Introduction to Data Science and Engineering

- Bayes theorem

Zhenqin (Michael) Wu / 吳楨欽

School of Computing and Data Science
University of Hong Kong

Slide deck originally created by RB Luo

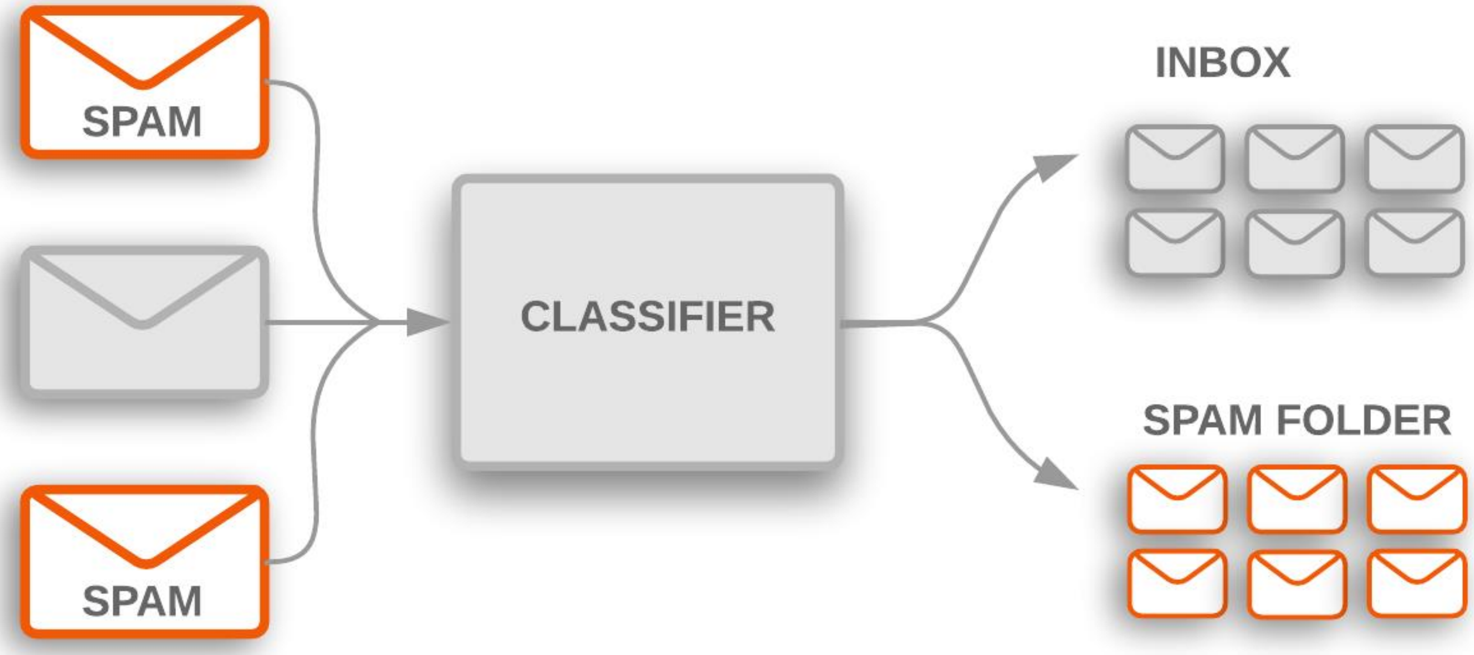
In this lecture

- Basics of Bayes theorem
- Case study: lung cancer and pulmonary nodules
- Bayesian versus frequentist

In this lecture

- **Basics of Bayes theorem**
- Case study: lung cancer and pulmonary nodules
- Bayesian versus frequentist

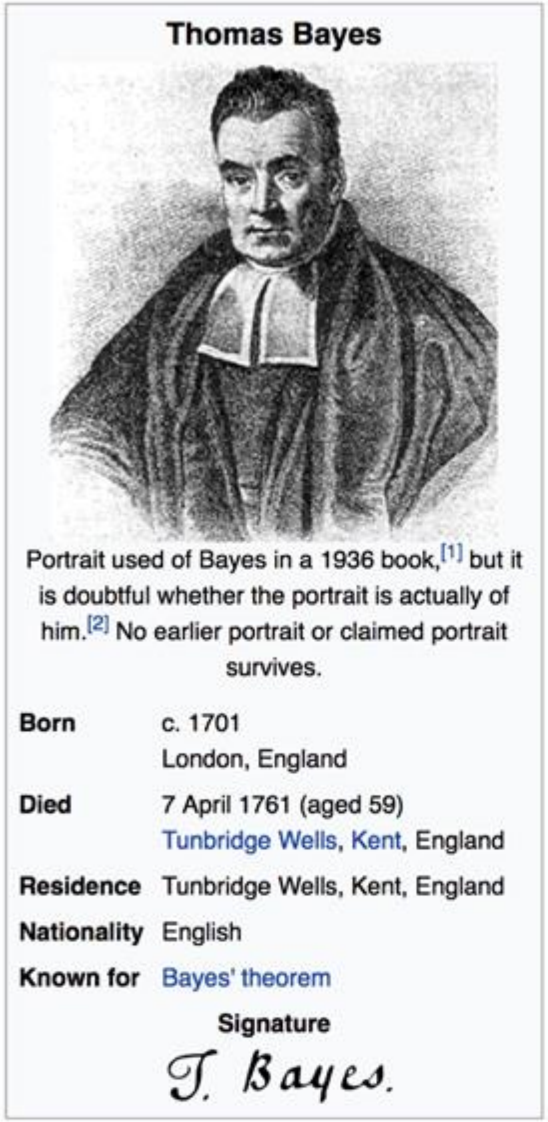
Email spam detection



Introducing Bayes' theorem

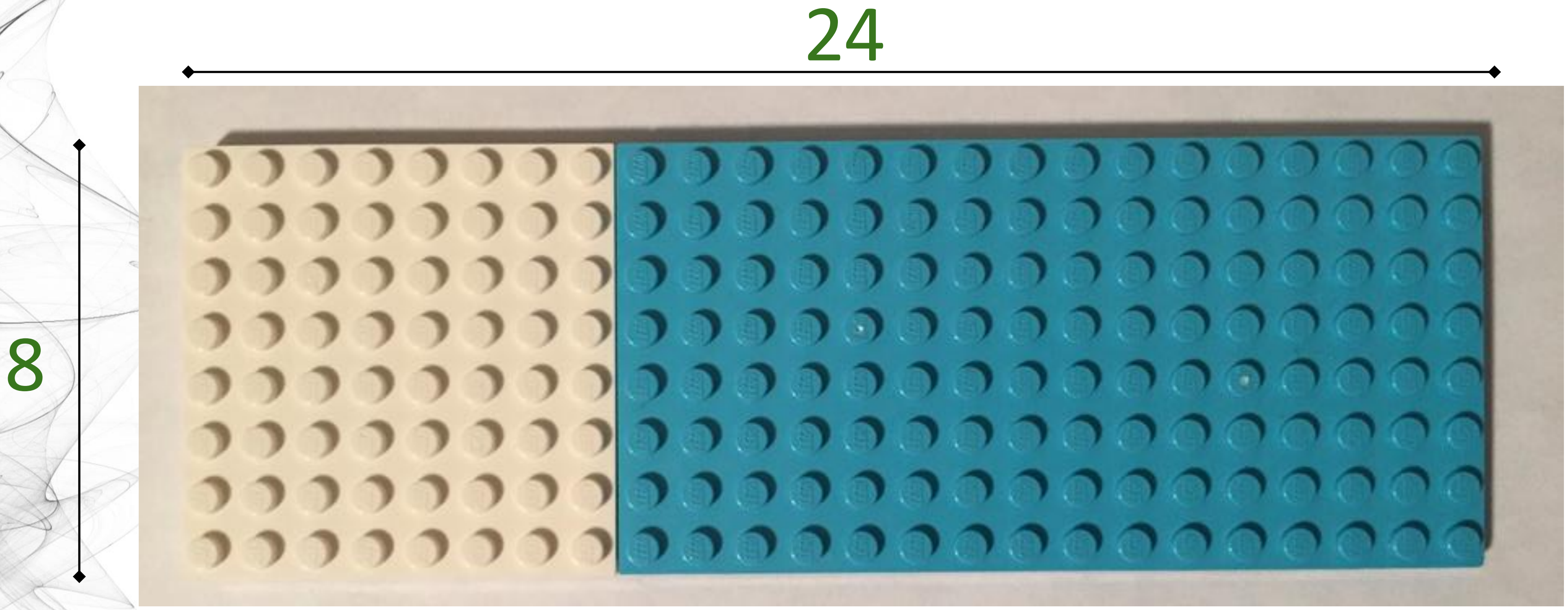
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- B is the observation
- A is the hypothesis/theory
- $\boldsymbol{P(B)}$**: evidence**, overall probability of observing B, regardless of theories.
- $\boldsymbol{P(A)}$**:** **prior**, with no observation, our belief about the theory;
  - Flipping a coin has a 50/50 chance on getting head/tail;
- $\boldsymbol{P(B|A)}$**: likelihood**, if theory A is true, how likely is that we observed B.
- $\boldsymbol{P(A|B)}$**:** **posterior**, after observing B, our belief about theory A.
  - After observing 5 tails in a row, maybe chance of getting tail is higher than 50/50



https://en.wikipedia.org/wiki/Thomas_Bayes

Bayes' theorem with Legos

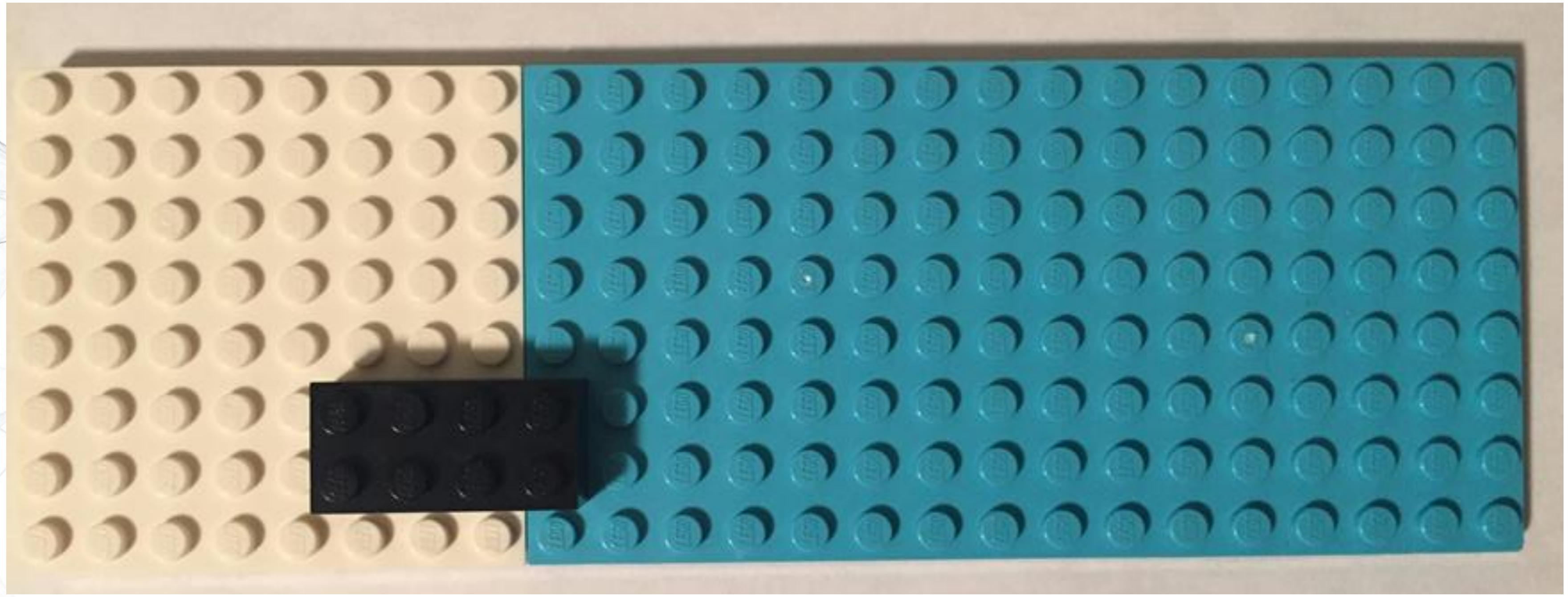


We have two grand theories: white and blue, blue seems more likely

$$P(\text{blue}) = 128/192 = 0.67$$
$$P(\text{white}) = 64/192 = 0.33$$

Inspired by https://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego

Bayes' theorem with Legos

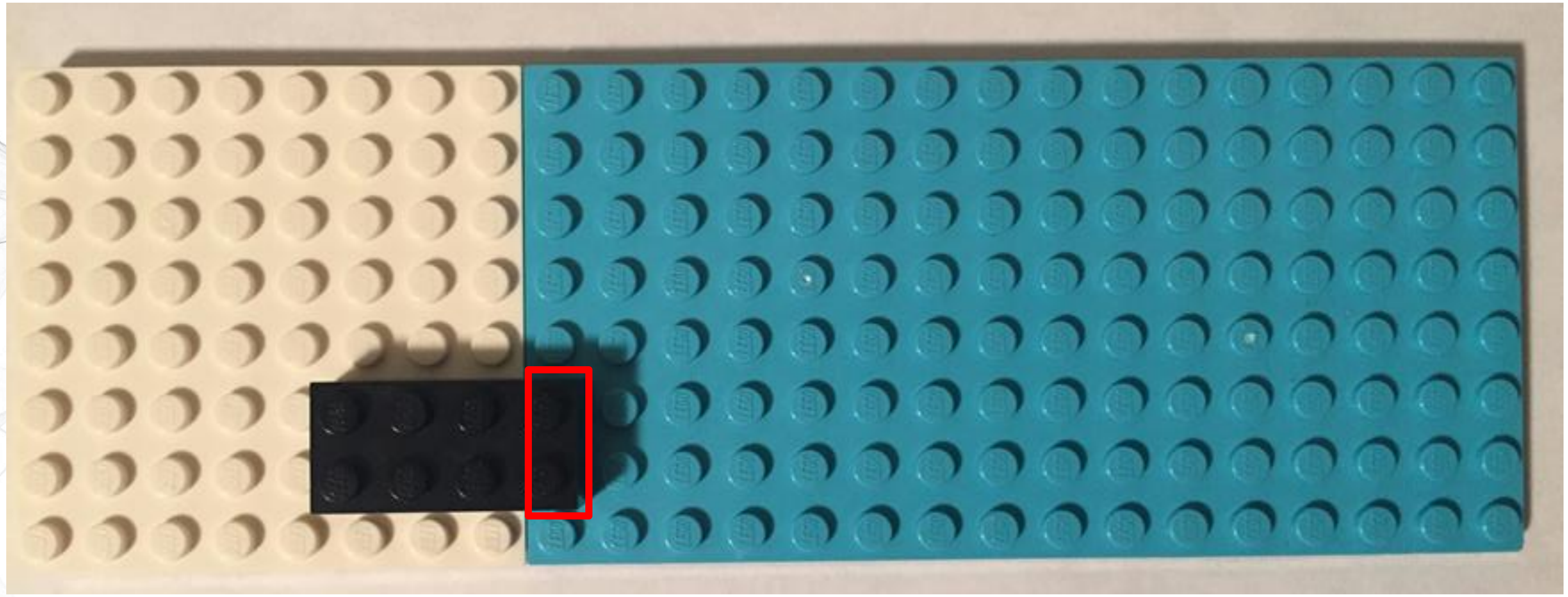


Now we have observed black: landed on a black unit

Regardless of what underlying theory is, the chance of this observation is:

$$P(\text{black}) = 8/192 = 0.042$$

Bayes' theorem with Legos

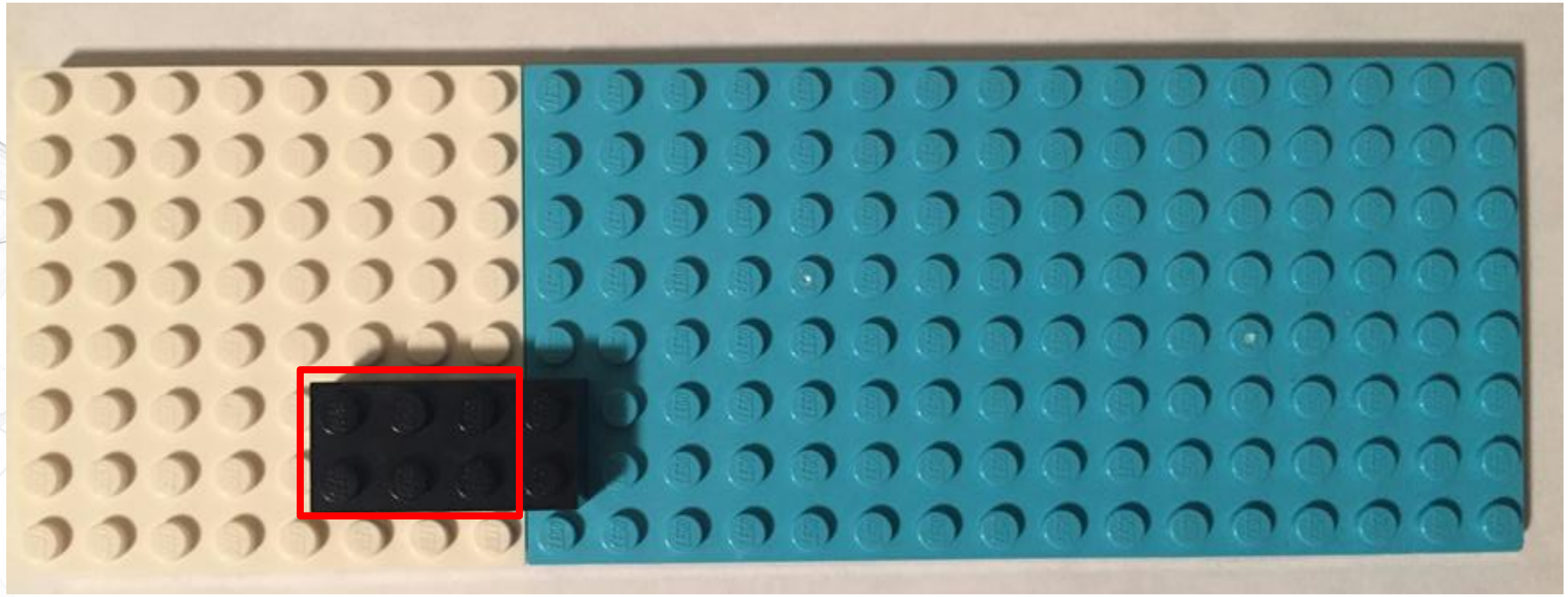


If we are in blue territories, the chance of observing black is:

$$P(\text{black}|\text{blue}) = 2/128 = 0.016$$

This is kind of unlikely?

Bayes' theorem with Legos

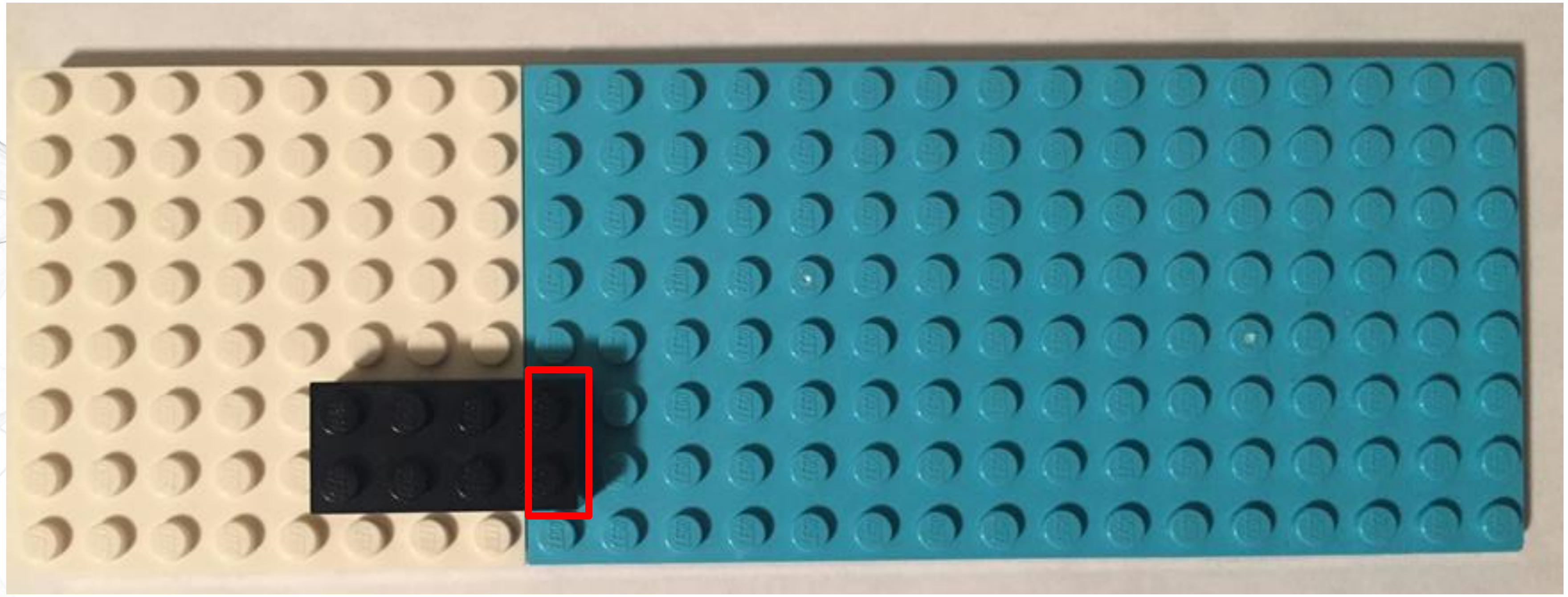


If we are in white territories, the chance of observing black is:

$$P(\text{black}|\text{white}) = 6/64 = 0.094$$

Maybe more likely?

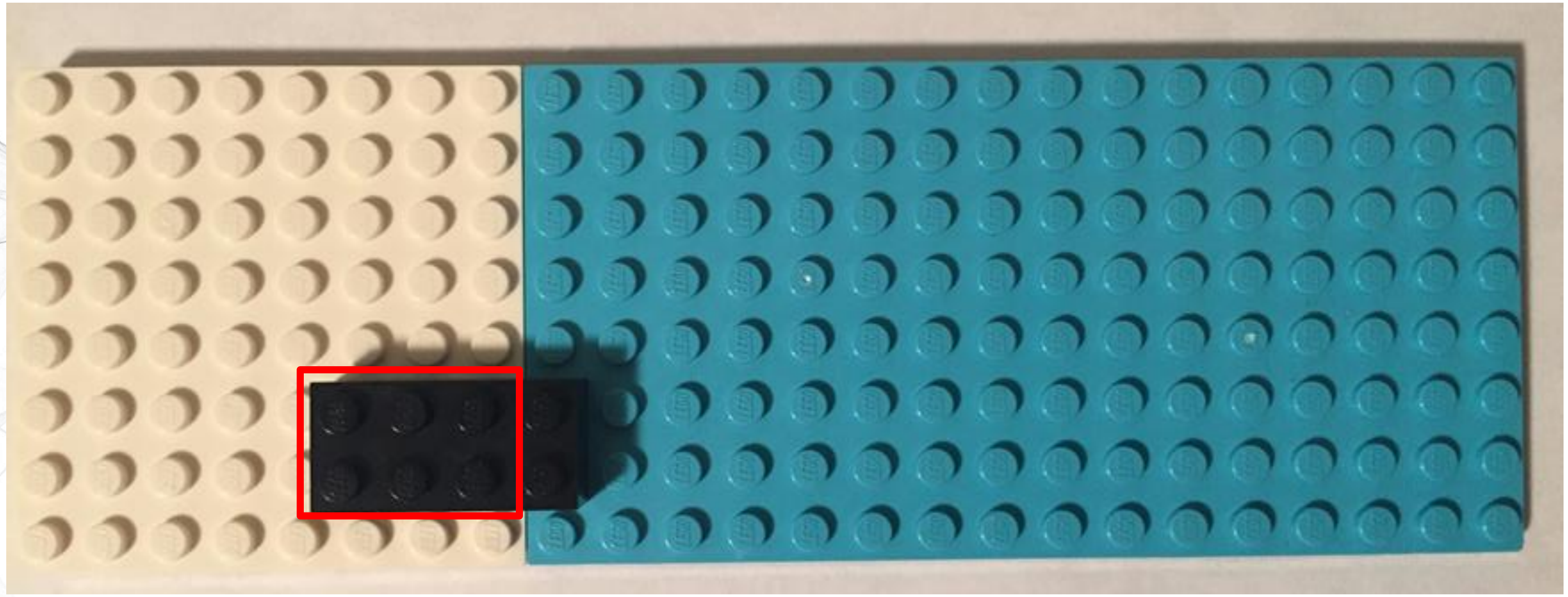
Bayes' theorem with Legos



Given the observation of black, how likely is theory blue true?

$$P(\text{blue}|\text{black}) = \frac{P(\text{black}|\text{blue})P(\text{blue})}{P(\text{black})} = \frac{0.016 \; * 0.67}{0.042} = 0.25$$

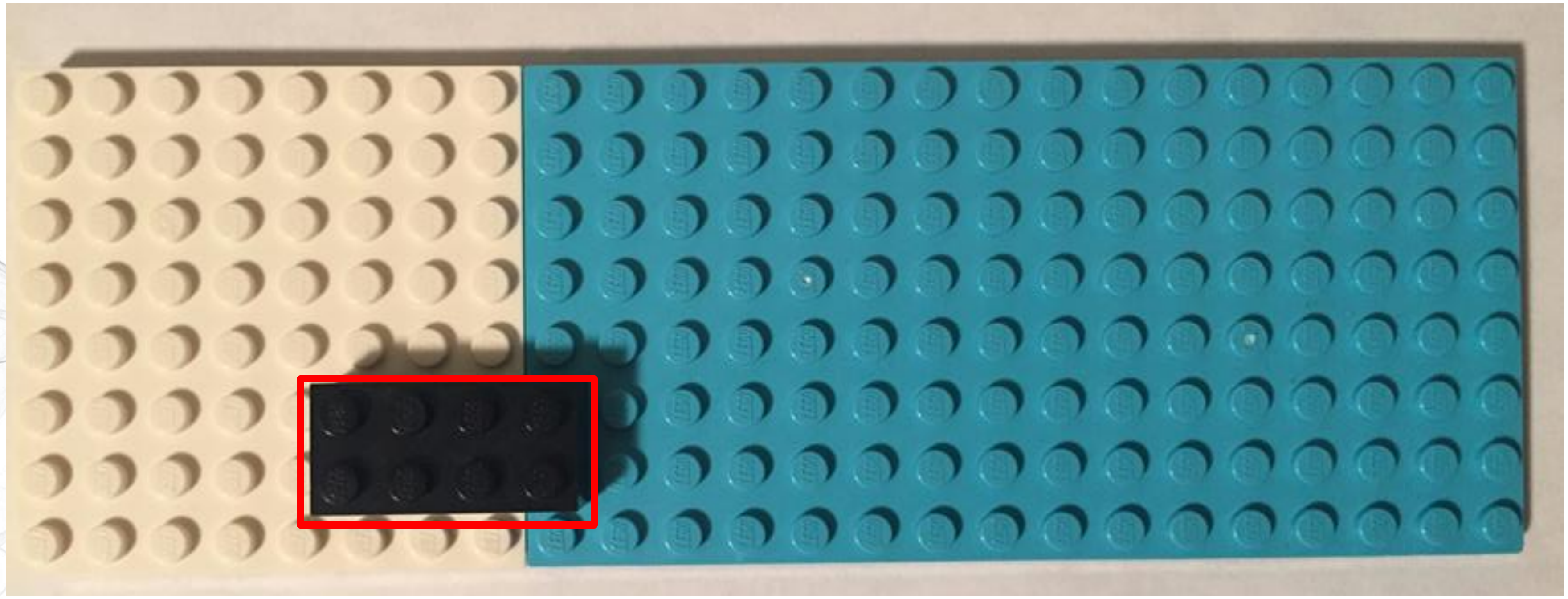
Bayes' theorem with Legos



Given the observation of black, how likely is theory white true?

$$P(\text{white}|\text{black}) = \frac{P(\text{black}|\text{white})P(\text{white})}{P(\text{black})} = \frac{0.094 \ * 0.33}{0.042} = 0.75$$

Bayes' theorem with Legos



Key observations:

- Probability of theory white increased from 1/3 (**prior**) to 3/4 (**posterior**)
  - We narrowed down our field of view from the entire space to the observed space
- This is consistent with the observation that 6 out of 8 units of the black lego have a white background.

In this lecture

- Basics of Bayes theorem
- **Case study: lung cancer and pulmonary nodules**
- Bayesian versus frequentist

Case study: pulmonary nodule

- Age-standardized incidence rate of lung cancer in China: 36.71 per 100,000
- Overall prevalence of pulmonary nodules: 0.27
- Sensitivity of detecting lung lesions in actual cases of lung cancer: 94.4–96.4%

Cao, Maomao, and Wanqing Chen. "Epidemiology of lung cancer in China." *Thoracic cancer* 10.1 (2019): 3-7.
Chen, Dan, et al. "Prevalence and management of pulmonary nodules: a systematic review and meta-analysis." *Journal of thoracic disease* 16.7 (2024): 4619.
Rubin, Geoffrey D. "Lung nodule and cancer detection in computed tomography screening." *Journal of thoracic imaging* 30.2 (2015): 130-138.

Case study: pulmonary nodule

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

- B: Identified a pulmonary nodule in health screening
- A: Lung cancer
- $\boldsymbol{P(B) = 0.27}$
- $\boldsymbol{P(A) = 0.000367}$
- $\boldsymbol{P(B|A) = 0.95}$ (95% of lung cancer patients will have positive results in a CT examination for pulmonary nodules)
- $\boldsymbol{P(A|B)} = ?$
  - CT screening identified a pulmonary nodule, how likely is it due to lung cancer?

Cao, Maomao, and Wanqing Chen. "Epidemiology of lung cancer in China." *Thoracic cancer* 10.1 (2019): 3-7.
Chen, Dan, et al. "Prevalence and management of pulmonary nodules: a systematic review and meta-analysis." *Journal of thoracic disease* 16.7 (2024): 4619.
Rubin, Geoffrey D. "Lung nodule and cancer detection in computed tomography screening." *Journal of thoracic imaging* 30.2 (2015): 130-138.

Case study: pulmonary nodule

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

- B: Identified a pulmonary nodule in health screening
- A: Lung cancer
- $\boldsymbol{P(B) = 0.27}$
- $\boldsymbol{P(A) = 0.000367}$
- $\boldsymbol{P(B|A) = 0.95}$ (95% of lung cancer patients will have positive results in a CT examination for pulmonary nodules)
- $\boldsymbol{P(A|B) = \frac{0.95 * 0.000367}{0.27} = 0.0013}$
  - CT screening identified a pulmonary nodule, how likely is it due to lung cancer: 0.13%

Cao, Maomao, and Wanqing Chen. "Epidemiology of lung cancer in China." *Thoracic cancer* 10.1 (2019): 3-7.
Chen, Dan, et al. "Prevalence and management of pulmonary nodules: a systematic review and meta-analysis." *Journal of thoracic disease* 16.7 (2024): 4619.
Rubin, Geoffrey D. "Lung nodule and cancer detection in computed tomography screening." *Journal of thoracic imaging* 30.2 (2015): 130-138.

Case study: pulmonary nodule

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

- B: Identified a pulmonary nodule in health screening
- A: Lung cancer
- Potential bias: $\boldsymbol{P(A)}$ is derived from the entire population, $\boldsymbol{P(B)}$ is derived from a subset of the population that did health screening (will you do health screening before 20-year-old?).
- If we condition everything on the subset of population that do regular health screening, $\boldsymbol{P(A)}$ is likely higher. In a report from Shanghai, $\boldsymbol{P(A)\sim 0.4}\%$, $\boldsymbol{P(A|B)\sim 1}\%$
- Many more factors:
  - Age
  - Smoking history
  - Size, type, and number of pulmonary nodules

赵俊松, et al. "上海 22351 例无症状体检者低剂量 CT 肺癌筛查及随访结果初步分析." *诊断学理论与实践* 18.2 (2019): 183-188.

Case study: pulmonary nodule

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

- B: Identified a pulmonary nodule in health screening
- A: Lung cancer
- “93 million CT examinations performed in 62 million patients in 2023 were projected to result in approximately 103 000 future cancers”
- This is one cancer every ~900 CT scans, how can we balance the benefits and risks?

Smith-Bindman, Rebecca, et al. "Projected lifetime cancer risks from current computed tomography imaging." *JAMA internal medicine* 185.6 (2025): 710-719.

Visualize the example with Monte Carlo Simulation

```
> prevalence <- 100 / 100000  # 100 cases per 100k population
> N <- 1e6  # population size
> outcome <- sample(c('Disease', 'Healthy'), N, replace=TRUE,
+                   prob=c(prevalence, 1 - prevalence))
> sum(outcome == 'Disease')
[1] 1018
> sum(outcome == 'Healthy')
[1] 998982
```

- We start by randomly selecting 1M people from a population in which the disease in question has a 100 in 100,000 prevalence: common cancer has about this level of prevalence, note that the number is per year.

- Very few people have the disease.

Visualize the example with Monte Carlo Simulation

```
> test_acc <- 0.99
> test <- vector("character", N)
> N_D <- sum(outcome == 'Disease')
> test[outcome == 'Disease'] <- sample(c('+', '-'), N_D, replace=TRUE,
+                                      prob=c(test_acc, 1 - test_acc))
> N_H <- sum(outcome == 'Healthy')
> test[outcome == 'Healthy'] <- sample(c('-', '+'), N_H, replace=TRUE,
+                                      prob=c(test_acc, 1 - test_acc))
> table(outcome, test)
         test
outcome        -      +
  Disease     13   1005
  Healthy 989024   9958
> 1005 / (1005 + 9958)
[1] 0.09167199
```

With a highly accurate test (correct 99% of the time), there is still only a 10% chance that one with a positive test result actually has the disease:

- Largely caused by the amount of false positives

Visualize the example with Monte Carlo Simulation

```
> test_acc <- 0.95
> test <- vector("character", N)
> N_D <- sum(outcome == 'Disease')
> test[outcome == 'Disease'] <- sample(c('+', '-'), N_D, replace=TRUE,
+                                      prob=c(test_acc, 1 - test_acc))
> N_H <- sum(outcome == 'Healthy')
> test[outcome == 'Healthy'] <- sample(c('-', '+'), N_H, replace=TRUE,
+                                      prob=c(test_acc, 1 - test_acc))
> table(outcome, test)
         test
outcome        -      +
  Disease     57    961
  Healthy 949232  49750
> 961 / (49750 + 961)
[1] 0.01895052
```

With a slightly less accurate test (correct 95% of the time), this number drops to 1~2%:

- Because of even more false positives

In this lecture

- Basics of Bayes theorem
- Case study: lung cancer and pulmonary nodules
- **Bayesian versus frequentist**

Bayesian versus Frequentist

- You have a coin that is biased:
    - When flipped, heads appear more often than tails. Now you have flipped the coin 2 times. It ends up head 2 times.

- What is the underlying probability p of getting a head in the coin toss?

Bayesian versus Frequentist

- We can just use observations to estimate p:

$$\hat{p} = \frac{2}{2} = 1$$

- As the number of observations increases, $\hat{p}$ will converge to the ground-truth p (the Law of Large Number)

- But do you believe that there is an underlying ground-truth p?

Bayesian versus Frequentist

- Alternatively, we know that given p, the probability of observing two heads is:

$$P(2\ \mathrm{heads}|p) = p^2$$

- Let's say we do not have any presumption on p, so the prior is uniform:

$$P(p) = constant$$

- Let's adapt the Bayes theorem:

$$P(p|2\ \mathrm{heads}) = \frac{P(2\ \mathrm{heads}|p)P(p)}{P(2\ \mathrm{heads})} \propto p^2$$

Bayesian versus Frequentist

- Let's adapt the Bayes theorem:

$$P(p|2\text{ heads}) = \frac{P(2\text{ heads}|p)P(p)}{P(2\text{ heads})} \propto p^2$$

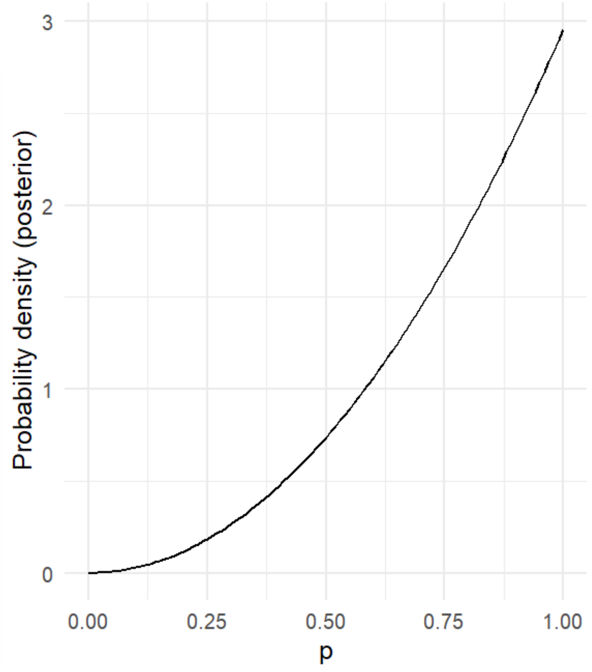
- We know that $p \in [0, 1]$:

```r
df <- data.frame(p=seq(0, 1, 0.01))
df <- df |> mutate(posterior=p**2)

step_size <- 0.01

# Approximate the area under the curve
total_area <- sum(df$posterior * step_size)
df <- df |> mutate(posterior = posterior / total_area)  # Normalize

df |> ggplot(aes(p, posterior)) +
  geom_line() +
  ylab("Probability density (posterior)") +
  theme_minimal()
```



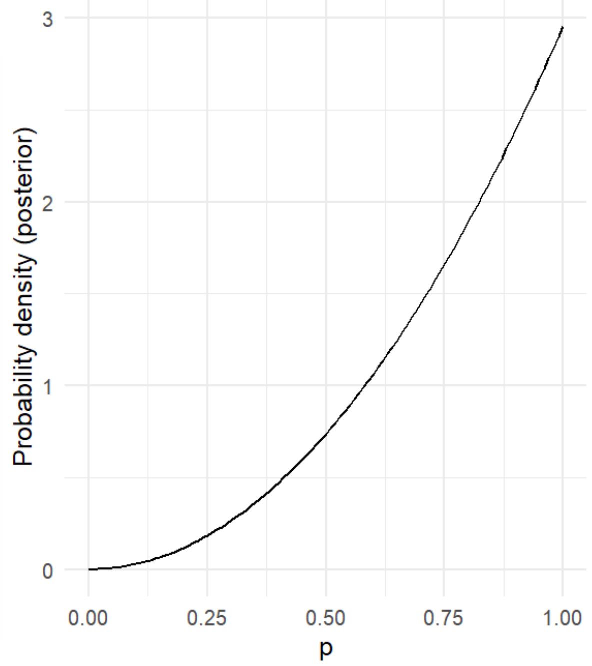
Bayesian versus Frequentist

```r
df <- data.frame(p=seq(0, 1, 0.01))
df <- df |> mutate(posterior=p**2)

step_size <- 0.01

# Approximate the area under the curve
total_area <- sum(df$posterior * step_size)
df <- df |> mutate(posterior = posterior / total_area)  # Normalize

df |> ggplot(aes(p, posterior)) +
  geom_line() +
  ylab("Probability density (posterior)") +
  theme_minimal()
```



- The maximum a posteriori (MAP) estimate is aligned with the frequentist view
- However, the expectation is 0.75

Are you a Bayesian or a Frequentist?

- You have a coin that is biased:
  - when flipped, heads appear more often than tails. Now you have flipped the coin 14 times. It ends up head 10 times.

- Question: *If you get two heads in a row in the next two tosses, you win. Will you bet on it?*

Are you a Bayesian or a Frequentist?

- Frequentist:

$$\hat{p} = \frac{10}{14} = 0.714$$

- The chance of two heads in a row will be:

$$P(2\text{ heads}) = \hat{p}^2 = 0.51 > 0.5$$

- You should bet.

Are you a Bayesian or a Frequentist?

- Bayesian:

$$P(10 \text{ heads out of } 14 \text{ tosses}|p) = \binom{14}{10} p^{10}(1-p)^4$$

- Let's still use uniform prior:

$$P(p) = constant$$

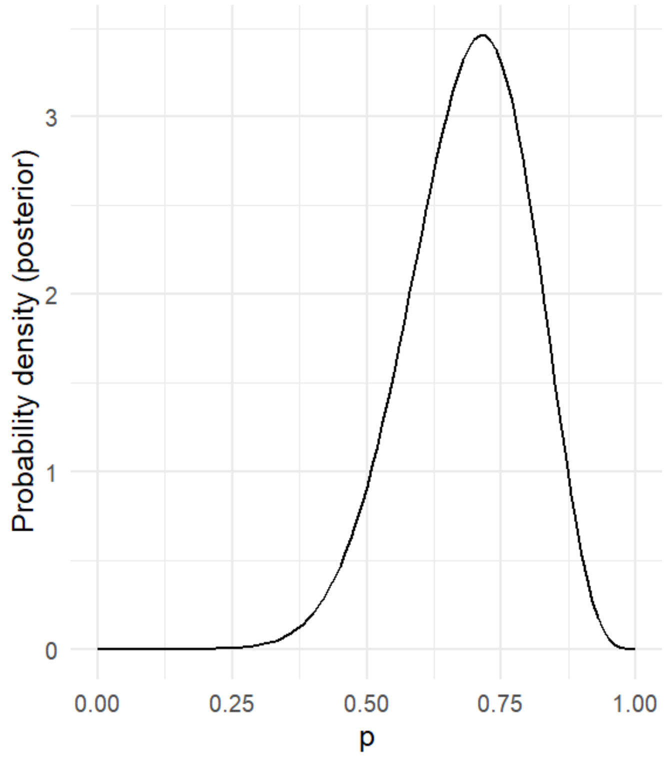
Are you a Bayesian or a Frequentist?

```r
cons = length(combinations(14, 10))
df <- data.frame(p=seq(0, 1, 0.01))
df <- df |> mutate(posterior=cons * ((1 - p)**4) * (p**10))

step_size <- 0.01

# Approximate the area under the curve
total_area <- sum(df$posterior * step_size)
df <- df |> mutate(posterior = posterior / total_area)  # Normalize

df |> ggplot(aes(p, posterior)) +
  geom_line() +
  ylab("Probability density (posterior)") +
  theme_minimal()
```



Are you a Bayesian or a Frequentist?

```
> df |> arrange(desc(posterior)) |> head(3)
     p posterior
1 0.71  3.457011
2 0.72  3.455273
3 0.70  3.435506
```

- Again, MAP is aligned with frequentist result (this is often true if no specific prior is used)

```
> sum(df$posterior * (df$p**2)) / length(df$posterior)
[1] 0.4804892
```

- If we use expectation over the full distribution:

$$P(2\ \text{heads}|\text{p}) = 0.48 < 0.5$$

- You should not bet.

This example is adapted from: https://www.behind-the-enemy-lines.com/2008/01/are-you-bayesian-or-frequentist-or.html

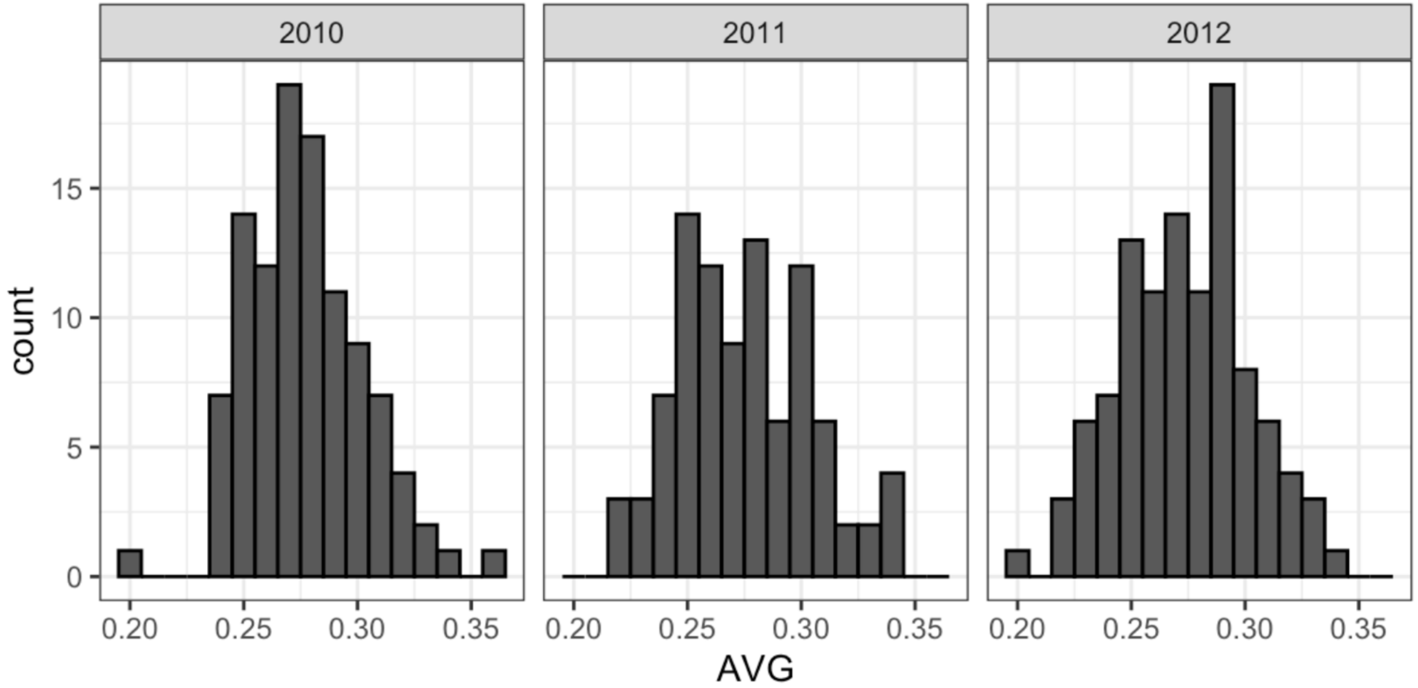
When prior matters

- In professional baseball, success rate of batting measures how good a player is. An AVG of 0.45 is extremely high.
- From a frequentist point of view:
    - Estimated average batting success rate: $9/20 \ = \ 0.45$
    - Estimated SD is $\sqrt{\frac{(1 - 0.45) * 0.45}{20}} = 0.111$
    - 95% CI is: $[0.228, 0.672]$

Month	At Bats	H	AVG
April	20	9	.450

When prior matters

- Estimated average batting success rate: $9/20 = 0.45$
- Estimated SD is $\sqrt{\frac{(1-0.45)*0.45}{20}} = 0.111$
- 95% CI is: $[0.228, 0.672]$



When prior matters

- From a Bayesian point of view:
  - We have a prior on batting AVG based on historical data
  - Given the player's outstanding performance, we will update our prior and derive a posterior for this player's batting AVG:

$$\mathrm{E}(\mathrm{AVG}|9\text{ out of }20) = 0.285$$
$$\mathrm{SE}(\mathrm{AVG}|9\text{ out of }20) = 0.026$$

- Note how different this is from the CI of $[0.228, 0.672]$
  - The frequentist view has a larger uncertainty, because our approach does not consider the vast amounts of observations from the past

Are you a Bayesian or a Frequentist?

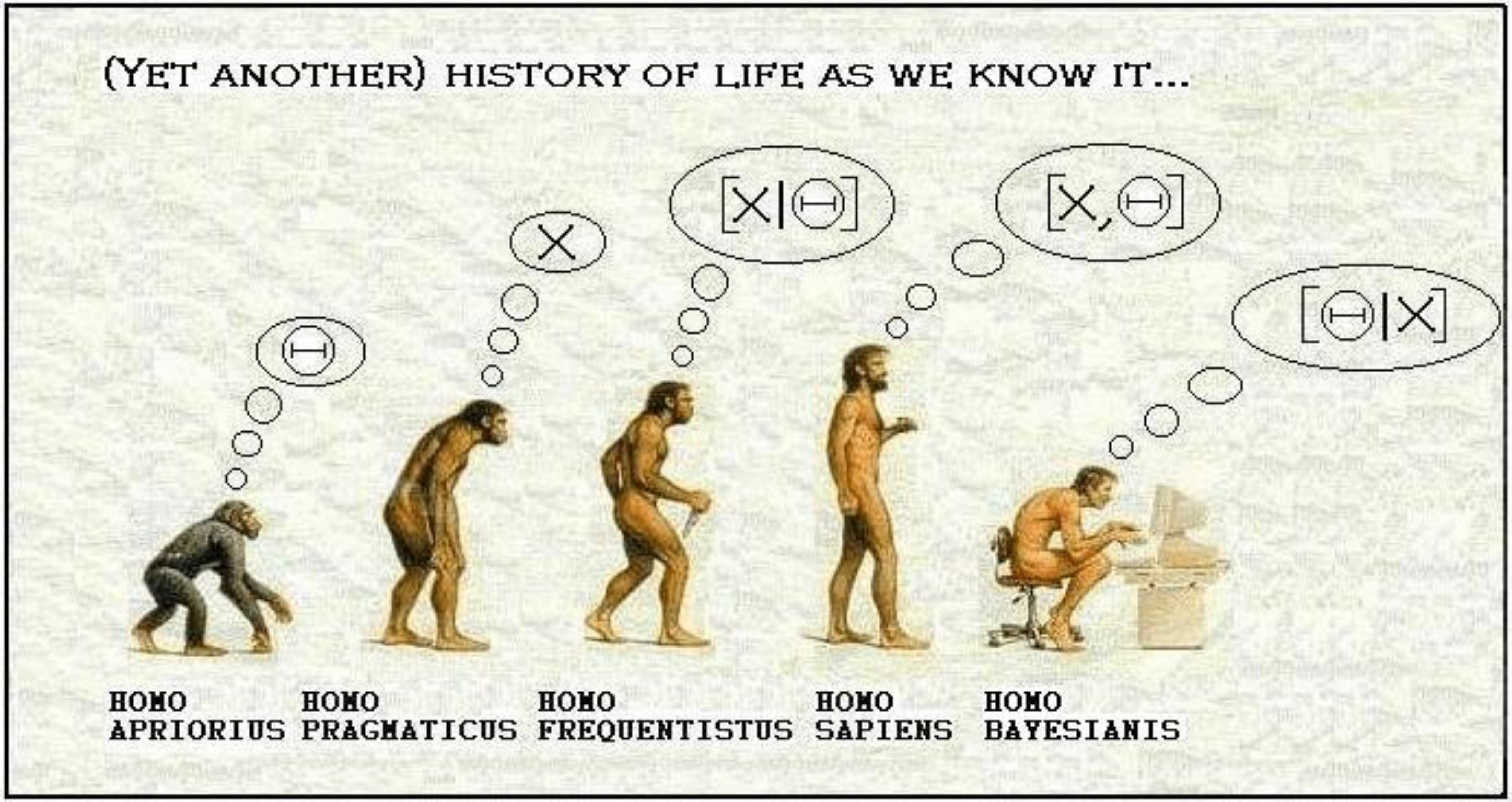
- Frequentist view:
  - Yes, there is always an underlying ground truth probability distribution
  - Observations were randomly sampled from the distribution, and we use them to approach the underlying probability distribution
- Bayesian view:
  - No, there is no fixed "ground truth" probability; they should be treated as random variables
  - Observations were real. We update our belief based on observations.

Are you a Bayesian or a Frequentist?

- Terms commonly used in Frequentist view:
  - Sampling distribution
  - Hypothesis testing: null hypothesis, alternative, **p value**
  - Confidence interval
  - Power
- Terms commonly used in Bayesian view:
  - Prior
  - Posterior
  - Likelihood
  - Credible interval: Bayesian equivalent of a confidence interval

Bayesian statistics in practice

- Sally Clark case and Meadow's law: https://en.wikipedia.org/wiki/Sally_Clark
- Widely used in machine learning and AI: Bayesian Neural Networks, Bayesian Optimization
- Healthcare: diagnosis based on evidence, personalized medicine
- Finance: risk modeling for financial product
- Spam email detection
- Voice recognition



I have a theory on how the universe work

Reality is the only truth

The likelihood of reality given theory

Based on reality, how likely is our theory right?

https://phylonetworks.blogspot.com/2014/08/the-evolution-of-statistical.html

In this lecture

- Basics of Bayes theorem
- Case study: lung cancer and pulmonary nodules
- Bayesian versus frequentist