



大数据技术及其水利应用

河海大学 朱跃龙

2017年3月30日 南京

主要内容

1

大 数 据 概 念

2

大 数 据 技 术

3

大 数 据 水 利 应 用

第一部分

大 数 据 概 念

大数据概念

- 大数据（Big Data）最早由全球知名咨询公司麦肯锡提出
- 2008年9月，《科学》杂志发表文章“Big Data：Science in the Petabyte Era”，“大数据”这个词开始广泛传播。

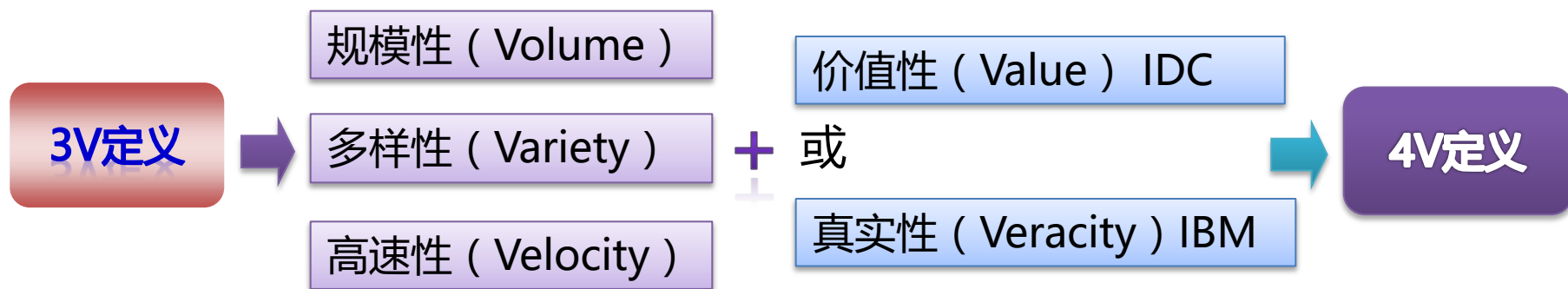


大数据概念

- 美国2012年出台《大数据的研究和发展计划》
- 英国2013年出台《英国数据能力发展战略规划》
- 澳大利亚2013年发布了《公共服务大数据战略》
- 日本2013年制定《创建最尖端IT国家宣言》
- 2015年8月19日国务院发布《关于促进大数据发展的行动纲要》；各省相继推出计划，例如，《上海推进大数据研究与发展三年行动计划》、《江苏省大数据发展行动计划》、《重庆市大数据行动计划》和《贵州省大数据产业发展应用规划纲要》等。

大数据概念

大数据意指一个超大的、难以用现有常规的数据库管理技术和工具处理的数据集。

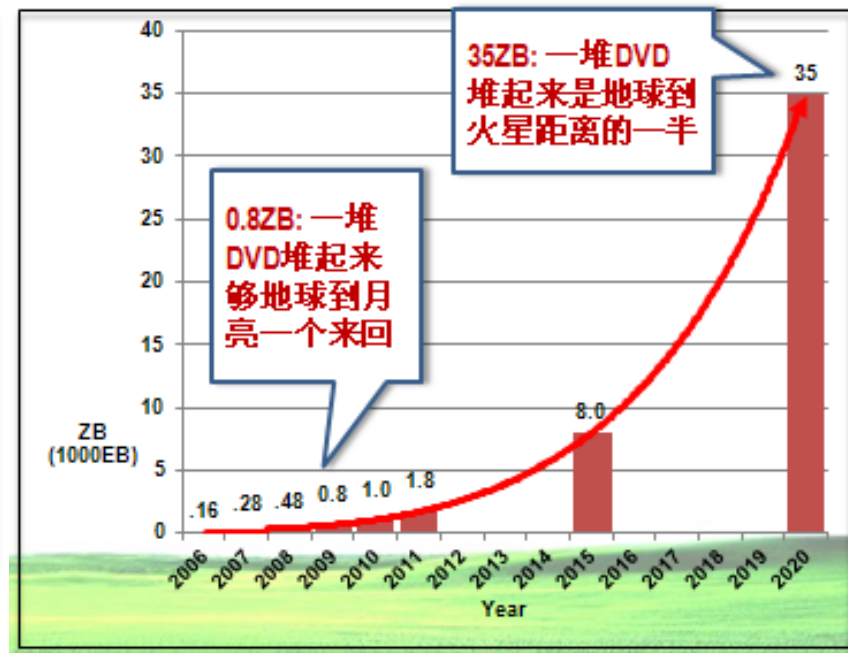


对于大数据尚未有一个公认的定义，不同的定义基本是从大数据的特征出发，加以阐述归纳给出定义。比较有代表性的是3V定义，以及新增加的4V定义（4V定义中的第四个V的含义还未统一）。

大数据概念

单位	定义	字节数 (2进制)	字节数 (10进制)
Kilobyte (千)	1024-Byte	2^{10}	10^3
Megabyte (兆)	1024- Kilobyte	2^{20}	10^6
Gigabyte (吉)	1024-Megabyte	2^{30}	10^9
Terabyte (太)	1024-Gigabyte	2^{40}	10^{12}
Petabyte (拍)	1024-Terabyte	2^{50}	10^{15}
Exabyte (艾)	1024-Petabyte	2^{60}	10^{18}
Zettabyte (泽)	1024-Exabyte	2^{70}	10^{21}
Yottabyte (尧)	1024-Zetabyte	2^{80}	10^{24}

大数据



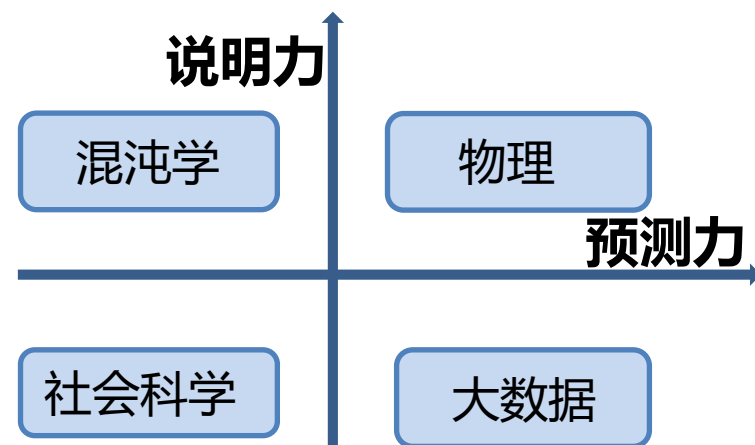
未来10年全球数据量将以40+%的速度增长，2020年全球数据量将达到35ZB（35,000,000PB），为2009年（0.8ZB）的44倍。

大数据概念

计算机图灵奖获得者Jim Gray：提出了一种新的数据探索型研究方式，称之为科学研究的“第4种范式”（The Fourth Paradigm）



- 科学的两个主要目标：解释（说明）与预测。
- 大数据方法解释能力弱（Google翻译不懂语法）
- 大数据方法预测能力强（特别适合做情报分析）
- 大数据技术在智能性问题的出色结果（如IBM沃森、Google的AlphaGo）令人“难以理解”，正是大数据方法的特征。



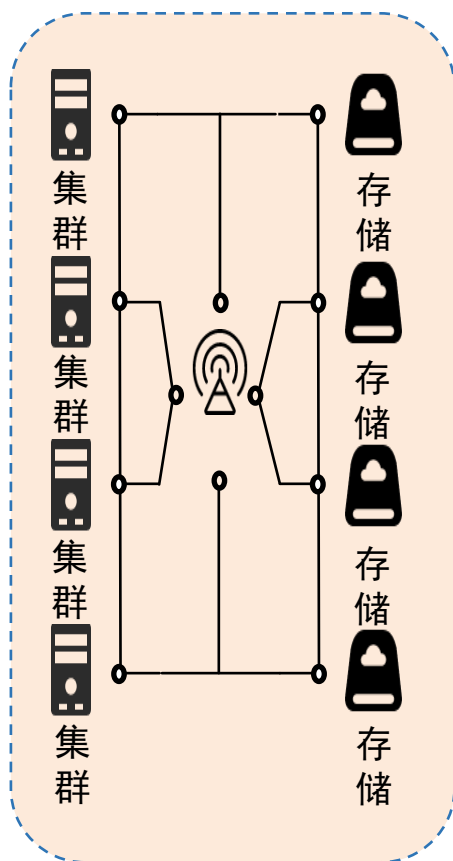
第二部分

大 数 据 技 术

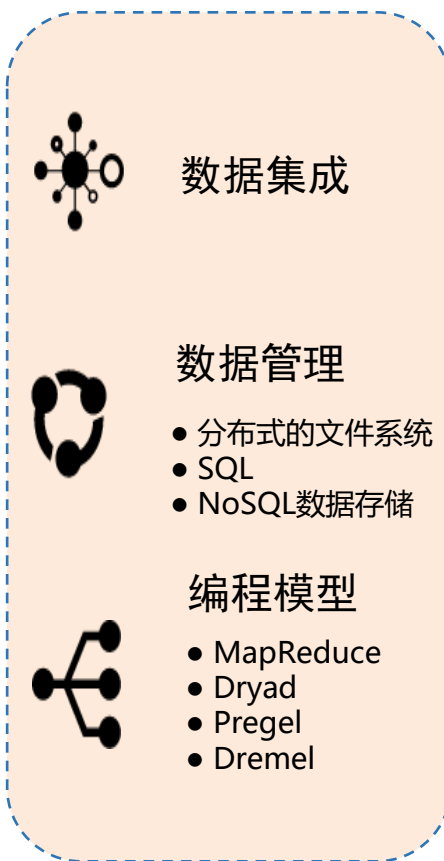
大数据技术

大数据系统的层次结构 / Layered Architecture

基础设施层



计算层



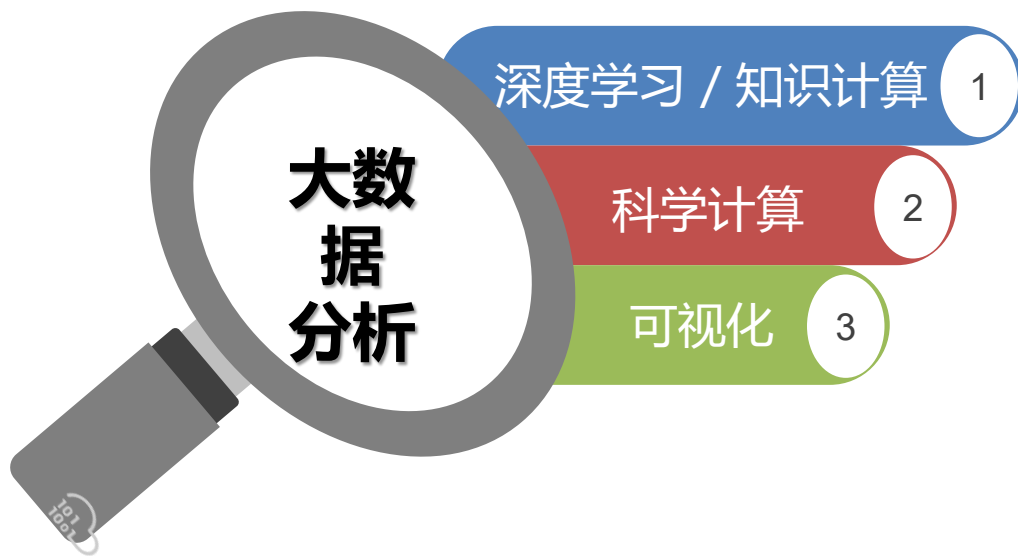
应用层



大数据技术

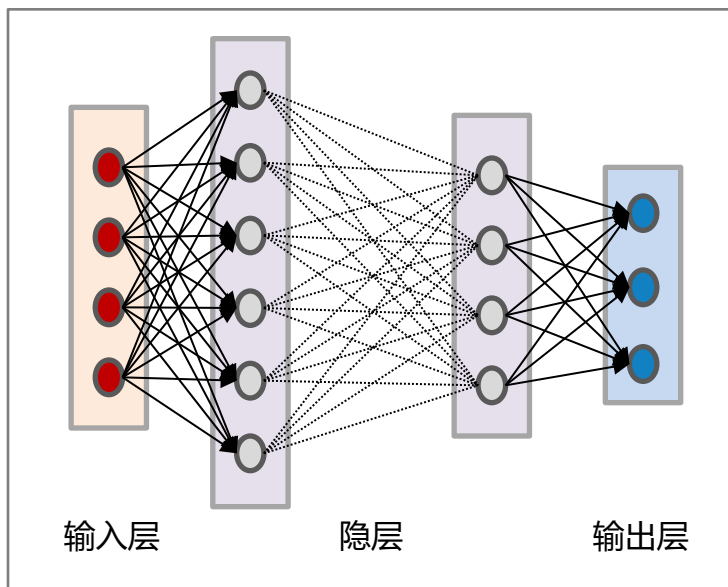
大数据分析 / Analysis

要挖掘大数据中所蕴涵演变规律和有价值的信息，就必然要对大数据进行内容上的分析与计算，即大数据分析。



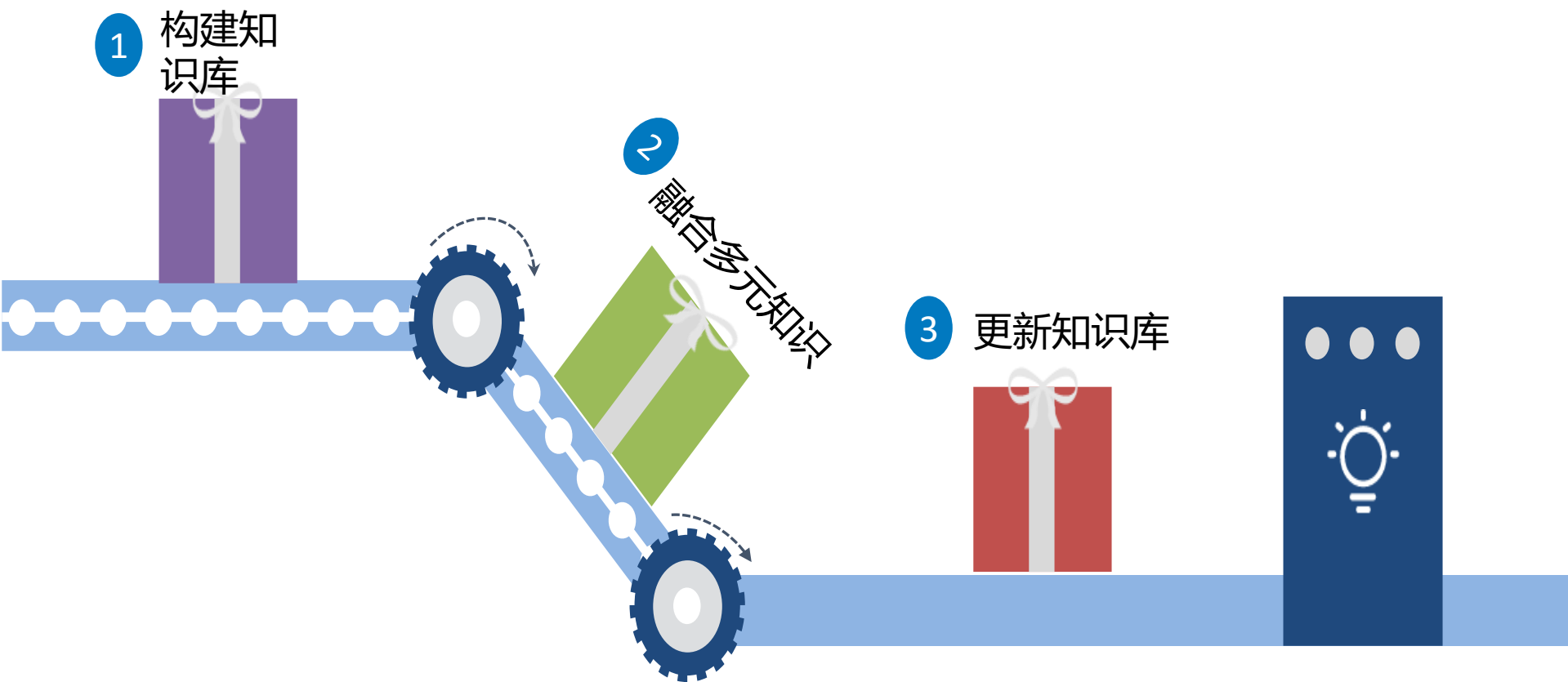
深度学习 / Deep learning

深度学习是一种基于机器学习、数据挖掘技术以及神经网络理论，分析大数据潜在价值的过程，利用层次化的架构学习出对象在不同层次上的表达，从而解决复杂抽象的问题



知识计算 / Knowledge Computing

基于大数据的知识计算是大数据分析的基础。目前，世界各国各个组织建立的知识库多达 50 余种，相关的应用系统更是达到了上百种。



科学计算 / Scientific Computing

由于大数据存在复杂、高维、多变等特性，如何从真实、凌乱、无模式和复杂的大数据中挖掘出人类感兴趣的知识，迫切需要更深刻的机器学习理论进行指导，使用系统科学、人工智能、数据挖掘等科学计算，从而解决政治、经济、文化等领域复杂性社会问题的一种理论和方法论体系

大数据聚类

聚类所面临的不仅是数据量越来越大的问题，更重要的还是数据的高维问题。

大数据关联分析

解决大数据的关联分析主要有两种途径：
并行和增量。

大数据并行算法

把传统机器学习算法运用到大数据环境中，
一个典型策略是对现有学习算法并行化。



大数据分治策略

数据分治与并行处理策略是大数据处理的基本策略。

大数据特征选择

大数据存在复杂、高维、多变等特性，采用降维和特征选择技术以降低大数据处理难度。

大数据分类

- 支持向量机分类
- 决策树分类
- 神经网络与极端学习机

可視化 / Visualization

大数据可视化，不同于传统的信息可视化，面临最大的一个挑战就是**规模**，如何提出新的可视化方法能够帮助人们分析大规模、高维度、多来源、动态演化的信息，并辅助作出实时的决策，成为了这个领域最大的挑战。



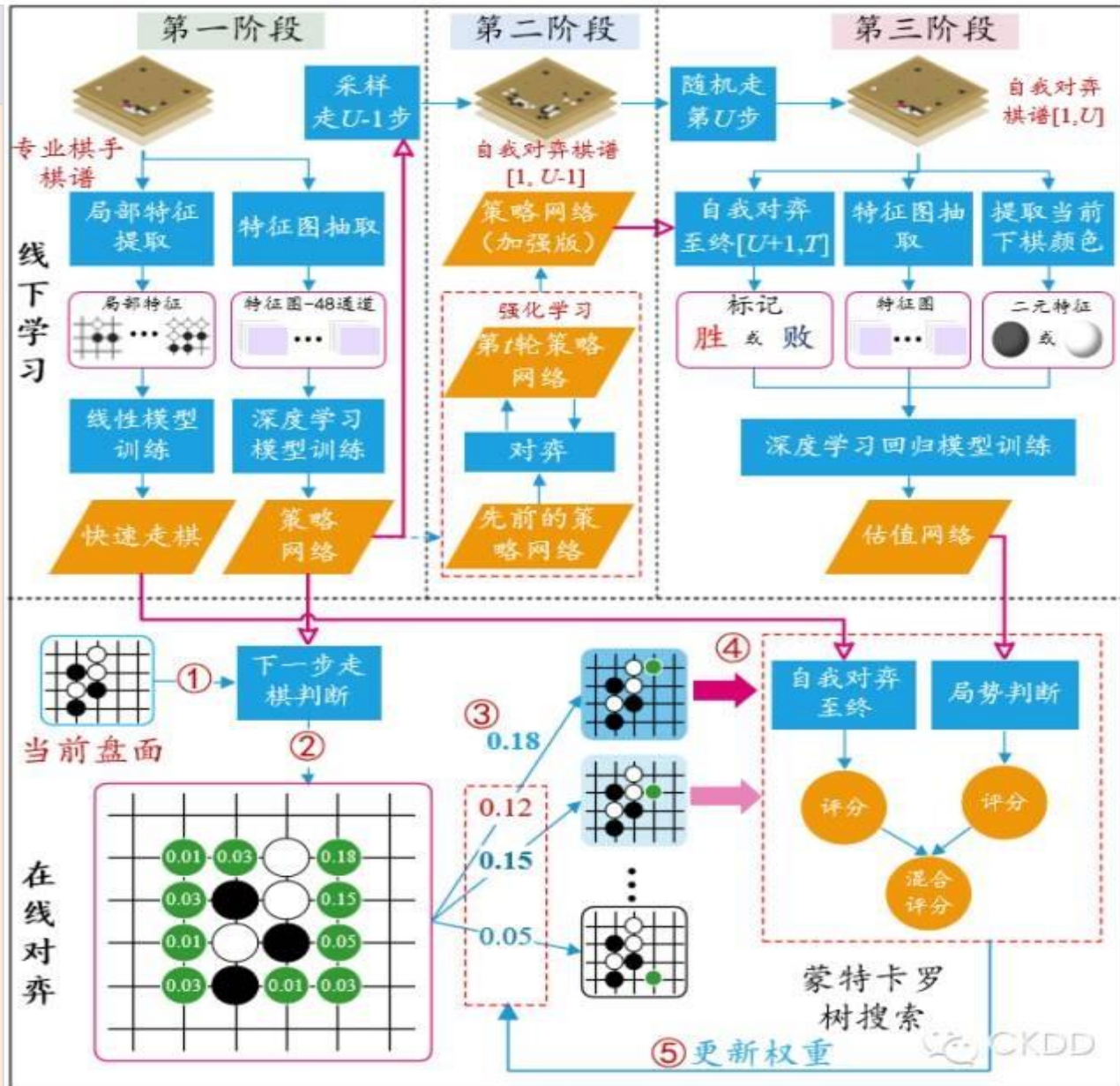
神奇的 AlphaGo

线下部分

- 专业棋谱训练
- 策略网络互博
- 价值网络生成

线上部分

- 提取棋盘特征
- 估计落子概率
- 估计走势权重
- 蒙特卡罗搜索



线下部分，离线学习

- 利用3万多幅专业棋手对局的棋谱来训练两个网络。一个是基于全局特征和深度卷积网络(CNN)训练出来的**策略网络**(Policy Network)，其主要作用是给定当前盘面状态作为输入，输出下一步棋在棋盘其它空地上的落子概率。另一个是利用局部特征和线性模型训练出来的**快速走棋策略**(Rollout Policy)。策略网络速度较慢，但精度较高；快速走棋策略反之。
- 利用第t轮的策略网络与先前训练好的策略网络互相对弈，利用增强式学习来修正第t轮的策略网络的参数，最终得到增强的策略网络。
- 先利用普通的策略网络来生成棋局的前U-1步(U是一个属于[1, 450]的随机变量)，然后利用随机采样来决定第U步的位置(这是为了增加棋的多样性，防止过拟合)。随后，利用增强的策略网络来完成后面的自我对弈过程，直至棋局结束分出胜负。此后，第U步的盘面作为特征输入，胜负作为label，学习一个价值网络(Value Network)，用于判断结果的输赢概率。用3000万盘棋局训练价值网络。

线上部分，在线对弈

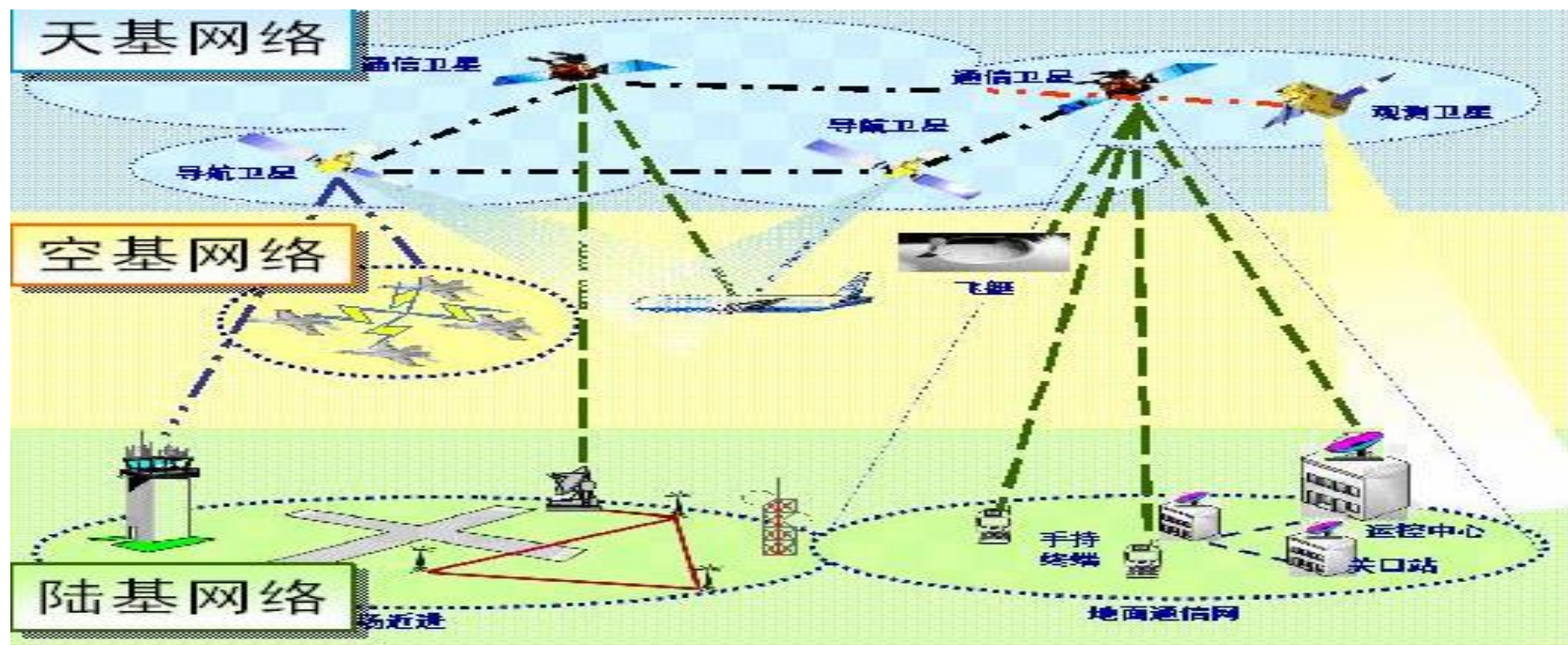
- 根据当前盘面已经落子的情况提取相应特征；
- 利用策略网络估计出棋盘其他空地的落子概率；
- 根据落子概率来计算此处往下发展的权重，初始值为落子概率本身。实际情况可能是一个以概率值为输入的函数。
- 利用价值网络和快速走棋网络分别判断局势，两个局势得分相加为此处最后走棋获胜的得分。
- 利用第四步计算的得分来更新之前那个走棋位置的权重；从权重最大的那条边开始继续搜索和更新。当某个节点的被访问次数超过了一定的门限值，则在蒙特卡罗树上进一步展开下一级别的搜索。

第三部分

大数据水利应用

大数据水利应用

众所周知，我国的水问题依然严峻，因此国家投入了巨额资金对水利设施进行修建改造，对水文、气象等水利多要素进行观测。如：国家防汛抗旱指挥系统工程、国家水资源监控能力建设工程、地下水监控工程、中小河流监测预警工程、山洪灾害风险监测预警工程、病险水库加固工程等。初步形成了天空地网观测与数据采集体系。



大数据水利应用

根据国家民用空间基础设施中长期发展规划（2015-2025年），我国卫星遥感空间系统包括：

陆地观测卫星系列：高分辨率光学观测星座、中分辨率光学观测星座、合成孔径雷达（SAR）观测星座、地球物理场探测卫星。

海洋观测卫星系列：海洋水色卫星星座、海洋动力卫星星座、海洋监视监测卫星。

大气观测卫星系列：天气观测卫星星座、气候观测卫星星座、大气成分探测卫星。

目前，可应用于水利行业的卫星有：GF-1/2卫星、HJ-1A/B卫星、ZY-3卫星、ZY-102C卫星、SAR系列卫星、以及MODIS卫星等。

卫星型号	重要参数		
HJ-1A/B	30米分辨率，	幅宽700公里，	重访周期4天
ZY-102C	2.36米分辨率，	幅宽54公里，	重访周期3天
ZY-3	2.1米分辨率，	幅宽51公里，	重访周期5天
GF-1	2米分辨率，	幅宽800公里，	重访周期4天
GF-2	0.8米分辨率，	幅宽45公里，	重访周期5天

大数据水利应用

- **天** 利用卫星观测，如：HJ、MODIS、ZY、高分卫星等，获得卫星遥感影像。
- **空** 空中观测，如：雷达测雨，无人机遥测等。
- **地** 各类地面观测设施，如：降雨、流量、水位、水工建筑物应力应变、位移、温度等。
- **网** 通过网络获取的数据，如：雨量预报，气象等。
- **专题数据** 如：DEM、气象、土地利用等专题数据。

水利数据量十分巨大，增长非常迅速。以水文数据为例，水文关系数据近100TB，矢量与遥感影像数据近10PB(10000TB)，每年新增数据>1PB(1000TB)。

大数据水利应用：洪水预报

我国江河众多，如果把中国的天然河流连接起来，总长度可达43万公里，可绕地球赤道10.5圈。其中流域面积在50平方公里以上的河流有4.5万多条；流域面积在1000平方公里以上的河流1300多条，流域面积大于1万平方公里的河流超过100条。

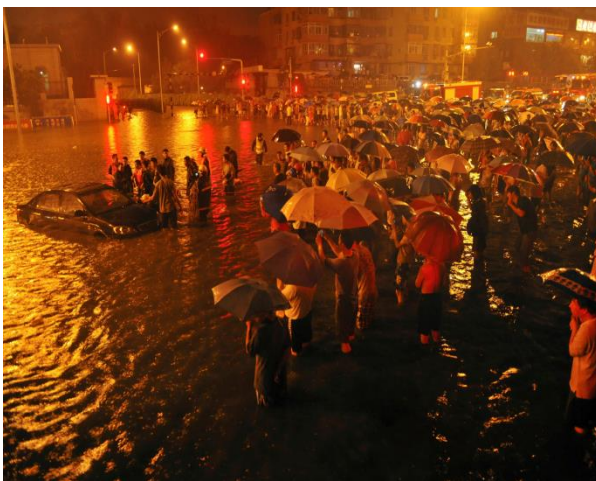


大数据水利应用：洪水预报

众多江河，也为防洪带来了巨大压力。1950年以来，因洪涝灾害死亡人口，50-70年代约为每年6000人，80-90年代约为每年4100人，本世纪前10年平均为每年1600人。直接经济损失占GDP的2%，是美国的近百倍。



流域大洪水



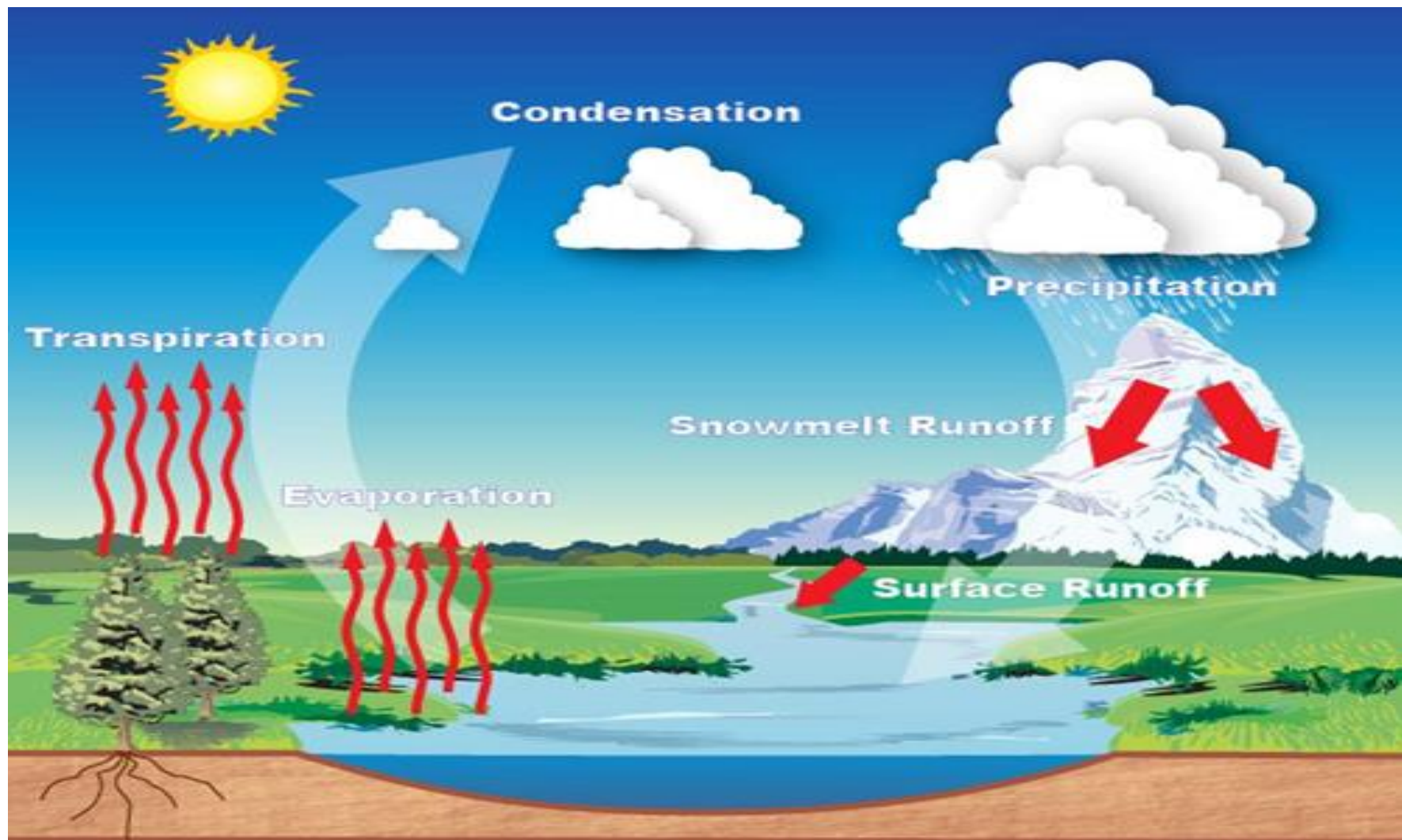
城市大暴雨



山洪泥石流

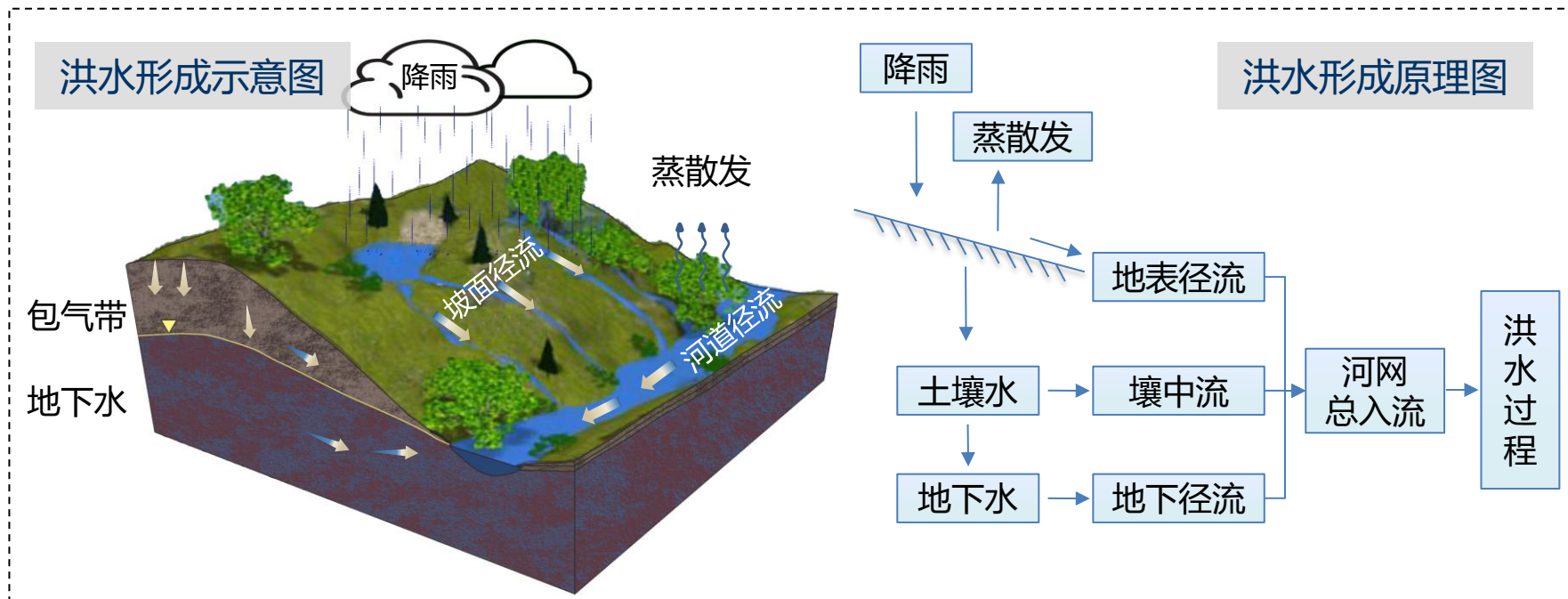
大数据水利应用：洪水预报

认识水循环演变规律与孕灾机理，对防洪减灾工作十分重要



大数据水利应用：洪水预报

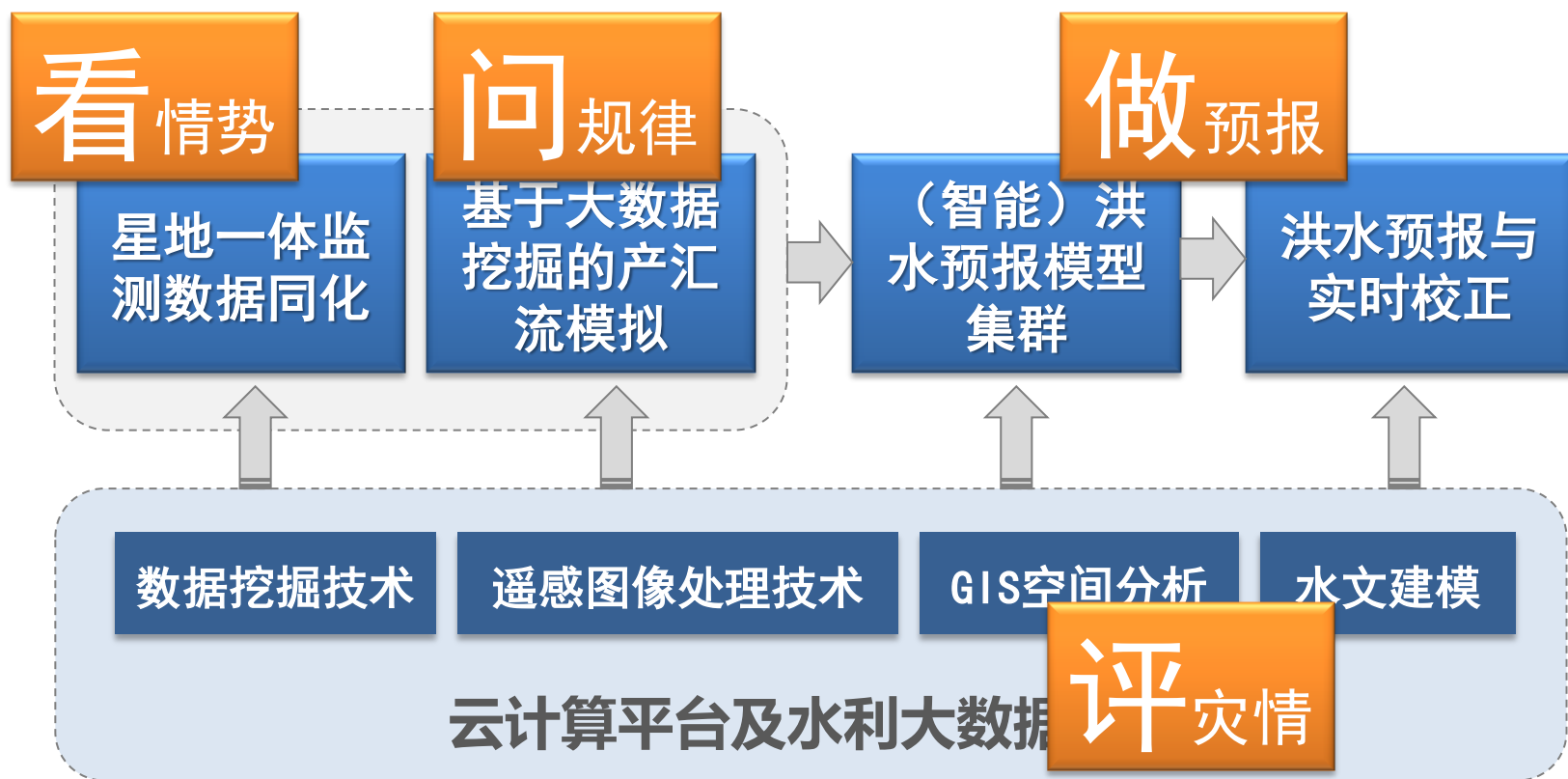
洪水预报是防洪减灾最重要的非工程措施，高精度预报是建立在暴雨洪水形成的水文多要素物理过程精细刻划的基础上。



精细刻划的核心是水文、水利和气象模式的结合，其难点在于需要跨越不同模式之间的尺度差异，即将气象要素尽可能准确地降尺度到不同河道上，并且河道上下游的计算也有差异。此外，为了确保降雨和径流预报的准确性，需要同化大量雷达降水等气象资料和其他实时观测数据等。在确保业务预报时效性的前提下，需要有效解决大数据和大计算量下的一系列问题。

大数据水利应用：洪水预报

利用**大数据挖掘新技术**对预报**众多影响因素**进行分析，挖掘大数据中蕴含的水文时空变化特征，通过**多源信息同化**提高模型输入参量的时空分辨率，结合**智能水文规律发现**突破水文预报的**产汇流空间变异大、下垫面变化非平稳**等瓶颈，提高水文预报的精度、延长预见期。



大数据水利应用：洪水预报

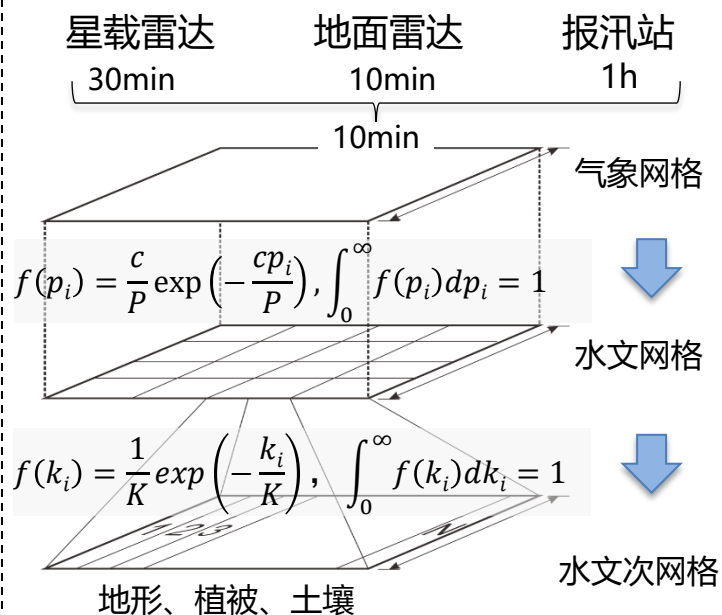
基于云平台的星地一体水利监测业务化系统，能够有效提高洪涝灾害、突发水事件监测预警时效性和精度



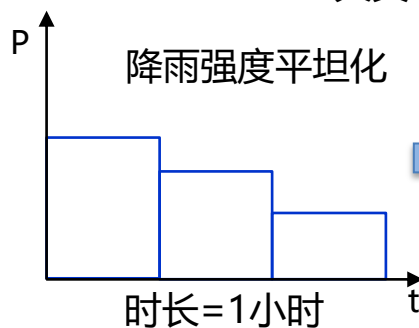
大数据水利应用：洪水预报

雨量站的监测数据难以用来确定暴雨中心和高精度雨量场，基于星载雷达、地面雷达和雨量站的监测数据，采用多源雨量的融合算法和考虑地形的降尺度方法，构建高精度分钟级雨量场，精准定位暴雨中心

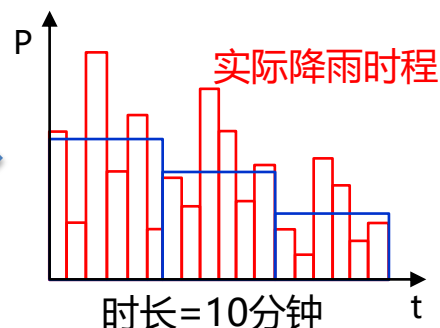
多源雨量融合与降尺度



常规方法

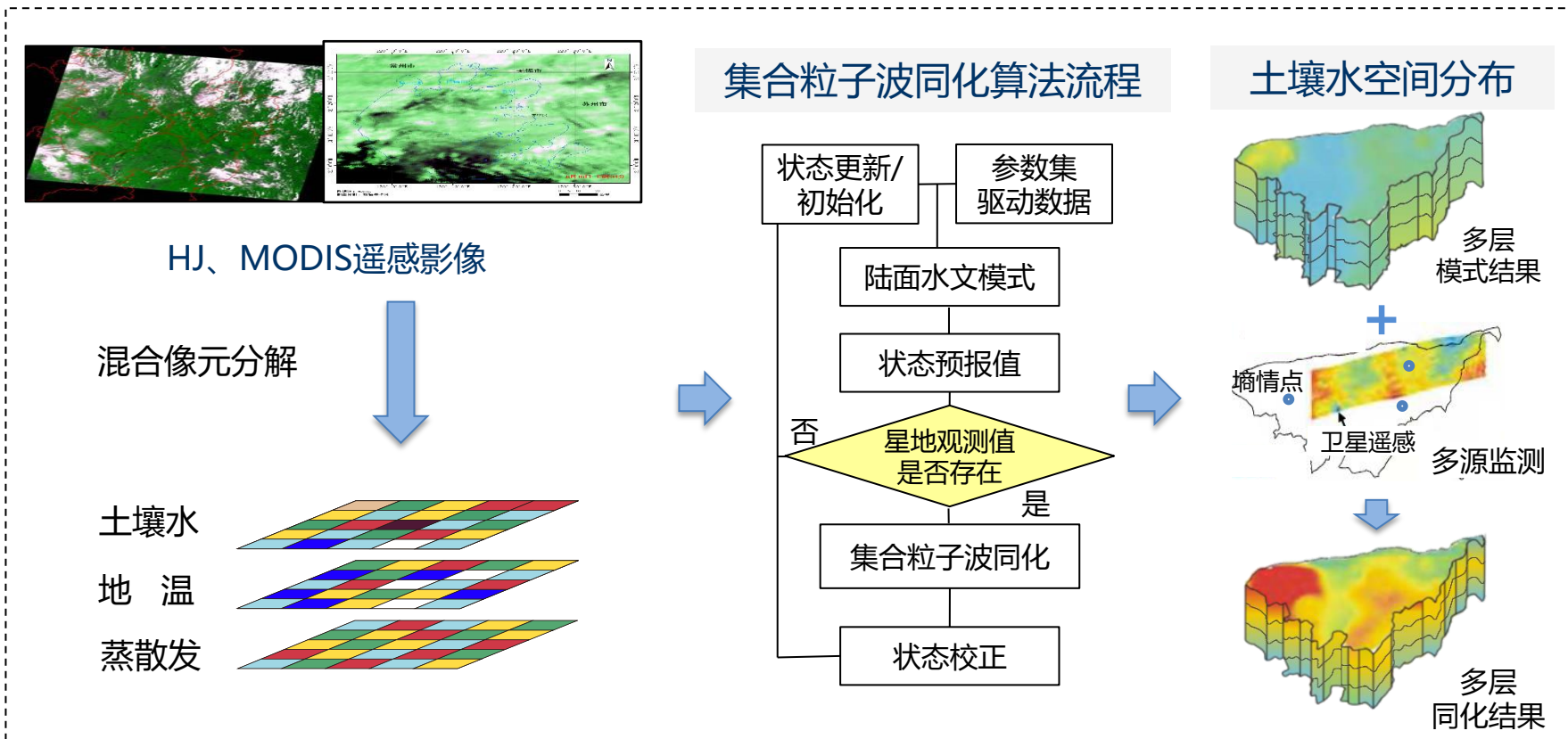


改进方法



大数据水利应用：洪水预报

仅利用地面墒情站监测信息，难以估算流域尺度土壤水。基于HJ星、MODIS遥感影像，采用混合像元动态分解进行土壤水遥感反演；结合陆面水文模式进行多要素耦合的土壤水分层同化，实现土壤水精准计算



大数据水利应用：洪水预报

针对水体提取易受薄云、山体阴影影响，提出了自适应特征选择的水体提取方法，提高了塘坝等小水体计算精度

提取关键算法

图像分割：概率密度函数的核密度估计

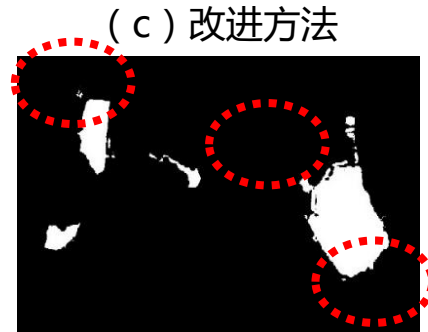
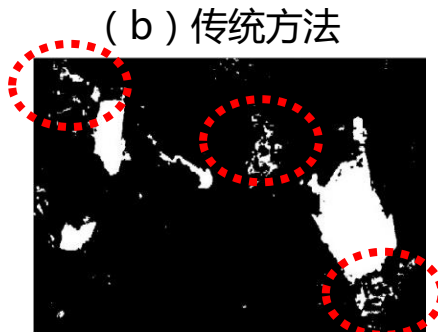
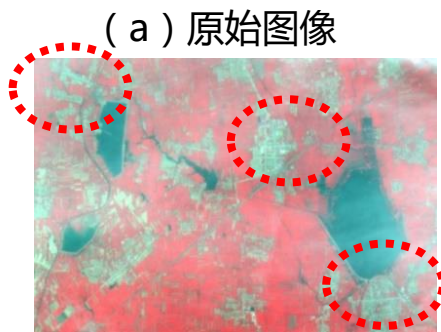
$$\hat{v}f_{h,k} = \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right) \right] \left[\frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x \right]$$

自适应特征选择：改进最小描述长度准则聚类分析

$$\Delta l_{c_j} = -K - n_j \log_2 p_j + \sum_{k=1, k \neq j}^m n_{jk} \log_2 \left(\frac{n_k + n_{jk}}{|I|} \right) + \sum_{x \in c_j} \sum_{i=1}^d \frac{(x_i - c_{ik})^2 - (x_i - c_{ij})^2}{2(\ln 2)\sigma^2}$$



方法效果对比

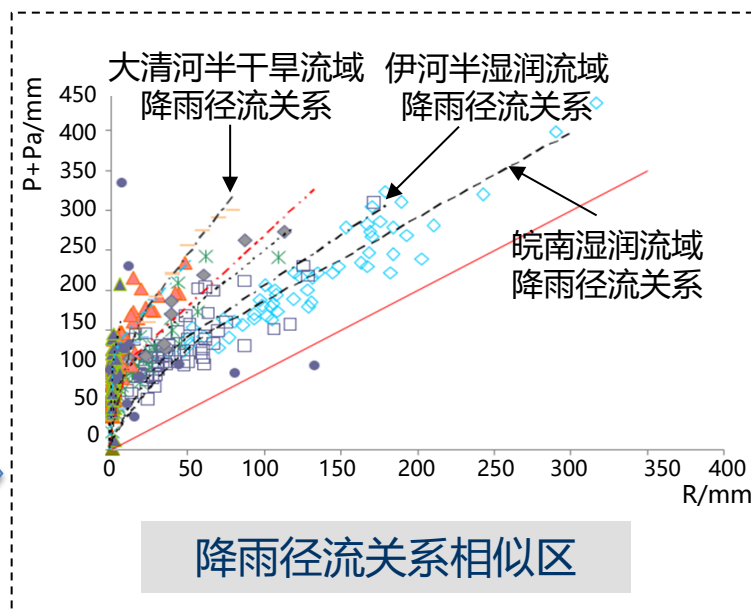
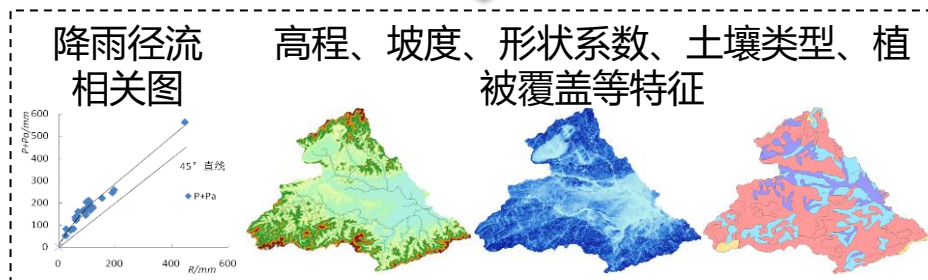


面积 $\geq 50\text{m}^2$ 的小型水体提取精度 $\geq 85\%$

大数据水利应用：洪水预报

针对无资料区径流预报难题，基于流域相似度挖掘方法，确定降雨径流响应的相似区，实现降雨径流关系向无资料区的移植

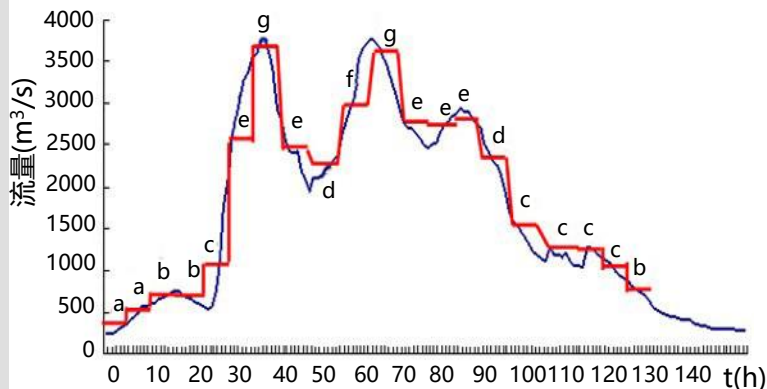
37个湿润、半湿润半干旱流域，2000多场洪水的降雨径流关系，地形地貌、土壤植被等32类下垫面特征



大数据水利应用：洪水预报

典型洪水特征模式挖掘 针对暴雨-历史洪水相似问题，进行暴雨洪水响应语义概化挖掘，构建暴雨洪水特征模式库，为实时洪水预报提供智能化校正依据

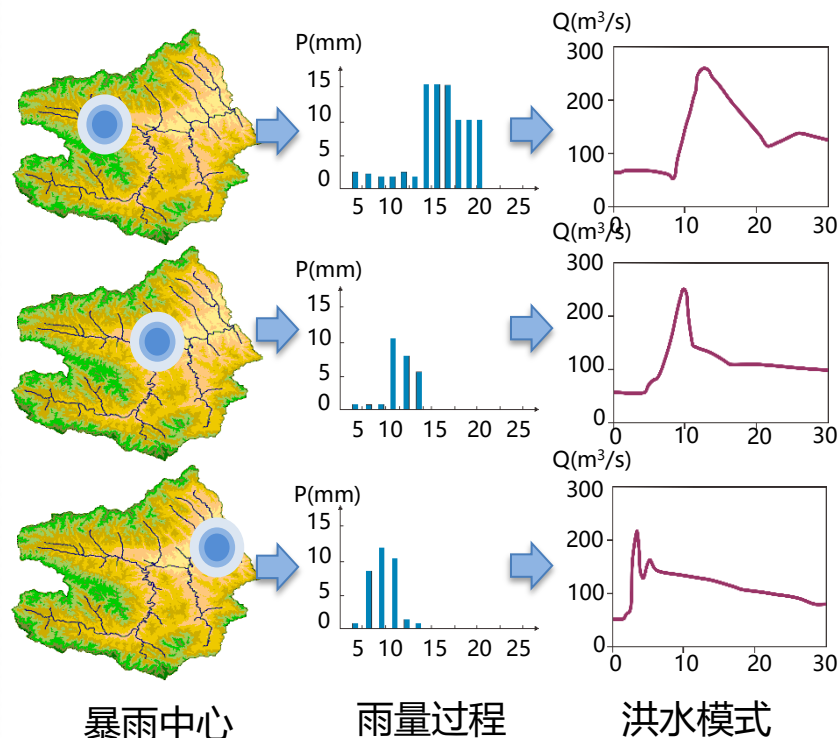
洪水语义概化



洪水模式库

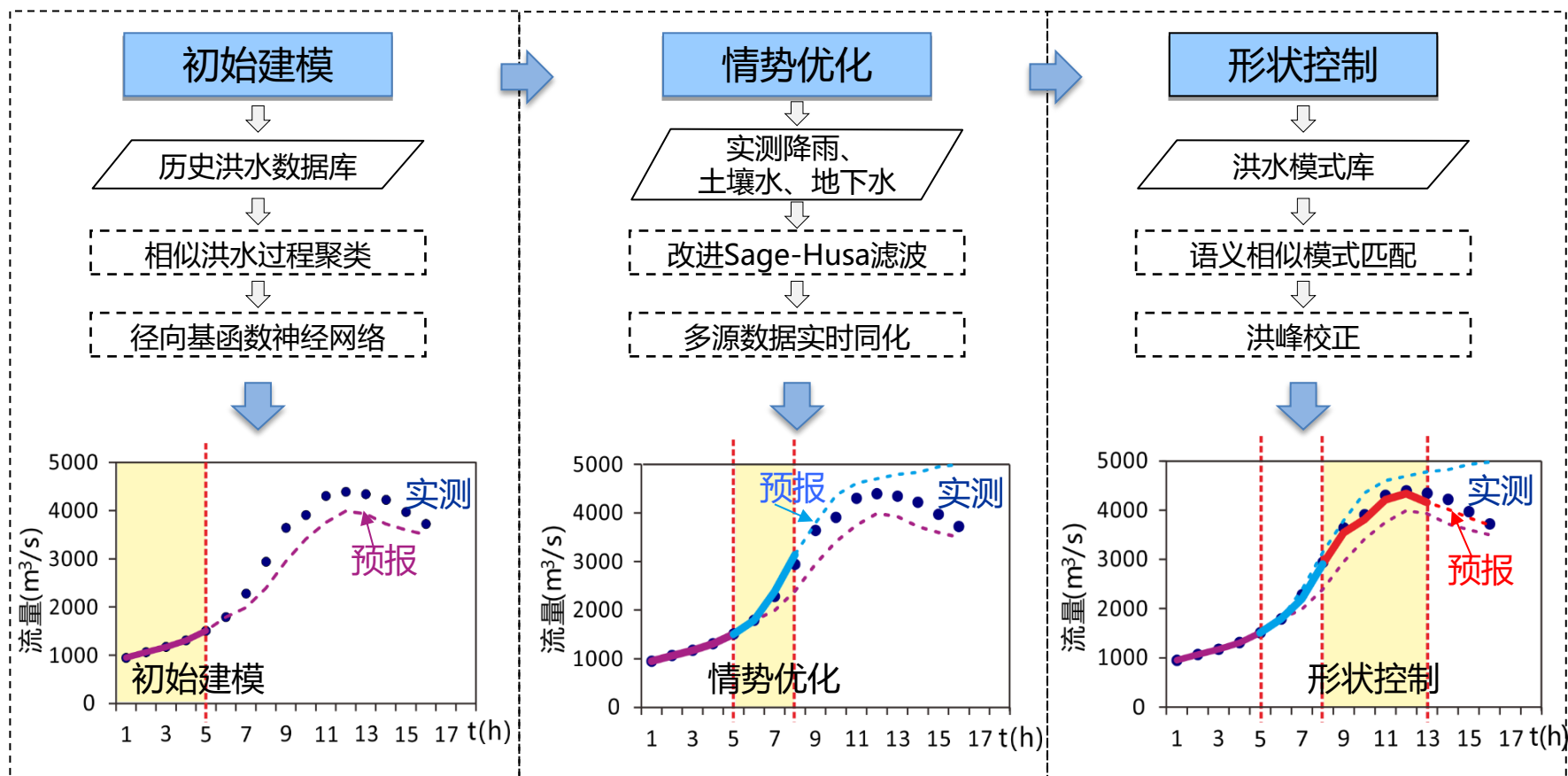
天气形势(前期降水、气温、天气环流...)
降雨时空分布(C_v, C_s , 暴雨极值(1,3,5,7天)...)
洪水过程(洪峰、峰现时间、洪量...)
洪水情势(行洪区域、区域间洪水...)
下垫面地理特征(地形指数、植被指数...)

暴雨洪水时空模式



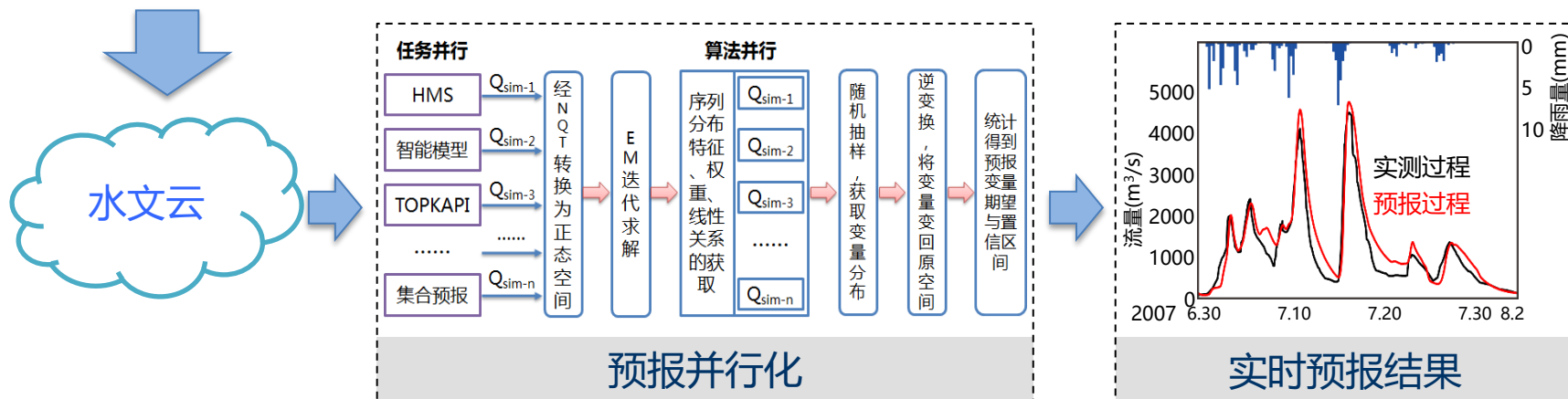
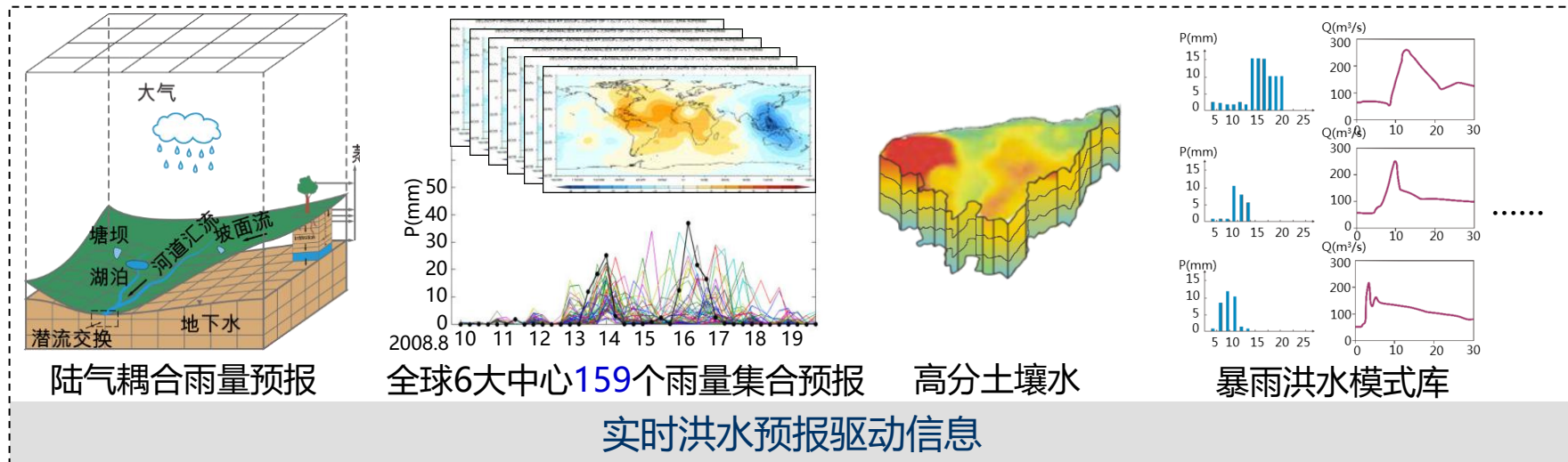
大数据水利应用：洪水预报

洪水变化情势复杂，需要在涨水、洪峰、退水阶段考虑不同因素的影响，实现过程自适应的智能预报；并基于洪水模式库进行洪峰预报校正，提高洪峰预报精度



大数据水利应用：洪水预报

基于云平台，实现大数据驱动的实时洪水预报，提高预报精度和预见期，降低不确定性



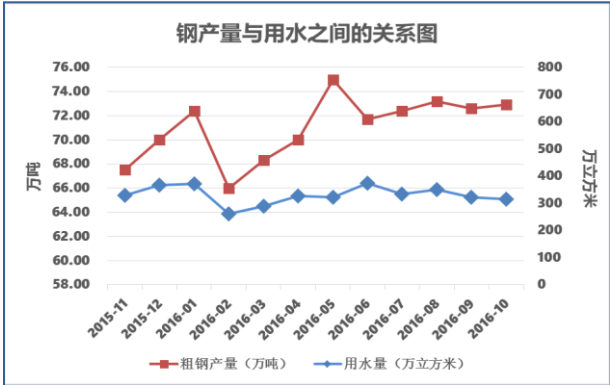
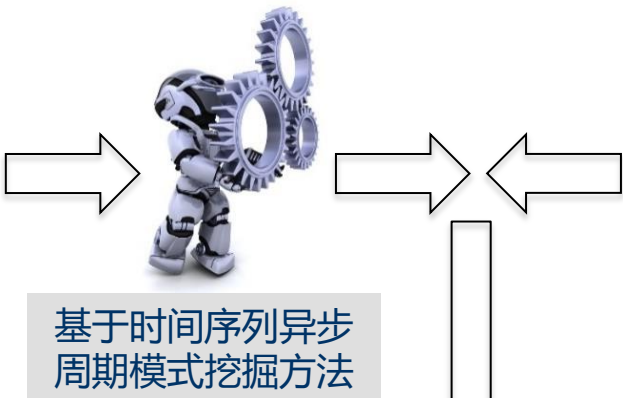
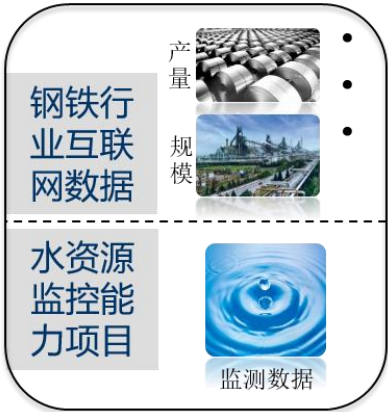
大数据水利应用：水资源监控

发现监控对象与应监控对象相比大量缺失，仅河北省钢铁企业按照规模应当监控69家，实际上只监控了15家，约缺3/4。发现钢铁产量数据与用水量数据关系不稳定，说明许可证、取水点、用水户、水表等信息不清，甚至不排除企业用水存在表外数据。

发现地方上报的监控对象的基础信息与实际差别较大，甚至有的达到几千倍差距，说明地方水利部门不掌握实际情况，也不排除隐瞒不报。

大数据水利应用：水资源监控

通过钢铁产业网站的互联网数据，利用文本挖掘技术分析出钢铁企业规模、产量等信息，与水资源监控能力项目的取水户用水监测数据，进行协同对比分析，对节水数据进行跨行业核查。



领域知识

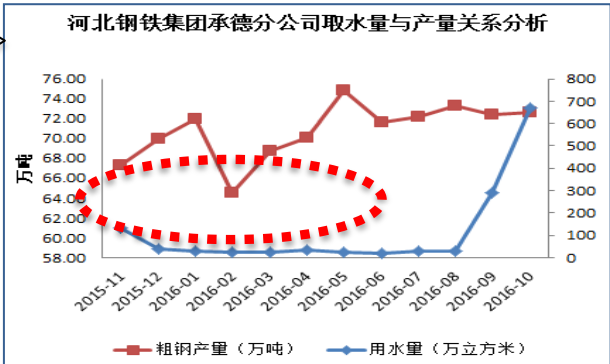
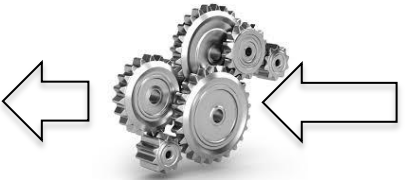
原因1：

已监控对象数量远小于应监控数量

序号	类别	企业数量	监控数量	监控百分比
1	10<P<50	1	0	0%
2	50<P<100	2	1	50%
3	100<P<500	50	8	16%
4	500<P<1000	7	4	57%
5	>1000	6	2	33%

原因2：

地方水利部门不掌握实际情况，也不排除隐瞒不报



异常模式

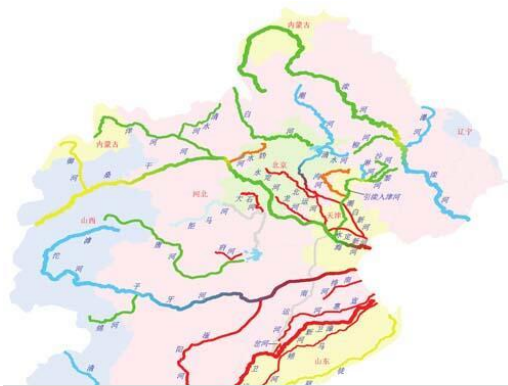
大数据水利应用：污染负荷监测

基于多部门跨层级的环渤海污染负荷监测

- 摸清污染底数是渤海治理的首要问题，涉及水利、环保、海洋、住建等中央部门，涉及山东、河北、天津、辽宁等地方各级政府。
- 利用水利部门的分省入海水量、住建部门的城镇污水排污河道、环保部门的排污河道水质评价数据、海洋部门的主要排污口、以及省市部门的主要污染源数据，进行数据综合分析与处理，形成多部门跨层级的环渤海污染负荷监测产品。



国家海洋局陆域入海排污口



环保部海河流域
地表水水质评价

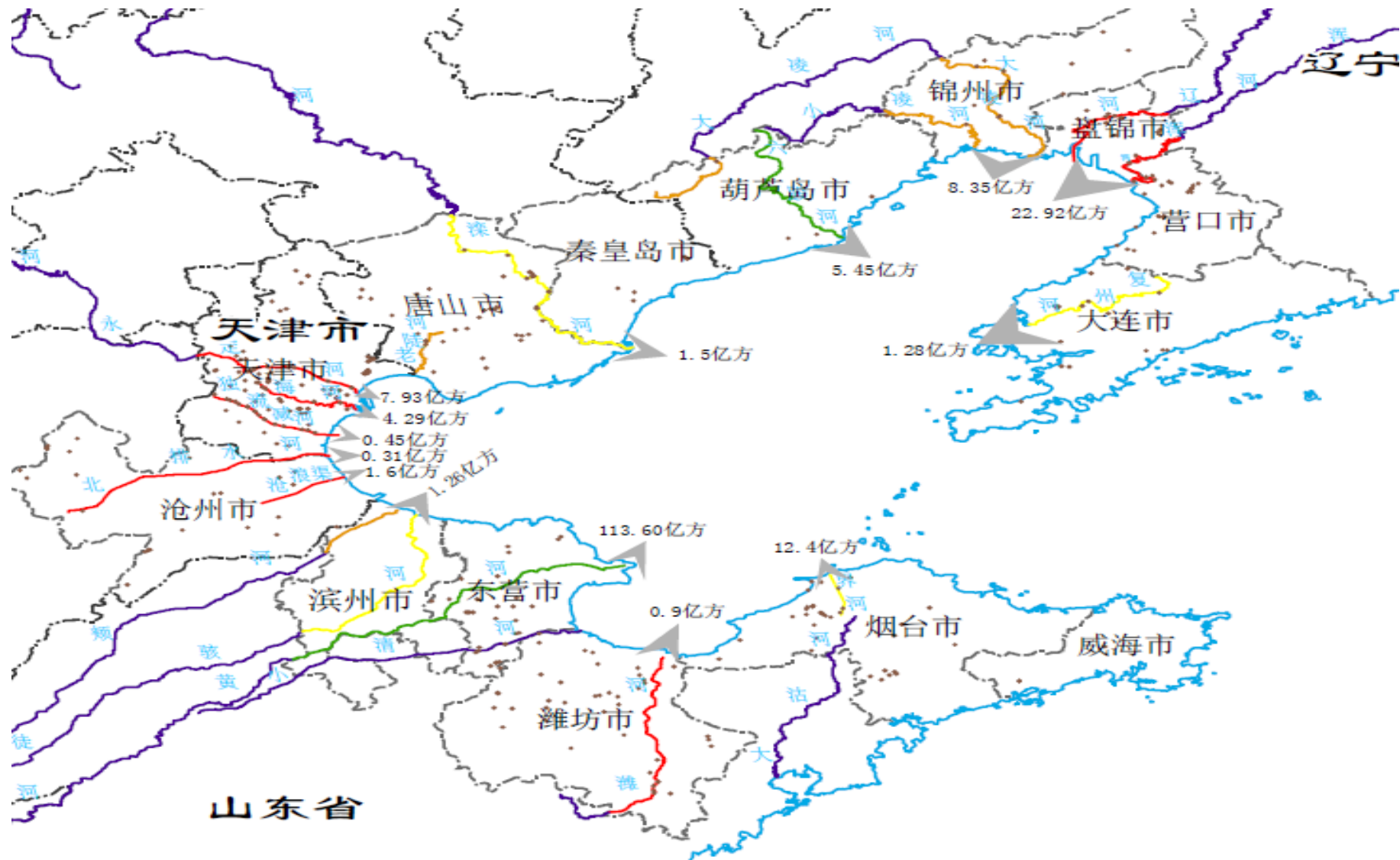


环渤海污染负荷监测产品

获得国家信息中心及国家发改委的高度认可

大数据水利应用：污染负荷监测

基于多部门跨层级的环渤海污染负荷监测



大数据水利应用：水环境舆情监控

2015年11月23日晚约9至10点，甘肃陇南市西和县陇星铋业有限公司发生尾矿库泄漏事故，大量含铋尾沙流入太石河，太石河在甘肃境内流入西汉水，并在陕西略阳汇入嘉陵江，进入四川境内，波及三省。铋矿尾沙的泄漏导致很多下游很多水域的水质污染，铋浓度超标7.9倍，对河流水生物以及下游人民的饮用水都带来了很大的影响。

大数据水利应用：水环境舆情监控

数据收集：基础地理数据、卫星遥感数据、水文数据以及气象数据，从环保等部门获取下游断面水质监控数据，实时采集社交平台互联网数据。

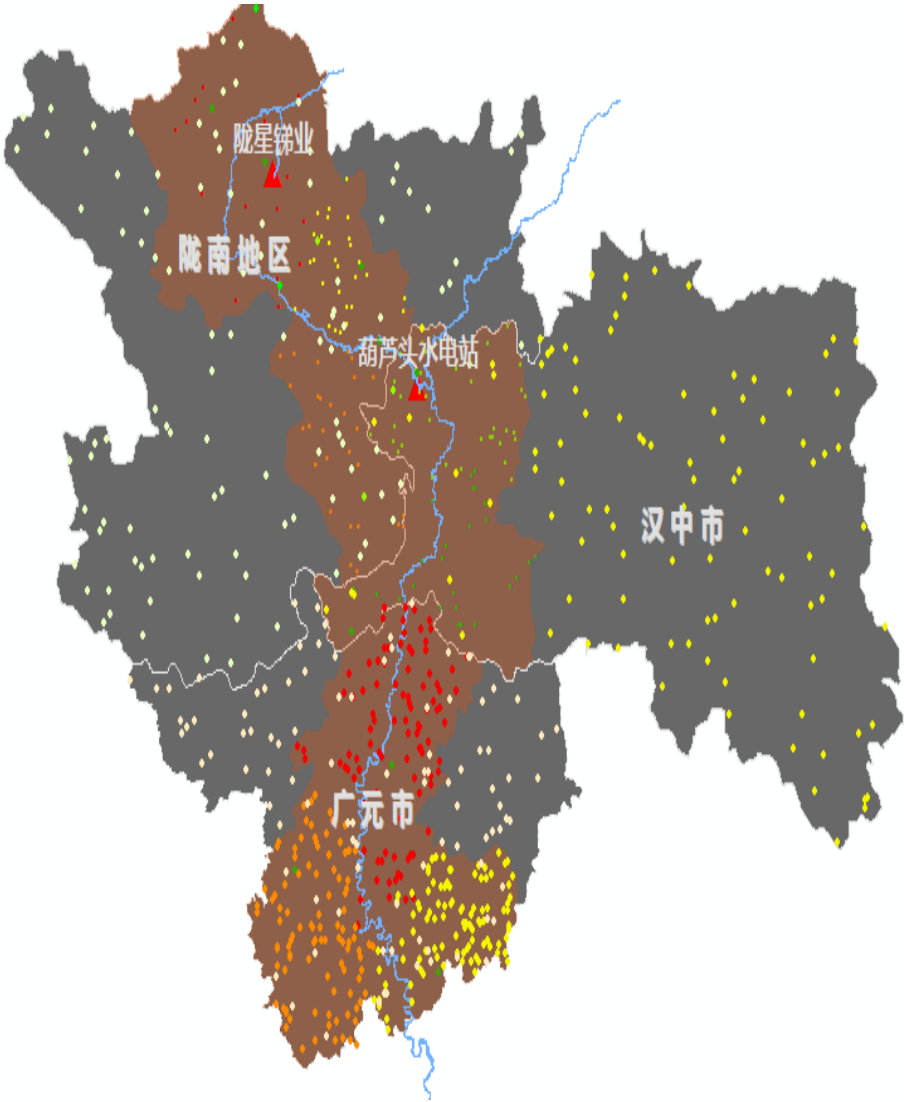
分析处理：已有数据制作基础地图，互联网社交数据建立模型判断正负能量，并对互联网数据进行空间化处理,形成舆情分析产品。

序号	断面名称	污染物	浓度	是否达标
1	西汉水甘陕入境	镉	0.1071	超标20.4倍
2	葫芦头水电站库区	镉	0.1544	超标29.9倍
3	西汉水嘉陵江入口	镉	0.0034	达标
4	嘉陵江陕西出境	镉	0.002	达标

来源	网名	发布评论内容
微信	王祖新	政府会追责吗？这种事发生了政府也有责任，监管不力。
	我在故我思6093	不知道还有别的地方有这种污染吗？
	东海弄潮	当地救灾工作又会涌现出不少先进人物吧？
	半截香烟	这种企业不把他罚款到破产有什么用/
QQ空间	五个叉终于	：已经停水了，不知道处理得怎么样了
	沧桑话悲凉	下西村停水要停到几时啊？已经一周了，受不了，受不了，受不了！

大数据水利应用：水环境舆情监控

曝光量		137万	这条微博在多少个微博用户的页面上显示
 情感值		36	对该微博的转发文本进行自然语言处理后,得出的情感倾向判断
 内容评价		40	消息传播的穿透力
 用户总体评价	用户质量	74	二级粉丝计算
	用户活跃度	29	用户平均每天发微博数
	加V比例	7.4%	新浪认证用户
	参与用户数	68	新浪反垃圾系统过滤之后的用户总数
	水军	疑似	机器操作非真实用户
 短链点击数		79	短链接的点击数与分享数

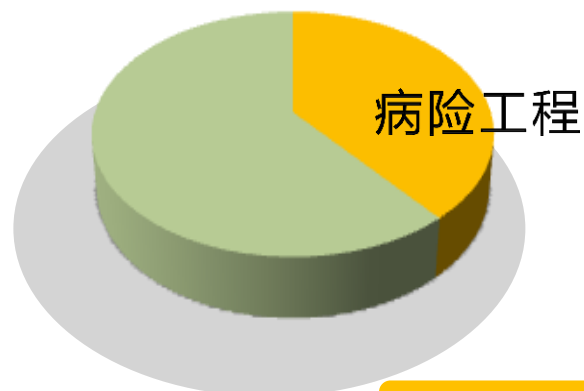


大数据水利应用：工程安全

目前，水电站工程安全监测体系已采集了海量数据，并且数据量持续快速增长。但是，如何实时高效提取和利用海量监测数据，综合分析评判大坝运行工况，为工程运行维护人员提供决策支持，成为问题。特别是，工程内置传感器超出服役年限后，如何评价大坝安全性成为难题。

- ◆ 1954年至2010年共有**3515座**水库大坝失事
- ◆ 2010年有**11座**水库大坝失事

病险水工程统计



病险水库超过 **三分之一**

大型混凝土双曲拱坝自动化监测数据

监测仪器及测点1万余支（个）

- 采集频率：每隔4小时采集一次
- 每天增加：60000余条记录
- 数据量：10年即达到TB级

水库地震监测
水情自动测报
枢纽区变形监测
拱坝安全监测
引水发电系统监测
边坡和抗力岩体监测坝
内温度裂缝专项监测
枢纽区安全监测.....

大坝激光三维变形测量数据

坝顶GNSS变形监测数据

裂缝变形测量数据.....

大数据水利应用

受自然条件的制约和社会经济高速发展的影响，我国水问题十分特出。应用的目的是打造精准水治理、形成多方协作的治理新模式。通过水利多源大数据的协同互联和开放共享可以有效消除信息孤岛，通过大数据的挖掘分析技术和人工智能应用可建立更加精准的水利专业模型，通过云计算支撑的模型校核率定和融合应用技术可提高更高精度的水利产品服务。更好地为社会经济发展做出新贡献。

敬请雅正 欢迎交流
谢谢！