主持人：徐亚波应该是广州的，我觉得他跟我们分享的不仅仅是技术，刚才我们在聊组织、企业变革，他讲的过程中听得比较多的是技术上的东西，我们似乎经常听到，但是如何对你的组织，如何对你的企业进行革命与时俱进，甚至做痛苦的去做一些定，我相信不仅仅是我们公司，很多公司都遇到这样的问题，内部变革的时候主力不知道怎么做，谢谢他给我们一个很好的体例。

接下来有请广州佩升云数网络科技有限公司副总经理吴长江《在大数据语境下数据获取方式创新和整合》。

吴长江：大家下午好，我是来自佩升云数的吴长江，很高兴有机会跟大家分享我们在数据收集以及大数据分析方面的想法、探索和阶段性成果。

我想跟大家分享我对数据应用以及分析的理解。我们调研行业实际上就是在跟信息接收者和信息产生者打交道，人每时每刻都通过视觉、听觉感知器获得信息，获得信息之后大脑会对它进行处理，进行消化、提取、识别、重构形成自己的想法、意念、观念，甚至会形成信念，对他的消费行为产生消费决策。

在这个发生的同时，这并不是终点，他把信息收集到脑子之后，他还会有一个动作出去，所以他也是信息的生产者跟发出者，也会通过人际沟通、自媒体，通过采用声音、文字、图片、图像等方面把意思表达出去。在自我表达的同时，能够对他所在的圈层产生影响。

我们这个行业做的东西是什么？我们要把信息收集回来，我

们要看消费者在做消费决策的时候，他的心理的动线是什么，他心智的模式是什么，以及他的行为路线是什么样的。

我们要看哪些信息更重要？这些信息和他的营销决策，以及购买行为之间，它们相互关系和影响机制是什么样的，我们会向客户做建议。一是内向的，怎么样提高他决策的质量，怎么样使他的落地更加到位，提高工作效率以及效果。二是向外的，就是要告诉他们应该生产什么样的信息，这些信息以什么样的形式、介质、载体让受众知道，从而影响他决策过程，然后再介入、操纵他的购买行为，当然这个操纵是在善意和法律范围内做事情。我们还要帮助客户监控用户所传递出去的信息，去做管理，做引导，让人际信息传递放大，能向比较正向和良性的方向走。

我们还要关注行业的发展和行业的信息，以及竞争对手的信息，使我们的客户能够跟上趋势，使他的营销策略在这么多的噪音里面更加有锐力度和竞争力。

回过头来看，信息收集就会变得非常关键，现在信息的捕获手段基本上就是大数据和调研产生的小数据。大数据和小数据的关系这两天是热词，我也想分享我的看法。我觉得大数据跟小数据之间并不是替代的关系，而是强互补的关系，为什么有这样的结论？

第一，这两种数据的特质和它回答的问题是不同的，大数据更多是行为数据，在回答 What、Where、How 的时候有自己的优势，特别是这几个问题的答案加上在实际上变化的时候，能更加

体现它的优势。但是在回答 why 和 how 其他问题的时候是非常无力的，这正好就是我们调研的优势。我们调研不光是了解行为，还要挖掘行为背后的原因，一连串的东西都可以说出来。

第二，我们再往深层次考虑，大数据和小数据的区别在什么地方？我的结论是这样的，它们的底层思维和应用场景是有差异的，我们在做调研的时候，我们会对行业有非常大的信心，我们会认为人的行为是由他的思想决定，如果我们能摸到他的思想，能够介入他的思想，改变他的行为，我就能改变他的行为。所以我要收集这样的信息，我要找自己的机会。这样出来的数据会更加深层，所以能够在策略方面应用更加有利。大数据底层的思维逻辑和信念就是相关，小数据是归因，大数据是找相关。举一个例子，前段时间 P2P 比较火的时候，他们用大数据的方式识别潜在客户，寻找防欺诈的东西，现在有一批潜客，这批潜客长成这样，我就推定长这样的人就可能会成为欺诈客户或者潜客，真的是这样吗？效果确实非常好，我觉得这些数据应该用在战术层面或者在销售促进上面。

第三，我们要做分析，做分析就是看数据之间的关系。我们的调研数据，特别是定量数据是绝对同源的，数据之间的交互分析怎么玩都可以，所以会有故事和解释，会更加支撑战略层面的决策。大数据当然会有局部的同源，但是在整体看起来，会有一些需要再思考的地方。

第四，我们所说的代表性问题，在收集数据的时候有两个要

考虑，第一个就是数据质量，第二个就是数据的代表性，能否投射，能够让我对总体做认识，基于这个认识作出决策。我们的调研实际上有比较坚实的理论基础做支撑，我们可以通过抽样方法、题库的改进提高可信度，使用者和我们的信心都有。但是对于大数据就会存在一些问题。

第五，投入和产出方面，因为调研已经发展了几十年，整个生态已经非常成熟，我们很细的环节都有公司来做，非常成熟。所以提供这样的比较高质量的数据，基本上一个比较好的研究经理就可以打天下了，他的能力可以覆盖到设计、数据收集、数据处理、分析和展示。对于大数据来讲就有很大的不同，它需要我们做的是知道这个信息在哪，这些信息是以什么样的介质来承载，我怎么样把信息抽取出来，变成我能够读和能够进行数据分析的东西，而且我们还要考虑怎么把这些数据变成让机器认识，通过它来高效进行分析。所以他需要的服务提供者，它的能力边界就要比那个宽很多，我认为这也是一个非常重要的点。

我们今天要分享的是什么？我们怎么样把信息从介质里面提出来，怎么让人能看懂。另外，我们怎么让机器能够看懂，这是我们努力的方向。我们公司是从大事着眼，从小事着手进行研发。

接下来分享语音数据的应用和文本分析。

语音方面，我们更多是在调研方面的应用，包括定性、定量研究。定量研究就是开放题的语音化，另外就是笔录的生成，我

们这两个系统都已经开发出来，已经在用了。

语音答题系统方面，我们这个语音答题系统，实际上是我们公司在线调研系统的一部分，是在它的基础上开发出来的。我们的在线调研系统在被访者端可以支撑三大系统，三个屏都是可以使用的。另外它也会有一些特点，应用场景比较多，所有在线调研都存在这样的功能。

我想跟大家分享另外一点，我们也有做数据中台业务，在跟客户沟通的时候，我们会发现除了数据孤岛、数据融通等等问题之外，还有一个很大的问题，就是数据会有残缺，他们在业务流程沉淀下来的工作流的数据是很完整的，它的销售数据也是比较丰富的。但是在消费者的数据方面很少，特别是跟消费者的互动界面方面。实际上有很多客户有获取这种数据的能力，我们在线调研系统可以赋能客户拥有这个能力，让他自己产生数据，使整个业务闭环形成，所以数据中台的推进难点跟这个有关系。

另外，分发渠道也比较丰富，设计方面我们也做了努力，增加了人性化的设计，会有一些动画在里边，好比说图可以是动图，不是静态的。还有排序题，点完之后并没有在原位，可以跑到前面，会自动进行排序。

基本分析功能和可视化已经实现，在分析方面，我们不管是现在的大数据风格，我会分成两类：一类是大数据类，很酷，很炫，数据非常多。我们看新闻的时候，我的动作往往是绕行的，太多东西使我的焦点失交，使我的分析失焦，我没有办法得出一

个结论。二是 BI 的公司，他们也很好，但是他们也会有他们的问题，他们也是很重，虽然已经往轻量化方向走了。客户在做分析的时候，需求是比较简单的，现在做了非常轻量化的工具，非常方便的做多层次的交叉、过滤，用起来非常方便。

（PPT 图）就是我们公司在线问卷，大家可以体验一下，我希望大家能够找到一些不同。

语音 AI 问卷更多的针对开放题，之前我们研究过其他封闭题的应用，封闭题效率不太适合。开放题一般来讲回答的问题会比较开放，我们很难囚禁他所有的想象，所以必须在开放题去实现。我们做研究设计的时候一般很小心，尽量控制开放题的量，为什么？一是开放题出来之后，获取数据的方式一般要么是一问一答，要么就是自填，这两个都会存在信息流失或者不完整的问题。人讲话一般的语速是 100-200 个字，记录就是 30-40 个字，差不多就是 2 秒钟写一个字，会有 5 倍的差异，信息会有流失。而且在操作过程中来不及，会让被访者等一下再记，那个时候受访者思路被打断，存在信息流失的可能，或者说自己少讲一点，这是访问的方式存在的可能的问题。

另外，自填也会有一定风险，人现在越来越懒，让他写字很难，而且有的字也根本不会写。

开放题的信息拿回来之后，分析是一个非常头疼的事，在拿到数据报告的时候，谁在决定最后数据报告拿到的时间点，实际上是开放题，后面需要编码，需要复核，需要录入，语音 AI 问

卷系统想要解决这部分所有的问题，我们还想去到编码。

（PPT 图）就是我们 AI 问卷的组成，包括四个部分，包括收音，把被访者回答的内容收回来，还有做语音文件的内部传送，还要做后期语音识别、编码工作。

语音识别我们都是用大场的语音识别引擎，从我们测试的结果来看，他们做出来的结果我觉得还是非常让人佩服的，它的准确率一般在 80%以上，我估计大家有很多人已经调用过这些大场的语音转录引擎的 API，可能大家觉得可读性不好，问题在什么地方？一是里面会有噪音，一个极端的例子，有的时候我们把语音输进去，大段的文字是没有的，甚至 80%都没有，这跟语音的质量有直接关系，所以就需要降噪，降噪有很多的软件和工具。比如说音乐行业发展很好，有很多东西都可以拿来用，当然它不能解决所有问题，还需要用算法进行辅助。

我们做了这个工作之后，它的正确率还是不能去到我们能够接受的点上，为什么？准确率很高，达到 90%，但是读起来感觉还是读不懂，为什么？因为里面的“主角”丢失了。从自然语言角度讲，更多是命名是题词，比如说品牌、人名、英文等内容。英文其实是很让人头疼，中文可能不同方言对同一个字会有比较一致化的读音，但是英文不同，不同人的单词发音也不同，有时候对字母也不一样。当这些东西混在一起的时候，会更加有问题。所以这不是大场的问题，应该是我们做垂直应用的时候应该解决的问题。所以，我们有我们的算法，在引擎方面我们用的是集成

应用，在后边会有文本纠错算法。

自动编码部分，后面我会有详细的介绍，主要跟大家说我们最后输出的结果和人工的记录方式，采用同样的编码程序，其他的工序都不变。在这种对比之下，我们会发现实际上它确实在信息量上边会有流失，增益挺能打动人的，有 30%的提升。

文本分析部分，我们已经做了六年，AI 方面我非常赞成之前楚总的看法，有时候你会觉得 AI 很好，有时候又觉得不好，有时候确实有很大的问题。AI 在文本分析领域里边比其他地方难很多，比如说语音识别、图像识别比较好，因为它最底层的单位就是数据，比如说声波、像素点，像素点有三个通道再加上其他的深度，直接就变成了数据，直接是可读、可分析。但是文本分析就不一样了，非常复杂，一个词在不同的的语境之下可能就不一样，不同搭配又会出现千差万别的意思，消歧在文本分析当中是一个非常大的问题。

回过头来说，我们的能力积累下来会有四个方面：一是爬虫系统，我们爬虫这一块可以做到全网爬虫，能够实现所见即所得，包括 APP，当然不是所有的 APP，我们试过大部分的 APP 是没有问题的。二是文本分析系统，下面会有技术体系的介绍。三是词库，现在文本分析非常不成熟，很多部分是需要有行业的词库，文本分析在现在的阶段，如果文本分析要做出性能比较好的模型出来，技术重要，经验会更加重要。四是可视化。六年当中，我们积累了很多的技术，促使我们一步一步往前走。（PPT 图）就

是我们数据抓取的能力。

文本分析系统，如果说对 AI 的技术有所了解的人，就会发现文本分析系统的技术体系非常复杂、繁杂。在一些关键的点上边，就分词这一个事来讲都没有一个通用的算法，就算集中之后的算法都不能达到非常好的效果。所以垂直领域应用是难题，词库就变得非常关键了。但是可喜的是现在的一些大场在这块开发出性能非常优越的技术模型，我们都用过了，确实让人耳目一新，性能的提升不是几个点的问题，当然我们解决的问题未必所有的都是一样的，因为文本分析在技术方面还有另外一个特点，它对特殊需求、特别的应用场景，你要用的模型是不一样的，在做迁移的时候，不同的项目之间的技术迁移有时候难度很大，复制性真的不高。

自动编码方面，主要有四个流程：一是对语料解析。二是文本表示。三是模型训练。四是自动编码。

第一步，非常关键，就是说在一句话里边，或者在一个段落里边，要把关键的语句和关键词找出来。当你找关键词的时候，实质上你就在审视某些信息，包括它的词性、语境、上下文的信息。所以我们在用语言表达的时候，我们不仅仅是凭字面值判断它的意思，必须要看它的前后和它所在的语境，所以它的信息就非常复杂。即使拿词来讲，都会是一个很大的问题。你现在是用一个词表示，还是双词做表示，或者用更多的词做表示，在不同的应用场景下效果不一样，有的场景下我们就不用词，直接拿字

了，这也是我们非常吃惊的地方，这个方面的性能非常优越。

第二步，我们把词找出来之后，这些信息机器是读不懂的，我们要把它转成机器能够读懂的东西，把它变成数据，这个时候就要用专有技术来做，现在基本上主流就词的向量，技术上是通用的，但它对算力有一定的要求。一般来说一个词可能得用几百个数字表示，当你再加上其他信息的时候，我们分析的时候会加上词的信息，词性的信息、正向和负向都要加进去，判断组合起来效果更好，还是单用这个词会更好。

第三步，建模。自然语言处理感觉会更复杂，你可能要把模型拆成几层，这个几层不是我们说的神经网络会有多少层的网络，它根本不是这个级别的，会比它更高一层，解决不同的问题要有共同的模型去解决，它们前后顺序是什么样的，都要有需求来做帮助，它肯定会有错误的结果，怎么有试错模型提高它的准确率。所以在应用的时候，它是一个技术整体的组合、集成。即使是在某一个节点上也要用这个技术来做，我们在比较大的项目里边有300 多个最底层的细项指标，我们用了规则、机器学习和深度学习，在机器学习我们差不多试了十几个模型，最后剩下八个模型。深度学习是用了三个，不是说不同的需求场景跟需求所需要匹配的技术不一样，这个之外还有另外一个很重要的点，在这一道题里面，不同的码要用不同的模型。

自动编码，可以输出结果，现在我们可以做到 91%的准确率。整体来讲，在做编码的时候有很大的问题，就是码不平衡，在看

开放题出来结果的时候，你可以看到有的码很大，有的码很小。这些小的码，一般 3%以上的码都要出的，会导致信任级很少。另外一个障碍，好几百个样本量没有办法做出编码的，因为自动编码是通过学习来的，学习的前提就是要给这些模型喂语料，喂标注进去。

展示方面，也就是可视化方面，现在的展示方式，特别是大数据风格的展示会有很大的问题。大家可能去读文章的时候，会有一个非常漂亮的图，但是这个图里面的信息量实在太大了，基本上我们很难根据它得出结论，这是大数据风格的特点。

另外，它是断的，在上一页跟下一页之间是断掉的。我自己的感觉，大数据本来最底层分析逻辑就是相关，展示的又是断掉的，我们怎么去诠释和使用？这个时候就需要考虑一个问题。现在我们做的展示，基本的思路也是从这个角度去考虑的。我拿的是电商数据，电商整个网页，包括标题、关键参数、详情页、消费者之后的评论，这些信息我都融到这一个图里面去。

你在去看这个信息的时候，你可以把相关的信息同步展示到这里面，做分析不需要跳来跳去，或者在同一个面板里面不同图之间的切换，这样理解我认为会更加友好，当然它不是最完美的。我们放在一个特定场景，比如说护肤品行业的某个细分品类，我们非常关注它，我们想看这个品类是什么样的。我们首先会关注到底在这个小的细分市场里面会有哪些品牌，点开之后我们就可以看到不同的血统的品牌在里面，在不同血统品牌里面，来自这

个国家的品牌能够有多少个，（PPT 图）展示的数据是它 SKU 的数量，我们还有其他的同样展示方式，也可以从销售额、价格、销量角度去看，然后看它们的聚集度是什么样的。现在我只放到三级，其实可以无限的放下去，可以把每个 SKU 具体情况列出来，可以直接看到 SKU 里面具体的信息。

我们会考虑到底在这个细分市场的获利能力是什么样的，这跟价格有直接的关系，然后我们去看价格。可以看到它的分段，也可以看到下面更细的分段，它的颗粒度可以自己做定义。我们还会看它的销量，看看它的盘子有多大，它的销售额能够有多少。在看的过程中你可以不断往下走，在走的过程中，你可能看到它的上一级是归属到信息的哪一个方面，我们也可以看到不同的方面，可以看到有哪些品牌在，也知道它的销量怎么样，价格分布怎么样，产品下面的信息，它有什么样的卖点，用户使用之后的感受是什么样，我们可以在这个图里面全部展示，不需要再做跳转。

另外，当对这个细分市场有了解之后，我们就要聚焦到某个品牌，从品牌的角度看，这个起点你可以任意切换，你可能还会有其他的角度，如果你想开发一个新品，你还可以市面上拿这个来做卖点的新品有哪些，有哪些品牌，价格是什么样，销量又是什么样的，当把实际信息加入进去之后，我们就知道这个品类的销量随着时间有什么样的变化，是不断往上走的品类，还是在没落的品类。

从品牌的角度，我们看到品牌下边有多少 SKU，它的价格、销量是什么样的，还有它的功能，消费者使用之后的口碑，还有使用的应用场景。

我的介绍基本到此结束，有很多同行到我们展台上讨论，特别是语音文件方面给了我们很多建议，非常感谢大家，希望这次大会结束之后不是结束，是我们沟通和合作的起点，非常感谢大家！