



# Auto ML

Michael Lin  
Senior Solutions Architect  
AWS



## Immersion Day

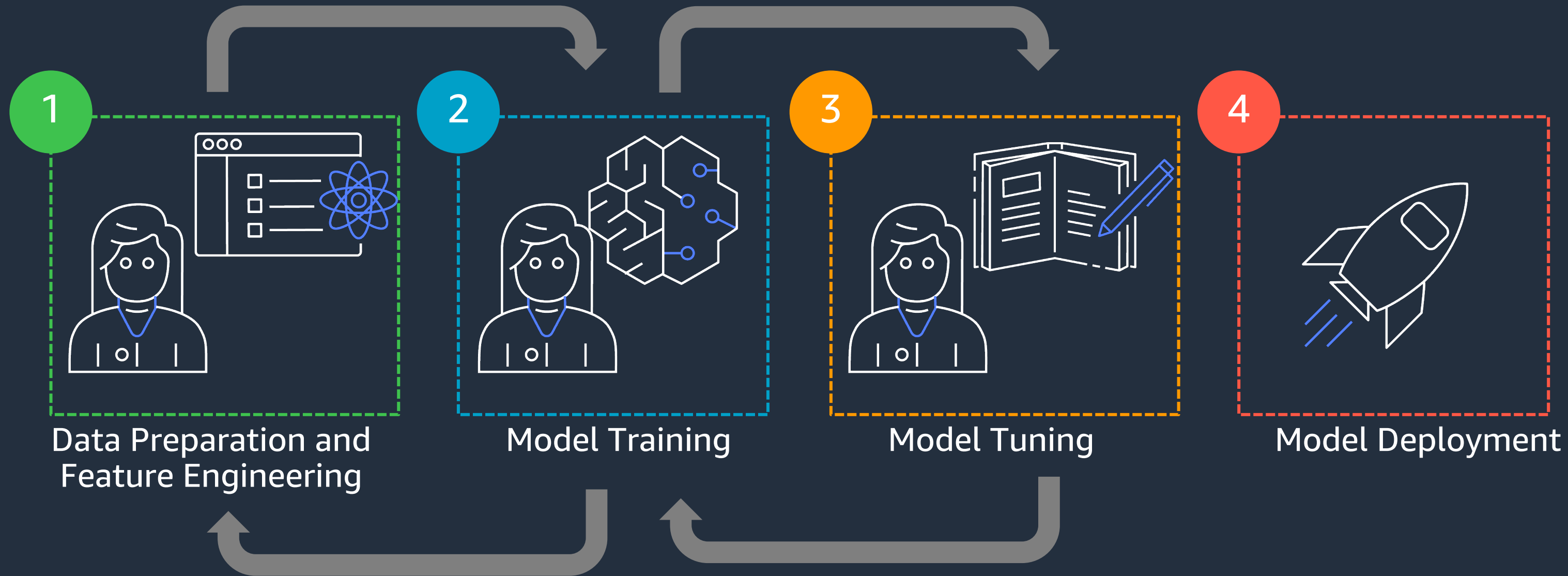
---

# Amazon SageMaker Autopilot



Empowering customers to extract business value from their data quickly, accurately, in a highly scalable way, using the power of machine learning.

# Why building ML models is hard...



# AutoML with Amazon SageMaker Autopilot

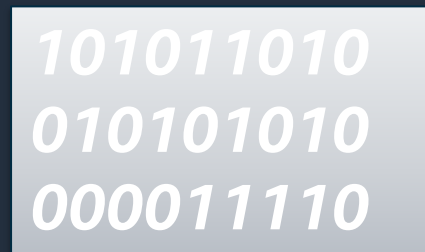
SageMaker Autopilot covers all steps

- *Problem identification*: looking at the data set, what class of problem are we trying to solve?
- *Algorithm selection*: which algorithm is best suited to solve the problem?
- *Data preprocessing*: how should data be prepared for best results?
- *Hyperparameter tuning*: what is the optimal set of training parameters?

Supported **algorithms** at launch:

Linear Learner, XGBoost

# How Amazon SageMaker Autopilot works:



```
101011010
010101010
000011110
```

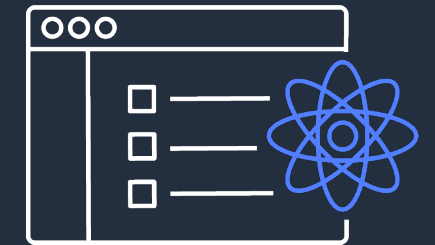
1. Bring data



2. Run Autopilot



3. Deploy model



4. Make predictions

# Amazon SageMaker 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 &  
model tuning  
automatically done



Visibility and  
control

Get notebooks for your  
models with source  
code



Recommendations  
and optimization

Get a leaderboard &  
continue to improve  
your model

# Build Your Models Securely

Complete Integration with Amazon SageMaker Security and Compliance Features



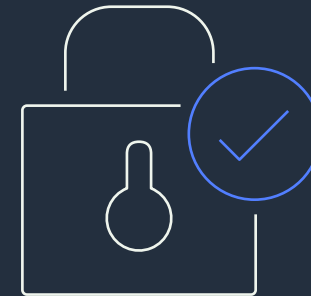
## Access control

Use IAM Policies to control access to data, models, and endpoints



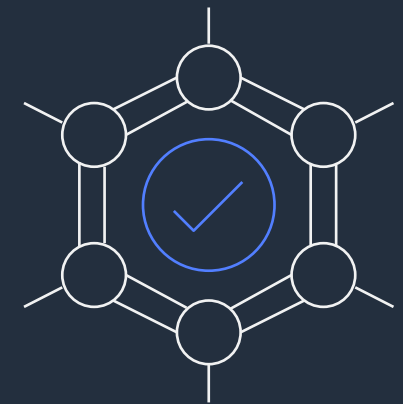
## End-to-end encryption

Support for PrivateLink VPC endpoints, encryption at rest and in-transit



## Secure auditing

CloudWatch and CloudTrail integration



## Compliant with multiple standards

SOC, PCI, FedRAMP, HIPAA, and more



# Autopilot for data exploration

## Dataset Exploration Notebook:

- Dataset statistics: row-wise and column-wise
- Suggested remedies for common data set 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



# Autopilot for model candidates

## Fully runnable Model Candidate Notebook:

- Data transformers
- Featurization techniques applied
- Override points:
  - Algorithms considered
  - Evaluation metric
  - Hyper-parameter 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
  },
})
```

# Thank you

Michael Lin

[linmicht@amazon.com](mailto:linmicht@amazon.com)

