BDTC 2016中国大数据技术大会
Big Data Technology Conference 2016

DI：基于SPARK的交互式数据探索与建模系统

熊永平

北京邮电大学

大数据分析需求变化

- 企业级应用中，BI已经满足不了业务人员的数据分析需求
- 用户越来越希望用高级数据分析方法提升业务能力
- 开发数据价值依赖于大量的懂业务的数据分析人员
- 需要强大易用的大数据分析系统

大数据分析需要掌握这么多装备...

- 数据分析往往利用多种编程语言或系统管理不同的数据挖掘任务和机器学习流程
- 需要掌握从分布式系统到数据分析等众多门槛很高的工具和技能

为什么不能有这样的分析工具？

- 针对普通的数据分析人员
- 几乎不需要编程开发分布式程序
- 提供直观易用的图形化系统界面
- 提供可扩展的数据分析手段
- 可处理大规模数据集
- 计算能力可线性增加
- 部署运维简单

技术选型：Spark

性能 →
- 基于内存的架构极大的减少了磁盘I/O
- 通用任务上20-100x速度的提升

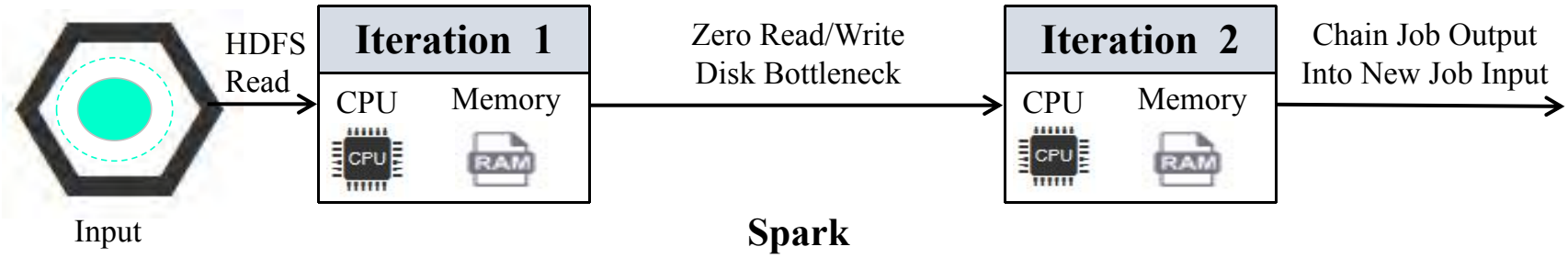
高效 →
- 精简且表达力强大的语法(如Spark2.0的Dataframe)
- 统一的编程模型
- 能用主流的编程语言—Java，Phthon，Scala
- 新工具减少使用的障碍（Spark2.0支持SQL2003）

扩展性 →
- 通过增加机器计算能力实现整体分析能力的线性扩展
- 单个计算节点失效后自动重算
- 利用自动持久化确保整个计算过程可靠完成

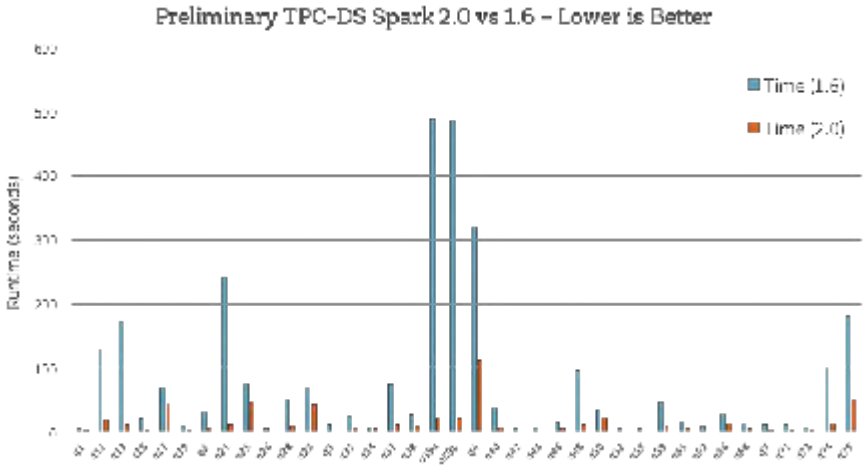
利用Hadoop资产 →
- 和已有的Hadoop生态系统能很好的工作

持续改进 →
- 大量的社区贡献者持续的对Spark技术栈组件进行持续的快速更新

Spark2.0的性能

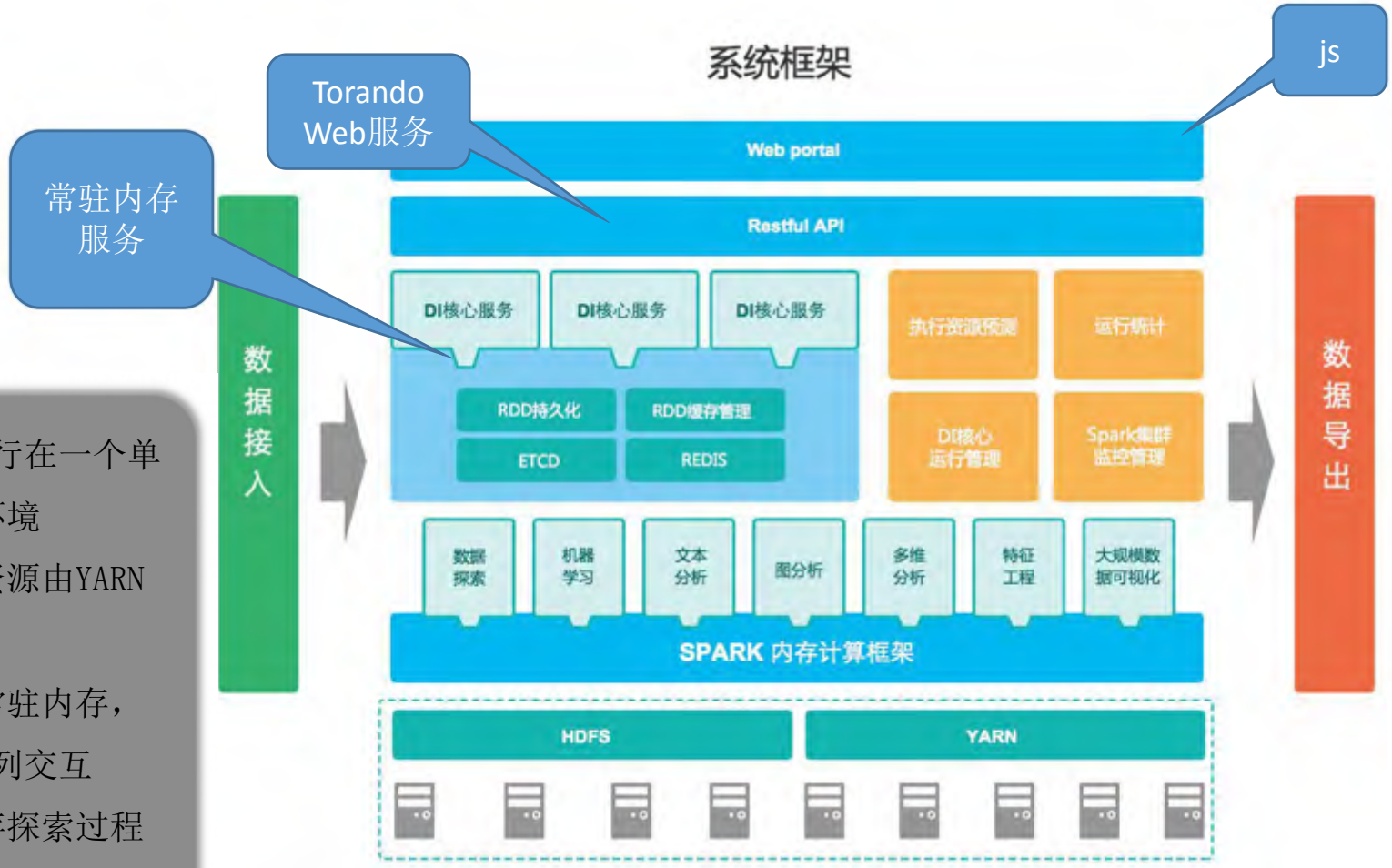

HDFS
Read
Iteration 1
CPU
Memory
Zero Read/Write
Disk Bottleneck
Iteration 2
CPU
Memory
Chain Job Output
Into New Job Input
Input
Spark


Parquet
in 1.6
11M
rows/s
Parquet
in 2.0
90M
rows/s
Spark 1.6
14M
rows/s
Spark 2.0
125M
rows/s

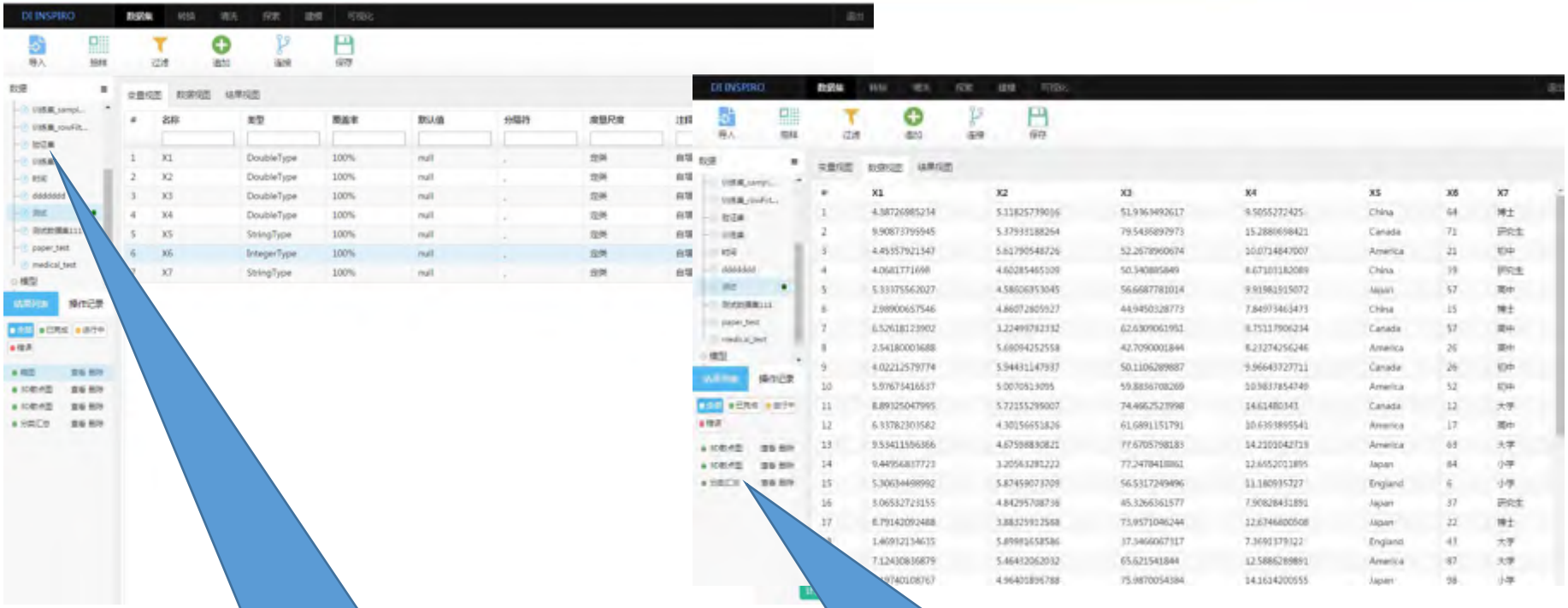

Preliminary TPC-DS Spark 2.0 vs 1.6 – Lower is Better
Runtime (seconds)
Time (1.6)
Time (2.0)

DI：交互式数据探索框架

- 每个工程运行在一个单独的Spark环境
- Spark环境资源由YARN分配调度
- DI和Spark常驻内存，通过消息队列交互
- 利用RDD保存探索过程中的各种中间表



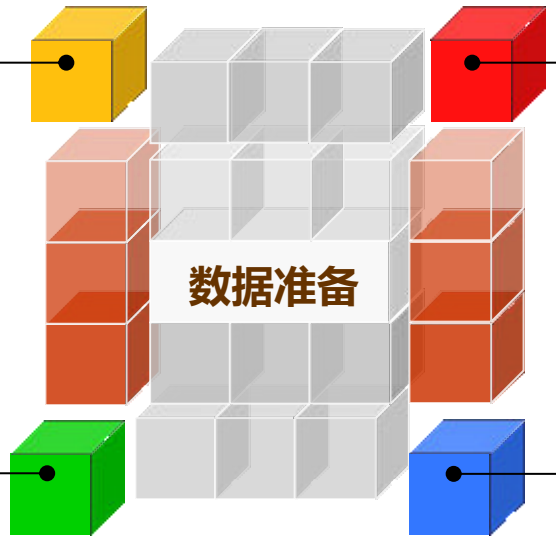
DI概览



- 操作对象抽象为表
- 函数式编程思想
- 所有算子不改变原表数据
- 增加列或生成新表
- 新表单独保存

- 算子操作异步运行
- 记录算子流程DAG
- 流程持久化，实现自动批处理

数据转换



▪ 抽样过滤

- 均匀/随机/分层抽样
- 多条件组合过滤
- 根据距离/密度/局部离群因子/类离群因子/离群点检测的聚类过滤离群点

▪ 变量构造

- 根据已有变量拆分/组合/抽取/运算生成新变量
- 支持数学/统计/文本/日期等100多种函数运算及其逻辑组合
- 根据离散或连续分布函数/等差/等比/日期分布等生成基础列数据

▪ 去重合并

- 左右连接Join
- 取交集/并集
- 合并Merge
- 多列联合主键去重
- 识别并移除冗余

▪ 归一化分箱

- 按根据字典数据替换
- 不同变量进行取值区间归一化
- 连续变量离散化等频/等距/Bootstrap/聚类等分箱
- 利用变量统计值填充空值

典型功能

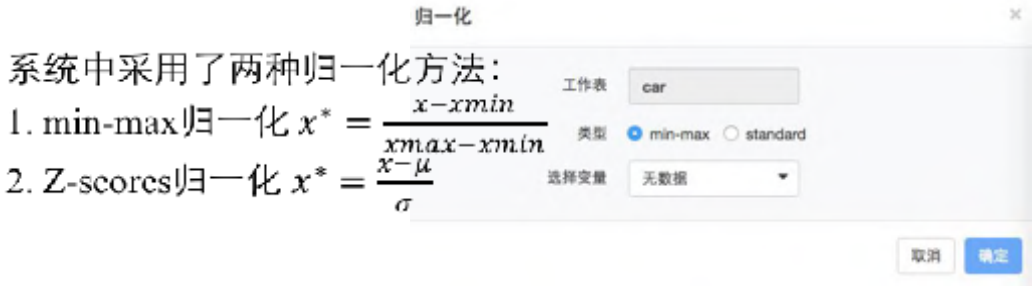
等距分箱

age	sex	ill	data	money	equiBucket_age_1
53	男	手术后恶性肿瘤化学治疗	2013/11/28	5398.08	37.6-56.400000000000006
44	女	恶性肿瘤维持性化学治疗	2013/11/27	5234.93	37.6-56.400000000000006
60	男	恶性肿瘤维持性化学治疗	2013/11/27	4660.67	56.400000000000006-75.2
61	男	肺恶性肿瘤	2013/11/25	4237.37	56.400000000000006-75.2
58	女	肺恶性肿瘤	2013/11/29	3714.02	56.400000000000006-75.2
73	男	慢性肾脏病5期	2013/11/28	7380.61	56.400000000000006-75.2
78	女	肺炎	2013/11/16	21720.56	75.2-94.0000000001
63	女	慢性肾脏病5期	2013/12/2	1532.93	56.400000000000006-75.2
50	男	肺恶性肿瘤	2013/11/20	11407.83	37.6-56.400000000000006
43	女	肺恶性肿瘤	2013/11/29	2181.65	37.6-56.400000000000006
75	男	肺部感染	2013/11/25	11048.23	56.400000000000006-75.2
62	男	慢性肾脏病4期	2013/11/18	9315.45	56.400000000000006-75.2
74	男	肺部阴影	2013/12/2	7714.72	56.400000000000006-75.2
48	男	肾性贫血	2013/12/4	9267.5	37.6-56.400000000000006
22	女	肺炎	2013/11/28	10447.17	18.8-37.6
76	男	非典型分枝杆菌感染	2013/11/26	9020.44	75.2-94.0000000001

归一化

系统中采用了两种归一化方法：

1. min-max归一化 $x^* = \frac{x - xmin}{xmax - xmin}$
2. Z-scores归一化 $x^* = \frac{x - \mu}{\sigma}$



文本抽取

X7	extract1_X7_1
博士	-
研究生	-
初中	中
研究生	-
高中	中
博士	-
高中	中
高中	中
初中	中
初中	中
大学	-
高中	中
大学	-
小学	-
小学	-
研究生	-
博士	-

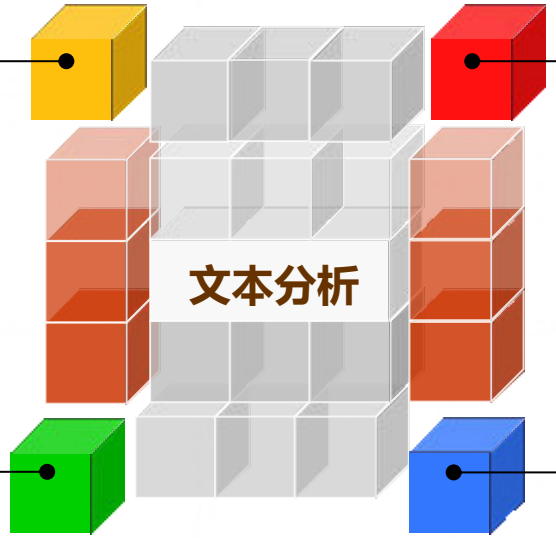
文本分析

▪ 词句段切分

- 支持多种不同的词语切分方法和用户词典加载
- 提供段落和句子切分功能
- 提供新词发现功能，利用互信息熵、条件随机场CRF等算法

▪ 特征词抽取

- 提供TFIDF统计计算
- 利用TextRank算法抽取特征词
- Word2vec词向量模型



▪ 实体识别

- 提取标准实体,如人名、地名、时间、日期和物理量等
- 提取特定领域的概念

▪ 文本挖掘

- 文本分类
- 文本聚类
- 文档矩阵主题分析

利用文档预处理、自然语言处理、主题检测等功能分析文本数据，便于数据分析人员处理非结构化文本数据。

从目标文本中按某种算法提取关键词

- TFIDF算法

$$\text{词频}(TF) = \frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}}$$

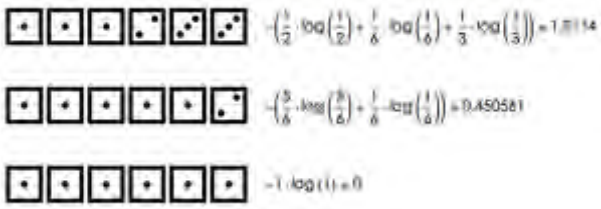
$$\text{逆文档频率}(IDF) = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}\right)$$

$$TF - IDF = TF * IDF$$

- TextRank算法
  - 将待抽取关键词的文本进行分词
  - 固定长度为n(通常取5)的窗口，在一个窗口中的任两个单词对应的节点之间存在一个无向无权的边
  - 基于窗口分割后的边构建TextGraph图，利用PageRank计算每个节点的重要性
  - 得到最重要的k个节点即提取出了k个关键词

典型算子：新词识别

- 信息熵定义



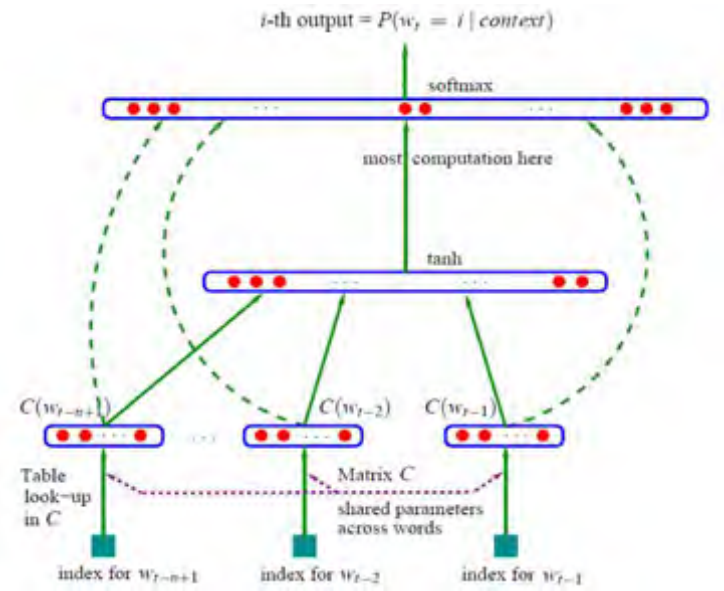
- 新词识别
  - 词频
  - 自由度（片段所有可能左右邻的混乱程度）
    - 例句：利用公用配电**负荷**历史负载率以及中长期配电**负荷**预测结果，结合配变的容量，指出**负荷**容量不足和容量过剩的配变。
    - 片段“负荷”的所有左邻字实例为{电，出}
      熵为 $-(\frac{1}{2}\cdot\ln(\frac{1}{2})+\frac{1}{2}\cdot\ln(\frac{1}{2}))\approx 0.693$
    - 片段“负荷”的所有右邻字实例为{历，预，容}
      熵为 $-(\frac{1}{2}\cdot\ln(\frac{1}{2})+\frac{1}{3}\cdot\ln(\frac{1}{3})+\frac{1}{3}\cdot\ln(\frac{1}{3}))\approx 1.08$
    - “负荷”的右邻字比左邻字更丰富更灵活
  - 凝合度
    - 令n为文本的长度，令f(x)为字符串x在文本中出现的次数
    - 令p(x)为f(x)/n，即字符串x出现的概率
    - 定义“负载率”的可拆分度为：
      - max(p(负)·p(载率) / p(负载率), p(负载)·p(率) / p(负载率))

- 利用知网电力领域期刊全文数据50G文本数据
- 抽取新词1100个

俄统国际　生物质燃料　热电联产　能源消费　电价基　固体生物质　抽水蓄能　农村电气化　生物甲烷　联合循环　资产融资　乙醇燃料　国际电力　运输燃料　南卡　生物质柴油　光伏组件　并网光伏　全球光伏　卡奥拉巴萨　生物质供热　并网太阳能　能源国际　燃料混合　矿产能源部　燃料车　税收激励　埃克森美孚　特许权招标　亚联邦　太阳能法令　径流式　燃能系统

典型算子：词语语义距离分析Word2Vec

- 每个词表示为一个高维向量
- 利用多层神经网络和统计语言模型训练出词的向量表示
- 利用空间向量模型Cos计算距离



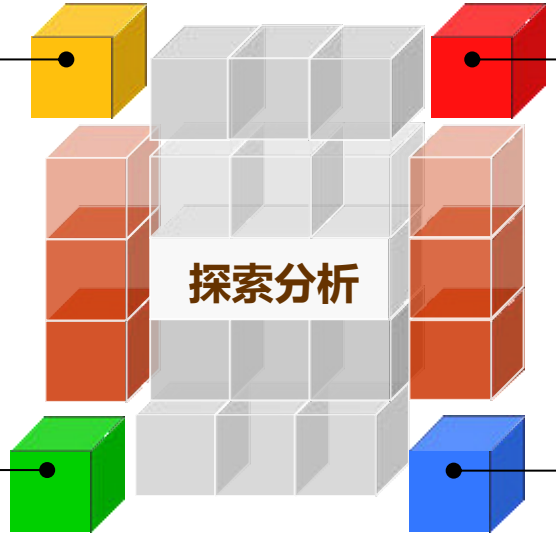
探索分析

▪ 变量统计

- 描述性数据统计，包括常见统计量、分位数等
- 提供变量概率密度和直方图，以及分布推断
- 实现变量信息熵和信息值IV计算
- 多维分组汇总OLAP

▪ 行相关分析

- 关联规则分析，支持频繁项集的交互式筛选
- 对连续和离散变量的 K 均值聚类，自动估计最佳聚类数量，输出整个数据集的聚类归属和距离测量值
- 超过20种相似性距离测量方法



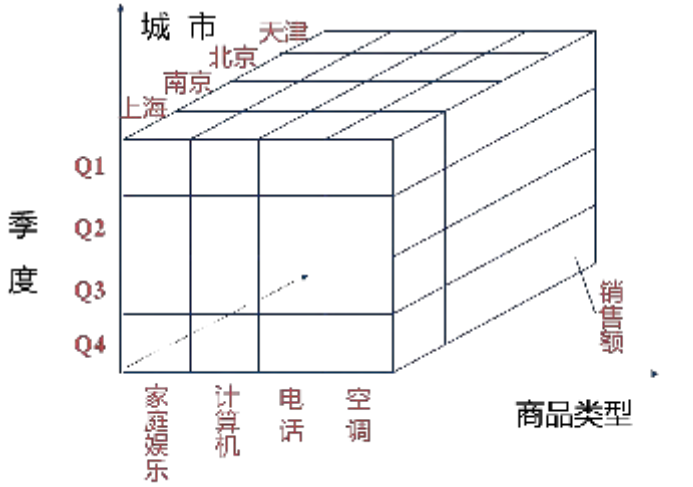
▪ 特征降维（列相关分析）

- 变量聚类
- 主成分分析PCA
- 支持维度的层次划分
- 支持多种度量的计算函数

▪ 离散相关分析

- 相关性度量：pearson、spearman、余弦相关系数、互信息等
- 共现矩阵分析
- 自相关矩阵分析
- 互相关矩阵分析

- 利用cube进行分组汇总



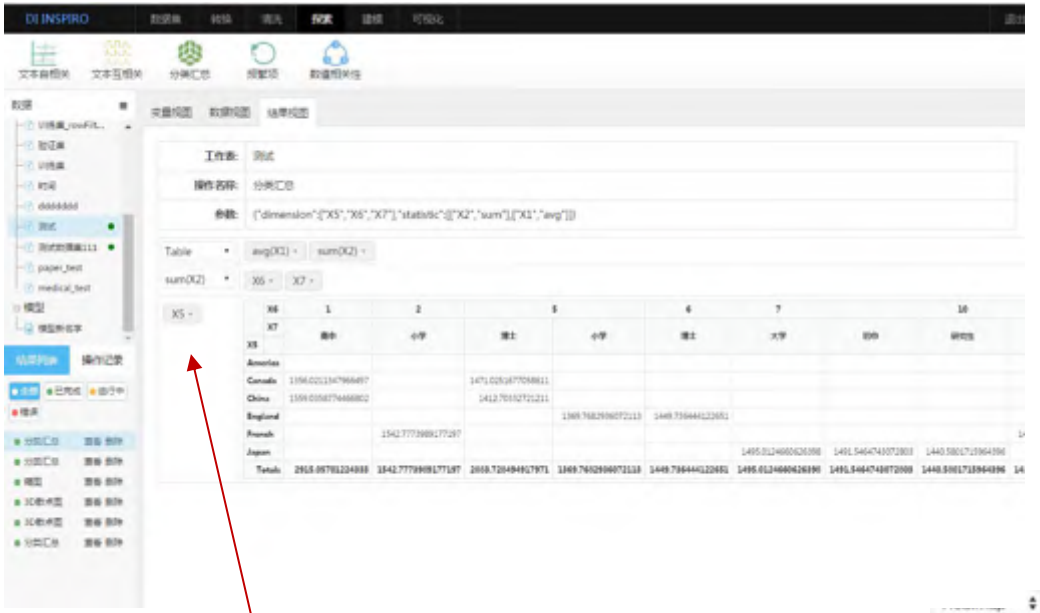
按季度时间、城市地区和商品类型三个维销售的数据

Item	Color	QtySum
Table	Blue	123
Table	Red	223
Chair	Blue	101
Chair	Red	210

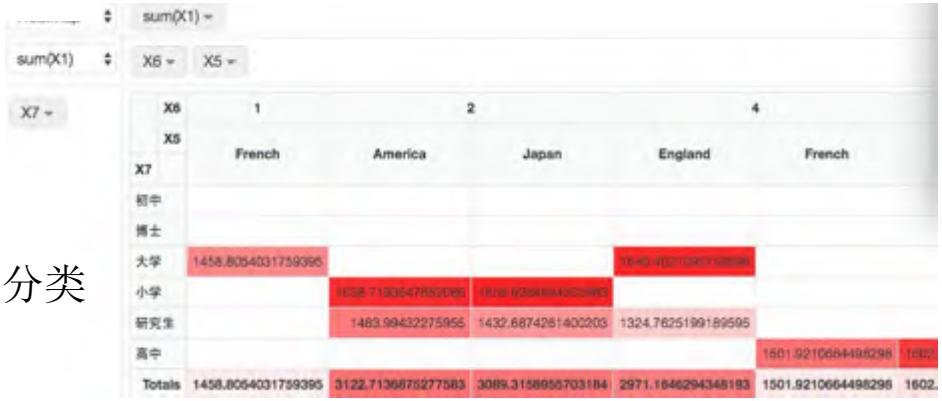
Item	Color	QtySum
Table	Blue	123
Table	Red	223
Table	ALL	347
Chair	Blue	101
Chair	Red	210
Chair	ALL	311
ALL	Blue	225
ALL	Red	433
ALL	ALl	658

- 提供了对数据进行多种维度、多种度量方法进行汇总展示的功能。
- 快速获取数据的宏观统计信息。

分组汇总示例



支持维度、指标的交互式拖拽和实时计算



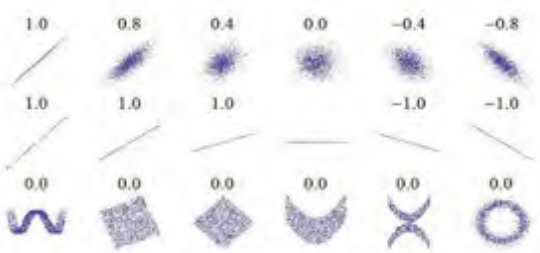
X5 X6 X7是维度，以交叉表的方式展现了分类汇总的统计信息

- Pearson相关系数

用来衡量两个数据集合是否在一条线上面，它用来衡量变量间的线性关系。

本质上是去中心化后的余弦相似度

$$\rho_{x,y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$



- Spearman相关系数

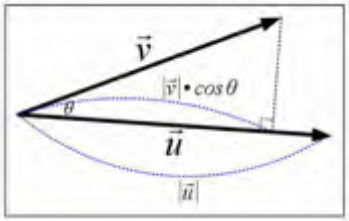
用来衡量两个变量之间的关联程度与方向

$$r_s = \frac{l_{pq}}{\sqrt{l_{pp}l_{qq}}}$$

- 余弦相似度

通过计算两个向量的夹角余弦值来评价相似度

$$\cos\theta = \frac{a\cdot b}{\|a\|\cdot\|a\|}$$



- 欧氏距离

$$EUCLID = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

- Minkowski距离

$$MINKOWSKI(x,y) = \sqrt[p]{\sum_{i=1}^{k}|x_i - y_i|^p}$$

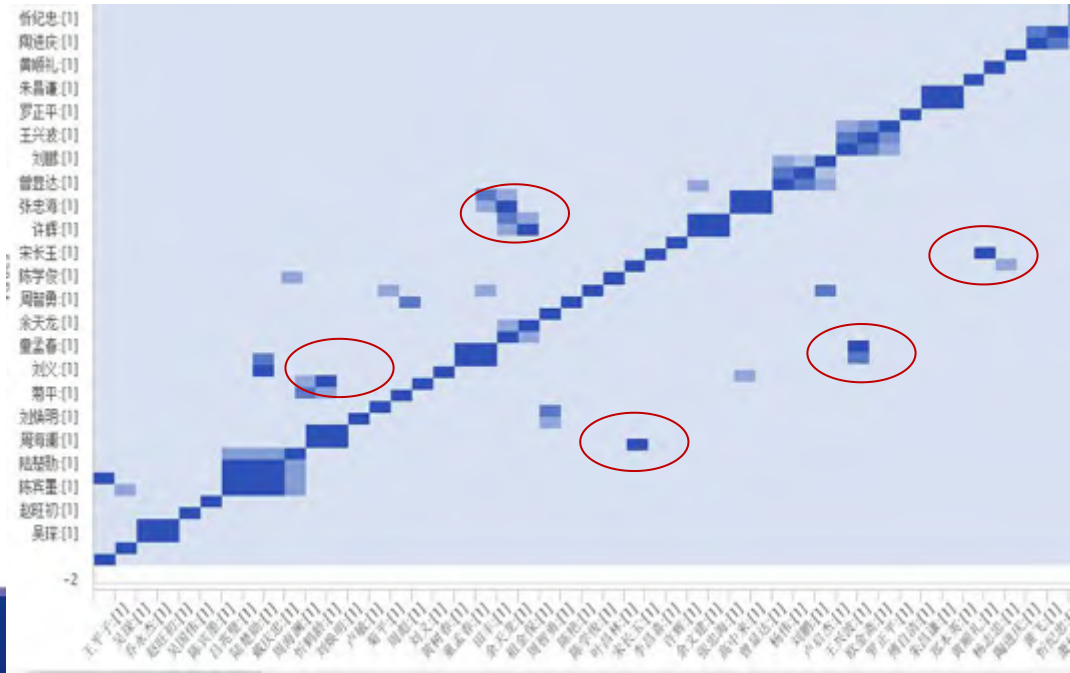
典型算子2：特征降维（列相关）

- 用途
  - 数据建模则需要快速挖掘出与目标相关度最高的特征
  - 建模前无法确定输入变量对目标变量的预测能力，把可能的基础变量和衍生变量放到模型中，但这些变量特别是衍生变量之间可能会存在相关性，会导致模型的多重共线性，从而造成模型整体预测能力的下降。
  - 例子
    - 保险业务
    - 保单号、被保险人、车牌号、保费收入、起保日期、已赚保费
- 解决方法
  - 主成分分析/卡方检验/岭回归等（无法解释）
  - 变量聚类
    - 采用相关系数，夹角余弦和列联系数来寻找反映元素之间亲疏关系的统计量，然后依据此把变量分为若干组。
    - 把高度相关的变量聚到一组，每一组内的变量之间信息重合度很高，互补性很弱，而组间的变量相关性则很低，信息重合度很弱，互补性很强。
    - 从每一组选择一两个最具有代表性的变量代表整个类别，参与建模。

问题：科研合作分析

论文数据

publication_time	title	abstract	author
2000_1期	主编致辞	-	-
2000_1期	大功率汽轮机低压损调节阀的试验研究	介绍通过吹风试验研究,得到在所有工况下都...	王平子:[1]
2000_1期	四杆同源机构的构成特性和传动角特性	四杆曲线同源机构与原始机构的对应杆长具...	吴琛:[1],乔永杰:[1]
2000_1期	循环流化床锅炉分离器的研究	着重介绍国内外气固分离器的结构及使用情...	赵旺初:[1]
2000_1期	300/600MW燃煤电厂输煤控制系统设计及...	在总结上海工业自动化仪表研究所从事火力...	吴国伟:[1]
2000_1期	我国发电设备产业产品与技术发展预测	对今后10～15年内,我国发电设备产业的产品...	陈宾墨:[1],吕兆璧:[1],陆楚勋:[1],戴庆忠:[1]
2000_1期	论超临界火电机组的发展	简要论述了世界上超临界火电机组的最新发...	周海澜:[1],忻鹤龄:[1]
2000_1期	水轮机环列导水叶栅网格生成技术及应用	网格生成是流场数值模拟的一个关键环节,网...	刘焕明:[1]
2000_1期	水轮发电机临界失步问题的导纳分析及应用	从同步电机电势相量图与功角特性出发,用导...	卢敏:[1]



探索的是作者之间的合作关系
相关性方法支持pearson、cos

典型算子3：离散自相关矩阵分析

原始数据表

	C1	C2	C3	C4	C5
R1	X_1	Y_1	A_1,A_2,A_3	B_1,B_2,B_3	Z_1
R2	X_2	Y_2	A_2,A_3	B_2,B_4	Z_2
R3	X_3	Y_3	A_1,A_4,A_5	B_2,B_3,B_6	Z_3
R4	X_4	Y_4	A_2,A_5	B_1,B_4	Z_4
R5	X_5	Y_5	A_3,A_4	B_1,B_5	Z_5

投影

矩阵变换

	R_1	R_2	R_3	R_4	R_5
A_1	1	0	1	0	0
A_2	1	1	0	1	0
A_3	1	1	0	0	1
A_4	0	0	1	0	1
A_5	0	0	1	1	0

共现相关性

	A_1	A_2	A_3	A_4	A_5
A_1		1	1	1	1
A_2	1		2	0	1
A_3	1	2		1	0
A_4	1	0	1		1
A_5	1	1	0	1	

cos相关性

	A_1	A_2	A_3	A_4	A_5
A_1	1	0.408	0.408	0.5	0.5
A_2	0.408	1	0.667	0	0.408
A_3	0.408	0.667	1	0.408	0
A_4	0.5	0	0.408	1	0.5
A_5	0.5	0.408	0	0.5	1

pearson相关性

	A_1	A_2	A_3	A_4	A_5
A_1	1	-0.167	-0.167	0.167	0.167
A_2	-0.167	1	0.167	-1	-0.167
A_3	-0.167	0.167	1	-0.167	-1
A_4	0.167	-1	-0.167	1	0.167
A_5	0.167	-0.167	-1	0.167	1

应用：保险欺诈合谋分析

理赔数据表

案件编号	人员	车牌	地点	金额
344561	段建华，张华，许卫	湘A2BA32，湘AA1391，湘ZG00069	板仓南路	20000
344562	罗坚，肖蓉	湘J7ZH83，湘AL5S85	开元西路	50000
344563	王丽萍，刘双泉	湘A65N90，湘A1661K	寿昌路	100000
344564	彭发兵，周辉，苏英雄	湘A2ZB92，湘B2HL12，湘A2KA19	人民路	70000
344565	张斌，王丽萍，	湘AT8137，湘A65N90	湘江东路	10000

矩阵变换

	344567	344568	344569	344570	344571
谢前	0	1	1	0	1
敬春桥	0	0	0	1	0
罗坚	0	1	0	0	1
肖蓉	0	0	1	0	0
刘双泉	0	1	0	0	0

	谢前	敬春桥	罗坚	肖蓉	刘双泉
谢前	1	-0.612	0.667	0.408	0.408
敬春桥	-0.612	1	-0.408	-0.25	-0.25
罗坚	0.667	-0.408	1	-0.408	0.612
肖蓉	0.408	-0.25	-0.408	1	-0.25
刘双泉	0.408	-0.25	-1	-0.25	1

pearson相关性

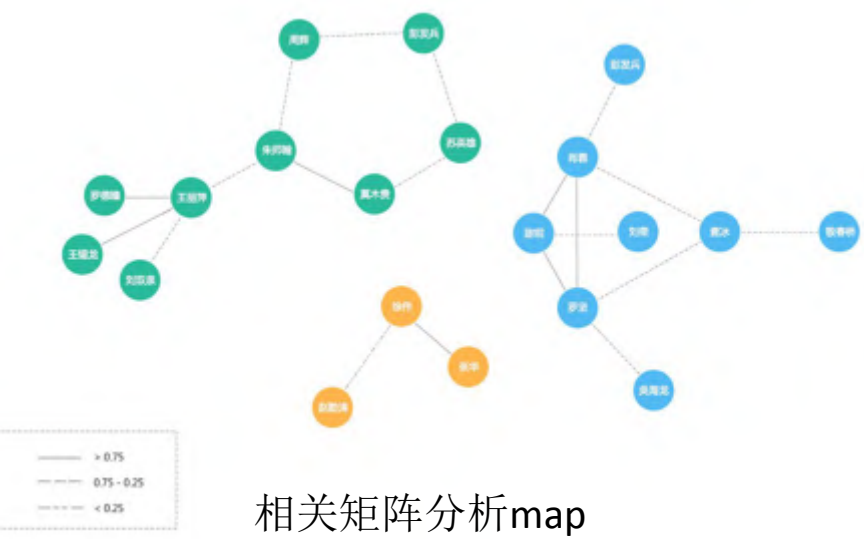
	谢前	敬春桥	罗坚	肖蓉	刘双泉
谢前		0	2	1	1
敬春桥	0		0	0	0
罗坚	2	0		0	1
肖蓉	1	0	0		0
刘双泉	1	0	1	0	

共现相关性

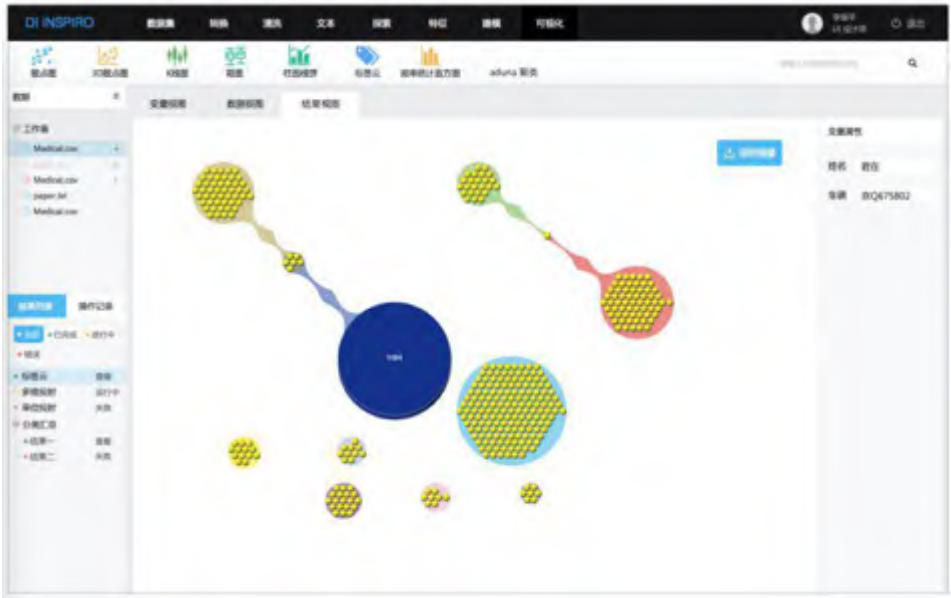
cos相关性

	谢前	敬春桥	罗坚	肖蓉	刘双泉
谢前	1	0	0.816	0.577	0.577
敬春桥	0	1	0	0	0
罗坚	0.816	0	1	0	0.707
肖蓉	0.577	0	0	1	0
刘双泉	0.577	0	0.707	0	1

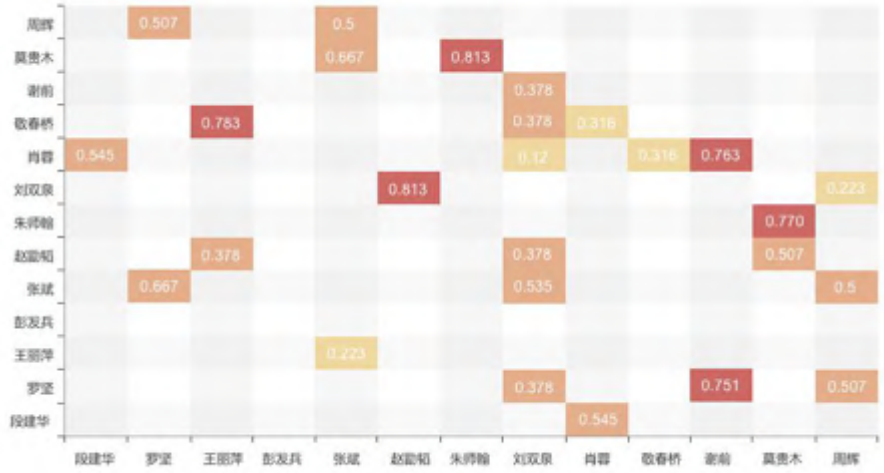
合谋关系分析可视化



相关矩阵分析map



Aduna图合作关系可视化



车辆自相关分析

原始数据表

案件编号	人员	车牌	地点	金额
344573	周辉，朱师翰	湘A65N90，湘A1661K	桂花路	20000
344574	苏英雄，莫贵木，罗坚，肖蓉，谢前	湘A2KA19，湘AT0101，湘J7ZH85，湘AL5S84，湘AA0390	人民路	50000
344575	张斌，刘荣	湘A632AK，湘AUQ852	开元西路	100000
344576	罗德臻，王丽萍	湘A9HT68，湘A65N90	北斗路	70000
344577	吴海龙，张华	湘AVB590，湘AA1391	北斗路	10000

矩阵变换

	344573	344574	344575	344576	344577
湘ZG0069	0	0	0	0	0
湘A2KA19	0	1	0	0	0
湘AA0390	0	1	0	0	0
湘A65N90	1	0	0	1	0
湘AA1391	0	0	0	0	1

cos相关性

	湘ZG0069	湘A2KA19	湘AA0390	湘A65N90	湘AA1391
湘ZG0069	1	0	0	0	0
湘A2KA19	0	1	1	0	0
湘AA0390	0	1	1	0	0
湘A65N90	0	0	0	1	0
湘AA1391	0	0	0	0	0

典型算子4：离散互相关矩阵分析

原始数据表

	C1	C2	C3	C4	C5
R1	X_1	Y_1	A_1,A_2,A_3	B_1,B_2,B_3	Z_1
R2	X_2	Y_2	A_2,A_3	B_2,B_4	Z_2
R3	X_3	Y_3	A_1,A_4,A_5	B_2,B_3,B_6	Z_3
R4	X_4	Y_4	A_2,A_5	B_1,B_4	Z_4
R5	X_5	Y_5	A_3,A_4	B_1,B_5	Z_5

投影

矩阵变换

	B_1	B_2	B_3	B_4	B_5
A_1	1	2	2	0	0
A_2	1	2	1	2	0
A_3	1	2	1	1	1
A_4	1	1	1	0	1
A_5	1	1	1	1	0

共现互相关性

	A_1	A_2	A_3	A_4	A_5
A_1		3	3	3	3
A_2	3		5	0	0
A_3	3	5		2	0
A_4	3	0	2		3
A_5	3	0	0	3	

cos互相关性

	A_1	A_2	A_3	A_4	A_5
A_1	1	0.7	0.783	0.849	0.849
A_2	0.7	1	0.894	0.566	0.849
A_3	0.783	0.894	1	0.791	0.791
A_4	0.849	0.566	0.791	1	0.8
A_5	0.849	0.849	0.791	0.8	1

pearson互相关性

	A_1	A_2	A_3	A_4	A_5
A_1	1	0.25	0.354	0.548	0.548
A_2	0.25	1	0.707	-0.548	0.548
A_3	0.354	0.707	1	0	0
A_4	0.548	-0.548	0	1	-0.2
A_5	0.548	0.548	0	-0.2	1

应用：从报案地点看合作欺诈关系

理赔数据表

案件编号	人员	车牌	地点	金额
344561	段建华，张华，许卫	湘A2BA32，湘AA1391，湘ZG00069	板仓南路	20000
344564	彭发兵，周辉，苏英雄	湘A2ZB92，湘B2HL12，湘A2KA19	人民路	70000
344574	苏英雄，莫贵木，罗坚，肖蓉，谢前	湘A2KA19，湘AT0101，湘J7ZH85，湘AL5S84，湘AA0390	人民路	50000
344563	王丽萍，刘双泉	湘A65N90，湘A1661K	寿昌路	100000
344581	张华，莫木贵，许卫	湘AA1391，湘AT0101，湘ZG00069	板仓南路	100000

矩阵变换

	板仓南路	开元西路	寿昌路	人民路	湘江东路
湘ZG0069	3	0	0	0	0
湘A2KA19	0	0	0	2	0
湘AA0390	0	0	0	1	0
湘A65N90	0	0	1	0	1
湘AA1391	3	0	0	0	0

Cos互相关

	湘ZG0069	湘A2KA19	湘AA0390	湘A65N90	湘AA1391
湘ZG0069	1	0	0	0	1
湘A2KA19	0	1	1	0	0
湘AA0390	0	1	1	0	0
湘A65N90	0	0	0	1	0
湘AA1391	1	0	0	0	1

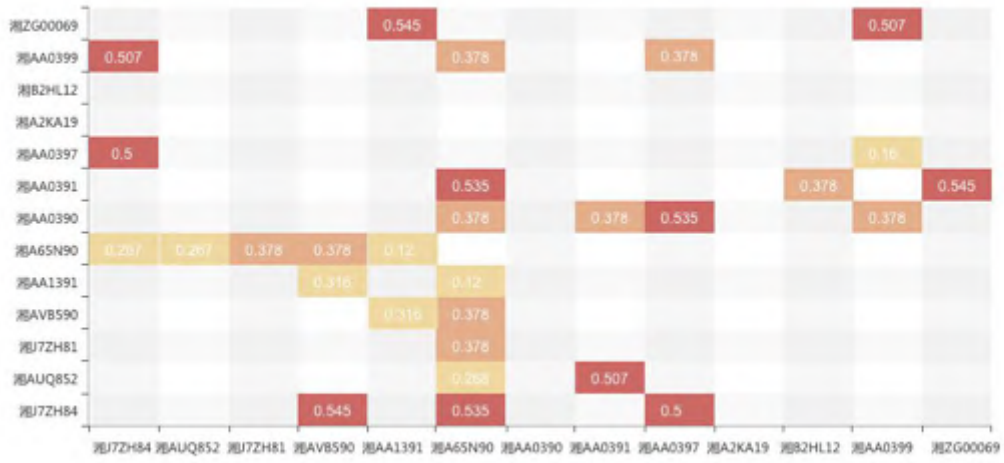
	湘ZG0069	湘A2KA19	湘AA0390	湘A65N90	湘AA1391
湘ZG0069	1	0	0	0	0
湘A2KA19	0	1	1	0	0
湘AA0390	0	1	1	0	0
湘A65N90	0	0	0	1	0
湘AA1391	0	0	0	0	0

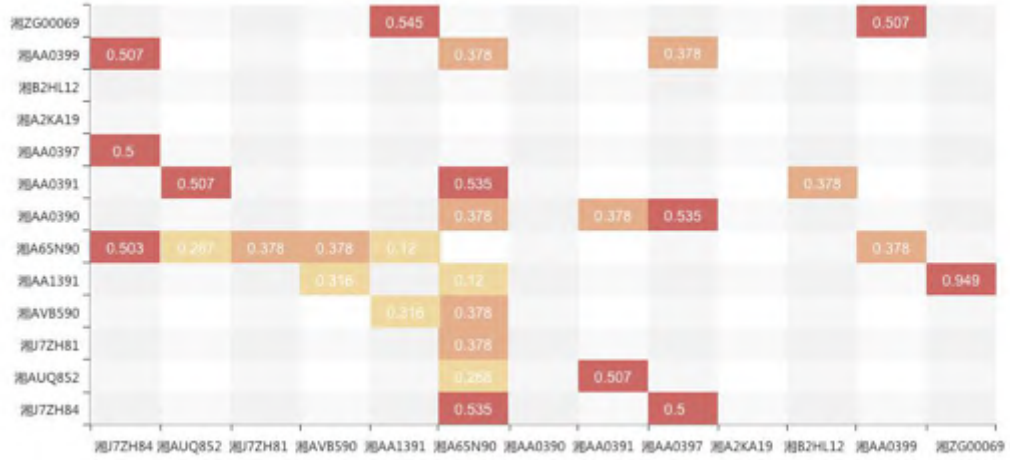
cos自相关

	湘ZG0069	湘A2KA19	湘AA0390	湘A65N90	湘AA1391
湘ZG0069		0	0	0	3
湘A2KA19	0		1	0	0
湘AA0390	0	1		0	0
湘A65N90	0	0	0		0
湘AA1391	3	0	0	0	

共现互相关

可疑车辆可视化


湘ZG00069
湘AA0399
湘B2HL12
湘A2KA19
湘AA0397
湘AA0391
湘AA0390
湘A65N90
湘AA1391
湘AVB590
湘J7ZH81
湘AUQ852
湘J7ZH84
湘J7ZH84 湘AUQ852 湘J7ZH81 湘AVB590 湘AA1391 湘A65N90 湘AA0390 湘AA0391 湘AA0397 湘A2KA19 湘B2HL12 湘AA0399 湘ZG00069

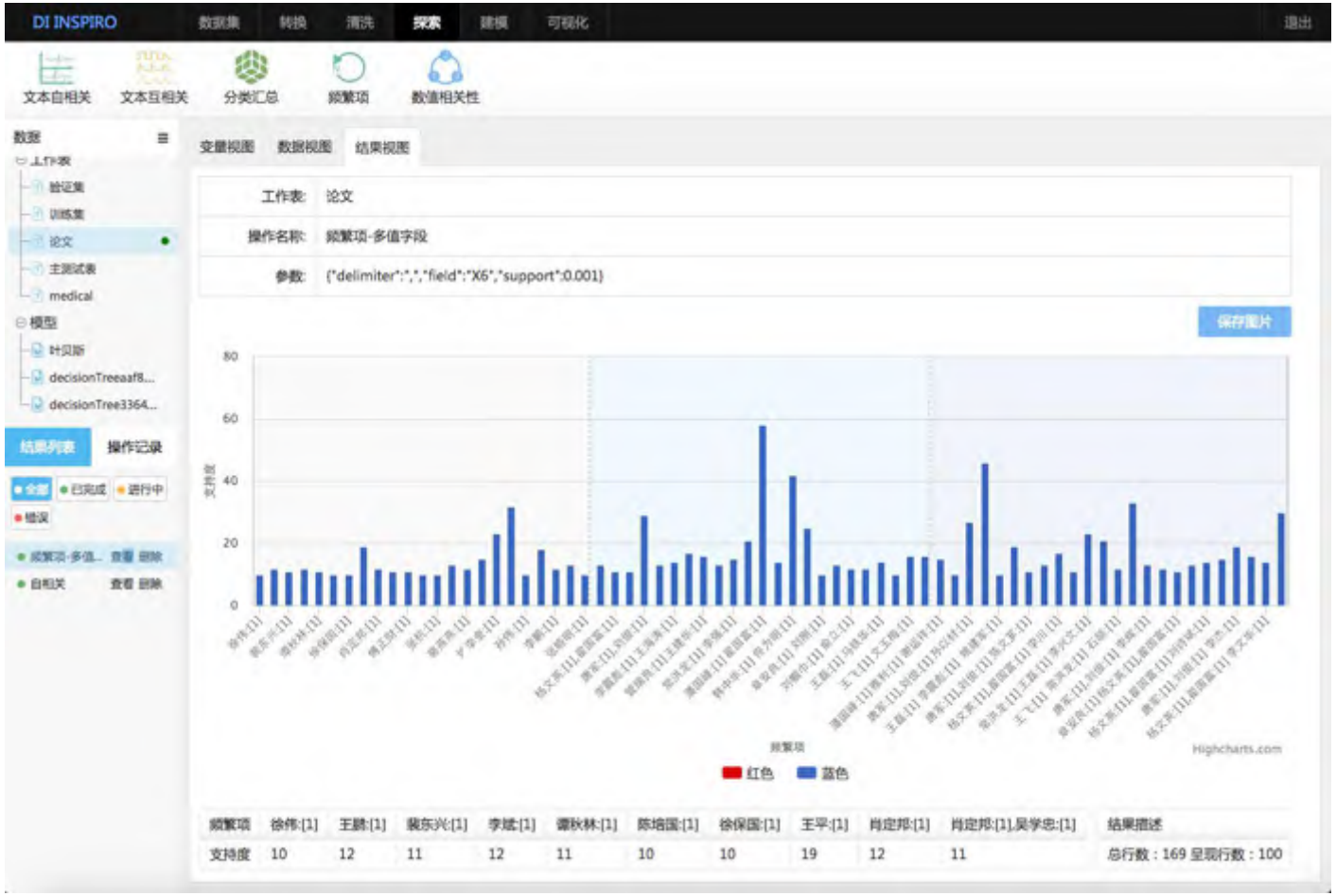

湘ZG00069
湘AA0399
湘B2HL12
湘A2KA19
湘AA0397
湘AA0391
湘AA0390
湘A65N90
湘AA1391
湘AVB590
湘J7ZH81
湘AUQ852
湘J7ZH84
湘J7ZH84 湘AUQ852 湘J7ZH81 湘AVB590 湘AA1391 湘A65N90 湘AA0390 湘AA0391 湘AA0397 湘A2KA19 湘B2HL12 湘AA0399 湘ZG00069

离散矩阵分析

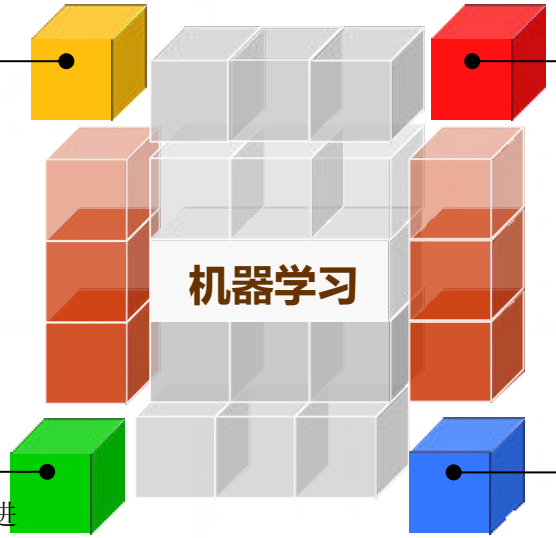
- 实质
  - 二分图
- 很多应用领域
  - 医疗领域
    - 药品关联分析
  - 公共安全
    - （相同时间/机场）乘坐相同航班的同乘分析
    - （相同时间/地点）的紧密通话客户分析
  - 科技领域
    - 研发相类似技术领域的竞争对手分析

典型算子5： 频繁项

探索记录属性间的关联性



数据建模



▪ 线性模型

- 支持线性回归和逻辑回归模型，支持任意程度嵌套效果、多项式和曲线效果
- 支持自动模型选择，提供丰富的模型诊断结果和自动模型评估。

▪ 决策树和随机森林

- 支持包含分类和连续特征的分类树和回归树
- 提供成本复杂性、C4.5 和减少误差的自动修剪并基于保留最优树
- 支持二分变量、名义变量和连续变量的随机森林、自动组合多个决策树预测单个目标
- 自动分配独立模型训练任务，自动智能调整参数设置确定最佳模型

▪ 模型管理

- 利用分区数据集对预测模型结果进行测试比对，快速确定最佳模型
- 利用提升表、ROC 图 表、协调统计和错误分类表，通过可视化评估和验证指标来验证结果
- 支持模型的持久化存储

▪神经网络和支持向量机

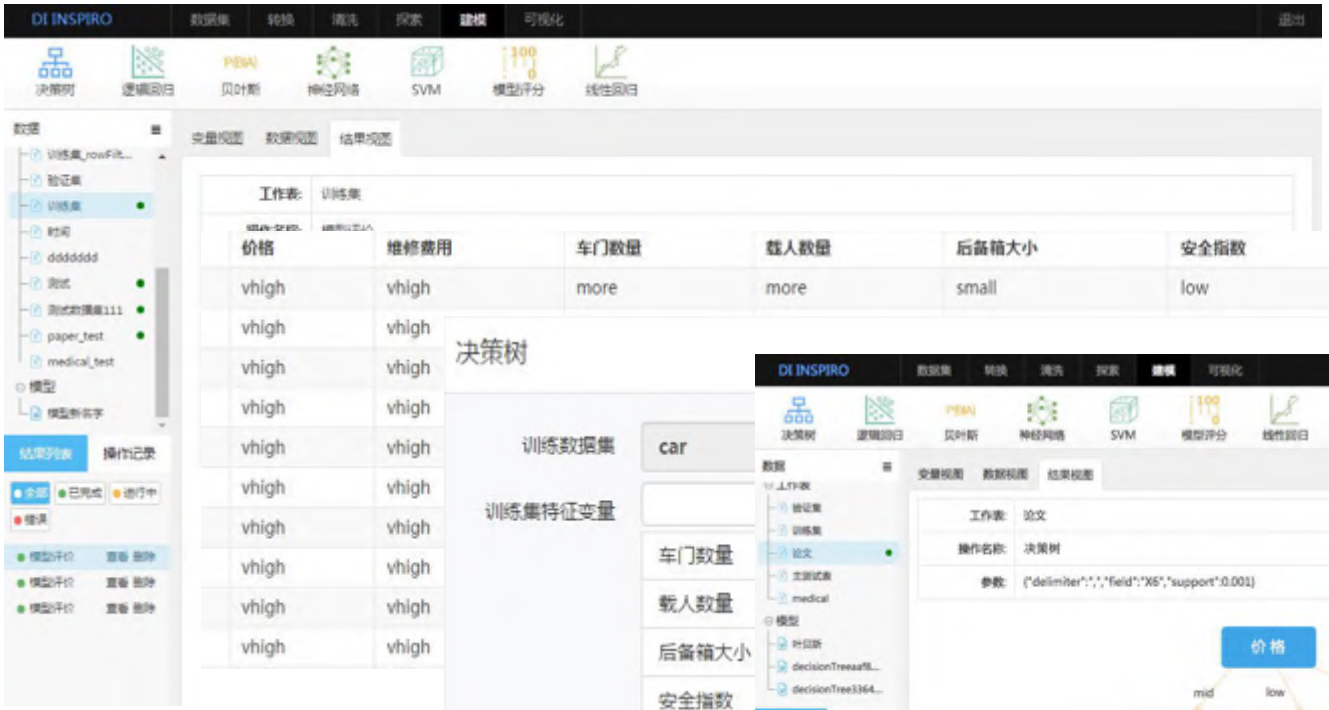
- 支持二分变量、名义变量和连续变量的神经网络
- 提供智能默认的大部分神经网络参数,如激活和误差函数，定制神经网络结构和加权
- 支持二分变量的支持向量机模型，线性和多项式内核模型训练

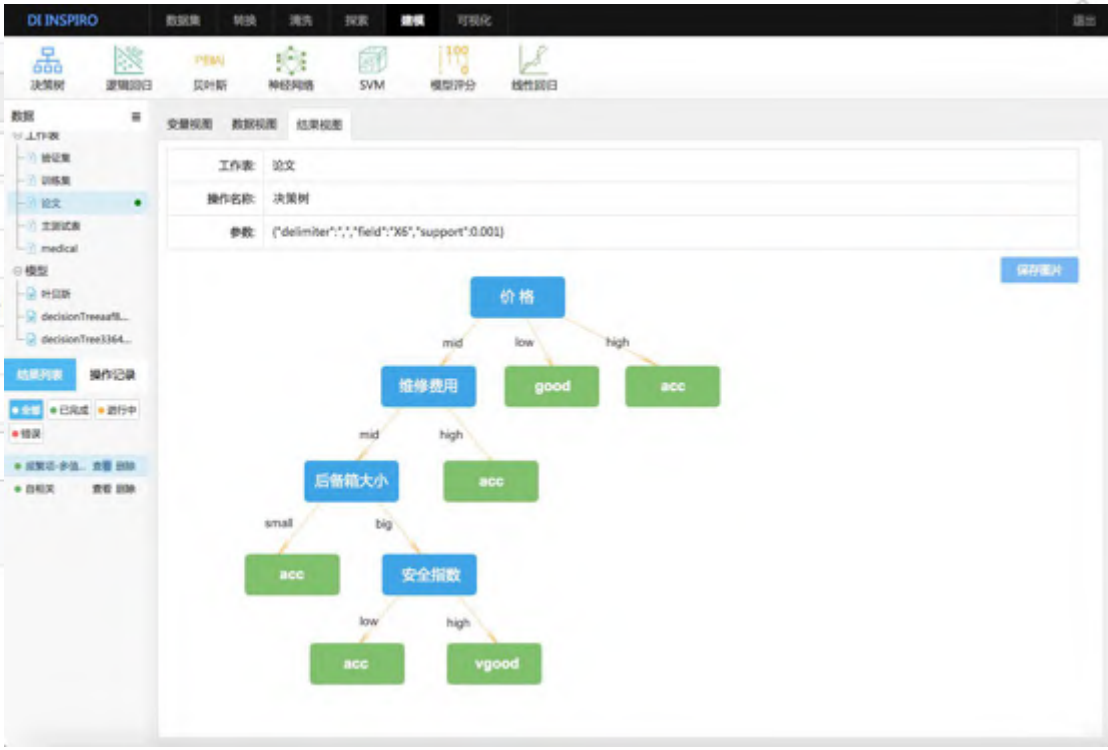
模型建立--无需编写代码，在web页面可配置模型参数进行学习、评价、调优、存储

模型评价—在界面上即可对生成的模型进行评价。

模型预测---在界面上选取保持的模型及要进行预测的数据集即可生成预测数据

决策树模型

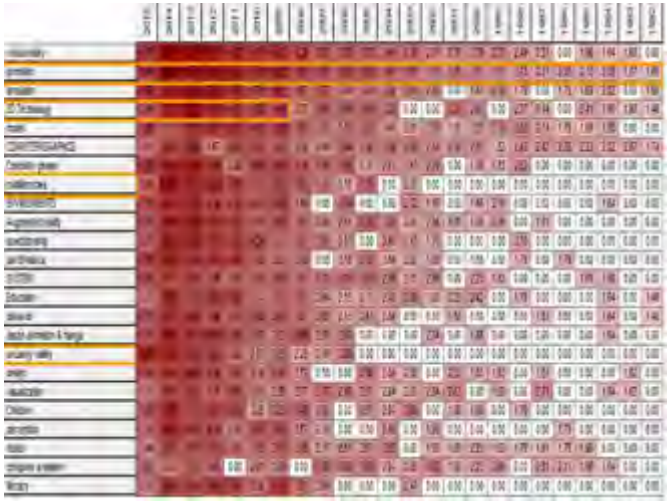




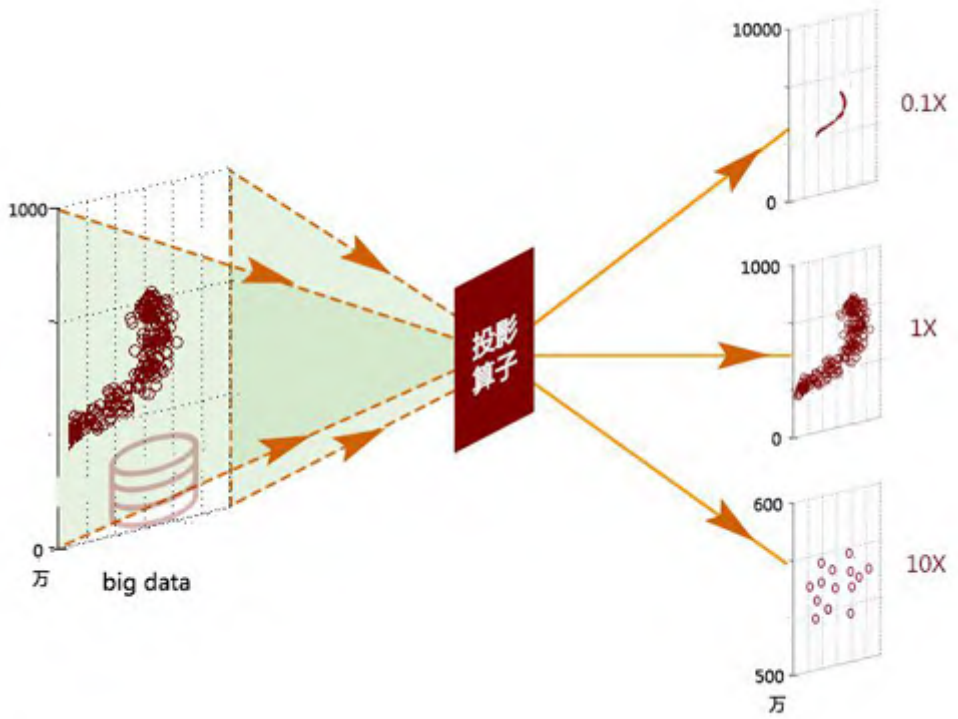
在真实汽车数据集上利用决策树
发现用户对车的评价

数据可视化

- 散点图
  - 2D、3D散点图
    - 快速的对大规模数据生成非失真的可视化图形，并支持无级缩放
- 聚类图
  - 直观查看数据的聚类结果
- 热力图
  - 相关性分析结果展示
- 基本图形
  - 点线柱图
  - 箱图、直方图—方便了解数据的总体分布



无级投影的数据可视化



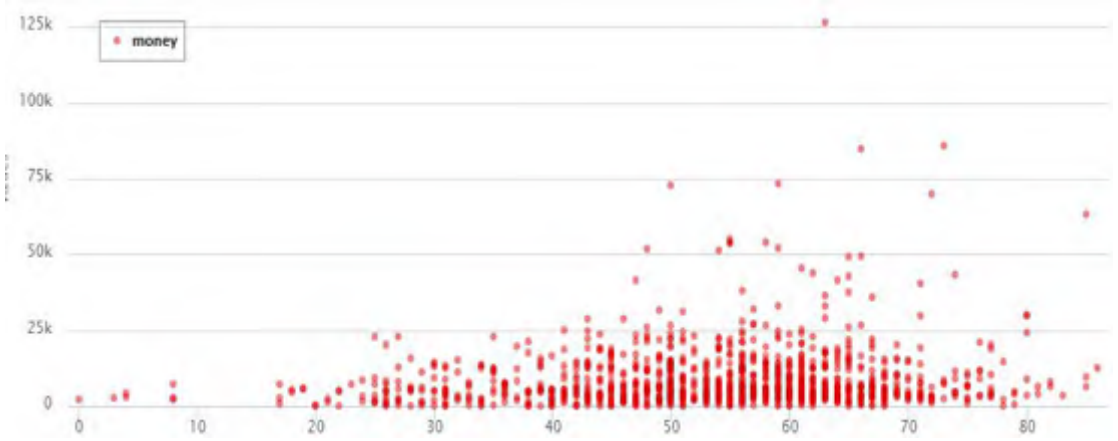
- 基于保持数据概率分布不变的思想
- 将原始数据根据缩放级别和距离远近将原始数据映射成特定显示区域的矩阵

数据分布可视化

医疗数据（真实数据）

年龄	性别	疾病种类	就诊日期	医疗费用
53	男	手术后恶性肿瘤化学治疗	2013/11/28	5398.08
44	女	恶性肿瘤维持性化学治疗	2013/11/27	5234.93
60	男	恶性肿瘤维持性化学治疗	2013/11/27	4680.67
61	男	肺恶性肿瘤	2013/11/25	4237.37
58	女	肺恶性肿瘤	2013/11/29	3714.02
73	男	慢性肾脏病5期	2013/11/28	7380.61
78	女	肺炎	2013/11/16	21720.56

患者年龄和医疗费的相对分布？

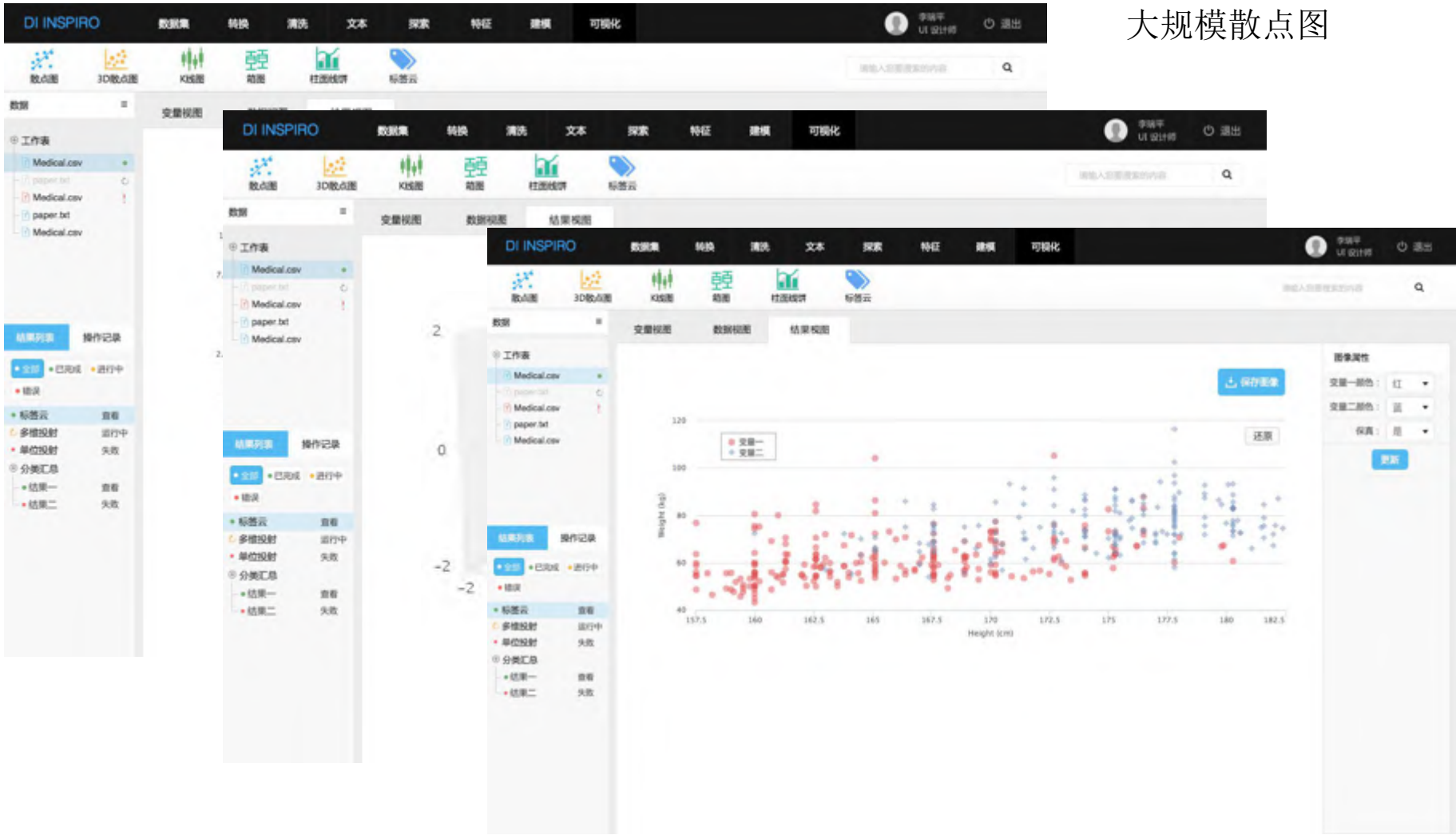


X轴：年龄，
Y轴：医疗费用
是否保真：系统会自动选用合适的粒度展示两者之间的关系。

上图反映了：

1. 患者主要集中在40岁~50岁之间
2. 医疗费用会在40~50岁之间出现大幅度波动，且总体趋势升高

可视化示例



大规模散点图

- DI设计目标
  - 旨在为普通数据分析人员提供直观易用的大数据分析处理工具
  - 无需掌握复杂的分布式集群和编程，而是直接利用web操纵大规模数据集
- 特性
  - 图形化Spark大数据分析系统
    - 基于SPARK实现的图形化交互式大数据分析系统
  - 插件式架构设计
    - 插件式设计，支持分析算子的动态加载和删除，具有高度扩展性
  - 分析功能服务化设计
    - 数据分析功能提供Restful API接口，便于与其它系统集成
  - 自适应的交互响应速度
    - 利用数据集分析历史，根据待分析文件规模对所需计算资源估计并动态分配执行资源，确保交互响应速度

开发状态和下一步计划

- 开发状态
  - 主要开发语言：python、scala
  - 核心代码行数：3500行
- 项目数据
  - 从2016年1月提出构想到开发出第一版，耗时8个月，目前还没有完全稳定
  - 2017年初计划开放公测和免费使用
  - 等代码优化到一定程度后，计划将项目开源

谢谢！

www.datainsight.tech

BDTC 2016中国大数据技术大会
Big Data Technology Conference 2016