



电子信息工程中的数学模型和方法

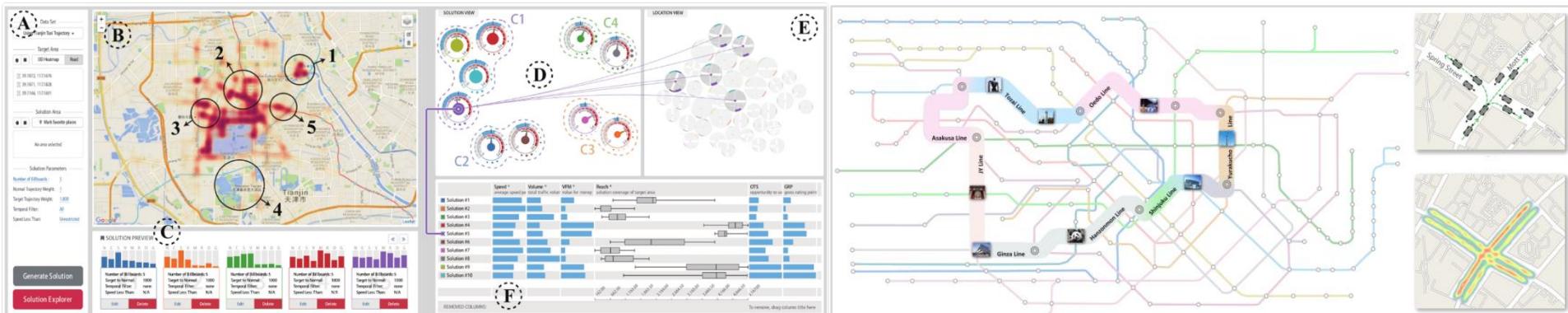
可视化建模（上）

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>



数据可视化与可视分析

什么是数据可视化？为什么我们要用数据可视化？

<https://github.com/hongxin/vizmodeling>

引子



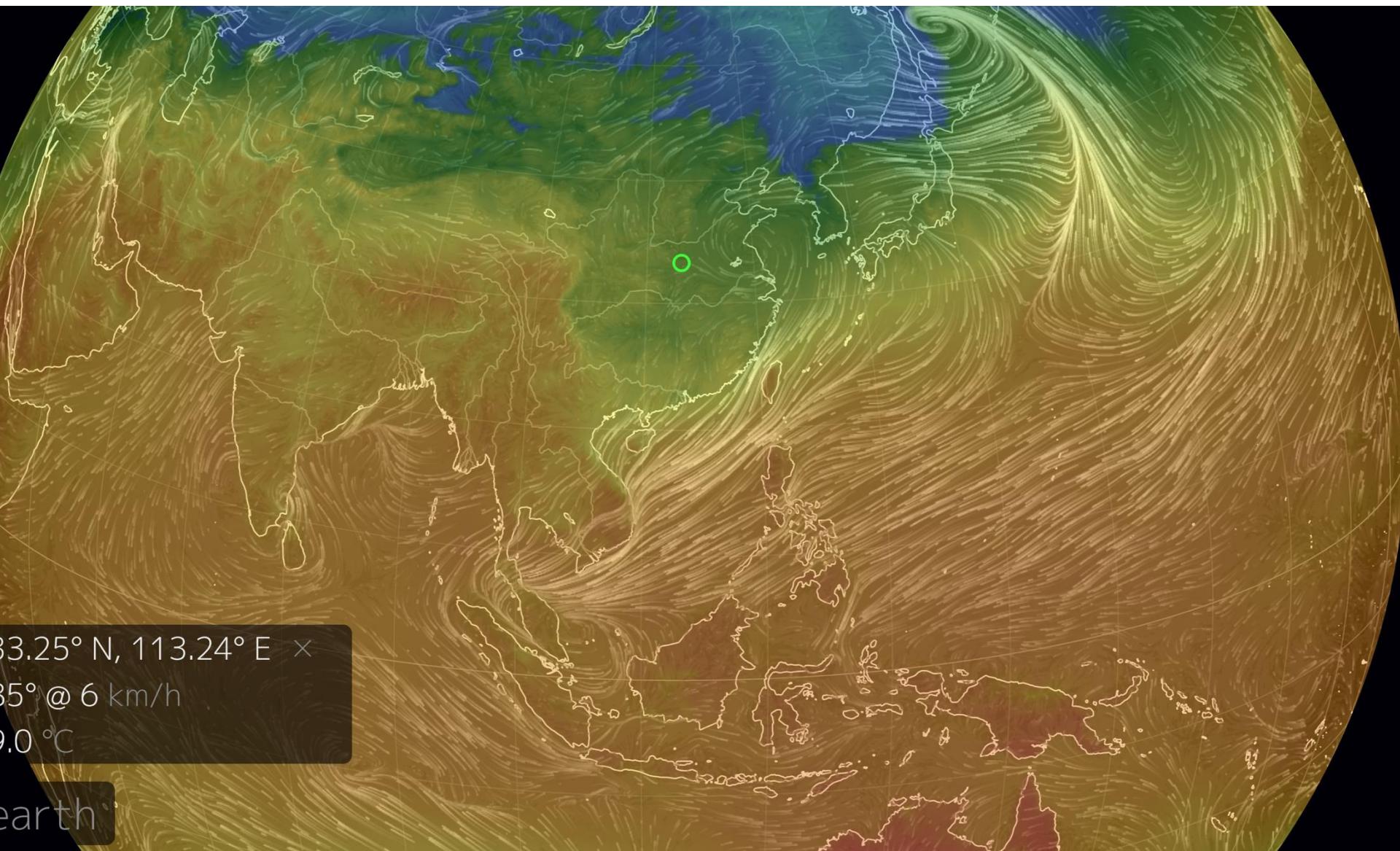
2017年双11的区域经济闪电图，将实时订单数据与物流干线结合展示

例子 - 飞行模式可视化



例子 - 全球气象可视化

<https://earth.nullschool.net>





例子 - 《本草纲目》可视化

药症方关联的中医药古籍 交互可视分析方法

吴泓嘉, 张弛, 张宏鑫, 陈为, 夏佳志

ChinaVis 2023

ID 1239

本文提出了一种药症方关联的
中医药古籍交互可视分析方法

吴泓嘉, 张弛, 张宏鑫, 陈为, 夏佳志. 药症方关联的中医药古籍交互可视分析方法[J].
计算机辅助设计与图形学学报. DOI: 10.3724/SP.J.1089.2023-00623

什么是数据可视化？

数据可视化

■ 创建并研究数据的视觉表达 (Visual Representation)

- 输入: 数据 (data)
- 输出: 视觉形式 (visual form)
- 目标: 深入理解 (insight)



数据

视觉形式

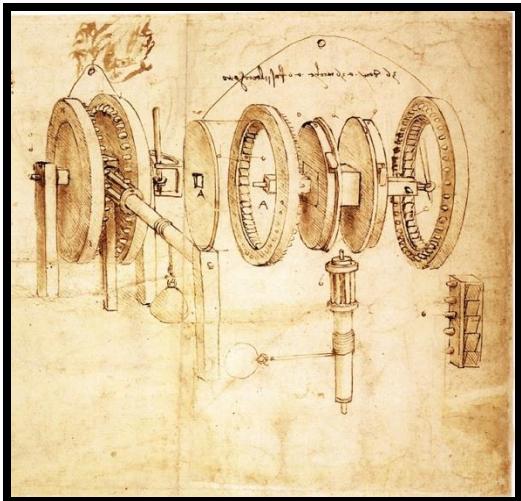
深入理解

数据可视化的主要任务

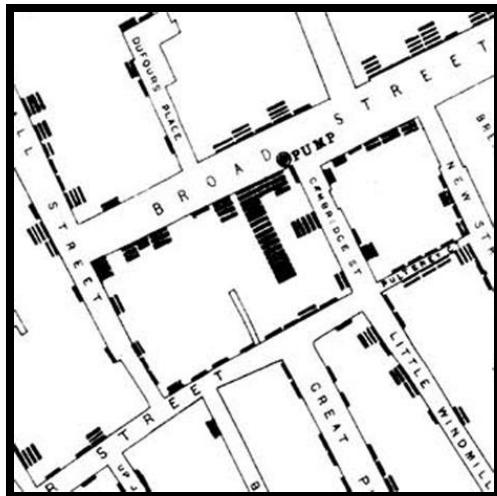
■ 表示数据 – Represent

■ 分析数据 – Analyze

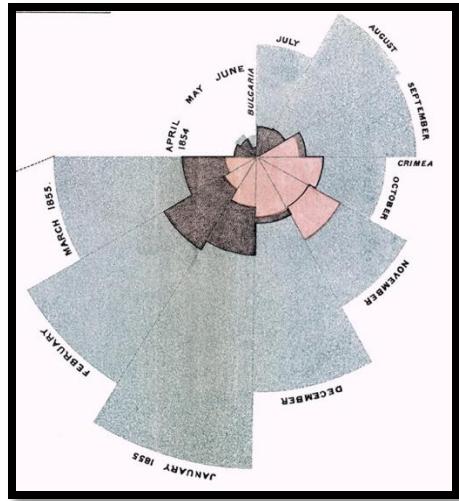
■ 交流数据 – Communicate



三维素描图



霍乱病例的分布



英国东征士兵死亡原因



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dessiné par M. Minard, Inspecteur Général des Ponts et Chaussées en sa qualité

Paris, le 20 Novembre 1869

Les nombres d'hommes perdus sont représentés par les largures des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui me serviront à dresser la carte me sont pris dans les ouvrages de M-M. Chiers, de Léger, de Tercenac, de Chambray et le journal médical de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps de l'Armée Napoléon et du Maréchal Davout qui avaient été détachés sur Moscou et Malibor au commencement avec Oudiné et Wrede, avaient toujours marché avec l'armée.

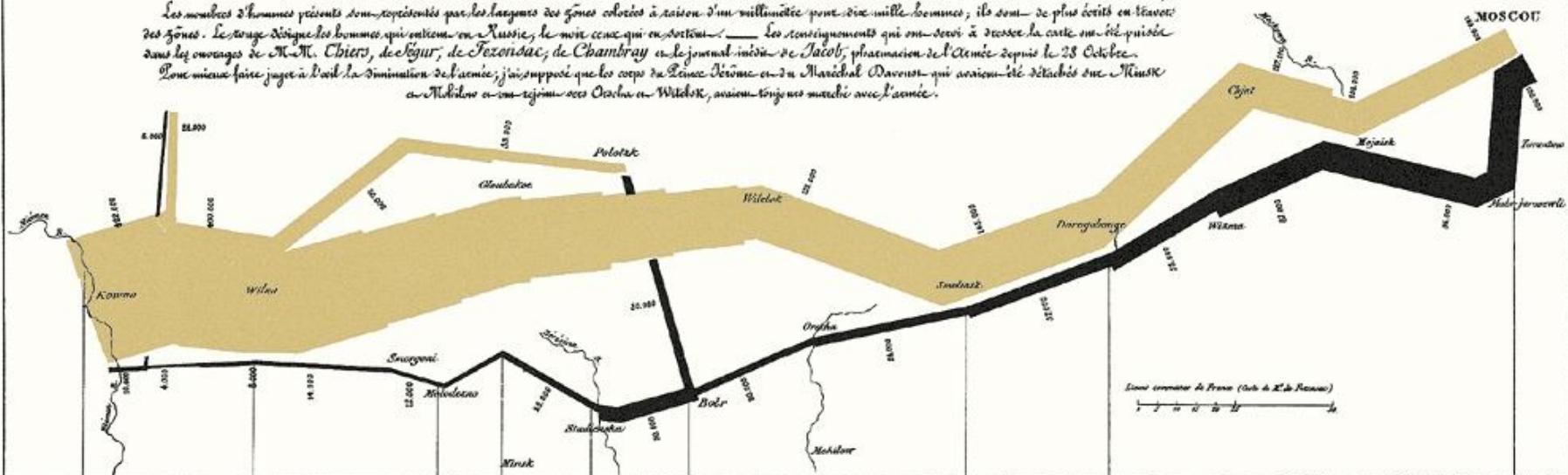
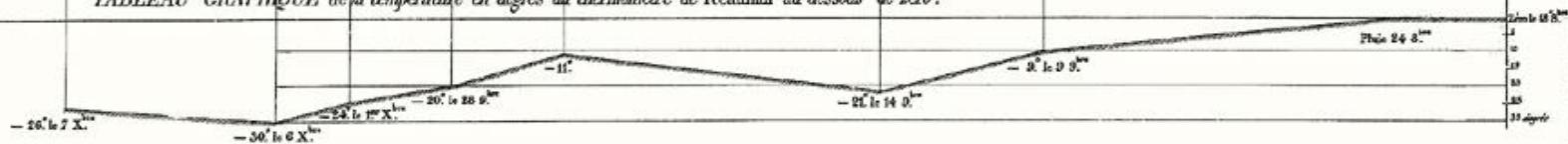


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Casques passent au gelé
le Nôtre, gelé.

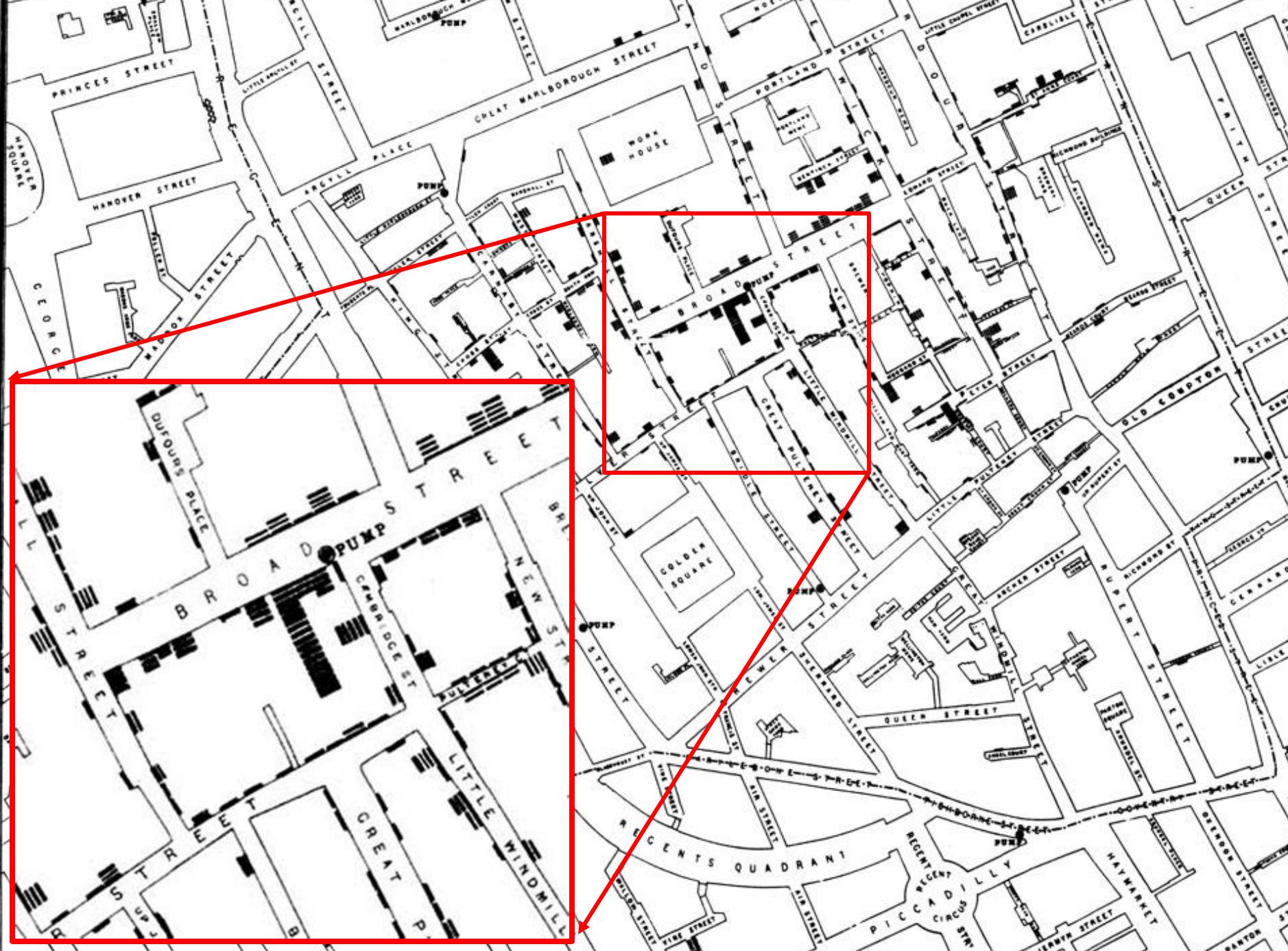


Imprimé par Bony, 8, Rue St Martin, 2^e arrondissement de Paris.

Dep. Int. Repaire et Marque.

C. J. Minard, 1869

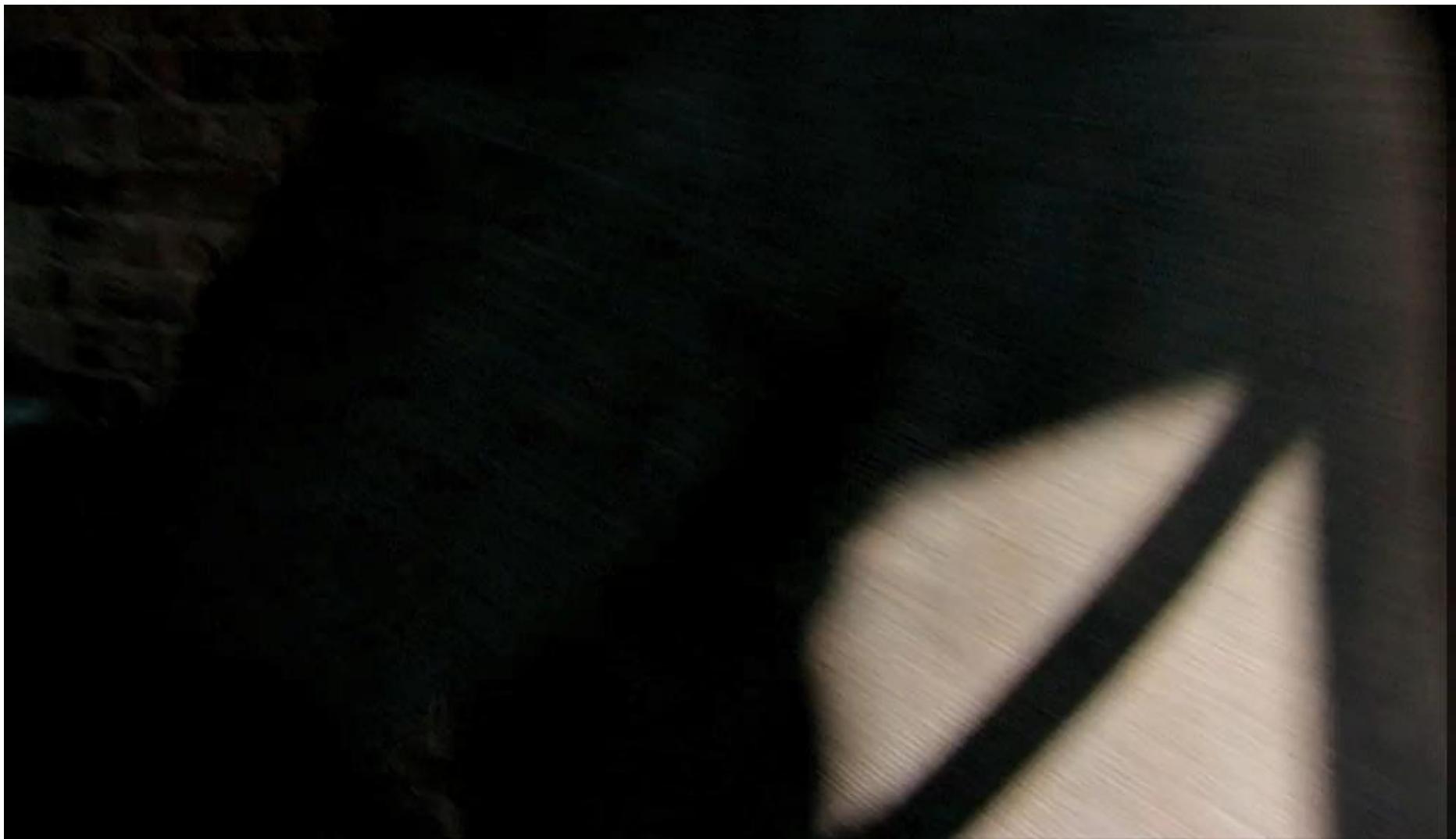
E. Tufte, Writings, Artworks, News





浙江大學
ZHEJIANG UNIVERSITY

Hans Rosling' TED talk



为什么要用数据可视化？



为什么要用数据可视化

- 人类认知存在天生的缺陷

- Change blindness

- 统计数据存在盲区

- Anscombe's Quartet

人类认知的缺陷



**Watch what happens as the
unsuspecting pedestrian
provides directions.**

Daniel J. Simons and Daniel T. Levin,
Failure to detect changes to people during a real world interaction, 1998



统计数据的盲区

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

$$\mu_X = 9.0 \quad \sigma_X = 3.317$$

$$\mu_Y = 7.5 \quad \sigma_Y = 2.03$$

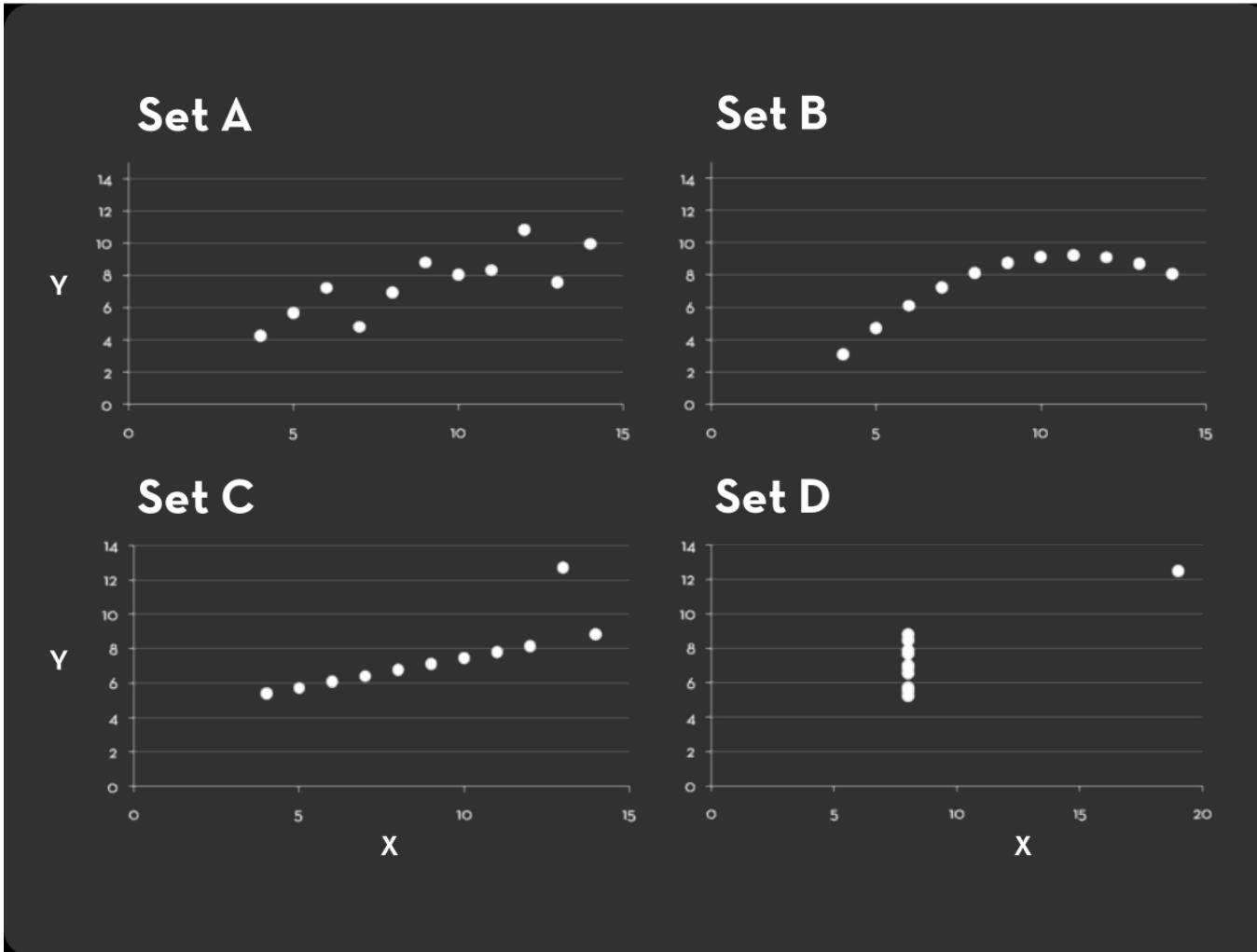
Linear Regression

$$Y = 3 + 0.5 X$$

$$R^2 = 0.67$$

[Anscombe 73]

统计数据的盲区





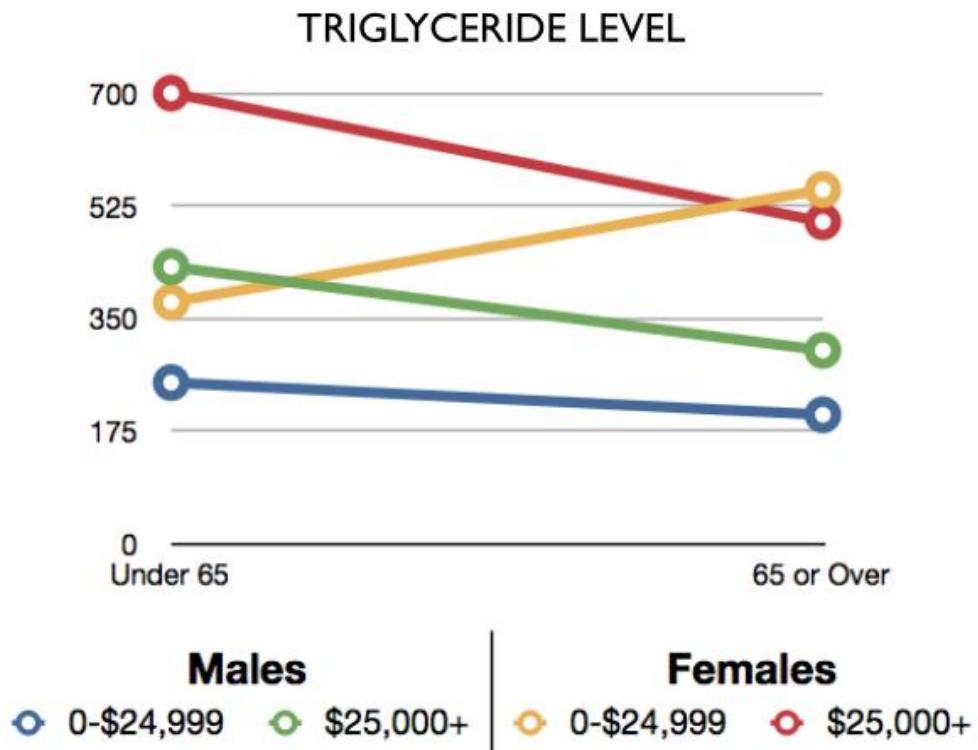
小测试

三酸甘油脂水平在哪个年龄段和收入水平的人群中
随年龄增长表现出不同的趋势？

收入水平	男性		女性	
	65岁以下	65岁及以上	65岁以下	65岁及以上
0-\$24,999	250	200	375	550
\$25,000+	430	300	700	500

数据可视化的力量

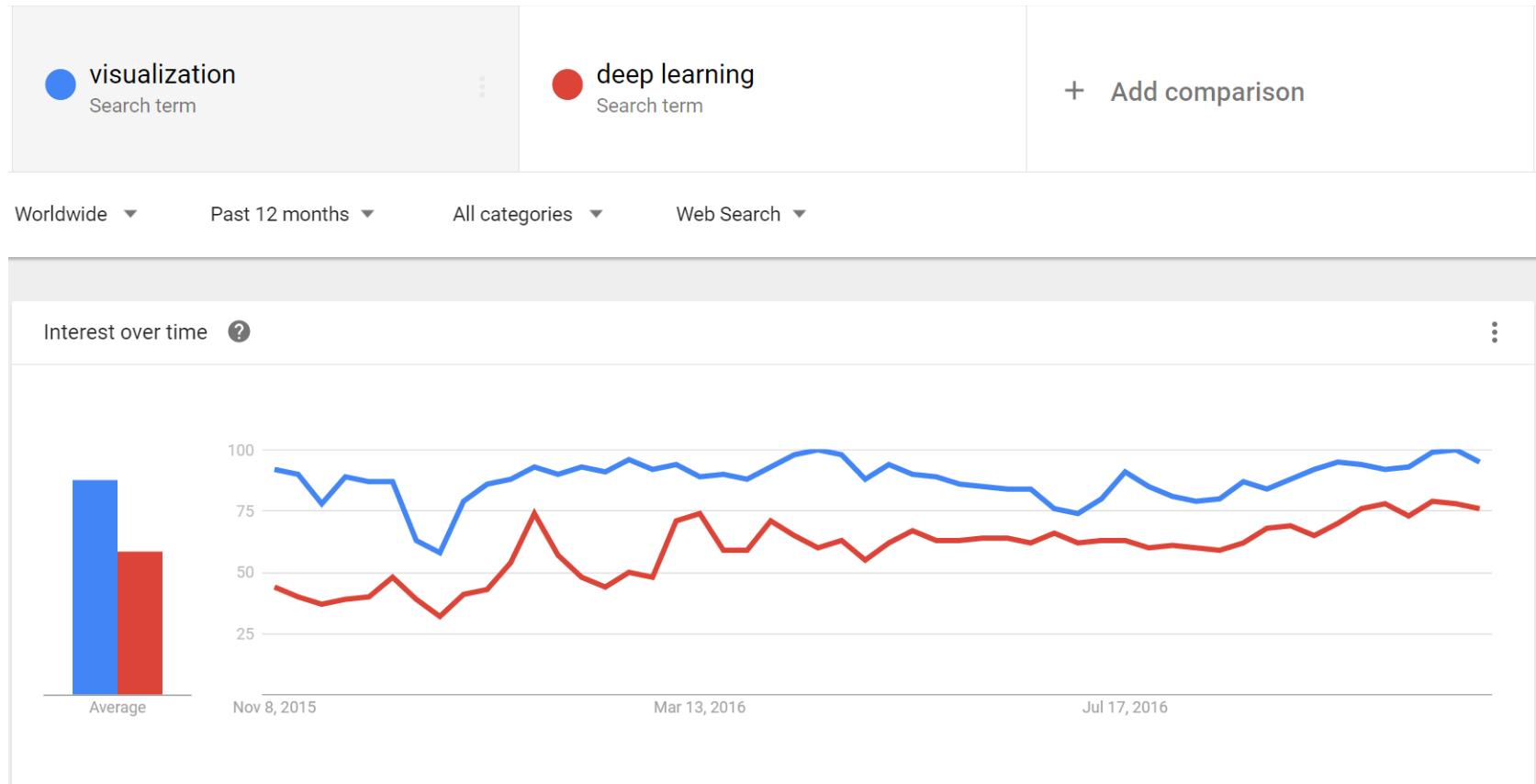
三酸甘油脂水平在哪个年龄段和收入水平的人群中随年龄增长表现出不同的趋势？





可视化热度

可视化与深度学习



可视化是一个跨学科领域

可视化与图形学



图形学

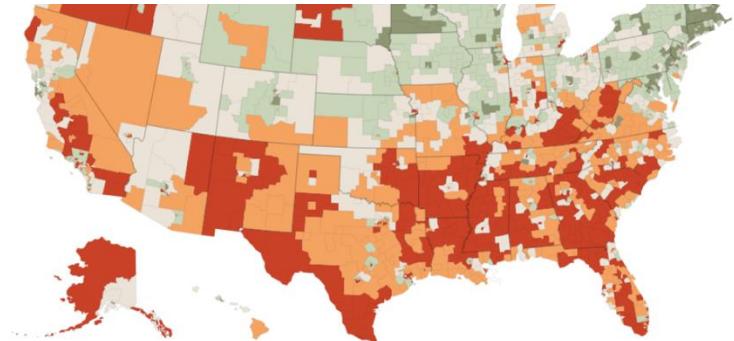
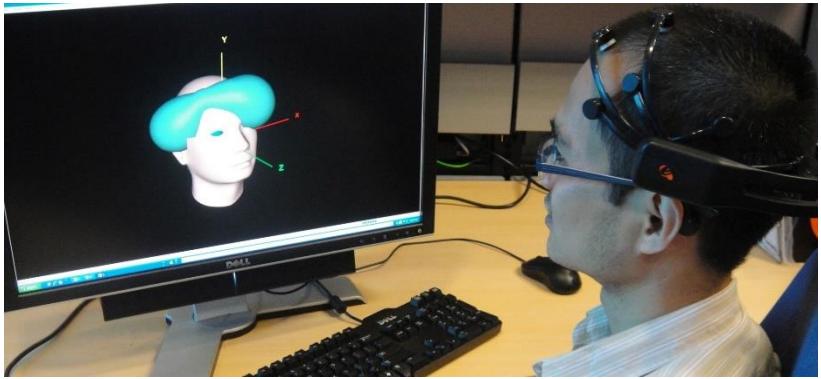
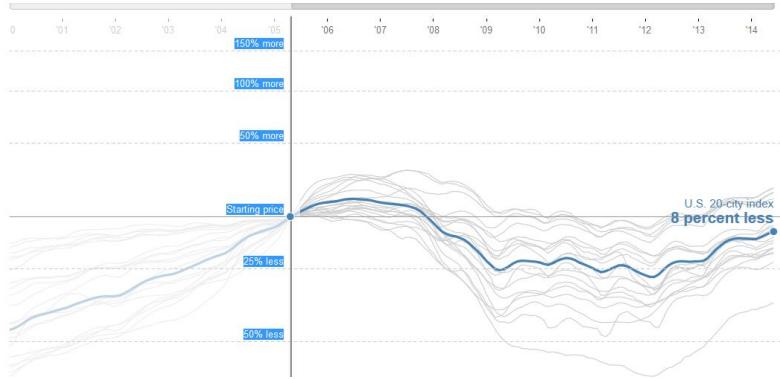
- 照片真实感
- 模拟
- 现实世界
- 视觉媒介

可视化

- 描述性
- 增强理解
- 传达信息

可视化与人机交互

- 可视化主要处理数据
- 人机交互主要关注人与机器之间的交互行为



可视化与数据挖掘

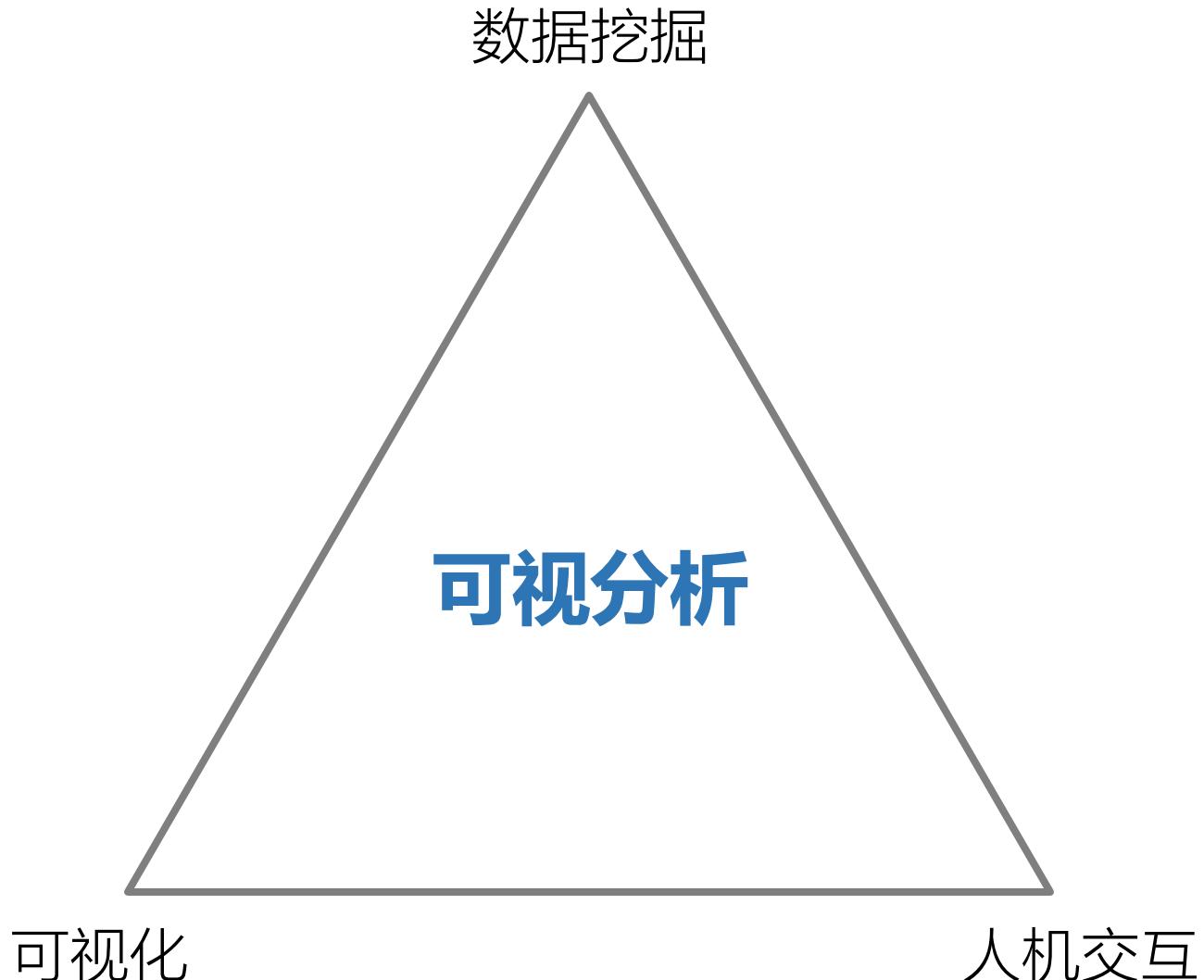
- 数据挖掘 (Data Mining) 更关注自动化的算法
- 可视化通过让人类参与其中实现交互式的数据分析

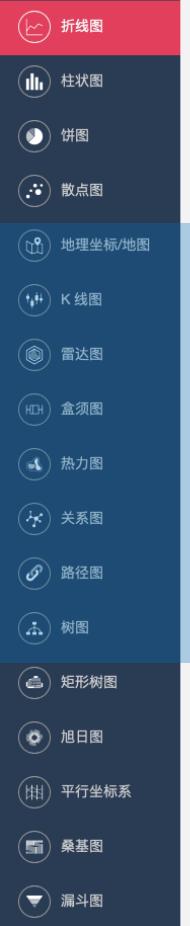


(港科大) 屈华民教授对可视化与数据挖掘的观点



可视分析





折线图 Line

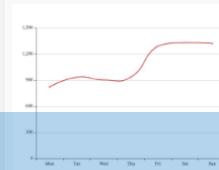
Basic Line Chart



Basic area chart



Smoothed Line Chart



Stacked area chart



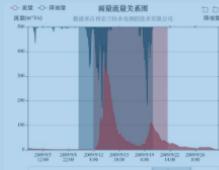
Stacked Line Chart



Area Pieces



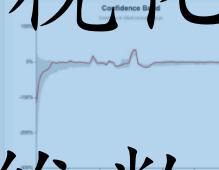
Rainfall



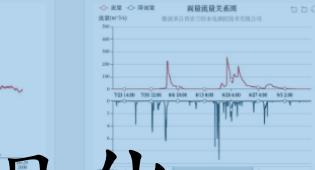
Large scale area chart



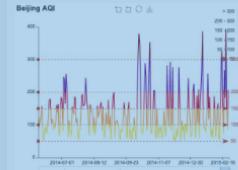
Dynamic Data + Time Axis



Rainfall and Water Flow



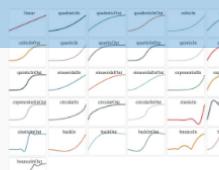
Beijing AQI



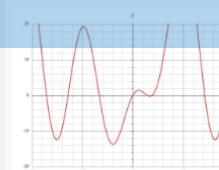
Try Dragging these Points



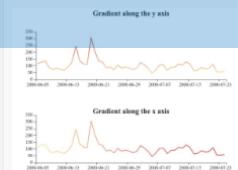
Line Easing Visualizing



Function Plot



Line Gradient



Custom Graphic Component



Line Chart in Cartesian Coord...



Log Axis



Temperature Change in the c...



Line with Marklines



Click to Add Points



Two Value-Axes in Polar



Two Value-Axes in Polar

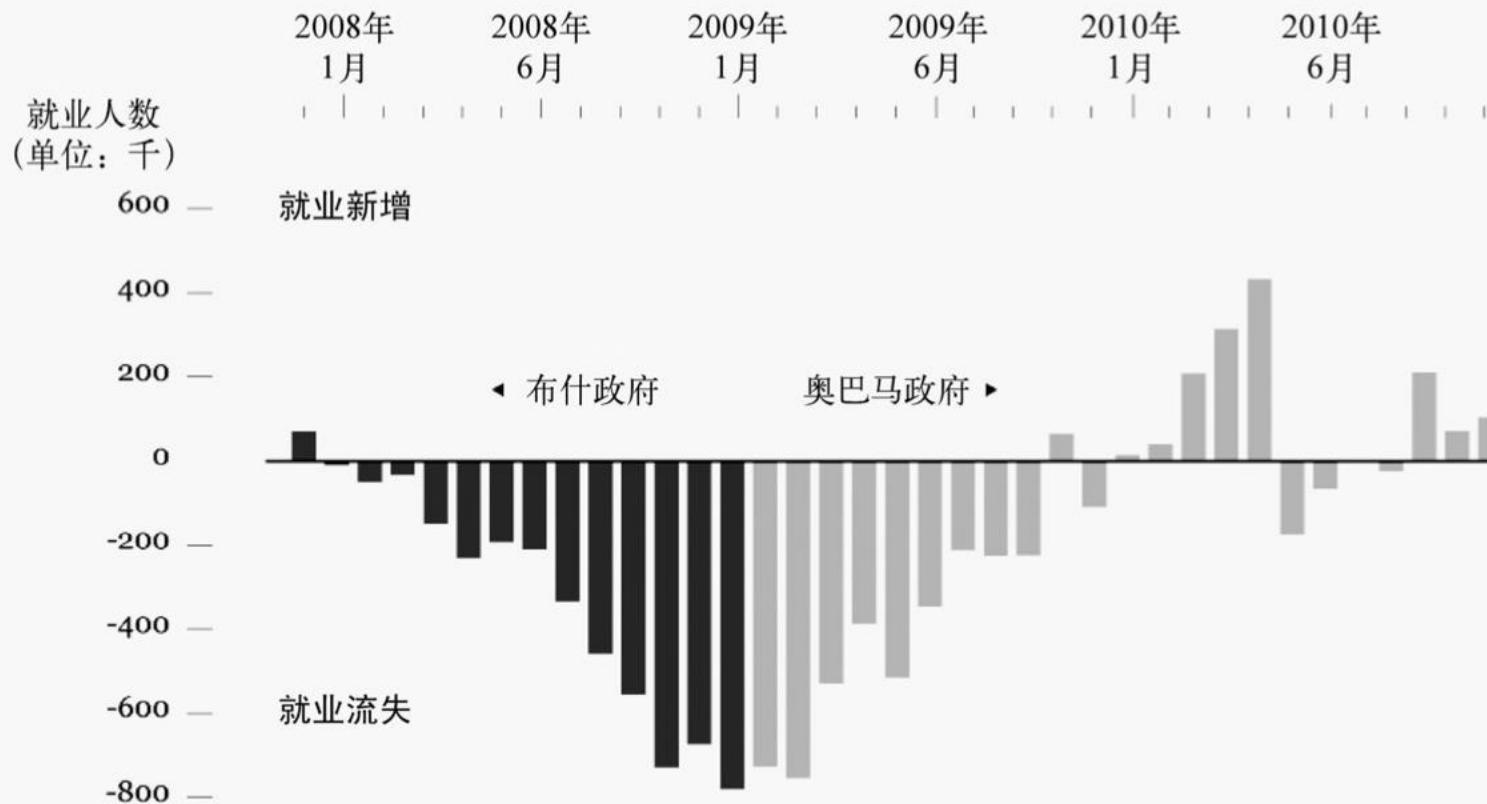


以Echarts为例 <https://echarts.apache.org/zh/index.html>

可视化建模 1. 一维数据可视化

一维数据可视化

美国新的就业机会

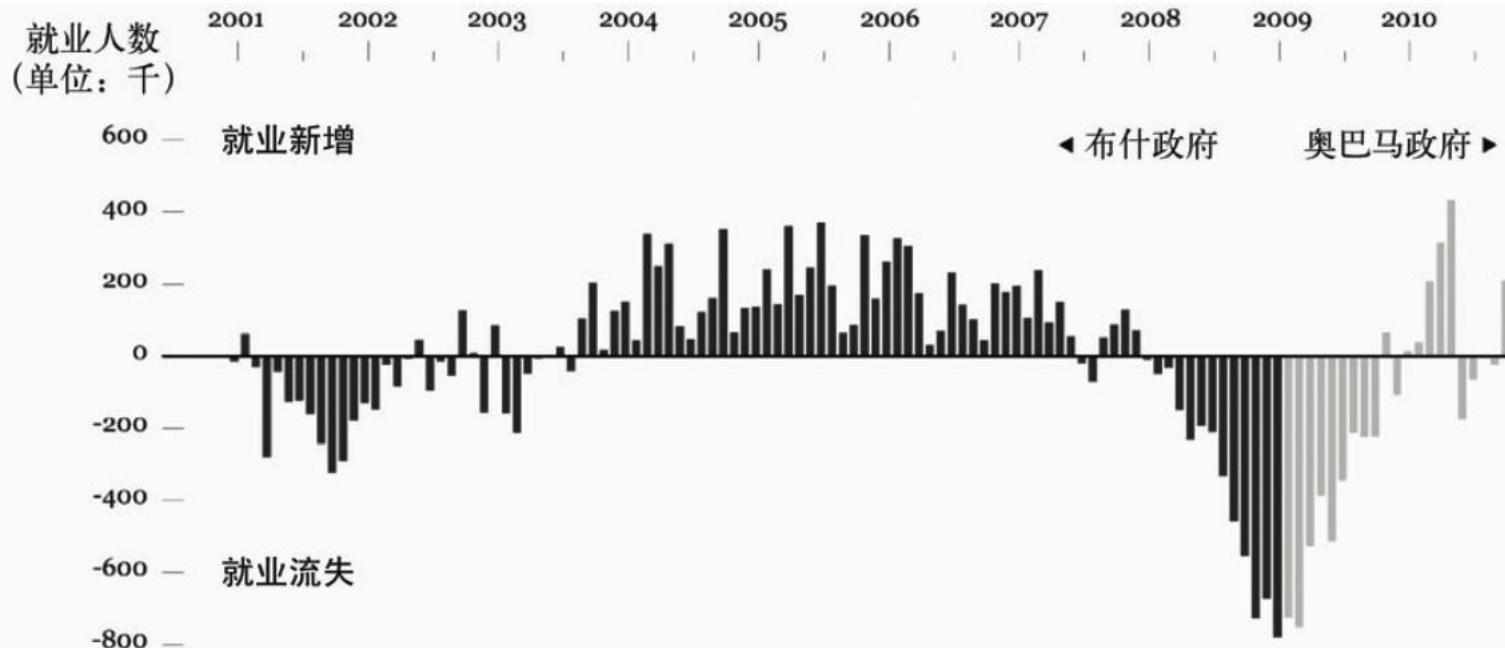


Source: Bureau of Labor Statistics | Nathan Yau

巴拉克·奥巴马执政后失业的变化

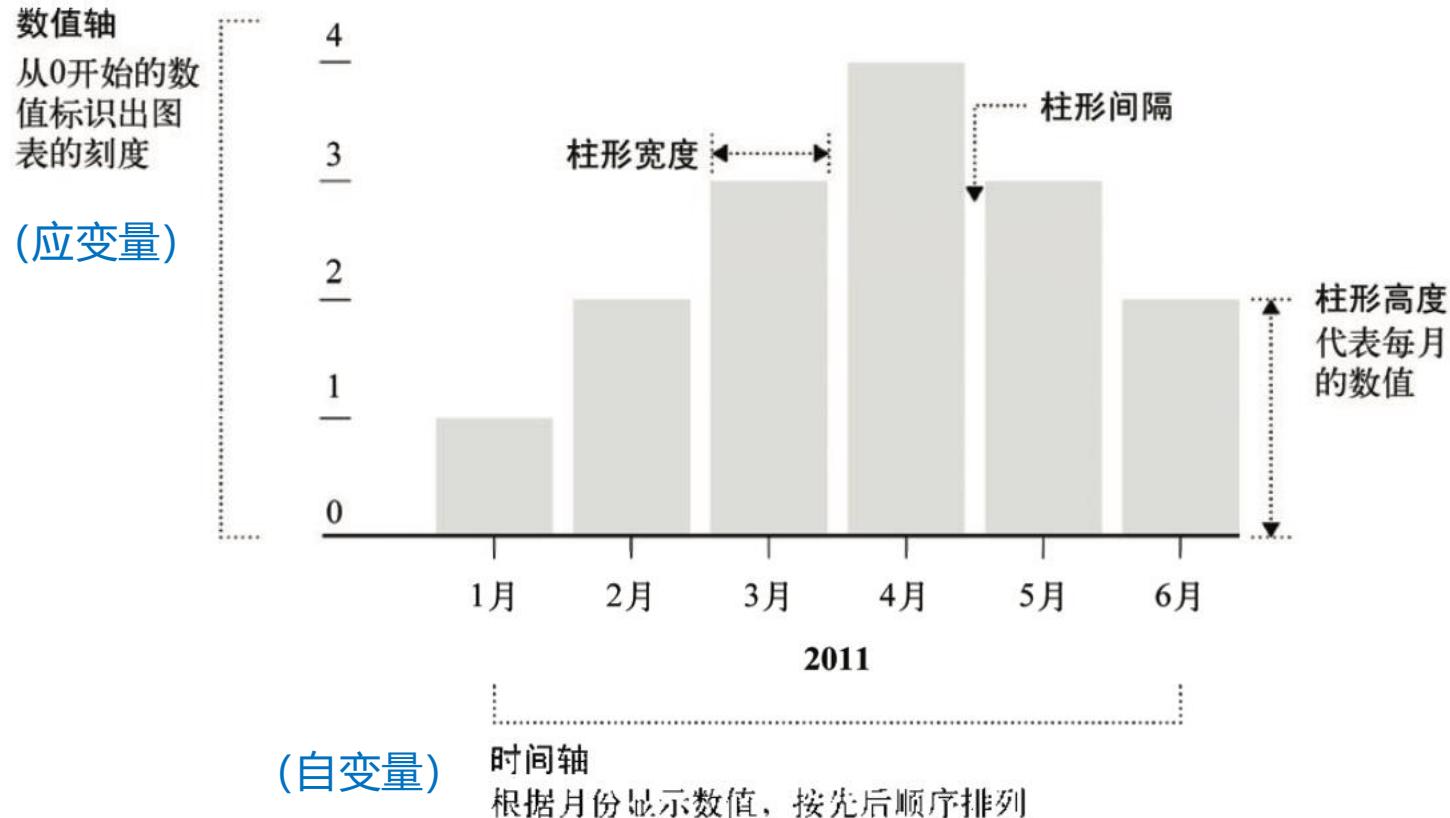
一维数据可视化

美国新的就业机会



2001—2010年的美国失业变化

一维数据可视化 - 柱形图



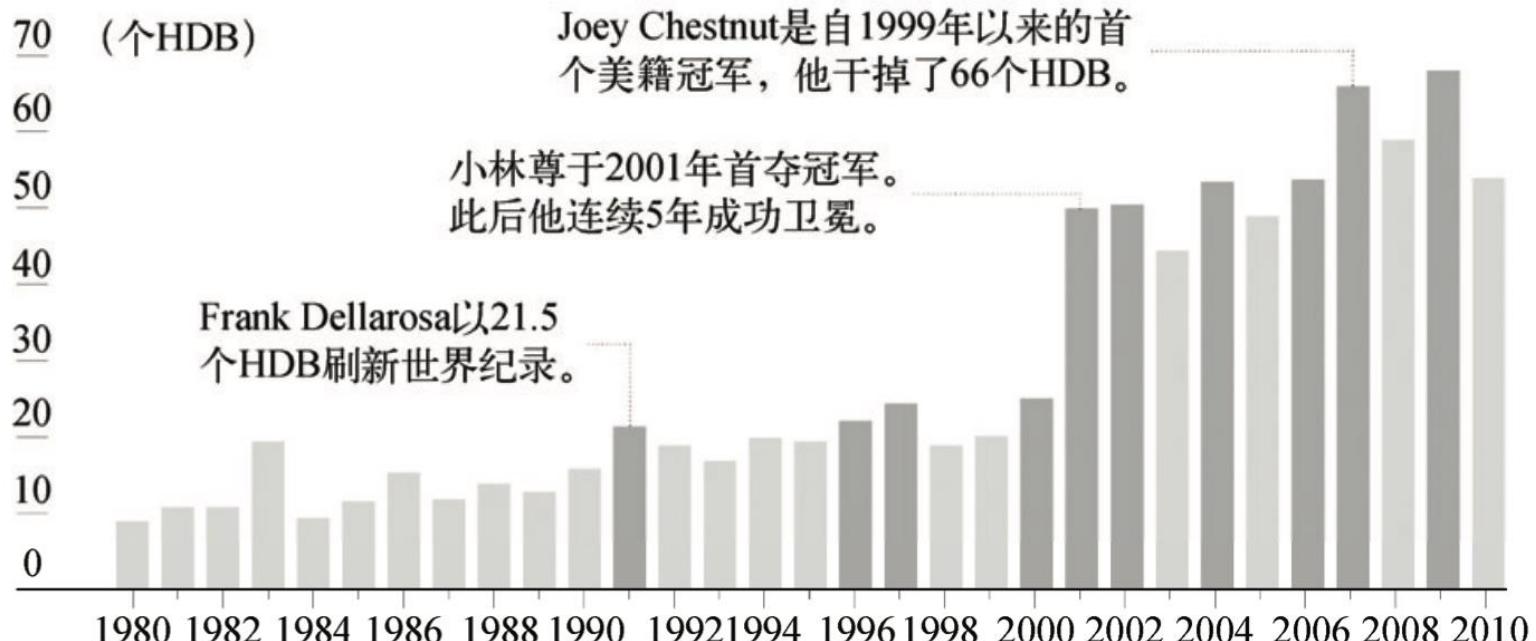
柱形图的基本框架

一维数据可视化 - 柱形图例



热狗大胃王

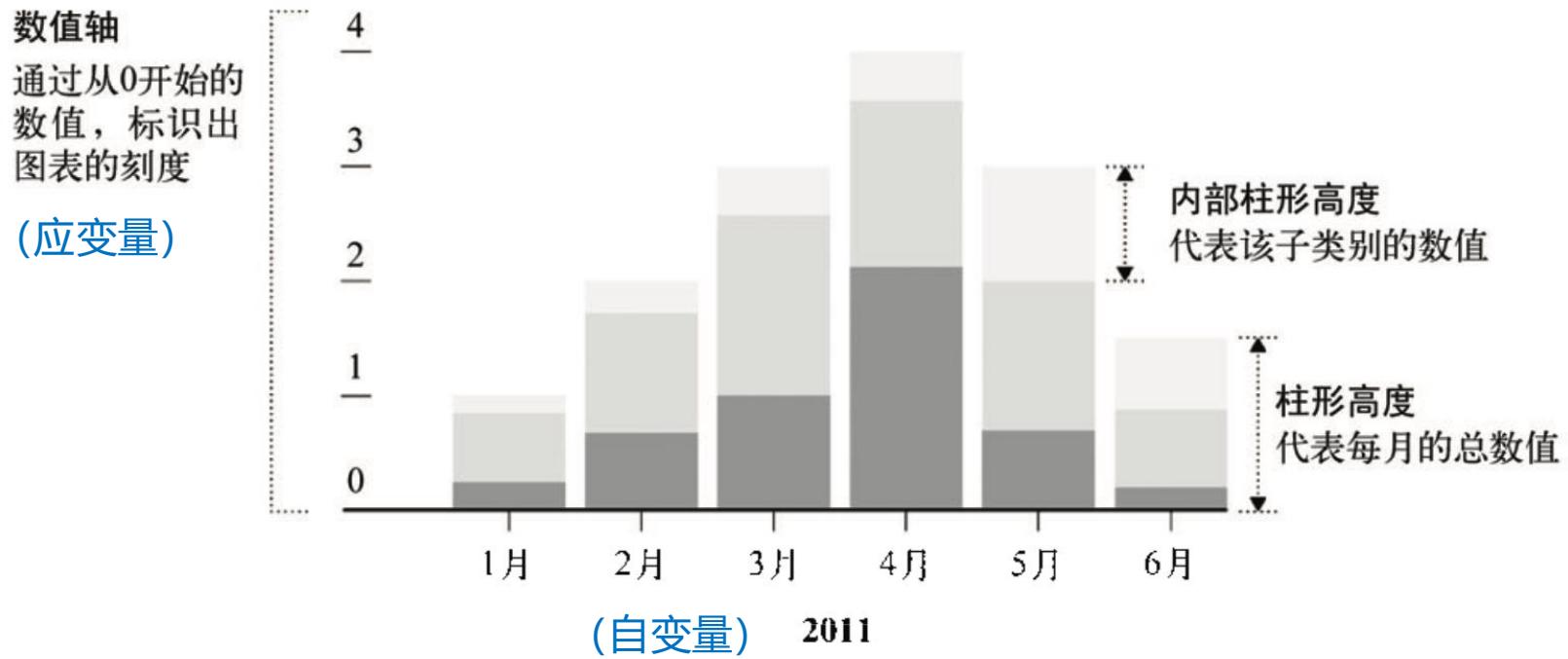
每年7月4日举办的“内森杯”热狗大胃王比赛始于20世纪初，但直到2001年才开始获得广泛关注。当时来自日本的选手小林尊横空出世，一举将该项世界纪录刷新至之前的两倍多。绿色显示的柱形表示新纪录。



Source: Wikipedia | Nathan Yau

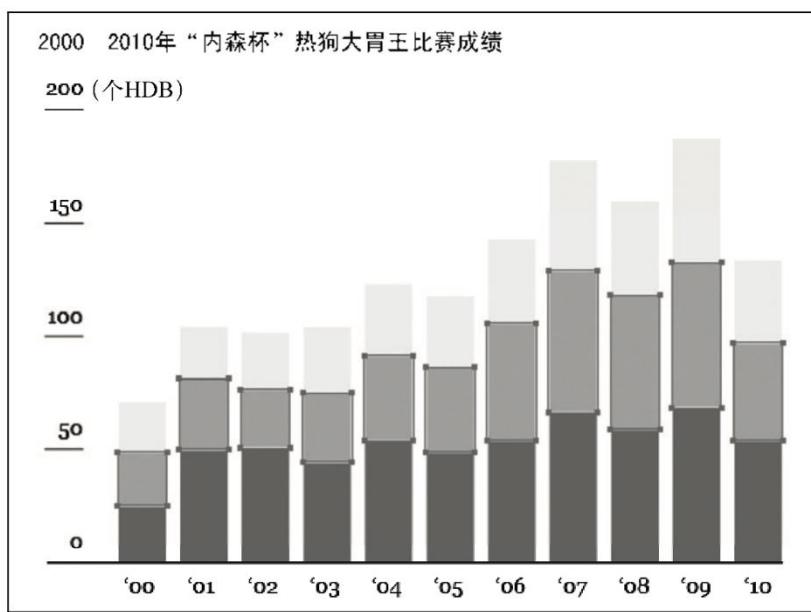
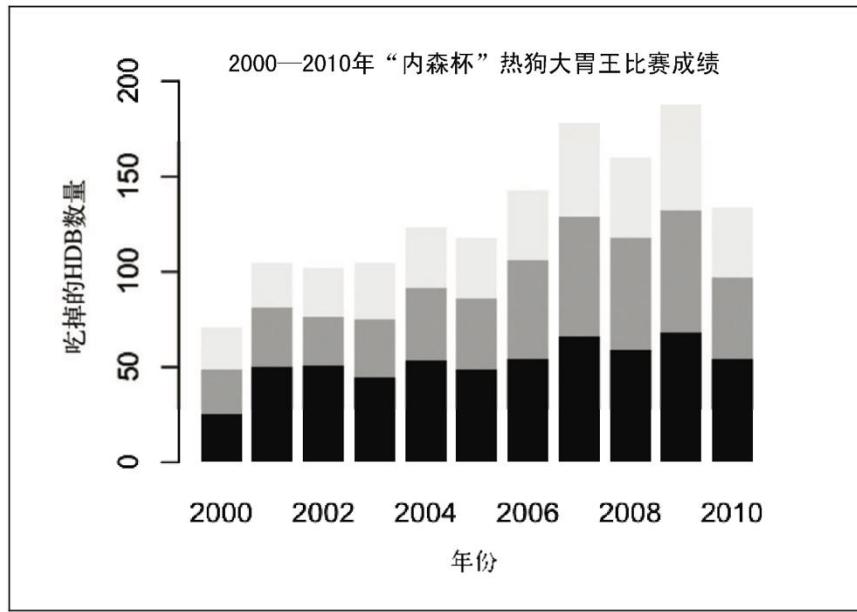
“内森杯” 热狗大胃王比赛成绩的柱形图

一维数据可视化 - 堆叠柱形图



堆叠柱形图的基本框架

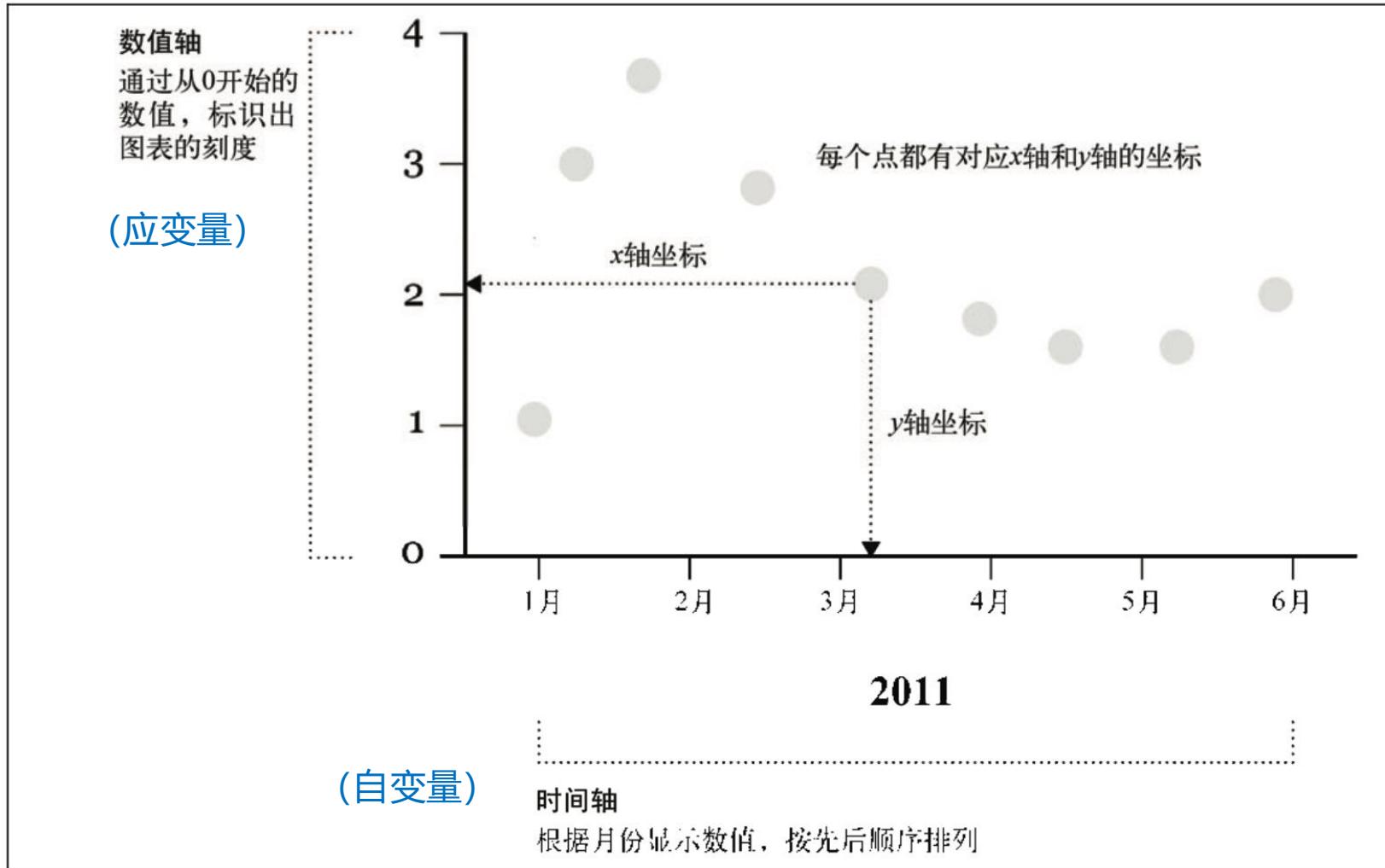
一维数据可视化 - 堆叠柱形图例



R语言创建的堆叠柱形图
(提示：调用[barplot](#)函数)

美化后的堆叠柱形图
(提示：使用[Illustrator](#))

一维数据可视化 - 散点图



散点图的基本框架



一维数据可视化 - 散点图例

订阅人数增长

2010年1月，通过RSS和邮件订阅FlowingData网站的人数增长到了27 611，使之成为第10个连续增长超过10%的月份。

30 (千名订阅者)

25
20
25 047

27 611
(+10%)

15

10

5

0

1 5 10 15 20 25 30
2010年1月

数据报告错误
12日及13日，来源的数据报告出现
错误。订阅人数相差超过17 000。

Source: Feedburner | Nathan Yau

在R中创建并在Illustrator中设计得到的散点图

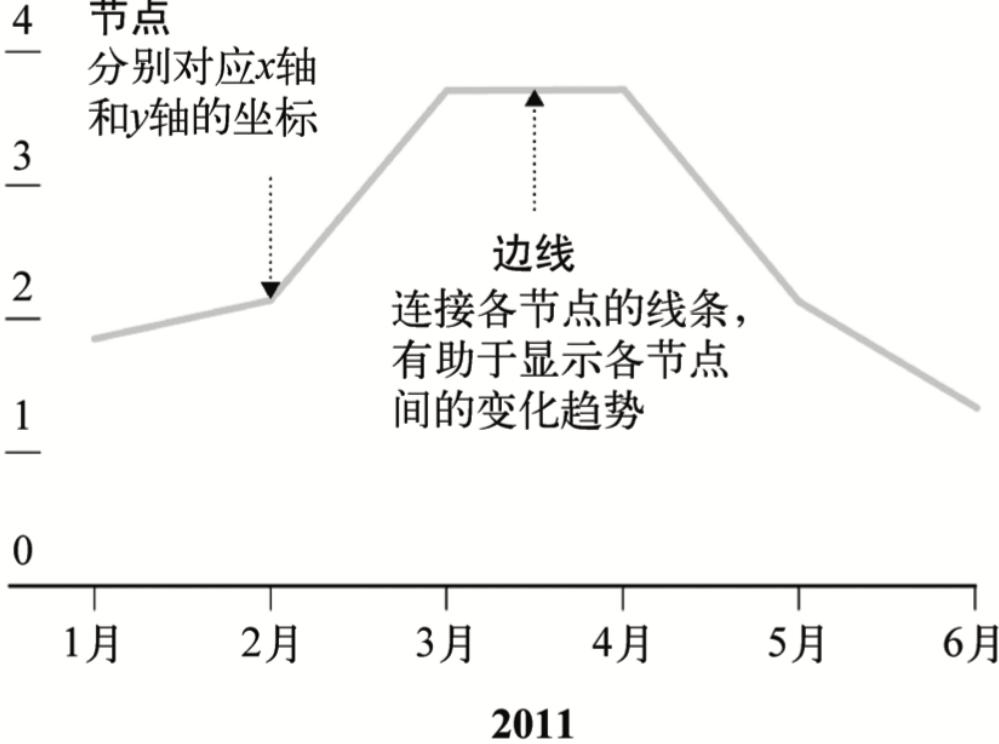
一维数据可视化 - 折线图

数值轴

通过从0开始的数值，标识出图表的刻度

节点

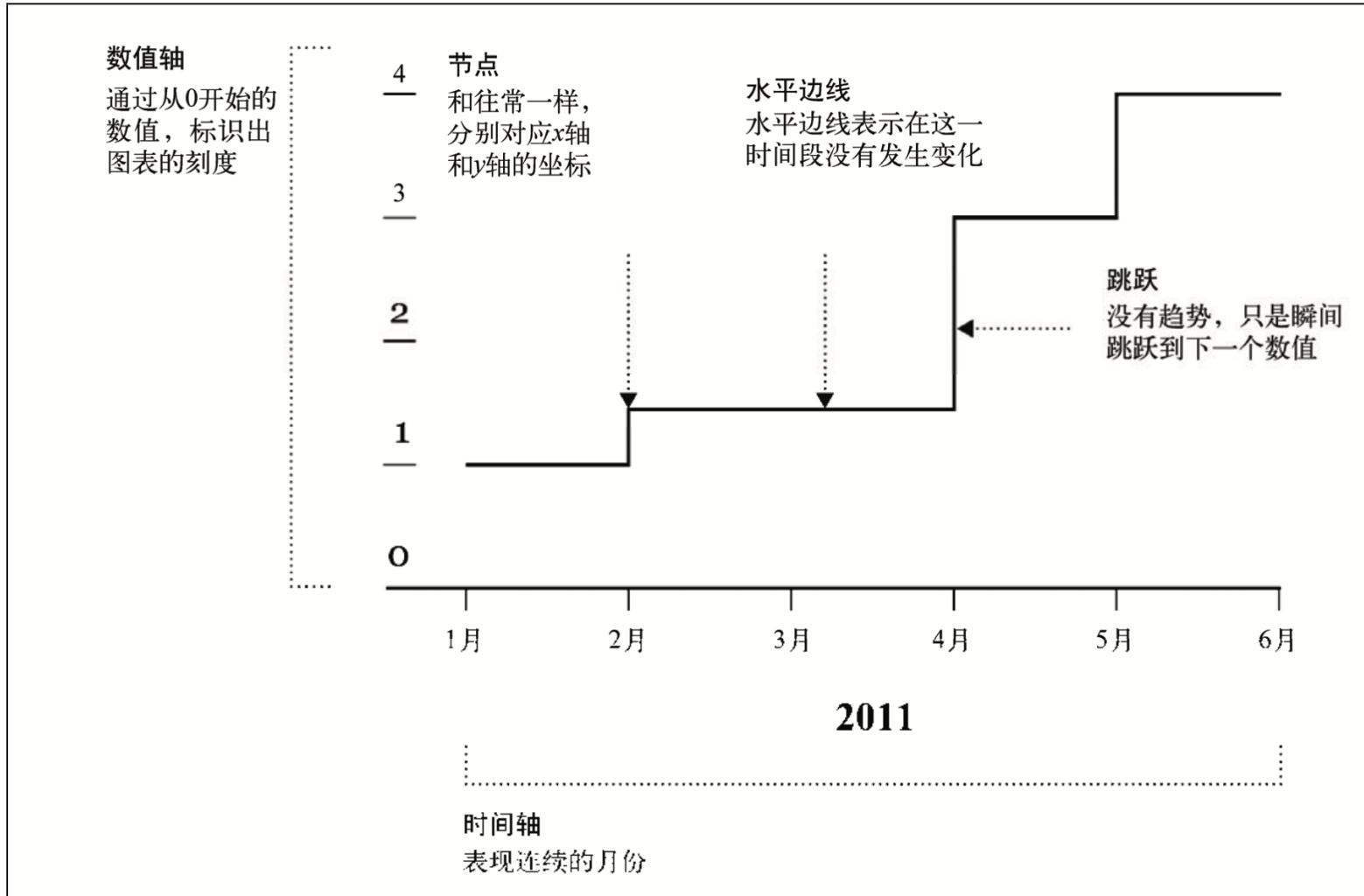
分别对应x轴和y轴的坐标



时间轴
表现连续的月份

思考：延续型和离散型表达的异同点

一维数据可视化 - 阶梯图

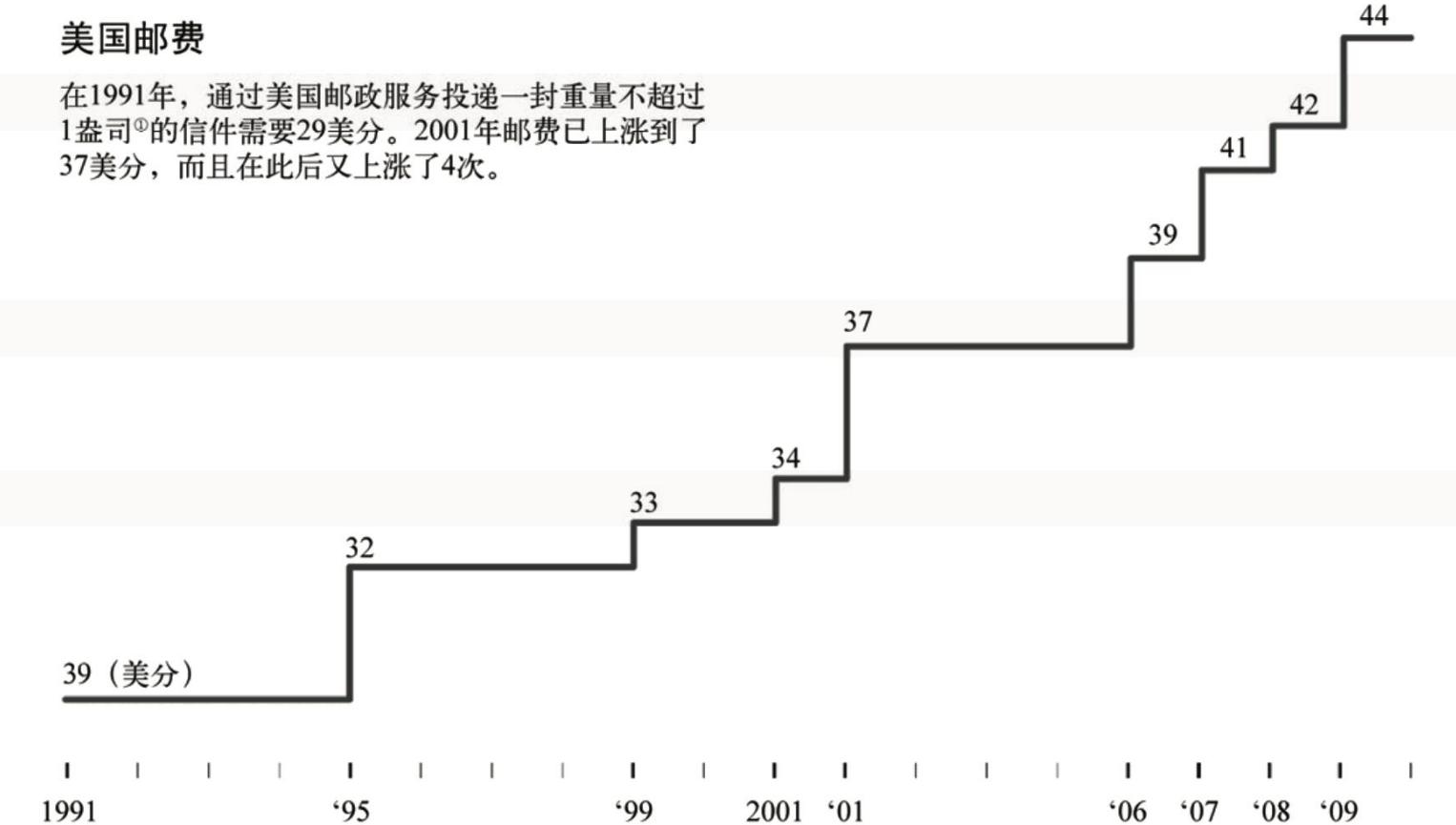


思考：延续型和离散型表达的异同点

一维数据可视化 - 阶梯图例

美国邮费

在1991年，通过美国邮政服务投递一封重量不超过1盎司^①的信件需要29美分。2001年邮费已上涨到了37美分，而且在此后又上涨了4次。

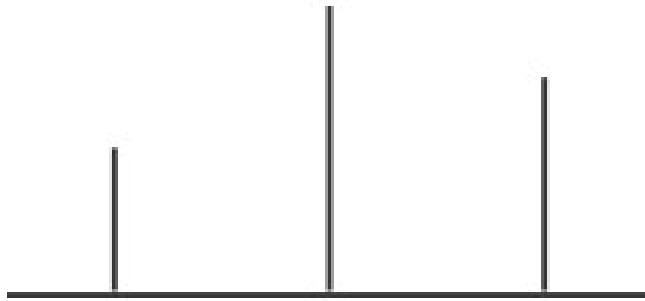


显示美国邮费变化的阶梯图



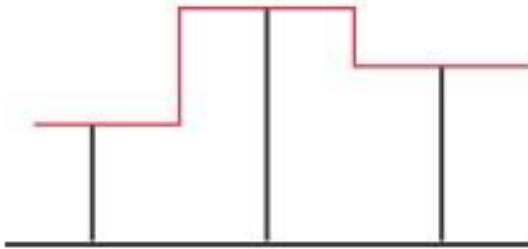
一维数据可视化 - 插值

Goal: Interpolate Values



一维数据可视化 - 插值

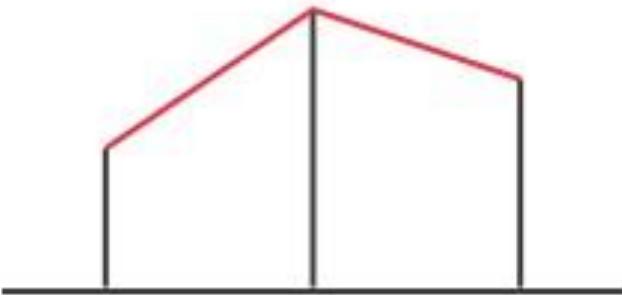
Nearest Neighbor Interpolation



Problem: values not continuous

一维数据可视化 - 插值

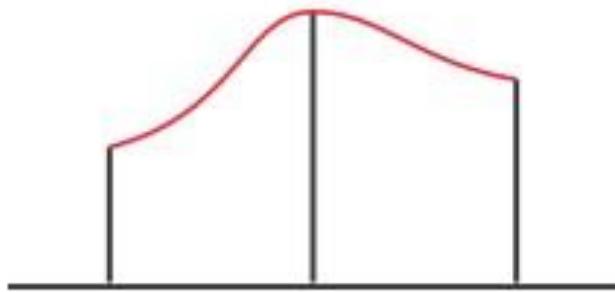
Linear Interpolation



Problem: derivatives not continuous

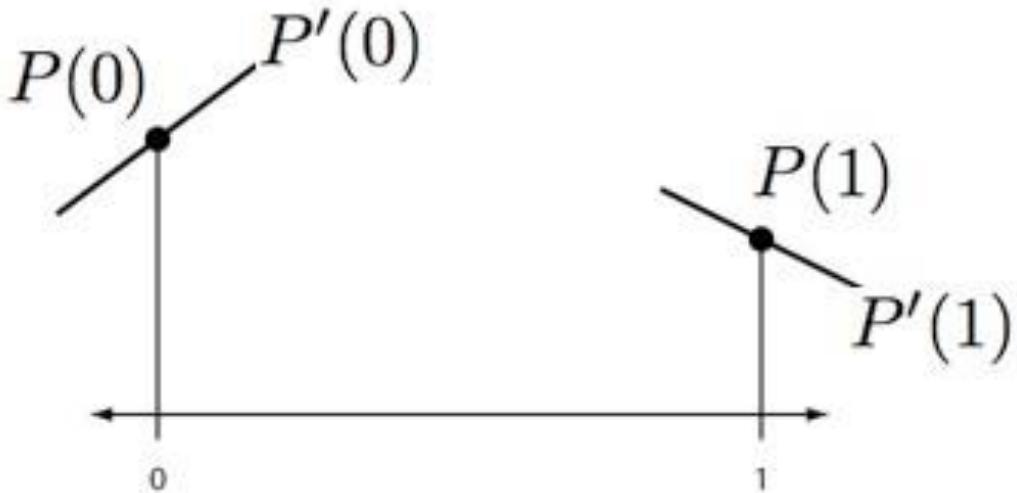
一维数据可视化 - 插值

Smooth Interpolation?



一维数据可视化 - 插值

Cubic Hermite Interpolation



Given: values and derivatives at 2 points



Cubic Polynomial Interpolation

Assume cubic polynomial

$$P(t) = a t^3 + b t^2 + c t + d$$

Why? 4 constraints => need 4 degrees of freedom



一维数据可视化 – 插值

Cubic Hermite Interpolation

Assume cubic polynomial

$$P(t) = a t^3 + b t^2 + c t + d$$

$$P'(t) = 3a t^2 + 2b t + c$$

Solve for coefficients:

$$P(0) = h_0 = d$$

$$P(1) = h_1 = a + b + c + d$$

$$P'(0) = h_2 = c$$

$$P'(1) = h_3 = 3a + 2b + c$$



一维数据可视化 – 插值

Matrix Representation

$$h_0 = d$$

$$h_1 = a + b + c + d$$

$$h_2 = c$$

$$h_3 = 3a + 2b + c$$

$$\begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$



一维数据可视化 – 插值

Matrix Representation of Polynomials

$$P(t) = [a \ b \ c \ d] \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix}$$

Hermite Basis Functions

$$[a \ b \ c \ d] \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix} = [h_0 \ h_1 \ h_2 \ h_3] \begin{bmatrix} H_0(t) \\ H_1(t) \\ H_2(t) \\ H_3(t) \end{bmatrix}$$

$$P(t) = \sum_{i=0}^3 h_i H_i(t)$$

Matrix Representation

$$\begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

Solve for a, b, c, d

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \end{bmatrix}$$

Inverse Matrix



一维数据可视化 – 插值

Matrix Inverse

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Change Basis

$$[a \ b \ c \ d] \begin{bmatrix} 0 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix}$$

{ }

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Change Basis

$$[a \ b \ c \ d] \begin{bmatrix} 0 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix}$$

{ }

$$[h_0 \ h_1 \ h_2 \ h_3]$$



一维数据可视化 – 插值

Change Basis

$$[a \ b \ c \ d] \begin{bmatrix} 0 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix}$$

$\underbrace{\hspace{10em}}_{[h_0 \ h_1 \ h_2 \ h_3]} \quad \underbrace{\hspace{10em}}_{\begin{bmatrix} H_0(t) \\ H_1(t) \\ H_2(t) \\ H_3(t) \end{bmatrix}}$

一维数据可视化 – 插值

Hermite Basis Functions

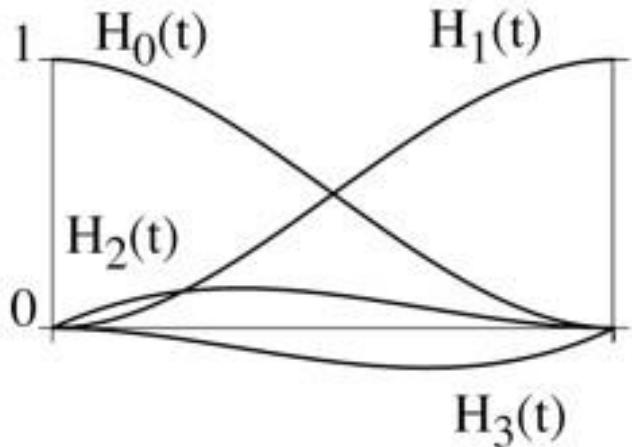
$$\begin{bmatrix} H_0(t) \\ H_1(t) \\ H_2(t) \\ H_3(t) \end{bmatrix} = \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix}$$

$$H_0(t) = 2t^3 - 3t^2 + 1$$

$$H_1(t) = -2t^3 + 3t^2$$

$$H_2(t) = t^3 - 2t^2 + t$$

$$H_3(t) = t^3 - t^2$$

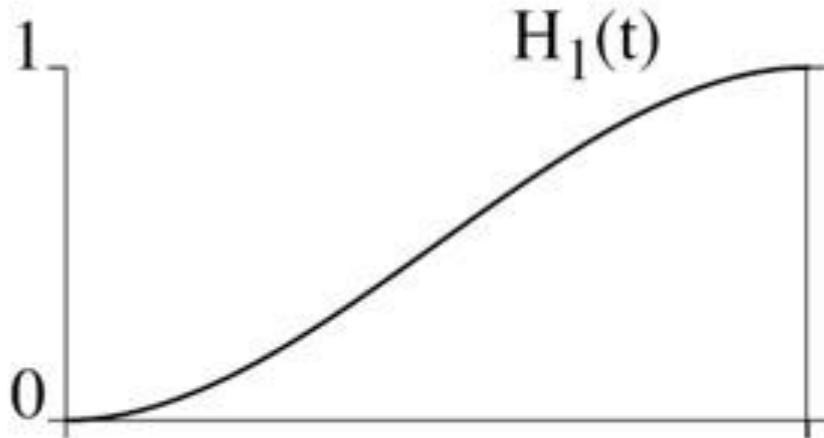


一维数据可视化 – 插值

Ease

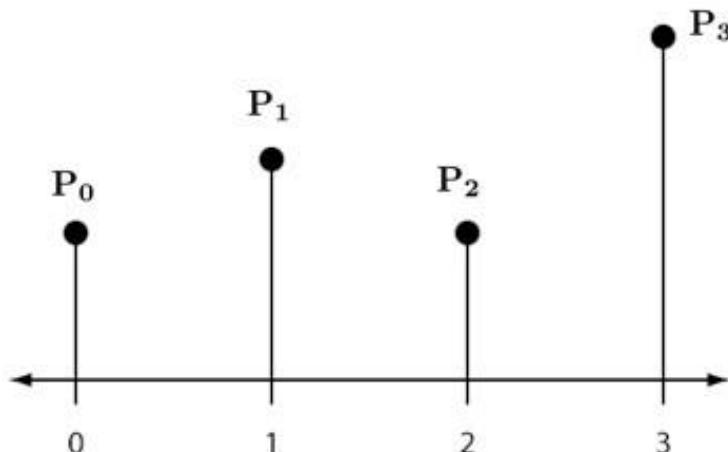
A very useful function

In animation, start and stop slowly (zero velocity)

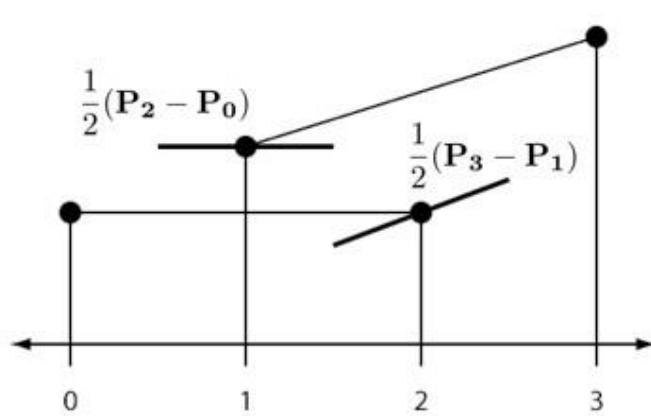
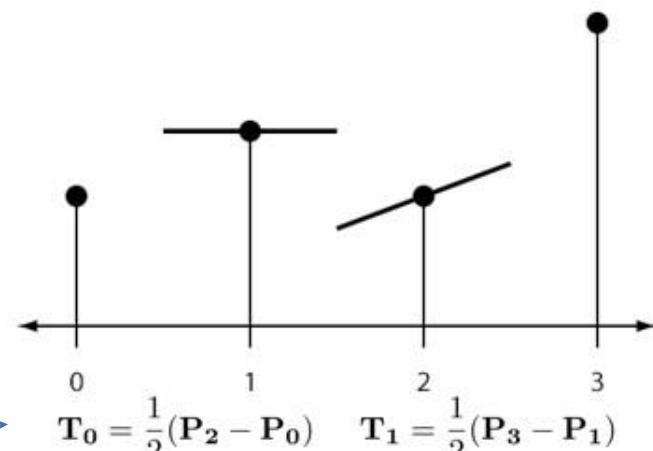
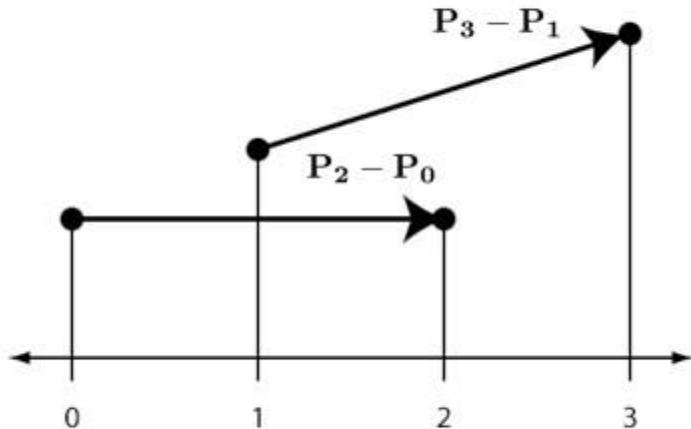


$$H_1(t) = -2t^3 + 3t^2 = t^2(3 - 2t)$$

一维数据可视化 – Catmull-Rom插值



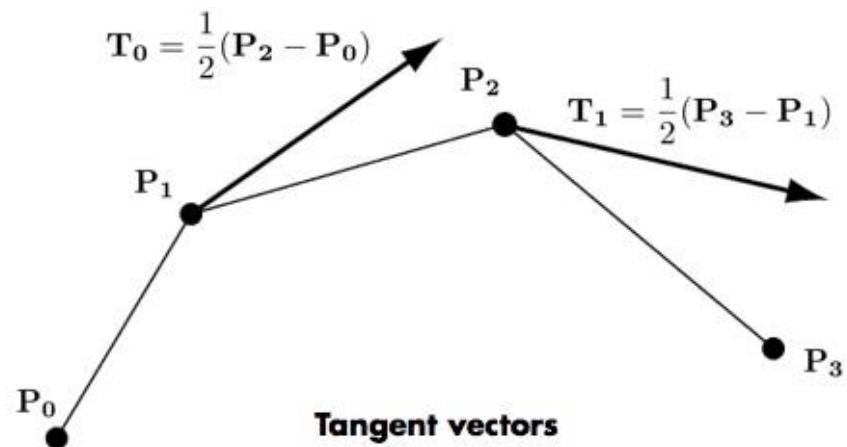
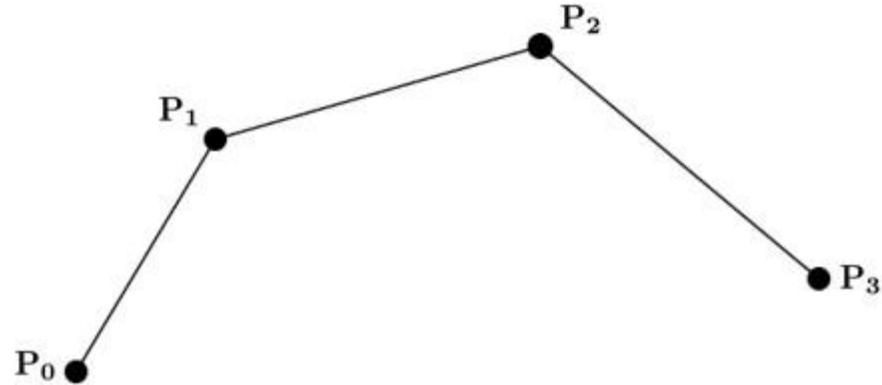
Interpolate points smoothly
Slopes not given though



一维数据可视化 – Catmull-Rom插值



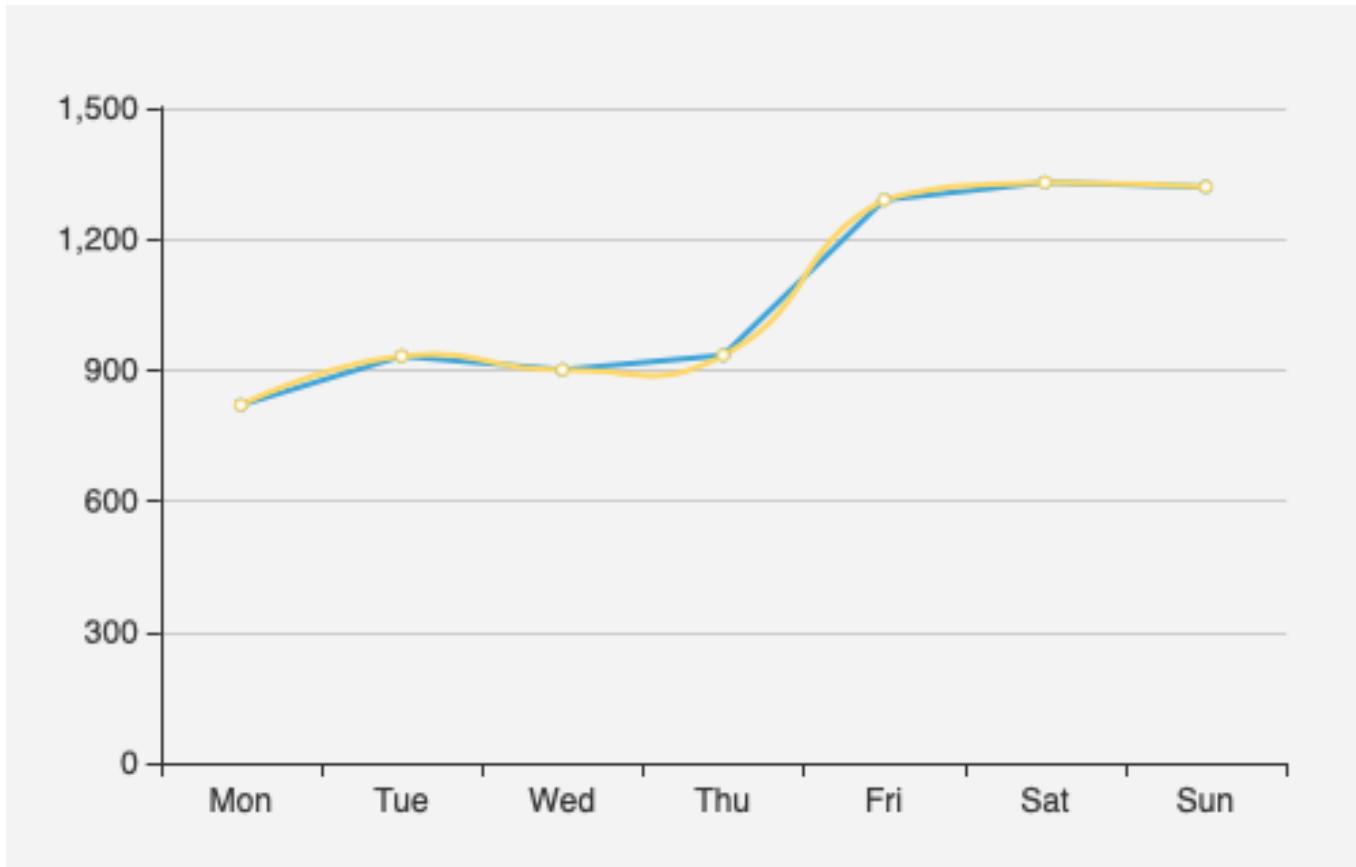
Catmull-Rom Interpolation



We can interpolate points as easily as values

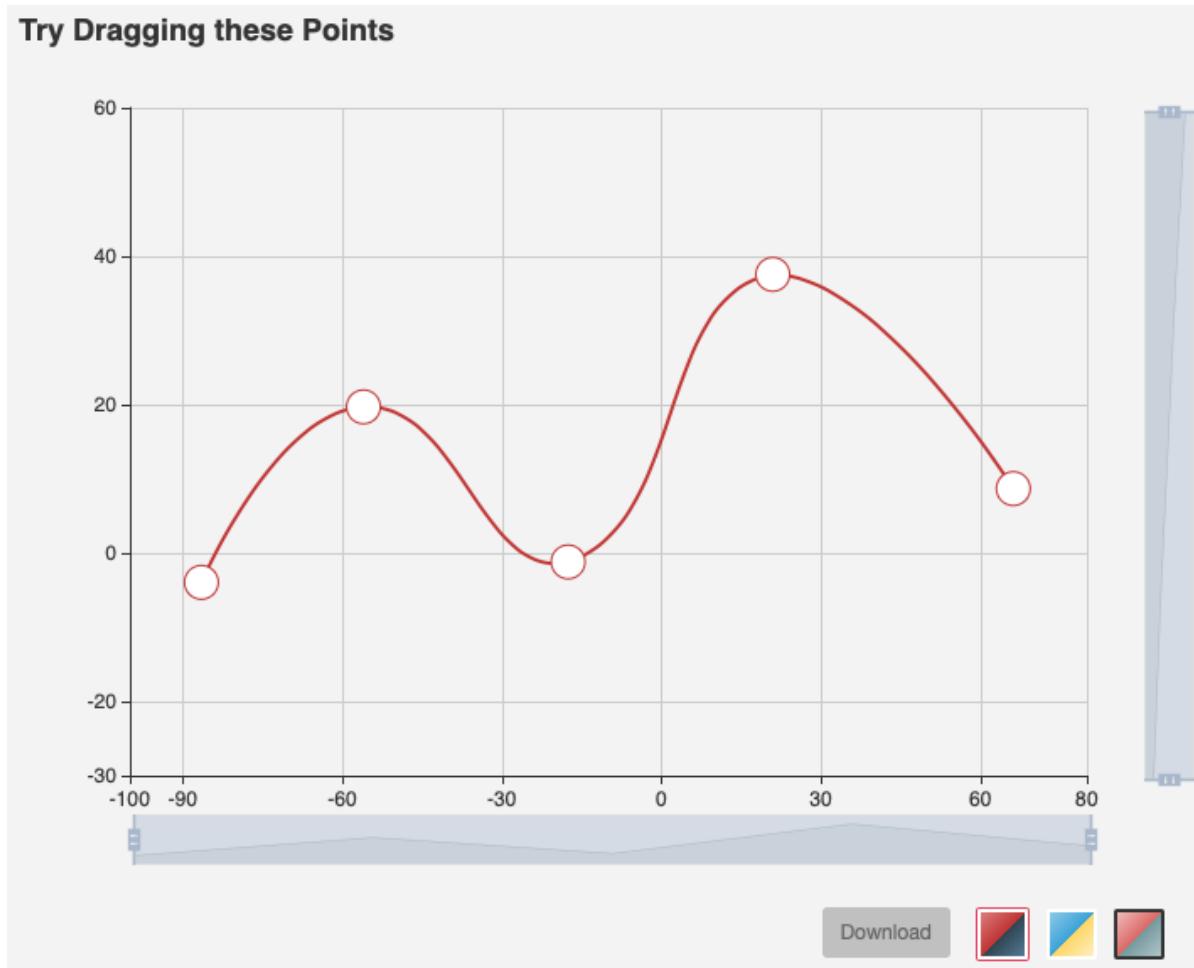
$$\mathbf{p}(t) = (2t^3 - 3t^2 + 1)\mathbf{p}_0 + (t^3 - 2t^2 + t)\mathbf{m}_0 + (-2t^3 + 3t^2)\mathbf{p}_1 + (t^3 - t^2)\mathbf{m}_1$$

一维数据可视化 – Catmull-Rom插值例



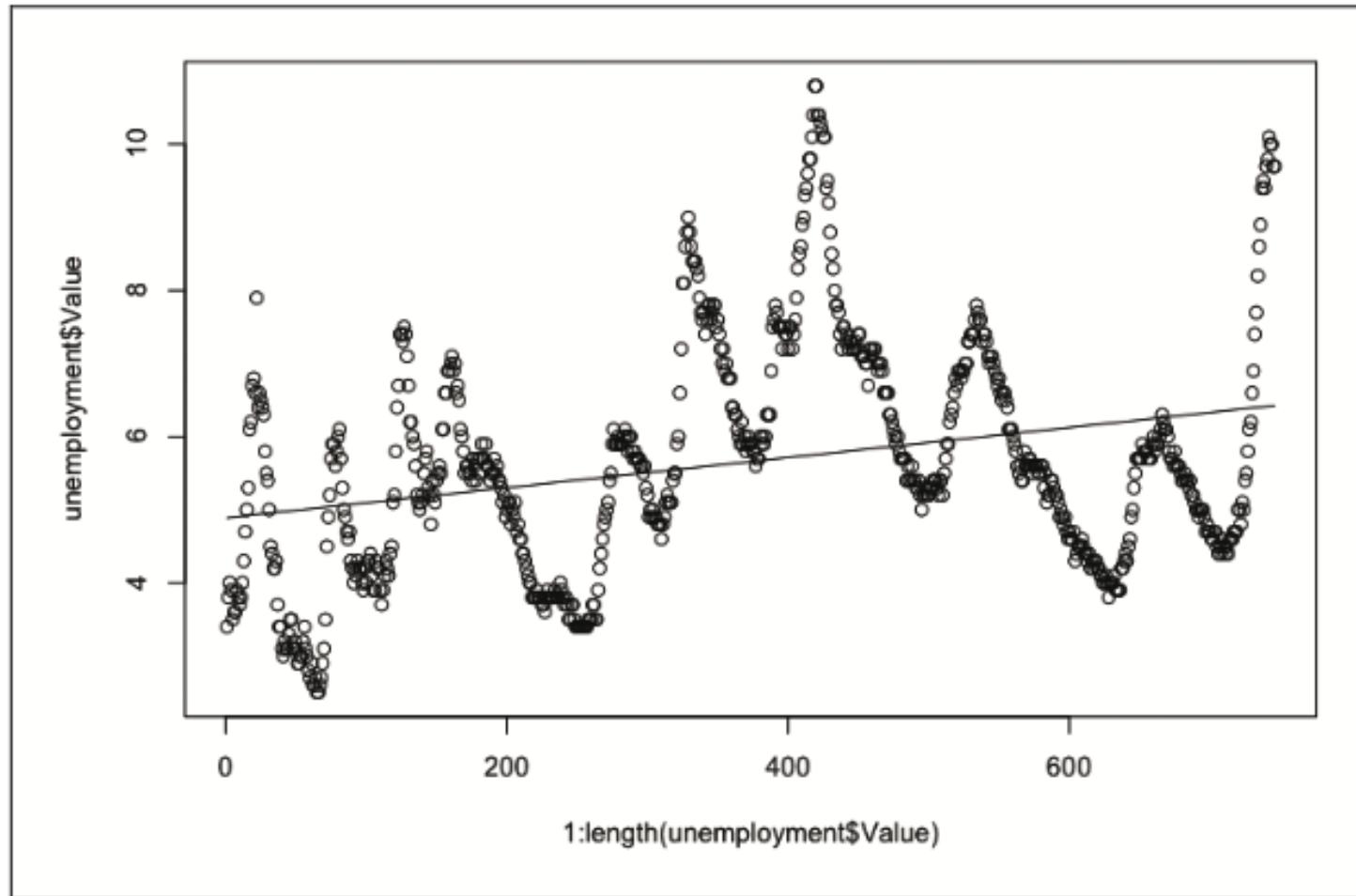
Echarts中的例子

一维数据可视化 – Catmull-Rom插值例



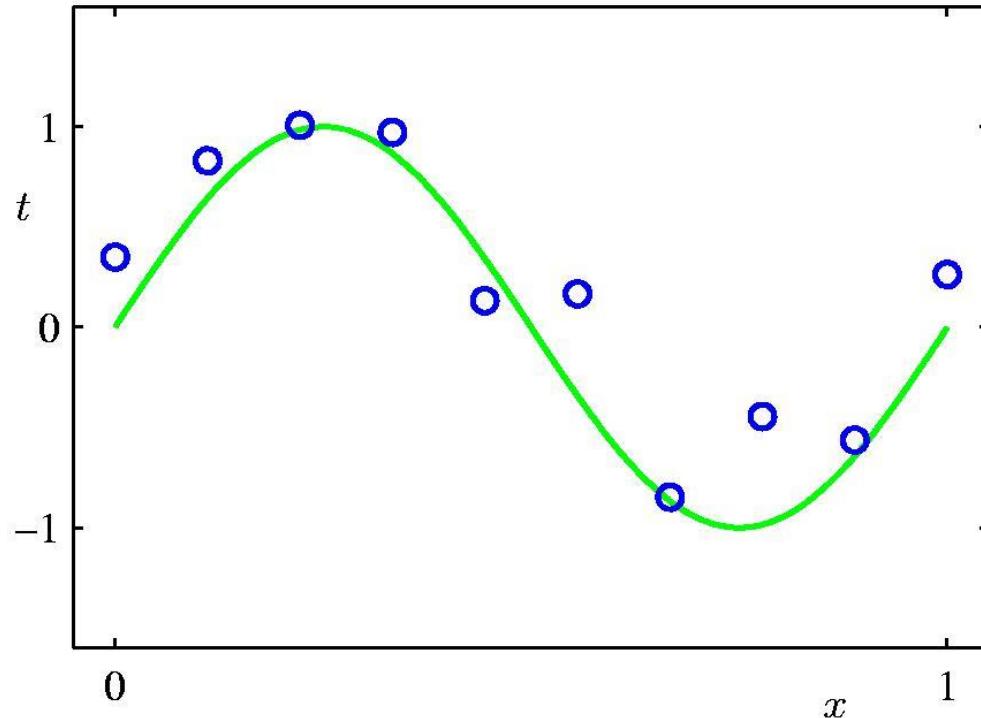
Echarts中的例子
<https://echarts.apache.org/zh/index.html>

一维数据可视化 - 拟合



失业率数据的散点图与直线拟合

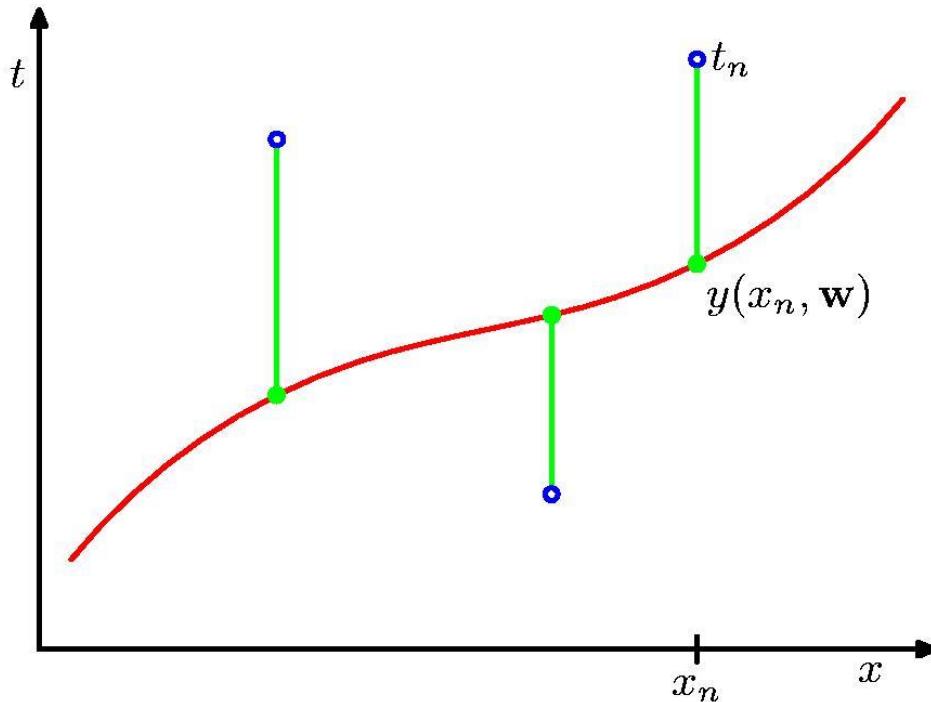
一维数据可视化 - 多项式拟合



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j$$

多项式函数与多项式基

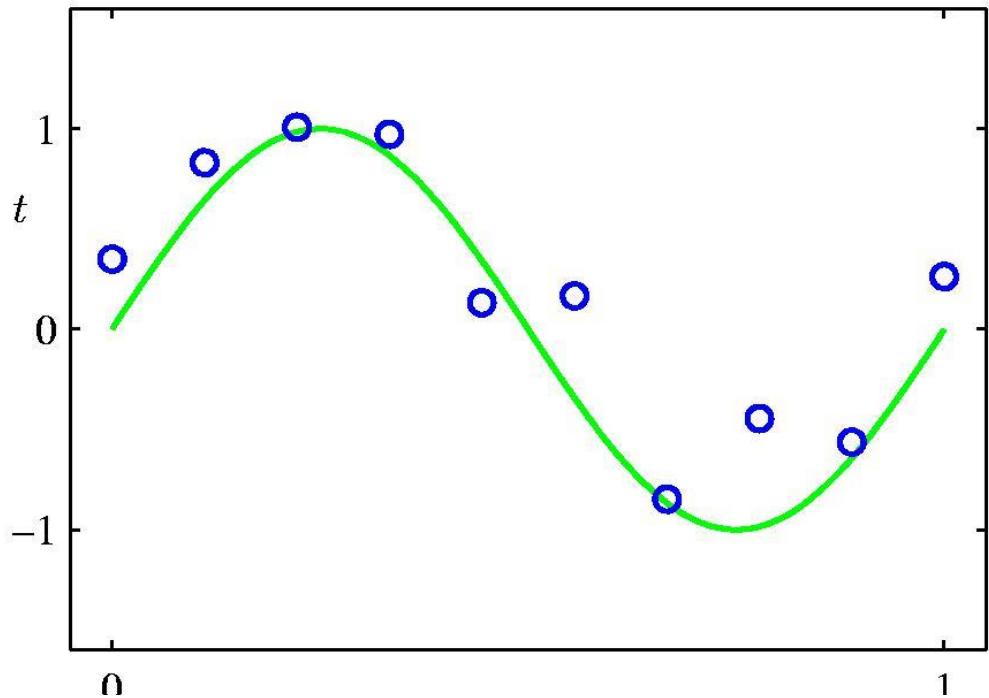
一维数据可视化 - 多项式拟合



$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$

平方和误差函数

一维数据可视化 - 多项式拟合



$$\mathbf{p}(x) = \begin{pmatrix} x^0 \\ x^1 \\ \vdots \\ x^M \end{pmatrix}$$

$$\mathbf{w} = \begin{pmatrix} w^0 \\ w^1 \\ \vdots \\ w^M \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y^0 \\ y^1 \\ \vdots \\ y^N \end{pmatrix}$$

$$y_j = \mathbf{p}(x_j) \cdot \mathbf{w}$$

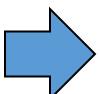
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j$$

$$\mathbf{y} = \mathbf{P}(\mathbf{x}) \cdot \mathbf{w}$$

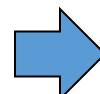
$$\arg \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{P}(\mathbf{x}) \cdot \mathbf{w}\|^2$$

$$\mathbf{w} = (\mathbf{P}^T \mathbf{P})^{-1} \mathbf{P}^T \mathbf{y}$$

Matrix form

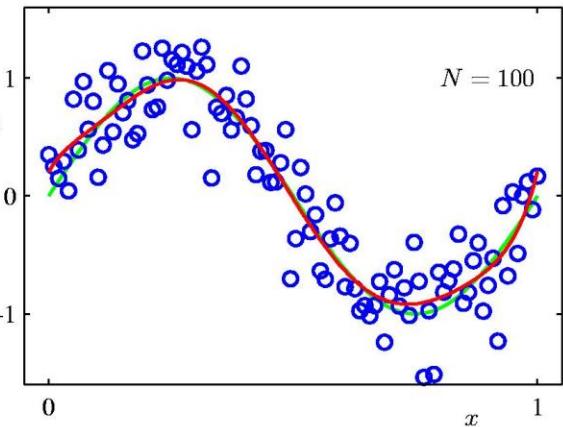
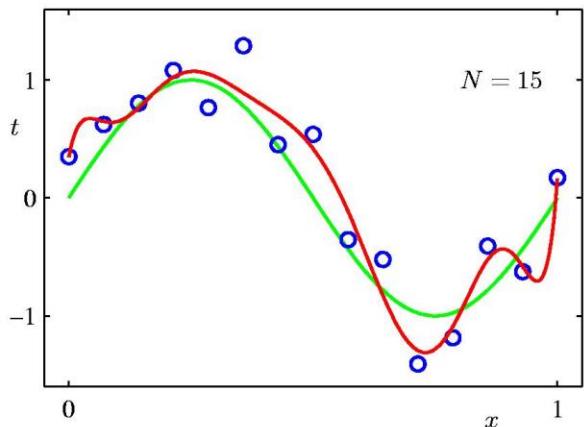
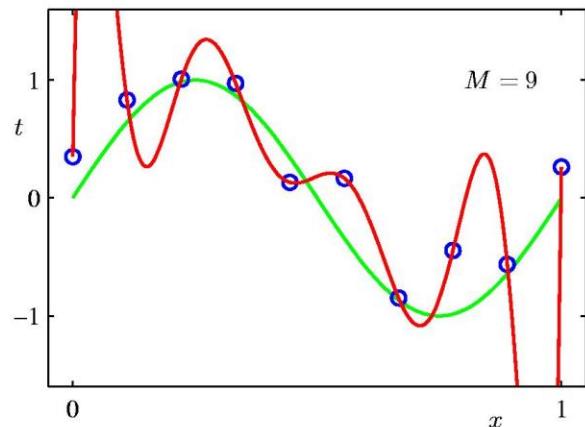
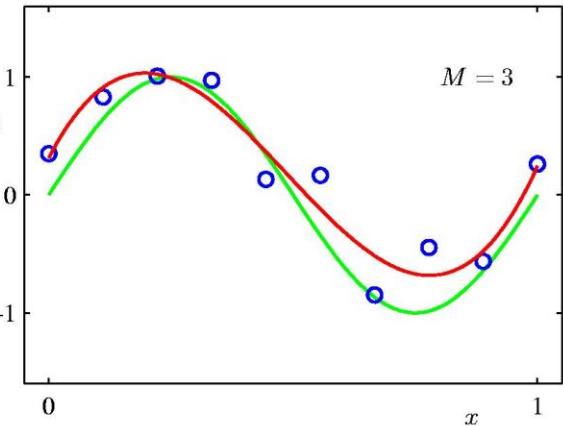
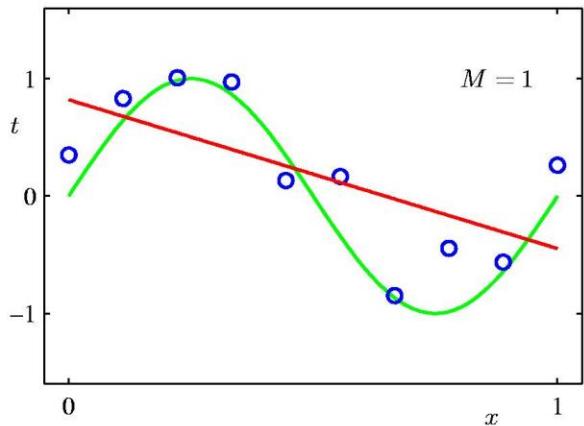
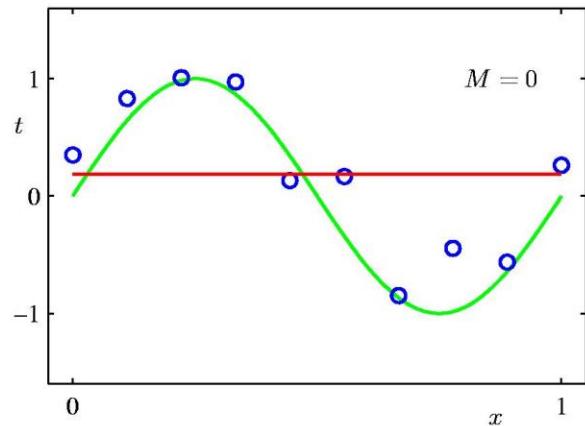


Least squares



Normal equation

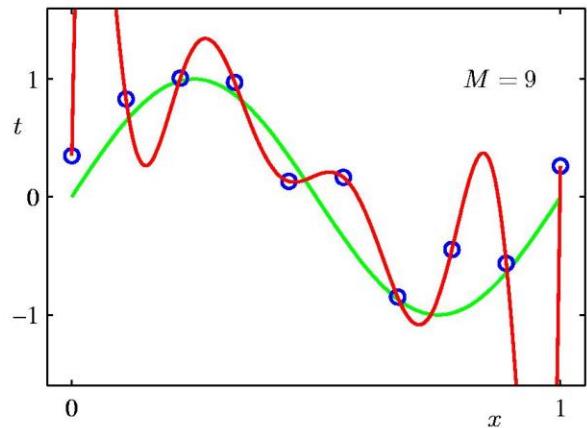
一维数据可视化 - 多项式拟合



$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$

平方和误差函数

一维数据可视化 - 过拟合现象

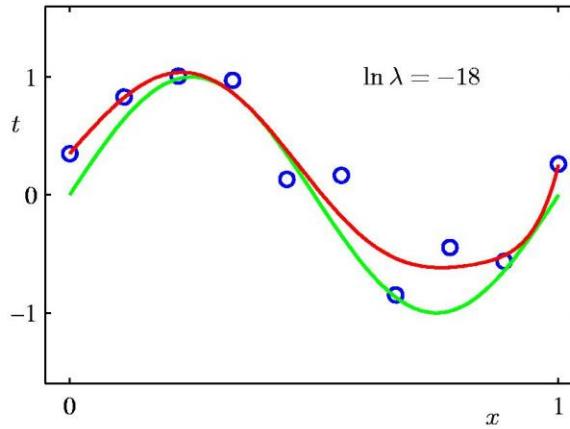
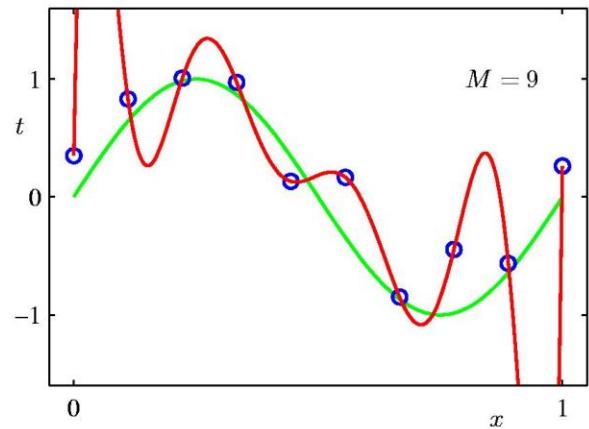


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

平方和误差函数

一维数据可视化 - 过拟合现象及其解法



$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 \quad \tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

平方和误差函数

正则函数校正



一维数据可视化 - LOESS拟合

Taylor & Francis Online

Home ▶ All Journals ▶ Journal of the American Statistical Association ▶ List of Issues ▶ Volume 74, Issue 368 ▶ Robust Locally Weighted Regression and S

Journal
Journal of the American Statistical Association ▶
Volume 74, 1979 - Issue 368

Enter keywords, authors, DOI, C

1,494 Views
4,571 CrossRef citations to date
9 Altmetric

Theory and Method
Robust Locally Weighted Regression and Smoothing Scatterplots
William S. Cleveland
Pages 829-836 | Received 01 Mar 1978, Published online: 05 Apr 2012
Download citation
References Citations Metrics Reprints & Permissions Get access

Abstract

Select Language ▾
Translator disclaimer

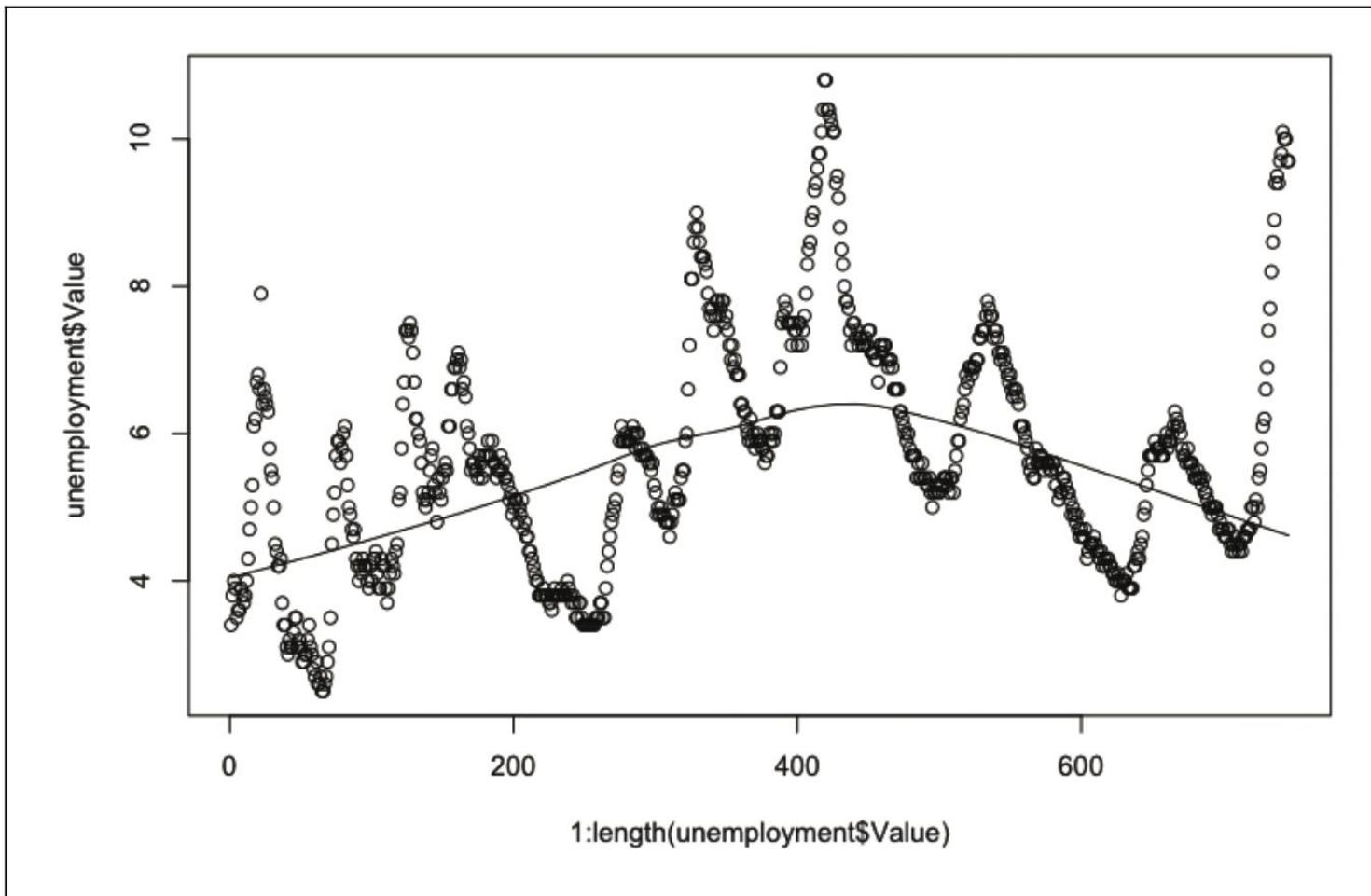
The visual information on a scatterplot can be greatly enhanced, with little additional cost, by computing and plotting smoothed points. Robust locally weighted regression is a method for smoothing a scatterplot, (x_i, y_i) , $i = 1, \dots, n$, in which the fitted value at x_k is the value of a polynomial fit to the data using weighted least squares, where the weight for (x_i, y_i) is large if x_i is close to x_k and small if it is not. A robust fitting procedure is used that guards against deviant points distorting the smoothed points. Visual, computational, and statistical issues of robust locally weighted regression are discussed. Several examples, including data on lead intoxication, are used to illustrate the methodology.

Key Words: [Graphics](#), [Scatterplots](#), [Nonparametric regression](#), [Smoothing](#), [Robust estimation](#)

William Cleveland

Robust Locally Weighted Regression and Smoothing Scatterplots
Journal of the American statistical association 74.368 (1979): 829-836.

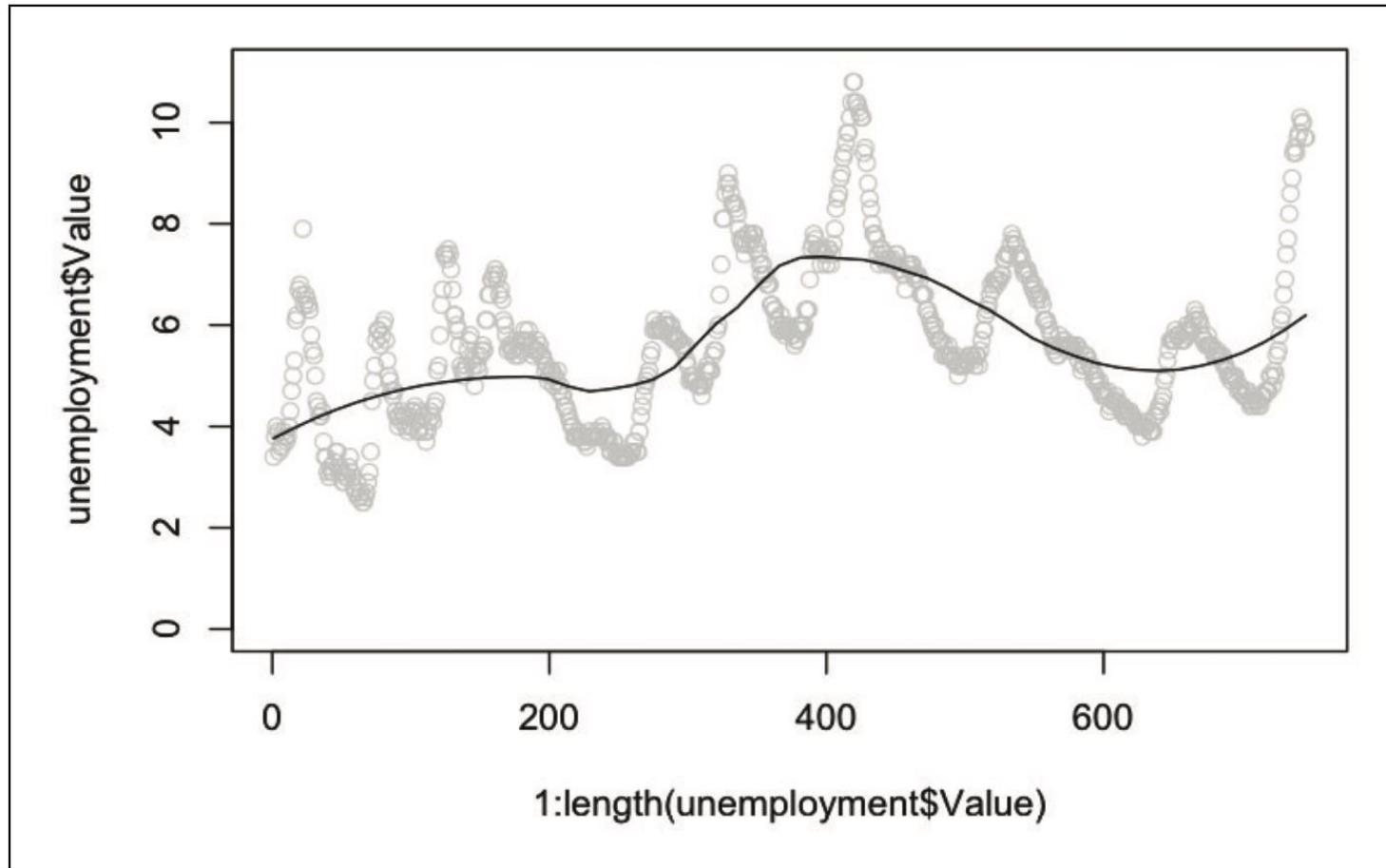
一维数据可视化 - LOESS拟合例



R语言实现：

```
scatter.smooth(x=1:length(data$Value), y=data$Value)
```

一维数据可视化 - LOESS拟合例



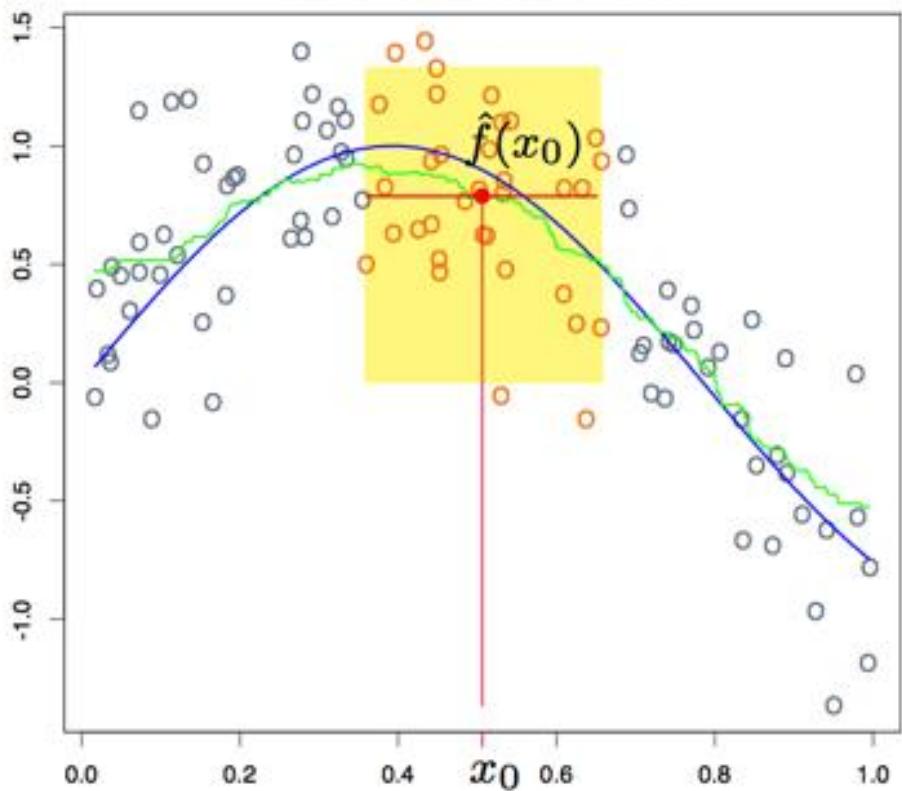
R语言实现 (更多参数调整) :

```
scatter.smooth(x=1:length(data$Value), y=data$Value,  
degree=2, span=0.5)
```

一维数据可视化 - LOESS拟合



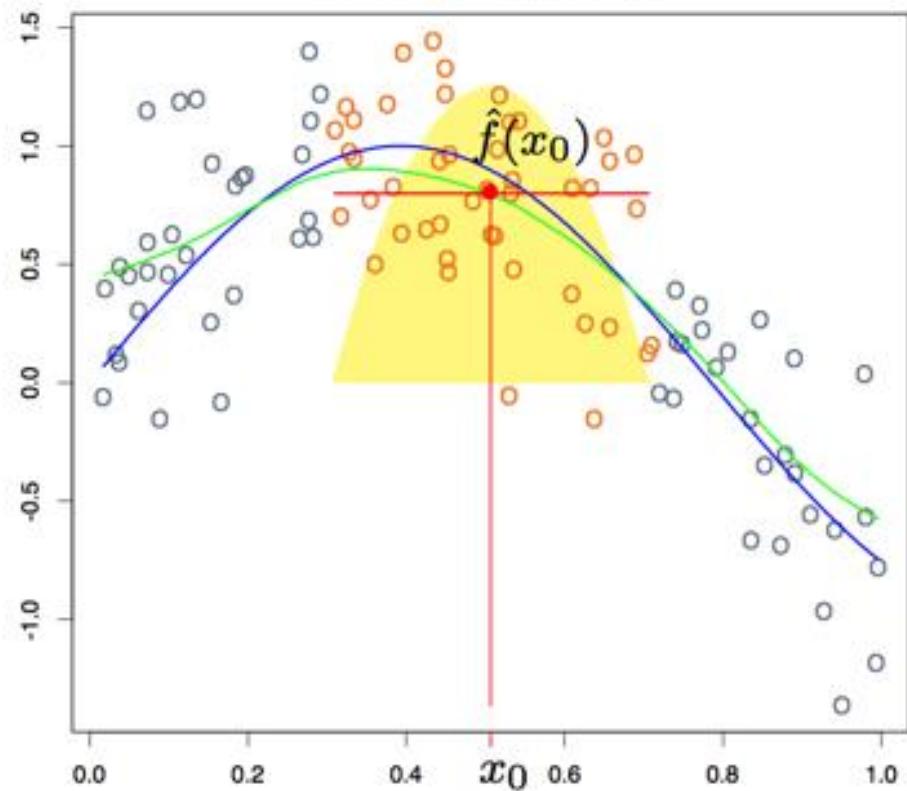
Nearest-Neighbor Kernel



K-NN平均回归造成结果不连续

$$\hat{f}(x) = \text{Ave}(y_i | x_i \in N_k(x))$$

Epanechnikov Kernel



Kernel方法带来了连续光滑的结果

$$\hat{f}(x_0) = \frac{\sum_{i=1}^N K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^N K_\lambda(x_0, x_i)}$$

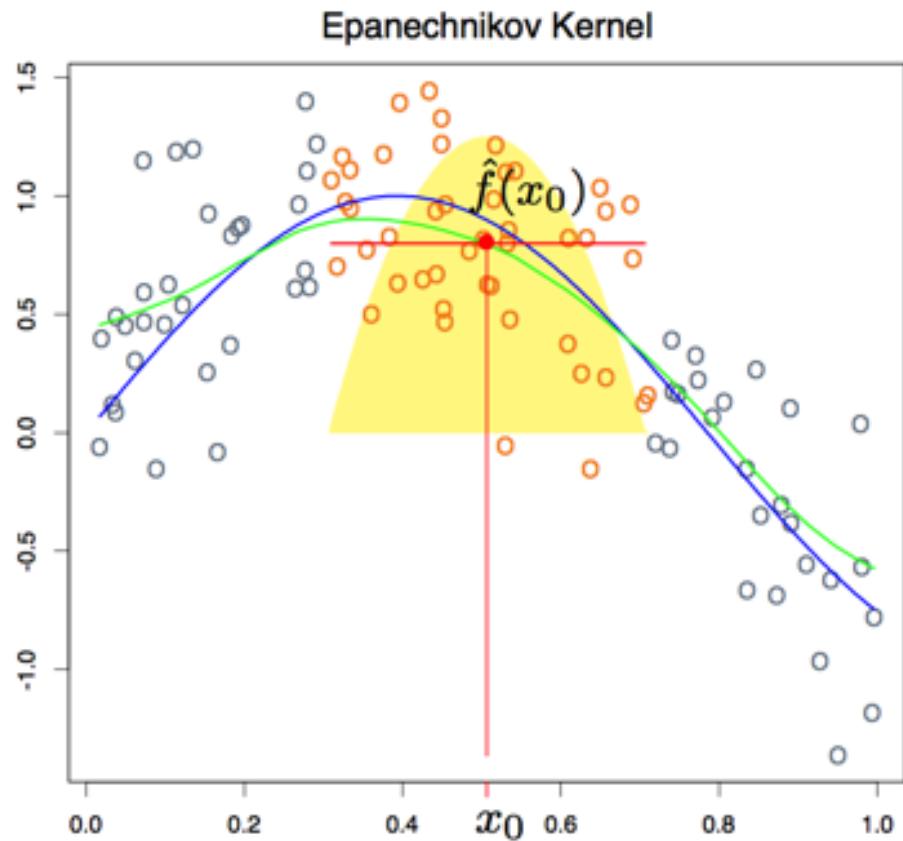
一维数据可视化 - LOESS拟合

$$\hat{f}(x_0) = \frac{\sum_{i=1}^N K_\lambda(x_0, x_i)y_i}{\sum_{i=1}^N K_\lambda(x_0, x_i)}$$

$$K_\lambda(x_0, x_i) = D\left(\frac{|x_0 - x_i|}{\lambda}\right)$$

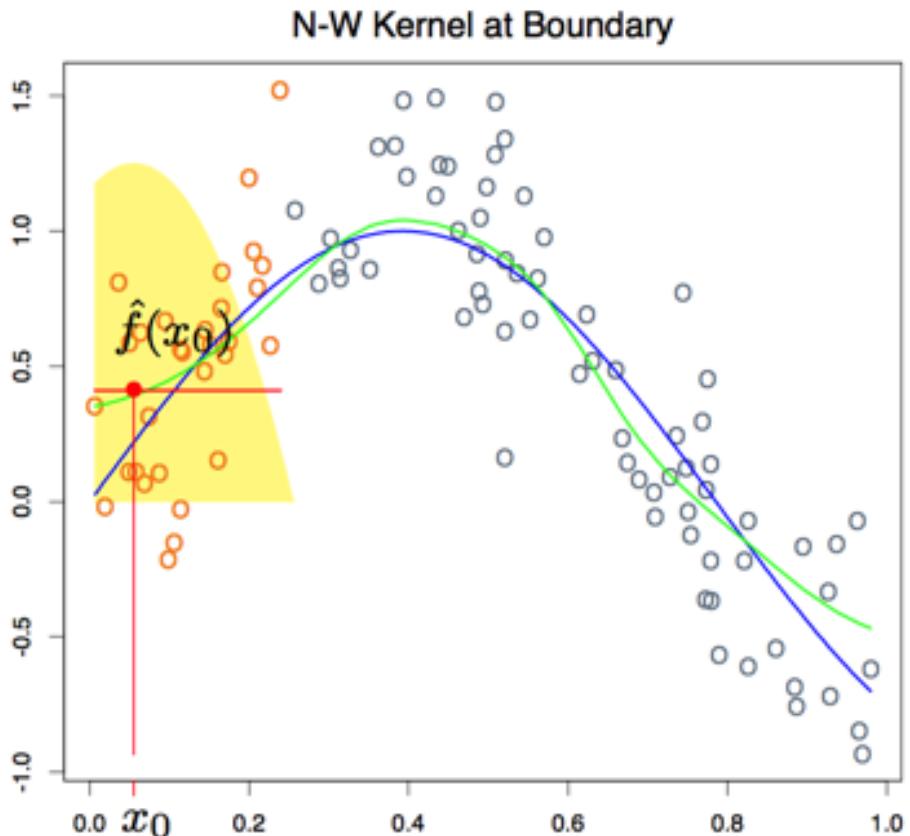
$$D(t) = \begin{cases} \frac{3}{4}(1-t^2) & \text{for } |t| < 1 \\ 0 & \text{otherwise} \end{cases}$$

Epanechnikov 二次kernel函数

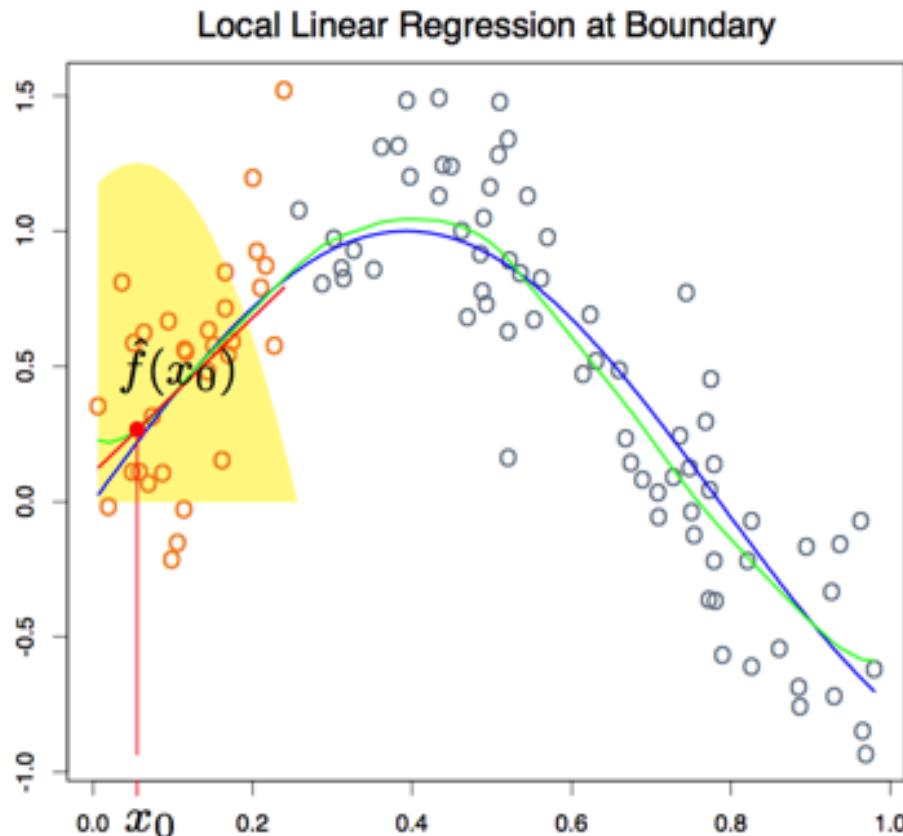


kernel方法带来了连续光滑的结果

一维数据可视化 - LOESS拟合



Kernel方法存在边界问题
开始与结束区段的点
其左右邻域是不对称的
导致了平滑后的值偏大或偏小



需要对权值做再修正

一维数据可视化 - LOESS拟合

$$\min_{\beta} \sum_{i=1}^N K_\lambda(x_0, x_i) [y_i - \sum_{j=0}^d \beta_j x_i^j]^2$$

$$B = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \cdots & x_N^d \end{pmatrix}$$

$$W_{x_0} = \begin{pmatrix} K_\lambda(x_0, x_1) & 0 & \cdots & 0 \\ 0 & K_\lambda(x_0, x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_\lambda(x_0, x_N) \end{pmatrix}$$

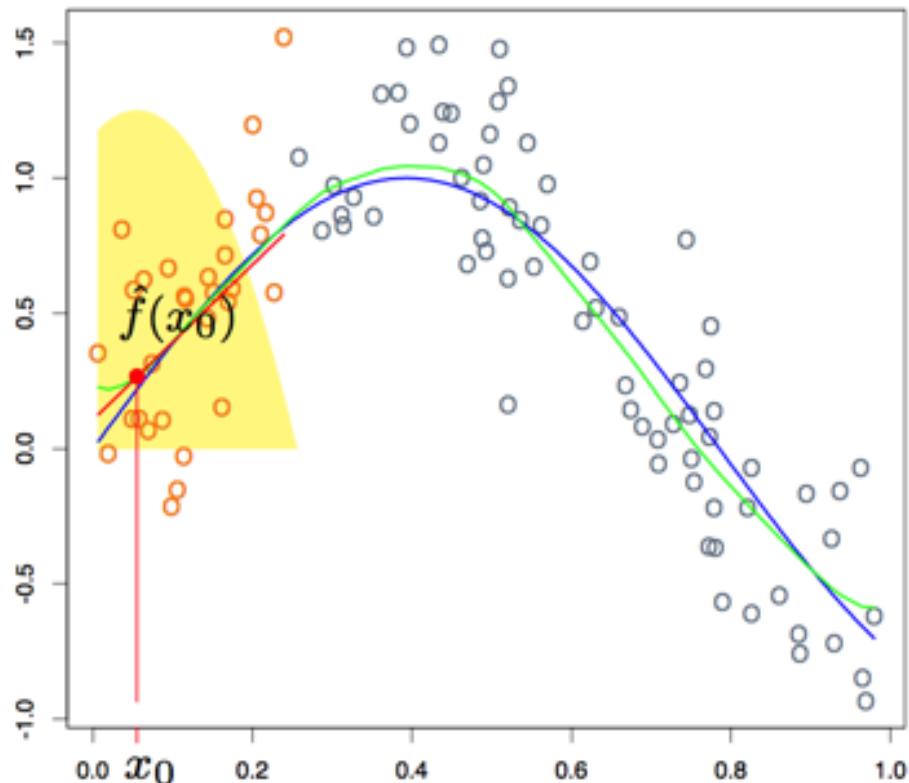
$$\Delta = (\beta_0, \beta_1, \dots, \beta_d)^T$$

$$Y = (y_1, y_2, \dots, y_N)^T$$



$$\min_{\Delta} (Y - B\Delta)^T W_{x_0} (Y - B\Delta)$$

Local Linear Regression at Boundary



需要对权值做再修正

$$\hat{f}(x_0) = \sum_{j=0}^d \hat{\beta}_j x_0^j$$

一维数据可视化 - LOESS拟合

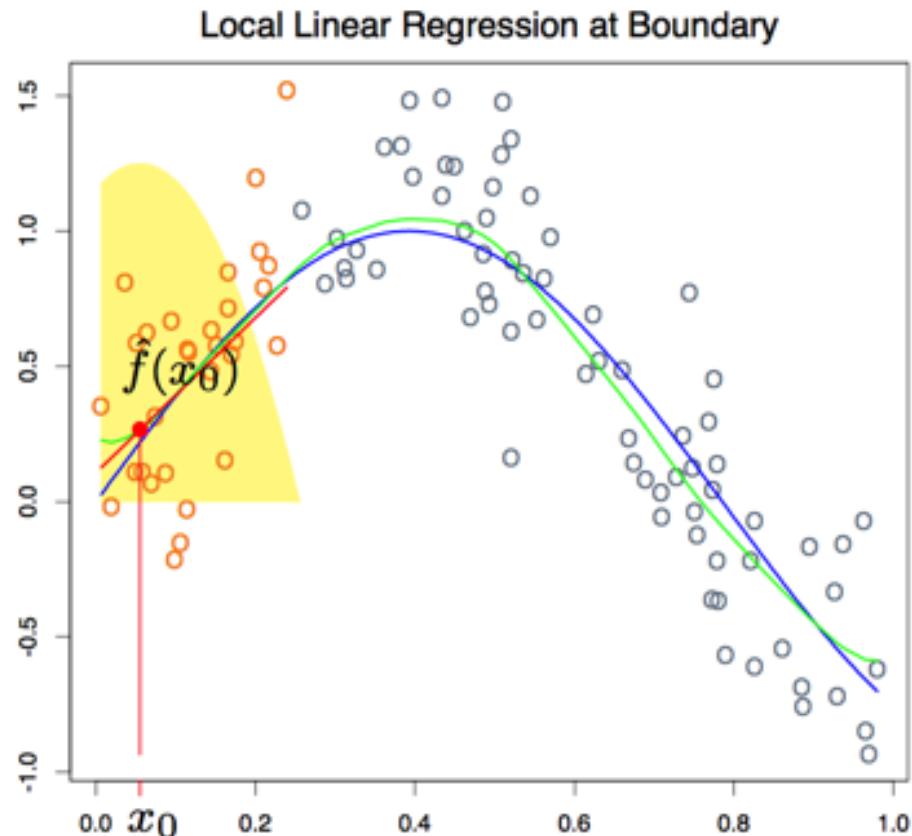


$$\Delta = (B^T W_{x_0} B)^{-1} (B^T W_{x_0} Y)$$

$$\begin{aligned} e(x_0) &= (1, x_0, \dots, x_0^d) \\ \hat{f}(x_0) &= e(x_0)(B^T W_{x_0} B)^{-1} (B^T W_{x_0} Y) \\ &= \sum_i w_i(x_0) y_i \end{aligned}$$



$$\min_{\Delta} (Y - B\Delta)^T W_{x_0} (Y - B\Delta)$$



$$\hat{f}(x_0) = \sum_{j=0}^d \beta_j x_0^j$$

一维数据可视化 - Robust LOESS拟合

计算残差

$$e_i = y_i - \hat{f}(x_i)$$

根据残差计算Robust Weight

$$\delta_i = B(e_i/6s)$$

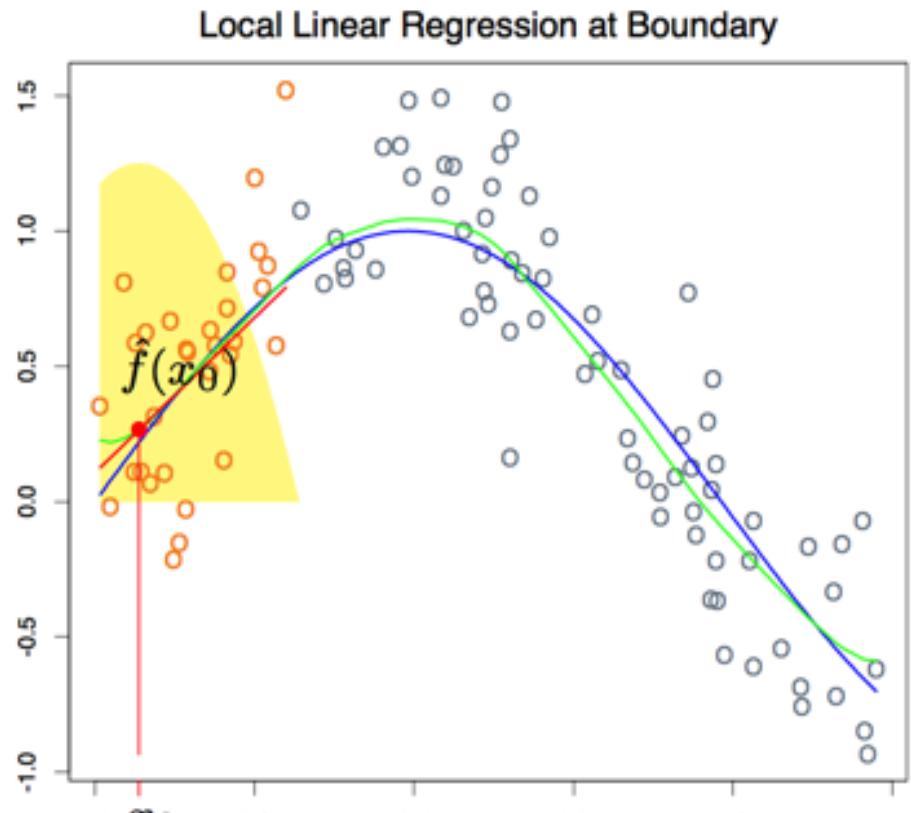
其中

s 为残差绝对值序列 $|e_i|$ 的中位数

B 函数为bisquare函数

$$B(u) = \begin{cases} (1 - u^2)^2 & \text{for } 0 \leq u < 1 \\ 0 & \text{for } u \geq 1 \end{cases}$$

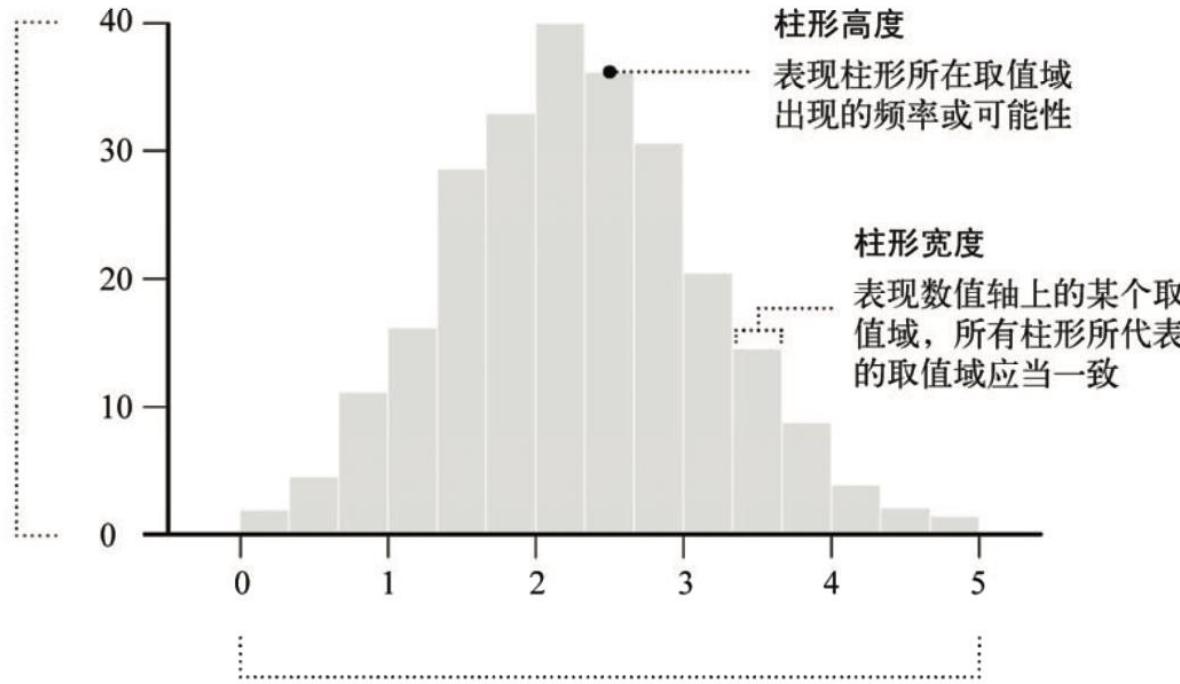
用robustness weight乘以kernel weight作为新的weight



剔除残差较大的异常点
对于回归的影响

一维数据可视化 - 直方图

频率或可能性
根据群体中给定的比例值进行尺度标识



柱形高度
表现柱形所在取值域
出现的频率或可能性

柱形宽度
表现数值轴上的某个取
值域，所有柱形所代表
的取值域应当一致

数值轴
表现某一个延续
性变量的数值

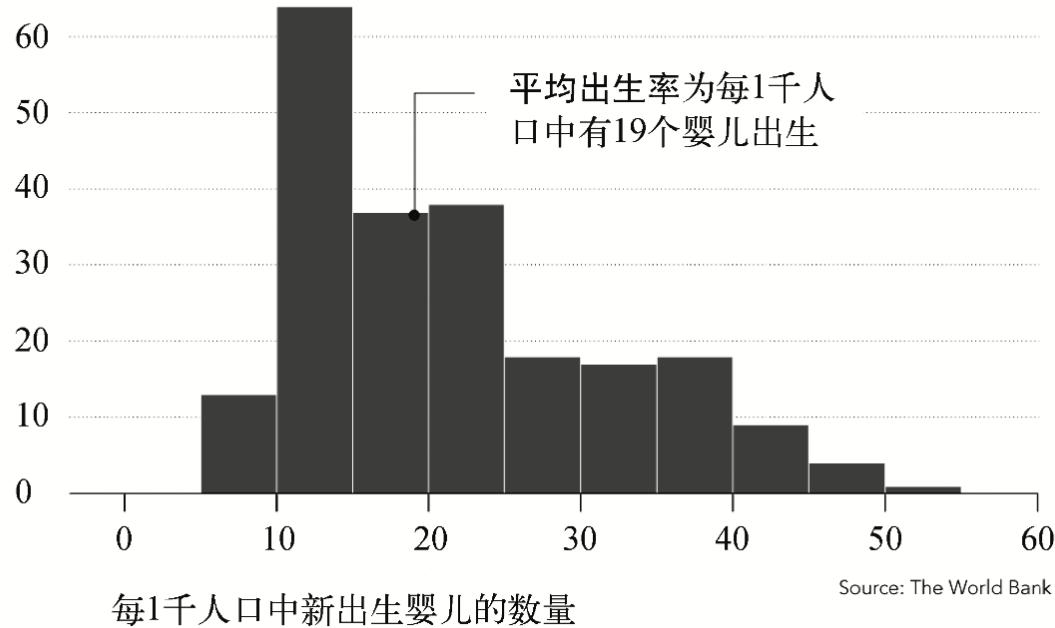
直方图的基本框架

一维数据可视化 - 直方图例

全球出生率分布图

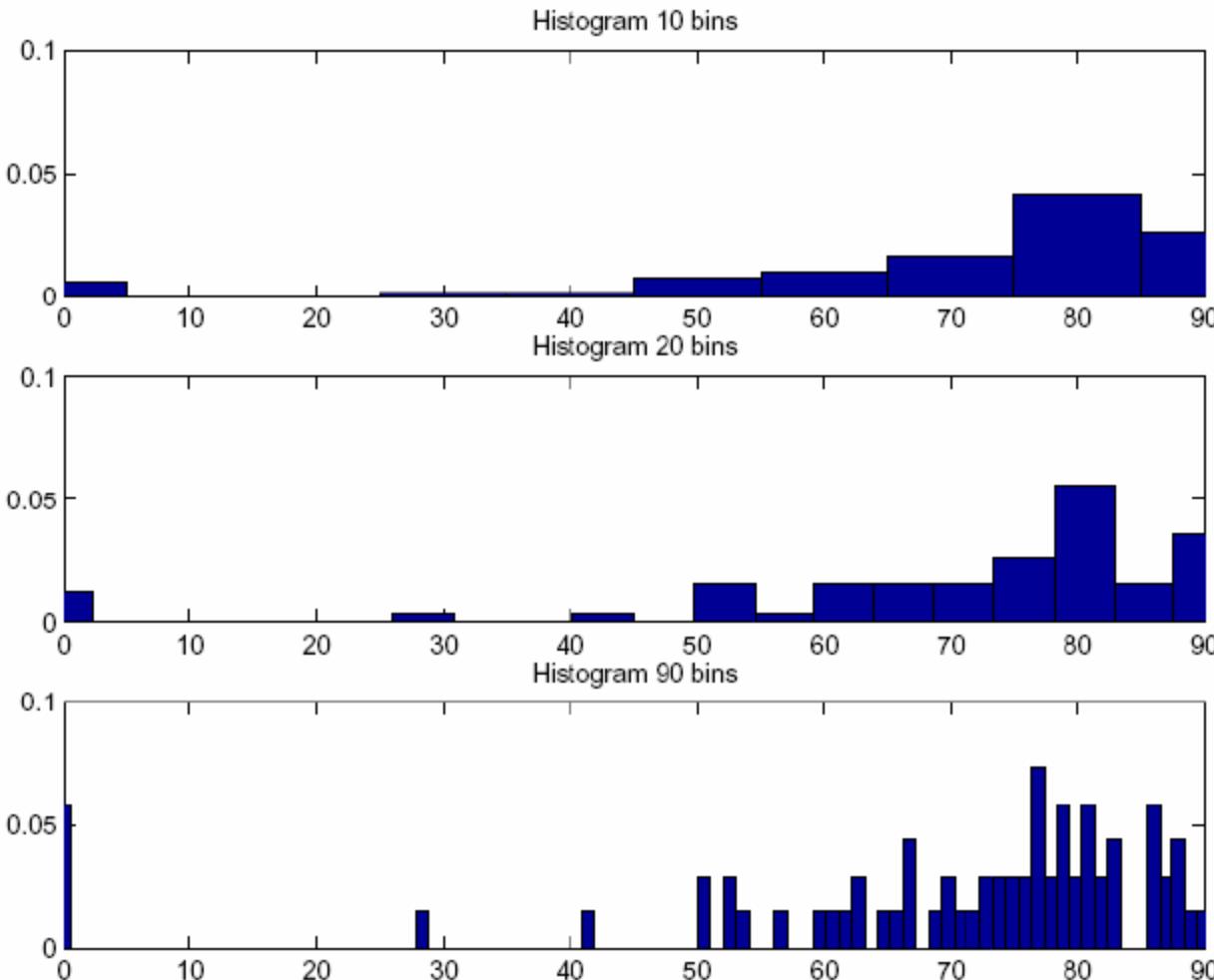
2008年，绝大多数国家每1千人口中新出生婴儿的数量都少于25。但仍然还有很多发展中国家的女性倾向于要更多的孩子。

国家数量



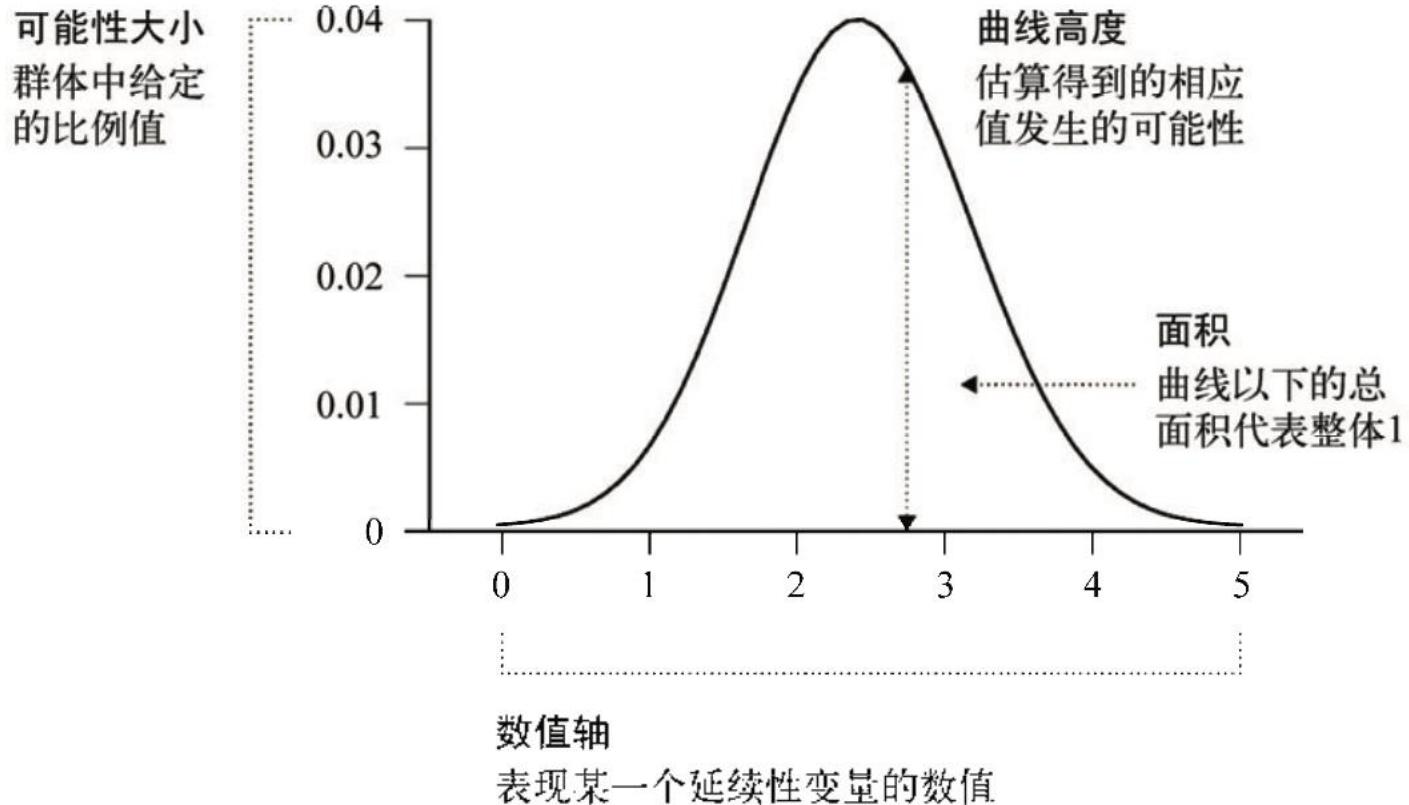
全球出生率分布图

一维数据可视化 - 直方图的缺点



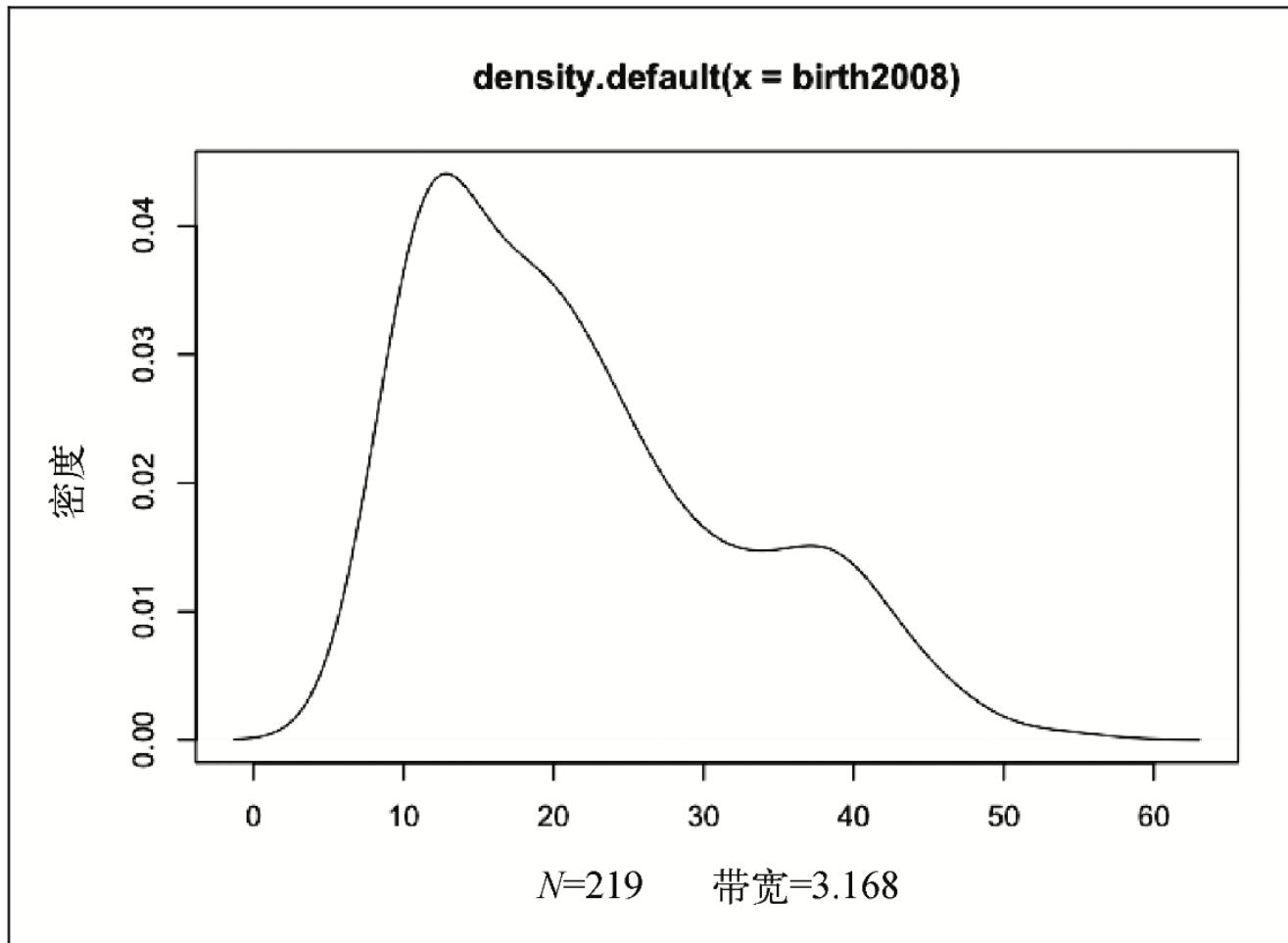
不同采样率下的直方图

一维数据可视化 - 密度图



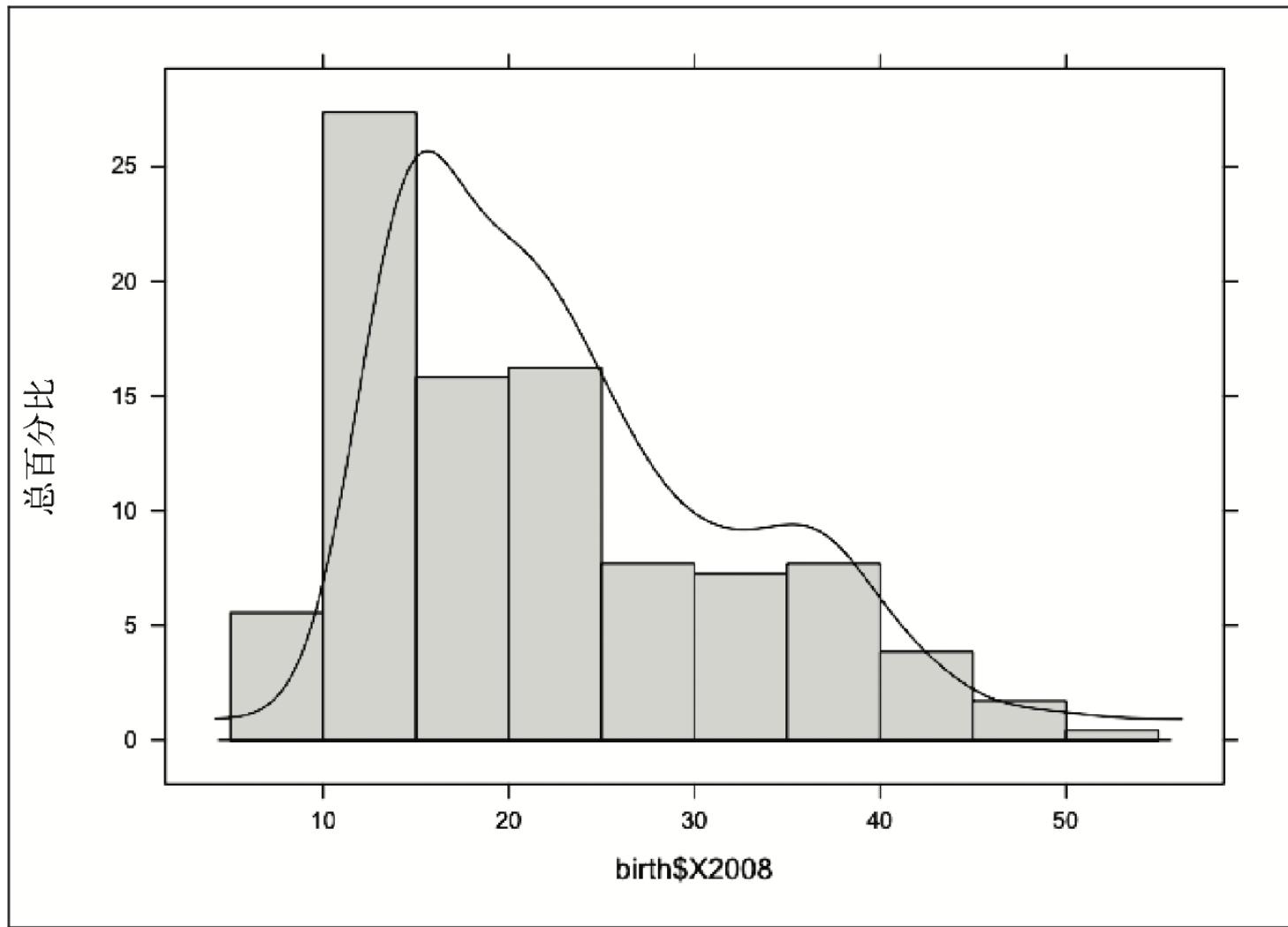
密度图的基本框架

一维数据可视化 - 密度图例



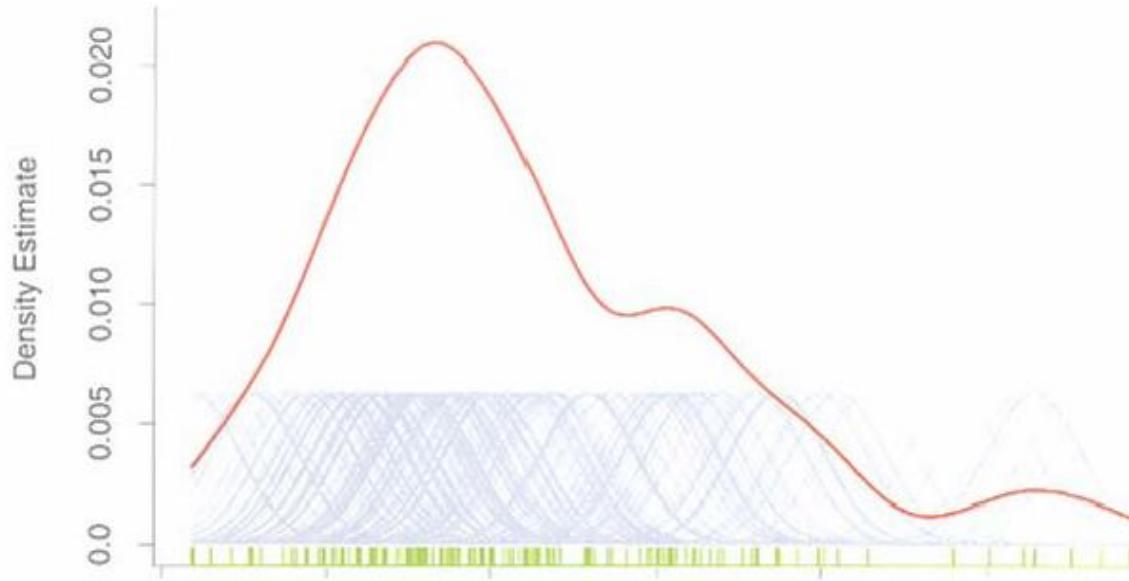
全球出生率的密度图

一维数据可视化 - 图例



全球出生率的直方图与密度图混合

Parzen example



from Hastie et al.

密度图与非参数估计



一维数据可视化 - 核密度估计

■ Multivariate kernel density estimation

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

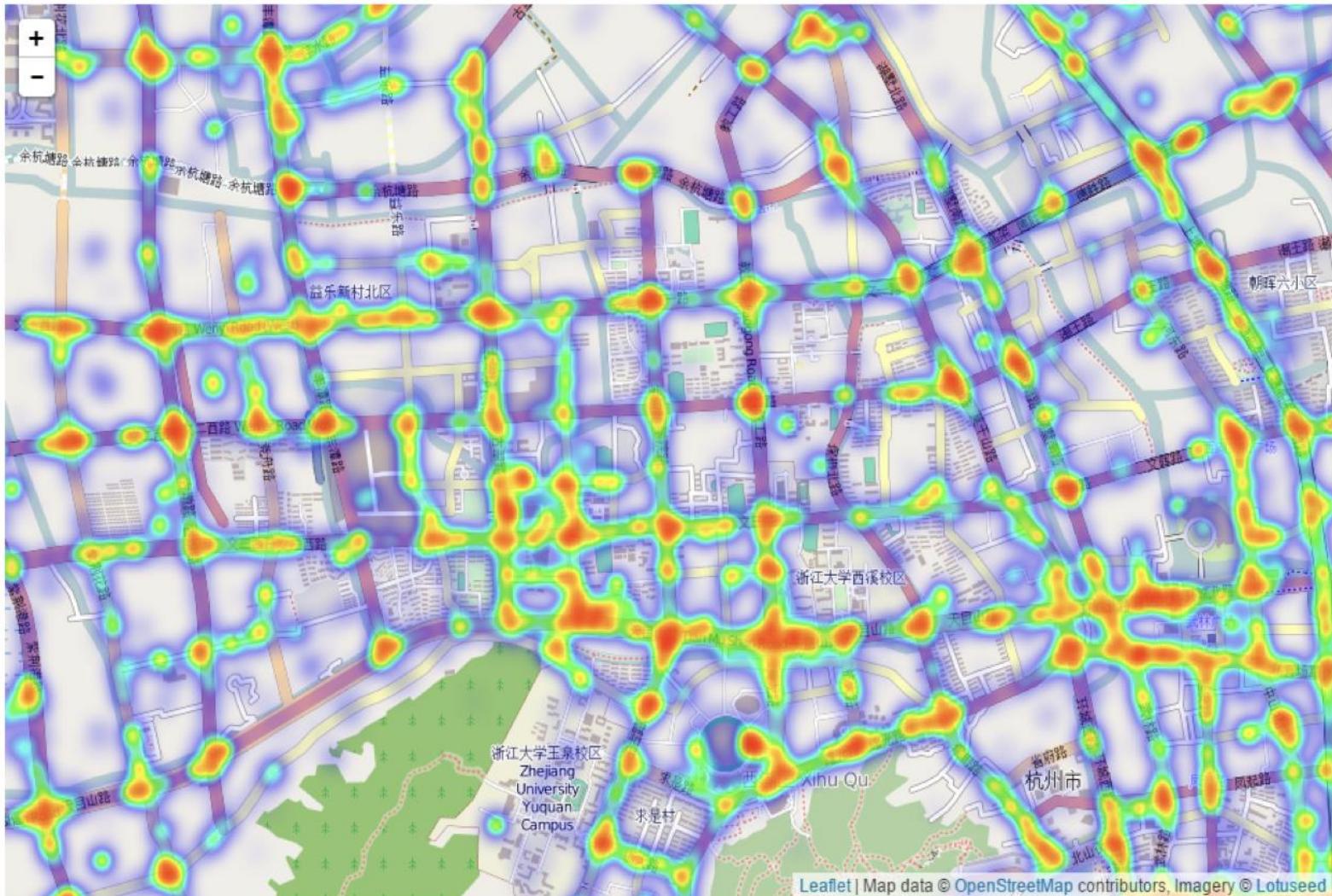
■ Kernels

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

- Epanechnikov

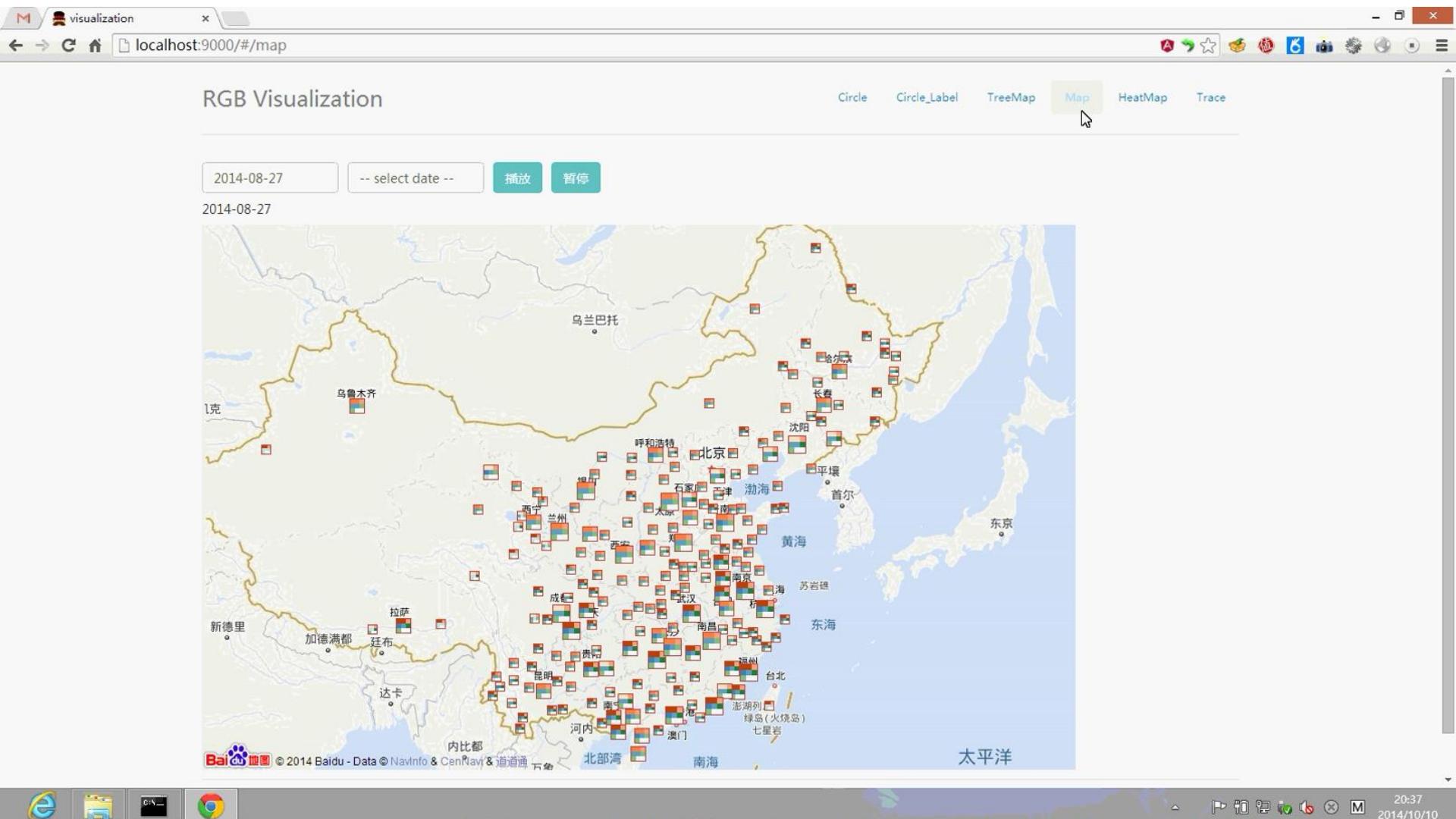
$$K_E = \begin{cases} 1/2c_d^{-1}(d+2)(1-\|\mathbf{x}\|^2) & \text{if } \|\mathbf{x}\| < 1 \\ 0 & \text{otherwise} \end{cases}$$

手机日志数据分析 - 热力图





手机日志数据分析 - 热力图





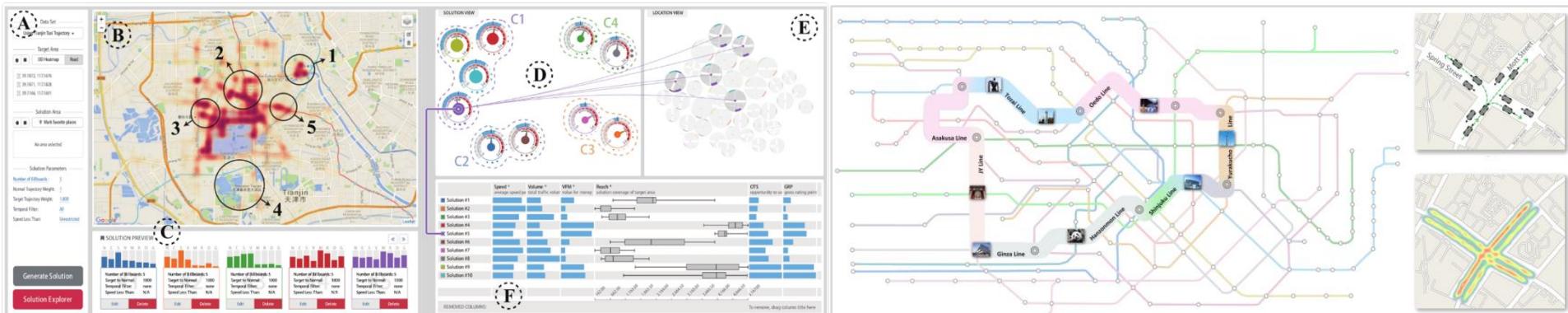
主要参考资料 《大数据可视化建模》

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/>