Data Labeling That You Can Feel Good About

dataengineeringpodcast.com/cloudfactory-data-labeling-episode-89/

Summary

Successful machine learning and artificial intelligence projects require large volumes of data that is properly labelled. The challenge is that most data is not clean and well annotated, requiring a scalable data labeling process. Ideally this process can be done using the tools and systems that already power your analytics, rather than sending data into a black box. In this episode Mark Sears, CEO of CloudFactory, explains how he and his team built a platform that provides valuable service to businesses and meaningful work to developing nations. He shares the lessons learned in the early years of growing the business, the strategies that have allowed them to scale and train their workforce, and the benefits of working within their customer's existing platforms. He also shares some valuable insights into the current state of the art for machine learning in the real world.

Announcements

- Hello and welcome to the Data Engineering Podcast, the show about modern data management
- When you're ready to build your next pipeline, or want to test out the projects you hear about on the show, you'll need somewhere to deploy it, so check out our friends at Linode. With 200Gbit private networking, scalable shared block storage, and a 40Gbit public network, you've got everything you need to run a fast, reliable, and bullet-proof data platform. If you need global distribution, they've got that covered too with world-wide datacenters including new ones in Toronto and Mumbai. And for your machine learning workloads, they just announced dedicated CPU instances. Go to dataengineeringpodcast.com/linode today to get a $20 credit and launch a new server in under a minute. And don't forget to thank them for their continued support of this show!
- Integrating data across the enterprise has been around for decades – so have the techniques to do it. But, a new way of integrating data and improving streams has evolved. By integrating each silo independently – data is able to integrate without any direct relation. At CluedIn they call it "eventual connectivity". If you want to learn more on how to deliver fast access to your data across the enterprise leveraging this new method, and the technologies that make it possible, get a demo or presentation of the CluedIn Data Hub by visiting dataengineeringpodcast.com/cluedin. And don't forget to thank them for supporting the show!
- You listen to this show to learn and stay up to date with what's happening in databases, streaming platforms, big data, and everything else you need to know about modern data management.For even more opportunities to meet, listen, and learn from your peers you don't want to miss out on this year's conference season. We have partnered with organizations such as O'Reilly Media, Dataversity, and the Open Data Science Conference. Coming up this fall is the combined events of Graphorum and the Data Architecture Summit. The agendas have been announced and super early bird registration for up to $300 off is available until July 26th, with early bird pricing for up to $200 off through August 30th. Use the code BNLLC to get an additional 10% off any pass when you register. Go to dataengineeringpodcast.com/conferences to learn more and take advantage of our partner discounts when you register.
- Go to dataengineeringpodcast.com to subscribe to the show, sign up for the mailing list, read the show notes, and get in touch.
- To help other people find the show please leave a review on iTunes and tell your friends and co-workers
- Join the community in the new Zulip chat workspace at dataengineeringpodcast.com/chat
- Your host is Tobias Macey and today I'm interviewing Mark Sears about Cloud Factory, masters of the art and science of labeling data for Machine Learning and more

Interview

- Introduction
- How did you get involved in the area of data management?
- Can you start by explaining what CloudFactory is and the story behind it?
- What are some of the common requirements for feature extraction and data labelling that your customers contact you for?
- What integration points do you provide to your customers and what is your strategy for ensuring broad compatibility with their existing tools and workflows?
- Can you describe the workflow for a sample request from a customer, how that fans out to your cloud workers, and the interface or platform that they are working with to deliver the labelled data?
    - What protocols do you have in place to ensure data quality and identify potential sources of bias?
- What role do humans play in the lifecycle for AI and ML projects?
- I understand that you provide skills development and community building for your cloud workers. Can you talk through your relationship with those employees and how that relates to your business goals?
    - How do you manage and plan for elasticity in customer needs given the workforce requirements that you are dealing with?
- Can you share some stories of cloud workers who have benefited from their experience working with your company?
- What are some of the assumptions that you made early in the founding of your business which have been challenged or updated in the process of building and scaling CloudFactory?
- What have been some of the most interesting/unexpected ways that you have seen customers using your platform?
- What lessons have you learned in the process of building and growing CloudFactory that were most interesting/unexpected/useful?
- What are your thoughts on the future of work as AI and other digital technologies continue to disrupt existing industries and jobs?
    - How does that tie into your plans for CloudFactory in the medium to long term?