深圳技术大学 大数据与互联网学院
SHENZHEN TECHNOLOGY UNIVERSITY College of Big Data and Internet

2023-2024秋季课程：数据科学与大数据导论

Introduction to Data Science and Big data

第三章：大数据处理基础

Chapter 3: Big Data Analytics Fundmentals

曹劲舟 博士 助理教授

深圳技术大学 大数据与互联网学院

caojinzhou@sztu.edu.cn

2023年9月

大数据与互联网学院 曹劲舟 版权所有

Outline

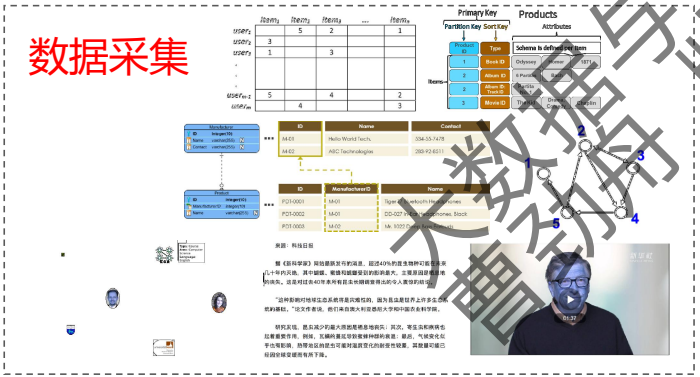
- Data Types and Sources 数据模型
- Data Collection 数据采集
- Data Preprocessing 数据预处理
- Exploratory Data Analysis 数据探索性分析

大数据与互联网学院
曹劲舟 版权所有

数据科学的工作流程

- 三个基本任务
  - 获取原始数据
  - 准备待分析数据
  - 针对特定问题进行数据分析

数据采集
数据准备
数据分析



数据准备

特征				标签
…	…	…	…	1
…	…	…	…	0

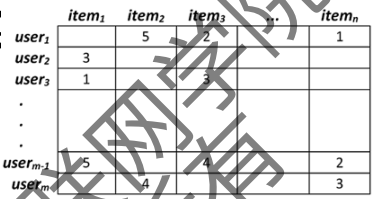
待分析数据

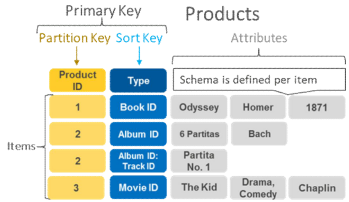
数据分析

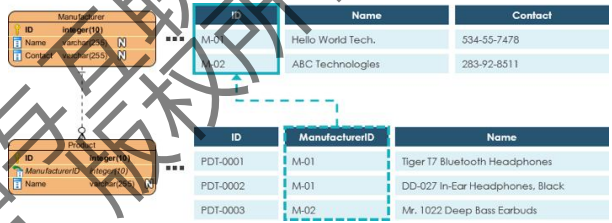
大数据与互联网学院 版权所有 曹剑舟

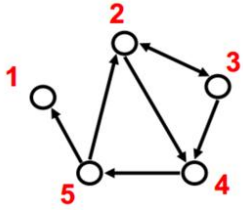
数据类型

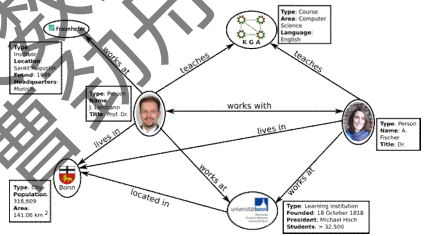
- Variety数据的种类繁多
  - 数组、矩阵
  - 键值对
  - 实体-关系表
  - 时序数据、流数据
  - 图数据
  - 文本数据
  - 多媒体数据
  - …











来源：科技日报

据《新科学家》网站最新发布的消息，超过40%的昆虫物种可能在未来几十年内灭绝，其中蝴蝶、蜜蜂和蜣螂受到的影响最大，主要原因是栖息地的丧失。这是对过去40年来所有昆虫长期调查得出的令人震惊的结论。

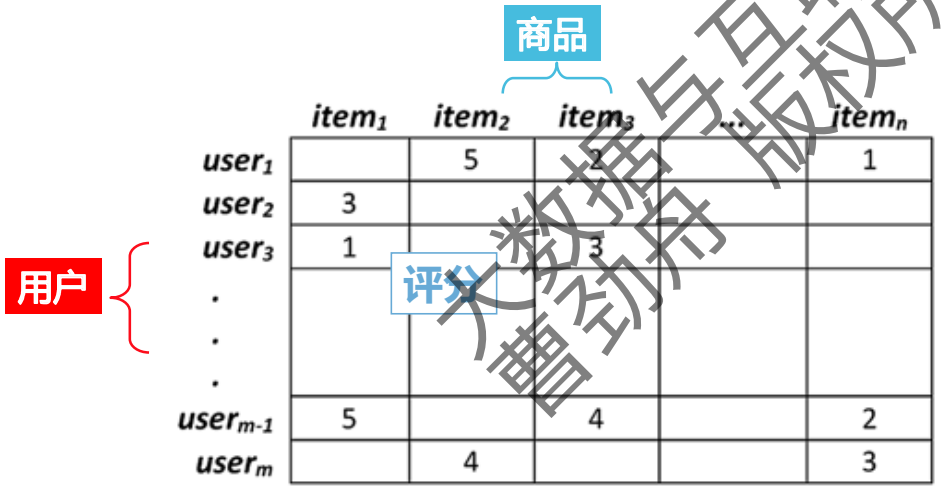
“这种影响对地球生态系统将是灾难性的，因为昆虫是世界上许多生态系统的基础。”论文作者说，他们来自澳大利亚悉尼大学和中国农业科学院。

研究发现，昆虫减少的最大原因是栖息地丧失；其次，寄生虫和疾病也起着重要作用，例如，瓦螨的蔓延导致蜜蜂种群的衰退；最后，气候变化似乎也有影响，热带地区的昆虫可能对温度变化的耐受性较差，其数量可能已经因全球变暖而有所下降。

大数据与互联网学院 曹鸿舟 版权所有

数据模型——数组与矩阵

- 数据项同类型，可以利用下标访问
  - 例子：NumPy的多维数组（ndarray）
  - 例子：推荐系统中的user-item矩阵



大数据与互联网学院 曹劲舟 版权所有

两个用户对三个商品打分：

- u_1 → 1 (5); 3 (2)
- u_2 → 2 (3); 3 (5)

请用NumPy构造矩阵

A. mat = np.array([[5,0,2],[0,3,5]])

B. mat = np.array([[5,np.nan,2],[np.nan,3,5]])

```
import numpy as np
mat = np.array([[5,np.nan,2],[np.nan,3,5]])
mat
```

```
array([[ 5., nan,  2.],
       [nan,  3.,  5.]])
```

数据模型——关系数据（Relational Data）

□简单的关系数据：单表数据

- ■行：表示一条记录（Record）
- ■列：表示一个属性（Attribute）

使用pandas表示单表数据

Team	Win	Loss	Win%
Houston Rockets	20	4	0.83
Golden State Warriors	21	6	0.78
San Antonio Spurs	19	8	0.7
Minnesota Timberwolves	16	11	0.59
Denver Nuggets	14	12	0.54
Portland Trail Blazers	13	12	0.52
New Orleans Pelicans	14	13	0.52
Utah Jazz	13	14	0.48

```
nba_df = pd.DataFrame ({'Team': team_col,
                        'Win': win_col,
                        'Loss': loss_col})
print (nba_df)
```

列标签

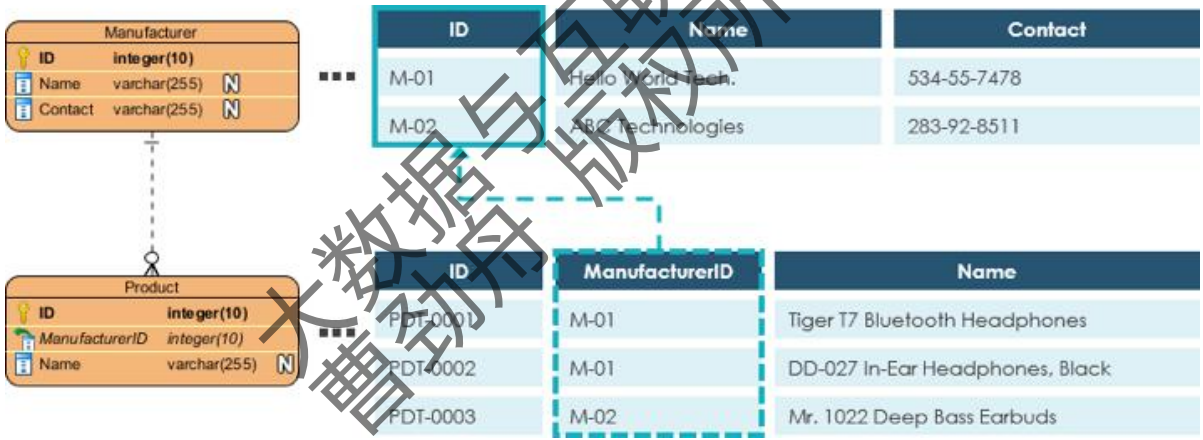
	Team	Win	Loss
0	Houston Rockets	20	4
1	Golden State Warriors	21	6
2	San Antonio Spurs	19	8
3	Minnesota Timberwolves	16	11
4	Denver Nuggets	14	12
5	Portland Trail Blazers	13	12
6	New Orleans Pelicans	14	13
7	Utah Jazz	13	14

行标签

大数据与互联网学院 曹劲骨 版权所有

数据模型——关系数据（Relational Data）

- 关系数据库：将数据表示为多个彼此可关联的表格
  - ER模型组织数据
  - 表格、属性、主外键



大数据与互联网学院 曹劲良 版权所有

数据模型——文本数据

- 自然语言是人们交流信息最为自然的表达方式
  - 互联网网页、论坛评论等
  - 企业文档
  - 聊天记录

来源：科技日报

据《新科学家》网站最新发布的消息，超过40%的昆虫物种可能在未来几十年内灭绝，其中蝴蝶、蜜蜂和蛾鲷受到的影响最大，主要原因是栖息地的丧失。这是对过去40年来所有昆虫长期调查得出的令人震惊的结论。

“这种影响对地球生态系统将是灾难性的，因为昆虫是世界上许多生态系统的基础。”论文作者说，他们来自澳大利亚悉尼大学和中国农业科学院。

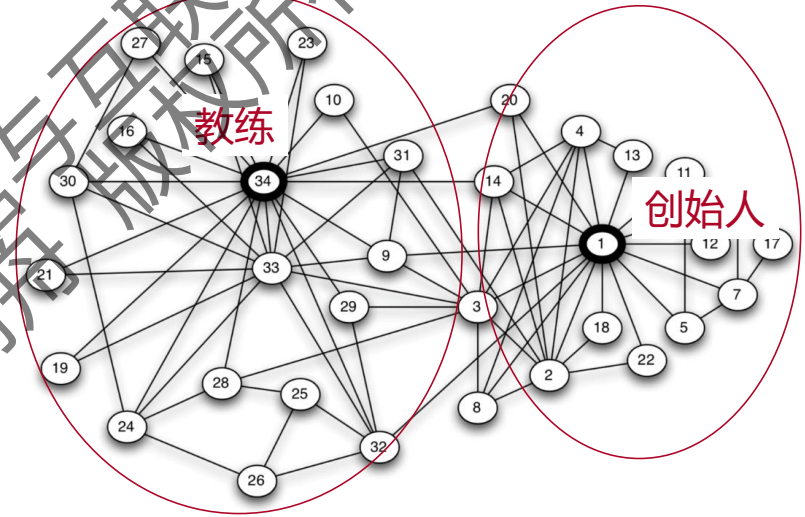
研究发现，昆虫减少的最大原因是栖息地丧失；其次，寄生虫和疾病也起着重要作用，例如，瓦螨的蔓延导致蜜蜂种群的衰退；最后，气候变化似乎也有影响，热带地区的昆虫可能对温度变化的耐受性较差，其数量可能已经因全球变暖而有所下降。

- 非结构化，给文本分析处理带来巨大挑战
- 理解词语、实体、句子、关系等
- 自然语言的语义鸿沟

大数据与互联网学院 版权所有

数据模型——图数据

- 顶点一般表示实体或者属性值
  - 顶点之间的边，表示被连接的两个顶点间的关系
  - 实例
    - 社交网络
    - 知识图谱



请你预言该俱乐部在不久的将来会：

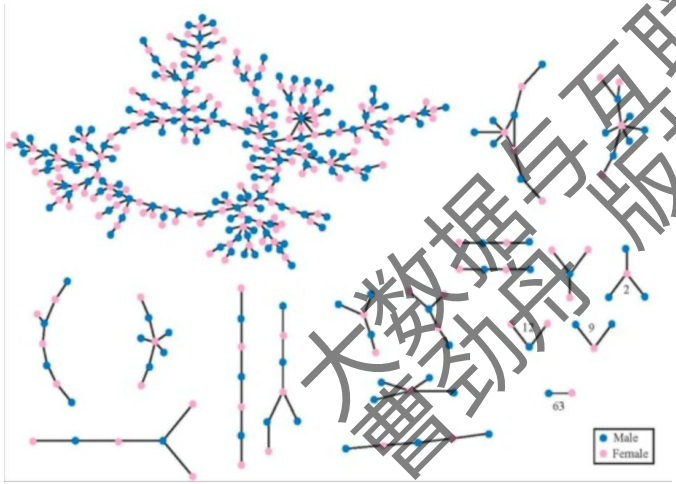
A. **分裂为两个俱乐部**

B. **团结在创始人的周围**

大数据与互联网学院 版权所有

数据模型——图数据

- 图数据：直观地理解群体的行为
  - 例子：高中生恋爱关系图（边代表二人在18个月内恋爱过）



Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks

Peter S. Bearman
Columbia University

James Moody
Ohio State University

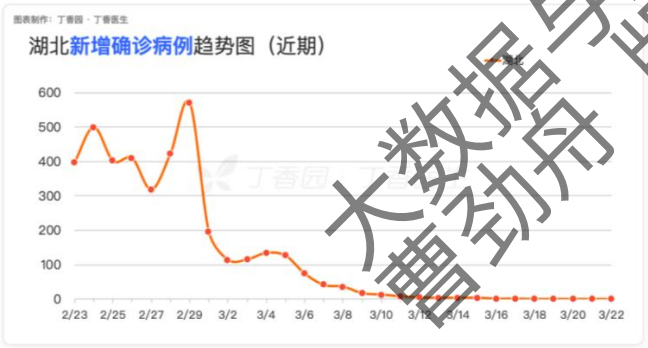
Katherine Stovel
University of Washington

July 2004 · American Journal of Sociology 110(1)
DOI:10.1086/386272

大数据与互联网学院
曹劲昌 版权所有

数据模型——时序数据

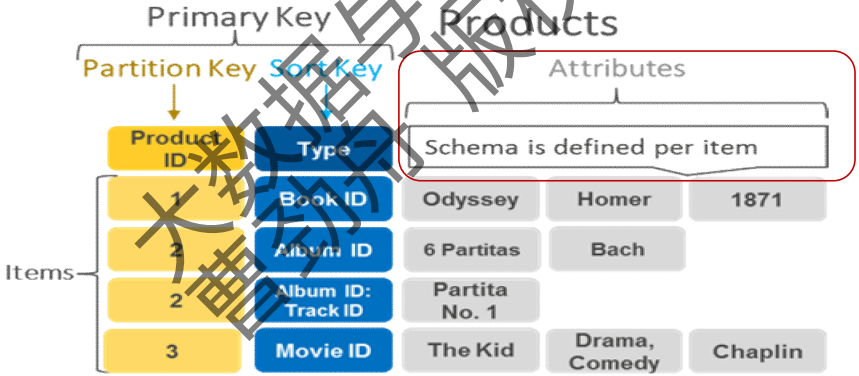
- 随时间不断变化或累积的数据
  - 每个数据项有时间戳
  - 关注一段时间内的数据值变化、关注异常值
  - 新的数据价值更高
  - 多用于监控传感等场景



大数据与互联网学院
曹劲舟 版权所有

数据模型——键值对

- 键值对灵活定义属性，每行可以有多个不同的属性
  - 例子：用户画像
  - 通过键直接访问值
  - 简单的如Hash table，Map等数据结构



大数据与互联网学院 版权所有 盗录必究

数据模型——多媒体数据（非结构型数据）

- 图像、视频、音频等
  - 多种媒体类型的混合
  - 更关注语义
  - 处理复杂，计算代价高
  - 数据量相对更大
  - 在自媒体应用中普遍存在

【简介】比尔及梅琳达·盖茨基金会联席主席比尔·盖茨12日在通过新华社独家发布的视频里说，过去一年里中国在促进全球发展方面继续作出重要贡献。具体聊了哪些贡献？快戳视频看看吧！

ISSN: 1077-3142

Computer Vision and Image Understanding

Editor-in-Chief: N. Paragios

> View Editorial Board

> CiteScore: 8.7 ⓘ Impact Factor: 3.121 ⓘ

大数据与互联网学院
曹劲舟 版权所有

半结构化数据

□半结构化数据

- □一种介于自由文本和结构化文本之间的数据，也是结构化数据的一种形式，但是其结构变化很大。
- □半结构化数据并不符合关系型数据库或其他数据表的形式，但包含相关标记，用来分隔语义元素以及对记录和字段进行分层，
- □因此也被称为自描述的结构。

大数据学院
曹劲舟版权所有

半结构化数据

- 以一个半结构化的数据为例，在招聘季，公司需要收集很多学生简历，每个学生简历都大不相同，有的很简单，比如只包括教育情况；有的则很复杂，比如包括实习情况、发表论文情况、出入境情况、户口迁移情况、党籍情况、技术技能等

大数据与智能学院 重动力 版权所有

半结构化数据

- 将半结构化数据化解为结构化数据——通常是对现有的简历中的信息进行粗略的统计、整理，在总结出简历中信息所有的类别的同时，考虑系统真正关心的信息。

姓名	学校	专业	政治面貌	外语能力	实践经历	备注
王明	中国人民大学	计算机应用技术	无	大学英语六级	XXXXXXXX	…
李丽	中国人民大学	计算机应用技术	中共党员	无	YYYYYYYY	…

大数据与互联网学院
曹劲昂 版权所有

图8.5 使用结构化表格形式保存半结构化数据

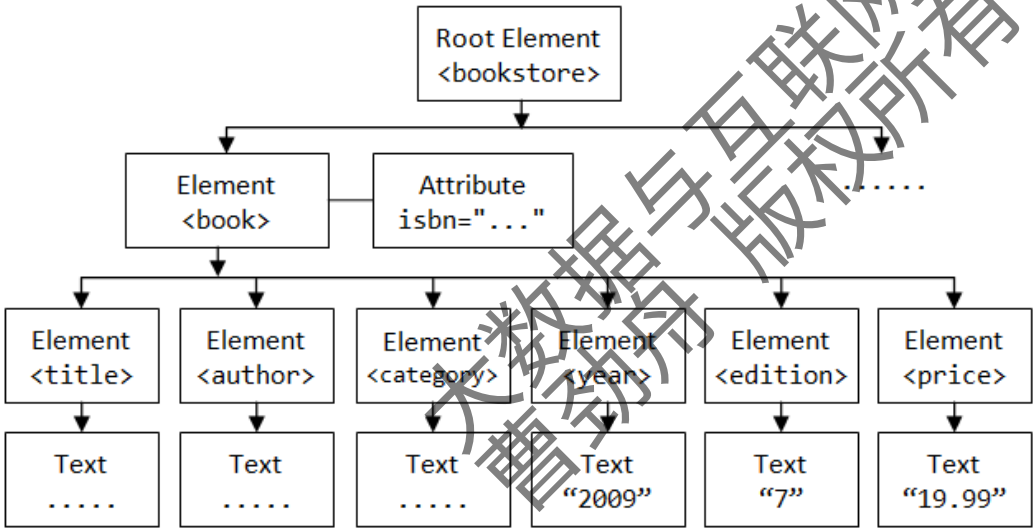
数据模型——XML

- XML（eXtensible Markup Language）是一种可扩展的标记语言，用于描述数据的结构和内容。XML被广泛用于在不同系统之间交换和存储数据。
- XML的数据结构提供了一种通用的方式来存储和传输数据，使得不同系统之间可以共享和解释数据。通过自定义标签和结构，XML可以适应不同的数据需求和语义。

大数据与互联网学院
曹劲丹 版权所有

数据模型——XML

DOM树状结构：




```
<学生>
    <姓名>李明</姓名>
    <学校>中国人民大学 </学校>
    <专业技能>XXXXXXXX</专业技能>
    <外语证书>XXXXXXXX</外语证书>
    <实践经历>XXXXXXXX</实践经历>
    <项目经历>XXXXXXXX</项目经历>
</学生>
```

```
<学生>
    <姓名>李丽</姓名>
    <学校>中国人民大学 </学校>
    <政治面貌>中共党员</政治面貌>
    <联系方式>YYYYYYYY</联系方式>
    <校内奖励>YYYYYYYY</校内奖励>
    <社会实践> YYYYYYYY </社会实践>
    <资格证书> YYYYYYYY </资格证书>
    <兴趣爱好> YYYYYYYY </兴趣爱好>
</student>
```

大数据与互联网学院 版权所有 盗窃必究

数据模型——XML

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- bookstore.xml -->
<bookstore>
    <book ISBN="0123456001">
        <title>Java For Dummies</title>
        <author>Tan Ah Teck</author>
        <category>Programming</category>
        <year>2009</year>
        <edition>7</edition>
        <price>19.99</price>
    </book>
```

文档头

注释

开始元素 “bookstore”

开始元素“book”

开始元素 “title”

结束元素 “title”

结束元素 “book”

大数据与互联网学院 曹劲舟 版权所有

数据模型——JSON

- JSON（JavaScript Object Notation）是一种轻量级的数据交换格式，常用于通过网络传输数据。它以简洁易读的文本形式表示结构化数据，通常由键值对（key-value pairs）组成。
- JSON的灵活性和易用性使其成为在不同系统之间交换数据的常见格式。
- JSON数据结构具有以下特点：
  - 键值对：JSON数据由键值对组成，键是一个字符串，值可以是字符串、数字、布尔值、数组、对象或null。
  - 嵌套结构：JSON允许在值中嵌套其他键值对，从而创建多层次的数据结构。
  - 数组：JSON支持数组，可以将多个值组合在一起，形成有序的列表。
  - 简洁性：JSON采用了一种紧凑的结构表示方式，使得数据更易于阅读和编写，同时也减少了数据传输的大小。
  - 平台无关性：JSON是一种与编程语言无关的数据格式，可以被多种编程语言解析和生成。

大数据与互联网学院 禁止未经授权转载

数据模型——JSON

JSON（Javascript Object Notation）：

```
{
  "name": "John",
  "age": 30,
  "isStudent": false,
  "hobbies": ["reading", "running", "gaming"],
  "address": {
    "street": "123 Main St",
    "city": "New York"
  },
  "scores": [95, 87, 92]
}
```

"name"、"age"、"isStudent"等都是键，对应的值可以是字符串、数字或布尔值。"hobbies"是一个数组，包含多个值。"address"是嵌套的键值对，表示一个地址对象。"scores"也是一个数组，包含多个分数值。

大数据与互联网学院 曹劲舟 版权所有

数据模型——HTML

- HTML（Hypertext Markup Language）是一种用于创建网页和网页应用程序的标记语言。HTML使用标签（tags）来描述文档中的结构和内容。
- HTML数据结构具有以下特点：
  - 标签：HTML使用标签来定义文档中的元素。标签由尖括号（< >）包围，并以起始标签和结束标签的形式出现。起始标签用于定义元素的开始，结束标签用于定义元素的结束。
  - 元素：HTML文档由一个或多个元素组成。元素由起始标签、内容和结束标签组成。标签定义了元素的类型，内容是元素包含的文本或其他元素。
  - 属性：HTML标签可以包含属性，用于提供关于元素的附加信息。属性以键值对的形式出现，位于起始标签中。
  - 嵌套结构：HTML允许元素在其他元素内部嵌套，形成层次结构。嵌套的元素可以形成父子关系，其中父元素包含子元素。
  - 文档结构：HTML文档通常由<html>标签作为根元素开始，包含<head>和<body>等子元素。头部部分（<head>）包含了文档的元信息和引用的外部资源，而主体部分（<body>）包含了显示在浏览器中的实际内容。

禁止用于商业用途

数据模型——HTML

```
<!DOCTYPE html>
<html>
<head>
  <title>My Webpage</title>
</head>
<body>
  <h1>Welcome to My Webpage</h1>
  <p>This is a paragraph of text.</p>
  <ul>
    <li>Item 1</li>
    <li>Item 2</li>
    <li>Item 3</li>
  </ul>
</body>
</html>
```

在上面的例子中，<html>是根元素，<head>和<body>是其子元素。在<head>中，<title>标签定义了文档的标题。在<body>中，<h1>标签定义了一个标题，<p>标签定义了一个段落，<ul>和<li>标签定义了一个无序列表。

HTML的数据结构描述了网页的结构、内容和语义，使得浏览器能够正确地解析和显示网页内容。

大数据与互联网学院
曹劲舟 版权所有

XML、HTML、JSON三者的差异

- 语法和标记：
  - XML使用自定义标签来表示数据结构和内容，标签需要成对出现，包括起始标签和结束标签。例如：<tag>content</tag>。
  - HTML也使用标签来描述文档结构和内容，但它具有一组预定义的标签，用于标识网页元素和样式。例如：<tag>content</tag>。HTML的主要目的是呈现结构化内容。
  - JSON使用键值对表示数据（花括号（{}）包围对象，方括号（[]）表示数组）。例如：{"key": "value"}。JSON的主要目的是数据交换和存储。
- 数据类型：
  - XML不限定数据类型，可以包含文本、数字、布尔值、日期等；可以通过定义数据模型（DTD、XML Schema）来约束数据结构和类型。扩展性高。
  - HTML主要用于描述网页结构和内容，常用于展示文本、图像、链接等网页元素。扩展性相对较低，通常通过CSS和JavaScript来增强交互性和样式。
  - JSON支持基本数据类型（字符串、数字、布尔值、null）以及数组和对象，是一种轻量级的数据交换格式，常用于跨平台数据传输和存储。
- XML适用于复杂的数据结构和语义要求，HTML用于构建网页和呈现内容，JSON用于轻量级的数据交换和存储。选择适当的格式取决于具体的应用场景和需求。

大数据与互联网学院 禁动产 版权所有

数据模型

- 大数据时代：**多模态数据并存**
  - 以关系数据为代表的结构化数据
    - 数据量占比低于20%
    - 数据价值相对高
  - 以文本、图数据为代表的非结构化数据
    - 数据量占比高于80%
    - 数据价值相对低
  - 需要融合结构化数据和非结构化数据
    - 信息抽取
    - 实体链接与数据融合

大数据与互联网学院
曹劲舟 版权所有

数据模型

- 数据模型小结
  - 不同类型的数据与数据模型
  - 人们如何理解与表达数据
  - 计算机如何存储与处理数据

大数据与互联网学院
曹劲舟 版权所有