



大数据近似计算



马 帅



BIG DATA
BRAIN COMPUTING
大数据科学与脑机智能高精尖创新中心



北京航空航天大学
BEIHANG UNIVERSITY

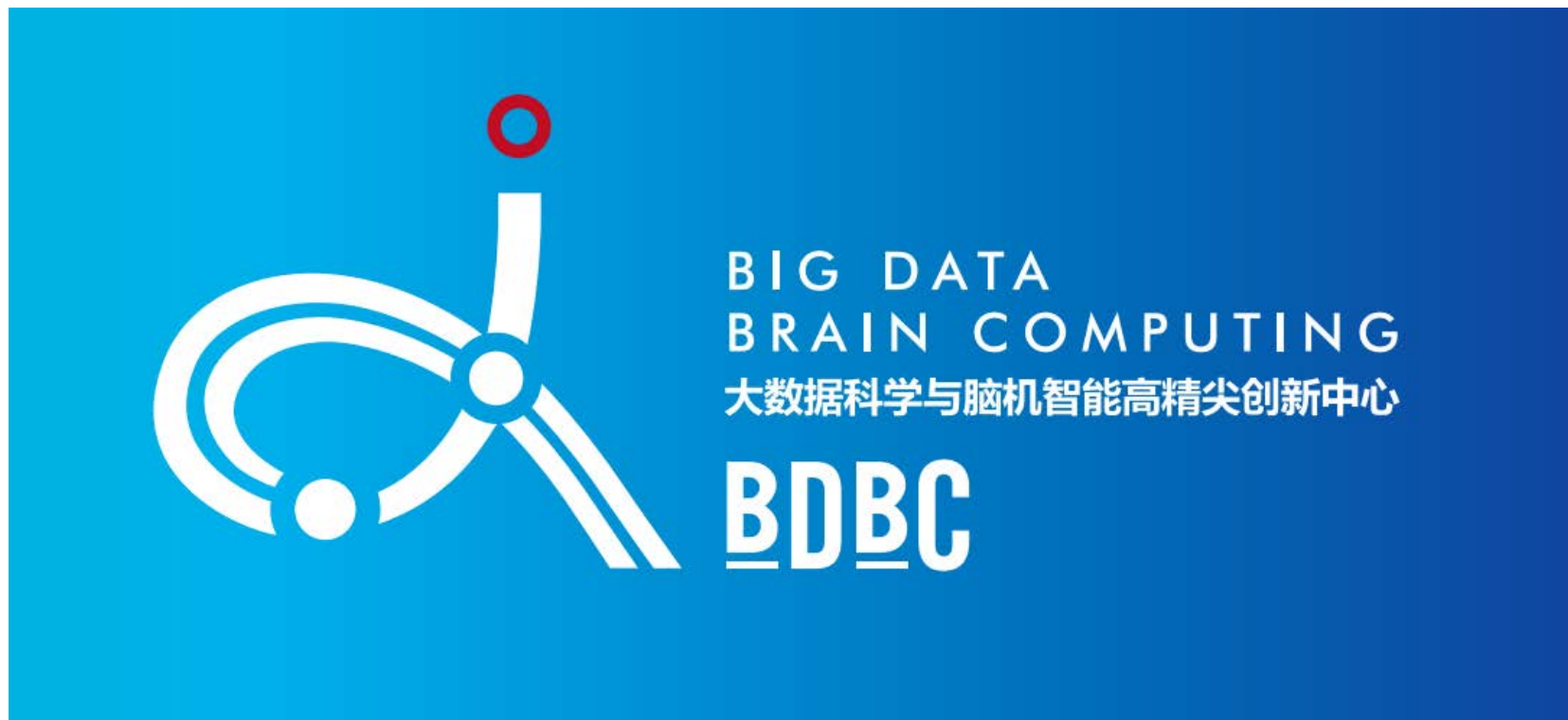


国家重点基础 research 发展计划

- 网络信息空间大数据计算的基础研究(2014-2018)
 - Chief Scientist: Prof. Jinpeng Huai.
 - 8 institutes involved
 - Focus on “computing theory and practice on Big Data”
 - <http://cnbigdata.org/>



北京市大数据科学与脑机智能创新中心



- 2015年，北京市首批北京高校高精尖创新中心
- **引领**未来数据科学与计算智能的研究与应用方向
- **加速**计算科学、数据科学与脑科学的交叉研究
- **促进**高效智能的下一代计算与数据分析技术创新
- 通过以数据为中心的智能机器、系统及应用**改变未来**

研究方向与机构设置



- **瓶颈1：计算的有效性遇到障碍**

- 计算的有效性：
- 认识数据的内在特征，复杂网络、数学（统计）方法



数据科学与计算智能

- **瓶颈2：能耗成为突出问题**

- 随着规模增大，调度复杂，计算系统功耗问题日益突出
- 传统存算分离的结构，产生大量的数据搬移开销
- 传统的计算和存储器件“功耗”不友好



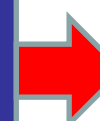
新型计算技术与系统

- **瓶颈3：学习效率和灵活性**

- 学习效率：需要大量的输入数据及标定数据，学习效率低
- 灵活性：普遍缺乏“类比、联想”等学习功能



认知机理与仿真



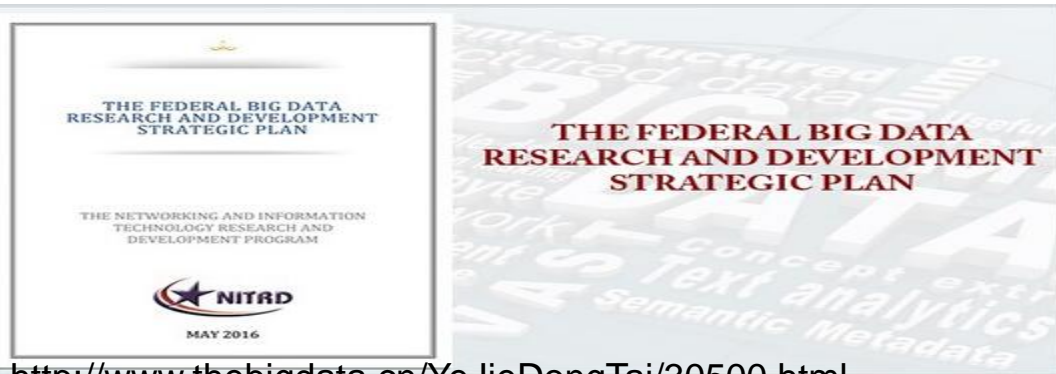
数据工程与脑机系统



大数据的政策与引导：国家大力支持

- 2012年3月29日，美国总统科技政策办公室OSTP（Office of Science and Technology Policy）宣布了每年投资两亿美元的“大数据研究计划”（Big Data R&D Initiative）
- 同天，我国科技部发布的“‘十二五’国家科技计划信息技术领域2013年度备选项目征集指南”把“大数据研究”列在首位
- 美国2016年5月发布《联邦大数据研究与开发战略计划》
- 其目标是对联邦机构的大数据相关项目和投资进行指导,主要围绕代表大数据研发关键领域的七个战略进行，包括促进人类对科学、医学和安全所有分支的认识;确保美国在研发领域继续发挥领导作用；通过研发来提高美国和世界解决紧迫社会和环境问题的能力。

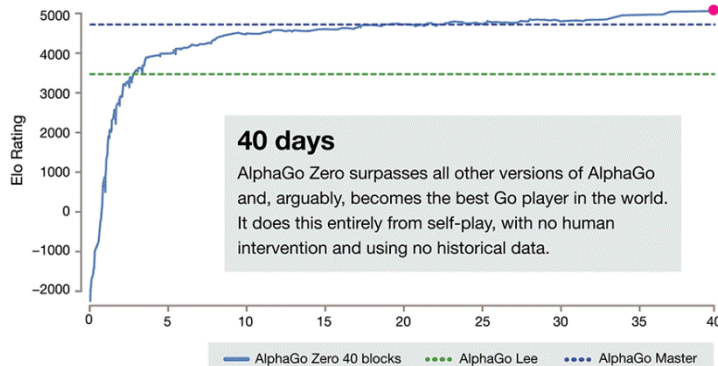
“数据是一项有价值的国家资本，应对公众开放，而不是把其禁锢在政府体制。”
——美国联邦政府



大数据的研究与应用：取得重大突破

- 过去5年大数据的研究，已经产生了重大突破，并在部分领域取得良好的应用

- 计算基础：云计算、深度学习
- 感知处理的角度：深度学习
- 知识组织与管理：知识图谱



- 基于数据产生知识的问答系统

- Watson DeepQA：智能搜索→知识引擎
- Apple Siri & Wolfram Alpha



IBM WATSON 系统介绍

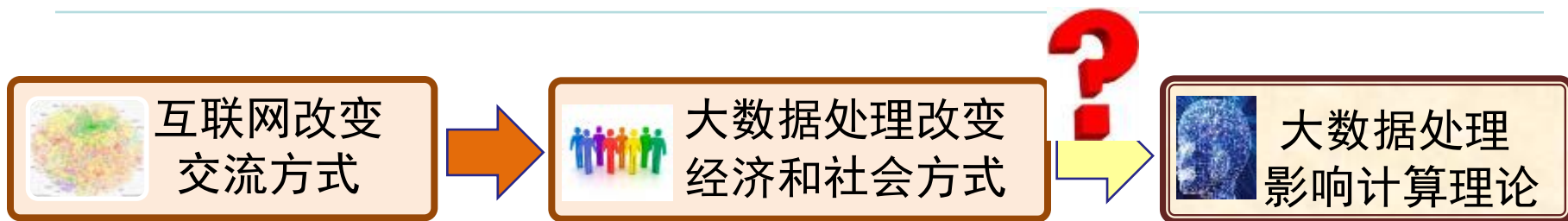
设计目标：设计一台能解答人类语言自然表达的提问，懂得分析大量非结构性数据，拥有自我学习能力，并能实时回应的计算机



IBM Watson 发展过程



大数据的科学价值：发现计算的规律



转变1: 抽样与全样

- 量大、快变，数据统计特征分布不均匀，传统方法不适用 — “尝菜”
- 从无序数据到信息关联

转变2: 精确与非精确

- 精确性不再是绝对追求目标，需对宏观趋势给出快速预测 — 买鞋
- 从信息关联到知识图谱

转变3: 因果与关联

- 仅需知其然，有时无需知其所有然，用于发现事实、预测未来 — 医学
- 形成关联网谱

??

如何理解数据是科学或工程技术？

是否本质上产生新的现象，或只是一种适应过程？

只有掌握规律才能更好理解BD，才能发挥更好价值

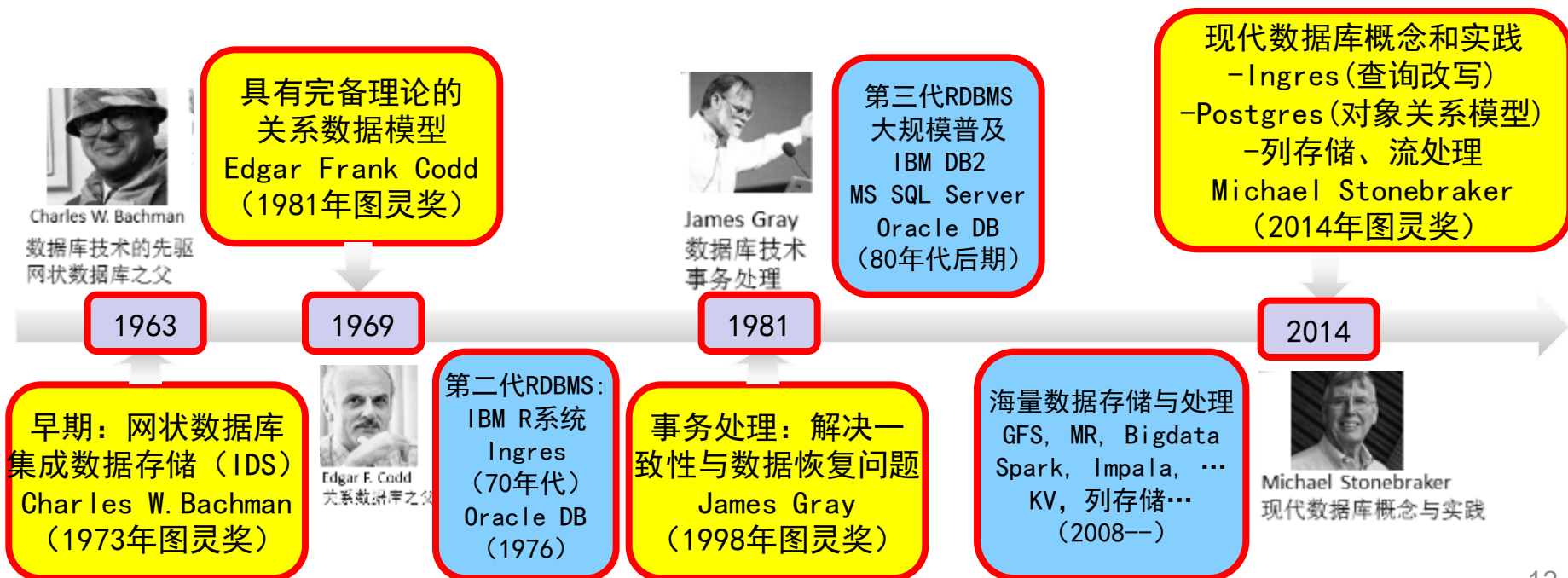


● 问题：是否有坚实的理论基础

● （大）数据科学是否能真的成为一种“科学”？

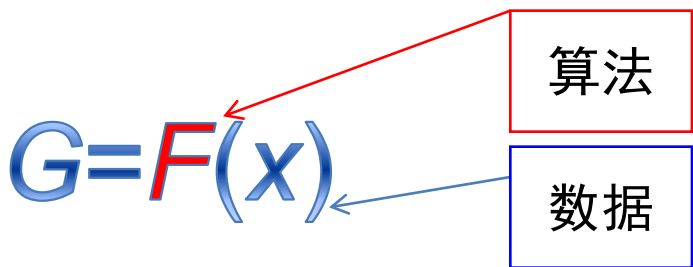
● 回顾数据库领域的发展历程

- 1960s前后，外部存储设备问世，催生数据管理需求，将数据库从文件系统中分离出来
- 1969年，具有完备理论的关系数据模型
- 1981年，事务处理
- 2000年以来，数据量持续增大带来的挑战





- **问题：** 是否有坚实的理论基础
- **（大）数据科学是否能真的成为一种“科学”？**
- 其中一个可能性：计算问题、复杂性与算法
 - 计算问题是计算机科学的本质问题，而算法是一切计算问题的核心



70年代前	• 算法研究
70年代	• 确定性多项式时间算法 • 发现NP困难性
80年代	• 随机化算法 • 随机性能加速算法
90年代	• 近似算法 • 后期发现近似困难性



John E Hopcroft Robert Tarjan (1986) Stephen Cook (1982) Donald Knuth (1974)



Leslie Valiant (2010) Manuel Blum (1995) Juris Hartmanis, Richard Edwin Stearns (1993)



21世纪大数据时代：计算复杂度与算法理论是否有新的理论问题和新方法？



大数据的查询近似

$$R = Q(D)$$

查询近似技术

主要思想：对一类查询复杂性高的查询语言 Q ，变换为一类查询复杂性低的查询语言 Q' ，并且尽量不影响查询结果的准确性。



挑战：平衡查询的复杂性和查询的准确性！

如一，强模拟图查询



子图同构^[11]:

- 给定模式图 Q , 数据图 G 的子图 G_s :
 - Q 图同构 G_s 如果存在一一映射函数 $f: V_Q \rightarrow V_{G_s}$ 满足:
 - ✓ 对 Q 中的任何顶点 u , u 和 $f(u)$ 有相同的标签
 - ✓ 边 (u, u') 在 Q 当且仅当 $(f(u), f(u'))$ 在 G_s
 - Q 子图同构 G , 如果 G 中存在如上子图 G_s

优点 : Q 和 G_s 一模一样

缺点 : NP完全问题 ; 最坏情况下指数个匹配子图 ; 约束过于严格

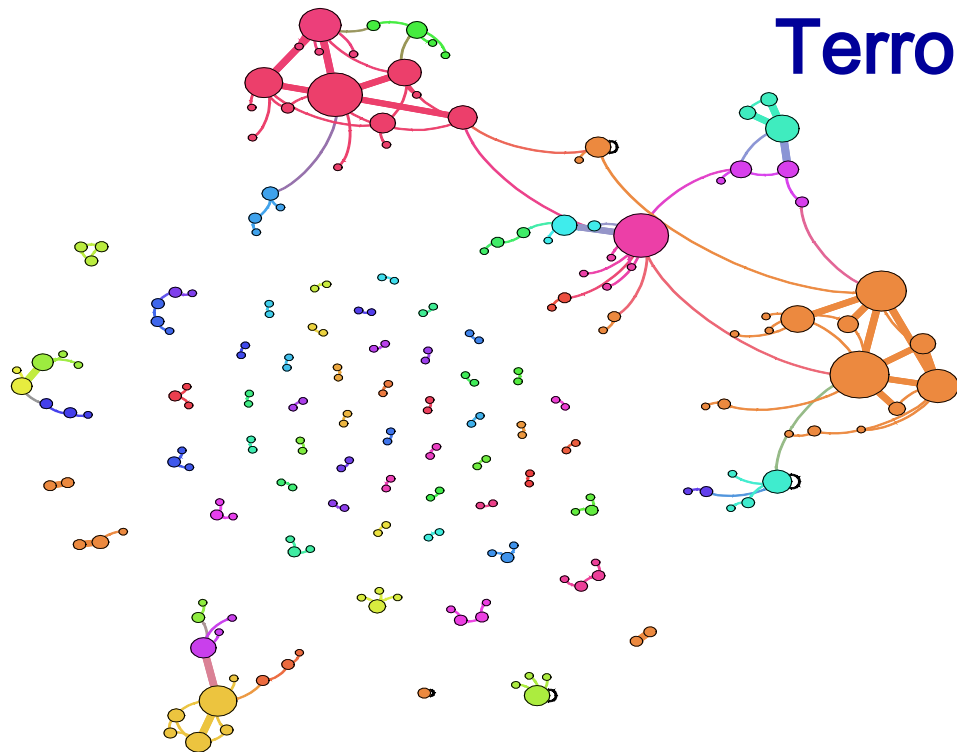
Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo. Strong Simulation: Capturing Topology in Graph Pattern Matching. **TODS 2014**.

Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Capturing Topology in Graph Pattern Matching. **VLDB 2012**.

如一，强模拟图查询



Terrorist Collaboration Network



“Those who were trained to fly didn’t know the others.
One group of people did not know the other group.”
(Osama Bin Laden, 2001)

Build upon (revised) strong simulation to aid the detection of homegrown violent extremists (HVEs) who seek to commit acts of terrorism in the United States and abroad,
Colorado State University, Benjamin W. K. Hung, Anura P. Jayasumana: Investigative simulation: Towards utilizing graph pattern matching for investigative search. ASONAM 2016.

如一，强模拟图查询



Matching	children	parents	connectivity	cycles
\prec	✓	×	×	✓ (directed), × (undirected)
\prec_D	✓	✓	✓	✓ (directed & undirected)
\prec_D^L	✓	✓	✓	✓ (directed & undirected)
\triangleleft	✓	✓	✓	✓ (directed & undirected)

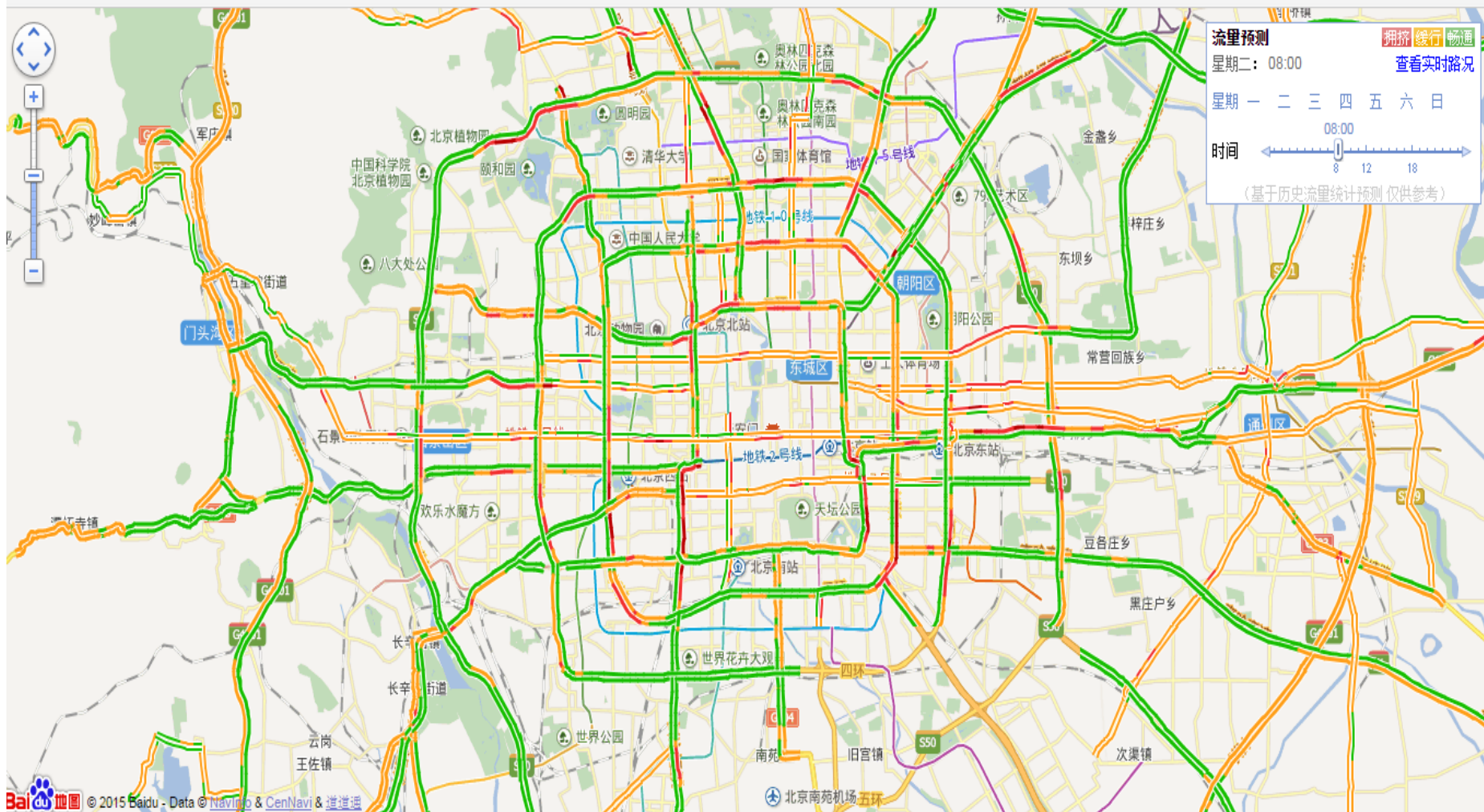
locality	matches	Bisimilar&b'ed-cycle
×	✓	×
×	×	×
✓	✓	×
✓	×	✓

查询结果保持70-80%子图同构结构，效率提高100倍！

如二，时态稠密图查询

Baidu 地图 实时路况

北京市 [选择城市]





如二，时态稠密图查询

- 筛选与验证的方法 (Filter-and-Verification)

10^4	10^5	10^6	...	10^8
5×10^2	5×10^3	5×10^4		5×10^6

过滤掉95%

- 数据驱动的查询近似方法(Data Driven Query Approximation)
 - 根据数据的特点选取k个(k是个小的常数，比如10或15)

10^4	10^5	10^6	...	10^8
k	k	k		k

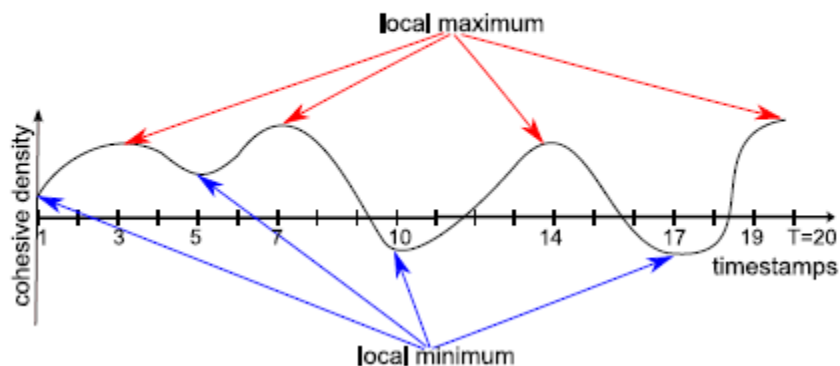
- 实验结果 (With the state of the art solution^[Bogdanov et al. 2011])

	准确性	效率
BEIJING DATA	100.25%	快4,870倍
SYNTHETIC DATA	99.69%	快1,468倍

如二，时态稠密图查询

在演化生物学中，**趋同演化**（Convergent evolution）指的是两种不具亲缘关系的动物/植物长期生活在相同或相似的环境，或曰生态系统，它们因应需要而发展出相同功能的器官的现象，即同功器官

Evolving convergence assumption



The p_{EC} are 96% on BEIJING DATA and 90% on average on all tested SYNTHETIC DATA, respectively, which justifies our observation of the evolving convergence assumption.



Proposition 2: To find the dense subgraph, we only need to consider the time intervals $[i, j]$ such that the cohesive density curve has a local maximum at certain point between i and j under the evolving convergence assumption. \square

Fact 2: Temporal subgraph $\mathbb{G}[i, j]$ ($i \leq j \in [1, T]$) with a higher positive cohesive density has a higher probability of containing a dense subgraph under the assumption of independent and identically distributed edge weights. \square



大数据的数据近似

$$R = Q(D)$$

数据近似技术

主要思想： 对一类查询复杂性高的查询语言Q，将查询数据D变换机器能够高效处理的较小量D'，并且尽量不影响查询结果的准确性。



二八定律： 在众多现象中，80%的结果取决于20%的原因

$$D = \text{HARD}(D) + \text{SOFT}(D)$$



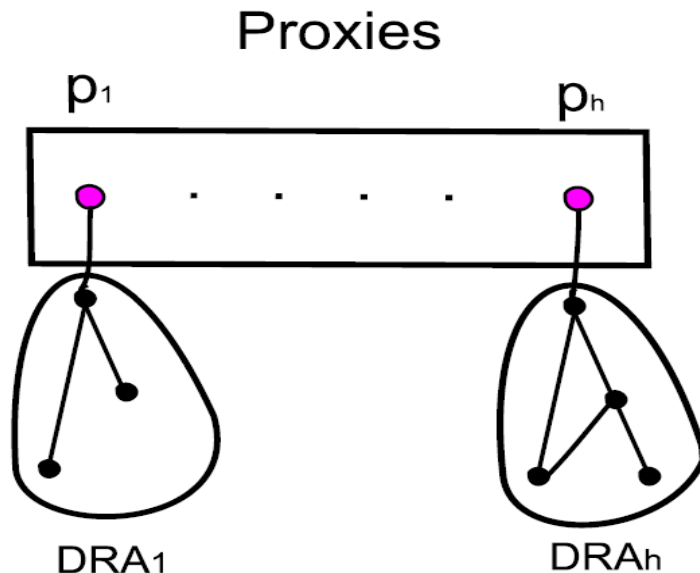
挑战： 平衡查询的效率和查询的准确性!



如一、最短路径/距离



- 针对有权无向图，提出“proxies”概念
- 每个proxy代表了一个子图DRA中的顶点(互补重叠)
- Proxies 能够在 $O(n)$ 时间内算出



关键性质: 给定图 G 、顶点 u 和 v 及其代理 u_p 和 v_p ，则：

$$(1) \text{path}(u, v) = \text{path}(u, u_p) + \text{path}(u_p, v_p) + \text{path}(v_p, v)$$

$$(2) \text{dist}(u, v) = \text{dist}(u, u_p) + \text{dist}(u_p, v_p) + \text{dist}(v_p, v)$$

在真实的公路网络和社会网络中，数据减少了1/3!
是一种针对最短路径/距离的轻量级的通用的数据缩减技术!

Shuai Ma, Kaiyu Feng, Jianxin Li, Haixun Wang, Gao Cong, and Jinpeng Huai, Proxies for Shortest Path and Distance Queries. **TKDE 2016**.

Shuai Ma, Kaiyu Feng, Jianxin Li, Haixun Wang, Gao Cong, and Jinpeng Huai, Proxies for Shortest Path and Distance Queries. **ICDE 2017 (TKDE Extended Abstract)**.



如二，网络链接预测

链接预测：

- n 个顶点网络， $O(n^2)$ 个可能链接
- CPU速度XGHz/s，假定1个机器时钟处理1个顶点对。

Network Sizes	1 GHz	3 GHz	10 GHz
10^6 nodes	1000 sec.	333 sec.	100 sec.
10^7 nodes	27.8 hrs	9.3 hrs	2.78 hrs
10^8 nodes	> 100 days	> 35 days	> 10 days
10^9 nodes	> 10000 days	> 3500 days	> 1000 days

多数链接预测算法仅仅预测一个可能链接子集，而不是整个网络所有可能的链接，如[Dashun et al. 2011, Chungmok et al. 2014].

Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, Albert-László Barabási: Human mobility, social ties, and link prediction. KDD 2011.

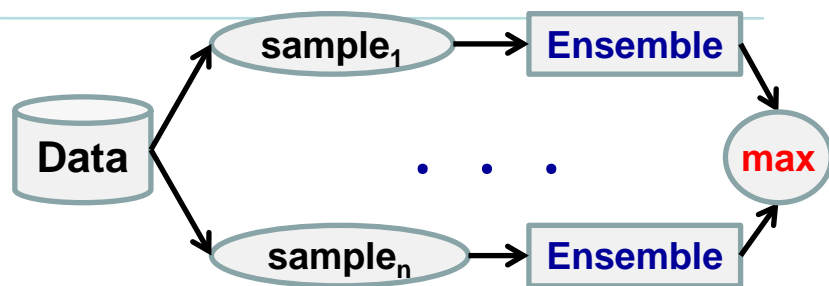
Chungmok Lee, Minh Pham, Norman Kim, Myong K. Jeong, Dennis K. J. Lin, Wanpracha Art Chaovalitwongse. A novel link prediction approach for scale-free networks. WWW 2014.

如二，网络链接预测



- 直接采用非负矩阵分解的代价高

- 效率低
- 数据越稀疏，效果越差



- 数据近似技术 (Ensemble Enabled Sampling)

- 采样要保证一定的覆盖率
- 基于链接预测特征的抽样 - Triangles
- 结合Ensemble的思想：链接e的预测分值是所有Ensemble中的最大值

PROPOSITION 2. The expected times of each node pair included in μ/f^2 ensemble components is at least μ .

小数据	准确性	大数据	效率
YouTube	高18%	Friendster	快31倍
Wikipedia	高16%	Twitter	快21倍

同时提高了准确性和检测效率！

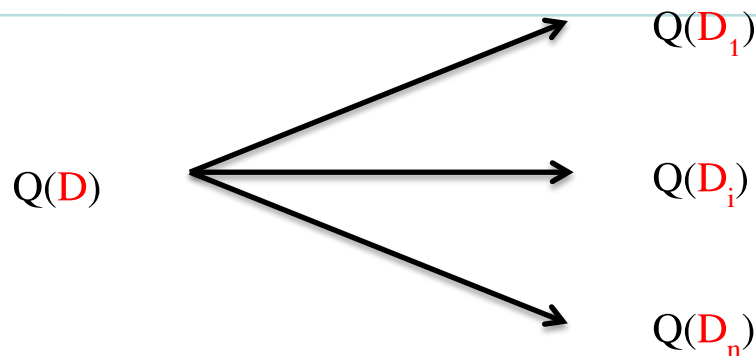
Liang Duan, Charu Aggarwal, Shuai Ma, Renjun Hu, and Jinpeng Huai, Scaling up Link Prediction with Ensembles, **WSDM 2016 - Big Data Algorithms Session**.

Liang Duan, Shuai Ma*, Charu Aggarwal, Tiejun Ma, and Jinpeng Huai, An Ensemble Approach to Link Prediction. **TKDE, 29(11): 2402-2416, 2017**.

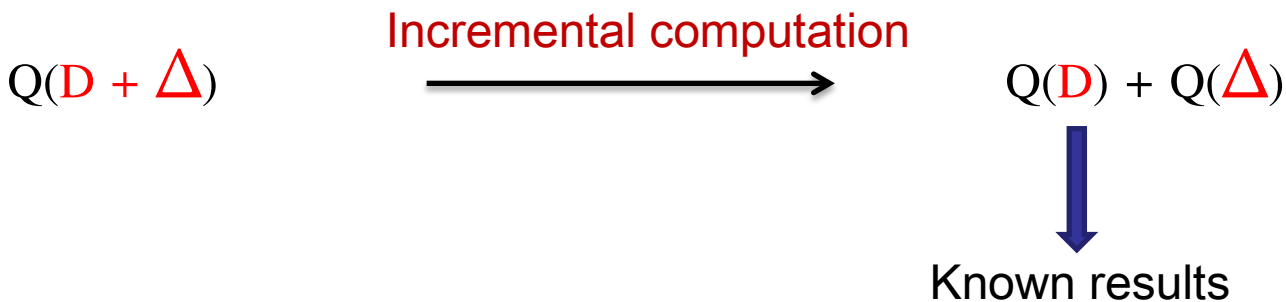


其它相关技术

- 分布式计算:



- 增量计算:



- 数据压缩:



- 数据索引: 空间代价、构建时间代价、查询效率提高



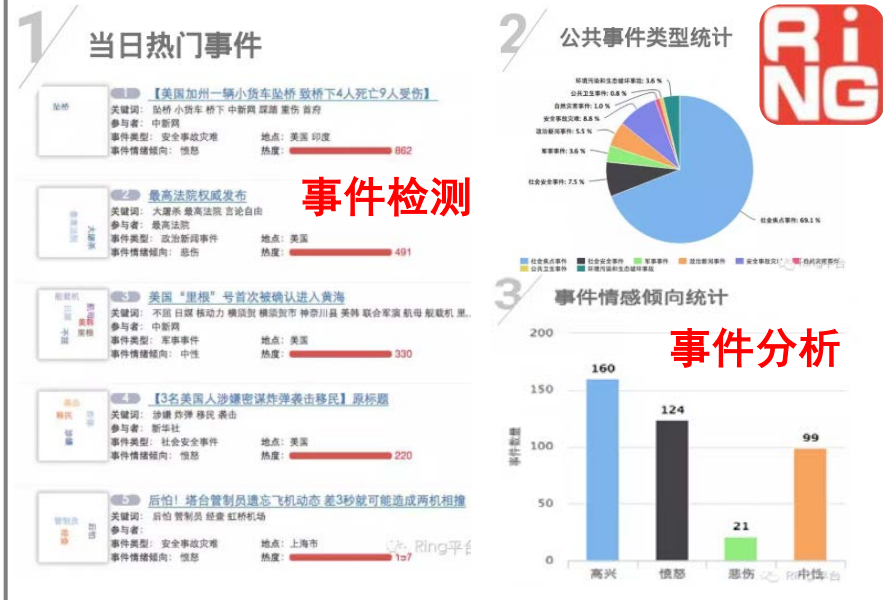
大数据技术的应用案例

BDA

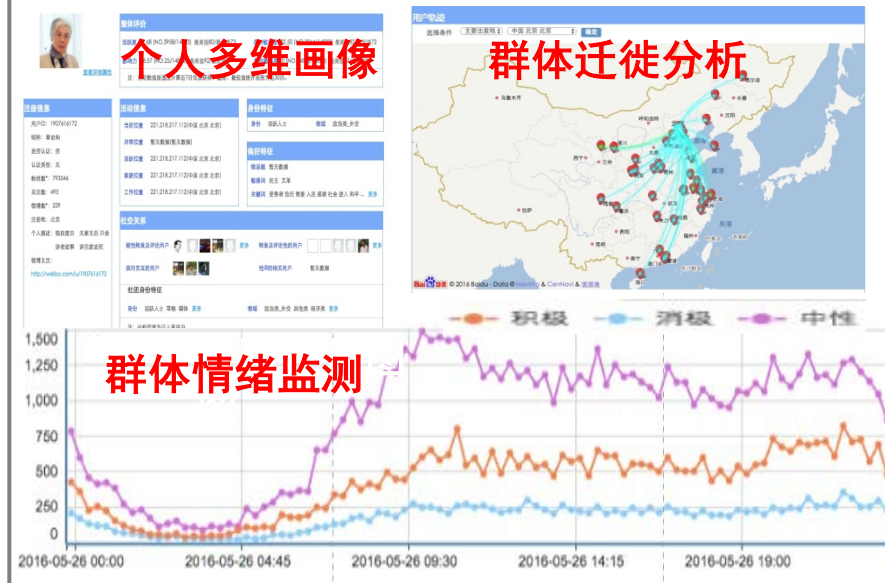
如一：面向公共事件的示范应用



安管中心：实时事件检测与分析系统



安管中心：微博人物画像系统

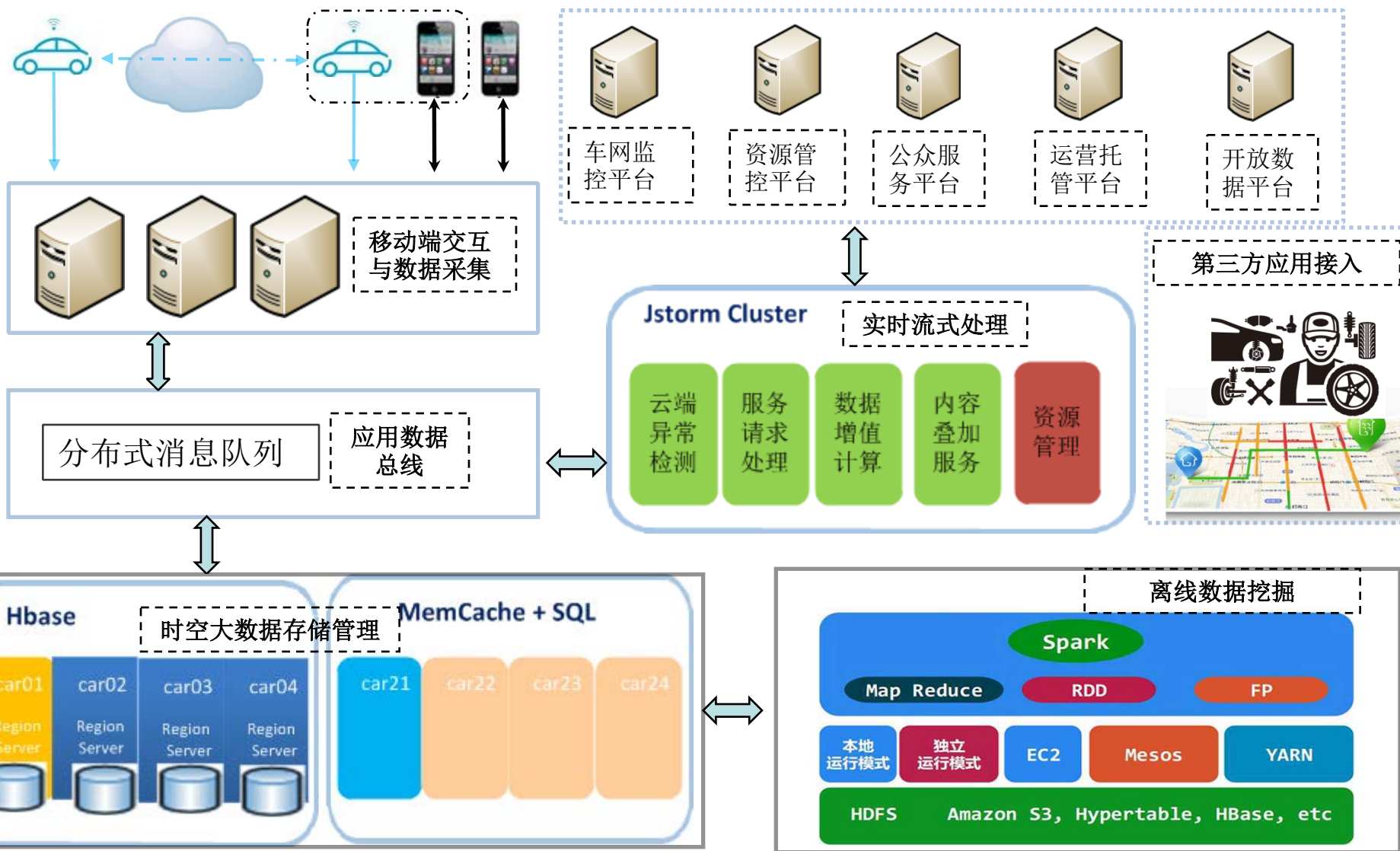


舆情大数据服务：已被国家信息安全管理中心采用，建立Ring微信公众号

研究工作：异质大数据获取、建模与深度分析，异构大数据存储与多模态计算，复杂系统敏捷定制，事件分析与预测。

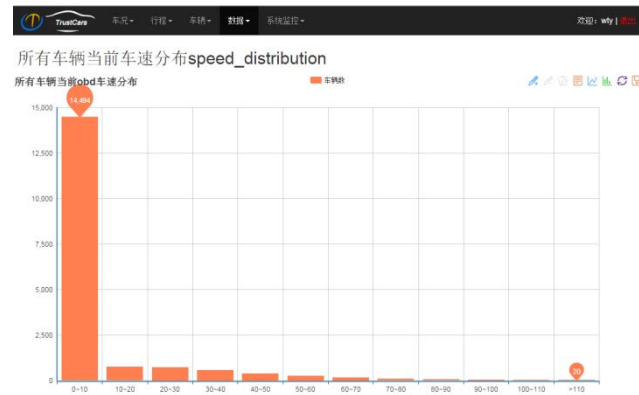
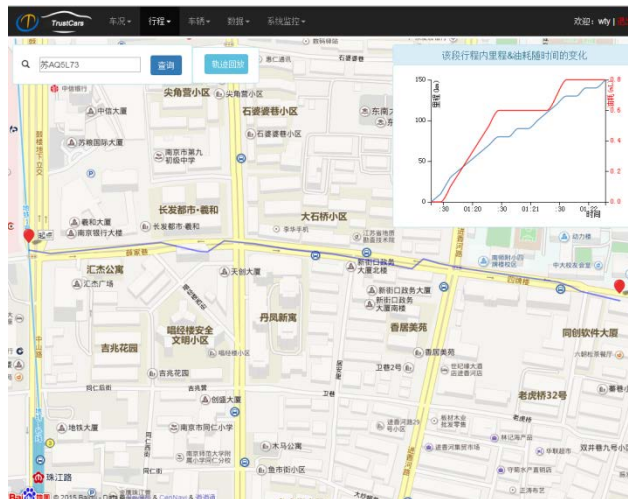
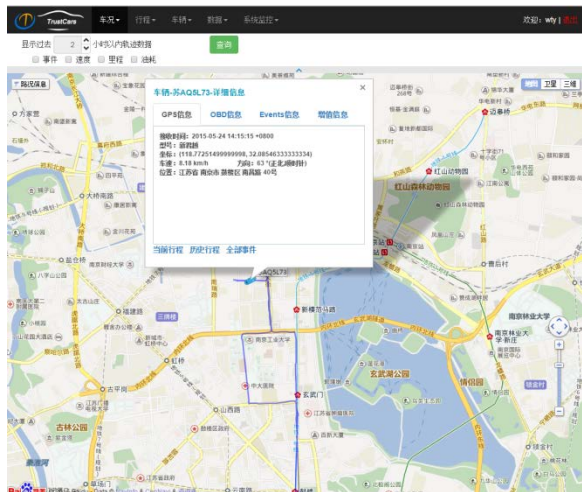
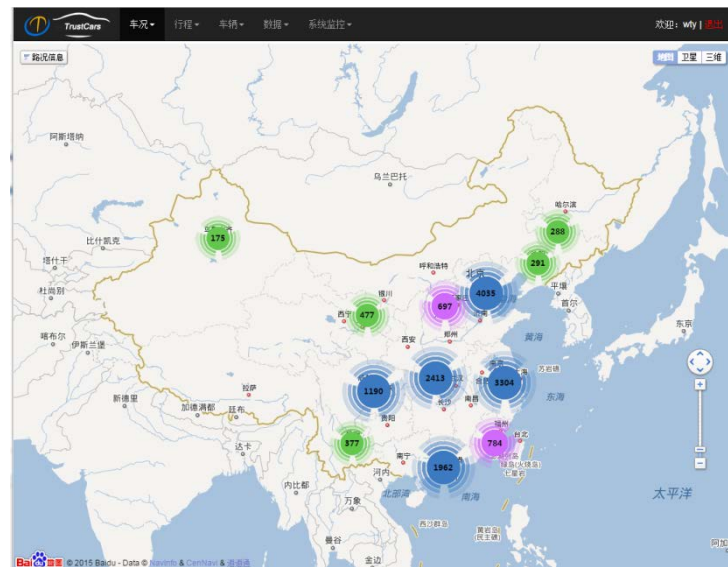
核心算法：Weiren Yu, Charu C. Aggarwal, Shuai Ma, Haixun Wang: On Anomalous Hotspot Discovery in Graph Streams. **ICDM 2013.**

如二：实时流式大数据处理平台

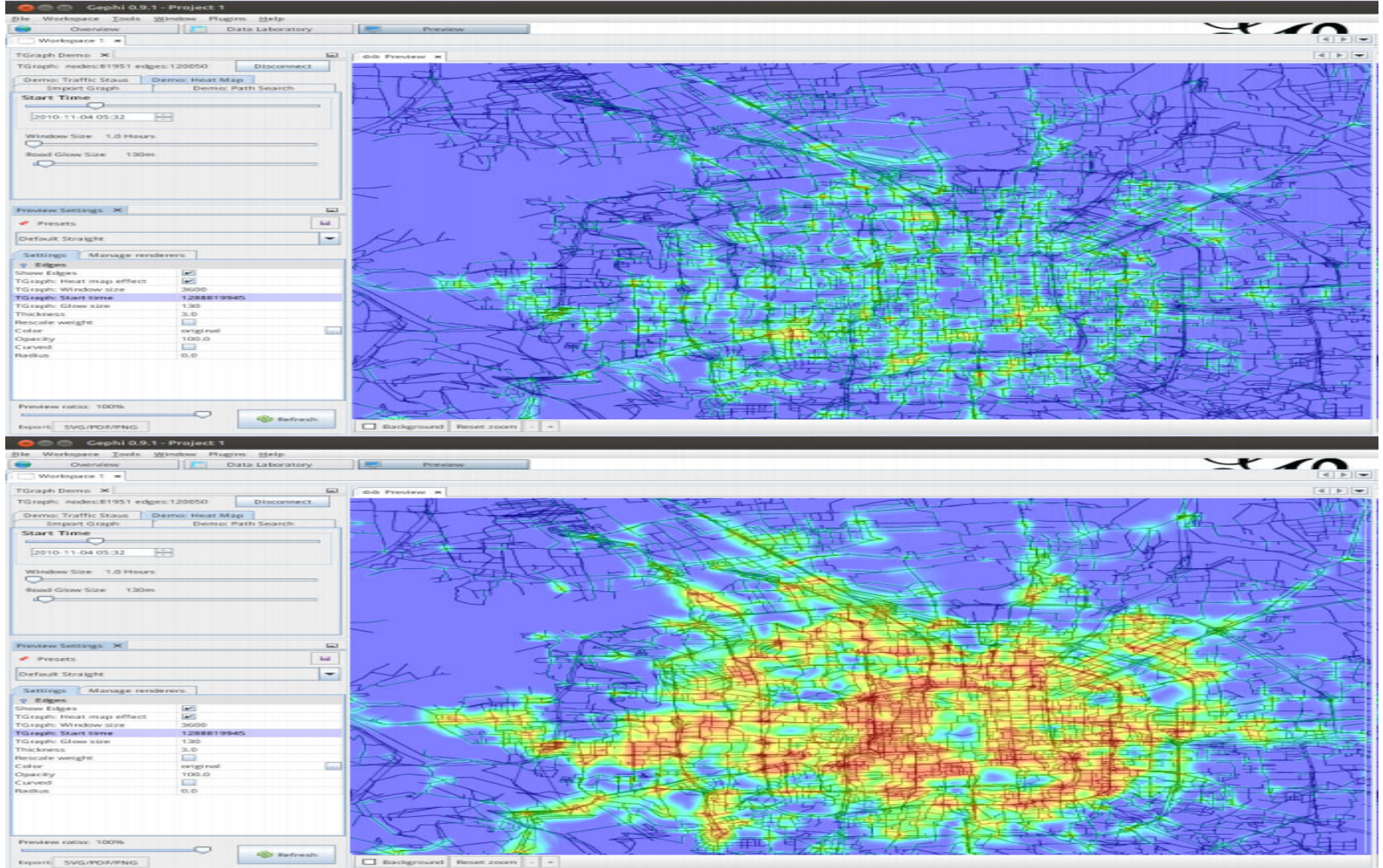


如二：实时流式大数据处理平台

- 车辆运营态势
- 单个车辆状态实时跟踪
- 历史运营分析
 - 线路、里程、油耗、行为
- 行程实时切分与速度分析
- 数据质量分析与系统状态监控



如三：TGraph时态图数据管理系统



如四：人类行为预测

- Prediction is now a developing science
- Success seems to be achieved most consistently when questions are tackled in multidisciplinary efforts that join human understanding of context with algorithmic capacity to handle terabytes of data - Special Issue: Prediction, Science, Feb. 3, 2017.



2016年美国大选预测失败原因？

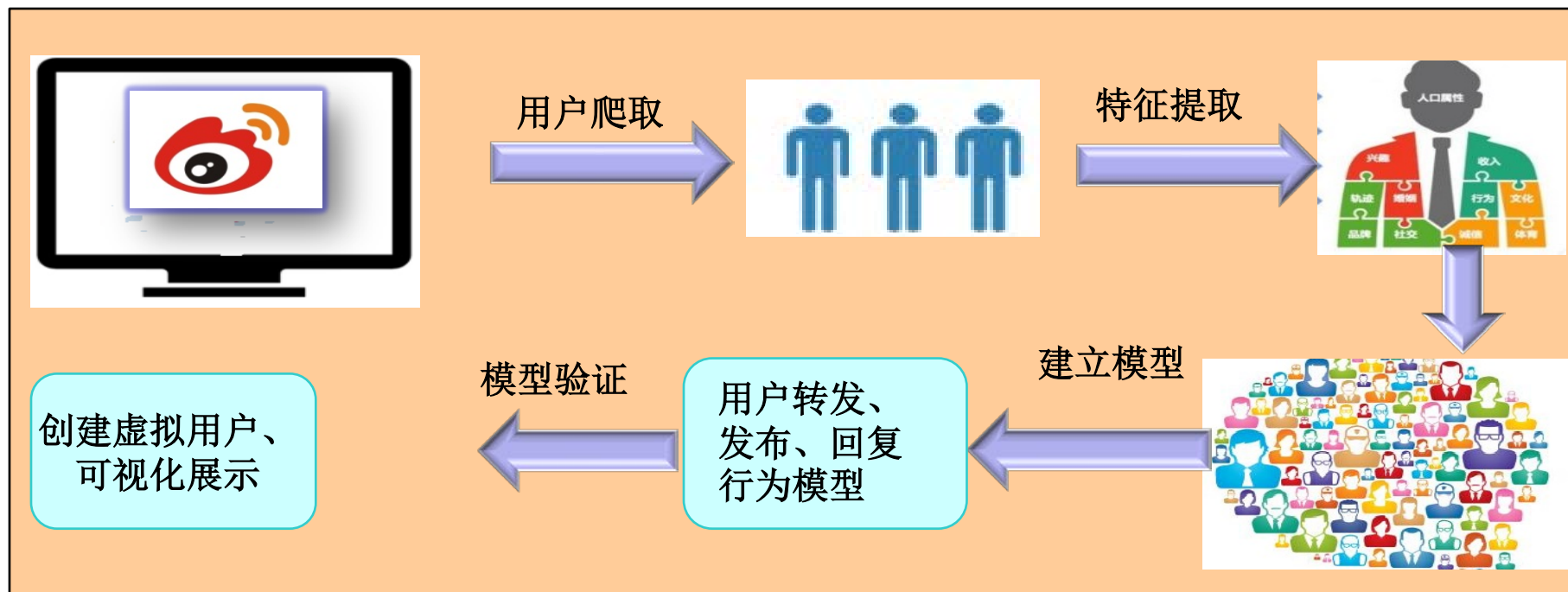
- Twitter机器人、雇佣军
- Shy Trump voter理论：支持者趋向于保持沉默

人类行为预测是下一个前沿：

- 更多的数据 \neq 更好的预测：噪音的存在
- 罕见事件、行为动态变化
- IoT的涌现导致多种异构数据爆炸



如四：基于微博数据的群体行为预测



目标： 本文以爬取的微博数据为基础，对微博用户进行特征提取，利用提取的用户特征进行聚类，得到统计特征差别明显的不同用户群体，针对不同群体，分析用户行为的影响因素，探究微博用户行为机理。

主要衡量指标： 行为建模准确度

Acknowledgements

Collaborators:

Charu Aggarwal, Sourav S Bhowmick, Yang Cao, Gao Cong, Liang Duan, Wenfei Fan, Kaiyu Feng, Haixing Huang, Renjun Hu, Jinpeng Huai, Jia Li, Jianxin Li, Xuelian Lin, Xudong Liu, Zhe Liu, Jinghe Song, Haixun Wang, Luoshu Wang, Tianyu Wo...

They are from:



THE UNIVERSITY of EDINBURGH



NANYANG
TECHNOLOGICAL
UNIVERSITY



THE OHIO STATE UNIVERSITY



Microsoft

Research
微软亚洲研究院



facebook



主页: <http://mashuai.buaa.edu.cn>

邮件: mashuai@buaa.edu.cn

地址:

新主楼G1122,
北京航空航天大学



Thanks!