前言

医疗行业是数据密集型产业，数据积累亘古存在。然而，在数据的应用水平上，医疗行业远远落后于互联网、金融和电信等信息化程度更好的行业。

峰瑞资本生物医疗技术团队从数据产生、数据处理、数据消费的角度分析了医疗数据产业链。分析显示，医院、诊所等专业医疗机构和保险机构仍然是医疗数据产生的最重要来源，来自手机 App 和可穿戴设备的数据开始提升数据的完整性、连续性和准确性；数据处理是个系统工程，包括清洗、整理、分析等标准环节，对数据结构化提出了

更高要求；截至目前，为医疗数据买单的是 B 端的医疗机构、药企和保险公司，让 C 端的病人和医生为数据付费目前还不现实。

美国的医疗体制相对市场化，对医疗体系的投入巨大，使其在技术、服务和流程等支柱产业，都可以成为中国医疗产业发展的远景参照物。近几年，医疗数据产业在美国发展迅速。峰瑞资本生物医疗技术团队挑选了 4 家有代表性的美国医疗大数据公司（Flatiron、IBM Watson Oncology、IMS Health Oncology、Palantir）做案例分析。

如果您在医疗健康领域有创业想法，可以与本文作者、峰瑞资本医疗组早期项目负责人

王蕾（lei@freesvc.com）和谭验（yantan@freesvc.com）联系。加入峰瑞资本前，王蕾曾任职于美国最大的医药咨询和市场调研公司 IMS Health，负责为国际和中国本土医药企业提供战略和战术咨询。谭验曾是 Tamr 早期员工，大数据整合平台公司 Tamr 由 2014 年图灵奖获得者、美国数据库专家 Michael Stonebraker 创办。

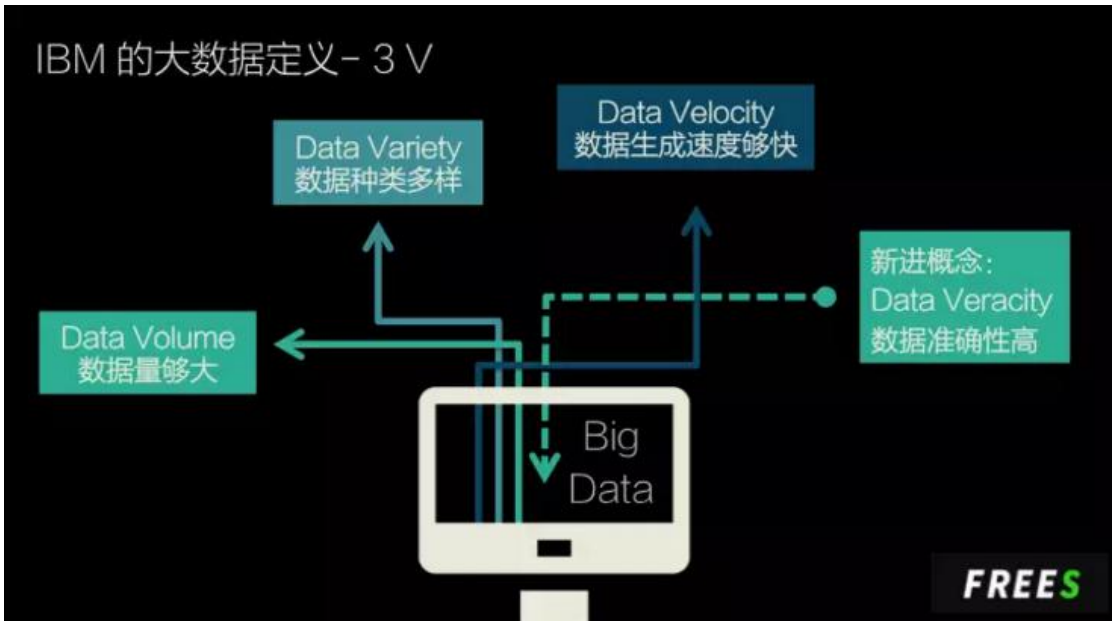
大数据产业的出现
和医疗数据投资策略分析

文 / 谭验（yantan@freesvc.com）
王蕾（lei@freesvc.com）

/ 01 /

IBM 用 3V 定义大数据

IBM 最早提出了大数据的 3V 定义。3V 是 Volume，Variety，Velocity。



Volume 比较好理解，因为大数据本身的“大” 代表了数据数量的巨大。数据量越

来越大的原因很多，其中一个是现在机器和网络每天都在生成大量的数据。据统计，我们现在每两天产生的数据量约等于自人类文明开始到 2013 年的数据量的总和。

第二个特征是 Variety，多样化。多样化主要指不同的数据来源和种类。传统意义上的数据主要来自类似 excel 的表格和数据库。现在人类能够分析各种形式和类型的数据，比如电子邮件、图片、视频、音频、监控仪器，等等。

第三个特征是 Velocity，即数据生成的速度。比如，互联网上数据的生成是以秒甚至

毫秒来计算的。再比如，基因测序仪、网络监控的录像，都在随时随地产生大量数据。

以上 3 个 V 是公认的大数据定义。在 2013 年波士顿的大数据峰会上，Express Scripts 的首席数据科学家 Inderpal Bhandar 提出了 Veracity 的概念。

Veracity 主要是指数据是否有偏差、数据噪声有多大，以及是否有异常值。当业界大量积累各种来源的数据时，数据是否准确变成一个非常重大的问题，否则最后就是“Garbage in，Garbage out”。

峰瑞观点（freesvc）

- ✓ 从以上对大数据的描述可以发现，大数据对数据存储、数据传输和数据处理这 3 方面的能力提出了挑战。
- ✓ 企业在数据产生和处理端也逐渐出现了一些变化。企业开始存储海量数据，数据传输并分布式地存储到数据中心，数据在

云端进行处理和分析，通过网络端进行数据的呈现并指导商业决策。

/ 02 /

大数据的产业链分析

得益于计算能力的快速增长、数据传输能力的增长和成本的下降，以及数据储存成本的下降，大数据获得了极大的发展。

上游数据的产生

大数据产业的最上游是数据的产生，这包括了数据的定义和数据的搜集。数据的定义顾名思义就是定义哪些是数据。例如在搜索广告出现之前，用户点击链接本身并不产生任何价值，也就不被定义为数据。数据定义产生之后，就开始快速、准确、有效地收集数据。

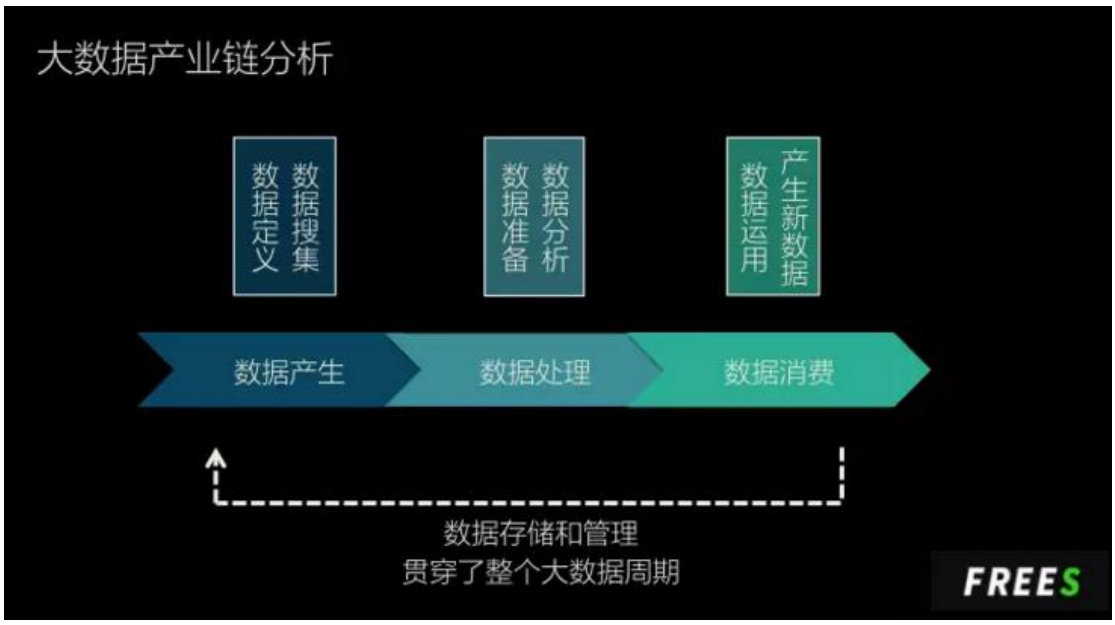
中游数据的处理

大数据产业的中游是数据处理，其中包括了数据的准备，例如数据清洗和整合，以及数据分析，例如数据建模、可视化呈现，等等。

下游数据的消费

大数据产业的最下游是数据消费，例如利用数据指导商业决策，指导商业决策之后产生的结果本身又成为了新的数据，因此数据的消费和数据的产生形成了一个闭环。

在整个大数据产业的所有环节中都存在数据存储和数据管理，这两个技术贯穿了整个大数据的周期。



/ 03 /

数据驱动型企业结构的分析

在一个通过数据驱动的商业环境中，企业组织或者技术组织结构一般分为以下 3 个逻辑板块。从底层到上层分别是 Data engineering（数据工程），Data sciences（数据科学）和 Decision sciences（决策科学）。

下层数据平台：通用性平台为主，完整解决方案，开源解决方案

最底层是工程性的工作，主要指对于数据底层的工程性技术解决方案，例如对原始数据进行清洗、验证和纠正，数据储存和调取。在这一层有很多的开源解决方案和系统集成服务商。

这一步的目的是收集和整理大量数据，把它变成便于数据科学家使用的方式。大部分企业或者工程师把 80% 的时间花在了这一步 。美国财富杂志前几天公布的数据显示，美国企业每年在大数据服务上的花费是 40 亿美金左右，其中 40% 花在了数据整合和清洗上。可以说，整个数据工程在时间和花费上都占据了很重要的位置。

中层算法和数据呈现：通用性算法接口，行业专业知识，开源解决方案

处于中间层的是数据科学，这可能是大家最常听到的一个领域。现在很热的人工智能、深度学习，都属于这一层。这一层的作用是通过数据建立起对某个问题的模型。比如说，通过历史数据建立起天气预报模型，或者通过大量病理数据建立起疾病的预测或者诊断模型。

开源社区的发展让很多非常复杂的算法模型变得非常容易使用，极大地促进了数据科学的发展。数据科学家可以很快地验证预测模型，并使用到实际的商业项目中。目前的

解决方案主要是开源方案，一些商业 API 以及企业内部的私有数据计算框架等等。

上层商业决策：深入的行业专业知识，商业洞察，内部决策和外部咨询

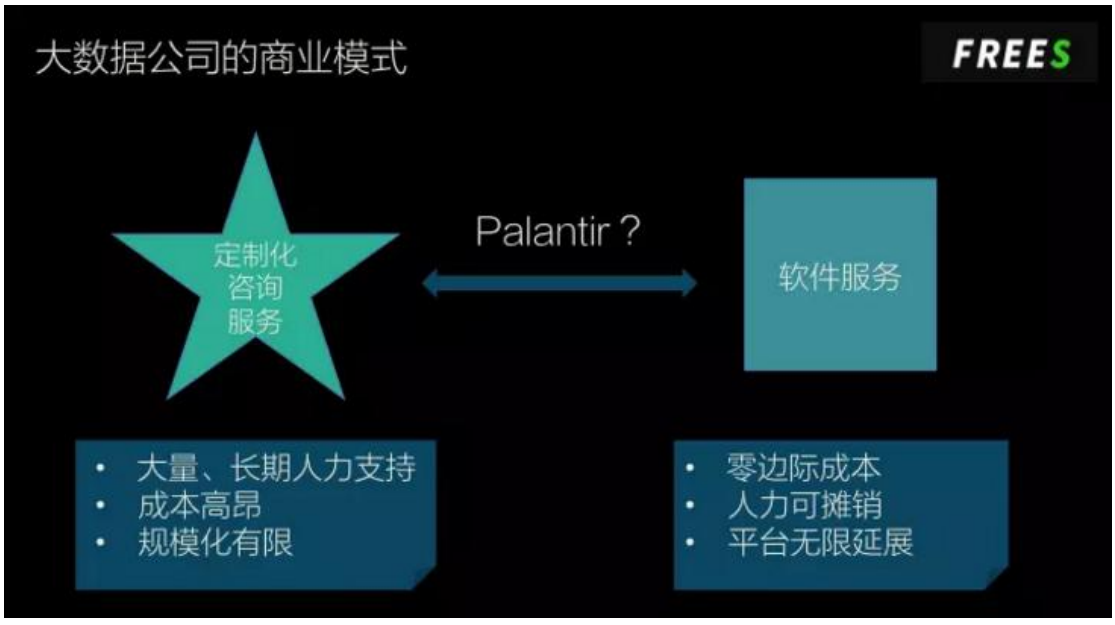
第三层是决策科学，它是数据的最顶层，也是实际产生商业价值的。比如我们预测明天要下雨，这个预测的价值在于，得到这个信息的商家第二天可以把伞放到更明显的地方，以增加购买量。这样就产生了商业价值。这只是一个简单的例子，实际情况要复杂很多。比如，很多游戏中，机器可以根据玩家玩游戏的时间、模式，来预测用户是否对游戏感兴趣，一旦发现玩家对游戏的兴趣正在

减弱，就会自动进行一些奖励措施，比如奖励装备、奖励点数来留住玩家，都是商业决策的范畴。

/ 04 /

大数据企业的商业模式：在咨询和软件服务中徘徊

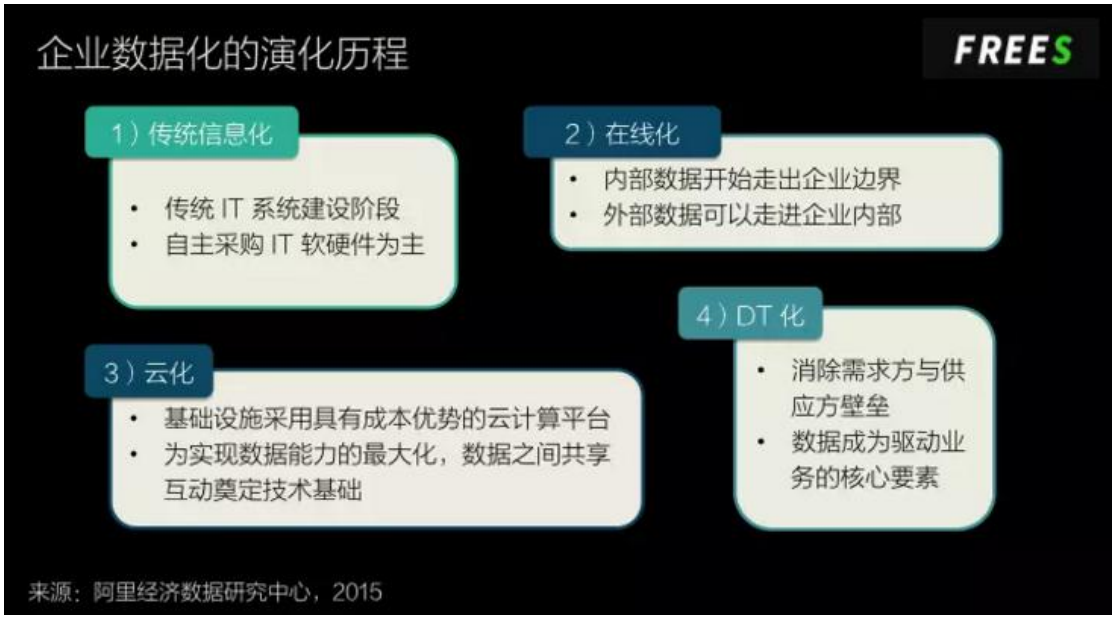
大数据的价值往往通过商业价值来体现，而不同公司的商业逻辑往往有很大的区别。因此，大数据公司往往在咨询模式和软件模式之间徘徊。



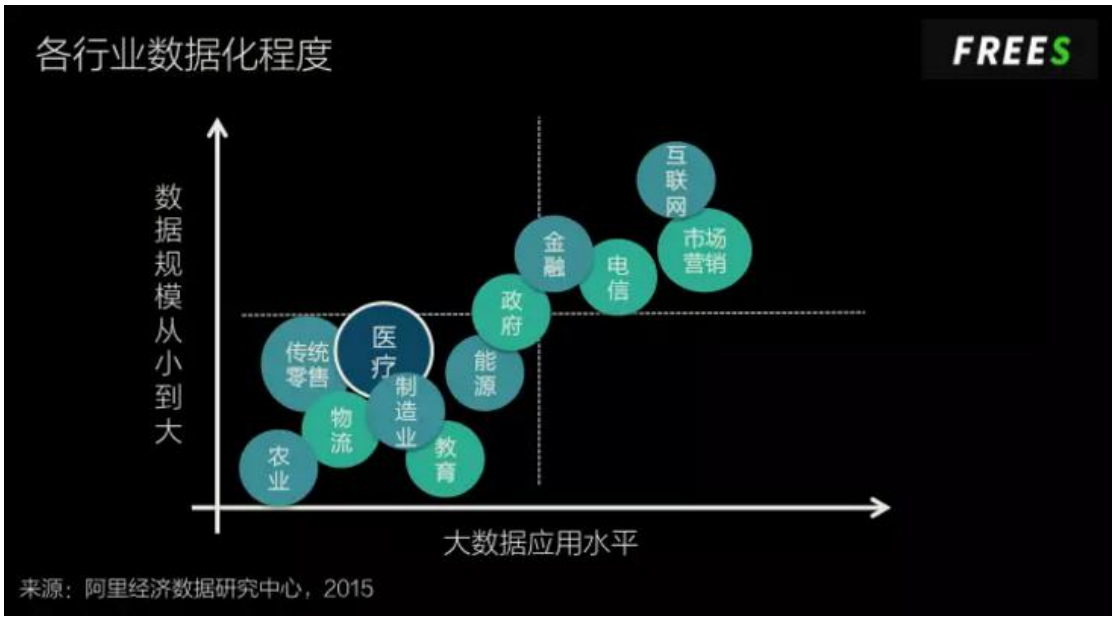
这两种商业模式不难理解，咨询有很强的可定制性，能够准确有效地解决公司的商业需求，但是需要大量和长期的人力支持，花费高，不容易规模化。软件服务则具有边际成本低、人力支持少、容易规模化的特点，但是它缺乏可定制性。很多时候企业并不能直接解决问题，所以面临难以销售的问题。

/ 05 /

企业数据化的演化历程：传统信息化，在线化，云化，数据化



企业数据化的演化历程：传统信息化，在线化，云化，数据化。



各个行业的数据化发展程度，因其行业特点而不同。相较于传统零售、农业和制造业，医疗行业在数据积累上有领先优势，但是在数据的应用水平上，医疗行业远远落后于互联网、金融和电信等信息化程度更好的行业。

峰瑞观点（freesvc）

通过分析各个行业数据化的程度看到：

- ✓ 互联网化程度越高的企业数据化水平越高
- ✓ 数据变现越容易的企业数据化程度越高
- ✓ 个性化需求越高的企业数据化程度越明显
- ✓ 数据储备量越大的企业数据化趋势越快

行业的数据化

- ✓ 受到商业变现能力和模式的驱动
- ✓ 依赖于底层基础设施的发展
- ✓ 依赖于行业数据的积累

/ 06 /

医疗数据产业链

接下来我们从数据产生、数据处理、数据消费的角度来分析医疗数据产业链。

目前，医疗数据的产生最大的来源是医院、诊所等专业医疗机构以及保险机构。这些数据包含了病理、临床、诊疗和理赔数据。随着移动医疗和智能硬件行业的发展，越来越多的数据开始来自手机 App 记录以及可穿戴设备，这些数据主要包含了人体的生命体征和行为数据，等等。这些数据有助于提升数据的完整性、连续性和准确性，并开始

得到重视。峰瑞资本投资的 Haalthy 已经在收集肺癌用户院外数据方面取得进展。

医疗数据的处理不仅包含清洗、整理和分析等标准环节，它还有其特殊性。例如，临床数据往往来自于电子病历等以自然语言描述的文本文件，且不同医疗机构或者医生对临床症状的描述往往存在一些细微差别，这对数据结构化提出了较高的需求。

医疗数据的消费端比较明确，在 C 端主要是病人和医生，B 端包括了医疗机构、药企和保险公司等。从目前的情况来看，通过 C 端来收费和变现比较困难，主要的商业模式还是围绕着 B 端开发。

/ 07 /

美国 Top 医疗大数据公司产品分析

近几年，医疗数据产业在美国发展迅速。这归功于电子病历在过去 10 年的逐步普及，以及包括医院、药厂和保险等机构对数据分析价值的高度认可。 除了传统的数据巨头 IMS Health ,一些新型数据公司和数据分析公司纷纷涌现。我们挑出 4 家有代表性的公司（Flatiron、IBM Watson Oncology、IMS Health Oncology、Palantir）来分析。

它们分别代表了当前医疗数据领域发展的大方向：基于肿瘤临床数据的事实；肿瘤人工智能辅助决策；肿瘤全景数据；医疗公众资源数据。

我们把重点放到肿瘤数据上。这个领域的诊疗过程复杂、不确定性高、治愈率低，市场价值巨大，因而，数据在这个领域的作用和

价值也得以突显和被重视。 其它疾病领域数据的方法论其实非常相似。

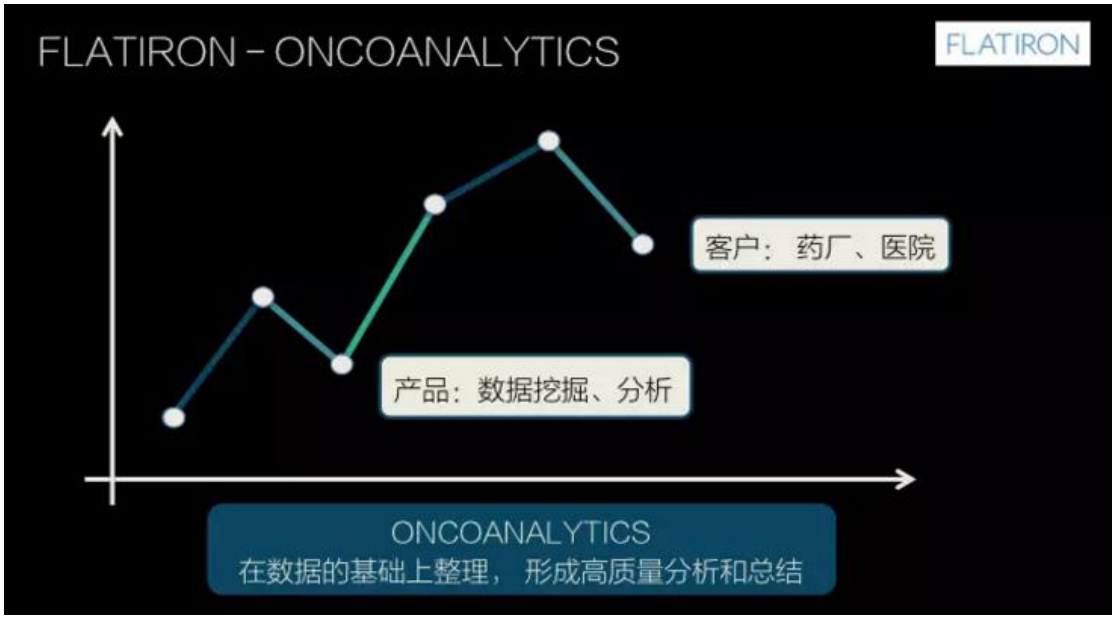
以 Flatiron 为例

创立于 2012 年的 Flatiron 是一家基于肿瘤病患的医疗数据分析公司。它接连获得顶级投资机构和药厂的融资，抗癌药巨头 Roche/Genetech 的参与充分说明机构方认可癌症临床数据对药品研发和市场指导的作用。 Flatiron 平台由行业领先的肿瘤学家、医生和工程师共同打造，在这个平台上医生可以记录、整理、追踪和分析自己病人的情况。

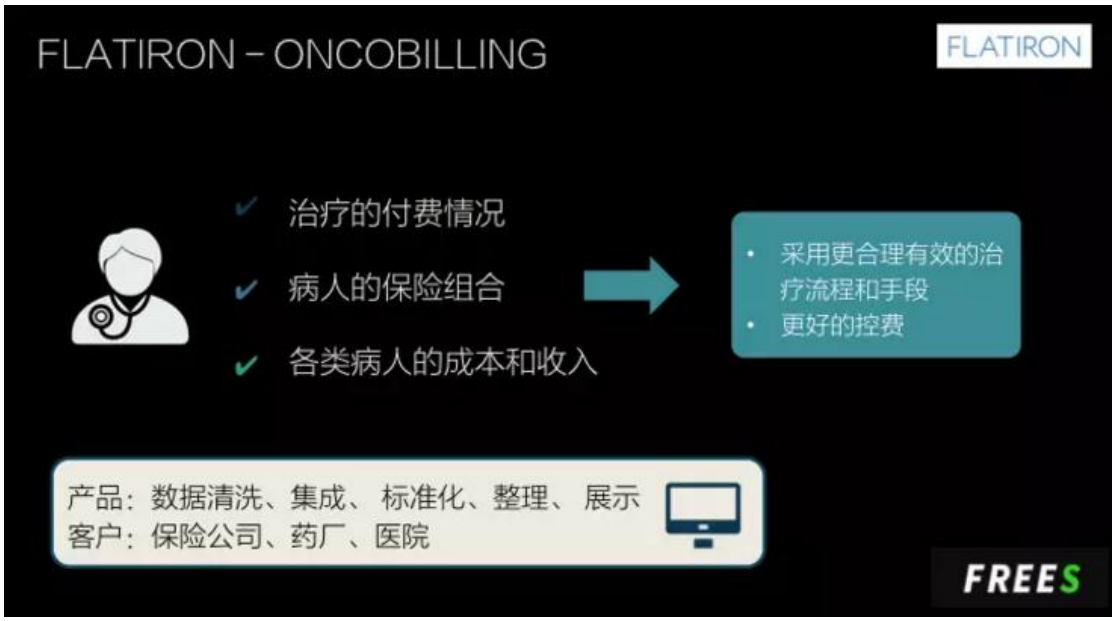
基于平台上收集到的信息， Flatiron 打造了几款主要产品。

FLATIRON- ONCOEMR 是一个癌症病人电子病历，它的主要使用方是医院和医生，药厂也会购买它后台的数据，然后自己做数据分析 ,或者通过第三方协议的形式由 IMS Health 帮助与其他数据进行整合。其它医疗数据分析和人工智能公司也是

FLATIRON- ONCOEMR 后台数据的使用者。



FLATIRON-ONCOANALYTICS 主要基于数据做整理，并形成高质量的分析和总结。比如，某种类型的病人的增长、正在治疗的病人的增长、存活率的跟进，这类产品能对医院与医生管理诊疗工作和病人提供商业和运营上的见解，受到医疗机构的欢迎。

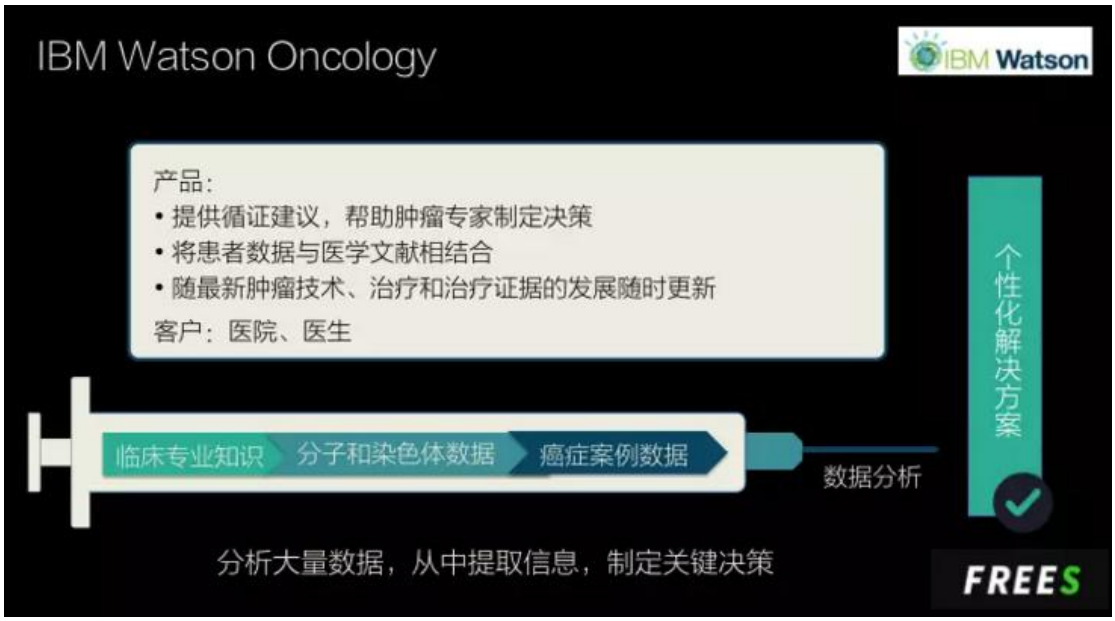


FLATIRON-ONCOBILLING 在医保、商保发达的美国用途广泛。在医院和医生端，FLATIRON-ONCOBILLING 清晰地了解治疗的付费情况、病人的保险组合，对各项治疗、各类病人的成本和收入，采用更合理有效的治疗流程和手段，以更好的控费；保险公司对这类产品的关注度更是毋庸置疑，大

量数据能为控费和更好的理赔设计提供支持。

和 Flatiron 一样，也有一些平台基于电子病历的数据积累，建立起过往没有的诊疗过程的数据挖掘。尽管它们是基于样本医院的病历， 但是已经足够大到提供统计学上有意义的 “怎样做” 和 “为什么” 的见解。

IBM Watson Oncology

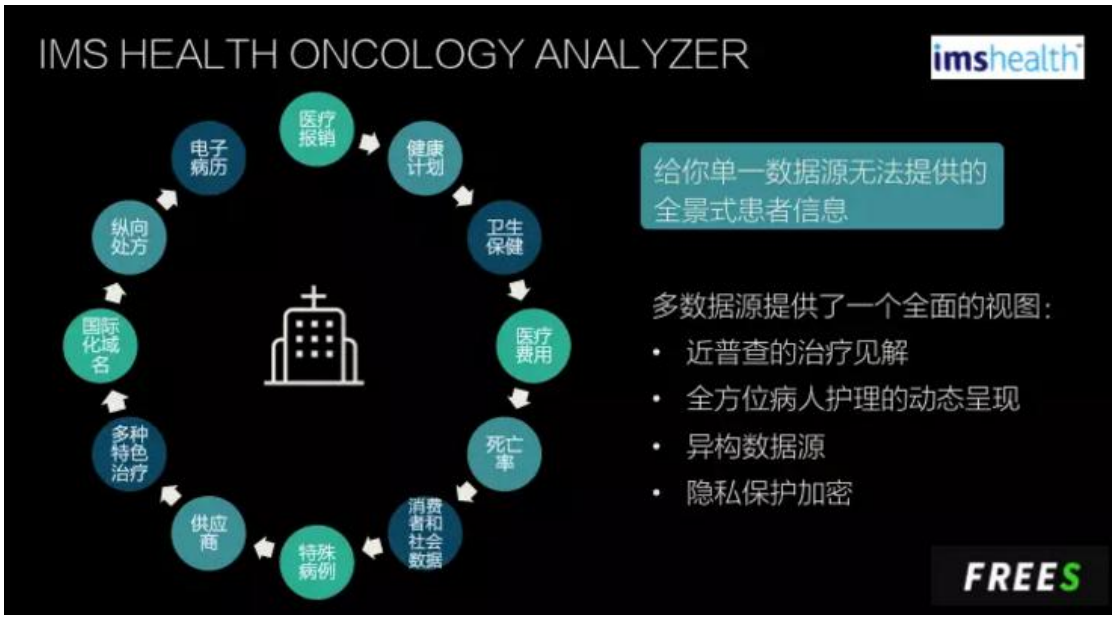


最大的私家癌症中心 MSKCC 与 IBM 合作，将临床专业知识、分子和染色体数据、以及大量癌症案例数据整合到一项循证解决方案中， 分析大量数据并从中提取重要信息，以制定出关键决策。

肿瘤学专家培训 Watson ,将患者的医学信息与大量的治疗方针、已发表的研究结果和

其他洞察力信息相对比，为医师提供个性化的、基于置信度的建议。Watson 的自然语言处理能力允许系统利用非结构化数据，例如杂志文章、医师的笔记、以及来自 National Comprehensive Cancer Network (NCCN) 的指导方针和最佳实践信息。

IMS Health Oncology Analyzer



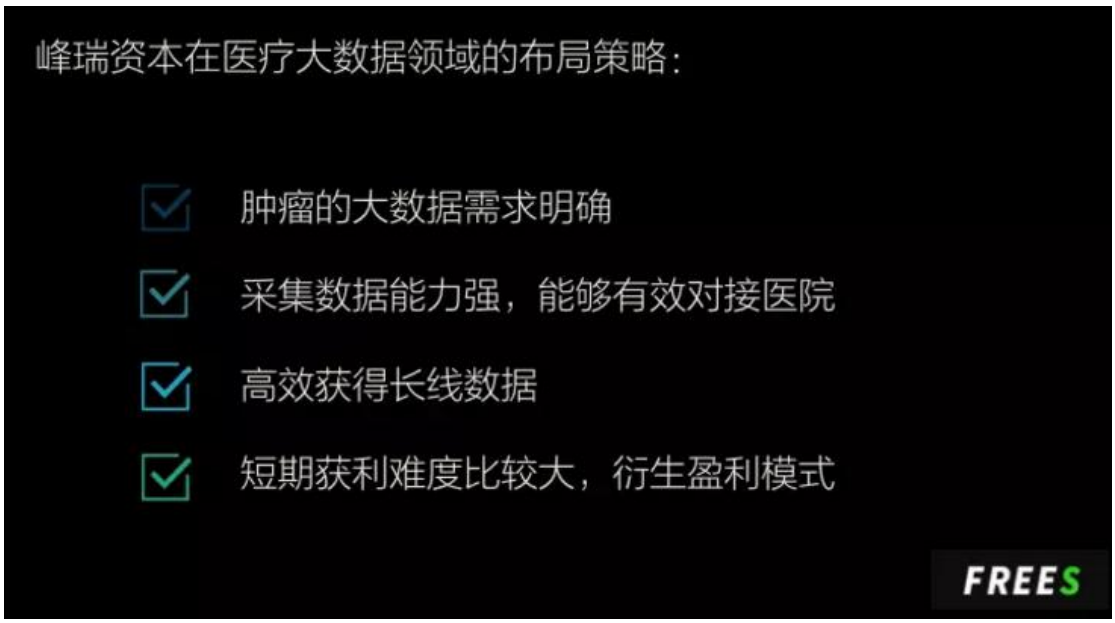
凭借庞大的用药和医生数据基础，结合丰富的医药咨询经验，医疗数据界的巨头 IMS Health 多年来一直在打造医药医疗全景数据图。没有任何一个数据源头能提供足够全面的信息，IMS 除了拥有巨大的数据量，在数据拼接和整合上也有丰富的经验，随着电子病历数据的引入和增长，IMS 致力于把药厂销量、销售到医疗机构的量、医疗机构用

药治疗情况以及病人保险付费情况全部串联到一起。

并购了 Quintile 以后，IMS 还能整合临床实验的数据。其咨询业务基于 IMS 自身汇拢的数据产生的见解，能够对数据业务带来良好正反馈。合并后近 200 亿美金的估值体现了市场对医疗数据价值的认可。

IMS 在世界范围内不断复制其美国模式，逐步形成自己的垄断地位。

Palantir 的模式在中国比较难于复制， 先不赘述。



峰瑞观点（freesvc）

了解了以上几家美国著名医疗数据公司后，我们回顾下之前的报告（我们曾经对比过中美医疗数据市场阶段的差距），并结合中国现有医疗数据项目的重点， 我们总结出中国医疗数据创业项目的 4 大方向：

1. 基于肿瘤临床数据的事实。大量创业项目从这个方向切入；

2. 肿瘤人工智能辅助决策。现在相对较难，因为是建立在 1 的基础上；

3. 肿瘤全景数据。和 1 类似，创业项目能获取到的其他数据比较少；

4. 医疗公众资源数据。中国的数据基础弱，这个方向可能需要国家和上层推动。