数据的定义

阿里云

数据（data）

- 用于表示客观事物的未经加工的原始素材；
  - 示例：12瓶饮料
- 不仅指狭义上的数字，也指具有一定意义的文字、字母、数字符号的组合；
  - 示例：猫狗的图片、书籍、音乐
- 客观事物的属性、数量、位置及其相互关系的抽象表示。
  - 示例：牡丹花是我国特有的木本名贵花卉，其花色艳丽，带有淡淡的清香。

- 在**计算机科学与技术领域**中，数据是指一切能够输入计算机中，且能被计算机程序所处理的符号的总称。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据类别——按字段分类

阿里云

按照字段类型分类是最基本的数据分类方式。

文本类

- 用于描述性字段；
- 非量化值，不可直接用于运算。

8月10日

时间类

- 用于描述事件发生的时间；
- 可直接用于运算。

8/10

数值类

- 用于描述可量化属性/编码操作；
- 可直接用于运算。

44783

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据类别——按数据结构类型分类

阿里云

按照数据结构类型分类是人工智能领域中**较为重要**的数据分类方式。

结构化数据
由统一的结构来逻辑表示和存储的数据。

非结构化数据
无预定义数据模型，不可直接用数据库逻辑来表现的数据。

半结构化数据
具有结构化形式，但并不符合数据模型结构。

姓名	座位号	考试成绩
小A	01	88
小B	02	76
小C	03	91

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据采集的定义

阿里云

随着网络和信息技术的不断提升，人类社会产生的数据量正在**呈指数级增长。**

数据采集

- 又称数据获取；
- 指利用装置从系统外部采集数据并输入到系统内部的技术；
- 对数据进行抽取、转换、加载操作；
- 目标是获得数据。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据采集的4种常用方法

阿里云

根据数据源的物理性质及数据分析的目标，用采取不同的数据采集方式。

网络数据采集

- 主要采集现实网页中的数据；
- 常用API法和网络爬虫法。

端侧数据采集

- 主要采集已转换成电信号的各种物理量；
- 常用摄像头、麦克风等端侧设备。

系统日志采集

- 主要采集用户行为日志、业务变更日志、系统运行日志；
- 常用WebAPI方式、Service Proxy方式、LCClient方式。

数据库采集

- 主要采集数据库中的数据；
- 常用MySQL、Oracle、NoSQL数据库。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据可视化的含义及典型图表

阿里云

- 借助于**图形化手段**对数据加以解释
- 在进行数据可视化中，经常使用的图表主要包括
  - **直方图**
  - **折线图**
  - **散点图**
  - **饼状图**
  - **箱线图**
  - 小提琴图
  - 雷达图
  - 热力图
  - 树状图
  - 漏斗图
  - 地理图...

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

常用的绘图库Matplotlib

阿里云

- 一个非常强大的 Python 画图工具，轻松地将数据图形化；
- Python 的绘图库，用来绘制各种静态，动态，交互式的图表；
- 将很多数据通过图表的形式更直观的呈现出来；
- 可以绘制线图、散点图、等高线图、条形图、柱状图、3D 图形、甚至是图形动画等。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

直方图的含义和绘图方法

阿里云

含义

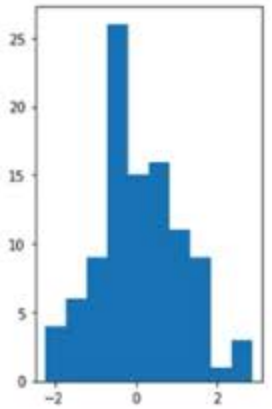
- 一种统计报告图；
- 由一系列高度不等的纵向条纹或线段表示数据分布的情况；
- 一般用横轴表示数据类型，纵轴表示分布情况。

绘图方法

- plt.hist(x, bins=10)
- x 是一维数组
- bins 代表直方图中的箱子数量，默认是 10

```
# 准备数据
a = np.random.randn(100)
s = pd.Series(a)

plt.figure(figsize=(10,5))
plt.subplot(131)
# matplotlib绘图
plt.hist(s)
```



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

折线图的含义和绘图方法

阿里云

含义

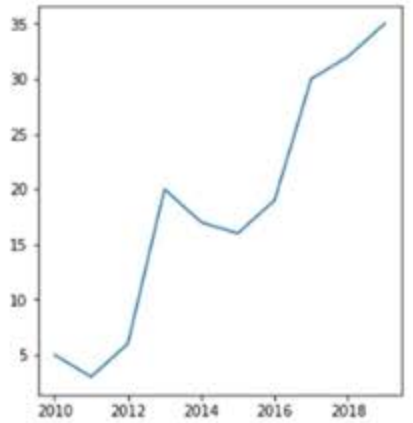
- 显示随时间而变化的连续数据；
- 适用于显示在相等时间间隔下数据的趋势；
- 类别数据沿水平轴均匀分布，所有值数据沿垂直轴均匀分布。

绘图方法

- plot(x, y)
- x 为 x 轴数据
- y 为 y 轴数据
- 数据可以列表或数组

```
# 数据准备
x = [2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019]
y = [5, 3, 6, 20, 17, 16, 19, 30, 32, 35]

plt.figure(figsize=(10,5))
plt.subplot(121)
# matplotlib 画图
plt.plot(x, y)
```



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

散点图的含义和绘图方法

阿里云

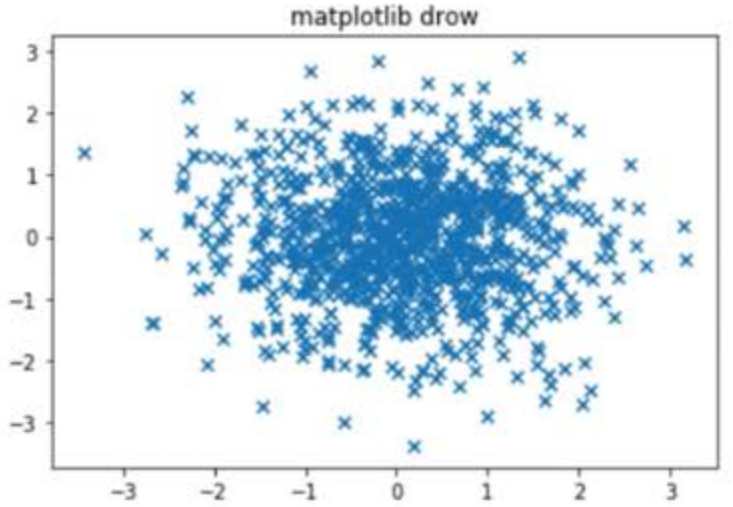
含义

- 数据点在直角坐标系平面上的分布图；
- 表示因变量随自变量而变化的大致趋势；
- 通常用于比较跨类别的聚合数据。

绘图方法

- plt.scatter(x, y)
- x、y表示长度相同的数组

```
plt.scatter(x, y, marker='x')
plt.title('matplotlib drow')
plt.show()
```



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

饼状图的含义和绘图方法

阿里云

含义

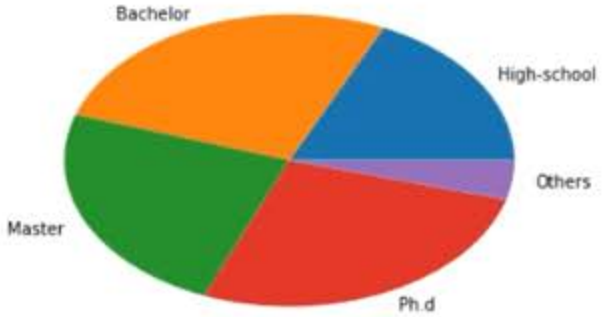
- 显示一个数据系列；
- 每个数据系列具有唯一的颜色或图案；
- 图表中绘制一个或多个数据系列；
- 表示某个数据系列中各项的大小与各项总和的比例。

绘图方法

- plt.pie(x, explode=None, labels=None)
- x：表示每个扇形的面积
- explode：表示各个扇形之间的间隔
- Labels：列表，各个扇形的标签

```
# 数据准备
nums = [25, 37, 33, 37, 6]
labels = ['High-school','Bachelor','Master','Ph.d', 'Others']

# 用Matplotlib画饼图
plt.pie(x = nums, labels=labels)
plt.show()
```



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

箱线图的含义和绘图方法

阿里云

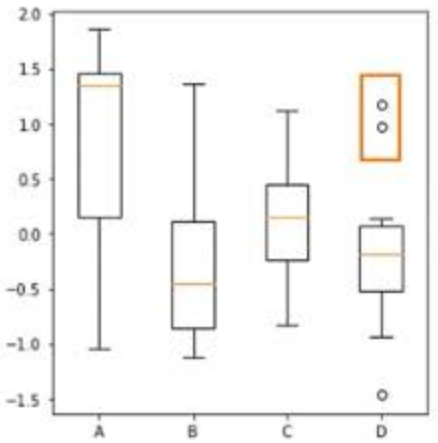
含义

- 用作显示一组数据分散情况资料的统计图；
- 能显示出一组数据的最大值、最小值、中位数、及上下四分位数。

绘图方法

- plt.boxplot(x, labels=None)
- x：表示每个需要绘制的数据
- Labels：列表，箱型线的标签

```
# 数据准备
# 生成0-1之间的10*4维度数据
data = np.random.normal(size=(10, 4))
labels = ['A', 'B', 'C', 'D']
plt.figure(figsize=(10,5))
plt.subplot(121)
# matplotlib绘制图
plt.boxplot(data, labels=labels)
```

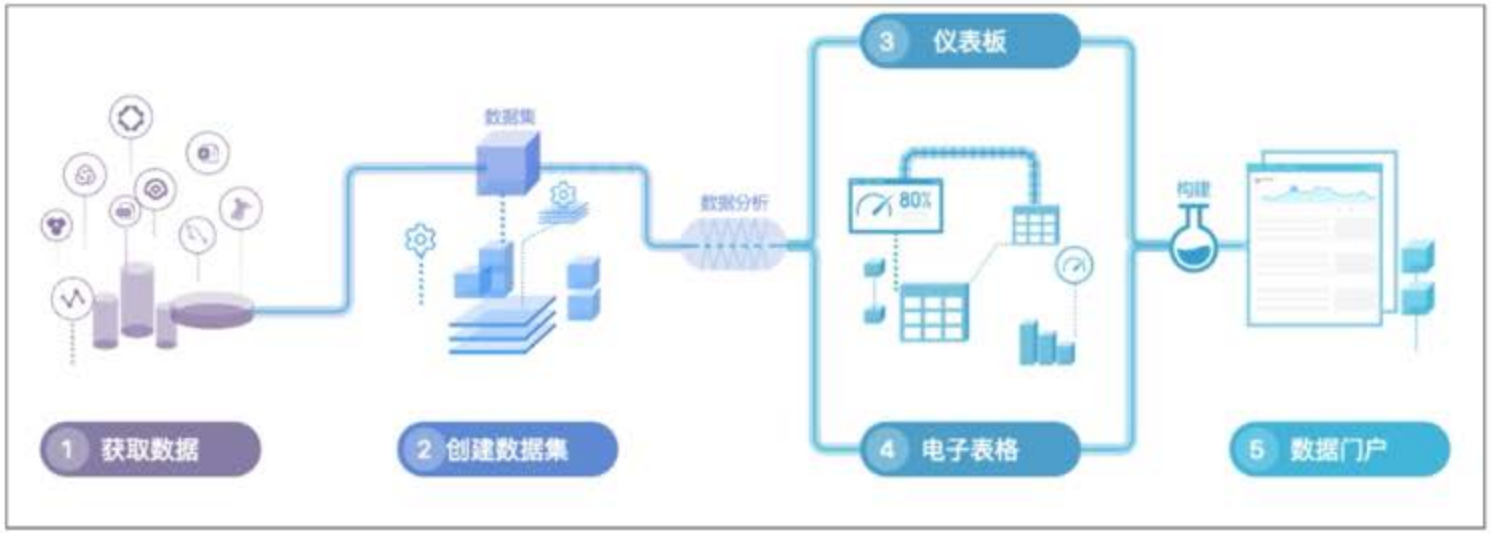


异常点

数据可视化工具QuickBI概述

阿里云

- Quick BI是一款全场景数据消费式的BI平台；
- 可以用于制作仪表板、电子表格以及有分析思路的数据门户；
- 可以将报表集成在业务流程中并分享给协作伙伴。



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

QuickBI的功能特点

阿里云

强大的数据引擎

无缝集成云上数据支持**多种数据源**：支持云数据库、关系型数据库、Hadoop、MPP等数据源。

快速搭建数据门户

快速搭建数据门户**拖拽式**操作、强大的数据建模、丰富的数据可视化图表，快速**搭建数据门户**。

数据分析和交互

智能**数据分析和交互**提供对话式智能机器人，满足智能数据洞察和数据预警需求。

安全管控数据权限

安全管控数据权限内置组织成员管理；同一份报表可以授权给不同的用户。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

QuickBI的应用场景-数据即时分析与决策

阿里云

某科技企业在业务数据化运营中，经常需对用户留存率、活跃率等进行数据报表分析。

取数难。业务人员需经常找技术写SQL取数查看各个维度的数据做决策。

报表产出效率低。后台分析系统的数据报表变更，编码研发周期长，维护困难。

图表效果设计不佳。使用HighChart等工具做报表，界面效果不佳，人力维护成本高。

使用Quick BI，数据展现丰富，操作便捷，
即时分析与即时决策快节奏，解决了以上问题

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

QuickBI的应用场景-报表与系统集成

阿里云

某运输公司期望用最低成本，最快速度搭建一个可展示、可分析的简易BI。

- 上手快上手简单，快捷，满足不同岗位的数据需求，学习门槛低。
- 极大提高看数据的效率与内部系统集成，可结合进行数据分析，极大提高看数据的效率。
- 统一系统入口解决员工使用多系统的麻烦，利于使用与控制。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据可视化工具DataV概述

- 使用可视化应用的方式来分析并展示庞杂数据的产品；
- 帮助非专业的工程师通过图形化的界面轻松搭建专业水准的可视化应用；
- 可满足会议展览、业务监控、风险预警、地理信息分析等多种业务的展示需求。

DataV的功能特点

阿里云

多种场景模板解决设计难题

- 提供指挥中心、地理分析、实时监控、汇报展示等**多种模版**

多种图表组件支撑数据展示

- 能够绘制包括**海量数据**的地理轨迹、热力分布等；
- 能够实现地理数据的多层叠加。

多数据源接入大数据计算强

- 能够接入**阿里云**的分析型数据库和关系型数据库；
- 支持**本地**上传、**在线**接入。

图形化搭建快速实现应用

- 提供**所见即所得**的配置方式；
- 只需要通过拖拽，即可创造出专业的**可视化应用。**

多分辨率适配灵活发布应用

- 提供**分辨率优化**功能；
- 能够**灵活发布分享。**

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

DataV的应用场景

阿里云

DataV数据可视化经过多年的可视化应用实践操作，已形成多样化的典型场景可供参考。

运营数据看板

地理数据看板

领导驾驶舱

指挥中心大屏

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据标注概述

阿里云

01 概念

- 通过分类、画框、标注等对语音、图片、文本数据进行处理，提高训练的准确度。

02 标注分类

- 包括语音标注、图片标注、文本标注等。

03 具体方法

- 通过画框，描点等方法对数据打标签，给后续处理提供训练信息。

04 应用场景

- 可以应用于语音识别、无人驾驶、证件识别等场景中。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据标注的重要性

阿里云

采集到的数据都需要进行数据标注后才能使用

在进行人工智能算法训练时，所训练数据的**质量**越高最后得到的模型预测效果越好

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

图像的标注方法

01 2D和3D边框

02 图像分类

03 直线和曲线

04 多边形

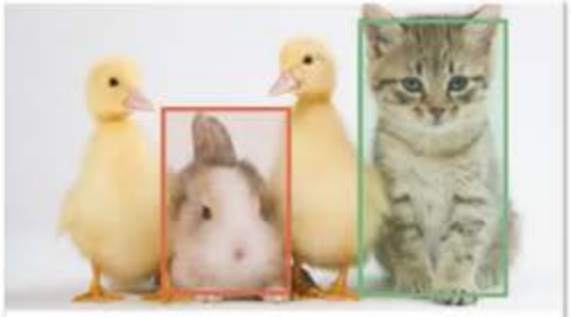
05 语义分割

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

图像的标注类别（1/2）

阿里云

目标检测



- 图像中的具体目标进行定位；
- 常用矩形框工具；
- 应用：车辆检测、行人检测、图片搜索等。

语义分割

- 识别标注图像中存在的内容及位置；
- 常用多边形描点工具、笔刷工具及超像素工具；
- 应用：自动驾驶、表情识别、服装分类等。

图像分类

- 从分类标签集合中，找到与输入图像内容相匹配的分类标签，并将其分配给该输入图像；
- 应用：相册图片分类、拍照识图、图片搜索等。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

图像的标注类别（2/2）

光学字符识别OCR

- 首先将输入图像中的文字转换为文本格式；
- 再根据文字信息类别对输入图像进行分组；
- 应用：证件识别、票据识别、车牌识别等。

图像综合标注

- 指在一组标签集合中，对输入图像的图片内容进行标签匹配；
- 应用：自动驾驶、内容审核及内容识别等。

常用的图像标注工具

阿里云

LabelImg

- 一种图形图像注释工具
- 主要用于图像分类和目标检测

Labelme

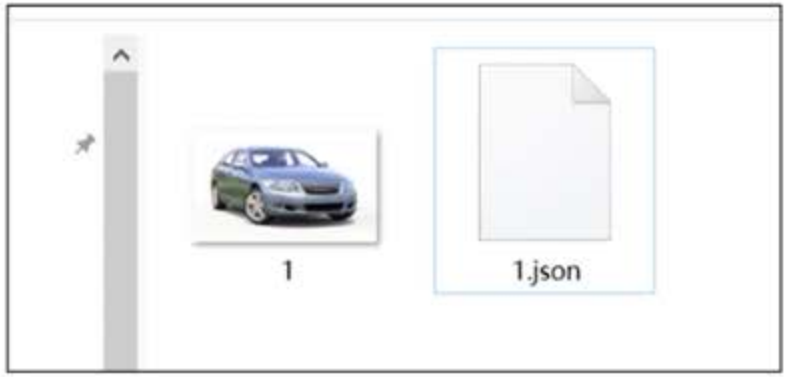
- 图形界面的图像标注软件
- 主要用于语义分割

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

图像标注示例

阿里云

使用Labelme多边形标注车辆，标注完成后会在当前图片路径处生成json格式文件;
内含目标物体的标注信息，包括标签、颜色等信息。



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

图像的标注质量标准

阿里云

矩形框标注：

需要让框刚好包围物体的边界。

多边形标注：

多边形的边框与物体的边缘紧密的贴合。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本的标注类别

常见的
文本标注类别

分类标注：对文本进行分类标注的过程。

实体标注：对文本信息中的通用实体进行标注。

词性标注：对词的性质进行标注，如名词（n）、动词（v）。

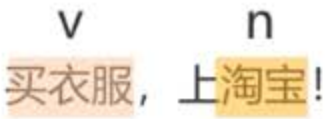
实体关系标注：对文本实体之间的关系进行标注。

服务很好！ 积极

配送太慢了！ 消极

分类标注

实体标注



词性标注

实体关系标注

文本的标注质量标准

- 文本标注要情感符合真实的句子情感
- 语义标注要标注正确的语义
- 多音字要符合字典中的读音。

- 对文本中的不感兴趣的内容进行删除。
- 将文本分成词语。
- 对词语进行词性的标注，比如形容词、名词、动词等。
- 去掉对文本的含义无用的词语，比如标点符号。

语音的标注类别

阿里云

- 对语音对应的文本信息进行关联，常用于语音识别，实时翻译等领域；
- 语音标注工具主要用于对数字化的语音信号进行分析、标注、处理及合成；

音频分类

对音频按照预设标签进行分类标记，支持单标签和多标签

音频分割

对音频数据集的内容进行分割并分段添加标签

音频识别

将音频内容的文字识别出来

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

语音的标注质量标准

阿里云

- 音频中的语音是否有效；
- 说话人的方言，标记是否有口音；
- 说话人的数量，标注语音内容的人数；
- 说话人的性别，标注第一个说话人的性别；
- 音频是否有明显的噪音，标注是否有噪音；
- 标注需要与发音内容完全一致，保证文字的正确性。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

数据标注的常用文件格式

阿里云

XML

- 一般数据集的标注格式为xml。

```
<annotation>
    <folder>JPEGImages</folder>
    <filename>00002.jpg</filename>
    <path>F:\Ali_Project\Project_code\2021_1_7code\VOCdevkit\VOC2021
\JPEGImages\00002.jpg</path>
    <source>
        <database>Unknown</database>
    </source>
    <size>
        <width>480</width>
        <height>364</height>
        <depth>3</depth>
    </size>
    <segmented>0</segmented>
    <object>
        <name>person</name>
        <pose>Unspecified</pose>
        <truncated>0</truncated>
        <difficult>0</difficult>
        <bndbox>
            <xmin>185</xmin>
            <ymin>62</ymin>
            <xmax>279</xmax>
            <ymax>199</ymax>
        </bndbox>
    </object>
</annotation>
```

JSON

- 每一个标注区域都有特定的属性与位置信息。

```
{
  "data": {
    "picUrl": "oss://****/pics/fruit/apple-1.jpg"
  },
  "label-****(标注任务ID)": {
    "results": [{
      "data": [{
        "id":"Inyumd-*****",
        "type":"image/rectangleLabel",
        "value":{
          "rotation":0,
          "x":40.65320610687023,
          "width":327.52035623409663,
          "y":5.762467474590647,
          "height":296.68117192104745
        },
        "labelColor":"#72bf7d",
        "labels":["apple"]
      }],
      "id":"44****",
      "type":"image"
    }]
  }
}
```

CSV

- 前四列为属性；
- 属于二值分类。

5	3.5	1.6	0.6	1
5.1	3.8	1.9	0.4	1
4.8	3	1.4	0.3	1
5.1	3.8	1.6	0.2	1
4.6	3.2	1.4	0.2	1
5.3	3.7	1.5	0.2	1
5	3.3	1.4	0.2	0
7	3.2	4.7	1.4	0
6.4	3.2	4.5	1.5	0
6.9	3.1	4.9	1.5	0
5.5	2.3	4	1.3	0

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

机器学习PAI平台的智能标注介绍

阿里云

- 支持图像、文本、视频、音频等多种数据类型的标注以及多模态的混合标注
- 提供了丰富的标注内容组件和题目组件
  - 图像类：图片OCR、目标检测、图像分类
  - 文本类：实体识别、文本分类、实体关系
  - 语音类：音频分类、音频分割、音频识别
  - 视频类：视频分类
- 如果预置的标注模板无法满足需求，还可以选择自定义模板

视频分类

对视频按照预设标签进行分类标记，支持单标签和多标签

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

机器学习PAI平台的数据标注步骤

阿里云

Step 1

数据准备

使用阿里的PAI平台的进行数据准备工作。

Step 2

创建标注任务

使用阿里的PAI平台的标注工具创建数据标注任务。

Step 3

处理标注任务

打开阿里云PAI平台的标注工具进行数据标注。

Step 4

导出标注结果

在指定OSS存储路径中，可以查看标注结果。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER