01

第1章 大数据技术概述

授课教师：[李阳源]
ywlee1212@gmail.com
主页：https://sites.google.com/view/dalguri



第1章 大数据技术概述


BIG
DATA
ANALYTICS
COMPLEX
SOCIAL
INTERNET
TECHNOLOGIES
STORAGE
SYSTEMS
SOFTWARE
INFORMATION
DATABASES
NETWORKS
RECORDS
SEARCH
PARALLEL
SAN
SIZE
RESEARCH
PETABYTES
INCLUDE
MANAGEMENT
SETS
NEEDED
TIME
LARGE
DISK
SHARED
EXAMPLES
EVERY
ALSO
TOOLS
TARGET
SENSOR
DEFINITION
MOVING
MANAGE
ARCHIVES
PRACTITIONERS
CAPTURE
BUSINESS
DISTRIBUTED
GENOMICS
TERABYTES
ABILITY
PERFORMANCE
STORE
LOGS
COST
MASSIVELY
GROW
DIFFICULTY
HUNDREDS
CAPACITY
BIOLOGICAL
PROCESSING
RELATIONAL
LARGER
CONTINUES
CITATION
TENS
SET
USE
GARTNER
WORKING
ELAPSED
TOLERABLE
CASE
MAY
CURRENT
WITHIN
USING
ONE
PROCESS



1.2 大数据概念

目录

CONTENT

大数据时代

大数据概念

大数据的影响

大数据关键技术

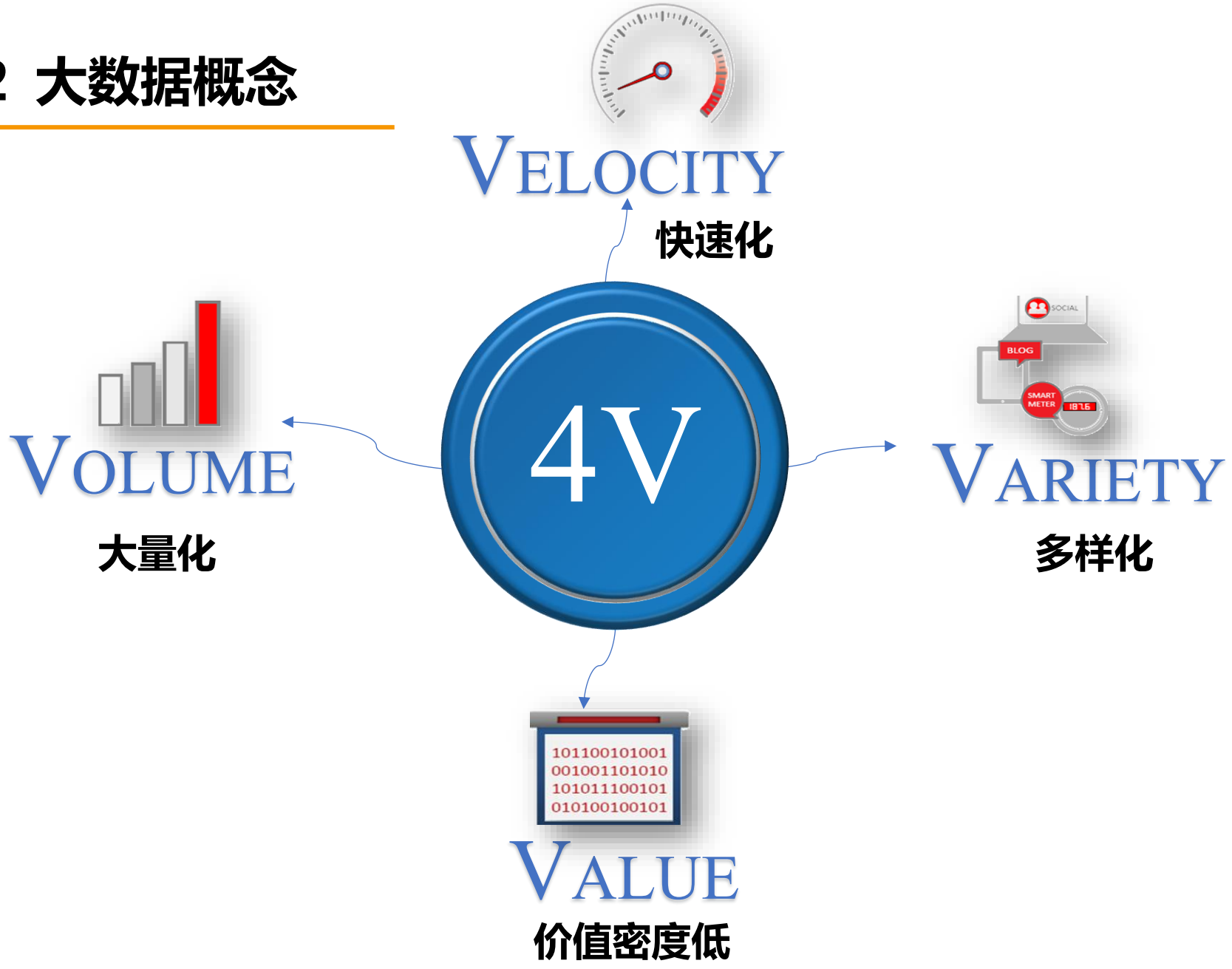
大数据计算模式

代表性大数据技术

什么是大数据?



1.2 大数据概念


VELOCITY
快速化
VOLUME
大量化
4V
VARIETY
多样化
SOCIAL
BLOG
SMART METER
101100101001
001001101010
101011100101
010100100101
VALUE
价值密度低

1.2 大数据概念



1. 数据量大
2. 数据类型繁多
3. 处理速度快
4. 价值密度低

1.2 大数据概念

数据量大



大量化 VOLUME

1.2 大数据概念





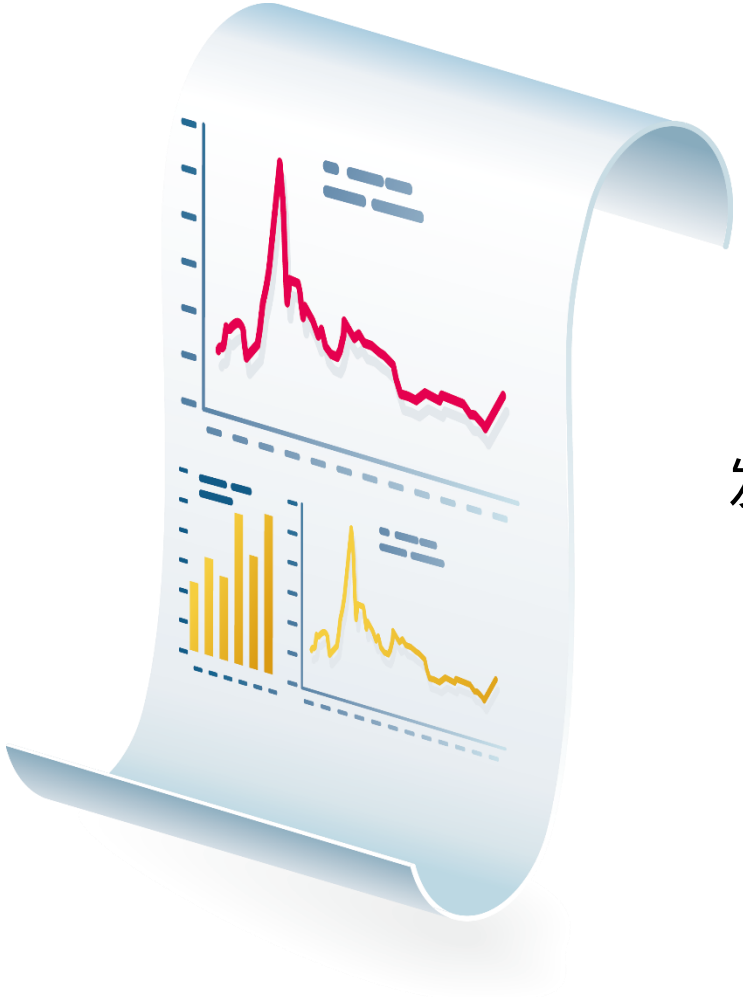
美国的一个非常知名的机构 **IDC**

1.2 大数据概念



每年都会对IT一些行业做一些调查

1.2 大数据概念



发布一些研究报告

1.2 大数据概念



大数据摩尔定律



根据IDC发布报告

人类社会数据每年50%的速度，**每两年就增长一倍**

1.2 大数据概念



数据量大

近两年
产生数据量

=

之前产生
所有数据总量之和

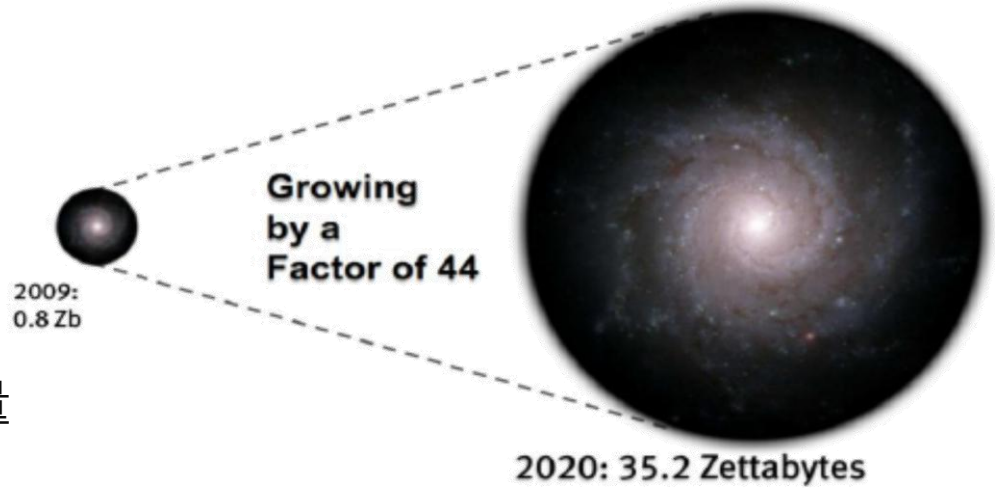
1.2 大数据概念

数据量大

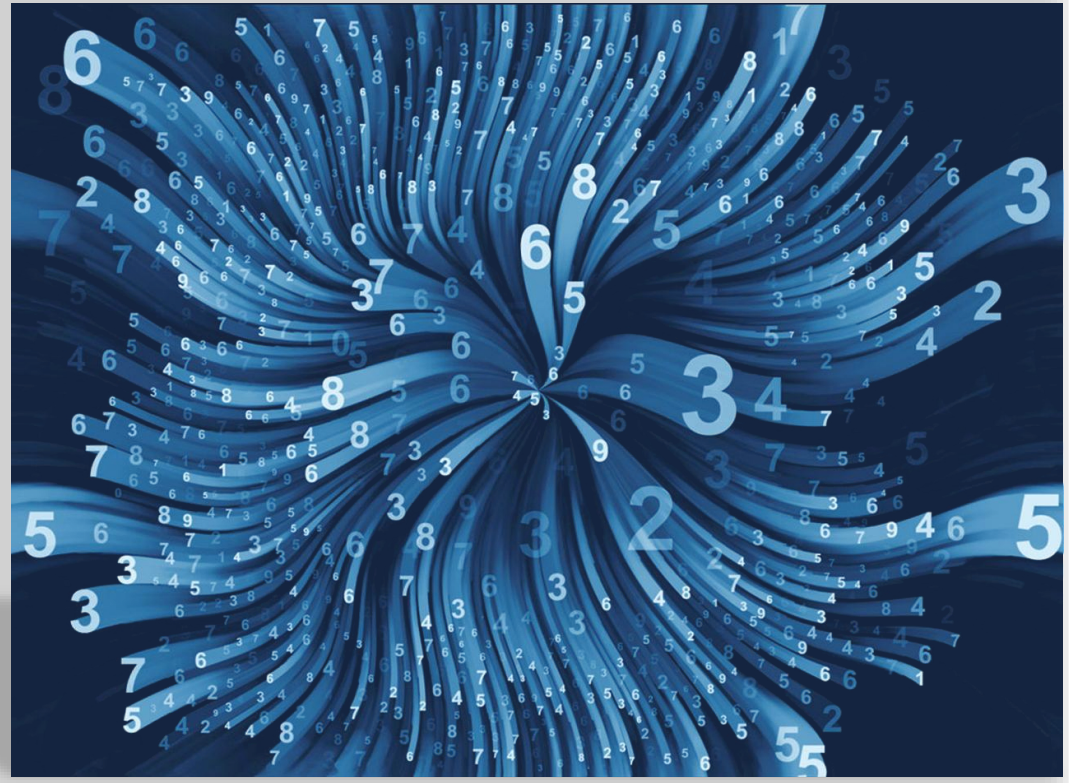
- 人类在最近两年产生的数据量

  相当于之前产生的全部数据量

- 到了2020年，全球将会有35ZB数据量



1.2 大数据概念



35ZB

是一个非常大的数据量

1.2 大数据概念



数据存储单位之间的换算关系

单位	换算关系
Byte（字节）	1 Byte=8 bit
KB（Kilobyte，千字节）	1 KB=1024 Byte
MB（Megabyte，兆字节）	1 MB=1024 KB
GB（Gigabyte，吉字节）	1 GB=1024 MB
TB（Terrabyte，太字节）	1 TB=1024 GB
PB（Petabyte，拍字节）	1 PB=1024 TB
EB（Exabyte，艾字节）	1 EB=1024 PB
ZB（Zettabyte，泽字节）	1 ZB=1024 EB

1.2 大数据概念

数据类型繁多



多样化 VARIETY

1.2 大数据概念



- 它的类型非常繁多

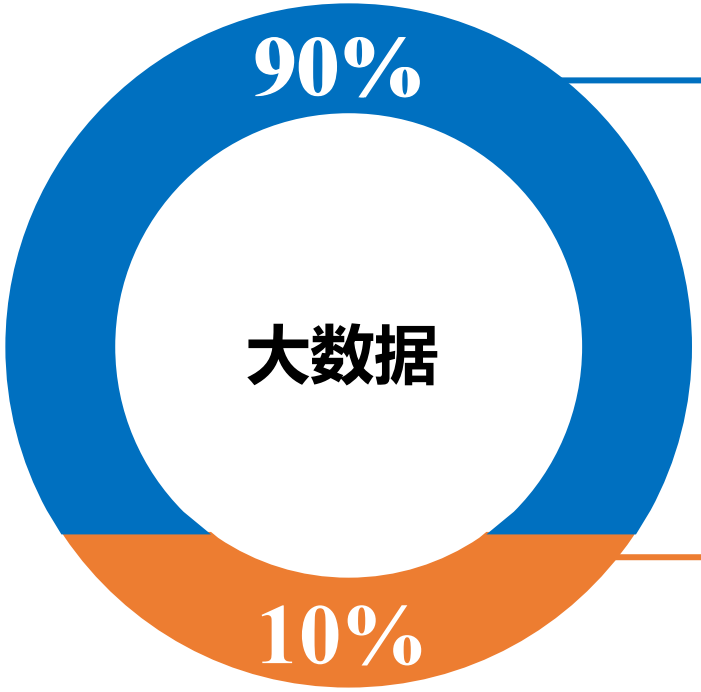
1.2 大数据概念

- 具有规范的行和列结构数据，称为结构化数据



学号	姓 名	年龄	性别	系名	年级
99004	宋明	19	男	计算机	99
99007	王小芬	18	女	经济学	99
99009	张文	19	男	自动化	99
…	…	…	…	…	…

1.2 大数据概念



非结构化数据

存储于非关系型数据库中

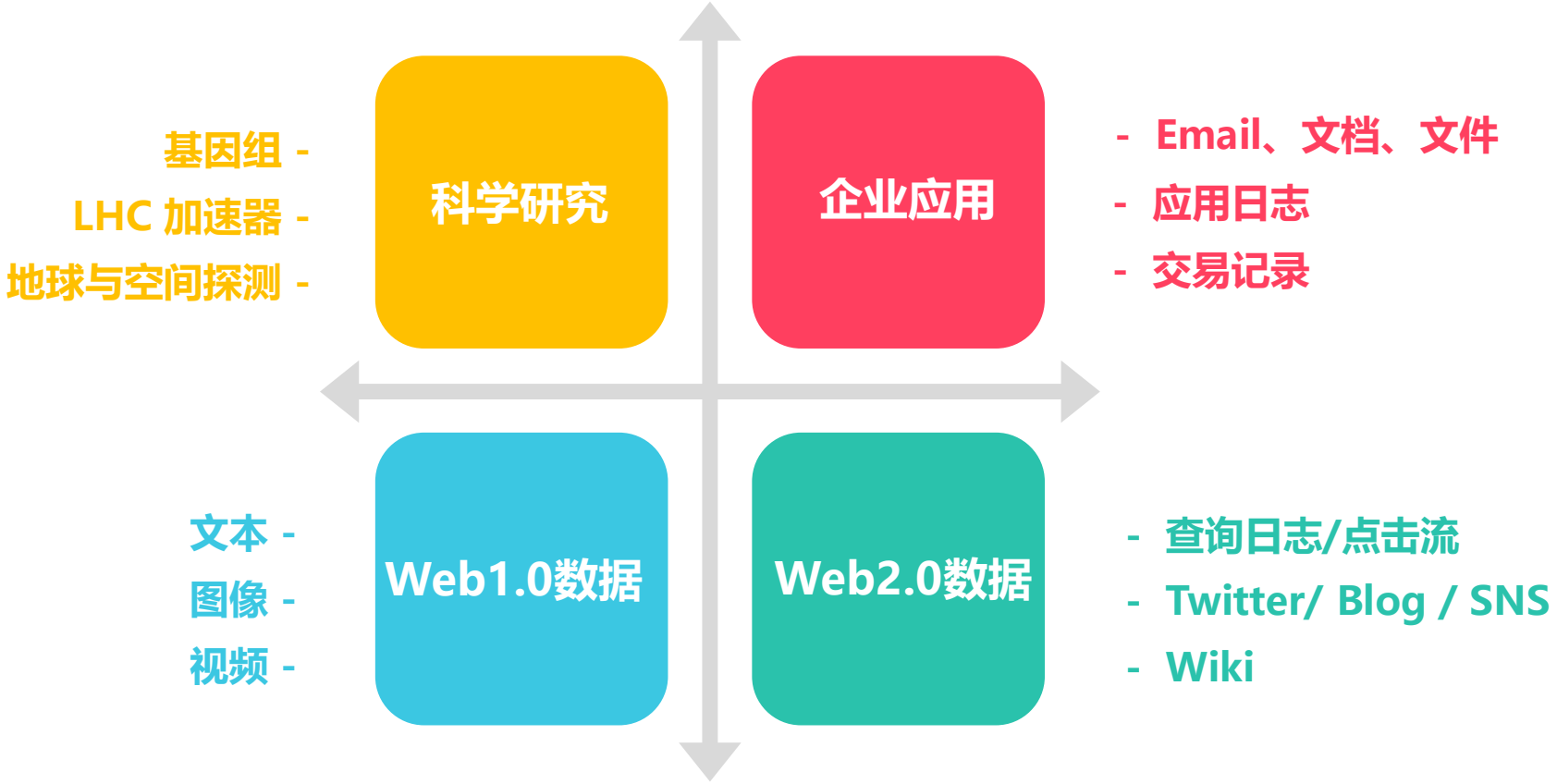
结构化数据

存储于关系型数据库中

1.2 大数据概念



数据类型繁多

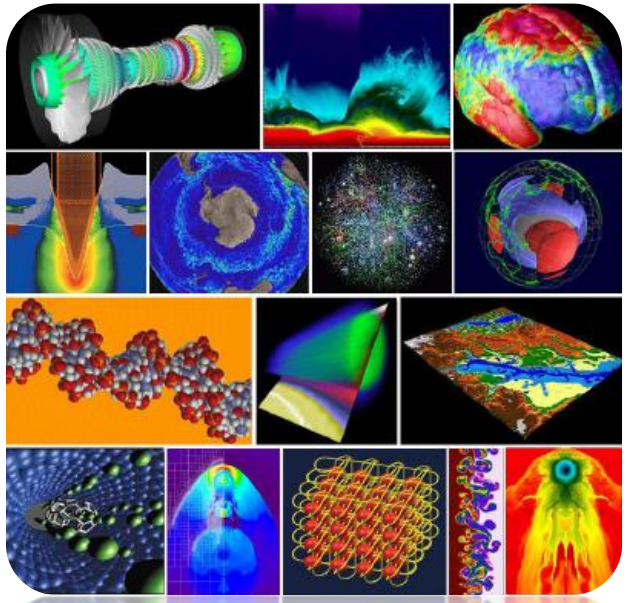


LHC（Large Hadron Collider）:大型强子对撞机。

1.2 大数据概念

数据类型繁多

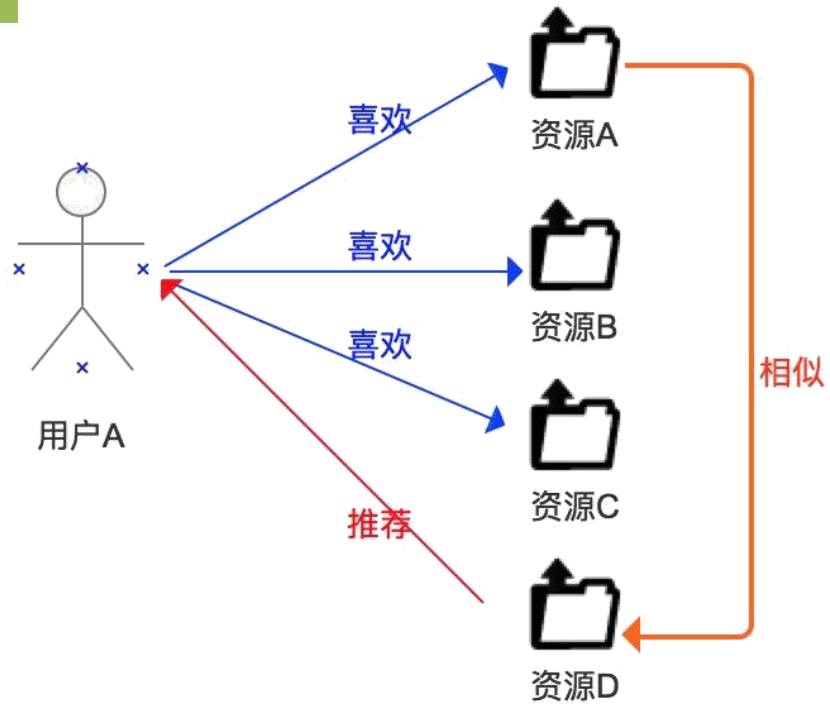




不同类型的数据

1.2 大数据概念

数据类型繁多





面对繁多数据类型，传统算法不再有效

1.2 大数据概念

数据类型繁多

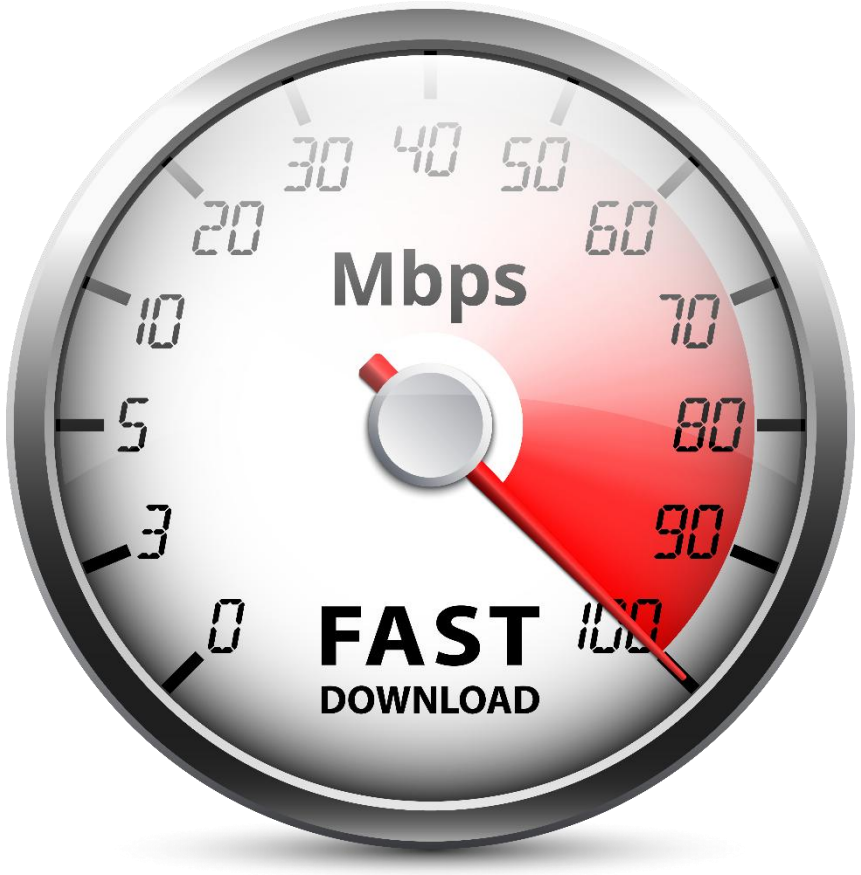


需要设计一些新型大数据算法处理多种类型数据

1.2 大数据概念



处理速度快

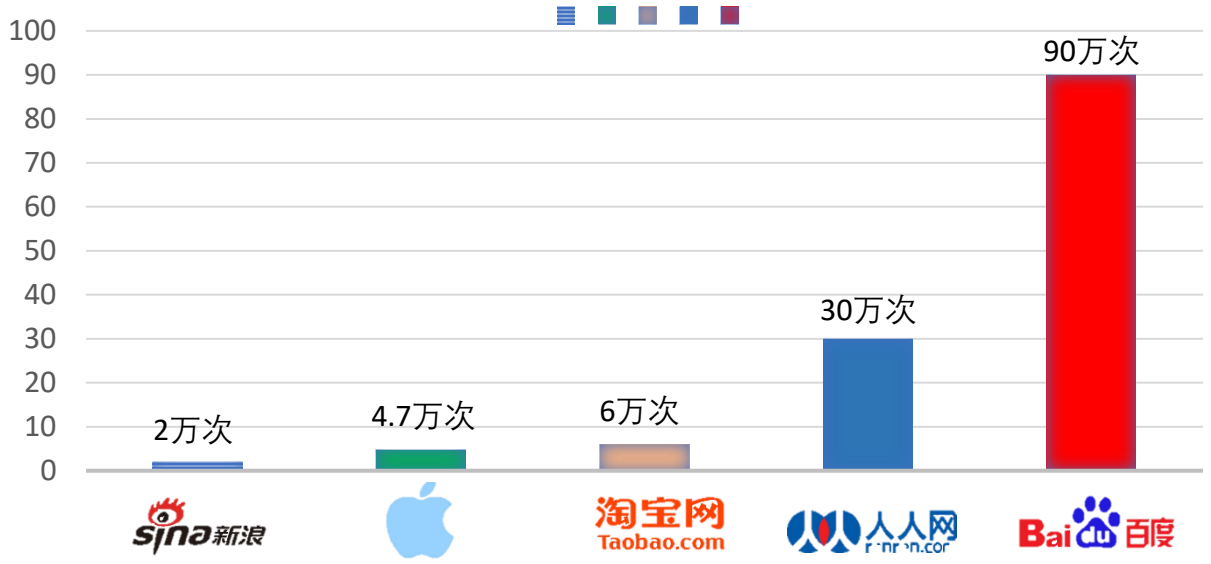


快速化　VELOCITY

1.2 大数据概念



1秒定律



一秒定律　从数据生成到决策响应仅需1秒

1.2 大数据概念



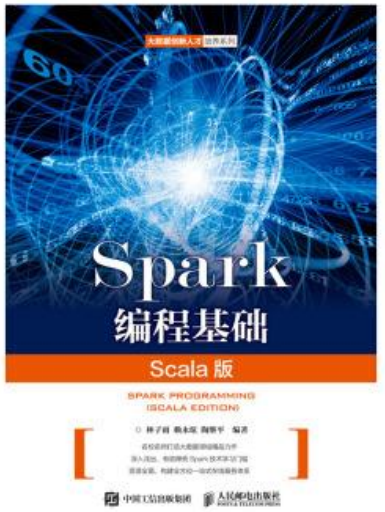








您感兴趣的



一秒之内给你响应

1.2 大数据概念

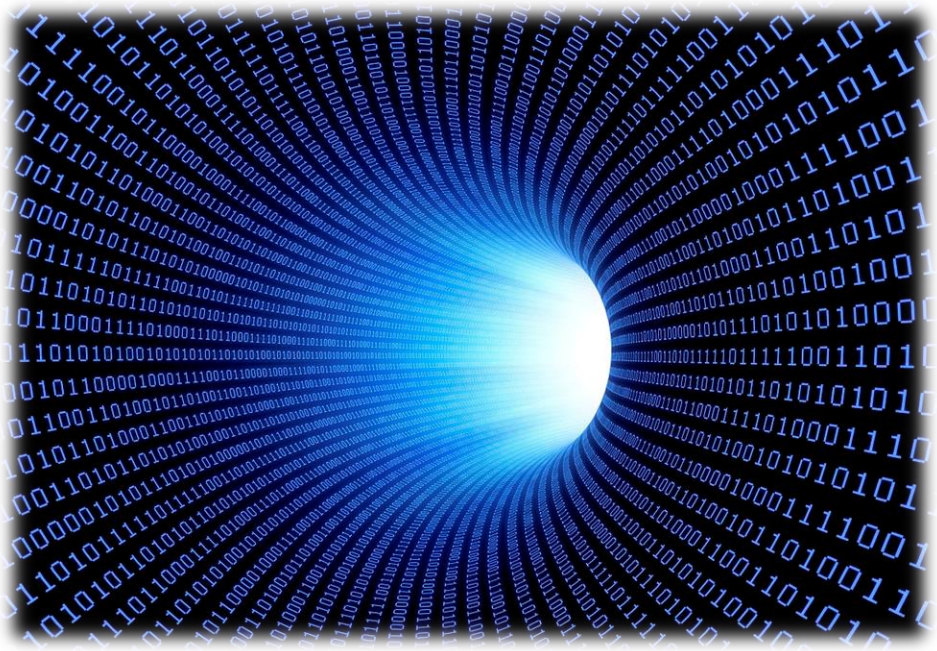






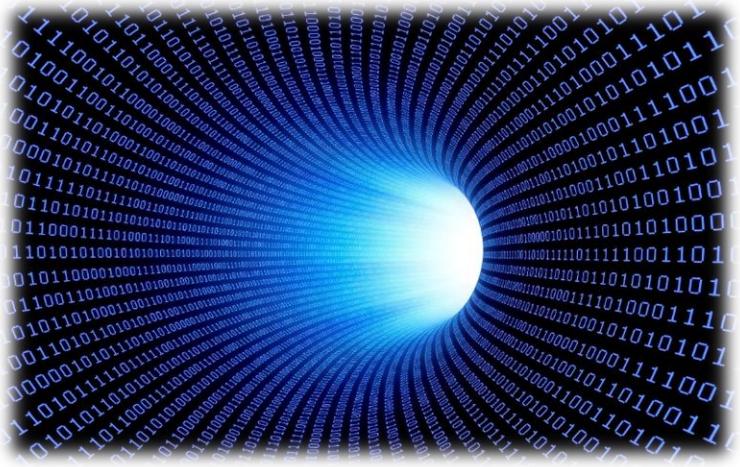
每一次鼠标点击

1.2 大数据概念



生成鼠标点击数据流

1.2 大数据概念



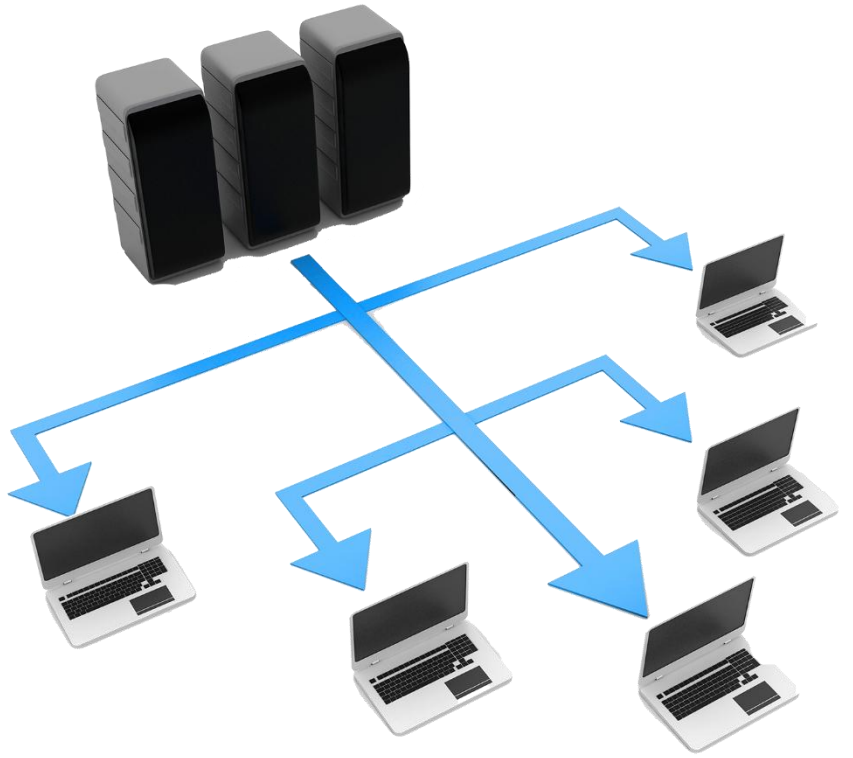
实时发送



生成鼠标点击数据流

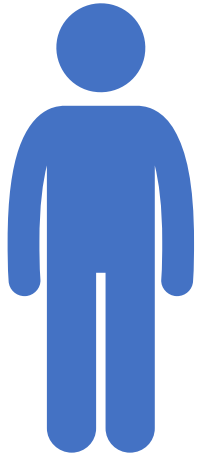
大数据分析的后台

1.2 大数据概念



对当前访问的信息进行实时的处理

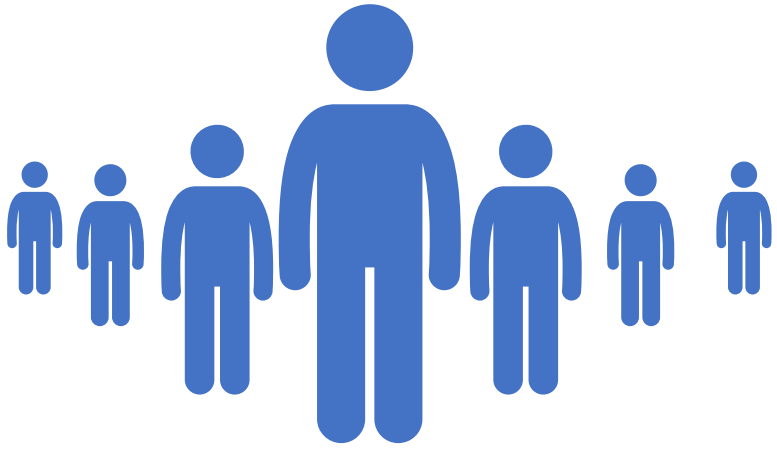
1.2 大数据概念



用户具备特征



进行匹配



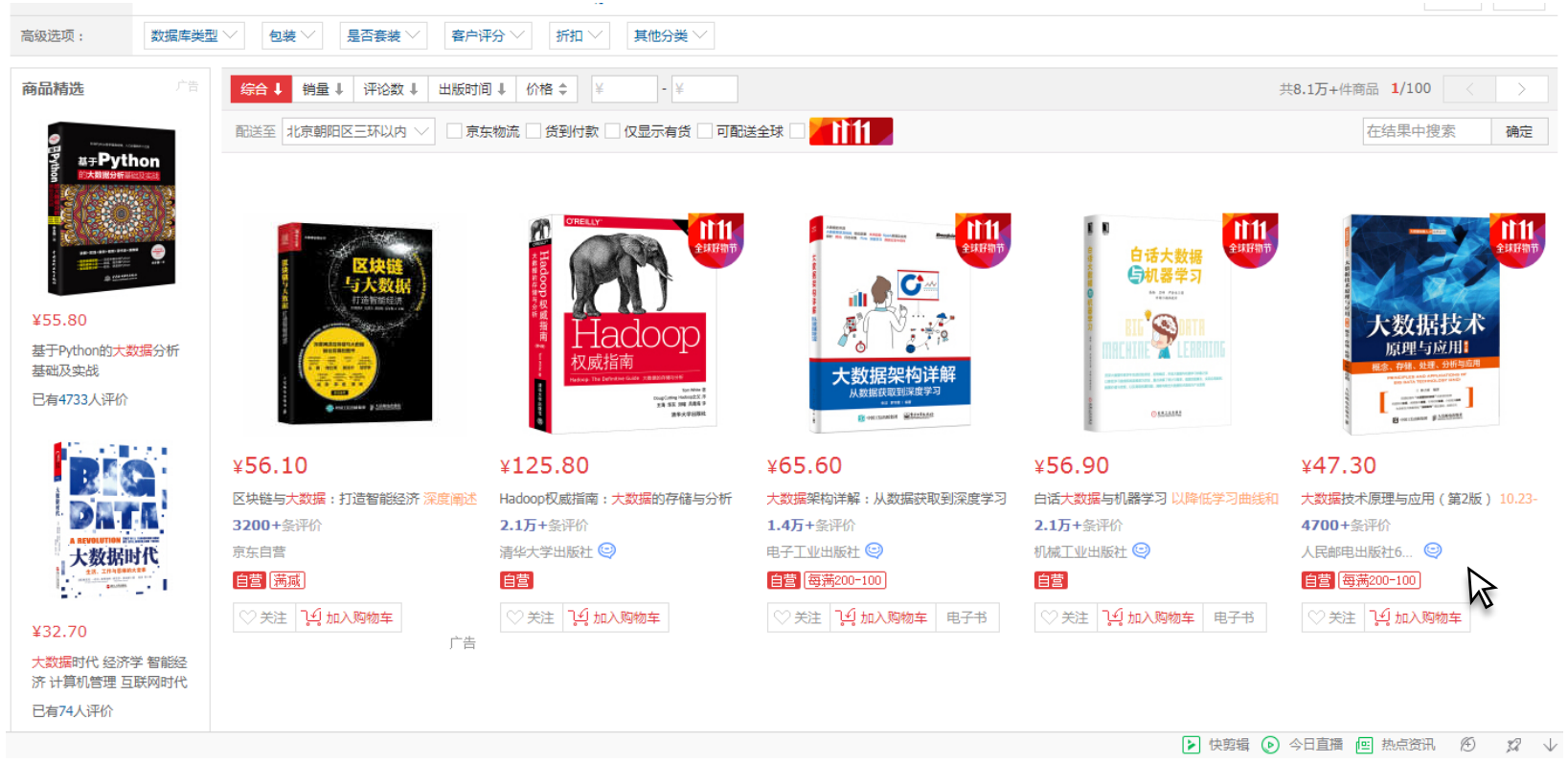
群体特征

1.2 大数据概念



群体已经买过的商品(你没有买过的商品)推荐给你

1.2 大数据概念



京东网站购物点击选择商品

1.2 大数据概念





实时推荐您可能感兴趣的商品

1.2 大数据概念

过了五秒十秒钟

您感兴趣的





一秒钟的响应 才能够具备它相应的商业价值

1.2 大数据概念



价值密度低



价值密度低　VALUE

1.2 大数据概念



价值密度低



1.2 大数据概念







价值密度低

1.2 大数据概念





数据

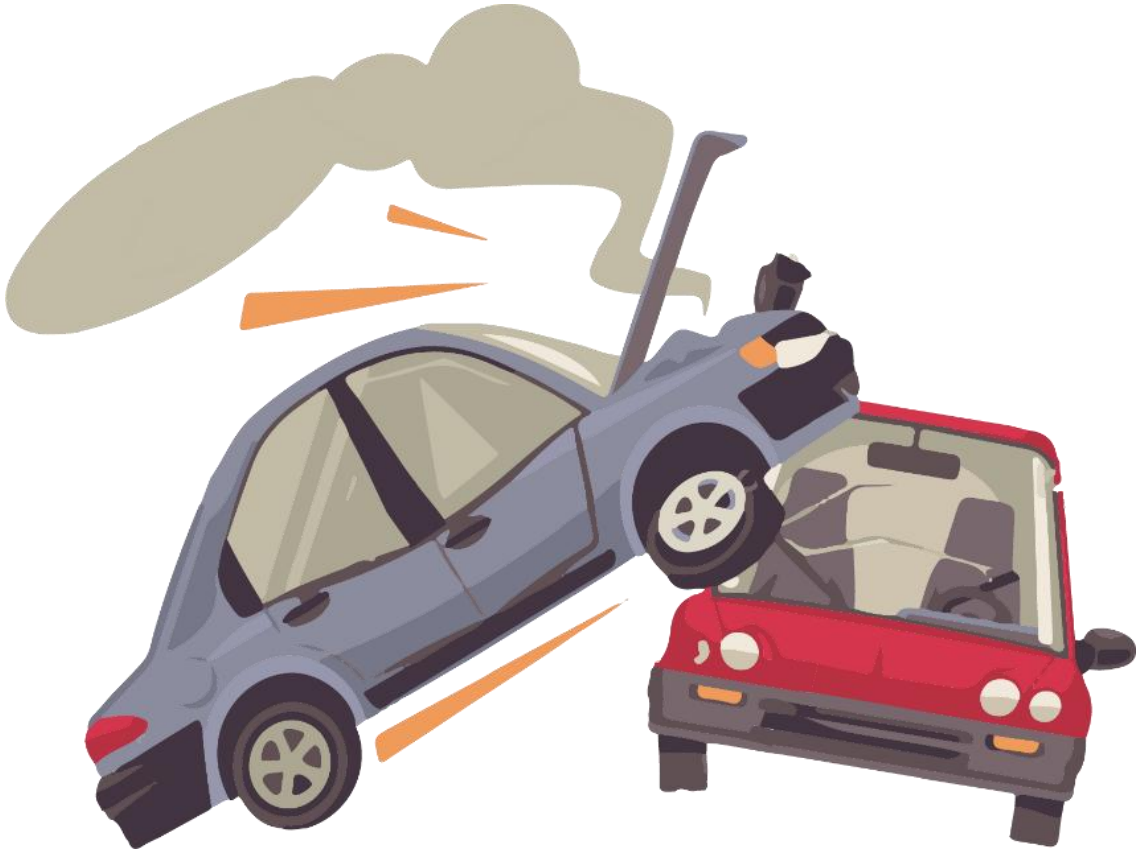
没发生任何的意外事件



垃圾数据

1.2 大数据概念

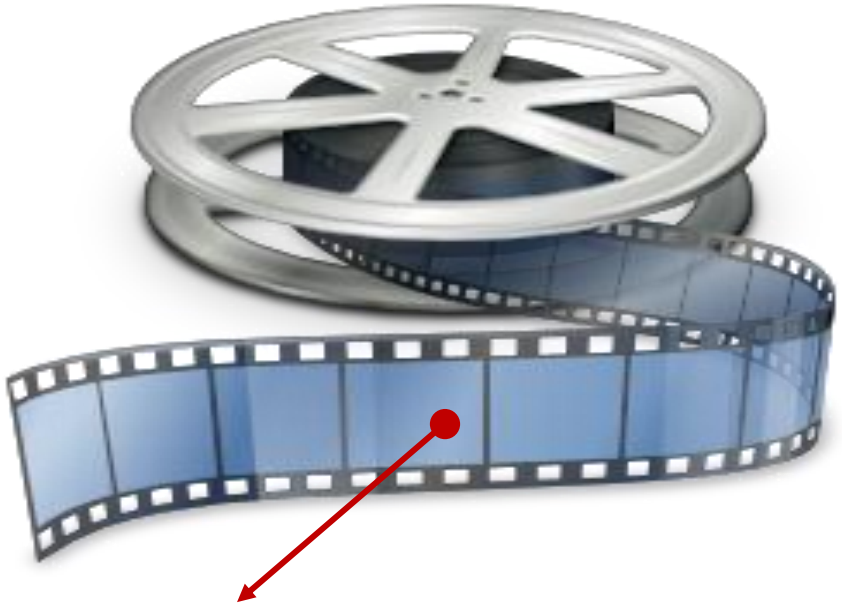




发生意外事件

1.2 大数据概念





- 意外事件瞬间的视频　　单点价值非常高

1.2 大数据概念



价值密度低　单点价值高

1.2 大数据概念理解测验（选择题）



第1题

以下哪一项**不属于**大数据的4V特征？

A. 数据量大（Volume）
B. 数据安全（Validity）
C. 数据类型多样（Variety）
D. 处理速度快（Velocity）



1.2 大数据概念理解测验（选择题）



第2题
在大数据中，“价值密度低”意味着什么？

A. 所有数据都很有价值
B. 只有结构化数据才有用
C. 大多数数据价值不高，少部分数据非常有价值
D. 数据越多，价值就越高

1.2 大数据概念理解测验（选择题）



第3题
根据 IDC 的研究，人类社会的数据每几年翻一倍?

A. 每年
B. 每 5 年
C. 每 3 年
D. 每 2 年

1.2 大数据概念理解测验（选择题）



第4题
哪种类型的数据属于非结构化数据?

A. Excel 表格
B. SQL 数据库中的订单记录
C. 图片和视频
D. 银行流水账单

1.2 大数据概念理解测验（选择题）



第5题
以下哪个例子最能体现 Velocity 的特征？

A. 存储海量历史档案
B. 实时推荐你可能喜欢的商品
C. 用硬盘长期备份文件
D. 打印纸质报告

1.2 大数据概念理解测验（选择题）



第6题
结构化数据主要以什么形式存储?

A. 图像文件夹
B. 视频流平台
C. 关系型数据库
D. 压缩包

1.2 大数据概念理解测验（选择题）



第7题
为什么说大数据中的 Variety 是一个挑战？

A. 数据太小，不好处理
B. 数据太整齐，无法分析
C. 数据类型繁多，传统算法难以处理
D. 所有数据都来自一个来源

1.2 大数据概念理解测验（选择题）



第8题

如果一段视频中只有短短几秒记录了事故现场，这种数据体现了什么？

A. Volume

B. Variety

C. Velocity

D. Value

1.2 大数据概念理解测验（选择题）



第9题
以下哪个数据单位最大？

A. GB
B. TB
C. EB
D. ZB

1.2 大数据概念理解测验（选择题）



第10题

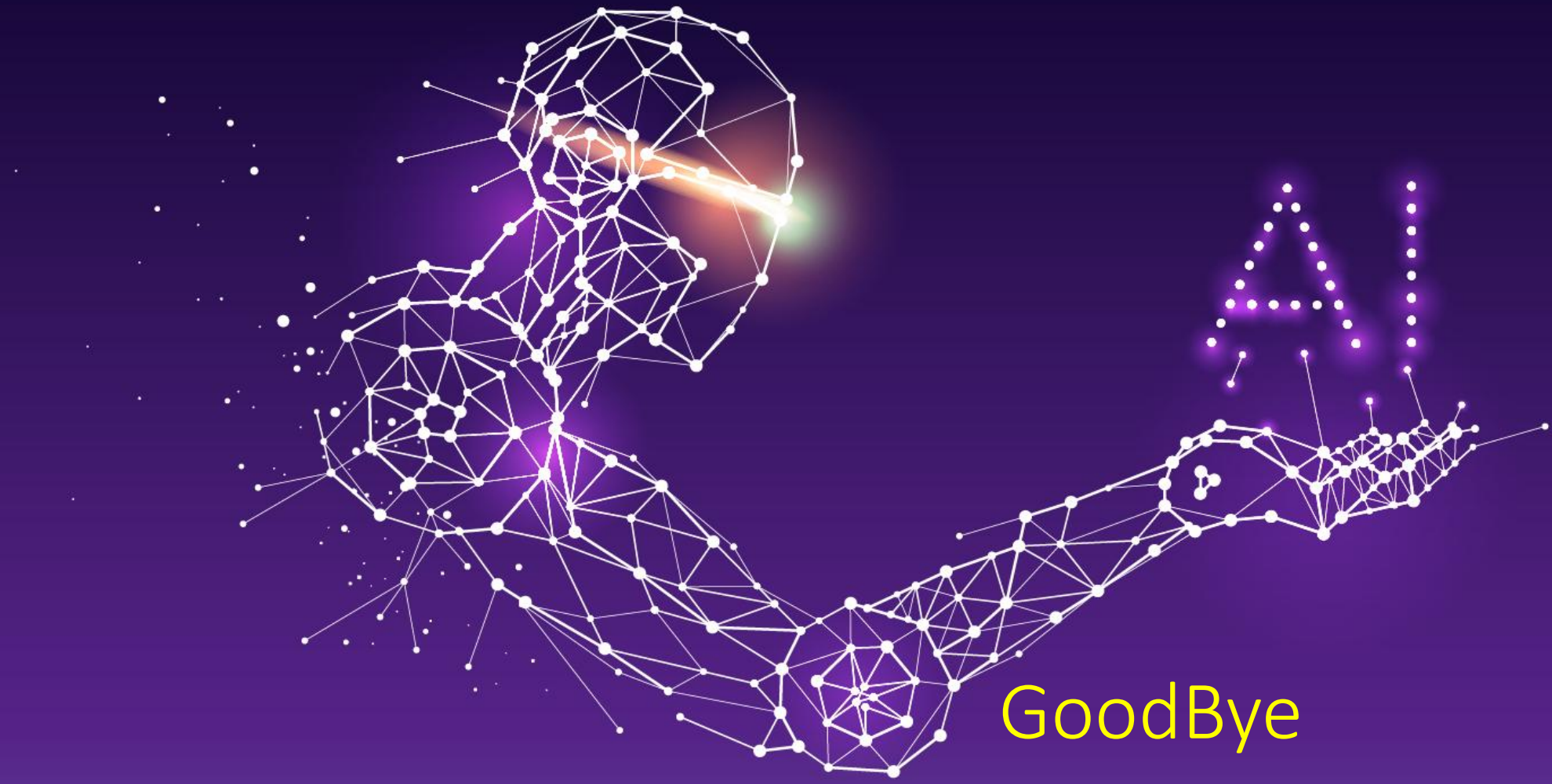
大数据推荐系统会根据什么来推荐你没买过的商品？

A. 你的家庭住址

B. 群体中其他相似用户的行为

C. 你最近的搜索次数

D. 你的电话号码



GoodBye