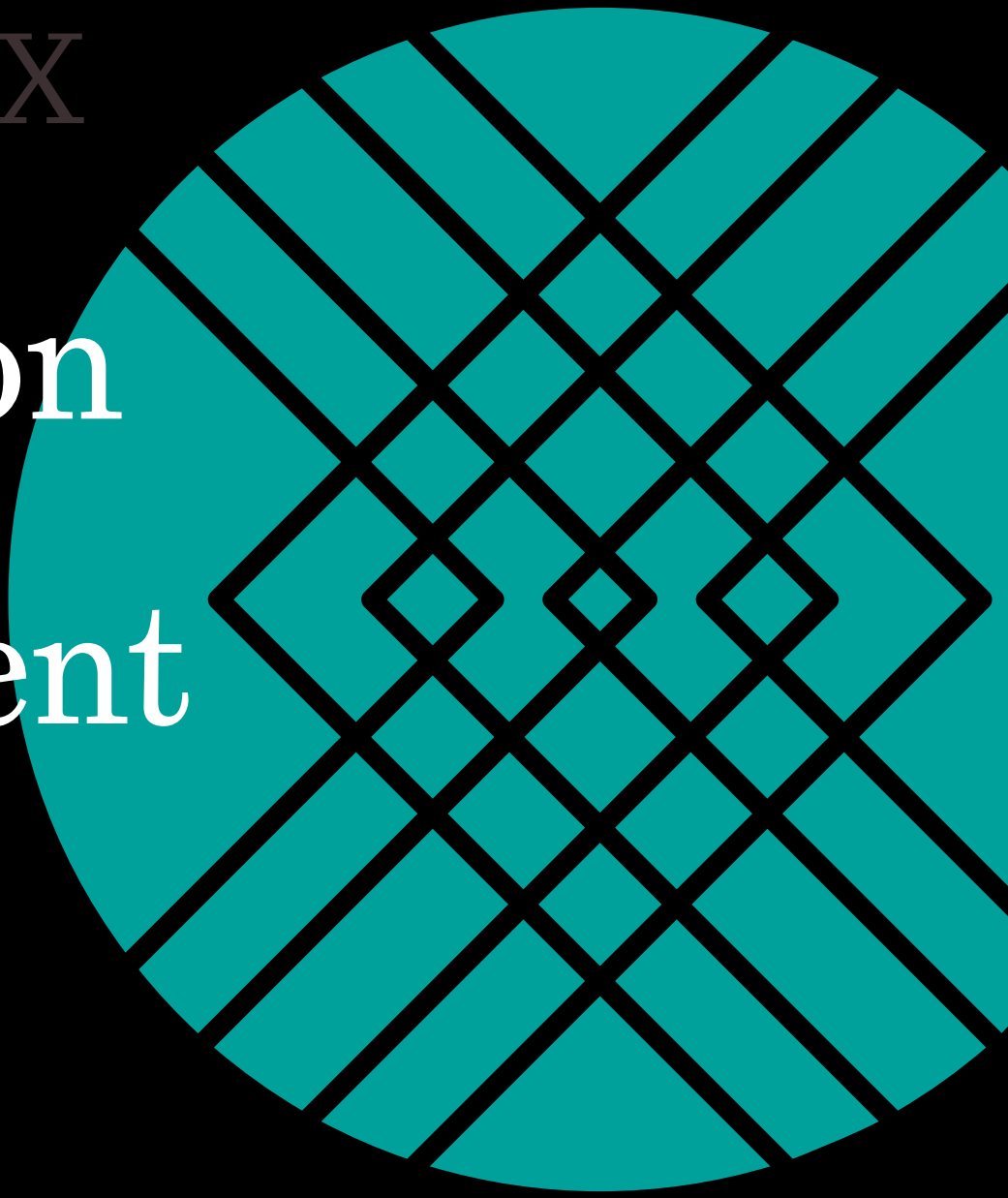
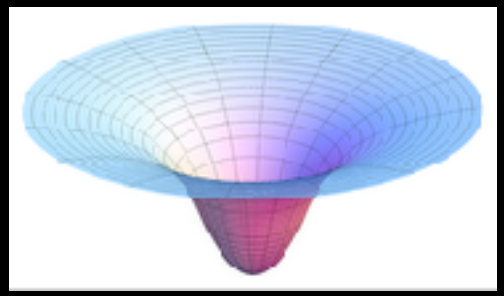
STITCH FIX

Production Model Deployment



Juliet Hougland
@j_houg
April 2018

whoami



STITCH FIX

cloudera

And other, failed startups

odiago

wibidata

Agenda

MODEL LIFECYCLE

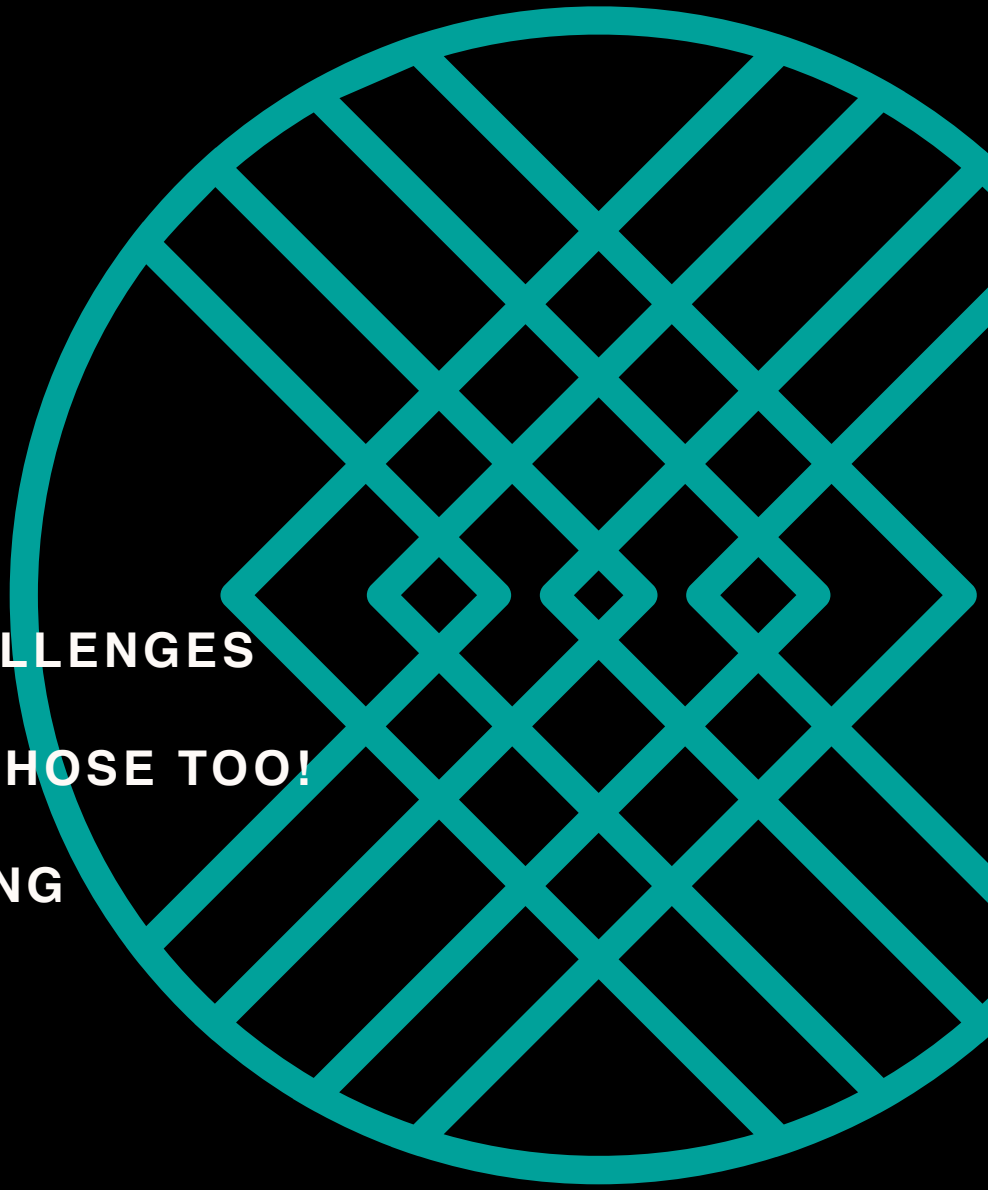
DEPLOYMENT CHALLENGES

SOLUTIONS

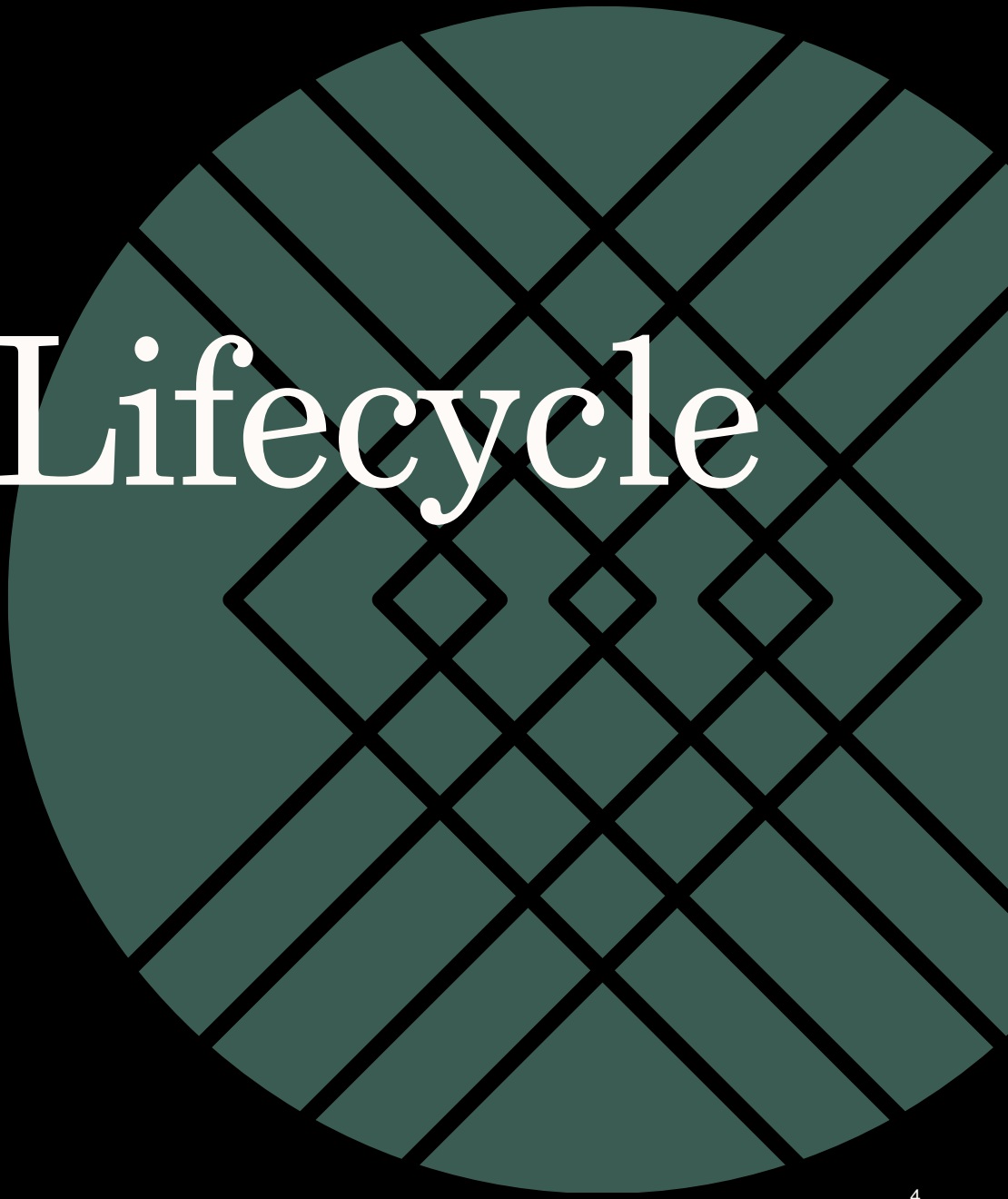
THAT THEMSELVES ARE CHALLENGES

THERE ARE SOLUTIONS TO THOSE TOO!

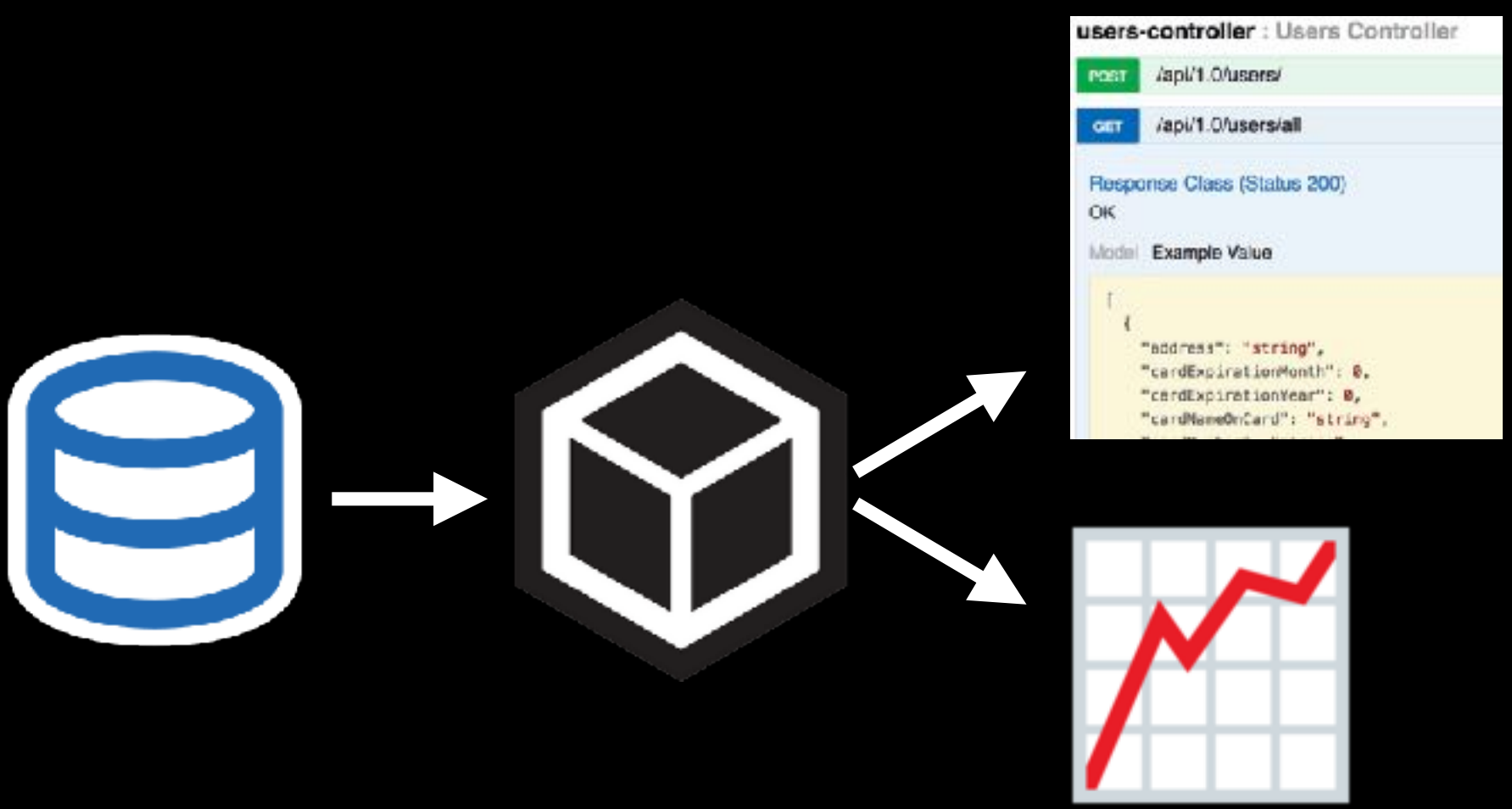
THEY ARE ALSO CHALLENGING



Model Lifecycle



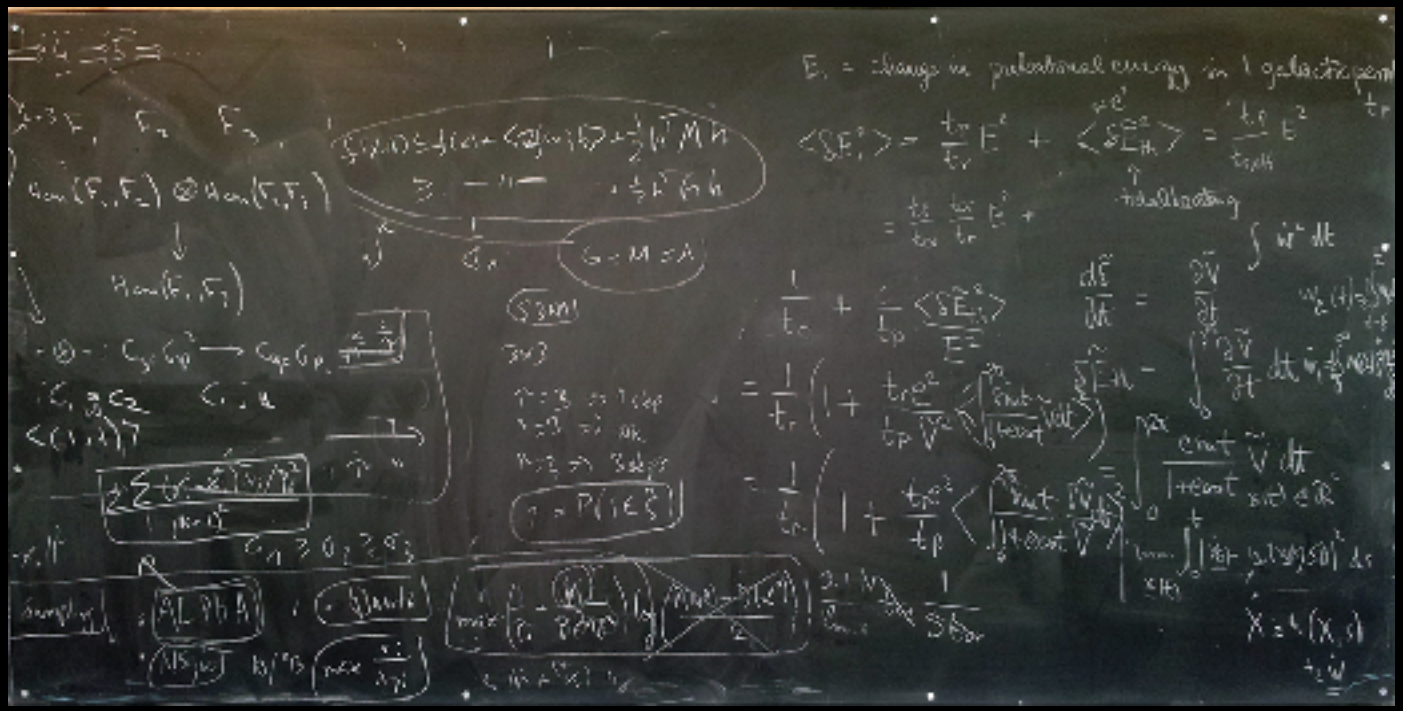
Black Box Data Scientists



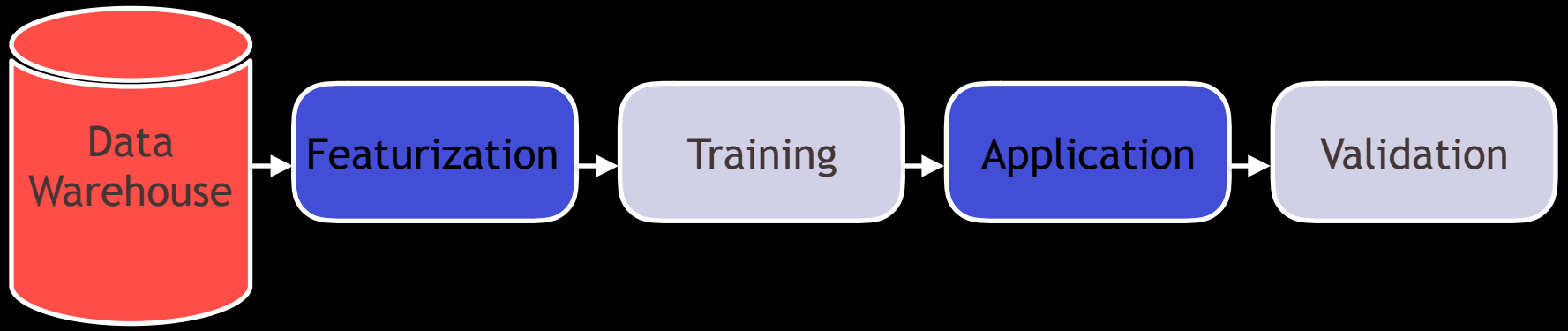
@j_houg

Black Box Data Scientists

The Black Box



@j_houg

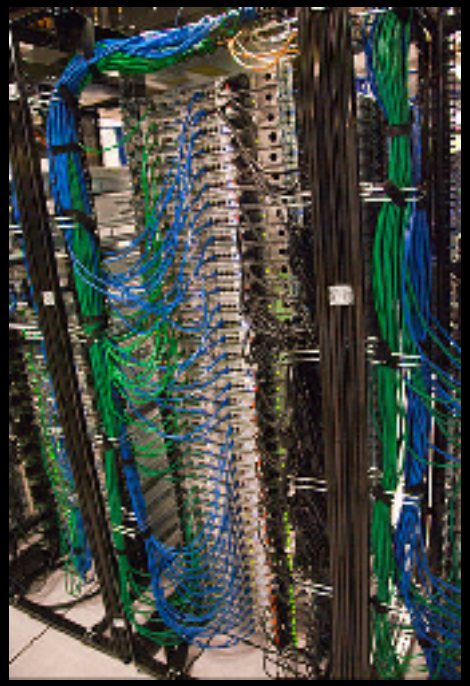

Data Warehouse
Featurization
Training
Application
Validation

@j_houg

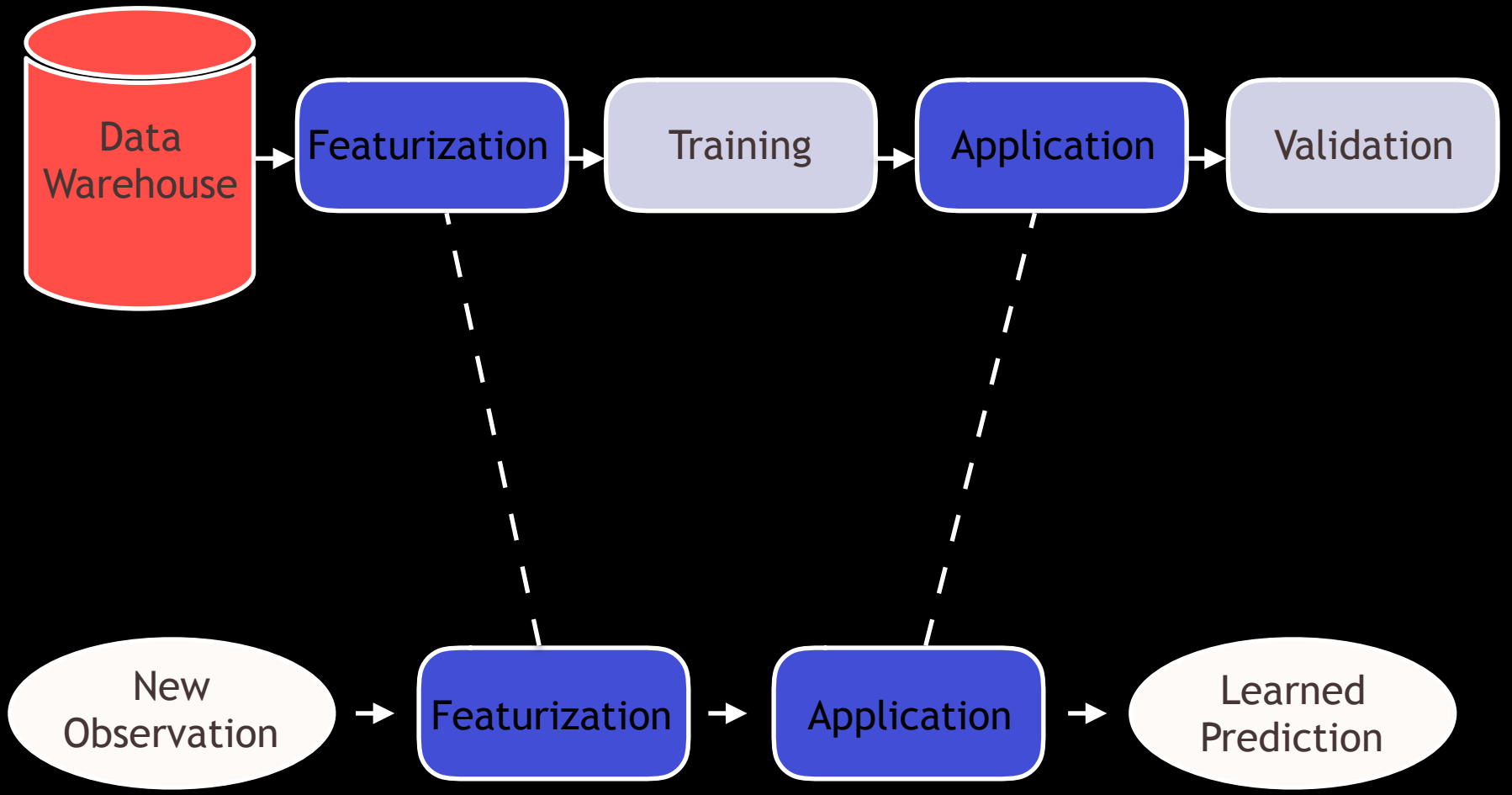
Deploying a Model

Share!

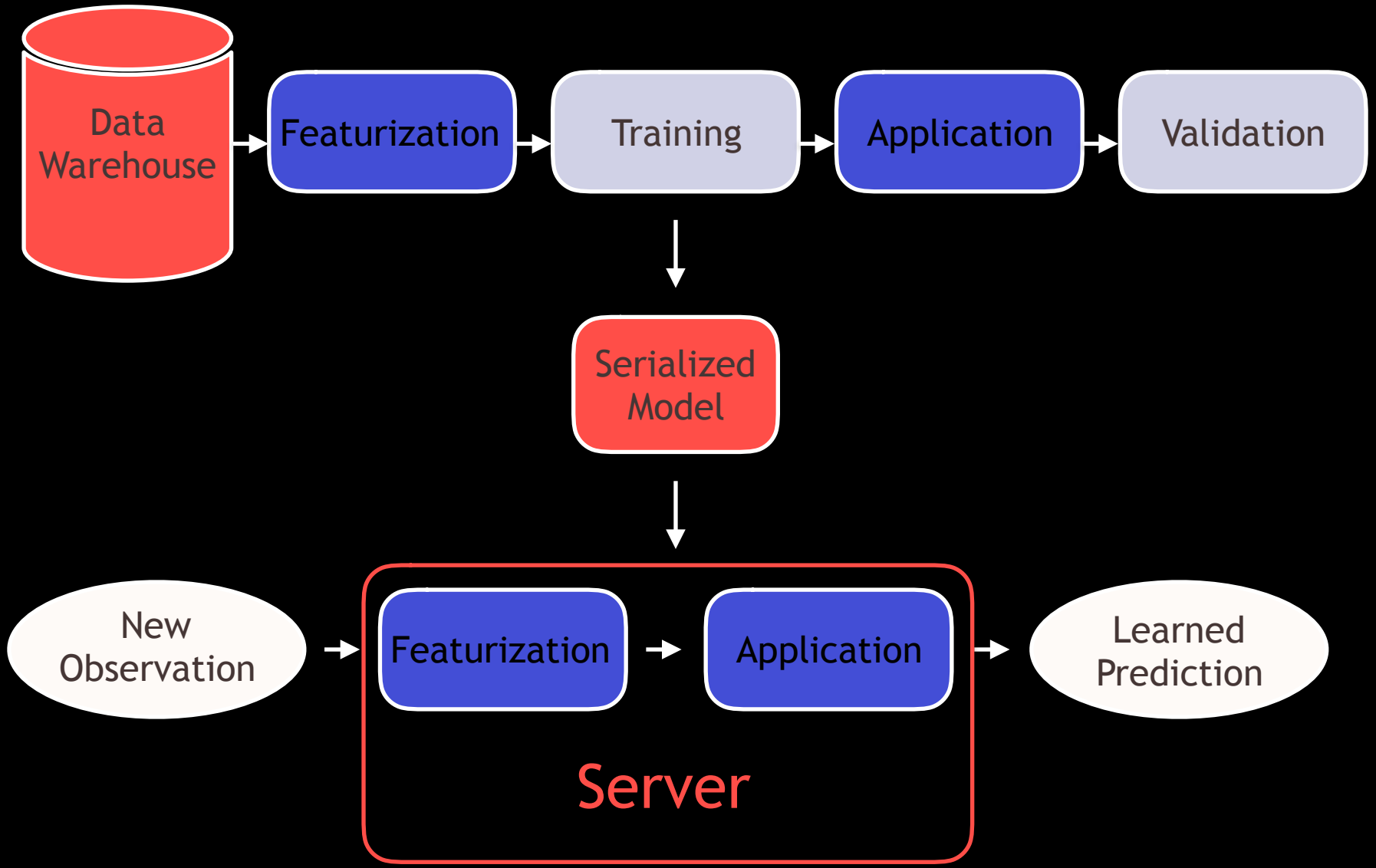
3 Modes of Deployment



What is a model?



Moving Models


Data Warehouse
Featurization
Training
Application
Validation
Serialized Model
New Observation
Featurization
Application
Learned Prediction
Server

@j_houg

Deployment Challenges & Solutions

3 Questions

Does your model do what you need?

Does it meet your engineering requirements?

Is your team organized to build and support it?

Function?

Useful?

George Box
(kind of) said
“All models are wrong, some are useful.”

Useful?

Continues to be useful?

Forever?

Distribution of features
Distribution of predictions

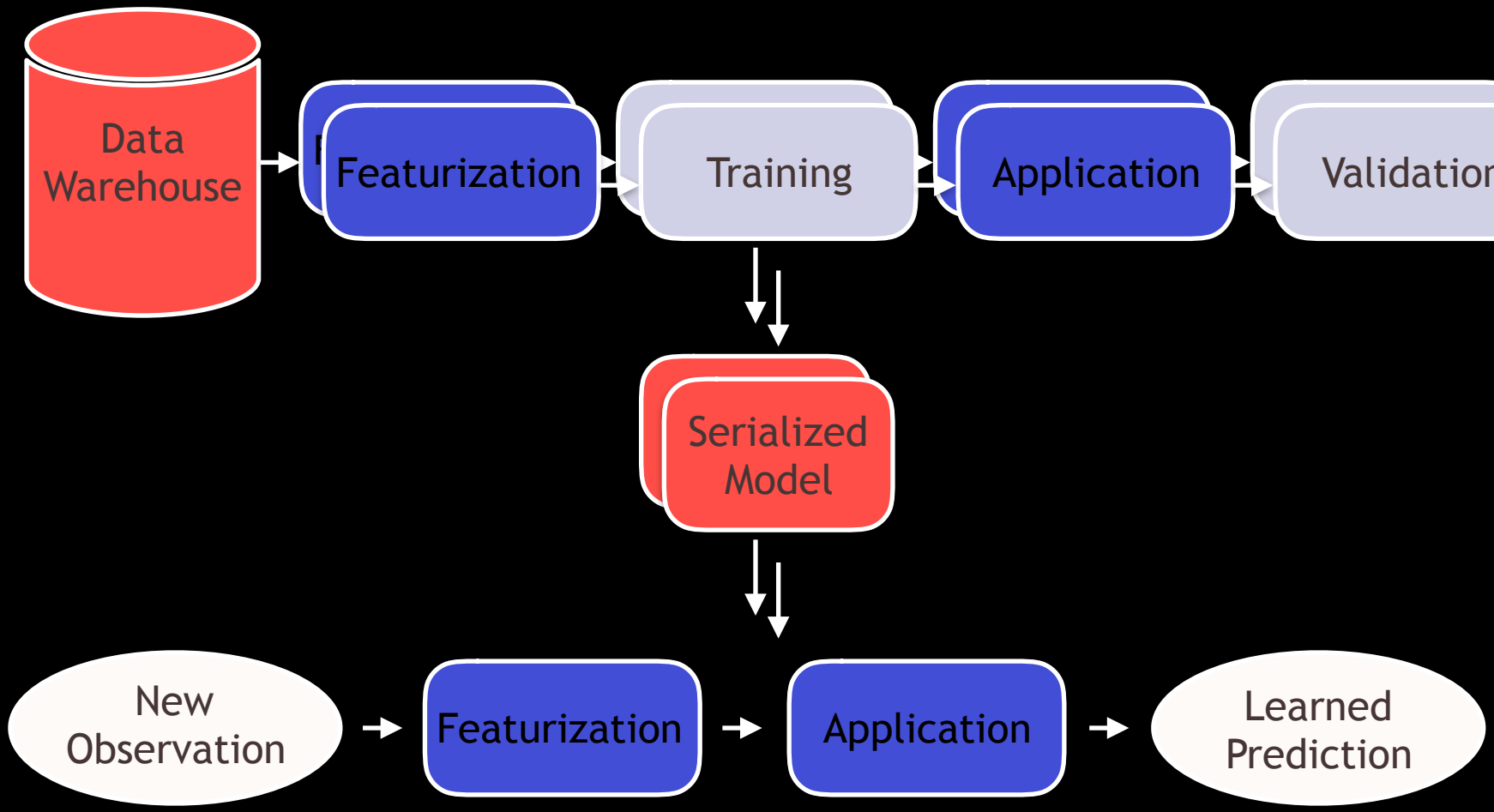
Continues to be useful?

Forever?

Distribution of features
Distribution of predictions

Pipelines

We don't deploy one model, we deploy the process for repeatedly making more


Data Warehouse
Featurization
Training
Application
Validation
Serialized Model
New Observation
Featurization
Application
Learned Prediction

When?

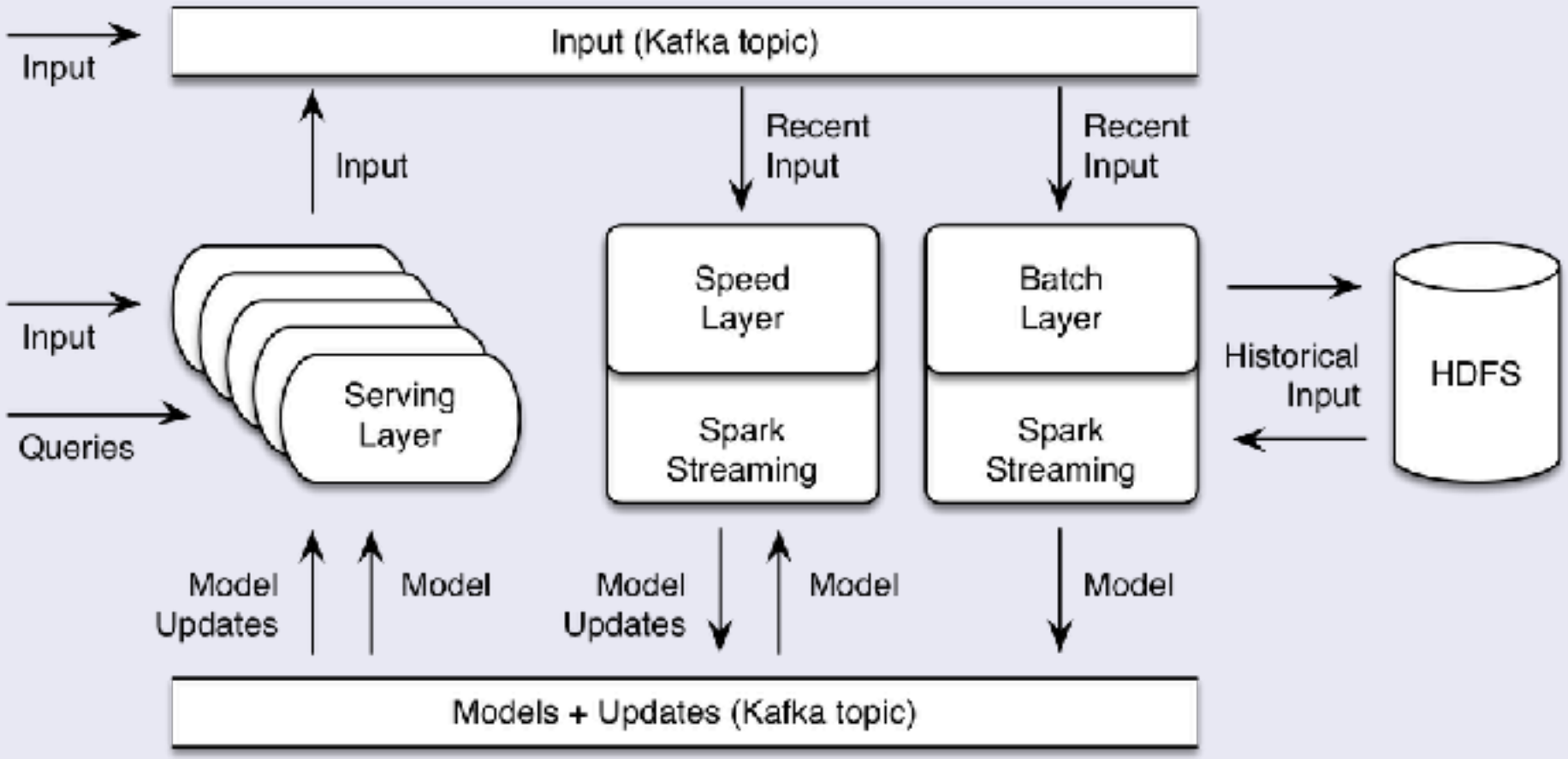
Nightly

Continuously

Scheduled Model Training

Cron

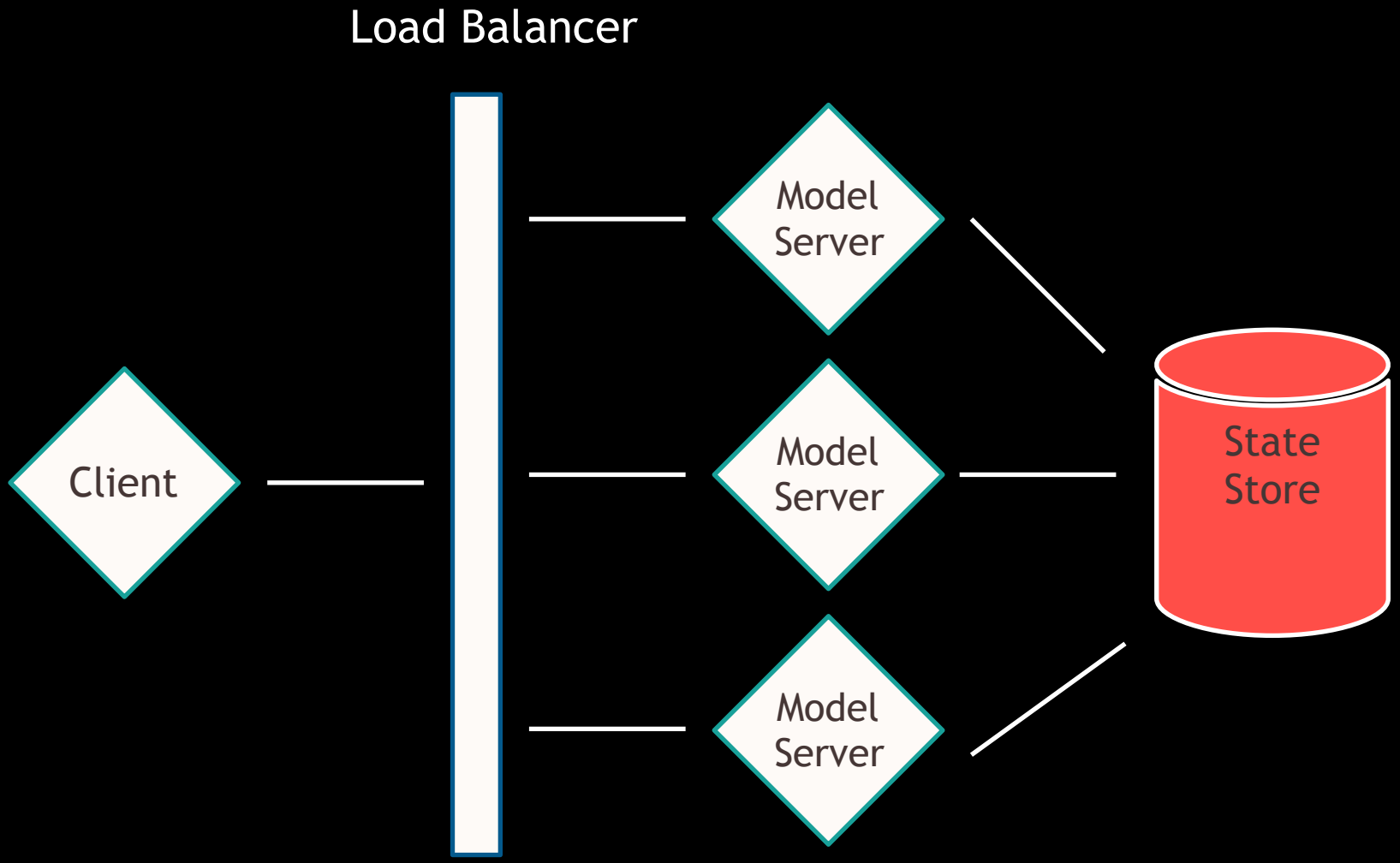
Luigi



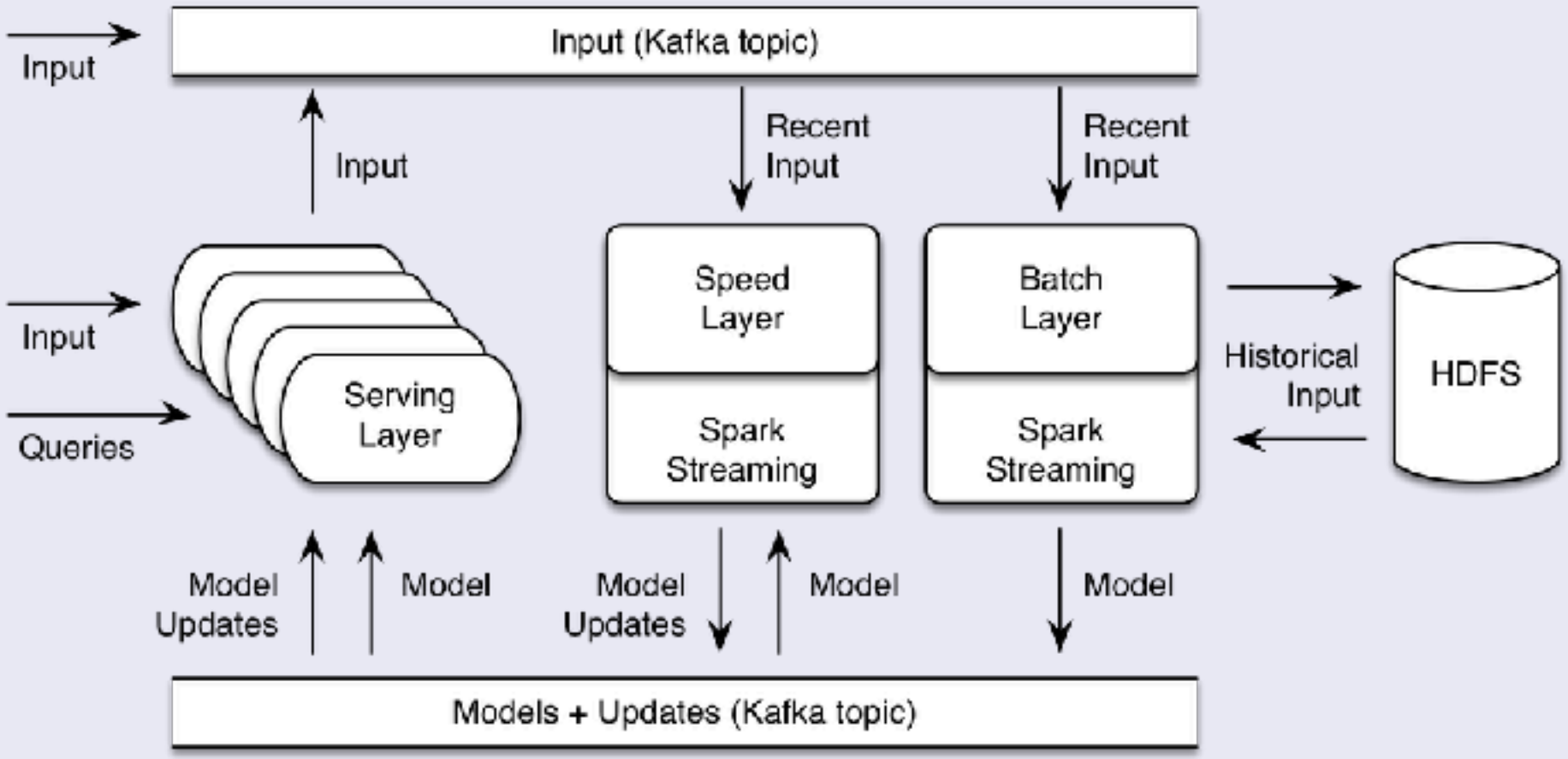
oryx.io and https://www2007.org/papers/paper570.pdf

Meets Requirements?

Throughput?


Load Balancer
Client
Model Server
Model Server
Model Server
State Store

Lambda & Throughput?

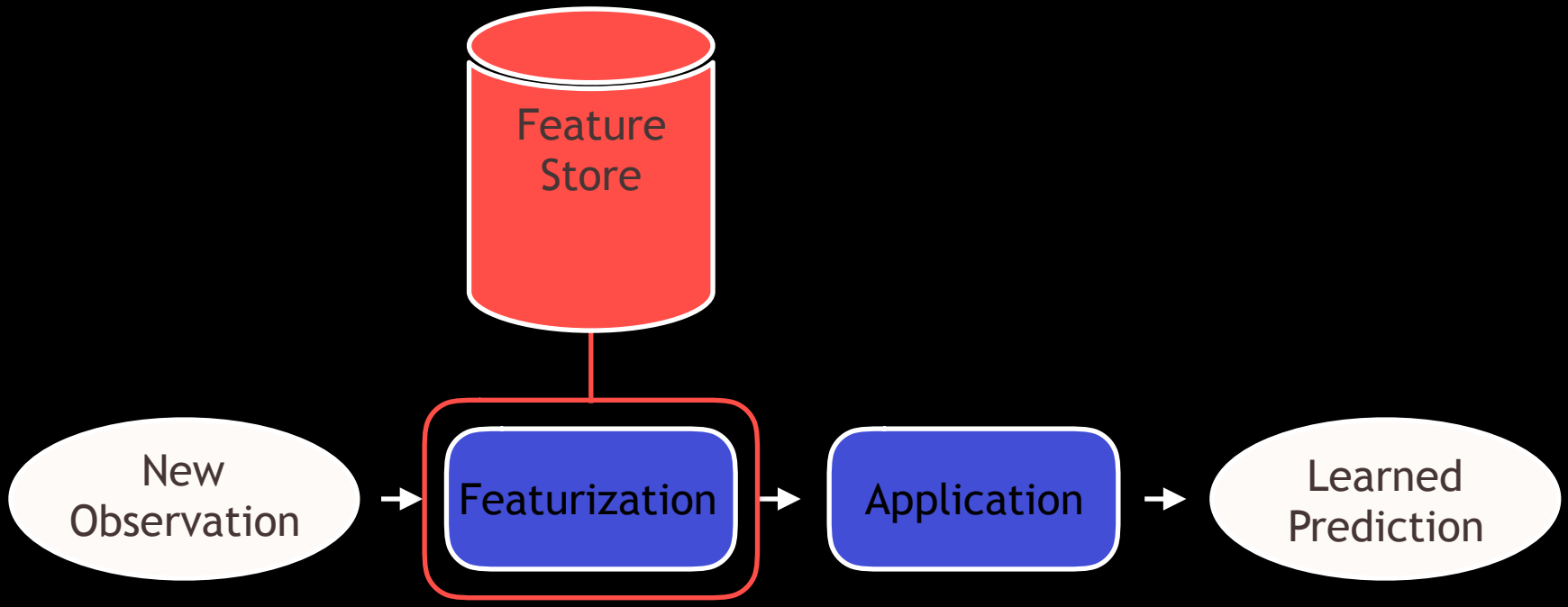


oryx.io and https://www2007.org/papers/paper570.pdf

Fast enough?

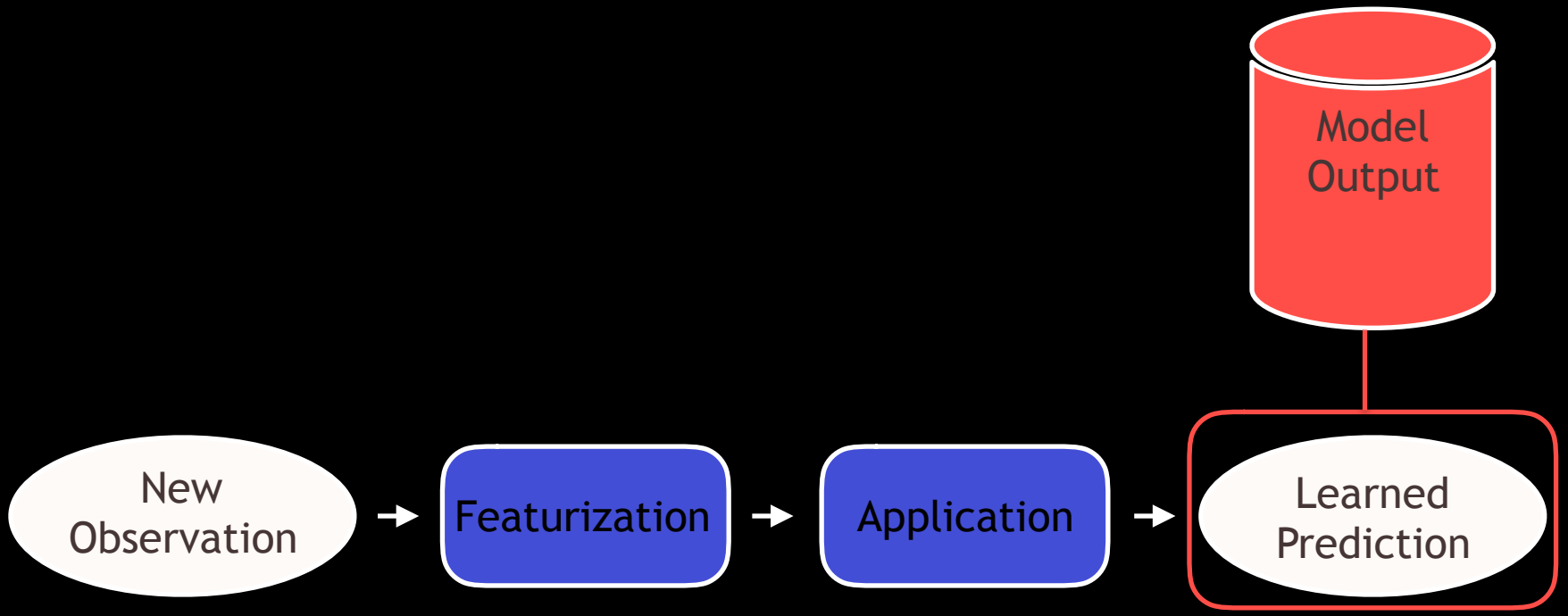
Fast enough?

Materialize model features

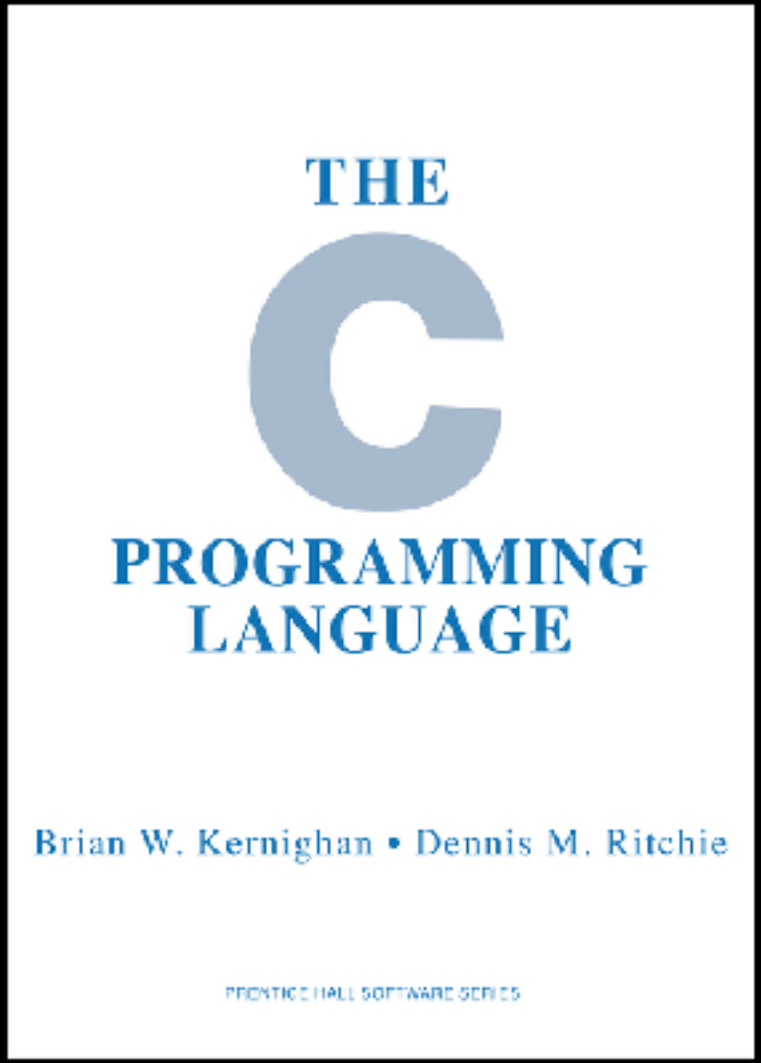


Fast enough?

Materialize model output



Fast enough?


THE
C
PROGRAMMING
LANGUAGE
Brian W. Kernighan • Dennis M. Ritchie

Model Handoff

Model Handoff

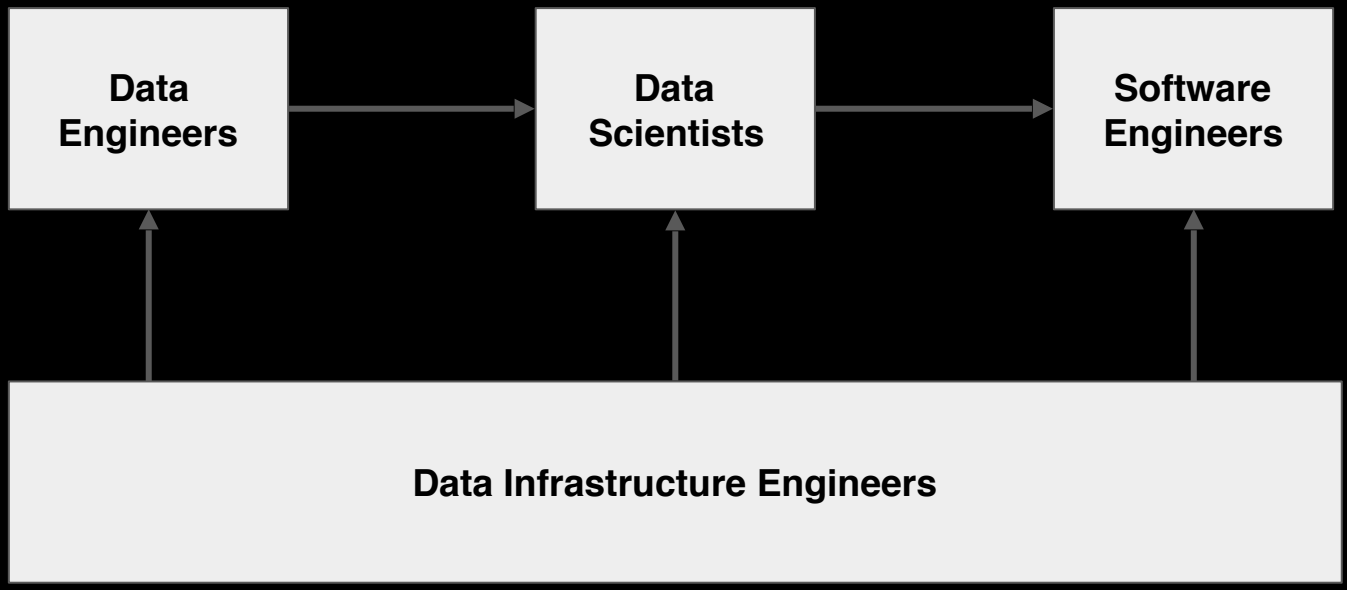
Error Prone


Jane Hu
me reflecting on my life thus far
764
people are talking about this

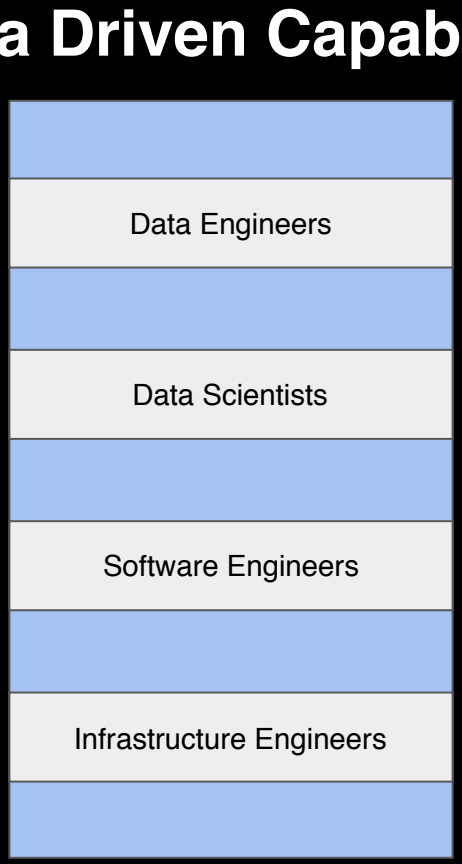
Conway's Law

"organizations which design systems ... are constrained to produce designs which are copies of the communication structures of these organizations."
— M. Conway

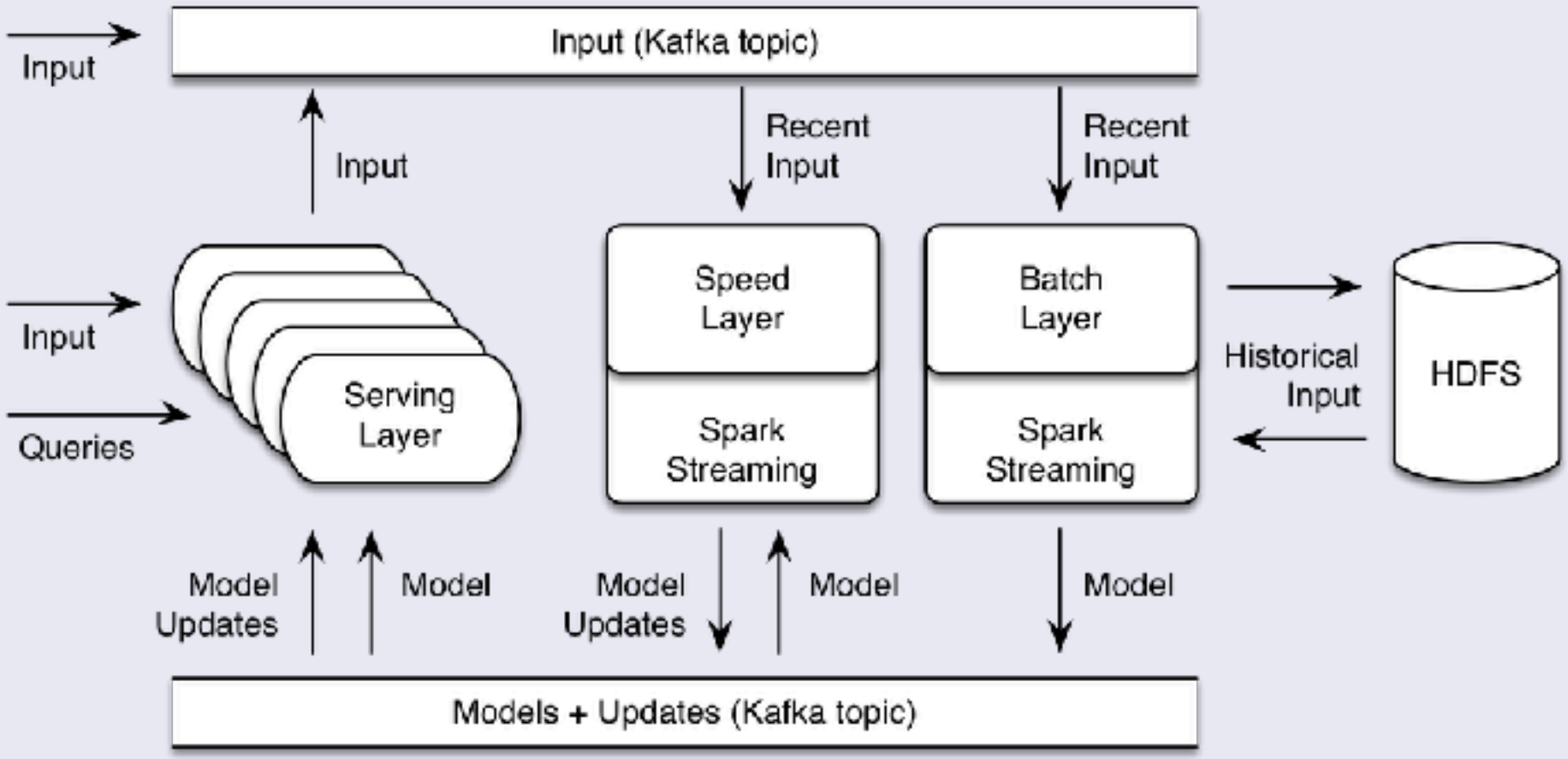
Typical Data Science Department



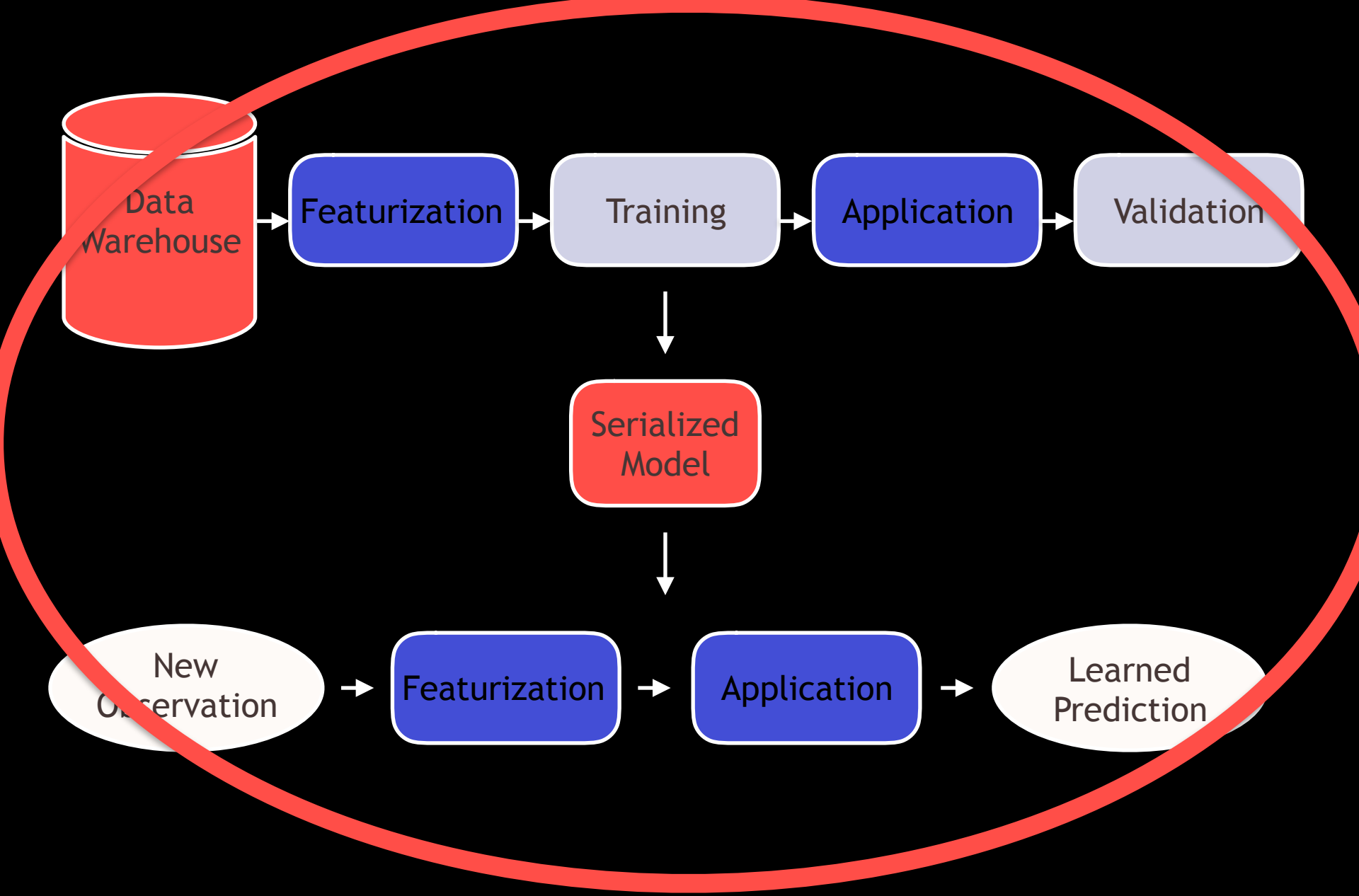
Typical Data Science Department



Lambda & Team Structure



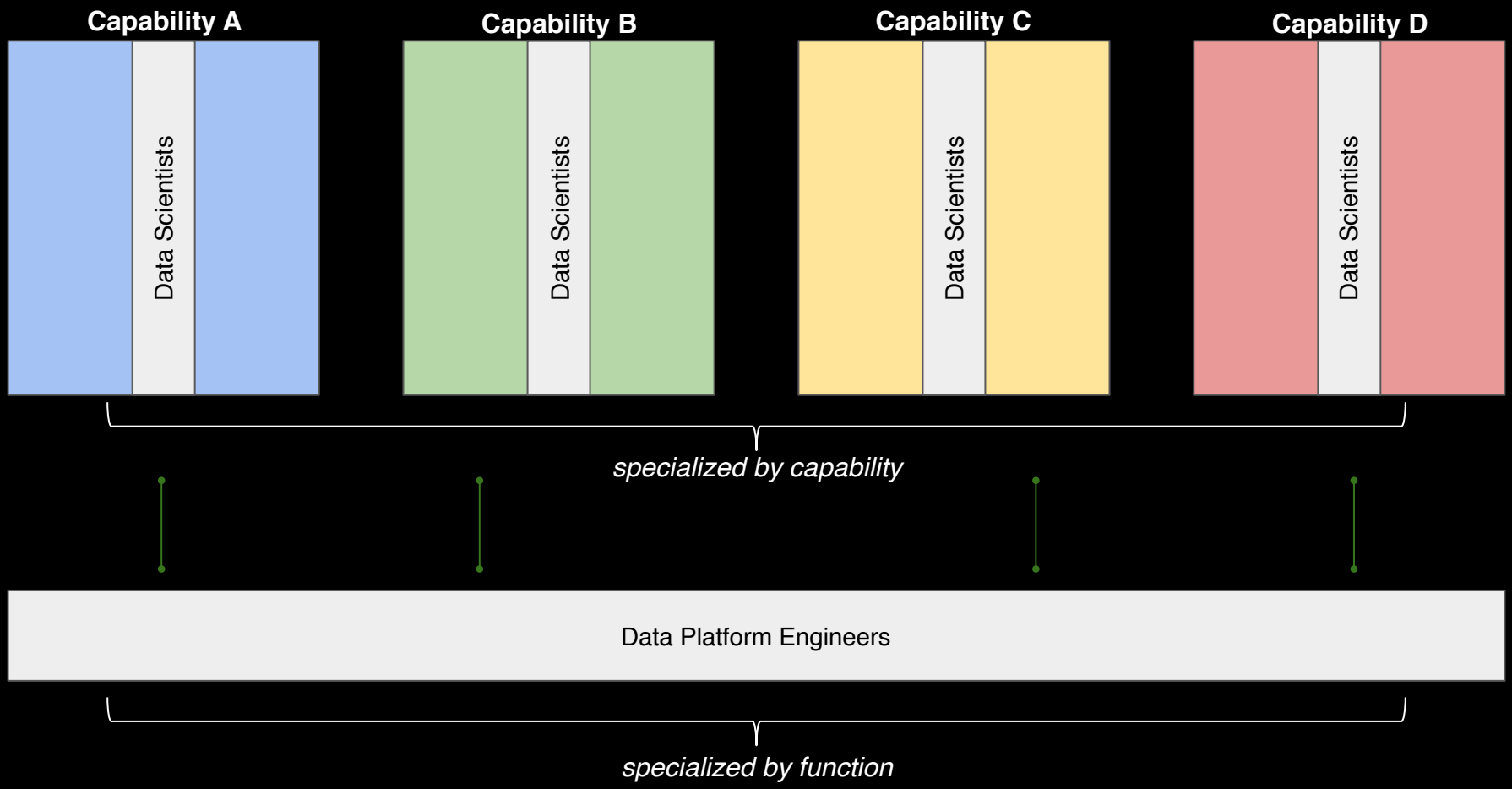
oryx.io and https://www2007.org/papers/paper570.pdf

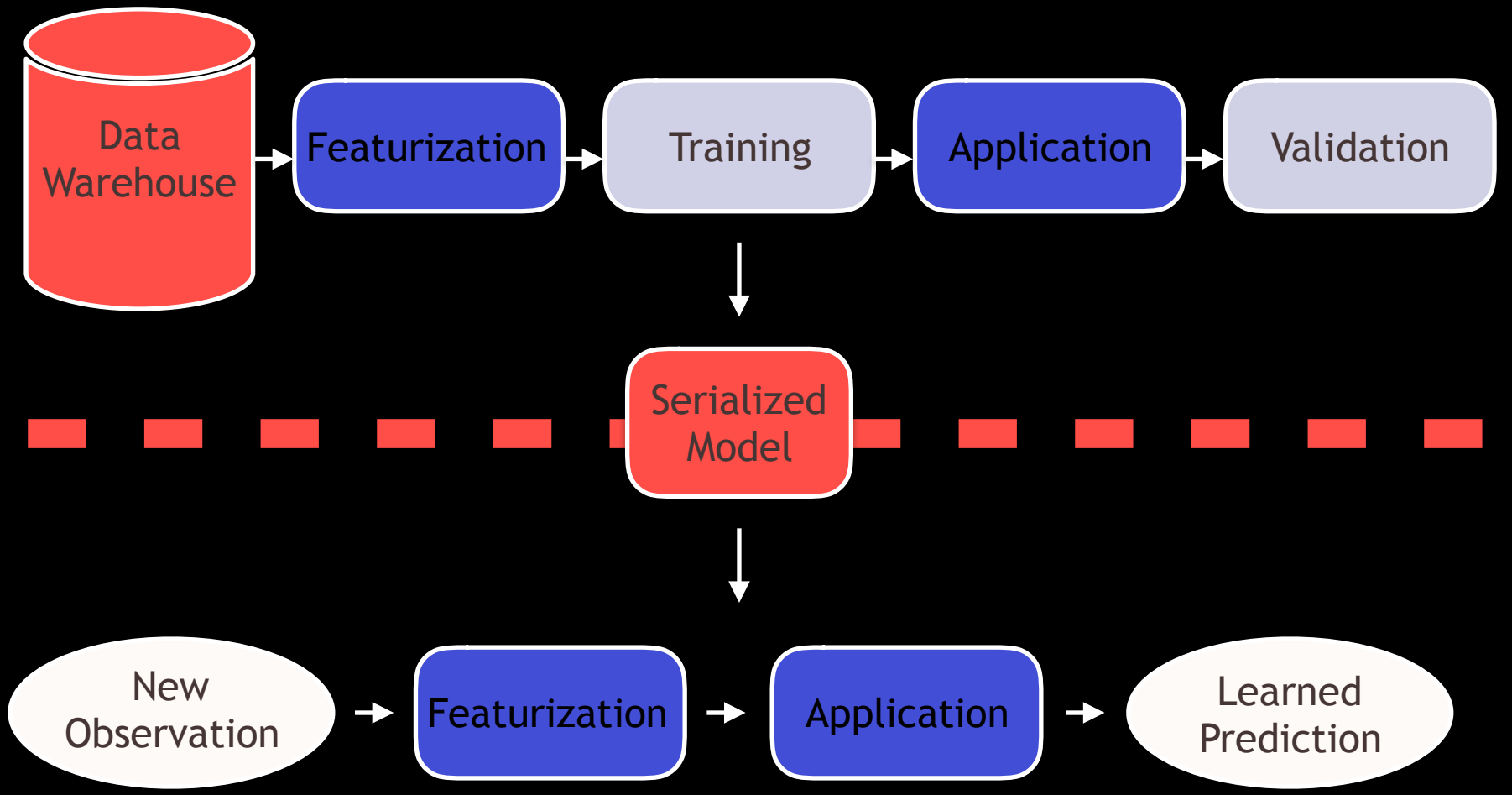

Data Warehouse
Featurization
Training
Application
Validation
Serialized Model
New Observation
Featurization
Application
Learned Prediction

@j_houg

Fix Model Handoff?

Capabilities



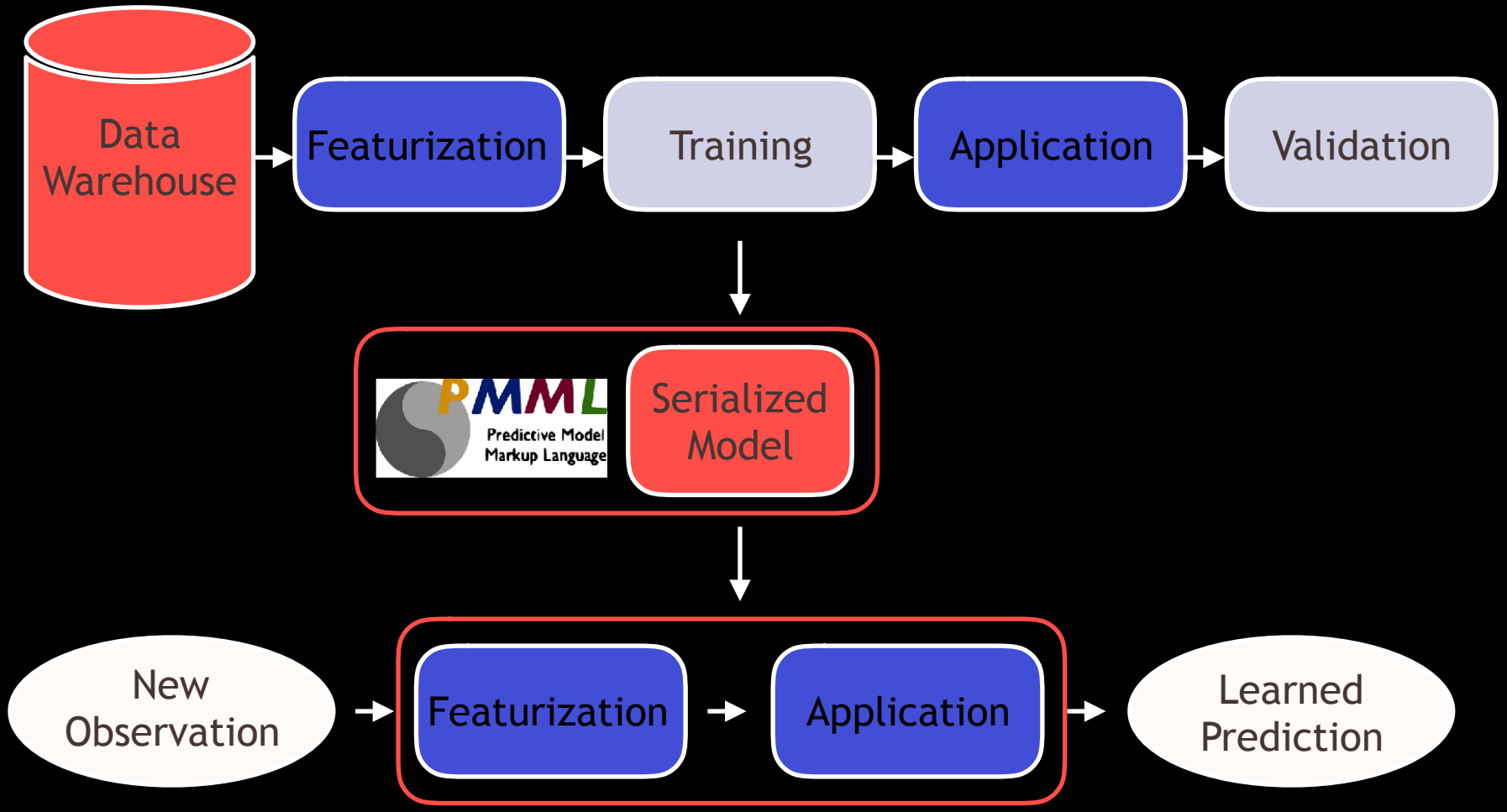

Data Warehouse
Featurization
Training
Application
Validation
Serialized Model
New Observation
Featurization
Application
Learned Prediction

Open Standards

Open Scoring

HTTP method	Endpoint	Required role(s)	Description
GET	/model	-	Get the summaries of all models
POST	/model	admin	Deploy a model
PUT	/model/${id}	admin	Deploy a model
GET	/model/${id}	-	Get the summary of a model
GET	/model/${id}/pmml	admin	Download a model as a PMML document
POST	/model/${id}	-	Evaluate data in "single prediction" mode
POST	/model/${id}/batch	-	Evaluate data in "batch prediction" mode
POST	/model/${id}/csv	-	Evaluate data in "CSV prediction" mode
DELETE	/model/${id}	admin	Undeploy a model

@j_houg

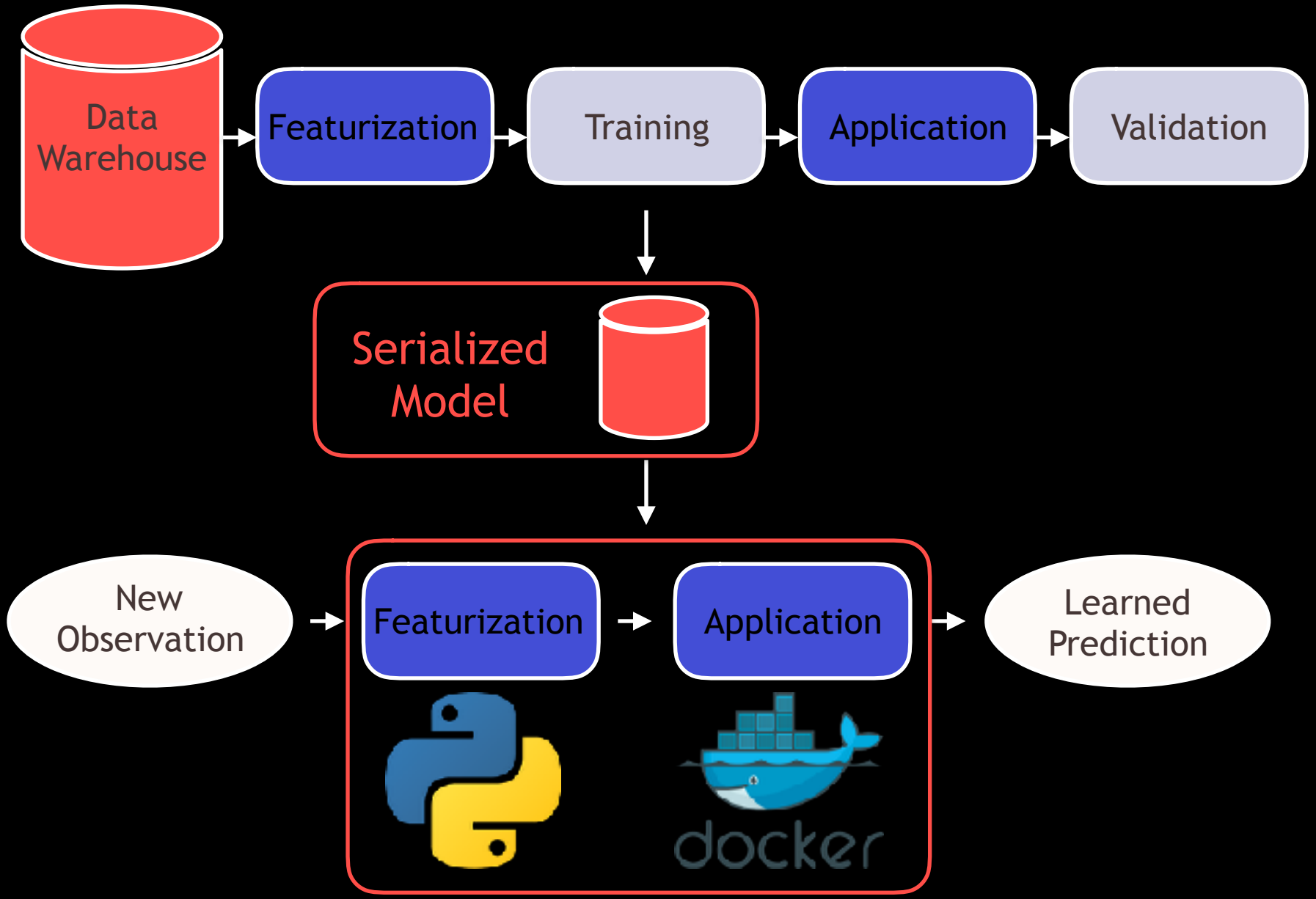


Open Scoring

Open Standards, Limited Choices

Openscoring REST API		PMML 3.0 through 4.2	Association Rules Decision Trees Clustering General Regression Mining Model Naïve Bayes k-NN Neural Network Regression Rule Sets Scorecards SVM

http://dmg.org/pmml/products.html


Data Warehouse
Featurization
Training
Application
Validation
Serialized Model
New Observation
Featurization
Application
Learned Prediction
docker

@j_houg

Does the model...

Do the thing you want it to?

- Functionality
- Usefulness
- Both continually

Meet your requirements?

- Throughput
- Latency
- Freshness

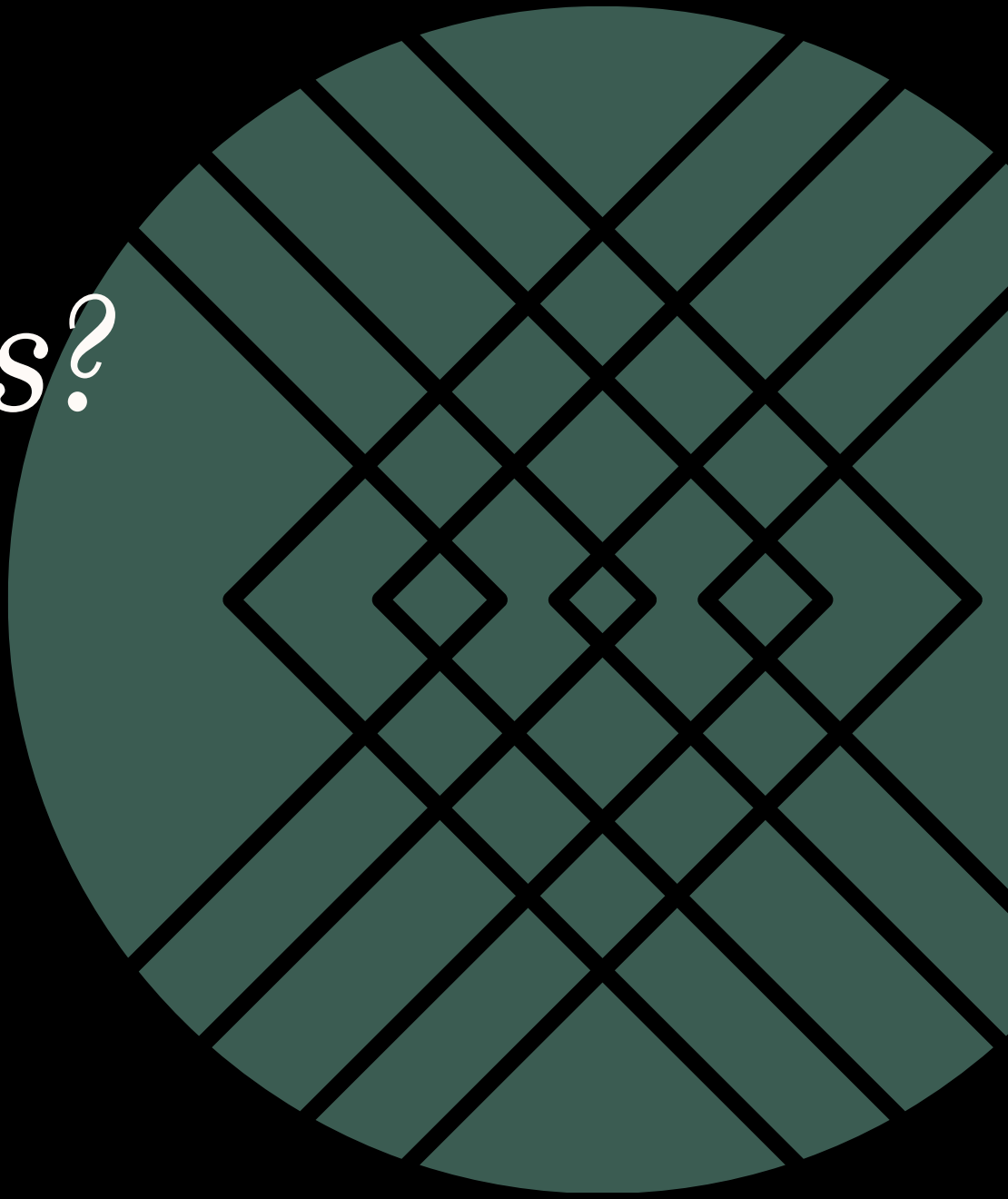
Is your team...

Organized in a way that supports your system and its requirements?

Thanks!

@j_houg

Questions?



@j_houg