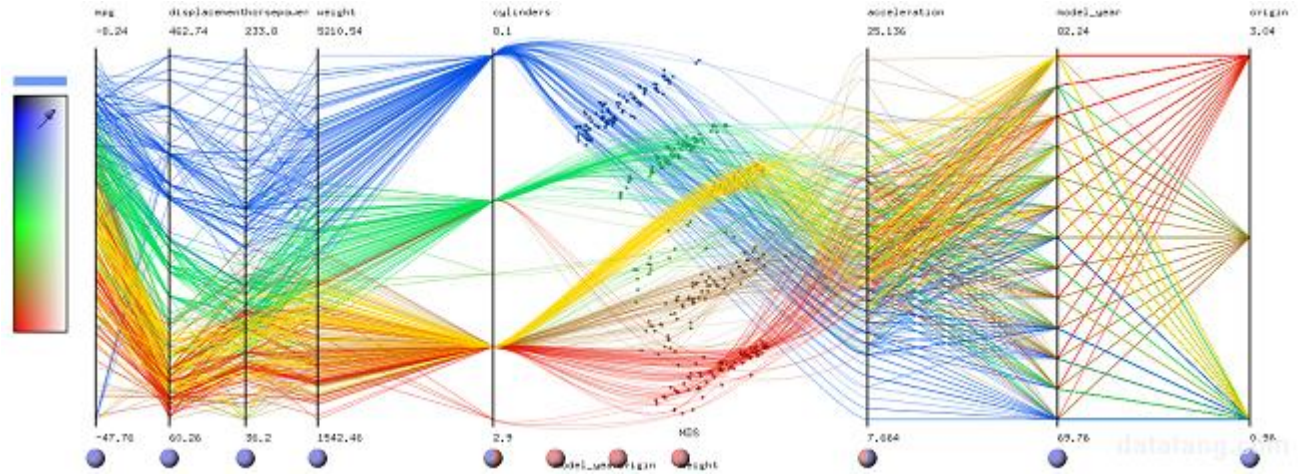
可视化

A picture is worth more than a thousand words.



什么是可视化

可视化（Visualization）是利用计算机图形学和图像处理技术，将数据转换成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。

为什么要可视化

大脑最容易接受辨认的信息就是颜色信息和形状信息，用不同色彩组成的图形是最容易被大脑所识别。

因此将数据进行可视化，主要是为了更好地展示数据里的信息。将数据里包含的信息以最直接最直观的方式展现给观察者。而这些信息最容易让观察者理解其含义。

为什么要可视化

1. 从天大东门向前走到鞍山西道。
2. 然后右转，直到走到南京路。
3. 右转走到滨江道

可视化的原则

在设计可视化图表时，我们应该遵循以下两个原则：

1. Expressiveness

A set of facts is expressible in a visual language if the sentences (i.e., the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

2. Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

图表属性

数据类型：Categorical, Ordinal, Interval, Ratio

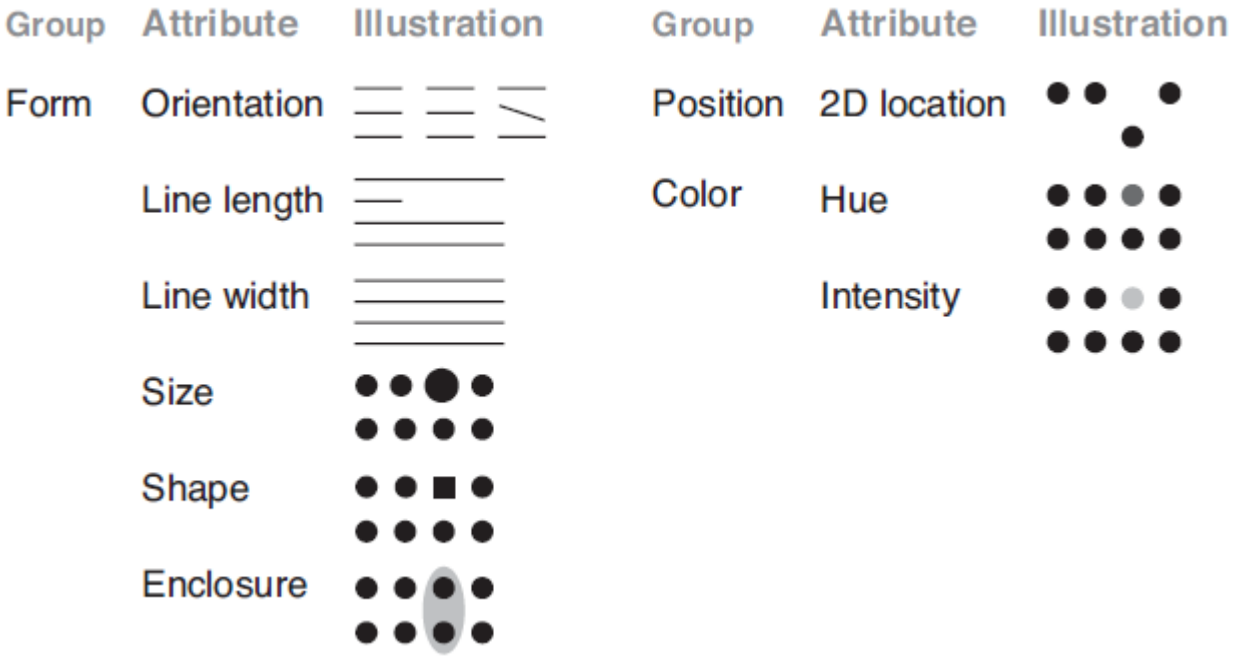
视觉属性：

① form

② color

③ position

④ motion



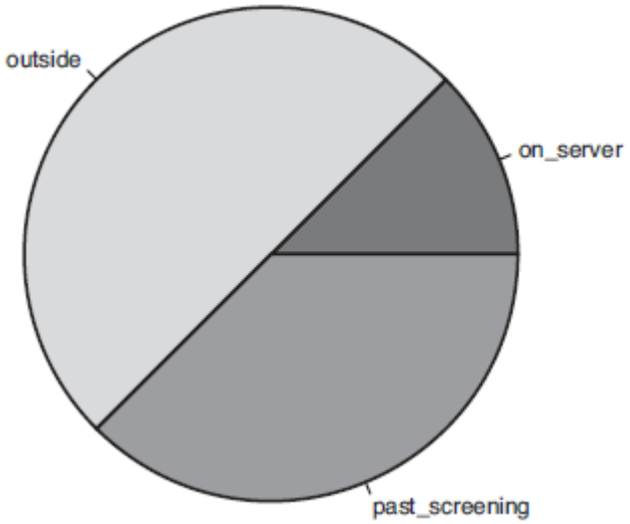
基本图形

1. 扇形图
2. 条形图
3. 折线图
4. 堆叠扇形图
5. 堆叠条形图
6. 堆叠折线图
7. 直方图
8. 箱线图
9. 散点图
10. 平行坐标
11. 连接图
12. 地区图（MAPS）
13. 矩形树状图（TREEMAPS）

扇形图

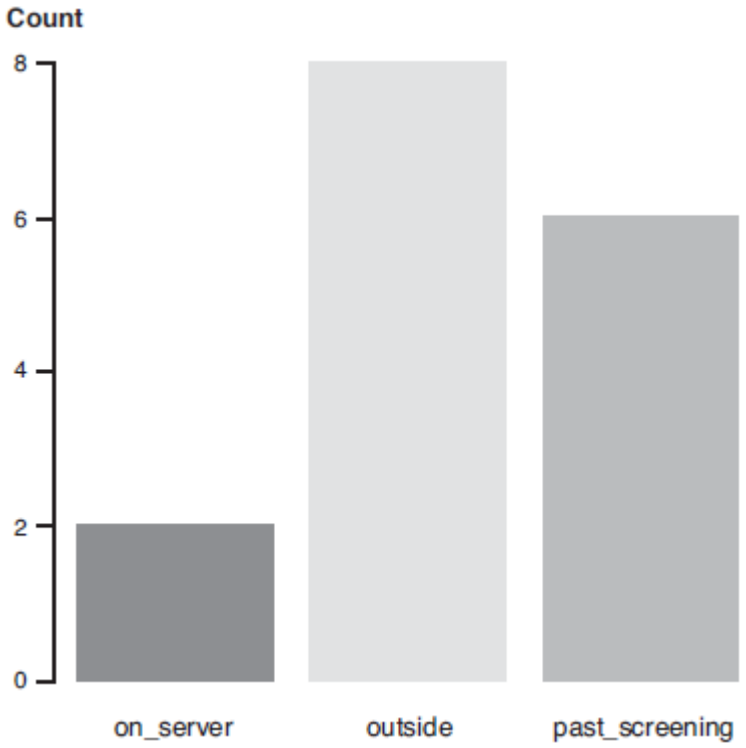
扇形统计图是用整个圆表示总数，用圆内各个扇形的大小表示各部分数量占总数的百分数。

通过扇形统计图可以很清楚地表示出各部分数量同总数之间的关系。



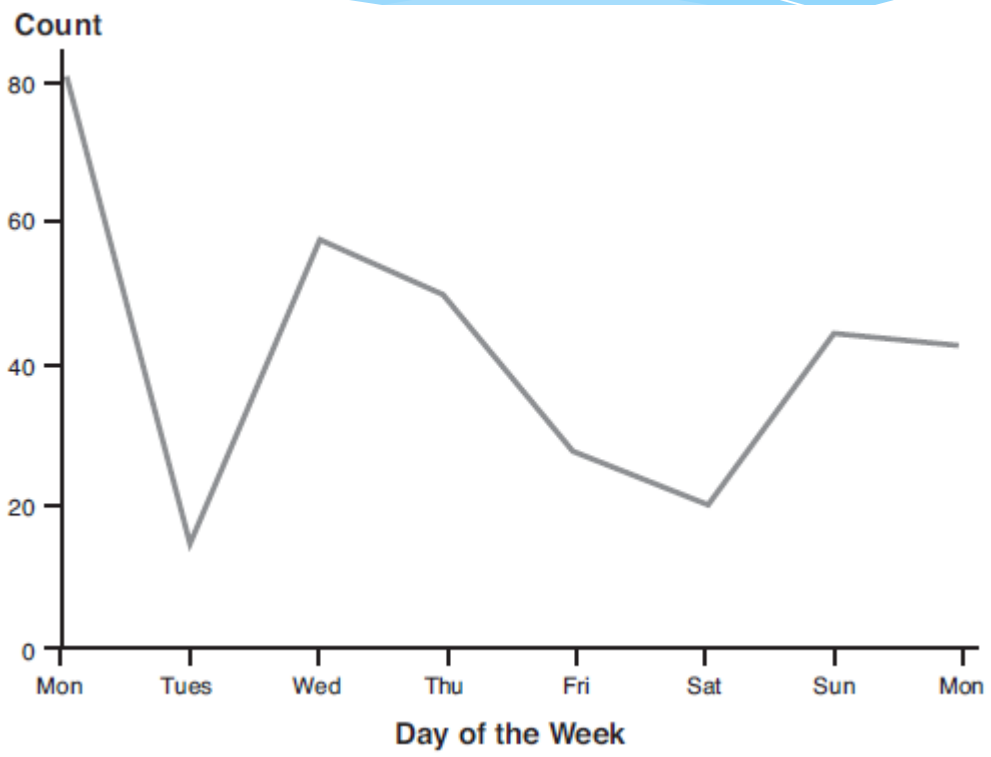
条形图

条形图用来显示不连接的且无关的对象的差别情况，这种图表类型的淡化数值随时间的变化而变化，能突出**数值**的比较。



折线图

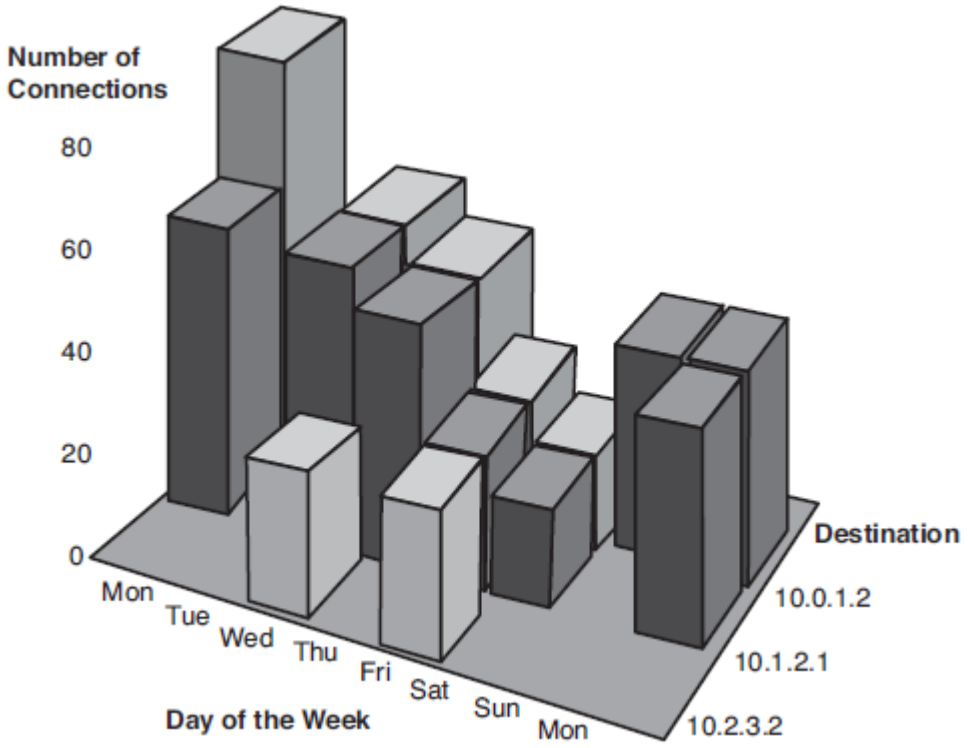
折线图可以显示随时间而变化的连续数据，因此非常适用于显示在相等时间间隔下数据的趋势。



3D条形图

和条形图相比，3D条形图能够显示二维甚至三维的数据。

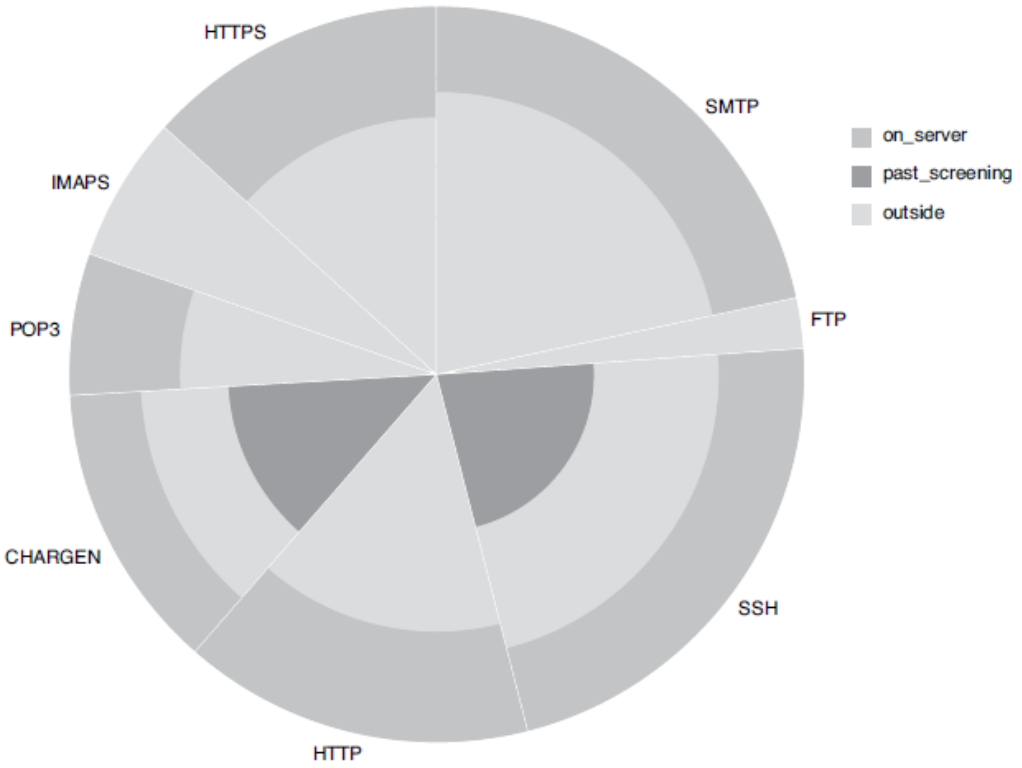
但是，3D条形图也有一个缺点，即有些条块可能被前面的挡住。



堆叠扇形图

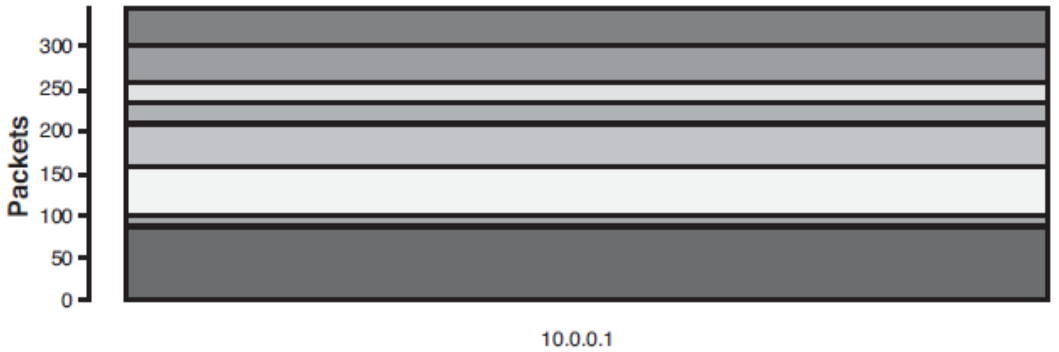
堆叠扇形图能表示二维的数据。

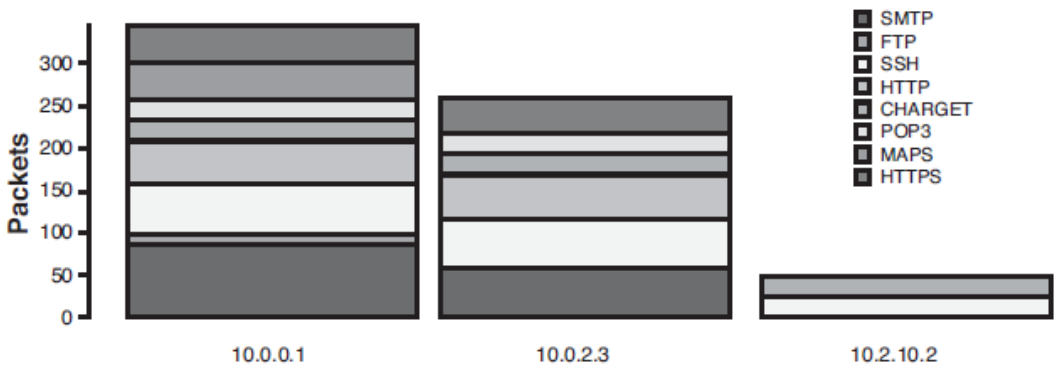
在每一个小的扇区里又可以表达一维数据。



堆叠条形图

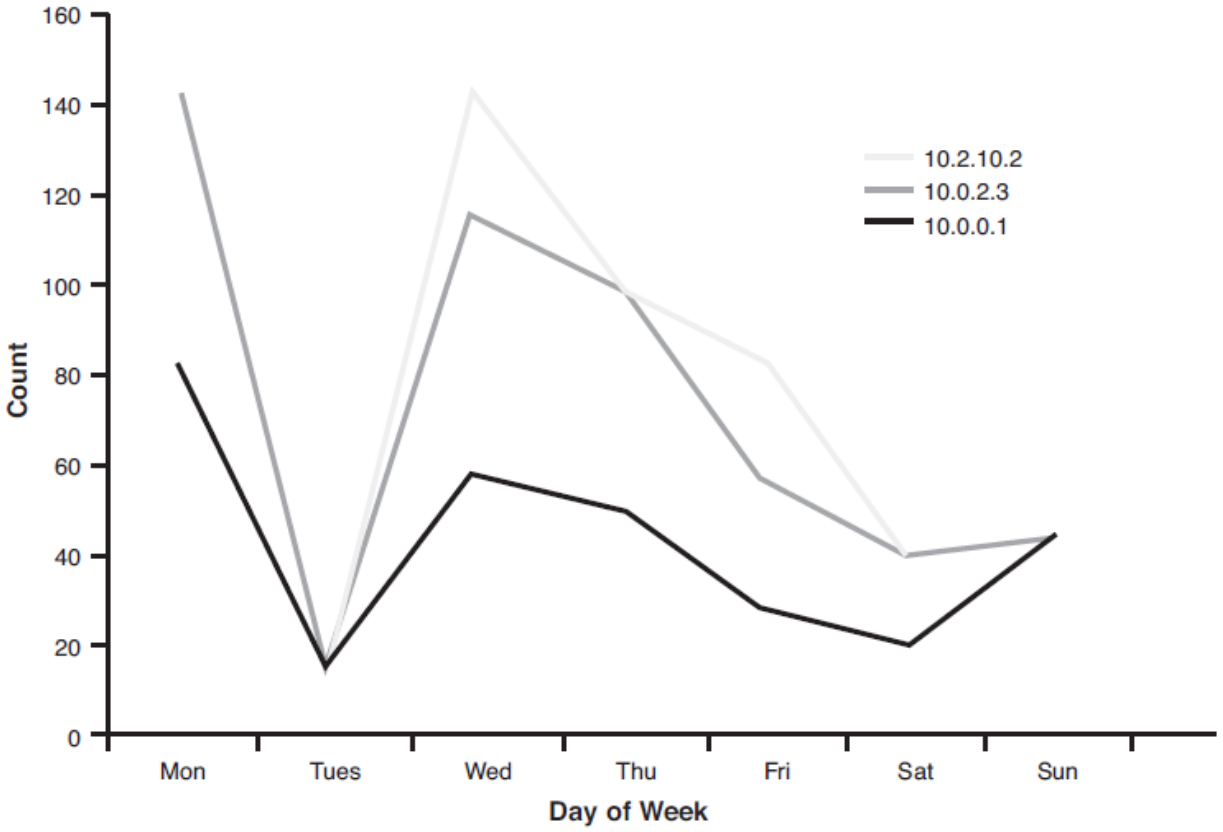
堆叠条形图在条形图的基础上增加了一维，可以表示二维数据。





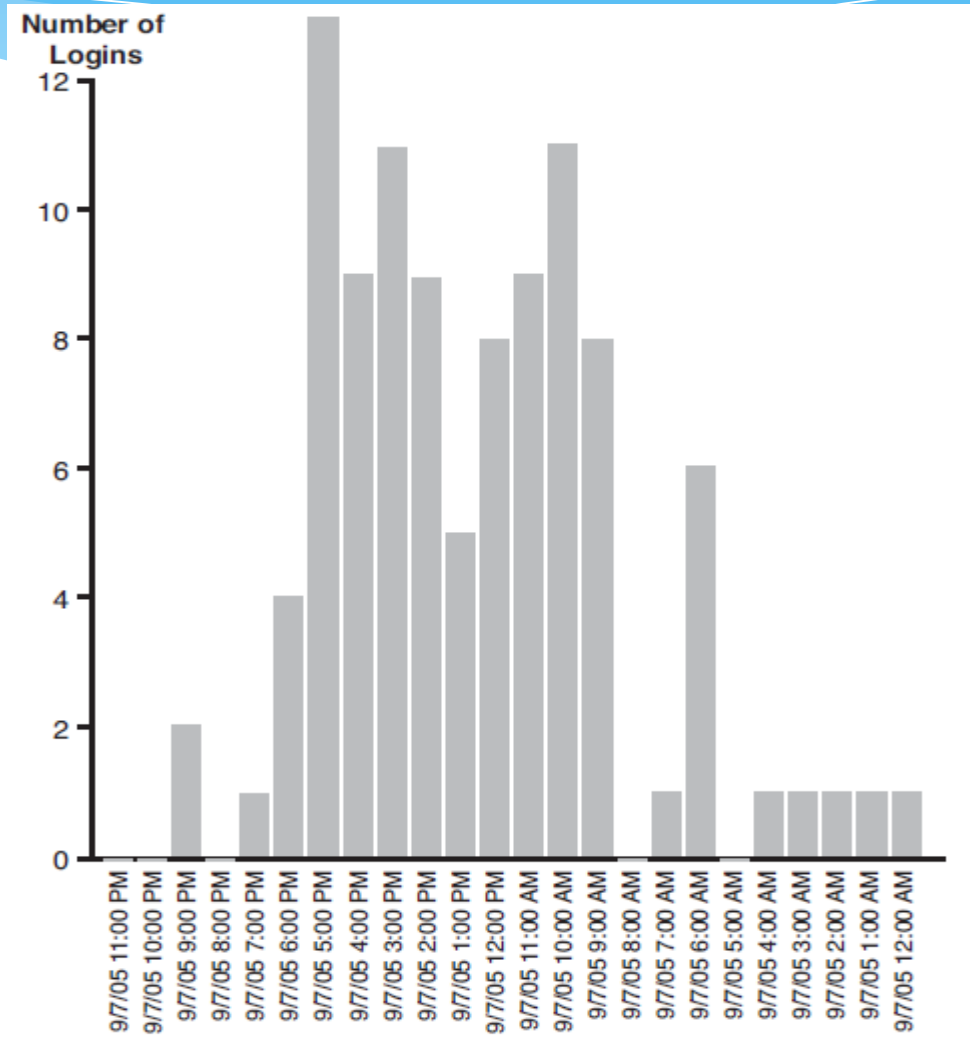
堆叠折形图

和其他堆叠图一样，堆叠折线图也是表示二维的数据，通过增加折线的数量，可以表示不同类型的数据。



直方图

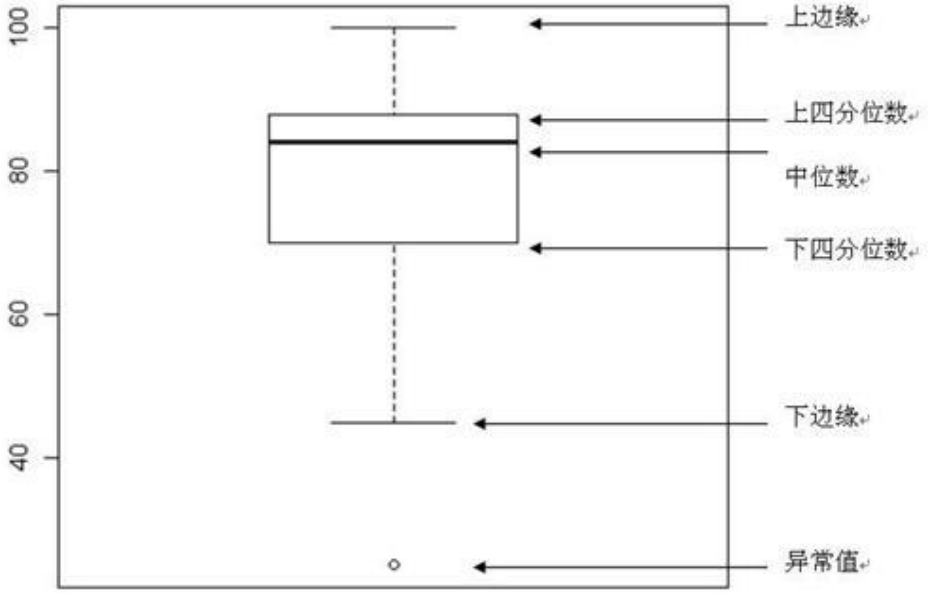
直方图一般用于表示数据的分布，和条形图相比，直方图用于表示连续数据。



箱线图

箱线图是一种用作显示一组数据分散情况资料的统计图。它包含以下几个数:

1. 下四分位数(Q1)
2. 中位数
3. 上四分位数(Q3)
4. 下边缘(Q1-1.5IQR)
5. 上边缘(Q3+1.5IQR)
6. 异常值



箱线图

1. 计算上四分位数（Q3），中位数，下四分位数（Q1）。

2. 计算上四分位数和下四分位数之间的差值，即四分位数差（IQR）Q3-Q1。

3 . 绘制箱线图的上下范围，上限为上四分位数，下限为下四分位数。在箱子内部中位数的位置绘制横线。

4.大于上四分位数1.5倍四分位数差的值，或者小于下四分位数1.5倍四分位数差的值，划为异常值。

5 .异常值之外，最靠近上边缘和下边缘的两个值处，画横线，作触须。

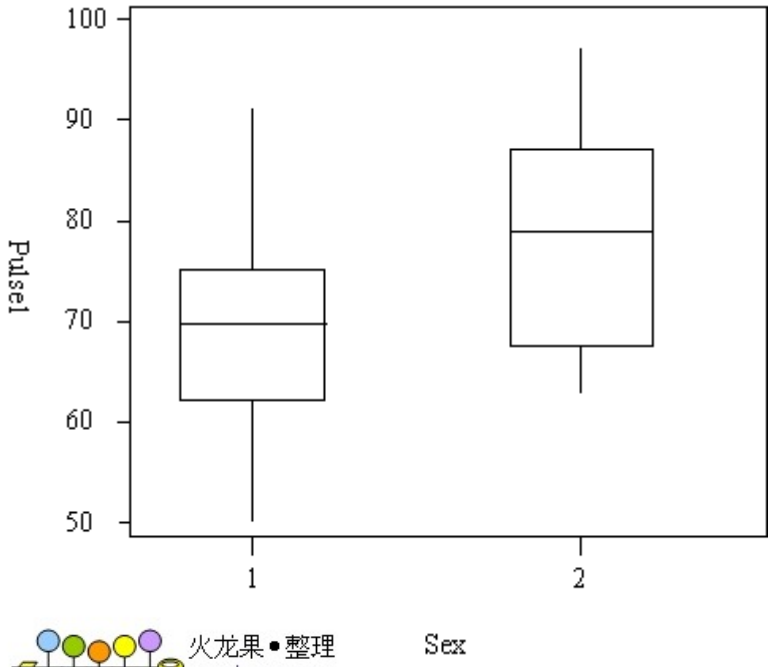
箱线图

箱线图的优势：

1.可以直观地看出数据的离散程度。

2.直观明了地识别数据批中的异常值。

3.利用箱线图判断数据批的偏态。



从该箱线图中可以得到如下信息：

1. 男性（1）的平均脉搏约为70，女性（2）的平均脉搏约为78左右，高于男性；

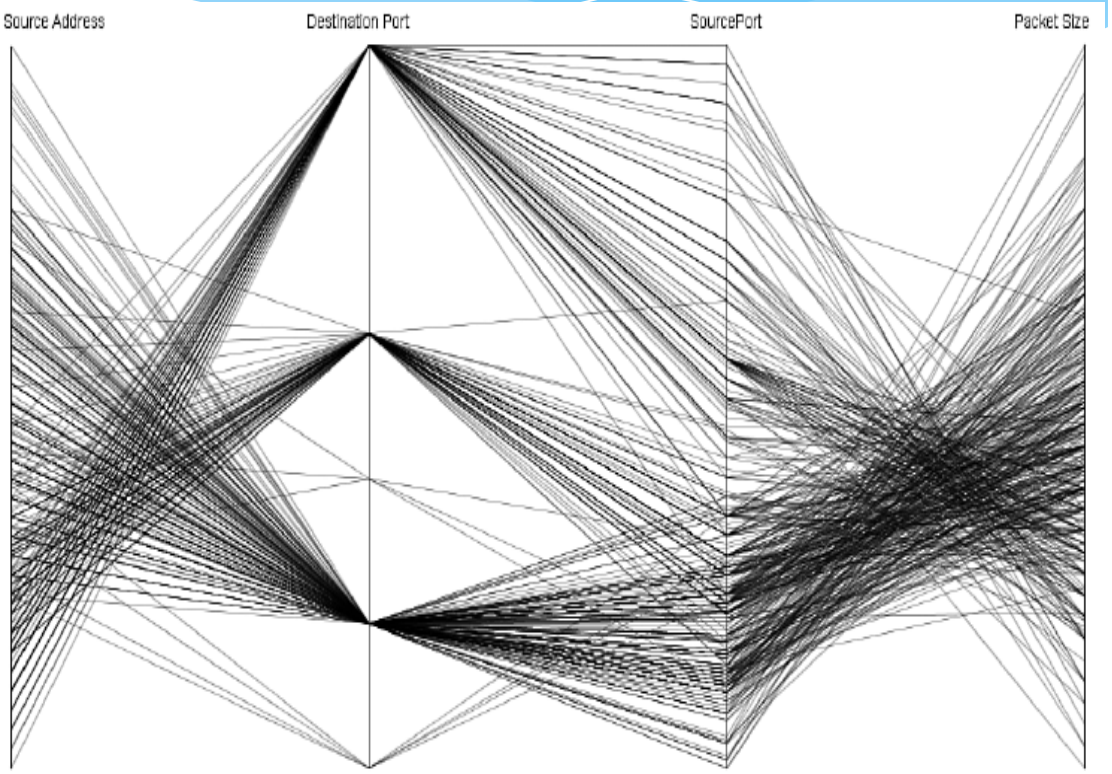
2. 男性脉搏的分布（箱体的高度）较为紧密，女性脉搏的分布比较分散；

3. 最大值出现在女性中，最小值出现在男性中；

4. 两组数据中都没有出现溢出值，表明分布比较正常。

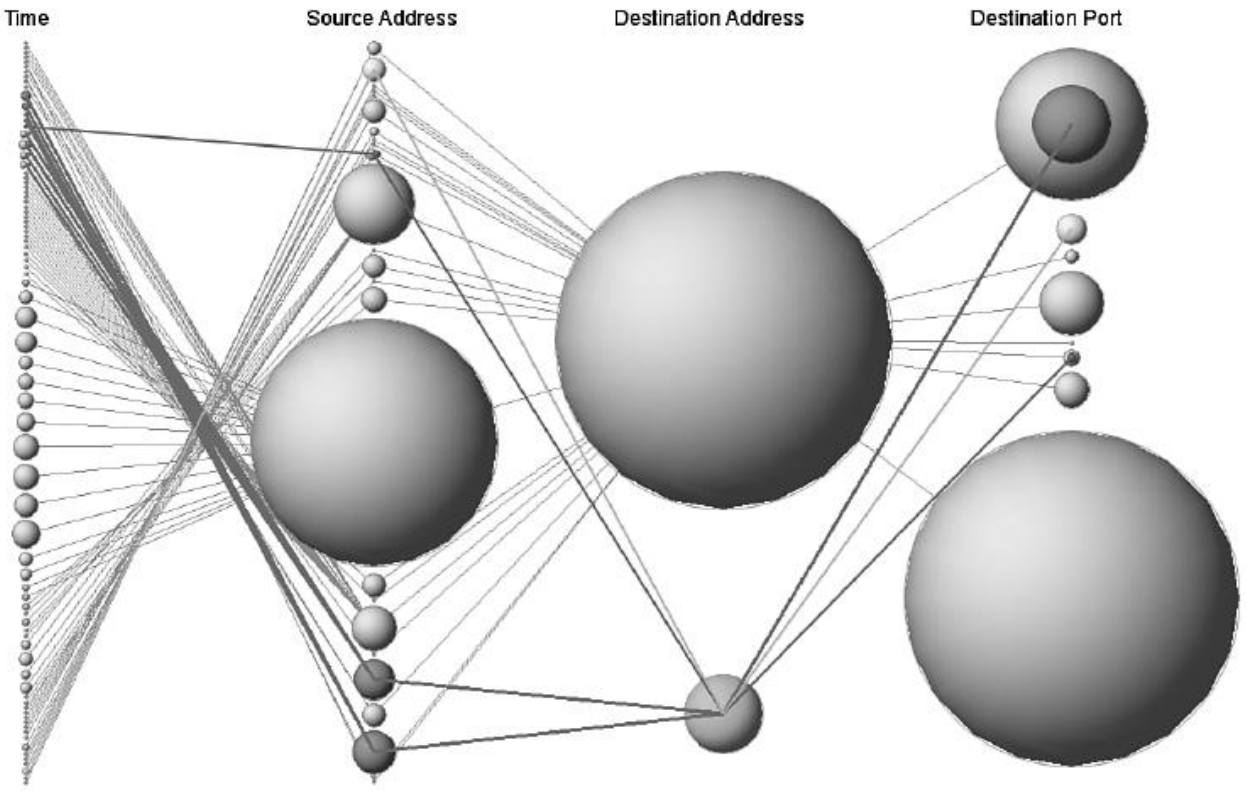
平行坐标

平行坐标用于对高维几何和多元数据 的可视化。为了表示在高维空间的一个点集，在N条平行的线的背景下，一个在高维空间的点被表示为一条拐点在N条平行坐标轴的折线，在第K个坐标轴上的位置就表示这个点在第K个维的值。



平行坐标

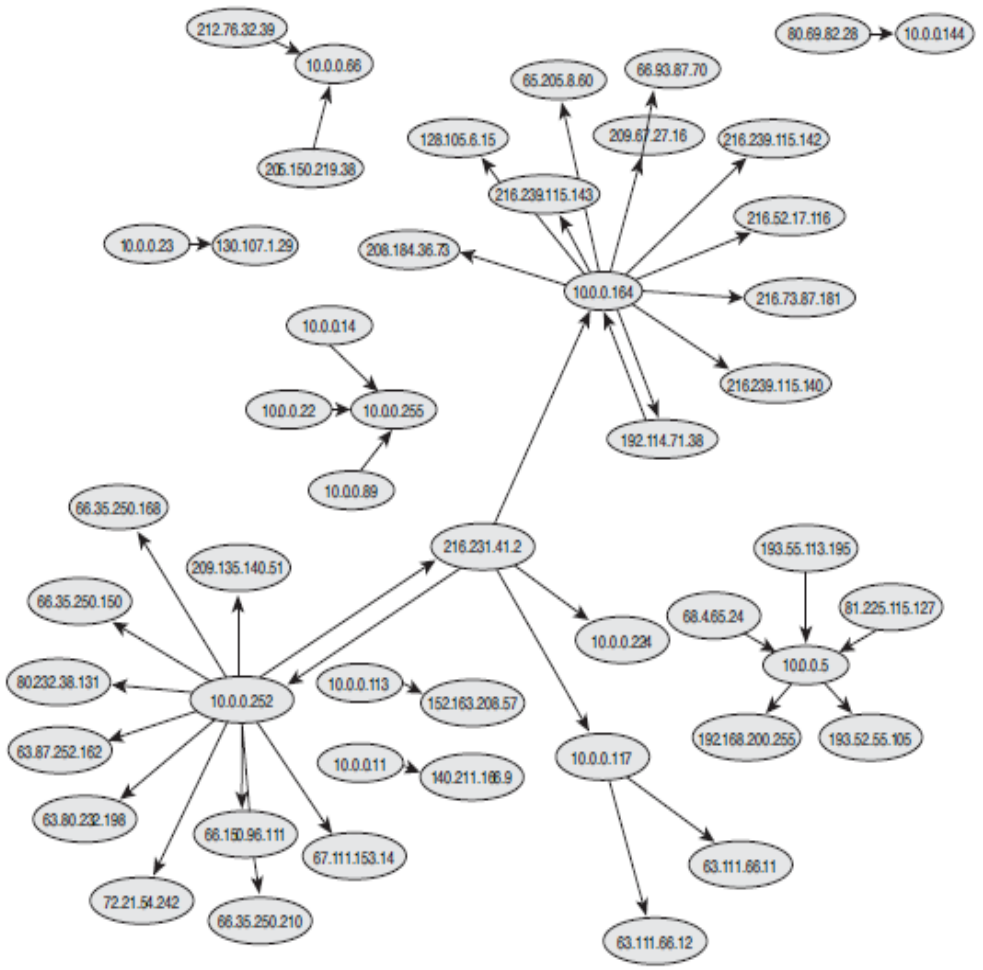
可以对平行坐标进行适当的改变，比如对一个轴上的数据进行统计，然后用不同大小的球来表示数量的不同。也可以对数据进行分组。



连接图

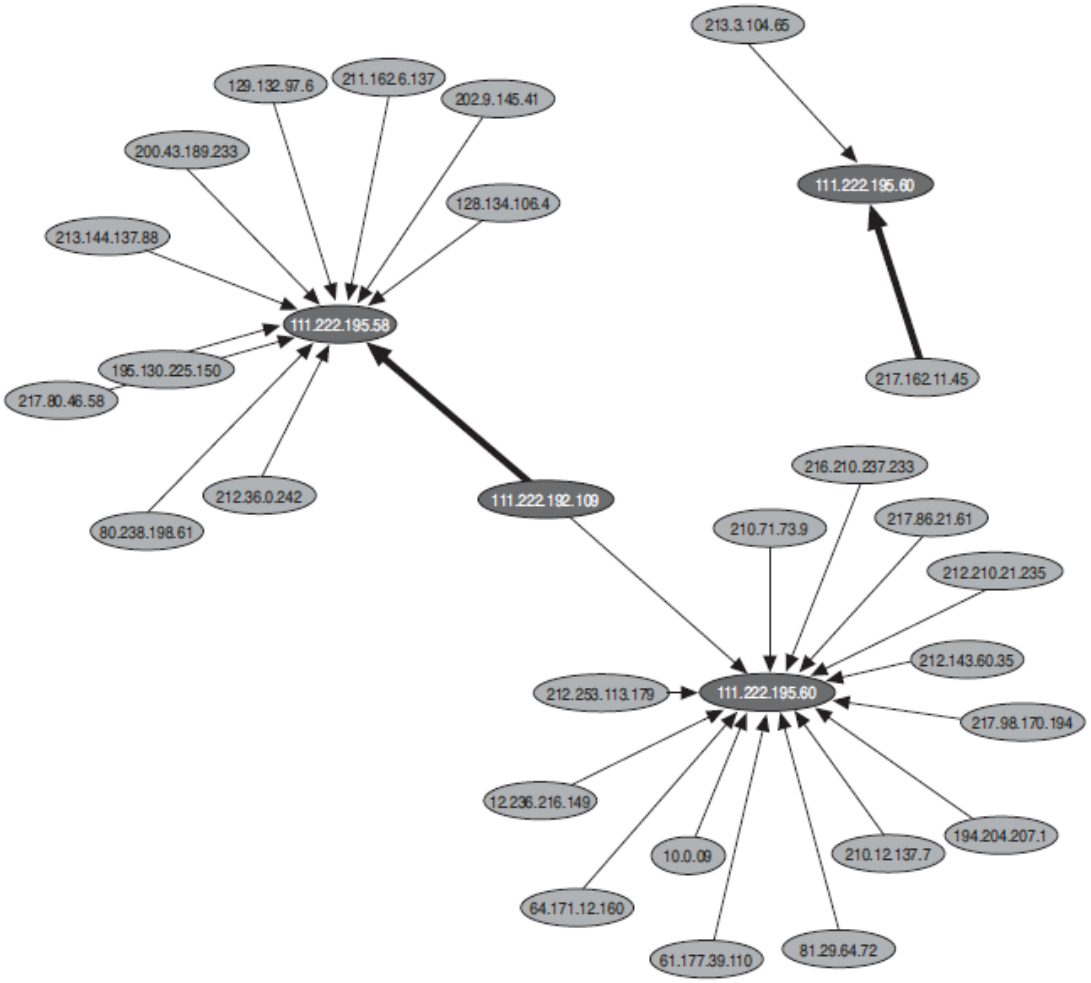
连接图是用结点和边组成的图，它主要用于描述结点之间的数据通信。

通过使用不同的颜色、形状和边的粗细来表示其他的信息。



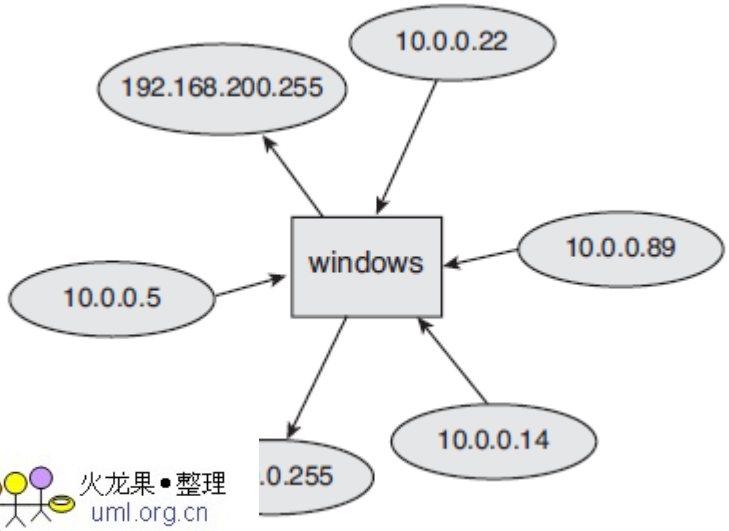
连接图

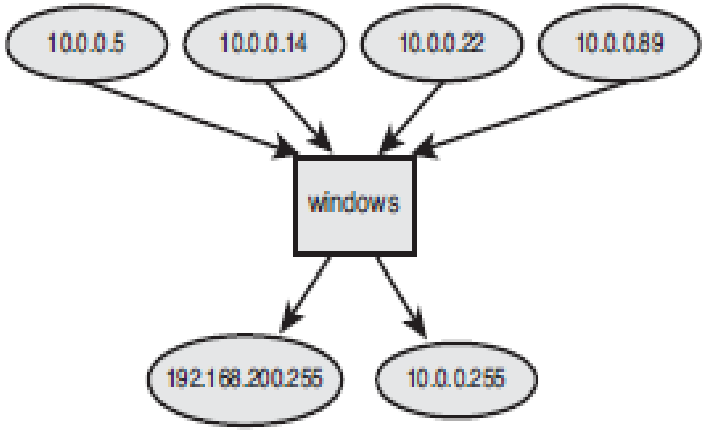
如右图，深色的表示连接到互联网的路由器，而浅色的表示连接到路由器上的电脑。两个结点之间的边的粗细可以表示数据流量的大小。



连接图

我们在画连接图时，有一个非常有挑战性的问题： 结点的层次布局问题。

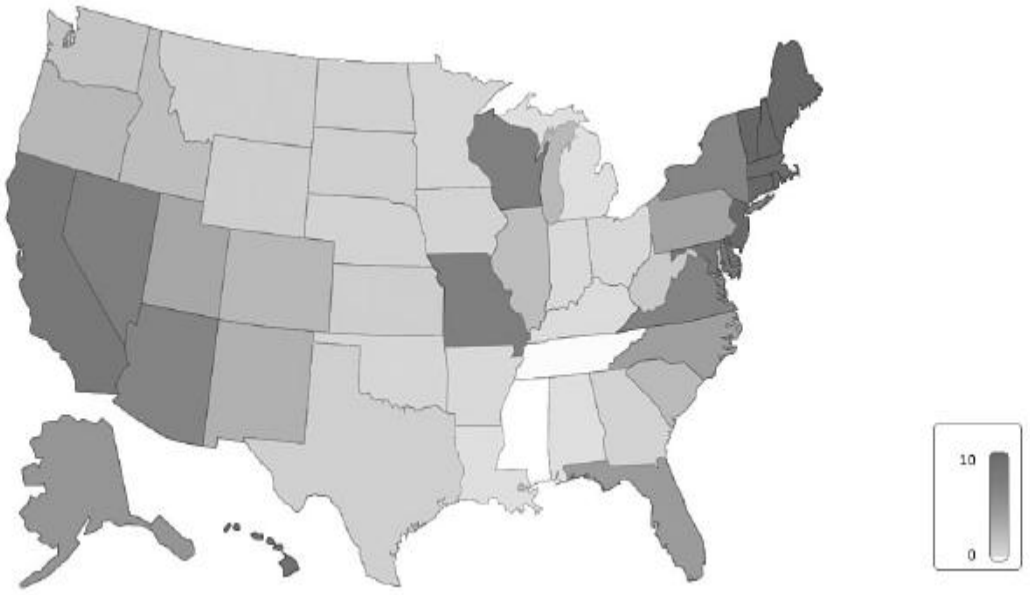




地区图(MAPS)

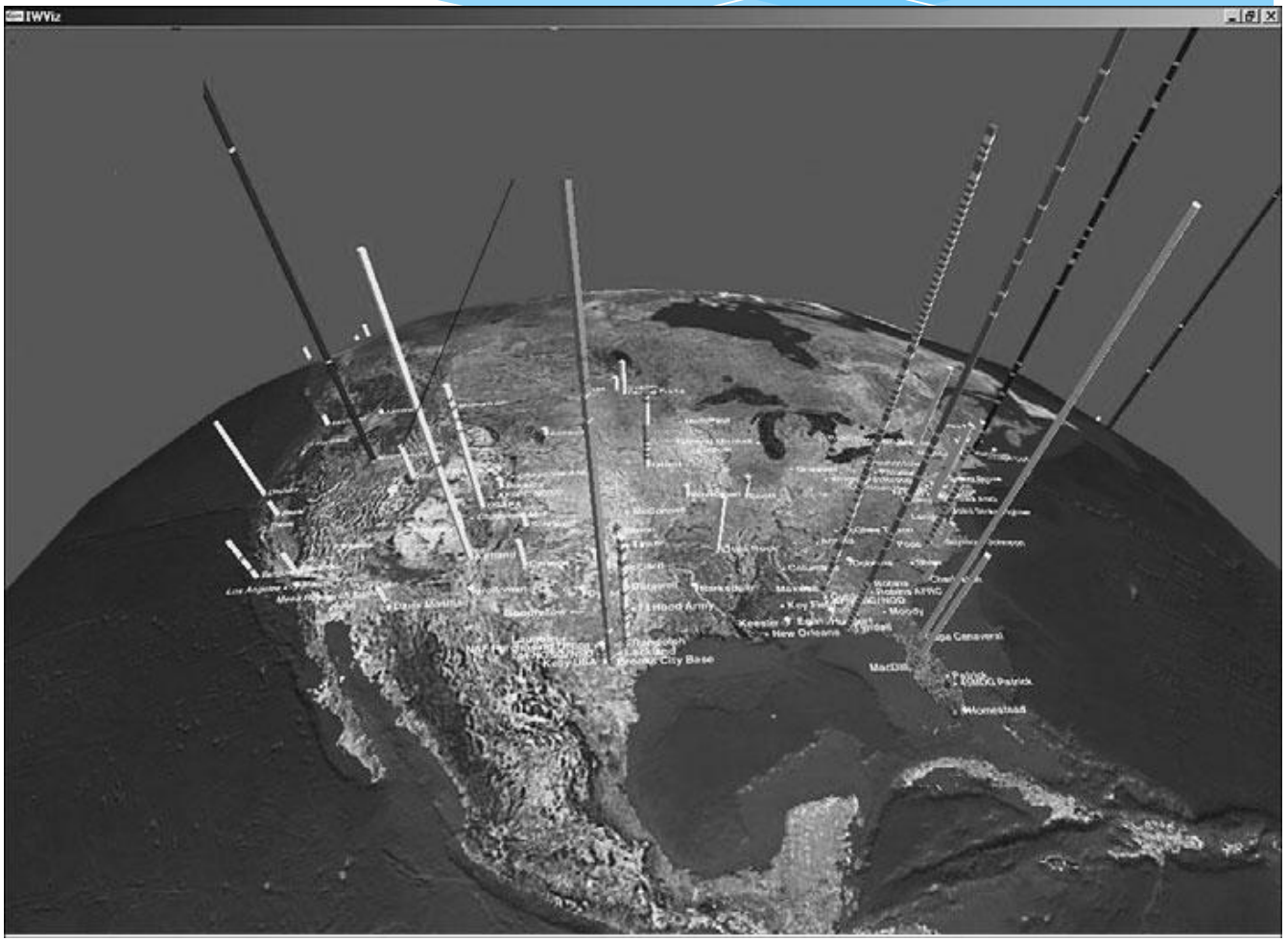
地区图主要用于表示那些与物理地点有关的数据。
如何在地图上表示数据?

右图是使用颜色来表示数据的大小，颜色越深，表示的数据越大



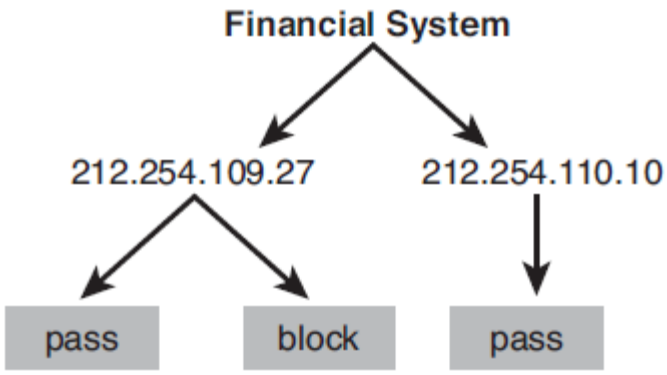
地区图(MAPS)

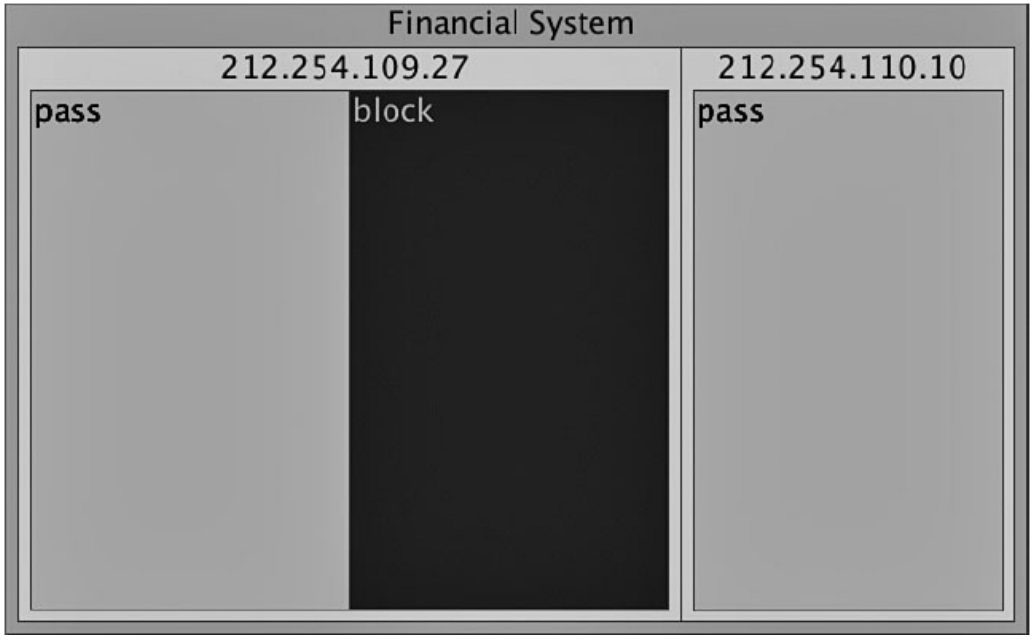
右图是在地图上使用图表来表示数据，每一个积木表示一个事件，而积木的颜色表示事件的严重程度。



TREEMAPS

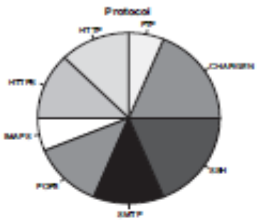
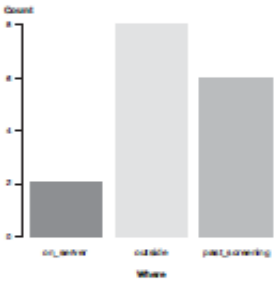
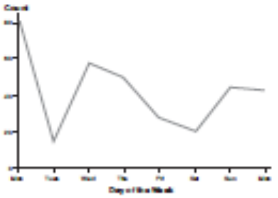
Treemaps能够表示多维、层次结构的数据。

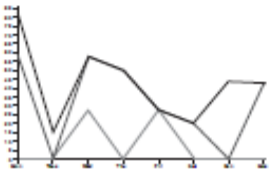


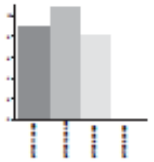
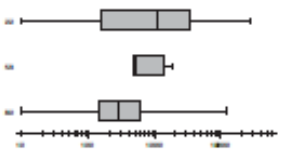
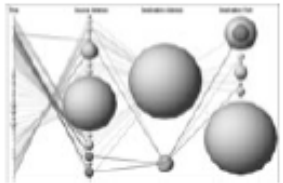


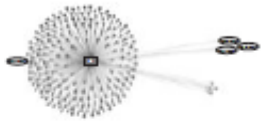
选择正确的图表

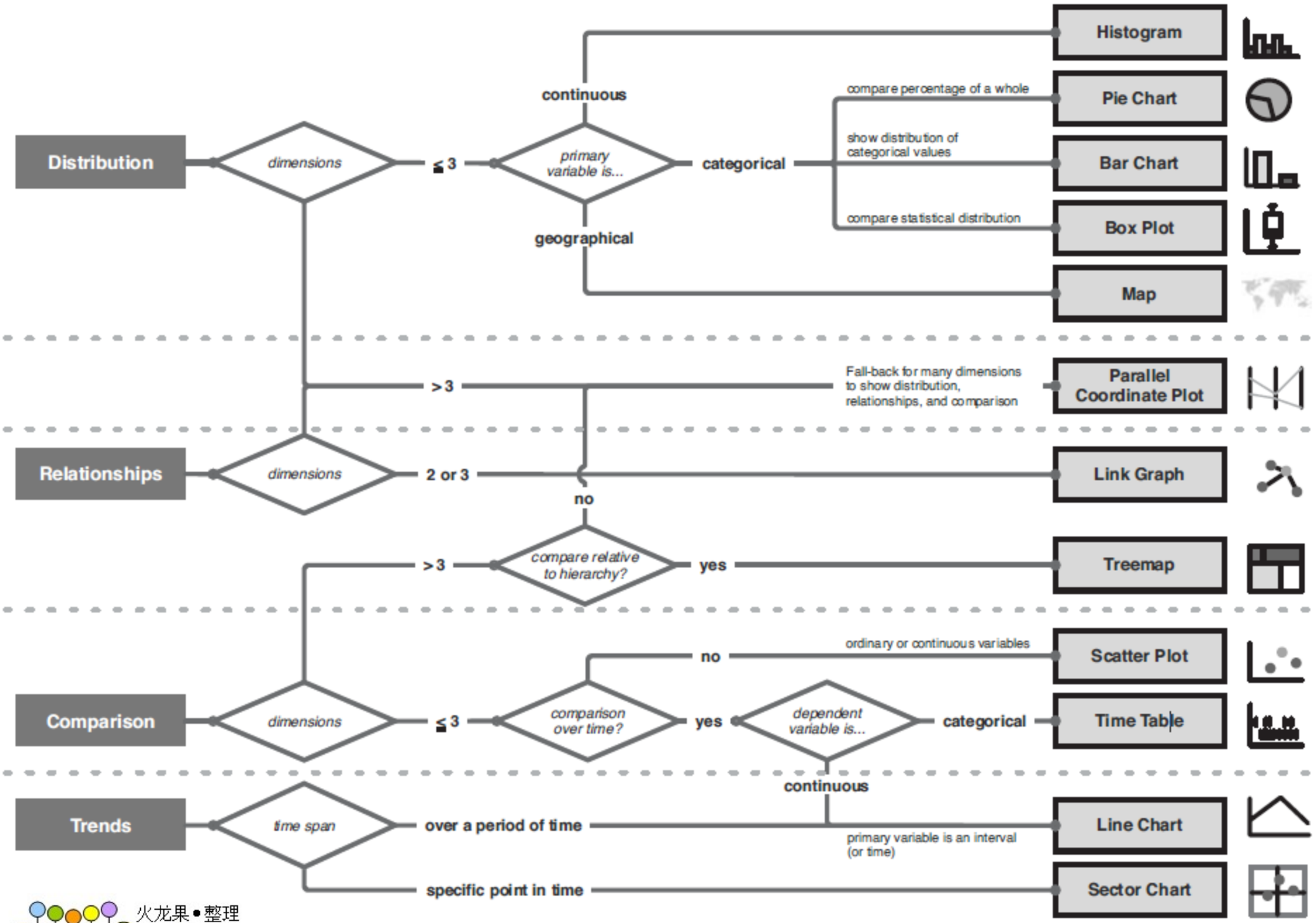
前面介绍了许多图表，但是每个图表都有优缺点，因此我们要根据我们要可视化的数据来选择正确的图表。下面是几个表格，记录了哪些数据应该选用什么图表。

Visualization Technique	Data Dimensions	Maximum Number of Data Values	Data Type	Use-Case	Example Application	Example Chart
Pie chart	1	~10	Categorical	Use to compare values of a dimension as proportions or percentages of the whole.	Proportion of application protocols.	
Bar chart	1	~50	Categorical	Use to show the frequency of the values of a dimension or the output of an aggregation function. Each bar represents a value. The height of the bar represents the frequency count of the value.	Number of bytes transferred per machine.	
Line chart	1	~50	Ordinal, interval	Use to show the frequency of the values of a dimension or the output of an aggregation function. The height of data points in the chart indicates the counts. The data points are connected by lines to help display patterns or trends.	Number of blocked connections per day.	

Visualization Technique	Data Dimensions	Maximum Number of Data Values	Data Type	Use-Case	Example Application	Example Chart
Stacked pie	2	~10 times 5	Categorical	Use to compare values of two dimension as proportions or percentages of each whole.	Based on the role of machines, identify the percentage of protocols used to connect to the machines.	
Stacked bar	2	~50 times 5	Categorical	Use to show the frequency of values or the output of an aggregation function for two dimensions. The chart represents one dimension as the bars. The second dimension is represented as subdivisions in the bars.	For each destination port, identify the role of the machines involved in the traffic. The role is determined by the protocols the machine was using.	
Stacked line	2	~50 times 10	Ordinal or interval for each of the data series	Use to show the frequency of values or the output of an aggregation function for multiple dimensions.	Number of attacks per day across multiple locations.	

Visualization Technique	Data Dimensions	Maximum Number of Data Values	Data Type	Use-Case	Example Application	Example Chart
Histogram	1	~50	Ordinal or continuous	Use to indicate the shape of the distribution of values.	Distribution of number of logins over period of a day.	
Box plot	2	~10	Continuous, categorical	Use to show distribution of values. The categorical dimension can be used to split into multiple box plots for comparison.	Distribution of packet size in traffic.	
Scatter plot	2 or 3	Thousands for each dimension.	Continuous, continuous	Use to examine how two data dimensions relate or to detect clusters and trends in the data.	Show communication patterns of machines by plotting the endpoints along with the destination ports they accessed.	
Parallel coordinates Compare relative frequency per dimension and relationships across dimensions.	*n*	Thousands for each dimension. Up to 20 dimensions.	Any	Use for visualizing multidimensional data in a single plot.	Analyzing firewall rulesets to show for each rule what traffic is affected.	

Visualization Technique	Data Dimensions	Maximum Number of Data Values	Data Type	Use-Case	Example Application	Example Chart
Link graph	2 or 3	Without aggregation: 1000	Any, any	Use for visualizing relationships among values of one dimension and across multiple dimensions.	Identify the impact and extent of a compromise by visualizing communications of the compromised machine after the attack.	
Map	1	100	Coordinates, any	Use to display data relative to a physical location.	Number of trouble tickets per state.	
Treemap	n	10,000	Categorical, any	Use to visualize hierarchical structures in data. Enable comparison of multiple dimensions at once.	Assess risk by visualizing severities and criticalities of vulnerabilities per machine.	



谢谢