2018 第四届大型企业信息运维高峰会
The 4th Large Enterprise Information Operation and Maintenance Summit

基于机器学习的数据库智能化运维

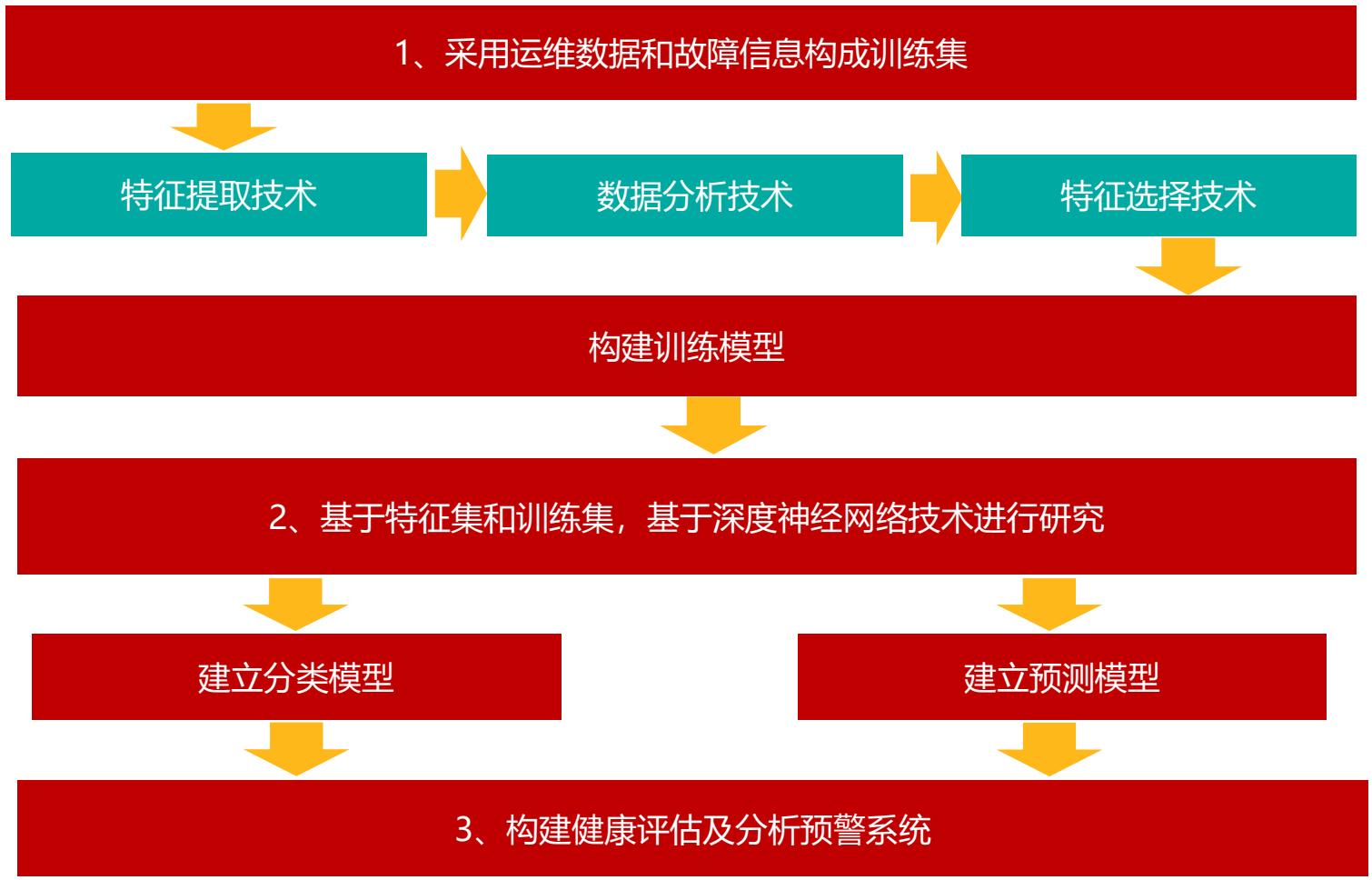
基石数据 徐戟

关于我和基石数据

- 徐戟：网名白鳝（QQ:62565），资深系统优化专家，从事信息系统建设与优化工作超过25年，著有《Oracle DBA优化日记》、《Oracle RAC日记》、《DBA的思想天空》等著作
- 南京基石数据/南瑞集成 技术总监
- 信息无障碍研究会专业顾问
- 南京基石数据：由南大尚诚、南瑞集团与徐戟技术团队联合出资成立的混合所有制企业，主要从事软硬件产品研发、技术咨询服务、数据运营等业务

问题提出-数据库运维面临的困境

机器学习-一扇新的大门


1、采用运维数据和故障信息构成训练集
特征提取技术
数据分析技术
特征选择技术
构建训练模型
2、基于特征集和训练集，基于深度神经网络技术进行研究
建立分类模型
建立预测模型
3、构建健康评估及分析预警系统

研究案例说明

- 从50个生产环境的数据库实例采集数据
  - 数据构成：15维负载数据、26维性能数据
- 运行数据的KPIs
  - 采集2017年1-8月的数据
  - 每个小时采集一个采样点
- 监督数据：使用基石D-SMART大师问诊系统作为监督数据
  - 负载分（LOAD SCORE） → 负载分类(LOAD LEVEL)
  - 性能分(PERF SCORE) → 性能分类(PERF LEVEL)
- 机器学习
  - 选择最有效的指标与最有效的算法建立分析模型

模型训练

数据说明

数据编号	f1	f2	……	fn	y	$\hat{y}$
t1					y_1	$\hat{y}_1$
t2					y_2	$\hat{y}_2$
……					… …	
tm					y_m	$\hat{y}_m$

$$y \in \{LoadScore, LoadLevel, PerfScore, PerfLevel\}$$

模型输出： $\hat{y} = f(\vec{f}), \vec{f} = (f_1, f_2, ..., f_n)$

Score为回归问题
Level为分类问题
(建立并训练模型，在给出时刻t各属性值，能输出该时刻的回归或分类值)

f1 … fn 的特征重要性的评估
以及选出较小的特征子集
以期降低复杂度
同时提高精确度
特征选择

特征选择

- 最大信息系数 – MIC

$$I(f_i;y)=\int_{f_i}\int_{y}\log\left(\frac{p(f,y)}{p(f)p(y)}\right)dfdy$$

- Relief-F 方法

$$\delta^j=\sum_i -diff(x_i^j,x_{i,nh}^j)^2+diff(x_i^j,x_{i,nm}^j)^2$$

- Lasso 方法

$$\min\sum_{i=1}^{m}(y_i-(\omega^T x_i+b))^2+\lambda\|\omega\|_1$$

snap_id	f1	f2	……	fn	y	$\hat{y}$
t1 (x1)					y_1	$\hat{y}_1$
t2 (x2)					y_2	$\hat{y}_2$
……					… …	
tm (xm)					y_m	$\hat{y}_m$

回归方法

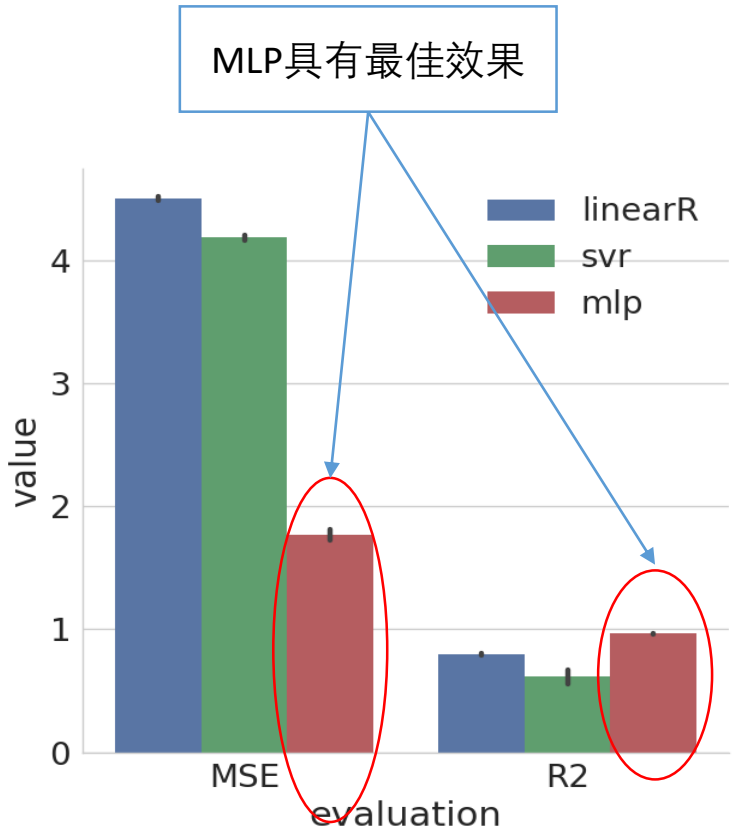
- 定义

$$y = f(x), y \in R$$

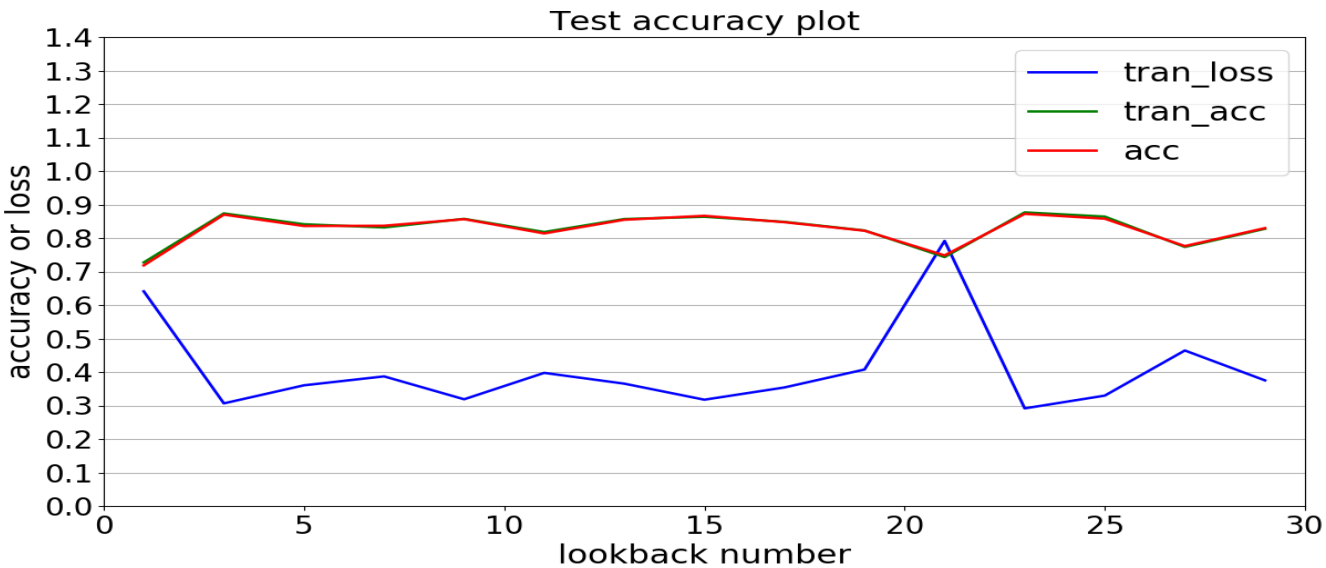
方法	学习对象	训练目标函数	优点
线性回归	$\hat{y} = \omega^T x + b$	$\min \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$	线性假设、简单
支持向量回归 (线性核)	$\hat{y} = \omega^T x + b$	$\min \frac{1}{2}\|\omega\|^2 + C\sum_i \ell_{\varepsilon,i}$	相对小的数据集上表现出很好的性能
多层感知机	各层连接权值	$loss = \min \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$	较为复杂，性能好

回归方法选择

- 采用的方法
  - 线性回归 -- LinearR
  - 支持向量回归 -- SVR
  - 多层感知机 – MLP
- 评价指标
  - 均方误差
  - 拟合优度可决系数R2得分



数据库运行状态预测

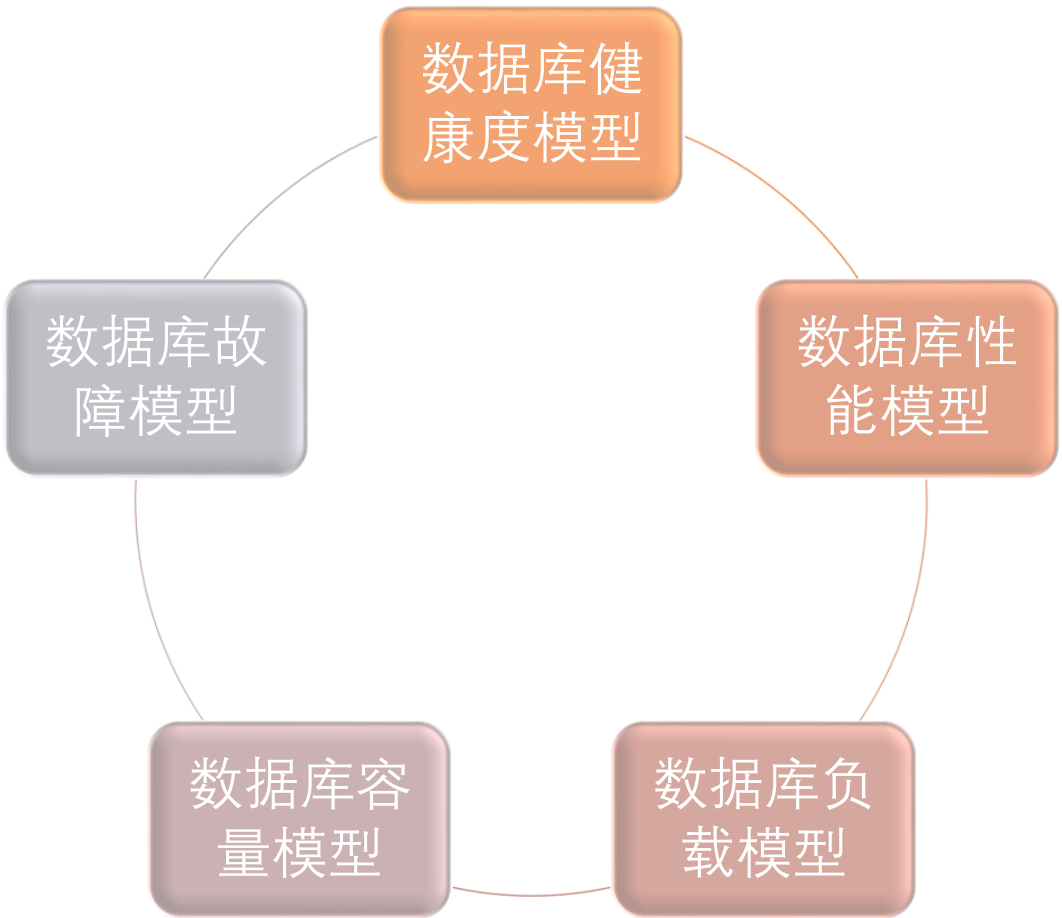


加入当前属性值后的分类曲线，回看之前23个样本预测未来一小时的数据库运行状态，并按照A/B/C/D进行评价，经过500轮训练后，模型预测的准确率为87.2%。

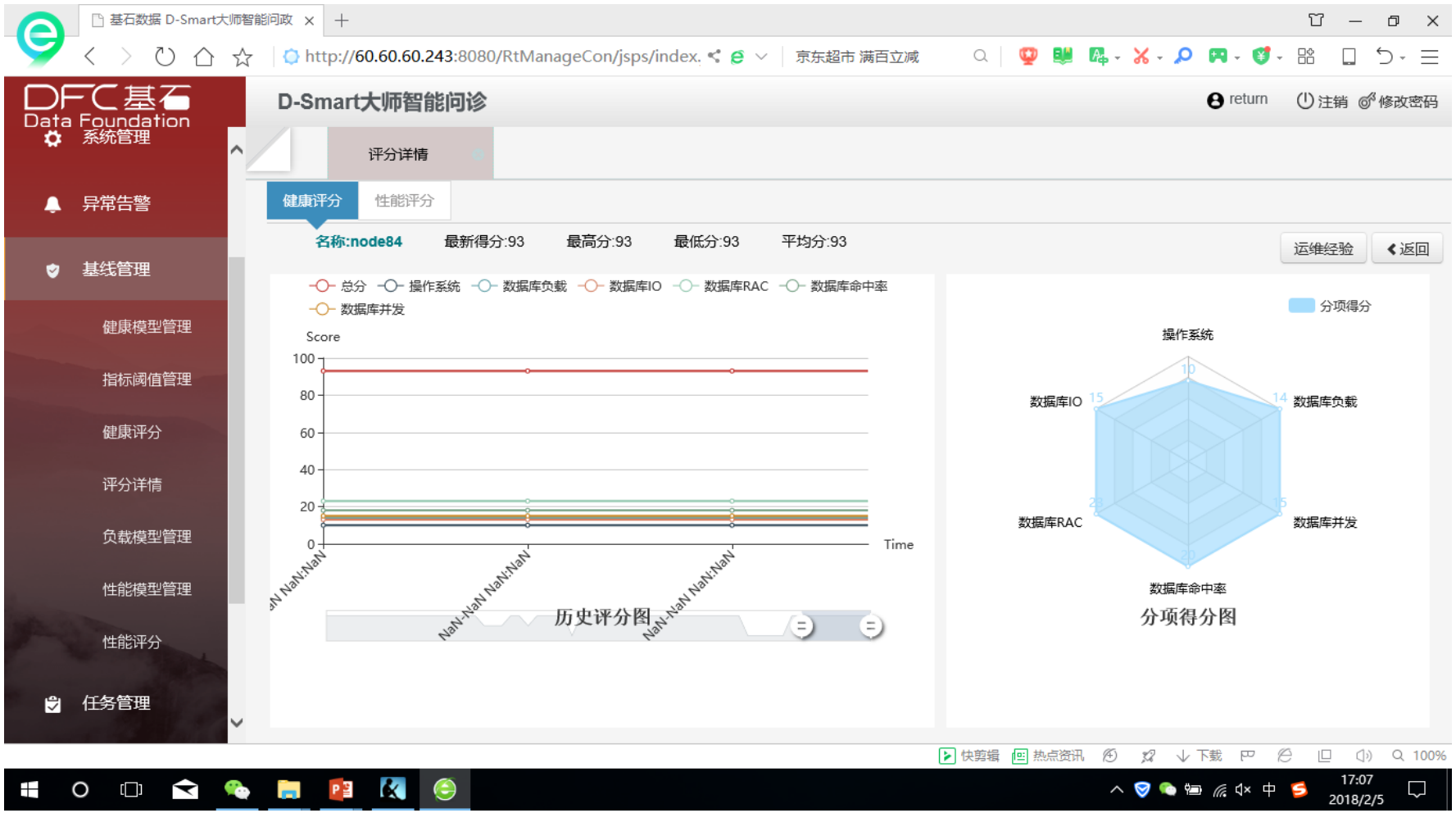
结论

- 使用神经网络对于SCORE类预测有较好的效果，通过3000多个样本验证，AI打分的误差在正负4之间
- 使用支持向量机和随机森林对LEVEL类预测有较好的效果
- 通过机器学习对D-SMART中的专家模型进行验证的结果是收敛的，说明目前的指标体系确实能够较为准确的反映出系统现状
- AI模型用于预测数据库未来运行状态的准确率已达到实用水平

数据库运维中的重要智能模型



健康与性能模型


基石数据 D-Smart大师智能问政
http://60.60.60.243:8080/RtManageCon/jsps/index.
京东超市 满百立减
DFC基石
Data Foundation
系统管理
D-Smart大师智能问诊
return
注销
修改密码
异常告警
基线管理
健康模型管理
指标阈值管理
健康评分
评分详情
负载模型管理
性能模型管理
性能评分
任务管理
评分详情
健康评分
性能评分
名称:node84
最新得分:93
最高分:93
最低分:93
平均分:93
运维经验
返回
总分
操作系统
数据库负载
数据库IO
数据库RAC
数据库命中率
数据库并发
Score
100
80
60
40
20
0
Time
历史评分图
分项得分
操作系统
数据库负载
数据库并发
数据库命中率
数据库RAC
数据库IO
分项得分图
快剪辑
热点资讯
下载
100%
17:07
2018/2/5

数据库状态巡检平台

外网门户　企业门户　移动终端 APP/微信..　大屏

巡检结论生成	巡检结论展示	巡检结论标注	巡检报告生成	历史数据分析
巡检数据采集	手工数据导入	巡检分析工具集	数据可视化工具	远程协作工具
系统配置管理	数据库健康评估	数据库性能评估	采集模板管理	巡检报告管理

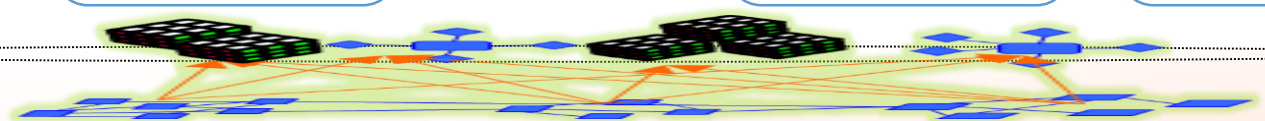
数据库性能打分引擎
数据库健康打分引擎

状态巡检规则引擎
巡检任务管理引擎

机器学习引擎

数据自动采集引擎
数据装载引擎

数据可视化引擎
数据分析处理引擎

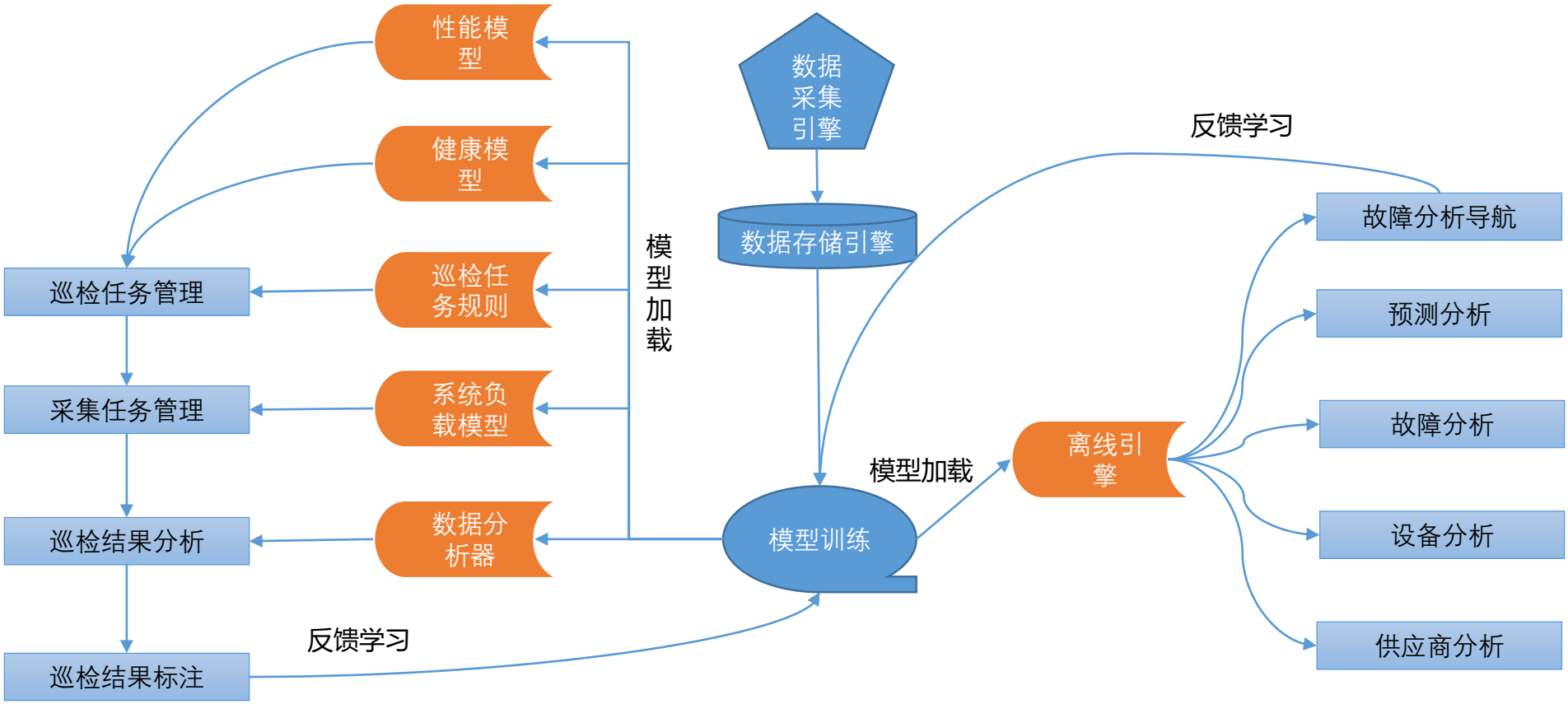


大数据平台

统一采集引擎

营销系统　财务系统　PMIS系统　ERP系统　基建管控系统

机器学习器在状态巡检中的应用



运维知识在运维工作中的使用方式演变

运维标准化文档

基于全文检索的运维知识库

基于知识图谱的知识精准搜索平台

基于机器学习的智能知识网络

智能知识自动化系统

运维知识自动化

运维知识最初的形态是方案与预案，标准化预案，标准化生产工艺等在前些年企业运维管理中发挥了巨大的作用

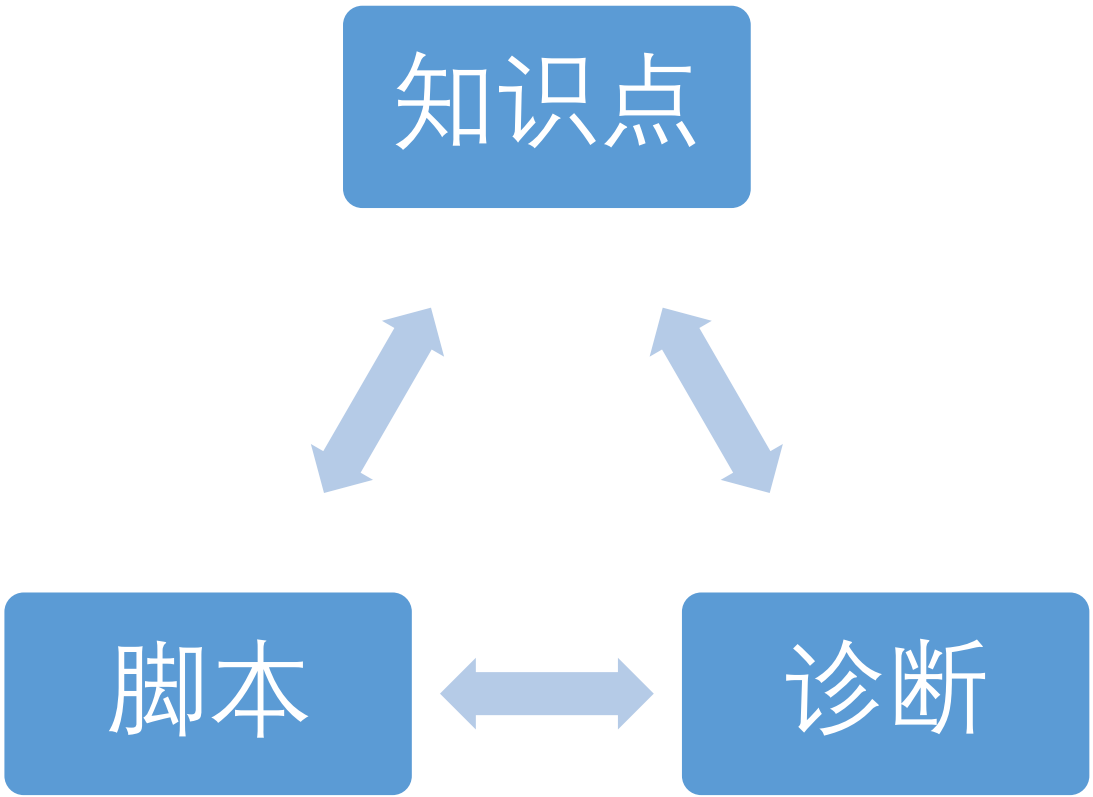
方案与预案很难组织与管理，于是出现了知识库系统，建立全文搜索引擎，对运维知识进行分类与导航、搜索

传统的知识管理系统是由一个个孤立的文档组成的，所以知识点之间的关系很难表示出来，于是出现了以自然语言处理NLP与知识图谱为基础的知识管理系统，可以把知识之间的相互关系串联起来

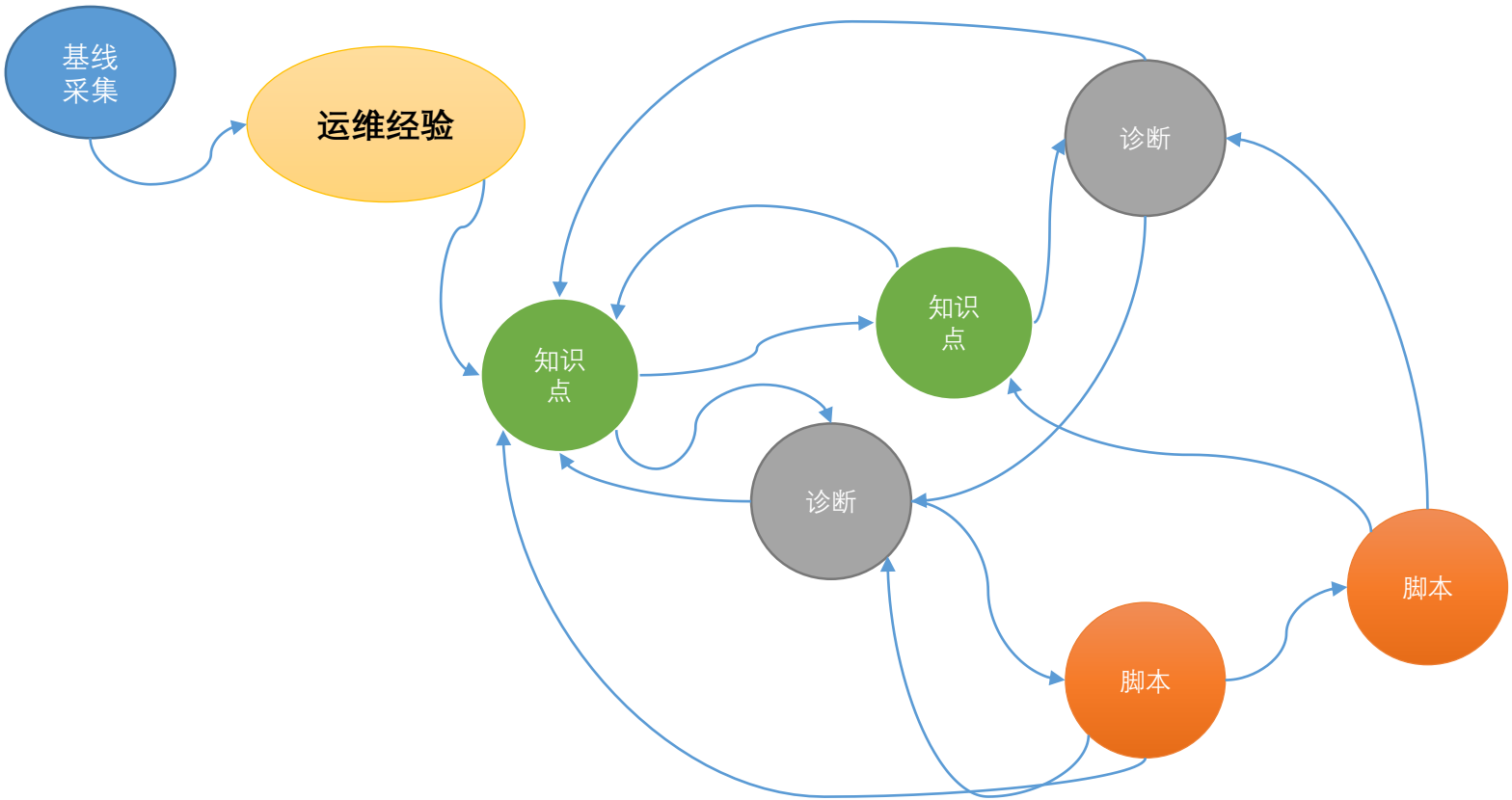
基于知识图谱的知识库也只是一个静态的资料库而已，如何让知识从死的文字变成能够直接帮助运维人员完成工作的工具，实现知识自动化，机器学习使这种想法成为可能

基于机器学习的知识自动化系统可以将专家的知识体系化的积累下来，通过自动化诊断和处理脚本完成自动化/半自动化的运维工作

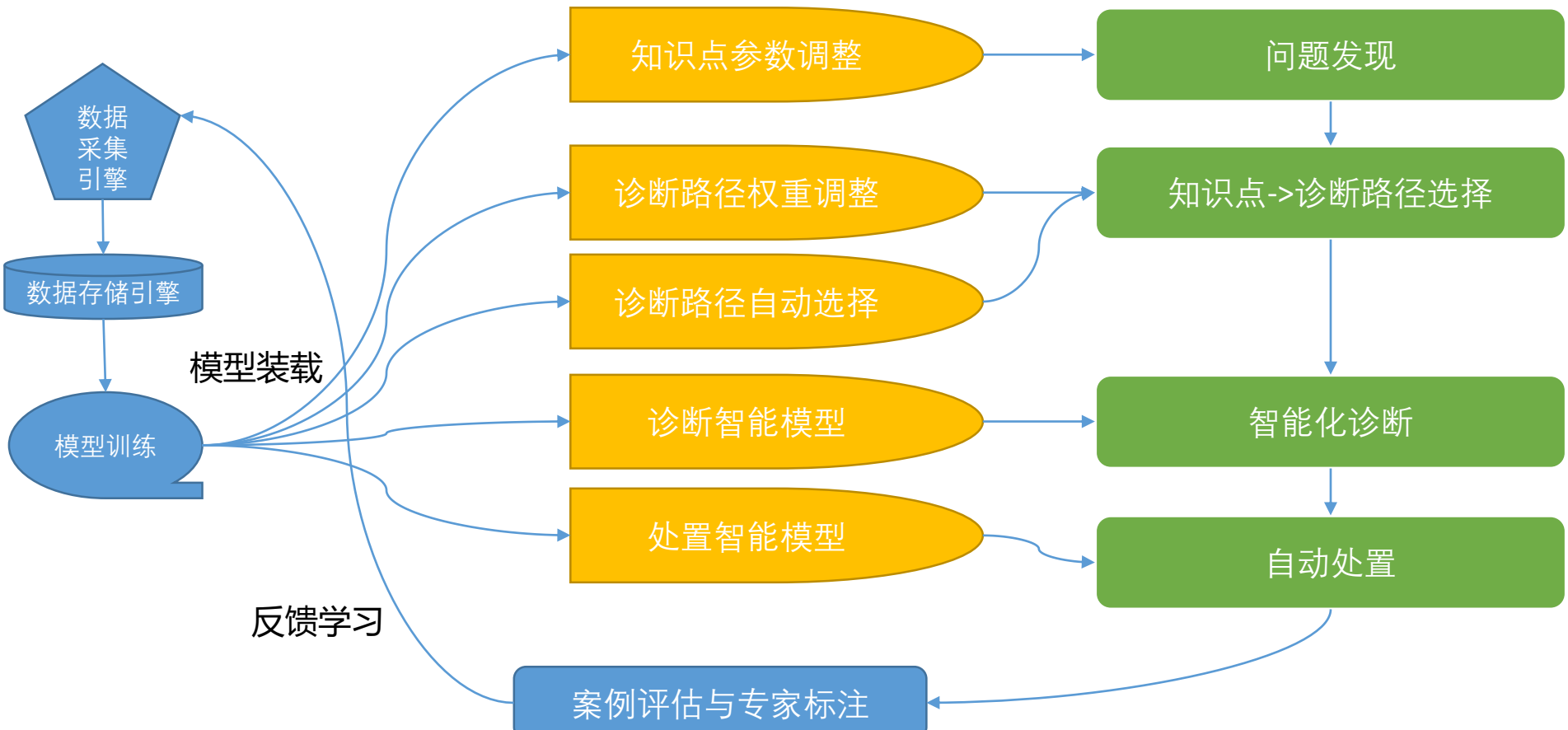
运维知识自动化数据模型



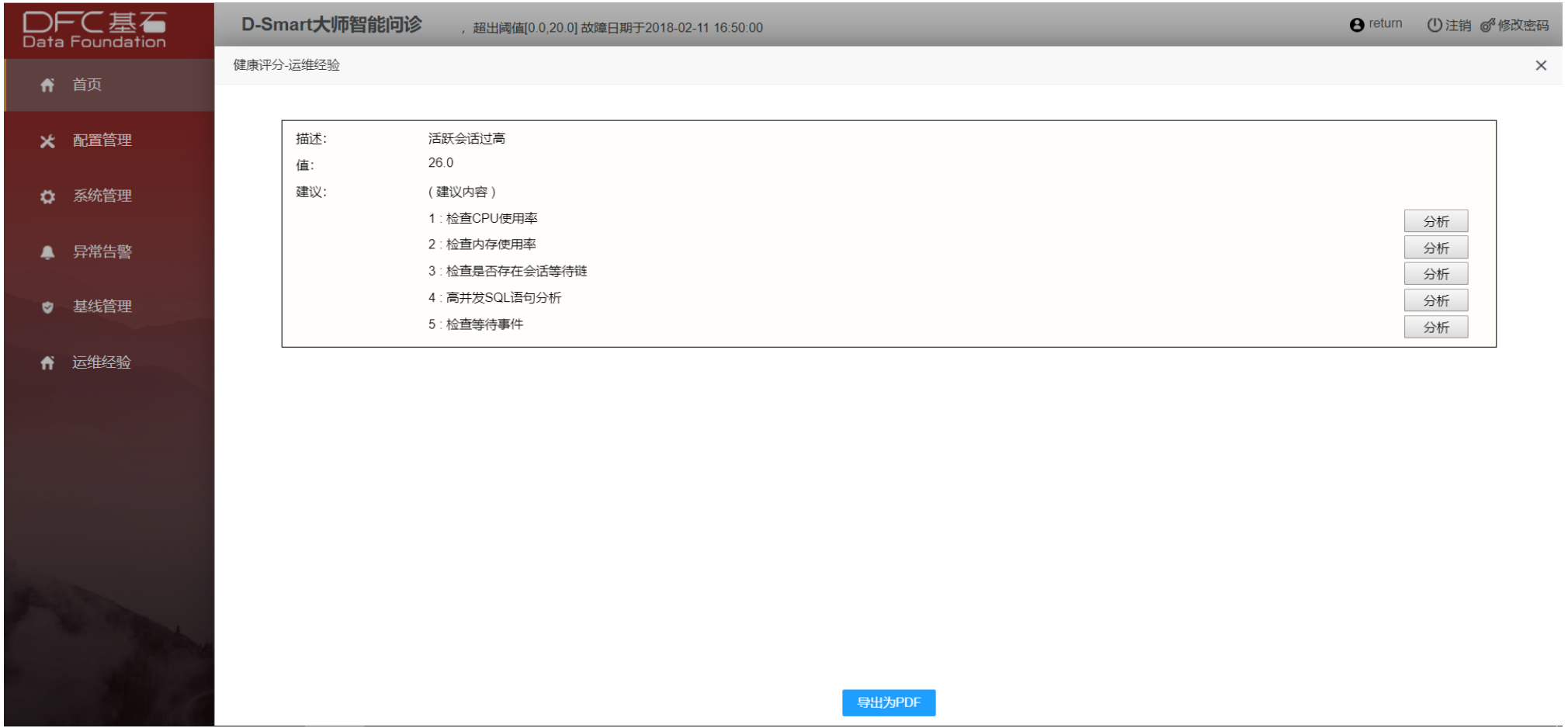
运维知识自动化数据模型举例



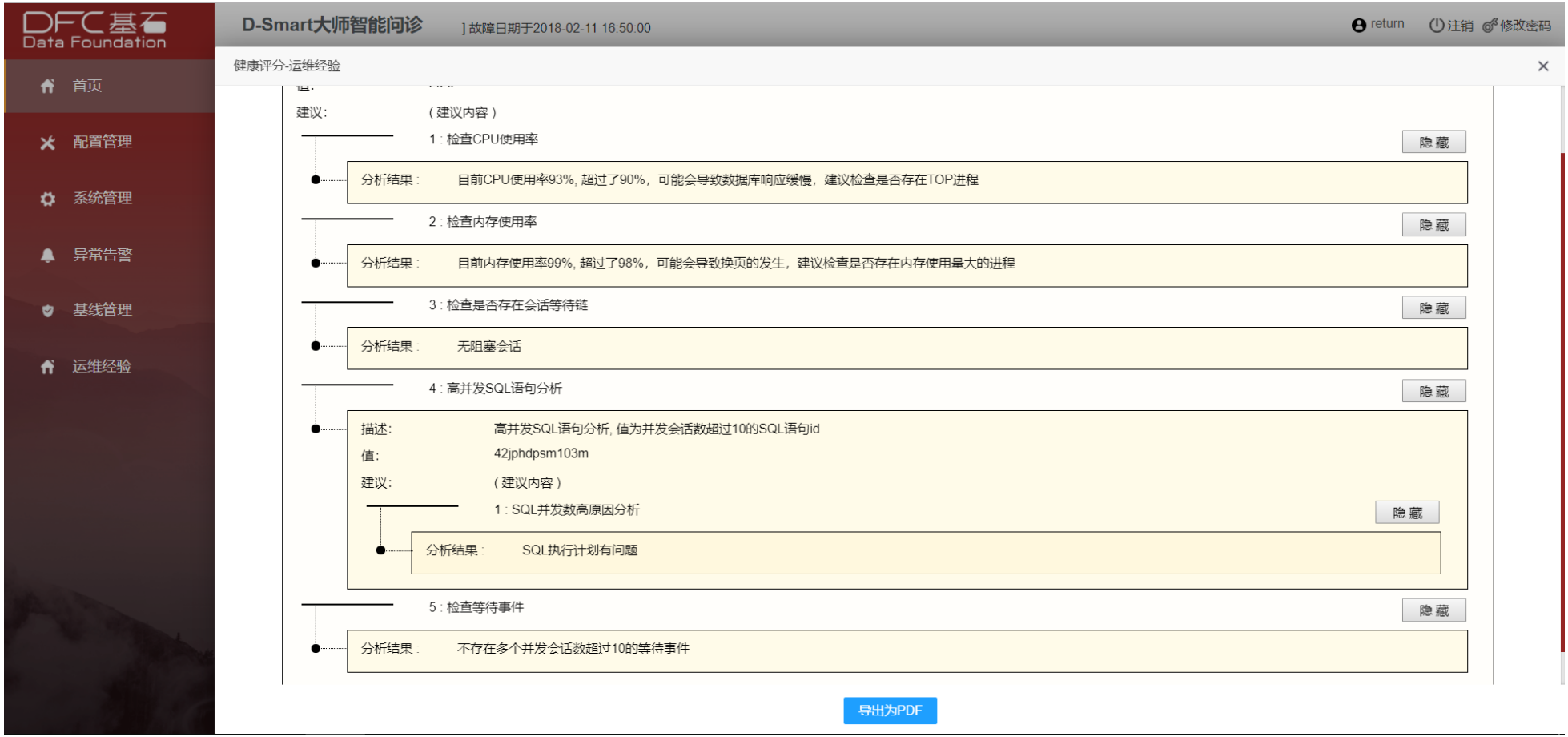
机器学习在知识自动化中的应用



D-SMART运维知识自动化功能



D-SMART运维知识自动化功能



运维知识自动化的价值

真正的可以积累与灵活应用知识的系统

像外骨骼机器人一样，知识自动化系统可以直接增强运维人员的能力，而且对运维人员的能力水平要求不高

实现长时间的知识积累，将核心价值从人转向组织

随着知识积累，平台能力越来越强，智能系统最终将超越人类的专家

传统自动化运维系统 VS D-SMART

传统运维自动化系统	D-SMART
运维自动化系统	运维**知识**自动化系统
以**指标、工具与运维场景**为核心的系统	以**运维知识、专家经验与智能模型**为核心的系统
比较容易解决**已知的，简单**的问题	更适合处理**未知的，复杂**的问题
通过提高**使用者的熟练度与能力**提高运维能力，人员变动后能力流失严重	通过**积累知识**提高企业的整体运维能力，人员变动后**能力保留在系统中**
只提供孤立的指标与基线，运维人员很难直接从中获得运维能力	提供**专家模型与智能模型**，针对指标与基线提供相关的运维知识，**通过运维知识去使用指标与基线**
系统只能通过版本升级提高系统自身的能力	通过知识积累，导入专家模型包，导入别人的经验以及机器学习，可以提升系统的运维能力

总结-智能模型对数据库运维的价值

使运维更有针对性，节约运维成本

减少对高水平DBA的依赖，使数据库运维成为可持续高价值的工作

提前发现问题并自动定位问题，真正做到防患于未然

积累运维数据，形成运维智能

THANKS