PostgresML

Why we put GPUs in the database

Postgres Conf 2024


AI isn't the future,
it’s here.

IT'S BEEN A LONG TIME COMING

Web 1990s — **(Big) Data** 2000s — **ML** 2010s — **AI** Today

AI is table stakes

Successful companies will incorporate AI into their core UX, if they haven't already

Personalization

Real-time personalization, knowledge based chat

Search

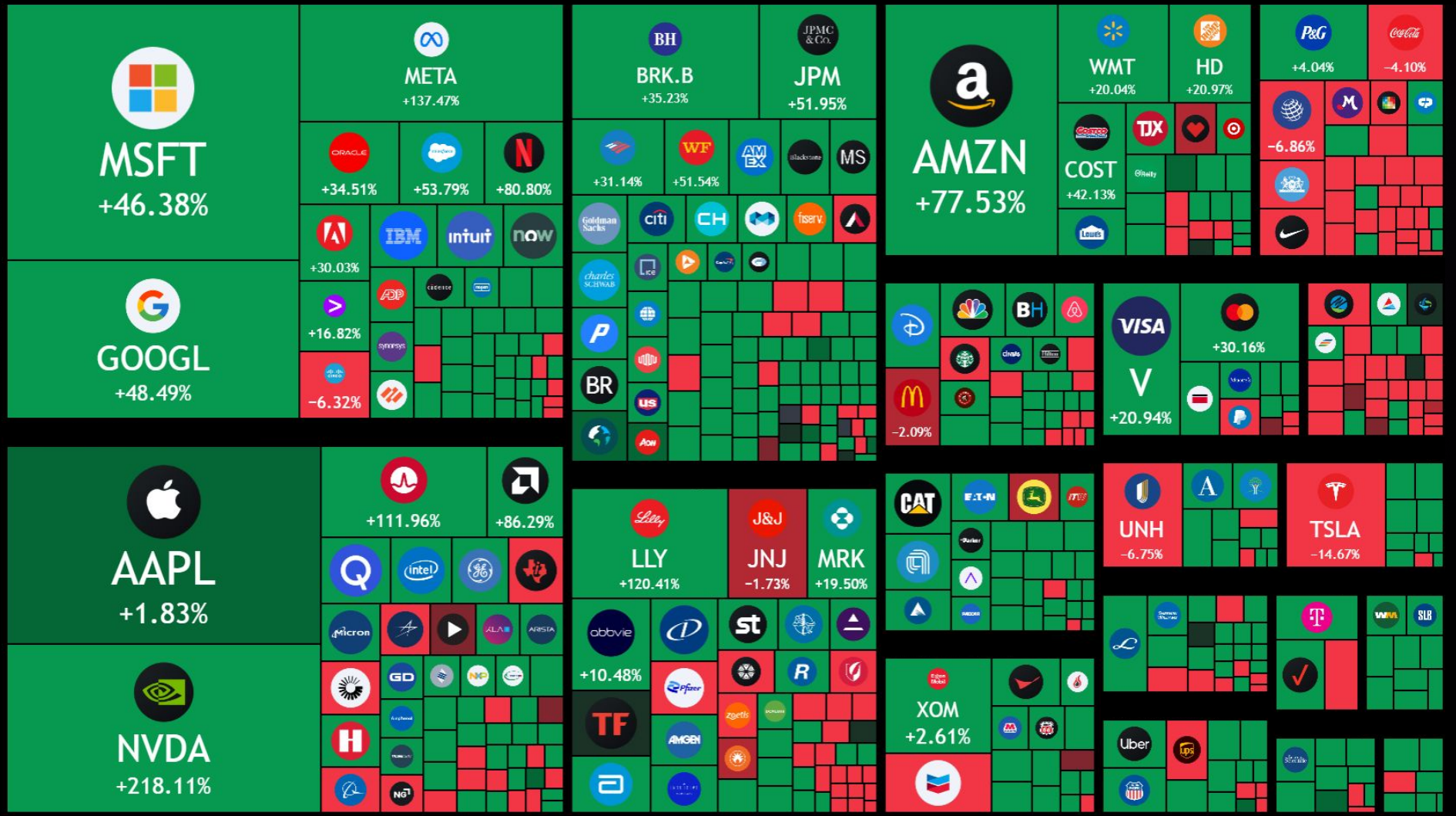
Proprietary data made findable and effective

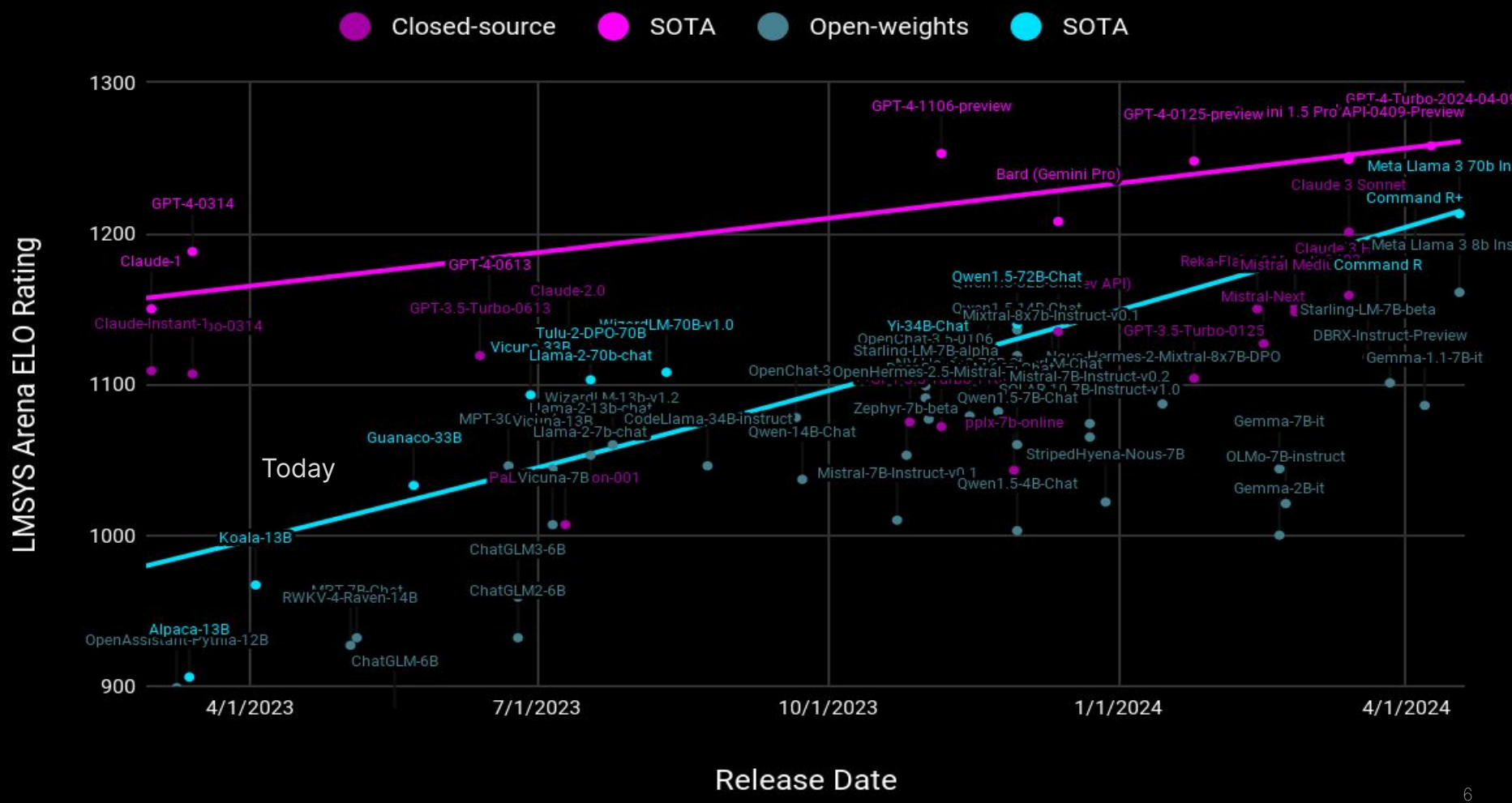
Anomaly detection

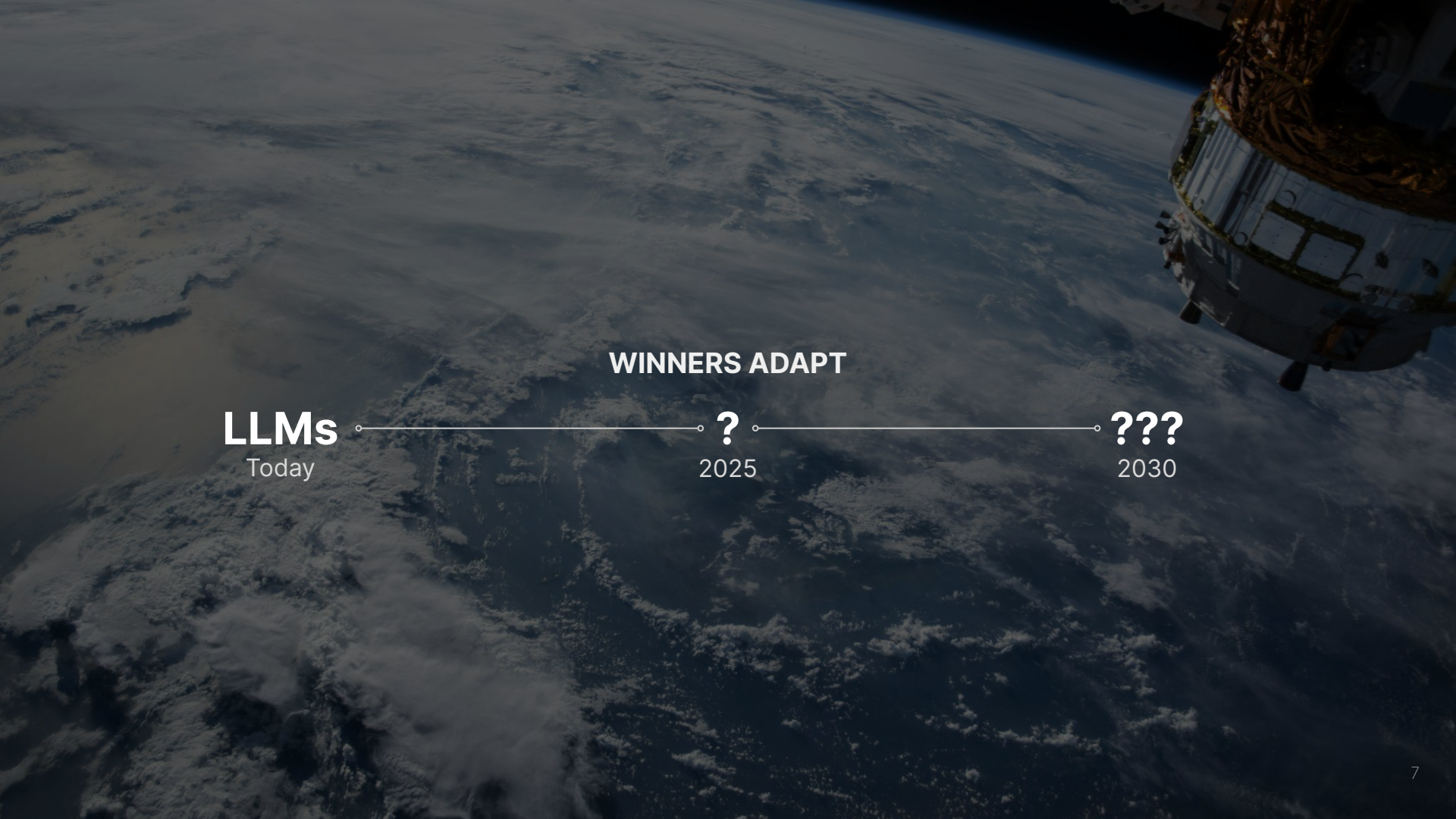
Rapid detection, response command and control

Fraud detection

Fraud reduced, countermeasures activated instantly

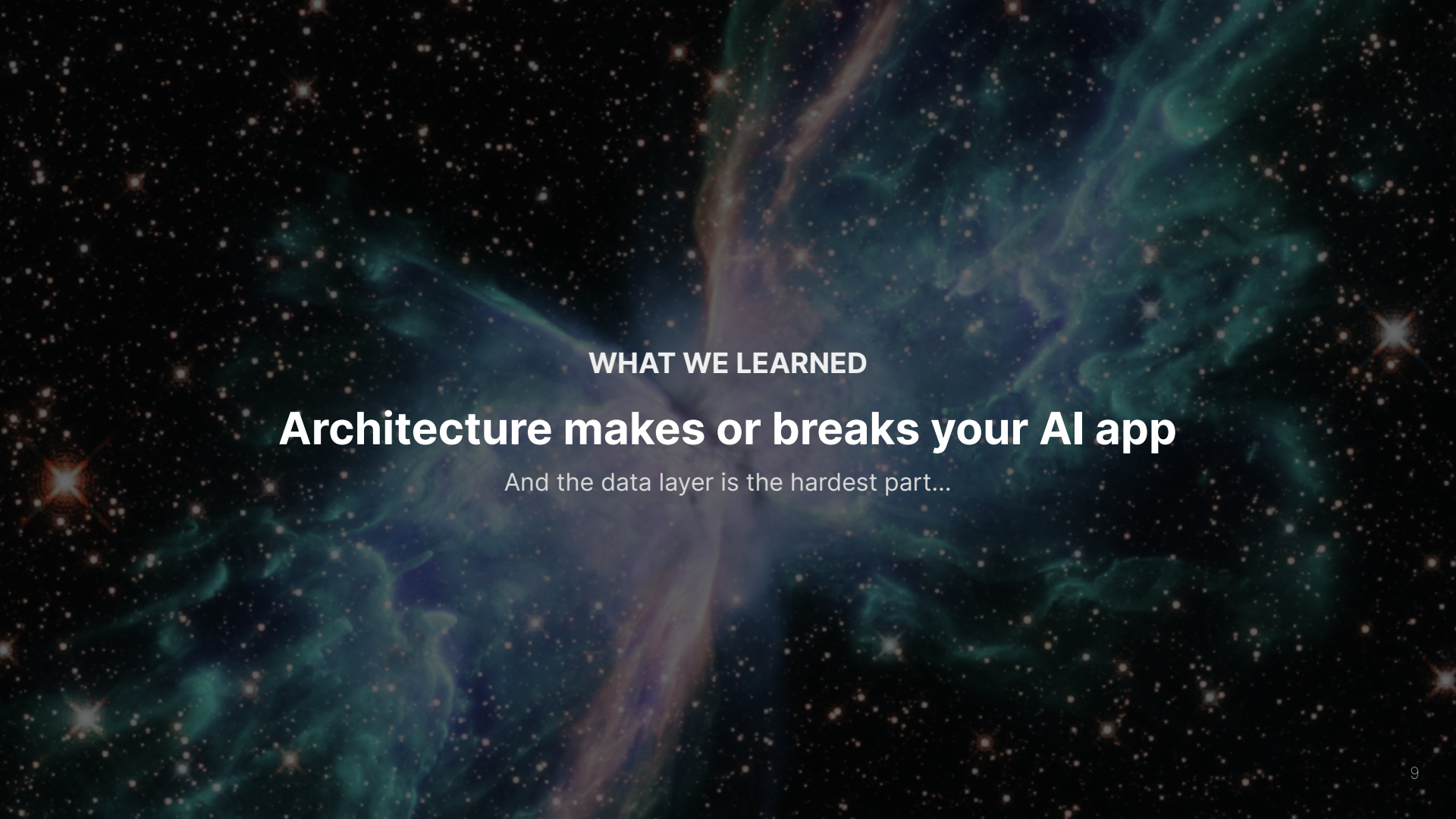

MSFT
+46.38%
GOOGL
+48.49%
META
+137.47%
+34.51%
+53.79%
+80.80%
+30.03%
+16.82%
−6.32%
AAPL
+1.83%
NVDA
+218.11%
+111.96%
+86.29%
BRK.B
+35.23%
JPM
+51.95%
+31.14%
+51.54%
LLY
+120.41%
JNJ
−1.73%
MRK
+19.50%
+10.48%
AMZN
+77.53%
WMT
+20.04%
HD
+20.97%
COST
+42.13%
+4.04%
−4.10%
−6.86%
−2.09%
V
+20.94%
+30.16%
XOM
+2.61%
UNH
−6.75%
TSLA
−14.67%


Closed-source
SOTA
Open-weights
SOTA
LMSYS Arena ELO Rating
1300
1200
1100
1000
900
4/1/2023
7/1/2023
10/1/2023
1/1/2024
4/1/2024
Release Date
Today
GPT-4-1106-preview
GPT-4-0125-preview
Bard (Gemini Pro)
GPT-4-0314
Claude-1
GPT-4-0613
Claude-2.0
GPT-3.5-Turbo-0613
Claude-Instant-1
Tulu-2-DPO-70B
Llama-2-70b-chat
Yi-34B-Chat
Mixtral-8x7b-Instruct-v0.1
Qwen1.5-72B-Chat
Starling-LM-7B-alpha
OpenChat-3
Guanaco-33B
Koala-13B
PaLM-Chat
Llama-2-7b-chat
Llama-2-13b-chat
CodeLlama-34B-Instruct
Qwen-14B-Chat
Zephyr-7b-beta
pplx-7b-online
Mistral-7B-Instruct-v0.2
Qwen1.5-7B-Chat
Mistral-Next
Starling-LM-7B-beta
GPT-3.5-Turbo-0125
DBRX-Instruct-Preview
Gemma-1.1-7B-it
Command R
Command R+
Meta Llama 3 8b Ins
Meta Llama 3 70b In
Claude 3 Sonnet
Gemini 1.5 Pro API-0409-Preview
GPT-4-Turbo-2024-04-0
Gemma-7B-it
OLMo-7B-Instruct
Gemma-2B-it
StripedHyena-Nous-7B
Qwen1.5-4B-Chat
Mistral-7B-Instruct-v0.1
ChatGLM3-6B
ChatGLM2-6B
ChatGLM-6B
RWKV-4-Raven-14B
Alpaca-13B
OpenAssistant-Pythia-12B

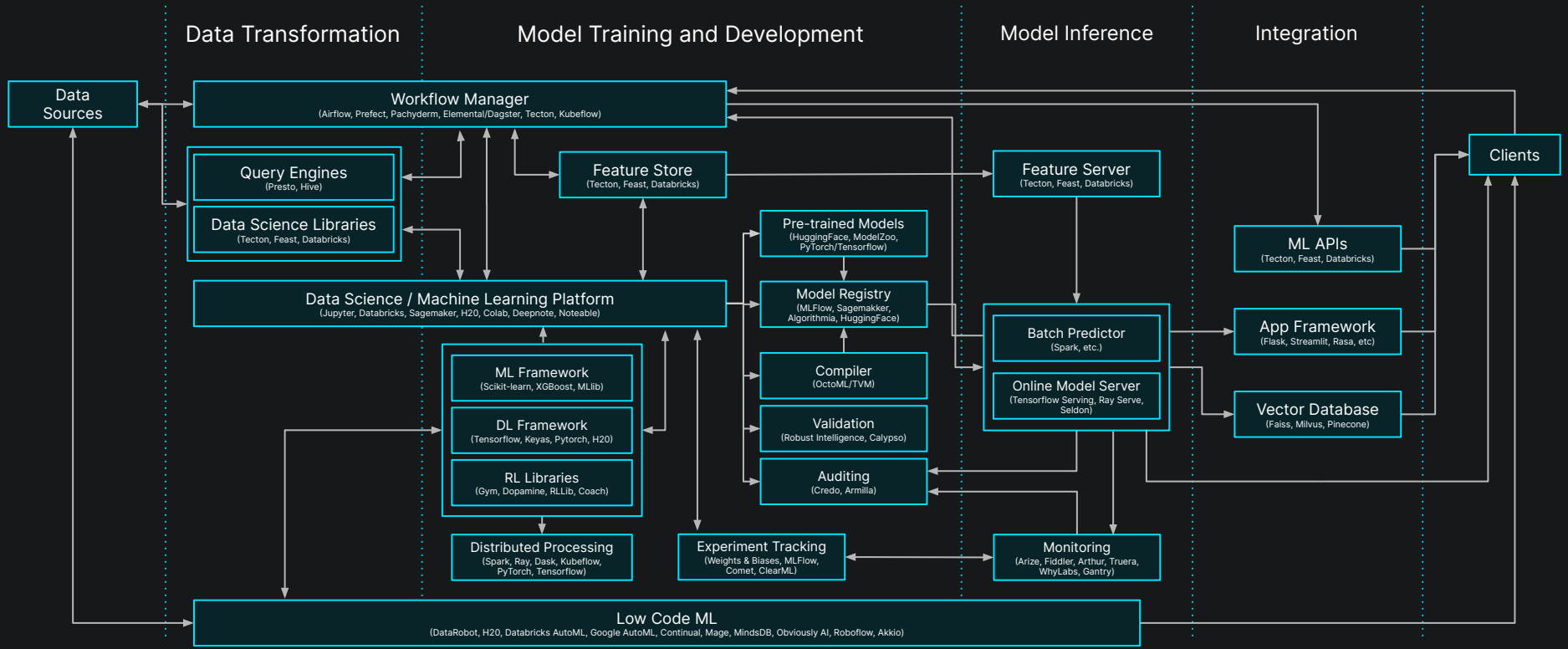

WINNERS ADAPT
LLMs
Today
?
2025
???
2030

Our story

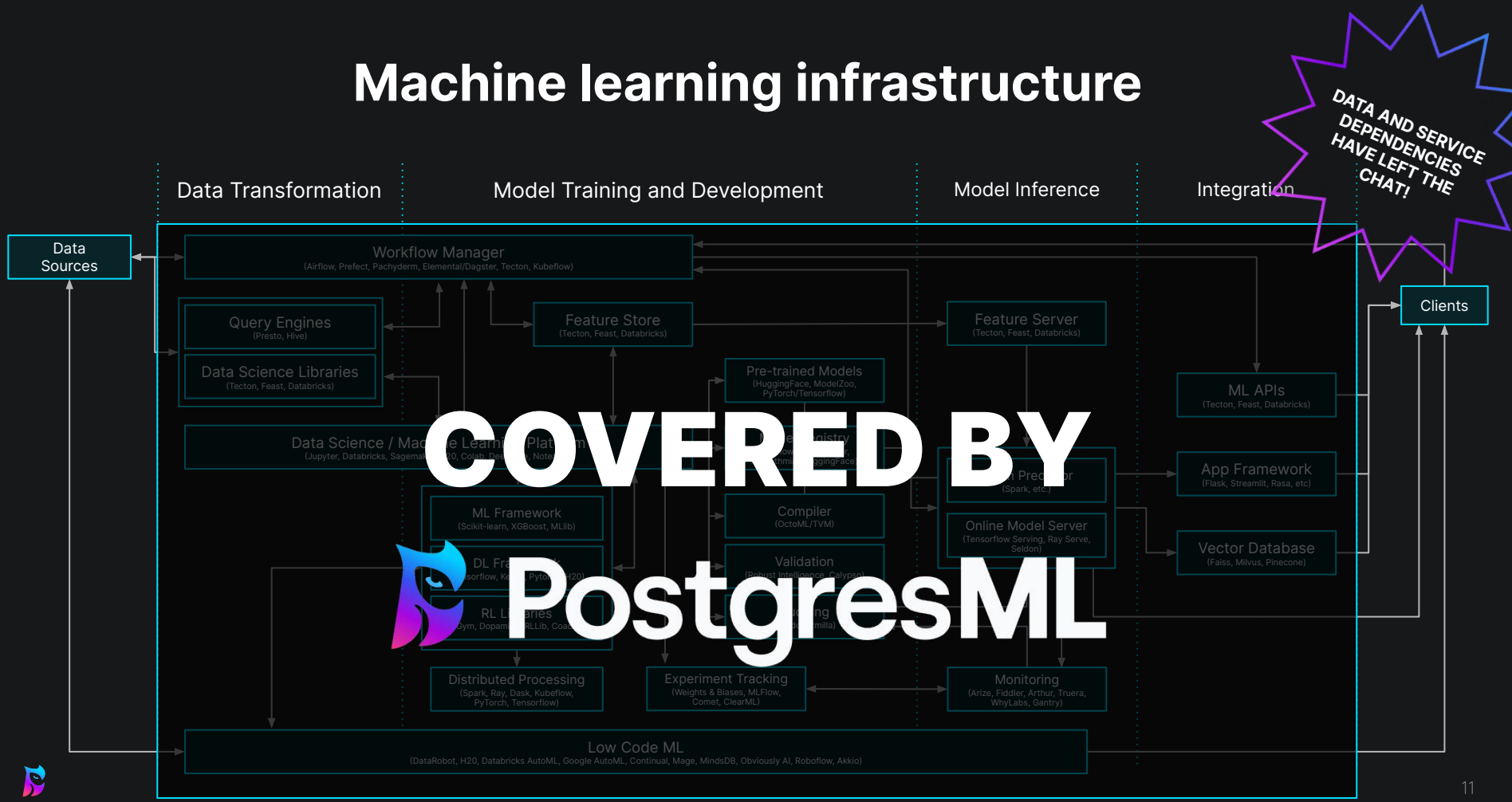
Founded by the engineers that built (then re-built) and scaled Instacart's machine learning platform

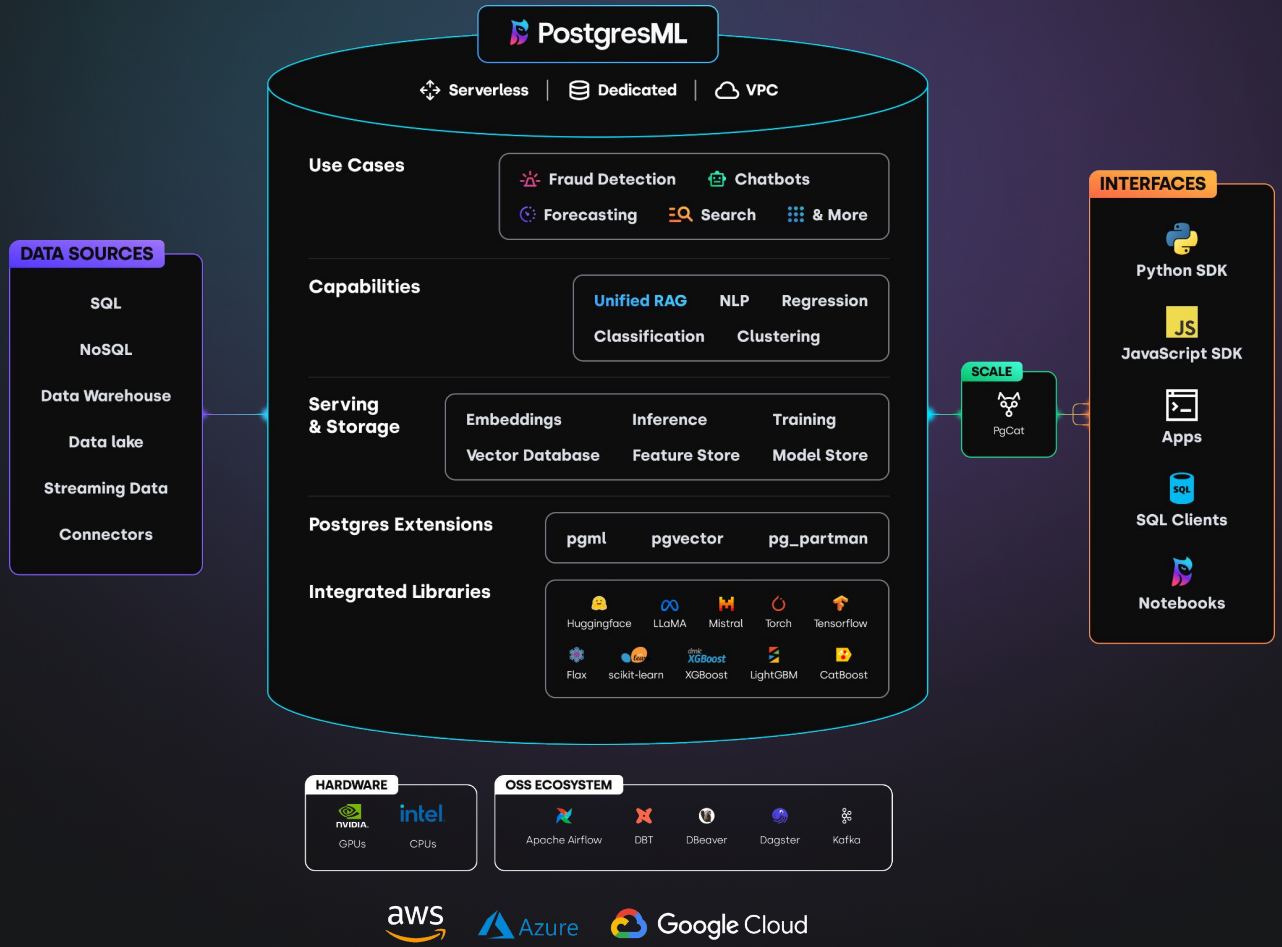

WHAT WE LEARNED
Architecture makes or breaks your AI app
And the data layer is the hardest part...

Machine learning infrastructure



Machine learning infrastructure




PostgresML
Serverless
Dedicated
VPC
Use Cases
Fraud Detection
Chatbots
Forecasting
Search
& More
Capabilities
Unified RAG
NLP
Regression
Classification
Clustering
Serving & Storage
Embeddings
Inference
Training
Vector Database
Feature Store
Model Store
Postgres Extensions
pgml
pgvector
pg_partman
Integrated Libraries
Huggingface
LLaMA
Mistral
Torch
Tensorflow
Flax
scikit-learn
XGBoost
LightGBM
CatBoost
DATA SOURCES
SQL
NoSQL
Data Warehouse
Data lake
Streaming Data
Connectors
SCALE
PgCat
INTERFACES
Python SDK
JavaScript SDK
Apps
SQL Clients
Notebooks
HARDWARE
GPUs
CPUs
OSS ECOSYSTEM
Apache Airflow
DBT
DBeaver
Dagster
Kafka
aws
Azure
Google Cloud

Move the algorithm to the data

4x faster

than HuggingFace + Pinecone
for a RAG chatbot

10x faster

than OpenAI for embedding
generation

10-20x faster

than MindsDB
(not really a DB, it's a microservice)

8-40x faster

than Python + Redis for an
XGBoost microservice

Don’t take our word for it

“...with @postgresml, we're not just streamlining processes but revolutionizing data handling and insights generation in one fell swoop.”

Co-Founder @ MedPiper Technologies, Inc

TRUSTED BY ENGINEERS AT

instacart sudo write OneSignal Alibaba vmware ZAPOJ

Join us

- Attend the pgml RAG session to see SOTA design
- Contribute to our open-source projects, including pg-cat

We're hiring:

Email: **Montana@postgresml.org**