



技术：应用场景下的 数据基本分析流程和分析方法

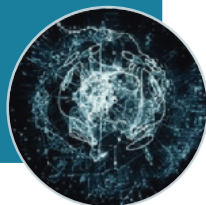
夏菁

浙江大学CAD&CG国家重点实验室
可视化与可视分析小组

内容大纲

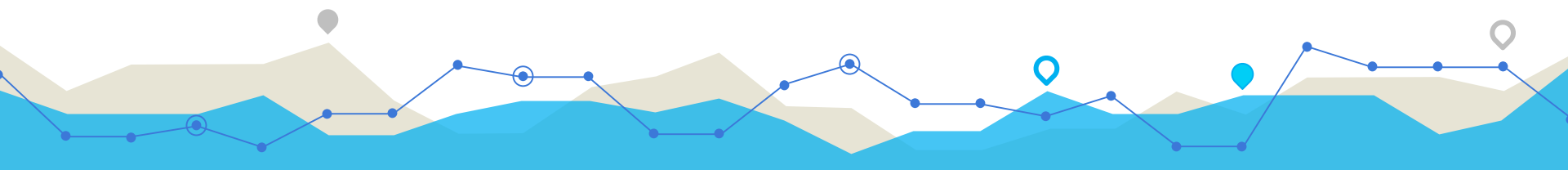
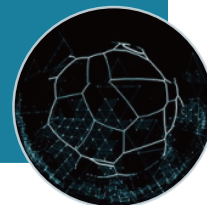
- 数据属性
- 数据质量
- 数据处理

基础



- 时序排名
- 数据清洗
- 统一数据平台

应用

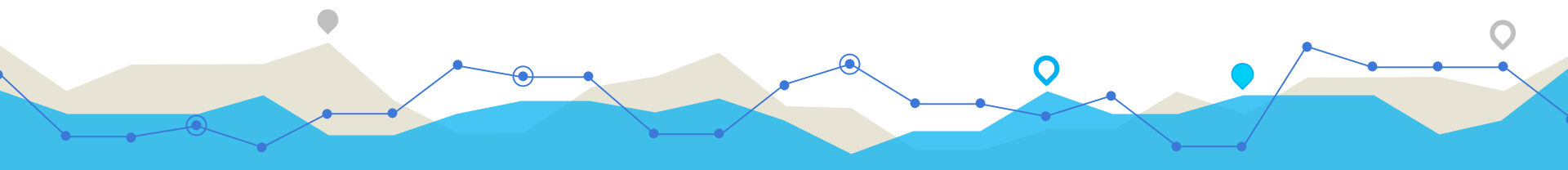




基础 1

数据属性

- 也可以称为变量、维度、特征等
 - 风速：大小(标量)和方向(多值矢量)
- 基本数据类型
 - 有序型
 - 时序、文本序列等
 - 类别型
 - 省
 - 数值型
 - 温度



数据质量

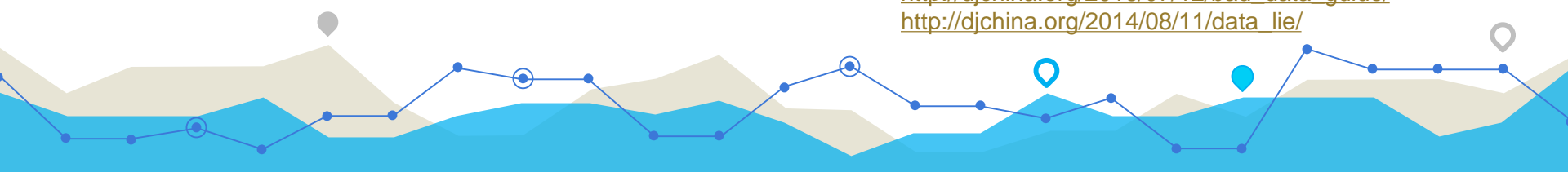
V	有效性	数据是否真实有效
A	准确性	数据是否精确，有无误差
B	可信性	数据来源和收集方式是否可信
I	一致性	数据(格式、单位等)是否一致
C	完整性	数据是否有缺失
T	时效性	数据保质期(相对分析任务)

Kandel S, Heer J, Plaisant C, et al. Research directions in data wrangling: Visualizations and transformations for usable and credible data[J]. Information Visualization, 2011, 10(4): 271-288.

The Quartz 坏数据手册

http://djchina.org/2016/07/12/bad_data_guide/

http://djchina.org/2014/08/11/data_lie/



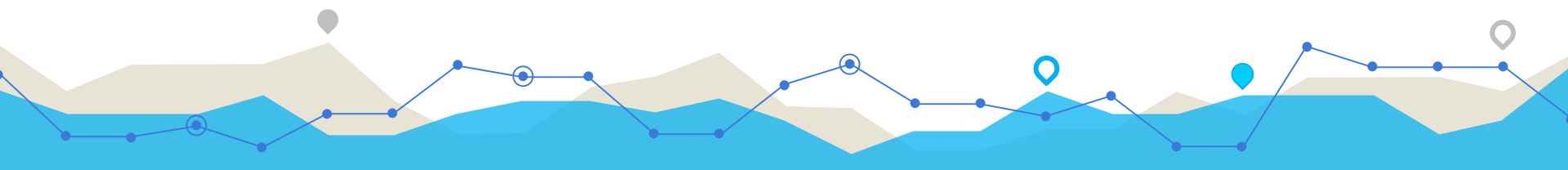
数据处理 —— 统计

方法

- 数据分布
 - 均值、方差、众数、分位数
- 回归方法
 - 线性回归
 - 逻辑回归
- 多元统计分析
 - 变量相关性

工具

- 软件类工具
 - SPSS
 -
- 脚本类工具
 - MATLAB
 - R
 - Python



数据分布

- 均值
 - 平均值

- 方差 $\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$

有序数列

3, 6, 7, 8, 8, 10, 13, 15, 16, 20

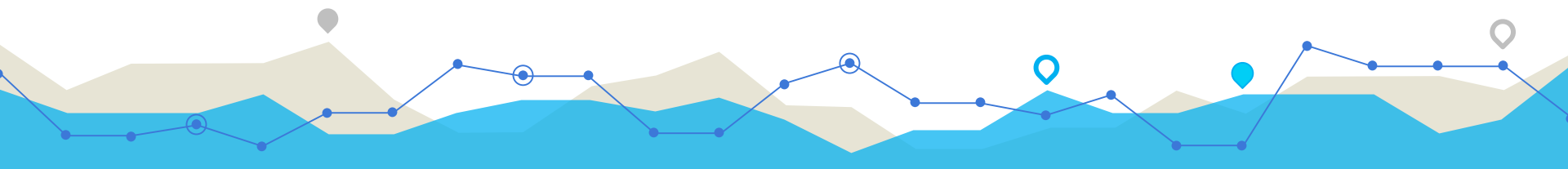
均值: 10.6

方差: 24.84

众数: 8

四分位数: 3、7、9、15、20

- 众数
 - 个数最多的值
- 分位数
 - 有序数列的第1, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ 个值和最后一个值



回归方法

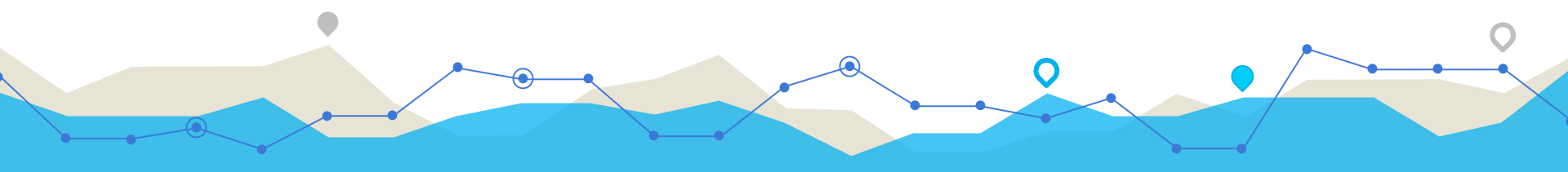
- 线性回归

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- 逻辑回归

$$P(Y = 1|X = x) = \frac{e^{x'\boldsymbol{\beta}}}{1 + e^{x'\boldsymbol{\beta}}}$$



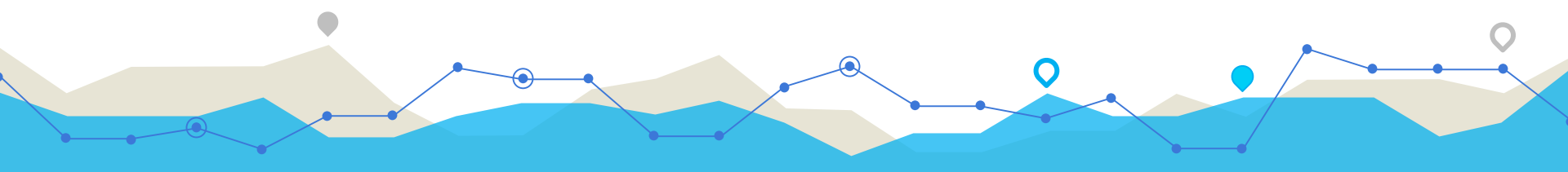
多元统计分析

- 协方差 $\text{cov}(X, Y) = E((X - \mu)(Y - \nu))$

- 相关系数 $\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$

变量独立 \Leftrightarrow 不相关 \Leftrightarrow 协方差为0 \Leftrightarrow 相关系数为0

- 互信息 $I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x) p(y)} \right)$



数据处理 —— 降维(略)

降维的本质：

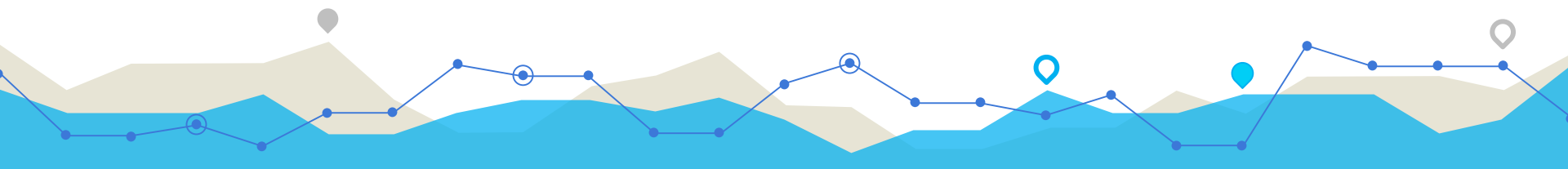
使数据在低维的距离尽量与在高维的距离保持一致

线性

- PCA、MDS、NMF、.....

非线性

- LLE、Iso-map、SOM、.....



数据处理 —— 相似度量

类别型

集合(杰卡德)
相似度

海明距离

高维数值型

曼哈顿距离
(L1范数)

欧氏距离
(L2范数)

夹角余弦

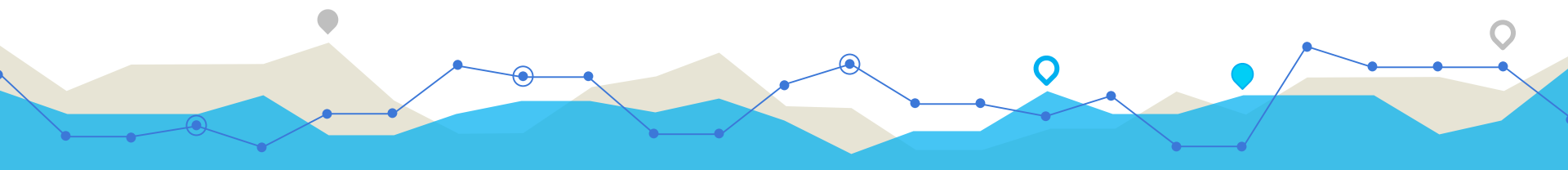
有序型

皮尔逊相关
系数

动态时间扭
曲

最大公共子
序列

自定义距离



类别型相似性度量

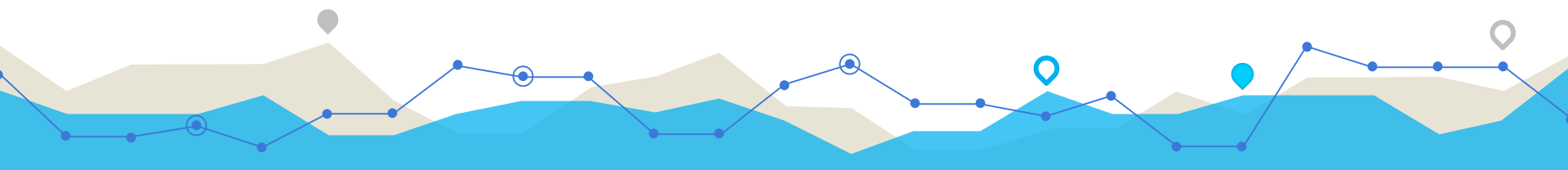
- 集合(杰卡德)相似度
 - 两个集合内容的相似性
 - 交集/并集

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- 海明距离
 - 对应位置不同字符的个数

1011101

1001101



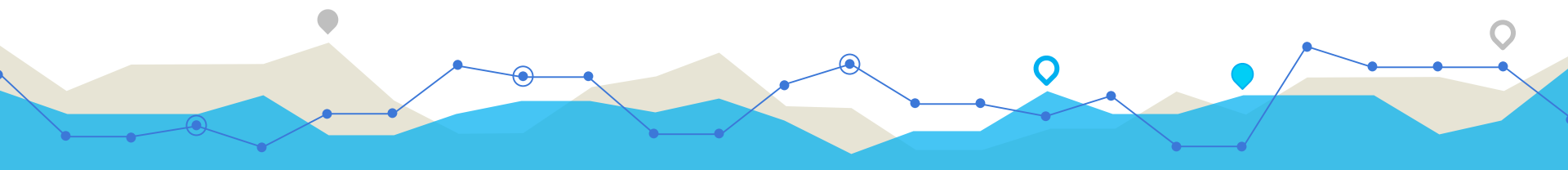
高维数值型相似性度量

- 曼哈顿距离(L1范数) $\sum_{i=1}^n |x_i - y_i|$

- 欧式距离(L2范数) $\left(\sum_{i=1}^n |x_i - y_i|^2 \right)^{1/2}$

- 夹角余弦

$$\cos(\theta) = \frac{\sum_{k=1}^n x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^n x_{1k}^2} \sqrt{\sum_{k=1}^n x_{2k}^2}}$$

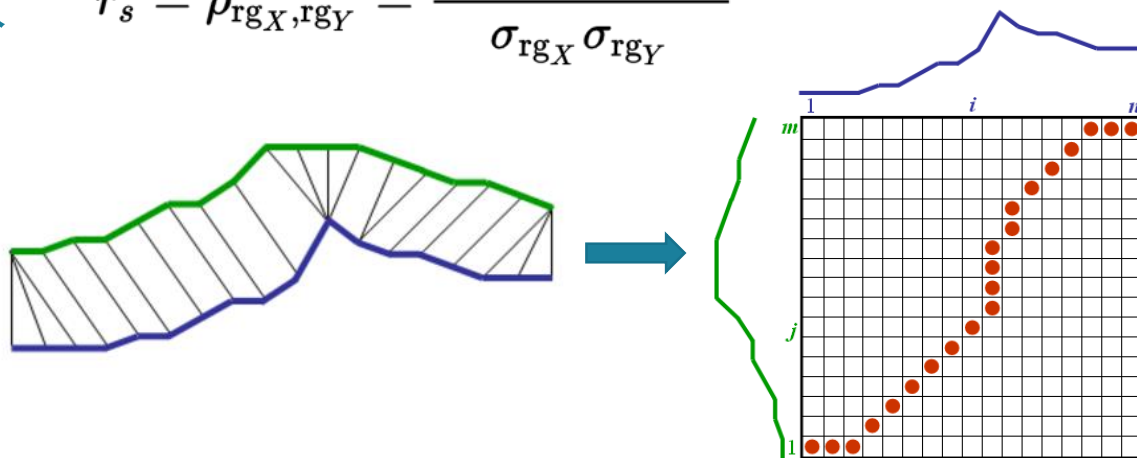


有序型相似性度量

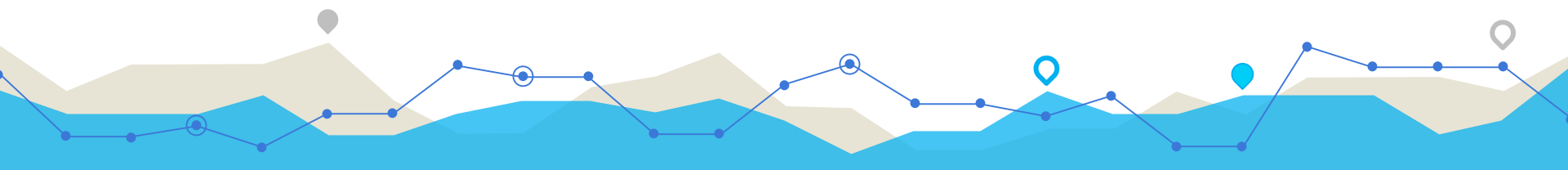
- 斯皮尔曼相关系数

$$r_s = \rho_{\text{rg}_X, \text{rg}_Y} = \frac{\text{cov}(\text{rg}_X, \text{rg}_Y)}{\sigma_{\text{rg}_X} \sigma_{\text{rg}_Y}}$$

- 动态时间扭曲

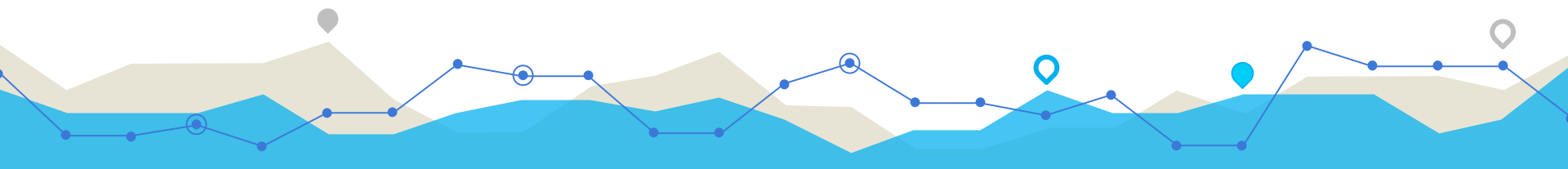


- 最大公共子序列



数据处理 —— 聚类

- K-均值家族
- 层次聚类
 - 自底向上
 - DBSCAN算法
 - 自顶向下
 - Graph-cut算法



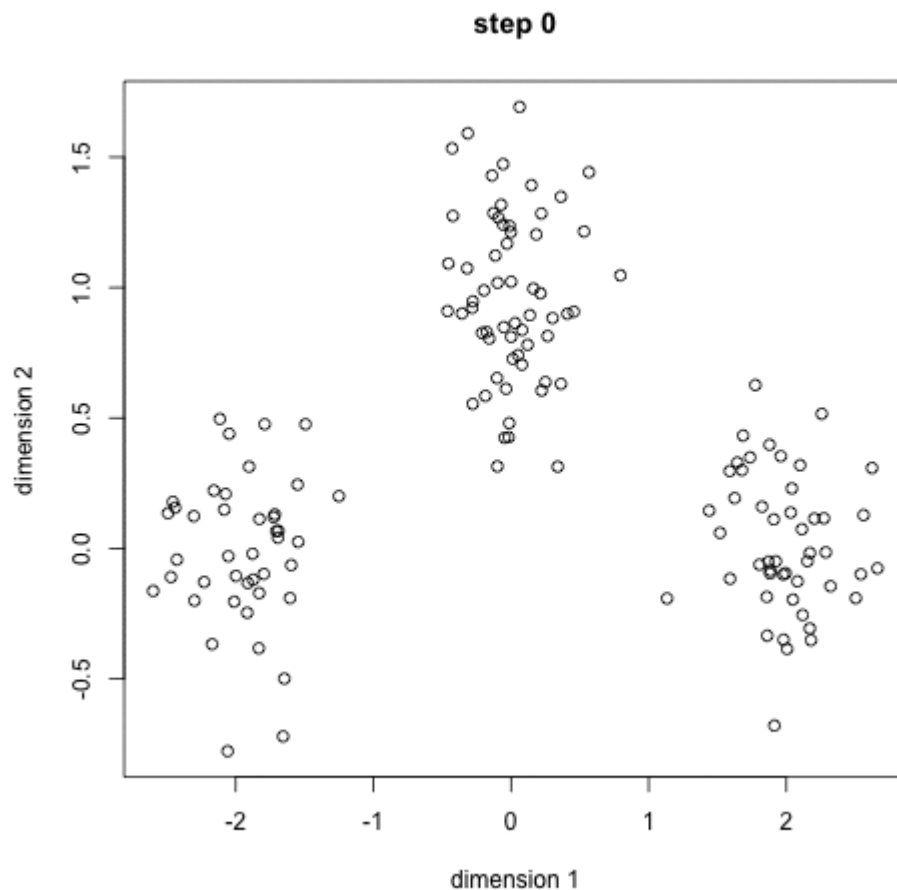
K均值

- K-means

- 随机定k个聚类中心
- 指定所有数据点类别
- 计算均值为新的聚类中心

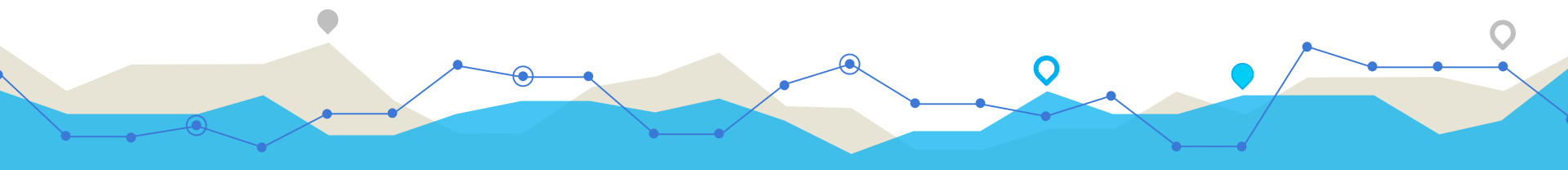
- K-medoids

- 聚类中心为数据点
- 适用于虚拟节点无法计算距离的情况



层次聚类

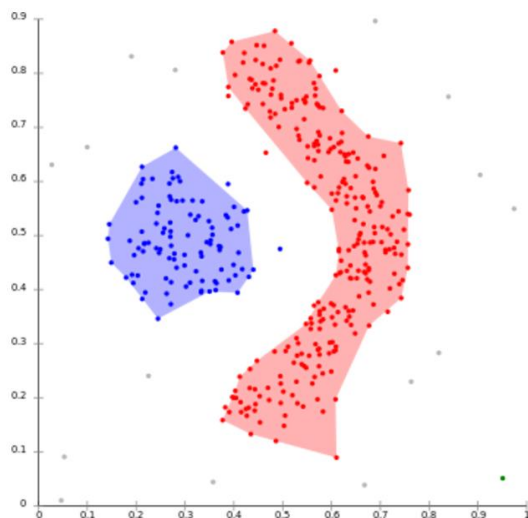
- 自底向上
 - DBSCAN
- 自顶向下
 - Graph-cut算法



DBSCAN

- 基于密度的带噪声空间聚类

Density-based spatial clustering of applications with noise

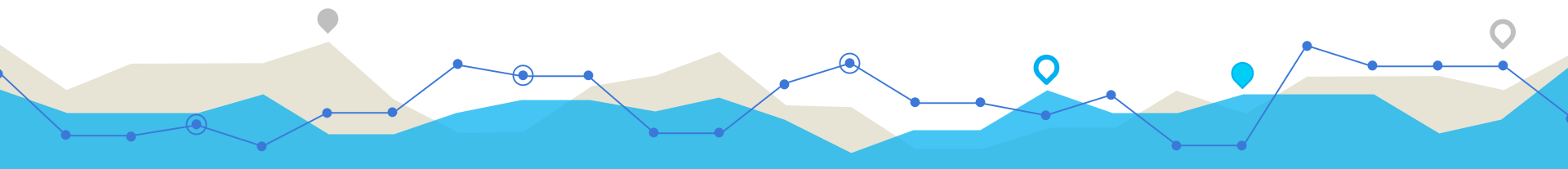
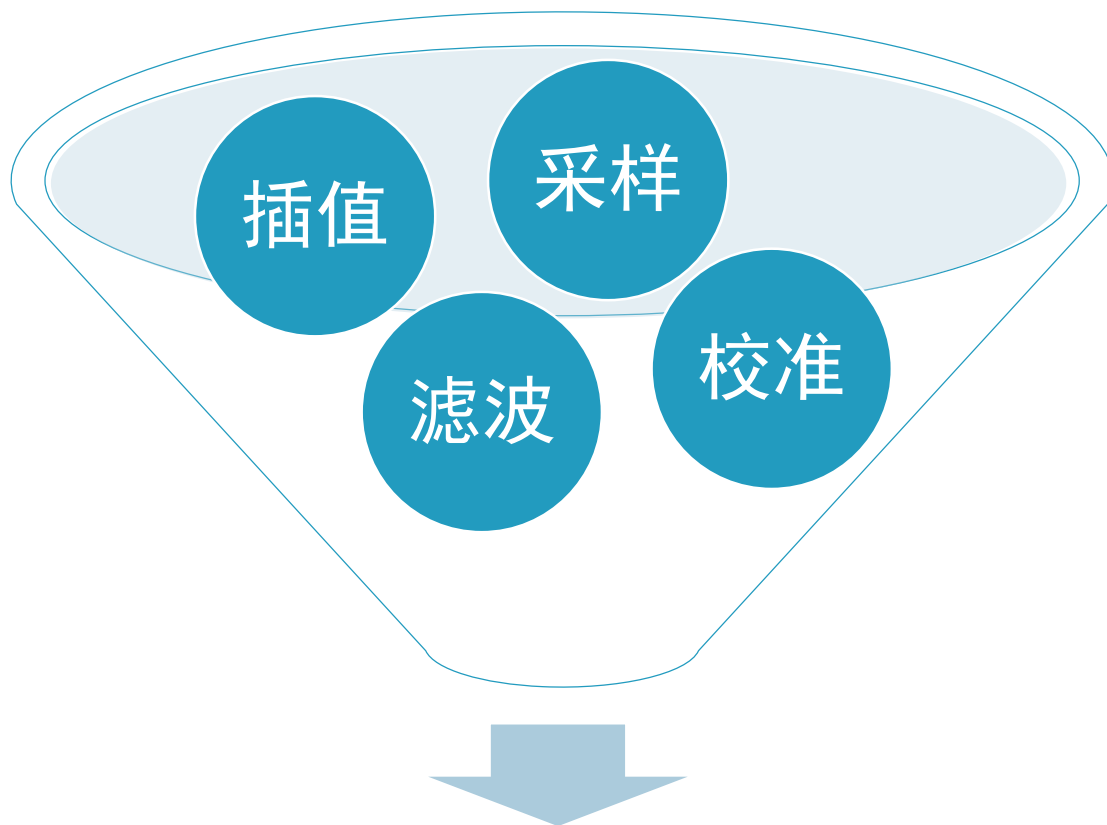


```
DBSCAN(D, eps, MinPts) {
  C = 0
  for each point P in dataset D {
    if P is visited
      continue next point
    mark P as visited
    NeighborPts = regionQuery(P, eps)
    if sizeof(NeighborPts) < MinPts
      mark P as NOISE
    else {
      C = next cluster
      expandCluster(P, NeighborPts, C, eps, MinPts)
    }
  }
}

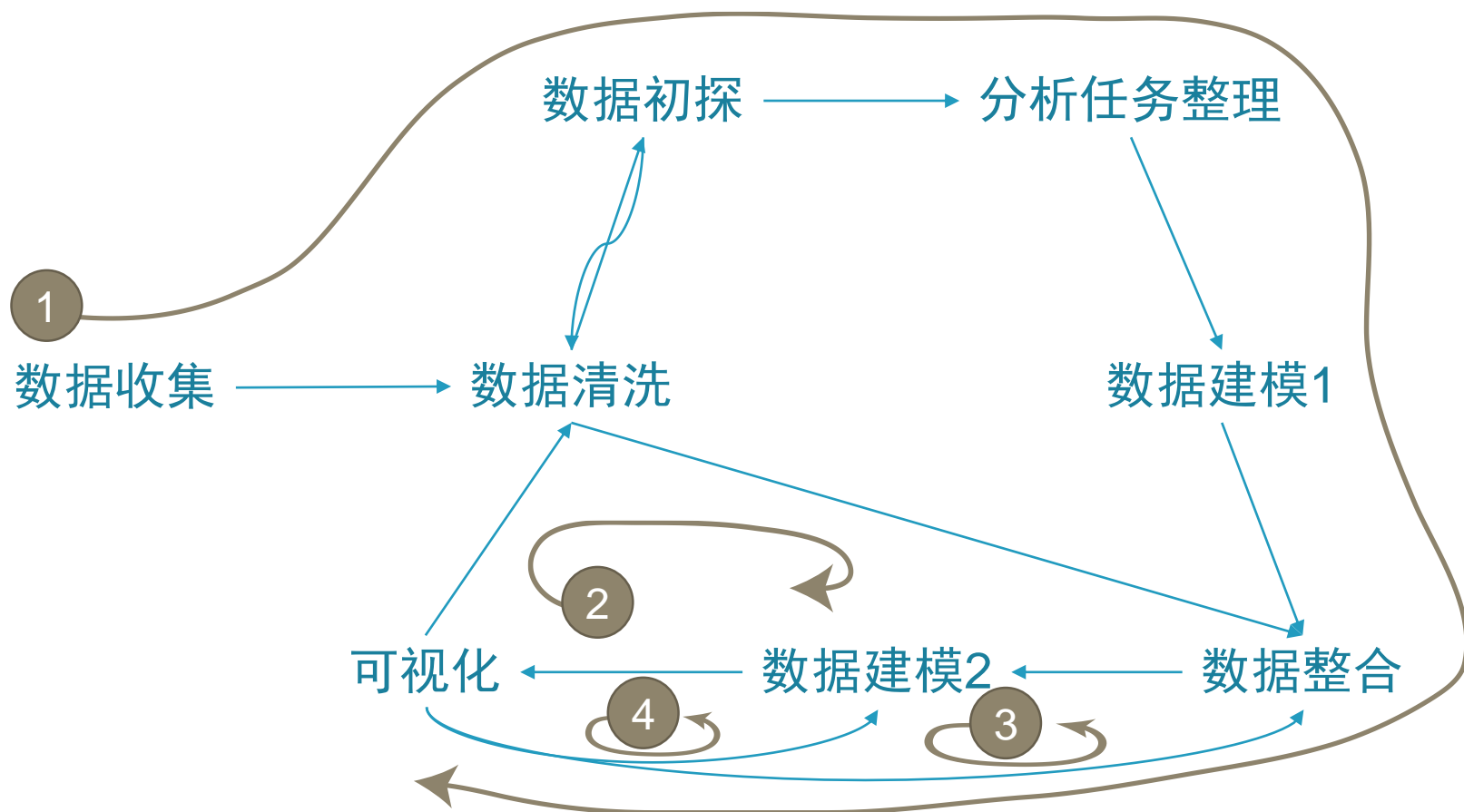
expandCluster(P, NeighborPts, C, eps, MinPts) {
  add P to cluster C
  for each point P' in NeighborPts {
    if P' is not visited {
      mark P' as visited
      NeighborPts' = regionQuery(P', eps)
      if sizeof(NeighborPts') >= MinPts
        NeighborPts = NeighborPts joined with NeighborPts'
    }
    if P' is not yet member of any cluster
      add P' to cluster C
  }
}

regionQuery(P, eps)
  return all points within P's eps-neighborhood (including P)
```

数据处理 —— 其它



数据分析流程

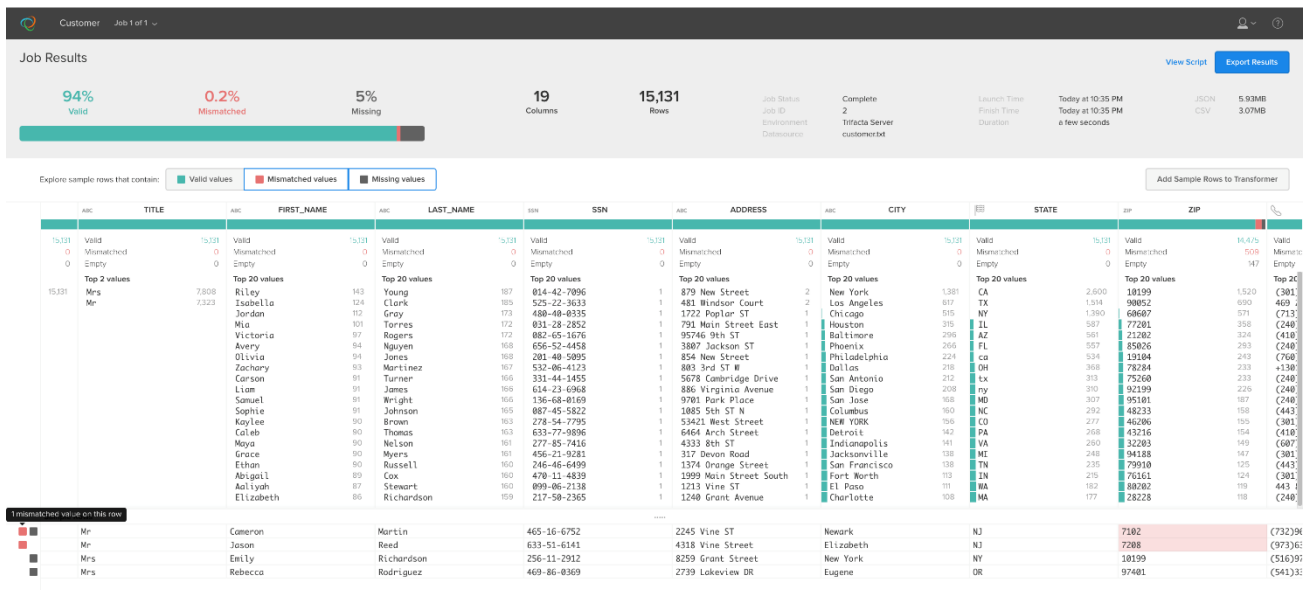




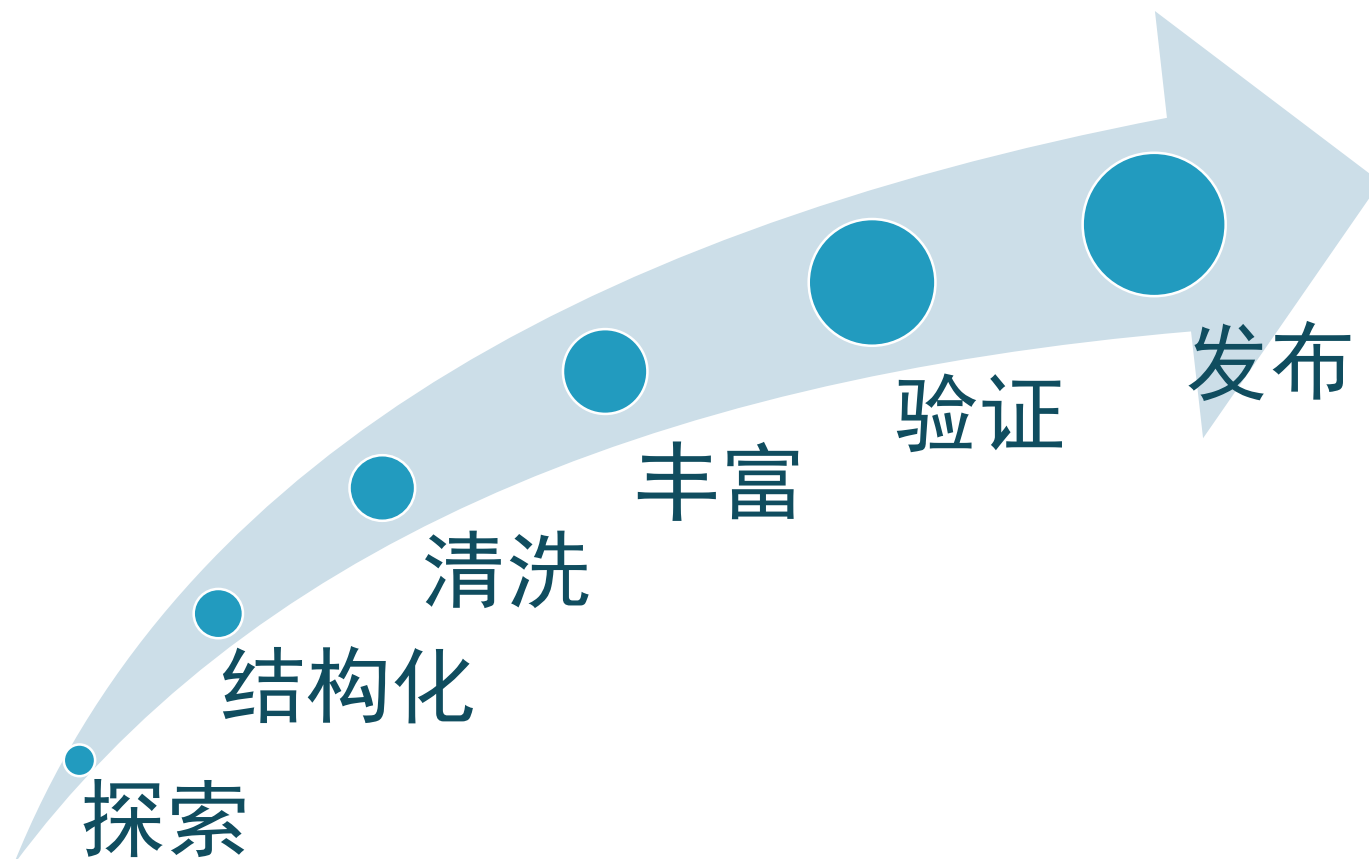
应用 2

免费数据清洗工具

- Trifacta (Data Wrangler)
- OpenRefine (Google Refine)



Trifacta —— 可视数据清洗工具



The screenshot shows a web browser window with the URL `https://23gateest.cloud.trifacta.com/datasets`. The page title is "Trifacta - Datasets". The main header features the Trifacta logo and navigation tabs for "DATASETS", "JOBS", and "SOURCES". A user profile icon and a help icon are also present. Below the header, the "Datasets" section displays "1 Dataset". A search bar labeled "Filter Datasets" and a "Create New Dataset" button are available. A table lists the dataset with columns: "Dataset Name", "Project Name", "Created", "Updated", "Datasources", and "Jobs". The table contains one entry: "USDA Farmers' Markets" with creation and update times of "Today at 3:20 PM". The "Datasources" column shows "USDA_Farmers_Market_201" with a count of "0". An "Action" dropdown menu is visible next to the dataset name, with a "Transform" button. The bottom of the image features a decorative graphic with a blue line graph and a tan area chart.

Trifacta - Datasets

<https://23gateest.cloud.trifacta.com/datasets>

TRIFACTA

DATASETS JOBS SOURCES

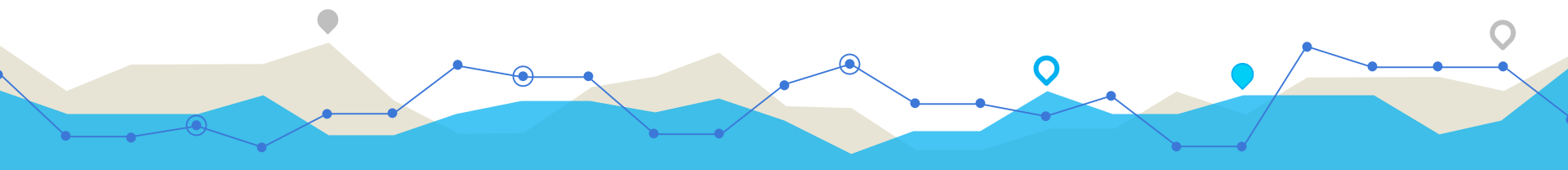
Datasets 1 Dataset

Filter Datasets Create New Dataset

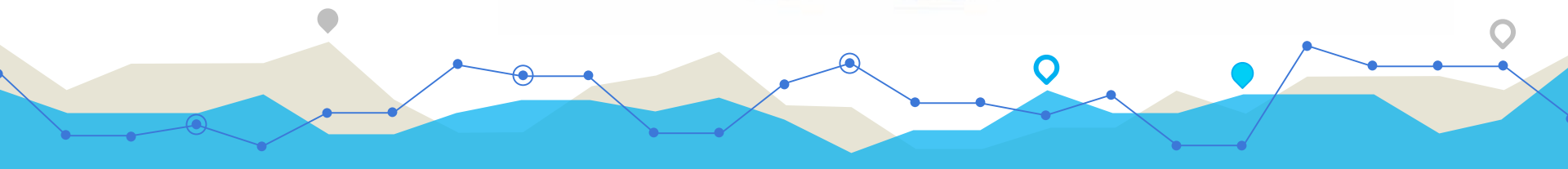
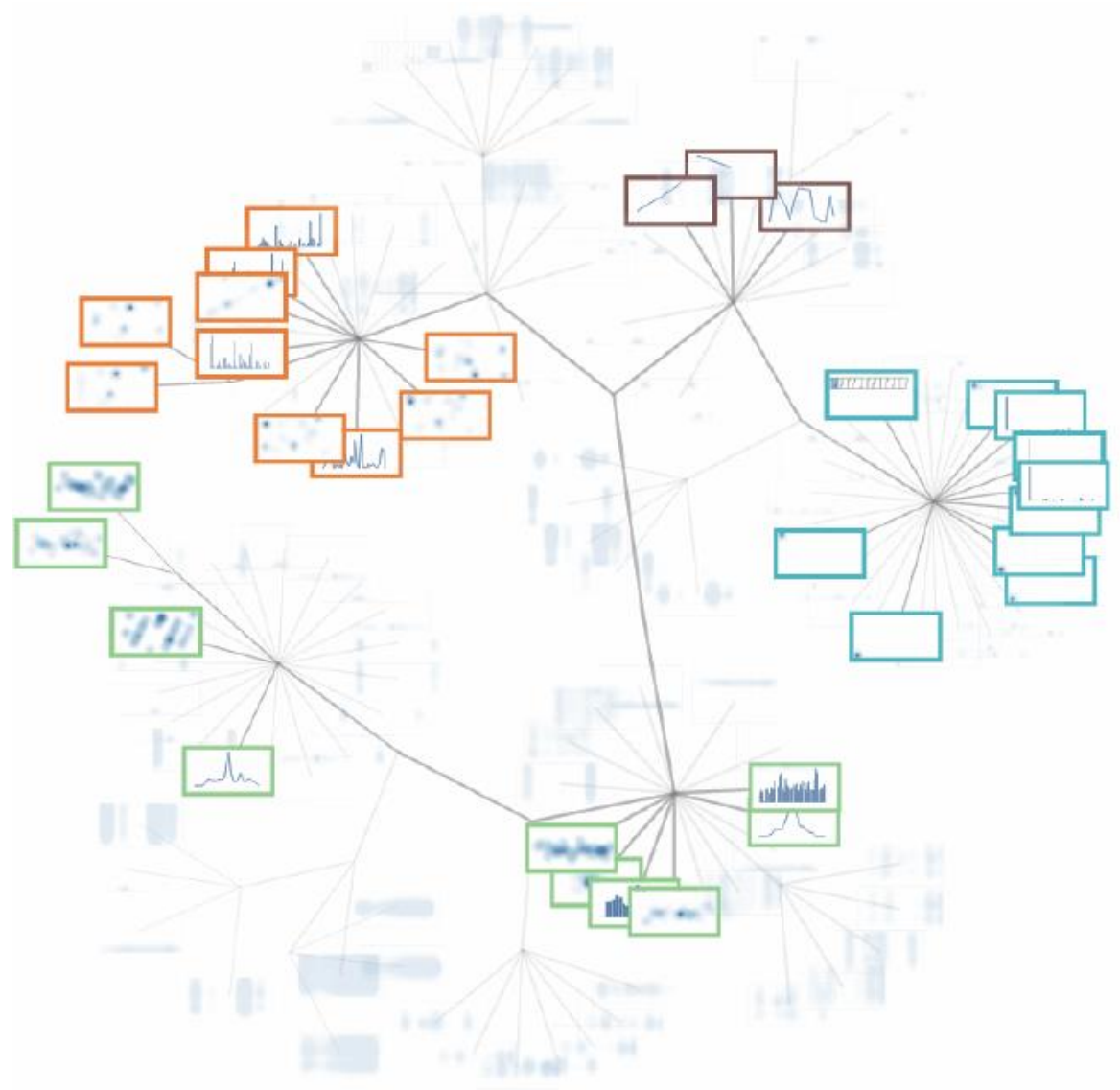
Dataset Name	Project Name	Created	Updated	Datasources	Jobs
USDA Farmers' Markets		Today at 3:20 PM	Today at 3:20 PM	USDA_Farmers_Market_201 0	Action Transform

数据初探 —— DimScanner

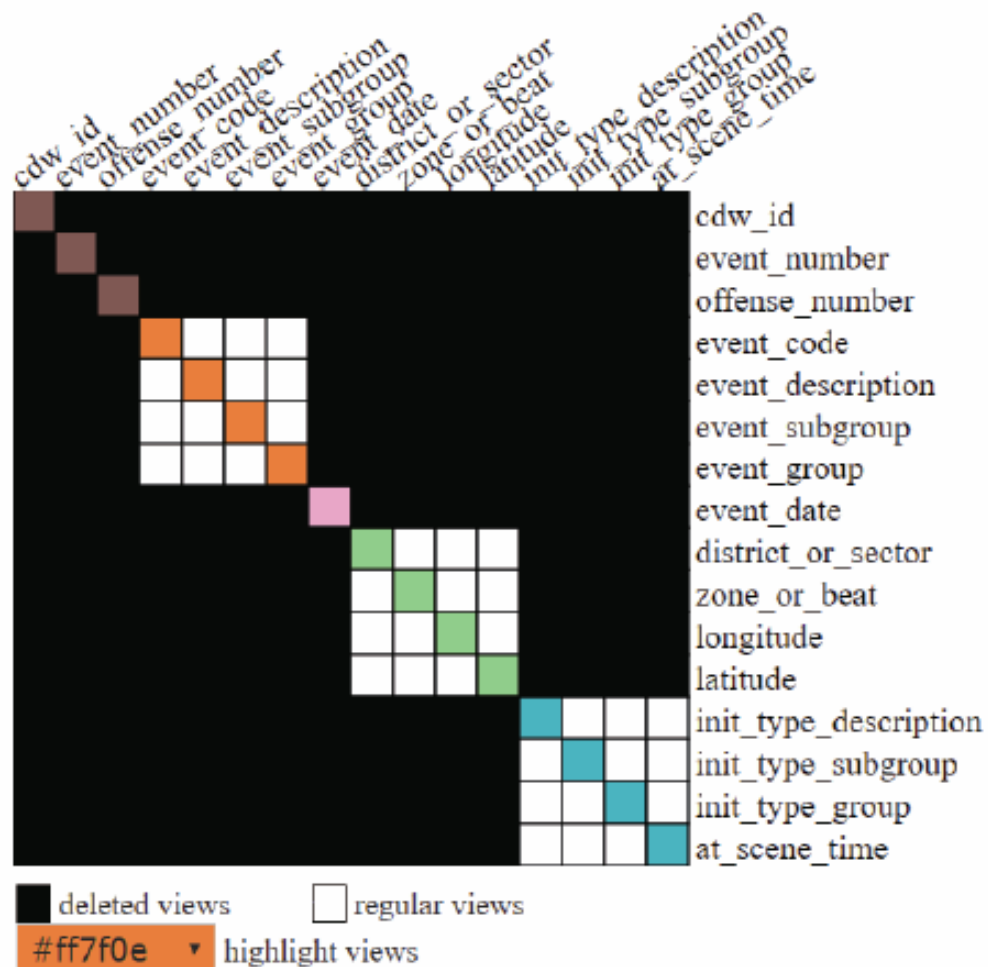
- 数据
 - 西雅图911报警数据
- 任务
 - 初探数据维度相关性



预览树

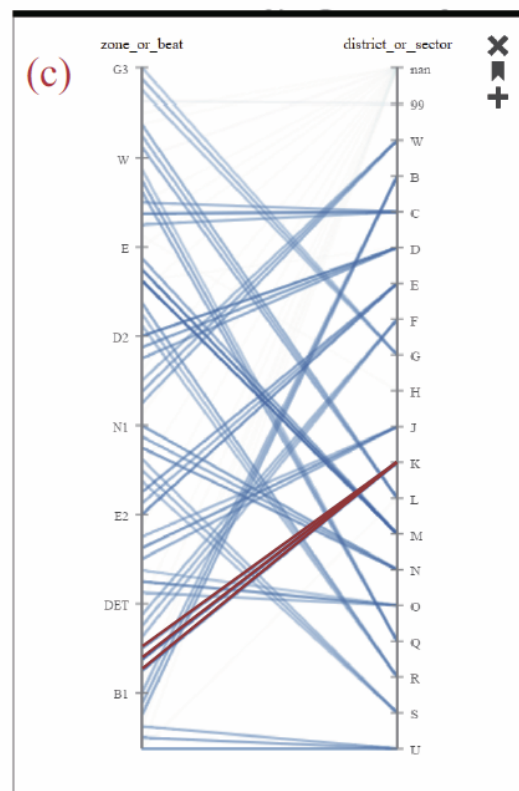
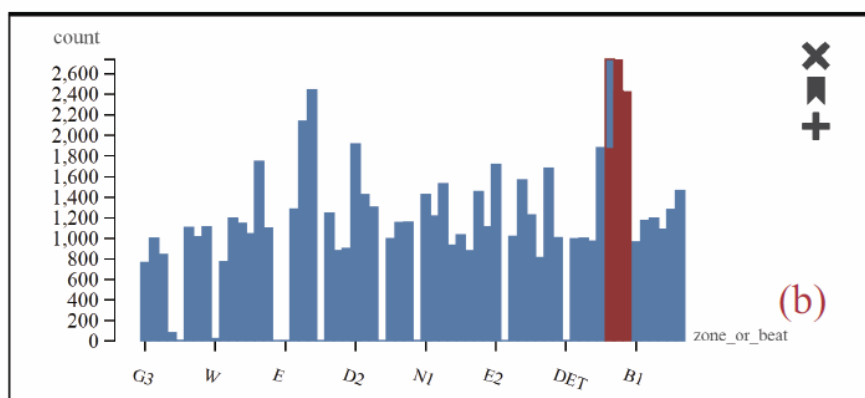
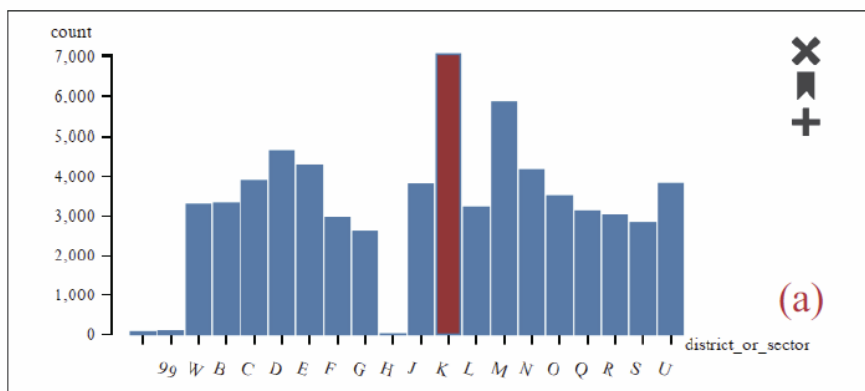


辅助交互



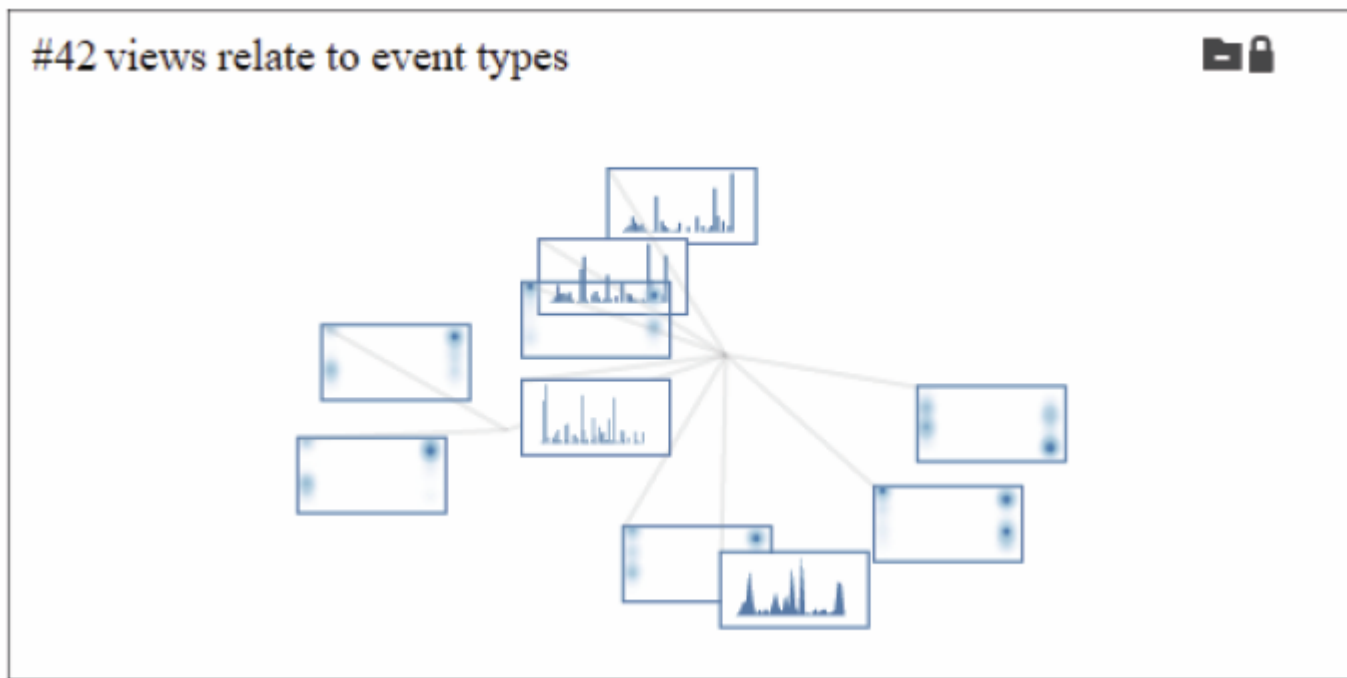
维度选择

辅助交互

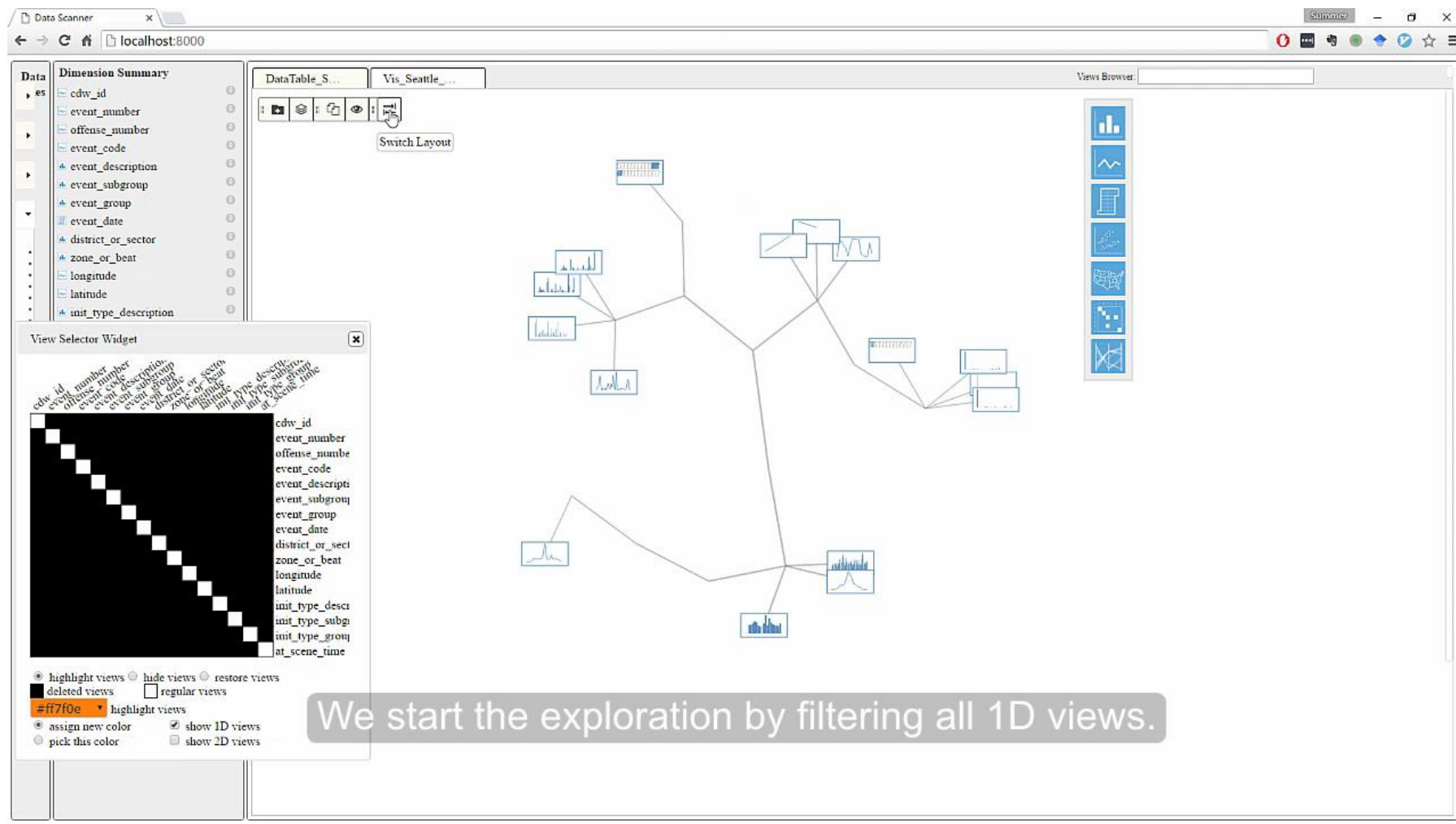


数据分布

辅助交互













数据概括

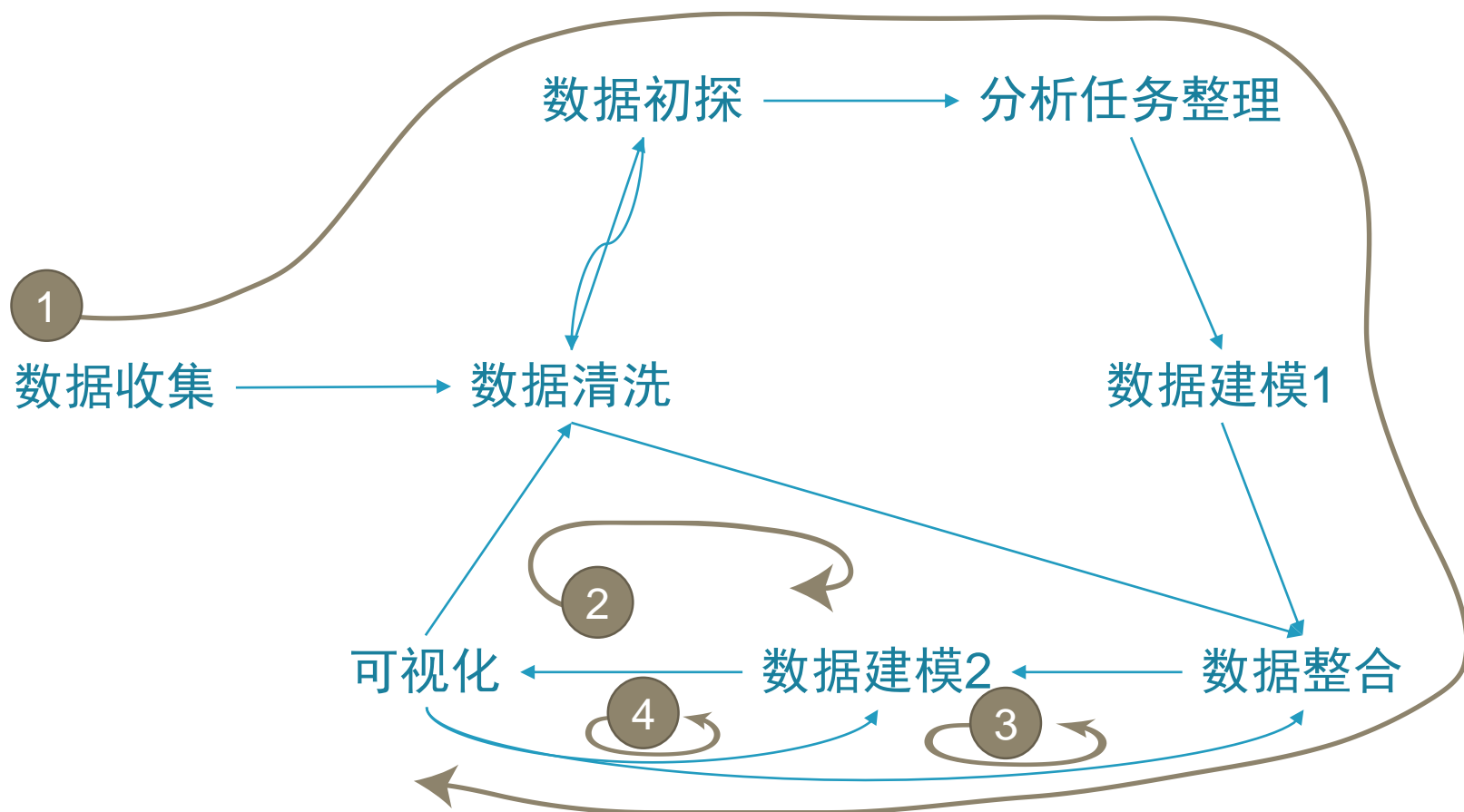


维基百科热词的时序排名可视化

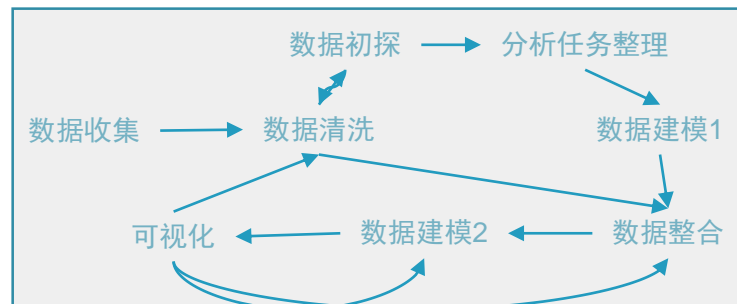
- 数据：维基百科时序排名
- 数据属性
 - 时间
 - 排名
 - 关系
- 任务
 - 可视化topk页面的走势
 - 分析topk页面之间的关系

Rank	Article	Class	Views	Image
1	<i>Pokémon Go</i>		4,778,652	
2	<i>Theresa May</i>		1,738,109	
3	<i>Mike Pence</i>		1,651,153	
4	<i>Sultan (2016 film)</i>		1,220,923	
5	<i>UFC 200</i>		1,139,080	

数据分析流程



数据处理



- 数据收集

- 维基百科访问日志

<https://dumps.wikimedia.org/>

- 数据清洗

- 删除乱码数据

- 数据初探

- 聚合页面点击率

- 数据清洗

- 删除index等无效页面

- 分析任务整理

- 可视化topk页面的走势
 - 分析topk页面之间的关系

- 数据整合

- 取每日的top-1000页面
 - 调用维基百科API获取页面跳转词，存储到数据库

- 数据建模

- 基于DTW的时序排名相似性建模

数据收集

- 维基百科访问日志

<https://dumps.wikimedia.org/other/pagecounts-raw/>

Page view statistics for Wikimedia projects

(For up-to-date information (outages, ...) about this dataset, please consult the [dataset's wiki page](#).)

Pagecount files per year

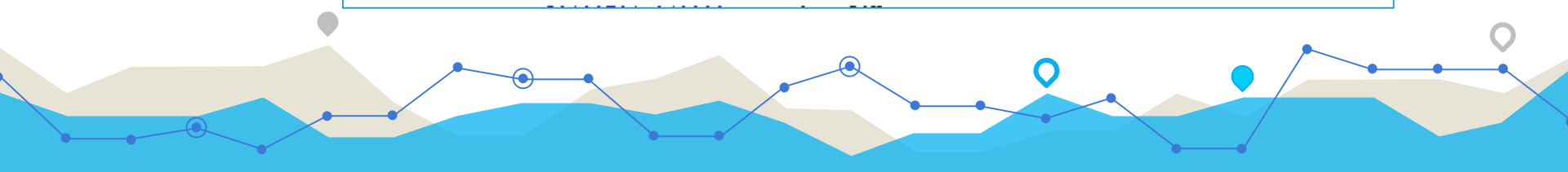
- [2007](#)
- [2008](#)
- [2009](#)
- [2010](#)
- [2011](#)
- [2012](#)
- [2013](#)
- [2014](#)
- [2015](#)
- [2016](#)

Index of page view statistics for 2016-07

Pagecount files for 2016-07

Check the [hashes](#) after your download, to make sure your files arrived intact.

- [pagecounts-20160701-000000.gz](#), size 72M
- [pagecounts-20160701-010000.gz](#), size 84M
- [pagecounts-20160701-020000.gz](#), size 90M
- [pagecounts-20160701-030000.gz](#), size 85M



数据整合

MySQL



每日top-1000页面

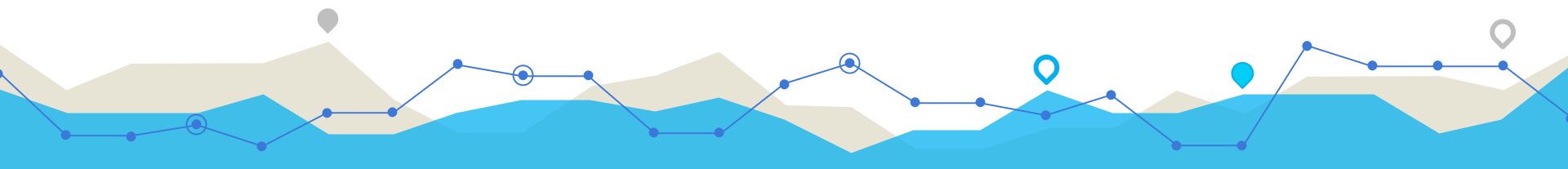
Neo4J



Pagelink关系页面

https://en.wikipedia.org/w/api.php?action=query&format=json&titles=The_Big_Bang_Theory&prop=links&pllimit=max

```
{"ns":0,"title":"Emmy Award"}  
{"ns":0,"title":"Entertainment Weekly"}  
{"ns":0,"title":"Euglossa bazinga"}  
{"ns":0,"title":"Experimental physics"}  
.....
```

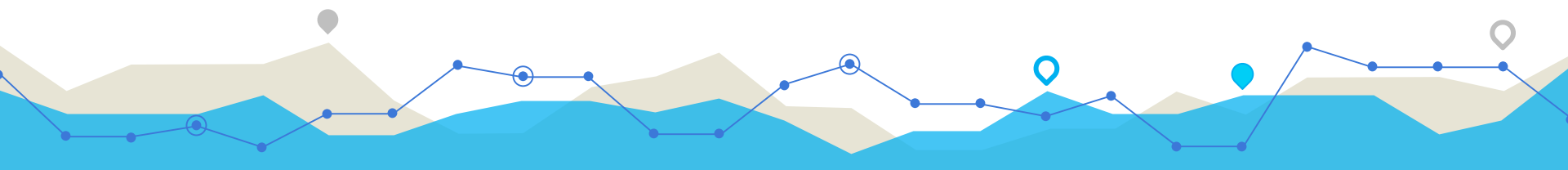


数据建模

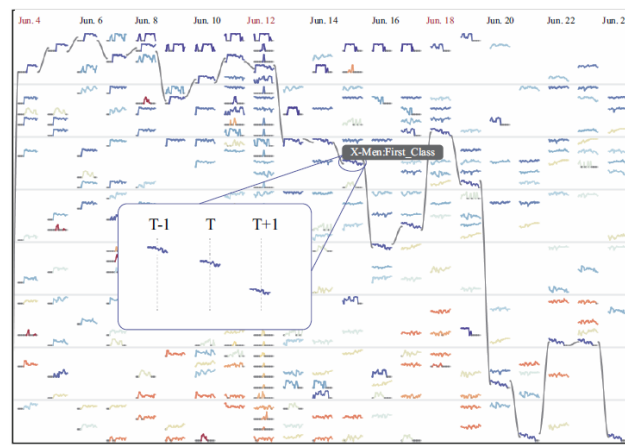
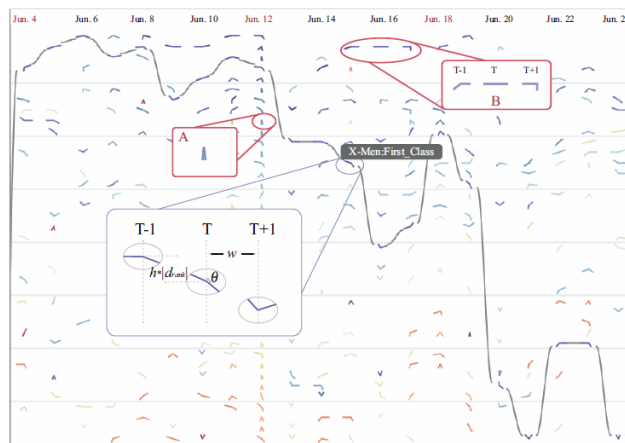
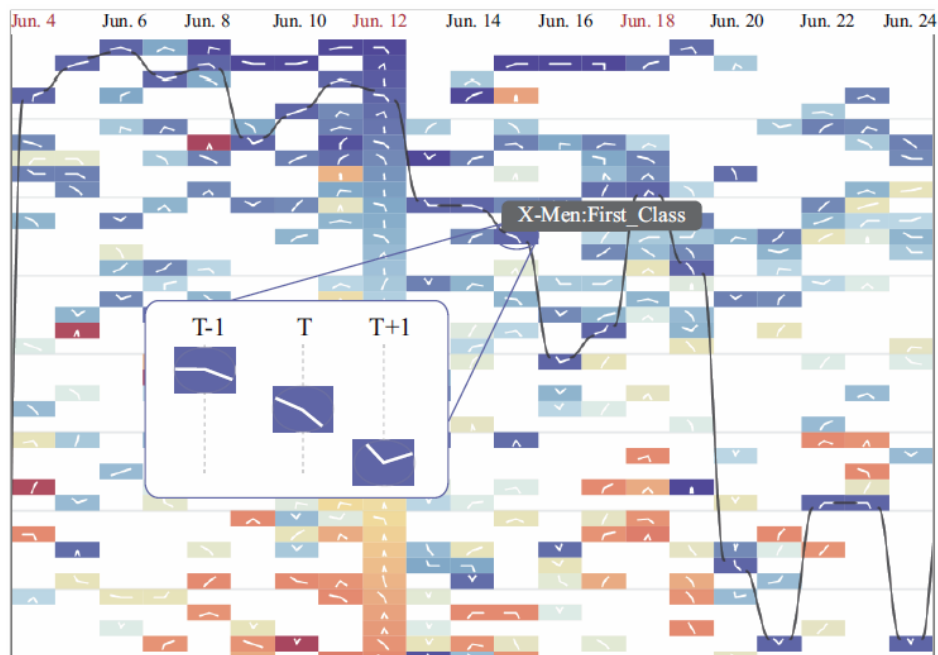
- 基于DTW的时序排名相似性建模

$$Dissim = \frac{w_{dtw} * f_{dtw} + w_{comp} * f_{comp} + w_{avgo} * f_{avgo}}{w_{dtw} + w_{comp} + w_{avgo}}$$

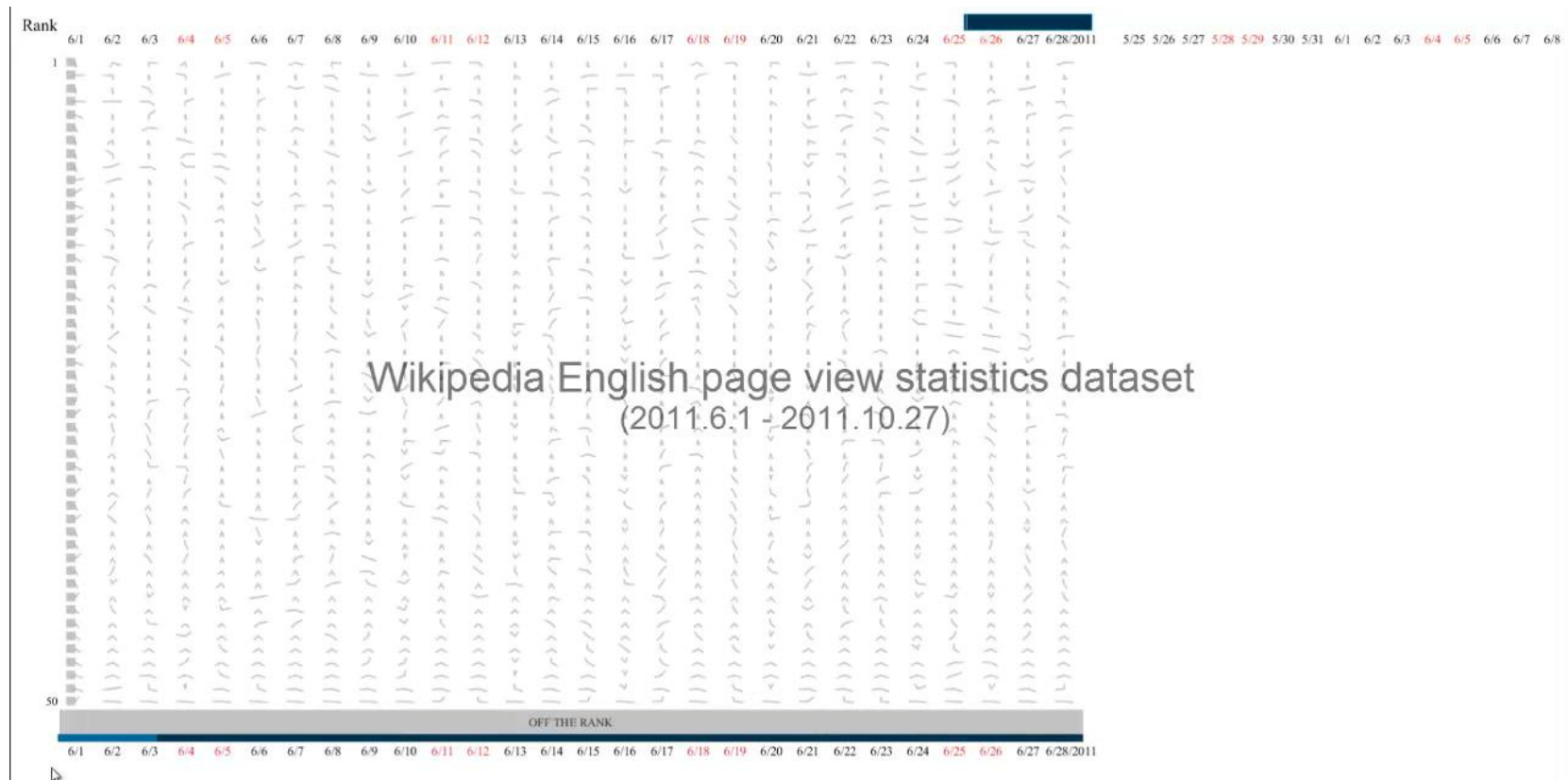
- 相似性是包括DTW因子(f_{dtw})、不连续排名损失因子(f_{comp})和平均排名因子(f_{avgo})的加权综合度量
- 某个页面的相关页面是以下页面的交集
 - 从top-1000中基于相似性查找得到的页面集合
 - 通过pagelink API找到的页面集合



时序排名——图元设计

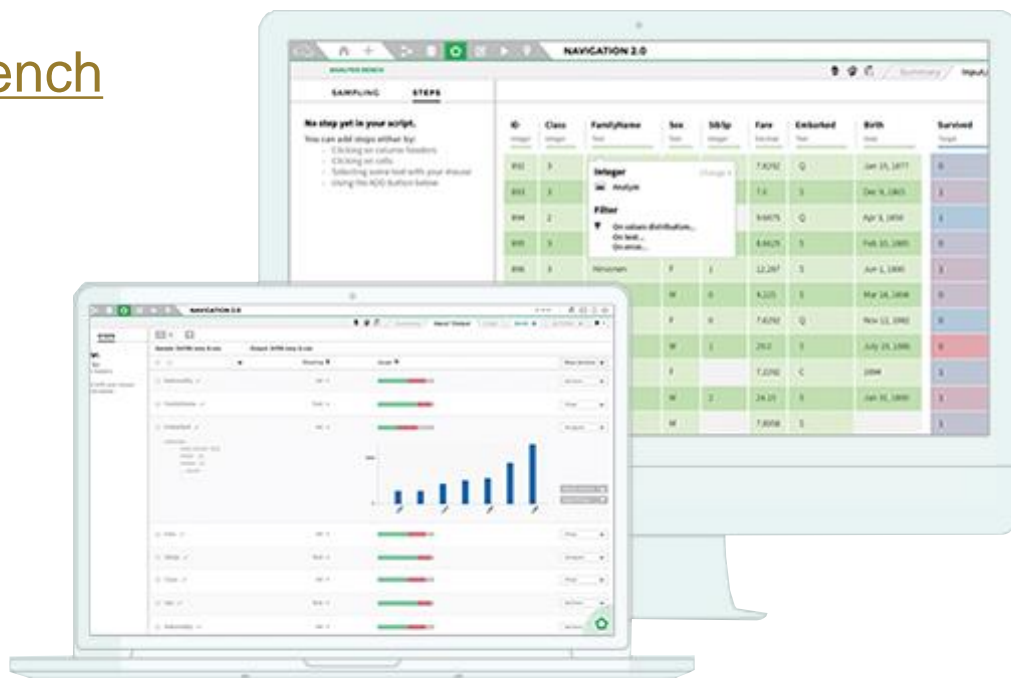


DEMO



一站式数据工具

- Dataiku
- DataScientistWorkbench



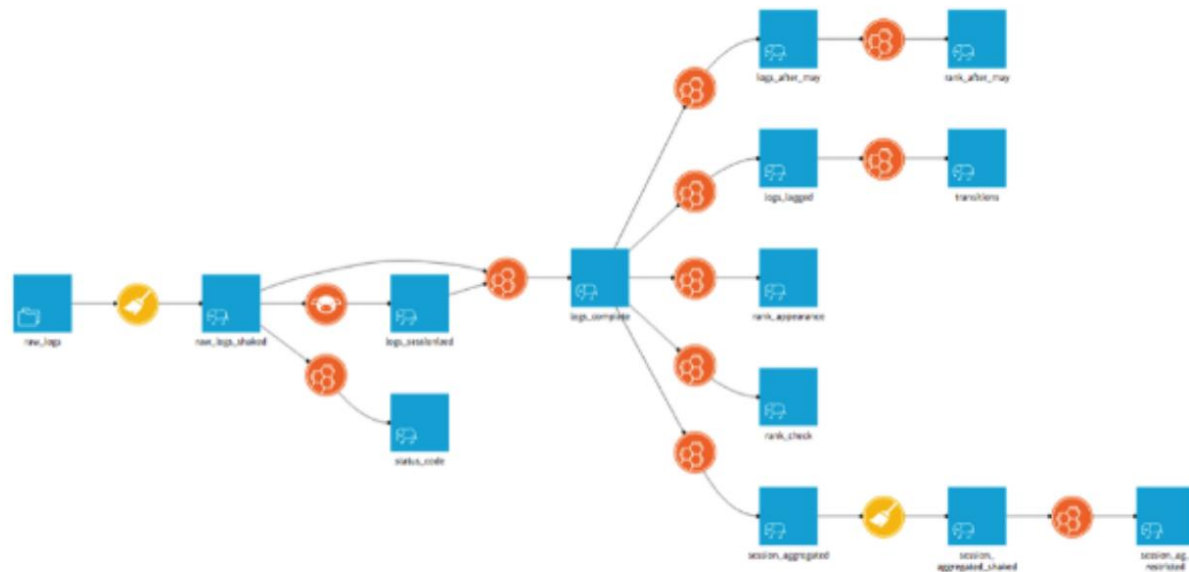
Dataiku —— 数据分析软件

• 功能

- 数据整合
- 数据清洗
- 可视化分析
- 机器学习
- 产品发布

• 特点

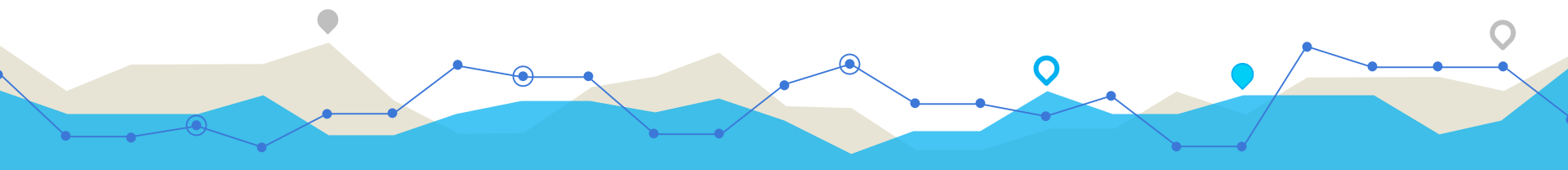
- 功能模块化
- 界面统一





Data Science Studio

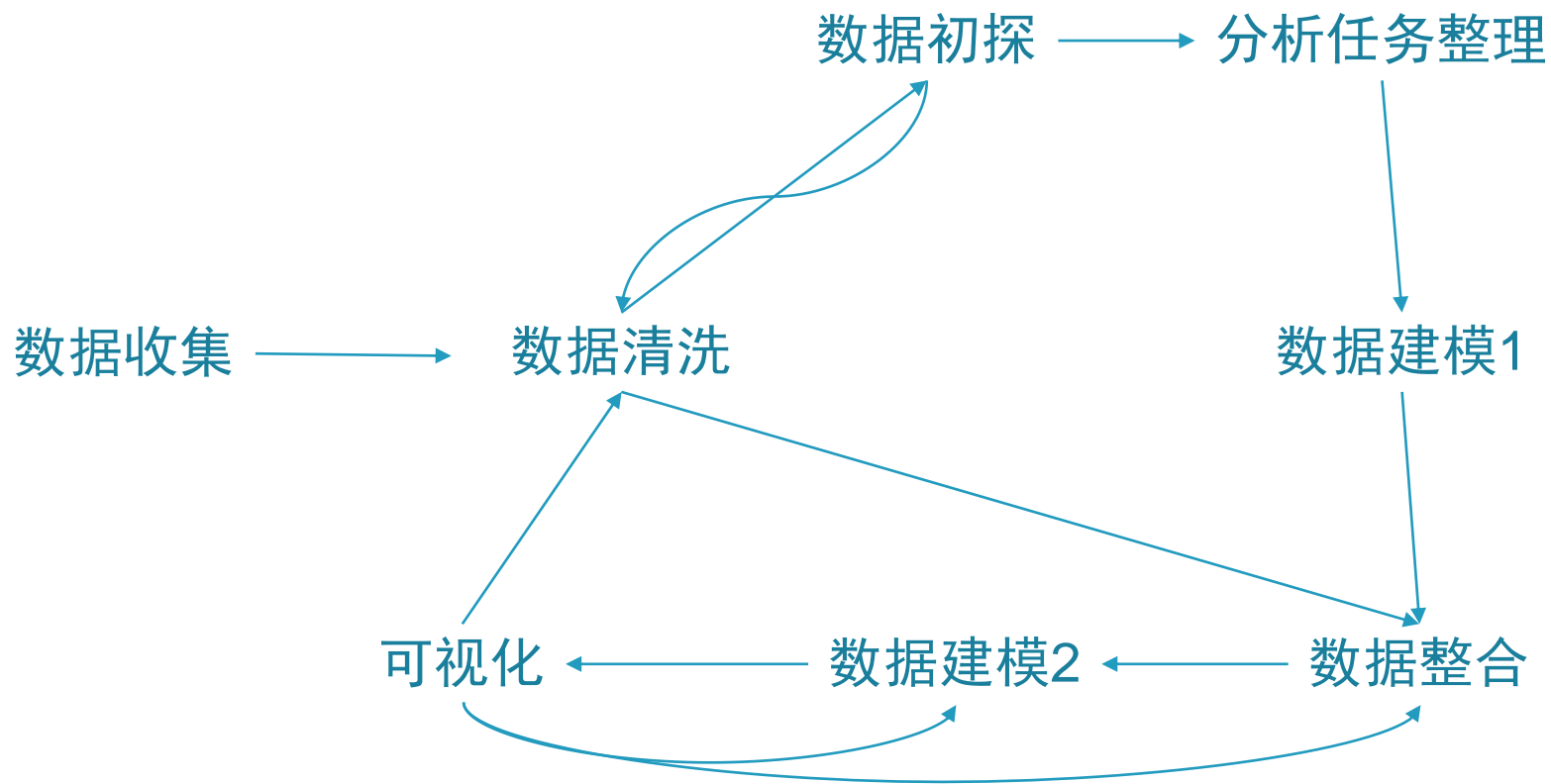
101 Tutorial Video – first steps in DSS



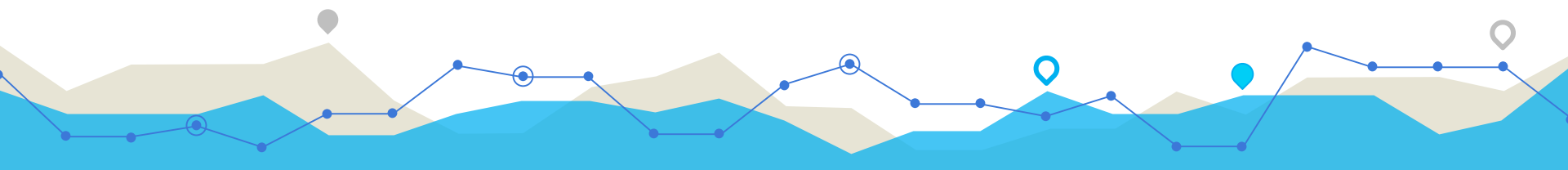
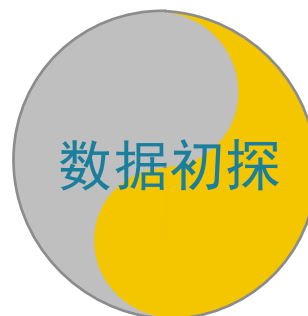
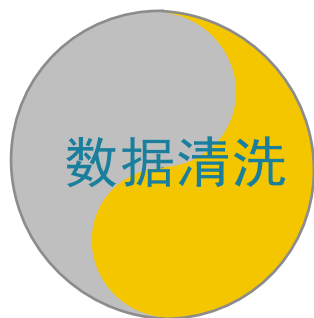
总结



总结



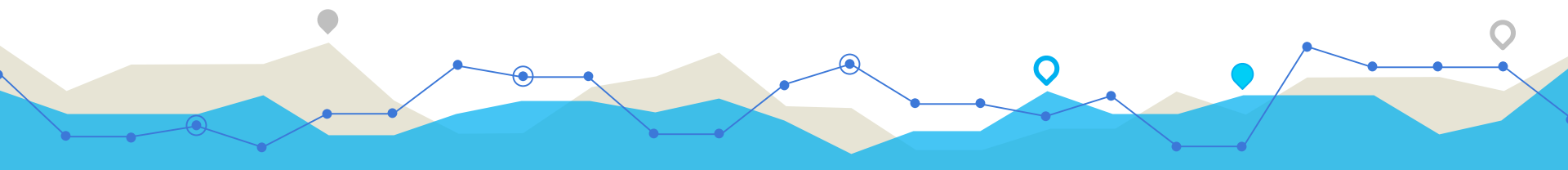
总结



总结

数据清洗和数据初探对于数据分析是至关重要的步骤

可视化的辅助能够降低数据清洗和数据初探的成本



谢谢

夏菁

 summer_179279 /

jjane.summer@gmail.com

