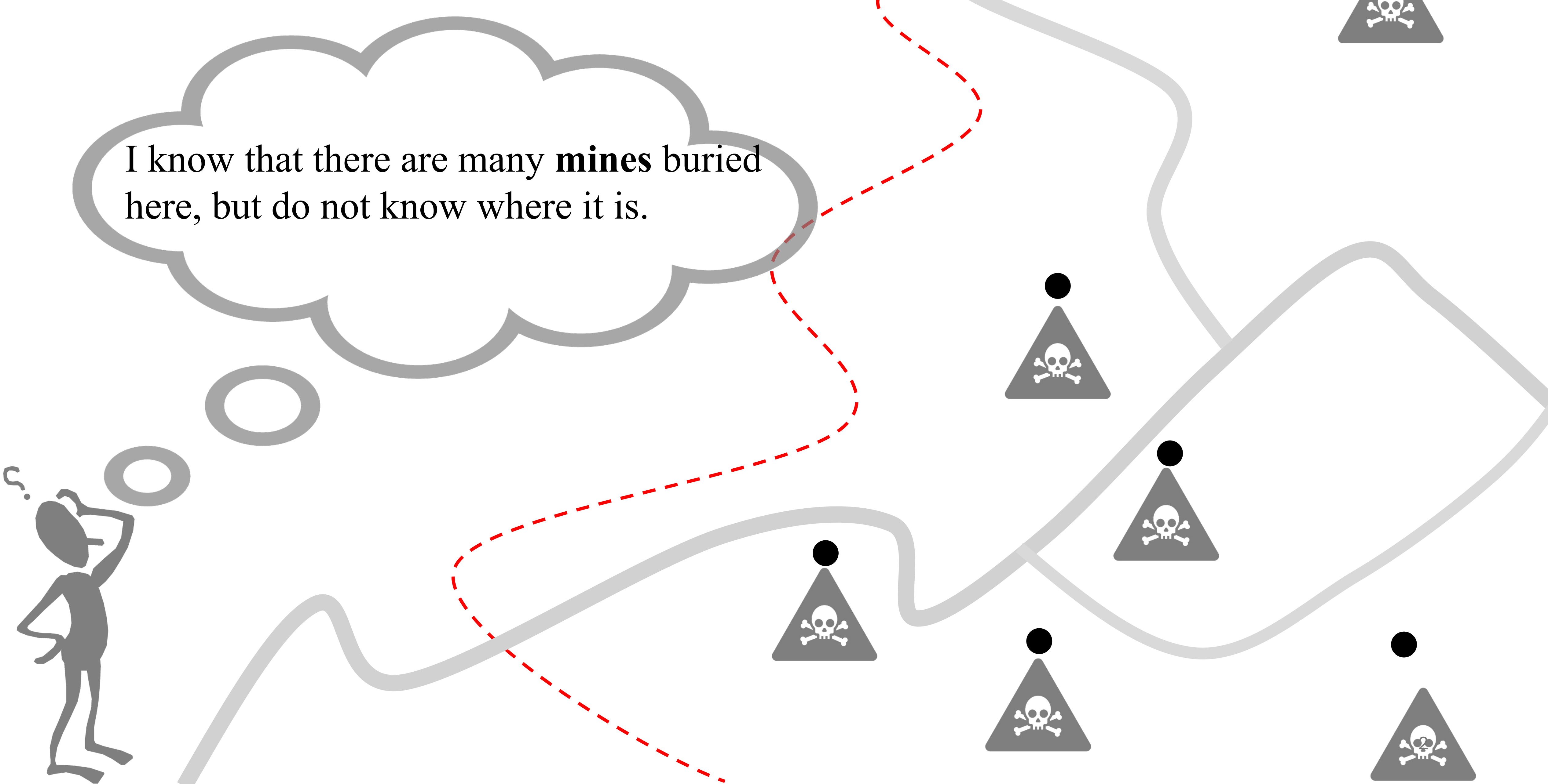
CS 461: Machine Learning Principles

Class 2: Sept. 9

Probability 101 and Machine Learning

Instructor: Diana Kim

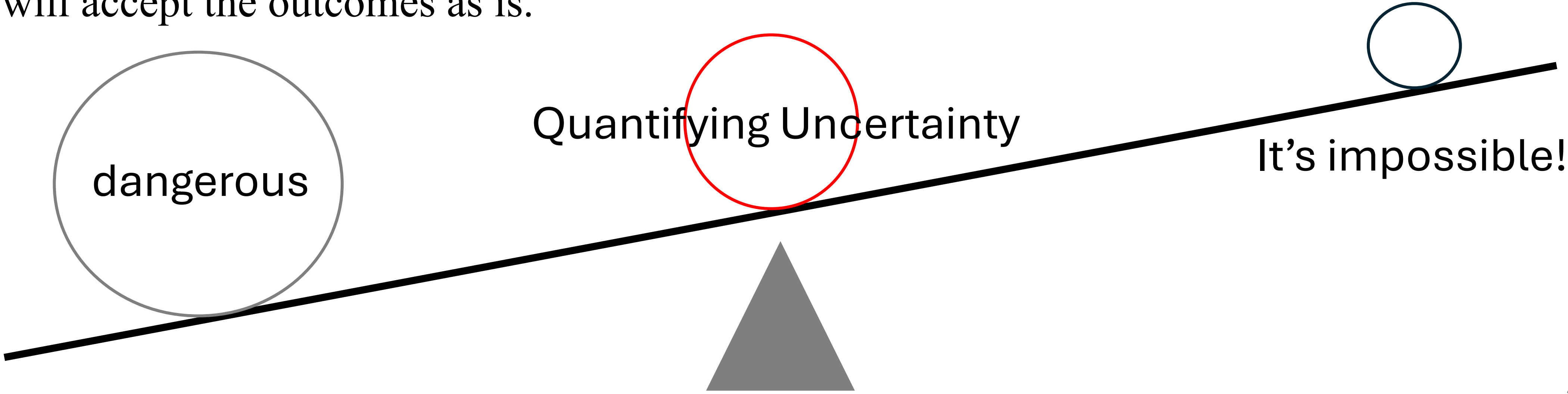
We live in an uncertain world.



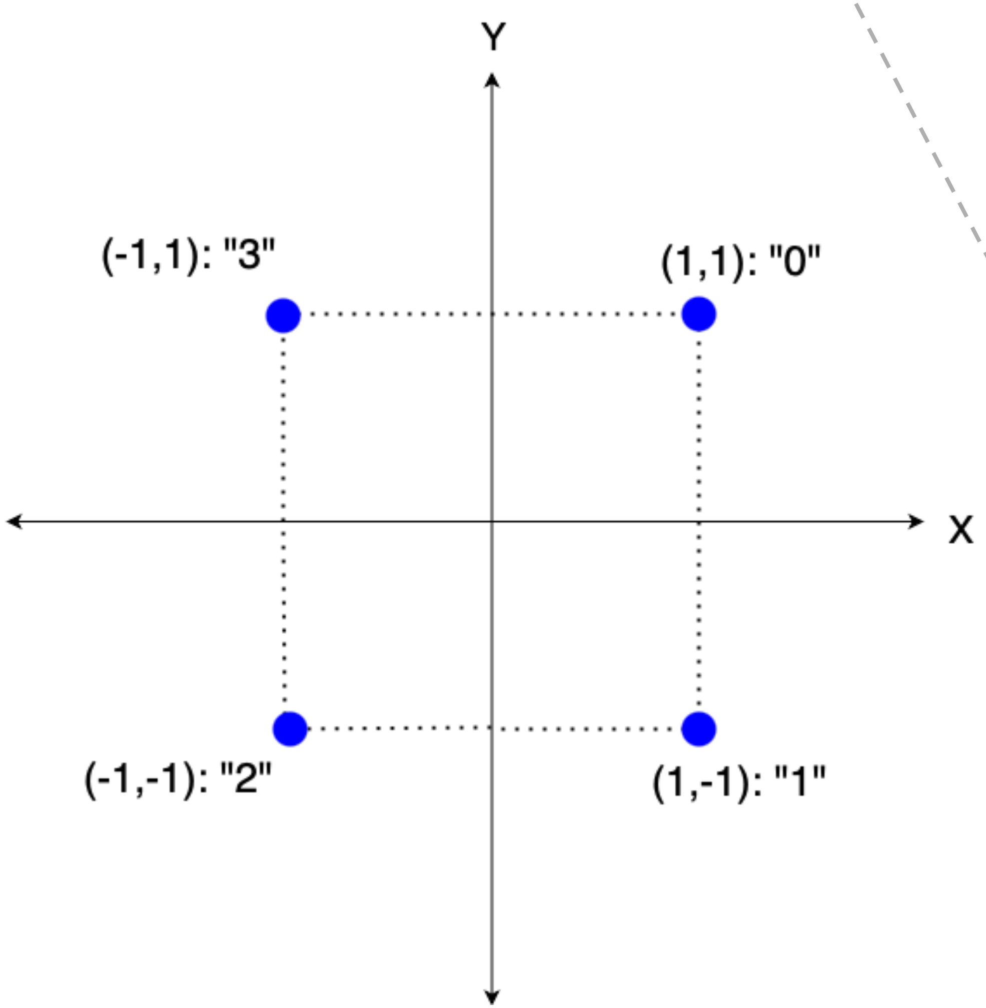
Two Extreme Strategies to Uncertainty

Omni-Knowledge World:
I know everything!
so can predict what will be happened.

Random World:
"Life is Random; no strategy.
I will accept the outcomes as is."



In the real world we have limited knowledge.



SENDER: Two Channels-X and Y

RECEIVER: One Channel-X

+ as miss to detect channel Y
+ as we don’t know the channel Y exist
+ we don’t know the map

- If we receive (1) “0” “1”
- If we receive (-1) “2” “3”

We have limited knowledge.
Our world is uncertain.
We need to measure uncertainty to make a rational/safe decision.

Probability Theory is to provide mathematical machinery to measure uncertainty associated events.

Probability in ML?

Probability in ML?

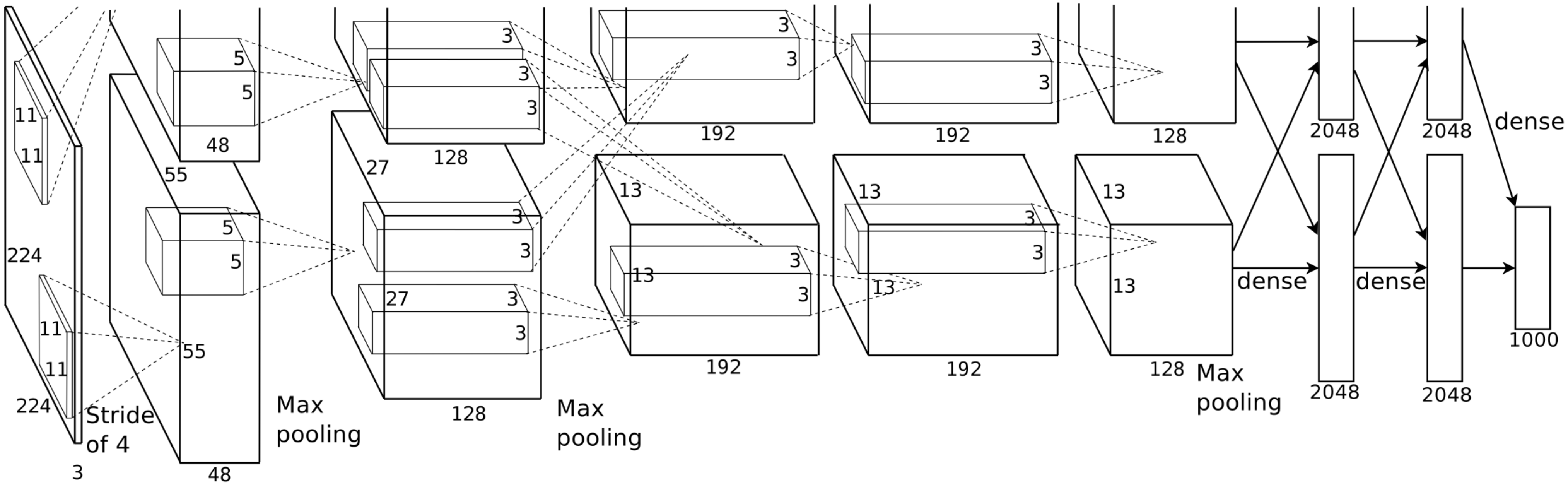
1. Probabilistic Modeling (target to learn):

- Discriminative modeling: $P(y|x)$ ouput layer: sigmoid /softmax
- Generative modeling: $P(x, y)$ encodes a joint density (implicit/ explicit)

Probability in ML?

1. Probabilistic Modeling: (target to learn a joint/ conditional density)

- Discriminative modeling: $P(y|x)$ softmax/ logistic



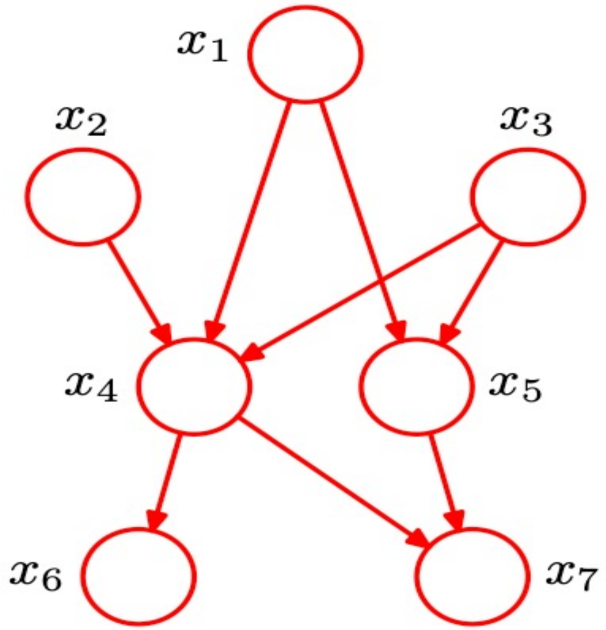
AlexNet (Deep-CNN) to learn $P(y = object\ class\ | \ x)$

From the paper “ImageNet Classification with Deep Convolutional Neural Networks” by Alex Krizhevsky et al.

Probability in ML?

1. Probabilistic Modeling: (target to learn joint/ conditional density)

- Generative Modeling: $P(x, y), P(x_1, x_2, x_3 \ldots, x_n)$



Bayes-net

from Bishop Figure 8.2

Bayesian net describing the joint distribution:

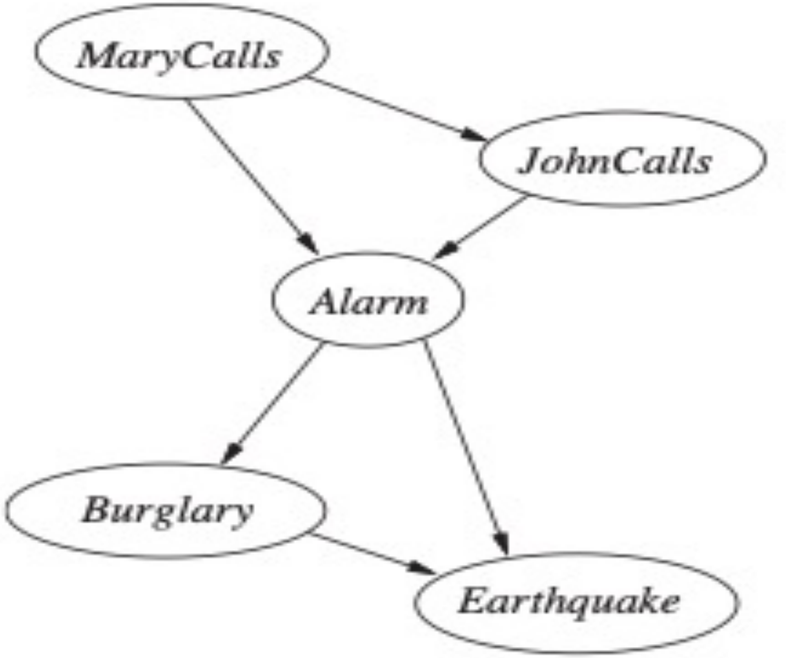
$$p(x_1, x_2, \ldots, x_7) =$$

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5).$$

Probability in ML?

1. Probabilistic Modeling: (target to learn joint/ conditional density)

- Generative Modeling: $P(x, y), P(x_1, x_2, x_3 \ldots, x_n)$



From Figure 14.3 "Artificial Intelligence, A modern Approach"

Probability in ML?

1. Probabilistic Modeling: (target to learn joint/ conditional density)

- Generative Modeling: $P(x, y), P(x_1, x_2, x_3 \ldots, x_n)$

from the paper "Generative Adversarial Nets" by Ian J. Goodfellow et al.

Generative Adversarial Nets (GAN) learns a joint density to generate new images not in the training set. The figure shows the sample from the model. Rightmost column shows the nearest training example of the neighboring sample.

Probability in ML?

1. Probabilistic Modeling: (target to learn joint/ conditional density)
2. Error modeling: even for non-probabilistic/deterministic modeling, we need to consider the random error possible in it and proceed ML or MAP estimation to $f_w(x)$ given data for x and y

$$y = f_w(x) + \varepsilon$$

errors from :
+ imperfect feature / data set
+ imperfect hypothesis space
+ measurement error
+ outlier/ (intrinsic error)
+ small number of data points

Probability in ML?

1. Probabilistic Modeling: (target to learn joint/ conditional density)
2. Error modeling: even for non-probabilistic/deterministic modeling, we need to consider the random error possible in it and proceed ML or MAP estimation for $f_w(x)$ given data for x and y

$$y = f_w(x) + \varepsilon$$

Learning $f(x)$ becomes an estimation problem
given observations for x and y and ε following a certain density.

ex) $\varepsilon \sim N\,(0, \sigma^2)$ then $y \sim$

Probability in ML?

1. Probabilistic Modeling: (target to learn joint/ conditional density)
2. Error modeling: even for non-probabilistic/deterministic modeling, we need to consider the random error possible in it and proceed ML or MAP estimation for $f_w(x)$ given data for x and y

$$y = f_w(x) + \varepsilon$$

According the density we assumed for ε,
learning $f(x)$ becomes an estimation problem given observations for x and y.
ex) $\varepsilon \sim N(0, \sigma^2)$ then $y \sim$

- $w = argmax\, P(y|w, x)$: Maximum Lliklihood Estimation (MLE) w as a fixed point
- $w = argmax\, p(w|y, x) = \dfrac{p(y|w, x)p(w)}{p(y|x)}$: Maximum A posteriori Estimation (MAP)

P(w) encodes prior knowledge/ expert knowledge

Probability in ML?

3. Making an optimal decision/ choice under an uncertain situation

Ex) Bayesian Binary Hypothesis Testing

- Hypothesis
$$\begin{cases} H_0 : Y \sim f(y|\, H_0) \\ H_1 : Y \sim f(y|\, H_1) \end{cases}$$

- Decision region
$$\begin{cases} y_0 = \{y|\delta\,(y) = 0\} \\ y_1 = \{y|\delta\,(y) = 1\} \end{cases}$$

- Average Risk

$$\begin{aligned} E[R] = \pi_0 \cdot E[R|H_0] + \pi_1 E[R|H_1] &= \pi_1 \cdot \int_{y_0} f(y|H_1)dy + \pi_0 \cdot \int_{y_1} f(y|H_0)dy \\ &= \pi_1 \cdot \int_{y_0} f(y|H_1)dy + \pi_0 \cdot (1 - \int_{y_0} f(y|H_0)dy) \\ &= \pi_0 + \int_{y_0} \pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0)dy \end{aligned}$$

Q: How should we set the decision rule for y_0?

we need to set decision region for y-0 getting inside of the integral negative

Probability in ML?

- Average Risk $E[R] = \pi_0 \cdot E[R|H_0] + \pi_1 E[R|H_1] = \pi_1 \cdot \int_{y_0} f(y|H_1)dy + \pi_0 \cdot \int_{y_1} f(y|H_0)dy$

$$= \pi_1 \cdot \int_{y_0} f(y|H_1)dy + \pi_0 \cdot (1 - \int_{y_0} f(y|H_0)dy)$$

$$= \pi_0 + \int_{y_0} \pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0)dy$$

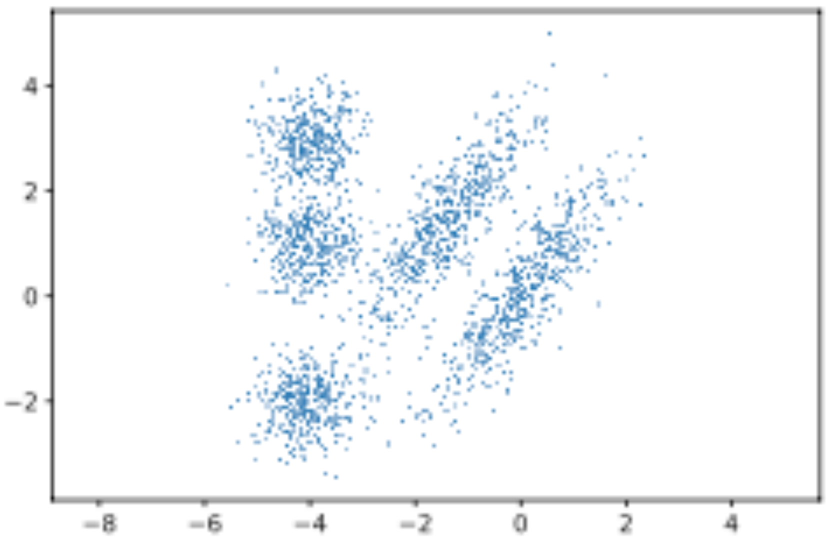
Q: How should we set the decision rule for y_0?

for y, if $\pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0) < 0$ then detect as H_0
else if $\pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0) \geq 0$ then detect as H_1

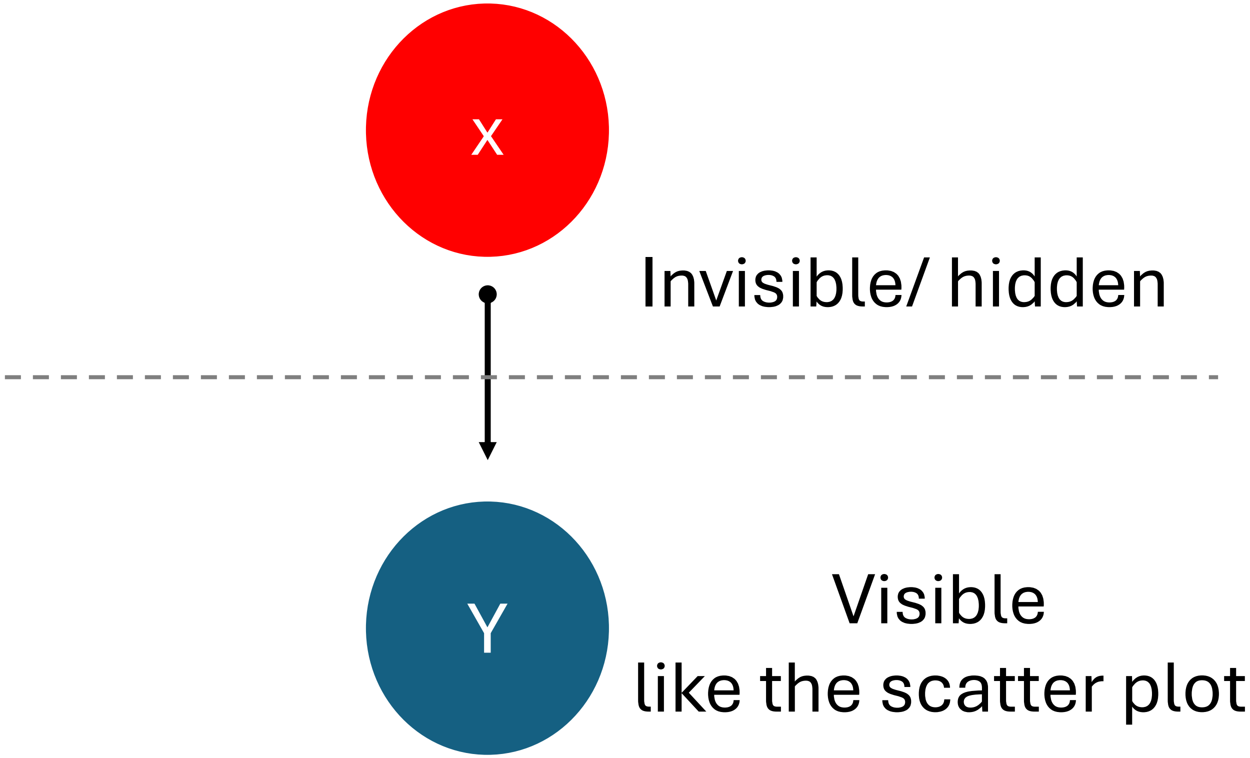
$$\frac{p(y|H_1)}{P(y|H_0)} \underset{H0}{\overset{H1}{\gtreqless}} \frac{\pi_0}{\pi_1}$$

Probability in ML?

4. Representation of hidden information
   Ex) Mixture Gaussian Density



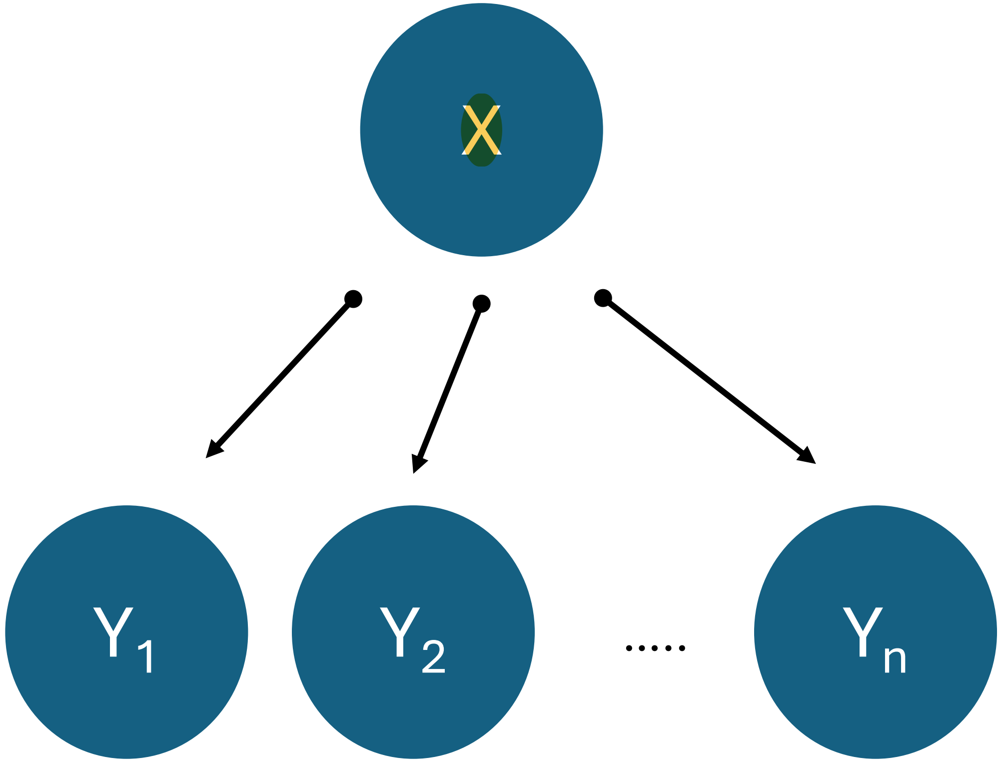
from Figure 3.11 Murphy, An Introduction



Q: Can you see a hidden variable in the 2-D scatter plot?
How would you encode the information?

Q: Naïve Bayes Classification?

Naive Bayes classifier
is a generative modeling because it need to learn joint density P(x,y1,y2,…Yn)



Probability 101

Probability Space $[\boldsymbol{\Omega}, \mathbf{2}^{|\boldsymbol{\Omega}|}, \mathrm{P}]$

[1] Experiment: any process of obtaining or generating an observation

Ex] Inspection of an instance item is defective or non-defective

[2] **Sample Space ($\boldsymbol{\Omega}$): a set** of all possible outcomes

Ex] $\boldsymbol{\Omega} = \{\text{non-defective, defective}\}$

[3] **Events Set: ($A \subset \boldsymbol{\Omega}$ or $A \in \mathbf{2}^{|\boldsymbol{\Omega}|}$): a set of** all possible subsets of Ω

Ex] $2^{|\Omega|} = \{\emptyset, \{\text{non-defective}\}, \{\text{defective}\}, \boldsymbol{\Omega}\}$

[4] **Probability Measure P[E]: a function** P: $\mathbf{2}^{|\boldsymbol{\Omega}|} \rightarrow [0, 1]$

Ex] P[{defective}] = monitor assembly line for a period of time,

compute the relative frequency.

Probability Measure follows the three axioms.

- Non-negativity: $P[A] \geq 0$
- Total Proablity: $P[\Omega] = 1$
- Countable Additive: $A_i \cap A_j = \phi \; if \; i \neq j \implies \mathrm{P}[\bigcup_k A_k] = \sum_k P[A_k]$

The three axioms derives corollaries.

- Non-negativity: $P[A] \geq 0$
- Total Proablity: $P[\Omega] = 1$
- Countable Additive: $A_i \cap A_j = \phi \; if \; i \neq j \implies \mathrm{P}[\cup_k A_k] = \sum_k P[A_k]$

- $P[A^c] = 1 - P[A]$
  by countable additivity and total probability, $P[A^c \cup A] = \mathrm{P}[A] + \mathrm{P}[A^c] = 1$
- $P[\phi] = 1 - P[\Omega] = 0$

Are we ready to define a measure P?

As equally likely outcomes, P[A] becomes counting problem.

$$P[A] = \frac{|A|}{|\Omega|}$$

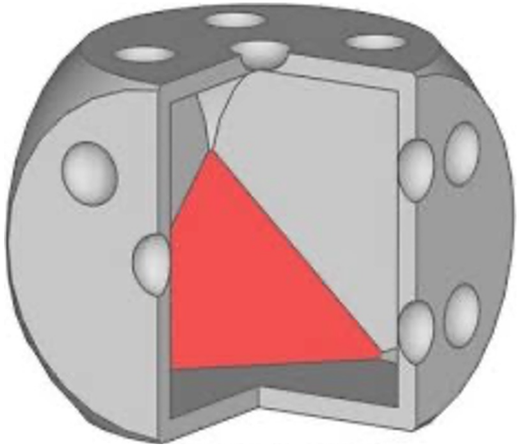
Ex] When tossing a fair coin N times, compute P [k times H]

$$P[\ k \text{ times H}\ ] = \binom{N}{k}\frac{1}{2^N}$$

- $|\Omega|$ = choose H or T N times : 2^N
- $|A|$ = choose k among different N without orders:

2000, Denver Native American One Dollar coin

However,
not always the outcomes are equally and likely.



Corner is loaded with lead!

However, if outcomes are not equally likely?

$$P[A] = \sum_{\omega_k \in A} P[\{\omega_k\}]$$

$$= \sum_{A_k \subset A} P[A_k], A_k \cap A_j = \emptyset \text{ if } k \neq j$$

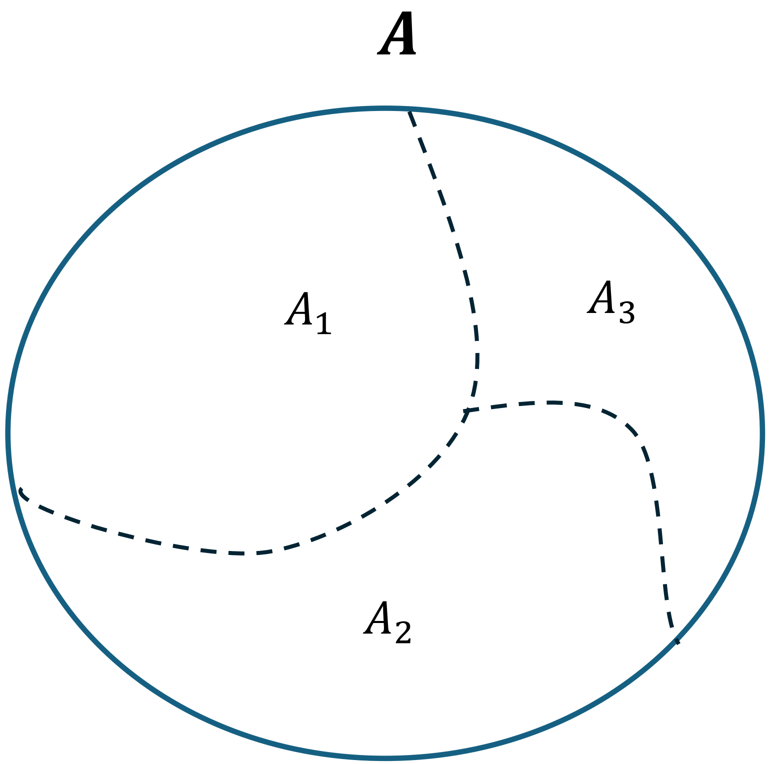


Fig: Event A can be divided into three disjoint sets.

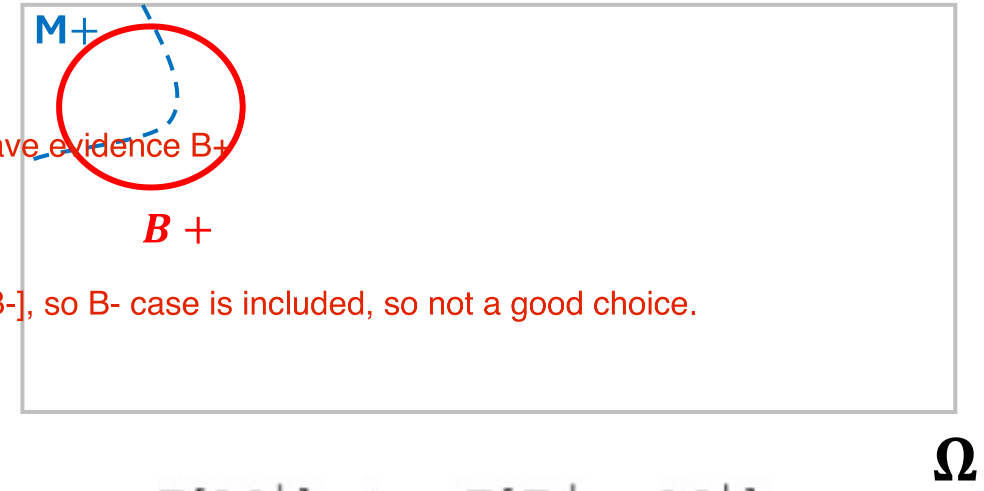
We can divide a complex event into disjoint events, which are tracible,
we can compute its probability in easier way often.

Conditional Probability

Conditional Probability:
[1] Computing Diagnostic Probability (Posterior Prob)

EX] **Breast Cancer** is a deadly disease that claims thousands of lives every year.
If you are a doctor who had a patient having a positive mammogram.
You know that mammogram accuracy is between 90% - 95%.
Which probability would you tell the patient?

- P[B+]=0.008? this may not be a good choice, we have evidence B+
- P[M accurate]=? P[M accurate] = P[M+ B+] + P[M-B-], so B- case is included, so not a good choice.
- P[B+|M+]? = 9%



$$P[M^+] : 1 = P[B^+ \cap M^+] : x$$

$$P[B+|M+] = \frac{P[B^+ \cap M^+]}{P[M^+]}$$

Joint probability and Conditional Probability

$$P[A \cap B] = P[A] \cdot P[B|A] = P[B] \cdot P[A|B]$$

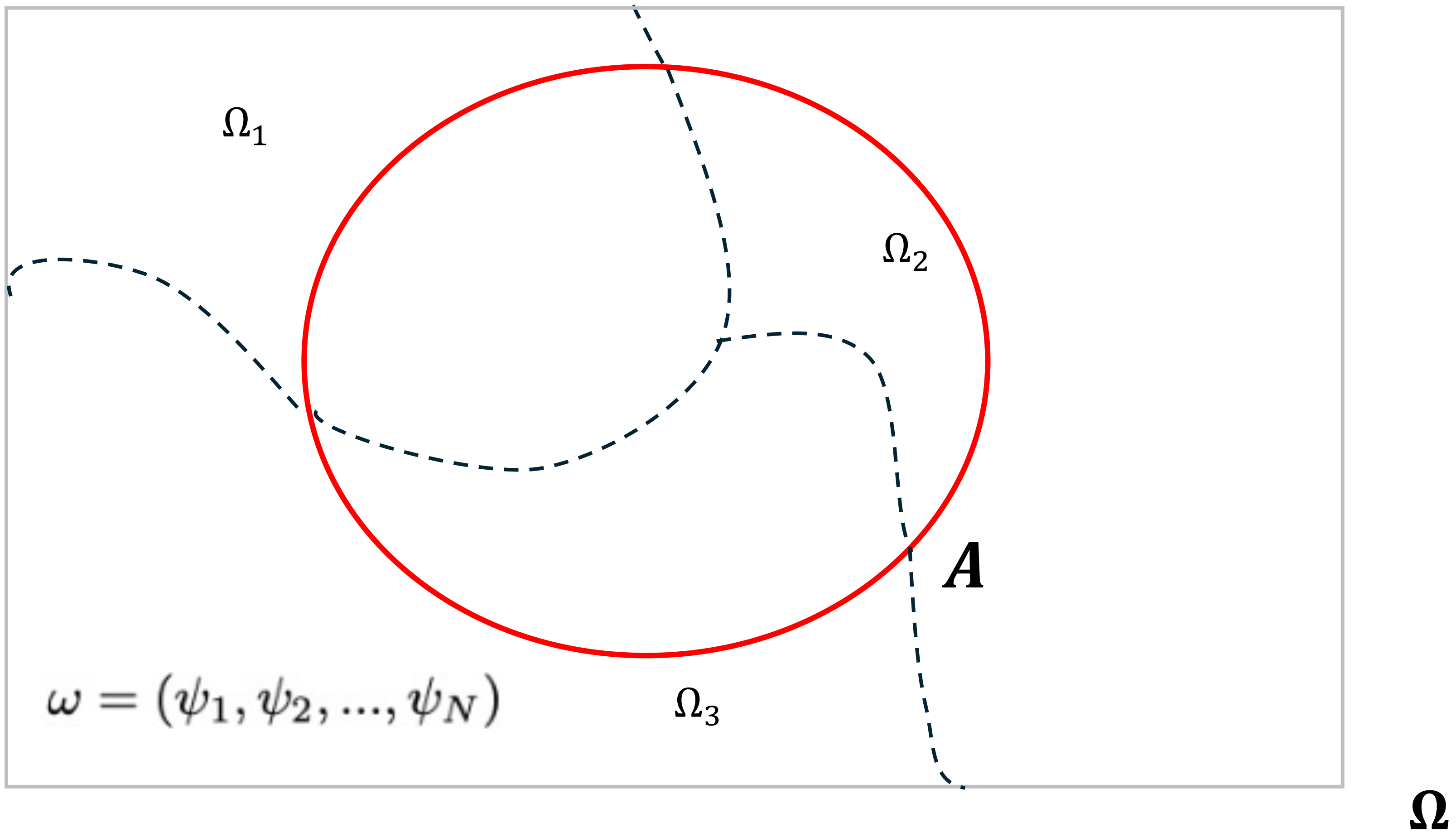
$$P[A \cap B \cap C] = P[A] \cdot P[B|A] \cdot P[C|A \cap B]$$

Independent Events ↔

$$P[A \cap B] = P[A] \cdot P[B]$$

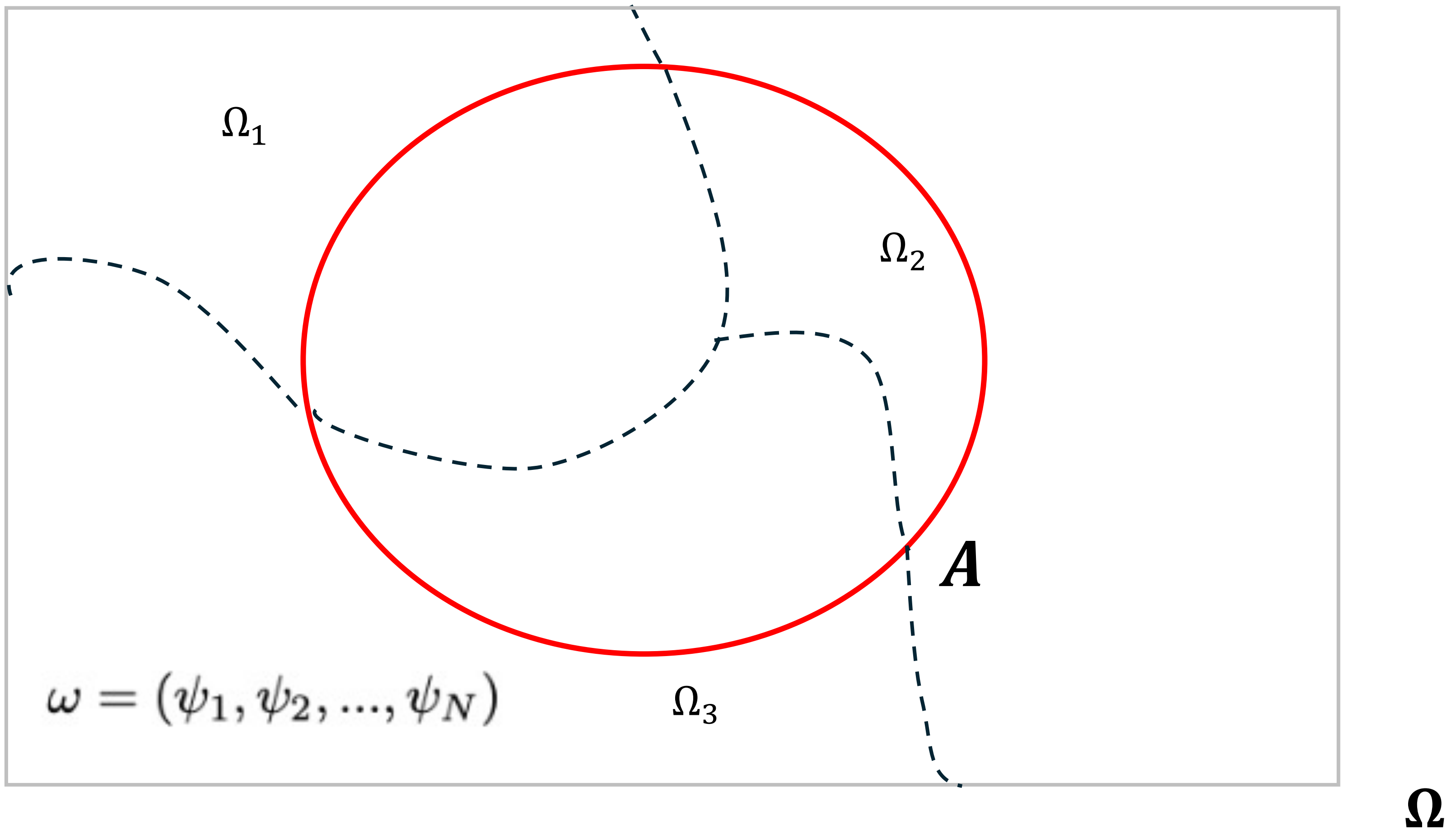
$$P[A \cap B \cap C] = P[A] \cdot P[B] \cdot P[C]$$

Partition the sample space Ω and measure the probability A



$$P[A] = P[A \cap \Omega_1] + P[A \cap \Omega_2] + P[A \cap \Omega_3]$$
$$P[A] = P[A|\Omega_1]P[\Omega_1] + P[A|\Omega_2]P[\Omega_2] + P[A|\Omega_3]P[\Omega_3]$$

Marginalization: Partition the sample space Ω and measure the probability A



$$P[A] = P[A \cap \Omega_1] + P[A \cap \Omega_2] + P[A \cap \Omega_3]$$
$$P[A] = P[A|\Omega_1]P[\Omega_1] + P[A|\Omega_2]P[\Omega_2] + P[A|\Omega_3]P[\Omega_3]$$

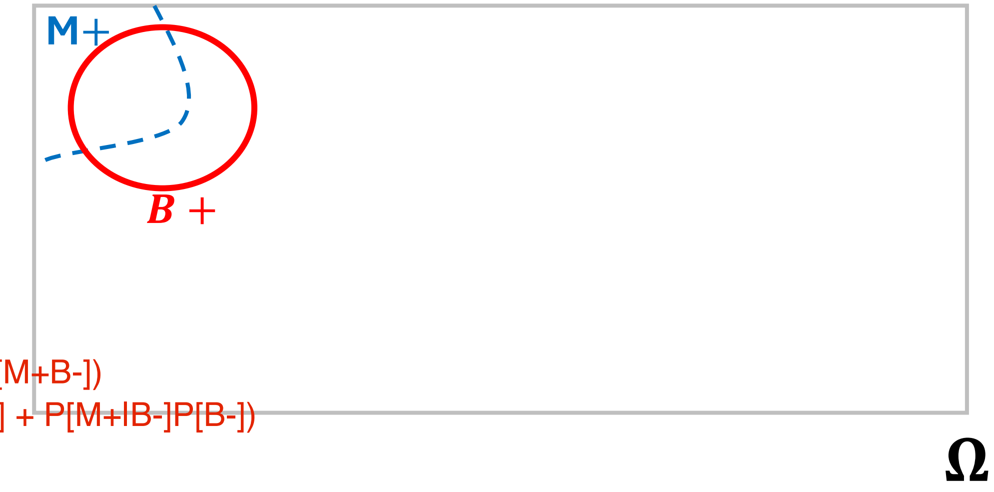
Revisit the breast cancer example and compute the diagnostic probability

Given Information:

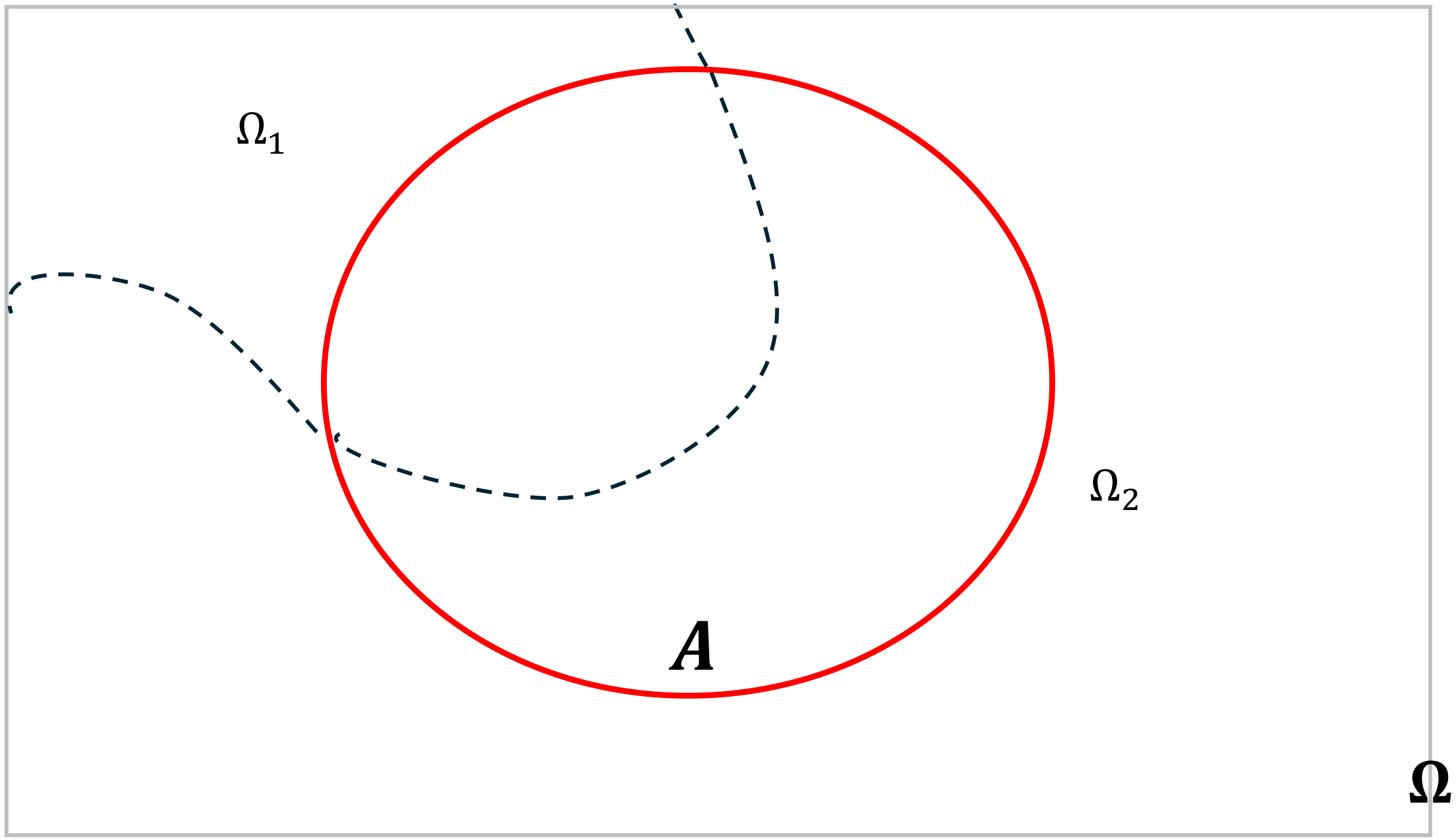
- P[B+]=0.008
- P[M+|B+]=0.9 and P[M+|B-]=0.07

- P[B+|M+] =?

P[B+M+]/ P[M+]
= P[M+|B+] P[B+]/ (P[M+B+] + P[M+B-])
= P[M+|B+]P[B+]/(P[M+|B+]P[B+] + P[M+|B-]P[B-])



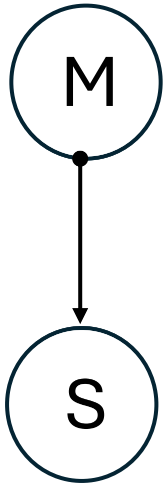
Bayes Theorem



$$P[\Omega_1|A] = \frac{P[A|\Omega_1] \cdot P[\Omega_1]}{P[A|\Omega_1] \cdot P[\Omega_1] + P[A|\Omega_2] \cdot P[\Omega_2]}$$

Computation of Posterior Probability and its Sensitivity to Prior Probability.

EX] Suppose there are two events in causal relationship like **meningitis** and **stiff neck**



prior knowledge

- $P[M] = 2.0e - 5$
- $P[S|M] = 0.7$
- $P[S|M^c] = 0.0099$

$$
\begin{aligned}
P[M|S] &= \frac{P[M \cap S]}{P[S]} \\
&= \frac{P[M \cap S]}{P[S \cap M] + P[S \cap M^c]} \\
&= \frac{P[S|M] \cdot P[M]}{P[S|M] \cdot P[M] + P[S|M^c] \cdot P[M^c]} = 0.0014
\end{aligned}
$$

Posterior P[M|S] is more fragile than the causal direction P[S|M] because of P[M].
P[M] is affected by epidemic, while causal P[S|M] reflects the way meningitis works.

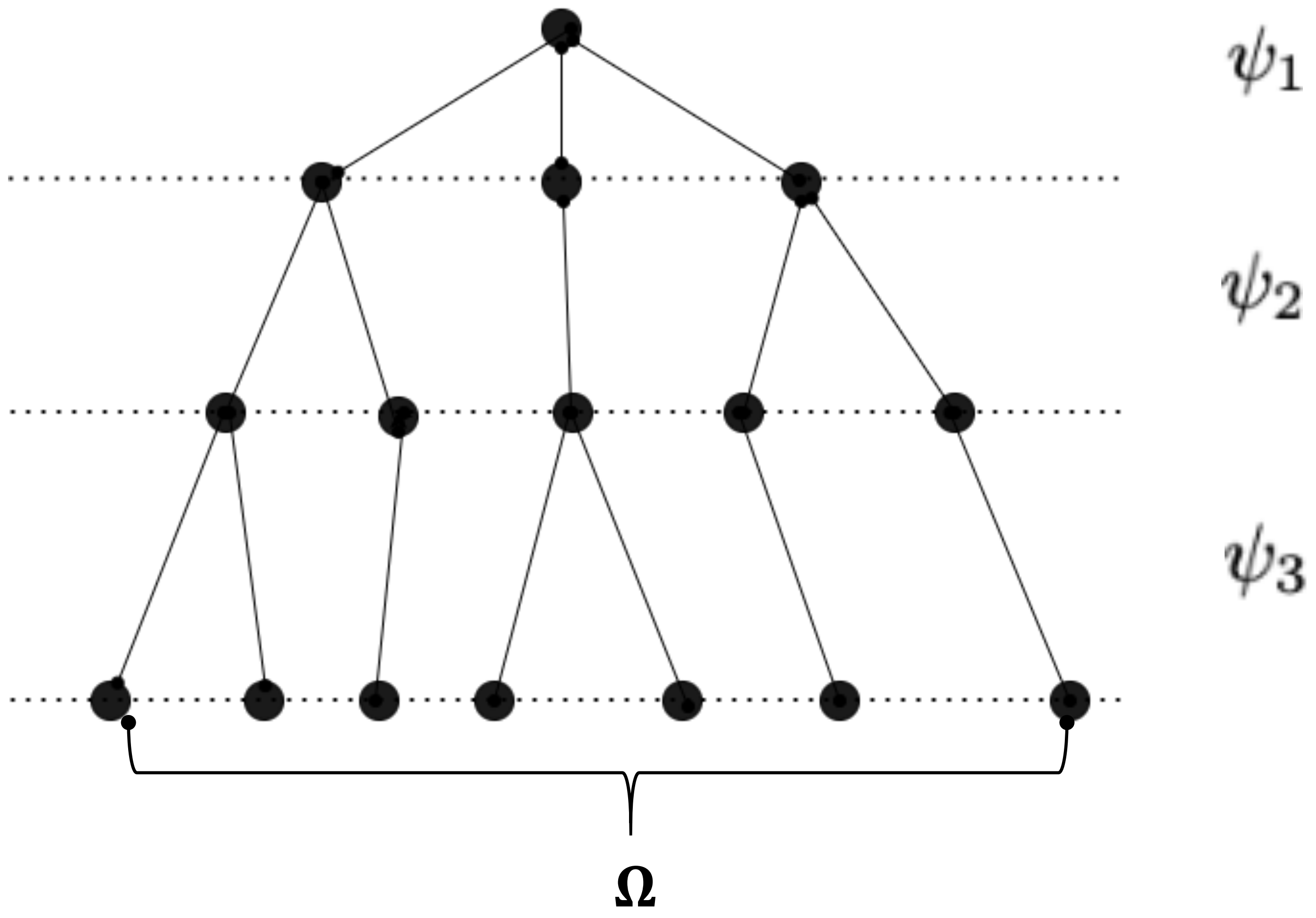
The Conditioning Probability for

- computing diagnostic probability (posterior probability)
- solving complex problem

Conditional Probability: TA recitation

[1] Solving Complex Problem (Tree Diagram)

We can draw a sample space in a tree diagram. $\omega = (\psi_1, \psi_2, ..., \psi_N)$



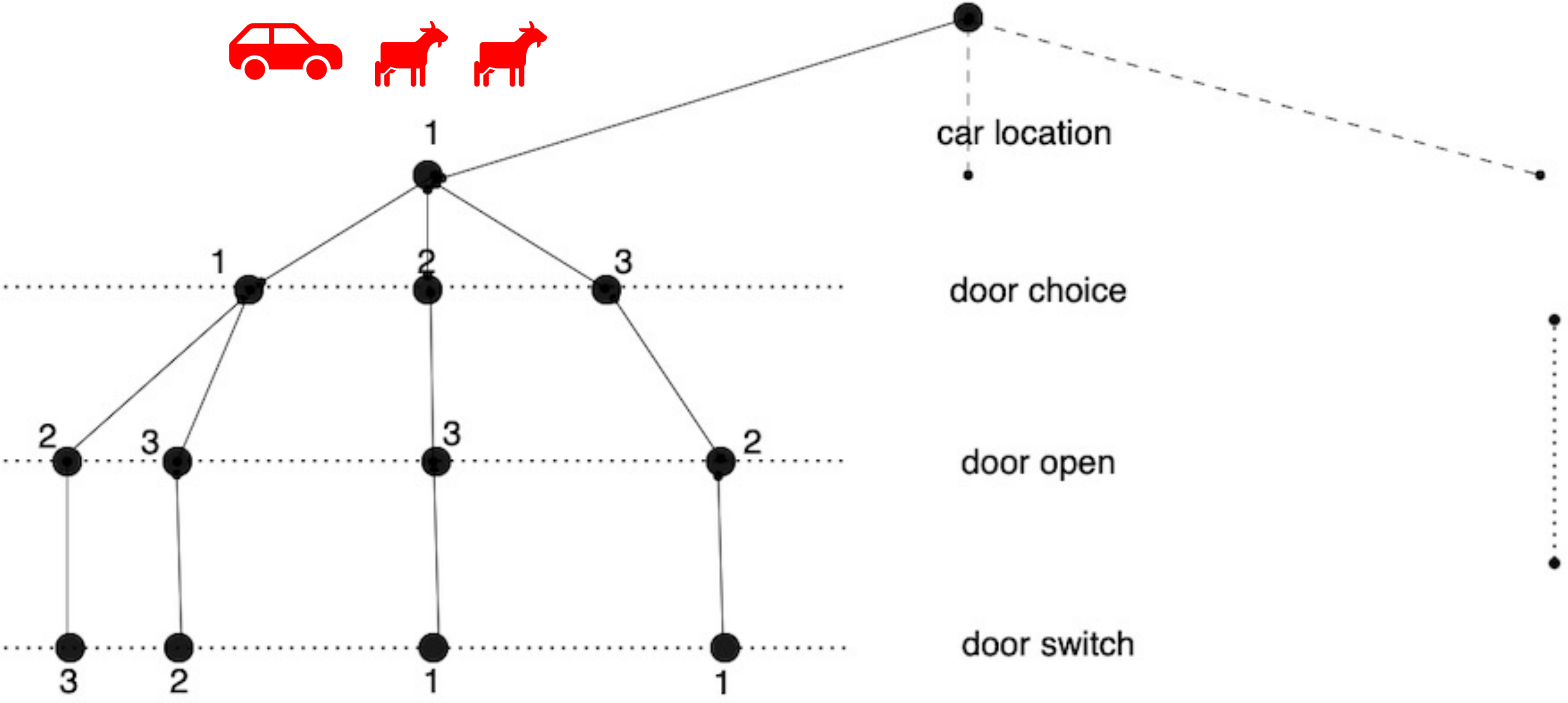
Conditional Probability:
[1] Solving Complex Problem (Tree Diagrams)

Monty Hall Problem

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say number 1, and the host opens another door, say number 2, which has a goat. He says to you, "Do you want to pick door number 3?" Is it to your advantage to switch your choice of doors?

$P_{\text{no-switch}}[\text{Win}]$ vs. $P_{\text{switch}}[\text{Win}]$

Monty Hall Problem



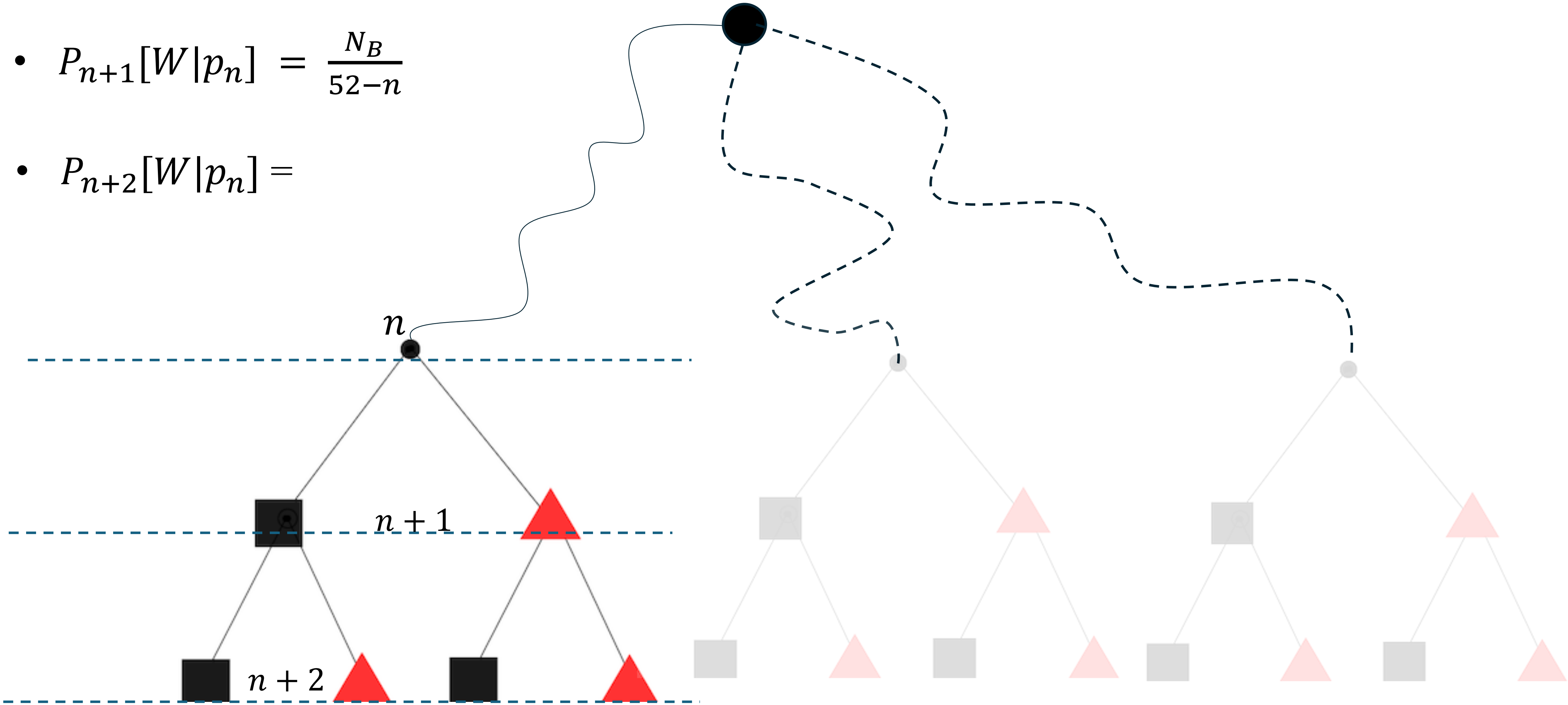
Conditional Probability: TA recitation

[2] Solving Complex Problem (Tree diagrams and Induction)

A game with a deck of 52 cards, 26 are red and 26 are black.
Two option of "taking" or "skipping" the top card.
If you skip the top card, then that card is revealed
and we continue playing with the remaining deck.
If you take the top card, then the game ends; you win if the card black, and you lose if it was red. If we get to a point where there is only one card left in the deck, you must take it.

- Prove that you have no better strategy than to take the top card.
- $P_1[W] = \frac{26}{52} = \frac{1}{2}$
- $P_{n+1}[W] = P_{n+2}[W]$

- $P_1[W] = \frac{26}{52} = \frac{1}{2}$
- $P_{n+1}[W|p_n] = \frac{N_B}{52-n}$
- $P_{n+2}[W|p_n] =$



Bayes Rule in ML as an Inference Method

Inference in Machine Learning

$$f_w(x) = y$$

\+ unknown
\+ inaccessible
\+ hidden

+accessible D: $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_{n-1}, y_{n-1}), (x_n, y_n)$

- ML assumes there exist a model that generates the data
- ML learns w from the data.
- Q: $argmax_w P(W|D)$

Bayes Rule in Machine Learning as an Inference Method

$$P(w|D) = \frac{p(w,D)}{P(D)} = \frac{p(D|w)p(w)}{p(D)}$$

Bayesian Probability	Frequentist Probability
+ quantification uncertainty + prior density (expert knowledge)	+ relative frequency as # trials goes ∞ + w exists as a fixed point

Q: The chances of detecting life on Mars?

This questions is asking Bayesian probability, which is belief, degree of certainty. This cannot be about relative occurrence.

Bayes Rule in Machine Learning as an Inference Method

$$P(w|D) = \frac{p(w, D)}{P(D)} = \frac{p(D|w)p(w)}{p(D)}$$

Frequentist vs. Bayes Estimation

- $w = argmax\ P(D|w)$: Maximum Lliklihood Estimation (MLE) Frequentist
- $w = argmax\ p(w|D) = \frac{p(D|w)p(w)}{p(D)}$: Maximum A posteriori Estimation (MAP)

Bayes

Going back to Probability 101

Random Variable

$X(\omega)\colon \Omega \to \mathbb{R}$

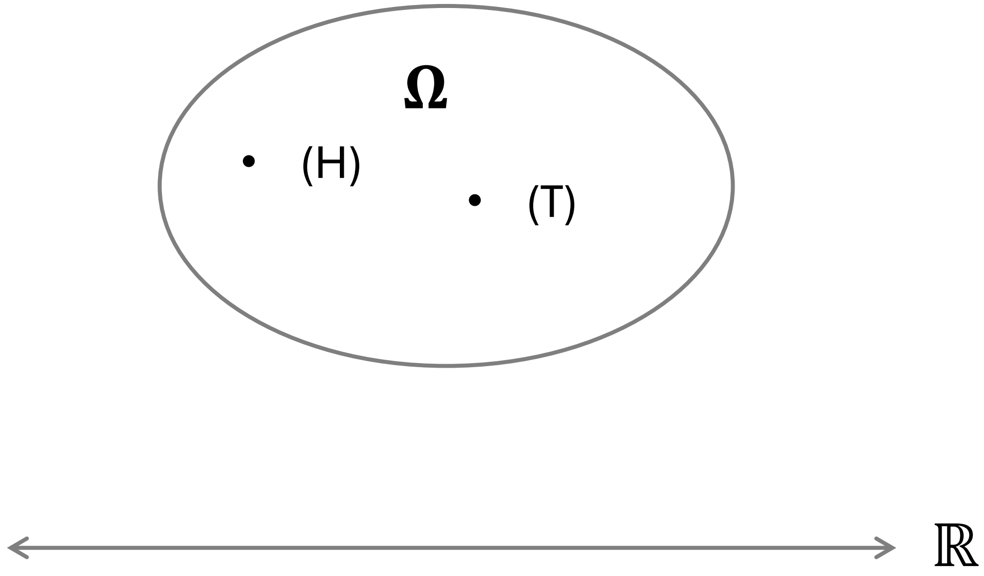
- Bernoulli
- Binomial
- Gaussian density

Random Variable $X(\omega): \Omega \to \mathbb{R}$

- Real-Valued Function
- PMF / PDF

(1) Bernoulli R.V

Ex] Suppose a coin tossed one time.
Let X be the indicator function for **Tail event.**



PDF / PMF as the derivative of CDF $F_X(x) = P[X \le x]$

$$
\begin{aligned}
P[x < X \le x + h] &= F_X(x+h) - F_x(x) \\
&= \frac{F_X(x+h) - F_X(x)}{h} \cdot h \\
&= F'_X(x) \cdot h \\
&= f_X(x) \cdot h
\end{aligned}
$$

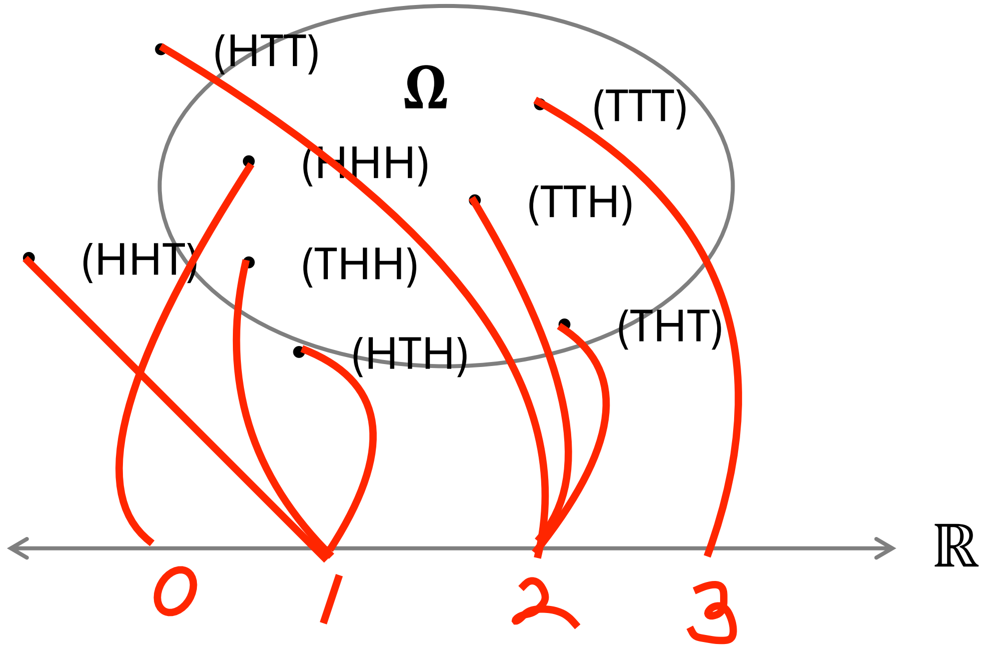
$$\int_a^b f_X(x)\,dx = F_X(b) - F_X(a)$$

Random Variable $X(\omega): \Omega \to \mathbb{R}$

- Real-Valued Function
- PMF / PDF.

(1) Binomial R.V

Ex] Suppose a coin tossed 3 time.
Let X be the count for **Tail event among N trials.**



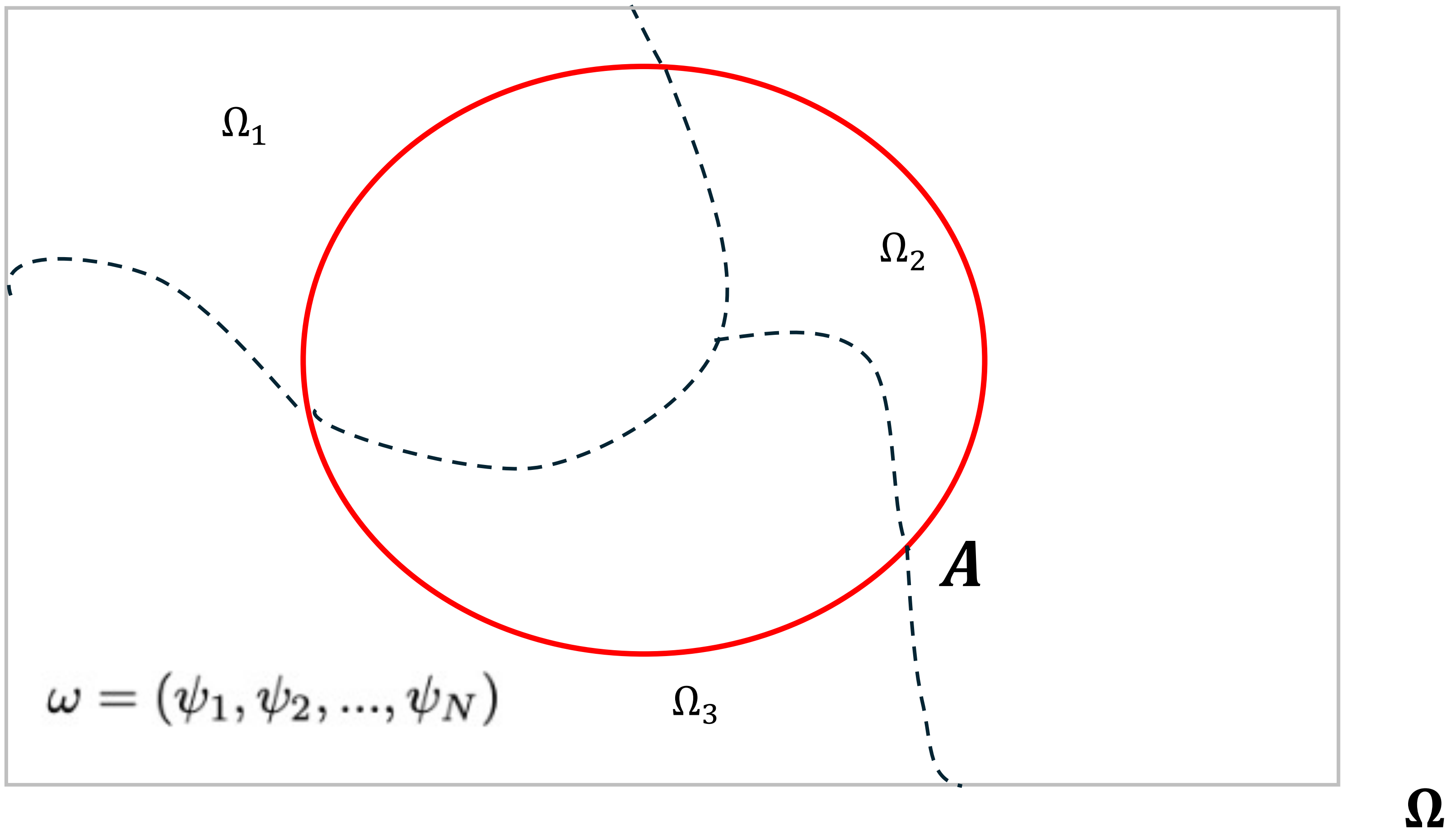
Again, Independence and Marginalization

Independent Events ↔

$$P[A \cap B] = P[A] \cdot P[B]$$

$$P[A \cap B \cap C] = P[A] \cdot P[B] \cdot P[C]$$

Marginalization: Partition the sample space Ω and measure the probability A



$$P[A] = P[A \cap \Omega_1] + P[A \cap \Omega_2] + P[A \cap \Omega_3]$$
$$P[A] = P[A|\Omega_1]P[\Omega_1] + P[A|\Omega_2]P[\Omega_2] + P[A|\Omega_3]P[\Omega_3]$$

X and Y are independent $\leftrightarrow$

$$P(x, y) = P(x) \cdot P(y) \quad \forall\, x, y$$

$$f(x, y) = f(x) \cdot f(y) \quad \forall\, x, y$$

EX1) $f(x, y) = \begin{Bmatrix} 2\, e^{-x} e^{-y}, 0 \le y \le x < \infty \\ 0 \qquad , elsewhere \end{Bmatrix}$ independent?

EX) $f(x, y) = \frac{1}{2\pi}\, e^{\frac{-(x^2 + y^2)}{2}}$?

Marginalization

$$P_X(x) = \sum_y P_{XY}(x, y)$$

$$f_X(x) = \int_y f_{XY}(x, y)\, dy$$

Mean and Variance
:two statistics to describe the behavior of a random variable

- Expectation / Mean

$$E[X] = \sum_x x\, P(X)$$

- Linearity of Expectation

$$E[aX + b] = \sum_x ax\, P(x) + bP(x) \; = aE[X] + b$$

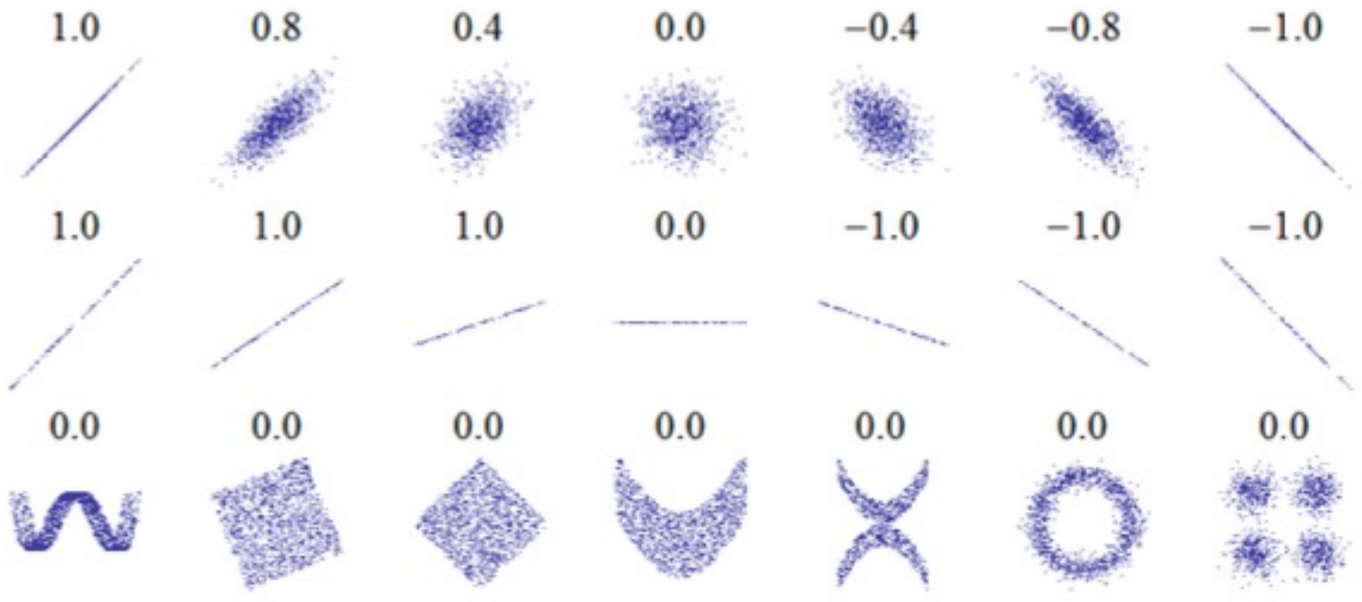
- Variance

$$VAR[X] = \sum_{x}(x - E[x])^2\, P(X) \;= E\left((X - E[X])^2\right) = \; E[\,X^2] - \; E[X]^2$$

$$VAR[aX + b] = \; a^2 VAR[X]$$

$$VAR[aX + bY] = E[(aX + bY \; - a\mu_X \; - b\mu_Y)^2]$$
$$= \; E[a^2(X - \mu_X)^{\,2} + b^2(Y - \mu_Y)^2 + \; 2ab(X - \mu_X)(Y - \mu_Y)]$$
$$= \; E[a^2(X - \mu_X)^{\,2} + b^2(Y - \mu_Y)^2 + \; 2ab(X - \mu_X)(Y - \mu_Y)]$$
$$= a^2 E[(X - \mu_X)^{\,2}] + b^2 E[(Y - \mu_Y)^2] + \; 2ab\, E[(X - \mu_X)(Y - \mu_Y)]$$
$$= a^2 VAR[X] + b^2 VAR[Y] + \; 2ab\, E[(X - \mu_X)(Y - \mu_Y)]$$

Covariance shows how X and Y linearly related



From Figure 3.1 Murphy, Introduction

This figures presents correlation coefficient $\rho = \frac{COV(X,Y)}{\sqrt{VAR(X)}\sqrt{VAR(Y)}}$ for several sets of data points.

Multiple Variable/ Random Vector

In practice,
we can access to the population only through data points.
Data is the realization of the repetitive process of the R.V.
One data point does not show much information,
but what if we observe 100, 1,000, 10,000 realization?

$$\vec{X} = (X_1, X_2, X_3, \ldots, X_N)$$

Random Vector $\vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}$

- Mean vector: $E[\overrightarrow{X}] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_D] \end{bmatrix}$

- Covariance Matrix $\mathrm{Cov}[\boldsymbol{x}] \triangleq \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\mathsf{T}}\right] \triangleq \boldsymbol{\Sigma}$

$$= \begin{pmatrix} \mathbb{V}[X_1] & \mathrm{Cov}[X_1, X_2] & \cdots & \mathrm{Cov}[X_1, X_D] \\ \mathrm{Cov}[X_2, X_1] & \mathbb{V}[X_2] & \cdots & \mathrm{Cov}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_D, X_1] & \mathrm{Cov}[X_D, X_2] & \cdots & \mathbb{V}[X_D] \end{pmatrix}$$

Q: The R.Vs are independent and identical,
how can we express the Covariance Matrix?