Data Starved? Let's Feed the Models

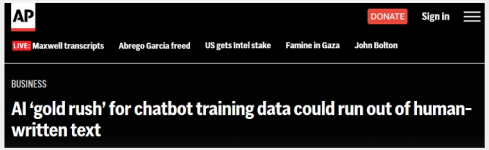
Introducing the next data frontier for AI development

Jessica Li Gebert | Data Strategy Consultant, Neudata
consulting@neudata.co | substack: @jessicalg | linkedin: /jessicajlg

September 17, 2025

Neudata® | THE AI CONFERENCE

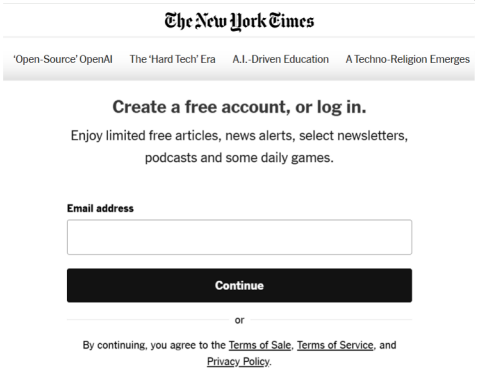
“We’re running out of data”

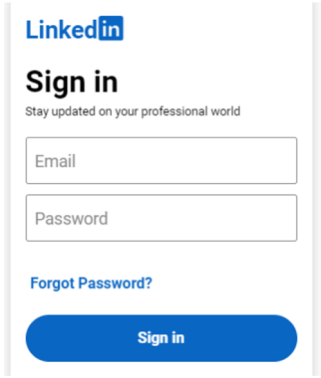


What happened to our traditional data sources?

Web-scraping

- Increasing anti-bot measures
- Increasing log-ins
- Internet *'tollgates'*





Public data / data commons

- Limited coverage in terms of domain-specific data, modalities, and languages

Published books

- Early legal decisions on copyright and fair use, pending appeals and formal legal guidance

Lawsuit	Thomson Reuters and West Publishing v. ROSS Intelligence	Bartz v. Anthropic	Kadrey v. Meta
Status	2020 - present, ongoing • Partial summary judgement granted Feb 11, 2025 • ROSS appealing as of Apr 2025	Aug 2024 - present, ongoing • Summary judgment on fair use granted Jun 23, 2025 • Piracy cause is still open	Jul 2023 - present, ongoing • Partial summary judgement Jun 25, 2025
Court	District Court, D. Delaware (Judge Stephanos Bibas)	District Court, N.D. California (Judge William Haskell Alsup)	District Court, N.D. California (Judge Vince Girdhari Chhabria)
Interim outcome	In favour of Thomson Reuters; Rejected ROSS fair use argument	In favour of Anthropic on fair use argument	In favour of Meta on fair use argument

CLOUDFLARE
Pay per crawl

ProRata.ai

TOLLBIT

For full analysis on AI copyright lawsuits, see substack: @jessicalg

Data acquisition paths forward

Synthetic data

- Not real-world human-generated data, constrains generalization of models
- Error propagation
- Limited market traction

Data partnership with content owners

- ✓ **Avoid protracted, costly 'webscraper – anti-webscraper war'**
- ✓ **Tested and trusted:** emerging data acquisition model used by foundation model makers

The next frontier: enterprise data

- ✓ **Diverse, new, human-generated data produced as part of an organization's operations**

Direct data licensing to reduce web-scraping

- Content owners/publishers are **ready and willing** to collaborate
  - 52 major AI data licensing deals since Nov 2022
  - 38 unique data providers, 12 unique model makers
  - Excluding Chinese AI market

Recent AI data licensing deals*

#	Licensee (Buyer)	Licensor (Seller)	Content type	Data modality	Content language(s)	Deal year	Data use case(s)	Deal size
1	Amazon	Shutterstock	Visual	Image, video	NA	2024	Model training	Est. total contract value: $25m - $50m
2	Apple	Shutterstock	Visual	Image, video	NA	2024	Model training	Est. total contract value: up to $50m
3	Google	AP	News media	Text	English	2025	Information retrieval and attribution	Undisclosed
4	Google	Reddit	User generated	Text, image, video (multimodal)	Multilingual	2024	Model training	Total contract value: $203m expected from AI data licensing, allegedly from Google deal. Reddit's IPO prospectus (2024) indicated $66.4m expected revenue recognition from the Google deal in 2024.
5	LG	Shutterstock	Visual	Image, video	NA	2023	Model training	Undisclosed
6	Meta	Reuters	News media	Text	English	2024	Information retrieval and attribution	Undisclosed
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

OpenAI Google shutterstock

AXIOS reddit Taylor & Francis Group an informa business

For full list of deals and more on AI data, see substack: @jessicalg

Case study 1: Trustpilot

Problem: open business model allow website users free access to all reviews without log-in, resulting in excessive unauthorized web-scraping

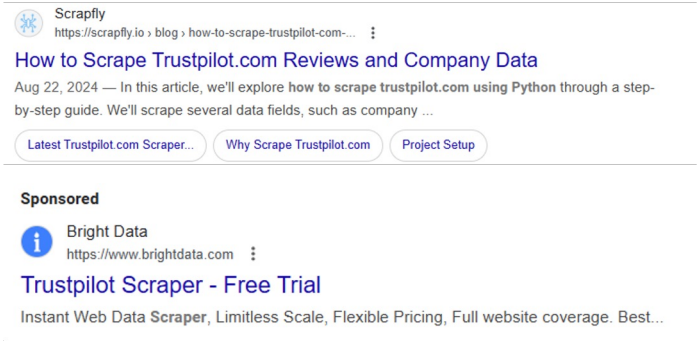
Our solution for Trustpilot:

If you can't stop them, join them

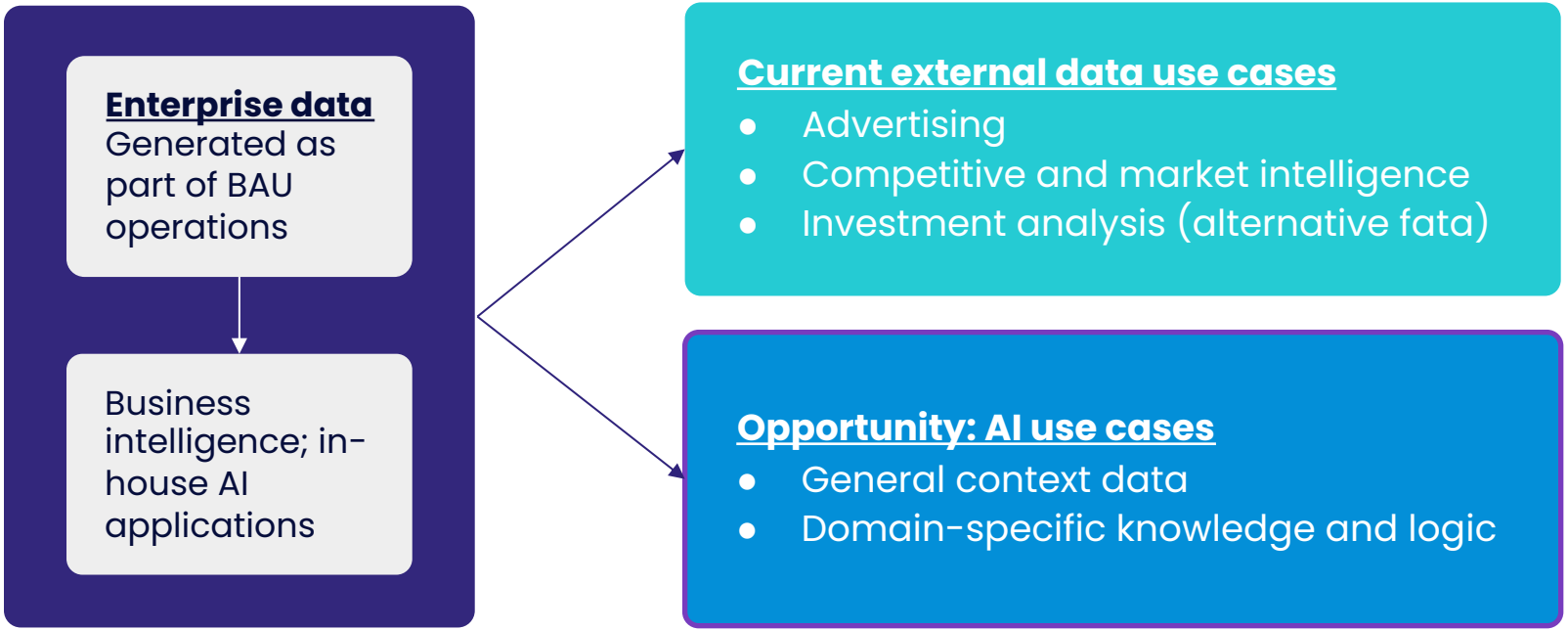
- Technical measures to minimize and deter bots
- Commercial collaboration - directly license reviews data to reduce the need for web scraping

Trustpilot

Trustpilot is a global review platform hosting over 150 million consumer reviews for more than 800,000 businesses, helping build trust and transparency across markets.



Enterprise data: the next data frontier for AI



Companies across various industries

Case study 2: global market research firm

Data value lies in content and medium

Core Business	Data Productization & Risk Mitigation	AI Use Cases
• Market research on international development, security, and consumer behaviours in developing countries in local languages • **Data assets:** voice recordings of research interviews in 100 low-resources languages and dialects	• GDPR-compliant or equivalent collection and disclosures • Personal identifier information completely removed • Truncate or modify (e.g. Pitch shift) voice recordings • Transcripts	★ Enrich multi-lingual LLMs, machine translation, ASR, small language models ★ Native voice models, multi-modal models ★ Sentiment analysis with cultural context; security, development and consumer behavior contexts ★ Voice commons

Case study 3: healthcare software provider

Domain-specific logics for vertical AI models.

Core Business

- Infrastructure and software solutions for digital communications (e.g. EHR, digital prescription) within the healthcare sector, including healthcare providers, pharmacies, PBMs.
- **Data assets:** Patient journey, prescription and fill of >90% of US population

Data Productization & Risk Mitigation

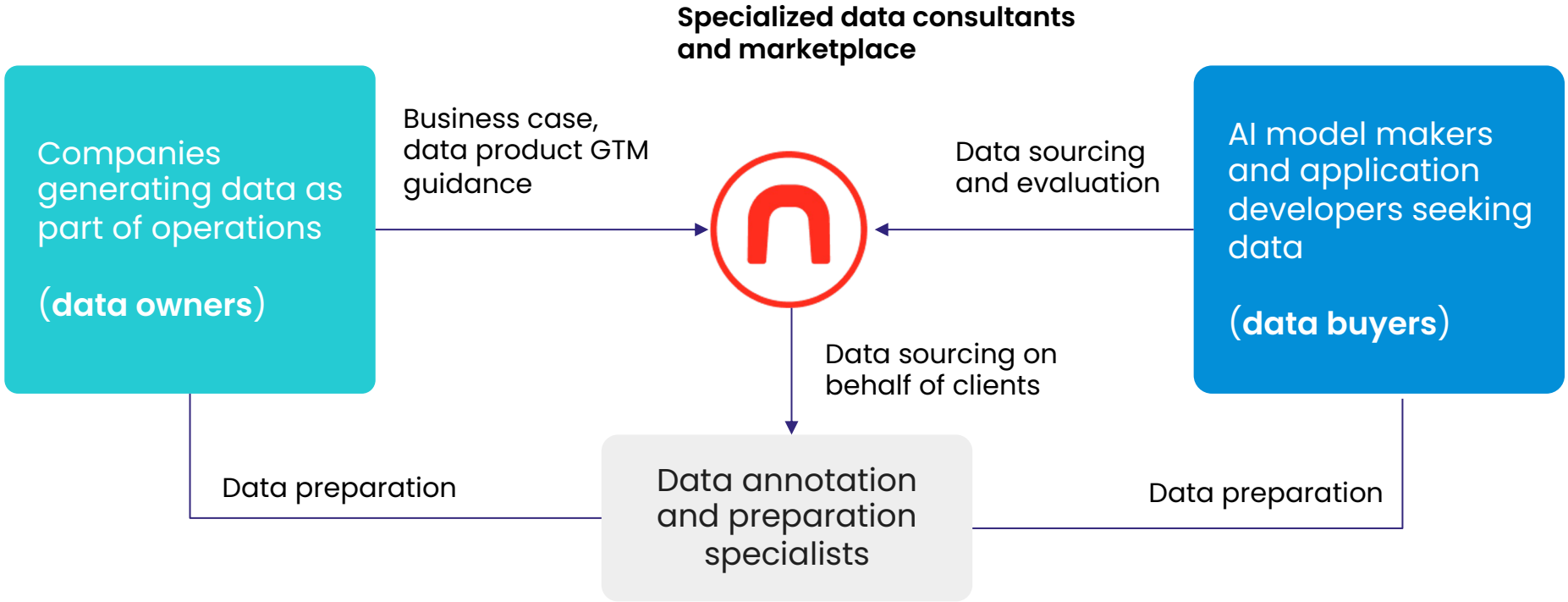
- HIPAA-compliant collection
- Personal identifier information completely removed
- Data aggregation and generalization (reduce data precision)
- Work with the right partners to plan and execute data GTM safely and compliantly

AI Use Cases

★ Healthcare function workflow and logic for healthtech AI

★ Digital healthcare companion, patient engagement AI tools

Accessing the next data frontier



Where do you fit in the enterprise data frontier? Get in touch below.

Thank You!

Jessica Li Gebert | Data Strategy Consultant, Neudata
substack: @jessicalg
consulting@neudata.co
Linkedin

Neudata

THE AI CONFERENCE