AI is Everywhere专场

Data in AI : Pain Points and Opportunities

堵俊平/Junping Du 2023年05月27日

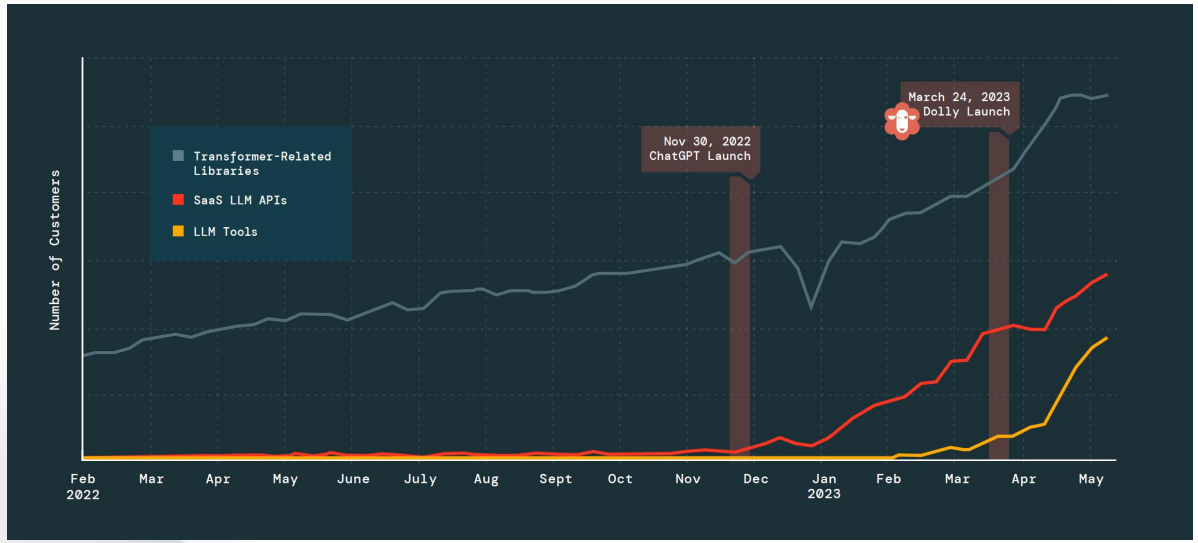
In Time We Live – “Big Bang” of AI

Buzzword:

Machine Learning, Deep Learning, AI Framework, Transformer, AIGC, LLM, AGI, etc.

Refer: LifeArchitect.ai/models

What We Focus – from Model Training to Model Eco-system



The number of companies using SaaS LLM APIs (used to access services like ChatGPT) has grown 1310% between the end of November 2022 and the beginning of May 2023

NLP accounts for 49% of daily Python data science library usage, making it the most popular application

Organizations are putting substantially more models into production (411% YoY growth) while also increasing their ML experimentation (54% YoY growth)

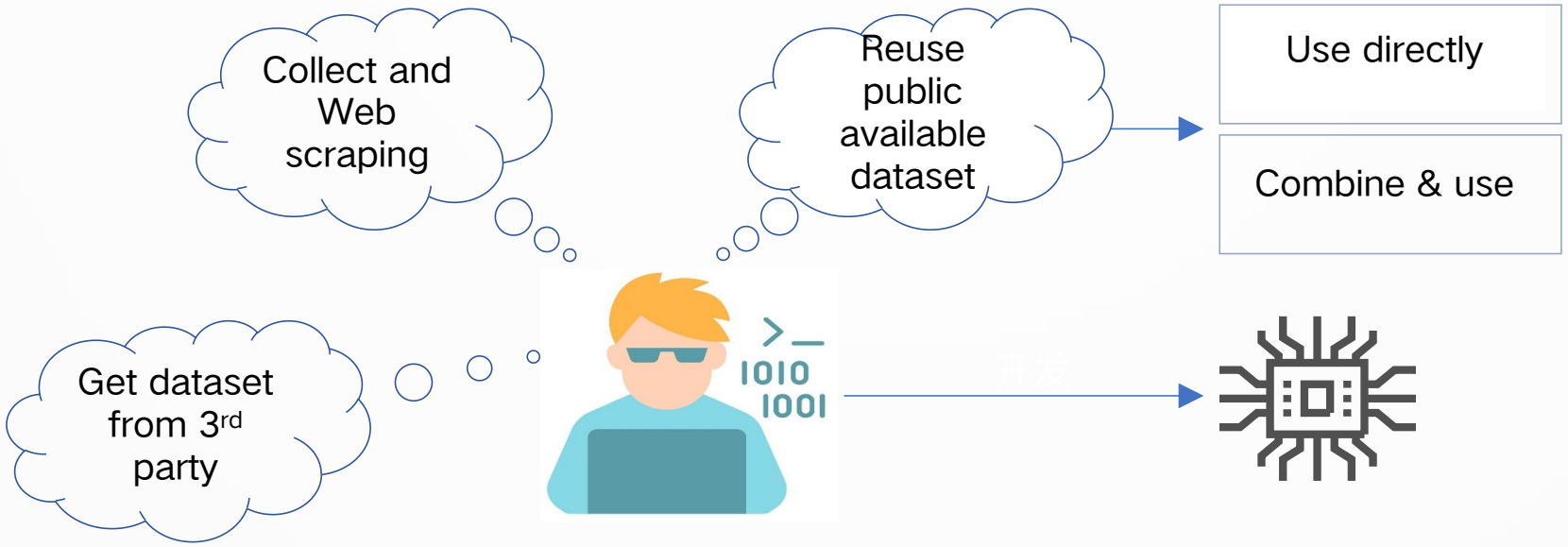
Organizations are getting more efficient with ML; for every three experimental models, roughly one is put into production, compared to five experimental models a year prior

Refer: 2023 state of Data + AI by Databricks

Data is underwater

Data is never be enough, until Model EOL

Data is NOT free, cost in whole life cycle



Data Dilemma: quantity or quality?

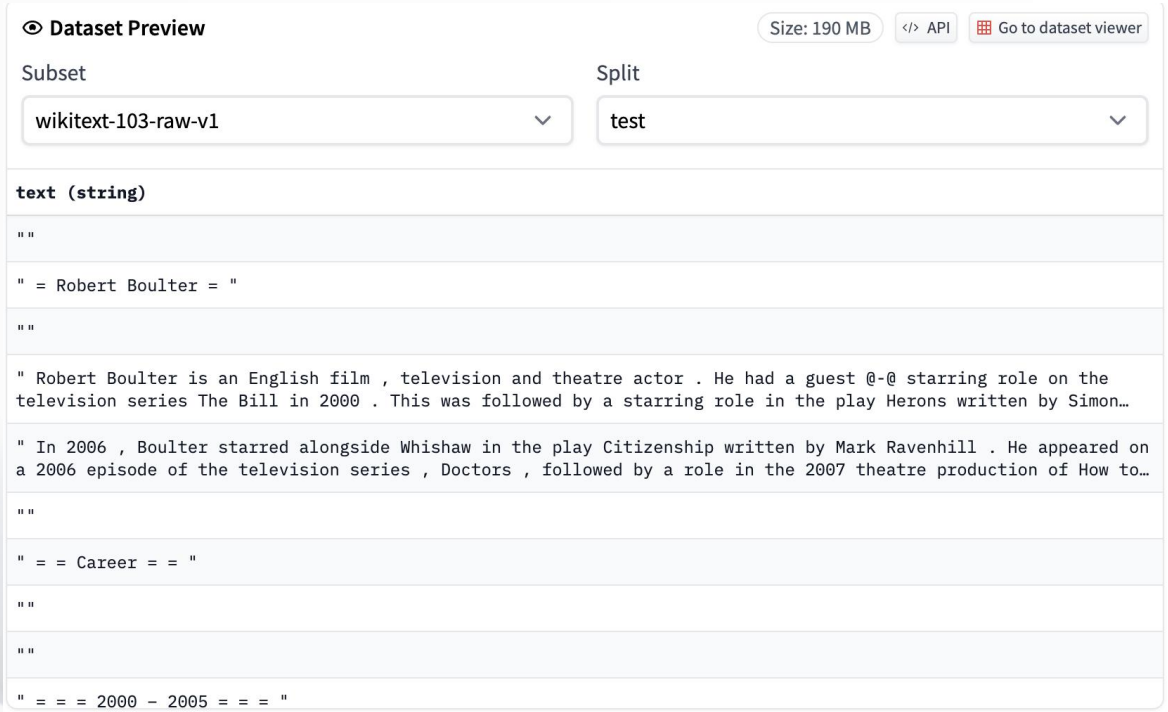
Quantity of data	Quality of data
Generalization	Accuracy
Robustness	Relevance
Rare occurrences	Bias and fairness

Ideal: to have large quantity of diverse, high-quality data

Reality: priority one over the other

Dataset discovery – metadata is important… but NO standard?

Data is not easy get discovered and usage without metadata standard
More data silos caused by organizational silos



Dataset Preview

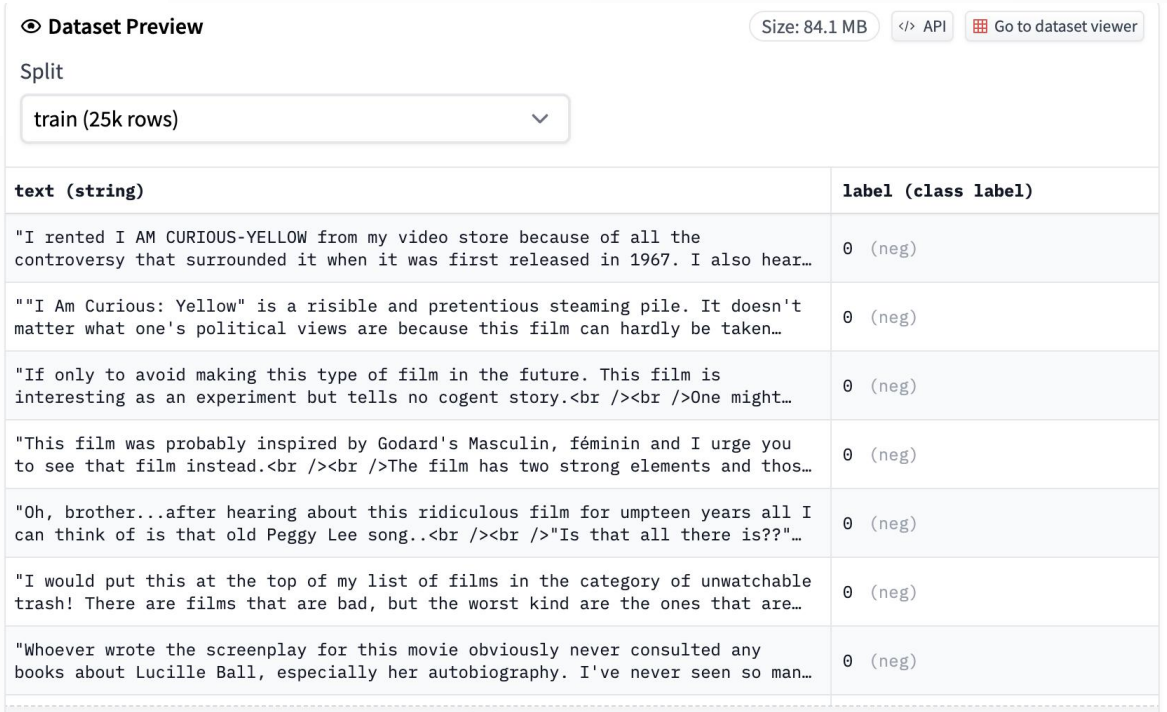
Size: 190 MB | API | Go to dataset viewer

Subset: wikitext-103-raw-v1

Split: test

text (string)
""
" = Robert Boulter = "
""
" Robert Boulter is an English film , television and theatre actor . He had a guest @-@ starring role on the television series The Bill in 2000 . This was followed by a starring role in the play Herons written by Simon…
" In 2006 , Boulter starred alongside Whishaw in the play Citizenship written by Mark Ravenhill . He appeared on a 2006 episode of the television series , Doctors , followed by a role in the 2007 theatre production of How to…
""
" = = Career = = "
""
""
" = = = 2000 – 2005 = = = "

https://huggingface.co/datasets/wikitext



Dataset Preview

Size: 84.1 MB | API | Go to dataset viewer

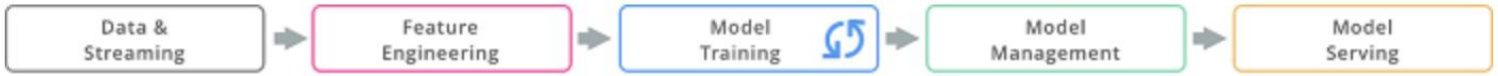
Split: train (25k rows)

text (string)	label (class label)
"I rented I AM CURIOUS-YELLOW from my video store because of all the controversy that surrounded it when it was first released in 1967. I also hear…	0 (neg)
""I Am Curious: Yellow" is a risible and pretentious steaming pile. It doesn't matter what one's political views are because this film can hardly be taken…	0 (neg)
"If only to avoid making this type of film in the future. This film is interesting as an experiment but tells no cogent story. One might…	0 (neg)
"This film was probably inspired by Godard's Masculin, féminin and I urge you to see that film instead. The film has two strong elements and thos…	0 (neg)
"Oh, brother...after hearing about this ridiculous film for umpteen years all I can think of is that old Peggy Lee song.. "Is that all there is??"…	0 (neg)
"I would put this at the top of my list of films in the category of unwatchable trash! There are films that are bad, but the worst kind are the ones that are…	0 (neg)
"Whoever wrote the screenplay for this movie obviously never consulted any books about Lucille Ball, especially her autobiography. I've never seen so man…	0 (neg)

https://huggingface.co/datasets/imdb

Governance, from Data to Model

Governance is required for datasets, features, models, etc.



How can one identify all models derived from a particular dataset?

Which features and model function have been used as input?

What was the training, test, and validation performance?

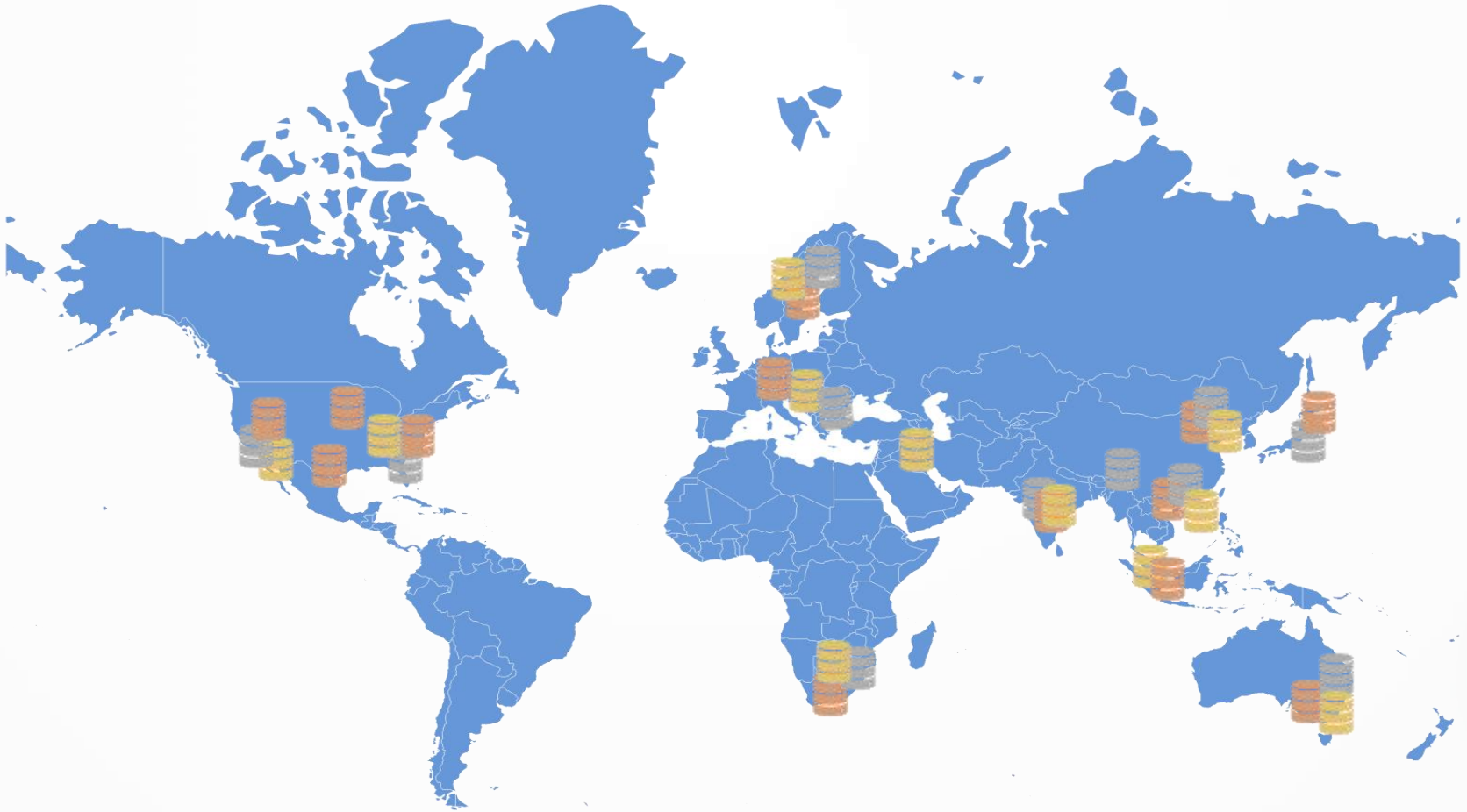
What type and amount of resources was required to train?

What model has been created and where has it been persisted?

…

New silos for data and model

Multi-Region, Cloud, Data Source, etc.



GOTC

DON'T PANIC

on Data & AI

datastrato.ai

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE