**MEDIUM**

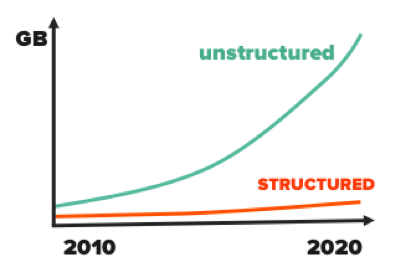
[Emily Watkins](https://medium.com/@emwatz?source=post_header_lockup)

March 27, 2018

# ****The time has come for ETL 2.0****

For decades, we’ve become comfortable with getting structured data ready to analyze. We’ve perfected the ETL processes for databases, and so we take it for granted that our relational data can be moved into data warehouses and used for business intelligence.

Unfortunately, that manageable structured data represents a smaller and smaller percent of enterprise data. **Unstructured data** is often at the heart of today’s companies — log files of user interaction, product images, audio call logs.



And there is no ETL for unstructured data.

We need to get images, audio, and logs into a format ready for analysis, but those unstructured file types are not interpretable by computer. The status quo for Extract/Transform/Load of unstructured data is: human effort. Humans interpret the content **one file at a time**, adding labels to give each file structure.

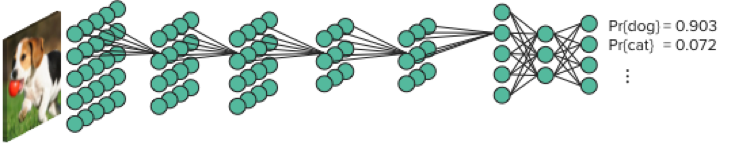
Sometimes the solution is in-house labeling — e.g. when hospitals change the clinical workflows so physicians draw a bounding box around each tumor in pathology images. Sometimes the solution is to outsource labeling — e.g. paying strangers on Amazon Mechanical Turk to describe content.

Neither of these solutions is efficient. They don’t scale well, they’re costly, and — worst of all — they’re prone to human error, resulting in less accurate downstream analysis. (A dataset of 10M videos, AMT could cost $3.3M and 78 person years to label.)

This madness has got to stop. We need a **modern ETL process** for unstructured data.

And what’s more modern (read: buzzword-y) than neural networks? :)

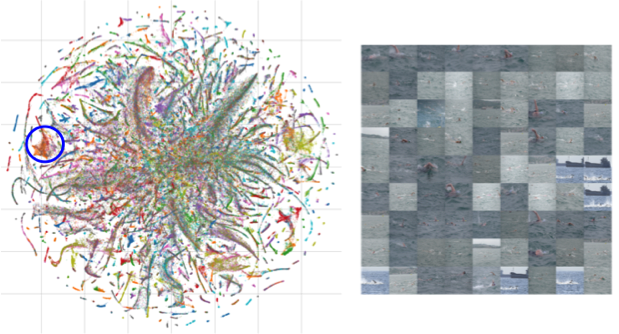
Under the hood, classification neural networks actually do a great job transforming unstructured data into structured data on their way to identifying content.



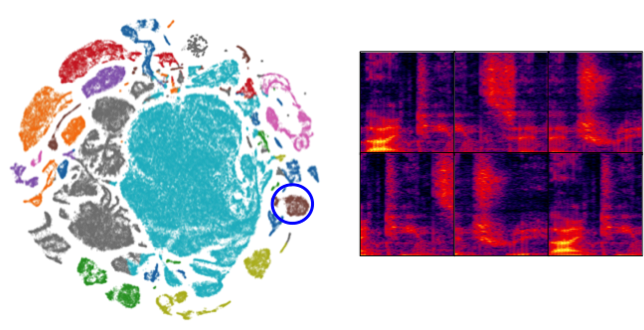
We should stop just before prediction of content and save off the structured representation of the original content.

If you create a database containing all these structured, vector-based image representations, you can start to search, explore, and analyze the dataset contents much more systematically. Instead of looking at a petabyte-scale dataset through the pinhole view of 1 image at a time, run deep-learning-based inference against the dataset once, then query the entire dataset at once.

“Do I have any videos of swimmers?”



“Do any of my Customer Support reps curse at customers?”



Using this AI-powered indexing method, you can also apply ETL 2.0 during data ingest for real-time anomaly detection and alerting.

“Have I ever seen this pattern before?”

“Does this video contain any people that shouldn’t be here?”

With an indexed version of their unstructured data, teams can get start to utilize and depend on their images, audio, video, and log data as much as they depend on their transactional databases today.

“Deep-learning based data curation is the indispensable tool I never knew I needed.” — Caroline Rowland, CEO of [New Moon](http://new-moon.co.uk/)

If you’re interested in learning more about unlocking your unstructured data with ETL 2.0, Ben Taylor ([Ziff.ai](http://ziff.ai/)) and I ([Pure Storage](https://blog.purestorage.com/industries/big-data-analytics/)) are at [GTC](https://www.nvidia.com/en-us/gtc/) all week talking about it. We have a live demo at Booth #509 and we’ll walk through some customer deployment examples during our presentation: [Eliminating Manual Data Labeling with AI-powered Data Curation](https://2018gputechconf.smarteventscloud.com/connect/sessionDetail.ww?SESSION_ID=161764) (3/28/18 @ 4–5 PM Marriott Ballroom 2).

Of course, we’d love to hear about any other ETL 2.0 solutions that you’re working on.