



# 多媒体信息表示

Representations for Multimedia Information



俞俊、高飞、谭敏、余宙、匡振中

{yujun, gaofei, tanmin, yuz, zzkuang}@hdu.edu.cn

<http://mil.hdu.edu.cn>

# 学习目标

- **视觉信息表示**

- 能够正确说出以下术语（概念）的含义：像素、分辨率、颜色通道；
- 能够简述三原色原理的内容及其生理基础；
- 给定一幅RGB彩色图像，能够将其转换为灰度图像；

- **自然语言**

- 能够正确说出以下术语（概念）的含义：**字典、词语、向量**；
- 给定一段文本和一组字典，能够计算BoW向量；
- 了解几类典型的文本表示模型

# CONTENTS

机器学习基本概念（回顾）

视觉信息

自然语言

小结

# CONTENTS

机器学习基本概念（回顾）

视觉信息

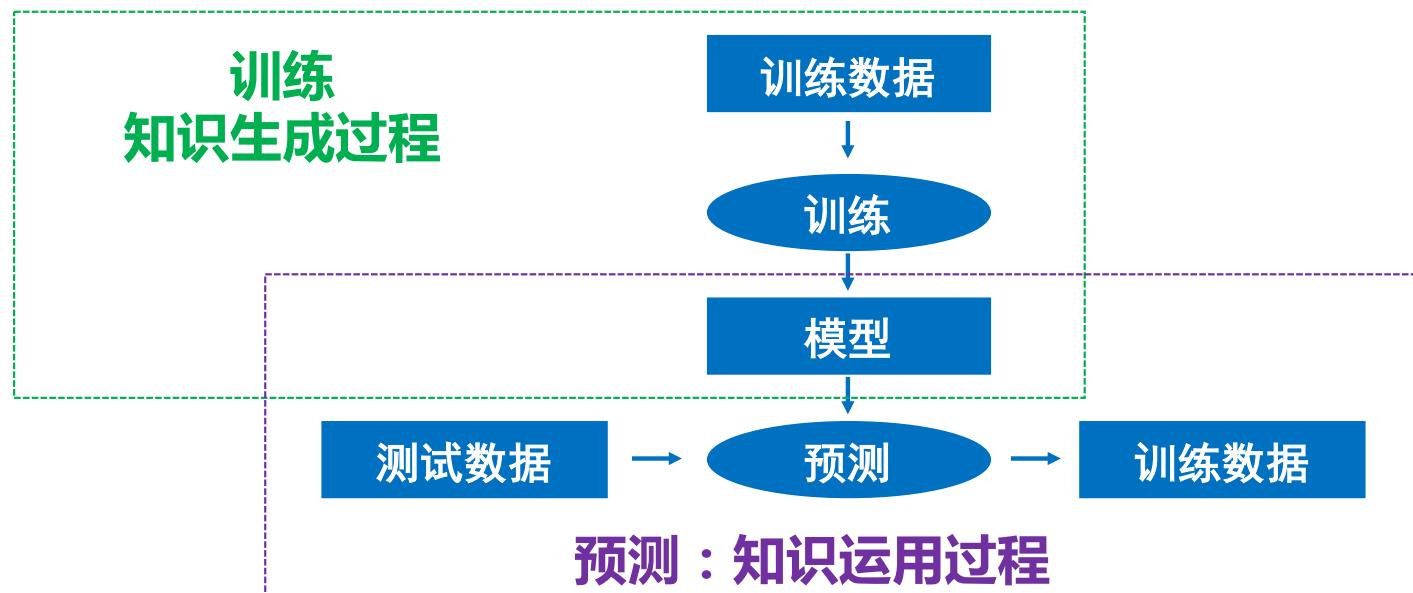
自然语言

小结

# 机器学习基本概念（回顾）

## • 问题1：机器学习定义是什么？并举例说明。

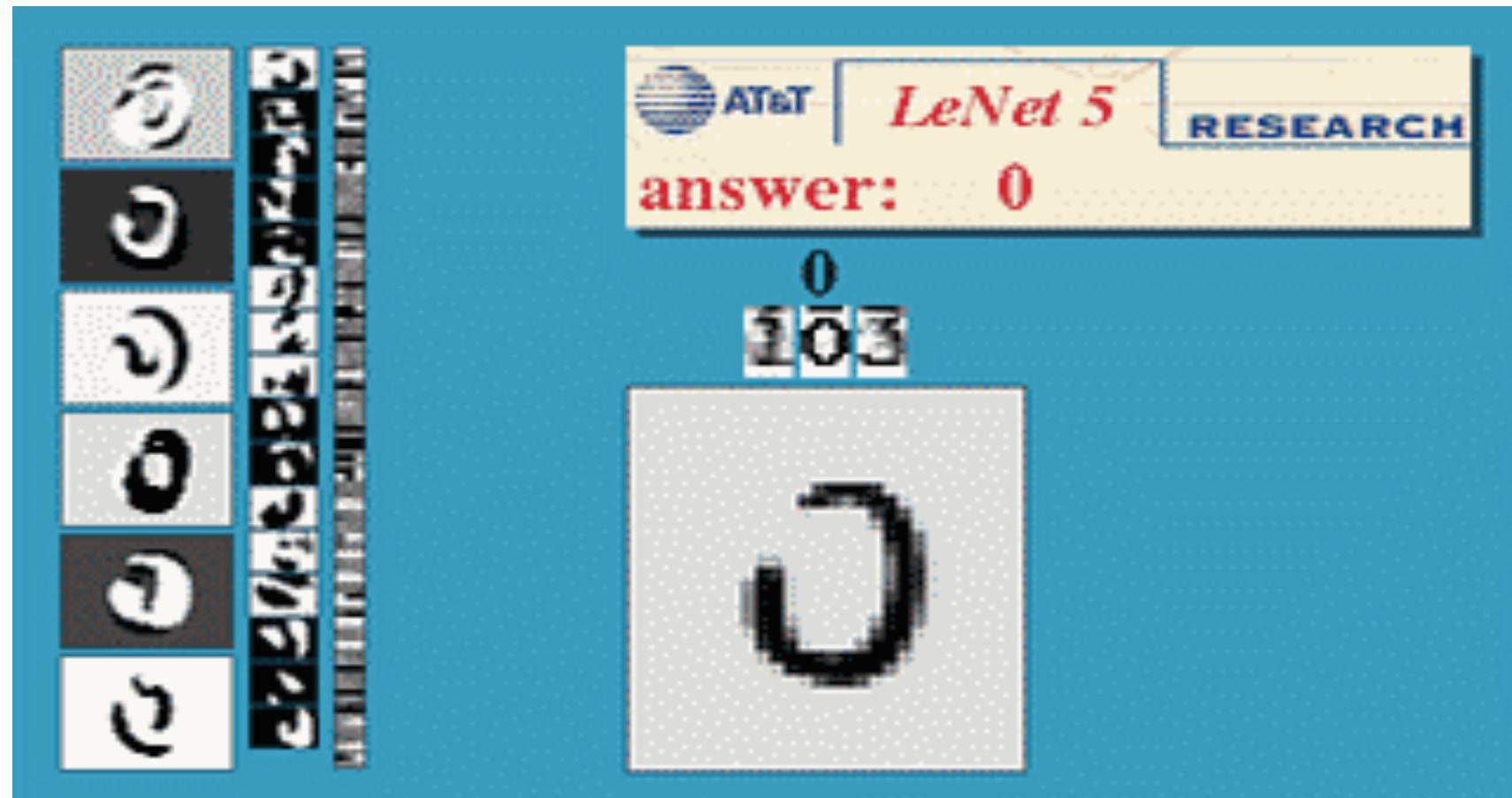
- Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
- 关键词：任务T，经验E，性能指标 P



# 机器学习基本概念（回顾）

- 问题2：解释以下术语，并举例说明。

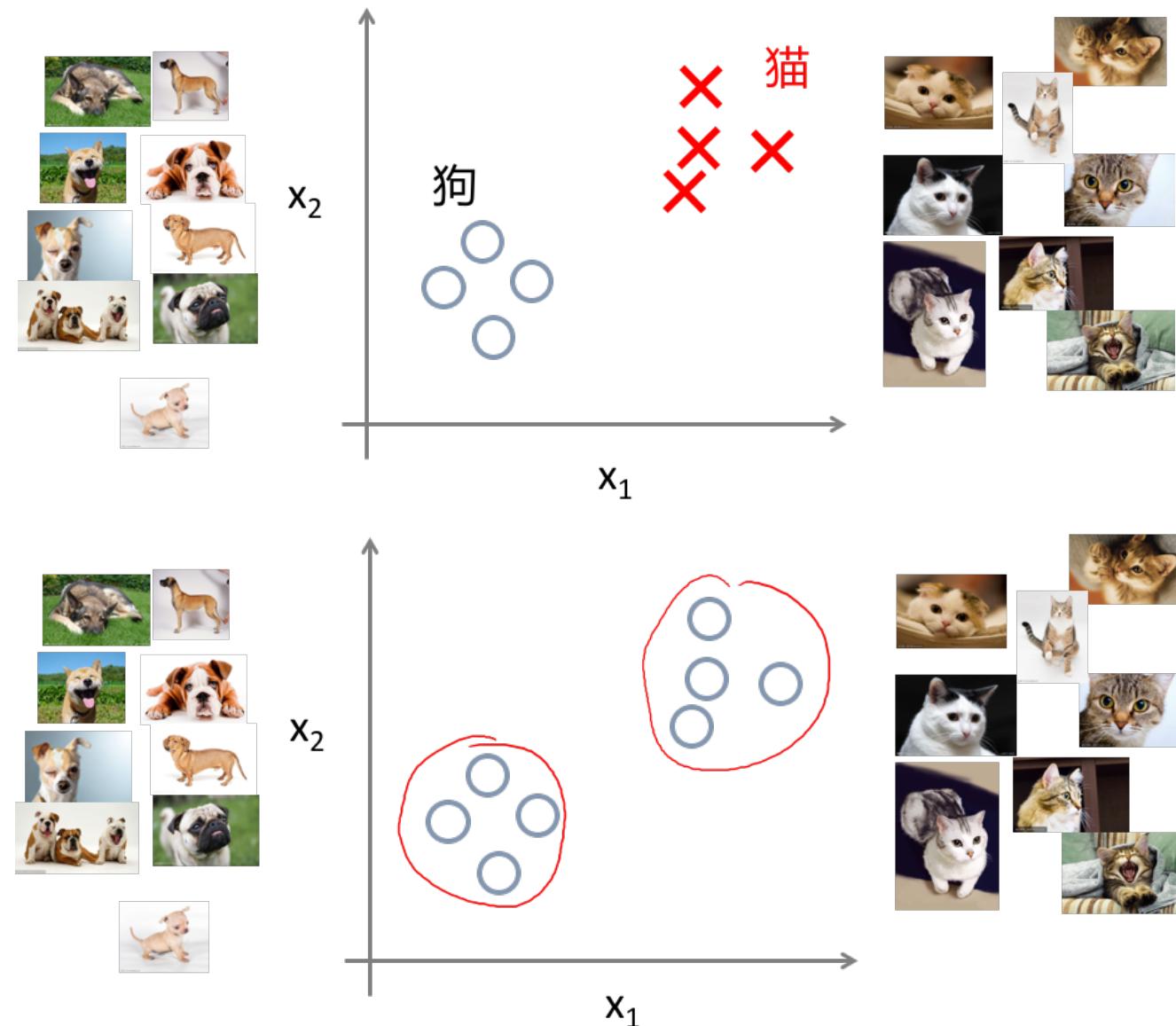
- 特征
- 训练
- 测试
- 分类
- 回归
- 排序



# 机器学习基本概念（回顾）

- 问题3：解释以下术语，并举例说明。

- 有监督学习
- 无监督学习
- 半监督学习



# CONTENTS

机器学习基本概念（回顾）

视觉信息

自然语言

小结

# CONTENTS

背景及意义

人怎么“看”？

机器怎么“看”？

小结（视觉信息）

# 背景及意义

- 视觉信息（图像、视频等）数据极具增长

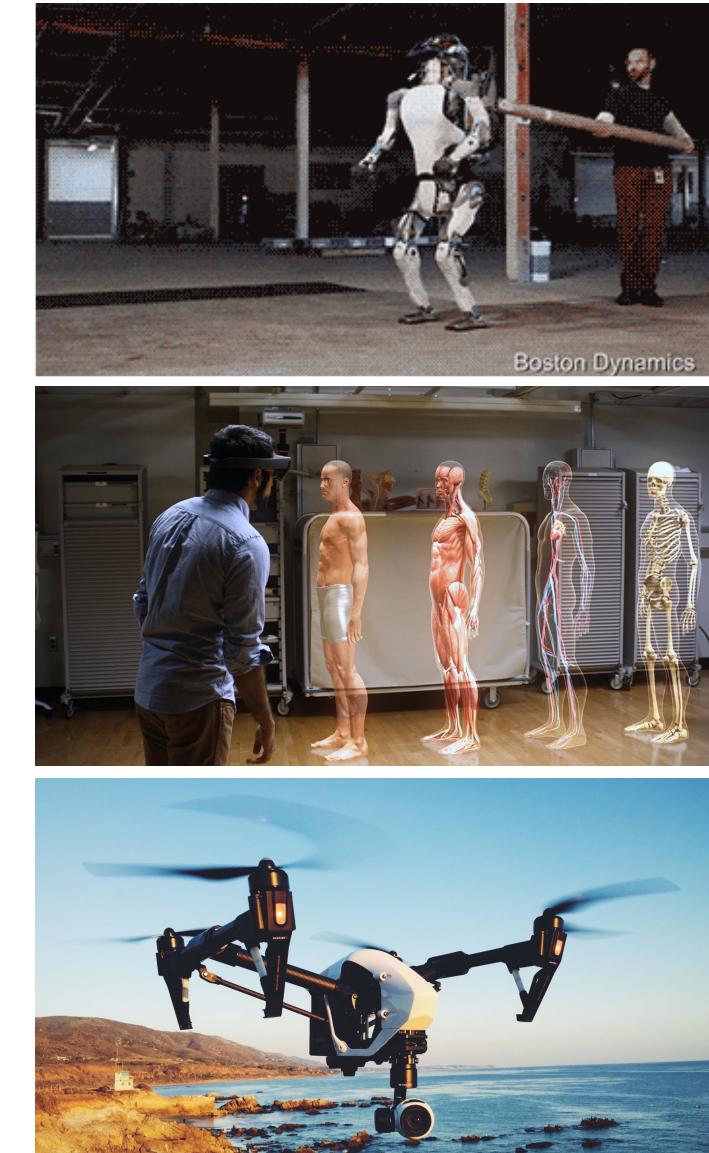
- 设备：相机、手机、平板电脑 ...
- 应用：社交网站、APP、直播、视频娱乐、 ...



# 背景及意义

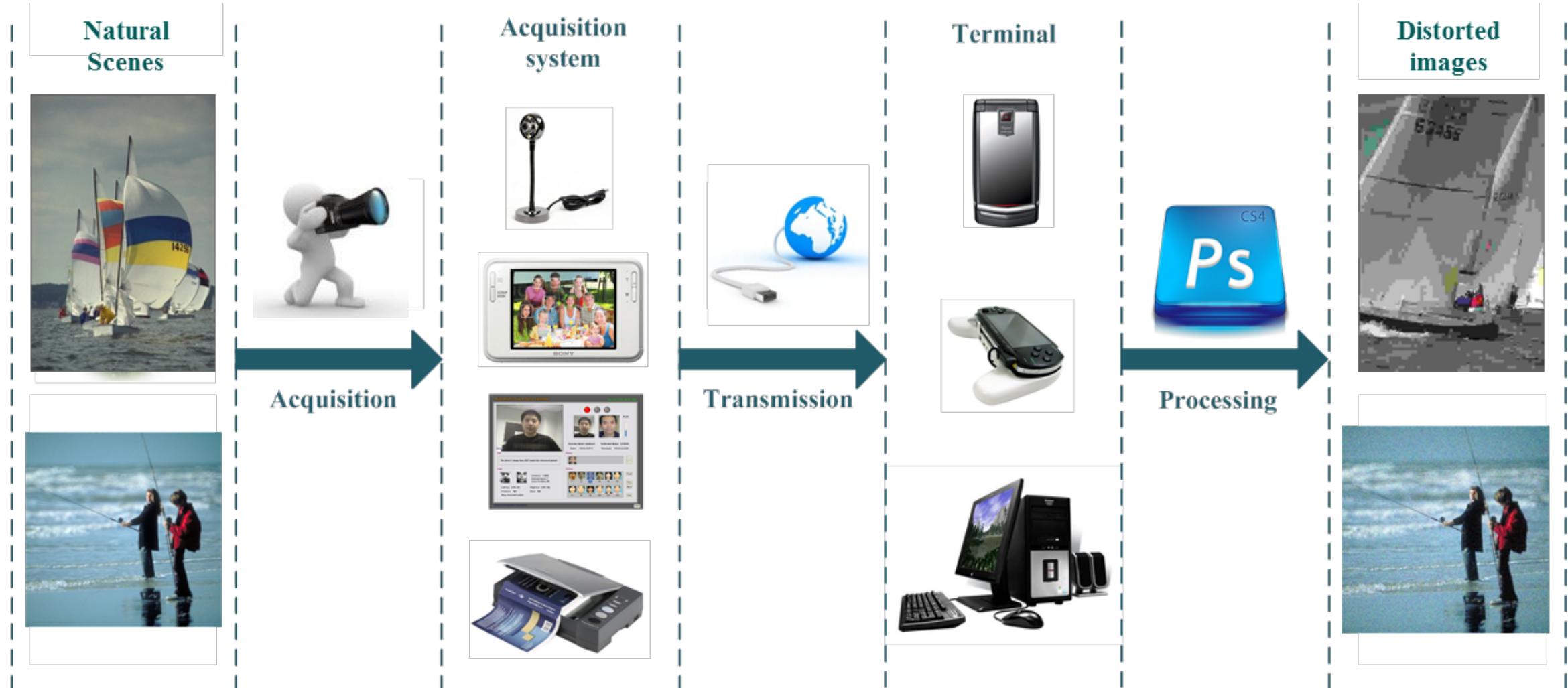
## • 应用领域

- 安防监控
- 图像搜索
- 工业视觉
- 人机交互
- 视觉导航
- 虚拟现实
- 生物医学
- 遥感测绘
- ...



# 背景及意义

- 视觉信息处理过程：获取、压缩、传输、重建、处理 ...



# 计算机视觉研究课题：人脸识别

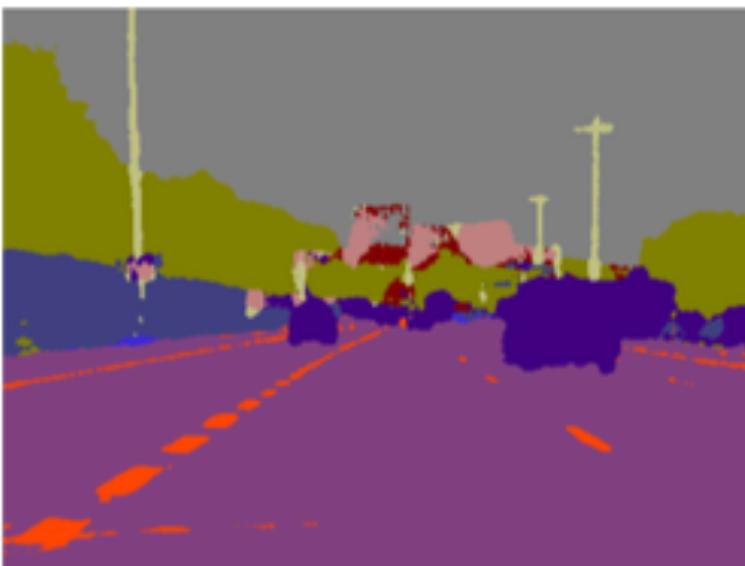
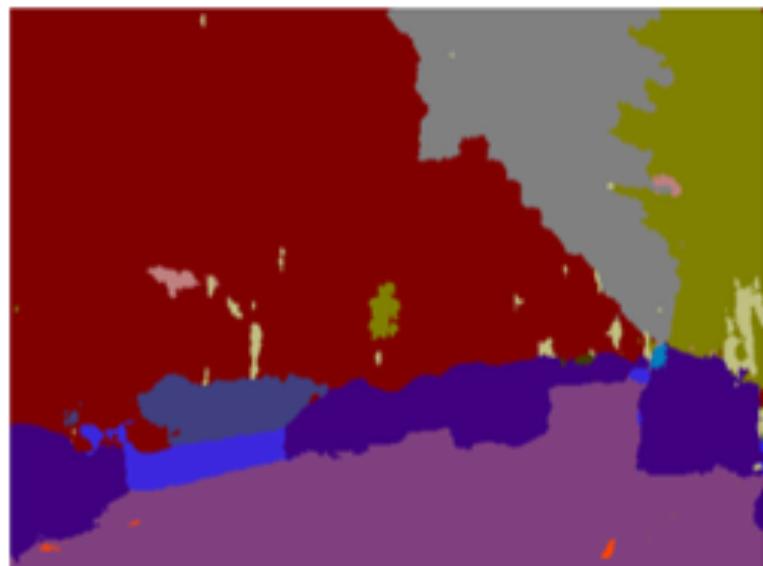


Probe Set



Gallery (at million scale)

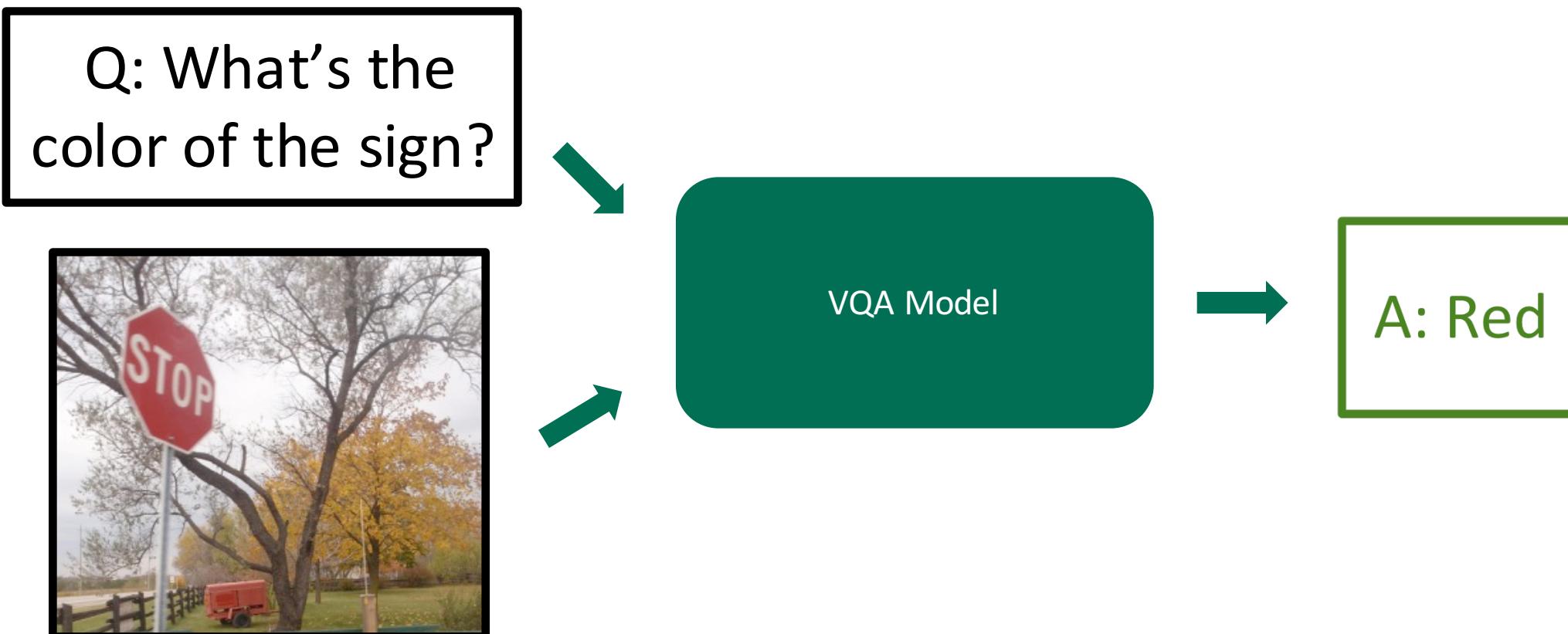
# 计算机视觉研究课题：图像分割



# 计算机视觉研究课题：视觉内容自动问答

## • 视觉问答(Visual Question Answering, VQA)问题描述：

- 给定任意图片，用户使用自然语言一个图像中内容相关的问题，算法自动给出自然语言描述的答案



# CONTENTS

背景及意义

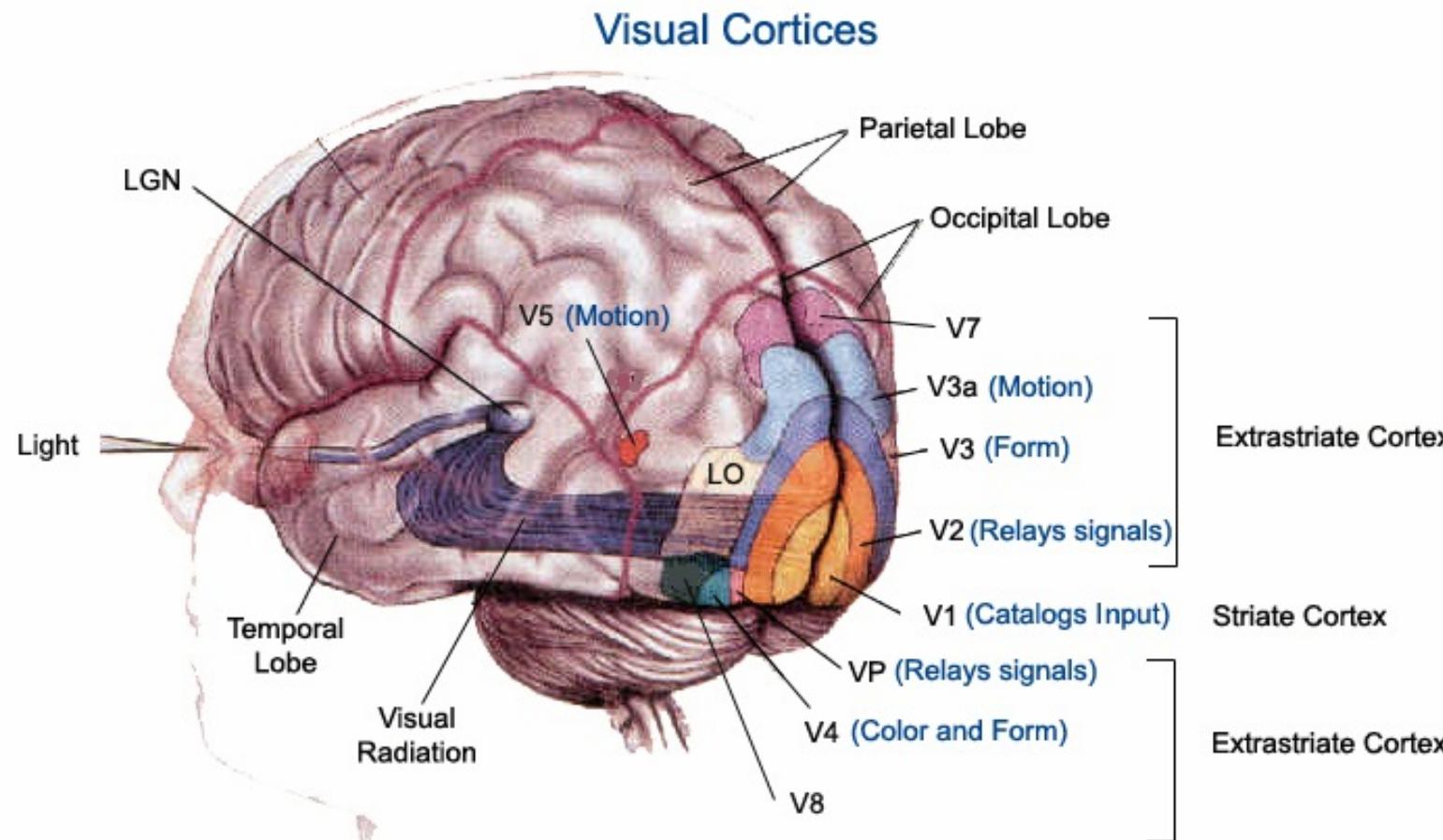
人怎么“看”？

机器怎么“看”？

小结（视觉信息）

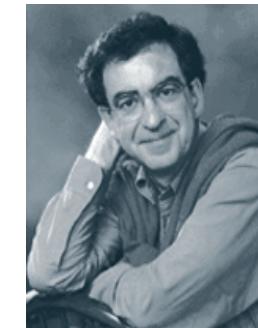
# 人怎么“看”？

## • 人类视觉系统 ( Human Visual System, HVS )



**Laurent Itti: GIST 模型**

Professor of computer science,  
psychology and neuroscience  
University of Southern  
California



**Tomaso Poggio: HMAX 模型**

Eugene McDermott Professor in the  
Brain Sciences and Human Behavior  
Director of the Center for Biological  
and Computational Learning at MIT

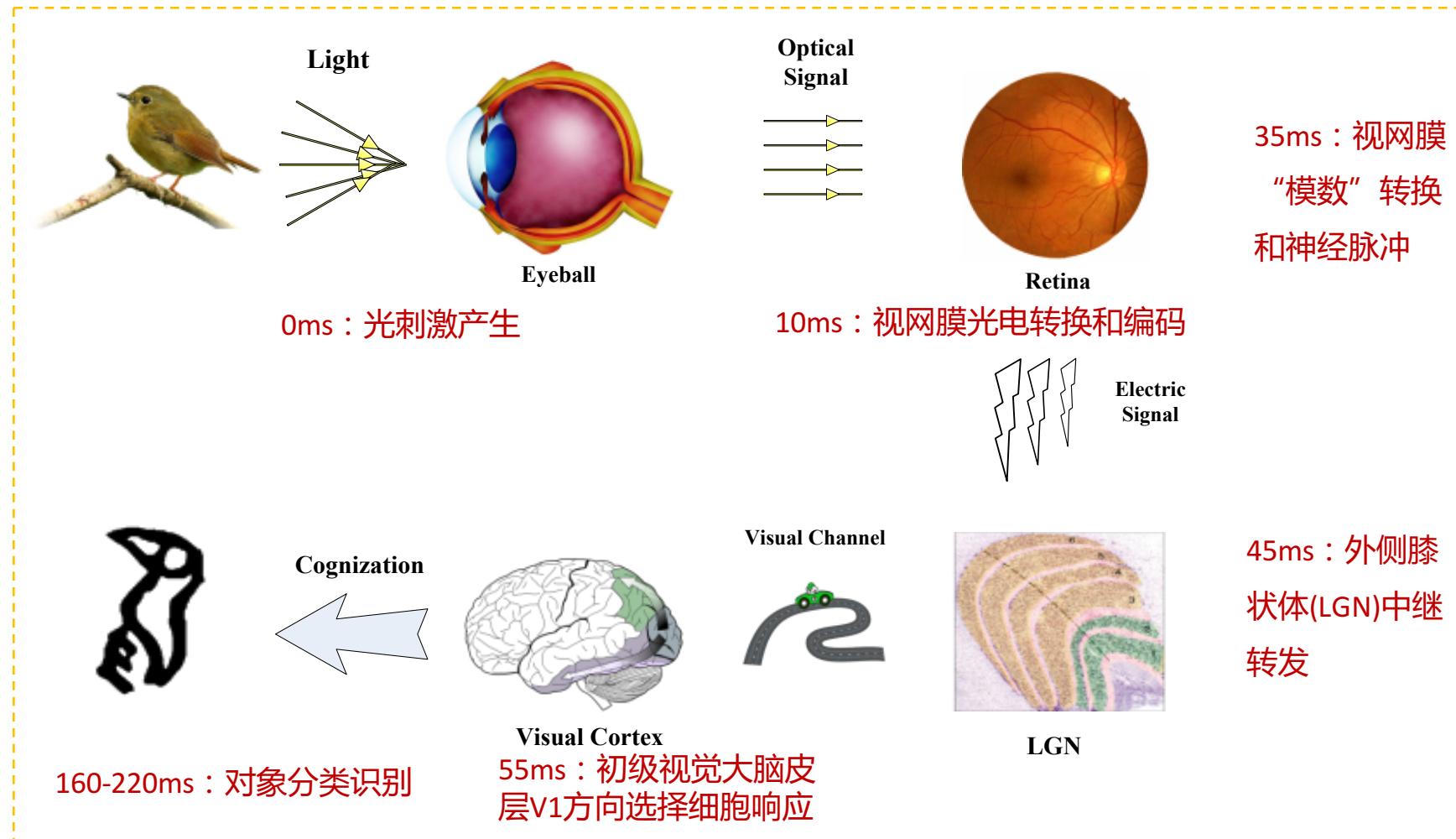
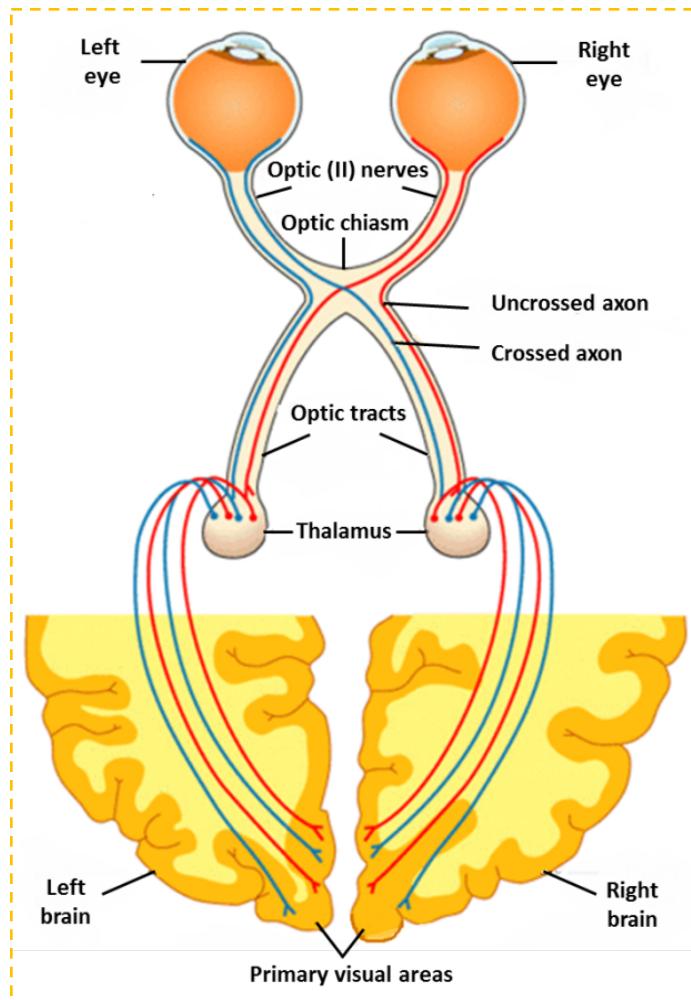
# 人类视觉系统

## • Human Visual System

[ Watson, Digital Images and Human Vision, 1993 ]

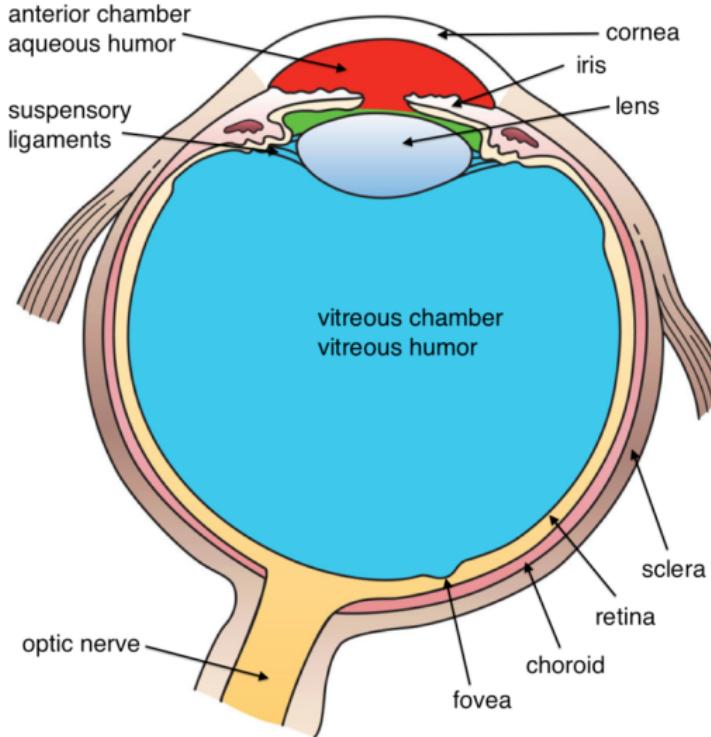
[ Wandell, Foundations of vision, 1995 ]

[ Marr, Vision 1982 ]

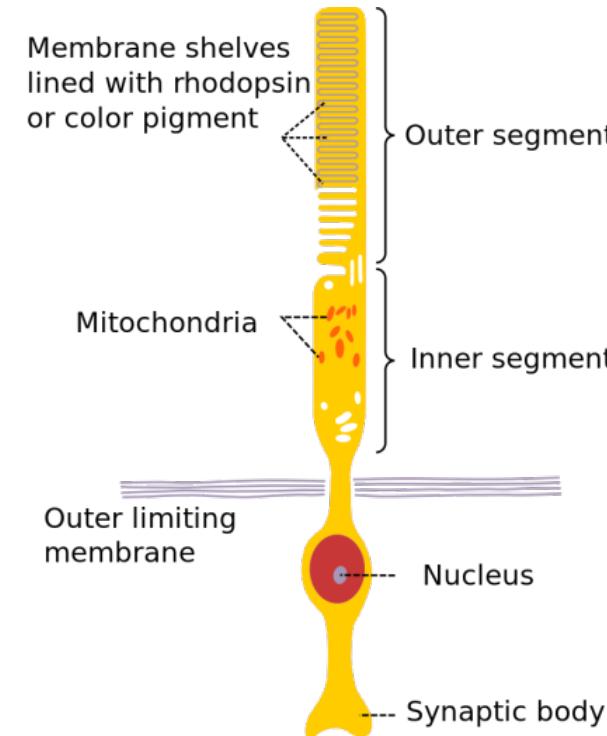


# 人类视觉系统

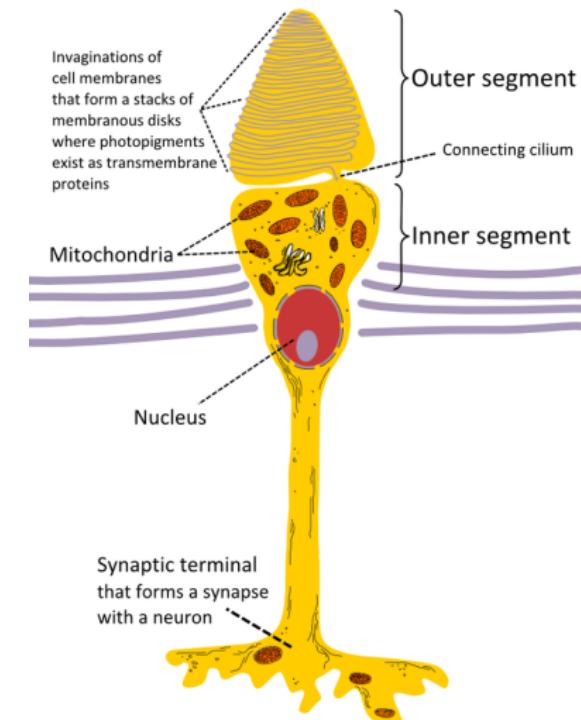
## • 视网膜中的视觉感知器：柱状细胞 Rods vs. 锥状细胞Cones



(a) 视网膜 Retina



(b) Rods(柱状细胞)



(c) Cones (锥状细胞)

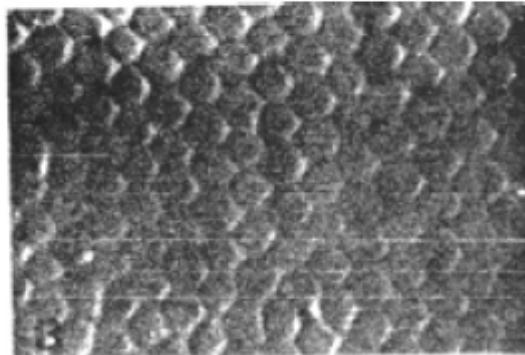
- 锥状细胞对于亮度不敏感，主要在高亮度的时候工作，而柱状细胞可以在亮度较低时工作；
- 锥状细胞可以感知图像中的细节信息以及快速的变化，因为其对于刺激的响应比柱状细胞快。

# 人类视觉系统

## • 人眼感知阵列

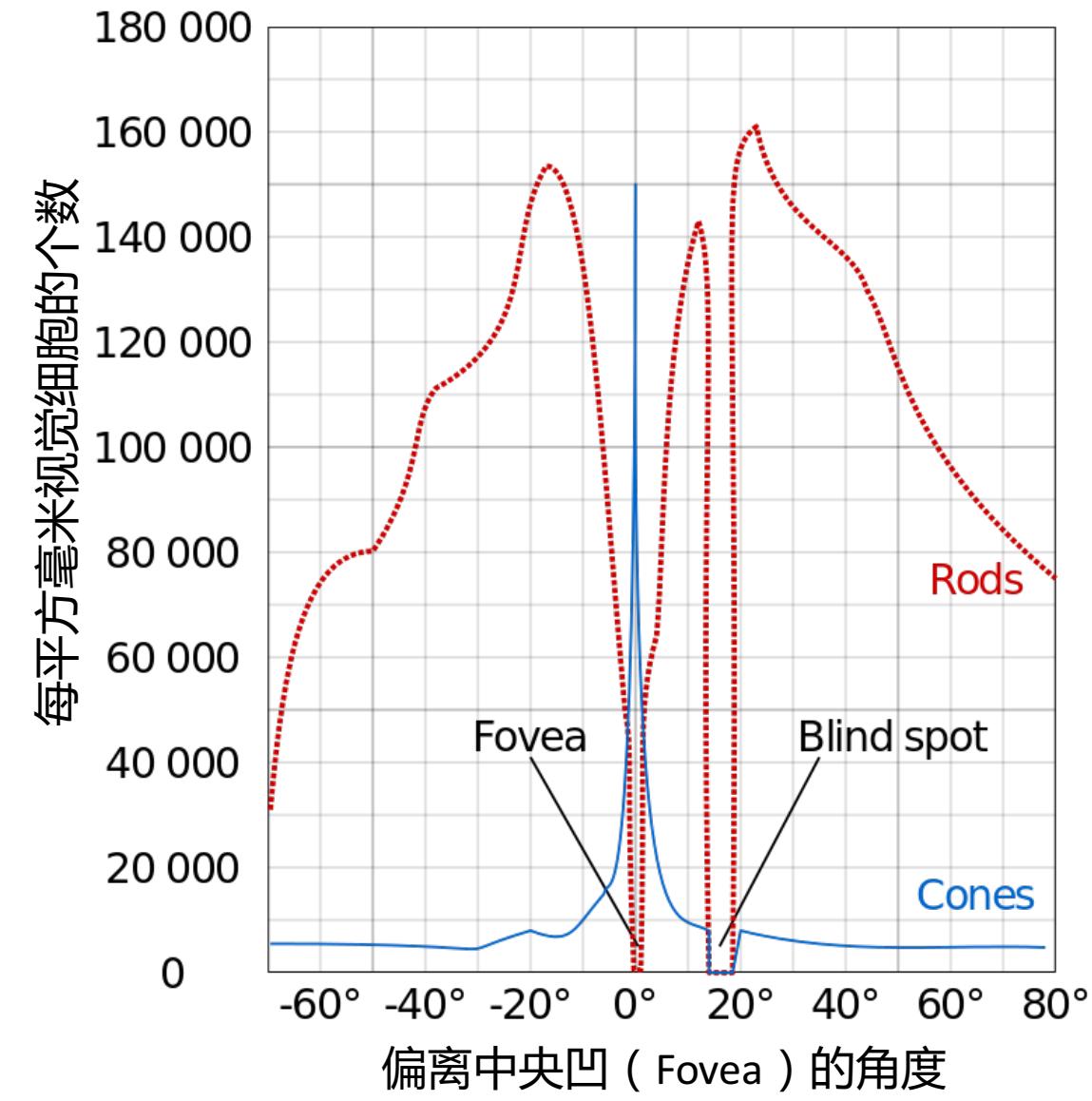
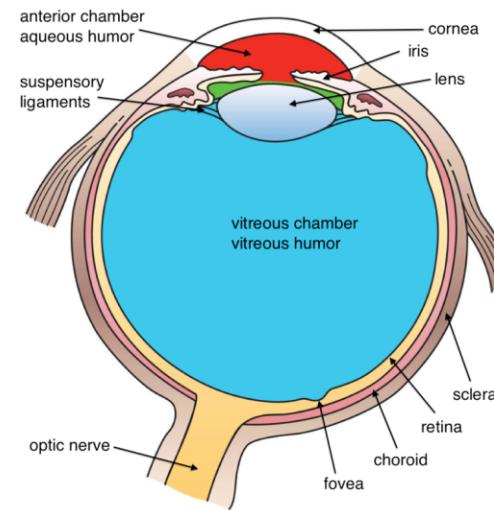
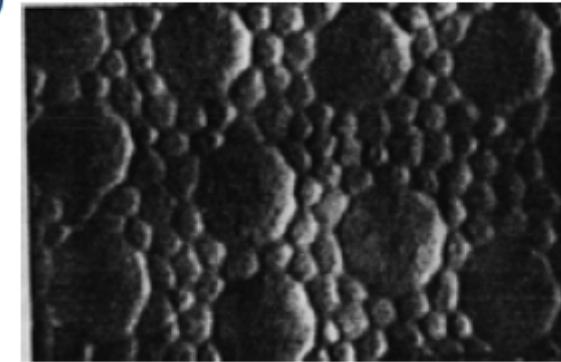
- 人眼平均包含：
- 5 百万柱状细胞
- 1亿万锥状细胞

(a)



10  $\mu\text{m}$

(b)



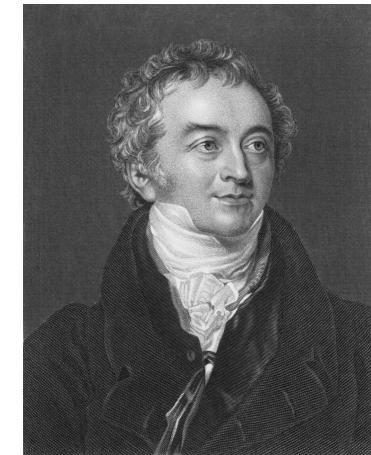
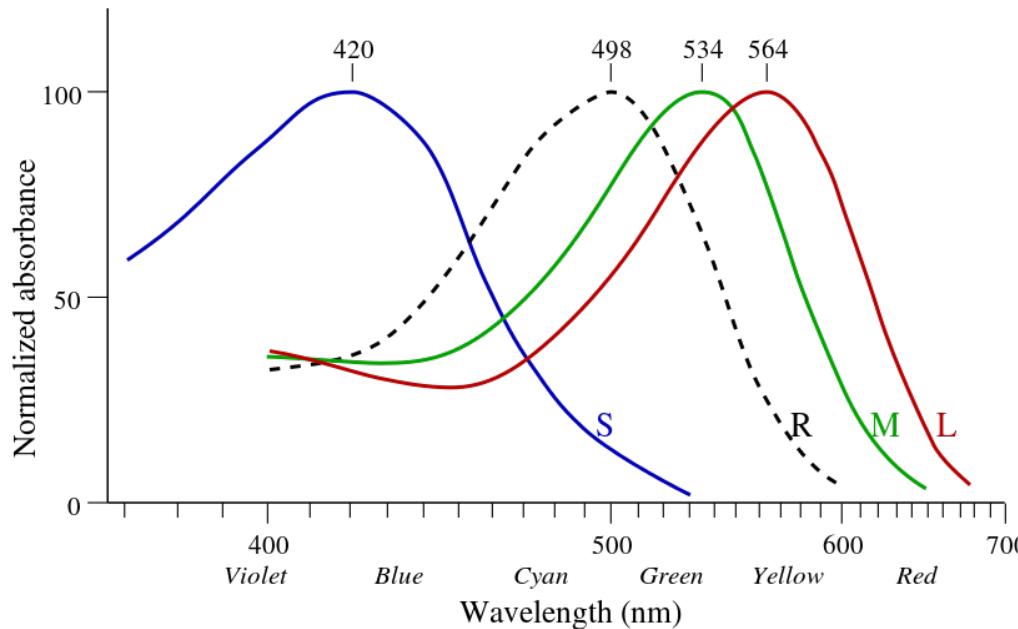
# 人类视觉系统

- **三原色原理 Trichromacy theory**

- 锥状细胞对于感知颜色至关重要

- **S-cones:** short wavelength
- **M-cones:** median wavelength, and
- **L-cones:** long wavelength

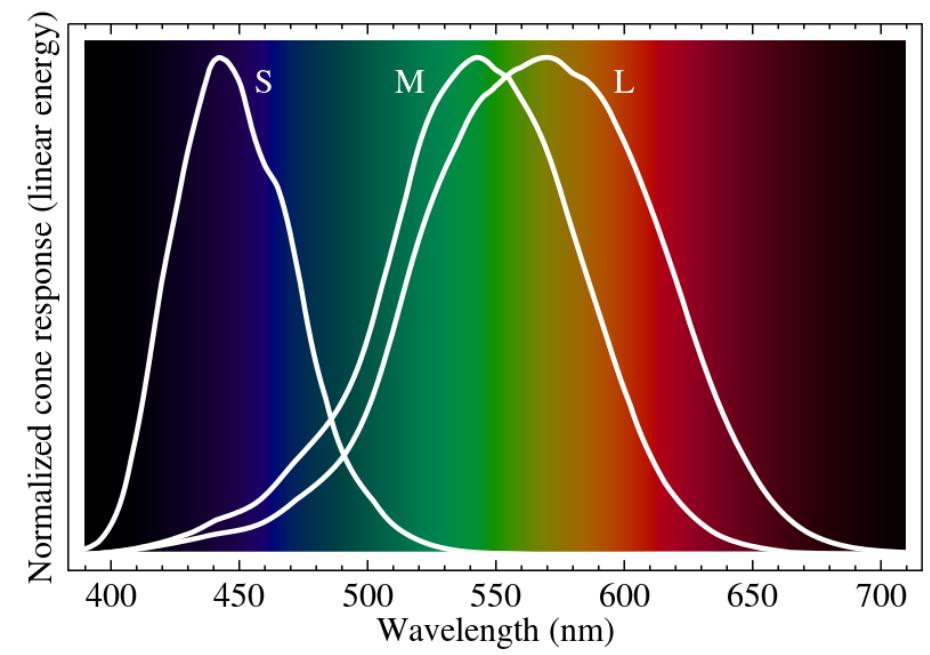
三种柱状细胞的波长响应；虚线为锥状细胞的波长响应。



Thomas Young  
(1773-1829)

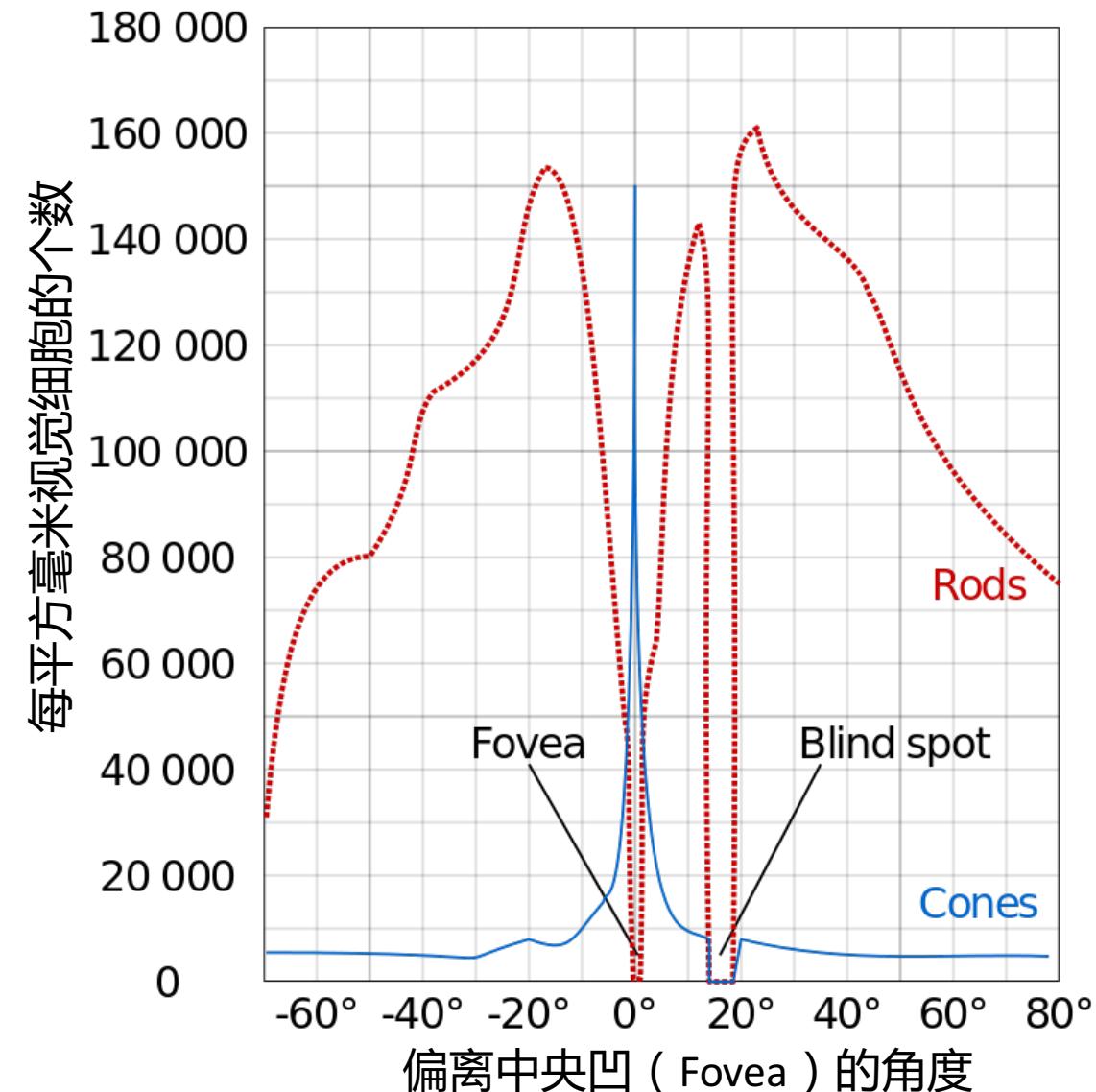


Hermann von Helmholtz  
(1821-1894)



# 人怎么“看”？

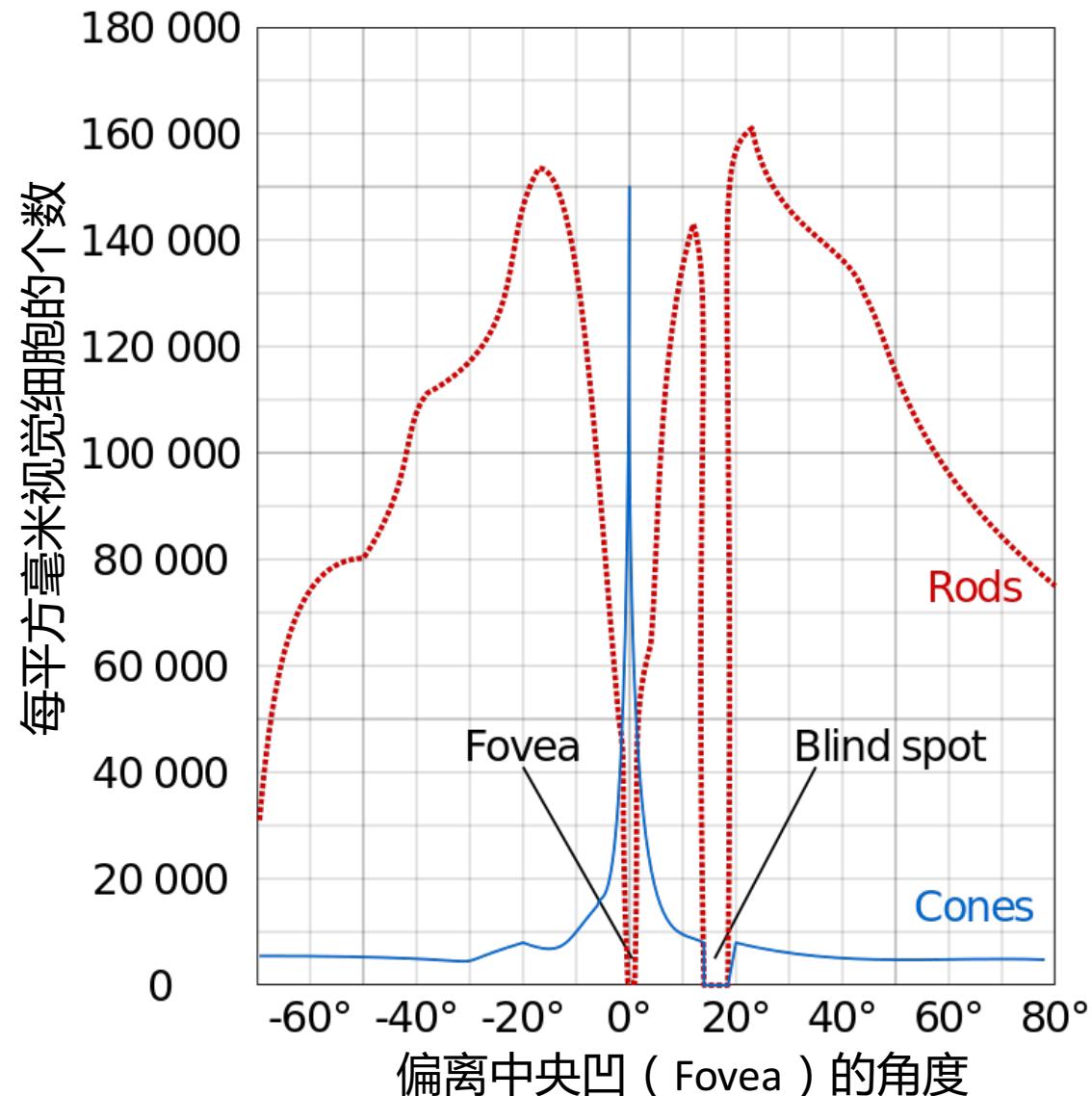
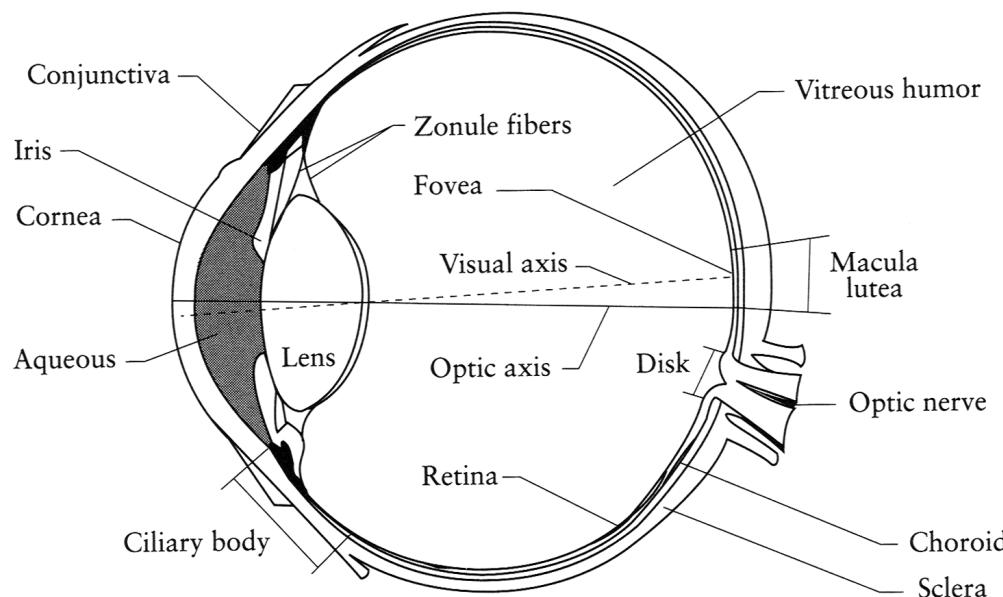
- 思考题：为什么人在看的时候，“盯着”的地方比较清晰，而周围的区域会模糊？



# 人怎么“看”？

## • 思考题答案：

- 与视觉细胞的分布有关。
- 盯着的地方，会投影在中央凹（Fovea）附近，视觉细胞比较密集，特别是锥状细胞（对细节比较敏感），所以有比较多的细胞对其进行感知，看得比较清晰。
- 而其他区域，投影到距离中央凹较远的区域，只有少量细胞进行感知，细节会丢失，就会模糊。



# CONTENTS

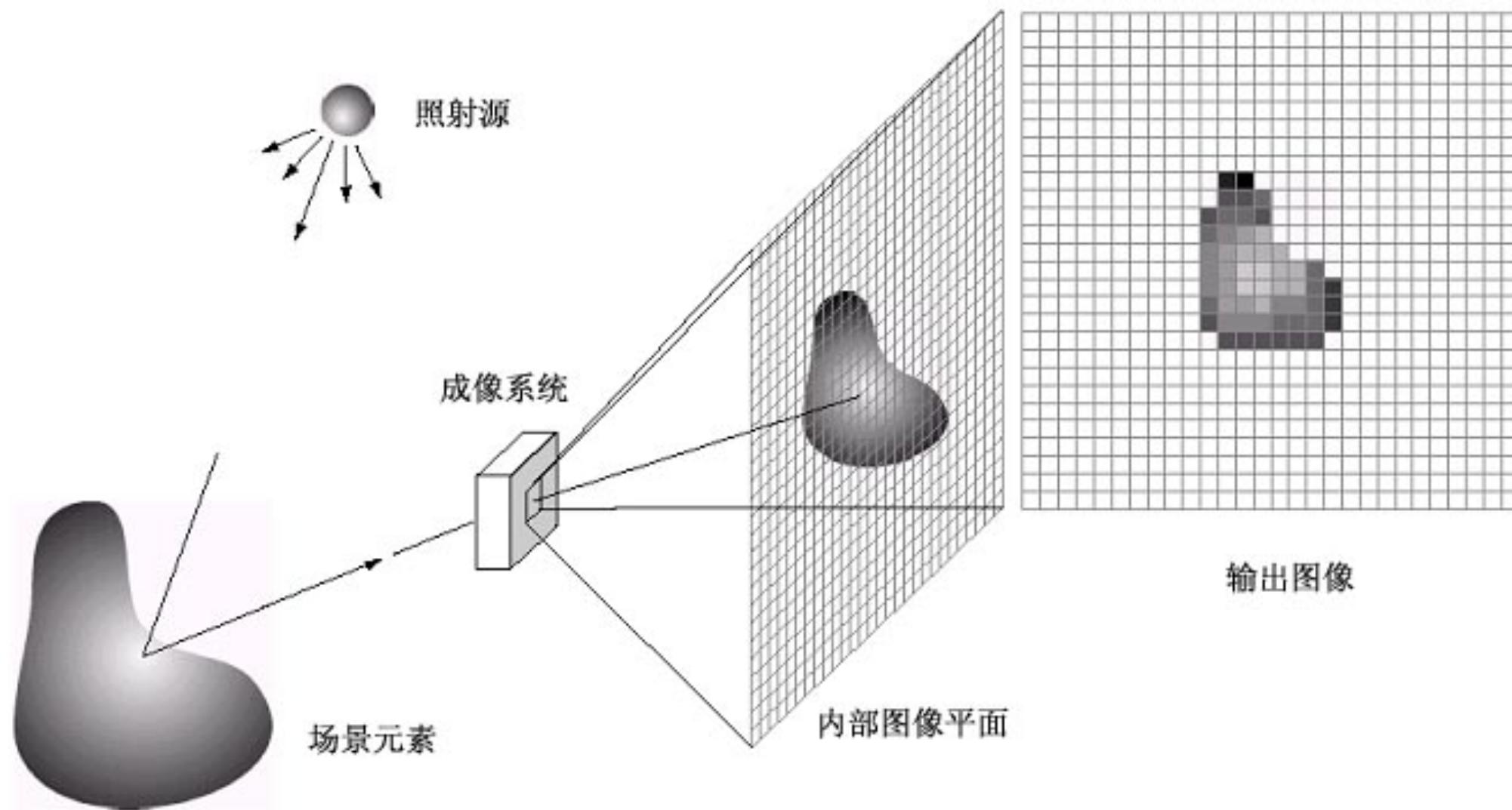
背景及意义

人怎么“看”？

机器怎么“看”？

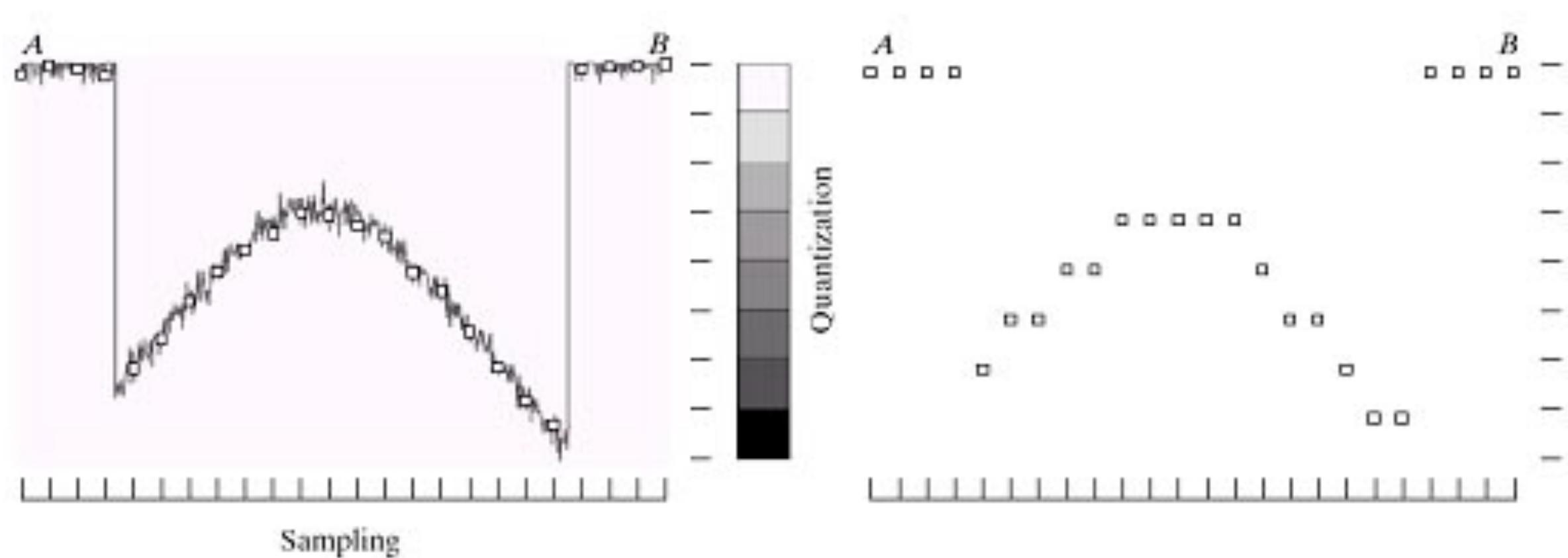
小结（视觉信息）

# 机器（计算机）怎么“看”？



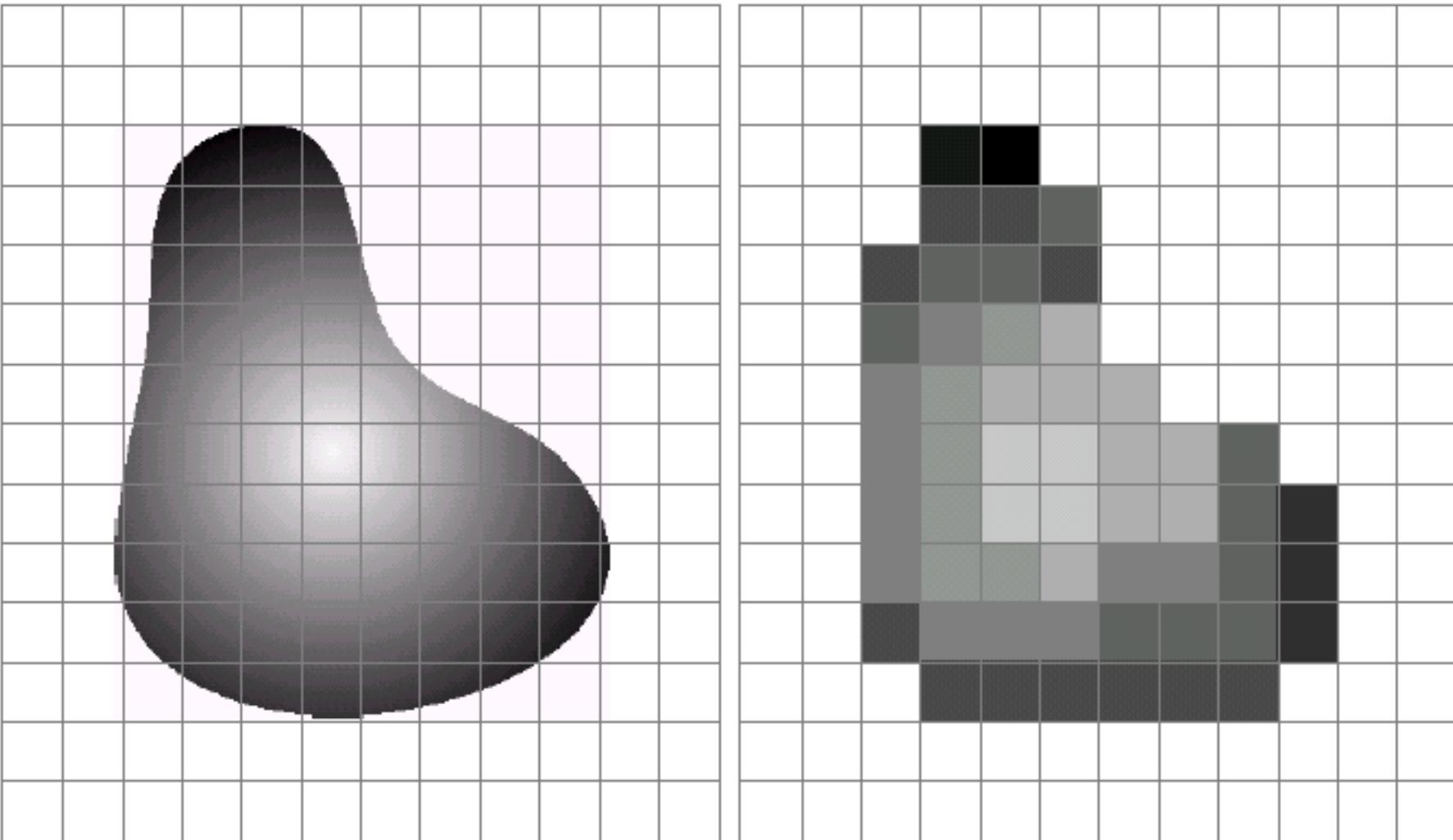
# 机器（计算机）怎么“看”？

- 模拟图像、数字图像
- 采样、量化



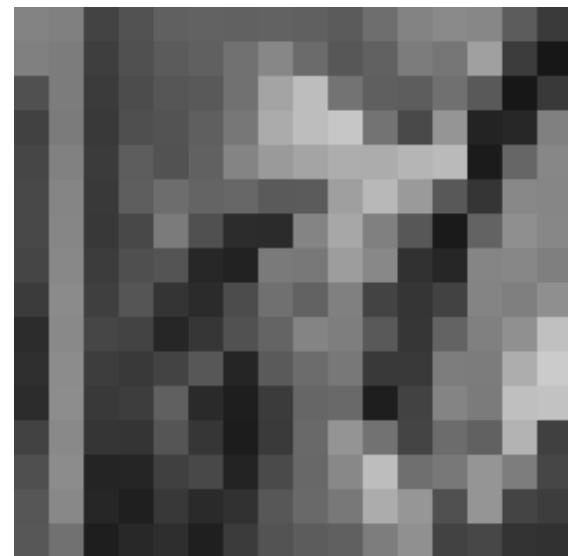
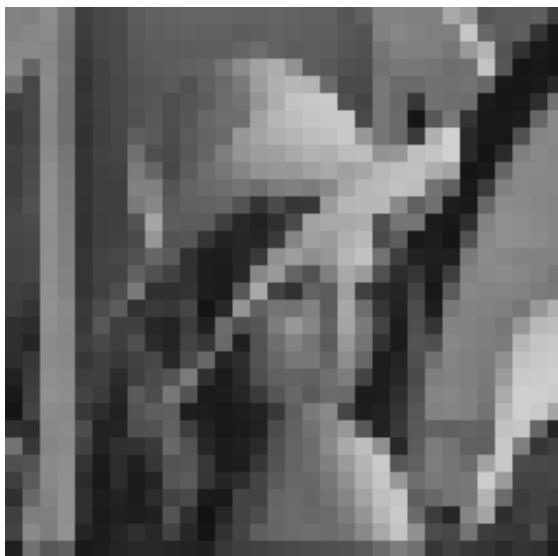
# 机器（计算机）怎么“看”？

- 模拟图像 vs 数字图像



# 机器（计算机）怎么“看”？

## • 采样点数与图像质量之间的关系



- (a) 采样点 $256 \times 256$ 时的图像
- (b) 采样点 $64 \times 64$ 时的图像
- (c) 采样点 $32 \times 32$ 时的图像
- (d) 采样点 $16 \times 16$ 时的图像

采样点数 = 分辨率

# 机器（计算机）怎么“看”？

- 量化级数与图像质量之间的关系



- (a) 量化为2级的Lena图像
- (b) 量化为16级的Lena图像
- (c) 量化为256级的Lena图像

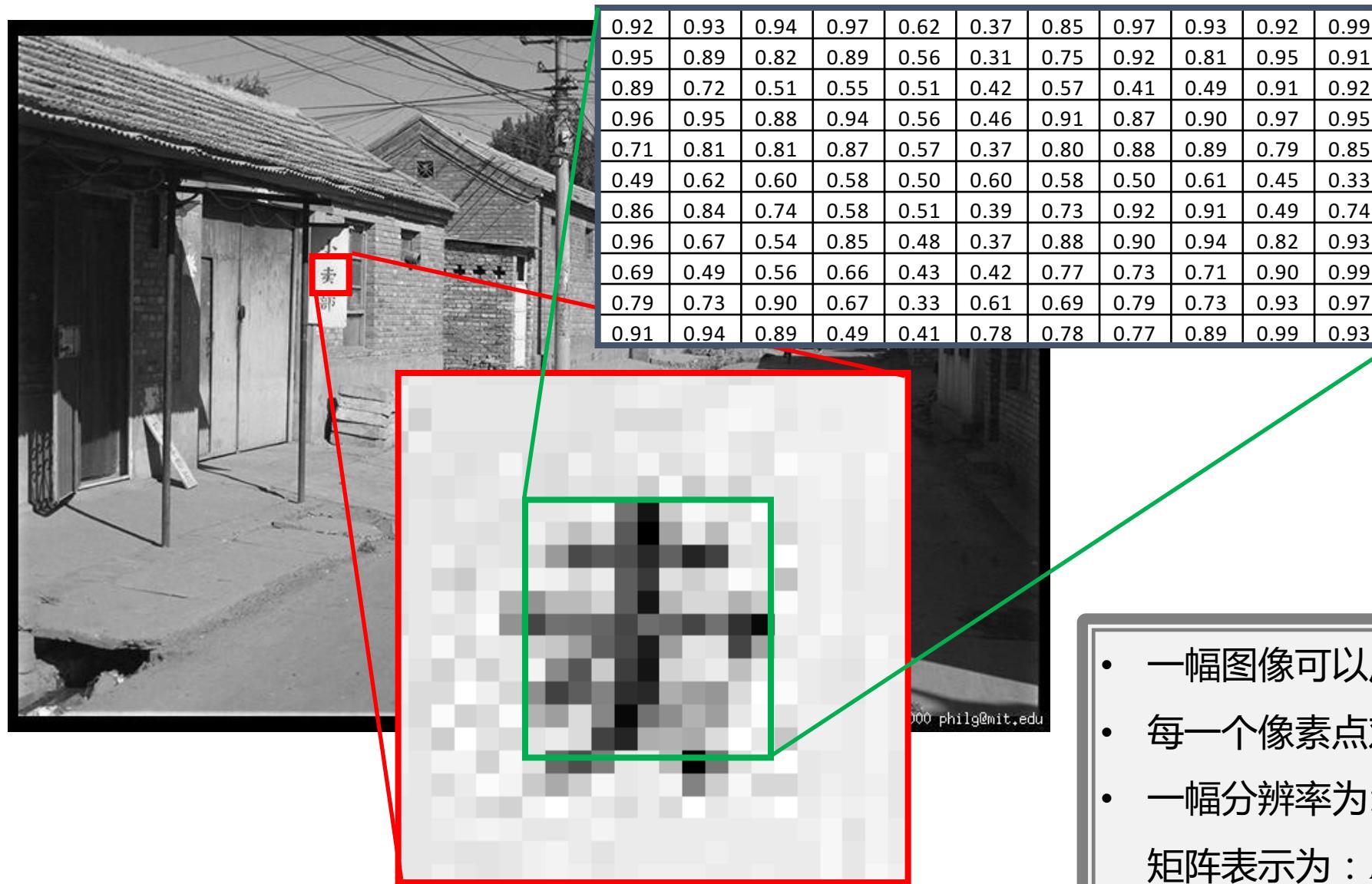
# 机器（计算机）怎么“看”？

- 一幅图像在计算机中是如何表示的呢？



copyright 2000 philg@mit.edu

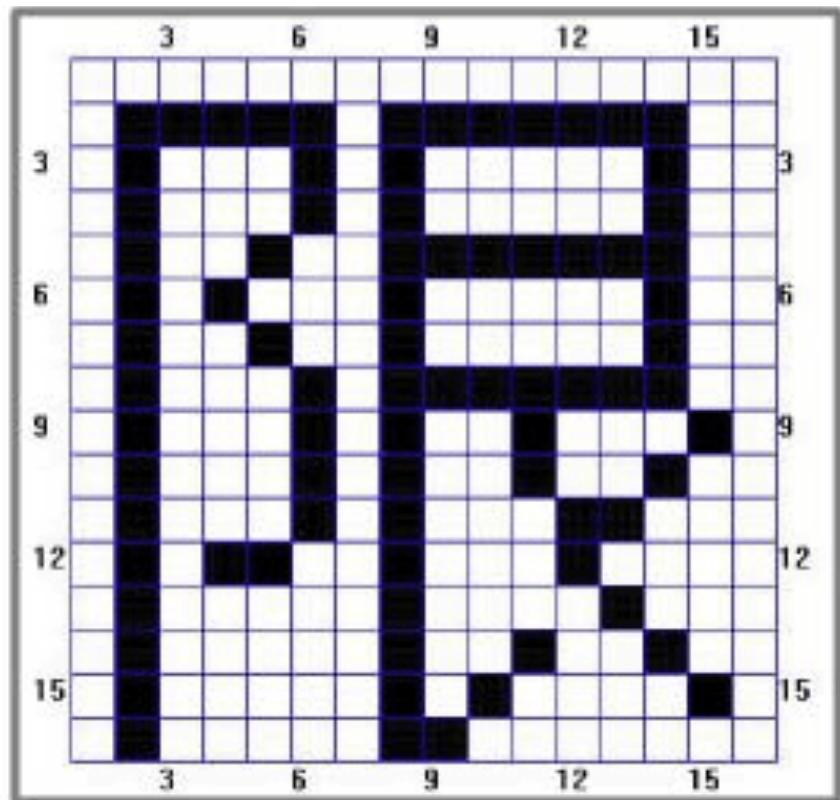
# 机器（计算机）怎么“看”？



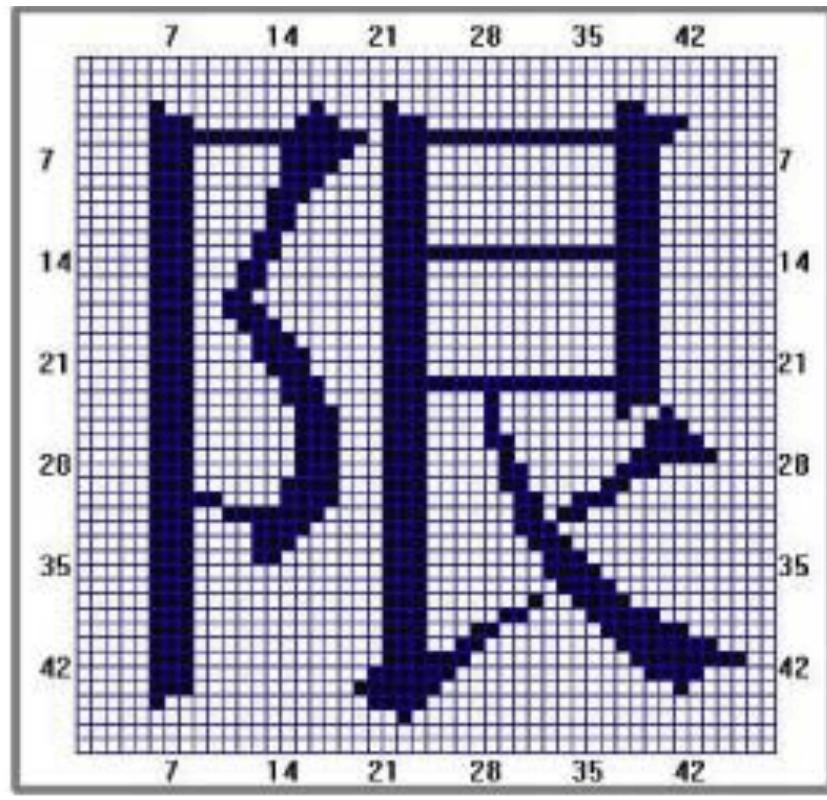
- 一幅图像可以用**矩阵**表示；
- 每一个像素点对应矩阵中的一个元素。
- 一幅分辨率为 $1024 \times 768$ 的图像，对应的矩阵表示为： $I \in \mathbb{R}^{1024 \times 768}$

# 机器（计算机）怎么“看”？

- 问题：以下两幅图像（点阵字库）的分辨率是多少？像素点个数是多少？



(图一)



(图二)

# 机器（计算机）怎么“看”？

• 例：手机配置中通常列出相机的以下参数：

- 1600万像素
- 18 : 9 全面屏



• 例：大家在观看视频时通常可以选择以下参数：

- 480P, 720P, 1080P

• 问题：这些参数是什么意思？

骁龙™ 660 高性能八核处理器	6GB 双通道 内存容量可达	256GB 闪存容量可达
5.99 英寸 In-Cell 屏幕	18:9 全面屏 2160 × 1080 分辨率	USB OTG 反向充电功能
3500mAh 大容量电池	18W QC3.0 安全快充	1200 万 + 500 万 后置双摄像头
1600 万像素 高清前置摄像头	指纹 + 人脸 便捷解锁	指纹支付 微信等方式

# 机器（计算机）怎么“看”？

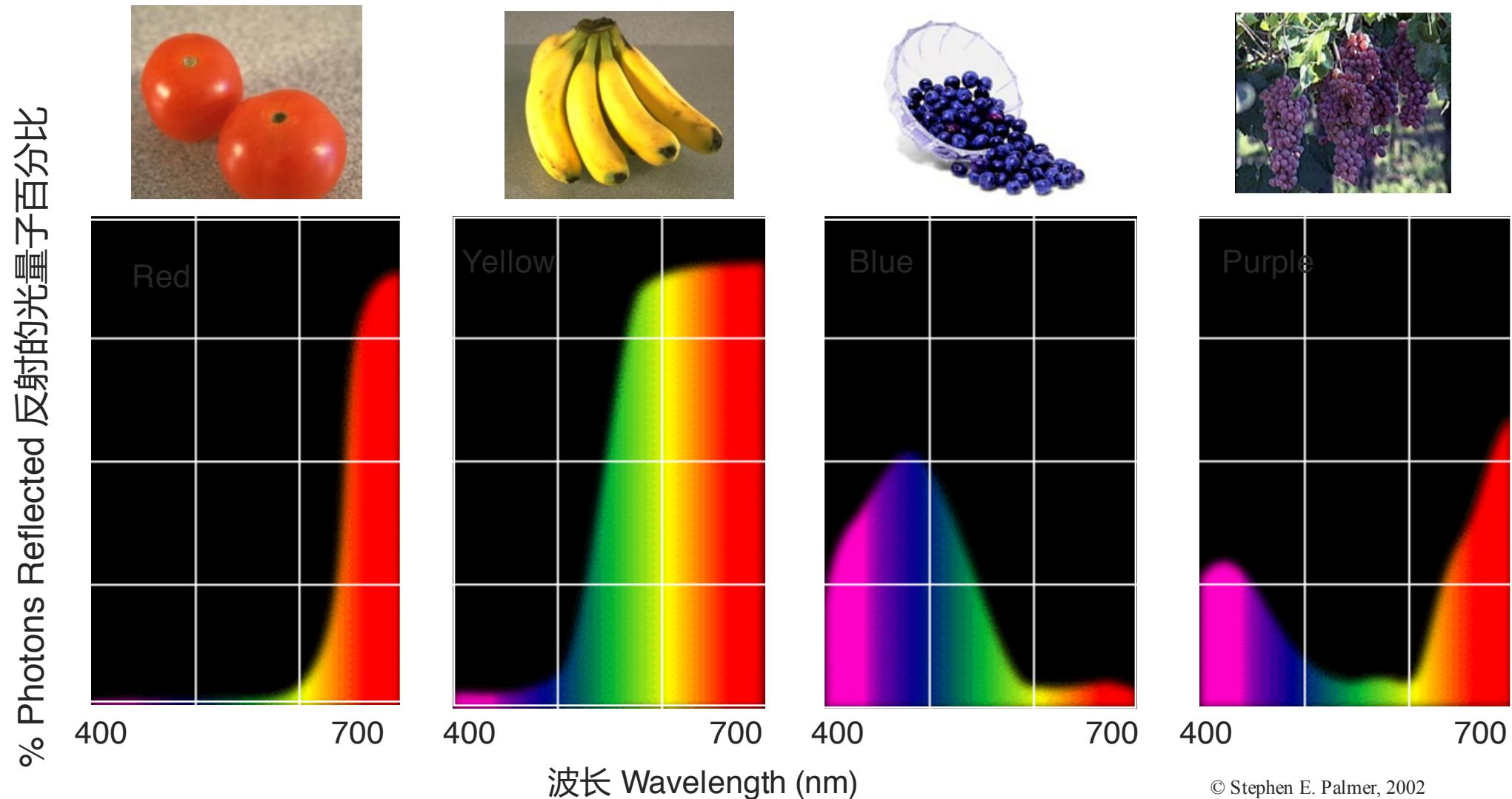
颜色是什么？



copyright 2000 philg@mit.edu

# 机器（计算机）怎么“看”？

- 物体表面反射光谱



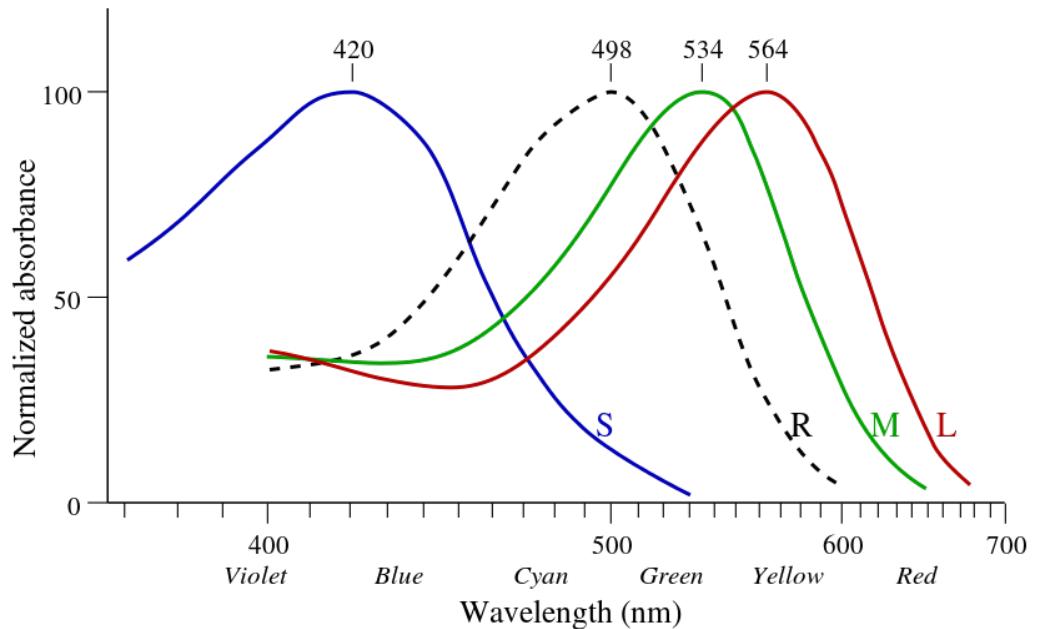
# 机器（计算机）怎么“看”？

## • 回顾：三原色原理 Trichromacy theory

- 锥状细胞对于感知颜色至关重要

- **S-cones:** short wavelength
- **M-cones:** median wavelength, and
- **L-cones:** long wavelength

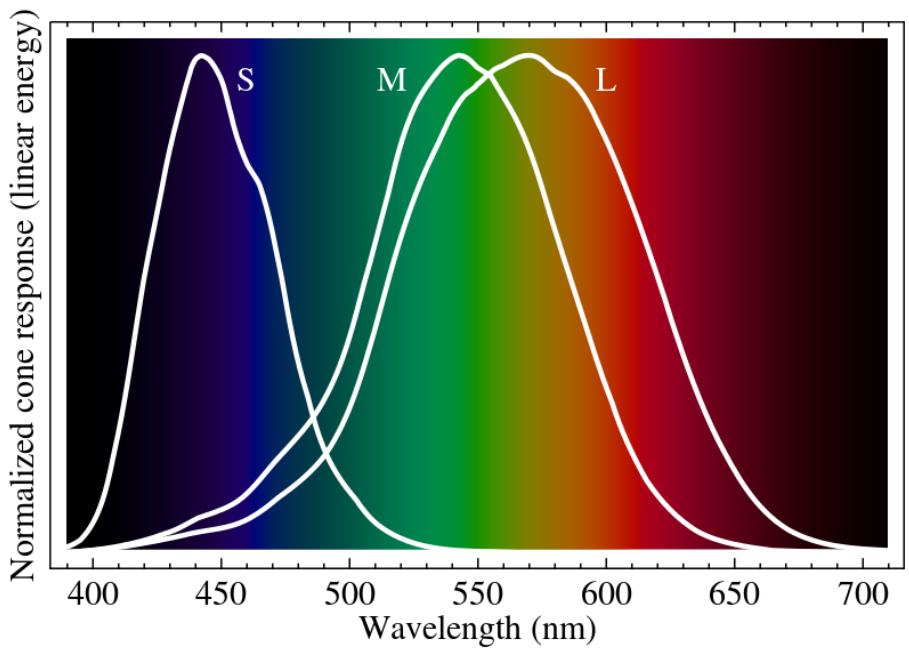
三种柱状细胞的波长响应；虚线为锥状细胞的波长响应。



Thomas Young  
(1773-1829)



Hermann von Helmholtz  
(1821-1894)

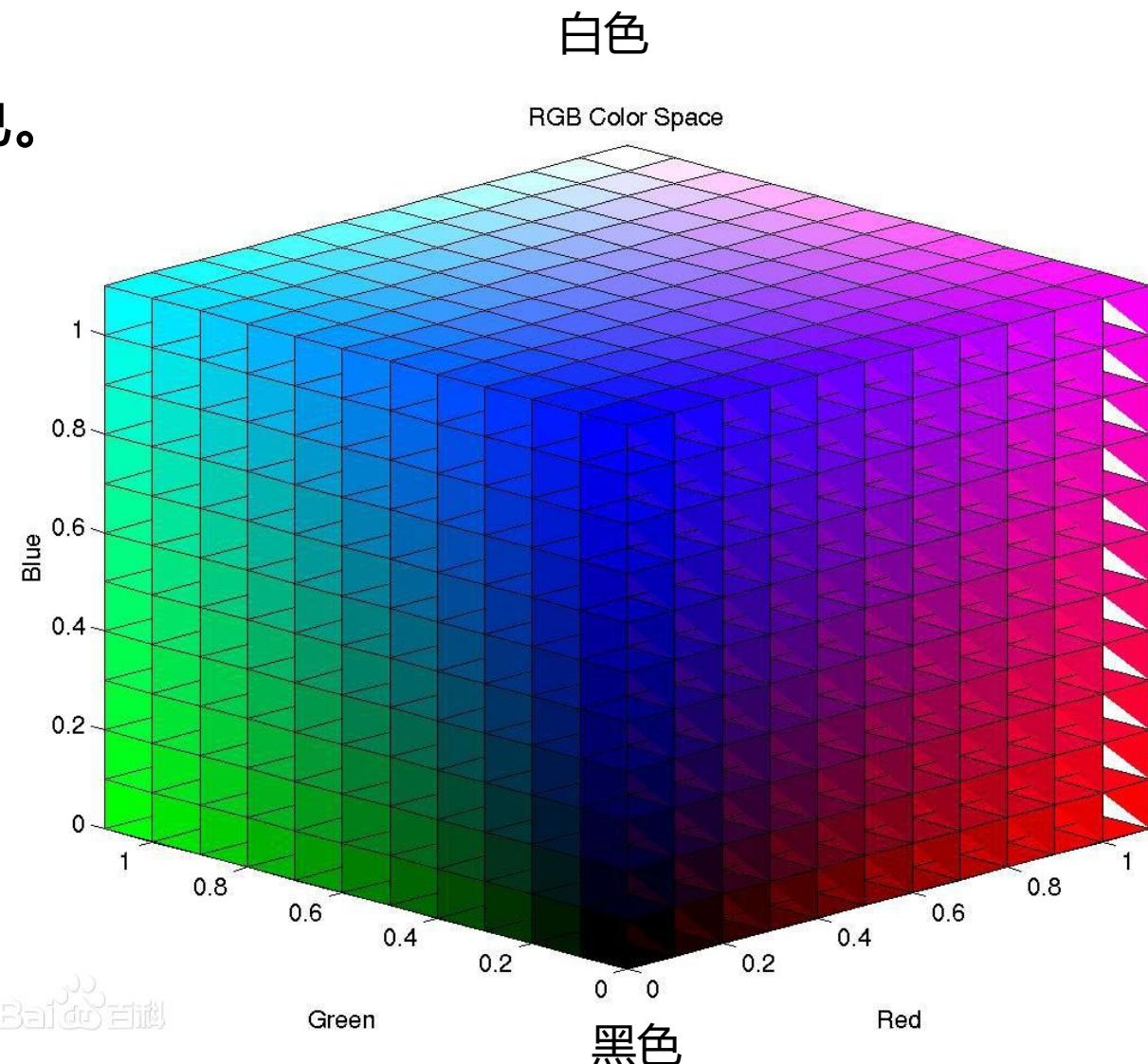


# 机器（计算机）怎么“看”？

## • RGB颜色空间

- 人类的眼睛可分辨约一千万颜色。

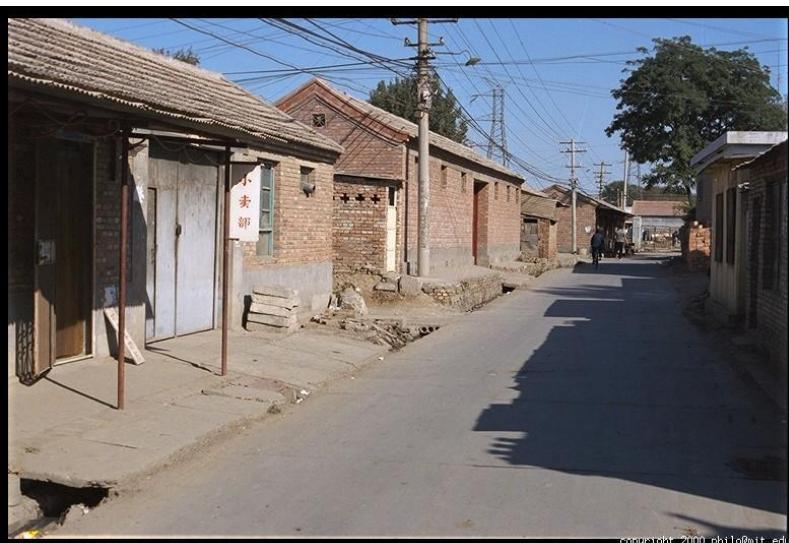
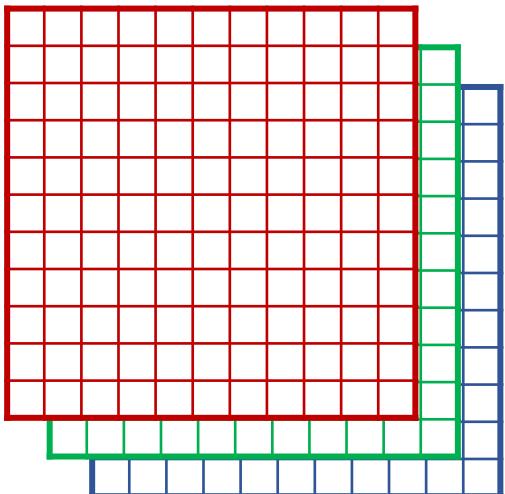
颜色名称	红色值 Red	绿色值 Green	蓝色值 Blue
黑色	0	0	0
蓝色	0	0	255
绿色	0	255	0
青色	0	255	255
红色	255	0	0
洋红色	255	0	255
黄色	255	255	0
白色	255	255	255



# 机器（计算机）怎么“看”？

- 彩色图像

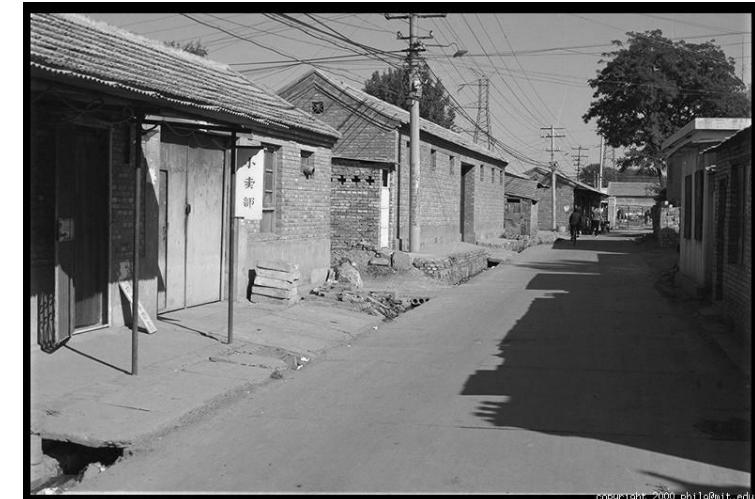
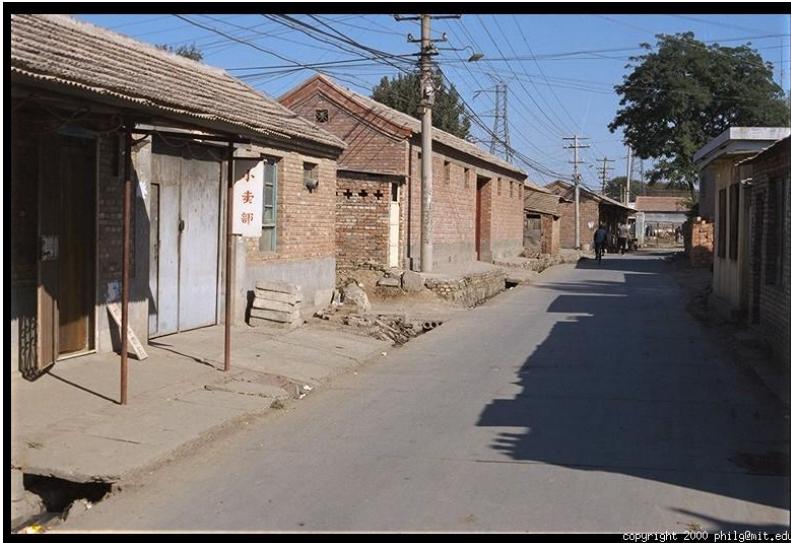
- 每个像素点由3个矩阵元素表示；
- 一幅分辨率为 $1024 \times 768$ 的彩色图像，对应的矩阵表示为： $I \in \mathbb{R}^{1024 \times 768 \times 3}$
- 术语：颜色通道



# 机器（计算机）怎么“看”？

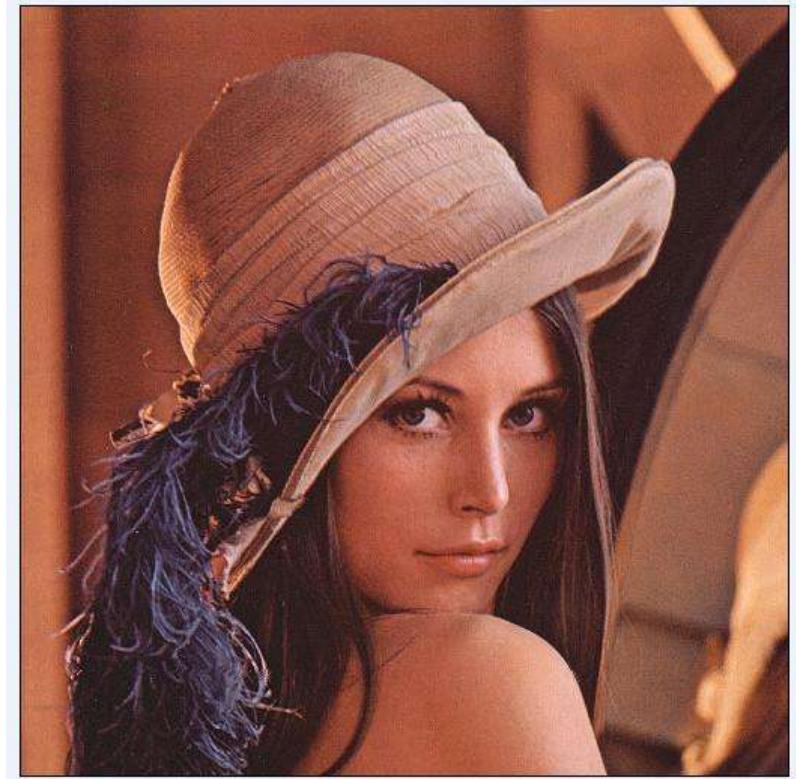
- 彩色图像 → 灰度图像

$$Y = \frac{R + G + B}{3}$$



# 机器（计算机）怎么“看”？

- 问题：一幅尺寸为 $1024 \times 1024$ 的RGB彩色图像，如果每个像素值由8个比特进行表示，则其文件大小为多少？（单位：字节 B）

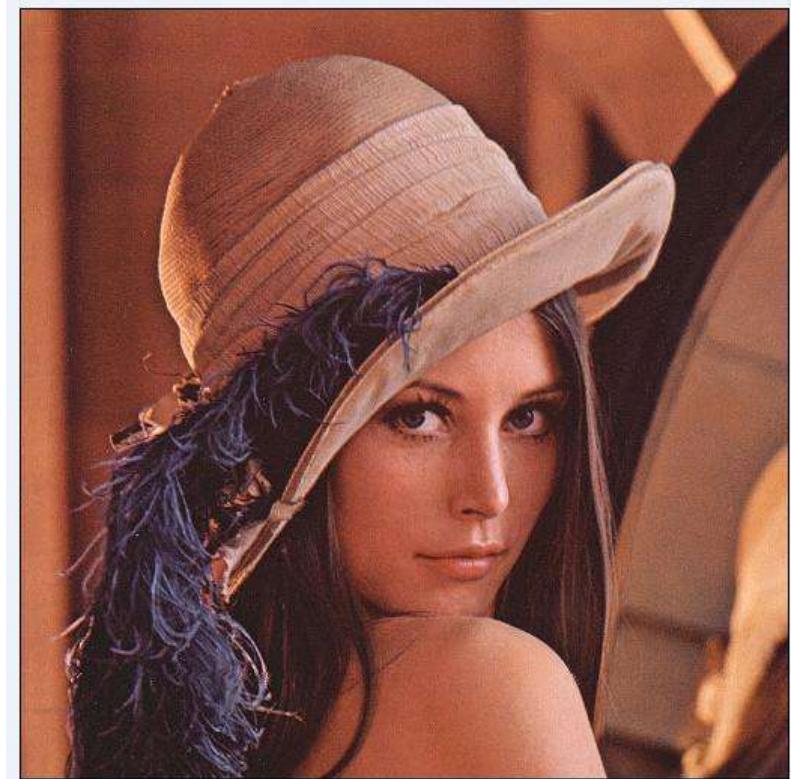


# 机器（计算机）怎么“看”？

- 问题：一幅尺寸为1024\*1024的RGB彩色图像，如果每个像素值由8个比特进行表示，则其文件大小为多少？（单位：字节 B）

- 答案：

- $= 1024 * 1024 * 8 * 3 / 8 \text{ B}$
  - $= 1024 * 1024 * 3 \text{ B}$
  - $= 3\text{MB}$
- 
- 注：1字节 = 8比特

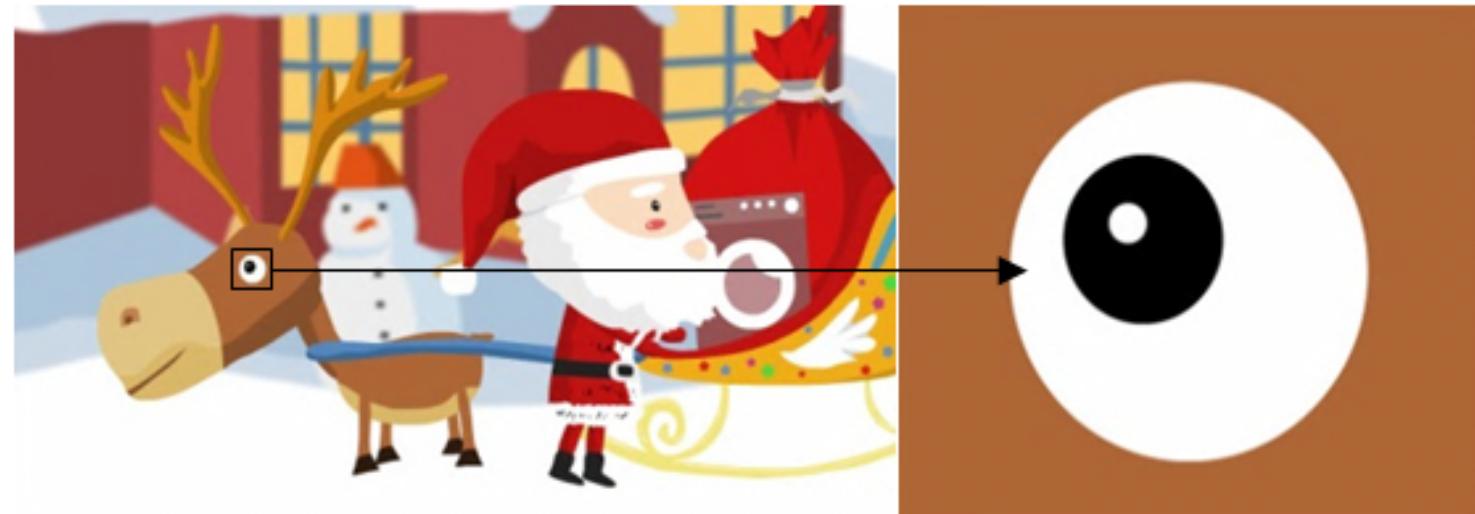


# 机器（计算机）怎么“看”？

## • 图像格式

### • 位图

- 记录每一个像素的颜色值，再把这些像素点组合成一幅图像；放大看到像素点。



(a) 矢量图局部放大

### • 矢量图

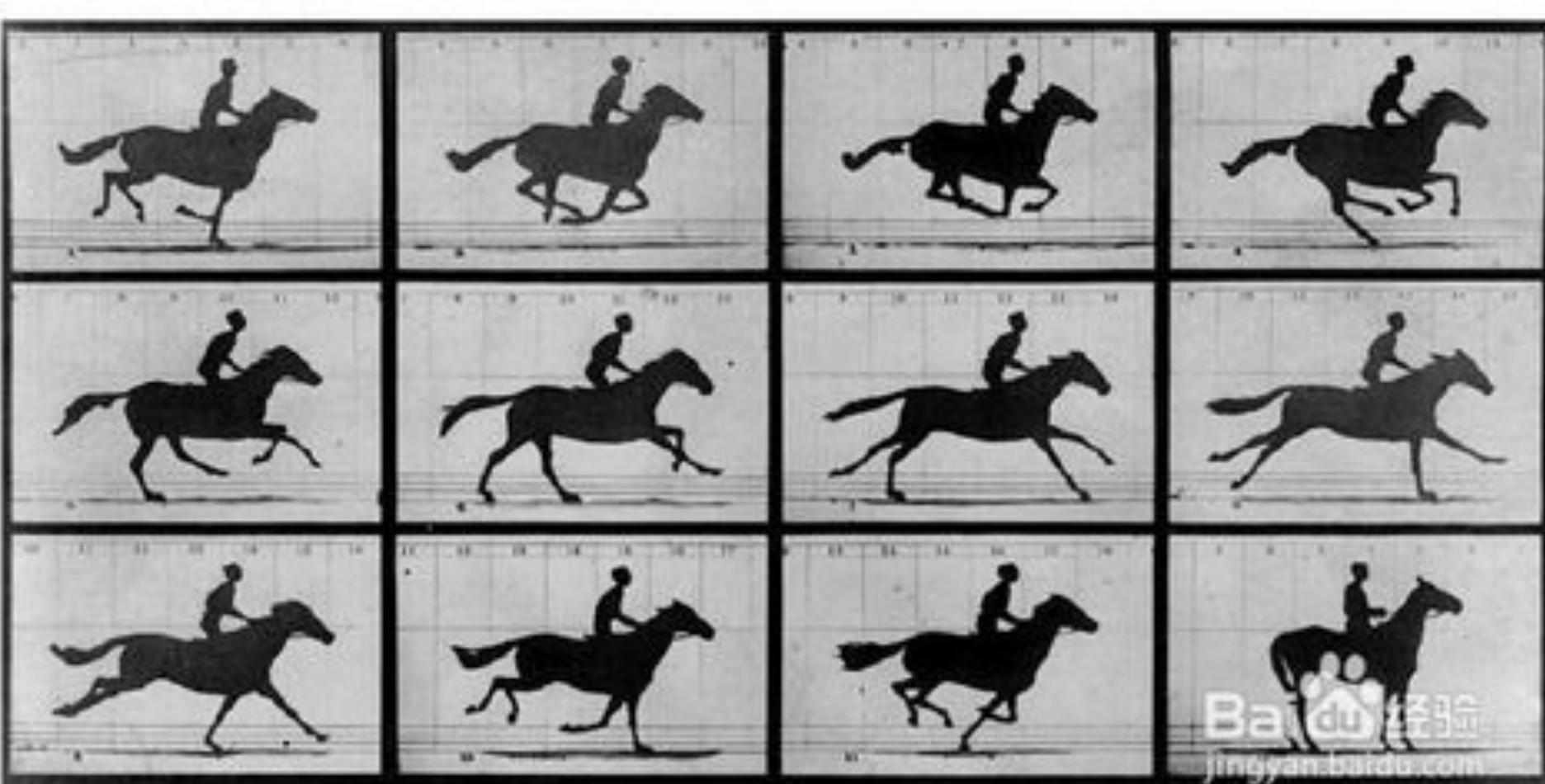
- 保存节点的位置和曲线、颜色的算法；放大不失真；
- 位图占用的存储空间较矢量图要大的多，而矢量图的显示速度较位图慢。



(b) 位图局部放大

# 视频

- 视频（Video）就是其内容随时间变化的一组动态图像，所以又叫运动图像或活动图像。



# 视频

## • 原理：视觉余像

- 当外界物体的视觉刺激作用停止后，在眼睛视网膜上的影象感觉不会马上消失，这种现象的发生是由于神经兴奋留下的痕迹作用，称为视觉余像或视觉暂留。
- 时值是二十四分之一秒；
- 电影的播放：一般为24帧/秒。
- 《比利·林恩的中场战事》，导演 李安
  - 120帧/秒；视觉效果更好



# CONTENTS

背景及意义

人怎么“看”？

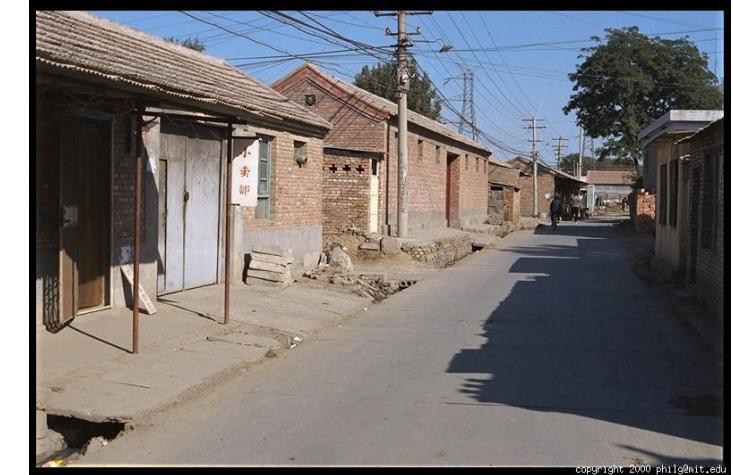
机器怎么“看”？

小结（视觉信息）

# 小结（视觉信息）

## • 本节内容：

- 人眼视觉系统：构成、锥状细胞、柱状细胞
- 三原色原理
- 概念：采样、量化、像素、分辨率、颜色通道



copyright 2000 phils@mit.edu

# 小结（视觉信息）

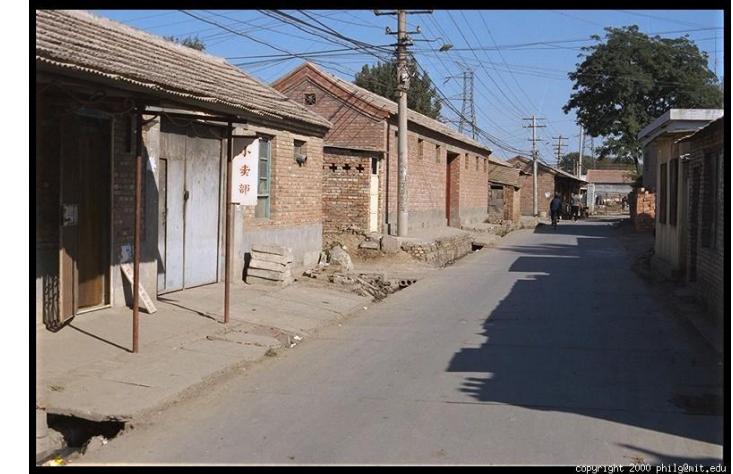
- 本节内容：

- 人眼视觉系统：构成、锥状细胞、柱状细胞
- 三原色原理
- 概念：采样、量化、像素、分辨率、颜色通道

- 本节内容：计算机怎么“看到”？

- 问题：计算机怎么“看懂”？

- 能够从图像的像素数据中推理得到什么？



copyright 2000 phils@mit.edu



# CONTENTS

机器学习基本概念（回顾）

视觉信息

自然语言

小结

# 背景介绍

- **计算语言学(Computational Linguistics)**

- 利用电子数字计算机进行的语言分析。已开发的领域包括**自然语言处理**，言语识别，自动翻译，语法的检测，以及许多需要统计分析的领域。

- **自然语言处理（Natural Language Processing, NLP）**

- 利用计算机为工具对人类特有的书面形式和口头形式的**自然语言**的信息进行各种类型**处理和加工**的技术

# 背景介绍

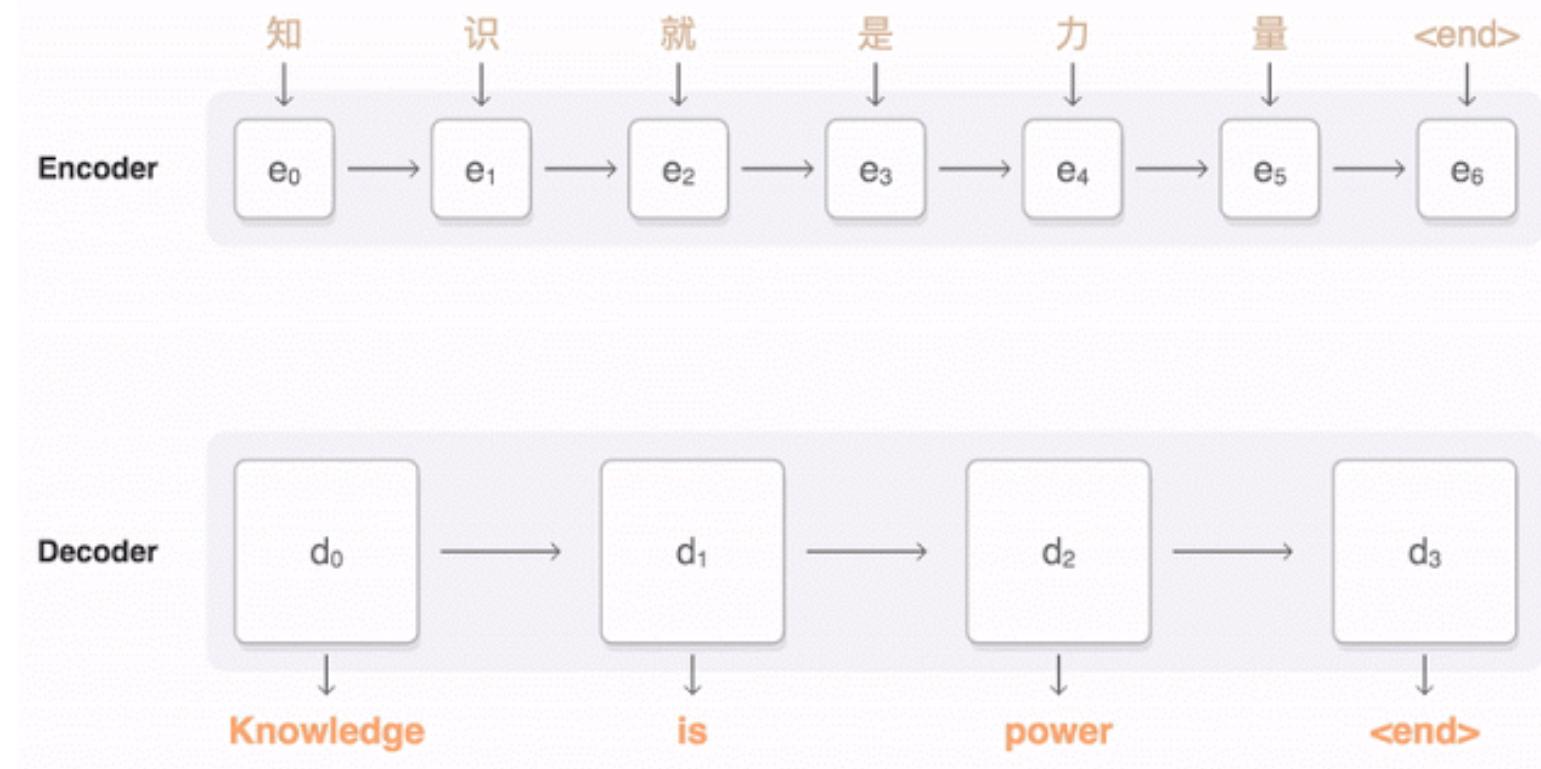
## • 自然语言处理的基本任务

- 词干抽取 ( {doing, does, did, done} → do )
- 词性标注 ( people → noun., like → vt., happy → adj. )
- 命名实体识别 ( “天安门” → 地点, “奥巴马” → 人名 )
- 中文分词 ( “今天天气不错” → “今天”, “天气”, “不错” )

# 背景介绍

- 自然语言处理应用

- 机器翻译



# 背景介绍

- 自然语言处理应用

- 文本摘要自动生成

原文：

海湾报刊对美国新当选总统克林顿，能否帮助振兴中东和平进程感到怀疑，但也确实看到了一丝希望。

=====

人类：海湾报刊对克林顿是否会恢复和平进程，持怀疑态度

机器：海湾新闻界对克林顿恢复和平进程的前景，持怀疑态度

# 背景介绍

## • 自然语言处理应用

### • 文本情感分析

最有用的好评	最有用的差评
<p>33/34 人认为此评论有用</p> <p>★★★★★ 很愉快的一次购物 在忐忑不安中下单 和等待中收到了 苹果5S金色的手机，之后验证后是国行的没错，心里十分感激亚马逊商城，一次愉快的购物开始了我对亚马逊商城的信任。下次还会来逛逛的，并分享给身边的朋友的~！~！顶你~！</p> <p>吴永权在3个月前发表</p> <p><a href="#">查看更多5星, 4星的评论</a></p>	<p>20/24 人认为此评论有用</p> <p>★★☆☆☆ 有问题啊。蓝屏，冲不了电 有问题，蓝屏，冲不进去，退货麻烦。还得去苹果售后，检测拿报告。这么大的问题，你们不能自己拿回去看看</p> <p>赵宇哲在2个月前发表</p> <p><a href="#">查看更多3星, 2星, 1星的评论</a></p>

Vs.

# 背景介绍

- **自然语言处理应用**

- 文本分类



# 背景介绍

## • 自然语言处理实际应用



同声传译



翻译助手

文本  
表示学习



记者



客服

# 文本表示学习

- 词袋模型 ( Bag-of-word, BoW )

- 例子

*“It was the best of times,  
it was the worst of times,  
it was the age of wisdom,  
it was the age of foolishness,”*

-----Charles Dickens

# 文本表示学习

- 词袋模型 ( Bag-of-word, BoW )

- 步骤1. 构建字典

$$D = \{ "It": 1, "was": 2, "the": 3, "best": 4, "of": 5, \\ "times": 6, "worst": 7, "age": 8, "wisdom": 9, \\ "foolishness": 10 \}$$

问题2：如果字典很大会面临什么问题？

- 步骤2. 字典编码

Doc1: *It was the best of times*

[1,1,1,1,1,1,0,0,0,0]



Doc2: *It was the worst of times*

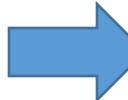
[1,1,1,0,1,1,1,0,0,0]

# 文本表示学习

- **信息稀疏存储**

- 正排表

**Doc1:** [1,1,1,1,1,1,0,0,0,0] {1,2,3,4,5,6}

**Doc2:** [1,1,1,0,1,1,1,0,0,0]  {1,2,3,5,6,7}

**Doc3:** [1,1,1,1,0,0,0,1,0,1] {1,2,3,5,8,10}

- 倒排表

it: **Doc1, Doc2 Doc3, Doc4**

was: **Doc1 Doc2, Doc3, Doc4**

times: **Doc1, Doc2**

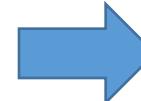
.....

# 文本表示学习

- 文本相似度计算
- 余弦相似度

$$sim(d1, d2) = \frac{d1 \cdot d2}{||d1|| ||d2||}$$

*It was the best of times*

Doc1: [1,1,1,1,1,1,0,0,0,0]  sim(Doc1, Doc2) = 5/6

*It was the worst of times*

Doc2: [1,1,1,0,1,1,1,0,0,0]

截然相反的两句话相似度非常高！

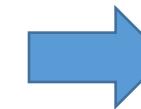
# 文本表示学习

- 文本相似度计算
- Jaccard相似度：

$$sim(Doc1, Doc2) = \frac{Doc1 \cap Doc2}{Doc1 \cup Doc2}$$

*It was the best of times*

**Doc1:** {1,2,3,4,5,6}



$$sim(Doc1, Doc2) = \boxed{5/7}$$

*It was the worst of times*

**Doc2:** {1,2,3,5,6,7}

# 文本表示学习

- **TF-IDF**
  - TF (Term Frequency)
    - 某一个给定的单词 $w$ 在该文件中出现的次数，通常会被归一化(一般是词频除以文章总词数), 以防止它偏向长的文件
  - IDF (Inverse Document Frequency)
    - 所有文档中包含单词 $w$ 的频率的倒数。总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到
  - $\text{TF-IDF} = \text{TF} * \text{IDF}$

$$\text{TF}(w, doc) = \frac{\text{doc中 } w \text{ 出现的次数}}{\text{doc中所有单词个数}}$$

$$\text{IDF}(w) = \log \frac{\text{文档总数}}{\text{包含 } w \text{ 的文档数} + 1}$$

# 文本表示学习

## • TF-IDF练习：求 Doc4的TF-IDF特征向量

- Doc1: It was the best of times,
- Doc2: it was the worst of times,
- Doc3: it was the age of wisdom,
- Doc4: it was the age of foolishness

第一步：构建词典  
第二步：对Doc4形成BoW特征  
第三步：TF(BoW)  
第四步：计算字典中次的IDF值  
第五步：TF-IDF ( BoW )

$D = \{"It": 1, "was": 2, "the": 3, "best": 4, "of": 5, "times": 6, "worst": 7, "age": 8, "wisdom": 9, "foolishness": 10\}$

# 文本表示学习

- BoW模型的缺点：

- 无法解决“一词多义”和“一意多词”问题
  - 一词多义：apple：水果 or 苹果公司？
  - 一意多词：car & automobile：都表示汽车

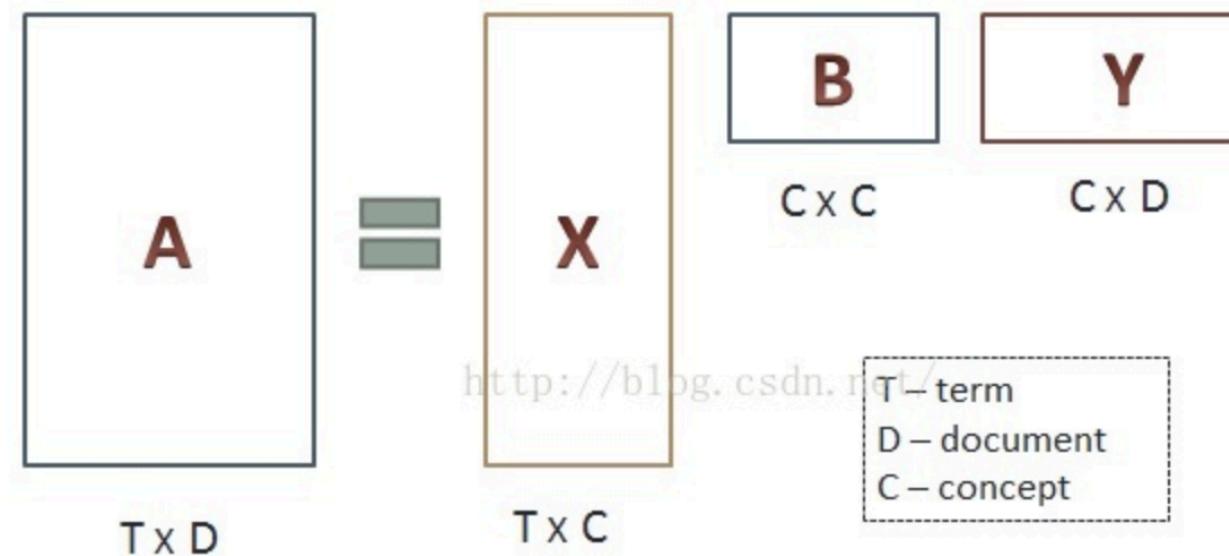
- 主题建模模型

- 基于矩阵分解：LSI
- 基于概率建模：pLSA, LDA

# 文本表示学习

## • 隐含语义分析 ( Latent Semantic Analysis, LSA )

- 对“文档-单词”构成的关系矩阵进行SVD分解，消除同义词、多义词的影响，对每个文档表达进行降维形成紧凑的隐语义（主题）表达



# 文本表示学习

- **LSA的缺点**
  - 缺乏严谨的概率学上的解释性，SVD分解非常耗时
- **pLSA**
  - 使用概率图模型对LSA进行修改，施加了主题在词上的多项式分布约束以及文档在主题上的多项式分布约束。
- **LDA**
  - pLSA模型的参数上施加了狄利克雷分布的先验

[1]. T. Hofman, Probabilistic latent semantic analysis, *Uncertainty in Artificial Intelligence (UAI)*, 1999  
[2]. D. Blei *et al.*, Latent dirichlet allocation, *Journel of machine learning research (JMLR)*, 2003

# 文本表示学习

- 词向量（word vector）：直接建模文本的表示特征难以把握文本的细粒度特性，人们开始研究单词级别的表示，在此基础上再进行文本表示
  - one-hot representation
  - word2vec
  - doc2vec
  - ...

# 文本表示学习

- **One-Hot Representation**

- 最基本的词表示方法，类似于前面讲的BoW

“话筒” 表示为 [0 0 0 1 0 0 0 0 0 0 0 0 ...]

“麦克” 表示为 [0 0 0 0 0 0 0 1 0 0 0 0 ...]

- 特征维度高，无法准确描述词之间的关系
- 期望有一种紧凑的低维的单词的**分布式表达(distributed representation)**

[0 0 0 1 0 0 0 0 0 0 0 0 ...]



[0.52 -0.12 0.896 -0.328]

# 文本表示学习

- **Word2vec**

- T. Mikolov在2013年提出的一个软件包，包含CBOW和Skip-gram两种方法
- 随机初始化所有单词的词向量，使用迭代法对初始结果不断进行更新，最终得到最优的结果

*queen + woman = king + man*

# 文本表示学习

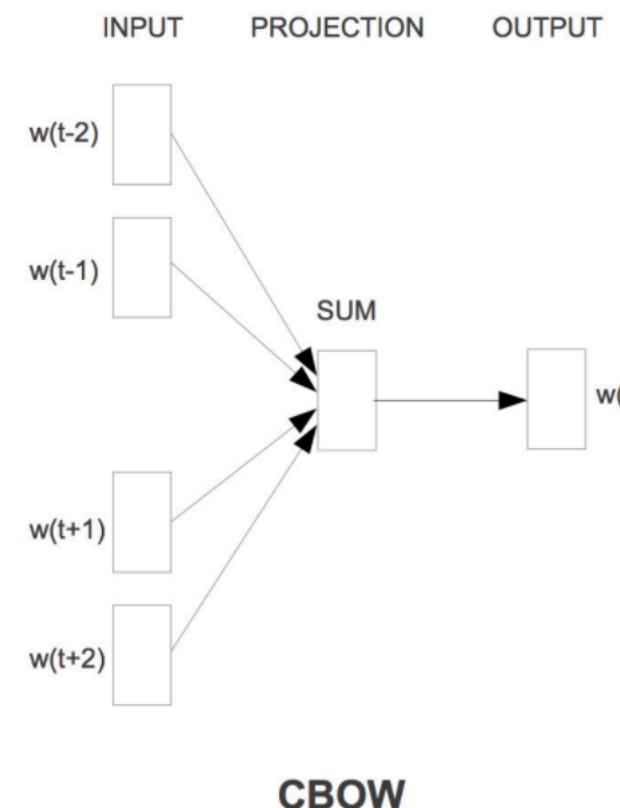
- **Continuous Bag of Word (CBoW)**

- 句子 : The cat *jumped* over the puddle.
- 窗口大小 : 2
- 中心词center : jumped
- 上下文context : {the、 cat、 over、 the}
- 用 $w^{word} \in \mathbb{R}^{|V|*1}$ 表示一个单词当前词的one-hot特征表示，例如 $w^{cat}$ 。

# 文本表示学习

- **Continuous Bag of Word (CBoW)**

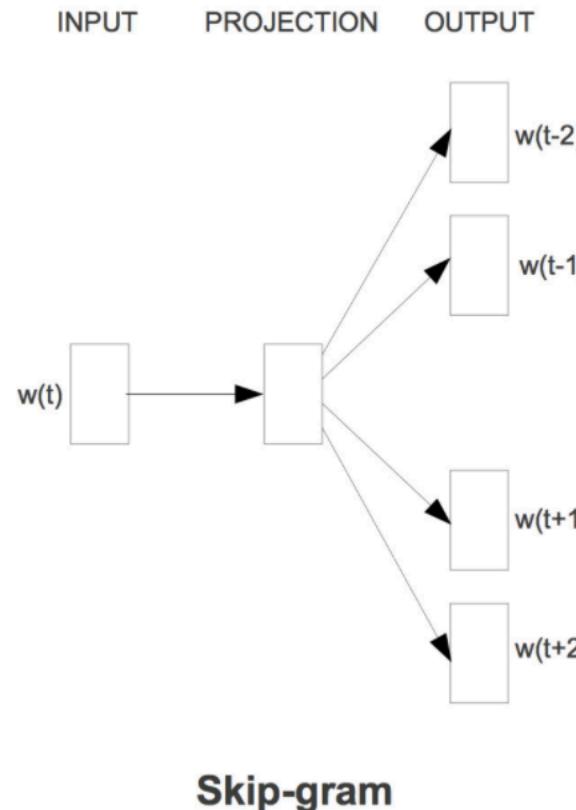
- 基于窗口内的上下文单词的词向量的加和，预测中心词  $P(\text{center}|\text{context})$



# 文本表示学习

- **Skip-Gram**

- 基于中心词的词向量，预测上下文单词  $P(context|center)$



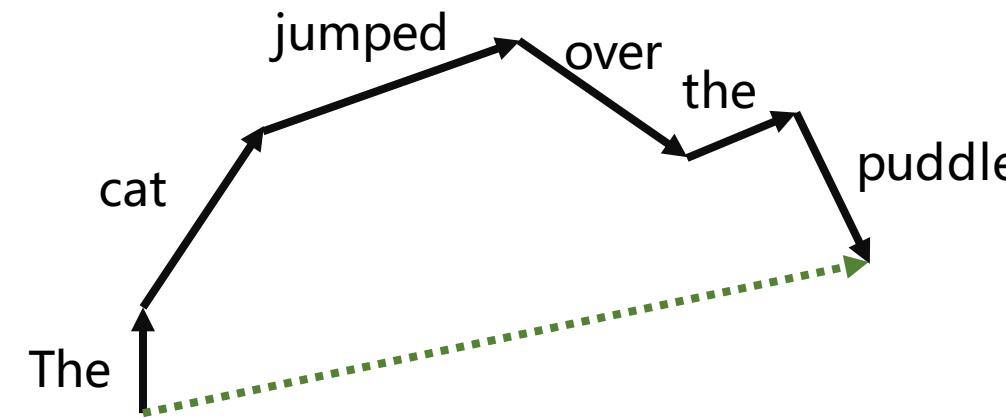
# 文本表示学习

- **CBoW vs. Skip-Gram**
  - Skip-Gram 更快
  - 由于CBOW有一个求平均的步骤，这一步给CBOW提供了一定的泛化能力，在小数据集上表现好于Skip-Gram
  - 还是因为CBOW有一个求平均的步骤，在数据量较大时，反而使得CBOW表现不如Skip-Gram

# 文本表示学习

- 基于word2vec的文本表示

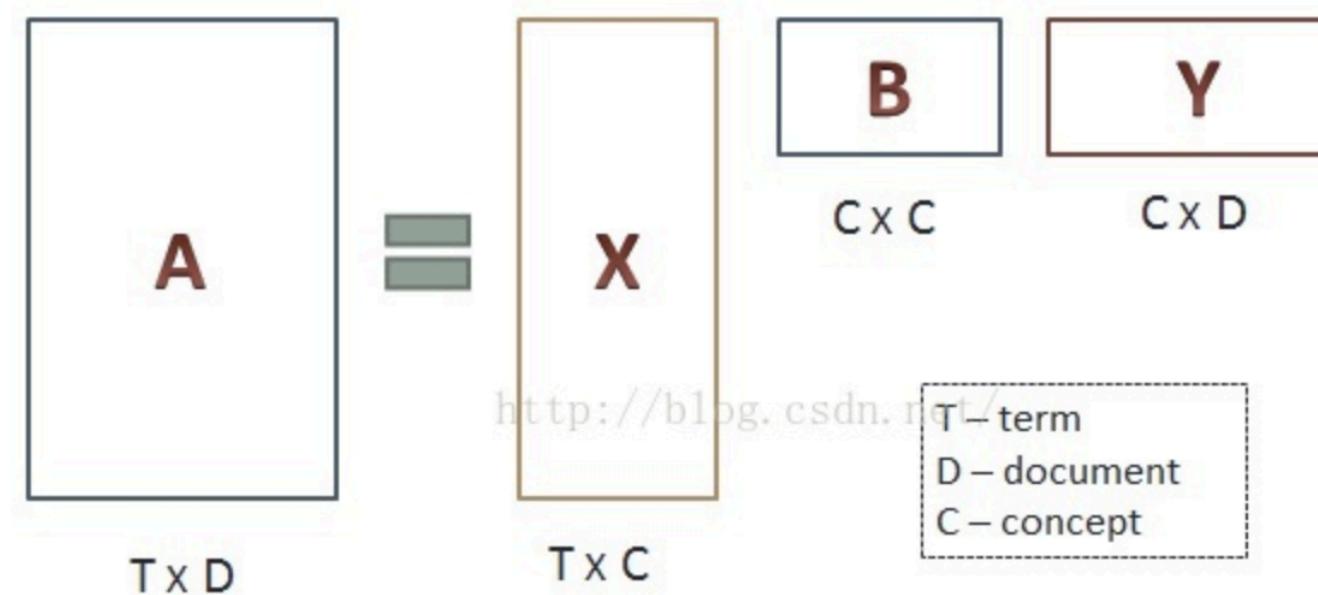
- 一般直接使用每个单词的词向量的加和表达文本



- 通常对短文本表示效果较好

# 文本表示学习

- 思考：word2vec和之前介绍LSA有什么关联？



# 文本表示学习

- **循环神经网络 ( Recurrent Neural Networks, RNN )**

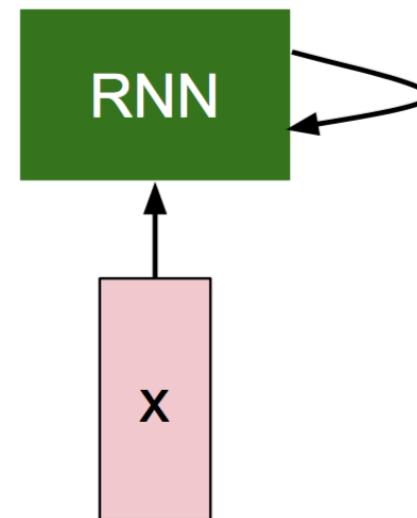
- 使用词向量在文本表示时丢失了**时序**的信息

“我喜欢那个女孩” != “那个女孩喜欢我”

- 在时序上展开的使用参数共享方式的神经网络结构

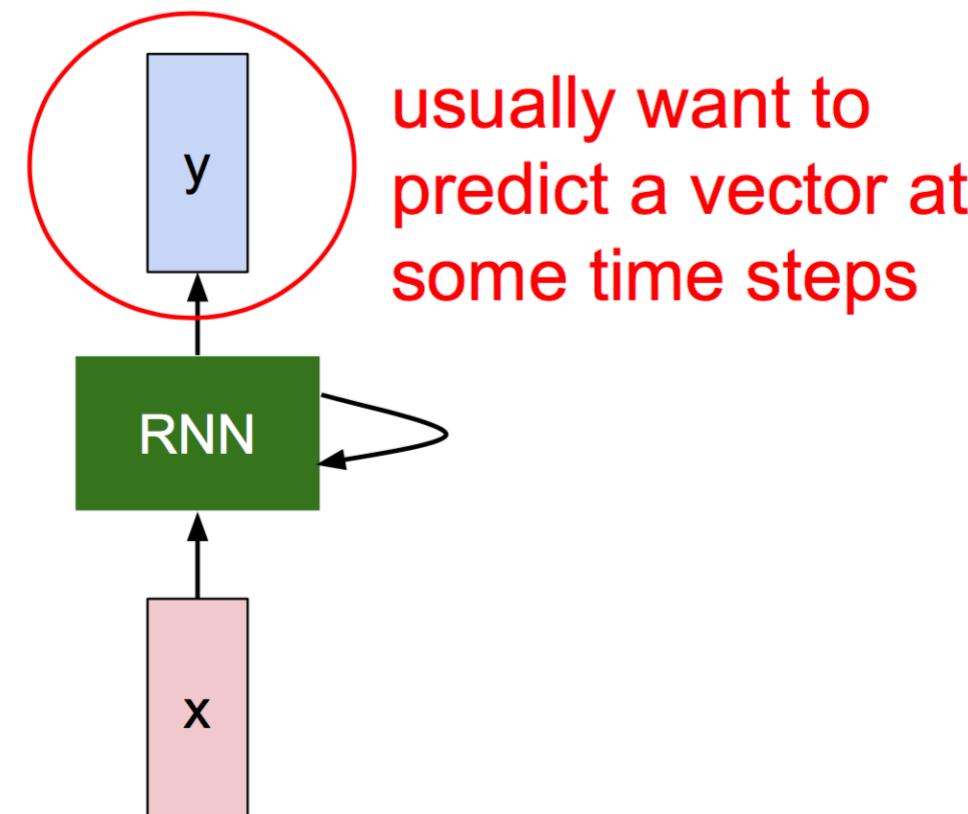
# 文本表示学习

- **循环神经网络 ( Recurrent Neural Networks, RNN )**



# 文本表示学习

- **循环神经网络 ( Recurrent Neural Networks, RNN )**

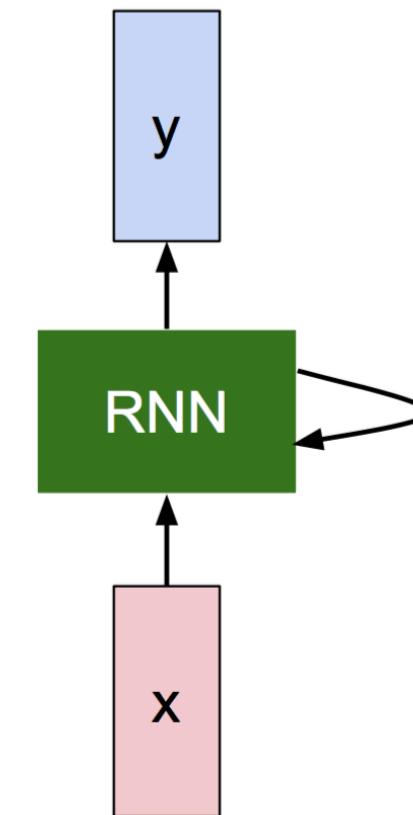


# 文本表示学习

- **循环神经网络 ( Recurrent Neural Networks, RNN )**

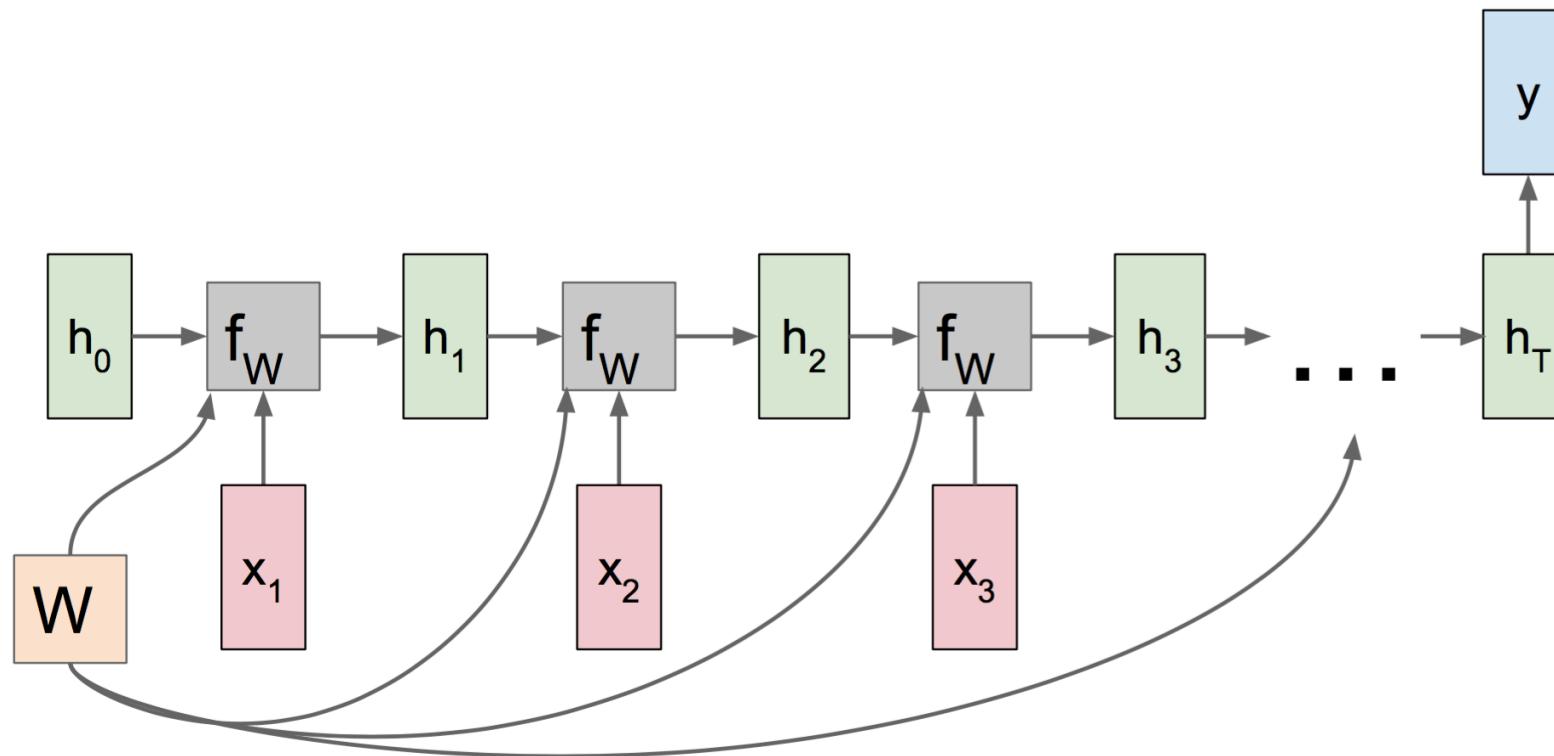
$$h_t = f_W(h_{t-1}, x_t)$$

new state      /      old state      input vector at  
some function      |      some time step  
with parameters W



# 文本表示学习

- **循环神经网络 ( Recurrent Neural Networks, RNN )**



# CONTENTS

机器学习基本概念（回顾）

视觉信息

自然语言

小结

END

# 多媒体信息表示

Representations for Multimedia Information



Quark



俞俊、高飞、谭敏、余宙、匡振中

{yujun, gaofei, tanmin, yuz, zzkuang}@hdu.edu.cn

<http://mil.hdu.edu.cn>