UNIVERSITY OF CAMBRIDGE

Introduction to Inference

Carl Henrik Ek - che29@cam.ac.uk

28th of September 2021

http://carlhenrik.com

Previous Session

- What does it mean to learn from data?

Previous Session

- What does it mean to learn from data?
- What are the conclusions we can draw?

This Session

- Formalise ML

This Session

- Formalise ML
- Probability Theory as one approach to encode beliefs

Statistical Learing Theory

Learning Theory

- $\mathcal{F}$ space of functions

Learning Theory

- $\mathcal{F}$ space of functions
- $\mathcal{A}$ learning algorithm

Learning Theory

- $\mathcal{F}$ space of functions
- $\mathcal{A}$ learning algorithm
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$

Learning Theory

- $\mathcal{F}$ space of functions
- $\mathcal{A}$ learning algorithm
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
- $\mathcal{S} \sim P(\mathcal{X} \times \mathcal{Y})$

Learning Theory

- $\mathcal{F}$ space of functions
- $\mathcal{A}$ learning algorithm
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
- $\mathcal{S} \sim P(\mathcal{X} \times \mathcal{Y})$
- $\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)$ loss function

Statistical Learning

$$e(\mathcal{S}, \mathcal{A}, \mathcal{F}) = \mathbb{E}_{P(\{\mathcal{X},\mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)\right]$$

Statistical Learning

$$
\begin{aligned}
e(\mathcal{S}, \mathcal{A}, \mathcal{F}) &= \mathbb{E}_{P(\{\mathcal{X},\mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)\right] \\
&= \int \ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y) p(x, y) \mathrm{d}x \mathrm{d}y
\end{aligned}
$$

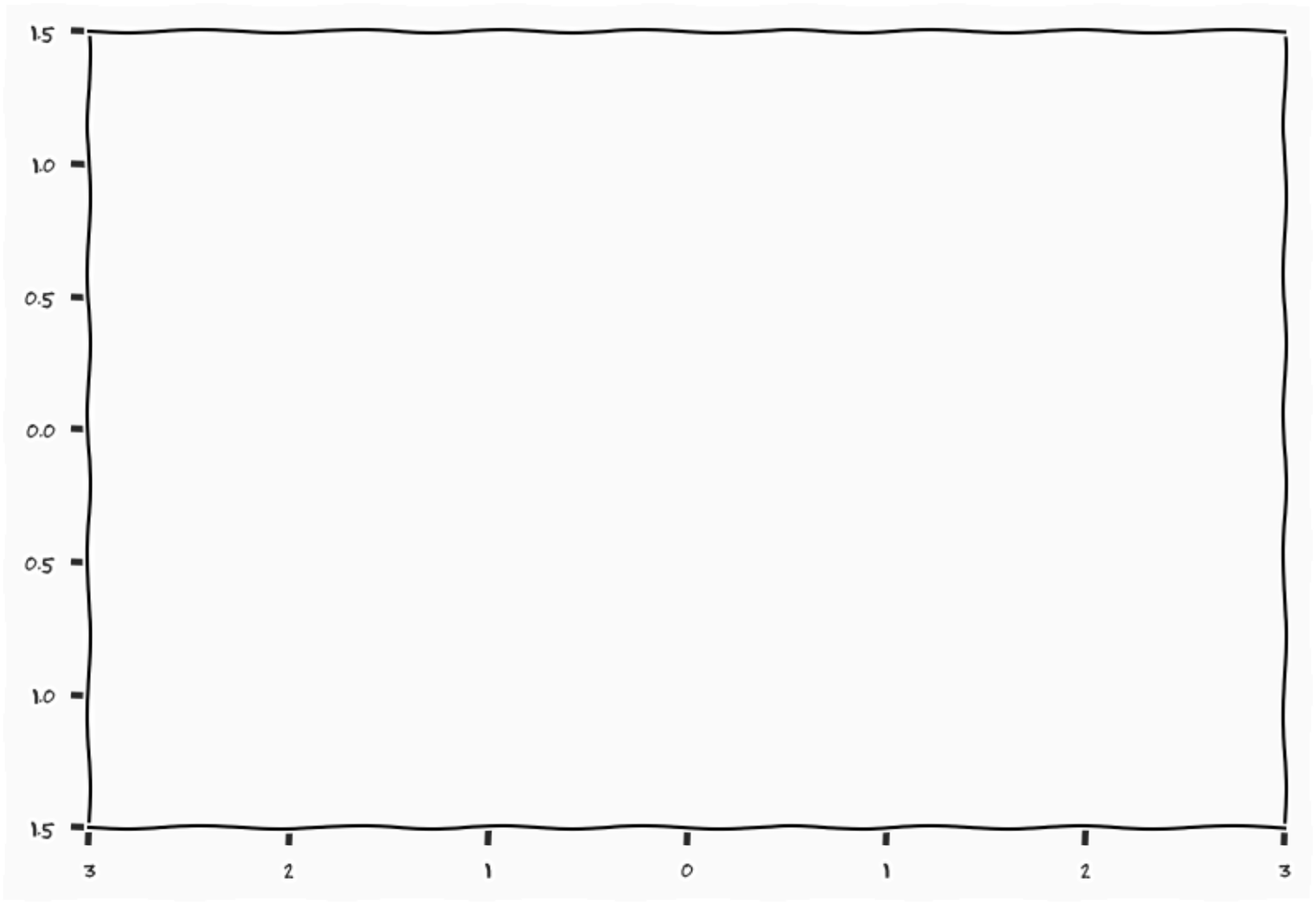
Statistical Learning

$$
\begin{aligned}
e(\mathcal{S}, \mathcal{A}, \mathcal{F}) &= \mathbb{E}_{P(\{\mathcal{X},\mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)\right] \\
&= \int \ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y) p(x, y) \mathrm{d}x \mathrm{d}y \\
&\approx \frac{1}{M} \sum_{n=1}^{M} \ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x_n, y_n)
\end{aligned}
$$

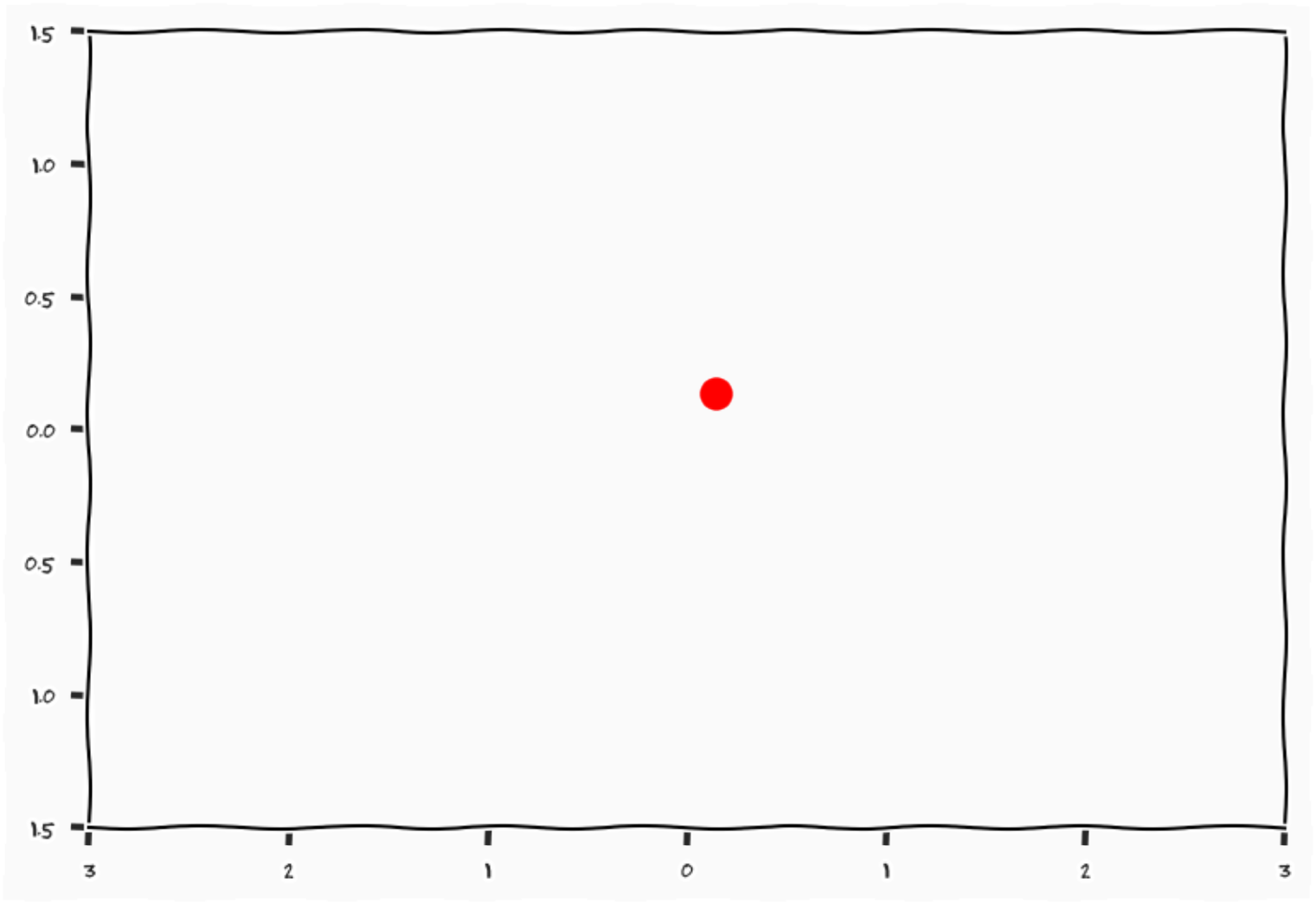
No Free Lunch

We can come up with a combination of $\{\mathcal{S}, \mathcal{A}, \mathcal{F}\}$ that makes $e(\mathcal{S}, \mathcal{A}, \mathcal{F})$ take an arbitary value

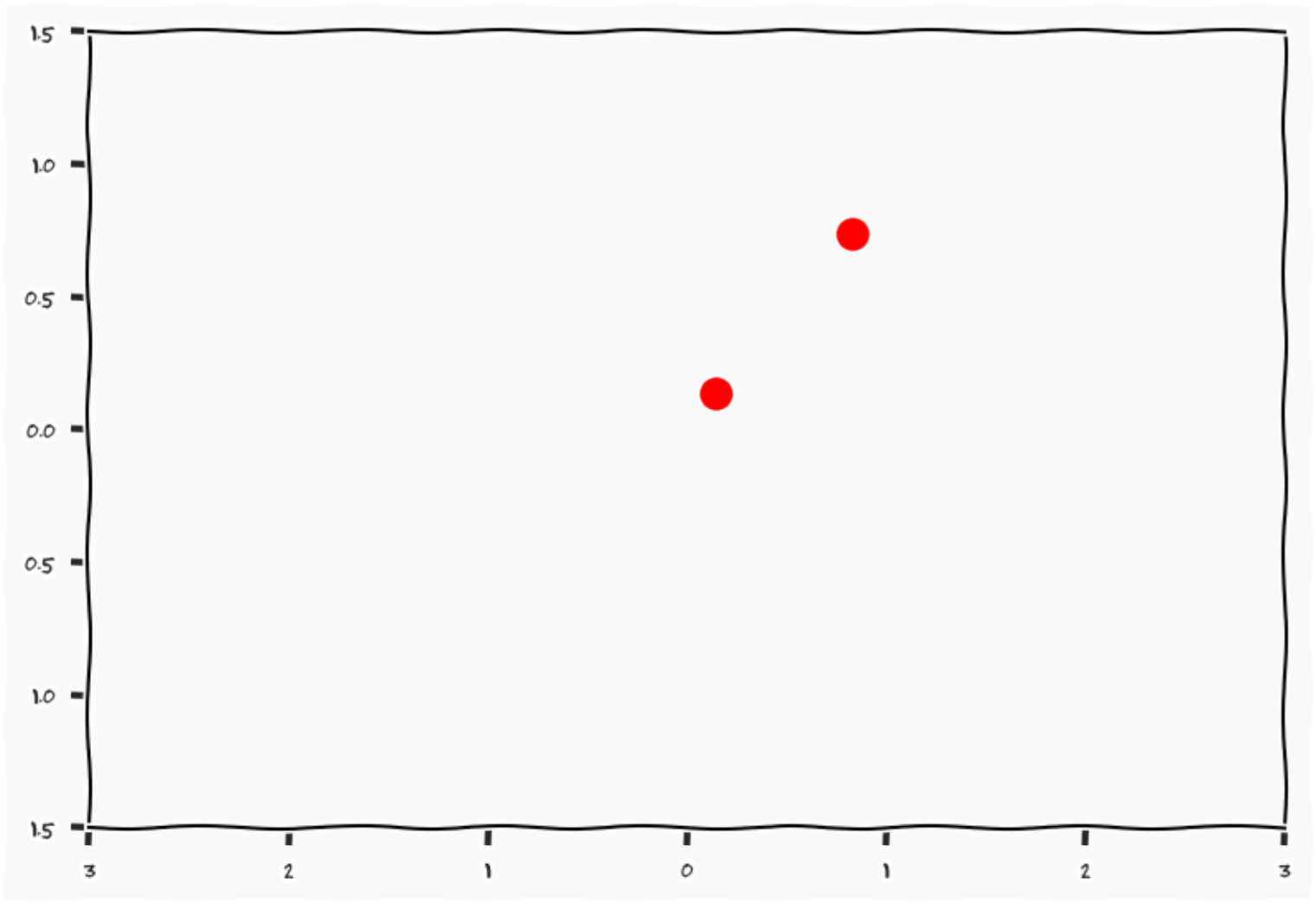
Example



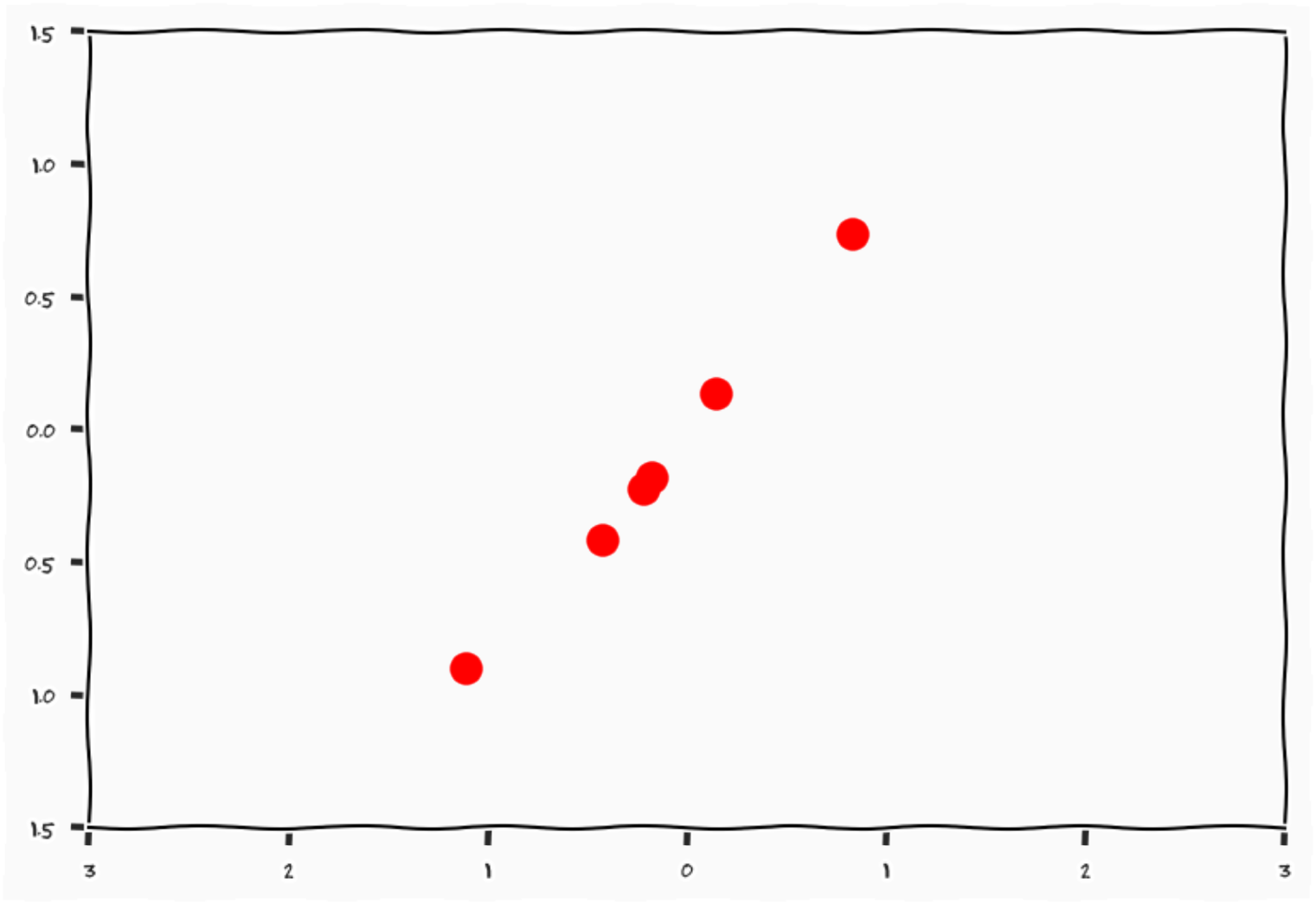
Example



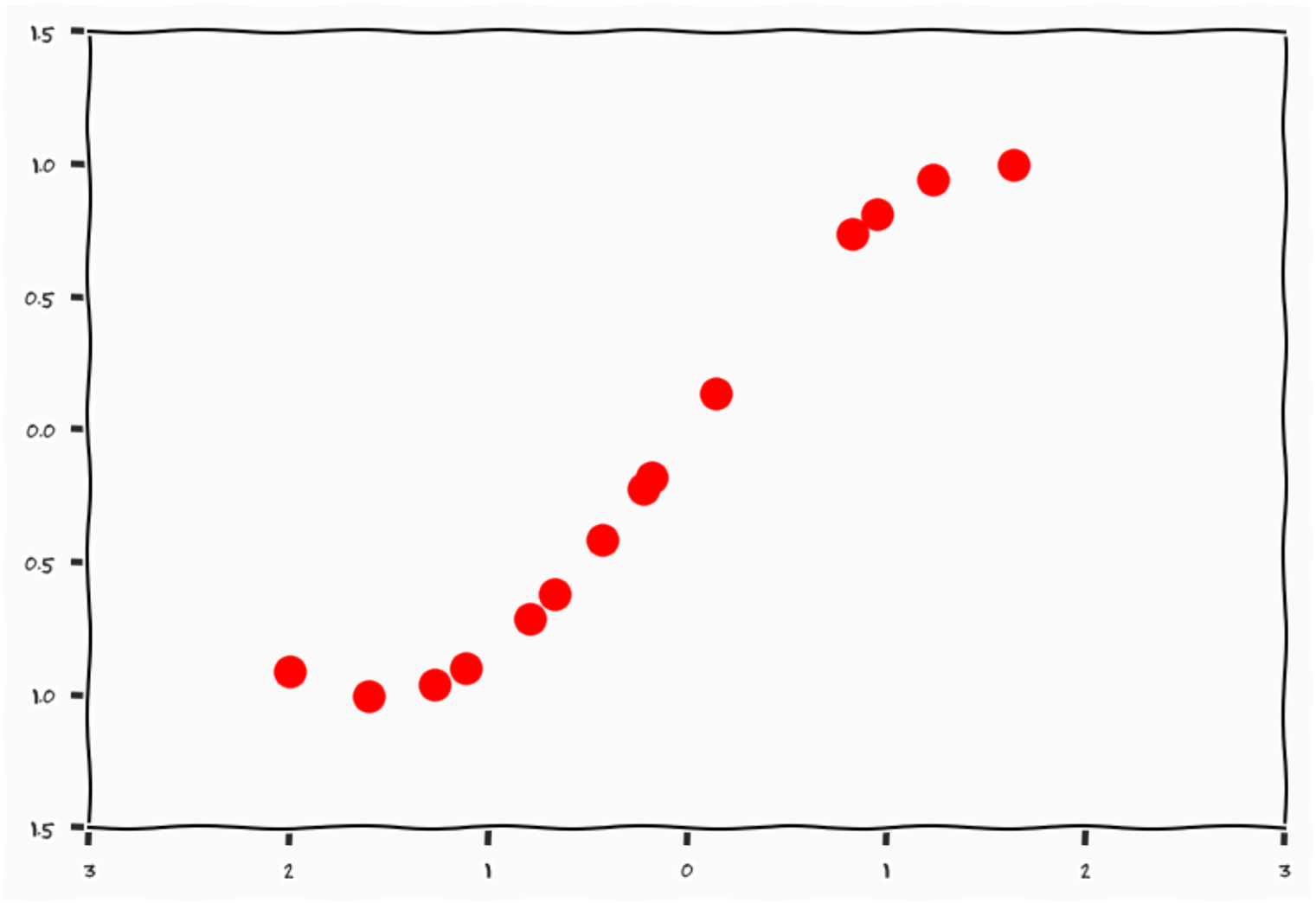
Example



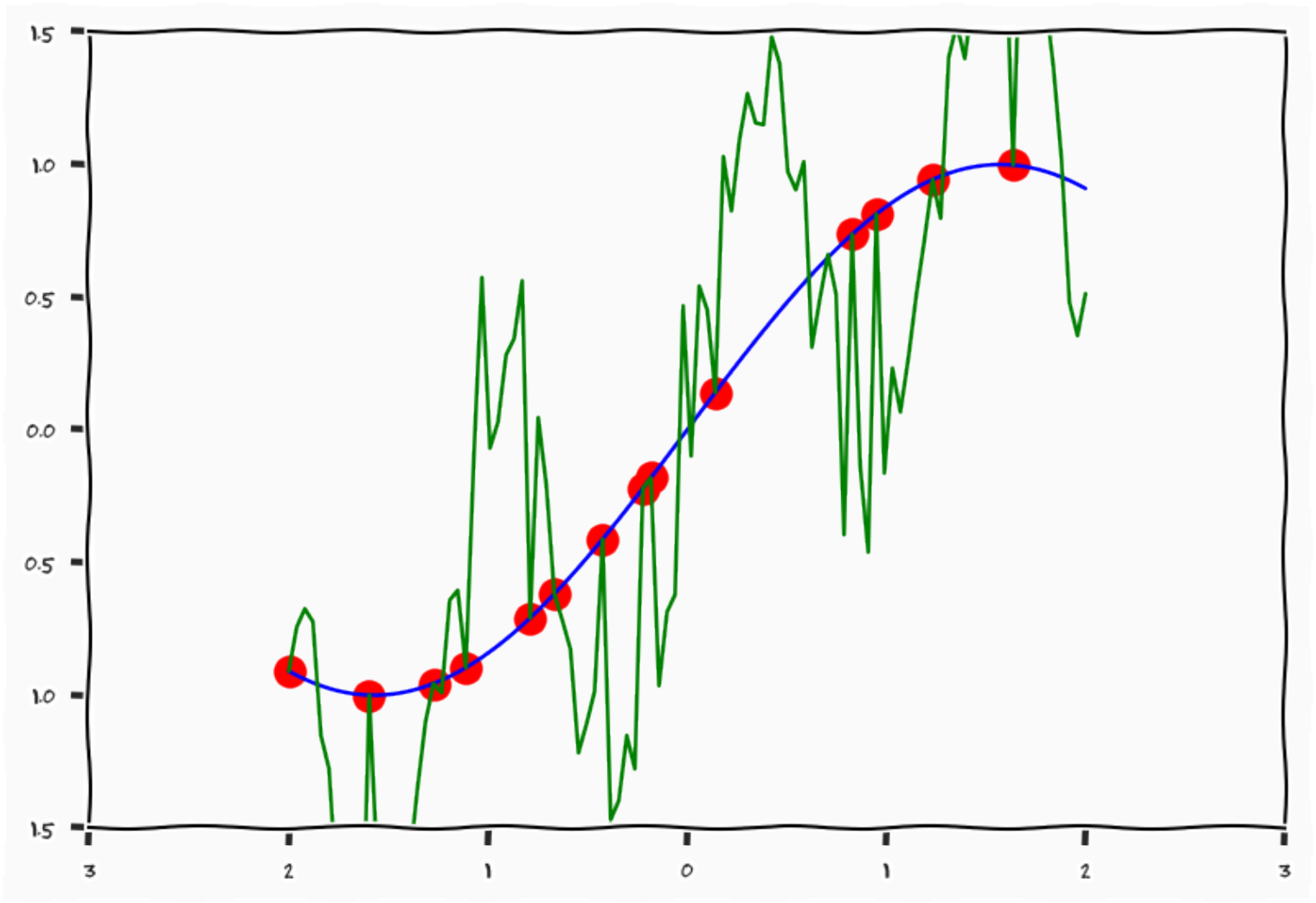
Example



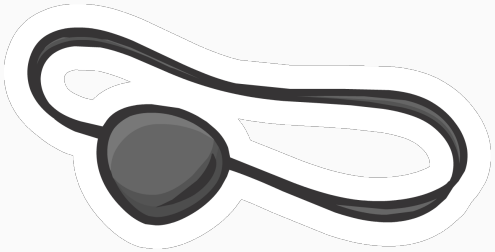
Example



Example



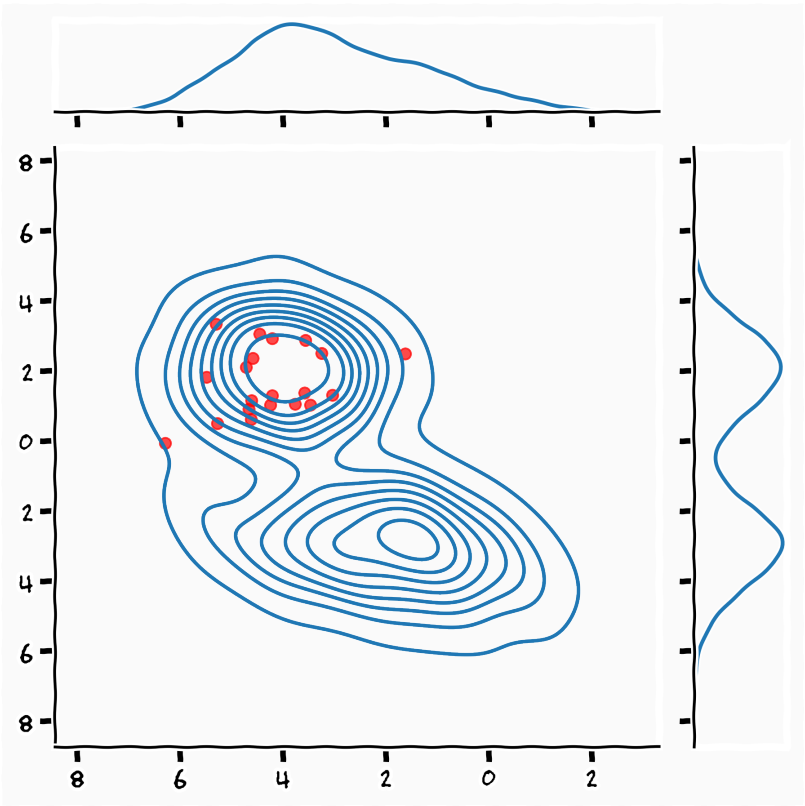
Assumptions: Algorithms



Statistical Learning

$$\mathcal{A}_{\mathcal{F}}(\mathcal{S})$$

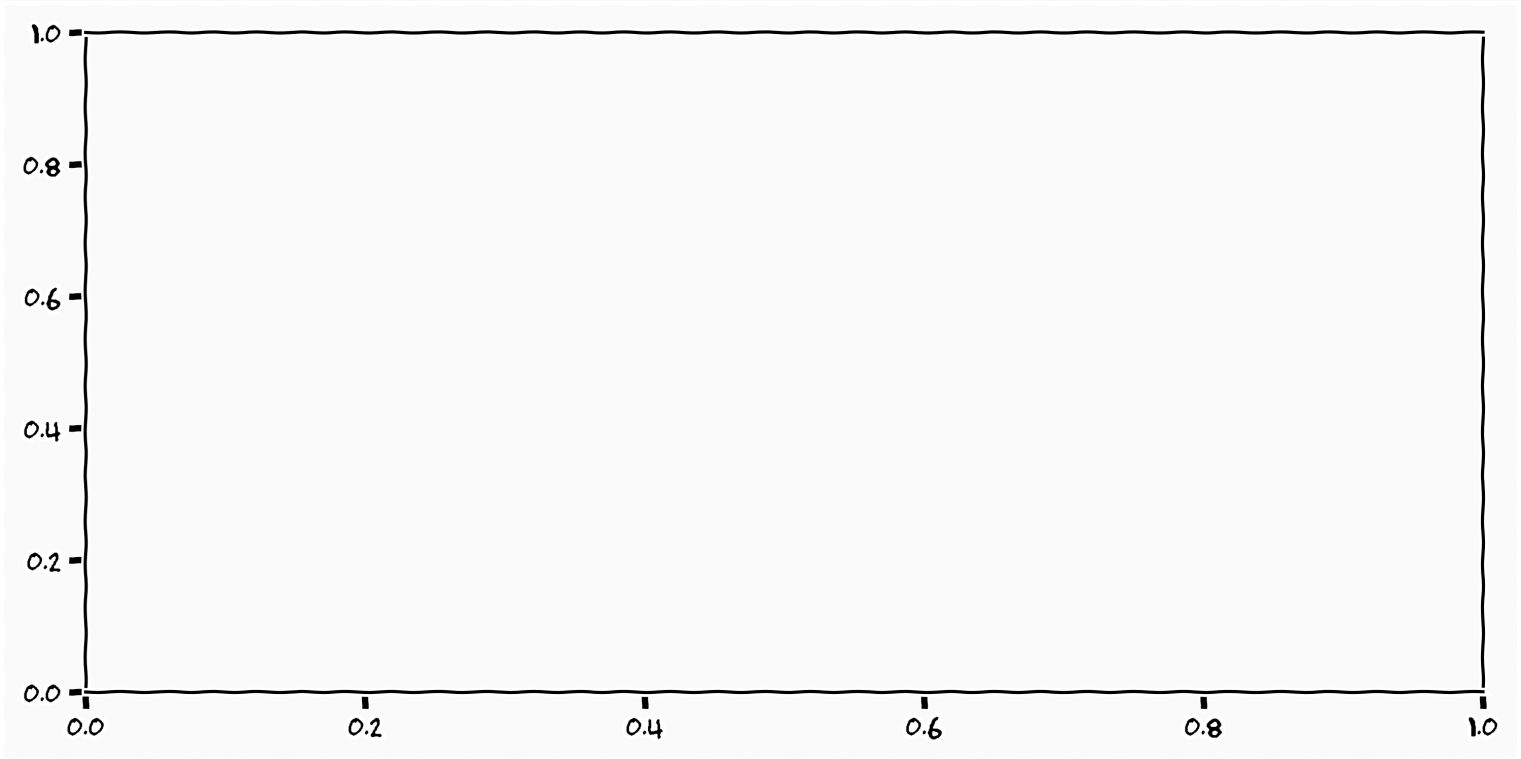
Assumptions: Biased Sample



Statistical Learning

$$\mathcal{A}_{\mathcal{F}}(\mathcal{S})$$

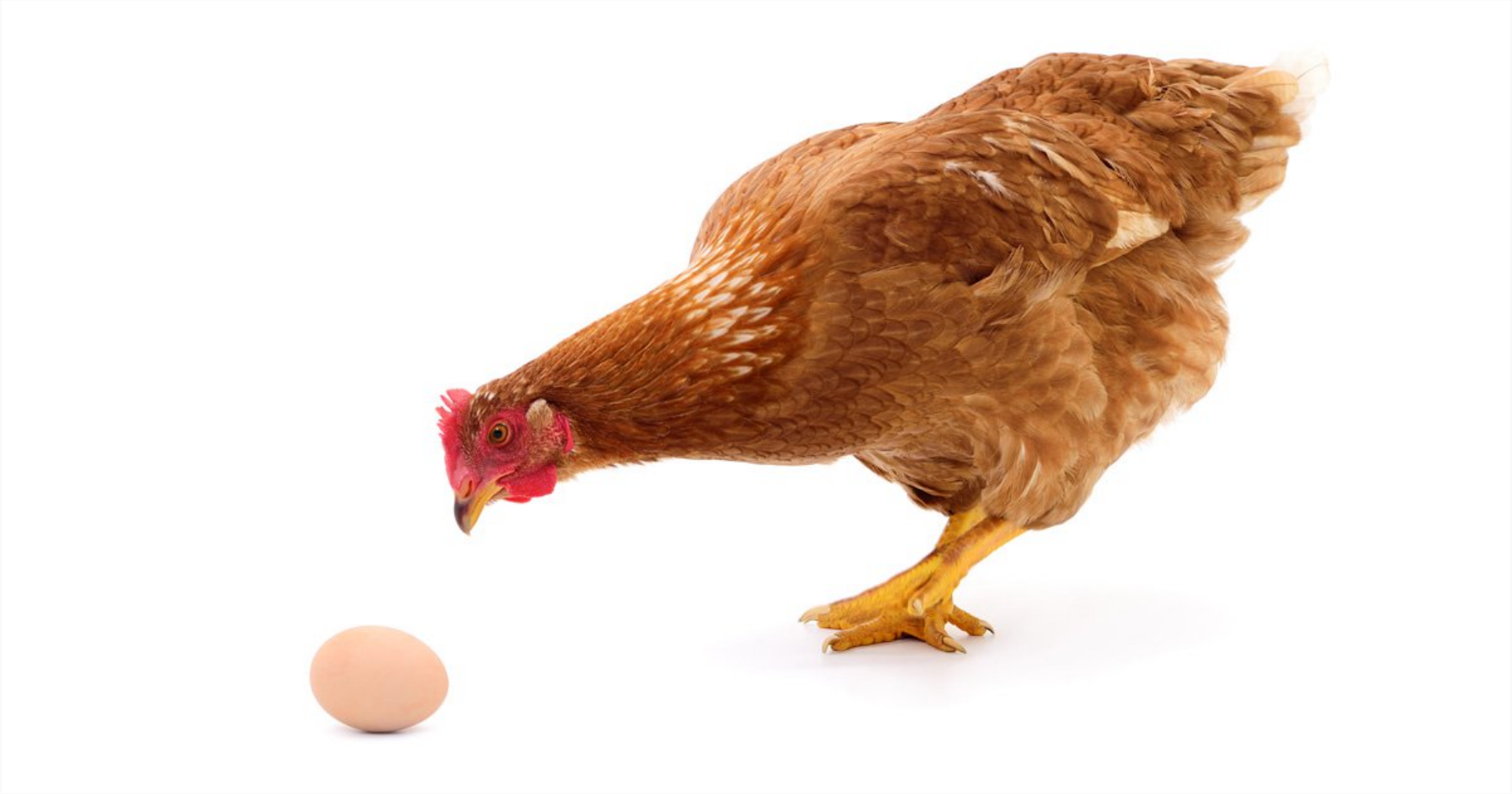
Assumptions: Hypothesis space



Statistical Learning

$$\mathcal{A}_{\mathcal{F}}(\mathcal{S})$$

Chicken and Egg



The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is

The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is
    - The opposite is true

The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is
    - The opposite is true
    - The best possible model has infinite support (nothing is excluded) but very focused mass

The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is
    - The opposite is true
    - The best possible model has infinite support (nothing is excluded) but very focused mass
- Observations cannot be argued with

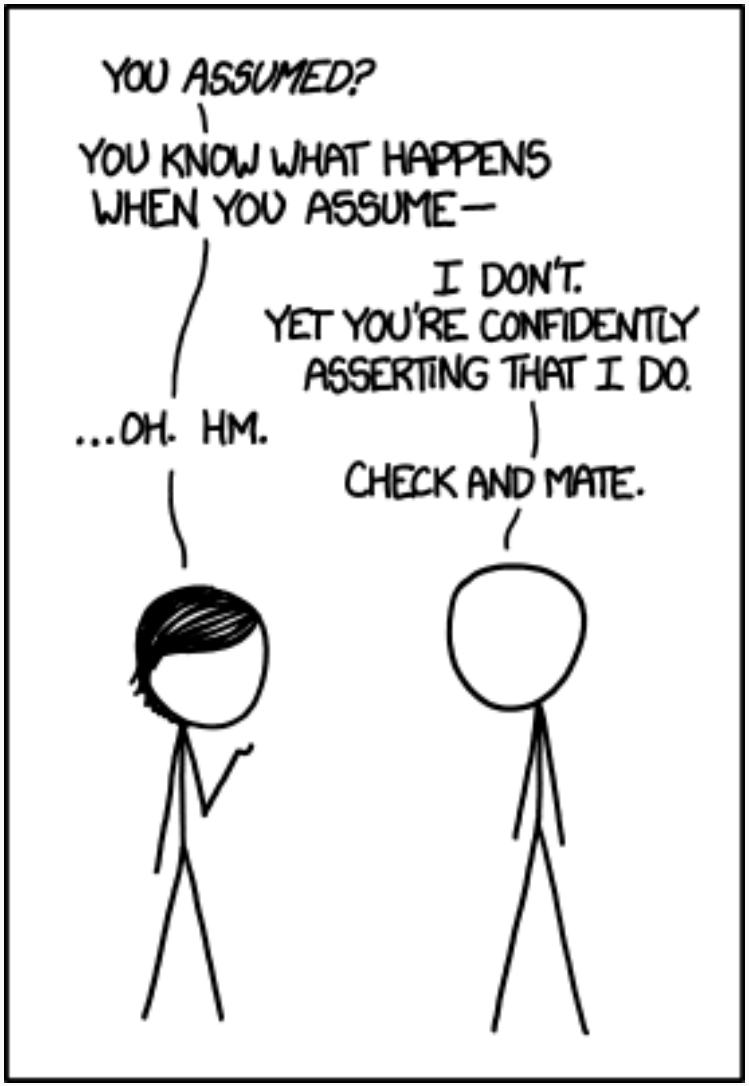
The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is
    - The opposite is true
    - The best possible model has infinite support (nothing is excluded) but very focused mass
- Observations cannot be argued with
- Intepretations of observations are relative to our assumptions

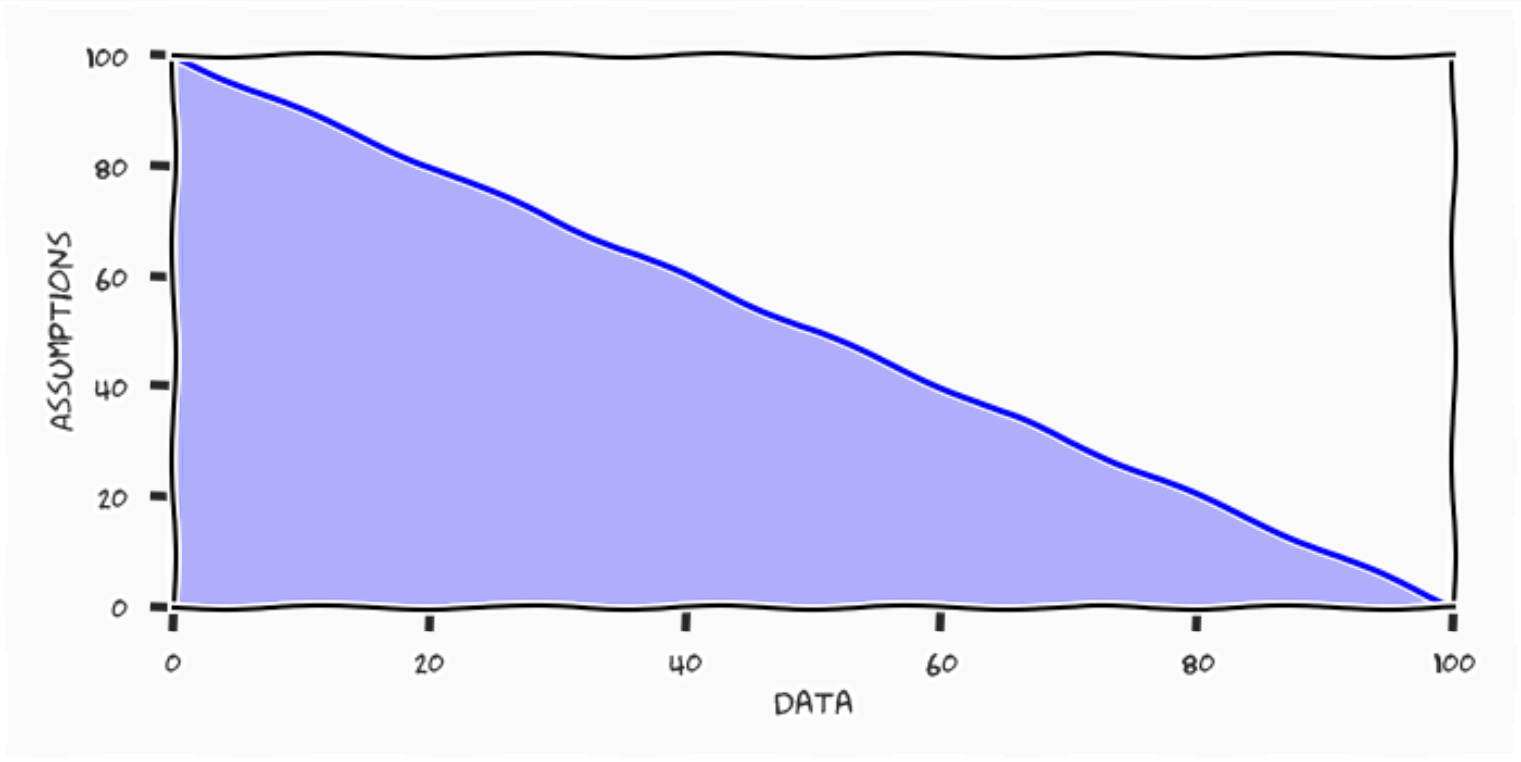
The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is
    - The opposite is true
    - The best possible model has infinite support (nothing is excluded) but very focused mass
- Observations cannot be argued with
- Intepretations of observations are relative to our assumptions
- *Your solution can only ever be interpreted in the light of your assumptions*

Assumptions/Beliefs



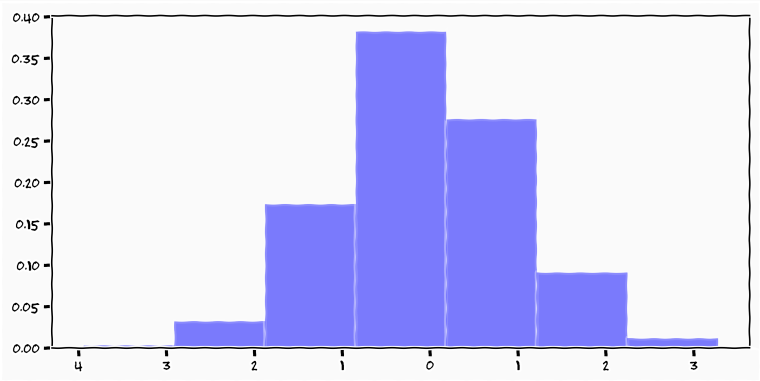
Data and Knowledge

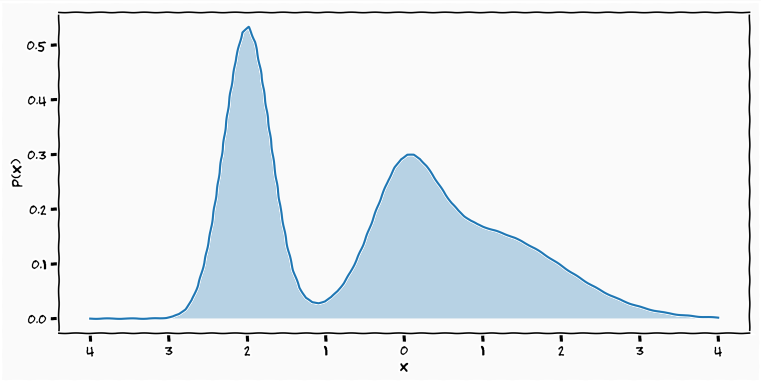


Statistical Modelling

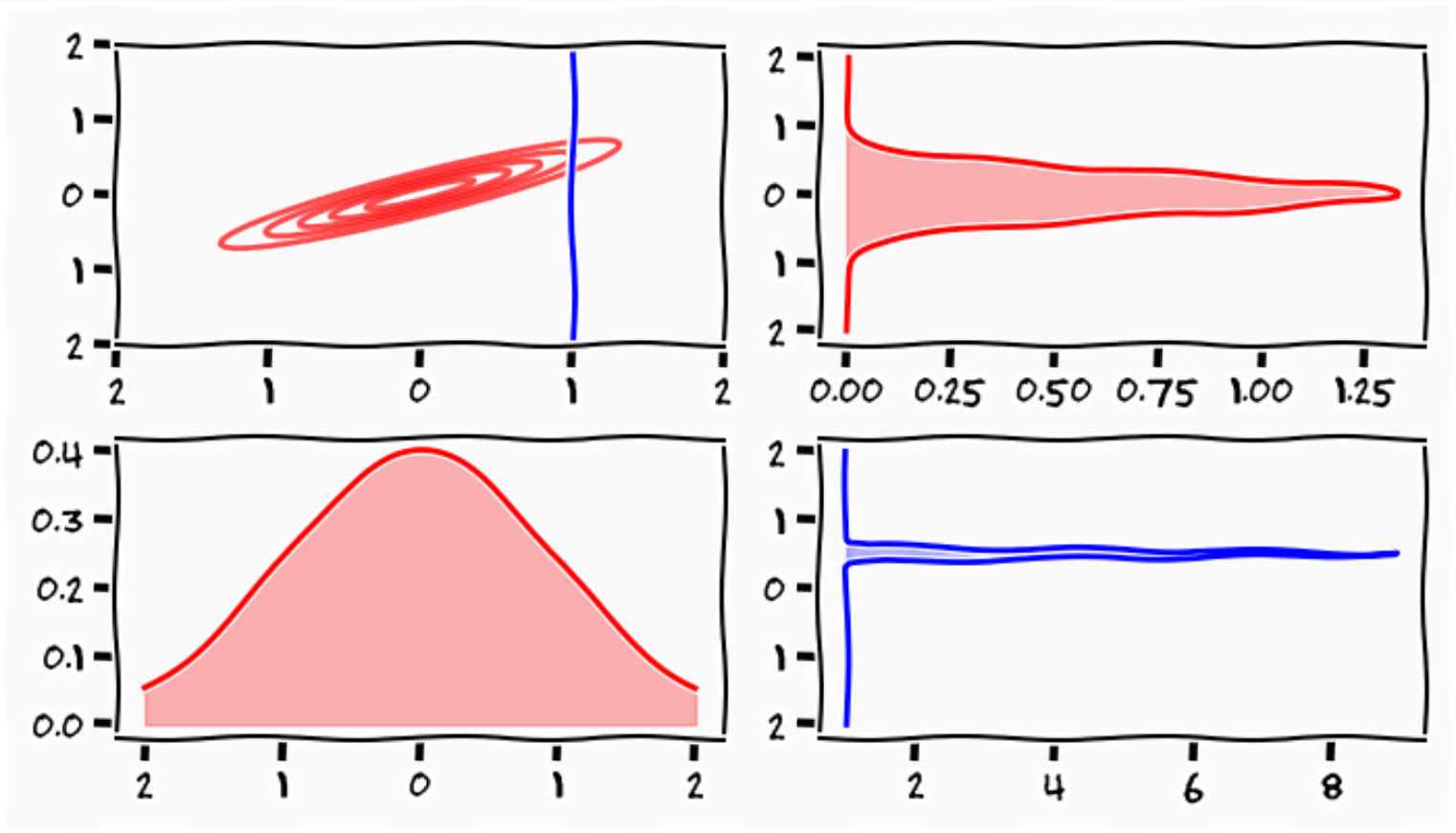
Encoding Beliefs

Probabilities





Basic Probabilities



Variables

Deterministic Variable

Code

```
int x = 3;
float y = 3.14;
```

Stochastic Variable

$$x \sim p(x)$$

$$y \sim \mathcal{N}(0, 1)$$

Rules of Probablity

Sum Rule

$$p(x) = \sum_{\forall y \in \mathcal{Y}} p(x, y)$$

Product Rule

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

Baye's "Rule"

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

Baye's "Rule"

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

Baye's "Rule"

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

Baye's "Rule"

$$
\begin{aligned}
p(x, y) &= p(y|x)p(x) \\
p(x, y) &= p(x|y)p(y) \\
p(x|y)p(y) &= p(y|x)p(x) \\
p(x|y) &= \frac{p(y|x)p(x)}{p(y)}
\end{aligned}
$$

Baye's "Rule"

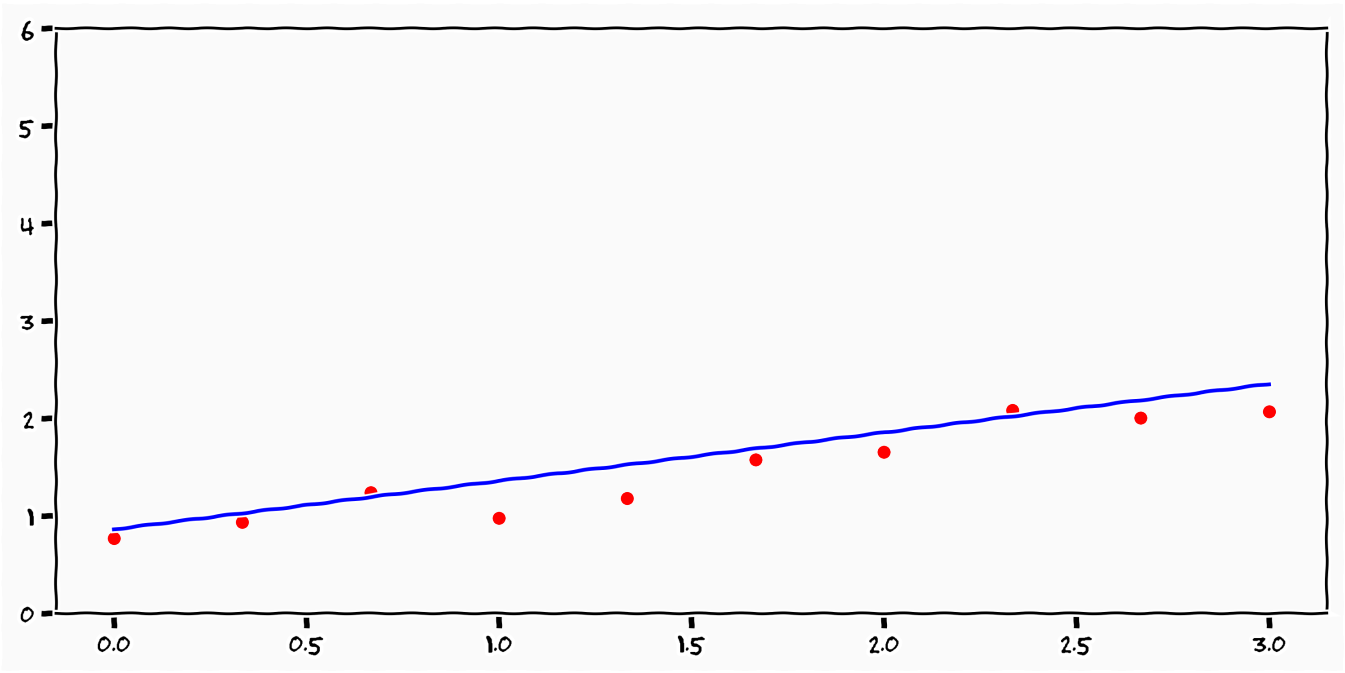
$$
\begin{aligned}
p(x, y) &= p(y|x)p(x) \\
p(x, y) &= p(x|y)p(y) \\
p(x|y)p(y) &= p(y|x)p(x) \\
p(x|y) &= \frac{p(y|x)p(x)}{p(y)} \\
&= \frac{p(y|x)p(x)}{\sum_x p(y|x)p(x)}
\end{aligned}
$$

Why get a tattoo?

"Model"

$$y = f(x)$$

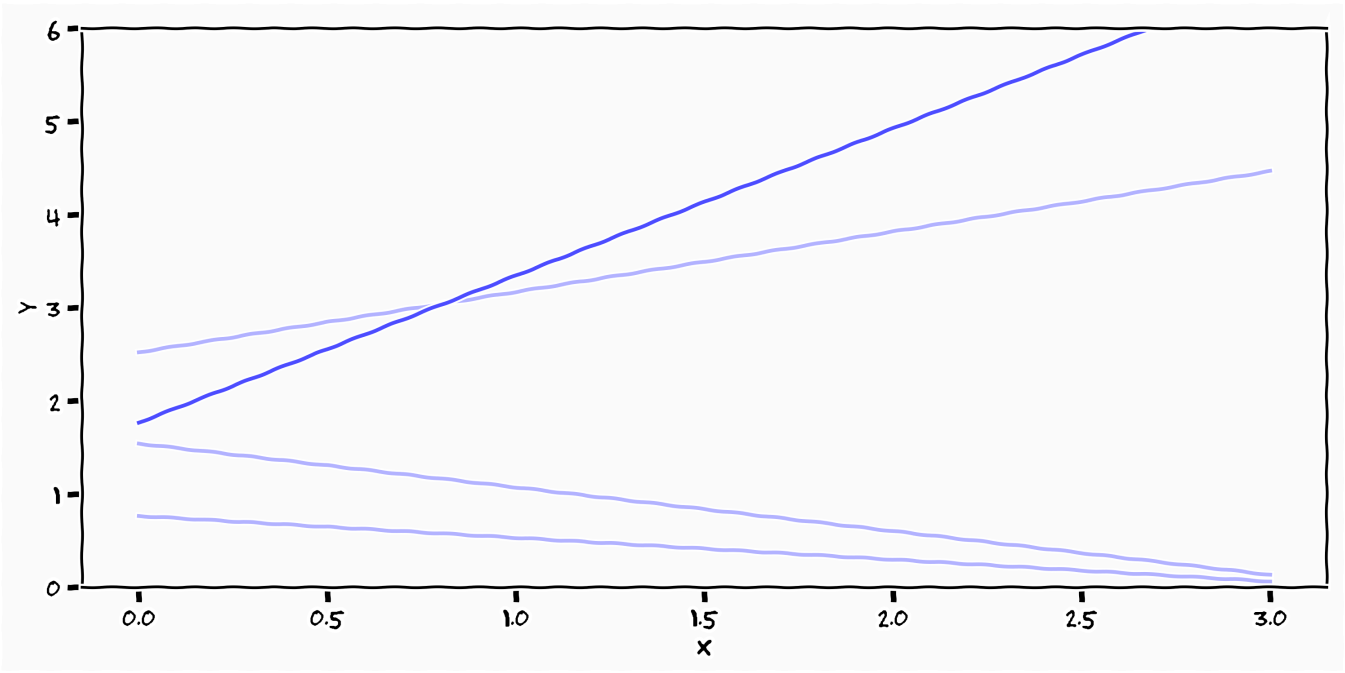
Likelihood



$$p(y \mid f, x)$$

How strongly do I believe that the data y that I see comes from the function f at input x

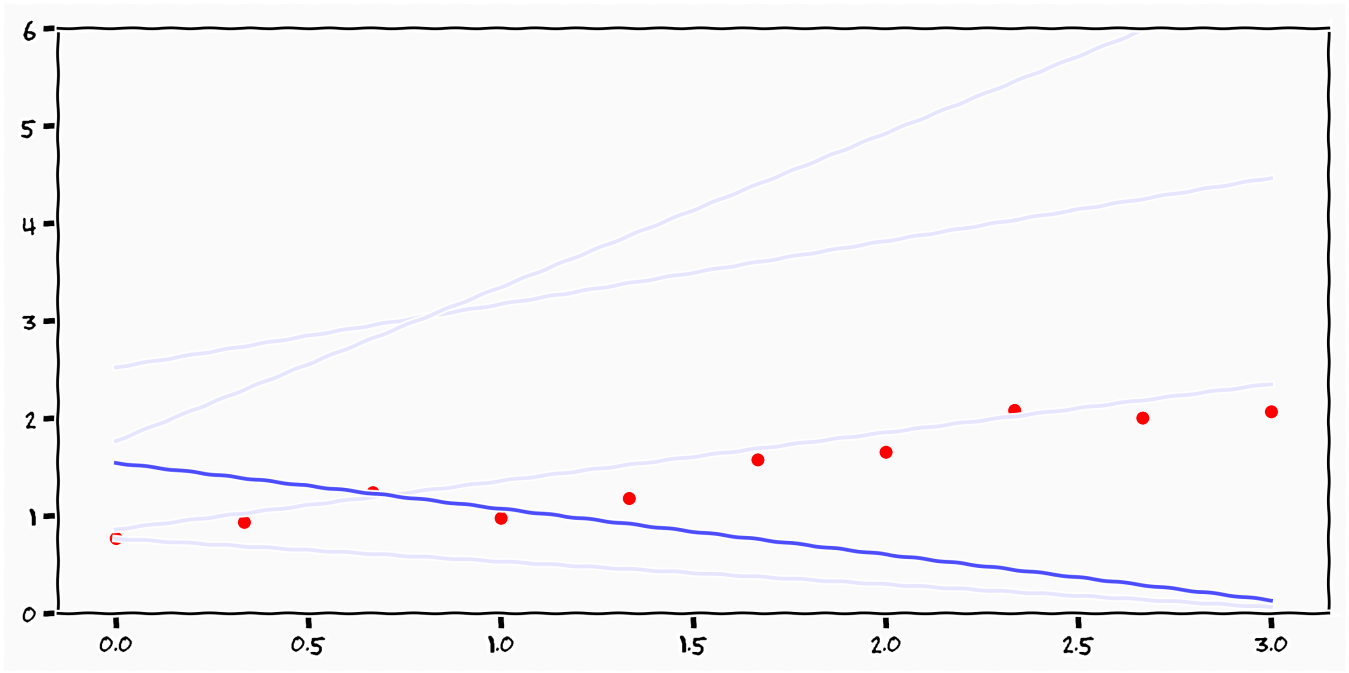
Prior



$$p(f)$$

Before I have seen the data how much do I believe in each possible function?

Posterior



$$p(f \mid y)$$

Given that I have seen observations y what is my updated belief about the function f

Bayes Rule

$$\underbrace{p(f \mid y)}_{\text{Posterior}} = \frac{\overbrace{p(y \mid f, x)}^{\text{Likelihood}} \overbrace{p(f)}^{\text{Prior}}}{\underbrace{p(y)}_{\text{Marginal Likelihood}}}$$

Model we define `Likelihood` and `Prior`

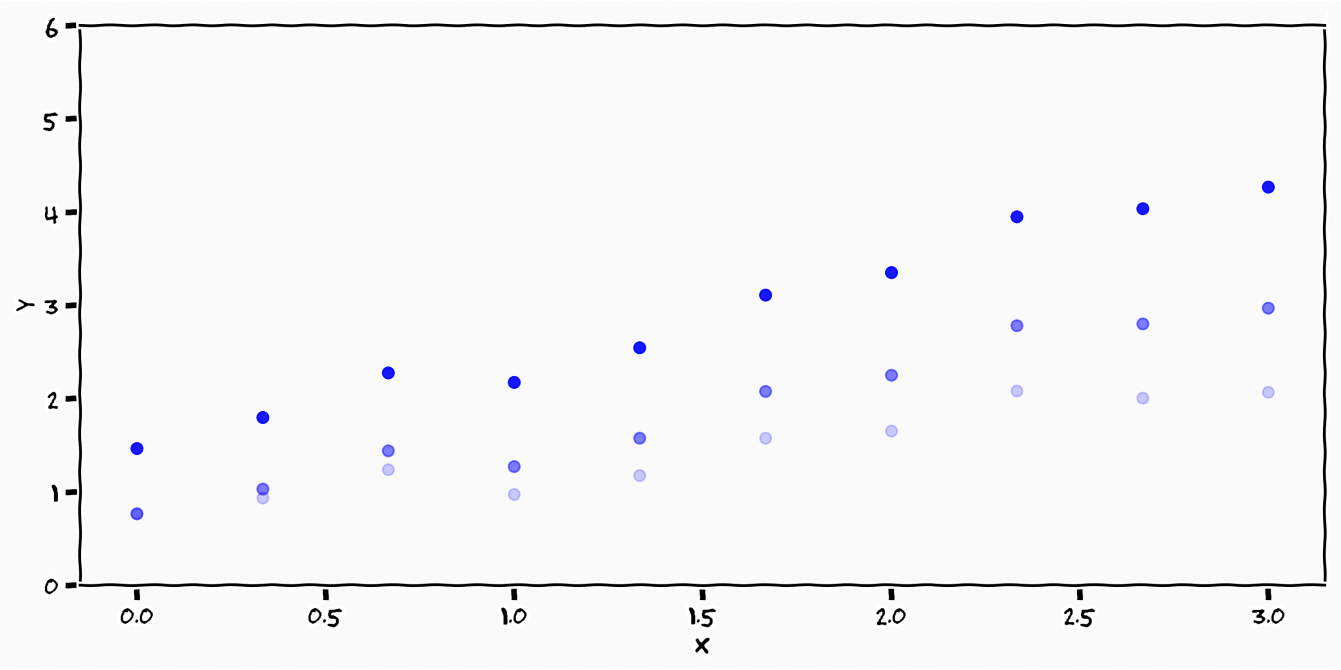
Inference we compute `Posterior` through *marginalisation* of our beliefs

Marginalisation

$$p(y) = \int p(f, y) \mathrm{d}f = \int p(y|f)p(f) \mathrm{d}f$$

- Marginalisation accounts for all your belief in a variable
- Importantly it does not "remove" the effect of the variable
- Marginalisation is an expectation over a conditional distribution

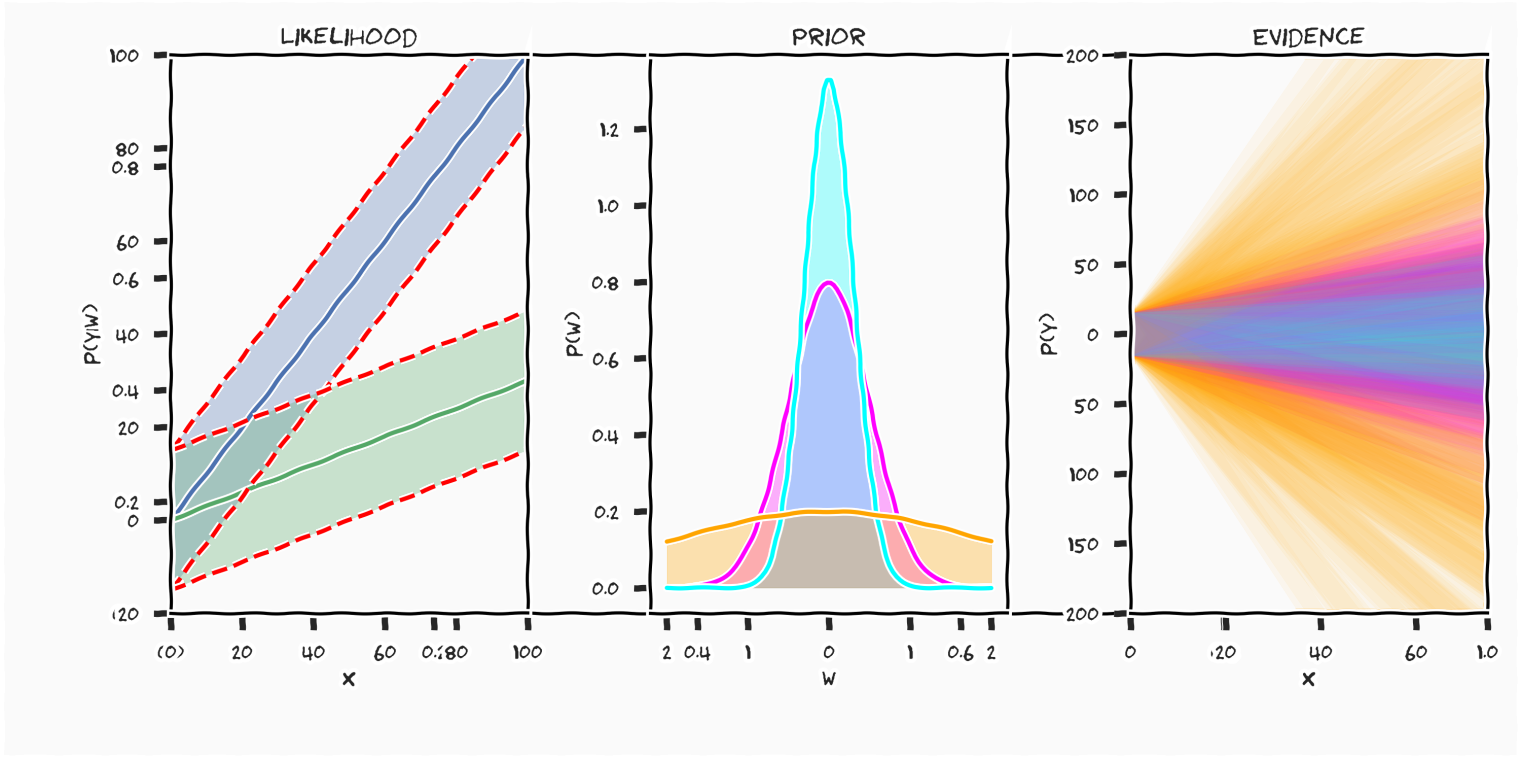
Marginal Likelihood/Evidence



$$p(y) = \int p(y \mid f) p(f) \mathrm{d}f$$

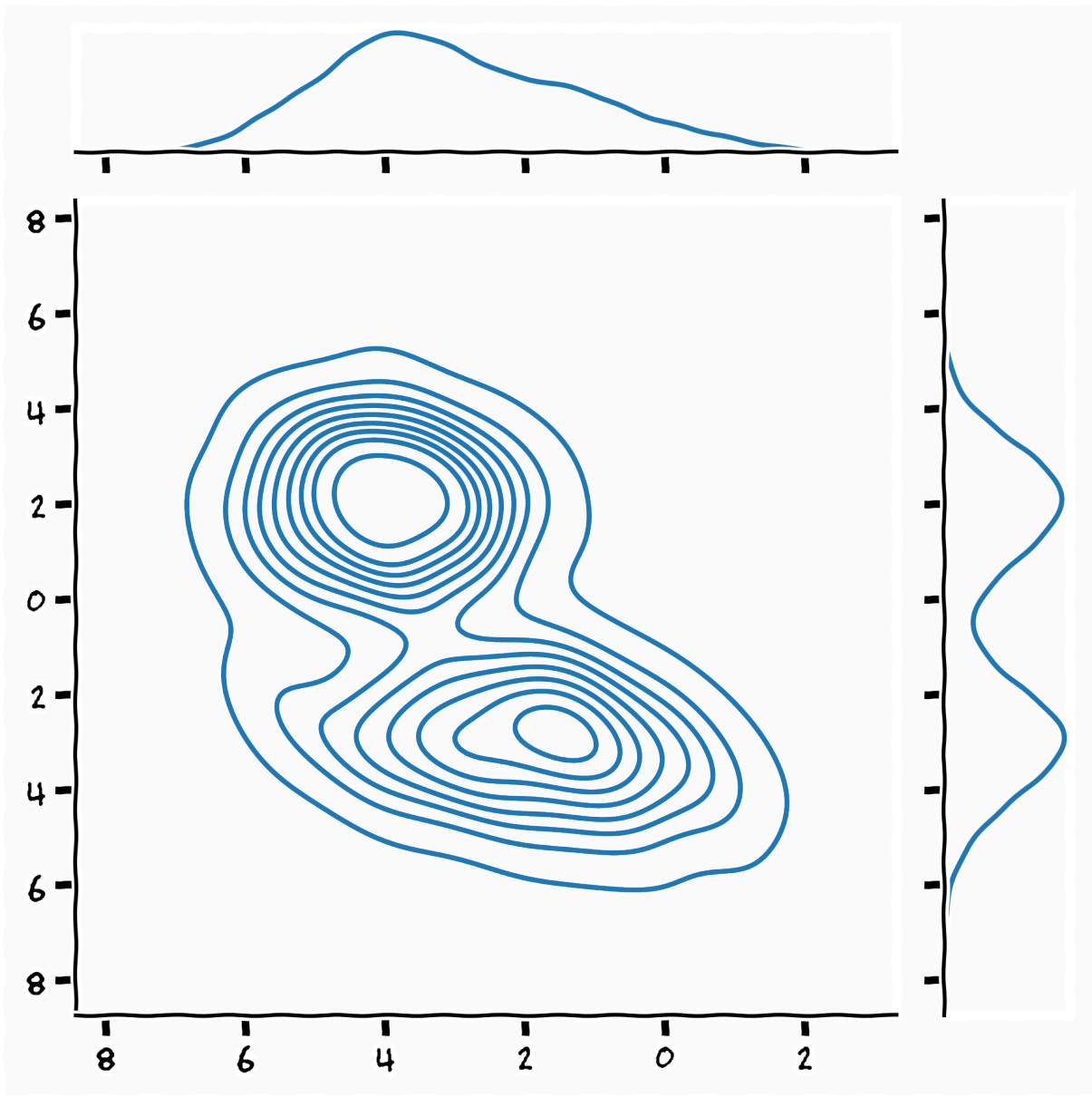
Given different data sets y_i how much evidence do they provide of the model?

Regression Model



$$y = x \cdot w \pm 15$$

Marginal Distribution



Marginalisation

Next time you want to give your friends a compliment, tell them that you have completely marginalised them from your life

Bernoulli Trial

Coin Toss[1]

- We want to figure out if a coin is biased based on data
- We will toss the coin N times

$$\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$$

- "Biasedness" of coin μ

[1] how much do you need to bend a coin to make it biased

Inference

$$p(\mu \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mu)p(\mu)}{p(\mathcal{D})}$$

Beliefs/Assumptions

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu)$$

- We completely trust our capability to see the side the coin landed on

Beliefs/Assumptions

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu)$$

- We completely trust our capability to see the side the coin landed on
- Tossing the coin does not significantly change the coin

Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

- Due to binary outcome

$$p(x = 0|\mu) = 1 - \mu$$

- Distribution

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

Conjugate Prior

Posterior

I want a prior distribution, such that when interacting with the likelihood to reach the posterior, the posterior takes the same functional form as the prior"

Conjugate Prior

Posterior

I want a prior distribution, such that when interacting with the likelihood to reach the posterior, the posterior takes the same functional form as the prior"

Posterior

I only want the data to be able to change the degree of my belief not its fundamental form

Conjugate Priors

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

- The `Beta` distribution is the conjugate prior of the `Bernoulli` likelihood parameter μ

Posterior

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} =\propto p(\mathcal{D}|\mu)p(\mu)$$

Posterior

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} =\propto p(\mathcal{D}|\mu)p(\mu)$$
$$= \prod_{i=1}^{N} \mathsf{Bern}(x_i|\mu)\mathsf{Beta}(\mu|a,b)$$

Posterior

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} =\propto p(\mathcal{D}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathsf{Bern}(x_i|\mu)\mathsf{Beta}(\mu|a,b) \\
&= \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}
\end{aligned}
$$

Posterior

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} =\propto p(\mathcal{D}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathsf{Bern}(x_i|\mu)\mathsf{Beta}(\mu|a,b) \\
&= \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \mu^{\sum_i x_i}(1-\mu)^{\sum_i (1-x_i)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}
\end{aligned}
$$

Posterior

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} \propto p(\mathcal{D}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathsf{Bern}(x_i|\mu)\mathsf{Beta}(\mu|a,b) \\
&= \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \mu^{\sum_i x_i}(1-\mu)^{\sum_i (1-x_i)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{\sum_i x_i}(1-\mu)^{\sum_i (1-x_i)}\mu^{a-1}(1-\mu)^{b-1}
\end{aligned}
$$

Posterior

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} =\propto p(\mathcal{D}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathsf{Bern}(x_i|\mu)\mathsf{Beta}(\mu|a,b) \\
&= \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \mu^{\sum_i x_i}(1-\mu)^{\sum_i (1-x_i)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{\sum_i x_i}(1-\mu)^{\sum_i (1-x_i)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{\sum_i x_i+a-1}(1-\mu)^{\sum_i (1-x_i)+b-1}
\end{aligned}
$$

Posterior II

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{\sum_i x_i + a - 1}(1-\mu)^{\sum_i (1-x_i)+b-1}}{p(\mathcal{D})}$$

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \mu)p(\mu)\mathrm{d}\mu$$

- we still do not know the normaliser/evidence?
- conjugacy means that we know its form

Posterior III

$$p(\mu|\mathcal{D}) \propto \mu^{\underbrace{\sum_i x_i + a - 1}_{a_n - 1}} (1-\mu)^{\underbrace{\sum_i (1 - x_i) + b - 1}_{b_n - 1}}$$

- we know the normaliser to a `Beta` distribution
- now we can avoid computing the integral

Posterior III

$$p(\mu|\mathcal{D}) = \frac{\Gamma(a_n + b_n)}{\Gamma(a_n)\Gamma(b_n)}\mu^{a_n}(1-\mu)^{b_n-1}$$

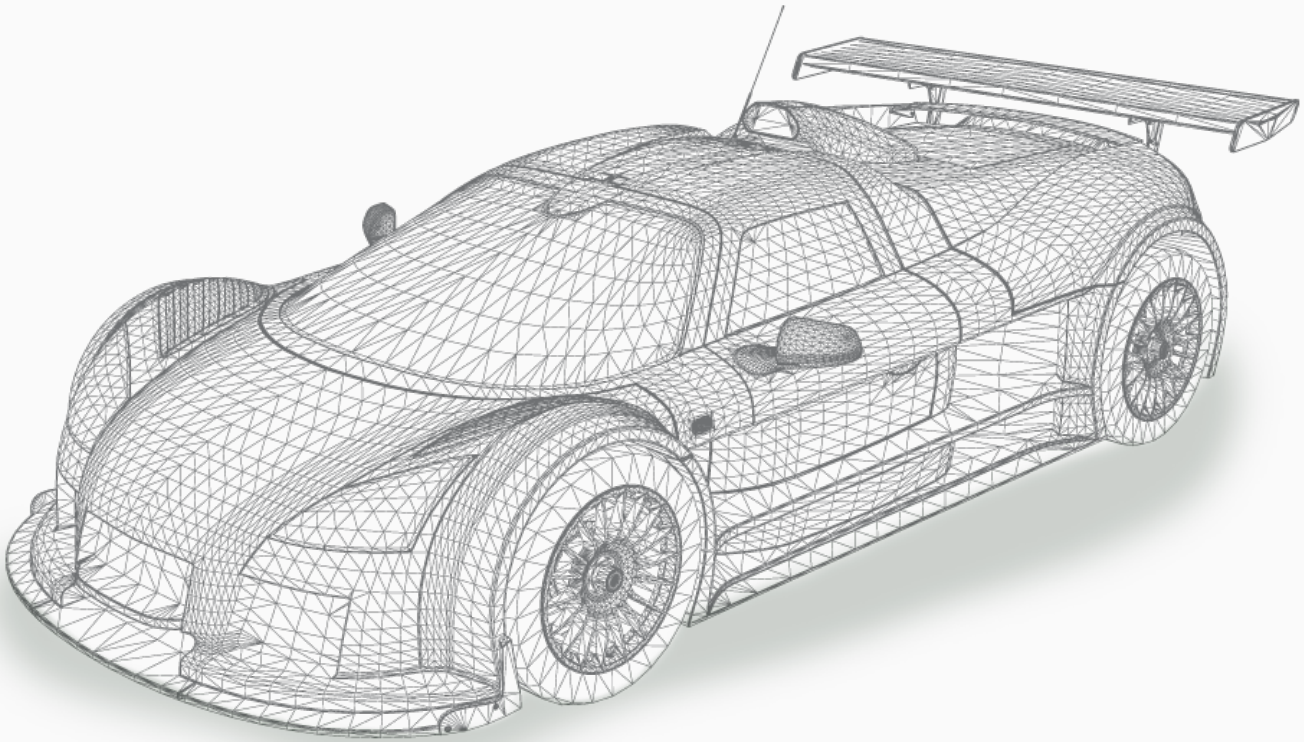
- we know the normaliser to a `Beta` distribution
- now we can avoid computing the integral

Posterior III

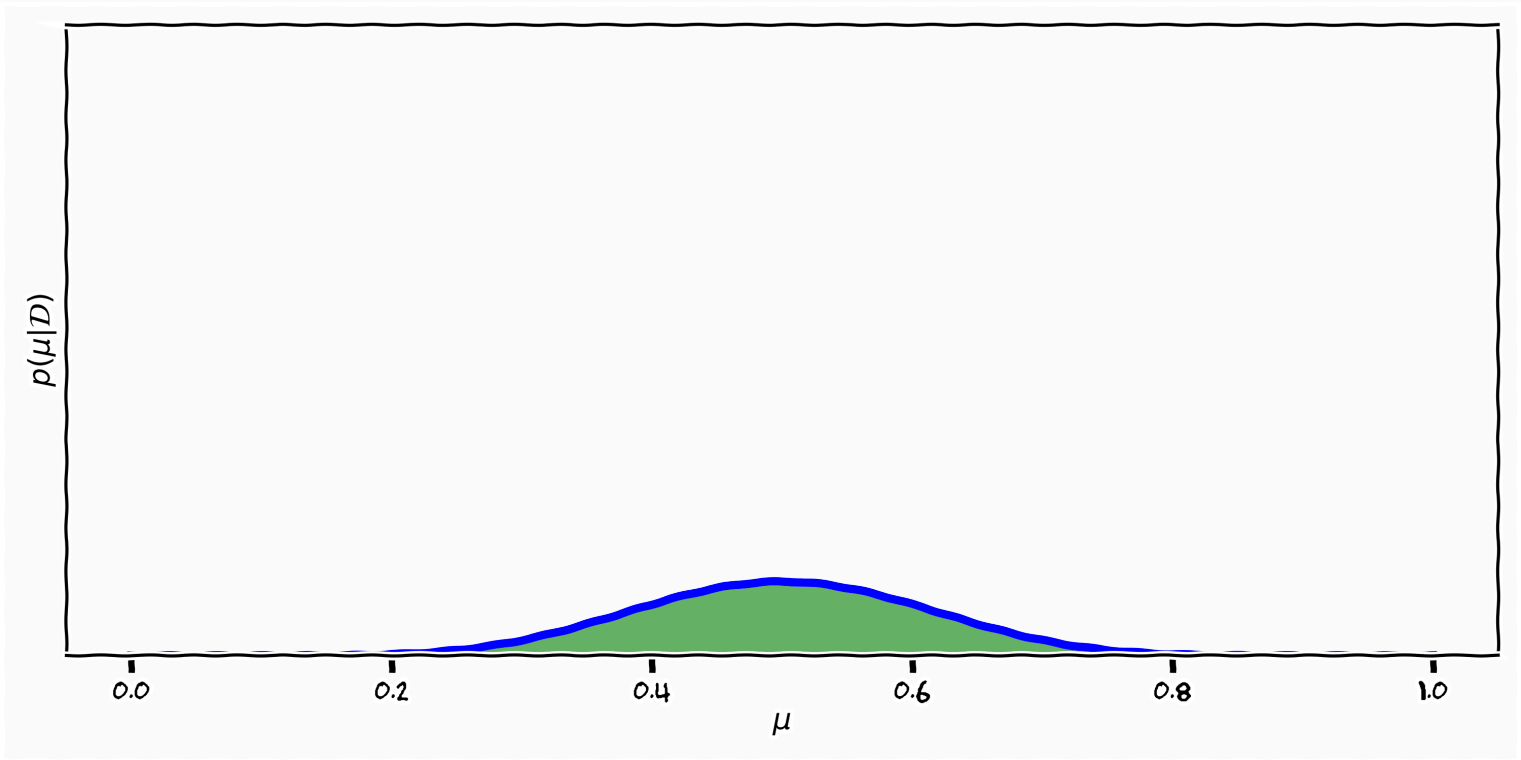
$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \frac{\Gamma(a_n + b_n)}{\Gamma(a_n)\Gamma(b_n)}\mu^{a_n}(1-\mu)^{b_n-1} \\
&= \mathsf{Beta}(\mu|a_n, b_n)
\end{aligned}
$$

- we know the normaliser to a `Beta` distribution
- now we can avoid computing the integral

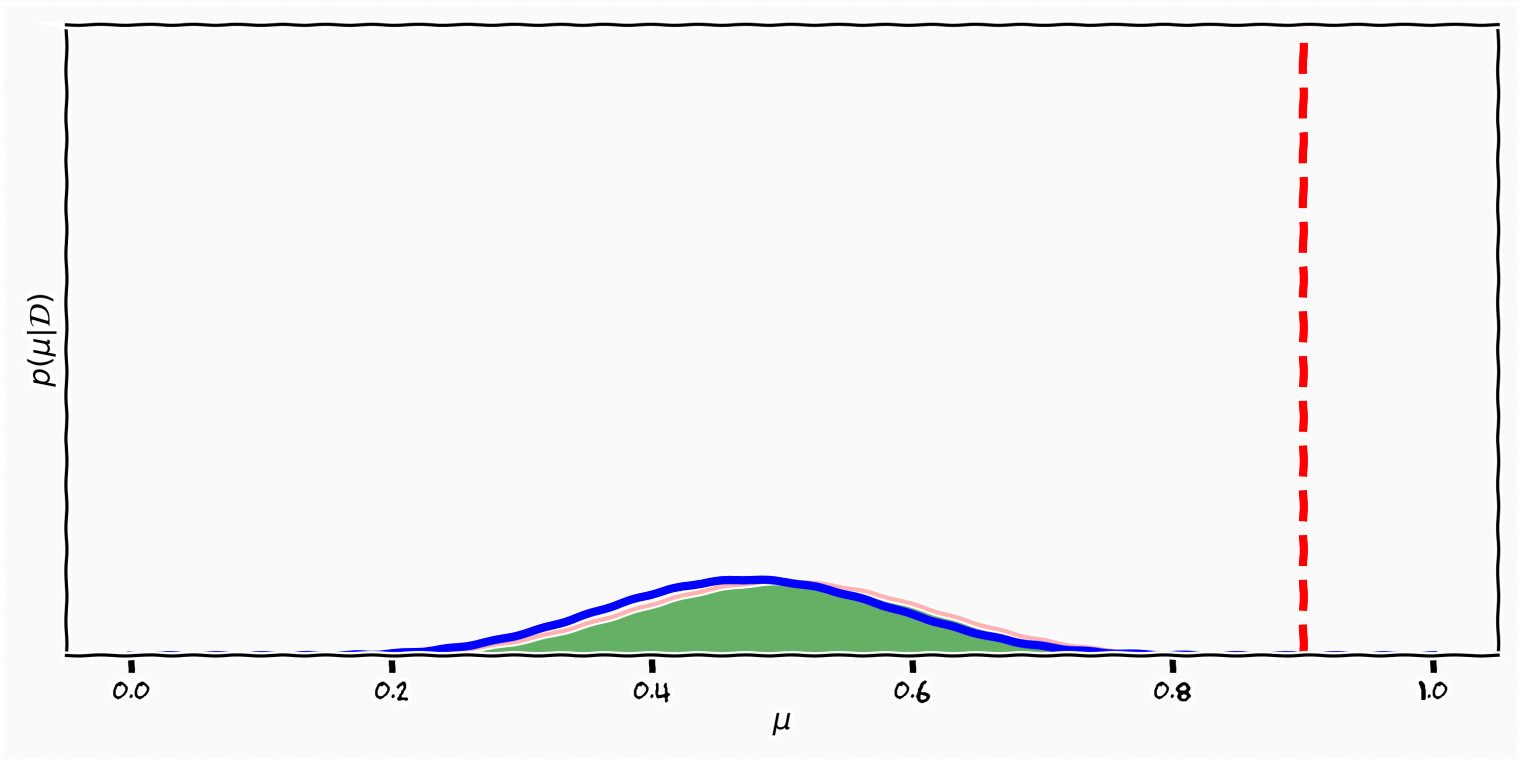
Why its awesome to work in Computer Science!



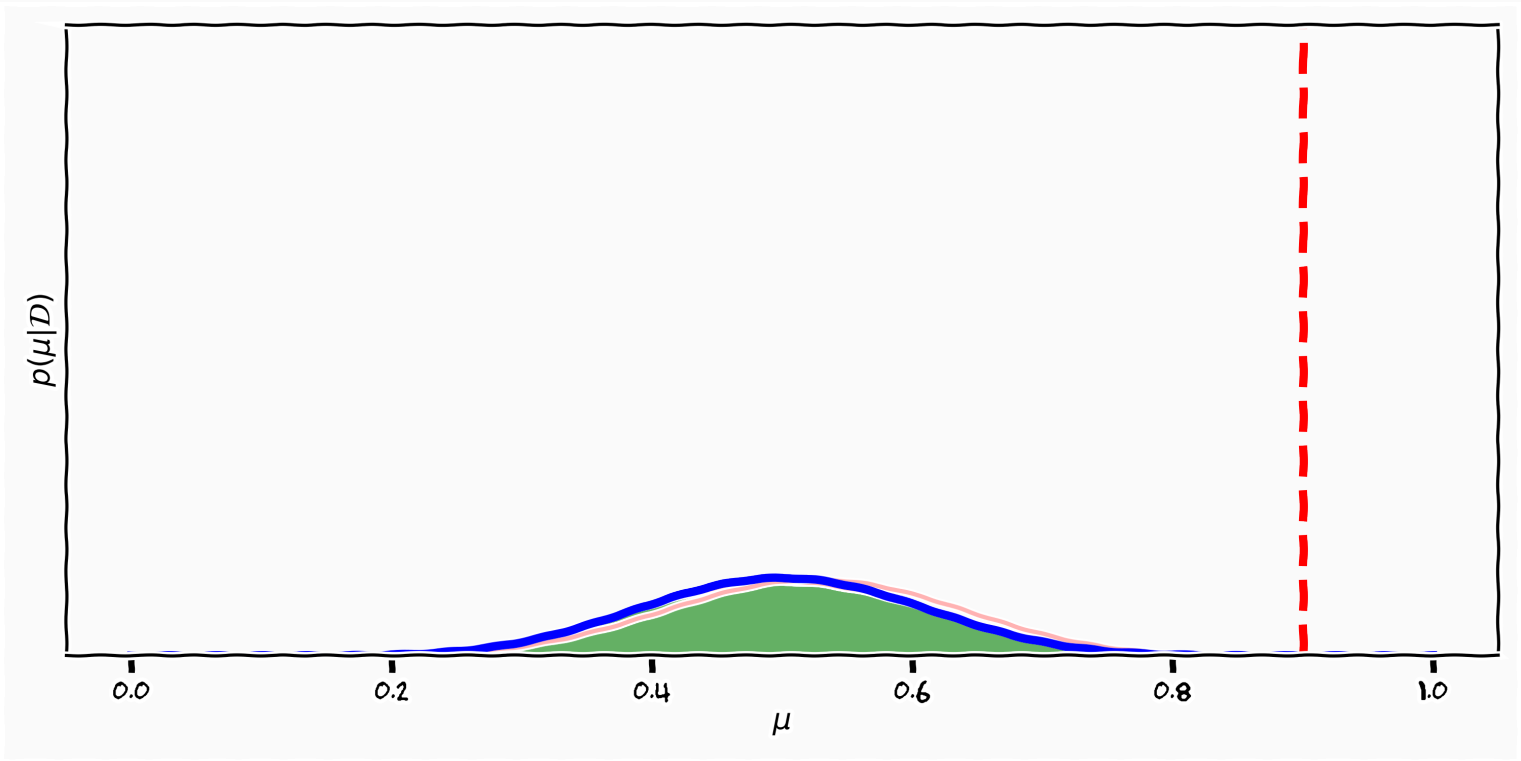
Experiments



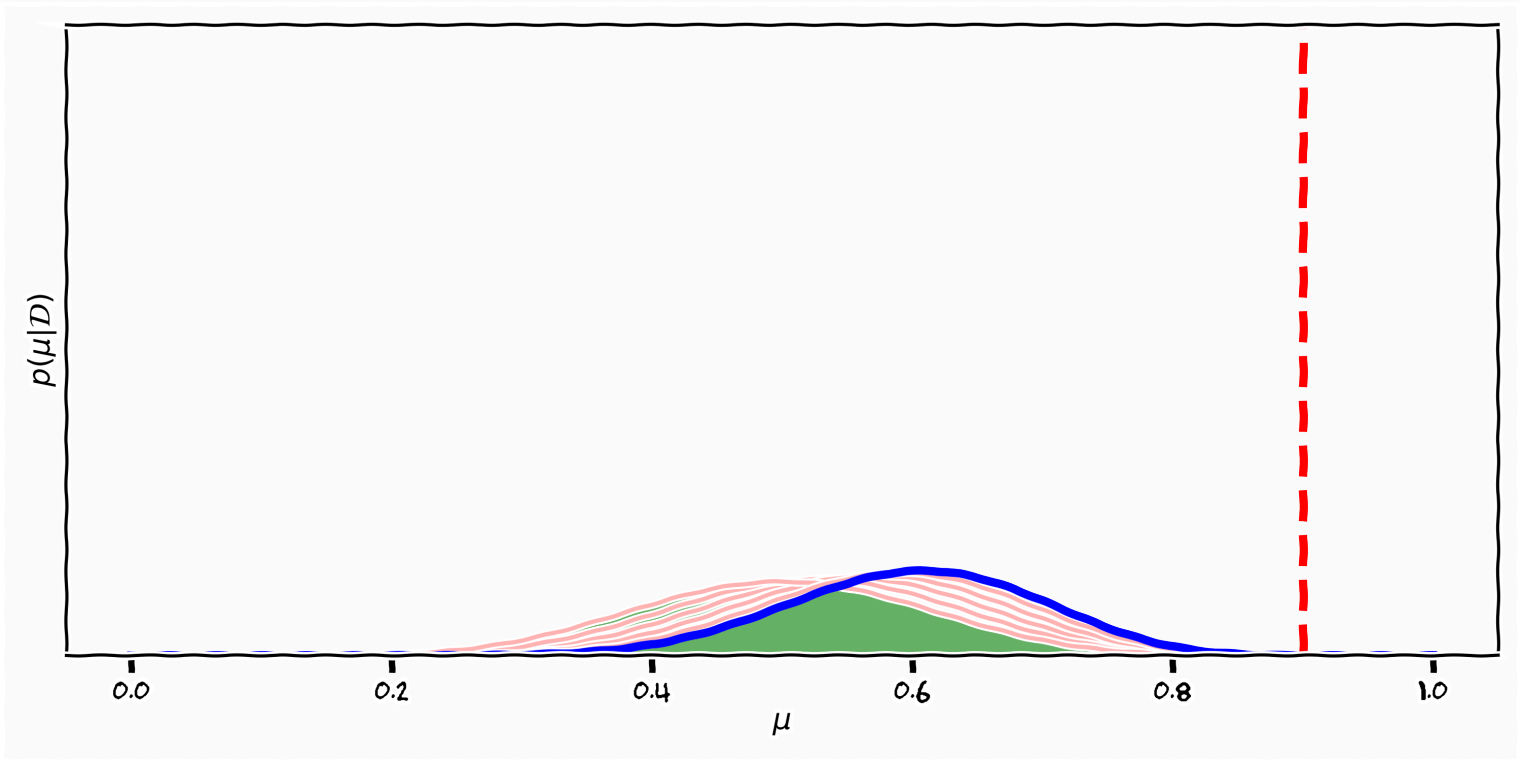
Experiments



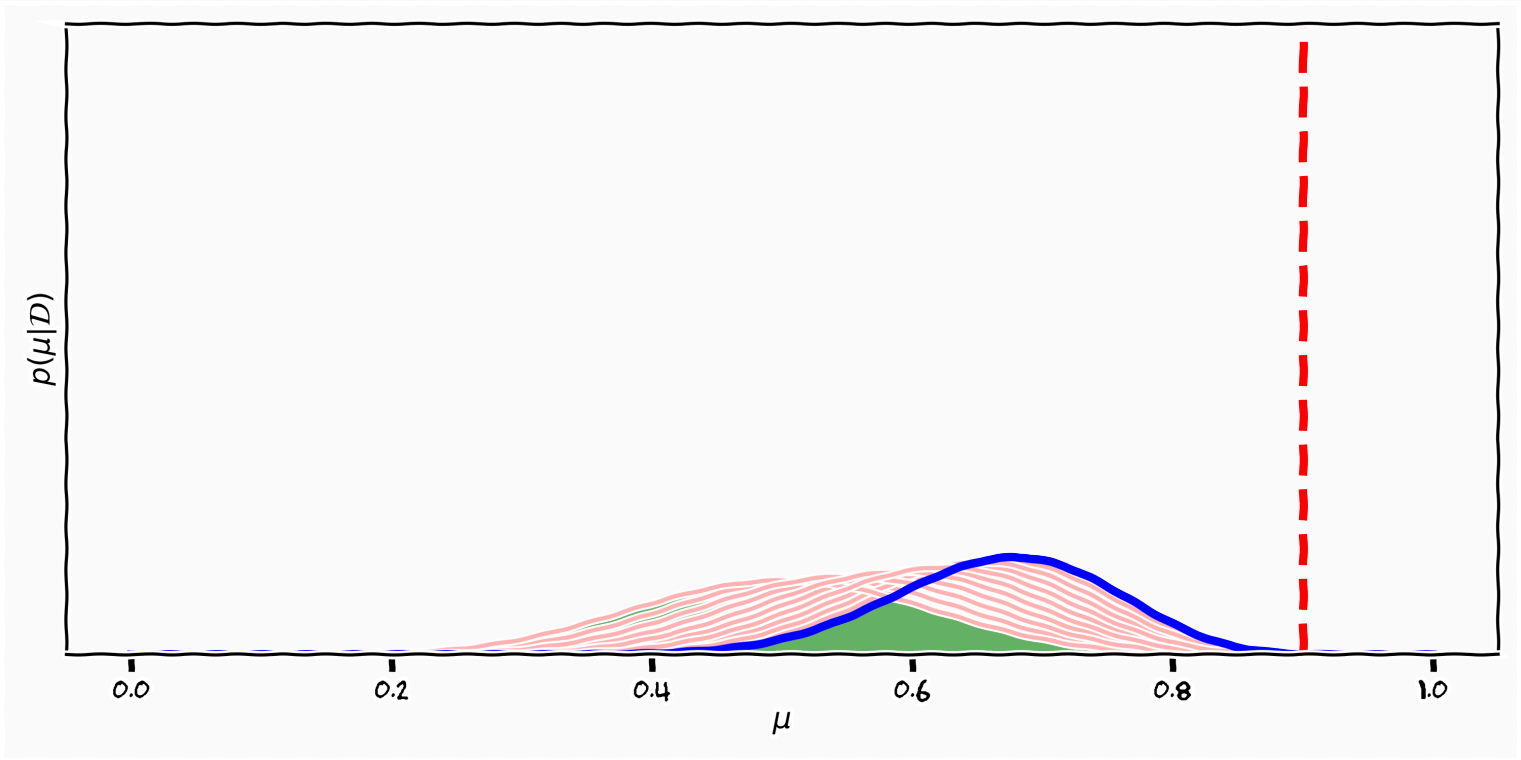
Experiments



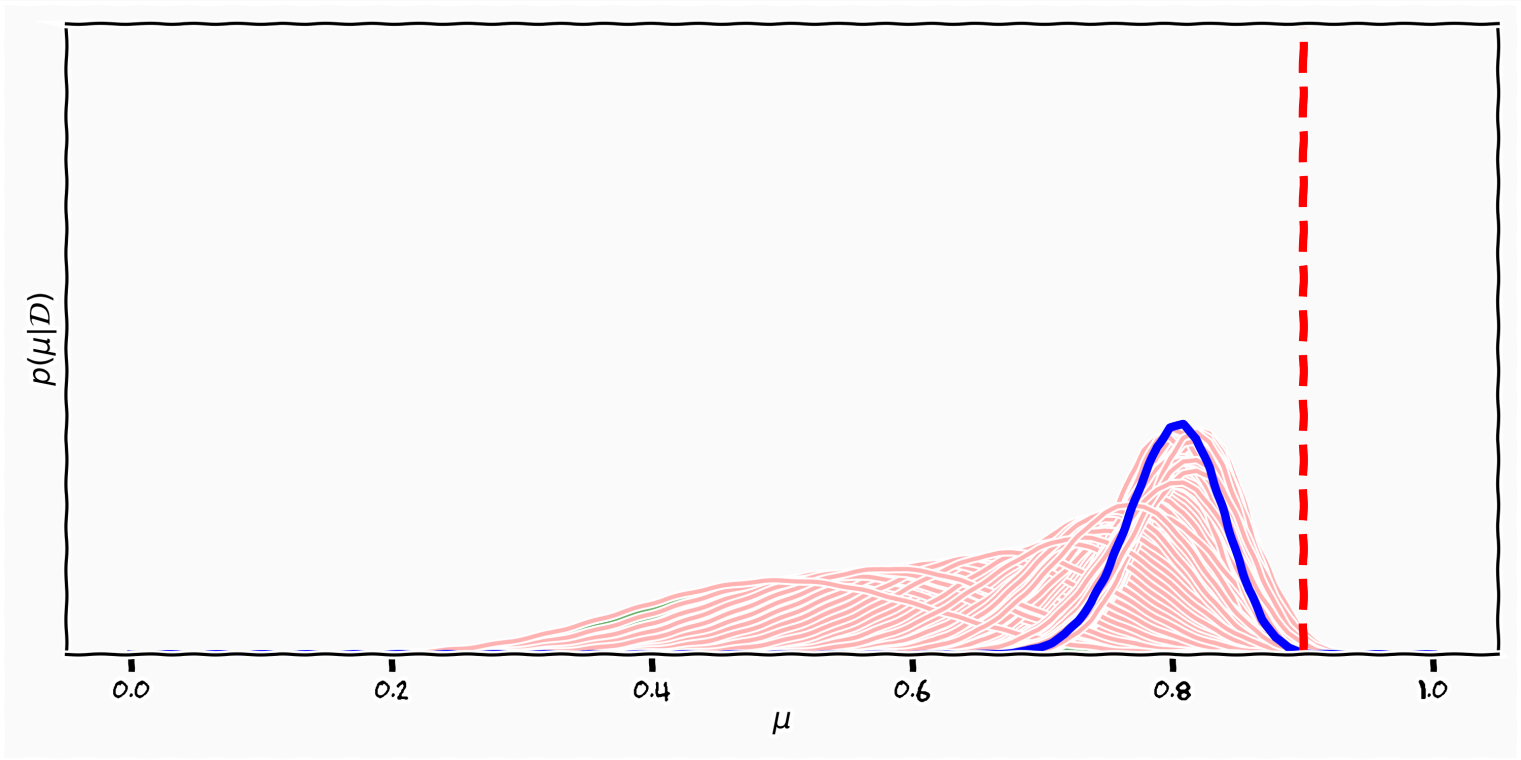
Experiments



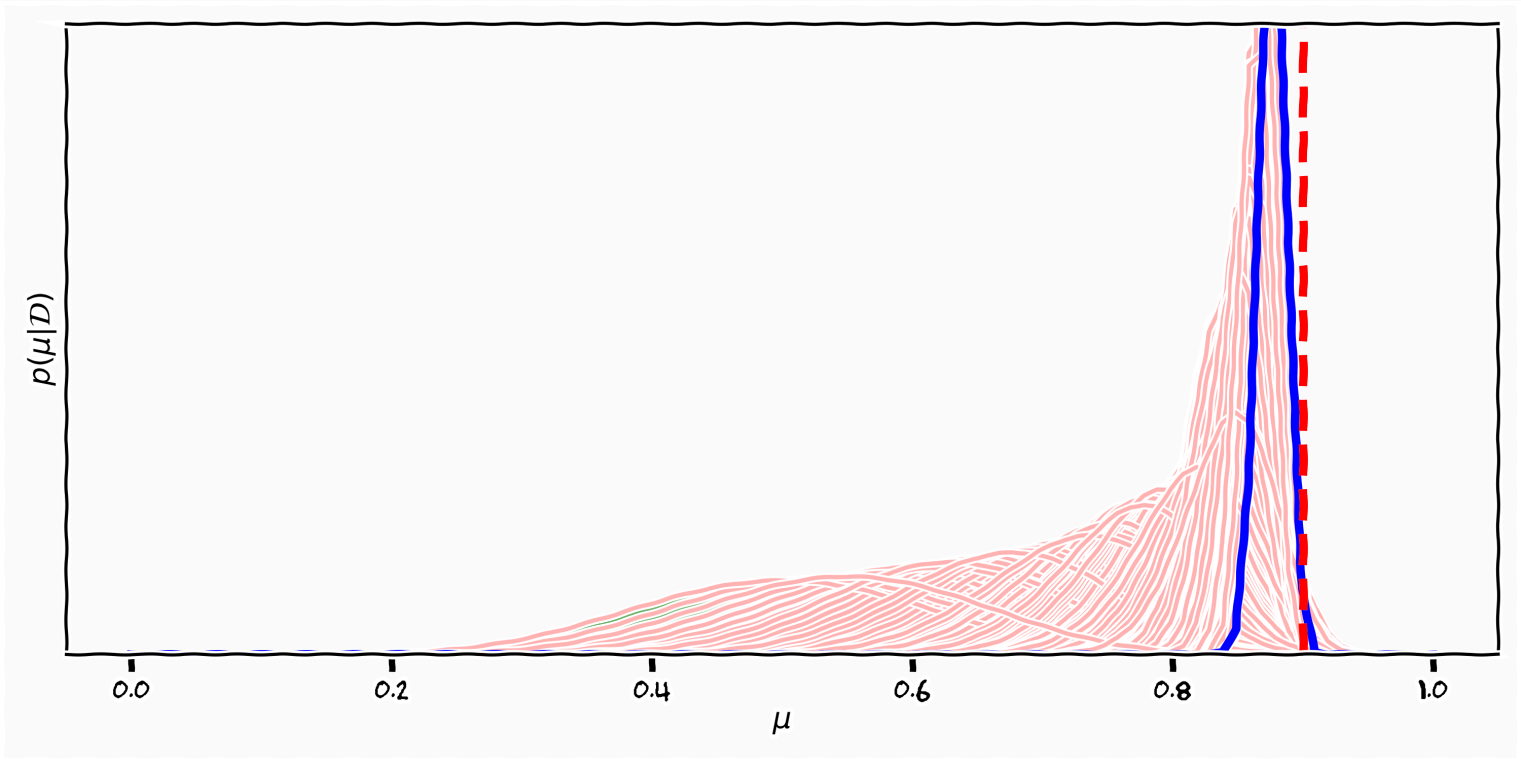
Experiments



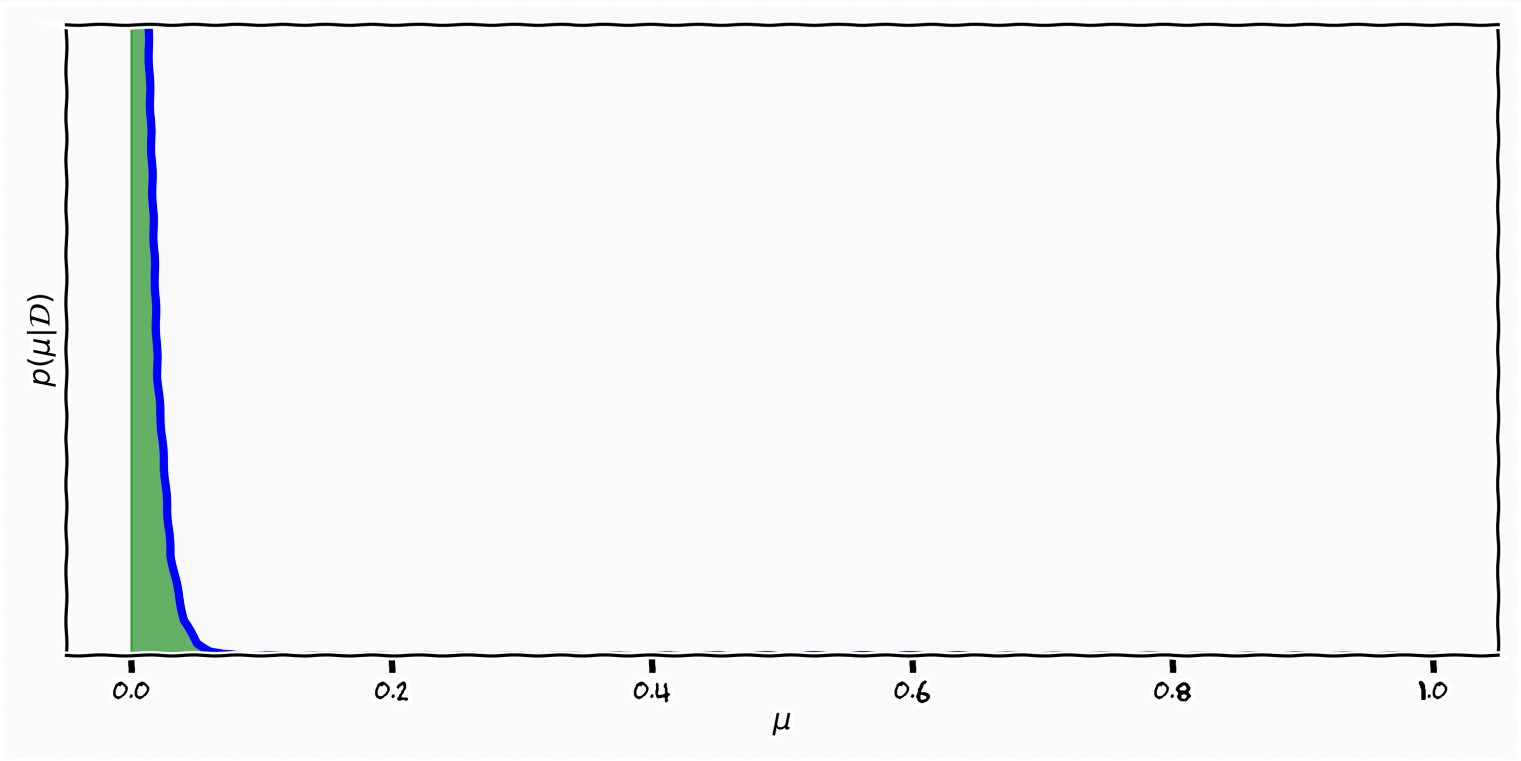
Experiments



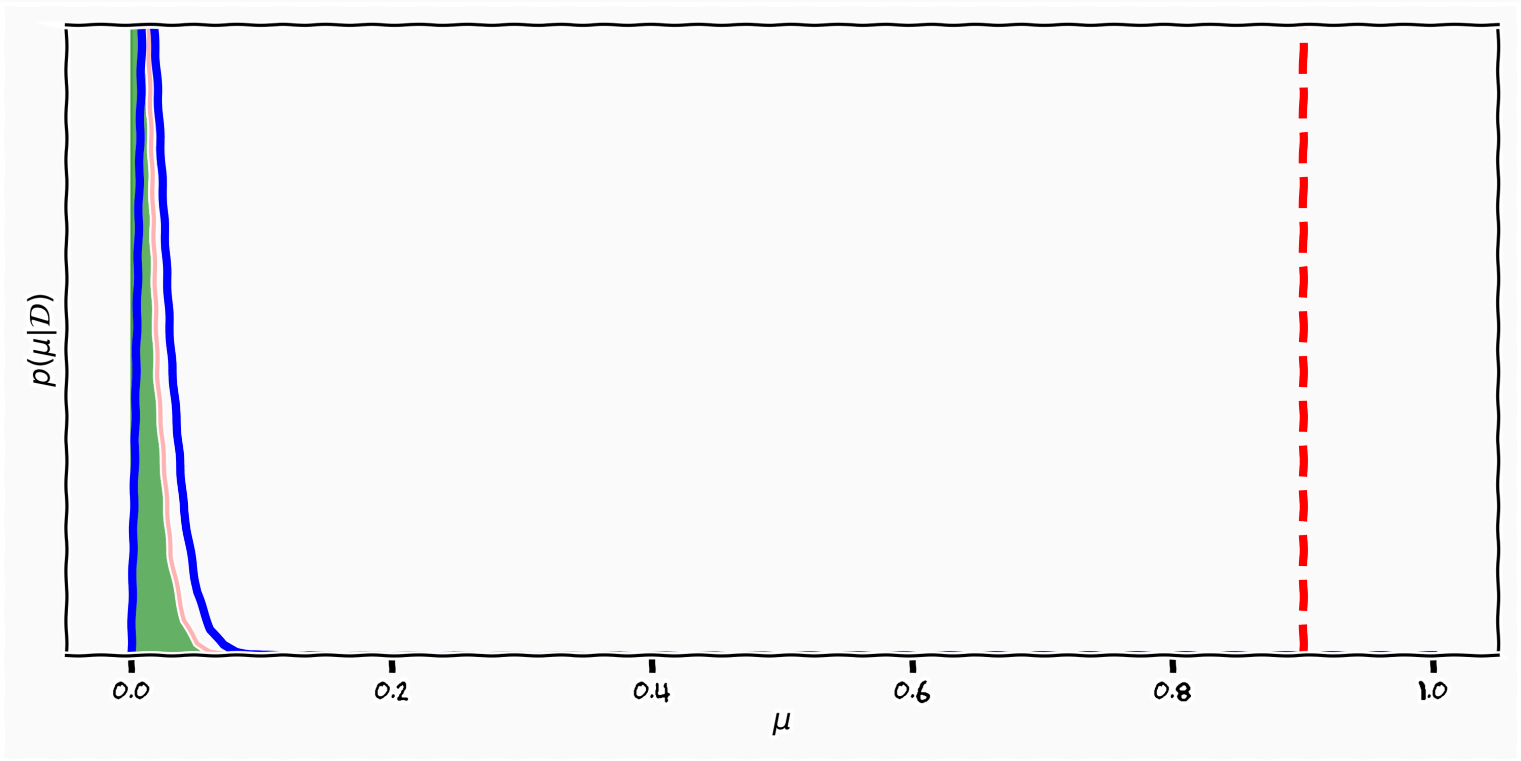
Experiments



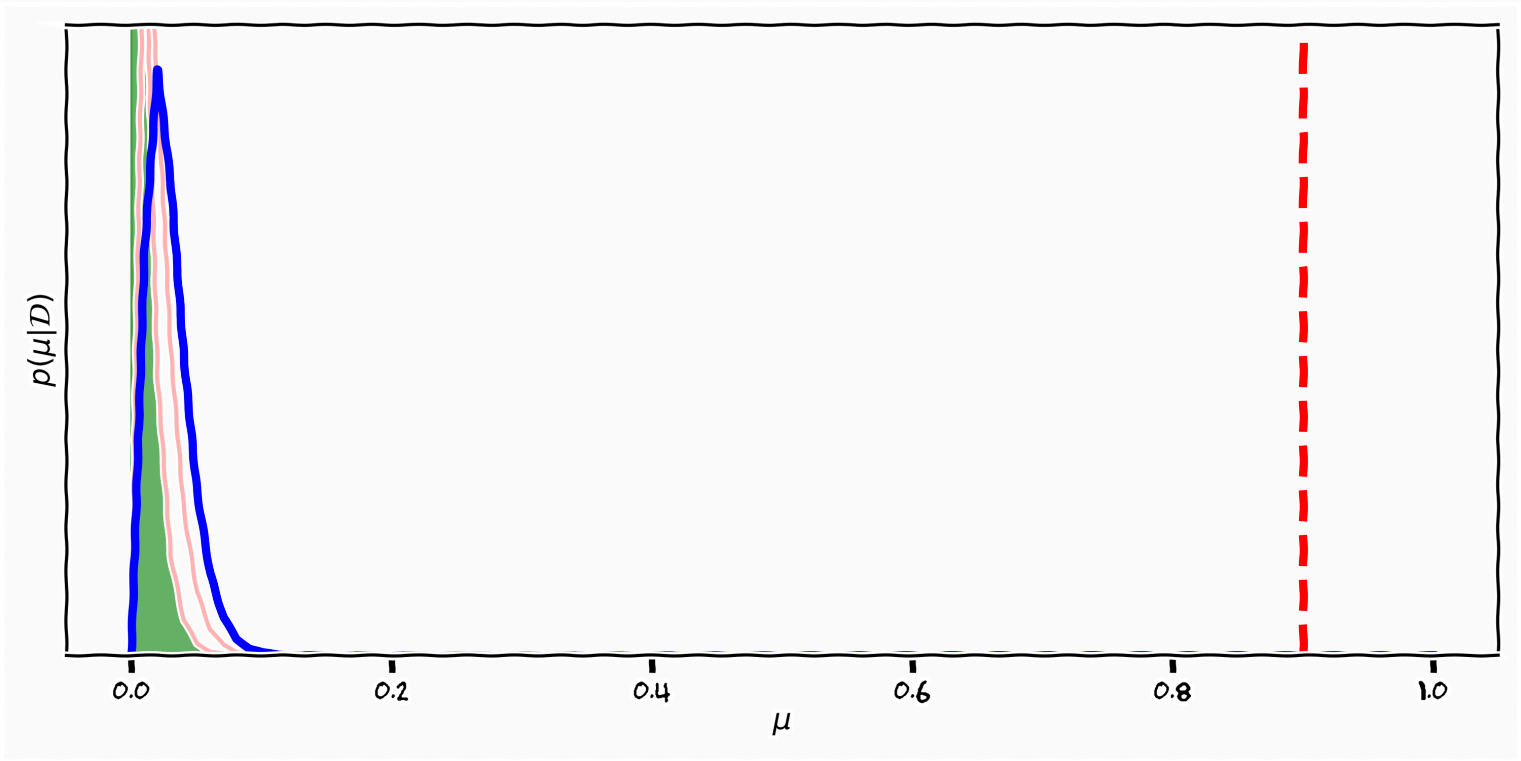
Experiments



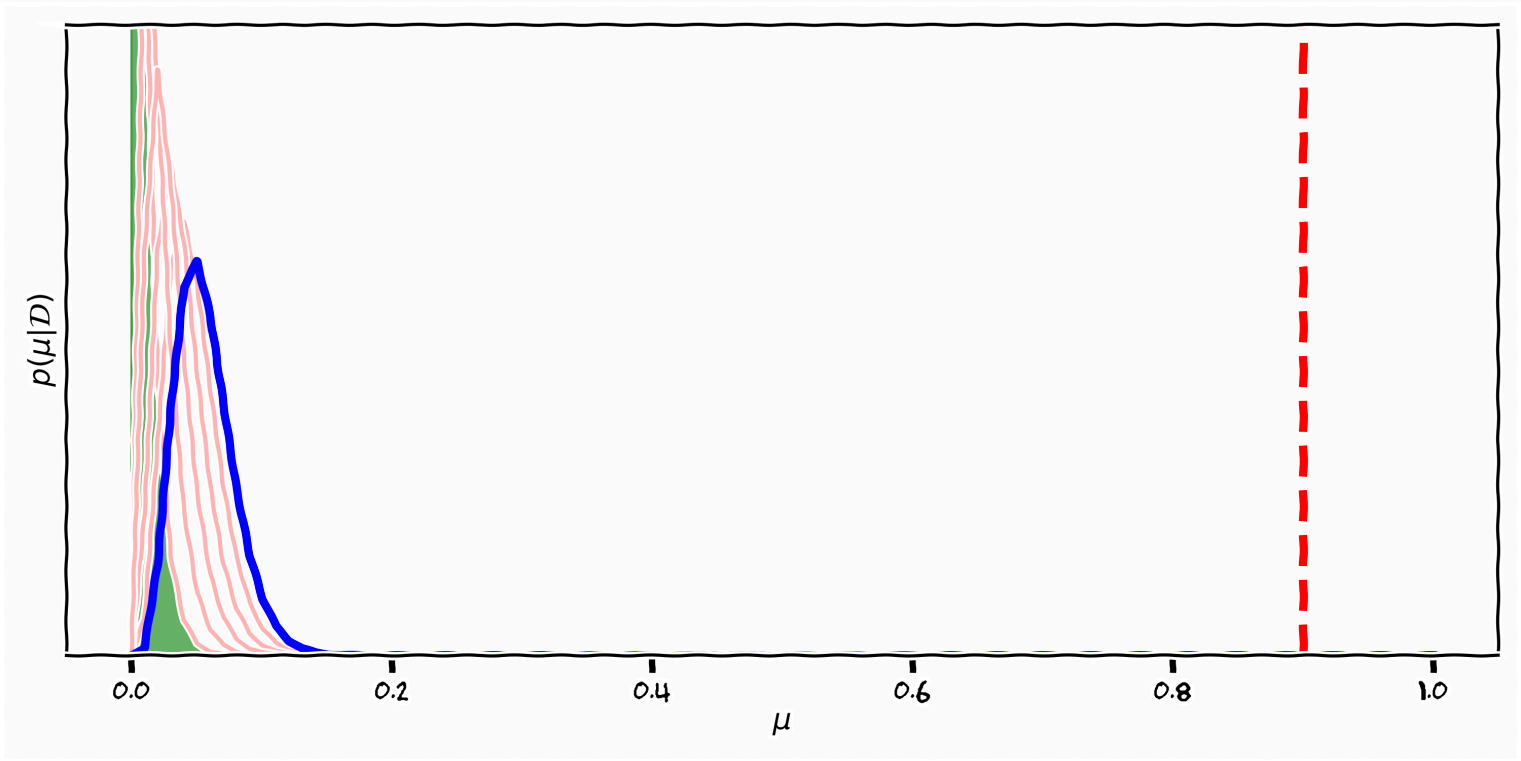
Experiments



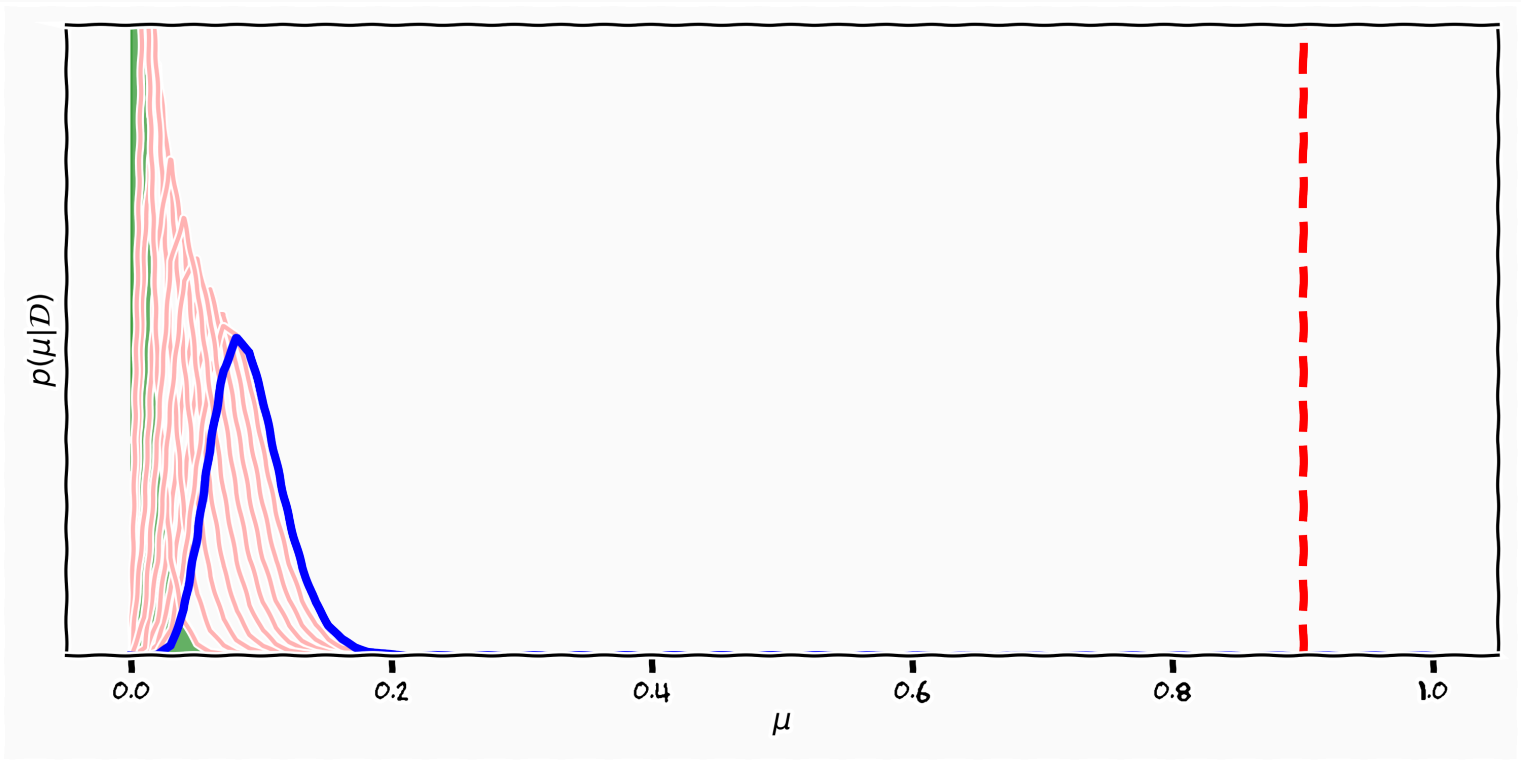
Experiments



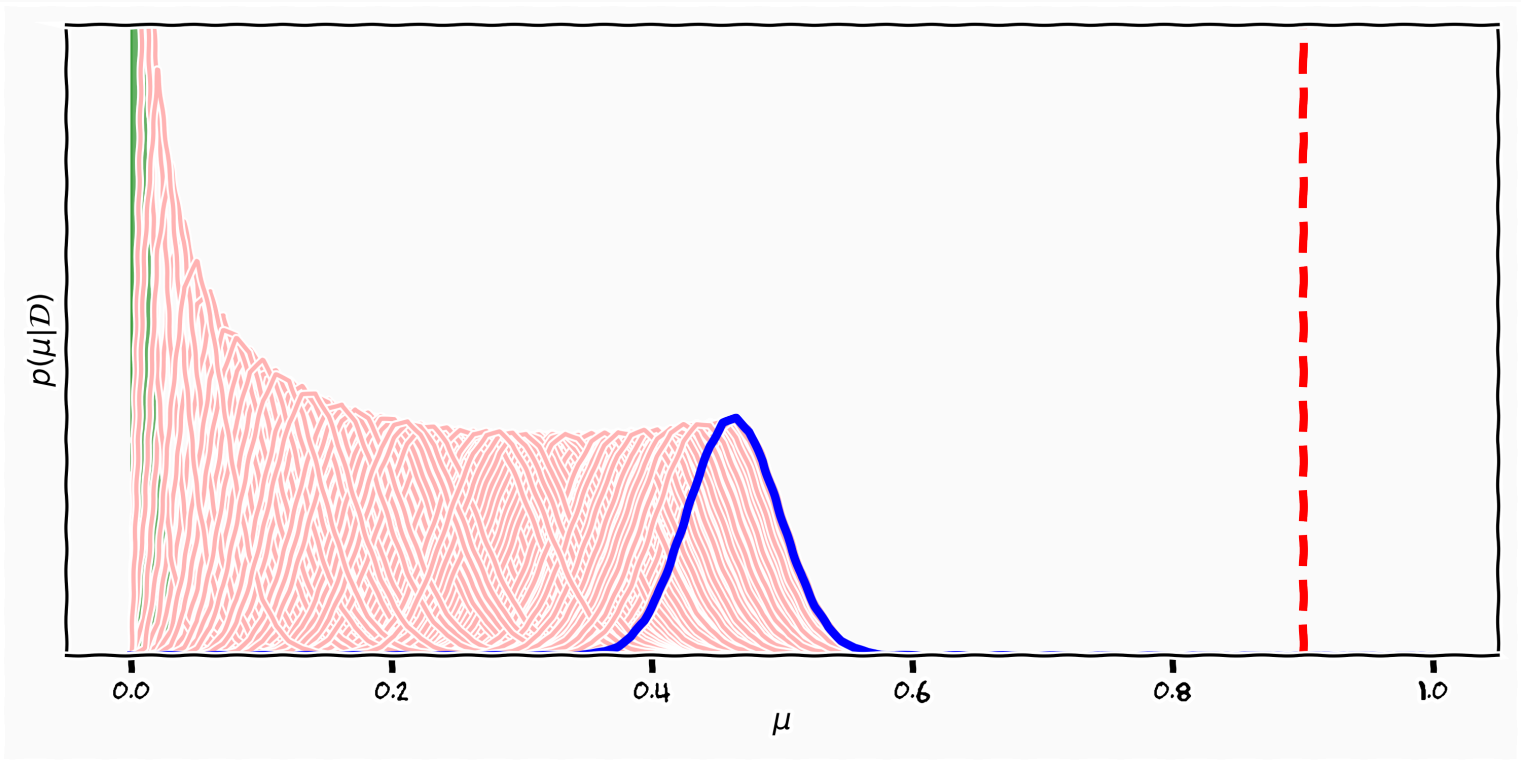
Experiments



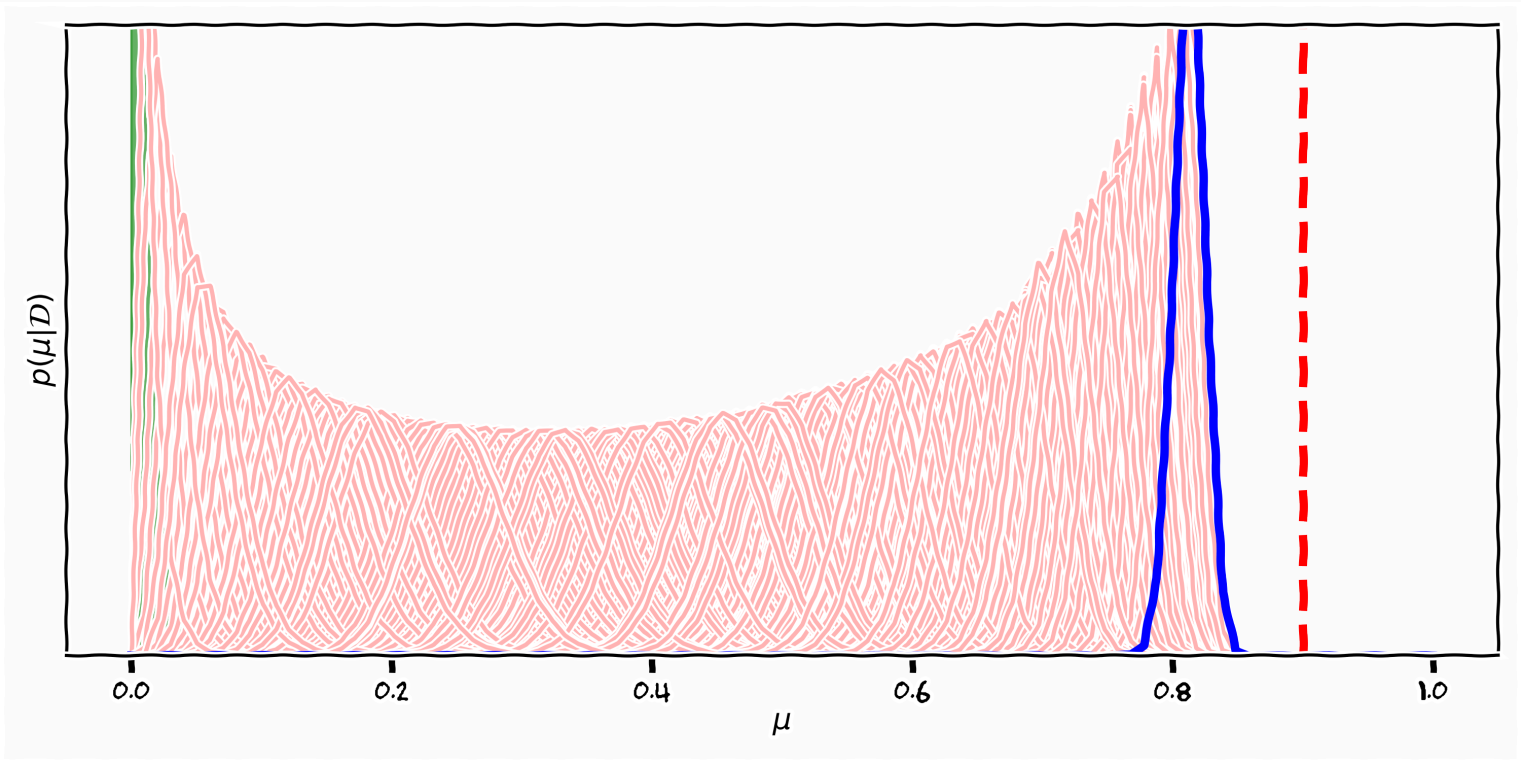
Experiments



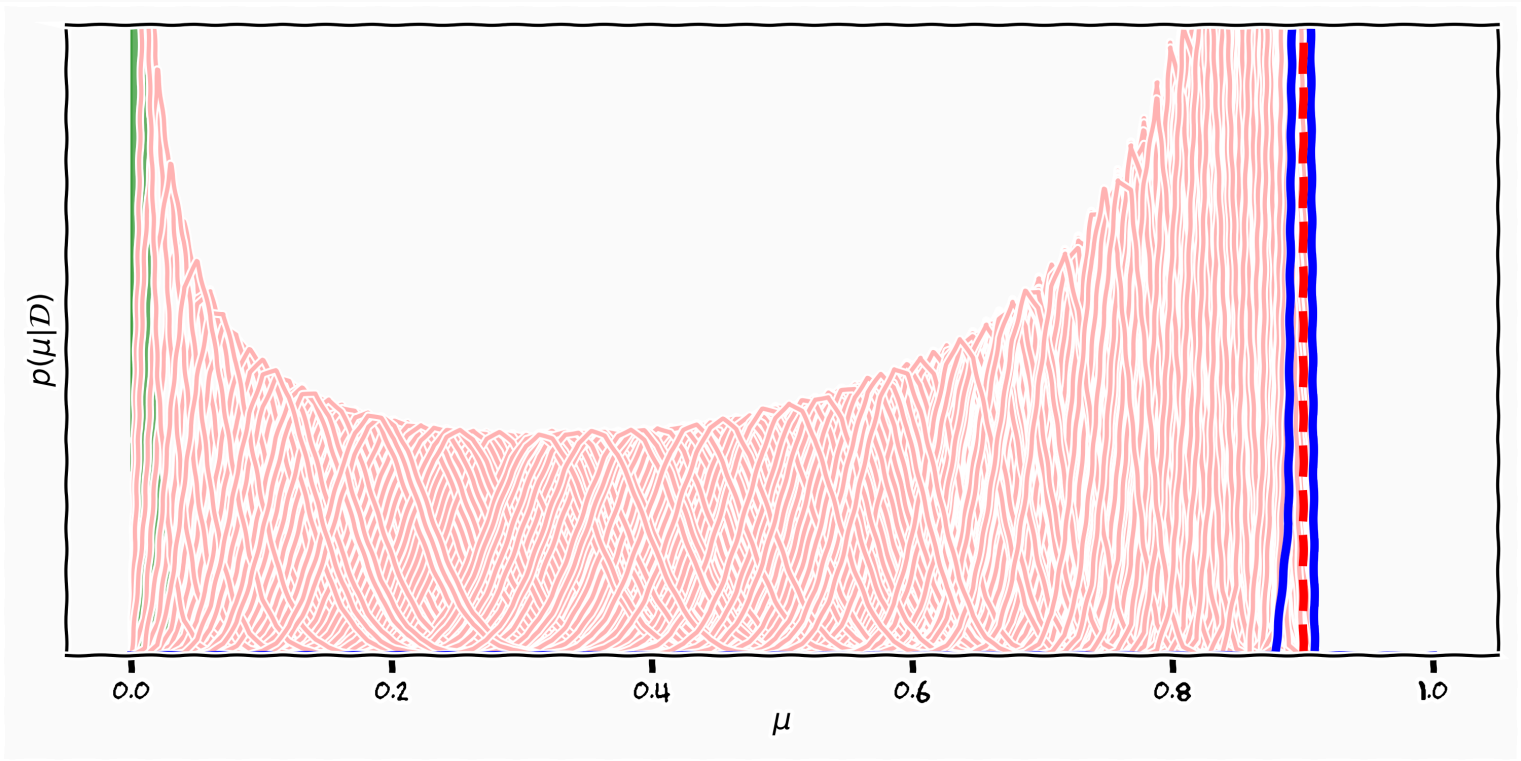
Experiments



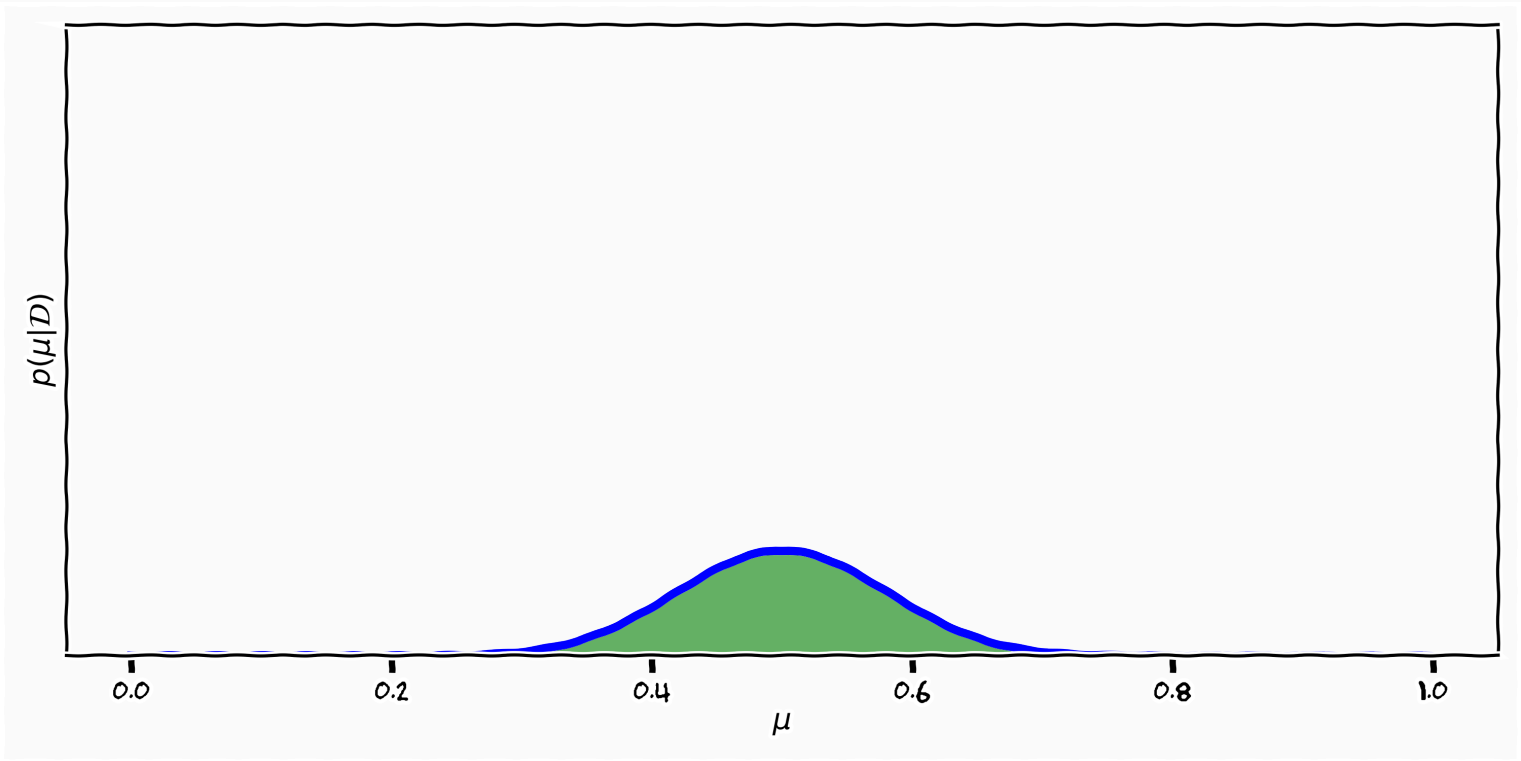
Experiments



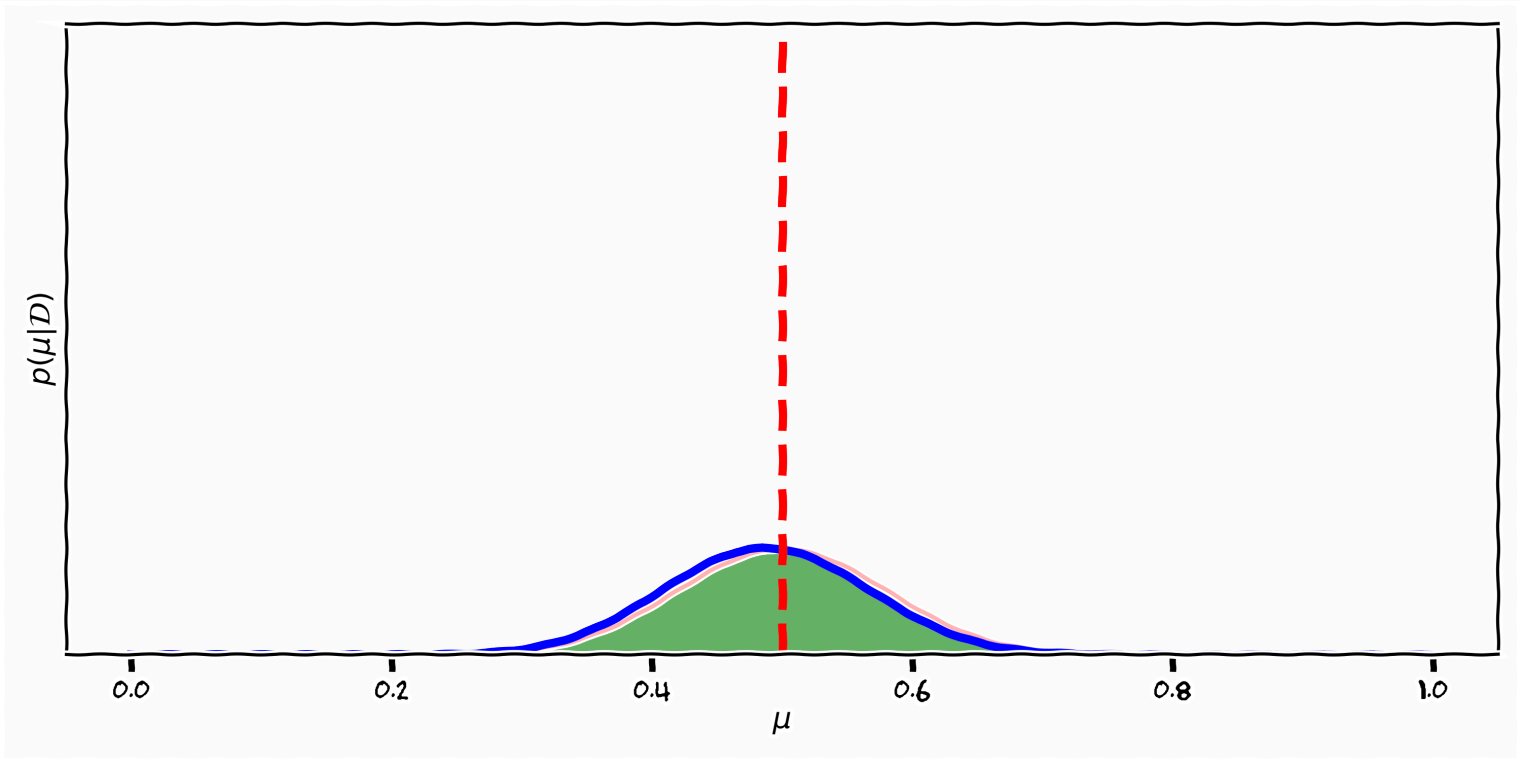
Experiments



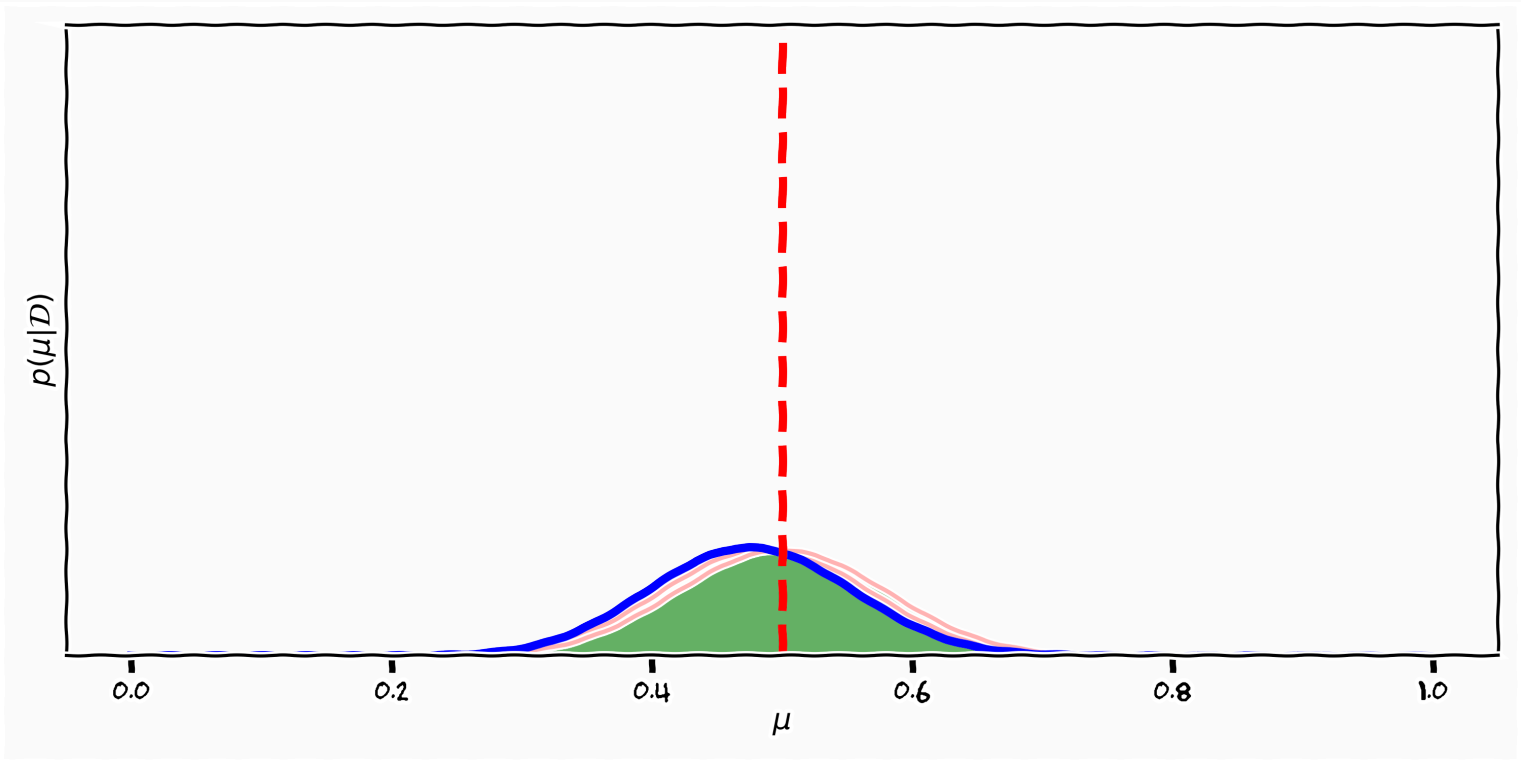
Experiments



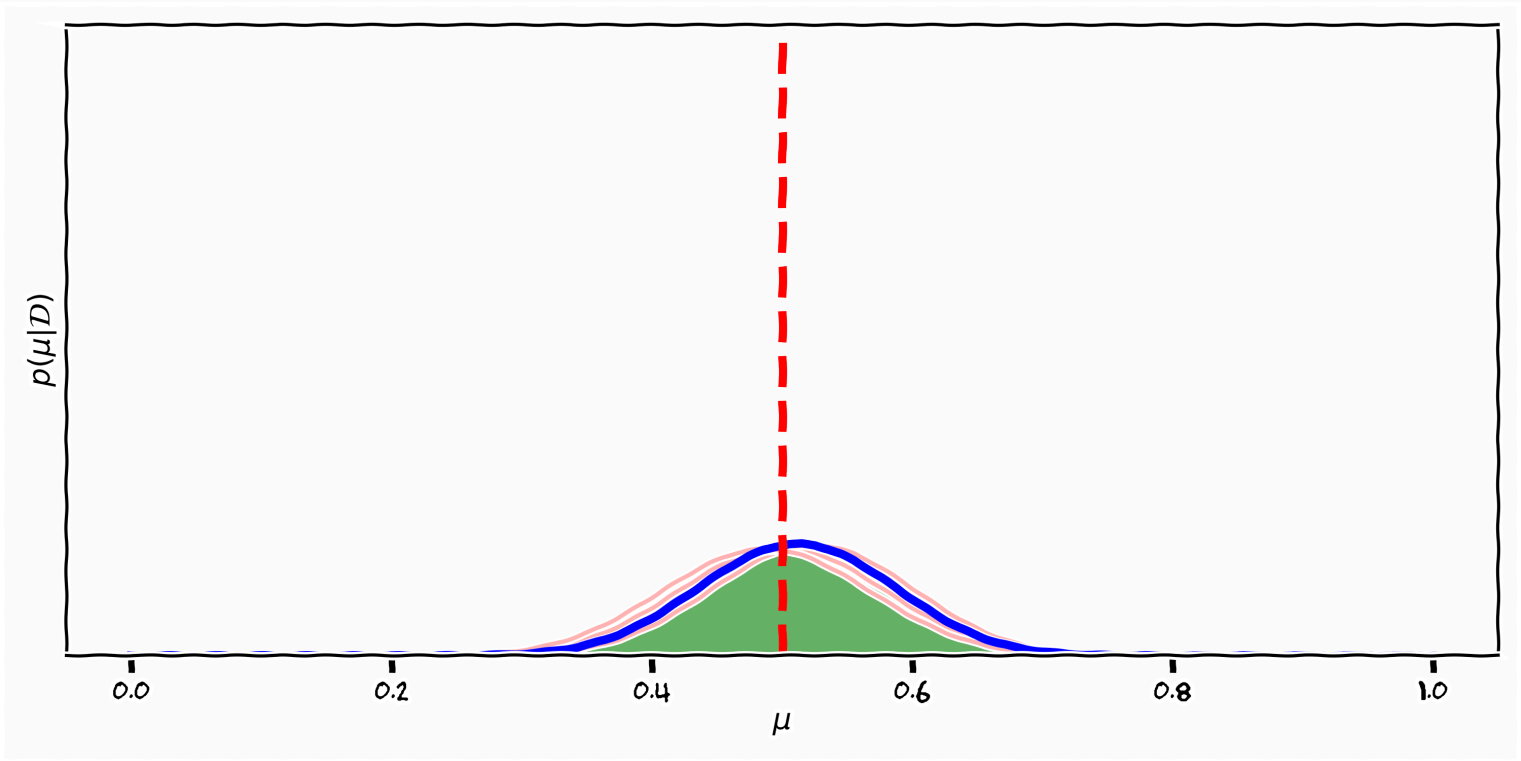
Experiments



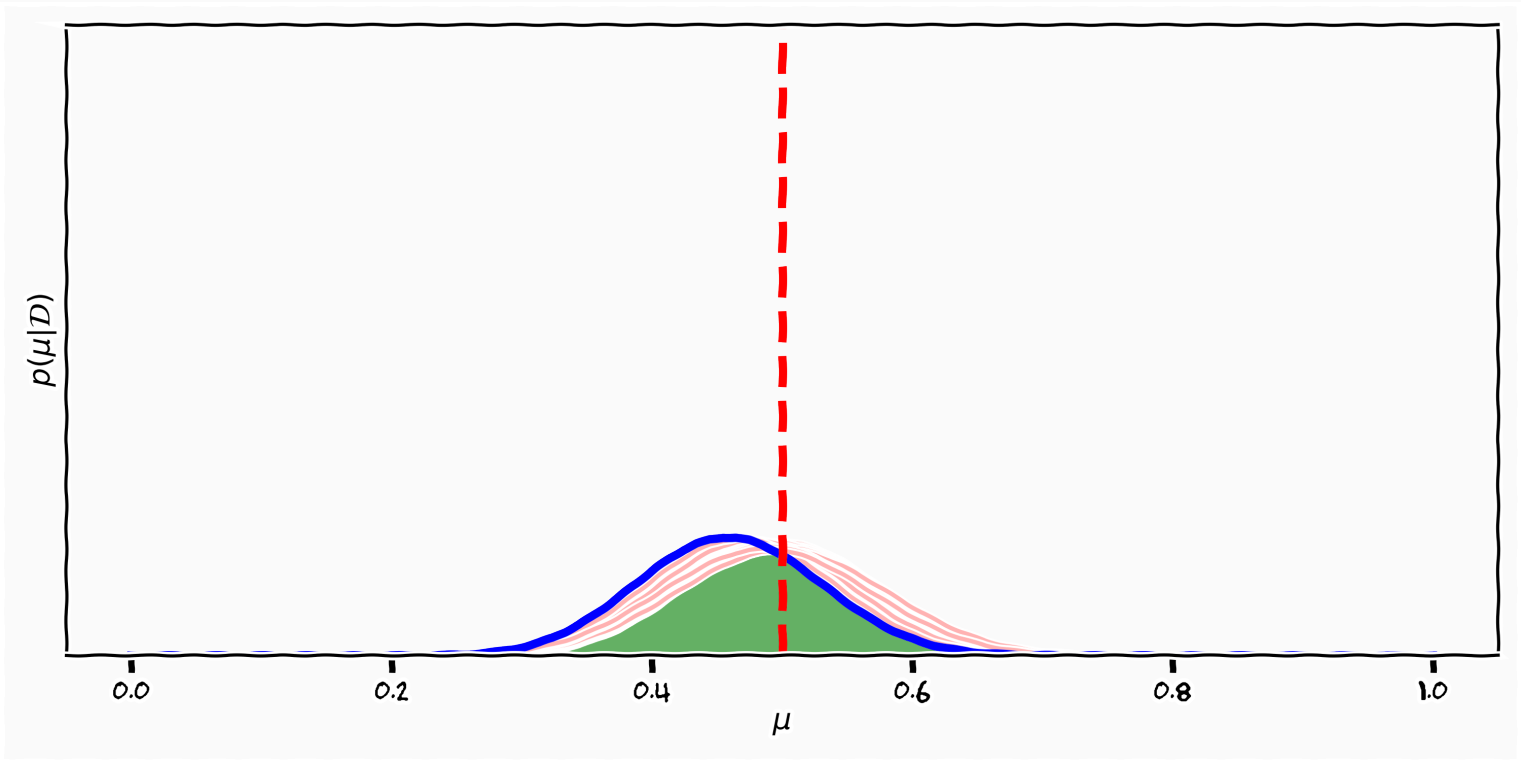
Experiments



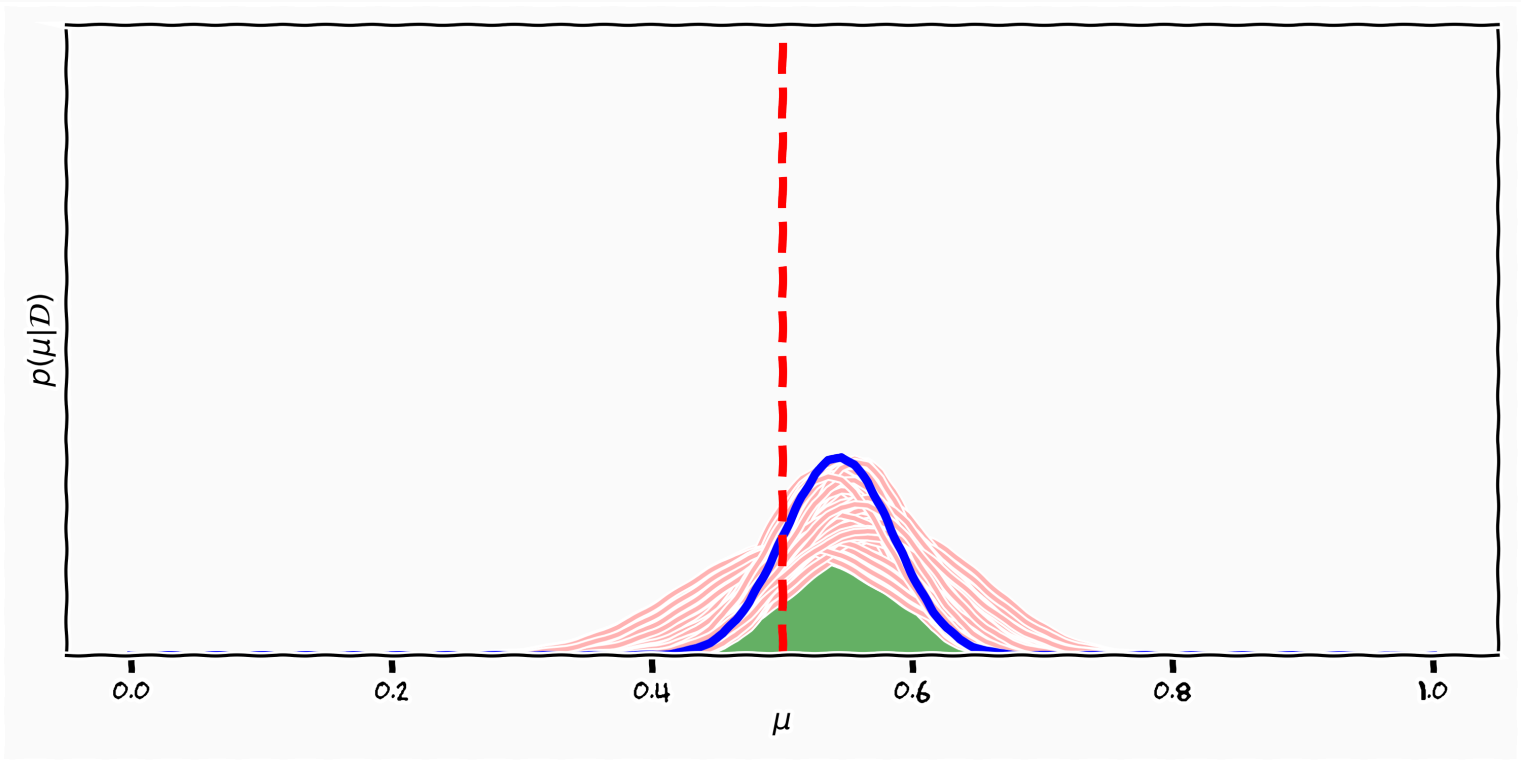
Experiments



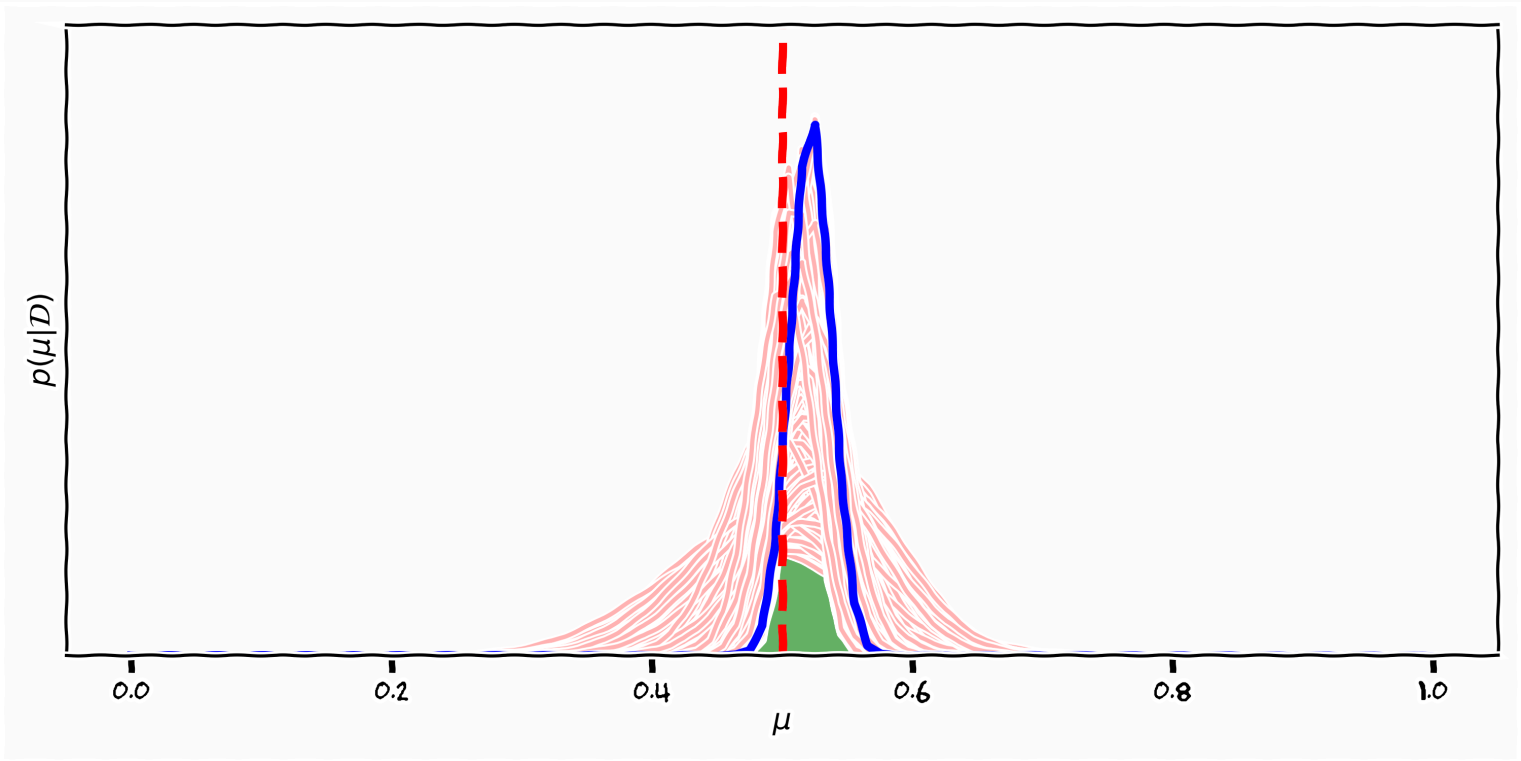
Experiments



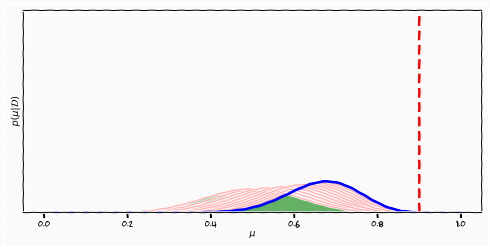
Experiments

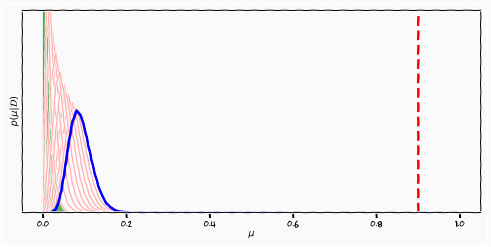


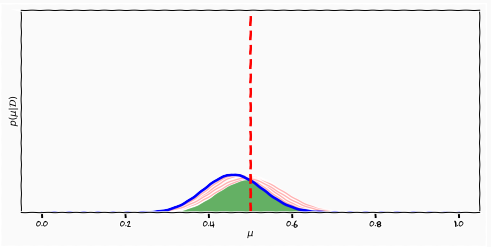
Experiments



Summary



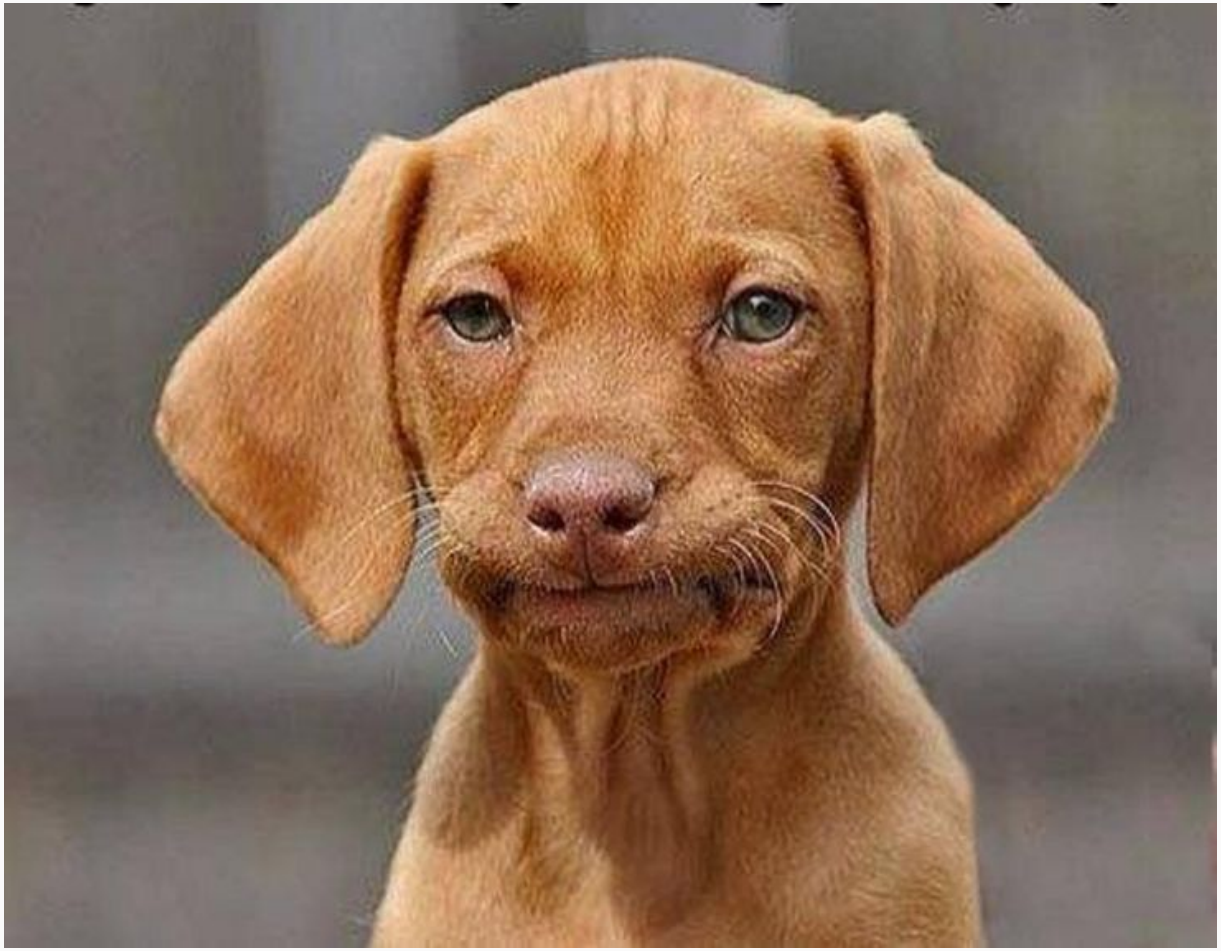




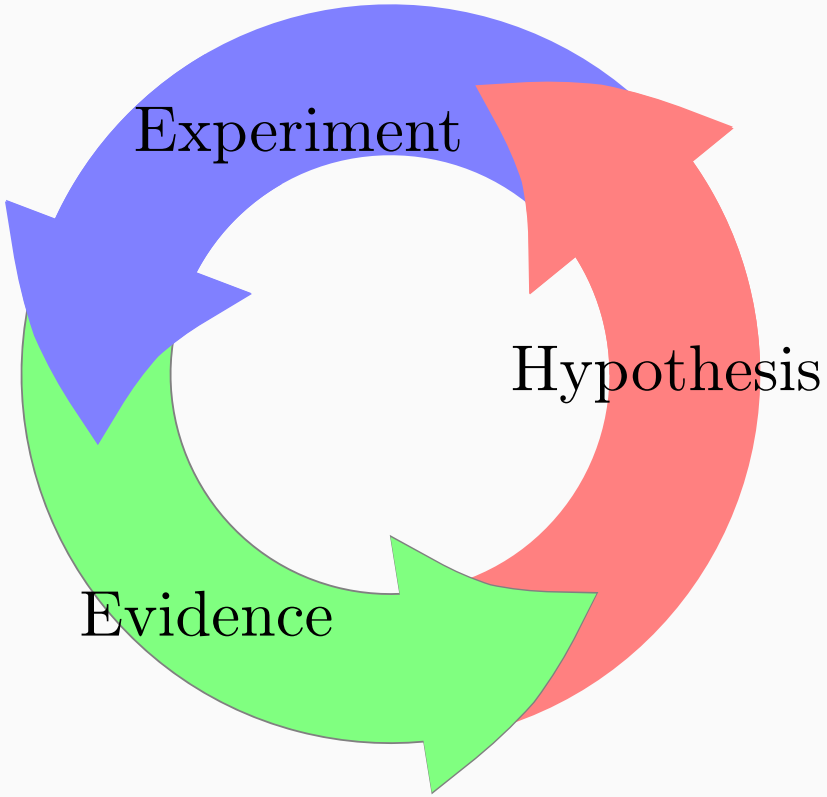
Summary

- It is impossible to differentiate solutions without assumptions
- Our "solution" can only be interpreted in the light of our assumptions
- Posterior inference "balances" the information in the data with the assumptions

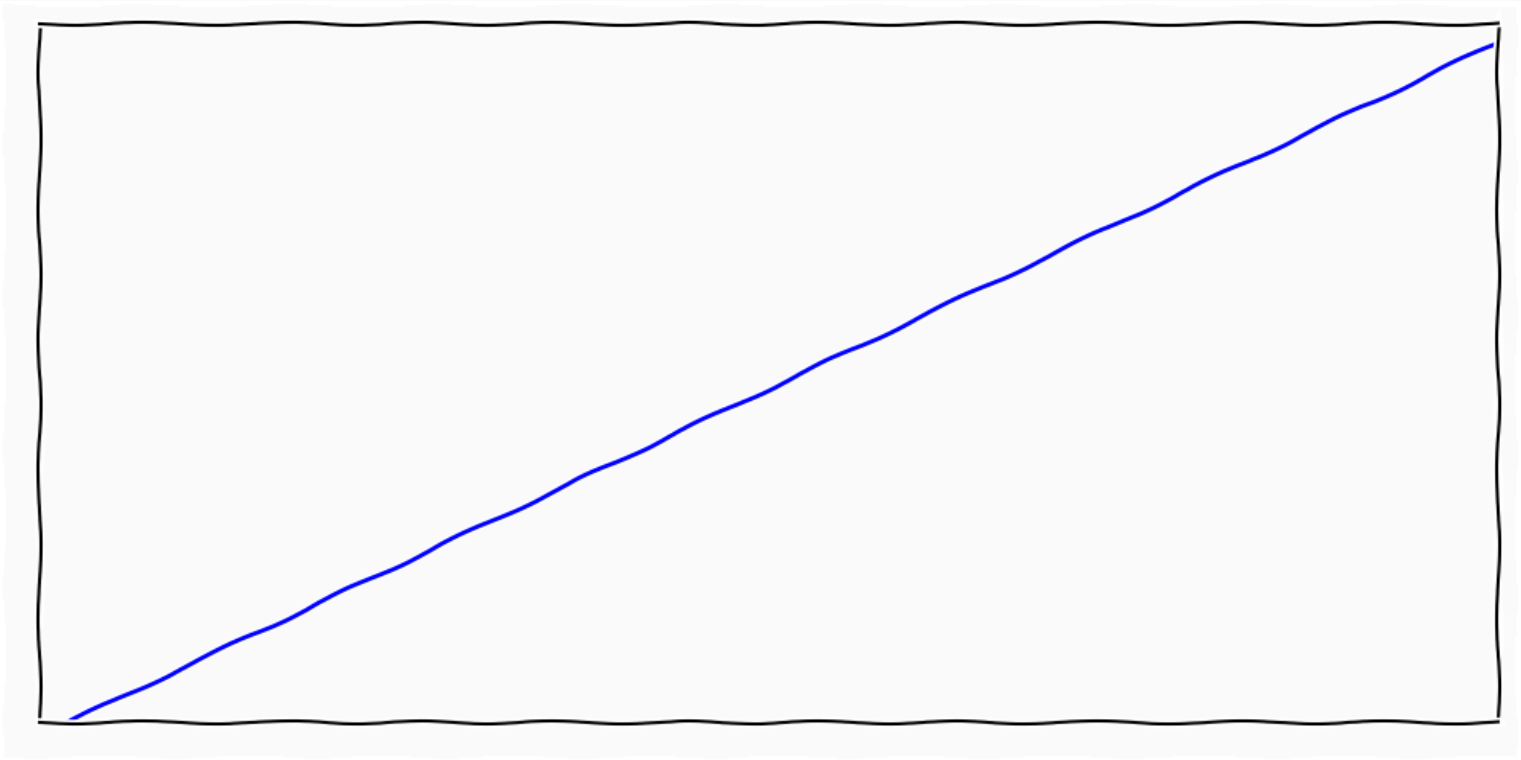
Reflections



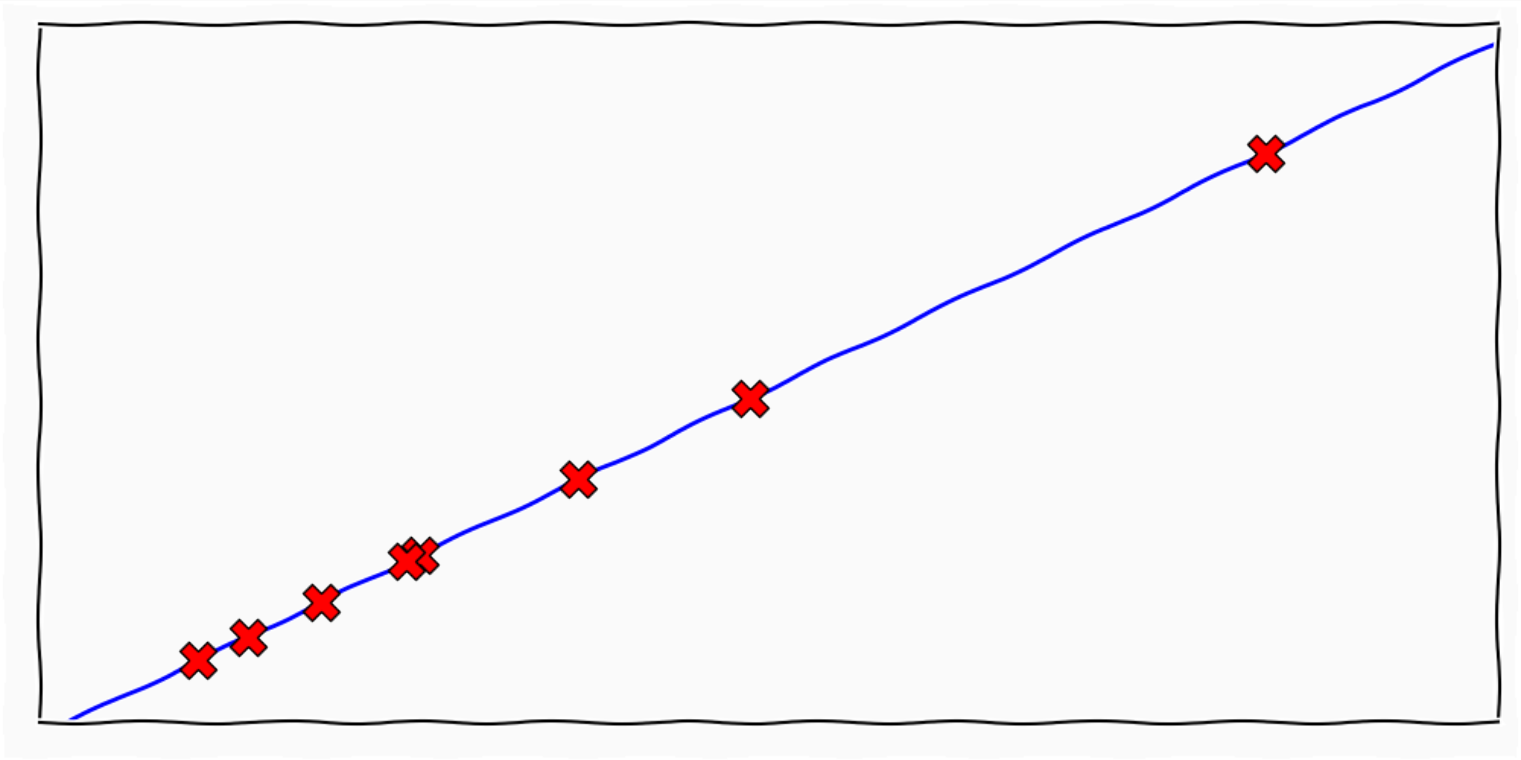
Science?



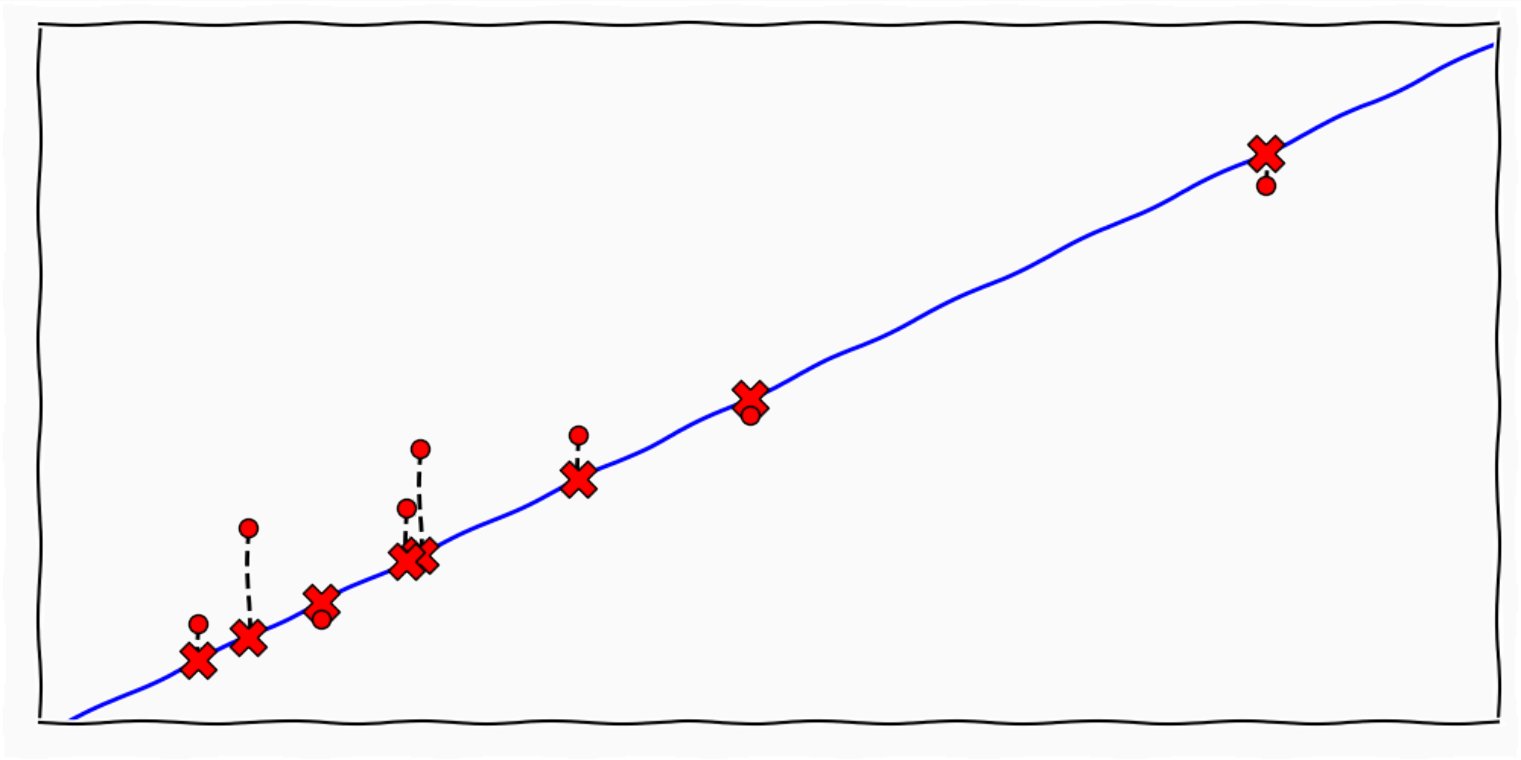
Linear Regression: Modelling



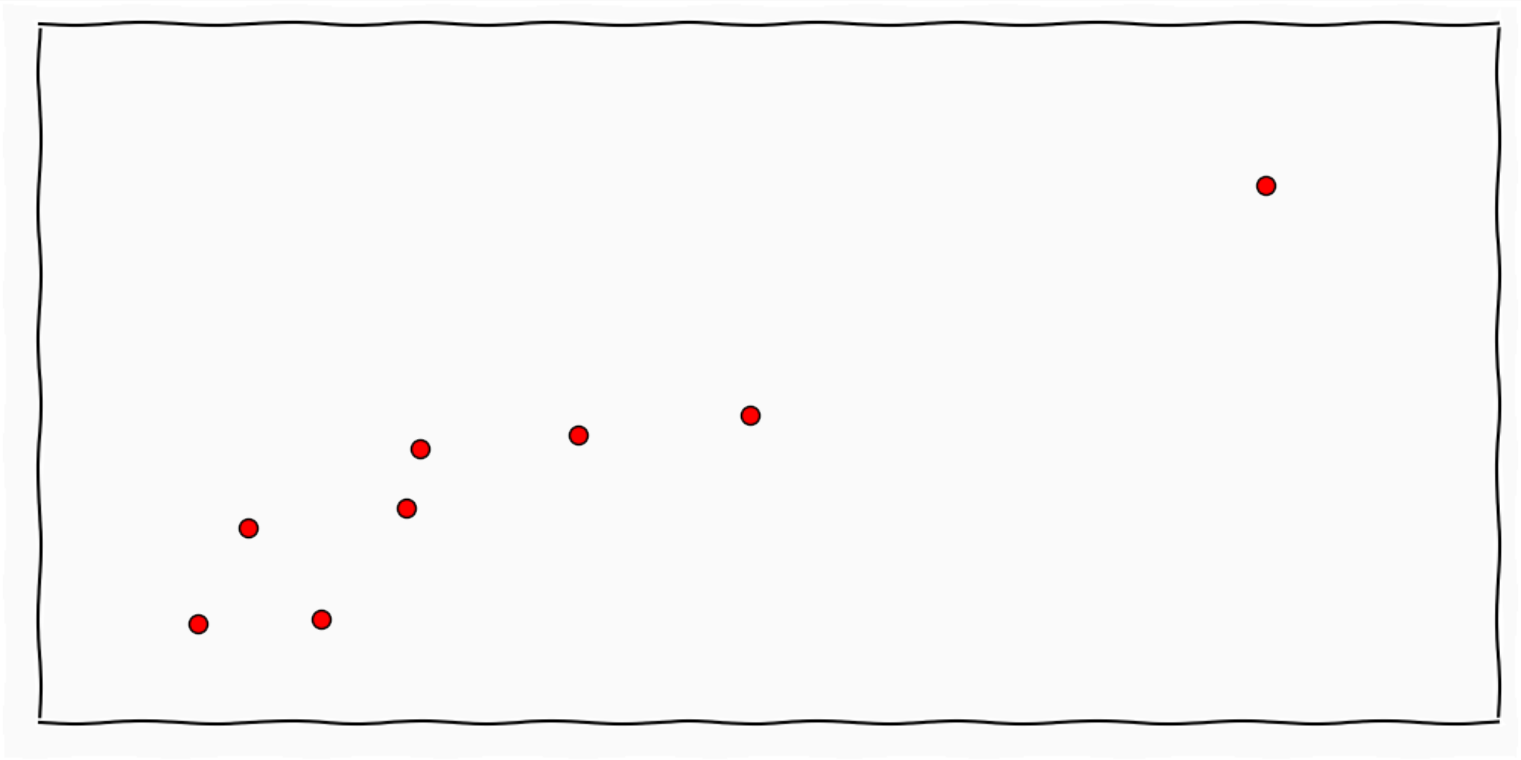
Linear Regression: Modelling



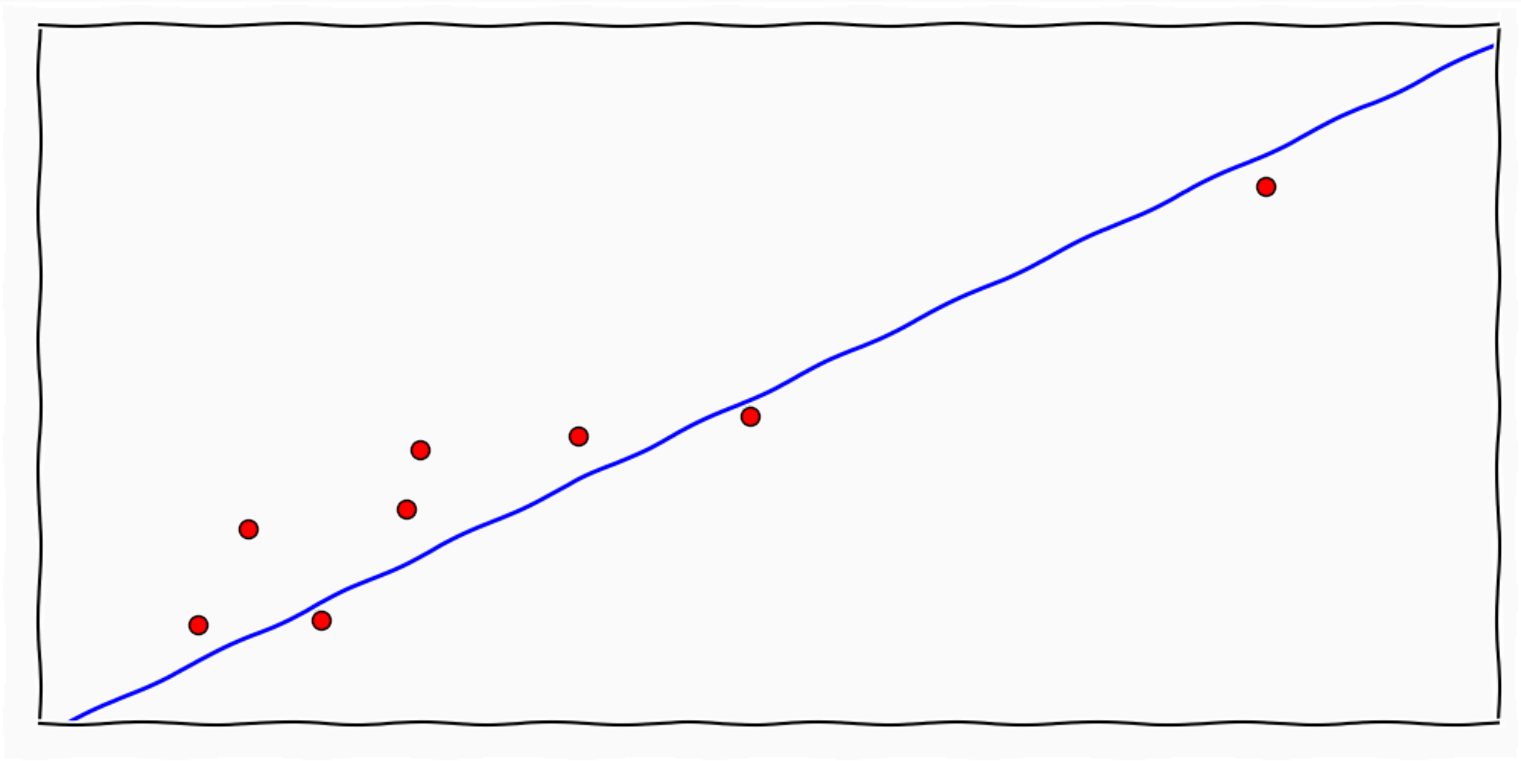
Linear Regression: Modelling



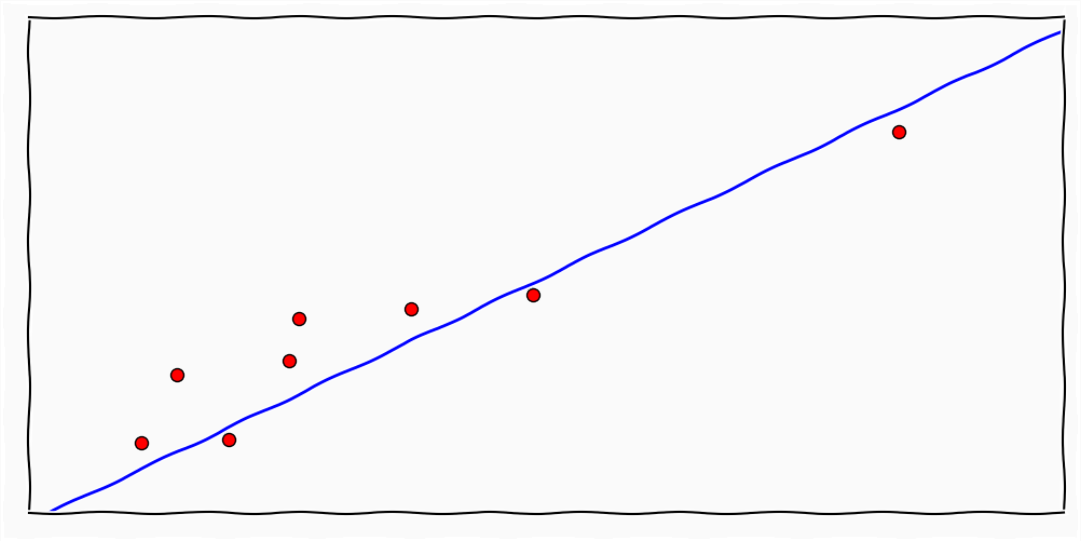
Linear Regression: Modelling



Linear Regression: Modelling

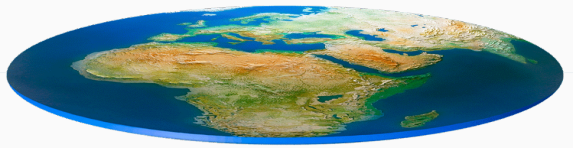


Linear Regression: Modelling



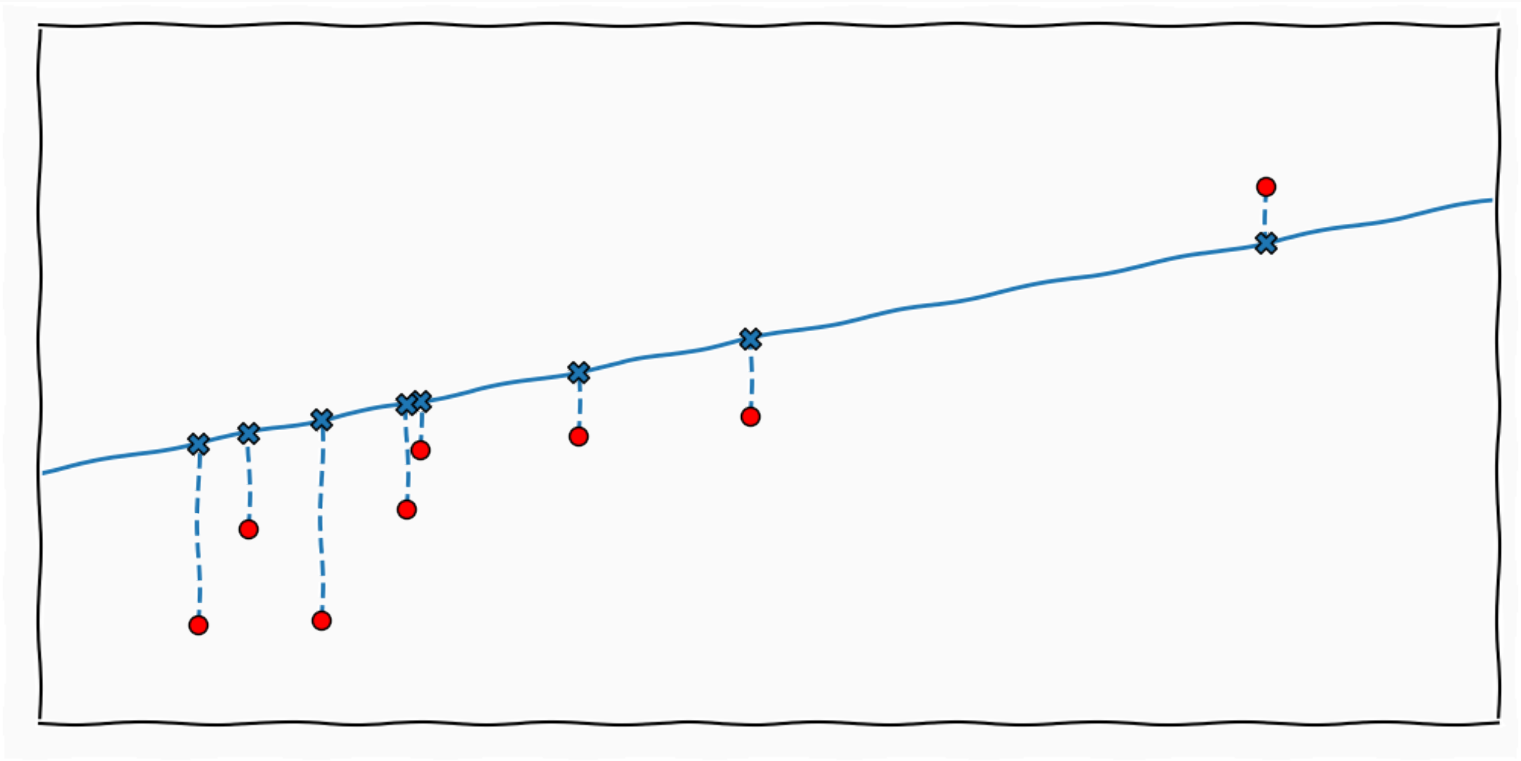
$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$$

Likelihood

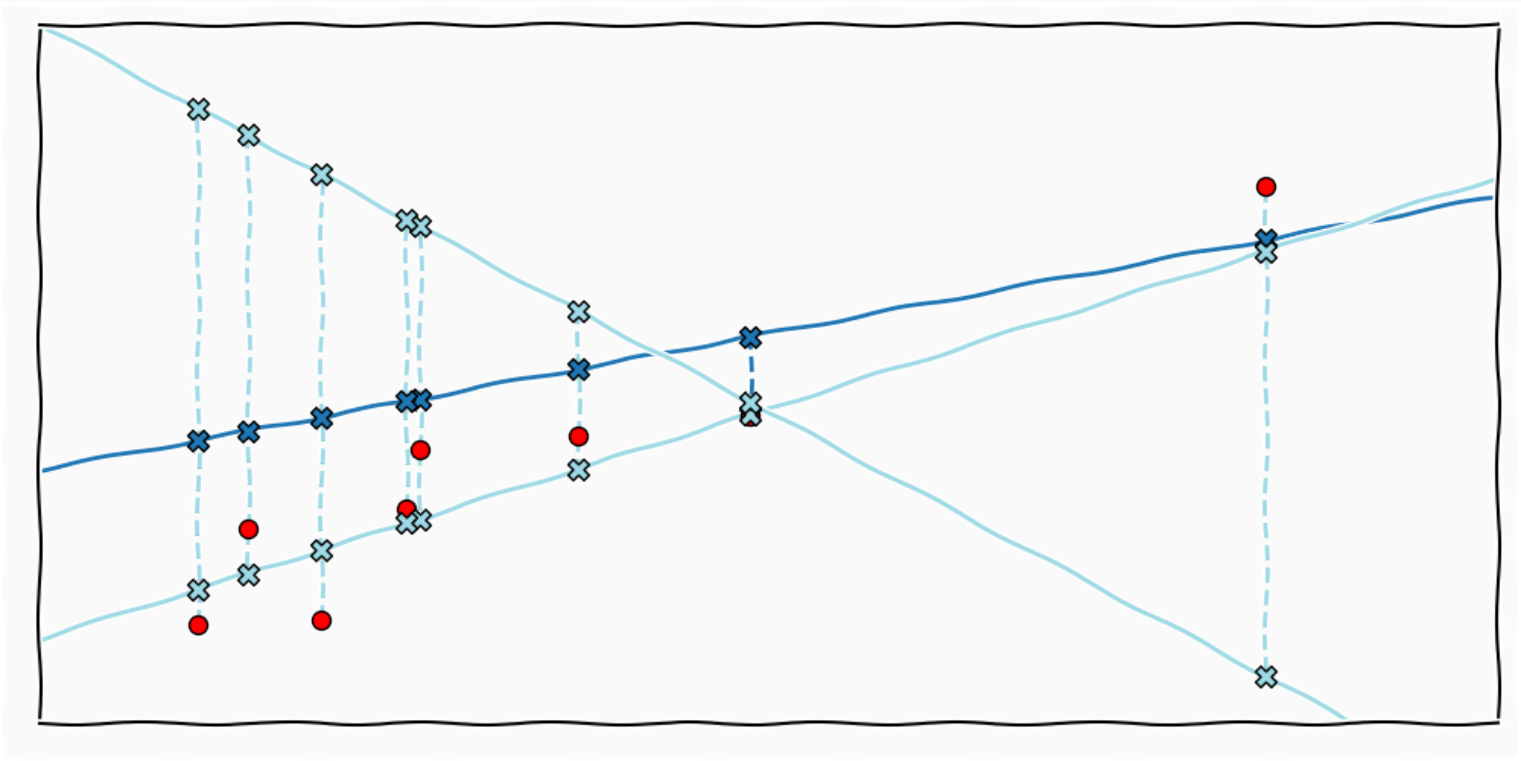


"Given the premise that the earth is flat, how supportive do I believe observations y are of this?"

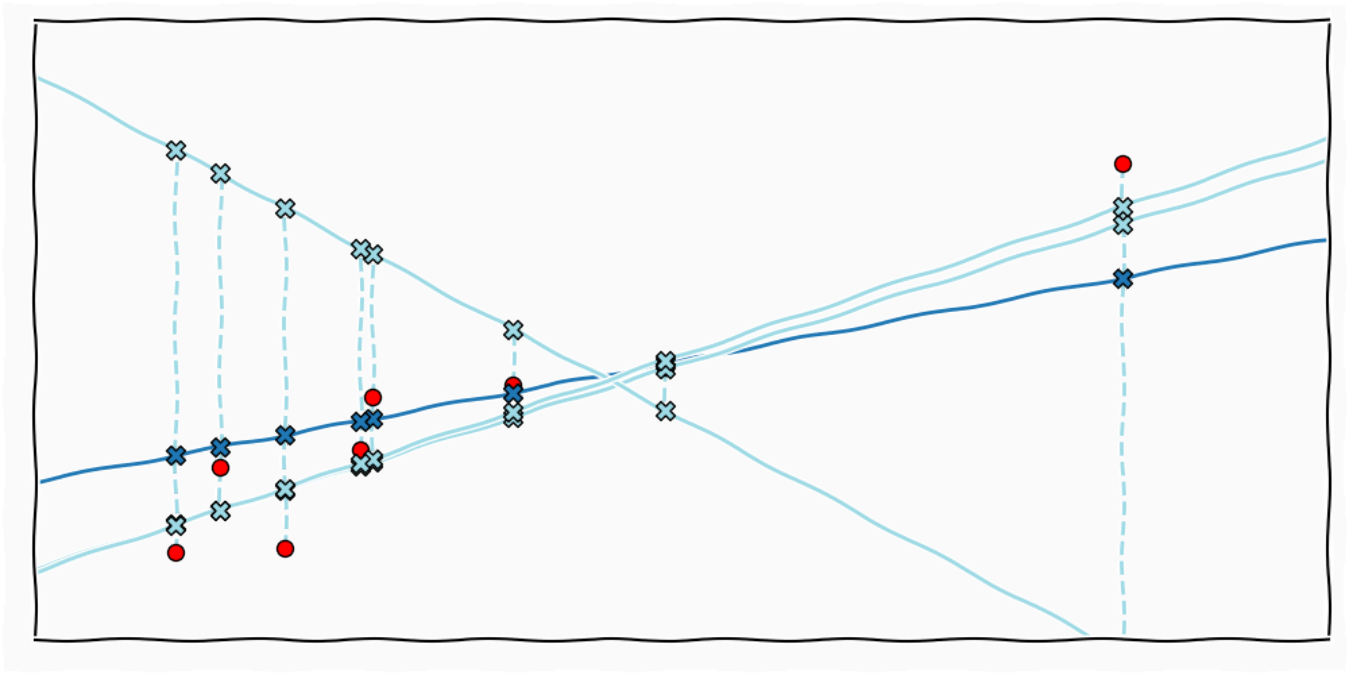
Linear Regression: Modelling



Linear Regression: Modelling

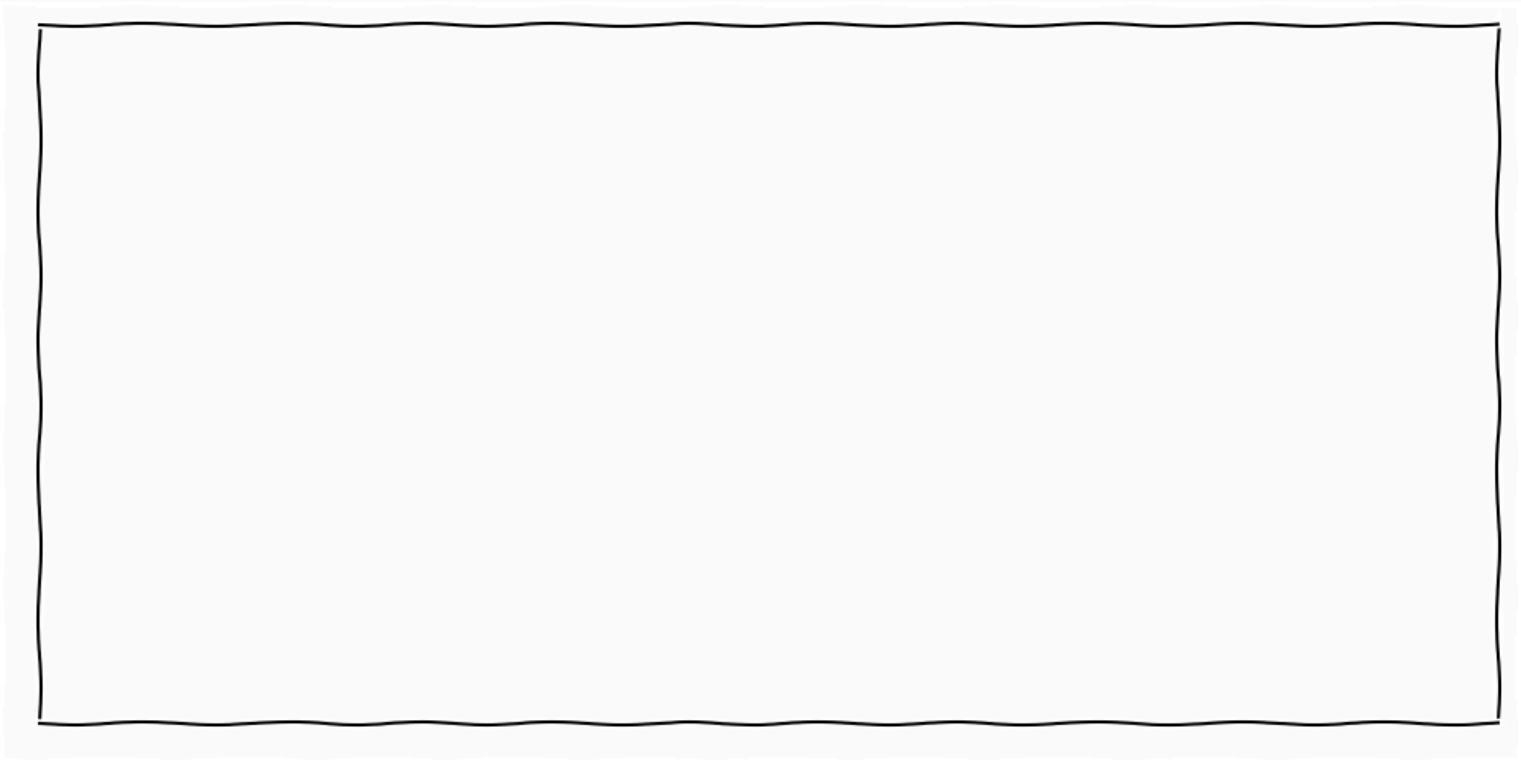


Linear Regression: Likelihood

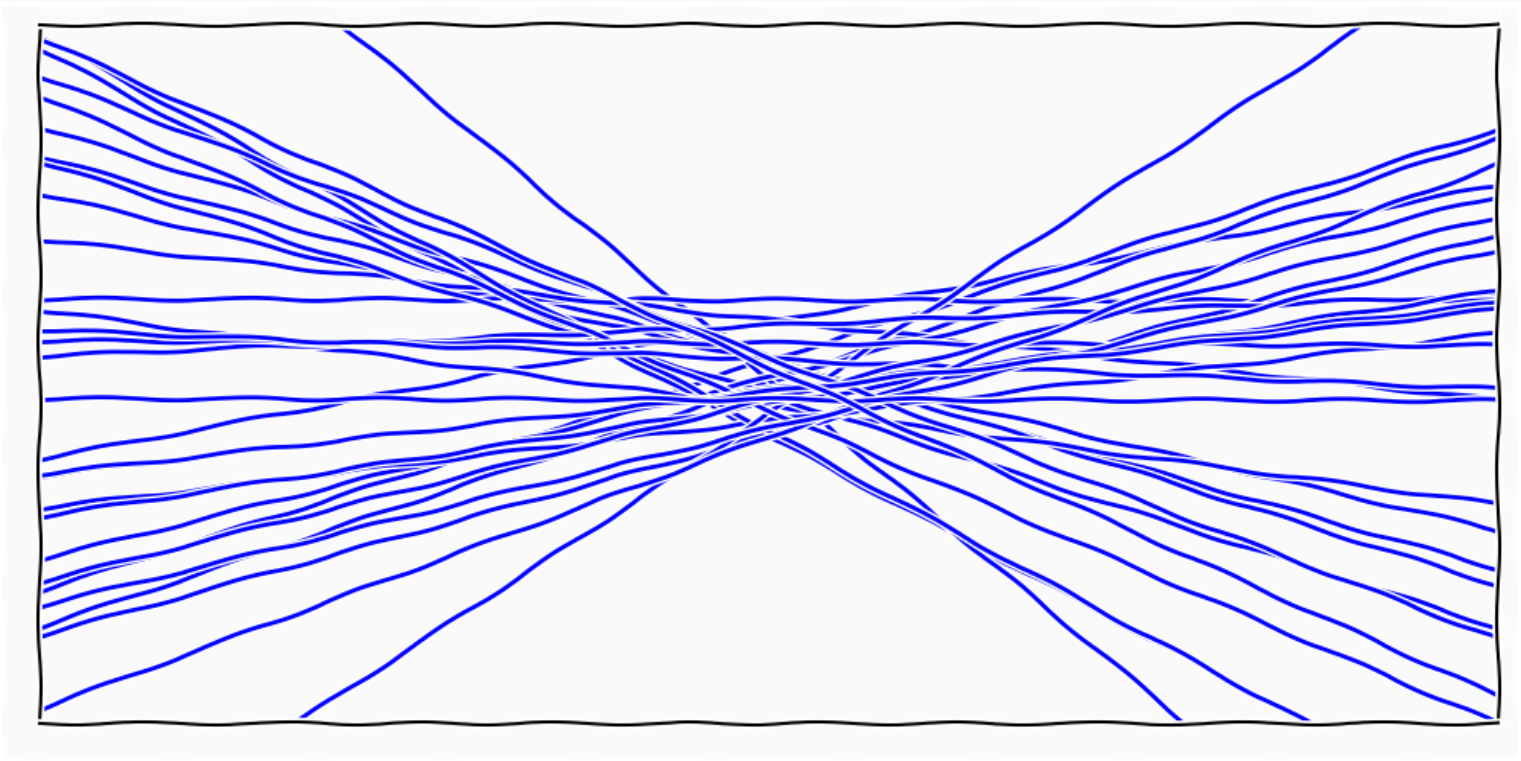


$$p(y|\mathbf{w}, \mathbf{x}) = \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)$$

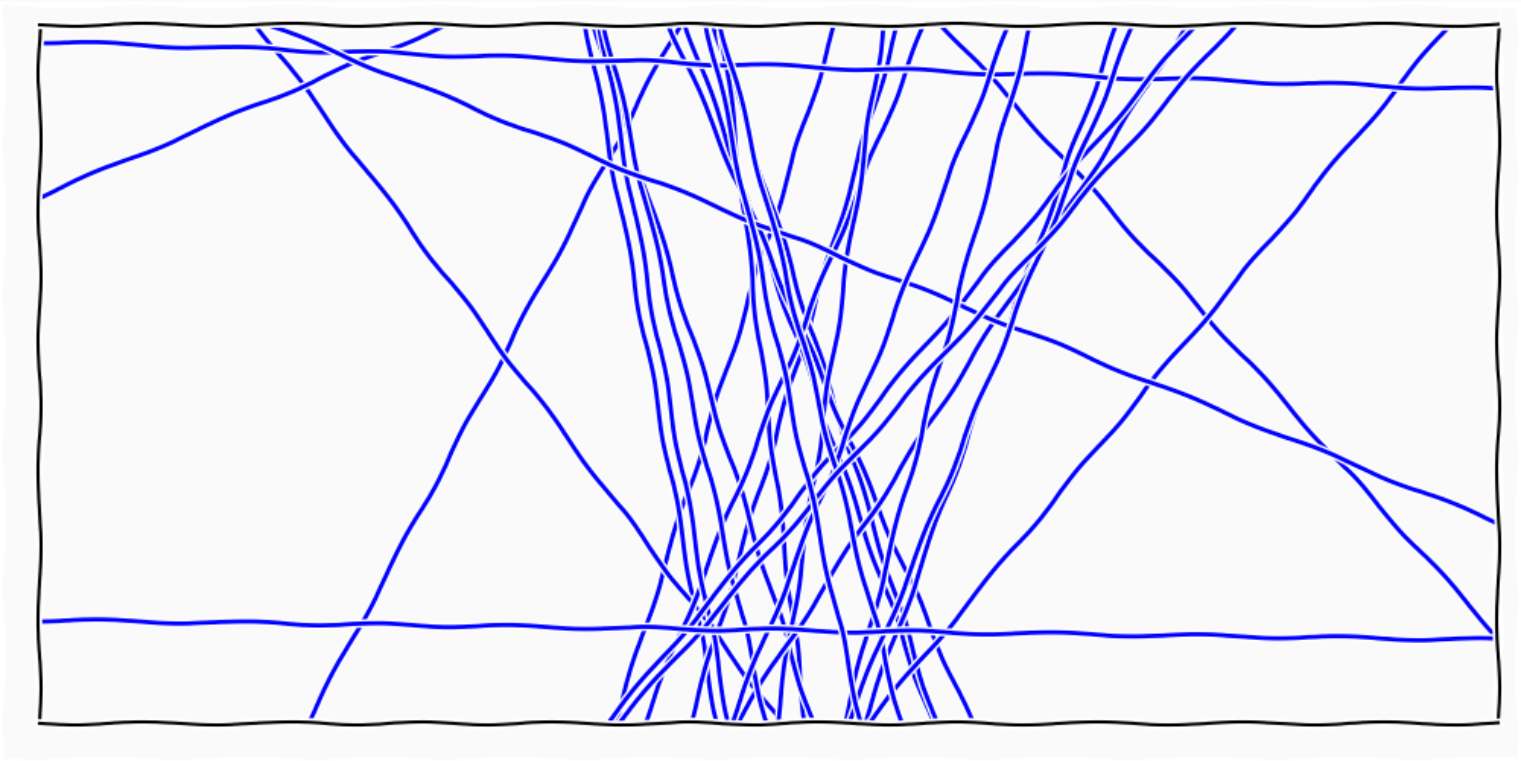
But I don't believe in a flat earth



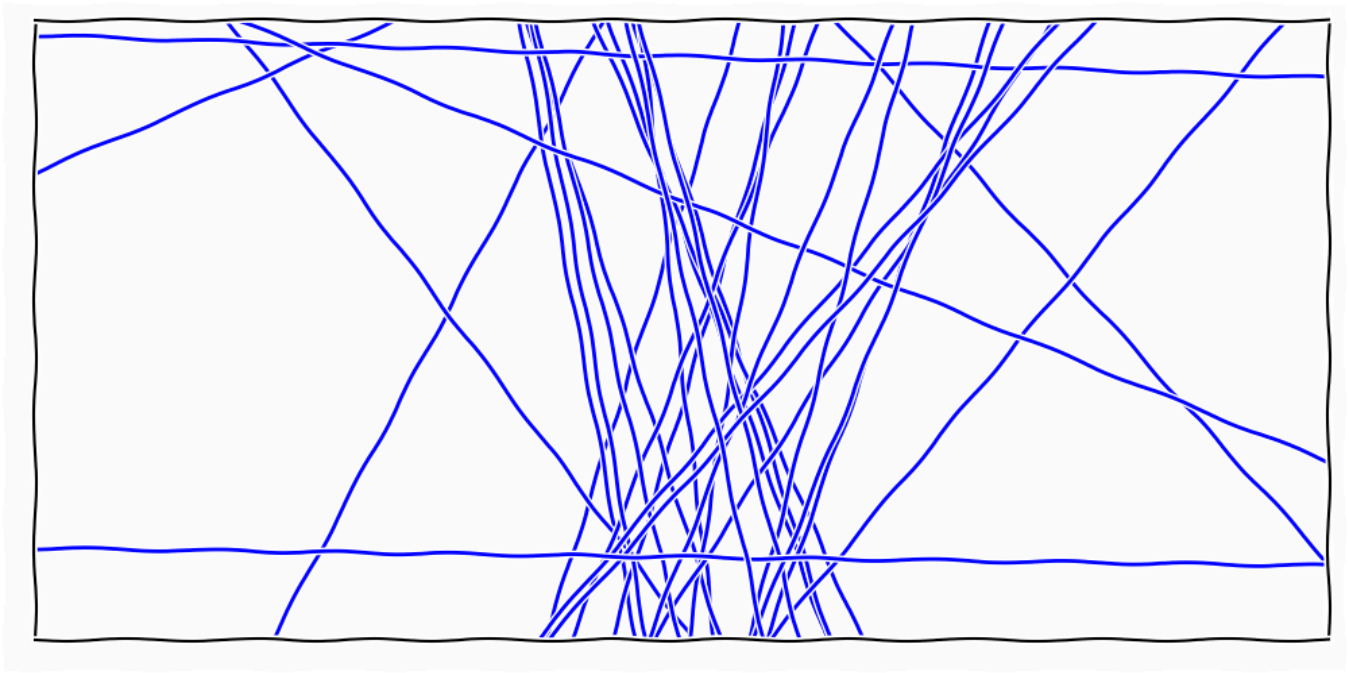
Prior



Prior

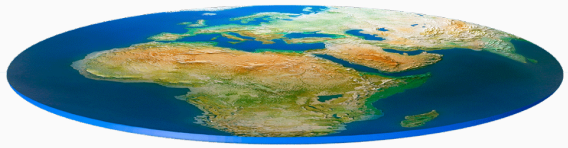


Prior



$$w \sim \mathcal{N}(0, 2)$$

Prior



"Well this is how much credibility belief I give to the hypothesis that the earth is flat"

Two opposing theories

That French Dude Again

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.
– Laplace Laplace, 1814

Semantics

$$p(w \mid y) = \frac{p(y \mid w)p(w)}{\int p(y \mid w)p(w)\mathrm{d}w}$$

Likelihood How much evidence is there in the data for a specific hypothesis

Prior What are my beliefs about different hypothesis

Posterior What is my updated belief after having seen data

Evidence What is my belief about any observations

Challenges

- Acquire knowledge
- Mathematically formulate knowledge
- Acquire data
- Inference
- Interpret Results

Conjugacy

Model and Inference

$$p(\mathcal{D}, \theta) = p(\mathcal{Y}, \mathcal{X}, \theta) = p(\mathcal{Y} \mid \mathcal{X}) p(\mathcal{X} \mid \theta) p(\theta)$$

$$p(\theta \mid \mathcal{D})$$

Inference

Churn the handle

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mu)p(\mu)}{\underbrace{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}_{\text{This is hard}}}$$

Inference

Churn the handle

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mu)p(\mu)}{\underbrace{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}_{\text{This is hard}}}$$

Conjugacy

- We know the functional form of the posterior

Inference

Churn the handle

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mu)p(\mu)}{\underbrace{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}_{\text{This is hard}}}$$

Conjugacy

- We know the functional form of the posterior
- We know that the posterior is proportional to the likelihood times the prior

Inference

Churn the handle

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mu)p(\mu)}{\underbrace{\int p(\mathcal{D}|\mu)p(\mu)\mathrm{d}\mu}_{\text{This is hard}}}$$

Conjugacy

- We know the functional form of the posterior
- We know that the posterior is proportional to the likelihood times the prior
- *Use these facts to avoid the integral*

Conjugate Priors

- Most distributions are parametetrised using exponentials
- Exponential family natural parametrisation

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})e^{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})}$$

- Conjugate prior

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\chi})^{\nu}e^{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}}$$

Conjugate Models

Conjugate Priors

Beyond Conjugacy

$$p(y) = \int p(y \mid x) p(x) \mathrm{d}x$$

- Analytically intractable
- Computationally intractable

Summary

Summary

Beliefs/Assumptions think of probabilities as quantifications of beliefs

Probability theory operationalise learning

Inference provides a semantic

Tomorrow

Part I A Story about three donkeys

Part II Machine Learning in a practical setting

Part III Practical session

References

References

 Laplace, Pierre Simon (1814). *Théorie analytique des probabilités*. Mme. Ve. Sourcier.