10-315: Intro to Machine Learning

MACHINE LEARNING DEPARTMENT

# Probability Review

Siddharth Ancha

# Sources

**Chris Cremer's slides**

https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial1.pdf

**Jeff Howbert's slides**

https://courses.washington.edu/css490/2012.Winter/lecture_slides/02_math_essentials.pdf

**Aarti Singh's slides**

https://www.cs.cmu.edu/~aarti/Class/10315_Fall19/lecs/Lecture2.pdf

# Sample Space

- The sample space Ω is the set of possible outcomes of an experiment. Points ω in Ω are called sample outcomes, realizations, or elements. Subsets of Ω are called Events.

- Example. If we toss a coin twice then Ω = {HH,HT, TH, TT}. The event that the first toss is heads is A = {HH,HT}

- We say that events A1 and A2 are disjoint (mutually exclusive) if $A_i \cap A_j = \{\}$
  - Example: first flip being heads and first flip being tails

# Probability

- We will assign a real number P(A) to every event A, called the probability of A.
- To qualify as a probability, P must satisfy three axioms:
  - Axiom 1: P(A) ≥ 0 for every A
  - Axiom 2: P(Ω) = 1
  - Axiom 3: If A1,A2, . . . are disjoint then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

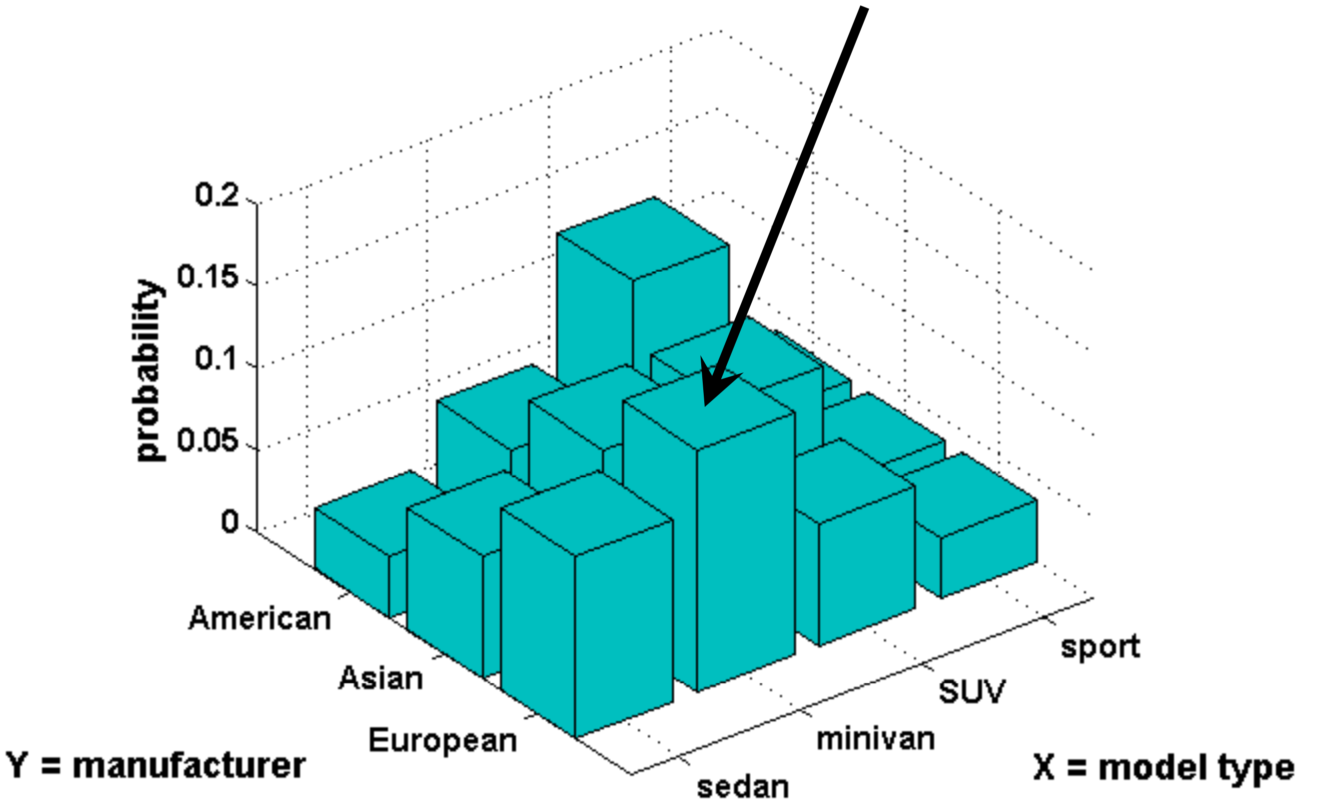
# Joint and Conditional Probabilities

- Joint Probability
  - $P(X,Y)$
  - Probability of X and Y

- Conditional Probability
  - $P(X|Y)$
  - Probability of X given Y

# Example of multivariate distribution

**joint probability: *p*( *X* = minivan, *Y* = European ) = 0.1481**



# Multivariate probability distributions

- *Marginal* probability
  - Probability distribution of a single variable in a joint distribution
  - Example: two random variables *X* and *Y*:

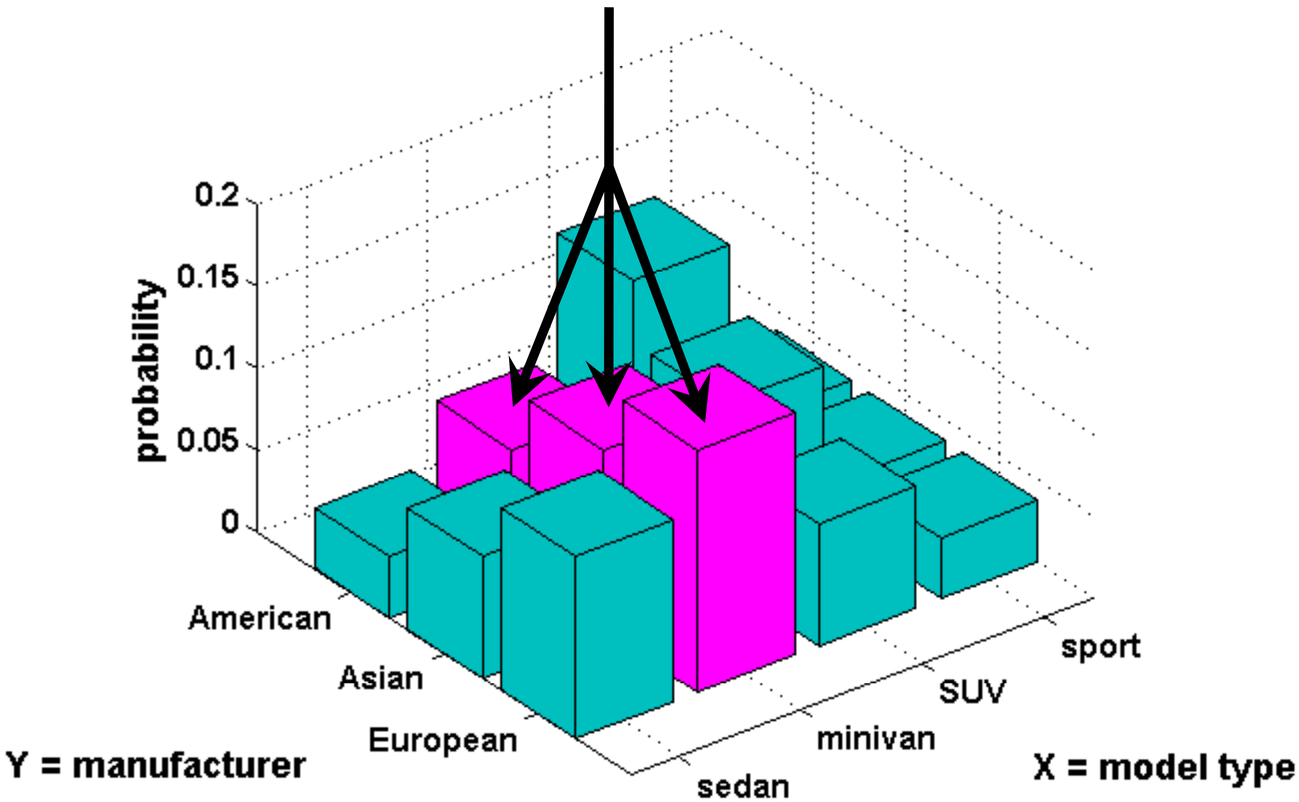
$$p(X = x) = \sum_{b=\text{all values of } Y} p(X = x, Y = b)$$

- *Conditional* probability
  - Probability distribution of one variable *given* that another variable takes a certain value
  - Example: two random variables *X* and *Y*:

$$p(X = x \mid Y = y) = p(X = x, Y = y) / p(Y = y)$$

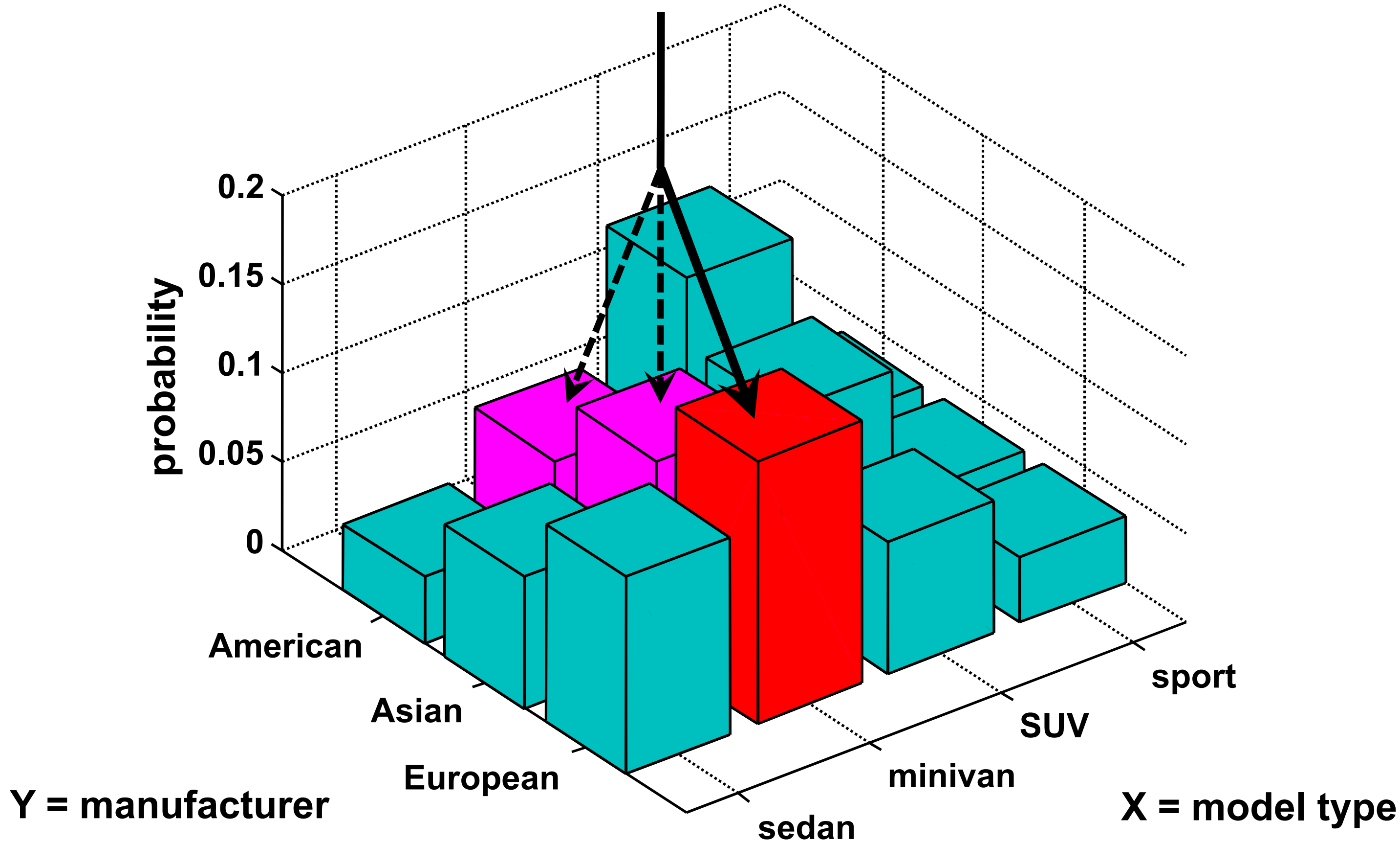
# Example of marginal probability

**marginal probability: *p*( *X* = minivan ) = 0.0741 + 0.1111 + 0.1481 = 0.3333**



# Example of conditional probability

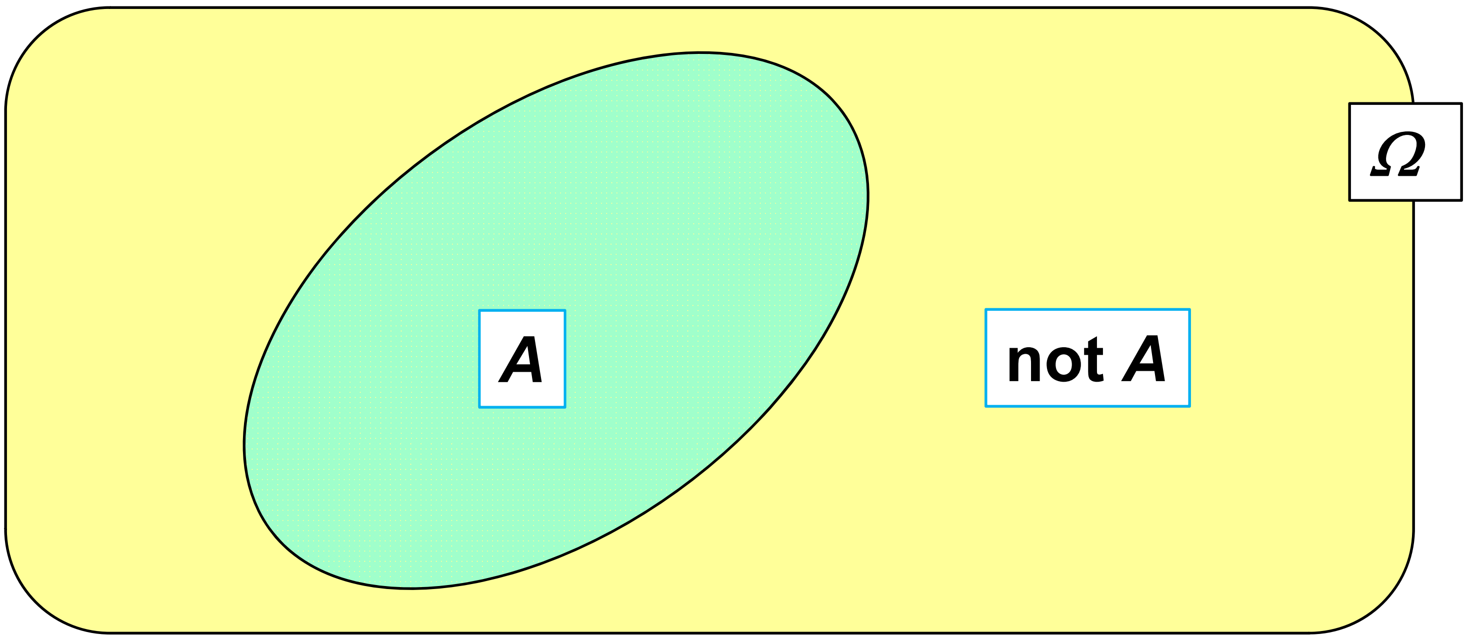
**conditional probability: $p( Y = \text{European} \mid X = \text{minivan} ) =$**
**$0.1481 / ( 0.0741 + 0.1111 + 0.1481 ) = 0.4433$**



# Complement rule

Given: event *A*, which can occur or not

$$p(\text{ not } A) = 1 - p(A)$$



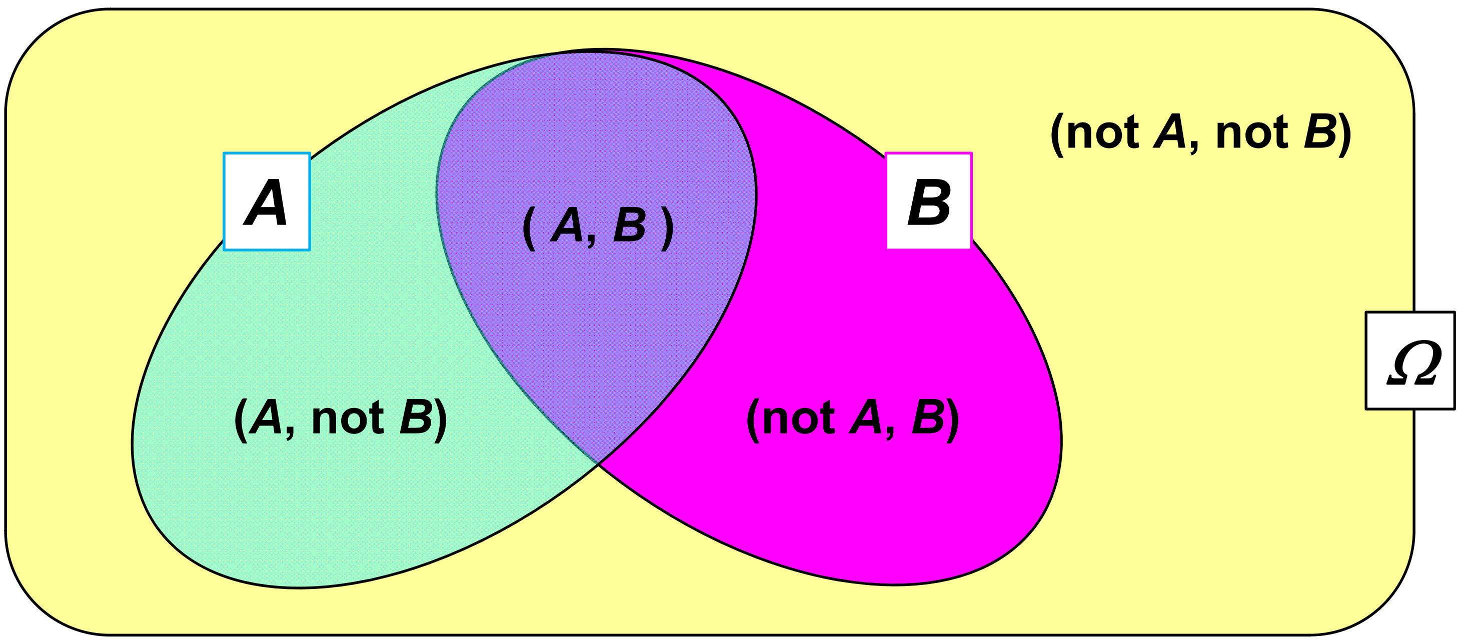

areas represent relative probabilities

# Product rule

Given: events *A* and *B*, which can co-occur (or not)

$$p(A, B) = p(A \mid B) \cdot p(B)$$

(same expression given previously to define conditional probability)



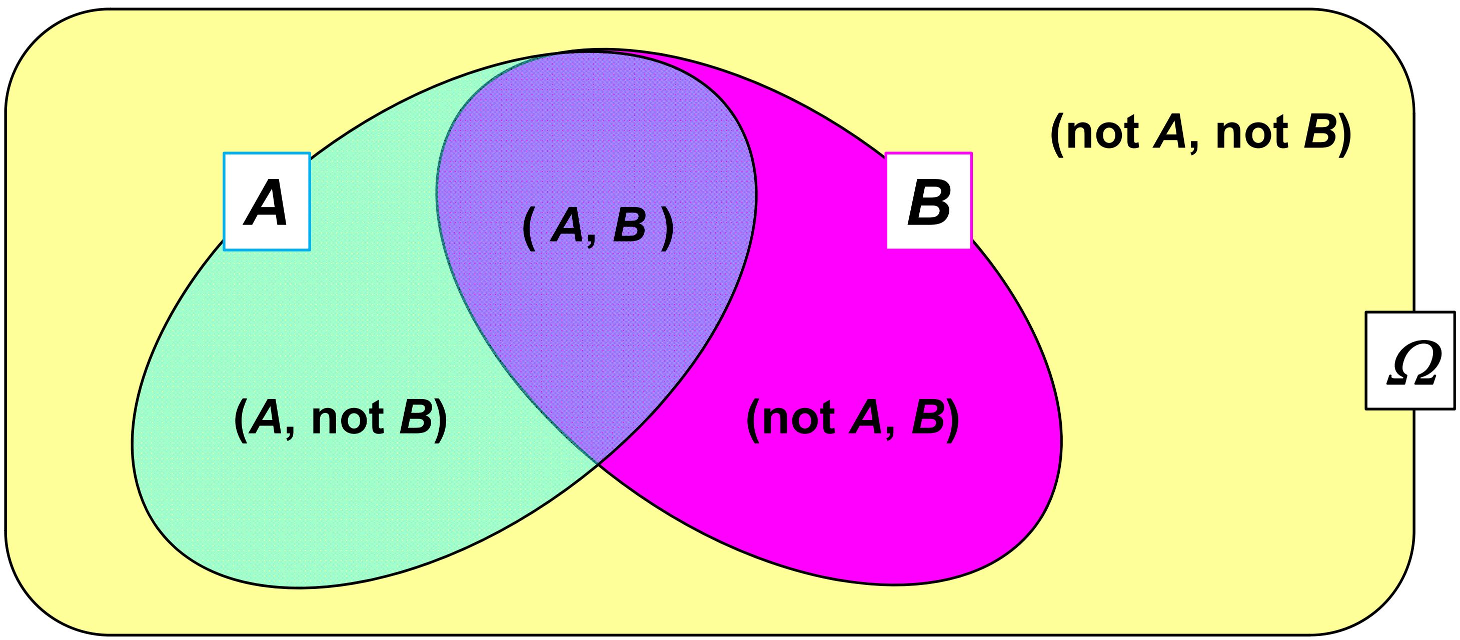

areas represent relative probabilities

# Rule of total probability

Given: events *A* and *B*, which can co-occur (or not)

$$p(\ A\ ) = p(\ A, B\ ) + p(\ A, \text{not } B\ )$$

(same expression given previously to define marginal probability)




areas represent relative probabilities

# Marginalization and Law of Total Probability

- Marginalization (Sum Rule)

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

- Law of Total Probability

$$p(x) = \sum_y p(x \mid y) \cdot p(y)$$

# Indicator Random Variables

Let $A$ be an event.

Then $\mathbf{1}_A$ is random variable such that -

$$\mathbf{1}_A \begin{cases} 1 \text{ if } A \text{ is True} \\ 0 \text{ if } A \text{ is False} \end{cases}$$

Example, if $A$ is the event that tossing biased coin (with bias $p$) results in heads, then

Bernoulli $(p) = \mathbf{1}_A$

**Important!**

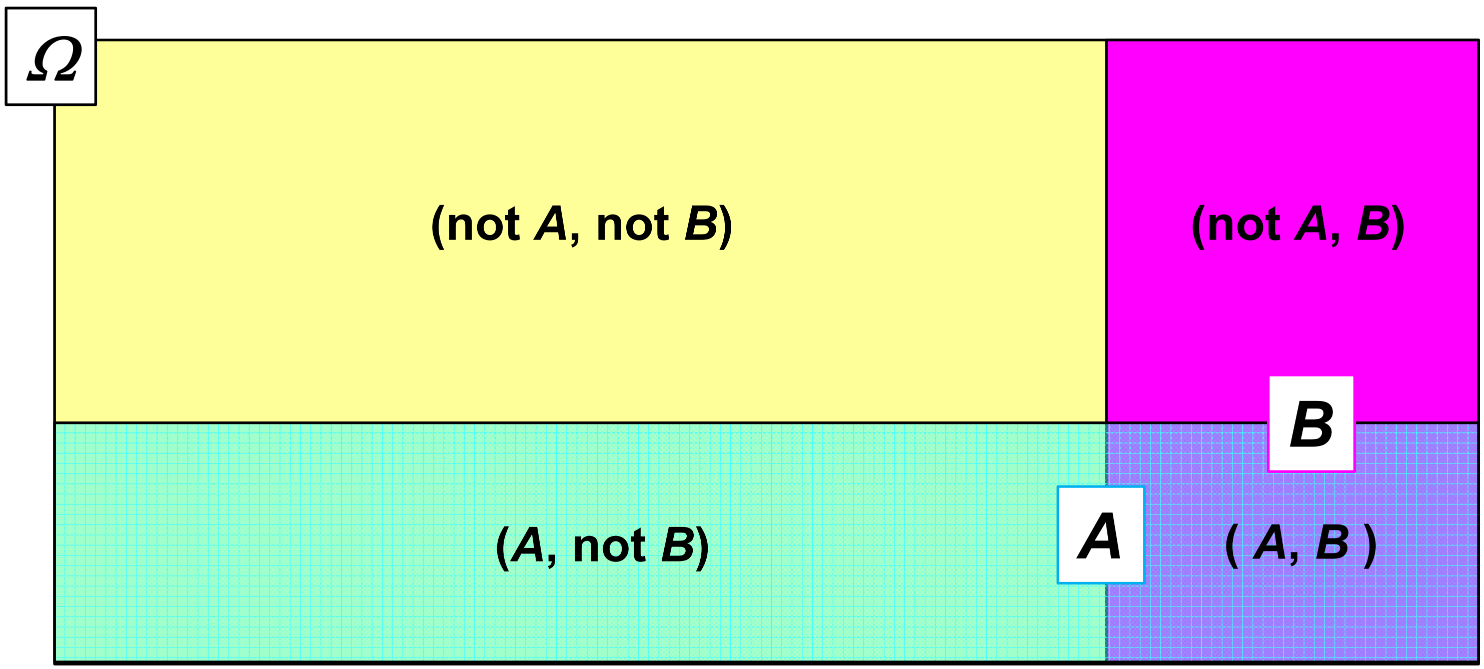
$$\boxed{\mathbb{E}\ [\mathbf{1}_A] = P(A)}$$

Why? $\mathbb{E}\ [\mathbf{1}_A] = 1.P(A) + 0.P(\neg A) = P(A)$

# Independence

Given: events *A* and *B*, which can co-occur (or not)

$$p(A \mid B) = p(A) \quad \text{or} \quad p(A, B) = p(A) \cdot p(B)$$




areas represent relative probabilities

# Independent and Conditional Probabilities

- Assuming that $P(B) > 0$, the **conditional** probability of A given B:
- $P(A|B)=P(AB)/P(B)$
- $P(AB) = P(A|B)P(B) = P(B|A)P(A)$
  - Product Rule

- Two events A and B are **independent** if
- $P(AB) = P(A)P(B)$
  - Joint = Product of Marginals

- Two events A and B are **conditionally independent** given C if they are independent after conditioning on C
- $P(AB|C) = P(B|AC)P(A|C) = P(B|C)P(A|C)$

# Example

- 60% of ML students pass the final and 45% of ML students pass both the final and the midterm *
- What percent of students who passed the final also passed the midterm?

* These are made up values.

# Example

- 60% of ML students pass the final and 45% of ML students pass both the final and the midterm *
- What percent of students who passed the final also passed the midterm?

- Reworded: What percent of students passed the midterm given they passed the final?
- $P(M|F) = P(M,F) / P(F)$
- $= .45 / .60$
- $= .75$

* These are made up values.

# Bayes' Rule

P(A|B) = P(AB) /P(B) (Conditional Probability)

P(A|B) = P(B|A)P(A) /P(B) (Product Rule)

P(A|B) = P(B|A)P(A) / Σ P(B|A)P(A) (Law of Total Probability)

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

$$P(B) = \sum_j P(B \mid A_j) P(A_j)$$

**Rev. Thomas Bayes**

# Bayes' Rule

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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

$$\text{Posterior probability} \propto \text{Likelihood} \times \text{Prior probability}$$

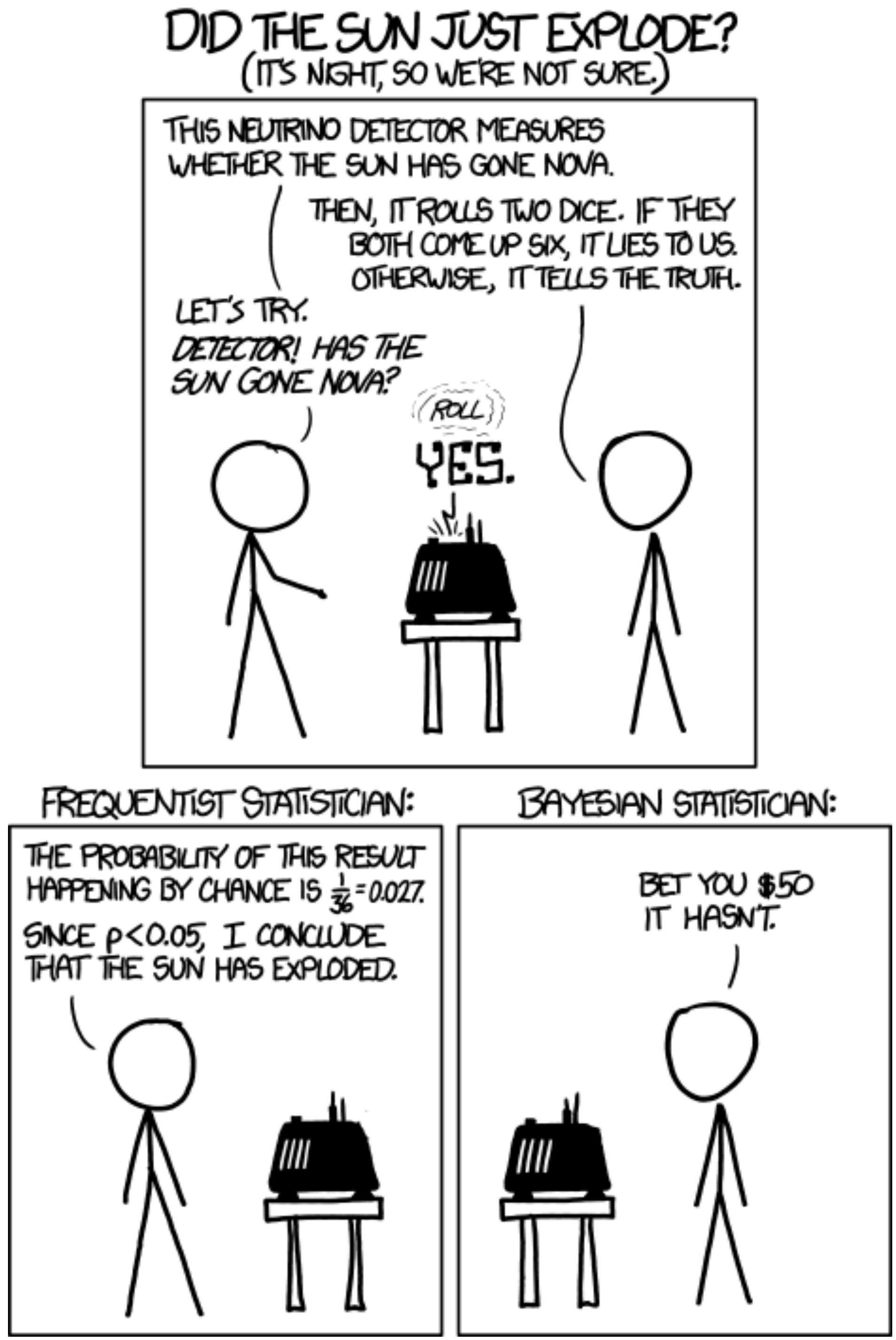



Image from xkcd.com

# Example of Bayes rule

- Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman is forecasting rain for tomorrow. When it actually rains, the weatherman has forecast rain 90% of the time. When it doesn't rain, he has forecast rain 10% of the time. What is the probability it will rain on the day of Marie's wedding?
- Event *A*: The weatherman has forecast rain.
- Event *B*: It rains.
- We know:
  - $p( B ) = 5 / 365 = 0.0137$ [ It rains 5 days out of the year. ]
  - $p( \text{not } B ) = 360 / 365 = 0.9863$
  - $p( A \mid B ) = 0.9$ [ When it rains, the weatherman has forecast rain 90% of the time. ]
  - $p( A \mid \text{not } B ) = 0.1$ [When it does not rain, the weatherman has forecast rain 10% of the time.]

# Example of Bayes rule, cont'd.

- We want to know $p( B \mid A )$, the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes rule:

1. $p( B \mid A ) = p( A \mid B ) \cdot p( B ) / p( A )$
2. $p( A ) = p( A \mid B ) \cdot p( B ) + p( A \mid \text{not } B ) \cdot p( \text{not } B ) =$
   $(0.9)(0.014) + (0.1)(0.986) = 0.111$
3. $p( B \mid A ) = (0.9)(0.0137) / 0.111 = 0.111$

- The result seems unintuitive but is correct. Even when the weatherman predicts rain, it only rains only about 11% of the time. Despite the weatherman's gloomy prediction, it is unlikely Marie will get rained on at her wedding.

# Discrete vs Continuous Random Variables

- Discrete: can only take a countable number of values
- Example: number of heads
- Distribution defined by probability mass function (pmf)
- Marginalization: $$p(x) = \sum_{y} p(x, y)$$

- Continuous:  can take infinitely many values (real numbers)
- Example: time taken to accomplish task
- Distribution defined by probability density function (pdf)
- Marginalization:

$$p(x) = \int_{y} p(x, y)\, dy$$

# Probability Distribution Statistics

- Mean: E[x] = μ = first moment = $\int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x$ Univariate continuous random variable

  = $\sum_{i=1}^{\infty} x_i\, p_i$ Univariate discrete random variable

- Variance: Var(X) = $\mathrm{E}\left[(X-\mu)^2\right]$

$$
\begin{aligned}
&= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + (\mathrm{E}[X])^2 \\
&= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2
\end{aligned}
$$

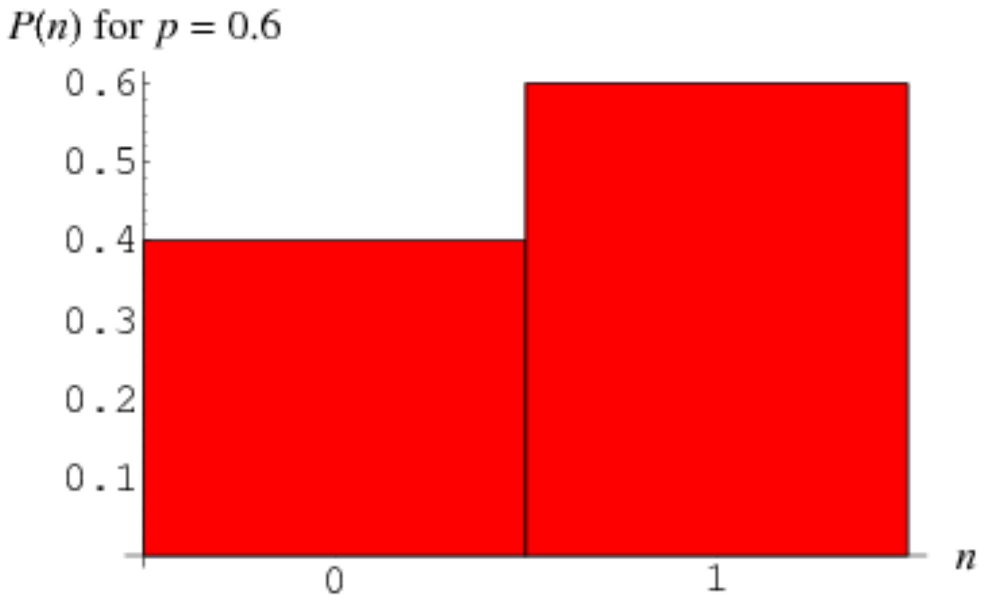
- Nth moment = $\int_{-\infty}^{\infty} (x-c)^n\, f(x)\, dx$

# Bernoulli Distribution

- Input: x ∈ {0, 1}
- Parameter: $\mu$

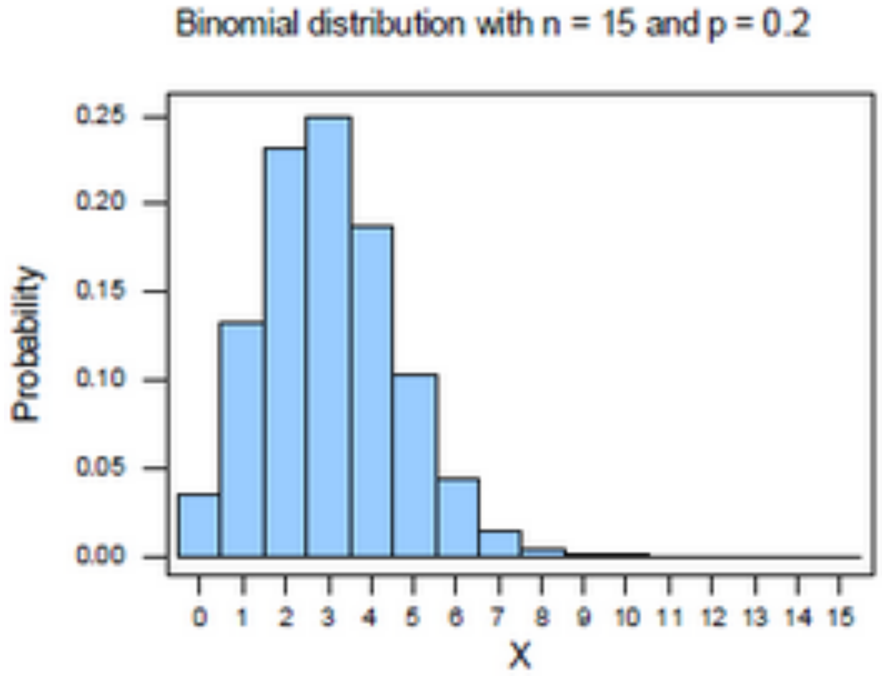
- Example: Probability of flipping heads (x=1)

$$\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$




- Mean = E[x] = $\mu$
- Variance = $\mu(1 - \mu)$

# Binomial Distribution




- Input: m = number of successes
- Parameters: N = number of trials

  μ = probability of success
- Example: Probability of flipping heads m times out of N independent flips with success probability $\mu$

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

- Mean = E[x] = N$\mu$
- Variance = N$\mu(1-\mu)$

# Multinomial Distribution

- The multinomial distribution is a generalization of the binomial distribution to k categories instead of just binary (success/fail)
- For n independent trials each of which leads to a success for exactly one of k categories, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories
- Example: Rolling a die N times

# Multinomial Distribution

- Input: $m_1 \ldots m_K$ (counts)
- Parameters: N = number of trials

  $\boldsymbol{\mu} = \mu_1 \ldots \mu_K$ probability of success for each category, $\Sigma\boldsymbol{\mu}=1$

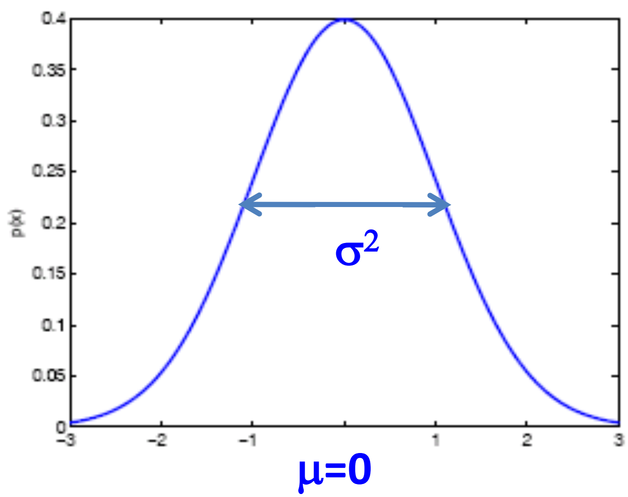
$$\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

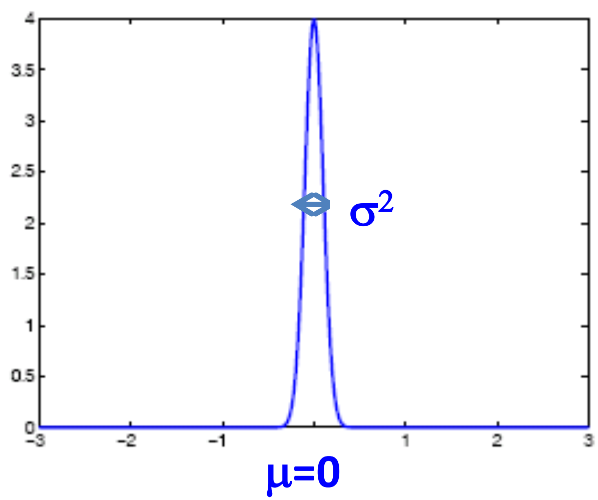
- Mean of $m_k$: $N\mu_k$
- Variance of $m_k$: $N\mu_k(1-\mu_k)$

# 1-dim Gaussian distribution

X is Gaussian N(μ,σ²)

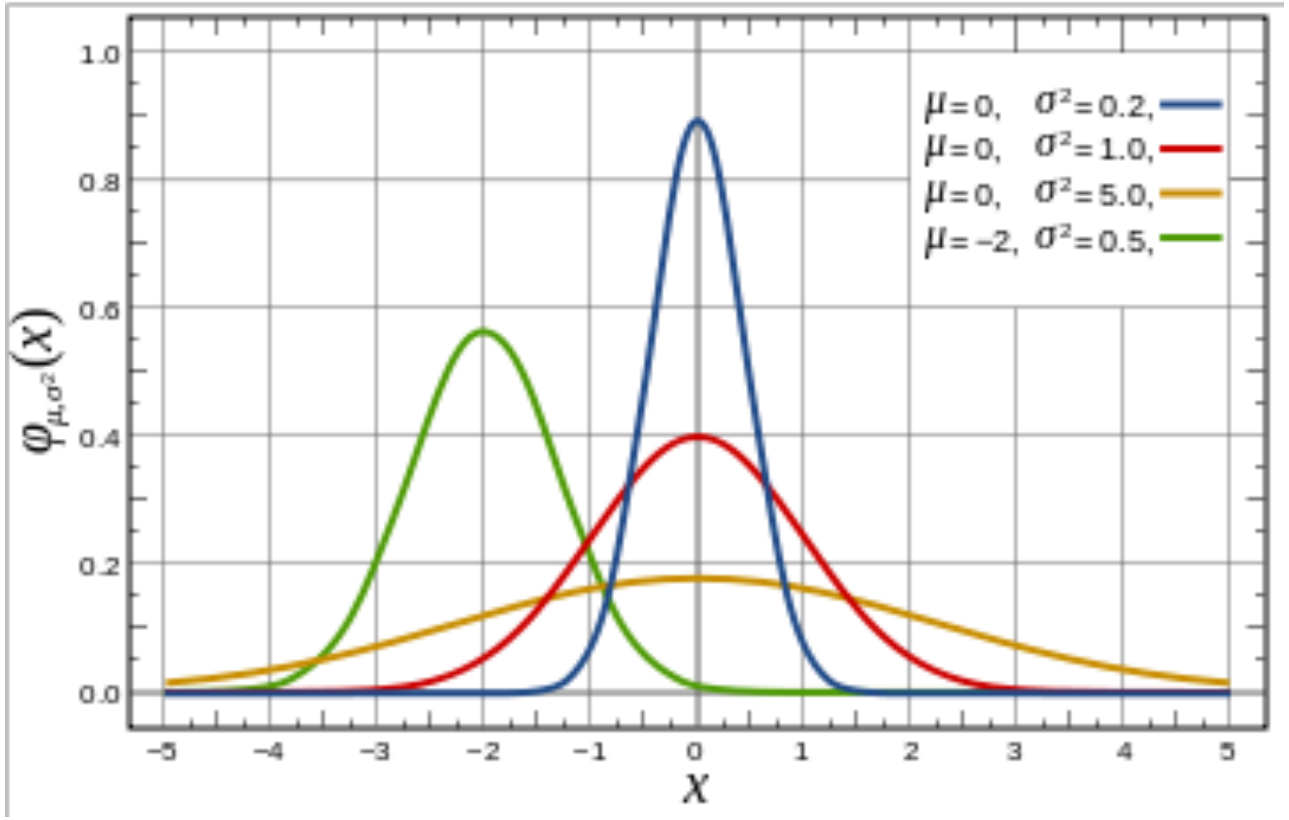
$$P(X = x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



# Gaussian Distribution

- Gaussians with different means and variances



# d-dim Gaussian distribution

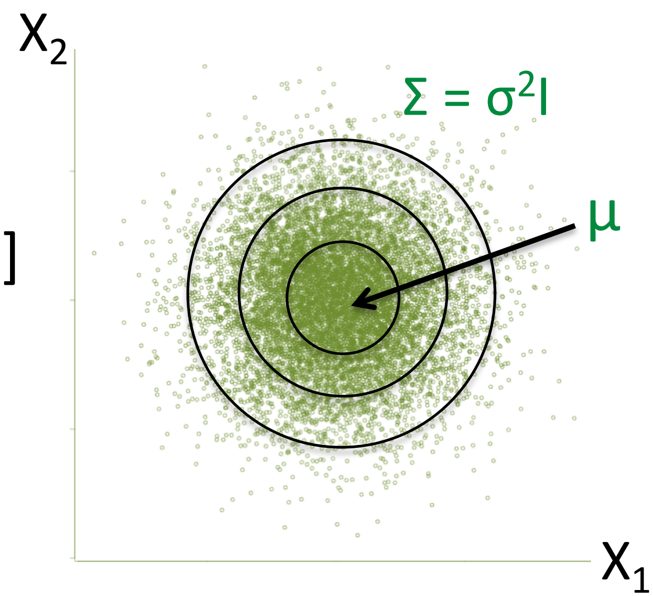
X is Gaussian N(μ, Σ)

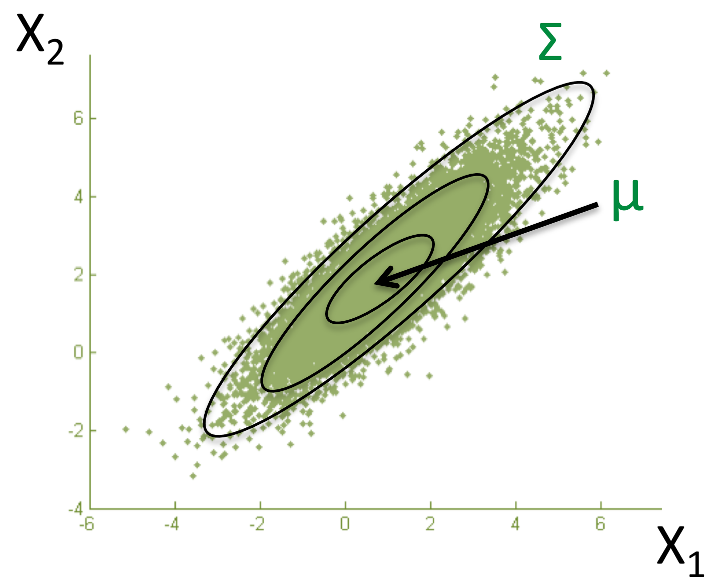
μ is d-dim vector, Σ is dxd dim matrix

$$P(X = x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

d=2

$X = [X_1; X_2]$



# Questions?