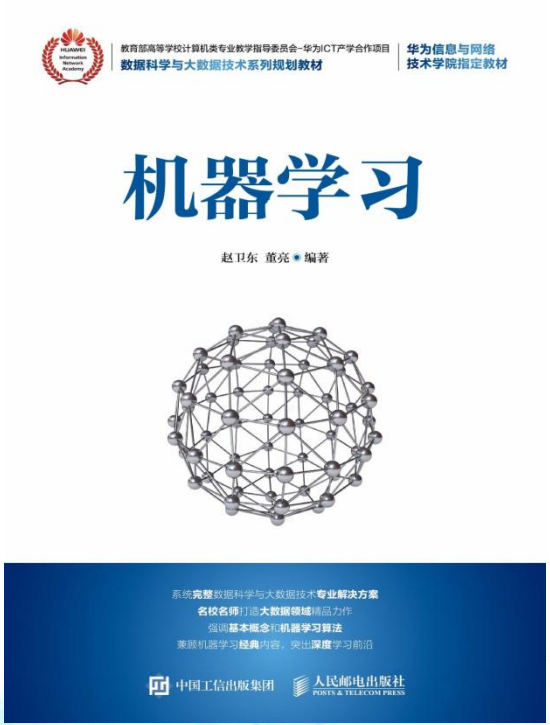


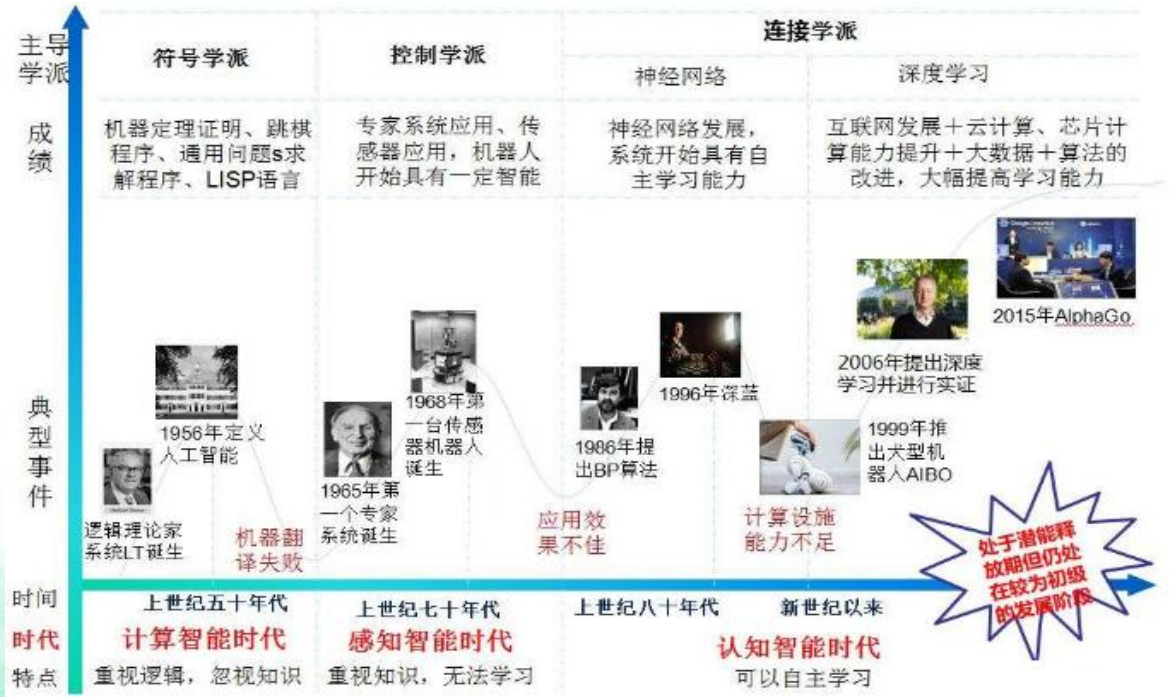
实践驱动的机器学习课程建设

赵卫东 复旦大学

wdzhao@fudan.edu.cn

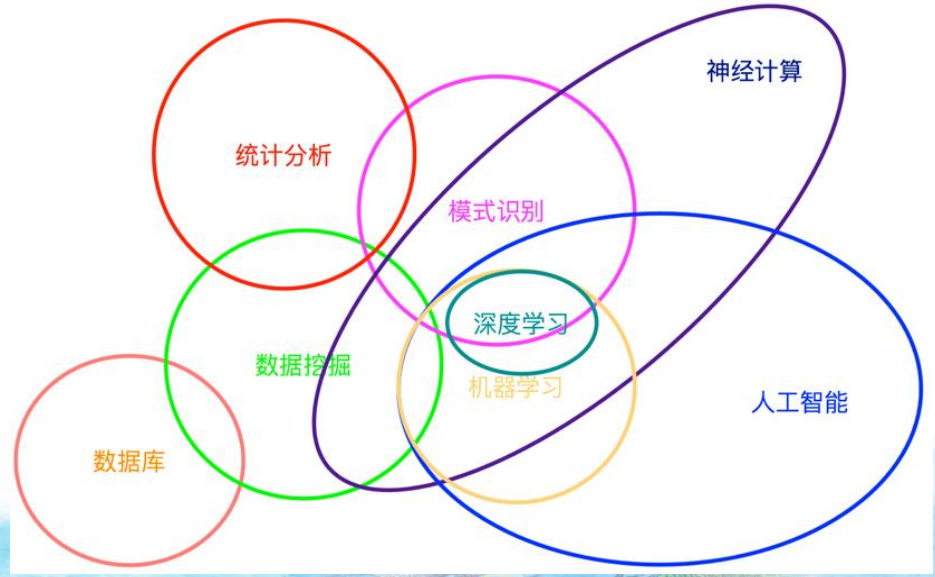
机器学习在人工智能发展中的作用

- 机器学习是人工（机器）智能的重要技术基础，也是智能科学的核心。
- 人工智能的4次浪潮基本与机器学习的方法有关。
- 人工智能的突破在于机器学习算法的进步。



机器学习的复杂性

- 机器学习主要的理论基础涉及到高等数学、线性代数、概率论与数理统计、数值逼近、最优化理论、数据库、计算复杂理论等，核心要素是数据、算法和模型。
- 机器学习的应用范围非常广泛，涉及各行各业。
- 机器学习技术的不成熟性（过高的期望）。



机器学习的演变（1）

- 机器学习的起源可以追溯到上世纪50年代早期人工智能的符号演算、逻辑推理、自动机模型、启发式搜索、模糊数学、专家系统、神经网络等，乃至最近的深度学习算法等。
- 强烈的概率逻辑和归纳逻辑，对大量、高质量数据的依赖过高。

机器学习阶段	年份	主要成果	代表人物
人工智能起源	1936	自动机模型理论	Alan Turing
	1943	MP模型	Warren McCulloch、Walter Pitts
	1951	符号演算	John von Neumann
	1950	逻辑主义	Claude Shannon
	1956	人工智能	John McCarthy、Marvin Minsky、Claude Shannon
人工智能初期	1958	LISP	John McCarthy
	1962	感知器收敛理论	Frank Roseblatt
	1972	通用问题求解(GPS)	Allen Newell、Herbert Simon
	1975	框架知识表示	Marvin Minsky
进化计算	1965	进化策略	Ingo Rechenberg
	1975	遗传算法	John Henry Holland
	1992	基因计算	John Koza
专家系统和知识工程	1965	模糊逻辑、模糊集	Lotfi Zadeh
	1969	DENDRA、MYCIN	Feigenbaum、Buchanan、Lederberg
	1979	ROSPECTOR	Duda

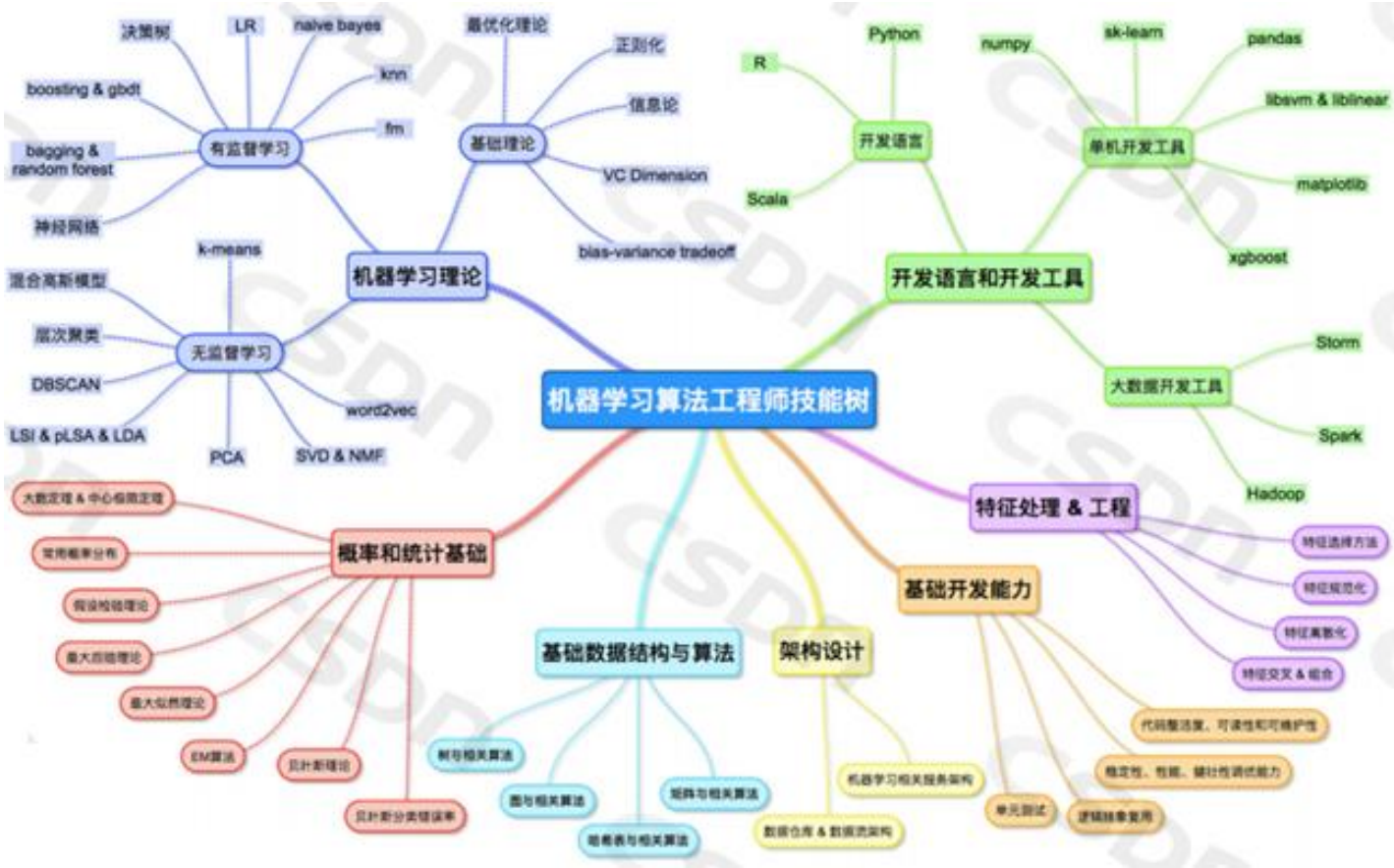
机器学习的演变（2）

神经网络	1982	Hopfield网络	Hopfield
	1982	自组织网络	Kohonen
	1986	BP算法	Rumelhart、McClelland
	1989	卷积神经网络	LeCun
	1998	LeNet	LeCun
	1997	循环神经网络RNN	Sepp Hochreiter、Jurgen Schmidhuber
分类算法	1986	决策树ID3算法	J. Ross Quinlan
	1988	Boosting算法	Freund、Michael Kearns
	1993	C4.5算法	J. Ross Quinlan
	1995	AdaBoost算法	Yoav Freund、Robert Schapire
	1995	支持向量机	Corinna Cortes、Vapnik
	2001	随机森林	Leo Breiman、Adele Cutler
深度学习	2006	深层神经网训练方法	Geoffrey Hinton
	2012	谷歌大脑	Andrew Ng
	2014	生成对抗网络GAN	Ian Goodfellow

机器学习人才培养的难题

- 数理要求高
- 学习成本高（算法、编程语言、平台、应用领域知识等）
- 跨学科综合能力
- 实践机会少

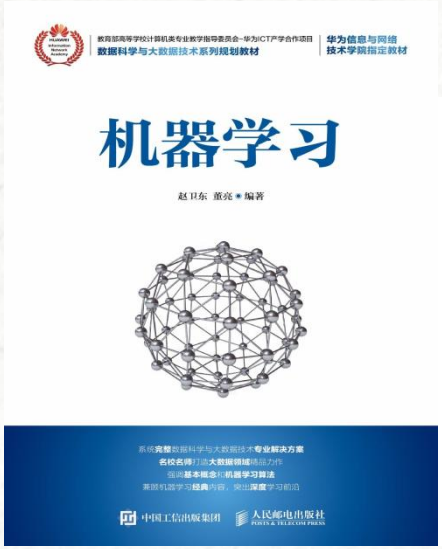
- Python
- R
- TensorFlow
- Caffe
- 开源社区Github

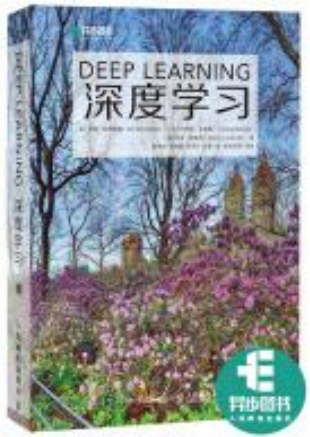


机器学习平台多样　选择困境？

《机器学习》教材的特点

- 《机器学习》教材相对国内主流教材的特点：
- 强调基本概念和常用机器学习算法，简明易懂，系统性强
- 兼顾机器学习经典内容，突出深度学习前沿
- 重视案例和实践教学





《机器学习》教材的内容（1）

机器学习

目 录 CONTENTS

第1章 机器学习概述 ····· 1
1.1 机器学习简介 ····· 2
1.1.1 机器学习简史 ····· 2
1.1.2 机器学习主要流派 ····· 3
1.2 机器学习、人工智能和数据挖掘 ····· 5
1.2.1 什么是人工智能 ····· 5
1.2.2 什么是数据挖掘 ····· 6
1.2.3 机器学习、人工智能与数据挖掘的关系 ····· 6
1.3 典型机器学习应用领域 ····· 7
1.4 机器学习算法 ····· 13
1.5 机器学习的一般流程 ····· 20
习题 ····· 21
第2章 机器学习基本方法 ····· 23
2.1 统计分析 ····· 24
2.1.1 统计基础 ····· 24
2.1.2 常见概率分布 ····· 29
2.1.3 参数估计 ····· 30
2.1.4 假设检验 ····· 32
2.1.5 线性回归 ····· 33
2.1.6 逻辑回归 ····· 35
2.1.7 判别分析 ····· 37
2.1.8 非线性模型 ····· 38
2.2 高维数据降维 ····· 39
2.2.1 主成分分析 ····· 39
2.2.2 奇异值分解 ····· 42
2.2.3 线性判别分析 ····· 43
2.2.4 局部线性嵌入 ····· 46
2.2.5 拉普拉斯特征映射 ····· 47
2.3 特征工程 ····· 49
2.3.1 特征构建 ····· 49
2.3.2 特征选择 ····· 50
2.3.3 特征提取 ····· 51
2.4 模型训练 ····· 51
2.4.1 模型训练常见术语 ····· 51
2.4.2 训练数据收集 ····· 51
2.5 可视化分析 ····· 52
2.5.1 可视化分析的作用 ····· 53
2.5.2 可视化分析方法 ····· 53
2.5.3 可视化分析常用工具 ····· 54
2.5.4 常见的可视化图表 ····· 56
2.5.5 可视化分析面临的挑战 ····· 66
习题 ····· 66
第3章 决策树与分类算法 ····· 68
3.1 决策树算法 ····· 69
3.1.1 分支处理 ····· 70
3.1.2 连续属性离散化 ····· 76
3.1.3 过拟合问题 ····· 78
3.1.4 分类效果评价 ····· 83
3.2 集成学习 ····· 87
3.2.1 装袋法 ····· 87
3.2.2 提升法 ····· 88
3.2.3 GBDT ····· 90
3.2.4 随机森林 ····· 91
3.3 决策树应用 ····· 93
习题 ····· 96
第4章 聚类分析 ····· 97
4.1 聚类分析概念 ····· 98
4.1.1 聚类方法分类 ····· 98
4.1.2 良好聚类算法的特征 ····· 99
4.2 聚类分析的度量 ····· 100
4.2.1 外部指标 ····· 100
4.2.2 内部指标 ····· 101
4.3 基于划分的聚类 ····· 103
4.3.1 k-均值算法 ····· 103
4.3.2 k-medoids 算法 ····· 108
4.3.3 k-prototype 算法 ····· 108
4.4 基于密度的聚类 ····· 109
4.4.1 DBSCAN 算法 ····· 109
4.4.2 OPTICS 算法 ····· 111
4.4.3 DENCLUE 算法 ····· 112
4.5 基于层次的聚类 ····· 115
4.5.1 BIRCH 聚类 ····· 115
4.5.2 CURE 算法 ····· 118
4.6 基于网格的聚类 ····· 121
4.7 基于模型的聚类 ····· 121
4.7.1 概率模型聚类 ····· 121
4.7.2 模糊聚类 ····· 126
4.7.3 Kohonen 神经网络聚类 ····· 126
习题 ····· 132
第5章 文本分析 ····· 134
5.1 文本分析介绍 ····· 135
5.2 文本特征提取及表示 ····· 135
5.2.1 TF-IDF ····· 136
5.2.2 信息增益 ····· 136
5.2.3 互信息 ····· 137
5.2.4 卡方统计量 ····· 138
5.2.5 词嵌入 ····· 138
5.2.6 语言模型 ····· 139
5.2.7 向量空间模型 ····· 141
5.3 知识图谱 ····· 142
5.3.1 知识图谱相关概念 ····· 143
5.3.2 知识图谱的存储 ····· 144
5.3.3 知识图谱挖掘与计算 ····· 145
5.3.4 知识图谱的构建过程 ····· 146
5.4 词法分析 ····· 151
5.4.1 文本分词 ····· 151
5.4.2 命名实体识别 ····· 154
5.4.3 词义消歧 ····· 155
5.5 句法分析 ····· 155
5.6 语义分析 ····· 157
5.7 文本分析应用 ····· 158
5.7.1 文本分类 ····· 159
5.7.2 信息抽取 ····· 161
5.7.3 问答系统 ····· 162
5.7.4 情感分析 ····· 163
5.7.5 自动摘要 ····· 164
习题 ····· 165
第6章 神经网络 ····· 166
6.1 神经网络介绍 ····· 167
6.1.1 前馈神经网络 ····· 167
6.1.2 反馈神经网络 ····· 169
6.1.3 自组织神经网络 ····· 172
6.2 神经网络相关概念 ····· 173
6.2.1 激活函数 ····· 173
6.2.2 损失函数 ····· 176
6.2.3 学习率 ····· 178
6.2.4 过拟合 ····· 180
6.2.5 模型训练中的问题 ····· 181
6.2.6 神经网络效果评价 ····· 184
6.3 神经网络应用 ····· 184
习题 ····· 188
第7章 贝叶斯网络 ····· 189
7.1 贝叶斯理论概述 ····· 190
7.2 贝叶斯概率基础 ····· 190
7.2.1 概率论 ····· 190
7.2.2 贝叶斯概率 ····· 191
7.3 朴素贝叶斯分类模型 ····· 192
7.4 贝叶斯网络推理 ····· 195
7.5 贝叶斯网络的应用 ····· 200
7.5.1 中文分词 ····· 200
7.5.2 机器翻译 ····· 201

《机器学习》教材的内容（1）

7.5.3 故障诊断 201
7.5.4 疾病诊断 202
习题 204

第 8 章 支持向量机 205

8.1 支持向量机模型 206
8.1.1 核函数 206
8.1.2 模型原理分析 207
8.2 支持向量机应用 210
习题 215

第 9 章 进化计算 216

9.1 遗传算法的基础 217
9.1.1 基因重组与基因突变 217
9.1.2 遗传算法实现技术 218
9.1.3 遗传算法应用案例 222
9.2 蚁群算法 223
9.3 蜂群算法 225
习题 227

第 10 章 分布式机器学习 229

10.1 分布式机器学习基础 230
10.1.1 参数服务器 230
10.1.2 分布式并行计算类型 231
10.2 分布式机器学习框架 232
10.3 并行决策树 238
10.4 并行 *k*-均值算法 238
习题 240

第 11 章 深度学习 242

11.1 卷积神经网络 243
11.1.1 卷积神经网络简介 243
11.1.2 卷积神经网络的结构 244
11.1.3 常见卷积神经网络 246
11.2 循环神经网络 254
11.2.1 RNN 基本原理 254
11.2.2 长短期记忆网络 260
11.2.3 门限循环单元 263
11.3 深度学习流行框架 264
习题 265

第 12 章 高级深度学习 266

12.1 高级卷积神经网络 267
12.1.1 目标检测与追踪 267
12.1.2 目标分割 270
12.2 高级循环神经网络应用 272
12.2.1 Encoder-Decoder 模型 272
12.2.2 注意力模型 273
12.2.3 LSTM 高级应用 274
12.3 无监督式深度学习 275
12.3.1 深度信念网络 275
12.3.2 生成对抗网络模型 277
12.4 强化学习 277
12.5 迁移学习 279
12.6 对偶学习 282
习题 283

第 13 章 推荐系统 284

13.1 推荐系统基础 285
13.1.1 推荐系统的应用场景 285
13.1.2 相似度计算 286
13.2 推荐系统通用模型 288
13.2.1 推荐系统结构 288
13.2.2 基于人口统计学的推荐 288
13.2.3 基于内容的推荐 289
13.2.4 基于协同过滤的推荐算法 290
13.2.5 基于图的模型 292
13.2.6 基于关联规则的推荐 293
13.2.7 基于知识的推荐 299
13.2.8 基于标签的推荐 300
13.3 推荐系统评测 301
13.3.1 评测方法 301
13.3.2 评测指标 302
13.4 推荐系统常见问题 306
13.5 推荐系统实例 309
习题 318

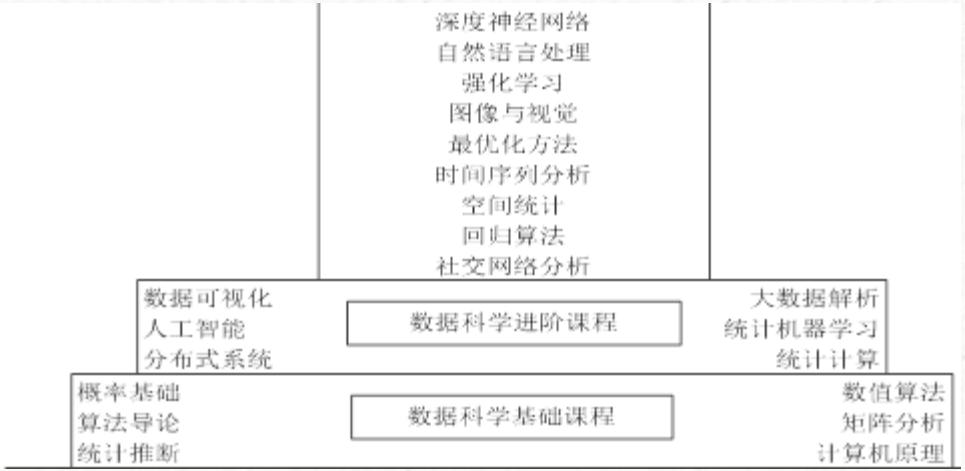
第 14 章 实验 319

14.1 华为 FusionInsight 产品平台介绍 320
14.2 银行定期存款业务预测 321
14.2.1 上传银行客户及存贷款数据 322
14.2.2 准备存款业务分析工作区 322
14.2.3 创建数据挖掘流程 323
14.2.4 定期存款业务模型保存和应用 330
14.3 客户分群 333
14.3.1 分析业务需求 333
14.3.2 上传客户信息数据 335
14.3.3 准备客户分群工作区 336
14.3.4 创建数据挖掘流程 337
14.3.5 客户分群模型保存和应用 344

参考文献 347

《机器学习》预修课程

- 扎实的高等数学、统计学、线性代数等数学基础
- 掌握Python编程语言
- 分布式计算理论（建议），针对大数据机器学习
- 相关应用领域的基本了解



《机器学习》实践教学

资讯类推荐系统实例

项目沉浸式教学的实施难点：

项目从哪里来？企业内部创新科创项目以及与企业的合作横向课题。

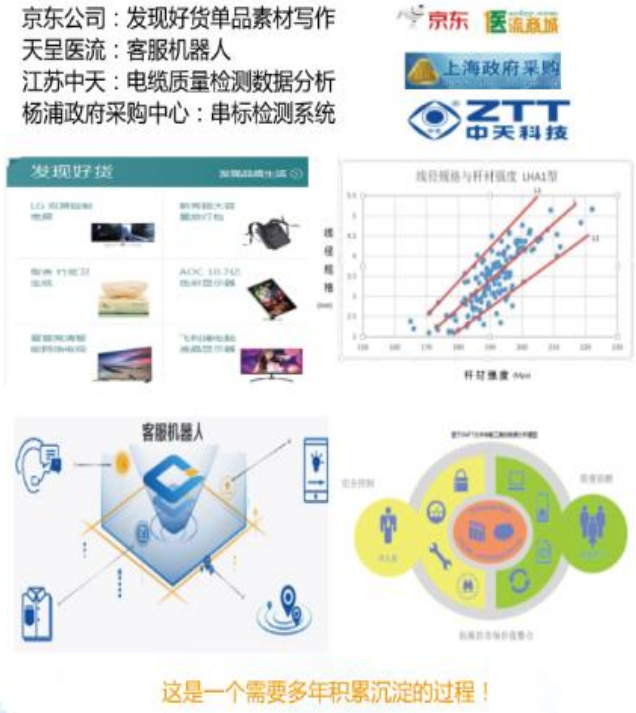
我们近2年与企业的典型合作项目

京东公司：发现好货单品素材写作

天呈医流：客服机器人

江苏中天：电缆质量检测数据分析

杨浦政府采购中心：串标检测系统

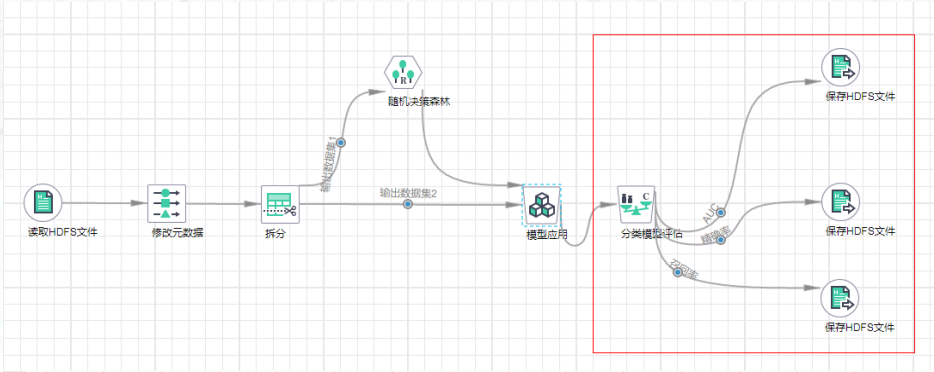


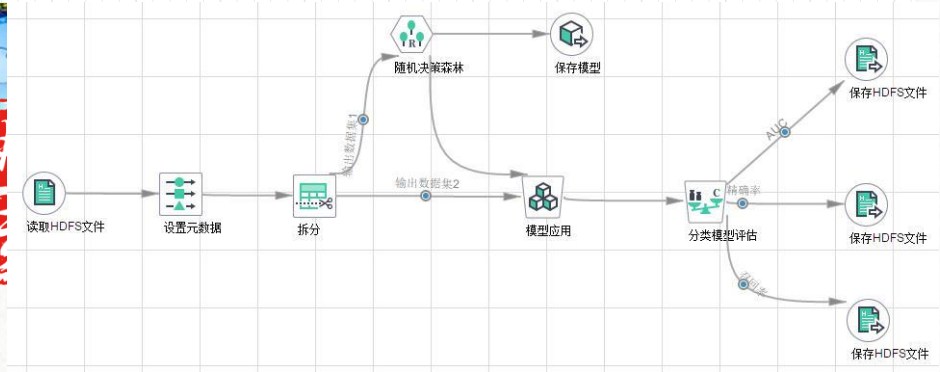
这是一个需要多年积累沉淀的过程！

特别致谢

- 基于华为MLS产品平台的实践过程，通过分析银行定期存款业务和对客户进行分群两个案例说明算法应用过程。其中银行存款业务分析中运用了逻辑回归、随机森林等算法，在客户分群的案例中主要基于聚类算法对客户进行分类。

华为FusionInsight产品平台






THANK
YOU