DTCC

2017第八届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2017

工业大数据技术与实践

王晨

清华大学 大数据系统软件国家工程实验室

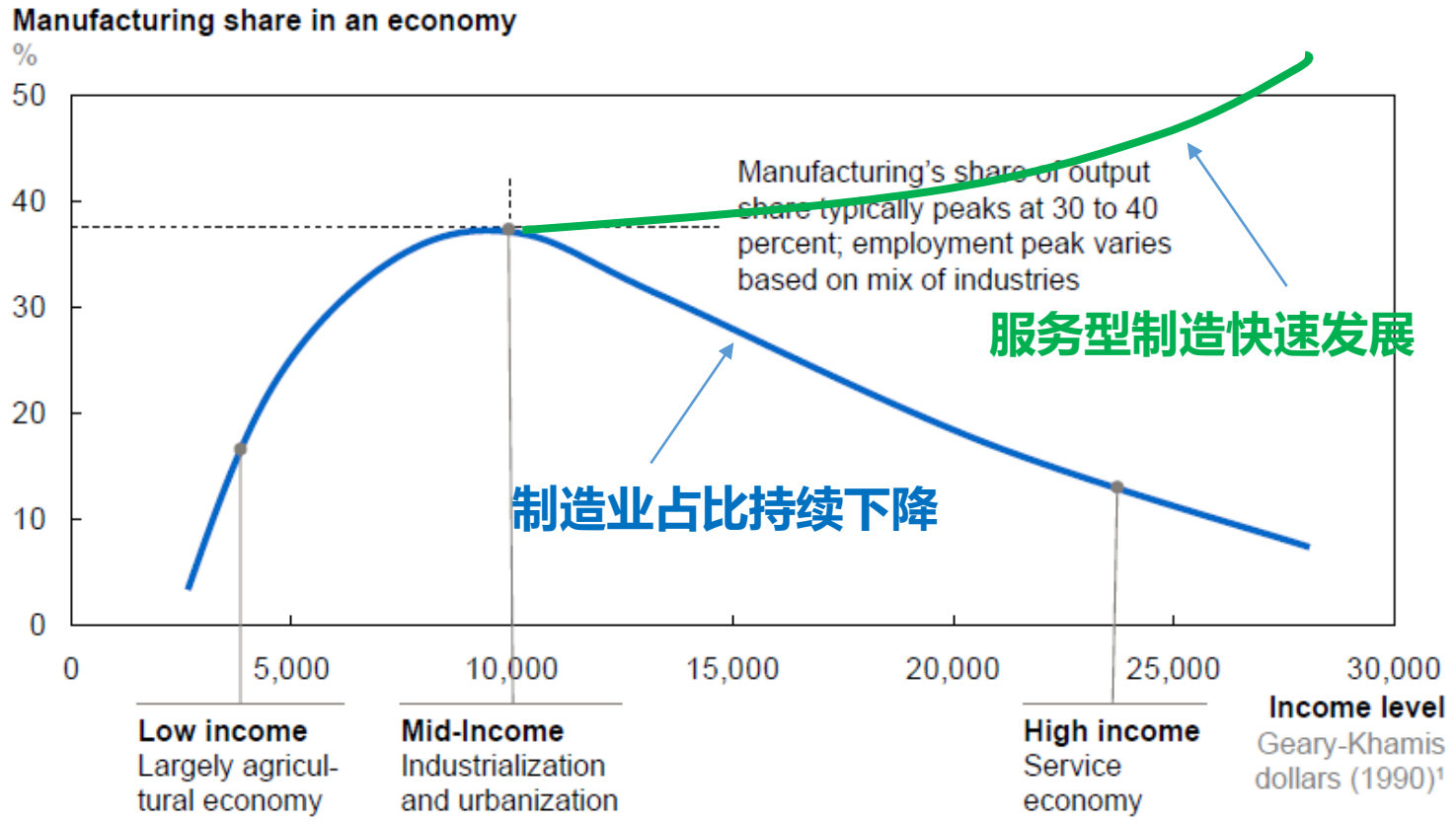
北京工业大数据创新中心

国家重点研发计划“面向高端制造的大数据管理系统”技术团队

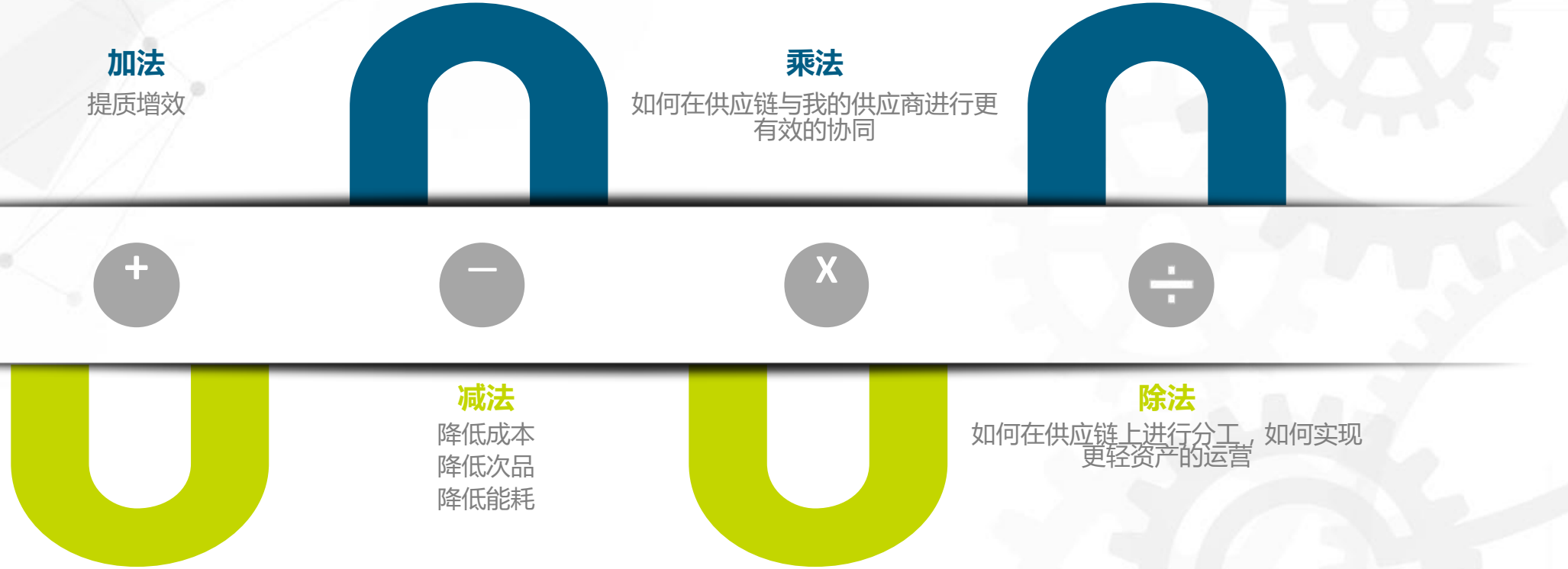
工业大数据概念

升级转型-全球制造业发展的必然趋势

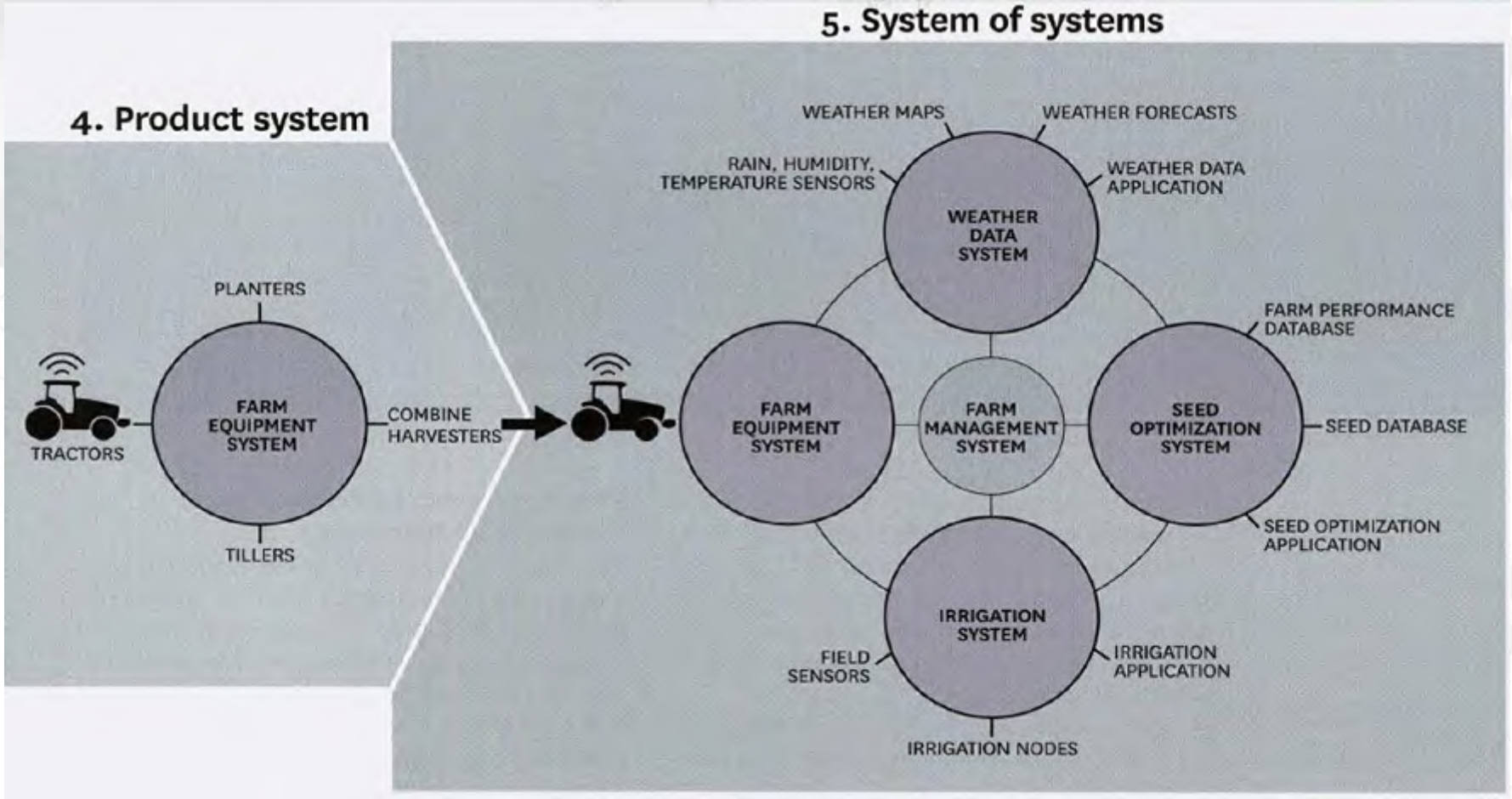
Gartner 2012：Manufacturing the Future : The next era of global growth and innovation



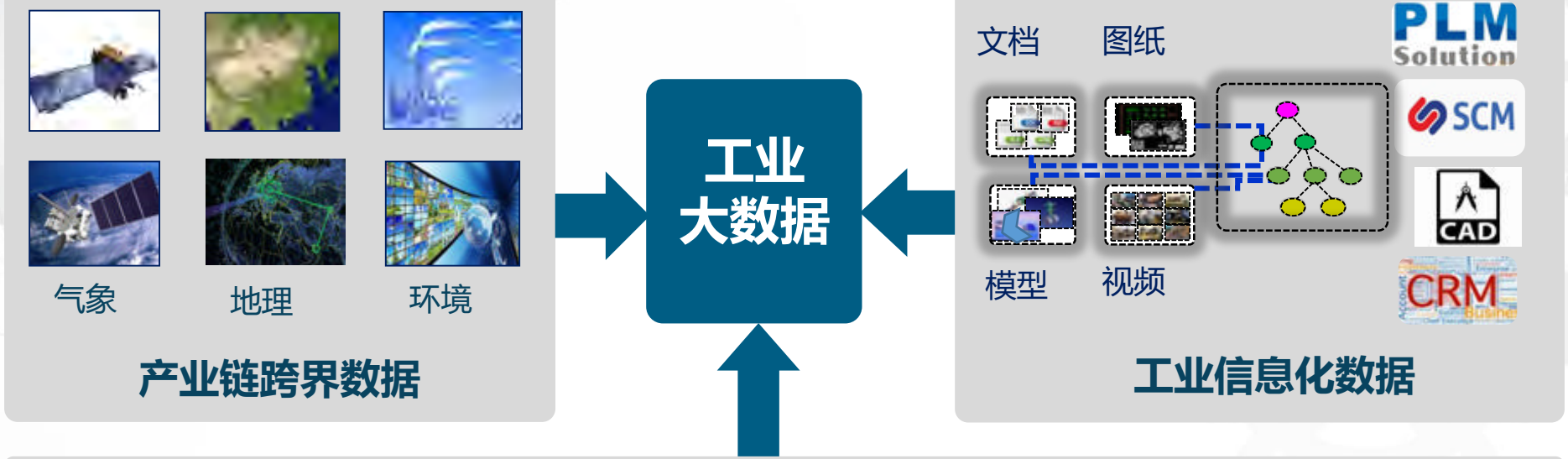
工业大数据的业务目标


加法
提质增效
乘法
如何在供应链与我的供应商进行更有效的协同
+
−
X
÷
减法
降低成本
降低次品
降低能耗
除法
如何在供应链上进行分工，如何实现更轻资产的运营

工业大数据的方向


5. System of systems
4. Product system
WEATHER MAPS
WEATHER FORECASTS
RAIN, HUMIDITY, TEMPERATURE SENSORS
WEATHER DATA APPLICATION
WEATHER DATA SYSTEM
PLANTERS
FARM PERFORMANCE DATABASE
TRACTORS
FARM EQUIPMENT SYSTEM
COMBINE HARVESTERS
FARM EQUIPMENT SYSTEM
FARM MANAGEMENT SYSTEM
SEED OPTIMIZATION SYSTEM
SEED DATABASE
TILLERS
SEED OPTIMIZATION APPLICATION
IRRIGATION SYSTEM
FIELD SENSORS
IRRIGATION APPLICATION
IRRIGATION NODES

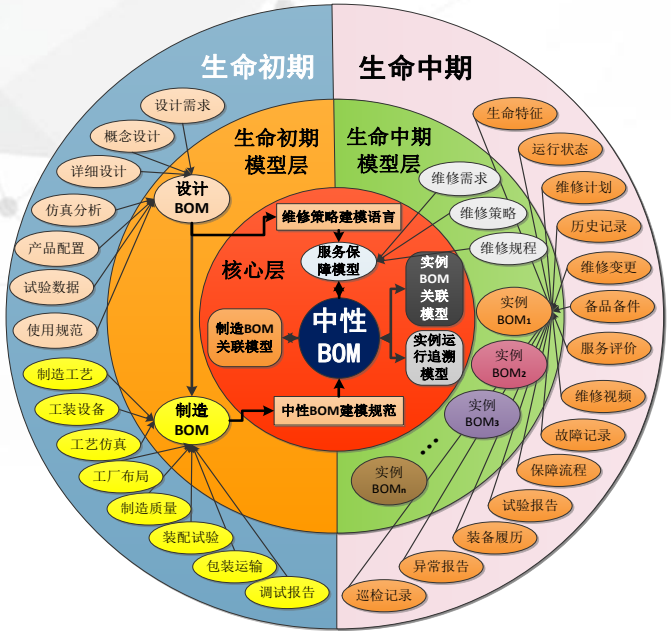
工业大数据的来源


气象
地理
环境
产业链跨界数据
工业
大数据
文档
图纸
PLM
Solution
SCM
CAD
CRM
模型
视频
工业信息化数据
机器设备数据
Pressure
Speed

工业大数据的特点

以多种类非结构化工程数据、过程与BOM图数据、高端装备监测时序数据为代表的高端制造业大数据

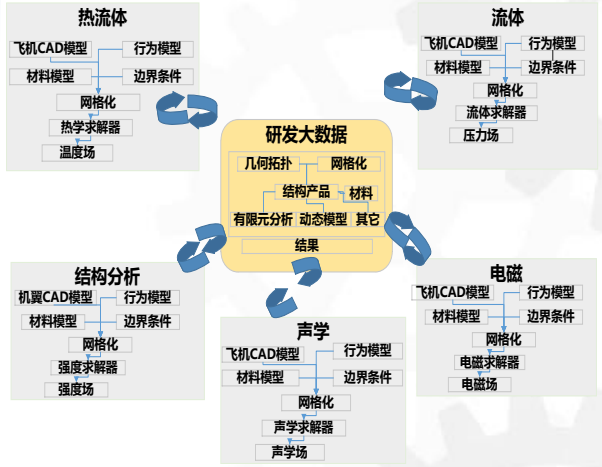
呈现 “多模态，高通量，强关联”特性



海量高速

机器7*24产生

采集频率高，数据量大



多学科异构数据信息交互模型

数据模态多样,结构关系复杂

典型高端制造企业数据类型可达300余种

汽轮机35万个零部件数据

数据通量大

50Hz，500测点/台，2万台风机

最高可达数千万数据点/秒

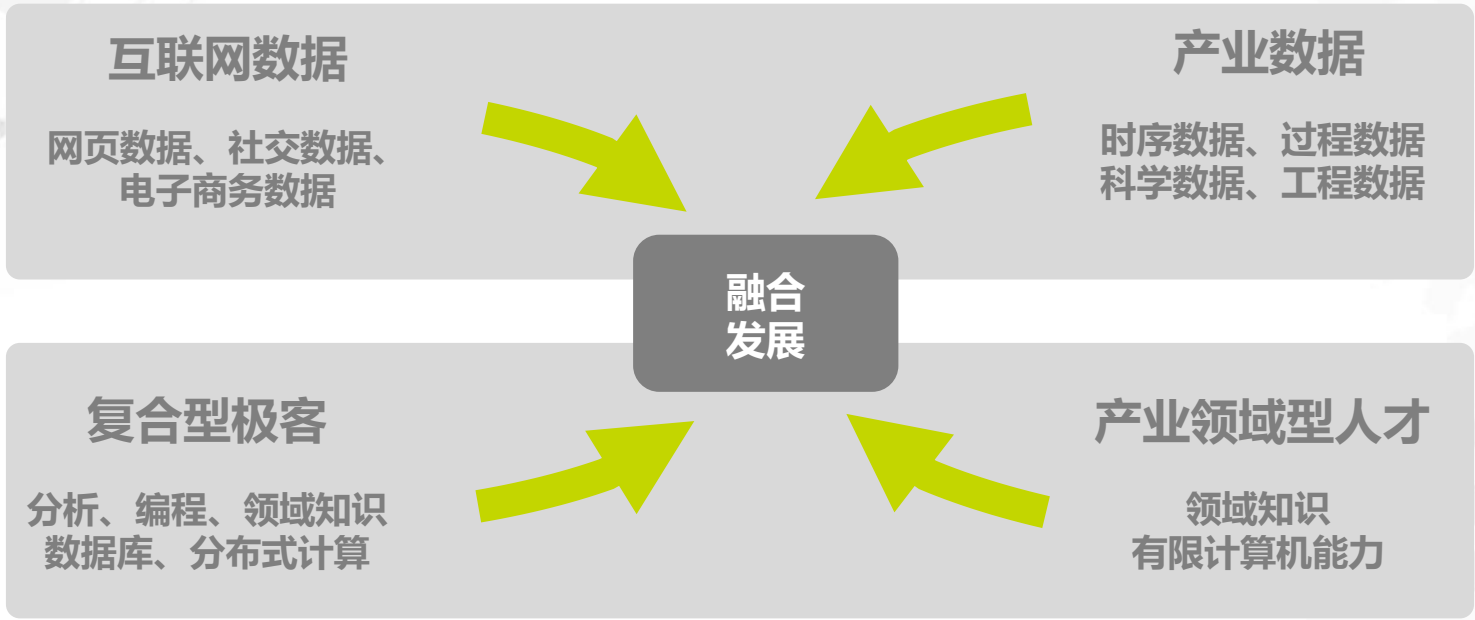
协作专业多

飞行器研发相关专业200多类

工业大数据技术挑战

大数据正在从消费互联网向产业互联网渗透

网页搜索 ● 电子商务 ● 政府治理 ● 公共服务 ● 工业4.0 ● 现代农业 ● 自主可控



国家重点研发计划-面向高端制造的大数据管理系统

大数据驱动的航天航空装备创新研发与应用示范

基于大数据的“互联网+制造”应用示范

标准规范、评测基准和测试工具

高端制造大数据管理系统

一体化管理

非结构化数据管理引擎

图数据管理引擎

时序数据管理引擎

键值数据库*

关系数据库*

高端制造大数据系统管理工具

工业大数据工作步骤

2

- 机器数据建模与元数据管理
- 数据质量分析
- 数据关联与语义集成

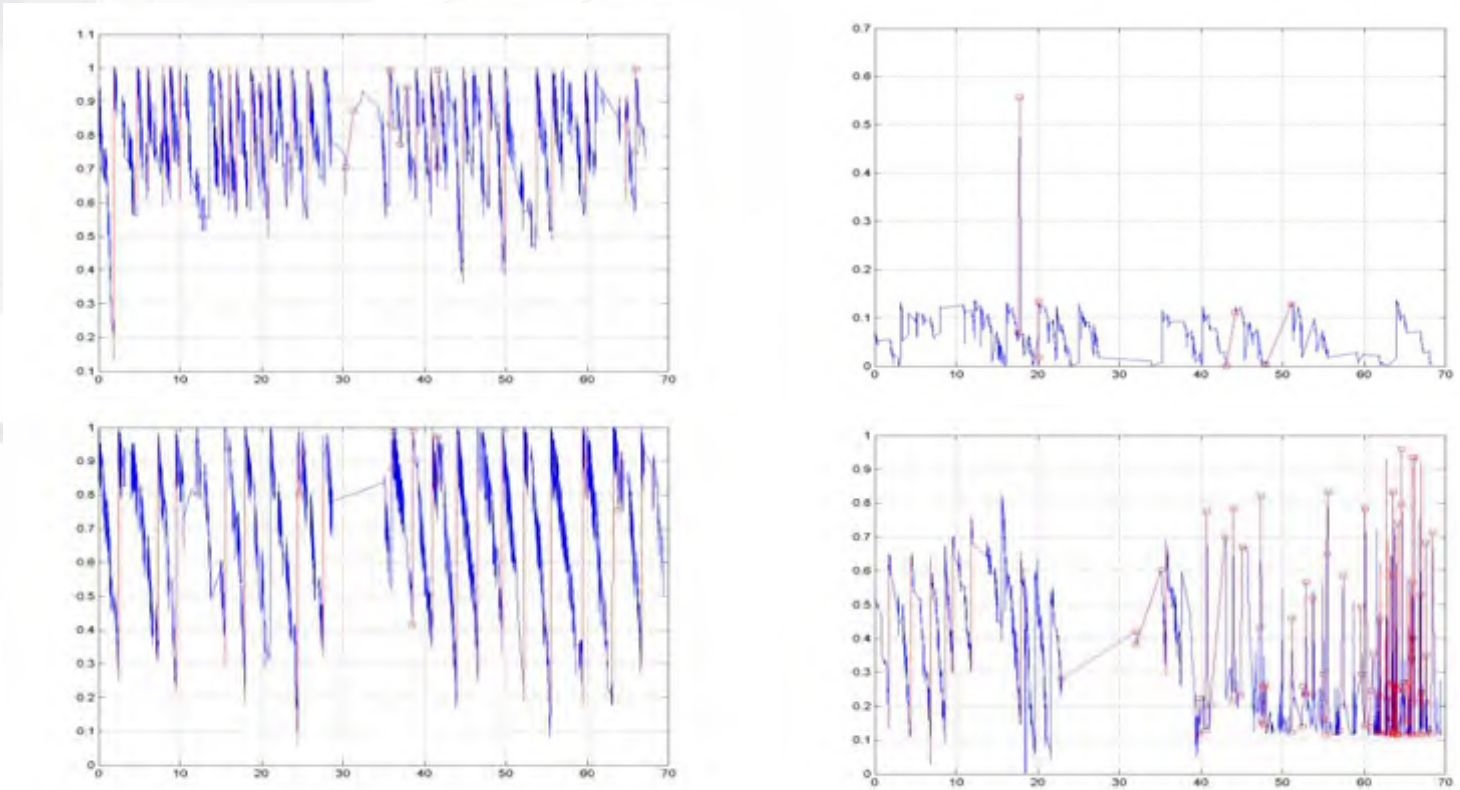
3

- 数据探索与可视化
- 数据分析
- 结果反馈

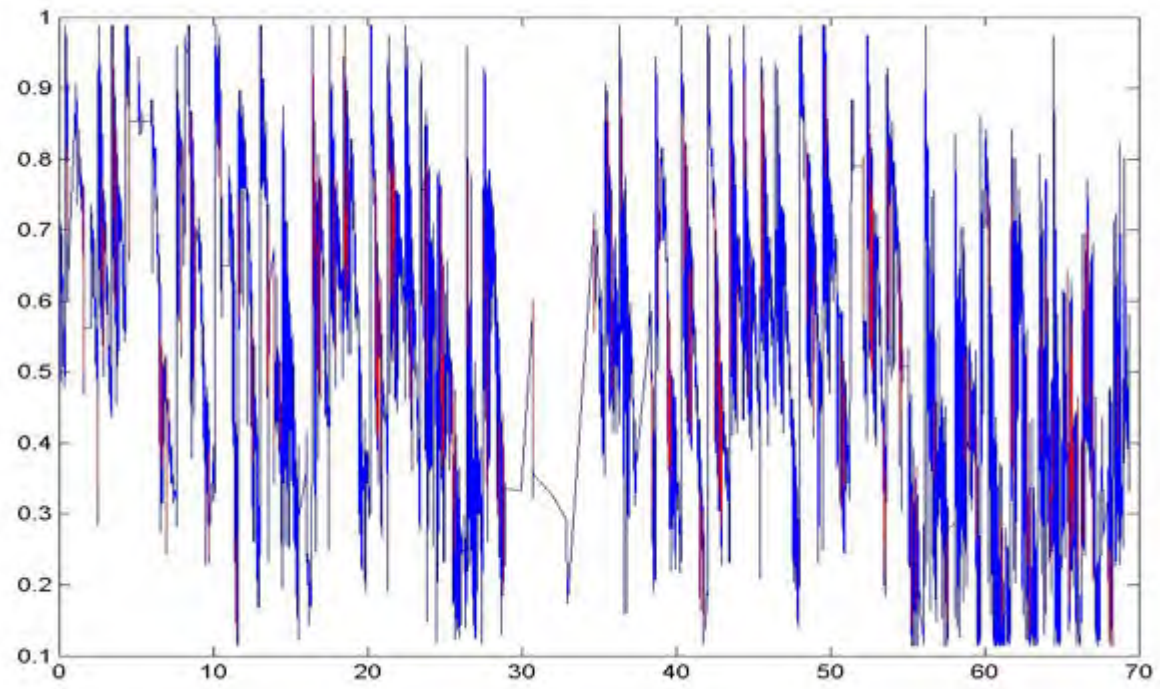
1

- 建立数据采集体系
- 持续采集与清洗
- 工业数据存储

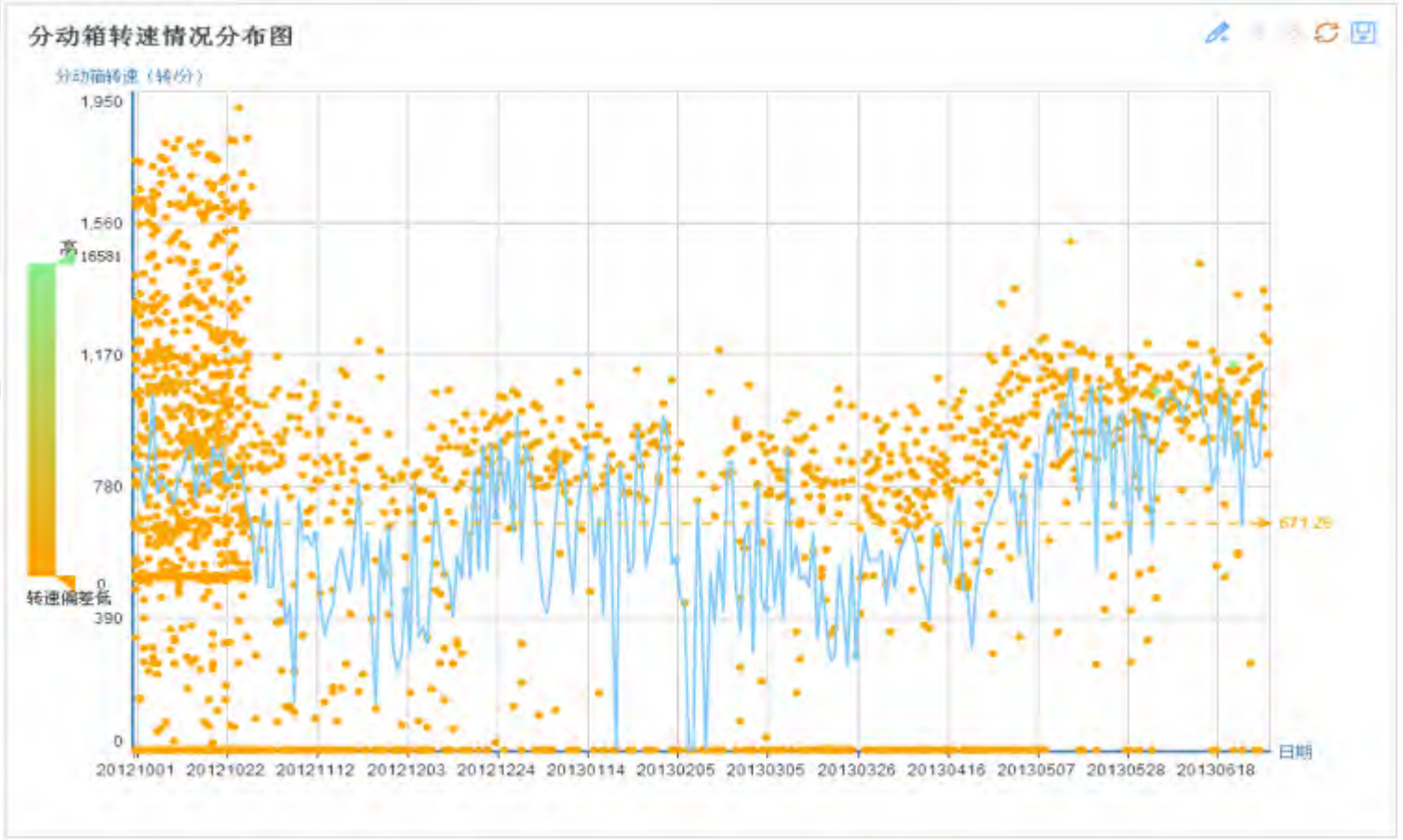
油位分析示例



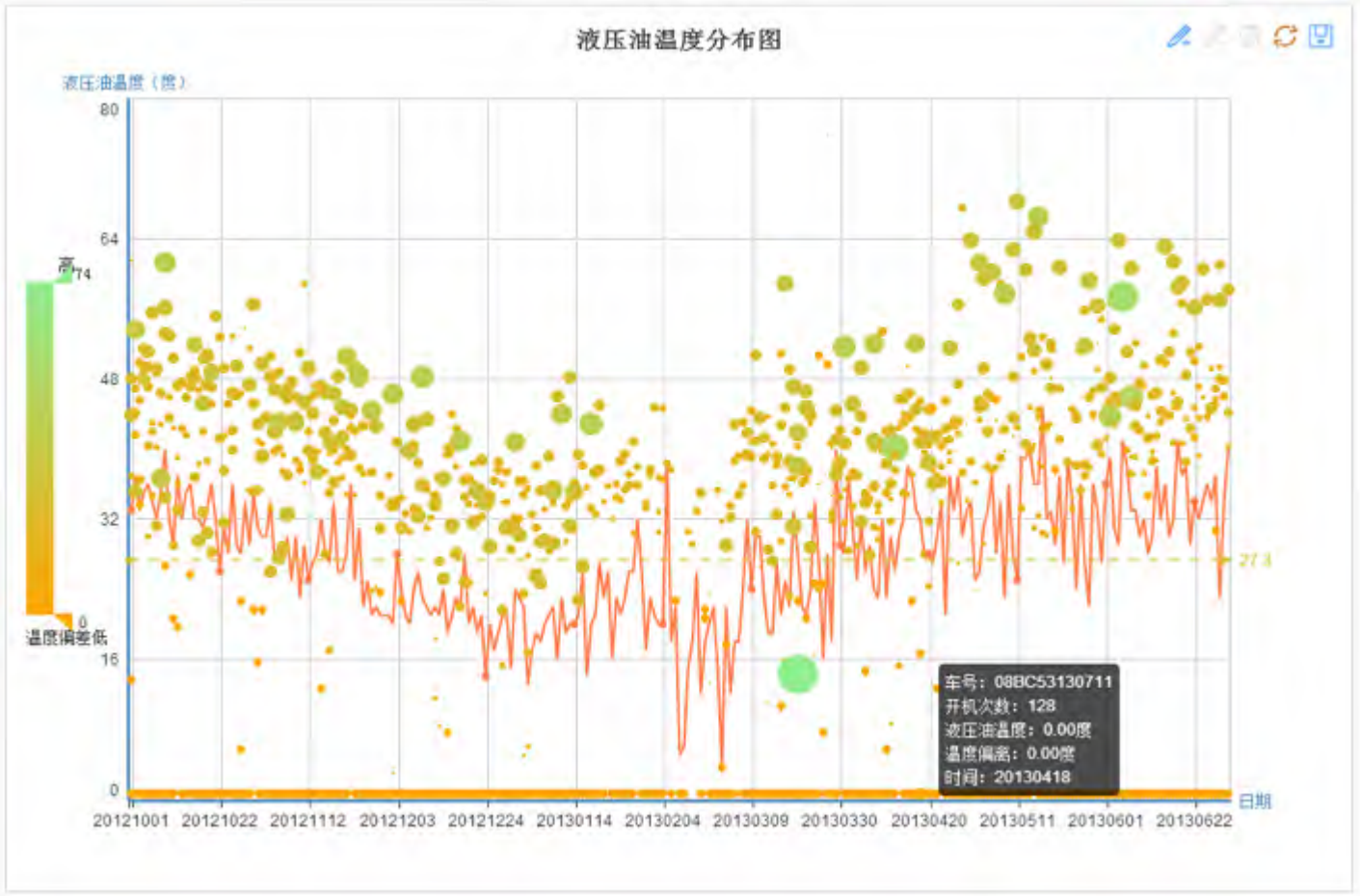
现实



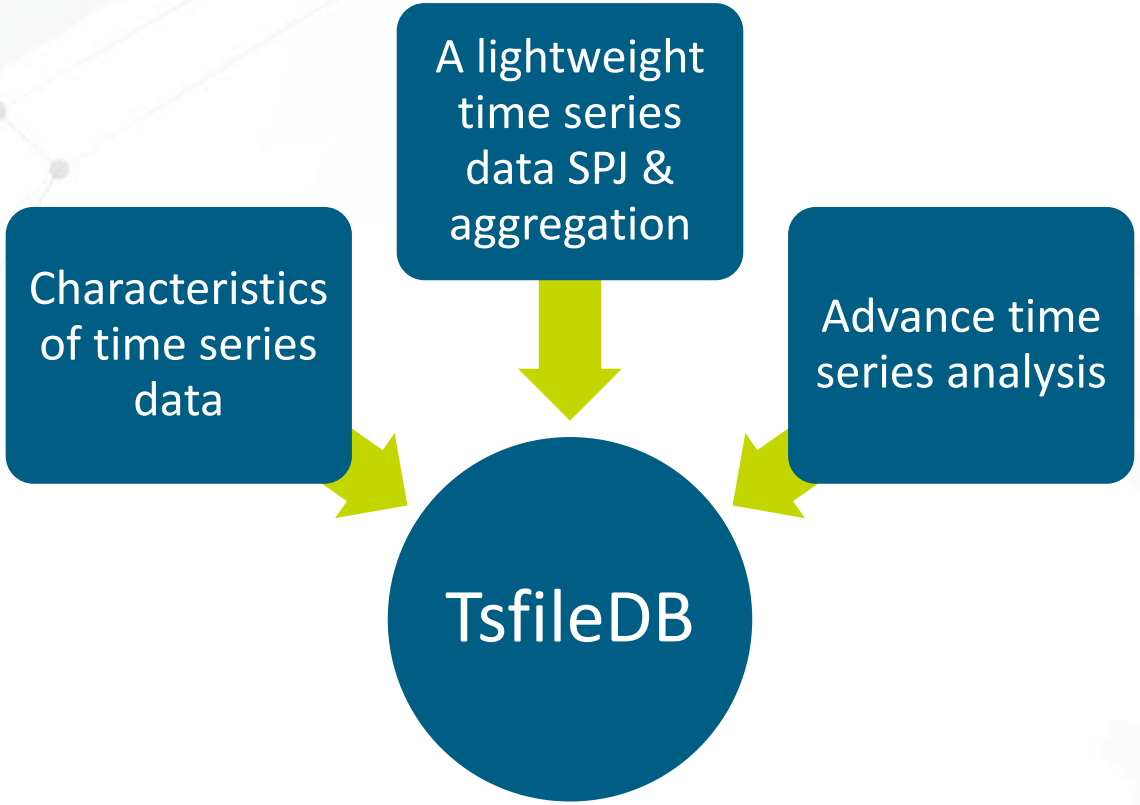
现实


分动箱转速情况分布图
分动箱转速（转/分）
1,950
1,560
1,170
780
390
0
671.29
日期
20121001
20121022
20121112
20121203
20121224
20130114
20130205
20130305
20130326
20130416
20130507
20130528
20130618

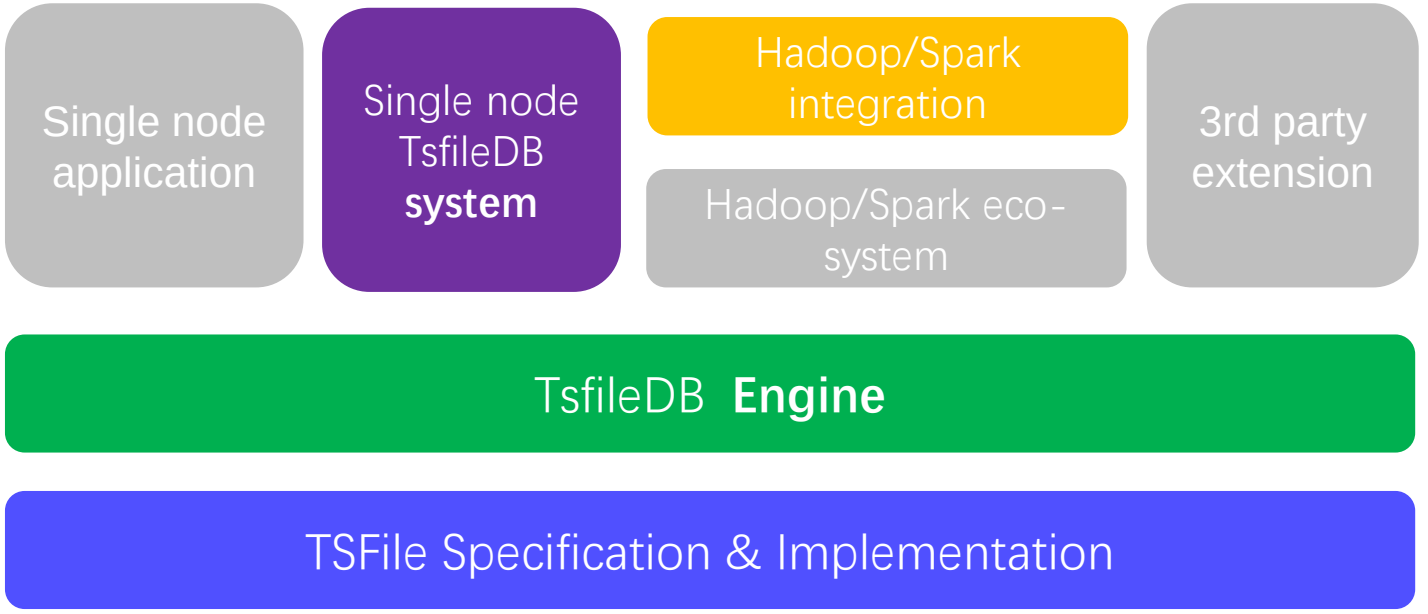
现实


液压油温度分布图
液压油温度（度）
80
64
48
32
16
0
高
74
0
温度偏差低
27.3
日期
20121001
20121022
20121112
20121203
20121224
20130114
20130204
20130309
20130330
20130420
20130511
20130601
20130622
车号：08BC53130711
开机次数：128
液压油温度：0.00度
温度偏差：0.00度
时间：20130418

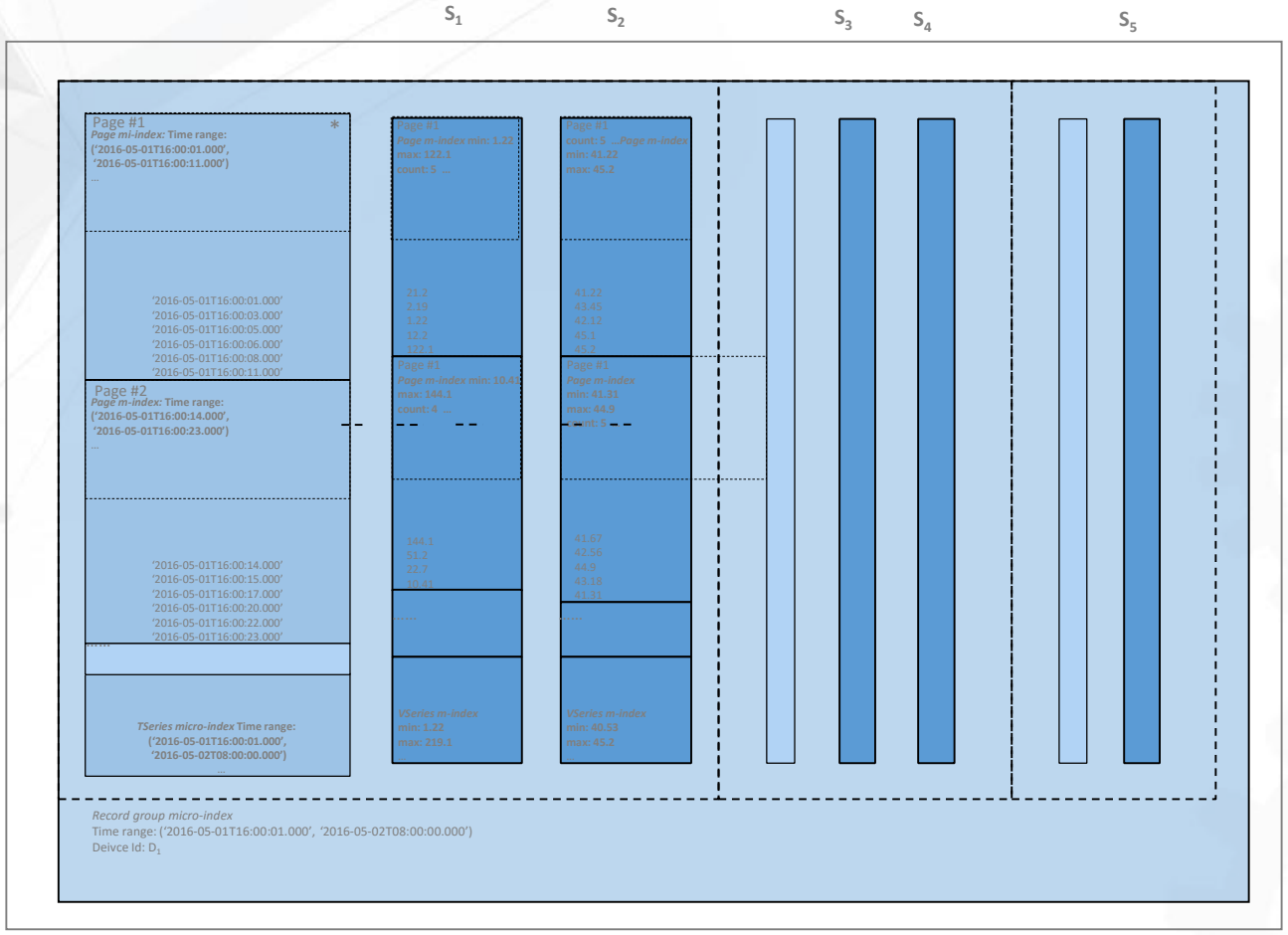
Motivation


A lightweight time series data SPJ & aggregation
Characteristics of time series data
Advance time series analysis
TsfileDB

Architecture



Data Partitioning



Encoding

TS2Diff encoding – Optimized for timestamps

- Unified support of fixed frequency times series or irregular frequency time series

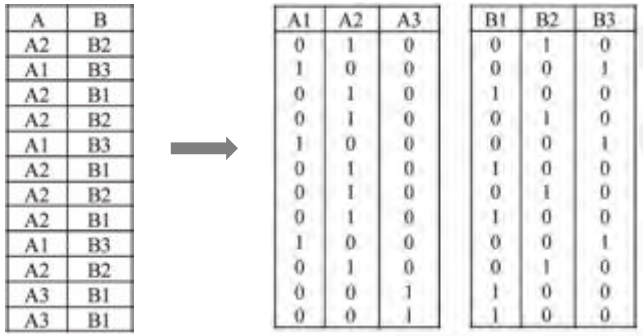
128, 136, 144, 152, 160, …
8, 8, 8, 8 → 1st difference is constant.
0, 0, 0 → 2nd difference is **1-bit** storage needed!

128, 135, 143, 154, 163, …
7, 8, 11, 9 → 1st difference is not constant though
1, 3, -2 → 2nd difference is **2-bit** storage needed!

RLE encoding - for consecutively repeated values

BitPacking encoding - for squeezing out wasteful bits when storing switch values

Bitmap encoding – for enum-type values



A	B
A2	B2
A1	B3
A2	B1
A2	B2
A1	B3
A2	B1
A2	B2
A2	B1
A1	B3
A2	B2
A3	B1
A3	B1

A1	A2	A3
0	1	0
1	0	0
0	1	0
0	1	0
1	0	0
0	1	0
0	1	0
0	1	0
1	0	0
0	1	0
0	0	1
0	0	1

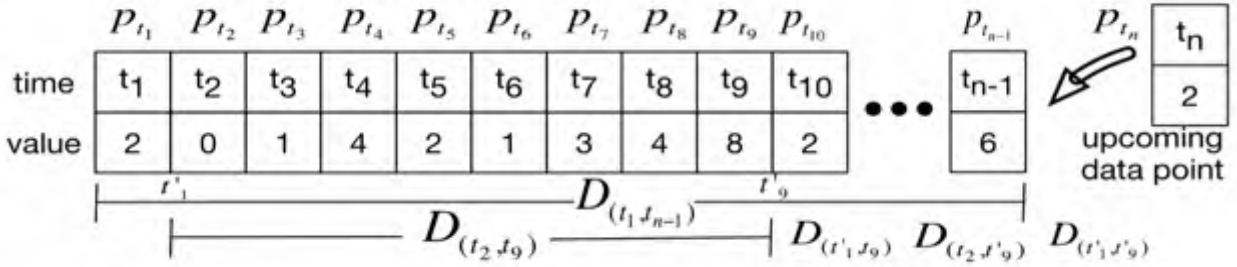
B1	B2	B3
0	1	0
0	0	1
1	0	0
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0
1	0	0
1	0	0

Delta encoding – for timestamps, values with a stable increasing/decreasing step

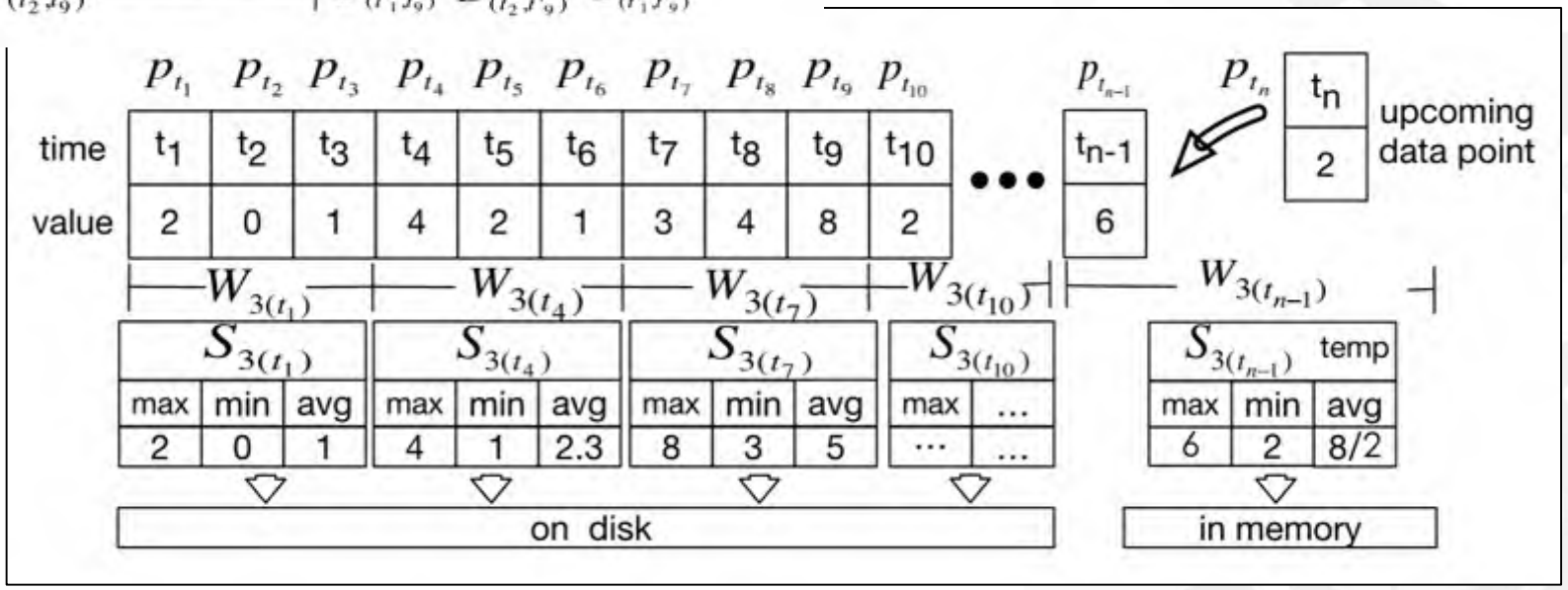
Dictionary encoding - for values with a almost fixed set

… …

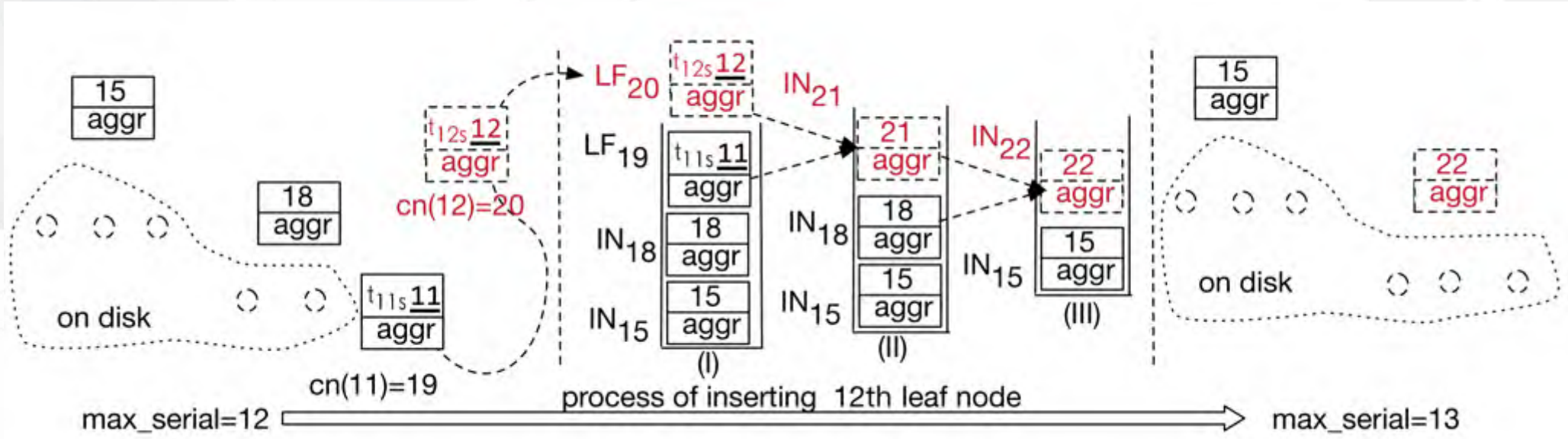
Window summary



Pre-calculate
window summary

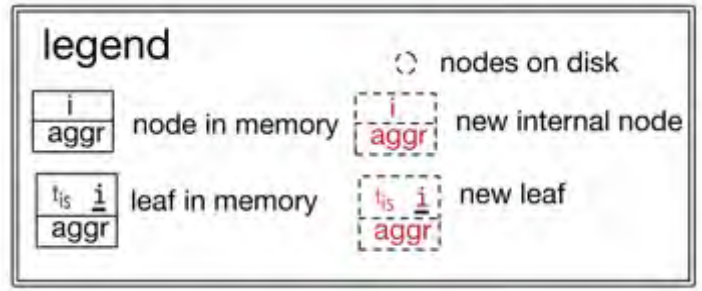


Insertion of PISA

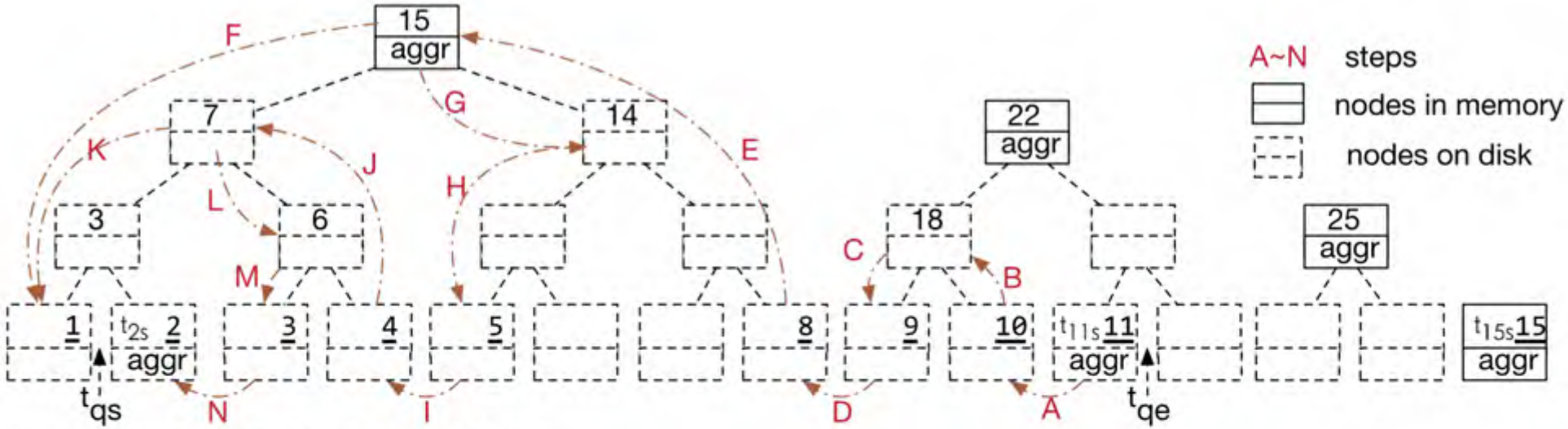


Combination of trees

$$(1) Tree_i, Tree_j \in PISA \wedge i < j,$$
$$(2)\ \nexists Tree_m \in PISA \wedge i < m < j,$$
$$(3)\ depth(Tree_i) = depth(Tree_j).$$



Query for $Q_{\mathbb{A}(qs,qe)}$



Change history of the candidate set (Cs):

query on disk

$(t_{qs}, t_{qe}) \dashrightarrow (LF_2, LF_{19})$

$LF_2 = IO.gt(qs)$

$LF_{19} = IO.ls(qe)$

Cs= $\{LF_{19}\}$

in memory (ABC)

11->10->18->9

(9>2)

Cs= $\{LF_{19}, IN_{18}\}$

in memory (DEFGH)

9->8->15->1 (1<2)

15->14->5 (5>2)

Cs= $\{LF_{19}, IN_{18}, IN_{14}\}$

in memory (IJKLM)

5->4->7->1 (1<2)

7->6->3 (3>2)

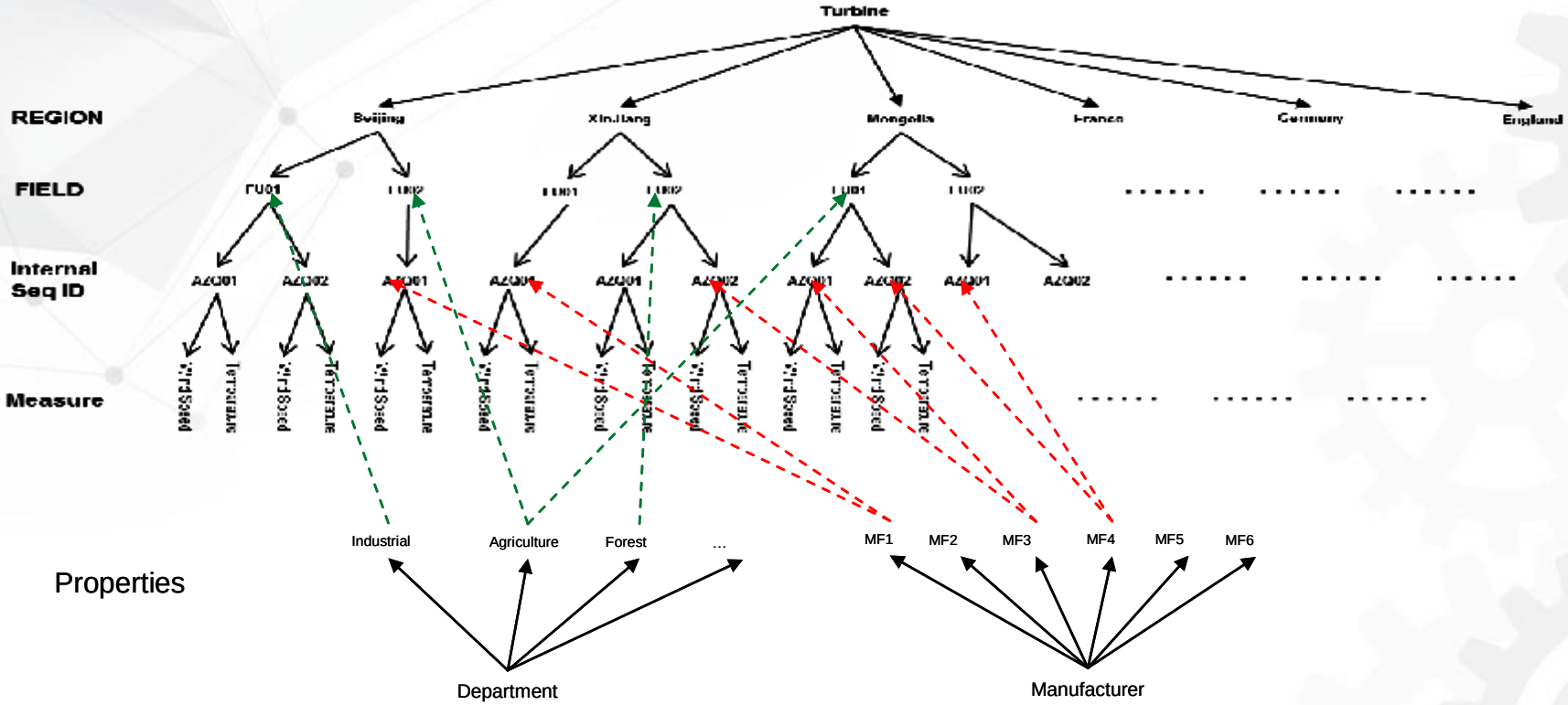
Cs= $\{LF_{19}, IN_{18}, IN_{14}, IN_6\}$

in memory (N)

3->2

Cs= $\{LF_{19}, IN_{18}, IN_{14}, IN_6, LF_2\}$

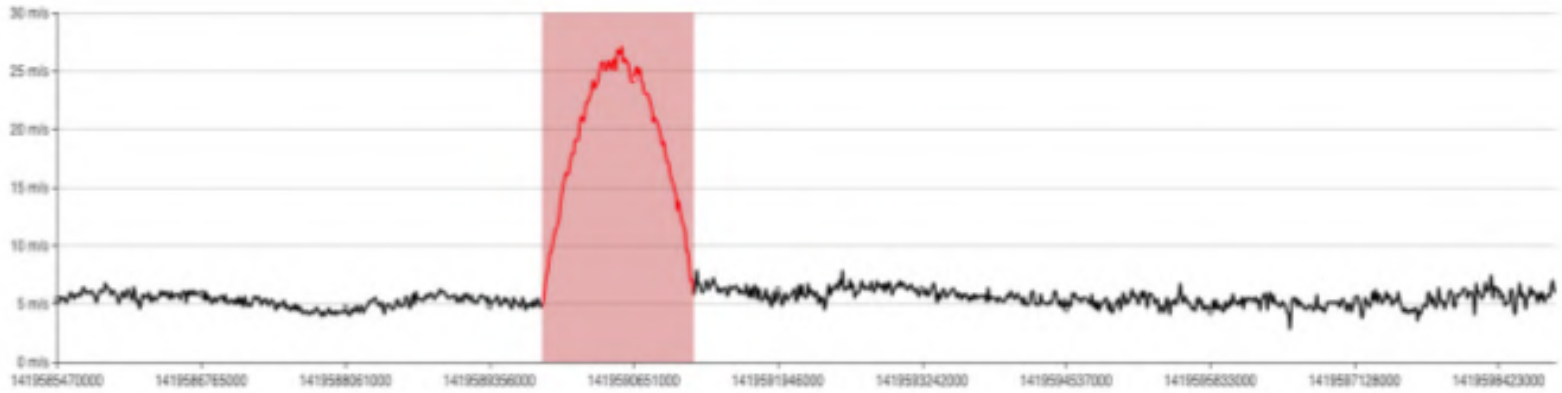
Metadata Model



Properties

Subsequence Matching

- 时间序列子序列相似性查询
  - 异常检测
  - 历史数据分析
  - 传感器数据监测
  - 轨迹识别
  - …
  - 其他时间序列挖掘算法的基础



```
SELECT wind_3s FROM china.farm1.tb2
WHERE time > t1 AND time < t2
AND wind_3s LIKE PATTERN(7.2,..,20.3,..,6.0)
```

Index Structure and Matching Algorithm

- 基于文件的索引结构
  - 比R-tree索引更加高效
  - 适合大规模数据处理
- 查询匹配算法
  - 支持不同长度的查询
  - 提前终止等优化

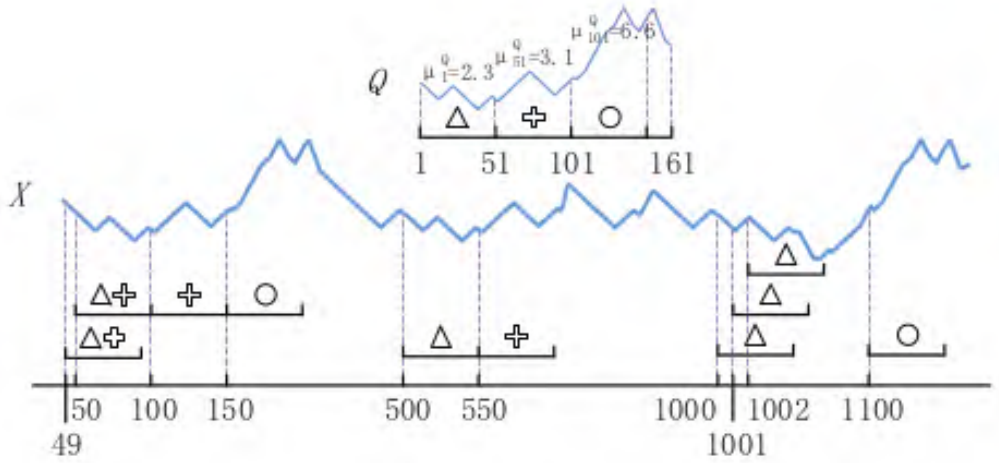


Figure 1: Illustrative example

Index

	μ	Window
△	1.7	1000
	1.8	1001
	1.9	1002
✚	2.4	500
	2.6	50
	2.8	49
	3.0	100
	3.3	550
	...	...
○	6.1	150
	7.2	1100

Table 1: Index structure

Key	Value
[1.5, 2.0)	[1000, 1002]
[2.0, 3.0)	[49, 50], [500, 500]
[3.0, 4.0)	[100, 100], [550, 550]
.....	.....
[6.0, 7.5)	[150, 150], [1100, 1100]

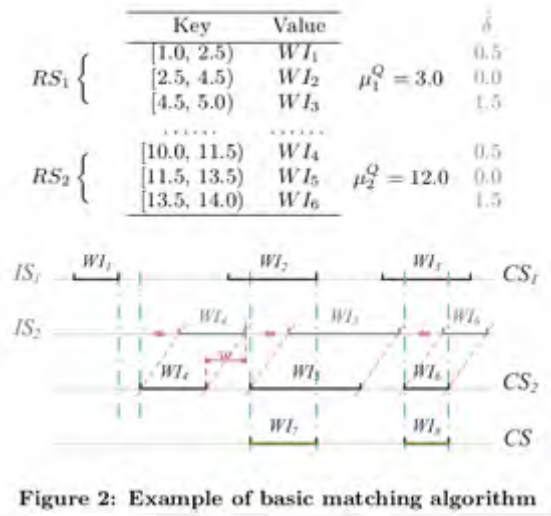
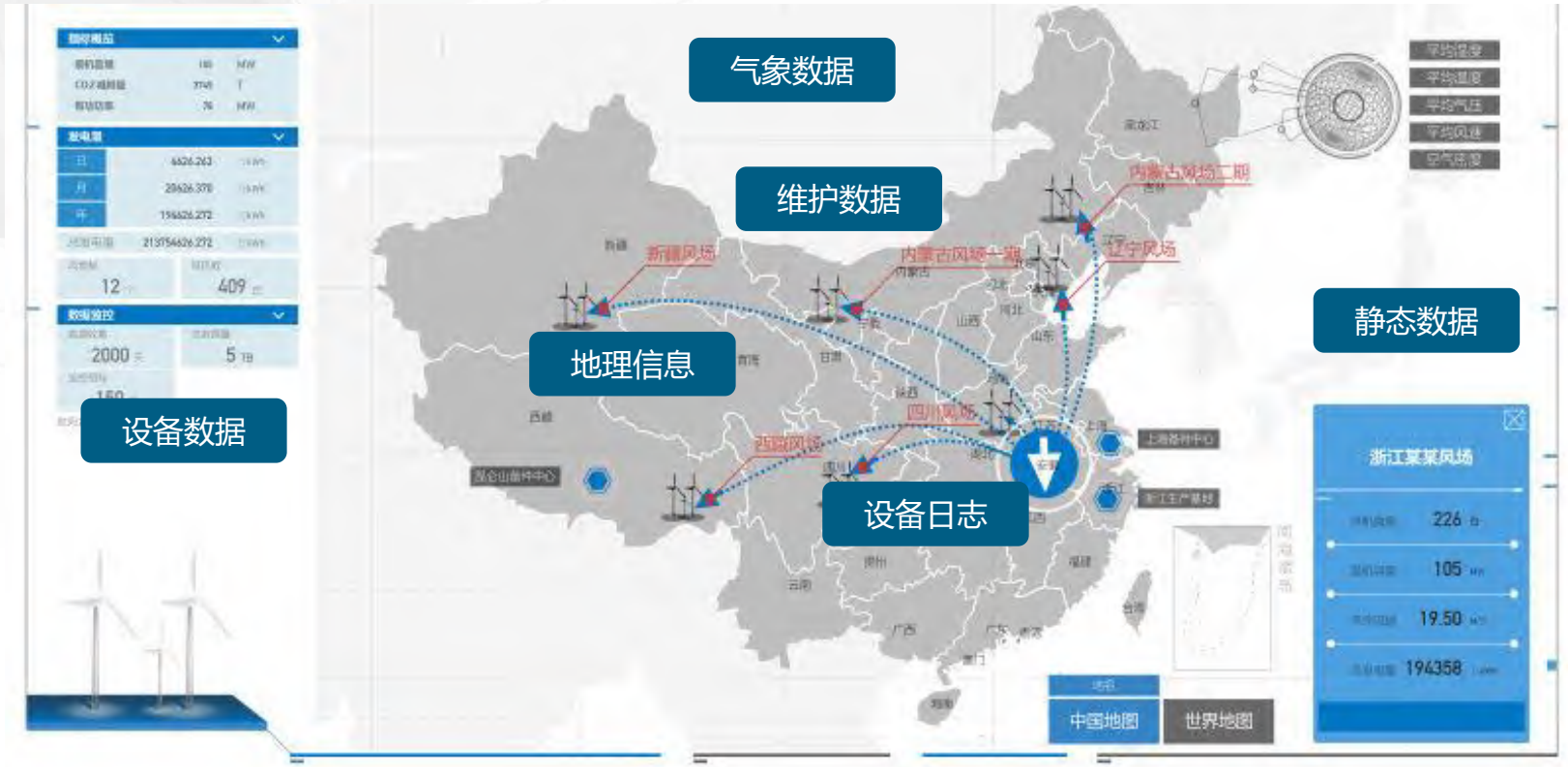


Figure 2: Example of basic matching algorithm

工业大数据应用案例

数据整合

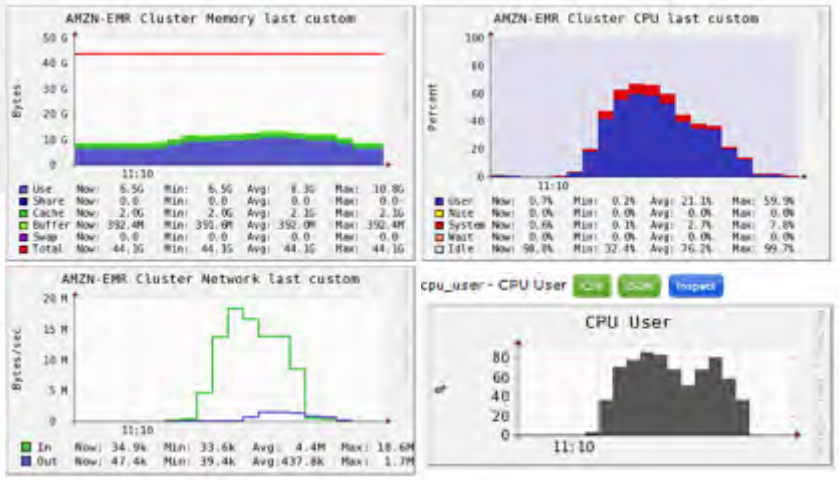


分析模型示例：预测性模型

对风校正优化

风场工况数据的并行化加工

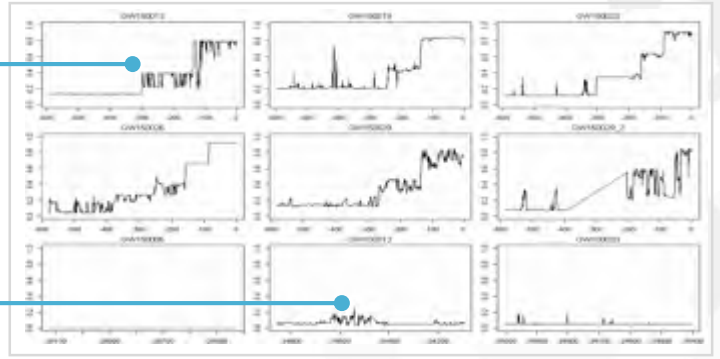
齿形带预警

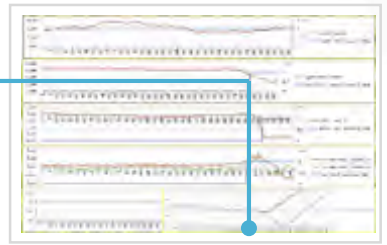


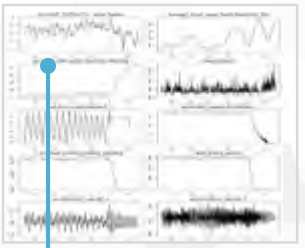
SCADA数据分析模型可在实际断裂前90小时预警

预警模型没有带来虚假预警（底部3台风机至今未发生断裂）

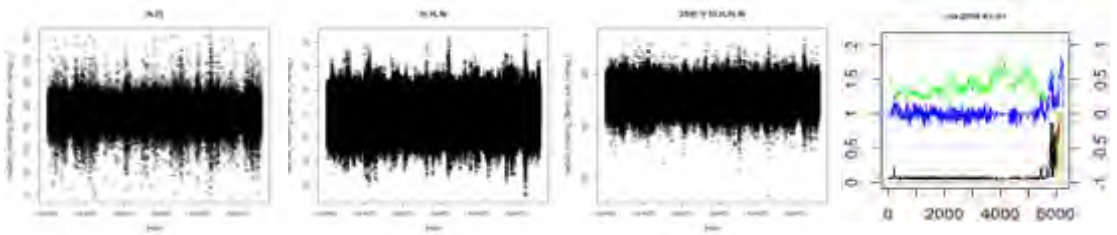
20ms异常模式检测算法可将目前的停机时间再提前0.6s







发现部分风场长期存在的异常震荡

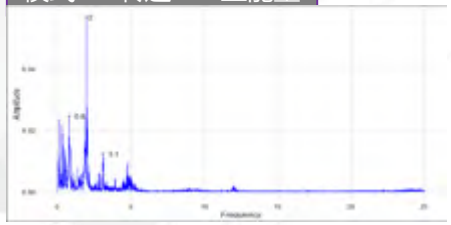


风机叶片结冰预测

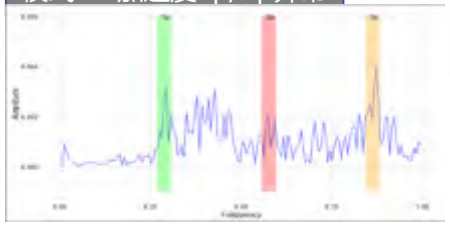
分析模型示例：异常模式检测

叶片失效模式（4种）

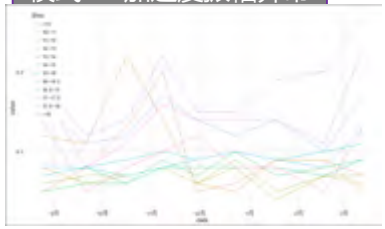
模式1：转速2Hz主能量



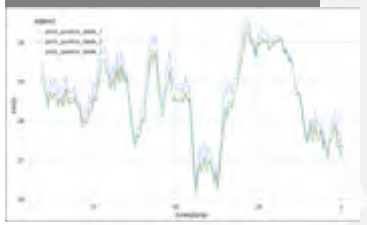
模式2：加速度1p/3p异常



模式3：加速度振幅异常

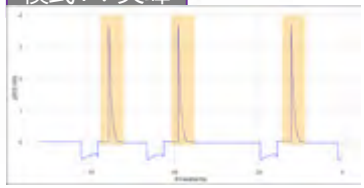


模式4：桨距角跟随异常

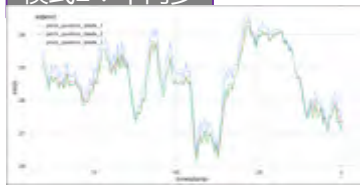


变桨速率异常模式（7种）

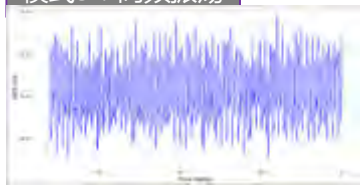
模式1：尖峰



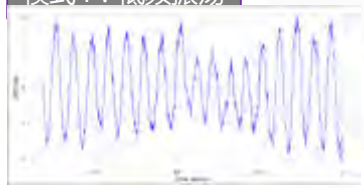
模式2：不同步



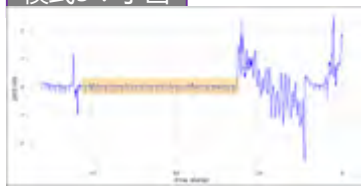
模式3：高频振荡



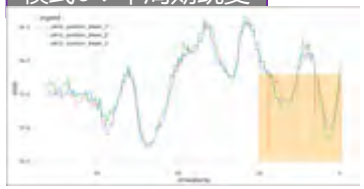
模式4：低频振荡



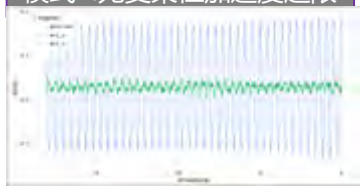
模式5：小齿



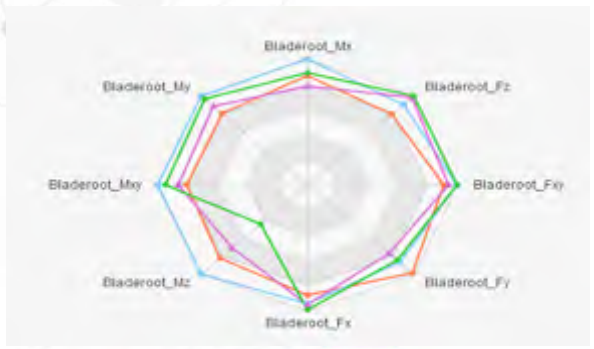
模式6：单周期跳变

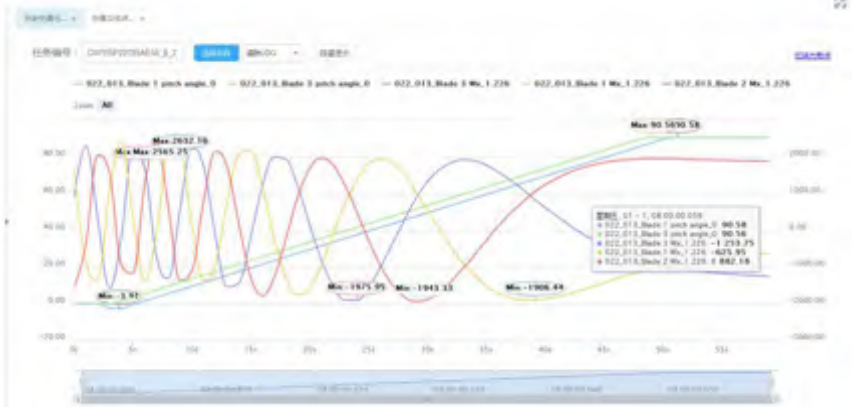


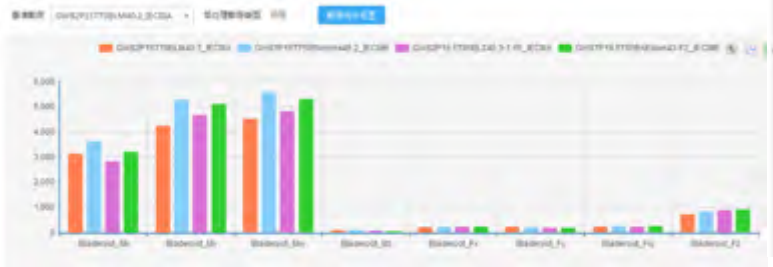
模式7:无变桨但加速度超限



研发数据 — 载荷仿真数据分析

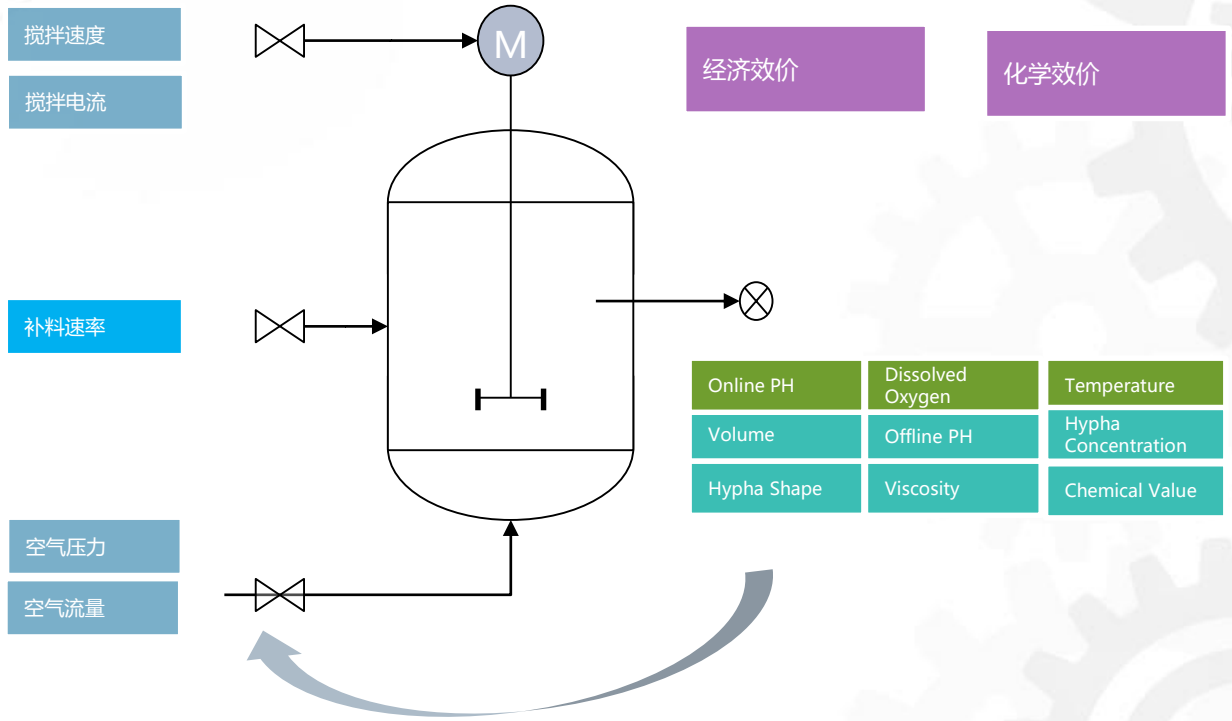






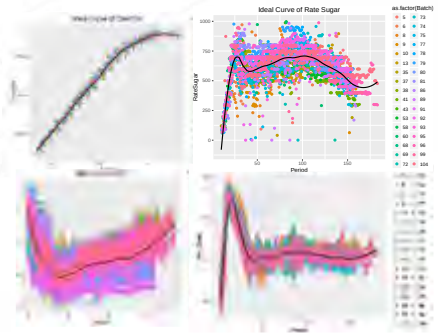
示例：生物发酵过程的智能控制

- 业务目标：智能发酵罐
  - 精益控制
  - 原料产地分析
  - …
- 业务问题：发酵过程的精益控制
  - 最佳经验控制曲线
  - 自适应控制

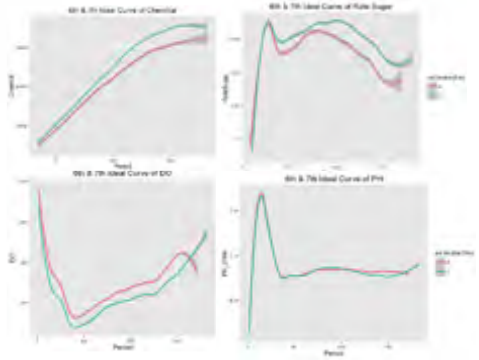


精益控制模型

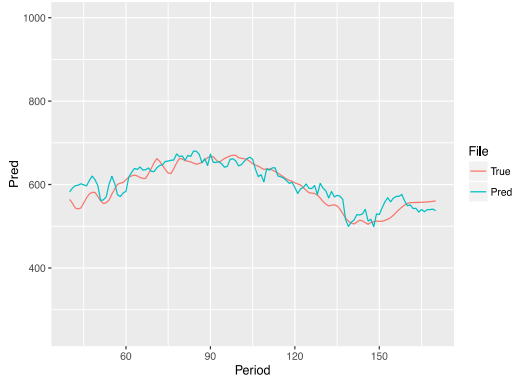
从最佳批次学习经验控制曲线



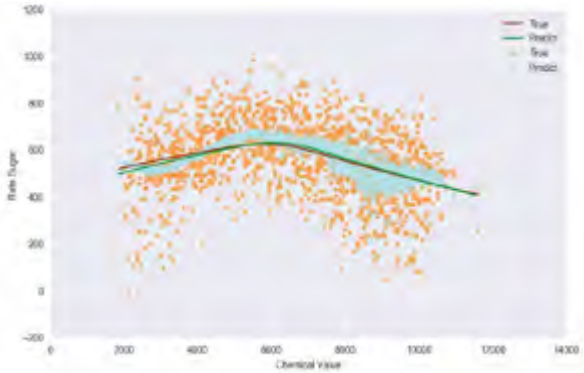
经验控制曲线学习算法可以在线学习
（比如原料发生了改变）



自适应控制效果与专家接近



但数据模型的波动性比专家要小很多



鸣谢

- 清华大学与北京工业大数据创新中心各位同仁

- 国家重点研发计划 “面向高端制造的大数据管理系统” 技术团队
  - 复旦大学、武汉大学、北京大学
  - 中国人民大学
  - 哈尔滨工业大学、西北工业大学

THANKS

SequeMedia 盛拓传媒 IT168.com ITPUB ChinaUnix.net