Executive Briefing: What it takes to use machine learning in fast data pipelines

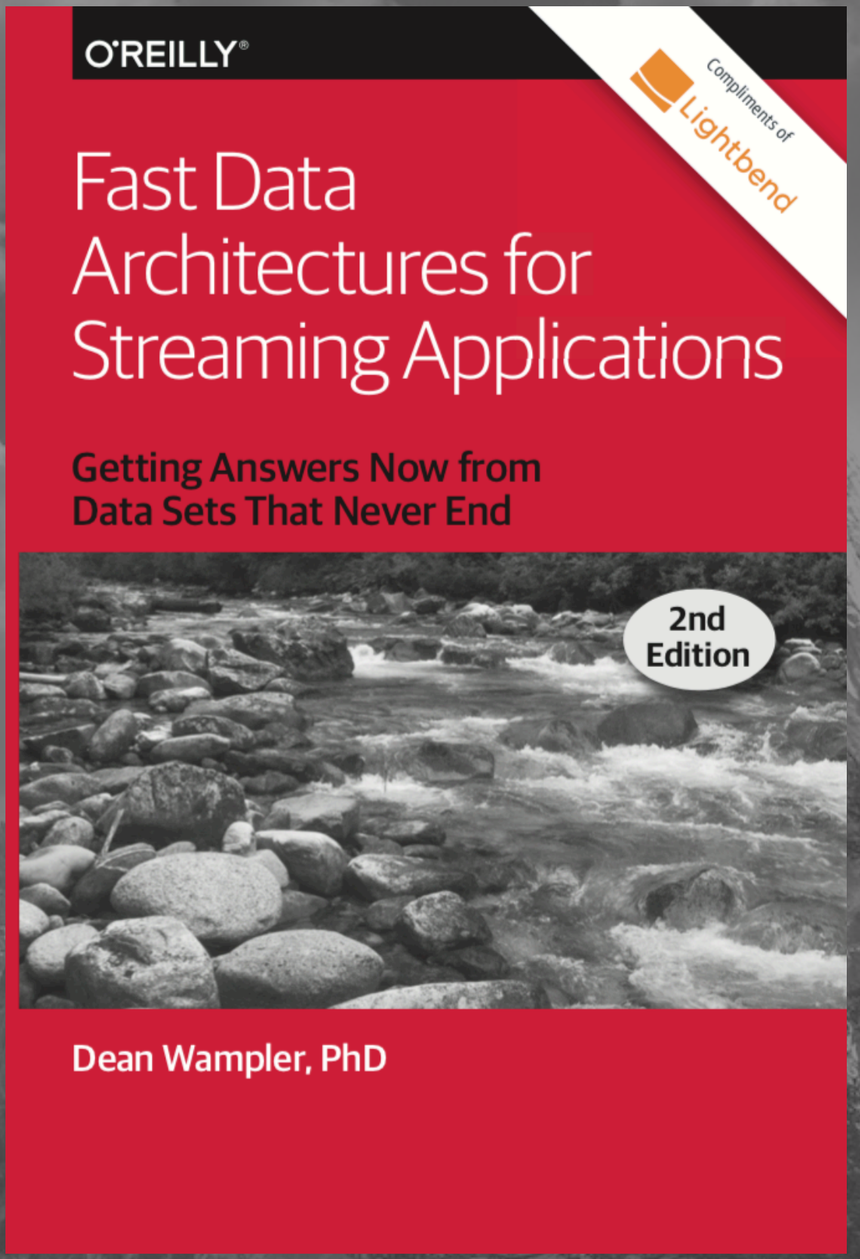
Dean Wampler, Ph.D.
dean@lightbend.com
polyglotprogramming.com/talks

Lightbend

@deanwampler

Data Streaming, in General

lbnd.io/fast-data-ebook



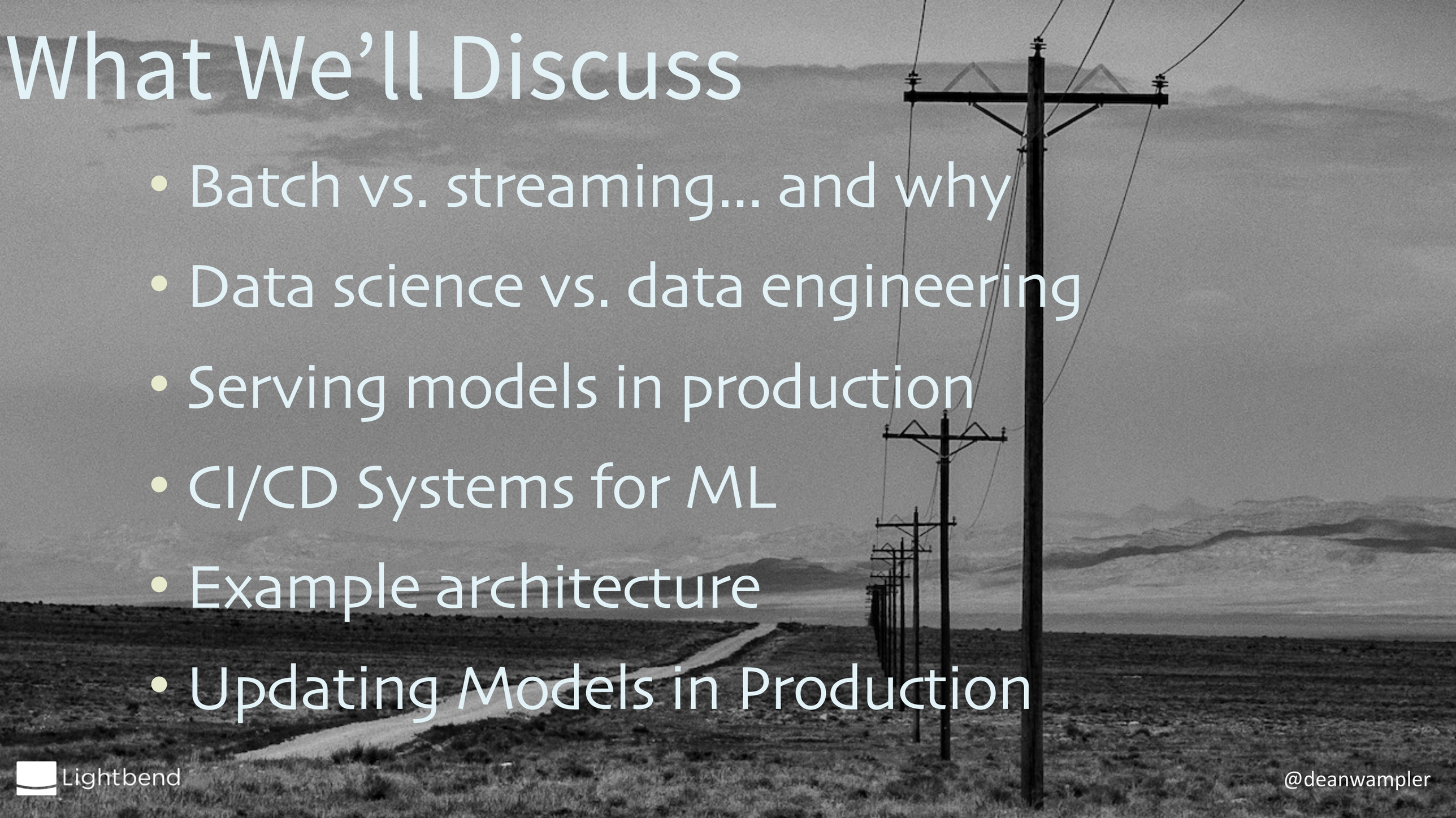
Lightbend

@deanwampler

lightbend.com/pipelines

Lightbend Pipelines

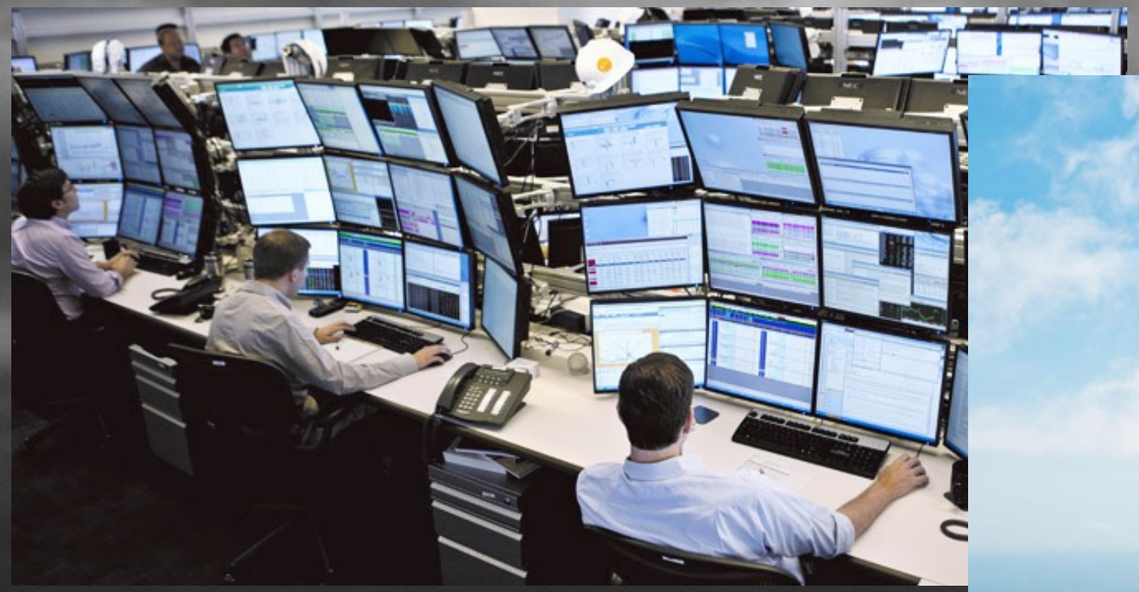
What We'll Discuss



- Batch vs. streaming... and why
- Data science vs. data engineering
- Serving models in production
- CI/CD Systems for ML
- Example architecture
- Updating Models in Production

Batch vs. streaming… and why

@deanwampler



Finance

Telecom

Medical

Energy

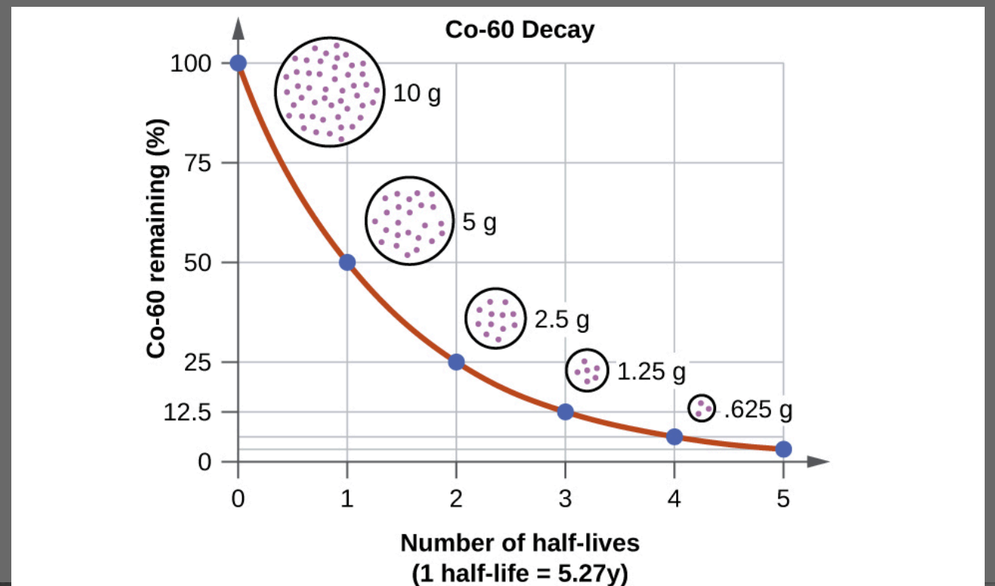
... and IoT

State of the art phone!

Mobile

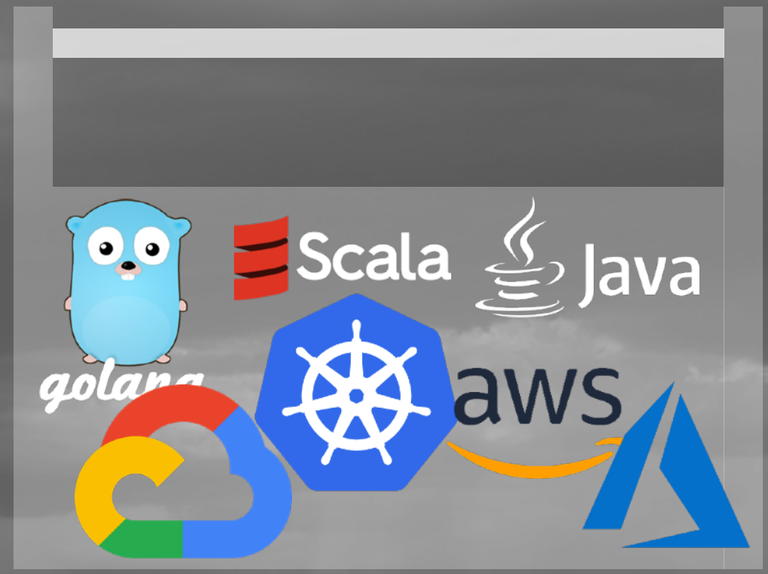
@deanwampler

Information value has a half life; it decays with time




Data Science vs.
Data Engineering
Lightbend
@deanwampler

Data Science toolbox



Software Engineering toolbox

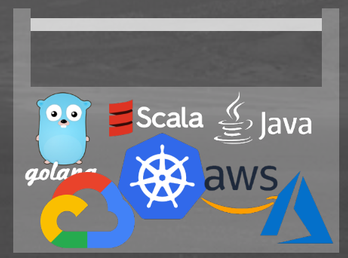
Data Scientists

- Comfortable with uncertainty
- Less process oriented
  - Iterative, experimental

Data Engineers

- Uncomfortable with uncertainty
- Process oriented
  - Agile Manifesto
  - … which does not mention data!

https://derwen.ai/s/6fqt

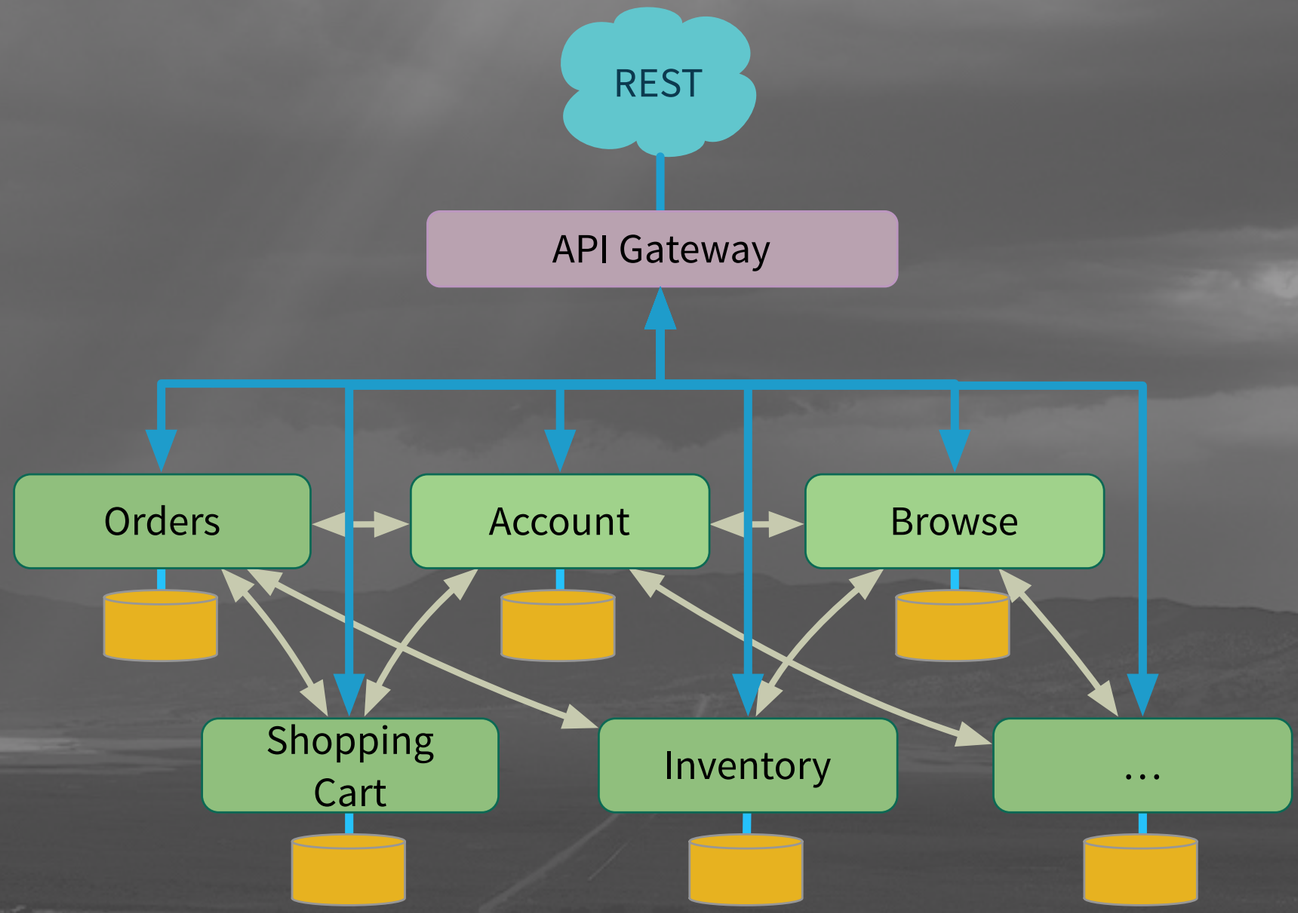


Streaming Imposes New Requirements

If you run something long enough, all rare problems eventually happen!

- Reliability - fault and "surprise" tolerant
- Availability - "always on"
- Low latency - for some definition of "low"
- Scalability - up and down
- Adaptability - ideally without restarts

Reminds Me of Microservices

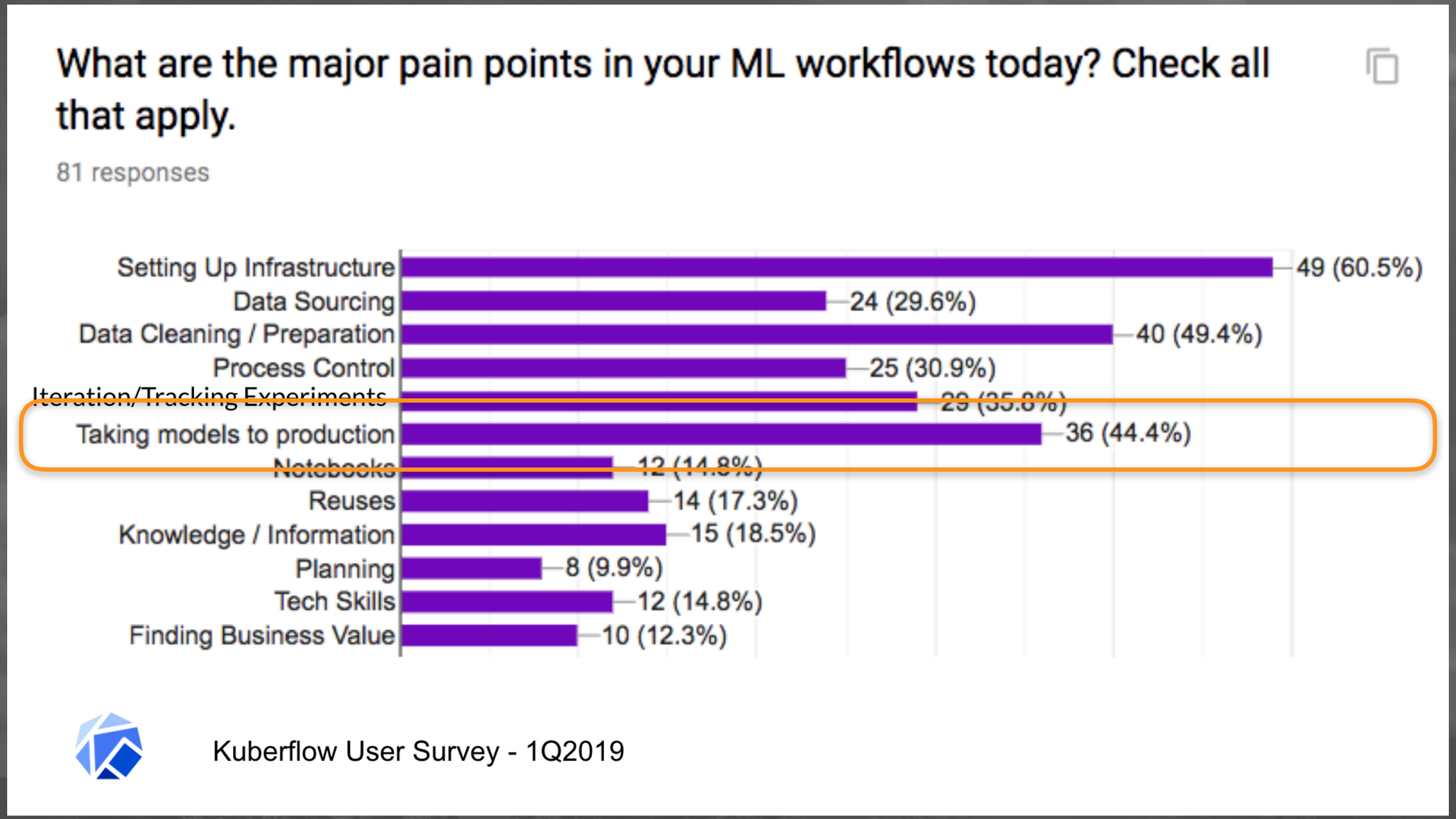


Serving Models in Production

Lightbend

@deanwampler

A Recent Kubeflow User Survey



From this Kubeflow Overview

Lack of Tool/Process Integration

- ~60% worry about missed opportunities
- ~50% worry about loss of data team productivity
- ~45% worry about slow time-to-market
- ~40% worry about customer dissatisfaction

From a recent Lightbend survey

Can You Answer this Question?

- Why did the model reject that loan application?

First, which model was it?

- Which version of the model was used?
- How was it trained?
- When was this model deployed?
- …

CI/CD Process Required (1/3)

- Version control - for models and code
- Automation - builds, tests, other quality checks, artifact management & delivery
  - Necessary for reproducibility

CI/CD Processes Required (2/3)

- Auditing
  - Which model used to score this record?
  - Which records used to train this model?
  - Who accessed this model and when?

CI/CD Processes Required (2/3)

- Auditing
  - Which model used to score this record?
  - Which records used to train this model?
  - Who accessed this model and when?

Models Are Data

CI/CD Processes Required (2/3)

GDPR - What if a customer asks you to delete their data? Do you also delete the models trained with that data?

- Auditing
  - Which model used to score this record?
  - Which records used to train this model?
  - Who accessed this model and when?

CI/CD Processes Required (3/3)

- Monitoring
  - Resource utilization changes?
  - Quality metrics:
    - Match performance during training?
    - Concept drift?

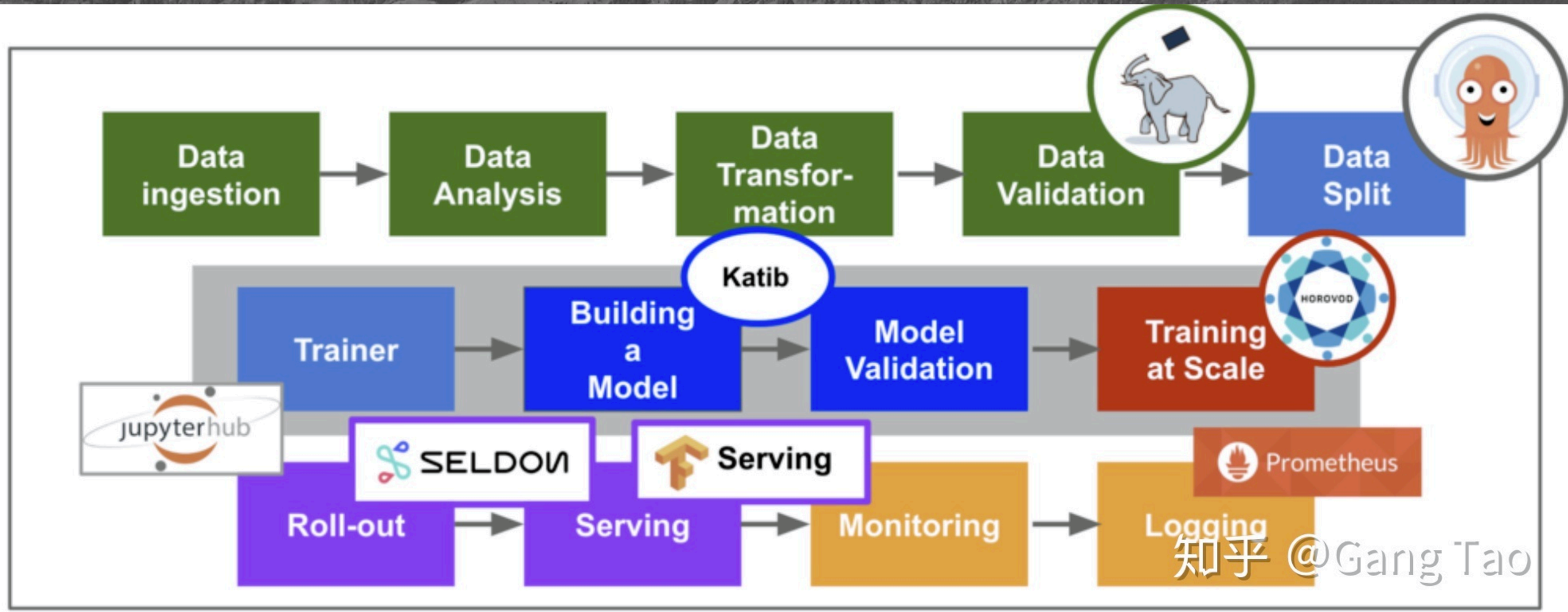
CI/CD Systems for ML?

Lightbend

@deanwampler

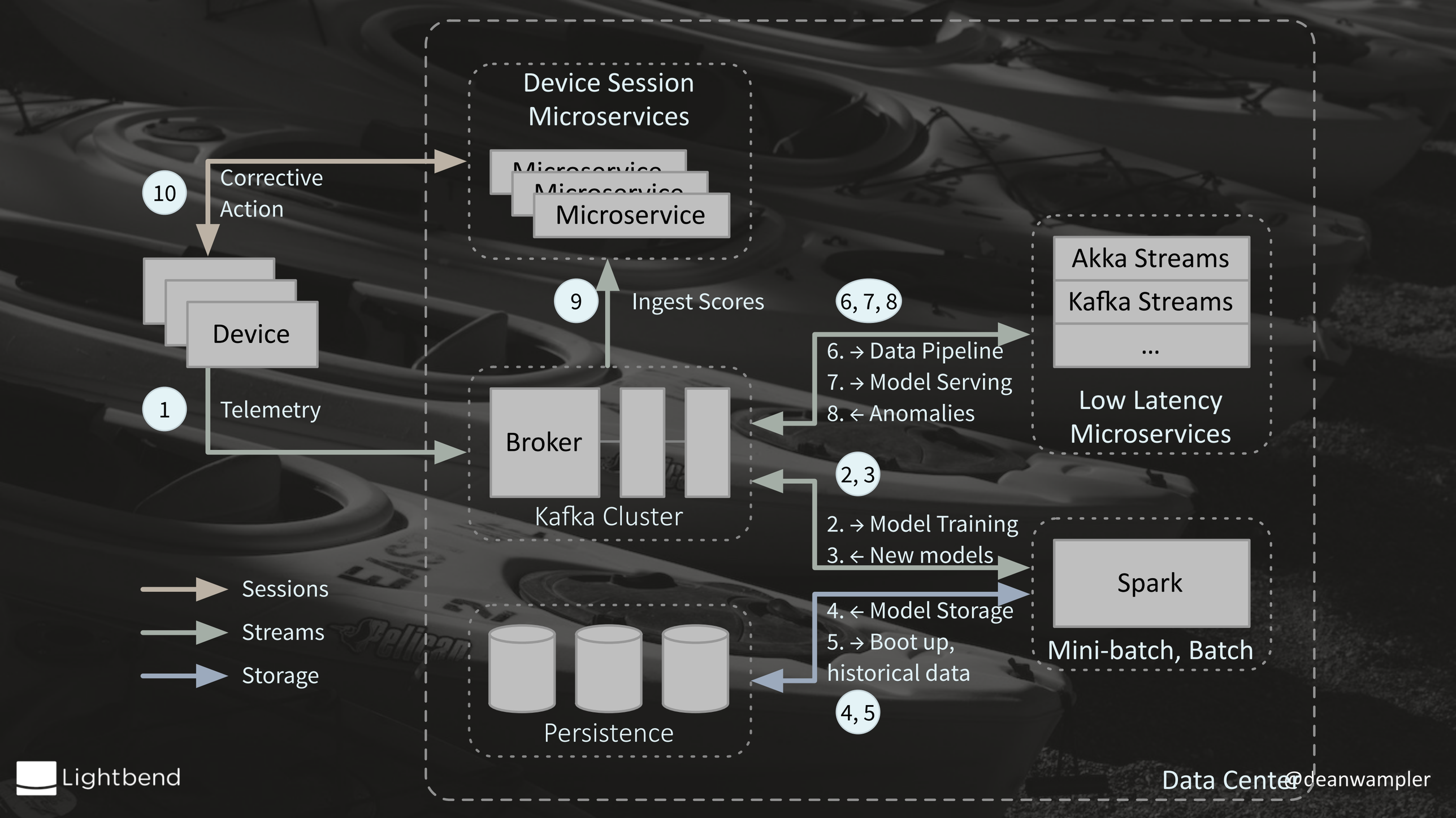

• Kubeflow - for Kubernetes
• SageMaker - for AWS users
• MLFlow - from the Spark community
• … plus emerging vendors

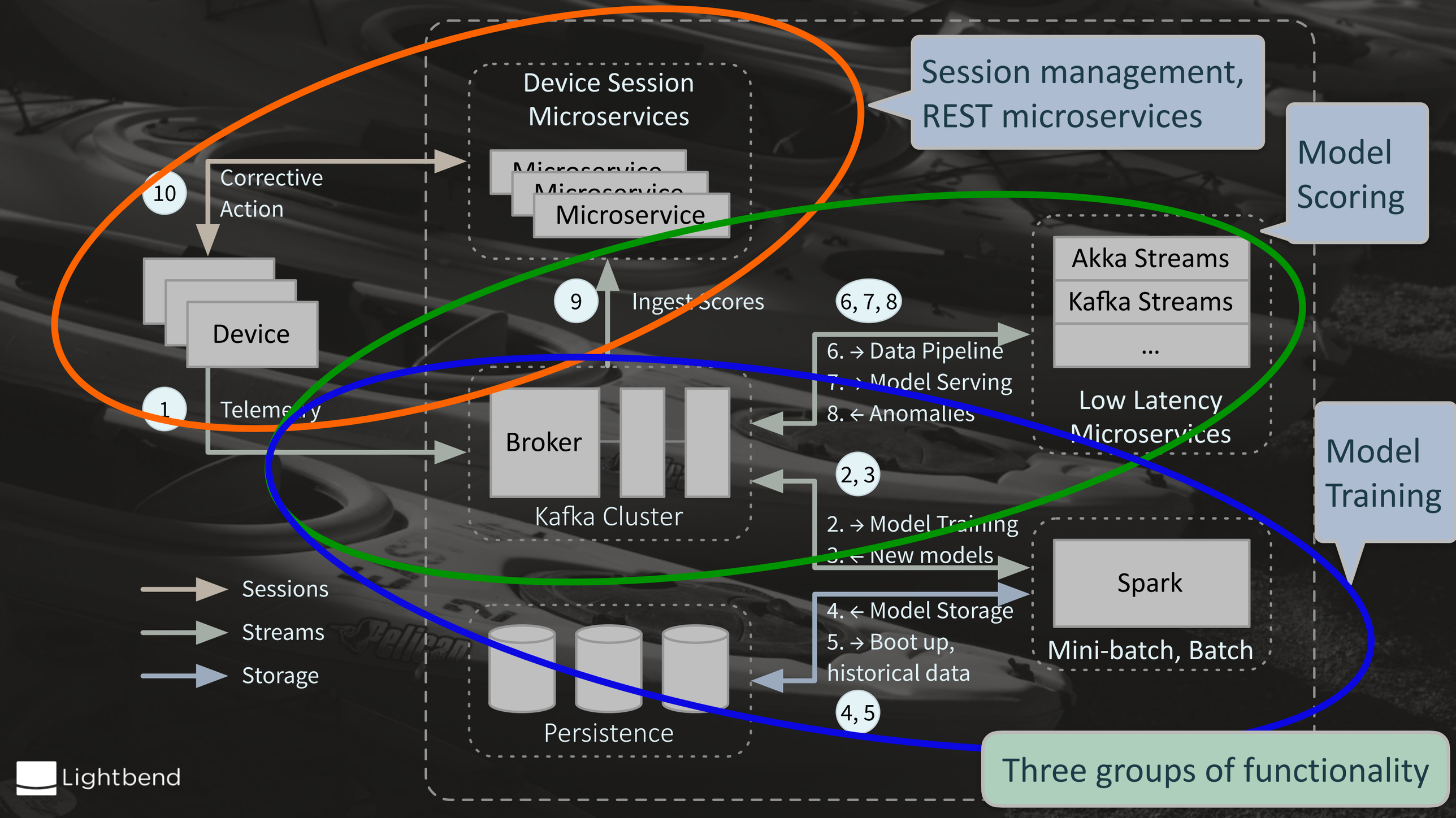
Systems - Kubeflow

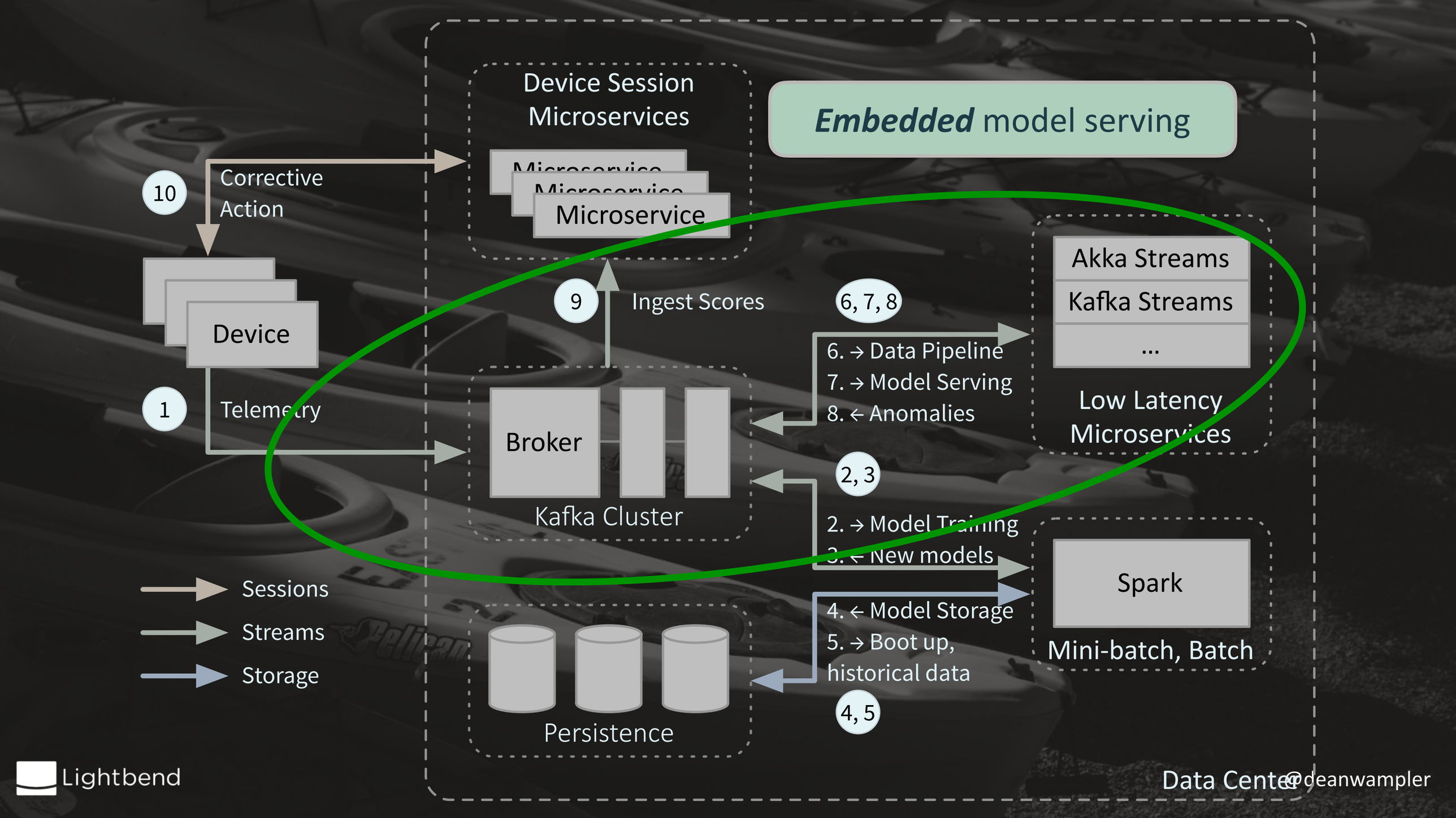


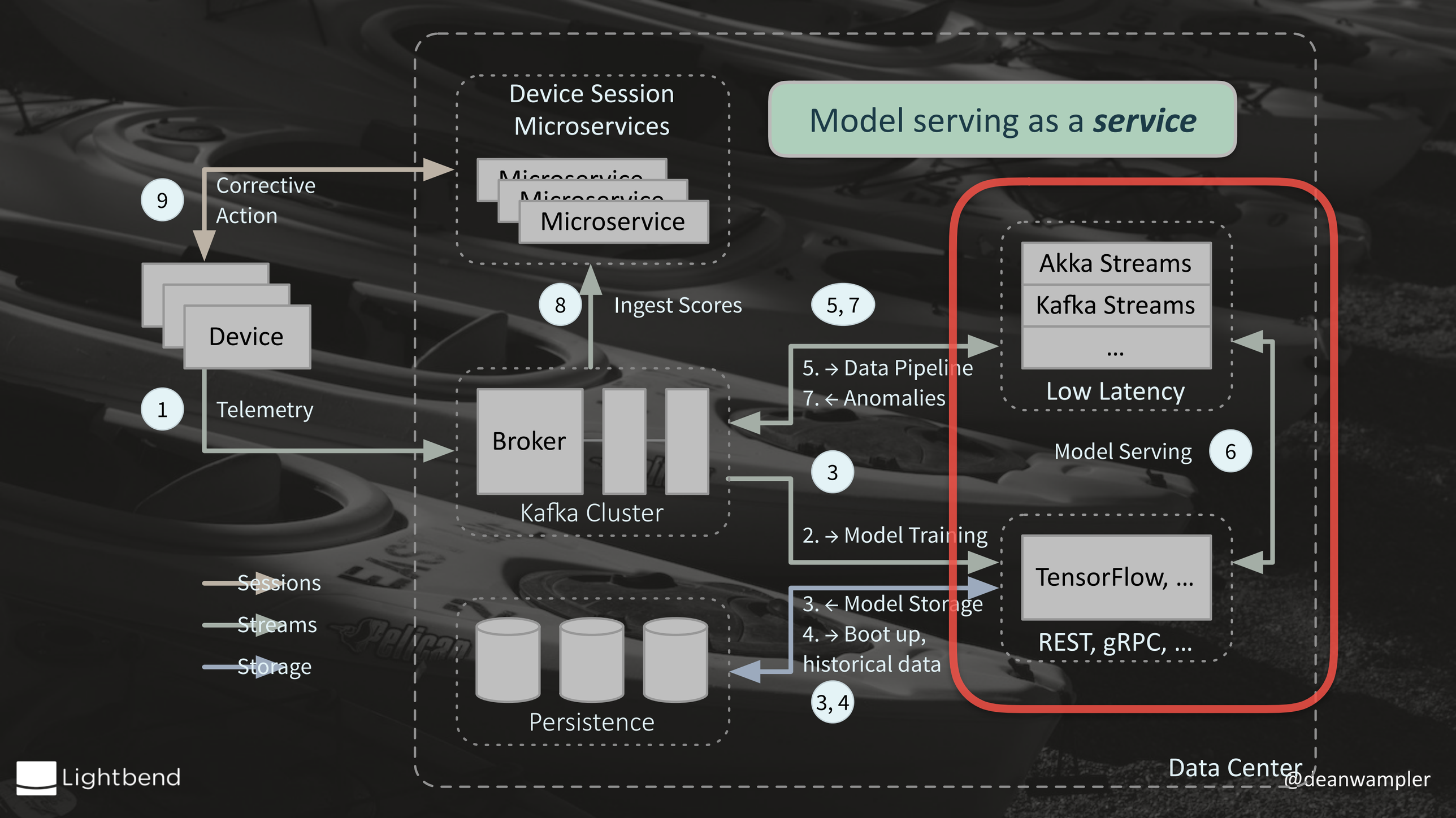
An Example

An Example

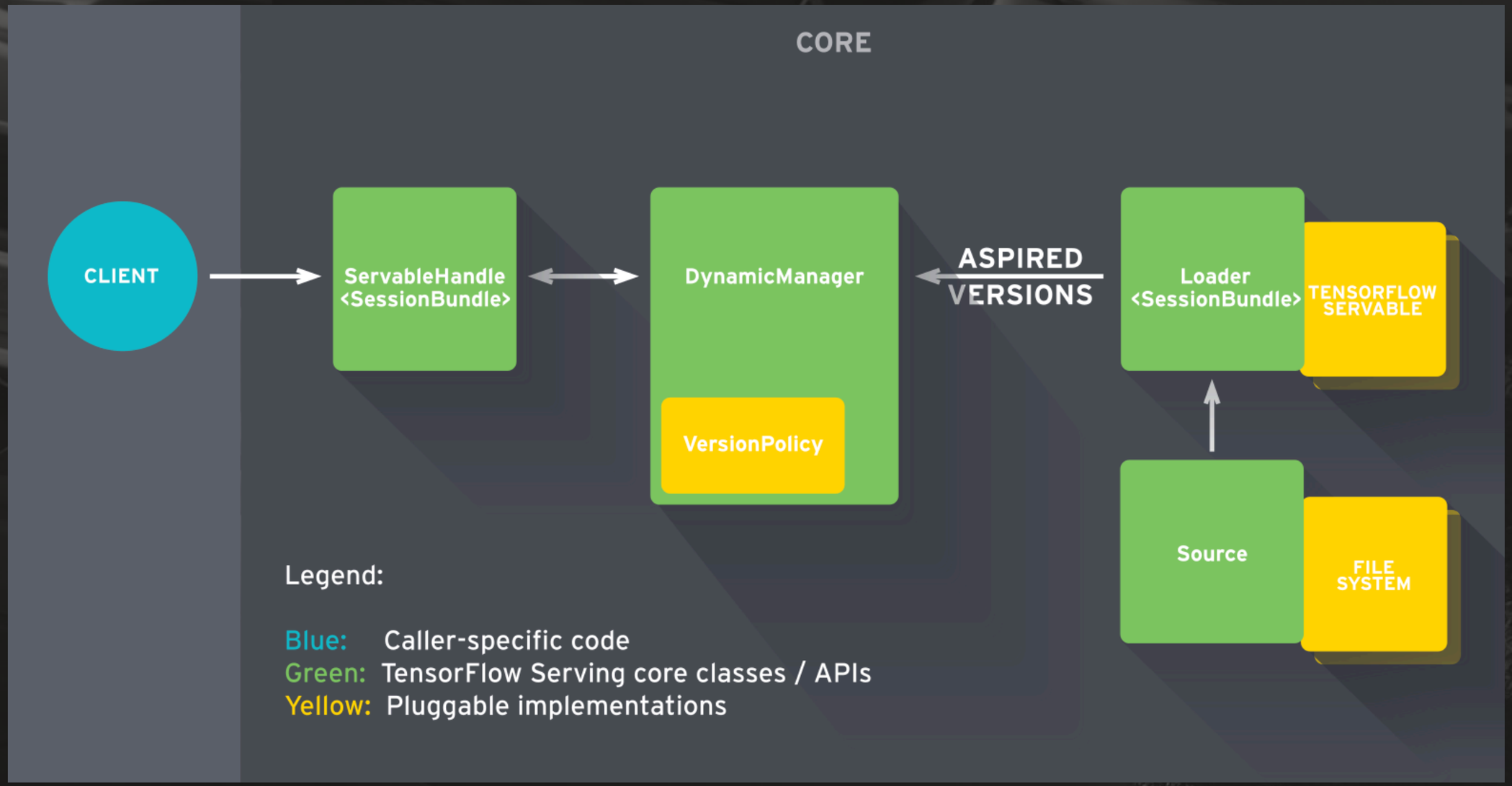

Device Session Microservices
Microservice
10
Corrective Action
Device
1
Telemetry
9
Ingest Scores
Broker
Kafka Cluster
6, 7, 8
6. → Data Pipeline
7. → Model Serving
8. ← Anomalies
Akka Streams
Kafka Streams
...
Low Latency Microservices
2, 3
2. → Model Training
3. ← New models
4. ← Model Storage
5. → Boot up, historical data
4, 5
Spark
Mini-batch, Batch
Sessions
Streams
Storage
Persistence
Data Center
Lightbend
@deanwampler


Session management, REST microservices
Model Scoring
Model Training
Device Session Microservices
Microservice
Microservice
Microservice
10
Corrective Action
Device
1
Telemetry
9
Ingest Scores
6, 7, 8
6. → Data Pipeline
7. → Model Serving
8. ← Anomalies
Akka Streams
Kafka Streams
...
Low Latency Microservices
Broker
Kafka Cluster
2, 3
2. → Model Training
3. ← New models
4. ← Model Storage
5. → Boot up, historical data
Spark
Mini-batch, Batch
4, 5
Persistence
Sessions
Streams
Storage
Three groups of functionality
Lightbend


Device Session Microservices
Microservice
Microservice
Microservice
Embedded model serving
10
Corrective Action
Device
1
Telemetry
9
Ingest Scores
Broker
Kafka Cluster
6, 7, 8
6. → Data Pipeline
7. → Model Serving
8. ← Anomalies
Akka Streams
Kafka Streams
...
Low Latency Microservices
2, 3
2. → Model Training
3. ← New models
4. ← Model Storage
5. → Boot up, historical data
4, 5
Spark
Mini-batch, Batch
Persistence
Sessions
Streams
Storage
Data Center
Lightbend
@deanwampler

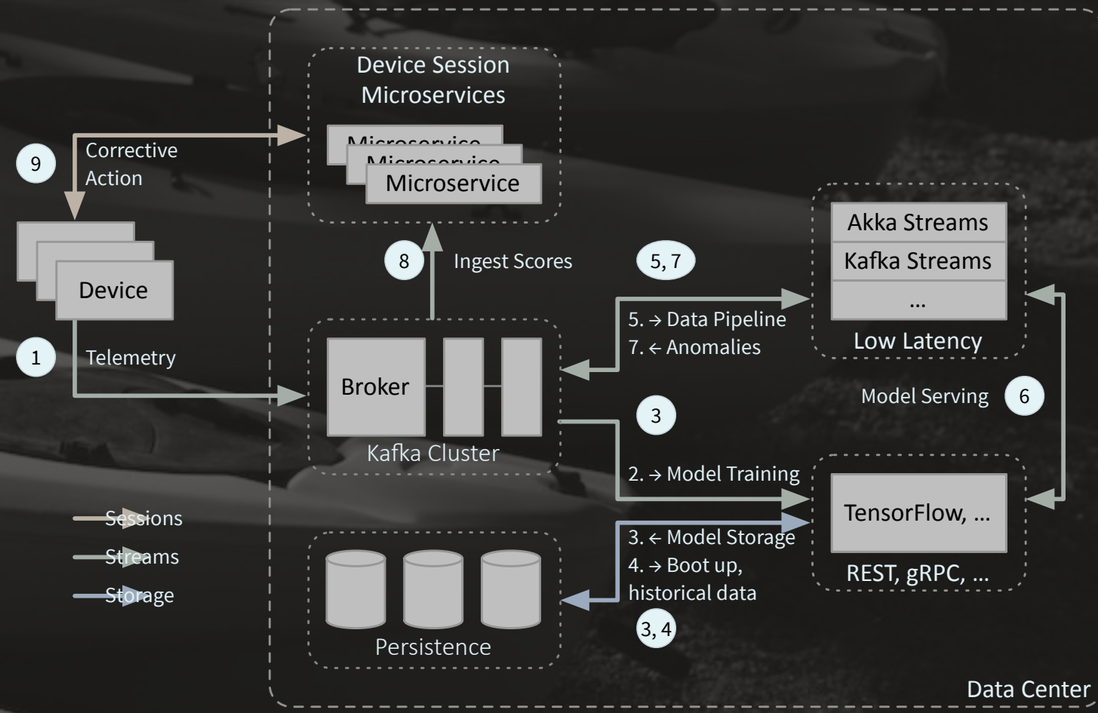

Model serving as a *service*
Device Session Microservices
Microservice
Microservice
Microservice
9
Corrective Action
Device
1
Telemetry
8
Ingest Scores
Broker
Kafka Cluster
5, 7
5. → Data Pipeline
7. ← Anomalies
3
2. → Model Training
3. ← Model Storage
4. → Boot up, historical data
3, 4
Persistence
Akka Streams
Kafka Streams
...
Low Latency
Model Serving
6
TensorFlow, ...
REST, gRPC, ...
Sessions
Streams
Storage
Data Center
Lightbend
@deanwampler

TensorFlow Serving



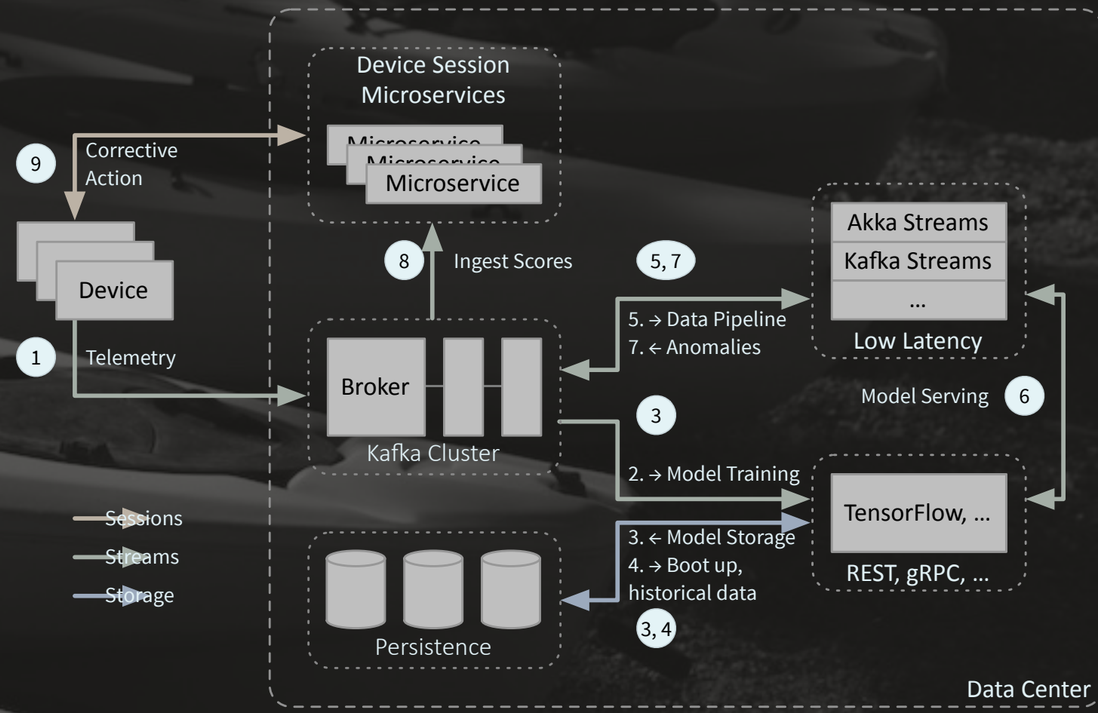
Model Serving as a Service

- Pros:
  - A familiar integration pattern
  - Decouples "concerns": AI tools, scalability, upgradability, …



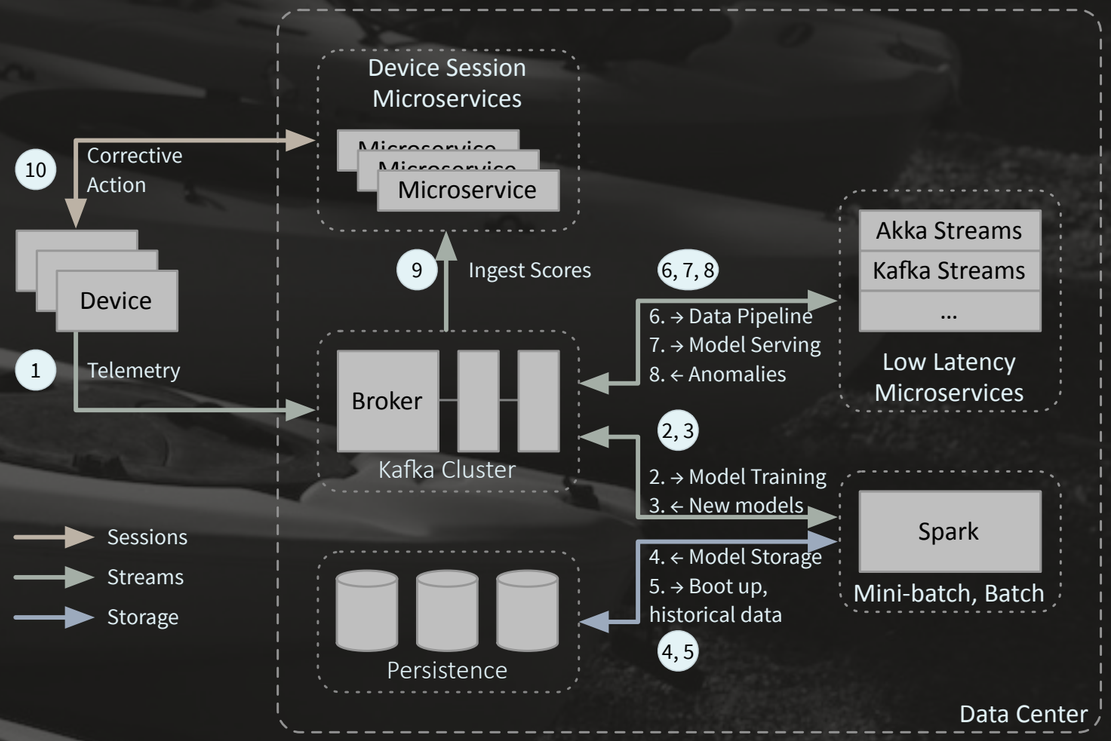
Model Serving as a Service

- Cons:
  - Overhead of invocation, e.g., REST
  - ML Pipeline becomes a unique production work flow



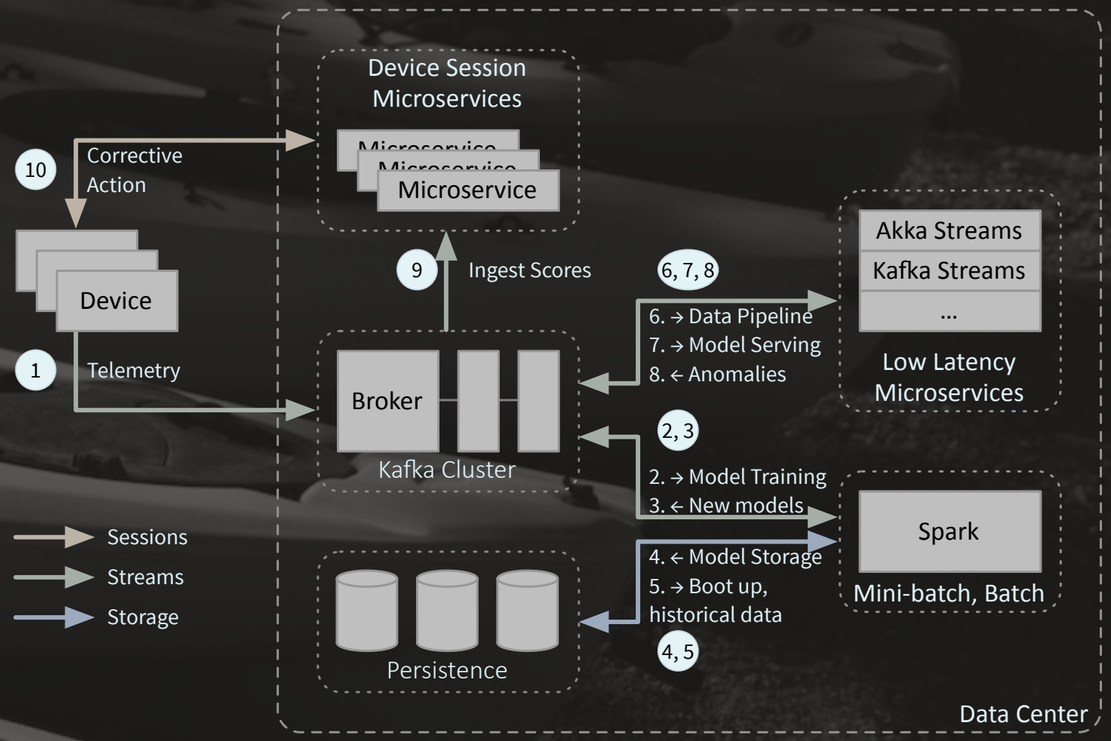
Embedded Model Serving

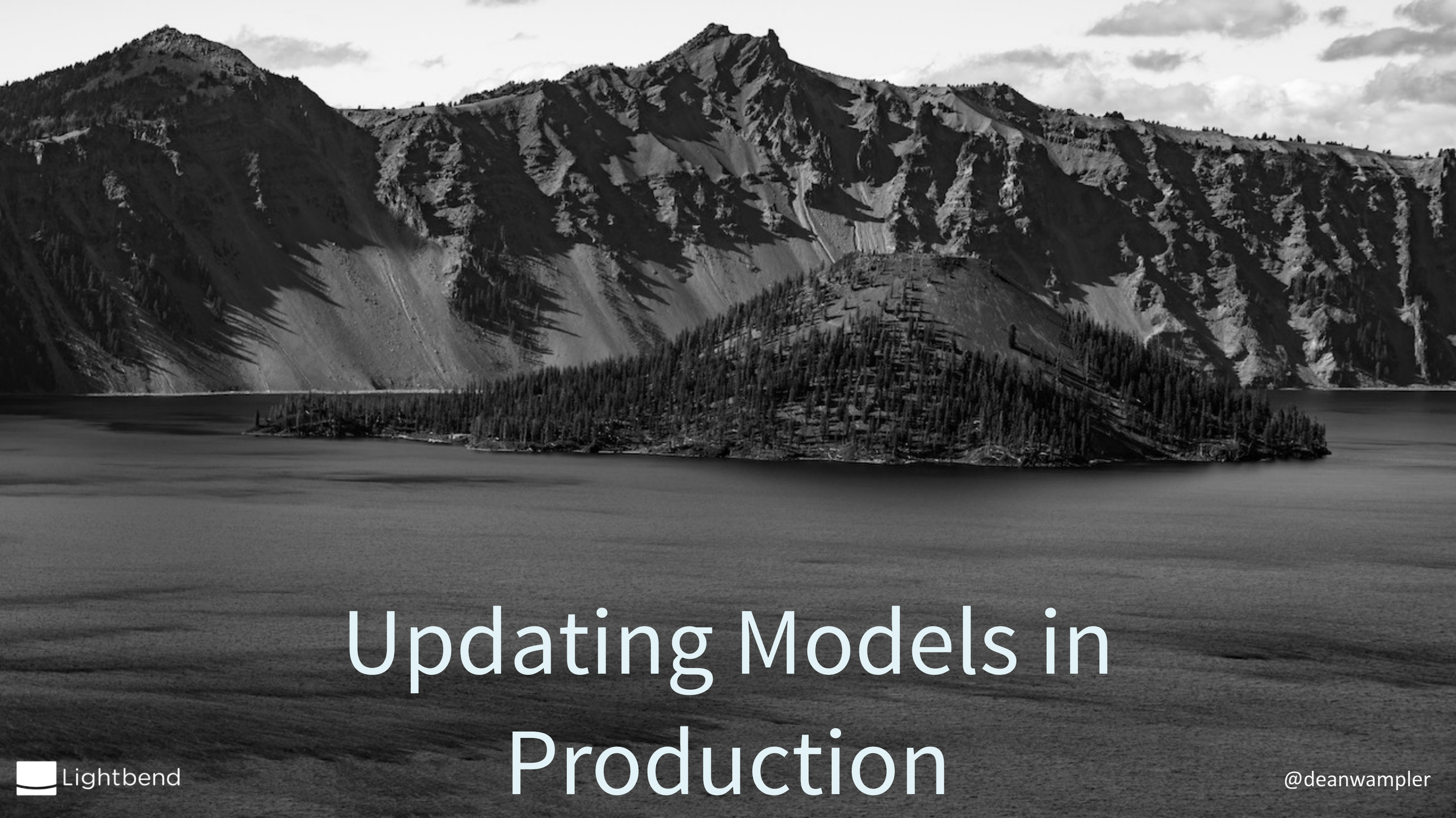
- Pros:
  - Lowest scoring overhead - interprocess communication only used for model updates
  - Performance tuning focuses on one system, the data pipeline



Embedded Model Serving

- Cons:
  - Model parameters must be serialized
  - Model serving library must be "compatible" with training system:
    - Algorithms, quality





Updating Models in Production

Lightbend

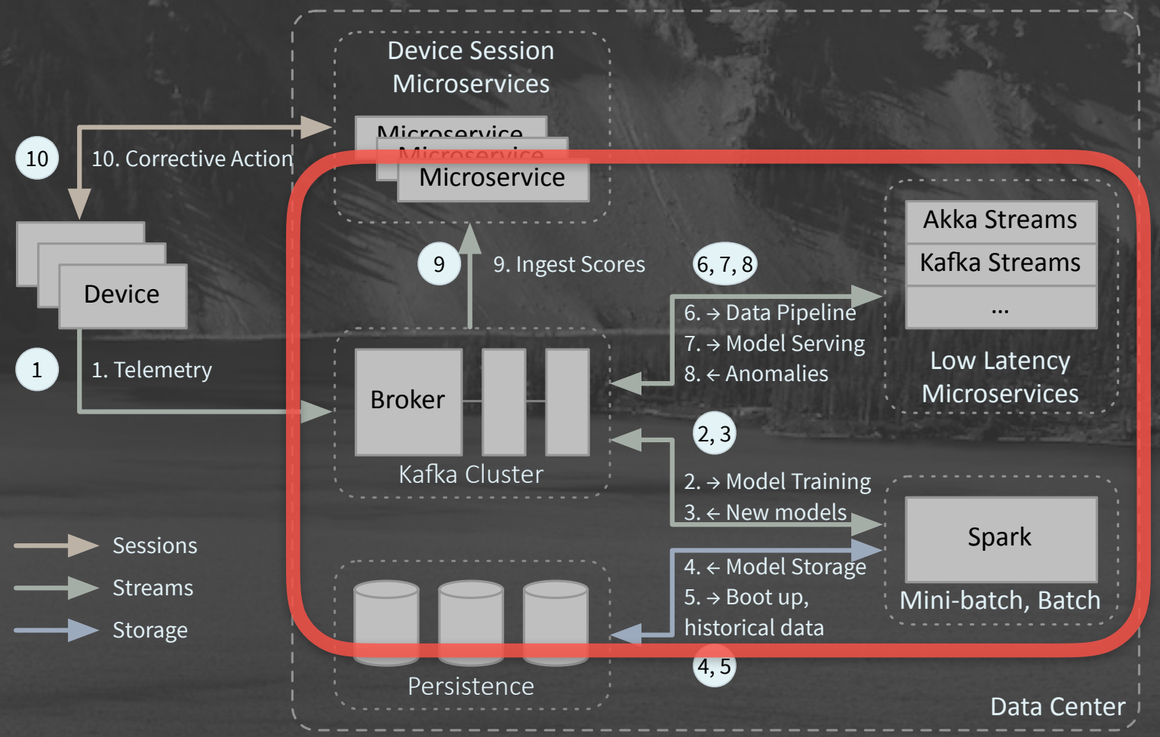
@deanwampler

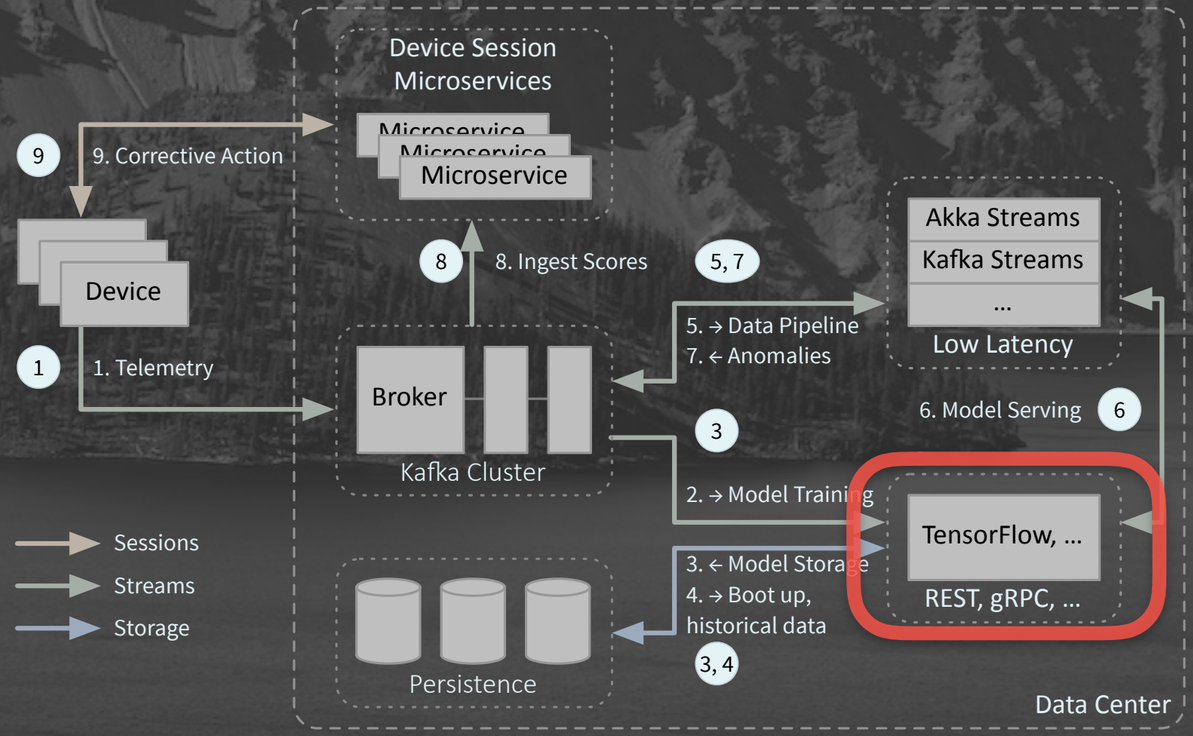
Model Updates

- Concept Drift - models grow stale
  - They have a half life, too
  - So, periodically retrain, then serve the new model, ideally without downtime

Retraining Considerations

- How do you measure model quality?
- What's the trade-off between model performance vs. retraining cost?
- How far back in the data set do you go when training?





Complex to update embedded models!

Model updates can be straightforward

Lightbend

@deanwampler


Dusty Milky Way
Mars last Summer

References

- Ideas:
  - Ben Lorica on 9 AI Trends
  - Paco Nathan's Data Governance Talk

You can get these slides with the links here:
polyglotprogramming.com/talks

References

- Ideas:
  - O'Reilly Radar: Data, AI, others
  - distill.pub
  - The Algorithm
  - The Gradient

References

- A few research papers, etc.
  - Incremental training
    - an example
  - Continual learning
  - Explainability

References

- Kubeflow
- MLFlow
- DVC
- AWS SageMaker
- Fiddler (explainable AI)

References

- General Information about Stream Processing
  - My O'Reilly Report on Architectures
  - Streaming Systems Book
  - Stream Processing with Apache Spark
  - Designing Data-Intensive Apps book

References

- Other Talks
  - Strata Talk on ML in a Streaming Context
  - Stream All the Things! (video)
  - Streaming Microservices with Akka Streams and Kafka Streams (video)

References

- Tutorials
  - Model serving in streams
  - Stream processing with Kafka and microservices

Please Provide Feedback!

Cyberconflict: A new era of war, sabotage, and fear

See passes & pricing

David Sanger (The New York Times)
9:55am-10:10am Wednesday, March 27, 2019
Location: Ballroom
Secondary topics: Security and Privacy

Add to Your Schedule
Add Comment or Question

Rate This Session

We're living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you're often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

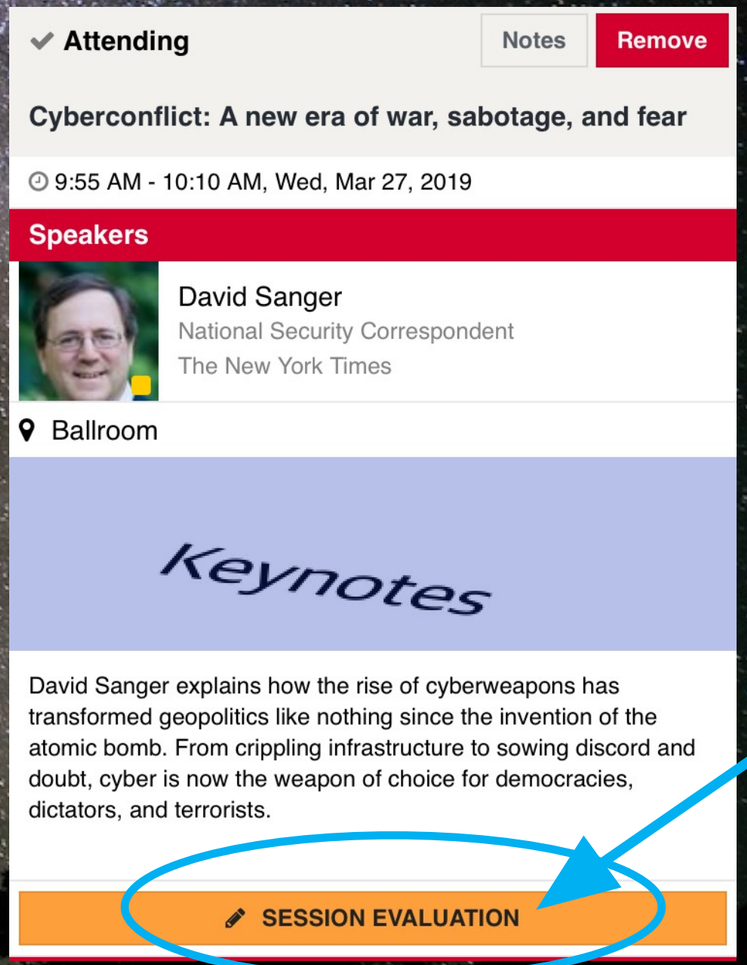
David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we're uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger

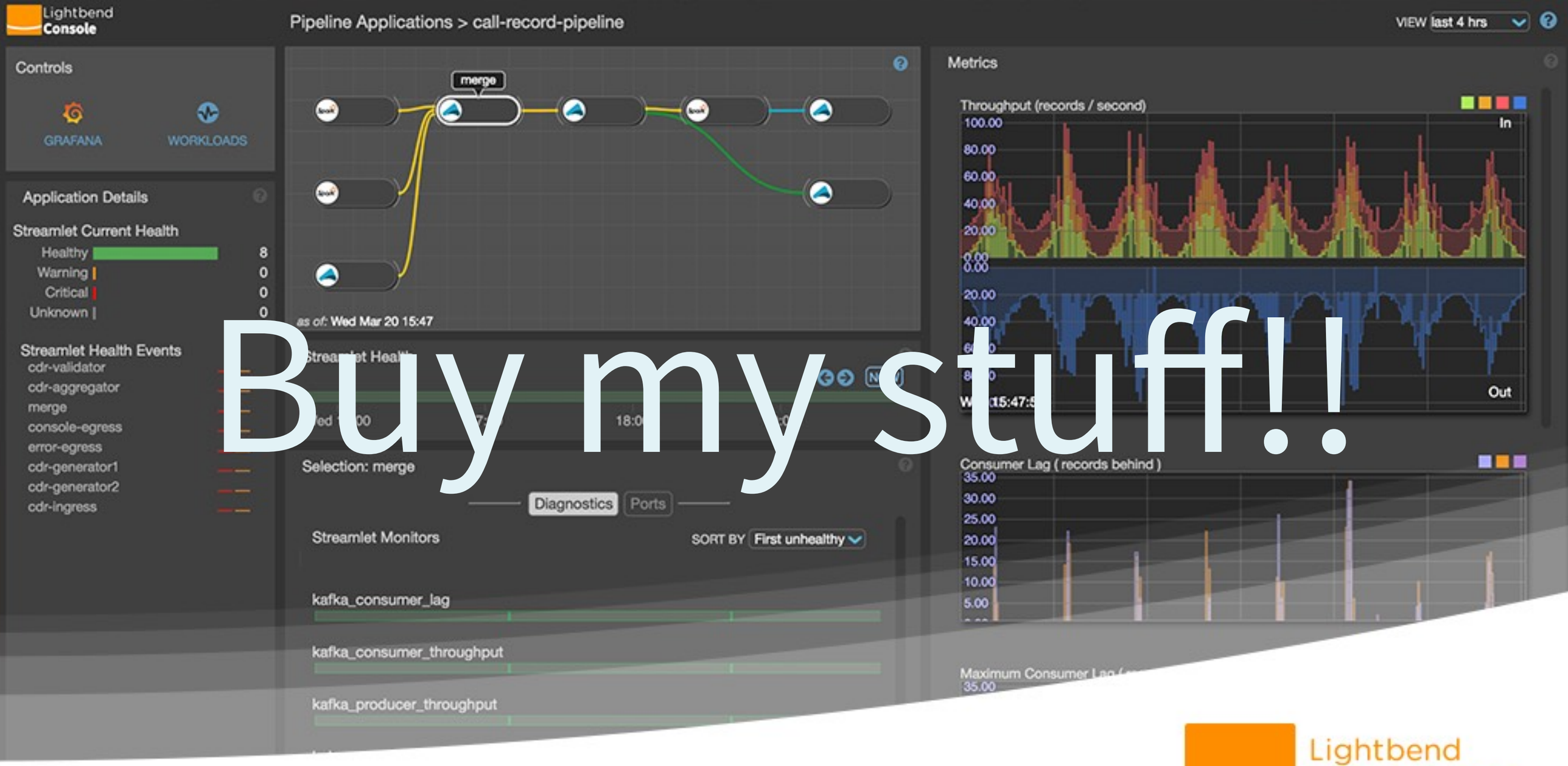
The New York Times

David E. Sanger is the national security correspondent for the *New York Times* as well as a national security and political contributor for CNN and a frequent guest on *CBS This Morning*, *Face the Nation*, and many PBS shows.

Session page on conference website



O'Reilly Events



Buy my stuff!!

lightbend.com/pipelines

Lightbend Pipelines

Questions?

Dean Wampler, Ph.D.
dean@lightbend.com
@deanwampler

lightbend.com/pipelines

polyglotprogramming.com/talks