可视化分析

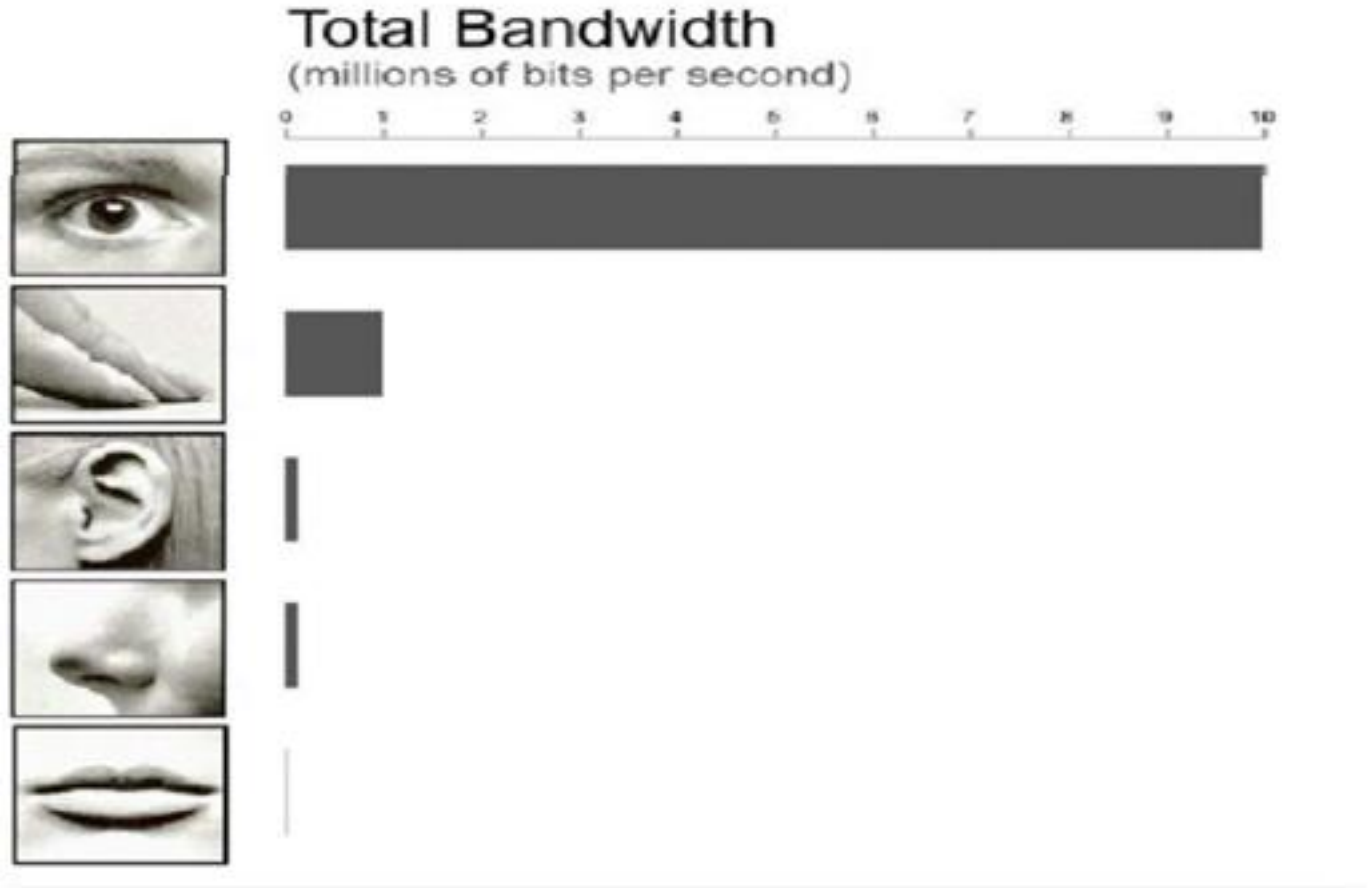
什么是数据可视化？

- **数据可视化**是利用计算机图形学和图像处理技术，将数据转换为图形或者图像在屏幕上显示出来进行交互处理的理论方法和技术。
- 数据可视化是一种能很好展示数据、处理数据的方法。

- 数据可视化不仅是一种工具和技术，同时作是一种表达数据的方式，它是对现实世界的抽象表达。
- 数据可视化像文字一样，为我们讲述各种各样的故事。

为什么要进行数据可视化？

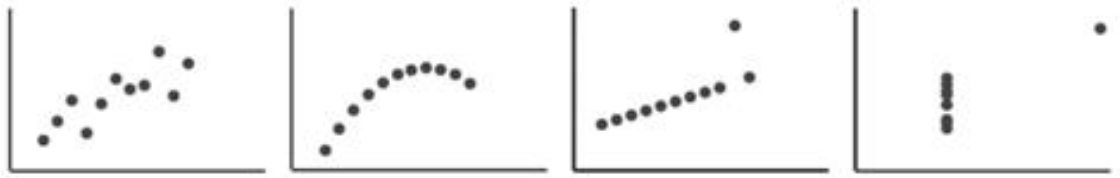
1、我们利用视觉获取的信息量，远远比别的感官要多得多。



为什么要进行数据可视化？

2、数据可视化能够帮助我们对数据有更加全面的认识。

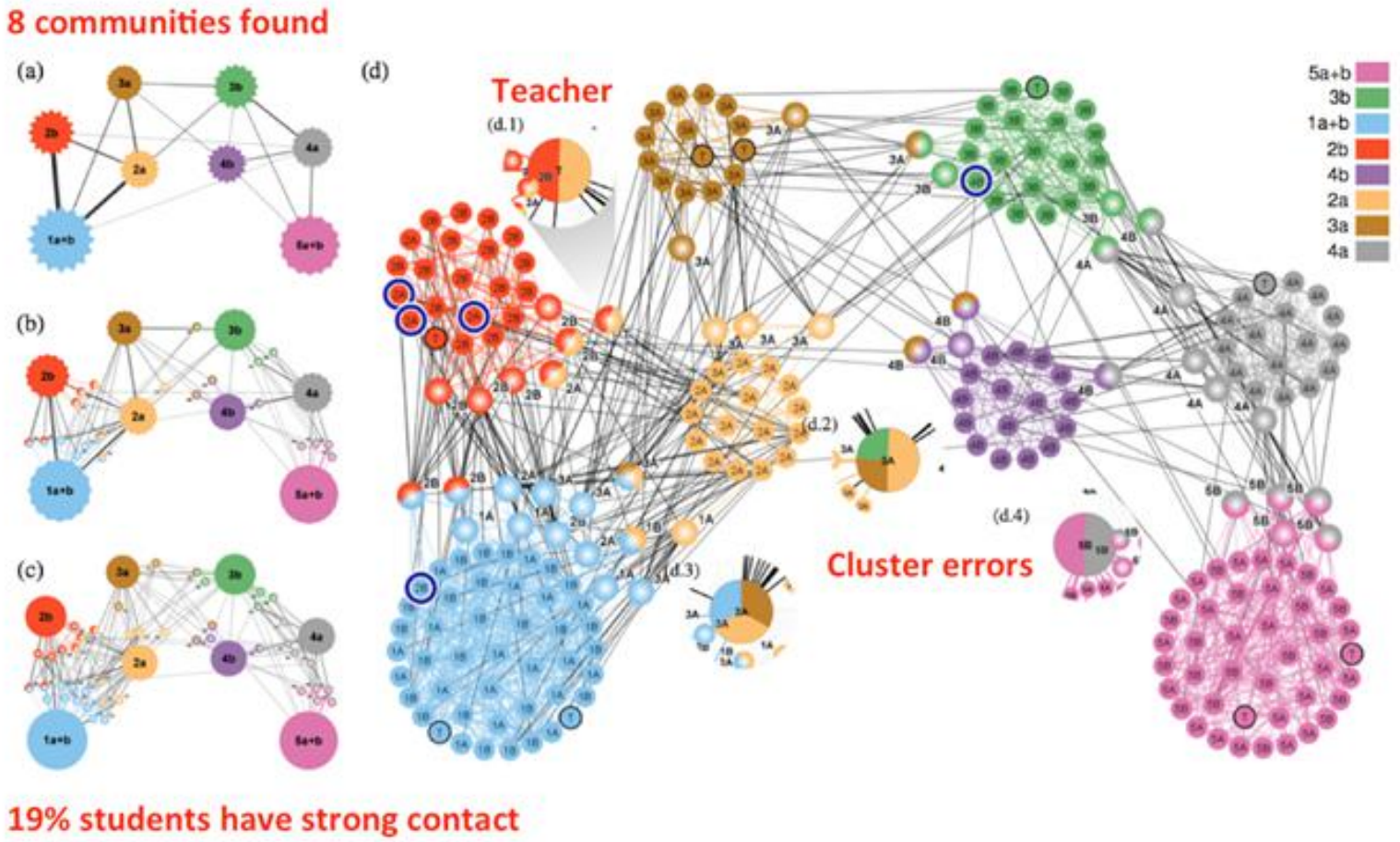
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	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



Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: y = 3.00 + 0.500x

为什么要进行数据可视化？

3、数据可视化能够在小空间中展示大规模数据。

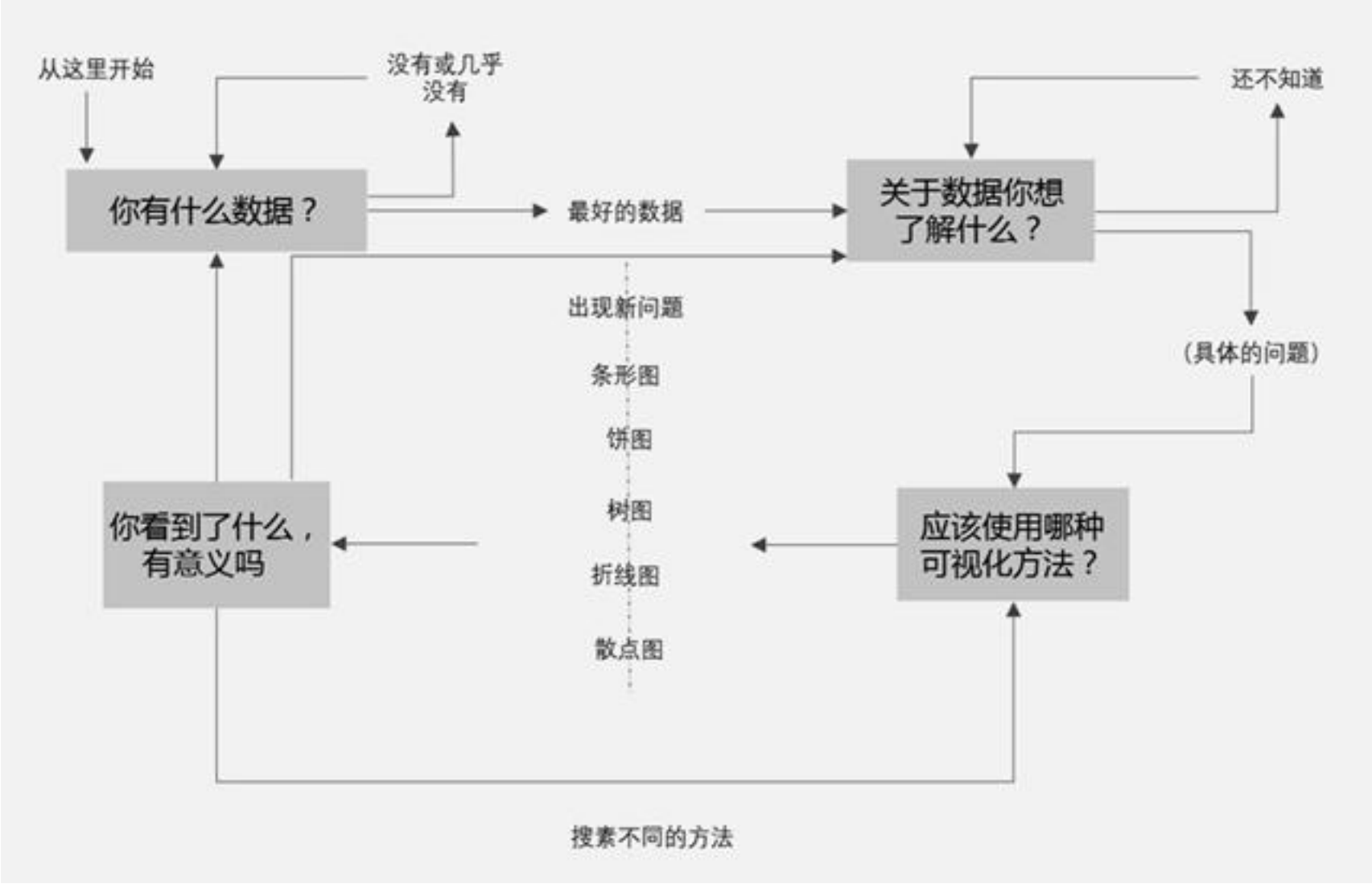


为什么要进行数据可视化？

4、人类大脑在记忆能力的限制。

人类的记忆能力是有限的，我们不可能记住所有的数据，单纯地记忆数据特诊对我们来说也是不小的挑战。俗话说的好，百闻不如一见。如果能够将数据总结到一张图表中，我们通过图像记忆，能更好地帮助我们记忆。

数据可视化过程


从这里开始
没有或几乎
没有
还不知道
你有什么数据？
最好的数据
关于数据你想
了解什么？
出现新问题
（具体的问题）
条形图
饼图
树图
你看到了什么，
有意义吗
应该使用哪种
可视化方法？
折线图
散点图
搜索不同的方法

可视化分析

- 可视化分析是一种数据分析方法，利用人类的形象思维将数据关联，并映射为形象的图表。人脑对于视觉信息的处理要比文本信息容易得多，所以可视化图表能够使用户更好地理解信息，可视化分析凭借其直观清晰，能够提供新洞察和发现机会的特点活跃在诸多科学领域

可视化分析的作用

- 在数据分析中，通过绘制图表更容易找到数据中的模式。传统的数据分析方法存在一些局限性，需要借助于分析师丰富的分析经验。可视化分析方法将数据以图像的方式展现，提供友好的交互，还可以提供额外的记忆帮助，对于将要分析的问题，无需事先假设或猜想，可以自动从数据中挖掘出更多的隐含信息
- 在机器学习领域，缺失数据、过度训练、过度调优等都会影响模型的建立，可视化分析可以帮助解决其中一些问题
- 可视化分析在机器学习的数据预处理、模型选择、参数调优等阶段也同样发挥重要作用。在数据建模的过程中，容易辨别出数据的分布、异常、参数取值对模型性能的影响等

可视化分析的作用

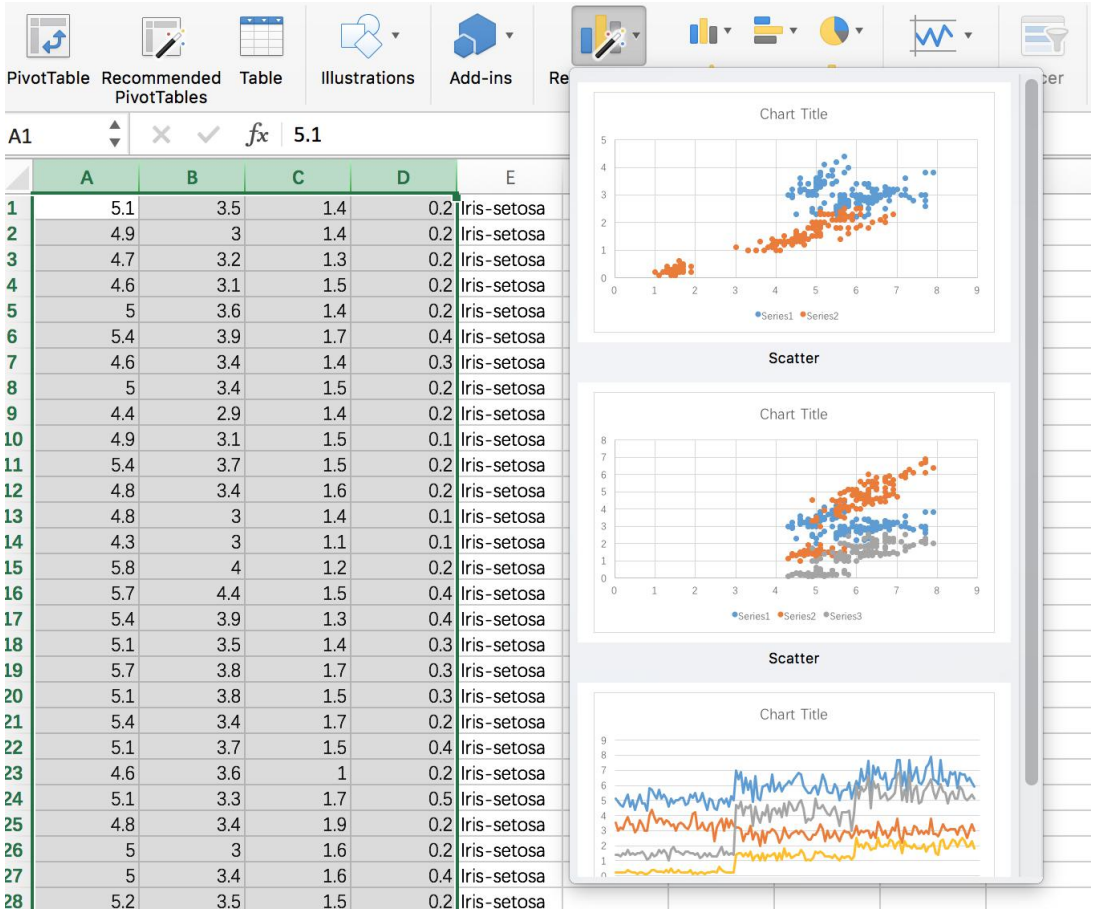
- 在分析结果展示时，通过建立可视化仪表板，组合多幅可视化图表，从不同的角度来描述信息，全方位展示分析结论
- 除了辅助数据分析之外，可视化分析为看似冰冷的数据带来更多趣味性，直观清晰的表达拥有更多的受众。在信息传播领域，可视化结果的独特风格（颜色、线条、轴线、尺寸等）不仅将有用的信息展示出来，更像是种精美的艺术品，让数据展示也变得更加富有情感

可视化分析方法

- 为了获得易于理解的可视化结果，人机交互很重要。可视化分析的常用方法大致可以划分为三个层次：领域方法、基础方法以及方法论基础
- 领城方法：根据数据的来源领域以及数据的性质进行可视化，包括地理信息可视化、空间数据可视化、文本数据可视化、跨媒体数据可视化、实时数据可视化等
- 可视化基础方法：包括统计图表、视觉隐喻。常见的统计图表有柱状图、折线图、饼图、箱图、散点图、韦恩图、气泡图、雷达图、热地图、等值线等，不同的统计图表有各自的适用场合
- 可视化分析的方法论基础是视觉编码，视觉编码是指受众对于接收到的视觉刺激进行编码，所以视觉编码的关键在于使用符合目标用户人群视觉感知习惯的表达方法，鉴于视觉感知习惯往往与一个人的知识、经验、心理等多种特异性的因素相关，而且视觉感知是一种视觉信息直接映射与信息提取、转换、存储、处理、理解等后续活动结合而成的过程

可视化分析常用工具

- Excel
- Tableau
- Raw
- Chart.js
- Processing
- Wordle
- Orange
- Facets
- Python、R语言库：
  - matplotlib、Seaborn、Pyecharts、ggplosts



3、数据可视化

下面从最常用和实用的维度总结了五种数据可视化方法：

一、面积&尺寸可视化

对同一类图形（例如柱状、圆环和蜘蛛图等）的长度、高度或面积加以区别，来清晰的表达不同指标对应的指标值之间的对比。这种方法会让浏览者对数据及其之间的对比一目了然。制作这类数据可视化图形时，要用数学公式计算，来表达准确的尺度和比例。

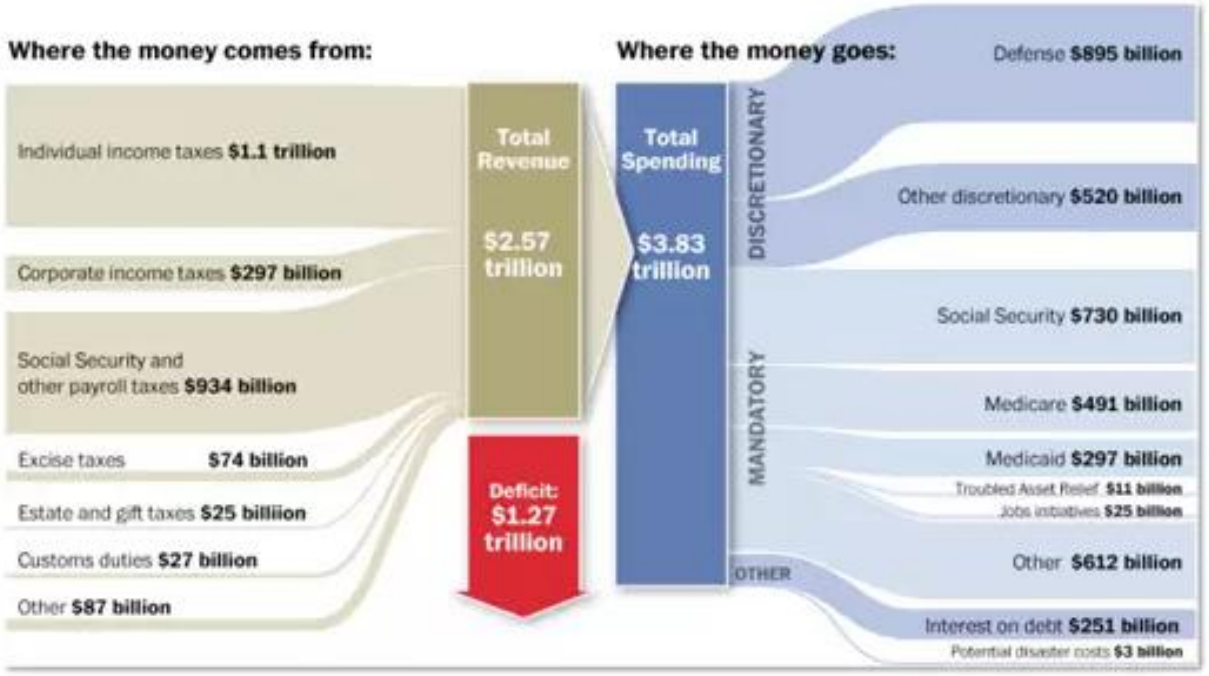
a: 天猫的店铺动态评分

天猫店铺动态评分模块右侧的条状图按精确的比例清晰的表达了不同评分用户的占比。从下图中我们第一眼就可以强烈的感知到5分动态评分的用户占绝对的比例。

一、面积&尺寸可视化

b: 联邦预算图

如下图，在美国联邦预算剖面图里，用不同高度的货币流清晰的表达了资金的来源去向，及每一项所占金额的比重。



一、面积&尺寸可视化

c: 公司黄页-企业能力模型蜘蛛图

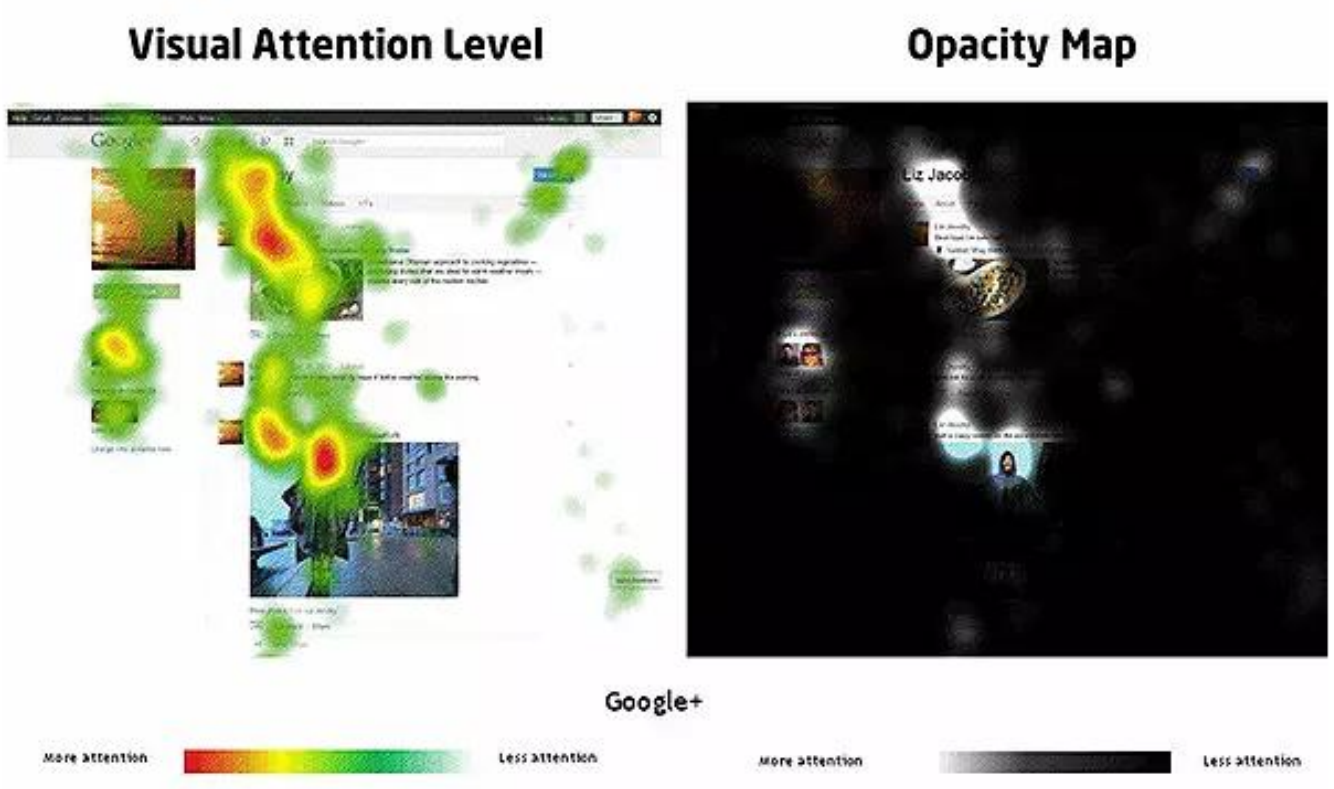
如下图，通过蜘蛛图的表现，公司综合实力与同行平均水平的对比便一目了然。

二、颜色可视化

通过颜色的深浅来表达指标值的强弱和大小，是数据可视化设计的常用方法，用户一眼看上去便可整体的看出哪一部分指标的数据值更突出。

a: 点击频次热力图

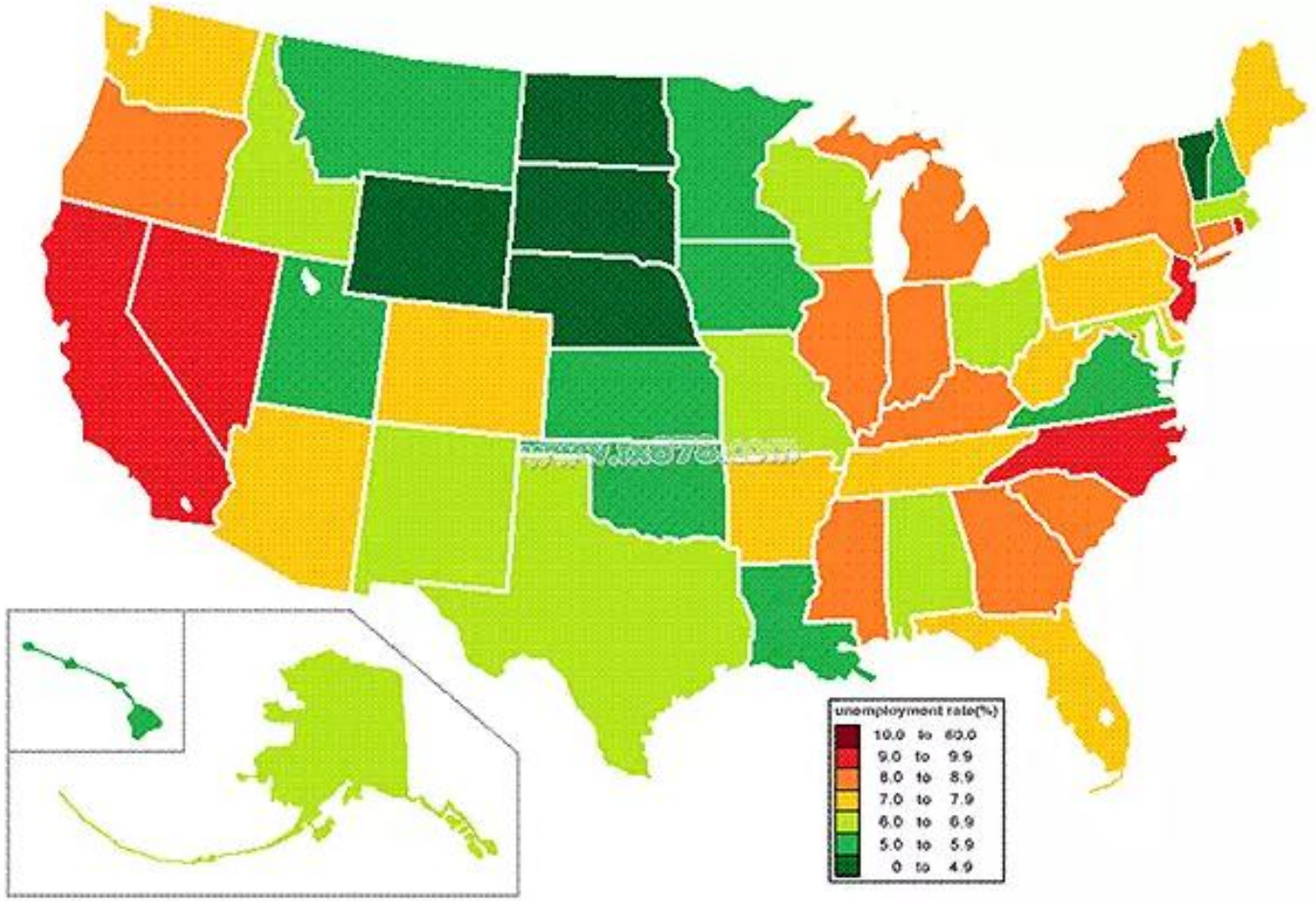
比如下面这张眼球热力图，通过颜色的差异，我们可以直观的看到用户的关注点。



二、颜色可视化

b: 2013年美国失业率统计

在图中可以看到，通过对美国地图以州为单位的划分，用不同的颜色来代表不同的失业率等级范围，整个的全美失业率状况便尽收眼底了。



二、颜色可视化

c: 美国手机用户城市分布

图中红点是用iPhone的人，绿点是用安卓的人。这两张在微博上看到的图，第一张是美国一个城市的一览，第二张图特写了纽约的市中心，尤其是曼哈顿地区。我们可以看到在市中心和主干道的人用iPhone居多，而用安卓的人都在郊区。这也引起了人们的热议，有的说在美国富人都住郊区别墅，所以富人爱用安卓手机；有的反驳说曼哈顿地区的人几乎都用iPhone，说明富人喜欢用iPhone手机。不管结论如何，都足以说明用户都被这些图所吸引，所以可视化的方式效果真的很直观。

注：科学家统计了2年里30亿条含有地理数据的twitter推文，根据客户端总结出来的数据。

三、图形可视化

在我们设计指标及数据时，使用有对应实际含义的图形来结合呈现，会使数据图表更加生动的被展现，更便于用户理解图表要表达的主题。

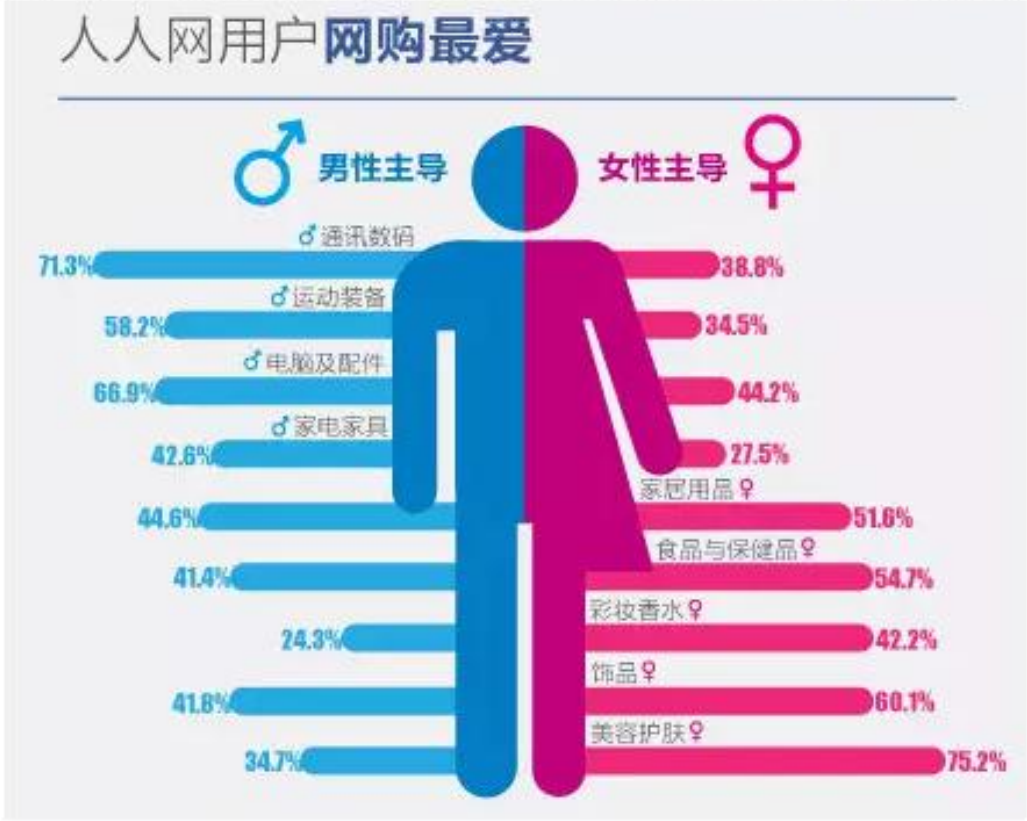
a: iOS手机及平板分布

如下图所示，当展示使用不同类型的手机和平板用户占比时，直接用总的苹果图形为背景来划分用户比例，让用户第一眼就可以直观的看到这些图是在描述苹果设备的，直观而清晰。

三、图形可视化

b: 人人网用户的网购调查

下图可以看出，该数据可视化的设计直接采用男性和女性的图形，这样的设计让分类一目了然。再结合了颜色可视化（左面蓝色右面粉色），同时也采用了面积&尺寸可视化，不同的比例用不同长度的条形。这些可视化方法的组合使用，大大加强了数据的可理解性。



四、地域空间可视化

当指标数据要表达的主题跟地域有关联时，我们一般会选择用地图为大背景。这样用户可以直观的了解整体的数据情况，同时也可以根据地理位置快速的定位到某一地区来查看详细数据。

四、地域空间可视化

a: 美国最好喝啤酒的产地分布

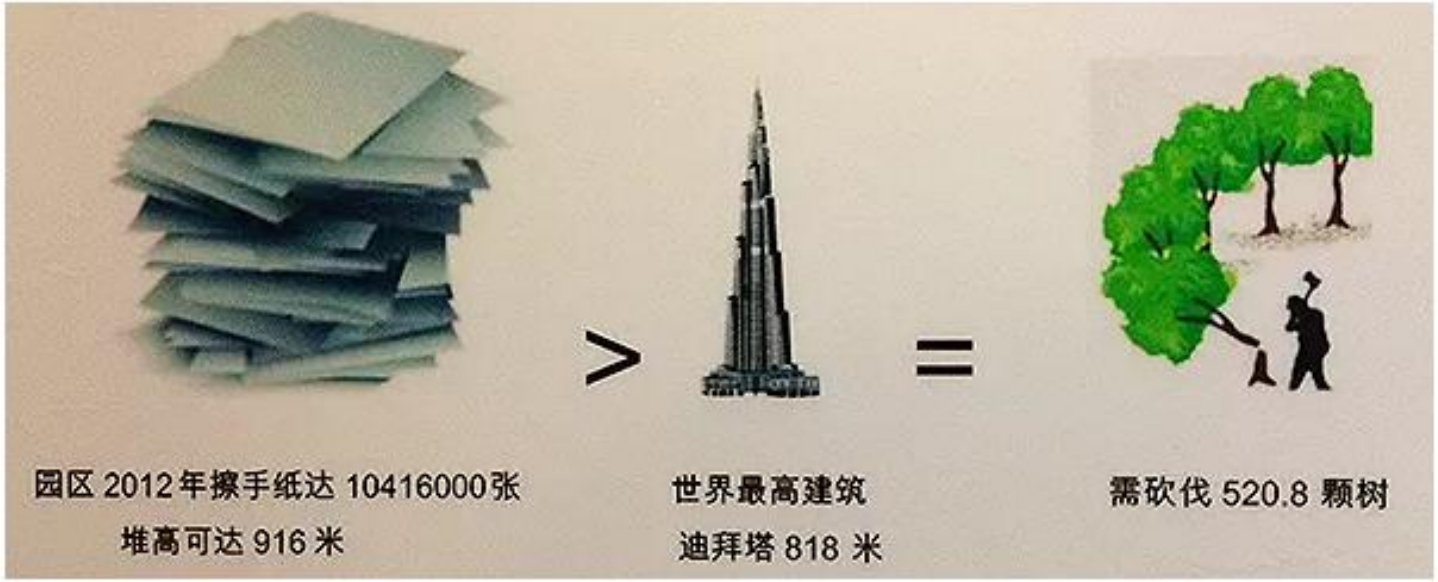
下图中，通过以美国地图为大背景，清晰的记录了不同州所产啤酒在1987-2007年间在美国啤酒节中获得的奖牌累计总数。再辅以颜色可视化的方法，让用户清晰的看到美国哪些州更盛产好喝的啤酒。

五、概念可视化

通过将抽象的指标数据转换成我们熟悉的容易感知的数据时，用户便更容易理解图形要表达的意义。

a: 厕所贴士

下图是厕所里贴在墙上的节省纸张的环保贴士，用了概念转换的方法，让用户清晰的感受到员工们一年的用纸量之多。如果只是描述擦手纸的量及堆积可达高度，我们还没有什么显性化概念。但当用户看到用纸的堆积高度比世界最高建筑还高、同时需砍伐500多颗树时，想必用户的节省纸张甚至禁用纸张的情怀便油然而生了。所以可见用概念转换的方法是多么的重要和有效。



五、概念可视化

b: Flickr云存储空间达1TB的可视化描述

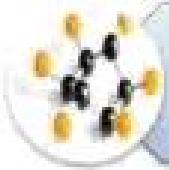
Flickr对云存储空间升至1TB确实是让人开心的事情，但相信很多人对这一数量级所代表的含义并不清晰。所以Flickr在宣传这一新的升级产品时，采用了概念可视化的方案。从下图可以看出，用户可以动态的选择照片的大小，之后Flickr会采用动态交互的方式计算和显示出1TB能容纳多少张对应大小的图片。这样一来，用户便有了清晰的概念，知道这1TB是什么量级的容量了。

图表的作用

迅速传达信息

直接关注重点

更明确地显示其相互关系



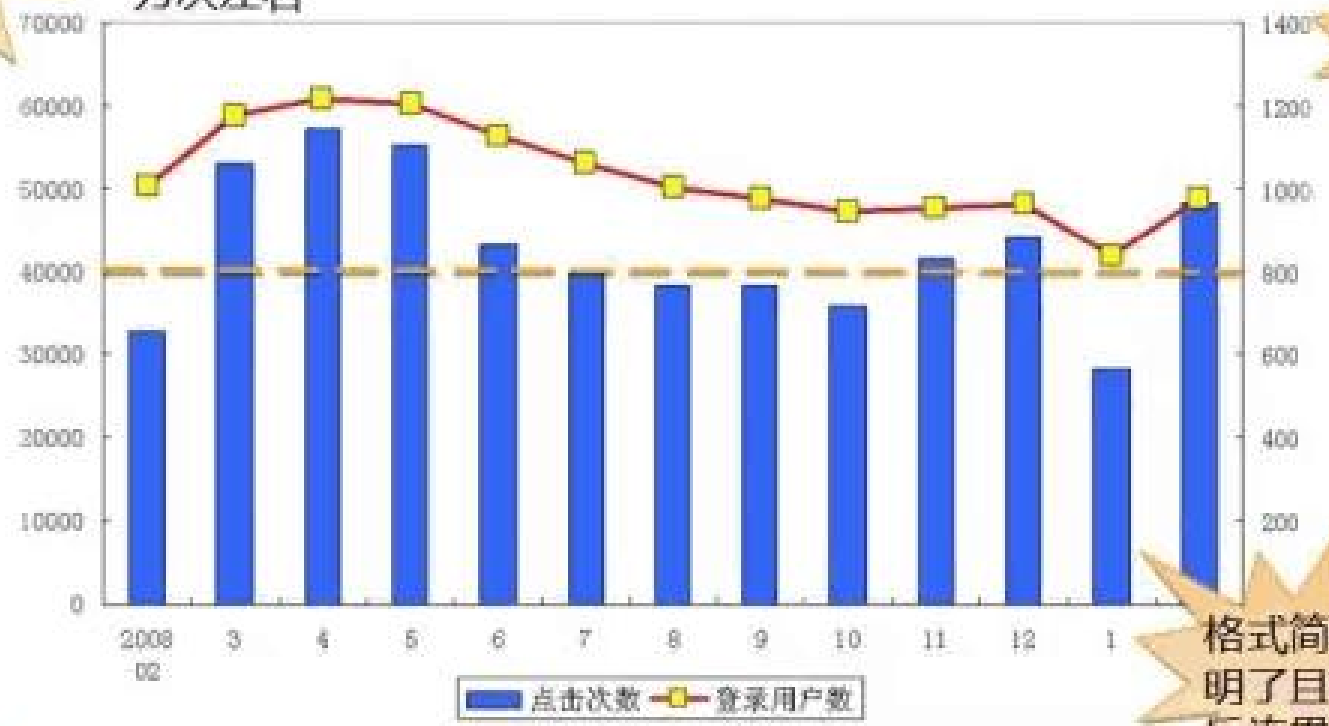
使信息的表达鲜明生动

成功图表的几个关键因素

图表与标题
相辅相成

每张图表都
表达一个明
确的信息

除春节、国庆点击明显下降外，每月都保持在8百人，4万次左右

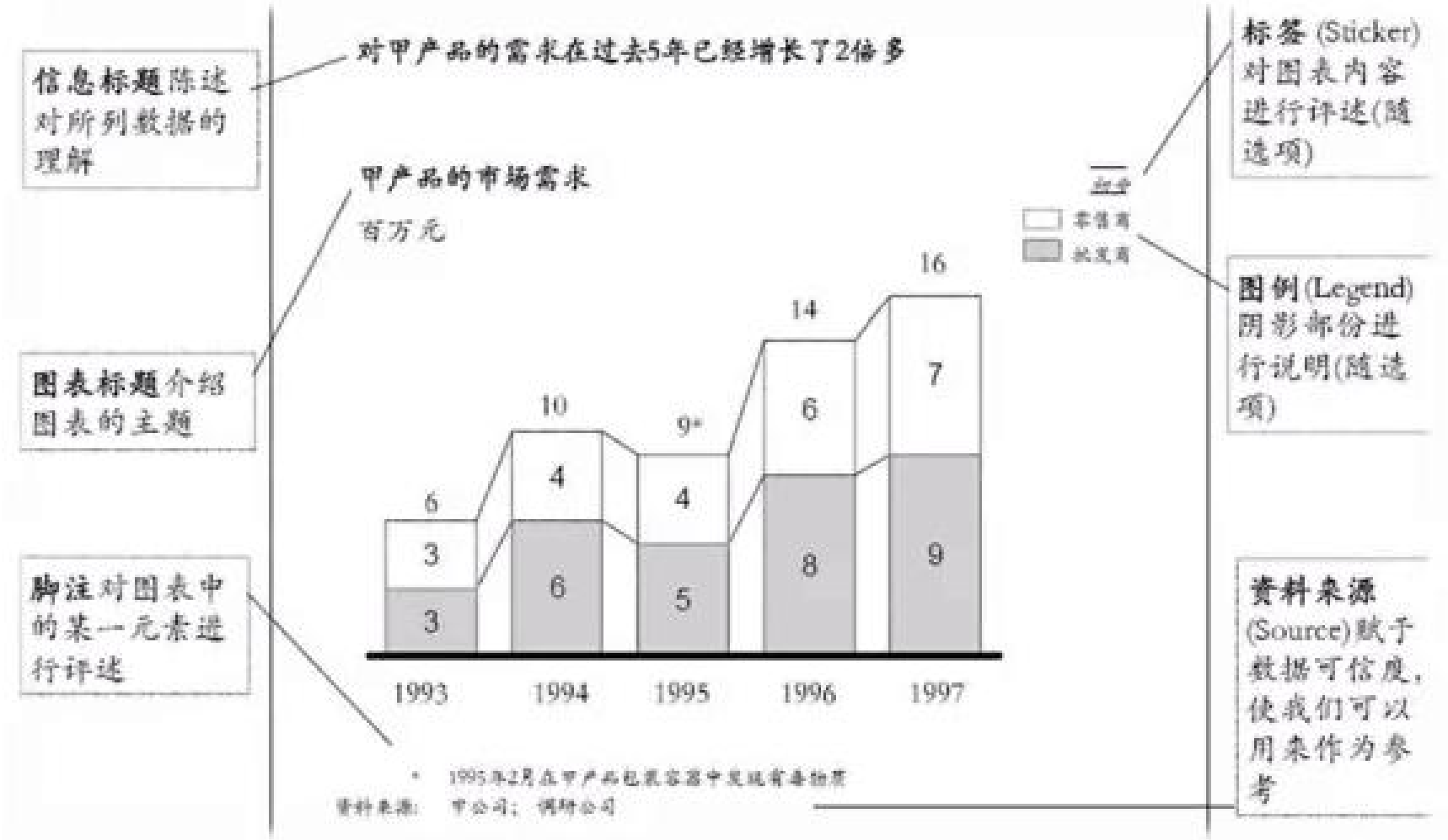


清晰
易懂

格式简单
明了且前
后连贯

少而精

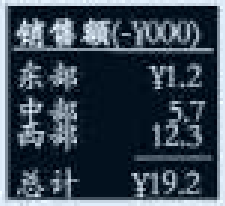
一个好的图表应该遵循一定的标准格式



图表类型

数据类图表

数值？



表格

饼图

条形图

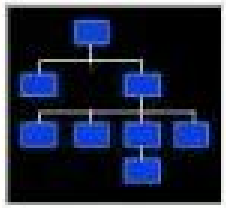
柱形图

线性图

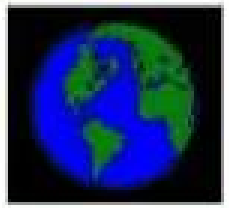
散点图

概念类图表

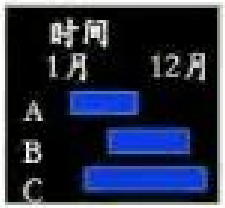
人员？ 地点？ 时间？



组织架构图

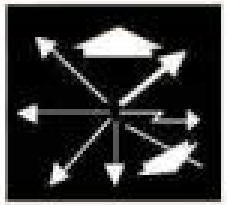


地图

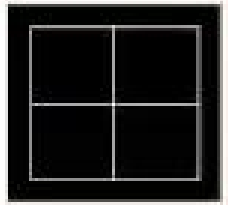


甘特图

流程图



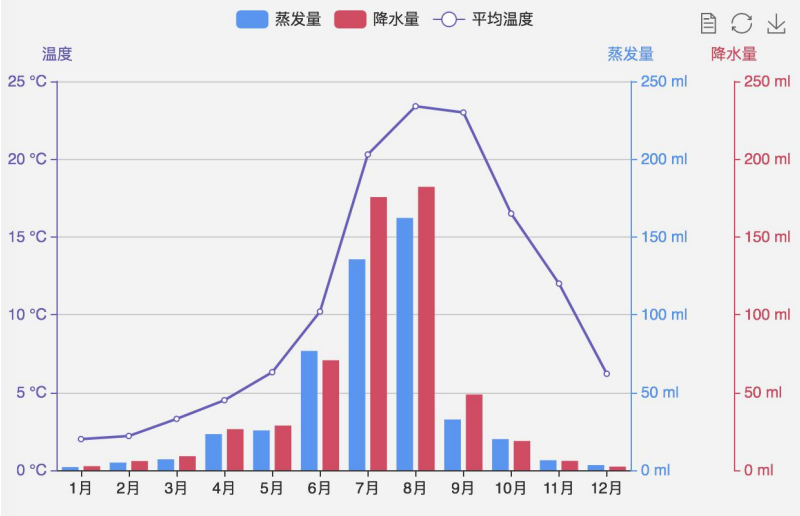
概念

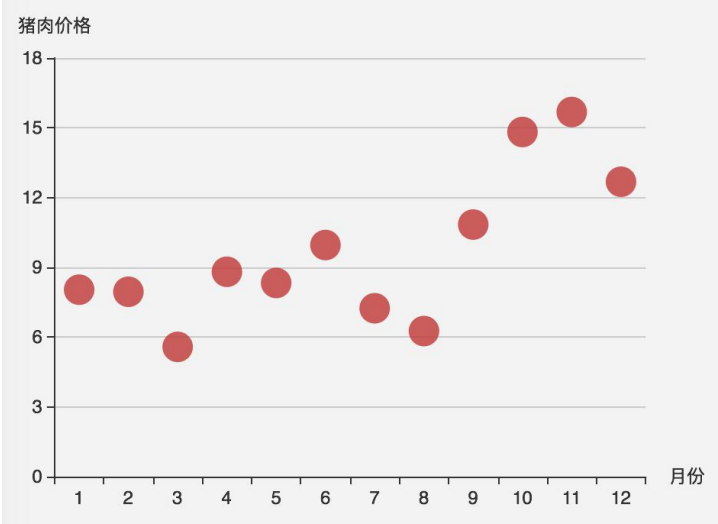


矩阵

常见可视化图表

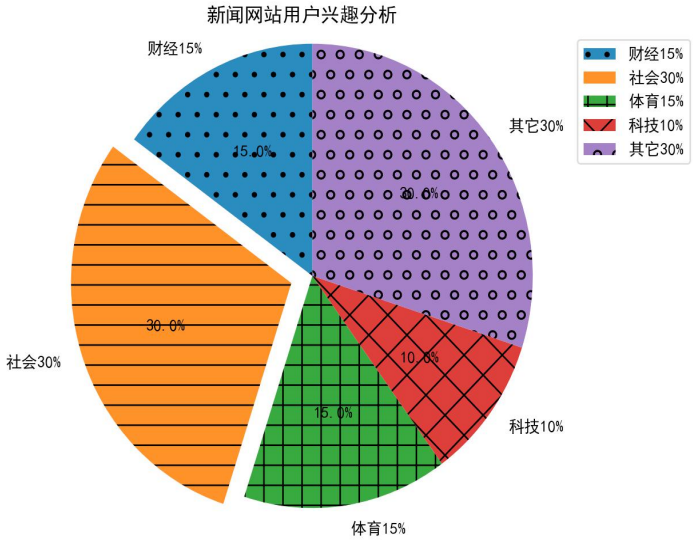
- 时间序列可视化



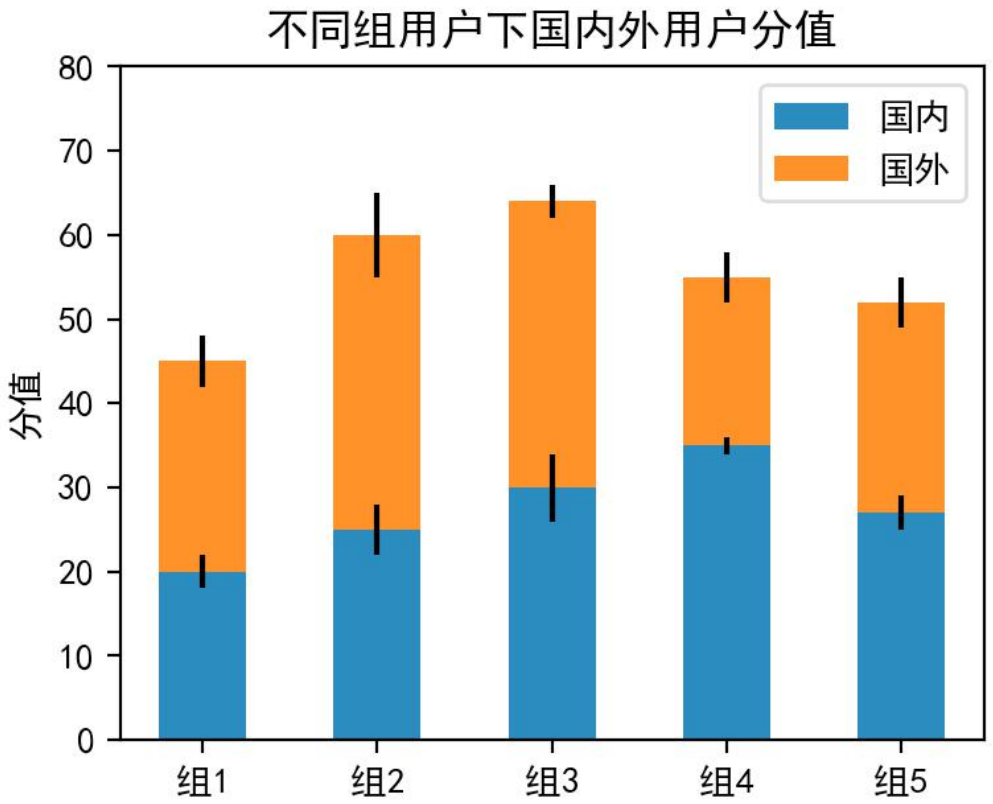


常见可视化图表

```
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif']=['SimHei'] #用来正常显示中文
plt.rcParams['axes.unicode_minus']=False #用来正常显示负号

labels = '财经15%', '社会30%', '体育15%','科技10%', '其它30%`   #初始化参数autopct为显示的百分比样式
sizes = [15, 30, 15, 10, 30]
explode = (0, 0.1, 0, 0,0)#突出第2项
fig1, ax1 = plt.subplots()
pie = ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',shadow=False,
startangle=90)

patches = pie[0]    #设置分块的填充模式
patches[0].set_hatch('.')
patches[1].set_hatch('-')
patches[2].set_hatch('+')
patches[3].set_hatch('x')
patches[4].set_hatch('o')

plt.legend(patches, labels)
ax1.axis('equal')
plt.title('新闻网站用户兴趣分析')
plt.show()
```



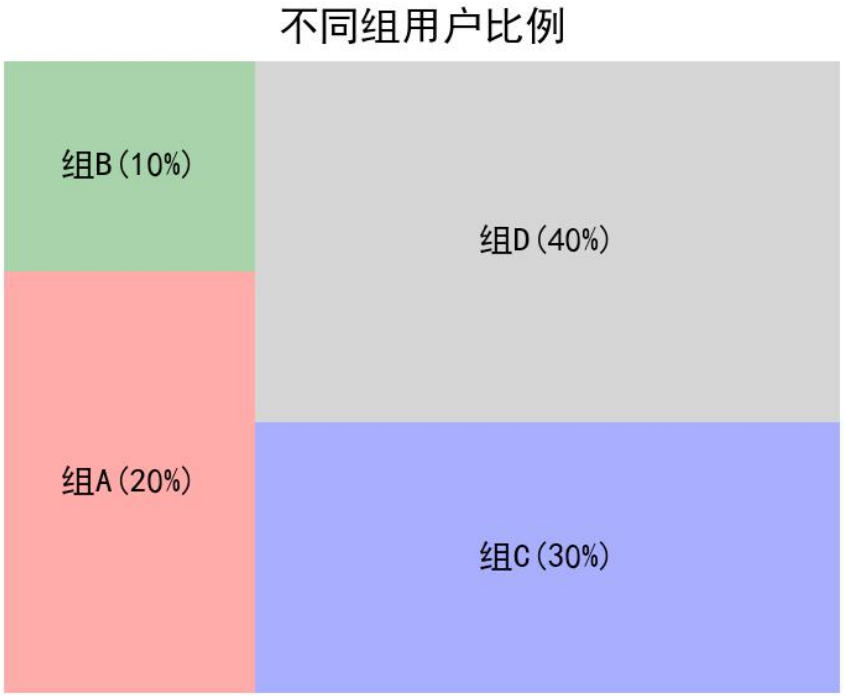
常见可视化图表

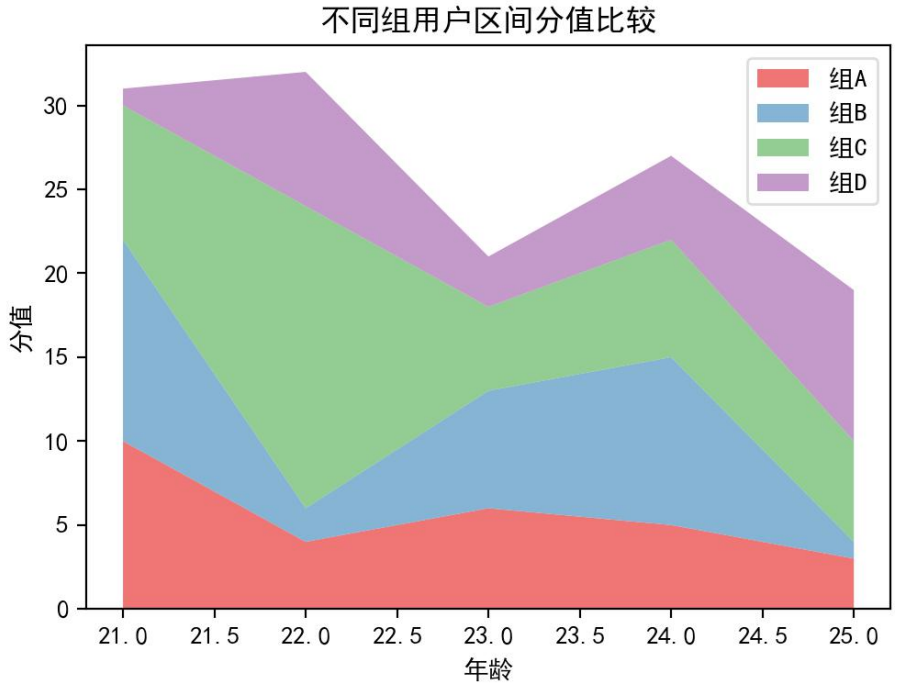



```
import numpy as np
N=5
inMeans=(20,25,30,35,27)
outMeans=(25,35,34,20,25)
inStd = (2,3,4,1,2)
outStd=(3,5,2,3,3)
ind=np.arange(N)     #Bar坐标位置
width=0.5      #Bar的宽度
#使用plt.bar()方法生成两个国人和国外两组柱子
p1=plt.bar(ind, inMeans, width, yerr=inStd)
P2=plt.bar(ind,outMeans,width,bottom=inMeans,yerr=outStd)
#查看不同组用户的总分值基础上，查看组内不同类别的用户分值占比情况
plt.ylabel('分值')
plt.title('不同组用户下国内外用户分值')
plt.xticks(ind, ('组1', '组2', '组3', '组4', '组5'))
plt.yticks(np.arange(0, 81, 10))
plt.legend((p1[0], p2[0]), ('国内', '国外'))
plt.show()
```

常见可视化图表

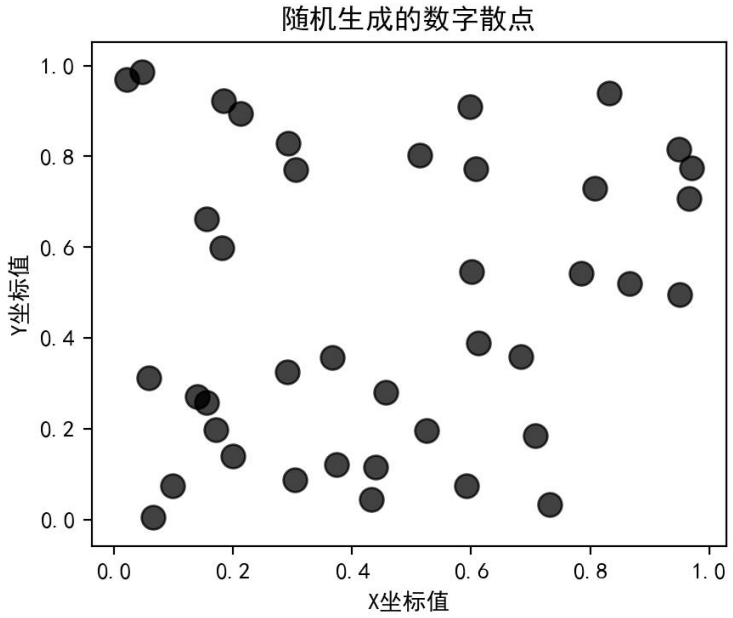
- 比例的可视化

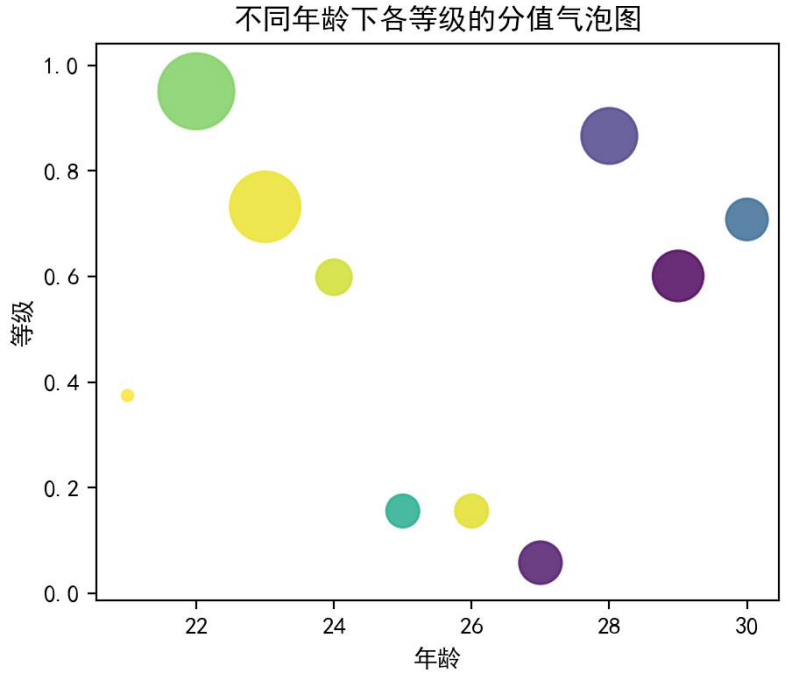




常见可视化图表

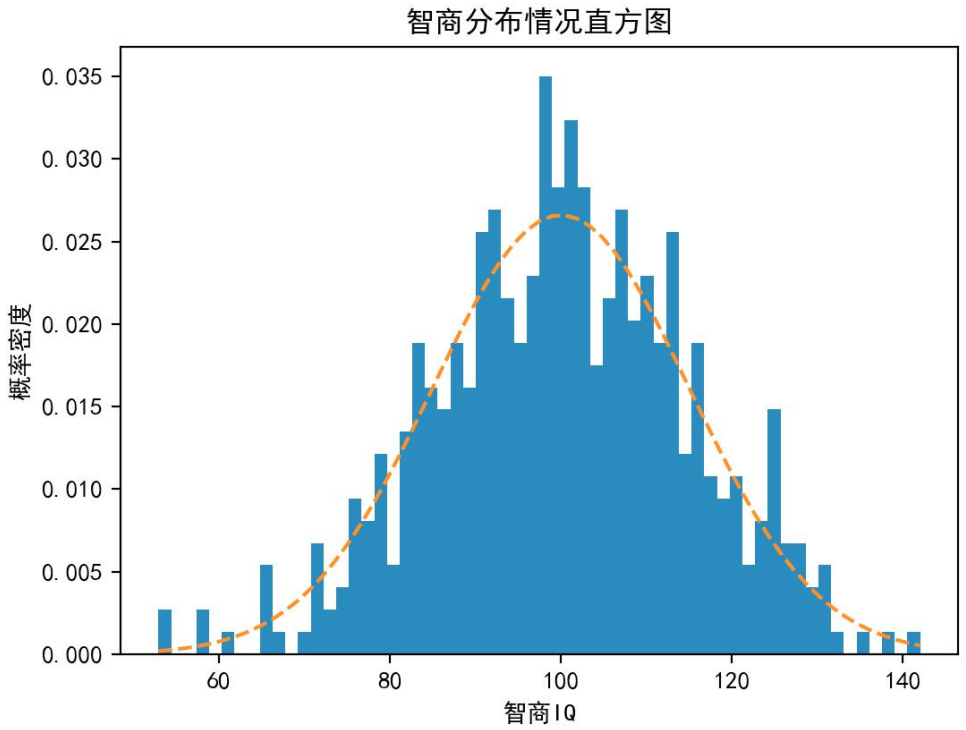
- 关系可视化

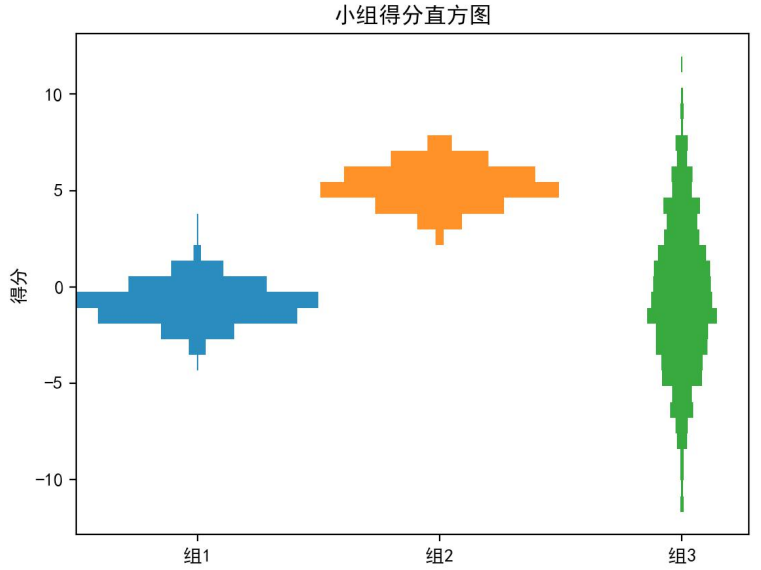




常见可视化图表

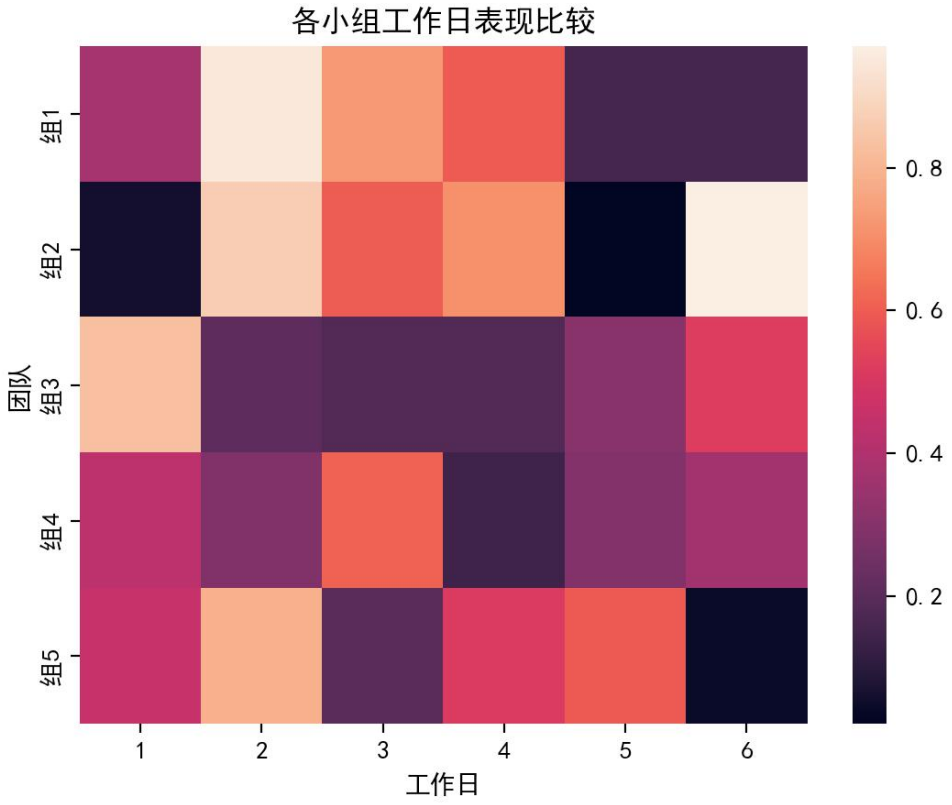
- 关系可视化





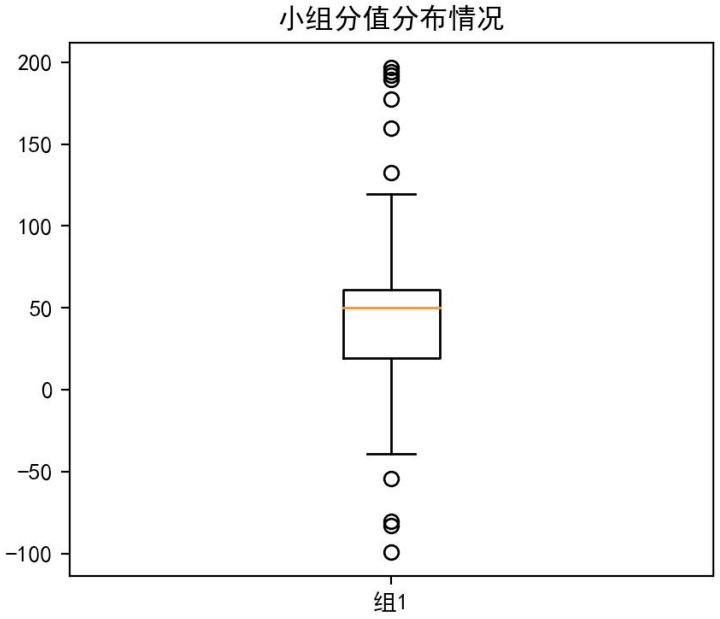
常见可视化图表

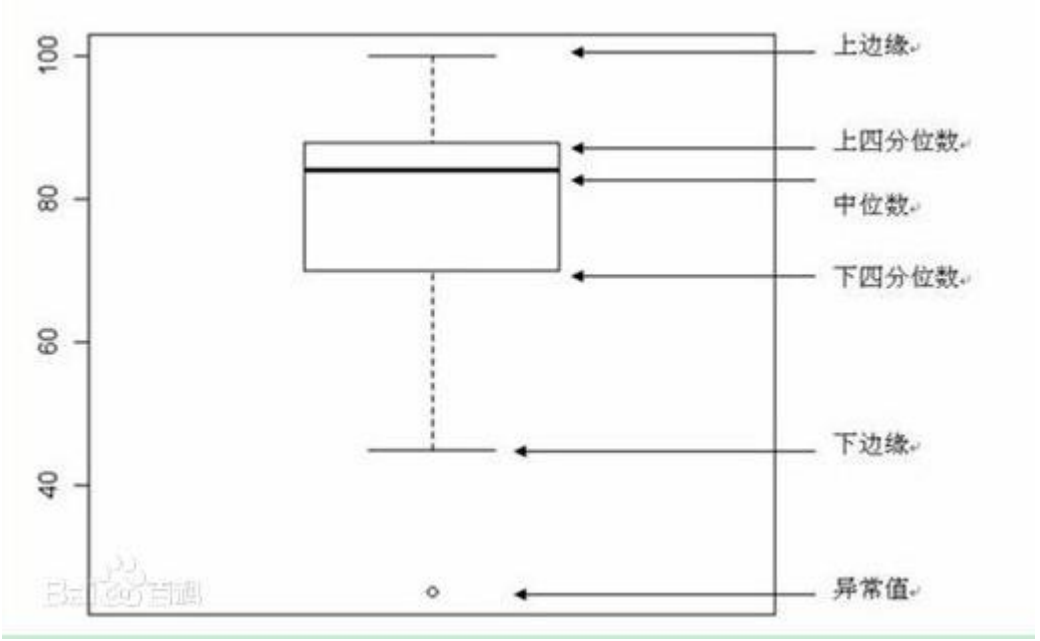
- 差异可视化



常见可视化图表

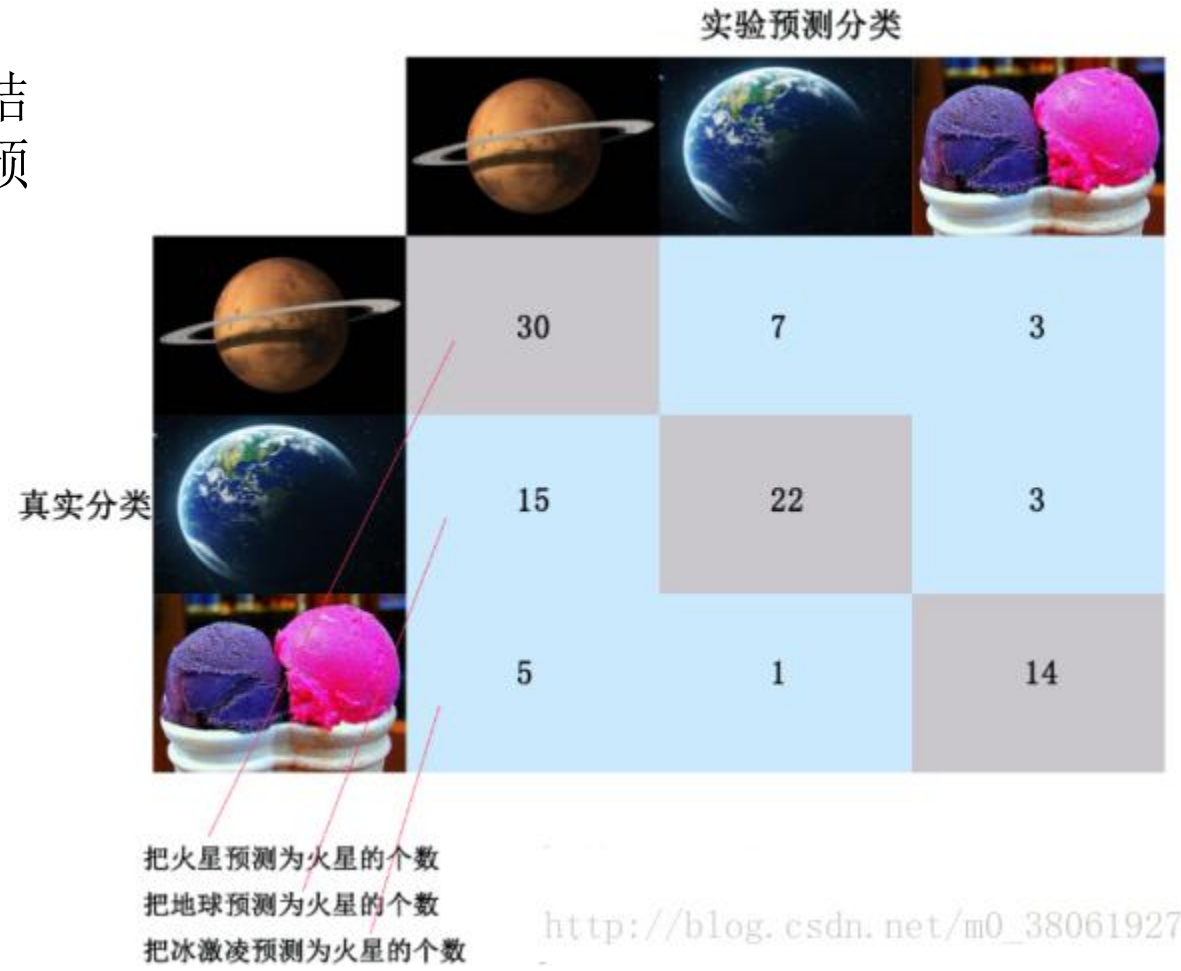
- 箱线图


小组分值分布情况
200
150
100
50
0
−50
−100
组1


100
80
60
40
上边缘
上四分位数
中位数
下四分位数
下边缘
异常值

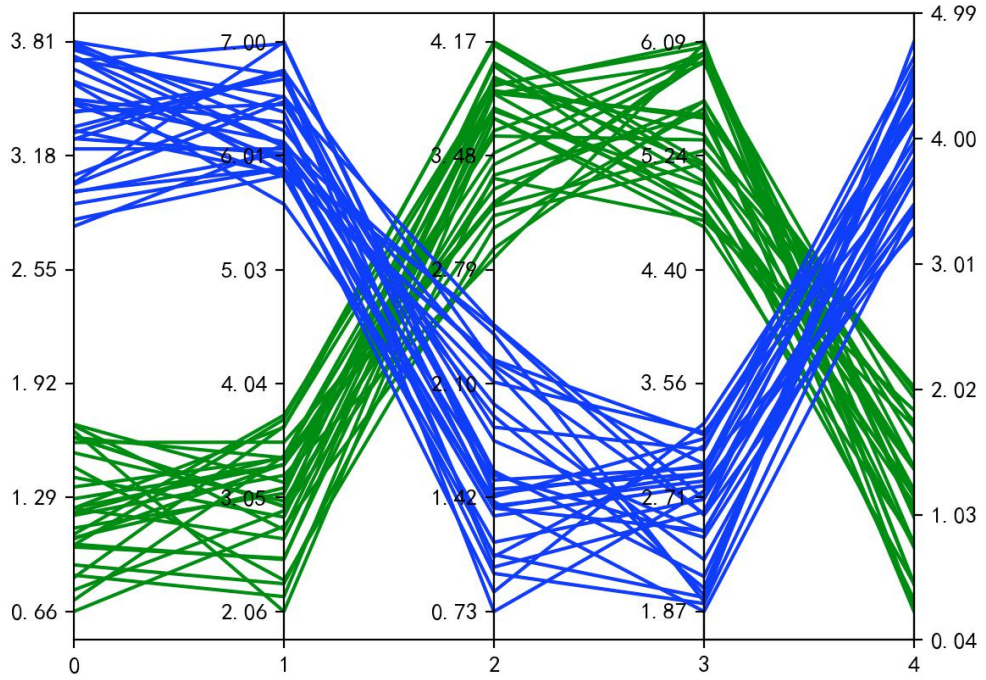
常见可视化图表

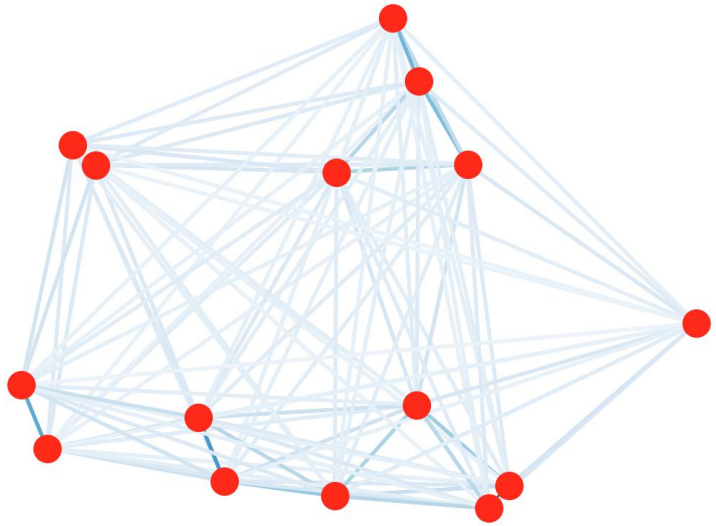
- 混淆矩阵 Confusion Matrix
  - 其中灰色部分是真实分类和预测分类结果相一致的，绿色部分是真实分类和预测分类不一致的，即分类错误的。



常见可视化图表

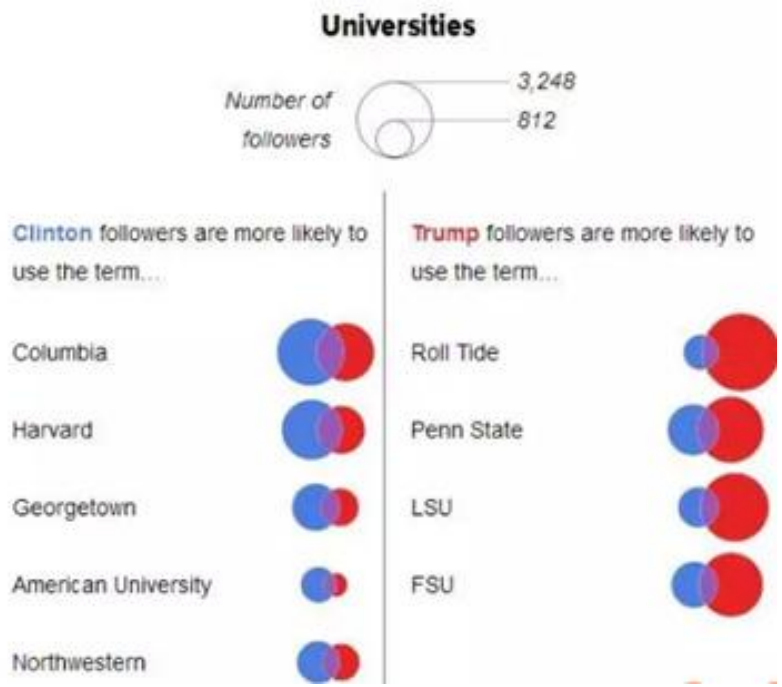
- 差异可视化





常见可视化图表

- 空间关系可视化



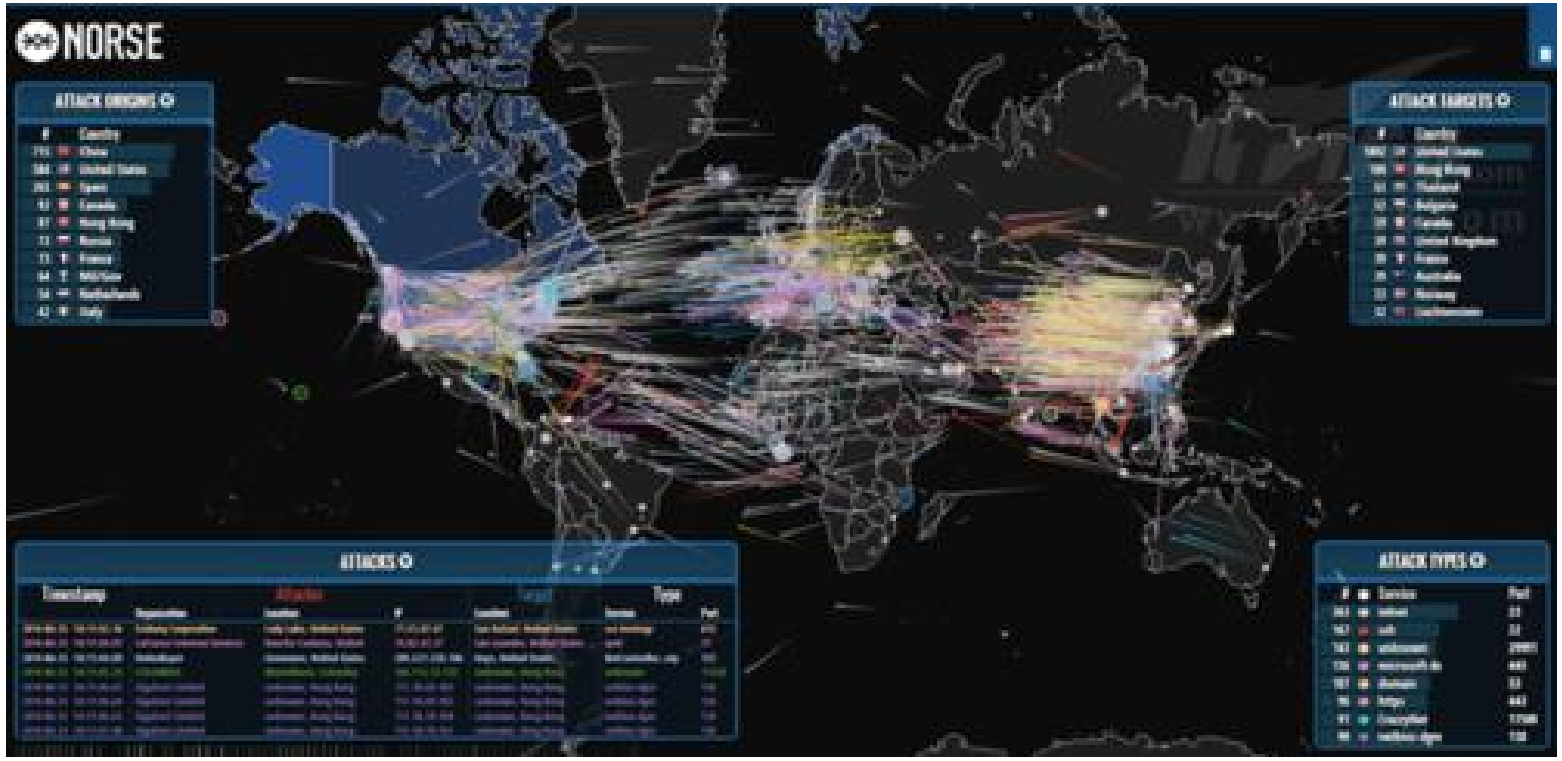
income

	clinton	trump	other/no answer
under $30,000 **17%**	**53%**	41%	6%
$30k-$49,999 **19%**	**51%**	42%	7%
$50k-$99,999 **31%**	46%	**50%**	4%
$100k-$199,999 **24%**	47%	**48%**	5%
$200k-$249,999 **4%**	48%	**49%**	3%
$250,000 or more **6%**	46%	**48%**	6%

24537 respondents

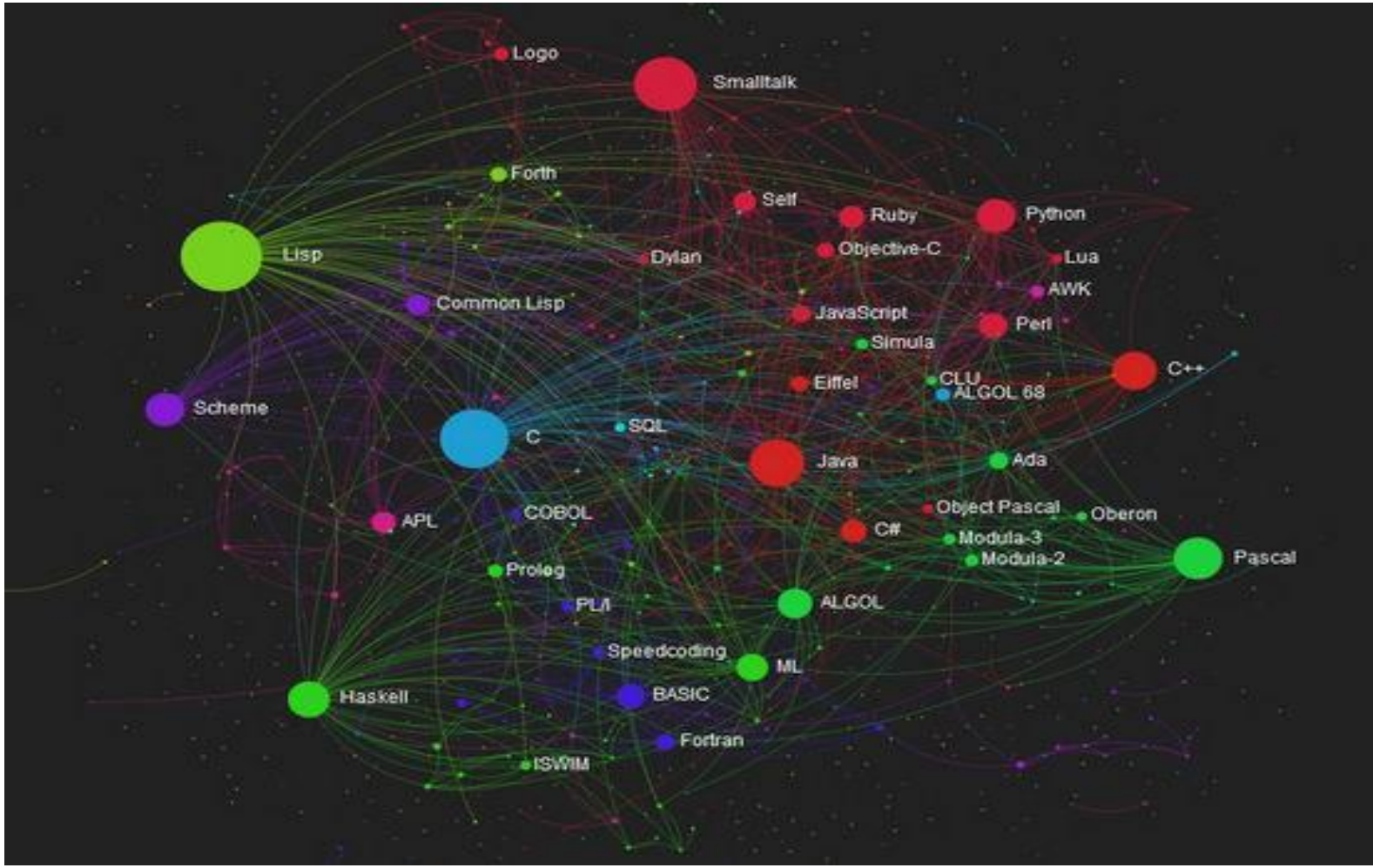
可视化案例：全球黑客活动

- 安全供应商Norse打造了一张能够反映全球范围内黑客攻击频率的地图（http://map.ipviking.com）。地图中的每一条线代表的都是一次攻击活动，借此可以了解每一天、每一分钟甚至每一秒世界上发生了多少次恶意渗透。



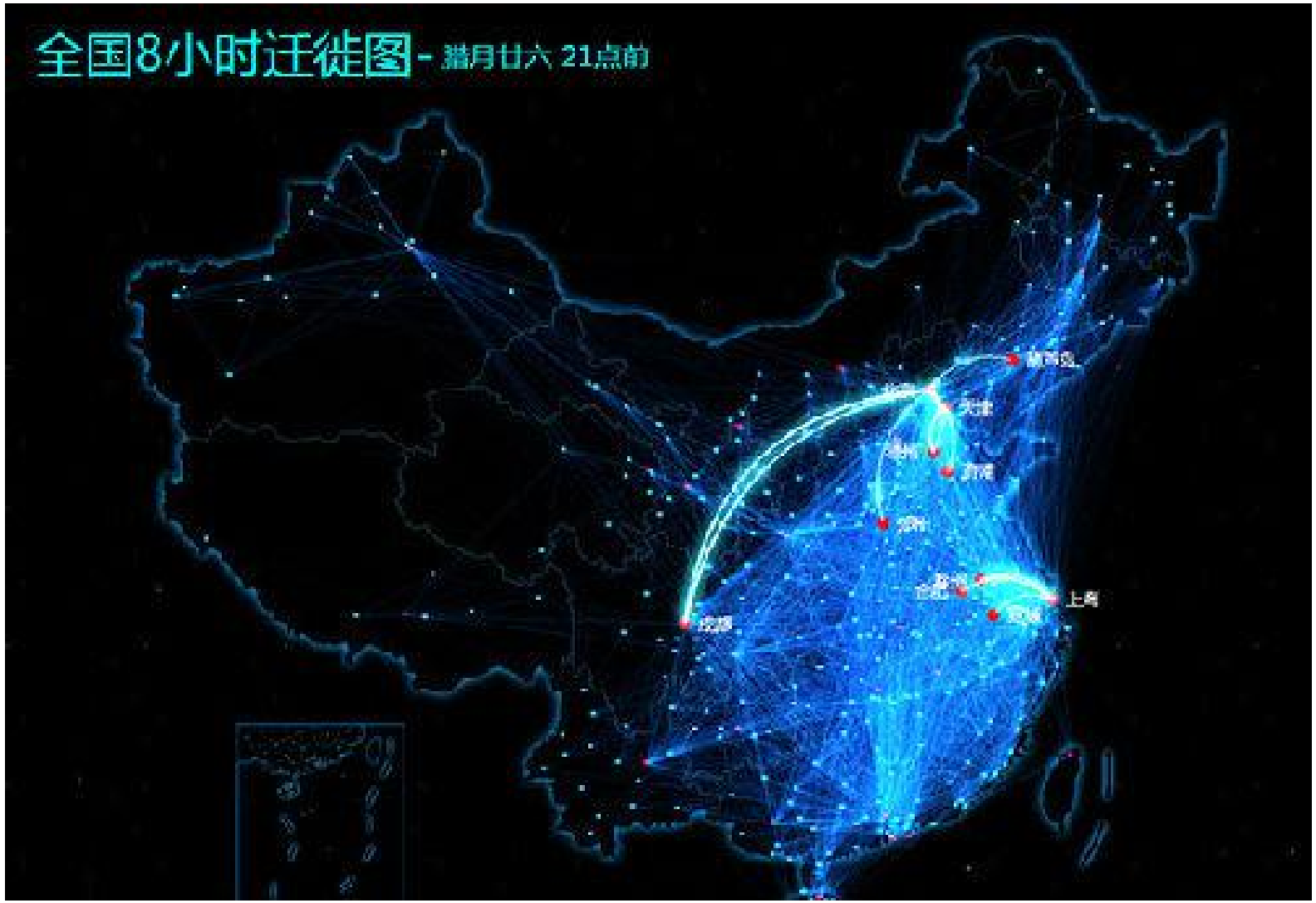
可视化案例：编程语言之间的影响力关系图

Ramio Gómez利用来自Freebase上的编程语言维护表里的数据，绘制了编程语言之间的影响力关系图，图中的每个节点代表一种编程语言，之间的连线代表该编程语言对其他语言有影响，有影响力的语言会连线多个语言，相应的节点也会越大。



可视化案例：百度迁徙

2014年，央视与百度合作，启用百度地图定位可视化大数据播报春节期间全国人口迁徙情况，引起广泛关注。



可视化分析面临的挑战

- 进行可视化分析时挑战主要来自于两个方面：数据和可视化结果
- 数据层面的挑战包括数据的来源不唯、数据质量良莠不齐、数据整合困难等挑战。信息时代数据更新飞快、体量大，对可视化分析速度要求越来越高。分析过程涉及领域广而繁杂，对于数据的专业解读带来挑战
- 在可视化结果层面，数据集中样本的相关性导致视觉噪声的大量出现，面临降噪的挑战。 受限于设备的长宽比、分辨率、现实世界的感受等，可视化图表中大型图像的感知的挑战；受限于可视化的算法以及硬件的性能，及时响应，高速图像变换的挑战；专业领域不同带来的可视化需求不同，最大限度地满足受众视觉喜好的挑战
- 此外还有可视化分析流程的优化，可视化分析工具的可操作性等等。