint:* Related to the data collection process



How the data was collected should inform how we clean/preprocess the data

Data cleaning to be continued...

The data cleaning journey is to be continued as you work through lab 1 on your own.

Remember: there is more than one right way to clean the dataset

Some of your to-dos:

- Merge log and network data (don't forget to remove duplicate copies)
- Deal with voltages (e.g., convert them to the same scale/unit)
- Identify and remove erroneous measurements or "outliers"
- Other data cleaning steps that you think are appropriate – you have free reign!
  - There are more issues than what has been discussed in class

Document your data cleaning steps and explain **why** you chose to do it that way

- This includes documenting and justifying data cleaning steps that we've done in class
- If you don't like how we cleaned it in class, that's also totally fine. Document and justify.

Recap + Next Time

Recap

- **Data cleaning** is a highly iterative process.
- My two cents:
  - Don’t be afraid to ask lots of questions. Better to ask than to assume (more likely than not, incorrectly)
  - Read all documentation

Next Time

- Exploratory data analysis
  [chapter 5 from VDS textbook]