浙江大学
ZHEJIANG UNIVERSITY

电子信息工程中的数学模型和方法

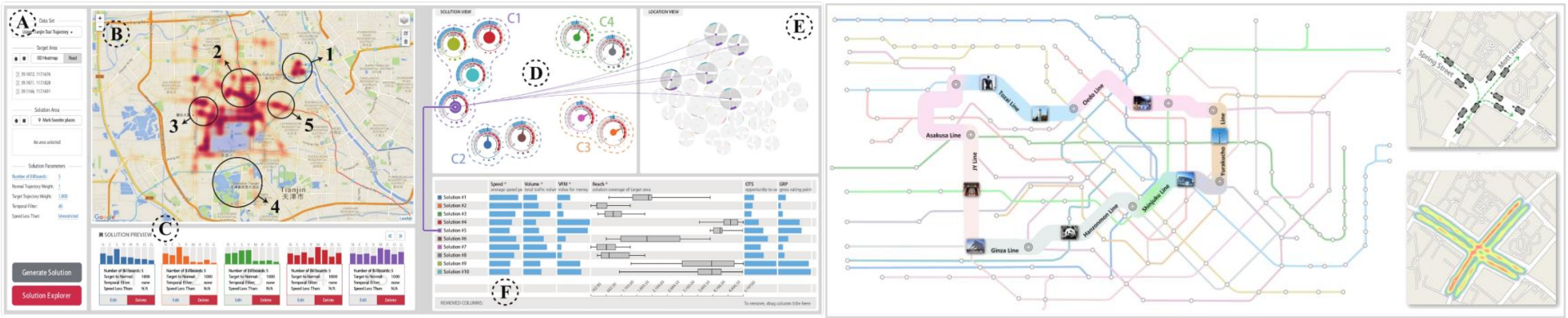
可视化建模（下）

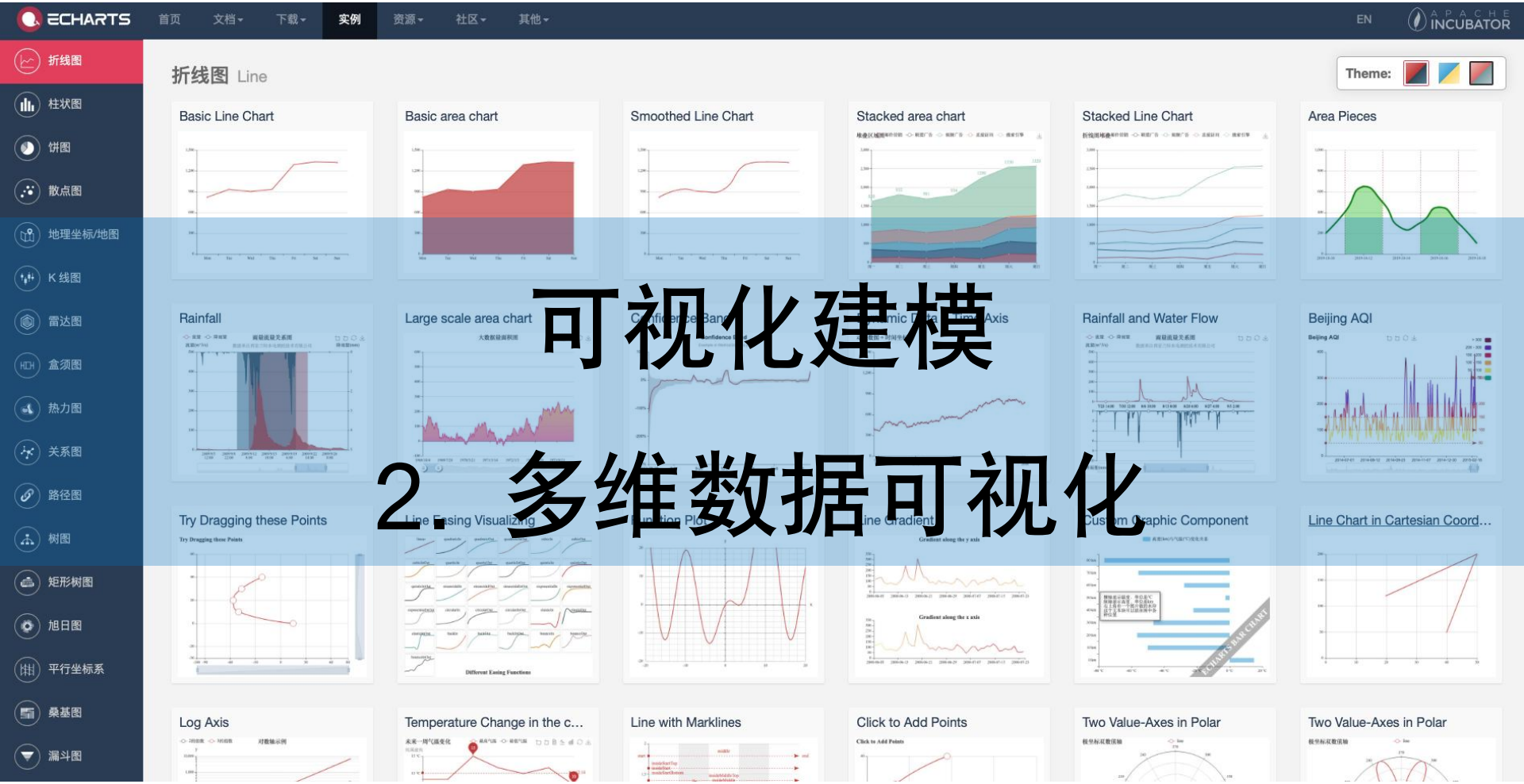
张宏鑫

浙江大学计算机学院 CAD&CG全国重点实验室

zhx@cad.zju.edu.cn

http://www.cad.zju.edu.cn/home/zhx





可视化建模

2. 多维数据可视化

多维数据可视化 – 数据案例

- 任務：由Bike Sharing Dataset中的各項外在因素(如天候，假日)與租借量的數據來建立預測模型，以預測新的外在因素下可能的租借量
- 資料集：https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
  - 本檔案是由自行車共享系統收集而來，共兩個檔案：
    - hour.csv: 記錄2011.01.01~2012.12.30每小時的外在因素及租借量，共17,379筆資料
    - day.csv: 為hour.csv資料的匯總（以日為單位）

Bike-Sharing数据集

多维数据可视化 – 数据案例

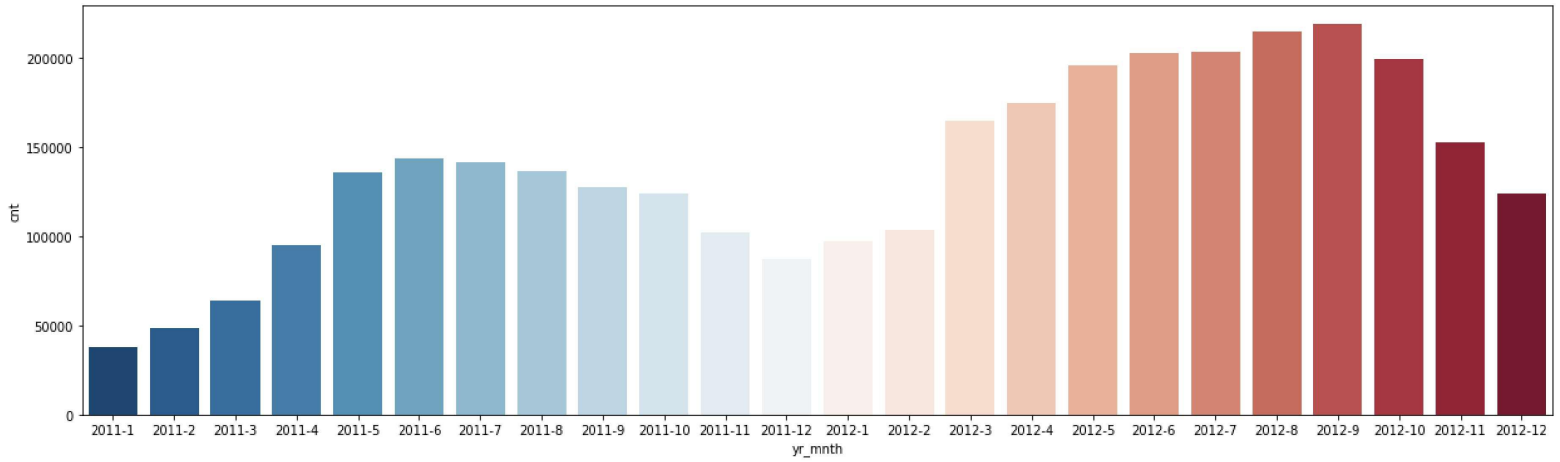
	屬性	資料型態	描述
1	instant	數值	record index
2	dteday	數值	date
3	season	類別	season (springer，summer，fall，winter)
4	yr	類別	year (2011, 2012)
5	mnth	類別	month (1 to 12)
6	hr	類別	hour (0 to 23)
7	holiday	數值	weather day is holiday or not
8	weekday	數值	day of the week
9	workingday	數值	if day is neither weekend nor holiday is 1, otherwise is 0.
10	weathersit	類別	Clear，Mist，Light Snow，Heavy Rain
11	temp	數值	Normalized temperature in Celsius. The values are divided to 41 (max)
12	atemp	數值	Normalized feeling temperature in Celsius. The values are divided to 50 (max)
13	hum	數值	Normalized humidity. The values are divided to 100 (max)
14	windspeed	數值	Normalized wind speed. The values are divided to 67 (max)
15	casual	數值	count of casual users
16	registered	數值	count of registered users
17	cnt	數值	count of total rental bikes including both casual and registered

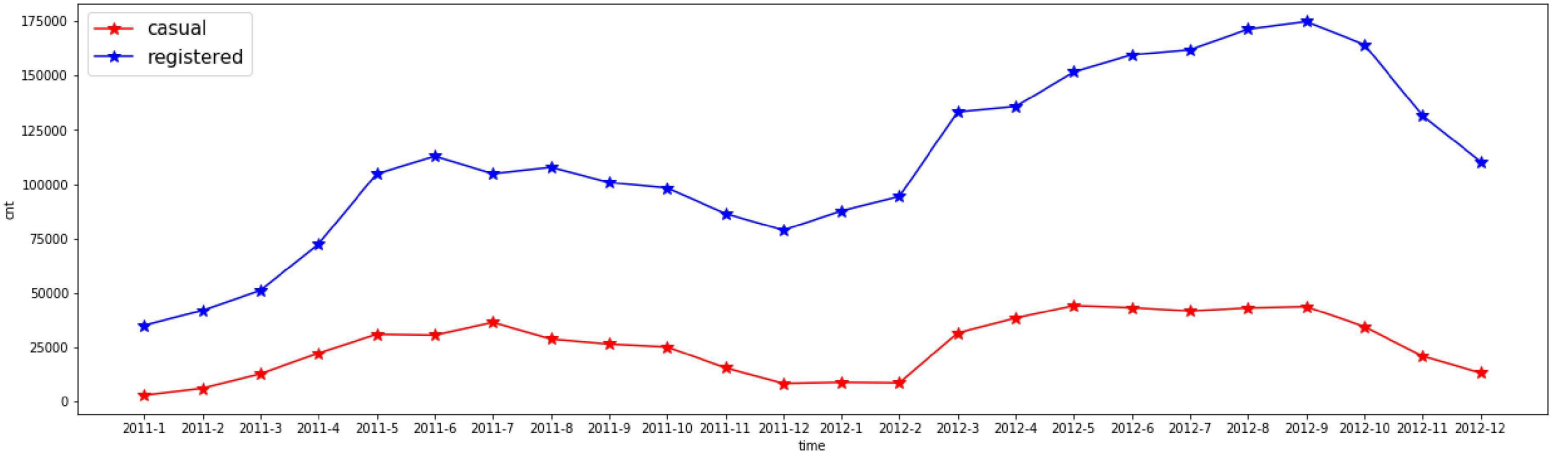
特征

标签！？

Bike-Sharing数据集

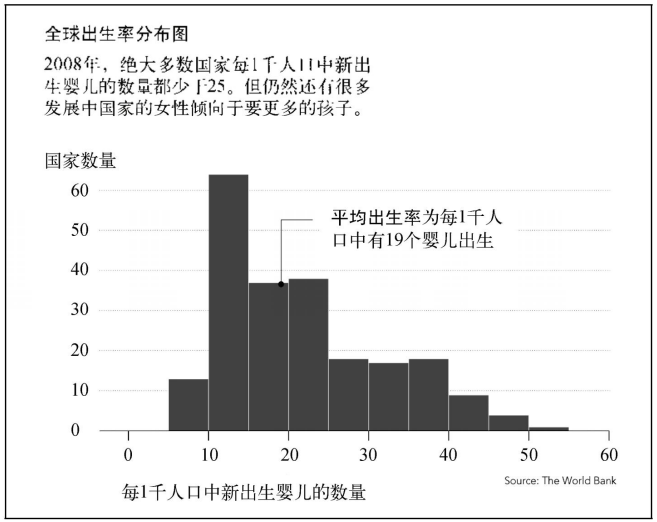
多维数据可视化 – 数据案例





Bike-Sharing数据集

数据可视化 – 图例



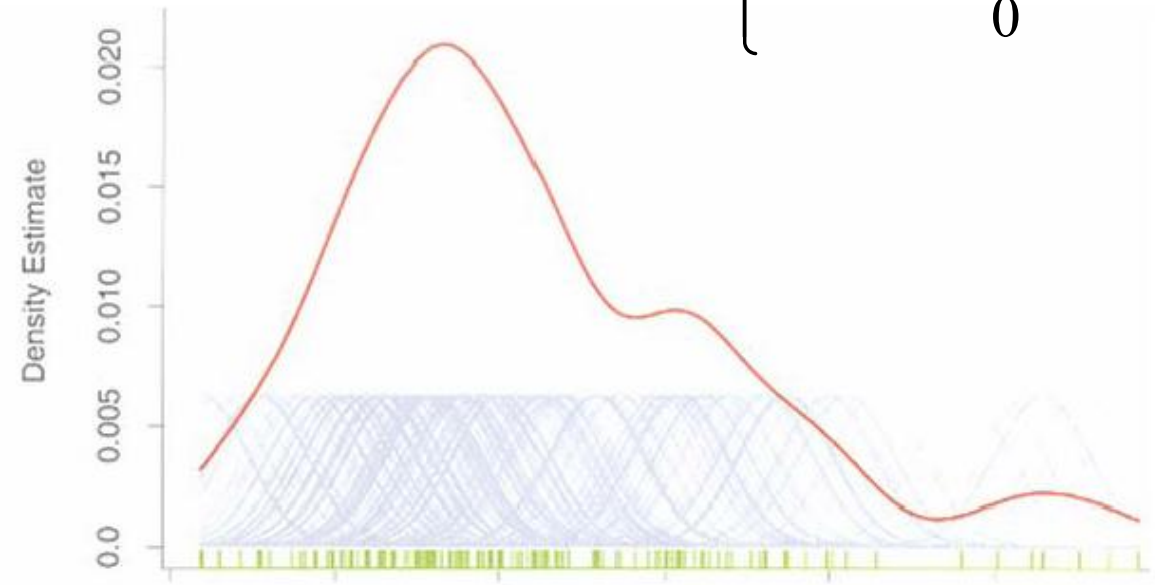
全球出生率的直方图与密度图混合

多维数据可视化 – 密度图原理

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \frac{1}{h} K(\frac{x - x_i}{h})$$

$$K_N = (2\pi)^{-d/2} \exp(-\frac{1}{2}\|\mathbf{x}\|^2)$$

$$K_E = \begin{cases} 1/2c_d^{-1}(d+2)(1-\|\mathbf{x}\|^2) & \text{if } \|\mathbf{x}\| < 1 \\ 0 & \text{otherwise} \end{cases}$$



from Hastie *et al.*

密度图与非参数估计

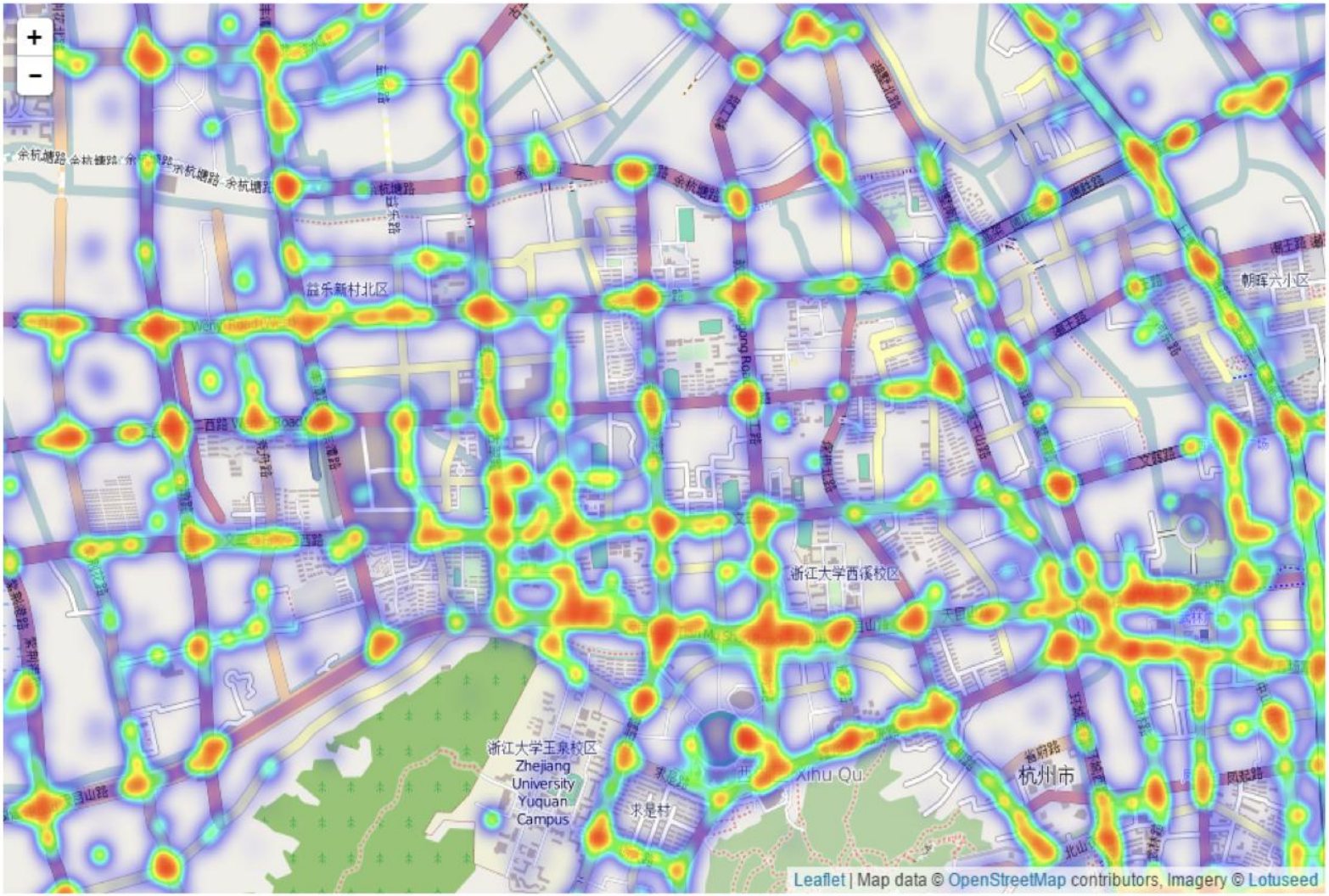
多维数据可视化 – 数据案例

usr_id	用户 id
install_ids	安装的应用程序 id 列表
date	日期
latitude	经度坐标
longitude	纬度坐标

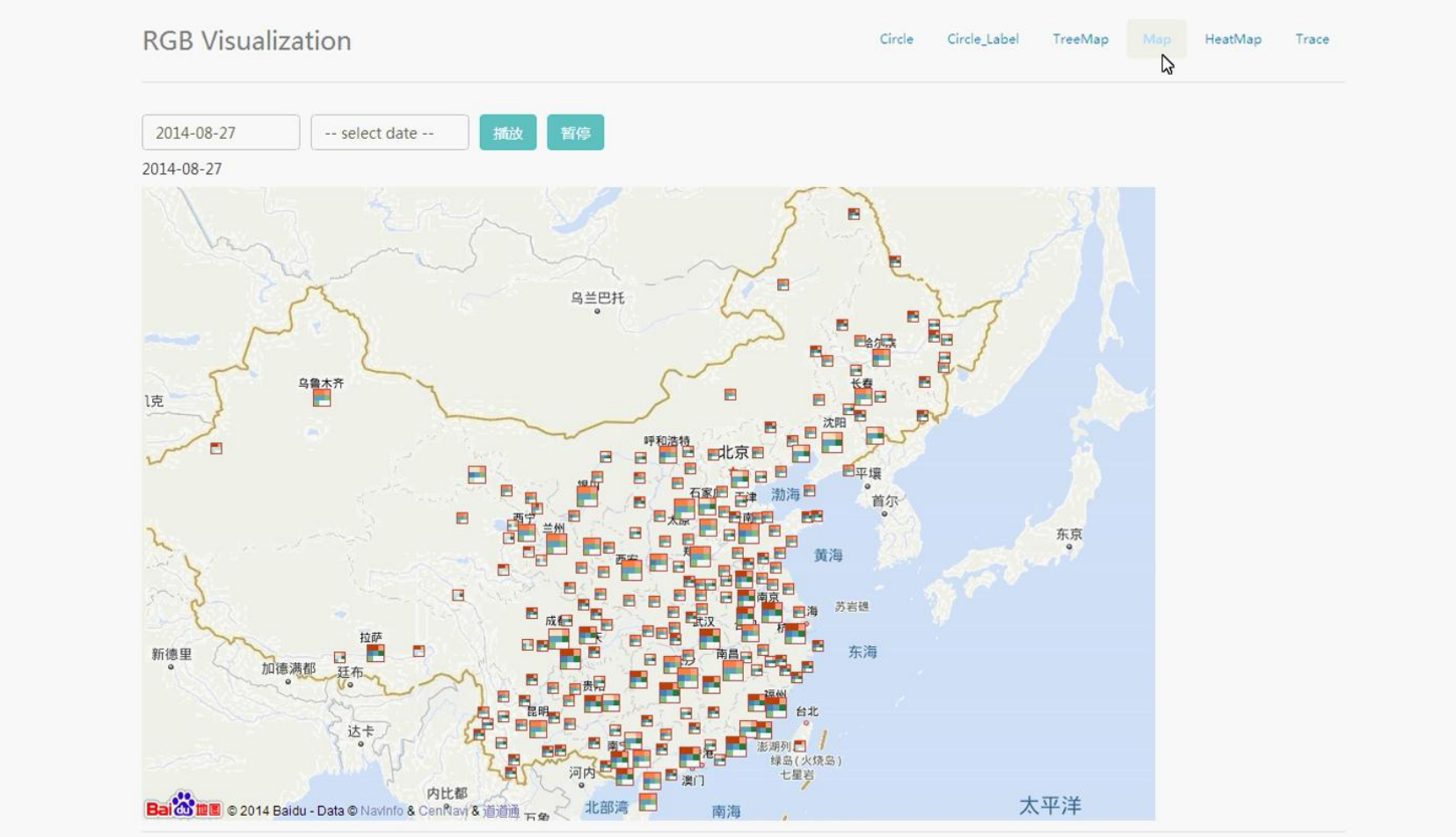
app_id	应用程序 id
category_names	分类标签名称列表

手机日志数据分析

手机日志数据分析 – 热力图



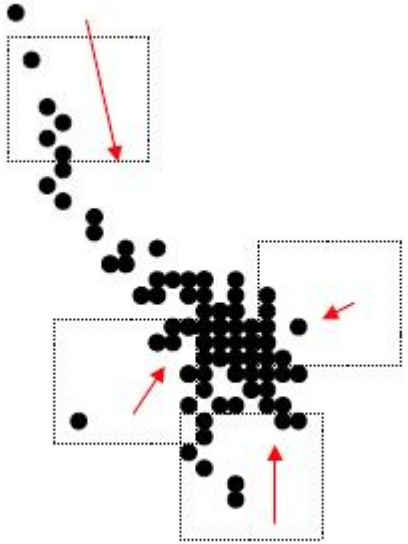
手机日志数据分析 – 热力图



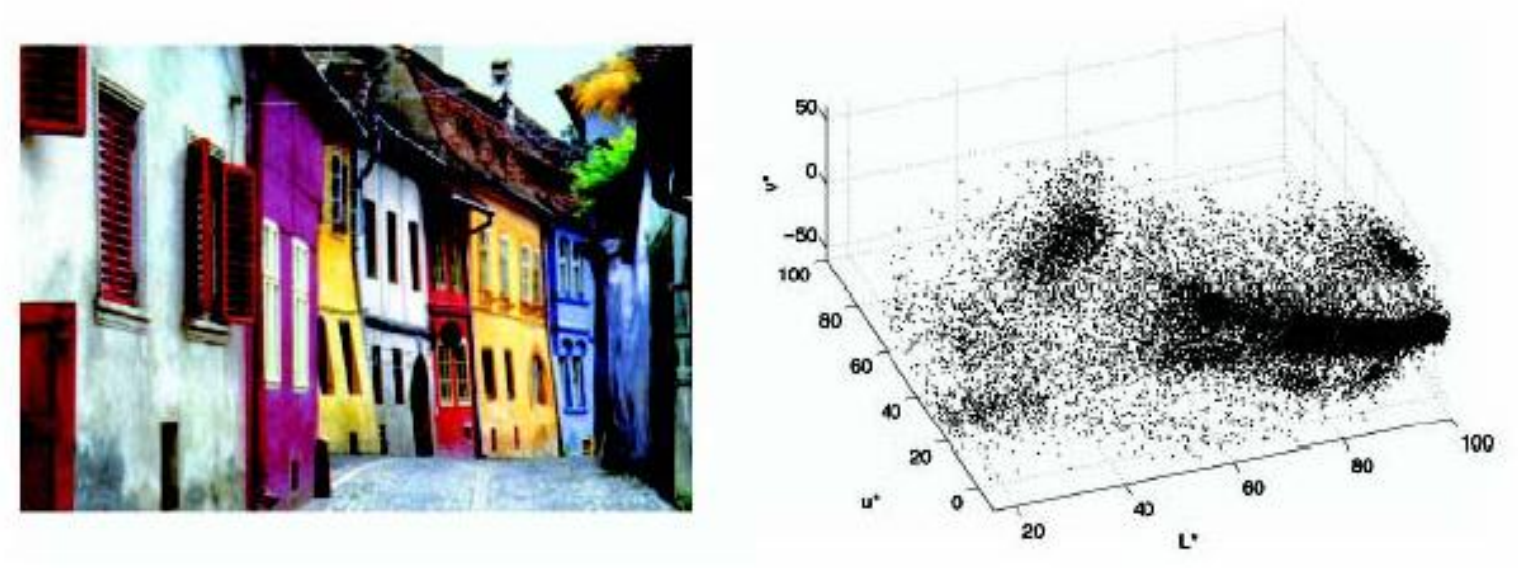
张宏鑫,盛风帆,徐沛原,汤颖. 基于移动终端日志数据的人群特征可视化. 软件学报,2016,27(5):1174-1187

多维数据可视化 – 热力图

- 如何计算热点集中区域?
- Non-parametric method to compute the nearest mode of a distribution
  - Density increases as we get near “center”

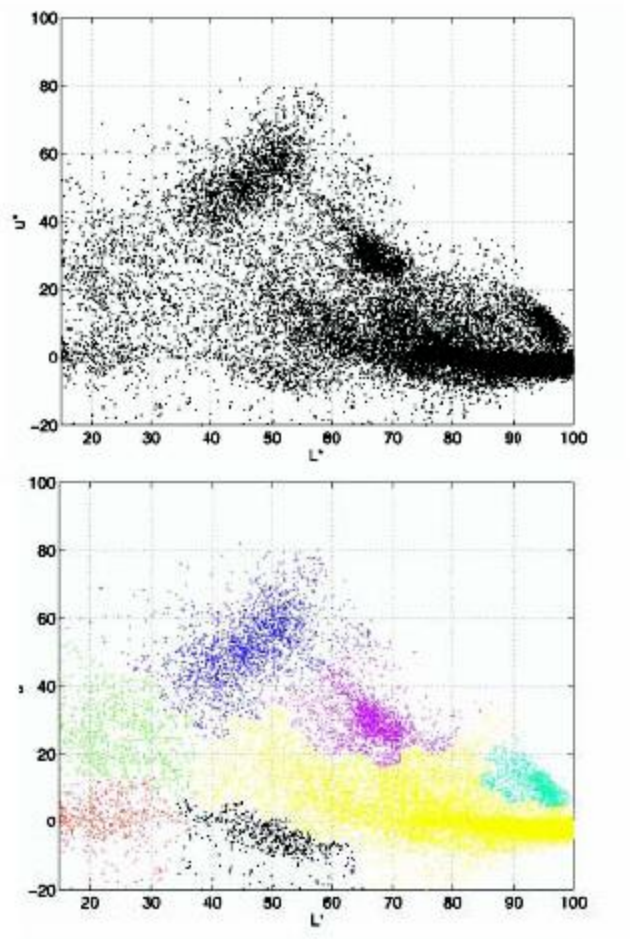


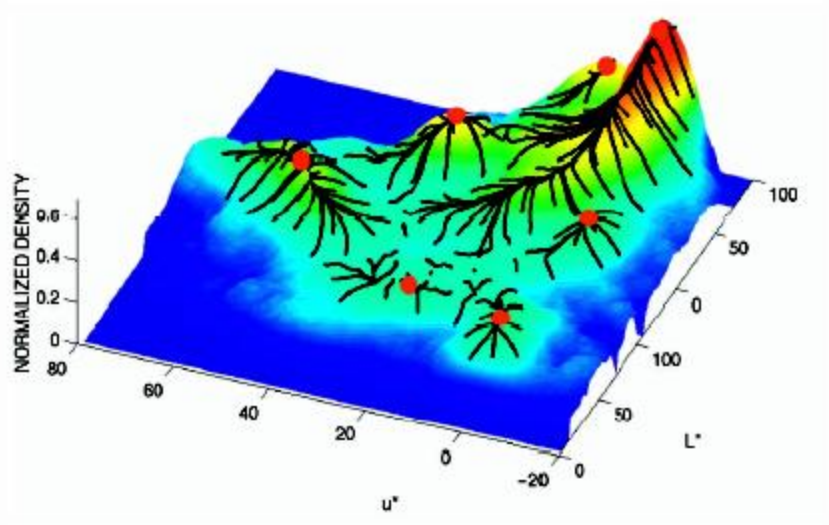
多维数据可视化 – 核密度估计



图像与其LU三维颜色直方图

多维数据可视化 – 核密度估计





图像的LU三维颜色直方图
局部模式

多维数据可视化 – 核密度估计

- Multivariate kernel density estimation

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \frac{1}{h} K(\frac{x - x_i}{h})$$

- Kernels
  - Gaussian $K_N = (2\pi)^{-d/2} \exp(-\frac{1}{2}\|\mathbf{x}\|^2)$
  - Epanechnikov

$$K_E = \begin{cases} 1/2c_d^{-1}(d+2)(1-\|\mathbf{x}\|^2) & \text{if } \|\mathbf{x}\| < 1 \\ 0 & \text{otherwise} \end{cases}$$

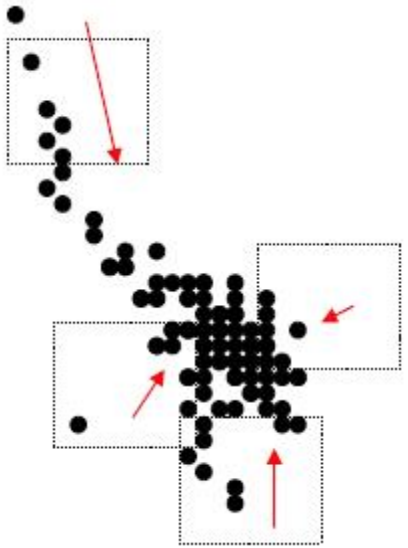
多维数据可视化 – Mean-Shift

- Gradient computation
  - For symmetric kernel

$$\hat{\nabla} f(\mathbf{x}) = \frac{2}{nh^{d+2}} \sum_{i=1}^{n} K_N\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \left[\frac{\sum_{i=1}^{n} \mathbf{x}_i K_N\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^{n} K_N\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)} - \mathbf{x} \right]$$

- Always converges to the local maximum!

多维数据可视化 – Mean-Shift



- Give a point x
  1. Compute the mean shift vector

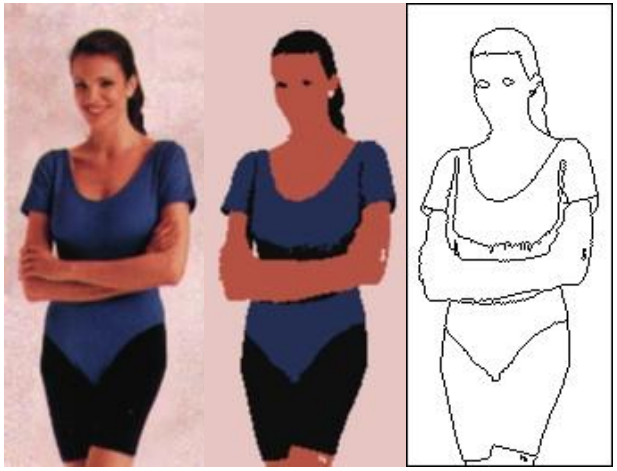
$$\hat{\nabla} f(\mathbf{x}) = \frac{2}{nh^{d+2}} \sum_{i=1}^{n} K_N\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \left[\frac{\sum_{i=1}^{n} \mathbf{x}_i K_N\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^{n} K_N\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)} - \mathbf{x} \right]$$

  2. Translate density estimation window:

$$\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \hat{\nabla} f(\mathbf{x}^{(t)})$$

  3. Iterate steps 1. and 2. until convergence i.e., $\hat{\nabla} f(\mathbf{x}) \to 0$

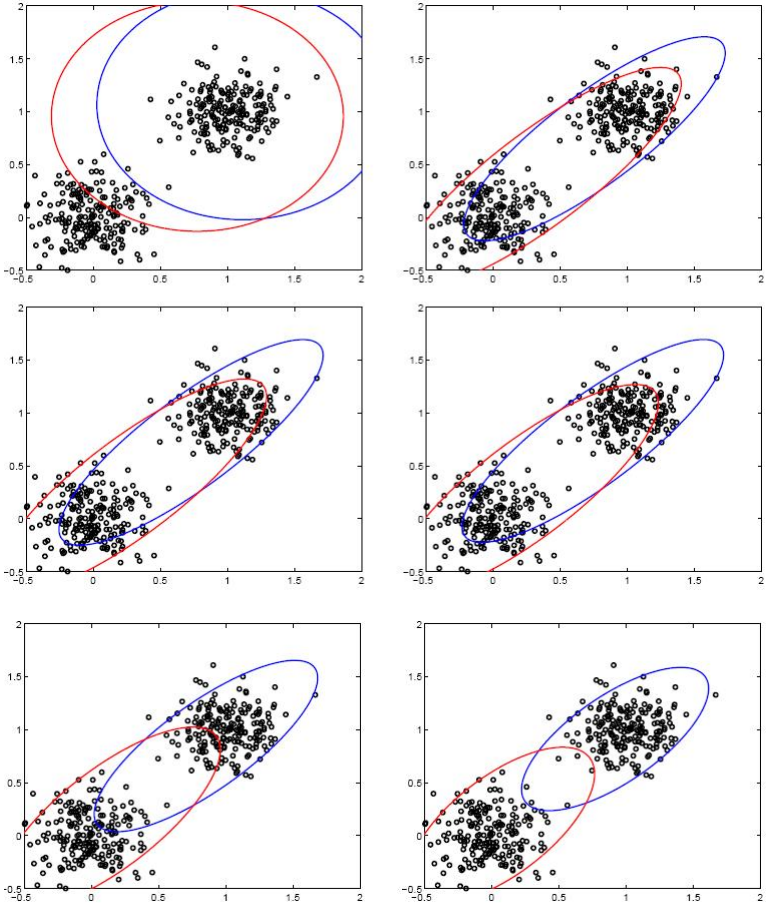
多维数据可视化 – 数据聚类

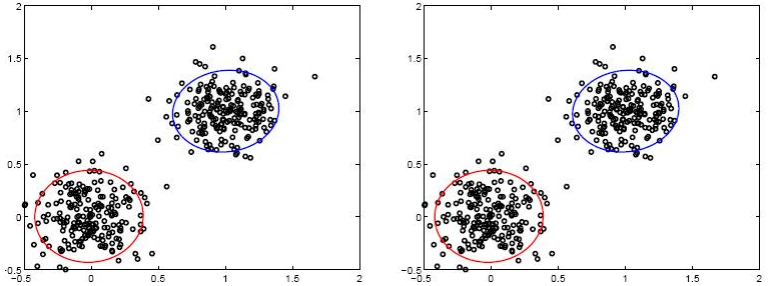


Mean-shift segmentation

多维数据可视化 – 数据聚类 - MOG

- Mixture of Gaussians and EM algorithm





多维数据可视化 – 数据聚类 - MOG

- Mixture distribution:
  - Assume $P(\mathbf{x})$ is a mixture of K different Gaussians
  - Assume each data point, x is generated by 2-step process
    - Choose one of the K Gaussians as $N(\mu_z, \Sigma_z)$ label z
    - Generate $\mathbf{x}$ according to the Gaussian

$$P(\mathbf{x}) = \sum_{z=1}^{K} P(Z = z \mid \pi) N(\mathbf{x} \mid \mu_z, \Sigma_z)$$

- What object function shall we optimize?
  - Maximize data likelihood

多维数据可视化 – 数据聚类 - MOG

E-step: softly assign examples to mixture components

$$\hat{p}(j|i) \leftarrow P(y_i = j|\mathbf{x}_i, \theta), \quad \text{for all } j = 1, 2 \text{ and } i = 1, \ldots, n$$

M-step: re-estimate the parameters (separately for the two Gaussians) based on the soft assignments.

$$\hat{n}_j \leftarrow \sum_{i=1}^{n} \hat{p}(j|i) = \text{Soft } \# \text{ of examples labeled } j$$

$$\hat{p}_j \leftarrow \frac{\hat{n}_j}{n}$$

$$\hat{\mu}_j \leftarrow \frac{1}{\hat{n}_j} \sum_{i=1}^{n} \hat{p}(j|i)\, \mathbf{x}_i$$

$$\hat{\Sigma}_j \leftarrow \frac{1}{\hat{n}_j} \sum_{i=1}^{n} \hat{p}(j|i)\, (\mathbf{x}_i - \hat{\mu}_j)(\mathbf{x}_i - \hat{\mu}_j)^T$$

多维数据可视化 – 数据聚类 – K-means

Given data <x_1 … x_n>, and K, assign each x_i to one of K clusters, C_1 … C_K , minimizing $J = \sum_{j=1}^{K} \sum_{x_i \in C_j} ||x_i - \mu_j||^2$

Where μ_j is mean over all points in cluster C_j

K-Means Algorithm:

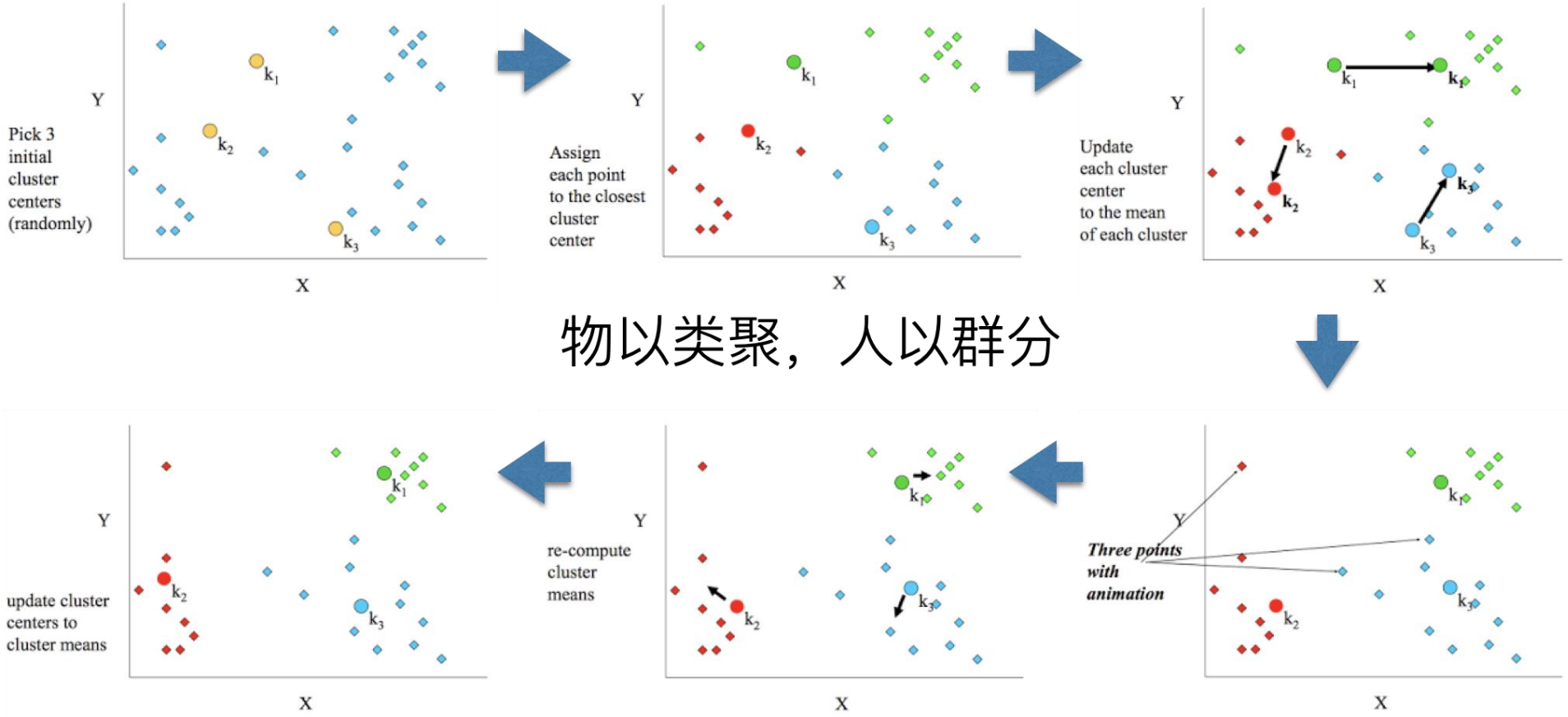
Initialize $\mu_1 \cdots \mu_K$ randomly

Repeat until convergence:

1. Assign each point x_i to the cluster with the closest mean μ_j
2. Calculate the new mean for each cluster

$$\mu_j \leftarrow \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

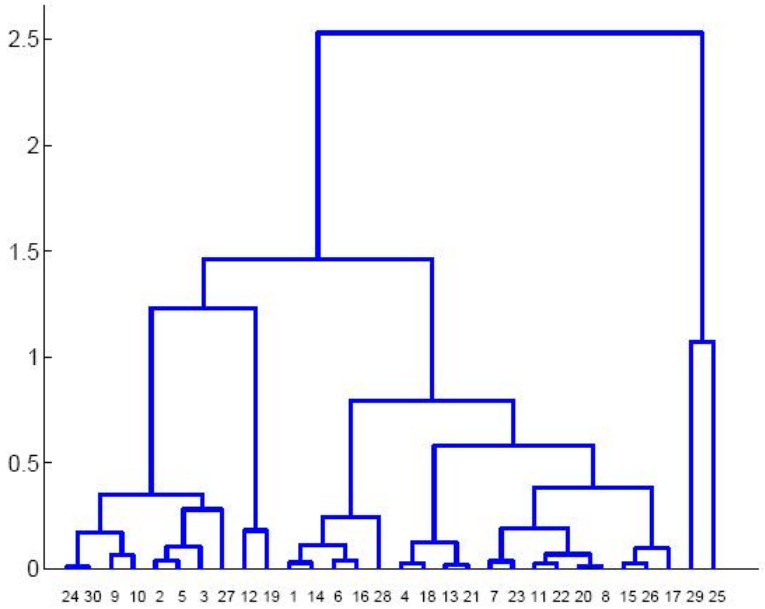
多维数据可视化 – 数据聚类 – K-means



物以类聚，人以群分

多维数据可视化 – 数据层次聚类

- A dendrogram representation of hierarchical clustering



The height of each pair represents the distance between the merged clusters; the specific linear ordering of points is chosen for clarity

自底向上的聚类方法
（bottom-up）

多维数据可视化 – 数据层次聚类

- Hierarchical agglomerative clustering: we sequentially merge the pair of "closest" points/clusters

- The procedure
  1. Find two closest points (clusters) and merge them
  2. Proceed until we have a single cluster (all the points)
- Two prerequisites:
  1. distance measure d(xi, xj) between two points
  2. distance measure between clusters (cluster linkage)

自底向上的聚类方法
(bottom-up)

多维数据可视化 – 数据层次聚类

- A *linkage* method: we have to be able to measure distances between clusters of examples C_k and C_l

a) Single linkage:

Nearest neighbor

$$d_{kl} = \min_{i \in C_k, j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j)$$

b) Average linkage:

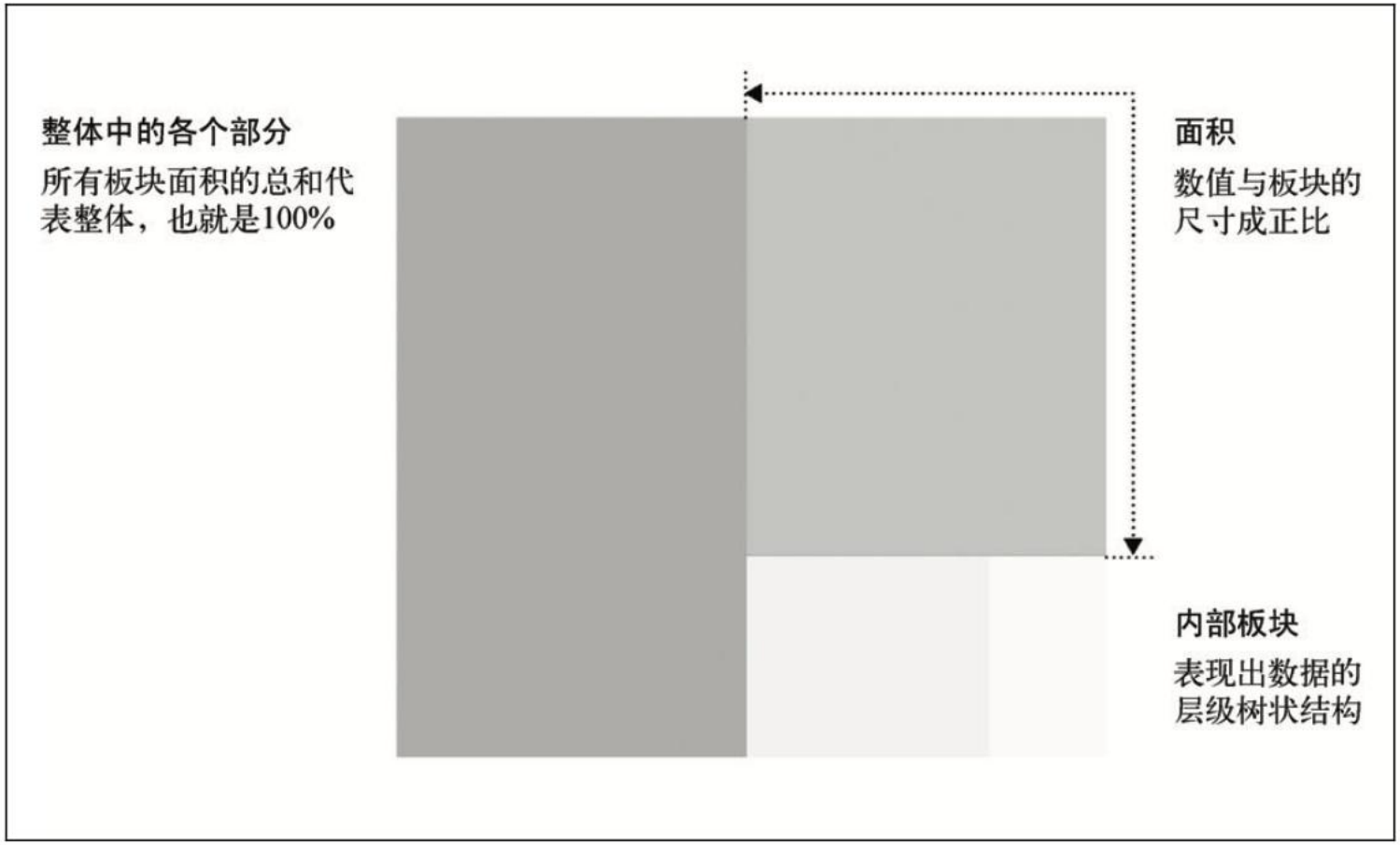
$$d_{kl} = \frac{1}{|C_l|\,|C_k|} \sum_{i \in C_k, j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j)$$

c) Centroid linkage:

$$d_{kl} = d(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_l), \quad \bar{\mathbf{x}}_l = \frac{1}{|C_l|} \sum_{i \in C_l} \mathbf{x}_i$$

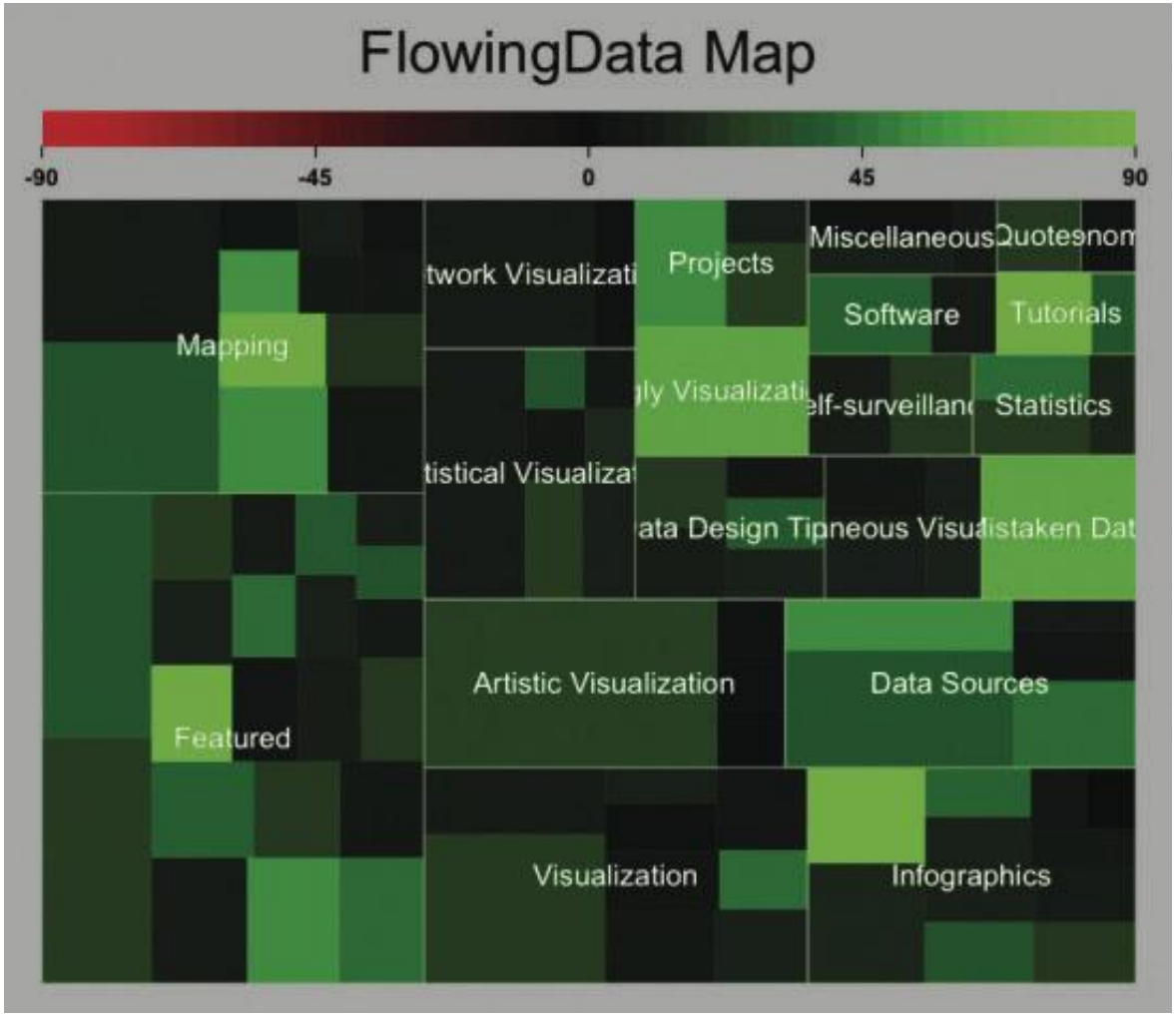
自底向上的聚类方法
（bottom-up）

多维数据可视化 – 数据层次 - Treemap



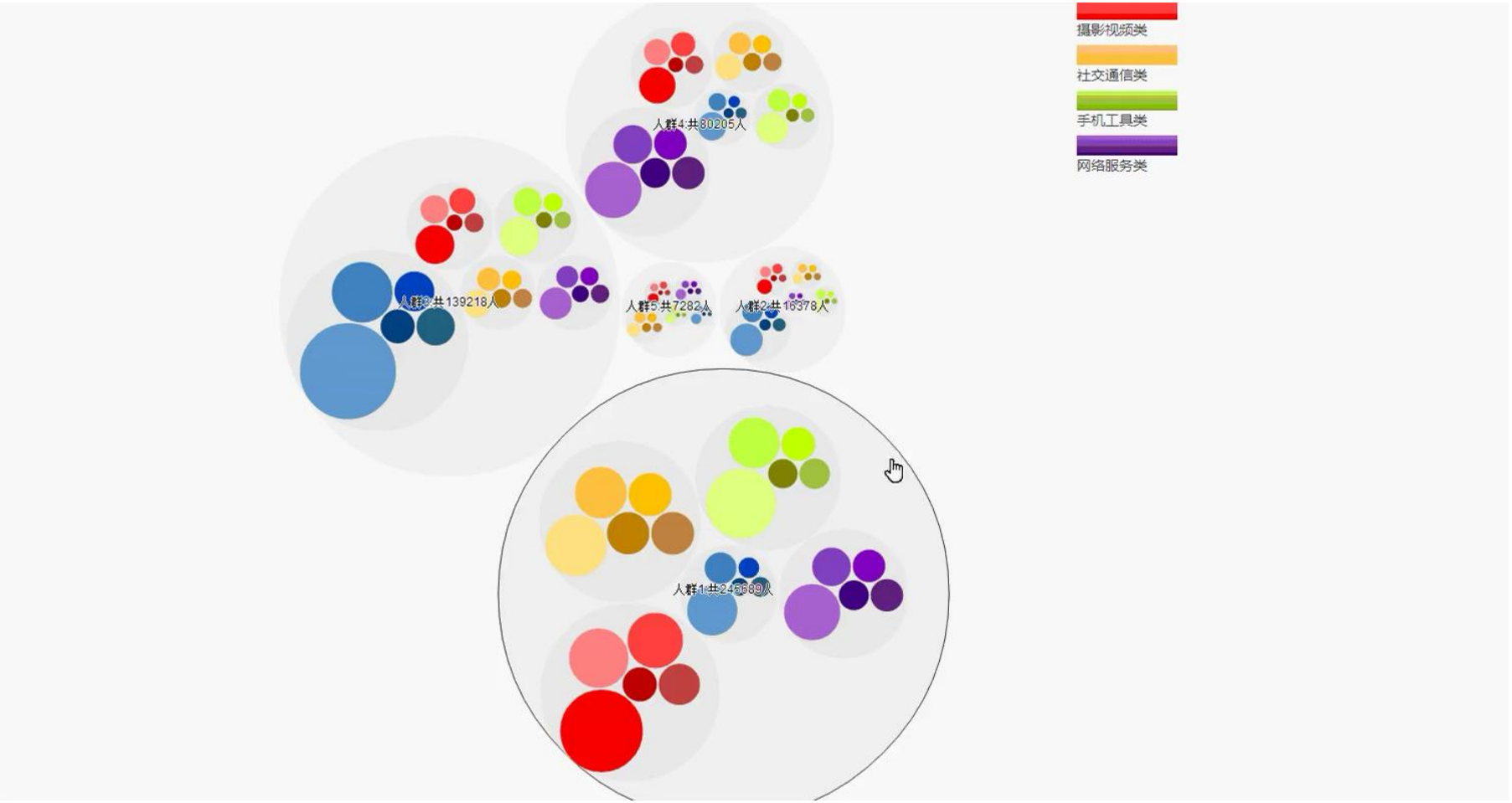
Treemap的基本框架

多维数据可视化 – 数据层次 - Treemap



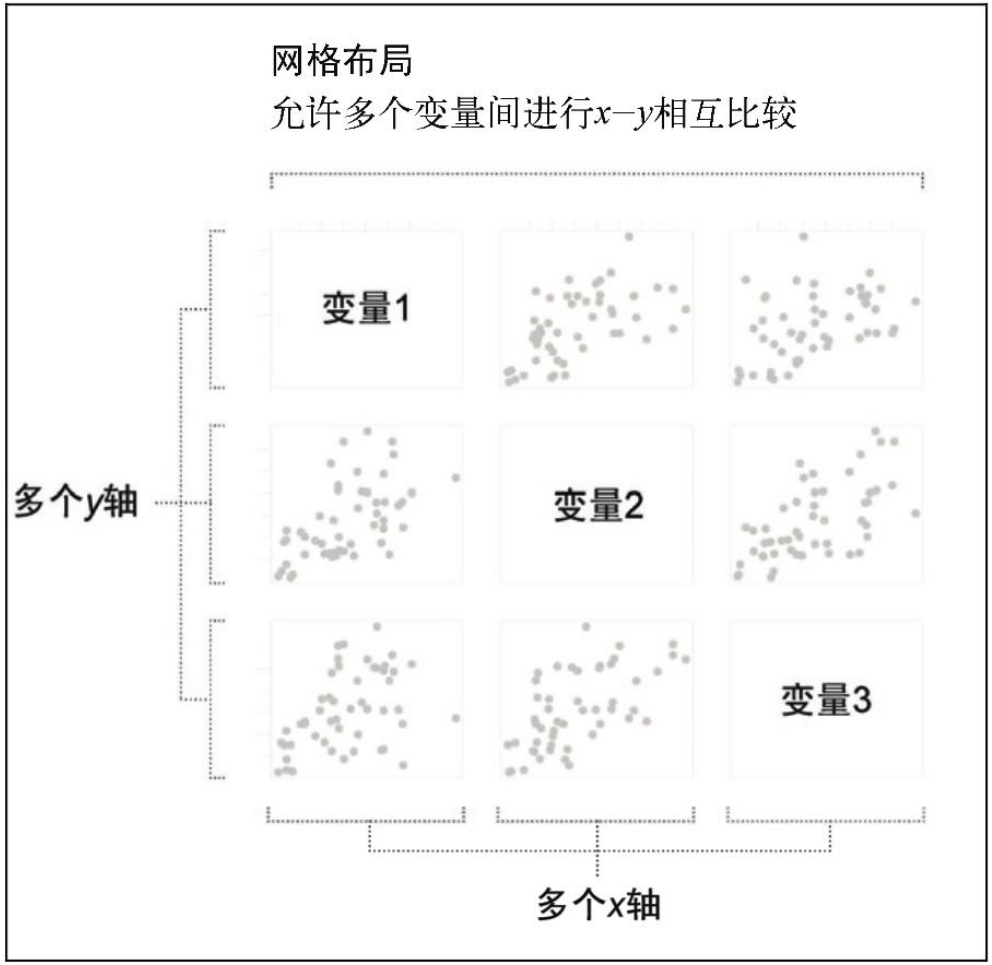
Treemap的图例

多维数据可视化 – 数据层次 – 气泡图



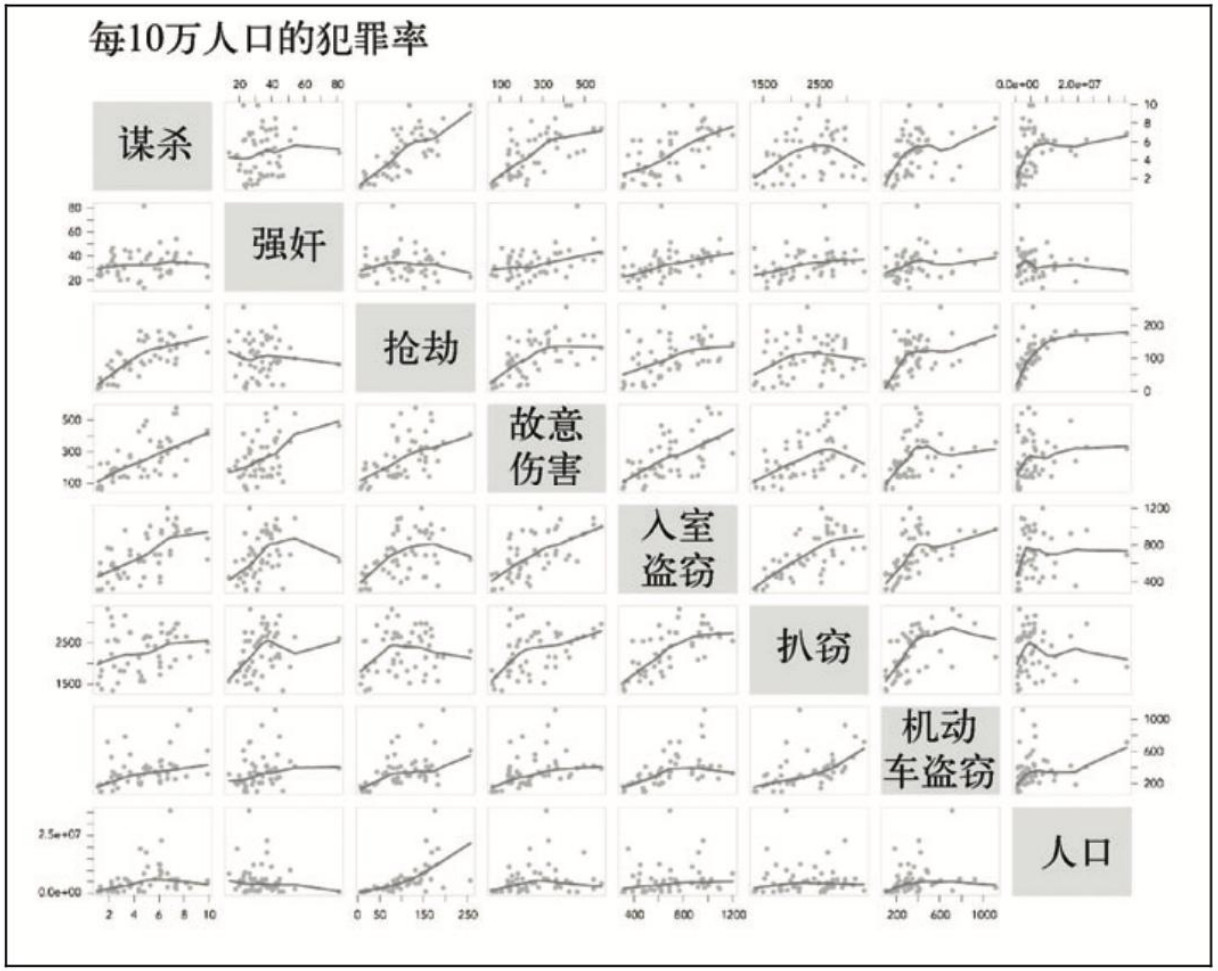
层次气泡图与Voronoi图的例子
根据手机应用情况对人群进行聚类

多维数据可视化 – 散点图矩阵



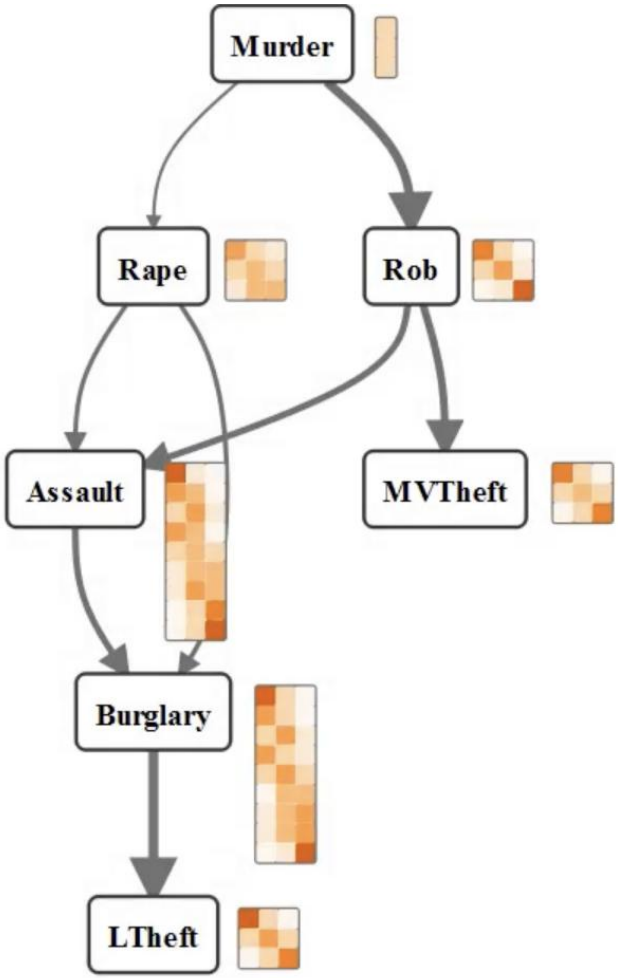
散点图矩阵的基本框架

多维数据可视化 – 散点图矩阵图例



美国各州犯罪率的散点图矩阵

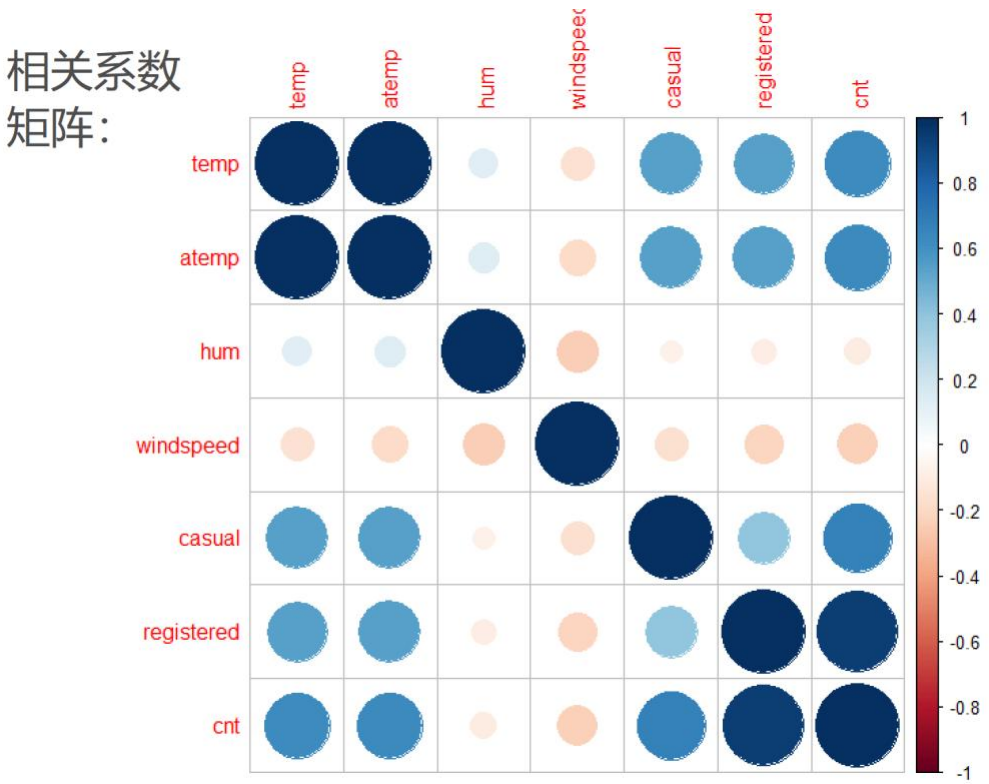
多维数据可视化 – 散点图矩阵图例



美国各州犯罪率数据分析：采用贝叶斯网络方法的分析结果

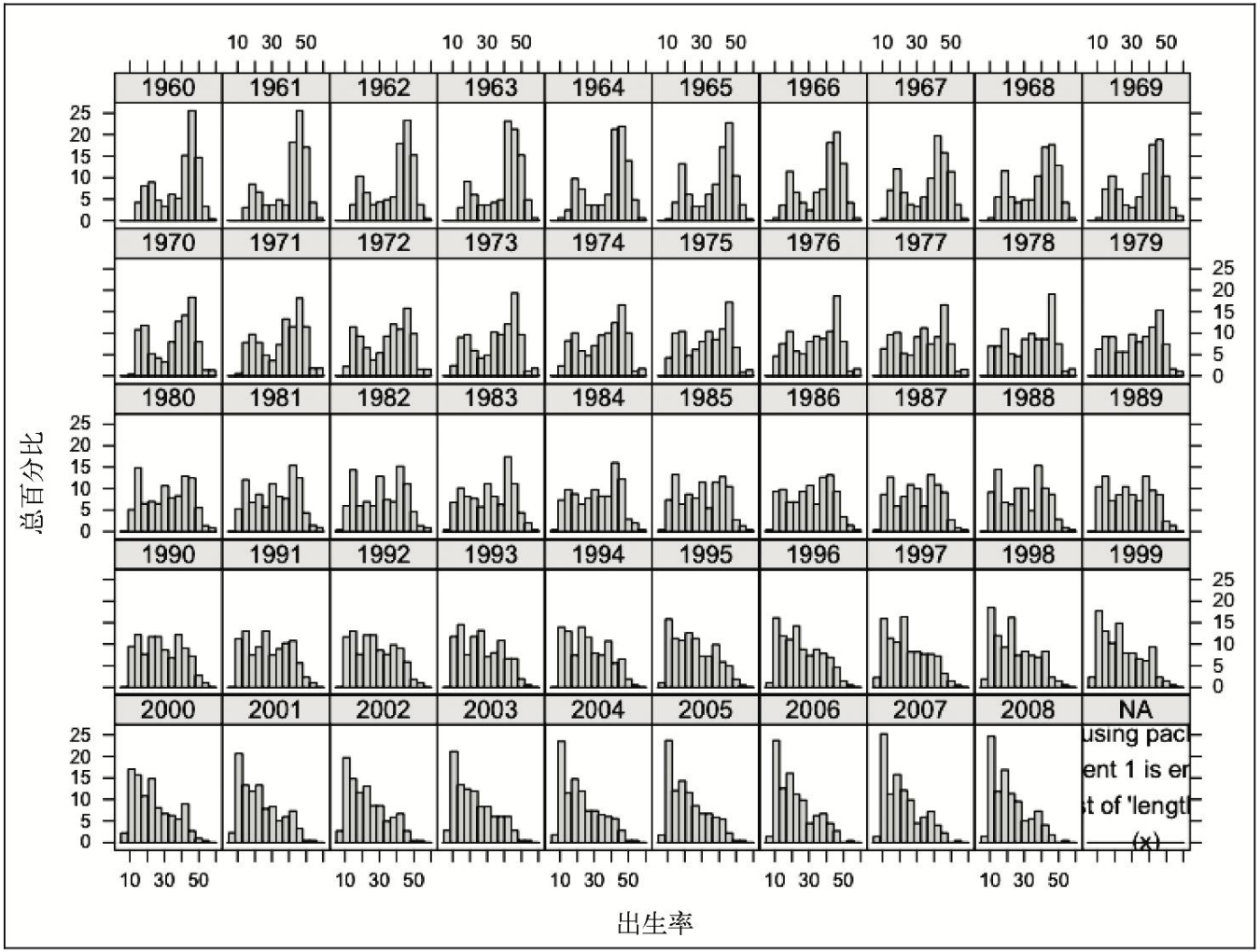
BN-Mapping:基于贝叶斯网络的地理空间数据可视分析

多维数据可视化 – 关联分析



Bike-Sharing数据集

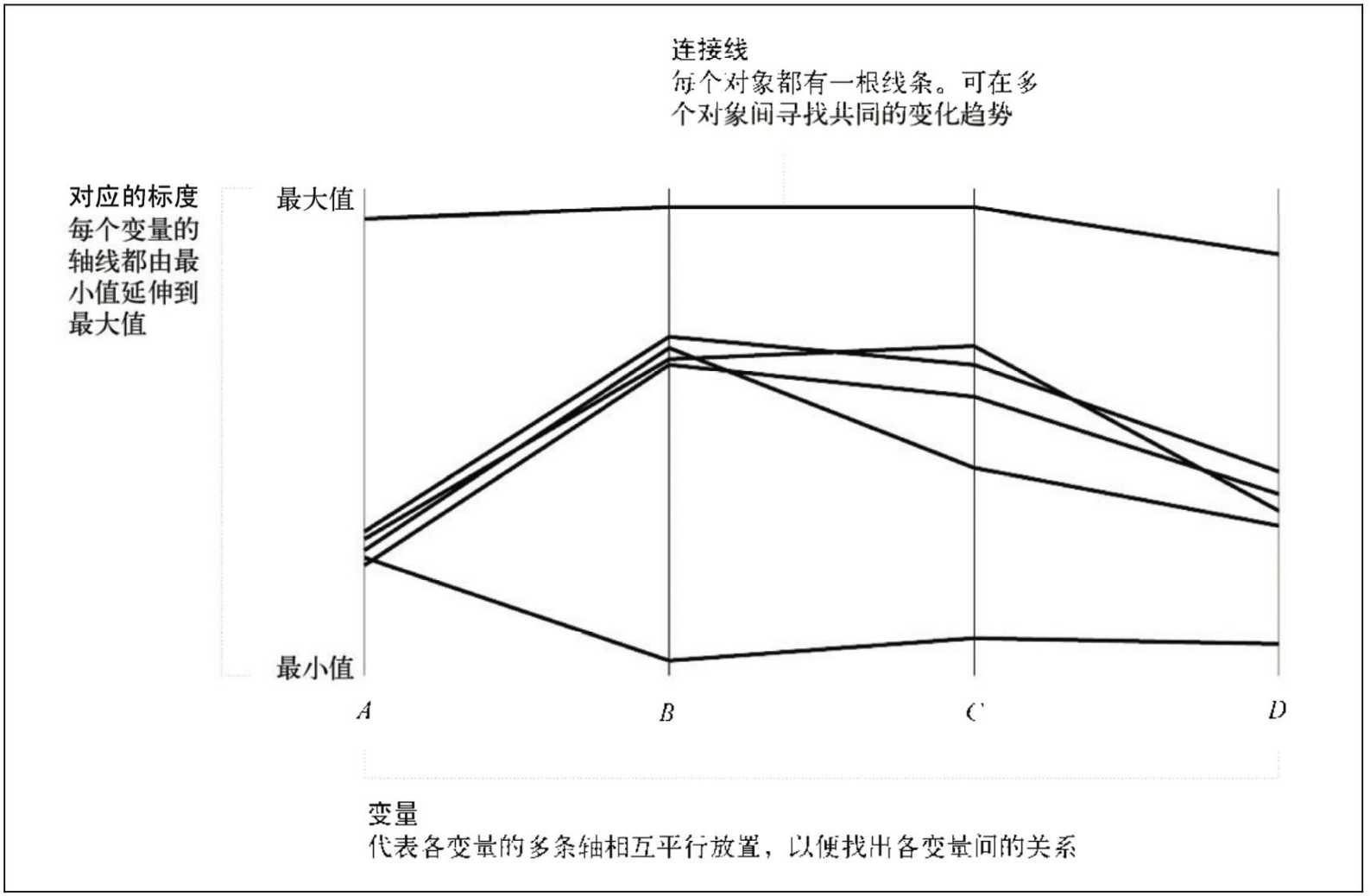
多维数据可视化 – 直方图矩阵图例



全球出生率的直方图矩阵

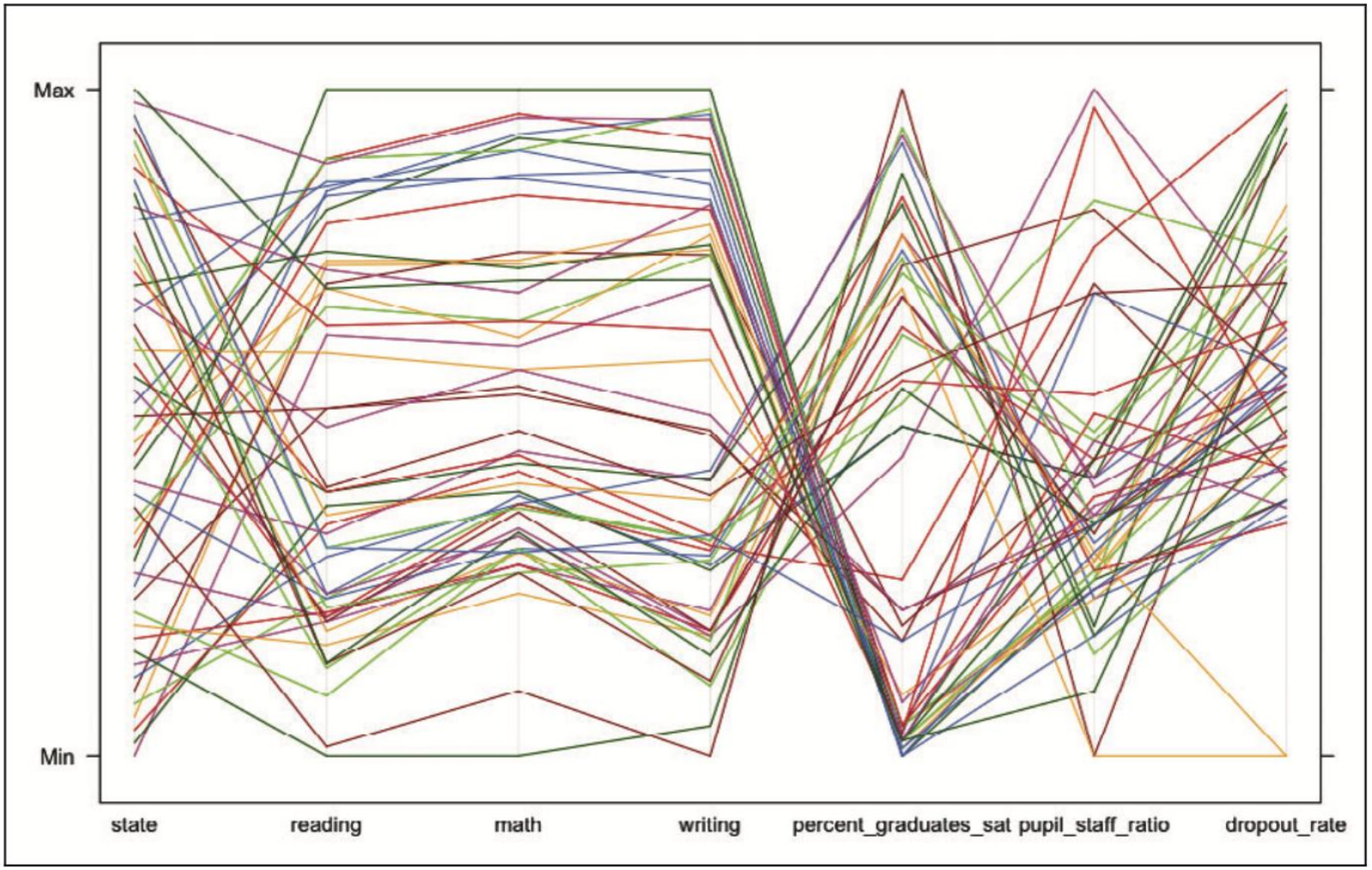
(思考题：我国建国以来的人口出生率变化趋势分析)

多维数据可视化 – 平行坐标图



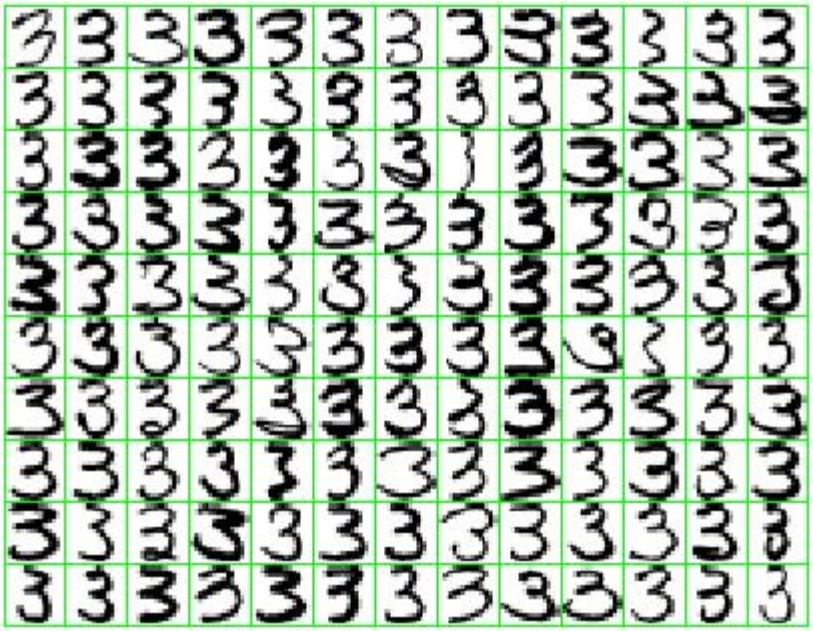
平行坐标图的基本框架

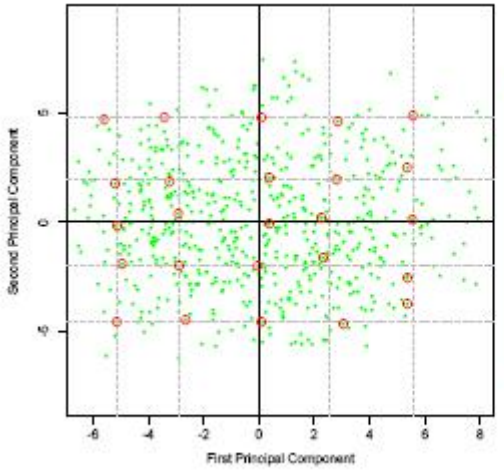
多维数据可视化 – 平行坐标图例

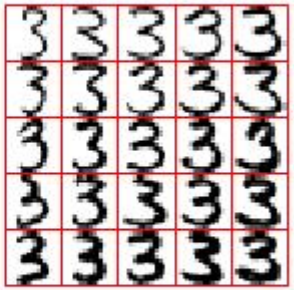


美国各州高中生SAT平均成绩数据

多维数据可视化 – 降维

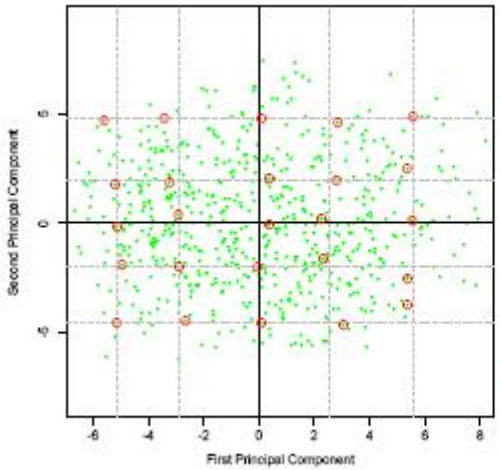


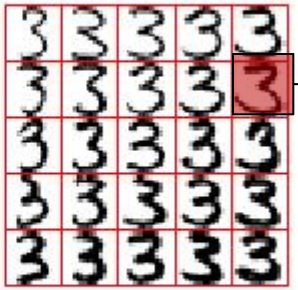




手写数字数据集的可视化

多维数据可视化 – 降维 – 特征抽取





Two-component model has the form

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$$= \text{  } + \lambda_1 \cdot \text{  } + \lambda_2 \cdot \text{  }.$$

$$y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_d \end{pmatrix}_{d\times 1}$$

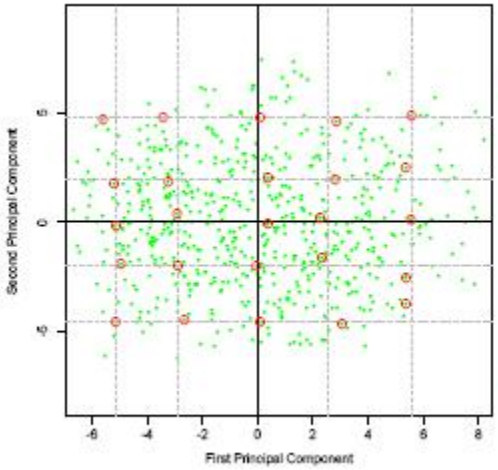
constant　Low frequency component　High frequency component

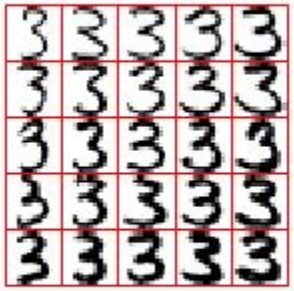
多维数据可视化 – 降维 – 特征抽取

$$X = \begin{pmatrix} x_{0,0} & x_{1,0} & x_{2,0} & \cdots & x_{N-1,0} \\ x_{0,1} & x_{1,1} & x_{2,1} & \cdots & x_{N-1,1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{0,p-1} & x_{1,p-1} & x_{2,p-1} & \cdots & x_{N-1,p-1} \end{pmatrix}_{p \times N}$$

130 threes, a subset of 638 such threes and part of the handwritten digit dataset. Each "three" is a 16 × 16 grayscale image, and the variables x_j, $j = 1, \ldots, 256$ are the grayscale values for each pixel.

多维数据可视化 – 降维 – 特征抽取





Two-component model has the form

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$$= \bar{x} + \lambda_1 \cdot \quad + \lambda_2 \cdot \quad .$$

Here we have displayed the first two principal component directions, v_1 and v_2, as images.

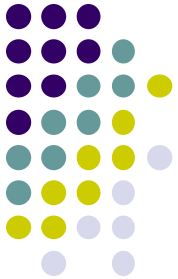
$$\boldsymbol{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_d \end{pmatrix}_{d\times 1} \approx \begin{pmatrix} \bar{x}_0 \\ \bar{x}_1 \\ \vdots \\ \bar{x}_d \end{pmatrix}_{d\times 1} + w_1 \begin{pmatrix} h_1(x_0) \\ h_1(x_1) \\ \vdots \\ h_1(x_d) \end{pmatrix}_{d\times 1} + w_2 \begin{pmatrix} h_2(x_0) \\ h_2(x_1) \\ \vdots \\ h_2(x_d) \end{pmatrix}_{d\times 1}$$

w and **X** are both unknown !

$$\boldsymbol{y} \approx \bar{\boldsymbol{x}} + w_1 \boldsymbol{x}_1 + w_2 \boldsymbol{x}_2 =: \boldsymbol{X}^\top \boldsymbol{w}$$

$$\arg\min_{\boldsymbol{X},\boldsymbol{w}} \left\| \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{w} \right\|$$

Solution:
Singular Value Decomposition



Let $\hat{Y}$ be the centered $d \times N$ data matrix (assume $N > d$).

$$\sum_i \mathbf{y}_i = \mathbf{0}$$

$$\mathbf{Y} = \begin{pmatrix} y_{1,1} & y_{2,1} & y_{3,1} & \cdots & y_{N,1} \\ y_{1,2} & y_{2,2} & y_{3,2} & \cdots & y_{N,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{1,d} & y_{2,d} & y_{3,d} & \cdots & y_{N,d} \end{pmatrix}_{d \times N} = \mathbf{USV}^\top$$

$\boldsymbol{y}_2$

Singular values

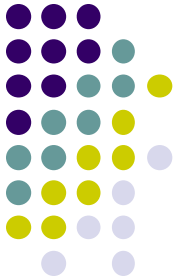
Unitary Matrices

is the SVD of $\hat{\mathbf{Y}}$, where

- ***U*** **is *d* × *d* orthogonal, the left singular vectors.**
- ***V*** **is *N* × *N* orthogonal, the right singular vectors.**
- ***S*** **is *d* × *N* diagonal, with $s_1 \geq s_2 \geq \ldots \geq s_d \geq 0$, the singular values.**

- ✓ **The SVD always exists, and is unique up to signs.**
- ✓ **The columns of V are the principal components**

Solution:
Singular Value Decomposition



Let $\hat{Y}$ be the centered $d \times N$ data matrix (assume $N > d$).

$$\sum_i \mathbf{y}_i = \mathbf{0}$$

$$\hat{\mathbf{Y}} = \begin{pmatrix} y_{1,1} & y_{2,1} & y_{3,1} & \cdots & y_{N,1} \\ y_{1,2} & y_{2,2} & y_{3,2} & \cdots & y_{N,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{1,d} & y_{2,d} & y_{3,d} & \cdots & y_{N,d} \end{pmatrix}_{d \times N} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$$

$\mathbf{y}_2$

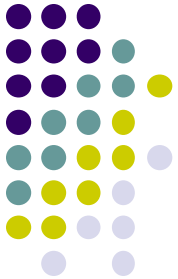
Singular values

Unitary Matrices

$$\arg\min_{\mathbf{X},\mathbf{W}} \left\| \mathbf{Y} - \mathbf{X}^\top \mathbf{W} \right\|$$

$$\mathbf{X} = \mathbf{U}^\mathbf{T}, \mathbf{W} = \mathbf{S}\mathbf{V}^\mathbf{T}$$

Principle component analysis



- Given data $\mathbf{Y}$
  - find transform $\mathbf{X}$ as well as feature $\mathbf{W}$

$$\arg\min_{\mathbf{X},\mathbf{W}} \left\| \mathbf{Y} - \mathbf{X}^{\mathrm{T}} \mathbf{W} \right\|_F$$

$$\mathbf{X} = \mathbf{U}^{\mathrm{T}}, \mathbf{W} = \mathbf{S}\mathbf{V}^{\mathrm{T}}$$

  - Given a new data $\mathbf{y}_{\mathrm{new}}$ we fix transform $\mathbf{X}$, then:

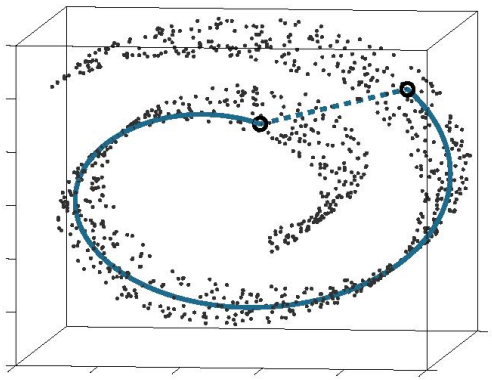
$$\mathbf{w}_{new} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 \cdot \mathbf{y}_{new} \\ \mathbf{u}_2 \cdot \mathbf{y}_{new} \\ \vdots \\ \mathbf{u}_p \cdot \mathbf{y}_{new} \end{pmatrix}$$

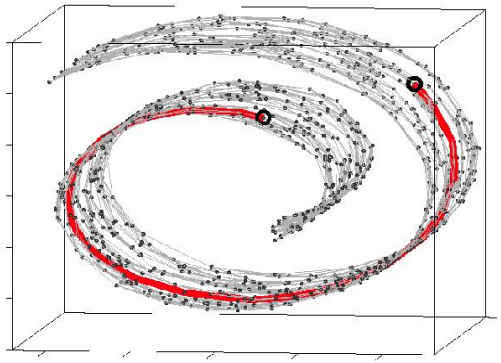
$\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ are principle components (rows of $\mathbf{X}$)

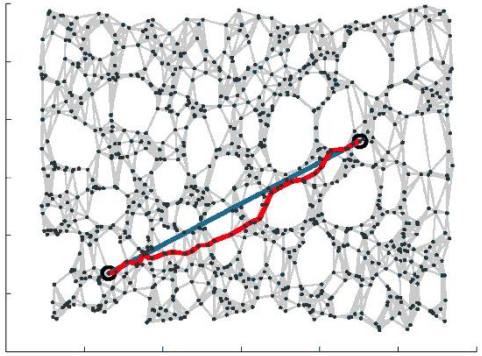
多维数据可视化 – 降维 – ISOMAP

- Preserve the **intrinsic geometry** of the data.
- Use the **geodesic distances** on manifold between all pairs.

Three steps algorithm

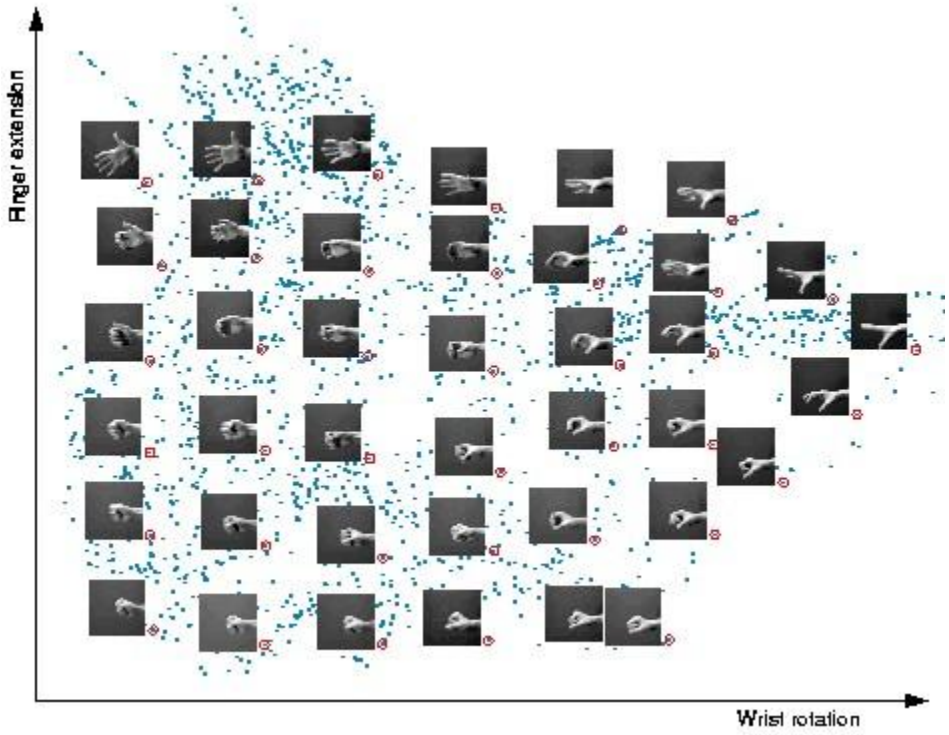






多维数据可视化 – 降维 – ISOMAP

- *N*=2000 images 64x64 pixels
- K=6

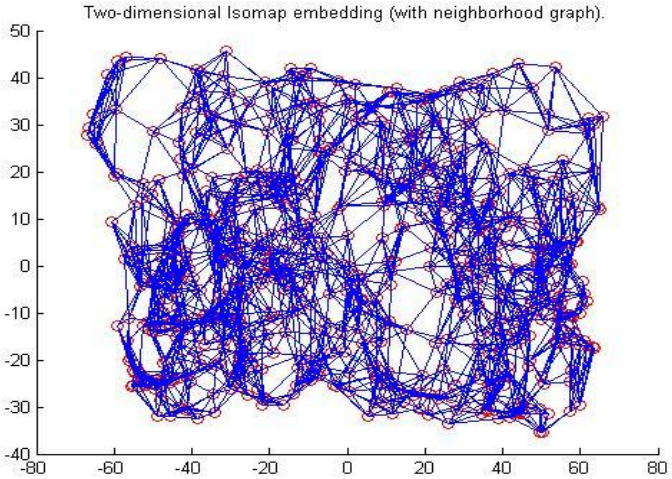


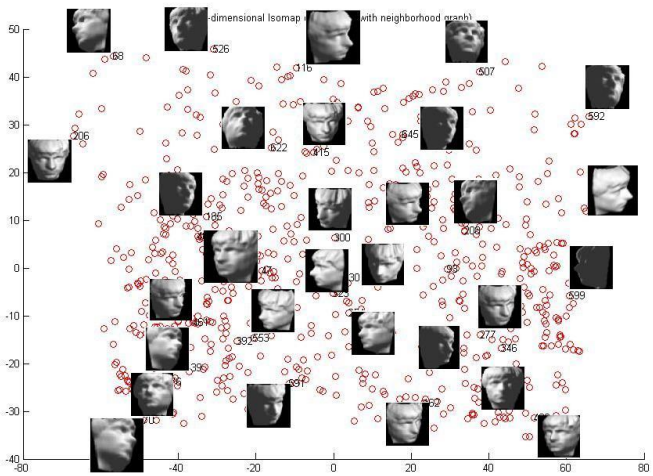
多维数据可视化 – 降维 – ISOMAP

Input: 698
images of 64x64

K=7, d=2

Outputs:





多维数据可视化 – 降维 – 更多方法

- https://cs.nyu.edu/~roweis/lle/publications.html

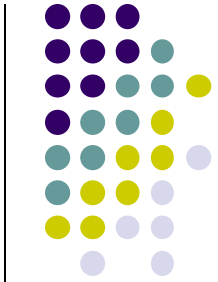
- Nonlinear dimensionality reduction by locally linear embedding. Sam Roweis & Lawrence Saul. Science, v.290 no.5500 , Dec.22, 2000. pp.2323—2326

- The Manifold Ways of Perception. (Cognition Perspectives in same issue) H. Sebastian Seung & Daniel D. Lee. Science, v.290 no.5500 , Dec.22, 2000. pp.2268—2269

- An Introduction to Locally Linear Embedding. Lawrence Saul & Sam Roweis

t-SNE

t-Distributed Stochastic Neighbor Embedding



t-SNE用于手写数字图像的2维投射

SNE

Stochastic Neighbor Embedding

$$p_{j|i} = \frac{exp(\frac{-|x_i - x_j|^2}{2\sigma_i^2})}{\sum_{k \neq i} exp(\frac{-|x_i - x_k|^2}{2\sigma_i^2})} \qquad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$$

$$q_{j|i} = \frac{e^{-|y_i - y_j|^2}}{\sum_{k \neq i} e^{-|y_i - y_k|^2}} \qquad q_{ij} = \frac{q_{i|j} + q_{j|i}}{2N}$$

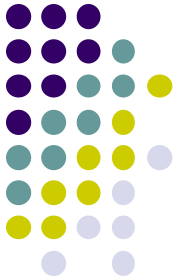
$$min\ KL(P||Q) = \sum_{i,j} p_{ij} log \frac{p_{ij}}{q_{ij}}$$

$$\frac{\partial C}{\partial y_i} = 2 \sum_j (p_{ij} - q_{ij})(y_i - y_j)$$

$$Y^t = Y^{t-1} + \eta \frac{\partial C}{\partial y_i} + \alpha(t)(Y^{t-1} - Y^{t-2})$$

t-SNE

t-Distributed Stochastic Neighbor Embedding



$$p_{j|i} = \frac{exp(\frac{-|x_i - x_j|^2}{2\sigma_i^2})}{\sum_{k \neq i} exp(\frac{-|x_i - x_k|^2}{2\sigma_i^2})}$$

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$$

$$q_{ij} = \frac{(1 + |y_i - y_j|)^{-1}}{\sum_{k \neq l}(1 + |y_k - y_l|)^{-1}}$$

$$q_{ij} = \frac{q_{i|j} + q_{j|i}}{2N}$$

$$min\ KL(P||Q) = \sum_{i,j} p_{ij} log \frac{p_{ij}}{q_{ij}}$$

$$q_{ij} = \frac{f(|y_i - y_j|^2)}{\sum_{k \neq l} f(|y_k - y_l|^2)} \qquad f(x) = \frac{1}{1 + x^2}$$

t-Distribution

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + |y_i - y_j|^2)^{-1}$$

$$Y^t = Y^{t-1} + \eta \frac{\partial C}{\partial y_i} + \alpha(t)(Y^{t-1} - Y^{t-2})$$

t-SNE

t-Distributed Stochastic Neighbor Embedding

t-SNE visualization of CNN codes
with 50,000 ILSVRC 2012 validation images

多维数据可视化 – 非结构化数据

usr_id	用户 id
install_ids	安装的应用程序 id 列表
date	日期
latitude	经度坐标
longitude	纬度坐标

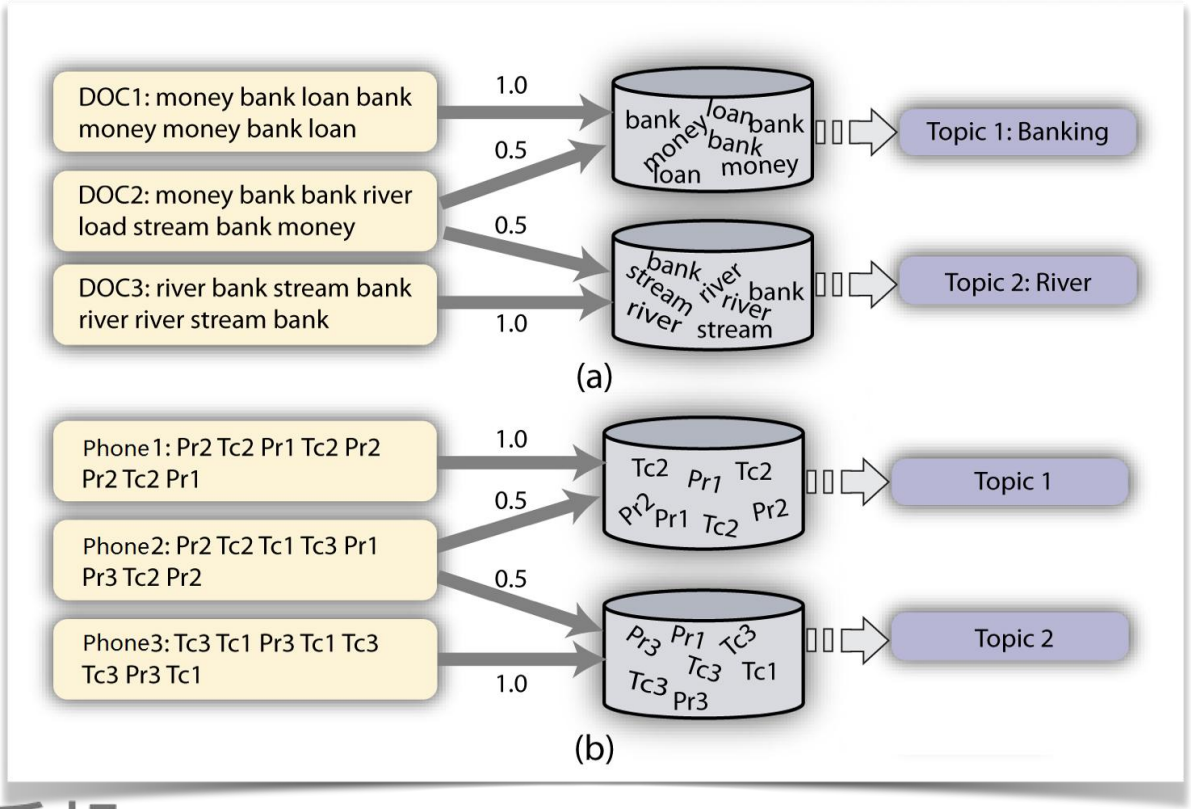
app_id	应用程序 id
category_names	分类标签名称列表

手机日志数据分析

难点：每个人手机里安装的应用个数不同，难以统一处理

多维数据可视化 – 非结构化数据

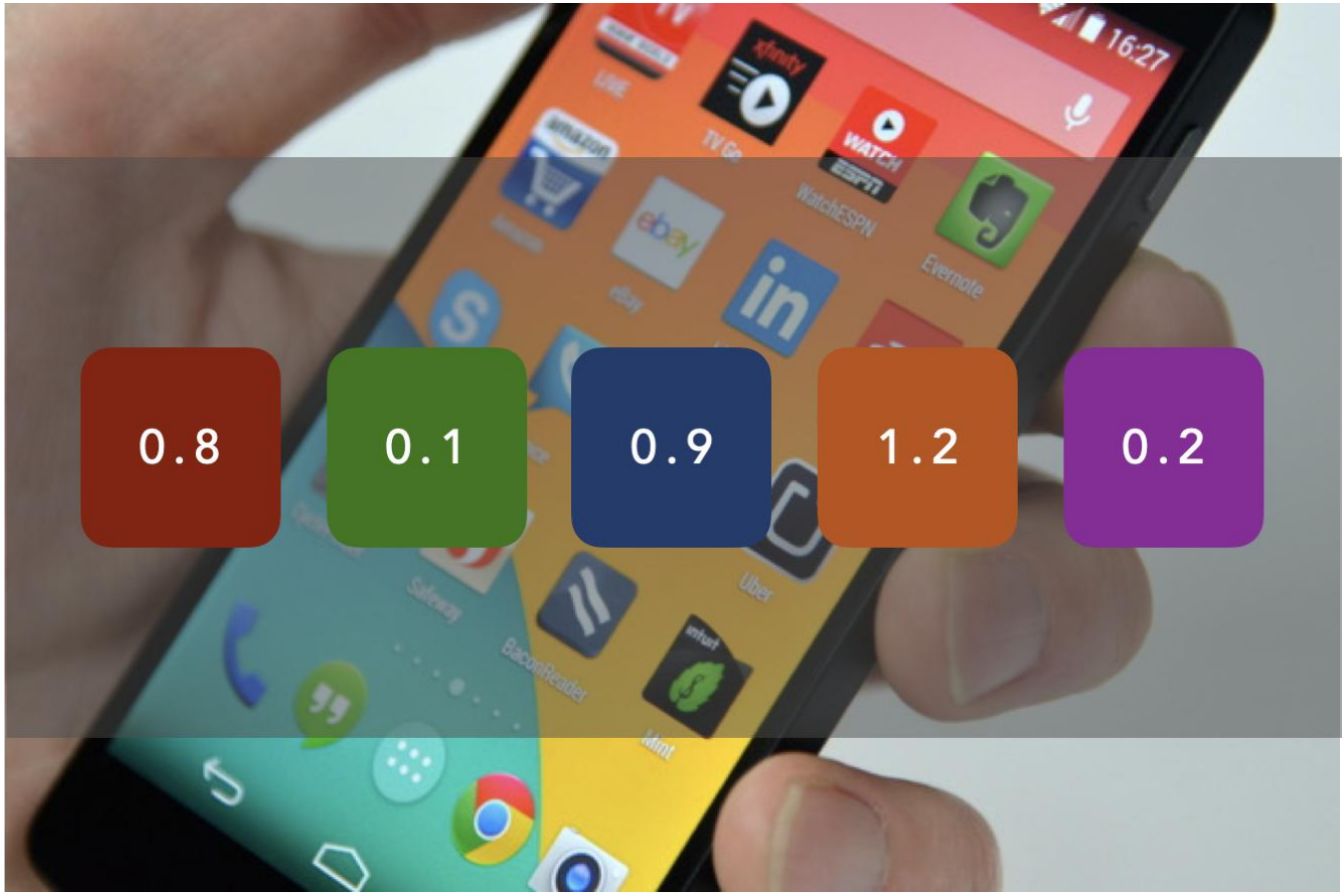
文章 单词 主题



手机用户 应用 特征

LDA主题模型

多维数据可视化 – 非结构化数据



手机日志数据分析

解决方案：通过LDA等方法进行统一的数据量化

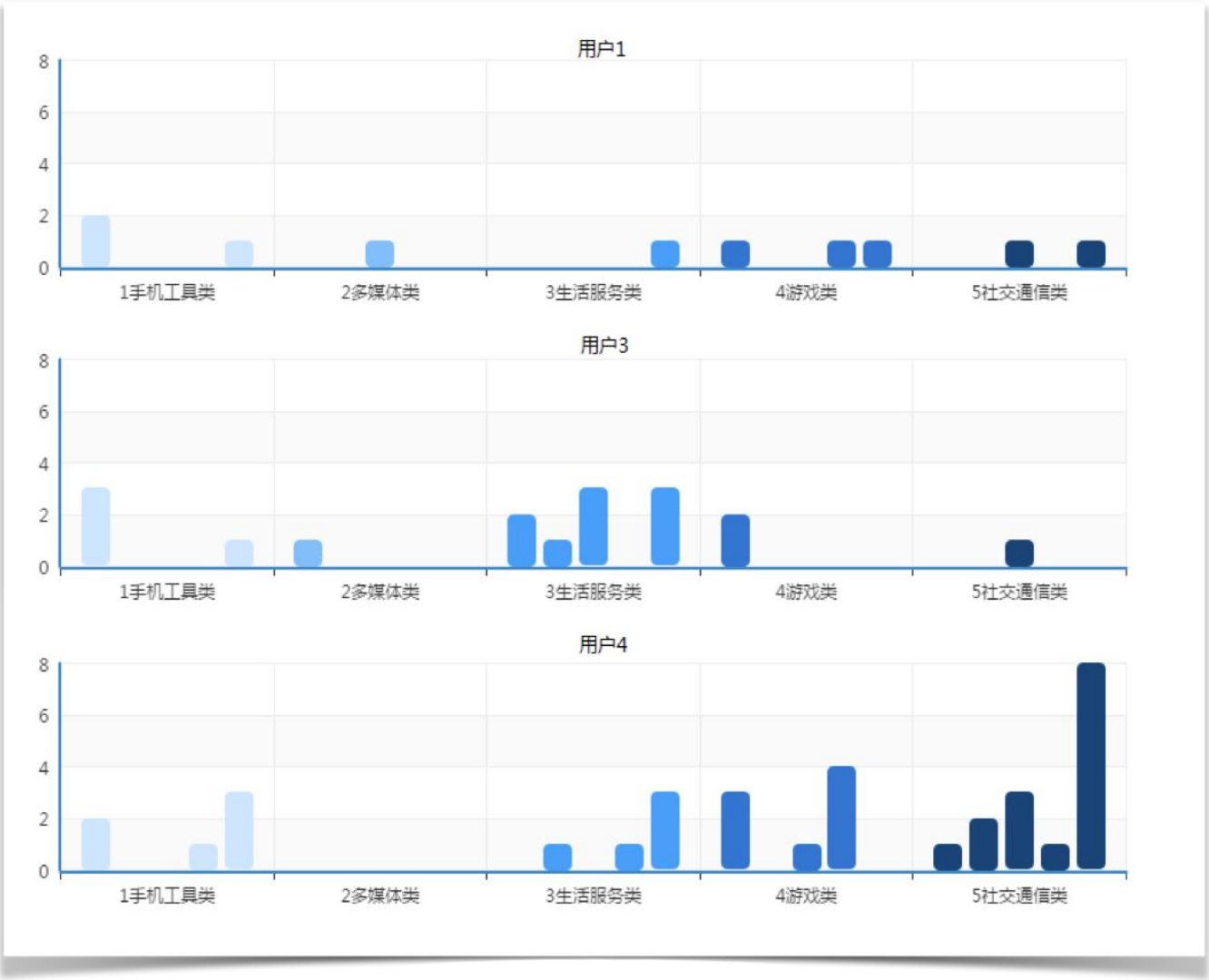
多维数据可视化 – 非结构化数据

采集了21295名手机用户的数据

主题 1	主题 2	主题 3	主题 4	主题 5
手机工具	多媒体	生活服务	游戏	社交通信
系统工具	音乐视频	综合服务	创意休闲	汉化软件
社交通信	视频播放	购物理财	益智棋牌	即时通讯
通信聊天	音乐音频	购物支付	体育竞速	通信聊天
即时通讯	摄影美化	学习阅读	动作射击	社交网络

计算后的主题构成

多维数据可视化 – 非结构化数据



用户1

没有明显倾向性

用户3

比较注重生活

用户4

社交达人

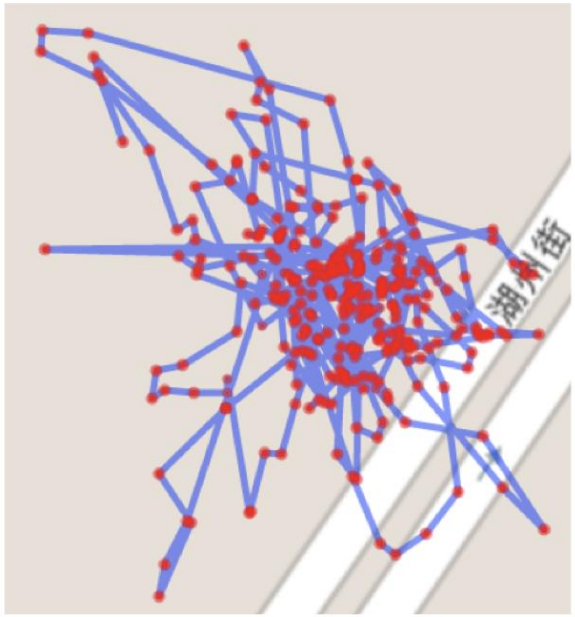
多维数据可视化 – 非结构化数据

杭州市

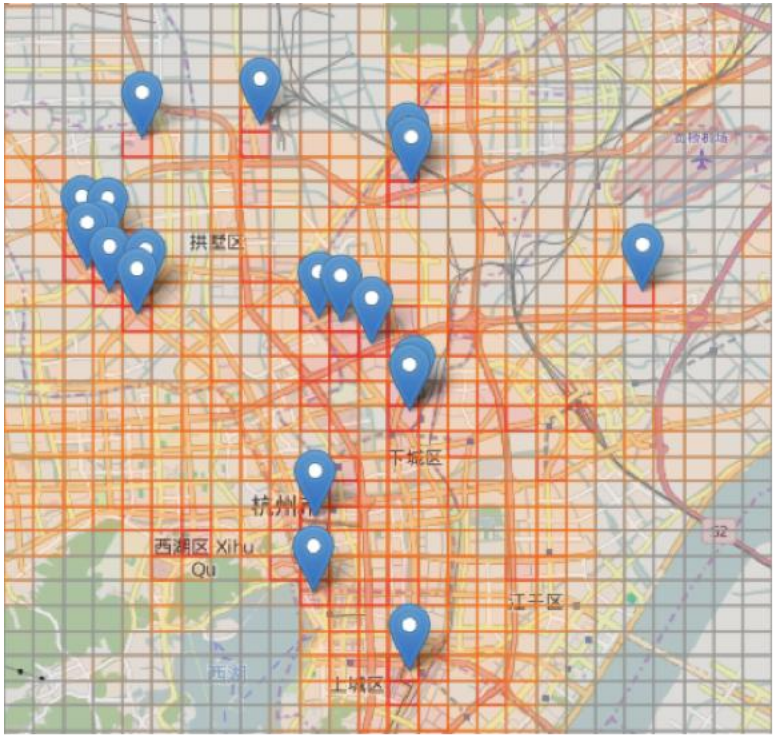
每天10GB数据

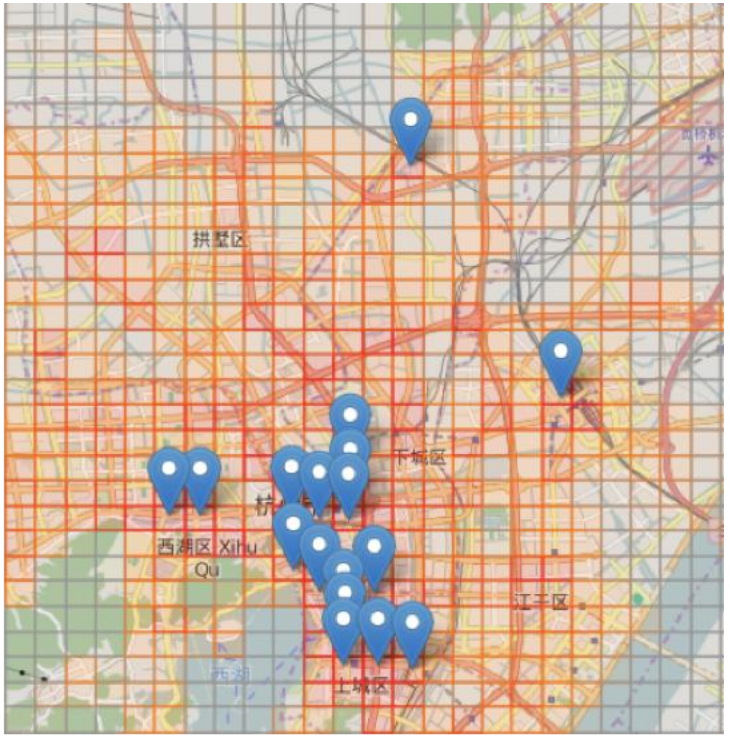
1个月为300GB

车辆10万+



多维数据可视化 – 非结构化数据



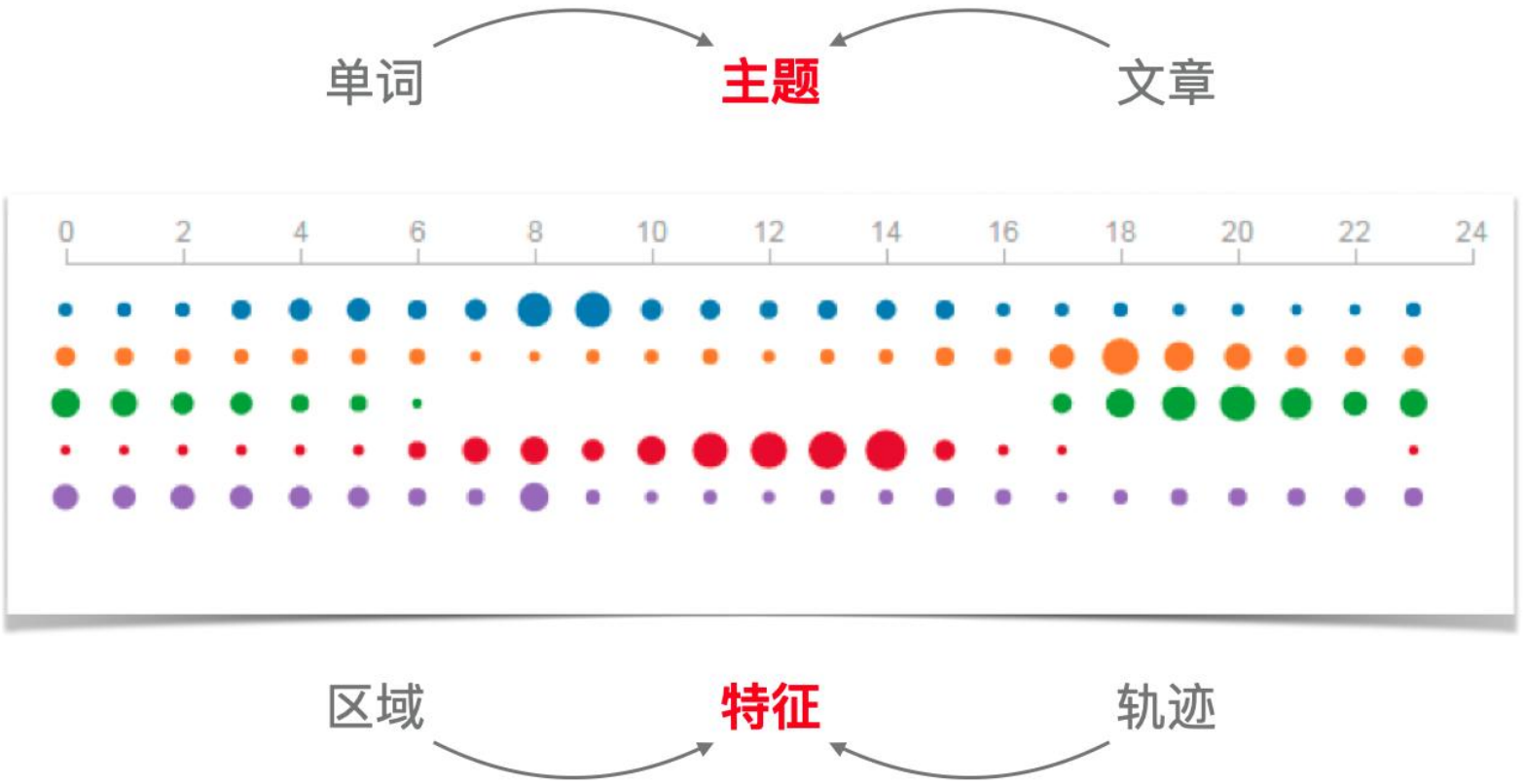


凌晨 2:00　　每天10GB数据　　下午 6:00

对轨迹数据进行时空网格量化

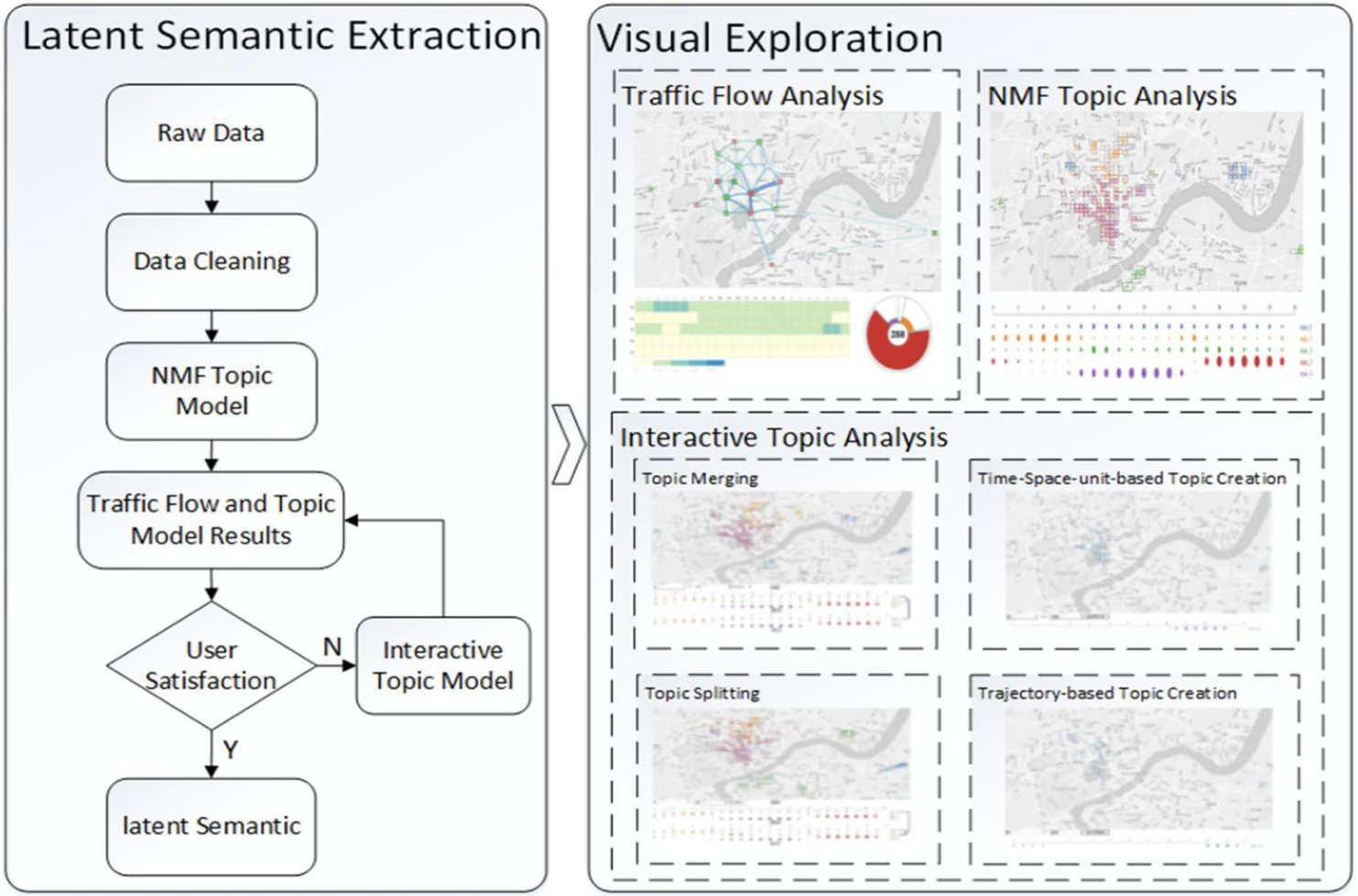
多维数据可视化 – 非结构化数据

智慧城市：轨迹数据分析：主题量化（CHINAVIS2017）



*LDA*主题模型

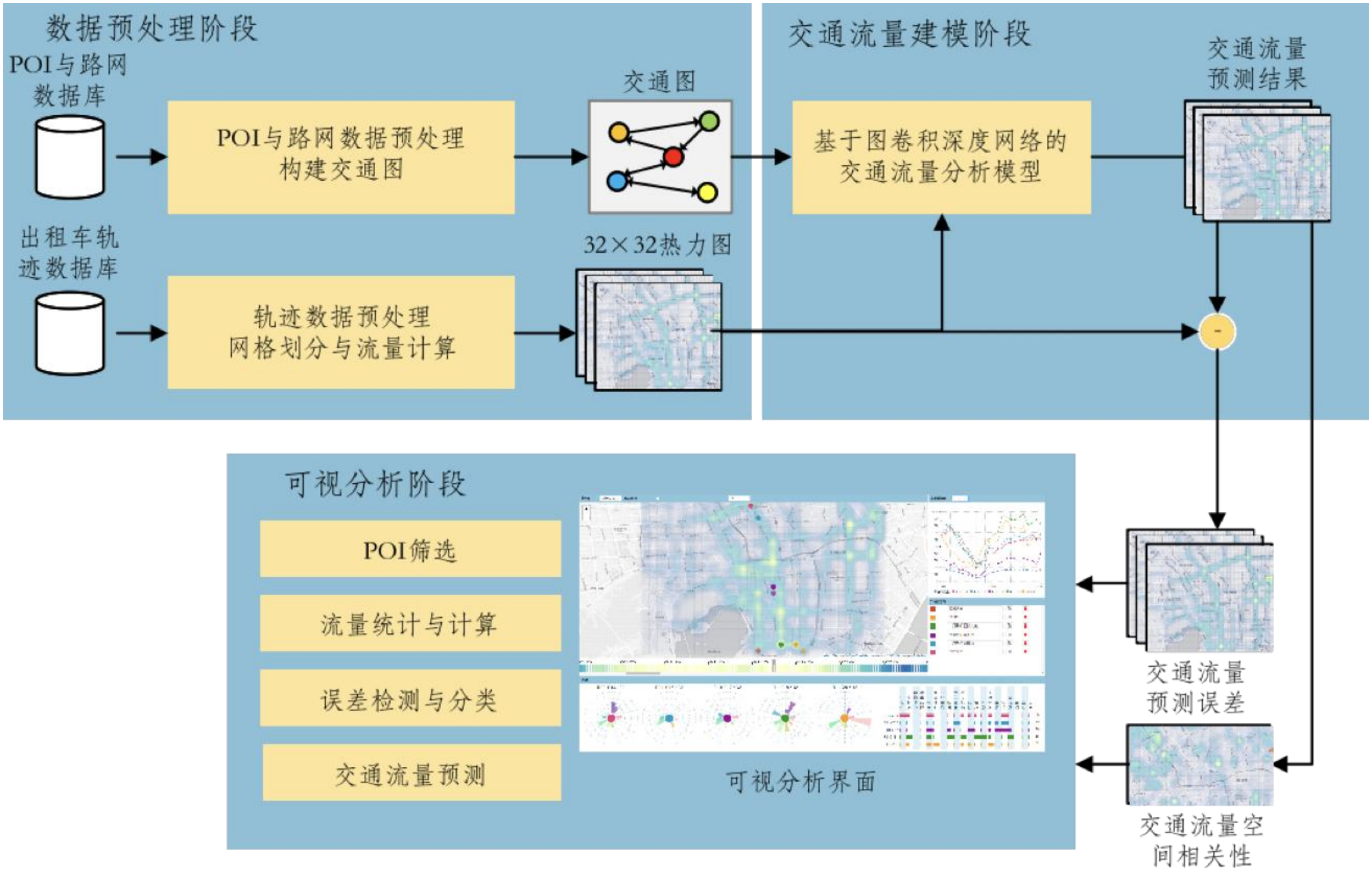
多维数据可视化 – 非结构化数据 – 组合



系统整体架构设计

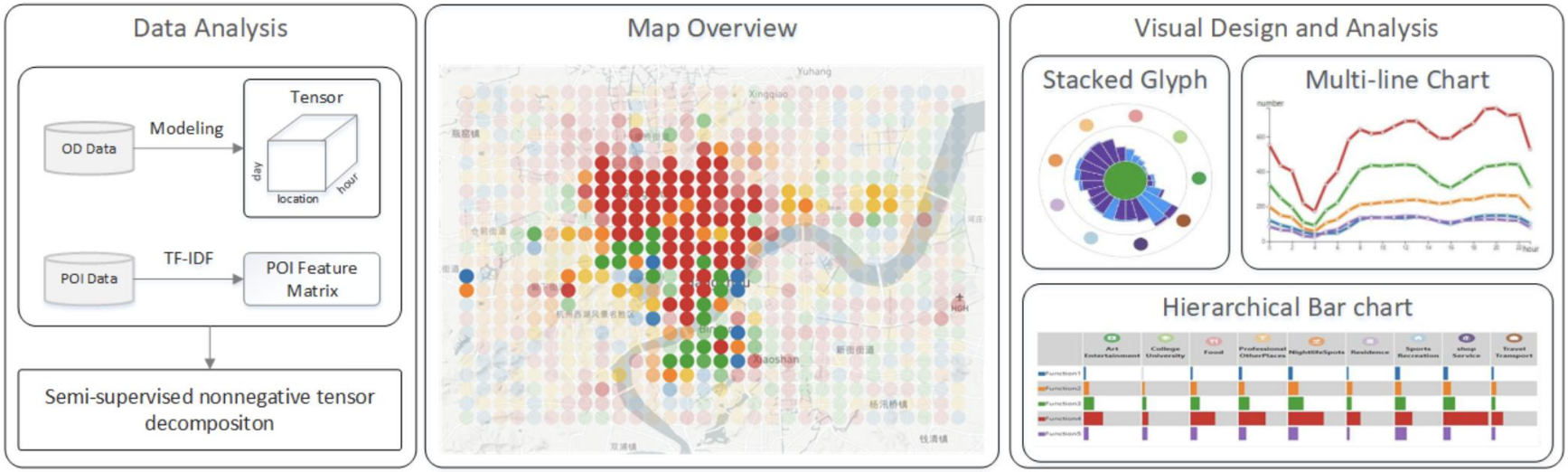
（JOV 2020）

多维数据可视化 – 非结构化数据 – DL



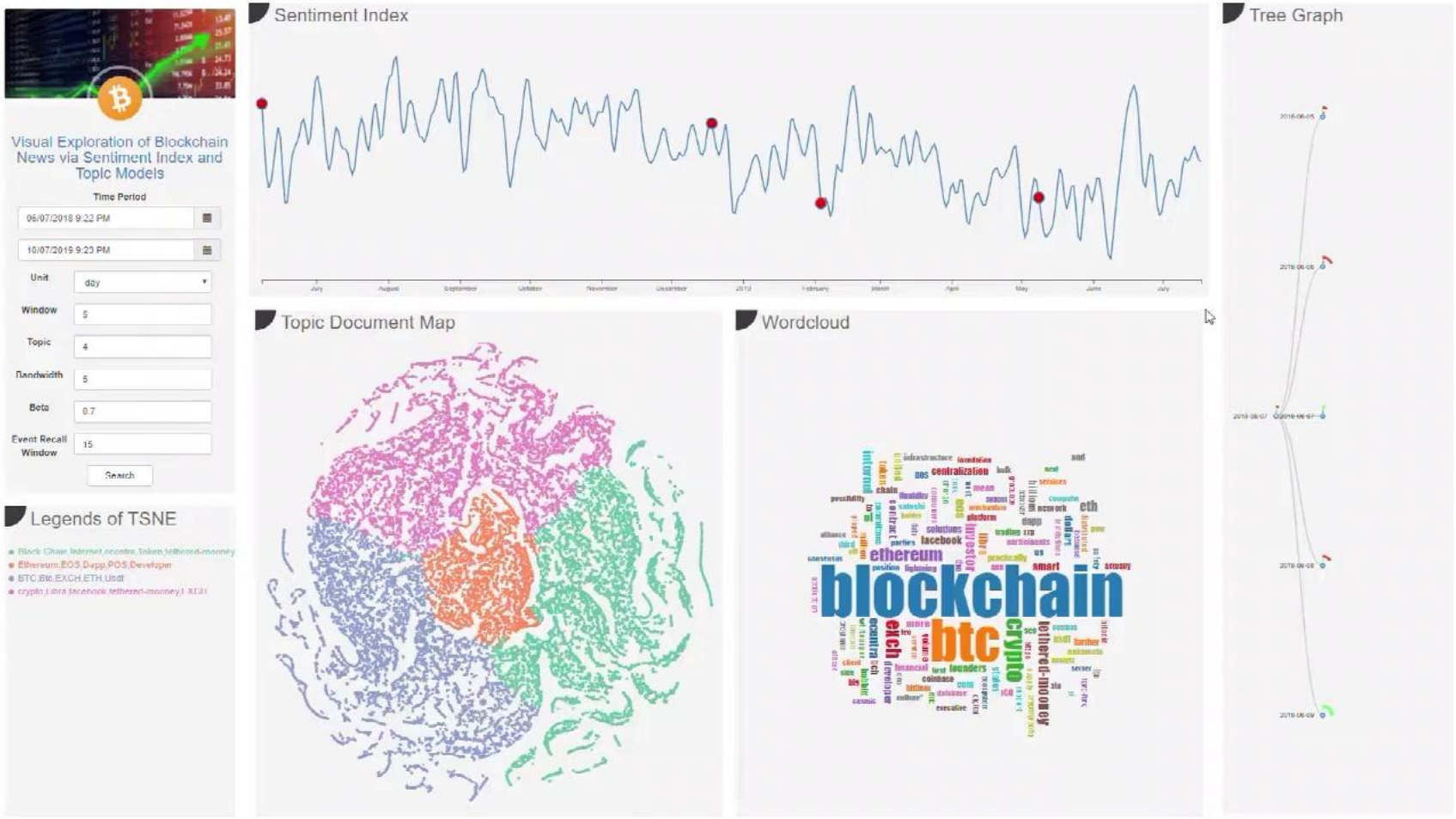
智慧城市：交通流量精准预测与分析：图卷积方法（2020）

多维数据可视化 – 非结构化数据 – 组合



Visual analysis of traffic data via spatio-temporal graphs and interactive topic modeling (ChinaVis 2018, JOV)

多维数据可视化 – 非结构化数据 – 组合



案例：数字货币新闻可视化

Visual exploration of Internet news via sentiment score and topic models (CVMJ 2021)

多维数据可视化 – 非结构化数据 – 组合

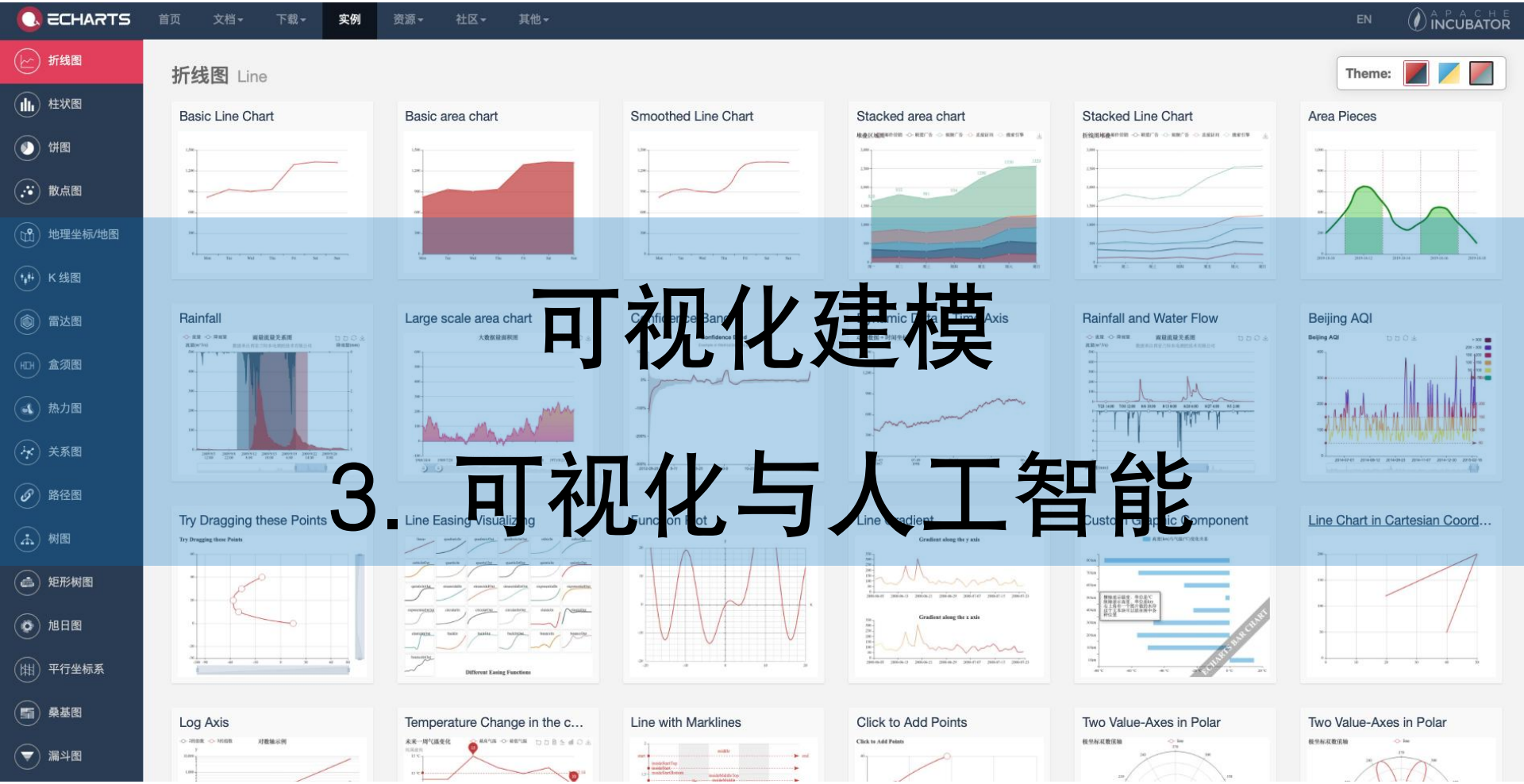
案例：在线啦啦操教学数据可视化

DanceVis: toward better understanding of online cheer and dance training (ChinaVis 2021, JOV)

多维数据可视化 – 组合 – 数据大屏

参考文献

- Visual analysis of traffic data via spatio-temporal graphs and interactive topic modeling (ChinaVis 2018, JOV)
- Visual analysis of traffic data based on topic modeling (ChinaVis 2017, JOV)
- Visual analysis of cloud computing performance using behavioral lines (IEEE TVCG 2016)
- BN-Mapping:基于贝叶斯网络的地理空间数据可视分析 (计算机学报 2016)
- 基于移动终端日志数据的人群特征可视化 (软件学报 2016)
- Visual exploration of Internet news via sentiment score and topic models (CVMJ 2021)
- DanceVis: toward better understanding of online cheer and dance training (ChinaVis 2021, JOV)

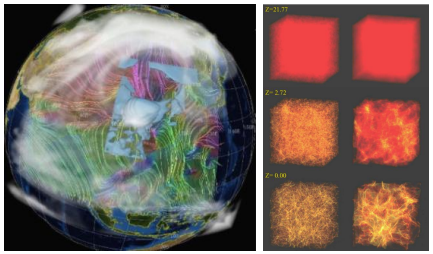


可视化建模

3. 可视化与人工智能

大数据可视化的重要应用领域

大科学



大安全

大工程

物联网与智慧城市

互联网与社交媒体

从智能角度对人工智能的分类

- **领域人工智能**
- 依葫芦画瓢、任务导向，如Deep Blue和AlphaGo

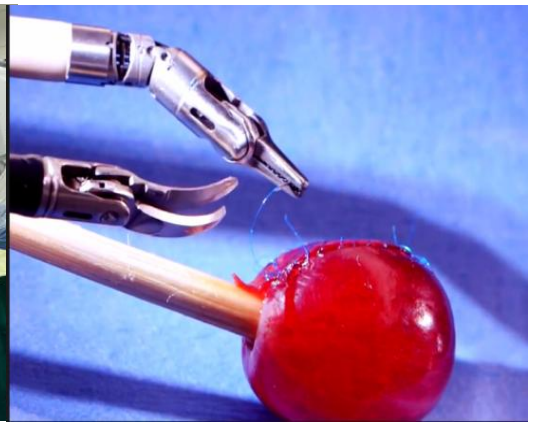
- **通用人工智能或跨领域人工智能**
- 举一反三、从经验中学习，如“人类”智能

Traditional AI	Artificial General Intelligence
Focus on having knowledge and skills	Focus on acquiring knowledge and skills
Action acquiring via programing	Ability acquiring via learning
domain-specific ability via rule-based and exemplar-based	general ability via abstraction (intuition) and context (common sense)
Learning by data and rules	**Learning to learn**

从智能角度对人工智能的分类

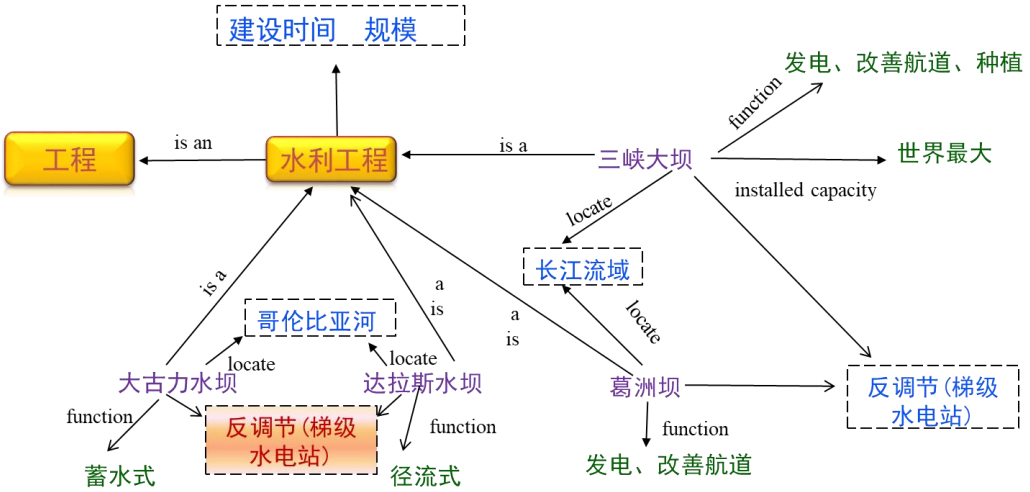
- **混合增强人工智能**
  - 多种智能体的混合形式
    - 人类智能+机器智能：如达芬奇手术机器人
    - 人、机、物、网互联：如智慧城市系统

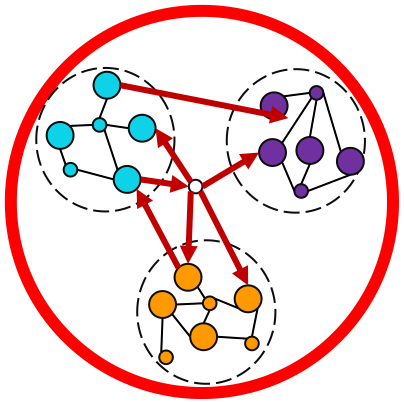
智慧城市与智能服务



Da Vinci手术机器人及其缝制葡萄皮

AI历史上的主流方法(1)：用规则教





符号逻辑表示下的推理

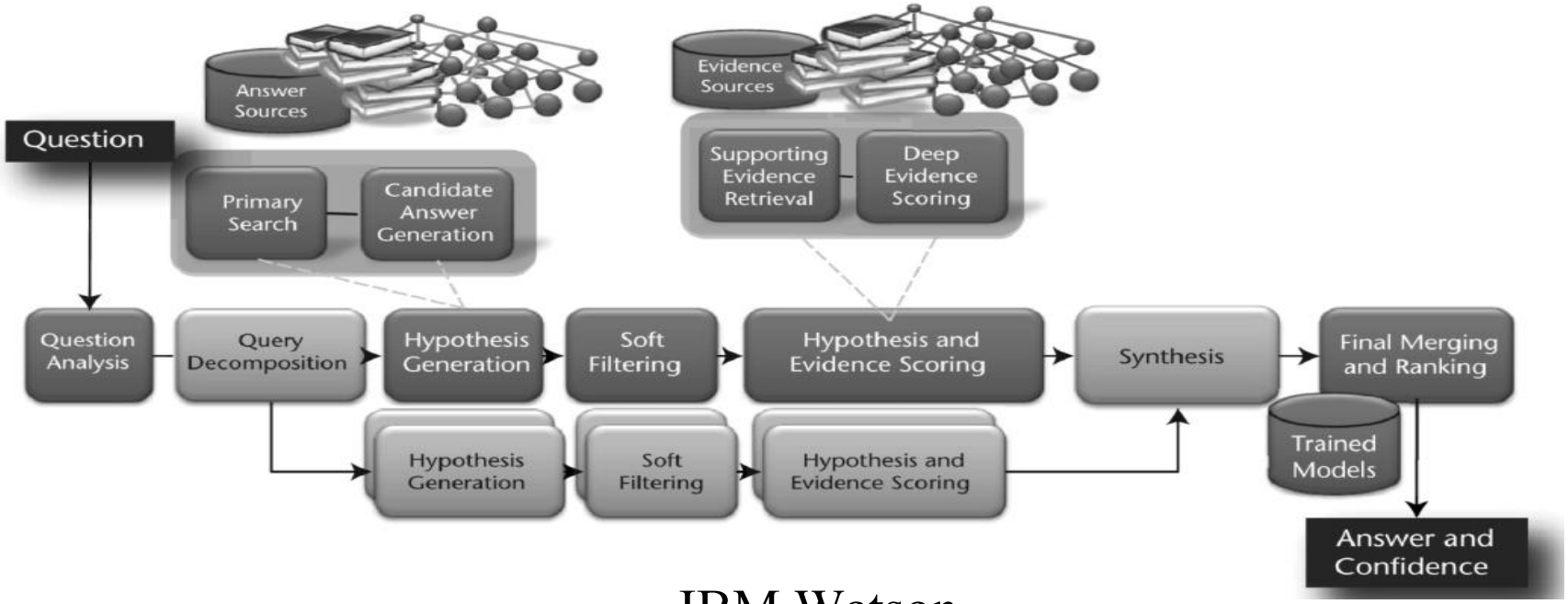
知识图谱

用规则教

符号主义人工智能
(Symbolic AI)

AI历史上的主流方法(1)：用规则教

1997年IBM“深蓝”和2011年IBM“沃森”是以推理为核心的人工智能的代表。



IBM Watson

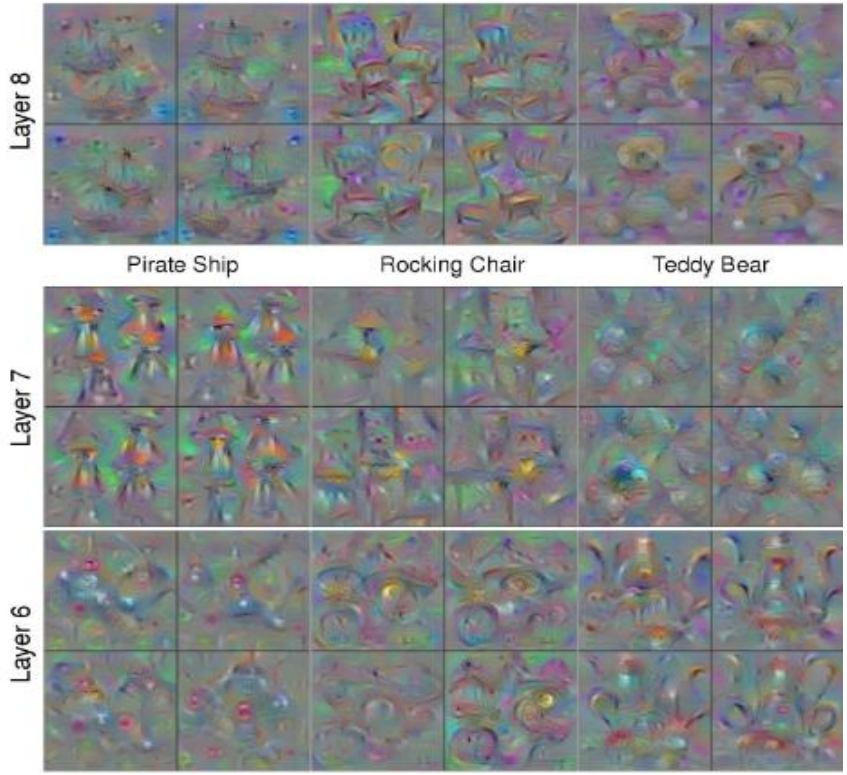
David Ferrucci, et.al., Building Watson: An Overview of the DeepQA Project, AAAI 2010

AI历史上的主流方法(2)：用数据学

- 数据驱动的人工智能（data-driven AI）

挖掘数据所蕴含的内在模式

用大数据学（有监督）

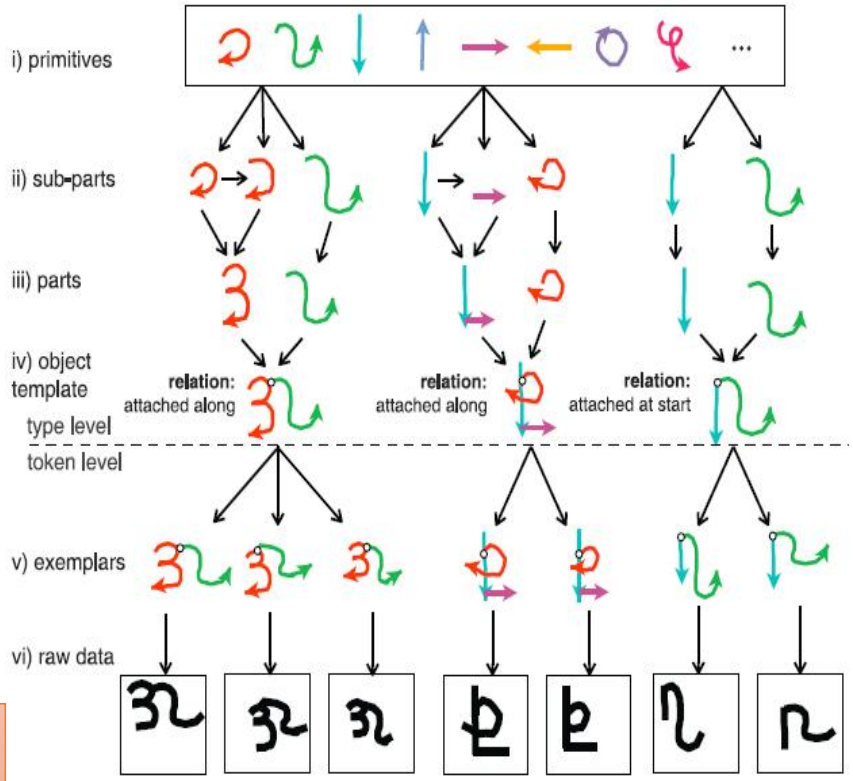


Deep Visualization
(ICML DL Workshop 2015)

AI历史上的主流方法(2)：用数据学

稀缺数据

贝叶斯概率学习



Probabilistic Program Induction, *Science*, 2015

解析数据所蕴含的内在模式

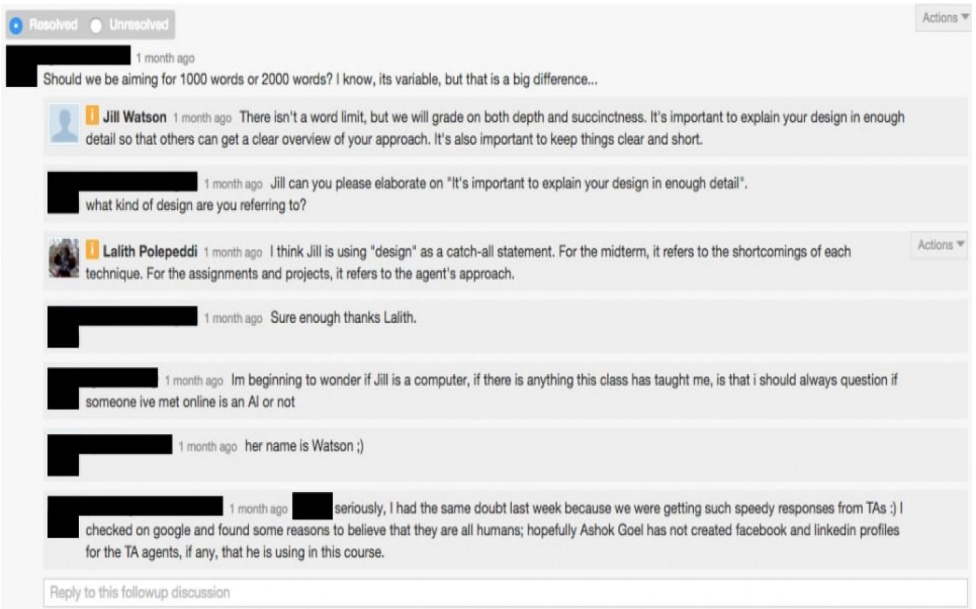
辅以组合性、因果性等先验知识

用小数据学(无监督/半监督)

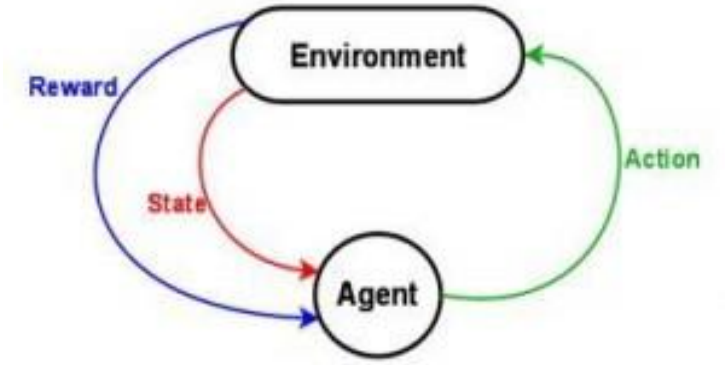
AI历史上的主流方法(2)：用数据学

- 美国佐治亚理工的人工智能助教：Jill Watson
  - 在线课程“Knowledge Based Artificial Intelligence (KBAI)”，400个学生，一个学期约有1万个问题（主讲：Prof. Ashok K. Goel）
  - 基于IBM’s Watson平台，基于2014年以来的4万多个问题-答案进行训练
  - Jill Watson于2016年春季上岗，三个月下来，未被学生发现

“Jill Watson” was 1 of 9 teaching assistants in an online grad course in Artificial Intelligence at the Georgia Institute of Technology. She performed admirably, and nobody suspected she wasn’t a human. The only hint might have been that she responded perhaps a little too promptly to student questions – and she single-handedly answered 40% of them



AI历史上的主流方法(3): 从经验中学



从经验中的策略学习

用问题引导（反馈牵引）

Science | DOI:10.1145/2949662 Marina Krakovsky

Reinforcement Renaissance

The power of deep neural networks has sparked renewed interest in reinforcement learning, with applications to games, robotics, and beyond.

EARLY THIS TIME DEEPMIND has announced an amazing accomplishment in game-playing computers in recent months, people have taken notice.

First, the Google-owned, London-based artificial intelligence (AI) research center wowed the world with a computer program that had taught itself to play nearly 50 different 1980s-era Atari games—from Pong and Breakout to Pac-Man, Space Invaders, Boxing, and more—using as input nothing but pixel positions and game scores, performing at or above the human level in more than half these varied games. Then, this January, DeepMind researchers impressed experts with a feat in the realm of strategy games: AlphaGo, their Go-playing program, beat the European champion in the ancient board game, which poses a much tougher AI challenge than chess. Less than two months later, AlphaGo scored an even greater victory: it won 4 games in a best-of-5 series against the best Go player in the world, surprising the champion himself.

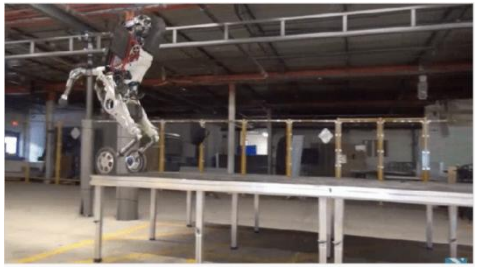
The idea that a computer can learn to play such complex games from scratch and achieve a proficient level elicits gee-whiz reactions from the general public, and DeepMind's triumphs have heightened academic and commercial interest in the AI field behind DeepMind's methods: a blend of deep neural networks and reinforcement learning called "deep reinforcement learning."

"Reinforcement learning is a model of learning where you're not given a solution—you have to discover it by trial and error," explains Sridhar Mahadevan, a professor at the University of Massachusetts Amherst, a long-time center of research into reinforcement learning.

12 COMMUNICATIONS OF THE ACM | AUGUST 2016 | VOL. 59 | NO. 8

Reinforcement Renaissance, ***Communications of the ACM***, 2016,59(8):12-14

AI历史上的主流方法(3): 从经验中学



Boston Dynamics'
Handle robot

67 fixed-wing UAV swarm
prototype in Zhuhai

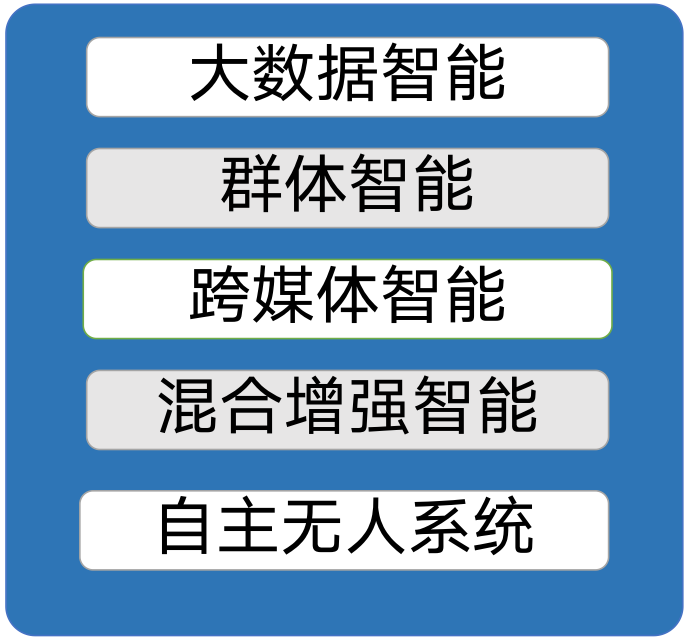
Brooks, R. A. (1986), A robust layered control system for a mobile robot, IEEE Journal of Robotics and Automation, RA-2:14-23

AI历史上的主流方法小结

学习模式	优势	不足
用规则教	与人类逻辑推理相似，解释性强	难以构建完备的知识规则库
用数据学	直接从数据中学	以深度学习为例：依赖于数据、解释性不强
用问题引导	从经验中进行能力的持续学习	非穷举式搜索而需更好策略

从数据到知识与能力，能力增强是最终目标
三种学习方法的综合利用值得关注！

新一代人工智能中五个智能方向



新一代人工智能中五大智能方向

- 推动人工智能发生如下跃变：
  - 从人工知识表达技术到大数据驱动知识学习
  - 从处理类型单一的数据到跨媒体认知、学习和推理
  - 从追求“机器智能”到迈向人机混合的增强智能
  - 从聚焦研究“个体智能”到基于互联网络的群体智能
  - 从机器人到自主无人系统的跨越

大数据智能

- 针对“数据驱动与知识引导相结合”以及“直觉感知可视交互”等特点，在计算框架、知识计算、博弈对决和交互可视等方面进行技术研发和平台研制。

- 形成从数据到知识、从知识到智慧的能力，打穿数据孤岛，形成链接多领域的知识中心，支撑新技术和新业态的跨界融合与创新服务。

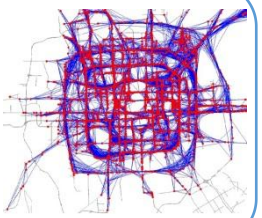
- 在创新设计、智慧医疗、智能经济和社会治理等方面进行示范。

可视化与知识工程：大数据智能

信息空间

人类社会

物理空间



三元空间

对话交互

可视化交互

增强现实

基础架构和平台

整合数据驱动、符号逻辑和概率等模型

知识计算引擎

支持直觉感知和深度推理的模型与方法

在线博弈与智能模拟

开放环境中博弈与神经元交互模型

决策与服务

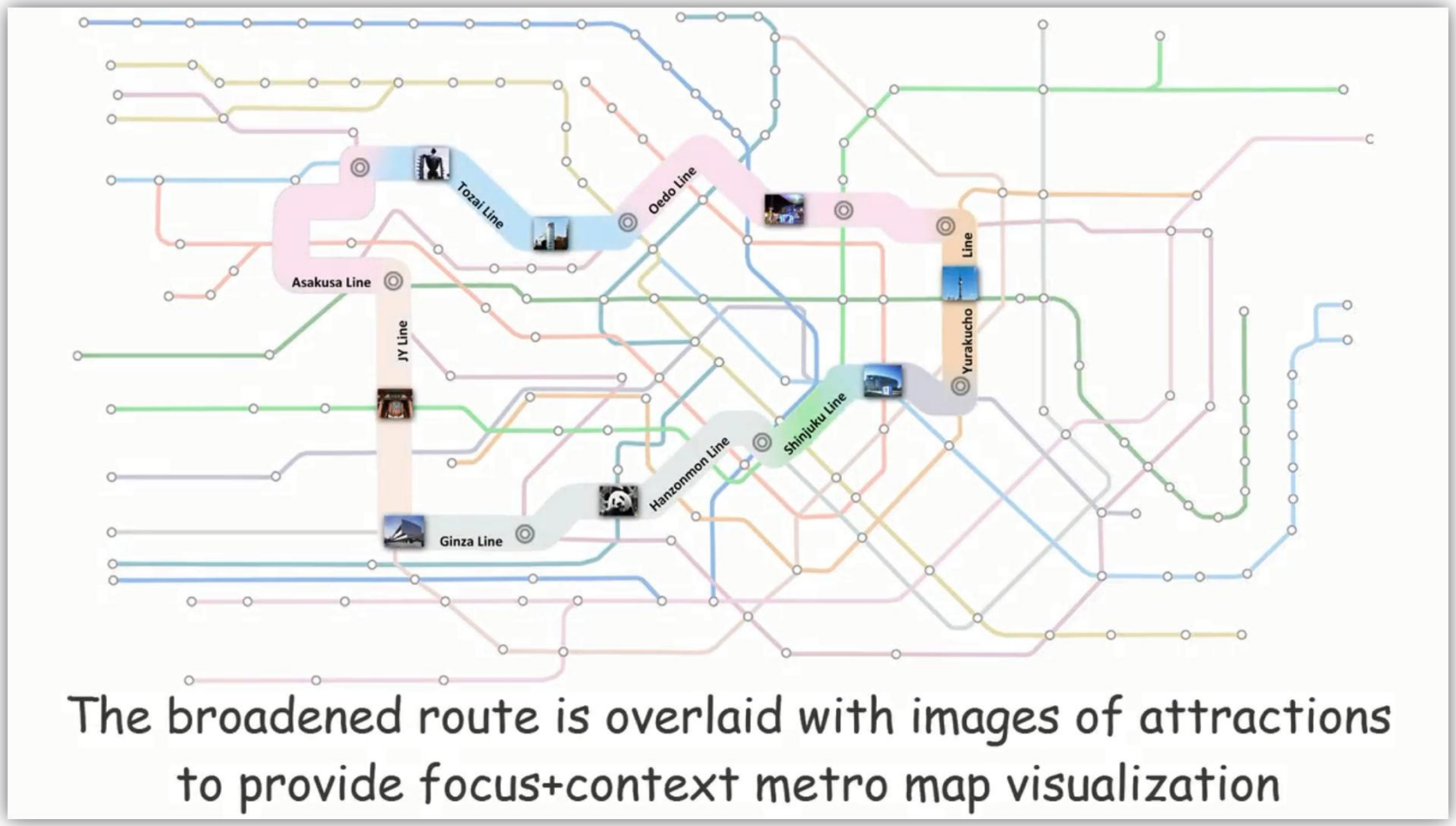
创新设计与经济机器人

单元小测试

- 开始：9:05
- 截止：9:50
- 持续：45分钟
- 尝试：1次

- 题目：20道选择题，考察概念为主
- 分值：100分

测试开始


Tozai Line
Oedo Line
Asakusa Line
JY Line
Line
Yurakucho
Shinjuku Line
Hanzonmon Line
Ginza Line
The broadened route is overlaid with images of attractions
to provide focus+context metro map visualization

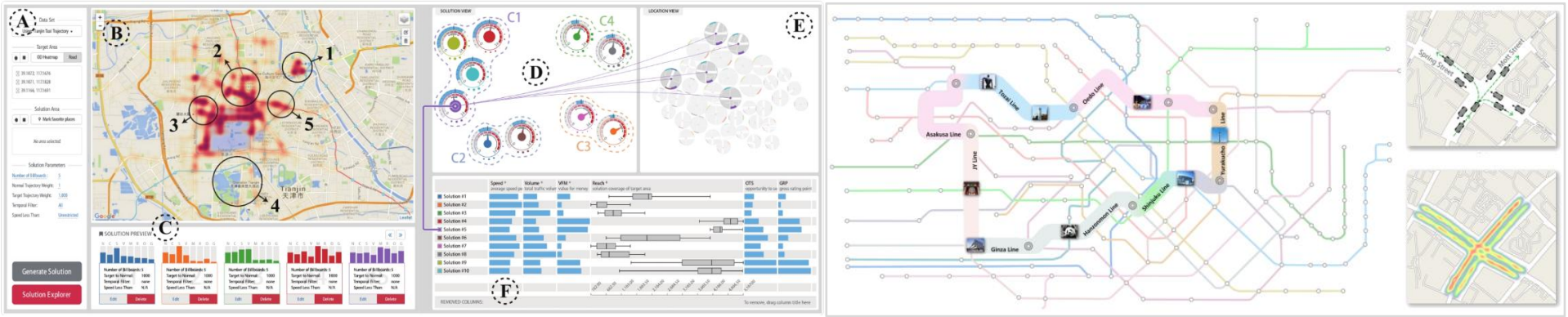
主要参考资料
《大数据可视化建模》

陈为

浙江大学计算机学院

chenwei@cad.zju.edu.cn

http://www.cad.zju.edu.cn/home/chenwei



谢谢

Thank You

微博：@浙大张宏鑫

邮件：zhx@cad.zju.edu.cn

主页：http://person.zju.edu.cn/zhx

手机：13958011790

微信：timothykull

开源：https://github.com/hongxin/vizmodeling