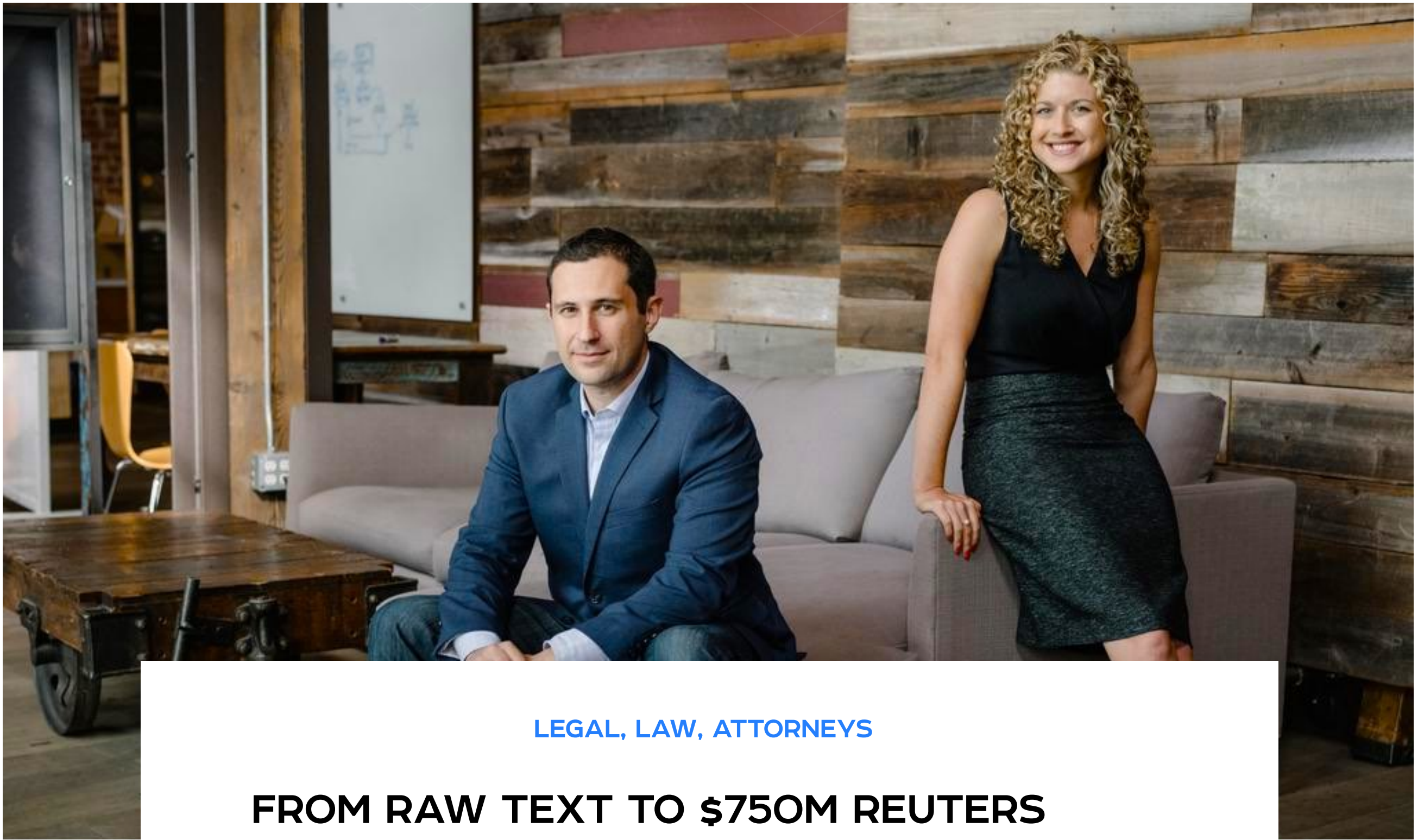




DATAOX

DATA THAT LEADS



LEGAL, LAW, ATTORNEYS

**FROM RAW TEXT TO \$750M REUTERS
ACQUISITION: COMPREHENSIVE LEGAL DATA
PROCESSING FOR AI INNOVATION**



750M

acquisition by
Thomson Reuters

50

all US states
covered

ML/AI-READY

structured data
pipeline



**CLIENT
LOCATION**

United States



**COMPANY
SIZE**

51-200



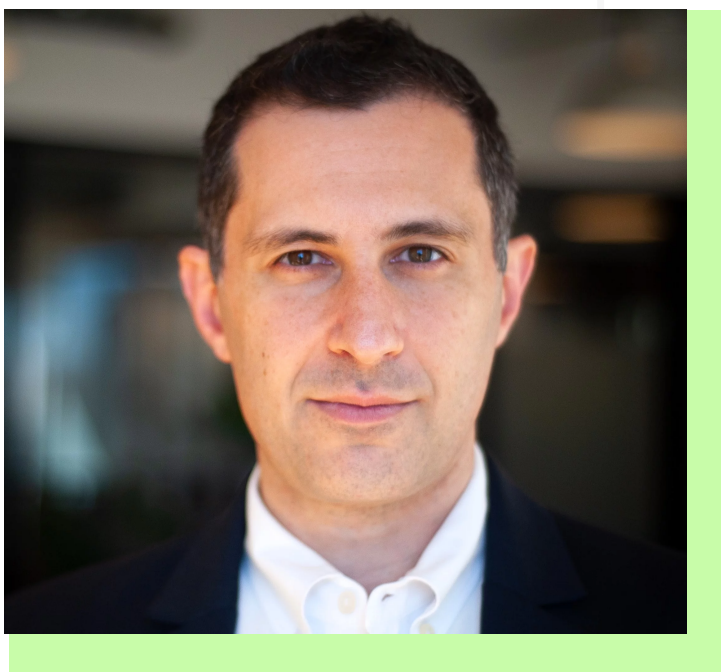
**PARTNERSHIP
DURATION**

up to 3 years

CHALLENGE

CaseText, a pioneering legal technology company, was developing an advanced AI engine that required massive amounts of structured legal data for training.

The challenge was enormous: they needed to scrape, parse, and structure all statutes and case law from all 50 U.S. states, transforming raw legal texts without machine-readable structure into a unified, well-defined format suitable for machine learning applications.



JAKE HELLER

Founder & CEO

CLIENT QUOTE

Absolutely amazing experience working with this team. Their developers are on-par with some of the best in Silicon Valley. They know their craft. They are fantastic communicators, and they care about doing good work. I would (and will) hire again.

SERVICES DELIVERED

SOLUTION

DATA SERVICES

- Data delivery
- Web scraping
- Data enrichment

```
obj.1): "translation"
obj.1): "protector"
obj.1): "verifier"
obj.1): "followers_count"
obj.1): "friends_count"
obj.1): "listed_count"
obj.1): "favourites_count"
obj.1): "statuses_count"
obj.1): "created_at"
obj.1): "utc_offset"
obj.1): "time_zone"
obj.1): "geo_enabled"
obj.1): "lang": "en"
```

TECHNOLOGIES

CORE STACK

BACKEND



Python



Java

DATA PROCESSING

ETL
Parsers
Data Structuring

TEXT PROCESSING

Regular Expressions
Text Mining

DATA EXTRACTION & AUTOMATION

WEB SCRAPING

Custom scrapers
for legal document sources

DOCUMENT PROCESSING

Multi-format parsing
for statutes and case law

ML PIPELINE

Structured data delivery
for AI training

RESULTS & IMPACT

PLATFORM SCALE & PERFORMANCE

- All 50 US states - complete statutory coverage achieved
- Raw text transformation into machine-readable, unified format
- Multi-source integration - expanded beyond statutes to additional legal information sources
- ML/AI-ready dataset - structured data pipeline feeding CaseText's internal engine

OPERATIONAL TRANSFORMATION

Complete legal data infrastructure supporting CaseText's AI capabilities

Ongoing scraper maintenance ensuring data freshness and system reliability

Consistent data structuring enabling advanced legal research and AI training

STRATEGIC PARTNERSHIP VALUE

\$750M USD acquisition by Thomson Reuters - validating the data-driven value

Up to 3 years of continuous collaboration with ongoing maintenance

Foundation for legal AI innovation - powering CaseText's research platform

Reliable technical infrastructure supporting legal technology advancement

Continuous system enhancement adapting to new legal information sources



hello@data-ox.com

Book a 30-min call

NEED HELP WITH DATA SERVICES?
GET IN TOUCH NOW