



P A R T N E R S

## MLOops:

A review of the ML stack in industry and the problems that ML stakeholders and tool builders must solve to drive AI adoption

---

**SARAH CATANZARO**

*General Partner*  
Amplify Partners



**who am I?**

# Who am I and why am I here?

# Agenda

## A SNAPSHOT

**01**

Snapshot of ML in industry (beyond FAANG/MANGA)

## THE ML STACK

**02**

A taxonomy of applications, platforms and infrastructure

## KEY CHALLENGES

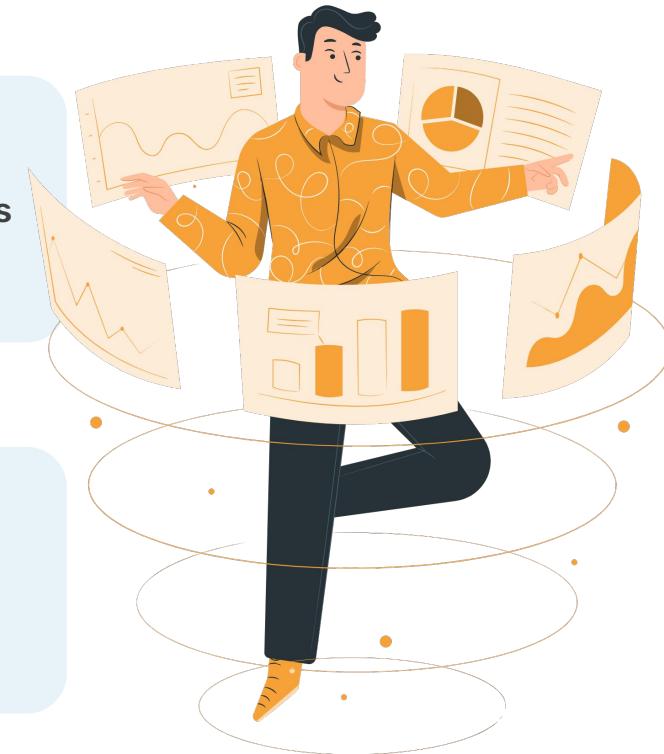
**03**

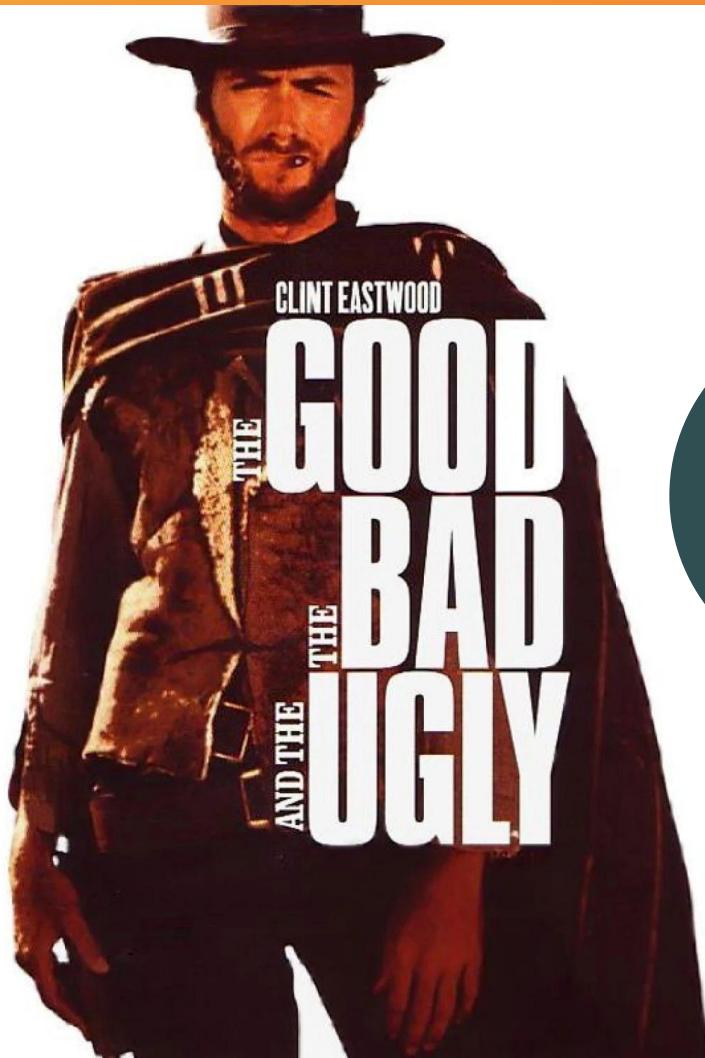
Key challenges ML practitioners face

## KEY CHALLENGES

**04**

Key challenges ML tool builders face





A snapshot of ML  
in industry

# A Snapshot of ML in Industry

High ROI use cases are emerging among mid-market and public companies

E-commerce

Marketplaces

Finance

Real Estate

Compare Similar Items

Image	Name	Price	Rating	Reviews	Action
	Wayfair Basics Metal Bed Frame by Wayfair Basics®	\$48.99	★★★★★	8647	Add
	Hiett Platform Bed Frame by Alwyn Home	\$76.99	★★★★★	10904	See Details
	Blough Heavy Duty Bed Frame by Alwyn Home	\$85.99	★★★★★	4672	See Details
	Higbee Steel Bed Frame by Alwyn Home	\$78.99	★★★★★	5116	See Details
	Vela Heavy Duty Bed Frame by Alwyn Home	\$69.99	★★★★★	1651	See Details

ML-driven SaaS applications

The image displays two side-by-side screenshots of software interfaces. The top interface shows a grid of numerous small, square portrait photos of people, likely a facial recognition or user identification system. Below this grid is a larger, more detailed view of a single person's face, possibly for closer inspection or analysis. The bottom interface is a more complex dashboard, featuring several charts, graphs, and text boxes. One chart on the left shows a distribution of data points. A central text box discusses 'Automatic Repricing Based on Inventory' and 'the AI automatically updates prices based on current inventory levels'. Another section on the right is titled 'Sku Status' and includes a 'Reprice Now' button. The overall theme of both interfaces is the integration of machine learning with business operations.

## A SNAPSHOT OF ML IN INDUSTRY

**Structured Data**



**Embedding**

(structured/semi-structured data)



**Text and visual data**



## A SNAPSHOT OF THE ML INDUSTRY

**Mid-market and public companies have hired ML practitioners and plan to expand their team, but are still experimenting with different organizational models.**

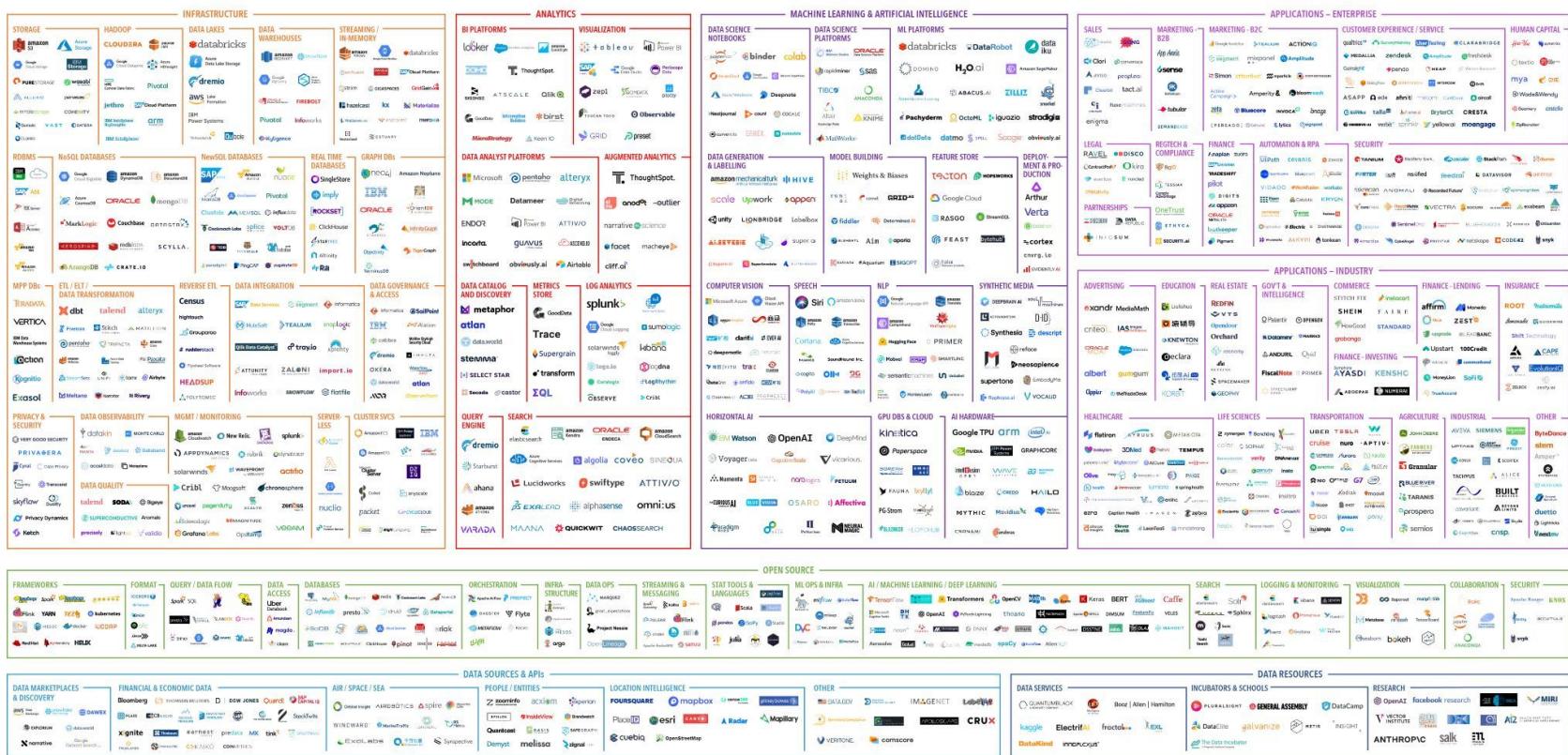
**DATA SCIENTIST AFTER WRITING A FLASK APP**

You know, I'm something  
of a software engineer myself

imgflip.com

# The ML stack; a taxonomy

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



Version 1.0 - September 2021

© Matt Turck (@mattturck), John Wu (@john\_d\_wu) & FirstMark (@firstmarkcap)

mattturck.com/data2021

FIRSTMARK  
EARLY STAGE VENTURE CAPITAL

Amplify Partners - Proprietary and Confidential

# The ML stack; a taxonomy

## MODEL OPERATIONS

MODEL MONITORING



MODEL OPTIMIZATION & DEPLOYMENT



MODEL COMPLIANCE



CONTINUOUS LEARNING



## BUILD & DEPLOY

DATA SCIENCE FRAMEWORK



EXPERIMENT TRACKING



VERSION CONTROL



AUTO ML



DISTRIBUTED TRAINING



## DATA MANAGEMENT

### DATABASES

CLOUD DATA WAREHOUSE

DATA LAKE

VECTOR DATABASE

### FEATURE MANAGEMENT

FEATURE STORE

FEATURE /METRICS LAYER

### DATA LABELING

LABELING SERVICE

WEAK SUPERVISION

# Data Management

## Databases

### DATA WAREHOUSE



### DATA LAKE



### VECTOR DATABASE

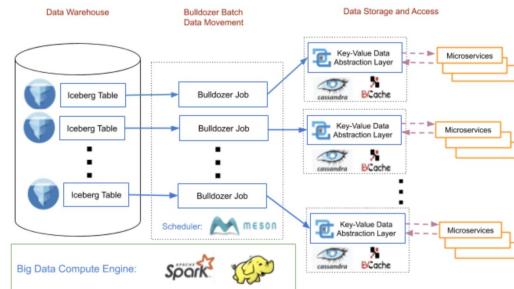


## Feature Management

### FEATURE STORE



### FEATURE LAYER/PROXY



## Data Labeling

### LABELING SERVICES



### WEAK SUPERVISION

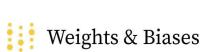


# Build and Deploy

## DATA SCIENCE FRAMEWORK



## EXPERIMENT TRACKING



## VERSION CONTROL



## AUTOML



## DISTRIBUTED TRAINING



# Model Operations

## Model Monitoring



## Model Optimization & Deployment

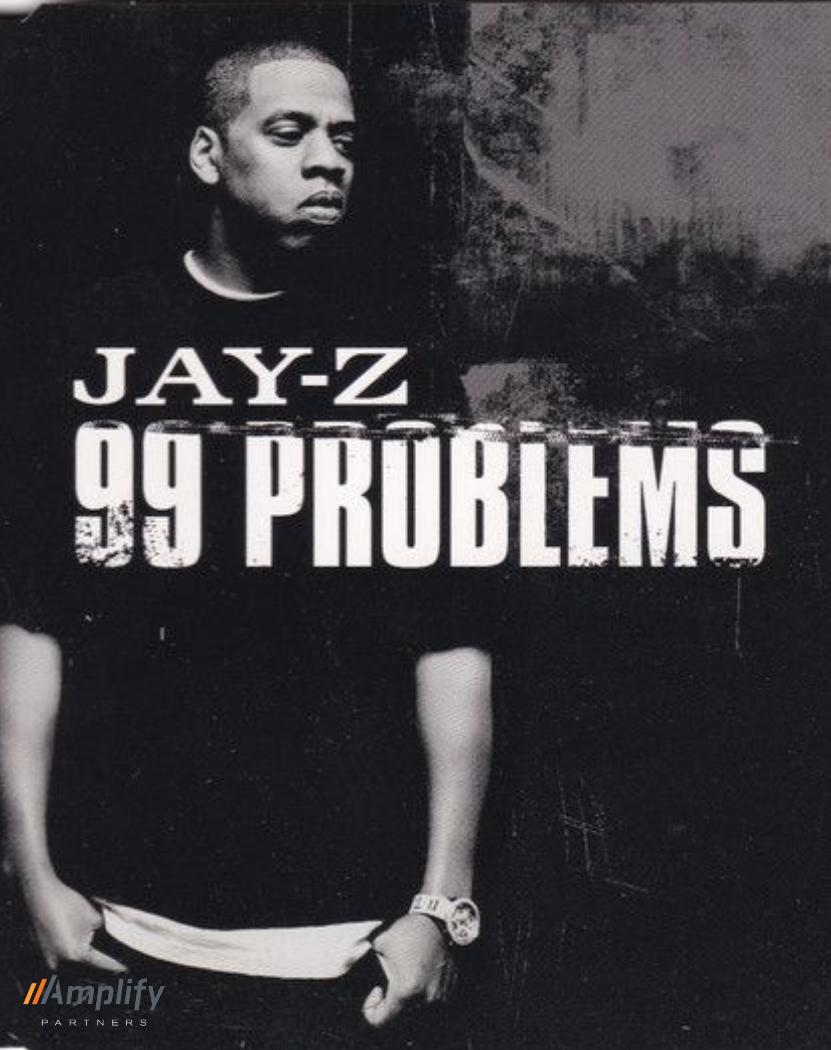


## Model Compliance



## Continuous Learning





# Challenges that ML practitioners face

# SCOPING ML PROJECTS

---

## HOW CAN YOU ANTICIPATE THE:

- Likelihood of success?
- Resources (data, compute, time) needed?

## HOW CAN YOU SCOPE ML EXPERIMENTS TO:

- Avoid expending unnecessary effort on projects that aren't feasible
- Leverage engineering resources effectively

Nice paper

---



GitHub  
Link

---



Written in  
your favourite  
framework

---



Runs smoothly on  
your system  
without error or  
dependency issues

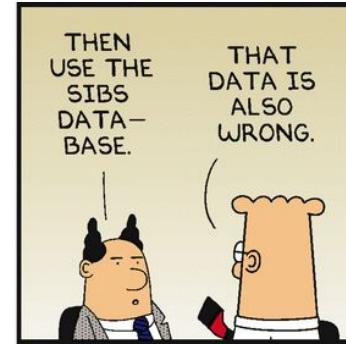
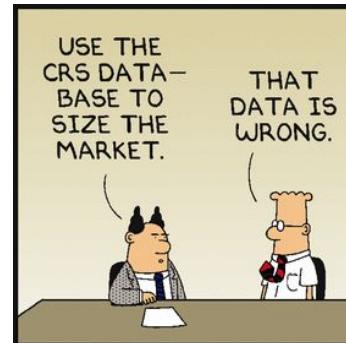
---



# ACCESSING AND ITERATING ON DATASETS

## PRACTITIONERS NEED:

- Access to high quality data
- Tools for data-centric ML
- Unstructured data wrangling

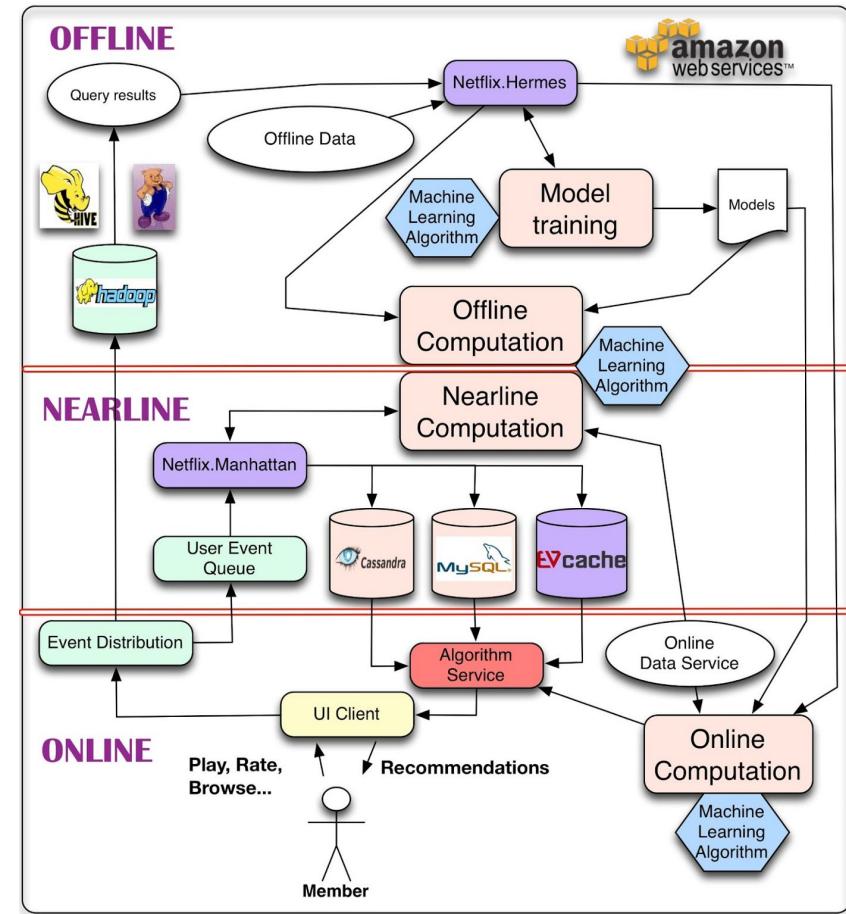


# BUILDING SOFTWARE APPLICATIONS WITH ML

Practitioners must build software applications for which a prediction service (ML model) is one component

## NEED TO INTEGRATE WITH:

- Data management systems
- Other [micro]services (e.g. authentication)
- Developer tools (e.g. CI/CD)



# MANAGING INFRASTRUCTURE

- DS should spend time on modeling and analyses NOT managing infrastructure and environments!





INTRODUCING

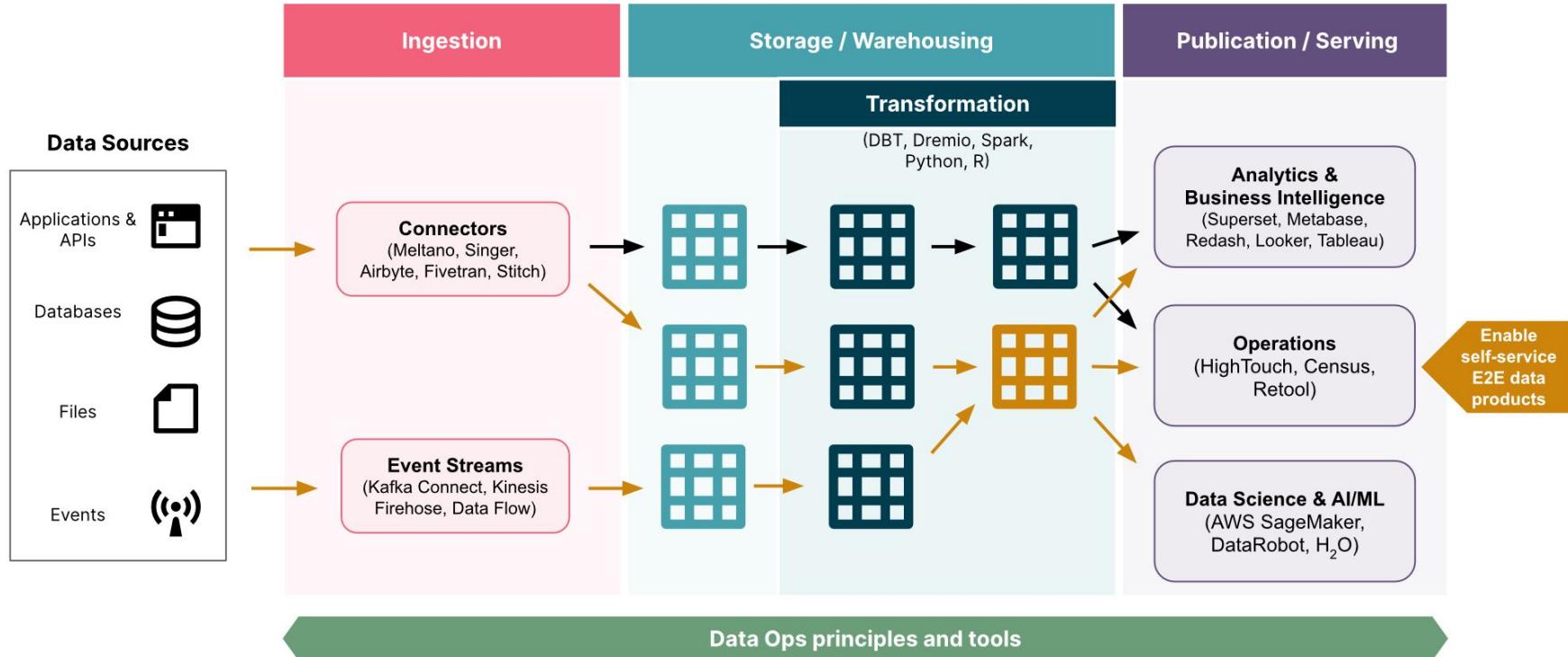
# Modal Labs





# Challenges ML Tool Builders Face

# Comparing the “Modern Data Stack” and ML Stack



**NO DOMINANT DESIGN  
FOR THE ML STACK**



**NO system of record**



**NO best practices or  
standards**



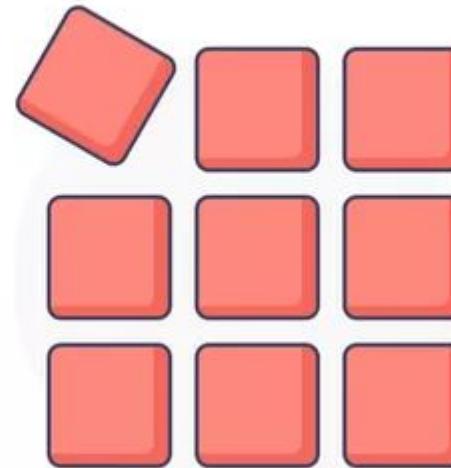
**NO clear roles and  
responsibilities**

# No partnership ecosystem: Why everything becomes an end-to-end platform

**BEST OF BREED  
STACK**



**ALL IN ONE SUITE**



# ONE ML STACK OR MANY?

- Different applications may have different requirements
- Different model architectures may have different requirements

