Data Mining Techniques

CS 6220 - Section 3 - Fall 2016

Lecture 3: Probability

Jan-Willem van de Meent
(*credit*: Zhao, CS 229, Bishop)

Project Vote

1. *Freeform*: Develop your own project proposals
   - 30% of grade (homework 30%)
   - Present proposals after midterm
   - Peer-review reports
2. *Predefined*: Same project for whole class
   - 20% of grade (homework 40%)
   - More like a “super-homework”
   - Teaching assistants and instructors

Homework Problems

Homework 1 will be out today (due 30 Sep)

- 4 or (more likely) 5 problem sets
- 30% - 40% of grade (depends on type of project)
- Can use any language (within reason)
- **Discussion is encouraged, but submissions must be completed individually**
  (absolutely **no** sharing of code)
- Submission via *zip* file by **11.59pm** on day of deadline
  (no late submissions)
- Please follow *submission guidelines* on website
  (TA's have authority to deduct points)

Regression: Probabilistic Interpretation

Log joint probability of *N* independent data points

$$\begin{aligned}
\log p(y_1, \ldots, y_N) &= \sum_{n=1}^{N} \log p(y_n) \\
&= -\frac{1}{2}\left[N \log(2\pi\sigma^2) + \sum_{n=1}^{N} \frac{(y_n - \boldsymbol{w}^\top x_n)^2}{\sigma^2} \right] \\
&= -\frac{N}{2}\left[\text{const} + E(\boldsymbol{w})\right]
\end{aligned}$$

$$\boxed{\operatorname*{argmax}_{\boldsymbol{w}}\; p(y_1, \ldots, y_N; \boldsymbol{w}) \;=\; \operatorname*{argmin}_{\boldsymbol{w}} E(\boldsymbol{w})}$$

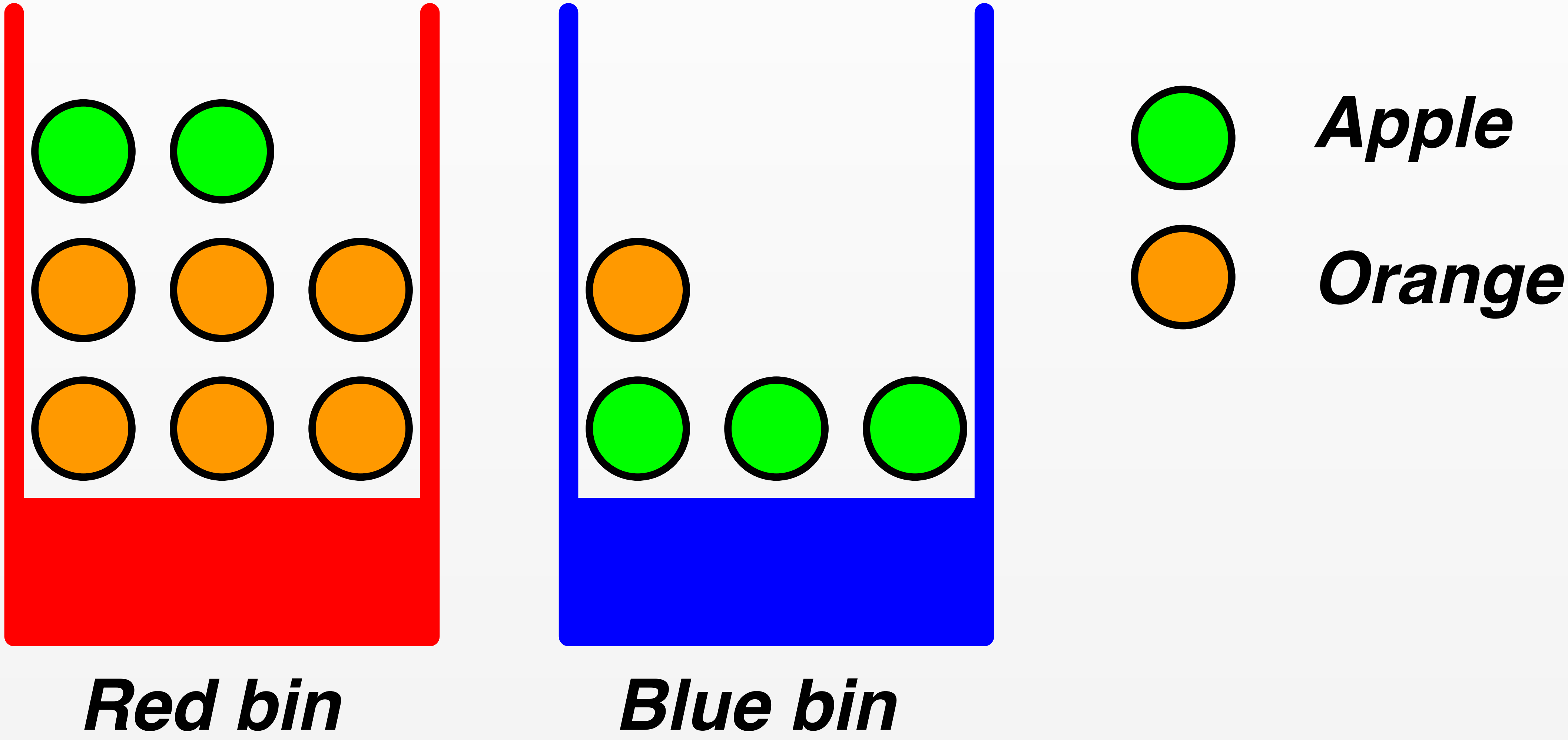
Maximum Likelihood

Probability

Examples: Independent Events

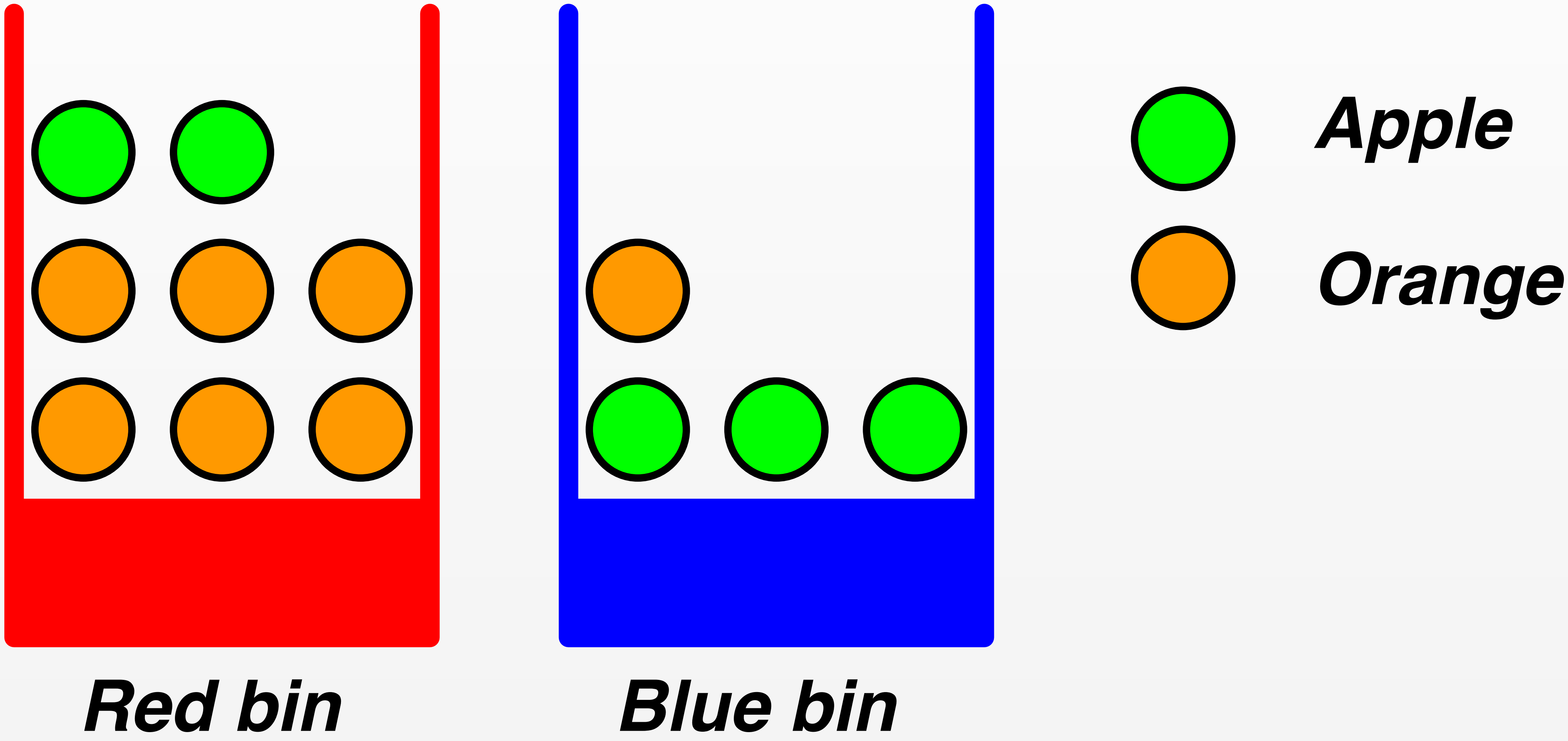
1. What's the probability of getting a sequence of 1,2,3,4,5,6 if we roll a dice six times?
2. A school survey found that 9 out of 10 students like pizza. If three students are chosen at random with replacement, what is the probability that all three students like pizza?

Dependent Events



If I take a fruit from the ***red*** *bin,*
what is the probability that I get an ***apple****?*

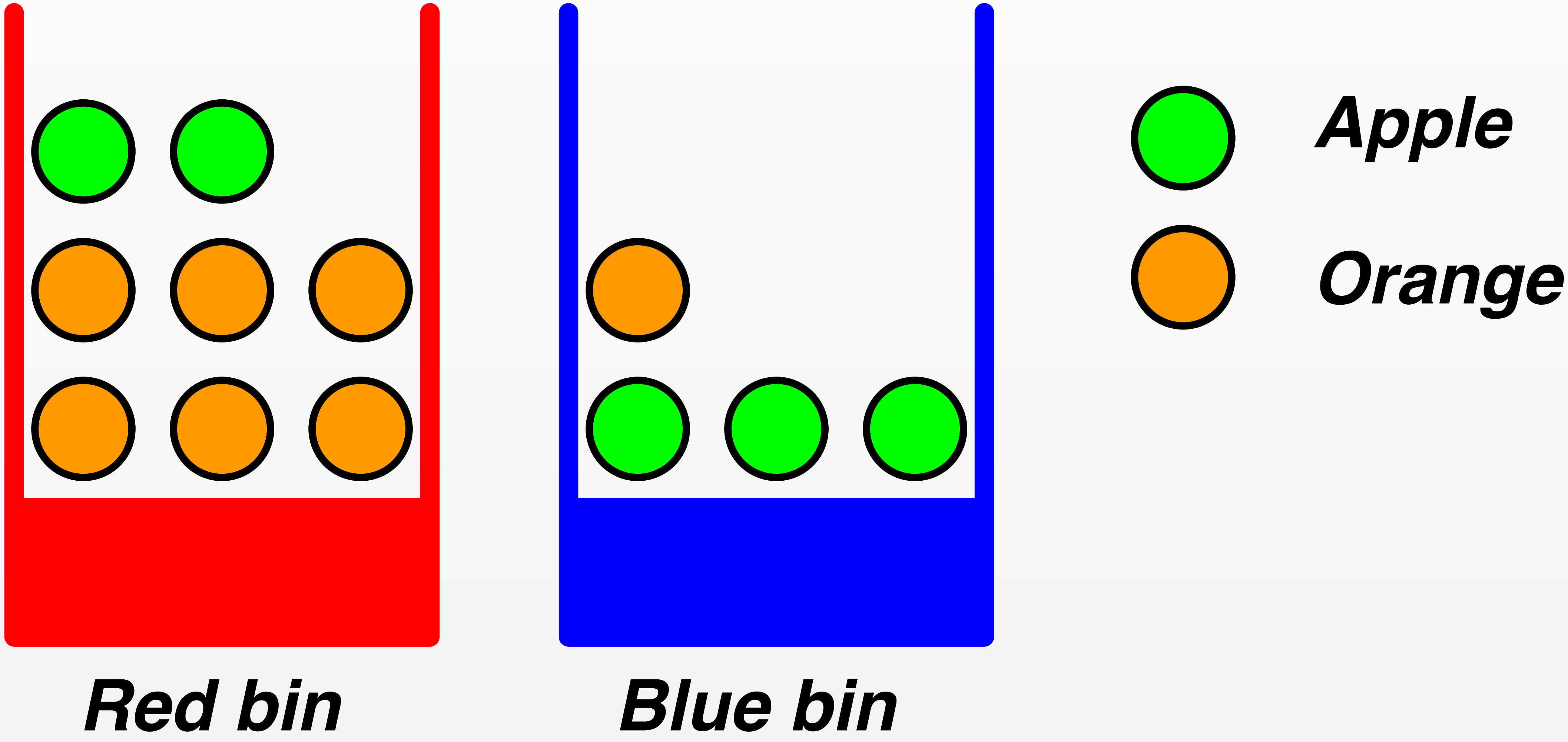
Dependent Events



Conditional Probability

P(fruit = apple | bin = red) = 2 / 8

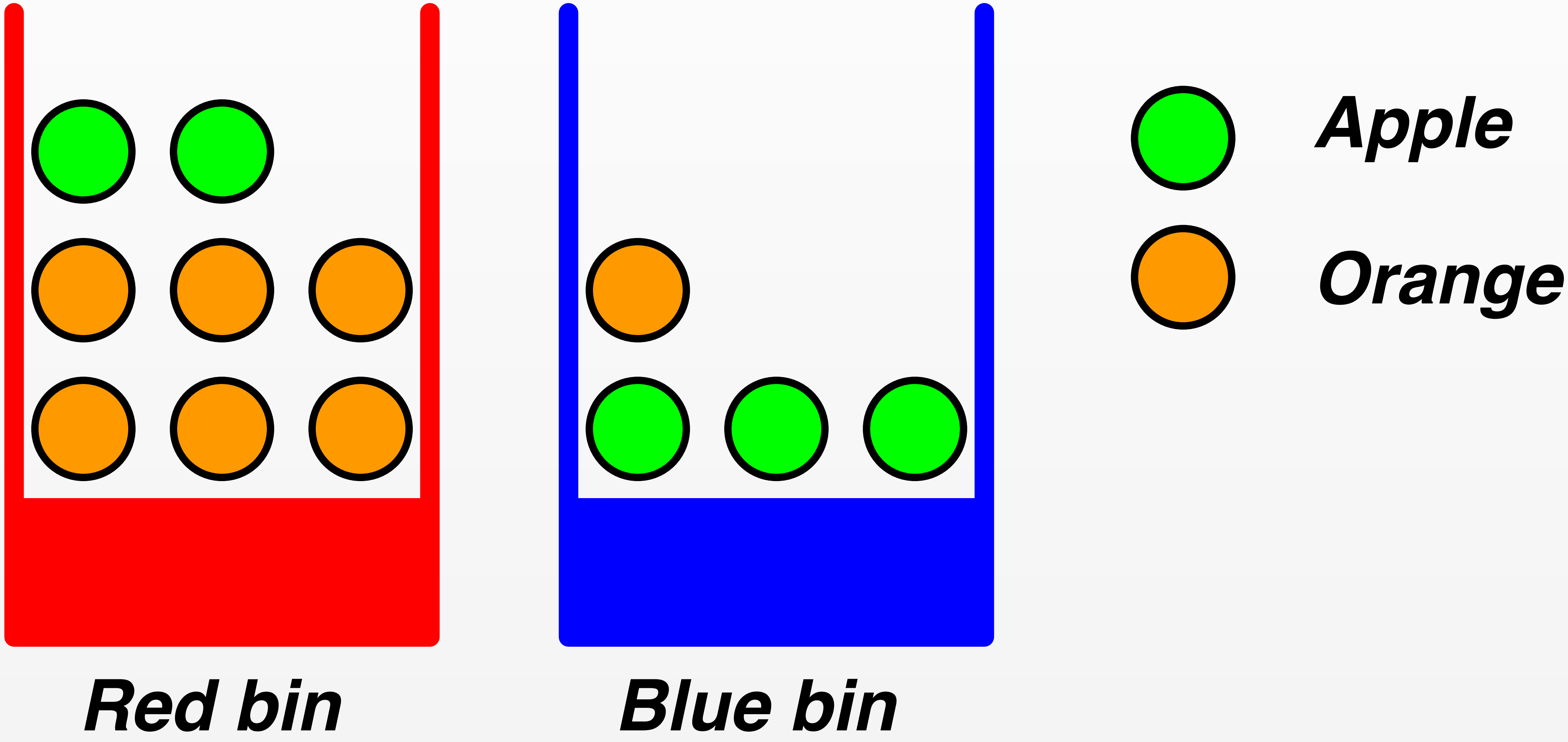
Dependent Events



Joint Probability

P(fruit = apple , bin = red) = 2 / 12

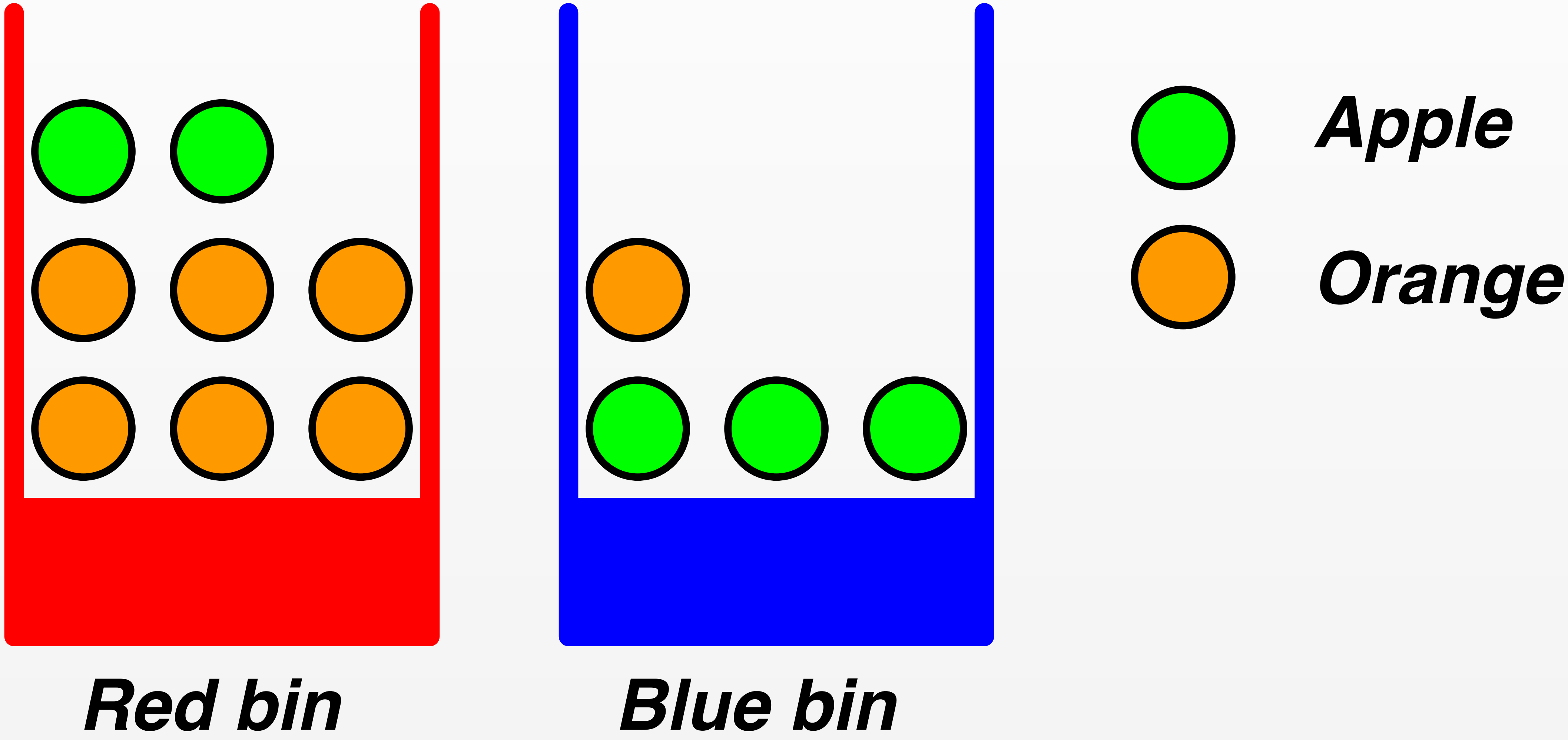
Dependent Events



Joint Probability

P(fruit = apple , bin = blue) = ?

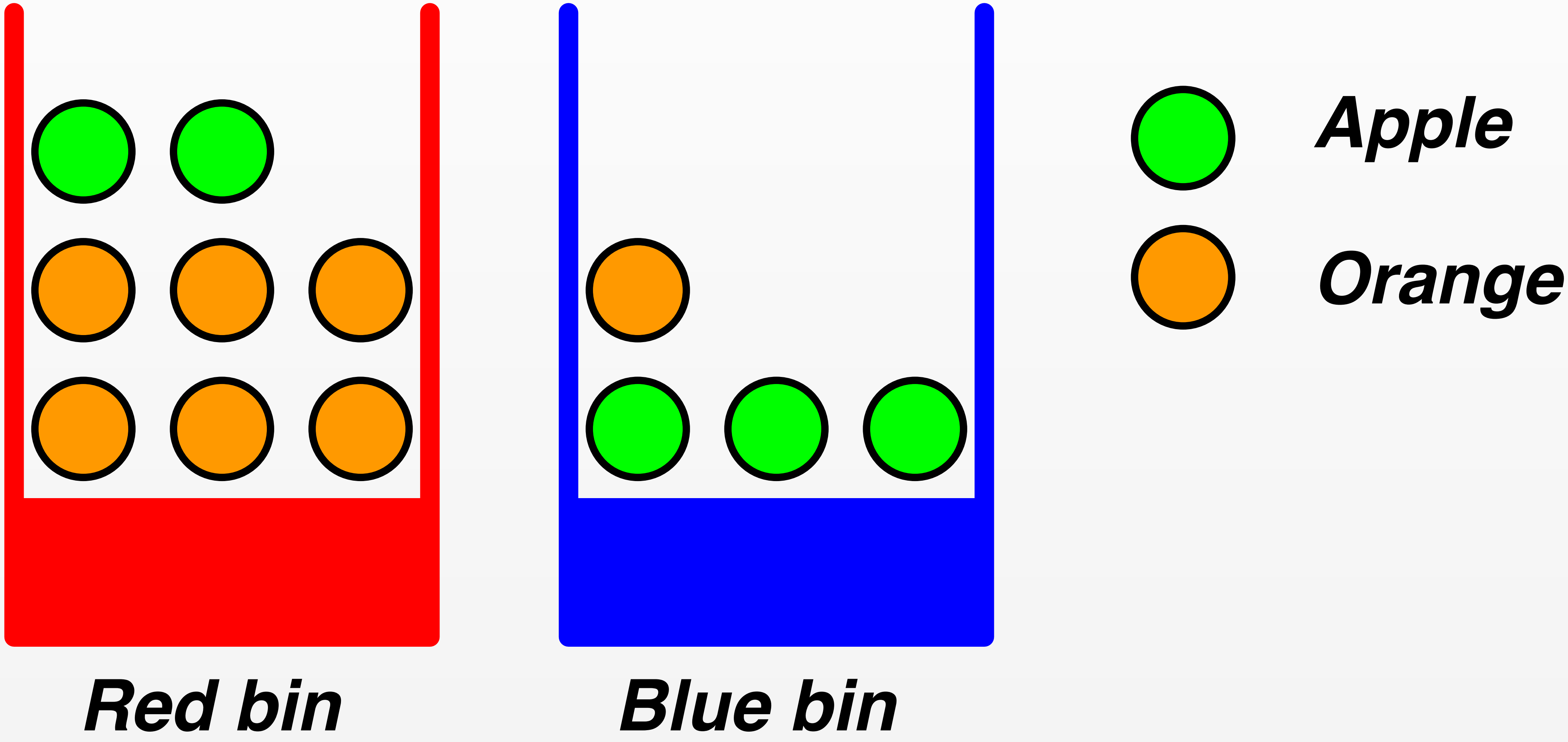
Dependent Events



Joint Probability

P(fruit = apple , bin = blue) = 3 / 12

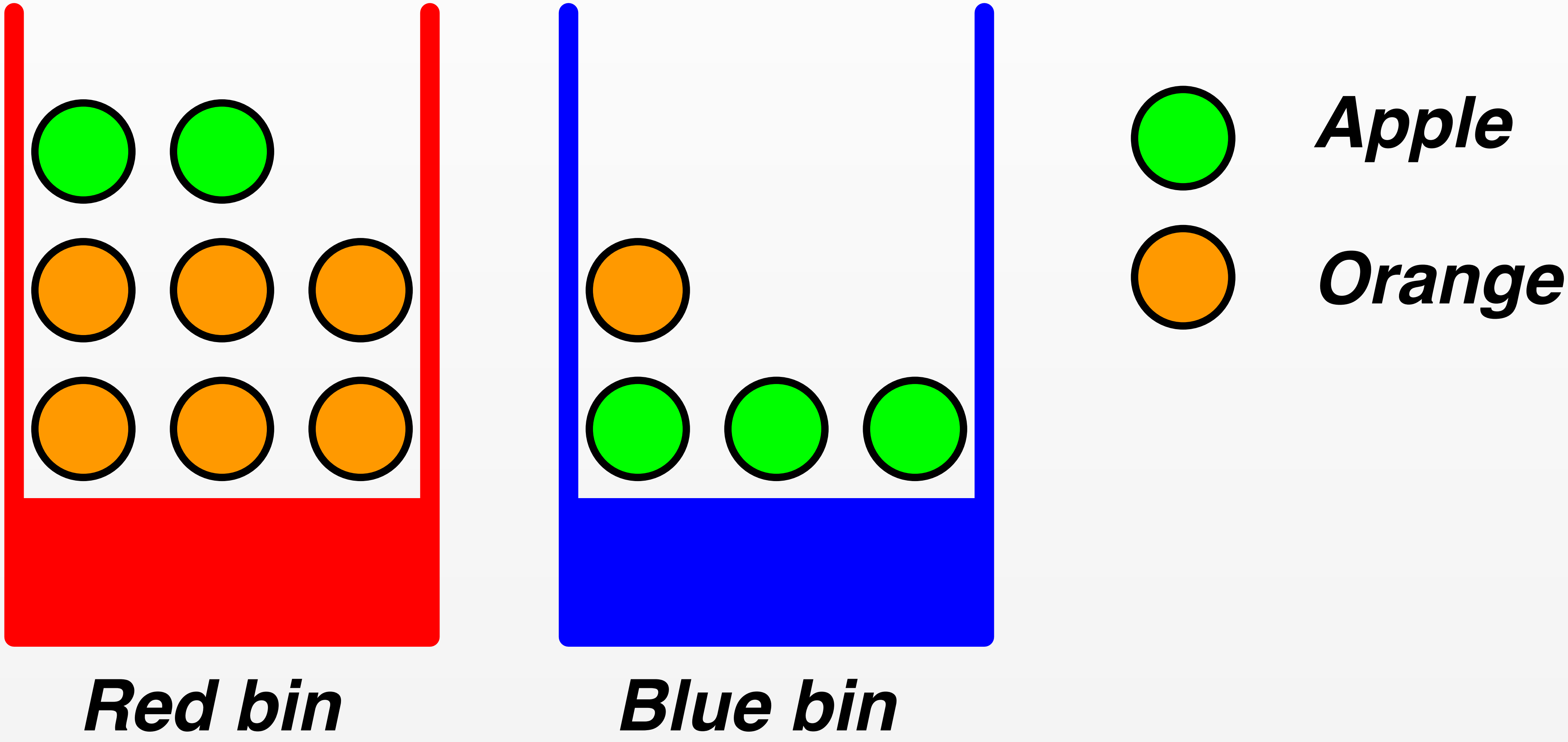
Dependent Events



Joint Probability

P(fruit = orange , bin = blue) = ?

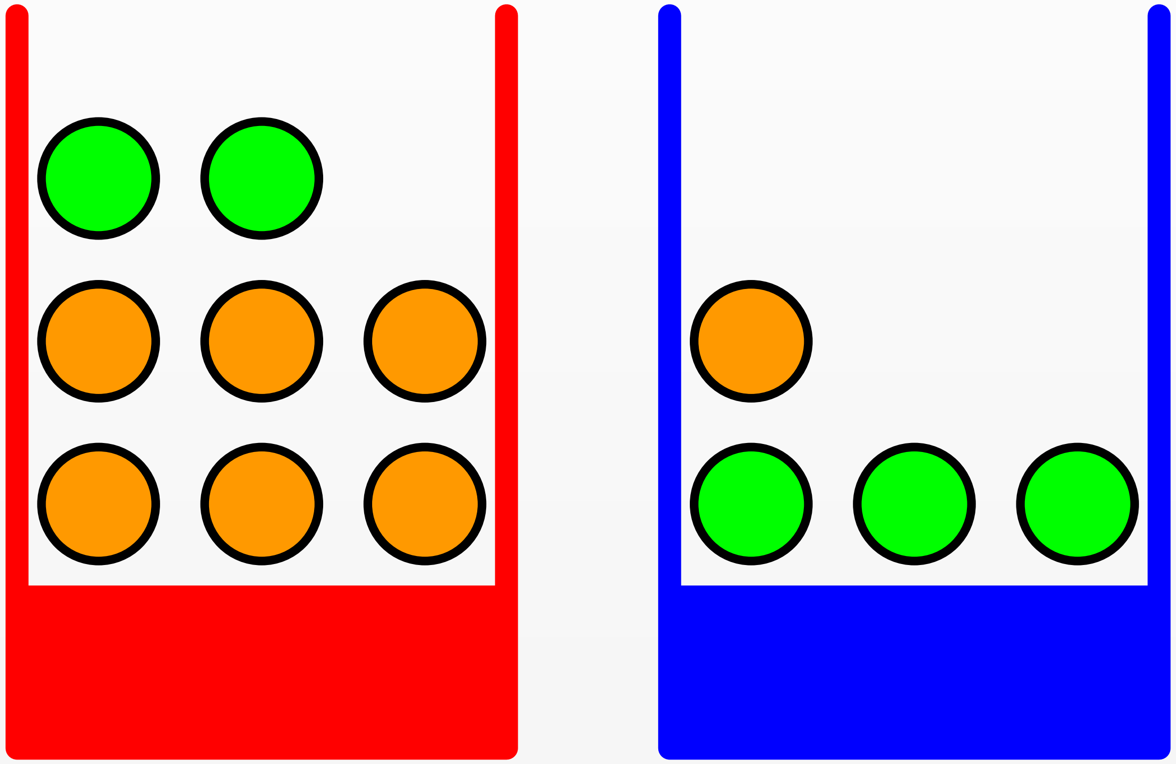
Dependent Events



Joint Probability

P(fruit = orange , bin = blue) = 1 / 12

Two rules of Probability



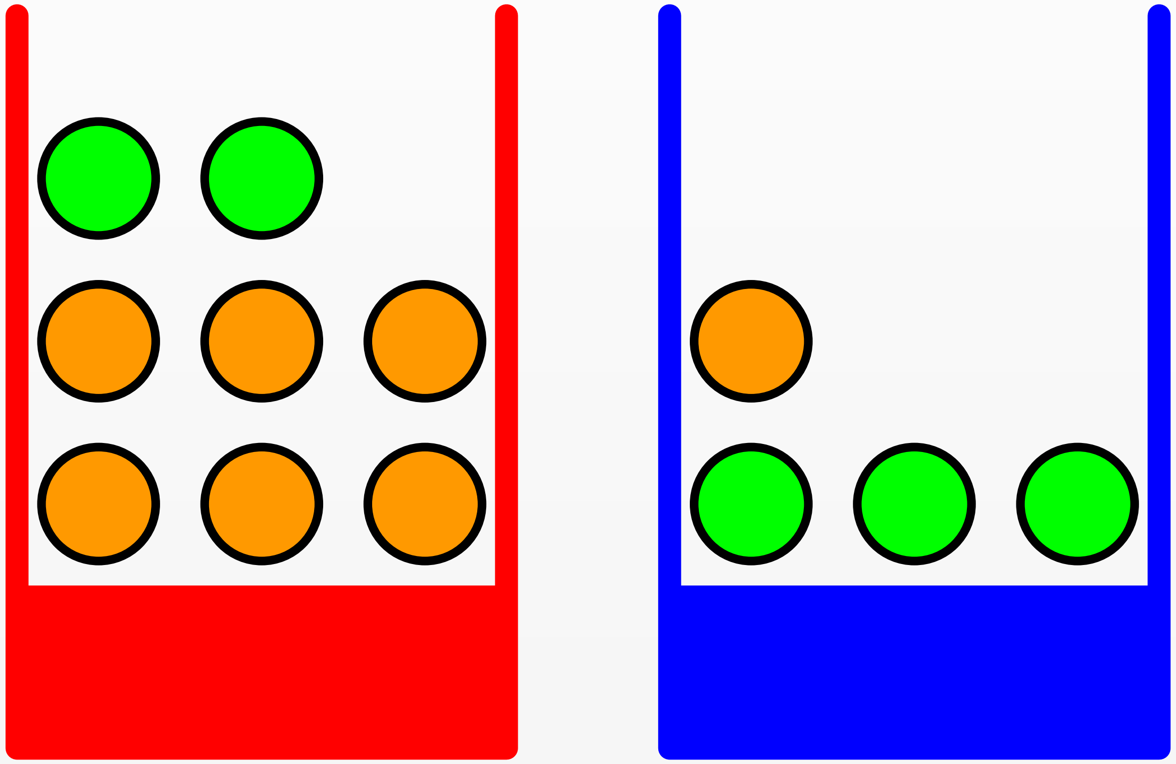
1. Sum Rule (Marginal Probabilities)

P(fruit = apple) = P(fruit = apple , bin = blue)

+ P(fruit = apple , bin = red)

= ?

Two rules of Probability



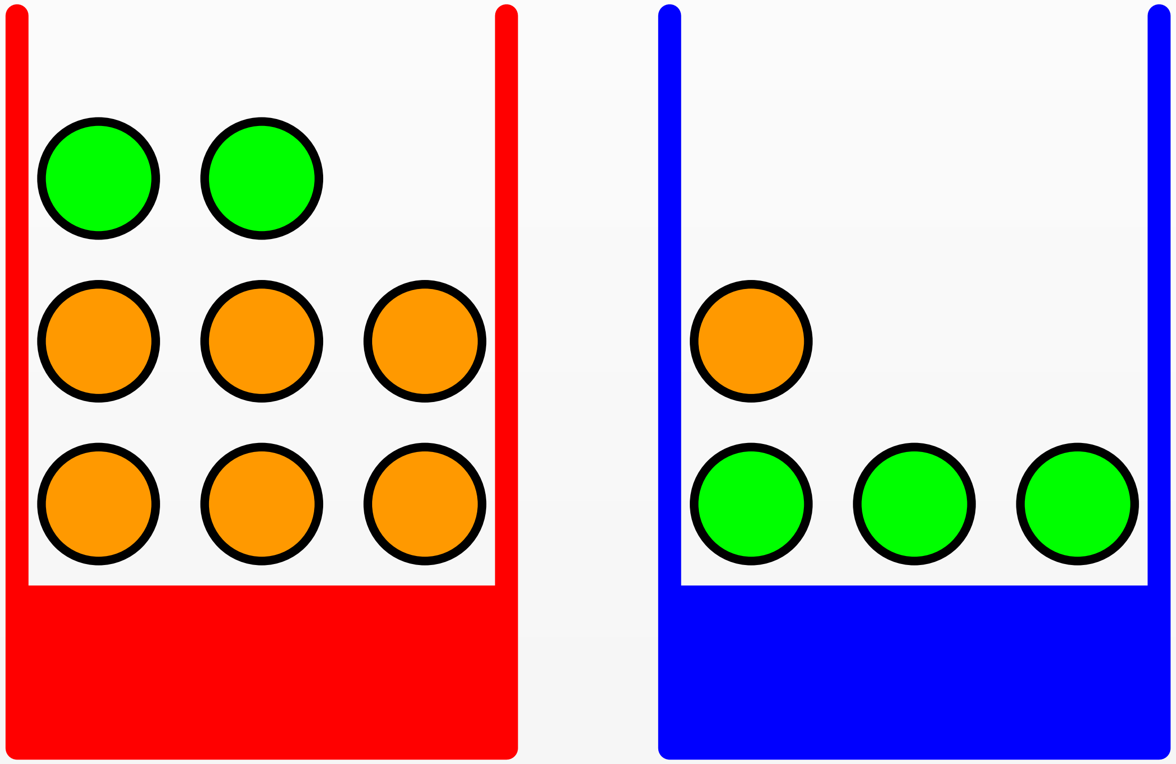
1. Sum Rule (Marginal Probabilities)

P(fruit = apple) = P(fruit = apple , bin = blue)

+ P(fruit = apple , bin = red)

= 3 / 12 + 2 / 12 = **5 / 12**

Two rules of Probability



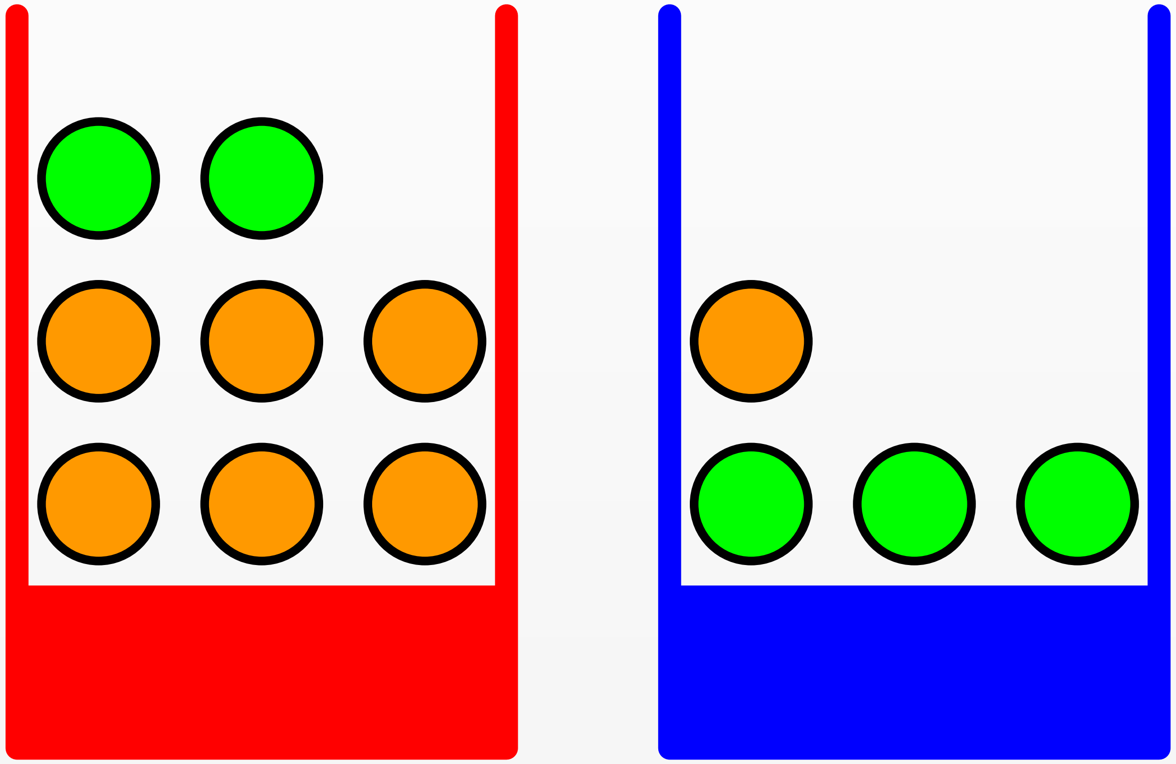
2. Product Rule

P(fruit = apple , bin = red) =

P(fruit = apple | bin = red) p(bin = red)

= ?

Two rules of Probability



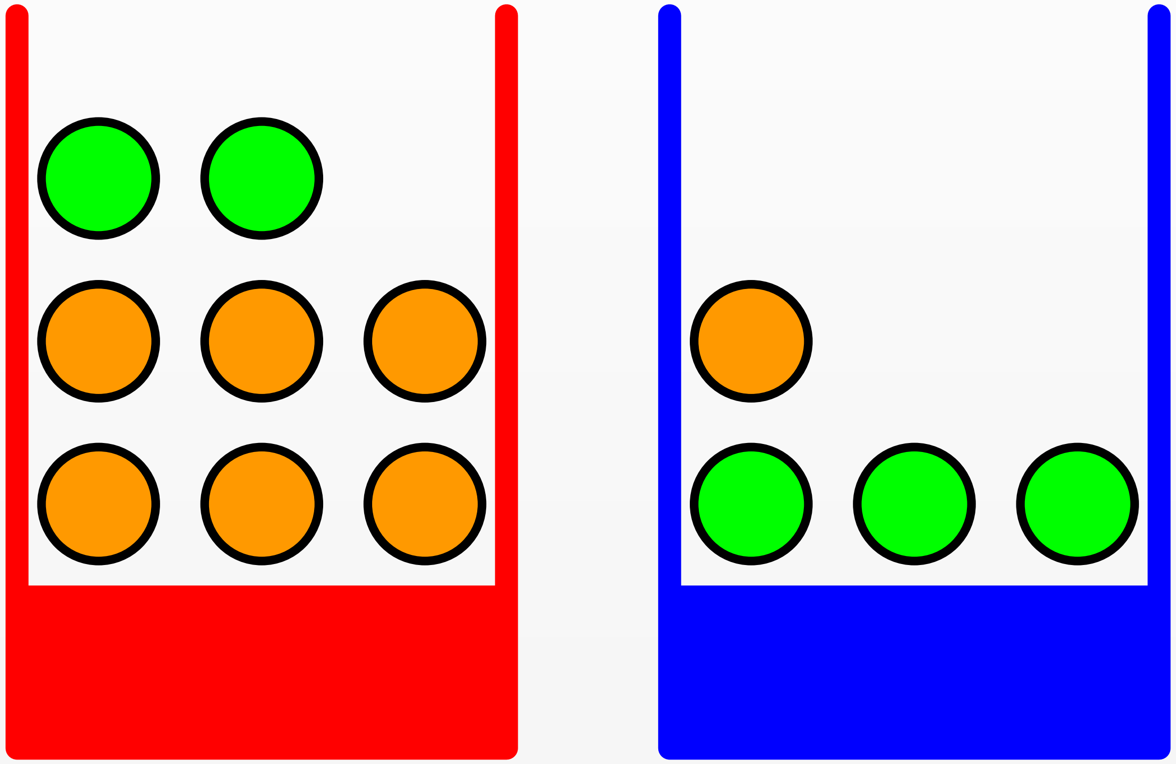
2. Product Rule

P(fruit = apple , bin = red) =

P(fruit = apple | bin = red) p(bin = red)

= 2 / 8 * 8 / 12 = **2 / 12**

Two rules of Probability



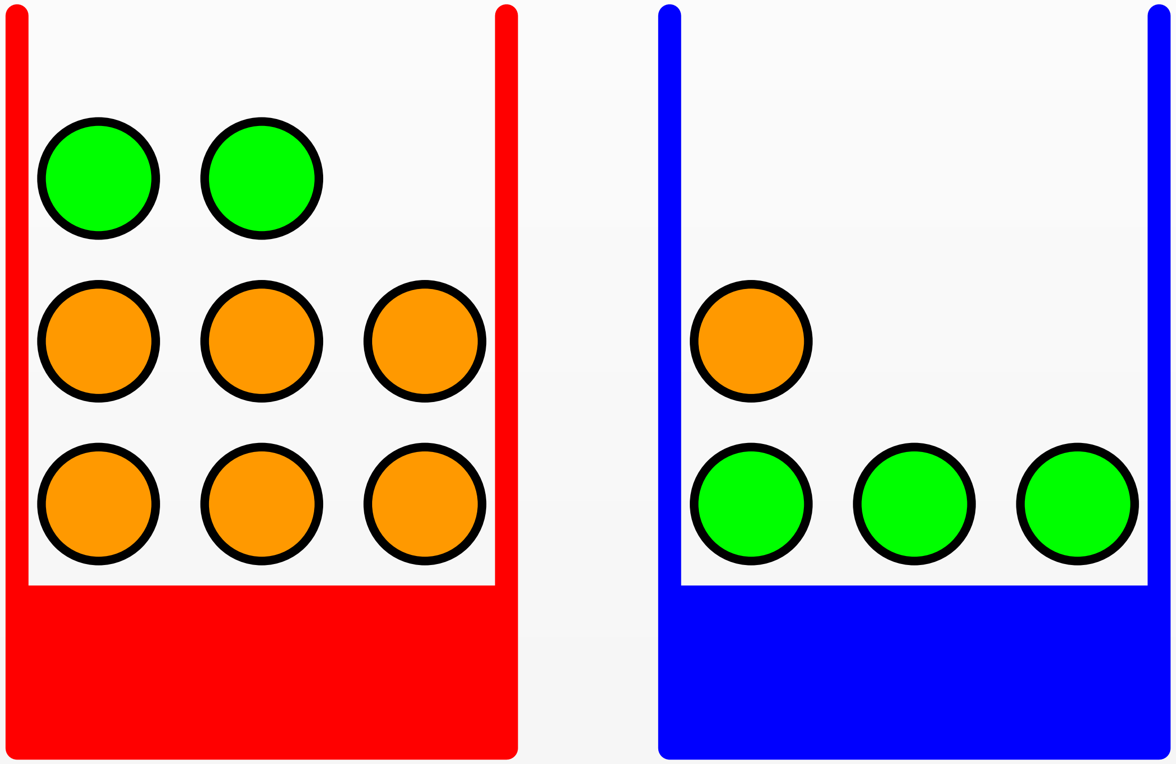
2. Product Rule (reversed)

P(fruit = apple , bin = red) =

P(bin = red | fruit = apple) p(fruit = apple)

= ?

Two rules of Probability



2. Product Rule (reversed)

P(fruit = apple , bin = red) =

P(bin = red | fruit = apple) p(fruit = apple)

= 2 / 5 * 5 / 12 = **2 / 12**

Bayes' Rule

$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \underbrace{p(y \mid x)}_{\text{Likelihood}} \underbrace{p(x)}_{\text{Prior}} / p(y)$$

Sum Rule: $p(y) = \sum_{x} p(y, x) \quad p(x) = \sum_{y} p(y, x)$

Product Rule: $p(y, x) = p(y \mid x) p(x) = p(x \mid y) p(y)$

Bayes' Rule

$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \underbrace{p(y \mid x)}_{\text{Likelihood}}\underbrace{p(x)}_{\text{Prior}}/p(y)$$

$p(x)$	*Probability of rare disease:* 0.005
$p(y \mid x)$	*Probability of detection:* 0.98 *Probability of false positive:* 0.05
$p(x \mid y)$	*Probability of disease when test positive?*

Bayes' Rule

$$\underbrace{p(x\,|\,y)}_{\text{Posterior}} = \underbrace{p(y\,|\,x)}_{\text{Likelihood}}\underbrace{p(x)}_{\text{Prior}}/p(y)$$

$$p(y,x) = p(y\,|\,x)p(x)$$

*0.99 * 0.005 = 0.00495*

$$p(y) = \sum_{x} p(y,x)$$

*0.99 * 0.005 + 0.05 * 0.995 = 0.0547*

$$p(x\,|\,y)$$

0.00495 / 0.0547 = 0.09

Measures

Elements of Probability

- *Sample space Ω*
  The set of all outcomes ω ∊ Ω of an experiment
- *Event space F*
  The set of all possible events A ∊ F, which are subsets A ⊆ Ω of possible outcomes
- *Probability Measure P*
  A function *P: F* → *R*

Axioms of Probability

- A probability measure must satisfy

  1. $P(A) \geq 0 \; \forall \; A \in F$
  2. $P(\Omega) = 1$
  3. When A_1, A_2, … disjoint

$$P(\cup_i A_i) = \sum_i P(A_i)$$

Corollaries of Axioms

- If $A \subseteq B \implies P(A) \leq P(B)$
- $P(A \cap B) \leq \min\left(P(A), P(B)\right)$
- $P(A \cup B) \leq P(A) + P(B)$ (Union Bound)
- $P(\Omega \setminus A) = 1 - P(A)$
- If $A_1, \ldots, A_k$ is a disjoint partition of Ω, then

$$\sum_{i=1}^{k} P(A_k) = 1$$

Conditional Probability

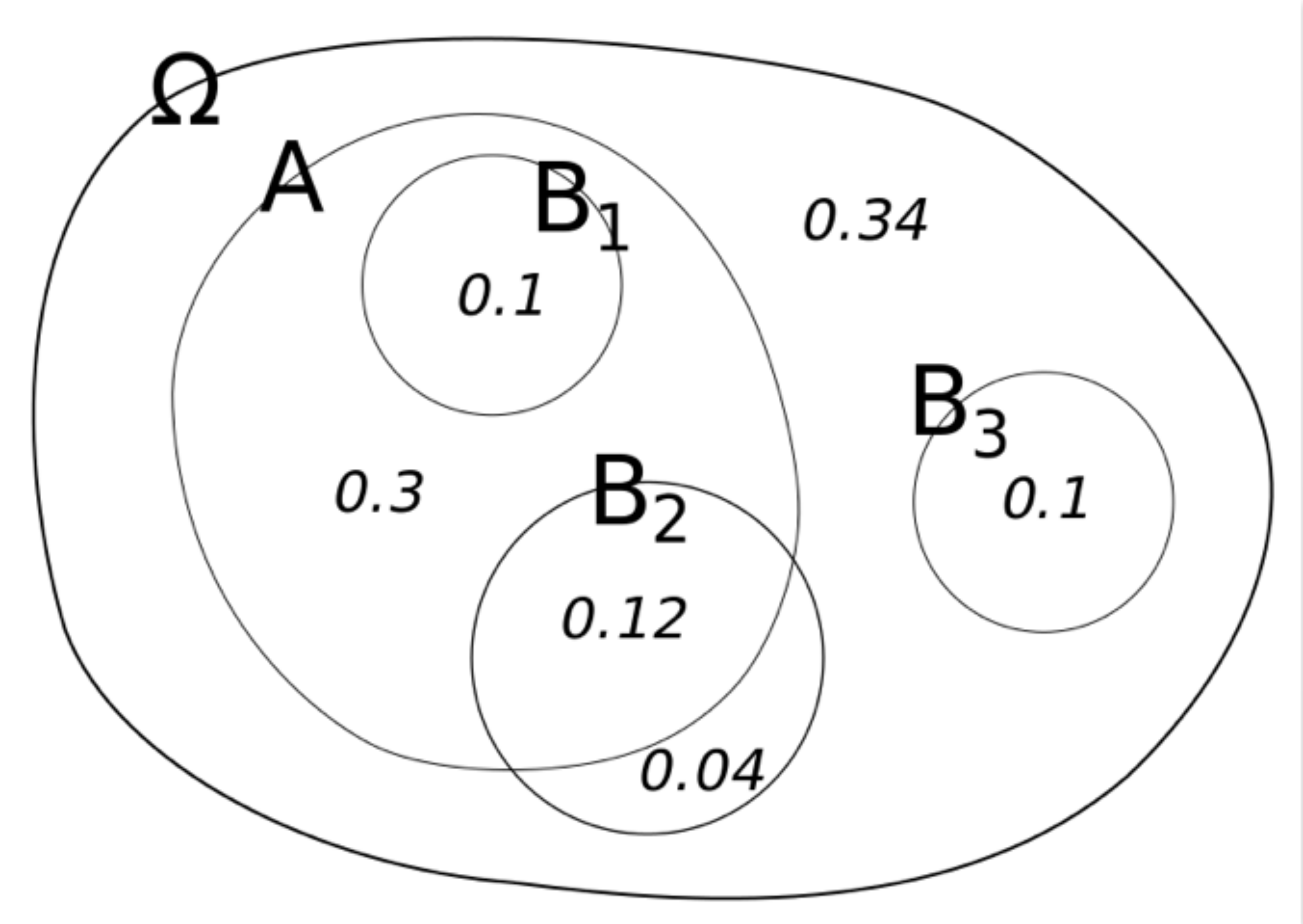
- *Conditional Probability*
  Probability of event *A*, conditioned on occurrence of event *B*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

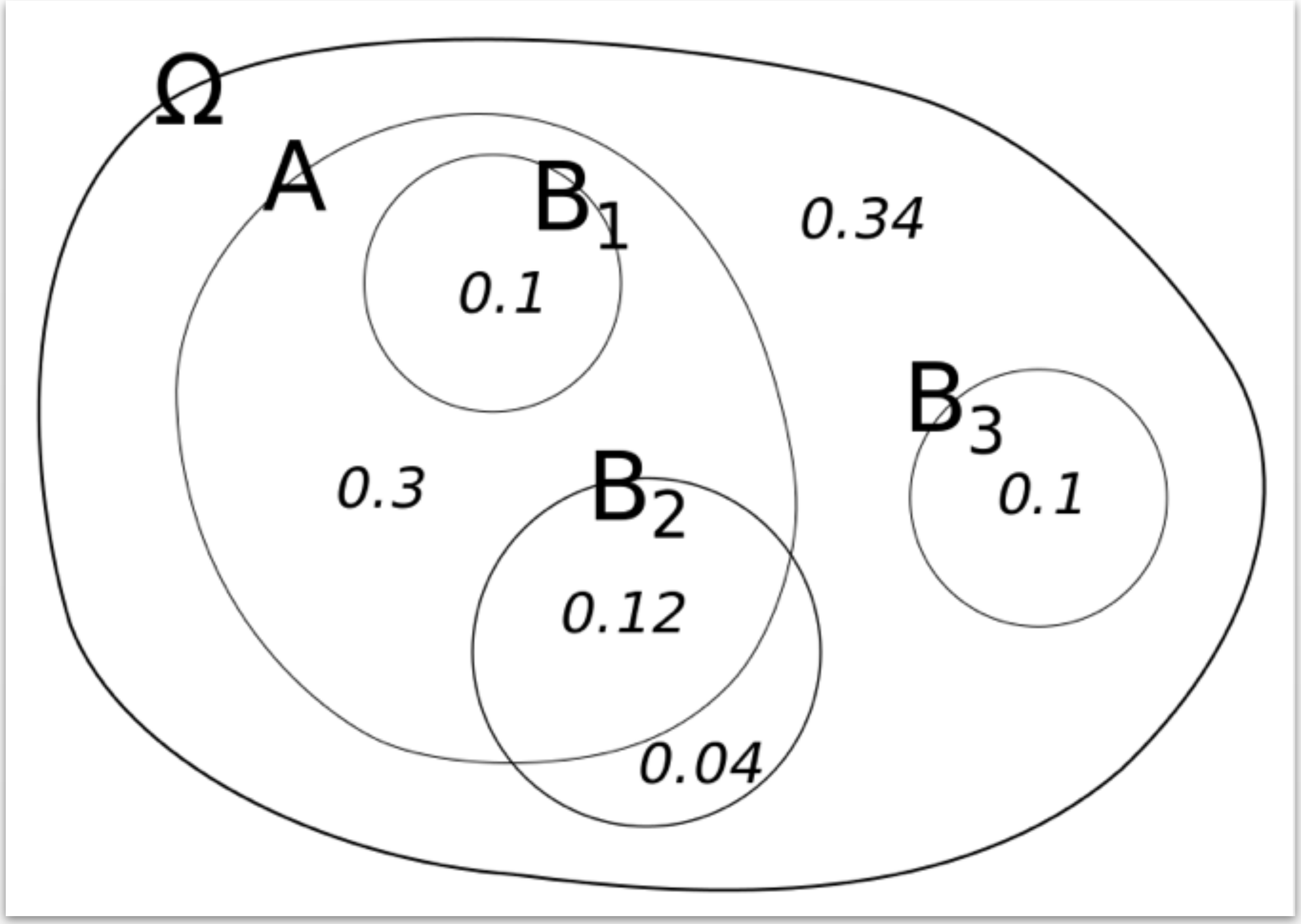
- *Conditional Independence*
  Events *A* and *B* are independent iff
  - $P(A \mid B) = P(A)$

  which implies
  - $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$

Conditional Probability

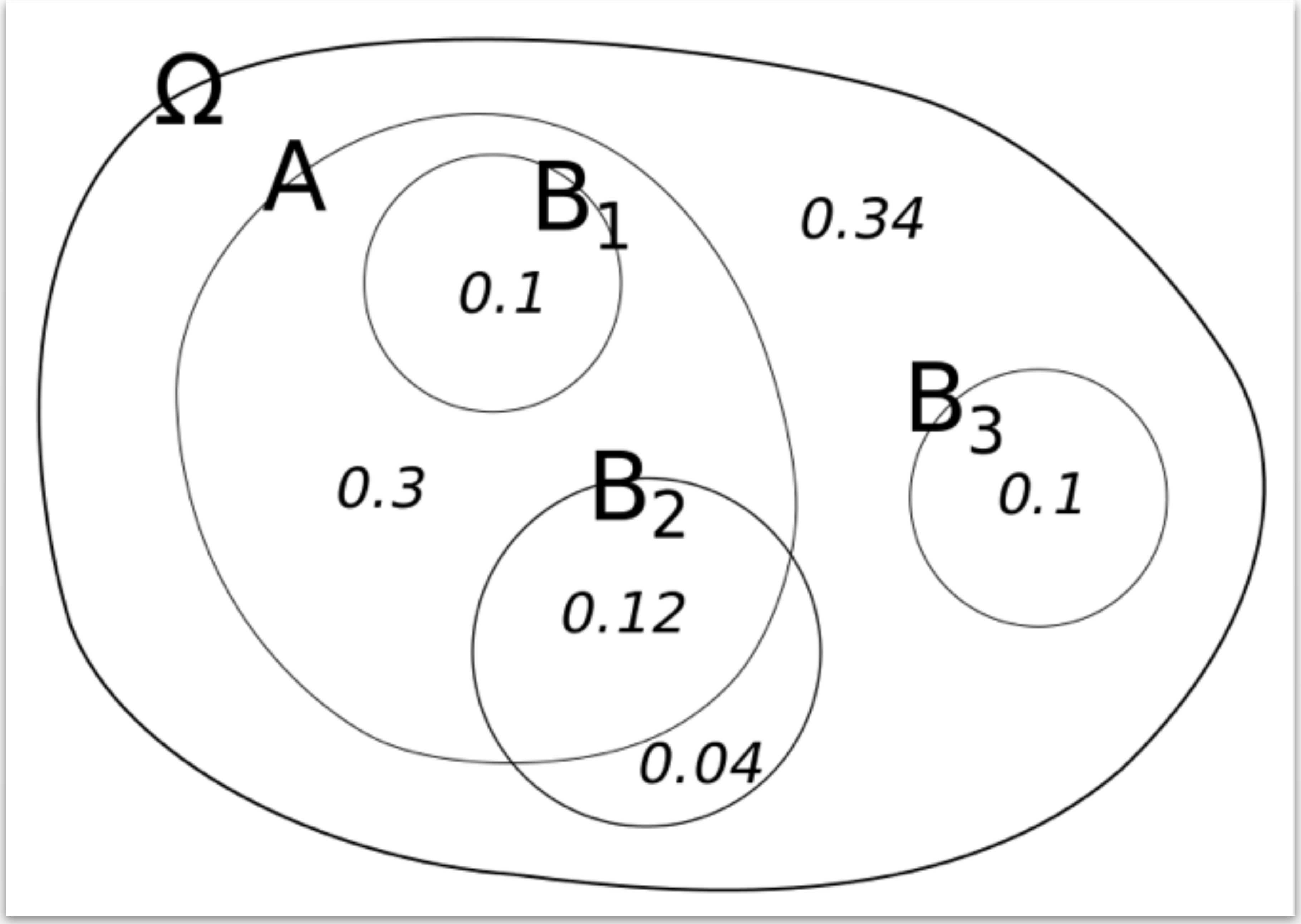


Conditional Probability



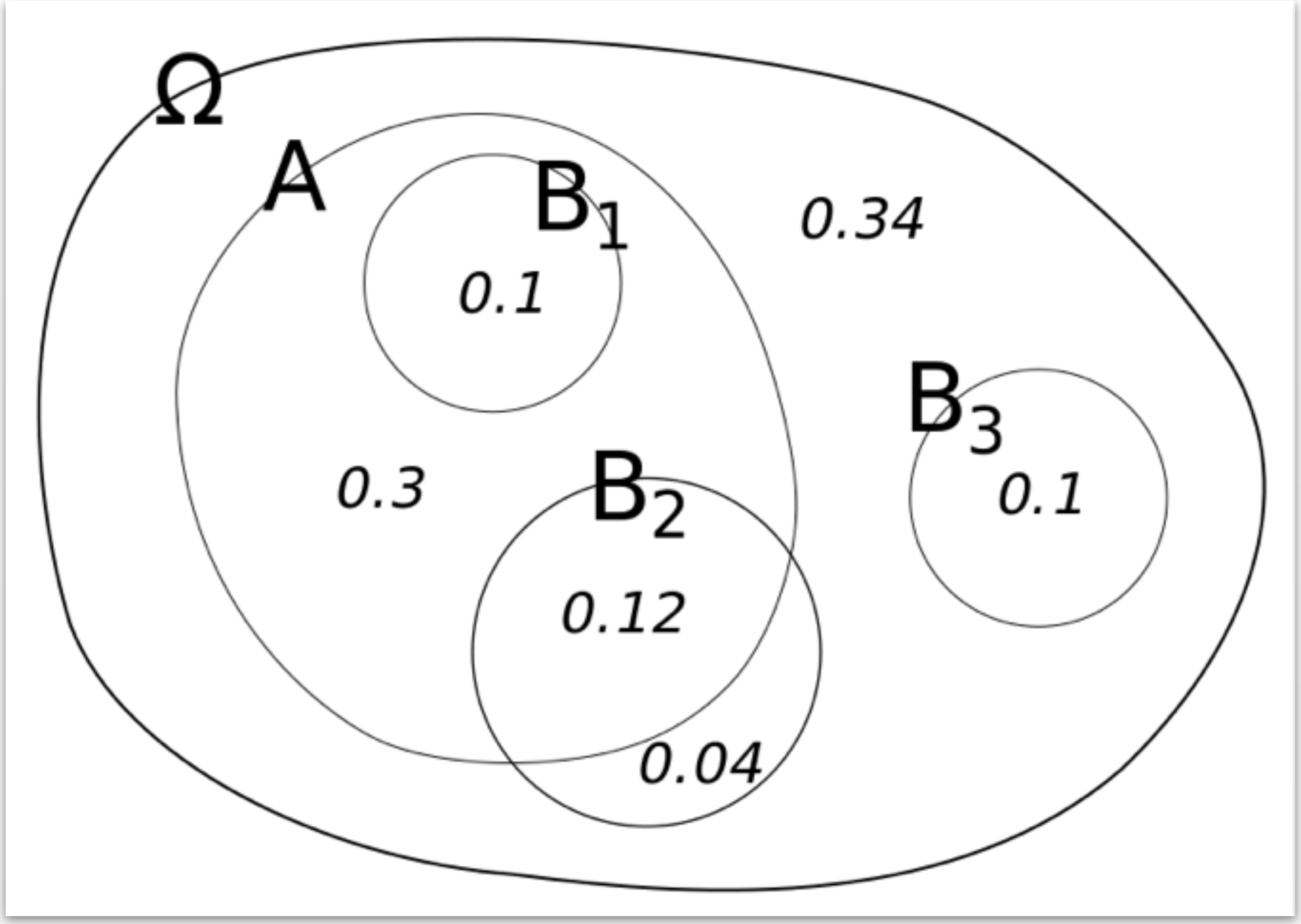
What is the probability P(B3)?

Conditional Probability



What is the probability $P(B_1 \mid B_3)$*?*

Conditional Probability



What is the probability $P(B_2 \mid A)$?

Examples: Conditional Probability

1. A math teacher gave her class two tests.
    - *25% of the class passed both tests*
    - *42% of the class passed the first test.*

What percent of those who passed the first test also passed the second test?

2. Suppose that for houses in New England
    - *84% of the houses have a garage*
    - *65% of the houses have a garage and a back yard.*

What is the probability that a house has a backyard given that it has a garage?

Random Variable

- A random variable X, is a function X: $\Omega \rightarrow R$

Rolling a die:

- X = number on the die
- $p(X = i) = 1/6 \qquad i = 1,2,\ldots,6$

Rolling two dice at the same time:

- X = sum of the two numbers
- $p(X = 2) = 1 / 36$

Probability Mass Function

- For a discrete random variable X, a PMF is a function p: $R \rightarrow R$ such that

  $$p(x) = P(X = x)$$

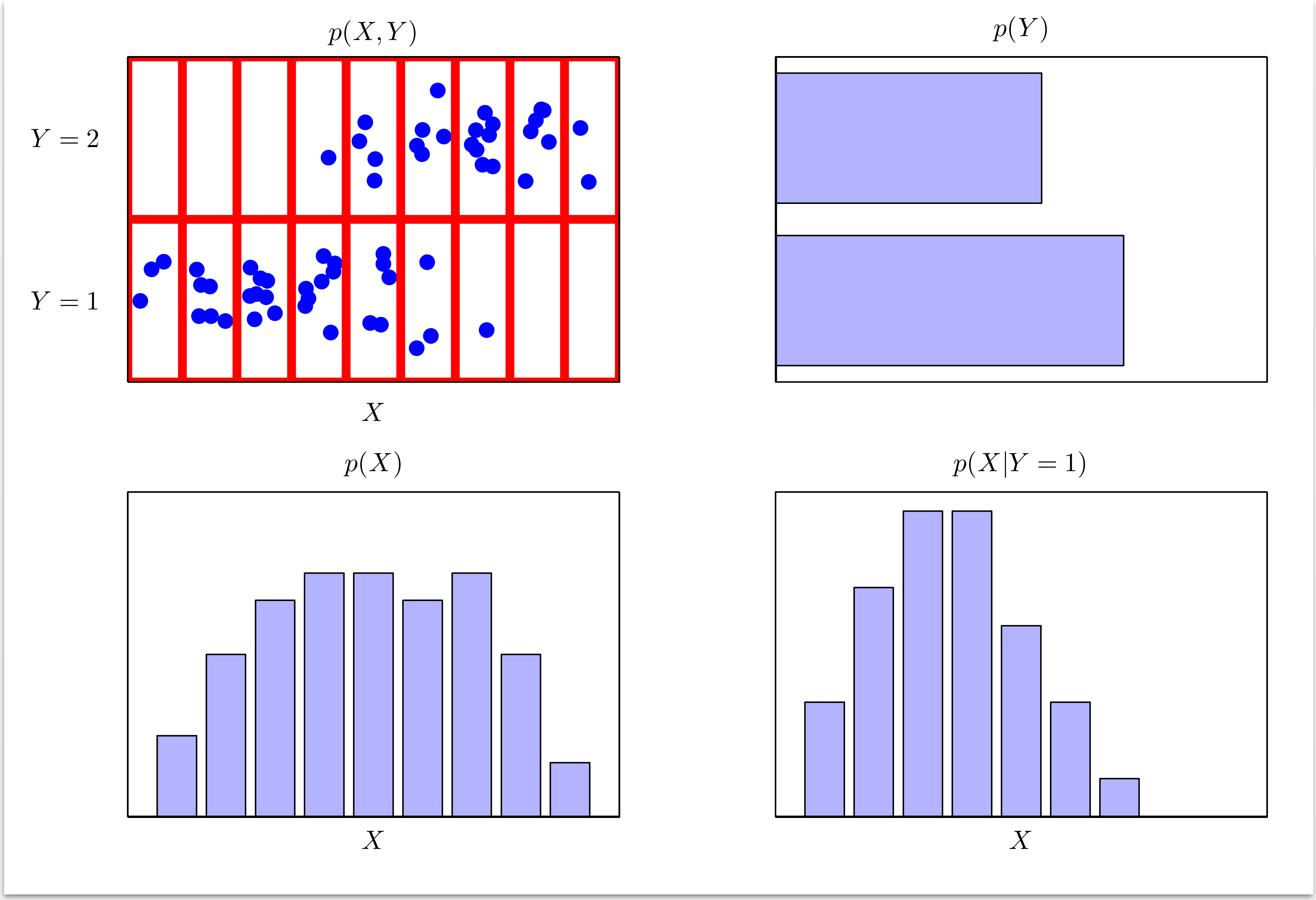
Rolling a die:

- X = number on the die
- $p(X = i) = 1/6 \qquad i = 1,2,\ldots,6$

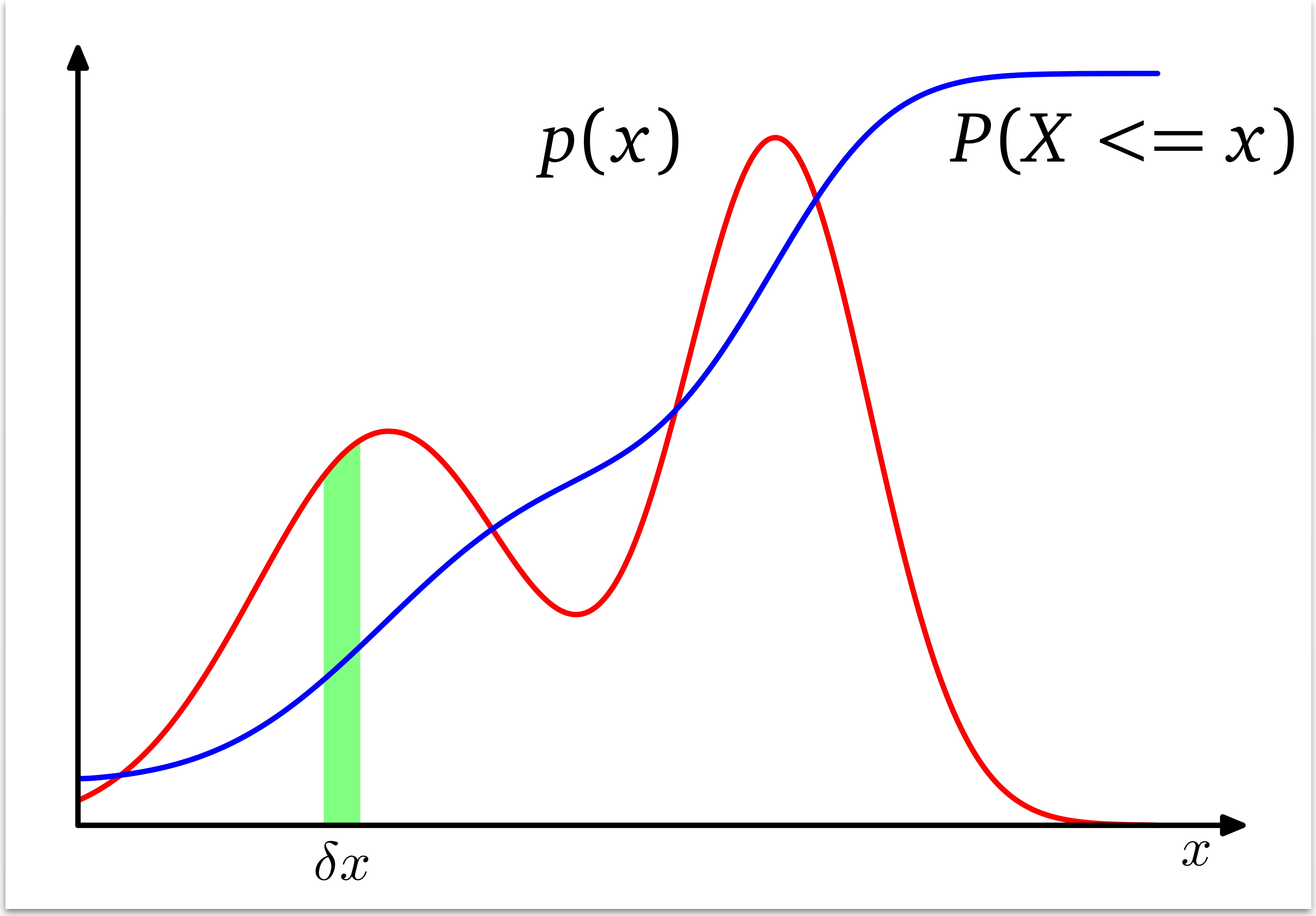
Rolling two dice at the same time:

- X = sum of the two numbers
- $p(X = 2) = 1 / 36$

Continuous Random Variables



Probability Density Functions



$$p(x) = \lim_{\delta x \to 0} \frac{P(X <= x + \delta x) - P(X <= x)}{\delta x}$$

Expected Values

Statistics

$$\mathbb{E}[X] = \sum_x p(x)\, x$$

$$\mathbb{E}[X] = \int dx\, p(x)\, x$$

Machine Learning

$$\mathbb{E}_{p(x\,|\,y)}[f(x)] = \sum_x p(x\,|\,y)\, f(x)$$

$$\mathbb{E}_{p(x\,|\,y)}[f(x)] = \int dx\, p(x\,|\,y)\, f(x)$$

Expected Values

Statistics

$$\mathbb{E}[X] = \sum_x p(x)\, x$$

$$\mathbb{E}[X] = \int dx\, p(x)\, x$$

Machine Learning

$$\mathbb{E}_x[f(x)\,|y] = \sum_x p(x\,|\,y)\, f(x)$$

$$\mathbb{E}_x[f(x)\,|\,y] = \int dx\, p(x\,|\,y)\, f(x)$$

Expected Values

Mean

$$\bar{X} = \mathbb{E}[X]$$

Variance

$$\mathrm{Var}[X] = \mathbb{E}[(X - \bar{X})^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Covariance

$$\mathbf{\Sigma}_{i,j} = \mathrm{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \bar{X}_i)(X_j - \bar{X}_j)]$$

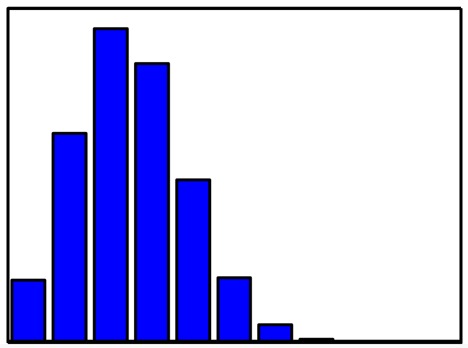
Conjugate Distributions

Bernoulli

$$\begin{aligned}
\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\
\mathbb{E}[x] &= \mu \\
\mathrm{var}[x] &= \mu(1-\mu) \\
\mathrm{mode}[x] &= \begin{cases} 1 & \text{if } \mu \geqslant 0.5, \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

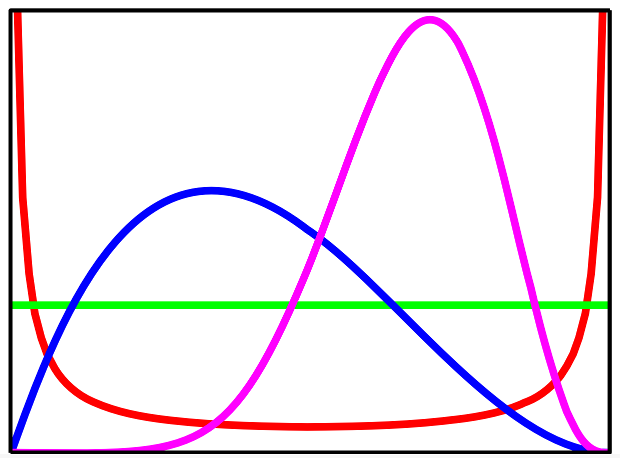
$$x \in \{0,1\} \qquad \mu \in [0,1]$$

Binomial



$$
\begin{aligned}
\mathrm{Bin}(m|N,\mu) &= \binom{N}{m}\mu^m(1-\mu)^{N-m} \\
\mathbb{E}[m] &= N\mu \\
\mathrm{var}[m] &= N\mu(1-\mu) \\
\mathrm{mode}[m] &= \lfloor (N+1)\mu \rfloor
\end{aligned}
$$

Beta



$$
\begin{aligned}
\mathrm{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} \\
\mathrm{mode}[\mu] &= \frac{a-1}{a+b-2}.
\end{aligned}
$$

Conjugacy

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^{m}(1-\mu)^{N-m}$$

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

$$\begin{aligned} p(\mu\,|\,m) &= \frac{p(m,\mu)}{p(m)} \\ &\propto \mathrm{Bin}(m\,|\,N,\mu)\mathrm{Beta}(\mu\,|\,a,b) \\ &\propto \mu^{m+(a-1)}(1-\mu)^{(N-m)+(b-1)} \end{aligned}$$

Conjugacy

$$\mathrm{Bin}(m|N,\mu) \quad = \quad \binom{N}{m}\mu^{m}(1-\mu)^{N-m}$$

$$\mathrm{Beta}(\mu|a,b) \quad = \quad \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

$$p(\mu\,|\,m) = \frac{p(m,\mu)}{p(m)}$$

$$\propto \mathrm{Bin}(m\,|\,N,\mu)\mathrm{Beta}(\mu\,|\,a,b)$$

$$\propto \mu^{m+(a-1)}(1-\mu)^{(N-m)+(b-1)}$$

$$\boxed{p(\mu\,|\,m) = \mathrm{Beta}(a+m, b+(N-m))}$$

Conjugacy

$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \underbrace{p(y \mid x)}_{\text{Likelihood}}\underbrace{p(x)}_{\text{Prior}}/p(y)$$

Example: Biased Coin

y	Observed data (flip outcomes)
x	Unknown variable (coin bias)

Conjugacy

$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \underbrace{p(y \mid x)}_{\text{Likelihood}}\underbrace{p(x)}_{\text{Prior}}/p(y)$$

Example: Biased Coin

$p(y \mid x)$	Likelihood of outcome given bias
$p(x)$	Prior belief about bias
$p(x \mid y)$	Posterior belief after trials

Conjugacy

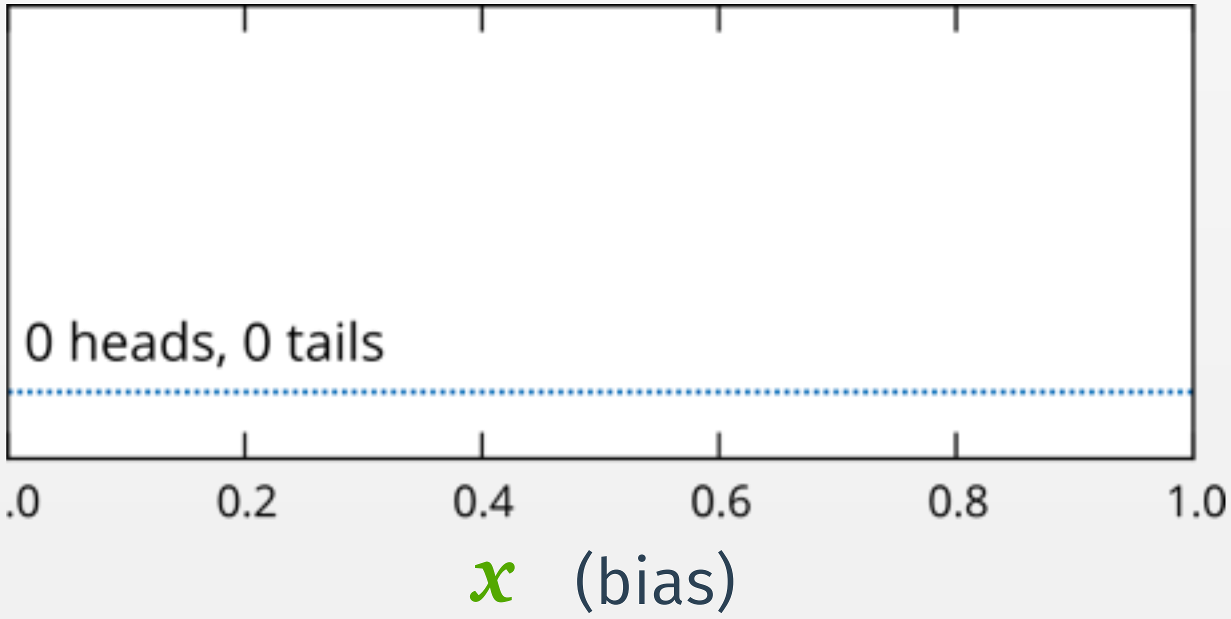
$$p(x \mid y) = p(y \mid x)p(x)/p(y)$$

Posterior

Likelihood

Prior

$$p(x) = \mathrm{Beta}(x; 0, 0)$$

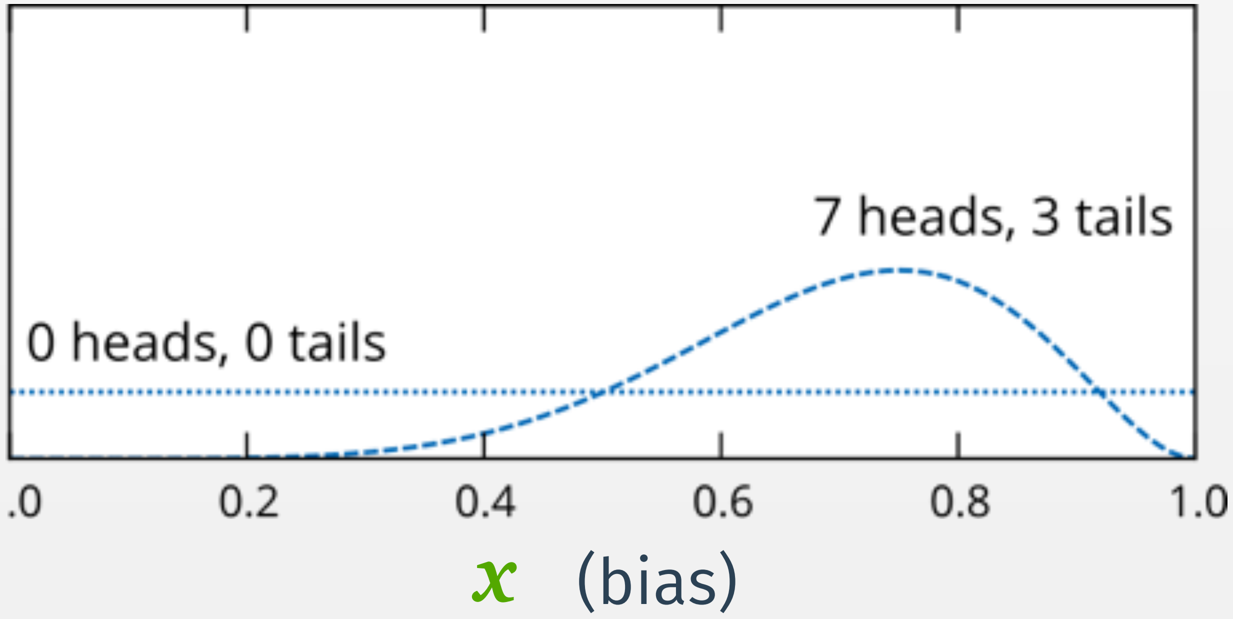


Conjugacy

$$p(x \mid y) = p(y \mid x)p(x)/p(y)$$

Posterior Likelihood Prior

$$p(x \mid y) = \mathrm{Beta}(x; 7, 3)$$



Conjugacy

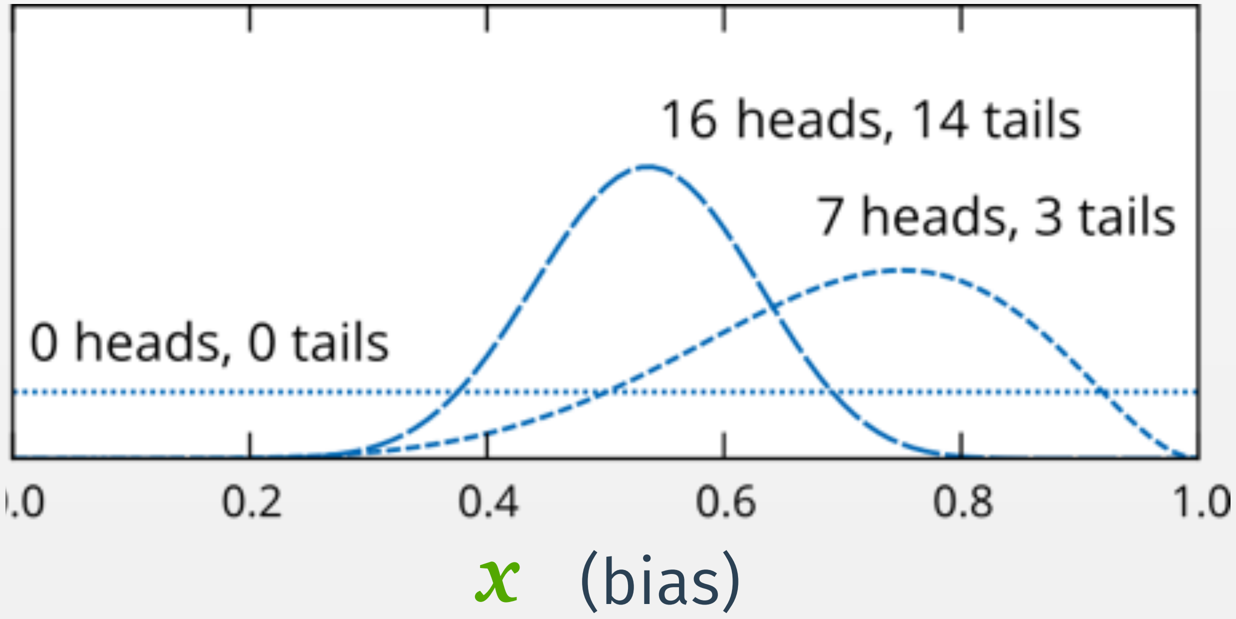
$$p(x \mid y) = p(y \mid x) p(x) / p(y)$$

Posterior

Likelihood

Prior

$$p(x \mid y) = \mathrm{Beta}(x; 16, 4)$$



Conjugacy

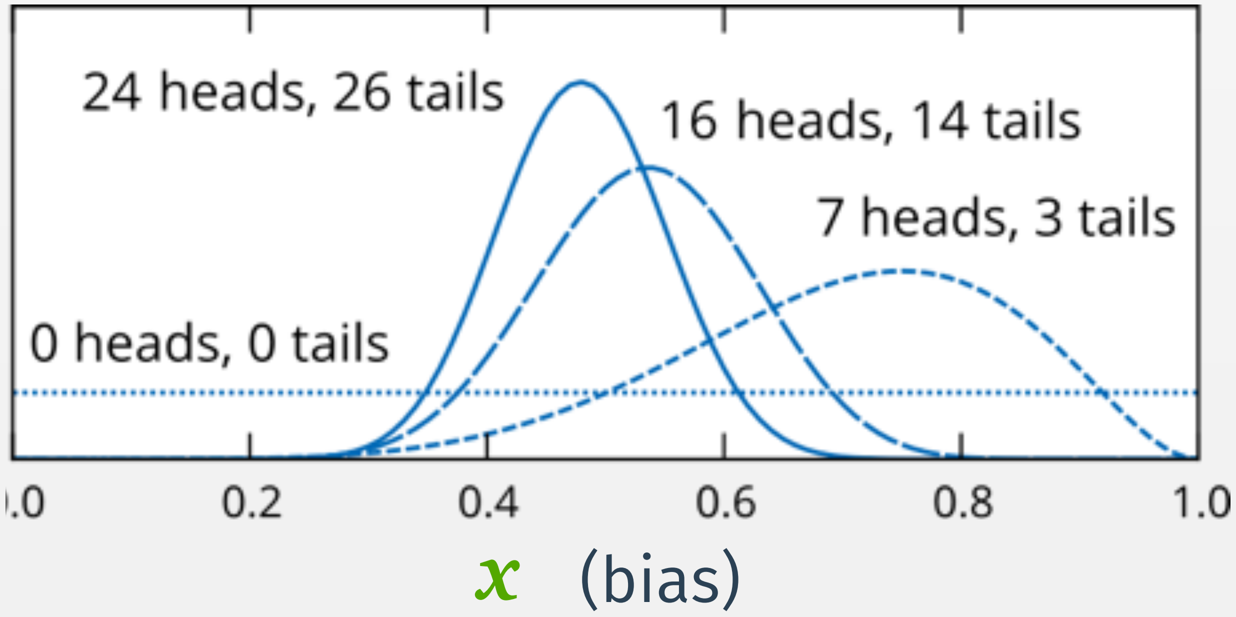
$$p(x \mid y) = p(y \mid x)p(x)/p(y)$$

Posterior

Likelihood

Prior

$$p(x \mid y) = \mathrm{Beta}(x; 24, 26)$$



Discrete (Multinomial)

$$
\begin{aligned}
p(\mathbf{x}) &= \prod_{k=1}^{K} \mu_k^{x_k} \\
\mathbb{E}[x_k] &= \mu_k \\
\mathrm{var}[x_k] &= \mu_k(1-\mu_k) \\
\mathrm{cov}[x_j x_k] &= I_{jk}\mu_k
\end{aligned}
$$

Discrete (Multinomial)

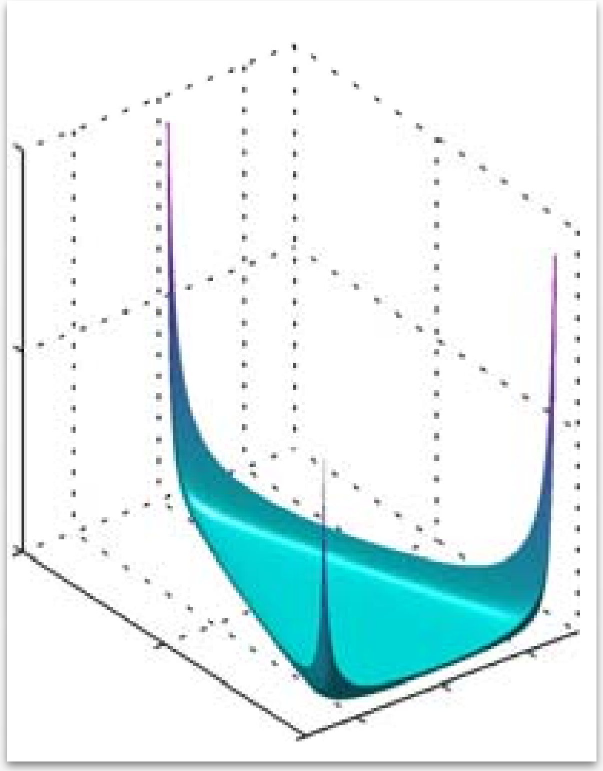
$$\begin{aligned} p(\mathbf{x}) &= \prod_{k=1}^{K} \mu_k^{x_k} \\ \mathbb{E}[x_k] &= \mu_k \\ \mathrm{var}[x_k] &= \mu_k(1-\mu_k) \\ \mathrm{cov}[x_j x_k] &= I_{jk}\mu_k \end{aligned}$$

Dirichlet

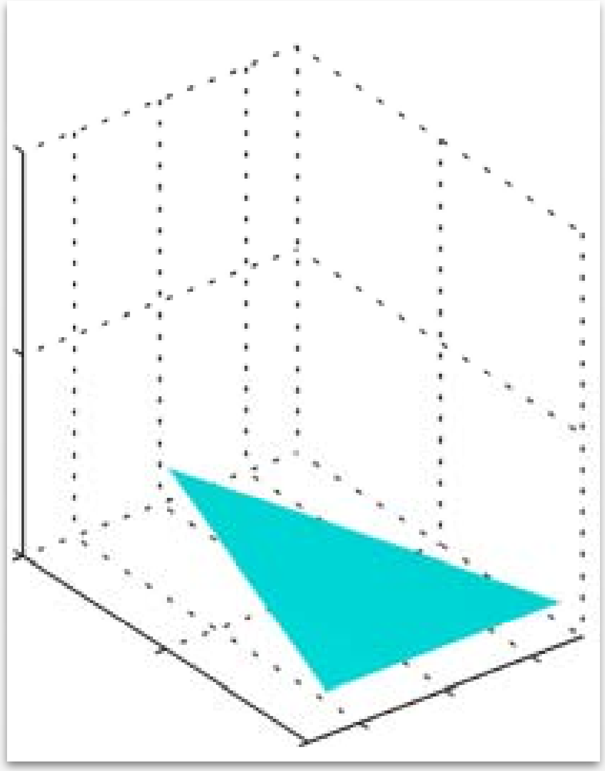
$$\begin{aligned}
\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) &= C(\boldsymbol{\alpha}) \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \\
\mathbb{E}[\mu_k] &= \frac{\alpha_k}{\widehat{\alpha}} \\
\mathrm{var}[\mu_k] &= \frac{\alpha_k(\widehat{\alpha} - \alpha_k)}{\widehat{\alpha}^2(\widehat{\alpha} + 1)} \\
\mathrm{cov}[\mu_j \mu_k] &= -\frac{\alpha_j \alpha_k}{\widehat{\alpha}^2(\widehat{\alpha} + 1)} \\
\mathrm{mode}[\mu_k] &= \frac{\alpha_k - 1}{\widehat{\alpha} - K}
\end{aligned}$$

Dirichlet

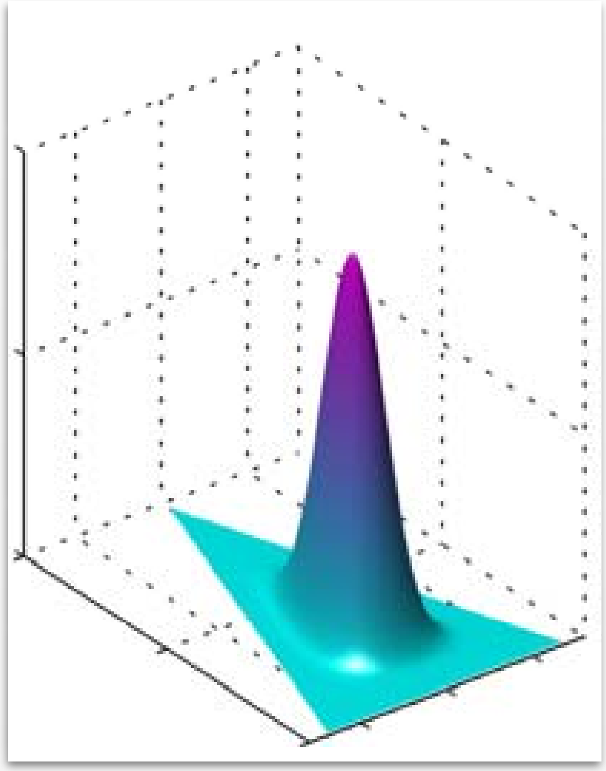
$\alpha = (0.1, 0.1, 0.1)$



$\alpha = (1, 1, 1)$



$\alpha = (10, 10, 10)$



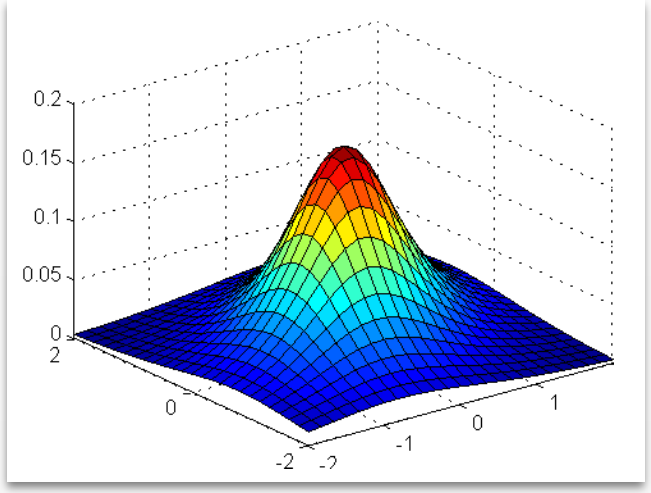
$$p(\boldsymbol{\mu}) = \mathrm{Dir}(\boldsymbol{\mu}; \boldsymbol{\alpha})$$

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \mathrm{Mult}(\boldsymbol{x}; \boldsymbol{\mu})$$

$$p(\boldsymbol{\mu}|\boldsymbol{x}) = \mathrm{Dir}(\boldsymbol{x}; \boldsymbol{\alpha} + \boldsymbol{x})$$

Multivariate Normal

$$
\begin{aligned}
\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\
\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu} \\
\mathrm{cov}[\mathbf{x}] &= \boldsymbol{\Sigma} \\
\mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu}
\end{aligned}
$$



$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}) \\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b},\mathbf{L}^{-1})
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\},\boldsymbol{\Sigma})
\end{aligned}
$$

Bayesian Linear Regression

Prior and Likelihood

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

$$p(\boldsymbol{y} \mid \boldsymbol{w}, \alpha, \beta) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1}\boldsymbol{I})$$

Posterior

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \alpha, \beta) \propto p(\boldsymbol{y} \mid \boldsymbol{w}, \alpha, \beta) p(\boldsymbol{w} \mid \alpha)$$

Maximum A Posteriori (MAP) gives Ridge Regression

$$\operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w} \mid \boldsymbol{y}, \alpha, \beta) = \frac{\beta}{2} \sum_{n=1}^{N} (\boldsymbol{w}^\top \boldsymbol{x}_n - \boldsymbol{y}_n)^2 + \frac{\alpha}{2} \boldsymbol{w}^\top \boldsymbol{w}$$