



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



海量®
HYLANDA



数据科学实践： 公共安全事件研究

OmniEye, 上海交通大学

团队：陈夏明 (队长), 强思维, 王海洋, 孙莹, 石开元

指导老师：上海交通大学网络信息中心 金耀辉 教授

jinyh@sjtu.edu.cn

大数据竞赛概况

大数据到数据科学的演变

数据科学实践：公共安全事件研究

数据的社会价值

大数据竞赛概况

大数据到数据科学的演变

数据科学实践：公共安全事件研究

数据的社会价值

参赛动机

- OMNILab : 开放移动网络创新实验室
- 来自OMNILab的五位**爱好数据**的骚年



陈夏明
博士生



强思维
博士生



孙莹
硕士生



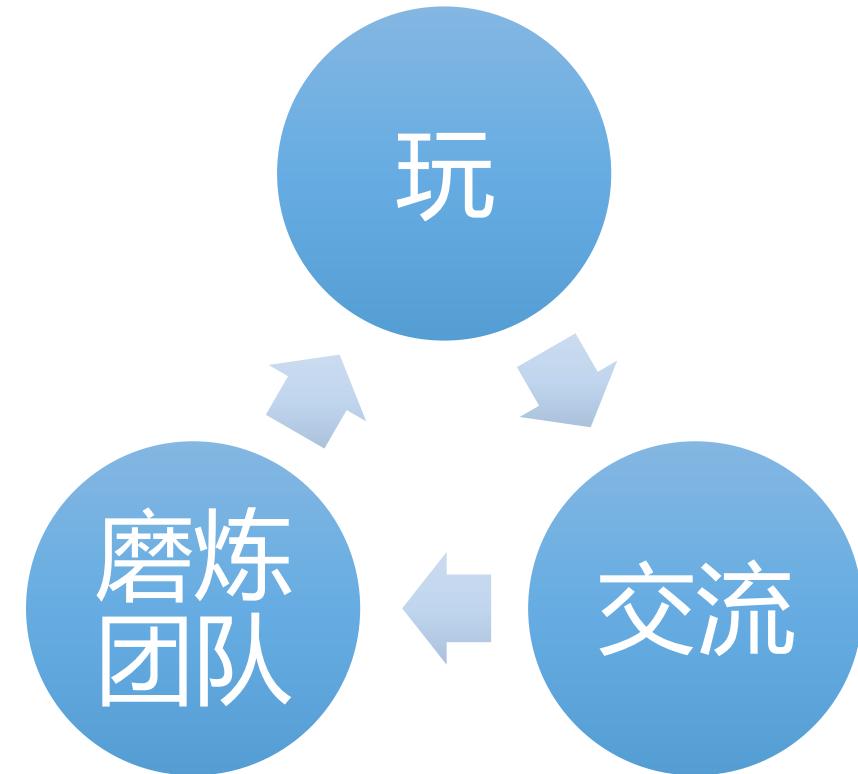
石开元
硕士生



王海洋
博士生

参赛动机

- OMNILab : 开放移动网络创新实验室
- 来自OMNILab的五位**爱好数据**的骚年



赛题选择

- 选择开放、创新、有挑战的题目来做
- 数据集保留时空维度信息
- 实践**大数据**背景下的数据思维方式



系列危害公共安全事件
关联关系挖掘和预测

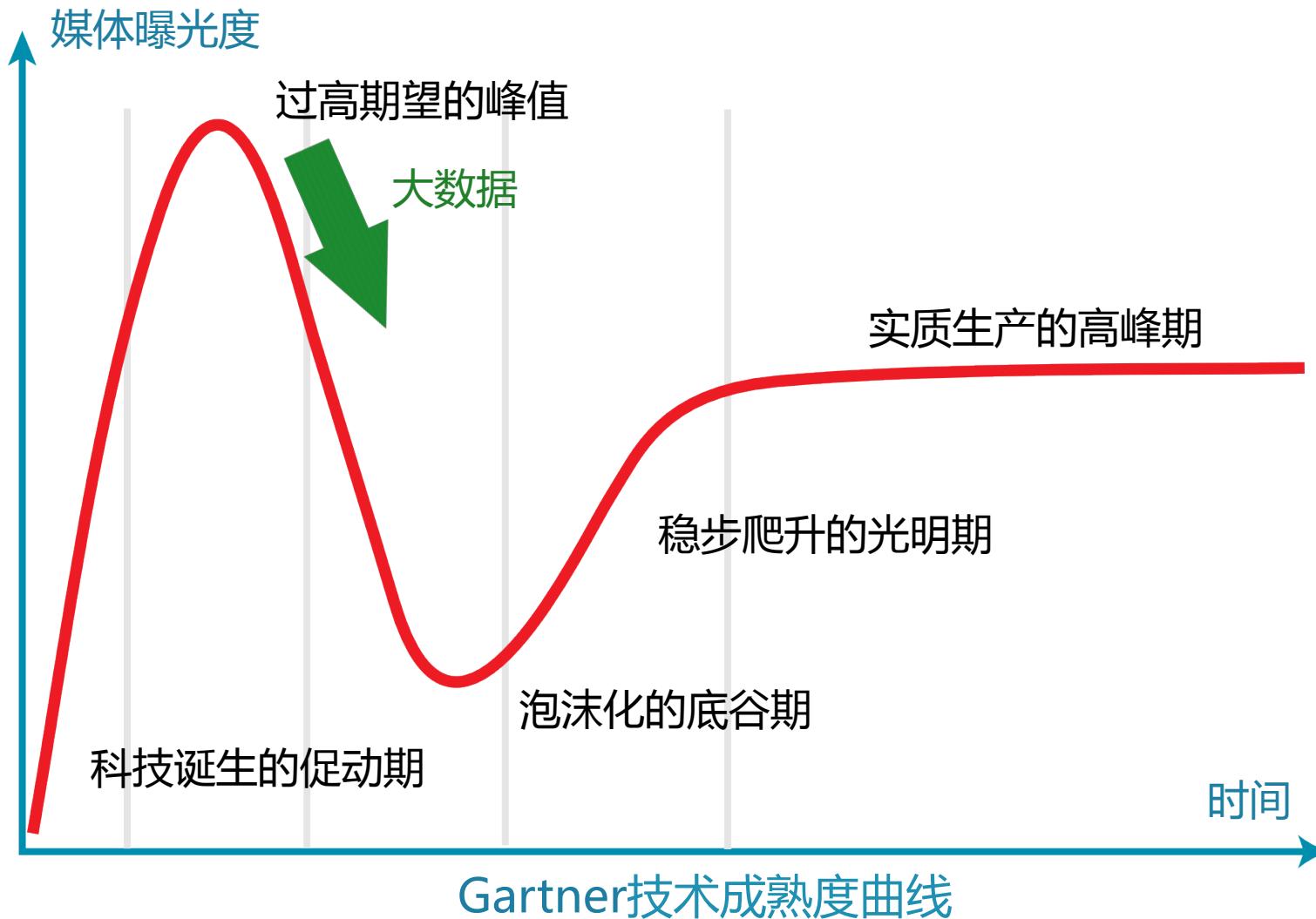
大数据竞赛概况

大数据到数据科学的演变

数据科学实践：公共安全事件研究

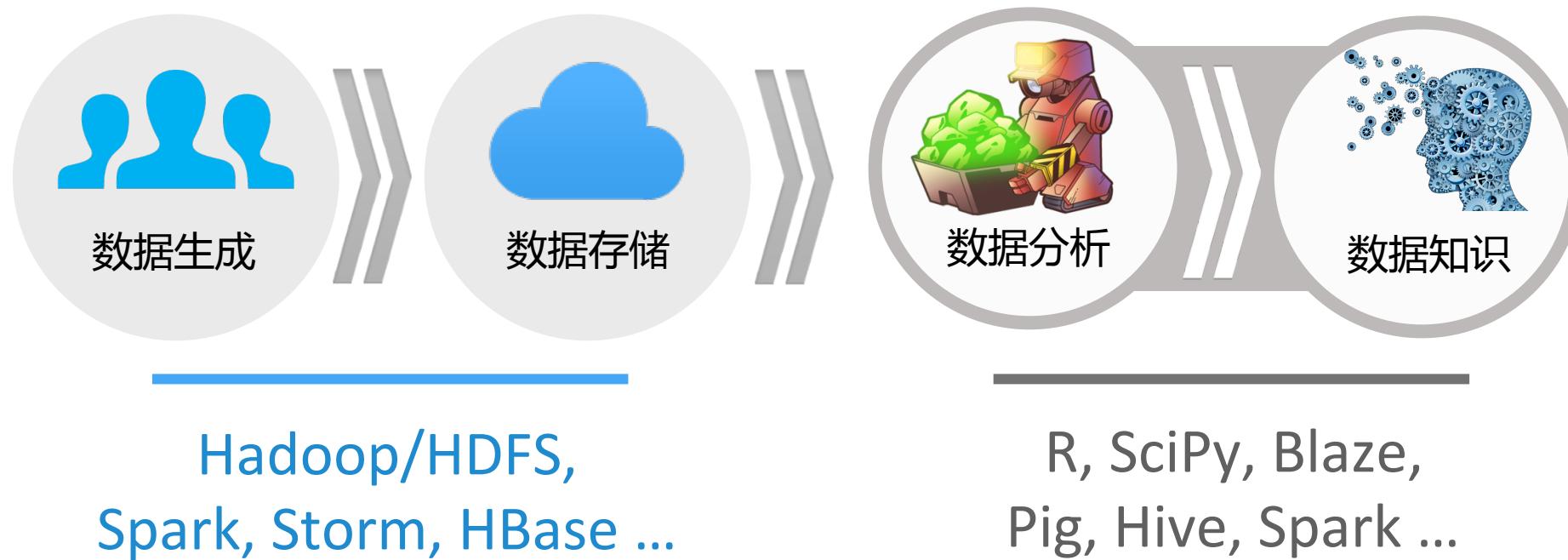
数据的社会价值

大数据接地气



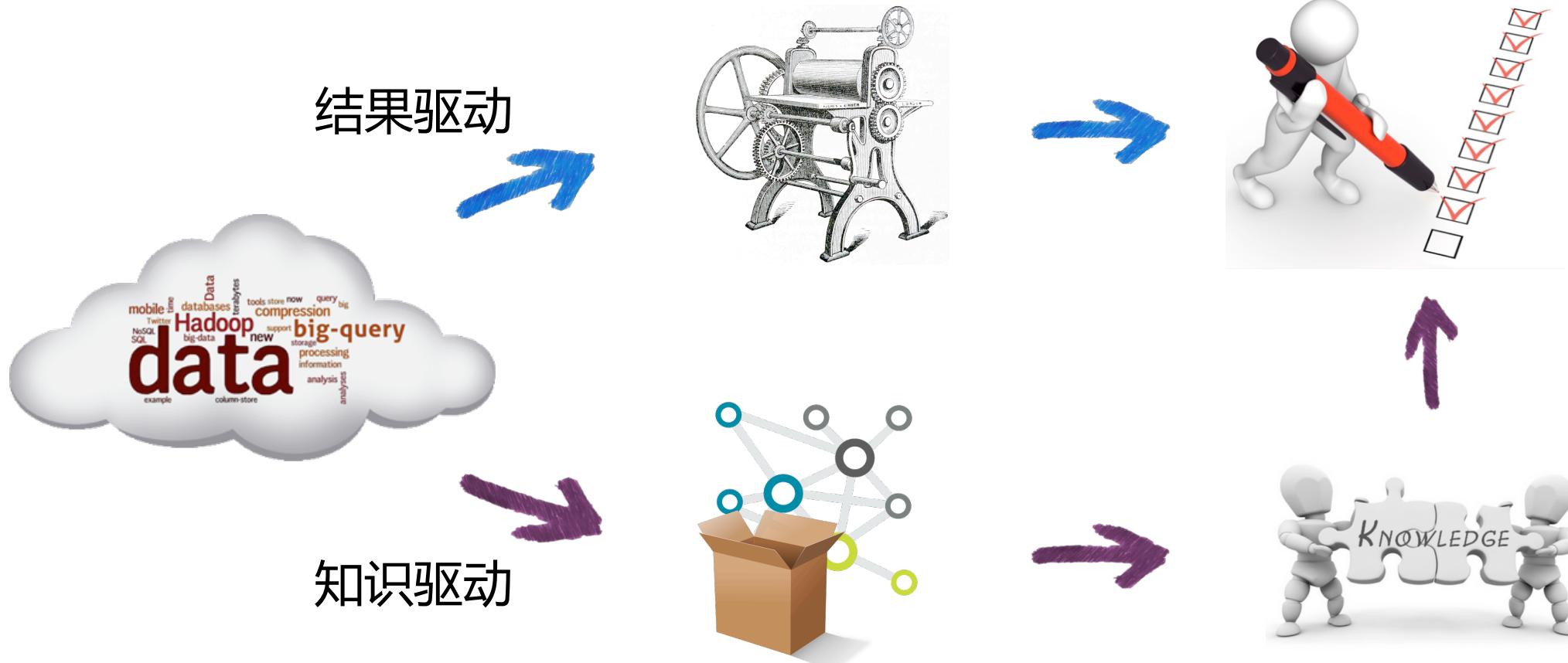
从大数据到数据科学

- 大数据解决数据的**管理问题**
- 数据科学解决数据的**价值问题**



从大数据到数据科学

- 数据科学注重解决问题的**过程**，而不仅仅是结果



大数据竞赛概况

大数据到数据科学的演变

数据科学实践：公共安全事件研究

数据的社会价值

公共安全事件

- 2014年7月17日至24日,北京于7天内发生6起危害公共安全事件。
- 媒体**大规模报道、网民舆论**——负面信息传播泛滥的温床。
- 了解危害公共安全事件在互联网上的**触发、传播机理**，找到相关事件间的影响关系和共性，是意义重大的研究课题。



竞赛目标

基本任务

1. 数据清洗，剔除杂质
2. 自定义标签，事件提取

核心任务（可选）

3. 同系列事件间相互触发关系研究
4. 不同系列事件间共性分析
5. 事件预测

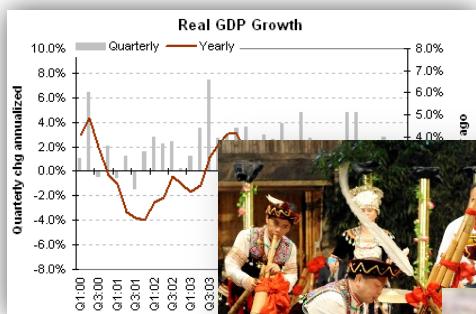
知识驱动的数据分析过程



- 数据获取及预处理
(任务1)

数据集及预处理

- 数据清洗：去重和纠错
- 数据扩充 (Data Enrichment)



GDP



民族节日



城市坐标

新闻/微博数据 (54万条)



媒体/用户数据 (24万条)

中国各省市行政区划一览表
表中记录了中国各省市行政区划

中国各民族传统节日总表
表中记录中国所有传统节日

新闻数据集
新闻数据集人工标注，从比赛提供的所有新闻中随机抽取3千条进行人工标注 -numkey...

微博数据集
微博数据集人工标注，从比赛提供的所有微博中随机抽取3千条进行人工标注 -numkey ...

知识驱动的数据分析过程



- 数据获取及预处理
(任务1)
- 新闻分类/事件聚类
(任务2)

新闻分类

公交车
爆炸事件

暴恐
事件

校园砍
杀事件

新闻分类

- 挑战一：不同媒介(体)的报道方式不同

公交车
爆炸事件

暴恐
事件

校园砍
杀事件

媒体名称	发布时间	新闻标题
新华网	2013-12-14	河南光山县发生校园伤害案 22名学生被砍伤
新浪微博	2013-12-15	目前，22名被砍伤的学生中，有7名学生因伤势严重转院治疗。此外，还有1名群众及1名小学生因伤势严重，仍在光山县人民医院的重症监护室进行治疗。愿平安！

新闻分类

- 挑战一：不同媒介(体)的报道方式不同
- 挑战二：新闻媒体报道角度不同

公交车
爆炸事件

暴恐
事件

校园砍
杀事件

媒体名称	发布时间	新闻标题
网易新闻	2013-07-28	新疆莎车县发生暴恐案件，造成37人死亡，13人受伤
人民网	2013-07-28	新疆莎车：暴恐分子袭击军警击毙暴徒59人
四川在线	2013-07-28	新疆莎车暴恐袭击案件告破

新闻分类

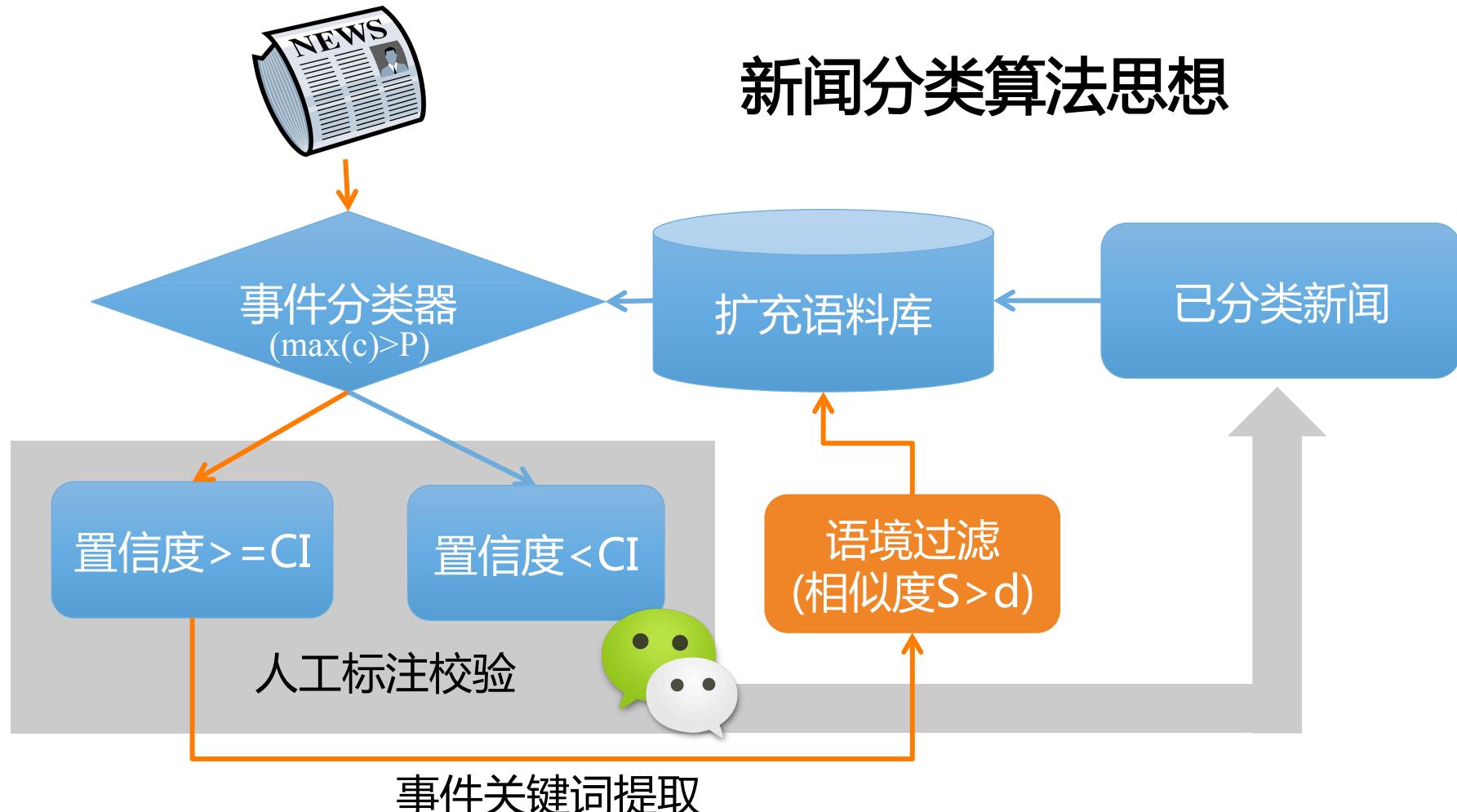
- 挑战一：不同媒介(体)的报道方式不同
- 挑战二：新闻媒体报道角度不同
- 挑战三：新闻事件和衍生事件的关联



媒体名称	发布时间	新闻标题
第一金融网	2013-04-26	新疆巴楚15名警察社区工作人员遭暴徒袭杀
人民网	2013-04-26	习近平批示新疆巴楚县暴力事件 对案件善后作指示
人民网	2013-04-27	习近平:要使暴力恐怖分子成为"过街老鼠 人人喊打"

语境过滤的新闻分类算法

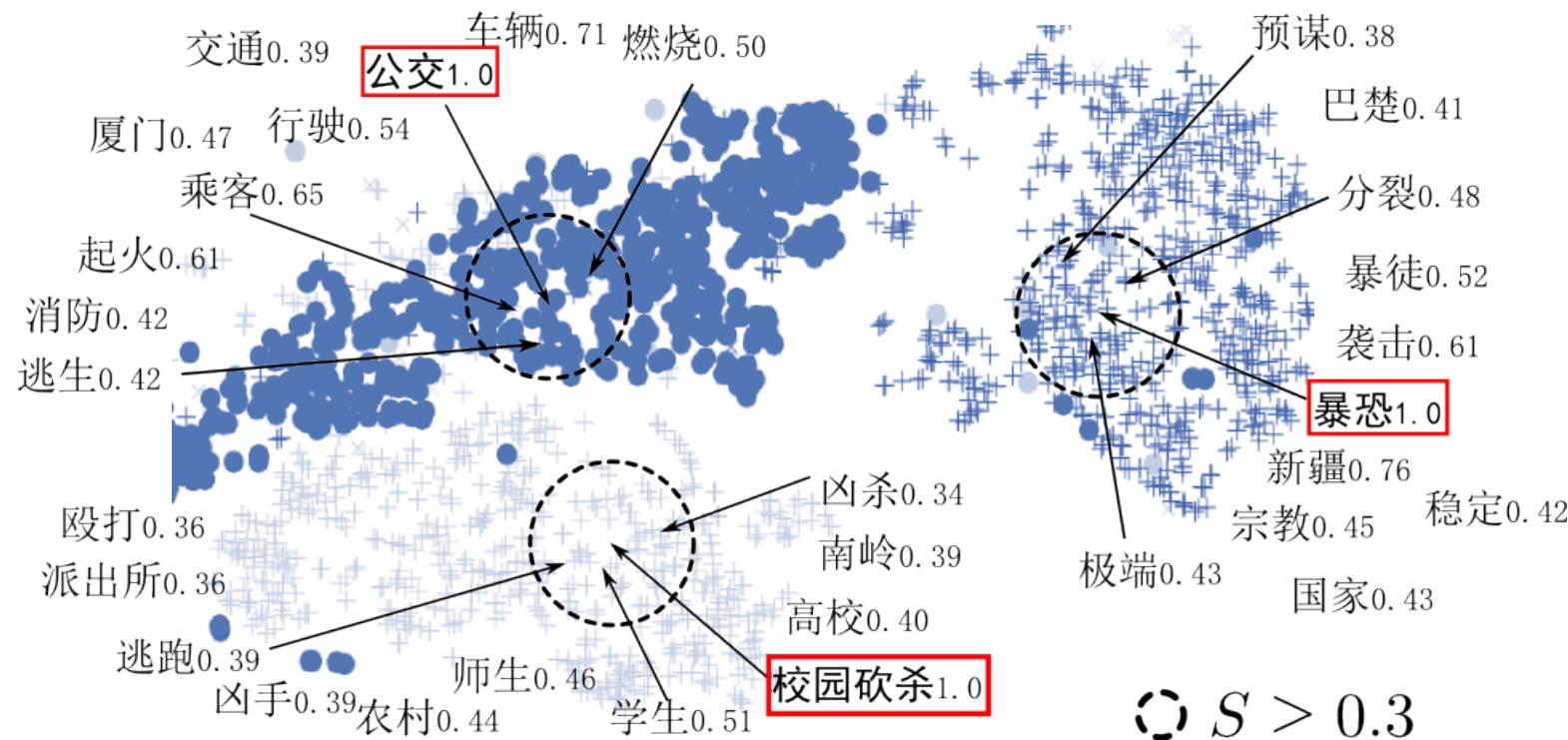
News Classification with Context Filtering



语境过滤

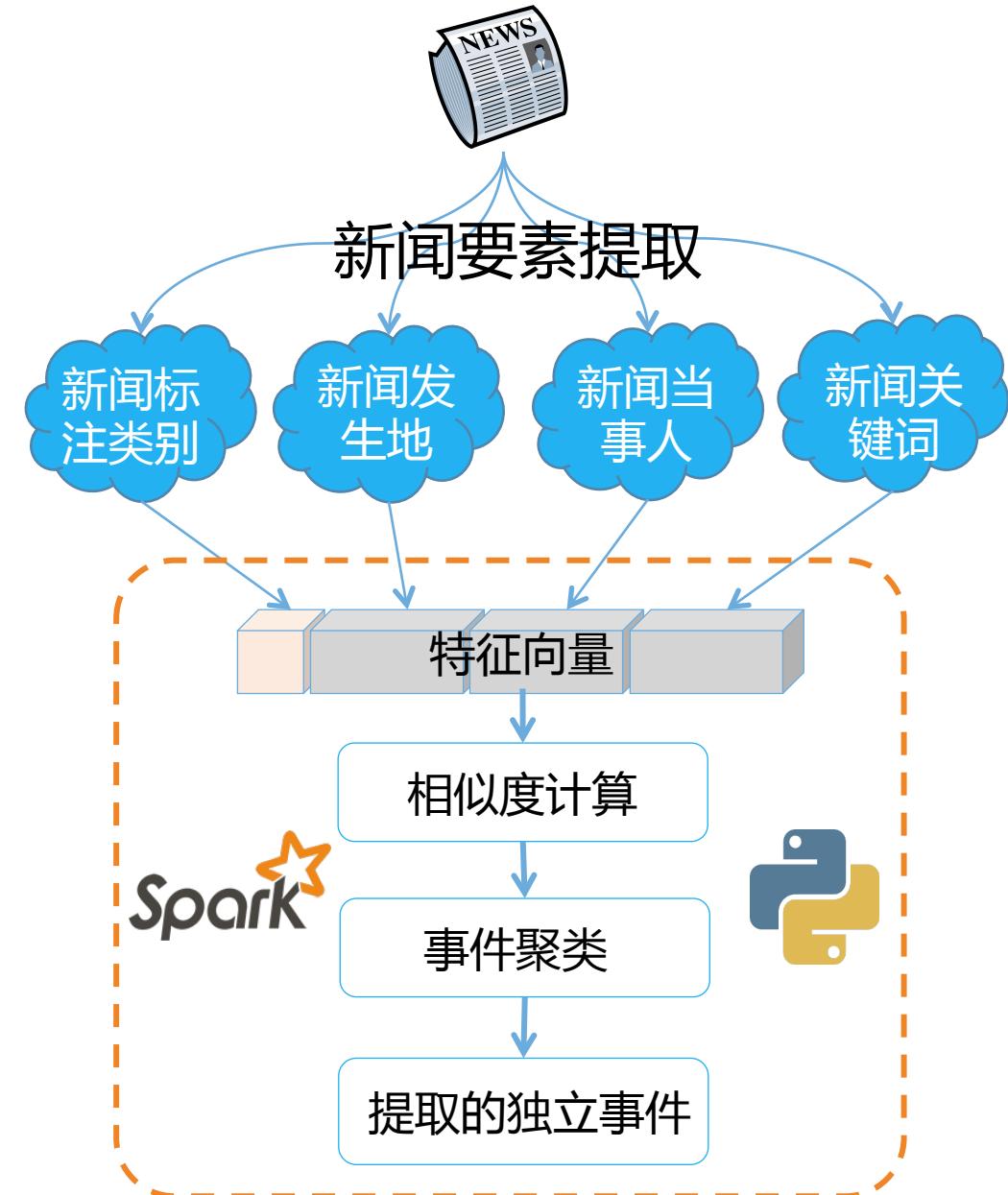
- 事件描述的语境特征

- 以无监督的方式自主学习(基于Google word2vec)
- 从传统词频统计到词语语境关联



事件聚类算法

- 独立事件聚类及Spark并行处理
 - 多角度新闻要素提取



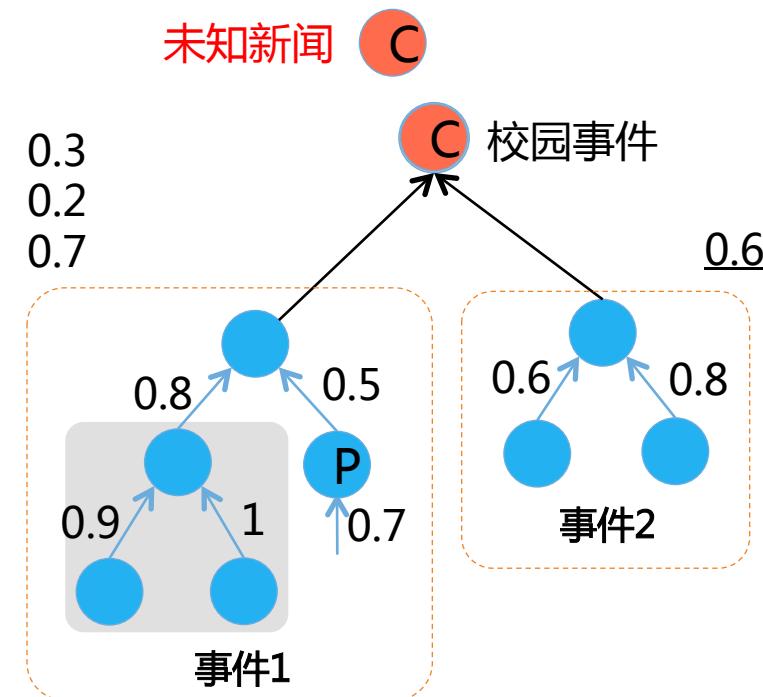
事件聚类算法

- 独立事件聚类及Spark并行处理
 - 多角度新闻要素提取
 - 采用并行和事件树结构进行优化

算法约束条件

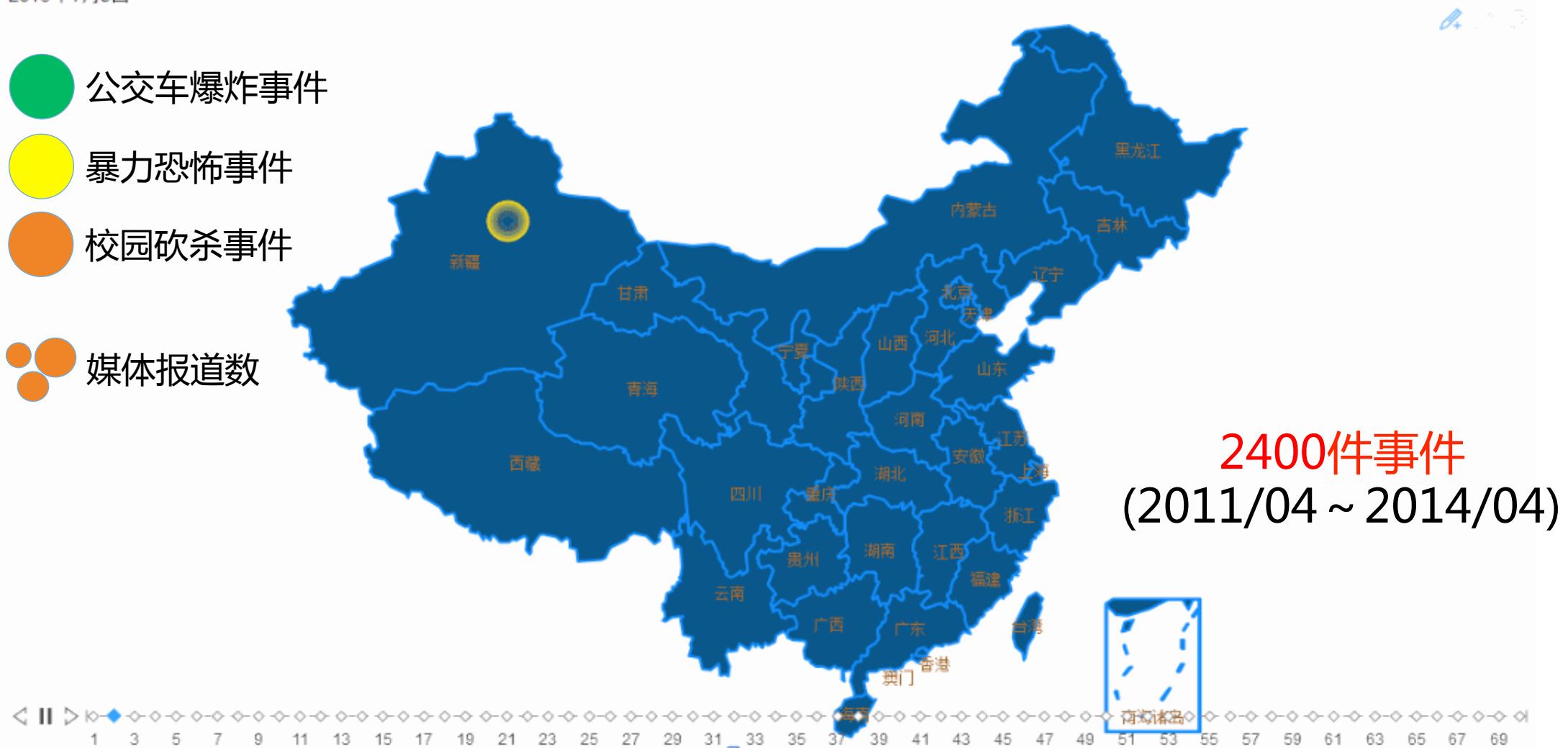
- $t_P \leq t_C$
- $\text{Sim}(P, C) \geq \text{Sim}(X, C)$
 $(\forall X, t_X < t_C)$

事件聚类算法Spark 并行演示



事件聚类结果可视化

2013年1月8日



知识驱动的数据分析过程



- 数据获取及预处理
(任务1)
- 新闻分类/事件聚类
(任务2)
- 事件关联分析
(任务3,4)

特征提取

- 季节
- 民族节日
- 工作日、休息日

时间



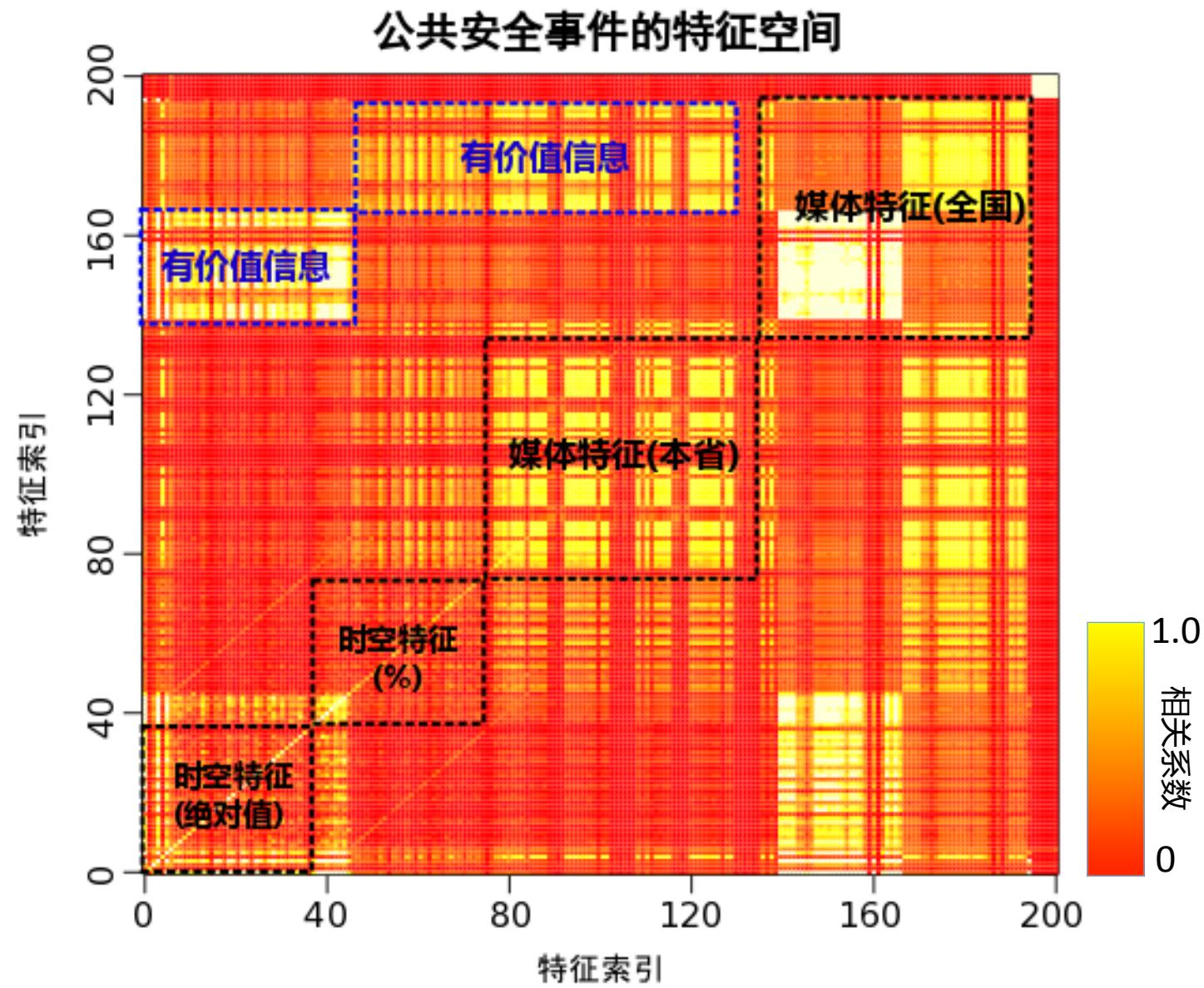
- 省、市地理划分
- 城市GDP
- 人口、民族组成

空间



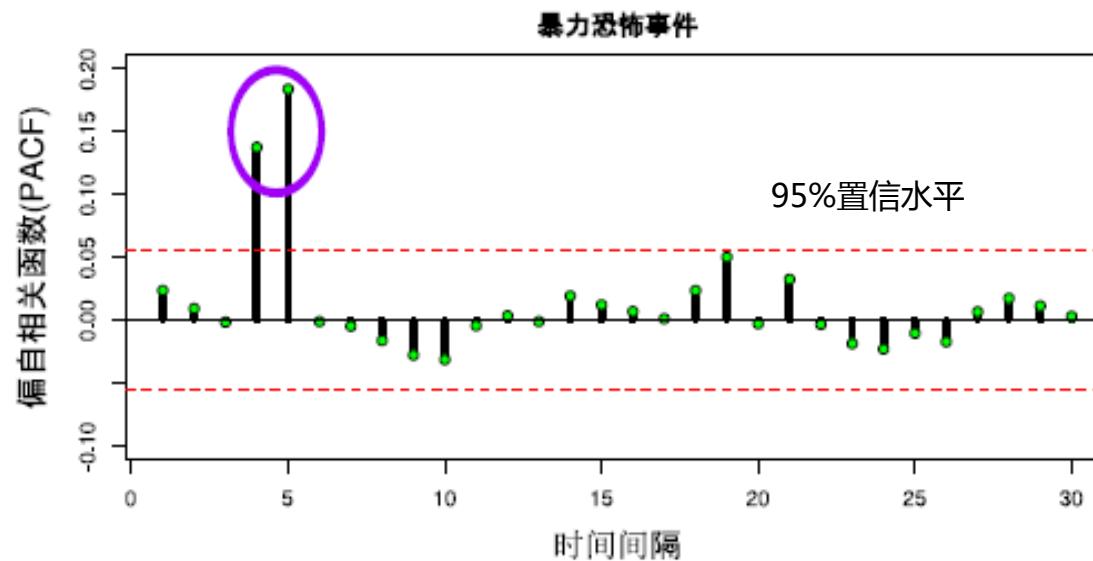
- 新闻报道
- 微博舆论
- 正负情感

媒体

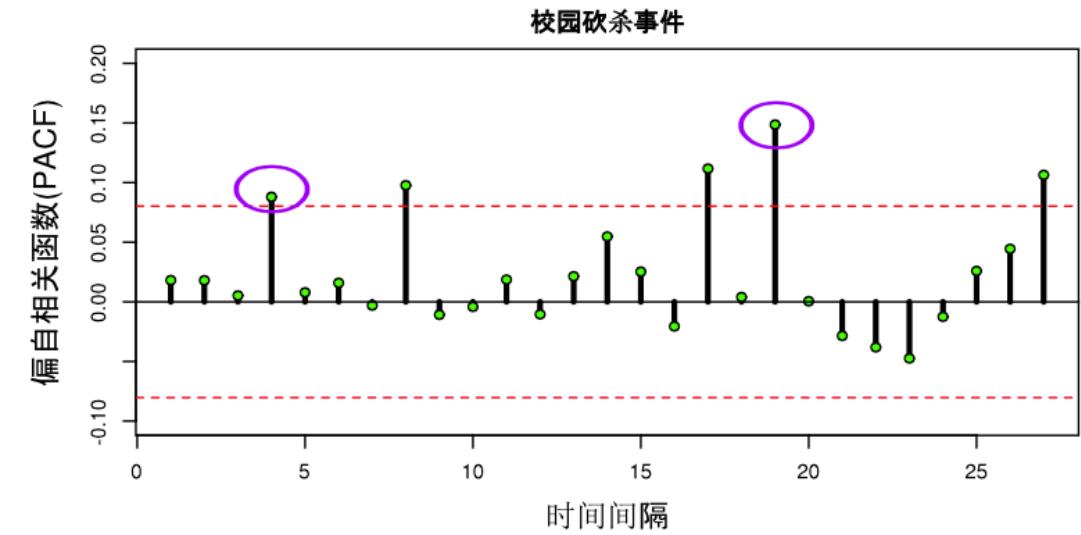
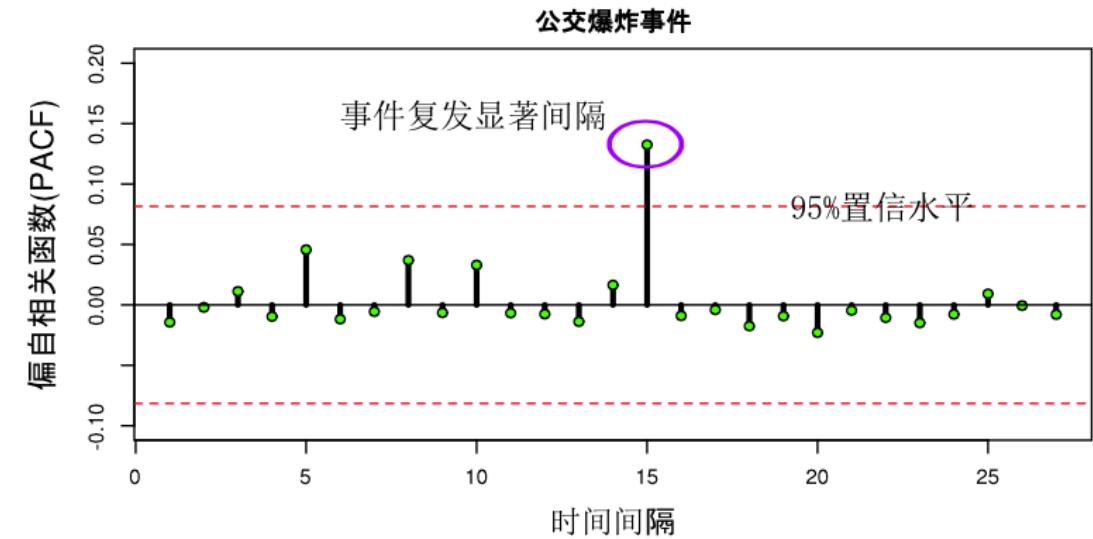


关联分析

- 同系列事件触发关系
 - 时间触发关系——事件频次自相关分析



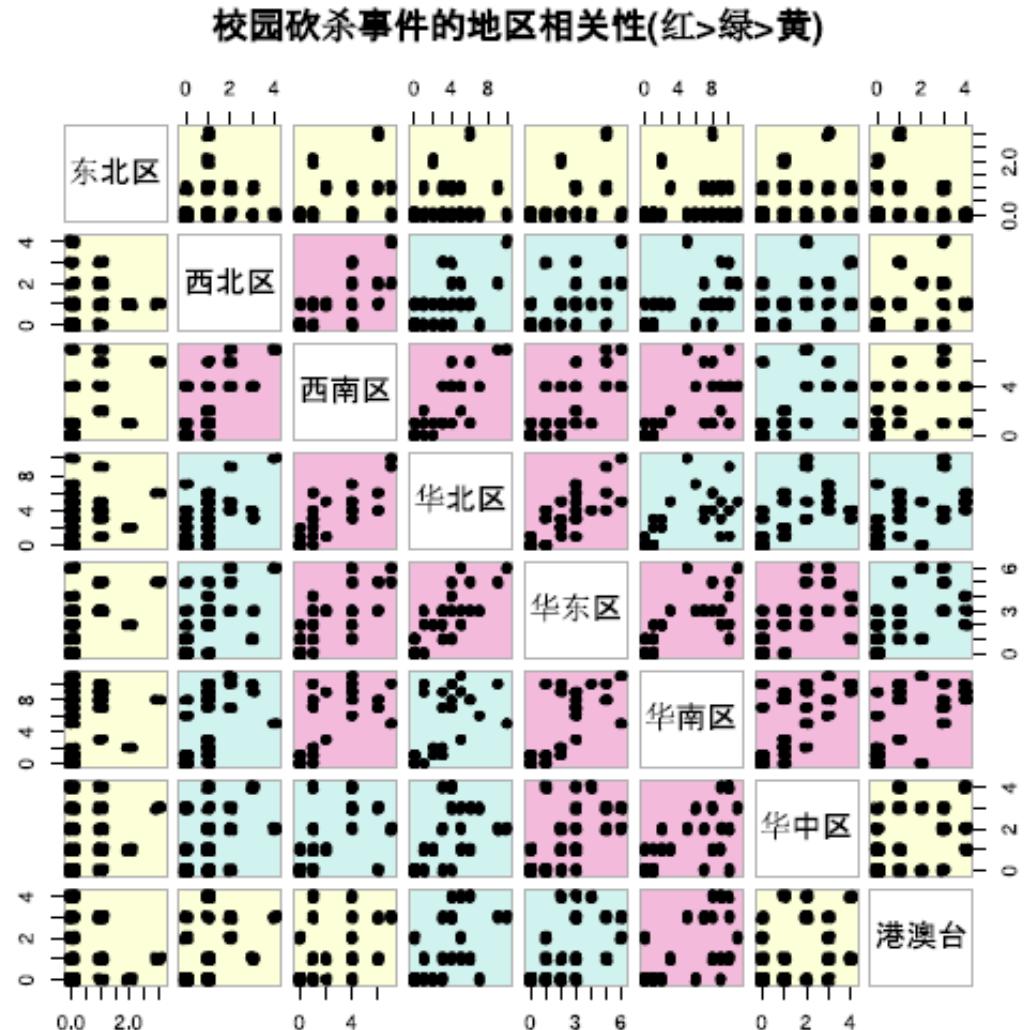
暴力恐怖事件每隔5天复发概率最高



关联分析

- 同系列事件触发关系
 - 时间触发关系——事件频次自相关分析
 - 空间触发关系——最大信息量相关系数(MIC)

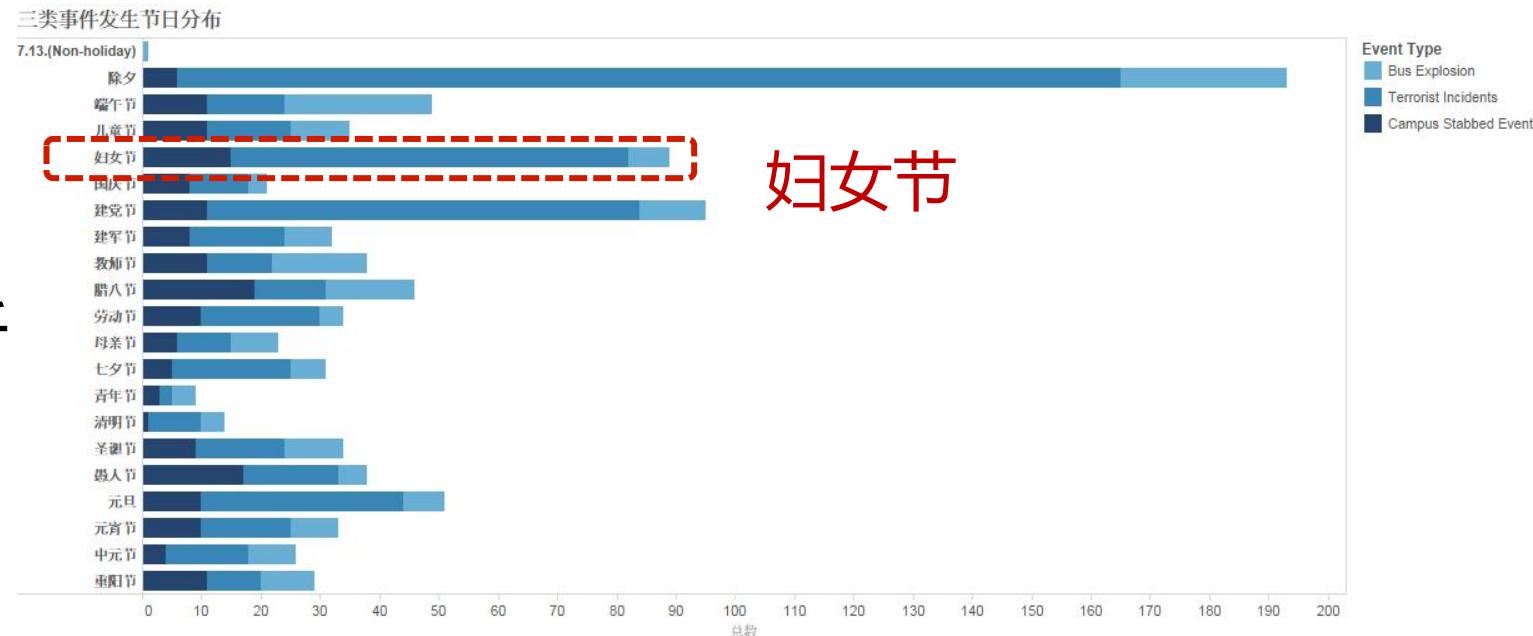
校园砍杀事件发生次数
在临近区域相似度较高



* MIC: Reshef et al. "Detecting Novel Associations in Large Data Sets". Science 334 (6062)

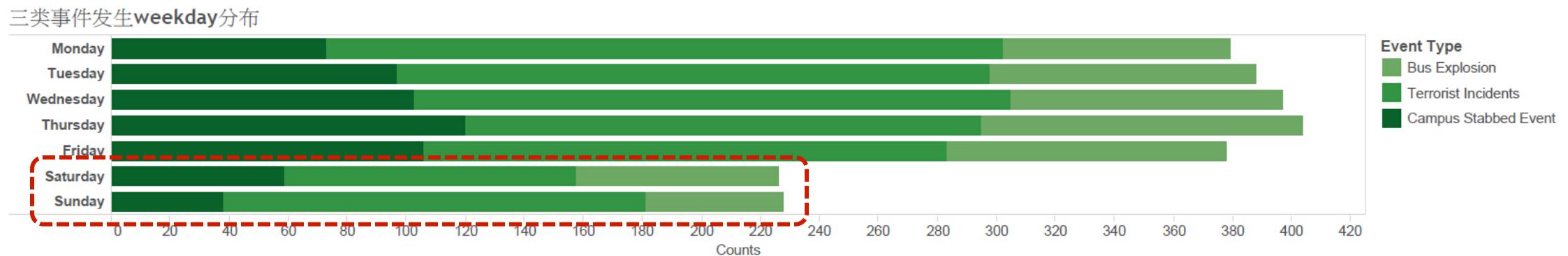
关联分析

- 不同系列事件共性分析
 - 时间特征



妇女节

元旦，除夕，建党节等是三类事件的多发时段

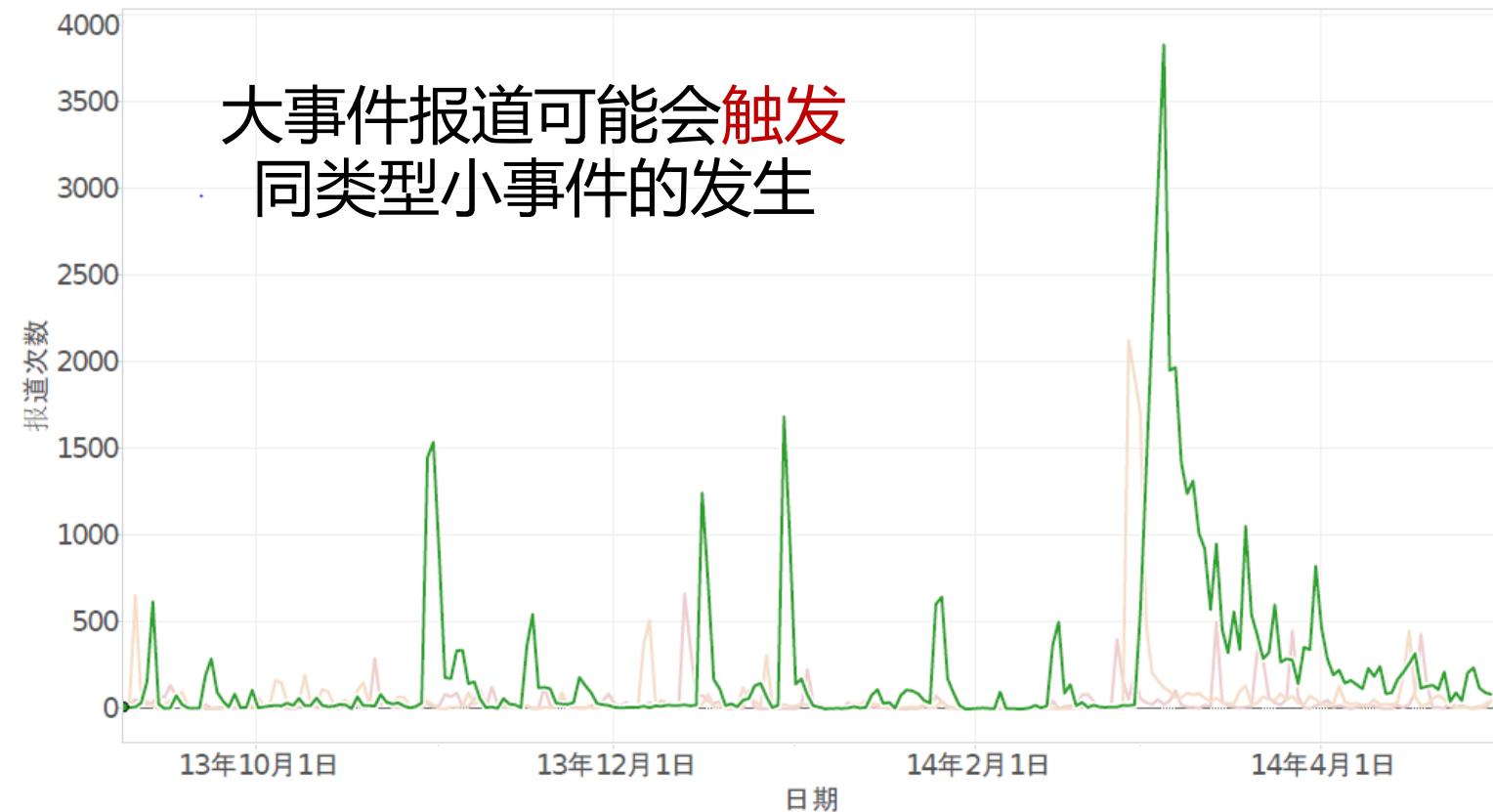


三类事件在周末发生频次低于工作日

关联分析

- 不同系列事件共性分析

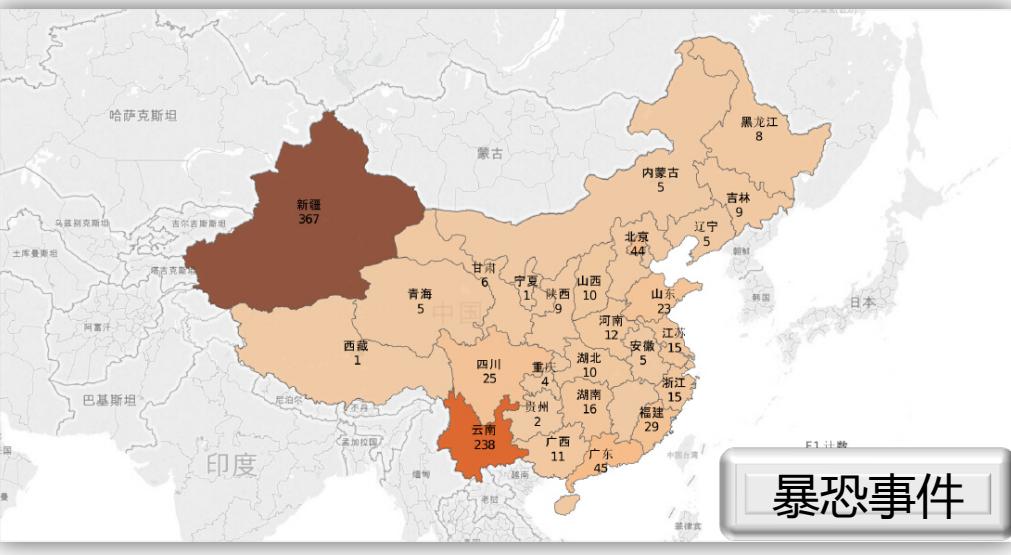
- 时间特征
- 媒体特征



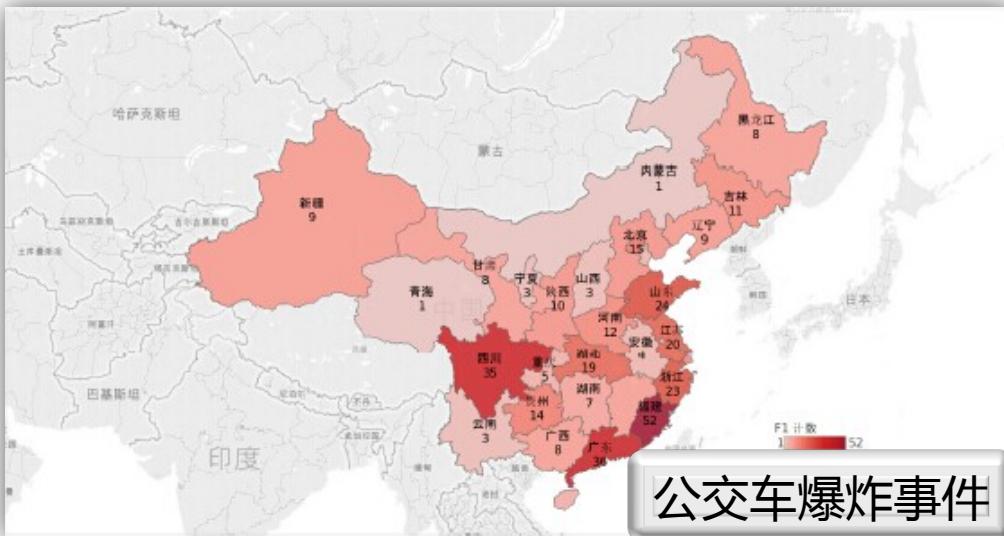
关联分析

• 不同系列事件共性分析

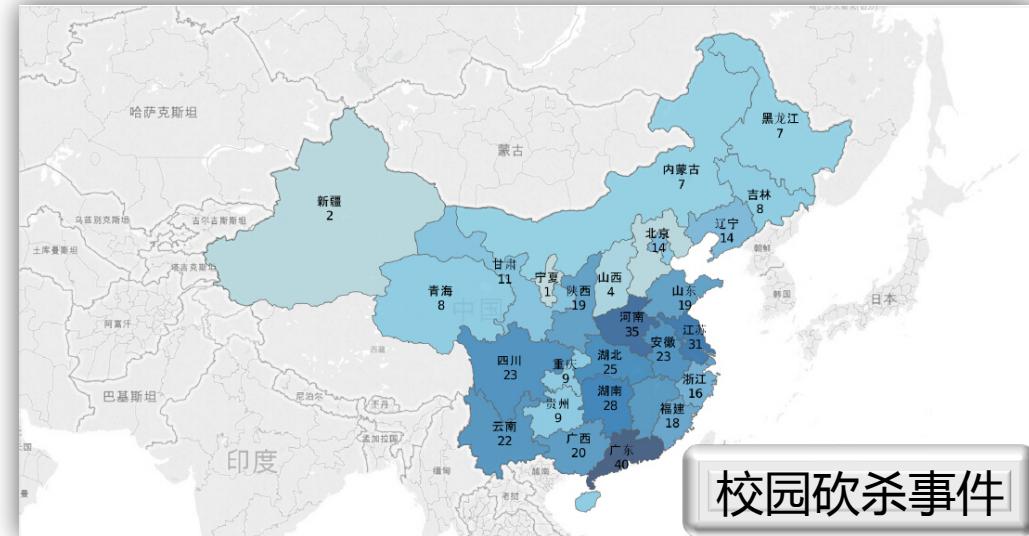
- 时间特征
- 媒体特征
- 空间特征



集中性的空间分布

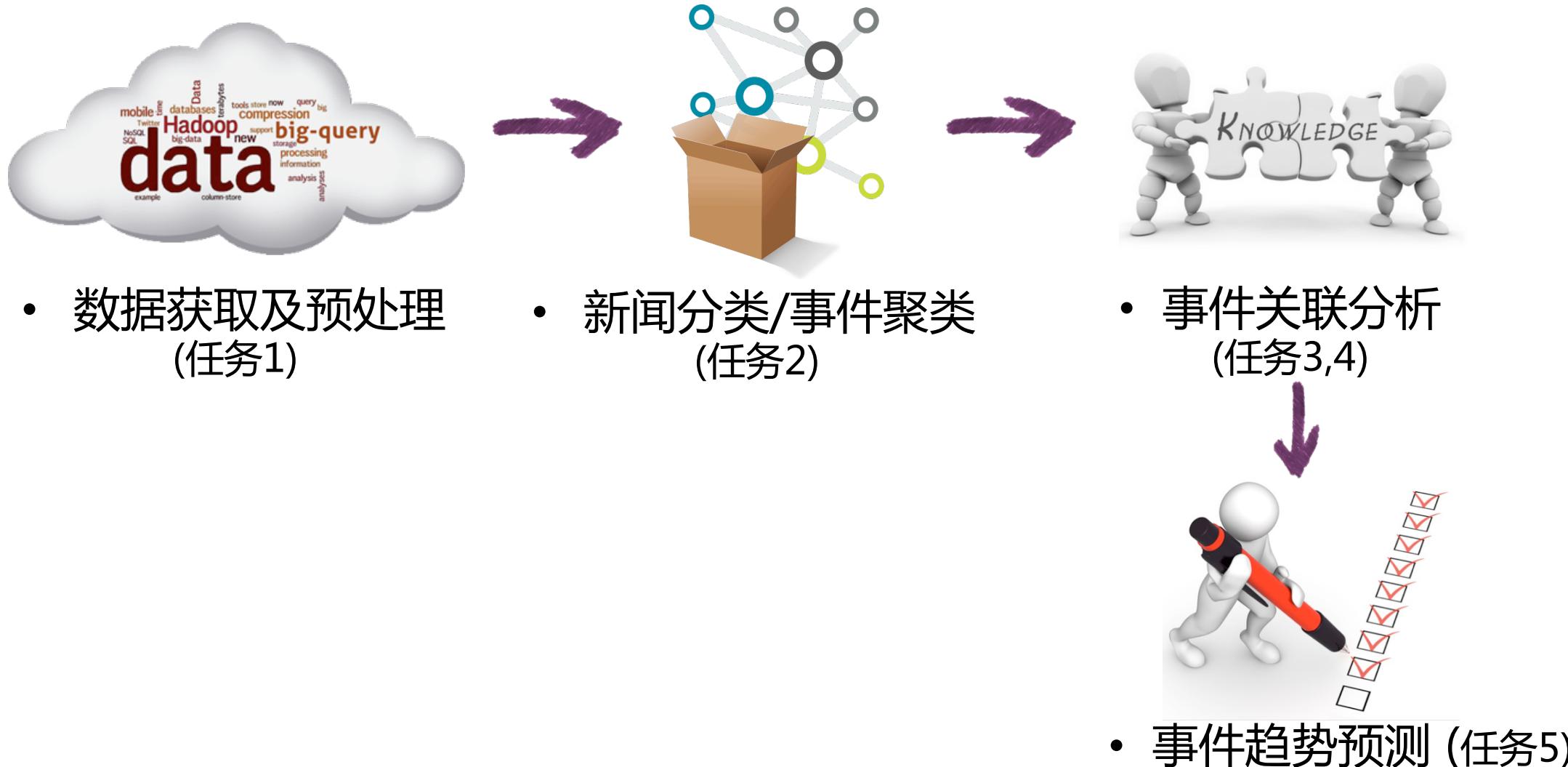


公交车爆炸事件



校园砍杀事件

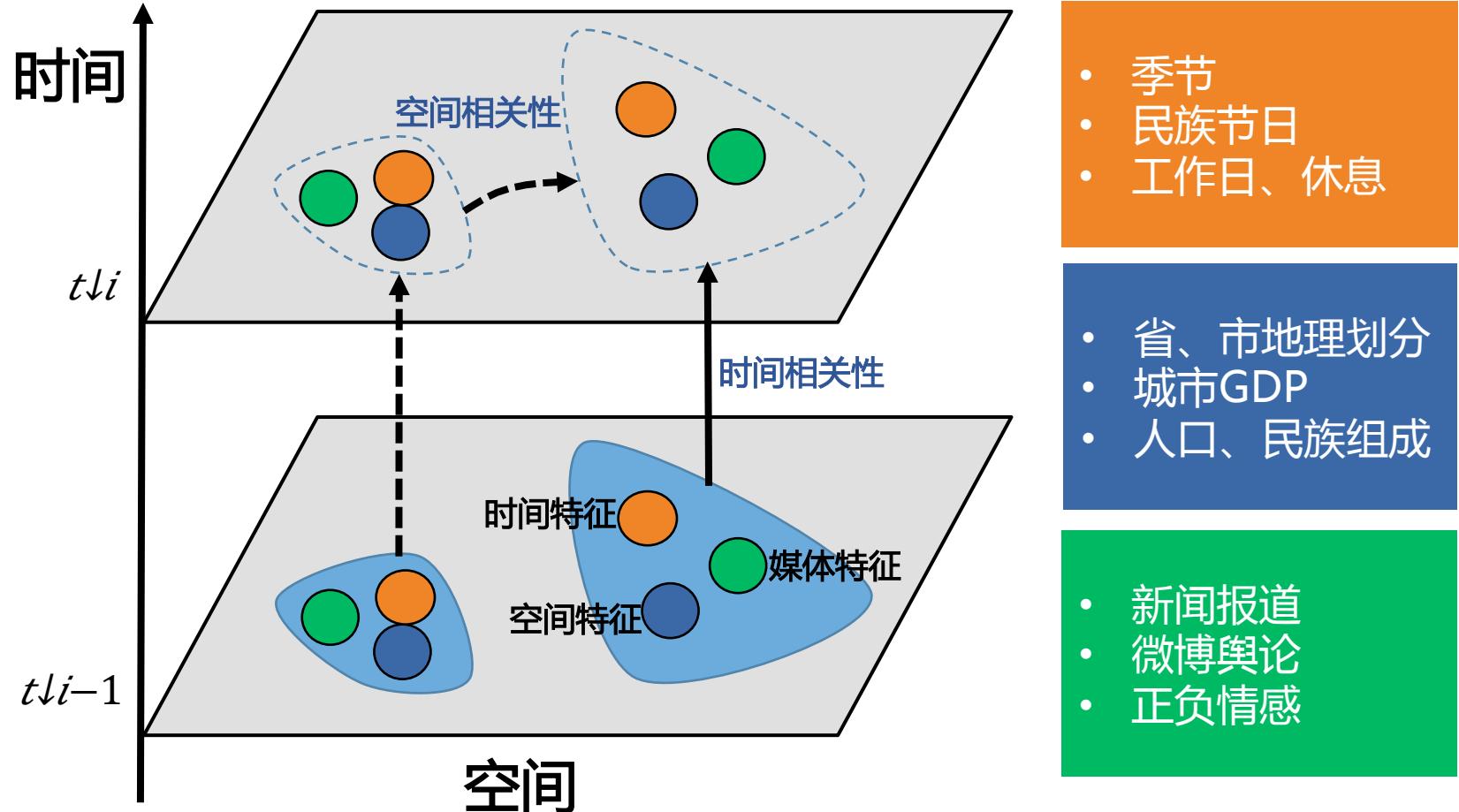
知识驱动的数据分析过程



事件预测

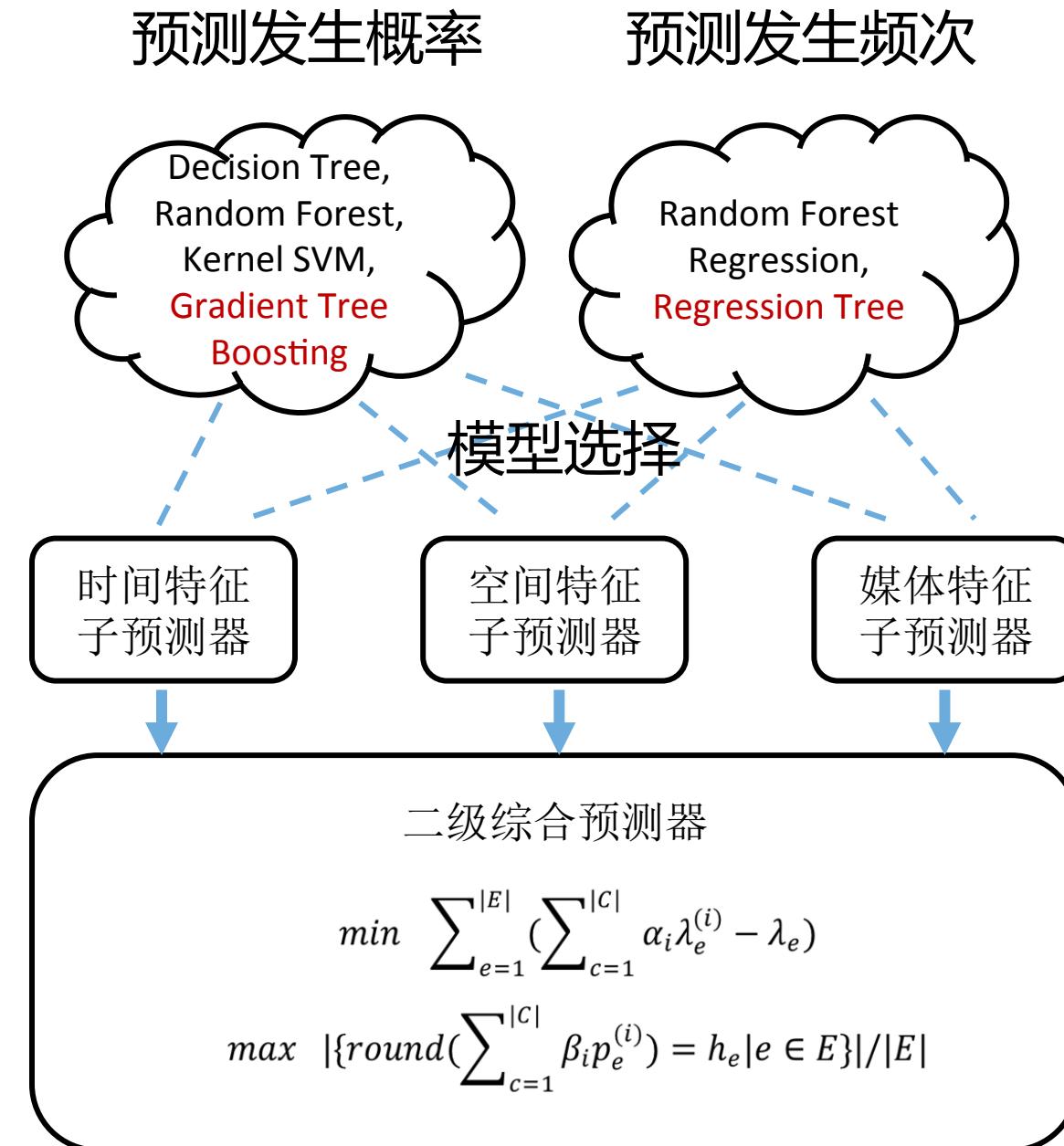
- 预测目标

- 未来时间段内事件是否发生(0/1)
- 未来时间段内事件发生的次数



事件预测

- 预测模型建立
 - 多维度特征分类建模
 - 时间特征具有最好的预测效果



算法评估

媒体名称	发布时间	新闻标题	事件类型
搜狗新闻	2013-12-16	河南砍学生男子患20年癫痫病	校园砍杀
新华网	2013-01-23	面包车自燃 公交司机徒手拔断着火线路帮助灭火	无
天津在线	2014-03-19	乌鲁木齐发生持械袭警案 嫌犯被民警当场击毙	暴恐事件

新闻分类
算法评估

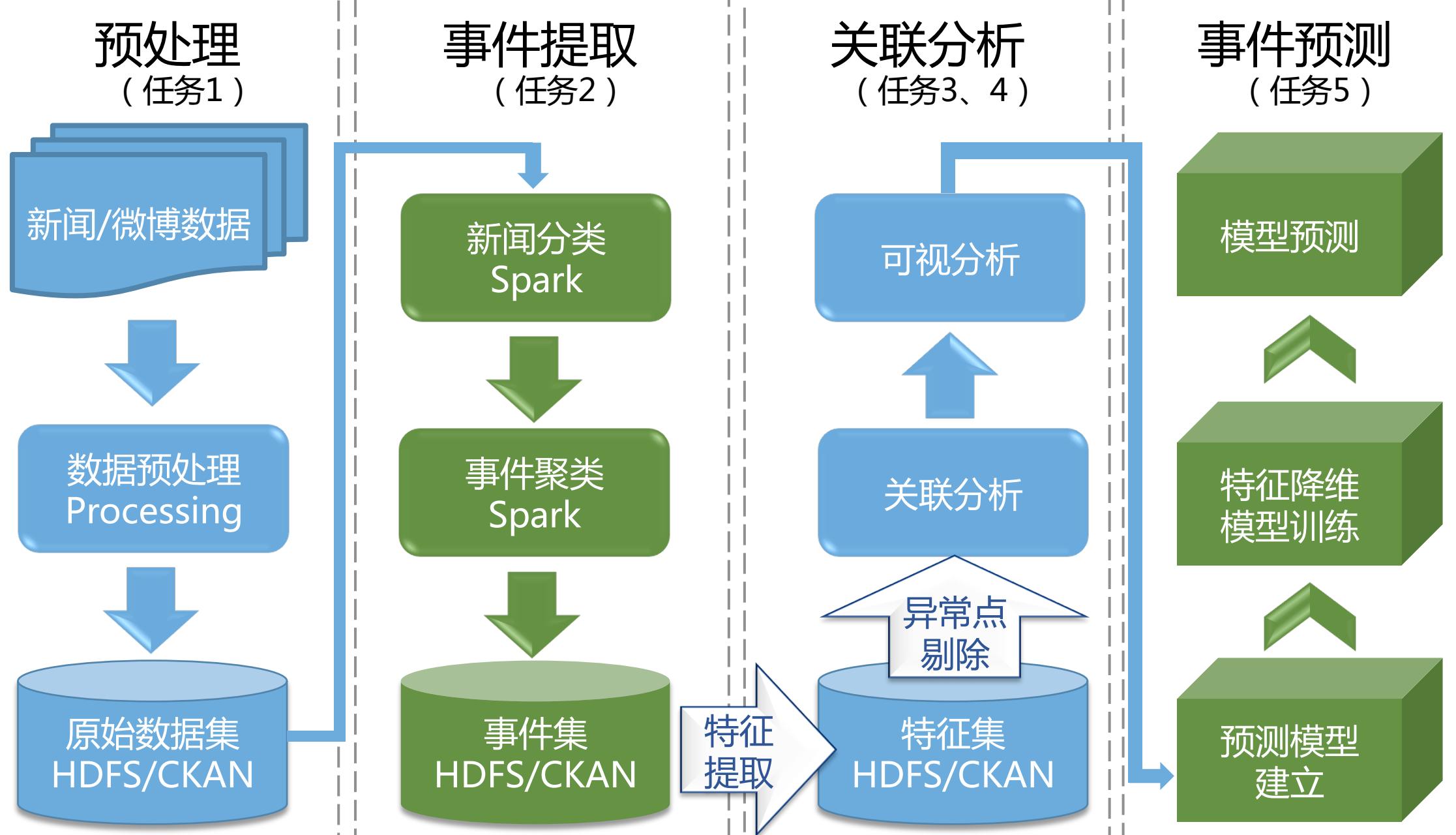
准确度 ~ 95%

事件聚类算法评估

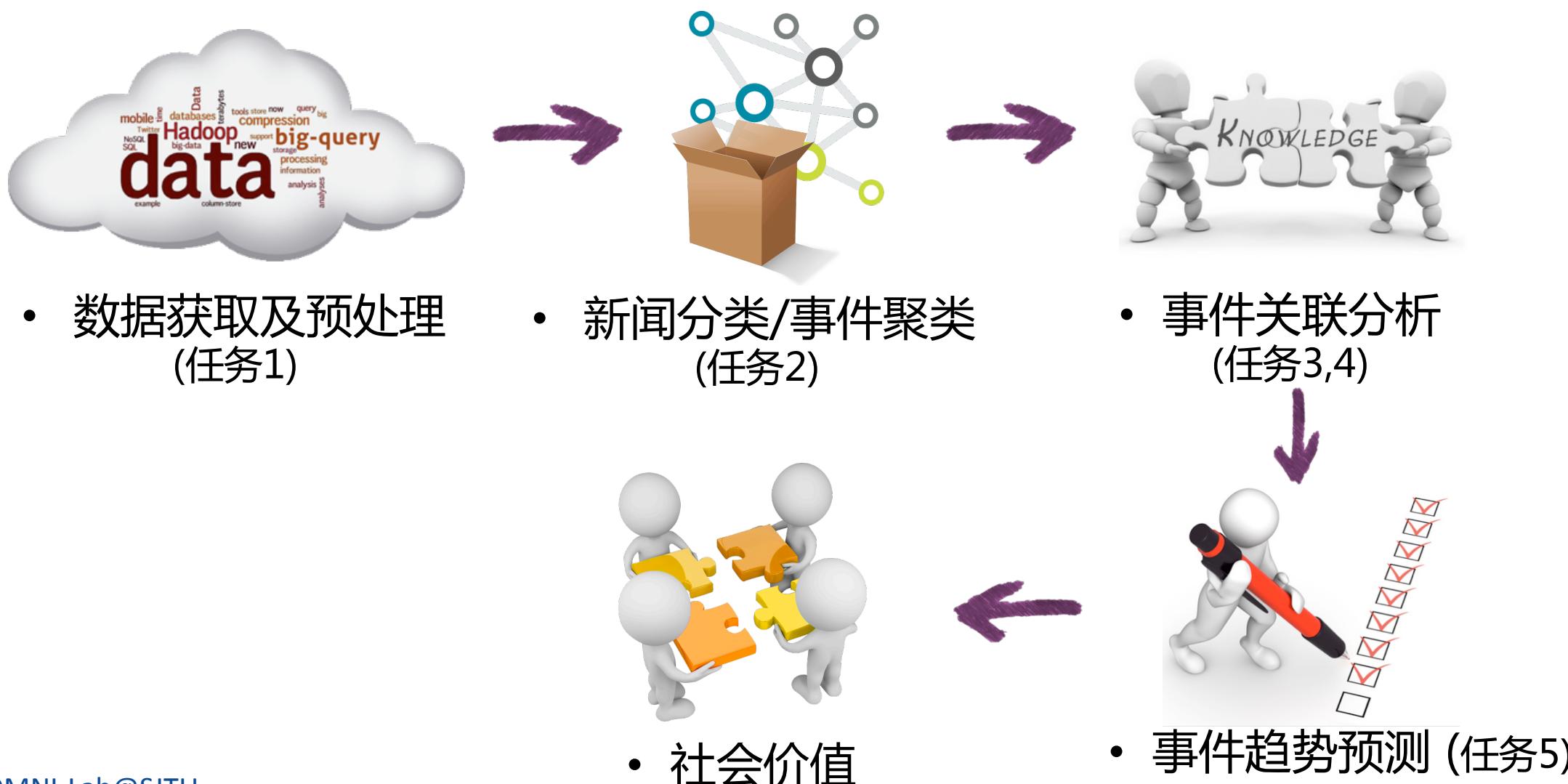
事件类型	误报率	漏报率
公交车爆炸事件	14.28%	12.09%
暴恐事件	12.39%	14.05%
校园砍杀事件	14.10%	11.54%

事件预测算法评估

评估方法	准确率	预测频次误差
测试集验证	64.50%	0.8956
留一验证	82.34%	0.5250
K-Fold	82.34%	0.5234
滑动窗口	75.27%	0.5525



知识驱动的数据分析过程



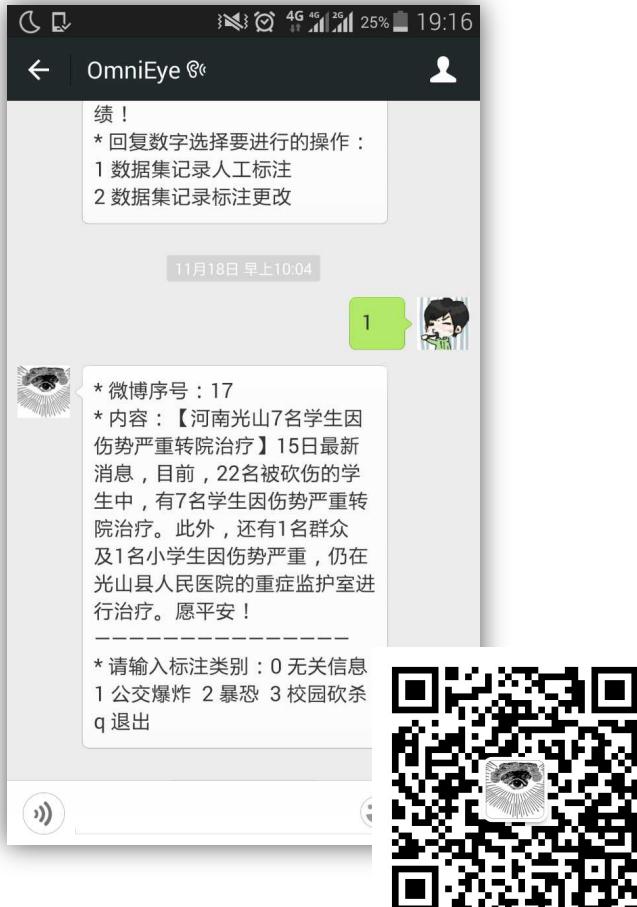
大数据竞赛概况

大数据到数据科学的演变

数据科学实践：公共安全事件研究

数据的社会价值

社会众包标注



The image displays three overlapping windows. The top window is titled 'Your Plugin Code' and contains tabs for 'HTML5', 'XFBML', and 'IFRAME'. It includes a placeholder text 'Include the JavaScript SDK on your page once, i' and a large block of JavaScript code:

```
<div id="omni-root">
  <script>{<div id="omni-root">
var js, fjs = d.getElementsByTagName(s);
if (d.getElementById(id)) return;
js = d.createElement(s); js.id = id;
js.src = "//data.sjtu.edu.cn/zh_CN/sdk/
fjs.parentNode.insertBefore(js, fjs);
}(document, 'script', 'omnieye-jssdk'));
```

The middle window is a 'Please sign in' form with fields for '邮箱地址', '密码', and '验证码' (with the value '0713'). Below the form is a news summary: '新疆巴楚15名警察社区工作人员遭暴徒袭杀' and '本报讯, 4月26日晚, 位于新疆维吾尔族自 [阅读更多](#)'. The bottom window is a modal for selecting event types, with radio buttons for '公交车爆炸事件', '暴恐事件', and '校园砍杀事件', and a large blue '登录' button.

开放数据共享

The screenshot shows the homepage of the OMNILab Open Data Sharing Platform. The header features the CKAN logo and navigation links for '数据集' (Dataset), '组织' (Organization), '群组' (Group), and '关于' (About). A search bar is also present. The main visual is a dark background with a 3D geometric cube pattern. The text 'OMNILab开放数据共享平台' (OMNILab Open Data Sharing Platform) is prominently displayed in the center, along with the tagline 'Every Sharing Makes The World Better'. Below this, there's a 'Join Now!' button and the text '上海交通大学网络信息中心'. At the bottom, it says '我们现在已经拥有' (We now have) followed by 'OmniCKAN 数据统计' (OmniCKAN Data Statistics) showing 6 datasets, 2 organizations, 0 groups, and 0 related items. It also encourages users to '加入我们, 贡献你的力量' (Join us, contribute your strength) and '在广阔的开放数据平台中挖掘无限可能' (Explore infinite possibilities in the vast open data platform).

The screenshot shows a CKAN dataset page for the '第二届中国大数据技术创新大赛' (Second China Big Data Innovation Competition). The top navigation bar includes '登录' (Login) and '注册' (Register). The page title is '数据集 / 第二届中国大数据技术创新大赛'. On the left, there's a sidebar with '第二届中国大数据技术创新大赛' and a '追从者' (Follower) count of 5. It lists social sharing options for Google+, Twitter, Facebook, and other platforms, along with an 'OPEN DATA' link. The main content area displays several datasets with their file types (e.g., CSV, XLS), descriptions, and download links. These include:

- 新闻数据集人工标注 (News dataset manual annotation)
- 微博数据集人工标注 (Weibo dataset manual annotation)
- 中国各省市行政区划一览表 (List of China's provinces and districts)
- 中国各民族传统节日总表 (List of China's traditional ethnic festivals)
- 中国各省市别称、气候、方言一览表 (List of China's provincial nicknames, climates, and dialects)

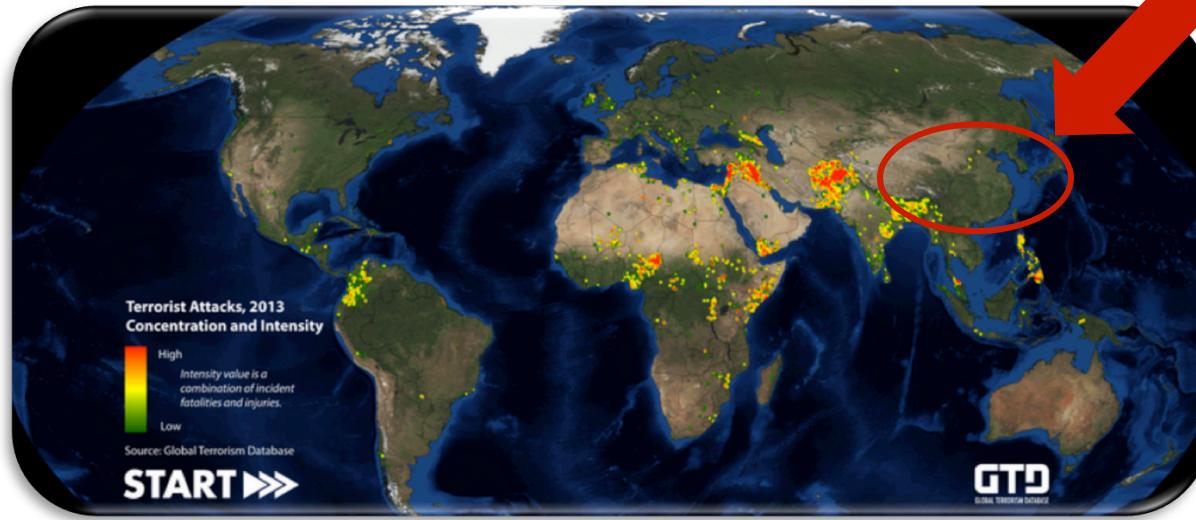
<http://data.sjtu.edu.cn>

中国地区数据稀缺！

未来

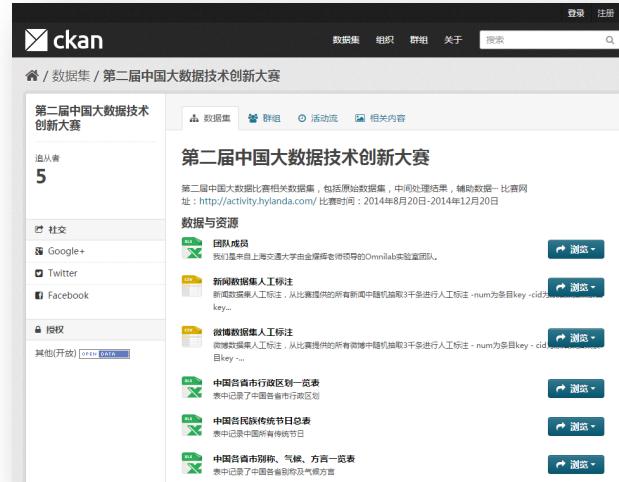


马里兰大学 (UMD)
<http://www.start.umd.edu/gtd/>



- 基于开放数据平台
- 自动爬取网络新闻报道、微博媒介传播
- 语境过滤 & 事件聚类 & 众包标注

数据生产力



CTD

CHINA TERRORISM DATABASE

致谢

- 海量 HYLANDA 大数据情报服务平台的 数据支持；
- 暨南大学应急管理学院 陈玉梅老师的 指导建议；
- 南风窗杂志社 戴玉老师提供的 新闻背景；
- 上海交通大学网络信息中心 金耀辉老师的 悉心指导；
- OMNILab 团队的所有成员。

请批评指正！
谢谢！