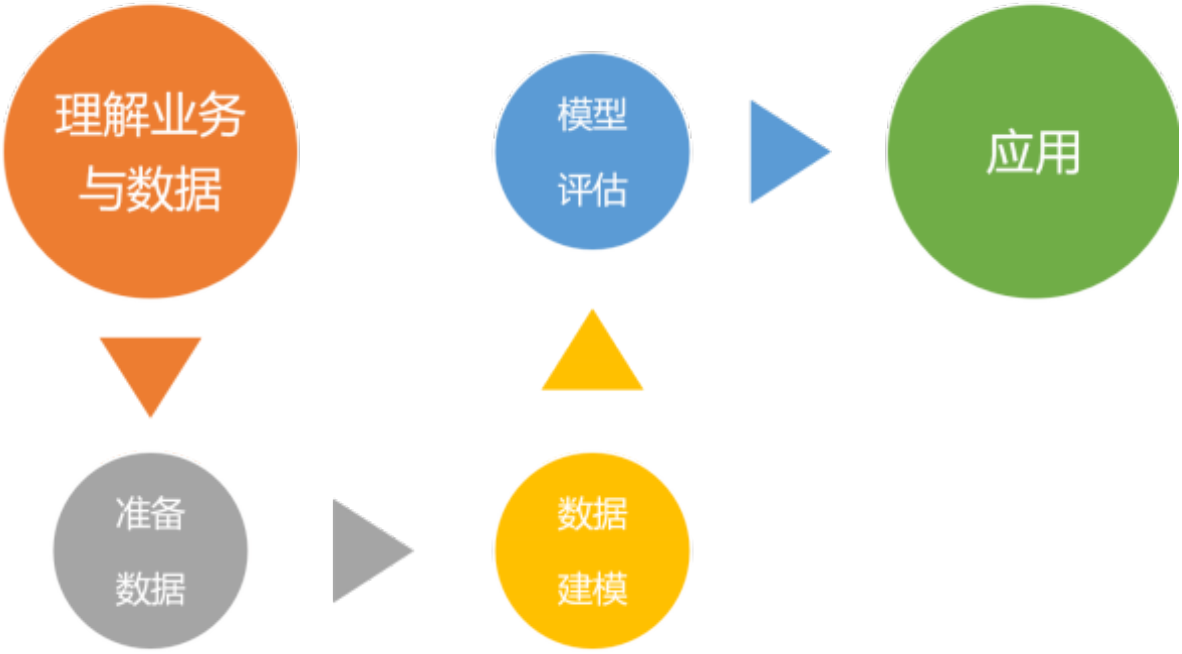
从这一课时开始，我们就要学习数据挖掘的具体步骤了。

这里面的每一个步骤看似都是循规蹈矩的，但是在实际的工作中，通常都会有各种各样的限制，遇到各种各样的问题，我会尽量把自己工作中的一些惨痛经历和教训总结出来。如果你看后能有一点点感悟或者收获，甚至可以给你带来一点点快乐，那我的痛苦就没有白费。

根据 CRISP-DM（Cross-Industry Standard Process for Data Mining，跨行业数据挖掘标准流程）这个前人总结的方法论，我们在开始做数据挖掘的时候需要有一些前置的准备。



数据挖掘的流程图

这些准备有两个方面：**业务理解**和**数据理解**。

数据挖掘是一种方法，在这个方法中虽然说技术是一个重点，但归根到底，你的方法要去解决问题，空有技术而没有问题是没办法去实施方法的。所以，在开始数据挖掘的时候，要确保你对业务及其数据有充分的理解。如果你仅仅按照自己的想法，而没有对业务与数据进行充分理解，很可能导致你的理解与业务有较大的偏差，所做出的结果与业务的需求不符合导致返工，或者是因为对数据的了解不全面导致缺乏关键数据而影响最后的结果，甚至是任务难以进行，最后项目流产。

经过这一步的准备，要明确你对于要解决的问题的所有可知信息。但是，要想真的理解业务、理解数据，我觉得还有一个更重要的，那就是思想准备，所以，在这里我把前置准备分成了三项：**思想问题**、**业务背景与目标**、**把握数据**。

思想问题

1. 避免对业务的轻视

要做什么样的人，**要先去按照那样的人去思考**，而不是要先像那样的人。比如，你要健身减肥，很多人都是先去买一套健身服装、健身器材，然后办一张健身房卡，但是自己的思想并没有转变成一个健身人士的思想，那么最终健身减肥的效果也就可想而知了。所以你做数据挖掘，一定要避免这样的思想问题——我学了很多的算法，穿着程序员的衣服，背着程序员的电脑，我就是一个优秀的数据挖掘工程师了。

这个事情看起来好像很容易，很多人会说：“我当然知道业务的需求是什么样的，不然我干吗要做数据挖掘呢？”

在实际的公司组织中，有些公司的数据分析师或者算法工程师是与业务部门在一起的，而有些则是分开的。其实不管是否在同一个部门，如果数据挖掘人员没能够**真正理解业务场景与挖掘需求**，很有可能会与业务人员产生分歧，以至于你觉得做了很多的工作，每天都加班加点，你的挖掘项目充满了技术含量，数据充分、算法丰富、结果翔实，最后业务却不认账，说你做得不好，效果没有达到预期，做出来的东西对业务的价值不大。

轻视业务是很容易犯的问题。

我在刚做这些工作的时候也犯过不少类似的错误。一个业务方提需求，而我也没有仔细去询问业务的真正想法，按照我自己对业务的理解去做模型。最后呢，如果拿准确率这样的线下指标去看似乎效果很不错，但是放进业务去做测评却无法满足业务的需要，导致不得不推倒重做。

所以要始终牢记，数据挖掘的本质是一种方法，这个方法要去解决问题，一定要源于业务需求，服务业务需求。如果脱离具体业务去做，那做得再好、再漂亮也没有太大的价值。

如果要做一个成功的数据挖掘项目，你就要去深入**学习业务**，明白业务的关键点，在项目的需求阶段与业务方进行充分的沟通，在发现偏差时**及时调整**，甚至在制定 OKR 的时候也与业务方来共同制定，这样在做项目的时候才不会出现南辕北辙的问题。不要觉得你是一名技术人员，学习业务对你帮助不大。

2. 明白可以为和不可以为

做技术，有一个很容易进入的误区，那就是相信 “**技术万能**”，技术可以解决一切问题。一个业务需求来了，你明白了业务的要求以及目标，**还需要明白数据挖掘要解决的点在哪里。**

“技术万能” 的工程师会相信数据挖掘或者说算法可以解决任何问题，对于各种各样的难题，认为只要有数据就可以解决。

然而事实上，技术在业务上绝不是万能的，甚至由于公司不同、业务不同、流程不同，所做的数据挖掘流程和数据挖掘目标也千差万别。

比如说你在做一个 OTA 酒店消歧的项目，在酒店业务中有一个痛点就是不同的供应商提供的酒店信息可能存在一些区别，以至于需要消耗大量的人工去做比对决策。

这个问题场景看似很符合数据挖掘的解决范畴，然而实际上却有各种各样的情况：

- 数据可能是残缺的导致无法使用算法处理；
- 不同供应商提供的同一家酒店名称可能是中文，可能是英文，甚至是日文、法文、泰文，不同语种间无法使用同样的模型来解决，而如果每一个语种都做一个模型又没有足够的数据做支撑；
- 不同的供应商提供的信息可能是不对等的，有的供应商会提供电话和邮件，有的则没有这个字段，然而这些对准确率影响也很大，这也无法使用一套解决方案来完全解决所有问题。

总结一下，说数据挖掘不是万能的主要有两个原因。

- **第一个是数据不完美。**

虽然数据挖掘的理念很美好，但现实总是残酷的，完美的数据是不存在的，至少现在是不存在的。每一个公司都只是掌握了部分数据，有些甚至没有多少数据，还需要去外面爬取数据来进行处理。从总体上来看，似乎我们的数据量很大、很充足，但数据的真实性、准确性、完整性具体到每一条数据的时候或者某一个需求的时候，是不完美的，甚至是匮乏的。你要解决这些问题需要付出大量的工作，甚至超出了业务本身，这就会造成入不敷出的情况，这个项目开展的必要性可能就要受到质疑了。

- **第二个是业务条件不完美。**

数据挖掘项目通常都是跨团队的协作项目，譬如说我带领的数据挖掘部门既不生产数据，也不存储数据，但是我们却是对数据处理应用最多的部门，这当然有赖于业务部门以及其他技术部门之间的配合和协作。

再说上面那个酒店业务的例子，如果我们投入大量的人力，比如说经过长达一年的研究和调试，当然可以产出一个效果更好的模型，甚至是直接输出结果。然而这个周期对于业务来说太久了，业务没有办法等待一年以获得一个足够好的结果。况且由于数据的限制，这一年后半段时间的努力可能只有很小幅度的提升。所以我们又跟业务进行了更深入的讨论，以更改目标，最后确定了目标是提升酒店运营人员的效率，而不是直接输出一个完整的结果。

这样我们就可以在三个月甚至是更短的时间内产出一个效果还不错，而业务也可以使用的模型方案，同时加入业务开发的一些流程，最后我们的项目降低了酒店运营 60% 的人力成本。

所以说，**数据挖掘只能在有限的资源与条件下去提供最大化的解决方案**，不要忘记我们的初衷，去解决业务需求，而不是以 "**技术万能**" 为导向，最终导致项目陷入泥潭消耗过大，甚至是项目流产，而偏离了我们解决业务问题的初心。

要避免这种问题的产生，需要与业务方进行深入的沟通，同时对你所掌握的数据有充分的认识，对业务的难点和重点有明确的区分。建立需求多方评估机制，让业务专家与技术专家参与进来，评估需求的合理性以及你的数据情况，确认问题是否可以通过数据挖掘得到有效解决；或者是对需求进行拆解，以最大化在数据限制和业务限制前提下的项目效果。

业务背景与目标

解决了上面两个想法上的误区后，要明确你的业务需求就变得相对简单了。因为你的思想已经发生了变化，你已经是一个具备业务视角的数据挖掘工程师了。

自然而然的，你在进行数据挖掘之初就要去明确业务背景和业务目标，更好地契合业务的需要。

数据挖掘是一种方法，需求的产生必然是因为某种分析需求、某个问题或者某个业务目标的需求。如果你一开始就不能对问题进行准确的定位，那么后面该如何使用合适的数据选择合适的算法，都是无稽之谈。

假设你现在是一个自媒体平台，自媒体作者会在上面发布文章，很多用户会来看这些内容，从而产生互动行为，比如说点赞、收藏、分享、评论等。这些会刺激作者继续创作，而作者持续发布好内容又会吸引更多的用户来浏览，在这个环节中，作者的贡献是一个重要的部分。

所以这里业务提出了一个需求，要对发布内容的自媒体做一个**贡献度评级模型**。这个目标虽然确定了，但是这个贡献度到底该如何去衡量？对于一个作者来说，他的贡献度体现在他的内容上，但是 CTR（点击率）高的内容贡献度高，还是有独特观点的内容贡献度高？或是能引发讨论的内容贡献度高，还是技术深度更深的内容贡献度高？是发布内容的频率高贡献度高，还是发布的内容够长贡献度高？这里面还有很多需要思考的事情。

如果数据挖掘工程师自行去理解业务，那很可能出现偏差，和业务方的需求产生分歧。所以这时你就应该展开沟通，并成立专家小组来对目标进行评审。

在沟通的过程中了解到，我们的业务背景是在打造品牌影响力的时候，发现很多用户对我们的内容产生了质疑，因为有些作者为了提升自己的点击率，故意使用一些标题、低俗图片等手段，并且频繁地生产无意义的内容，造成了用户的反感。在这样一个背景下，你的业务方希望能够对作者形成一种分级制度，让那些写有深度、有内涵内容的作者能够被评为更高的级别，而那些标题党作者的级别则会降低。

那么接下来，我们对数据的收集就要围绕着这个点去展开了。

把握数据

在核对好需求之后，紧接着你就要对数据来进行了解，这一步有点类似可行性分析。巧妇难为无米之炊，如果你的数据无法支撑挖掘需求，那这件事也就没办法解决了。

所以作为一个数据挖掘工程师，还需要对你要用到的数据了如指掌。收集、存储、转换数据都是十分重要的环节，如果这些环节存在问题，那么整个项目的进度都会被拖慢。

在这个步骤中，你要考虑所有可以用到的数据，哪些可以用来做你的模型，可以用来回答业务的问题，哪些对这个需求来说是没有什么价值的。

数据的质量、数量、可靠性、完整性，以及部分数据缺失是否会导致模型的效果不好，某些关键数据是否会严重影响业务问题等，这些问题都是你需要搞清楚的。

从粗粒度到细粒度，你对数据的认知应该有这样几个层级。

1. 是否有数据

是否有这样一个数据集来支持你做这样一个模型，来完成这样一个需求，来回答业务的问题。

2. 有多少数据

光有数据还不行，还得看到底有多少数据，是一条、十条，还是一万条、一亿条，数量的不同也会影响你的处理方式。

3. 是什么样的数据

亦或者说，你的数据都有哪些属性可以被用到。比如上面的例子，一条数据就是一个作者所写的一条内容、各种互动指标等信息。到了这一层，你需要考虑的是这些维度是否可以支持完成业务需求，是否与所提出的问题有关系。

4. 标签

这个是针对特定的项目，比如说有监督学习任务，那么每条数据都需要有结果的标注，这也是你的模型或者算法要学习的结果。如果只有数据，而没有对应的标签标注，那机器学习也是比较困难的。

确认了这些数据的问题后，你才可以说对数据有了初步的了解，接下来，可以按部就班地进行下一步：**准备数据**了。

看完了这一课时的内容后，不知道你是不是改变了对数据挖掘的看法？想想自己做的业务需求是什么样子的，你是否能做好对应的准备呢？如果你是业务方，那又该如何跟工程师做好沟通呢？

总结

这一课时讲解了数据挖掘步骤的第一步，如果用一个词来总结的话，那就是 “做好准备”。这里所讲的准备分成三个方面：

- **思想准备**，确保自己已经具备了一个专业的数据挖掘工程师的思维模式；
- **理解业务**，确保与业务需求方的充分沟通，对业务需求的充分理解，知道什么可以做，什么不可以做；
- **理解数据**，确保对可以掌握的数据有全面的了解，知道哪些数据有用，哪些数据没用。

我觉得作为一个工程师，通常在沟通方面可能会有所欠缺，比如我自己就是这样。所以如果你想要在沟通方面有所提升，我可以推荐两本书给你《非暴力沟通》《高难度沟通》，有时间可以读一下哦。