



下载APP



## 32 | 数据之美：如何选择合适的方法对数据进行可视化处理？

2020-09-11 月影

跟月影学可视化

[进入课程 >](#)**讲述：月影**

时长 11:20 大小 10.38M



你好，我是月影。

我们知道，可视化包括视觉和数据两大部分。通过前面的课程，我们完成了可视化中视觉呈现部分的学习，学会了用某种技术把数据展现给用户，产生丰富的、生动的、炫酷的视觉效果。今天，我们正式进入数据篇，开始学习数据处理。

你可能会问，学习可视化设计一定要学会处理数据吗？答案是一定要学。因为在可视化项目中，我们关注的信息经常会隐藏在大量原始数据中，而原始数据又包含了太过丰富的信息。其中大部分信息不仅对我们来说根本没用，还会让我们陷入信息漩涡，忽略掉真正重要的信息。



因此，只有深入去理解数据，学会提炼、处理以及合理地使用数据，我们才能成为一名优秀的可视化工程师。

那数据究竟是怎么从原始数据中获取的，又是怎么被我们用可视化的方式表达出来的呢？其实方法有很多，不过这节课我先举三种方法，让你对可视化数据处理手段有一个全面的认知，后几节课我们再深入讲解一些比较通用的数据处理技巧。

## 从原始数据中过滤出有用的信息

首先，我们明确一点，在可视化中，我们处理数据的目的就是，从数据中梳理信息，让这些反应出数据的特征或者规律。一个最常用的技巧就是按照某些属性对数据进行过滤，再将符合条件的结果展现出来，最终让数据呈现出我们希望用户看到的信息。

这么说可能还不太好理解，我们来看一个简单的例子。假设现在有一个小公园，公园有四个区域，分别是广场、休闲区、游乐场以及花园。每天上午 8 点、中午 12 点、下午 6 点以及晚上 8 点这四个时间，公园管理处会通过航拍收集 4 个区域上人群的分布信息，得到每天人群分布的数据之后，公园管理者就能够利用这些数据来优化公园的娱乐设施了。

具体该怎么做呢？利用可视化来解决这个问题会非常简单，思路就是先把人群的分布数据绘制成合适的可视化图表，从中分析出人群分布的规律。

这里，我仿造了一组人群数据，将它放在 [🔗 GitHub 仓库](#)里，数据格式如下：

 复制代码

```
1  [{
2    "x": 456,
3    "y": 581,
4    "time": 12,
5    "gender": "f"
6  }, {
7    "x": 293,
8    "y": 545,
9    "time": 12,
10   "gender": "m"
11  }, {
12   "x": 26,
13   "y": 470,
14   "time": 12,
15   "gender": "m"
16  }, {
```

```
17   "x": 254,  
18   "y": 587,  
19   "time": 12,  
20   "gender": "m"  
21 }, {  
22   "x": 385,  
23   "y": 257,  
24   "time": 8,  
25   "gender": "m"  
26 },  
27 ...]
```

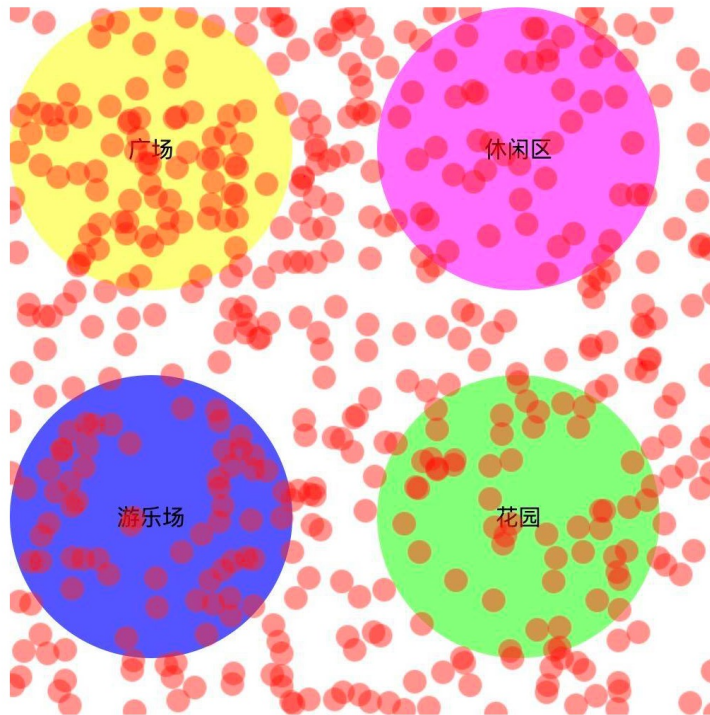
数据是 JSON 格式，数组中的每一项表示一个游客，x、y 是拍摄位置，time 是时间，gender 是性别。

想要表现人群分布的规律，我们可以用这个数据来绘制一个散点图，它能非常直观呈现出原始数据。绘制方法非常简单，就是根据 x、y 坐标将一个小圆点标记在公园的某个位置上，代码如下：

[复制代码](#)

```
1 function draw(data) {  
2   const context = canvas.getContext('2d');  
3   context.fillStyle = 'rgba(255, 0, 0, 0.5)';  
4   for(let i = 0; i < data.length; i++) {  
5     const {x, y} = data[i];  
6     context.beginPath();  
7     const spot = context.arc(x, y, 10, 0, Math.PI * 2);  
8     context.fill();  
9   }  
10 }  
11  
12 fetch('data.json').then((res) => {  
13   return res.json();  
14 }).then((data) => {  
15   draw(data);  
16 });  
17
```

最终绘制出来的效果如下图所示：

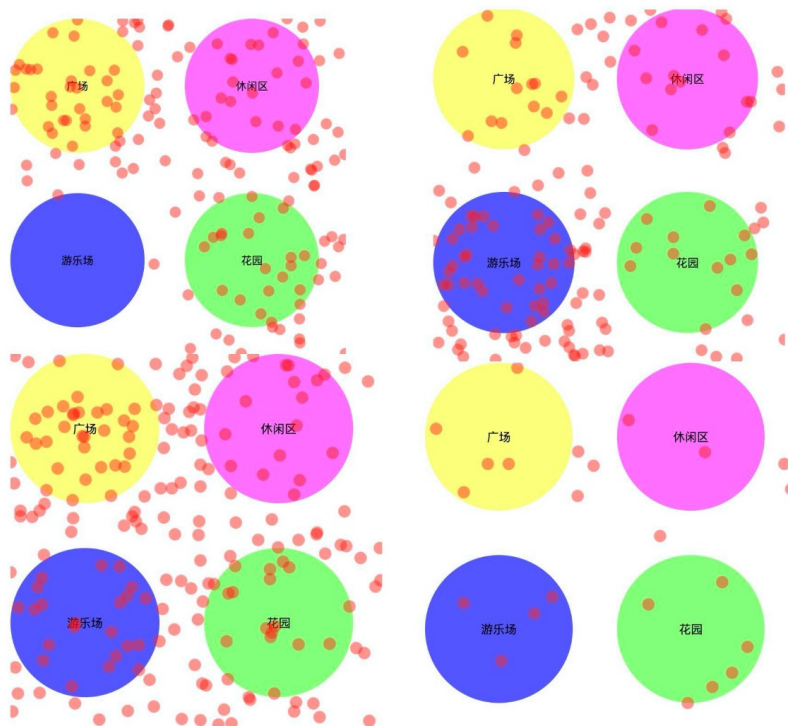


我们可以看到，这样绘制出来的分布图上显示了每天访问公园的人群，他们很均匀地分散在公园各处，似乎并没有什么特殊的地方。这其实是因为我们并没有根据其他属性来过滤这些数据。我们可以先试着根据时间来过滤。我们修改一下代码，给 draw 方法添加一个过滤函数。

[复制代码](#)

```
1 function draw(data, filter = null) {
2   if(filter) data = data.filter(filter);
3   const context = canvas.getContext('2d');
4   context.fillStyle = 'rgba(255, 0, 0, 0.5)';
5   for(let i = 0; i < data.length; i++) {
6     const {x, y} = data[i];
7     context.beginPath();
8     const spot = context.arc(x, y, 10, 0, Math.PI * 2);
9     context.fill();
10  }
11 }
12
13 fetch('data.json').then((res) => {
14   return res.json();
15 }).then((data) => {
16   draw(data, ({time}) => time === 8);
17   // draw(data, ({time}) => time === 12);
18   // draw(data, ({time}) => time === 18);
19   // draw(data, ({time}) => time === 20);
20 });
```

把数据按照 8 点、12 点、18 点、20 点分别过滤之后，我们就能得到不同时间的游客散点图，如下图所示。



我们先看左上角，也就是 8 点钟的时候，游客大部分会集中在广场，结合这个时间点，他们可能是在晨练，而游乐场几乎没有游客，因为 8 点的时候游乐场还没开始营业。接着我们来看右上角，也就是中午 12 点，这个时候游客大部分集中在游乐场，说明此时是游乐场的高峰时间。然后是左下角，18 点的时候，你会发现一天中这个时间的游客数量是最多的，并且，集中在广场上的游客也最多，我们推测他们正在进行健身活动。最后，右下角也就是 20 点，这个时候公园临近关门，所以游客已经很少了。

就这样，我们得到了不同时间段，游客集中活动的场所。接下来，我们可以再把性别这个属性加上，看看还有什么分布规律，修改后的代码如下所示。我们用蓝色标记男游客，用红色标记女游客。

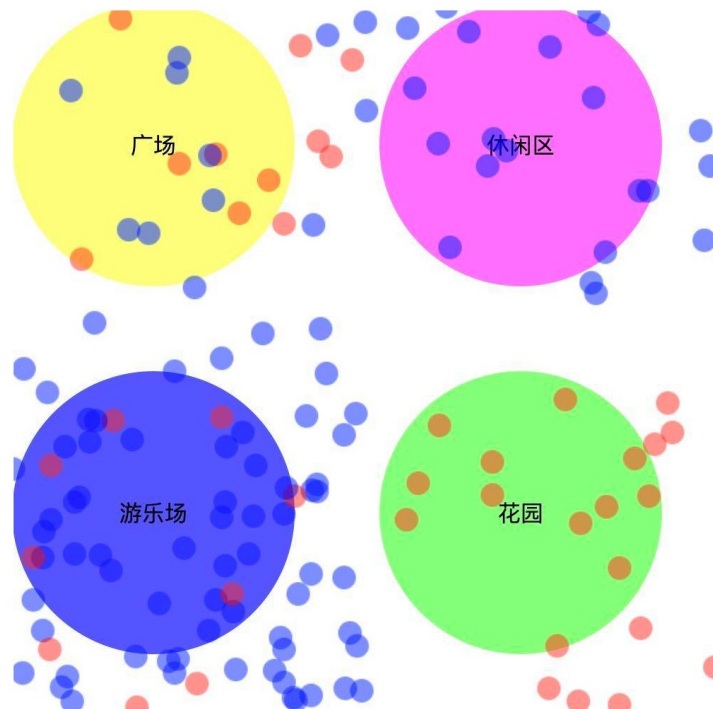
[复制代码](#)

```
1 function draw(data, filter = null) {  
2   if(filter) data = data.filter(filter);  
3   const context = canvas.getContext('2d');  
4   for(let i = 0; i < data.length; i++) {  
5     const {x, y, gender} = data[i];  
6     context.fillStyle = gender === 'f' ? 'rgba(255, 0, 0, 0.5)' : 'rgba(0, 0,  
7     context.beginPath();  
8     const spot = context.arc(x, y, 10, 0, Math.PI * 2);  
9     context.fill();
```



```
10    }  
11  }  
12  
13  fetch('data.json').then((res) => {  
14    return res.json();  
15  }).then((data) => {  
16    draw(data, ({time}) => time === 12);  
17  });  
18
```

标记完我们再来看一下 12 点的游客散点图。



我们看到，集中再游乐场和休闲区的主要是男游客，而女游客更喜欢呆在花园。这可能是和游乐场的游乐设施以及休闲区的设计有关。

到这里，我们关于公园游客的可视化分析就告一段落了。通过我们分析得出的规律，对游乐场改进游乐设施和日常管理是有实际的参考作用。

因为这里的数据是我仿造的数据，所以不一定符合真实情况，不过这并不重要。通过这个例子，我主要是想让你体会数据可视化分析的一般过程，通常我们通过数据过滤和展示，从中提取出有用信息，以便于做出后续的决策，这就是数据可视化的价值所在。只要数据是客观的，分析过程是合理的，那数据表现出来的结果就是具有实际意义的。

## 强化展现形式让用户更好地感知

在前面的例子里，我们用散点图呈现游客信息并从中分析出有用的内容，这种形式直观有效，但是展现形式略显单调。除了合理的数据分析以外，数据可视化有时候通过强化展现形式，让用户更好地感知数据表达的内容。这样能够帮助需要关注该数据的用户，更好地把握整体信息。

我在 GitHub 看到一个非常合适的例子，我们来看一下。

空气质量和我们生活质量息息相关，那在过去的几年里，雾霾和蓝天交替，成为我们生活的一部分。近几年来，国家一直在大力治理空气污染。那在这种情况下，我们的空气质量到底有没有变好呢？

一名 [@亚赛](#) 同学写了一个 [@北京空气质量 \(2015-2018\) 的可视化展现](#)，利用每天北京空气的 AQI 值绘制了色条，他还用心地让每一天对应了一个地标的当天实拍照片。这不仅增加了项目整体的趣味性，也强化了用户的直观认知。



北京空气质量 (2015-2018) 可视化展现 (图片来源: wangyasai.github.io)

如上面示意图所示，我们将每一年的数据按照 AQI 排序后，可以明显看出灰色的区域在逐年减少，这说明北京的空气质量的确是在逐年好转的。

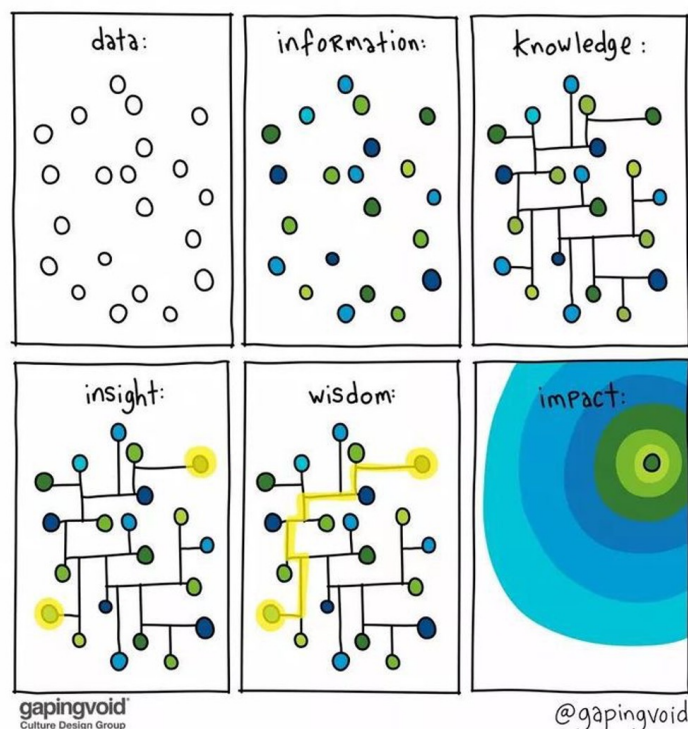
这个项目的代码在亚赛同学的 [GitHub 仓库](#)里，我就不拿出来细讲了。因为这个可视化效果的实现原理并不复杂，而且这节课我们更应该学习和理解用数据进行可视化展现的思路，而不是代码实现细节。当然，如果你想搞懂代码，那你可以深入分析一下 GitHub 仓库里的源码。这个项目代码具体实现是依赖一个叫做 p5.js 的图形库，它也是一个很棒也很有趣的图形库，用来学习可视化也非常合适，如果你有兴趣可以去看一下。

如果你想要亲自动手实现一个这样的可视化项目，我也建议你借鉴亚赛同学这个项目中的数据和思路，对它进行一些改进。

## 将信息的特征具象化

到现在为止，你可能认为可视化都是需要使用真实数据来呈现的，数据越真实、越详细，可视化效果呢就越好。如果有了这个想法，说明你有一点陷入到思维定式中了。实际上，有时候我们并不要求数据越真实越详细，甚至不要求绝对真实的数据，只需要把数据的特征抽象和提取出来，再把代表数据最鲜明的特征，用图形化、令人印象深刻的方式呈现出来，这就已经是成功的可视化了。

比如下面这张图，就用可视化的方式解释了数据、信息、知识、见解、智慧和阴谋论。这种可视化呈现的数据并不是真实、准确地，而是带有趣味性的，通过对信息特征进行抽取，让看的人形成了一种视觉认知。

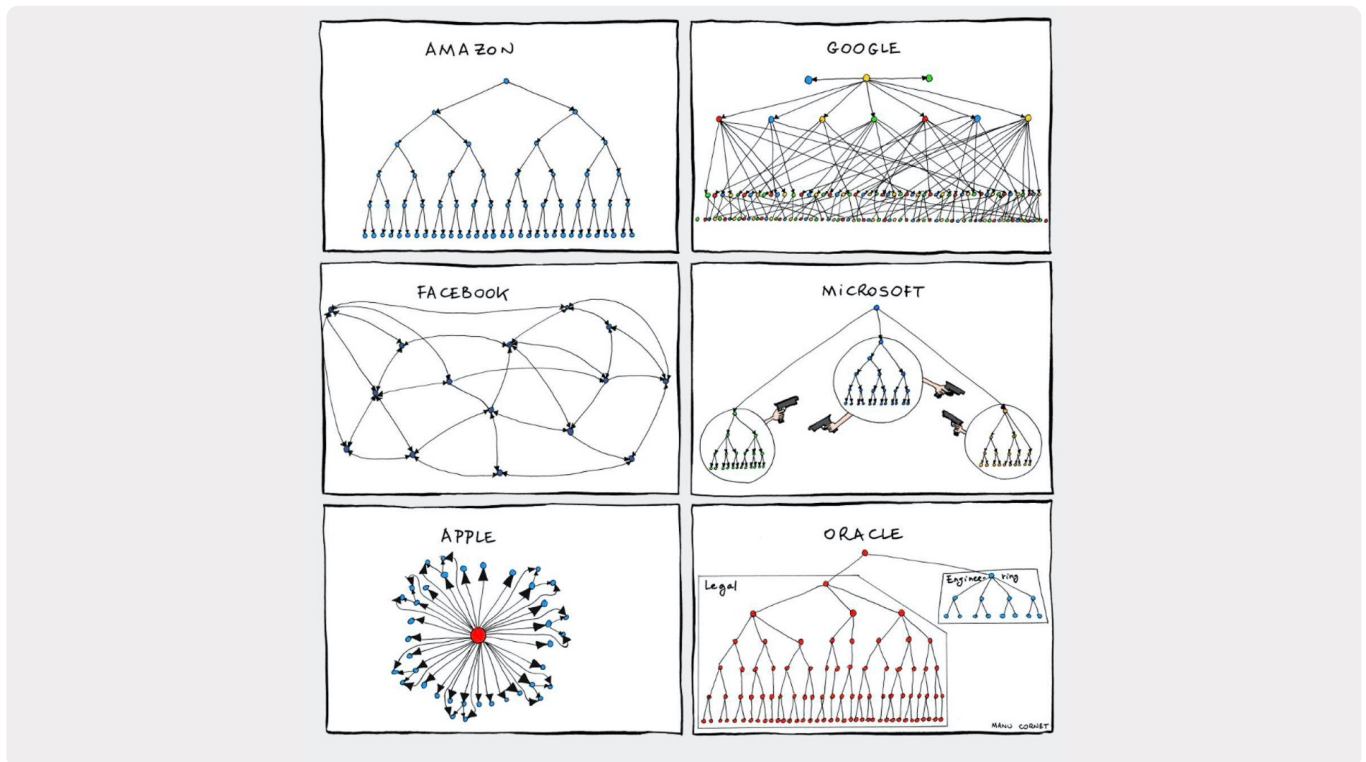


数据-信息-知识-洞见-智慧-影响力(图片来源: gapingvoid)



其实这样的可视化例子还有很多，比如 Matt Might 教授绘制的 “[图解博士是什么](#)” 也非常有趣。由于图片很长，我在这里就不列出来了。虽然它的数据不是基于海量数据提取的，但却是一组概念的具象化，所以它毫无疑问也是一个非常成功的可视化方案例。

除此之外，Manu Cornet 的组织架构图，也用非常形象的方法绘制出了各个知名公司的组织架构差异。它的数据当然也不是各个公司详细的组织架构数据，而是根据每个公司组织架构特征直接图形化形成的。



(图片来源: bonkersworld)

看了前面这些例子，你可能会有疑问，第三种方法似乎和原始数据并没有关联，而是直接用信息特征来完成的可视化，那第一、二种方法和数据处理过程和它又有什么关系呢？

实际上，我们使用第三种方法，也就是信息特征具象化的前提，是我们真正掌握了我们需要的信息特征，而这些特征的提取和掌握，正是通过前面两种方法迭代出来的！用一句话总结就是，数据可视化本身是一个不断迭代的过程。

具体过程是，我们先进行原始数据的信息收集和分类处理，再通过原始方法表达出有用的信息，接着通过强化展现形式，让信息的核心特征变得更加鲜明，经过这一轮或者几轮的迭代，我们就可能拿到最本质的信息了，最终我们再把这些信息具象化，就可以达到令人印象深刻的效果了。

所以，对原始数据进行不断迭代，就是数据可视化的基本方法论。我希望你能牢牢记住这句话，并且在实践中认真去做。

## 要点总结

这节课，我们主要讲了对数据进行可视化处理的三种常见方法。

第一种，是从原始数据中过滤出有用的信息。这是数据可视化处理的第一步，也是最基础的方法。第二种，是强化数据的展现形式，让用户更好地感知我们要表达的信息。这是我们在第一步的基础上对数据进行的加工处理。而第三种，是把数据的特征具象化，然后用图形表达。这是我们在第一、二步的基础上，对数据进一步的抽象和提取。如果达到这一步，我们甚至有可能完全脱离原始数据，不依赖原始数据，而是着眼于数据特征的表现形式。

这三种方法层层递进，是数据可视化的基本方法论，而数据可视化本身，其实就是使用这些方法对数据信息进行不断迭代和构建的过程。


## 小试牛刀

你能借助我们前面说的北京空气质量，这个可视化例子中的代码，实现一个你想要的可视化展现吗？你可以不完全按照亚赛同学的方法，多加些自己的创意，以及我们前面学过的图形学技巧，相信你能做出非常好的效果。当然了，我也知道这个挑战有点难，但整个实现的过程能让你把学到的图形学知识融会贯通，所以我还是建议你尝试一下。


欢迎在留言区和我讨论，分享你的答案和思考，也欢迎你把这节课分享给你的朋友，我们下节课见！


---

## 源码

 课程中完整实例代码

## 推荐阅读

[1]  北京空气质量（2015-2018）的可视化展现

[2]  图解博士是什么

提建议

更多课程推荐

# 程序员的数学基础课

在实战中重新理解数学

黄申

LinkedIn 资深数据科学家



涨价倒计时 

今日秒杀 **¥79**, 9月11日涨价至 **¥129**

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 31 | 针对海量数据，如何优化性能？

下一篇 加餐一 | 作为一名程序员，数学到底要多好？

## 精选留言

 写留言

由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。

