Useful Data

Data is pretty topical these days. Everybody, including I’m sure, some automata, are talking about big data and AI. I wanted to make a couple of very general observations.

If you have a lot of data but they all look the same, or are pretty close to each other, then the information the data contain is very little. In statistics, we require that when more and more data are added to the dataset, the variability of the data as a whole must grow, not diminish. Otherwise the explanatory or predictive power of our model is weakened. It’s a technical point but one that is illuminating when the data fail to meet the standard. Inferences begin to vary wildly as new data is added.

If we have multiple variables or factors in our model, then these factors must be sufficiently independent of one another. If we have 2 factors which are somehow related, then they can be replaced by one factor. If they are not, then the model again behaves very erratically.

The art in the science is discerning how diverse the data are while remaining relevant. Diverse but irrelevant data are as useless as concentrated and correlated data.

This very narrow phenomenon in statistics can be generalized. If you observe someone only when they are hungry you might conclude that they are a glutton.