Twitter Thread by Santiago



Santiago
@svpino

Machine learning education is broken.

If you are preparing for a research position, you are good. If you are looking to get out there and start solving problems, not even close.

Here are some thoughts so you can get ahead.

↓

Most classes, courses, and books cover the same road.

They start with a dataset. They finish with a working model. The focus is always on everything that happens in between.

Dataset → Model.

This is great, but not enough.

Real-life situations rarely start with a dataset, and they never end after you finish building your model.

Applying machine learning successfully is hard.

Here are a few examples that you should keep in mind.

First challenge: Properly framing up the problem.

If you don't understand the problem, you can't determine what data you need. If you don't understand the data, you can't build a good model.

I've never seen a company that had their data ready to go.

In fact, most of them don't even have data at all and need you to determine what exactly they should start collecting.

You usually have to go Problem → Potential Solution → Data.

Another fun challenge: Getting the data from the point of origin to a place where you can start using it.

Who's putting together the pipeline to move the data? Who's building processes to clean it and get it ready?

Talk about deploying models, and people roll their eyes.

I've talked to a lot of data scientists that have no idea how to get this done. They struggle to look past a Jupyter notebook, and they have to learn on the job.

Another gap:

Everyone is laser-focused on building models that solve problems, but almost nobody looks at them to help other people do their job better.

Combining models with humans unlocks a lot of value.

There are many more gaps that we need to fill.

From drift monitoring to automatic retraining processes, bias mitigation, and everything in between.

Heck, even camera selection before building a deep learning model for computer vision is a common issue!

A few programs out there are starting to adventure outside of the "Dataset → Model" approach.

I hope more join the party.

The more we cover, the better we can tackle the problems waiting for us.

If you are getting ready to go out there, going above and beyond to close these gaps represents an incredible opportunity.

Show up with a broader understanding of what it takes to get the work done, and companies will throw a ridiculous amount of money at you.

If you are looking for more information about machine learning in the real world, follow me @svpino, and I'll give you something to think about every week.

We can do this together. One tweet at a time.

You don't make up the data. You collect it.

It's not uncommon to start working on a project and be 1-2 years away from having the necessary data to solve their problem.

Plant the seed today, so you can harvest it when ready.

https://t.co/6WaSNwNfJa

> How should we approach if there is no data available? if you 'make' data then it may not represent actual production data, and model also may not behave as expected.
>
> — Sumedh Patil (@sumedh_bp) May 5, 2021