工业大数据分析的误区与建议

来源：昆仑数据 K2Data（微信公众号：k2datas）
作者：田春华

前言

作为数据价值变现的核心技术手段之一，大数据分析的作用被广泛宣传甚至神化。对于工业大数据分析，产业界存在有不少困惑。

是不是把商业大数据分析照搬过来就是就足够了？只要有了海量数据，大数据分析是不是不需要任何假设前提了？是不是机理模型或领域经验就不重要了？工业大数据分析有没有典型的范式来指导实际操作？

从行业数据分析实践者的角度，本文**上篇**剖析工业大数据分析的常见误区与正确的价值变现之路；**下篇**归纳了工业大数据的典型分析范式，归纳为 6 类算法应用模式、4 种融合模式和 3 类业务应用模式。

上篇：工业大数据“大，不一样”

在与工业企业的交流中，笔者感受到业界对大数据分析的期望与“神化”。谓之“神化”，是由于大数据应用在国内外实践产生的案例，在提质增效及个性化服务方面，产生的利润与之煽动的蝴蝶效应，让有些工业企业以为只要安装了传感器，能把数据采集下来，就能让数据说话，就能从上千种因素中定位出故障原因，就能精准指导研发、生产、运营。甚至误认为经典的机理模型或多年积累的经验不再重要。

然而脱离机理与领域知识的大数据分析结果常常是“你以为你以为的不是你以为的”。

工业大数据的“小”与“大”

从传统大数据 3V（Volume, Velocity, Variety）或 4V（Veracity）度量角度来看，工业数据当然属于大数据的范畴，在体量上甚至超过互联网大数据[1]。然在数据分析中仍不时感觉到工业数据之“小”，主要体现在 3 个方面。

1）价值密度：王建民教授曾指出[2]，相对于产品图纸、工艺设计等传统“小”数据，工业“大”数据的价值密度低。工业大数据分析无法脱离这些基础信息的支撑，不举小数据之“纲”，难行大数据之“目”。

2）大数据永远是物理世界的“小”样本：以 SMT(Surface Mount Technology)生产线为例，最终产品质量由工艺参数、材料特性、生产设备等上千个参数共同影响，生产检测大数据仅仅覆盖了很小的参数组合空间（curse of dimension）。并且不是所有关键因素都有测量，测量值也不一定能反映分布式参数系统的全部（比如回流焊的温度监测值并不等于电路板的表面温度）。工业数据分析更需要利用先验知识缩小搜索空间，同时保持一种“大胆探索、小心求证”的态度。

3）对分析有直接意义的样本比例通常很小：工业通常是运行在设计的常态模式下，对不期望的干扰因素会进行很多压制，造成绝大部分数据对应非常相似的环境与过程。特别对于故障分析、残次品因素分析等大数据分析，样本不均衡程度非常高(biased data)。虽然物理系统相对社会系统更容易做一些控制性实验，但由于很多工业领域控制实验（比如风机叶片断裂、油气管道泄漏等）成本或风险太高，实际上也很难提供足够的异常情形样本。

因此，工业大数据的“大”不能仅从数据量、数据类型、产生速度、质量等角度来看，而应考虑以下两个方面。

1）维度之大：风力发电机组的健康分析应该从时间（过去故障记录、整机性能演化等）、空间（相同机型在不同风场的表现）、环境（气象、地理）、业务运作（设计、维修、限电等）等多个维度综合来看。独立看似异常的事件，很多其实是正常业务操作引起的（如风机功率低可能是由于启动限功率运行模式以降低对居民区的影响）。对于工业数据，更应构建全面的上下文(context model)，才有可能分析出一些有价值的结果。

2）先验知识基础之大：工业领域通常有大量的机理模型、专家经验的深厚积累，可以为数据分析缩小参数空间、提供有用的特征变量（如齿轮箱震动的倒谱参数），数据分析也应思考如何有这些基础更好的互动与融合，以期创造更大的价值。

工业数据分析与商业数据分析：一字之别？

当前很多流行的大数据理念来自于互联网和商务领域，不少分析技术也是针对商业大数据。但工业大数据与商业大数据在很多地方存在比较大的差别，郭朝晖等行业专家对此从不同角度进行了深刻剖析[2,3]，我们将其归纳为如下表所示的四个维度[4]。

	工业大数据	商业大数据
研究对象	以物理实体与环境为中心 （Cyber-Physical-People）	以互联网支撑的交互 （Cyber-Cyber-People）
现有基础	中/微观机理模型与定量领域知识 在当前基础上前进“半”步都很困难	宏观理念与定性认识 存在广阔的提升空间
新驱动力	新的感知技术 产品的服务化转型	新的交互渠道 （如社交媒体）
对分析的期望	因果关系才有用 模型的高可靠性 （很难接受概率性的预测）	相关性关系就非常有帮助 大数原则

1）研究对象不同：工业领域以物理系统（物理实体或环境）为中心，研究动态过程的规律和因果关系，而商业大数据以人造系统（人或流程）为研究对象，试图理解其中的行为模式。当然，工业领域的一些简单产品（如个人电子消费品）制造业和商业产品在产品定义、营销和售后有不少相似之处，但对于复杂产品（如高端装备、高精度制造），区别是非常显著的。

2）现有基础不同：在工业领域，人们对生产过程的研究一般比较深入，形成了很多系统化的中观、微观机理模型，领域知识也比较丰富。客观来讲，对物理系统本身的突破性知识发现难度很大。工业数据中体现出来的规律常常难以突破现有生产技术人员的认知范围。与之相比，商业领域中仅存在一些宏观理念，定性描述人的行为偏好和经济活动规律，给大数据分析留有广泛的提升空间。

3）新的驱动力不同：感知技术的发展和普及是工业大数据的驱动力，现有的工控技术很难处理大数据量的挑战，大量的监测数据也为大数据分析带来与业务数据融合分析的机会。而互联网的发展为企业带来与客户交互的新渠道，极大

促进了商业大数据分析的发展。工业领域的大数据大多是具有时空信息的结构化数据，且背后有明确的物理结构（如系统动力学、网络拓扑关系等），对时间序列、时空模式、序列模式等结构模式挖掘非常重要。而商业大数据分析大多集中在结构化的数据仓库表或非结构化数据（如文本、视频），数据间除了实体关系和部分时空信息外，结构性关系较弱。

4）对分析技术的要求不同：工业系统的实时性高，动态性强，对分析结果的精度要求高，很难接受概率性预测，而商业应用常遵循大数原则，概率性的分析就可以为运营提供很大的帮助。不同工业应用场景对技术指标的要求也不同，比如在风机领域，大部件的故障检测报警已经在 PLC 中实现，大数据分析只有提前若干小时的故障预警才有意义；油气管道泄漏检测中，泄漏发生后的及时报警也很有意义，但其要求零漏报、极低的误报(管道深埋地下，误报会给一线工作人员带来很大工作量)；在抽油机监测分析中，可容忍分析算法对一些罕见或复杂故障类型的无法研判（类似漏报），但分析算法可以研判的出示功图异常的的准确率应该是 100%（这样就可以降低 70~80%的重复性工作）。

工业数据分析的价值实现之道

综上所述，工业大数据分析更应该抱着“小数据”的心态，敬畏机理模型和领域经验，把数据分析模型与机理模型充分融合。数据分析对工业领域知识的帮助主要体现在如下 3 个渠道：

1）物理过程和业务过程的融合。能将物理量与经营过程量（如产品质量、生产效率、设备可靠性等）的关系定量化，突破现有生产技术人员的知识盲点，实现过程痕迹的可视化。

2）对于物理过程环节，重视知识的“自动化”，而不仅仅是知识的“发现”。将领域知识进行系统化管理，通过大数据分析进行检索和更新优化；对于相对明确的专家知识，借助大数据建模工具提供的典型时空模式描述与识别技术，进行形式化建模，在海量历史数据上进行验证和优化，不断萃取专家知识，充分利用多维度融合带来的统计显著性（比如个别风场看似偶发的故障，在全体风场上可能有稳定的统计规律）

3）“软”测量。在工业应用中，不同过程量监测的技术可行性、精度、频度、成本差别较大，通过大数据分析，建立指标间的关联关系模型，通过易测的过程量去推断难测的过程量，提升生产过程的整体可观可控。

小结

如前所述，工业大数据分析更应秉承“小数据”思维，尊重机理模型和领域知识，利用数据分析技术手段，披沙简金，释放工业大数据的价值。为更明确指导工业大数据分析软件架构，下篇将从分析算法侧重点、分析模型与机理模型融合方式、业务应用场景等 3 个方面分享工业大数据分析的典型范式，敬请期待。

参考文献

[1]王建民，“大数据与智能制造”， RONG 系列论坛，2016 年 1 月 9 日
[2]王建民，“中国工业大数据的实践与思考”， 中关村大数据产业联盟论坛，2015 年 3 月 26 日
[3]郭朝晖，“别让商务大数据的思路，误了工业大数据”，物联网智库，2015 年 11 月 23 日
[4]王晨、杨良、田春华等，“工业大数据发展历程”，《2015 年中国大数据技术与产业发展报告》第 9 章。

下篇：工业大数据的分析范式

前言

上篇文章解读了工业大数据分析的特点，指出工业大数据分析应该注重与机理模型的融合，充分利用领域先验知识。那么，工业大数据分析是不是存在典型的模式，可促进不同领域分析模型的借鉴和复用？

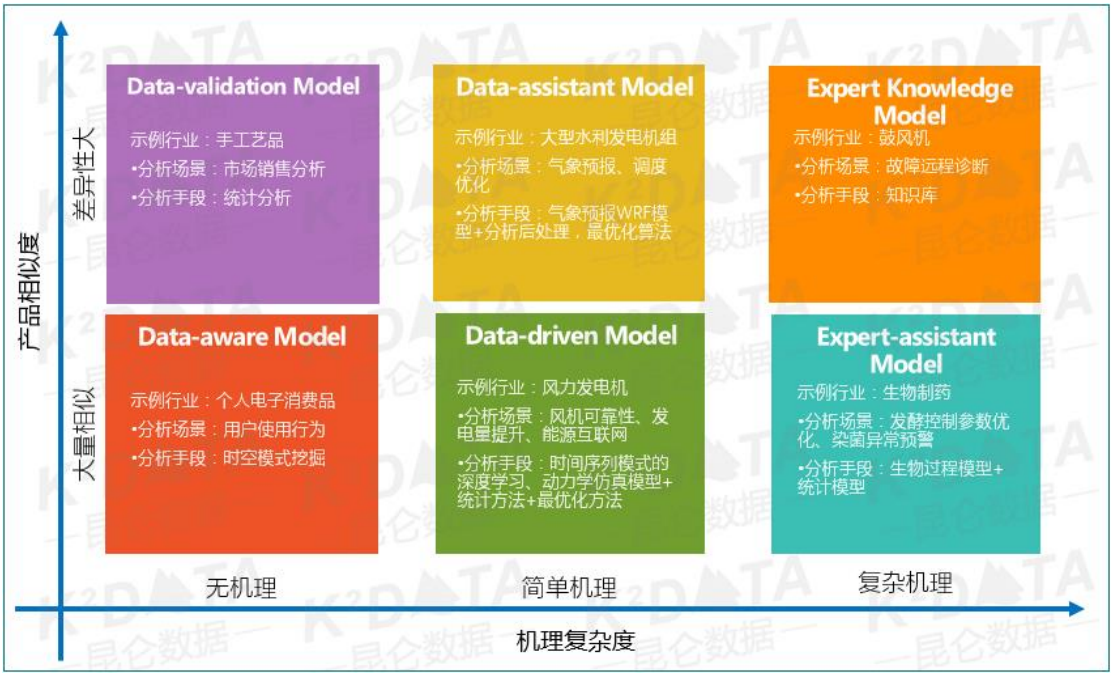
本篇将尝试从分析算法的应用侧重点、分析模型与机理模型融合方式、业务应用场景等三个维度归纳工业大数据分析的典型范式。

6 类算法应用范式

数据分析本质上是一种统计手段，需要足够的样本才有可能发挥显著作用。另外，数据分析作为探索未知的一种技术手段，它的作用也与机理复杂度密切相关。这里从产品相似度、机理复杂度两个维度，将分析算法应用分为 6 类范式。

1）从工业产品的相似度来看，可分为大量相似产品（如风力发电机）和少量定制化产品（如就地建设的化工反应塔）。相似产品在数据分析时可以充分利用产品间的交叉验证，而少量定制化产品应深度挖掘时间维度。

2）从产品机理的复杂性来看，有无需机理模型的 black-box 产品（如电子消费品，通常不会深入元器件内部去分析）、简单明确机理产品（如风力发电机）、复杂机理产品（如鼓风机、化工厂）。复杂机理产品在工业大数据分析时，应更加重视机理模型和专家经验的融入。



6 类算法应用范式图解

4 种融合范式

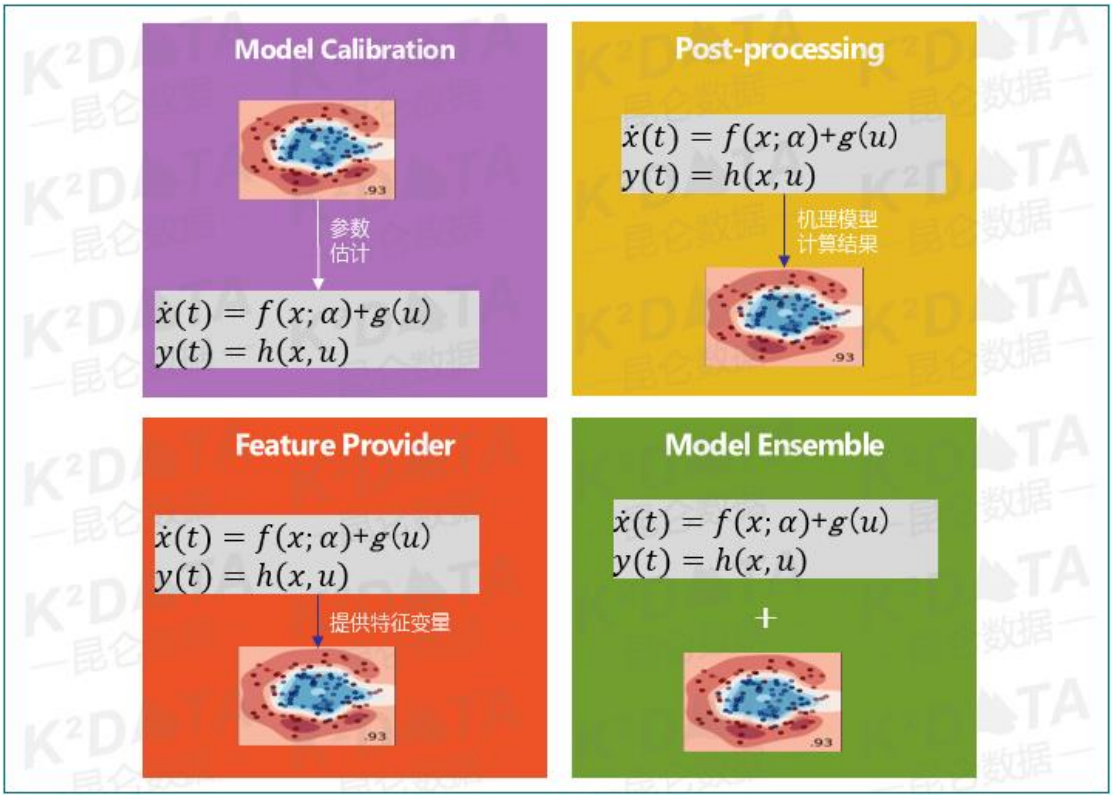
分析模型与机理模型的融合可以分为 4 种范式：

1）分析模型为机理模型做 model calibration，提供参数的点估计或分布估计。例如 Kalman 滤波。

2）分析模型为机理模型做 post-processing。比如，利用统计方法对 WRF 等天气预报模型的结果做修正或多个机理模型综合，提高预测的稳定性。

3）机理模型的部分结果作为分析模型的 feature。例如，在风机结冰预测中，计算风机的理论功率、理论转速作为数据挖掘模型的重要特征。

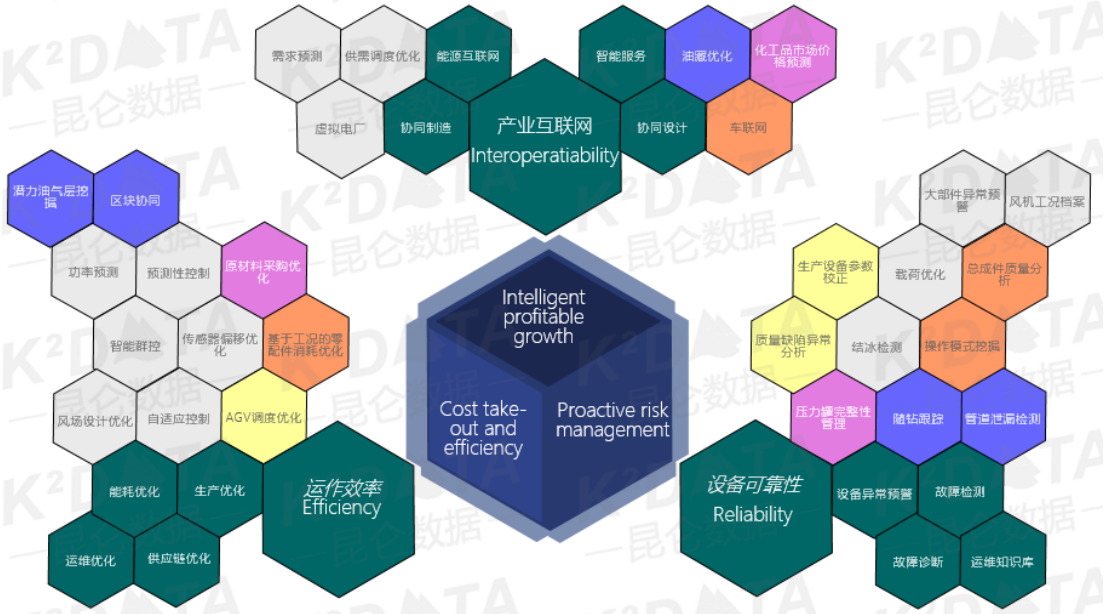
4）分析模型与机理模型做 ensemble。比如，在空气质量预测中，WRF-CHEM/CMAQ 等机理模型可及时捕获全局动态演化过程，而统计模型可对局部稳态周期模式有较高精度的刻画，model ensemble 可有效融合两类模型的各自优势。



4 种融合范式图解

3 类业务应用范式

通过对复杂过程的演化过程和上下文的全面深入刻画，工业大数据对产品/设备可靠性、运作效率、产业互联网等 3 类业务应用场景都有很大促进作用。一些行业的典型工业大数据分析场景如下图所示。



小结

工业大数据分析能否真正落地，取决于能否创造经济价值。价值的持续创造，必须与生产/管理流程和上下文相结合，必须理解工业的特点、工业数据的特征和工业界的特殊要求。

这些特殊性决定了工业大数据分析的思路和方法有别于商务大数据，更应以“小数据分析”的心态，融合机理模型和领域经验。在分析模式上，本文将工业大数据分析归纳为 6 类算法应用范式、4 种融合范式和 3 类业务应用范式，以期促进不同行业分析模型的复用。

作者介绍

田春华：昆仑智汇数据科技（北京）有限公司首席数据科学家。2004 年 1 月清华大学自动化系博士毕业。2004 年-2015 年在 IBM 中国研究院，负责数据挖掘算法研究和产品工作，在高端装备制造、石油石化、新能源、航空与港口等行业，帮助中国、亚太、欧美领先企业，成功实施资产管理、运营优化、营销洞察等各类数据分析项目。发表学术论文（长文）82 篇（第一作者 42 篇），拥有 36 项专利申请（10 项已授权）。研究兴趣是数据挖掘算法与应用。