Data Labeling: Creating Ground Truth

Astasia Myers   [Follow]

Apr 15, 2019 · 4 min read

Last year we published our ML workflow landscape. One category we've seen continued interest in is data labeling, the process of attaching meaning to different types of digital data like audio files, text, images, and videos. Our research suggests themes in the data labeling segment include: 1) data is the new oil, 2) dark data is valuable, 3) deep learning algorithms are a driver, 4) hand labeling can be expensive, and 5) automation is important. The data labeling market is over $1.5B today and expected to grow at a double-digit CAGR over the next five years. We categorize +30 offerings that represent different approaches. We are excited about innovation in the space and look forward to speaking with startups offering data labeling solutions.

Over the past few months we spoke to dozens of data scientists and ML engineers about data labeling. Our conversations unearthed five themes:

1) **Data is the new oil.** Data is a foundational component of training models. It is a day one concern. Andrew Ng stated data is the rocket fuel needed to power the ML rocket ship.

2) **Dark data is valuable but largely inaccessible.** Dark data like scientific articles, government reports, medical images, etc. are hard

Discover Medium

Welcome to a place where words matter. On Medium, smart voices and original ideas take center stage - with no ads in sight. Watch

Make Medium yours

Follow all the topics you care about, and we'll deliver the best stories for you to your homepage and inbox. Explore

Become a member

Get unlimited access to the best stories on Medium — and support writers while you're at it. Just $5/month. Upgrade

About   Help   Legal