### 大数据可视化与可视分析-李国政-第10周 星期一 第3大节.mp4

00:00  
谁知好的我差不多有10多个人。差不多。谁知数据一二20四，还有谁知我们同时也是一样。

00:37  
我们是什么？好好的。谁知。

00:49  
好的。

00:52  
好，各位同学下午好，我们开始我们的第三次课程。

00:58  
是这样的。

01:01  
我想首先请问一下大家有读好作业的吗？读好作业可以举个手。

01:06  
那么我们有几个人？

01:10  
然后有比如说决定好我们去做什么方向吗？有吗？没有是吧？你们会觉得上次我记得是小东同学提到一个问题是吧？你当时说搜集数据比较困难，是不是？

其实我后面也去找了一下，我发现其实还是有很多的可能性，大家可以我发现有一个这个网站还是它叫合金数据，然后我可以打开给大家看一下，其实有很多有趣的话题。

01:44  
我们

01:46  
我可以先发到群里，大家先。

01:51  
谢谢大家。

01:58  
当然我只是找了很小，也没有说很深入的去搜集数据。

02:06  
但是我。

02:07  
发现网上的数据还是有很多的，然后它的主题也很有趣，其实也算是一个大家可以去寻找一个方向，比如说这边大家可以看到这边有数据集对不对？还有很多数据当然也有数据可视化。

02:23  
你可以看到这边。

02:23  
我们可以点开来看一下。

好。

02:31  
你怎么会觉得开放开发。

02:45  
我老家什么设备，刚刚还是好好的。

02:50  
谁知

02:58  
我把链接发到群里，大家可以先看一下。

03:03  
数据谁知谁知。

03:32  
我就知道了。

03:33  
谁知我刚刚试了还是可以的，启动非常快，这边争取的刚开。

03:54  
但是比较暗你知道吗？

03:56  
太大了，数据的他们的产品在工作任务的企业拿到了人生的光。

04:08  
这样好一点吗？能看到吗？

04:11  
好好。

谁知。

04:28  
能看到我就随手搜了一下，我发现很多的数据，其实大家不用去处理数据，也能够去找到很多的来源，比如说这个平台上我发现有很多有趣的方向。

04:42  
然后这个。

04:44  
它的首页，然后我们可以去看他说数据转化的一个部分，然后其实这些数据都还是挺有意思的，比如说。

04:56  
这些数据。

04:58  
然后点开比如说我们可以去看，他自己也有一些初步的分析，但是我觉得我们做出来的结果肯定比他有趣的多得多，但是我们可以用它的方向，用他的数据，然后自己来去探索这些方向好了，然后我也找到了一些比较有趣的数据，可以跟大家去。

05:26  
分享

05:28  
整个数据平台，然后我看到不是最近新能源车特别火，我不知道大家有没有关注最近的北京车展，其实这个还是一个模块的方向，比如说小米服务器和我视频播放的点赞数据，我们都可以去看到，通过我们可以去看到，有的保的，有的是称赞他的，有的是去批评他的，他们不同的方向表示什么？

在b站上还有很多，比如说他们有很多分类，那up主他其实有一些比如说热门的方向，然后他的它的名字是什么，然后它的主题是什么？比如说它的播放量怎样，它的点赞量怎样，其实都是我们可以去来探索的方向。

当然我们还有历年不知道有没有大家关不关注百大up，然后其实现现在很多00后他们的目标现在就是我要成为百大up，这个也是一个很有趣的数据集，比如说他把历年来有百大up，他的它的点赞量、姓名、名字，然后主题以及一些简介，我们都把它放在上面，我可以去看一下什么，我可以去看过去这几年它的变化，然后当然如果自己熟悉的，如果我对这个方向特别感兴趣，我也可以自己来去想，我能从这个数据里面去发现一些什么。

07:05  
前面。

07:07  
都是比如说我从一些b站上来推过去的数据集，当然我前面也提到过，也有一些从我们的一些影视剧里面去看到人物之间的关系，我想去看一下这些人物之间它的关联关系是怎样的，然后辅助我来去看到整个很多剧情的发展，这个也是一个负面机会。

比如这个是权力的游戏，它的人物之间的关系，它的数据有什么呢？当然这个节点这个数据可能只是做一个初步，比如说它是84个节点，216条边可能比较基础一点，我我们可能在这个数据集上面进一步的去扩展，它还有边的属性，它是自然的父母关系，然后兄弟关系等等。

然后进一步的我们也可以去看中国的历史的文化里面它有怎样的数据集。看到比如说红楼梦它的事件事件之间的关系是怎样的，事件的介绍是怎样的，也在我们的数据集的这些提供的平台上也会存在，然后当然我们可能现在说新冠大家可能我不知道大家有没有陌生，可能回想不起那种魔幻的当时的情况，然后可以看到2019年的时候，新冠病毒它感染的死亡率是怎样的，还有空气质量的情况，这个数据比较大，我当时还仔细去看了一下这个数据，它是在2023年，比如说每一天它全国各地的污染的情况怎样，以及各个不同的污染物它的浓度情况是怎样，这样我可以看到什么？

不同的地区，比如说全国不同的地域，它们的污染物主要分布有什么差异？随着时间的变化，他们主要呈现的模式有什么不同？这样的话我是不是通过我对于数据的分析，可以给比如说我怎么去改善它的污染状况，一些很好的建议，通过这个数据其实可以是我们的一个方向。

有了这些数据集之后，大家可能会去想我要做到怎么样，他算是一个比较好的比较优秀的人的话，我说了不算，我们可以去看一下，这个是我们可视化领域一个非常资深的学者，他是我认为他是我们领域，比如说他的基础理论的一个奠基人，他是ubc的什么马尔教授，他是威力这本书的主编，当然我们课上介绍的很多的内容都是从威力这本书里面去就是说提取出来，然后来去总结出来，然后我可以去看一下他的课上他去选择哪些比较好的作品，作为他觉得认为是比较优秀的文化，结果我们来看一下。他的特点是他的课程，然后每个都会放在我他自己的主页上，当然大家也可以去看。

然后普拉达其实也是在我们领域是一个很优秀的这种教育者，因为他在每年我们为此紧急的会议上都会有有一个专门半天的时间来去有一个讨论介绍它可视化的基础，然后很多的人都没听。

我当然前面也介绍过，他的老师是无影响，或者是是在上海也是很有名的。然后这个是它的一个就说这是20是多少年，202023年23年的时候，他的课上然后选出了一些比较优秀的，比如他的作品，然后我们可以多打开几个去看一下。是。

11:10  
我先把。

11:12  
链接发到群里，大家可以自己玩一下现在。

好ok。

11:20  
没有发挥速度。我现在已经发到群里了。

11:28  
大家也可以自己来去看一下。

11:35  
比如说。

11:38  
我看一下，我记得有几个是特别有趣的，感觉这个是选。

11:46  
的

11:49  
这个是说我怎么去选择不同的游戏分辨率的问题，然后他有一些他选择的是不同游戏的特征，我去看，然后你可以看到它有一个很关键的部分，就是我可以交互的来去进行数据分析，比如说我可以看到选择一部分之后，我可以去看到这些游戏它比较倾向于哪些方面，当然这边也有一些分布等等。

12:14  
对。

12:16  
然后我还可以进一步的去看。

12:20  
2016年4月2号好不好？谁知到我们的情况。

好，2年7个。

12:43  
我自己也当然他做得非常好，比如说他在每一个的其实我们都知道获奖获奖理由，其实可以看到他的每一个获奖的作品中，就是how Free就是名人堂，然后我们可以去看他为什么获奖，他为什么做的好，他们说做的不好，其实这些都是通过大家的评价，通过塑造师的评价来去获得的。

这个是说还是他是说世界上所有的威尔达菲，这个是独角兽，对我们去看不同领域的它的独角兽，它的分布情况是怎样，比如它的投资机构有哪些，然后它的领域有哪些，它随着时间变化，它的市值是怎样的？

对，然后你可以看到我可以来去探索一下。这个屏幕屏幕的分辨率有点低。

13:44  
有种history。

13:46  
artificial、Intelligence，人工智能领域它的独角兽有哪些？

可以去看到字节跳动是一个非常非常离谱的等待，可以看这个就是字节跳动，然后字节跳动到现在还没有上市，所以但是它已经好像市值好像销售好像营收已经超过腾讯了，所以其实它又是特别离谱的，可能8月15号，然后它是由中国的某某些机构来去投资的，可以看到它市值的变化怎样，然后它的国家它在地球，但是世界上它的分布情况怎样，当然每我们可以看到这个网站上每一个它都有一个这个recall是什么？

我从中能够去得到什么有趣的人发有什么有趣的洞察，然后当然我自己去探索，我简单去探索了一点点，然后你就会发现不同的人有不同的发现，我只能发现这一部分，但是如果真的是做的人或者是熟悉这个人，他们就能够发现很多有意思的事情。

15:00  
然后。

15:06  
以上我就是给大家一些比较优秀的案例，然后以及比较有趣的数据，然后会给大家提供参考，这样的话也可以避免，比如大家去搜集数据的时间，去处理数据的时间，然后来去降低我们的这样的工作的难度。

对。我们有什么问题吗？当时老彭同学有什么问题吗？没有问题，你觉得那些数据你看到比较有趣的问题，你觉得还符合。

15:42  
你的想法吗？

15:44  
目前看到比较有趣的。

好的，其实数据还是有很多的，你可以去看，可以多去搜集一些，大家有什么问题吗？关于这些我们提供的数据和案例有吗？Ok如果没有问题的话，我们就进一步的去开始我们后面的课程的内容的介绍，然后我们上次主要是介绍到了可视化和可视分析的概念，我们当时讲到，我们可以先简单回顾一下我们上次课讲的内容，首先我们去用定义一个比较官方的定义来去把它结构来去解释，可是它的分析是什么？

比如说它是基于计算机的，它是一个可视化系统，它提供的是数据的视觉表达，然后帮助用户来去完成特定的任务，这些都是我们通过结构的结果来去提供的一些比较有趣的对于这个定义的一些解释。然后后面我们就简单介绍一下可视化分析它的发展历史情况是怎样。它是从1987年开始，然后直到现在分为很多个阶段，一开始主要是集中于科学主义文化，比如说人体或者是一些物理模拟产生的一些流程的数据等等。

然后到了20到了90年代的时候，我们会发现有很多的抽象的数据，比如说像。上节课请大家来去绘制的一些层次数据，当时我们的文件目录，电脑文件目录就是典型的红色数据结构，还有网络数据文本数据等等，这些数据就是导致了信息损耗的出现，然后进一步的到了2012年的时候，数据量进一步的去增大，到了大数据的时代，我们发现单纯通过通过一个简单的数据类型，没有办法去辅助用户完成一个具体的任务，这时候我们需要综合很多的数据来去实现数据的分析，这时候只是发出扮演的角色，角色就是我们将人跟机器，然后来去通过一个分析的试图来去实现协同，然后从而更好的去完成用户的分析任务。

讲到这里，我们这边就提出了在2010年的时候有n条k它是它是在欧洲的一个科学家来提出了分析的系统的框架图，然后这个框架图其实很抽象的来去概括或者分析它的概念是什么，就是他从一开始然后到数据，我们一开始数据肯定不是非常的完善，非常的整洁的数据，这时候我们需要对数据进行清洗，要对数据进行股权等等的策略，然后去得到一个比较干净的数据。

当然除了这个方面，我们一会在实例中我们也可以去看到数据它可以进行变换了之后，我们可以通过另外一个视角来看这个数据，你可以得到更多不一样的发现，这个就是数据变换的方面。

然后有了数据变化的方面，我们发现它有一个分叉，这个分叉就代表着两种不同的对于数据的分析的方式。第一种数据的分析方式就是通过儿化来去实现数据分析，我要通过数据，然后经过视觉映射得到儿化的结果，然后在可视化的结果里面，用户通过一个 is来去实现交互的探索。当然这个可能我们用人多了一点，其他方面我们大家可以使用数据挖掘的方法来去建立抽象的模型。

比如说现在之前的时候，我们可以通过基于这个模型多一点，而现在可能更多的用深度学习的模型，用深度学习的模型，这时候我们涉及到比如说参数的调整，比如说深度学习模型现在调参数一个甚至是一个最重要的工作的部分，对不对？现在已经发展了，这两者之间其实是会有协同，它的协同的地方在哪里？

第一个地方是当我们去构建这个模型的时候，我们需要对这个数据有一个初步的理解，因为我相信大家都去，比如说用机器学习方法之后，我们用这是相机或者我在推波，然后这时候我们会一个很重要的方式，就是我们要去提取teacher就是teacher and generation对不对？

这是一个很重要的方面，可能现在有了深度学习之后，你会发现可以说你认为是一个现在就已经他们它的重要性没有那么强，但是在那个时候我们其实很重要的时候，我们要去推选一个很重要的东西，就是要去理解数据，你要理解数据的话，你看到这个数据你怎么去理解？

你可能一个方面是通过把它变成把它去提取去统计，看一下不同的它的分布怎样，它的平均值怎样，它的帮它怎样的相关性是怎样。当然一个更直观的方法你可以把它转化出来，你可以更好的去理解这些数据。理解了这些数据之后，然后你会辅助大家去构建这个东西，这是第一个方面。

当然还有第二个方面就是你构建了这个模型之后，你也不知道这个模型它的效果怎么样，你需要去更直观的方法来去理解它。

当然现在有了深度学模型之后，其实你会发现去做深度学模型的可解释性是一个很重要的网民的，我其实分支现在是有很多研究者都在去做，我要去把模型的可以解释性去增强，然后有了这两个方面之后，我们可以去协同的去提取到猫尾，提取到数据中的知识，但是可能一开始的时候，你的知识并不是非常的完善，非常的完全的，你需要反过头来去修改数据变化的部分，然后再去重新的去走这个流程，所以这是一个循环，这是一个有一个人在回路中的被感动，这其实就是我不知道大家有没有听说过，客户的方式，这个就是我们一个典型的人在回路的数据分析的一个模型和框架，然后这是在2010年提出的，可以看到这个框架非常的简单，它其实它其实抽象了很多的东西，但是然后进一步的在2014年的时候，我们就有一些研究者他就进一步的把人跟机器，你说你看我要知道人做了什么，机器做了什么，所以这就有了这样的想法更详细一点的图，但是它基本框架没有发生变化，它的左侧还是数据，它的右侧还是knowledge，你会发现为什么我们这堂课我跟大家去强调，我们去做数据分析的最终的结果是要去得到数据的洞察，如果你觉得我们只是做出一些结果，比如说你把它化成82化成22，其实你并没有得到任何发现，你就没有完成从数据到knowledge的整个的流程。

所以说我们要完成这个流程的话，因为我们会发现我们其实就是在完成这样一个流程，从非法到然后这个图中它有怎样的变化，当然它的变化在于比如说我们可以把可视化的整个拍片的话加进来了，从数据到可以让他我没选，然后到租赁社，然后到没有，然后这样的一个流程到规划，这是它的系统，然后到某某的部分我们有了非法有了调整，这一块可以清洗，然后有了 three process的贝塔，就是我们去做数据处理传送贝塔，然后去建立model，然后得到然后当然到更重要的部分是在于这个地方，我们发现人他的这个部分更加的细化去得到到了以后你会发现管理间接得到管理，这是一个保卫生的出路，就是它分了三路，第一个路就是我先去探索，第二部分我先再去验证思路，第三个部分我去提取到位，然后这个是我在第二个抽第二个阶段怎么去抽象和分析框架的这样一个更详细的一个模型。

然后当然还有其他模型他们大同小异，他们的最终的结果其实可以看到最终的结果数据，其实从个人分析到自动的分析，它们之间是协同的，然后最终的结果就是为了去完成科学的决策。

看到这几个图之后，我不知道大家有没有更好的意识，有没有去修正，比如说大家之前对于可视化的一些观点，一开始大家可能认为可视化是什么，我就是将数人是在这边，然后机器是在这边，可是它是一个界面的，可是它就是一些就是一些视图，视图就是去为了把画作人就是去理解，但其实这个界面我帮你放在这个界面，它其实并没有那么窄，它其实覆盖了很多的地方，它其实覆盖到我们的交互，它又覆盖到我们的机器如何去计算。

所以其实数据可视化它的边界它其实并不是说只有一一一个竖向，它其实是比较宽的，就是宽的覆盖到左边又覆盖到右边。然后当然我们从最简单的过程来去看，其实格式化就是什么？我们将数据、任务和场景哪三个方面都如何来去得到一个数据可视化的结果，可视化的结果里面，我们需要去学习数据的映射，需要去学习数据的编码，也需要去学习一些辅助的算法，包括算法。

但是我希望大家记住这一句话，就是你构建的话，其实最根本上也是在去做很多方面的一个权衡，你要去权衡这三个方面，然后得到一个最好的结果，这个就是我们本堂课学习最重要的地方，最需要去学习的一个一个核心和方法。

因为我们在后面再去介绍针对不同类型的表达方法的时候，你会发现没有银色子弹没有说一个主要方法能够解决所有的方式，没有能够解决所有的任务，适用于所有的数据，适用于所有的应用场景，这时候你需要根据这三个方面来去做一个如何去做选择这个最核心要做的事情，但是可能你会觉得为什么我要去做这件事情，为什么不用自动化的方法来行动？

当然企业有很多自动化的方法去辅助你来去做，这也是我们做研究的地方做研究的方式，做研究的时候要去解决的一些问题。

前面我们去介绍可是分析是什么，可是分析的框架部分我们来去看一个具体的就是分析的案例，这个案例是在2017年的时候的话，其他问题确认的是x波一BC也是我们领域最顶级的会议，上面的一个挑战赛，挑战赛的特点是啥？

它是由美国西北太平洋国家实验室每年都会自带你们去召开一个比赛，这个比赛其实它圈子其实它的源头是这样的，它的源头分析当时911事件，然后然后我们从需要去收集罪犯对吧？

然后好像好当时是斯坦福的这种学者，他其实没有去依赖于比如说他们的信息，他就从网上去搜这些信息，然后美国的中庭就发现其实他们收集的找到了罪犯的踪迹，其实跟他们这种FBI有专业渠道来源，其实发生了数据发现的那种冬季其实相差无几的，然后他们就意识到这个方面非常重要，然后美国西北太平洋国家实验室他们就是负责国家安全，然后由此他们就照设立了这样的一个比赛，然后他这个比赛每年都是基于真实的应用场景，应用场景，当然他们会去做一些比如说地名去做一些场景的变化，但是其实这些数据都是真实的。

然后这个应用场景是啥？让从世界上所有的研究人员可视化的研究人员来去分析，从这个数据里面去回答他的问题，然后这样一个比赛的背景，而2017年的时候这个比赛是啥？他们给了一些自然保护区的数据，这些数据非常多，那一年我也参加了，所以这就是我们的比赛。

我们的方案我现在讲，然后它的背景是在自然保护区里面，红鹳蓝雀是一个非常珍稀的保护动物，它在近年来数量大量的减少，然后鸟类科学家这个人他就会对保护区去展开调查，然后来去发现蓝雀它减少的原因是怎样，然后它基于的数据是自然保护区的相关的数据，交通的数据、环境监测的数据、遥感图像的数据，交通数据才有很多的车经过大洋马路去，然后环境监测我去看到不同的地方，还有一些传感系统监测站或者监测它的空气污染这些职能。

然后遥感图像就是我要去看它的图像，遥感它的差别，然后这个数据量很大，然后有个交通数据，这个园区它设置了各种各样的交通的监测站，然后每个监测站就类似于我们的高速公路上那种收费的路口，每次通过就会留下一些记录。

数据的细节就是我有40个监测站，13个月的跨度，17万条的交通记录，以及18多万18万车辆以上的不同不同类型的车，可以看到这个数据量很大了，对不对？

上次我记得我也问过一些同学，当时是哪位同学来说，说到一个我们说大数据是不是包括还是小东同学，当时你说的是数据量是不是要求很大，对不对？这个数据量我做大了，我觉得对还可以小东同学满意了，然后我们可以来去看一下，有了这个数据之后，我们的目的是啥？我要去发现蓝雀减少的原因，其实你并没有什么明确的分析的任务，就是这个方向其实还是很泛的，你要从这个数据里面来去提取，选你可以做一个数据侦探，然后来去怎么去发现异常的模式，这个是你面临的这些数据，如何能高效的去归纳出常见的交通模式，并发现出一些可疑的车辆，这个是第一个问题。

然后当时我们就去做了一个很我觉得现在来看他也不是比较杂乱，但其实我现在对这个方案还是很满意的，其实通过这个过程也让我学习到了很多，这个是我们的不同的视图，这些视图可以关联起来去分析，然后这个是自然保护区它每个点它的监测站每一条路画出来的路就是一个车辆的轨迹，这是车辆的列表，然后你可以去看到它在不同时间点上它有没有一些活动，然后这是时间的视图，然后等等可以看，然后我们可以一步一步的来看，比如说到因为这是一个时空的数据对吧？

还有空间的信息，这节课下来后面我们会去介绍不同的数据类型，其实可以大家可以先去记得这个数据类型，然后反过头来跟着我们的思路来去看这些数据它属于哪些不同的类型，然后他有空间的数据，他肯定有路网来搞清楚的路网车辆的轨迹，这是第一个方面。

第二方面它有时间的信息，因为它是在一段时间里面车辆的交通轨迹和周期性的规律，然后它要有序列的信息，它是轨迹的监测站的序列，因为它每经过一个监测站它都会留下一条记录。Ok正常来说那一条轨迹我们就应该这么来去看，他的人就把它画在地图上就完了，很多的轨迹我就可以都把它画出来，这是一个最正常不过的思路。

我们换个角度来去看这些数据，这就是我跟大家说的数据变换发生重要的意义所在。

我们把这个数据把它变成一个好，把这个监测站变成一个维度，我们可以统计每辆车通过监测站的次数，这样的话其实。

32:55  
就是车。

32:56  
这是多少个监测站，然后这样的话就形成了一个高维的数据对吧？

有了高维的数据，你学完这堂课学到高维数据的时候，你会发现有了高维数据最正常不过的是它的方法就是我们去做投影，因为我们高维数据其实在四十维的空间，我们人我们生活在三维空间里面，其实超过四维，超过三维我们就没办法去理解了，那那么对我们就要把它去降三维二维然后去看，然后我们降维的结果可以发现它有几个更显眼的方式，结果就是它有几个不同的序列，对吧？123456789 10是吧？

10个区分。

然后我们可以去看这10个聚类分别是啥，比如说这是我们把10个聚类把它给看，把它给画出来，在这个监测站上去画出来，可以去看到，可以去看到1个怎样的有趣的共同点，比如说这10个聚类，大家可以看一下，其实这就需要人的一些知识，这个知识是啥？

我发现其实很多的病人他都通过了一个舍近求远的思路，比如说从这儿到这儿，你正常来说你就通过这儿过去就好了，对吧？

从这儿过去就好了，但是往往都舍近求远，然后绕了一圈之后到了目的地，这时候作为分析人员，他很可能他就会有一些疑问，他就会有一些自己的判断，为什么你要舍近求远呢？

这是一个，然后答案我们可以去通过自己的知识来去看到，这是一个交通的要道，然后这个部分是只有护理人员才能通过的，通物流，所以人比如说正常的车辆，一般的车辆它要通过从这儿到对面的话还要绕一大圈，它不可能经过这儿。

所以这次我们得到的知识，记住这个知识，后面我们发现有，这个是我们的一个解释。

然后我们进一步的除了这石堆之后，我们发现还有一小部分。

35:03  
那一。

35:05  
小部分我们把它给挖出来，发现它其实从这儿它就没有绕道，它就直接过去了。

他就是一个违禁的车辆，就是说没有他不是护林员的车，但是他通过了护林员的通道，而且这是从时空轨迹上来去看，而且从空间轨迹上来去看，从时间层面上来去看，我们发现它其实在每个月它有1~3次的行动，仅在凌晨的2:00~5:00的时候行动，而且行程不超过一个小时。

这时候我们人是有知识的，人是有自己的判断的，你就会发现为什么他会做这样的异常很异常对吧？

每个月会很少，然后时间都是在半夜，然后行程非常快去快回，然后进一步的我们从这个空间上会发现它从工业园区到了湖边，这时候人就有判断了，其实这些车辆是不是在运输一些违禁的一些污染物去排放，有了这个判断之后你进一步的去验证，比如说通过新闻报道来去验证禁止碳排放，所以禁止排放，所以他就会偷偷的去运到湖边把污染物排放出来。

然后当然进一步的去验证，我们发现工业园区内它的排放水平是正常的，但是到了湖边它是不正常的。

然后进一步的通过遥感图像发现湖边的植被在生长，化合物的含量在升高，这个。大家可以看到我们这个思路就很符合我们前面所说的这个图，我们有了发明，然后我们找到一块，然后找到一块之后，我们初步得到knowledge，然后后面还要去验证它的假设，然后就得到了这样一个知识，这个就是比较适合用我们而分析用人单独做的方法来去完成的数据分析的工作。

它的特点是我们的任务是不确定的，我们的数据是非常丰富的，非常非常很多，然后同时人也有自己的知识，这时候机器可以比如说我们可以用一些投影的算法，用一些聚类的方法等等辅助我们去得到发现，但是人是最终去判断这个发现来去指导你要去探索的方向，然后最终得到这个结论。

因为如果假设我假设正常，你无论是现在哪怕是就拆PPP大家觉得很透明，PPT4他觉得很聪明，但其实他完全没有这种当然我们现在也在尝试做科研的时候也在去用，比如说我们能不能用这种做基数的逻辑分析能力来去实现数据的探索，这是我们做科研的一些探索，但是他完全是做不了这种很复杂的这种数据的探索的任务，所以这时候人还是很重要的，人要去做人最擅长的事情，来去辅助我们去做很多的东西。

好的，我们完成了我们上节课应该完成的目标就是说我们要去讲清楚我们的规划是什么，然后用案例来去讲清楚规划是什么。而同时我们来去介绍一些比较大家可能去探索的数据集和我们的比较好的一些可视化的案例，大家有问题吗？主要什么？

38:46  
 Ok好的。

38:53  
我们下面就进一步的来讲，

39:05  
谁知

39:09  
我们本次课的第二个部分是来去介绍正式的去介绍我们的如何去完成这样的一个映射，这样的一个映射过程是我们的核心，我们可以先把用户的任务应用的场景先把它给去掉，我们只从数据的角度来去看，单纯从数据我们如何去得到一个比较好的转化结果。

说到这里大家应该能想到，比如说我们刚刚开始讲例子的时候，它有不同的数据类型，还有一些所记的数据，还有一些时间的数据，还有一些多维的数据等等，这些数据的类型就决定了我们要怎样去做视觉映射，视觉的编馆的模式。

然后当然上节课我也请一些同学在我们的课上去画了一些案例，当时我记得我给大家说的时候，我说他比较是比如说节点连接图，它更适合去看到整个层次结构它拓扑结构的信息，然后比如说它的冰柱图它更适合它既能够保持它的属性的信息，它也能够看到它拓扑结构的信息，我是怎样得到这样的结论的，我们这节课主要是回答这样的问题。

首先提到可视化的数据，我们先把这个数据先把它分为从两个方面来讲，第一个方面就是它的数据集的类型，第二个方面就是它的数据的问题，其实这两方面的关系就是我们将这个数据集的类型可以把它解构为这个数据的类型，其实每个数据集都是有很多的数据来去组成，这个就是我们的特点，我们主要考虑的是这种结构化的数据，结构化的数据它主要有这几类，有推广有表格数据，表格数据它可能它进一步的它其实可以抽象为多维的数据，然后还有网络数据，它更多的关联代表的是这种节点链接这种关系关联关系。

比如说我们在刚刚在数据示例里面去讲到了，红楼梦它的不同事件之间的关联，其实就比如说比较典型的网络数据，包括比如说权力的游戏，他们人物之间的关联关系也是网络数据。

然后第二第三个是产品的草数据是这种大家可能觉得更陌生一点，因为可能在我们真实场景中不太容易用得到，但是它跟科学数据比较相关，比如说我们的监测，比如说我们的空气污染监测，其实监测的时候，我们是通过在不同的点上来进行采样得到，每个点上采样得到的一个数据，然后来去构建这样的一个数据集。

然后还有一些时空的数据，这种及格的数据时空的数据，然后对于这种结构化的数据，大家可能会想到，比如说我更想去看这种非结构化数据，我想去看文本看视频看图像的，我这样的数据我应该怎样去处理，其实一般对于非结构化的数据，我们都要把它转化成结构化的数据来去进一步的去可视化。

比如说对于拿文本来去举例，我们可以通过自然语言处理的方法，然后来去汲取实体之间的关系，那是不是就把它转换成这种节点连接的数据？然后当然我们可以通过一些情感的识别来去看，它在不同的态度上，不同的情感上它的数值是怎样的，它的分布情况是什么？但是我们把它转换成这种多维的数据，然后当然我们可以进一步的去挖掘他的情感关键词、概念类别，其实他的这种多维的数据它的就更丰富一点，它就不只包含一些数据性，它可能包含一些类别性的数据等等。

然后对于这就意味着其实你把它其实将文本从非结构化转换成结构化数据之后，你就能够去提出很多的文化方法，比如说这个是一个文本化的供述，他其实就收集了从同质化的一个领域中，其实大所以的这种文本的可视化的一些方式，它有400倍的规模，然后它其实都是将文本转化成不同的结构化数据，进一步的来去设计的它的方法。

43:49  
这个事情对。

43:52  
然后对于数据类型来说，数据集的类型我们有这几类推广或者卖货，有场数据，有几何的数据，然后进一步的对于数据的数据的类型来说，我们可以把这些数据把它解读为很多的方面，而且很多的方面就能够覆盖掉上面的所有数据结构类型，还有数据项有属性、有链接、有位置、有网格，就是马尔他我们刚刚说的老师他来去解构的这种数据的组织方式，我就列举了一个表格来去介绍那些数据集跟数据之间的关联关系是怎样。

比如说表格其实毋庸置疑，它每一行是一个数据项，每一列是一个属性对吧？它其实就代表了数据项目属性。

然后网格网络网络其实比如说我们拿一个举例，拿这种权力的游戏来去举例，每一个人其实他都代表的是一个数据项目，每一个人他其实有很多的属性，比如泰勒斯他的生命长度是多少，他是男是女，然后他在第几集出现，他参与过哪些事件都是它的属性。然后还有一个更重要的方法方面就是它的链接，因为网络其实顾名思义它最重要的就是节点与节点间的连接关系对不了？当然还有它的位置，比如说也有一些网络它是有位置，比如说在时空中网络，比如说我们的拿一个举例，比如说在这种空间上分布的节点，它也有一些关联关系，这可能构成了一些广告，比如说我们的。

45:38  
搜索媒体，比如社交媒体，然后人跟人之间通讯它有网络，但是人它是有一个位置的，就是这个位置。

好的，大家就这样的分析。

45:48  
谁知你在为什么？好好。什么事去的大，谁知去了，谁知。

46:58  
谁知。

47:01  
这个里面的情况可能是他的内容，我们查一下业务。

谢谢的好。的谁知来看起，谁知是。是情况不是非常好。我跟你说谁知可以。谁知20你看这个东西。

48:19  
谁知。

48:24  
也是这样的情况，数据。

48:33  
谁知。

48:47  
看看谁知，谁知。

谁知谁知。

49:16  
谢谢。

49:18  
可能的清楚以后自己出来了，谁知谁知这个问题好的我们的企业，我们的社会，谁知谢谢。谁知。

谁知好。

50:07  
谁知老师的东西。

50:13  
谁知上涨的谁知好好。

谁知谁知，

50:44  
谁知谢谢。

50:51  
好。

50:55  
我们刚刚去介绍了数据集的类型和数据的类型，然后我们用网络网表格来去概括一下数据类型数据集类型之间的关联关系。我们刚才讲到了它有4种不同的数据集的类型，然后5种不同的数据的类型，然后比如说表格还有数据向同行来去看，同业来看不同的数据的属性，而对于网络来说，它其实最重要的特点就是数据项和个体，然后以及个体间的关联关系，就是它的链接，然后当然每个个体上还有很多的属性，然后每个链接上它当然也可以有属性，然后以及它同时每个个体它可能有一个位置的属性，所以这个地方它其实可有可无的。

然后当然厂比如说我们就拿我们去监测污染情况来去举例，然后它其实最重要的就是我要划分不同的网格来进行采样，然后去采样到不同的数据项，然后数据项上面有一些属性属性，一二维的可能是一个高维的张量等等，而进一步的还有一些几何的信息，它最重要的特征就是它有位置，它有了位置之后，然后每个点上它可能有一些属性的信息。然后我们来去介绍不同的这个就是数据项目数据属性，数据项目就是我们来去介绍不同的这种数据的数据的类型，它的一个逐个来去介绍，比如数据项目，我们刚说的就是一个实体，它是离散的，比如说患者，比如说汽车、股票、城市每个都是一个数据项，然后数据项里面它有一些属性，这个属性是通过测量观察记录得到的属性的信息，比如说对患者来说，他有身高血压，它的数据属性，然后汽车来说它的马力，它的制造的厂商就是它的数据属性，这个表格每一行就是一个说明项目每个月有一个它的数据属性的信息，然后链接就是说代表的是不同数据项之间的关联关系，比如说在社交媒体上它的关注的关系，比如说在蛋白质之间相互作用的关系，那位置它就是有空间位置，二维的有三维的，然后位置上的属性和经纬度网格，它其实代表着一种采购策略，我可以去划分不同类型的网络，比如说均匀的网格和非均匀的网格，比如说比如说一些曲线网络等等，它的应用产品，比如说MRI扫描的时候，我要去重建人体，我要去看人体上它不同的点，它的属性的信息是怎样，还有气象量，我们去看不同的空气污染情况，它的监测的结果。

以上就是不同的数，我们简单去介绍它们之间的关联关系，然后我们可以去介绍不同的数据集的属性，他们他们的特征以及可视化，比如说这个是平面的表格数据，它分为平表数据，也分为多维的数据。当然也有一些比如说层次表的数据，比如平均从平表数据，就是说它的表头是一个一维的，它每一页都是一些属性多维的表格，它就是我的建设多个，这是多维的表格。

当然还有层次的表格，比如说我们也对于层次表格来去开展一些工作，它的表头是具有层次结构的，它其实有相应的自己的挑战，其实层次表格数据它的管理数据管理能力更强，因为它可以把很多的表格数据的信息对吧汇总在一起，而同时它有自己的文化的挑战。

后面我们也会去介绍针对表格数据，针对这种高维数据的转化方法，其实最典型的这个就是平均坐标的转化形式，它的特征就是我有很多的维度，然后我每一套现在的这个事情，每一条线上它所取的点就代表了实体在属性上它的属性的信息大小是多少，然后把它连接起来，其实就代表数据集的结果反映了这个数据集的特征，这样的话你就可以去看到这个数据集上它的不同属性之间的关联关系，比如说它有正相关的，还有负相关的对吧？

它有这个就没有什么相关性，就没有这么强，对，然后这样的一些特征，然后这个是平行表坐标的这种形式。

当然还有一些是散点图矩阵的构造形式，我们先说比如说它的好处就是我可以去看相邻两个维度之间的关联关系是怎样的，但是我每个维度都想看的话，其实这个就不太适用，我更想把所有的维度之间的关联关系都把它展示出来，我就可以用平坐标散点图矩阵来去表达。

当然还有我保持表格的结构的信息，然后我将其转化为一些数值去映射到它的颜色上，这样的话你可能更强调的是每个属性之间的大小。

还有对于网络数据来说，我们更直观的方法就是采用这种节点链接的方法来去表达，每一个节点的代表实体，每一条边代表的是一个节点的实体与实体之间的链接关系。

对于网络来说，其实最重要的一个方面就是我如何把它变得更加的清晰，因为大家画图的时候会发现，假设你是手绘图，你会发现什么会影响我们观察这些交叉重叠，会影响我们对原始数据的理解。

对于图数据来说，我们更多的是研究我们怎么去做图布局的方法，图布局的算法，比如说我们把它有一些方法，比如说基于医疗项目组就把这些图把它转换成一个物理的弹簧的模型，然后每个节点之间节点与节点之间有一些斥力，然后每个边有一些相互的吸引之后，如何去维持这个系统达到一个平衡的状态，这样就得到一个比较美观比较清晰的总布局的结果。

后面我们也会去介绍。当然除了这种最简单的图之外，还有有效图，还有没有循环的图，就是树那树也有自己的布置的方式，这个就涉及到当然还有操作，就是说我每一个节点，我每一条边其实不一定连接两个节点，我可能连接很多的节点，这样的话可能节点链接图它就不太适合，实际上我们可以采用操作的方法还是比较难。

另外更进一步的还有二幅图，比如说购买的人跟商品之间的关联关系，其实就是一个典型的二幅图二幅图，二幅图表示的是我的一个集合是一个类型的节点，另一个集合是另一个类型的节点，然后在集合内部他们之间是没有关联的，只有在集合之间有关联的，比如说人购买商品，这个就是一个典型的二部分。

然后还有一些图上有些关节点，那些关节点就代表的是我把这个节点删掉之后，这个图它就不连通了，就导致了很多的非连通的图。对于关节点我们有更需要去更多的去关注。

当然除了疫苗上的问题之外，其实它还有这种矩阵式的，然后以及比如说数图的方法去布局重复结构图力导向的布局图布局跟矩阵式相关矩阵式的这种图布局的方法，它有自己的优缺点，比如这种方法它更容易去表达的是这样的一个它的边没有那么稀疏而没有那么密集，比如说比较它的节点数量没有那么多，然后我其实更好的去看到，比如说有它的集合一个聚类，这个图上更清晰的能够看得到，但是如果这个节点的数量特别多，其实我们后面去学习列表上同步值的方法，列表上同步值算法的时候，你就发现它欧文灯放了一个算法，当它的节点数量超过比如1000个，用我们的电脑来去算的话，它就会比较难得到1个比较好的数据方式，这时候我们可以更好的去采用相关举证的方法去做工具，然后因为它其实不涉及到这么高的计算复杂度，然后但是它的问题问题就在于比如说我去做举证的时候，它其实有很多的它比如说这个图特别的稀疏，你会发现更多的点它都没有很充分的利用得到，也有一些方法可以提高它的空间利用率。

对于树木来说，我们更多的去关注它的层次之间的关联关系，它柱子关系我们就可以采用，如果我更强调它给两个属性信息的时候，我们可以采用这种数轴的方法去补对。

数图的布局也有很多的算法，比如说它其实主要强调这两个方面，第一个方面是我这个数图，我希望这个节点更容易去交互一点，因为我们用传统的数图的布置方法，它更多会导致一些非常狭长的矩形，因为这些狭长的矩形它有很多的缺点，它不是没有办法很容易的去选择。

第二个方面是它也不容易去对比不同矩矩形之间的大小关系，我可以更我的布局方法可能更让它更接近一个东西，这是我的你解决的第一个问题。第二个问题是当我有很多的树图的时候，我假设去掉一个界定，然后又这个树图要重新布局，这时候我要去维持它的稳定性，就我去掉一个节点之后，这个图跟原来的图差别要没有那么大，就是让我更好的去追踪节点之间的关节点之间的对应性，这个就是数图它更多的研究的问题。

当然还有场的数据，那场就是说我的属性值和对象单元之间是关联的，它是来自于一些连续的数据，比如说它的温度、压力、风速等等，然后它主要是比如说也可以是一些测量或者模拟采样的数据，比如说这是大气，对这是大气的数据，然后这是洋流的数据，然后科学家跟海洋这些数据的时候，当然也可以能是模拟来，比如说我们在说你计算机去模拟一些数据，我们想把它汇聚出来的时候，我们就可以用这种场的方式来去采样，每个节点它可能是一个一维的标量，二维的矢量矩阵或者一个三维的张量，这些对应不同的可靠的方法，我们也后面会再去学习。然后每一个节点上它可能有很多不同的属性的类型，但你去采样的时候它也有很多的可能性，比如说它有单一的网格，比如说是均匀的网格，然后它是或者非的，但它都是直线材料或者非均匀的网格或者结构化的网格，它可以是曲线的，然后或者非结构化网格，比如说它是基于这种三角面片，然后再去采样的结果。

然后对于这种场的数据，它汇聚出来的结果其实就是这种体，其实很多都是起诉话题，体会这个的话就是感觉这种场数据场数据采样然后得到，然后它很重要的一个方面就是它的弹力函数，传递函数就是将三维的数据场的数值，然后转换成我们呈现在这个图像上它的一些光学的视觉的参数，然后这样的话就得到不同类型，比如说我去调整它的弹力函数，我更多的去看它的骨头，那么这时候我就把其他的部分的这种参数把它调调的比重更低一点，然后我就更多的去强调它的骨头的信息，然后可能更多的强调它的血管或者它的整体等等，这就是产品阐述它的作用。

然后除了场数据之外，还有几何的数据，几何数据更强调的是对象的形状，它有明确的空间的位置，还有点线曲线、曲面区域体系，然后其实在CAD中，在辅助建筑中是非常常用的。当然还有其他的数据的集合，比如说集合的数据，列表的数据集群的数据等等，然后这些都是一些一些没有前面那么长的，但是也是比较长的一些数据类型的数据结合的类型。

然后前面我们是通过大概的方式来去介绍不同的数据集合，它对应的文化形式有哪些最经典，给大家开了一点点小头，然后讲完数据集。

01:04:24  
数据。

01:04:25  
集合的数据来结构之后，我们可以先介绍数据，数据它有一个它的数据它的一个很重要的方面就是它的数据语义含义，比如说我们单纯的去拿到几个属性的合同，其实我们也不知道它的具体代表的含义是什么，所以我们需要通过每一个属性它的一些原数据来去更进一步的去描述它，比如说比如说集合我们去把它的上面的去概括之后，其实整个的含义就不清楚了，所以每一个数据它一定需要一些原数据这个数据才能完整。

数据的类型就这几类数据项目链接属性位置和网格，然后比如说我们先说数据属性的类型，这是一个非常重要的方面。每一个属性我们属性的类型我们分为三三个类别，第一个类别是定类型，就是类别性的数据属性，它的一个特点就是我是不可比的。比如说我的水，比如说我的水果的类别就是苹果香蕉还是之间是不可比的，比较大小是不能的对吧？男女也不能比较大小，电影的类型也不能比较大小，文件类型也不能比较大小，最重要的特征就是不可比，这是一个定类型。

然后定序型的有序型的我们又分为了4两类，第一类就是它可以比较大小，但是它不支持运算，比如说时间，当然时间差也可以，比如说它的衣服的尺码的大小，它其实是不支持，比如说 Smlm是不能减XL是不能减m的，对吧？这个是不能比较大小，比如说定量的，这是第二类。

第三类就是它既可以比较大小又能支持运算，比如说我们的长度重量，我都可以去进行加减运算，这些都是没问题的。

因为比如说一个苹果它是一一一斤，然后有一个100个苹果，它是把所有的它每个苹果它的属性加起来得到的结果是多少，这些其实是非常重要的，如果大家真的去做这种去做视觉的映射的时候，你会发现不同的视觉的属性，它适合的这种视觉的我们视频一会会讲到我们视觉的通道是有差异的，有了这些定义之后，我们就可以去选择更适合的属性来定义什么，然后这样就得到比较好的转化的结果。

当然除了这种基本的还之外，还有一些周期性的，比如说时间，比如每天的时间，它是一个周期性的，每天都在去循环，循环之后我们就会可能更多的去强调，比如说今天的8点跟明天8点这个区域，后天8点这个区域区别，这时候我们可以采用一些特殊的映射方法，然后比如说这种环境的映射方法，然后还有一些聚类的结构，有了前面的这种解数据集类型以及数据类型的解解释之后，我们可以回过头来去看这些基本的格拉形式，你就相当于看山不是山，现在我们可以再一眼去看看当时他这种格拉形式是怎么来去映射。

比如说最基本的折线图，这个是时间，这个是它的价格，然后当我们之前我们可能就觉得它其实是简单的一个的方法，但其实它的横轴x轴是时间，它是一个这种数值型的属性，然后y轴也是价格，它是一个数值型的属性，我们把它映射到它的空间信息当中，然后就得到了时间折线图。

然后当然还有一个这种最经典的这两个折线图，它是怎么去映射的？它每一个折线它代表的是一个折线图的信息的，横轴代表的是时间，纵轴代表的是它的属性的信息，每一个都是它的数字信息，对吧？

好，当然还有一个折线图它是这样子画的，然后它也是毕竟但是它在之间去映射它的颜色，去映射的是它的进口还是出口，这样我们就能看到不同进口进出口他们之间的差别，比如说之前是进口后面的出口，然后它的进出口的类别就是一个定类的属性，它不是一个定量，所以我就更适合英特尔它的颜色属性。

对，这个就是按年份 y轴代表的是它的货物量，然后颜色代表的是它的进出口进进出口的顺逆差，而朝代的宽度代表是顺逆差的大小，然后这样一个很好的结果。

然后当然对于数图来说，我们这个数字代表是它的股票数据还是股票的一个市场上它股票不同的股票它的涨跌情况以及它的市值的大小，它每一个红色和绿色代表着它涨跌的情况，它每一个矩形的大小代表的是这个公司，然后它的市值的大小，然后我们可以去看到市值它不同的板块市值大小之间的区别，然后颜色可以看到它涨跌情况的区别，这个映射方法我们怎么去结构呢？

其实它的位置就代表着它相互的关联关系，代表映射的是他父子之间的关系。

因为它包含关系，比如说这个是属于整个他们使用人死了，把你call x消费者，然后这样的话我就可以去看到它属于类别，就是在映射到这个位置这个位置范围内，然后它的颜色映射的是它的涨跌的情况，它的大小映射的是它的市值的大小这样。

然后我们可以看到不同的它股票的情况，涨跌的情况，通过这样一个图来去看这个。我们简单就把它结构得到的结果，比如说矩形的面积代表的是它的市值，矩形的位置代表是它市场部门市值，然后颜色代表是它损失的根本。争议就代表着他争议的量是怎样，他可能有一个增长下降的比例。

然后再回过头来我们去看拿破仑的行军图它是怎么去结构，比如说它是一个时空间的数据，因为它有一个精度维度的信息，左边代表是法国，右边代表是德国，然后宽度代表是它军队的数量，然后下面是它的温度的信息，它恒久代表其他时间，动作代表其他温度，然后把它合并起来就得到这样一个可靠的结果。

我们看了这么多之后，其实我看了这么多之后，你会发现现在我们看可视化形式，我们就不再是说看到它一个整体了，我们要看的是它原始的数据是怎样，它映射到怎样的视觉通道上去。

我们大家会去想我要怎么去决定视觉通道的映射方法，其实主要有两个方面，第一个方面就是文化的表达力，第二个方面是它的有效性。

第一个。

01:12:05  
是。

01:12:07  
视觉编码，其实它还是一个非常大的空间，因为你每一个不同的出品，它都可以映射到很多视觉通道，我怎么去决定？第一个就是它的一致性的原则，就是图像的属性，视觉变量应该与这个数据的属性之间是相互一致的。第二个是它的重要性，排序的原则就是我要去看到这就跟你的任务是相关的，如果你更多强调的是属性的大小的关系，数比如说它的一个数值型的属性，我就把这个数值型的属性去映射到我人更容易去辨别的这样的一个视觉通道，以更有效的方法去编码最重要的信息。

所以其实简单来说就是两个方面，第一个方面我要把所有的属性都考虑在内。第二方面我要是对于属性做一个排序，哪一个属性最重要，我要去把这些属性映射到更高效的视觉通道，这样得到的结果就是一个比较好的结果。它的表达力。

01:13:13  
是怎样来去度量的呢？

01:13:16  
这样是一个这是我们一个典型的例子，就是说在这个是128的数据，它每一个这是一个表格数据，每一行代表是一个数据对象，每一纵线代表是一个数据属性，然后它还有类别性的属性的信息，然后不能表达，比如说我们可以去看一些不太好的数据转让的形式，比如说它是这种多对一的数据转让的形式，然后我可以把它映射到一个映射到一个很一维的层面上，然后它的颜色代表着它的类别，然后它的每个点代表数据对象，然后这样的话其实你通过这个数据属性其实看不到任何的信息，当然你也可以把它映射到这样的一个结果上，就是说它不同的类别，然后它的花瓣不同的属性是怎样的，但是你通过这两个，我不知道大家之前有没有了解过属性数据的数据它的含义，其实我们更多的是去看它的花落和花瓣之间的关联关系是怎么样，这是我们的目的，但是如果我们只是把这样来去用这种不太好的表现形式把它画出来，其实你是得不到这样一个结果。

01:14:37  
然后。

01:14:41  
当然这是一个不太好的结果，然后还有不太好的结果，比如说条形图的长度，其实这个是不同的汽车它的一个它的一个不同的汽车，每个代表着一个不同的汽车，然后横向代表是不同的产出的国家，其实你会发现我把产出的国家把它映射到它的高度上，这样合理吗？

其实它其实就带来了不可一些不真实的内容，因为国家与国家之间是不可比的，你假设我们看到这个可是结果你有没有发现这边是比较高的，那边是比较低的，我会误以为这边比较好，但其实国家与国家之间并没有可比的这些属性的关联关系，所以你其实用这种格式化的运算方法就是不合理。

01:15:37  
对吧？是吧？

01:15:39  
这样的话就可以更好的去解释为什么我们要去选择这种更有效视觉工作的品质。

因为你选差了选错了这个通道会给你带来一些不必要的会导致一些不必要的不正确的理解。这个就是一个比较正确的结果，我们可以去看它的关联关系，这是前面我们主要是介绍的是它的表达力对吧？就是说你不能把原本有的信息把它给隐藏到像这样，当然你也不能把原本没有的信息你把它加上，会导致一些不必要的误解，这样就是一些不好的结果。

01:16:22  
还有。

01:16:24  
第二个方面就是它的有效性，有效性就是说我要对于不同的视觉通道，我要去识别到哪一个是重要的数据，哪一个是不重要的数据。然后有了这个不重要的属性，不重要的属性之后，我就可以来去更好的去选择我去定制的方式。

我们前面就介绍三种不同的视觉的数据的类型，比如说这是定量的，这是定性的，这是认识。这是可比可计算，这是可比的不能计算，这是不能比的。

然后这三种不同的视觉类型，它有不同的视觉通道的位置、长度、角度、倾斜度、面积、体积，然后以及它的密集的程度，然后饱和度、颜色，然后纹理连接关系，包含关系它的形状，这些不同的视觉通道，然后当时在18 19 86年的时候就有研究者他来去说不同的视觉通道我应该怎么去预测它比较合理，这样得到一个结果，他们当时是开展大量的用户体验说这样的视觉通道他怎么去在最适合的方式怎么去映射，比如说我们可以看到不同的这种定量的定序的和定类的这三个位置都是最有效的是吧？

所以说位置是一个非常宝贵的视觉通道，然后进一步的就可以看到对于定量的来说，它的最有效的是它的高度、长度、角度，然后与倾斜度。然后当然这种定性的它就有一些差别，比如说比如说我们的可以很好的去理解，比如说定性的我们的大小什么大小之间不可比，其实可以把它映射到它的深浅颜色上，其实是非常正确的一个选择。

然后但是对，但是对于定量的来说，把它映射到颜色上，它就不容易去比较它们相互之间的差别。

然后这是一些定类的这种属性，然后有了这个结果之后，我们就可以去做一些自动化的转让结果。当时这篇工作第二个的抗日精神，就是说我可以将有了这些数据，有了视觉通道，它的排序我就可以把它自动的去得到一些可视化的结果上得到这个结果上去，然后来去得到一些比较有效的可视化的形式。

当然现在也有很多的方法，我们如何去把映射的过程自动化，然后比如说这个是我们后可以去大家来可以去考虑的一个案例，这个是我们可以做一个小的测试，这个是二手车的一个案例，它有很多的属性、价格、里程、重量、车况等等，然后我们可以把它大家可以去想一下，我们应该怎么去采用什么样的方法来规划这样的一个数据。

然后可以考虑的就是第五条我们的视觉通道是这样的来去排序的，大家应该怎么来申报？F或者这样子也价格自然就是一个定量的是吧？可以加起来可以去计算里程，也是一个定量重量，也是定量车况的。什么原因？

01:20:08  
有特别。

01:20:10  
好好还算好不好？特别的差，然后这样的话你这样一个数据集，我应该采用什么样的方法来支持是吧？大家可以来去思考。对第一个我要去大家去提一提。

01:20:25  
谁知。

01:20:40  
我们要考虑它的表达力和有效性，表达力就是说我要把所有的信息都要排除，应适当去，但是有效性就是说我要把最重要的视觉通道去映射到最重要的最重要的属性映射最重要的视觉通道上去，最重要的视觉通道是怎样的？

是这个就是不同的属性的类型，然后它适合的这种通道是怎样的。让大家可以提问一下，就是说你不清楚，比如说否则与什么的之间有什么区别，当然大家觉得有什么区别，所以当时我也不清楚为什么他们两个还有区别？

其实position就是说我是对齐的，假设这是一个轴对吧？这个position比如说这个这个是它的位置。这个就是它的位置，但是任何事其实不一定对称，比如说这是我们的高度，但这个高度可以是这样的， A这个是b这个就是它的不同的高度，但其实你可以看到我是这两个对齐它就是它的不对称高，这两个不对齐不一定对齐，就是他们的我自然的假设和历史肯定是理论，它更高效，假设我的差别没有那么大，然后我这两个视觉通道差别很小，怎么说这个是让它差别，比如说它是1万差异可能会差别特别小，比如说二差异它可能会比较明显，他一个人这么高，一个人这么高。

01:22:36  
你。

01:22:37  
要去对齐的话你还能看到他是吧？

假设你是这么高，他们并没有对齐，你是这么高，你就分辨不出它们的差别，所以这就是为什么这种可以使它更理论它更高效一点。

但是角度它其实他相比认识来说就会更容易消化，然后知道其他的高度，角度就是语言品什么样的善意，怎么样老不他。

可以看一看，比如说这种波动，这个人还好。然后 l就是它面积，然后体积，其实这样这是一个其实前人他们做了很多的实验，得到了一个排序准则，这个有了这个排序准则，这个排序准则还是很重要的，因为我们就可以自动化的来进行治疗。

对，有很多的自动化的方法，你就可以清晰的可以得到。大家来看，就是说我们有处理器，我们应该怎么去映射比较好的结果？

位置一定是很重要的。然后然后还有什么你可以去看，因为有些还是重复的，你比如说昂高的通稿的可能不是很好，不是很适合去映射，面积也不是很适合映射等等，我们就可以来去想，我们原来还可以手画一下，其实一个很简单的测试，我只是想让大家去理解，我其实是要经过思考的过程来去得到一个比较好的结果的。

01:24:28  
好好。谁知

01:24:51  
谁知。

01:25:01  
不知道。

01:25:05  
你当然可以依据你自己的考虑，比如说在这个场景下，比如比如真正购买二手车，你觉得什么更重要，你觉得什么不重要，这就是跟任务是相关的，再把场景带入来去考虑。

01:25:24  
谁知42号。

01:25:30  
这也。

01:25:32  
是我们的。

01:25:33  
谁知。

01:25:36  
这个时候。

01:25:42  
等到46亿分钟的时间。

01:25:45  
的数据，谁知谢谢。

01:26:25  
我已经画出来了。因为。

01:26:32  
这个是。

01:26:35  
在这个场景下比较适合的这种咱们不同的场景和不同适合的形式，不同解决的任务。

比如说你要去做去买车，最重要的是啥呢？第一个就是价格对吧？因为我肯定是把价格放在第二第一位的，第二个就是我们排序了，第二个就是它的里程，因为它里程越长，说明。

01:26:59  
它这个车越旧。

01:27:03  
那么好，第三个是。第三个是它的重量，其实重量就在这个场景下就没有那么重要了。

至少对于我来说我觉得没有那么重要，而车况当然也是一个很比较重要，但是它是类别性的。

01:27:24  
它有几个不同的类别。

01:27:26  
那么这样的话我们就可以回过头来去看，其实每个点它有两个位置的属性点，如果你的属性是最重要的，你就把最重要的属性保护，把最重要的主线把给这两个重要的分配视觉通道，给这两个重要的主线，然后这样的话你就可以得到一个三点，一个这是一个它的里程就是它的价格，然后。

01:27:55  
然后。

01:27:57  
还有它的重量其实没有那么重要，其实你可以给一个我们可以回过头来去看这个数字性的属性，它还适合它能用到哪里，它的角度其实这个是不太适合的，没有可以去验证的地方，漏点没有它 Area就是它的面积可以用圆形来定表达，把它的大量车的重量是多少。

第三个方面就是它的一个类别性的属性，就是它的车况车况就是类别性属性的，你可以看到除了除了劳美劳贝塔就是它是一个定序型的属性标准，定序型的属性，因为它有一个从好到坏的区别，它除了位置之外，它的深浅，然后它的色彩其实也是一个比较好的方法方式，你就可以把它映射到它的颜色上去。

比如说红绿就表示车况比较好，红表示车况比较差，也符合我们自己对车况的理解。这时候有了这个之后，我们就可以来去选择这个车了，比如说我们更想去选择的是啥？价格比较低，然后公里数也比较少的，车况比较好的这样一个会有一个车，它还有一个看它的颜色就是红色的那种涂料，它有红色的涂料，其实可能比较好的，这样他可能也就得到你想要的结果了。

这个就是一个是非常小的案例，只是说咱们大家有一个思考的过程，大家以后思考的话，你就可以带着这个思考我有哪些属性，我属性映射到哪些不同的输液通道，然后这样得到同样的结果，这就是我们一个构建可靠的过程，这其实也是我们前面想说的有了数据，然后我单纯通过数据之后，我们进行一个老数据的转化结果上去，这样一个过程。

Ok其实我们本堂课主要就介绍了通过数据进行得到数据可视化结果这样一个这样一个不同的这样一个方式，然后大家需要掌握的是不同的数据的集的类型，有表格，有网络，更重要是它的关联关系，有采样，然后有几何，其实大家后面再去找数据的时候，你就可以带入到数据集的类型中来去思考你所找的数据，你感兴趣的数据是怎样的数据类型，还有哪样的数据的数据集的类型，对不对？哪样的数据的类型。

然后在后面我们去介绍的时候，我们会分别去介绍不同数据集的类型，它有哪些哪样的可视化的方法，有了可视化方法之后，你就可以来去开展你的课程，你的作业设计，然后对这个就是我们本堂课要介绍的东西，然后大家再去映射的时候，你还要去考虑的两个方面，第一个方面就是它的。第一个方面就是他的表达能力，第二个方面就是它的有效性，就是说你不能无中生有，或者说他没有这个数据，你要把这个数据映射上来，但是你也不能说忽略了，你也不能说有了数据，你也没有运营这个游戏。

然后当然你有决定了去映射哪些数据之后，你要去对于这些数据的属性进行排序，将更重要的属性去映射到更高效的视觉通道上，然后这样的话你就可以将这个范围，我们可视化空间的范围能够进一步的去缩小之后，你从这个空间中去找到一个比较适合的可视化的结果。对，当然其实大家觉得这个过程我还要去思考，这个过程是比较繁琐的，大家可以去看这个论文是。

01:32:04  
在。

01:32:06  
80年代的时候，其实我们就有很多的研究者在去考虑，我们怎么采用一些自动化的方法来去得到比较高效的转化形式，当然现在这个方面也一直在去探索的问题。

这个就是我们针对数据来去介绍相关内容，我们本堂课还有几分钟的时间。

01:32:32  
44。

01:32:33  
分钟的时间是吧？我们可以可能开始下一个，可能讲的不是太多。

01:32:41  
开始下一个步骤。

01:32:43  
我们可以来去讨论一下，比如说大家更感兴趣什么，然后我可以给大家出主意，比如说你们更感兴趣什么主题，然后有哪些不同的数据，然后我们来可以来去探讨探索探讨一下。

因为我觉得大家可以不用当做你们去完成一个作业，你们可以把我当成也是你们的队友，我们来去共同的去完成你们想然后去完成了数据可视化的工作，对。

然后大家有没有什么想聊就聊的，就说你更感兴趣什么，你更感兴趣，你有什么感兴趣的方向，或者你更感兴趣主题，比如说你特别喜欢看短视频。

01:33:28  
这个没问题。那么我们可以看到这个事情也大量。谁知。

01:33:44  
销售情况是指什么的？销售情况，专辑的销售情况。你有找到相关的数据吗？

01:33:54  
像什么网上有人支付？刚开始就我具体的这个问题。

01:34:03  
可以挺好，所以你更感兴趣的是音乐对吧？你想去看不同类型不同人的音乐，他们的点赞量消费情况一点，然后比如说他几乎出专辑的情况怎么样是吧？

对，你就可以带着这个数据来去考虑，带一个数据来考虑我们今天学的内容，它属于哪些不同类型的数据，哪些都可以，为什么有兴趣了，然后进一步的考虑它是怎样进行设计，然后以及你还要带着你的任务来，我觉得你更感兴趣什么样的任务，你更想从这些数据中去发现什么，这也是一个驱动力来去做这件事情。

01:34:40  
好，谢谢你说一说。

01:34:42  
你给我好好的谢谢你们都非常好，能够有效。还有我突然想到大家可能想过这个怎么去实现，比如说我们在课上是看到ubc他们去做一些比较好的案例，对不对？这些案例是基于什么来去实现的？其实都是基于教育思维的一个垃圾作用，就是第一次为我们。

01:35:11  
注册这样的国家的谁知。

01:35:17  
这个是当时他提出的时候还在斯坦福，但是现在他在华盛顿大学就是一个可视化领域非常非常厉害的人，非常厉害的老师他提出来的，其实这个库他其实改变了可视化的实现的方式，他一开始我们实现都是通过对大家来去实现。

01:35:37  
有了这个之后我们就可以通过依附于案子来去实现它。

01:35:42  
主要的编程语言是教会，然后后面我会来带大家去学，大家一开始不会没关系，我们会在课上去教大家一个例子，所以我会尽量的把课程，比如说我们要学的一些内容可以尽量往前赶一赶，给大家在课上去学编程的时间多留一点。

01:36:00  
带大家去完成多少年的历史。

01:36:03  
然后其实我们在这边所学习的，比如说这些文化的形式，你完全就可以通过我们在课上学的东西就可以去构建出来这种。

好的，我们今天的课程就到这里。

01:36:17  
然后还有提醒大家一下，我们补课应该是在这周日对吧？

01:36:20  
周日上周四的课，所以大家不要忘记，好谢谢。

01:36:25  
谢谢大家给到我们的这种认真的方向。

01:36:30  
好像不是下周就是下周的价格的报价。

01:36:35  
已经下周的10个点就下节课谁知好。

01:36:39  
现对下节课是下周一，然后下面首先基本上讲对，就是下周的三天给他的信息这么说。

01:36:47  
然后我就是把这个问题说过了。

01:36:50  
他虽然说我的力量。

01:36:52  
我知道我们还追着好处，但是好像也不是说尽可能越多越好，比如说你当时如果。

01:37:00  
说我所以也采用一种精神，完成了一个三年的时间目标，这个效果的非常成功。

对。

01:37:08  
你这句话说得非常对，下节课我们介绍其实很多时候带着妈妈不希望你们三五年的。

01:37:17  
因为他们会有一些多的，对，所以你三个字你就他把关系当地。

01:37:22  
的宣传是的。

01:37:24  
这是另外一个原则，我刚才讲的就是他会有一些想法。

01:37:28  
谁知然后来。

01:37:31  
而且这个地方的什么的。

01:37:34  
我没有我妈出去。

01:37:36  
比如说家里的东西，包括我们都是发展很好，谁知。谁知。

01:37:47  
我们没问题。

01:37:50  
小小小。

还有什么来说说什么。

01:37:56  
而且包括企业的这种感觉了。

01:38:01  
这个数据都了。

01:38:03  
所以你觉得是适合做的导出的规定用的效果了，将整个人将发现。

01:38:17  
大家注意这种系统的情况要不要在15月11日还有昨天。

01:38:26  
怎么问一下，不好意思。

01:38:28  
我们下节课我们都可以。

01:38:30  
你给我们的绩效。

01:38:33  
的新项目我们应该会好起来。

01:38:38  
行了，谁知。

01:38:40  
然后到了。

01:38:41  
目前的所以要做设计的谁知。

01:38:50  
而且的小的好。

01:38:54  
是这些。