DATA+AI
SUMMIT 2022

Scaling Real-Time ML at Cash App with Tecton

Michael Barnathan
Director of Applied ML, Cash App

Mike Del Balso
Co-Founder, Tecton

ORGANIZED BY databricks

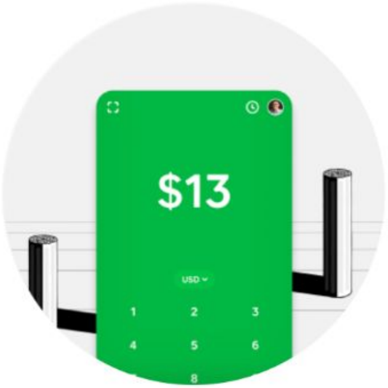
Introduction to the Speakers

Michael Barnathan
Head of Applied ML
Cash App

Mike Del Balso
Co-founder & CEO
tecton

Intro to Cash App

Not just payments!



Send

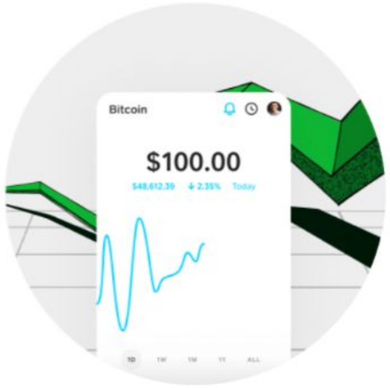
Pay anyone, instantly.

Spend

Design a debit card to match your style.

Bank

Speed up your direct deposits.[1]



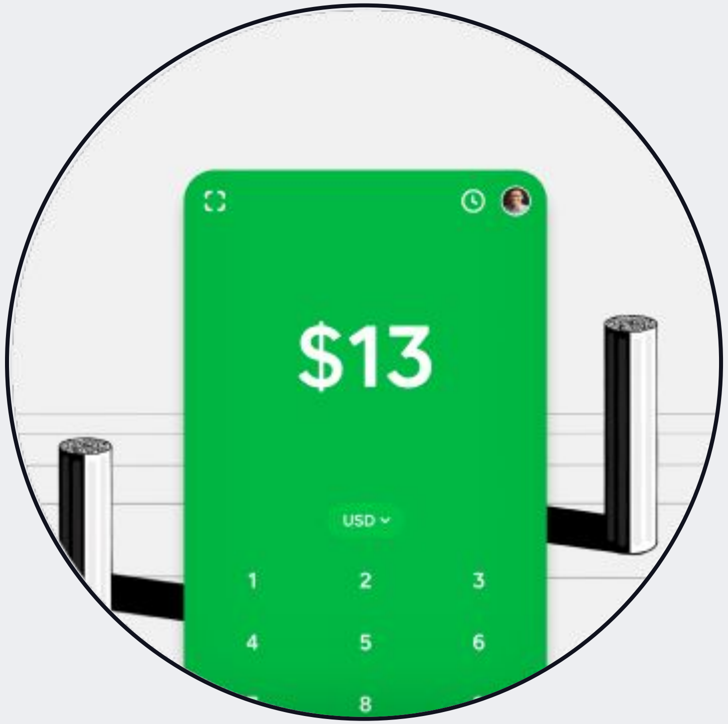
Invest

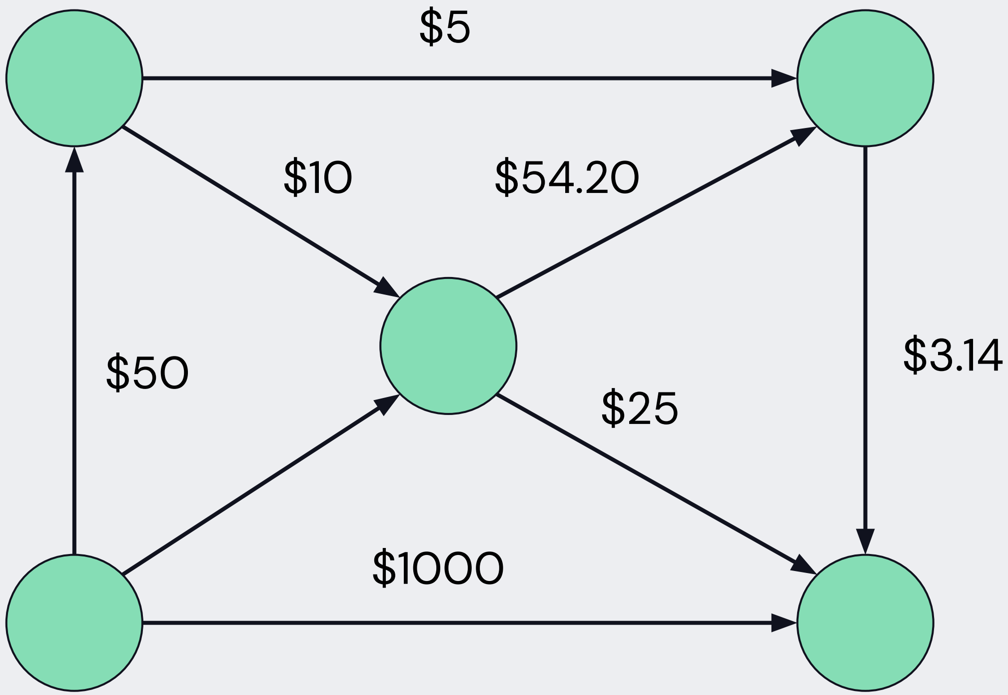
Buy stocks and bitcoin with as little as $1.[2]

Cash App's goal is to "redefine the world's relationship with money"

Send

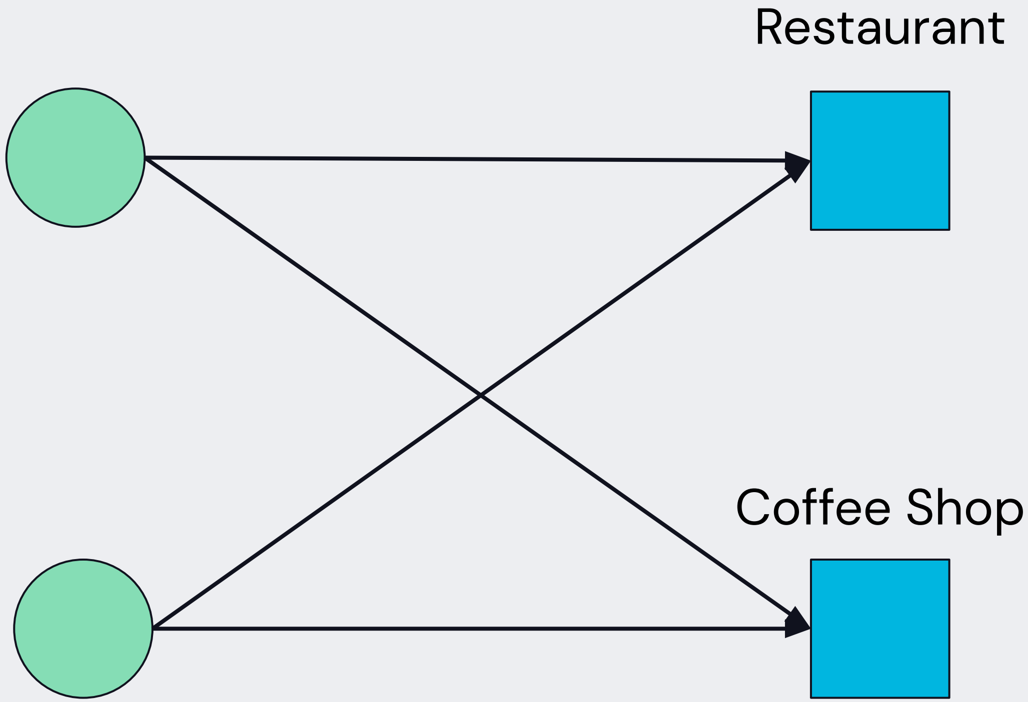
Huge payment graph incoming





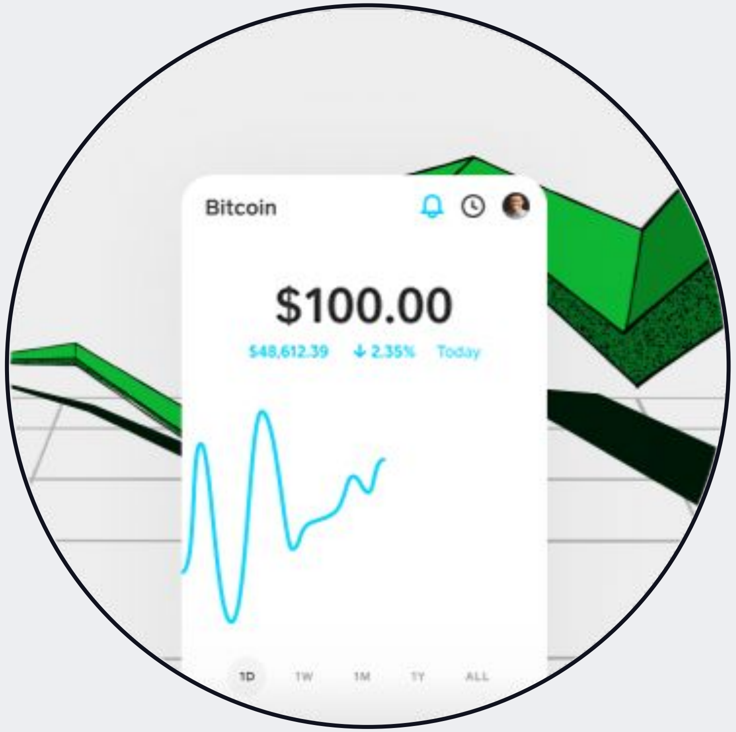
Spend

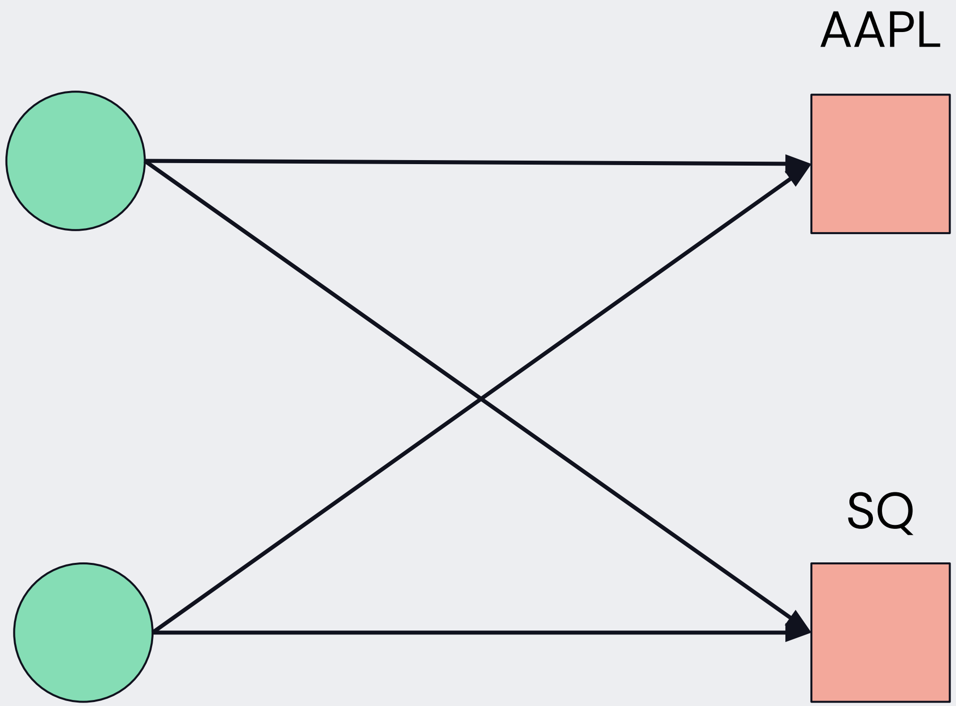
Cool, now it's bipartite!



Invest

User-to-asset reasoning





Why are search and discovery important?

Search:

1. Significant boost to conversion rates when the result you want is in the top three
2. You can use distances in the search space to limit expensive postprocessing or filtering to promising candidates
3. Search queries are an additional indicator of user intent

Discovery

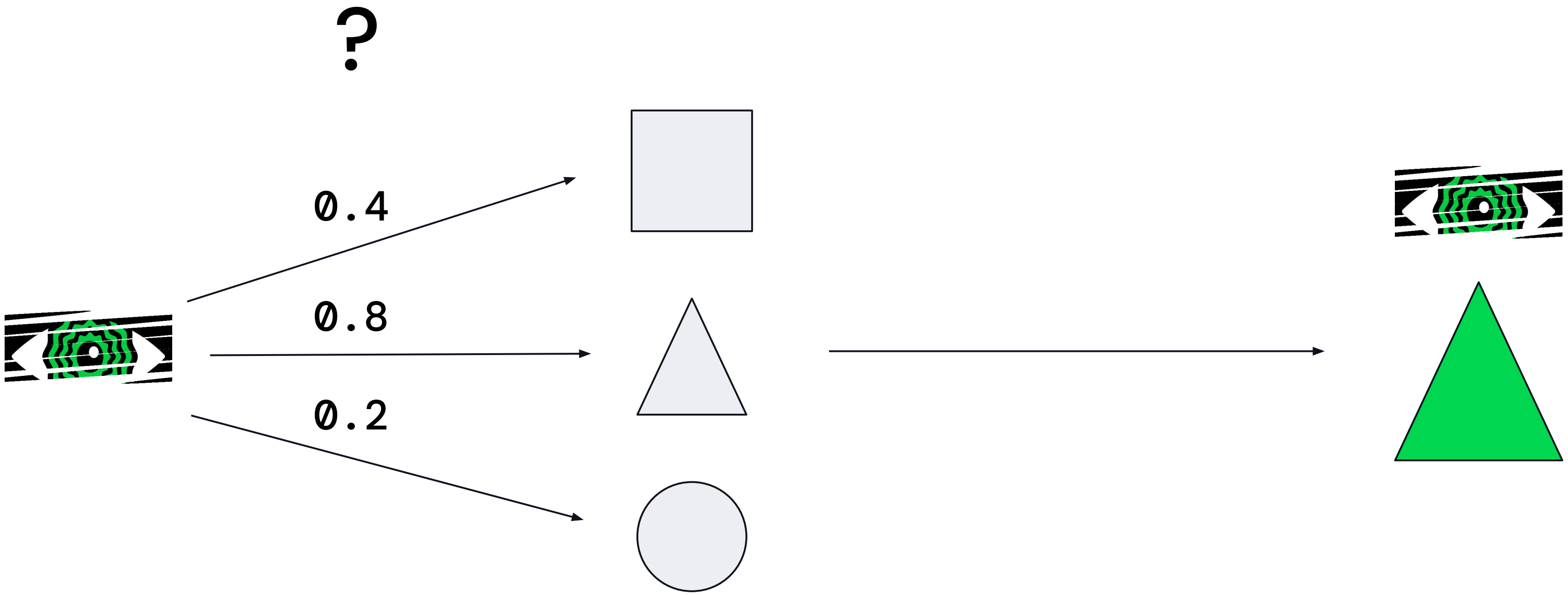
1. Cohesive UX: user's past actions influence their experience in the app
2. The right functionality is "just there" if we predict intent accurately

What is the Recommendation task?

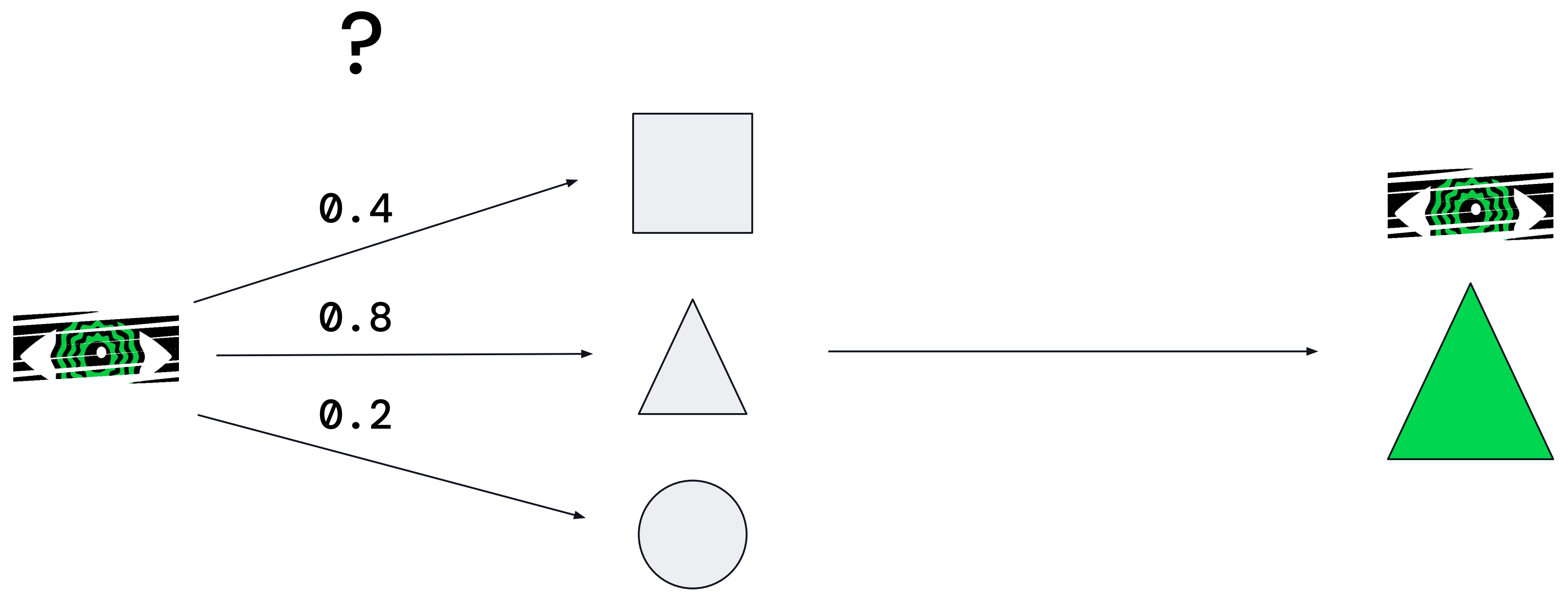
Bob the data scientist wants to do recommendation

What does Bob do?

Join and rank two entity types

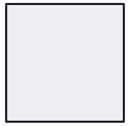


Simple, right?

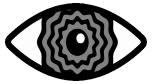


Maybe not so much

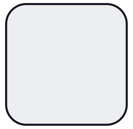
Step 1: Featurize both entities



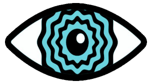
= ⟨0.8, 0.1, 0.5⟩



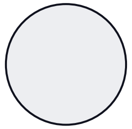
= ⟨165, 42, 42⟩



= ⟨0.3, 0.1, 0.4⟩



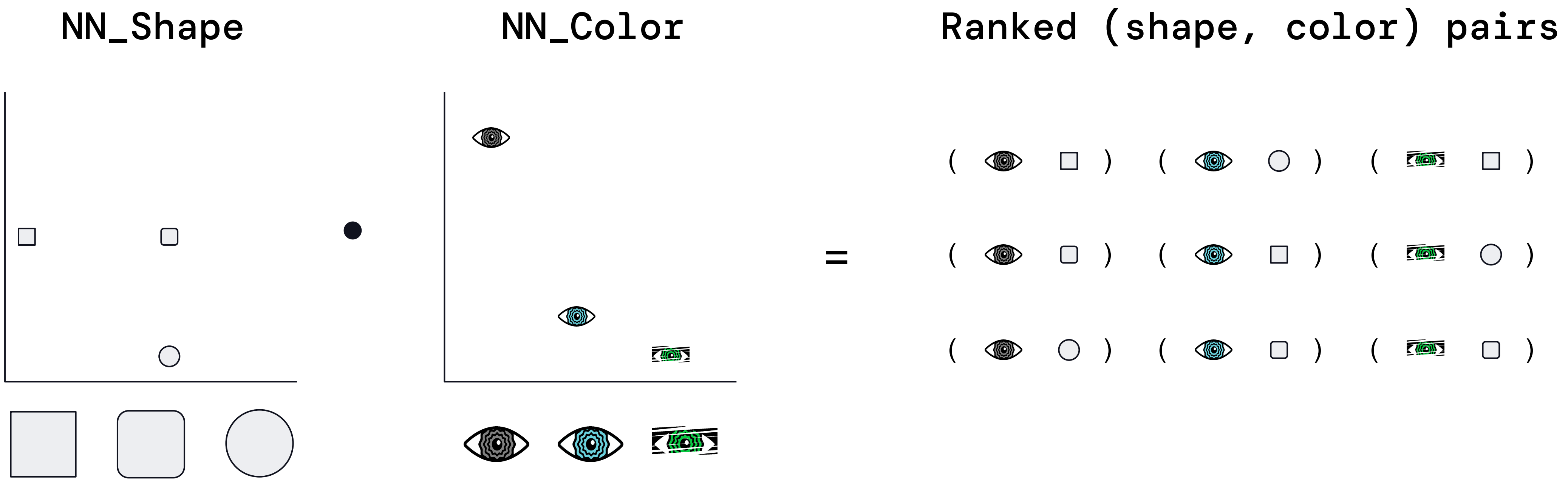
= ⟨0, 0, 128⟩



= ⟨0.1, 0.9, 0.2⟩

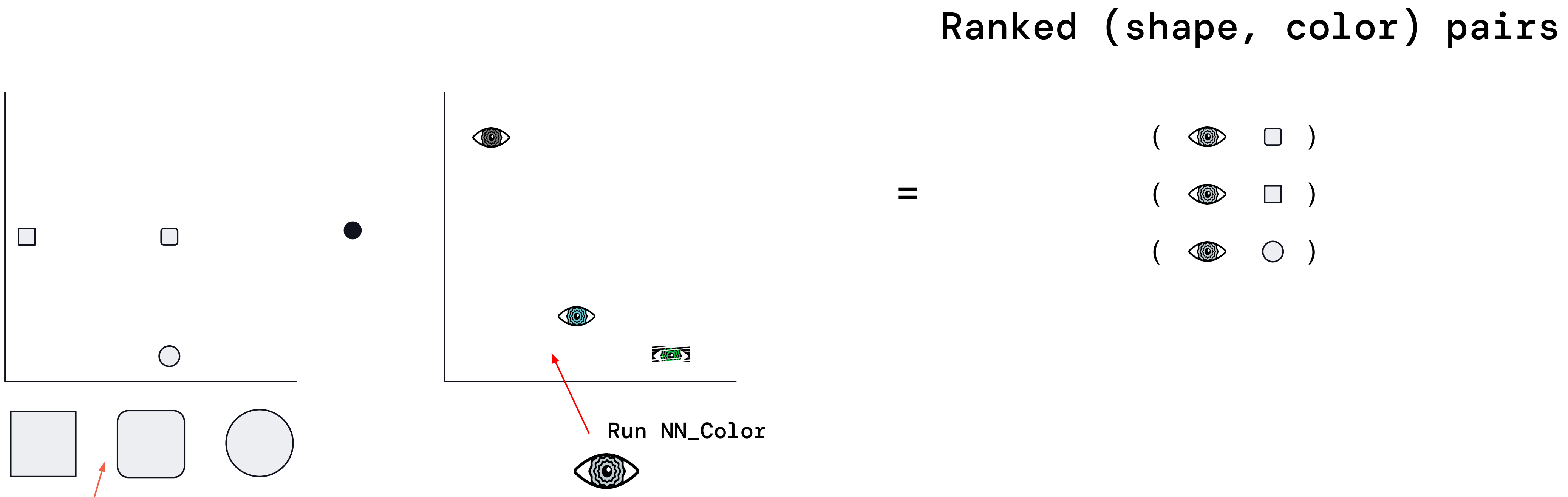
= ⟨0, 128, 0⟩

Step 2: Generate joint embedding



[Two tower models in Tensorflow](#)   ([Or you can always use matrix factorization](#))

Step 3: Retrieval



Shape embedding is the same for each user; you can cache the vectors

Congratulations, Bob! You're hired.

Now do recommendations in Cash App

Bob’s dealing with some serious scale…

- 44 million monthly active users as of Q4 2021
- $12b in revenue as of 2022
- Can you do 100k+ QPS to feature store and model hosting pipelines?
  - With end-user acceptable (say <200ms) latency?
  - And at least 3 nines of uptime?
- Data generally needs to be recent, if not real-time

Factorization of a $44m^2$ matrix is impossible (without creating and exploiting sparsity); we looked at the embedding approach

But this isn’t just a technical problem

The Organizational:

1. Privacy / Protecting PII. This is critical!
2. Team Ownership - core rec engine vs use case
3. Support and maintenance
4. Understanding the pipeline, running experiments

Uh oh. This is hard.

Our preexisting infra wasn't a good match

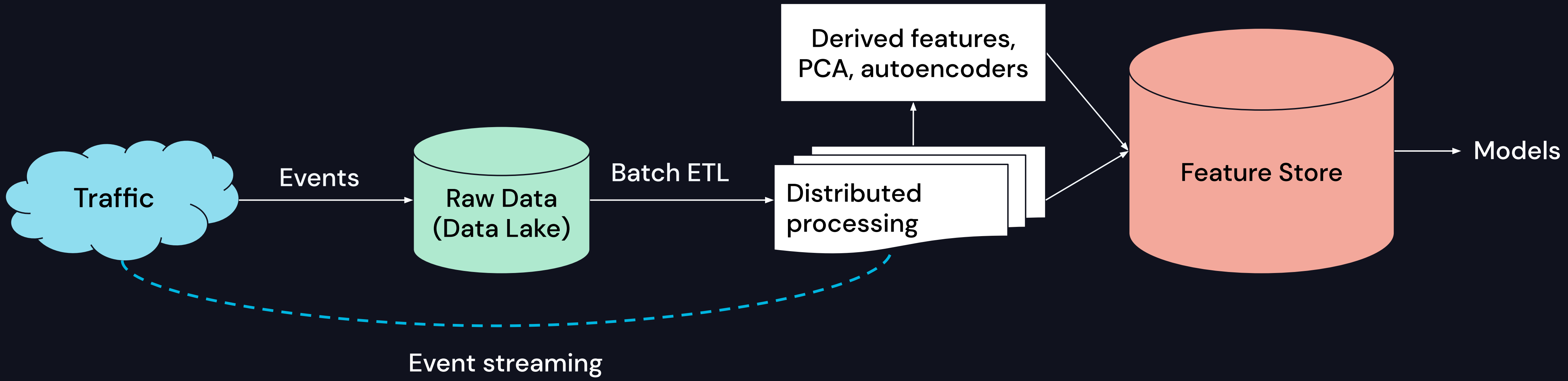
Issue

- Our existing feature store wasn't designed for this level of throughput!
- Calling our model hosting service also incurred network + serialization costs
- Feature caching – workable, but traded off performance for feature freshness
- Existing infra couldn't handle array-valued features, which are required to store embeddings
- Difficult for scientists; eng support

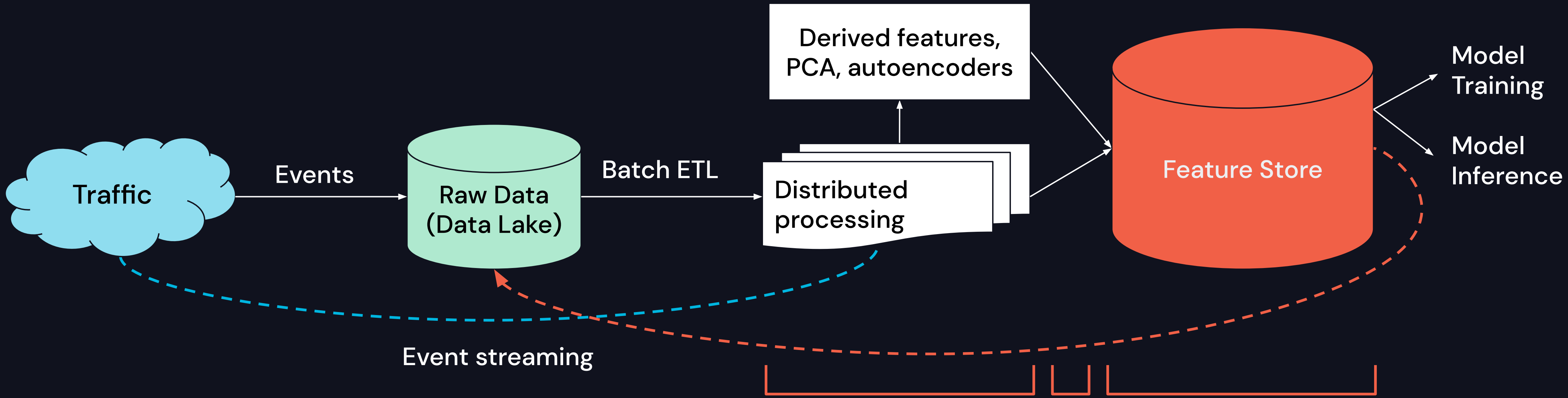
Effect

- Some requests > 1s
- Extra latency
- Features delayed by 30 min or more
- Couldn't use for recommendation
- Less eng team bandwidth

So we looked at typical feature pipeline architectures



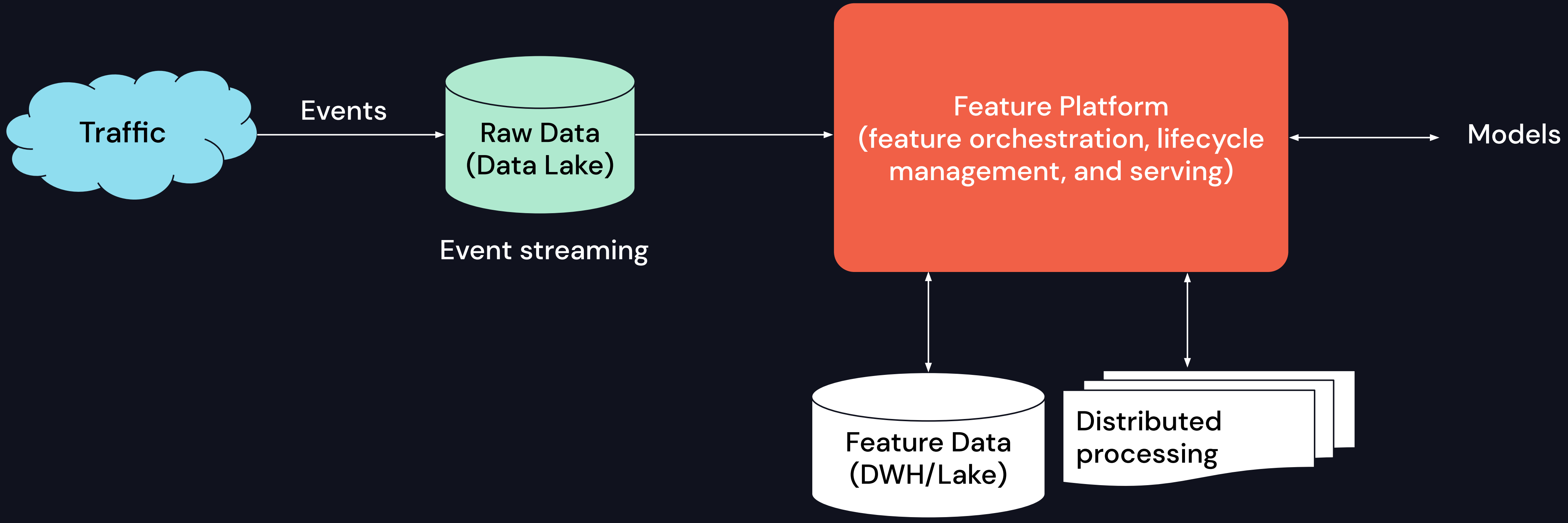
The typical architecture has some significant challenges too



- Logging
- Debugging
- Orchestration
- Caching
- Maintenance

- Latency
- Serialization
- Type conversion
- Network bandwidth

So we looked to a more comprehensive "feature platform" architecture



What’s a “Feature Platform”?!

Feature platforms power the data flows in ML applications

A feature platform:

- Supports the whole feature lifecycle: development, compute, backfill, storage, serving, logging, sharing
- Implements and orchestrates efficient ML data flows (like feature compute and compex retrieval)
- Operates high-reliability real time feature serving and compute for online ML applications
- Solves collaboration and governance problems from operational ML applications

Feature platforms power the data flows in ML applications

How you use it:

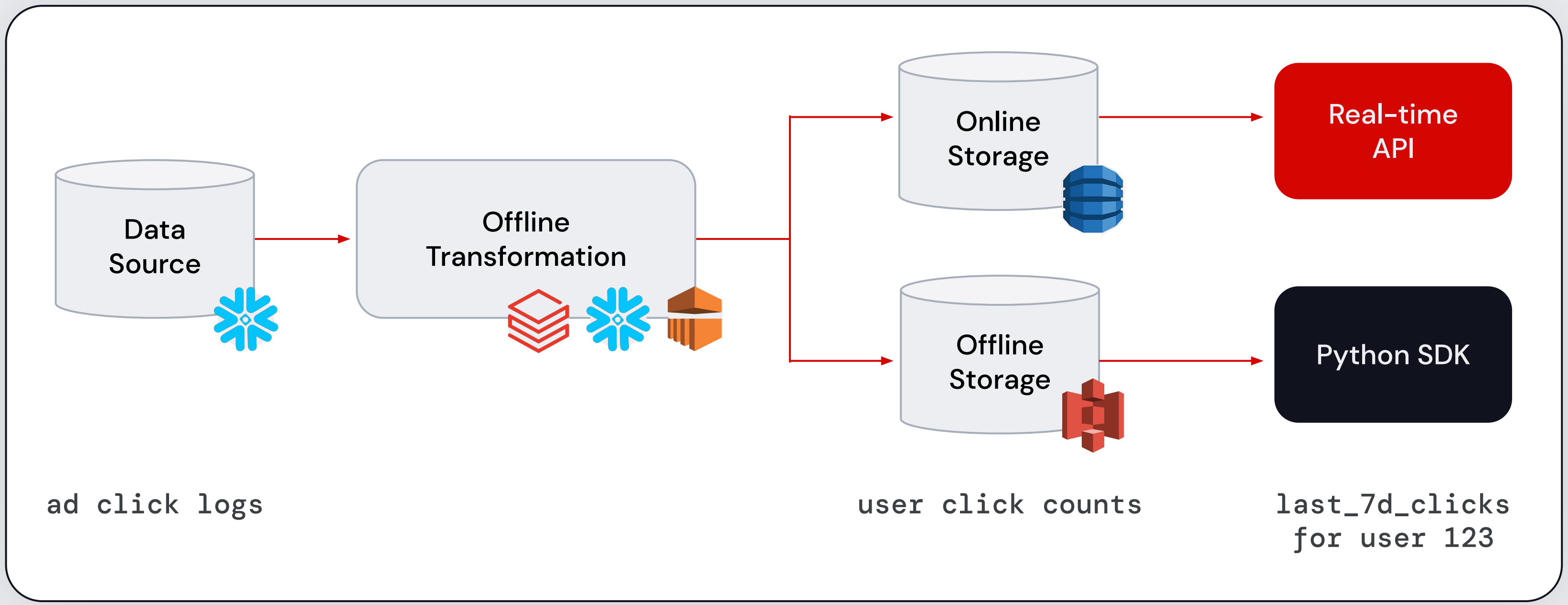
1. Define your features
2. Tecton orchestrates all the dataflows for your features
    - Backfills old feature values for training
    - Generates point-in-time accurate training datasets
    - Computes and serves fresh values for real-time inference
    - Logs served features / observed labels for later model training
    - Monitors feature data for drift / quality / staleness
3. Train models
4. Make predictions in production!

Simple definitions → production features in minutes

Simple Feature Definition
.py file

```
// Declarative Feature Definition
@feature_view(
    inputs=[ad_impressions],
    window='7d',
    entities=[ad],
    online=True,
    offline=True,
    mode="sql"
)
def
ad_ctr_preformance_7_days(ad_impressions):
    return f"""
        SELECT
            ad_id,
            feature_end_time,
            sum(clicked) as last_7d_clicks,
            count(*) as last_7d_impressions
        FROM
            {ad_impressions}
        GROUP BY
            1, 2
    """
```

Compiles to Physical Pipeline



1) Feature dev workflow: manage features as code

1 Write Feature Definitions

```
// Declarative Feature Definition
@feature_view(
    inputs=[ad_impressions],
    window='7d',
    entities=[ad],
    online=True,
    offline=True,
    mode="sql"
)
def
ad_ctr_preformance_7_days(ad_impressions):
    return f"""
        SELECT
            ad_id,
            feature_end_time,
            sum(clicked) as last_7d_clicks,
            count(*) as last_7d_impressions
        FROM
            {ad_impressions}
        GROUP BY
            1, 2
    """
```

2 Test changes in private workspace

Commit feature to local Git branch → Apply changes to private Tecton workspace → Test and validate feature

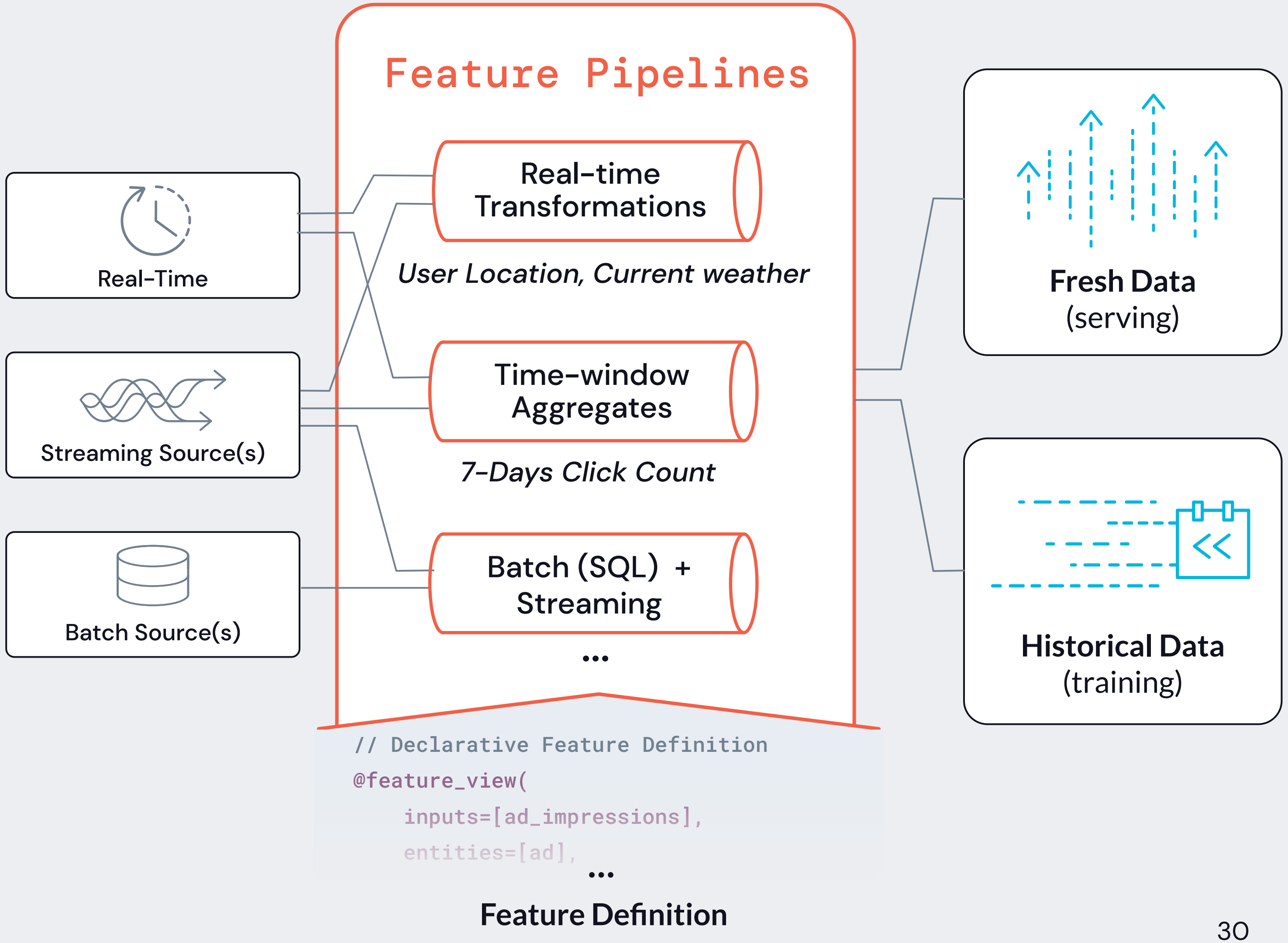
3 Deploy feature to production

Merge local Git branch to master → Apply changes to production Tecton workspace → Monitor feature

2) Feature Pipelines: Transform feature data reliably

Fully-automated ML data pipelines

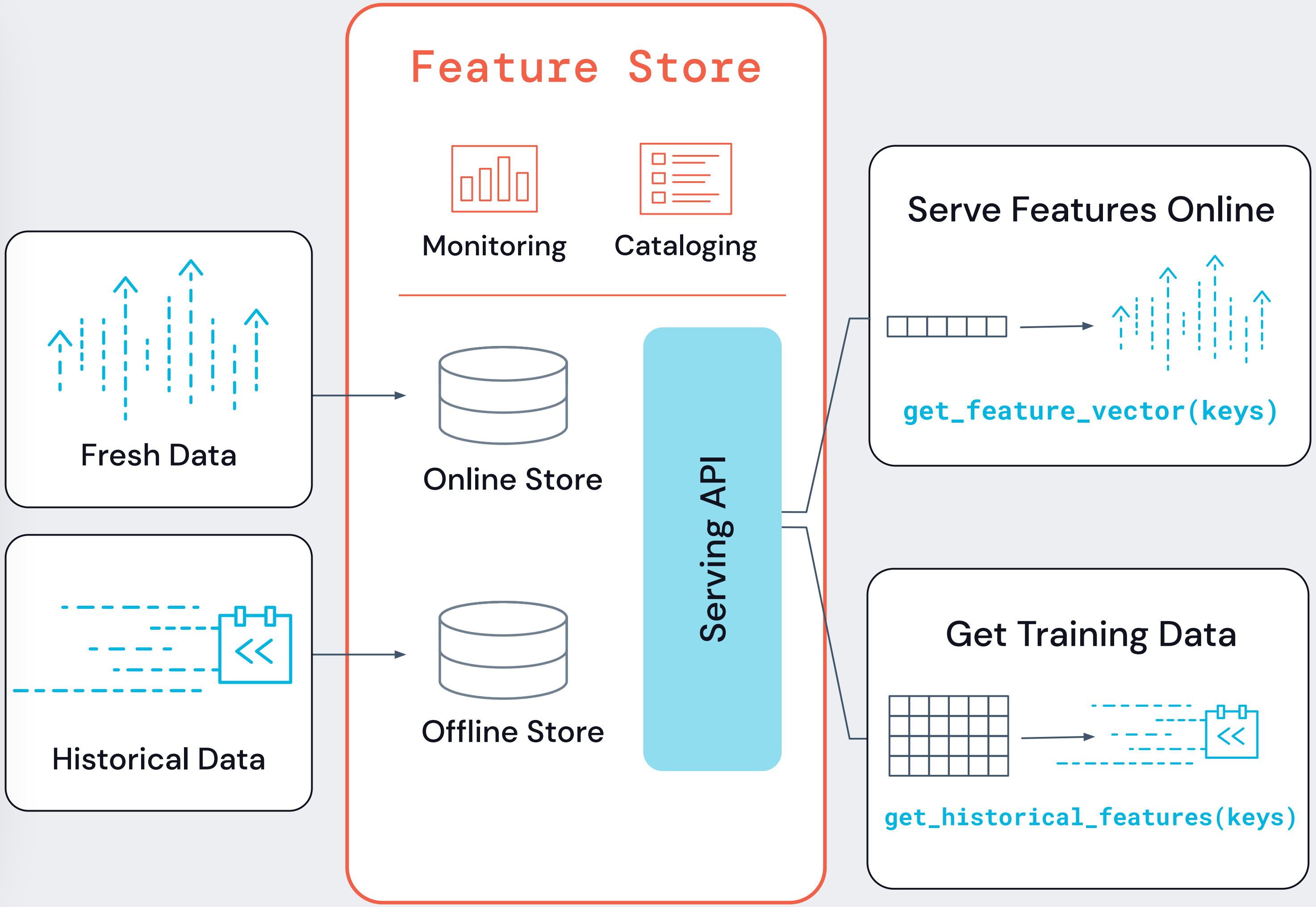
- Orchestrates reliable compute of fresh feature values
- Easy to build batch, streaming, realtime features
- Simple and optimized common features like time-window aggregates
- Automated backfilling



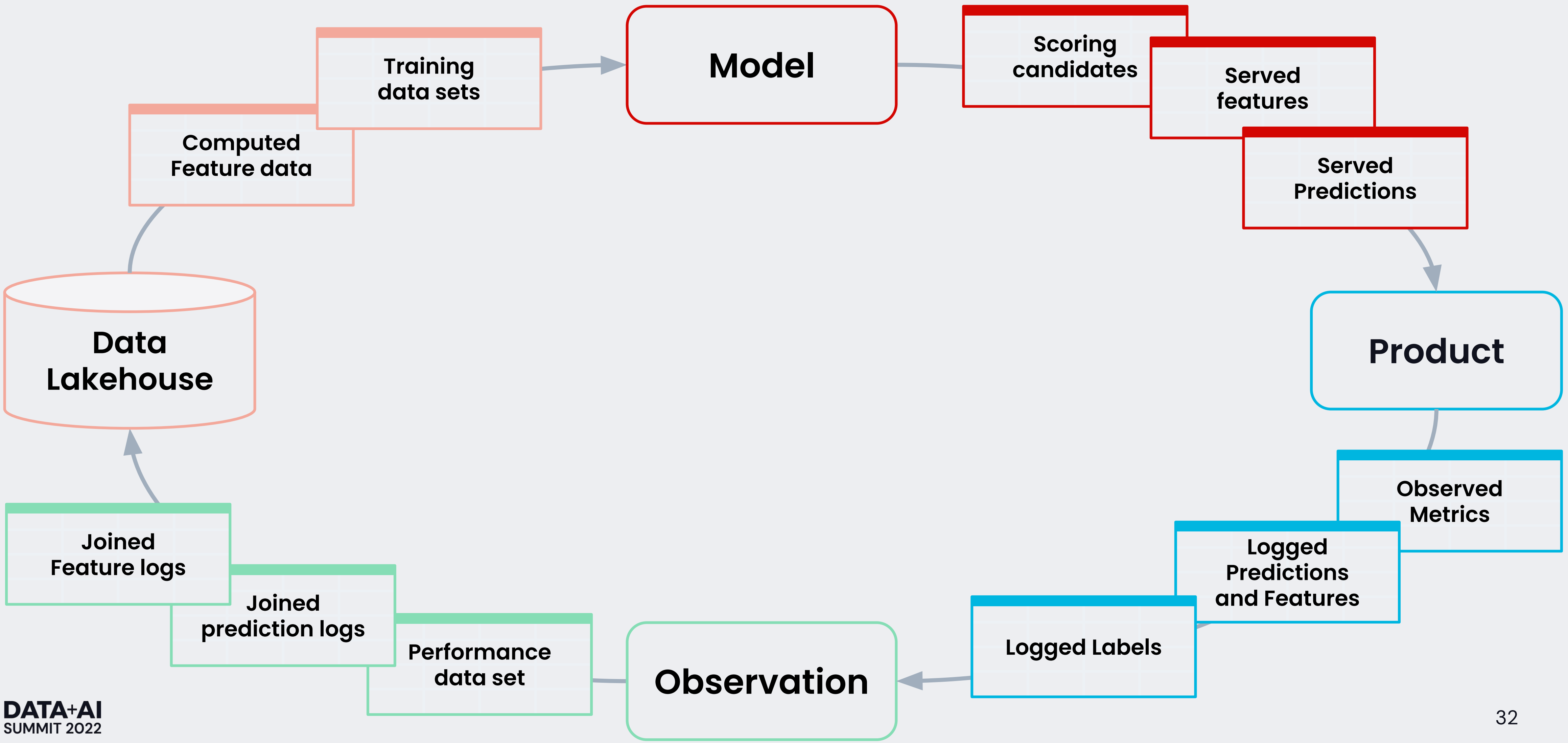
3) Feature Store: Store and Serve features at scale

Serve accurate data for training and online inference

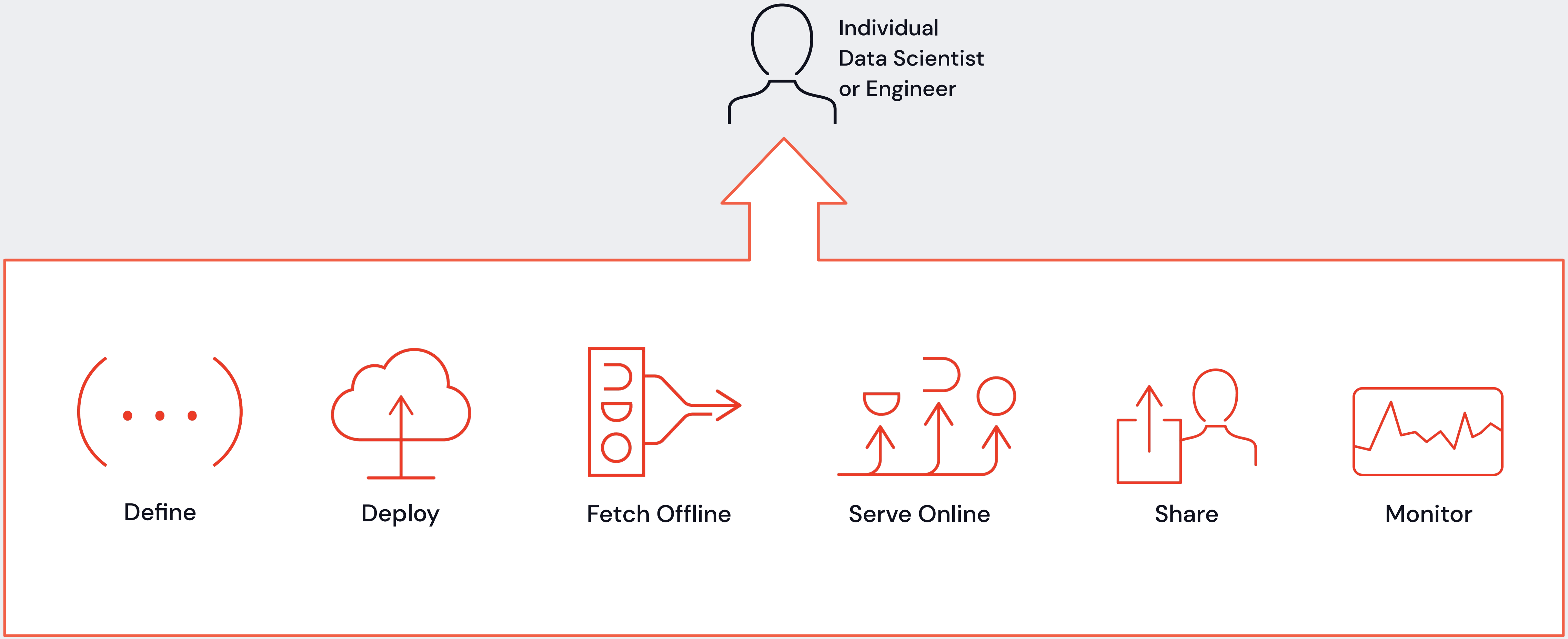
- Ensure consistency between online and offline data
- Serve features online at very low latency and very high scale reliably
- Store historical feature values and retrieve feature data with point-in-time accuracy
- Log served values
- Monitors data and service levels



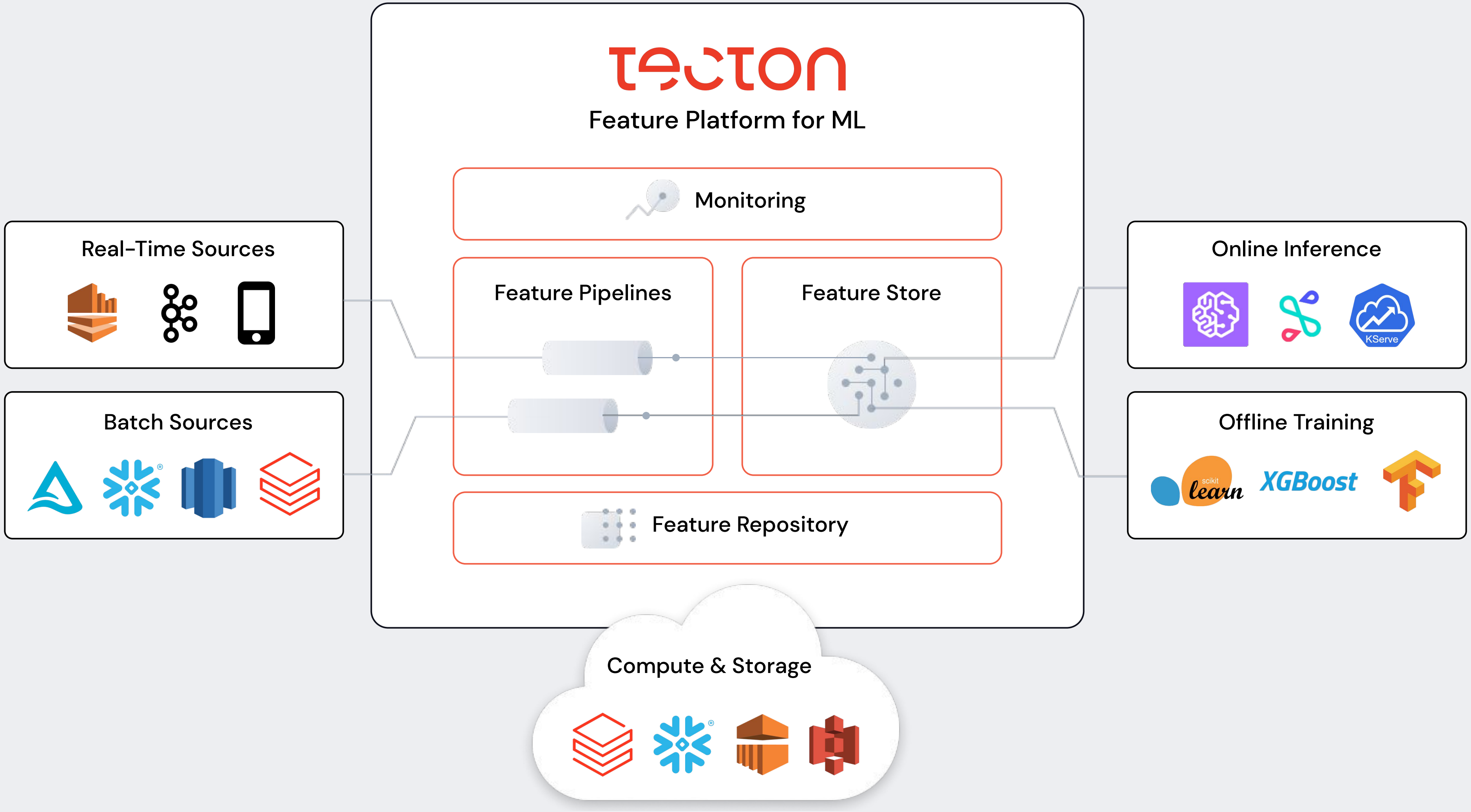
The Feature Platform manages data across the entire ML lifecycle



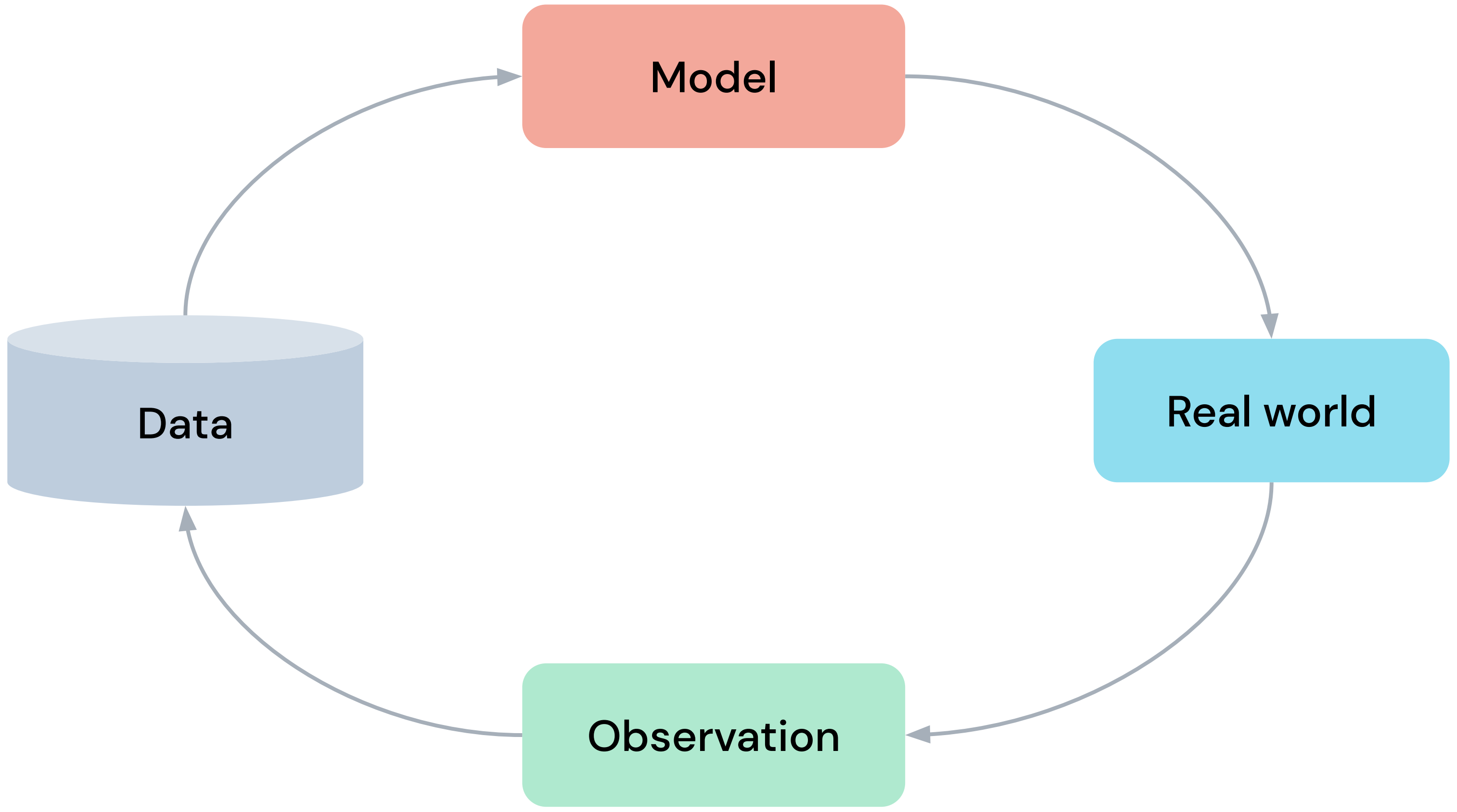
Now adding features to a production model is easy for any team member



One system to manage features across the entire ML Lifecycle

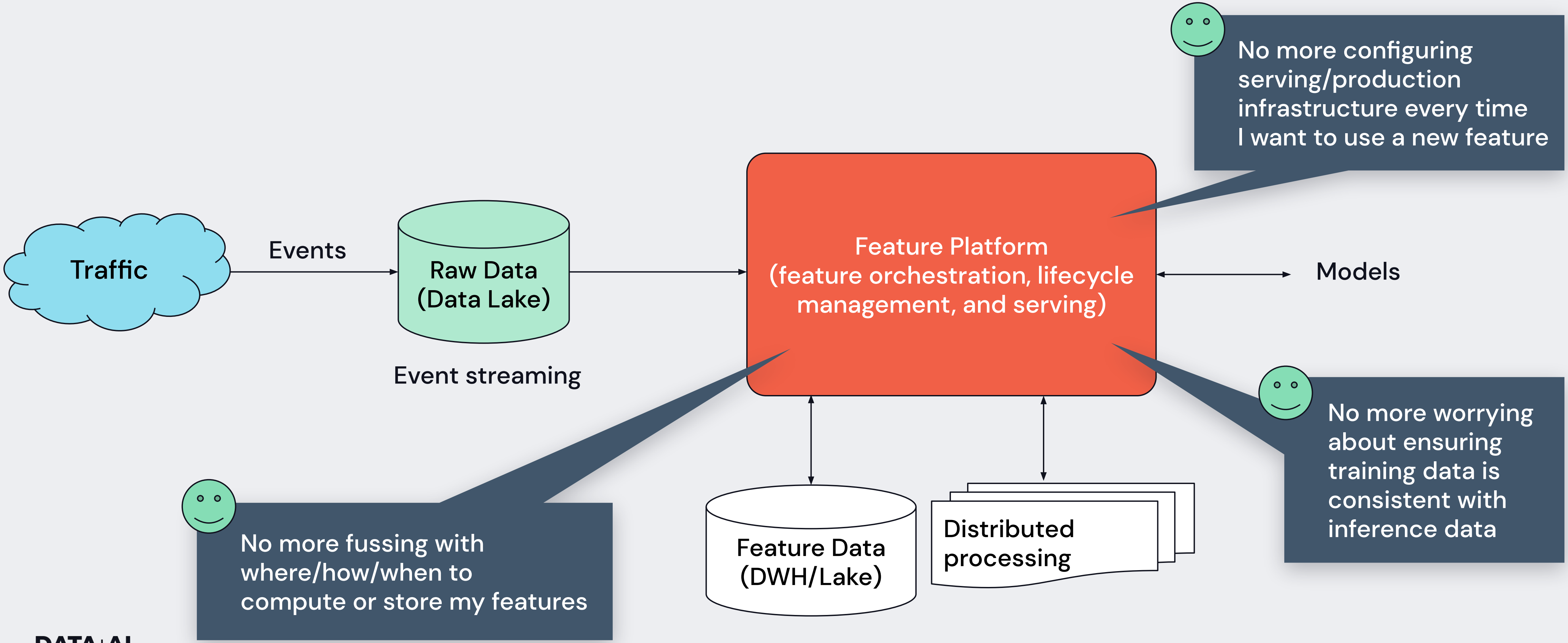


Holistic feature management enables an ML flywheel with compounding returns



So how does this apply to payments recommendation?

This sped up our ML teams, both DS and eng!



Nice outcomes

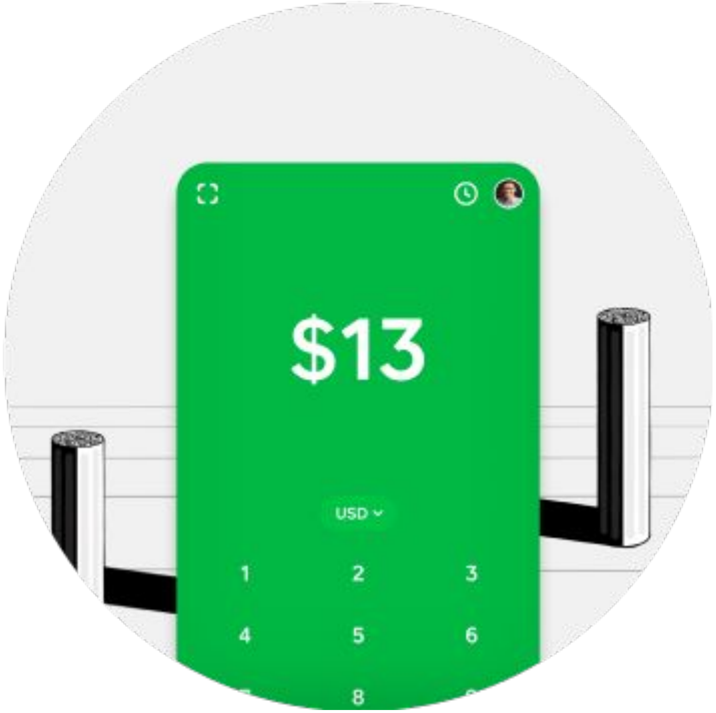
Technical

- Simplifies our ecosystem
- Data and compute were kept “close”, 30ms reduction in network latency
- Eliminated 100ms serialization overhead between compute and feature layer
- Fewer SEVs / less maintenance overhead

Organizational

- One system, no ownership questions
- Easier for scientists to plug in directly without requiring engineering support
- Can focus more on end to end SLAs, high level business logic
- Locating the data needed to trace and respond to events becomes easier
- Configure a feature in hours, not days!

Happily ever after



Faster iteration = happier modelers

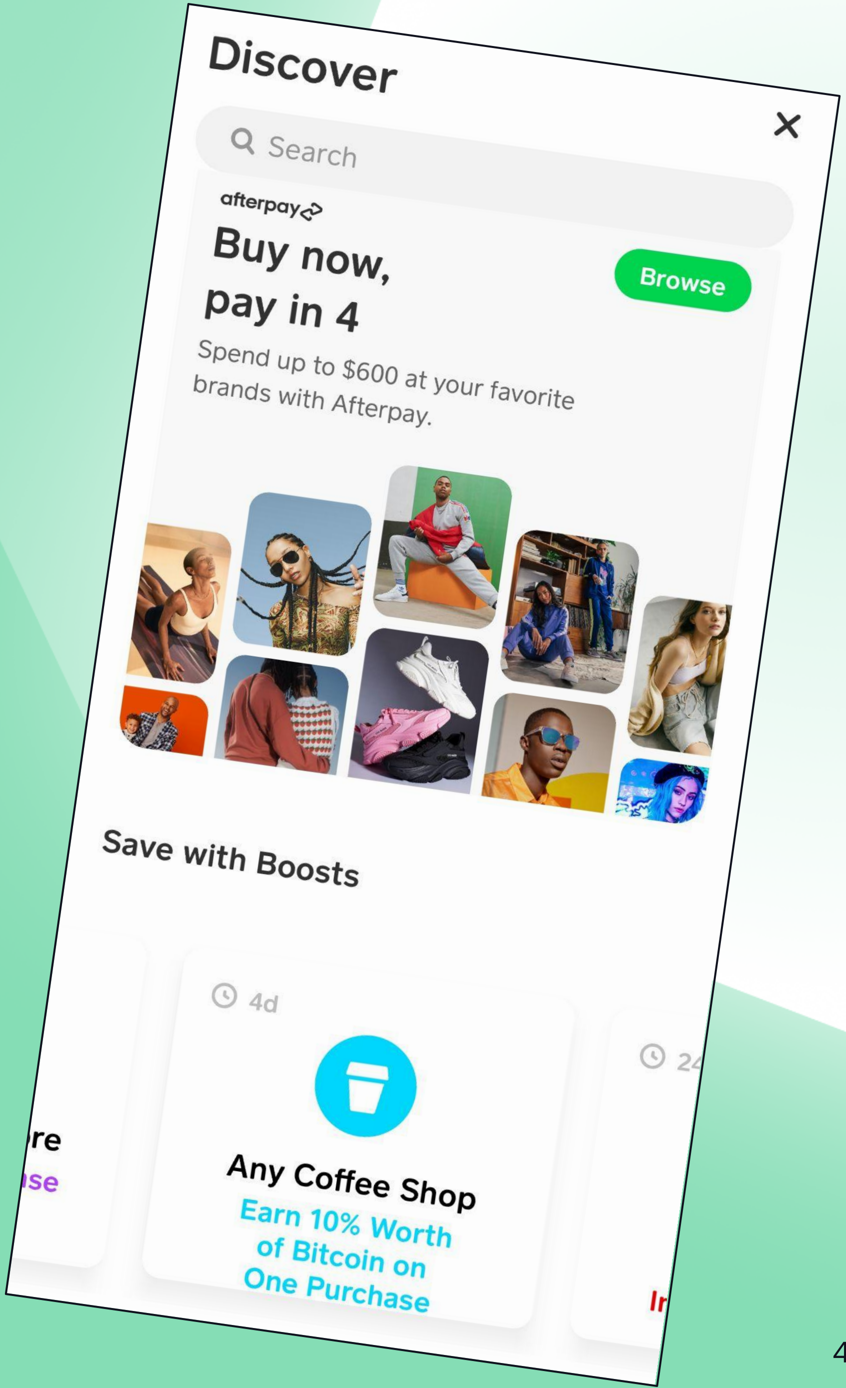
Features get easier = happier data scientists

More focus on business logic = happier engineers

Lower latency = happier users

If you think this is cool, join us and see more!

cash.app/careers



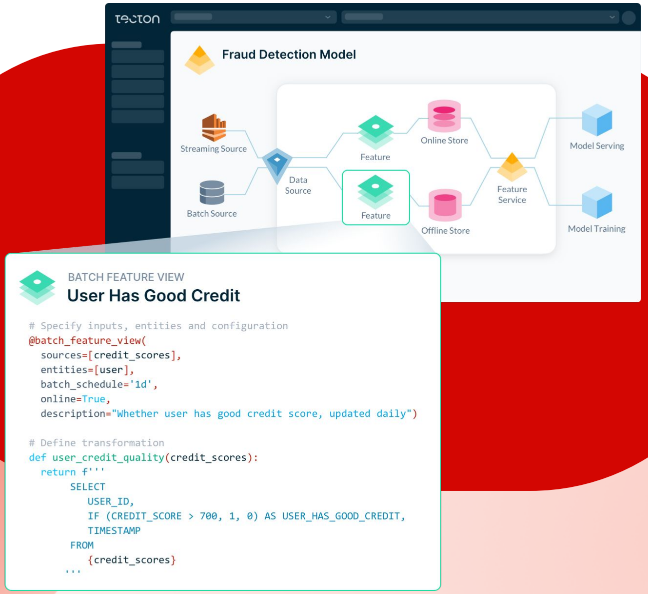
tecton

Visit booth 127

for a chance to win
a Skydio 2+ drone!

Request a free trial

www.tecton.ai



DATA+AI SUMMIT 2022

Thank you