SOFTWARE

FROM BIG DATA TO GOOD DATA

22.12.2022

I think the AI Community is slowly getting at the point where epidemiologists had already been two decades ago, so Ng:

“In many industries where giant data sets simply don’t exist, I think the focus has to shift from big data to good data. Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn.”
—Andrew Ng, CEO Landing AI

or Bickson: Large Image Datasets Today Are a Mess

“We were surprised to find that there are 1.2M pairs of identical images in ImageNet-21K. Most of them are exact duplicates which add no information to the data but waste on storage and compute. In addition 104,000 train/val leaks were identified by comparing similar images across the train and validation subsets.”
—Danny Bickson, CEO Visual Layer