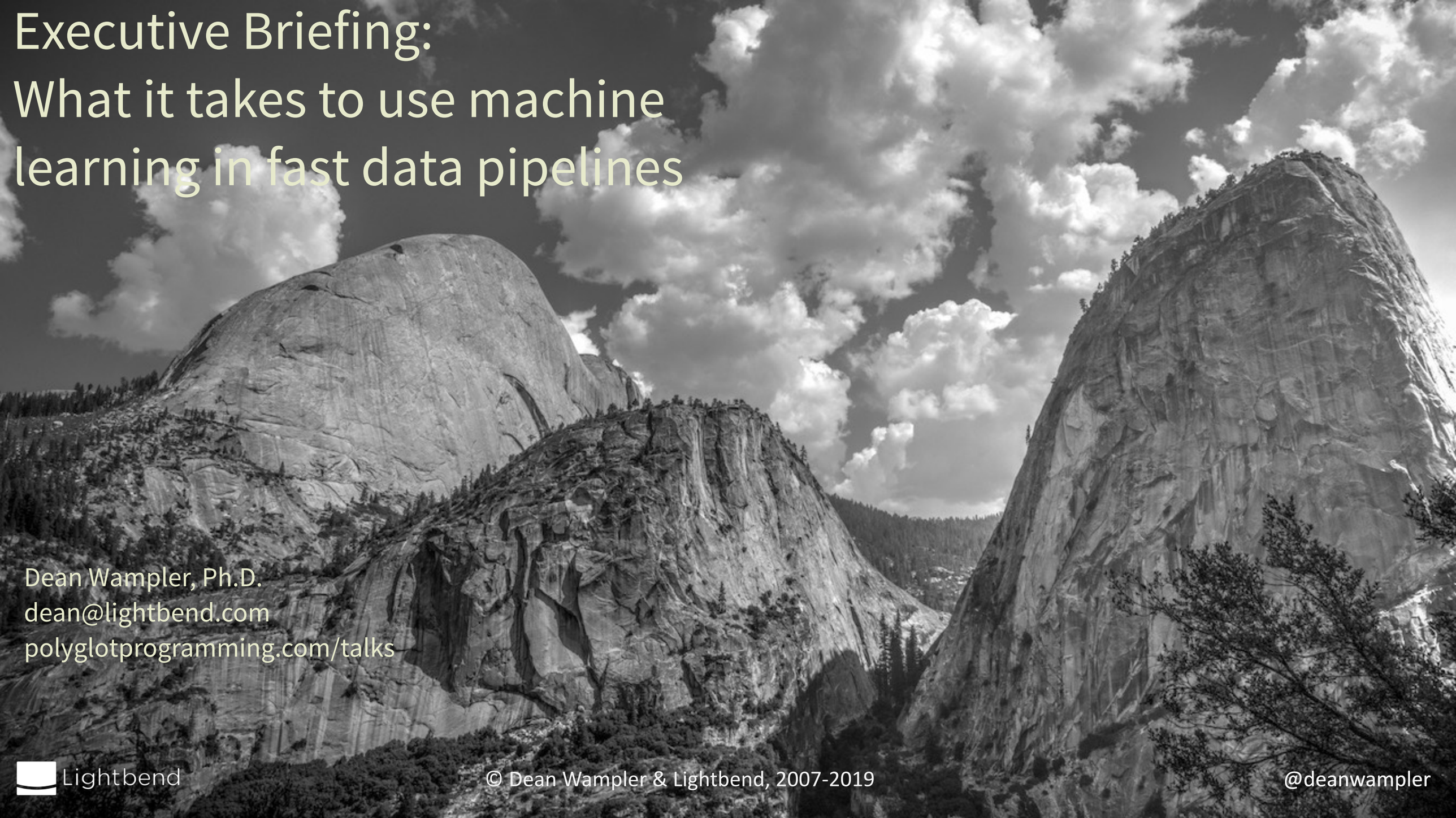
Executive Briefing: What it takes to use machine learning in fast data pipelines



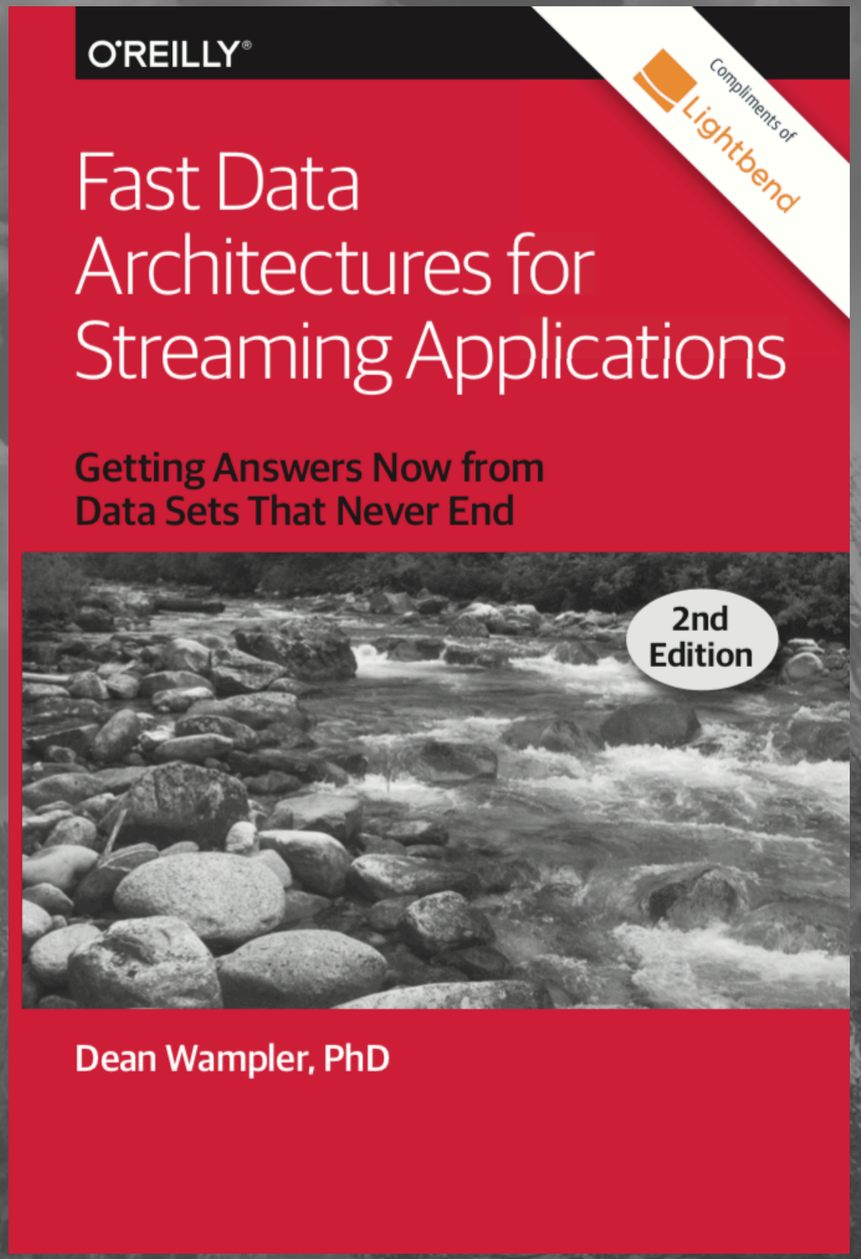
Dean Wampler, Ph.D.
dean@lightbend.com
polyglotprogramming.com/talks

Lightbend

@deanwampler

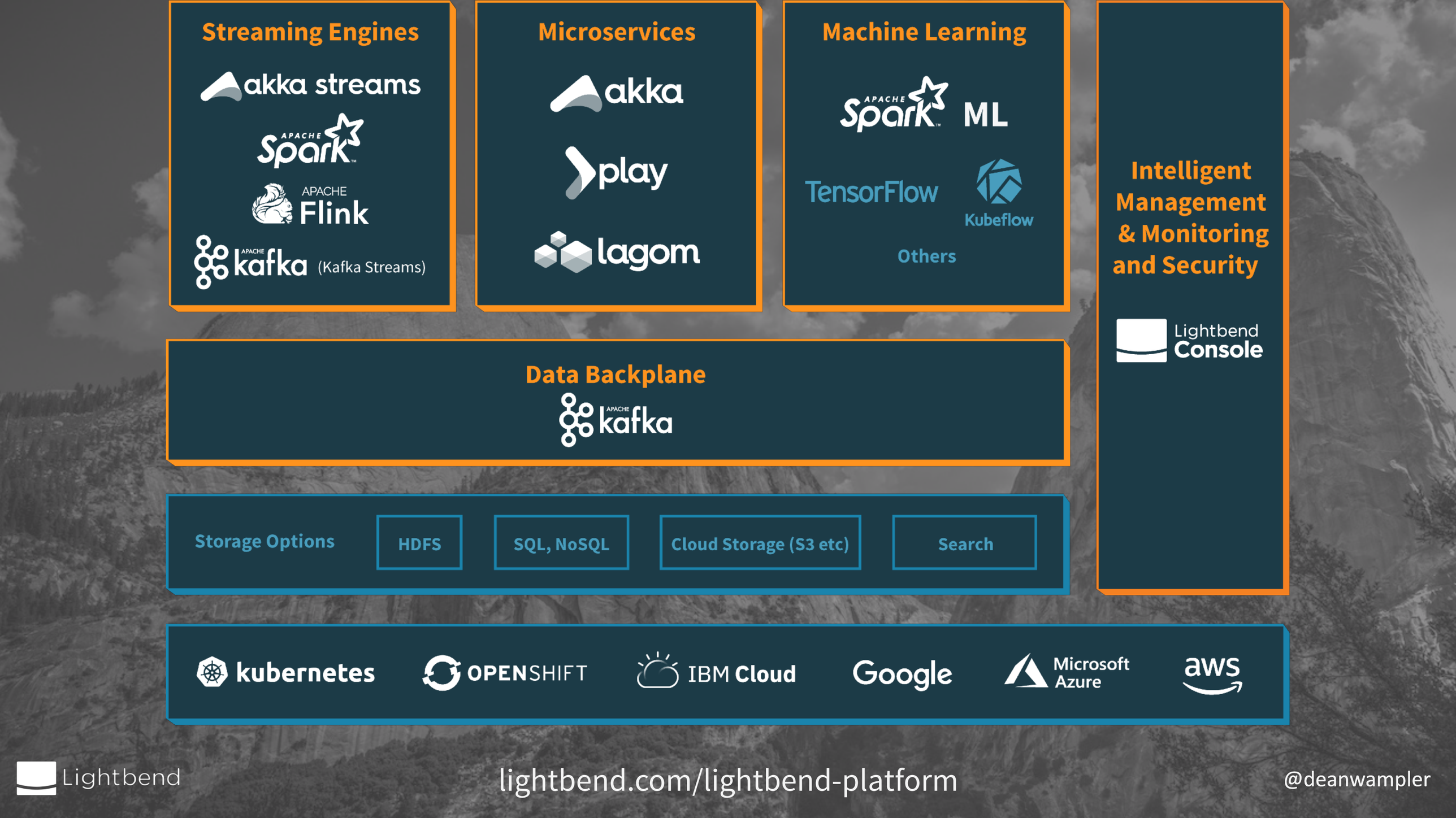
Data Streaming, in General

lbnd.io/fast-data-ebook



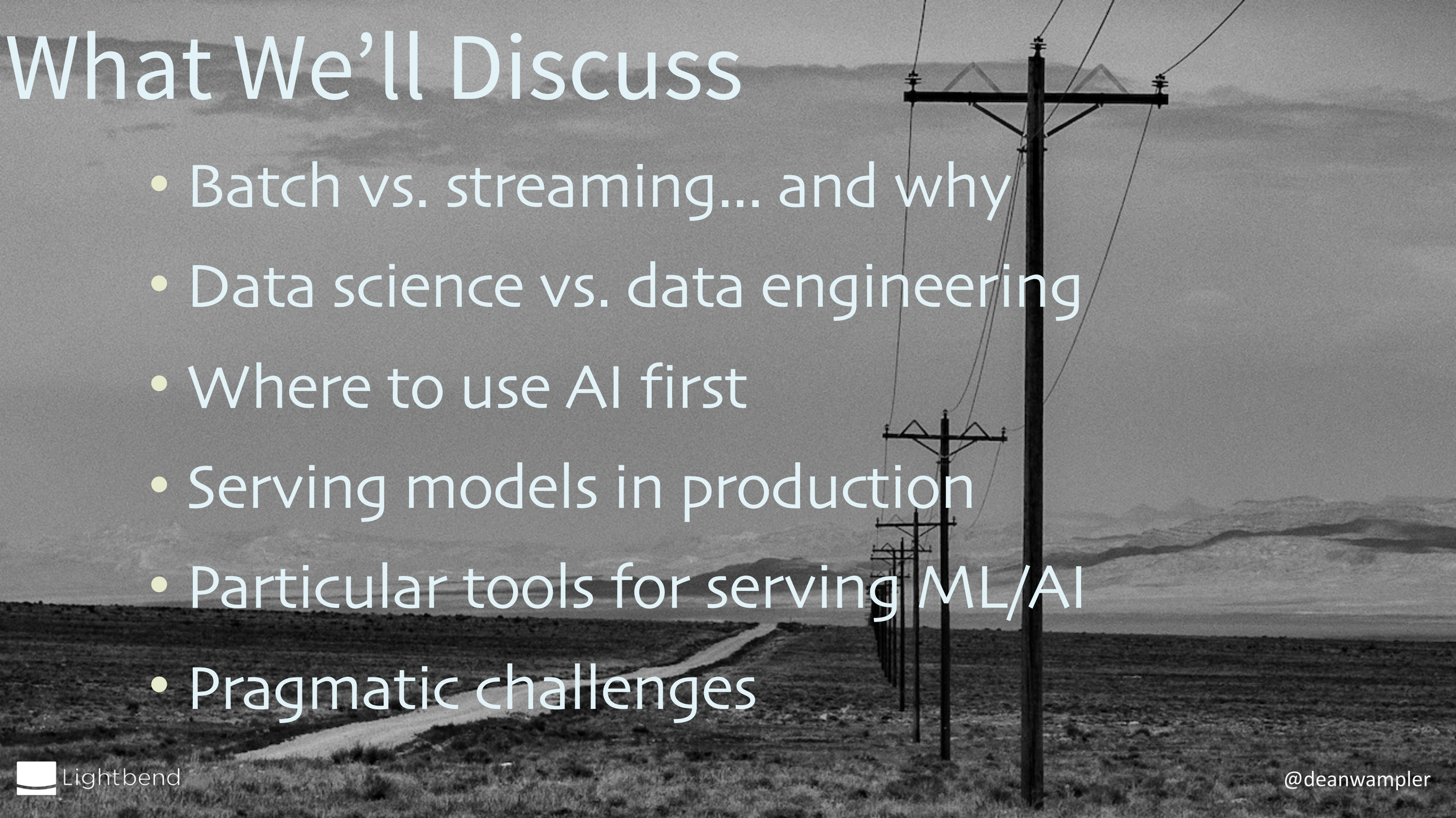
Lightbend

@deanwampler


Streaming Engines
akka streams
Spark
Flink
kafka (Kafka Streams)
Microservices
akka
play
lagom
Machine Learning
Spark ML
TensorFlow
Kubeflow
Others
Intelligent Management & Monitoring and Security
Lightbend Console
Data Backplane
kafka
Storage Options
HDFS
SQL, NoSQL
Cloud Storage (S3 etc)
Search
kubernetes
OPENSHIFT
IBM Cloud
Google
Microsoft Azure
aws
Lightbend
lightbend.com/lightbend-platform
@deanwampler

What we're up to at Lightbend…
lightbend.com/lightbend-pipelines-demo

Lightbend Pipelines

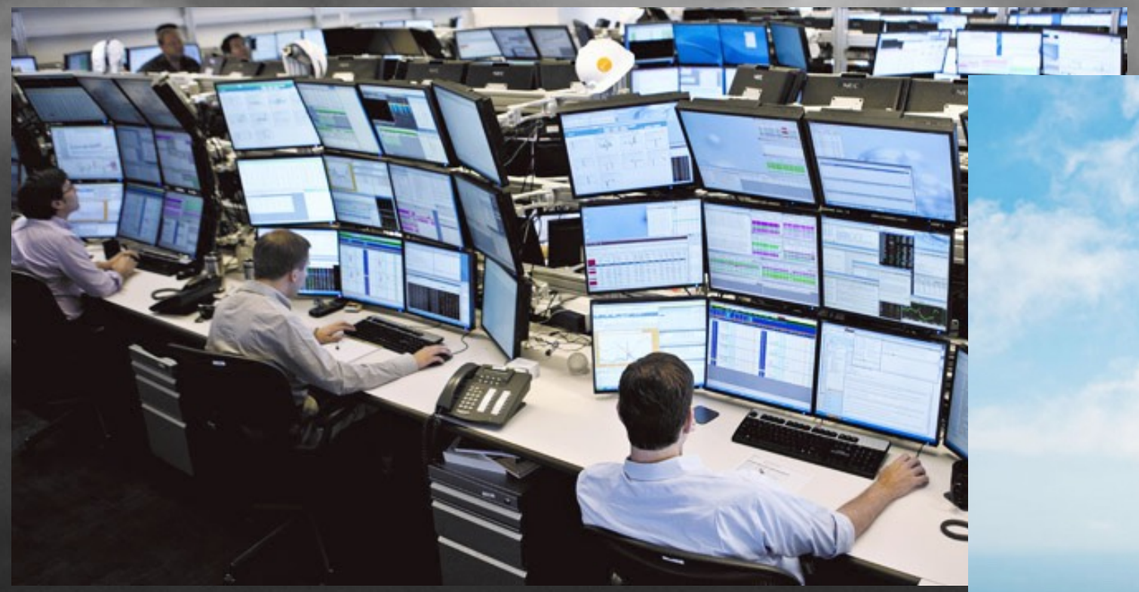


What We'll Discuss

- Batch vs. streaming... and why
- Data science vs. data engineering
- Where to use AI first
- Serving models in production
- Particular tools for serving ML/AI
- Pragmatic challenges

Batch vs. streaming… and why

@deanwampler



Finance

Telecom

Medical

Energy

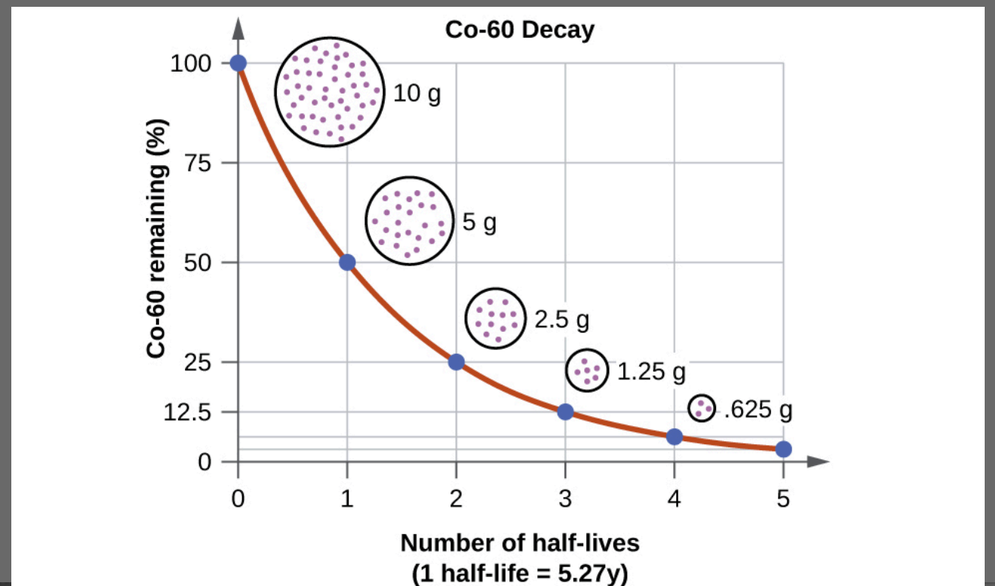
... and IoT

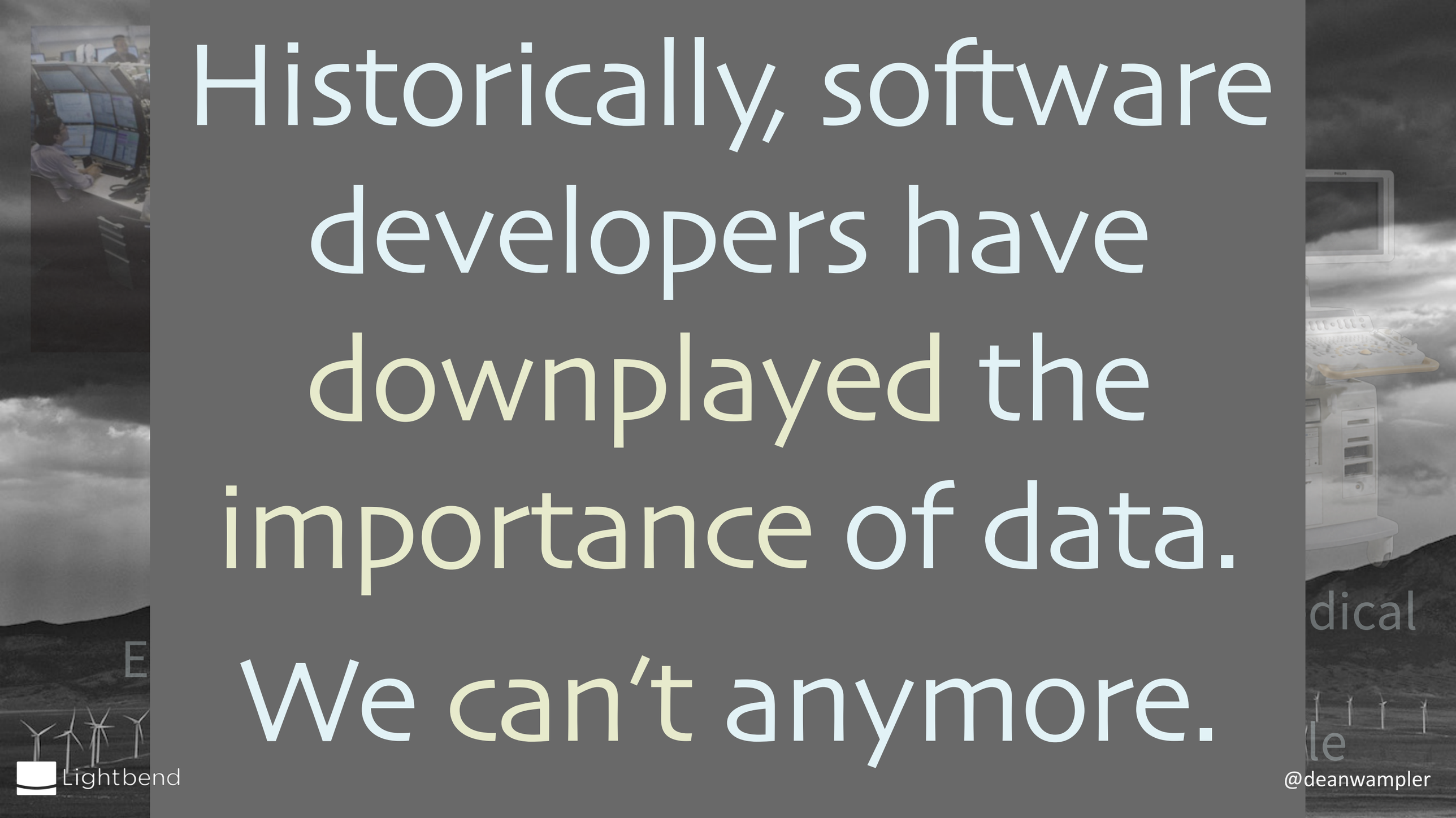
State of the art phone!

Mobile

@deanwampler

Information value has a half life; it decays with time



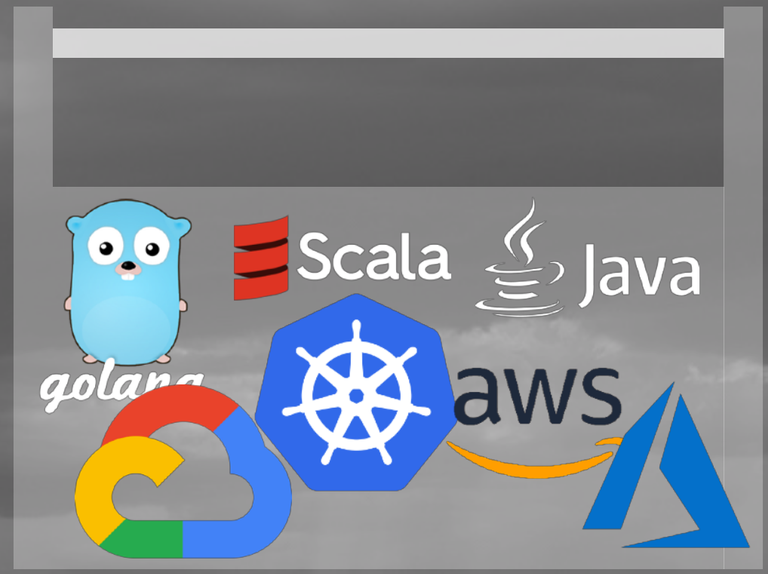


Historically, software developers have downplayed the importance of data.

We can't anymore.


Data Science vs.
Data Engineering
Lightbend
@deanwampler

Data Science toolbox

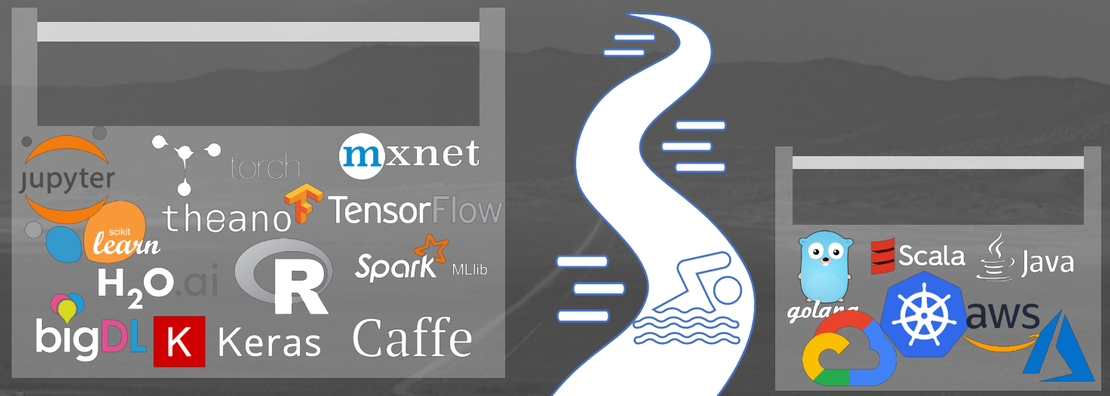


Software Engineering toolbox

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

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

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



Where to Apply AI First

Lightbend

@deanwampler

- "AI in the enterprise will build upon existing analytic applications."
  - Hybrid systems: enhance existing analytics with ML/AI models.

https://www.oreilly.com/ideas/9-ai-trends-on-our-radar

- Should we integrate legacy "expert systems" and how?

Serving Models in Production

Lightbend

@deanwampler

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

CI/CD Processes Required (1/3)

- Version control - for models and code
- Automated builds, tests and other quality checks, artifact delivery
- Launch Configurations: "dark" launches, A/B and other testing scenarios

CI/CD Processes Required (2/3)

- Monitoring
  - Performance overhead
  - Quality metrics match expectations from training?

CI/CD Processes Required (3/3)

- 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 (3/3)

- Au
  - V
  - V
  - V

Models

Are

Data

But What's Different? (1/2)

- Automation of model search, experimentation
  - E.g., hyperparameter tuning
- Data safety, fairness, and lineage
- Automation needs to measure reliability, SLAs,
- Reproducibility

https://www.oreilly.com/ideas/9-ai-trends-on-our-radar

But What's Different? (2/2)

How to integrate the extra indeterminacy in a DevOps world that prefers determinacy?

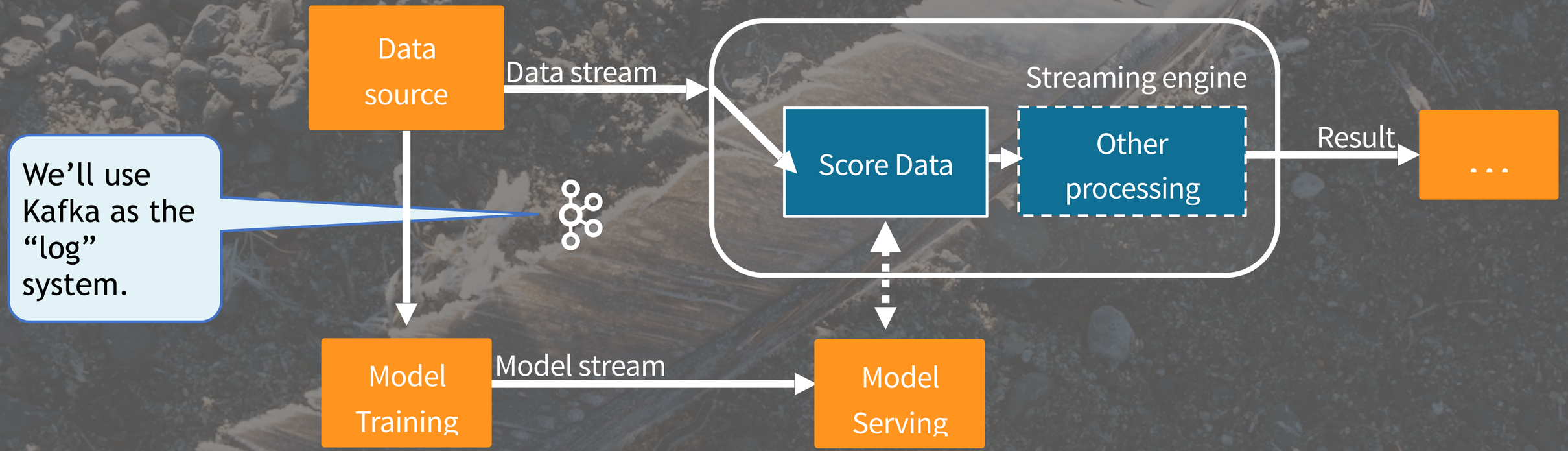
https://www.oreilly.com/ideas/9-ai-trends-on-our-radar


Tools

Tools

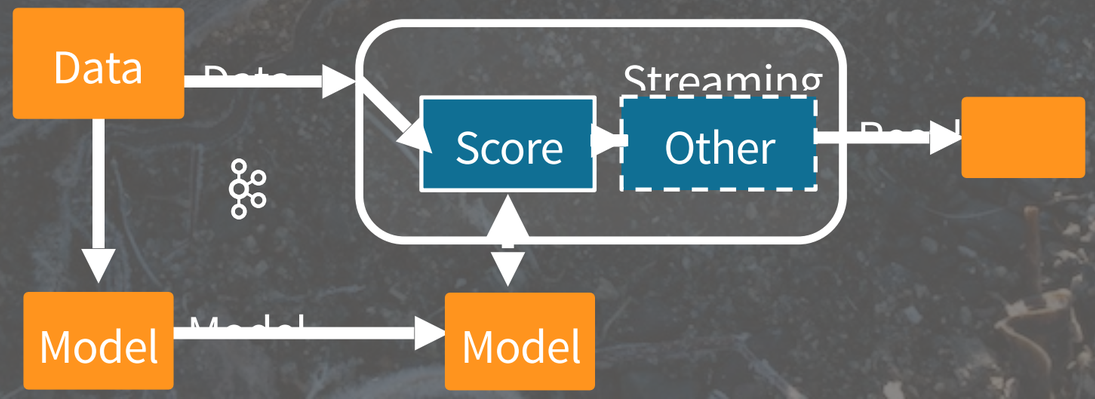
- Approaches:
  - Model Serving as a Service
  - Embedded within Microservices
- Systems
  - Kubeflow, MLFlow, …

Model Serving as a Service



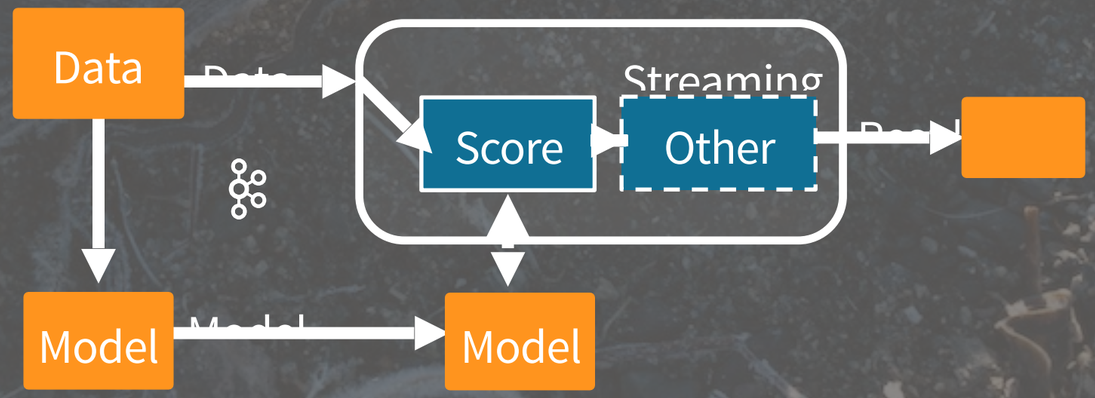
Serving

- Pros:
  - Decouples "concerns": tools, scalability, upgradability, …
  - One system for training and scoring

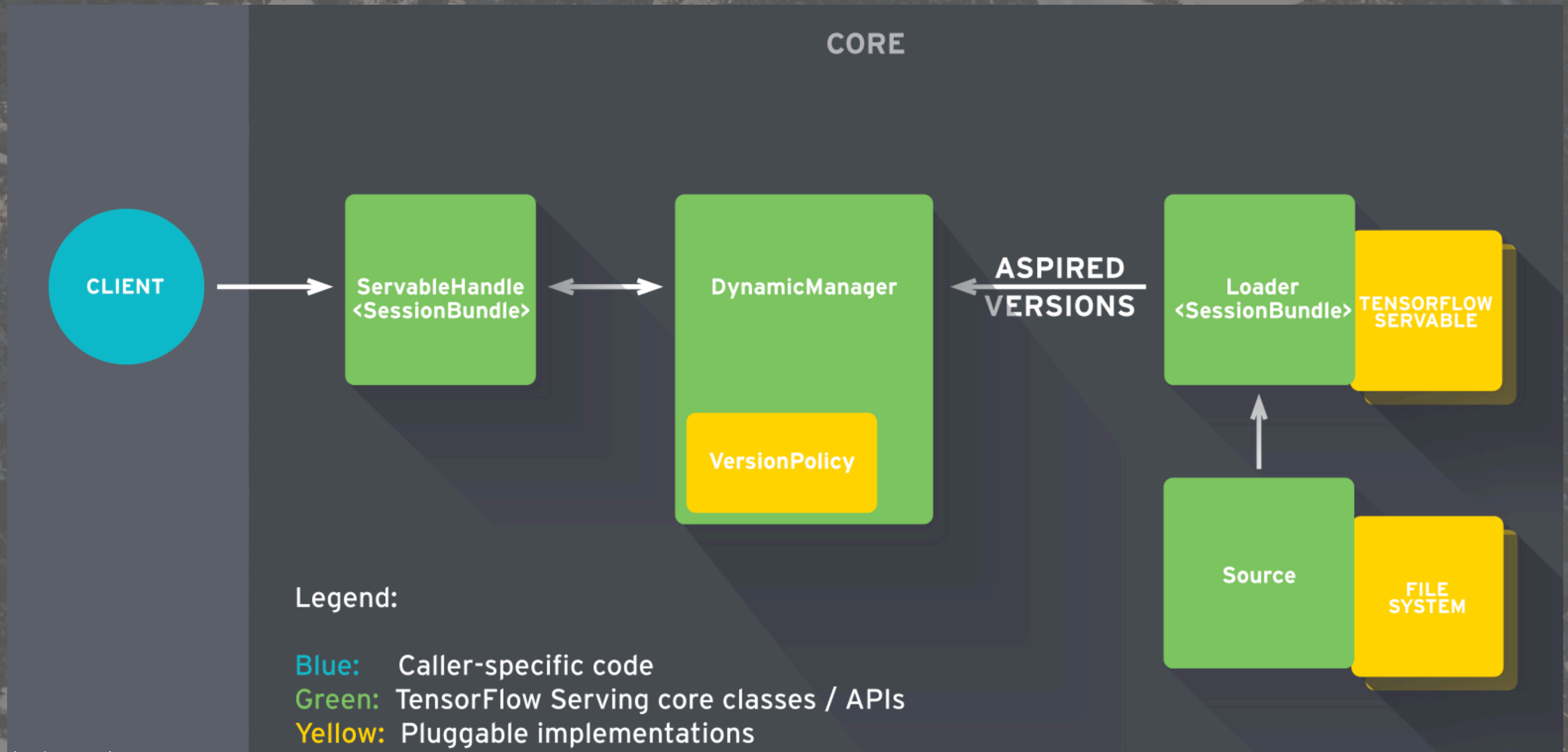


Serving

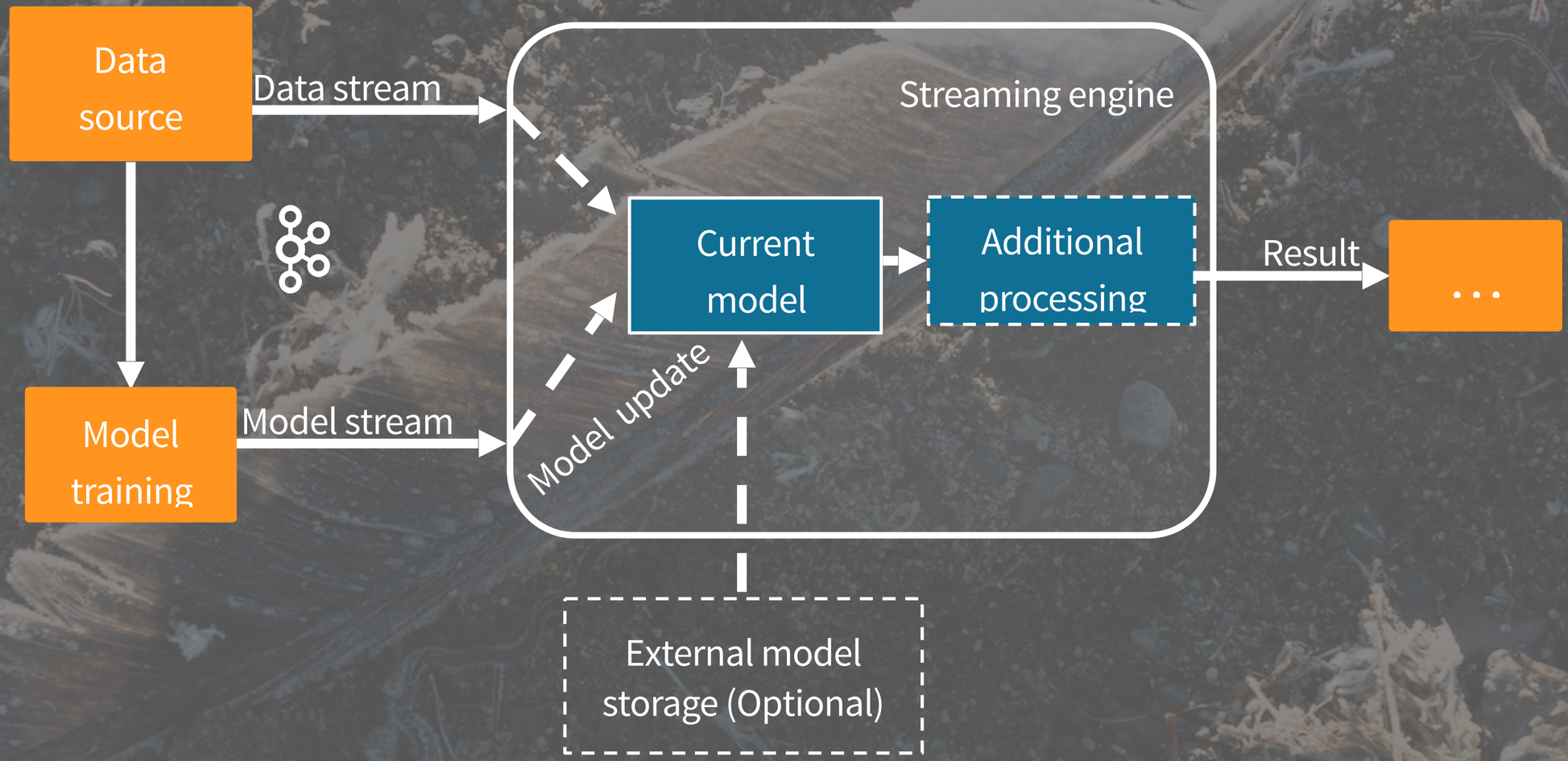
- Cons
  - Overhead of invocation, e.g., REST
  - ML Pipeline becomes a unique production work flow



TensorFlow Serving

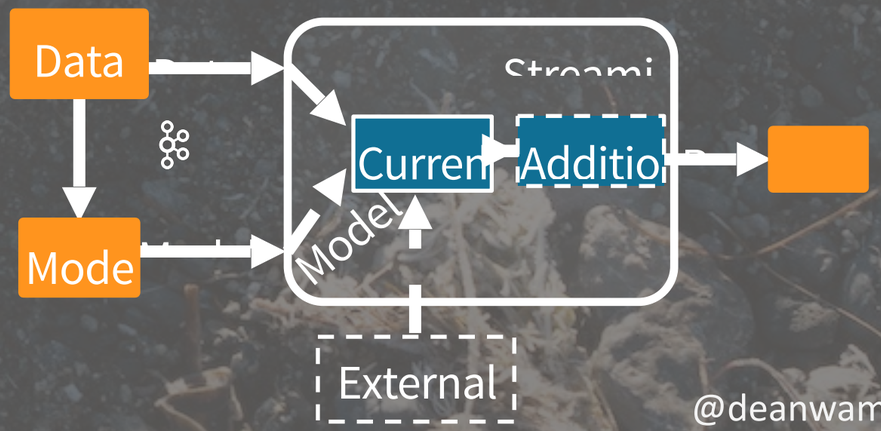


Embedded in Microservices



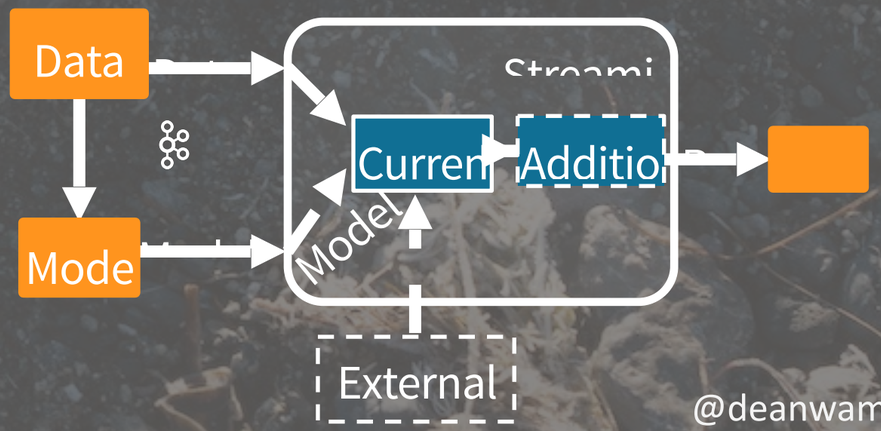
Embedded

- Pros:
  - Minimal interprocess communication overhead
  - Performance tuning focuses on one system, the data pipeline



Embedded

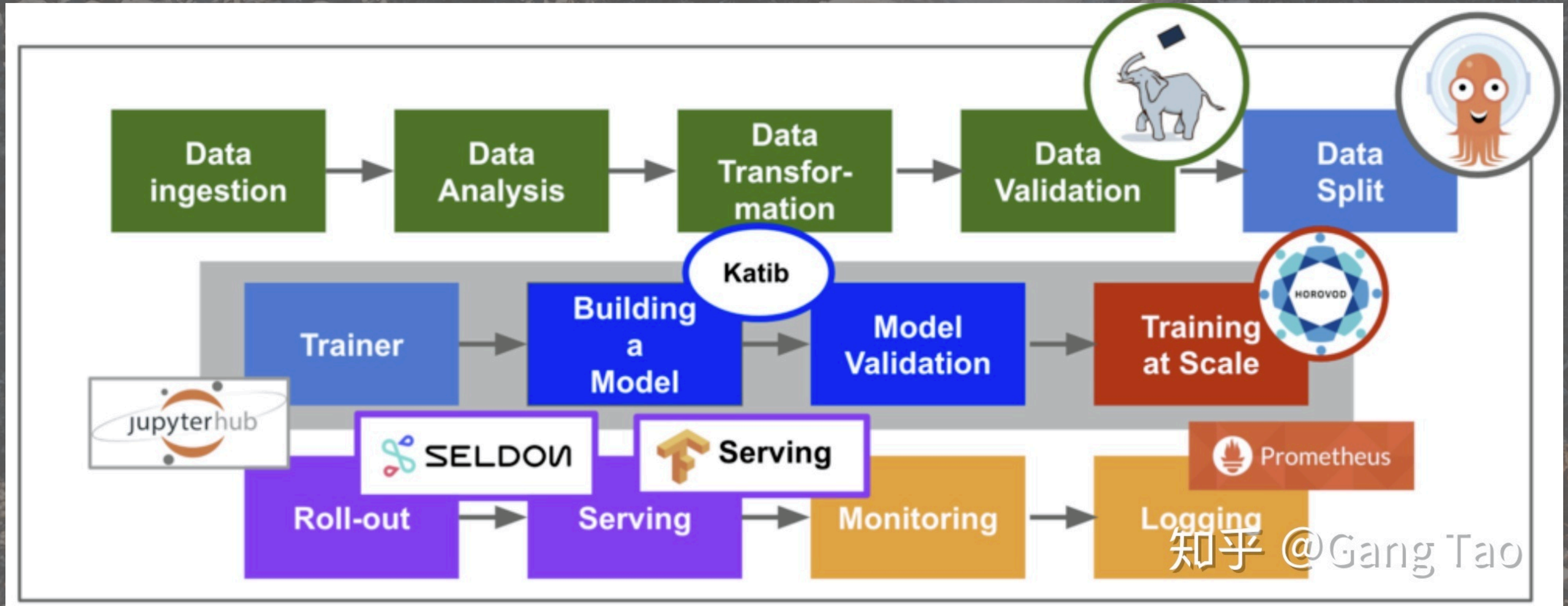
- Cons
  - Model parameters must be serialized
  - Model serving library must be "compatible" with training system:
    - Same algorithms
    - Same performance



Systems

- Kubeflow - for example
  - For Kubernetes
- Others
  - MLflow
  - AWS SageMaker
  - …

Systems - Kubeflow



Pragmatic Challenges

Lightbend

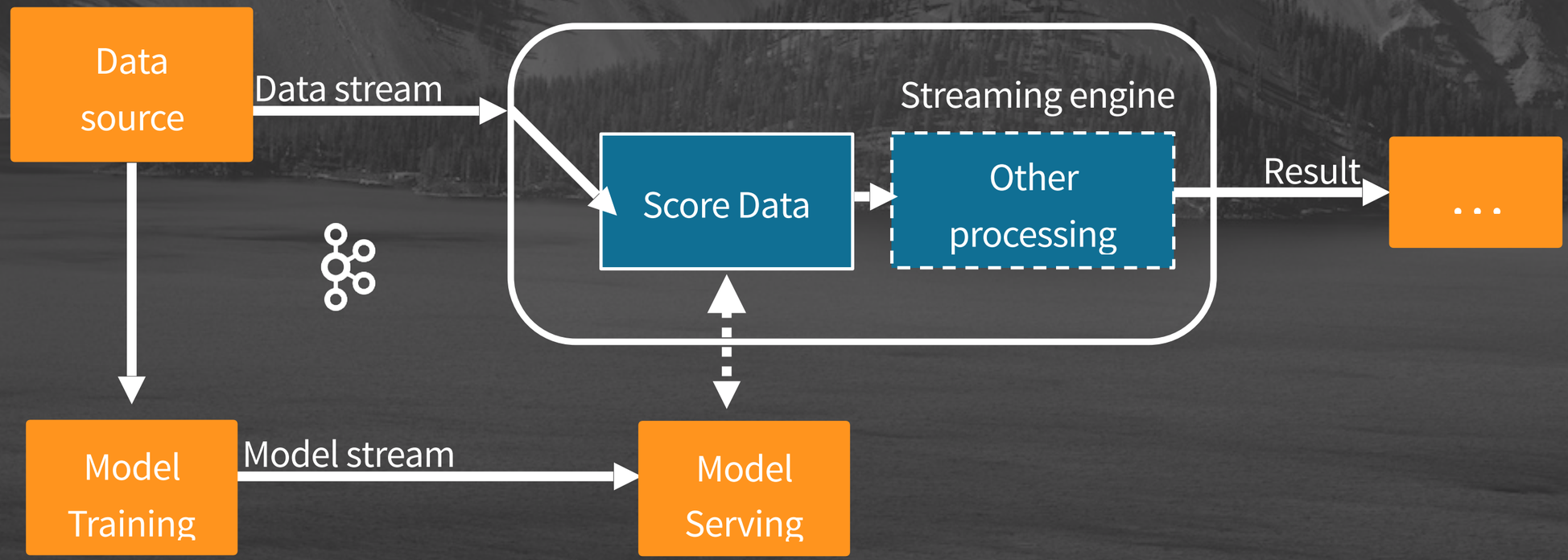
@deanwampler

Model Updates “in situ”

- “Concept Drift” - models grow stale
  - They have a “half life”, too
- Periodically retrain then serve the new model
  - (Continuous training a research topic)

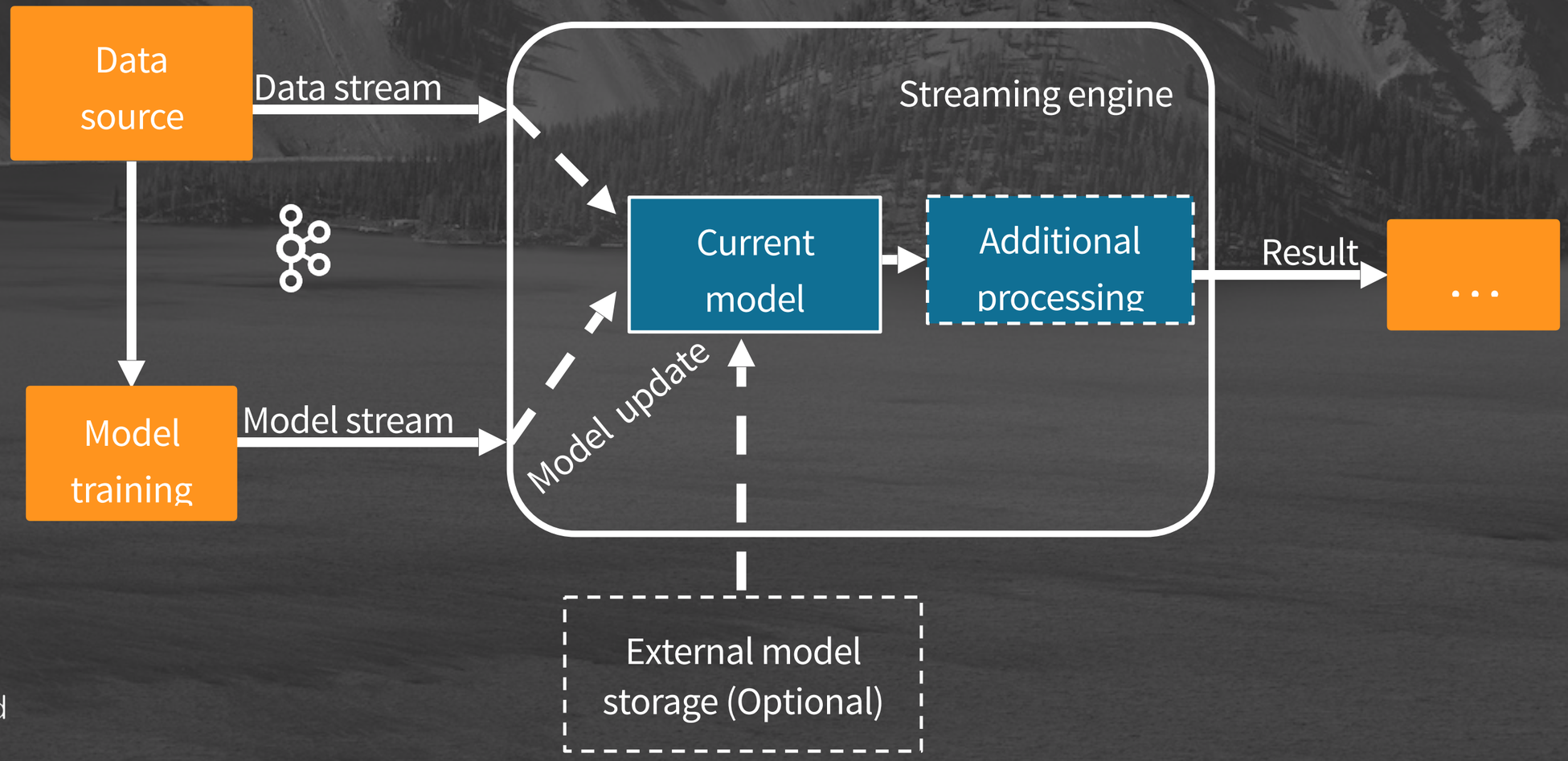
Model Updates “in situ”

- Model updates and MSaaS



Model Updates “in situ”

- Embedded model updates



Auditing

- Kind of Model
- Parameters and hyperparameters
- When trained
- Data used for training
- When deployed, undeployed, etc.
- …

Auditing

- Quality metrics
- Serving metrics (how many records, scoring times…)
- Provenance of decision to retrain
  - The metrics gathered above that were used to decide when to retrain

Retraining Considerations

- Trade off model performance vs. retraining cost
- What data set? How far back do you go?

Retraining Considerations

- Approaches:
  - Mature: "periodic" Batch jobs
  - Online and incremental algorithms
  - Future: Auto-adaptive ML (i.e., continual learning). Successor to AutoML


Dusty Milky Way
Mars last Summer

References

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

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

References

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

References

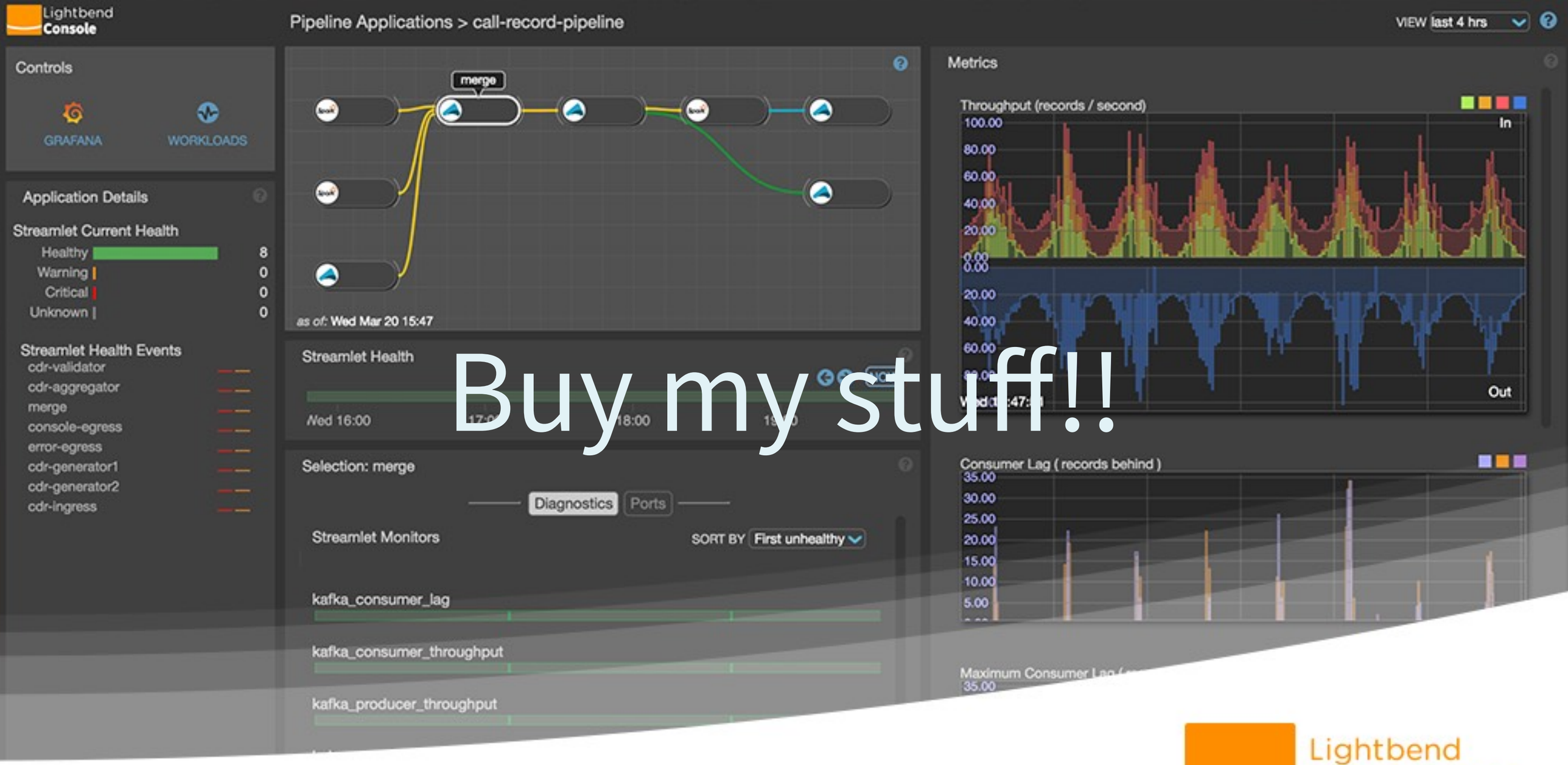
- General Information about Stream Processing
  - My O'Reilly Report on Architectures
  - Streaming Systems Book
  - Stream Processing with Apache Spark
  - Designing Data-Intensive Apps book

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



Buy my stuff!!

What we're up to at Lightbend…
lightbend.com/lightbend-pipelines-demo

Lightbend Pipelines

Questions?

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

lightbend.com/lightbend-platform
polyglotprogramming.com/talks