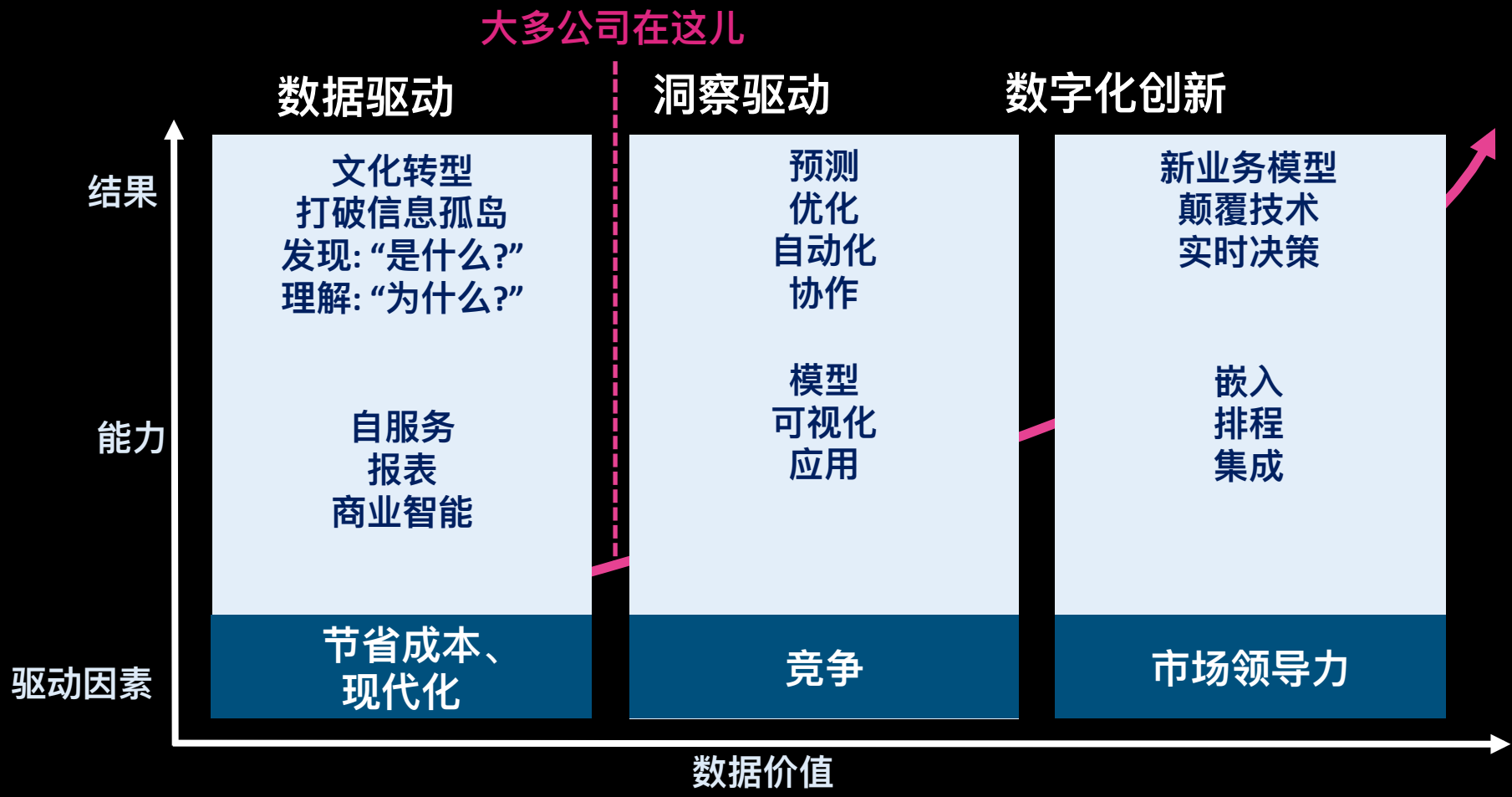
机器学习在企业级场景中的实践与探讨
Enterprise Machine Learning Practices

Jerry Zhu 朱辉
IBM机器学习全球研发总经理
IBM中国开发中心认知计算实验室总经理

IBM



提升数据价值的关键因素

访问数据

不同来源和结构的数据获取和存储、提供统一访问接口

治理数据

组织、管控数据以提升数据质量，安全性和可信性

获取洞察

从数据中快速准确的学习和获取洞察与智能，帮助业务转型

访问

混合数据管理

Write Once, Access Anywhere
with a common access layer to promote application independence

治理

统一的数据管控和集成

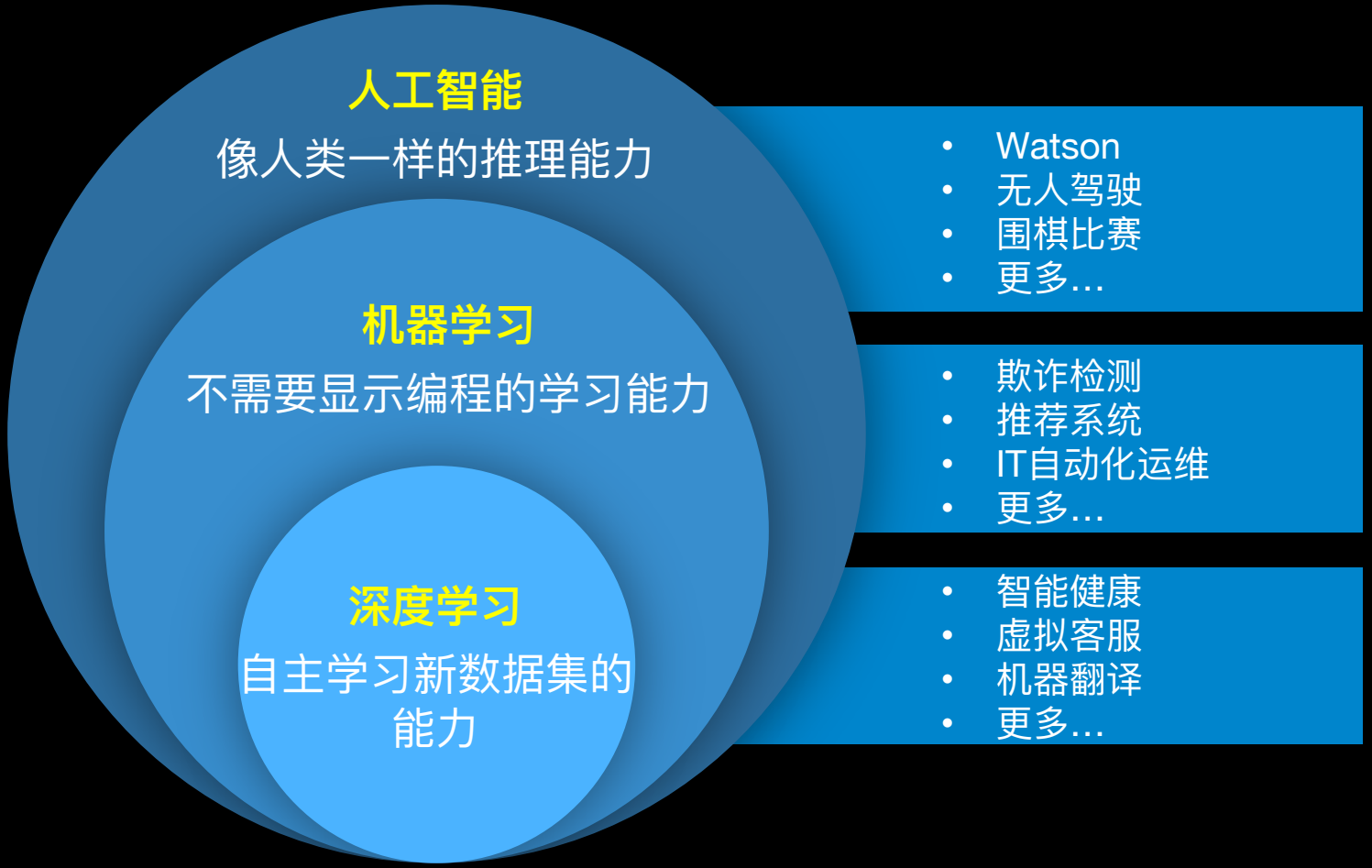
Prepare, Publish, Protect
your data to drive insights while mitigating compliance risks

分析

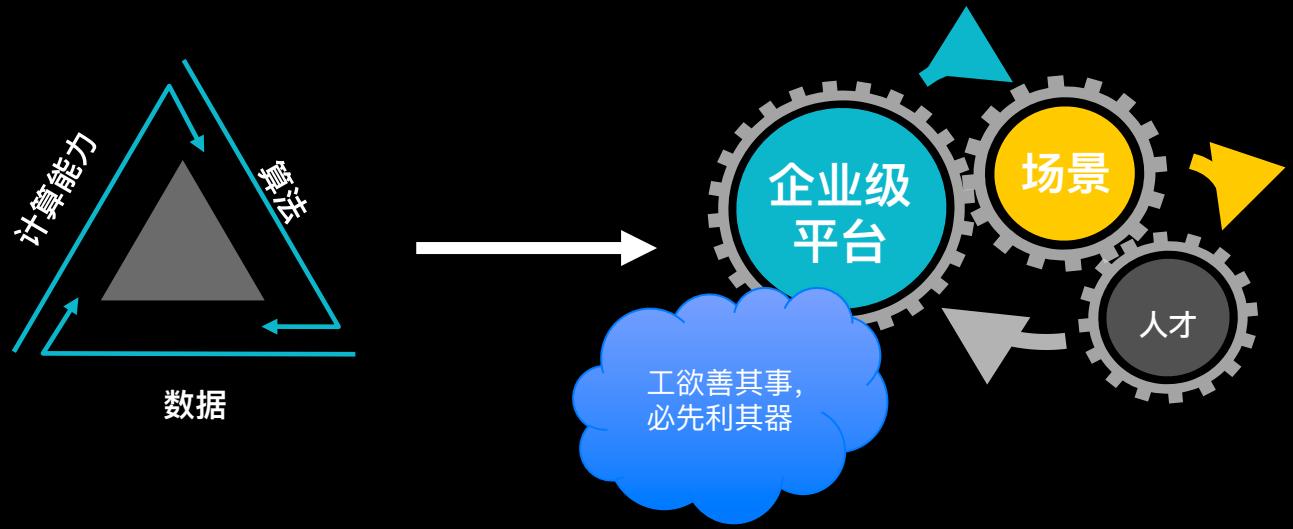
数据科学和认知商业

Describe, Predict, Prescribe
to understand the current, predict the future and change the outcome

融入 **机器学习**
无缝集成 **云上和云下**
基于 **开源技术**

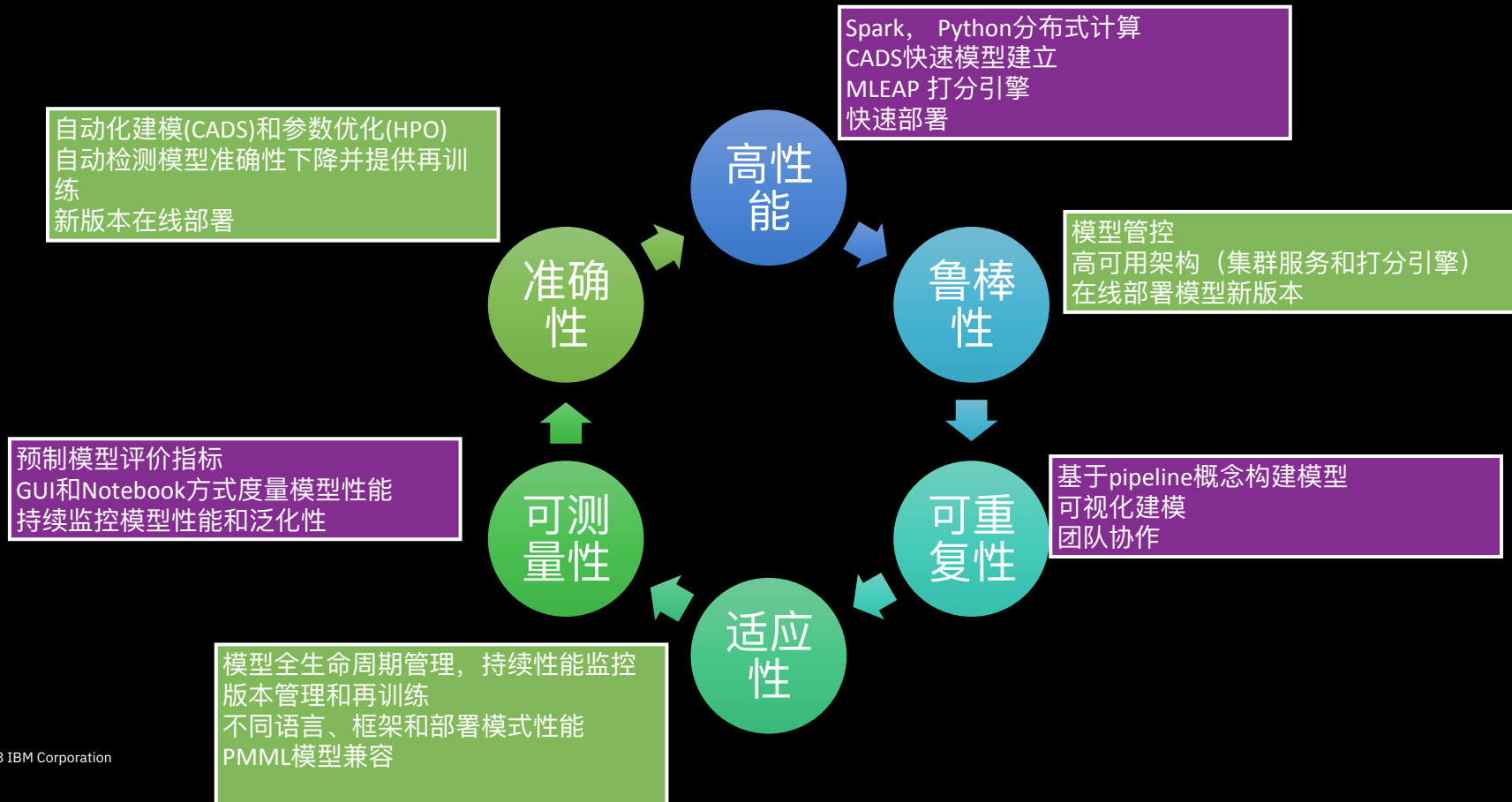


Ⅱ企业级机器学习的工程化挑战



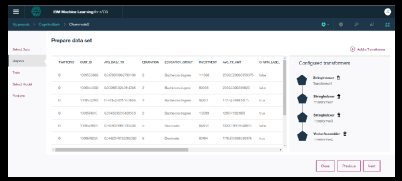
常规机器学习流程及挑战

企业级机器学习考量点 –
模型可操作性 (Model Operationalization)



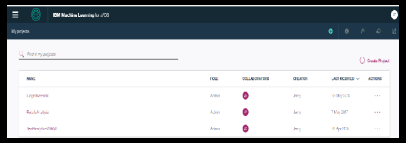
机器学习全生命周期管理与协作

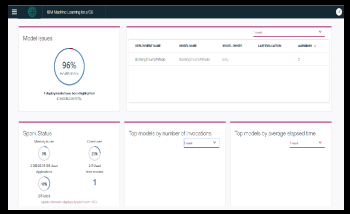
通过一个监控仪表盘管理企业内部所有部署的模型



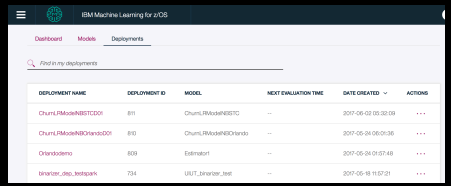
业务分析师

数据科学家





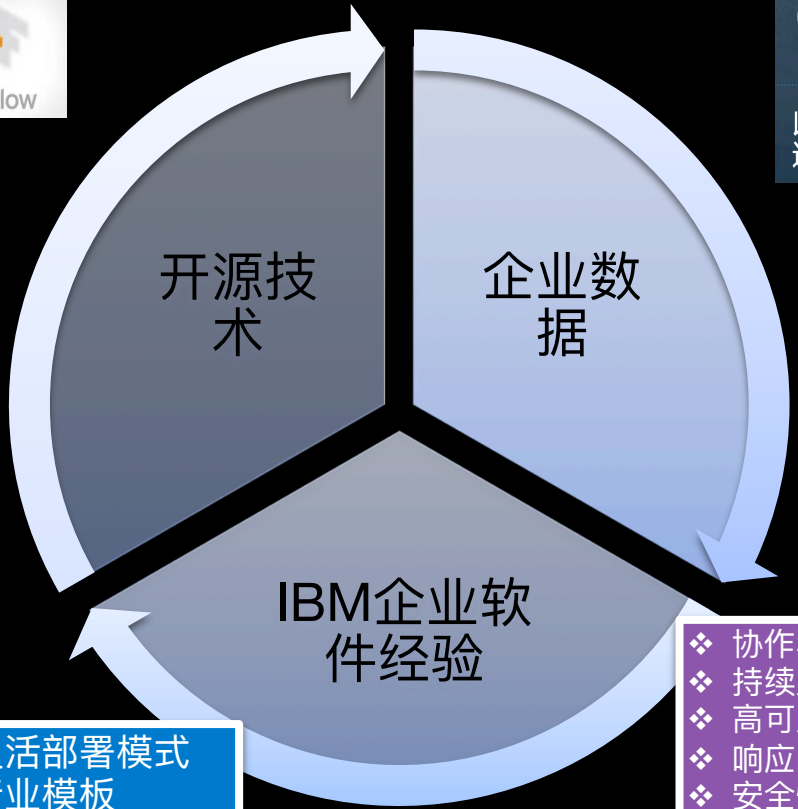
机器学习工程师



应用开发工程师

IBM 企业级机器学习平台

结合最新开源软件、企业数据和*IBM*特有的企业软件经验



全球 **90%** 以上的数据无法通过互联网获取

统一治理

覆盖所有数据的更佳洞察及合规

混合数据管理

统一数据及内容上云之路

- 协作和生命周期管理
- 持续监控和模型优化
- 高可用性
- 响应时间，高吞吐
- 安全性

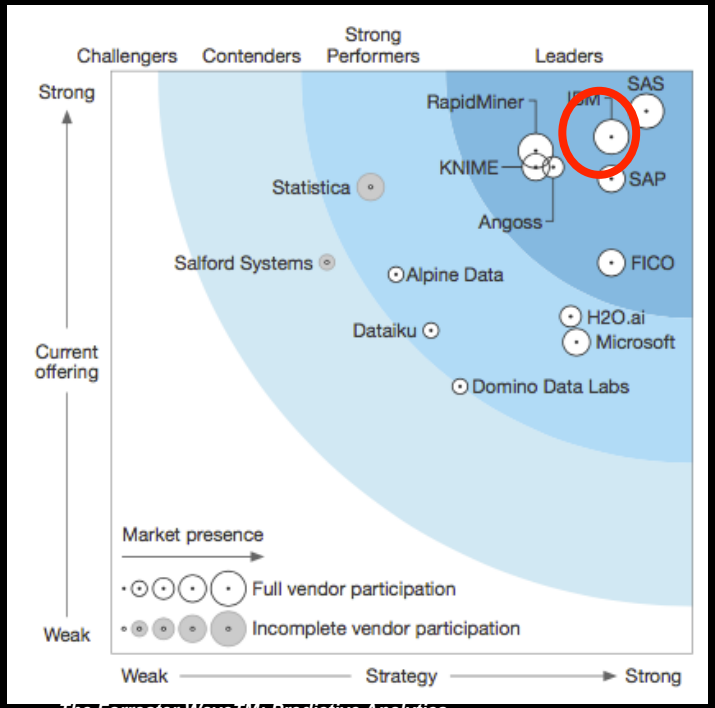
- 灵活部署模式
- 行业模板

NoSQL

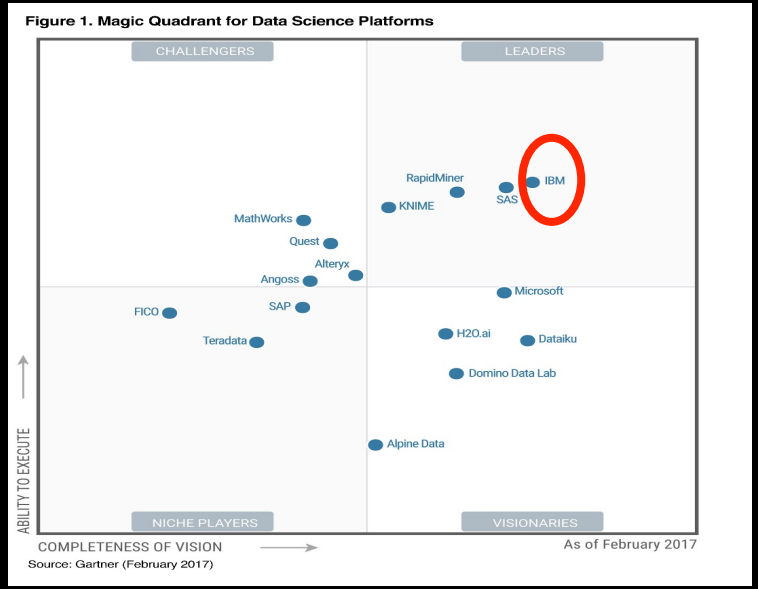
IBM 数据科学平台:
Gartner 和 Forrester 报告双料冠军

IBM被2017 Gartner 数据科学魔力象限评为最具远见和执行力的领导地位

.



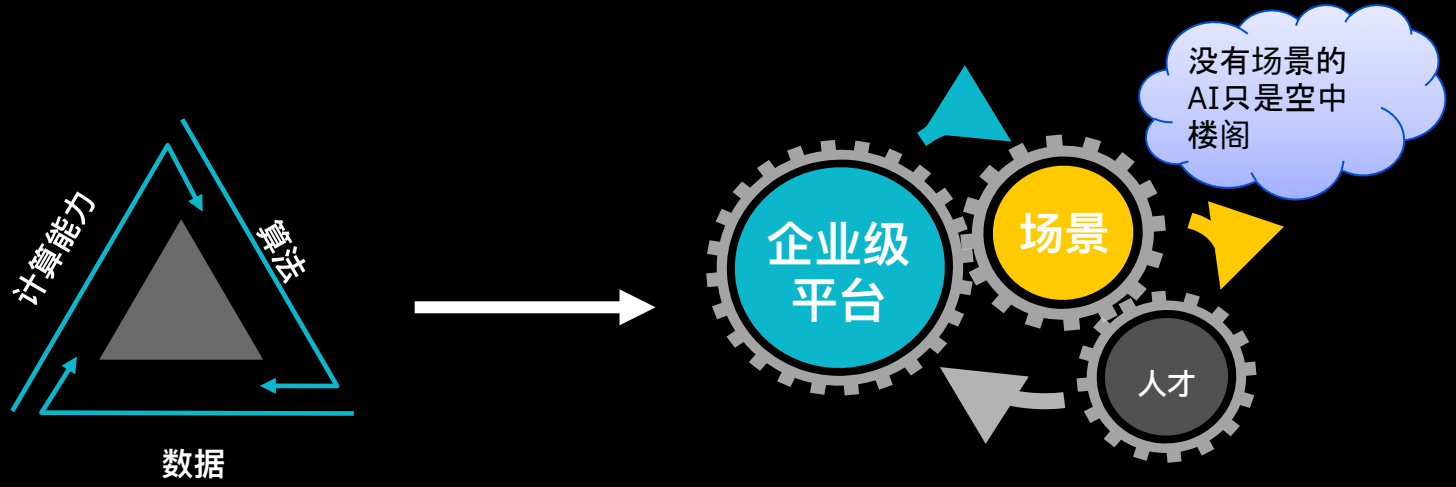
The Forrester Wave™: Predictive Analytics And Machine Learning Solutions, Q1 2017



Source: Gartner Magic Quadrant for Data Science Platforms,

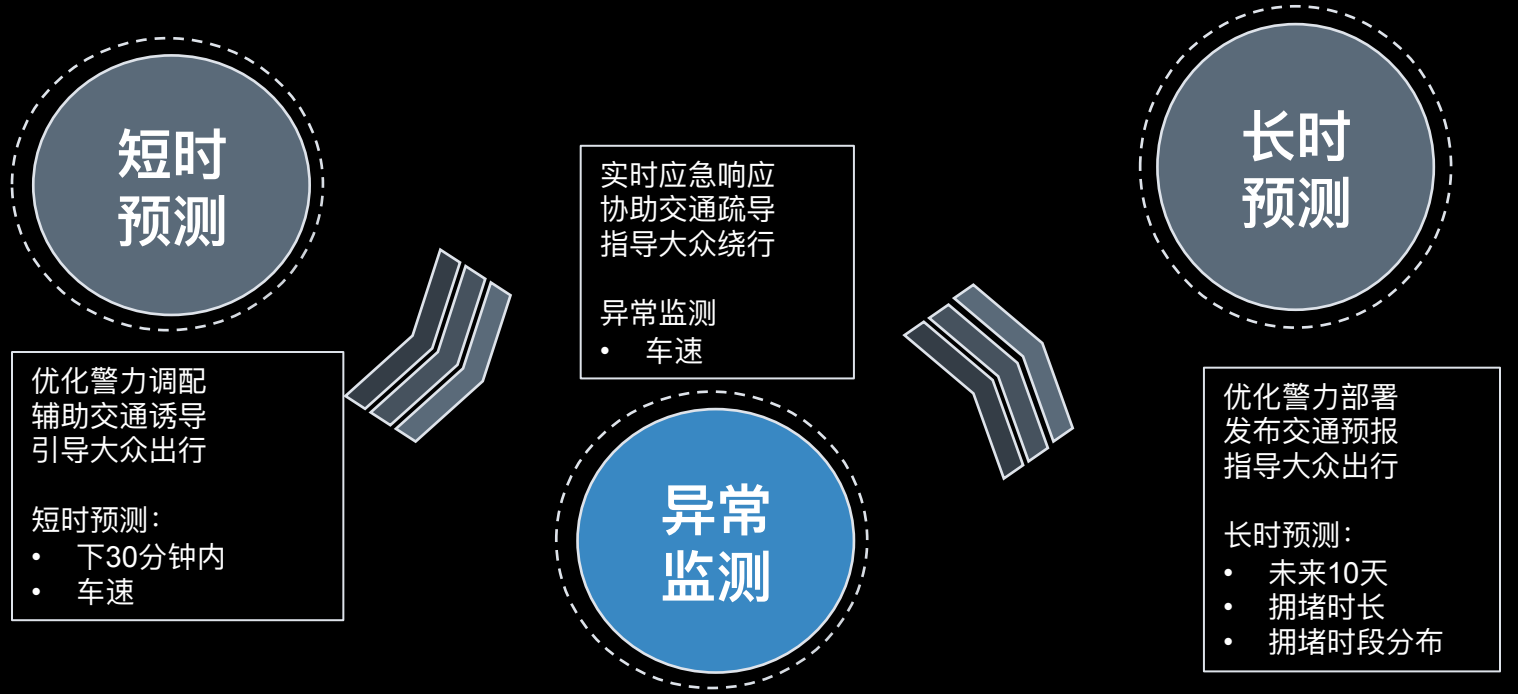
Forrester: IBM拥抱开源，发布的Data Science Experience 云服务能快速创建基于Spark集群的开源 Jupyter 和 Rstudio notebooks

Ⅱ企业级机器学习的工程化挑战



人工智能助力交通管控 (1)

人工智能助力交通管控 (2)

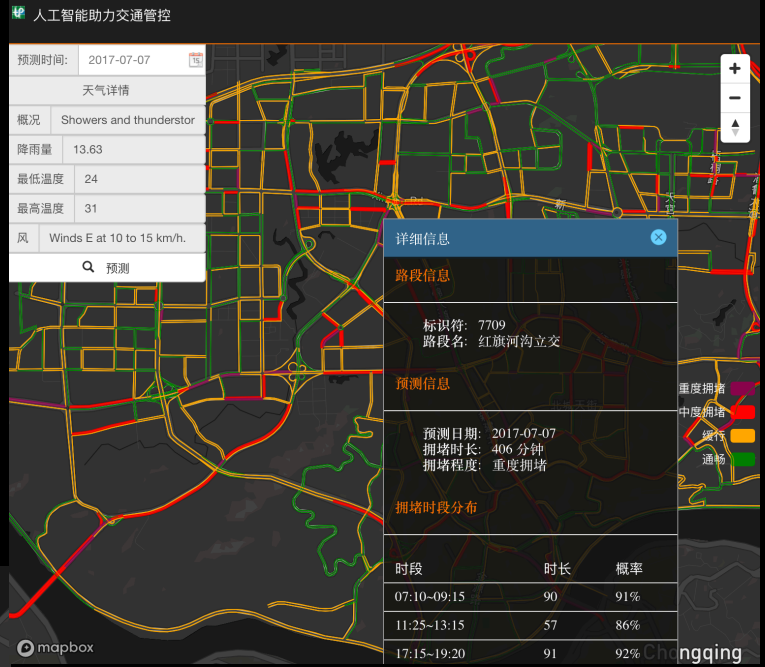


人工智能助力交通管控 (3)

异常监控预警 | 实时路况预报 | 7日路况预报 | 常发拥堵查询 | 拥堵趋势查询 | 拥堵对比查询 | 拥堵事件查询 | 拥堵模式查询 | 路网堵点查询

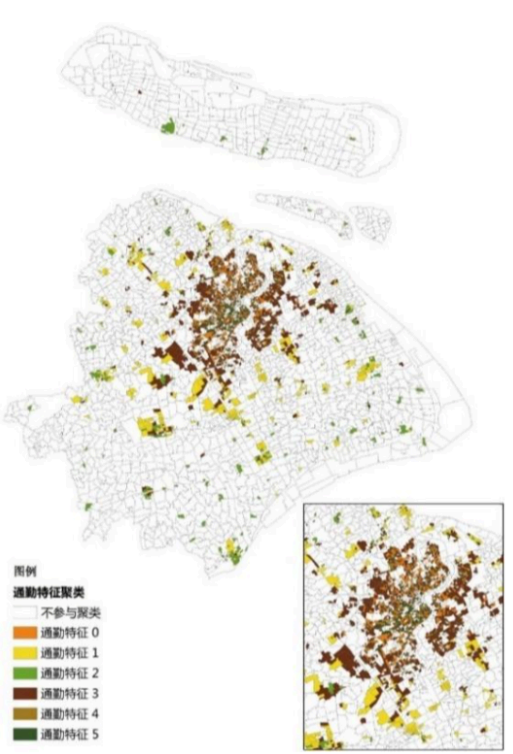
人工智能助力交通管控

人工智能助力交通管控 (4)



机器学习辅助城市规划

- 融合来自政府和企业的普查、交通、通信、房价等数百项数据，构造多维城市运行指标。
- 利用机器学习算法，对城市进行多角度的现状评估、风险预警，决策模拟。
- 应用1:
- 根据城市地块的就业类、居住类几十项定性指标，研究各类地块在城市中的空间分布。
- 应用2:
- 从就业、居住功能角度分类描述特定地块。
- 应用3:
- 模拟政策对地块指标带来的数量变化，并预测政策是否会使地块发生质的变化。



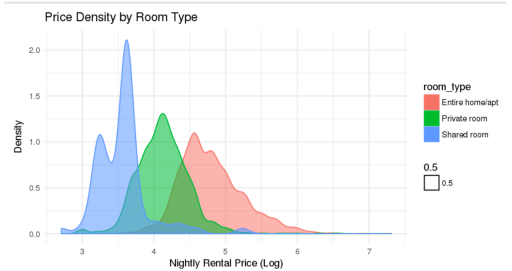
基于通勤特征的居住空间聚类结果空间分布

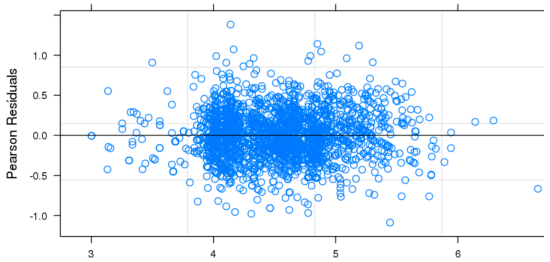
IBM 机器学习平台提供一个数据科学家的协作平台，能帮助我们快速探索数据并创建模型， 并将高效模型服务到同衡的客户。

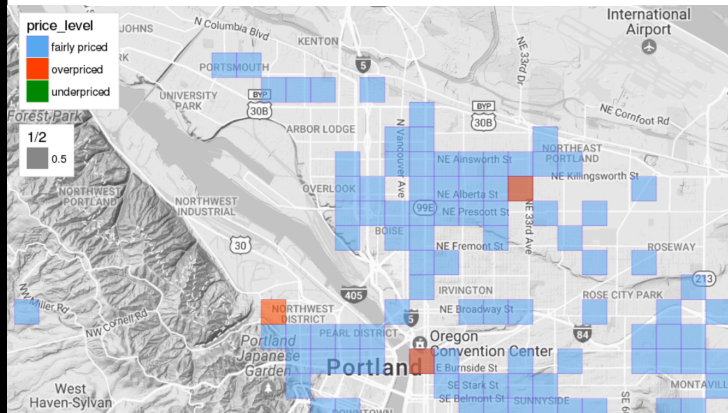
-- 李栋博士 清华同衡创新研究院 常务副主任

机器学习预测租房价格

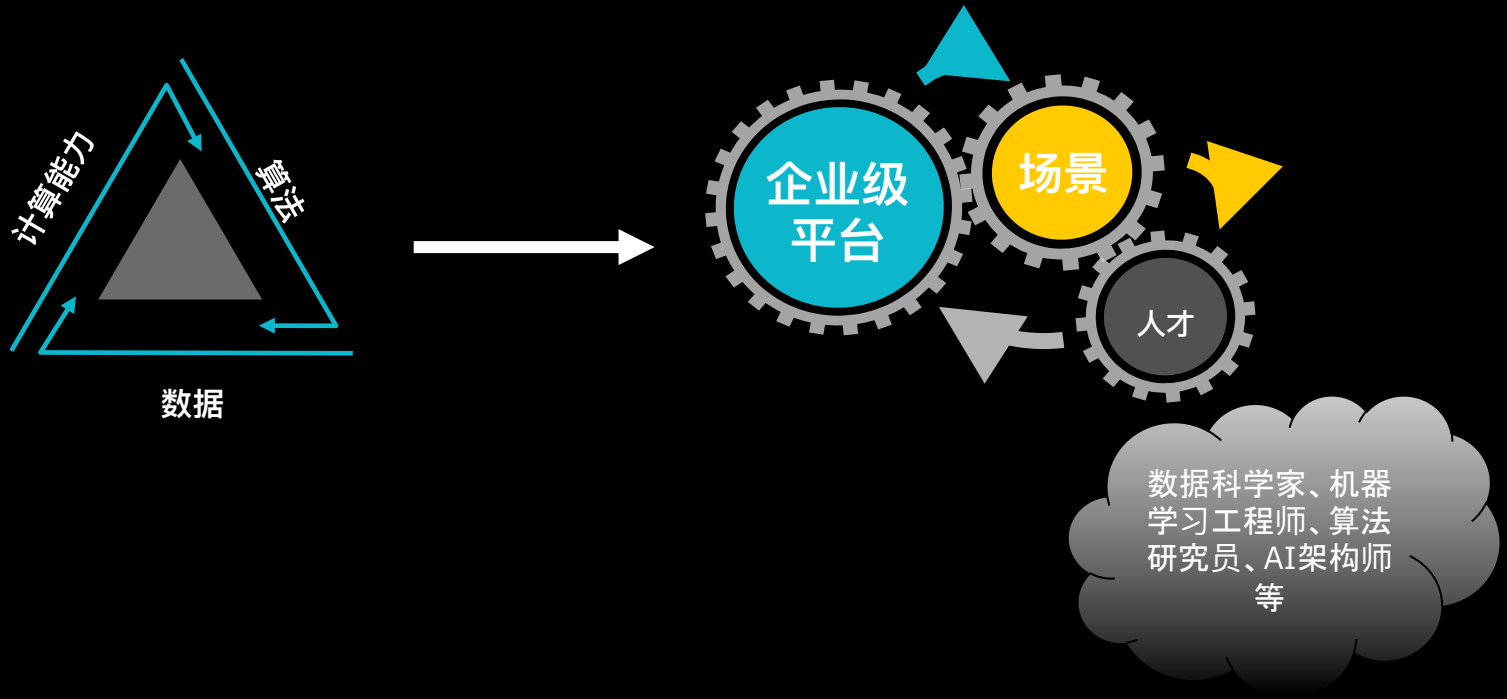
- 业务场景
  - 去到一个新的城市， 怎么从Airbnb上选择性价比高的房子
  - 出租房屋时， 如何才能定合适的价格呢?
- 机器学习方法
  - 数据获取-- 从Airbnb 上抓取出租房屋信息和价格
  - 数据清洗– 原始数据来自于互联网，需要进行异常值处理、缺失值填充等
  - 数据探索– 基于区域、房屋类型、价格、可住人数等进行分析
  - 特征工程– 数据分组、分箱、特征选择
  - 模型选择和超参数调优 – RF, Xgboost, Mixed-effects Model 等
- 业务价值
  - 分析发现被低估价值的出租房
  - 在Airbnb上找到性价比高的租房
  - 为自己的房子出租前定合适的价格







Ⅱ企业级机器学习的工程化挑战



2

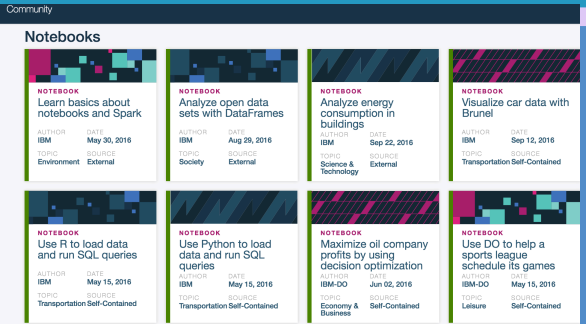
IBM机器学习中心

IBM Machine Learning Hub

学习 创新 应用

Learn Innovate Apply

- 你的数据+场景+数据科学家 *AND*
- *IBM*数据科学家+*IBM*机器学习平台



3

产品自带案例和应用

1

完成向AI时代的转型不仅仅涉及到技术...

...还有同样重要的思想和文化