Knowing Your Data

Machine learning and data science are all about the data. If data quality is poor, even the most sophisticated analysis would generate only lackluster (乏善可陳) results.

Example (A tale). To further illustrate the importance of these issues, consider the following hypothetical situation. You as a data scientist receive an email from a medical researcher concerning a project that you are eager to work on.

> Hi,
> I've attached the data file that I mentioned in my previous email. Each line contains the information for a single patient and consists of five fields. We want to predict the last field using the other fields. I don't have time to provide any more information about the data since I'm going out of town for a couple of days, but hopefully that won't slow you down too much. And if you don't mind, could we meet when I get back to discuss your preliminary results? I might invite a few other members of my team.
>
> Thanks and see you in a couple of days.

Despite some ambiguities, you proceed to analyze the data. The first few rows of the file are as follows:

```
012  232  33.5  0   10.7
020  121  16.9  2  210.1
027  165  24.0  0  427.6
⋮
```

A brief look at the data reveals *nothing strange*. You put your doubts aside and start the analysis. There are only 1000 lines, a smaller data file than you had hoped for, but two days later, you feel that you have made some progress. You arrive for the meeting, and while waiting for others to arrive, you strike up a conversation with a researcher who is also working on the project. When she learns that you have also been analyzing the data from the project, she asks if you would mind giving her a brief overview of your results.

> **Researcher**: So, you got the data for all the patients?
> **Data Miner**: Yes. I haven't had much time for analysis, but I do have a few interesting results.
> **Researcher**: Amazing. There were so many data issues with this set of patients that I couldn't do much.
> **Data Miner**: Oh? I didn't hear about any possible problems.

Researcher: Well, first there is field 5, the variable we want to predict. It's common knowledge among people who analyze this type of data that results are better if you work with the log of the values, but I didn't discover this until later. Was it mentioned to you?

Data Miner: No.

Researcher: But surely you heard about what happened to field 4? It's supposed to be measured on a scale from 1 to 10, with 0 indicating a missing value, but because of a data entry error, all 10's were changed into 0's. Unfortunately, since some of the patients have missing values for this field, it's impossible to say whether a 0 in this field is a real 0 or a 10. Quite a few of the records have that problem.

Data Miner: Interesting. Were there any other problems?

Researcher: Yes, fields 2 and 3 are basically the same, but I assume that you probably noticed that.

Data Miner: Yes, but these fields were only weak predictors of field 5.

Researcher: Anyway, given all those problems, I'm surprised you were able to accomplish anything.

Data Miner: True, but my results are really quite good. Field 1 is a very strong predictor of field 5. I'm surprised that this wasn't noticed before.

Researcher: What? Field 1 is just an identification number.

Data Miner: Nonetheless, my results speak for themselves.

Researcher: Oh, no! I just remembered. We assigned ID numbers after we sorted the records based on field 5. There is a strong connection, but it's **meaningless**. Sorry.

Although this scenario represents an extreme situation, it emphasizes the importance of "**knowing your data**."