大数据分析计算机基础

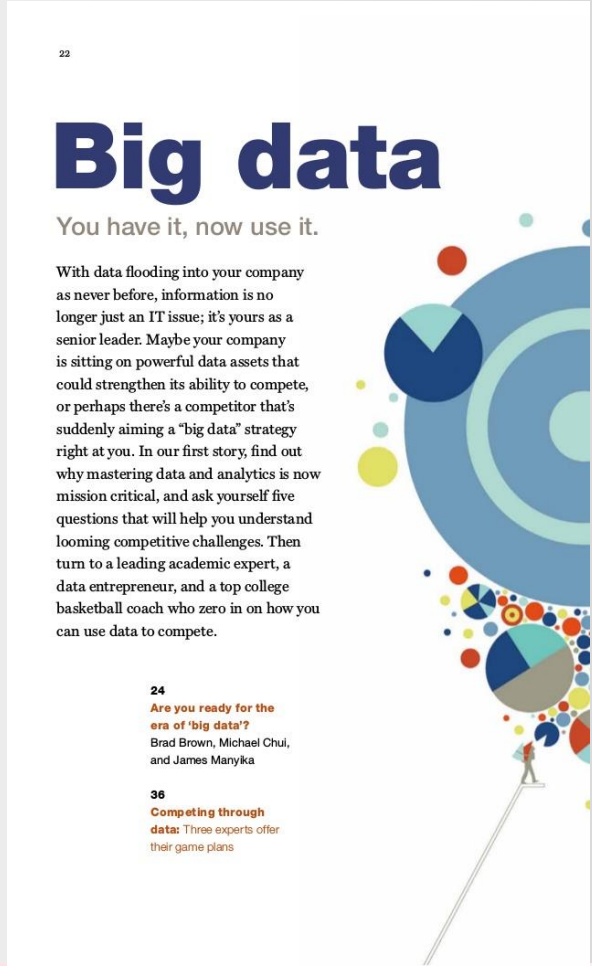
Computer Foundation for Big Data Analysis

北京大学·信息科学技术学院

邓习峰(dengxifeng@pku.edu.cn)

初识大数据：大数据的一些事儿

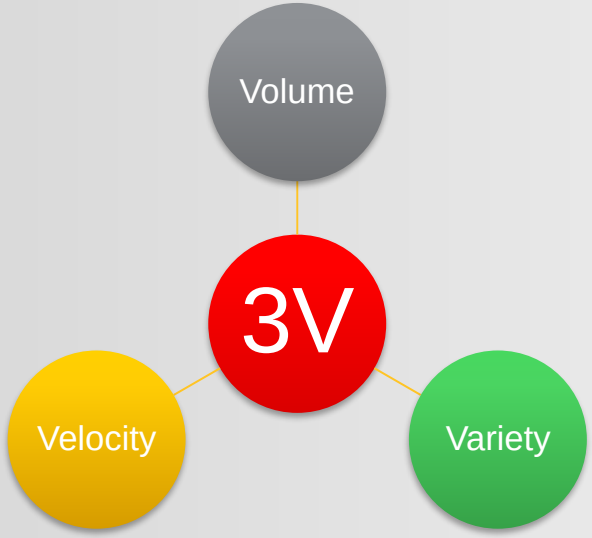
- 2008年9月，《自然》（Nature）刊登了一个名为“Big Data”的专辑；
- 2011年11月，麦肯锡(McKinsey)发布《大数据：创新、竞争和生产力的下一个前沿》；
- 2013年3月29日，美国总统奥巴马政府宣布推出“大数据研究和发展计划”(Big Data Research and Development Initiative)；
- 2015年8月31日，中国国务院正式印发《促进大数据发展行动纲要》。

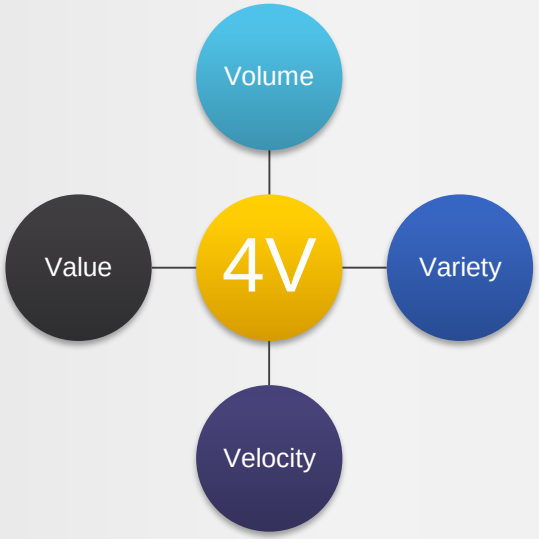


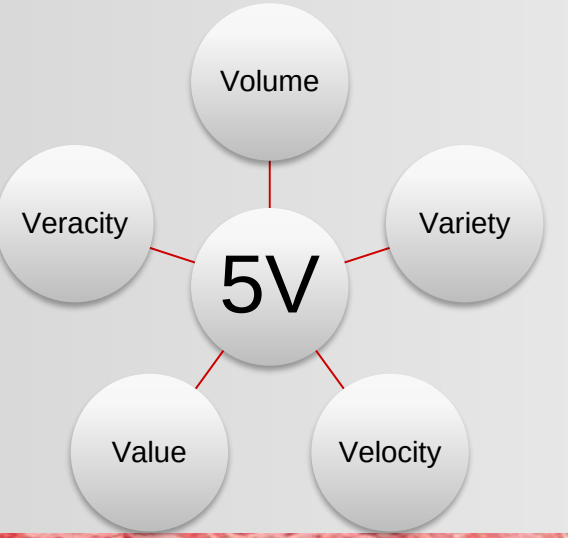
初识大数据：术语与口语

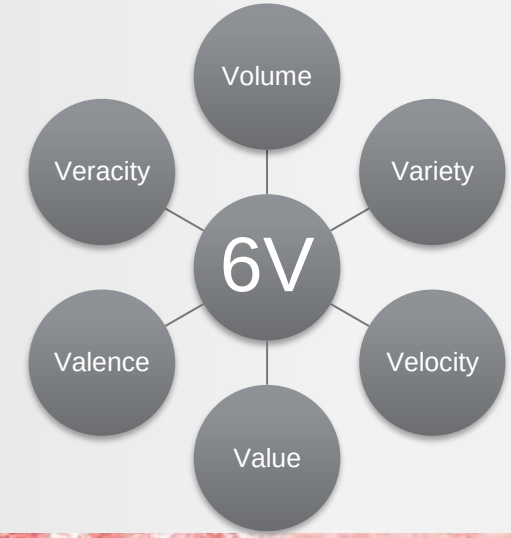
大量数据（huge data）不是大数据（big data）！

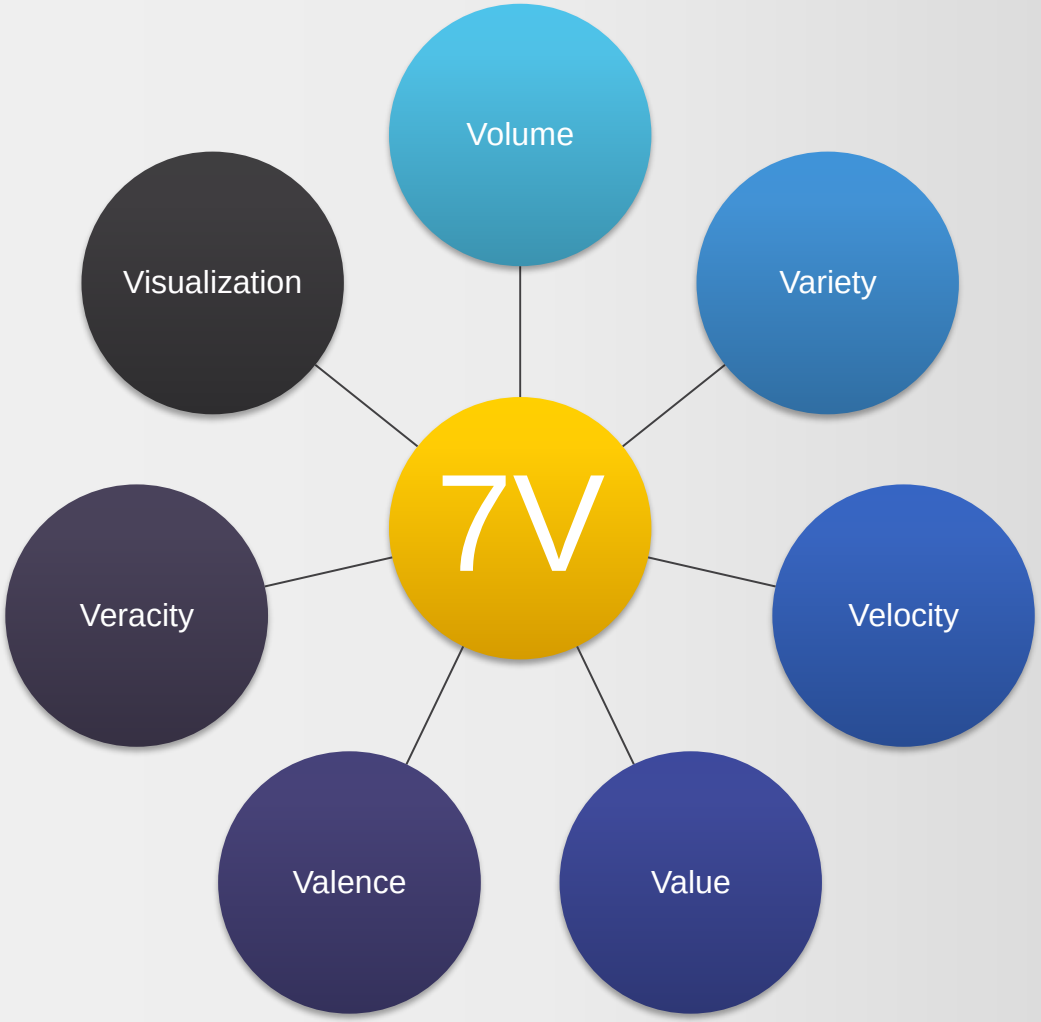
初识大数据：多少Vs？


Volume
3V
Velocity
Variety


Volume
Value
4V
Variety
Velocity


Volume
Veracity
Variety
5V
Value
Velocity

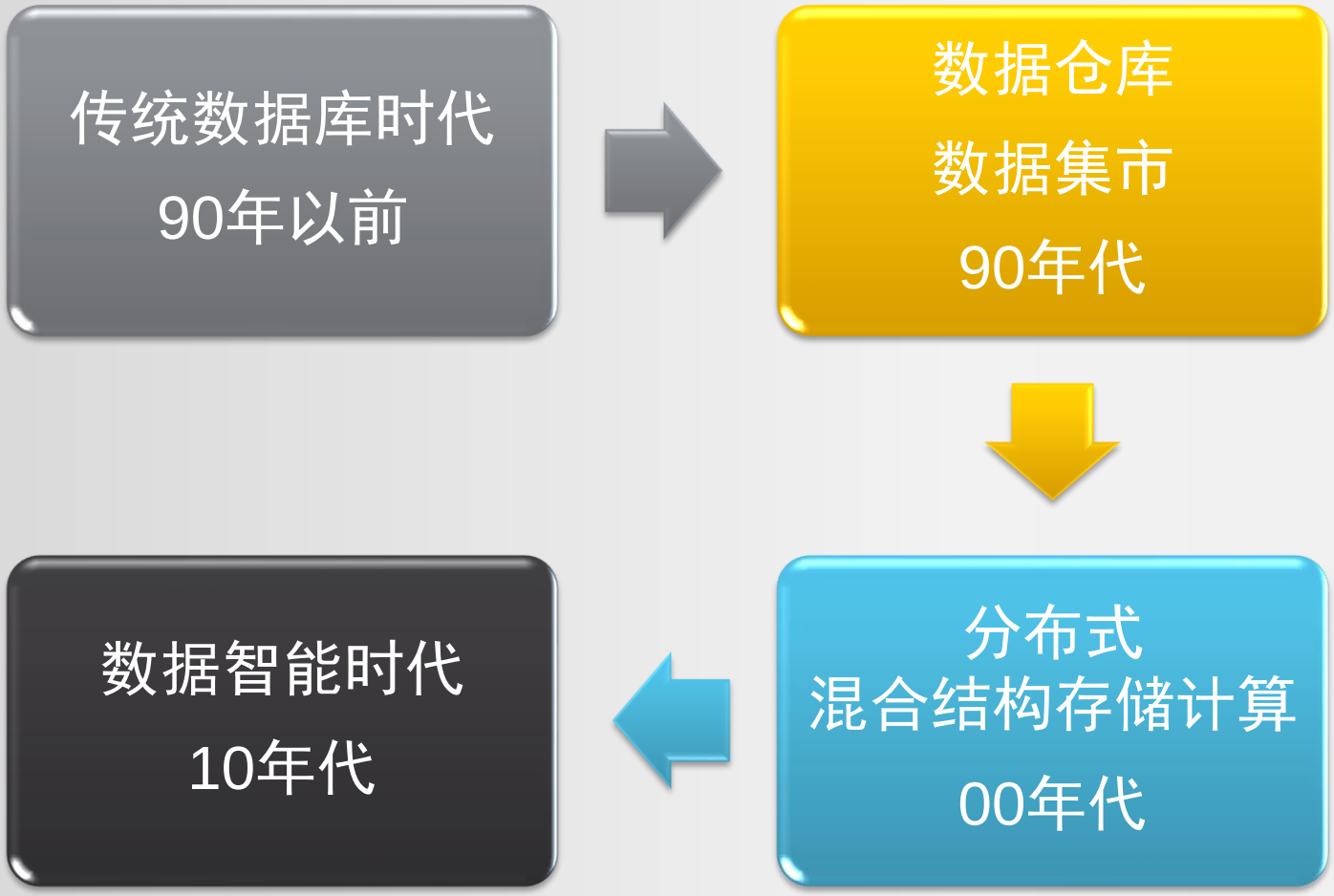

Volume
Veracity
Variety
6V
Valence
Velocity
Value


Volume
Visualization
Variety
7V
Veracity
Velocity
Valence
Value

初识大数据：多少Vs？

- Volume：大容量；
- Velocity：高速性，意指数据快速变化，因时而变；
- Variety：多样性，即不同类型的数据整合在一起；
- Value：价值密度低；
- Variability：可变性，不一致性，意指不同来源的数据，不同时间的数据，数据意义可能不一样；
- Veracity：真实性。要考虑数据源的真实性，完整性；
- Visualization：可视性。可视化是大数据研究的重要手段，也因为数据量大是重要挑战；
- Validity：有效性。数据治理、数据清理的重要性；
- Vulnerability：漏洞。大数据有秘密泄露风险。脱敏、消密异常重要；
- Volatility：挥发性易失性。多长时间的数据与当前决策有关，即多长时间的数据。
- Valence：联结性。数据与数据之间的关系。

初识大数据：发展阶段划分



- 传统数据库时代：上世纪90年代以前，关系数据库（relational database）在数据存储管理上占据主流，结构化数据（行列数据）管理为主。关系数据库更多是面向事务，可称为业务信息化；
- 数据仓库时代：数据仓库（Data Warehouse）由比尔·恩门（Bill Inmon）于1990年提出。相对于关系数据库，数据仓库面向主题，与此相关的概念有Data Mining、数据集市（Data Mart）以及商业智能（BI，Business Intelligence）等，强调了数据的应用分析，利用数据决策；
- 分布式混合存储计算：由于互联网的出现，数据量和非结构化数据存储面临挑战，以Hadoop存储和Spark计算为重要标志；
- 数据智能时代：以2011年5月麦肯锡(McKinsey)发布《大数据：创新、竞争和生产力的下一个前沿》为标志，尤其是紧随其后的AI技术的发展，更是为大数据发展添加动力。

初识大数据：大数据来源

初识大数据：万物皆联

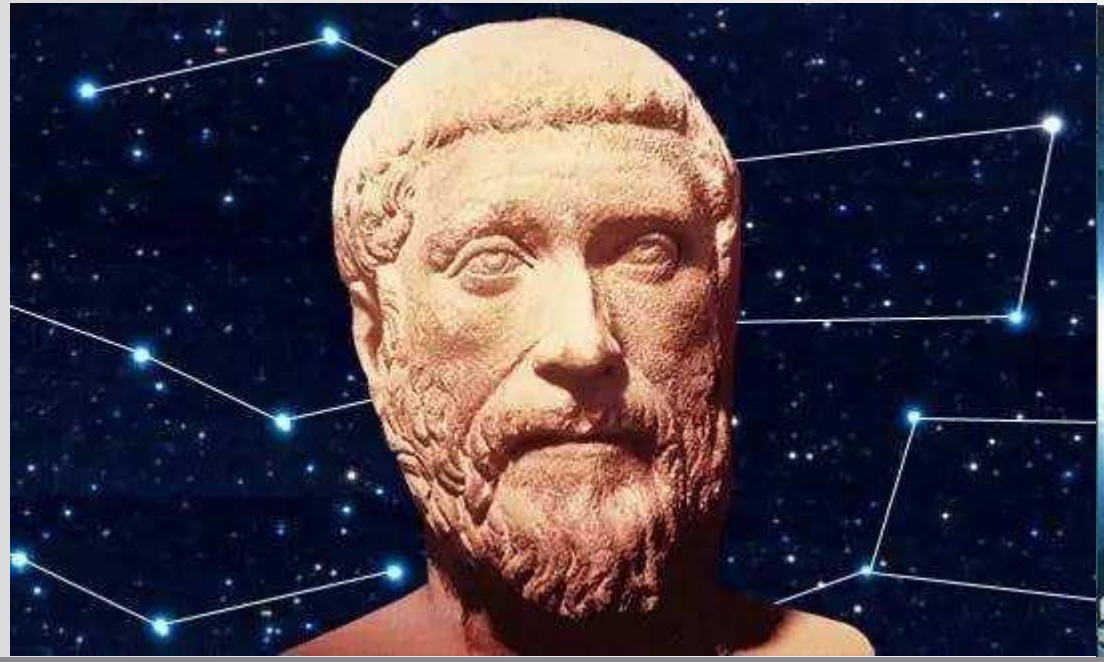
联者数也

IoT=Internet of Things，物联网

传感器(Sensor)：将物理量(光热气力磁等等)转换为电信号(多为电信号)，然后转换为数字信号并进一步传输存储处理等，是物联网的基础。

感 → 转 → 传 → 算

初识大数据：



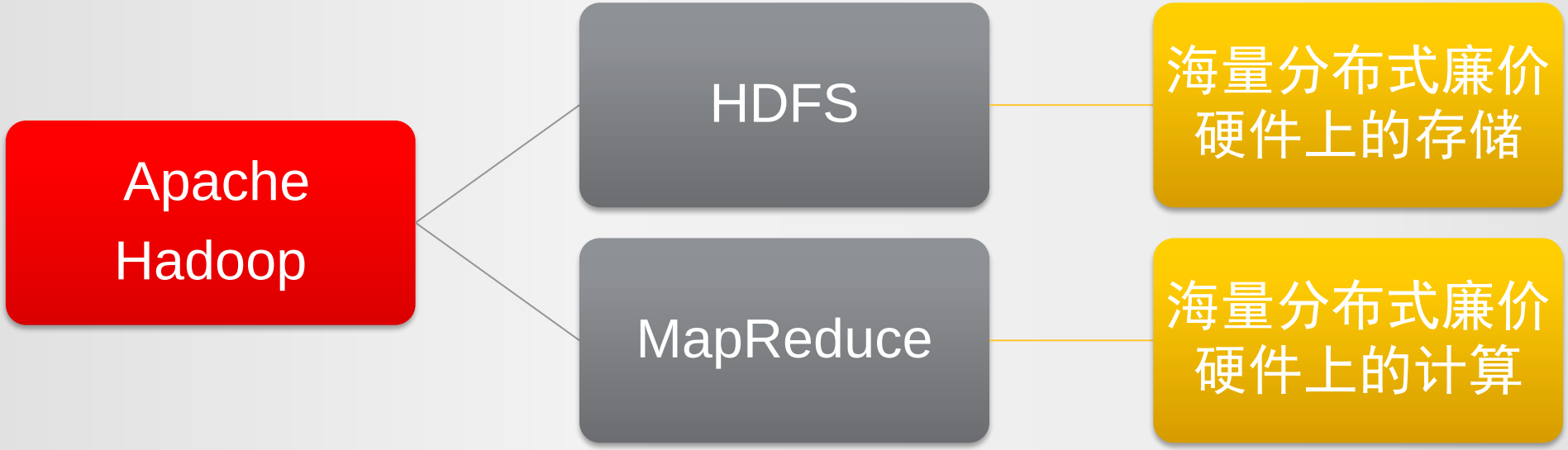
万物皆数 一切皆算

All is number! All is computing!

初识大数据：技术挑战→大存储

- 2010年时，主流硬盘容量是1TB，目前是5-10TB；
- 以北京公共交通刷卡数据为例，每次交易大约产生200B数，每天大约4000万次交易，即每日产生约8GB数据，每月约240GB，每年约3000GB，即3TB数据。这仅仅是刷卡交易数据，没有包含线路等数据；
- 对于双十一等场景，短时间内产生的数据，更是巨量。

传统存储和计算面临巨大挑战



初识大数据：技术挑战→大计算

大数据高相关于大计算，大计算的实现依赖大内存，或高性能CPU。单机内存难以大规模扩展，多机尤其廉价通用计算机集群，对大计算的实现有重要意义，Apache Spark是基于分布式内存的大计算。和Apache Hadoop相比，Apache Spark有如下特点：

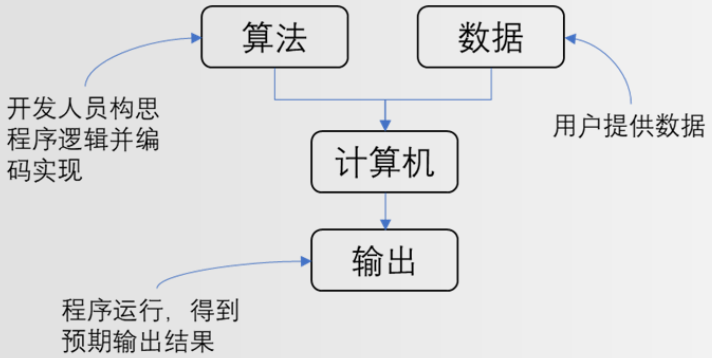
- Apache Spark是分布式计算分析框架，专门用来对分布式存储的数据进行计算处理，不直接支持外部持久存储（外存如硬盘等）；
- Apache Hadoop是两步计算磁盘存储，而Apache Spark是多步计算内存存储。Hadoop可以大致分为Map阶段（数据筛选）和Reduce阶段（合并计算），计算完成后需要存储到磁盘系统之中，然后开始其他MapReduce。而Spark则MapReduce后，其结果保存到内存之中，然后开始新的MapReduce，因此Spark计算效率更高。当内存空间不够时，则可以缓存到磁盘系统；
- Apace Spark是内存集群计算，可在内存集群中将数据集缓存在内存中，以缩短访问延迟；
- Apache Hadoop本质上数据存储基础设施，而Apace Spark则是内存集群计算基础设施，而二者面向不同的目标。

初识大数据：技术挑战→算力(GPU)

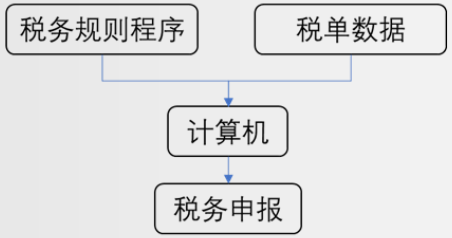
GPU是Graphics Processing Unit，即图形处理单元，传统上用于图形图像多媒体处理，其特点是并行处理能力强、计算能效比高，并且有很大的存储带宽。现阶段大量人工智能（机器学习）模型训练与推理、高性能计算等，往往是大数据流应用，用GPU解决这类问题，就比CPU效率更高，它对于用传统语言编写的软件形式的计算有较好的支持，具有高度的灵活性。

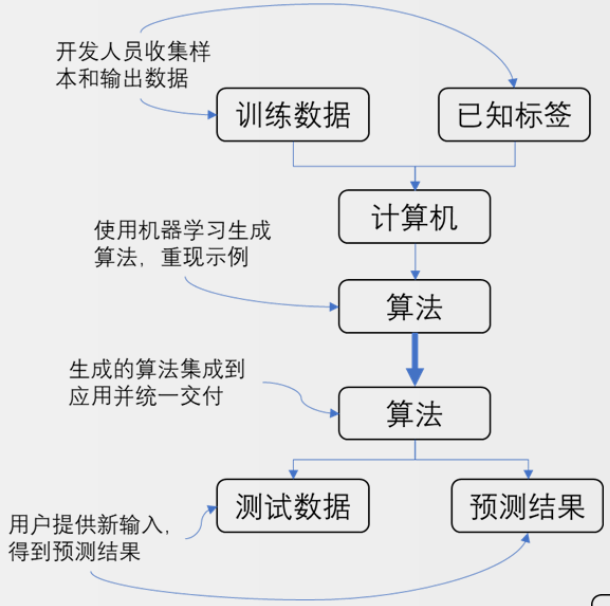
- Apache Hadoop是大磁盘存储体系、Apache Spark是大内存体系、GPU是大计算体系；
- GPU天然拥有大量并行处理能力，是专用处理器，相对于其他解决方案更加通用；

初识大数据：技术挑战→AI算法

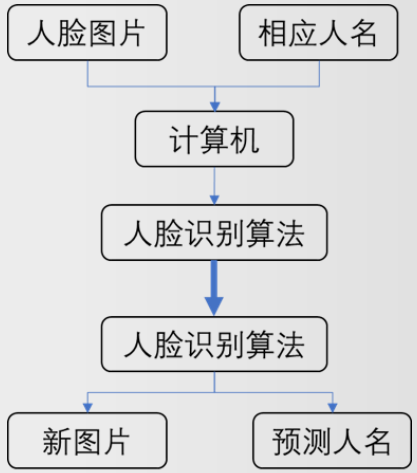


传统编程范式





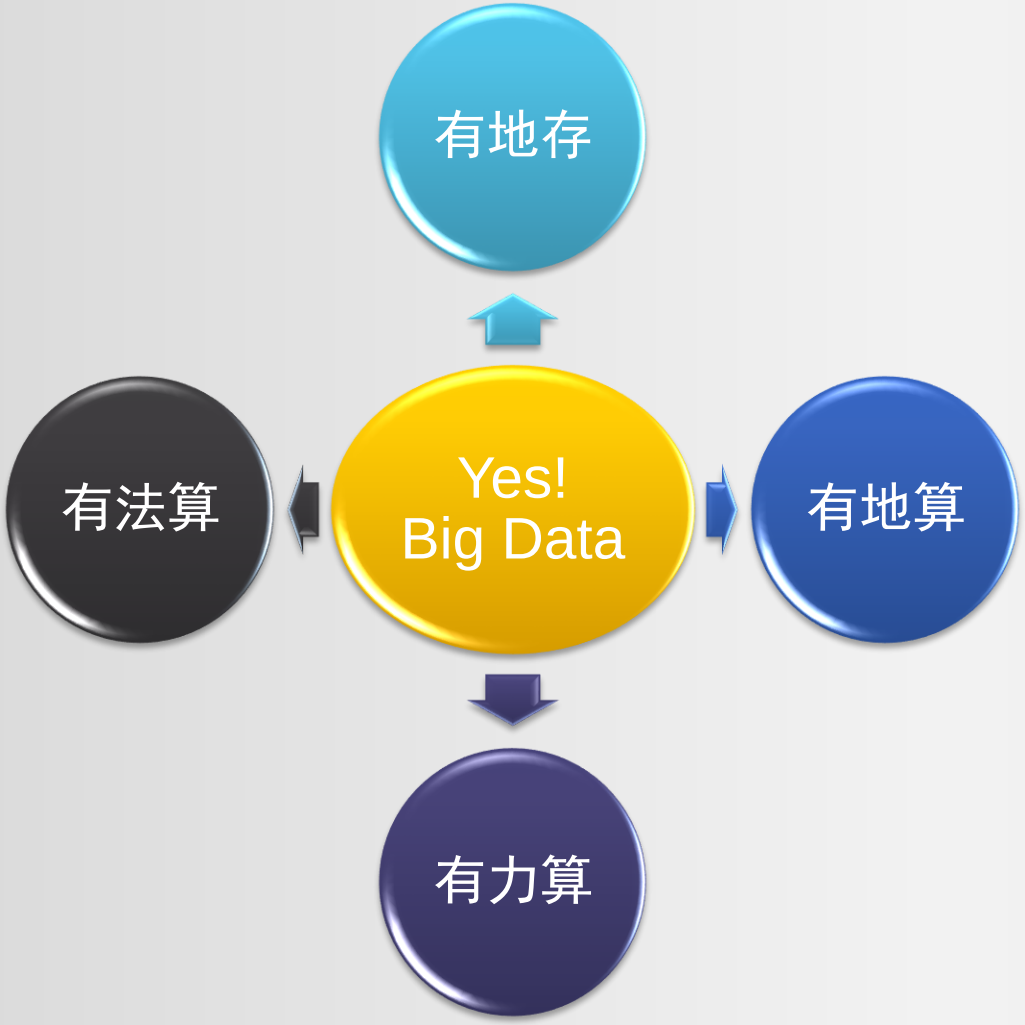
机器学习范式



初识大数据：技术挑战→AI算法

- **训练集**：其功能是拟合模型，通过设置参数，训练模型，与**验证集**相结合，选定同一参数不同取值，拟合出一个或多个模型。**极大量重复使用！**
- **验证集**：其功能是在**训练集**训练出的**多个模型**中寻找最佳模型(同一参数不同取值)，即使用各个模型对验证集数据进行预测，并记录模型准确率，并选出效果最佳的模型所对应的参数，相当于调整模型参数。可以认为验证集和训练集是训练的不同环节。**多次使用不断调参！**
- **测试集**：通过训练集和验证集得出**最优模型**后，使用测试集进行模型预测。用来衡量该最优模型的性能和价值，评估模型最终泛化能力。**一次使用！**

初识大数据：小结



- 有地存：Apache Hadoop分布式外存储系统；
- 有地算：Apache Spark分布式内存系统；
- 有力算：GPU并行计算机能力；
- 有法算：基于大数据的AI算法。

初识大数据：大数据与云计算

虽然Apache Hadoop和Apache Spark是开源软件，GPU可以购买，但以此组建大数据分析系统，也并不容易，也并不能更好发挥系统能力，**云计算(Cloud Computing)**应时应用而生，实现资源集约化，降低使用门槛。**弹性计算(Elastic computing)**是云计算的核心，可快速扩展或减少计算处理、内存和存储资源以满足不断变化的需求，而无需担忧用量高峰的容量计划和工程设计。现阶段所说的云服务已经不单单是一种分布式计算，而是分布式计算、效用计算、负载均衡、并行计算、网络存储、热备份冗杂和虚拟化等计算机技术混合演进并跃升的结果，云计算形成计算能力极强的系统，可存储、集合相关资源并可按需配置，向用户提供个性化服务。

通常，它的服务类型分为三类，即基础设施即服务(IaaS，Infrastructure as a Service)、平台即服务(PaaS，Platform as a Service)和软件即服务(SaaS，Software as a service)。

初识大数据：案例1：搜狗输入法

词语：
计算而得
计算语言学

数据源：
Web渠道

更新：
基于网络经常更新

界面：
多变化，少传统工程模样

初识大数据：案例2：公交刷卡数据

基本数据：

- 上车：交易线路编号、交易车站号、交易时间、卡类型（学生、老年优待、常规）、交易金额（此处为0）、交易状态
- 下车：交易线路编号、交易车站号，交易时间、卡类型（学生、老年优待、常规）、交易金额（此处为实际金额），交易状态

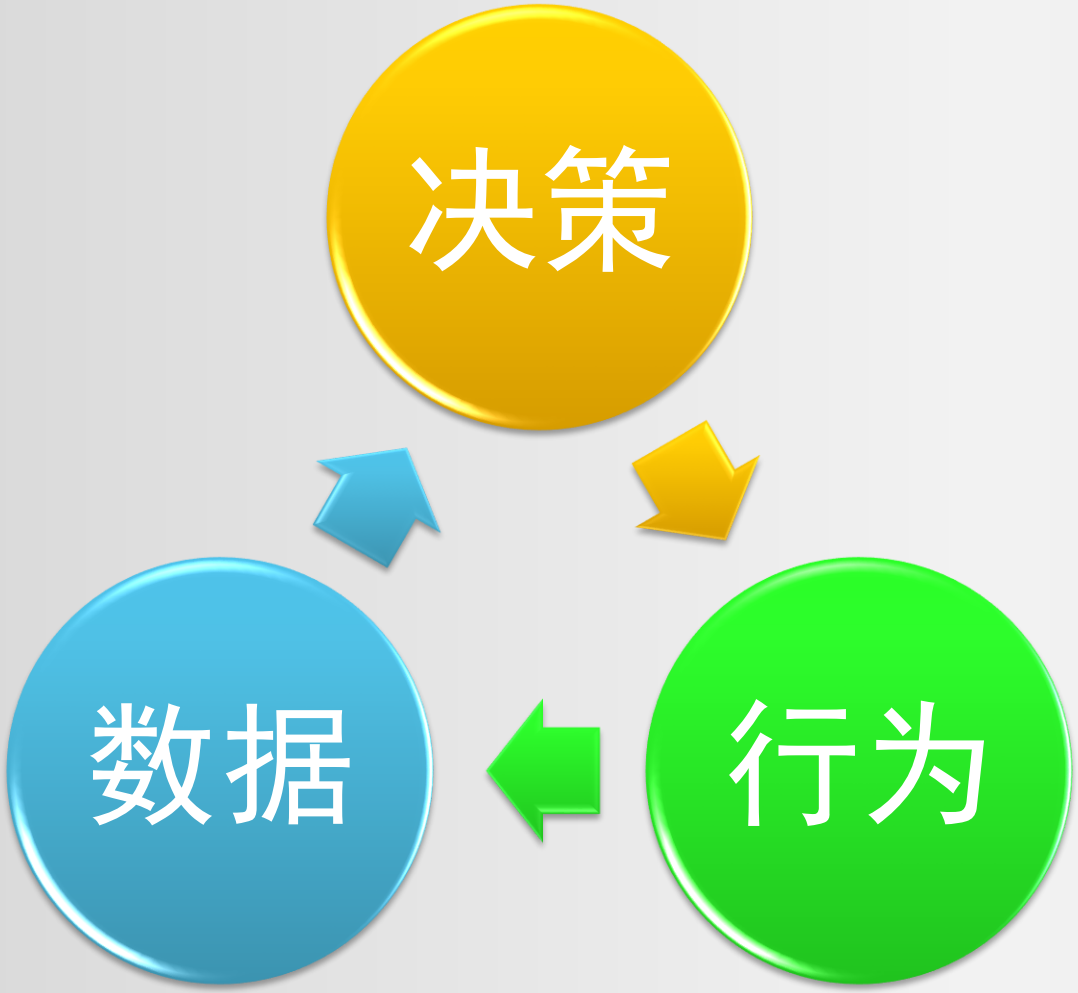
这些数据有哪些用途？如何更有用途？

初识大数据：手机信令数据

手机信令数据可以分为：话单数据（通话或短信数据）、PS域信令数据（上网信令）、CS域信令数据（基站切换、位置更新、开关机、位置区切换等）。总是包含ISMI号码（由服务商基于此映射为手机号）、时间戳、位置区编号、事件类型等。话单数据还包括：主叫、被叫、开始时间、结束时间、资费等。以此类推。

这些数据有哪些用途？如何更有用途？

初识大数据：基本理念



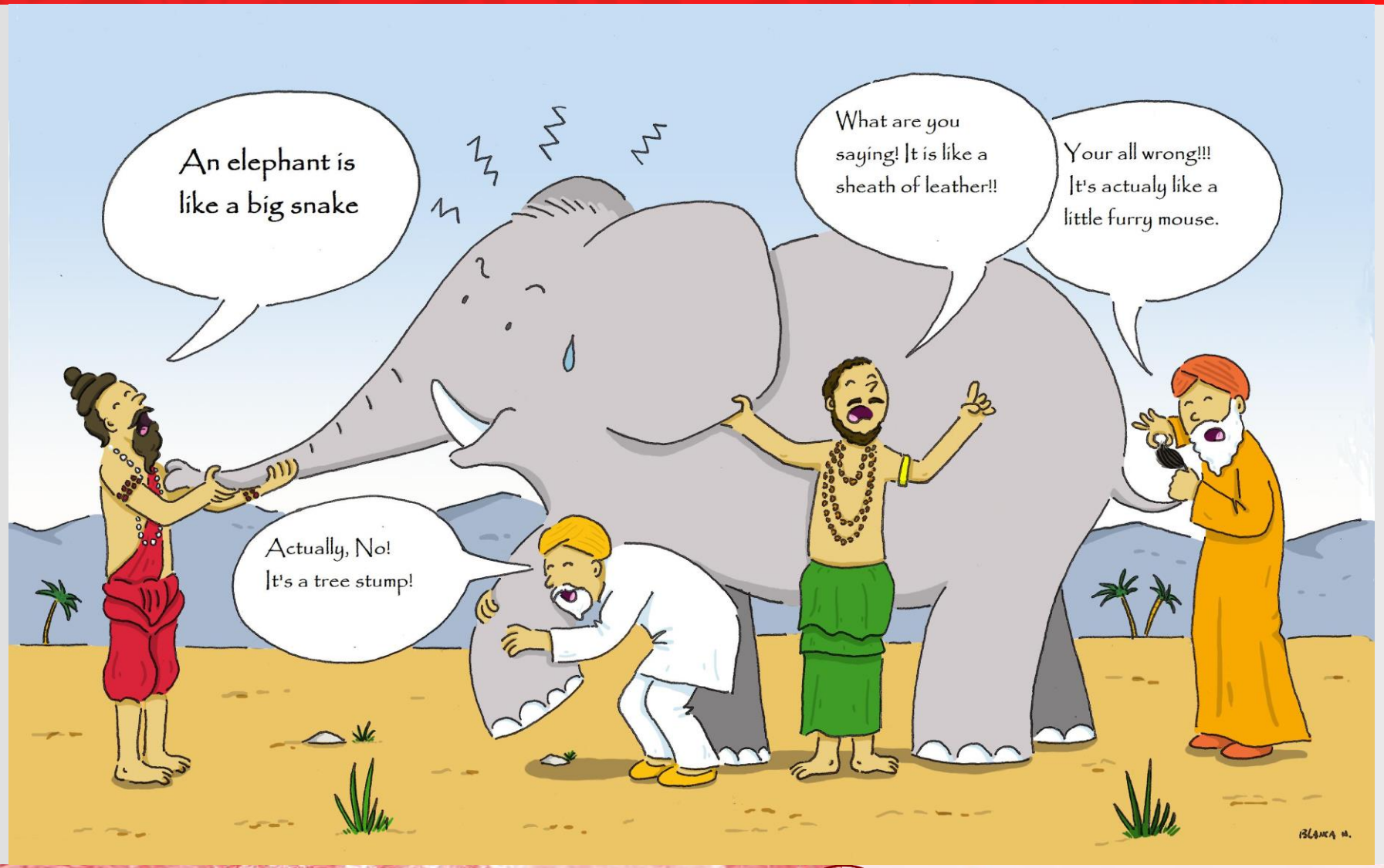
- 决策产生行为，行为产生数据，数据影响决策；
- 大数据为人工智能奠定基础；
- 人工智能本质上是发现人类行为要素以及要素影响的数学方式；
- 核心：行为要产生可记录并存储的数据。如果没有记录和存储数据，就不能为决策提供帮助，这是重大损失。

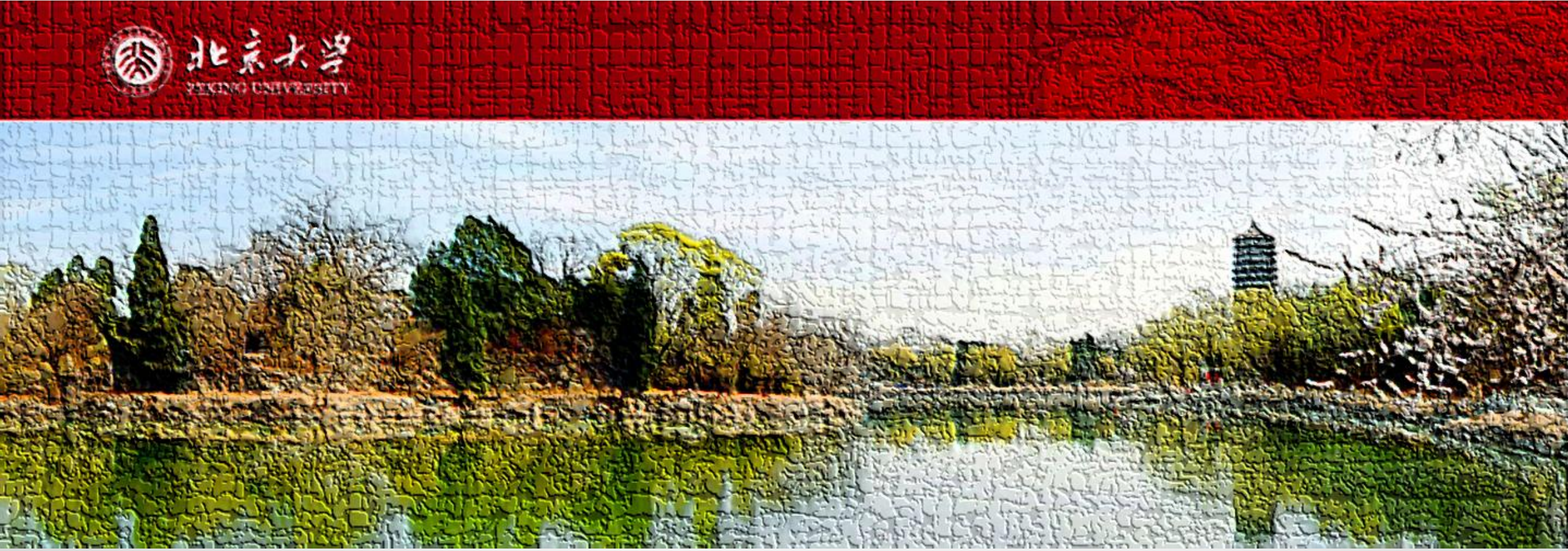
初识大数据：观念转变

大数据的三个颠覆性观念：

- 是全部数据，而不是随机采样；
- 是大体方向，而不是精确制导；
- 是相关关系，而不是因果关系。

初识大数据：数据若水


An elephant is like a big snake
Actually, No! It's a tree stump!
What are you saying! It is like a sheath of leather!!
Your all wrong!!! It's actualy like a little furry mouse.



博学之 审问之 慎思之 明辨之 笃行之

The End!