GENERAL ASSEMBLY

BAYESIAN INFERENCE

Matt Brems

DSI+

LEARNING OBJECTIVES

- Understand how Bayes' Theorem connects to Bayesian inference.
- Describe how the prior and likelihood influence the posterior.
- Describe the posterior distribution.

RECALL BAYES' THEOREM

Bayes' Rule

"probability of A given B"

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

- $P(A)$ is the probability that A occurs given no supplemental information.
- $P(B|A)$ is the likelihood of seeing evidence (data) B assuming that A is true.
- $P(B)$ is what we scale $P(B|A)P(A)$ by to ensure we are only looking at A within the context of B occurring.

BAYES' THEOREM

$P(jar\ 1 \mid V)$ $P(jar\ 2 \mid V)$

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypotheis)}{P(data)}$$

- I have two jars: Jar 1 and Jar 2.

- Jar 1 contains 20 chocolate and 20 vanilla cookies.
- Jar 2 contains 10 chocolate and 30 vanilla cookies.

- I pull a vanilla cookie out of a jar. My goal is to figure out which jar I pulled the cookie from.

BAYES' THEOREM

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypotheis)}{P(data)}$$

↳ any number b/w 0 and 1.

- I have a coin with some probability of "flipping heads." Call this $prob$.

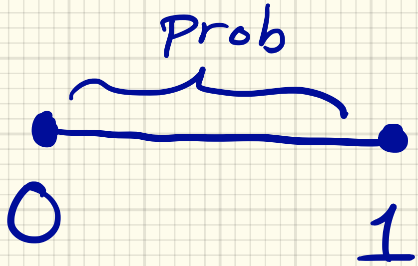
- I flip a coin once and flip heads.

↳ data: flipped heads

- My goal is to understand what $prob$ is.

Cookie hypotheses: {jar 1, jar 2}

Coin hypotheses: $[0, 1]$



WHAT IF WE LOOK AT ALL POSSIBLE HYPOTHESES?

Prob is continuous

time consuming!

$$P(prob = 0|data) = \frac{P(data|prob = 0)P(prob = 0)}{P(data)}$$

$$P(prob = 0.001|data) = \frac{P(data|prob = 0.001)P(prob = 0.001)}{P(data)}$$

$$\vdots$$

$$P(prob = 1|data) = \frac{P(data|prob = 1)P(prob = 1)}{P(data)}$$

dist'n

dist'n

what are all values of prob & their frequencies?

WHAT IF WE LOOK AT ALL POSSIBLE HYPOTHESES?

- Instead of manually writing out every possible hypothesis (time-consuming, impossible every time we want to learn about a continuous parameter), what if we combined each of these individual probabilities into one distribution?

$$P(prob = 0|data) \frac{P(data|prob = 0)P(prob = 0)}{P(data)}$$

$$\Downarrow$$

It is common to use "f" here, even though our distributions may be different.

$$f(prob|data) = \frac{f(data|prob)f(prob)}{f(data)}$$

→ f(prob) is the pdf of prob

BAYES' THEOREM: PARAMETERS

θ = theta

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Rightarrow f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

- $f(\theta)$ is the distribution of θ given no supplemental information.
  - "Prior Distribution of θ"
- $f(y|\theta)$ is the likelihood function relating y and θ.
  - "Likelihood"

given every value of prob, how likely is it that we observe our data?

- $f(y)$ is the normalizing constant to ensure $f(\theta|y)$ is a valid probability distribution.
  - "Marginal Likelihood of y"

Parameters: attributes of a population

μ: the average amount of student debt in a population

σ: the standard deviation of commute time of a population

β_1: the effect of a one-unit change in X_1 on Y.

Reference!

BAYESIAN INFERENCE IN TWO BULLET POINTS

- In Bayesian inference, parameters are not fixed numbers. They have **distributions**!
    - A parameter's distribution takes on the set of all hypotheses, and how frequently we observe each hypothesis!

- Our **goal** is **ALWAYS** to find the **posterior distribution** of our parameter.
    - That is, what is the set of all hypotheses and how frequently we observe each hypothesis after taking our data into account?

FREQUENTIST VS. BAYESIAN INFERENCE OF PARAMETERS

- Frequentist inference and Bayesian inference have different interpretations of probability, and these interpretations give rise to different methods of analysis.

	Definition of Probability	Interpretation of Probability	Parameter
Frequentist	$P(A) = \lim_{n \to \infty} \frac{\#\ of\ times\ A\ occurs}{n}$	long-run behavior	fixed number
Bayesian	$P(A) =$ how likely we believe A occurs	degrees of belief	distribution

FREQUENTIST VS. BAYESIAN INFERENCE OF PARAMETERS

- I want to study student loan debt among college graduates.
- Parameter: The average student loan debt among graduates, denoted μ.
- Goal: Learn about μ.

$[0, \infty)$

- **Frequentists** treat μ as **fixed**: $\mu = \$39{,}000$.
  - Confidence Interval: We are 95% confident that the true population mean is between \$38,000 and \$40,000.

- **Bayesians** treat μ as following a **distribution**: $\mu \sim N(\$39000, \$500)$
  - Credible Interval: There is a 95% chance that the true population mean is between \$38,000 and \$40,000.

"PROPORTIONAL TO"

Goal: understand the posterior dist'n

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\cancel{f(y)}} \propto f(y|\theta)f(\theta)$$

- We often ignore the $f(y)$ component in the denominator and simply say that the posterior $f(\theta|y)$ is **proportional to** $f(y|\theta)f(\theta)$.

- Why?

"PROPORTIONAL TO": AUTOCORRECT EXAMPLE

- I type the word "radom" into my phone. My phone has to decide to leave the word as "radom," change to "radon," or change to "random."
- What are my hypotheses?

  radom, radon, random

- What is my data?

  radom (+ include preceding words)

- (As always...) what is my goal?

  understand posterior $f(\theta \mid \text{'radom'})$

Example adapted from "Bayesian Data Analysis 3": http://www.stat.columbia.edu/~gelman/book/

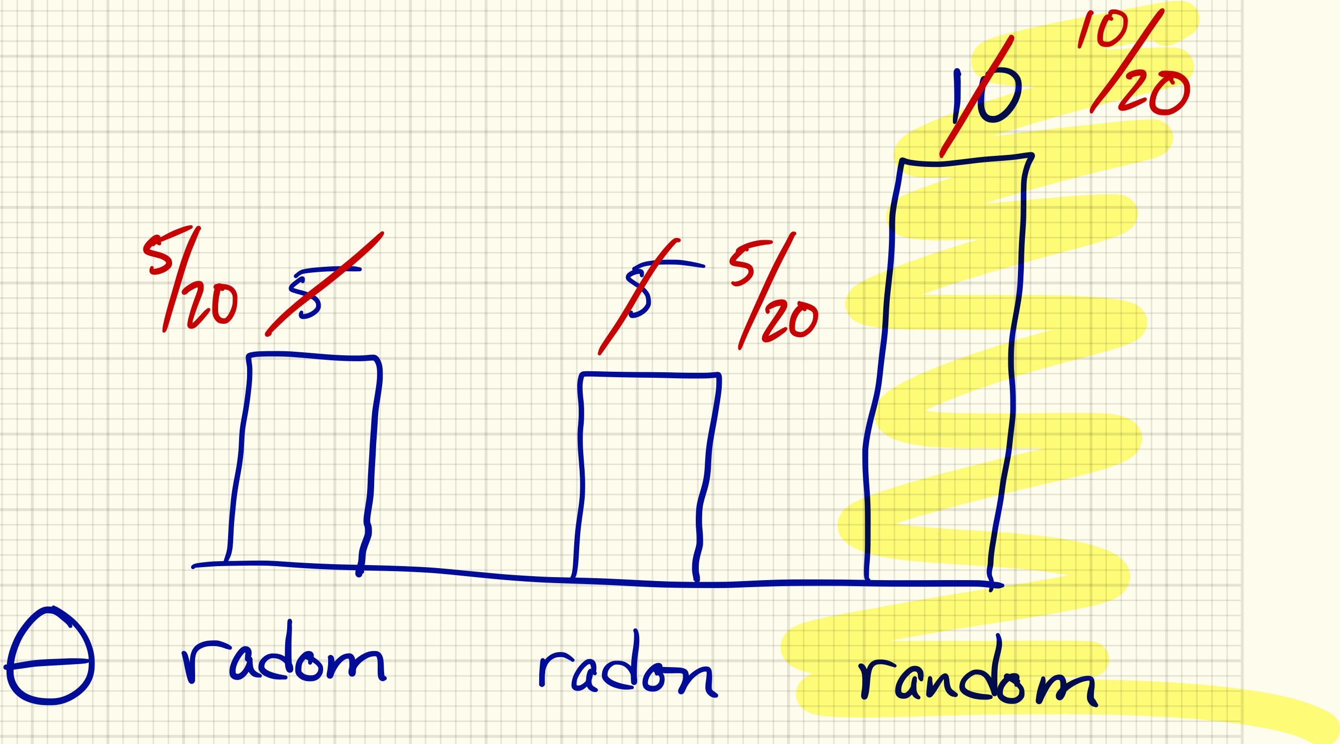
“PROPORTIONAL TO”

- I type the word “radom” into my phone. My phone has to decide to leave the word as “radom,” change to “radon,” or change to “random.”

- If we have three values of θ and we calculate:
  - $P(\theta = radom|y) \propto P(y|\theta = radom)P(\theta = radom) = 5$
  - $P(\theta = radon|y) \propto P(y|\theta = radon)P(\theta = radon) = 5$
  - $P(\theta = random|y) \propto P(y|\theta = random)P(\theta = random) = 10$

…it’s very easy for us to convert $f(\theta|y)$ into a valid probability distribution.

$20 = 5 + 5 + 10$



POSTERIOR DISTRIBUTION

- The posterior distribution $f(\theta|y)$ represents all possible values of θ (our hypotheses!) and how frequently we observe each of these values, given the data we've observed.
  - The posterior distribution is a **complete summary of our parameter of interest $\boldsymbol{\theta}$ that takes into account our data $\boldsymbol{y}$.**

Goal!

POSTERIOR DISTRIBUTION

- The posterior distribution $f(\theta|y)$ represents all possible values of θ (our hypotheses!) and how frequently we observe each of these values, given the data we've observed.
  - The posterior distribution is a **complete summary of our parameter of interest $\boldsymbol{\theta}$ that takes into account our data $\boldsymbol{y}$.**

- In order to construct this posterior distribution $f(\theta|y)$, we need two things:
  - $f(\theta)$, the prior distribution of θ.
  - $f(y|\theta)$, the likelihood of observing the data y under some model.

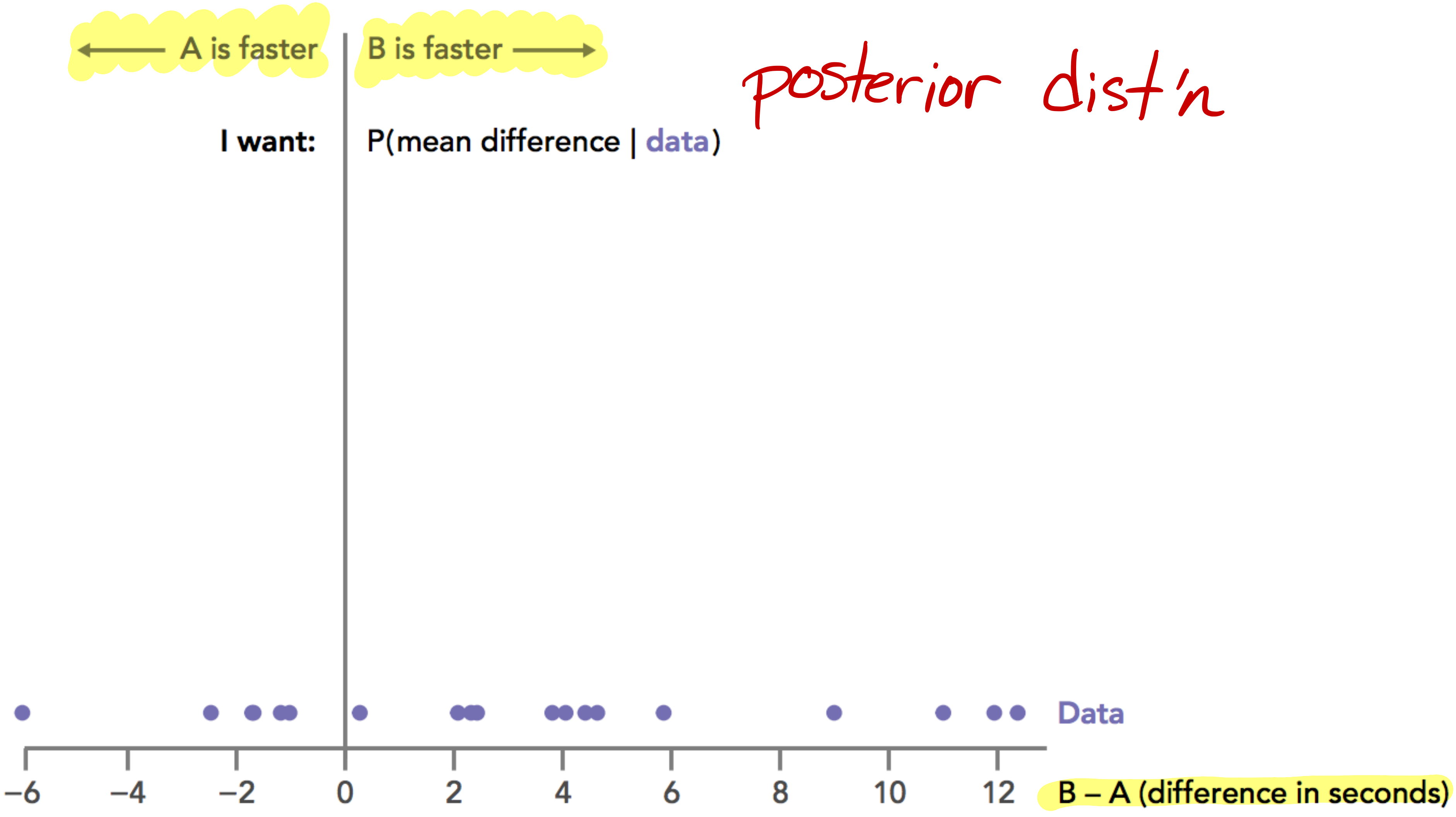
- We can think of our posterior distribution $f(\theta|y)$ as a combination of our data and our prior.

  $ \propto $

  - $f(\theta|y) \propto f(y|\theta) \times f(\theta) = likelihood \times prior$

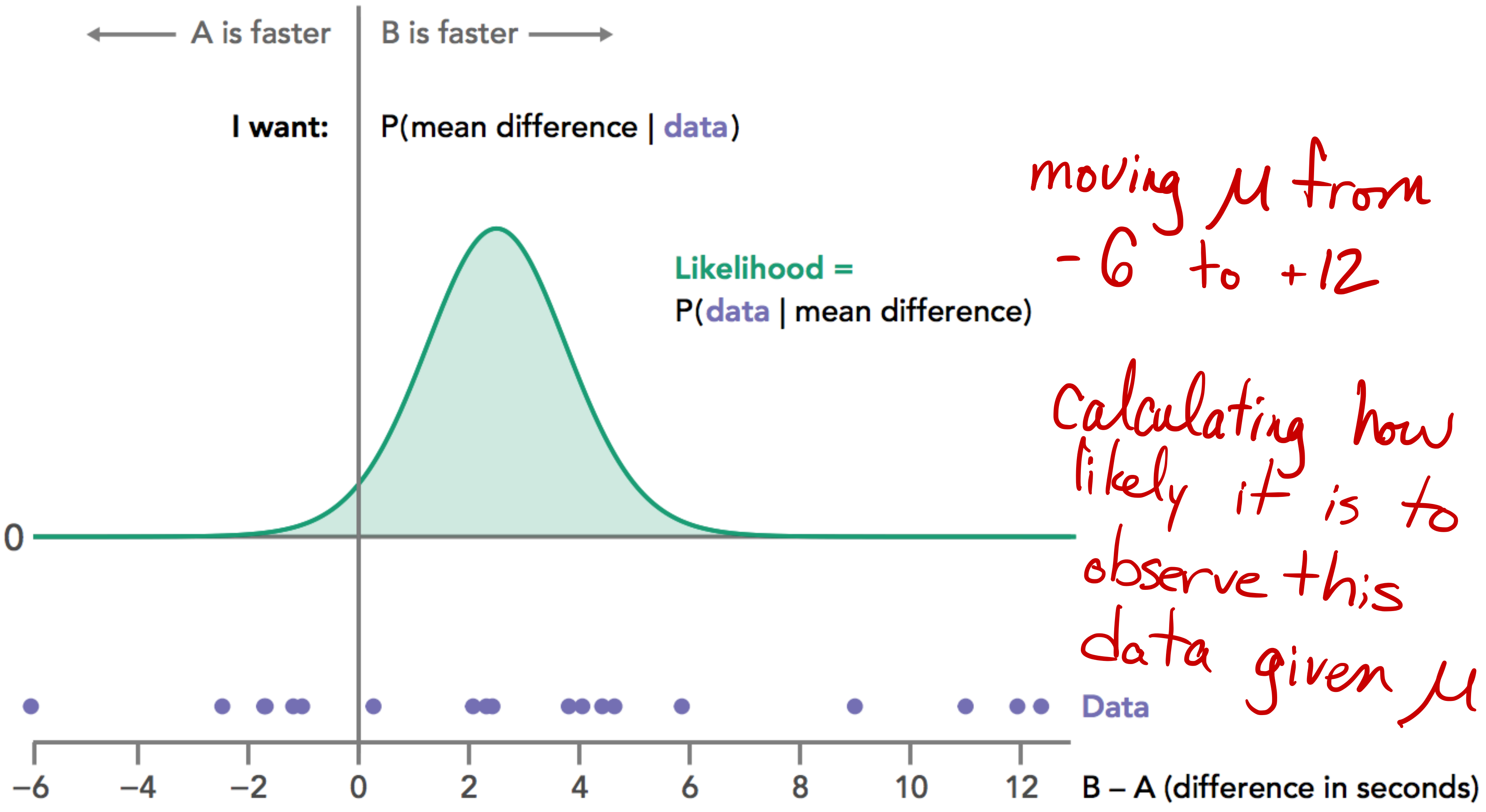
    proportional to

FREQUENTIST VS. BAYESIAN



Source: http://www.mjskay.com/presentations/openvisconf2018-bayes-uncertainty-2.pdf

FREQUENTIST VS. BAYESIAN

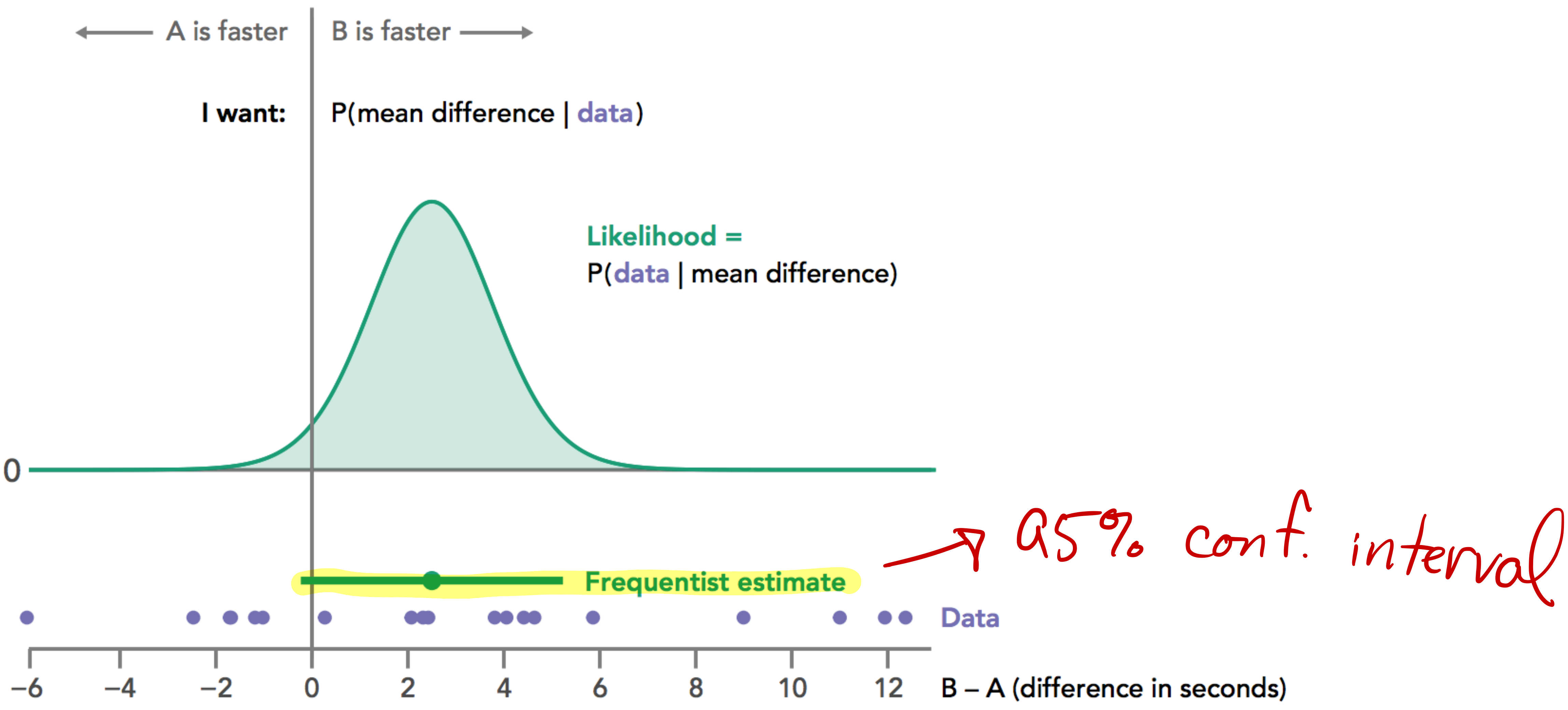


moving μ from -6 to +12

Calculating how likely it is to observe this data given μ.

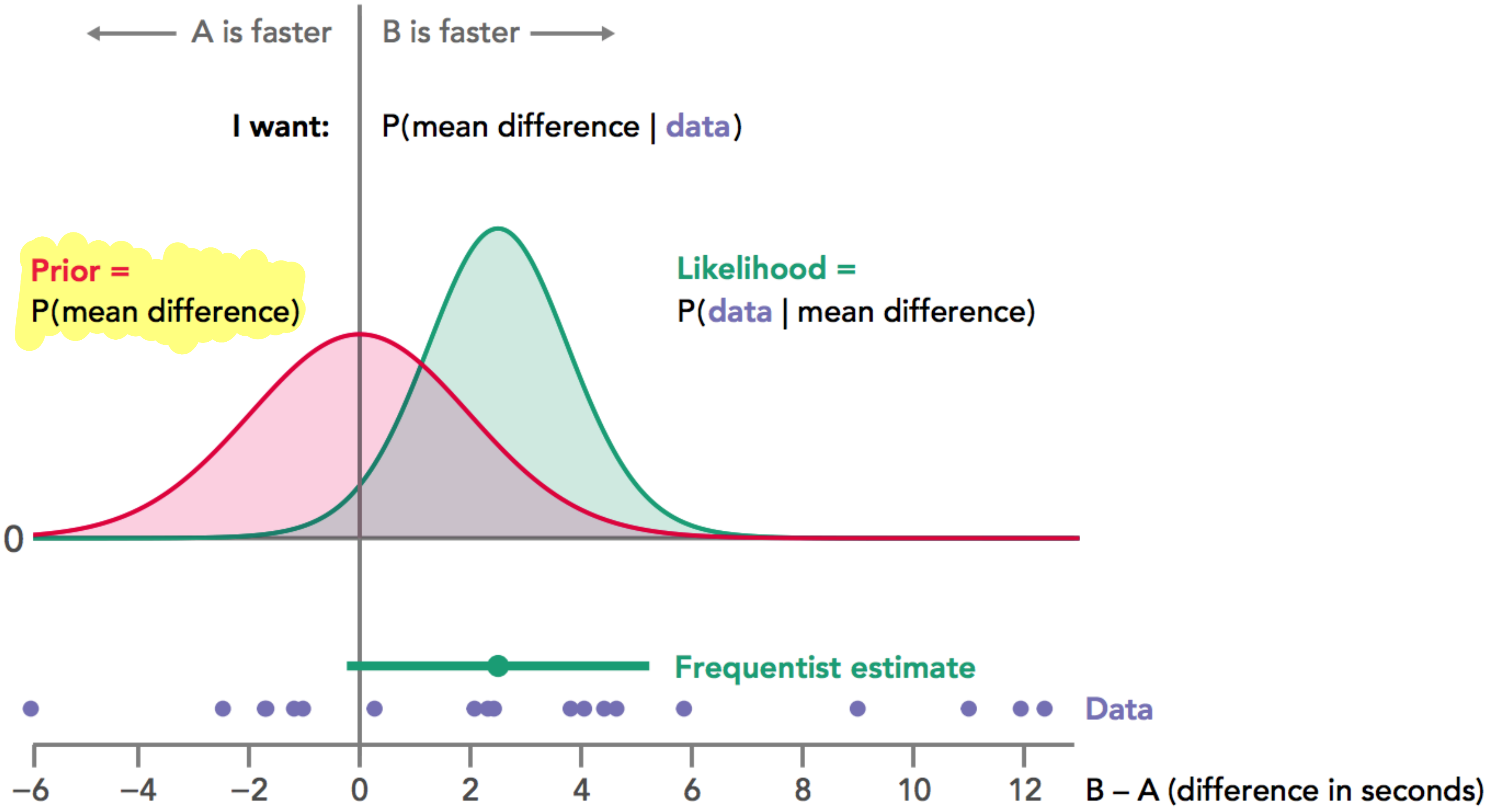
Source: http://www.mjskay.com/presentations/openvisconf2018-bayes-uncertainty-2.pdf

FREQUENTIST VS. BAYESIAN



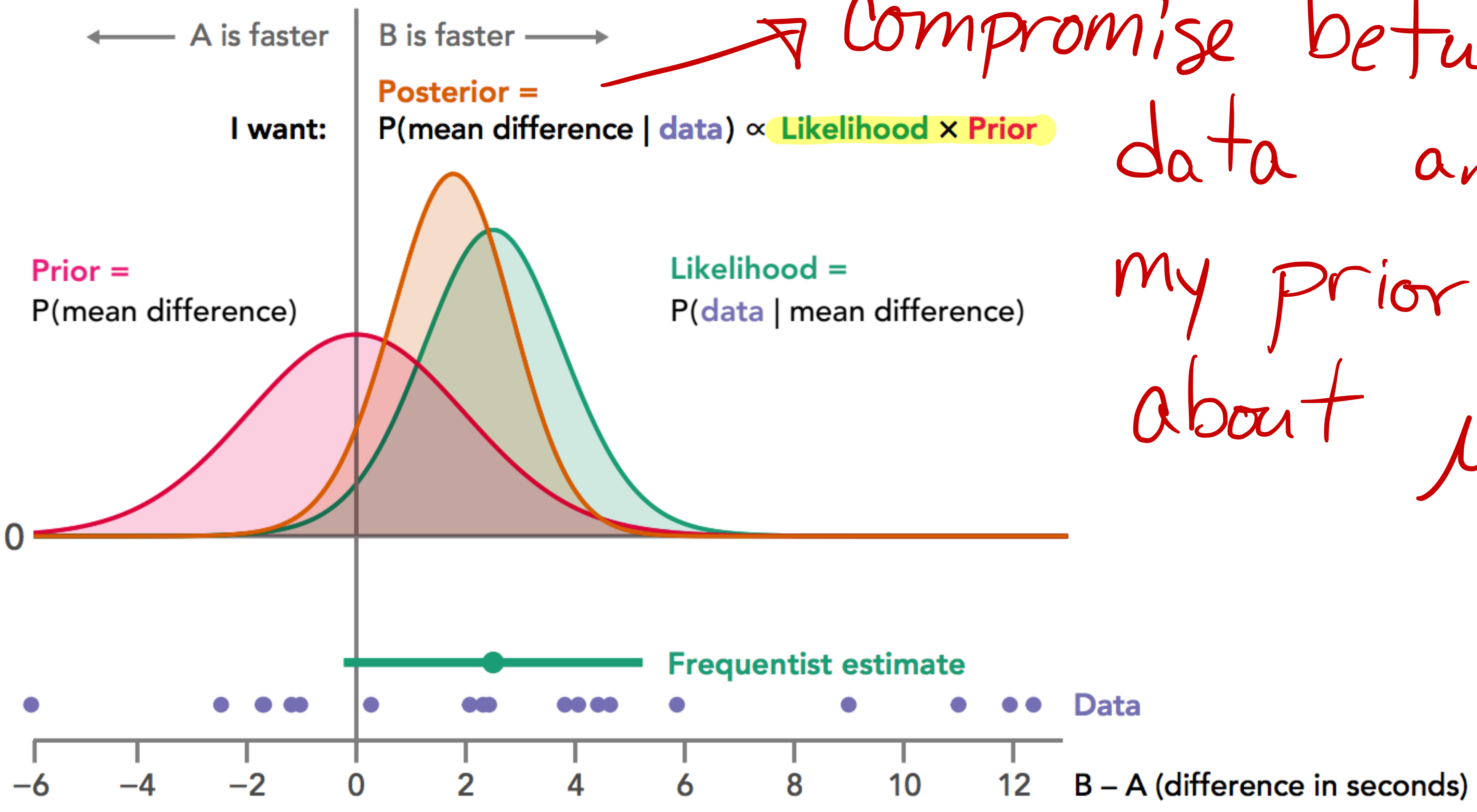
Source: http://www.mjskay.com/presentations/openvisconf2018-bayes-uncertainty-2.pdf

FREQUENTIST VS. BAYESIAN



Source: http://www.mjskay.com/presentations/openvisconf2018-bayes-uncertainty-2.pdf

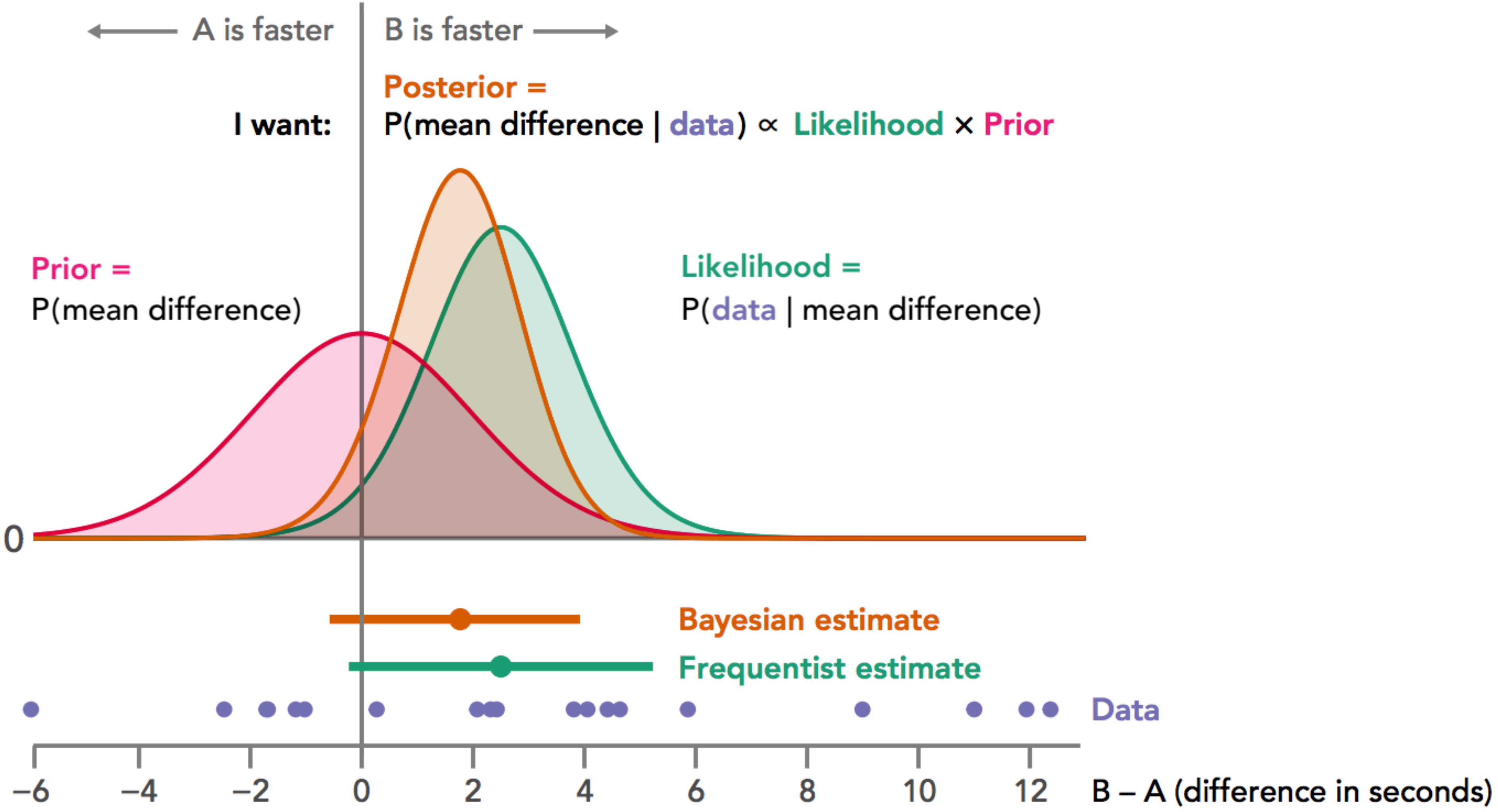
FREQUENTIST VS. BAYESIAN



compromise between my data and my prior beliefs about μ.

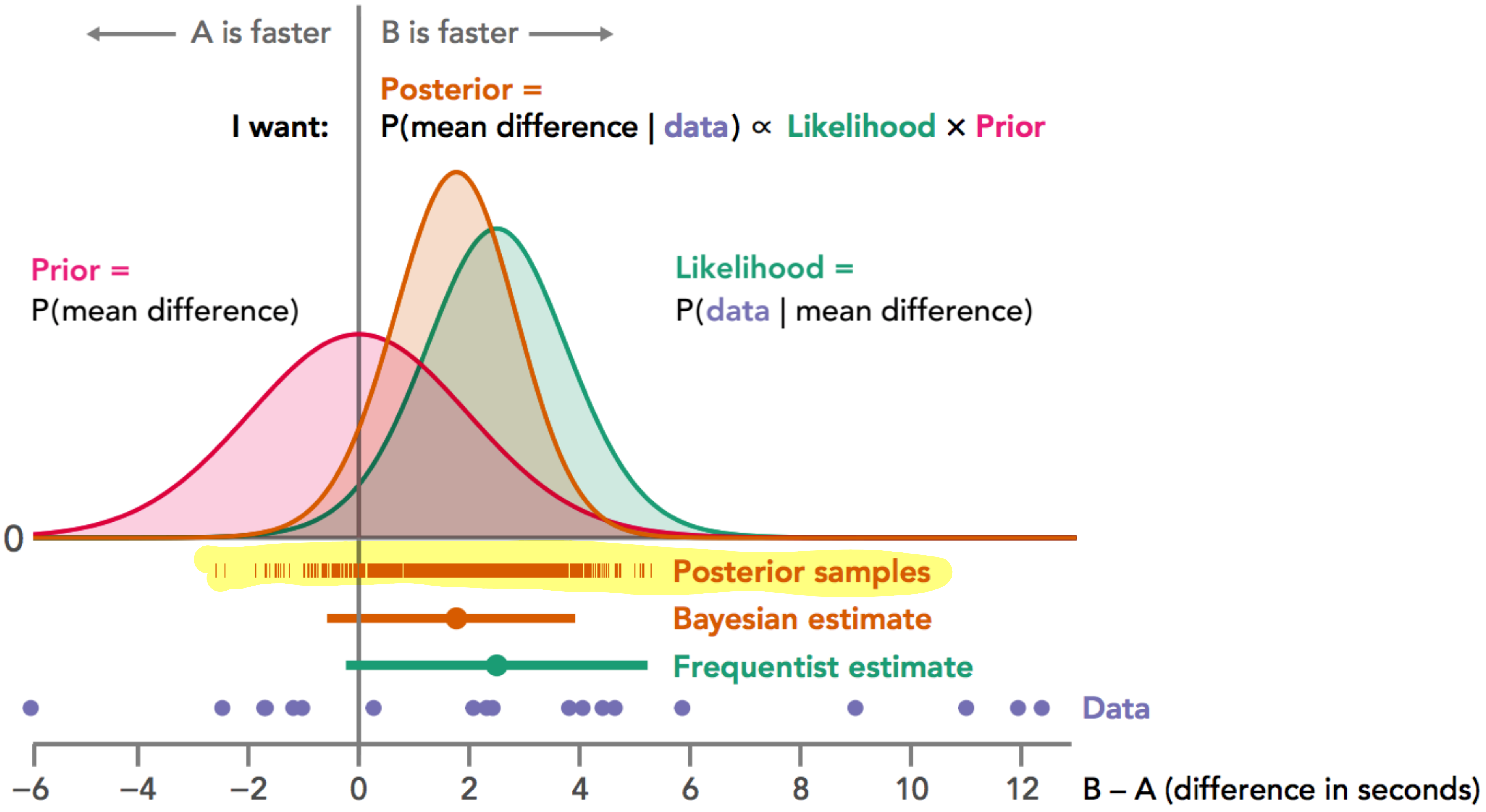
Source: http://www.mjskay.com/presentations/openvisconf2018-bayes-uncertainty-2.pdf

FREQUENTIST VS. BAYESIAN



Source: http://www.mjskay.com/presentations/openvisconf2018-bayes-uncertainty-2.pdf

FREQUENTIST VS. BAYESIAN



Source: http://www.mjskay.com/presentations/openvisconf2018-bayes-uncertainty-2.pdf

REFERENCE: STATISTICAL DISTRIBUTIONS

Distribution	Support	Continuous vs. Discrete	Common Use Case
Normal	$(-\infty, \infty)$	Continuous	θ = everything else
Exponential	$[0, \infty)$	Continuous	θ = time until event
Gamma	$[0, \infty)$	Continuous	θ = time until event
Beta	$[0,1]$	Continuous	θ = prob. of event
Binomial	$\{0, 1, \dots, n\}$	Discrete	θ = number of events
Poisson	$\{0, 1, \dots\}$	Discrete	θ = number of events
Negative Binomial	$\{0, 1, \dots\}$	Discrete	θ = number of events

Consider using this as a reference!

BONUS SECTION

ESTIMATING A PRIOR DISTRIBUTION

PRIOR INFLUENCE ON THE POSTERIOR

- We can think of our posterior distribution $f(\theta|y)$ as a combination of our data and our prior.
  - $f(\theta|y) \propto f(y|\theta) \times f(\theta) = likelihood \times prior$

- If our prior is too specific, then our posterior will be “dominated by” the prior.

- If our prior is too vague, then our posterior will be “dominated by” the data through the likelihood.

PRIOR INFLUENCE ON THE POSTERIOR

- If our prior is too specific, then our posterior will be "dominated by" the prior.
- If our prior is too vague, then our posterior will be "dominated by" the data through the likelihood.

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

- If $P(A) = 0, P(A|B) = 0$.
- If $P(A) = 1, P(B|A) = P(B) \Rightarrow P(A|B) = 1$.

TERMS

- Improper Priors
  - Priors that are not valid probability functions.

- Uninformative Priors
  - Includes minimal information about θ (i.e. physical limitations)

- Informative Priors
  - Includes prior knowledge about θ by taking past data and information into account. (i.e. scientific research)

BAYESIAN & FREQUENTIST STATISTICS

- Say we want to conduct inference on μ, the mean height of American adults.
  - Recall: A prior summarizes our beliefs about μ before observing any data.
  - What is an example of an **improper prior**?
  - What is an example of an **uninformative prior**?
  - What is an example of an **informative prior**?

BAYESIAN & FREQUENTIST STATISTICS

- Frequentist analysis makes no assumptions about the prior distribution of the parameter.

- You can think of a completely flat Uniform, improper prior distribution - this is equivalent to frequentism!

SPECIFYING THE LIKELIHOOD

DEFINITIONS

- We can think of our posterior distribution $f(\theta|y)$ as a combination of our data and our prior.
    - $f(\theta|y) \propto f(y|\theta) \times f(\theta) = likelihood \times prior$

- We want our likelihood to reflect the model that allows us to observe the data we observe.
    - If my data was observing k heads out of n coin flips, the Binomial distribution is probably a good model for how many heads I observe.
    - If my data was observing the number of people who visit my website in a fixed amount of time, the Poisson or Negative Binomial distribution might be a good model.

LIKELIHOOD PRINCIPLE

- The likelihood principle tells us that the data influences our posterior distribution **only** through the likelihood function.
  - The data should not influence our posterior distribution through the prior!

LIKELIHOOD PRINCIPLE

- The likelihood principle tells us that the data influences our posterior distribution **only** through the likelihood function.
  - The data should not influence our posterior distribution through the prior!
    - We may estimate a prior distribution from a pilot study or previous knowledge, but the data for our experiment/analysis should only affect our posterior through the likelihood!

CONJUGACY

- Certain likelihood functions give rise to particularly nice posterior distributions.
  - Normal prior, Normal likelihood ⇒ Normal posterior.
  - Beta prior, Binomial likelihood ⇒ Beta posterior.
  - Gamma prior, Poisson likelihood ⇒ Gamma posterior.

- This is called **conjugacy**.
  - Prior and posterior follow the same parametric distribution.

CONJUGACY

- Conjugacy used to be a very important concept in statistics. Why?

CONJUGACY

- This requires a working knowledge of common statistical distributions, your data-generating process, and your subject area.
  - "Think Bayes!" walks through these well.

WHAT HAPPENS WITHOUT CONJUGACY?

- Suppose I want to conduct inference on some parameter θ.
  - Before observing data, I **really** believe that θ follows a Wishart distribution.
  - I **really** believe that my data generating process $y|\theta$ follows a Cauchy distribution.

WHAT HAPPENS WITHOUT CONJUGACY?

- Suppose I want to conduct inference on some parameter θ.
  - Before observing data, I **really** believe that θ follows a Wishart distribution.
  - I **really** believe that my data generating process $y|\theta$ follows a Cauchy distribution.

- Strategy 1: Instead of picking Wishart/Cauchy distributions, I pick distributions that might reflect the real world less in order for my prior and likelihood to "play nicely" together.

WHAT HAPPENS WITHOUT CONJUGACY?

- Suppose I want to conduct inference on some parameter θ.
  - Before observing data, I **really** believe that θ follows a Wishart distribution.
  - I **really** believe that my data generating process $y|\theta$ follows a Cauchy distribution.

- Strategy 2: Monte Carlo simulations!

CALCULATING THE POSTERIOR

SIMULATING THE POSTERIOR

$$f(\theta|y) \propto f(y|\theta) \times f(\theta)$$

1. Specify $f(y|\theta)$ and $f(\theta)$.

2. Simulate one value from $f(\theta)$, called θ'.

3. Using the value θ', find and plot the height of $f(y|\theta')$.

4. Repeat this large number of times.

SIMULATING THE POSTERIOR

$$f(\theta|y) \propto f(y|\theta) \times f(\theta)$$

- Once we've simulated the posterior distribution, we can do whatever we want to do with it.
  - Estimate the average value of θ.
  - Estimate the median value of θ.
  - Estimate the range of the middle 95% values of θ.