### 大数据可视化与可视分析-李国政-第10周 星期四 第4大节.mp4

00:00  
谁知应该。

他容易出错，谁知好，谢谢。改变了。这个标准，谁知谁知好。

01:08  
的各位采访我们在上课。

01:11  
首先我想问一下大家现在组队情况咋样？有考虑过吗？就是有对的举手还是不少，还是感觉有一半人。大家决定好我们做什么题目了吗？

01:36  
我在群里发了一个链接，就是一个腾讯文档，因为我们可以在这里问一下，

01:43  
有没有选择。

01:47  
好对应或者那种情况。

01:49  
从对这种情况。

01:52  
的话就是大概方向，就像说我打电话，或者你遇到什么困难吗？因为困难就跟我们没有把握，能够找到相对应的还是没有把握找的。

02:05  
数据一样有。

02:07  
个什么特点，最终的情况的小问题不可能这样的，我们西部的这个是吧？人口新增人口有多少，我们要考虑的其实是什么？城市。

02:21  
我们的迁出人口的全部的出了一些新发展战略，但是确实还要考核他们找到一个相应的基础。

02:29  
是你说的数据确实是一个很重要的方面，就是你很多时候要先从数据出发，你看你能找到什么数据，然后再说这个内容要去做什么事情。

02:41  
不可以。

02:43  
如果没有一个明确，其实我们选择数据的时候也可以选择那些特别繁琐，要整理很长时间才能得到的数据，这样也会后面你也会觉得调整的不同。

02:58  
明白。

证明的。

03:03  
也可以。

03:04  
我觉得也没问题，领导自己地方什么人群就大概。

03:10  
的保障没有。

03:12  
可能是我们的谁知还有谁知也是我们打造我找到了也是人情关系，所以我们以后也可以交付一个情况，

03:26  
他要签这种数据信息的，我们可以解决这种。

03:32  
然后这个问题的沟通情况也不可能有做了。

谁知。

03:41  
我记得我之前的时候有一个本科毕设的东西做了一下北京市房价变化。

03:48  
情况的这样一个运作。

03:50  
预测未来的变化情况，然后发现前面的变化情况。

03:55  
你看又到了谁知

04:07  
我们之前说的是这周你要把你的队伍好，对不对？

然后你尽可能的要把你的主题确定下来，哪怕不是一个特别确定的主题，一个大概的方向也可以，然后你可能找一些可选的数据，然后我发在了那个群里，然后大家至少这周我看到把大家所有的人都找到一个队，然后填上去好吧？

我发到群里的链接大家都看到了，这样的话我这样做的目的是为了大家不要在最后的时候，你们觉得等到最后我们啥要去看了，这时候大家才开始做，你们也会觉得很慌乱。好的。然后这是第一件事情。

第二件事情是我们上节课介绍了怎么去编程，对不对？大家后续有没有什么问题？有吗？

05:12  
有。

05:12  
问题的话举手可以说一下。我们上节课看了什么？第一个就是我们首先去。先去说原始的数据怎么变，怎么映射到视觉元素上，它是基于迪斯瑞的就是三个小球。第二个就是基于这三个小球，我们去做一个简单的概括，就是它有轴，然后它有视觉映射的方式，然后他就能够这是一第一个第二点。

第三点就是我们基于该报去做一个该报的没事，然后它可以去交互，它支持你来去选择，然后上节课的时候我也给大家留了一个课下的小练习，你们可以自己的来去做一下，比如说选择某一次get boss的某一部分，然后我可以把选择的那部分它的分布能够画得出来，这样的话你能够做到这一点，我觉得你已经能够去做其他事情就可以自己想到什么就可以做什么。

我上节课之后我就把那个代码发到了群里，大家有看吗？有看吗？

06:28  
谁知。

06:33  
我们如何基于Skype到他们去，然后去画一个柱形图。

06:40  
对不对？

06:42  
你找到单位产品，好好，

06:46  
谁知谁知什么？

谁知。

07:06  
你刷出来之后，你看到82是选中了多少？82，它的代码我忘了，在周二的时候也发到了群里面，其实它是基于我们课上讲，我们这边的报没去死代码的去扩展额度的，这个是它的代码，它无非它增加的部分就是下面做包差的这些部分，然后他首先去选择这些数据，选中的数据拿上来，然后来去调用这个函数。

这个函数我觉得也很一样的，所有的都是一样的，先去得到它的比例，确定它的数据，然后对数据验做变换，确定它的比例尺，然后再去把数据去映射到视觉元素上，然后把它画出来，对，然后就得到这个结果对吧？当然没有问题对不对？

有问题的话我们可以现在来说好的，我是觉得如果大家能够做，如果现在没有问题的话，我觉得应该你们去做大作业就没有问题了，这个很简单的事情。

08:19  
我可以然后我们。

08:27  
来去我们继续讲我们今天的话题。

08:31  
谁知我们上。

08:37  
节课我们主要是讲的迪斯瑞对不对？

我们上节课在开始的时候说了，其实主要编程有两个方面的两两个类型，第一个类型是并不是的编程，它决定的是从数据到文化的构建步骤，当然第一次也不是一一蹴而就的，一开始就得到了第四位它也有一个逐渐的发展过程，一开始的时候它可能是模板式的，我们只是套用一个模板让它构建起来就比较固定，难以去拓展。

第二个步骤就是我们发展到我可以去构建这个判断，因为它判断也是从数据到可视化的步骤，对不对？这样的话我就可以去控制PAD里面的每一个模块，然后提供一些支持，提供了支持之后，我就可以来去进一步的去有更好的灵活性。

但是帕克拉它也有一个问题，就是我的我的可控的力度还是很粗，这时候我没办法去支持非常广泛的十二形式，这时候就有人思考，因为可视化它最重要的源是视觉图源对不对？视觉元素，这时候我就去控制这些视觉元素的映射方式，然后来去最终来去得到表达形式。

所以就得到了不是说位置不是位置是基于Java的，然后他也是去将数据绑定到数据源路上，然后进一步的他说Java其实不好调控，而且不易于传播，这时候随着外部编程的发展，然后大家就把视觉元素去映射到外部编程里面的倒退上，然后这样的话你就可以很灵活，同时你也很适合传播你的可视化结果，然后就要这样就发展到了第四位。

但是这些都是去控制可视化构建的过程，对不对？然后后面构建过程，我们上节课学了之后，你会发现迪斯瑞它其实代码量也很大，我也需要去学习很多对于不会编程的人来说，其实也是比较困难的一件事情。

同样还是斯坦福的那一群人，他们就想到我不去控制它的步骤去控制它直接得到的结果，我就是通过一些描述式的方法描述我构建的结果，是因为可视化通过这些比如说8块，我直接去描述数据，从什么映射到了什么，那不就构建得到了可视化结果结果了，所以他们就想到了这样的方式，最经典的就是 we got it同样是一个很有影响力的工作。

其实我觉得威刚艾特的影响力不在于说不像DC那样，我支持了很多很灵活的表达形式的构建，而威刚I它其实支持了很多的对于可视化的研究的工作，比如说文化的推荐，或者我们去做一些人机协同的去构建可视化，这时候威刚来它的作用就会很强大，因为它是基于尖子来去直接去生存可视化的。

他的4个作者同样现在 open的，他现在是MIT之前的版图，然后装备他现在在编u然后这个人他名字特别难读，然后他现在也在1个可视化的1个公司里面，然后直接被列入是第四位的作者，这4个人同样他们也去提出了描述式的语法障碍空间。

 We got it它同样有一个非常优秀的社区。

他的工作是这样的，比如说8块我要去画8块，我们刚才也看到你画8块的话，因为迪斯瑞你会需要写很大很多的代码，但是你用微钢来的只需要去写这样一个真的文件就可以了，非常简单只有几行，而且写这样的文件其实是不需要有编程能力的人来去做的，它的检测文件里面有这样的一些模块，就是首先它的密码是啥？

它叫维格拉斯的哪一个版本，它的discretion，然后最核心的就是它的非法以及它的一个口令的方式，它对它就是我就描述这两个部分，我就生成可视化形式，其实它这一段就是它一个历史的，然后它的帽儿我去决定它的规划，然后决定它的比如以后面的方式就得到了它的规划是吧？

然后还有一个挂牌的东西，它的固定方式是它的x映射的是的FA然后他也只能用顺序型的，所以我的APP这样排列起来，label airway if you speak，然后它的y轴就是我的映射的是属性b而b是一个数值型的属性，我去映射方式配合对象，然后就得到的话，当然你可以可以看到它非常简单。我觉得其实描述式的语法你可以这么去理解它，其实这个摘的。

13:57  
构建这个东西。

14:00  
这个是它形式的一个参数规范，I你可以想象成一个功能非常强大的函数。

14:05  
你把这些参数传递。

14:07  
到这个函数里面，这个函数就可以去构建可刷形式，构建82个。

那么所以它最核心的地方就在于你这个函数是怎么写的。其实我们能想到如果你要去构建一个这个函数的话，其实最容易的方式就是我相信如果是一个写代码的话，我们最简单的方法是啥是啥？就是这个参数其实这个函数里面有很多的switch，然后我去l然后我来去判断这个参数是什么，然后我来去构建怎么样调，用什么样的函数，然后去怎么怎么去画对不对？

但这就是一个这样传统的这种方式，基于模板的文化构建的方式，比如说在我们在Excel里面也可以插入很多的这样形式的，比如说巴沙尔这些都是可以的， Line I82等等，但是你可以看到他们的构建方法都是一个模块一个模块一个模块，你选择了一个模块构建对应模块的表达形式就好了。

所以它这种的方式是抽象层次非常低的，就相当于我刚刚说的Excel，我去判断一下它变量，然后我就调用变量对应的渲染的函数，然后得到它他的问题是啥？如果你的这种形式，你不可能一个都给它遍历遍历一遍对不对？你这样的话，假设你出现了你有100个不让停车，你出现101个可能又要重新写，这是非常维护性很差，然后可扩展性很差的一种方式。

然后我们其实比较赖它基于的工作叫然后这个工作是在1999年的。

16:17  
时候有一个。

16:19  
来统计学家他所提出的，当然他在2021年的时候就过世了，他其实做过态度的副总裁，然后对他同时也是一个大学的老师，他对他最大的贡献其实就是提出了的干扰装备是这本书。他我们去支持我们去实现可视化描述性语法的一个基础。

对这本书非常厚特别厚，这本书它主要提出了什么？它主要就是将提出了一个高级别的抽象的图表和描述方式。

我们可以举一个例子，就是他把图表变成一个数据加几何分析的视觉特征的表述，几何图形的视觉特征它就它不是一个的去遍历它，而是它把它抽象成一些基本的特征骨折，比如说这个一个很典型的骨折的行为方式，比如说横轴是视觉元素的特征，有位置、认识、长度、角度、方向这几个大小等等，然后纵轴就是它的它在哪里画，比如说它是笛卡尔系、坐标系、笛卡尔坐标系还是极坐标系还是地理位置等等，你这样的话你就能看到它拼成了很多的复杂图形。

它这样这还只是说我们选择了两个维度来去拼，如果你把它所有的抽象的维度都把它组合起来的话，它的能构建的复杂形式就会非常的广泛。比如说我们举个例子就是说一看二坐标系下，然后位置很显然就是三点，极坐标系下你也可以构成。

那迪卡诺威旗下任何的80毫升，咱可以变成微调微调查，然后或者是迪卡罗来记一下这种的叫做加货，比如说大家去看玩游戏的时候，不同的人他有不同的能量值，我要去看这个人他这个角色，他在不同的维度上他的能力强与否，我这种人用这种思考来去做，他就是在极坐标系下，然后通过direction的方式来进行一个然后在那当然他便到依靠的关系还是变成一个蓝色等等。

其实有了这个组合之后，你就会发现其实它的构建方式就会变得灵活很多，这还是威刚I的1个技术，同样威刚那种也不是刚刚那4个人，他们一下就提出来了，他们也经历了很长时间的一个探索的过程，一开始它是基于的观察花费出这本书来去做的，当然这个不是他们组的人做了，是体育高速也是一个很经典的描述。

描述式语的语法它是基于r语言，你可以看到描述式的语法它是不受限于某一个具体的语言的，不像DC瑞他一定用 Gs来去做。

其实描述的语法，因为你只是一个js，你可以用c加加的去画家，画得更快，你就可以用这样的话，或者Python对Python更熟悉，你也可以用Python来去实现，所以它的可迁移性能力比较强，这个是基于r语言的，它基于r语言，r语言是用统计的人你可以看到它构成的可视化形式都是一些返点图，然后折线图这些统计的图表它也没有那么灵活。

然后后来在14年的时候，在 u大在斯坦福的那几个人，他们就提出了微感是发表在幼儿位置上14年的文章。你可以看到微杆非常长，同样的花哨，我们刚刚看到只有几行对不对？就这么几行，但微杆这么长，伟大的能力远比伟大y要更强，但是它的长它的编写的难度会更大，然后这时候伟大的时候他还不能支持交互，然后后面他又提出了对x这个就是说我能够去描述这个图无源的映射过程了对吧？

我还要去描述用户的交互，我把用户的交互描述出来之后，用户也可以支持在这个上面去进行实时的交互，比如说我去刷选或者我去点选或者我去怎样的交互，其实用 Vr微感也能够去支持。

有了图源的映射，有了交互，ok你把它组合起来，然后就得到了一个很强大的这种可视化的描述式的预防。

然后威刚案子的提出主要就是说他觉得威刚太复杂了，然后这时候同时威刚的能力它仅仅限于一个单页面的这种格拉形式，这时候就提出了非常nice这样一个很灵活的很简单的这样一个规划，聊城市的地方，然后后来当然阿尔本的作者他后面在2023年的时候，到了MIT之后又提出了MIT的没搞大一点这样一个描述式的立法，但是它并没有前面的用的那么广泛了。

主要维刚艾特是在2017年的时候我们领域的该配合。其实当时我看到这个论文获得最佳论文的时候，我还觉得很诧异，我说为什么他能获得最佳论文呢？但是现在来看，其实它确实推动了很大一部分工作的发展，比如说至少用到只要你来去做推，比如说做推荐或者做一些这种方面的研究，其实威刚IP一定是绕不过去的，当然我现在做工作很多工作也是基于这个IP来去做。你可以看到这也是一个很好非常好的社区，他就把他不只是提出来了，而且他确实是可用，我觉得这是迪斯瑞跟威刚I跟v港他做得非常好的一点，确实是能让人都用得起来，所以它。大家可以来去自己来去玩一下玩玩去。

22:46  
大家。

22:47  
也可以直接收收费到那里就可以了。

22:49  
我们看到的是去年 ok我们可以看到

23:06  
可以看到被告外的里面，这些都是我们基于障碍能够去实现，当然有同学会问我直接用WiFi来去做我的大作业可以吗？其实理论上也是可以的，但是你会发现它的表达能力也远没有迪斯瑞更强，它可以支持一些很简单的非常好的图片，但是复杂一点的还是没有去支持的。

比如说我们去。他也可以去比如说来去嵌到一些ps里面去做，要猜测怎么样去做都可以，比如说大家以后去做图，完全可以用v8y的，全我们可以去看，比如说这种82次有这么多类型的都是我觉得都记不住。

23:58  
拉条子。

24:02  
你看这种人也很简单，就是他简单的一个对应的问题，就是导致我可控能力会比较差，这是一个推到就是。

24:14  
我的权衡。

24:25  
讲比如说刚刚有同学说我要去做卖对不对？你可以去看违章y也可以支持，然后你把它打开以后。

比如这就可以去画一个最典型的美国地图，然后我可以去映射这样的一些数据，然后把它映射到这个地方。代码有什么限制？其实你看它的包括政策的时候，它主要有两个方面，第一个方面就是它的SARS没水平，我要去对于数据进行变换。

第二方面就是我们怎么去映射 Mark，给你处理三个方面，Mark就决定了我的视觉元素是什么，你扣点就是我的视觉视觉映射方式是如何。

其实我们最近还做了一个工作，比如说我可以去用大模型，你去问大模型一个问题，然后 ll然后下载平均3.54都可以，你来去问一个问题，然后你让他去生成一个v8I其实他的能力也是非常强的，也几乎能够达到相比于现有的一些LP的方法，其实觉得能够生成。

25:46  
v8的。

25:46  
准确率和有效性也是很不错的。

当然也还有很多的方法来去提升它，比如说我们现在用的re g的方法，我们可以把这种基于这个道理问题，然后来去提升大模型，去创建为高大上的能力，也是一些研究的工作可以去做的。

26:05  
我可以好的。

26:10  
好的，你看一下这个例子，看完这个例子之后，你有没有一个感觉说其实这些例子都很简单，一般都是一些普通的差，对不对？那我相信大家我们之前上课的时候，我请大家来去画勾，请4位同学来去画勾。

比如说你有一个文件夹系统，它是一个层次结构的这样的一个层次结构的数据，我怎样能够把它用VR来去实现，至少在例子里面是没有对吧？那说明这种方式，但是用第四位是绝对能够实现，说明这种方式它就会有受限的表达。

我们就到这里就说描述式的语法是否具有足够的表达能力，我们来去看一下，比如说这是单元的数据可视化，就是每一个视觉元素数据我映射到一个视觉元素上，我可以得到这样的数据视觉的分化分化结果，这个是领域相关的基因数据的，我可以把基因去映射到跟领域相关的一些复杂结构上。

比如说我们的层次数据可视化，我们前面所提到的那些层次的层次数据可视化的不同的形式。

27:35  
好玩。

27:37  
其实刚刚看到那些例子之后，你会发现所有的突然形式它都不支持这样的，我们现在在屏幕上所展示的这种智能化的一个所以也有一些工作，比如说22年的一一个历史领会的文章，就再去研究说有没有他的表达能力是如何的，他得到的结论就是说并没有一个语法能够将所有的转化形式都能够创建出来。

既然描述式的语法很有效，但是它的表达能力又有限，所以一个自然的方法就是我去做这种领域性的这种描述式的语法，比如说我刚刚举的这个例子，所有的单元可视化它也有自己的描述式的语法，基因数据可视化也有，很多数据可视化也有，他有现在调研了之后有57个不同的斋粉，基于斋粉的文化的描述性，我们可以来去看一下，举一个例子，举一个红色数据本上描述去语法的一个例子，我们可以看一下，假设我们针对一种类型的这种复杂形式，我们应该怎么去把它能不能去构建出一个专门针对这种类型可视化的一种描述式的语法。

其实这些都是重复的数据可视化，比如说我们刚刚说的文件管理系统，前面大家也画过，这种是适合去展示它的。

我认为用缩进的方式，然后水平的来去布局它的标签，然后这个是我的一个适合去布局大规模层次数据规划的，这个是大家前面画过的树图，这个是新闻的数据，这个是一种统计，反演出我可以去重点去强调它们之间的连接关系，这个是一种。

29:36  
有这种。

29:37  
使用这种方法和原嵌套的方式来去表达它的层次结构，很多的不知道什么事情。

层次数据它当然普查形式会非常多，比如说它的数据很复杂，还有拓扑结构也有检验属性，首先它数据比较多，其实它的映射方式很多，我们刚刚前面说的它的有不同的适合用于不同的视觉元素，以及不同的视觉通道，而会导致我们的设计空间很大。然后它具有很各异的形式，比如前面我们看到过现在有330多个不同的回答方法，当然这些可查方法它适合用于不同的场景，比如说这个PAD它其实适合用于表现家族处理，所以它有两个，一个父亲一个母亲有个孩子，然后适合用这本书。

而这个是说它有一个持续的关系，我适合用y二这种丰富的数据不容易，这个适合用一个大规模的投资数据，因为它适合可以在横向纵向来去很方便进行压缩，所以说它因为这些方法的提出，它一定是有它的原因的。但是现有的方法，比如说基于VIP，我们刚刚所说的它是不适合去用，没有办法去支持的，比如用第四位它的难度比较大，这就是为什么我们要去做研究领域寂寞，也针对某些领域的规划的一些方法，所以我就可以去构建一个统一的数字化的框架，但是它的挑战就是数的可视化数量很大，我们如何去覆盖大量的数，可按形式它的形式各异，我如何去提取出数复杂形式中公共的部分，其实我们就是要去总结出来它一个就相当于类似于刚刚我们说的，不然的话瓦菲斯那几个轴，我们把它总结出来之后你拼接出来，那就构成了一个很大的空间，所以我们就可以来去解构这个数的规划，比如说第一步，第一步结构方法，因为层次数据它有一个嵌套的关系，所以我们可以把它由数和化解构为数和化的单元，比如说这是一个数和二单元，它其实也可以把它当做一个一个数字形式来去看的，然后我们有了这个结构之后，我就可以把一个数结构成为很多的数的单元，有了这个数单元之后，我们就可以把针对数可视化布局的问题转化为数单元布局的问题，然后进一步的对于数单元的布局，我们可以去定义刚刚说的坐标系，然后他们有我们的0号坐标系及坐标系，这些不同的坐标系对不对？

这些坐标系这是一个维度。

还有其他的维度就是视觉元素，就是我们说的 Mark是什么，你的数据映射到了什么样的视觉元素上，比如节点映射到不同的视源，当然边你也可以预测到不同的视源路上对不对？

还有一个方面就是它的布局是怎样的，布局是一个很重要的方式，现有的方法我们它仅仅关注的是父子的一个关系，我们通过引入有这个方法，通过引入朴素的集合，然后来去建立父子关系和兄弟关系，这样的话他就可以将这前面我们说的是有哪些视觉元素，然后主要是将在视觉元素应该放到什么样具体的位置上去，这时候它就可以去描述这些视觉元素之间的相对关系，然后来去决定它的位置，而他的关系有父子关系，这是在层次数据里面很重要的一个关系，还有他的兄弟关系就是说相邻同层的节点之间它们是怎样的，有了这两个关系之后，我们可以基于现有的人形式来去总结，比如说他们父子关系有一个包含被包含关系或者并列的关系，那父兄弟关系我们同样也有很多关系，这些关系可以覆盖到他们任意两两之间的关系之后，得到的效果是什么，你就可以把一个数总结分解构成为这样的组组件，你可以组合这些组件就构成了这样的复杂形式，比如说节点链接数，你就是构组组合，比如说一看517，然后视觉元素是圆形，节点是直线，然后兄弟关系父子关系兄弟关系是这样，然后你把它组合，但你可以切换，比如说我们去切换一块坐标系，你会发现它可以把一个数切换到一种极坐标系的这种树形结构上去，当然还可以进一步的去切换，比如说我们去切换它的视觉映射方式，我把它的边把它映射到取消，然后同时我去映射到一个圆形上方形上就是这种旅行上，然后我就可以得到这种趋势图的是吧？

所以这样的话你会发现我们需要编码了，但是它的表达能力同样还是很遗憾的，很强大的，他是怎么去做的，有了这些描述之后，你就可以转化一些线性的约束，它的元素，它的一些属性就是这个元素它放在什么位置，它的宽是多少，它的高是多少，然后每一个关系它都可以转化为一组的约束，有了这个约束之后你可以去求解约束，然后就得到比如说包含关系它的约束，他的父亲是他整个数单元，他的父亲的宽度是整孩子的宽度加上左边的距离加上右边的距离这样一个约束，有了这个约束之后，你就可以去构建很多的复杂形式，比如说这样的一个描述性语法，它可以构建100多种不同的数字化，这是一个树壳上的集合，这些都是我们去构建出来的。

当时第一节课的时候，大家在课上去画了一些图案形式，都是能够通过这种语法的东西构建出来。

比如说这种吹map的形式，当时画过有的同学是画过微调的形式，这种扇形的形式或者科技的形式都是可以的，当然它还可以进一步的去辅助我们来去探索不同的数字化形式的设计空间，比如说我还可以去多凑合，比如说这是两个不同的数，我可以去把一个组的数据去换成另一个部分之后，那就得到了另外一个数。

我们当然可以探索新颖的可查性，比如说这是之前没有的，我就可以通过拼拼凑凑的组合得到，同样它也还是基于Jason，这是Jason的一个语法，然后我就可以构建这样的处理。当然有了加分之后，这就是为什么我说这种描述式的语法它还是很有用的一个原因，你确实是可以支持让有了这个语法之后，你就可以在这个语法之上去构建一些交互式的构建的方法。

比如说这些都是我们指定的一些组件，你通过拖拖拽拽这些组件之后，那就可以实现这种文化的构建了。这种会远比你去写代码或者写摘分的方式更灵活。

37:05  
通过的然后不好意思。

37:18  
好的。

37:20  
我们现在去介绍跟大家去介绍了不同的可视化的编程的方法，这种编程的方法有描述式的，有命令式的，那描述式的它很简单，但是它的表达能力很有限，命令式的比如说第四位，它的相对还是更繁琐一些，你的学习曲线会更陡峭一点，但是他的表达能力还是很强大的，你能够支持使用第四位来去实现，就是你想到的任何的表达方式。

所以大家可以在以后，如果我们去有这种场景的话，你可以根据你的场景来去选择，比如说你构建了很经典，你也不需要来去重复的道轮，你基于这种描述式的语法，也可以去实现你的规划的实现你的规划这一点以上我们跟大家去介绍了关于在规划中它的编程的方式，我们继续来我们讲介绍下面的内容。

好的，我们前面主要讲跟大家讲了啥，讲前面几节课讲的第一件事情就是可视化是什么，可视分析是什么，对吧？第二件事情是说我认识人，那认识到其实人的有些他的认知上的能力是有限的，所以需要我们在设计需要制作可视化的时候需要去注意。

然后第三个方面就是我们怎么去实现从数据去映射到最终视觉表达的一个过程，这时候大家有了这个能力之后，我觉得下一步就要学习的能力是应该你能够去做，那是你不知道数据能够映射到什么对不对？我们前面跟大家说的有了数据，有的是觉得有了场景，然后有了你的任务，你要去映射到一个非常高效的视觉表达上去。

从此之后我们来去学习的，我们针对不同的数据，我们应该怎样去有。

39:44  
哪些不同的选择。

39:45  
什么样的高校人家可抓你是它是低效，它适合于不同的任务，我们后续的工作后续的课程就来去学习这样一件事情。

本节课我们先学习的是高维数据可视化这样一个很经典也很重要的一项一个类型，什么是高维数据呢？高维数据它大家应该记得我们去划分这个数据的那种不同的类型，其实第一个就是数据表格数据和高维数据，如果大家还记得的话，这个是最重要的，高维数据就是说顾名思义就是它有很多的维度对吧？

那表格数据它每有很多的行有一行，然后一列一列就是一个维度，然后一行就是一个数据对象，这样的话这样一个表格数据它其实就构成一个高度，这个表格数据它的维度很低，但是它只有三个维度三个对象，但是如果你有很多的列很多的属性，这样的话它就能够被称作一个高维数据。

高维数据它的问题是什么？我们的屏幕空间是二维的，我们能够感知到的空间，其实我们生活在三维的空间里面加上时间的话，其实我们是一个四维的空间，对不对？所以其实在这种二维三维空间中来去可视化高维数据，你就会感觉到这是一个非常困难的问题，就像非常乏力，你也不知道怎么去做。

那就像是这样，你有了这样一个很巨大的表格，如何从巨大的表格里面去获取模式呢？我可以大家来去先思考一下，比如说针对这个表格，你觉得你能构建什么样的可视化形式，然后来去探索这个表格？就这样一个表格，你的任务可以是你自己，比如说你自己设想任务也可以，自己设想场景也可以，有了这样一个场景，你觉得你可以用什么样的普查形式来去探索？

我们大家可以相互讨论，也可以调研资料，然后可以去想一下，我们一会儿来可以来去交流这件事情。

这个表格我当然是随便选的，它其实每一页你就可以认为它是一个水平，它是年1960年到974年，然后这是不同的国家。

42:27  
比如说。

42:29  
开门是吧？

它是你可以想象可能它是一个 Gdp。这样的一个数据，你觉得假设让你去分析这个这个表格的话，你应该去用什么样的表达形式，你能想到一个什么样的表达形式呢？

42:50  
不可能你。

42:51  
觉得这是一种什么样的。

42:57  
因为我觉得其实从可视化我们可能更关心的是它的一个趋势，所以我可能会选择这种。

43:04  
 ok折线图。

43:07  
你觉得我也不清楚。

43:12  
趋势，因为它有时间的，我觉得好像也挺顺理成章的，因为有时间，所以大家会想到任何任务的可以。

还有吗？除了人员之外，不能只有一个可靠性是吧？我觉得大家看看在我上我这节课之前肯定也有其他的了解过的形式结论。

43:37  
柱状图是包括柱状图。

43:40  
你觉得它柱状图主要来去分析什么？

43:45  
对这个应该 ok变化。

43:54  
的趋势也对确实。

43:57  
但。

43:59  
但他肯定没有那种折线，如果你想分析趋势的话。

44:02  
可能人员会发生的你不能吃，然后其他的还有什么，比如说

44:08  
表面下方的都可以。

44:10  
没有还没有。

44:13  
谁知没有了。

44:19  
我们自身也不可能的事情发生。

你觉得。

44:23  
镶嵌图他要看一个这种镶嵌图，你觉得镶嵌图如果觉得可以节省点。

44:34  
谁知我看到他们的过程。对镶嵌同时能够去分析。

44:41  
比如说它的阿尔法业务的它有一个异常值高的异常能力的异常是吧？

44:47  
过去看不到。

44:48  
你觉得。

44:53  
都不同意。

44:54  
其他人呢？

44:58  
谁知。

44:59  
其他人。

45:01  
你觉得应该比较关键，清平。

45:11  
Ok这个很好。

45:12  
其实。

45:13  
因为你可以把它当做一个nature，对不对？你可以把它映射到颜色上，然后来去看它的颜色的。

45:20  
给他申请，你能看到它的温度情况。

45:24  
其实。

45:26  
刚刚有的同学说的是一个其实自然的行为，自己为什么。肯定也能够去映射到一个地理空间上来，有大量的分布，对吧？好的，大家先下课休息。

45:38  
谁知我说过什么，谁知。

45:53  
谁知。

45:58  
我想问的是谁知不要吃什么，谁知数据。那么早晨感谢其实我们谁知和工作情况，谁知什么都有，房子可能不清楚，不能这么要求，谁知谁知我们可以管理这个事情。对。谁知谁知是吧？谁知。最好的数据觉得很好了。其实谁知好好的谁知道情况这个问题谁知谁知这个钱的谁知我们这块的工作其实挺好的，一直是可能的。

谁知计划的把我们的谁知深圳这句话谁知谁知数据谁知，谁知。谁知谁知，谁知。谁知你这都不看了。谁知也可以，对吧？数据谁知他可能没有什么事情，他就比较清楚。谁知这个东西怎么了？谁知。

50:51  
房地产的使用，谁知。

50:55  
好的。

51:00  
我们继续来讲。

其实大家就是想的形式还是很丰富的，但是其实也并没有超出我感觉大家确实应该想的形式，肯定是我们这节课不会去讲的这种形式。

还有其他人有想法可以去说一说吗？其实我这边拍了一下，大家可以想的形式有8串什么的，我们可以去看他的。

51:34  
数量的变化。

51:35  
82分钟是最经典的，然后啪啪一下，比如说它有不同的属性，然后我想去看，比如说某一个国家属性，它的占比情况是如何的，我也给你发一下，当然我可以用这种蓝叉来去做，因为我有一个时间的序列，所以说我想去看它的变化情况，还有一些同学他们比如说提到我可以用这种把它当成一个没去，我就把它映射到颜色上，其实一眼就能看到整个从时间从国家上分布情况是如何的这样的一个结果。

但是但是当然我用没去的这种方式还是能够把每一个坐标都能够展，每个数值都能够展现出来，但是用拍摄跟用蓝色，其实这些格拉形式它当然展现得非常清楚，比如说你是多高多低，你占比多少，你的趋势是多少都很清楚，但通常都是呈现的是数据的一个总结的信息，比如说你的182处应该往往都是大，你假设你要去描述整个的这样一个高维数据的话，你肯定是要把如果把这些维度都加起来，然后得到的话得到它的总和，然后来去看总和之间的区别，否则的话你就要花很多82的对吧？

一个国家一个82，一个国家一个82。所以它往往都是一些数据的总结，它并没有去体现出这种属性之间的相关关联关系。其实这就是我们这节课想要从高维数据可视化里面最想强调的部分就是两个属性之间的关联关系，我们如何去呈现出来？如何呈现两个变量之间的关系呢？如果有两个变量，你怎样去确定怎样去映射这两个变量呢？可以有还是同样的，我们可以来去考虑我们的 Mark跟我们的微弱卖品的方式，就是使产品这个问题要钱。

53:54  
这个图。

53:55  
源跟属性的位置是怎么决定的？比如说两个属性，数字性的属性，我们自然它把它映射到位置上去，对不对？

然后这样的话我们可以把它连接起来，得到一个用line来去映射这个对象，当然我们可以映射到两个坐标上，我们可以用这个点的位置来去映射这个数据对象，我们可以分别的去介绍这两种不同的方法，第一种方法就是我确定了这两个属性，然后我可以把这两个属性连接起来，然后这个线就代表的是对象，但是这种方式仅仅适用于它的属性，有两个的情况录入，假设我有100个，我有200个，我有50个30个，我用这种方式怎么去表现呢？

让大家觉得有可能去能够去做吗？我现在我们就是把它应确定了它的位置之后，然后把它连接起来，然后这个连接的图源就代表的是这个数据对象，假设你有50个主题，你应该怎么做？是不是？有。大家有什么想法吗？当然你用这种方式自然是比较困难的，但一个很巧妙的想法就是我把这两个坐标轴我把它平行的来去放，平行的来去放之后，我还是把它连接起来，对不对？

这是y这是x然后你这是两个不同的属性，你这样连接起来，假设你有5个属性的话，你就这样平行的去排就好了，这样的话你是可扩展的。

所以这个方法就叫做平行坐标，平行坐标它是一个很高效很非常巧妙的方法，它是如果你提到高维数据的话，其实这个方法是很典型的一个方法，它的怎么去理解它，你就想象成把这2个属性我把它扩展到6个属性7个属性上去就好了。

这个属性就是这样去连接的，我每一个数据对象比如说是加拿大，对这是车，上次我们不是看过这个车的数据对吧？Mpg气缸，然后它的马力，这是它重量、加速度、年、年、年代然后考虑的东西。

56:43  
对不对？

56:44  
这个可能是一个类别的属性。

我觉得这样，然后这就代表是一辆汽车，一辆汽车有很多的属性，然后把它连接起来，然后这样的话就得到了。

56:56  
一个汽车的。

56:57  
就得到了汽车，这条线就是描述的是汽车，它能够好处就是我其实有了这个之后，我就可以去看这两个数据之间的关联关系。

57:11  
有这种这样的。

57:12  
数据平均坐标。

57:16  
尽管它非常高效，但是它其实也存在着一些问题，我们可以来去分析它的问题是什么。

其实平行坐标它可以体现它的相关性，但是它主要体现在两个相邻轴之间的关系，两个相邻轴你才能看到它是正相关还是负相关的。当然它有可扩展性，但是它也是有限的可扩展性，比如说50个维度，然后1~5000个数据，如果你有更多的数据更多的维度的话，它也是慢慢去扩展。

第三个方面是它的交互很重要，它可以刷、选、排序、过滤，这些都是很重要，然后它也有一些算法能够去支持，让品种要变得更高效。前面我们看到这样一条，他还是很清楚的，对不对？但是如果我们这样，我们把这有很多个车，我们把很多个车都给画出来，可能你就得到了这样的格拉结果了。这样的核查结果他的问题是由这里对不对？你是不是就觉得它相互重叠会非常的严重，所以这种遮挡这种重叠就会导致数据中重要的模式都被隐藏掉了，这是品种的一个非常严重的问题。

当然解决这个问题的方法，我们可以来去通过一些筛选的方法去做，比如说我们可以去看这种用第一次未来去实现的平均坐标一个例子。

58:57  
谁知要坚持，谁知谢谢。

59:48  
我把发到群里了。这个就是第一次位置可以看到，第一次一点看到我们这边这样的这个就是一个苹果得到挖掘，这个就是我们刚刚说的汽车的图，汽车的图你看这边非常的重叠很严重，那好处就是你可以来去通过交互的方法来去选择它，比如说我只想看30，我想看啥？

我想看到的。

01:00:22  
虽然讲对于。

01:00:25  
气缸数量比较多的，它的分布情况是怎样的，你就可以这样去选择去探索它，你还可以去选择其他的，比如说它的。年份比较数量比较小了。

01:00:44  
数量比较大。

01:00:46  
的比较好，这时候我们可以来去选择，可以在不同轴上去选择，进一步的去确定好我希望得到的车是怎样的？

比如说有的同学他希望看的车是气缸数量比较多，我可以去定位到7道都算，比如说我提高数量比较多，然后钱比较低。

01:01:18  
然后。

01:01:20  
法律比较强，

01:01:24  
送达比较高。

01:01:26  
Ok其实。

01:01:28  
你这样的话，你就能够选择到很筛选掉很大一部分数。

01:01:33  
那个公司的车。

01:01:37  
然后进一步的你就可以来去确定我要去选择的车是什么。

交互的方法是一种非常是对于层次平行坐标来去探索的一个非常重要的方面，因为它可以有效的去避免减轻重叠带来的问题。

01:01:57  
钱。

01:01:59  
当然你还可以去进一步的去着色，来去让这些您坐标显得不同类别能够体现的更加清楚。除了交互之外，如果我们想直接的去直观的去看投资结构数品种高的数据，这时候我们应该有什么样的方法能够去减缓或者是避免这种重叠带来的问题？

那一个方法就是我们可以去对于高维数据来去做聚类，聚类之后我们可以对于聚类进行着色，比如说蓝色它就是一类绿色的这一类，红色的内容，大家也可以控制聚类的结果。

有了这个类别之后，你就可以更清楚的去看到，比如说这边其实它就分了三波，第一波位置跟加速度它成反比，然后还有一个就是绿色的，它其实相互可能没有那么明显的这种反比的关系，这样比使得更清楚。

然后除此之外你也能看到，其实大概就是以分类的来去做分类的，当然在不同的类别里面它也有一些加法，就是红色里面有一些绿色，可以进一步的去探索为什么红色里面这种绿色的是什么，可以进一步的去看到跟其他。

01:03:28  
然后。

01:03:30  
通过这个图你也能看到它主要分类的这种指标，基于c的来去做分类的，其他的它当然也有相关性，但相关性没有那么强，相比于这个图之外，这个图肯定是我能够更清楚的去了解这个数据了，对吧？

除此之外还有我可以在平行坐标里面去加上平行坐标的中位数是怎样的，就是一个类别，哪一个线能够更好的去代表维度，这样的话你就能够更容易地去探索到顶级坐标理念。

它代表这一类它的特征是什么？比如说红色这一类，它的公因数是这样的，它 Mpg跟c分别是成反比，然后这样的一个变化，然后绿色是这样，蓝色是说这个跟马力差不多，马力跟它的重量也差不多这样的一个类别。

当然除了这种方式之外，你可以看到平行坐标它的连接关系是直线，那直线呢就导致我相互之间它的重叠会更严重一点，进一步的就会有一些边捆绑的算法，它我不用指向我用一些曲线来去表示，那曲线就可以将一一簇能够更好地绑定在一起，我就使得我更好的去看到这一群它代表的含义是什么。

这群他就避免了，比如说这本来是交叉非常严重的，这样的话你就能够使得它交叉不那么严重，然后因为属于同一库的，它就可以通过边捆绑的方法把它绑定在一起，避免了边的相互之间的重叠与交叉。这是当然编头脑怎么做的，其实就很简单了，如果大家在图形学课上肯定学过贝塞尔曲线之类的，它通过增加一些控制点的方法使得这个直线变得圆滑平滑，然后平滑之后你可以通过控制点的位置，比如说这本来是极限的，但是你把它增加上一些控制点，控制点它会往搭载的中心来去收缩，这就使得你得到了边得到的连线，它就会使呈现出边捆绑的这种效果。

这个算法也是很简单，除了这种方式之外，我将平衡还有什么问题你可以看一下。

如果这样这是一个line，直线穿过line穿过这个点的一个直线，如果是这种平行坐标，如果它是直线的话，你会很难去判断 line a是这样走的还是这样走的，就是line ABC，其实它都进入到连到第一个轴上，但是出了第二个轴的时候，你也分不清这个是line a还是b还是c了，所以这就使得你追踪平行坐标就会比较困难，这也是平行坐标的一个问题。

使用这种平滑的曲线同样能够解决这个问题。

它的好处就是我可以增加一些控制点，然后增加控制点之后，我在连线上更加的平滑，然后我就可以去避免用户很难去追踪到 a是怎么去连接的这种问题，这个方法也是通过贝塔曲线来去实现的，我可以在这个地方去增加控制力，然后如果a是向上升的跟跟平滑在控制点上曲率就会不一样，所以我有了这种平滑的曲线，我就可以来去追踪，你是这么来去玩折过来的，可以这样来去折过来，你是这样来折过来的。避免了它的混淆的问题。

除了上面的之外，我们还有其他的方式，比如说我们增加一些康拓，然后来去看到不同类别他们的分布情况是如何的，他们相互之间的差异在哪里，然后是这样的得到这样类似的结果。

这是平行坐标的第一个问题，就是它们如果连接之后，会导致我的连线之间会相互的重叠，相互的遮挡，平行坐标还有第二个问题，就是大家可以看到我们把轴把它展开了之后，因为我使用什么来去表达这些属性之间的相关关系，我们使用连线的方向的，比如说它如果是平直的这样连的，那就说明它们是正相关，如果是这样的上下这样错开的这种变化，它就是负相关了，那么这种相关关系只会体现在相邻的两个轴上，对吧？

所以这就会影响我来去探索整个数据的所有的模式。

当然你肯定是希望看到数据中存在相关关系的这种两个属性，如果是这样的话，你就要反复的去交换不同的轴的位置，然后使得它们之间的相关关系能够更明显，然后得到像得到这样的结果。

如但是这个事情就是比较困难的，因为对于n维的数据，它其实有n个阶层的排列的方式，比如说七维它已经排列方式非常多，这样的话我们应该怎么去选择一个最优的排列方式，能够最好的去展现出高维数据中任意两个属性之间的相关性呢？这就是平行坐标要解决的第二个问题。

我们要去衡量属性之间的相关关系，但是没有衡量属性之间的相关关系，大家统计学里面有一个就是12根系数怎么去衡量变量之间的相关性是怎样的，这就是它的衡量的公式，对不对？

比如说要相关性比较好的，这就从三点情况能看到它都是表现这种间接相关的趋势，在平行坐标里面怎么去呈现它，几乎都是如果会相关的应该是完全水平，一点都没有交叉，这种侠锗对已经很好，这种方法就这样的一个结果，我要怎样去选择更好的一个平行坐标的排列方法，现有的一些研究工作也不说现有的其实很早了，他们的一些研究工作就尝试去解决过这种问题。

他的其实很简单，想法非常简单，我觉得大家都能想到这种方式，我无非就是去选择一个计算一些任意两个属性之间的相关矩阵是怎样的，相关矩阵比较大的，它的相关性这矩阵就因为每这个矩阵表现出任意两个属性之间，它的相关系数是如何，是大还是小。第二方面我去构建一个挂这个矩阵，后面我们学到图可视化的功能，知道吧？图它既能使用 note就是节点链接图来去表达，同样它也可以使用这种举证的方法的需要。

然后有了这种图的表达方式之后，我可以运行这种这个图布局的方式，图布局的模型，然后这样的话是实现更好的用户可控，而进一步我可以采用这种运营商算法，运营商算法你可以这么去想，运营商算法不是一个人他运营商他他如何怎么去便利，要便利所有的不重复的去便利到所有的城市，可能是拿到的收益是最大的，对吧？

其实你就相当于每一个维度就是一个城市，然后我去遍历所有的维度只见一次，然后我得到了权重，相互从a一到普通，比如城市一到城市二就是维度1~52，他走的路相当就可以用相关系数来去表达，然后相关系数当然是越大越好，然后我就希望得到一个对象最大的这样一个结果，然后进一步的我可以得到这个结果之后，我按照属性来去布局，从平移坐标，然后这样的话就能够来去得到一个比较优的平移坐标，当然用户也可以依据这个。

01:12:28  
用户。

01:12:29  
的探索的探索这种用户的期望来去改变。

我去布局的方布局的平构，布局坐标轴的顺序。

01:12:41  
其实。

01:12:42  
这是一个很很简单的方法，然后当然很简单的方法，在2012年的时候也发表到了我们领域的一个会议上，对。这个也是一个国际会议。然后可以看一下这个步骤，就是这个是完全杂乱的去排列的一个平行坐标。你可以看到它其实相关关系相关性没有那么大，整个都是一团乱麻，我也不知道是怎样，对吧？然后它有那么多的属性提升相关性是怎样的，我来去连接起来，然后你进一步的去运行它，你可以得到一簇，就是说它们相互之间相关性是比较大的属性，这是一个不同的因素，就是一些小的就比较零散。

而有了这个之后，你可以去选择一个旅行上的路径，然后来去走过所有的这些走过所有的。

01:13:38  
是。

01:13:40  
属性，然后你有了这个路径之后，这个路径就是你排列的一个品种道路，那相对于

01:13:48  
这个概念相当于。

01:13:50  
对于是不是他清楚了很多是吧？我能够很明显的去看到他分了两次对不对？在一跟二的时候分了两次，然后在三的时候他可能就没有那么明显，这样的话你就能够很有效的去得到探索，得到这样一个。平行坐标，这样一个比较有效的布局的方法。

怎么说？这个就是我们针对平行坐标这样这种方式来去实现高维数据可视化的平行坐标的复杂形式来去分析探索高维数据。

当然我相信大家在选择你们数据的时候，一定会用到这种平行坐标的，一定会有高维数据，其实你就可以使用高平行坐标的坐标方法来去去表现它，然后去看重它。其实现有的也有一些level能够支持我们，就这样做。其实平行坐标还有很多的变形，我们后续也会来去介绍。

所以它的变形都是在一些研究论文里面去提出的。

谁知这是我们第一种方式，第一种方式就是我确定了两个属性的位置之后，我把它连接起来，这个对象就映射到连线上去了，对吧？这是平行坐标，平行坐标就把两个轴把它给平行起来，然后来去做第二种方式。我不把它映射到连线上，我的Mark就变了，我就是把它映射到点上点的位置来去决定。

两个属性是什么，比如说就是一个然后媒体就知道说很快他的就点在这儿，就决定这个点，这个对象他的第一个维度是x区是x第二个维度它这个区域是y。

基于这种映射方法，我们怎样去表现我们的然后结果这个就是一个返点图的事，那就是我们上节课去讲的返点图，你有了这个两个维度，你就可以把它得到一个返点图，但是同样返点图还是它能够表现的是两个维度，我如果有很多个维度的话应该怎么去做？

我们上节课也提到过，我们刚才带大家去实现过返点读举证，那就是可以将任意高维数据中任意两个维度之间的属性能够表现出来。这种散点图它的轴映射的是数字型的属性，它的点映射的是两个数，表示数据项的两个属性，然后可以利用点的大小点的颜色来应用更多的属性。

比如说我们前面介绍过的从200个国家，1980年、18、一八几年到2000年200年，国家的人口的寿命跟它的收入它的变化的情况，大家应该记得那个视频对不对？

那个视频就是一个典型的高维数据的一个视频，其实它就是将时间去映射到了时间上，就是动画上它的大小颜色都有自己的含义，颜色映射的是最多数的大洲，它的大小映射的是人口数量和中国的大是中国。这个是印度的一个然后它的位置横轴映射的是他的收入，纵轴映射的是他的年龄，然后我们可以看到这些人他随着时间他琢磨右左下角一直变到了右上角，然后它的变化范围也去映射频度的差距是如何的，这又是一个非常经典的展点图的形式来去。

01:18:15  
讲一个数据故事。

01:18:17  
那散点图它注定还是说我只是去表达了两个属性之间的关联关系，我可以使用散点图矩阵来去表达。

有第一个属性的话，我任意两个属性之间的关系是如何的，可以看到它的对角线上的位置，其实都是因为它都是相同的属性，导致其实它的含义是没有特别多的含义。

同时沿着对角线对称的两个返点图矩阵，其实它也是有一些重复的例子在，因为他们采用两两表达两个属性之间的相关性，只用一个三角形里面就可以得到。

所以这就是比较好看的一个三点图矩阵，就是上节课我们看到的这就是花的内容。

但散点图矩阵有什么问题，大家可以来去想一下，你们觉得散点图矩阵有什么样的问题我们需要来去解决，好好。

01:19:31  
你好。

01:19:35  
大家记得我们说平行坐标它有什么样的问题，对不对？平行坐标的问题是第一个它如果太多了，它可能会相互重叠，它相互重叠之后导致他们之间的模式都不清晰了，这是他的问题。除了这个散点，除了平行目标之外，除了这个问题之外，平行目标还有一个问题。 Ok，我不好意思。

是的，我有了平行坐标之后，我只能去表达任意两个属性之间的相关关系，对不对？所以我也很难去所以我需要来去对于平行坐标来去排序排序，因为我只能看到相邻两个轴之间的相关性。对于散点图矩阵来说，它的问题是什么？你觉得它的问题是什么？

01:20:32  
谁知。

01:20:32  
或者没有想法。

01:20:44  
我如果。

01:20:46  
要。

01:20:48  
分析就是项目评估坐标，它可以去对多个属性共同聚聚类分组这样。

01:20:56  
一些网站3点。

01:20:57  
多就无法进行这样子。

01:21:01  
我如果要两个以上之间。

01:21:03  
的属性之间的关系的话，产品就没有办法。

01:21:14  
其实大家想问题的时候，可能所有的数据在维度比较小，数据量比较小的时候，它往往都是比较好，所有的理想现实都很美好。

但是一旦它的维度比较多，它的数据量比较多的时候，它的难难度就上来了，比如说我们三点规矩的现在还有4个还好是吧？假设我们有那么多，假设有215位，其实用平行坐标来说，你毕竟还是一个215个轴，这样的一个这样的一个轴的连接关系，但是你如果是215厘米算一个215×215的这样一个短点图的话，其实得到的就会非常你每一个散点图你还是要占有一定的空间，如果它的维度特别大，每个空间就很小，然后所以你就看不到它的它的具体每一个两个属性之间的相关性。

所以总结来说，其实这种返点图它也是有限的，可扩展性，当然它能够去表示高维了，但是它的维度还是有限的，它同样需要结合用户的交互来去做，然后以及如果它的维度比较高的时候，你要去结合欧克斯的3.0矩阵，而你关注到某些散点数矩阵，其他的采用一些简单的方式来去概括。

01:22:38  
同样。

01:22:39  
去选择维度选择顺序，这个是维度比较多的情况，除了维度比较多之外，你还可以去考虑它的数量比较多的时候你应该怎么去做。

假设我的维度还行，我的维度有五维，我的数量假设有1万多个点的话，其实大家知道用svg来去画图，每一个节点都是一个图源，你的svg上其实画不了几千个图源，如果你画几千个图源，你的网页去渲染就会比较卡很卡顿，其实你也没办法去及时去做这样的事情。

所以说如果你的数据量比较大的情况下，你可以使用这种比如说三点，第二次我没去的热烈出来去表达，说每一个我只表现的是它的属性的分布情况，而不是具体的每一个点我都要能够表达出来，这样的话它会更高效一点。

然后在对称的这种空间里面，你能看到你的属性，它的分布情况是如何的，它的因为本身这个空间在原始的散点图矩阵里面是比较浪费的，你可以看到它属性分布，然后同时你再去探索的时候，你也可以结合属性分布来去探索其他部分这些属性的它的位置。

这是散点图，散点图矩阵当然比较大的情况，你可以使用这种数这种关联关系矩阵来去配合散点图矩阵来去使用，比如说我们就去计算任意两个属性之间它的相关性的大小，计算好了之后，我可以把选择某一部分，然后来去使用单点数据这样的去呈现。

01:24:30  
高铁是按时给母亲，谁知。

01:24:40  
对于散点图举证来说，其实还有一个问题，我这边是散点图举证，当我去选择一些选择一组数据之后，可以我怎样在这个散点图举证之间去做一些很很高效的漫游现有的一些工作，他也提出了一些交互的方法，比如说这个工作是发表在20多少年，是在2008年的时候在顶会的获奖的论文，然后这篇工作主要是说我这个散点图矩阵，我在探索的时候，我如何保持用户的关注，然后能够比如说我选中了一部分数据之后，我能够持续的去追踪这部分数据，在不同的机构散点图上的分布情况，我们可以来去看一下它的对工作就是工作的不行，而这里面就是三点不举证，用户可以去选择某一个，然后你可以在。

01:25:43  
里面去实现。

01:25:45  
然后这边你可以在3:00举证里面去做危险，我可以来去浏览面临不同的反应，然后同时看到三点图之间的随着微信，然后你能看到三点不是怎样变化，他这种动画方法就能够辅助你来去追踪，它叫做柔林和大雅，就是掷骰子，你可以把这个散点图想象成一个散骰子的一面，它就掷骰子的时候，你能够去追踪一些点，它是怎样变化？

01:26:17  
好，谢谢。

01:26:27  
比如说从这一个点我变到另外一个点的时候，你可以看到他的一些点子，比如说你追踪一个点它是能够追踪得到的，它的位置在哪里？比如说我选择了一些点之后，我看节点的分布情况如何，你也能在不同的视角里面去追踪的到，谢谢。是吧？好。

01:27:06  
的谁知。

01:27:13  
可以在一个平板的交互系统里面，我如何来去实现这种全面布局的探索，是这样的。假设你选择这个数字，它在这个维度上比较低，它从这个角度更不能单独支持。

然后你看这三组在不同的角度上，他们之间的关联关系是怎样的，你就能够很好的去追踪。

01:28:11  
谁知谁知数据并且也有共同的政策，这么的好的 ok成本的事情，好的，谁知我们来继续去看一些其他的方式，是谁知

01:29:15  
对于散点图矩阵来说，其实他们散点图矩阵跟平行坐标它的一个问题就是我重点去强调的是任意两个维度之间的关联度，所以我们针对这个问题，也有一些其他的工作再去解决这类的问题。

01:29:39  
对吧？

中国人民共和国中国的时刻，我们可以。

01:29:54  
去看。

01:29:56  
一个很典型的。

01:29:58  
这个我其实在一开始的时候会讲到，我其实可视化里面有一个最重要的方法，就是要去做权衡，就是没有一个银色子弹能够解决所有的可视化里面的数据分析的问题。

在高维数据这时候我们就遇到了这个事情，就散点图矩阵跟平行坐标，它都有自己的好处跟坏处。这时候比如说散点图，它能够更好的去分析表现出这两个维度之间的这两两个维度之间的关联关系是如何的。对于一个平行坐标是更能够去反映出两个维度之间的关联关系如何。对于这种返点图来说，它更好的一个效果是说我们能够更好的去看到，比如整体的分布情况是如何的，这个就是它的好处。

所以基于这个好处，我们就可以做一个很好的巧妙的切入，就是我在散点图矩阵里面去把平坐在平行坐标里面去，把散点图矩阵嵌入进去，嵌入进去之后它的好处是什么？在这时候我们就可以把比如说我们这边只能去看到两个维度之间的关系，但从这个地方返点图里面，我们就能看到高维数据它大概能够分为几类，然后基于这几类的关系，我们就可以来去自然的进行探索。

我们可以去看一下这个工作他是这样去做的。这是顶坐标是吧？它进行着色了，然后着色了之后，它可以采用一些弯折的方式，然后使得我们更好的去居中繁衍图，平行坐标它每一条线它的连接情况，ok这时候我针对的是这样这两个维度，我现在只是能看到这两个维度之间的关系，对吧？

我可以把为了我看到全局的关系，我可以把很多的维度把它结合进来之后，这个答案我们就能够去看到这些选择的维度，整体的这样一个分布情况，比如说它就分为了这三类，然后随着你的维度的增加和减少，然后它的结果也在会呈现不一样的变化，然后基于这个我可以进一步的去着色它份额这三类来去有不同的颜色，然后有了不同的颜色我可以去分析，而不同的类别它的分布情况是如何的？

是这个就是我来去借助返点图举证来去结合返点图矩阵和平行坐标来去实现这样的一个很有效的方法。

当然返点图矩阵还有一个问题，我们这时候其实很多时候我们刚刚说的返点图矩阵，它更多的是看任意两个数数量化的数，数值型的属性它们之间的关联关系，对吧？

但是其实有很多的属性，它不是数值性的，而是类别性的，这时候你想到其实有了类别型之后，我可以看到我的得到的结果就是这样的，比如说这是区域，这是这是貌似这是月份，然后你这样的话你得到的结果其实都没有什么，没有不像你想要去看它们之间的相关性一样得到的返点，所以这时候是如果它的类别是一个它的属性是一个类别型的，你使用这种返点图，它的意义就会大大的降低，所以我们可以进一步的去控制每个散点图矩阵里面，它不一定使用散点图，可以使用其他的开发形式，比如说对于类别来说，我更强调的是它的分布，这时候我就可以采用这种去改变每一个散点图里面的这种具体的可视化表达，来去提高它的这样的可视化的效果。

进一步的你还可以还会发现一个问题，返点图里面还有问题。另外一个问题，它在对角线的地方，它其实都是一些，因为它都是相同的图形，所以它对角线的返点图其实是没有什么含义的。这时候我可以将对角线的地方变成一些分布，然后更好的去利用这个空间，而同时这个位置上下的位置其实它都是浪费的，所以你可以把它变成一个上三角或者下三角的矩阵，来去提升它的空间利用率，这些都是对于散点图的一些改进。

大家可以来去自己来去考虑，比如说对于散点图来说，我们还有哪些不同样的改进的方式？

这其实有一个好的想法，你甚至都可以去发出一篇很好的论文，其实这些工作都是我们刚刚说的工作，就是一个在我们顶会上2013年发表的一篇论文，然后它主要是解决的是不是针对这种数量数值型的指标，而是针对这种类别性的指标，我应该怎样使用返点工具，给一个好的想法，你会发现它非常的巧妙，但同时只要给你说了之后，你也很容易的能够做到，这就是一个。

01:35:18  
很好的一些东西。

01:35:22  
好的，我们还有其他的这种针对平行坐标上的一些改进的方法，然后我们下节课也会再进一步的提升。然后我们下节课也会重新的去复习一下，针对高维数据这类的手段方法有什么，然后以及除了散点流矩阵和平行坐标之外，我们也会去介绍投影，比如说高维的投影算法有哪些，比如说MD ST CTC内还有优迈普这些不同的方法，他们之间的好处和坏处等等。好，我们这节课就到这里，谢谢大家。

01:35:52  
分享。

01:35:54  
谁知现在好的吃饭里面的比较快的过程就是说我们在不同的工作不能做的说话的。

01:36:11  
比如说是是。

01:36:13  
什么。

01:36:14  
规范化统一化。

01:36:16  
的角度来说，这只是数据方面的问题来感受一下。就是说我们中心坐标的很重要的各种属性的数据的使用有没有所谓的规范，会议的话就。

01:36:29  
会对我可视化结果。

01:36:31  
产生我们讲的是错误的，其实。

01:36:34  
你的规划应该是不同的，对，就是因为它的范围值。

01:36:40  
域是吧？

对。

01:36:42  
但是其实它不同的属性之间它相互之间没有影响，因为它都是一个轴，一个轴映射的是一个属性。

01:36:49  
所以你可以看这个结果。

01:36:55  
比如说。

01:36:57  
你可以看到它从上站到下，从上站到下面。

01:37:02  
从上面来看，他们相互之间是相互不影响，比如这个是200对吧？

其实相互之间一点影响都没有。

01:37:08  
这也是它的一个好处。

01:37:11  
它可以直接去忽视它具体数据表现。

01:37:16  
它在相对的区域中的位置这种表现。

01:37:20  
谁知。

01:37:27  
可能另外。

01:37:29  
一个就可能。

01:37:30  
这个就不是说的问题，就是说如果说他的也不能说影响程度，就是说他如果不是一个均匀的资源影响，比如说我怎么去表达。

01:37:46  
我的意思，就是说。

01:37:49  
如果说我01和0.9之间的差距比我0.1和0之间的差距要大得多，这个是不是不是？

应该不是，我们说实话1和。

01:38:00  
0.9之间的差距跟。

01:38:02  
对0.80对他不是一个均匀的差距的高铁。

01:38:08  
数字化控制情况，

01:38:12  
你想表达的。

01:38:14  
是说。

01:38:17  
你可以选择什么不同的必要去表达它。

01:38:22  
你总有一个规律。

01:38:23  
你得总结出这个规律是怎样的，它到底是越大，它相应的差距越是0.9和0.1之间的差别。

01:38:33  
是比0.9跟0.8之间的差别是要更小的。

01:38:40  
这样的话你会去选择这样的数据。

01:38:42  
要把它映射到什么。

01:38:43  
没有什么比如说闹死掉，这个影响。

01:38:48  
就是说它的映射公司或者说走。

01:38:50  
哪里可以去做这个产品的。

好好谢谢。