《计算概论A》课程 计算机基础

信息、数据与计算机

李　戈

北京大学 信息科学技术学院 软件研究所

2010年10月28日

北京大学

信息与数据

- 信息（information）
  - 信息是事物运动的状态与方式;
- 数据（data）
  - 是对客观事物的符号表示，是通过物理观察得来的事实和概念，是关于现实世界中的地方、事件、其他对象或概念的描述。
- 信息不同于数据，
  - 数据是记录信息的一种形式，同样的信息也可以用文字或图像来表述。
  - 通俗讲：有意义的数据是对信息的表达;

计算机中的信息与数据

- 计算机只能处理数据
  - 只有给特定的数据赋予特定的含义，计算机才能具备“处理”信息的能力；

- 数据（data）
  - 在计算机科学中是指所有能输入到计算机并被计算机程序处理的符号介质的总称。

信息的表现形式

- 常见的信息表达的形式
  - 文字
  - 图片
  - 图像
  - 声音
  - 其他…

击剑、柔道、网球、举重、拳击、游泳

Fencing, judo, tennis, weightlifting, boxing, swimming

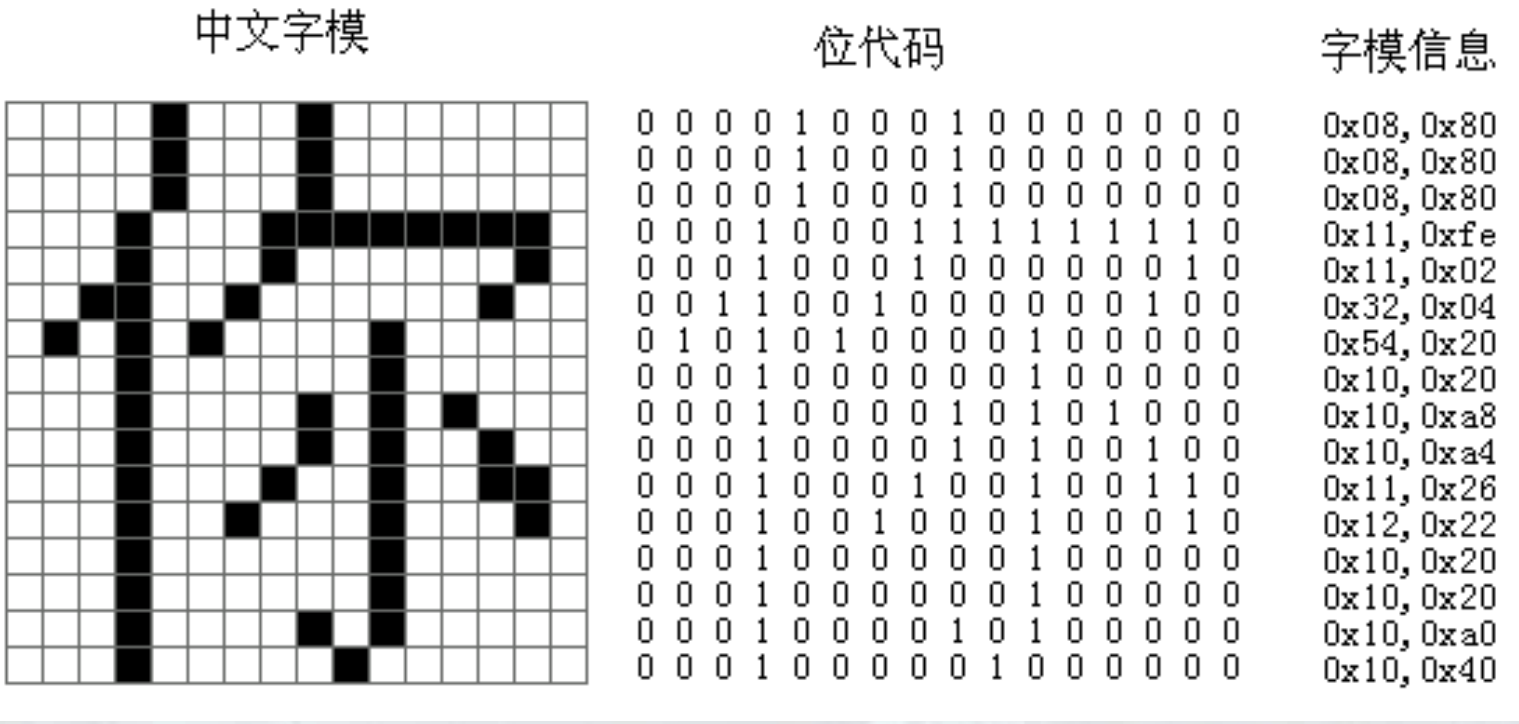
西文字符的编码

- **ASCII**
  - 美国国家标准信息交换码

    American national Standard Code for Information Interchange
  - 每个字符用7位二进制数表示
    - 7位二进制共有128种状态(27= 128)，可表示128个字符，7位编码的取值范围为0000000～1111111
  - 在计算机内，每个字符的ASCII码用1个字节(8位)来存放，标准的ASCII码字符集包括了128个字符。
    - 96个可打印字符
      - 常用字母、数字、标点符号等
    - 32个控制字符

中文字符的编码

- 汉字的字形码
  - 字形码也称字模码，它是标识汉字输出形式的编码。随着汉字字形点阵和格式的不同，汉字字形码也不同。

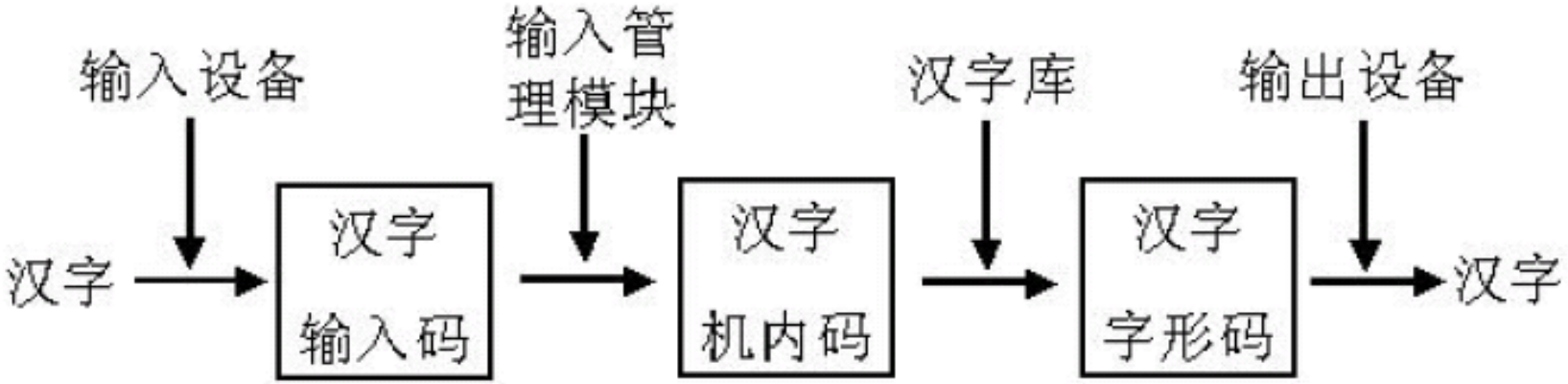


中文字模	位代码	字模信息
	0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0	0x08, 0x80
	0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0	0x08, 0x80
	0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0	0x08, 0x80
	0 0 0 1 0 0 0 1 1 1 1 1 1 1 1 0	0x11, 0xfe
	0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0	0x11, 0x02
	0 0 1 1 0 0 1 0 0 0 0 0 0 1 0 0	0x32, 0x04
	0 1 0 1 0 1 0 0 0 0 1 0 0 0 0 0	0x54, 0x20
	0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0	0x10, 0x20
	0 0 0 1 0 0 0 0 0 1 0 1 0 1 0 0	0x10, 0xa8
	0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0	0x10, 0xa4
	0 0 0 1 0 0 0 1 0 0 1 0 0 1 1 0	0x11, 0x26
	0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0	0x12, 0x22
	0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0	0x10, 0x20
	0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0	0x10, 0x20
	0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0	0x10, 0xa0
	0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0	0x10, 0x40

北京大学

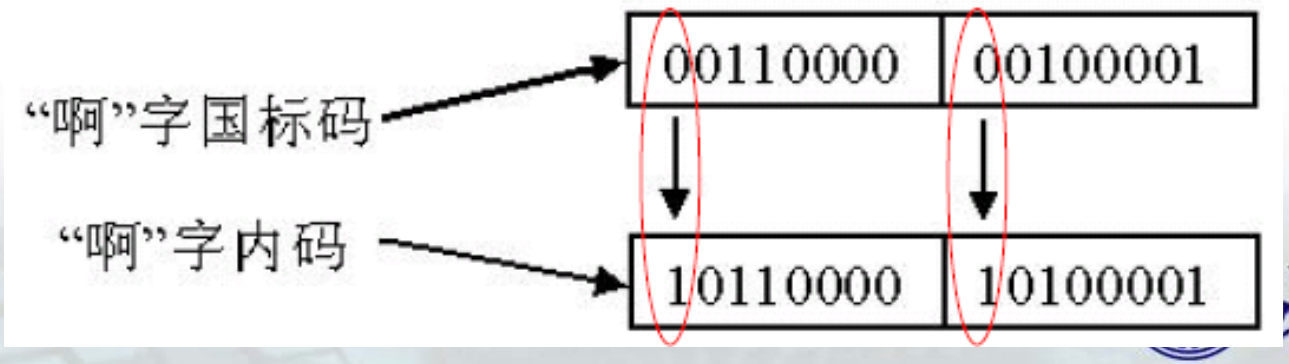
中文字符的输出过程

- 汉字字库
  - 全部汉字字形码的集合叫汉字字库。
  - 可分为软字库和硬字库。
    - 软字库以文件的形式存放在硬盘上;
    - 硬字库则固化在一个单独的存储芯片中，通常称为汉卡。



中文字符的编码

- GB2312-80汉字编码（国标码）
  - 《通用汉字字符集(基本集)及其交换码标准》
  - 共收集汉字7445个，无法用7位/8位表示；
  - 两个字节表示一个字符，每个字节最高位为0
- 汉字内码
  - 汉字在计算机内部存储、处理和传输用的信息编码。
  - 它必须与ASCII码兼容但又不能冲突。
    - 把国标码两个字节的最高位 置 “1”



图像的编码

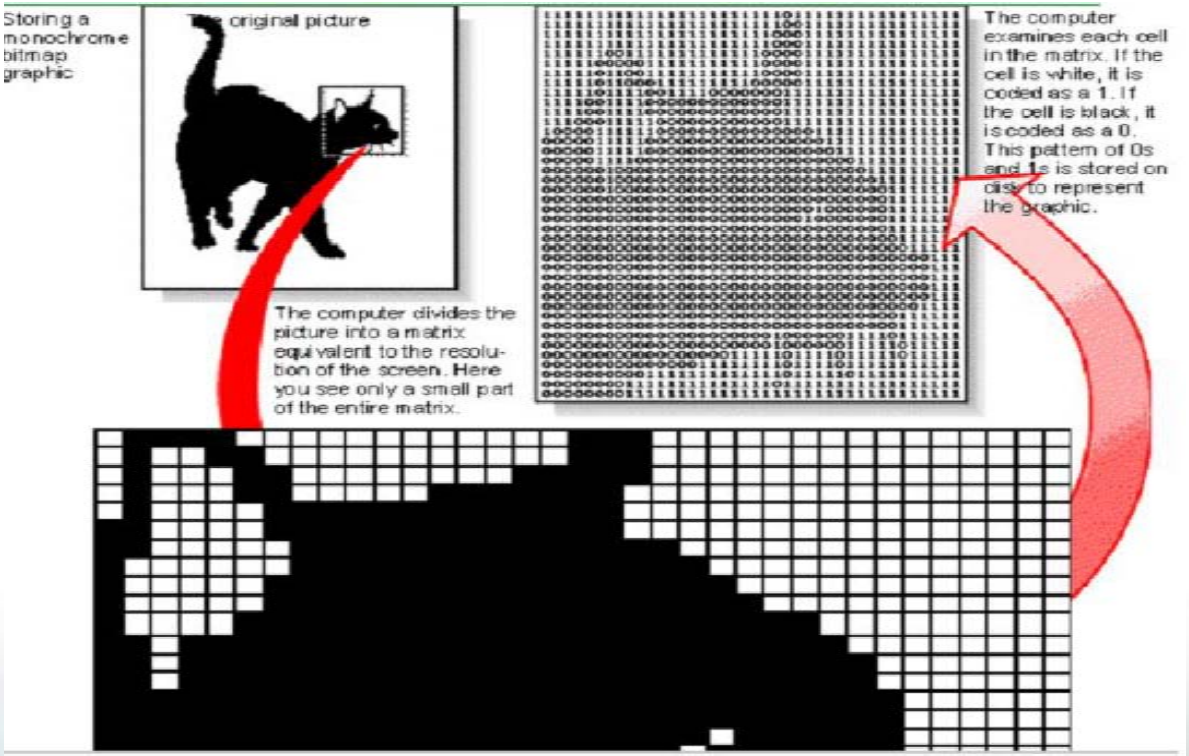
- 所有精美图片、影像都是由0，1代码组成。

图像的编码

- 图像的分类
  - 根据图像生成的原理不同，可以分为
    - 位图
    - 矢量图

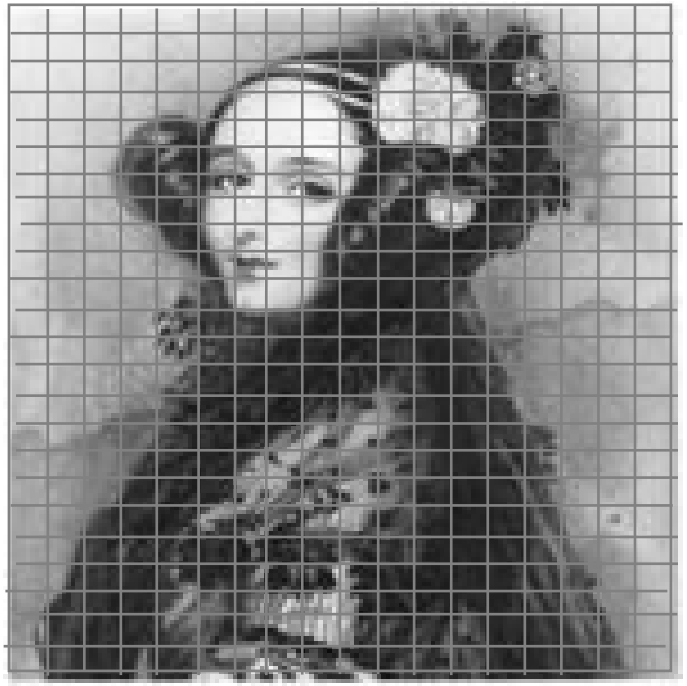
图像的编码

- 位图图像(bitmap)
  - 亦称点阵图像或绘制图像，由称作像素的单个点组成，各个点可以进行不同的排列和染色以构成图样。



图像的编码 —— 灰度

- 灰度图像---有灰度梯度的图像
  - 4个灰度级
    - 每个像素需要2个bit
  - 16个灰度级
    - 每个像素需要4个bit
  - 256个灰度级
    - 每个像素需要8个bit



- 1024*1024的显示器需要显示存储区

1M * 8 bit = 8 Mbit = 1M 字节

北京大学

图像的编码 —— 颜色

- 用数字表示颜色

计算机中颜色显示的差别

	一个象素(即点)用几位二进制表示	可以表示的颜色数
4位图	4位	2^4 =16
8位图	8位	2^8 =256
16位图	16位	2^{16} =65536
24位图(真彩色图)	24位	2^{24} =16777216

北京大学

相关概念

- 分辨率
  - 分辨率是一个表示平面图像精细程度的概念，通常它是以横向和纵向点的数量来衡量的，表示成水平点数×垂直点数的形式。
  - 在一个固定的平面内，分辨率越高，意味着可使用的点数越多，图像越细致。
  - 显示分辨率
    - 显示分辨率是显示器在显示图像时的分辨率，分辨率是用点来衡量的，显示器上这个“点”就是指像素(pixel)。显示分辨率的数值是指整个显示器所有可视面积上水平像素和垂直像素的数量。
    - 例如800×600的分辨率，是指在整个屏幕上水平显示800个像素，垂直显示600个像素。

图像的编码

- 真彩色图像
  - 每个像素用三原色组成，每个原色有256=2^8个级别，因此表示一个像素的颜色需要3*8=24个bit（3个字节）。
  - 对于1024*1024的图像，其显示存储区需要

    1M个像素 * 3字节 = 3M 字节
  - 图形工作站：

    每个原色有65536=2^{16}个级别，其显示存储区需要

    3 * 16 * 1024 * 1024 = 6 M 字节
  - 图像压缩技术非常重要
    - 压缩后的文件格式： .jpg; .tiff; .gif; .fpx; .raw 等

图像的编码

- 矢量图像
  - 矢量图像由一系列可重构的指令构成，计算机存储的是画图的指令而不是像素本身。
  - 矢量图像的优点：
    - 存储空间远远小于位图图像。
    - 绘图方便
    - 修改方便

  - 矢量图像的扩展名通常有*.wmf, .dxf, .mgx, .cgm

视频图像

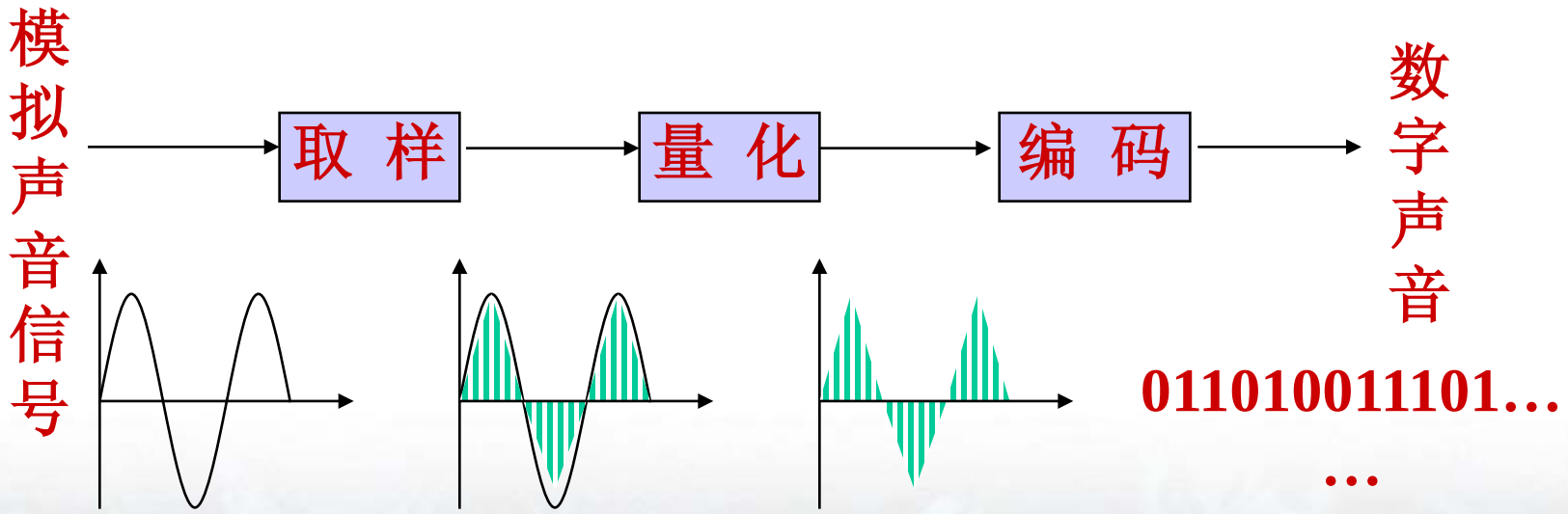
- 视频图像
  - 由一系列帧组成，每一帧都是一幅静止图像，连贯的视频图像每秒需要显示>24帧。
  - 存储空间
    - 设每幅图像均采用24位真彩色，图像大小为640 × 480，每秒播放24帧，连续播放10分钟，则需要

$$640 \times 480 \times 3\text{byte} \times 25\text{帧} \times 10\text{分钟} \times 60\text{秒} = 13824\text{M}$$

  - 视频图像的扩展名通常有 .rm; .mpg等

声音的数字化

- 声音信号的数字化：
  - 将模拟声音信号转换成数字编码形式；
- 声音信号数字化的过程：
  - 取样、量化、编码



北京大学

声音的数字化

- 声音的采样
  - 每秒钟对声音测量的次数，叫采样频率。
    - 采样频率越高，声音效果越好。
  - 采样频率以Hz单位，例如：
    - 每秒中采样1000次，则为1KHz.
- 声音的存储
  - 音乐CD的采样频率为44.1KHz，设有一首歌曲长度为4分钟，双声道采样，每个采样值用2字节记录，则需要的存储空间大小为：

    4分钟 × 60秒 × 44100次 × 2字节 × 2声道 = 36M
  - 声音文件的扩展名： .wav, .midi, .au, .voc, .mp3

数据的压缩

- 压缩
  - 将原始数据重新编码，达到减少文件大小的目的，便于存储和传输。
- 为什么要压缩
  - 采用数字化方法得到的原始文件 数据量巨大;
  - 实时处理数字化视频、音频数据，会严重占用计算机资源;
  - 如果能够实现“实时”压缩、解压可以节省大量计算机资源;

数据的压缩

- 为什么能压缩

  因为，原始信息源数据存在着大量冗余

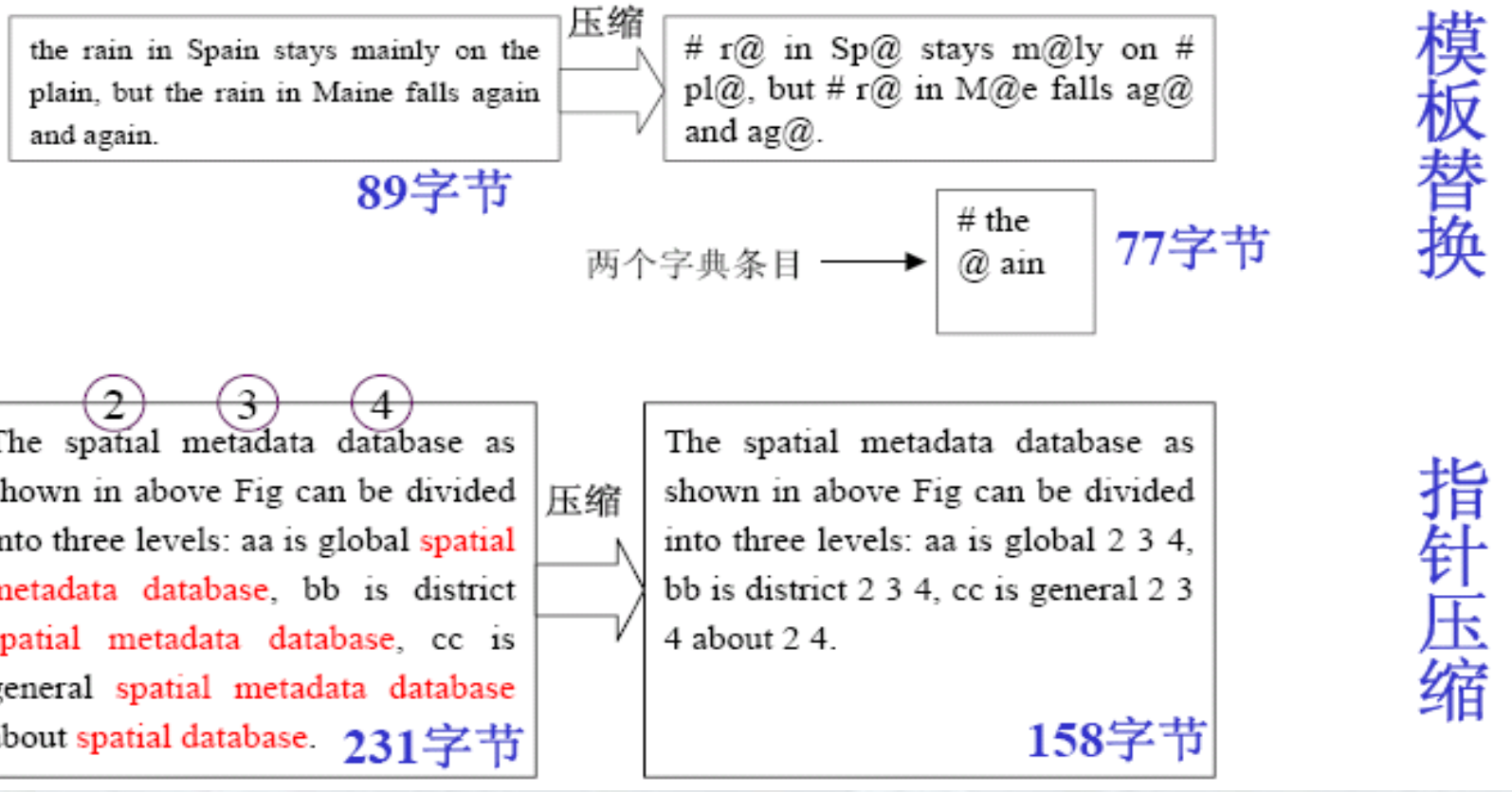
  - 空间重复
    - 单张画面（如静态画面）中的很多部分往往有相同的颜色和图像，这种相关性称为空间相关；
  - 时间重复
    - 在动画或影视图像（动态画面）中，相邻的两帧图像之间产生的变化往往很小，存在大量相同的数据，这种相关性称为时间相关；
  - 人类视觉/听觉器官的不敏感性
    - 对边缘剧变不敏感
    - 对亮度信息敏感而对颜色分辨力不敏感

数据的压缩

- 有损压缩
  - 利用人类视觉/听觉器官的不敏感性，在压缩过程中损失一定的信息，以换取更大压缩比的压缩方式；
  - 常见的声音、图像、视频压缩基本都是有损压缩：mp3; divX; jpeg; rm; rmvb; wma; wmv等；
- 无损压缩
  - 利用数据的统计冗余进行压缩，可完全恢复原始数据而不引起任何失真的压缩方式；
  - 压缩率是受到数据统计冗余度的限制，一般为2:1到5:1，
  - 适用于文本数据，程序和特殊应用场合的图像数据(如指纹图像,医学图像等)的压缩.

数据压缩

■ 文本数据压缩



数据压缩

- 常见的图像压缩格式
  - JPEG图像格式：扩展名是JPG，其全称为Joint Photograhic Experts Group。它利用一种失真式的图像压缩方式将图像压缩在很小的储存空间中，其压缩比率通常在10:1～40:1之间。
  - TIFF图像格式：扩展名是TIF，全名是Tagged Image File Format。它是一种非失真的压缩格式(最高也只能做到2～3倍的压缩比)能保持原有图像的颜色及层次，但占用空间却很大。
  - GIF图像格式：扩展名是GIF。它在压缩过程中，图像的像素资料不会被丢失，然而丢失的却是图像的色彩。GIF格式最多只能储存256色。

数据压缩

- 常见的图像压缩格式

  MPEG（Moving Pictures Experts Group）动态图像专家组 的系列标准:

  - MPEG-1标准（1992年）用于数据传输速率为1.5Mb/s的数字存贮媒体活动图像及其伴音的压缩标准。
  - MPEG-2标准（1994年）针对高清晰度电视（HDTV）的视频及伴音信号的压缩标准;
  - MPEG-4标准（1998年）用于超低速传输率运动图像和语言的压缩标准;
  - MPEG-7是“多媒体内容描述接口标准”。

数据压缩

- 常见的音频压缩格式
  - **MPEG-1 Audio Layer 3 （MP3）**
    - 有损压缩方式，它丢弃掉音频数据中对人类听觉不重要的数据，从而达到了小得多的文件大小。
    - 在MP3中使用了许多技术其中包括心理声学以确定音频的哪一部分可以丢弃。
  - **WAV**
    - Microsoft公司开发的一种WAV声音文件格式
    - 它合RIFF (Resource Interchange File Format)文件规范，用于保存Windows平台的音频信息资源。

数据的管理

- 数据库（database）
  - 长期储存在计算机中，有组织、可共享的数据集合;
  - 按组织方式可分为:
    - 层次数据库、网状数据库、关系数据库、面向对象数据库
- 数据库管理系统(DBMS)
  - 辅助对数据库进行管理的软件系统;
  - 常见的DBMS:
    - Oracle, SQL Server, MySQL, DB2… …

数据的管理

- 关系型数据库
  - 以行和列的形式存储数据，一系列的行和列被称为表，一组表组成了数据库。

学生登记表

学　号	姓　名	年　令	性　别	系　名	年　级
95004	王小明	19	女	社会学	95
95006	黄大鹏	20	男	商品学	95
95008	张文斌	18	女	法律学	95
…	…	…	…	…	…

数据的管理

- SQL

  结构化查询语言(Structured Query Language)

  - SQL语言包含4个部分:
    - 数据查询语言（SELECT语句）

      ```
      SELECT customer_id, first_name FROM
      Customer_Data
      WHERE first_name = "Frankie"
      ```

    - 数据操纵语言（INSERT, UPDATE, DELETE语句）
    - 数据定义语言（如CREATE, DROP等语句）
    - 数据控制语言（如COMMIT, ROLLBACK等语句）

好好想想，有没有问题?

谢　谢　！

北京大学