HelloBI 公开课

数据科学方法论及案例分享

曾勇华

数据科学家及机器学习解决方案架构师

IBM 机器学习及数据科学部门

IBM

大纲

- 数据科学及机器学习简要
  - –数据科学
  - –机器学习
  - –数据分析方法学
- 数据科学工具箱
- 数据科学案例分享

What is Data Science?

Data science, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to Knowledge Discovery in Databases (KDD).

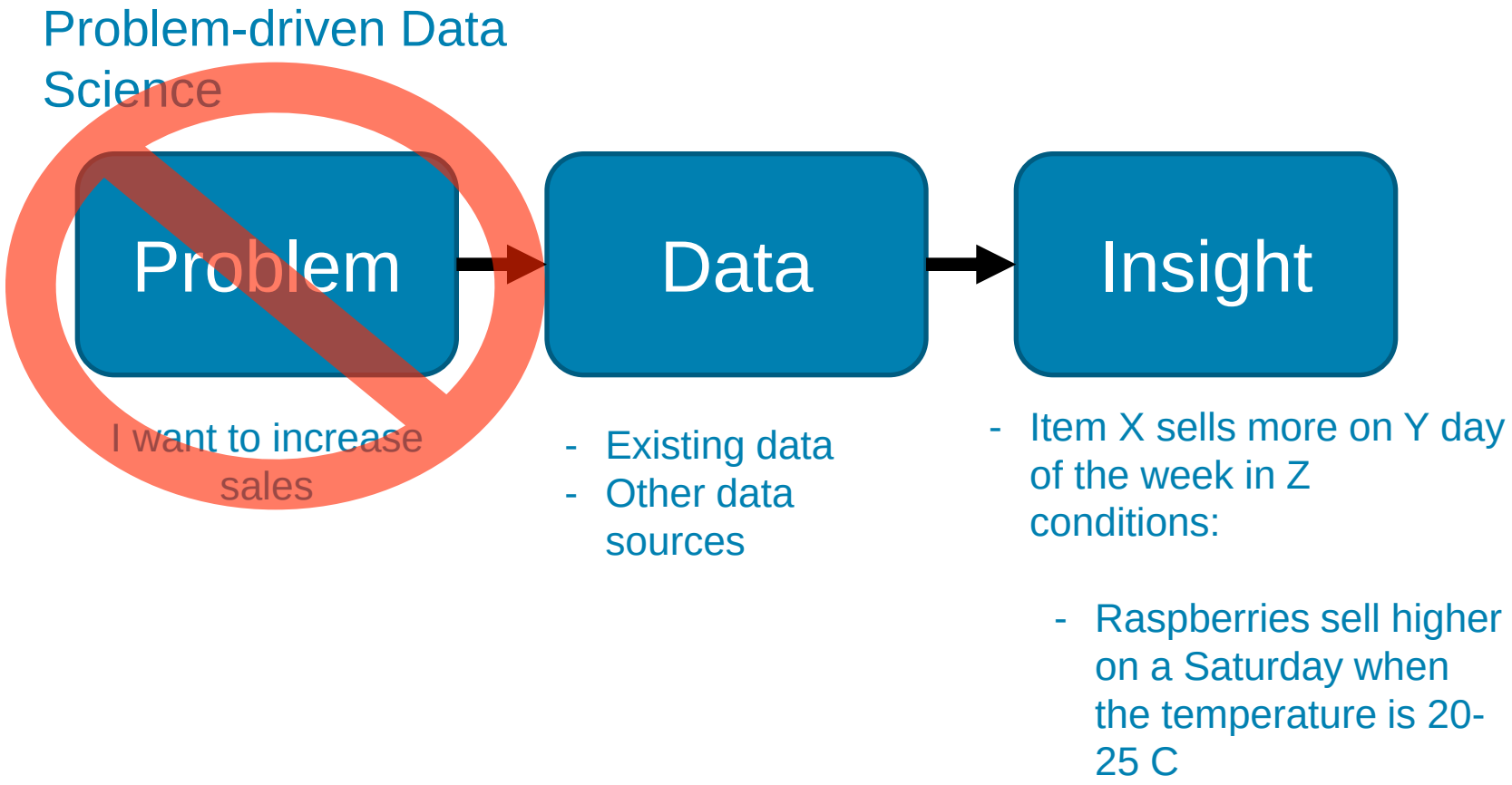
Problem-driven Data Science

I want to increase sales

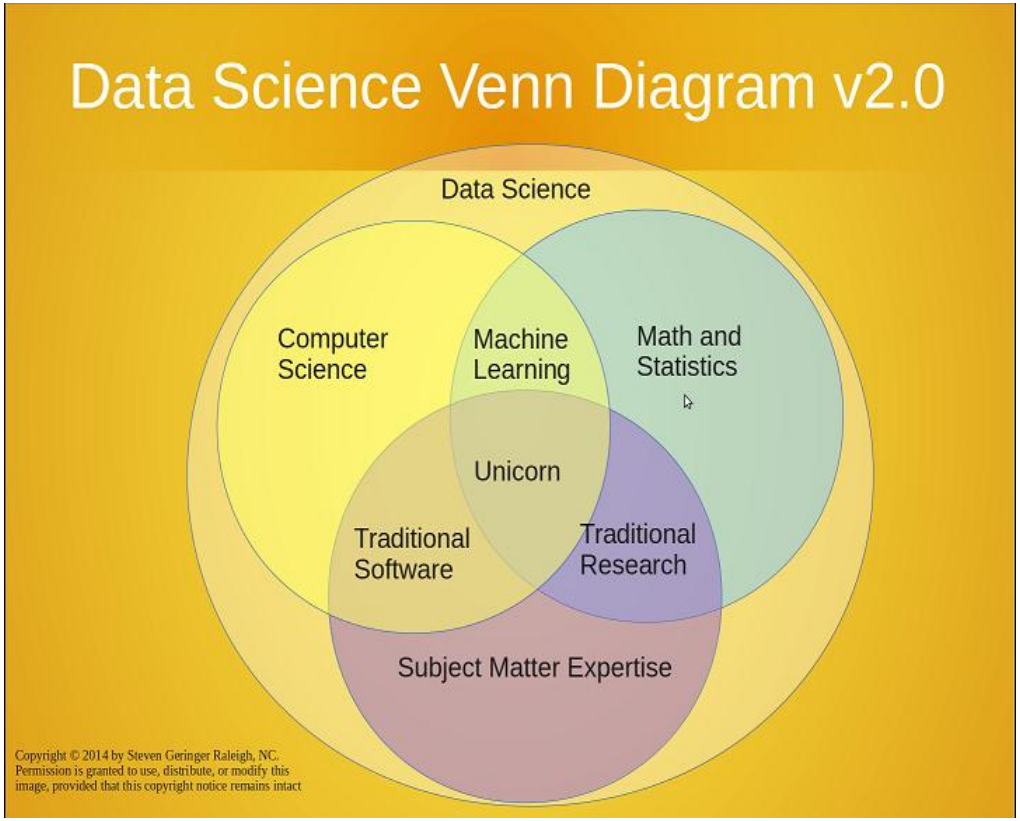
- Existing data
- Other data sources

- Item X sells more on Y day of the week in Z conditions:
  - Raspberries sell higher on a Saturday when the temperature is 20-25 C

What is Data Science?



What is Data Science?



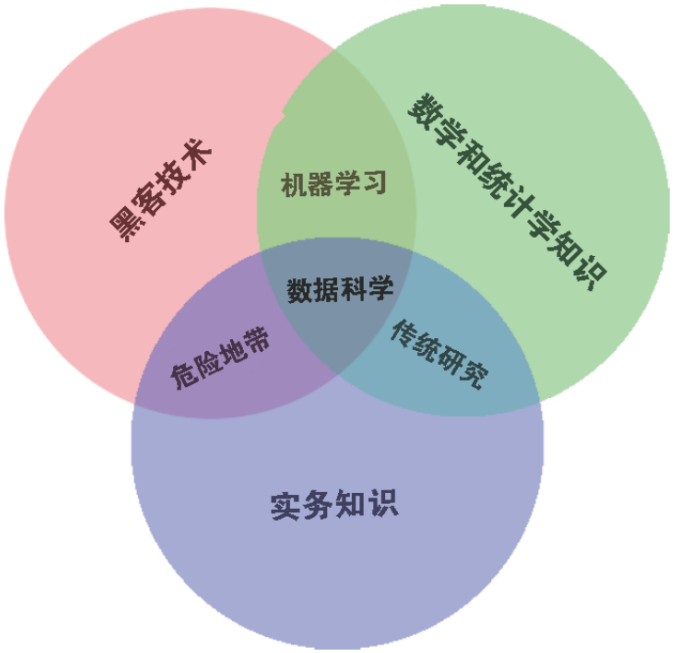
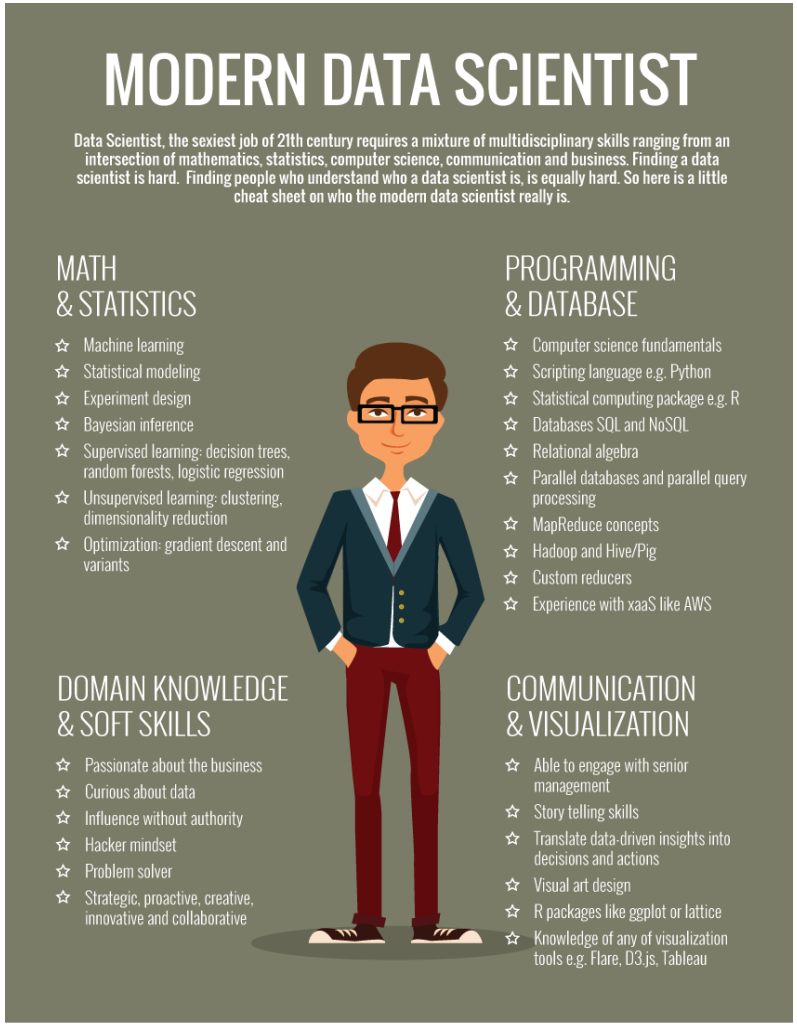


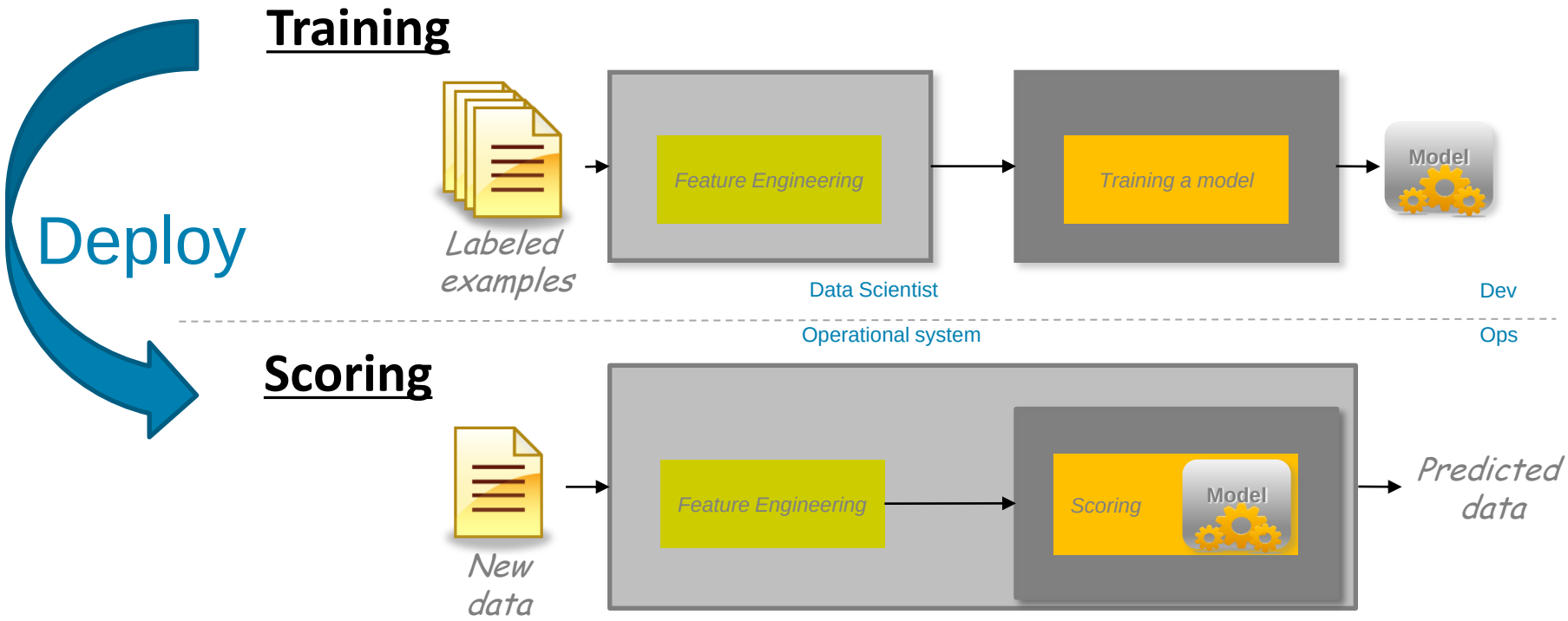
图1-1：Drew Conway的数据科学韦恩图

What is Data Science?



What is Machine Learning (机器学习概要)

a TrainOps (DevOps) story

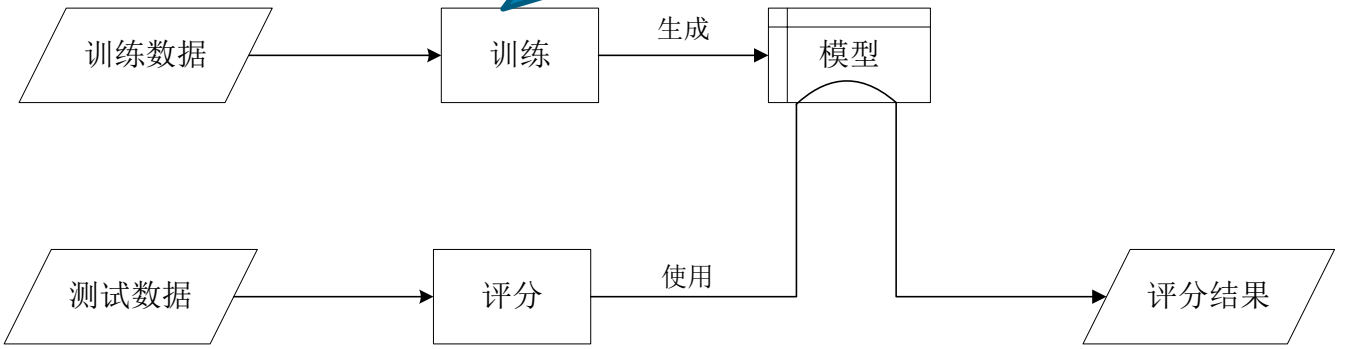


模型创建例子 – 挑西瓜

色泽	根蒂	敲声	好瓜？
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

使用机器学习算法：特征选择，决策树，Boosting，调参……

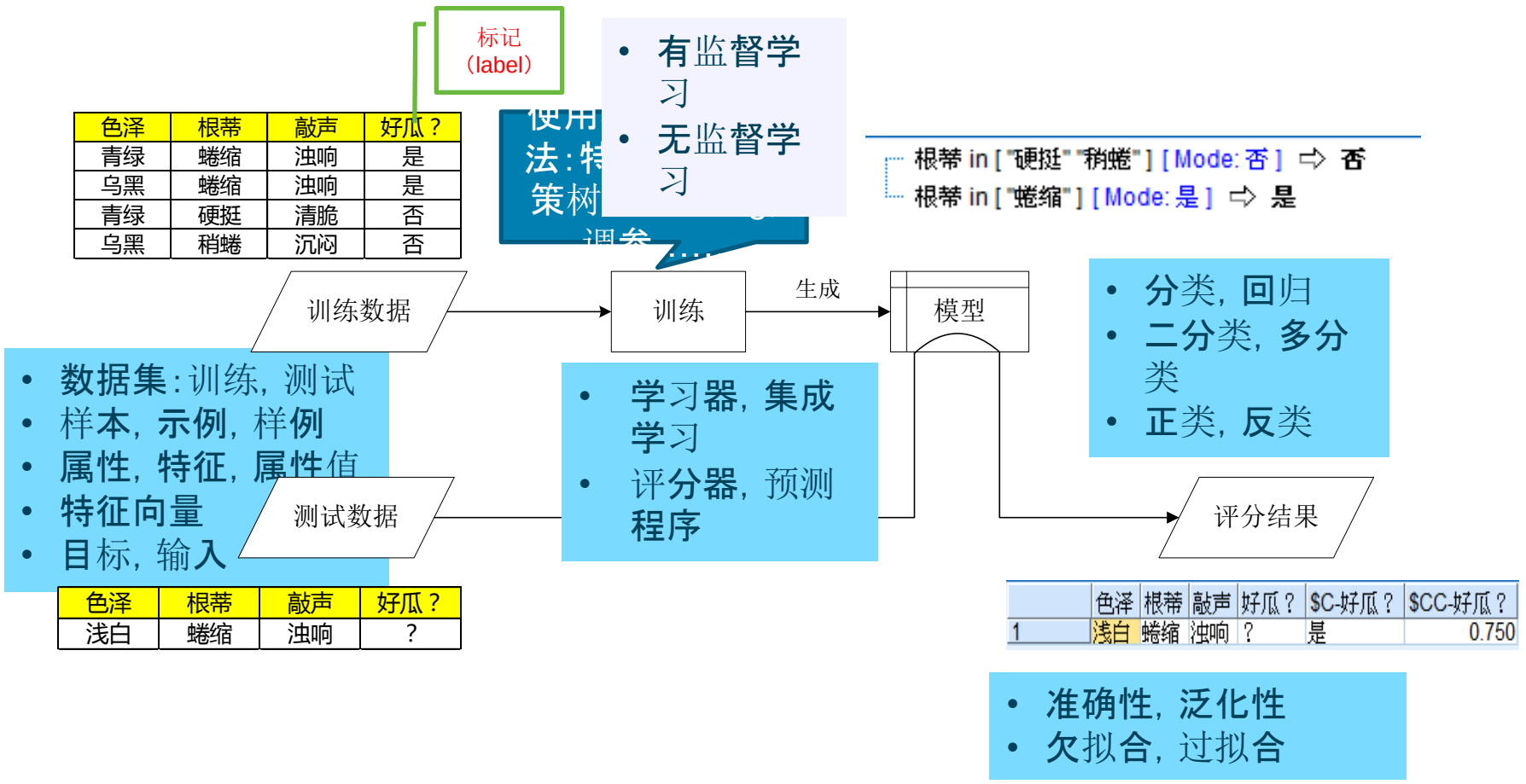
根蒂 in ["硬挺" "稍蜷"] [Mode: 否] ⇨ 否
根蒂 in ["蜷缩"] [Mode: 是] ⇨ 是



色泽	根蒂	敲声	好瓜？
浅白	蜷缩	浊响	？

	色泽	根蒂	敲声	好瓜？	$C-好瓜？	$CC-好瓜？
1	浅白	蜷缩	浊响	？	是	0.750

机器学习主要概念



机器学习算法的基本分类

有监督学习：从标签化训练数据集中推断出函数的机器学习任务。
- 线性模型
- 决策树
- 神经网络
- 支持向量机
- 贝叶斯分类器

无监督学习：设计分类器时候，用于处理未被分类标记的样本集。
- 聚类
- 任务与奖罚

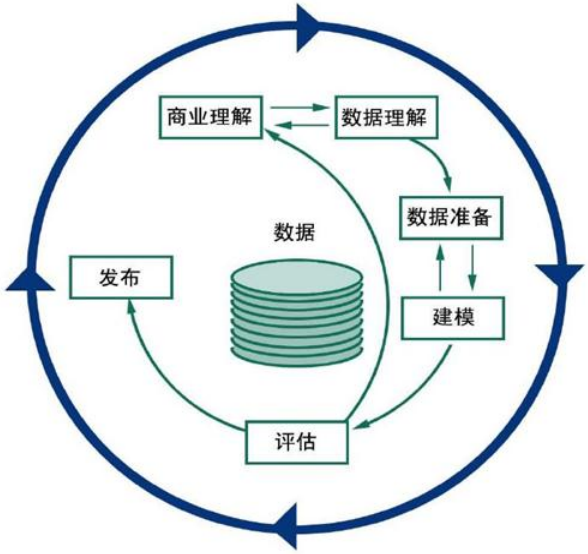
半监督学习：一部分有标记

- Linear regression, logistic regression, SVM
- Decision trees & ensembles
- Naïve Bayes
- Clustering – KMeans, Gaussian mixtures, PIC, etc
- Recommendation (ALS)
- Frequent pattern mining

数据科学和机器学习的方法学

数据科学各阶段任务重点

1. 业务理解
   - 确定业务目标
   - 评估资源和收益
   - 确定数据挖掘目标
   - 制定项目计划
2. 数据理解
   - 收集初始数据
   - 了解数据定义
   - 探索数据
   - 验证数据质量
3. 数据准备
   - 选择数据
   - 清理数据
   - 构建新数据
   - 集成数据
   - 格式化数据

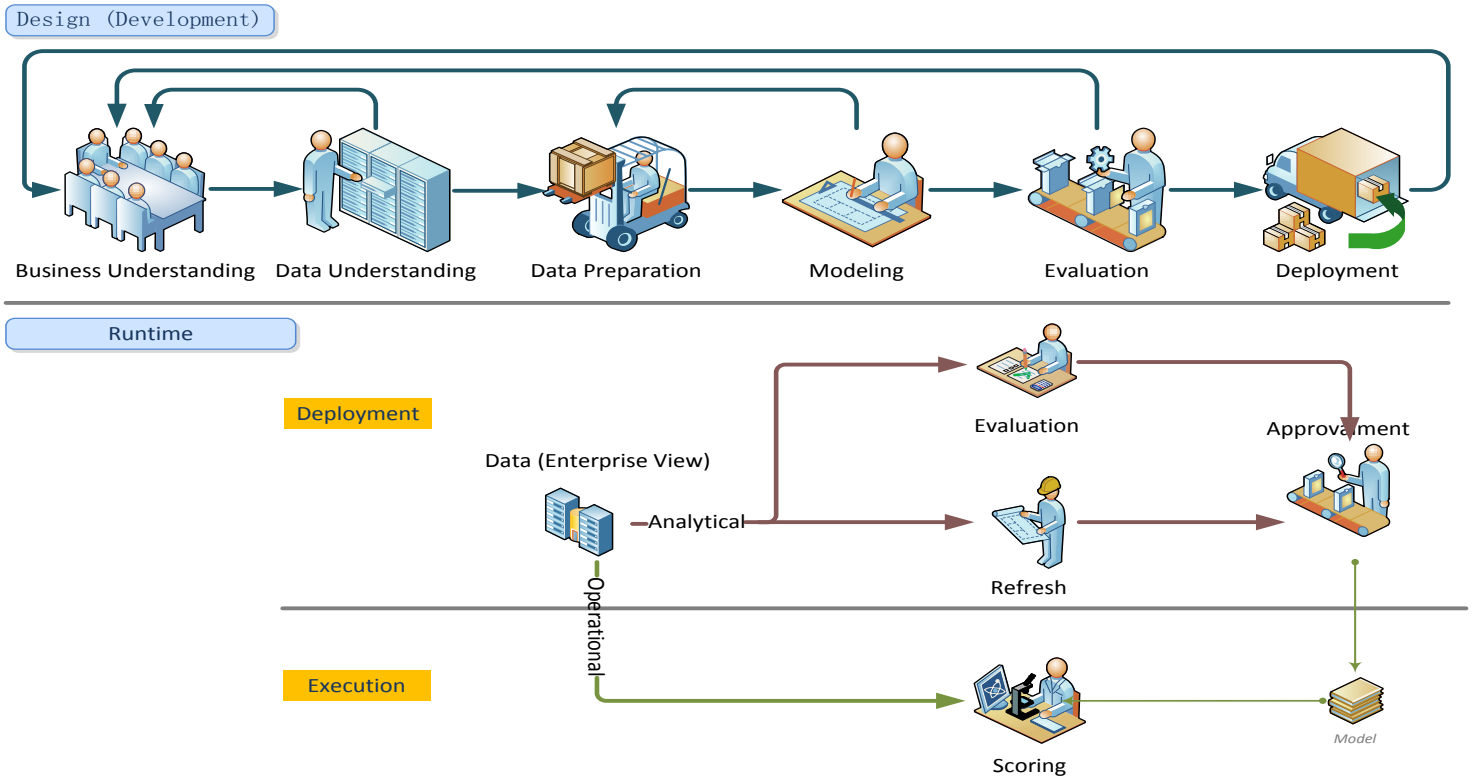


涉及到的人员：

- 业务决策者，业务专家
- 数据所有者， 业务系统专家
- 数据科学家， 数据工程师

4. 建模
   - 选择建模技术
   - 生成测试设计
   - 构建模型
   - 评估模型
5. 评估
   - 评估结果
   - 审核过程
6. 部署
   - 制定部署计划
   - 计划监视和维护
   - 生成最终报告
   - 执行最终项目审核

数据科学(机器学习模型)流程和角色



数据科学工具箱

- IBM Data Science Experience
- 数据科学体验

- IBM Machine Learning
- 机器学习平台

IBM Data Science Experience 数据科学体验

学习

提供各种级别的学习资料与路径

创建

提供最好的开源工具并结合IBM自有能力，为创建先进的数据产品提供支持

协作

社区与社交社交属性提供有效的协作

ALL YOUR TOOLS IN ONE PLACE

DSx主要特性

IBM Data Science Experience

社区

- 教材与数据集
- 连接数据科学家
- 提问
- 文章与论文
- 复制与分享项目

开源

- Scala/Python/R/SQL
- Jupyter and Zeppelin* Notebooks
- RStudio IDE and Shiny apps
- Apache Spark
- Your favorite libraries

IBM 提供的能力

- 数据预处理/Pipeline UI *
- 自动数据准备与建模*
- 高级可视化*
- 模型管理与部署
- 模型API文档*
- Spark云服务

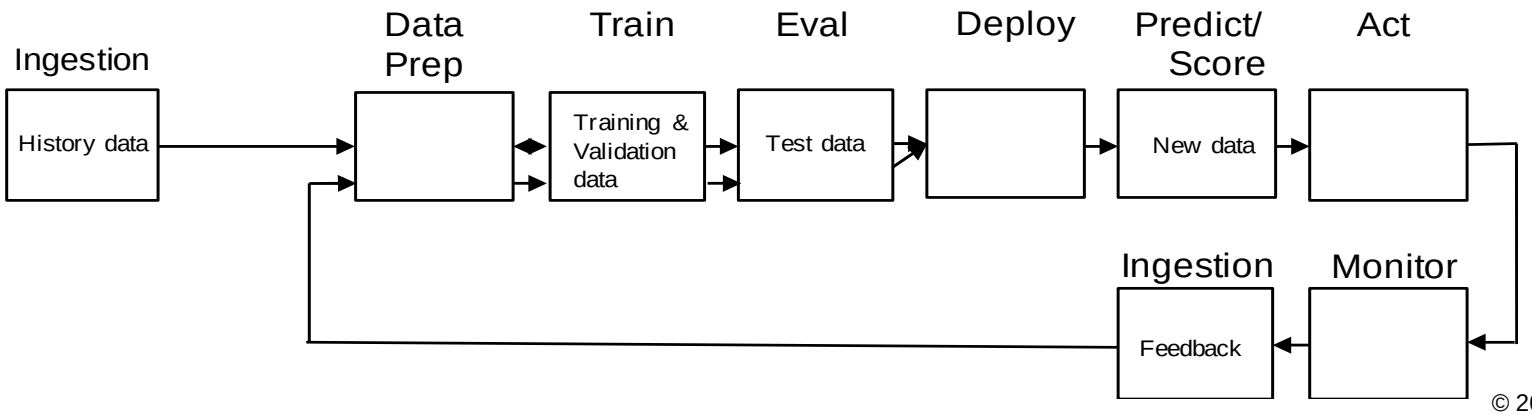
From Tic-Tac-Toe, Quiz Shows to Cancer Treatments

IBM - a pioneer of advanced Machine Learning (ML) solutions

Value - *Reduced business risks / costs and increased business opportunities*

Depth and Breadth - Covers full spectrum of analytics techniques required

- Machine Learning
  - Training infrastucture, Deployment infrastructure
- Optimization, Monitoring and Feedback Loop
- Collaborative, consumable, automated development environment.



持续智能：
机器学习让一切都简单了

在整个企业中融入持续智能

灵活：为您的业务选择正确的语言以及机器学习框架和平台。

高效：让经验丰富的数据科学家和新手都更富有成效。

可信：自信地部署各种洞察，因为知道它们是从最新的数据和趋势中生成的。

IBM Machine Learning for z/OS – 企业级机器学习平台

Notebook 和可视化建立模型

Cognitive Assistant for Data Scientists (CADS)

模型部署

模型管理

持续监控和反馈

IBM Machine Learning for z/OS 组件

案例一： 在线零售– 提升销量

- **业务问题：**

户外用品零售商的户外保护产品线销量下降， 希望能知道下降原因和如何提升销量

- **解决方案：**

分析销售数据，找到相关性并制定营销策略

- IBM Data Science Experience

案例二： “认知银行” – 了解你的客户

- **业务问题：**

我愿意提供优惠来防止客户流失，问题是我不知道谁会流失，过去都是流失之后我才知道的

- **解决方案：**

用机器学习找到现有已流失客户的特点，用来预测现有客户的流失可能性

- **知道之后应该怎么做：**

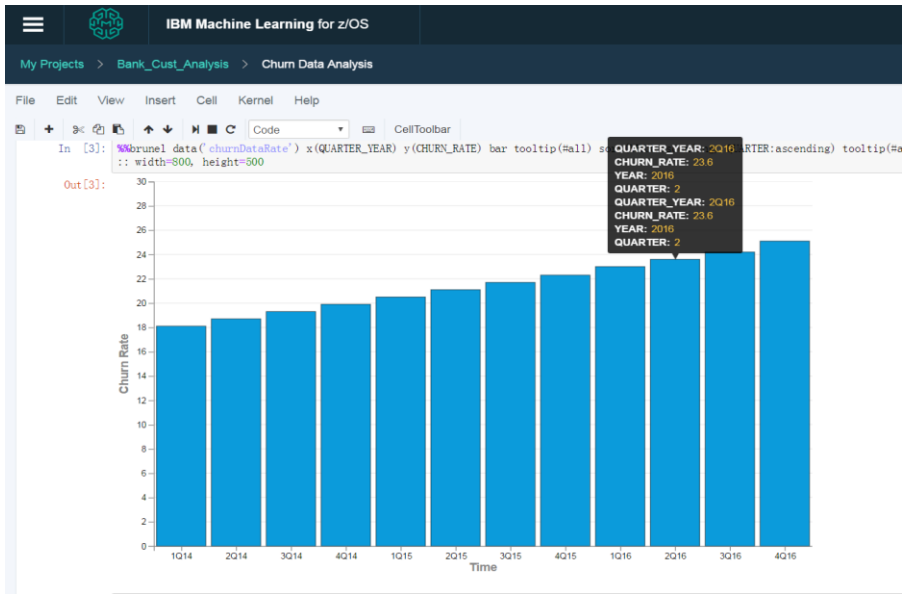
推荐系统之“认知银行”，以后介绍完整场景

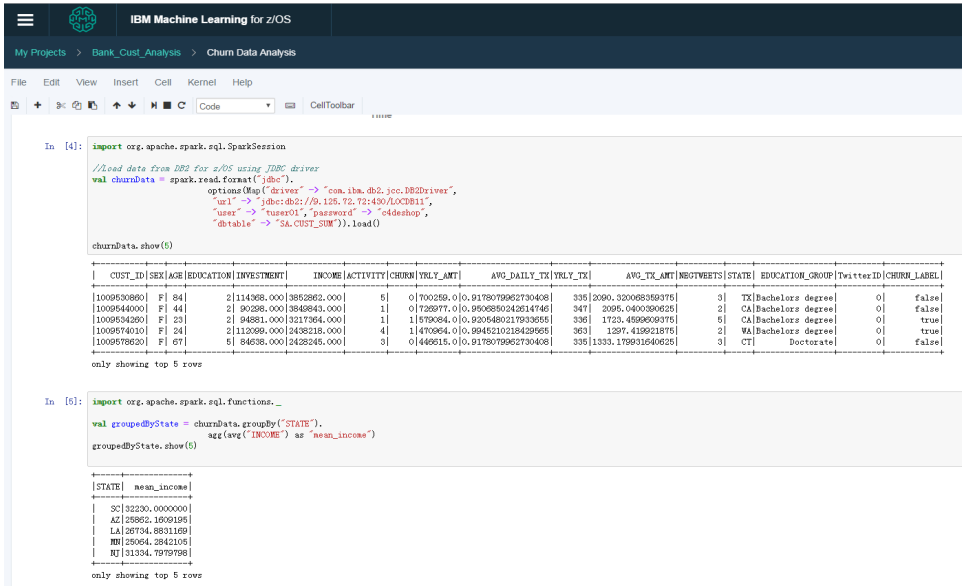
- IBM Machine Learning Platform

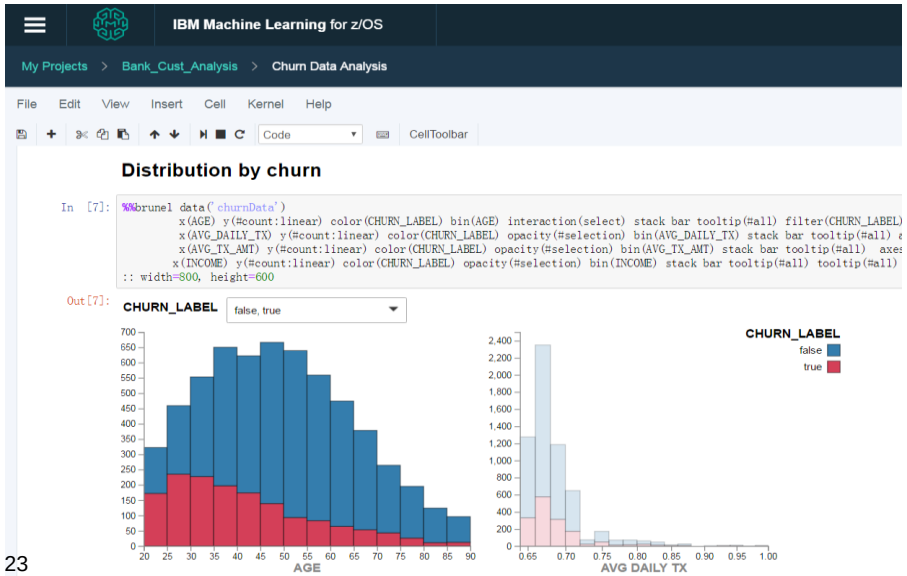
数据科学机器学习 Meetup 线下活动

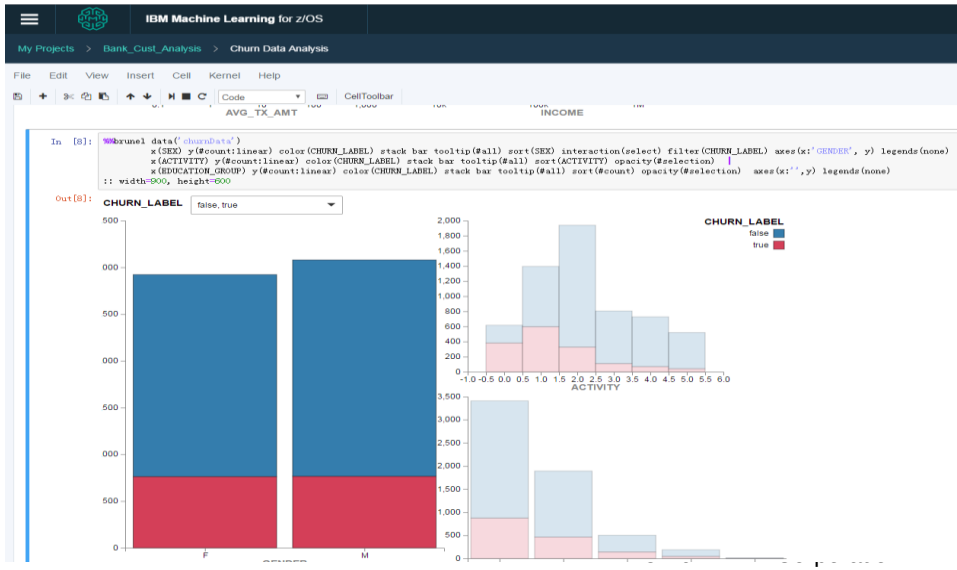
- 3月19号 Sunday 北京
- 技术分享+ 动手营
- 注册Data Science Experience
  - –http://datascience.ibm.com/registration/stepone
- 了解Spark ML, Scala, Python
- 了解Brunel 数据可视化包
  - –http://datascience.ibm.com/blog/brunel-interactive-visualization-in-jupyter-notebooks-2/
- 关注微信群获取最新信息

查看客户流失现状，数据列表和图表展示









数据获取和处理

IBM Machine Learning for z/OS

My Projects > Bank_Cust_Analysis > Churn Data Analysis

File Edit View Insert Cell Kernel Help

Code CellToolbar

In [4]:
```
import org.apache.spark.sql.SparkSession

//Load data from DB2 for z/OS using JDBC driver
val churnData = spark.read.format("jdbc").
                      options(Map("driver" -> "com.ibm.db2.jcc.DB2Driver",
                       "url" -> "jdbc:db2://9.125.72.72:430/LOCDB11",
                       "user" -> "tuser01","password" -> "c4deshop",
                       "dbtable" -> "SA.CUST_SUM")).load()

churnData.show(5)
```

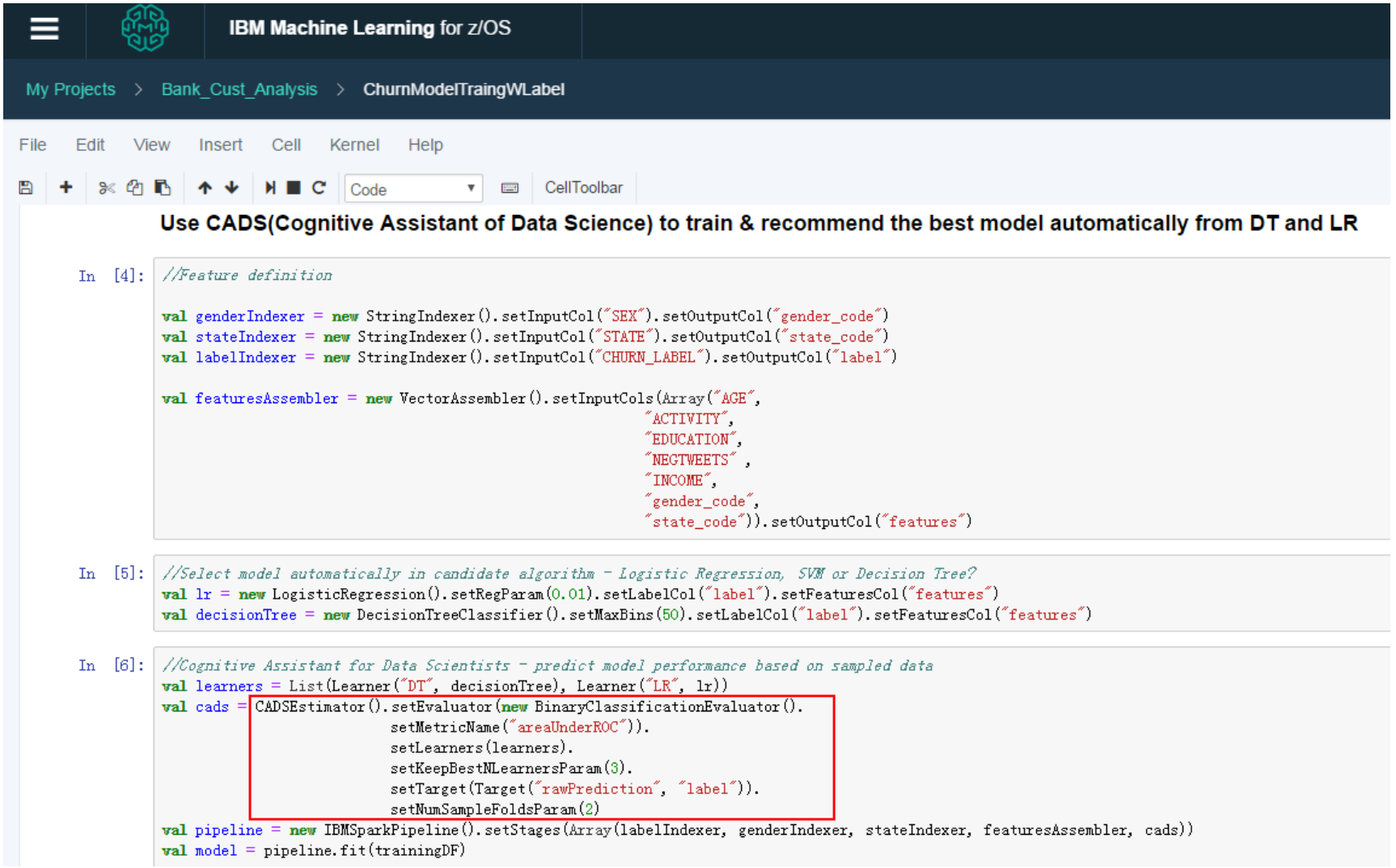
```
+----------+---+---+---------+----------+----------+--------+-----+--------+--------------------+--------+--------------------+---------+-----+----------------+---------+-------------+
|   CUST_ID|SEX|AGE|EDUCATION|INVESTMENT|    INCOME|ACTIVITY|CHURN|YRLY_AMT|       AVG_DAILY_TX|YRLY_TX|          AVG_TX_AMT|NEGTWEETS|STATE| EDUCATION_GROUP|TwitterID|CHURN_LABEL|
+----------+---+---+---------+----------+----------+--------+-----+--------+--------------------+--------+--------------------+---------+-----+----------------+---------+-------------+
|1009530860|  F| 84|        2|114368.000|3852862.000|       5|    0|700259.0|0.9178079962730408|     335|2090.320068359375|        3|   TX|Bachelors degree|        0|        false|
|1009544000|  F| 44|        2| 90298.000|3849843.000|       1|    0|726977.0|0.9506850242614746|     347|  2095.0400390625|        2|   CA|Bachelors degree|        0|        false|
|1009534260|  F| 23|        2| 94881.000|3217364.000|       1|    1|579084.0|0.9205480217933655|     336|  1723.4599609375|        5|   CA|Bachelors degree|        0|         true|
|1009574010|  F| 24|        2|112099.000|2438218.000|       4|    1|470964.0|0.9945210218429565|     363|   1297.419921875|        2|   WA|Bachelors degree|        0|         true|
|1009578620|  F| 67|        5| 84638.000|2428245.000|       3|    0|446615.0|0.9178079962730408|     335|1333.179931640625|        3|   CT|       Doctorate|        0|        false|
+----------+---+---+---------+----------+----------+--------+-----+--------+--------------------+--------+--------------------+---------+-----+----------------+---------+-------------+
only showing top 5 rows
```

In [5]:
```
import org.apache.spark.sql.functions._

val groupedByState = churnData.groupBy("STATE").
                      agg(avg("INCOME") as "mean_income")
groupedByState.show(5)
```

```
+-----+-------------+
|STATE|  mean_income|
+-----+-------------+
|   SC|32230.0000000|
|   AZ|25862.1609195|
|   LA|26734.8831169|
|   MN|25064.2842105|
|   NJ|31334.7979798|
+-----+-------------+
only showing top 5 rows
```

模型选择 – 数据科学家的认知助手



模型评估： 按目标类型对模型的性能指标进行计算和显示

Evaluate the trained model and draw the ROC curve

In [1]:
```
import com.ibm.analytics.ngp.pipeline.evaluate._
import com.ibm.analytics.ngp.pipeline.evaluate.JsonMetricsModel._
import spray.json._

val metrics = Evaluator.evaluateModel(MLProblemType.BinaryClassifier,model,testDF)

println(s"Binary Metric: ${metrics.asInstanceOf[BinaryClassificationMetricsModel].toJson}")

//Saving the model on file system if needed:
//model.saveToLocalPath("bfusr21/churnModel", "/home/bfusr21/model.tar.gz")

Connections.setEnvironment("dev")
Connections.setMetaServiceHost("http://9.30.166.110:12501")
model.save("steve/ChurnCADSModel")

println("Model saved successfully, you can view and deploy in the model management dashboard")
```

```
Binary Metric: {"recallByThreshold":[{"threshold":1.0,"metric":0.9018691588785047},{"threshold":0.0,"metric":1.0}],"precisionByThreshold"
R":0.9572567559904366,"fMeasureByThreshold":[{"threshold":1.0,"metric":0.9437652811735943},{"threshold":0.0,"metric":0.37842617152961977
reshold":1.0,"metric":1.0},{"threshold":1.0,"metric":1.0}],"areaUnderROC":0.9495121043325667}
Model saved successfully, you can view and deploy in the model management dashboard
```

In [8]:
```
val rocCurve = metrics.asInstanceOf[BinaryClassificationMetricsModel].roc.map{ case ThresholdMetricModel(x, y) => (x,y)}
```

In [9]:
```
val rocDF = spark.createDataFrame(rocCurve).
                  withColumnRenamed("_1", "FPR").
                  withColumnRenamed("_2", "TPR")
rocDF.show(3)
```

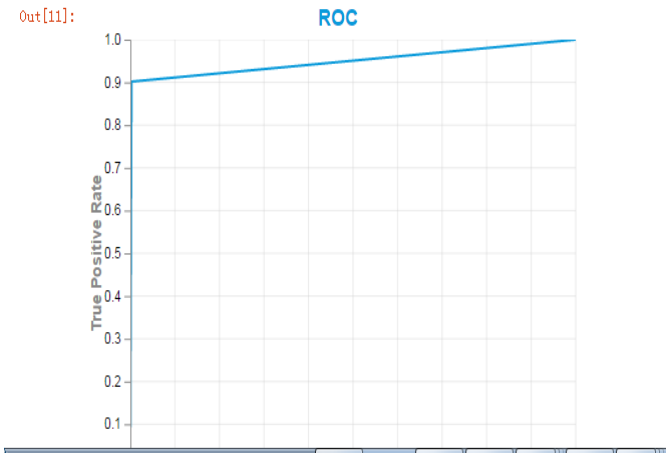
```
+--------------------+------------------+
|                 FPR|               TPR|
+--------------------+------------------+
|                 0.0|               0.0|
|0.002844950213371266|0.9018691588785047|
|                 1.0|               1.0|
+--------------------+------------------+
only showing top 3 rows
```

In [10]:
```
%AddJar -magic https://brunelvis.org/jar/spark-kernel-brunel-all-2.3.jar -f
```

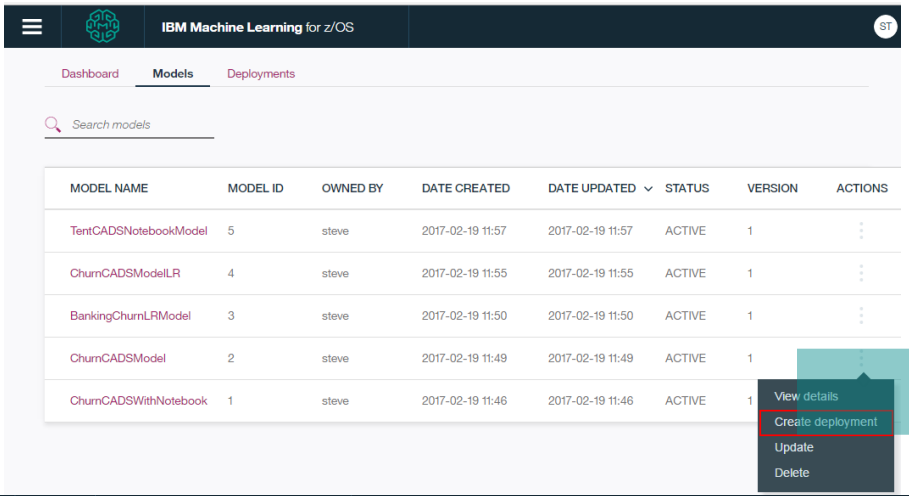
```
Starting download from https://brunelvis.org/jar/spark-kernel-brunel-all-2.3.jar
Finished download of spark-kernel-brunel-all-2.3.jar
```

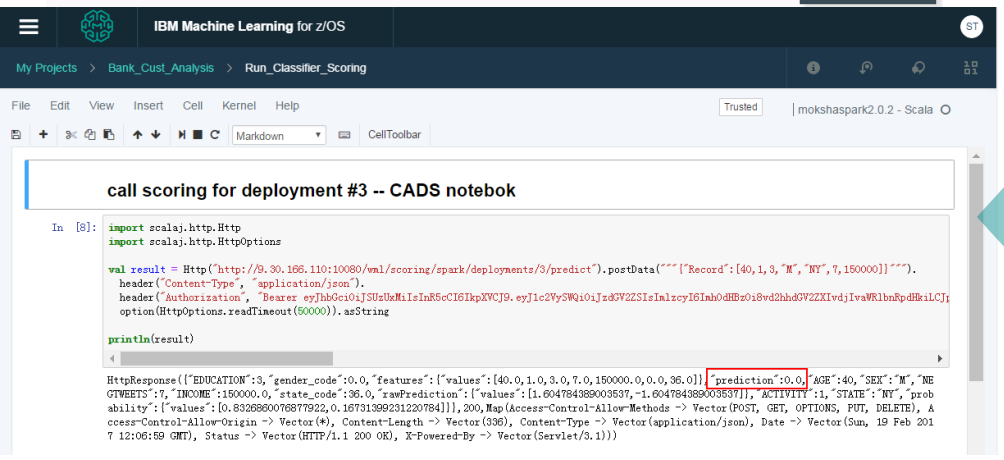
In [11]:
```
%%brunel data('rocDF') x(FPR) y(TPR) line tooltip(#all) axes(x:'False Positive Rat
```

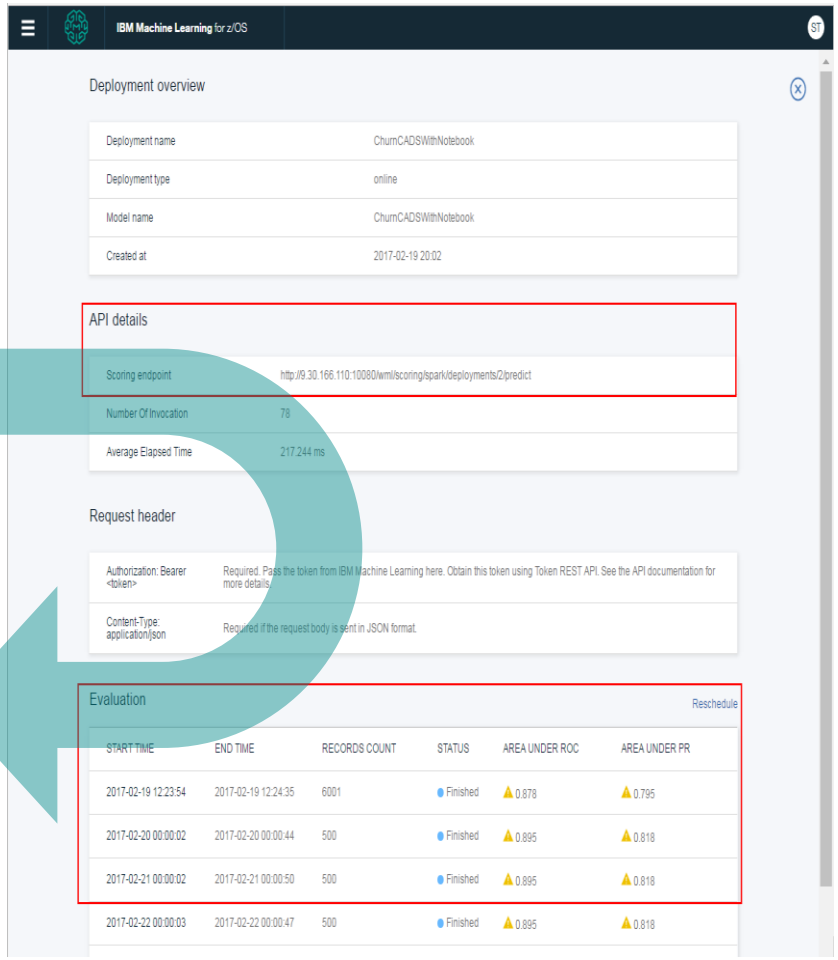
Out[11]:



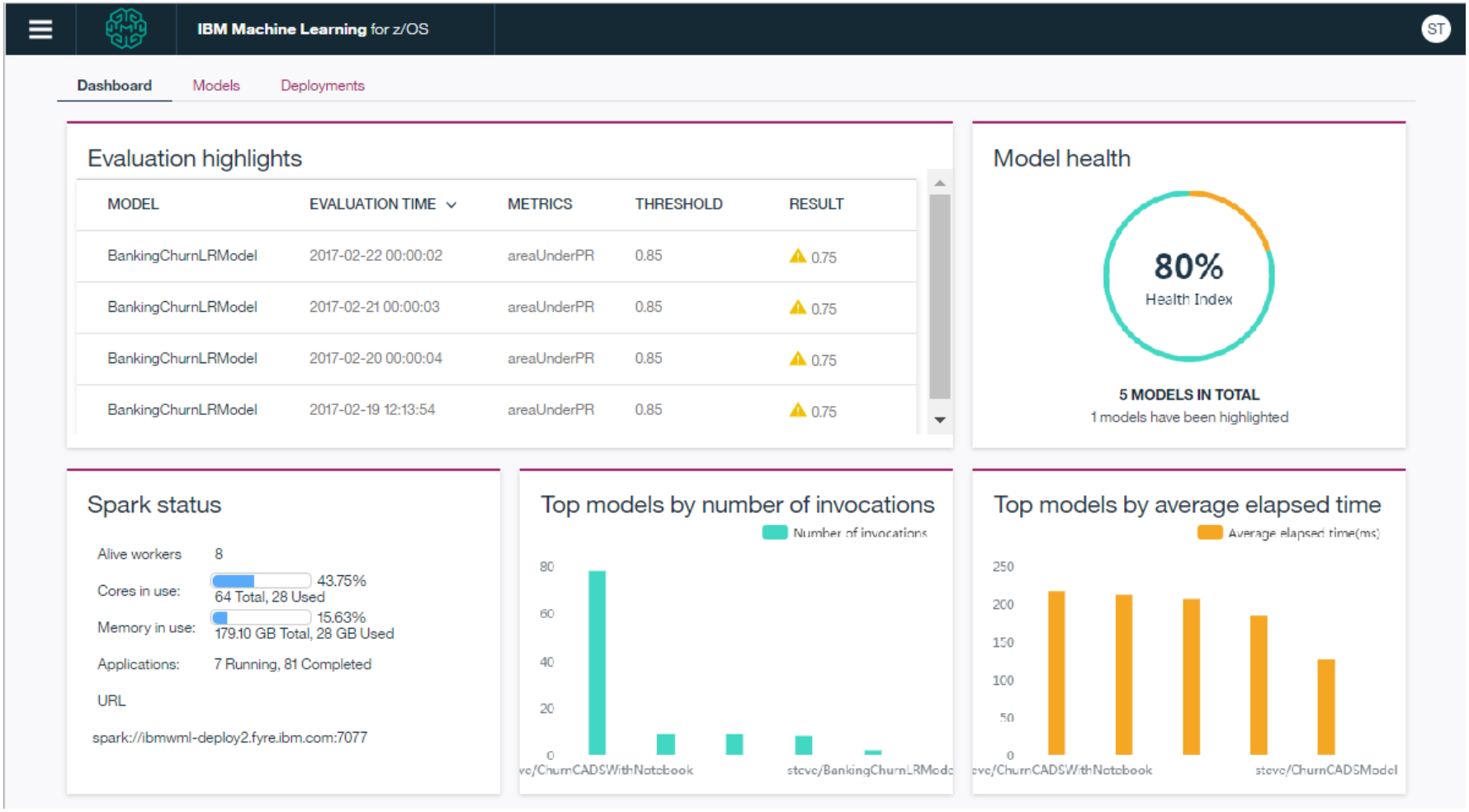
模型一键发布，提供API接口可供应用实时调用







模型运行监控：实时显示系统和模型的健康情况



机器学习产品好处：加速数据科学家端到端的工作

- 机器学习平台部署在云上，方便数据和计算资源企业内共享
- 方便数据获取、探索、预处理
- 支持数据科学家选择喜欢的工作方式：编码，流程图等
- 集成SparkML等多种开源算法
- 辅助评估多种模型、特征、参数，选择最佳组合
- 基于Spark平台，支持分大数据布式计算
- 一键部署，直接提供API可以快捷地集成到业务系统里
- 方便企业管理成千上万个模型，持续监控性能

机器学习产品来救场

- 核心算法改用IBM机器学习产品来支持，辅助选择模型和特征，快速建立模型
- 基于大数据量的基础上直接训练和预测
- 调整业务假设，合并模型结果
  - 上次成功转化的客户列表
  - 目前官网客户列表

数据量

硬件

算法

最佳实践

- 成熟的业务优化逻辑+灵活的算法实现+加入系统业务流程
  - 与业务专家一起找到可以利用机器学习智能优化的流程节点
  - 选择合适的机器学习平台实现算法
  - 将智能计算和预测结果无缝嵌入现有系统流程