.NET Conf China
2022

如何快速成为机器学习专家
Azure AutoML 介绍

张海
神州数码-云架构师

Microsoft® Most Valuable Professional

神州数码集团

About Digital China Group

收入
1000 亿
集团及相关企业收入

厂商合作伙伴
300+

专业技术人员
1200+

员工
~5000

渠道合作伙伴
30000+

- 神州数码集团股份有限公司（股票代码：000034.SZ，简称：神州数码），其名字源于Digital China，数字中国。

- 从2000年成立伊始，努力践行“数字中国”之理想，神州数码致力于用数字技术推动中华民族的伟大复兴，是中国领先的云管理服务及大数据应用解决方案提供商。二十年耕耘，神州数码始终坚持“理念领先、技术领先、实践领先”，以自主创新核心技术赋能产业数字化转型，在云管理服务、大数据应用、鲲鹏产业生态建设等领域持续深耕并领跑。

- 以“云+大数据”为核心，神州数码已构筑起基于云原生体系的全栈云服务能力，提供面向不同架构的云管理服务和一站式云网融合服务，在数据脱敏、智能数据管理平台等技术方面具有行业领先水平。目前，神州数码已在金融、医疗、制造、文旅、零售、汽车等多个行业为企业客户提供了全生命周期数字化转型服务，成为多家世界500强客户中国区域的首选服务商。

目录

机器学习简介

- 痛点

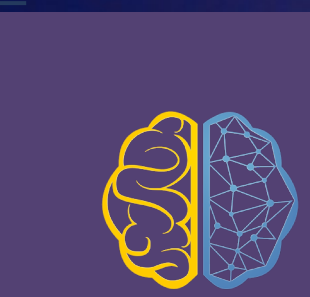
使用 Azure AutoML

- Demo

结合 ML.NET

- Demo

机器学习简介



What?

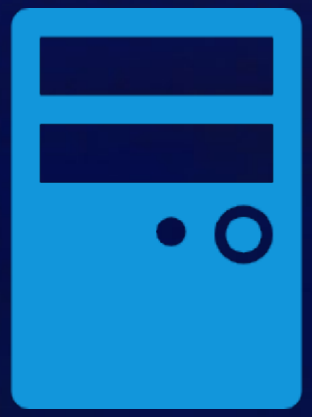
Why?

数据
输出
(答案)
机器学习
程序
(规律)
您最近查看的商品和相关推荐
热销商品
kindleunlimited
流浪地球
刘慈欣获奖作品
kindleunlimited
月亮
与六便士
kindleunlimited
kindleunlimited
大美人

How!

痛点

时间有限



资源不足

专业知识

特征工程

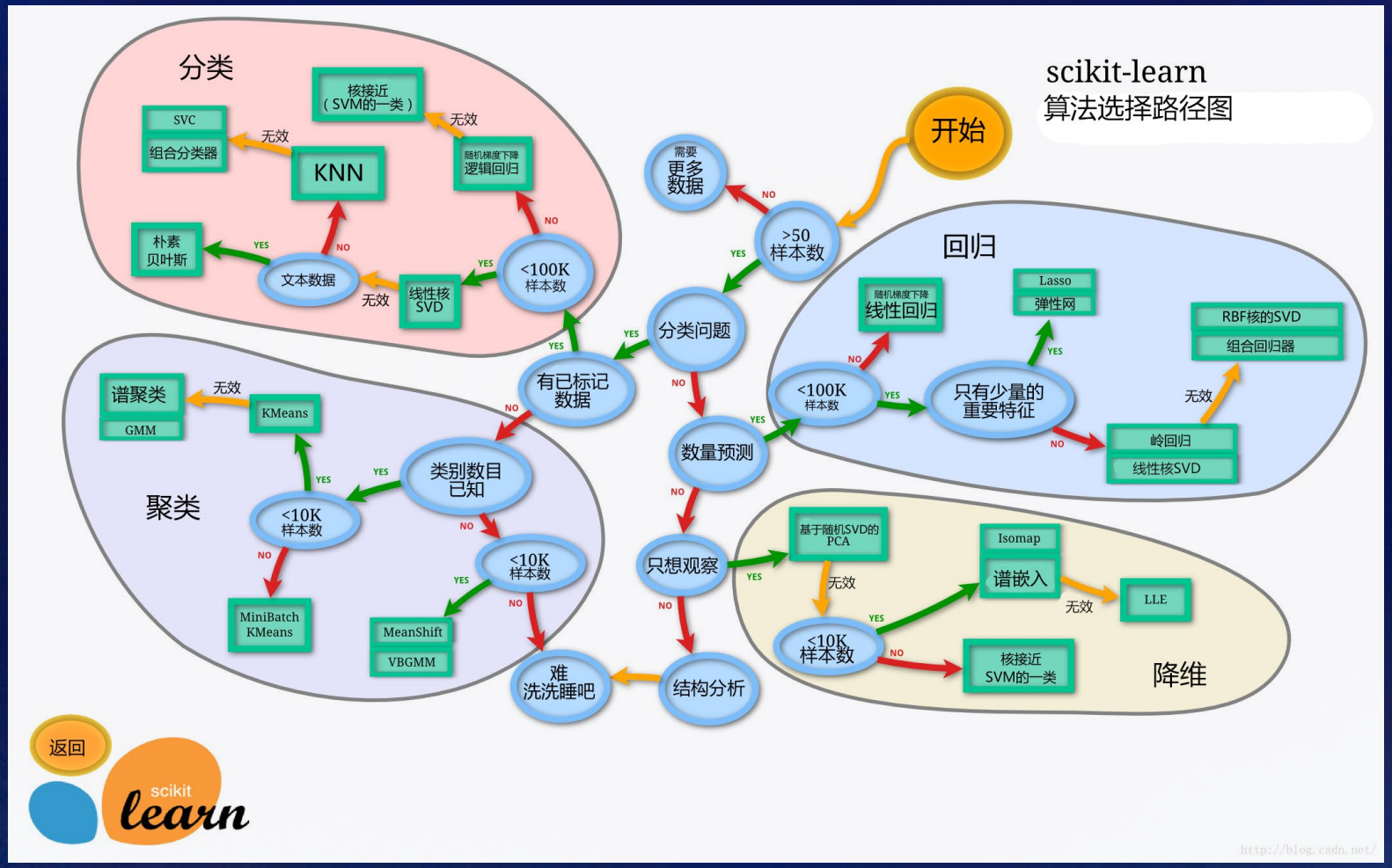
原始数据

```
0 : {
 house_info : {
 num_rooms: 6
 num_bedrooms: 3
 street_name: "Shorebird Way"
 num_basement_rooms: -1
 ...
 }
}
```

特征向量

```
[
 6.0,
 1.0,
 0.0,
 0.0,
 0.0,
 9.321,
 -2.20,
 1.01,
 0.0,
 ...,
]
```

选择算法

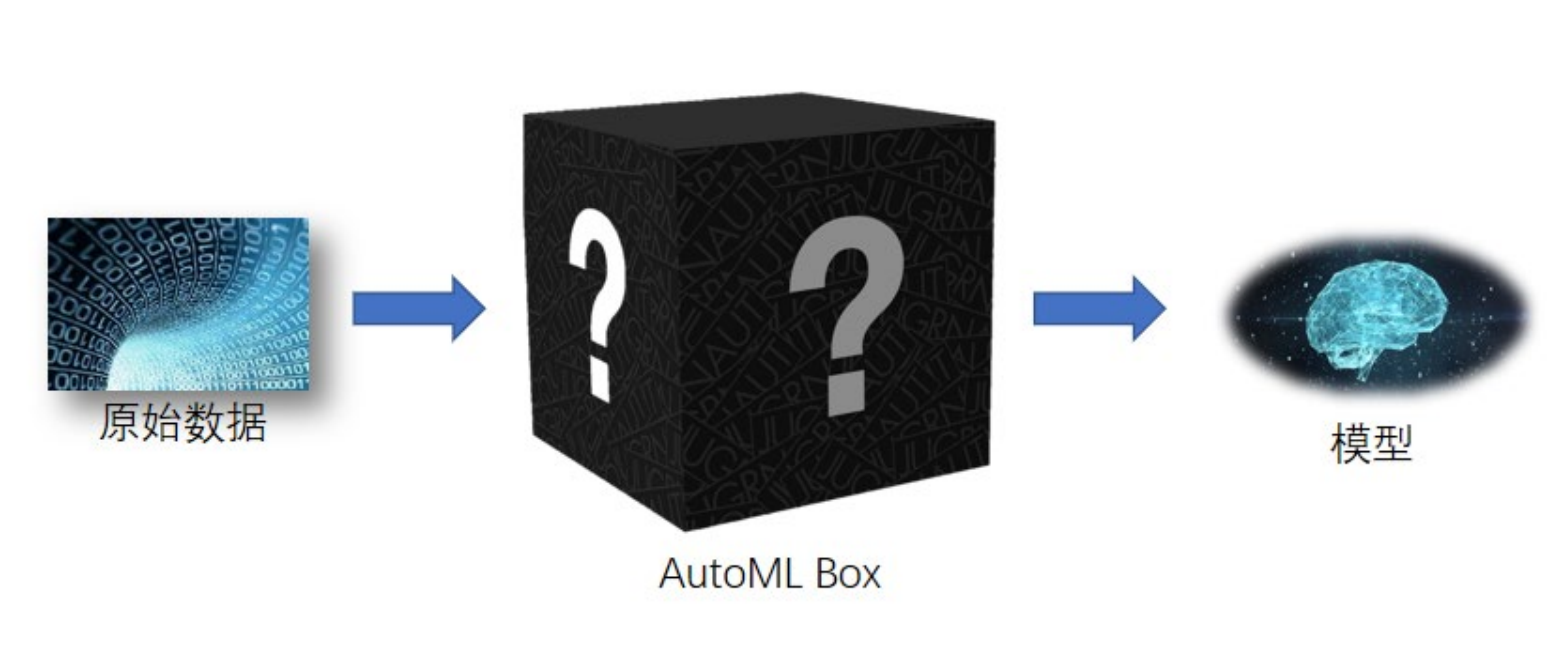

scikit-learn
算法选择路径图
开始
分类
核接近
（SVM的一类）
SVC
组合分类器
KNN
随机梯度下降
逻辑回归
朴素
贝叶斯
文本数据
线性核
SVD
<100K
样本数
无效
YES
NO
需要
更多
数据
>50
样本数
分类问题
有已标记
数据
回归
随机梯度下降
线性回归
Lasso
弹性网
RBF核的SVD
组合回归器
<100K
样本数
只有少量的
重要特征
岭回归
线性核SVD
数量预测
聚类
谱聚类
GMM
KMeans
类别数目
已知
<10K
样本数
MiniBatch
KMeans
MeanShift
VBGMM
只想观察
基于随机SVD的
PCA
Isomap
谱嵌入
LLE
核接近
SVM的一类
降维
难
洗洗睡吧
结构分析
返回
scikit
learn
http://blog.csdn.net/

超参数优化

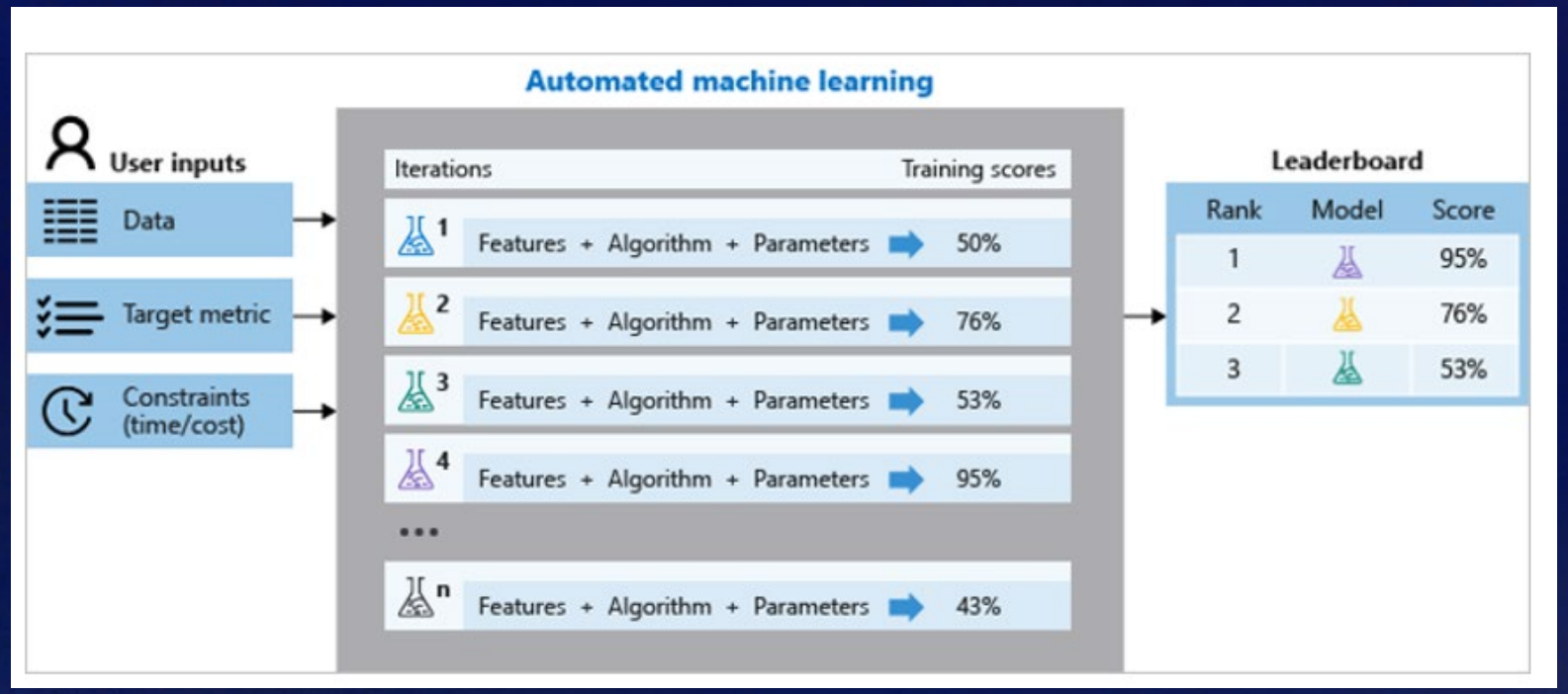
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=2, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```

使用 Azure AutoML

AutoML（Automated Machine Learning）



Azure AutoML



Azure AutoML 适用场景

分类：预测新数据将属于哪些类别。 比如疾病预测

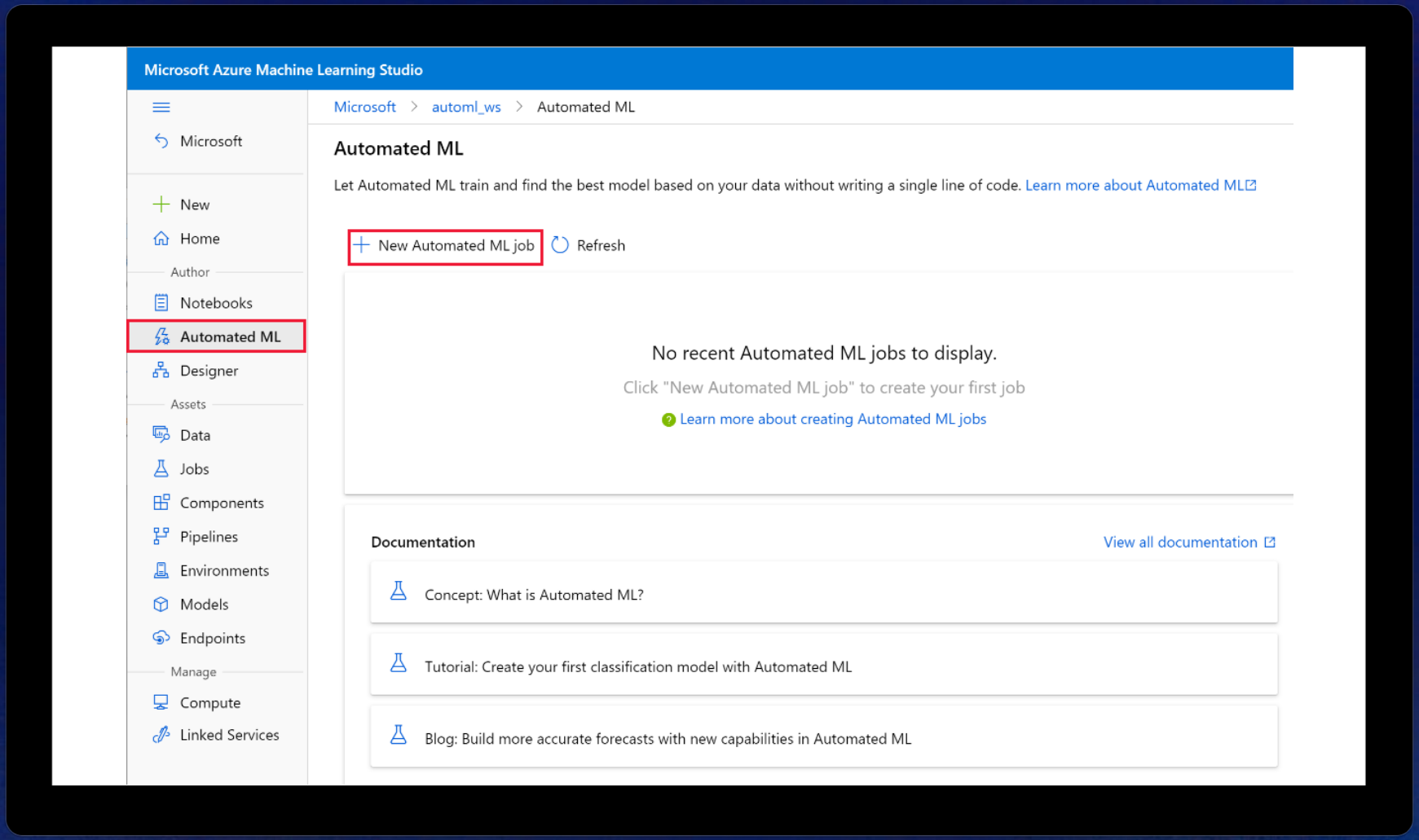
回归：预测数字输出值。比如预测房价

时序预测：“透视”过去的时序值，使其成为回归的附加维度。比如销售预测

计算机视觉（预览阶段）/ NLP自然语言处理（预览阶段）

.NET Conf China

Demo

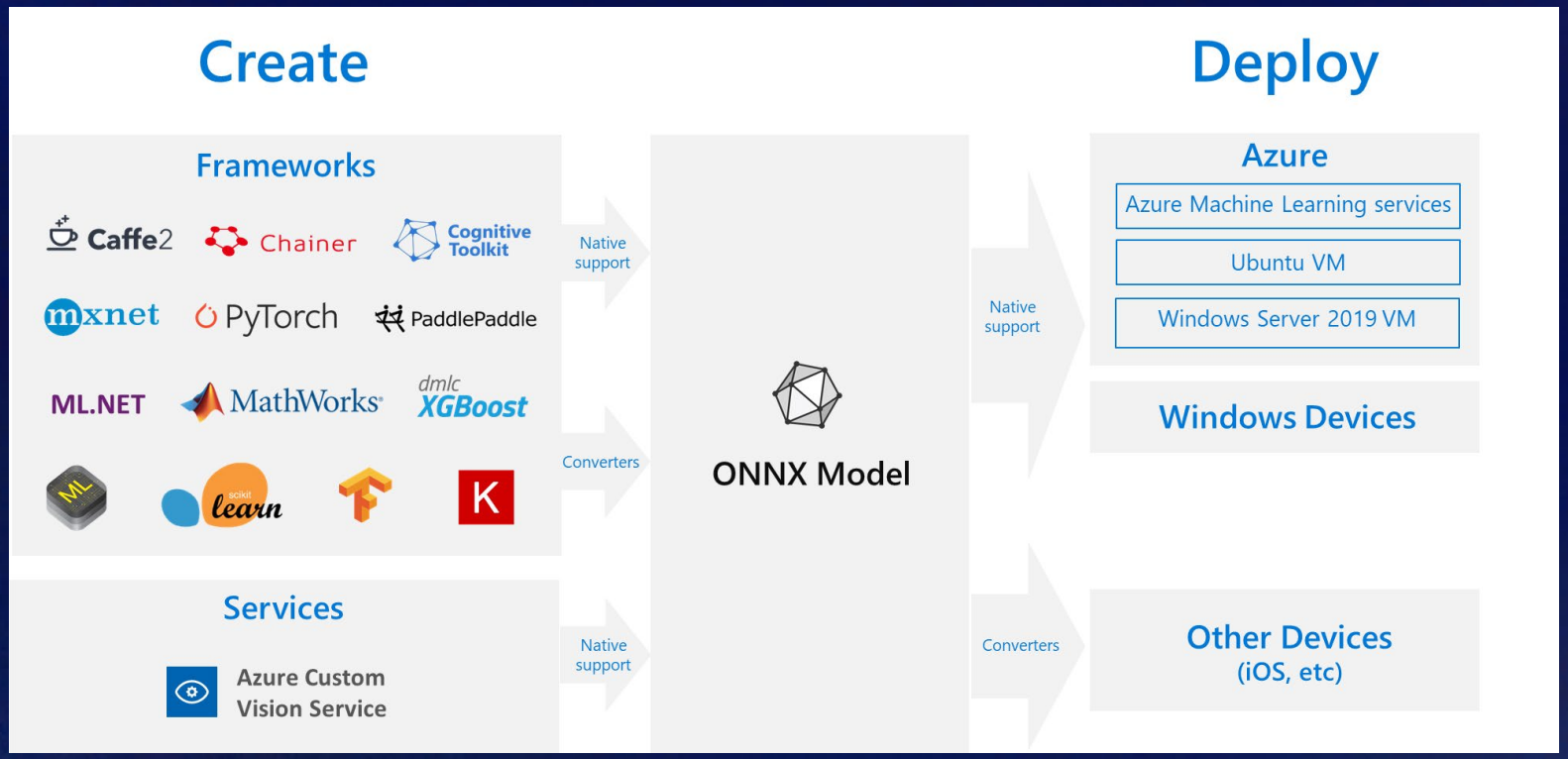

Microsoft Azure Machine Learning Studio
Microsoft
automl_ws
Automated ML
Microsoft
New
Home
Author
Notebooks
Automated ML
Designer
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Linked Services
Automated ML
Let Automated ML train and find the best model based on your data without writing a single line of code. Learn more about Automated ML
New Automated ML job
Refresh
No recent Automated ML jobs to display.
Click "New Automated ML job" to create your first job
Learn more about creating Automated ML jobs
Documentation
View all documentation
Concept: What is Automated ML?
Tutorial: Create your first classification model with Automated ML
Blog: Build more accurate forecasts with new capabilities in Automated ML

结合 ML.NET

ML.NET

https://dot.net/ml

ONNX（Open Neural Network Exchange）



生成ONNX模型

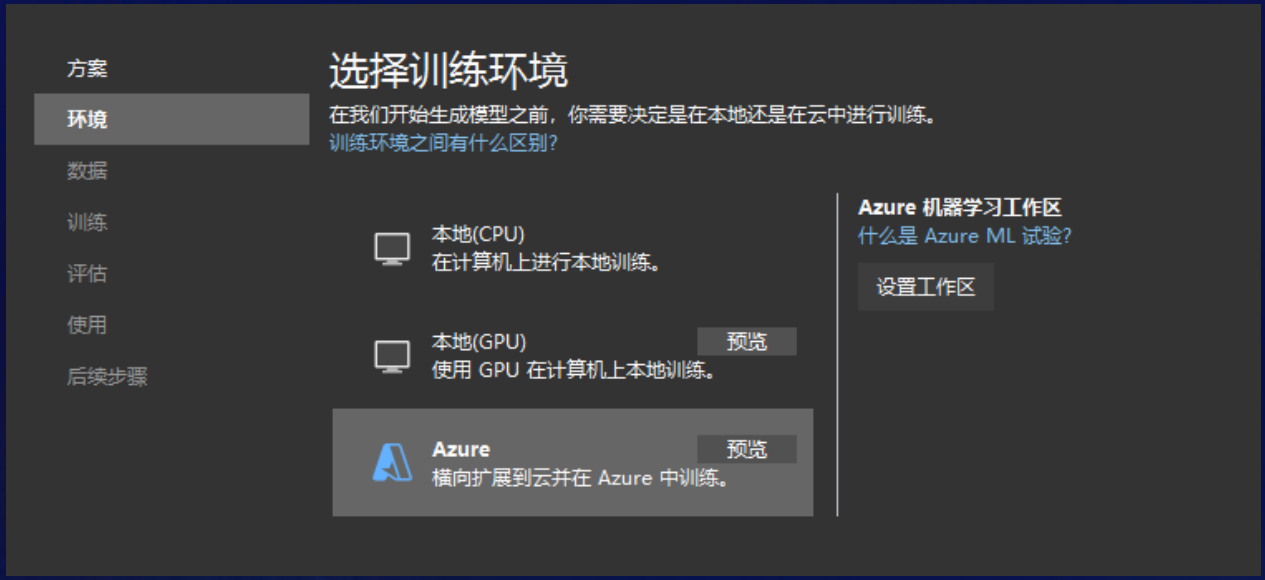
Azure ML Notebook（Python）

```
automl_config = AutoMLConfig(
    task="classification",
    debug_log="automl_errors.log",
    compute_target=compute_target,
    experiment_exit_score=0.9984,
    blocked_models=["KNN", "LinearSVM"],
    enable_onnx_compatible_models=True,
    training_data=train_data,
    label_column_name=label,
    validation_data=validation_dataset,
    **automl_settings,
)
```

在AutoML Configuration中添加对ONNX的支持

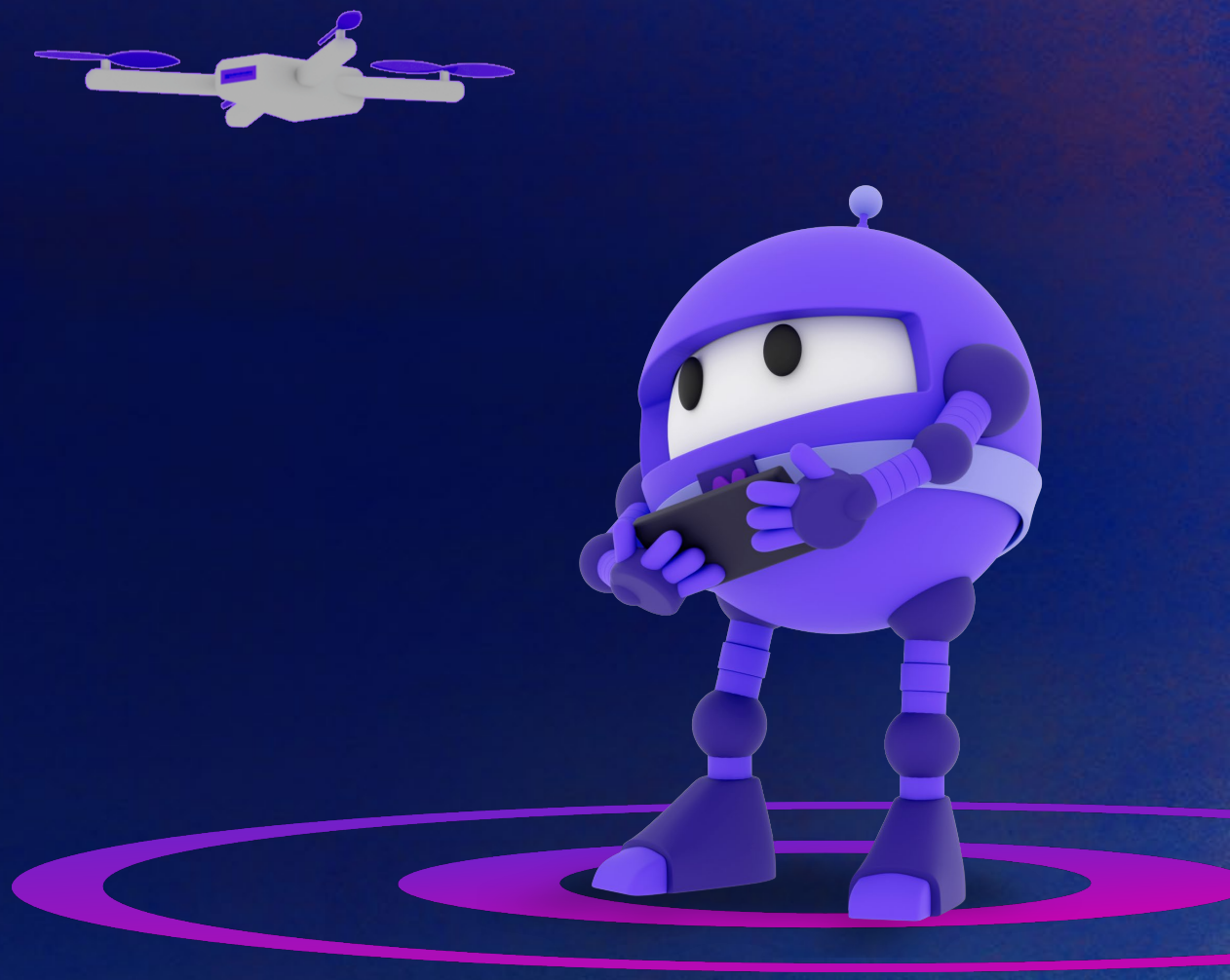
生成ONNX模型

ML.NET Model Builder



选择 Azure □ □ □ □ □ □

Demo

方案
环境
数据
训练
评估
使用
后续步骤
选择一个方案
使用你的数据进行训练
以下方案使用自动化 ML 来训练数据并选择最适合该数据的模型。
详细了解如何在 Model Builder 中使用自己的数据进行训练。
数据分类
将表哥数据分为 2 个以上的类别，例如预测评论中的情绪是积极的还是消极的。
本地
值预测
通过数据预测数值(回归)，例如根据大小、位置等特征预测房价
本地
图像分类
将图像分为 2 个以上的类别，例如预测某张图像是狗还是猫。
Azure
本地
建议
为特定用户生成建
建议的产品。
本地
反馈

Thank you!