Stanford eCorner

DIY Datasets

27-02-2019

URL: https://stvp.stanford.edu/clips/diy-datasets

Sophisticated algorithms only achieve useful results when combined with robust data. Verge Genomics founder and CEO Alice Zhang describes how her team built large datasets to amplify and validate the company’s computational work.

Transcript

- Machine learning in tech is fundamentally different from machine learning in biology because most of human biology is still unknown, that's to say, there's a huge missing data problem.. And we encountered this problem when we actually were looking at the datasets available, and we saw that these datasets were vastly underpowered and poorly designed and could not be used for machine learning.. And what's important here is that the sophistication of your algorithms is irrelevant if you don't have enough data from which to train and to learn.. So we embarked on this two-year internal data generation initiative where we now generate all of our own proprietary patient datasets.. And we do this by partnering with over a dozen different hospitals, brain banks, and universities to actually source thousands of patient brains after they've died, and we sequence that internally to create now one of the largest patient training datasets in the world for ALS and Parkinson's disease.. But while training data is really important for generating predictions, for every AI company you need to have validation data.. So we take all of these predictions, and we actually test them in our biology labs to create a huge body of validation data that feeds back into the algorithms and constantly improves them over time and across diseases as well.. And this critical for any company in AI and biology because AI cannot be a black box in biology.. You need to have a way to actually see if the predictions work and to let those predictions guide the improvement of the actual algorithms themselves...