郭朝晖：工业大数据的特征、方法与价值创造

来源：数据观

作者：数据派

本讲座选自宝钢中央研究院首席研究员郭朝晖于 2015 年 12 月 30 日在清华大数据“技术·前沿”讲座上所做的题为《工业大数据的特征、方法与价值创造》的演讲。

郭朝晖 ,现为宝钢中央研究院首席研究员、教授级高工。分别于 1990、1994、1997 年在浙江大学应用数学、化学工程和自动化专业获得学士、硕士和博士学位。1997 年加盟宝钢，2005 年晋升教授级高工。长期从事信息、模型、自动控制、大数据等领域的技术研发工作。曾先后担任中国工业与应用数学学会副理事长，中国现场统计学会第八届理事会理事，上海工业与应用学会常务理事，上海人工智能学会理事，上海交大、浙江大学、宝钢人才开发院兼职教授，东北大学兼职博导，宝钢集团党外知识分子联谊会会长，上海市知联会理事，并曾担任全国总工会十四大代表，中央企业青联委员。出版《管中窥道 :技术创新的观念与方法》等著作。

主持人王建民：首先代表软件学院、代表清华大学数据科学研究院、工业大数据中心欢迎咱们各位老师、朋友、同学来参加我们今天下午的这个报告会。今天是 2015 年的倒数第二天，我们马上就迎来了 2016，那么在 2015 当中有一个很热的词，在中国，就是“中国制造 2025”。最近大家也看到我们的中央工作会议之后，特别对中国制造这样的一个转型升级的大命题进行了非常深入的探讨。与之对应的大家知道就是工业 4.0 和美国的工业互联网。前两天，也就是这个星期六，清华大学发起成立了叫做工程科技创新联盟，在这个联盟发起的仪式和会议上，苗圩院长、周济院长和清华大学的孙家广院士分别作了主题发言。在这个会议上，周济院长对中国制造 2025 又作了特别深入的解读，中国制造 2025 总结出来的三句话就是智能产品、智能生产和智能服务。在这个会议中，在孙家广院士的主题报告当中特意把工业大数据作为主要的汇报内容，就是信息化和工业化深度融合的抓手，大家看，形态上边是一个工业互联网，但是这个工业互联

网的背后的大脑和智慧的来源还是工业大数据。前面一年来，或者一年多来大家对工业大数据也有这样或者那样的看法，很多人会质疑说工业界有没有大数据，工业界的数据是不是大数据，工业界的大数据怎么样发挥它的价值和作用，我想这都是摆在我们今天中国经济转型发展过程当中不可回避的问题。正好上个月我有幸见到了宝钢的首席分析师郭朝晖郭总，应该说我们有很多的观点非常相近，我觉得特别是郭总在企业一线积累了好多的非常深刻的，但是讲起来又非常生动的例子，我觉得值得到清华大学，特别是和我们的教师、和我们的同学进行分享。当然，今天我也看到好多来自于清华以外的老师和同仁，这也是非常非常高兴的。

郭朝晖：谢谢王老师、谢谢大家。我发现我作报告有个特点：就是我准备得越好，讲得越差。为什么呢？因为报告越重要，准备越认真；但越是重要的报告，我就越紧张。所以，有时候准备得好，讲得反而不好。清华是咱们国家的顶尖大学，在这里作报告我刚到特别荣幸，也感到特别紧张。所以，下面讲得不好的地方请大家多多原谅，因为我实在是很认真地准备的。

常见的“失败”：不了了之

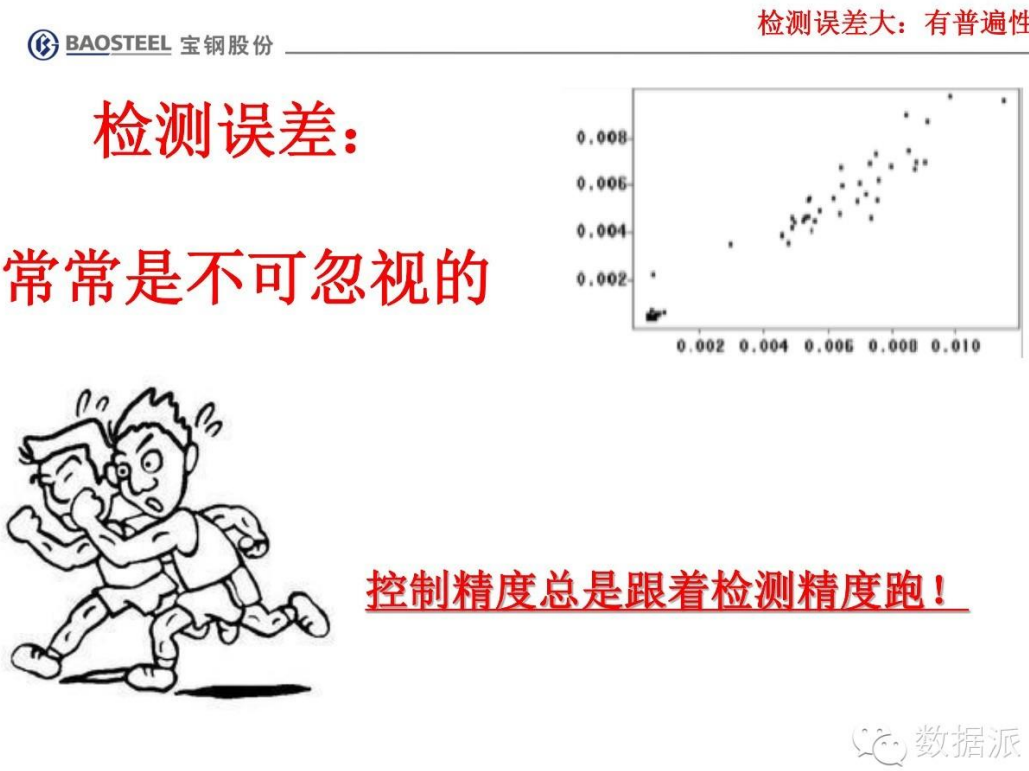
工业大数据以及数据挖掘技术，很少能取得预想中的成功。

关于工业数据处理的问题。我 20 多年前读硕士的时候，我的导师胡上序先生就有这么一个领域，希望通过工业数据的分析来提高我们工业的水平。但几十年下来，我却常常发现这么一种现象：当你立一个项目的时候把它说得非常好，好像什么事情都做得了 ;但当结束的时候 ,你却发现只能得到一个不理想的结果。所以，我们的现实和理想往往有很大的差别。某种意义上来讲，这样的结果就是失败了。这种失败的表现就是不了了之，你不能说他一点都没得到，但得到的跟想象的相差太远。

为什么会有不了了之呢？其实这里面有一个原因，就是我们在谈到数据应用的时候常常说得好的一面。比方说，我们说数据当中有信息、数据当中有知识、数据是有用的。但我们又常常忽略它的另外一个方面，比方说数据有假的、有错的、有偏差很大的，你得到的很多东西可能是局部性的、暂时性的；或许你得到了一个正确的结果，但它却是很平庸的，人家会对你说：“我早就知道了，这是常识，你告诉我有什么用呢？”等等。再就是：你告诉人家一个事，人家说：“真的吗？”你说：“我也不确定”，“算了，不确定我也不敢用。”经常由于这样的一些原因，我们的大数据分析之梦最终不了了之。

目录

今天我给大家汇报的内容大概包括这么三个方面的内容，第一，我先谈一谈工业大数据到底有什么样的特点；再介绍一下我们分析工业大数据当中有什么样的方法，最后针对工业大数据的价值创造，谈一点自己的想法。



工业大数据有什么特点呢？下面我举一个例子。

钢铁行业的学者们很早之前就想建一个模型，用于描述钢铁的成分、工艺和它的力学性能之间的关系。这是人类 60 年多年前的梦想。

从某种意义上说，建立这样的模型很容易：只要有了数据，直接回归就是了。但大家都会发现一个问题：模型的精度总是提不高。而且，更糟糕的是：张三、李四、王五会得出来差异非常大的模型；各种经验公式遍天飞，但是没有哪个人能证明自己的就比别人好。大家都在想：如何才能得到高精度的模型呢？

2002 我接手这个题目。因为我是做数学的出身，首先想到的是模型的存在性：高精度的模型是否存在呢？

很快，我发现：高精度的模型是不存在的。

为什么不存在呢？因为模型输入参数的误差太大，输入参数的误差大，精度自然不可能很高。

我是这样证明参数输入误差高的：同一个参数测两次，观察测量误差。我发现，同一炉钢成分的测量误差，和对应钢种中不同炉次的成分波动基本上处在一个等级上。这意味着：测量误差和测量值的信息量差不多。那么你设想一下：怎么样可能会得到高精度的结果呢？

这就意味着：如果你想得到一个很高精度的结果，证明你比我强得多，这是不可能的。彼得.德鲁克说：做正确的事，正确地做事。而只有能做成的事，才是正确的事。所以，去追求过高精度是错误的。

可能有人会问：测量结果和信息量为什么如此接近？这是偶然的，还是必然的？我想了一下，认为这种事情很容易发生。为什么呢？大工业生产总想越稳定越好、控制精度越高越好。但精度到了一定程度就提不上去了——导致精度难以提升的原因，往往是触及到检测误差的范围。换句话说，检测误差成为控制瓶颈的时候，就不可能再进一步把它的控制精度提高了。这就意味着：在一个工作点附近，检测误差和它本身的分布处于同一个量级上。这样，所以检测误差总是跟着自己的控制精度一起跑。

$$y = bx,\ \hat{y} = b(\hat{x} + \eta) + \xi$$

$$E(\hat{b}) = b\frac{Dx}{Dx + D\eta}$$

检测误差造成“有偏估计”，进而无法外推。
进一步的分析表明，一般会低估50%-70%。

这样的事情会导致很多的问题。比如，过去我们建模型的总有一种观念：模型精度越高越好，精度高到一定程度时，会逼近真实的对象。但是，如果是我刚才说的情况，就不一样了。

估计大家都学过最小二乘法，知道最小二乘法是“无偏估计”。所谓“无偏估计”，就是样本量很多的时候，回归系数就逼近真实。但是，不知大家有没有注意到一个条件：得到上述结论的前提，是自变量的误差可以忽略。

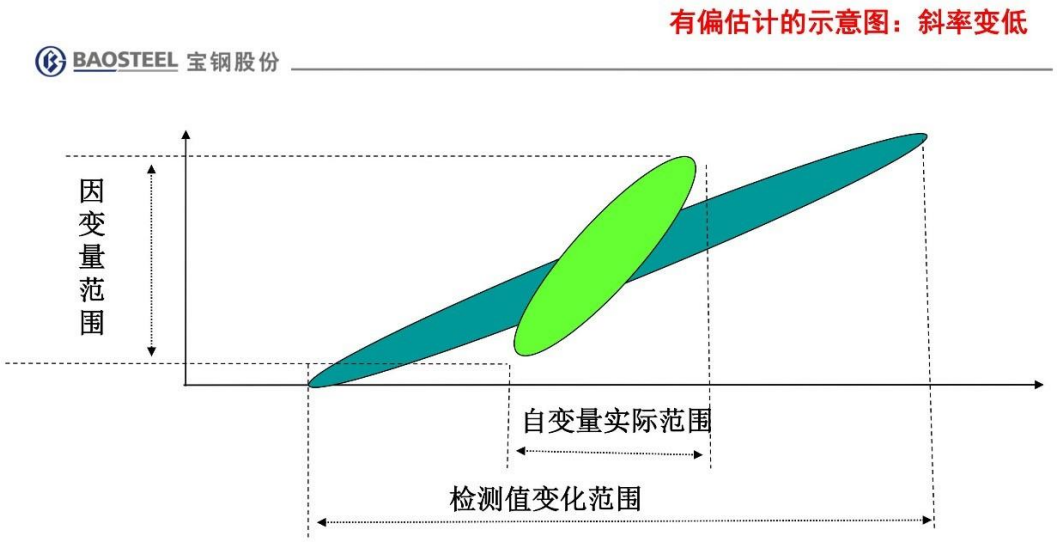
传统的最小二乘为什么没有考虑自变量的检验误差呢？道理很简单，因为回归这套理论，原本并不是用于工业数据分析的。它是产生于实验设计等问题。在那些情况下，自变量的检测误差往往是可以忽略不计的。但是我们这里变得不能忽略不计了。

这会发生什么情况呢？我们假设有一个函数 y=b*x。自变量误差很小时，用最小二乘法得出来 E(b)=b。但是，如果自变量误差显著时，你得到的结果比 b

要小（我指的是绝对值）。这就意味着：误差最小的那个模型其实是不真实的；说的严重点，甚至可以说是错误的。

这个问题我也疑惑了很久：误差最小有什么不好呢？

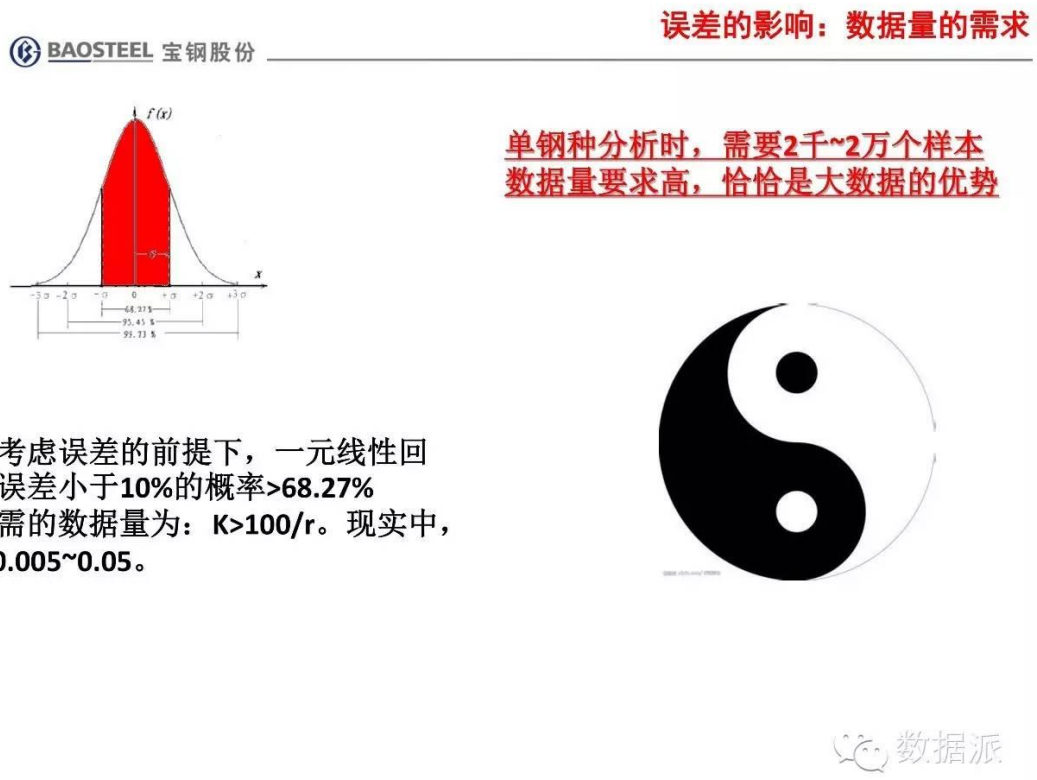
后来我想明白了：变量分布条件不变的情况下，不管模型对错，误差最小是最好的；但是当数据的分布特征一旦改变，误差就会立刻变大。所以，这种“误差最小”的模型不能用来预测大范围的结果，也不能用来控制和优化对象。



下面给一个形象的说法。本来 x 分布是在较小的范围内。当 x 有检测误差的时候，x 范围就变大了。于是，统计结果的斜率就变低了。这个偏差有多少呢？我大体算了算，大体有 50-70%。

请大家注意：在这个例子里，我是假设对象就是简单的一元线性模型。这么简单的东西，“误差最小”的观念都会失灵；那么更复杂的，比方说神经元，可能得到真实的东西吗？

这就意味着用 :误差最小化的一切优化方法可能都会存在跟真实性偏离的问题。但是，当我们需要用模型进行控制或设计时，我们需要真实性，而不是误差最小。这个结论其实很糟糕：意味着很多常见的办法都失效了。



另外一点，当检测误差比较大的时候，统计结果就比较难以稳定。我做了一个测算：误差小于 10%的概率大于 66%时，大概需要 2000 到 20000 个样本。这就为工业当中要想得到稍微稳定一点的结果你必须要大的数据量 ,小的是不行的。

练太极拳的，身体差

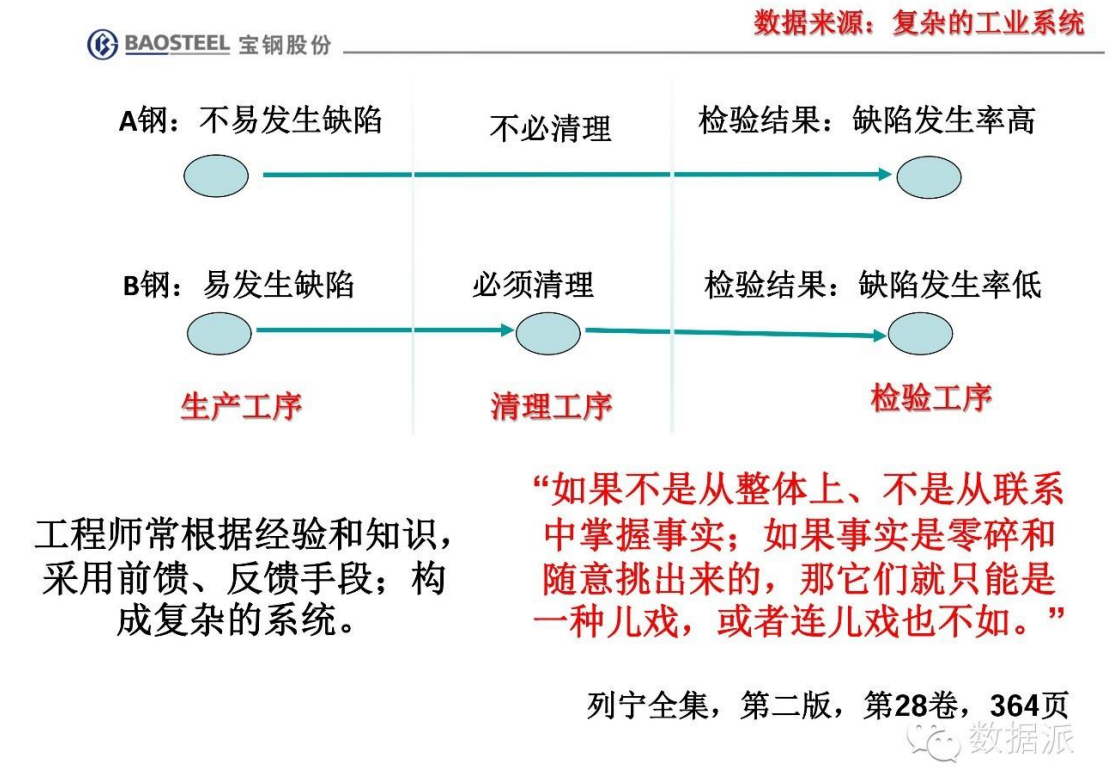
工业过程分析一定要重因果
否则很难达到可靠的要求

杀了公鸡，太阳照样升起

世界上有三种谎言:谎言、弥天大谎和统计学

工业过程还有一个特点，相关性和因果性常常是不一样的。

什么叫相关性呢？比如公鸡一叫太阳升起，这有相关性，但是它不是因果，因为你把公鸡杀了，太阳照样会升起。这就像我刚才说的：认真准备反而可能讲不好一样。类似的例子有很多，练太极拳的身体都往往不好，为什么呢？因为身体不好的人往往才去练太极拳。英国有位前首相说过：“世界上有三种谎言：谎言、弥天大谎和统计学。”如果统计方法用不好，你的结论可能是完全错误的。



这里举一个工业界的例子。比方说 A 钢种不太容易发生缺陷，所以不对它进行清理，直接检验。B 钢种容易发生缺陷，必须清理之后再进行检验。如果你统计分析时，把中间这个过程略掉，你会发现：A 钢种发现缺陷率高，B 钢种发生缺陷率低。这样，结论和实际正好是相反的。

这种现象不是偶然发生的，而是经常发生的。

工业系统是根据人们的认识设计的复杂的人造系统。工程师常常根据自己的经验和知识采用前馈、反馈的手段。特别地，如果已知某个变量（如钢种）对质量有重大影响的话，一定会设法把影响降低（如清理）。所以常常有前面这样的现象。所以几乎是必然发生。

现在经常有人说大数据，只要碎片化就可以了，但对工业大数据真的不一定合适。

列宁说过一句话：“如果不是从总体上，不是从联系中掌握事实，如果事实

是零碎的和随意挑出来的，那它只能是一种儿戏，甚至连儿戏也不如”。

我一个感受，做数据分析其实非常之难的，为什么难？你每天都跟各种各样的假象做斗争。你不知道谁是假象的话，你根本啥都没法办。我曾经跟我的一个徒弟说：做数据分析是异常驱动的。也就是说，如果数据展现的现象跟你想象的不一样，它里面就可能包含有用的东西。但是，这里有个前提：你要知道什么是意料之中。我常说：有意料之中才有意料之外。如果你对专业领域不熟悉，就没有“预料之中”，那“意料之外”也往往只是无知的表现。

所以，做数据分析的人必须了解工业实际。反之，如果不了解工业实际，发现一个问题，就要跟专家讨论半小时；再发现一个问题，继续去讨论半小时。问题是：别人没那么多时间来跟你啰嗦啊。

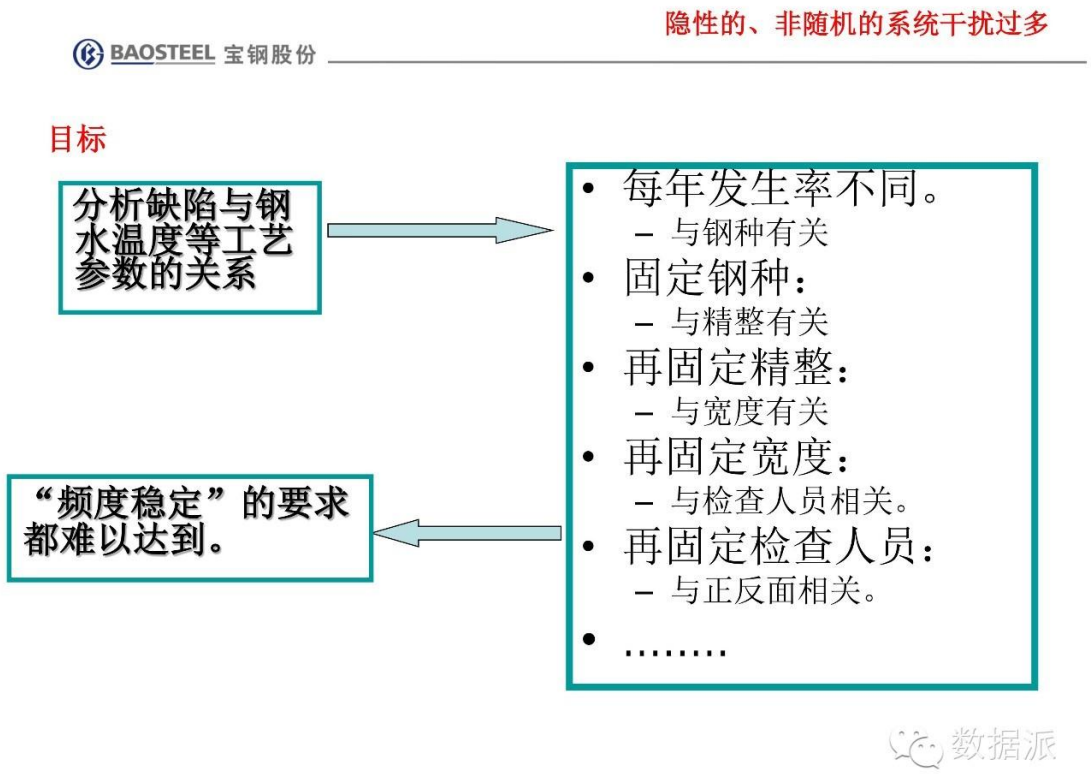
做分析麻烦之处，还在于很多“预料之外”是数据质量不好。

我曾经统计过两个钢种。我把预报误差特别大的拿来做分析。其中，从A钢

种抓取了 110 个特别大的；进一步的研究发现：有８５个数据本身含有某种数据严重异常，占预报失误的 80%。另一个钢种更高，占９２％。

这里又冒出来另外一个问题：很多人做出来的模型，正常情况下能预报，异常却预报不到了。这个事也很糟糕：人家希望你把异常的给抓出来，你只能预报正常的有什么用呢？这个原因，导致很多模型变得没用。

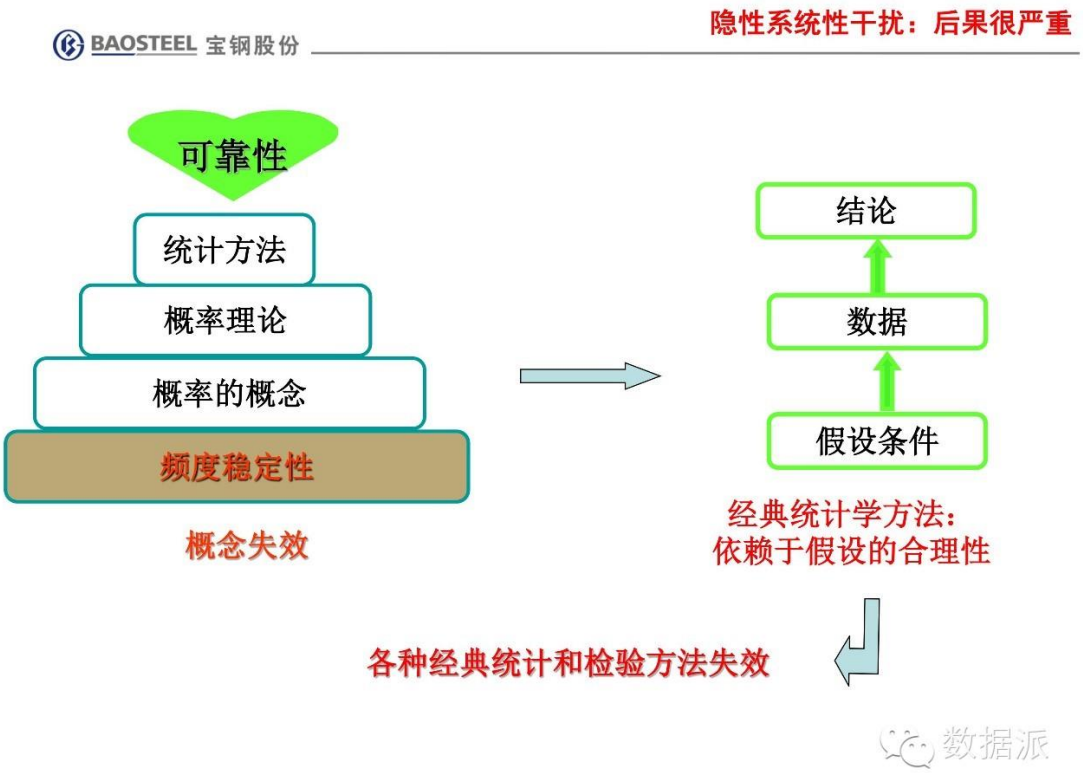
我们要记住：这往往是数据背后的原因，是很常见的。因为异常往往是有特殊的原因引起的，而如果你的系统中没有记录这个特殊的原因，那么你自然会出现这样的事情。



我再跟大家谈一个案例，谈谈工业数据分析的复杂性。

很早之前，我要研究钢坯的缺陷率 y 和生产温度 x 之间的关系。我拿几十万条数据分析，发现结果很不稳定。后来，人家告诉我：y 跟钢种有关，不同钢种的缺陷相差几十倍，建模时必须区分钢种。于是，我开始分钢种研究；然而，结

果还是不满意。后来我发现，y 与钢坯的正、反面相关，两面的缺陷率相差三四倍。那好，我再固定正反面继续分析；但遗憾的是：结果还是不稳定。这时，人家提醒我：你看看谁检验的，因为甲、乙、丙、丁四个班检验出来的缺陷相差三四倍……以此类推，不知要固定多少分组，直到每个分组里面几乎没有几个样本了。



这个案例高度我们 :工业系统中的系统性干扰非常多。如果没有意识到的话，怎么可能把它们两个变量之间的关系搞清楚呢？如果不排除系统性干扰，缺陷发生的频度就是不稳定的。所以，很早之前我就意识到：分析工业过程数据时，概率理论和统计方法不可滥用。

咱们再把问题稍微往远处扯一扯。实用的工业技术最重要的基础是什么？很多人经常忽视一个问题：可靠性。

神舟靠什么成功？

中华之星为何成流星

工业大生产相当复杂
从大数据淘金，仅靠相关分析可能是不够的

但可靠性实在是太重要了，给大家举正反两个例子。

大家知道，神舟飞船是一个很高级的技术。据说曾经这么一个故事：飞船安装过程中，一根头发掉到里面去了；然后他们决定停工，几十个人开了三天会，论证这根头发会导致什么样的后果。后来论证下来没事，才复工。反面的例子是中华之星：这是中国人自主研发的动车。就是因为测试时出了一点小问题，就被铁道部否决了。

上午跟莫老师交流的时候也谈到一件事 :我曾让一个非常优秀的同事开发一个程序。开发完成后 ,他拿过来问我是否可行。我说不行。他就问我有什么问题？我说： “我看不出有什么问题，但是我没法证明它是没有问题的” 。

在编写控制程序时 ,我往往要用 99%的精力去来想 1%的非正常状态如何处理；往往是一行功能性的程序，10 行防止错误的程序。说实话，如果程序有问题，出一次事故就吃不了兜着走了。所以工业界可靠是第一位的，你不产生效益

没关系，别把人家的设备搞坏掉了。

一个技术是否先进，最难做的往往也是可靠。我刚到宝钢时，有一位前辈问我：“小郭，那是学先进控制的，为什么不把先进控制技术用到宝钢呢？”我当时回答他：“条件不满足。”其实，一个企业之所以能用先进控制的前提它的设备先进、检测稳定，这时可能有好结果。如果设备、检测各个方面都是有问题的，那么用先进控制的结果可能会适得其反。所以能用先进控制技术是企业先进的一个结果、一个表现，不能为先进而先进。

我经常说，可靠性与价值往往是一个硬币的两个方面：可靠性要求高，它的价值才会高。比方说，我给你做一个预报，我告诉你按照我的预报做可以节省 100 万试验费。那么人家也可以告诉我：你预报错了，我亏 100 万试验费。所以，预报模型的价值和它可能产生的风险是同时存在的。

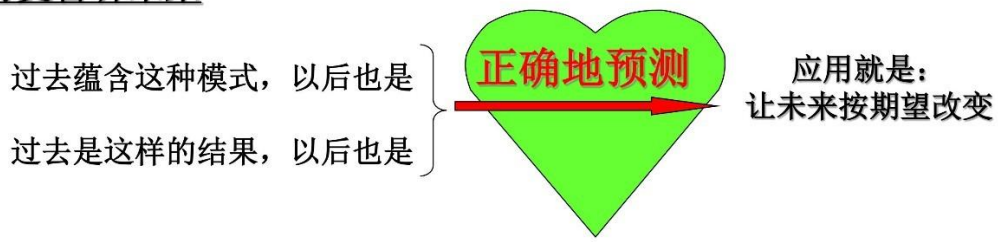
Ys=300+200*C+80*Mn+.....
Ys=230+800*C+40*Mn+.....

没有共识的预测

结果很可能是海市蜃楼

这就是我们常见的预报模型，每个人给的完全都不一样：你叫我去信谁？这样类似的模型出了上百年了，每个人都弄出一个来，但是同一个钢种不同的月份得出来也不一样。所以，它的问题是在于可靠性不够。我做出来的模型，有时候精度可能还不如简单的线性回归，但是它的可靠性提高了。

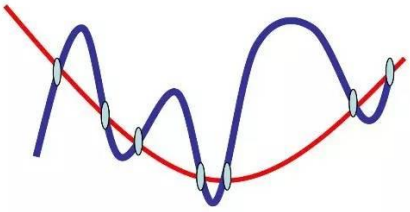
让历史告诉未来



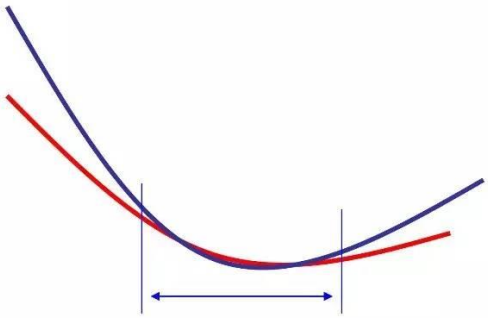
维克托·迈尔-舍恩伯格在接受记者采访时表示，大数据的核心是预测，其对人类行为以及社会问题的预测为人们津津乐道，而预测系统之所以能够成功，关键在于它们是建立在海量数据基础之上的。在不久的将来，现今许多单纯依靠人类判断力的领域都会被计算机系统所改变甚至取代，因为它为人类生活创造了前所未有的可量化的维度。大数据已经成为新发明和新服务的源泉，而更多的改变正蓄势待发。

特别是我们大家有一个观点，建模为什么重要？最重要它可以预测，特别是利用这个知识来改变世界，这是最重要的用途。好的模型不仅要预测未来，还要有外延性；不仅对建模数据管用，对新数据也要管用。这样的东西才有真正的价值。

数据不足或分布区域小时遇到的问题



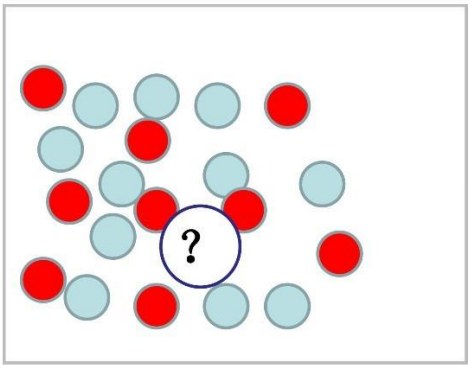
内部问题:过拟合

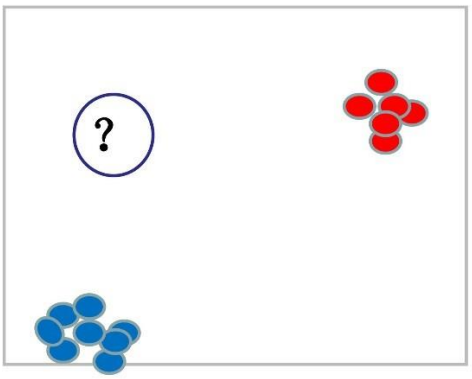


外部问题: 难外推

但是大家知道 ,过去我们往往强调精度 ,就会出现很多问题。一种是过拟合：老样本都预报正确，新的进来是不着调了。还有一种在里面是很对的，一到外面就发散了。这种事情很容易产生。

25年前的两个案例





当样本的数量少时
对已有样本的判断准确，绝不等于对新样本的判断正确

其实本人 25 年前做硕士的时候就遇到过这样的问题：你要多少精度我给你多少精度，但是我心里知道这个东西不靠谱，因为我不知道新的过来会怎么样。

有度难而无度易
《韩非子. 外储说左上》

知识往往需要超越已有认识，数据分析的结果才有使用价值。但人们对工业对象的认识却往往很深....

另外，工业数据分析困难还有一个原因。工业系统是人造系统，人家对这个对象研究得很透，像钢铁研究了几百年了。你说我发现的知识有用，你就得超越人家已有的知识。你告诉人家碳对强度有正作用，人家说这是废话，我知道 30 年了。你必须要比人家更高一层才能发挥你的作用，这也是难点所在。

工业数据分析：一般只能基于已有数据

数据分析常常不同于科学研究

往往只能利用现有数据，
而不是为证明结论进行试验。

另外，分析工业数据往往与实验设计不同：我做这个分析的时候，不知道能得到什么。很难给别人提要求，有什么数据用什么数据。

要求高

条件差

不满足要求
常常不了了之

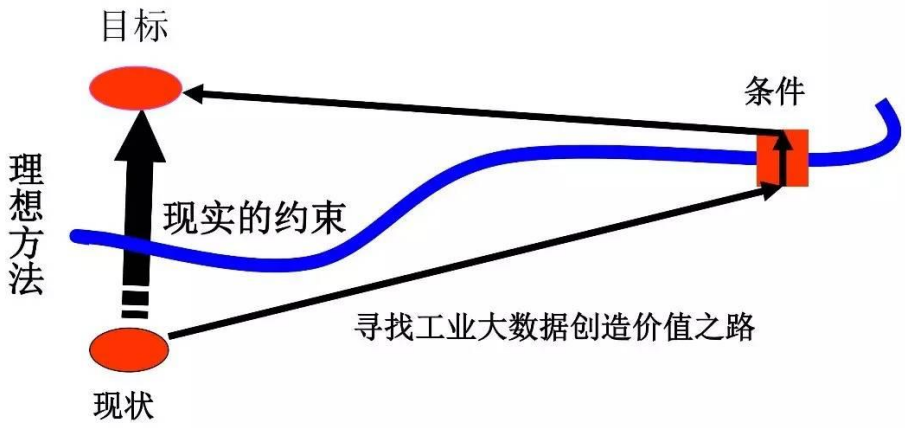
人们对工业数据分析的要求非常高，可靠性要求非常高，又要超越人，条件非常差，数据误差比较大，有时候分布也不合理，有的时候需要深入分析因果性.....许多工作最后似是而非，不了了之，就是这个原因。

以上是我给大家汇报的我对工业大数据特点的认识。

目录

第二点，我来谈谈工业大数据的方法。

方法：决定于目标、现状、条件与约束



- 提出宏大的目标、理想的思路都很容易，但达成目标很难。
- 理想方法无法达成理想目标，是因为遇到了某些隐性约束。
- 认清现实中的约束和条件。才能找到达到目标的可行方法

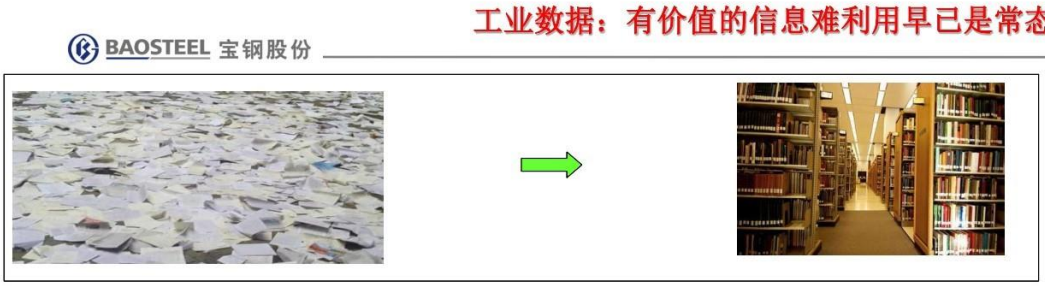
这是我经常跟大家谈的一张图 ,来描述技术创新的逻辑和思路。在这张图中，

蓝色的曲线代表一条河，右上角的方块代表桥。要求我们做得的是：找一条从现状到目标的最短道路（WAY）。

学究气太浓的人，常常执着于理论上的结论，如“两点之间直线最短”。的确，理论上的结论一定是正确的，但理论上的方法可行不可行则是另外一回事。正如图中所示：理论上的办法遭遇河流，无法过去。现实的办法，则是绕道远方的桥梁——这条路理论上不是最近的，确实现实可行的。

事实上，对于创新问题，理论上的方法一般是不可行的：创新是做别人没做成的事，理论上的办法常常最容易想到、能想到的人往往很多——如果现在这个技术还没做成，往往就说明理论上是走不通的。也就是说，在创新时，理论方向几乎必然遭遇困难。这个时候，必须借助特殊的条件，才能达到技术的目标。

只有现实中可行的方法，才是真正的好方法。所以，现实中，你的“水平”并不体现在对理论理解多少，而是对条件（桥梁）、约束（河流）、目标与现状的认识。



先说一下工业大数据的一些现状。

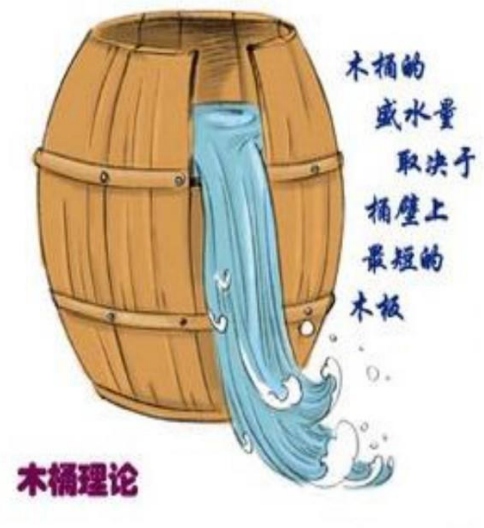
我经常听到一句话，说数据大得计算机存不下了。其实，多数情况不是计算机存不下了，而是少量的数据你都用好。比如每年我们有数以千计的质量异议，加在一起可能是数以万计。每件事背后都是一个案例、都有资金的损失。但是请问，计算机记得住吗？

当然，你可以做成文档记录下来。必要的时候可以去查。但是，你遇到问题的时候，计算机不会自动告诉你：过去发生过类似的事情，要当心。这样的能耐只有人才有。很多牛人之所以很牛，就是因为记住了这样的一些事。但这个人一退休，这个知识也就丧失掉了。

所以，对于这种碎片化的知识和相关数据，哪怕是一万条，计算机都不能很好地管理起来。另外，像前面说的这样，如何把几十万条数据中的规律挖掘出来并得到可靠的模型？这都是不容易的。

小米电视的段子

沙子再多 也无法经济地提炼宝藏



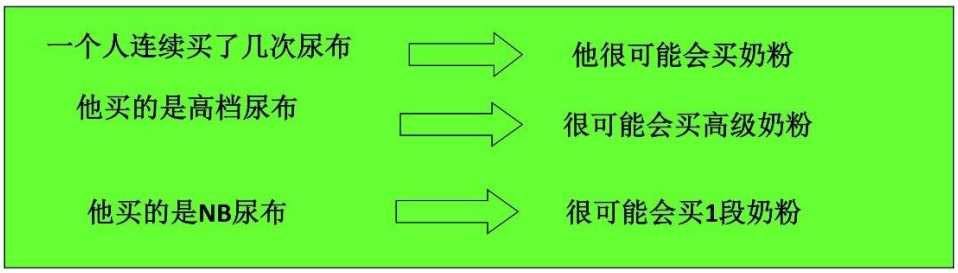
短板往往不在于计算机性能

我个人认为，要得到可靠的分析结果，缺少好的方法往往是真正的短板，计算机的性能往往并不是短板。

还有人认为数据多了就有用。不一定的。有个段子说，雷军做小米电视。他经常感到很困惑：小米电视的办法，跟小米手机一样，为什么就卖不出去呢？后来有人来给他点拨了一下 :请问买小米手机的人多少是家里有客厅的？这意味着数据再多，调查结果都可能是错的。这是短板是方法问题，数据多是不解决问题的。

另外，大家说大数据是沙里淘金。但如果随便给你拿一袋沙子，里面有万亿分之一的金子，你能淘得出来吗？如果一定要做，你淘金花的钱比这个金子要贵得多了，经济上是没有价值的。

相关性的价值：对商业和工业的价值不一样



相关就够了：
萝卜青菜各有所爱

ATM取款记录是秘密

这里我特别提醒一下，商务大数据和个人大数据是不一样的。与个人相关的大数据，相关性是很重要的概念，比如：一个人买了几次尿布，你可能说这个家伙老买尿布，是不是家里生孩子了？他可能买奶粉。而且看到买什么牌子尿布，大概知道他家的经济情况，可以推荐什么档次的奶粉；看他买多大的尿布，大概知道该给他推荐几段的奶粉比较合适。这就是相关性的价值。

但是工业上就不一样了：你买了我一吨的 Q235，我知道你干什么？我啥也不知道。所以，工业跟个人是不太一样的。人和人之间虽然有差别，但跟企业与企业之间的差别相比，还要小得多。

把荷花和仙人掌放在一个盆里；
浇水多也不好，浇水少也不好。

把与个人相关的商务和企业内部的
工业大数据相提并论
这样做也不对，那样做也不对。

所以我一个感觉，不能把工业大数据和商务、跟个人相关大数据混在一起。混在一起，强调相关性也不好，不强调也不好；强调因果性也好，不强调也不好。概念混了，就像把荷花和仙人掌养在一个盆子里面，浇水多也不好，浇水少也不好。

不是路到了尽头，而是到了该转弯的时候。或许我们要更换一种思维。比如，我们工业大数据要求的是什么？我们要求数据的完整性、真实性，这个东西是很重要的。

预则立，不预则废

问题表现在分析阶段，根子却在数据的收集与组织。

存储数据时的真实性，
组织数据时的结构化，
分析数据时的预处理。

首先是为人的分析创造条件
然后才肯能有自动化的分析

胜兵先胜而后求战
败兵先战而后求胜

孔子说：欲则立，不欲则废。孙武子说：胜兵先胜而后求战，败兵先战而后求胜。工业大数据也是这样。如果不是在开始的时候就把数据很好地组织起来，到了后面再努力也没用了。要想着把大数据用好，在收集和组织数据的时候就该想到它的目的。

IBM认为：真实性（Veracity）是当前企业亟需考虑的维度，将促使他们利用数据融合和先进的数学方法进一步提升数据的质量，从而创造更高价值。

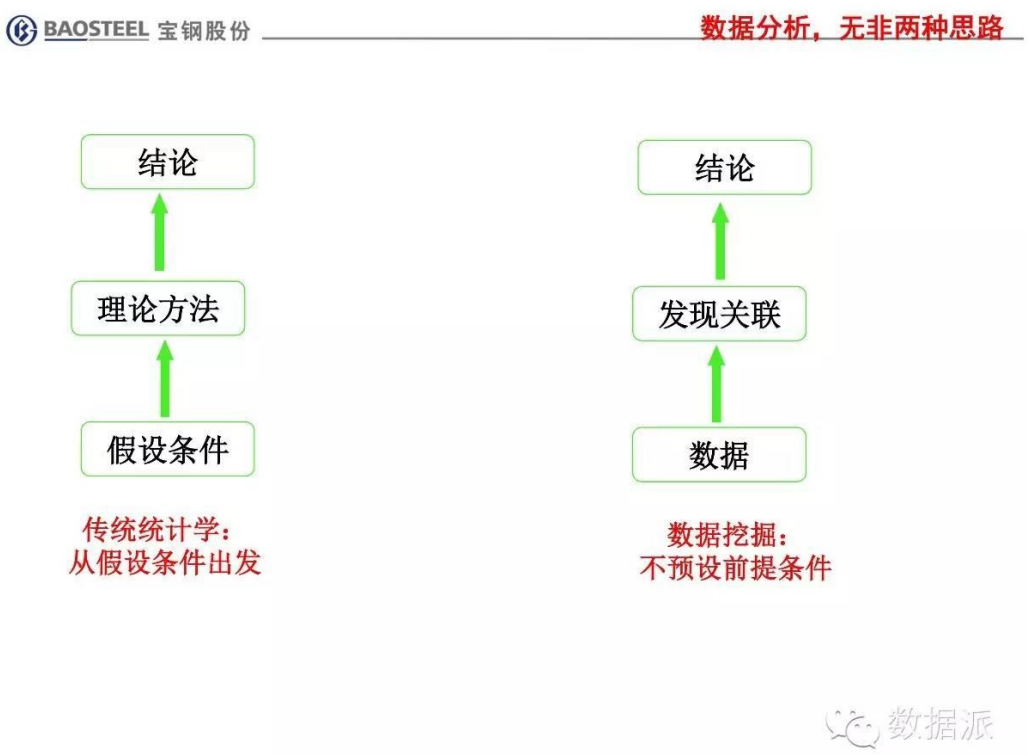
前辈王洪水先生认为：

真实性首先是完整性，数据之间的联系要尽可能完整地记录下来。

本人的一管之见：

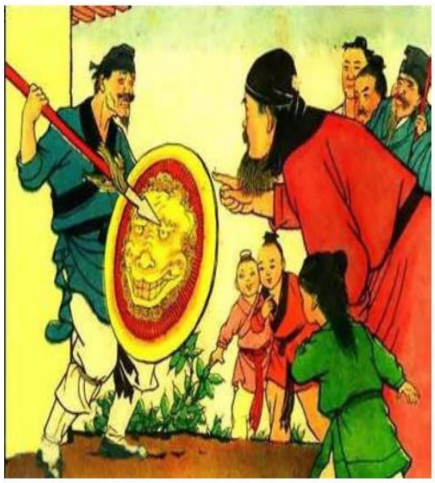
知道数据是怎么来的，有时候是分析问题的关键所在。

IBM 认为，应该把“4V 理论”改了一下。把其中一个 V 改成真实性，他认为真实性是当前企业急需考虑的维度，并且将促使他们利用数据融合和先进的数学方法进一步提升数据的质量，从而创造更高的价值。这段话读起来有点别扭，但说的理儿还是对的。就是数据的融合，特别是数据和人脑当中知识的融合，这是一个大有可为的一个地方。宝钢有个我非常尊重的前辈，叫王洪水先生，他说：“真实性首先是数据的完整性，数据之间的联系要尽可能地完成地记录下来。”本人在做数据分析的时候也有点想法：我不仅要知道数据是什么，更要知道数据是怎么来的。比方说，不仅要知道哪个字段是屈服强度，还要知道它是怎么取的，是横向取样还是纵向取样，是冷态取样和热态取样等等，不同情况下得到数据，虽然都叫做屈服强度，但内涵是不一样的。所以数据完整性不仅仅包含过程本身、对象本身，还要包含数据怎么来的。这样你在用的时候，才能识别一些假象，避免给误导。我们搞数据分析的整天就是跟假象做斗争。



可靠性如何获得？

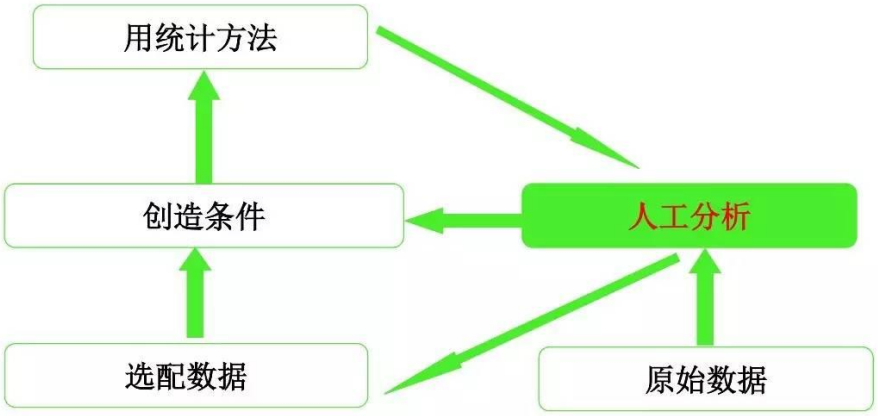
南开大学有位老先生，有这个一个观点，蛮有意思的。大体意思是，数据分析无非是两种办法：传统统计方法是先给出假设，结论的正确性决定于假设是否合理；现代数据分析方法是根据数据表现的结果直接给出结论，可靠性难说。但现实中我们发现：这两个方法都不好用。对于统计方法，我给不出合理的架势，而现代方法的可靠度又不够。



如果应用统计方法的条件是可以确认的，则统计结果就是可靠的。

应用统计学方法的条件，一般是不能确认的。

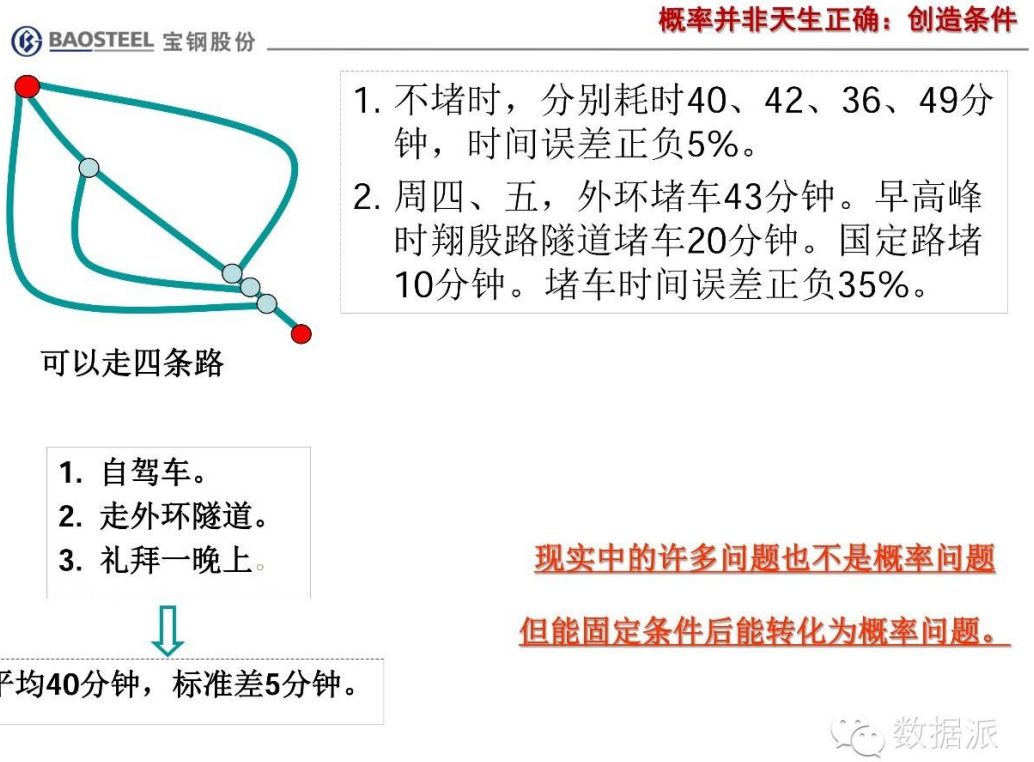
老先生给我的启示是：如果应用统计方法的前提条件是可以确认的，统计方法一定可以得到可靠的结果。我们的问题是：条件一般是不能确认的。那么，能否将注意力放在创在条件上呢？



统计和机理的结合

对条件的判断常常来自于数据之外的认知

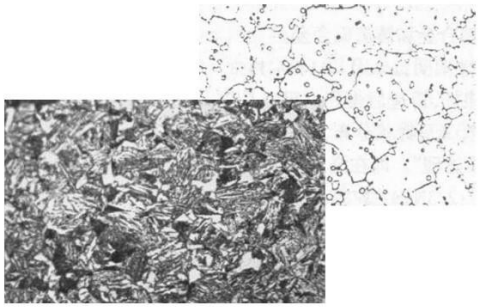
要把分析的重点转到幕后去，也就是说利用原始数据、通过人工分析给它选配数据，来创造统计上可行的条件，得到可靠的统计结果。注意：在这个人工分析过程中，很多知识来源于被分析数据之外的认知。



如果有人要问 :从宝山到浦东机场花多长时间？我认为这不是一个概率问题。如果变成一个概率问题的时候，你必须要说我走哪条路，从哪个地方走，什么时间段走。这些系统性干扰排出了，才是一个概率问题。

钢铁材料的问题：原理是一样的

C含量对强度的贡献多大？
平均：2-40Mpa/100ppm



1. 什么组织？
2. 什么强化机制？
3. 什么产品规格？
4. 什么工艺条件？
5. 取样方向？
6. ……

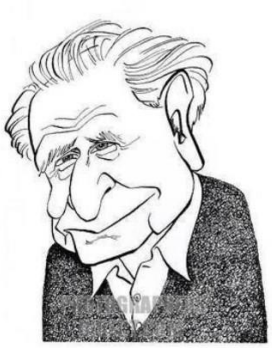
案例：620度卷取3mm的中低碳CMn钢为8.35

1. 数值固定、精确。
2. 影响因素可知。
3. 适用范围可知。
4. 其他干扰基本可忽略

也就是说：用统计办法的话，首先要把被后的系统干扰排除。

现实的可靠性：要放弃对绝对真理的奢求

数学规律：已知的正确，未知的也正确。
物理规律：已知的都正确，且竞争成功。
生物规律：80%的正确.
经济规律：50%的正确
社会规律:.....



证明数学定理，一个证明就够了，
证明历史事件，至少要5个证据

可能会有人抬杠：你的做法真正可靠吗？其实，现实中的可靠都是相对的。

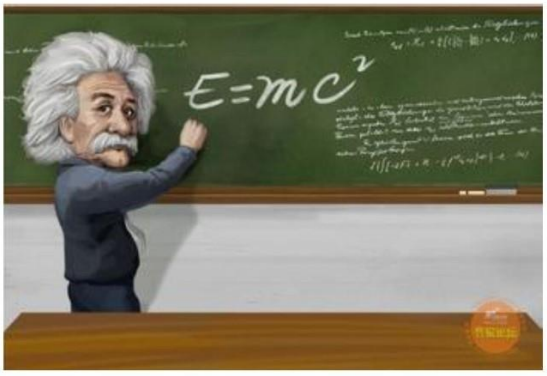
判断过程包含人为的因素，不是严格的数学证明。数据分析更类似一个发现科学规律的过程。学过科学哲学的人都知道：科学理论其实没法证明，只能证伪。

一个理论是否正确，不同的学科有不同的标准。有这么一个段子：“数学界的标准是：已知的要正确，未知的也要正确；物理学的标准是对已知的现象都能正确解释；生物学的标准是正确解释 80%的现象；经济学只要有 50%就可以了……”

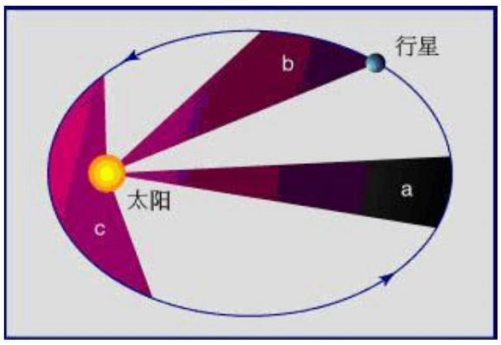
我想，为了得到更加可靠的结果，论证过程中就需要更多的数据，更多独立性的证据、更可靠的证据、更严密的论证链、有科学原理的解释，且没有明显的反例，这个时候我只好认定它就是比较好的了。

我们在认证一个结论时，尽量从多个维度验证；如果没有明显的例外，就认为它是可靠的了；在没有新的证据之前，找不到比这个更好的理论，就可以暂时采纳它。

可靠：验证精度及外延验证



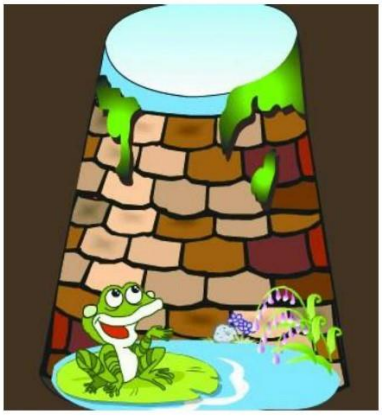
日心说的胜利：精度为王。



相对论的确立：预言成功

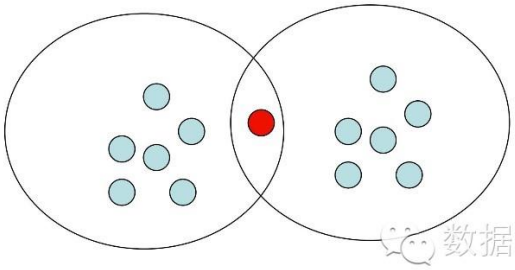
牛顿《自然哲学的数学原理》中就曾经提到：为什么做这几个假设？第一，这些假设足够简单且能解释问题，第二，现在没有发现跟它相违背的事实。

如何看待领域知识



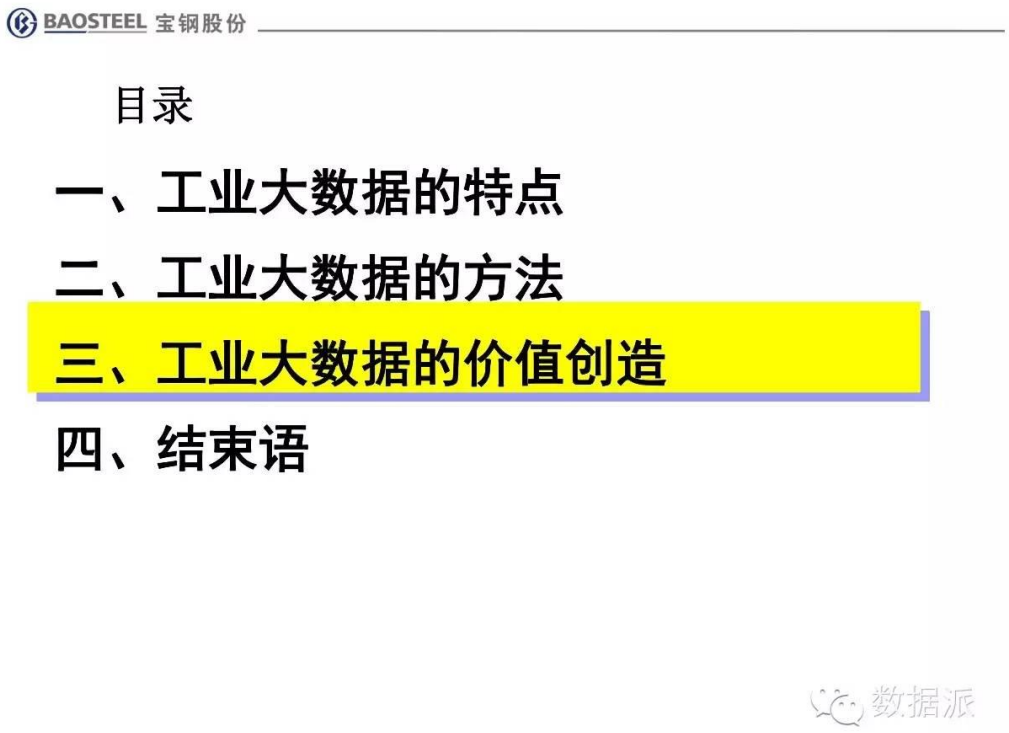
- 生产数据不包含设计依据。单纯从数据中分析问题，如坐井观天。
- 要完全用数据说话，数据需求必遭遇组合爆炸，数据永远不理想。
- 数据不理想时，会存在难以证实和选择的不同假说。机理用于选择更合适假说、客观上起到减少数据需求的作用。

我们认为“可靠”的结论，是理论和数据的共识



这里特别要说明一下：仅从数据本身就得到可靠的结论，往往是一种奢望。在做分析的时候一定要把人的知识和科学机理融合进去。一个可靠结论，既要能描述数据的实际特征，又要符合冶金机理。

从数据到数据的分析方法为什么会有问题呢？我的感觉是 :如果纯粹从数据上加以证明的话，一定会遭遇组合爆炸问题；要得到全面可靠的验证，数据永远是不够的。



现在，我跟大家谈谈价值创造。

其实关于技术创新，我在宝钢做了 20 年，经常感到很痛苦，为什么很痛苦呢？因为我们作为一个博士，很想做有技术先进性的东西。但现实当中，我们发现先进的东西往往不实用，实用的东西往往不先进。我们一直在很薄的夹缝中生存。尽管如此，我们不能放弃的底线是创造价值，因为我毕竟是企业的人。

只有创造价值

工业大数据才有生命力

技术才能真正在企业落地

——熊彼特（1883-1950）

只有将新技术用于经济活动并取得经济成功才算创新。

熊彼特说，只有将新技术运用于经济活动并且取得成功才创新。同样，只有创造价值，工业大数据才有生命力，才能真正在企业里面落地。这是必须坚持的一条原则。

寻找大数据落地的蓝海：场景很重要

一杯水

放在餐桌上是垃圾

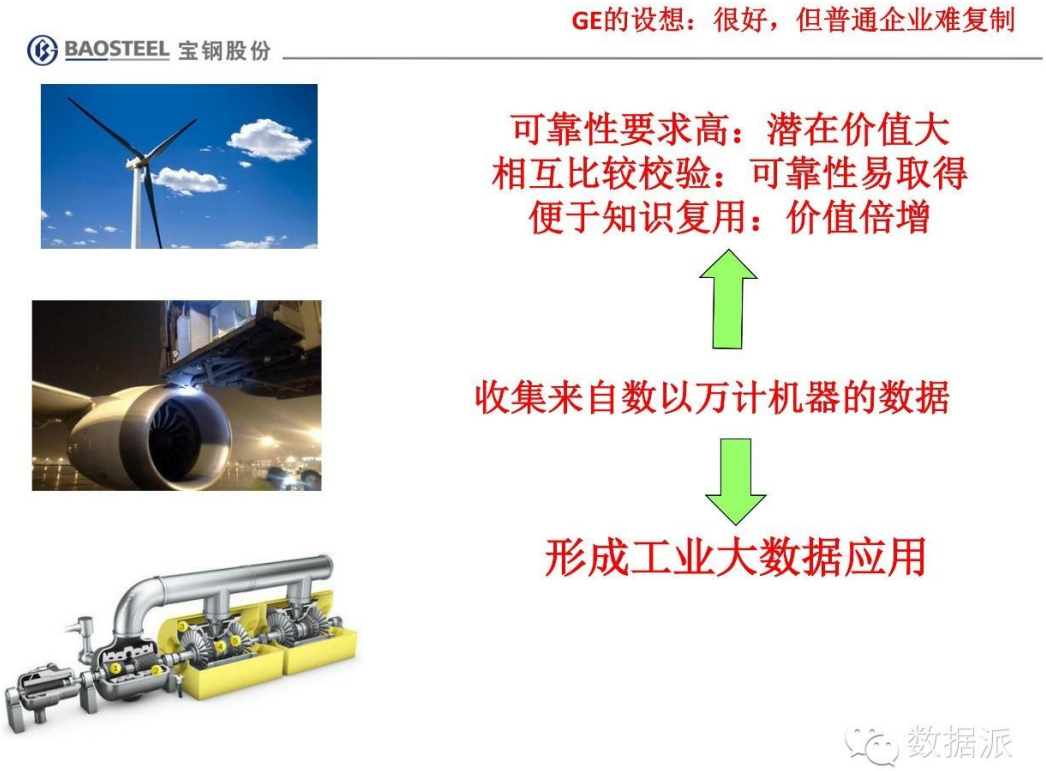
放在沙漠里则可以救命

新技术，要雪中送炭，不要锦上添花

价值决定于用户

要求和难度决定于场景

宝钢的老领导何麟生先生，今年快 90 岁了。我去探望他的时候，他跟我说：半杯水，剩在餐桌上是垃圾，放在沙漠中可以救人一命。换句话说：技术的价值它决定于用户，用户是怎么看待它的。我们做新技术要想创造价值，要做到雪中送炭，而不要锦上添花。所以，大数据能不能落地，关键要找到合适的场景，而不是技术本身是怎样的。



谈到工业大数据，很多人知道 GE 的设想。也就是说通过飞机发动机的大数据减少维修成本，来提高安全可靠性。这个例子很好，但要跟大家强调一下它的场景。

第一，航空发动机的成本很高，可靠性要求也很高，所以对它的相关工作能产生很大的价值。第二，从一台发动机的数据中发现的知识，可以用其他发动机来验证，提高可靠性；可以复制到成千上万台发动机上，发挥更大的价值。

但是，如果这个思路针对的是自家的一台重要机器，情况就完全不一样了。分析结果的可靠性、价值创造都不一样。

- 钢板要不要拆除？
  - 对过程数据的完整记录很重要
- 工艺工程师：每一个质量异议都分析。
  - 判别责任
- 河南近乎无人化的小企业
  - 数当然是必须的。
- 6sigma理论产生的背景：高质量要求
  - 用数据和事实说话，才能将次品发生率降低到百万分之3.4。

哪些场景会适应普通企业呢？我给大家举几个例子。

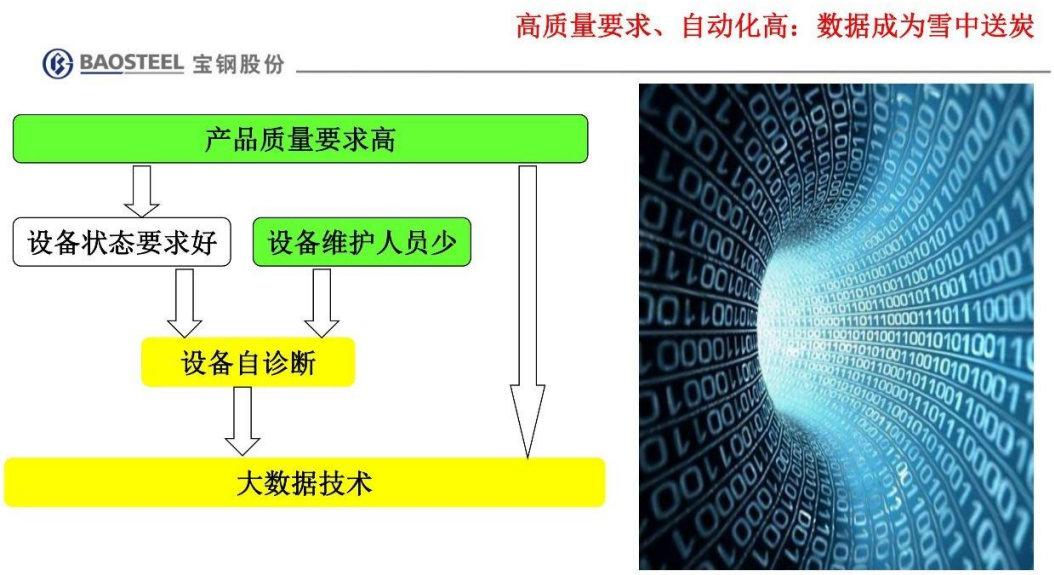
有人买了我们的钢，说我们的钢有问题，要我们赔 100 万。宝钢就说了，这不是我的问题。对方就说：可以把这块板子拆下来检验；但如果拆下来发现是你们的问题，就要赔 1000 万。后来，宝钢回家看相关的数据后，自信地说：你拆吧，肯定不是我们的问题。后来也证明了我们的判断。这就是数据的价值。没有数据，你怎么敢下这个结论？

河南有一个小厂，农民企业家开的。别看它是小厂，却几乎是一个无人工厂。为了保证质量，每一个环节的数据都记录下来，放到数据中心上。否则，没有人在那儿看着，产品出了问题怎么分析呢？

大家知道有一个理论叫 6sigma 理论，这个理论要求将次品率降低到百万分之三、四以下。这个理论有一个重要的观念，就是用数据和事实说话：降低到这么小的次品率，单靠设计是不行的，必须能够在生产中不断地优化。而改进的依

据是什么？必须是数据，

换句话说，在质量要求高，无人化的场景底下，数据变得非常重要。这时的数据就是雪中送炭。



反之，有些情况可能就变得不重要了。比如说为了降低成本，有时明知设备有问题都要带病工作。这时，开发依靠数据的智能诊断技术，价值就小了。

数据到底有没有用处，关键是用户对质量有没有高的追求。有高的追求的话，数据的价值自然会被带上去；反之，企业对质量不关注时，再有好的数据没用。

我想起一个更极端的例子，大家知道三鹿奶粉。厂里明明知道里面有三聚氰氨，还是要卖出去，更可气的是：石家庄政府甚至还包庇它！所以，从大局上看，政府要改革、要重视质量监管，数据才会重要。

某一个豆腐厂，工人操作常常不规范；质量卫生难保证。

老板安装了几个摄像头，引到监控室。虽然他也不怎么去看，但问题却解决了。

宝山有家豆腐厂，有 1000 来号工人。过去，有工人经常偷懒。于是，老板搞了一个摄像头，引到他的办公室里，产品质量和管理水平马上变好了。偷懒的人想到：万一被老板发现怎么办？这就是监控的作用。

咱们中国是刚刚起步于一个农业社会，人的纪律观念差、缺乏工匠精神。怎么才能应对新工业革命的挑战呢？我想，用大数据提高管理能力或许是个好的切入点。管理其实很重要：管理能力差导致质量差，质量差导又会成为技术创新的阻力。

王洪水先生说：

利用数据，把产生过程像录像一样记录下来。

1. 对抽象的研发、设计、采购、销售、制造、设备维护等诸多环节，大数据记录可以看做一种抽象的“录像”。对提高管理水平的价值是很大的。
2. 在某些传统企业，管理漏洞的浪费可能大于总利润。
3. 用大数据提高管理水平，或许适合很多中国企业。

工业企业中有很多工作流程。包括生产、采购、销售、服务、研发、设备维护等等。我想：利用数字化的办法，把这些流程的痕迹记录下来，再加上一些职能性的算法，评价这些正在进行的工作。就像录像一样，把工作的状况显性化，管理水平可能就会上去。

- 大数据是对（生产、研发、服务....）过程痕迹的数字化记录，以建立“用数据说话”的基础；常常是对数据资源的二次利用，主要通过间接途径创效。
- 将大数据与恰当的业务流程绑定，才能持续地创造价值。

- 工业大数据主要价值，或许不是发现规律性知识，而是用来提炼有用的信息，用于启动软件化、模型化的知识，推动智能化，并解决人的关注力瓶颈问题。
- 在上述逻辑中，规律性知识、流程元知识主要来自人脑；工业大数据主要起到验证、纠正和精确化知识的作用。
- 数据、分析方法与领域知识的深度融合是关键。

我有几个粗浅的想法，请大家批评指正。

如何看待工业大数据。工业大数据是对过程，生产、研发、服务过程的数字化记录，它的目的是建立“用数据说话”的基础。它常常是对数据资源的二次利用，不是为了大数据而大数据的，主要是通过间接的效益来创造价值。大数据要持续创造价值，最好能与日常业务流程绑定，才能持续地创造价值。

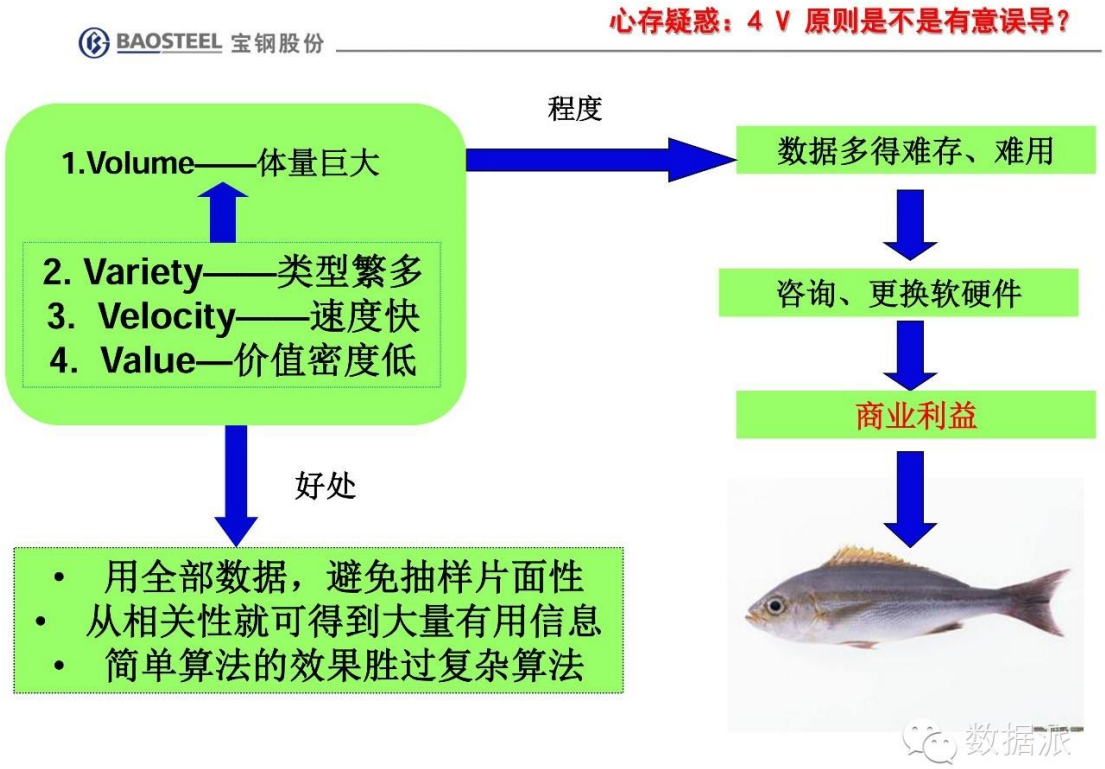
大数据与知识发现。我觉得工业大数据的主要价值或许不是发现规律性的知识。这种事情难度实在是太大了，我自己做了 10 年。用大数据提炼信息（如模式识别)或许是个更好的方向。如果能将分析结果与自动化系统、智能系统对接，就能持续创造价值。

一个企业什么资源是最缺乏的？是领导的关注力往往是最稀缺的资源：领导职位越高，他就越忙；忙会导致很多错误的决策、机会的遗失。如果能用大数据，把必要的信息提炼出来，让他在最紧急的时刻能够看到应该看的东西，就等

于扩大了领导的能力。

另外谈一个观点：规律性的知识是需要的，但是它主要是来自于人脑的。工业大数据的作用往往起到验证、纠正和精确化的作用。数据分析与领域知识能否深度融合，往往是用数据创造价值的关键。

特别提醒一下：我们一定要了解国情。我们的国情就是我们刚刚起步于一个农业社会。经常有人问：高大上的技术为什么不能落地？我的答案是：如果一个人营养不足是因为没钱买吃的，就不要给他推荐青岛大虾。其实，他也知道青岛大虾有营养，但他买不起。一定要知道这个逻辑，我们的技术才能落地。



如前所述，在工业界。我们遇到的问题是明知有知识挖掘不出来，瓶颈不是计算机本身。

所以，我对这个4V原则有点怀疑。在我看来，4V理论中，体量巨大（Volume）是核心：因为其他三个‘V’都是来支撑它的。如果认为体量巨大是关键问题，你就要付钱给软件公司，请他做咨询、要更换你的软件和硬件了。于是这些公司就开始赚钱了。

我觉得用好工业大数据请先把“大”字忘掉。我们首先应该关心价值创造的逻辑。把这个逻辑理顺了，就容易落地了。谢谢大家！

下面是问答环节

提问 A :您刚才讲的一个画面我很有感触 ,就是仙人掌和荷花 ,看似不能融，但是我觉得看似不能融的植物完全可以种在一个盆里 ,只是它里面的秘密我们没有发掘。如果荷花需要水大，仙人掌需要水少，他们两者之间隔了一层膜，这个膜不是塑料膜，而是生物膜，它可以过滤。当仙人掌需要水的时候，这种膜可以自动地释放水，当它不需要水的时候它可以屏蔽，我想我们人体当中肯定有这样的膜。我们有没有可能开发一种。再一个，我对人体的结构非常感兴趣，就是我们在探寻癌症的时候，有很多里边的信息全是假的，那我们怎么发掘？当你往里头越深的时候，你会觉得自然给我们的教训是非常无穷的。在这里边，我想问您的，最关键的发现最本质的东西，就像您在工业大数据当中，最本质的东西是什

么？这是我的问题。

郭朝晖：我先回答第一个问题，我觉得你说的其实是非常有道理的。有一个创造理论叫做 TRIZ。其中有一个重要的原理，叫做分离原理。我们经常会遇到各种矛盾，这样也不对，那样也不对，它有什么办法？它有分离，时间分离、空间分离、局部和整理分离，等等，您的这个思想其实就是分离原理。

我这里强调的是：不要把商务大数据的想法生搬硬套到工业大数据当中，它的里面有相当多不同的地方。我这次讲的工业大数据，主要是针对制造过程的。按照现在一般的提法，工业大数据并不仅限于制造过程，还包括销售采购、研发服务等过程。

提问 A：关键的想法和解决问题的方案，这里面关键的因素是什么？

郭朝晖：想法很重要。但能否解决问题，关键是要看能否有合适的条件。如果条件不具备,再好的想法都没有办法解决问题。

提问 A：还是不对，一个简单的例子，您知道爱因斯坦的相对论是怎么发明出来的吗？灵光乍现。其实简单的数据里边最值钱的，就像您这种顶级的人才，包括清华，直觉非常重要。美国一个未来学家说现在大数据的数据太多了，在这个当中发现最有价值的数据因素就是直觉，在某种程度上，你要先非常自信你自己的直觉。这种直觉在某种程度上会非常大程度节省你的人力、物力，还有前期的研发，挖掘这种直觉，在大学，还有在其他的研发单位非常地罕见。我接触过很多的人，包括诺贝尔奖获得者，问他也是哑口无言，但是我说完以后没有人敢反驳。

郭朝晖：这个问题我可以试着你回答您一下。咱们往往把科学和技术放在一起，其实科学和技术是完全两种不同的事情，科学是发现知识的，技术是解决问

题的。两者不同，如果把它们混淆在一起，就好像一个人既要长跑又要短跑，长跑和短跑是不是可以兼容？或者是不是都可以取得很高的成绩？有可能的，比如在我们班里，这里所谓长跑可能也是短跑冠军；但是放到奥运会上，你是一个长跑冠军，就一定不是短跑冠军。同样，技术创新对企业来说是一种投资行为，我们依赖的是可以有序地推进的东西，灵光的乍现是不可缺少，但是不是我们可依赖的事情，因为那是科学家做的事情。

提问 A：但是没有灵光乍现，你研发的时间会越来越长，浪费你的金钱。

提问 B：您说大数据怎么来的比数据本身还重要，还有就是饥荒和虾的问题，好的不一定是最适合的，包括我们自己要做一些选投入产出比最高的事情来做，包括企业和个人，甚至于家庭都是这样。我的问题就是，如果能有一个很小的案例的话，说一下大数据可以解决的问题的话，因为我是一个外行的人，会更好地理解大数据的作用。第二个就是做模型的人通常来说知不知道自己的漏洞在哪儿？

郭朝晖：我在做模型的时候，其实是一个学习的过程，你在永远探知你所不知道的东西。我刚才说的疑问的驱动，当你出现问题的时候，你才发现不断的你的知识是不够的，不断地补充进来，这是肯定的。

提问 B：有没有可能，假如我在做一个成果，实际上我交出去的时候，我是知道在哪个地方会存在问题。

郭朝晖：我做了一个模型，是预报性能的。但事实上，除了性能我还给出其他两个指标：预报误差的方差是多少，也就是说它的波动范围多大；可预见性如何，也就是说是不是吃得准。这样一来，预报结果其实就变成了预报均值。

任何模型肯定是有缺点的，世界上永远没有到处都适合的模型。给用户提供可靠性，除了尽量把模型做好之外，要告诉人家模型可能存在的问题、用在什么

地方合适。

提问 C：您好郭老师，今天听您的演讲感觉很有收获，您这个大数据是很接地气的。其实您刚才有一句话就是可靠性的问题，我觉得搞生产技术的最重要的就是必须能够运行。现在我做的是跟互联网搜索有关，主要也是跟分词有关，当然跟您这个不太对接。我也看了您的资料，在工业方面的大数据，包括我看过一个案例，广东有一家生产数据机床的，他利用大数据，在机床上进行了传感器的安装，每台机床销售到国内，甚至于国外，它通过互联网技术，直接把它的整个运行状态随时采集到它的数据中心，来观察它平时的运行状态，这算是工业上的一种应用。

郭朝晖：对，跟 GE 其实是同一种模式。

提问 D：你好，我是清华的研究生，刚才听您说工业大数据，我想有两个方向，我感觉您把它放到一起，但是从我的认为，我认为是不同的事情。第一件，您之前说数据用到工业生产上，我感觉首先做数据的事绝对不是站在很高的层面上考虑问题的。因为我记得有一句话叫闻名不如见面，但是我想后面加一句“见面不如实践”。再牛的一个学生天天去到工厂一线，你去亲自动手摸一下、感受一下，看看具体的问题在哪儿，因为你去分析这个数据也是没有意义的，你不知道它究竟是怎么样存在，你的数据采集和整个数据怎么来的，其实是每一个，你可能坐的是一个很高的职位，但是你必须得要进到一个很低的位置上去感受，没有这个感受，做工业数据分析是没有意义的。这是我说的第一个层面。第二个层面，工业大数据中的大数据讲的是什么？共享和融合，我想讲到大数据是不是应该多讲讲如何从宏观层面上，各个企业之间如何共享技术，人才的共享，还有资源的共享，比如说宝钢可能哪些擅长，鞍钢哪些擅长，怎么样互补优势，让短处

更加地减少，这可能是大数据，大家共享数据之后，分析每家的优势是什么。大数据可能要在这个层面上凸显。您讲的是从产业，然后到大数据，我感觉是不是把它们给说小了？

郭朝晖：关于大数据，王老师是有一个很地道的说法，而且引经论点，这个事情我想王老师，您能回答一下吗？王老师因为作过一个报告，我觉得说得非常好。

主持人：我觉得你是更关注在应用方面大数据的核心的科学问题上，怎么样和应用来结合？如果从刚才郭总说这件事情，大数据的定义上来讲，如果从计算机的角度来讲，我们是给了一个很窄的定义，现有的技术不能很好解决的数据才叫大数据。其实今天郭总讲的并不完全是一个，用一个补集来介绍，其实它在介绍一个全集，我不知道这个能不能回答你的问题。

提问 E：我今天听郭总讲话受益匪浅，第一，他把大数据分成两大类，一个是商业，叫消费类的，一类是工业类的。消费类的，他认为这里边的知识都比较常识性的，所以他知识的那个关联度比较强，而且比较直观，尿布和牛奶什么的，但有的人说那是瞎编出来的故事。工业大数据刚才实际上这位同学也是在分，工业大数据也分，有产品质量方面的，今天郭总就讲钢的质量怎么从大数据的分析当中寻找到模型，来预测新的钢种的生产质量、概率，工业大数据除了质量以外，还有别的，就像你说的，企业之间的合作，还有供应链的问题等等，我这个理解不知道对不对，是不是工业大数据还得细分？

郭朝晖：工业大数据原则来讲是包括四个部分，这不是我说的，是我工信部的人这么说的。这四个部分我把它简称为“3 1”，所谓“3”，对应工业 4.0 的三个维度的集成，即横向集成、纵向集成、端到端的集成。你说的那个维度更多

的是在横向集成，这是它其中一部分，这就是“3”。另外“1”是加到一个企业外部的数据，这是工业大数据的一个全集，是这个样子的。端到端主要是跟客户相关的那套东西，是商务的味道浓一些。我对制造过程的关心更加多一些。

提问 F：首先非常感谢郭老师，我们跟宝钢也有一些合作，我想问一个具体一点的问题，在刚才郭老师讲的大数据，其实我觉得把“大”字去到非常好。其实我们很多年前就搞数据推动，刚才郭老师也讲了，工业当中好多场景没有用到大数据，或者不需要用到大数据。我们跟宝钢当中有这么一个问题，宝钢在炼钢过程当中，从铁水到钢水，铁水烧出来之后，一罐运到钢厂，到钢炉之间，郭总应该也很清楚，刚开始提的要求，这边一罐铁水用这边一罐钢水，刚开始提的概念做了好几年没做出来，后来又改名了，叫做铁水和钢水对应，为什么做不到一一对应呢？就是由于这边烧出来一罐车的钢水，首先重量不同，这边 300 吨，这边也是 300，这边可能一下子变成 200 吨了，中间一处理又损失一部分，变成 200 半吨了，再去回转炉做的时候就不够了，另外成分也不好。这边需要这种成分的钢，你做出来的又不是这种成分的。我就在想，中间过程中可能温度又降了，又不能进了。这个过程当中，宝钢是信息化在全国做得最好的，有公司也给他们做了软件，这套软件也让工人去用，但是工人用了一段时间之后就不用了，别的环节都用得很好，但是铁钢一一对应的环节做了系统也不用，用了之后觉得也没用。咱们站在学术的角度来看这个问题，咱们看待这个问题很简单，用传统的方式就可以了，但是就是由于这些数据和你想的不一样，采回来的不是你要的，你要的也采不上来。称重也称不准，温度也测不出来或测不准。郭老师，如果说这个问题让您，您做十几年，可能做钢检测这一块，如果您给一个方案的话，您觉得怎么用大数据解决具体的案例？或者说您又能提出一套方法，给我们一个展

示，作为一个案例，刚才那位老师也说了，能不能把这个问题，您如果做大数据的话，您觉得怎么做可能会有效果，谢谢。

郭朝晖：这个问题我不是特别了解，就是铁水到钢水之间的东西，我想如果让我考虑这个问题的话，我首先问你第一个问题，你做这个东西价值何在？这是第一个。第二个，它的整个过程，我为了达到这个价值，它需要有什么样的功能？为了达到这个功能，我需要有什么样的数据？我对数据的要求是什么？我会这样给铺开，然后我会考虑每个地方的约束，比如说它的可行度，以及它花的钱，然后这么分解下来，哪套路是可行的，如果我确实找不到一条安全可靠，达到我功能要求的东西，这个事我宁可不做。大概就是这个逻辑。

整理：张嘉元

校对：祁德力、辛洪录

编辑：张梦