Chapter 3　Application

Unit 3.5 Big Data

备课时间：2019/10/27~2019/11/05

词汇与词组

1. **Big data** usually includes **data sets** with sizes beyond the ability of commonly used software tools to **capture**, manage, and process data within a **tolerable elapsed time**.

- **Big data**　大数据
- **data sets**　数据集
  1) Indoor Scene Recognition 室内场景识别数据集：包含 67 个室内类别，15620 个图像。2.4GB
     http://web.mit.edu/torralba/www/indoor.html
  2) FMA 是音乐分析的数据集，1000 GB
     https://github.com/mdeff/fma
  3) Open Images 数据集，包含 9,011,219 张图像的训练集，41,260 张图像的验证集以及 125,436 张图像的测试集
     https://storage.googleapis.com/openimages/web/index.html
- **Capture**　获取
- **tolerable**　可接受的
- **elapsed time**　运行时间

used to describe the **time that passes between** the **start** and **end** of a project or a computer operation, in contrast to the actual time needed to do a particular task which is part of the project

2. Big data “size” is a **constantly moving target**, as of 2012 ranging from a few dozen **terabytes** to many **petabytes** of data.

- **constantly moving target** 不断变化的指标
- **as of** 自……起；到……时候为止

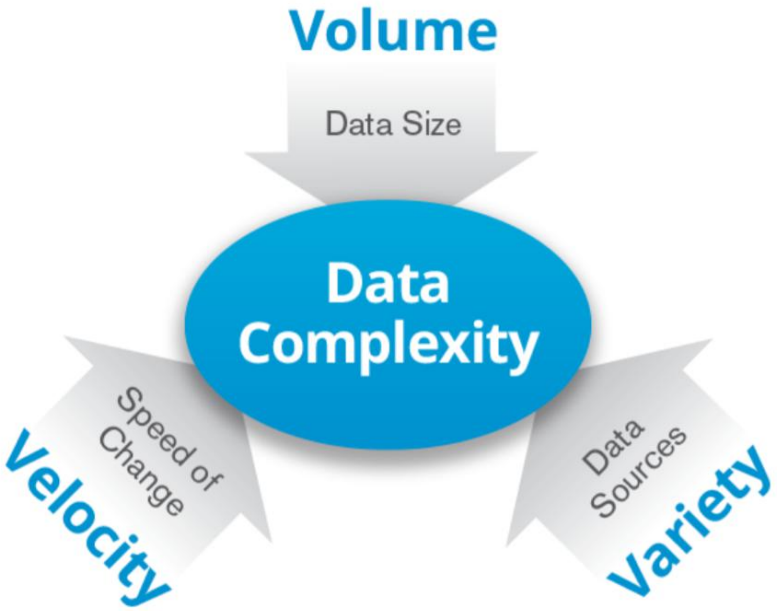
  It is effective as of a certain date.

  [本合同]视为在某年某月某日生效.
- **terabyte** /ˈterəbaɪt/太字节，TB；1TB=1024GB
- **petabyte** /'petəbaɪt/拍字节，PB；1PB=1024TB=2^{50} 字节

中文单位	中文简称	英文单位	英文简称	字节数
位	比特	bit	b	1/8
字节	字节	Byte	B	1
千字节	千字节	KiloByte	KB	2^10
兆字节	兆	MegaByte	MB	2^20
吉字节	吉	GigaByte	GB	2^30
太字节	太	TeraByte	TB	2^40
拍字节	拍	PetaByte	PB	2^50
艾字节	艾	ExaByte	EB	2^60
泽字节	泽	ZettaByte	ZB	2^70
尧字节	尧	YottaByte	YB	2^80

3. In 2012, **Gartner** gave its definition as follows, “Big data is high **volume**, high **velocity**, and/or high **variety** information **assets** that require new forms of processing to enable enhanced **decision making**, **insight discovery** and **process optimization**.”

“大数据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。”



- **Volume** /ˈvɒljuːm/ 量；体积；卷；音量
- **Velocity** /vəˈlɒsəti/ 速率；速度
- **Variety** /vəˈraɪəti/ 多样；种类；变化，多样化
- **Asset** /ˈæset/ 资产

  You do not need any guarantees or **asset** evaluation.

  你不需要任何担保或资产评估。
- **decision making** 决定，决策
- **insight discovery** 洞悉发现
- **process optimization** 流程（过程）优化

This model can facilitate the design, operation, **process optimization** and advanced control of a SMBC (simulated moving bed chromatography) unit.

该数学模型可以用来指导模拟移动床色谱装置的设计操作、**过程优化**与先进控制。

- **Gartner** 美国高德纳咨询公司

  https://www.gartner.com/en/information-technology/glossary/big-data

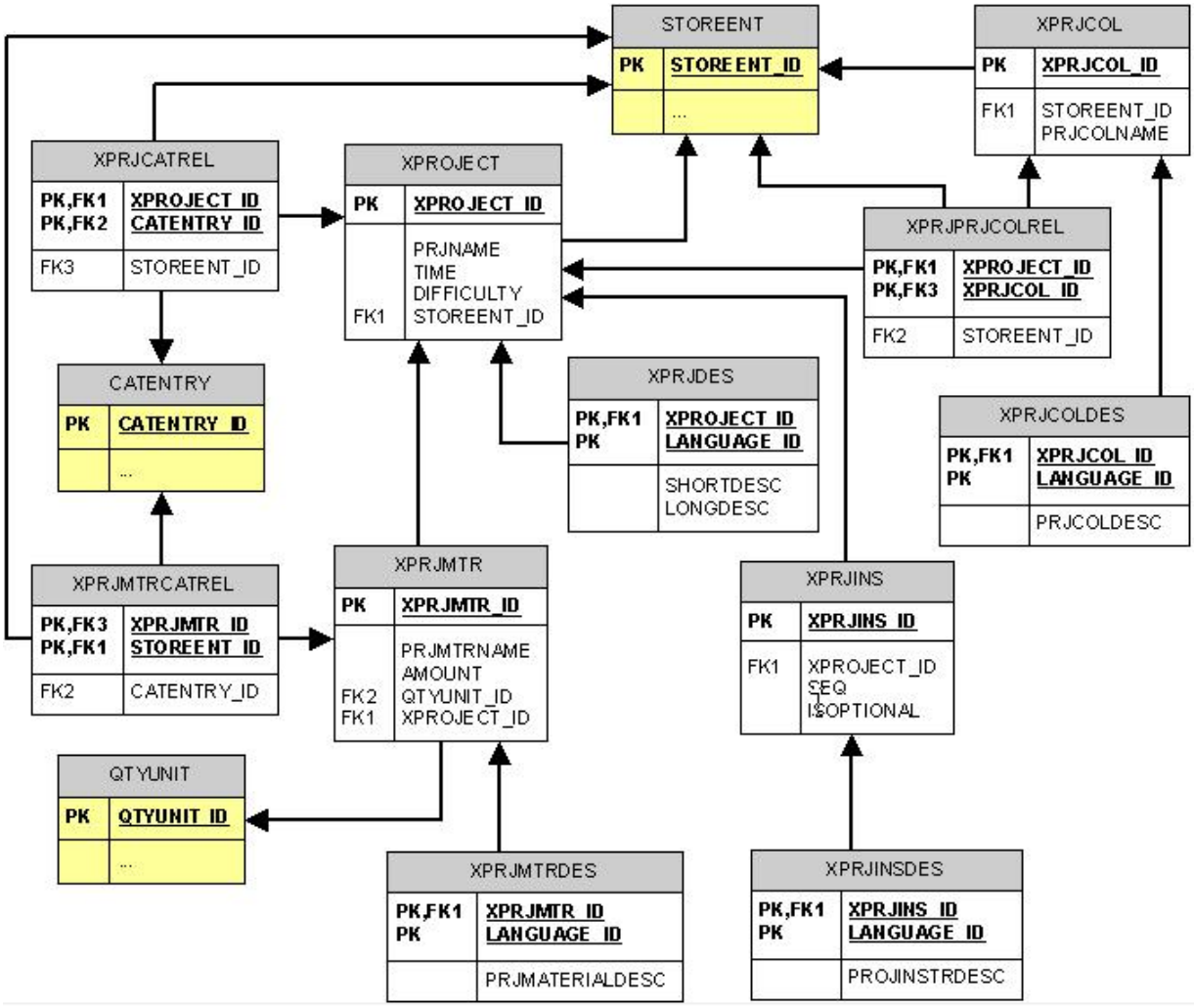
  Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.（不太一样）

  [作业]课文中大数据英文定义的原始出处在哪里？

4. By another definition, “Big data is a large volume unstructured data which cannot be handled by standard database management systems like DBMS, RDBMS or ORDBMS”.
   - **RDBMS**: Relational DataBase Management System 关系型数据库管理系统
   - **ORDBMS**: Object-Relational DataBase Management System 对象关系型数据库管理系统

   [Object-Relational Database] It is similar to a relational database, but with an **object-oriented database model**: **objects**, **classes** and **inheritance** are directly supported in database schemas and in the query language.

schema（/ˈskiːmə/，模式）是数据库的组织和结构，
模式中包含了 schema 对象，可以是
表(table)、
列(column)、
数据类型(data type)、
视图(view)、
存储过程(stored procedures)、
关系(relationships)、
主键(primary key)、
外键(foreign key)等。
数据库模式可以用一个可视化的图来表示，它显示了数据库对象及其相互之间的关系.



5. Volume—The quantity of data that is generated is very important in this context.
   - **Quantity** 数量，总量
   - **Generate** 产生，生成
   - **Context** 环境，语境，上下文（context-free grammar）
     Context-free grammars are named as such because any of the production rules in the grammar can be applied regardless of context—it **does not depend on** any other symbols that may or may not be **around a given symbol** that is having a rule applied to it.

6. This means that the category to which Big Data belongs is also a very essential fact that needs to be known by the data **analysts**.
   - **Analyst** /ˈænəlɪst/ 分析者

7. This refers to the **inconsistency** which can be shown by the data at times, thus **hampering** the process of being able to handle and manage the data effectively.
   - Inconsistency 不一致
   - Hamper /ˈhæmpə(r)/ 妨碍

8. **Veracity**—The quality of the data being captured can vary greatly.【获取的数据，在质量方面表现参差不齐】
   - **Veracity** /vəˈræsəti/ 真实，准确
   - **Vary** /ˈveəri/ 变化

9. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data.
   - **correlated** /'kɒrəleɪtɪd/ 有相互关系的

   correlation between inflation and money supply
   a link between A and B which can be regarded as causality

10. Big data **analytics** consists of 6 Cs in the integrated industry 4.0 and **Cyber Physical Systems** environment.
    - **analytics /ˌænəˈlɪtɪks/** 分析学
    - **Cyber Physical Systems** 信息网络系统

    6 Cs：

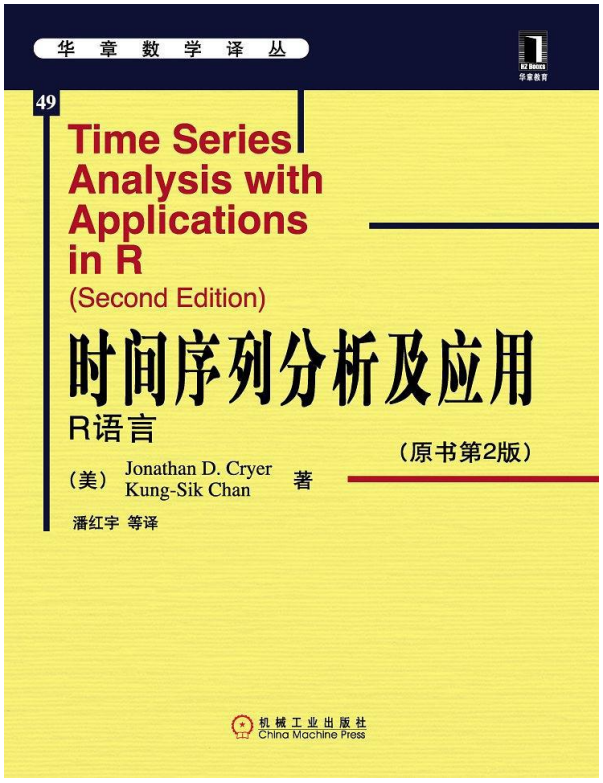
    Connection (sensor and networks) 连接
    Cloud (computing and data on the demand) 云
    Cyber (model and memory) 信息
    Content/Context (meaning and correlation) 内容
    Community (sharing and collaboration) 社区
    Customization(personalization and value) 定制化

11. Big data requires **exceptional** technologies to efficiently process large quantities of data within **tolerable elapsed times**.
    - **exceptional** 特别的
    - **tolerable** 可容忍的
    - **elapsed time** 运行时间，执行时间

12. A 2011 McKinsey report suggests suitable technologies include A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, natural language processing, signal process, simulation, time series analysis and visualization.
    - **McKinsey** 麦肯锡（公司）
    - **A/B 测试:**为 Web 或 App 界面或流程制作两个（A/B）或多个（A/B/n）版本，在同一时间维度，分别让组成成分相同（相似）的访客群组（目标人群）随机访问这些版本，收集各群组的用户体验数据和业务数据，最后分析、评估出最好版本
    - **crowdsourcing** 众包指的是一个公司或机构把过去由员工执行的工作任务，以自由自愿的形式外包给非特定的（而且通常是大量的）大众志愿者的做法
    - **data fusion** 数据融合指整合多个数据源以产生比任何单个数据源提供的更一致，准确和有用的信息的过程

- **genetic algorithms** 遗传算法(搜索+经验)
- **machine learning** 机器学习，主要研究如何让计算机具有能够自我学习的能力

- **natural language processing** 自然语言处理是人工智能和语言学领域的分支学科
  1) 有认知、理解、生成等过程
  2) 认知和理解是让计算机把输入的语言变成有意思的符号和关系，然后根据目的再处理
  3) 生成系统则是把计算机数据转化为自然语言
- **signal process** 信号处理在计算机科学、药物分析、电子学等学科中，指对信号表示、变换、运算等进行处理的过程
- **time series analysis** 时间序列分析



13. Multidimensional big data can also be represented as **tensors**, which can be more efficiently handled by tensor-based computation, such as **multilinear subspace learning**.

- **Tensor** 张量
  1) 一阶张量：一维数组，通常叫作向量（Vector）
  2) 二阶张量：二维数组，通常叫做矩阵（Matrix）
  3) 三阶张量：三维数组
  4) n 阶张量：n 维数组
- **multilinear subspace learning 多线性子空间学习，**通过直接映射高维张量数据到低维空间的一种降维（dimensionality reduction)方法

14. **Massively parallel-processing** 大规模并行处理

15. **Data mining** 数据挖掘，从大量的数据中通过算法搜索隐藏于其中信息的过程

16. **Distributed file systems** 分布式文件系统

17. **DARPA**：美国国防部预研项目局（Defense Advanced Research Projects Agency）

18. The **practitioners** of big data analytics processes are generally **hostile** to slower shared storage, **preferring direct-attached storage (DAS)** in its various forms from **solid state drive (SSD)** to high capacity **SATA** disk buried inside parallel processing nodes.
    - **Practitioner** 从业者
    - **Hostile** /ˈhɒstaɪl/ 敌对的
      be hostile to 对...有敌意
    - **Prefer** ： 偏爱
    - **direct-attached storage (DAS)** 直连存储
      指的是存储设备与主机直连的架构。如主机内部磁盘驱动器和与主机直连的外部存储
    - **solid state drive (SSD)** 固态硬盘
    - **SATA（Serial Advanced Technology Attachment）Disk** 串口硬盘

19. Exabytes 艾字节（EB）2^{60} bytes, or 1,024 petabytes.
20. Alphanumeric /ˌælfənjuː ˈmerɪk/ 含有字母数字的
    (of a character set, code, or file of data) consisting of alphabetical and numerical symbols
    For example, "1a2b3c" is a short string of alphanumeric characters.