Feature stores and evaluation stores: better together

Josh Tobin

gantry.io, UC Berkeley, Full Stack Deep Learning, Former OpenAI

Machine Learning is now a product engineering discipline

Mac w a

Chip Huyen @chipro · Oct 12, 2020

Machine learning engineering is 10% machine learning and 90% engineering.

98 642 7.8K

Elon Musk @elonmusk

Replying to @chipro

Yeah

8:09 PM · Oct 12, 2020 · Twitter for iPhone

104 Retweets **18** Quote Tweets **5.4K** Likes

How did we get here?

ML analytics
2000s

ML hype
2010s

ML products
2020s?

- Simple models run offline on medium to large datasets to produce reports
- Value comes from incorporating model insights into decisions

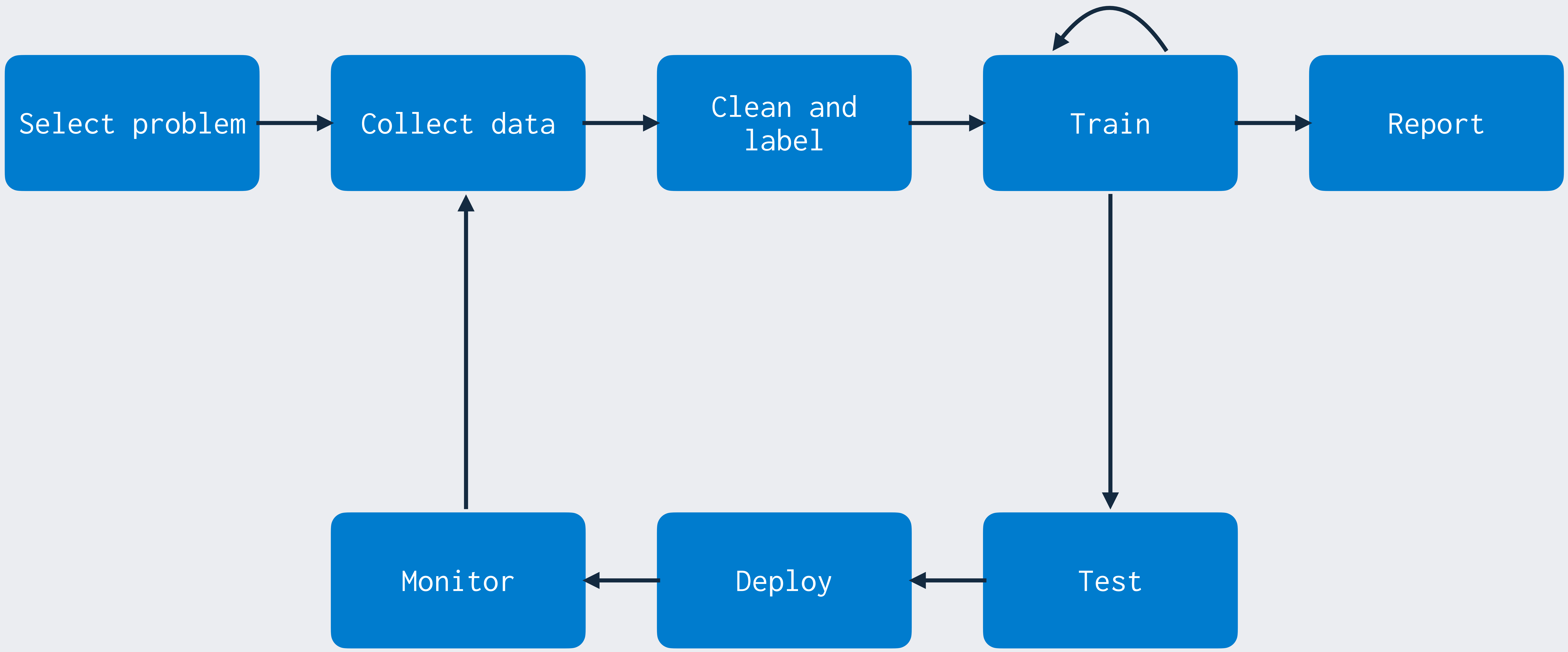
- Complicated models trained on massive datasets to produce papers
- Value comes from marketing potential of high-profile research output

- Reproducibility, scalability, and maintainability over complexity
- Value comes from models improving the business's products or services

ML products require a fundamentally new process

"Flat-earth" ML

ML products require a fundamentally new process



ML Product Engineering

ML teams that don't make the transition die

Uber sells ATG self-driving business to Aurora at $4 billion

By Krystal Hu, Tina Bellon, Jane Lanhee Lee

3 MIN READ

Montreal startup Element AI Inc. was running out of money and options when it inked a deal last month to sell itself for US$230-milion to Silicon Valley software company ServiceNow Inc., a confidential document obtained by the Globe and Mail reveals.

TECH • OPENAI

Buzzy research lab OpenAI debuts first product as it tries to live up to the hype

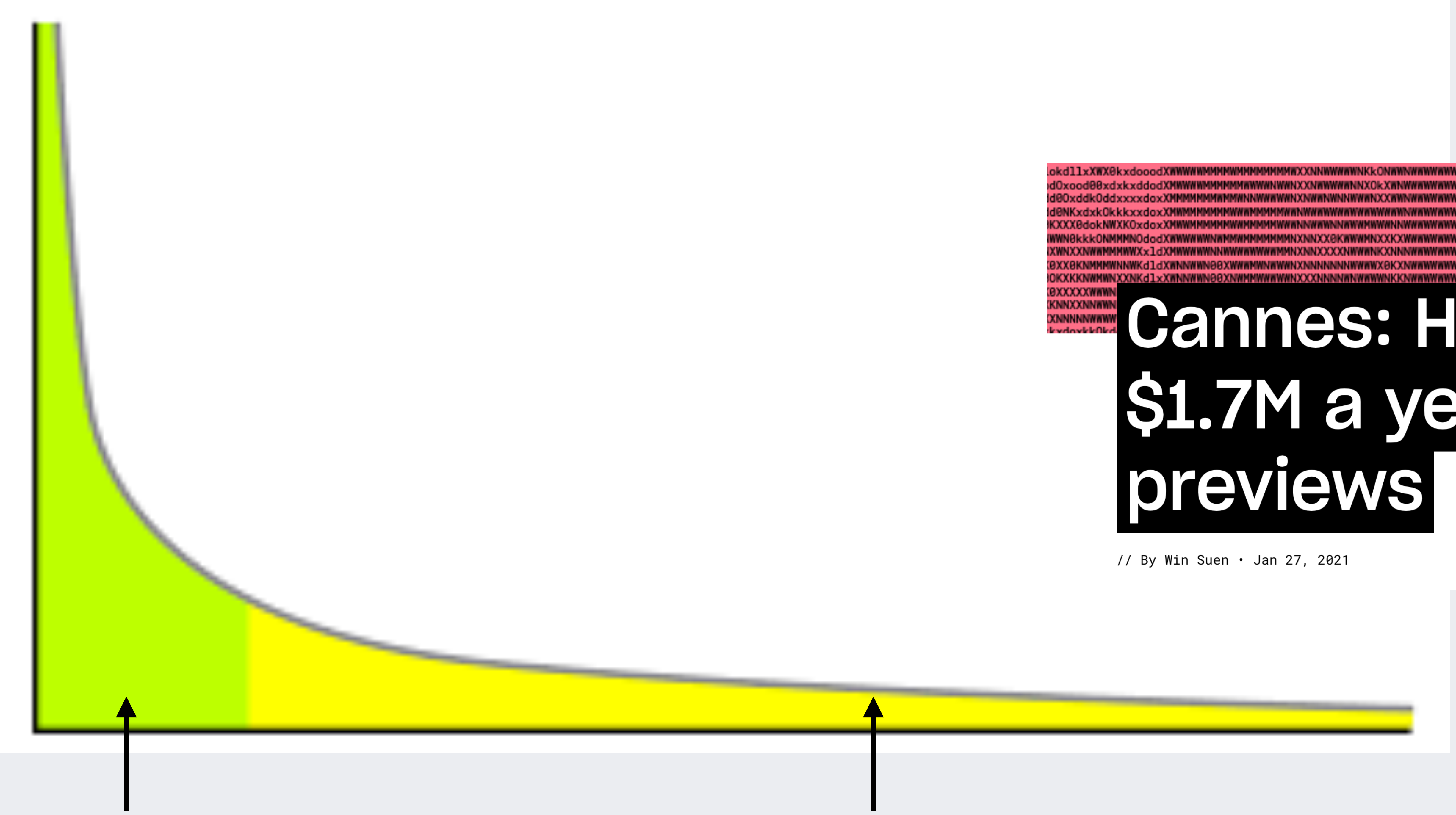
BY JONATHAN VANIAN
June 11, 2020 8:00 AM PDT

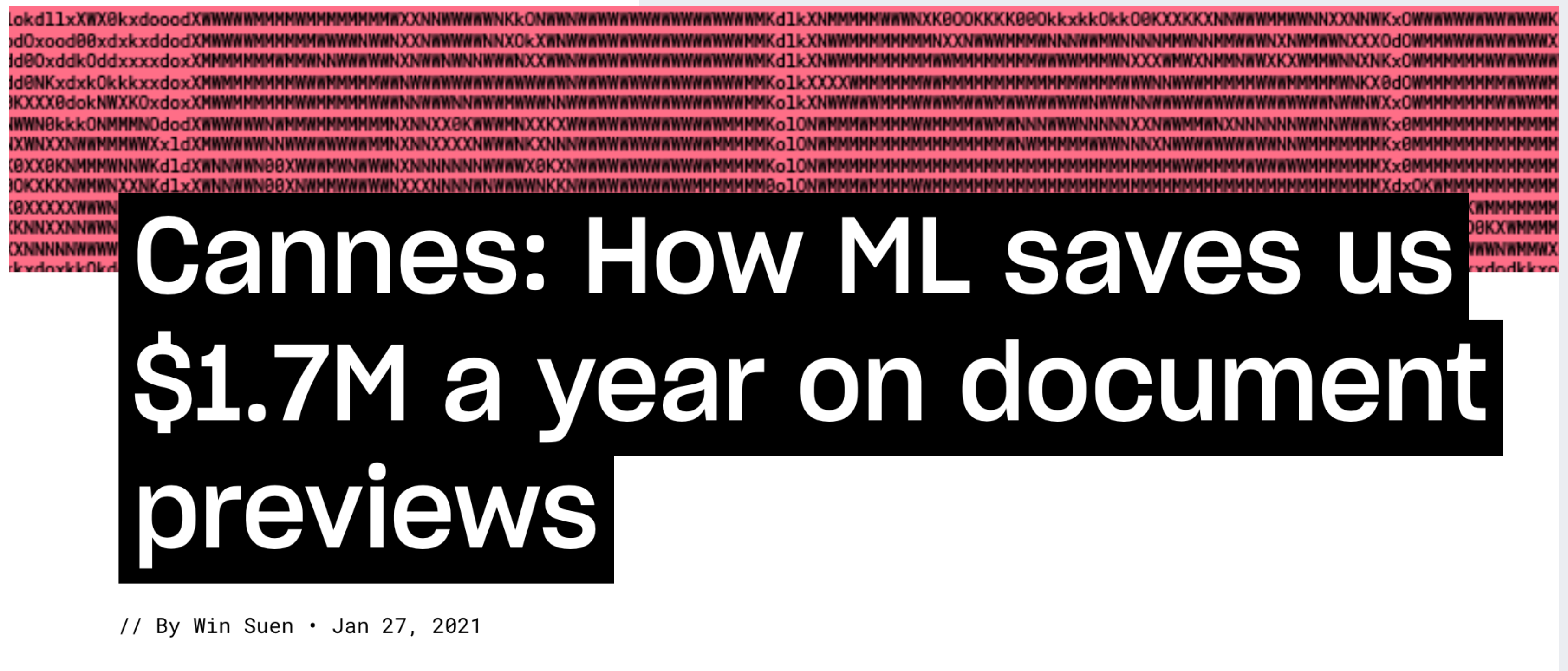
Of the 250 industrial firms Plutoshift surveyed,

- **over 72% found that they had taken far more time than anticipated** to implement the necessary data collection processes for applying machine learning.
- and perhaps as a result, **only 17% of those surveyed said they were actually at the full implementation stage of using A.I.**,
- while **about 70% said they were still studying what resources they'd need**, assessing possible business use cases, or conducting small pilot projects only.

Worryingly, **almost 20% of companies cited "peer pressure" as the reason they had embarked on A.I. projects**.

Those that make the transition will create amazing things





- Autonomous Vehicles
- Real-time translation
- Drug discovery

- Marketing automation
- Personalization
- Document understanding
- Etc

Unlike flat-earth ML, ML products often:

- Run online and in real-time
- Deal with constantly evolving data distributions
- Handle messy, long-tail real world data
- Make predictions autonomously or semi-autonomously

This implies new ops & infra demands

- Run online and in real-time
  Host and serve models with low latency

- Deal with constantly evolving data distributions
  Retrain models frequently, even continuously

- Handle messy, long-tail real world data
  Inspect your data scalable, manage slices and edge cases

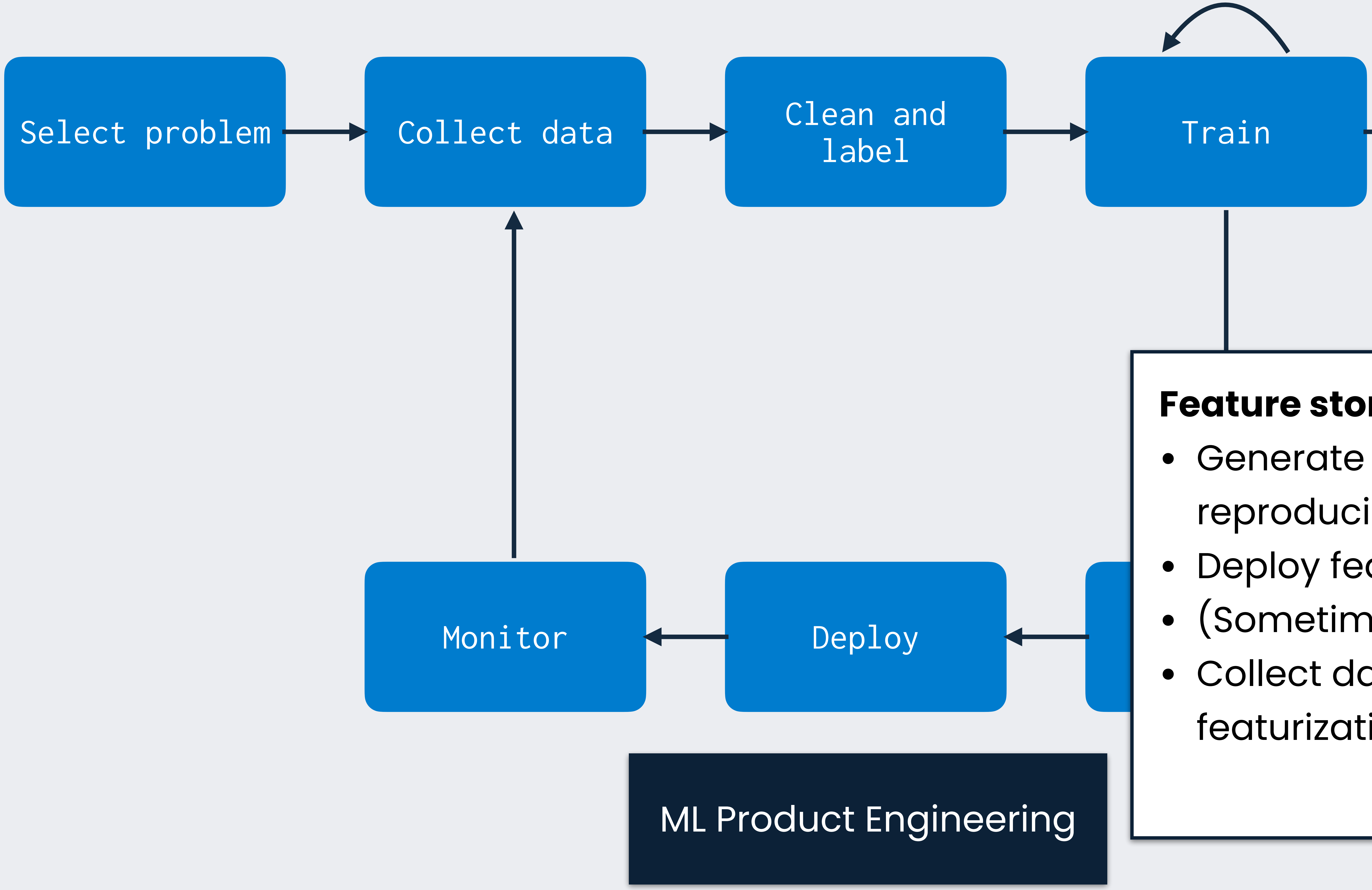
- Make predictions autonomously or semi-autonomously
  Quickly catch and diagnose bugs and distribution changes

How do feature stores fit in?

What does a feature store actually do?

- Define features consistently online and offline
- Make features available with low latency online
- (Sometimes) allow you to share features across the org
- (Sometimes) monitor features for drift and training-serving skew

How do feature stores fit in?



Feature stores

- Generate training sets efficiently & reproducibly
- Deploy features
- (Sometimes) monitor features
- Collect data & run it through the featurization pipeline

What don’t feature stores help with?

- Training & deploying models
- Deciding whether a model is good enough to deploy
- Deciding whether the new model is really better than the old one
- Deciding when a model needs to be retrained
- Deciding what data to collect, label, and retrain on

Ops

How you use your infrastructure to build better ML systems

Infrastructure

Tools to move models and data through their operational lifecycle

Feature store

Model training

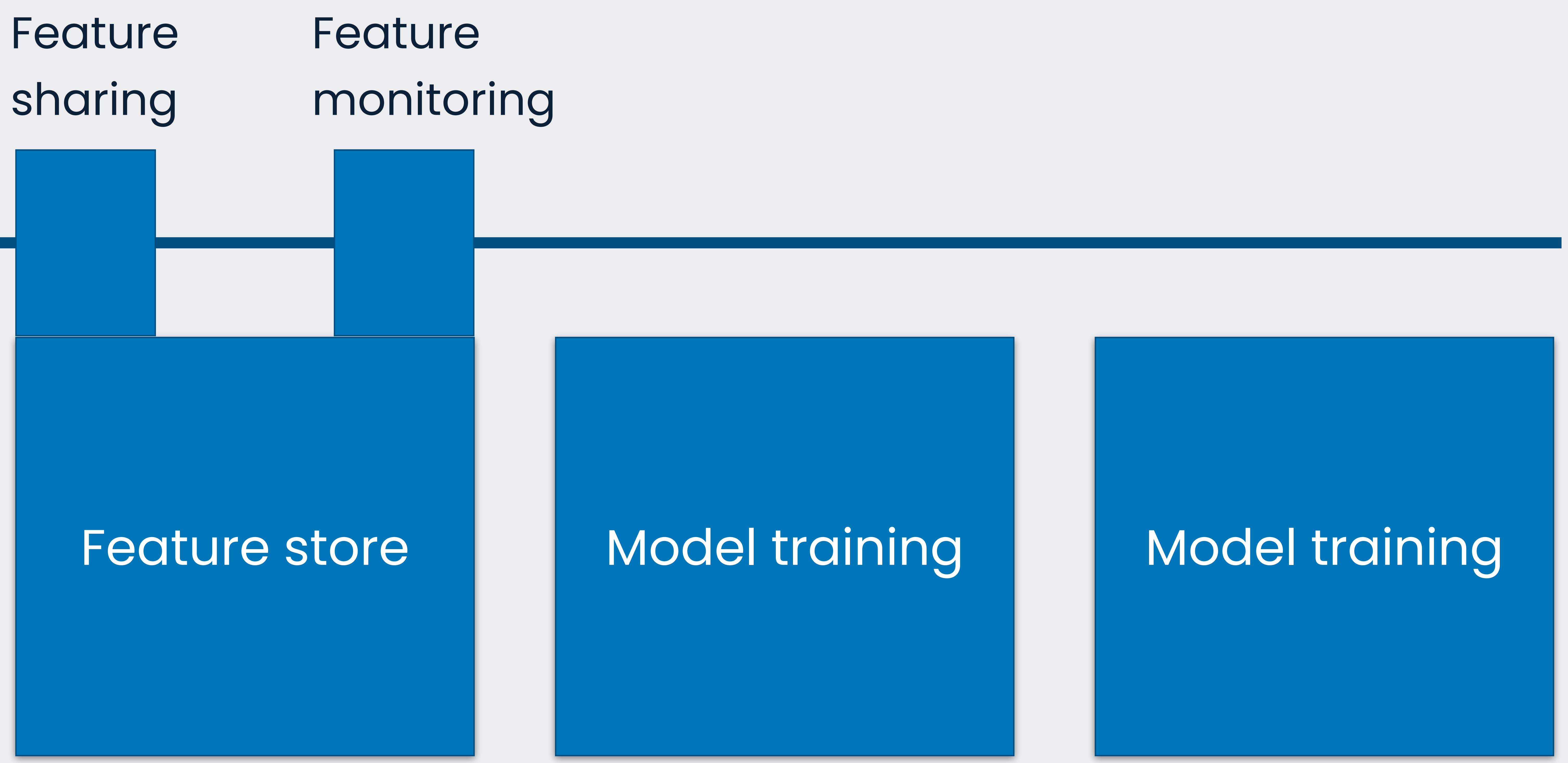
Model training

Ops

How you use your infrastructure to build better ML systems

Infrastructure

Tools to move models and data through their operational lifecycle



What don't feature stores help with?

- Training & deploying models
- Deciding whether a model is good enough to deploy
- Deciding whether the new model is really better than the old one
- Deciding when a model needs to be retrained
- Deciding what data to collect, label, and retrain on

What is an evaluation store?

The Evaluation Store

A central place to store and query **online and offline** *ground truth* and *approximate* **model quality metrics**

Data sources
Model training & evaluation
Production ML deployment
Labeling service
User-facing application
Business metrics
Raw inputs, predictions & feedback
GANTRY
Data storage
Data warehouse, data lake, feature store
Evaluation metrics for all data slices
Analysis & operational decisions
Exploration, debugging & reporting
Dashboard
SDK
Alerting
Pagerduty, Slack, email, etc
Workflows
• Trigger a retraining
• Label data from a particular slice
• Run an AB test
• Generate new test cases, etc

What could an eval store help you with?

- **Reduce organization friction.** Get stakeholders (ML eng, ML research, PM, MLOps, etc) on the same page about metric and slice definitions
- **Deploy models more confidently.** Evaluate metrics and slices consistently in testing and prod. Make the metrics visible to stakeholders
- **Catch production bugs faster.** Catch degradations across any slice, and drill down to the data that caused the degradation
- **Reduce data-related costs.** Collect and label production data more intelligently
- **Make your model better.** Decide when to retrain. Pick the right data to retrain on.

Querying the evaluation store

What form do queries take?

- Subset of models in the store
- Subset of metrics in the store
- Subset of slices in the store
- Specification of the window of data

Querying the evaluation store

What form do queries take?

- Subset of models in the store
- Subset of metrics in the store
- Subset of slices in the store
- Specification of the window of data

E.g.,

What is the importance-weighted average drift across all of my features in my production model in the last 60 minutes?

Monitoring

Querying the evaluation store

What form do queries take?

- Subset of models in the store
- Subset of metrics in the store
- Subset of slices in the store
- Specification of the window of data

E.g.,

How much worse is the my accuracy in the last 7 days than it was during training?

Monitoring

Querying the evaluation store

What form do queries take?

- Subset of models in the store
- Subset of metrics in the store
- Subset of slices in the store
- Specification of the window of data

E.g.,

How do all of the metrics compare for model A and model B across all slices in my main evaluation set?

Testing

Querying the evaluation store

What form do queries take?

- Subset of models in the store
- Subset of metrics in the store
- Subset of slices in the store
- Specification of the window of data

E.g.,

How do my business metrics compare for model A and model B in the last 60 minutes

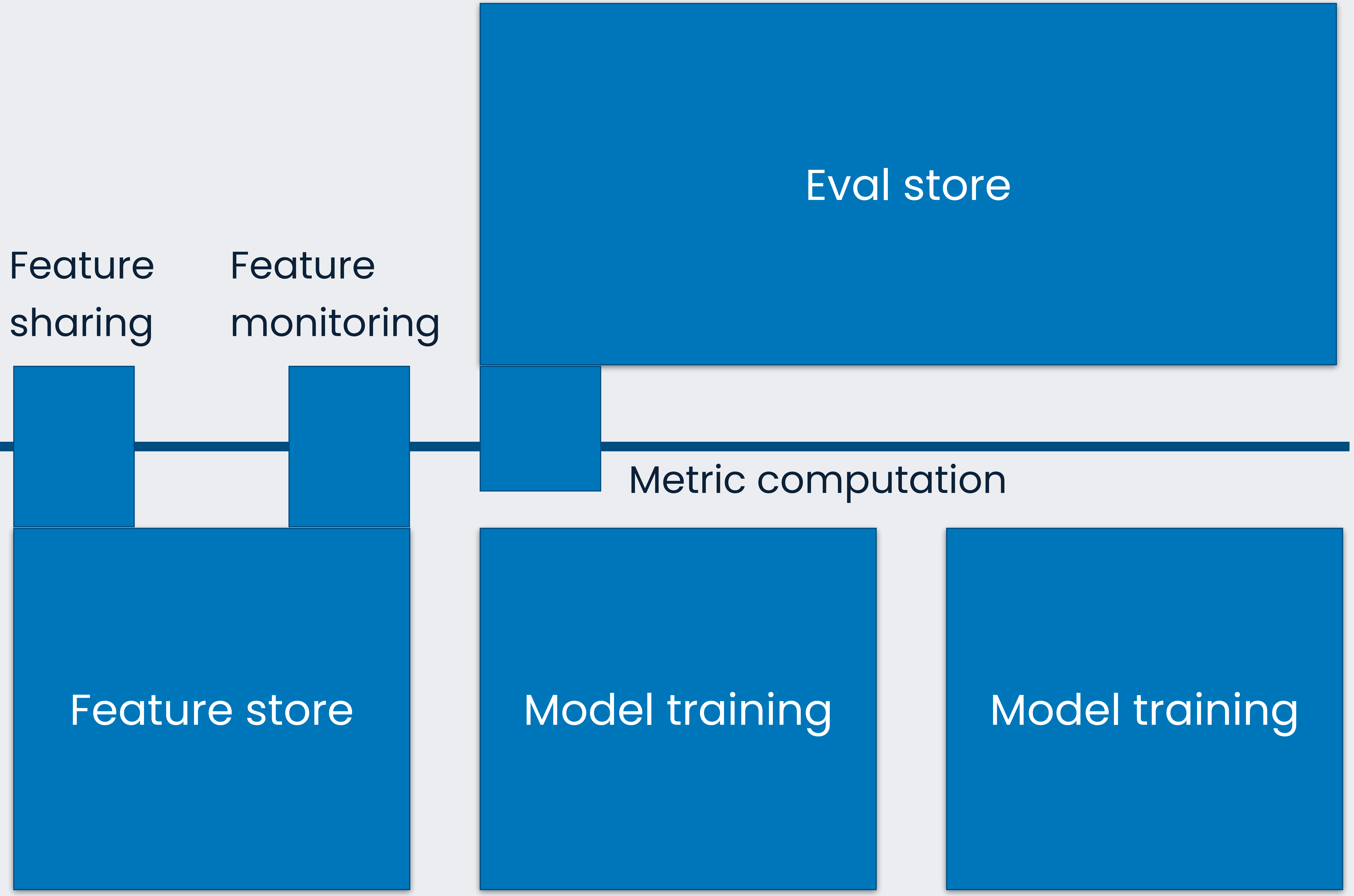
AB testing

Ops

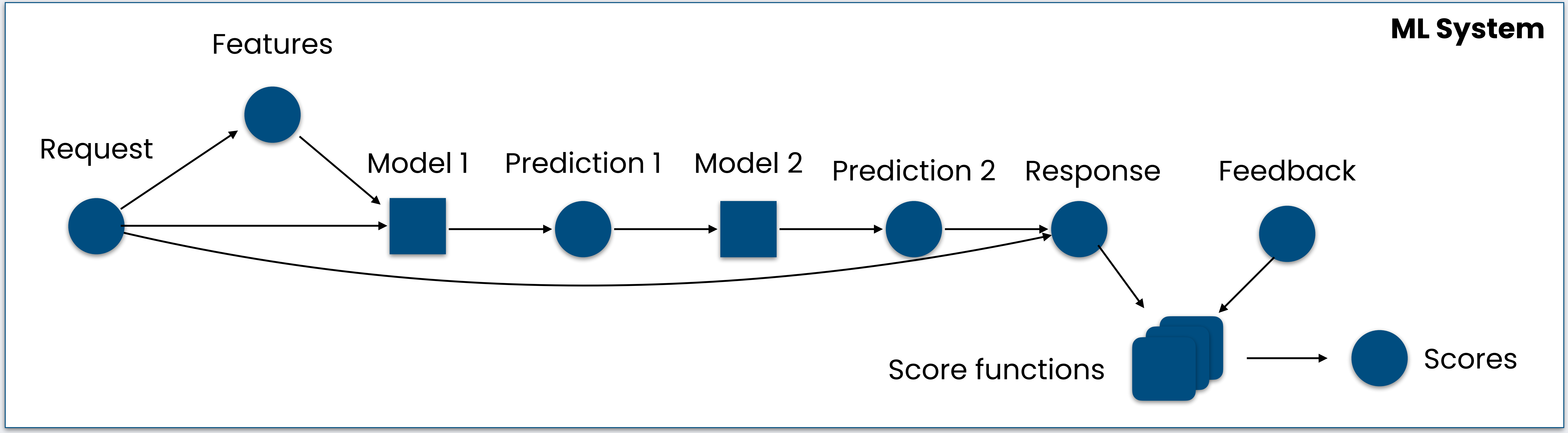
How you use your infrastructure to build better ML systems

Infrastructure

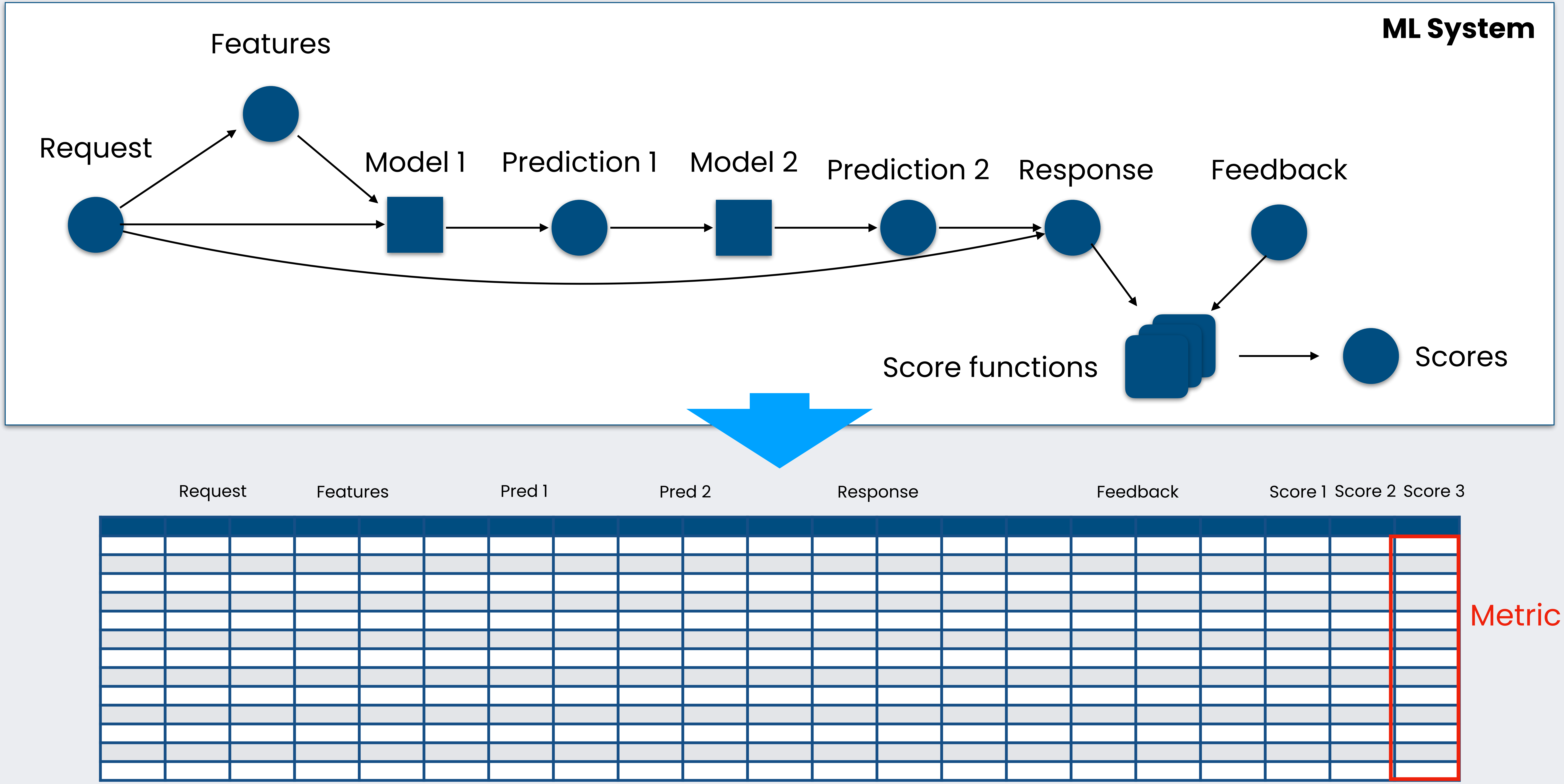
Tools to move models and data through their operational lifecycle



How does it work?



How does it work?



Shouldn't the feature store do this?

- Not all important data will pass through the feature store
  - Business metrics
  - Metadata that is useful for slicing
  - Images / text / etc
- Monitoring all of the features != monitoring the model
  - Performance drift is more important than feature drift
  - A "poor quality" feature has different effects on different models
  - Even if the features change, performance need not
  - How to disambiguate performance across the full ML pipeline?
- Different tooling and query patterns

Feature stores and eval stores should work together

- Eval stores are ops tools with their toes in the infra world. Feature stores are infra tools with their toes in the ops world.
- Eval stores should leverage feature stores for storage and querying of raw feature data
- Feature stores focus on features. Eval stores bring the context of the model and broader ML system, and can help your customers use the feature store more effectively

Continuous learning starts with continuous evaluation