



医学信息学 第二章

健康医疗云与大数据

医学信息学是以信息学、信息管理和信息技术为依托，研究医学领域中的信息现象和信息规律，用于医学决策和管理的一门交叉学科。学科以生命信息现象和人类社会活动中的医学信息为研究对象。

谭积斌

考勤

课堂签到







大数据



条件

万物互联



技术

云计算



目的

决策

大数据

Big Data

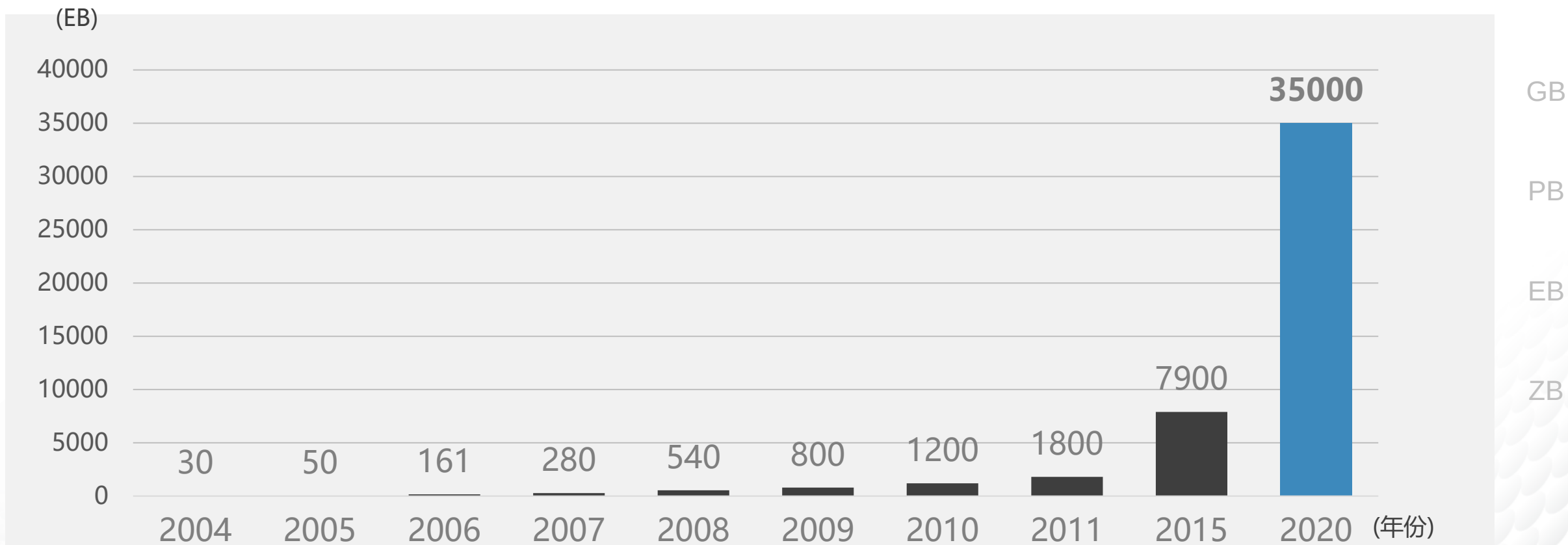
	传统数据	大数据
容量	GB	不断更新 (TB/PB)
产生速度	每小时, 每日...	非常迅速
形式	结构化	非结构化为主
数据源	集中式	完全分布式
数据整合	容易	困难
数据存储	关系数据库管理系统	分布式数据库系统

大数据

Big Data



杰姆·格雷 (Jim Gray) 提出著名的“新摩尔定律”，即人类有史以来的数据总量，每过18个月就会翻一番。



全球数据总量图

大数据

Big Data

	传统数据	大数据
--	------	-----

容量

GB

不断更新 (TB/PB)

互联网每天产生的全部内容可以刻满6.4亿张DVD

全球每秒发送290万封电子邮件，一分钟读一篇的话，足够一个人昼夜不停地读5.5年

Google每天需要处理24PB的数据

每天会有2.88万个小时的视频上传到YouTube，足够一个人昼夜不停地观看3.3年

网民每天在Facebook上要花费234亿分钟，被移动互联网使用者发送和接收的数据高达44PB

Twitter上每天发布5000万条消息，假设10秒就浏览一条消息，足够一个人昼夜不停地浏览16年

大数据到底有多大？

以上一组互联网数据

大数据

Big Data

传统数据

大数据

容量

GB

不断更新 (TB/PB)

存储的增长

医疗服务产生的数据总量(PB)



一个CT图像含有大约150MB的数据，而一个基因组序列文件大小约750MB，一个标准的病理图则大得多，接近5GB。

如果将这些数据量乘以人口数量和平均寿命，仅一个社区医院或一个中等规模制药企业就可以生成和累积达数个TB甚至数个PB级的结构化或非结构化数据。

到2020年，医疗数据将会急剧增长到35 ZB,相当于2009年数据量的44倍增长。

互联网

用户的在线活动很多实时刷新数据



物联网

物品需要实时在线反馈自身的数据



传统数据

大数据

产生速度

每小时，每日...

非常迅速



结构化数据

Structured data

数据清单，数据库存储的数据集，预设数据模型，规定数据格式和长度规范，主要用二维表的逻辑表示

字段1 字段2 字段3 字段4 字段5 字段6 字段7 字段8

	ID	Name	Age	Sex	Phone	Addr	Job	Price	Time
记录1▶	452126..	赵小涵	45	1	18678...	广西	教师	¥ 684	2020年
记录2▶	452126..	余潇潇	26	2	13548...	湖南	工人	¥ 6584	2020年
记录3▶	452226..	胡大大	48	2	18845...	湖北	农民	¥ 254.2	2019年
记录4▶	452436..	赵四	65	1	18756...	福建	律师	¥ 1024	2018年
记录5▶	452146..	杨小黑	35	2	13557...	广东	自由	¥ 523.5	2020年

*结构化数据较容易被检索和分析

非结构化数据



Unstructured data

非结构化数据则与结构化数据相反，是不规则的，没有语定义的数据模型。包括所有的办公文档、文本、图片、html、音视频等

*非结构化数据隐含的信息量非常丰富（显性/隐性）



传统数据

大数据

形式

结构化

非结构化为主



结构化数据

Structured data



你对医生满意吗？

总体：



ID	Name	Age	Sex	1	2	3	4	5
452126..	赵小涵	45	1	0	0	0	2	2
452126..	余潇潇	26	2	0	1	1	0	1
452226..	胡大大	48	2	1	1	0	1	5
452436..	赵四	65	1	0	1	0	5	0
452146..	杨小黑	35	2	0	2	2	4	0

非结构化数据

Unstructured data



你对医生满意吗？请你对医生进行简单的评价！



周医生态度很好，也很有耐心，检查过程较慢，等待时间太久，医生开了很多药。

大数据

Big Data

	传统数据	大数据
容量	GB	不断更新 (TB/PB)
产生速度	每小时, 每日...	非常迅速
形式	结构化	非结构化为主
数据源	集中式	完全分布式
数据整合	容易	困难
数据存储	关系数据库管理系统	分布式数据库系统

大数据

Big Data



条件

万物互联

技术

云计算

目的

科学决策

1

容量大

2

生长快

3

非结构

4

分布式

大数据是一种非常规数据集合，在万物互联时代下利用新模式和新技术来讲数据转化为辅助决策的信息资产

”

指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据4V特征

Big data 4V features



容量巨大



类型繁多



速度极快



价值密度低

大数据4V特征

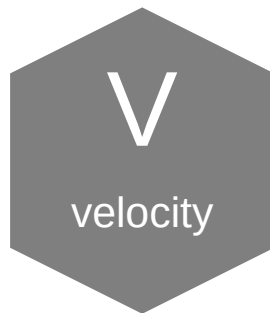
Big data 4V features



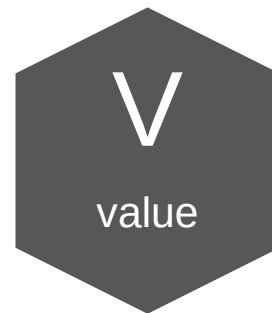
容量巨大



类型繁多



速度极快



价值密度低



变化性



真实性



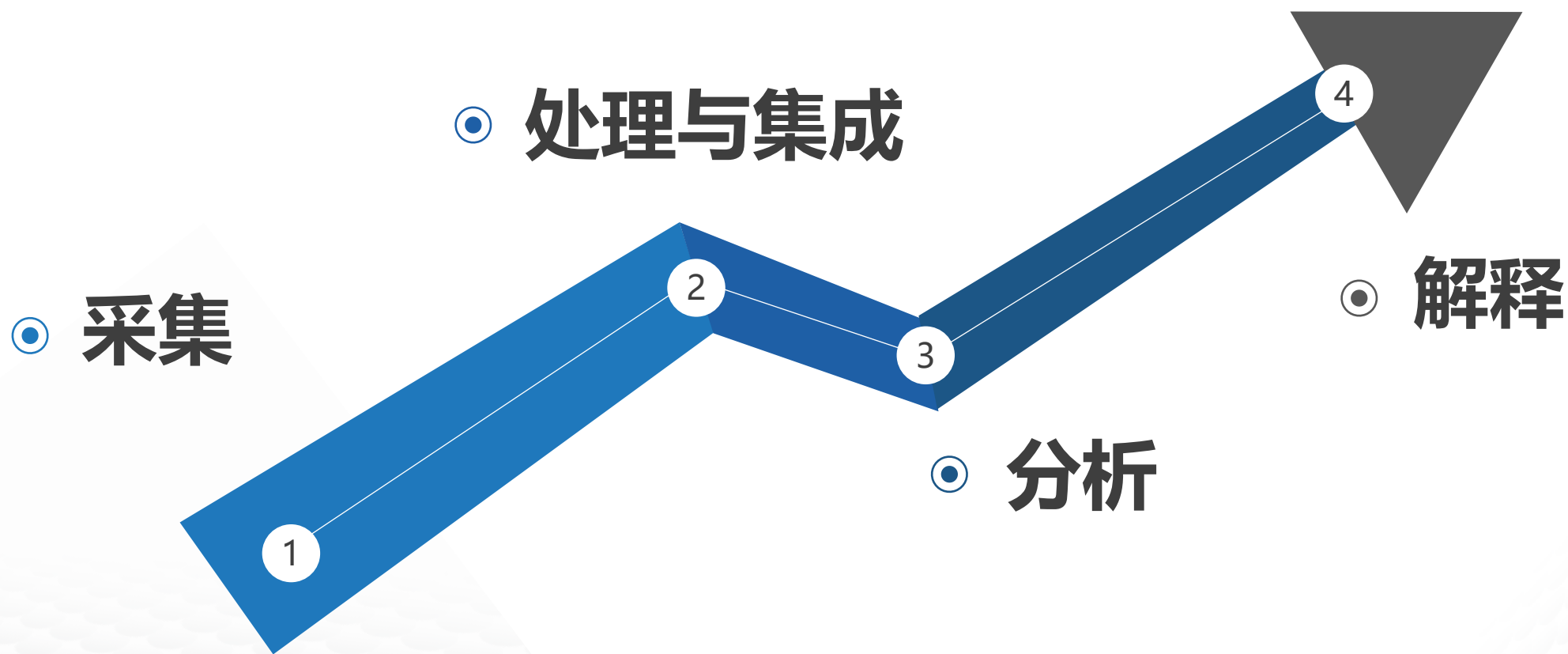
波动性

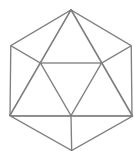


可视化

大数据处理基本流程

The basic process of big data processing





大数据 采集

从传感器和其它待测设备等模拟和数字被测单元中自动采集信息的过程。



采集什么数据？



怎么采集数据？

大数据

Big Data



智能终端拍照、拍视频



发微博、发微信

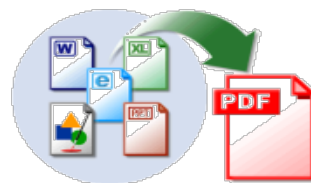
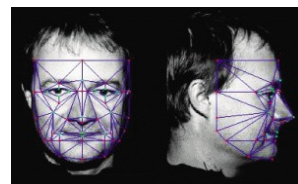


其他互联网数据

来自“大人群”泛互联网数据



来自大量传感器的机器数据



科学研究及行业多结构专业数据

随着人类活动的进一步扩展，数据规模会急剧膨胀，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的各行业累积的数据量越来越大，数据类型也越来越多、越来越复杂，已经超越了传统数据管理系统、处理模式的能力范围，于是“大数据”这样一个概念才会应运而生。



采集什么数据？

大数据采集的数据往往包括传感器数据和互联网数据，以互联网数据为主。互联网数据包括线上行为数据和内容数据。

行为事件

- 页面数据
- 交互数据
- 表单数据
- 会话数据

内容数据

- 应用日志
- 电子文档
- 机器数据
- 语音数据
- 社交媒体数据

网络数据

- 网络流量
- 带宽数据
-



怎么采集数据？

企业数据采集主要分为外部以及内部。内部（系统日志、行为事件）、外部（互联网数据）

系统内部采集数据

【数据来源】系统日志、用户浏览系统的行为事件、数据库

【常用方法】埋点

【常用工具】代码埋点

系统外部采集数据

【数据来源】社交媒体数据、其它平台/系统、用户群体

【常用方法】爬虫

【常用工具】第三方爬虫工具



用什么工具采集?

埋点

代码埋点。在某个控件操作发生时通过预先写好的代码来发数据的代码埋点;

【应用场景】个性化推荐（访问页、点击数、检索词、语音...）

国内的主要第三方数据分析服务商，如百度统计、友盟、TalkingData 等都提供了这一方案。Sensors Analytics 也一样提供了 iOS、Android、Web 等主流平台的代码埋点方案。

```
1 public void onResume() {  
2     super.onResume();  
3  
4     MobclickAgent.onPageStart("SplashScreen"); // 统计页面(仅有Activity的应用中SDK自动调用，不需要单独写。"SplashScreen"为页  
5     MobclickAgent.onResume(this); // 统计时长 5 | }  
6  
7 public void onPause() {  
8     super.onPause();  
9  
10    MobclickAgent.onPageEnd("SplashScreen"); // (仅有Activity的应用中SDK自动调用，不需要单独写) 保证 onPageEnd 在 onPause 之  
11    MobclickAgent.onPause(this); // 统计时长 5 | }  
12 }
```

10, 因为 MobclickAgent.onPause(this); 为页面名称，可自定义

复制



用什么工具采集?

网络数据采集方法



网络流量的采集

带宽管理技术

DPI / **DFI**

对已经采集到的数据进行适当的处理、清洗去噪及进一步集成存储，保证数据质量及可靠性。

常用的方法是在数据处理过程中涉及一些数据过滤器，通过聚类或关联规则方法将无用或错误的离群数据挑出来过滤掉，防止其对最终数据产生不利影响。

处理
集成





数据分析

传统数据处理分析方法有：数据挖掘、机器学习、智能算法、统计分析等。



数据解释

实质上就是数据的可视化，把复杂的数据关系以图形化的形式展现出来。

“

实操演练

《虚拟现实技术在医学教育中的研究现状和热点》

— 知识图谱 —

大数据处理基本流程

The basic process of big data processing



○ 采集

本文研究的数据来源于 CNKI “中国学术期刊全文数据库”，检索项为主题，时间跨度为1996年至今，检索词为“虚拟现实”并且“医学教育”、“VR”并且“医学教育”。共检索到期刊论文268篇，剔除会议、传记、刊误和社论等无关文献19篇，剩余有效文献249篇作为本研究对象。

- 明确研究内容 —— **虚拟现实在医疗教育领域的研究现状**
- 数据来源哪里？ —— **CNKI期刊全文数据库**
- 如何检索？ —— **检索关键词**
- 检索范围？ —— **时间**

◉ 处理与集成

书目共现分析系统



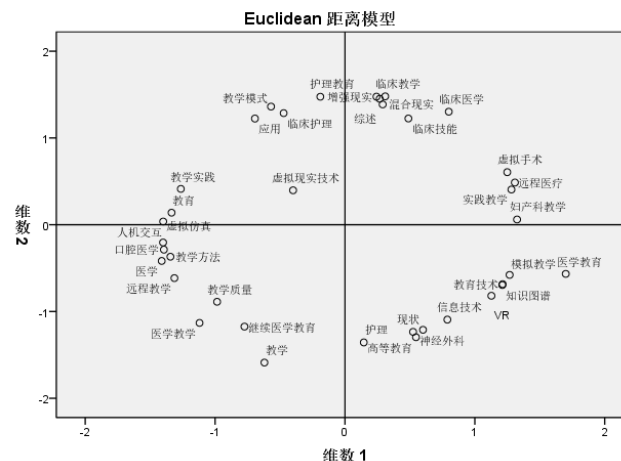
- 批量选择论文 —— 50/页
- 文献导出格式 —— **NoteFirst**
- 导入分析软件 —— **Bicomb**
- 提取关键词 —— 提取年份（年度发文量统计）
- 处理关键词 —— 处理冗余（高频关键词统计）

● 处理与集成

利用Bicomb的矩阵功能统计得到29 x 29的高频关键词共现矩阵，如表2所示。利用Ochia系数对共现矩阵进行包容化处理，实现从共现矩阵向相关矩阵的数值转换将关键词相异矩阵的数据导入到SPSS19.0中，以“Euclidean距离”为量度模型进行多维尺度分析得到本研究主题的高频关键词聚类图谱

高频关键词聚类图谱

1. 统计高频关键词共现矩阵
2. Ochia系数包容化处理得到相关矩阵
3. 相关矩阵转换为相异矩阵
4. Euclidean距离 多维尺度分析



大数据关键技术

Key technologies of big data



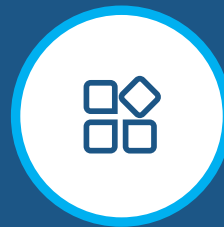
云计算

Cloud Computing



分布式文件系统

Google File System



处理框架

MapReduce



计算平台

Hadoop

你是一家饭店的老板，想要开发一个订餐系统



云计算

Cloud Computing

- 机房场地
- 服务器
- 光纤
- IT工程师，负责搭建环境、运维服务器
- 程序员，设计与开发订餐系统
- 运维人员，负责引流推广

买

租

成本

弹性

安全



云计算

Cloud Computing

你是一家饭店的老板，想要开发一个订餐系统



SaaS, 软件即服务

- 程序员，设计与开发订餐系统
- 运维人员，负责引流推广

PaaS, 平台即服务

- IT工程师，负责搭建环境、运维服务器

IaaS, 基础设施即服务

- 机房场地
- 服务器
- 光纤
-



云计算

Cloud Computing

公有云

- 由公有云厂商来提供云服务



私有云

- 企业自主搭建云服务
- 应用场景：大公司、跨国公司，方便各地员工随时随地安全访问公司业务系统

混合云

- 核心数据和应用放在私有环境
- 把硬件资源和基础软件配置托管给公有云厂商

讨论

课堂讨论



大数据关键技术

Key technologies of big data

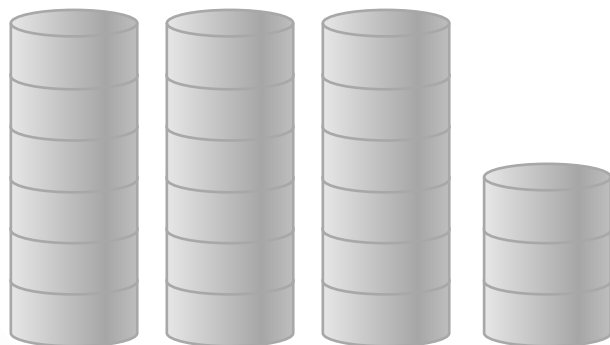


分布式文件系统

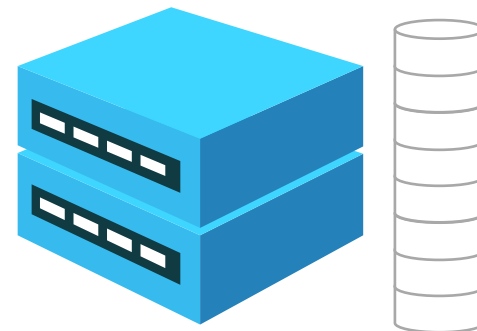
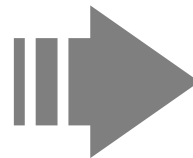
Google File System

大数据容量巨大，单台服务器的容量是有限的，即使外接扩容，也会存在数据安全风险。

2100Gb



800Gb



大数据关键技术

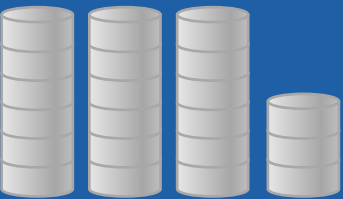
Key technologies of big data

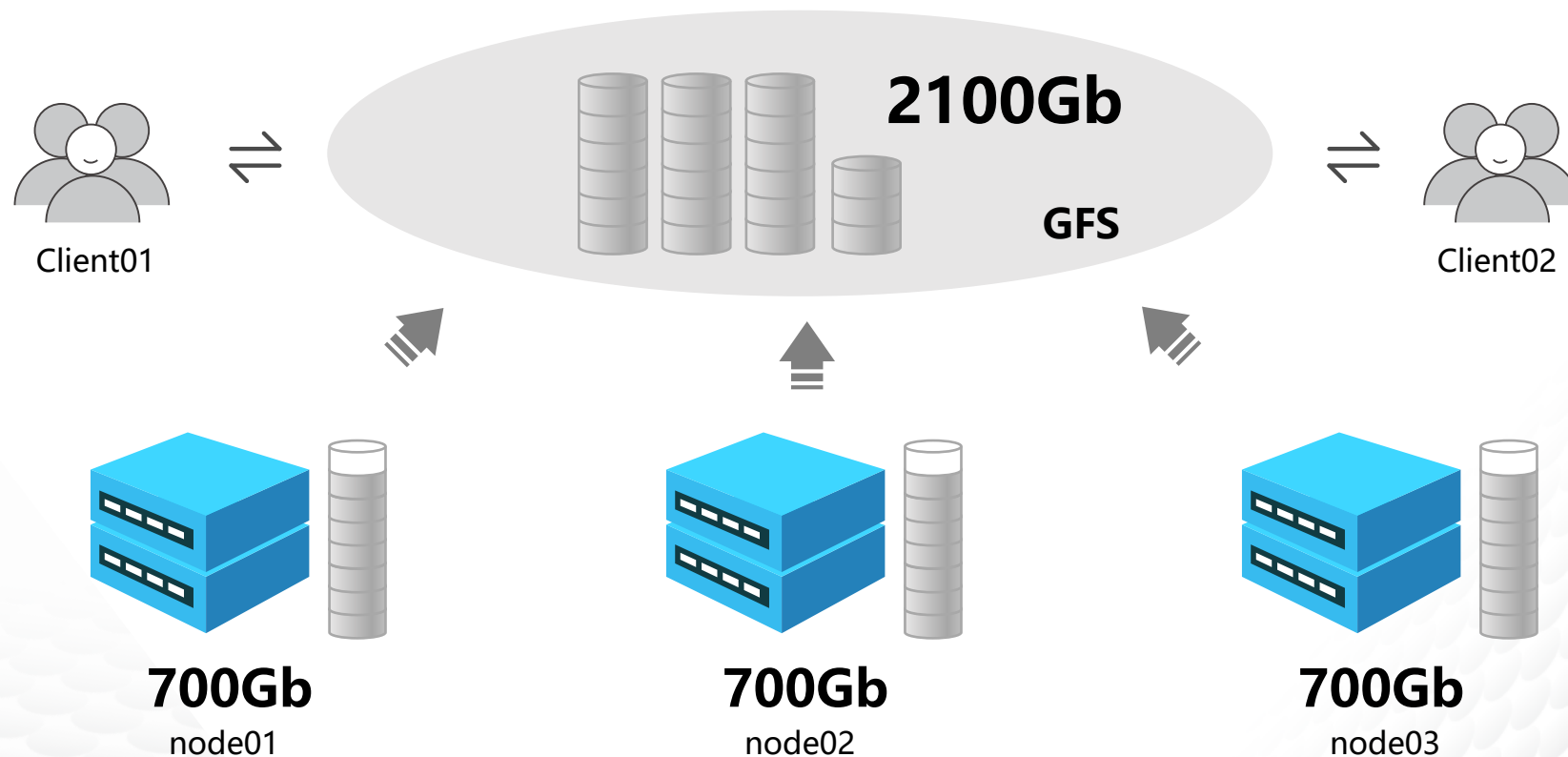
大数据容量巨大，单台服务器的容量是有限的，即使外接扩容，也会存在数据安全风险。



分布式文件系统

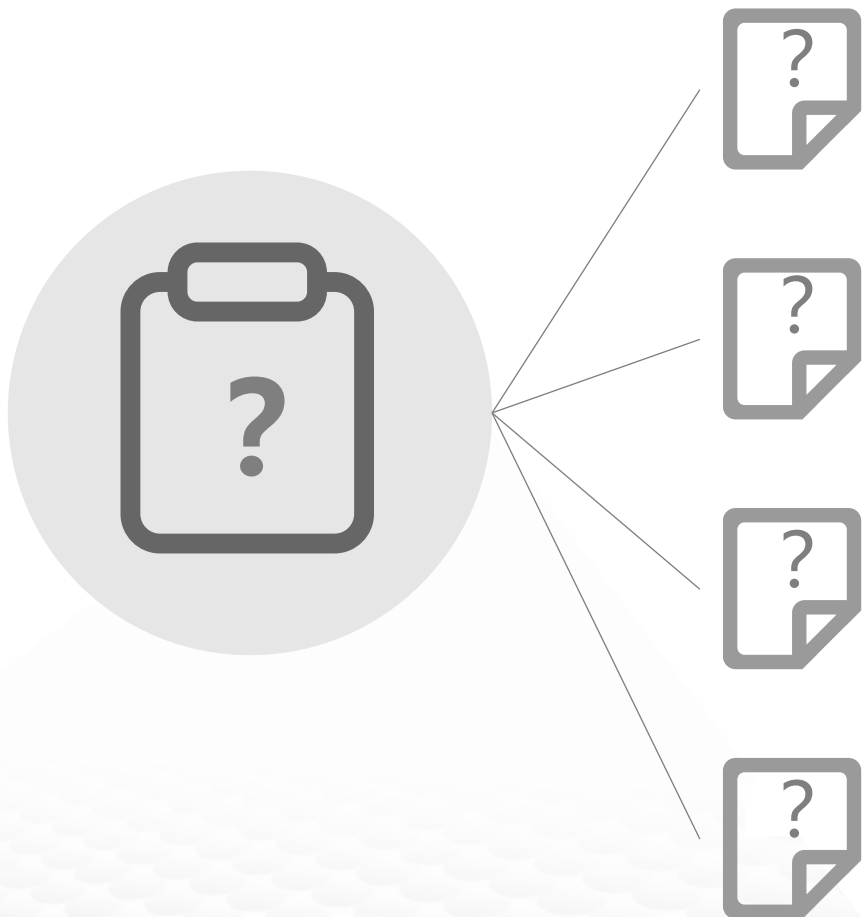
Google File System





Map

把大数据复杂任务划分为若干个简单任务来处理



处理框架

MapReduce

面向大数据并行
处理的计算框架

“分而治之”

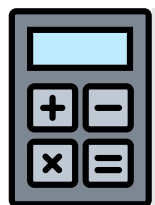
Reduce

对Map阶段的结果进行汇总



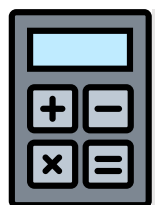
Map

把大数据复杂任务划分为若干个简单任务来处理



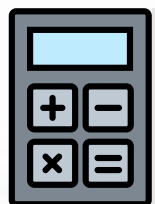
DataNode01

$1 + 2 + 3 + \dots + 300\text{亿}$



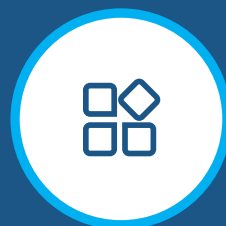
DataNode02

$1 + 300\text{亿} + \dots + 600\text{亿}$



DataNode03

$1 + 600\text{亿} + \dots + 1000\text{亿}$



处理框架

MapReduce

$1 + 2 + 3 + \dots + 1000\text{亿}$

Reduce

对Map阶段的结果进行汇总

数据结果01

+

数据结果02

=

+

数据结果03

总数据结果

大数据关键技术

Key technologies of big data

HDFS



GFS

MapReduce



MapReduce



计算平台

Hadoop

大数据关键技术

Key technologies of big data



医疗大数据

Medical Big Data



医疗大数据的特征包括那几个特征？

容量巨大、类型繁多、速度巨快、价值密度低



医疗大数据是服务于什么呢？

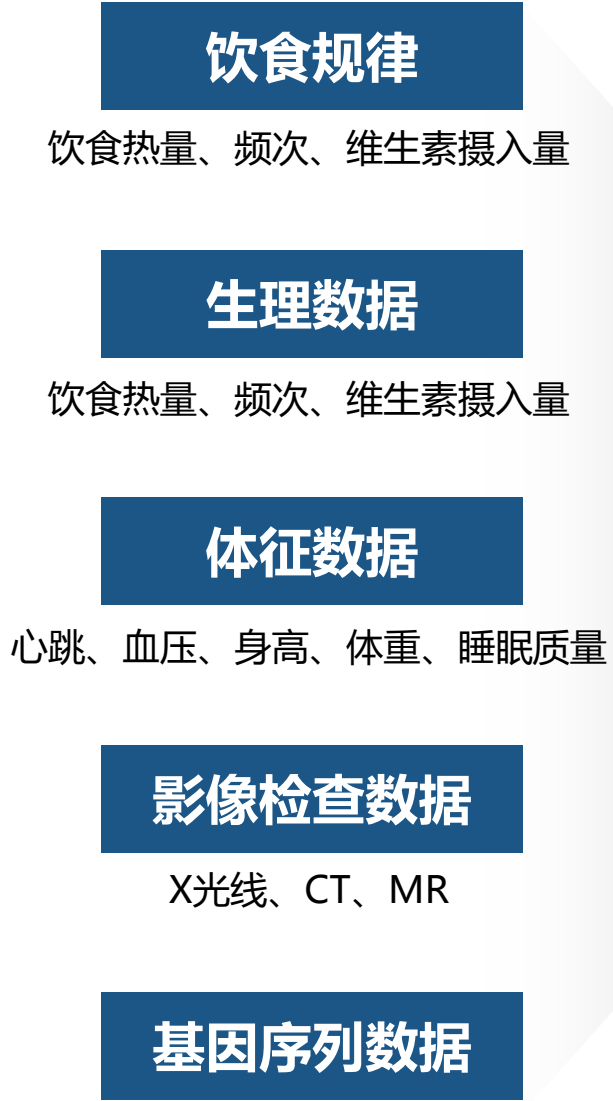
医生诊断、用药、管理决策



医学信息学的知识框架



研究对象	知识层次
社会群体	公共卫生信息学
患者个体	临床信息学
器官	影像信息学
组织	
细胞	生物信息学
亚细胞	
蛋白质	
基因	
分子	



医疗大数据分类

Medical Big Data

1

院内健康医疗大数据

产生于临床诊治、科研和管理过程，包括各种门诊记录、住院记录、影响记录、检验记录、处方记录、手术记录、随访记录和付费记录。有结构化的数据，也有非结构化的数据。非结构化的数据比较难以可视化。比如记录患者验血常规的数据，可以看出他的血常规数据变化情况。



医疗大数据分类

Medical Big Data

2

公共卫生健康大数据

公共卫生健康大数据主要应用在流行性传染病的时候检测，根据历史疫情大数据，建立传染病暴发、流行的预警模型，发现传染病种类及发生区域、高风险人群、病毒传染方式。数据主要来源于社交媒体、连锁药房、超市的销售记录等。



比如国家卫健委通过大数据发现，某个地区群众在这段时间内购买的感冒药较多，且在社交媒体上的言论与感冒相关的较多、家庭用电有小幅提升（居家较多），并且相邻区域的群众的数据也在攀升，就可以初步判断有传染病的发生。

但群众的数据还是比较难以收集的，比如有些人不开定位，就不知道他身处何处，也有些人，特别是老人不用手机，消费用的手机不是本人身份证办理的，就具有数据不准确的风险。

这次疫情，国家卫健委果断、及时建立了全国的“健康码”系统，搜集每一位公民的健康数据，而且这些数据就精确得多，包括体温、去向、时间等等，为什么每次都要你们扫码，就需要登记你们的去向和时间。如果你们去过的地方有传染患者，你的绿码马上就会变橙色甚至红色。就能及时控制疫情。

医疗大数据分类

Medical Big Data

3

移动互联健康大数据

主要是指可穿戴设备的健康数据监测。比如血压、心率、体重、睡眠质量等等数据，可以用智能手表、手机等可穿戴设备采集。



医疗大数据处理流程

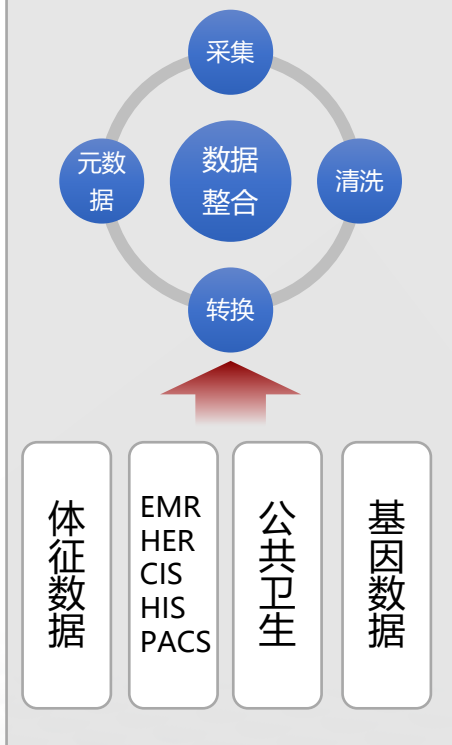
Medical Big Data



医疗大数据处理流程

Medical Big Data

异构健康医疗 大数据整合



多源异构数据的有效整合是大数据分析和挖掘的前提

多模态体征数据

电子健康档案

社交媒体

医疗信息系统

基因数据

格式不同

(结构化和非结构化)

来源多样

(Oracle、Mysql、SQL Server)



先于外部数据
源进行连接



解析数据源的
格式



进行数据采集
和转换



传输到目标系
统存储

数据整合的处理流程

医疗大数据处理流程

Medical Big Data

海量数据 统一存储

统一数据模型



海量数据统一存储是大数据分析的基础

在分布式文件系统中统一存储海量的数据，关键是要依据不同的数据模型进行存储，不能随意堆放，就像我们管理自己的文件一样，要有科学的命名习惯！实现对体征、体检、病历、住院、妇幼、疾控、社保等多美数据的统一存储。



负载均衡

传输的负载均衡



自动扩容

实现系统存储容量的
自动增长



医疗大数据处理流程

Medical Big Data

分布式并行 数据处理

ETL操作

实时排名

数据校验

异常分析

统计分析

数据迁移

高效地分布式数据处理是大数据分析的关键



医疗诊疗



疾病预测



医疗管理



医疗护理

大数据的医疗应用场景

频繁地处理**海量**数据

大数据的医疗数据情况

频繁地处理**复杂**数据

流处理

主要对动态产生数据进行实时计算并及时反馈结果，在数据有效时间内获取其价值

批处理

批处理方式的数据先被存储，然后再分析。MapReduce就是非常重要的批处理模型。

常见的分布式处理模式

医疗大数据处理流程

Medical Big Data

健康医疗大数据分析和挖掘

智能分析引擎

行为分析

健康预测

疾病控制

疾病趋势

药物研发

辅助治疗

业务支撑
决策支撑
科研辅助
管理支持

综合管理

公共卫生

变换共享

其它主题

健康医疗大数据分析的目的是从海量多元异构数据中挖掘、发现知识和获取对所关注问题的新见解和解决方案，以实现数据资产的巨大潜在价值，给数据拥有者及其他利益相关方都带来好处。

针对特定专业需求和应用场景，可以构建复杂的计算模型和学习算法，实现响应的大数据分析服务引擎，以为业务运行、管理决策服务…



应用案例

Application Case

1

助力基因检测服务

最广为人知的是美国好莱坞女星安吉丽娜·朱莉，在2013年经过检测她发现自身携带致癌基因——BRCA1基因，为防止罹患卵巢癌，于2015年切除了卵巢和输卵管。目前，国内外已经有多家基因检测机构，如我国的华大基因、贝瑞和康、美国的 23andMe、Illumina 公司等。



数据来源

检测数据：患者血清、口腔黏膜数据、基因测序等。

其它数据：体检数据、电子病历、遗传记录、患者调查、地理区域以及生活条件等。



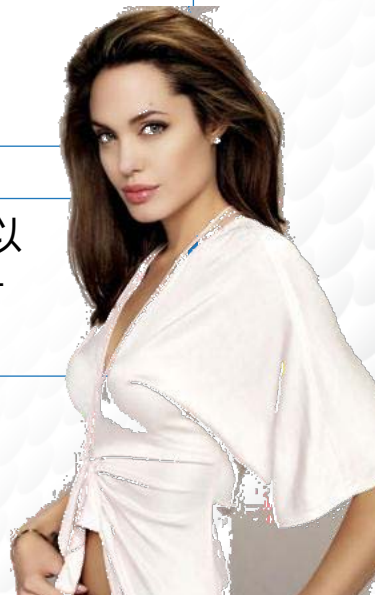
实现路径

采取患者样本做基因测序 → 得到基因序列 → 大数据与原始基因比对 → 大数据分析个体差异性 → 综合分析、诊断 → 制定个体化治疗方案



应用效果

肿瘤医院的病人中有 60%至 80%刚到医院时就已经进入中晚期，癌症早期的筛查可以帮助患者有针对性的改善生活习惯或者采取个体化的辅助治疗，有益于身体健康;同时将癌症扼杀在摇篮里，从而降低日后巨大的医药开支和生活困扰。助力个性化医疗。



应用案例

Application Case

2

支持管理决策判断

起因于国外医院的经验以及广东省人民医院各专业科室差异很大的病床使用率。长期以来，优势专业病源充足，病人候床情况严重，排队入院，相反有些专业空床情况明显，病床使用率仅 65%左右。



数据来源

患者数据：挂号数据、电子病历、患者基本数据等。

医院数据：各科室床位使用情况、诊疗活动、平均住院费用、平均住院周期等。



实现路径

对跨科收治病人之后的科与科之间的工作量、收入、支出、分摊成本等指标进行合理的划分，强化了入院处的集中床位调配权，解决病人入院排队情况。



应用效果

提高病床使用率。病床使用率由 87%提高到 92%，优势专业候床排队现象明显减少。

应用案例

Application Case

3

助力精准诊疗

认为患者的基因型、生活方式、身体特征、多重病患严重影响了治疗效果。提早根据患者的特征设计个性化的治疗方案将有助于降低成本，减少医疗事故。



数据来源

患者数据：基因数据、生活方式信息、身体特征数据、电子健康病历
医院数据：电子病历、检查系统数据



实现路径

根据患者基因信息和患者的其他特征预测各种治疗方案可能的副作用；选择更好的治疗方案，而不是尝试各种治疗方案；



应用效果

Mayo系统用来收集、存储个性化治疗所需要的数据，并为数据分析师提供分析数据的平台。通过分析病人的特征数据并匹配相似病例以帮助医师诊断。

应用案例

Application Case

4

助力预防传染病

2009年，Google比美国疾病控制与预防中心提前1~2周预测到了甲型H1N1流感爆发，此事件震惊了医学界和计算机领域的科学家，Google的研究报告发表在Nature杂志上。Google正是借助大数据技术从用户的相关搜索中预测到流感爆发。



数据来源

上网检索信息、用户上网行为、社交媒体



实现路径

借助大数据预测流感爆发分为主动收集和被动收集，被动收集利用用户周期提交的数据分析流感的现状和趋势，而主动收集则是利用用户在微博的推文、搜索引擎的记录进行分析预测。



应用效果

Google比美国疾病控制与预防中心提前1~2周预测到了甲型H1N1流感爆发

大数据面临的挑战

Challenges facing big data

数据安全与隐私问题

在互联网上浏览网页就会留下浏览痕迹，登录网站也需要输入重要信息（用户名、密码、身份证、手机号、住址、银行卡密码）；随处可见的摄像头和传感器会记录下个人行为 and 位置信息等，通过数据分析，数据专家就可以轻易挖掘人们的行为习惯和重要信息。





数据融合困难

大数据处理技术面临一个重要问题就是如何将个人、企业和政府的各种信息数据加以融合，但这些数据格式都不一致。为解决这个问题，需要用一种统一的数据格式，将人类社会、物理世界和网络空间联系起来，构建统一的信息系统。



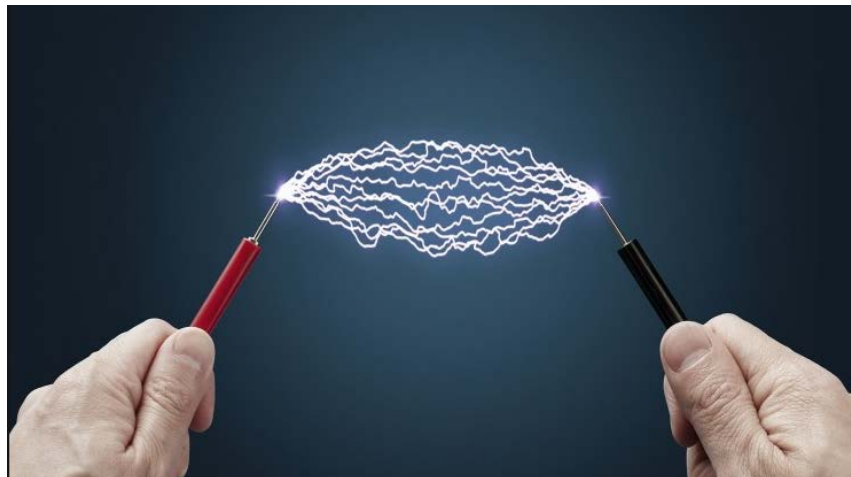
数据清洗困难

数据清洗：价值密度低是大数据的特征之一，信息量大意味着冗余的数据多，垃圾价值的泛滥。因此有必要对数据进行筛选、清理。否则过多的干扰信息一方面会占据大量的存储空间，造成浪费。另一方面这些垃圾数据会对有用信息造成甘涛，影响数据分析结果。



大数据耗能问题

Google数据中心耗电量每年3000kW左右，但只有6%-12%的电能被用来支撑大数据分析处理。绝大部分电能只是用来支撑很多闲置状态的服务器。



Finish

测验

课堂测试





《医学信息学》

本课至此·谢谢

医学信息学是以信息学、信息管理和信息技术为依托，研究医学领域中的信息现象和信息规律，用于医学决策和管理的一门交叉学科。学科以生命信息现象和人类社会活动中的医学信息为研究对象。

谭积斌