aws SUMMIT Switzerland

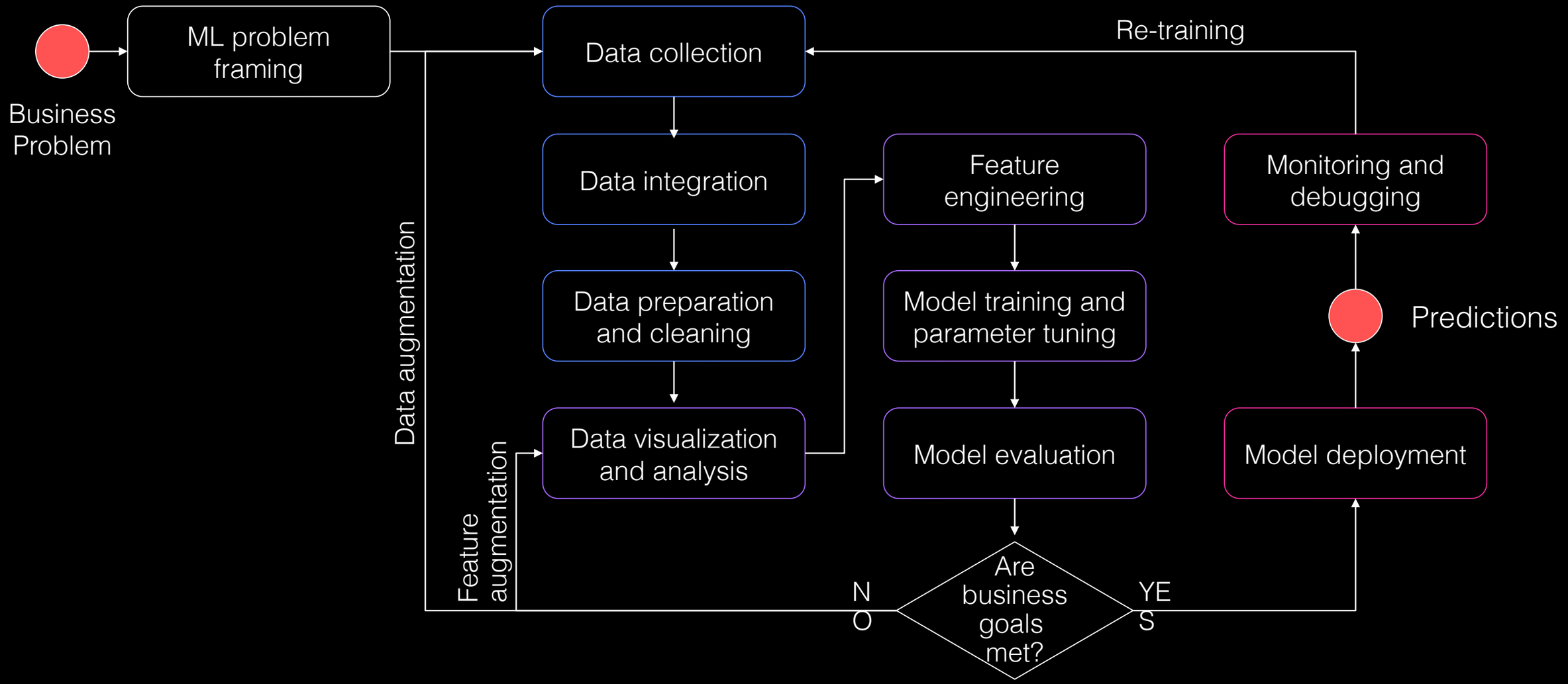
Build, Train and Deploy Machine Learning Models on Amazon SageMaker

Julien Simon
Global Evangelist, AI & Machine Learning
Amazon Web Services
@julsimon

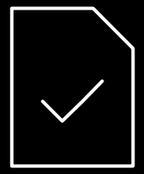
Stéphane Cheikh
Director, Portfolio Evolution using Artificial Intelligence
SITA

aws SUMMIT

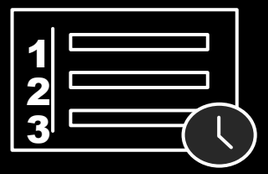
Machine learning cycle



Amazon SageMaker

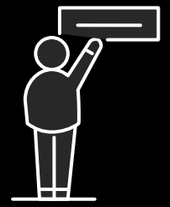


Collect and prepare training data



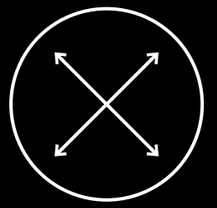
Choose and optimize your ML algorithm

Set up and manage environments for training



Train and tune ML models

Deploy models in production



Scale and manage the production environment

Same service and APIs from experimentation to production

intuit

tinder

KIA

CONVOY

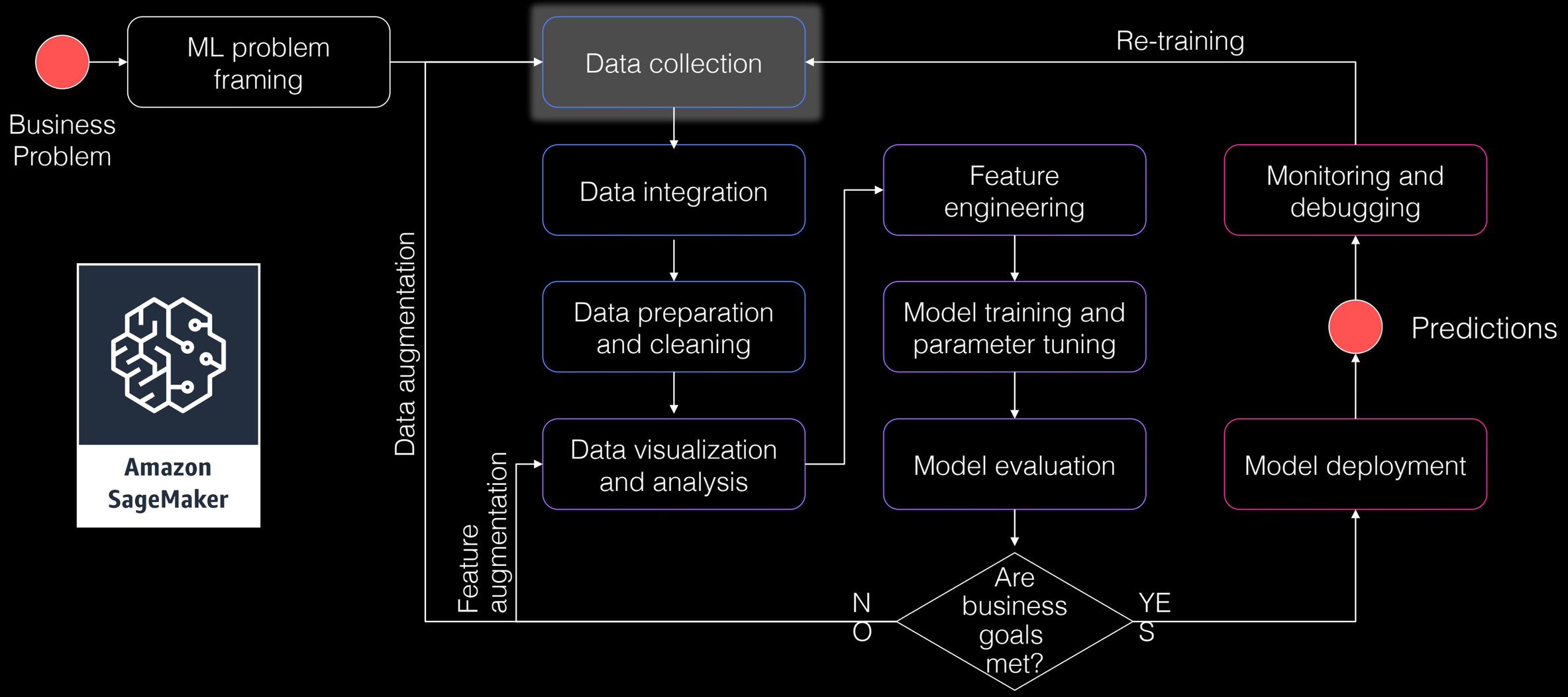
SIEMENS

DOW JONES

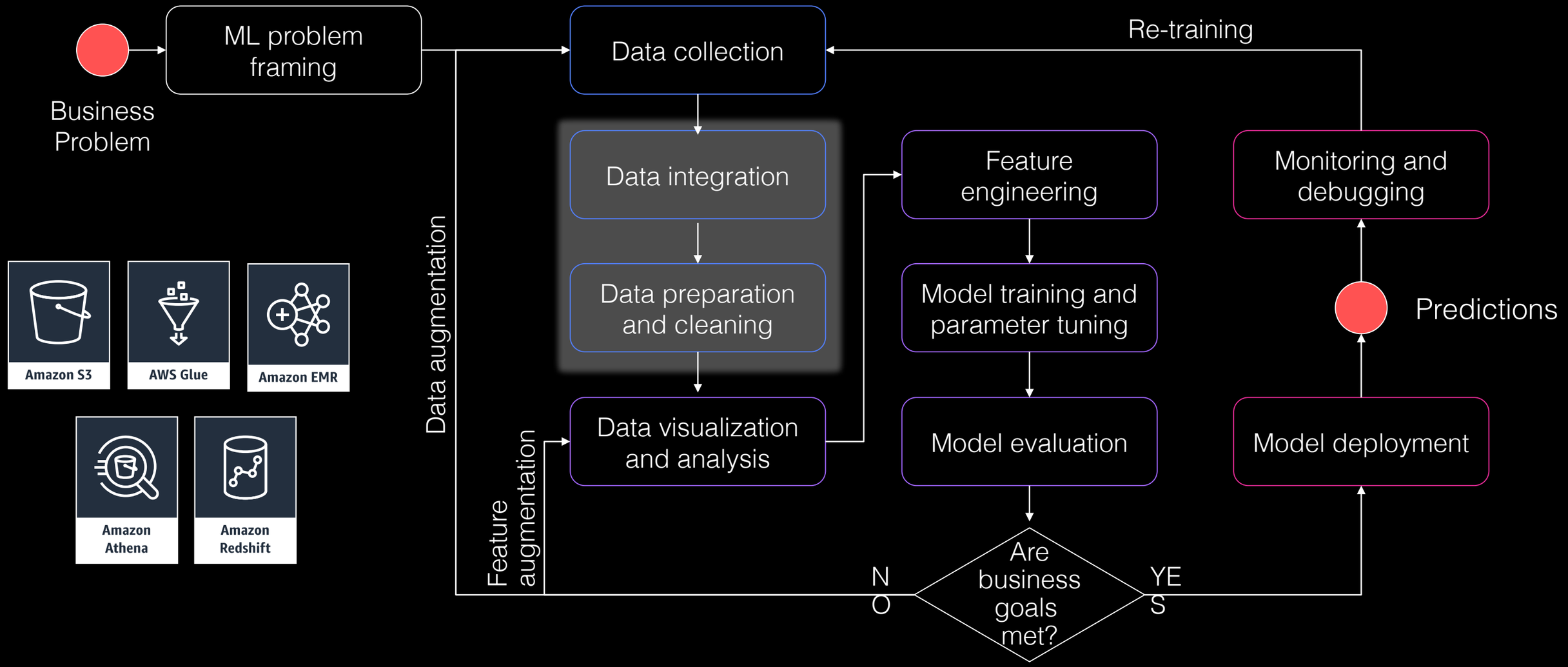
SONY

aws SUMMIT

Build your dataset

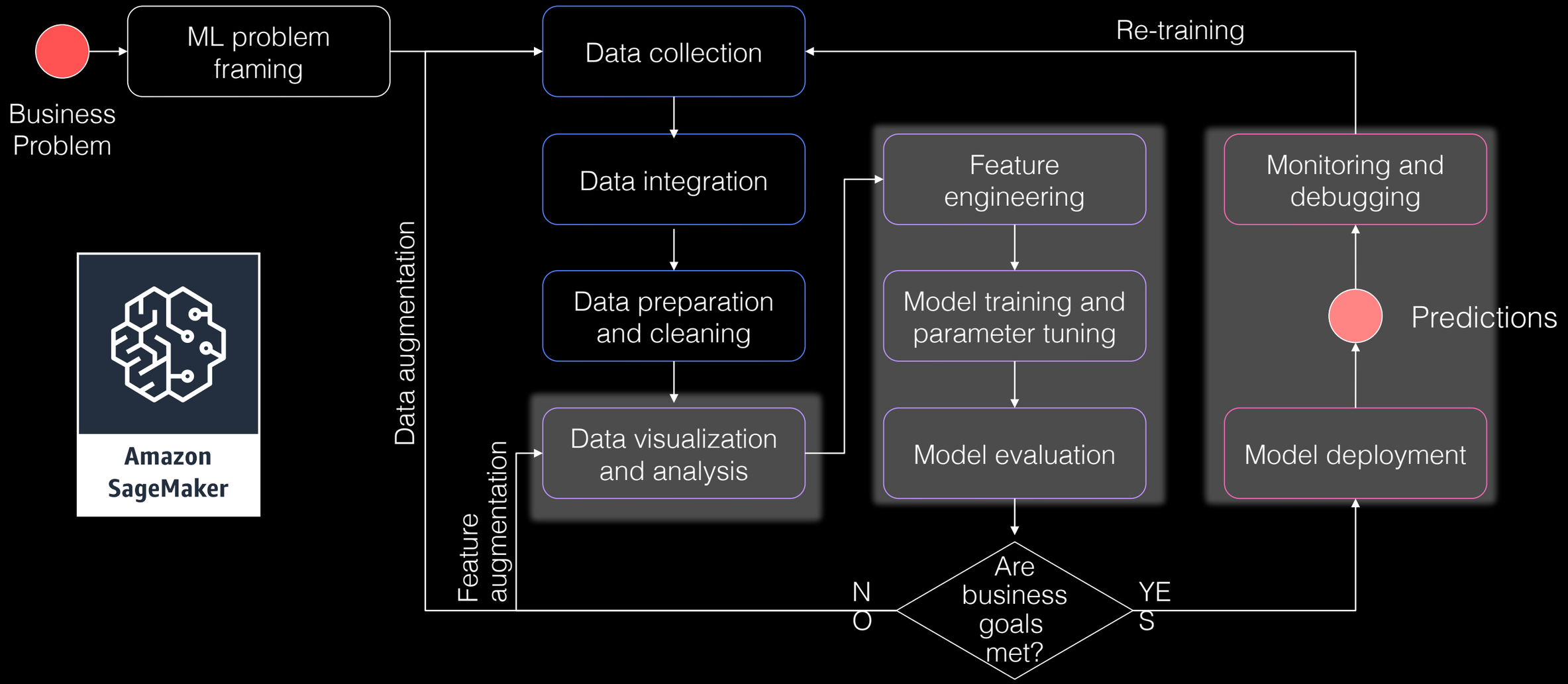


Prepare your dataset for Machine Learning



aws SUMMIT

Build, train and deploy models using SageMaker



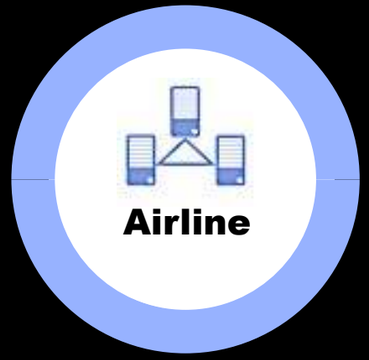
Flight Delay Prediction

Stephane Cheikh
Director, Portfolio Evolution using Artificial Intelligence

SITA

SITA at a glance

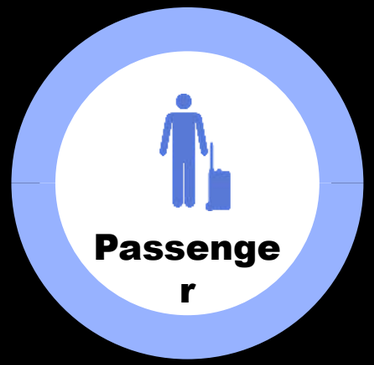
AIR TRAVEL SOLUTIONS



Airline operations

Airport operations,
Baggage processing,
Passenger processing

Border management



SITA Passenger Service System (SITA PSS)

Aircraft operations (cockpit and cabin services),
Connected aircraft

Cargo management, community integration, eCargo

aws SUMMIT

SITA

Disruption: Our Industry Issue

Airline Impact

- Delays
- Diversions & Cancellations
- Crew Time Limits
- Asset (Aircraft) Usage

Airport Impact

- Capacity
- Surges
- Out-of-Hours
- Extended Stays

Passenger Impact

- Baggage
- Accommodation
- Re-Booking
- Keeping Informed

Cargo Impact

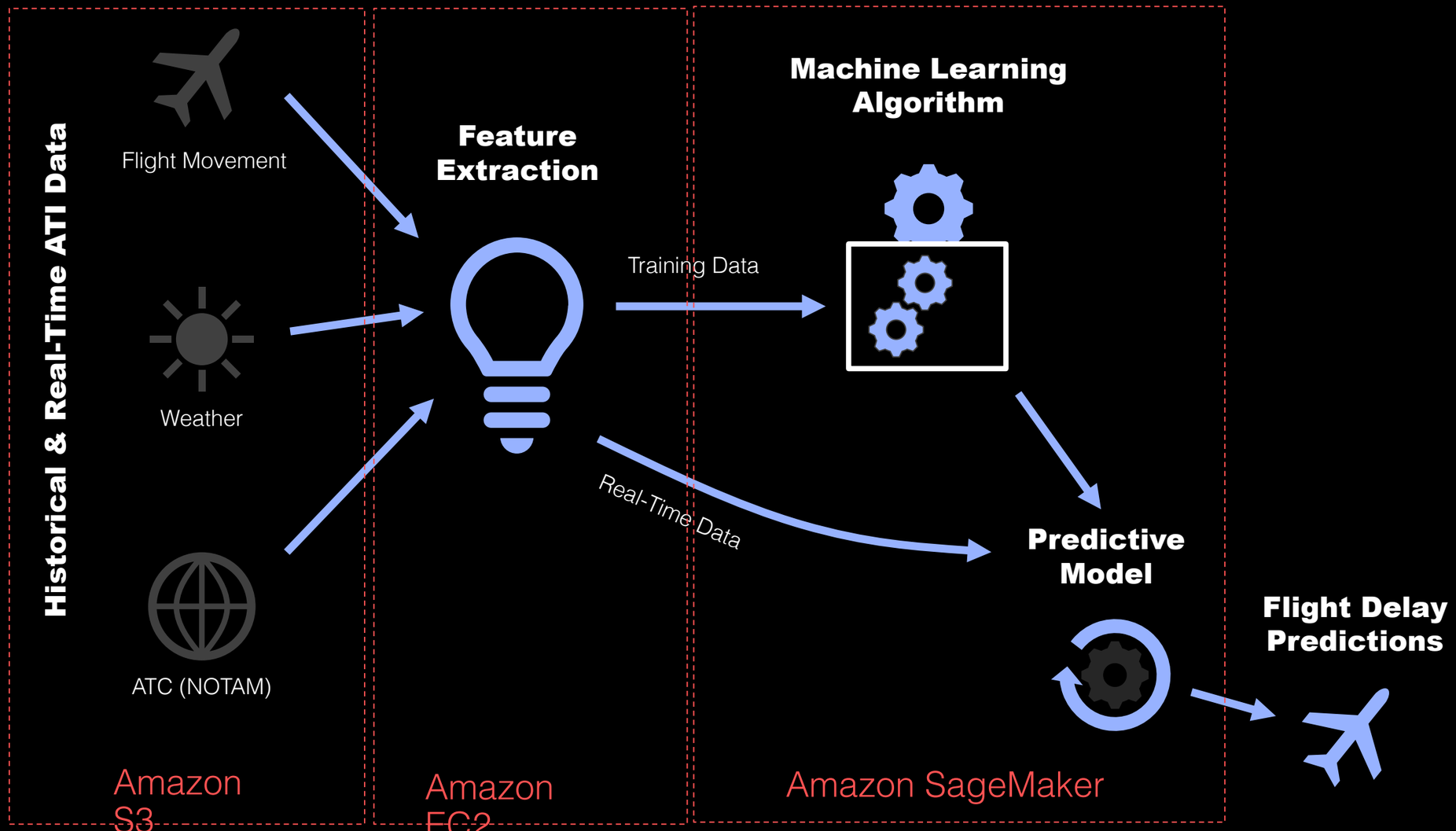
- Perished Goods
- Storage Capacity
- Delivery Delays

76% Global OTP

51 mins Average Flight Delay Time

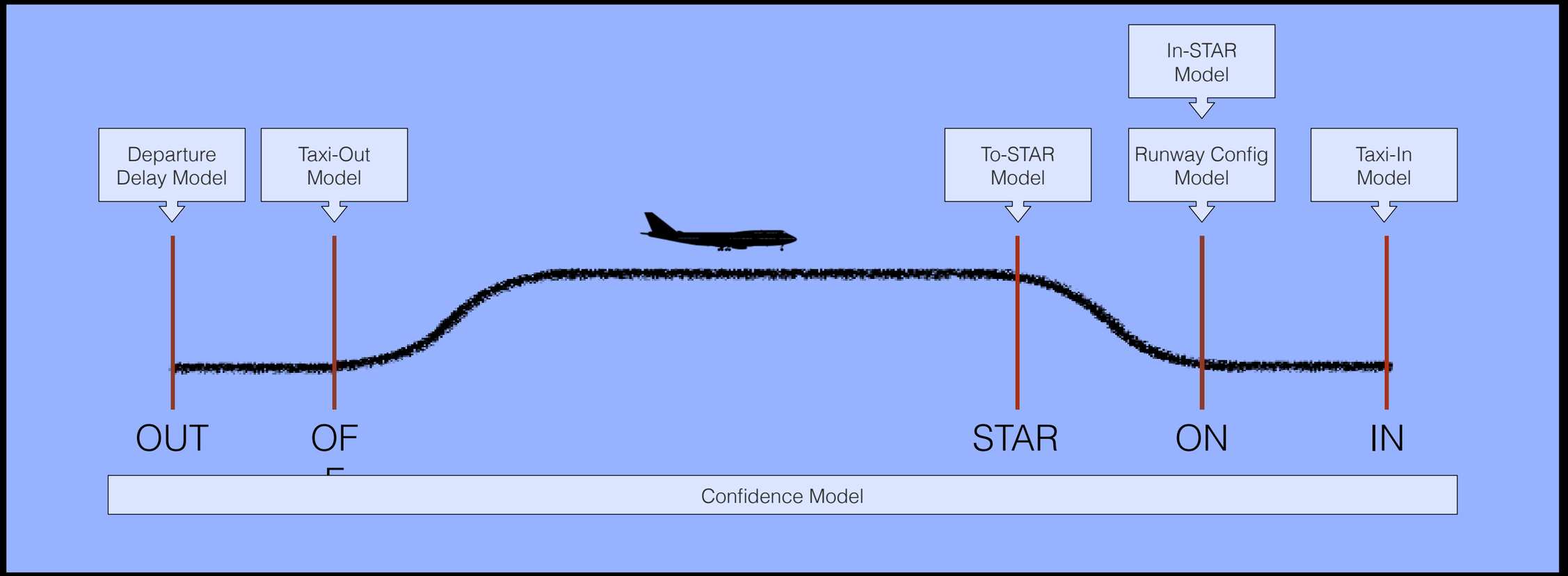
>$25Bn Global Cost of Delay

Applying AI to predict potential flight disruptions



Seven models working together

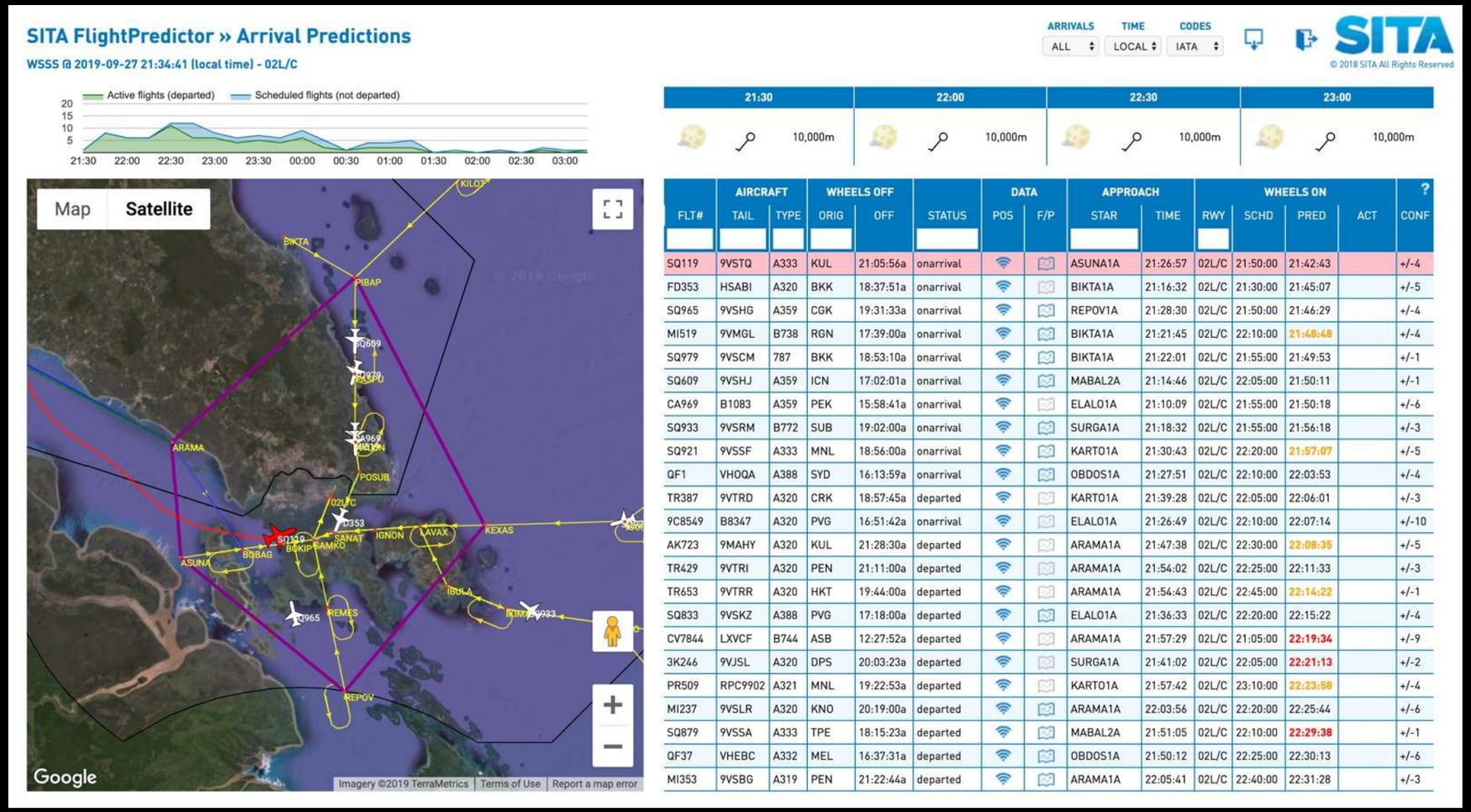
All machine learning models re-usable and adaptable for different airports



SITA

SITA Flight Predictor

Long look-ahead predictions of flight schedule deviations for airline and airport users



FLT#	AIRCRAFT TAIL	AIRCRAFT TYPE	WHEELS OFF ORIG	WHEELS OFF OFF	STATUS	DATA POS	DATA F/P	APPROACH STAR	APPROACH TIME	WHEELS ON RWY	WHEELS ON SCHD	WHEELS ON PRED	WHEELS ON ACT	? CONF
SQ119	9VSTQ	A333	KUL	21:05:56a	onarrival			ASUNA1A	21:26:57	02L/C	21:50:00	21:42:43		+/-4
FD353	HSABI	A320	BKK	18:37:51a	onarrival			BIKTA1A	21:16:32	02L/C	21:30:00	21:45:07		+/-5
SQ965	9VSHG	A359	CGK	19:31:33a	onarrival			REPOV1A	21:28:30	02L/C	21:50:00	21:46:29		+/-4
MI519	9VMGL	B738	RGN	17:39:00a	onarrival			BIKTA1A	21:21:45	02L/C	22:10:00	21:48:48		+/-4
SQ979	9VSCM	787	BKK	18:53:10a	onarrival			BIKTA1A	21:22:01	02L/C	21:55:00	21:49:53		+/-1
SQ609	9VSHJ	A359	ICN	17:02:01a	onarrival			MABAL2A	21:14:46	02L/C	22:05:00	21:50:11		+/-1
CA969	B1083	A359	PEK	15:58:41a	onarrival			ELALO1A	21:10:09	02L/C	21:55:00	21:50:18		+/-6
SQ933	9VSRM	B772	SUB	19:02:00a	onarrival			SURGA1A	21:18:32	02L/C	21:55:00	21:56:18		+/-3
SQ921	9VSSF	A333	MNL	18:56:00a	onarrival			KARTO1A	21:30:43	02L/C	22:20:00	21:57:07		+/-5
QF1	VHOQA	A388	SYD	16:13:59a	onarrival			OBDOS1A	21:27:51	02L/C	22:10:00	22:03:53		+/-4
TR387	9VTRD	A320	CRK	18:57:45a	departed			KARTO1A	21:39:28	02L/C	22:05:00	22:06:01		+/-3
9C8549	B8347	A320	PVG	16:51:42a	onarrival			ELALO1A	21:26:49	02L/C	22:10:00	22:07:14		+/-10
AK723	9MAHY	A320	KUL	21:28:30a	departed			ARAMA1A	21:47:38	02L/C	22:30:00	22:08:35		+/-5
TR429	9VTRI	A320	PEN	21:11:00a	departed			ARAMA1A	21:54:02	02L/C	22:25:00	22:11:33		+/-3
TR653	9VTRR	A320	HKT	19:44:00a	departed			ARAMA1A	21:54:43	02L/C	22:45:00	22:14:22		+/-1
SQ833	9VSKZ	A388	PVG	17:18:00a	departed			ELALO1A	21:36:33	02L/C	22:20:00	22:15:22		+/-4
CV7844	LXVCF	B744	ASB	12:27:52a	departed			ARAMA1A	21:57:29	02L/C	21:05:00	22:19:34		+/-9
3K246	9VJSL	A320	DPS	20:03:23a	departed			SURGA1A	21:41:02	02L/C	22:05:00	22:21:13		+/-2
PR509	RPC9902	A321	MNL	19:22:53a	departed			KARTO1A	21:57:42	02L/C	23:10:00	22:23:58		+/-4
MI237	9VSLR	A320	KNO	20:19:00a	departed			ARAMA1A	22:03:56	02L/C	22:20:00	22:25:44		+/-6
SQ879	9VSSA	A333	TPE	18:15:23a	departed			MABAL2A	21:51:05	02L/C	22:10:00	22:29:38		+/-1
QF37	VHEBC	A332	MEL	16:37:31a	departed			OBDOS1A	21:50:12	02L/C	22:25:00	22:30:13		+/-6
MI353	9VSBG	A319	PEN	21:22:44a	departed			ARAMA1A	22:05:41	02L/C	22:40:00	22:31:28		+/-3

aws SUMMIT

SITA

What we achieved

Reliable and accurate predictions are possible up to 6 hours out

- “Fine tuning” will ensure consistent and reliable performance of all models
- Incorporating winds aloft data will improve predictions at 6 hours out
- New departure delay model will also improve accuracy of longer lead-time predictions

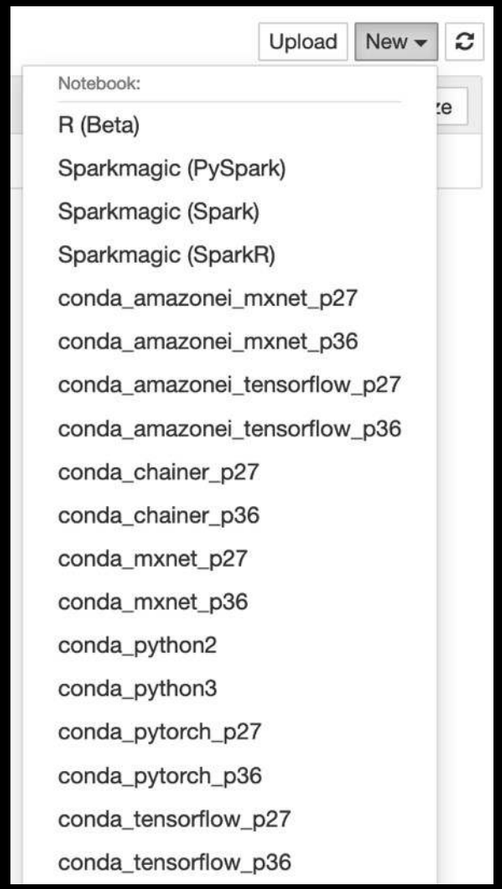
Prediction Time Interval	Prediction Accuracy
30 mins	±5 mins
1 - 2 hours	±10 mins
2 - 4 hours	±15 mins
6 hours	±15 mins

Building models

aws SUMMIT

Notebook instances

- Fully managed EC2 instances, from *ml.t2.medium* to *ml.p3.16xlarge*
- Pre-installed with Jupyter and Conda environments
  - Python 2.7 & 3.6
  - Open-source libraries (TensorFlow, Apache MXNet, etc.)
  - Beta support for R – NEW!
  - Amazon Elastic Inference for cost-effective GPU acceleration
- Lifecycle configurations
- VPC, encryption, etc.
- Get to work in minutes, zero setup



Model options

Training code

AWS Marketplace for Machine Learning: 250+ off-the-shelf algos and models

Factorization Machines
Linear Learner
Principal Component Analysis
K-Means Clustering
XGBoost
And more

Built-in Algorithms (17)

No ML coding required
No infrastructure work required
Distributed training
Pipe mode

mxnet Chainer TensorFlow PyTorch scikit learn dmlc XGBoost

NEW !

Built-in Frameworks

Bring your own code: Script mode
Open-source containers
No infrastructure work required
Distributed training
Pipe mode

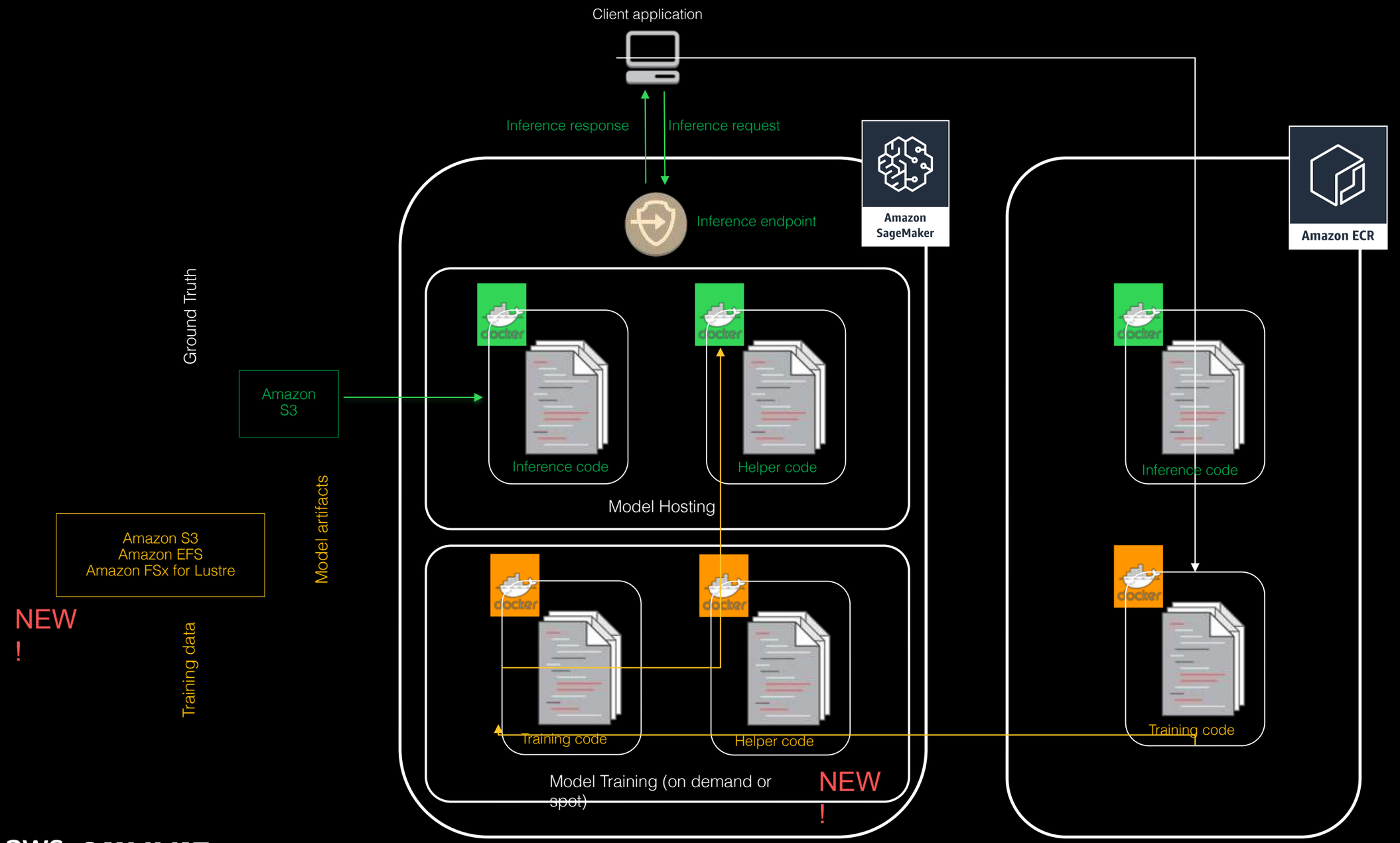
Bring Your Own Container

Full control, run anything!
R, C++, etc.
No infrastructure work required

aws SUMMIT

The Amazon SageMaker API

- Python SDK orchestrating all Amazon SageMaker activity
  - High-level objects for algorithm selection, training, deploying, automatic model tuning, etc.
    https://github.com/aws/sagemaker-python-sdk
  - Spark SDK (Python & Scala)
    https://github.com/aws/sagemaker-spark/tree/master/sagemaker-spark-sdk
- AWS SDK
  - Service-level APIs for scripting and automation
  - CLI: '*aws sagemaker*'
  - Language SDKs: boto3, etc.



aws SUMMIT

Built-in algorithms

Built-in algorithms

Orange: supervised, yellow: unsupervised

Linear Learner: Regression, classification	**Image Classification**: Deep learning (ResNet)
Factorization Machines: Regression, classification, recommendation	**Object Detection (SSD)**: Deep learning (VGG or ResNet)
K-Nearest Neighbors: Non-parametric regression and classification	**Neural Topic Model**: Topic modeling
XGBoost: Regression, classification, ranking https://github.com/dmlc/xgboost	**Latent Dirichlet Allocation**: Topic modeling (mostly)
K-Means: Clustering	**BlazingText**: GPU-based Word2Vec, and text classification
Principal Component Analysis: Dimensionality reduction	**Sequence to Sequence**: Machine translation, speech to text and more
Random Cut Forest: Anomaly detection	**DeepAR**: Time-series forecasting (RNN)
Object2Vec: General-purpose embedding	**IP Insights**: Usage patterns for IP addresses
Semantic Segmentation: Deep learning	

Demo:

Sentence classification with BlazingText

https://github.com/awslabs/amazon-sagemaker-examples/tree/master/introduction_to_amazon_algorithms/blazingtext_text_classification_dbpedia

Built-in frameworks

Built-in frameworks: Just add your code

- Built-in containers for training and prediction
  - Open-source, e.g., https://github.com/aws/sagemaker-tensorflow-containers
  - Build them, run them on your own machine, customize them, etc.

- Local mode: Train and predict on your notebook instance, or on your local machine

- Script mode: Reuse existing code with minimal changes

TensorFlow on AWS

C5 instances (Intel Skylake)

Training ResNet-50 with the ImageNet dataset using our optimized build of TensorFlow 1.11 on a c5.18xlarge instance type is designed to be 11x faster than training on the stock binaries

P3 instances (NVIDIA V100)

TensorFlow scaling efficiency with 256 GPUs

65 % Stock version

90 % AWS-optimized version

Apache MXNet: Deep learning for enterprise developers

Start with off-the-shelf models

- Gluon CV, Gluon NLP, Gluon TS
- ONNX compatibility

Fast and scalable training

- Keras-MXNet up to 2x faster than Keras-TensorFlow
- Near-linear scalability up to 256 GPUs
- Dynamic training

Easy deployment

- Java and Scala APIs
- Model Server

Demo:

Fashion-MNIST classification with Keras/TensorFlow

+ Script Mode
+ Managed Spot Training
+ Elastic Inference

https://aws.amazon.com/blogs/machine-learning/train-and-deploy-keras-models-with-tensorflow-and-apache-mxnet-on-amazon-sagemaker/

https://gitlab.com/juliensimon/dlnotebooks/tree/master/keras/05-keras-blog-post

Getting started

http://aws.amazon.com/free

https://ml.aws
https://aws.amazon.com/sagemaker

https://github.com/aws/sagemaker-python-sdk
https://github.com/aws/sagemaker-spark
https://github.com/awslabs/amazon-sagemaker-examples

https://gitlab.com/juliensimon/dlnotebooks

Thank you!

Julien Simon
Global Evangelist, AI & Machine Learning
Amazon Web Services
@julsimon

Stéphane Cheikh
Director, Portfolio Evolution using Artificial Intelligence
SITA

aws SUMMIT

Please complete the session survey.