7点开始

Apache Spark中国技术…

2496人

扫一扫群二维码，立刻加入该群。

Apache Spark中国技术社区

机器学习介绍与Spark MLlib实践

阿里云-E-MapReduce 江宇

2018.12.6

内容

1

机器学习介绍

2

Spark MLlib介绍

3

Spark MLlib实践

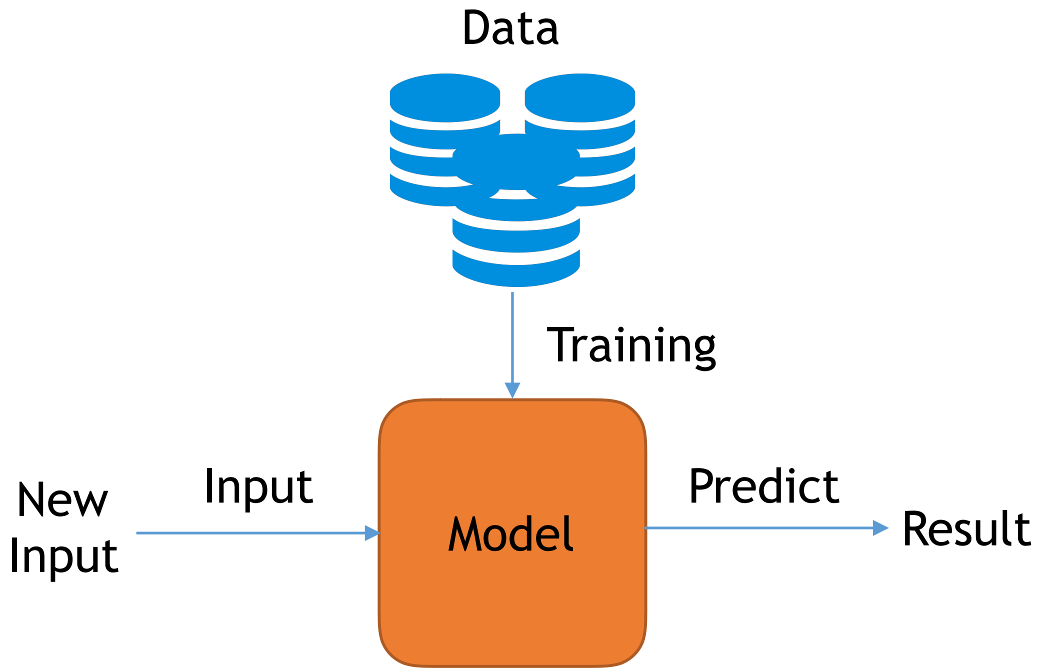
机器学习介绍

Part I

机器学习介绍

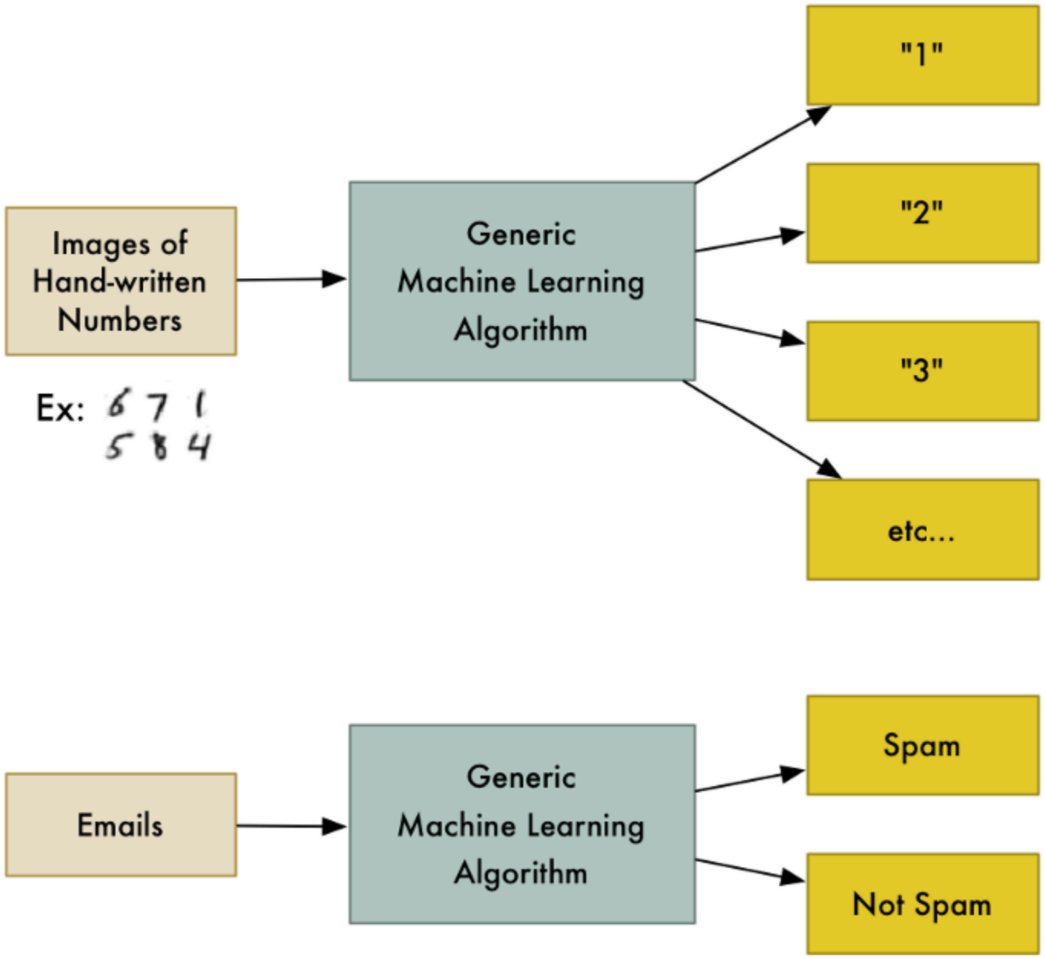
机器学习概念

Machine learning (ML)
is the study of algorithms and mathematical models that computer systems use to progressively improve their performance on a specific task.



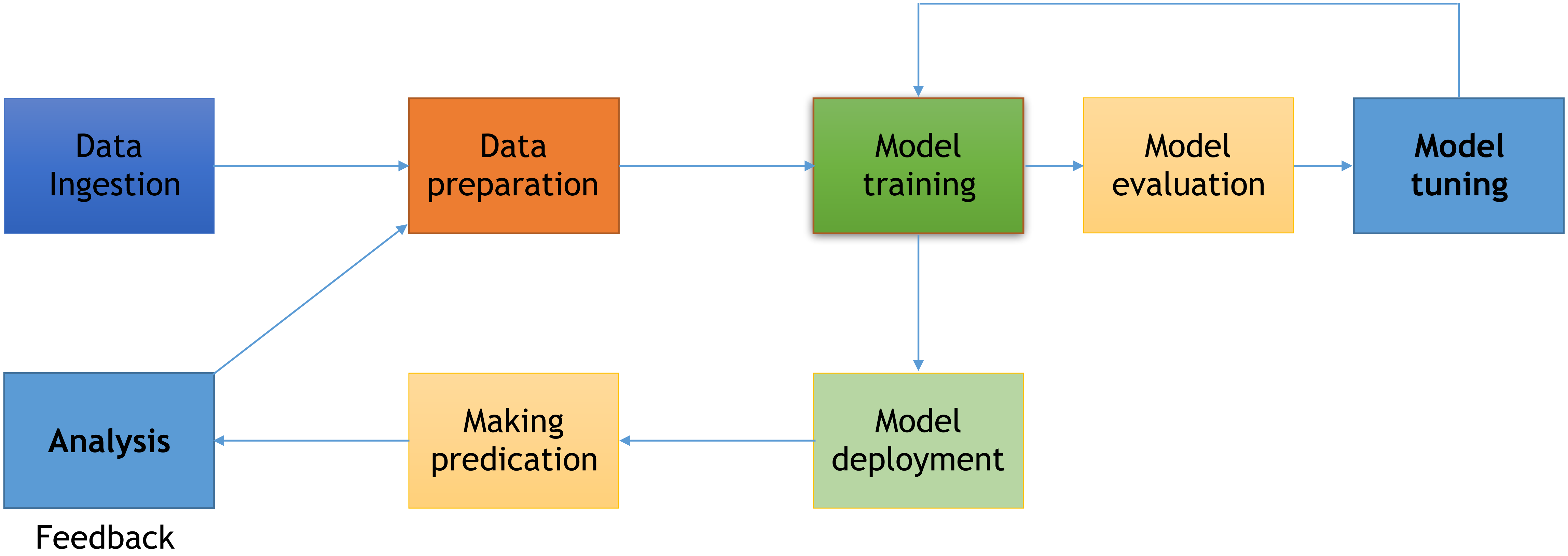
机器学习介绍

机器学习举例


Images of Hand-written Numbers
Ex: 6 7 1
5 8 4
Generic Machine Learning Algorithm
"1"
"2"
"3"
etc...
Emails
Generic Machine Learning Algorithm
Spam
Not Spam

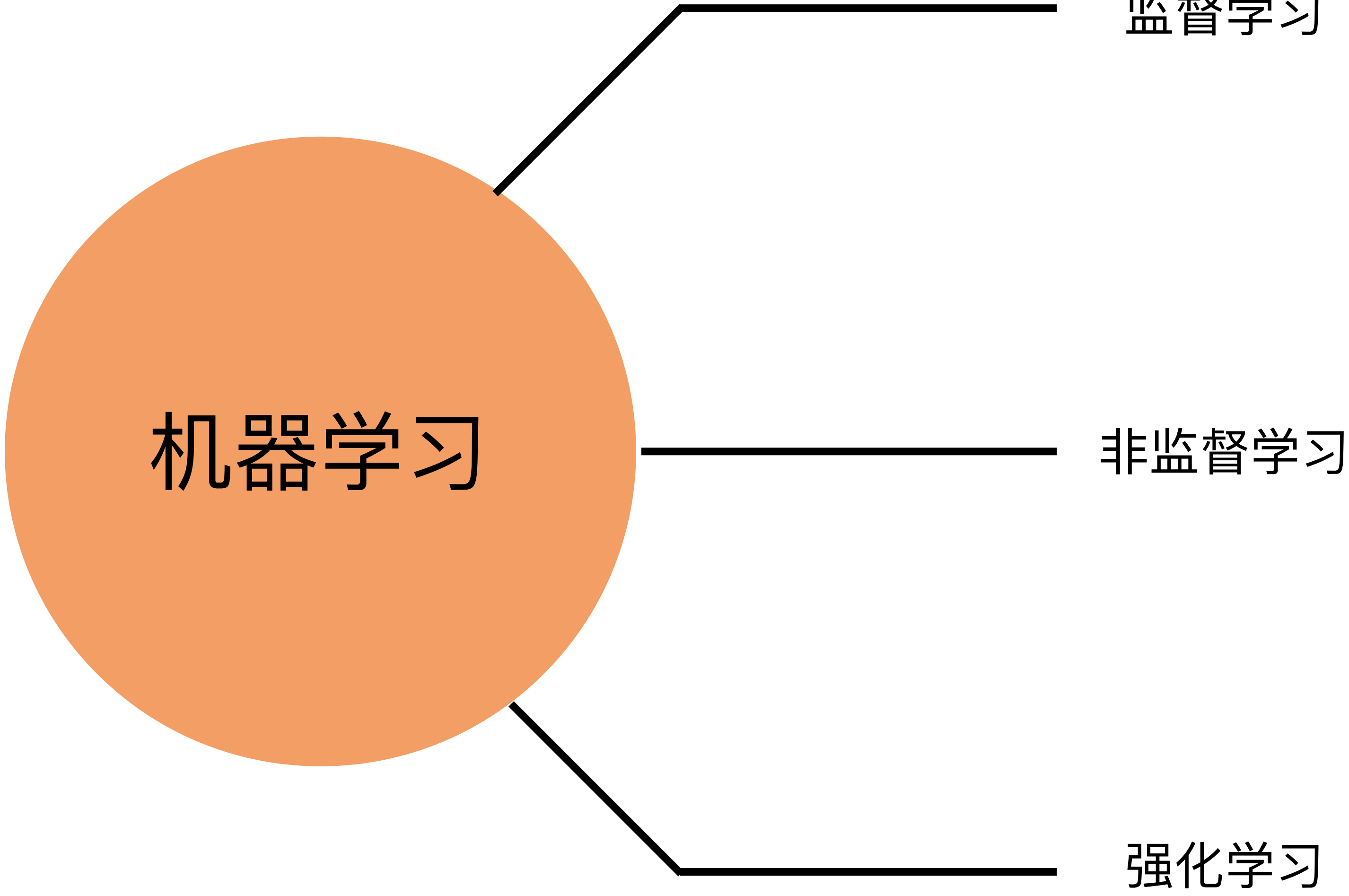
机器学习介绍

机器学习流程



机器学习介绍

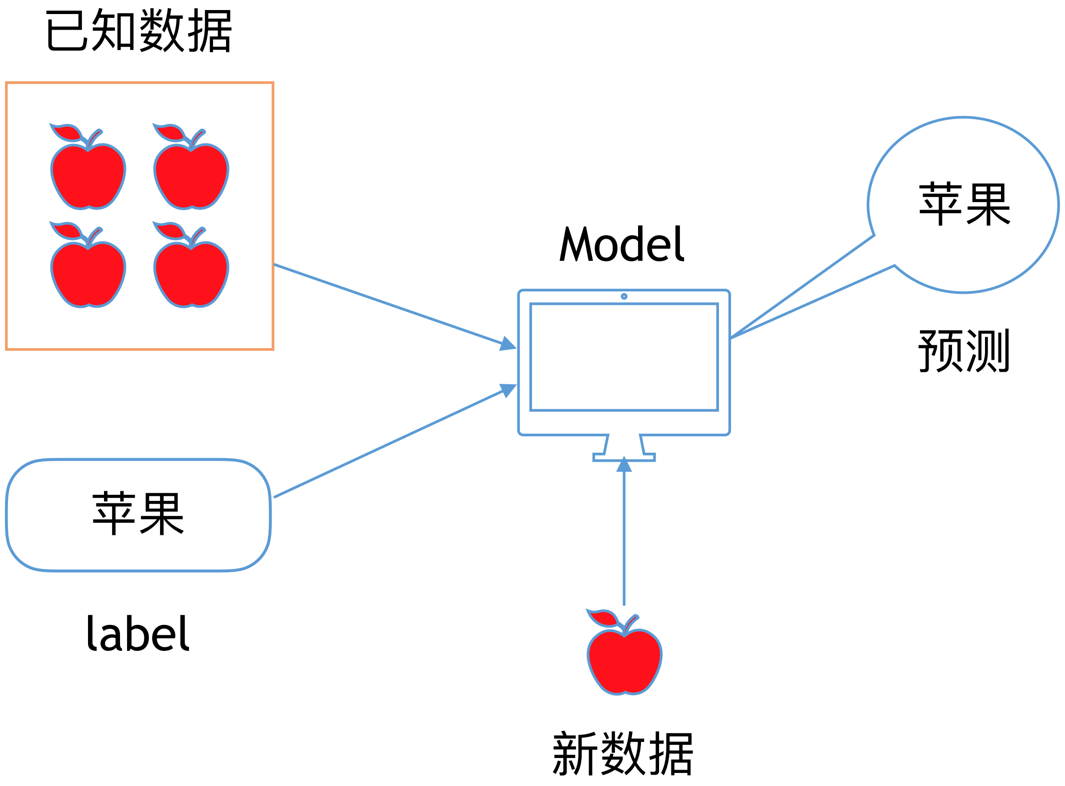
机器学习分类



机器学习介绍

监督学习

- label training
- 找到function 能够将input 映射到 output



机器学习介绍

监督学习基本概念

- **Input features:** $x^{(i)} \in \mathbb{R}^n, \quad i = 1, \ldots, m$
- **Output:** $y^{(i)} \in \mathbb{R}$
- **Model Parameters:** $\theta \in \mathbb{R}^n$
- **Hypothesis function:** $h_\theta(x) : \mathbb{R}^n \to \mathbb{R}$

$$h_\theta(x) = x^T\theta = \sum_{i=1}^{n} x_i\theta_i$$

机器学习介绍

监督学习基本概念

如何衡量假设函数(hypothesis function)的准确性?

Loss functions 代价函数

假设函数越“接近”实际值越好，代价函数“越小”
假设函数越“远离”实际值越差，代价函数“越大”

$$\ell\left(h_\theta(x), y\right) = \left(h_\theta(x) - y\right)^2$$

平方误差代价函数

机器学习介绍

监督学习求解流程

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \ell\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

1. 确定假设函数(hypothesis function)
2. 确定损失函数(loss function)
3. 确定优化算法(不断调整参数)

机器学习介绍

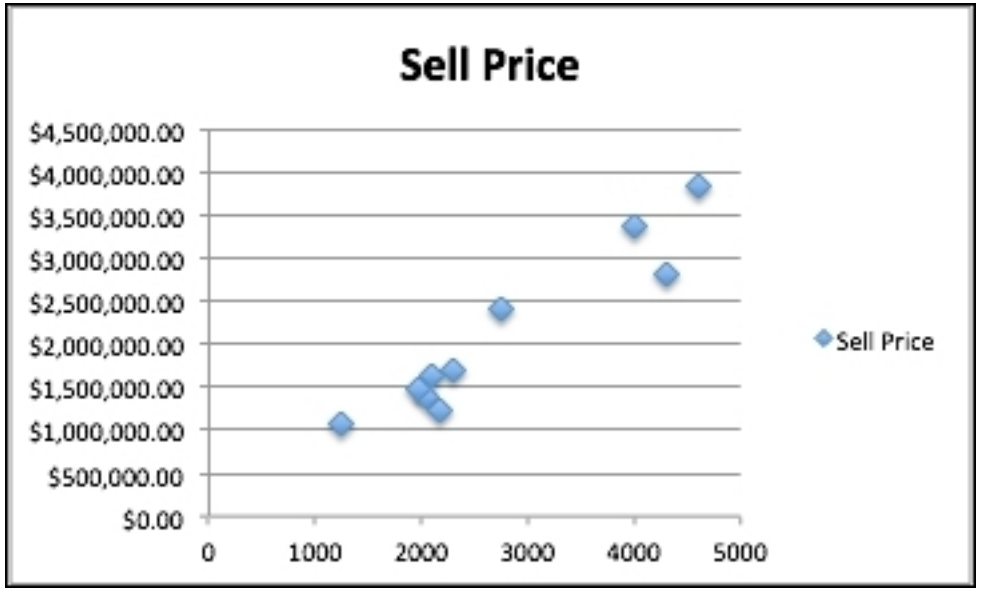
监督学习类型

1. 回归问题
2. 分类问题

机器学习介绍

监督学习—回归问题

House Size (sq ft)	Price ($)
2100	1,620,000
2300	1,690,000
2046	1,400,000
4314	2,000,000
1244	1,060,000
4608	3,830,000



1. 线性假设函数 $h_\theta(x) = x^T\theta$

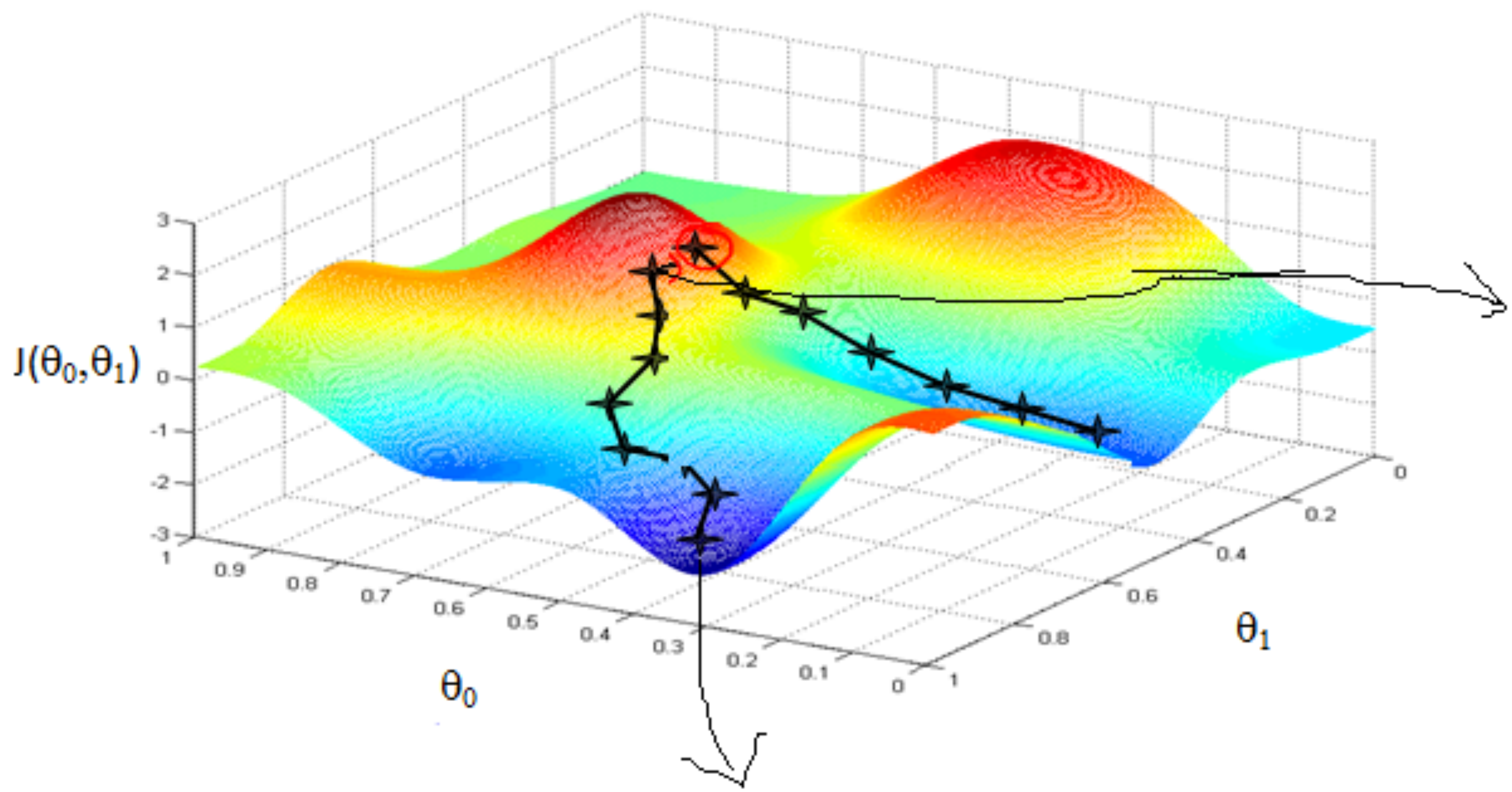
2. 平方损失函数 $\ell(h_\theta(y), y) = (h_\theta(x) - y)^2$

3. 优化方法　梯度下降法

Data URL：http://course1.winona.edu/bdeppa/Stat%20425/Datasets.html

机器学习介绍

监督学习–回归问题



初始点

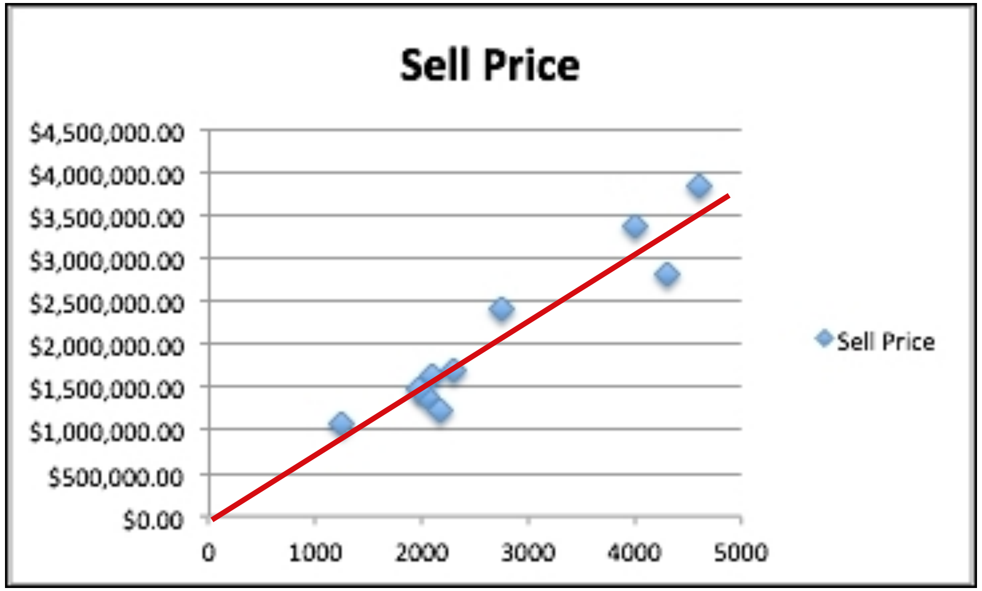
最小值

相应课程： https://www.coursera.org/learn/machine-learning

机器学习介绍

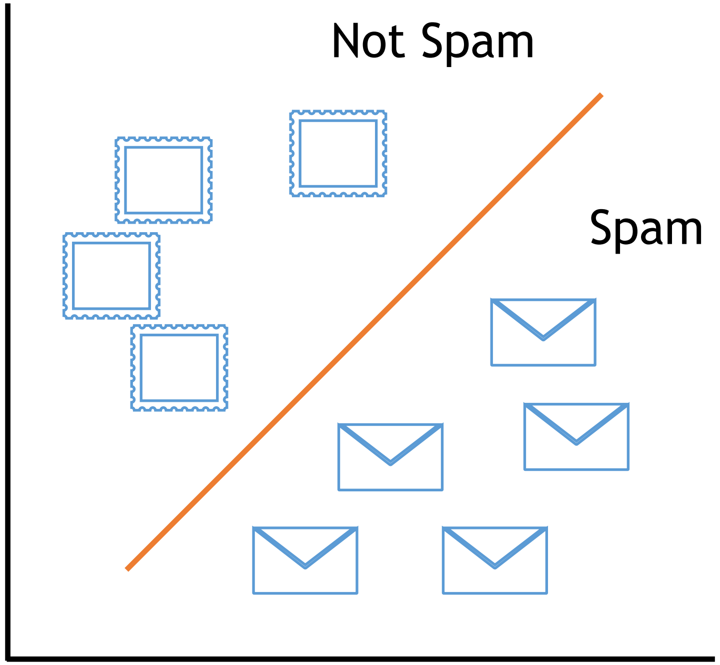
监督学习—回归问题

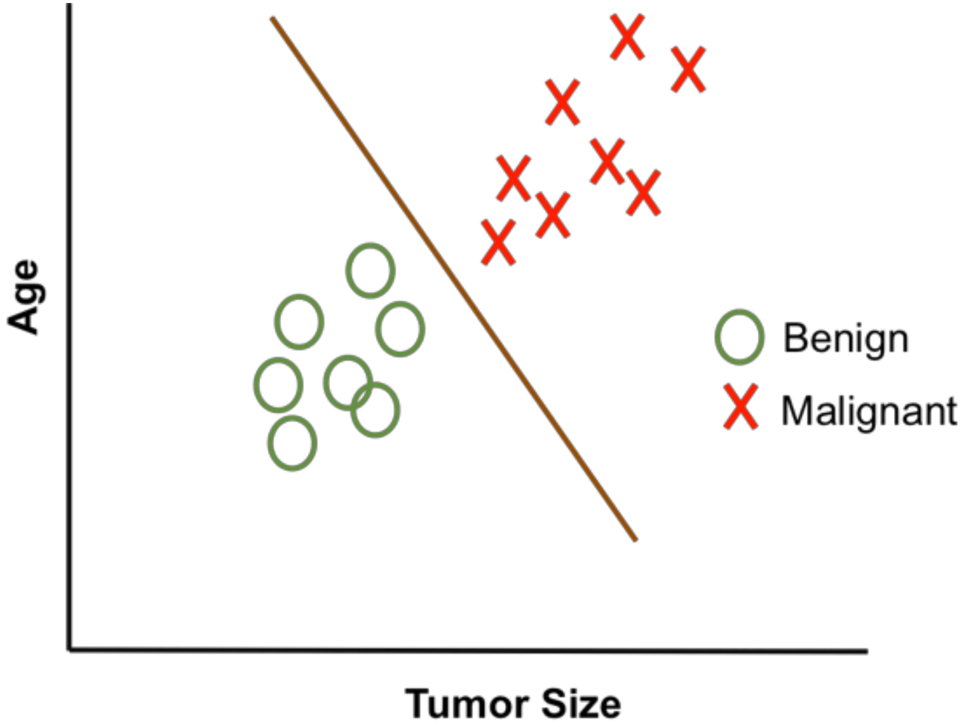
House Size (sq ft)	Price ($)
2100	1,620,000
2300	1,690,000
2046	1,400,000
4314	2,000,000
1244	1,060,000
4608	3,830,000



机器学习介绍

监督学习–分类问题

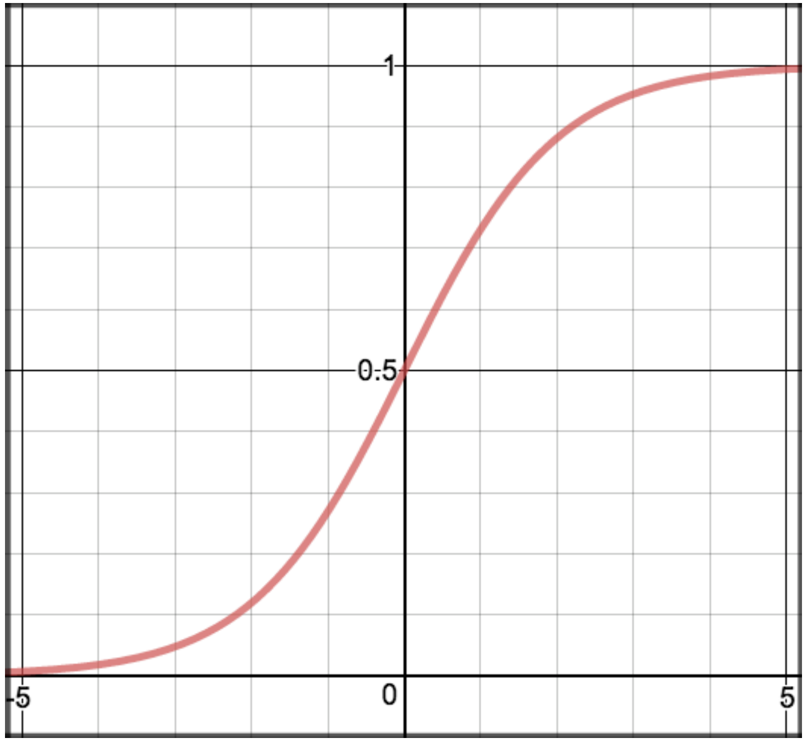




机器学习介绍

监督学习—分类问题

$$y = \frac{1}{1+e^{-x}},$$



Sigmoid function

确定假设函数(hypothesis function)

对于二分类问题，将input映射到0-1之间

0 negative
1 positive

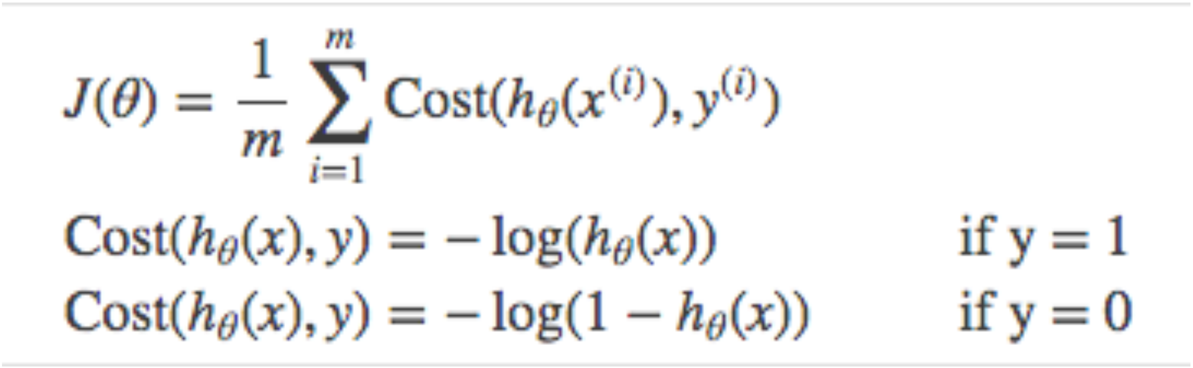
$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

机器学习介绍

监督学习–分类问题

确定损失函数(cost function)

平方损失函数？



$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if } \mathrm{y} = 1$$

$$\mathrm{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if } \mathrm{y} = 0$$

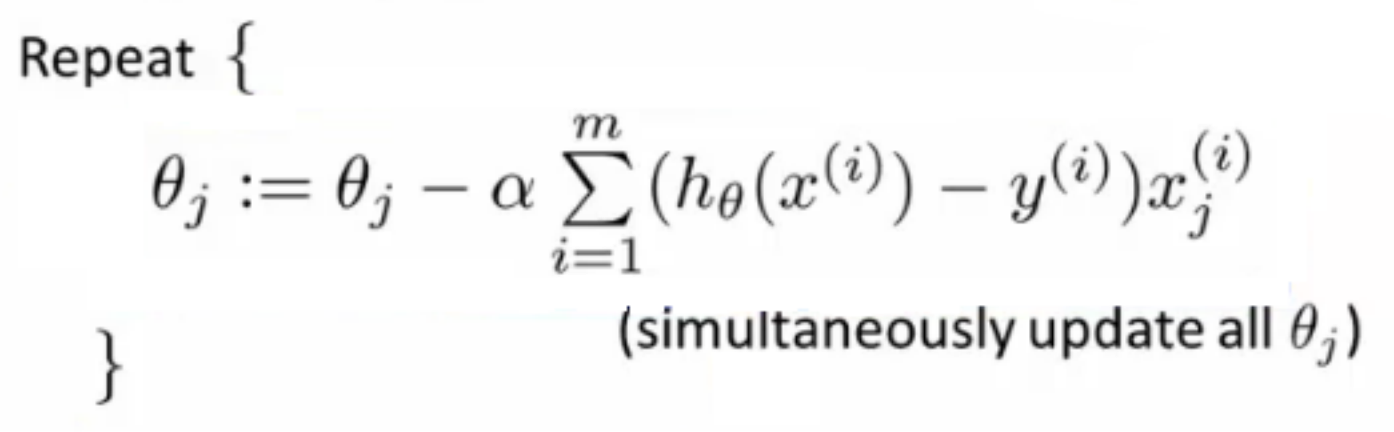
Cross-Entropy损失函数

https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#id11

机器学习介绍

监督学习–分类问题

确定优化方法



Repeat {

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

}

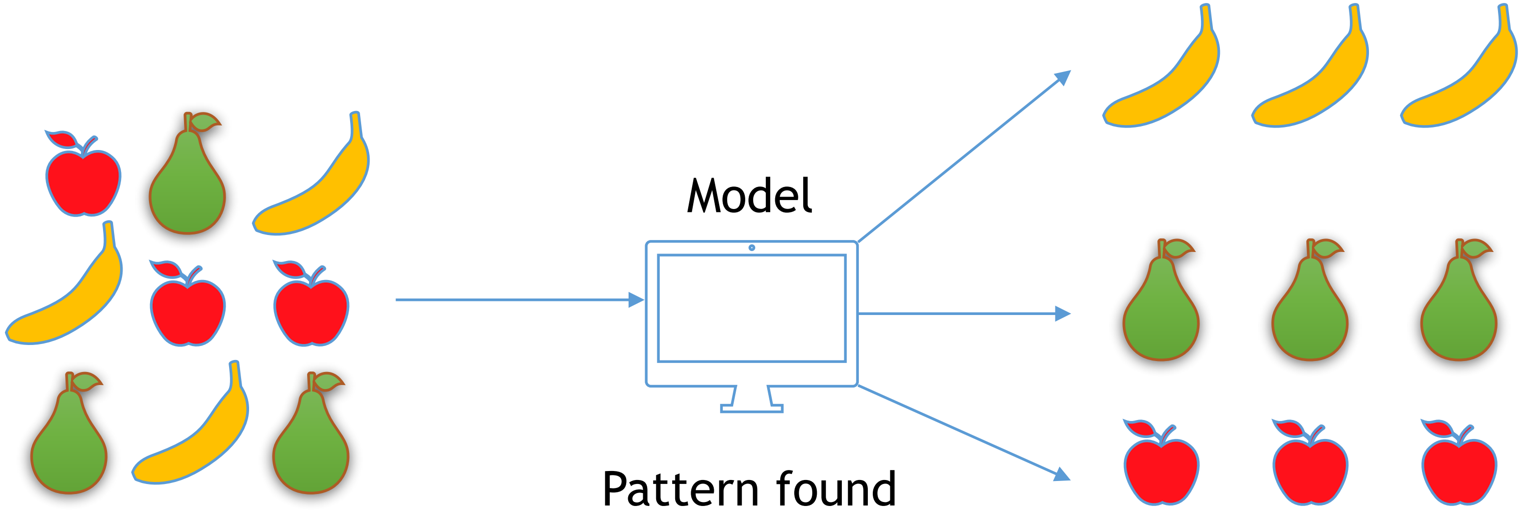
梯度下降法

相应课程： https://www.coursera.org/learn/machine-learning

机器学习介绍

非监督学习

- no label
- algorithms discover internal structures in data



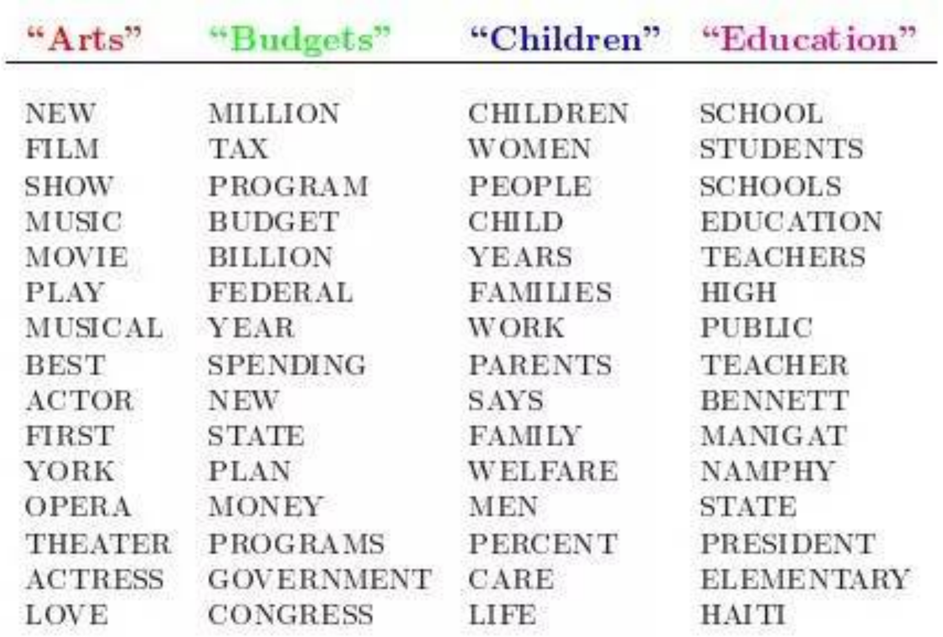
机器学习介绍

非监督学习

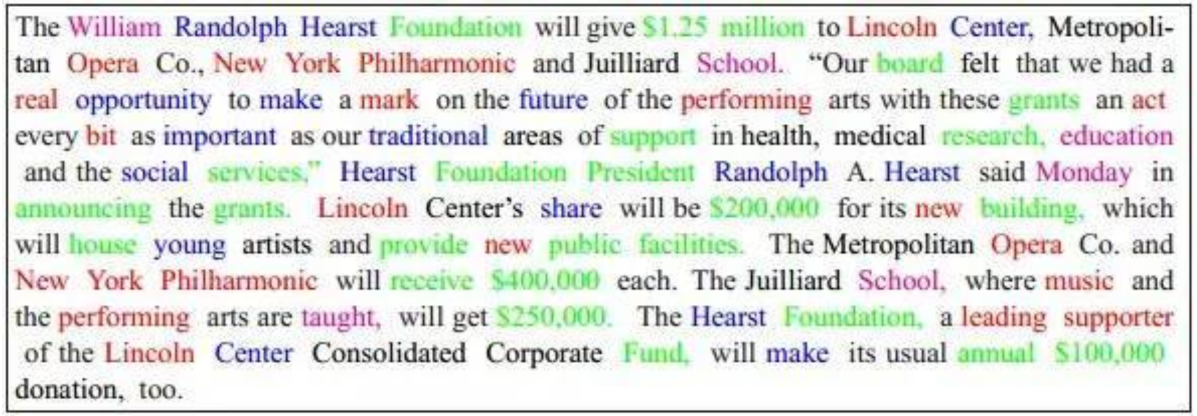
1. 聚类问题
2. 异常检测
3. LDA 主题模型

机器学习介绍

非监督学习–LDA简单介绍



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI



The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

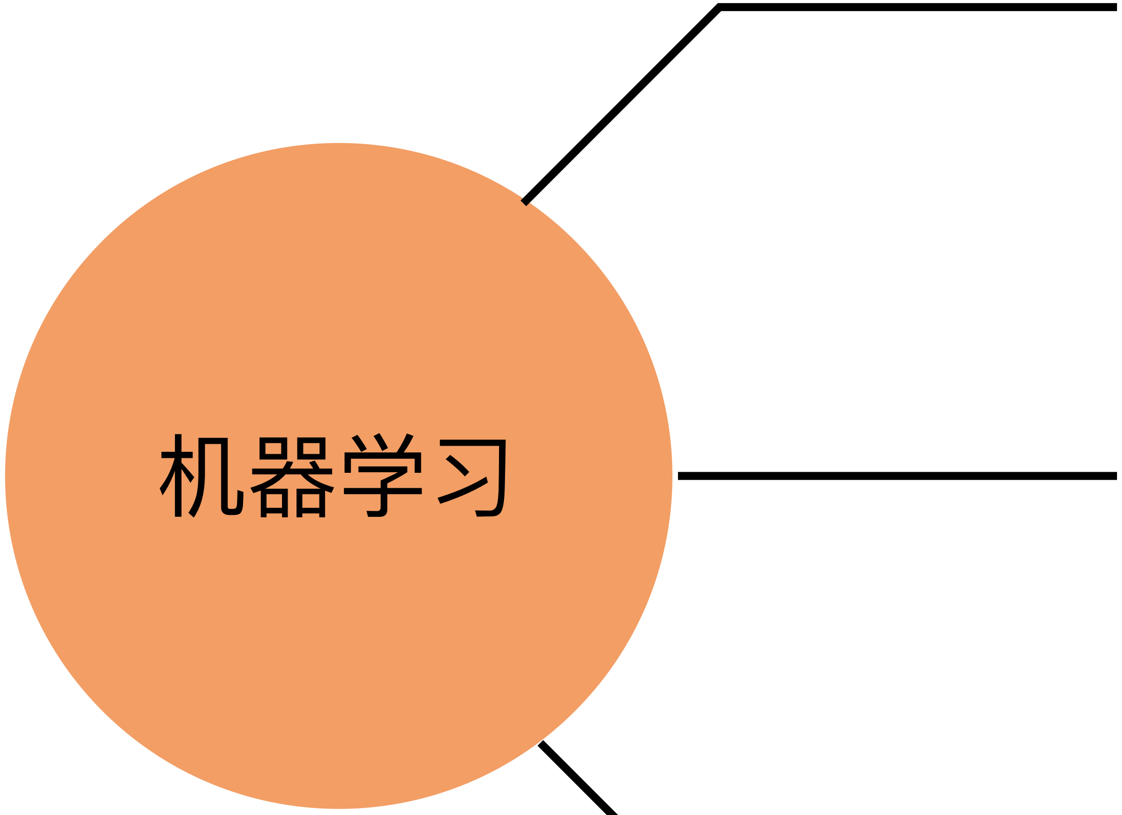
机器学习介绍

强化学习

强化学习就是通过不断与环境交互，利用环境给出的奖惩来不断的改进策略（即在什么状态下采取什么动作），以求获得最大的累积奖惩。

机器学习介绍

总结



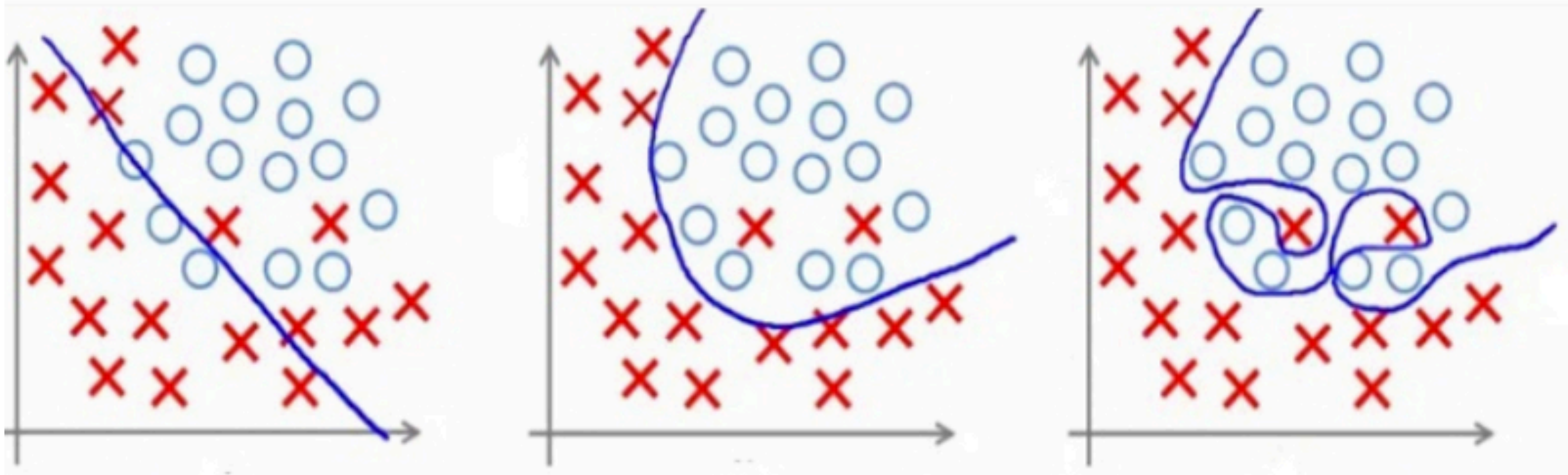
监督学习

非监督学习

强化学习

机器学习介绍

模型选择



Under-fitting

(too simple to explain the variance)

Appropriate-fitting

Over-fitting

(forcefitting -- too good to be true)

机器学习介绍

模型选择

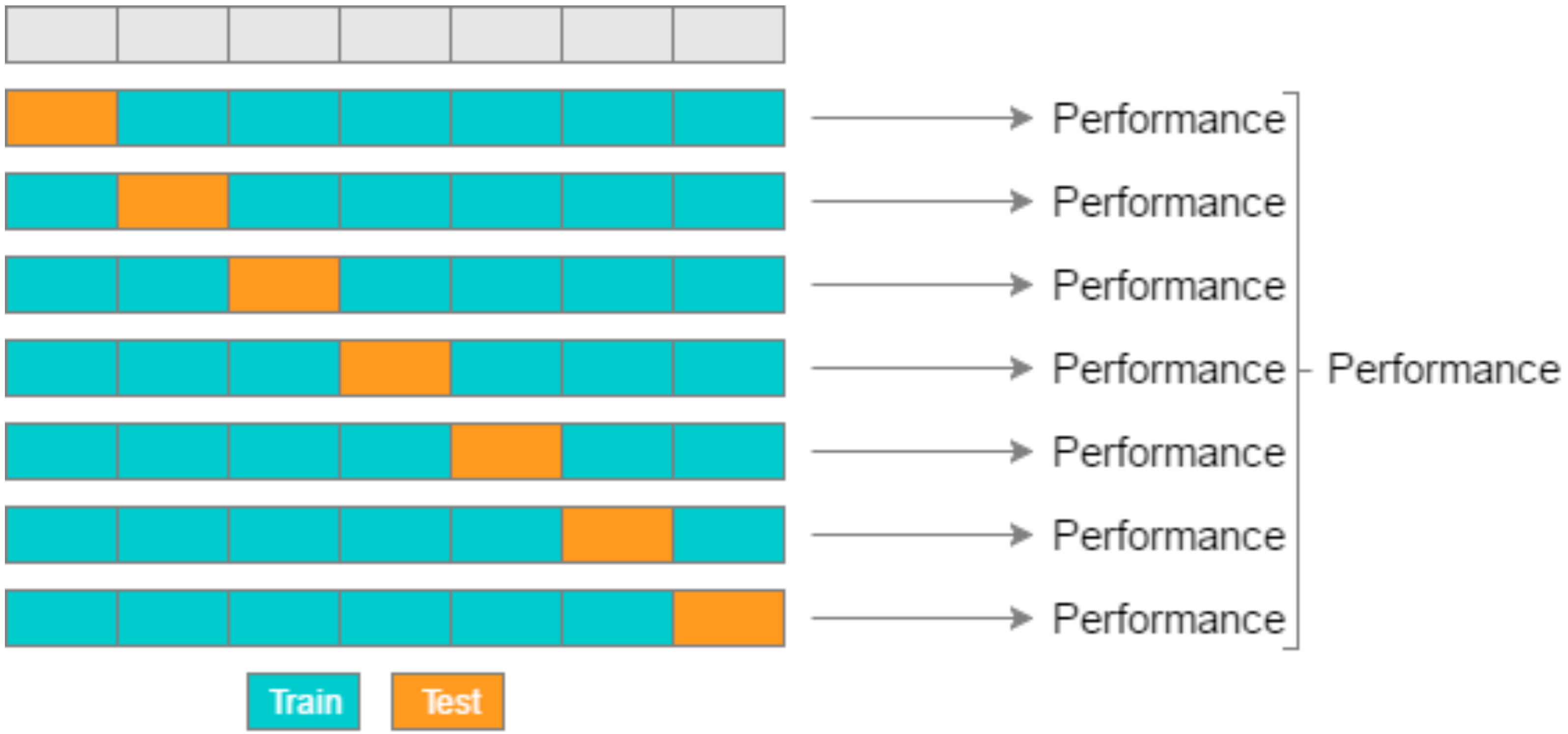
超参数：hyper parameters

- 梯度下降中的学习率
- 神经网络中的隐层规模
- k均值聚类中的簇数

机器学习介绍

模型选择

K-folds Cross Validation(K-折叠交叉验证)

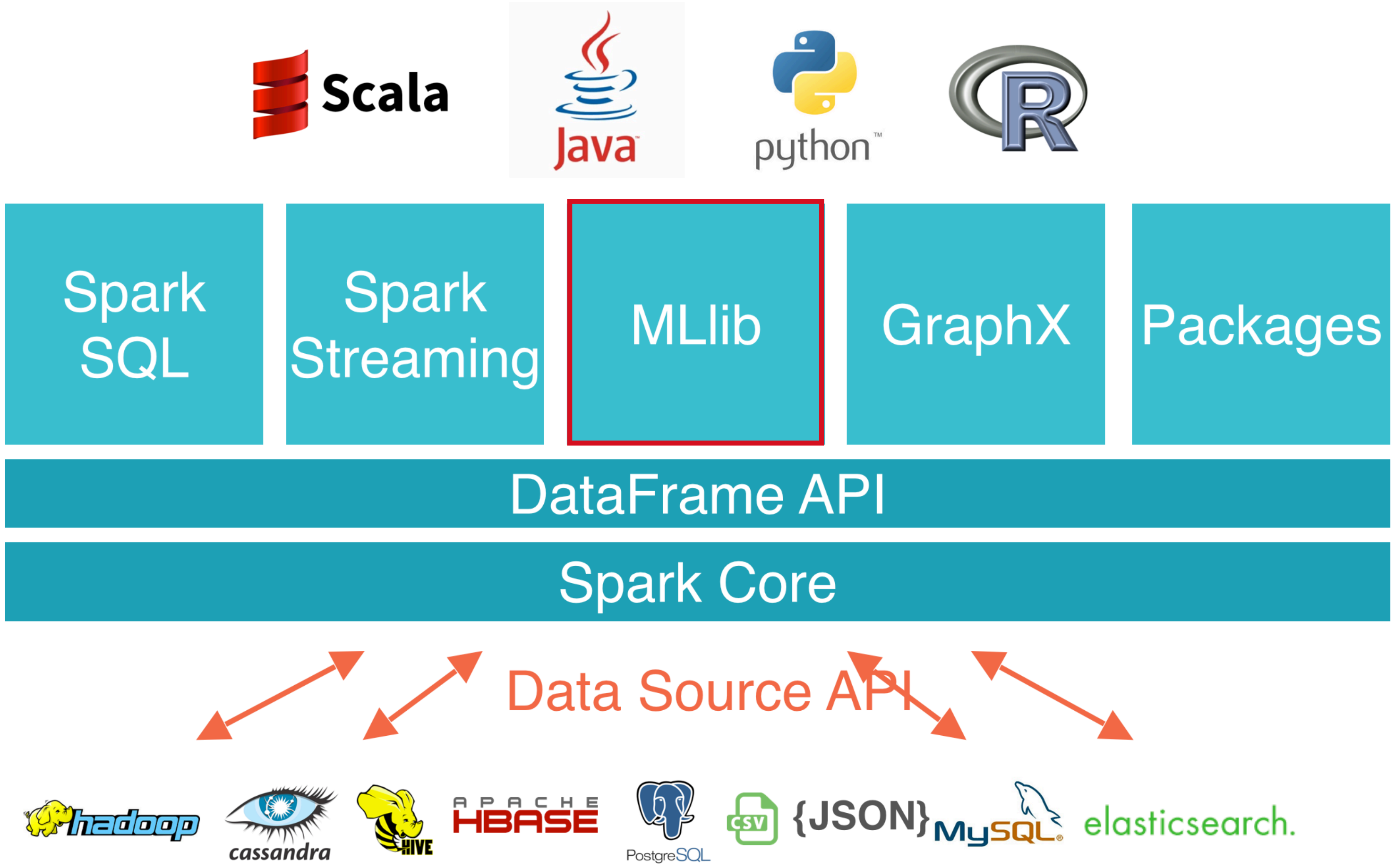


Spark MLlib介绍

Part II

Spark MLlib简介

MLlib



Spark MLlib简介

MLlib Component

Algorithms

- Classification
- Regression
- Clustering
- Recommandation

Pipeline

- construction
- Evaluation
- Tuning
- Persistence

Featurization

- Extraction
- Transformation

Utilities

- Linear algebra
- Statistics

Spark MLlib简介

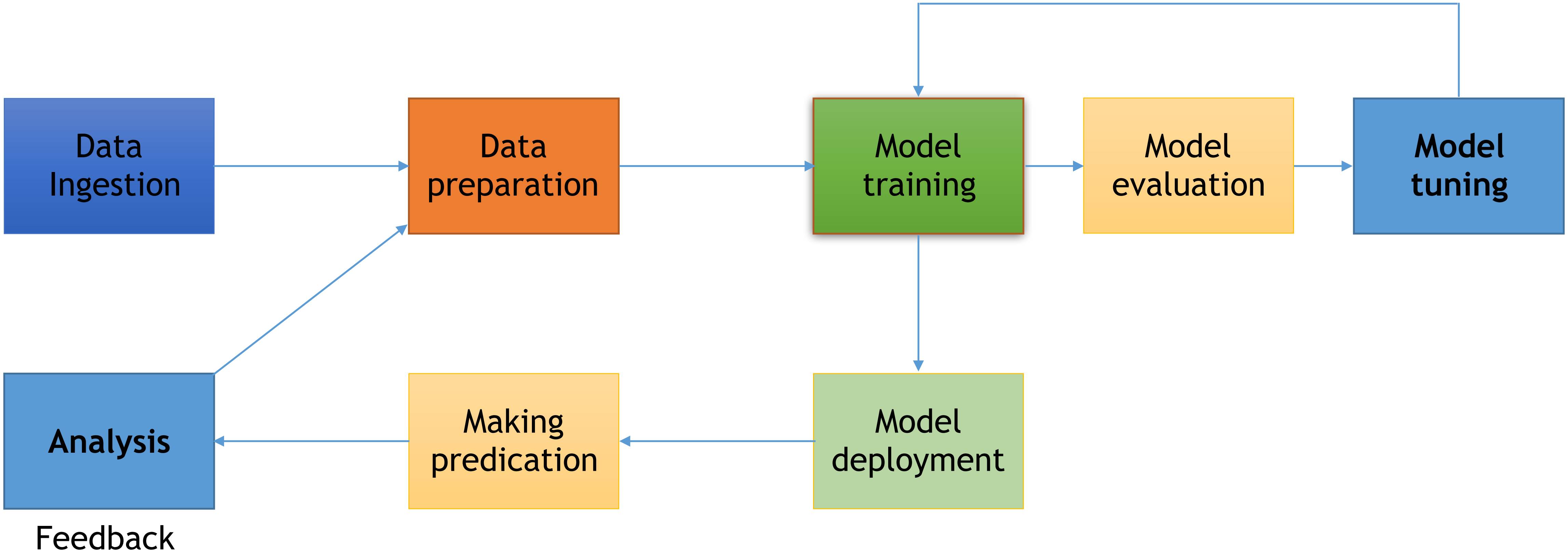
MLlib

spark.mllib　基于RDDs

spark.ml　　基于Data Frames

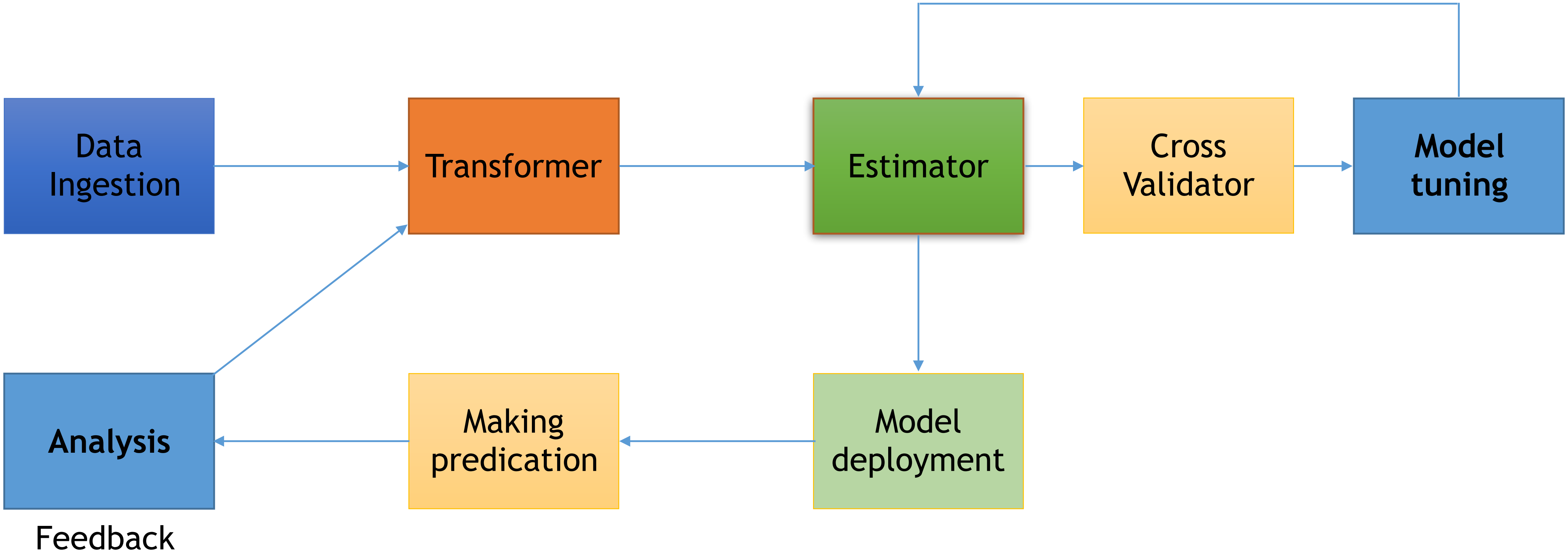
机器学习介绍

机器学习流程



Spark MLlib简介

MLlib流程

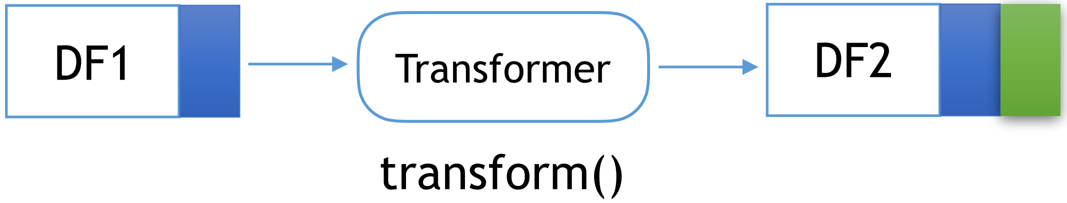


Spark MLlib简介

MLlib—Transformer

- 数据预处理
- 数据转换
- DataFrame 作为input，transform， 新的DataFrame作为output

例子：
1.Tokenizier   sentences —> words
2.VectorAssembler features columns —> vector column

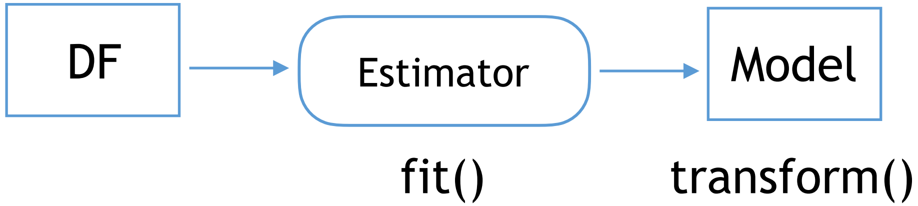


Spark MLlib简介

MLlib—Estimator

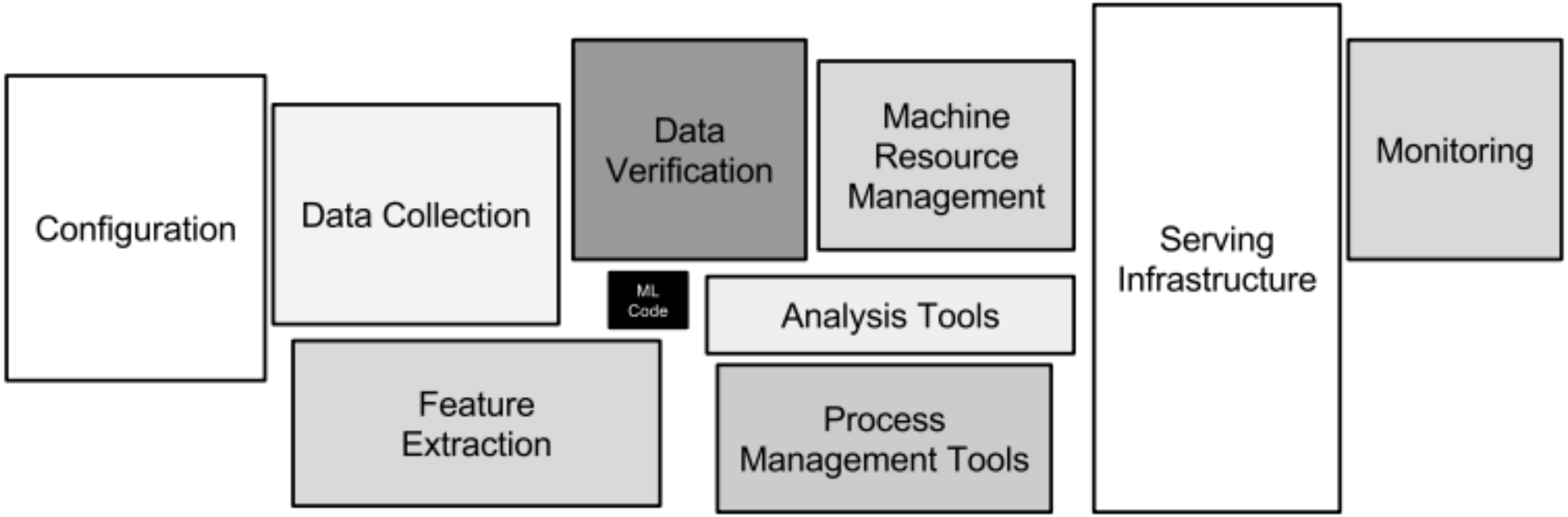
- ml algorithms abstraction
- trains(fit) on data
- DataFrame 作为input， fit， output 为Model(Transformer)

例子：
LogisticRegression -> LogisticRegressionModel



Spark MLlib简介

MLlib—Pipeline

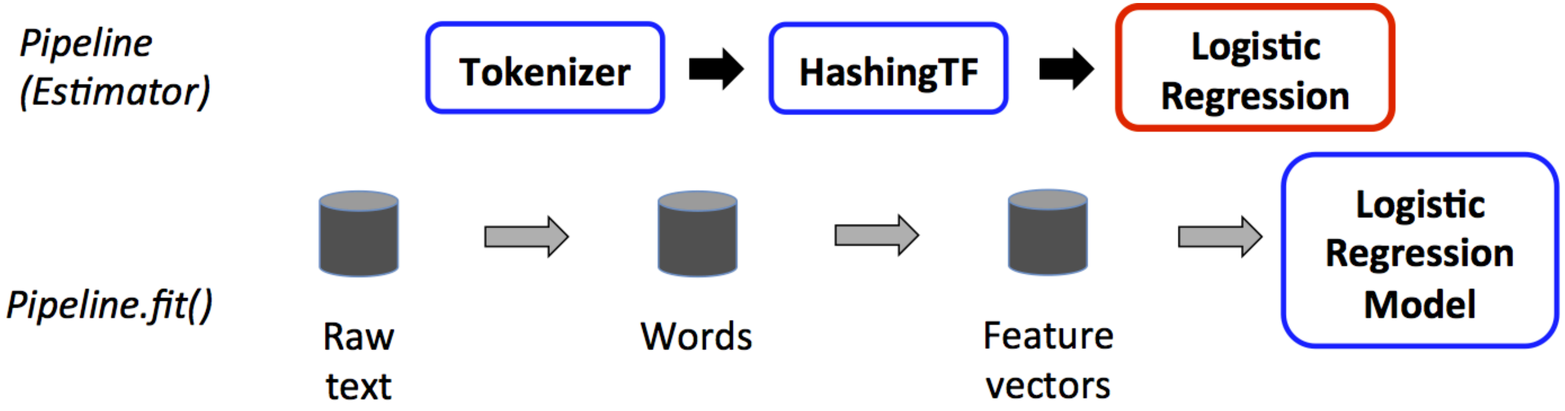


hidden technical debt in machine learning systems

Spark MLlib简介

MLlib—Pipeline

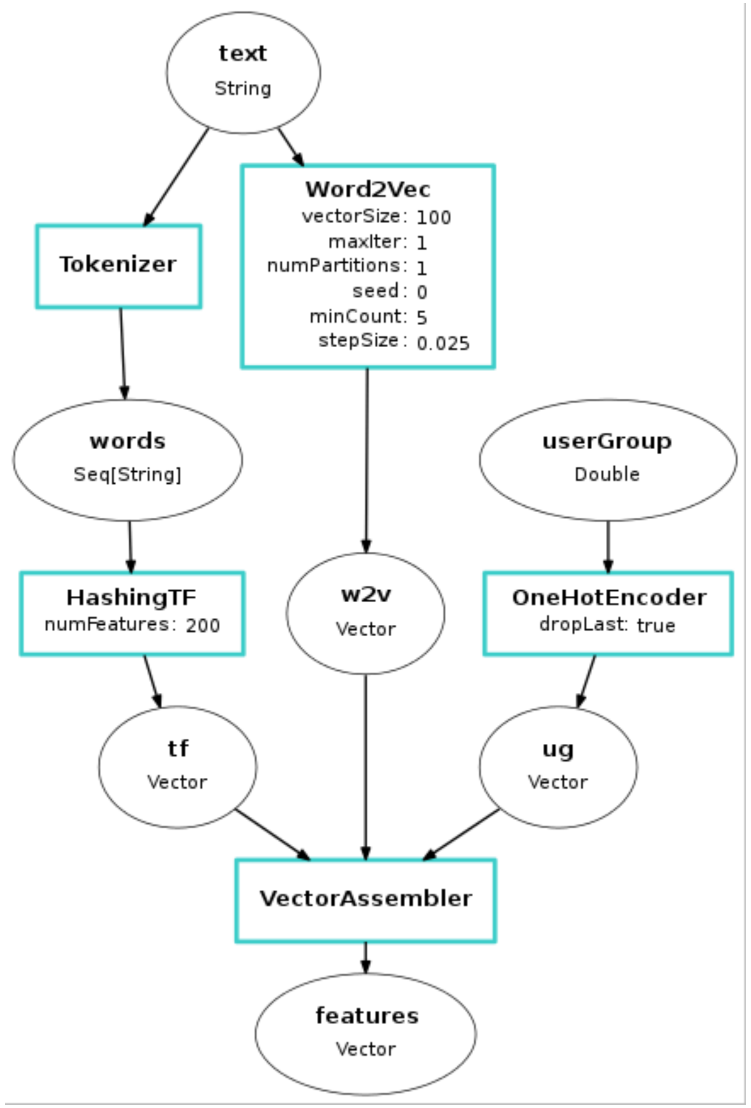
- ML workflow
- A set of stages
- Transformers & Estimators
- Persistent



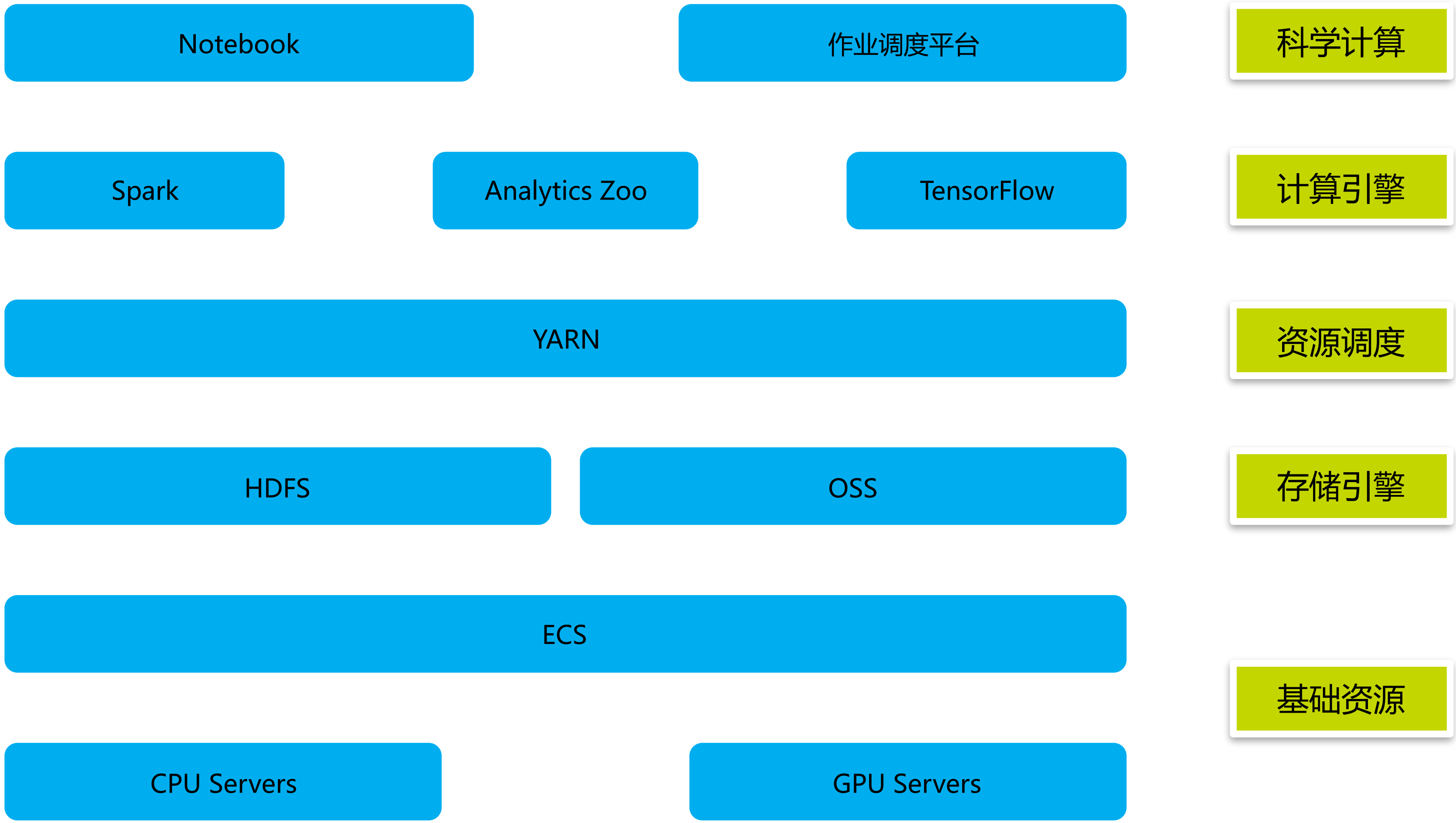
Spark MLlib简介

MLlib—Pipeline

- 现实中的Pipeline 非常复杂
- Pipeline DAG
- Pipeline ParamMap



阿里云EMR机器学习平台



阿里云EMR机器学习平台

- 整体资源管理(CPU,MEM,GPU)
- 统一的任务资源申请，资源隔离，日志查询
- Spark生态圈优化共享
- 便利的共享文件系统访问(HDFS,OSS)
- 资源监控与报警

阿里云EMR机器学习平台 – Analytics Zoo

- 重用现有的大数据工具(Spark, MR)构建应用
- 基于Apache Spark和BigDL的大数据分析+AI平台
- 内置大量模型与算法
- 提供Apache Spark高级的流水线支持，能够使用Data Frames 和ML Pipelines
- 阿里云EMR服务安装

Spark MLlib实践

Part III

Extra

Part IV

Extra

Python多版本支持

1.archives方法

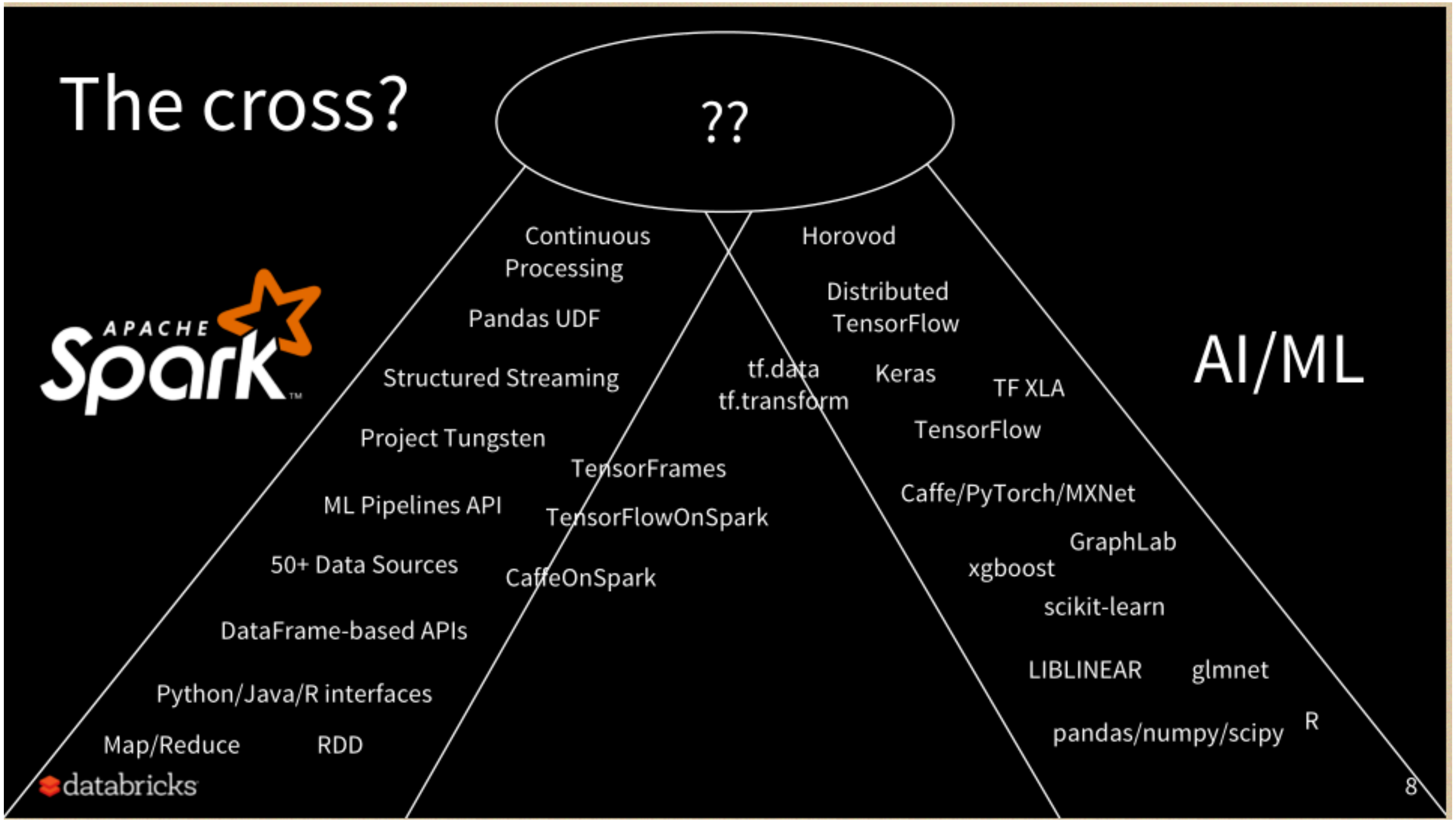
```
(1)zip -r py3.7.zip py3.7
(2)spark-submit --conf spark.yarn.appMasterEnv.PYSPARK_PYTHON=.python_alter/py3.7/bin/python \
    --master=yarn-cluster --archives py3.7.zip#python_alter  test.py
```

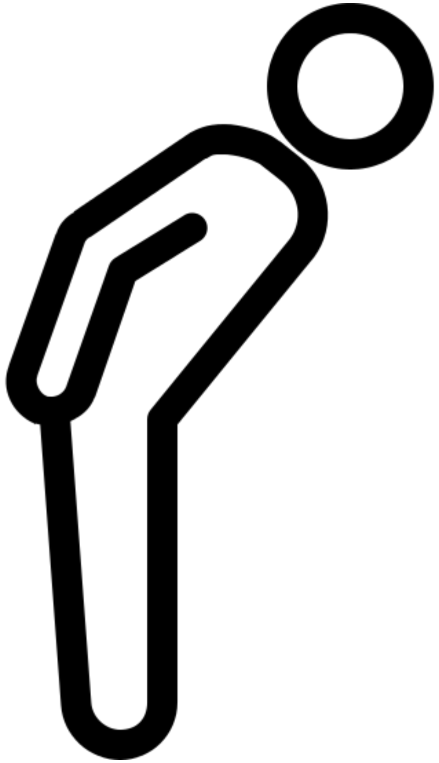
2.conda(虚拟环境)

```
(1)every slaves install conda and env (py2.7, py3.6)
(2)spark-submit --conf spark.pyspark.python=/usr/lib/anaconda/env/py3.6/bin/python \
    --conf spark.pyspark.driver.python=/usr/lib/anaconda/env/py3.6/bin/python \
    --master=yarn-cluster test.py
```

Extra

Spark MLlib与Deep Learning





谢谢！

欢迎投稿 Spark中国技术社区

Apache Spark中国技术…

2496人

扫一扫群二维码，立刻加入该群。

Apache Spark中国技术社区