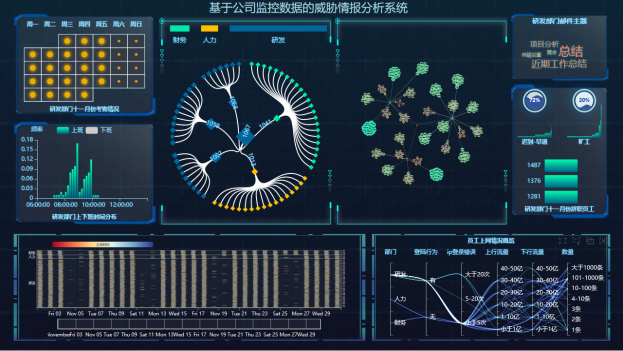


长春市长热线可视分析系统，基于的数据是长春12345热线投诉数据，我问为其做了8个可视分析系统帮助不同部门完成基础信息展示，预测，安全分析，政策分析。

以基础信息可视化系统为例，系统主要以统计信息的展示为主。系统由7个视图组成。具体来讲我们使用日历图展示投诉数量在不同月的分布，并提供选择月份操作。使用地图的热力图展示地区分布，使用树形图展示行业的层次结构，并用节点的大小代表投诉数量，使用文字云展示涉及单位的数量分布。考虑到用户群体我们设计的系统比较简单易于操作，可以看到那些月份，行业，单位，地区投诉数量分布。其他的系统主要也是展示为主，展示其他团队模型的结果。



另外一个值得说的系统是2018年我带队参加了Chinavis挑战赛并获得全局第一名。题目背景一家互联网高科技公司，出现情报泄露。请您设计并实现一套可视分析解决方案，有三个分析任务：

1. 分析公司内部员工所属部门及各部门的人员组织结构。
2. 分析该公司员工的日常工作行为，按部门总结和展示员工的正常工作模式。
3. 找出至少5个异常事件。

我们根据邮件主题使用随机森林算法将员工划分部门，接下来我们通过部门内部邮件往来确定每一个部门的组织结构。最后根据三个部门领导之间的邮件往来整合成整个公司的组织结构。在划分好部门后我们对每个部门分析它的日常模式，主要包括上下班时间，工作内容，tcplog和login情况，考勤情况四个方面。我们使用系统对不同类型的异常进行探索，包括登录异常，越权操作，隐藏上班记录，流量异常，辞职异常等，并分析其中存在的联系。

针对以上的需求我们设计并实现了一个完整的威胁情报分析系统。由基础板块和个人信息板块组成。

在基础板块中我们主要探索公司组织结构和部门日常模式。

时序热力图展示公司全体员工的30天的上下班的区间段；平行坐标图展示所有员工登录日志和TCPLOG日志信息；节点连接图展示部门内部的邮件往来；树形图展示整个公司的组织关系；日历图展示部门每天的上班人数；柱形图用于展示上下班时间段的分布情况；文字云展示部门常见邮件主题；水波柱形图展示部门的考勤情况。

同时在基础版块中我们找出感兴趣的员工在个人板块中进行细节探索，

阶梯瀑布图展示该员工本月上下班的情况；文字云展示该员工收发邮件的主题；流量图展示该员工上下行流量数据；极坐标堆叠柱状图展示多个协议登录成功和失败的数量； 列表视图展示用户浏览网页的记录；



题目背景：在一个虚拟的岛上，发生地震导致核工厂发生泄露。

数据：静态传感器和车载移动传感器的5天的核污染检测数据。

当然一些分析任务：

1、识别受污染的区域，不同区域测量值的时序变化。

2、分析传感器、区域的不确定性

3、找出重大事件，以及对不确定性的影响。

4、找出受污染的车

5、行动方案建议

为了解决上述的问题，我们也是集成了一个可视分析系统，由6个视图组成，包含丰富的交互，支持用户从概览到细节的分析。因为这个国际比赛还是偏学术的，我们在视图设计方面也非常注重创新性。简单介绍一下分析视图和分析流程：

折线图展示经过特定网格的传感器的检测值时序变化。

轨迹视图，通过动画展示，传感器地理空间的移动轨迹、和检测值的动态变化。其中一个point代表一个传感器，半径和内圆颜色编码污检测值。

传感器矩阵视图显示传感器的每小时辐射值平均值。由59\*120个矩形组成，X轴代表时间5天120小时。Y轴为59个传感器。其中一个矩形代表一个传感器某一小时的检测均值，颜色编码不同等级的检测值。一行矩阵代表一个传感器的5天120小时均值的变化。X轴就是按时间顺序，y轴的根据传感器的相似性排序。

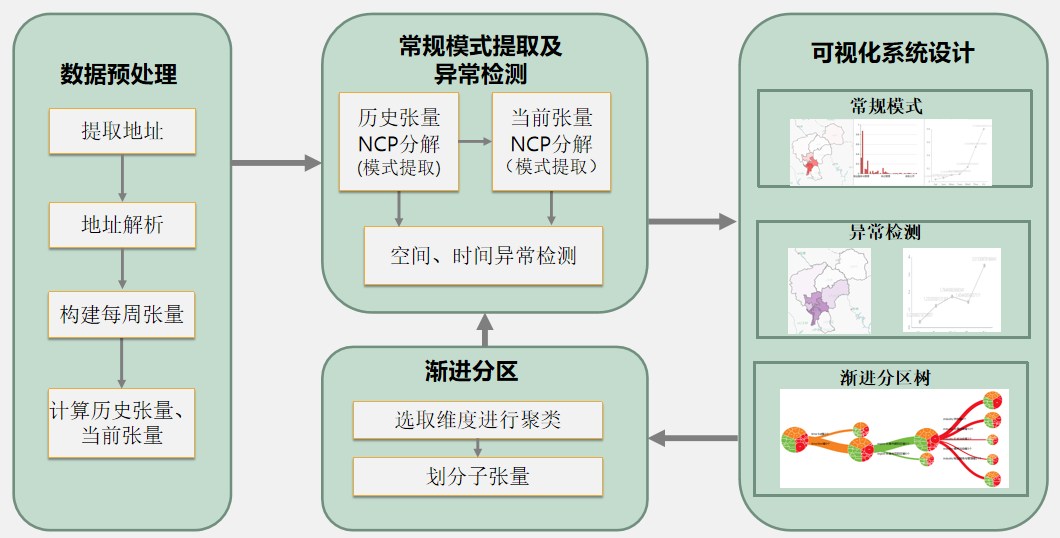
为了探究区域和传感器的不确定性，我们设计了传感器投影视图和网格概览视图。每个传感器和网格。分别为每个传感器和网格绘制不确定性glyph，描述了相应的均值，标准偏差和监测频次。

选择一个glyph后，它可以在细节检查视图中扩展，其中外环表示每小时的标准偏差，第二个环使用颜色显示每小时平均值，第三个图层中的径向条形图显示每小时监测频率。 第四层中的散点图显示每小时原始数据的分布，其中颜色编码点密度。

首先，在传感器矩阵视图中可以观察到，自4月8日上午10点开始，一些传感器开始监测持续的高辐射，因此我们推断第一次地震发生在这个时间段附近。

然后，刷选这个时间区间，并且可以在轨迹视图中找到高值主要集中在核电站周围的几个网格中。 同时，轨迹视图支持播放器的大部分功能，例如选择传感器，时间区间，倍速，暂停，播放，跳转等以便于详细探索。

之后我们可以选择感兴趣的传感器或网格进行细节视图的探索。



我的研究内容主要分成四块：数据预处理，常规模式提取与异常检测，可视化系统设计，渐进分区。数据预处理模块，因为我们用的数据是市长热线数据，我们需要在投诉内容中提取地址，再根据地址获取准确的经纬度。然后每一周（time epoch）构建一个三阶的张量，然后我们需要计算得到两个张量，一个是历史张量，一个是当前的张量。我们使用非负CP分解对两个张量进行分解提取常规模式，根据当前张量分解结果与历史张量的分解结果的差异对当前张量进行空间、时间异常检测。可视化系统一方面展示张量的常规模式，另一方面展示异常检测结果。因为对于全局张量进行常规模式提取和异常检测往往不能得到很好地效果，我们这里采用一种渐进式的思想，将大的张量不断划分成多个子张量，再对子张量分别进行探索。使用树型图展示渐进分区结果。

**研究背景与意义**

城市时空数据与我们每个人息息相关，对其进行常规模式提取与异常检测至关重要，能够辅助智慧城市建设，为市民生活保驾护航。

例如：政府工作人员根据市长热线的常规模式和最近时间异常情况，便可协调相关部门及时进行处理，防止事件恶化，改善民生。

但是城市数据通常具有多个维度，不同维度存在复杂的依赖关系，而且数据量巨大。传统基于统计学和及机器学习方法主要存在两个缺陷：

1、对常规模式和异常不能提供有效的解释。

2、大多数方法缺乏处理多维的能力。

所以我们引入可视分析方法。

**数据预处理：**

数据：

2016年长春市44个行业有大约30万条热线电话。

属性：时间，投诉内容，部门，行业，受理人等15个属性。

地址解析：

根据投诉内容提取地址，调用百度地图地址解析接口，将地址转化为经纬度。

构建每周张量：

2016年共有52周，每周一个张量（7天\*44行业\*10个行政区）—>热线数量

（行政区：朝阳,南关,宽城,二道,绿园,双阳,九台,德惠,农安,榆树）

2016年共有52周，每周一个张量（7天\*44行业\*1283个网格）—>热线数量

计算历史/当前张量

选取前42周张量加权平均得到历史张量Xhistory。

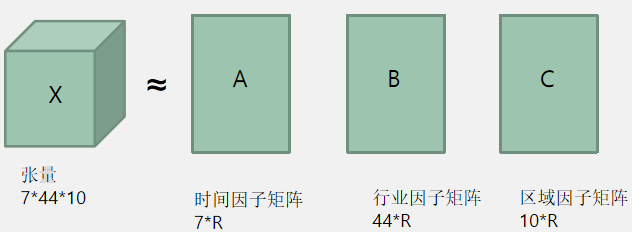
选取第43周张量作为当前张量Xcurrent。

**常规模式提取**

CP分解(CANDECOMP/PARAFAC Decomposition)

可以理解为对高维数据的一种低秩逼近或特征提取。通过分解可以得到R个模式，每个模式在三个维度都有一个特定分布，整理可以得到三个因子矩阵。

CP分解加非负约束即为NCP



A每一列代表该模式的时间分布，

每一行代表在该时间段，不同模式所占数量比例。

B每一列代表该模式的行业的分布，

每一行代表在该行业，不同模式所占数量比例。

C每一列代表该模式的区域的分布，

每一行代表在该区域，不同模式所占数量比例。

**异常检测**

时间i的异常分数

Ti=Distance(Ahistory[i,:],Acurrent[i,:])

区域i的异常分数

Si=Distance(Chistory[i,:],Ccurrent[i,:])

**渐进划分**

