如何成为一个数据侦探 – 可视分析技术入门

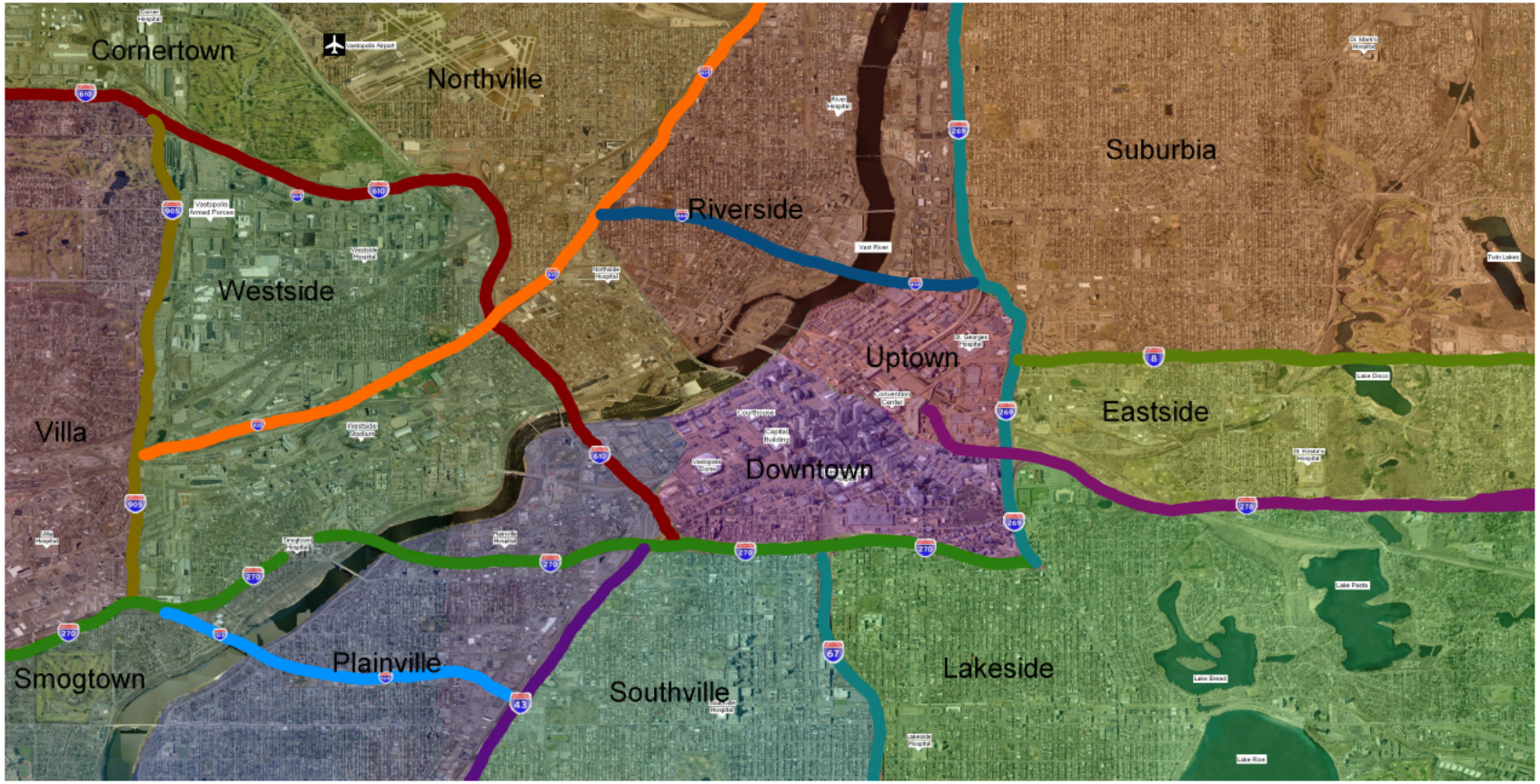
复旦大学 大数据学院 陈思明

http://simingchen.me

引子：调查一次流行病的爆发

- （注：数据来自于一个虚构的数据竞赛 IEEE VAST Challenge ）
- 拥有的数据：
  - 带有地理位置的Twitter数据
  - 一幅地图
- 你的角色：
  - 数据侦探

你拥有的地图


Cornertown
Northville
Suburbia
Riverside
Westside
Uptown
Eastside
Villa
Downtown
Smogtown
Plainville
Southville
Lakeside

你的任务

- 我们知道这个区域爆发了一场流行病（症状包括：发烧，发冷，盗汗，呼吸困难，腹泻等等），但除此之外别无所知

- 任务：
  - 找到爆发的开始时间和地点
  - 这个流行病是如何传播的？是否人传人？
  - 这个爆发有没有被控制住？如果需要投入人力物力进行救援，应该放到哪个区域？

第一步：破局 – 从数据入手

- 带有地理信息的Twitter数据
  - 它不是直接的疫情病例信息
  - 但是它也许可以反应出病例的分布与传播？
- 任务解析 – 时空探索
  - 时间分析：何时开始，何时爆发，趋势如何，是否被控制？
  - 空间分析：何地开始，何地爆发，趋势如何，何地较为集中？

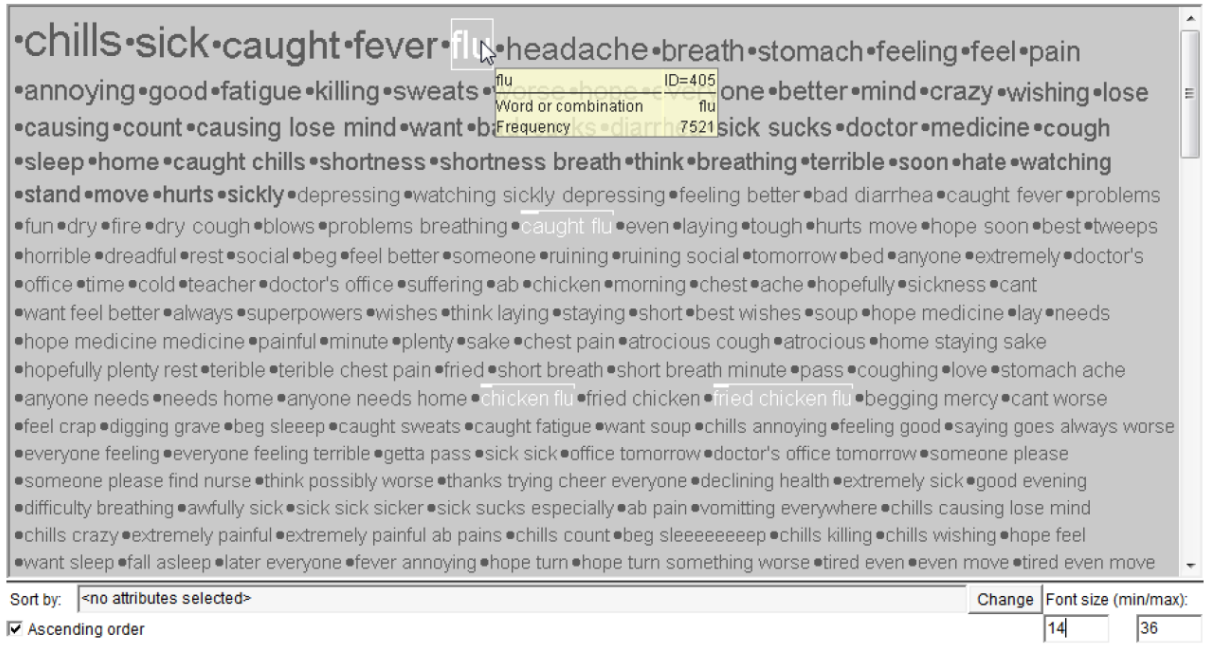
思考1： 如何建立Twitter数据与流行病之间的联系？

第一步：破局 – 从文本数据入手（2）

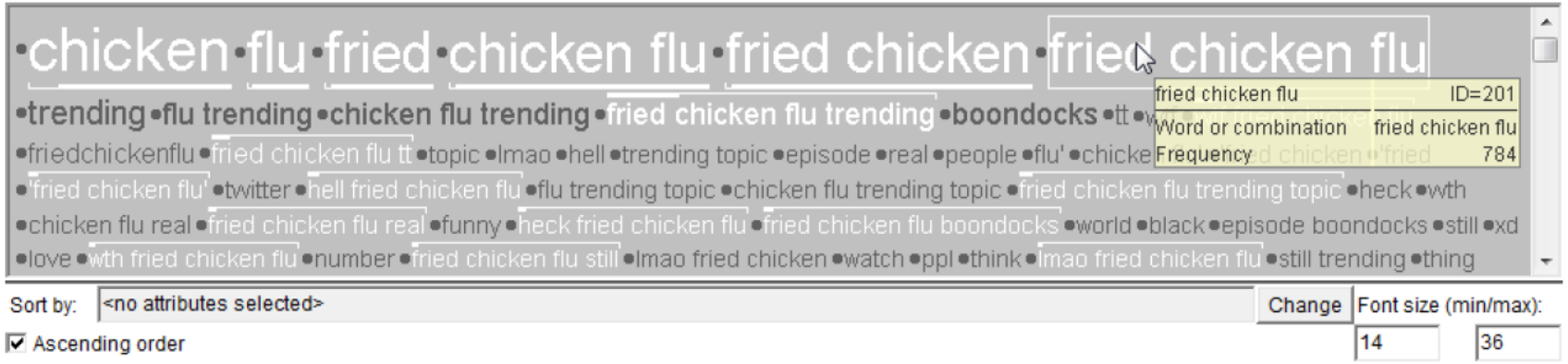
思考2： 从中获得什么样的信息？如何进一步改善？

第一步：破局 – 从文本数据入手（3）

- 数据处理：
  - 增加发现到的新的相关词汇（包括flu, stomach, sick, doctor），扩展搜索空间
  - 使用停用词：除去不相关的词

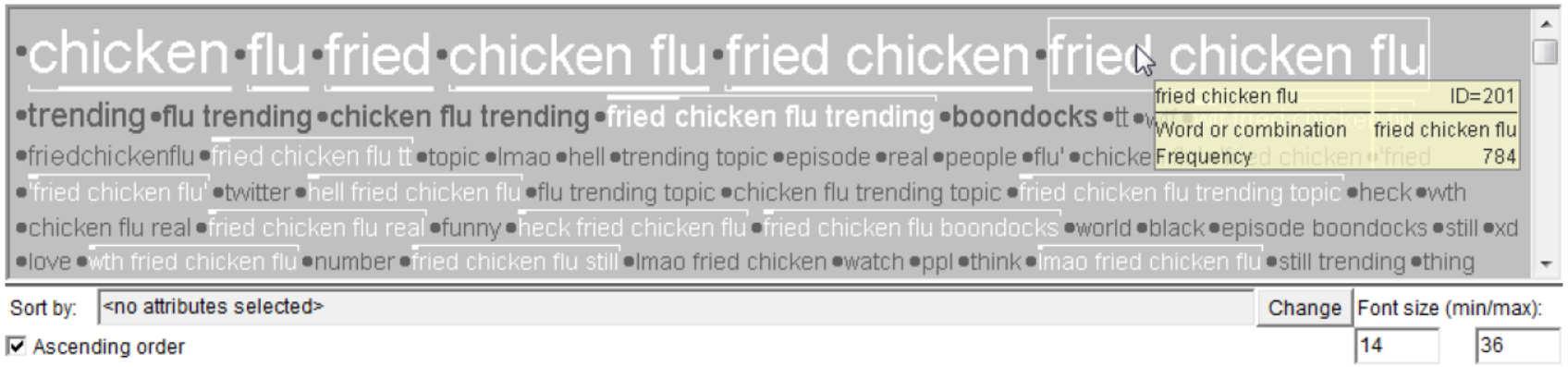


第二步：小心 – 数据中的陷阱



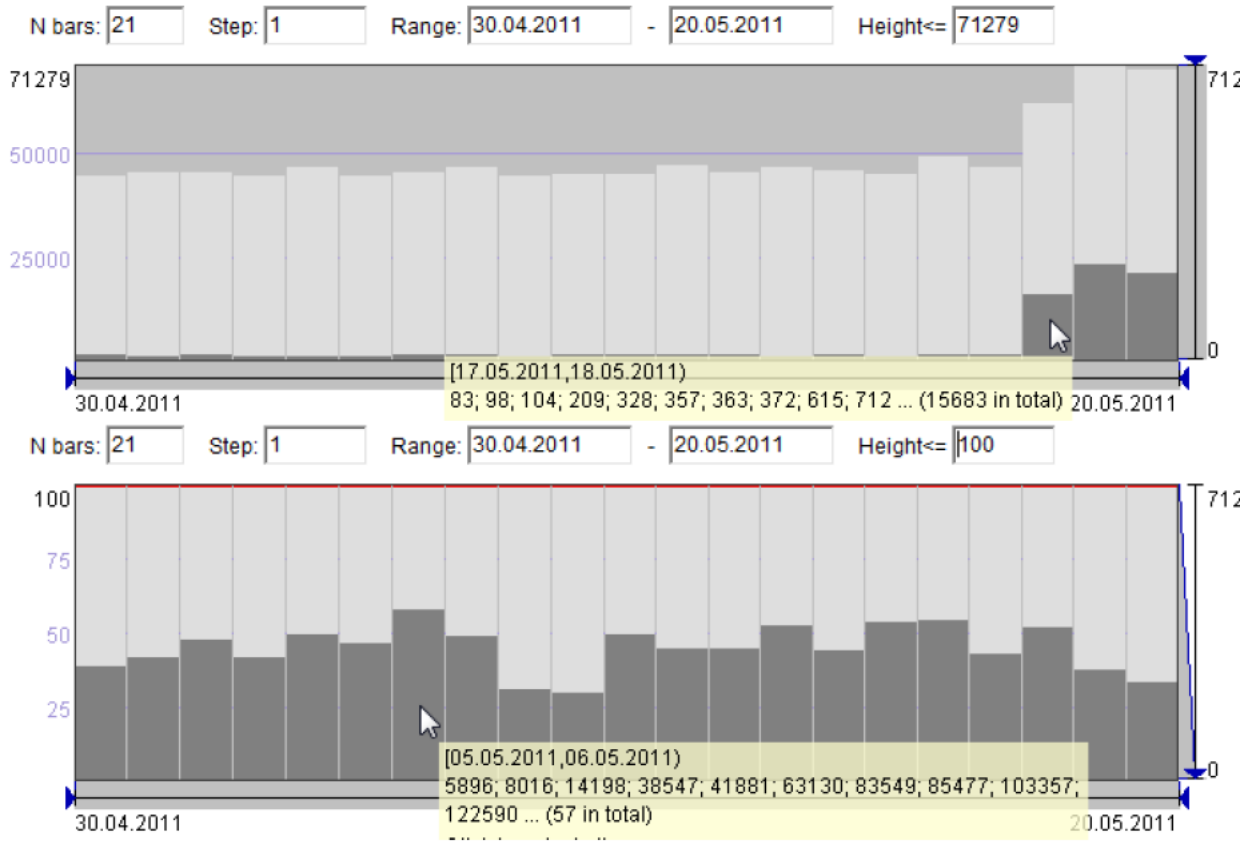
思考3： 你觉得这些关键词哪些有问题？哪些需要验证，如何验证呢？

第二步：小心 – 数据中的陷阱



- 禽流感？ 听起来不错，是不是这次疫情呢？
- 但“炸鸡流感”是什么？

第二步：小心 – 数据中的陷阱（2）

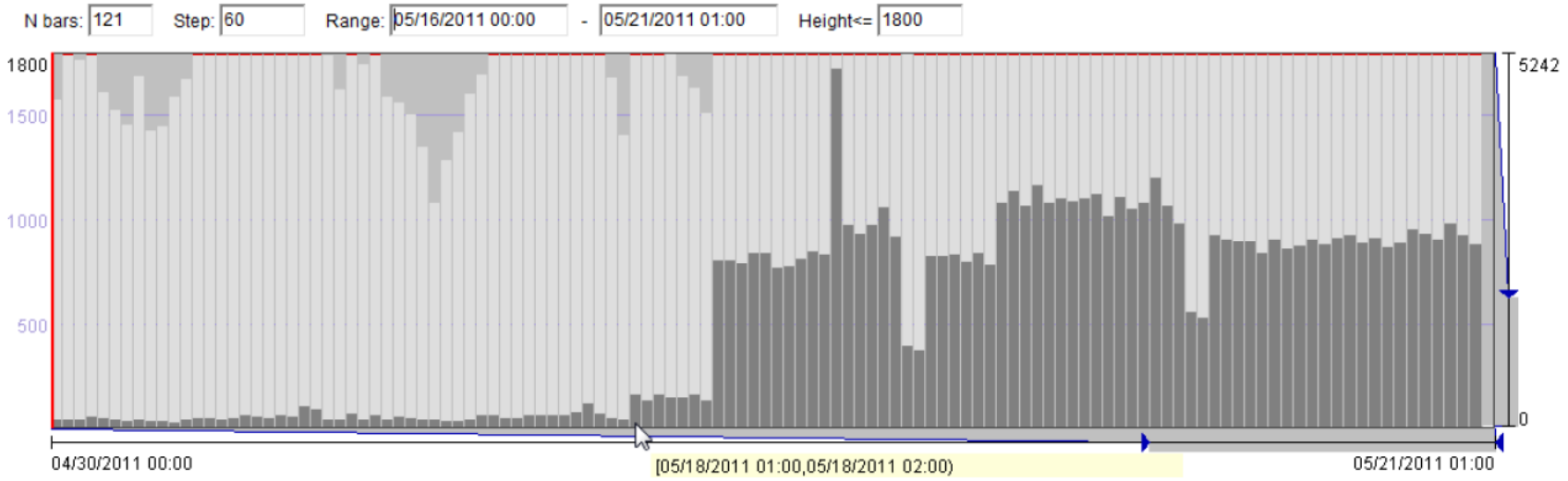


疾病相关Twitter

“Chicken” “Flu” 相关Twitter

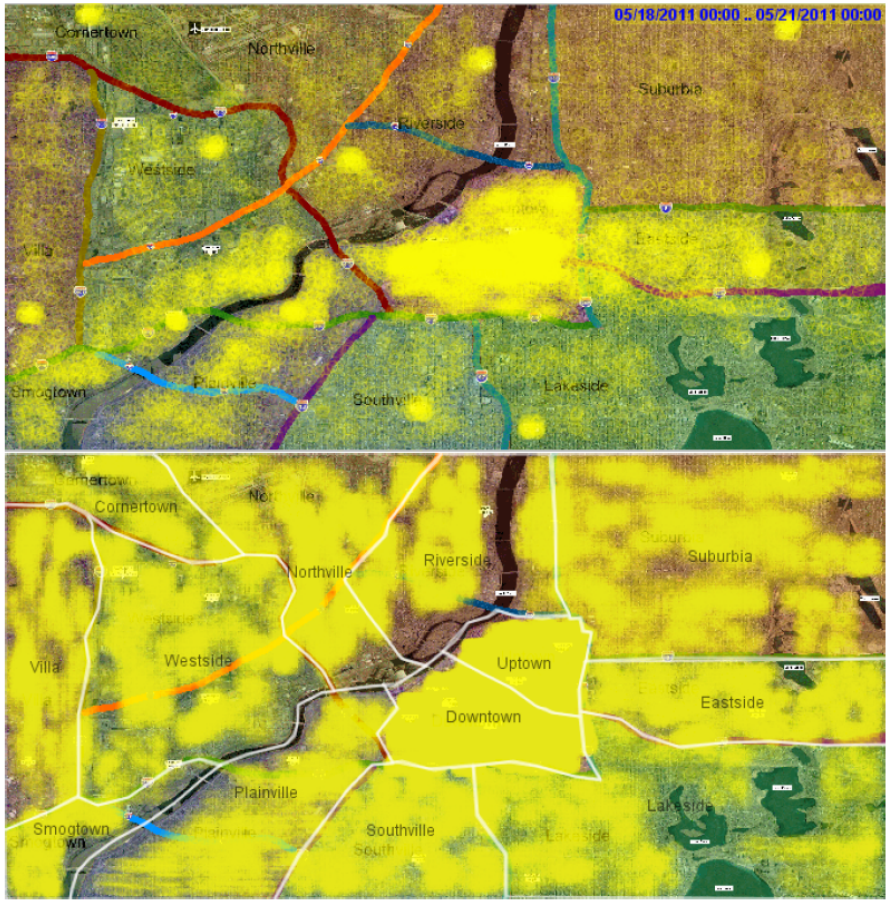
思考4：你发现了什么？那接下来如何进一步从时间角度探索疾病相关的Twitter？

第三步：关联 – 探索数据的时空特征



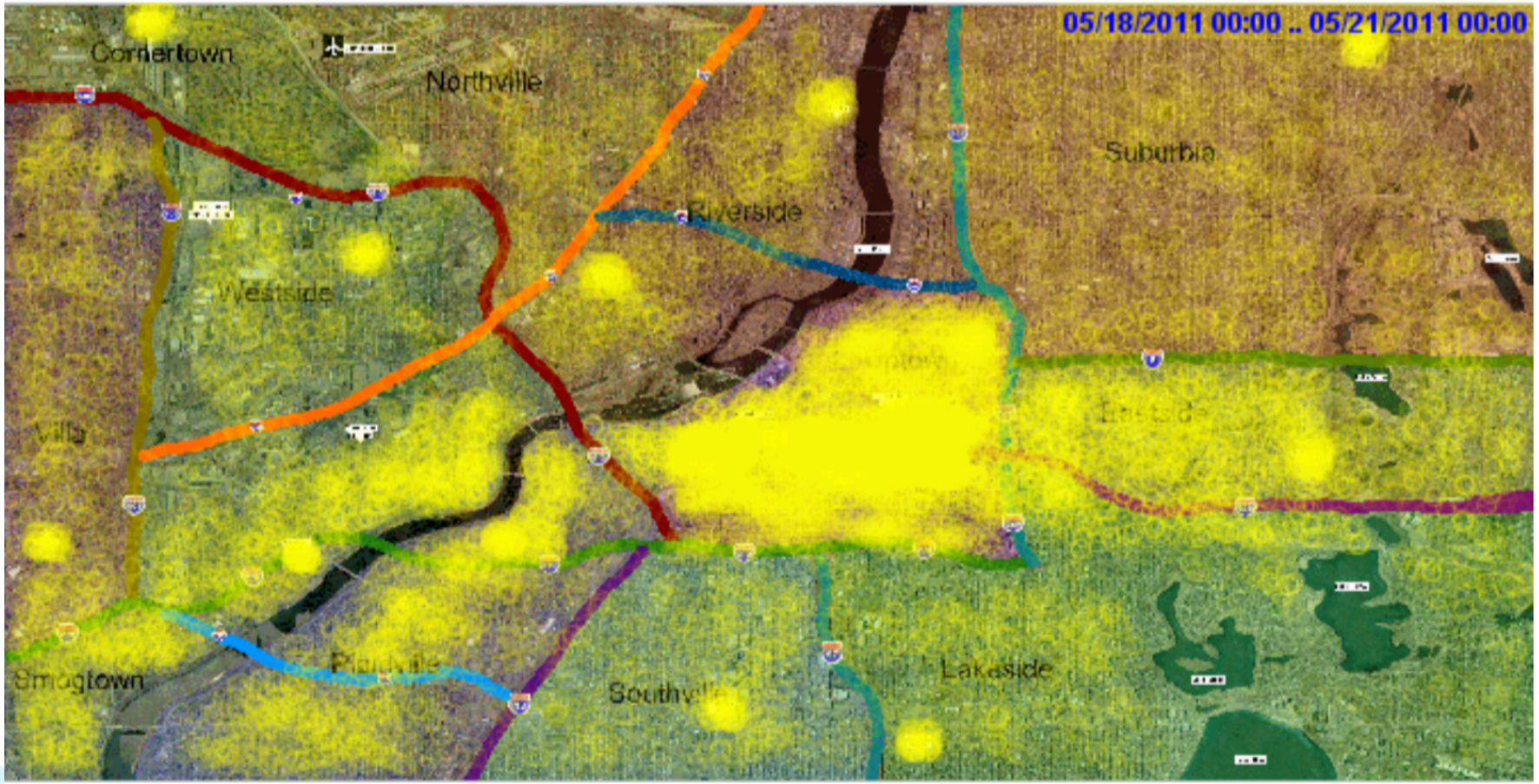
对时间特征进行细粒度的探索

第三步：关联 – 探索数据的时空特征（2）



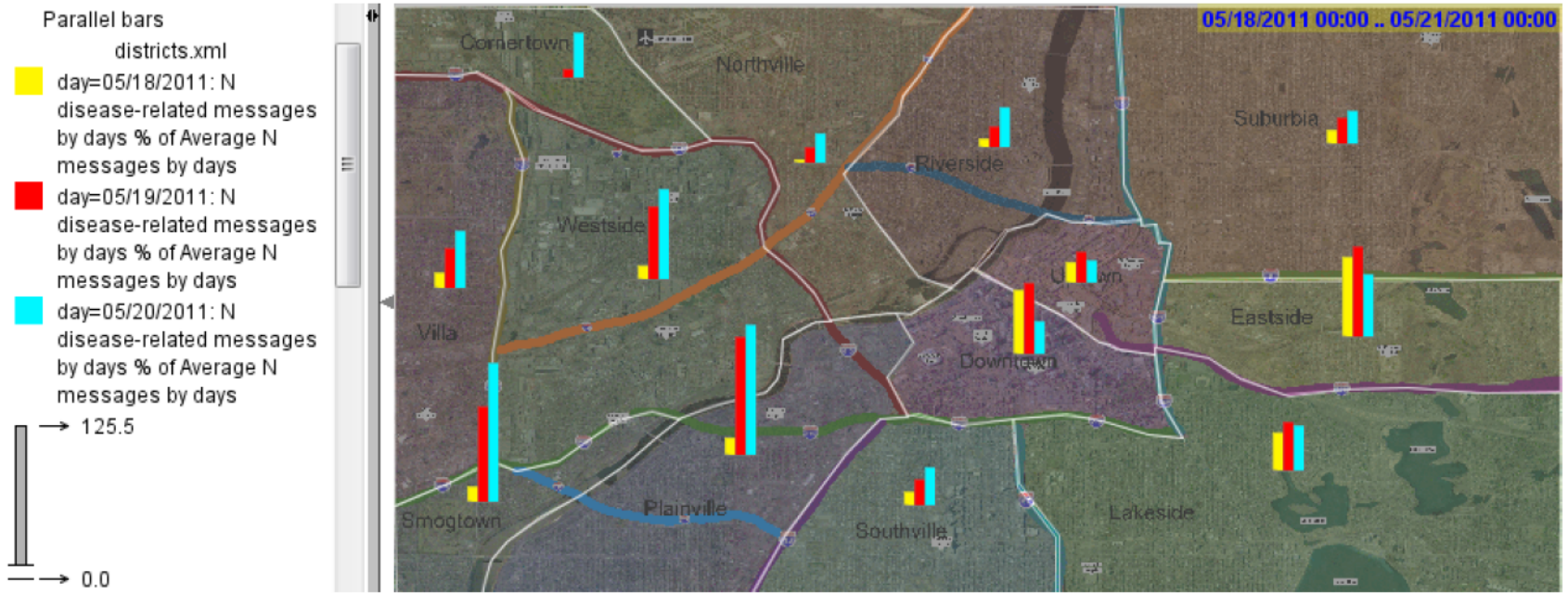
- 先前过滤出的Twitter的时空分布
- Twitter全集的时空分布

第三步：关联 – 探索数据的时空特征（2）



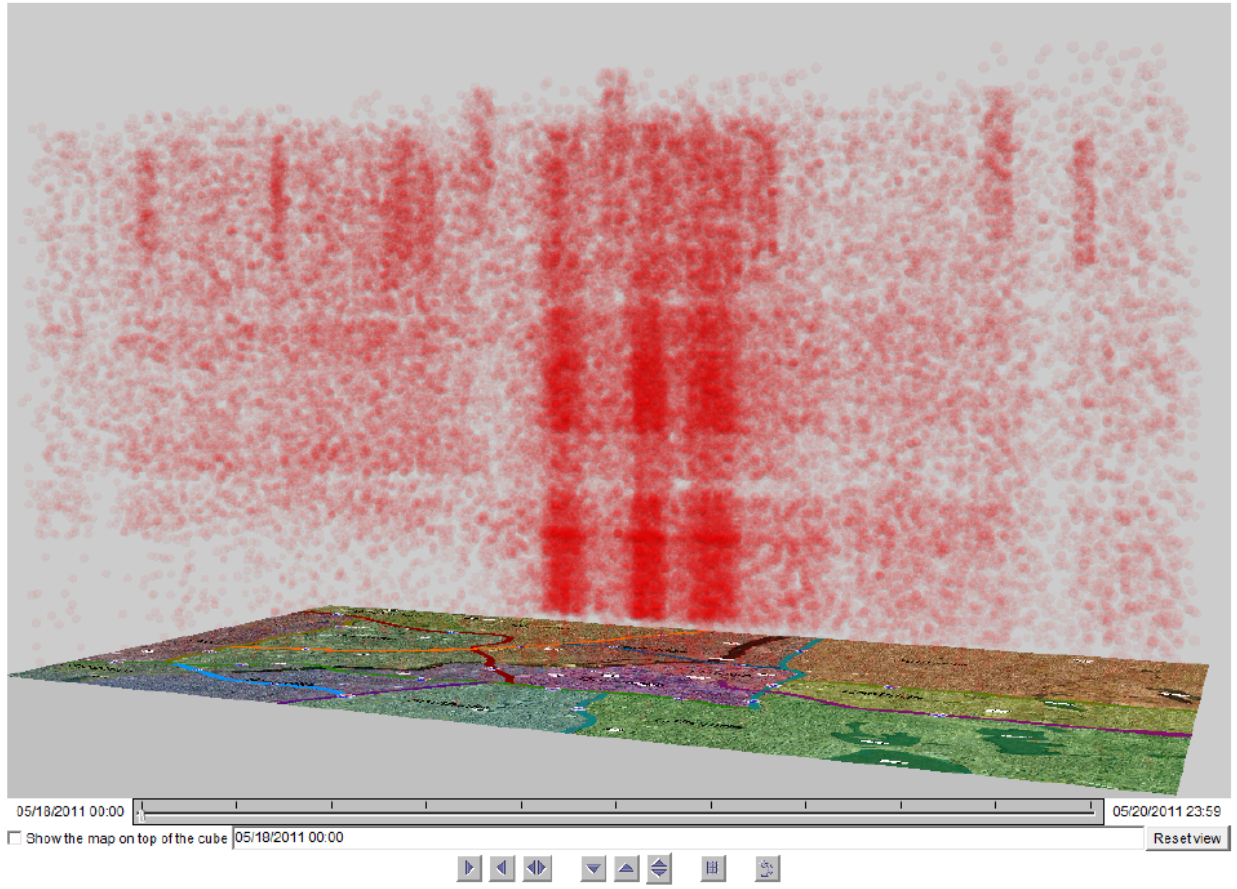
思考5：你发现了什么特征？

第四步 – 求证：你所看到的是事实吗？



思考6：关于疾病的时空分布，你有什么猜想？哪里最先爆发？

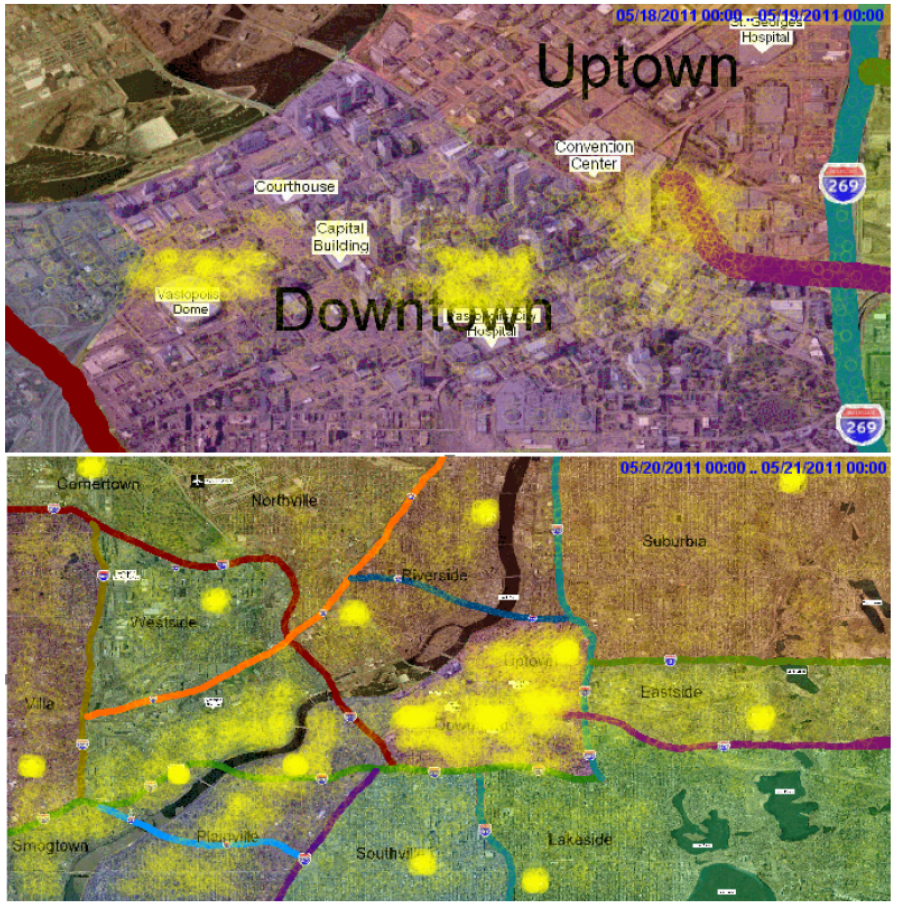
第五步 – 统观：从不同角度看问题



time

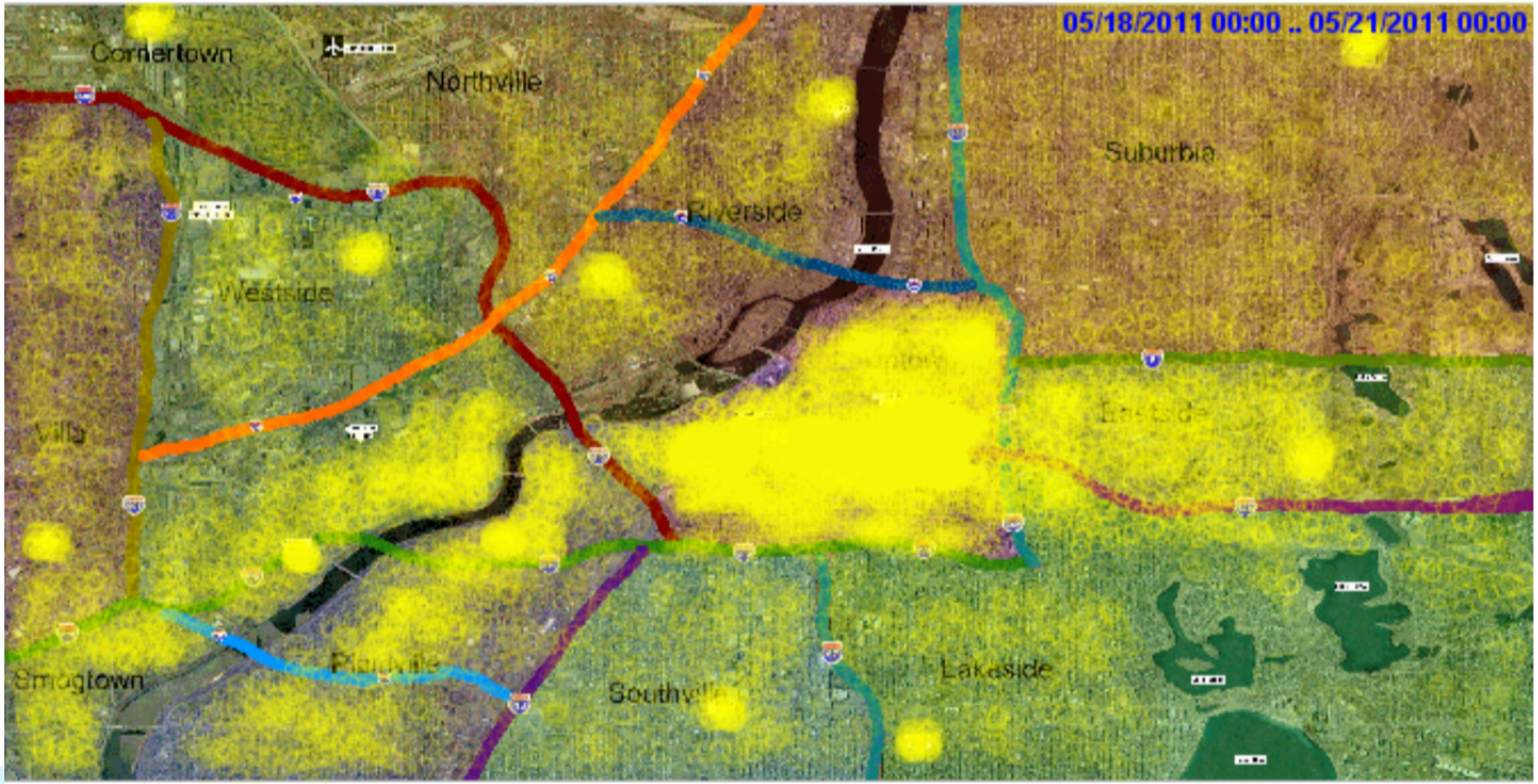
思考7：三维的视图，有什么好处，与坏处？

第五步 – 统观：从不同角度看问题（2）



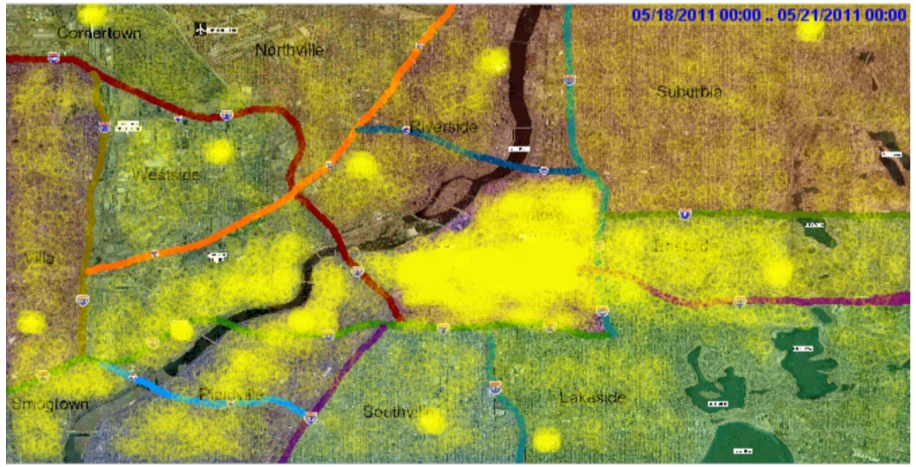
- 第一天的Downtown
- 第三天的全局

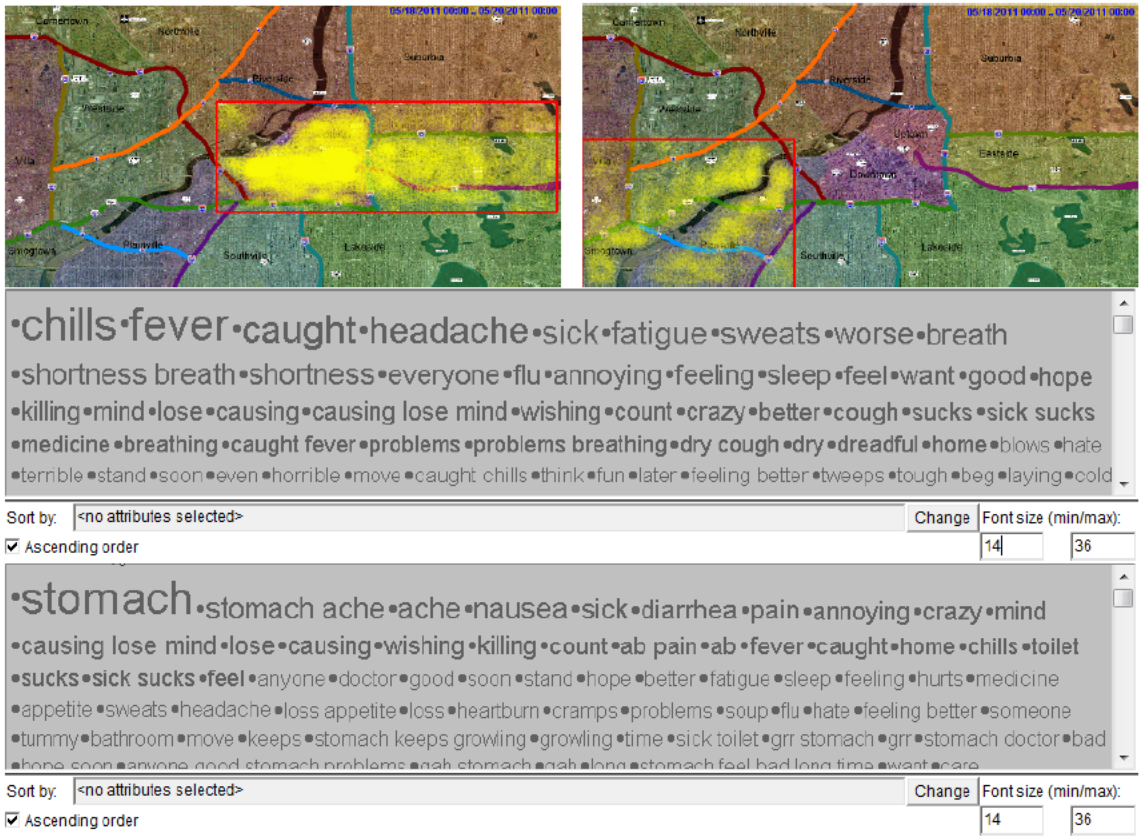
第六步－解局：条分缕析



思考8：还缺少什么？

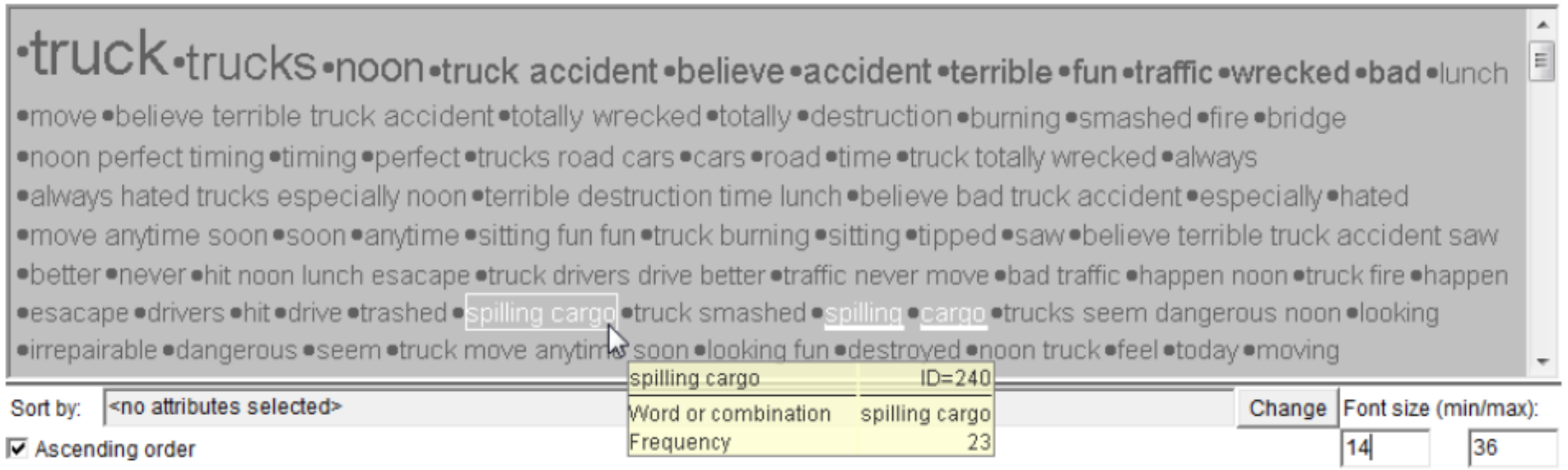
第六步 – 解局：条分缕析



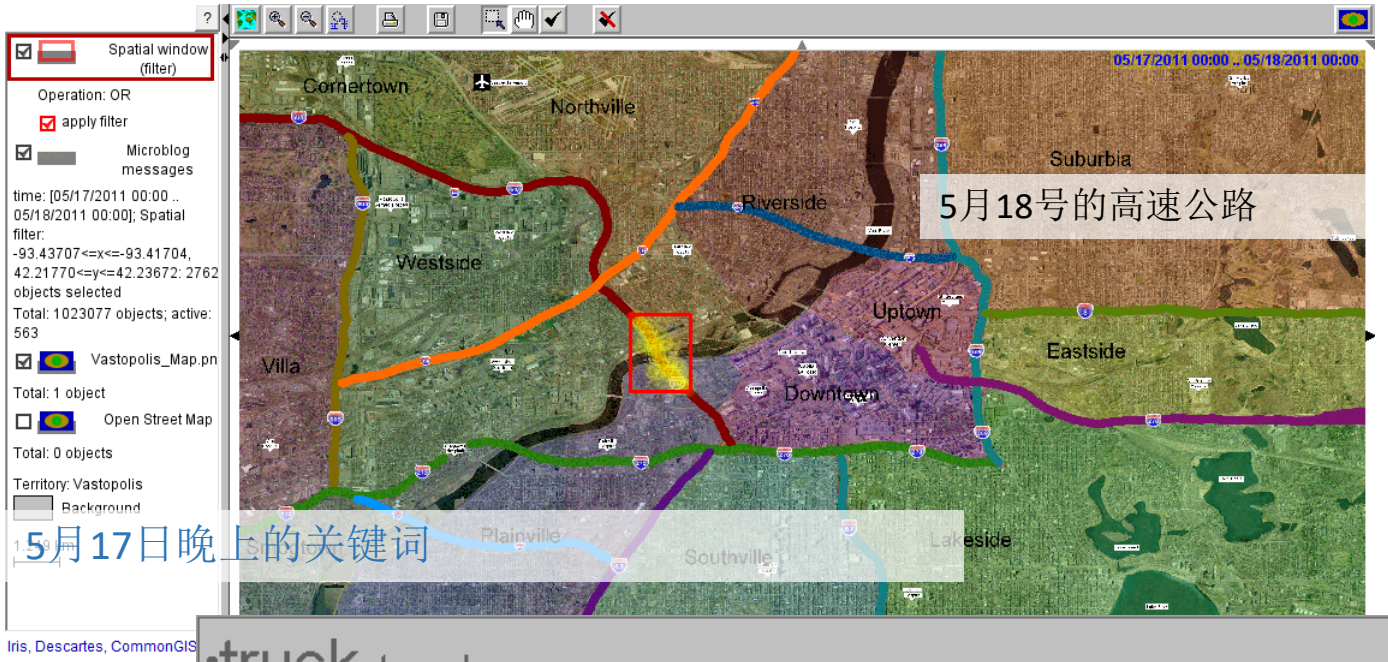


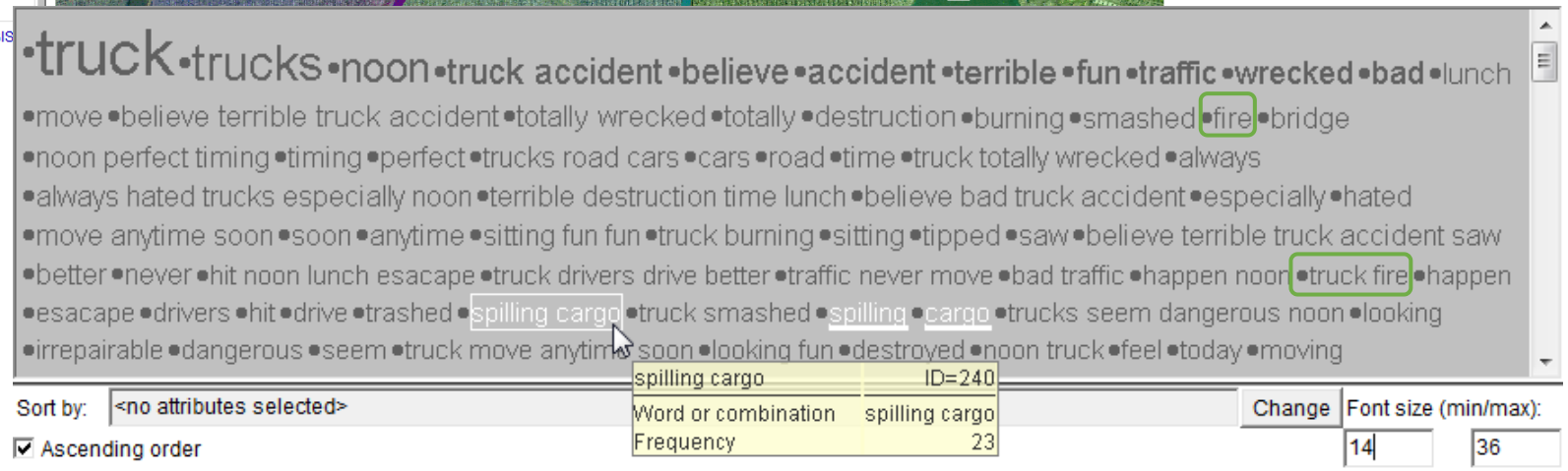
交互筛选技术

第六步 – 解局：条分缕析（2）

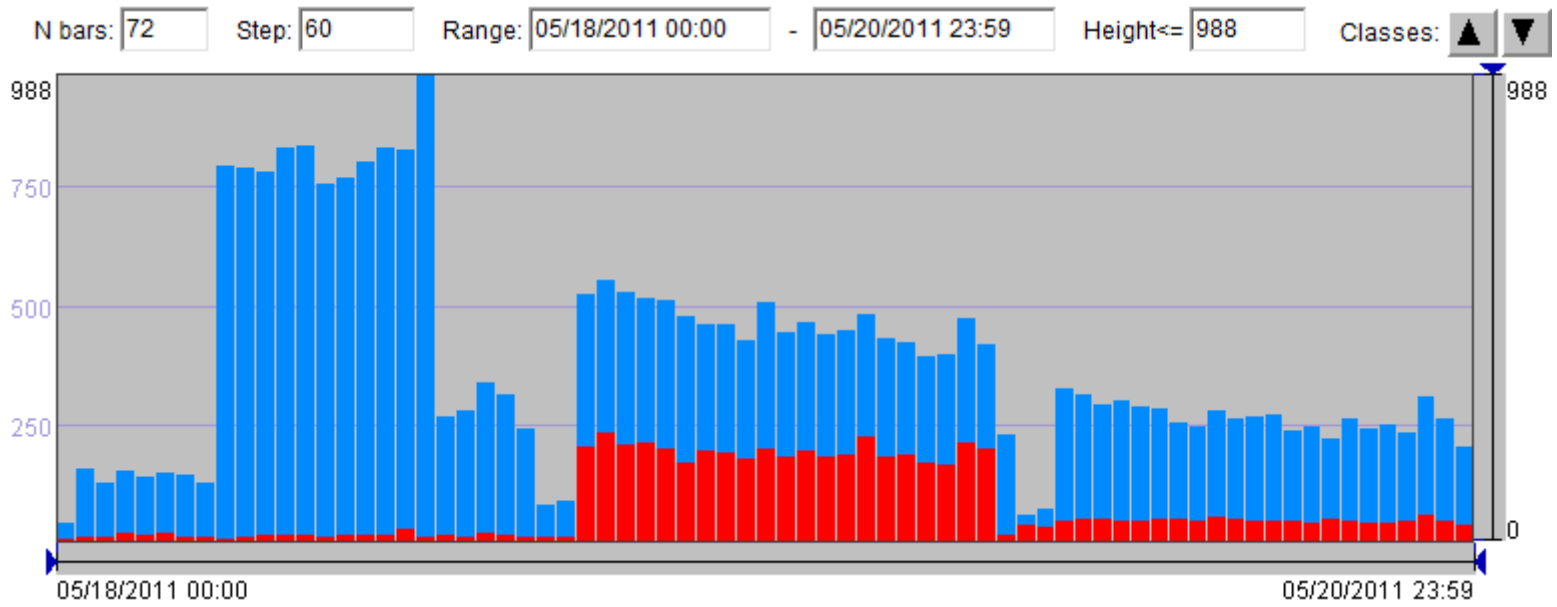

•truck •trucks •noon •truck accident •believe •accident •terrible •fun •traffic •wrecked •bad •lunch
•move •believe terrible truck accident •totally wrecked •totally •destruction •burning •smashed •fire •bridge
•noon perfect timing •timing •perfect •trucks road cars •cars •road •time •truck totally wrecked •always
•always hated trucks especially noon •terrible destruction time lunch •believe bad truck accident •especially •hated
•move anytime soon •soon •anytime •sitting fun fun •truck burning •sitting •tipped •saw •believe terrible truck accident saw
•better •never •hit noon lunch esacape •truck drivers drive better •traffic never move •bad traffic •happen noon •truck fire •happen
•esacape •drivers •hit •drive •trashed •spilling cargo •truck smashed •spilling •cargo •trucks seem dangerous noon •looking
•irrepairable •dangerous •seem •truck move anytime soon •looking fun •destroved •noon truck •feel •today •moving
spilling cargo ID=240
Word or combination spilling cargo
Frequency 23
Sort by: <no attributes selected>
Change
Font size (min/max):
Ascending order
14
36

第六步 – 解局：条分缕析（3）


Spatial window (filter)
Operation: OR
apply filter
Microblog messages
time: [05/17/2011 00:00 .. 05/18/2011 00:00]; Spatial filter: -93.43707<=x<=-93.41704, 42.21770<=y<=42.23672: 2762 objects selected
Total: 1023077 objects; active: 563
Vastopolis_Map.pn
Total: 1 object
Open Street Map
Total: 0 objects
Territory: Vastopolis
Background
Iris, Descartes, CommonGIS
05/17/2011 00:00 .. 05/18/2011 00:00
Cornertown
Northville
Suburbia
Riverside
Westside
Uptown
Eastside
Villa
Downtown
Plainville
Southville
Lakeside
5月18号的高速公路
5月17日晚上的关键词

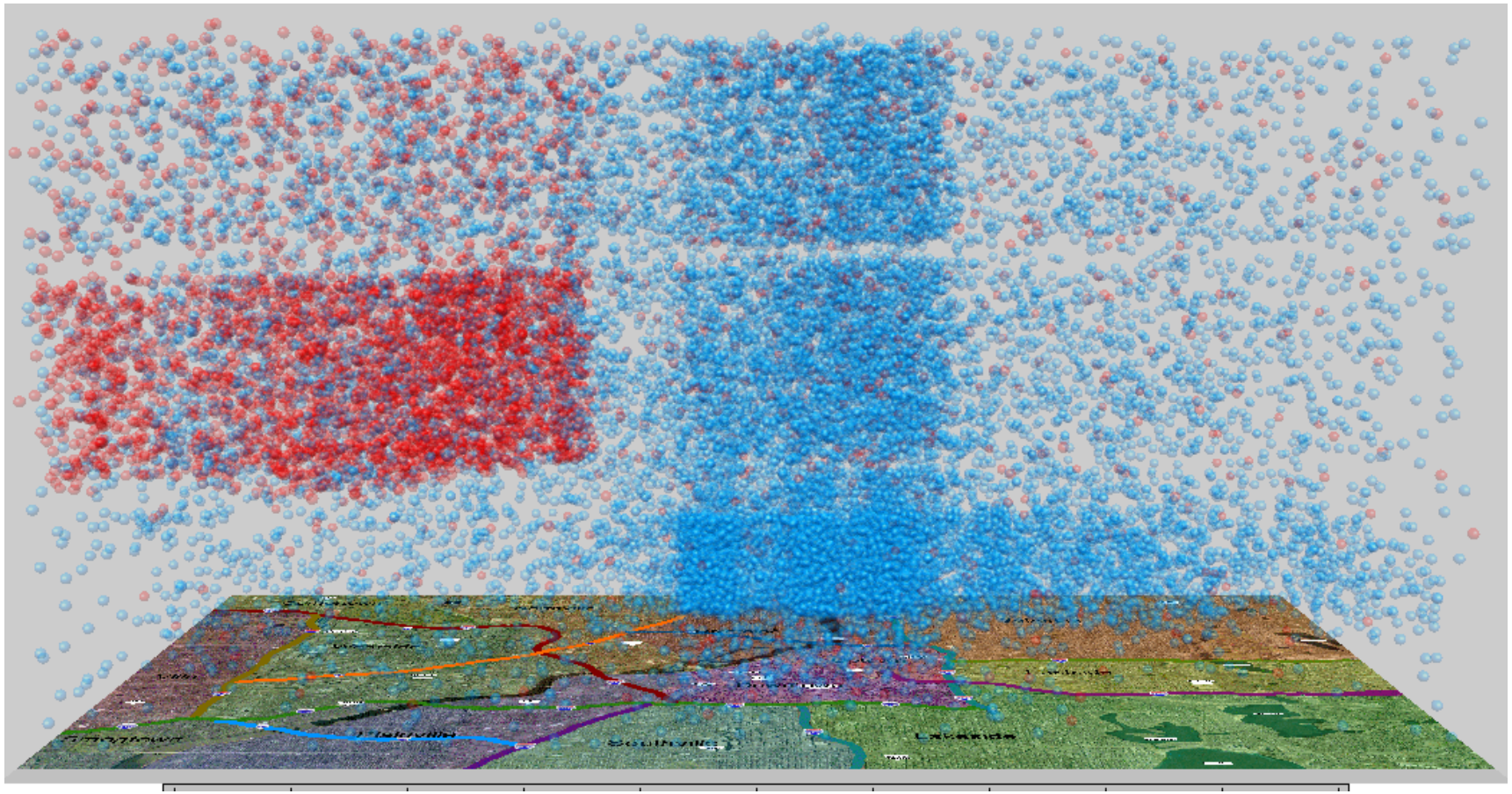

•truck •trucks •noon •truck accident •believe •accident •terrible •fun •traffic •wrecked •bad •lunch
•move •believe terrible truck accident •totally wrecked •totally •destruction •burning •smashed •fire •bridge
•noon perfect timing •timing •perfect •trucks road cars •cars •road •time •truck totally wrecked •always
•always hated trucks especially noon •terrible destruction time lunch •believe bad truck accident •especially •hated
•move anytime soon •soon •anytime •sitting fun fun •truck burning •sitting •tipped •saw •believe terrible truck accident saw
•better •never •hit noon lunch esacape •truck drivers drive better •traffic never move •bad traffic •happen noon •truck fire •happen
•esacape •drivers •hit •drive •trashed •spilling cargo •truck smashed •spilling •cargo •trucks seem dangerous noon •looking
•irrepairable •dangerous •seem •truck move anytime soon •looking fun •destroyed •noon truck •feel •today •moving
spilling cargo ID=240
Word or combination spilling cargo
Frequency 23
Sort by: <no attributes selected>
Change
Font size (min/max):
Ascending order
14
36

第六步 – 解局：条分缕析（4）



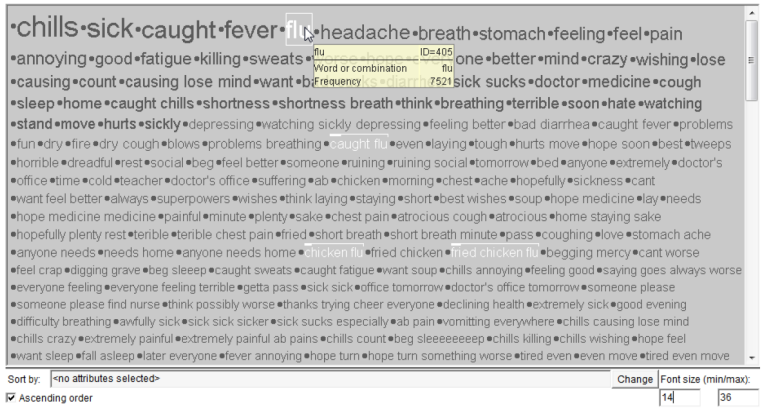
包含消化系统疾病的Twitter -> 思考8：是否人传人，疾病控制住了嘛？

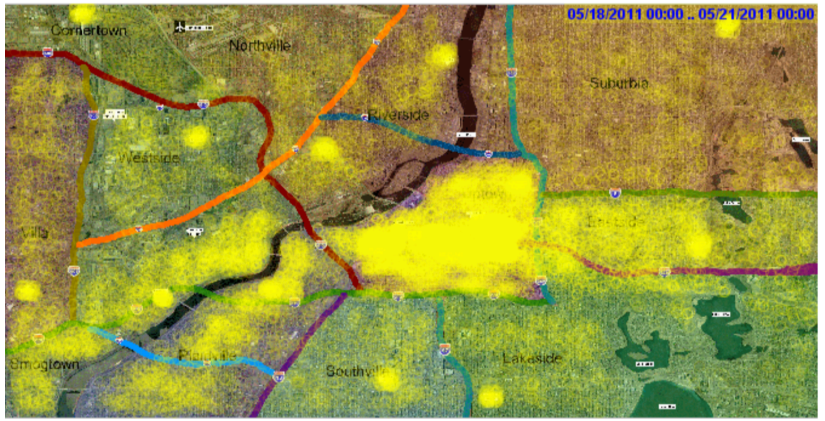
第六步－解局：条分缕析（5）

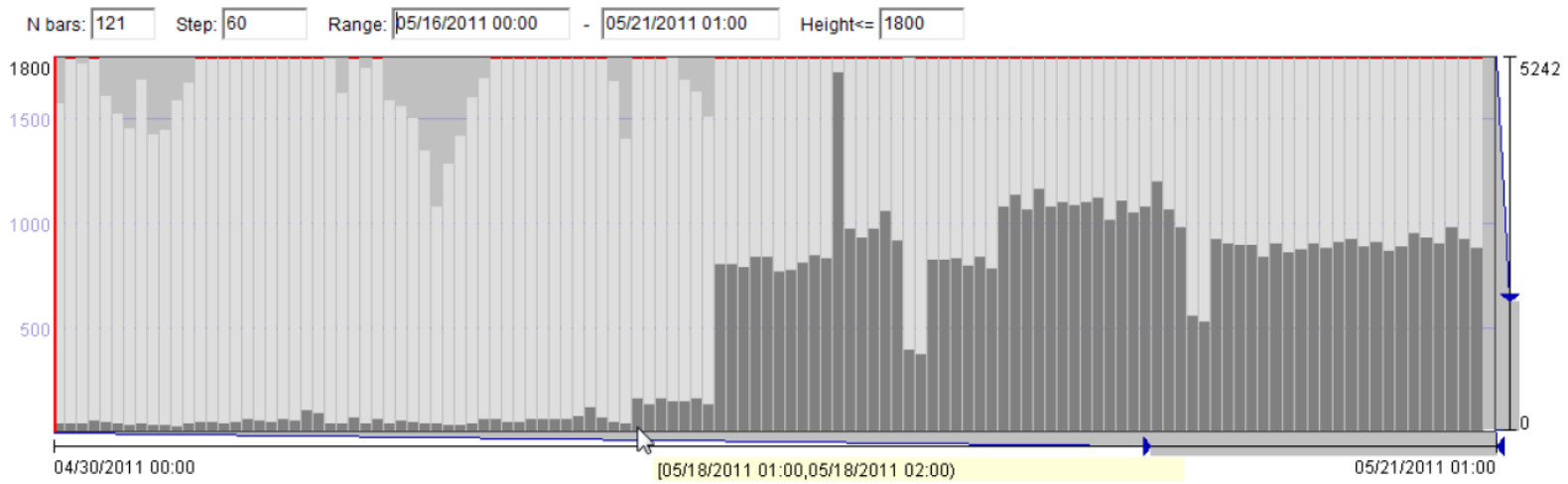


最后的思考

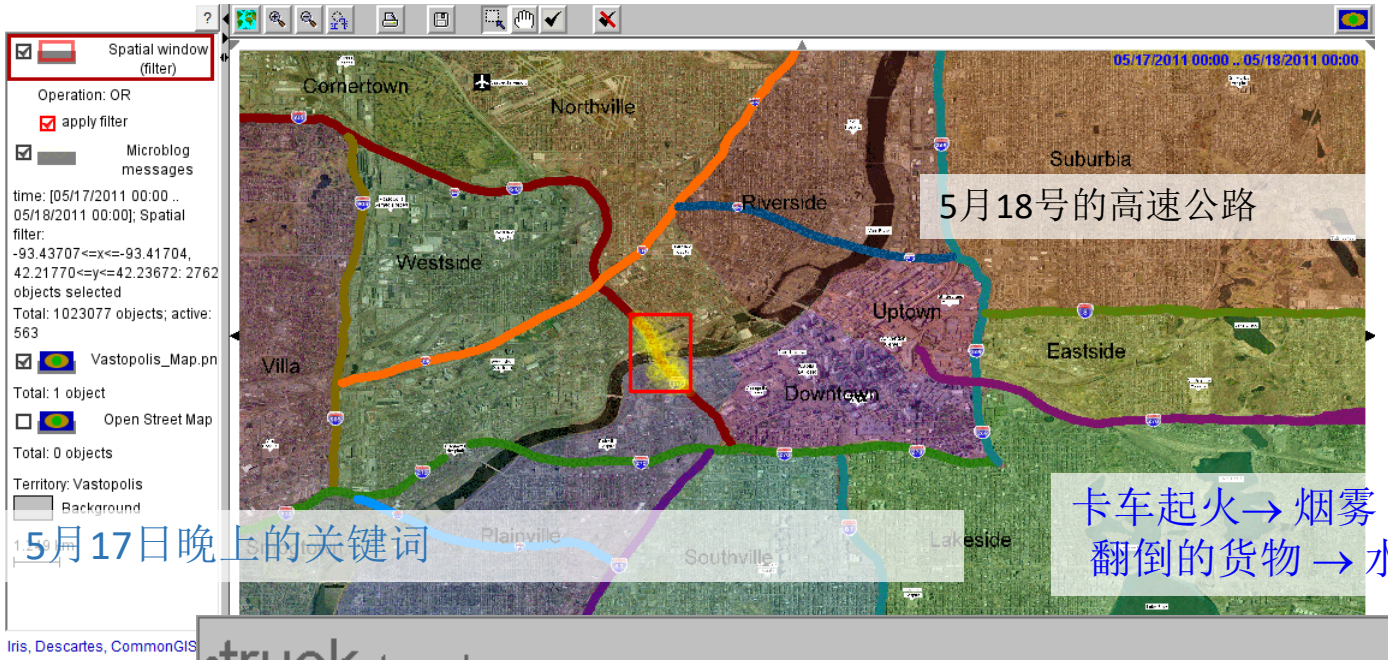
- 描述一下你探索出来的故事，把时间线和地点串起来？





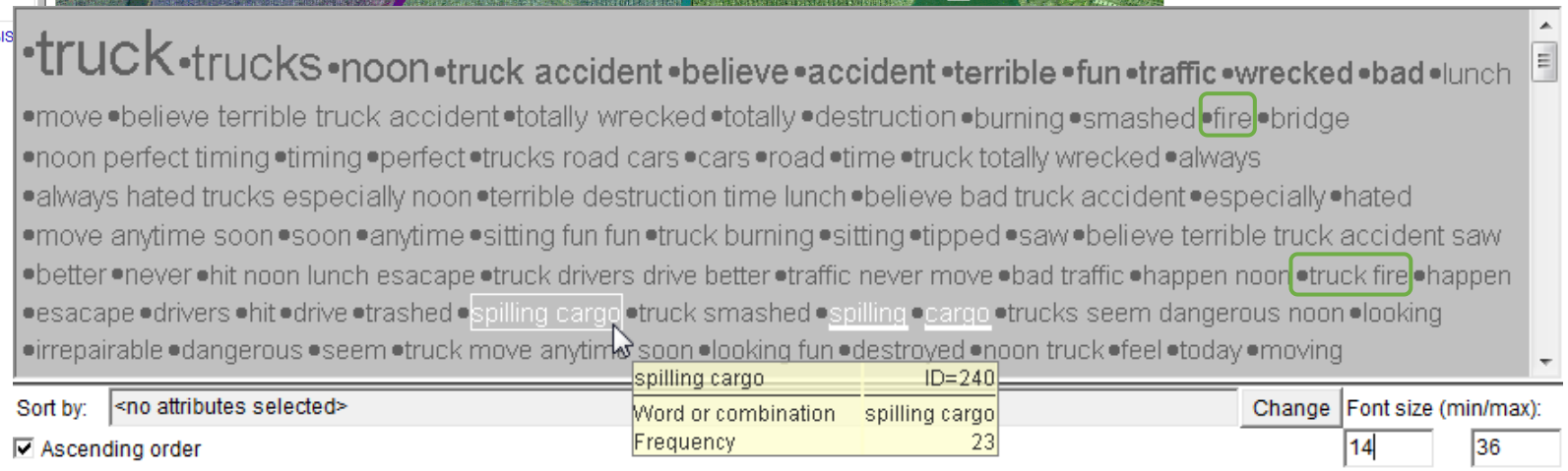


最后的思考



5月17日晚上的关键词

卡车起火→烟雾→风→呼吸系统疾病
翻倒的货物→水污染→消化系统疾病



总结

- 5月17日中午左右：市中心高速公路桥上发生一起交通事故，导致一辆载有有毒物质的卡车起火。西风将含有有毒颗粒的烟气移至城中、城东地区。受损汽车上的货物溅入河中，被河水流向西南方向移动。
- 5月18日：市中心和东部的许多人吸入有毒颗粒后生病，症状类似流感。
- 5月19日：在河岸边可能与水直接接触的人，因有毒物质溢入河中，得了胃病。
- 5月20日：得胃病的人明显减少。
- 5月19日-20日 新的流感症状病例不断出现，要求采取一些措施来清理。

回顾一下我们的分析

- 数据准备：选择潜在相关记录的子集
- 分析记录的时间分布
- 分析推文的空间分布
- 验证观察到的时空模式
- 分析和比较两个信息子集中关键词的频率分布
- 将观察到的模式与背景信息（风向、河流流量）联系起来，推理疾病传播机制。
- 根据观察到的时空模式，推测并寻找两种疾病的共同原因。

- 在每个步骤中，我们都在利用 **可视化** 进行辅助思考与决策。

所以我们为什么需要可视化与可视分析

- 人类的推理对于解决问题至关重要。
- 通过推理，人类从数据中提取有意义的信息，构建新知识。

* the entire perspective on a situation or issue
(Merriam-Webster dictionary)

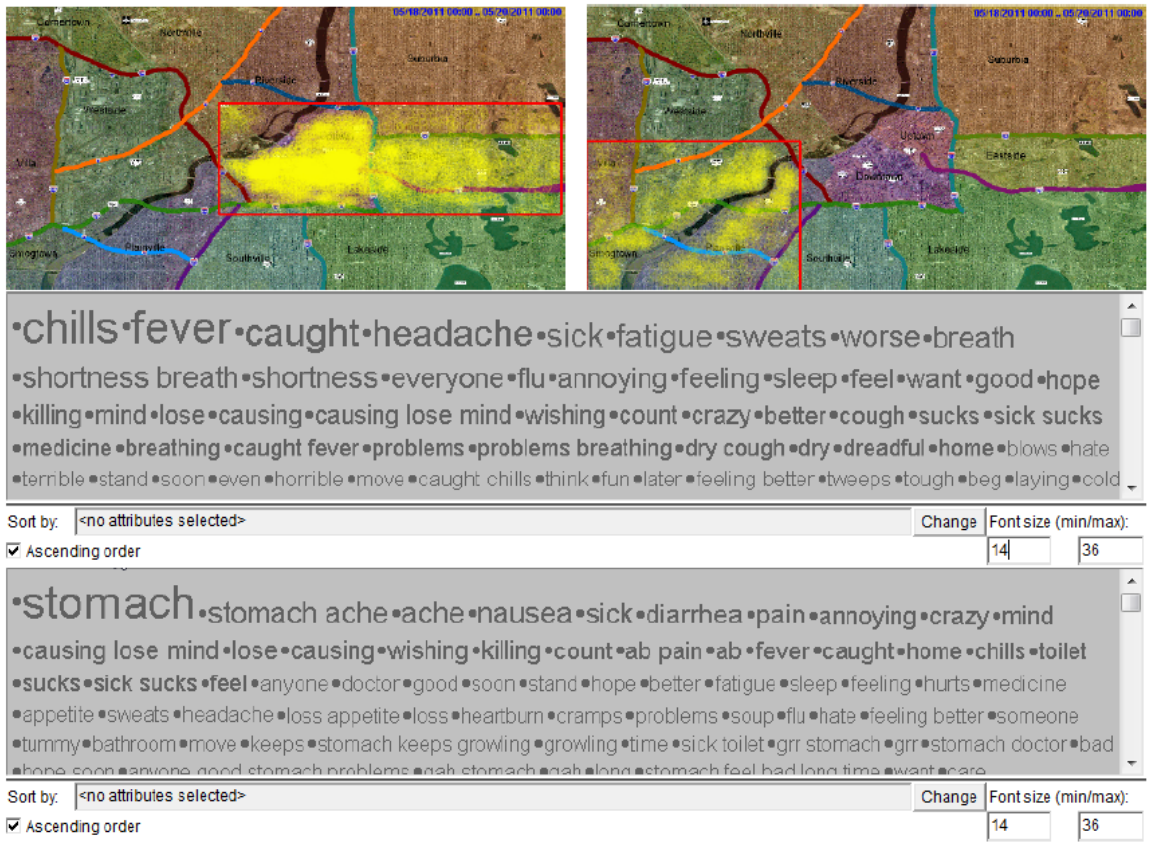
所以我们为什么需要可视化与可视分析

- 可视化表征通过有效地将信息传递给人的大脑，促进了人类的推理。人类分析人员经常通过考虑各种分布和识别模式来进行分析。
  - 在我们的例子中考虑的分布：时间分布、空间分布、时空分布。
  - 在我们的例子中发现的模式：时间趋势、空间聚类、时空聚类、时空趋势。
  - 需要对模式进行解释、验证，并与 "大局 “联系起来（可能包括一些背景知识）。

* the entire perspective on a situation or issue
(Merriam-Webster dictionary)

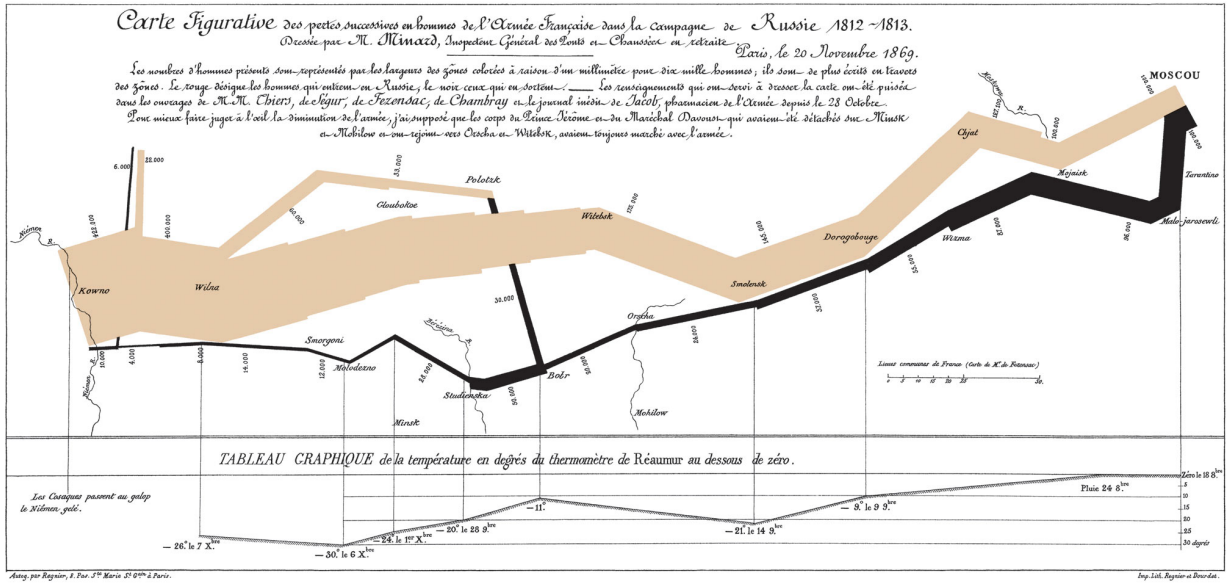
可视化

- 将数据转换为图形的学科，允许用户通过人机交互探索数据。



可视化 - 我的观点

- 表达性：一图胜千言
- 信息传递、产生知识
- 以用户为中心的设计：与领域专家、领域问题紧密结合
- 可交互性：支持迭代地探索、良好的用户反馈、自然人机界面
- 美观性：让用户有用的欲望



[Tobler et al. 1987]

呈现之外，建立假说

- 霍乱可视化
  - 空间分布探索
  - 空间关联探索
  - 建立假说

Cholera case mapping [Snow 1854]

可视分析

- 可视分析是利用交互的可视化界面进行分析推演的科学
  - 结合了自动算法
  - 新颖的交互界面
  - 视觉设计

- **可视分析的目的在于将人放在分析的重要位置中**

[Thomas et al. 2005]

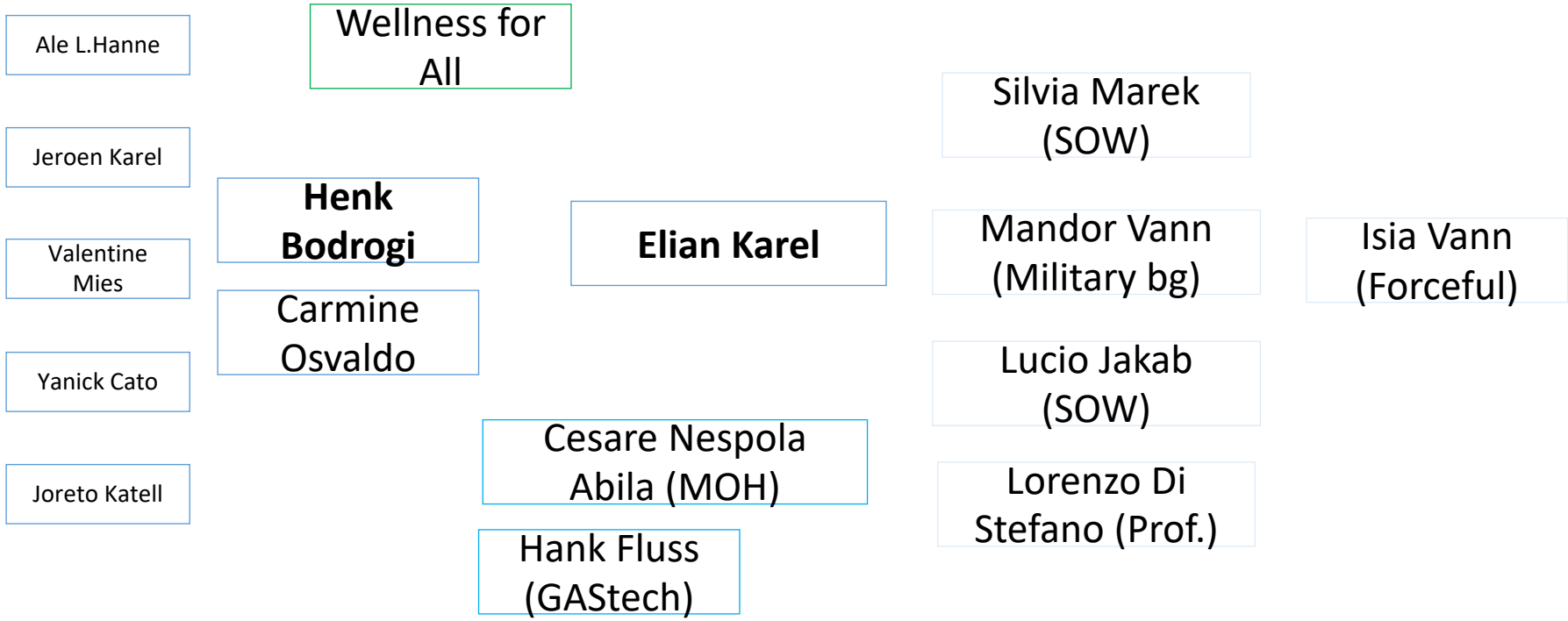
情报分析 - VAST Challenge 2014

- 背景介绍
- 分析任务与挑战
- 数据描述
- 可视分析
- 小结

背景介绍

- 一家Tethys国的跨国公司GAStech，已经在Kronos国发展了20年了。在这20年间，它主要经营天然气的开采工作，产生了可观的利润，并且它也与Kronos政府保持了良好而紧密的关系。但是，它却在环境保护上的作为却遭人诟病。
- 在2014年1月的一天，GAStech的高层举办庆祝典礼来庆祝公司成功的IPO，但在典礼的一半，突然有一些员工消失了！一个叫做Protectors of Kronos的组织被怀疑与这次失踪事件有关，但事情可能并不简单的是我们所看到的那样。
- 作为一名可视分析专家，你受雇与Kronos和Tethys国的法律部门，对整个事件进行评估，并且找出失踪的人在哪里，帮助他们重新回家。时间十分紧迫。

背景介绍- POK的起源与发展



1. Water Contamination (Elodis)
2. Forming of the Grassroots
3. Meetings with GAStech and govrn.
4. Forming an identity: POK
5. Membership Numbers Increase
6. Operations move to Abila
7. Additional issue (illegal alliance)
8. Arrested and violonce

可视分析任务与挑战

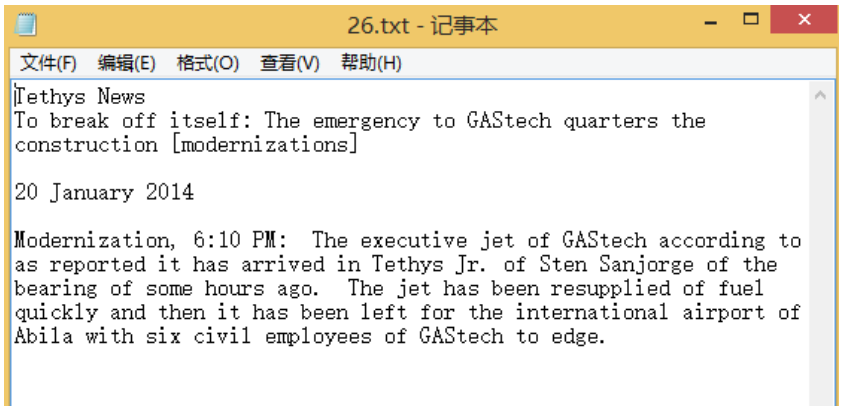
- 总体目标
  - 找出事情的真相，员工为何失踪？谁是主谋、动机为何？在合适的时间地点部署警力，救出被困人员
  - 整个1月份，各方势力如何周旋，究竟发生了什么？
- 子目标
  - 描述出POK的网络关系图，并且如何随时间变化？
    - 谁是领袖？谁是外延网络的一部分？组织架构如何变化？
    - POK与GAStech有什么潜在的关系？
  - 描述出2014年1月20-21日，发生了什么事件（时间地点人物）
  - 描述出2014年1月6日-19日，GAStech员工的日常行为，并找出哪些是异常举动？
  - 描述出2014年1月23日，发生了什么事件？（根据Streaming数据）

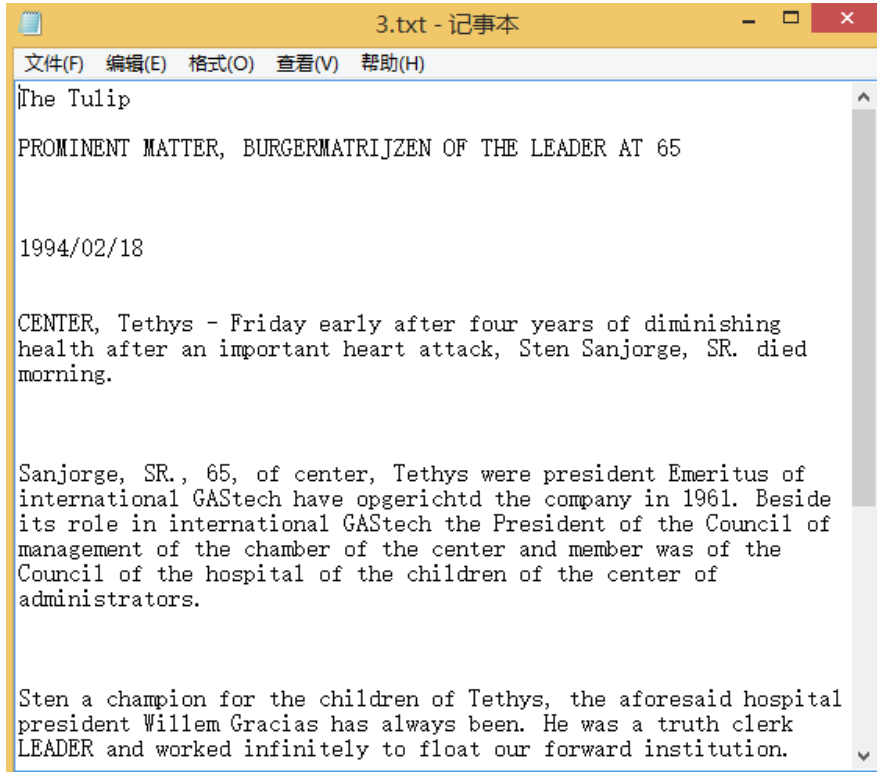
数据概览

- 1）新闻数据 – 文本（包括20年间的老新闻，以及绑架发生前后共3天的密集新闻报道）
- 2）GAStech公司数据 – 文档
  - 人物信息
    - 员工记录表
    - 部分员工简历
  - 组织架构信息
- 3）GAStech员工Email通信数据（仅标题）- 社会网络数据
- 4）GAStech员工车辆GPS轨迹数据 – 时空数据
- 5）GAStech员工刷卡账单、会员卡数据 – 日志数据
- 6）事发后相关的Twitter数据 – 社交媒体数据
- 7）消防车与警车出警调度数据 – 日志数据
- 8）背景资料
  - 国家简介、POK简介、地图、Abila城市路网数据

数据介绍（1）- 新闻数据

- 包括历史新闻与绑架事件前后的新闻（各一半左右）
  共845篇，时间跨度为20年





数据介绍（2.1）- GAStech公司数据：员工信息表

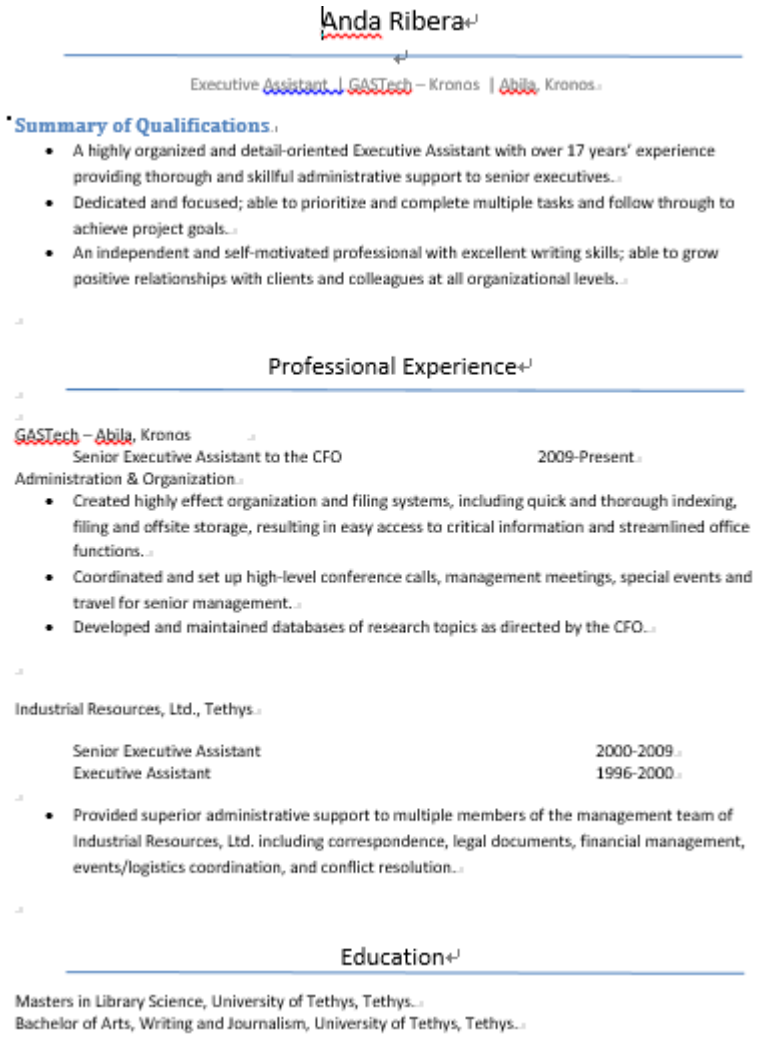
- 员工信息表
  - 基本信息：Name, birthdate, gender
  - 国籍信息：
    - Birth country, citizenship country, citizenship basis
    - Passport country, passport issue date, passport expiration date
  - 部门职位信息：
    - Current employment type (belong group of the company, e.g. administration, security, etc), current employment type, current employment start date
  - 邮件
  - 部队相关信息
    - Military Service Branch, military discharge type, military discharge date

LastName	FirstName	BirthDate	BirthCountry	Gender	CitizenshipCountry	CitizenshipBasis	CitizenshipStartDate	PassportCountry	PassportIssueDate	PassportExpirationDate
Bramar	Mat	1981/12/19	Tethys	Male	Tethys	BirthNation	1981/12/19	Tethys	2007/12/12	2017/12/11
Ribera	Anda	1975/11/17	Tethys	Female	Tethys	BirthNation	1975/11/17	Tethys	2009/6/15	2019/6/14
Pantanal	Rachel	1984/8/22	Tethys	Female	Tethys	BirthNation	1984/8/22	Tethys	2013/6/13	2023/6/12
Lagos	Linda	1980/1/26	Tethys	Female	Tethys	BirthNation	1980/1/26	Tethys	2009/11/1	2019/10/31
Mies Haber	Ruscella	1964/4/26	Kronos	Female	Kronos	BirthNation	1964/4/26			
Forluniau	Carla	1981/6/2	Kronos	Female	Kronos	BirthNation	1981/6/2			
Lais	Cornelia	1991/7/7	Kronos	Female	Kronos	BirthNation	1991/7/7			

CurrentEmploymentType	CurrentEmploymentTitle	CurrentEmploymentStartDate	EmailAddress	MilitaryServiceBranch	MilitaryDischargeType	MilitaryDischargeDate
Administration	Assistant to CEO	2005/7/1	Mat.Bramar@gastech.com.kronos			
Administration	Assistant to CFO	2009/10/30	Anda.Ribera@gastech.com.kronos			
Administration	Assistant to CIO	2013/10/1	Rachel.Pantanal@gastech.com.kronos			
Administration	Assistant to COO	2010/2/1	Linda.Lagos@gastech.com.kronos			
Administration	Assistant to Engineering Group Manager	2003/4/2	Ruscella.Mies.Haber@gastech.com.kronos	ArmedForcesOfKronos	HonorableDischarge	1984/10/1
Administration	Assistant to IT Group Manager	2005/3/7	Carla.Forluniau@gastech.com.kronos	ArmedForcesOfKronos	GeneralDischarge	2001/10/1
Administration	Assistant to Security Group Manager	2011/12/22	Cornelia.Lais@gastech.com.kronos	ArmedForcesOfKronos	GeneralDischarge	2011/10/1

数据介绍（2.2）- GAStech公司数据：部分员工简历

- 由Word格式存储，并无固定格式
- 主要包含信息
  - （职业规划）
  - 资质能力
  - 经验
    - 工作经验
    - 部队经验
  - 教育经历



Anda Ribera

Executive Assistant | GASTech – Kronos | Abila, Kronos

Summary of Qualifications

- A highly organized and detail-oriented Executive Assistant with over 17 years' experience providing thorough and skillful administrative support to senior executives.
- Dedicated and focused; able to prioritize and complete multiple tasks and follow through to achieve project goals.
- An independent and self-motivated professional with excellent writing skills; able to grow positive relationships with clients and colleagues at all organizational levels.

Professional Experience

GASTech – Abila, Kronos

Senior Executive Assistant to the CFO 2009-Present
Administration & Organization

- Created highly effect organization and filing systems, including quick and thorough indexing, filing and offsite storage, resulting in easy access to critical information and streamlined office functions.
- Coordinated and set up high-level conference calls, management meetings, special events and travel for senior management.
- Developed and maintained databases of research topics as directed by the CFO.

Industrial Resources, Ltd., Tethys

Senior Executive Assistant 2000-2009
Executive Assistant 1996-2000

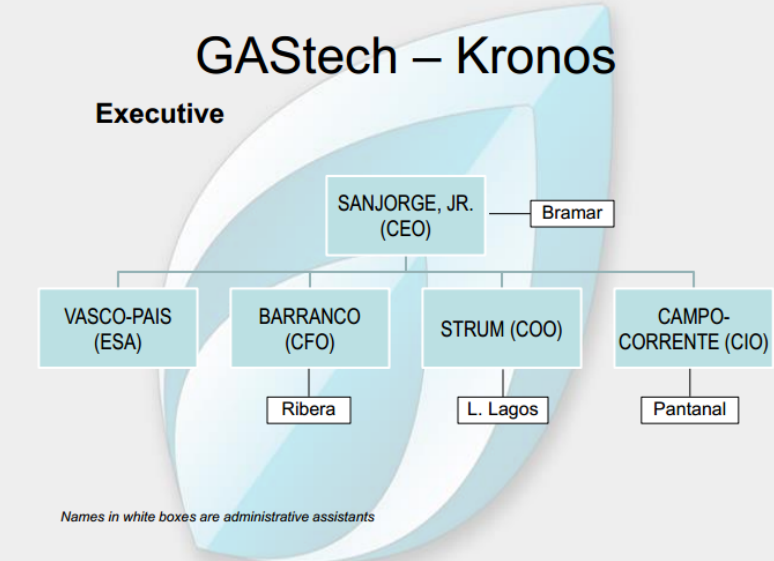
- Provided superior administrative support to multiple members of the management team of Industrial Resources, Ltd. including correspondence, legal documents, financial management, events/logistics coordination, and conflict resolution.

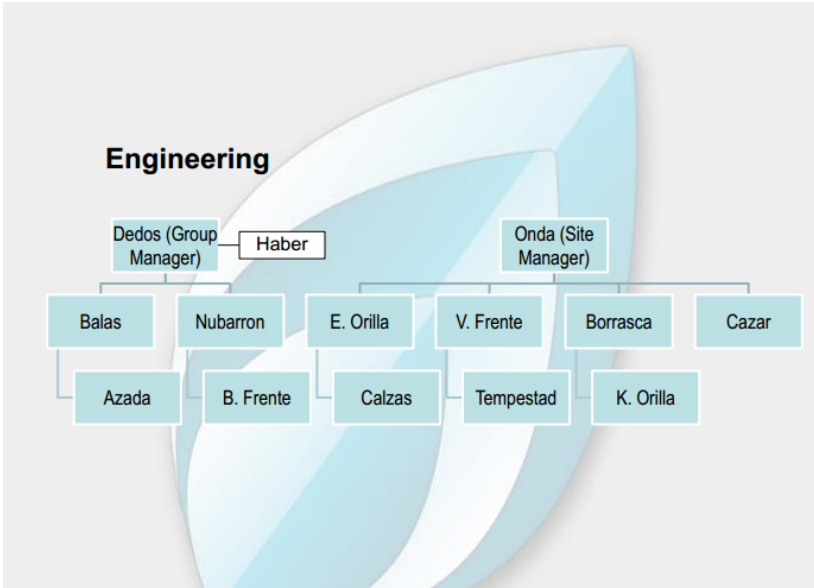
Education

Masters in Library Science, University of Tethys, Tethys.
Bachelor of Arts, Writing and Journalism, University of Tethys, Tethys.

数据介绍（2.2）- GAStech公司数据：组织架构表

- 包括以下部门
  - Executive, Engineering, Information Technology, Security, Facilities





数据介绍（3）- GAStech公司员工E-mail数据

- 邮件往来数据，共2周，1170封
- 包含
  - From
  - To
  - Time
  - Subject

From	To	Date	Subject
Sven.Flecha@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kronos	1/6/2014 8:39	GT-SeismicProcessorPro Bug Report
Kanon.Herrero@gastech.com.kronos	Felix.Resumir@gastech.com.kronos, Hideki.Cocinaro@gastech.com.kr	1/6/2014 8:58	Inspection request for site
Bertrand.Ovan@gastech.com.kronos	Emile.Arpa@gastech.com.kronos, Varro.Awelon@gastech.com.kronos,	1/6/2014 9:28	New refueling policies - Effective Februa
Valeria.Morlun@gastech.com.kronos	Dante.Coginian@gastech.com.kronos, Albina.Hafon@gastech.com.kron	1/6/2014 9:38	Route suggestion for next shift
Mat.Bramar@gastech.com.kronos	Rachel.Pantanal@gastech.com.kronos, Lars.Azada@gastech.com.krono	1/6/2014 9:49	Upcoming birthdays
Bertrand.Ovan@gastech.com.kronos	Emile.Arpa@gastech.com.kronos, Varro.Awelon@gastech.com.kronos,	1/6/2014 10:01	Don't text and drive!
Orhan.Strum@gastech.com.kronos	Stenig.Fusil@gastech.com.kronos	1/6/2014 10:25	Service anniversary
Varja.Lagos@gastech.com.kronos	Varja.Lagos@gastech.com.kronos, Hennie.Osvaldo@gastech.com.krono	1/6/2014 10:28	Patrol schedule changes
Brand.Tempestad@gastech.com.kronos	Birgitta.Frente@gastech.com.kronos, Lars.Azada@gastech.com.krono	1/6/2014 10:35	Wellhead flow rate data
Isak.Baza@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kronos	1/6/2014 10:43	RE: GT-SeismicProcessorPro Bug Report
Lucas.Alcazar@gastech.com.kronos	Isak.Baza@gastech.com.kronos, Lucas.Alcazar@gastech.com.kronos	1/6/2014 10:50	RE: GT-SeismicProcessorPro Bug Report
Linnea.Bergen@gastech.com.kronos	Rachel.Pantanal@gastech.com.kronos, Lars.Azada@gastech.com.krono	1/6/2014 11:00	RE: Upcoming birthdays

数据介绍（4）- GAStech公司员工车辆轨迹数据

- GAStech为员工配备车辆，但却又不相信员工，因此安装了GPS秘密记录了员工轨迹，地图为Abila市
- Sample Data:

```
Timestamp,id,lat,long
01/06/2014 06:28:01,35,36.0762253,24.87468932
01/06/2014 06:28:01,35,36.07622006,24.87459598
01/06/2014 06:28:03,35,36.07621062,24.87444293
01/06/2014 06:28:05,35,36.0762167,24.87425333
01/06/2014 06:28:06,35,36.0762137,24.87416677
```

- 数据信息
  - 1-3秒采样
  - 共685,170条数据，33MB

数据介绍（5）- GAStech公司员工刷卡账单与会员卡数据

- 包含何人、何时、何地、刷卡金额等数据，共1491条，跨度2周

```
timestamp,location,price,FirstName,LastName
1/6/2014 7:28,Brew've Been Served,11.34,Edvard,Vann
1/6/2014 7:34,Hallowed Grounds,52.22,Hideki,Cocinaro
1/6/2014 7:35,Brew've Been Served,8.33,Stenig,Fusil
1/6/2014 7:36,Hallowed Grounds,16.72,Birgitta,Frente
1/6/2014 7:37,Brew've Been Served,4.24,Sven,Flecha
1/6/2014 7:38,Brew've Been Served,4.17,Cornelia,Lais
1/6/2014 7:42,Coffee Cameleon,28.73,Linnea,Bergen
1/6/2014 7:43,Brew've Been Served,9.6,Mat,Bramar
```

数据介绍（6/7）- Twitter与出警日志 – 流式数据

- 数据流，混合了Twitter与火警、警察局出警日志
- Twitter：

type,date(yyyyMMddHHmmss),author,message,latitude,longitude, location
mbdata,20140123170000,POK,Follow us @POK-Kronos,,,
mbdata,20140123170000,maha_Homeland,Don't miss a moment! Follow our live coverage of the POK Rally in the Park!,,,
mbdata,20140123170000,Viktor-E,Come join us in the Park! Music tonight at Abila City Park!,,,
mbdata,20140123170000,KronosStar,POK rally to start in Abila City Park. POK leader Sylvia Marek to open with a speech.˜ #KronosStar,,,

- 出警日志：

ccdata,20140123170000,,KEEP THE PEACE-CROWD CONTROL/ABILA CITY PARK,,,Egeou St / Parla St

数据介绍（8）- 背景介绍

- 基本信息
  - 国家地图
  - 国家历史报告
    - Background
    - Geography
    - **Environment**
    - People and Society
    - Economy
    - Military
    - Transnational Issues

可视分析

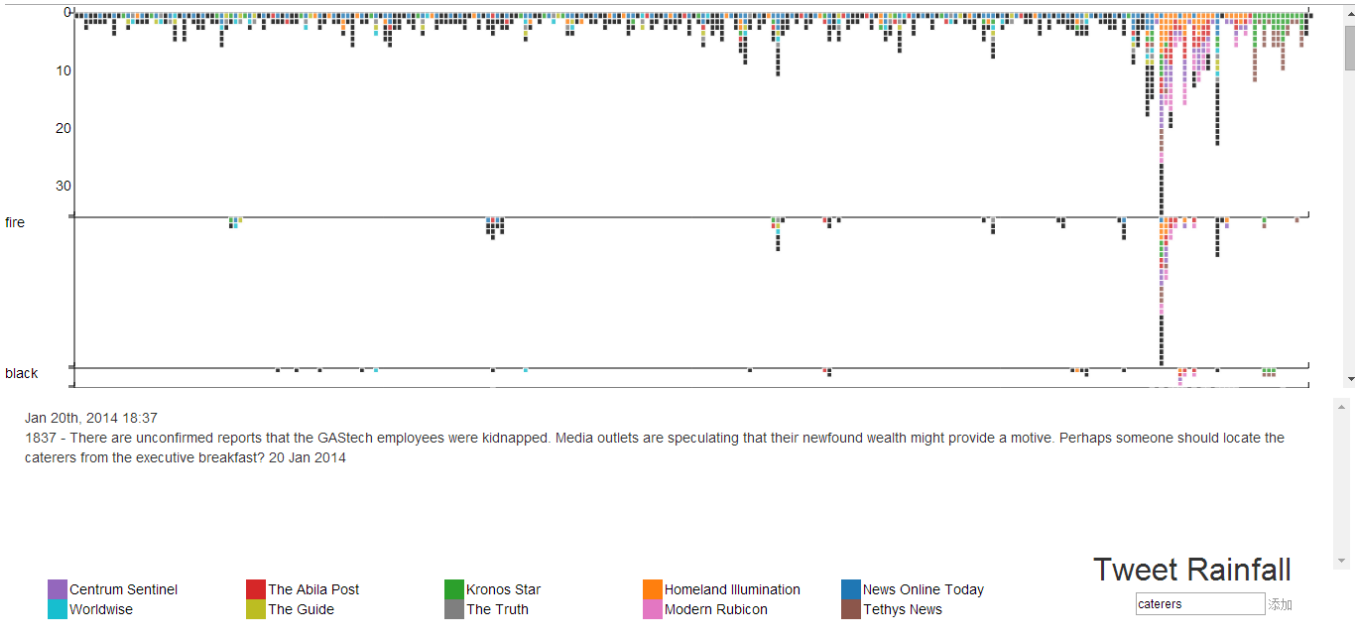
- 挑战
  - 多个数据源，不同类型数据
  - 时间跨度大（20年），粒度不同（在2014年1月有详细数据）
  - 数据中含有不确定性、数据缺失
  - 涉及分析种类较多、包括人物之间关系（社交网络）、轨迹刷卡（时空数据）、新闻文本（文本数据）、社交网络（实时流数据）等各类数据分析
  - 故事情结扑朔迷离
- 如何应对
  - 设计基本可视化方法 – 看懂数据
  - 深入可视分析交互 – 发现基本规律
  - 融合不同源数据 – 深入挖掘探索

可视分析

- 新闻文本数据可视分析
  - 对人物关系与事件脉络的探索
  - 把握数据时变规律
- 轨迹、账单数据可视分析
  - 时间、空间、人物行为可视化
  - 时空筛选与过滤
  - 不确定性、数据冲突处理
- Twitter、出警记录可视分析
  - 时变、话题演变规律
  - Streaming数据处理与分发

多数据融合可视分析
  - 以时间为纽带、多尺度分析

新闻文本数据可视分析（1）



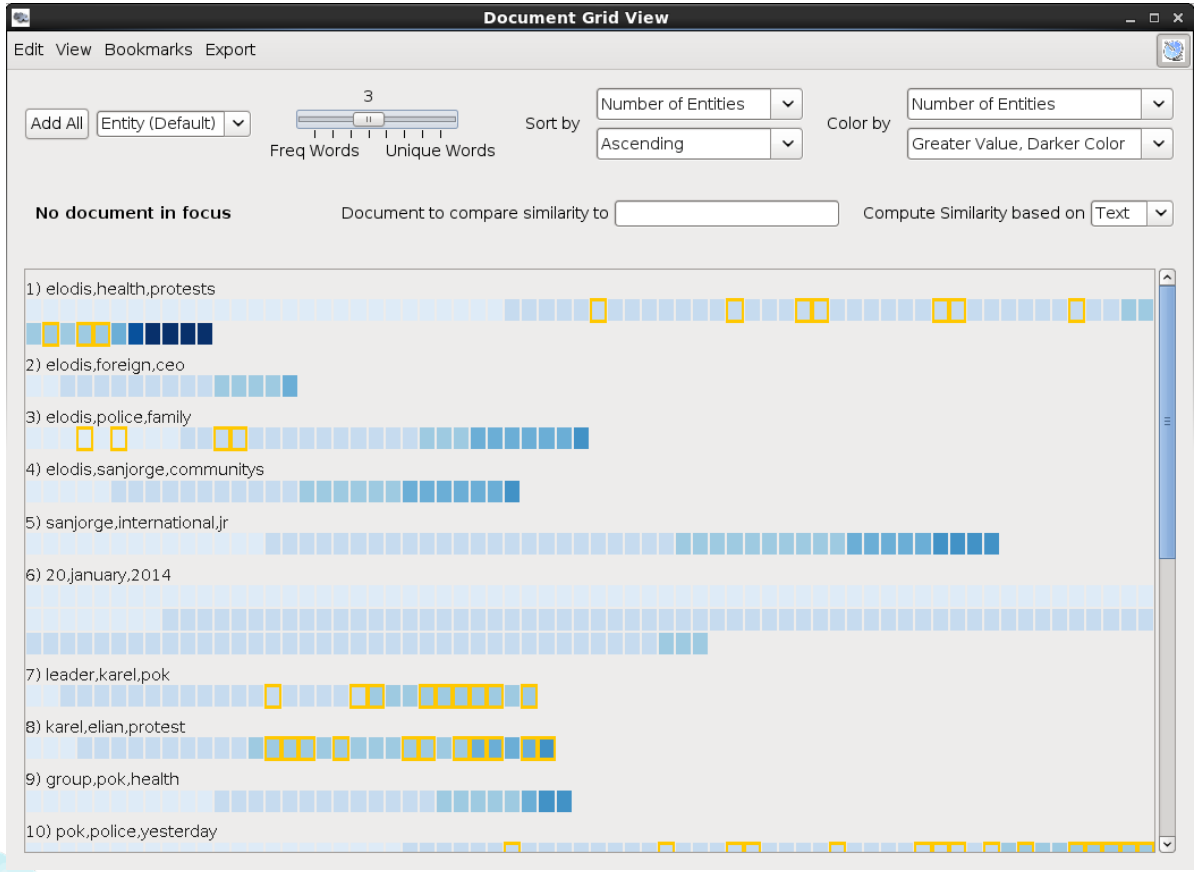
- 时变数据分析
  - 步骤1：将新闻数据预处理，提取出时间信息
  - 步骤2：将坐标轴变形，把历史数据与当今新闻（详细）以不同粒度表示，一个以天为单位，一个以小时为单位
  - 步骤3：用户可以点击新闻，查看具体文本，并且提取出感兴趣的实体（时间、地点、人物、关键词），输入搜索框，可以产生包含该搜索实体的新闻时间轴
  - 步骤4：重复步骤3可获得多个时间轴，并且鼠标点击可看到哪些关键词在同一篇文章中出现，右下角的图可视化也表示了两两间的关系强弱（用边的粗细表示）与搜索词出现频率（用点的大小表示）

- 发现
  - 1月20日新闻最为密集，1月19日和21日也较多
  - 通过阅读新闻，并且对其中的关键词进行搜索，可以辅助事件的判断
    - 搜索“Fire”
    - 搜索“Black”
  - 看出在1月20日上午发生了火警警报，但这是误报
  - 又可以看到黑衣人（people in black）被多次提到，有员工说这些人是酒会的服务人员
  - 通过以上探索可以得到相关事件发展情况

获得1月20日绑架经过

- 早上
  - GAStech高层与Kronos高层宴会与IPO庆祝典礼
  - 火警警报响起 – 消防车来 – 发现是假警报
- 中午、下午
  - 混乱中发现一些人员失踪，警方介入
  - 记者联系高层均无法接通电话，或者被秘书拒绝
  - 同时有两架神秘飞机飞离Abila机场
- 傍晚
  - GAStech陆续离开，6点左右最后一名员工被警方允许回家。
  - 警方召开新闻发布会 – 14名人员失踪
- 第二天
  - 警方改口，10人失踪
  - 记者报道GAStech CEO已经成功回国（Tethys）

新闻文本数据可视分析（2）-新闻事件与人物关系探索Jigsaw(1)



文本经过topic model生成的topic分类

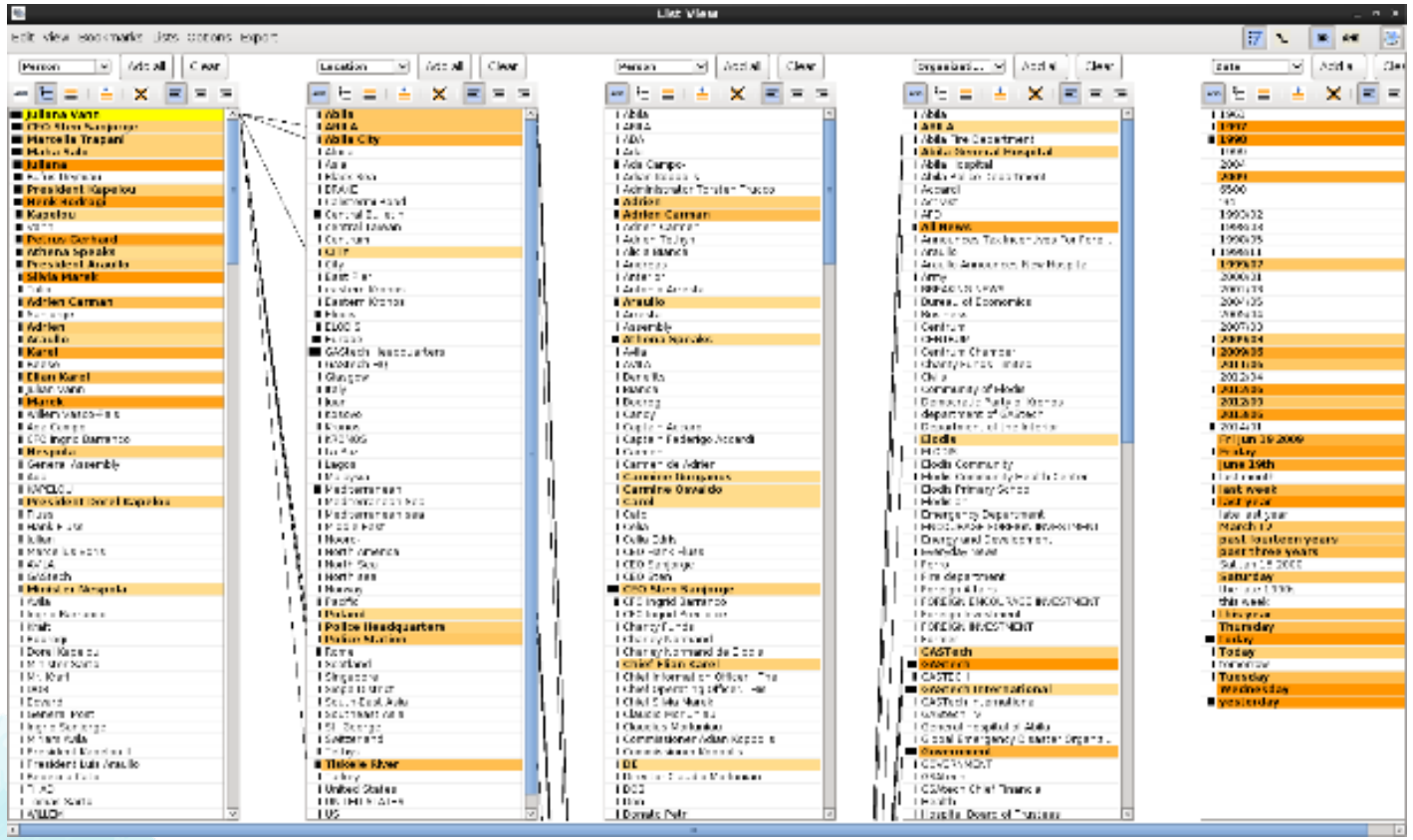
- 发现1：
  - 对文本进行topic model处理，划分出不同topic，
  - 我们可以找出一些有意义的分类
    - 健康
    - POK与逮捕
    - 天然气、税收
    - 毒品、棉花、天然气
    - 改革开放、前总统

新闻文本数据可视分析（2）-新闻事件与人物关系探索Jigsaw(2)

毒品分类

- 发现2：
  - 对于特定的topic，可以点击阅读原始文章；在文章中，人物、时间、地点、金钱等一些关键信息会被高亮
  - 我们发现在毒品topic下，一系列新闻报道了有一个叫做MDMC的技术，产生毒品在POK扩散开来，并且医院方面住院病人突然增加，许多人患有精神疾病，怀疑可能与该毒品有关
  - 点击高亮的词汇，也可以看到它们在不同topic之间的分布

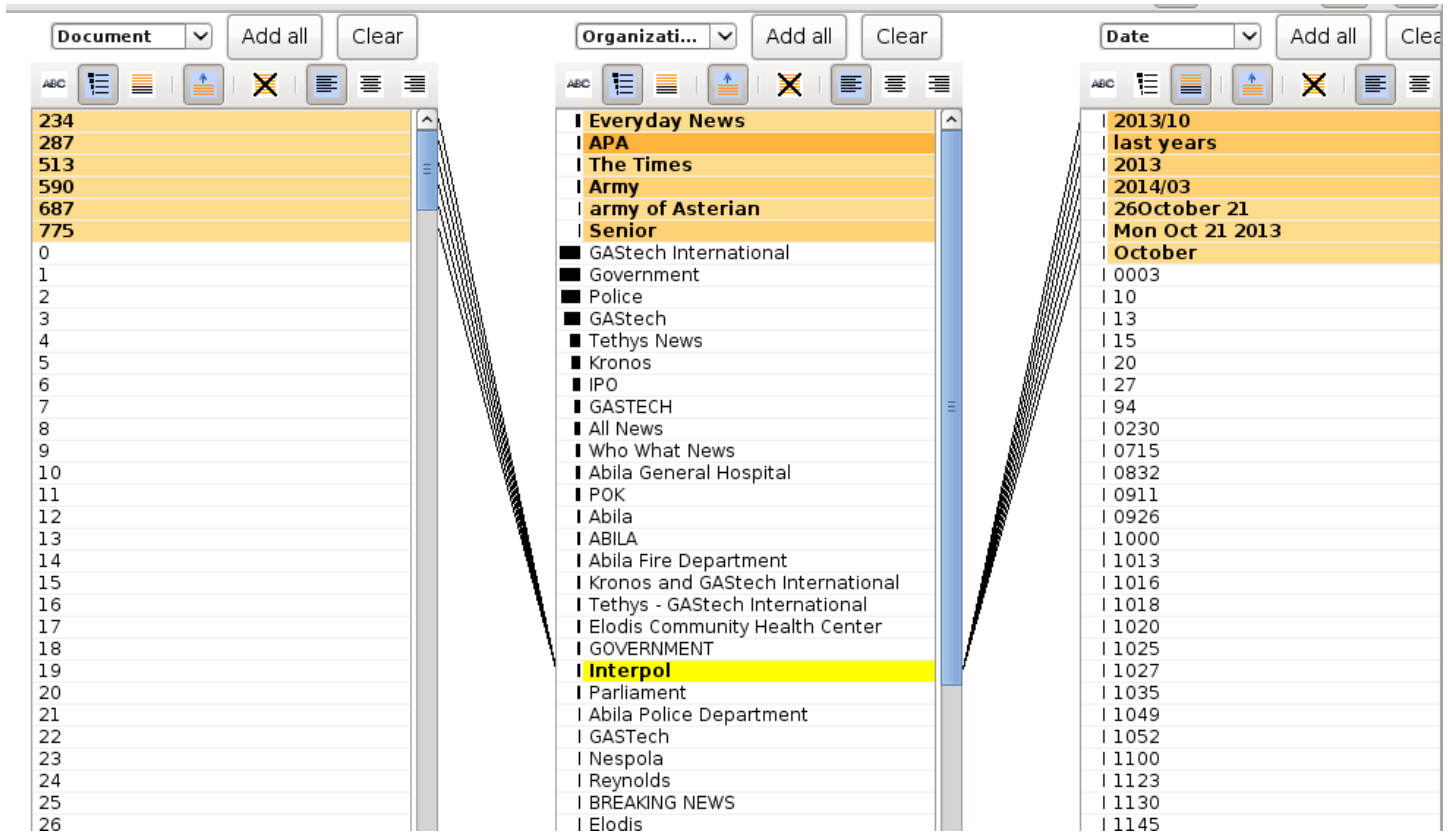
新闻文本数据可视分析（2）-新闻事件与人物关系探索Jigsaw(3)



ListView in Jigsaw

- 通过ListView探索
  - 对于不同的人物、地点、组织或者时间，点击它，与它相关（出现在同一篇文档）的条目会高亮，颜色表示相关程度
  - 可以多选，并且通过（and、ro）来寻找复杂相关关系
  - 找到后可以对应到文档视图中进行观察

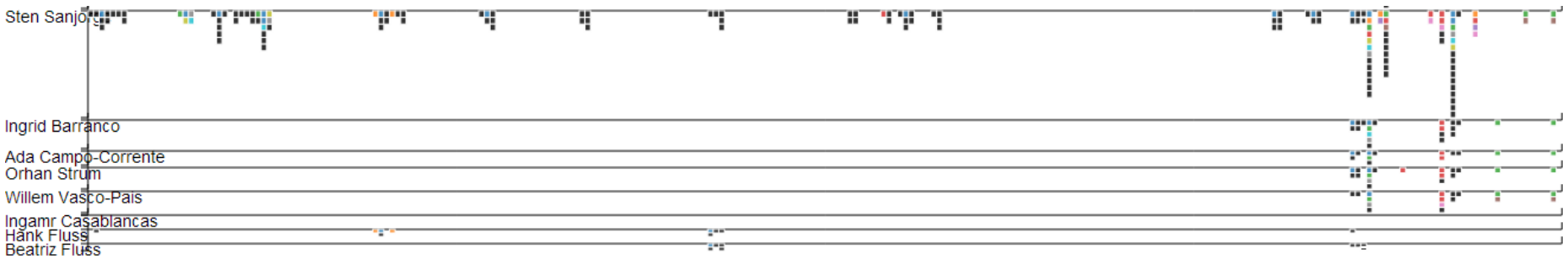
新闻文本数据可视分析（2）-新闻事件与人物关系探索Jigsaw(4)



通过Kronos市的病毒蔓延，关联到MDMC病毒技术，并关联找到国际刑警组织相关的公告，了解到新的组织APA的相关情况

- 发现4
  - 通过之前的对毒品话题的探索，我们发现国际刑警组织（Interpol）也介入其中
  - 于是我们对该组织进行关联探索，找到一些相关的文档（左），相关的组织（中）与发生事件的时间（右）
  - 通过Interpol相关的新闻，我们了解到另一个组织APA，是一个国际的贩毒团伙，受到国际刑警组织的稽查。

猜想情景（GAStech高层人员）



1）GAStech高层

人名	官职	事件
Sten Sanjorge Jr	CEO	2013年计划建设更多的国际拓展计划. 计划2013年末公司上市 妻子：Ingrid Sanjorge
Sten Sanjorge	前CEO	GAStech的创始人 1994年die
Ingrid Barranco	CFO	绑架事件中失踪
Ada Campo-Corrente	CIO	绑架事件中失踪
Orhan Strum	COO	绑架事件中失踪
Willem Vasco-Pais	Environmental Officer	绑架事件中失踪
Ingamr Casablancas	前任 Environmental Officer	
Hank Fluss	前任高层	创始人之一，死于2003年的心脏病。他在GAStech工作了40年，对于整个公司有着重要的贡献。
Beatriz Fluss	前任高层之妻	50-60岁 year old wife，拥有33%的股权

- 我们找到GAStech相关的人员，以及他们的相关关系
- 通过同样的方法我们也可以找到POK以及政府组织人员的分布

猜想情景（主要人物）

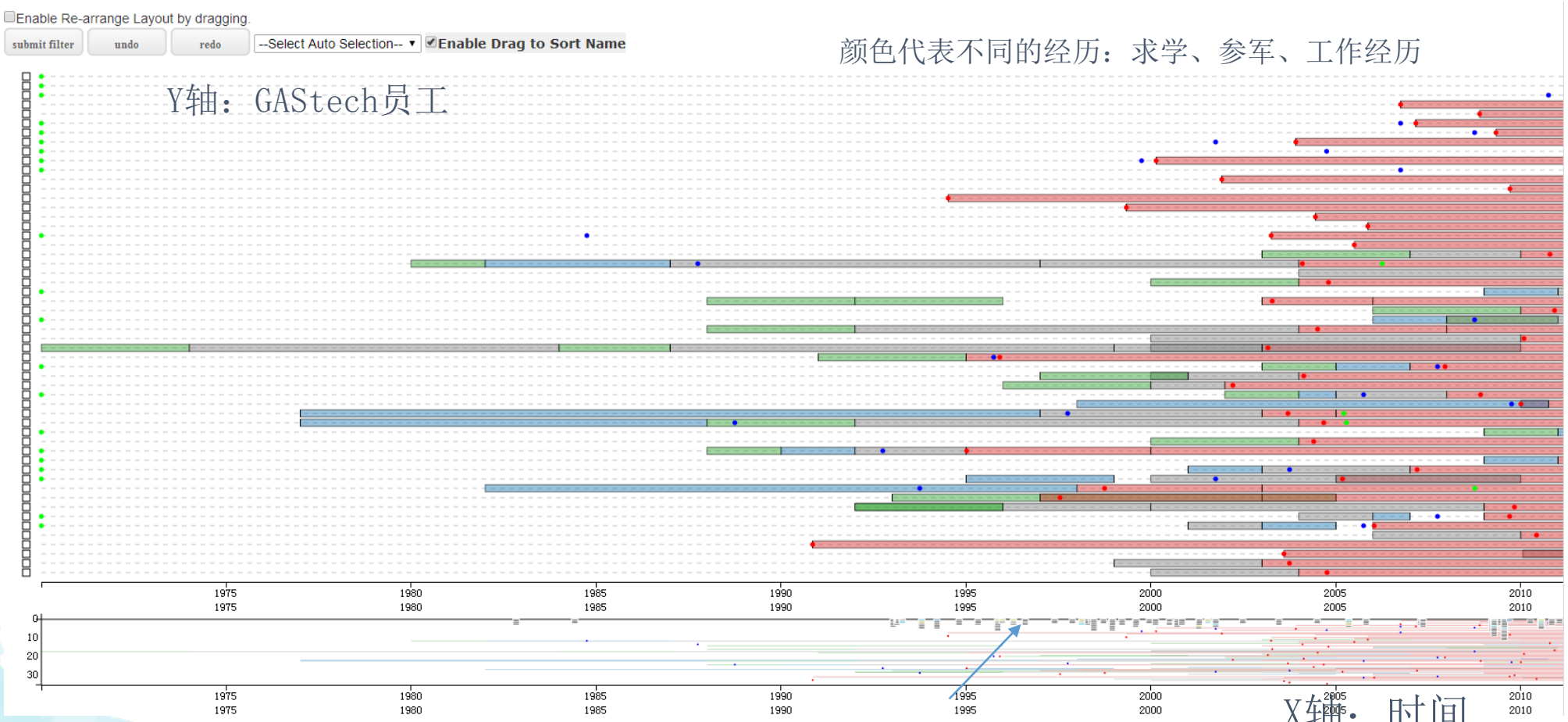
人名	官职	事件
Henk Bodrogi	第一任主席	“whose ill health forced him down from the position”
Elian Karel	第二任主席	1）March 12 2009入狱 2）his untimely death in 2009年 6月19日.（28岁）。Its death, thought with being a murder by 第二任主席带领下POK的目标： “All we want is to know how our taxes are being used. All we want is accountability for those who died because of greed, corruption and disregard of those who they should be protecting.“
Michale Kraft	Attorney of Karel	
Silvia Marek	第三任主席	21日发表第一篇演讲，被质疑领导力不强

人名	官职	时间	事件
Cesare Nespola	Minister of Health		1. announced that he would sponsor a bill to increase taxes on oil and gas development by an additional 10% 2. created a program of financial incentives designed to bring qualified physicians to practice medicine in Abil 3. 2001.06.12 die of heard attack 4. 1995年上任 5. 原来是一个医生 6. 2001年3月，其法案被废除了
Rufus Drymiau	spokesman of the government		
Carmine Gurganus	Government spokesman		
President Araullo	前任总统	YR1993	
President Dorel Kapelou II	现任总统	2000年上任	Araullo连任失败
~~	现任 Minister of Health		Kapelou的侄子
Tomas Sarto	Minister of interior		

2）POK的高层

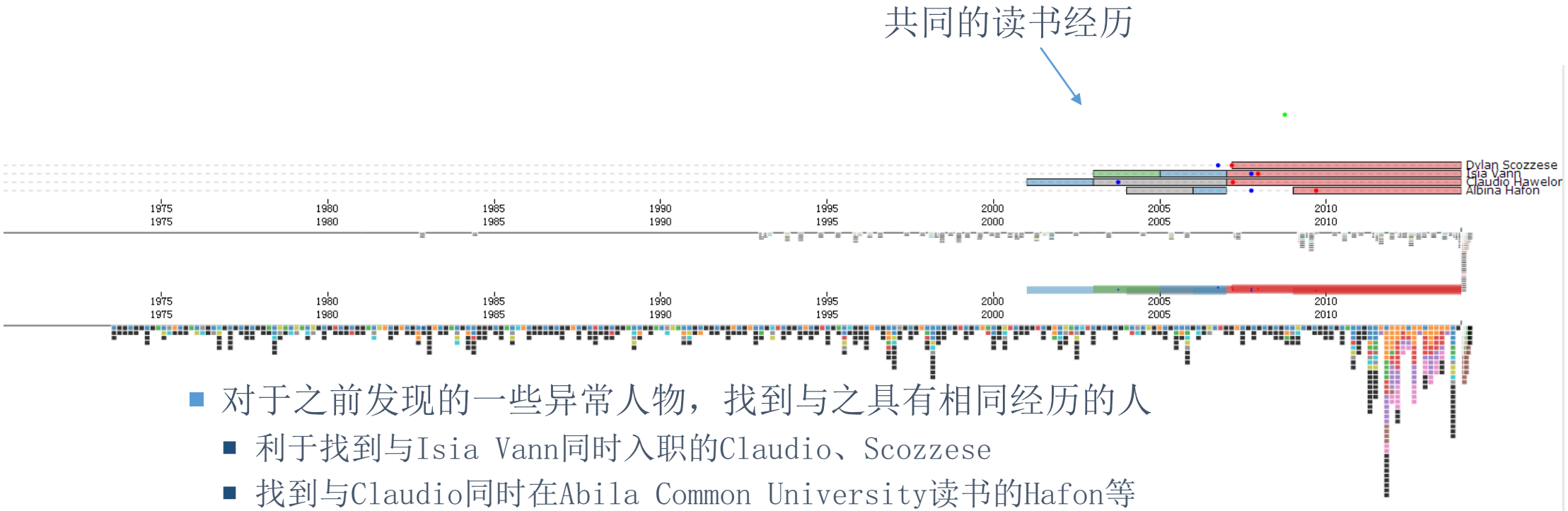
3）Kronos政府官员

人物关系可视分析－对发现的异常人物进行关系探索



通过刷选时间，获取子区间段的时间

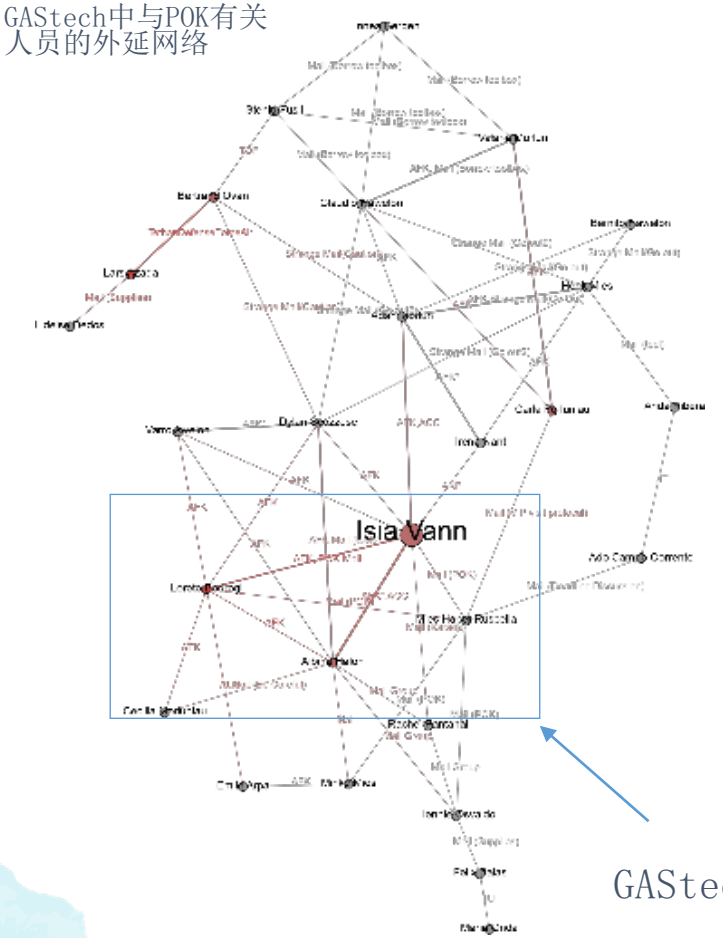
人物关系可视分析－对发现的异常人物进行关系探索（2）



- 对于之前发现的一些异常人物，找到与之具有相同经历的人
  - 利于找到与Isia Vann同时入职的Claudio、Scozzese
  - 找到与Claudio同时在Abila Common University读书的Hafon等
- 并且对于人物关系与新闻可以进行关联分析，探索在当年是否发生一些事件，与人物的经历有关

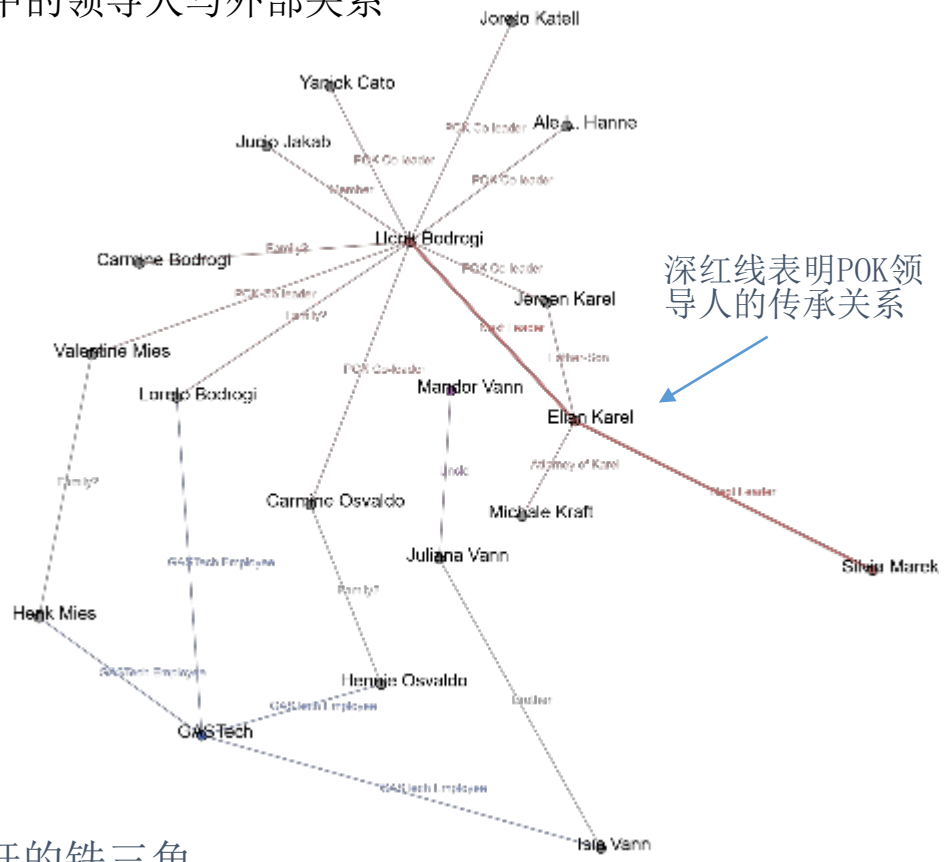
人物关系可视分析－对以上分析的人物关系进行整理

GAStech中与POK有关人员的外延网络



GAStech中内奸的铁三角

POK中的领导人与外部关系

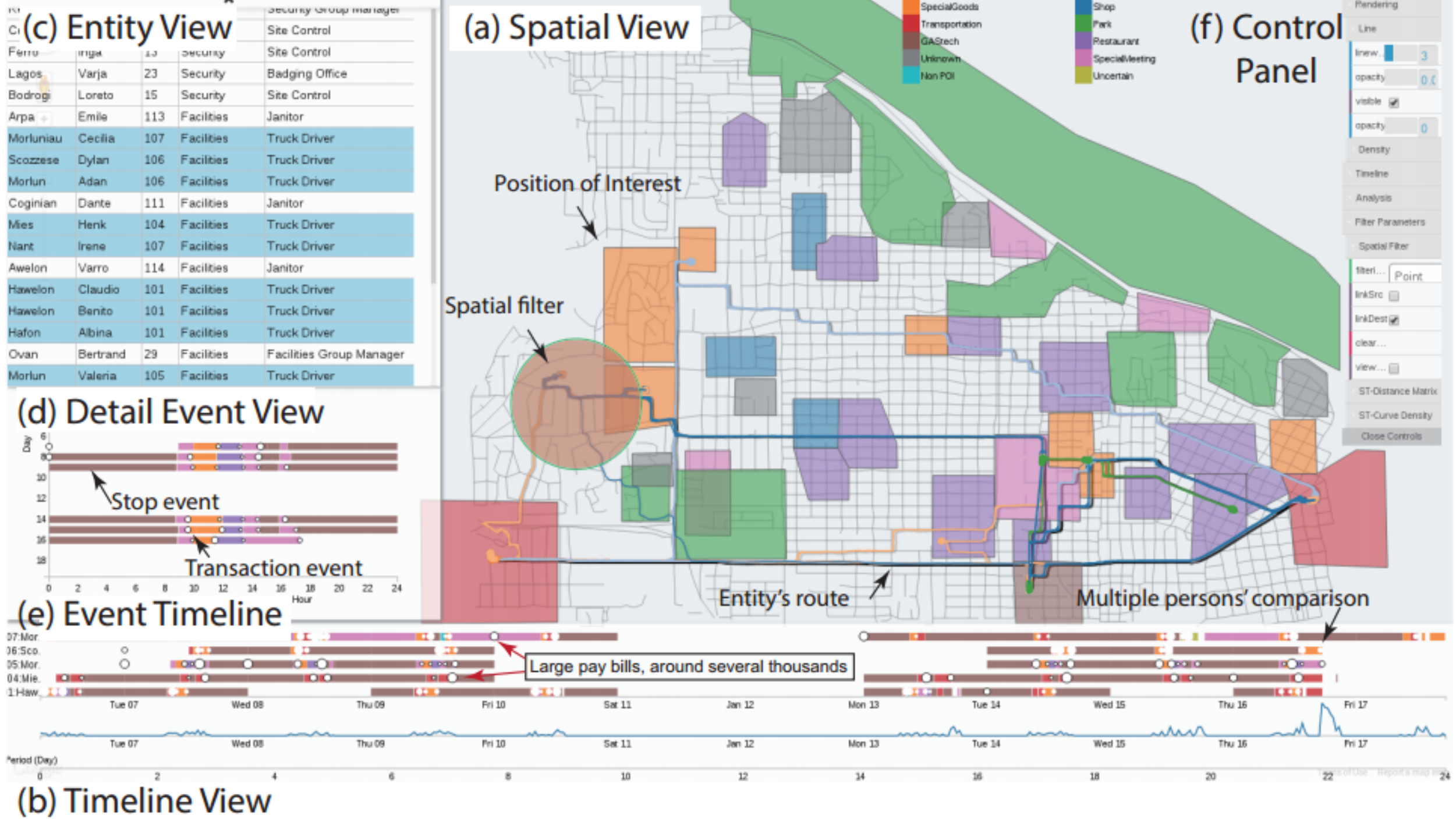


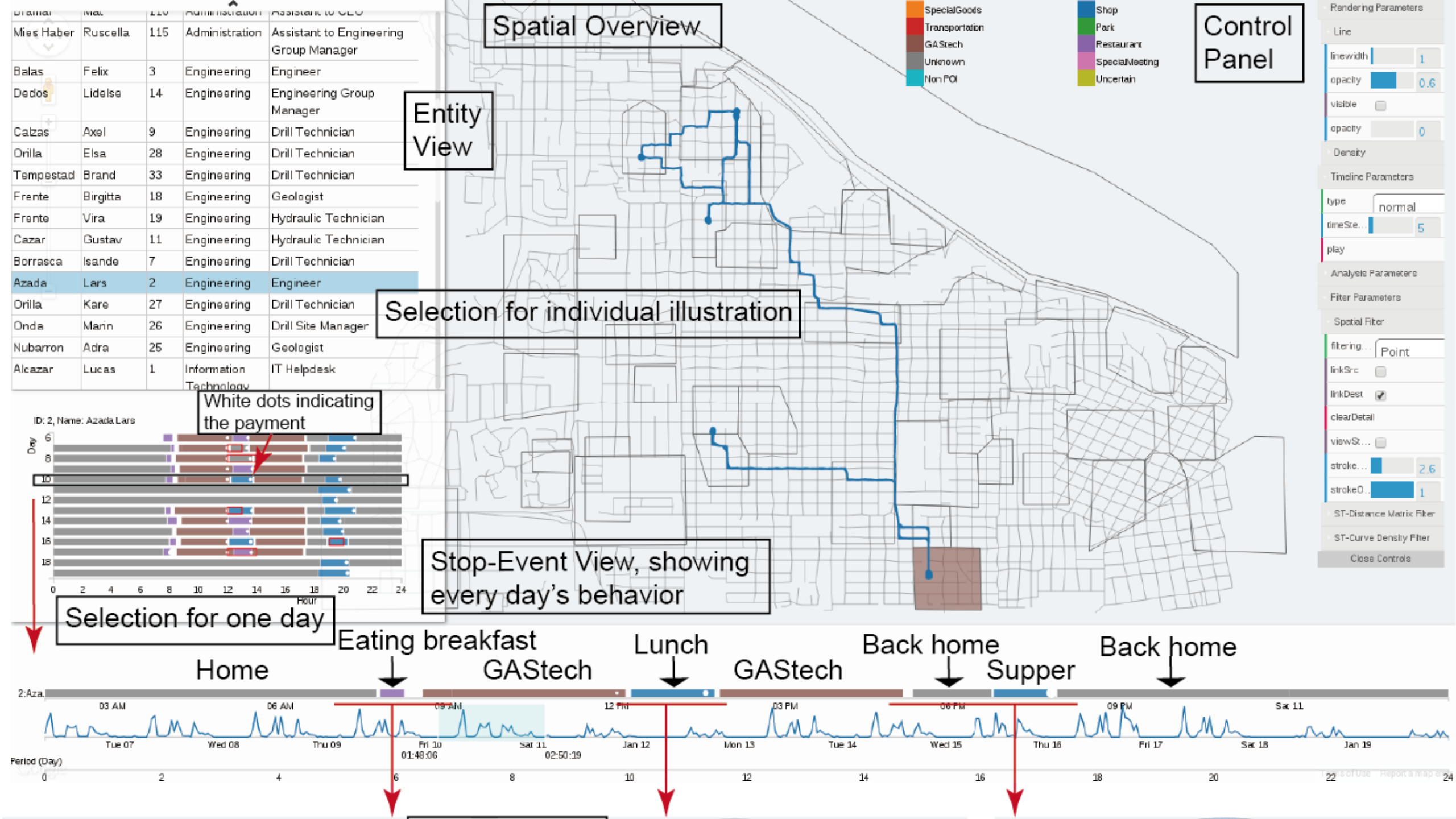
深红线表明POK领导人的传承关系

时空可视分析

- 正常模式的定义
  - 主要访问目的地
  - 访问时间
- 异常的定义
  - 访问特殊的地点
  - 在特殊时间访问地点
  - 聚集行为
  - 大型交易等等

- 基于多过滤器的时空探索
  - 时间过滤
  - 空间过滤
  - 实体过滤
  - 事件选择与比较

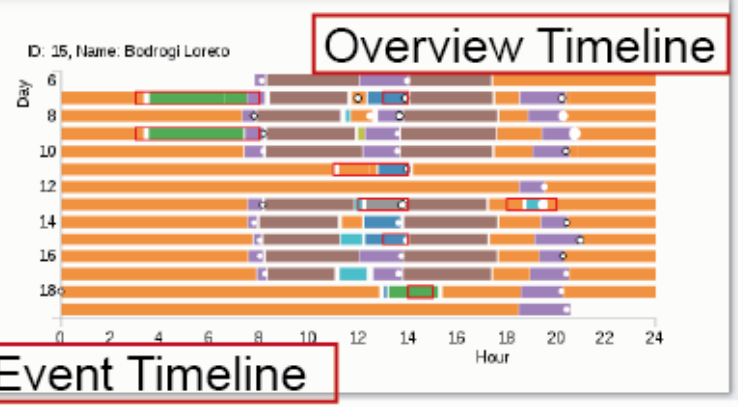

(c) Entity View
(a) Spatial View
(f) Control Panel
Position of Interest
Spatial filter
(d) Detail Event View
Stop event
Transaction event
Entity's route
Multiple persons' comparison
(e) Event Timeline
Large pay bills, around several thousands
(b) Timeline View



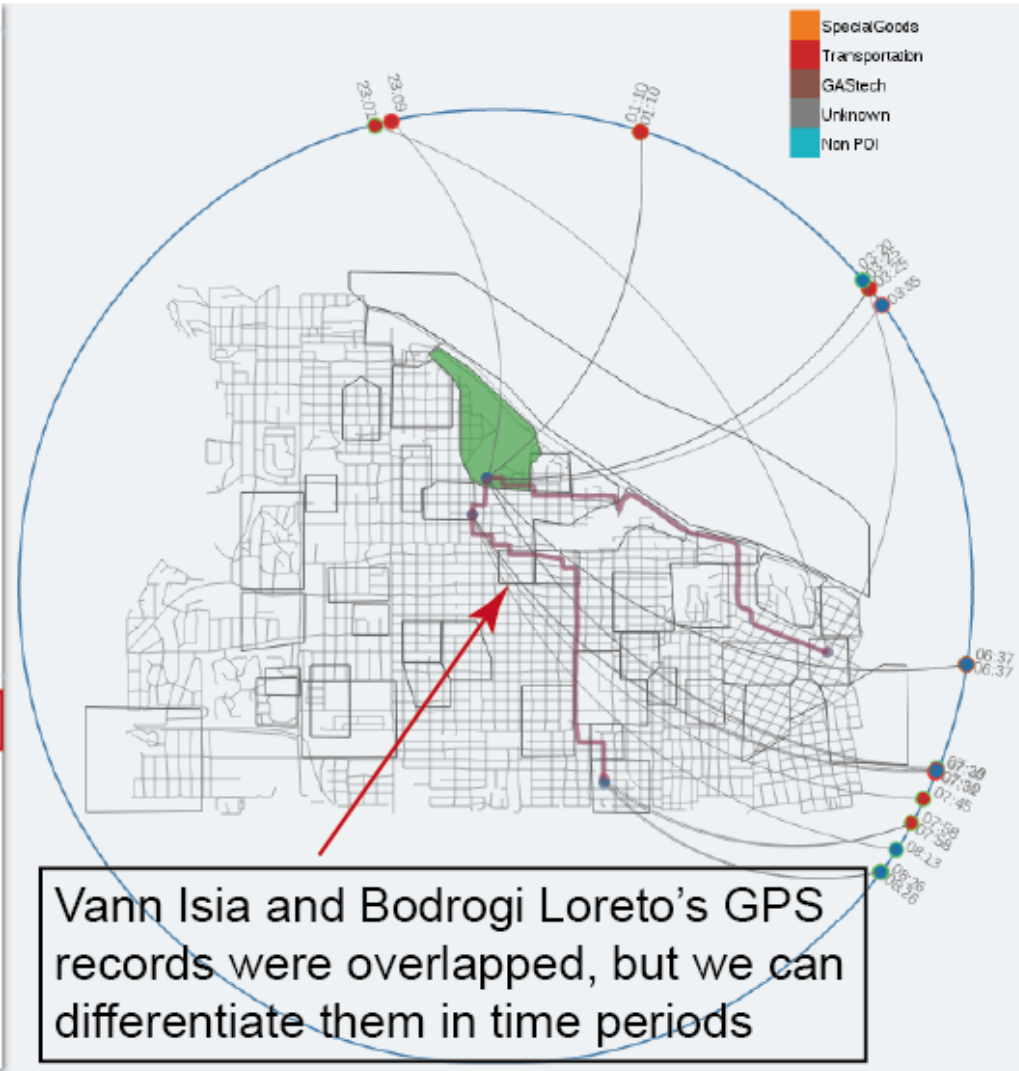
Mies Haber	Ruscella	115	Administration	Assistant to Engineering Group Manager
Balas	Felix	3	Engineering	Engineer
Dedos	Lidelse	14	Engineering	Engineering Group Manager
Calzas	Axel	9	Engineering	Drill Technician
Orilla	Elsa	28	Engineering	Drill Technician
Tempestad	Brand	33	Engineering	Drill Technician
Frente	Birgitta	18	Engineering	Geologist
Frente	Vira	19	Engineering	Hydraulic Technician
Cazar	Gustav	11	Engineering	Hydraulic Technician
Borrasca	Isande	7	Engineering	Drill Technician
Azada	Lars	2	Engineering	Engineer
Orilla	Kare	27	Engineering	Drill Technician
Onda	Marin	26	Engineering	Drill Site Manager
Nubarron	Adra	25	Engineering	Geologist
Alcazar	Lucas	1	Information Technology	IT Helpdesk

Event Overview
23:01, Jan.6 - 8:26, Jan.7

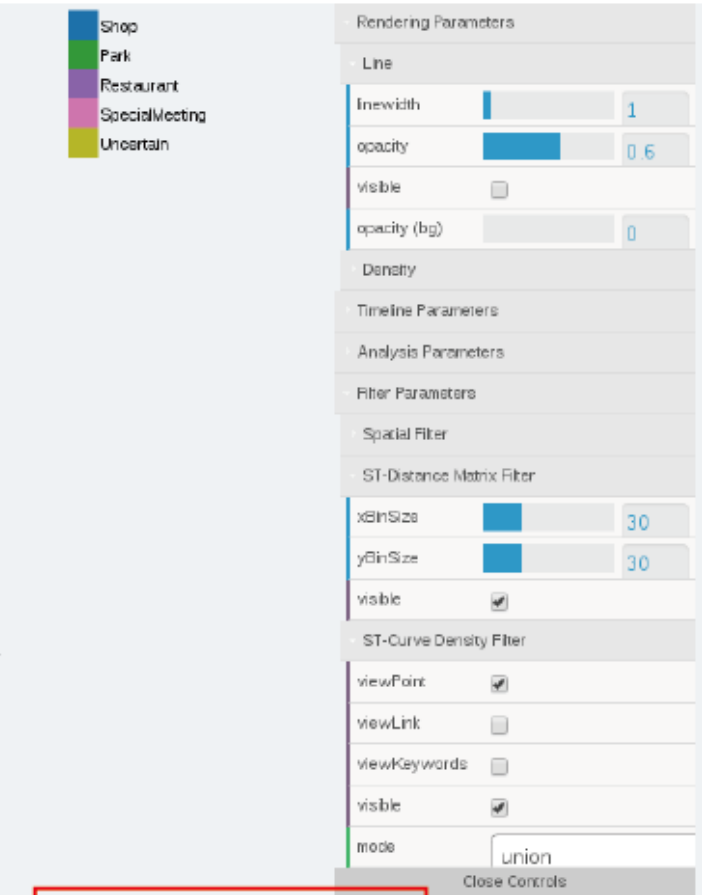
Fusil	Stenig	20	Security	Building Control
Herrero	Kanon	22	Security	Badging Office
Vann	Edvard	34	Security	Perimeter Control
Vann	Isia	16	Security	Perimeter Control
Mies	Minke	24	Security	Perimeter Control
Resumir	Felix	30	Security	Security Group Manager
Cocinaro	Hideki	12	Security	Site Control
Ferro	Inga	13	Security	Site Control
Lagos	Varja	23	Security	Badging Office
Bodrogi	Loreto	15	Security	Site Control
Arpa	Emile	113	Facilities	Janitor
Morluniau	Cecilia	107	Facilities	Truck Driver
Scozzese	Dylan	106	Facilities	Truck Driver
Morlun	Adan	106	Facilities	Truck Driver
Coginian	Dante	111	Facilities	Janitor
Mies	Henk	104	Facilities	Truck Driver
Nant	Irene	107	Facilities	Truck Driver



Overview Timeline



Vann Isia and Bodrogi Loreto's GPS records were overlapped, but we can differentiate them in time periods

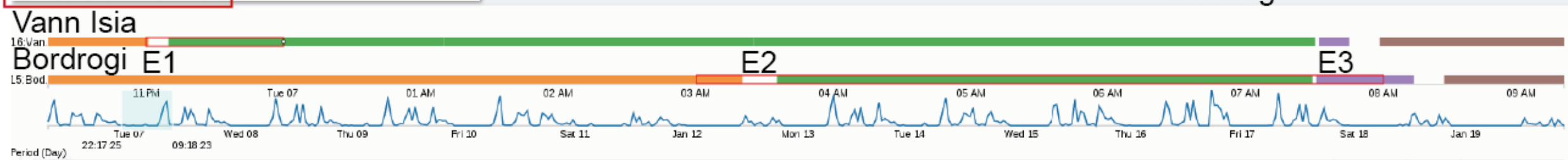


People Legend

Vann Isia

Bordrogi

Event Timeline



可疑人小结

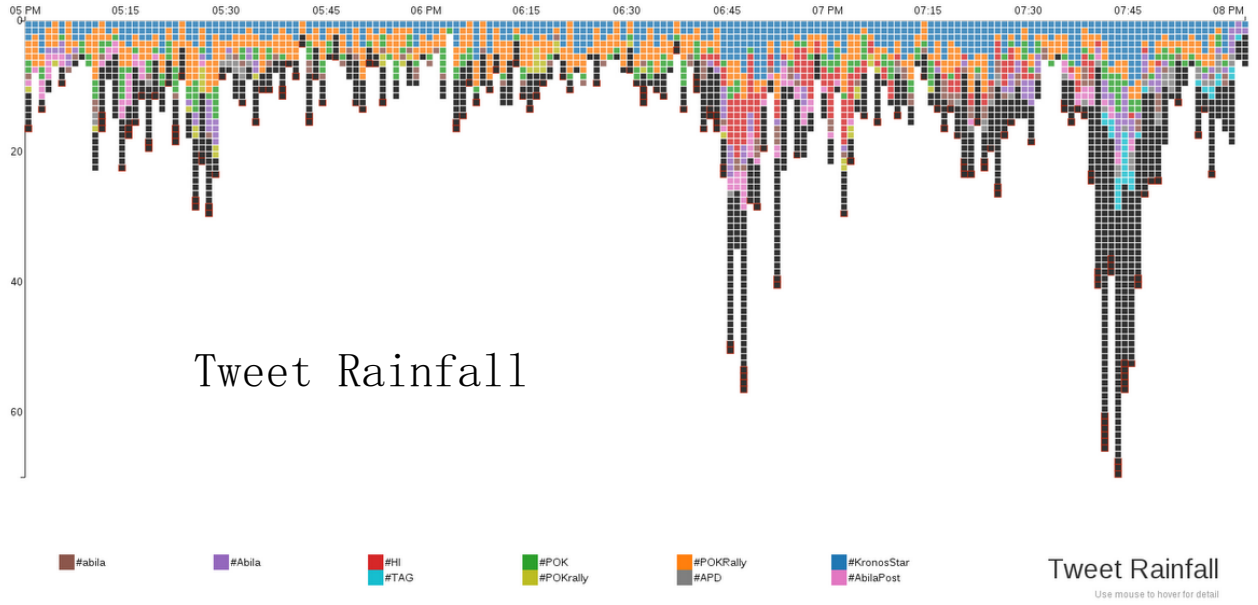
Alcazar	Lucas	1990/4/17	Tethys	Male	1990/4/17	Information Technology	IT Helpdesk	2010/11/30	Lucas.Alcazar@gastech.com.kronos			
Bodrogi	Loreto	1989/4/17	Kronos	Male	1989/4/17	Security	Site Control	2013/8/17	Loreto.Bodrogi@gastech.com.kronos	ArmedForcesOfKronos	HonorableDischarge	2008/10/1
Borrasca	Isande	1979/10/22	Tethys	Female	1979/10/22	Engineering	Drill Technician	2004/2/18	Isande.Borrasca@gastech.com.kronos			
Bramar	Mat	1981/12/19	Tethys	Male	1981/12/19	Administration	Assistant to CEO	2005/7/1	Mat.Bramar@gastech.com.kronos			
Mies	Henk	1984/9/23	Kronos	Male	1984/9/23	Facilities	Truck Driver	2013/7/12	Henk.Mies@gastech.com.kronos	ArmedForcesOfKronos	HonorableDischarge	2004/10/1
Mies	Minke	1992/11/19	Kronos	Male	1992/11/19	Security	Perimeter Control	2013/5/22	Minke.Mies@gastech.com.kronos	ArmedForcesOfKronos	GeneralDischarge	2011/10/1
Morlun	Valeria	1981/9/22	Kronos	Female	1981/9/22	Facilities	Truck Driver	2003/12/2	Valeria.Morlun@gastech.com.kronos	ArmedForcesOfKronos	HonorableDischarge	2001/10/1
Nubarron	Adra	1968/6/6	Tethys	Female	1968/6/6	Engineering	Geologist	1994/7/8	Adra.Nubarron@gastech.com.kronos			
Orilla	Kare	1964/2/29	Tethys	Female	1964/2/29	Engineering	Drill Technician	2005/11/12	Kare.Orilla@gastech.com.kronos			
Osvaldo	Hennie	1988/5/31	Kronos	Male	1988/5/31	Security	Perimeter Control	2011/6/7	Hennie.Osvaldo@gastech.com.kronos	ArmedForcesOfKronos	GeneralDischarge	2010/10/1
Ovan	Bertrand	1964/12/12	Tethys	Male	1964/12/12	Facilities	Facilities Group Manager	1998/10/1	Bertrand.Ovan@gastech.com.kronos	TethanDefenseForceAir	HonorableDischarge	1993/10/1
Scozzese	Dylan	1986/5/25	Kronos	Male	1986/5/25	Facilities	Truck Driver	2007/3/1	Dylan.Scozzese@gastech.com.kronos	ArmedForcesOfKronos	HonorableDischarge	2006/10/1
Tempestad	Brand	1979/5/14	Tethys	Male	1979/5/14	Engineering	Drill Technician	2004/6/12	Brand.Tempestad@gastech.com.kronos			
Vann	Edvard	1991/3/18	Kronos	Male	1991/3/18	Security	Perimeter Control	2013/8/15	Edvard.Vann@gastech.com.kronos	ArmedForcesOfKronos	HonorableDischarge	2011/10/1

流式数据分析

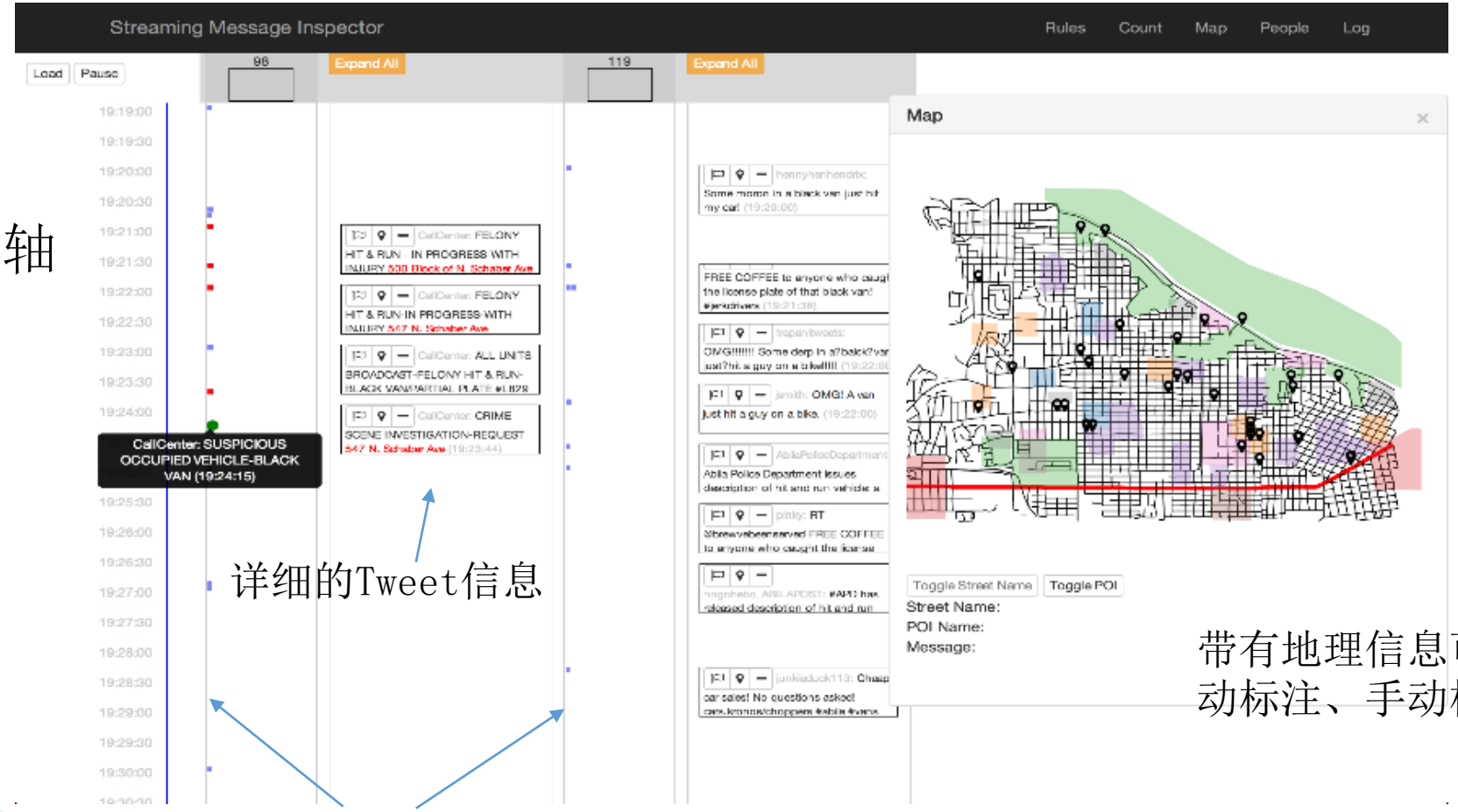
- 流式数据：对未知的不确定
  - 内容
  - 体量
  - 总体到细节？
- 消息
  - 结构化的元数据
- 有限时间的挑战

Twitter与出警记录流数据分析（1）整体分析

- 话题与时间演变分析
  - X轴是时间
  - Y轴是把每分钟的微博数据进行堆叠排列
  - 将不同的话题抽取，并用颜色来表示



Twitter与出警记录流数据分析（2）实时流数据分析



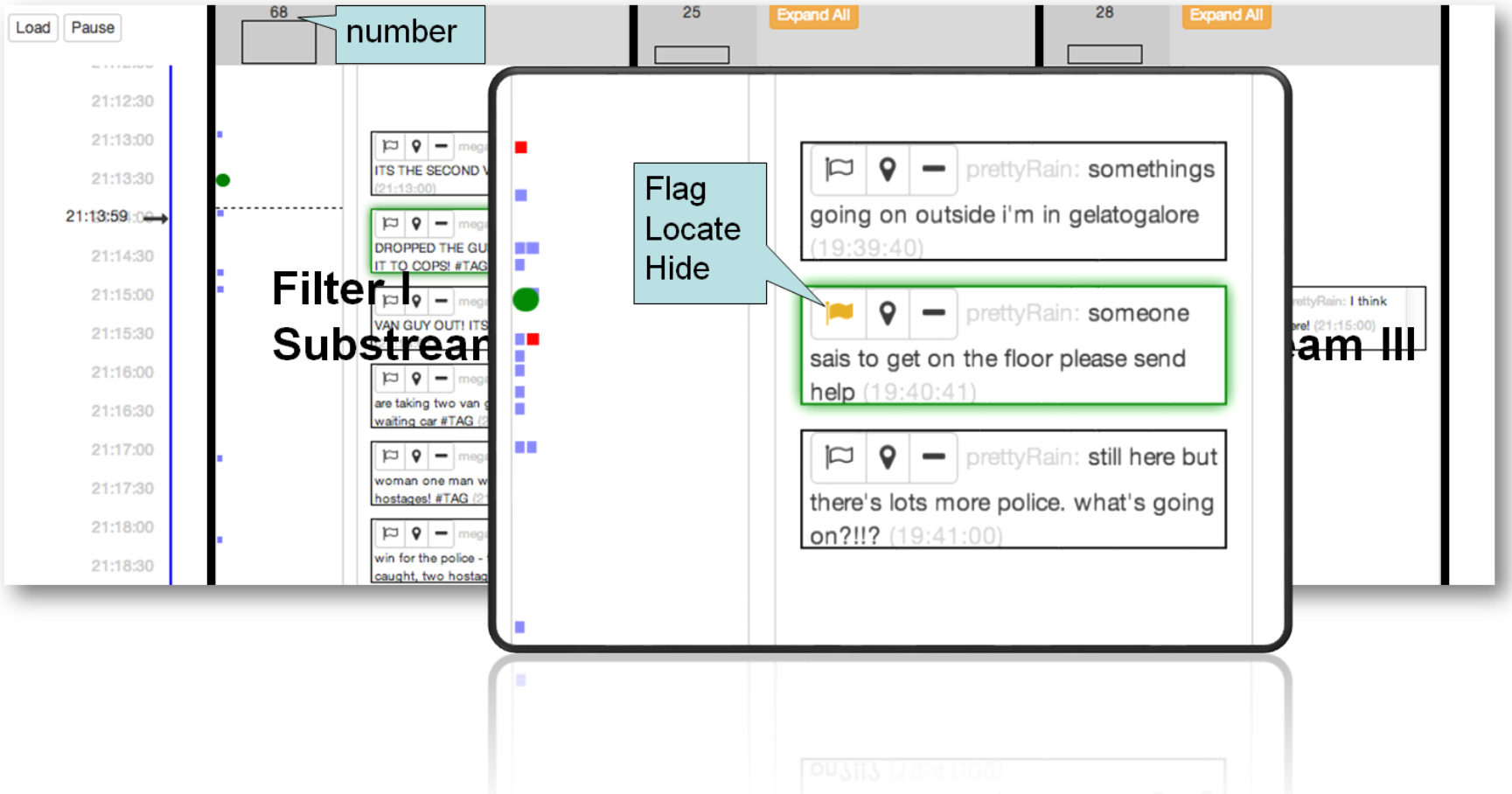
时间轴

详细的Tweet信息

带有地理信息可以自动标注、手动校准

总体的流数据，可以被用户筛选出不同的子数据流，并分发给不同的人合作分析

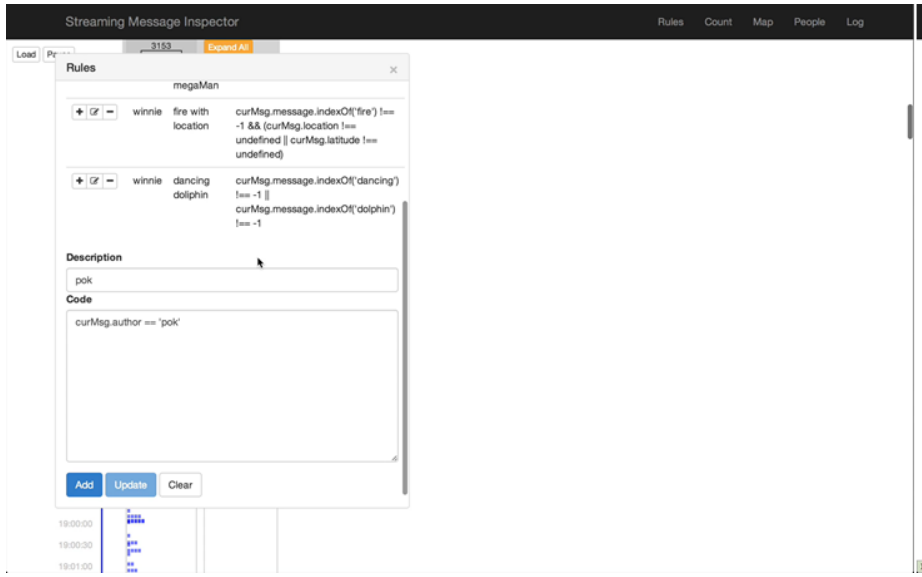
总控过滤器设计

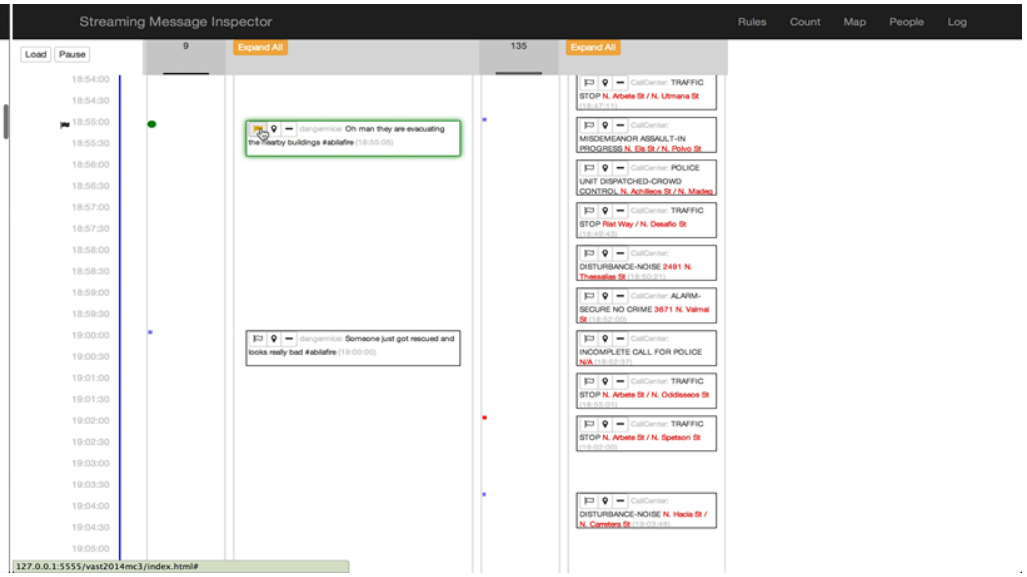

Load
Pause
68
number
25
Expand All
28
Expand All
21:12:30
21:13:00
21:13:30
21:13:59
21:14:30
21:15:00
21:15:30
21:16:00
21:16:30
21:17:00
21:17:30
21:18:00
21:18:30
Filter I
Substrean
Flag
Locate
Hide
prettyRain: somethings going on outside i'm in gelatogalore (19:39:40)
prettyRain: someone sais to get on the floor please send help (19:40:41)
prettyRain: still here but there's lots more police. what's going on?!!? (19:41:00)
eam III

过滤器细节

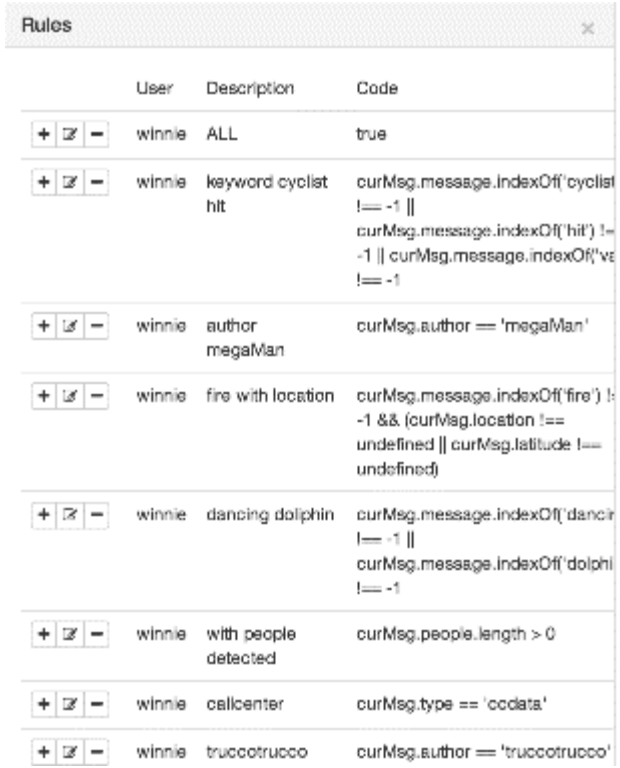
- 规则

映射


Streaming Message Inspector
Rules
Count
Map
People
Log
Rules
megaMan
winnie
fire with location
curMsg.message.indexOf('fire') !== -1 && (curMsg.location !== undefined || curMsg.latitude !== undefined)
winnie
dancing dolphin
curMsg.message.indexOf('dancing') !== -1 || curMsg.message.indexOf('dolphin') !== -1
Description
pok
Code
curMsg.author == 'pok'
Add
Update
Clear


Streaming Message Inspector
Rules
Count
Map
People
Log
Load
Pause
Expand All
127.0.0.1:5555/vast2014mc3/index.html#

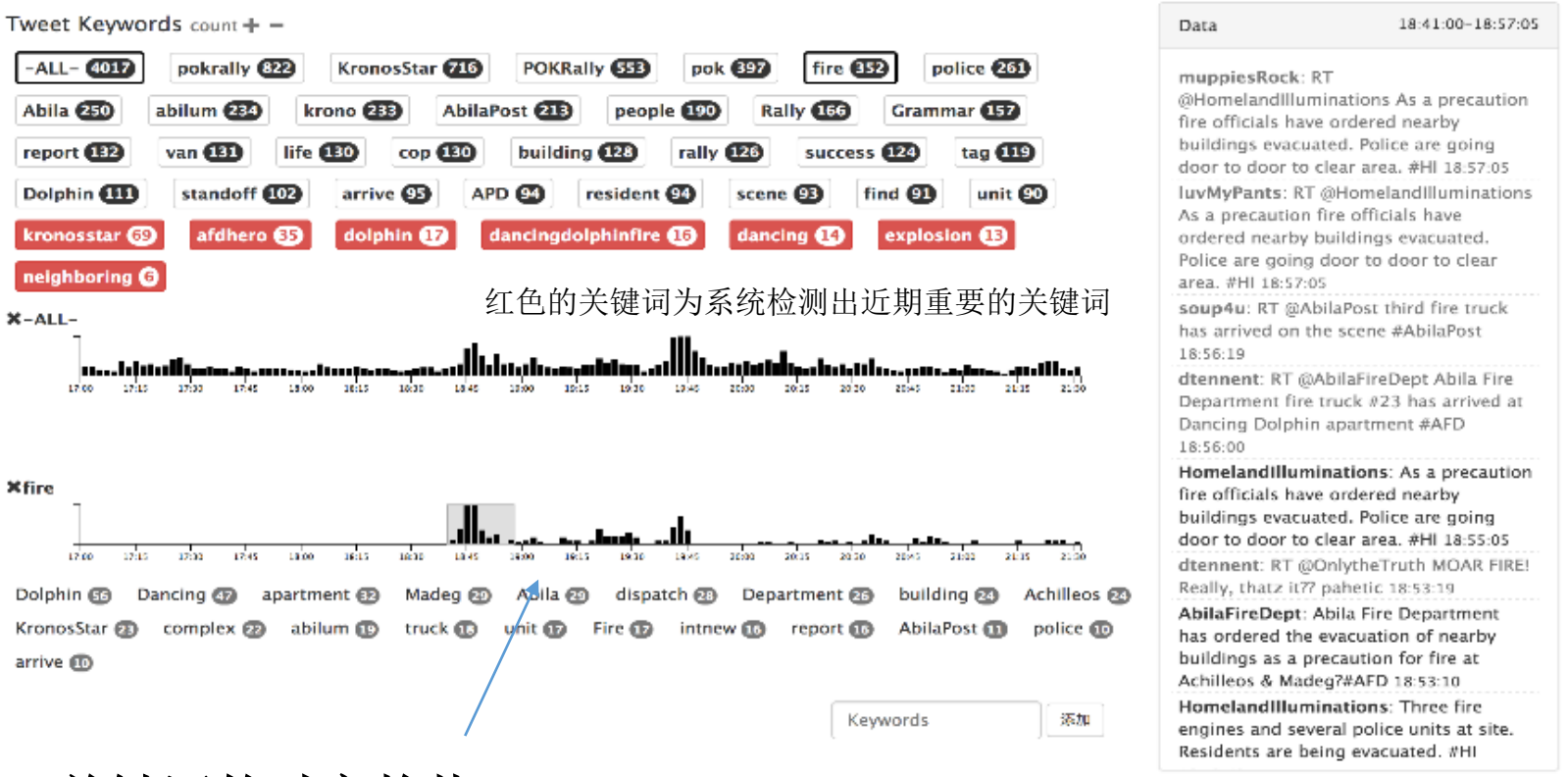
Twitter与出警记录流数据分析（3）实时流数据分析—筛选机制



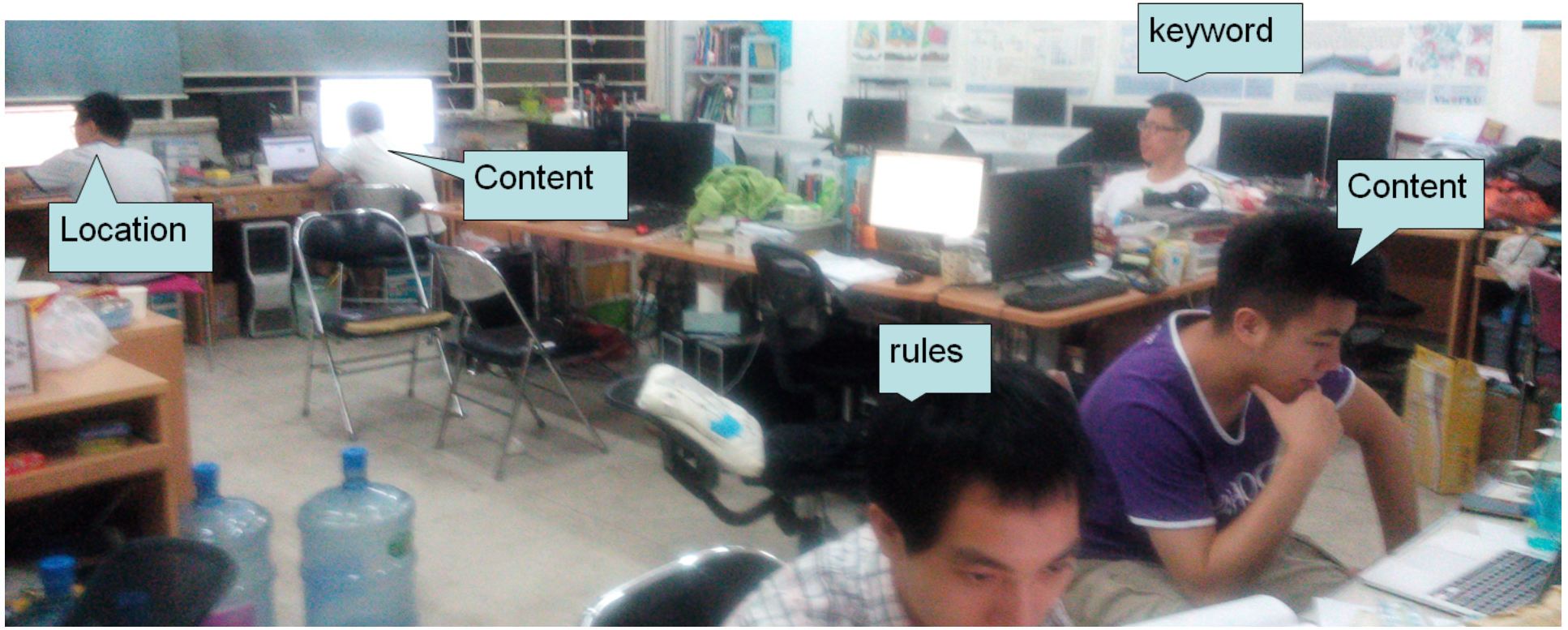
用户可以创建不同的过滤器：筛选出子数据流

相关人物信息会被自动标注，高亮出现次数

Twitter与出警记录流数据分析（3）实时关键词分析



多人合作分析



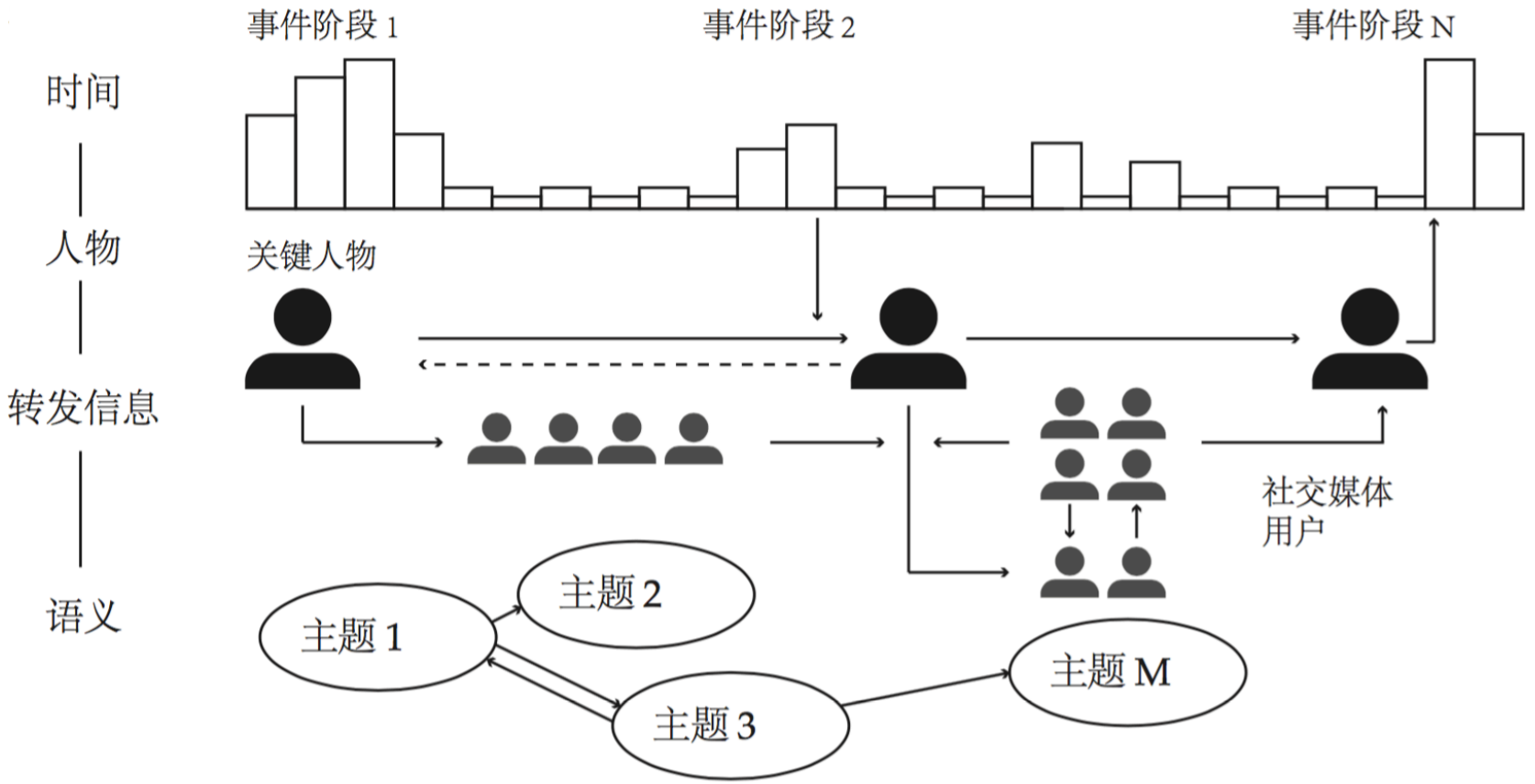
情报分析 - VAST Challenge 2014小结

- 不同源数据的融合
  - 数据预处理
  - 交互辅助探索
- 不同时间粒度的探索
  - 事件（短暂、瞬时）
  - 故事（长时间、多步骤）
- 关键角色关系的探索
  - 从多个信息源中搜索关系
  - “同时”出现在共同机构、学校、部队的隐形相关关系
- 基本规律与异常检测可视分析
  - 设置多维度（时间、空间、人物）的筛选器
  - 综合条件、与基本规律进行比对判断

对重大社会事件的动态演变及用户行为可视分析

- 研究动机
  - 社会事件频发，社交媒体扮演重要作用
  - 用于支持舆情分析的态势感知
- 挑战
  - 复杂的时变特征
  - 动态的用户行为与讨论
  - 如何探究与推断事件动态与用户行为的关联关系

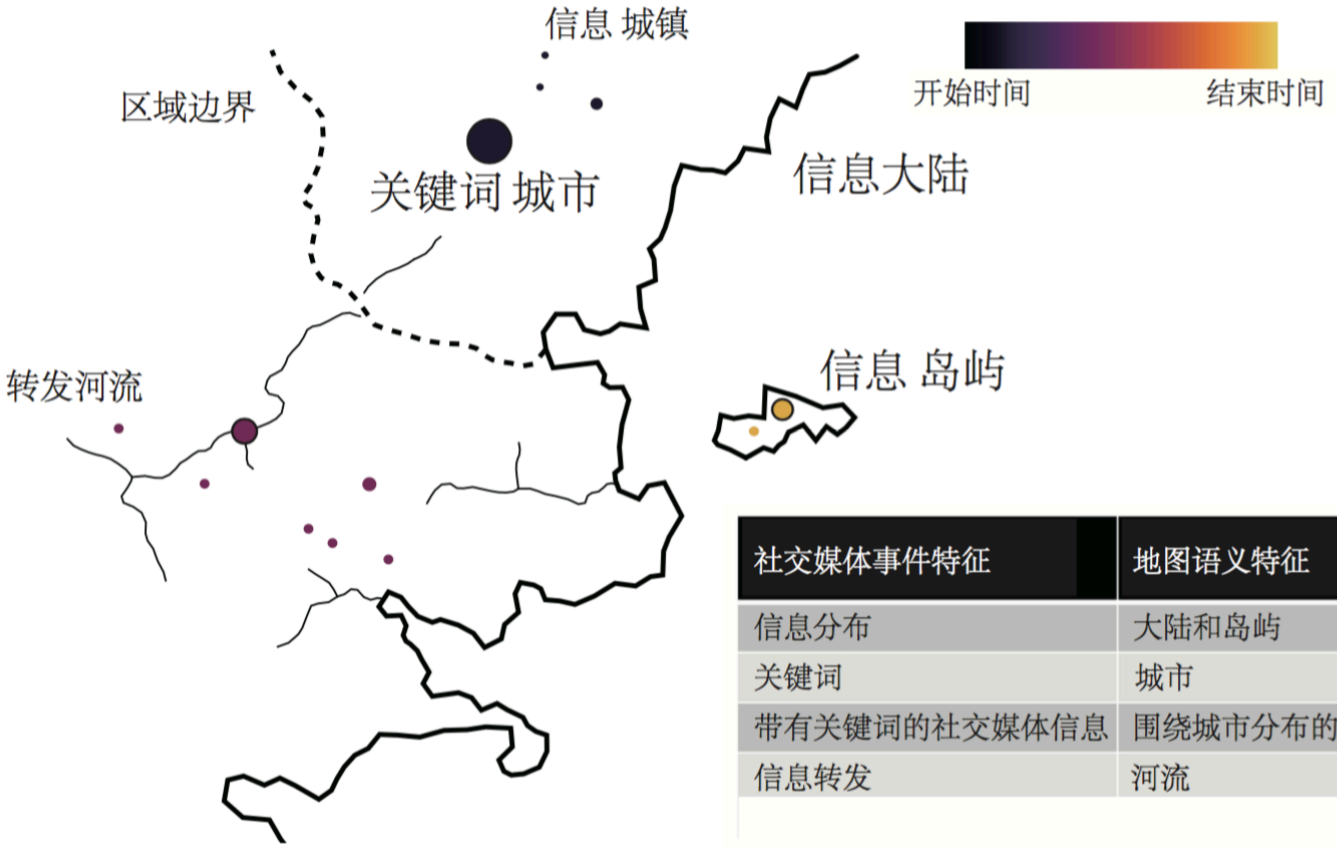
社会事件



分析任务

- D1：多个事件阶段
- D2：人物的影响力
- D3：由于发送与转发行为产生的信息传播
- D4：不同主题的动态变化

E-Map设计



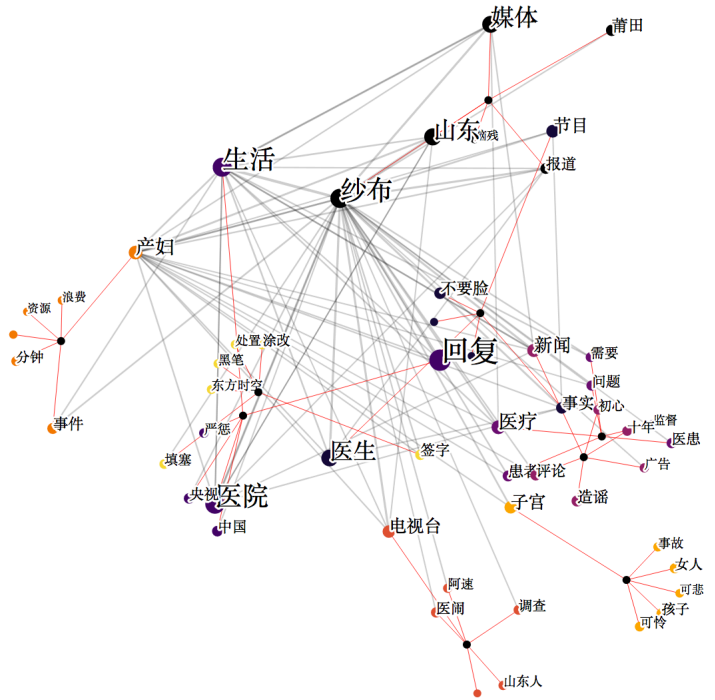
社交媒体事件特征	地图语义特征
信息分布	大陆和岛屿
关键词	城市
带有关键词的社交媒体信息	围绕城市分布的城镇
信息转发	河流

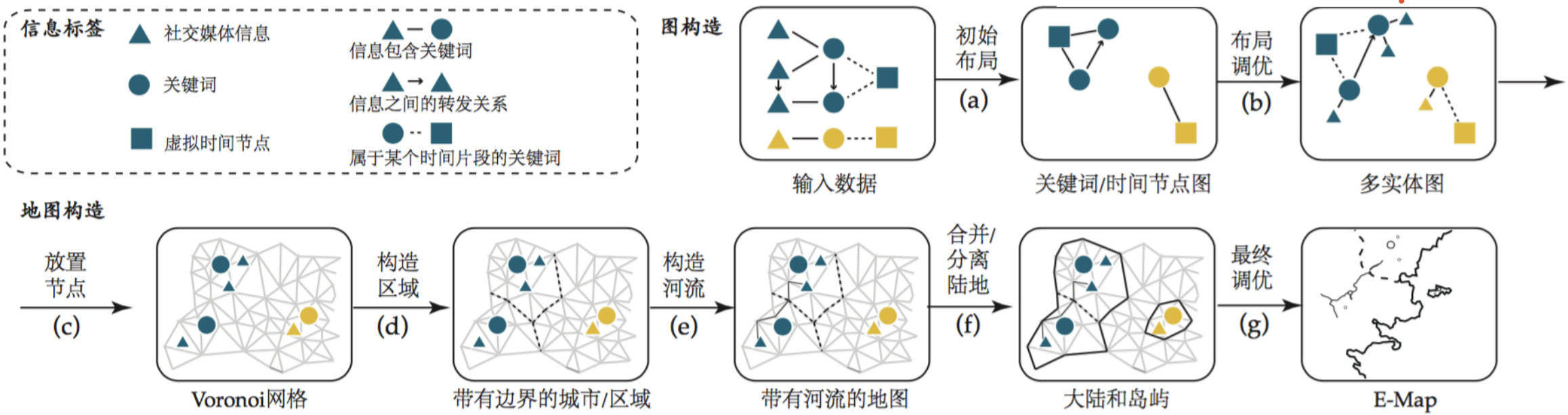
设计需求

- R1：提供语义与时间趋势上的事件总结。
- R2：根据信息的主题以及转发关系进行聚合。
- R3：检测与可视化不同时间段的主题
- R4：提取用户发送与转发信息的行为模式

E-Map构造算法

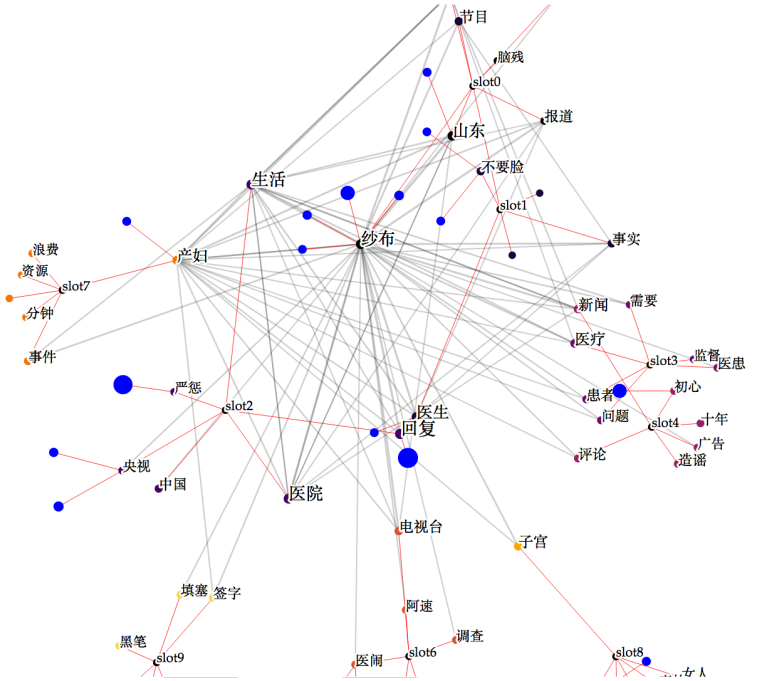
- 两步地图构造
  - 多实体图构造与布局
  - 地图生成

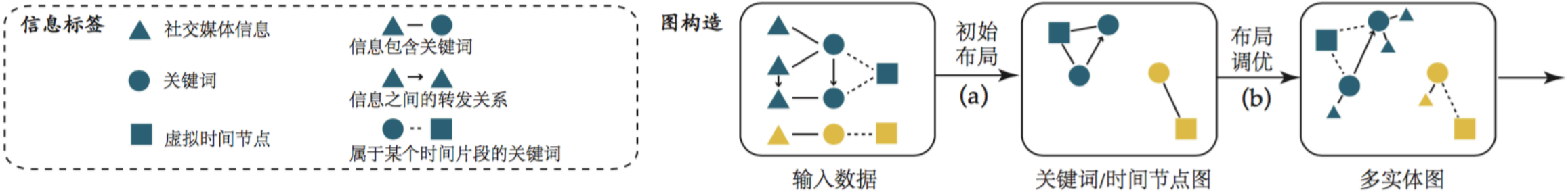




E-Map构造算法（1）

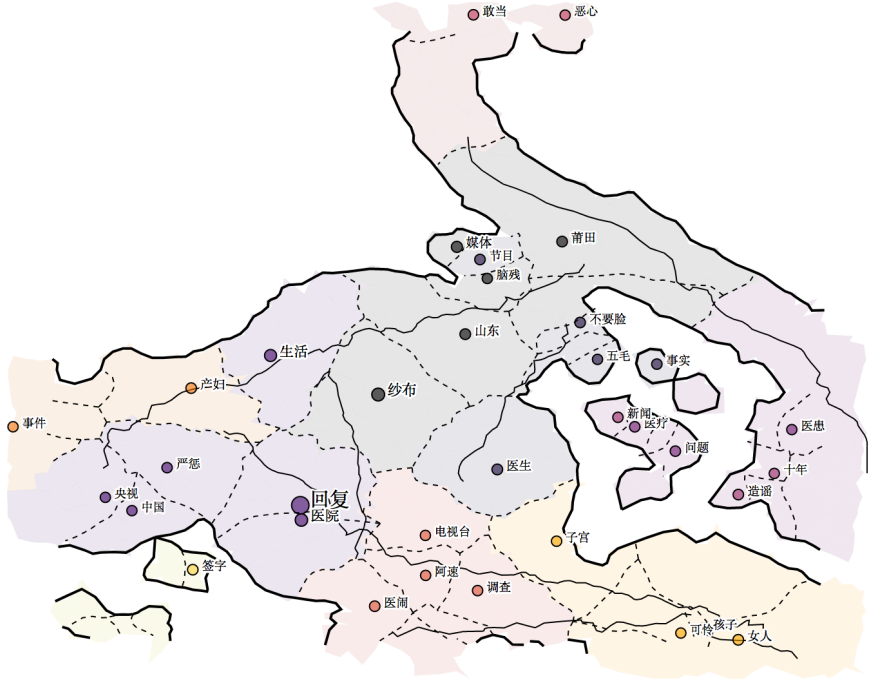
- 两步地图构造
  - 多实体图构造与布局
    - 关键词抽取（时变TF/IDF计算）
    - 基于稳定的力导向布局
    - 第一轮：关键词/时间 – 反映关键词的转发关系与时间特征
    - 第二轮：布局微调（弹开）以及加入信息节点，增加区域比例关系

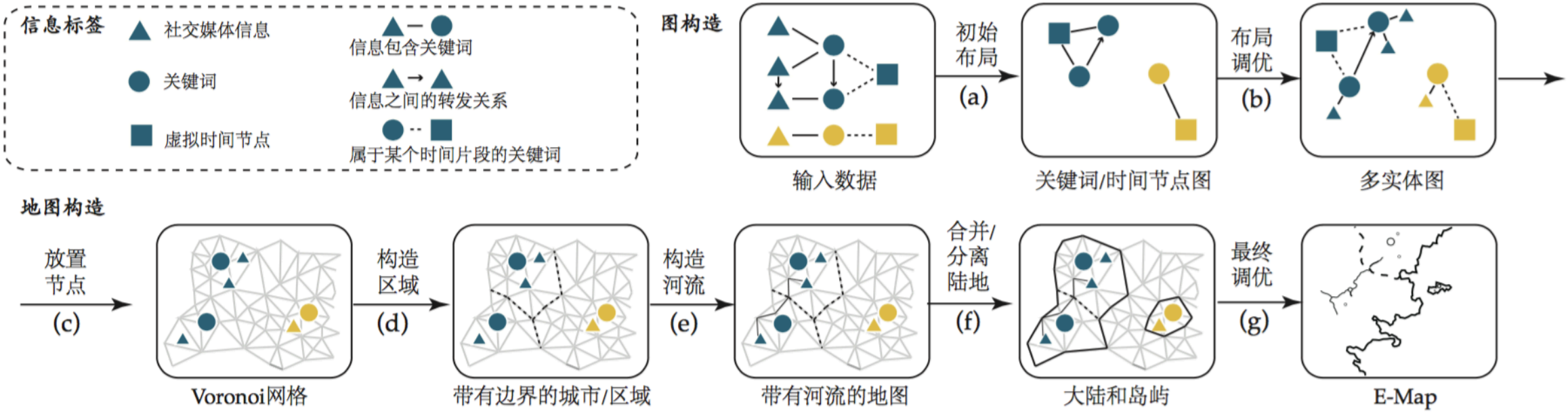


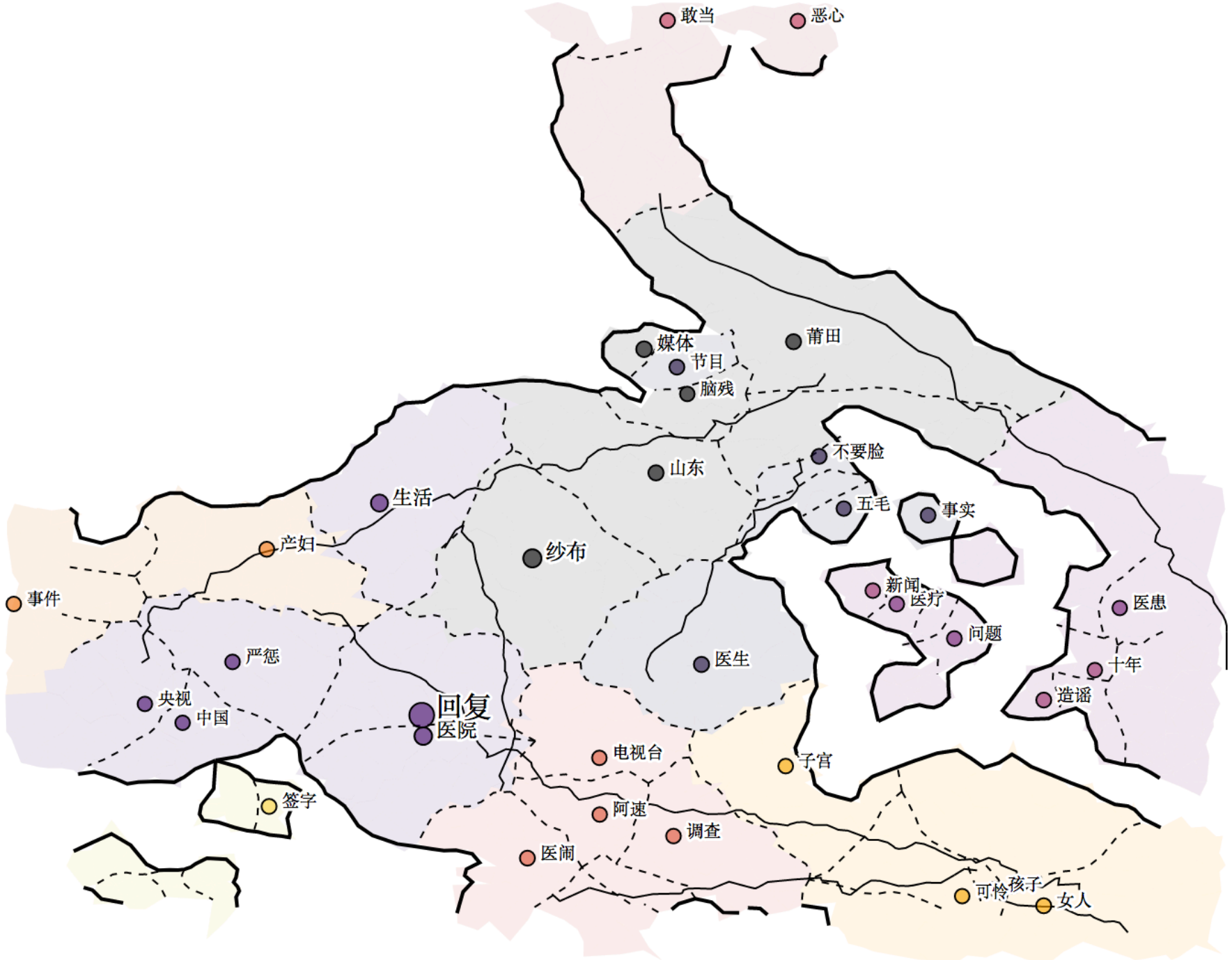


E-Map构造算法（2）

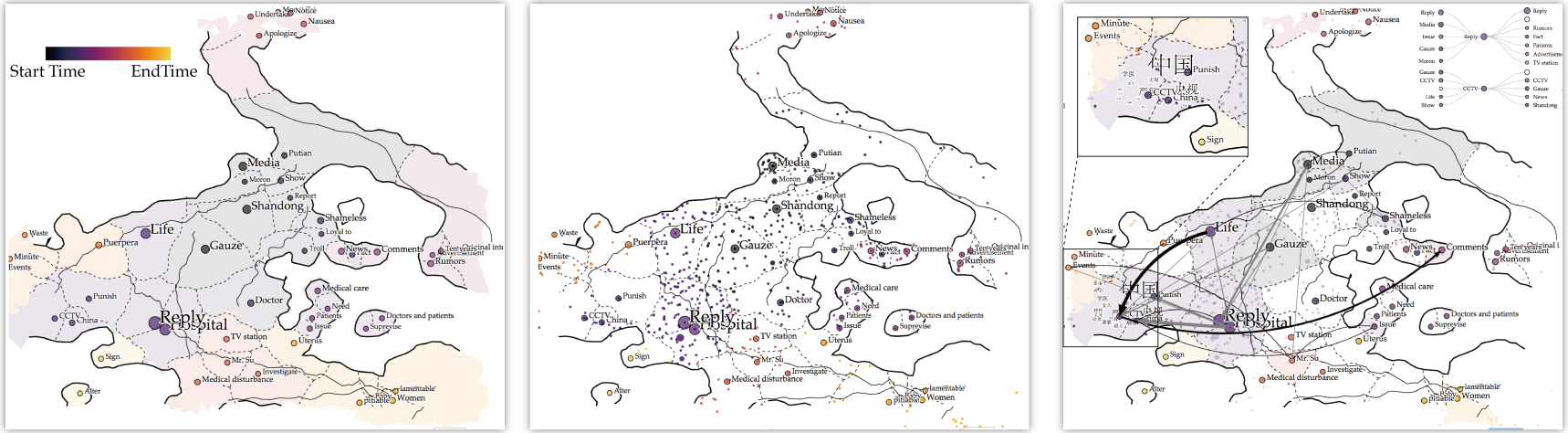
- 两步地图构造
  - 地图生成
    - 基于Voronoi网络（地图视觉效果、相邻矩阵关系计算方便）
    - 城市、区域、城镇边界构造 – 关键词分布、时间先后、相邻区域
    - 河流构造 – 转发关系
    - 大陆和岛屿 – 总体分布特征， 调优：河流平滑，边界腐蚀






敢当
恶心
媒体
节目
脑残
莆田
不要脸
山东
五毛
事实
生活
产妇
纱布
事件
新闻
医疗
医患
问题
十年
造谣
严惩
医生
央视
中国
回复
医院
电视台
子宫
签字
阿速
调查
医闹
可怜
孩子
女人

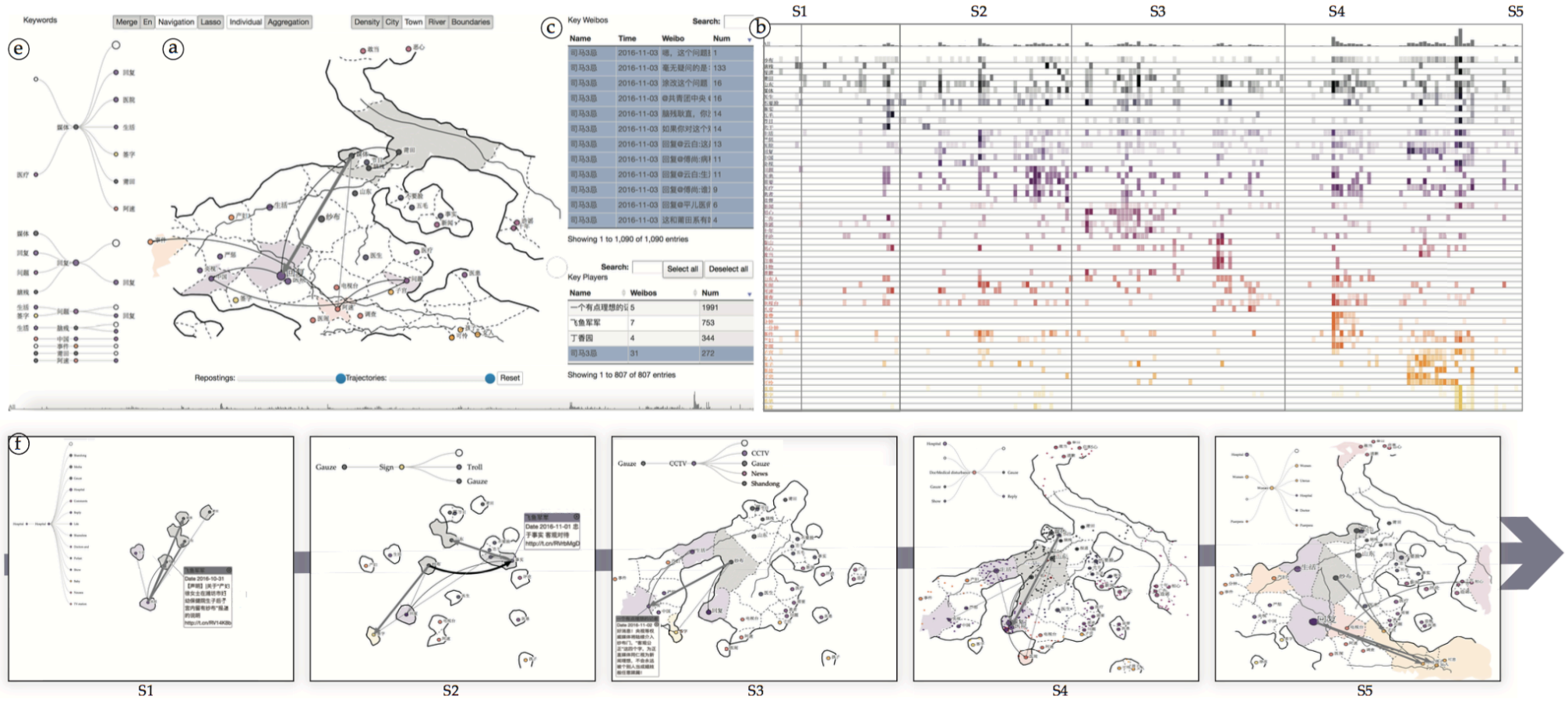
E-Map



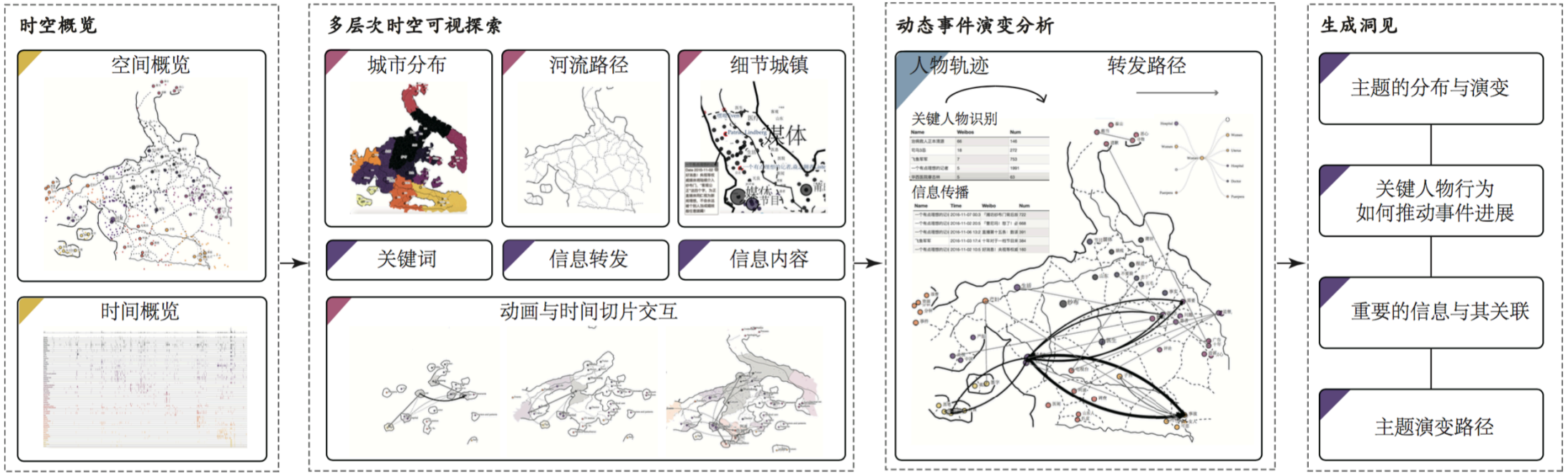
(a) (b) (c)

- 复杂、非结构化的信息 -> 结构化的、包含语义的可探索地图
- 用户的行为与关系 - >人群的移动与关联
- 支持时空可视分析

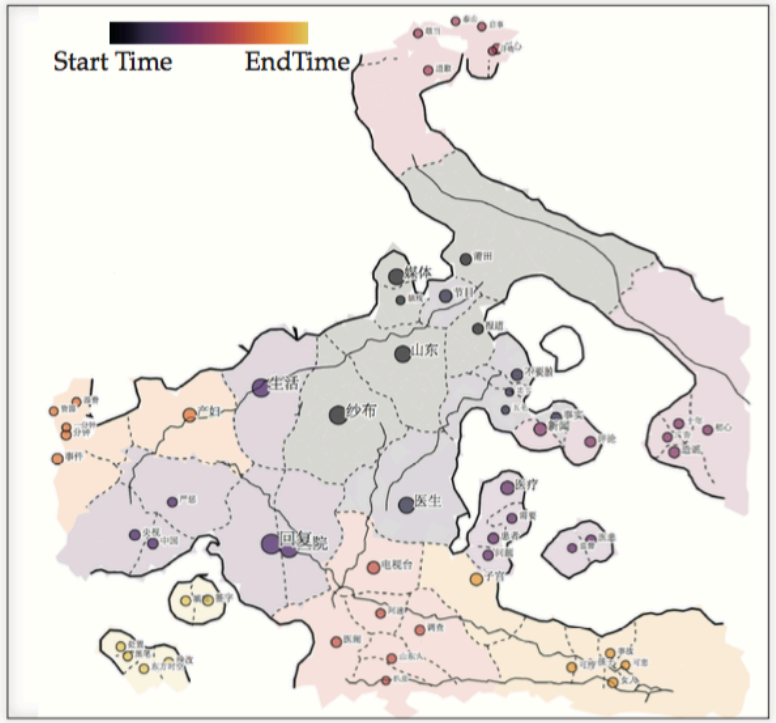
E-Map 可视分析系统

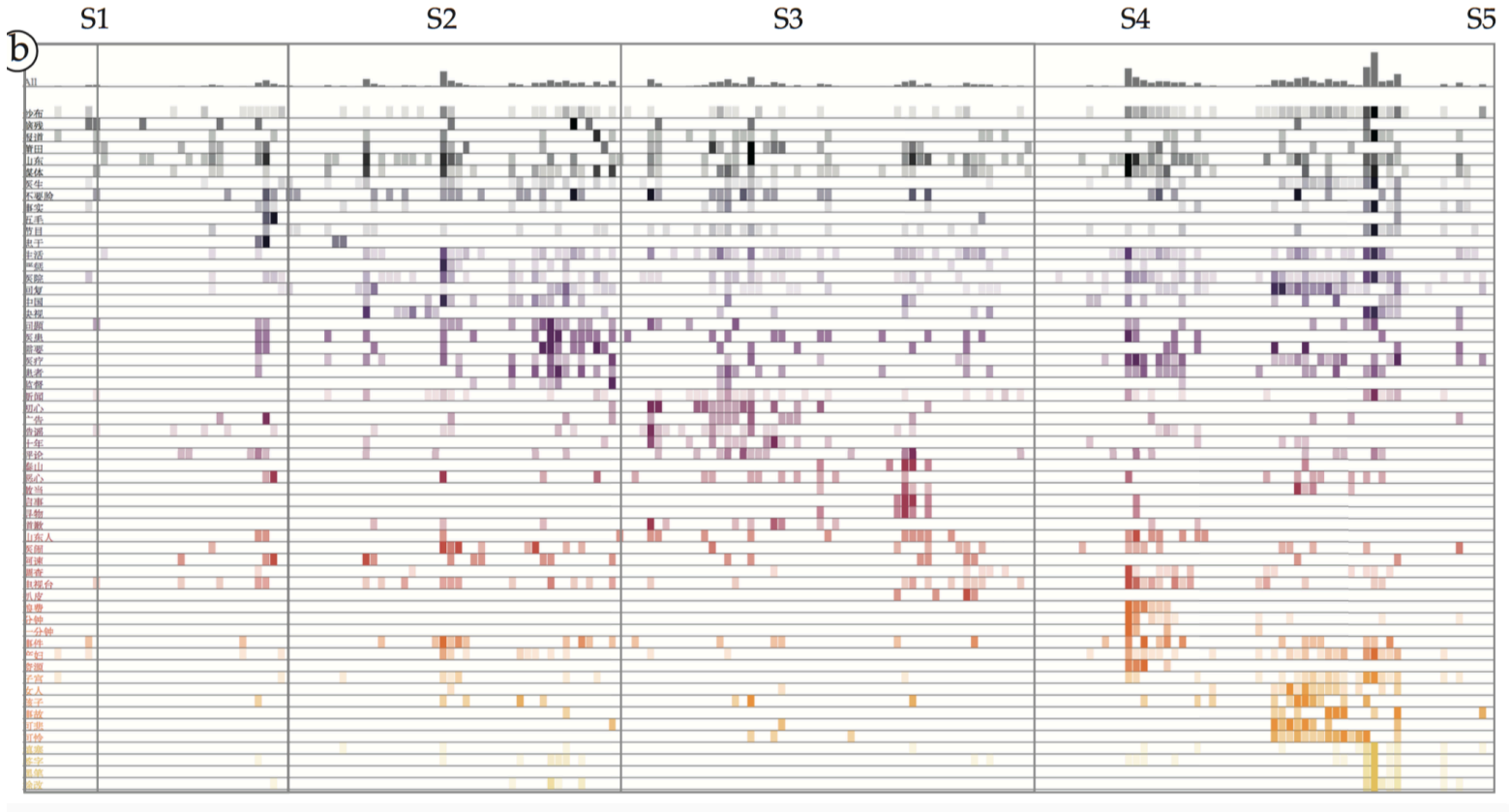

Keywords
Merge
En
Navigation
Lasso
Individual
Aggregation
Density
City
Town
River
Boundaries
Repostings:
Trajectories:
Reset
Key Weibos
Search:
Name
Time
Weibo
Num
Showing 1 to 1,090 of 1,090 entries
Key Players
Select all
Deselect all
Weibos
Showing 1 to 807 of 807 entries
S1
S2
S3
S4
S5
Gauze
Sign
Troll
CCTV
News
Shandong

E-Map 可视分析流程


时空概览
空间概览
时间概览
多层次时空可视探索
城市分布
河流路径
细节城镇
关键词
信息转发
信息内容
动画与时间切片交互
动态事件演变分析
人物轨迹
转发路径
关键人物识别
信息传播
生成洞见
主题的分布与演变
关键人物行为
如何推动事件进展
重要的信息与其关联
主题演变路径

E-Map 可视分析（1） – 时空概览

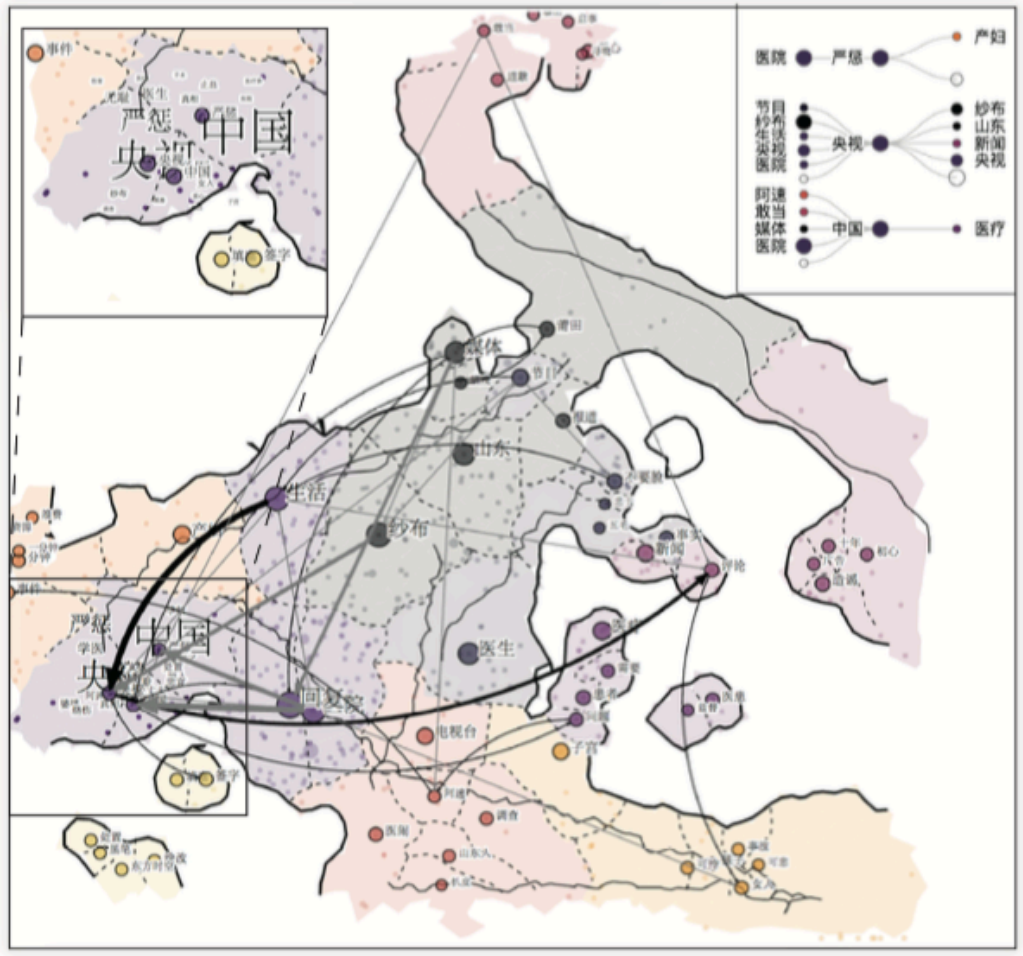




- 关键词的时序特征
- 地图的颜色

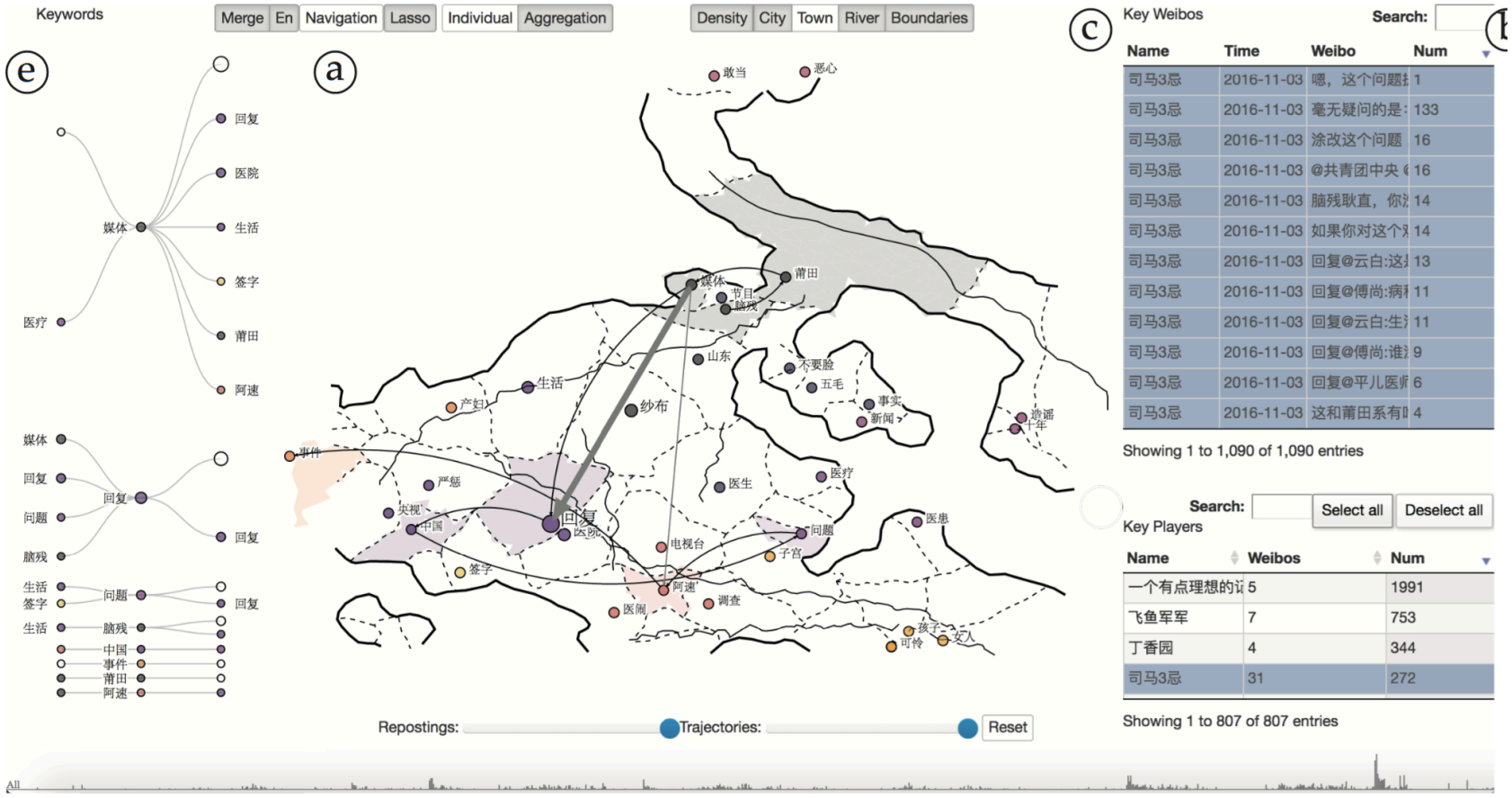
E-Map 可视分析（2）–多层次时空探索

- 时间轴的刷选与导航
- 地图的自由交互（导航、缩放
- 细节缩放与选择



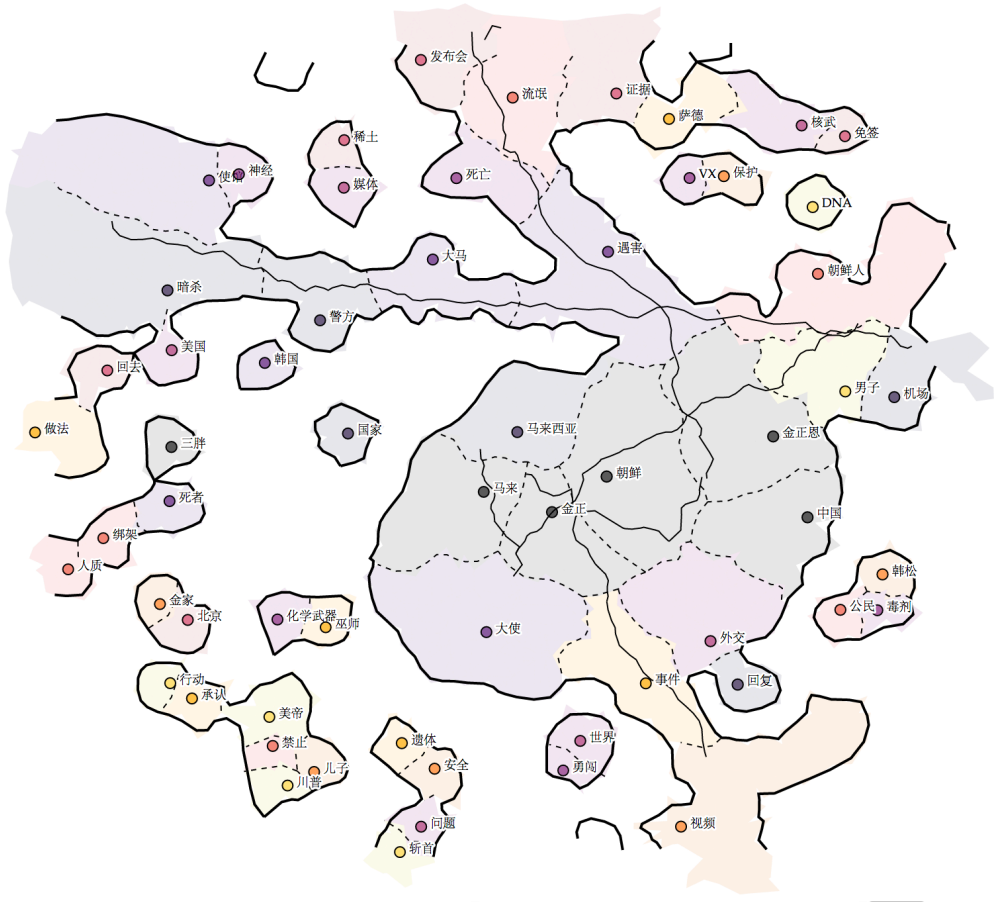
E-Map 可视分析（3） – 事件演变分析

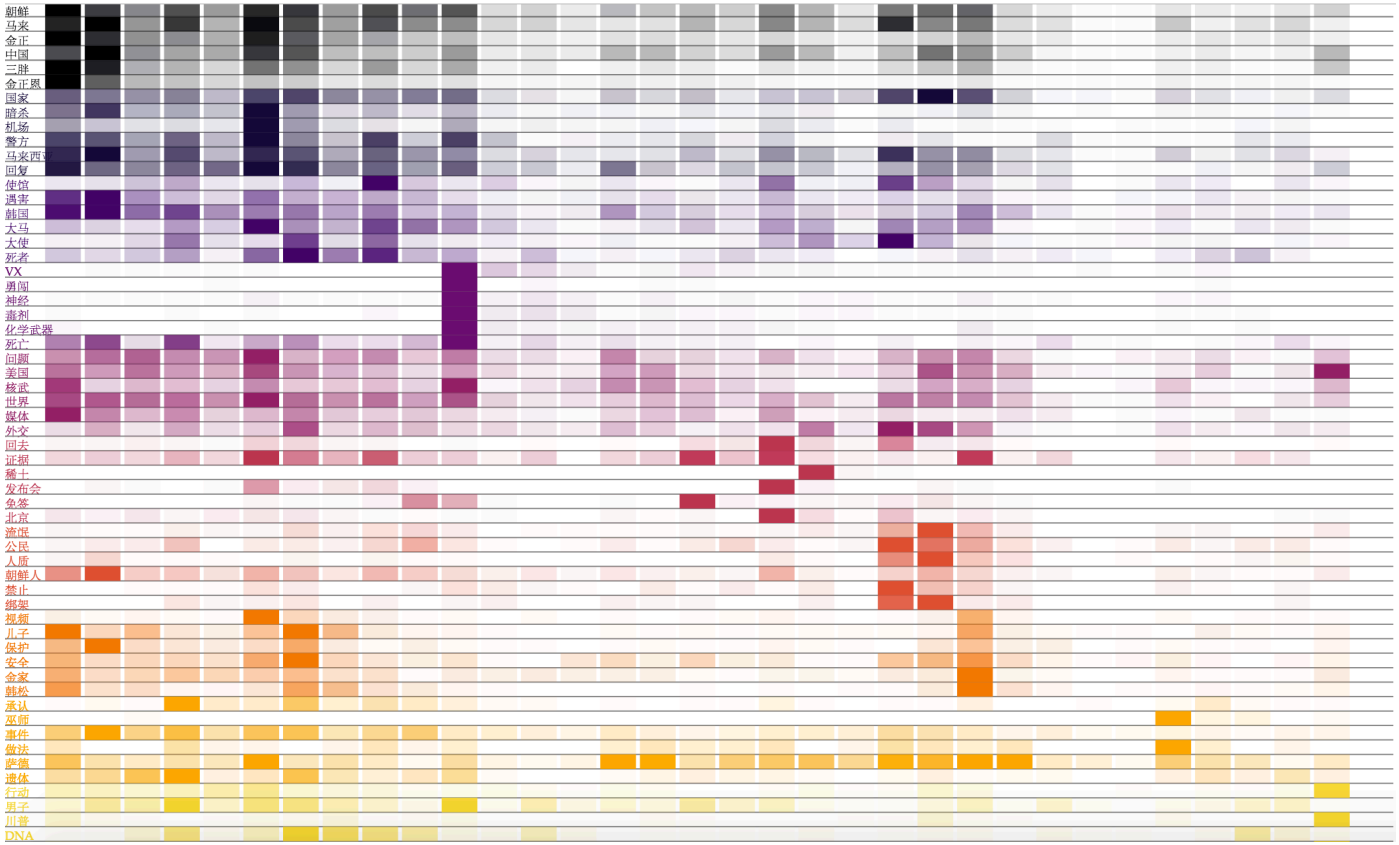
- 行为模式探索
  - 用户轨迹
    - 人群连续发送微博行为（话题的演变）
  - 用户连接关系
    - 微博的转发关系
- 聚合模式/细节模式

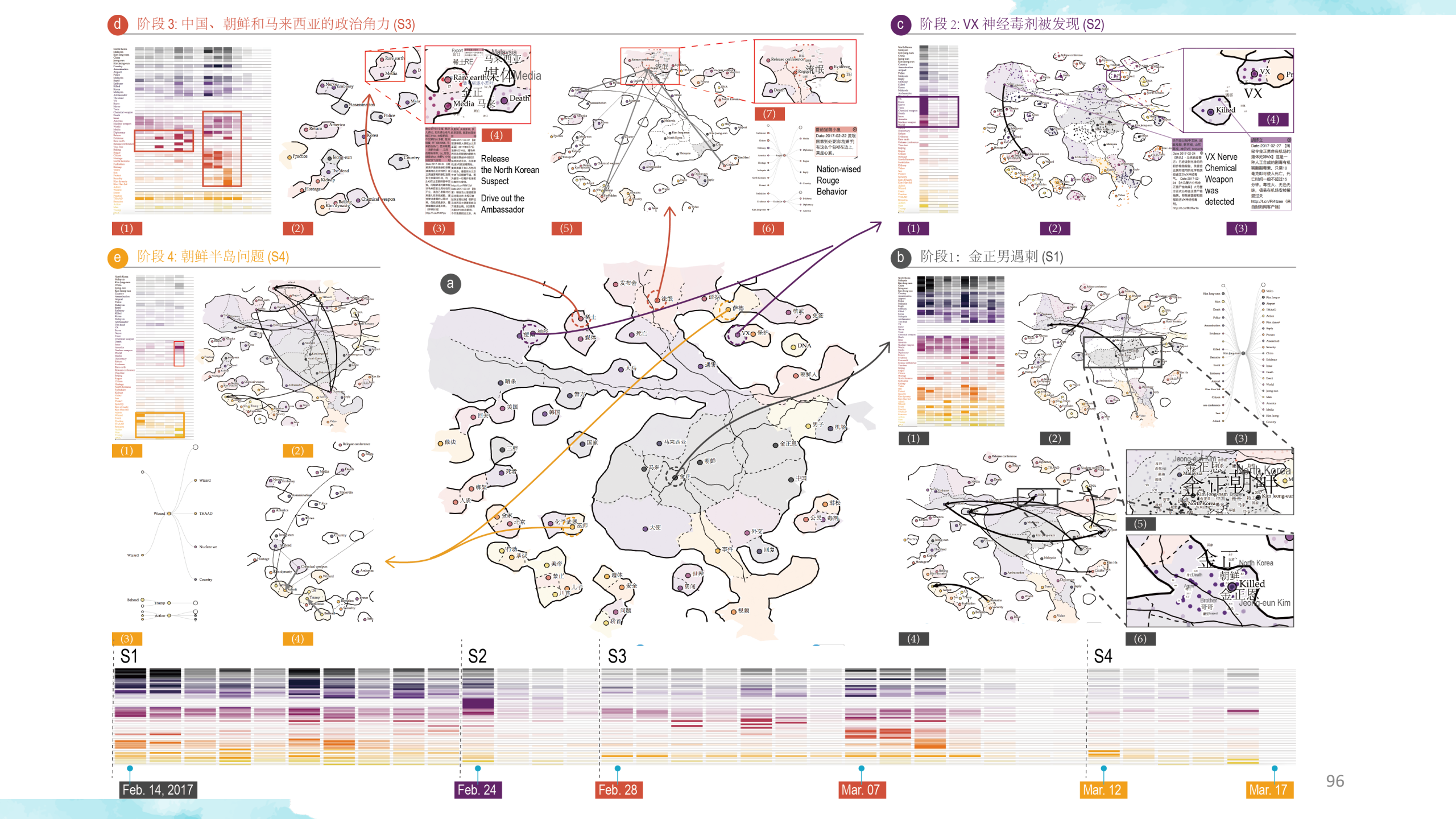


案例分析1 – 纱布门事件

案例分析2 – 金正男遇刺


发布会
流氓
证据
萨德
核武
免签
稀土
使馆
神经
媒体
死亡
VX
保护
DNA
遇害
大马
朝鲜人
暗杀
警方
美国
回去
韩国
男子
机场
做法
国家
马来西亚
金正恩
三胖
朝鲜
死者
马来
金正
中国
绑架
人质
韩松
金家
北京
化学武器
巫师
公民
毒剂
大使
外交
行动
承认
事件
回复
美帝
禁止
世界
遭体
安全
勇闯
儿子
川普
问题
视频
斩首


朝鲜
马来
金正
中国
三胖
金正恩
国家
暗杀
机场
警方
马来西亚
回复
使馆
遇害
韩国
大马
大使
死者
VX
勇闯
神经
毒剂
化学武器
死亡
回题
美国
核武
世界
媒体
外交
回去
证据
稀土
发布会
免签
北京
流氓
公民
人质
朝鲜人
禁止
绑架
视频
儿子
保护
安全
金家
韩松
承认
巫师
事件
做法
萨德
遭体
行动
男子
川普
DNA


d 阶段 3: 中国、朝鲜和马来西亚的政治角力 (S3)
c 阶段 2: VX 神经毒剂被发现 (S2)
e 阶段 4: 朝鲜半岛问题 (S4)
b 阶段1：金正男遇刺 (S1)
a
Release the North Korean Suspect
Drive out the Ambassador
Nation-wised Rouge Behavior
VX Nerve Chemical Weapon was detected
S1
S2
S3
S4
Feb. 14, 2017
Feb. 24
Feb. 28
Mar. 07
Mar. 12
Mar. 17

E-Map小结

- 一个基于社交媒体数据新颖的事件总结可视化形式
  - 自然地图
  - 灵活交互
- 一种探索社交媒体用户行为的时空可视分析方法
  - 动态时间切片
  - 人群移动轨迹与关联分析
- 基于社交媒体数据的真实事件分析

数据（侦探）科学家的素养

- 有数据的思维
  - 数据来源，分布特征，统计量
- 有科学的精神
  - 求证数据的背后的意义
  - 是否满足假说
- 有分析的目标
  - 清晰地解构你要分析的问题
  - 一步步地探索数据，求得结论
- 有实用的方法
  - 机器学习的算法，python的熟练掌握
  - 可视化的设计与实现，交互探索数据的能力

http://simingchen.me
http://fduvis.net
http://vis.pku.edu.cn