AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

AIM321

Use AutoML to create high-quality models

Kate Zimmerman (she/her)
Sr. Manager, Amazon Machine Learning Solutions Lab
Amazon Web Services

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Today’s agenda

Overview of AutoML at AWS

Deep dive into Amazon SageMaker Autopilot

Customers using AutoML

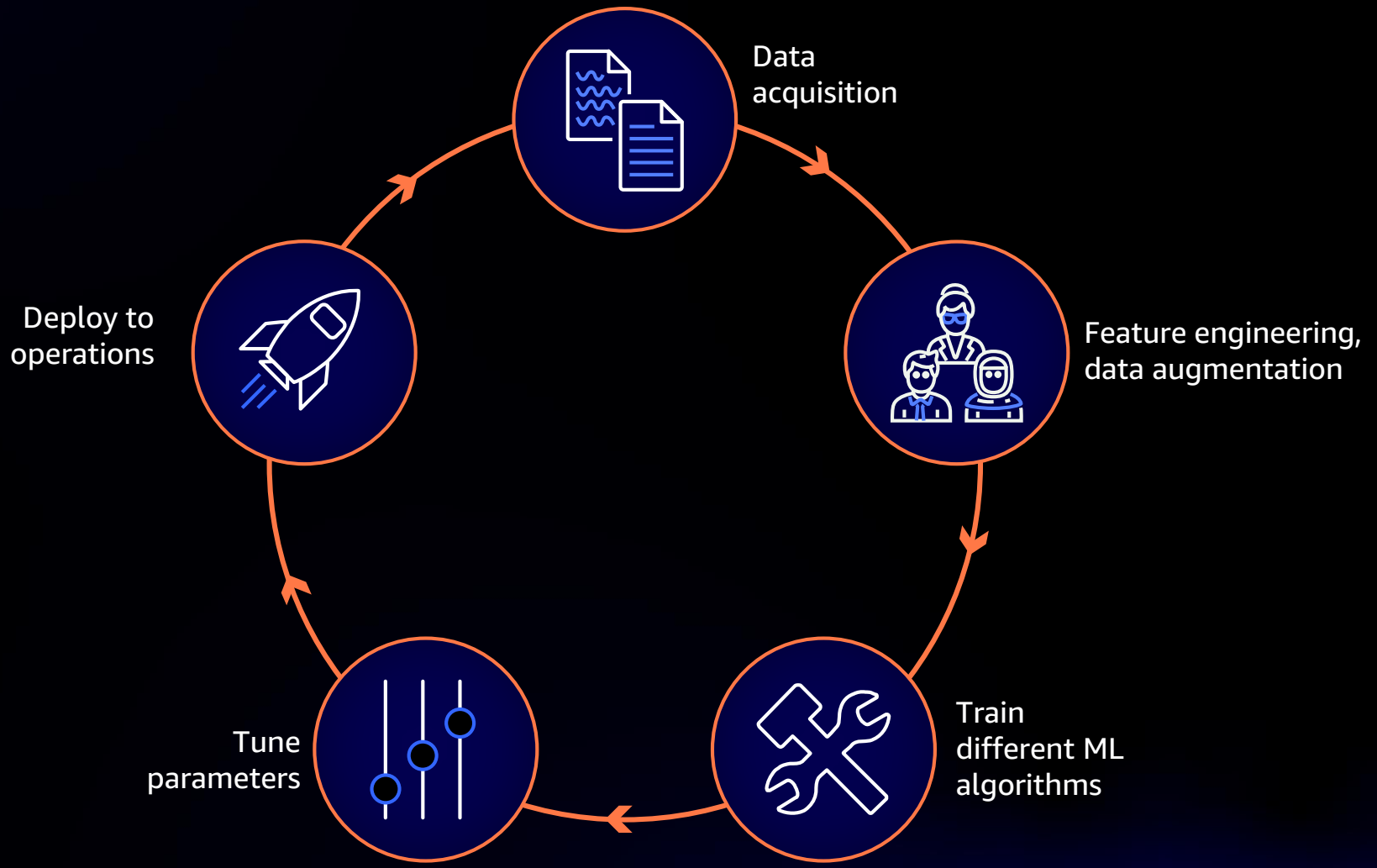
Autopilot demo

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Automatic machine learning (AutoML)

AutoML **removes** the tedious, iterative, and time-consuming work across the ML workflow, from data acquisition to model deployment

You can spend **less time** on low-level details and more time on using ML to improve business outcomes



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building models with AutoGluon



AutoGluon is an open-source AutoML toolkit built by AWS

You can train and deploy high-accuracy machine learning and deep learning models on text, image, and tabular data

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building models with Autopilot

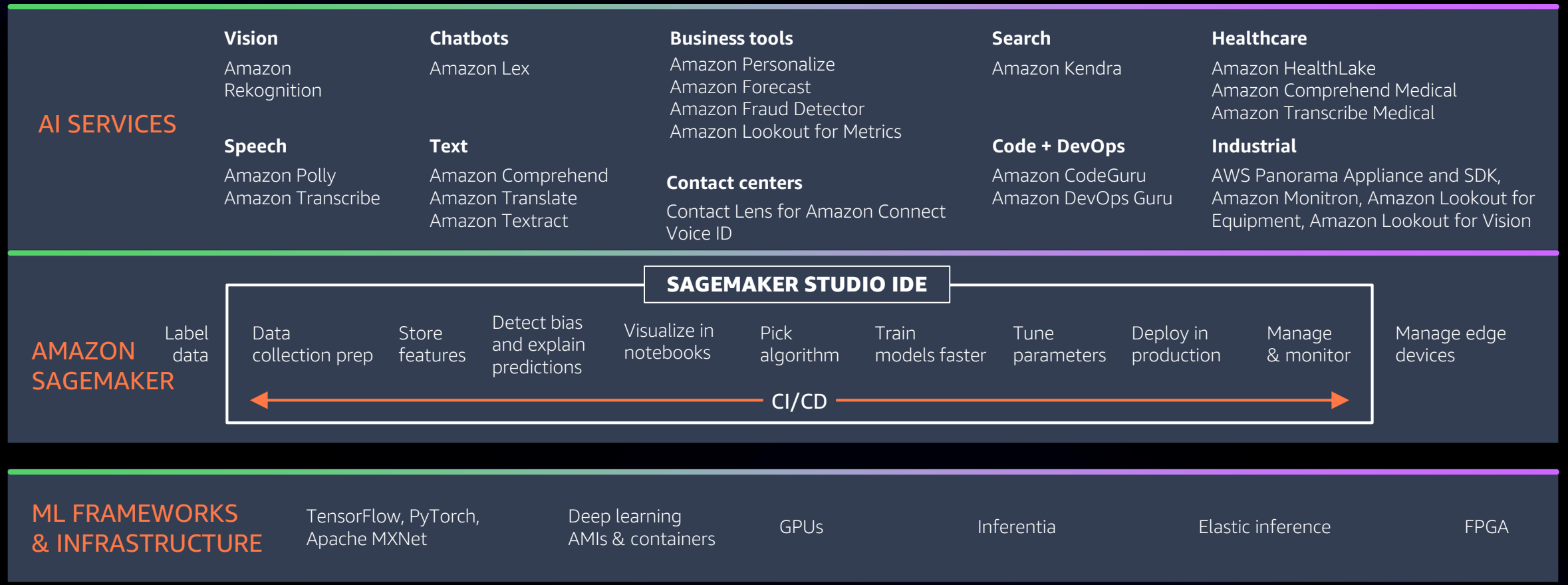


Autopilot is natively integrated with SageMaker and is the automated machine learning capability for tabular data that gives you complete visibility into your ML models

aws

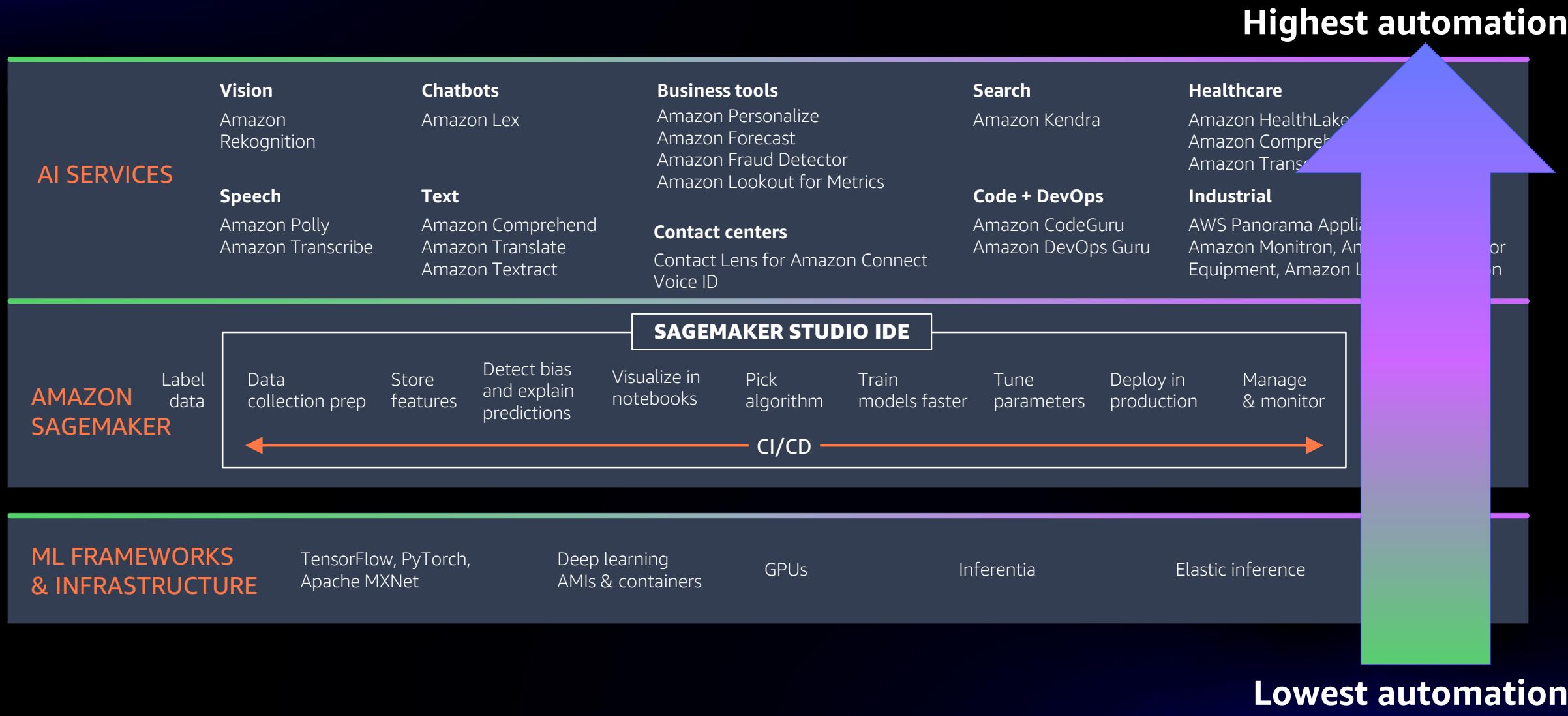
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AutoML is embedded in the AWS ML stack



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AutoML is embedded in the AWS ML stack



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI services: Easily add intelligence to applications

NO MACHINE LEARNING SKILLS REQUIRED



Vision



Chatbots



Business tools



Search



Healthcare



Speech



Text



Contact centers



Code + DevOps



Industrial

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker key benefits

The most complete ML service



Accelerate ML development

20+ tools covering the entire ML development lifecycle



Boost data scientist productivity

The world's first integrated development environment (IDE)



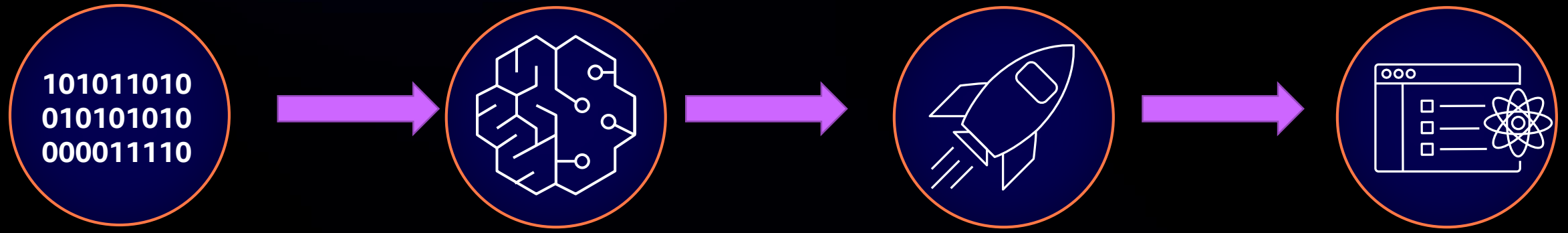
Reduce cost

Eliminate costs of writing custom integration code with integrated functionality optimized for ML

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How Autopilot works



1. Bring data
2. Run Autopilot
3. Deploy model
4. Make predictions

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tabular data as a common data type

Example tabular data

ID_codesort	var_0	var_1	var_2	var_3	var_4	target
test_0	11.0656	7.7798	12.9536	9.4292	11.4327	0
test_1	8.5304	1.2543	11.3047	5.1858	9.1974	0
test_2	5.4827	–10.3581	10.1407	7.0479	10.2628	0
test_3	8.5374	–1.3222	12.022	6.5749	8.8458	0
test_4	11.7058	–0.1327	14.1295	7.7506	9.1035	1
test_5	5.9862	–2.2913	8.6058	7.0685	14.2465	1
test_6	8.4624	–6.1065	7.3603	8.2627	12.0104	0

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Common tabular data use cases

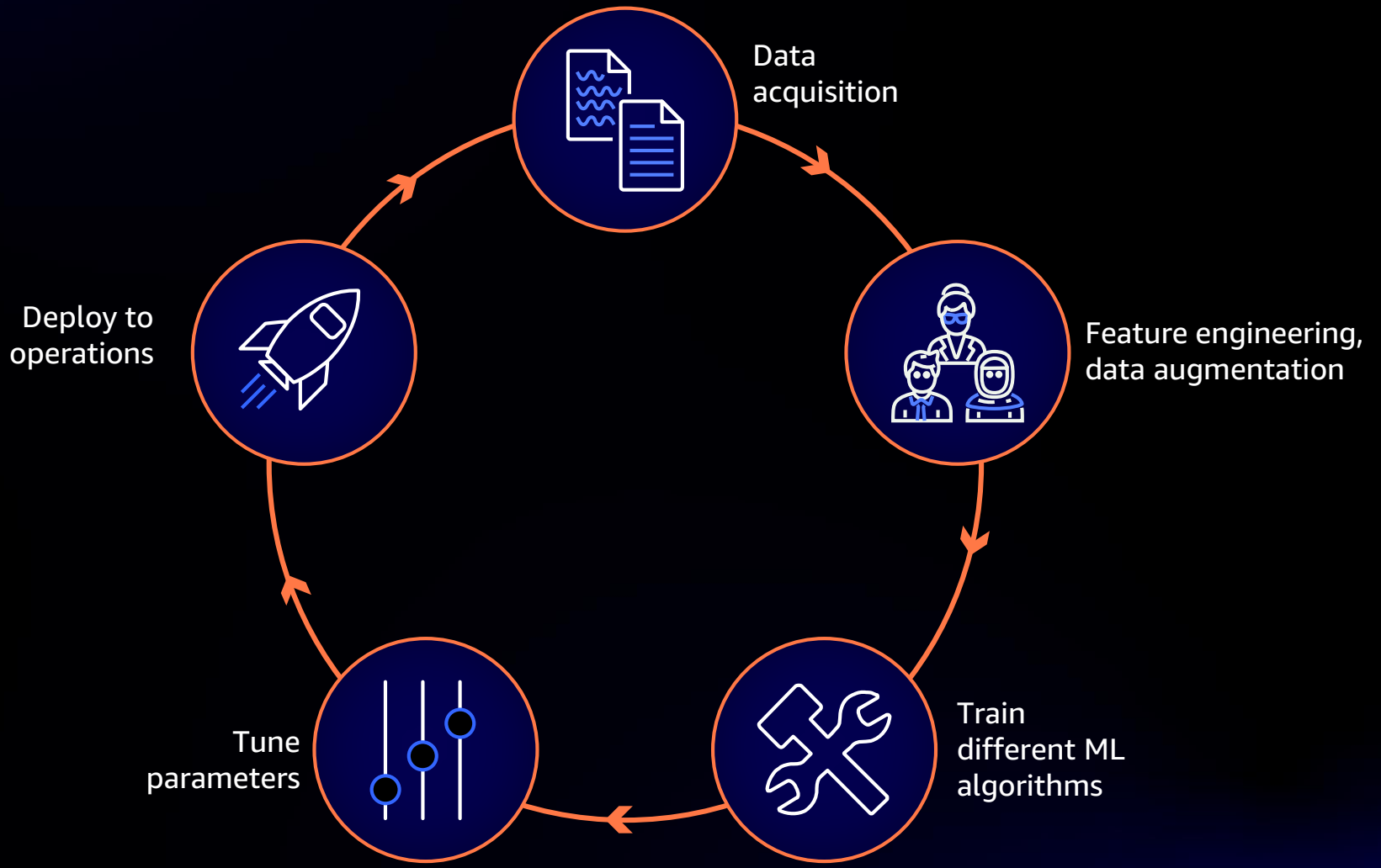
Example tabular data

ID_codesort	var_0	var_1	var_2	var_3	var_4	target
test_0	11.0656	7.7798	12.9536	9.4292	11.4327	0
test_1	8.5304	1.2543	11.3047	5.1858	9.1974	0
test_2	5.4827	–10.3581	10.1407	7.0479	10.2628	0
test_3	8.5374	–1.3222	12.022	6.5749	8.8458	0
test_4	11.7058	–0.1327	14.1295	7.7506	9.1035	1
test_5	5.9862	–2.2913	8.6058	7.0685	14.2465	1
test_6	8.4624	–6.1065	7.3603	8.2627	12.0104	0

- Direct marketing
- Customer churn prediction
- Customer lifetime value
- Credit risk prediction
- Sale prediction
- Price predictions
- Price elasticity for a customer
- Healthcare and life sciences
- Risk assessment
- Loan risk classification

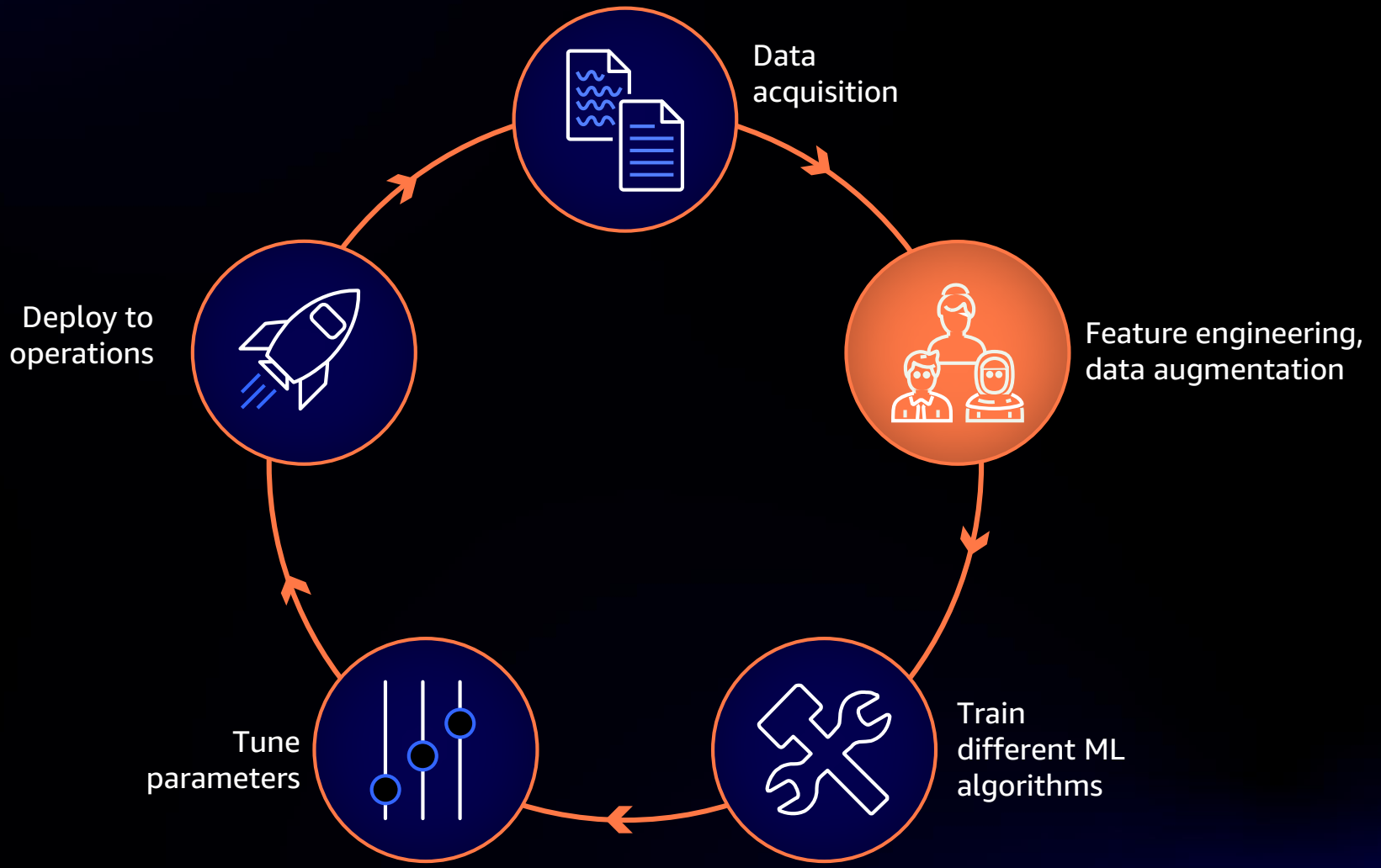
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Autopilot brings automation



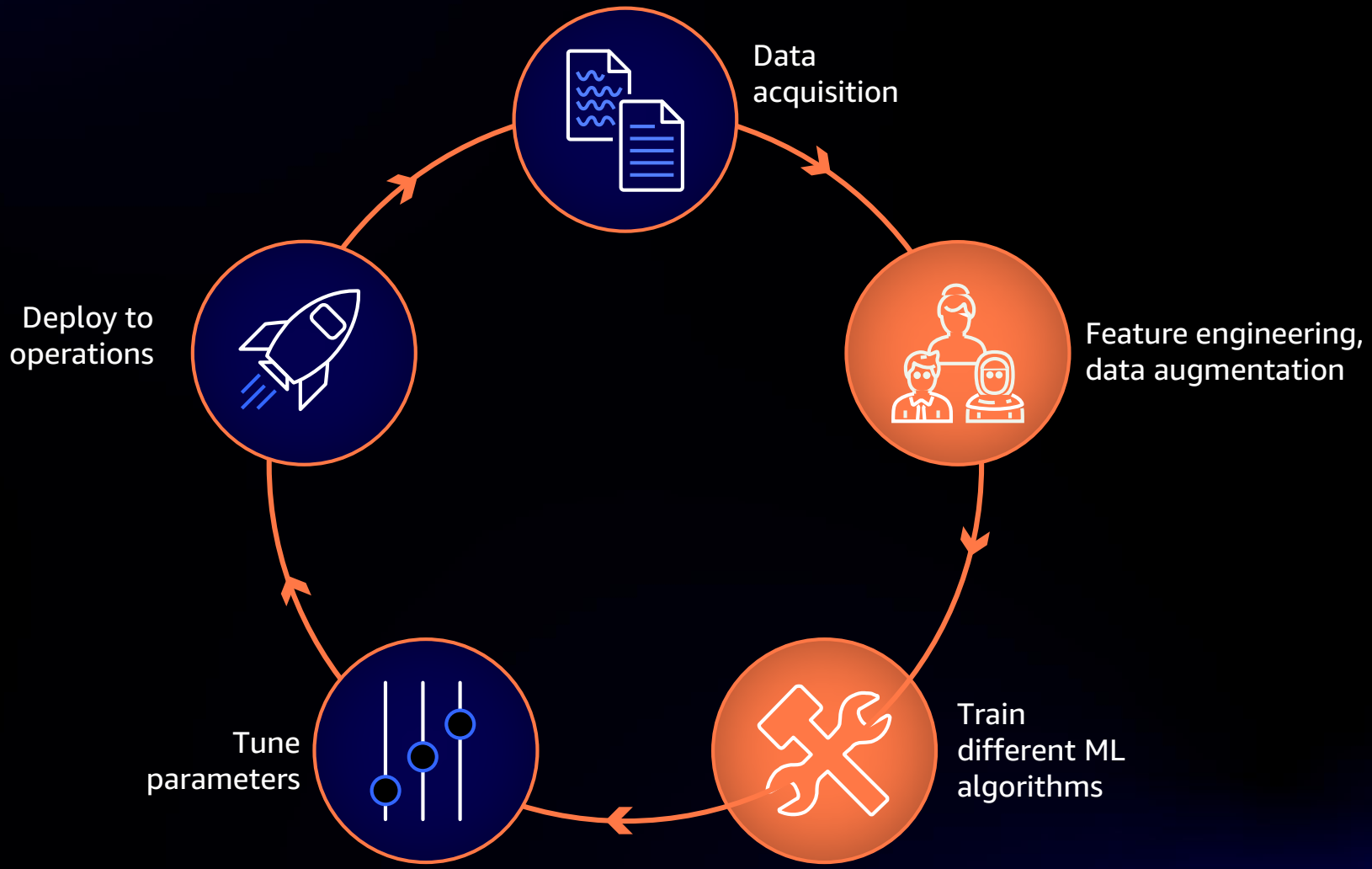
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Autopilot brings automation



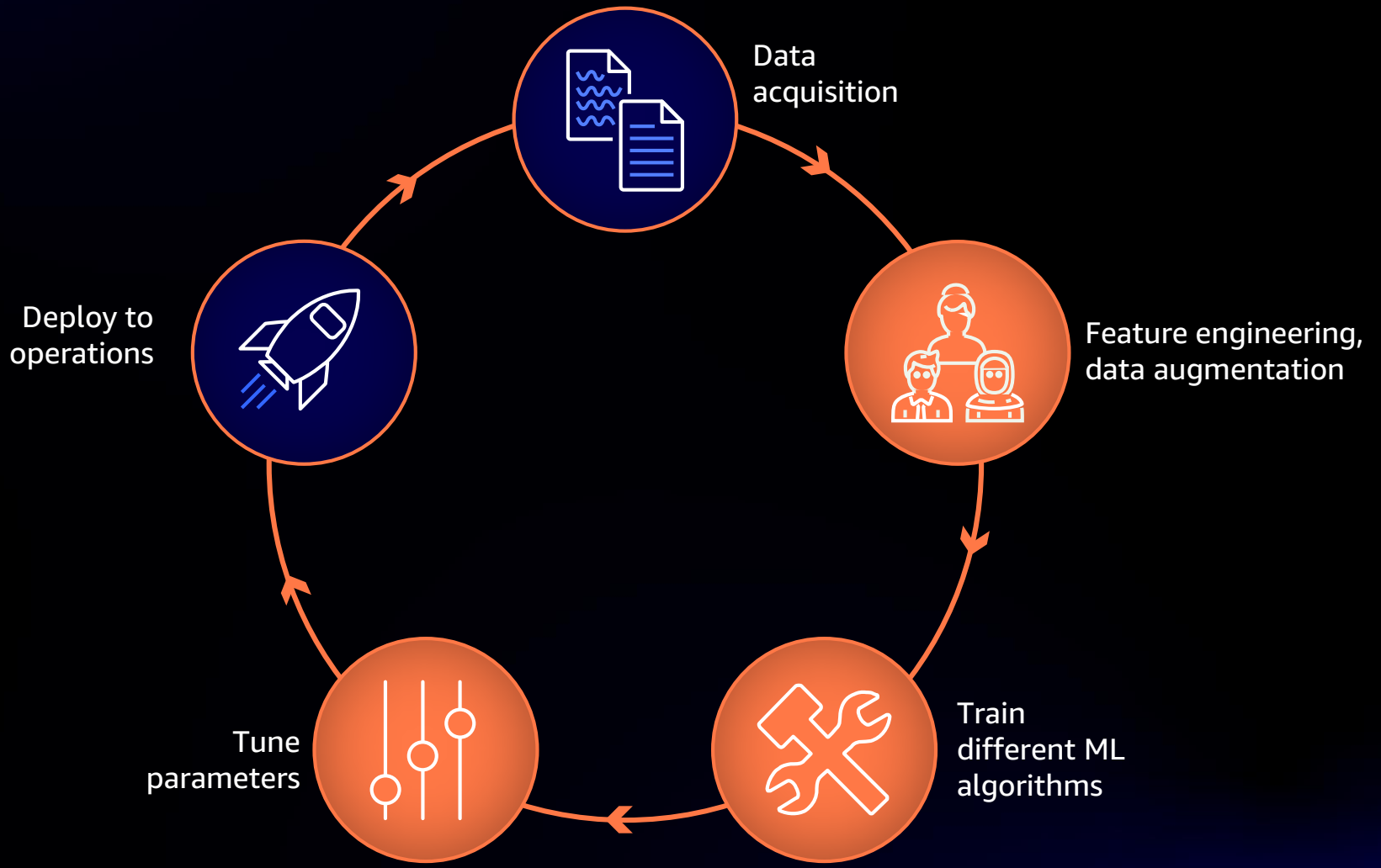
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Autopilot brings automation



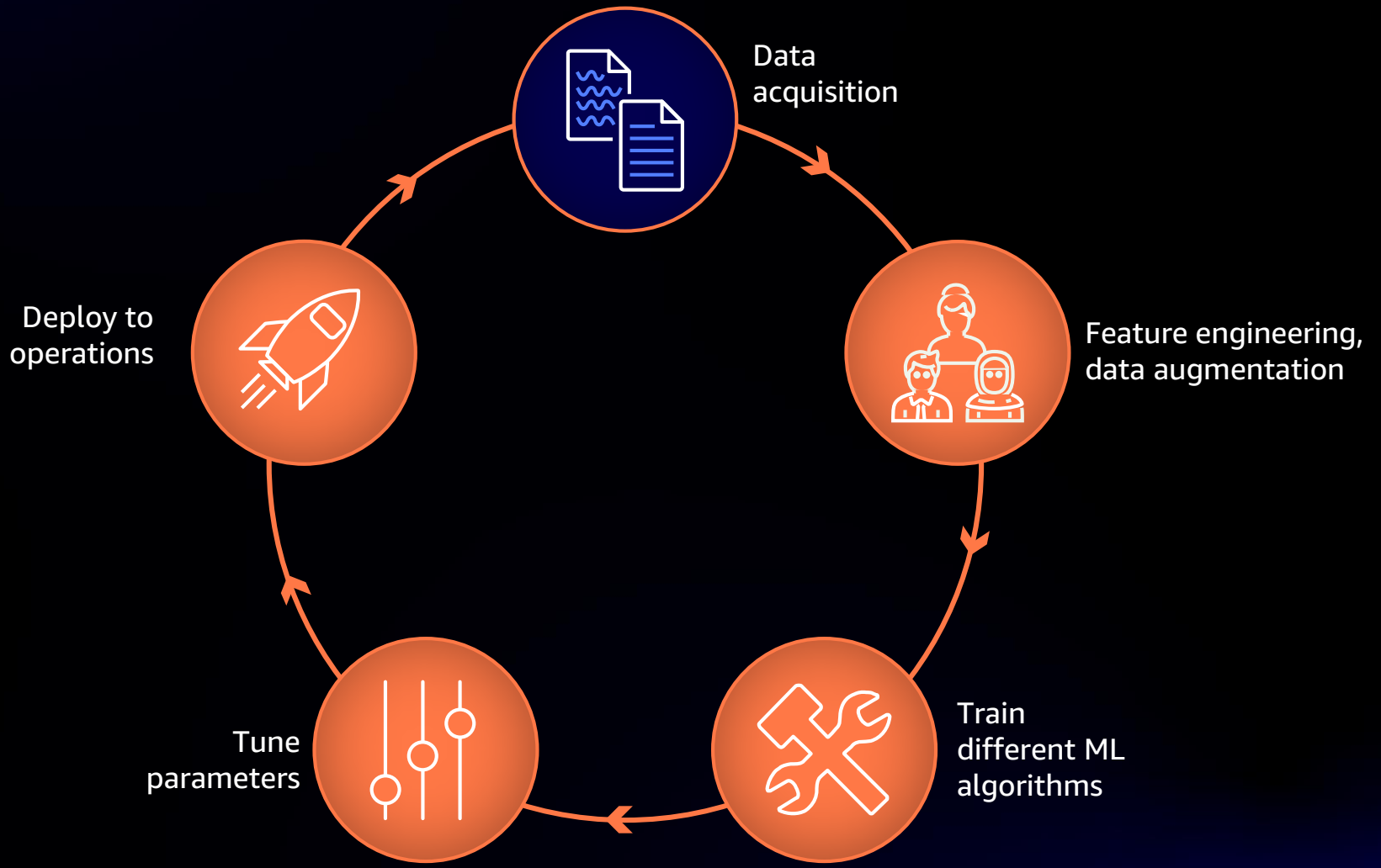
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Autopilot brings automation



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Autopilot brings automation



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

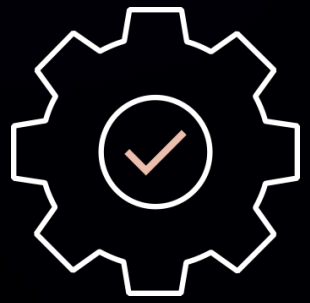
Autopilot core features

AUTOMATIC MODEL CREATION FOR TABULAR DATA WITH FULL VISIBILITY AND CONTROL



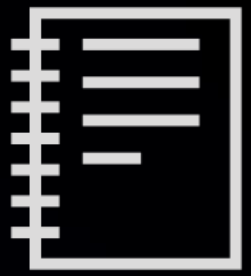
Quick to start

Provide your data in a tabular form and specify target prediction



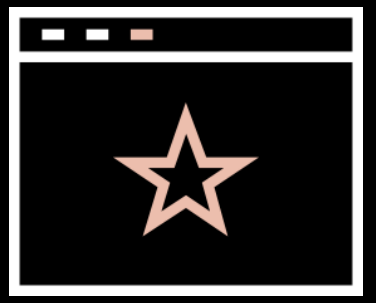
Automatic model creation

Get ML models with feature engineering and model tuning automatically done



Visibility and control

Get notebooks for your models with source code



Recommendations and optimization

Get a leaderboard and continue to improve your model

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Autopilot for data exploration

Dataset exploration notebook:

- Dataset statistics: row-wise and column-wise
- Suggested remedies for common dataset problems

Dataset Sample

The following table is a random sample of **10** rows from the training dataset. For ease of presentation, we are only showing **20 of the 21** columns of the dataset.

Suggested Action Items

- Verify the input headers correctly align with the columns of the dataset sample. If they are incorrect, update the header names of your input dataset in Amazon Simple Storage Service (Amazon S3).

	State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	...	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls
0	CO	76	408	412-4185	no	yes	26	214.6	110	36.48	...	87	17.44	134.6	140	6.06	8.1	2
1	NY	104	415	391-1793	no	yes	26	189.1	112	32.15	...	97	15.15	199.3	104	8.97	11.1	4
2	KY	122	408	392-1616	no	yes	27	253.7	84	43.13	...	109	19.48	190.5	123	8.57	9.2	5
3	NH	67	415	355-1113	no	yes	40	104.9	65	17.83	...	93	18.39	217.4	128	9.78	9.6	9
4	WI	153	510	349-3112	no	no	0	159.5	103	27.12	...	90	23.42	176.7	126	7.95	10.1	2
5	NH	146	510	345-2319	no	no	0	115.6	77	19.65	...	100	18.16	218.4	72	9.83	10.7	6
6	WV	63	510	328-9797	no	no	0	261.8	69	44.51	...	135	20.83	202.1	94	9.09	14.7	4
7	NH	90	408	393-7322	no	no	0	140.2	97	23.83	...	102	18.18	120.0	126	5.4	7.1	2

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Autopilot for model candidates

Fully runnable model candidate notebook:

- Data transformers
- Featurization techniques applied
- Override points:
  - Algorithms considered
  - Evaluation metric
  - Hyperparameter ranges
  - Model search strategy
  - Instances used

The SageMaker Autopilot Job has analyzed the dataset and has generated **10** machine learning pipeline(s) that use **2** algorithm(s). Each pipeline contains a set of feature transformers and an algorithm.

Available Knobs

1. The resource configuration: instance type & count
2. Select candidate pipeline definitions by cells
3. The linked data transformation script can be reviewed and updated. Please refer to the README.md for detailed customization instructions.

dpp0-xgboost: This data transformation strategy first transforms 'numeric' features using RobustImputer (converts missing values to nan), 'categorical' features using ThresholdOneHotEncoder, 'text' features using MultiColumnTfidfVectorizer. It merges all the generated features and applies RobustStandardScaler. The transformed data will be used to tune a *xgboost* model. Here is the definition:

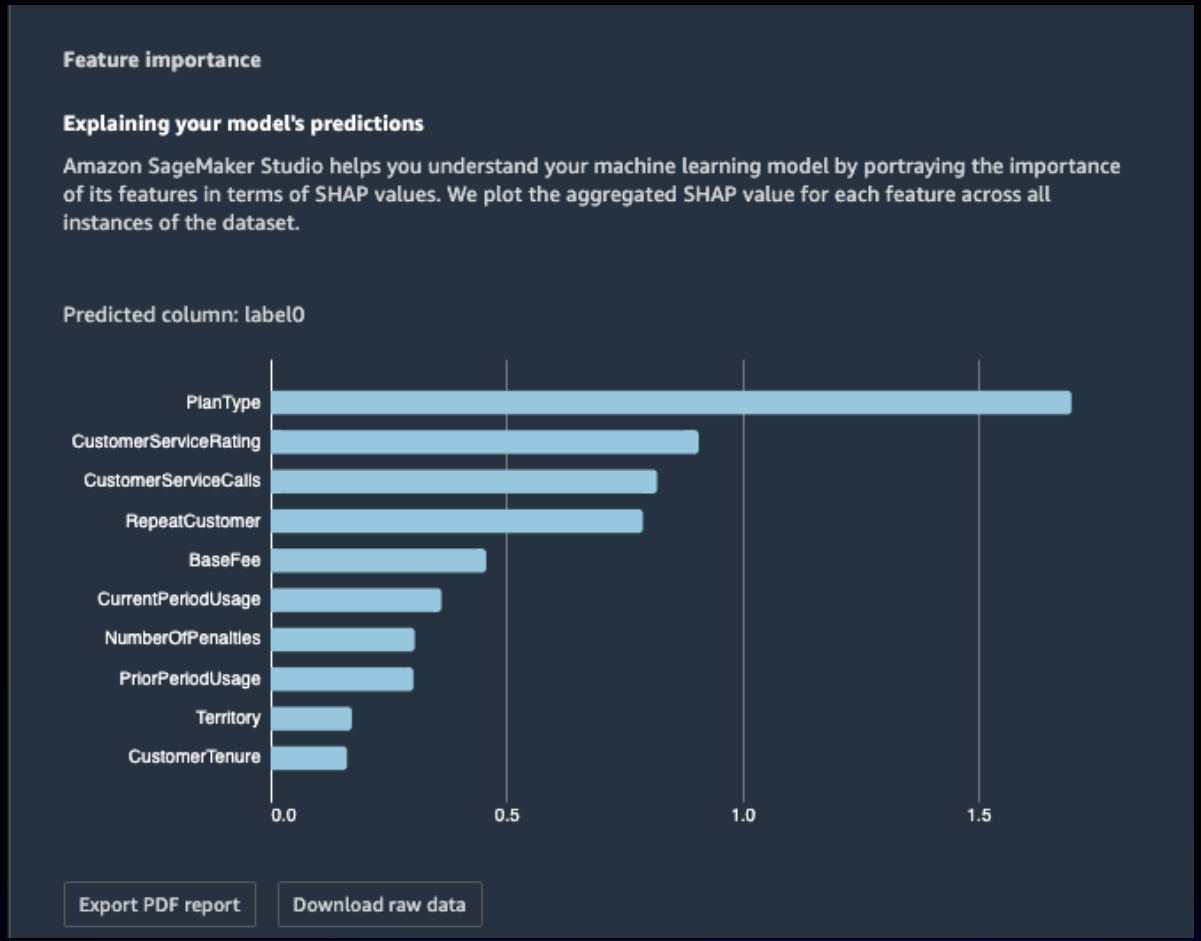
[ ]:
```
automl_interactive_runner.select_candidate({
    "data_transformer": {
        "name": "dpp0",
        "training_resource_config": {
            "instance_type": "ml.m5.4xlarge",
            "instance_count": 1,
            "volume_size_in_gb":  50
        },
        "transform_resource_config": {
            "instance_type": "ml.m5.4xlarge",
            "instance_count": 1,
        },
        "transforms_label": True,
        "transformed_data_format": "application/x-recordio-protobuf",
        "sparse_encoding": True
    }
```

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Autopilot for feature importance

Explainability report generated by Amazon SageMaker Clarify that identifies how each attribute contributes to the predicted result as a percentage

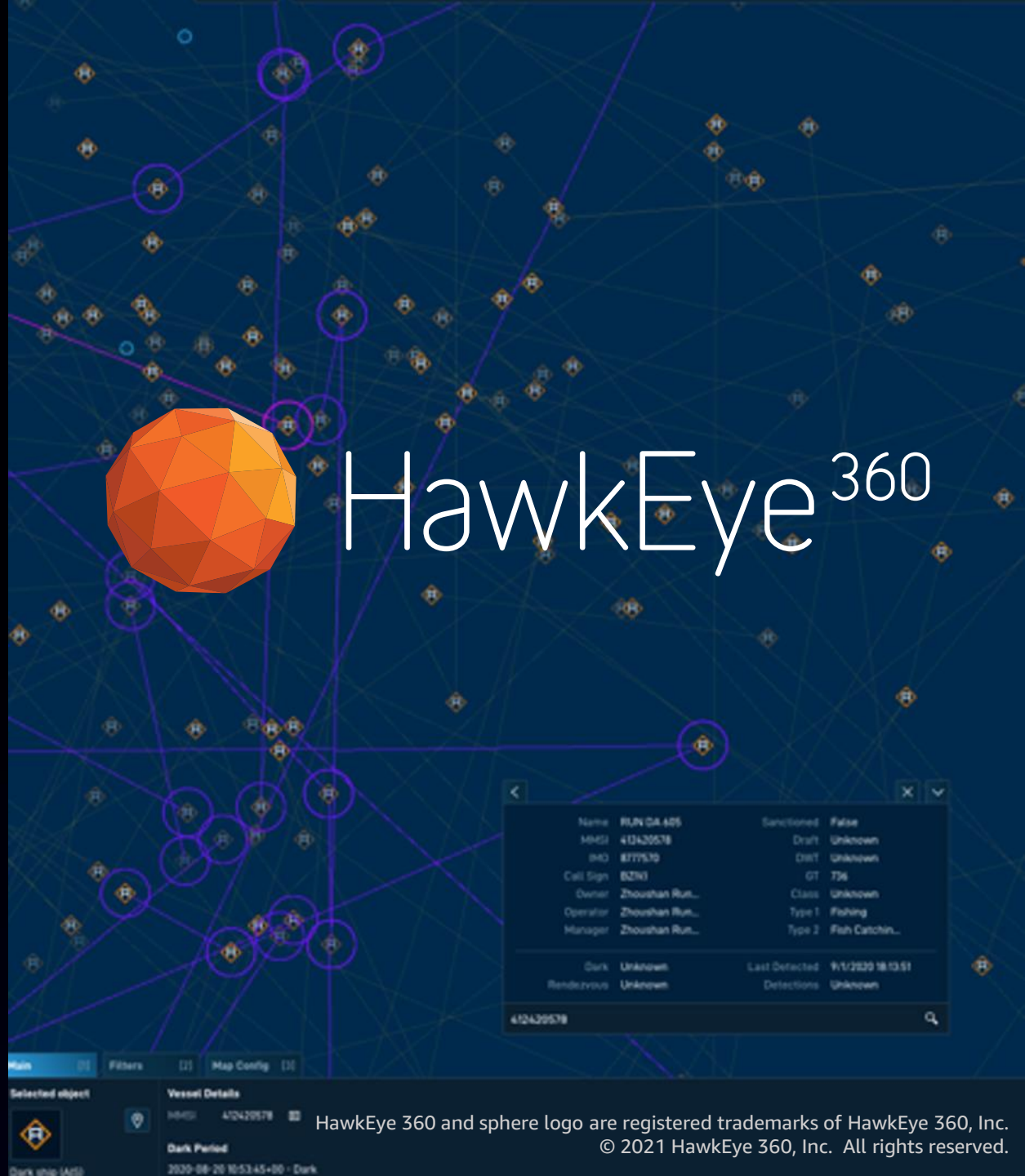
Higher percentages indicate a stronger impact on the model's predictions



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Anomalous maritime activity detection

HawkEye 360 partnered with the Amazon Machine Learning (ML) Solutions Lab to leverage a diverse set of variables, in combination with RF geolocations, to discover hidden patterns and automatically alert customers to anomalous maritime activity in their regions. HawkEye utilized Autopilot, Deep Graph Library, and Amazon Neptune to rapidly generate high-quality models for maritime vessel risk assessment.



aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

HawkEye 360 and sphere logo are registered trademarks of HawkEye 360, Inc.
© 2021 HawkEye 360, Inc. All rights reserved.

snowflake®



Amazon SageMaker

WESTERN UNION

"We have reduced our **model deployment process from weeks to days**, enabling us to rapidly improve our customer experiences with digital banking services that meet our most strategic initiatives."

THOMAS MAZZAFERRO
Chief Data Officer
Western Union

HARVEER SINGH
Chief Data Architect
Western Union

30+ data stores consolidated into a single source of truth

Millions of dollars saved annually by switching to Snowflake

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.