cloudburst

a website to aggregate data and redistribute it to people who hope to find modern solutions to urgent challenges

relevance

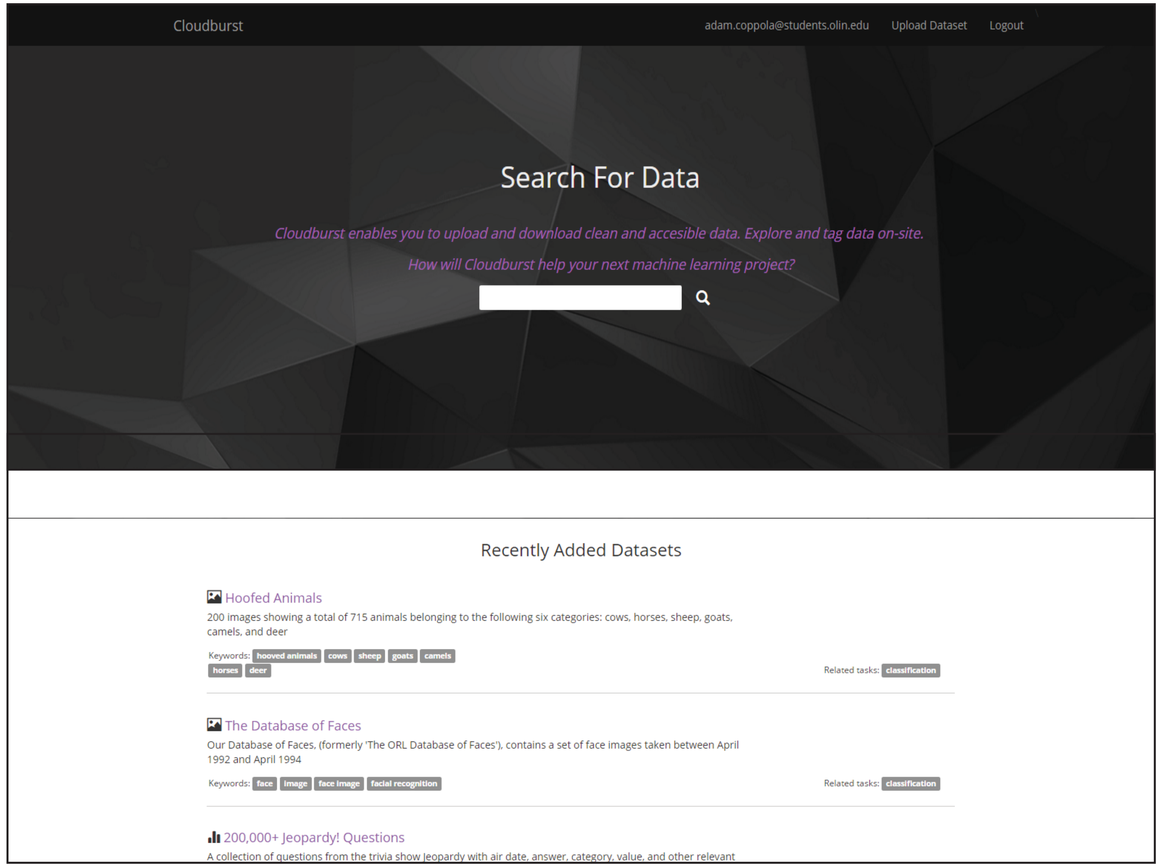
Relevance means something different to every ML user. Most begin their ML projects with a context and question, like “how can we better understand our customers to make money?” or “how can we reduce traffic buildup along the Mass Pike?”
These questions construct a need for a user. Their hope is to find a dataset that will address their questions, in order to make a predictive model. The ultimate value is in predictive power, but data is valuable because it enables this entire process.
Unfortunately for many users, their dreams are big but their options are limited. This is because our users’ questions ask modern questions, but corresponding novel datasets are hard to come by. Cloudburst’s sharing service hopes to break down these barriers for success.

quality

Data quality revolves around the form it takes.
On the surface, a dataset must be shaped in order to reflect the question being asked. For image data, this will be simple, because most questions will be about images. For text data, there exist various tagging and formatting schemas.
Poor online data formatting and organization was among the top concerns of our user group.
Beneath the surface, data quality is defined by what our users call “internal structure.” This simply means that the data represents some real phenomenon, because it follows some laws that may be learned by an ML model. While often harder to track, internal struture can often be judged by the extent to which a dataset has been useful in the past.

accessibility

Some ML users get data handed to them by their employer. These are not the users we want to consder. Instead, Cloudburst consider the users who do not have data creation abilities in the first place. Previously, these users may have had to search dozens of data websites in hopes of finding something to address their question. Then, even if they found a dataset that looks promising, they would have no way of understanding the internal structure apart from downloading it and working with it. This often leads to wasted time and useless downloads.
Cloudburst hopes to open up the data exploration process by putting it in-site. Cloudburst provides displays of the data and the associated readme, and as of May 2016, methods are in development for testing the data in-browswer.



“If the data is relevant such that the hypothesis, once arrived at, has power in some shape or form, that makes for a good dataset.”
ML user

advisor: John Geddes
liaison: Slater Victoroff | Diana Yuan
Adam Coppola | Deborah Hellen | Mitchell Kwock | Allison Patterson | Emily Tumang

“Computing hardware used to be a capital asset, while data wasn’t thought of in the same way. Now, hardware is becoming a service people buy in real time, and *data is becoming the asset.*”

erik brynjolfsson, *The Rise of Data Capital*

indico