Datamule

Making it easy to clean and enrich data at scale, starting with the Securities and Exchanges Commission.

Problem

- Valuable data is stored within unstructured text in file formats such as html, pdf, and xml.
- Post 2023, companies began using LLMs to convert messy data into usable data.
- This is expensive and does not scale.
- LLM costs are downstream of chip costs, which are downstream of fabs.
- Fabs take time and considerable capital expenditure to set up
- Scaling must be done using alternative methods.

Solution

- Algorithms
- Classical Machine Learning
- Surplus GPUs

Example: Algorithms

Customer wants to convert every company's annual report from 2005 into a cleaned dictionary representation. This is about 200,000 reports, totalling 20 million pages and 150 billion tokens.

- Commercial option (unstructured.io): $600,000.
- Standard LLM option (GPT-5): $200,000.
- Best case LLM option (Gemini 2.5 Flash Lite Mini) - $30,000.
- Our algorithmic solution (doc2dict) - less than $1.

Example: Classical Machine Learning + Surplus GPUs

Customer wants to extract the full name of every person mentioned within the text of the reports to create a network for fraud analysis. Ballpark 20 billion tokens.

- Google Cloud Natural Language Pricing: $21,745.
- AWS Comprehend: $21,750.
- Naive LLM cost (Gemini 2.5 Mini Flash Lite): $3,000.
- Our multi stage classical machine learning pipeline running on 85% spot AWS instances: $5.

Strategy

- The Securities and Exchanges Commission (SEC) corpus contains unstructured data in many file formats.
- Data derived from the SEC corpus is in high demand.
- Use the SEC Corpus as our starting point and niche.
- Start by selling data, transition into selling the pipeline.
- Charge 10-100x cost, while still being 100x or more cheaper than the competition.

Market Size

- Value of SEC data and derived products is in the tens of billions per year.
- Value of data cleaning and enrichment is unclear.
- Depends on growth of tech sector.

Competitive Advantage

- Many orders of magnitude cheaper.
- Classical ML easy to scale by spinning up new instances. Provided LLMs such as GPT-5 require connections or spend to raise rate limit.
- We control when we swap out ML models. Provided LLMs are constantly tweaked, requiring prompts and pipelines to be changed.
- Existing Open source Ecosystem.
- Ease of use.

Existing Open Source Ecosystem

- We have developed an open source (MIT License) ecosystem centered around document parsing and the SEC corpus.
- Many companies are building their products and services off of our ecosystem.
- This gives us distribution and consumer insights.

Ease of use

- We’ve used AWS and Google Cloud.
- They’re great.
- They’re also punishing for beginners, who frequently run up unexpected large cloud costs.
- In an era of vibe coding, we want to protect our users from making these mistakes.
- We make it easy to control spend.
- We provide a well-tested python interface to make it easy to focus on the good stuff—getting your product or service out.

Traction

- 100 paying customers for existing API.
- 230,000 downloads for our main open source package, datamule-python.
- 100s of research projects at universities on six continents use our ecosystem.
- 4-6 companies per month asking about enterprise support.

Datamule Cloud V2

- Existing cloud built to support researchers on a tight budget.
- We are now adding more offerings to support enterprise.
- We are also revamping our pricing system to enable us to offer enterprise support.

See: Datamule Cloud V2.

Team

John Friedman

- PhD at UCLA.
- Built data pipelines at Berkeley and MIT for economic research.
- Really likes data.