



# 首届工业大数据创新竞赛成果发布

Industrial Big Data Innovation Competition

主讲人：李铮  
中国信息通信研究院

创新引领 融通发展

2018 工业互联网峰会

INDUSTRIAL INTERNET SUMMIT 2018

创新引领 融通发展

2018 工业互联网峰会

INDUSTRIAL INTERNET  
SUMMIT 2018

# 目录

## Contents

1

工业大数据创新竞赛背景

2

竞赛开展情况

3

《工业大数据竞赛（2017）白皮书》介绍

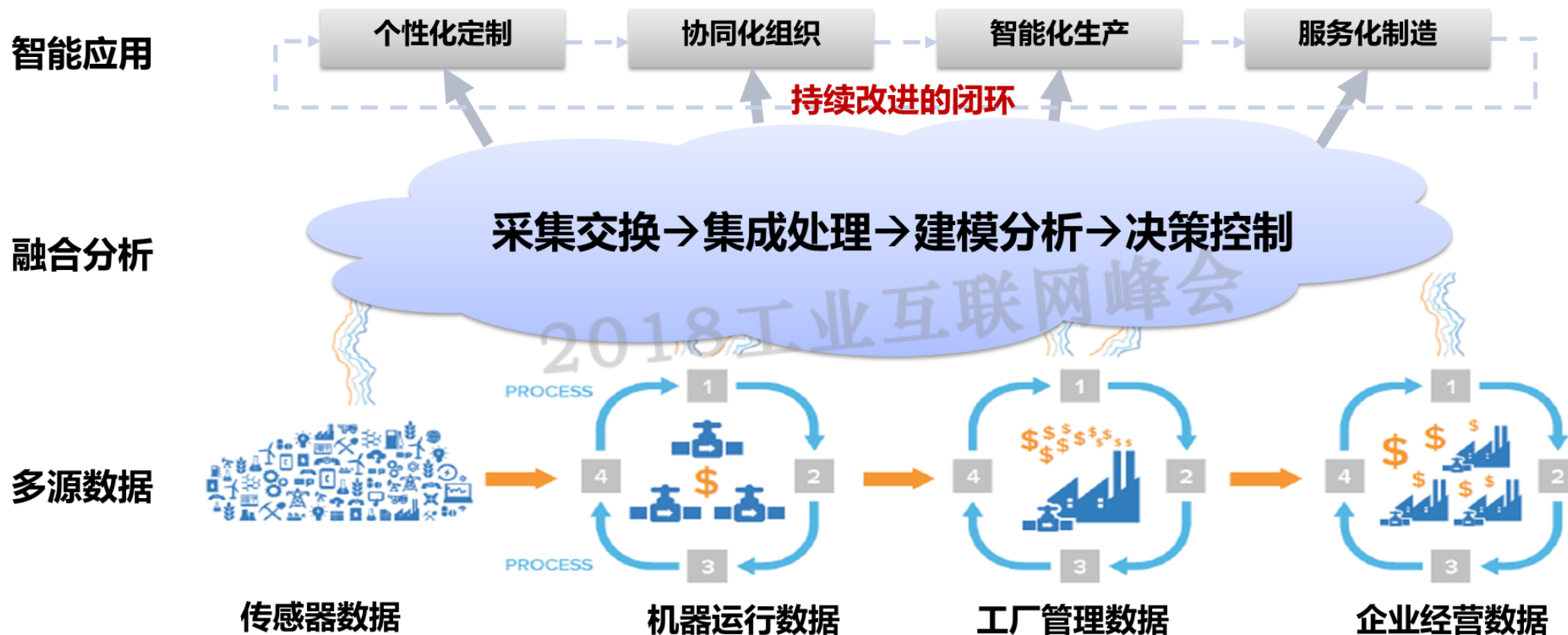
4

总结



# 工业大数据的认识——工业智能中枢

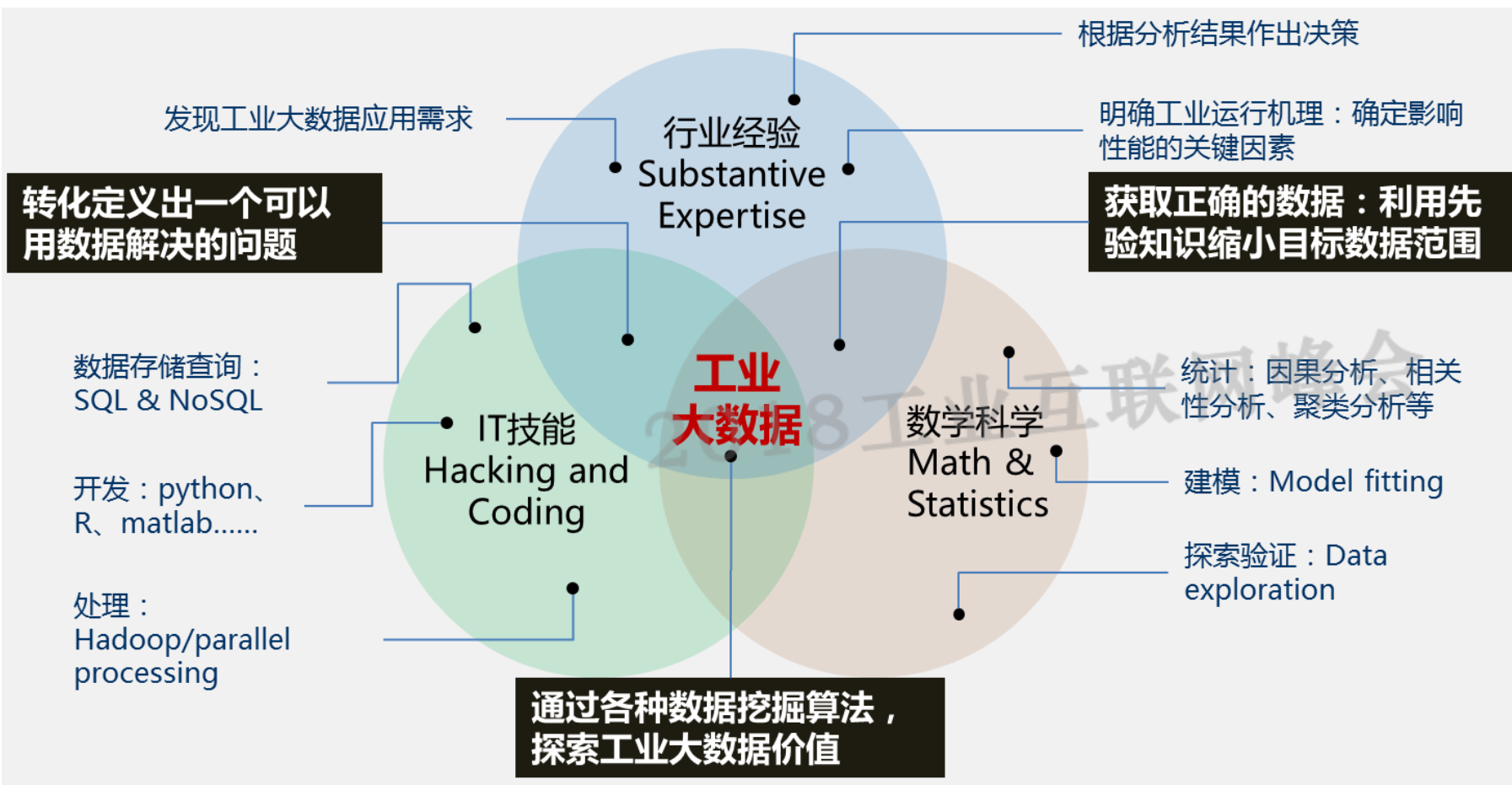
通过传感与智能控制系统开展泛在工业数据的实时采集和智能控制，打通数据获取、传输、分析、决策、执行的闭环系统，形成贯通工业全流程全领域的链条，构造持续改进的智能系统





# 工业大数据发展——多学科、跨领域融合

工业领域有大量机理模型、专家经验的深厚积累，其分析范式更加注重数据科学与行业经验的融合



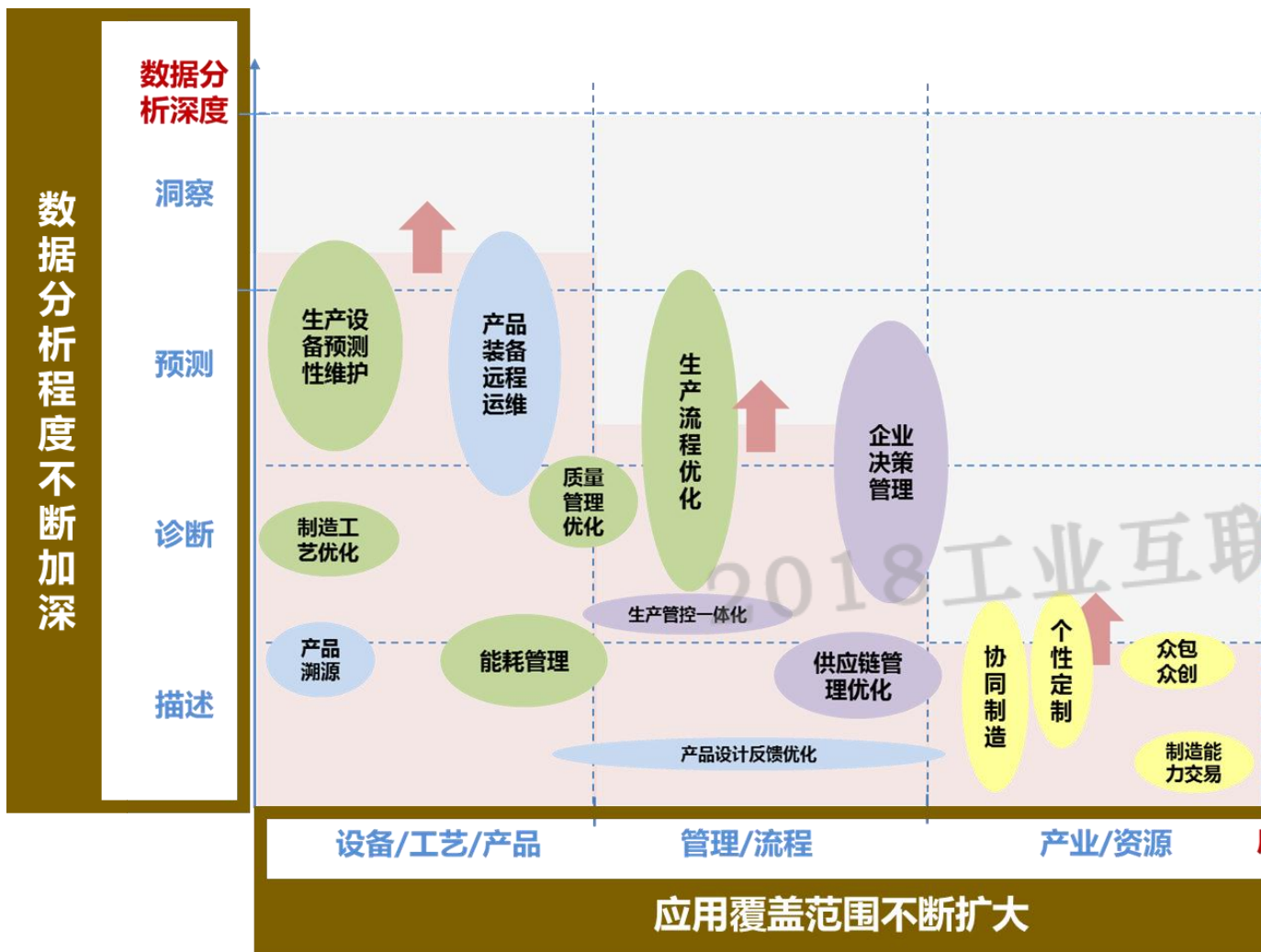
工业数据相比商业数据，具有**数据量大、实时性强、复杂性高、种类多样化**等特点

催生**IT技术、行业经验和数学科学融合、多方合作**的新生态





# 工业大数据应用——发现问题和解决问题



**1、我有数据，能做什么？**

**2、我知道要做什么，怎么做？**



# 竞赛——问题众包和人才众创新模式

## 制造企业提供数据源，求解真实应用问题

### PHM Data Challenge

设备健康状态评估、剩余生命周期预测等问题进行方法研究和测试论证

2017年 城轨车辆悬挂系统故障诊断

2016年 晶元化工一机械平面化系统健康状态追踪

2014年 某电池的剩余使用寿命预测和健康状态评估

2012年 轴承剩余使用寿命的预测

2011年 风力发电机传感器健康估计

2010年 高速数控铣床刀具剩余使用寿命的预测

2009年 齿轮箱关键部件诊断

2008年 涡轮发动机剩余寿命预测

	Flight Quest 1	Flight Quest 2
Industry domain	Aviation	
Data Type	Flight history, status, weather, route data	Flight history, status, weather, route and no-fly-zone data
Task	Predict runway & gate arrival time	Choose optimal
Participants	236 players on 173 teams	257 players across 223 teams
No. of entries	3067	3841
Competition length	4 months	6 months
Winning Method	Ensemble of gradient boosting & random forest models	Optimization Techniques
Prize	Total prize pool of \$250k	Total prize pool of \$250k



GE发起过多次数据竞赛，悬赏解决飞行路径规划、医疗大数据等问题

Kaggle大数据算法竞赛（众包）平台



# 中国首届工业大数据创新竞赛

**赋能与赋智  
构建工业大数据应用生态**

指导单位：工业和信息化部  
2017年7-12月

2018工业互联网峰会





# 首届工业大数据竞赛——主题

本次竞赛主题主要集中在设备故障预测和健康管理（PHM）领域

## 1 制造业与互联网融合特征明显

- 设备健康管理除了先进的算法工具以外，更重要的是结合工业场景和应用原理的领域知识，参赛者不仅要对智能算法非常了解，也需要学习设备的基本运行原理

## 2 理论基础较好，各领域参赛选手迅速熟悉相关机理

- 近几年美国举办的工业大数据竞赛主要围绕该领域进行。我国部分先进制造企业已将PHM作为工业大数据应用的重要方向，尤其是在能源、航空航天、交通等高端装备领域应用最为领先

## 3 该领域的数据分析结果评价指标体系成熟

- 终竞赛结果主要以准确率作为评价标准，最接近真实值的即为获胜方。如果出现一致值（概率非常小），则引入专家团队对分析方法进行评价

风机叶片结冰预测

风机齿形带故障分类





# 首届工业大数据竞赛——合作



**数据提供方：金风科技**

**北京工业大数据创新中心**

**北京天泽智云科技有限公司**

**选题、清洗、验证、评审、总结**

**星河互联**

**美国国家仪器 (NI)**

**美国国家科学基金会智能维护系统中心 (IMS)**



# 首届工业大数据竞赛——创新探索、实践指导与人才发展



## 创新探索

- 这次比赛的数据都是来源于真实的工业场景，数据来源非常宝贵，作为一个非常好的研究基础，有利于创新并做一些理论研究



## 实践指导

- 有益于产学研用紧密结合的紧密结合，加速技术转化落地，可以切实地解决很多工业中大家面对的一线问题，去到工业企业现场解决最头疼的问题



## 人才发展

- 从这次竞争的激烈程度来看，能发掘出很多优秀的工业大数据建模人才，可以将这些人才与更多的发展平台进行对接，为中国的工业振兴发挥更大的价值



# 首届工业大数据竞赛——成果

基于真实工业大应用场景，面向全社会通过竞赛方式征集解决方案

**13** 台风机 **6** 个月数据 **28** 变量

风机结冰故障

风机齿形带断裂

**1535** 人次参赛

**60%** 高校

**40%** 企业或其他

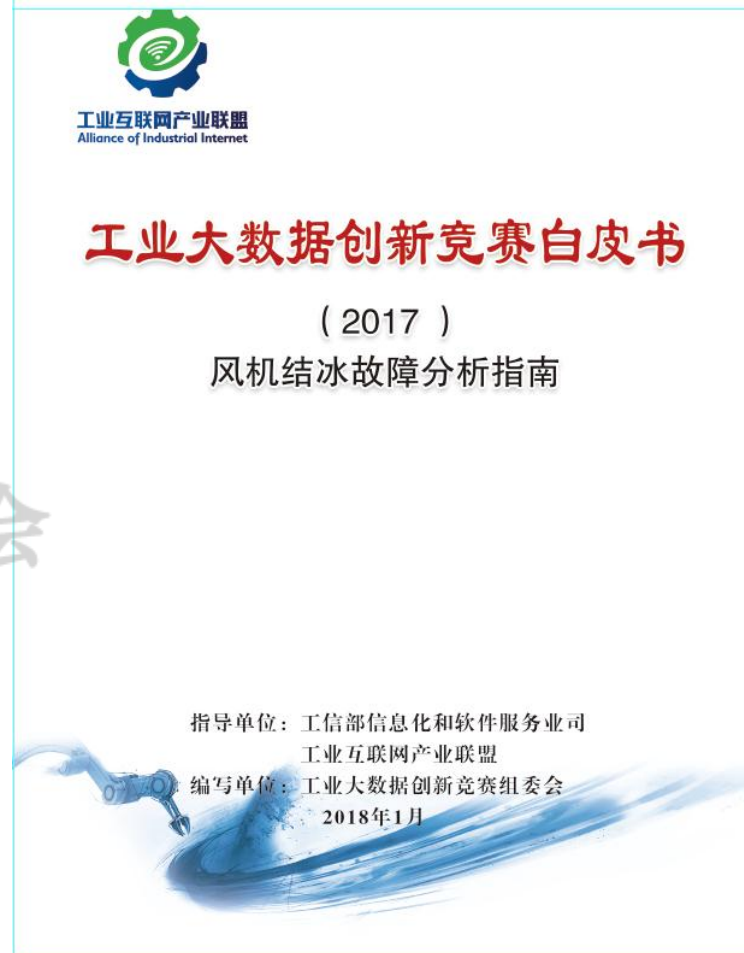
**12** 支优胜团队

**25%**

**75%**

形成工业大数据分析方法论

形成风机结冰故障分析解法集





# 工业大数据竞赛白皮书——编写

一、工业大数据创新竞赛概况 .....	1
(一) 数据经济的崛起与工业的变革 .....	1
(二) 工业大数据驱动制造业转型升级 .....	2
(三) 工业大数据创新竞赛开展情况 .....	4
二、工业智能分析方法论 .....	6
(一) 工业智能分析方法流程 .....	6
(二) 案例-风机结冰故障 .....	12
三、首届工业大数据创新竞赛解法集 .....	27
(一) 基于 CNN-LSTM 深度学习网络的风机叶片结冰预测 ...	27
(二) 基于物理原理+KNN 分类的混合预测模型 .....	41
(三) 基于领域知识特征构建和未来结冰概率估计的风机叶片 结冰预测 .....	48
(四) 基于数据驱动和非均衡数据学习的故障预测研究 ...	62
(五) 基于敏感特征的风机叶片结冰预测算法 .....	75
四、方法论总结 .....	94

《工业大数据创新竞赛（2017）白皮书》主要收录了工业大数据分析的**方法论与获奖的竞赛解法**，通过对建模思路及结果的对比，总结工业大数据建模的独有挑战、分析策略、与经验教训，达到**固化与推广竞赛成果、积极促进成果转化**的目的





# 工业大数据竞赛白皮书——工业大数据建模的挑战

## Broken

缺乏全面工况的数据，而且通常状态不平衡现象严重

## Background

对领域知识的要求很高，其作用体现在数据接入、预处理、建模、结果解释等各个方面



**CAUTION: BAD DATA**

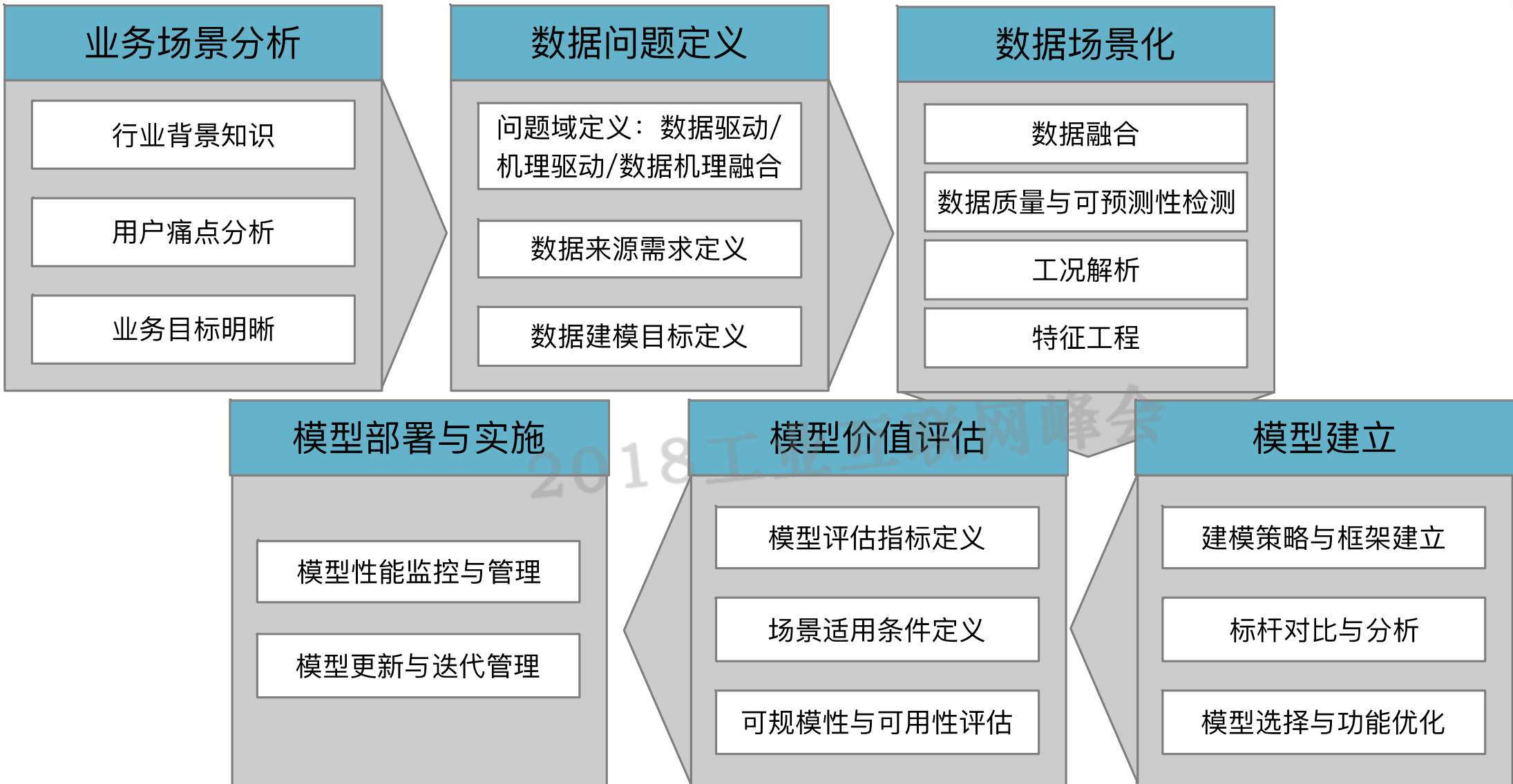


## Bad-Quality

工业数据建模对数据质量要求更高



# 工业大数据竞赛白皮书——工业大数据分析方法论





## 工业大数据竞赛白皮书——收录解法集

- 本白皮书收录了竞赛中第一个题目的5个获奖解法集，分别由这5个参赛队提供了他们的详细解法与结果讨论。

#	方法	团队
1	基于CNN-LSTM深度学习网络的风机叶片结冰预测	万腾科技
2	基于物理原理+KNN分类的混合预测模型	济中节能
3	基于多神经网络的结冰概率预测	富士康科技
4	基于领域知识特征构建和未来结冰概率估计的风机叶片结冰预测	浙江运达风电股份有限公司
5	基于数据驱动和非均衡数据学习的故障预测研究	北京邮电大学
6	基于敏感特征的风机叶片结冰预测算法	西安交通大学



# 工业大数据竞赛白皮书——解法特征

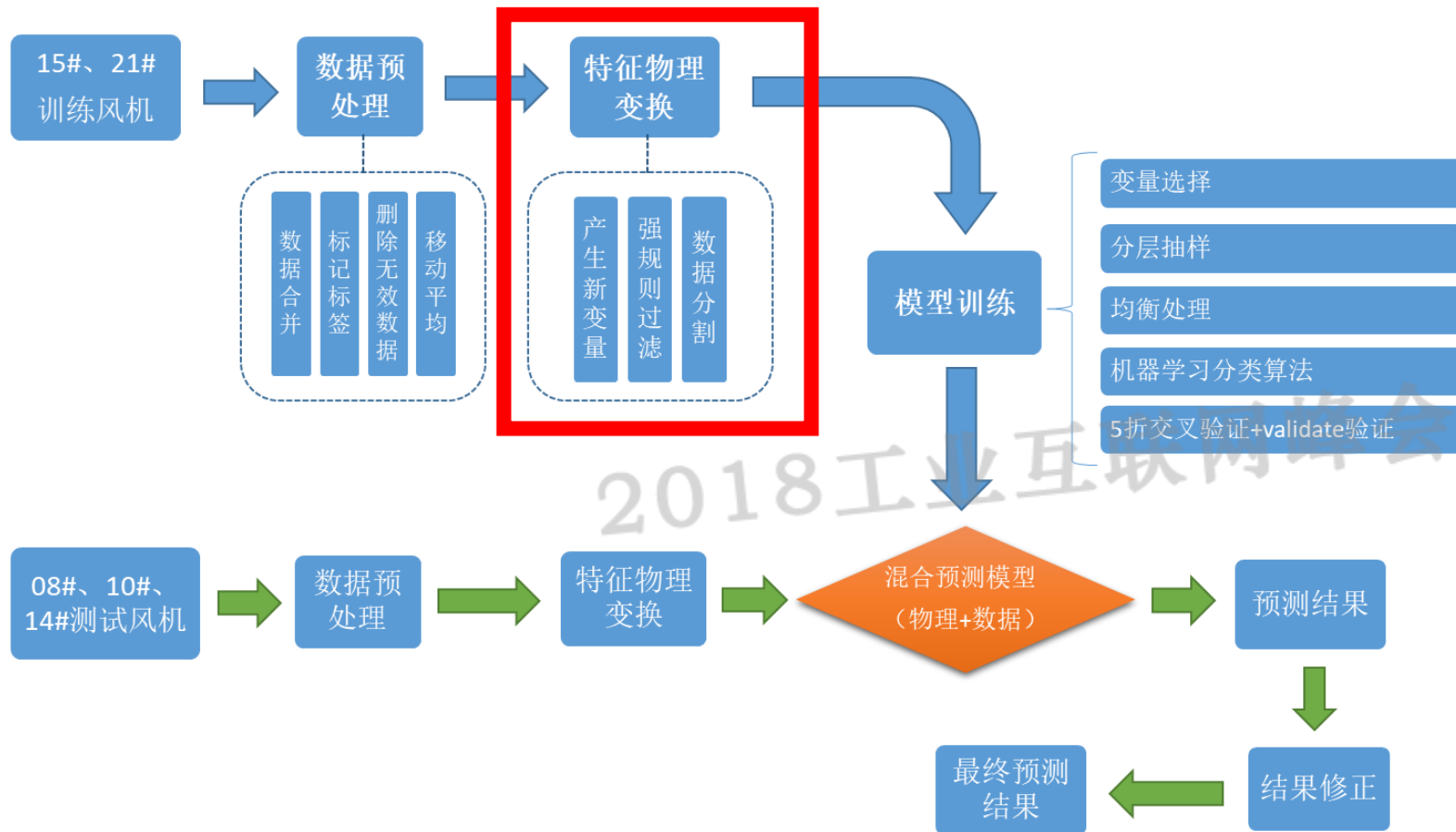






# 工业大数据竞赛白皮书——融入机理代表解法

## 基于物理原理 + KNN分类的混合预测模型 - 济中节能

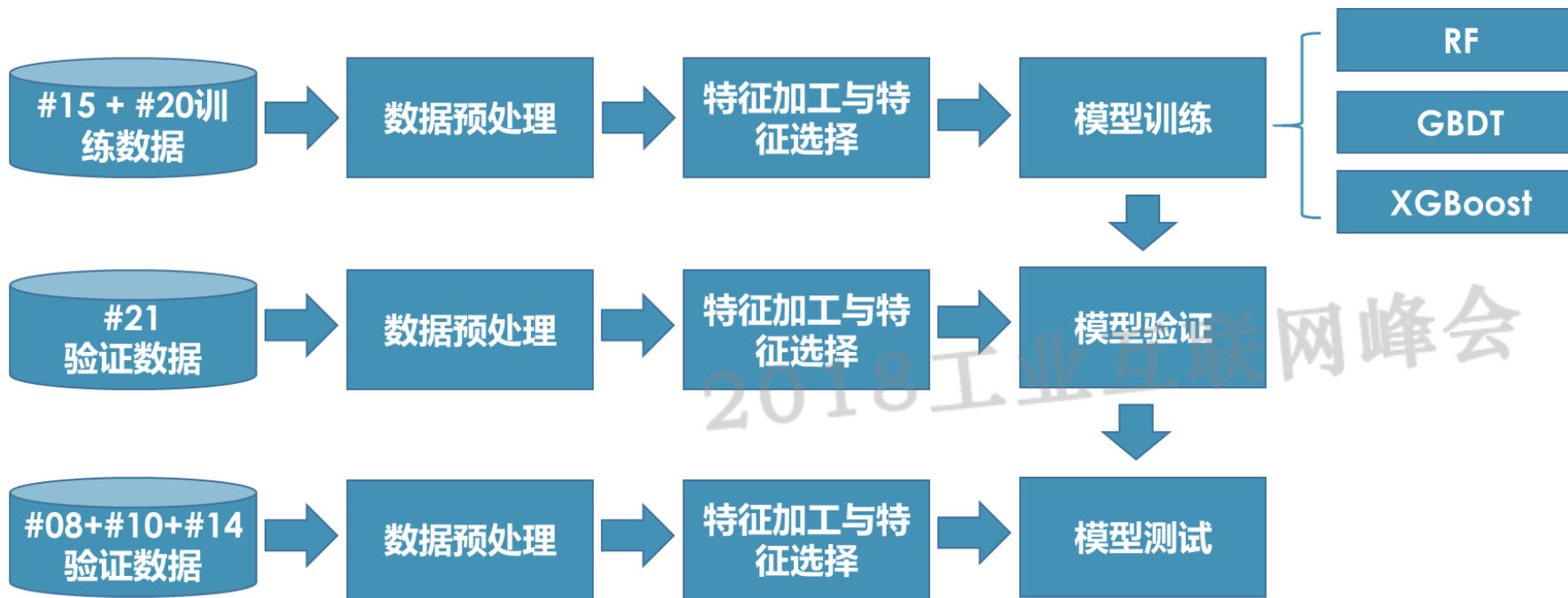


济中节能机理融入的接口在特征这一层，该团队的模型融入了风机领域知识，主要目的是在数据量不足的情况下增强模型的泛化能力，通过增加机理变量、强规则过滤和数据分割的方法对数据进行处理，所以最终取得比较理想的效果



# 工业大数据竞赛白皮书——数据驱动代表

## 基于数据驱动和非均衡数据学习的故障预测研究 - 北京邮电大学



学术组第一名北邮则是从数据的角度直接切入，在前期的数据观察上做了很多工作，同时也注意到了数据严重不平衡的现象，并给出了相应的应对方法。模型从机器学习角度出发，采用集成模型平均的建模策略，建立结冰预测模型



## 总结

数据的预处理地位在上升，数据的预处理对最后的预测结果影响不亚于模型本身

融入领域知识之后往往可以增强模型的泛化能力与准确性，其结果通常会优于没有领域知识的纯粹机器学习模型，尤其在数据质量不好的情况下

在增加机器学习模型复杂度后，其准确性有可能与“机理+简单模型”的结果相当。但是，其代价是模型训练时间随着其准确度的上升而显著增加，并且复杂性增加后其结果可解读性变差

2018工业互联网峰会

**《工业大数据创新竞赛白皮书（2017）》希望传递的理念是，对于不同的工业应用问题，要根据数据的质量、完整性、以及该问题对模型性能与准确度的要求，来灵活选择不同的建模策略，紧扣业务目标，用工程化的思维不断降低模型中的不确定性**



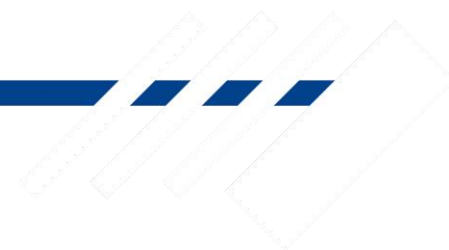
## 致谢

白皮书编写过程中获得了众多专家的指导与帮助，特别感谢孙家广院士、工信部信息化和软件服务业司谢少锋司长、安筱鹏副司长对竞赛开展和白皮书编写的全面指导

同时，美国辛辛那提大学美国辛辛那提大学智能维护系统（IMS）中心李杰主任、清华大学软件学院王建民院长、IIC技术工作组与架构任务组联执主席林诗万、中国信息通信研究院余晓晖总工对白皮书成稿过程中也提出了许多建设性意见，在此一并致谢

感谢北京天泽智云科技有限公司在组织参赛选手编制白皮书过程中所付出的努力





# THANKS

2018 工业互联网峰会

2018 INDUSTRIAL INTERNET  
SUMMIT 2018

主讲人：李铮

2018年2月1日