机器学习

朴素贝叶斯

涂文婷
tu.wenting@mail.shufe.edu.cn

回顾贝叶斯理论

- 贝叶斯理论

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)}$$

3Blue1Brown
中国官方账号

[【官方双语】贝叶斯定理，使概率论直觉化]

贝叶斯理论

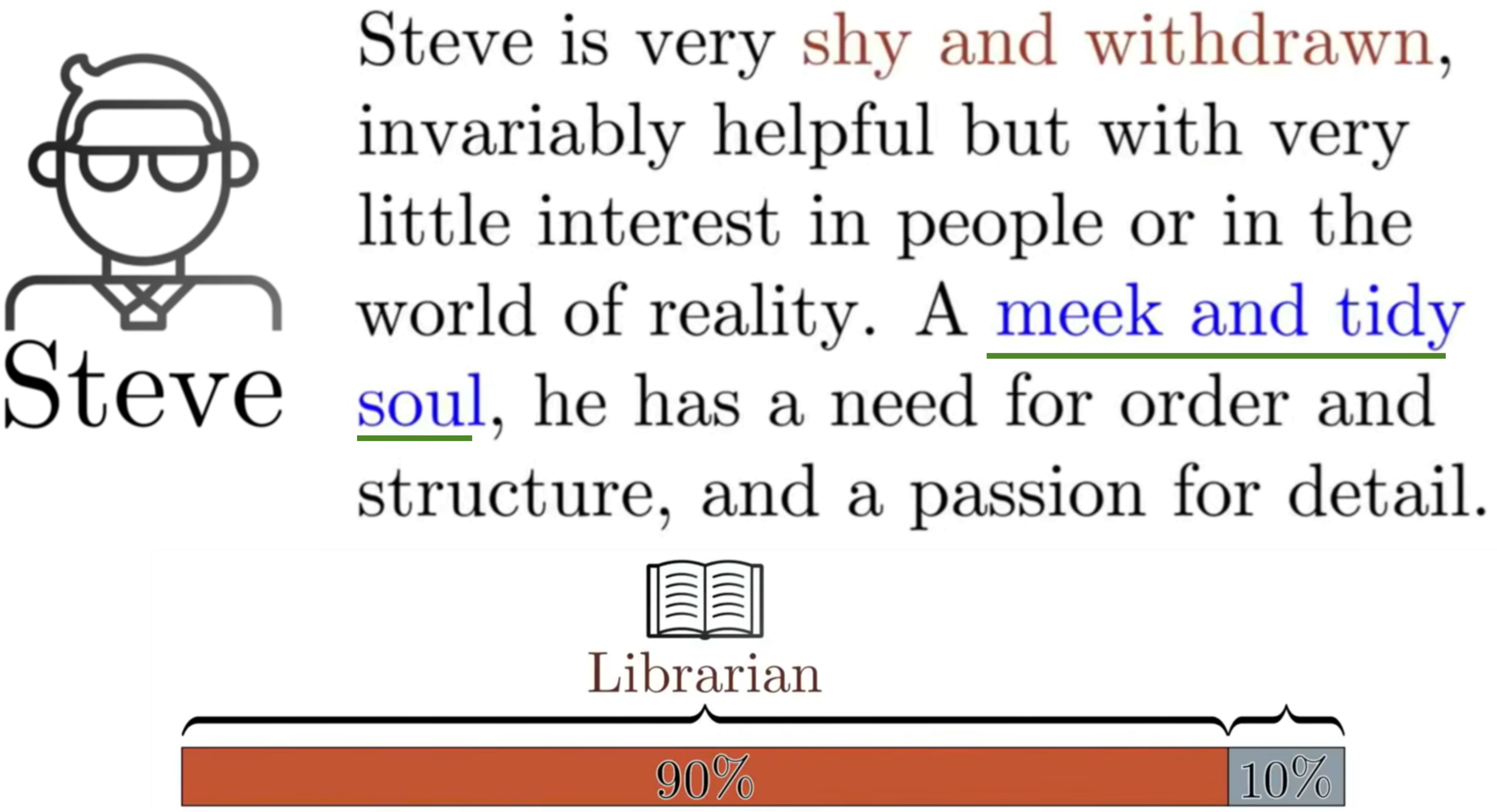
- Steve的身份

Steve

Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

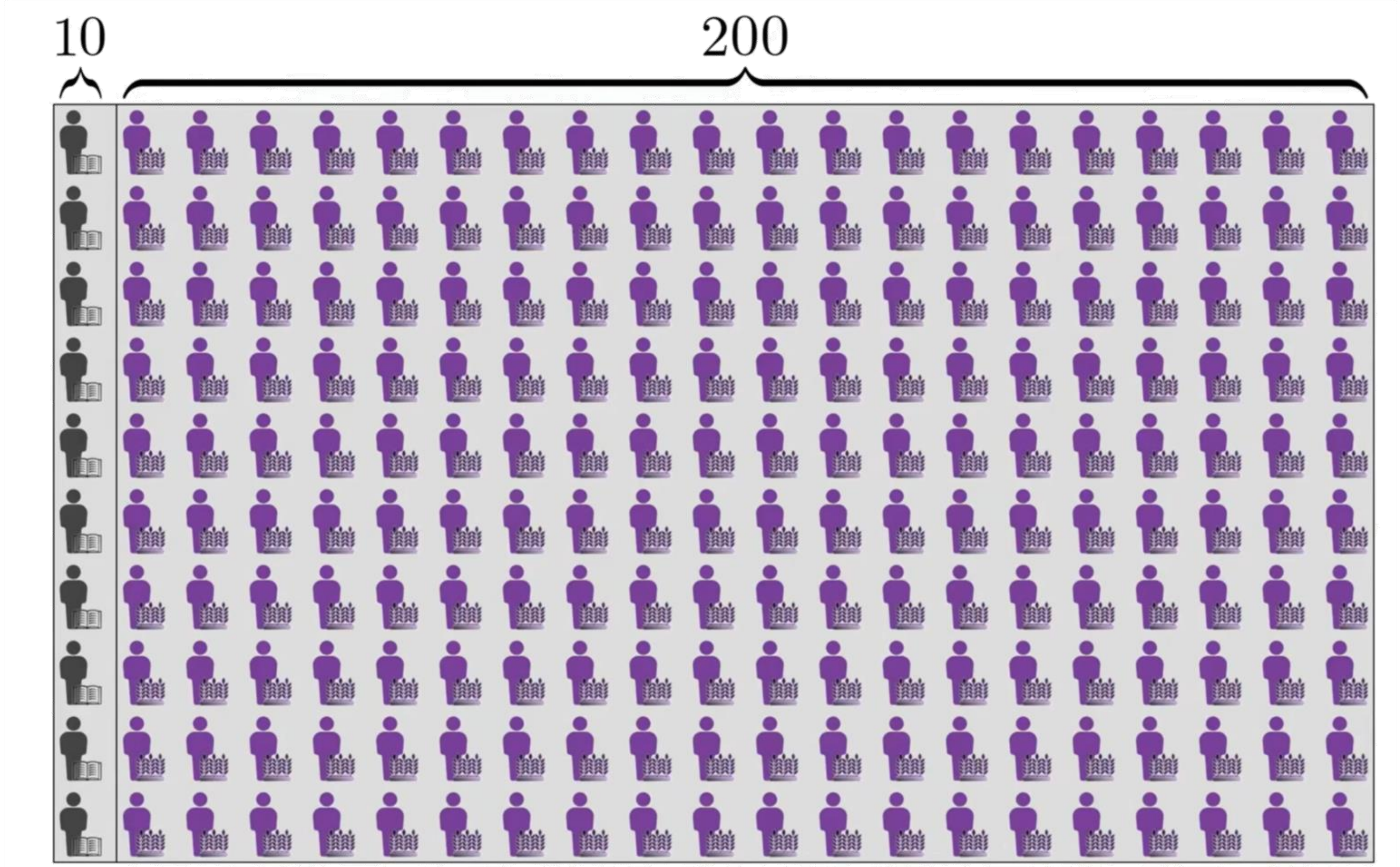
贝叶斯理论

- Steve的身份



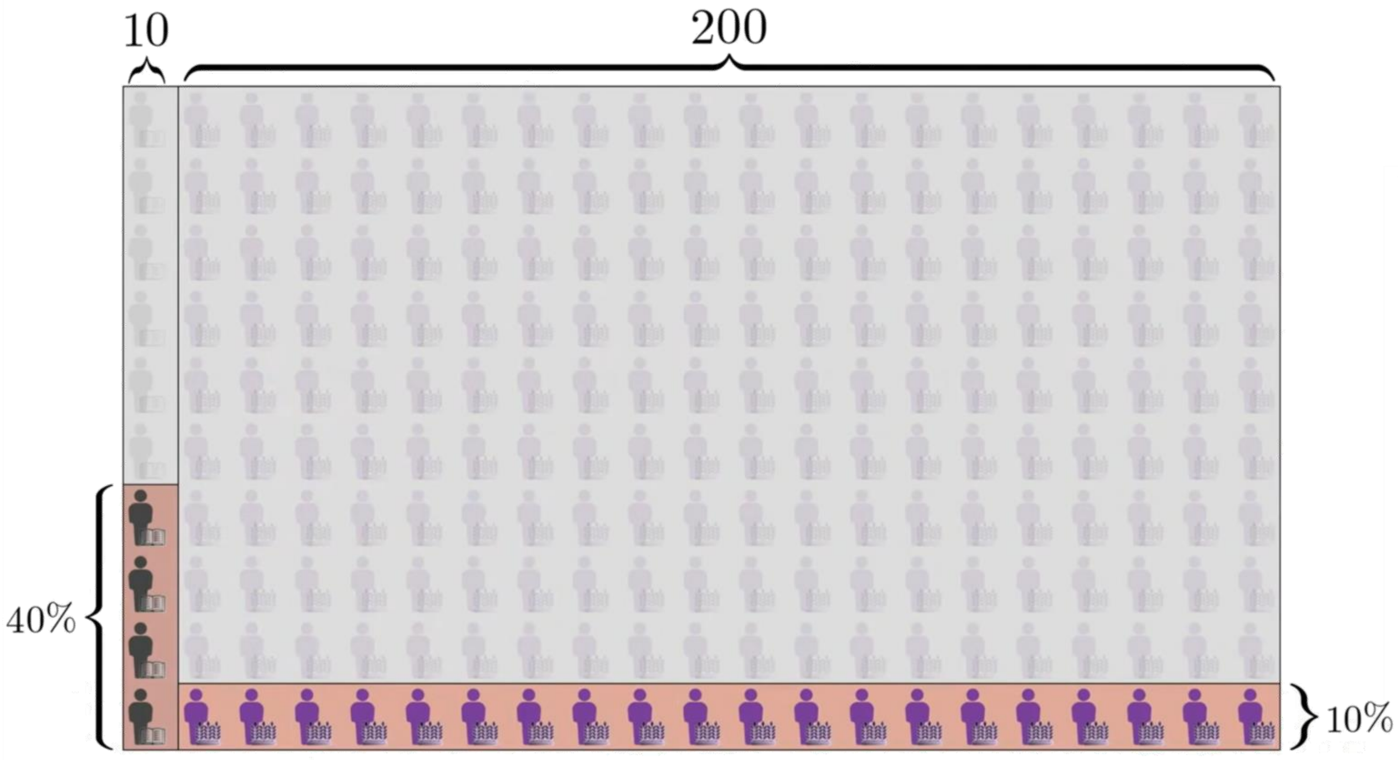
贝叶斯理论

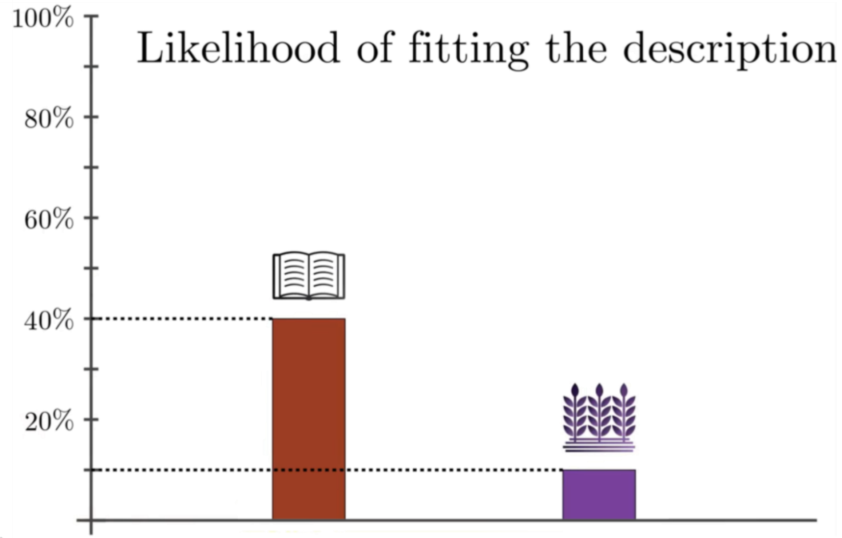
- Steve的身份



贝叶斯理论

- Steve的身份　　A meek and tidy soul

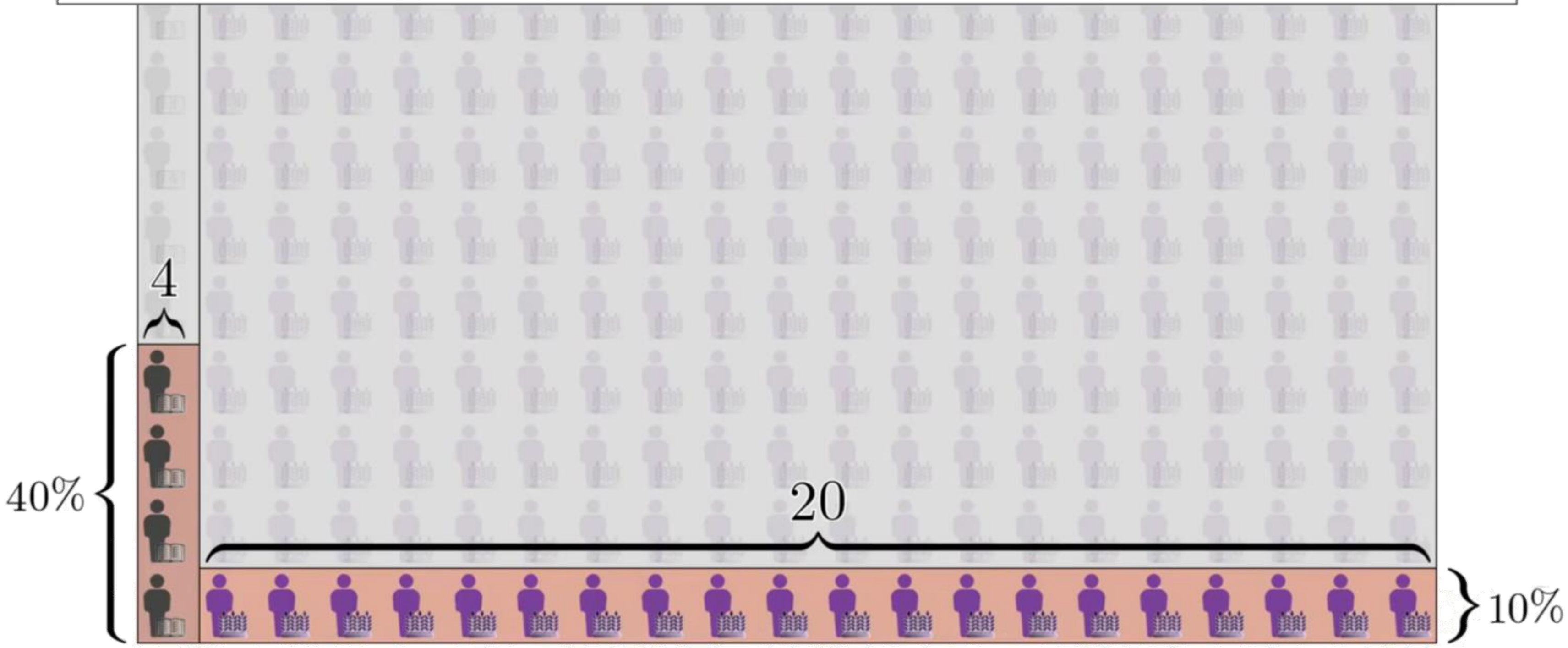




贝叶斯理论

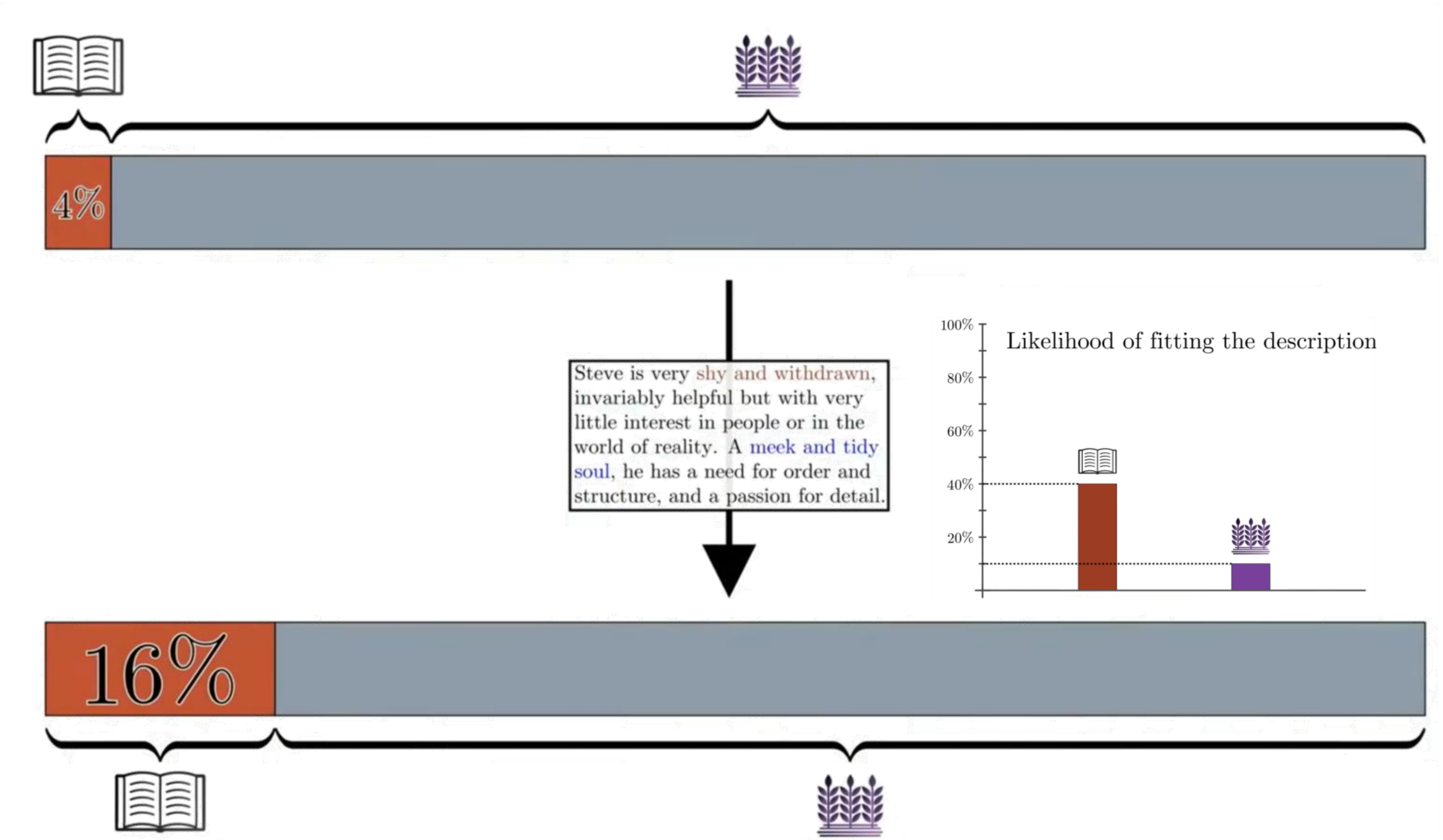
- Steve的身份

$$P\,(\text{Librarian given description}) = \frac{4}{4+20} \approx 16.7\%$$



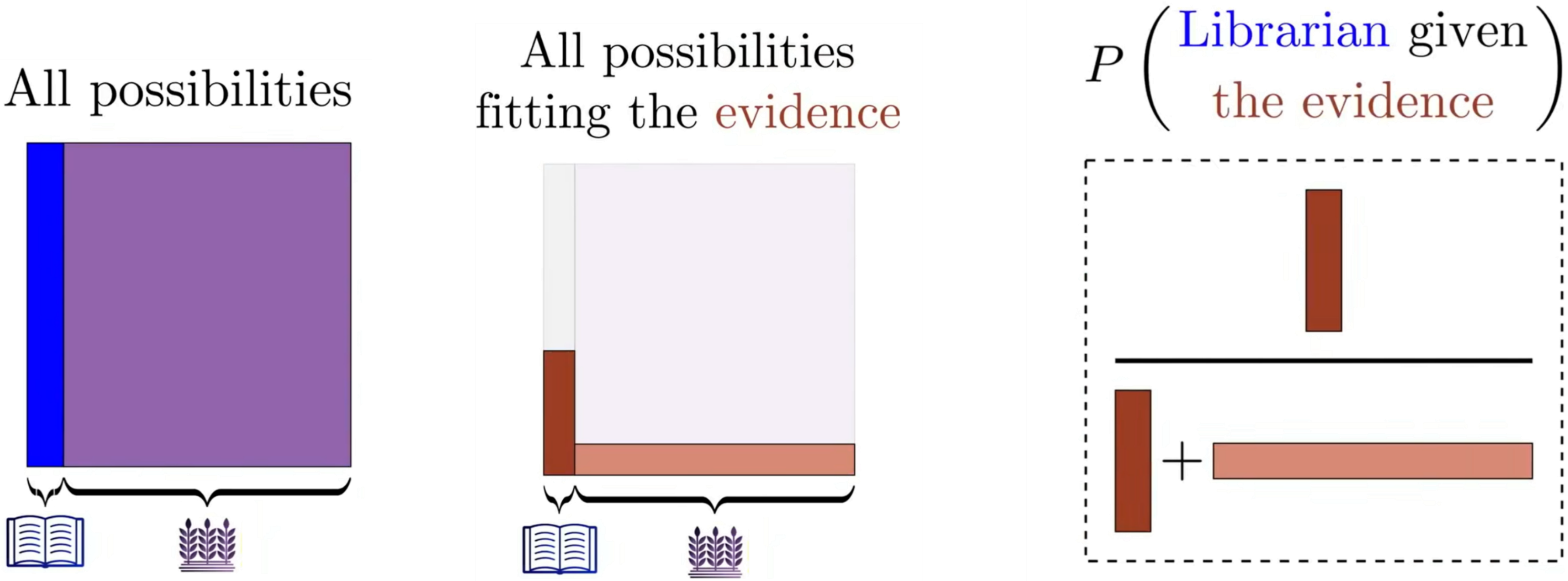
贝叶斯理论

- Steve的身份



贝叶斯理论

- Steve的身份



贝叶斯理论

- 贝叶斯公式

You have a hypothesis

You've observed some evidence

Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

You want

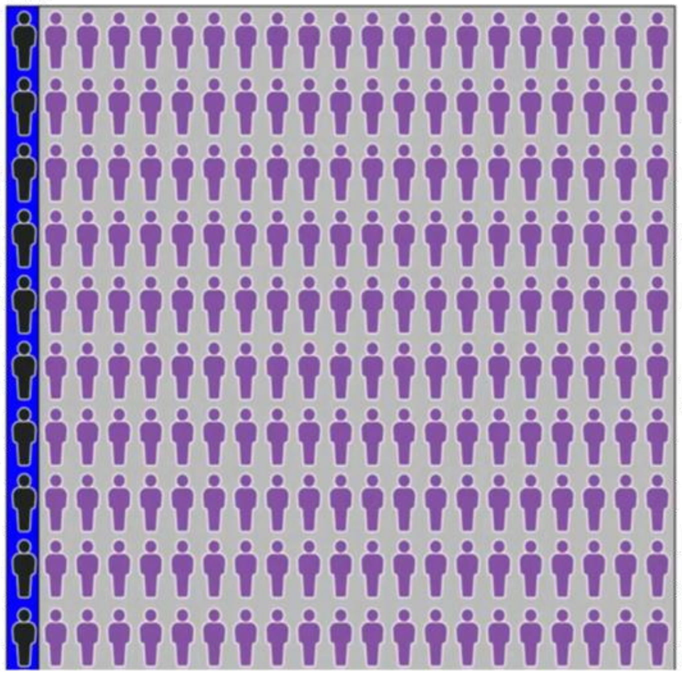
$$P\left(\begin{array}{c}\text{Hypothesis}\\ \textbf{given}\\ \text{the evidence}\end{array}\right)$$

$$P(H \mid E)$$

贝叶斯理论

- Goal: $P(H|E)$

"Prior" $\longrightarrow$ $P(H) = 1/21$



贝叶斯理论

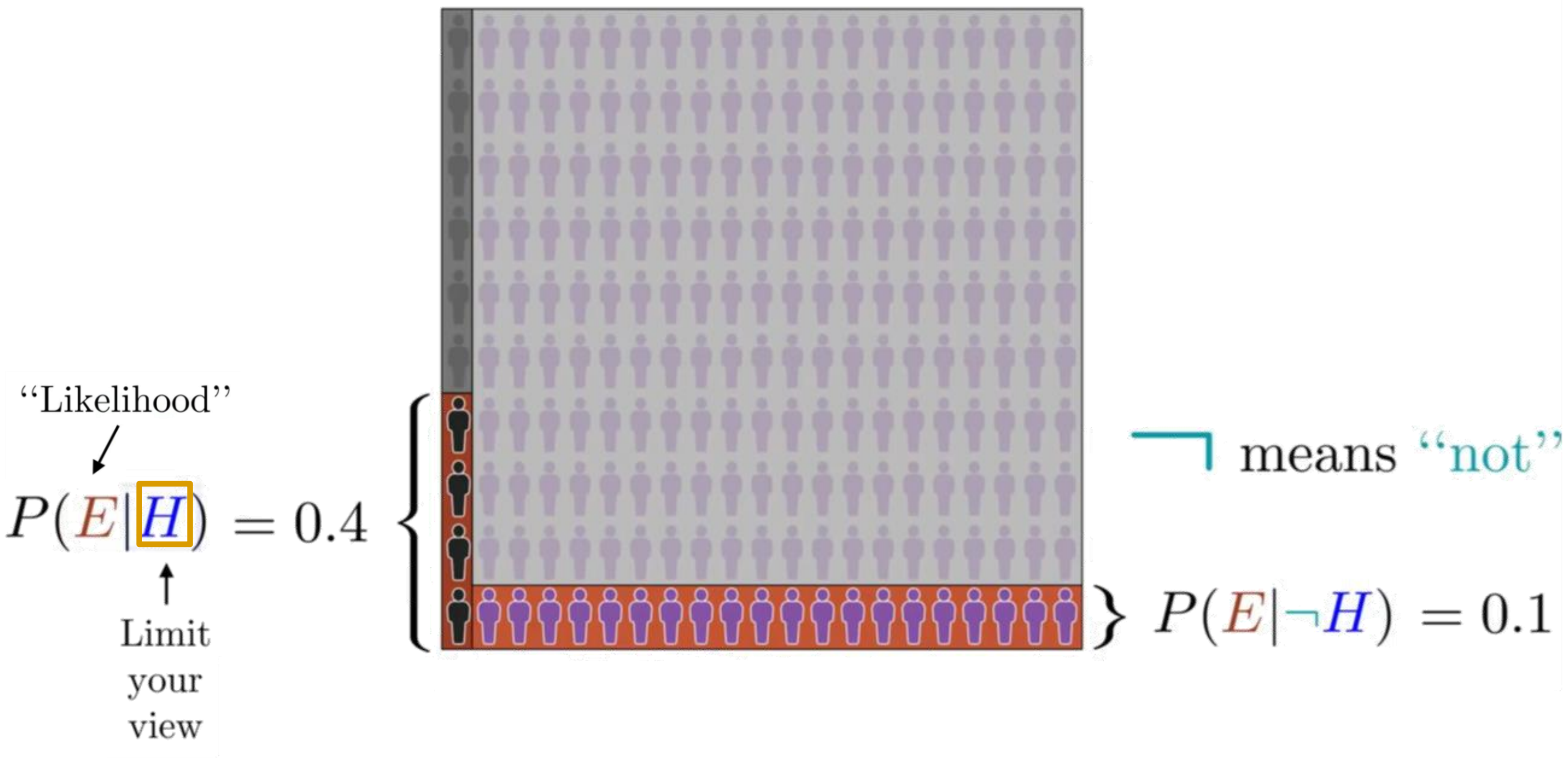
- Goal: $P(H|E)$

“Likelihood”

$$P(E|H) = 0.4$$

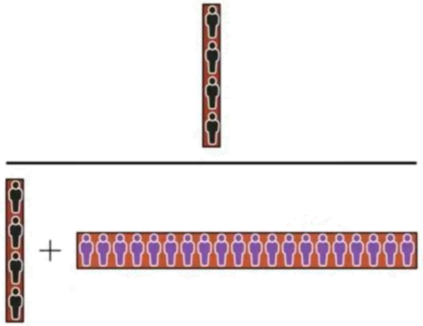
贝叶斯理论

- Goal: $P(H|E)$



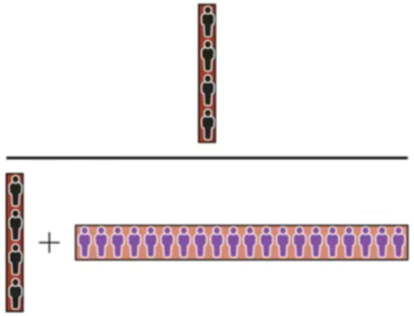
贝叶斯理论

- Goal: $P(H|E)$

$P(H|E) =$ 

贝叶斯理论

- Goal: $P(H|E)$

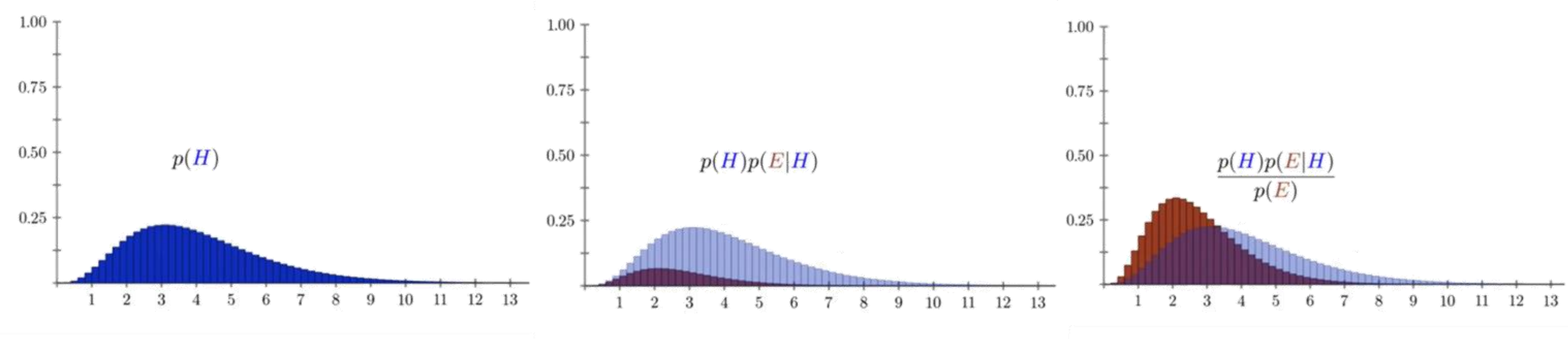
$P(H|E) =$  $= \dfrac{\cancel{(\#)}\,P(H)P(E|H)}{\cancel{(\#)}\,P(H)P(E|H) + \cancel{(\#)}\,P(\neg H)P(E|\neg H)}$

Bayes' theorem

"Posterior" $$P(H|E) = \boxed{\frac{P(H)P(E|H)}{P(E)}} = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(\neg H)P(E|\neg H)}$$

贝叶斯理论

- Goal: $P(H|E)$



什么是朴素贝叶斯算法

- 贝叶斯决策论

最小化分类错误率的贝叶斯最优分类器为

$$h^*(\boldsymbol{x}) = \underset{c \in \mathcal{Y}}{\arg\max}\, P(c \mid \boldsymbol{x})$$

即对每个样本$\boldsymbol{x}$，选择能使后验概率$P(c \mid \boldsymbol{x})$最大的类别标记

根据贝叶斯定理：

$$P(c \mid \boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x} \mid c)}{P(\boldsymbol{x})}$$

$P(c)$是类“先验”概率；$P(\boldsymbol{x} \mid c)$是样本$\boldsymbol{x}$相对于类标记c的类条件概率，或称为“似然”；$P(\boldsymbol{x})$用于归一化的“证据” 因子．对给定样本$\boldsymbol{x}$，证据因子与类标记无关，因此判断$P(c \mid \boldsymbol{x})$针对哪个类别最大的问题就转化为如何基于训练数据来估计先验$P(c)$)和似然$P(\boldsymbol{x} \mid c)$.

什么是朴素贝叶斯算法

- 朴素贝叶斯算法估计$P(c)$

类先验概率 $P(c)$表达了样本空间中各类样本所占的比例，根据大数定律，当训练集包含充足的独立同分布样本时， $P(c)$可通过各类样本出现的频率来进行估计：

$$\hat{p}(c) = \frac{|D_c|}{|D|}$$

什么是朴素贝叶斯算法

- 朴素贝叶斯算法估计$P(\boldsymbol{x} \mid c)$

估计后验概率$P(c \mid \boldsymbol{x})$的主要困难在于：类条件概率$P(\boldsymbol{x} \mid c)$ 是所有属性上的联合概率，难以从有限的训练样本直接估计而得.

朴素贝叶斯算法的精髓在于：
假设所有属性在给定类别的条件下相互独立 —属性条件独立性假设 ” *attribute conditional independence assumption*

$\hat{p}(\boldsymbol{x} \mid c)= \prod_{i=1}^{d} \hat{p}(x_i \mid c)$

什么是朴素贝叶斯算法

- 朴素贝叶斯算法估计$P(x_i \mid c)$
  - 当x_i为离散变量时，根据大数定律：

$$\hat{p}(x_i \mid c) = \frac{|D_{c,x_i}|}{|D_c|}$$

什么是朴素贝叶斯算法

- 朴素贝叶斯算法估计$P(x_i \mid c)$
  - 当x_i为连续变量时，利用最大似然法：

$$\hat{p}\,(\,x_i \mid c\,) \sim \mathcal{N}\left(\mu_{c,i}, \sigma_{c,i}^2\right), \text{记} \hat{\boldsymbol{\theta}}_{c,i} = \{\mu_{c,i}, \sigma_{c,i}^2\}$$

抽取训练集D中第c类样本的x_i值组成$D_{c,i} = \{\dots \mathrm{x}_i \dots\}$，在样本独立同分布假设下：

$$P\left(D_{c,i} \mid \boldsymbol{\theta}_{c,i}\right) = \prod_{\mathrm{x}_i \in D_{c,i}} P\left(\mathrm{x}_i \mid \boldsymbol{\theta}_{c,i}\right)$$

$$LL\left(\boldsymbol{\theta}_{c,i}\right) = \sum_{\mathrm{x}_i \in D_{c,i}} \log P\left(\mathrm{x}_i \mid \boldsymbol{\theta}_{c,i}\right)$$

然后估计

$$\hat{\boldsymbol{\theta}}_{c,i} = \underset{\boldsymbol{\theta}_{c,i}}{\arg\max}\, LL\left(\boldsymbol{\theta}_{c,i}\right)$$

有结论：

$$\mu_{c,i} = \frac{1}{\left|D_{c,i}\right|} \sum_{x_i \in D_{c,i}} \mathrm{x}_i$$

$$\sigma_{c,i}^2 = \frac{1}{\left|D_{c,i}\right|} \sum_{\mathrm{x}_i \in D_{c,i}} \left(\mathrm{x}_i - \hat{\mu}_{c,i}\right)\left(\mathrm{x}_i - \hat{\mu}_{c,i}\right)^{\mathrm{T}}$$

什么是朴素贝叶斯算法

- 拉普拉斯修正

$$\hat{p}(x_i \mid c) = \frac{|D_{c,x_i}|}{|D_c|} \xrightarrow[\text{Laplacian correction}]{\text{smoothing}} \hat{p}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

$$\hat{p}(c) = \frac{|D_c|}{|D|} \xrightarrow[\text{Laplacian correction}]{\text{smoothing}} \hat{p}(c) = \frac{|D_c| + 1}{|D| + N}$$

Multinomial作为似然分布
假设下的最大似然估计

Multinomial作为似然分布，
Dirichlet 作为先验分布
假设下的最大后验估计
（后验分布也为
Multinomial）

什么是朴素贝叶斯算法

- Naïve Bayes in Sklearn

  https://scikit-learn.org/stable/modules/naive_bayes.html

  参考实践篇：利用sklearn里的Naïve Bayes来实现西瓜分类

扩展：贝叶斯决策论

- 最大后验 → 最小风险

N种可能的类别标记：$\mathcal{Y}=\{c_1,c_2,\dots,c_N\}$
将一个真实标记为c_j 的样本误分类为c_i 所产生的损失：λ_{ij}
后验概率：$P(c_i \mid \boldsymbol{x})$

→

将样本$\boldsymbol{x}$分类为c_i所产生的期望损失(expected loss)，即在样本$\boldsymbol{x}$ 上的“条件风险”(conditional risk)：

$$R(c_i \mid \boldsymbol{x})=\sum_{j=1}^{N}\lambda_{ij}P(c_j \mid \boldsymbol{x}).$$

$$h^*(\boldsymbol{x})=\underset{c\in\mathcal{Y}}{\arg\max}\, P(c \mid \boldsymbol{x}) \quad \rightarrow \quad h^*(\boldsymbol{x})=\underset{c\in\mathcal{Y}}{\arg\min}\, R(c \mid \boldsymbol{x})$$

$$\lambda_{ij}=\begin{cases}0, & \text{if } i=j\\ 1, & \text{otherwise}\end{cases}$$