

# 国科大“大数据高端论坛”致辞

李国杰

首先，我代表此次论坛的主办方之一，中国科学院大学计算机与控制学院，感谢各位领导和专家学者在岁末年初的百忙之中冒着严寒来参加这次大数据高端论坛。上个月下旬，我因视网膜脱落做了右眼波切手术，手术后需要俯卧修养一个月，每天要眼睛朝下趴 7 个小时，所以今天不能全程参加大会聆听诸位的精彩报告，十分抱歉，会后我会认真拜读各位提供的 PPT 和有关资料。利用这个简短的致辞我讲三点意见。

- 第一点。这次会议不仅邀请了不少大学和科研单位的专家学者到会，还特意邀请了与大数据有关的企业和政府部门的领导与科技人员参加，这是因为大数据的兴起的主要动力是企业的经济利益。生物信息学、脑科学、空间科学等基础研究早就采用数以 PB 级计的大数据，但没有引发大数据的浪潮，大数据如今引起各方面的重视，主要是因为大数据蕴藏着巨大的经济价值。因此我国开展大数据研究一定要以企业为主体，要充分调动企业和经济界学者的积极性。

企业和政府部门开展大数据的研究和应用不必过于在意我们正在进行的数据分析是不是“大数据”，不必花精力对各种名词和定义做无谓的争论。实际上目前大多数单位还是处于“小数据”处理阶段，但只要在纵向上有一定的时间积累，在横向上有较丰富的记录细节，通过多个源头对同一个对象采集的各种数据的有机整合，进行认真仔细的数据分析，就可能产生大的价值。

大数据在经济领域应用的一个很有趣的案例是通过社交网络预测股价涨落。社交媒体监测平台 DataSift 监测了 Facebook IPO 当天 Twitter 上的情感倾向与 Facebook 股价波动的关联。在 Facebook 开盘前 Twitter 上的情感逐渐转向负面，25 分钟之后 Facebook 的股价便开始下跌。而当 Twitter 上的情感转向正面时，Facebook 股价在 8 分钟之后也开始了回弹。最终当股市接近收盘、Twitter 上的情感转向负面时，10 分钟后 Facebook 的股价又开始下跌。最终的结论是：Twitter 上每一次情感倾向的转向都会影响 Facebook 股价的波动。预测股价被公认是很难做到的事，大数据分析也能发挥作用，其他过去认为做不到的事也许大数据分析都能派上用场。因此我们的当务之急是多实践，实践出知识，实践

出理论。只有当一个一个领域的数据分析做得相当深入时，才能提出数据科学的共性问题。

- 第二点。我国人口最多，网民最多，许多人凭想象估计：每年新数据的增量我国应该在世界上名列前茅。但是，根据麦肯锡公司的报告，2010年美国新增数据量为 3500PB（3.5EB），欧洲为 2000PB，中国仅为 250PB。也就是说，我国每年新增数据量仅为美国的 7%，欧洲的 12%。仅就数据量而言，我国还不是数据的大国，在大数据领域，我国与世界先进国家并不是在同一条起跑线上，对我国在大数据方面的技术基础落后要有清醒的认识。当然，数据量不是越多越好，大数据的关键并不在数据原料的多少，而在于获取有效信息的能力和数据分析能力的高低。

虽然现在我国已形成“大数据热”，涉及大数据的会议很多。但政府、企业和学术界是不是真正理解大数据的重大意义还很难说。许多人对变革性技术的效益预计往往是，近期过于乐观，远期估计不足。如同云计算、物联网一样，近几年内大数据的市场规模不会很大，全世界大概只有数百亿美元。但大数据市场的增速将是传统 IT 产业的好几倍。开展大数据研究和应用一定要有长远眼光，对近期的市场不能估计过高，不能自己忽悠自己。

- 第三点。对政府部门提几点建议。
  - ✧ 数据既然是一种资源，就应当像管理国土资源、矿产资源一样，有一套有法必依的法规。对国家所有的共享数据资源，应积极推动数据源建设和开放共享。研究出台鼓励性或约束性政策，推动各级政府，尤其是地理信息、医疗、交通、教育等公共服务部门，在不泄露国家秘密和个人隐私的前提下开放数据，真正实现数据共享。对个人信息，也要积极推动有关的立法工作，探索通过技术标准、行业自律等手段解决法律出台前的个人信息保护问题。
  - ✧ 成立国家级的行业大数据共享联盟，使工业界、学术界及政府部门都能够参与进来，为学术研究和大数据的应用提供基本的数据资源。
  - ✧ 加大数据关键技术研究力度。建议国家启动大数据科学与工程研究计划，设立类似自然科学重大研究计划（基金重大）以及重大基础科学研究项目群（973 项目群或 863 重大项目）等专项资金。