Building Models

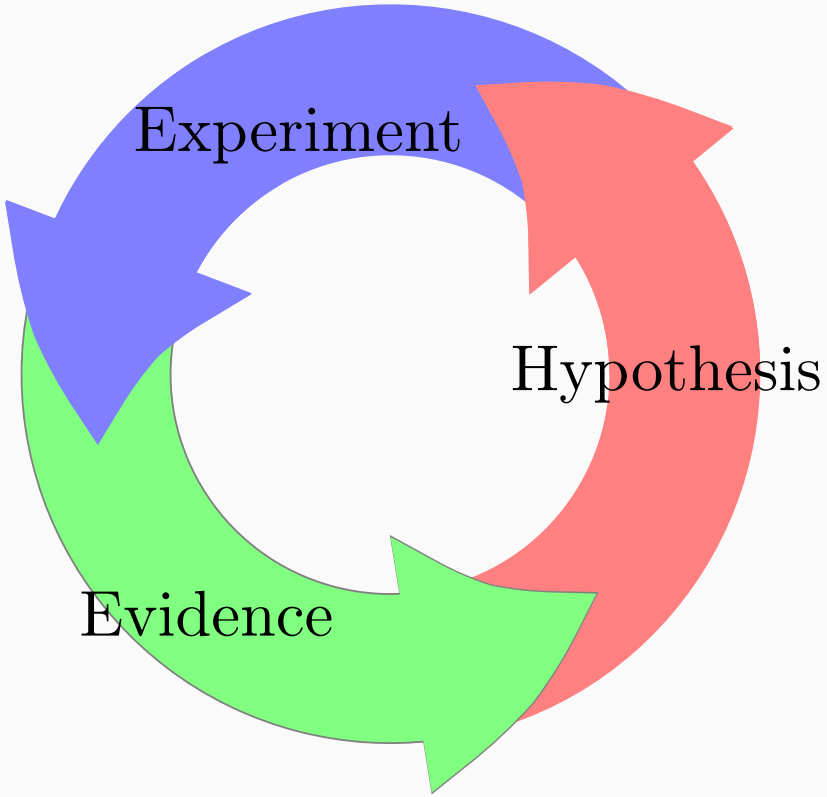
Carl Henrik Ek - che29@cam.ac.uk

29th of September 2021

http://carlhenrik.com

Recap

Science?



Scientific Modelling

"Scientific modelling is a scientific activity, the aim of which is to make a particular part or feature of the world easier to understand, define, quantify, visualize, or simulate by referencing it to existing and usually commonly accepted knowledge." [1]

[1] Wikipedia

Scientific Modelling

"Scientific modelling is a scientific activity, the aim of which is to make a particular part or feature of the world easier to understand, define, quantify, visualize, or simulate by referencing it to existing and usually commonly accepted knowledge." [2]

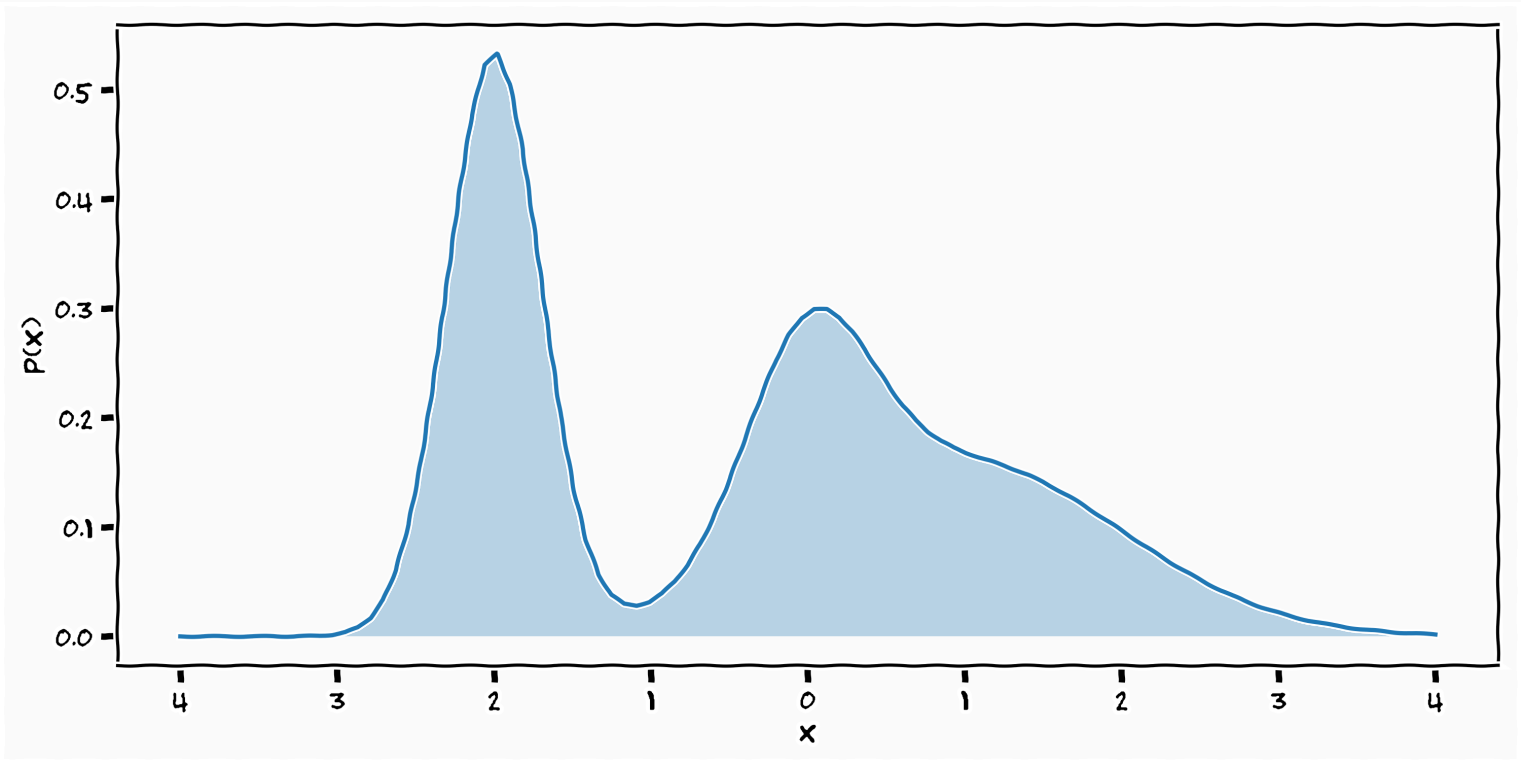
[2] Wikipedia

Machine Learning

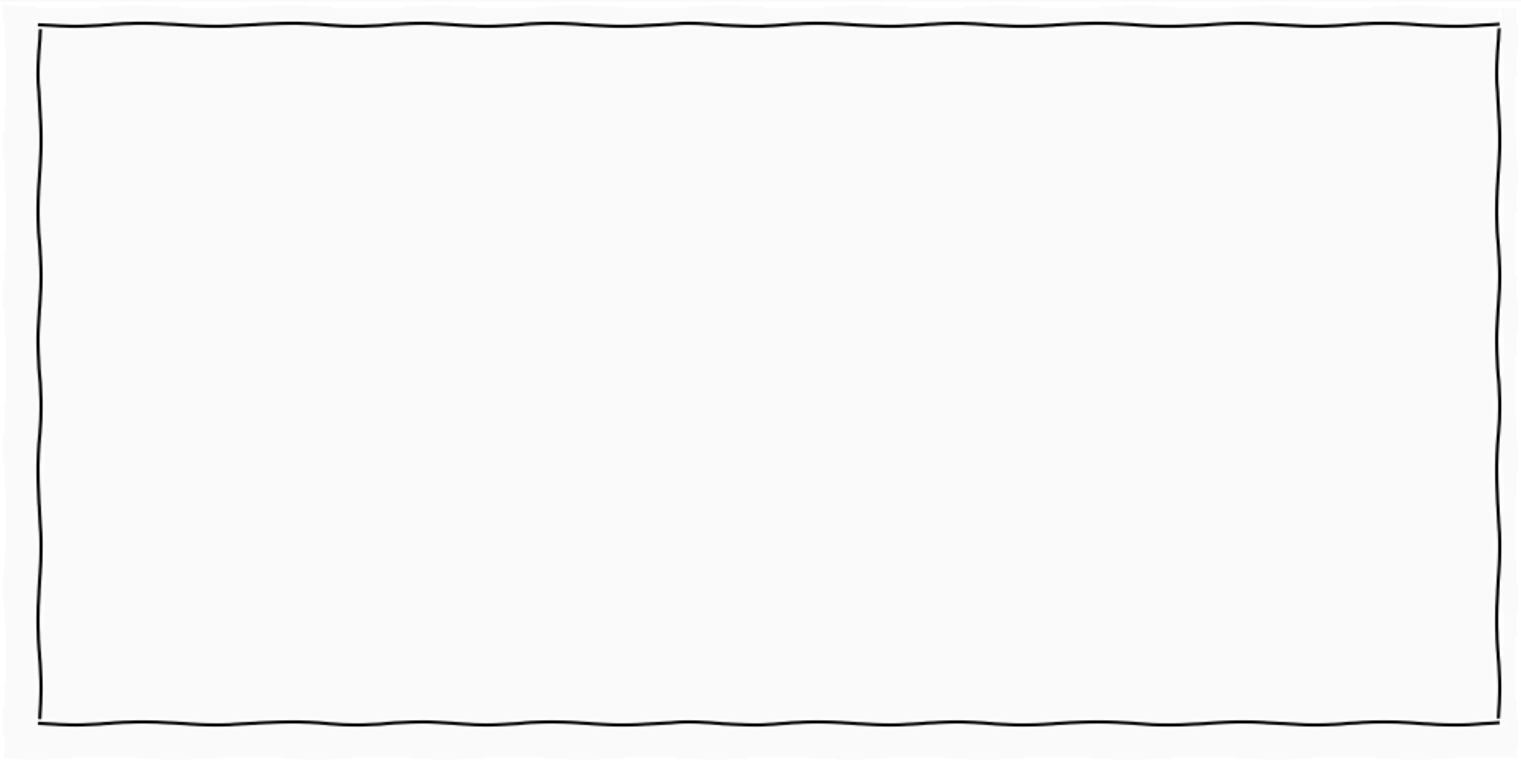
The Task of Machine Learning

1. How can we formulate beliefs and assumptions mathematically (Knowledge)
2. How can we connect our assumptions with data (Referencing)
3. How can we update our beliefs (Understand)

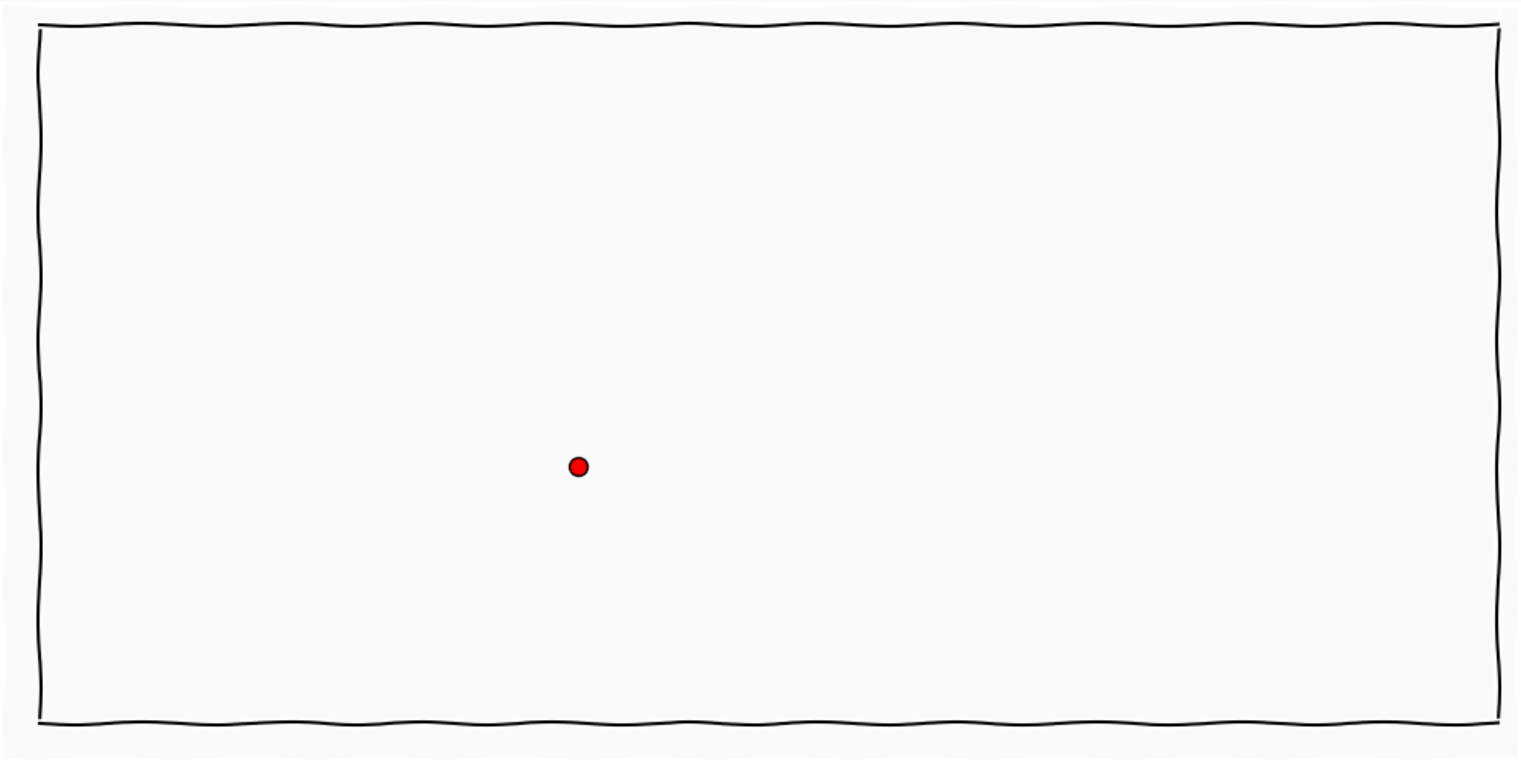
Probabilities



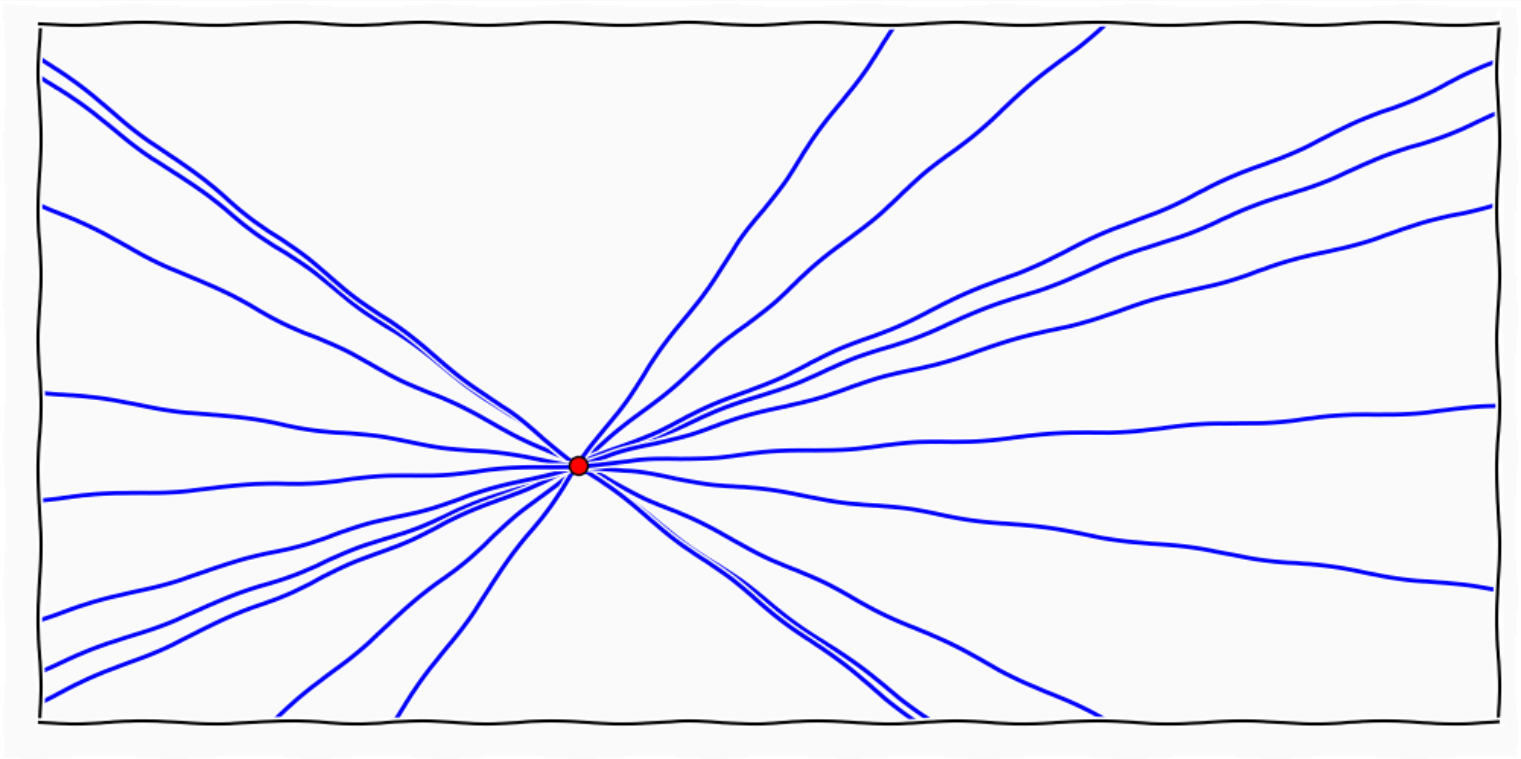
Linear Regression



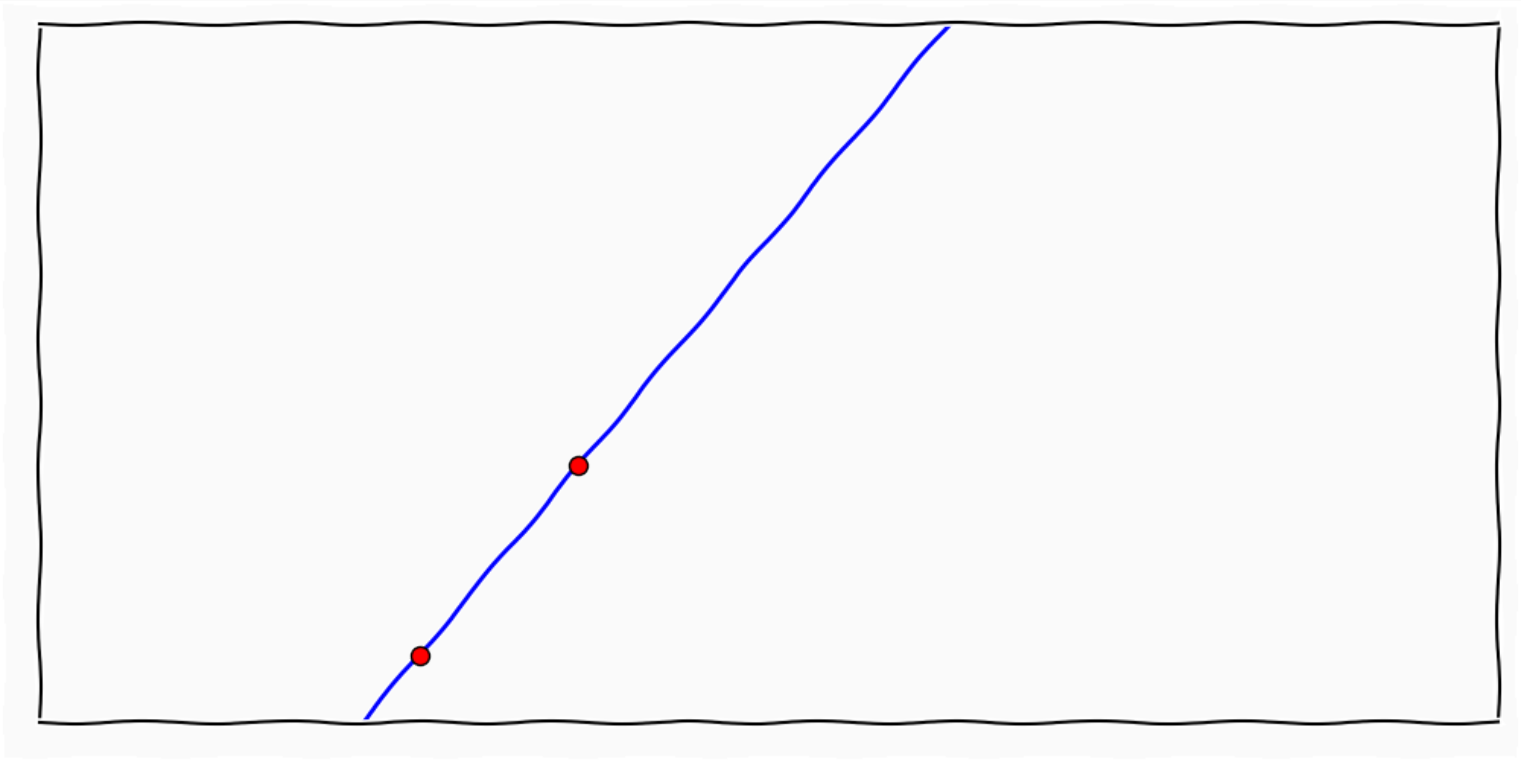
Linear Regression



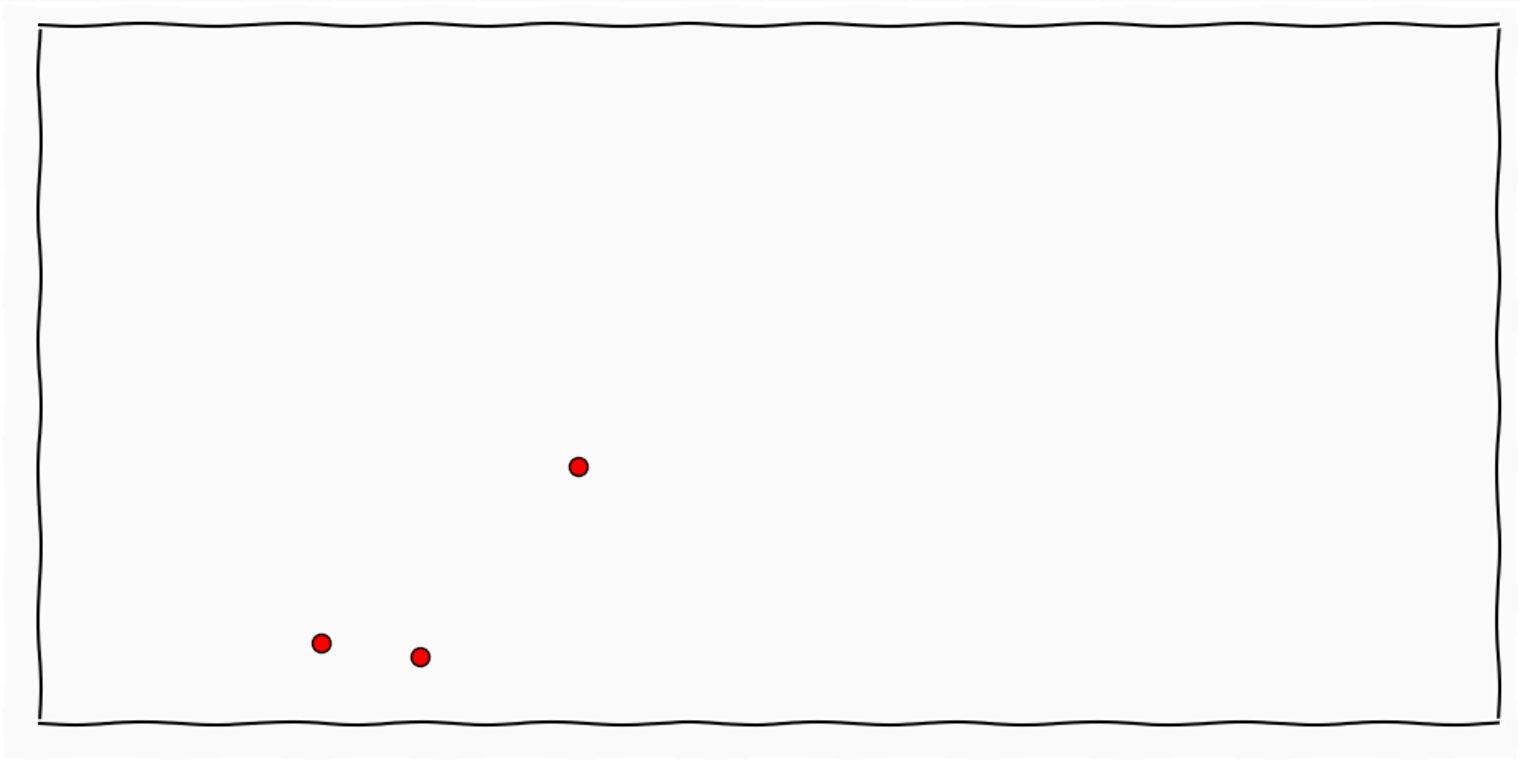
Linear Regression



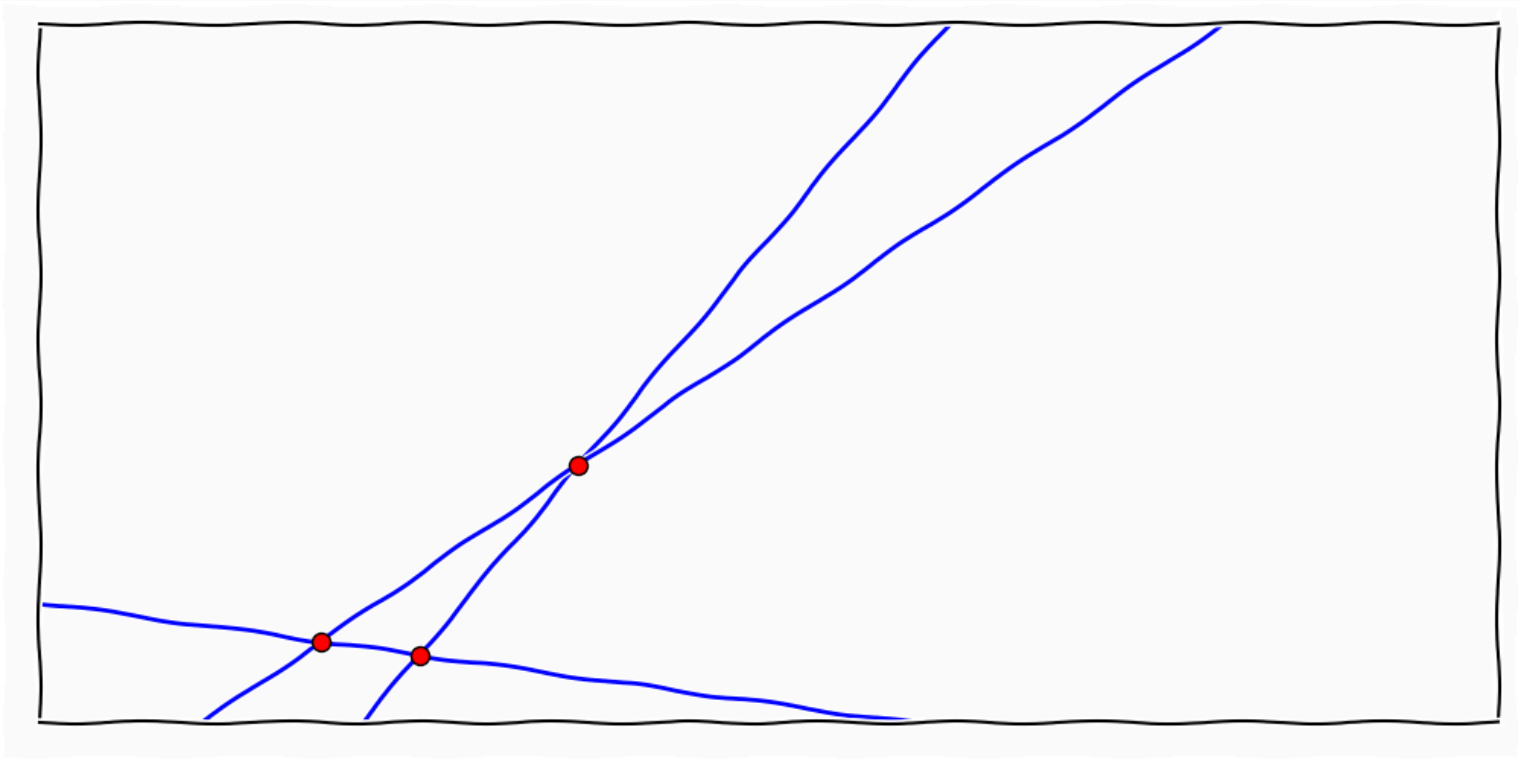
Linear Regression



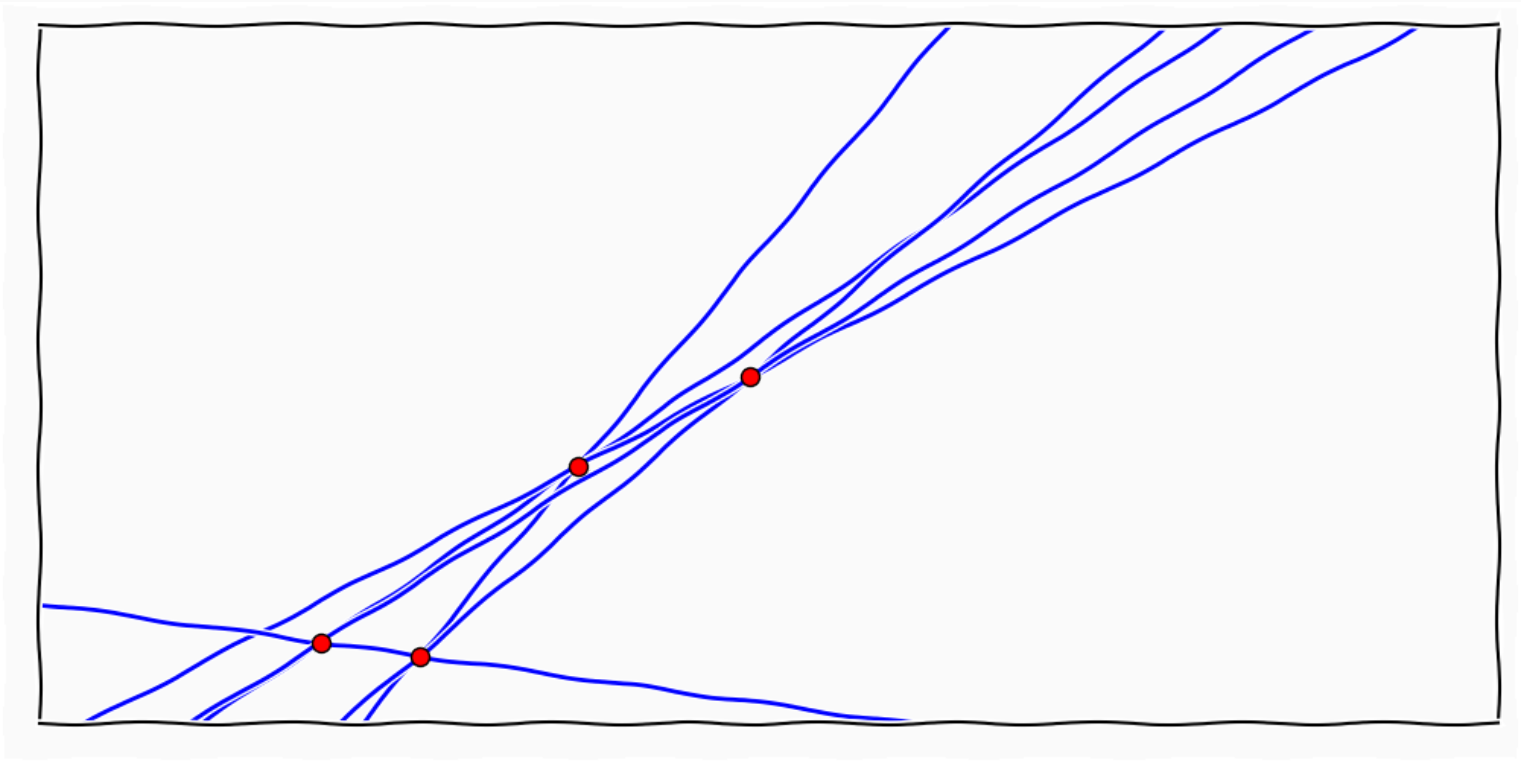
Linear Regression



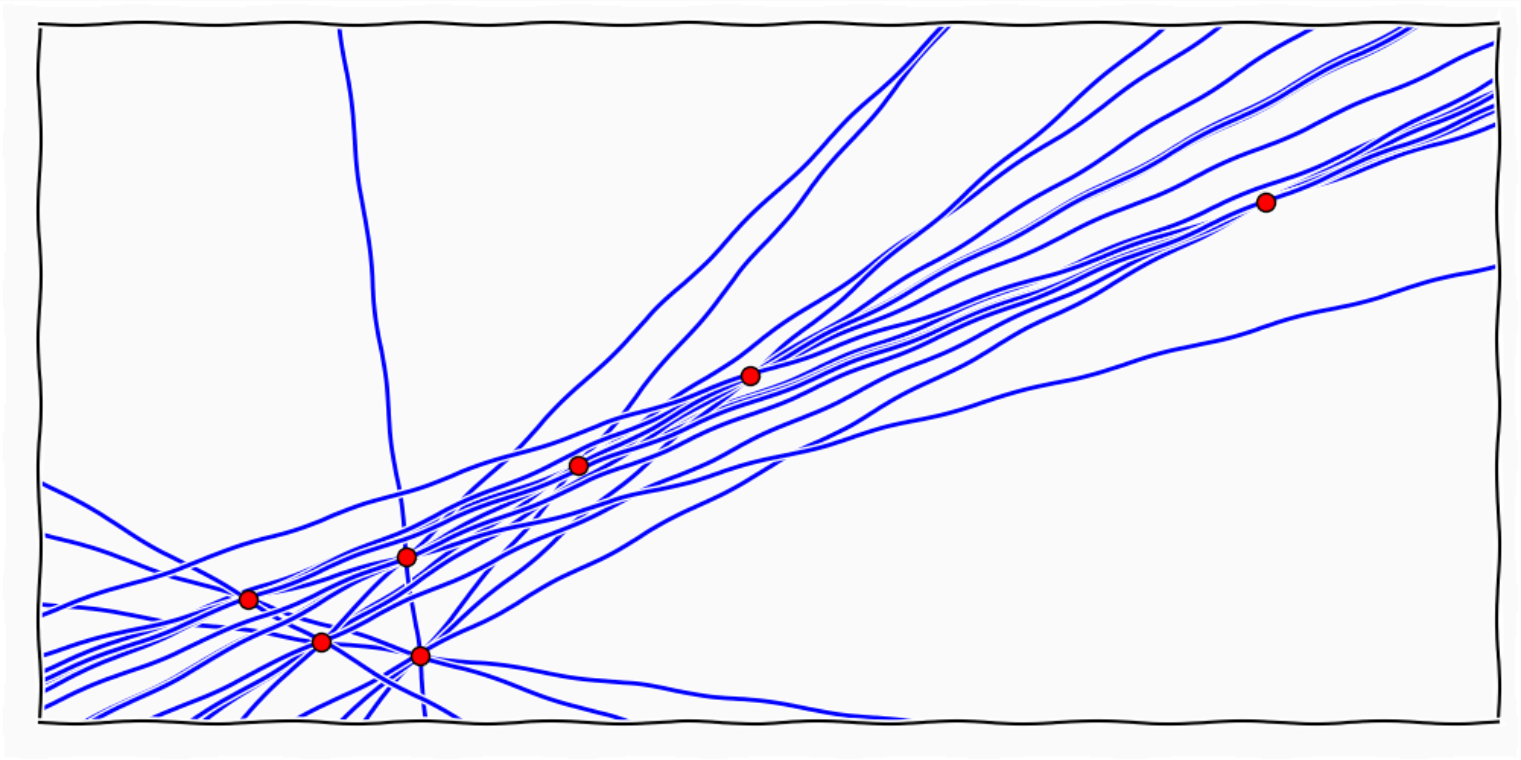
Linear Regression



Linear Regression



Linear Regression



Linear Regression: High School

$$\underbrace{\mathbf{A}}_{m\times n}\underbrace{\mathbf{x}}_{n\times 1} = \underbrace{\mathbf{b}}_{m\times 1}$$

- Over-determined $m > n$

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i}^{m} (b_i - \mathbf{A}_{i:}\mathbf{x})^2$$

Linear Regression: High School

$$\underbrace{\mathbf{A}}_{m\times n}\underbrace{\mathbf{x}}_{n\times 1} = \underbrace{\mathbf{b}}_{m\times 1}$$

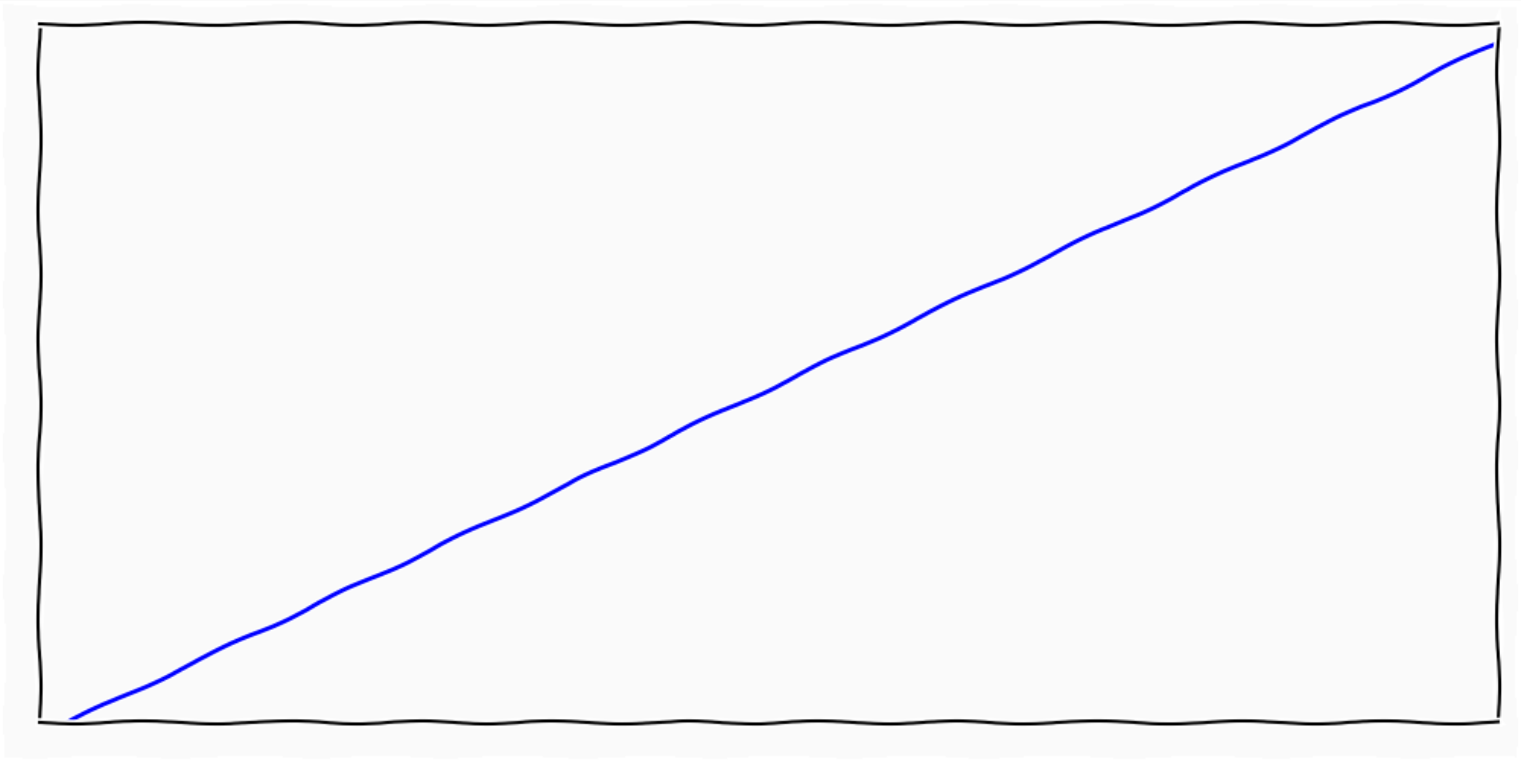
- Over-determined $m > n$

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i}^{m} (b_i - \mathbf{A}_{i:}\mathbf{x})^2$$

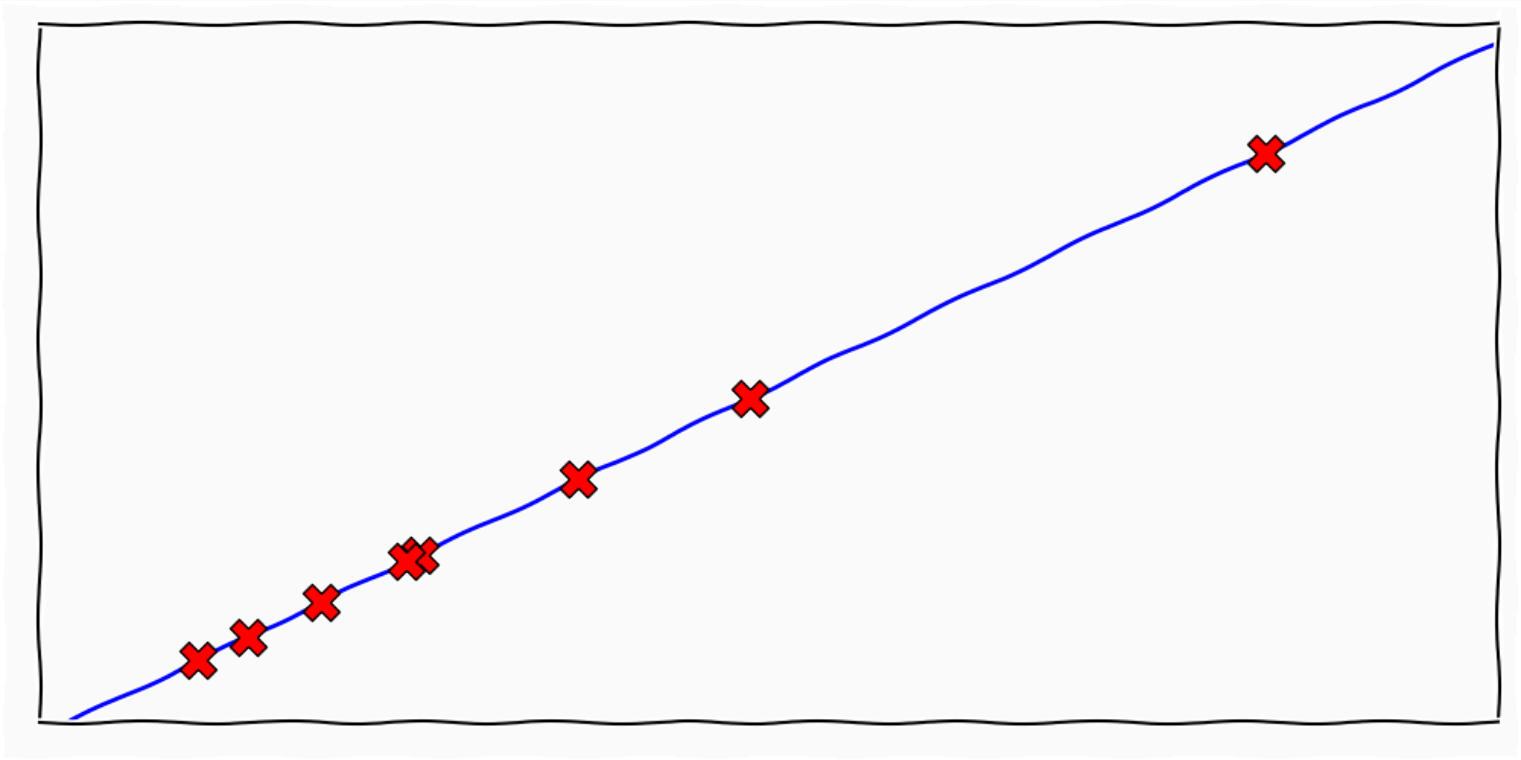
- Under-determined $m < n$

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i}^{m} (b_i - \mathbf{A}_{i:}\mathbf{x})^2 + \lambda \mathbf{x}^{\mathrm{T}}\mathbf{x}$$

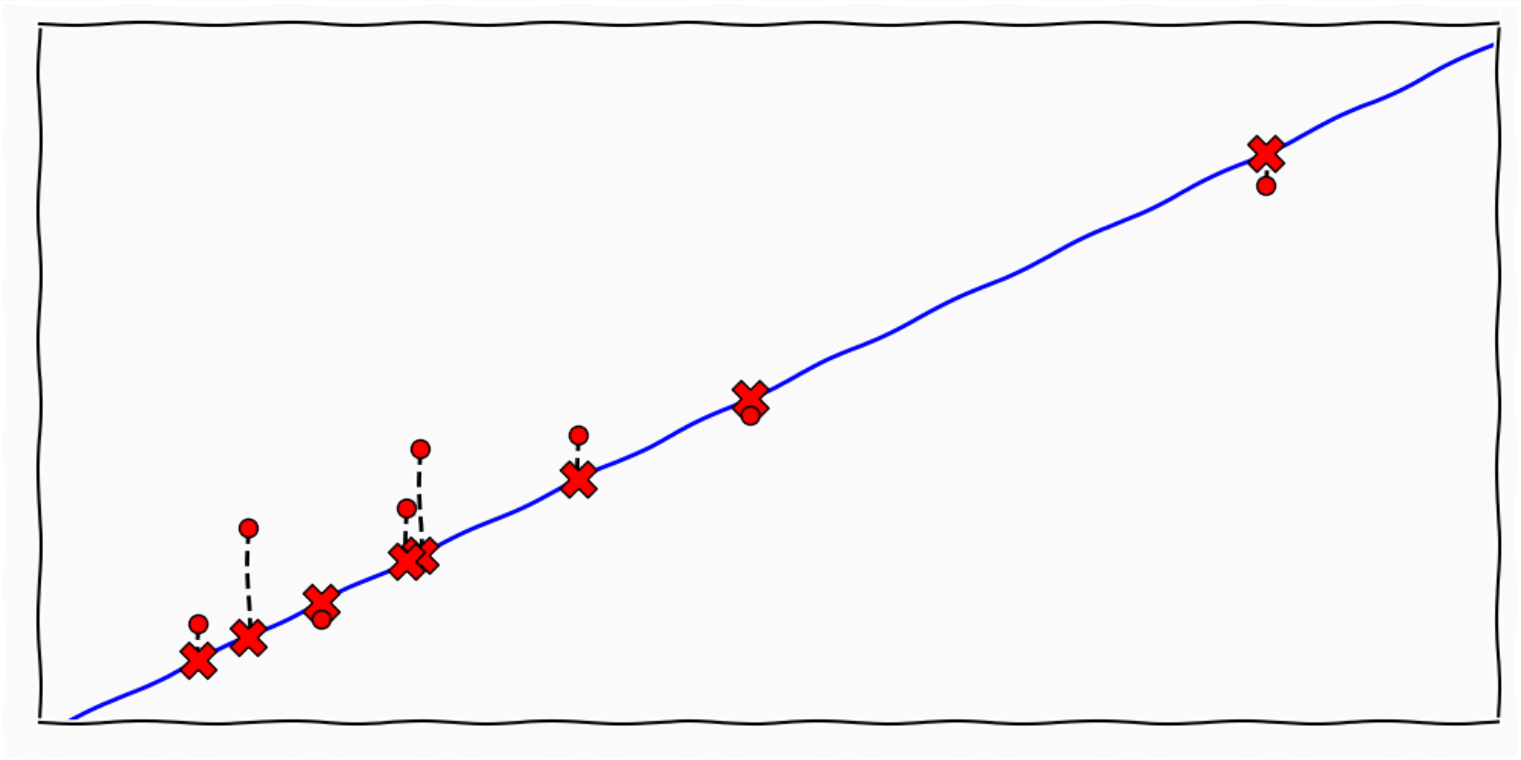
Linear Regression: Modelling



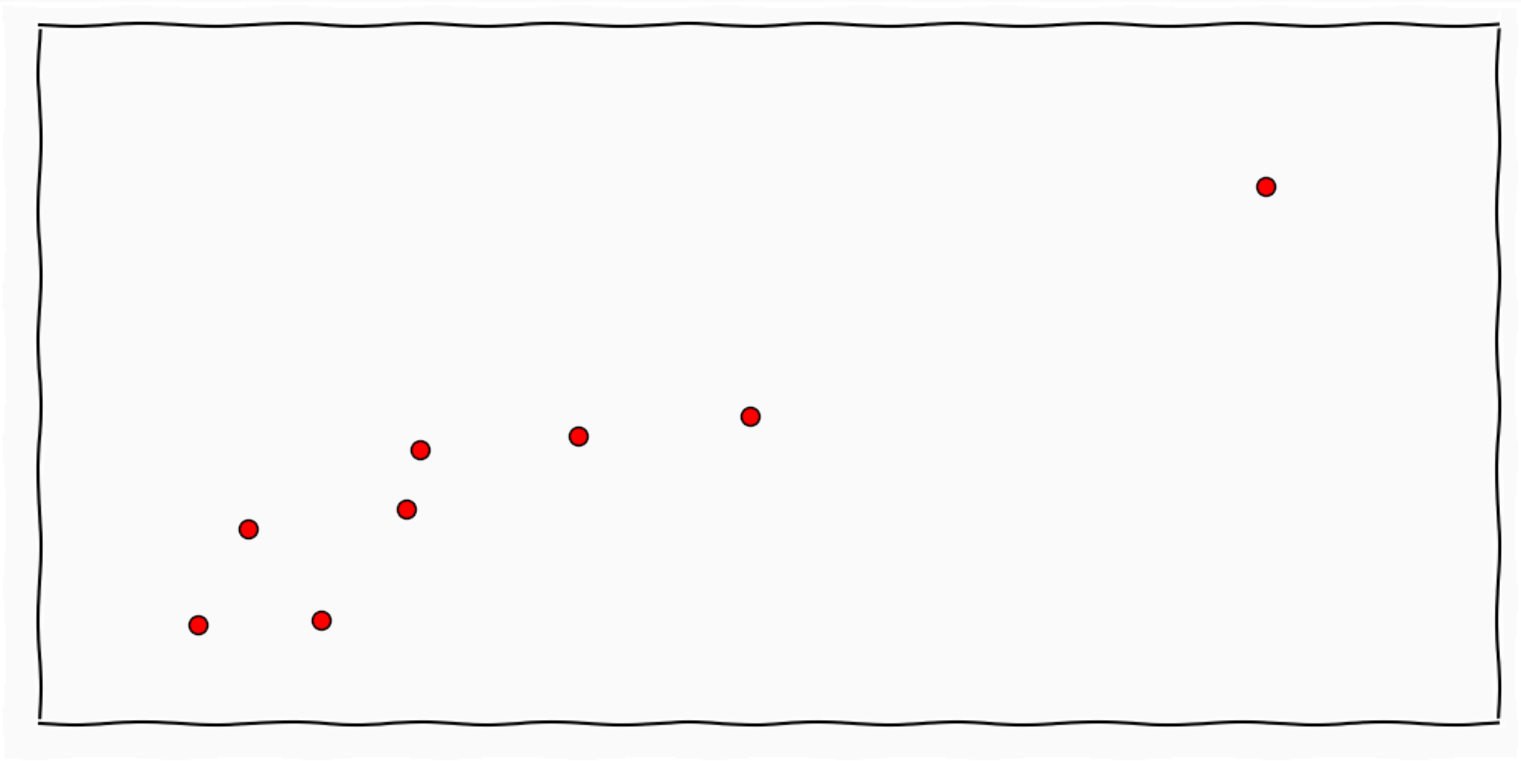
Linear Regression: Modelling



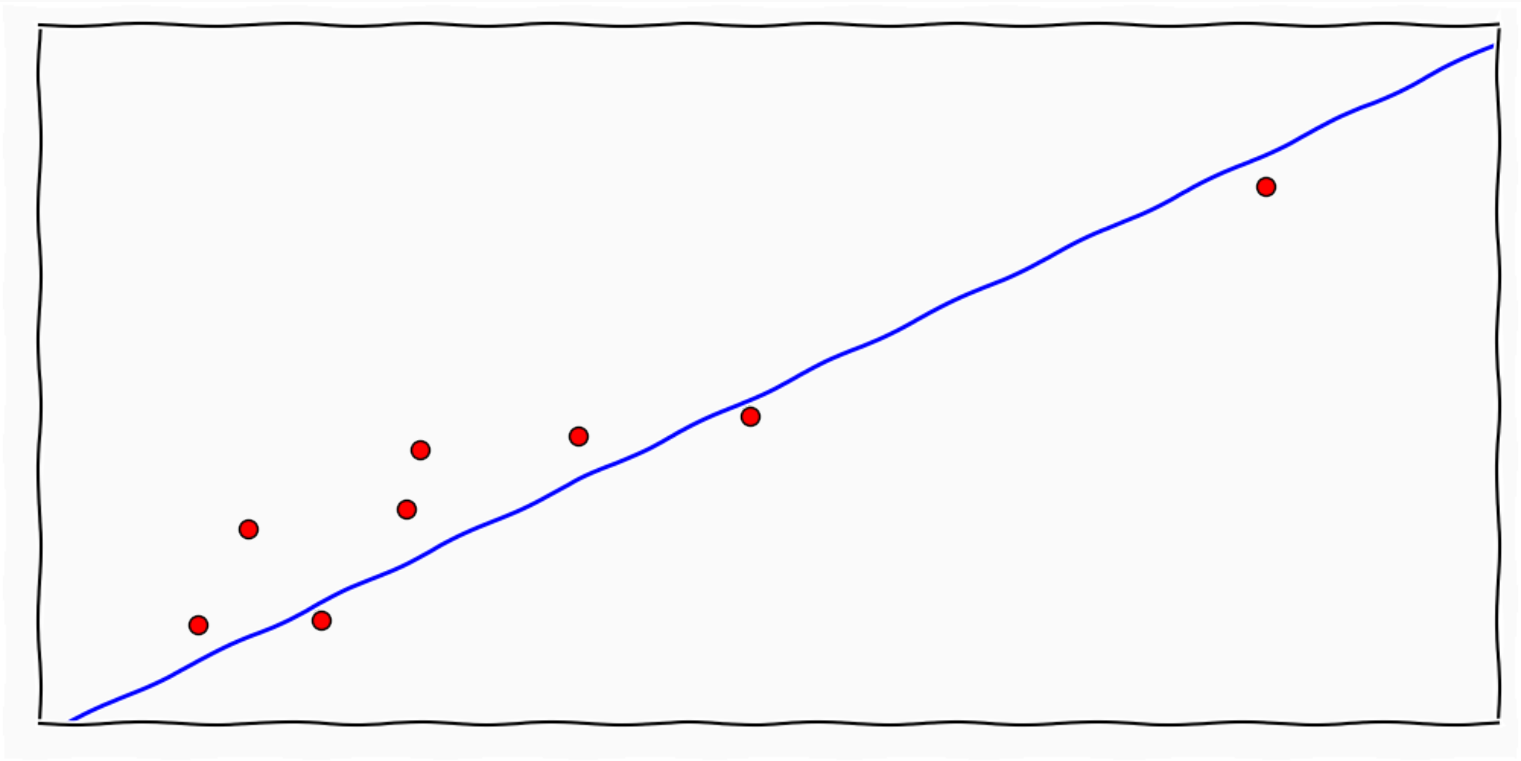
Linear Regression: Modelling



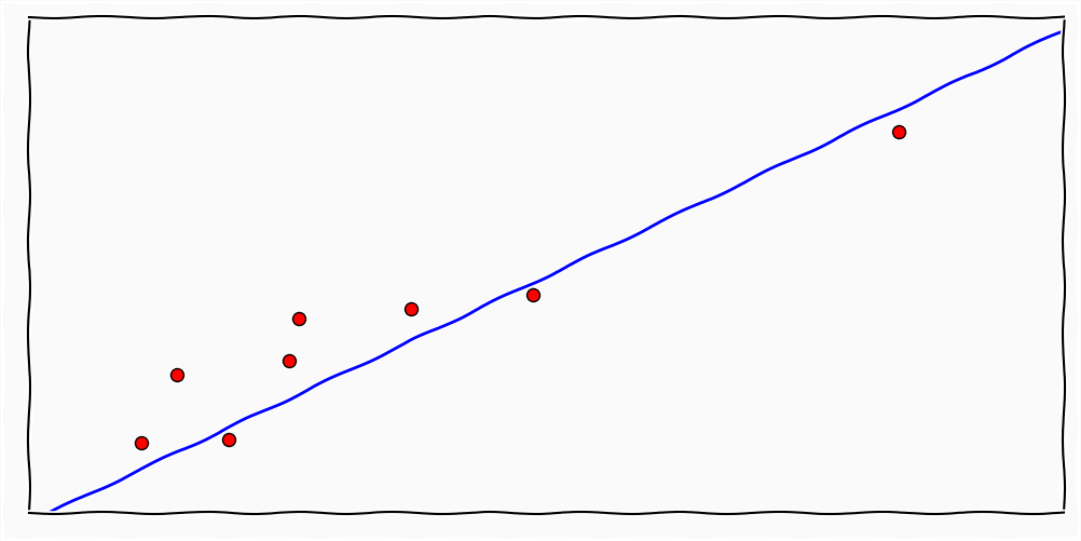
Linear Regression: Modelling



Linear Regression: Modelling



Linear Regression: Modelling



$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$$

Belief/Hypothesis

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

Belief/Hypothesis

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} = \epsilon$$

Belief/Hypothesis

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} = \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} \sim \mathcal{N}(\epsilon|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\epsilon-0)\beta(\epsilon-0)}$$

Belief/Hypothesis

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} = \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} \sim \mathcal{N}(\epsilon|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\epsilon-0)\beta(\epsilon-0)}$$

$$\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})\beta(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})}$$

Belief/Hypothesis

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} = \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} \sim \mathcal{N}(\epsilon|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\epsilon-0)\beta(\epsilon-0)}$$

$$\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})\beta(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})}$$

$$\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) = \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)$$

Belief/Hypothesis

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} = \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} \sim \mathcal{N}(\epsilon|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\epsilon-0)\beta(\epsilon-0)}$$

$$\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})\beta(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})}$$

$$\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) = \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)$$

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

Likelihood

- Likelihood

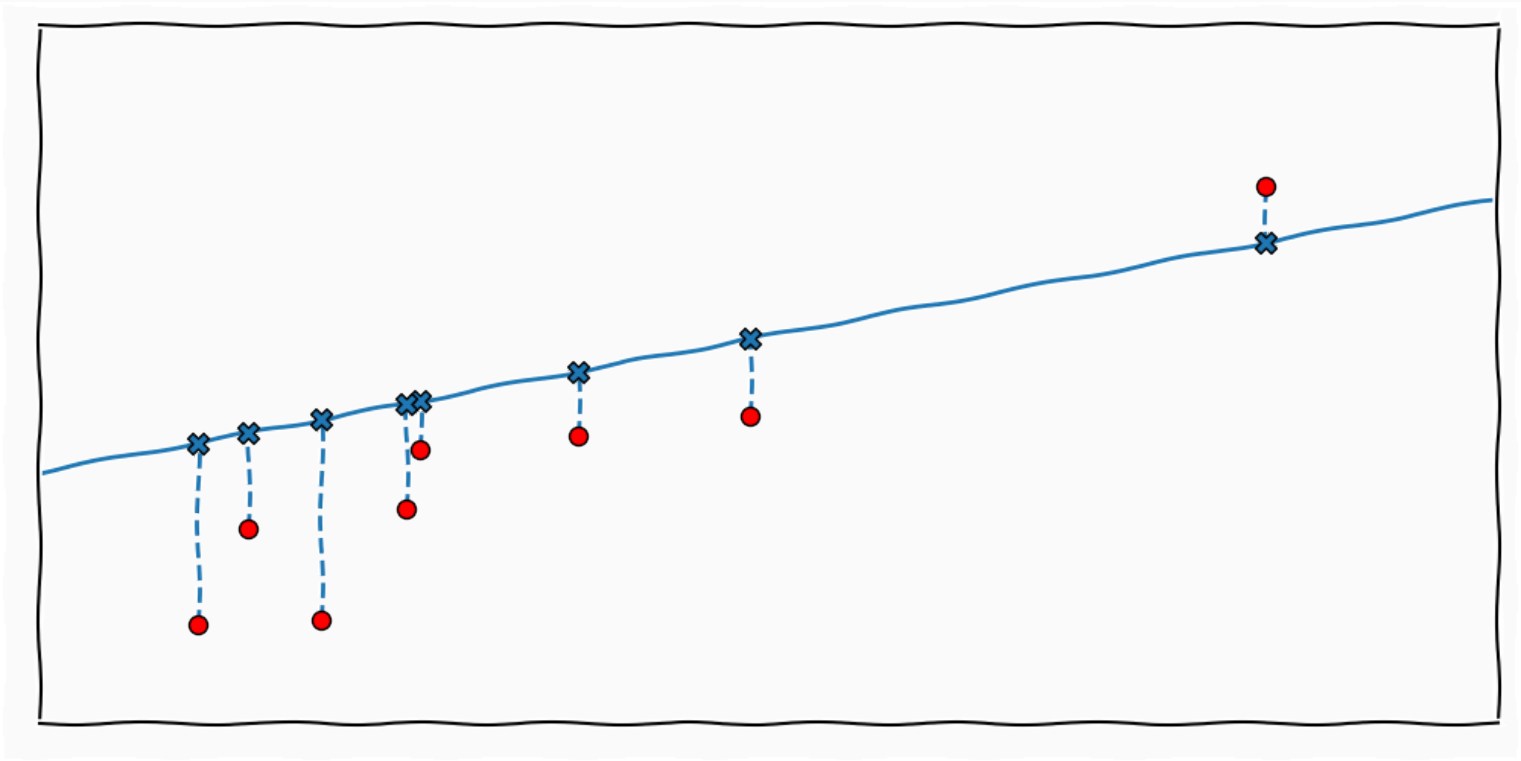
$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}\left(y|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right)$$

- Independence

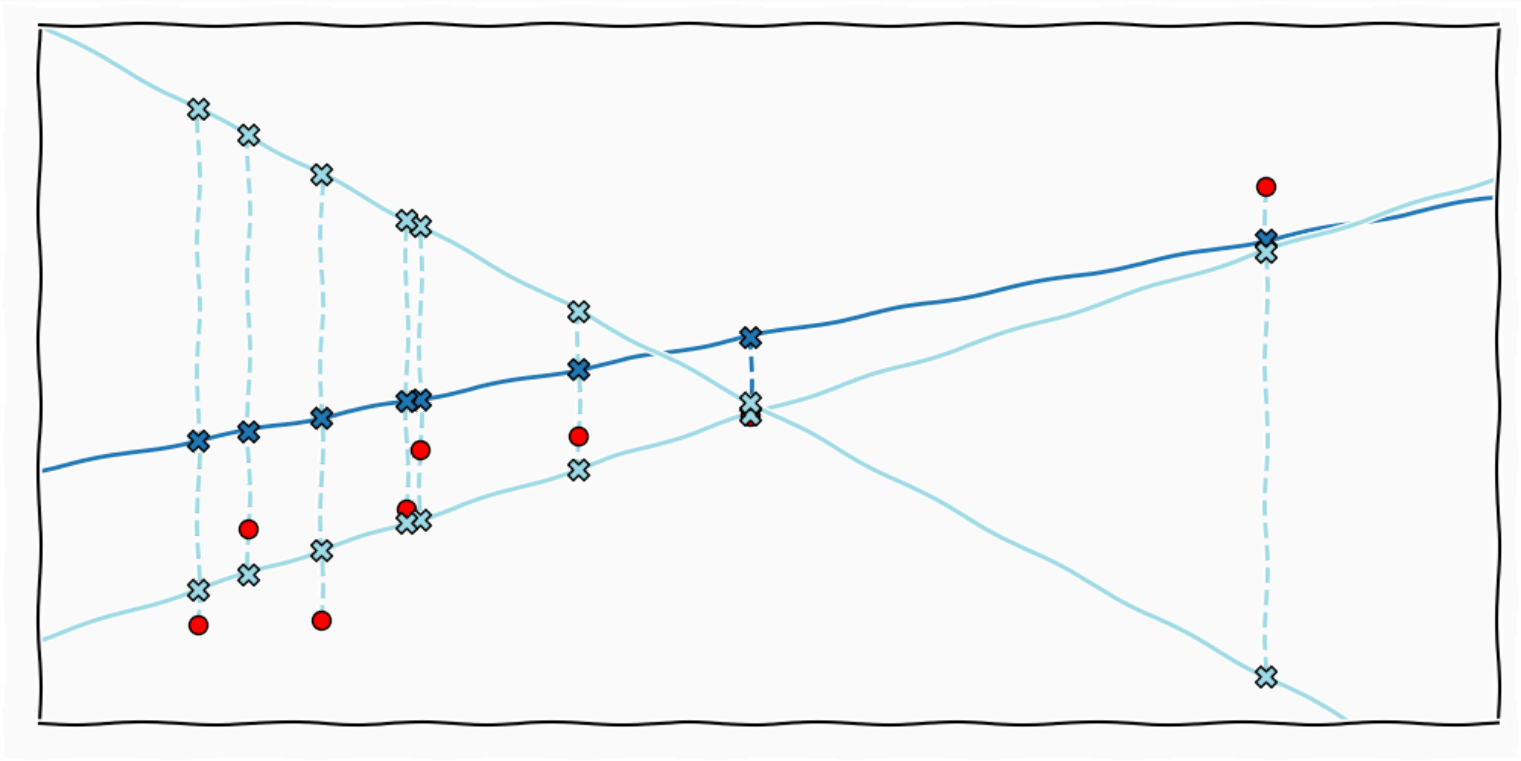
$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right)$$

Assume each output to be independent given the input and the parameters

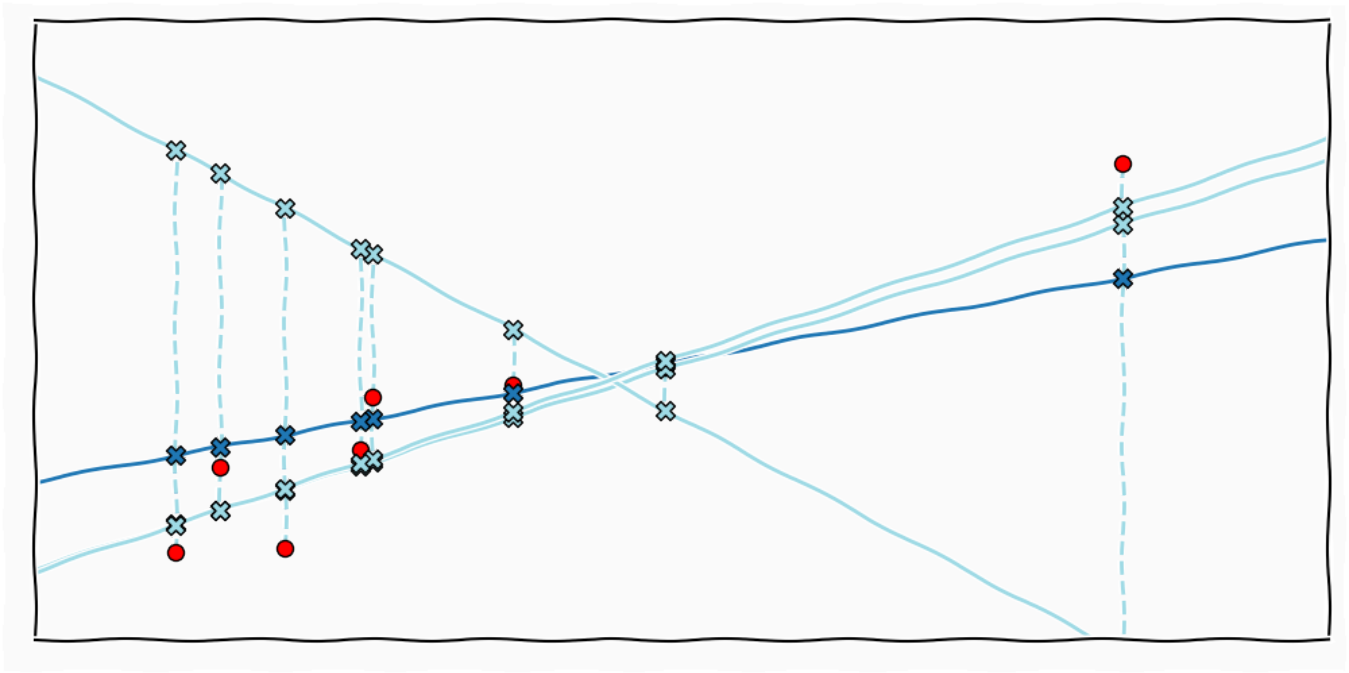
Linear Regression: Modelling



Linear Regression: Modelling

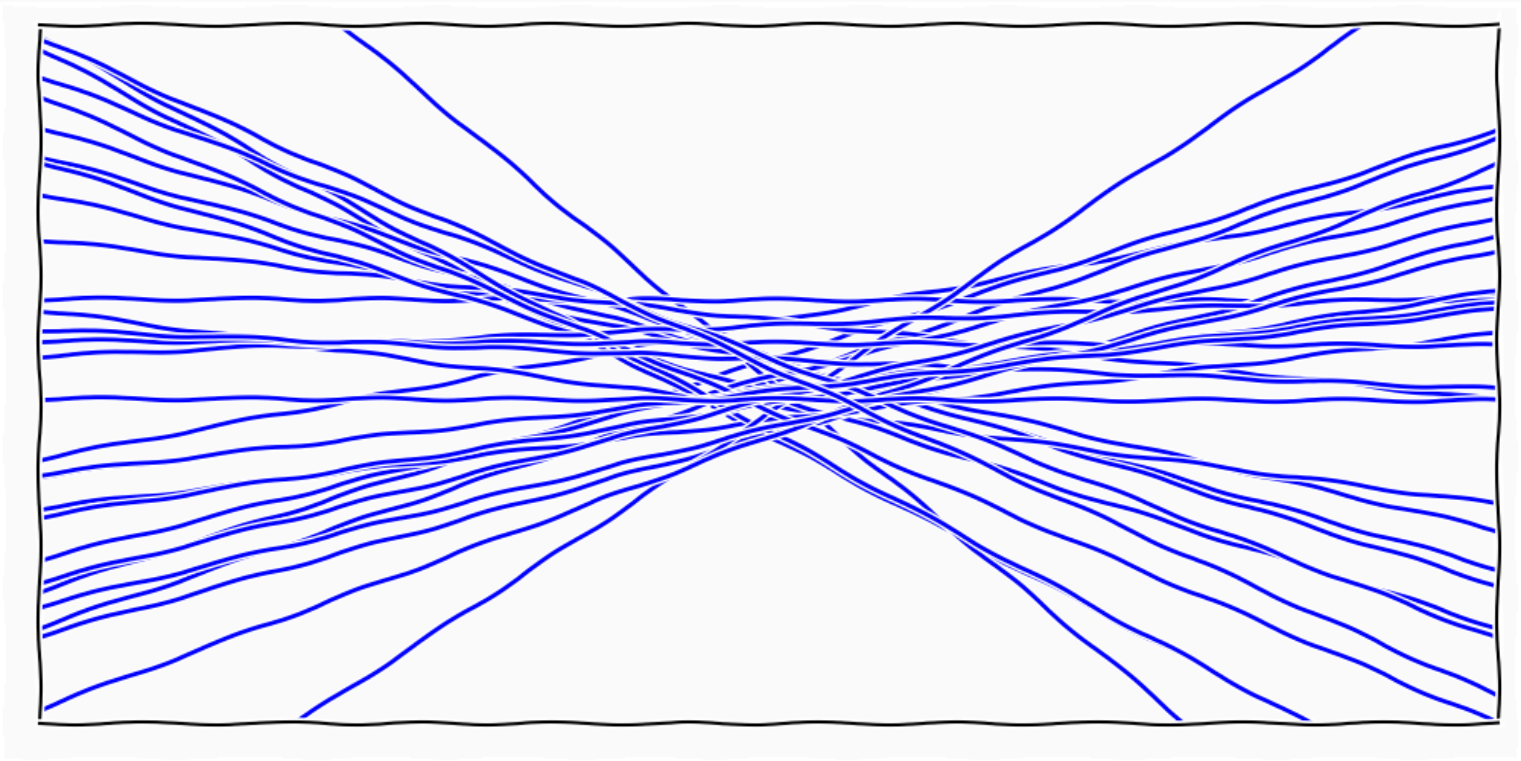


Linear Regression: Likelihood

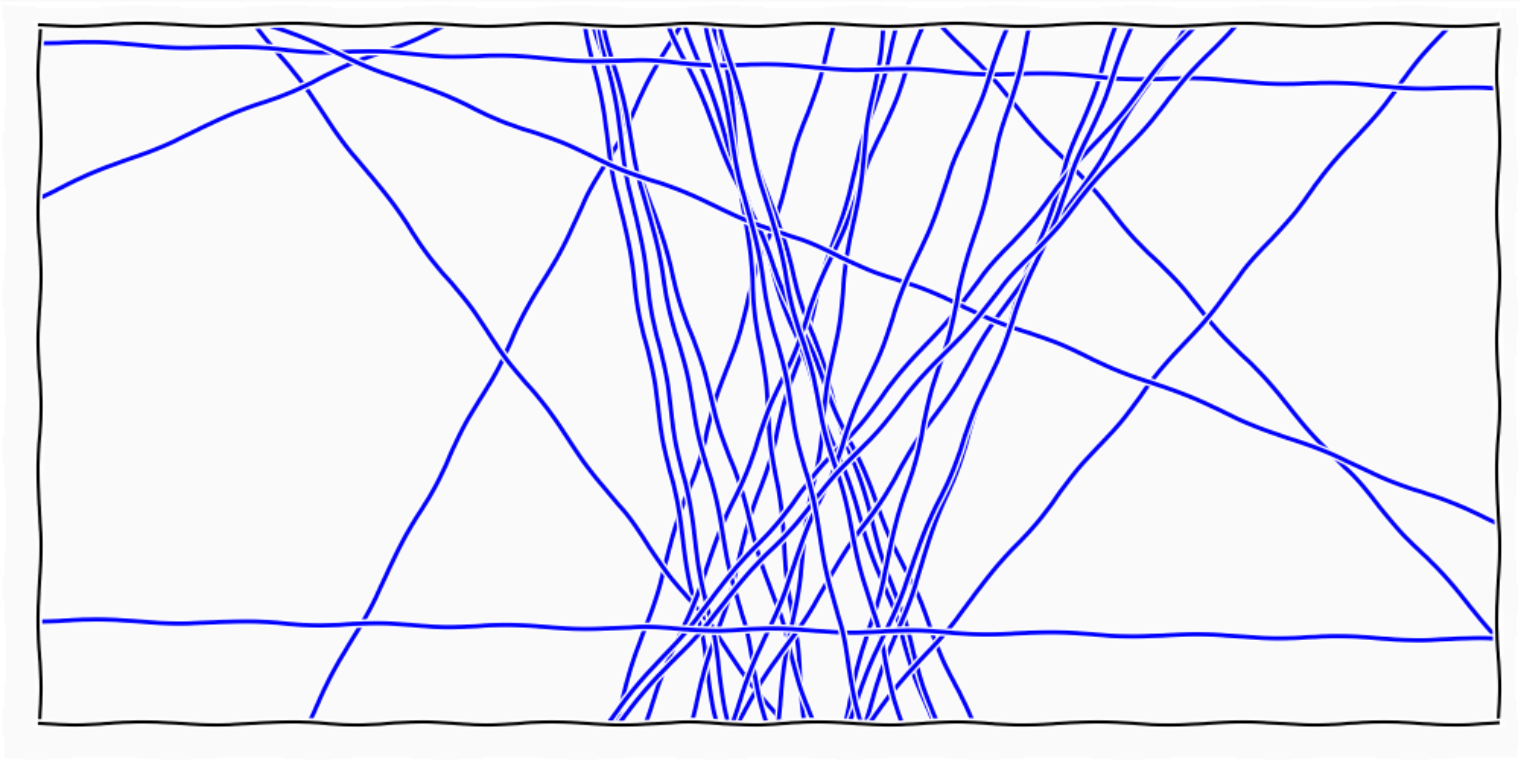


$$p(y|\mathbf{w}, \mathbf{x}) = \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)$$

Prior



Prior



Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

- Identification

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

- Identification

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

- Posterior

$$\mathbf{m}_N = \left(\mathbf{S}_0^{-1} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1}\left(S_0^{-1}\mathbf{m}_0 + \beta\phi(\mathbf{X})^{\mathrm{T}}\mathbf{y}\right)$$
$$\mathbf{S}_N = \left(\mathbf{S}_0^{-1} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1}$$

Posterior

- Assumption Zero mean isotropic Gaussian

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$

Posterior

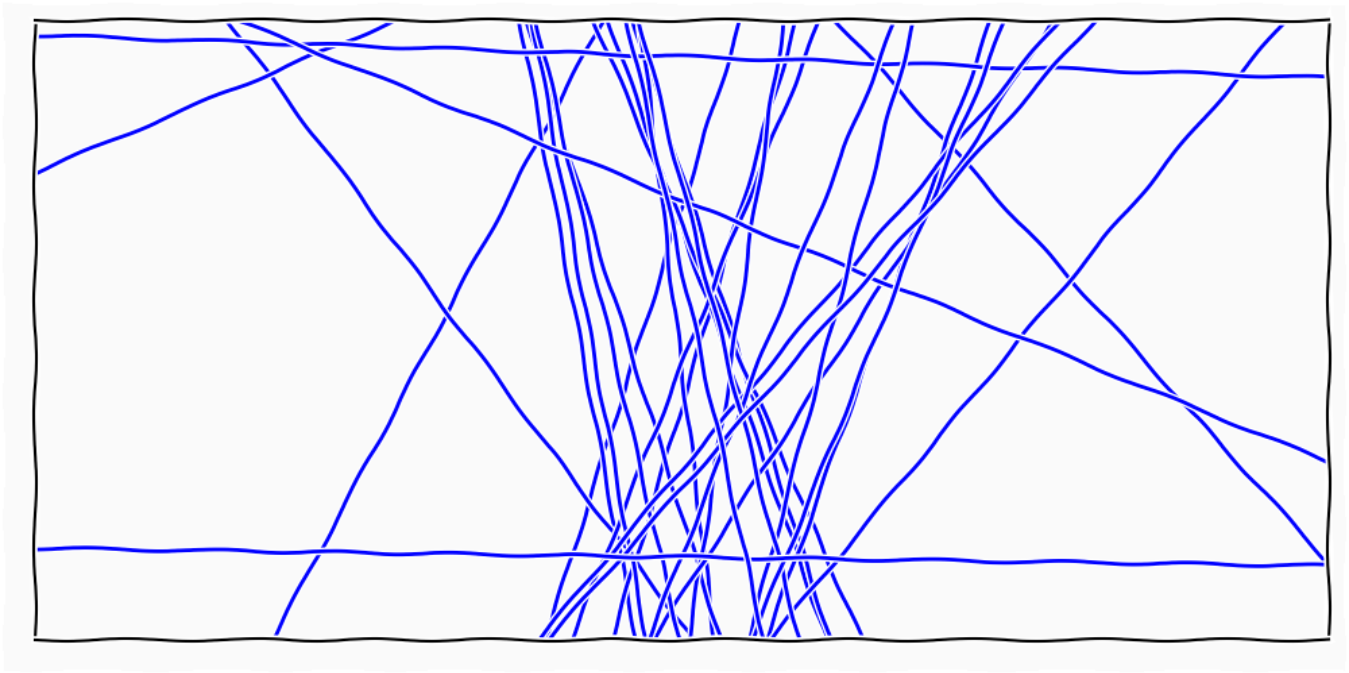
- Assumption Zero mean isotropic Gaussian

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$

- Posterior

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\beta \left(\alpha\mathbf{I} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1} \phi(\mathbf{X})^{\mathrm{T}}\mathbf{y}, \\ \left(\alpha\mathbf{I} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1})$$

Prior



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

Semantics

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

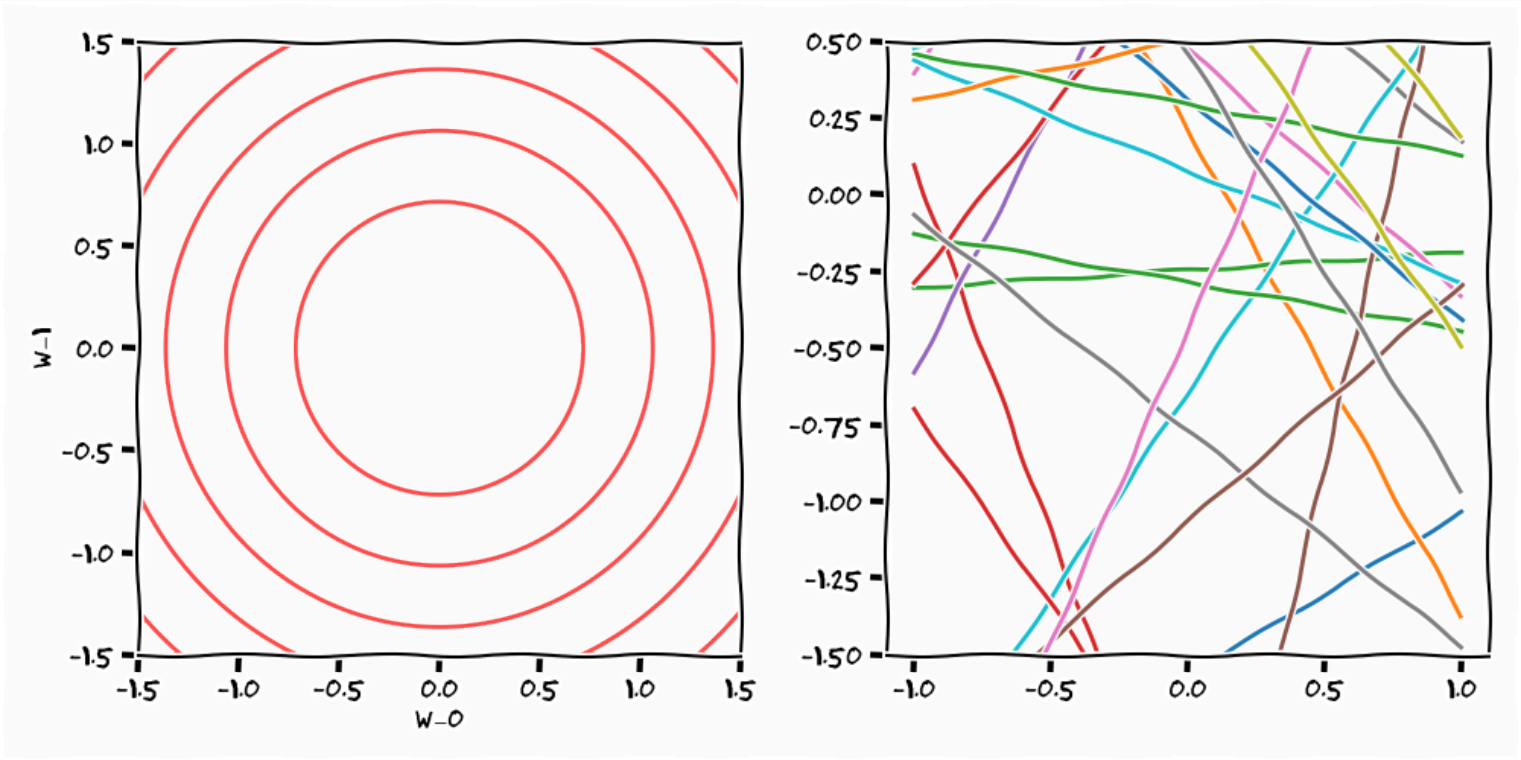
Likelihood How much evidence is there in the data for a specific hypothesis

Prior What are my beliefs about different hypothesis

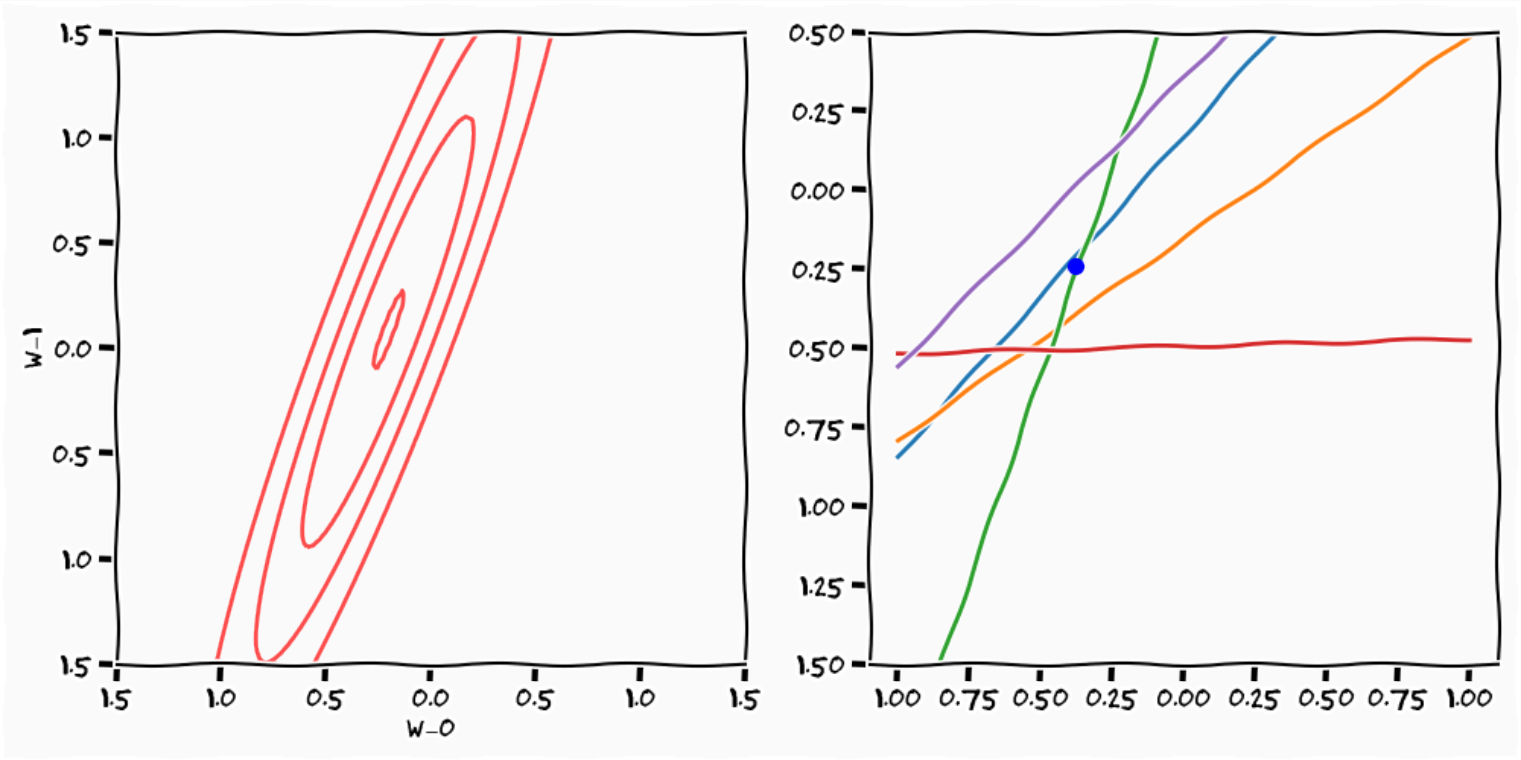
Posterior What is my updated belief after having seen data

Evidence What is my belief about any observations

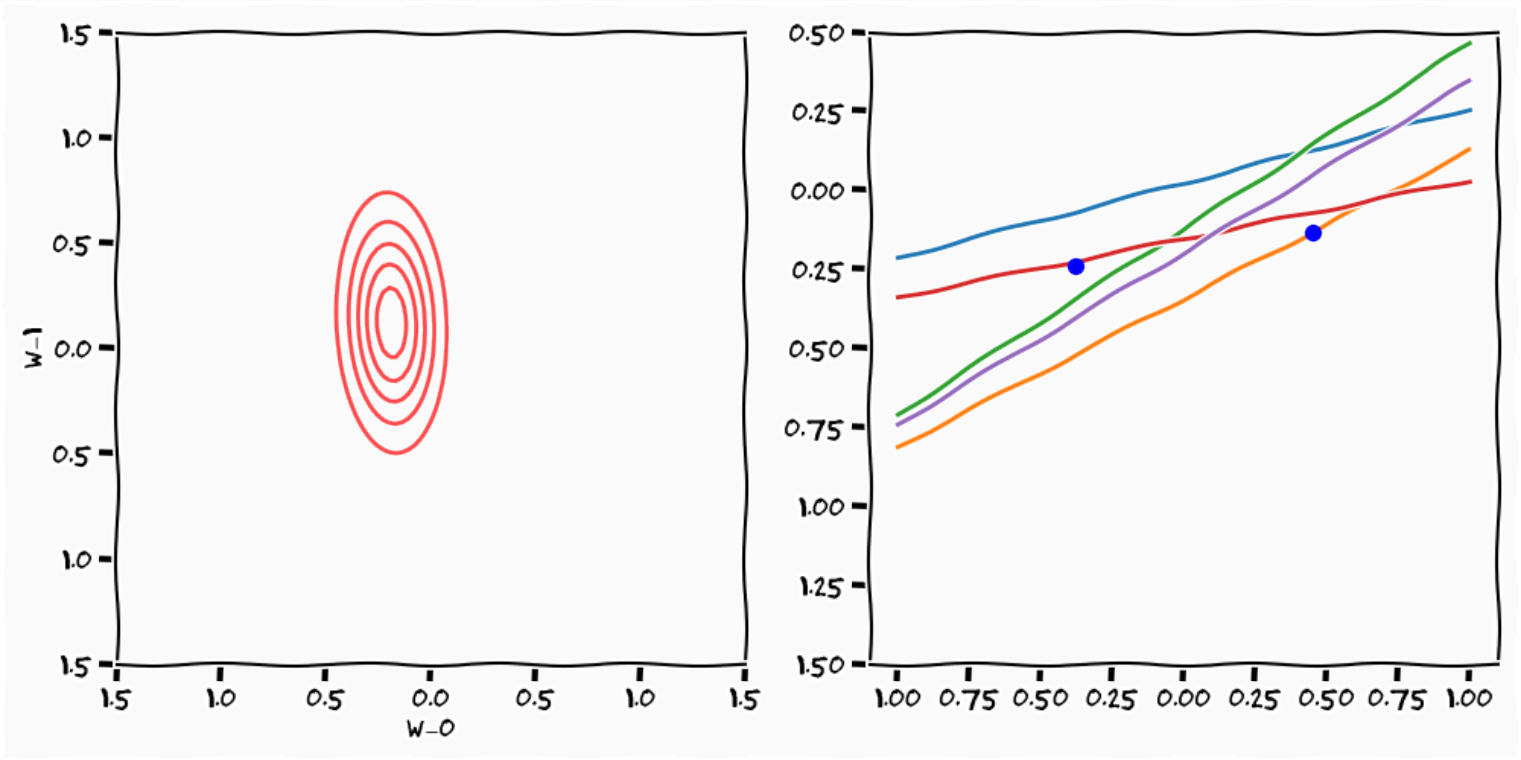
Linear Regression Example



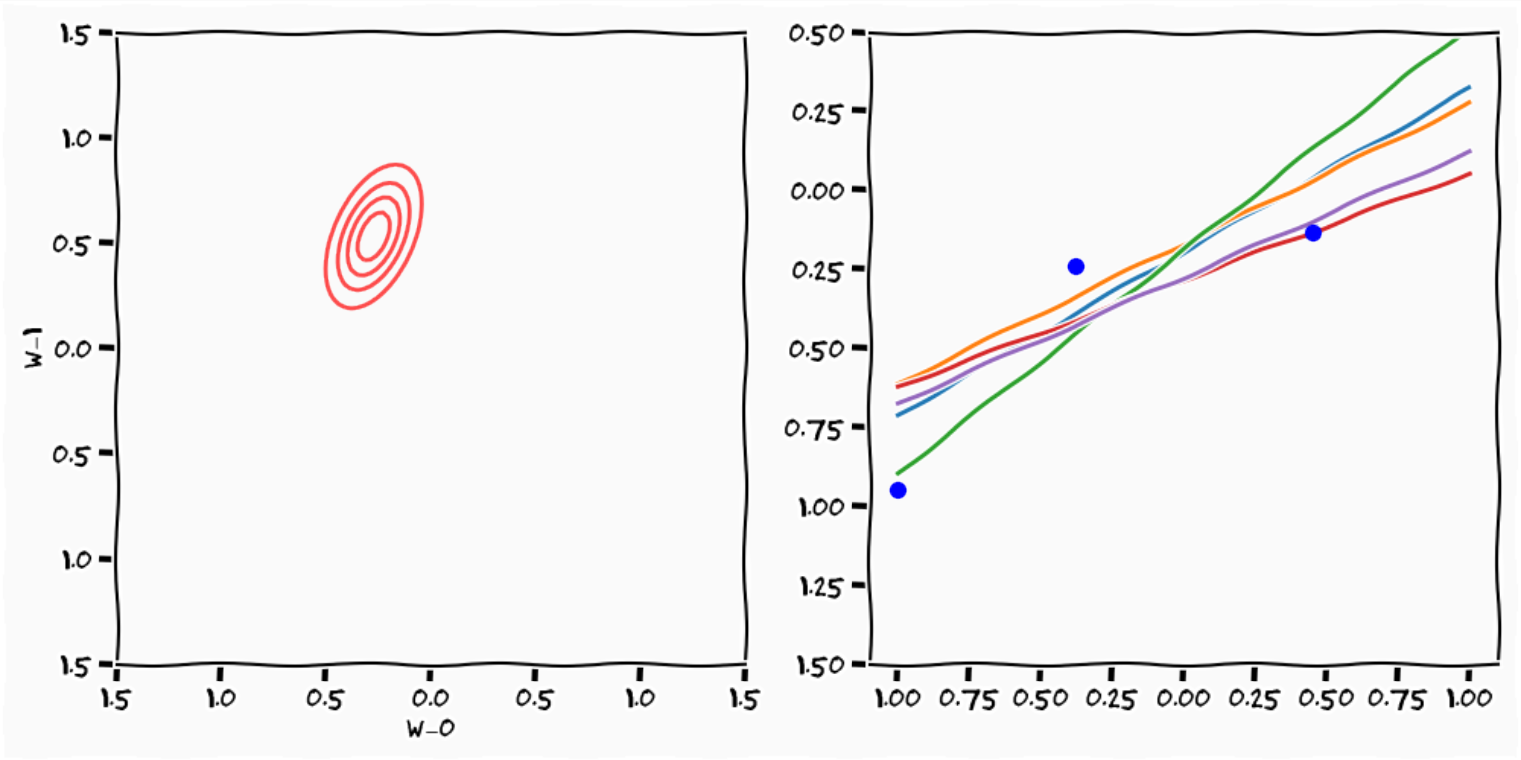
Linear Regression Example



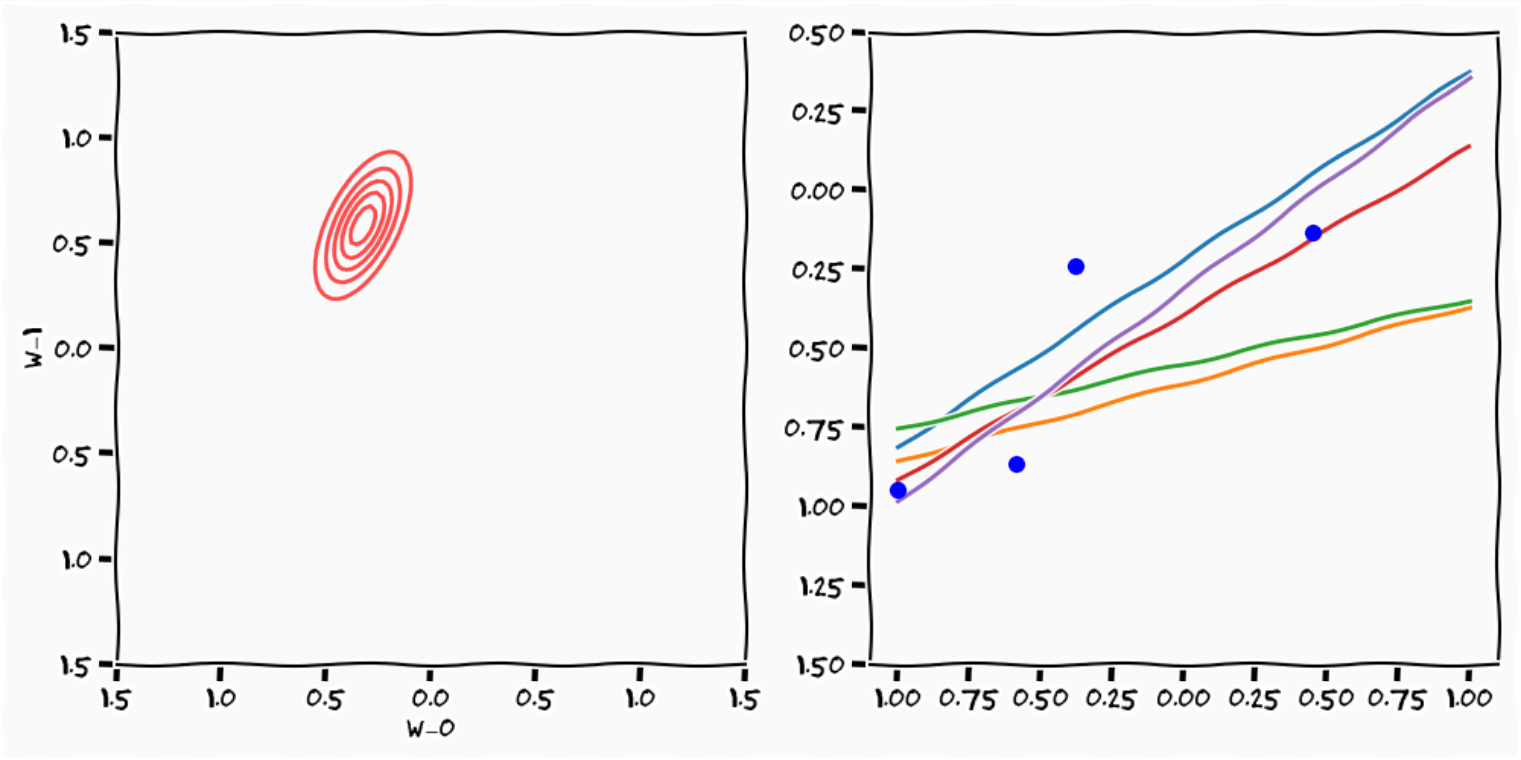
Linear Regression Example



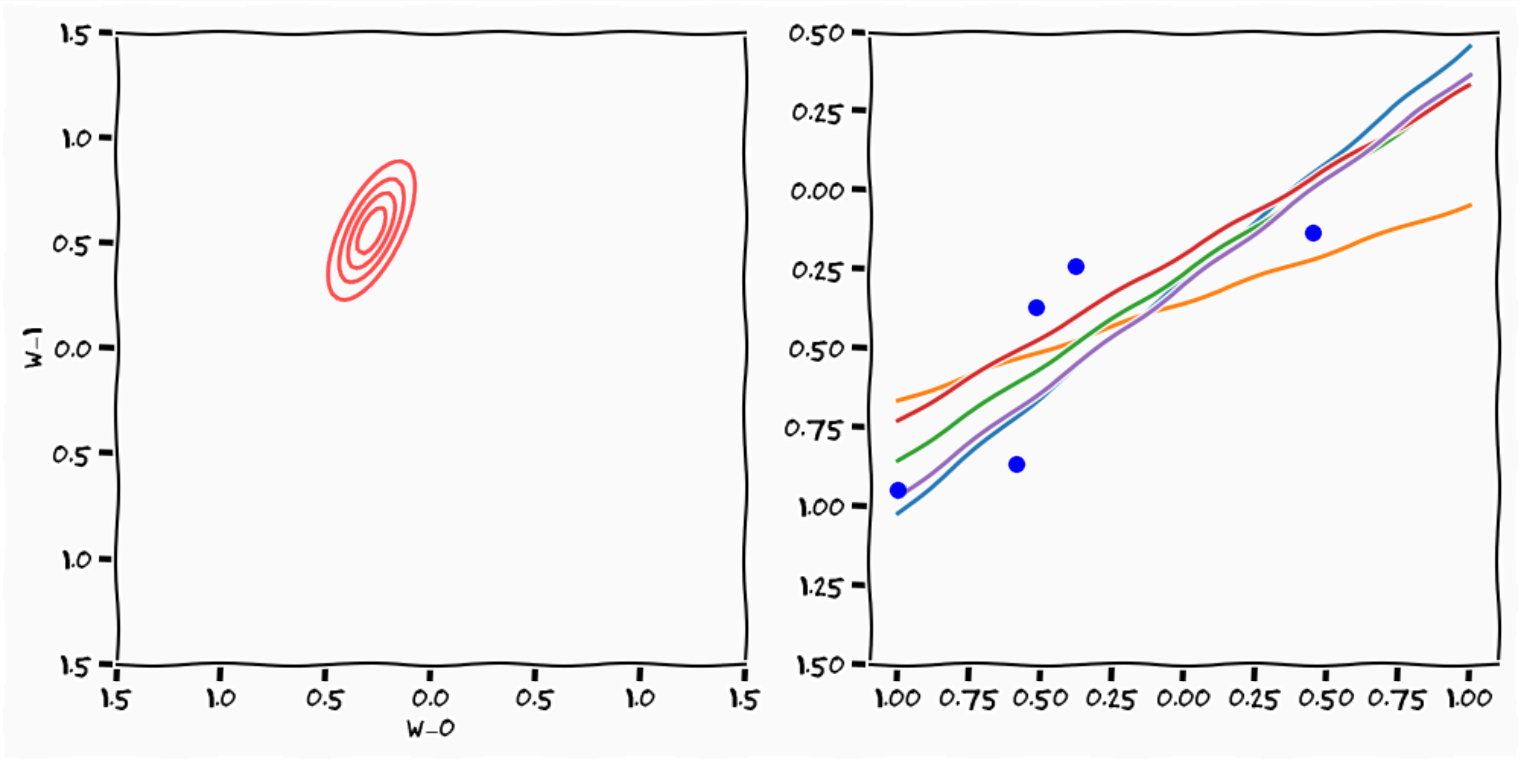
Linear Regression Example



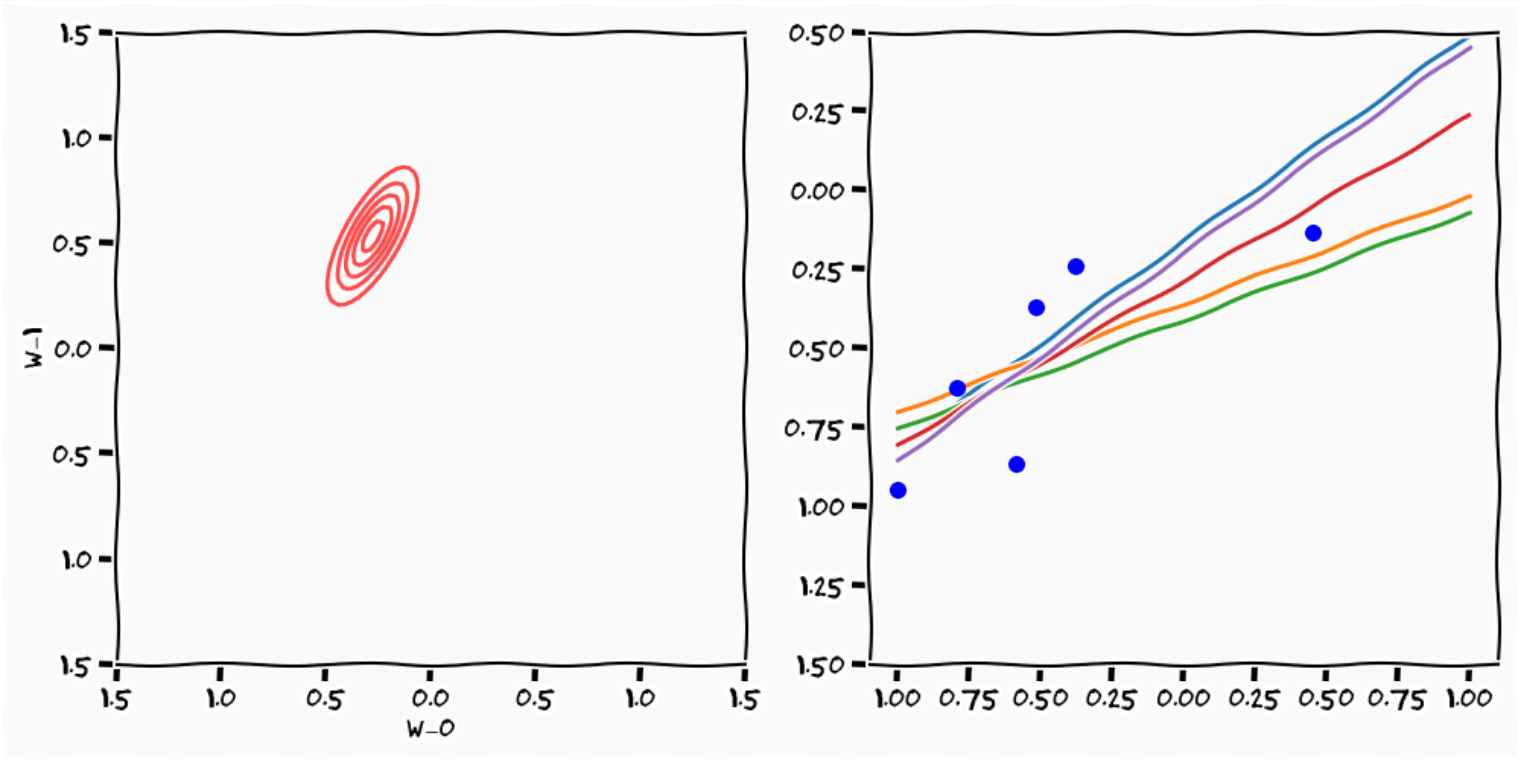
Linear Regression Example



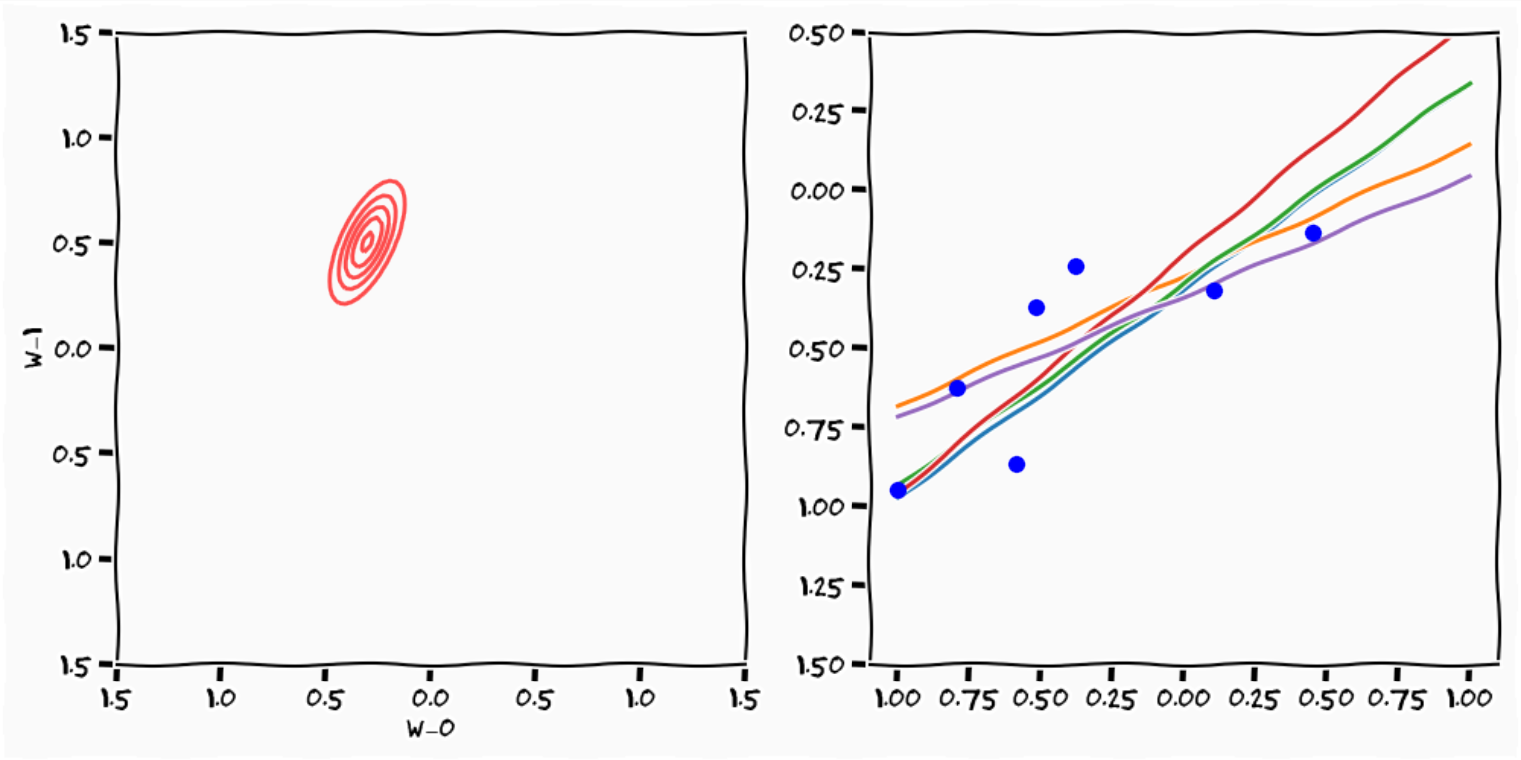
Linear Regression Example



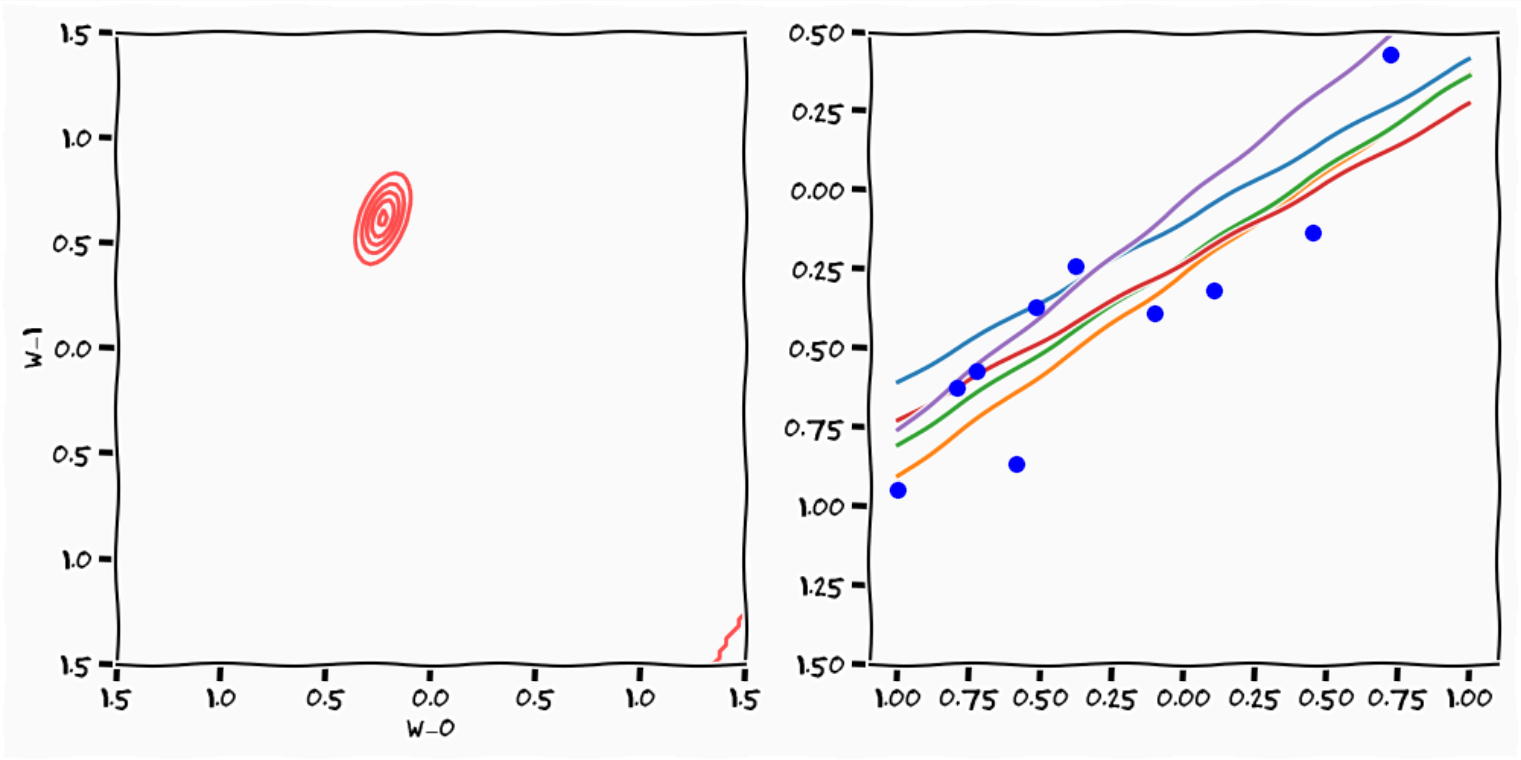
Linear Regression Example



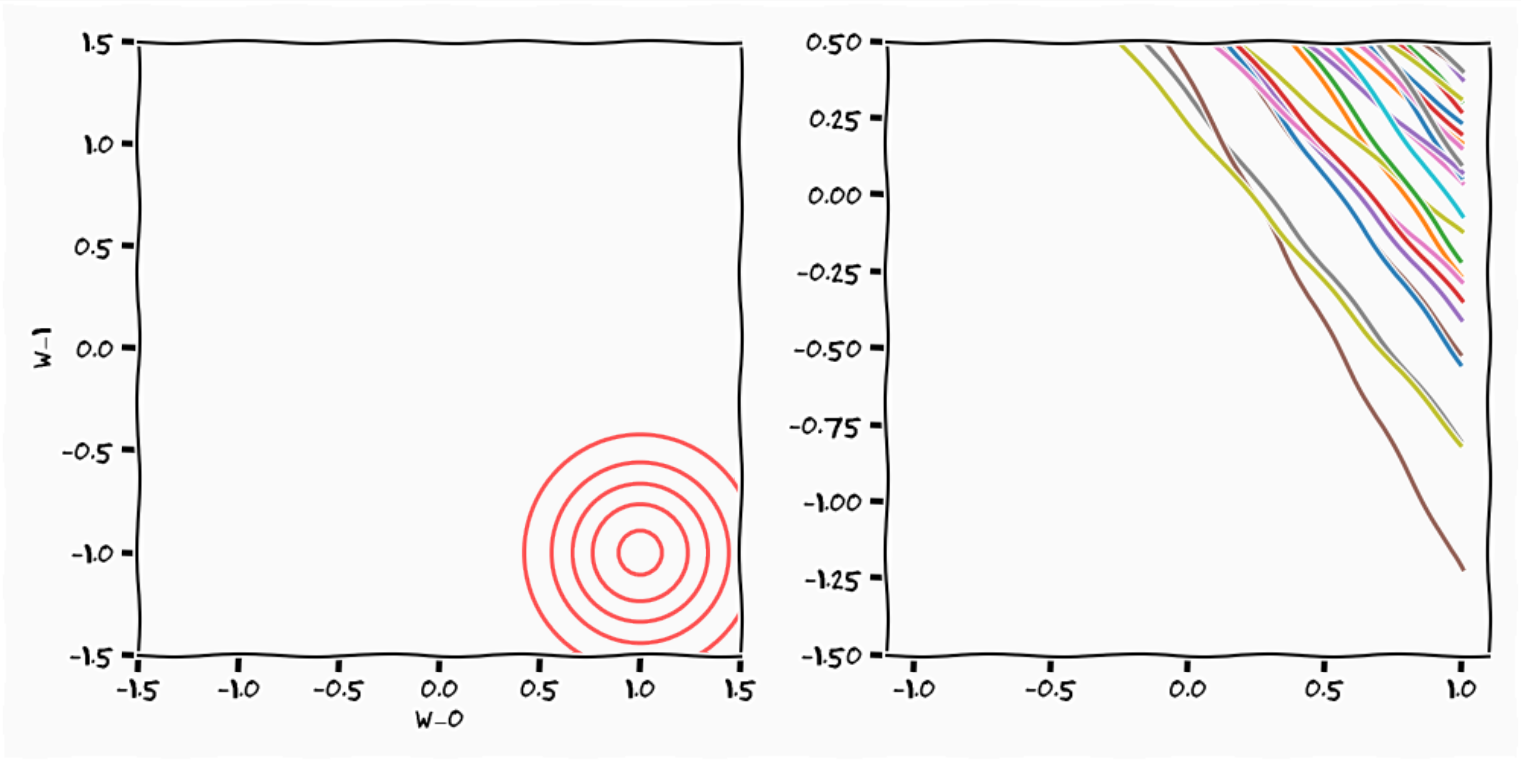
Linear Regression Example



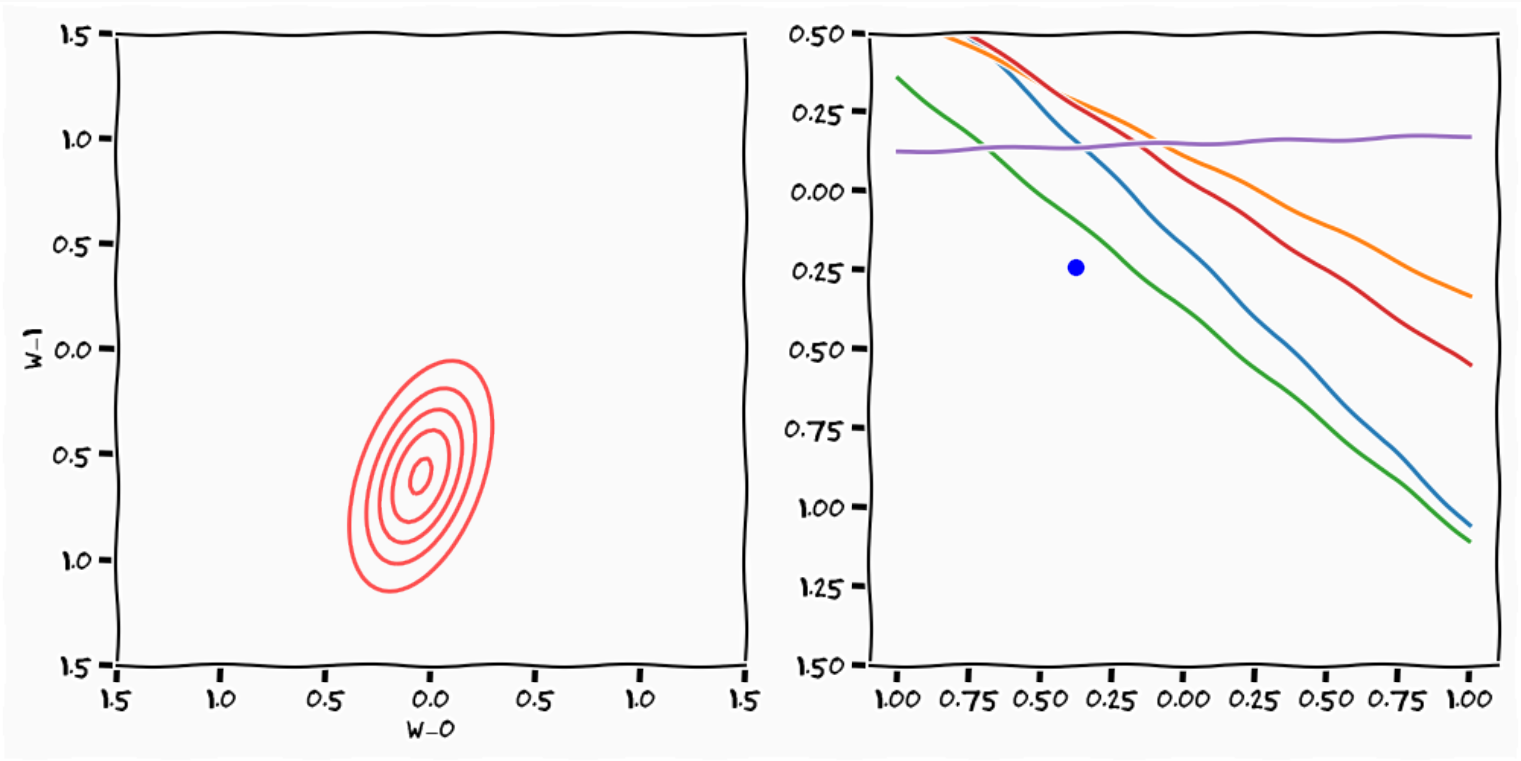
Linear Regression Example



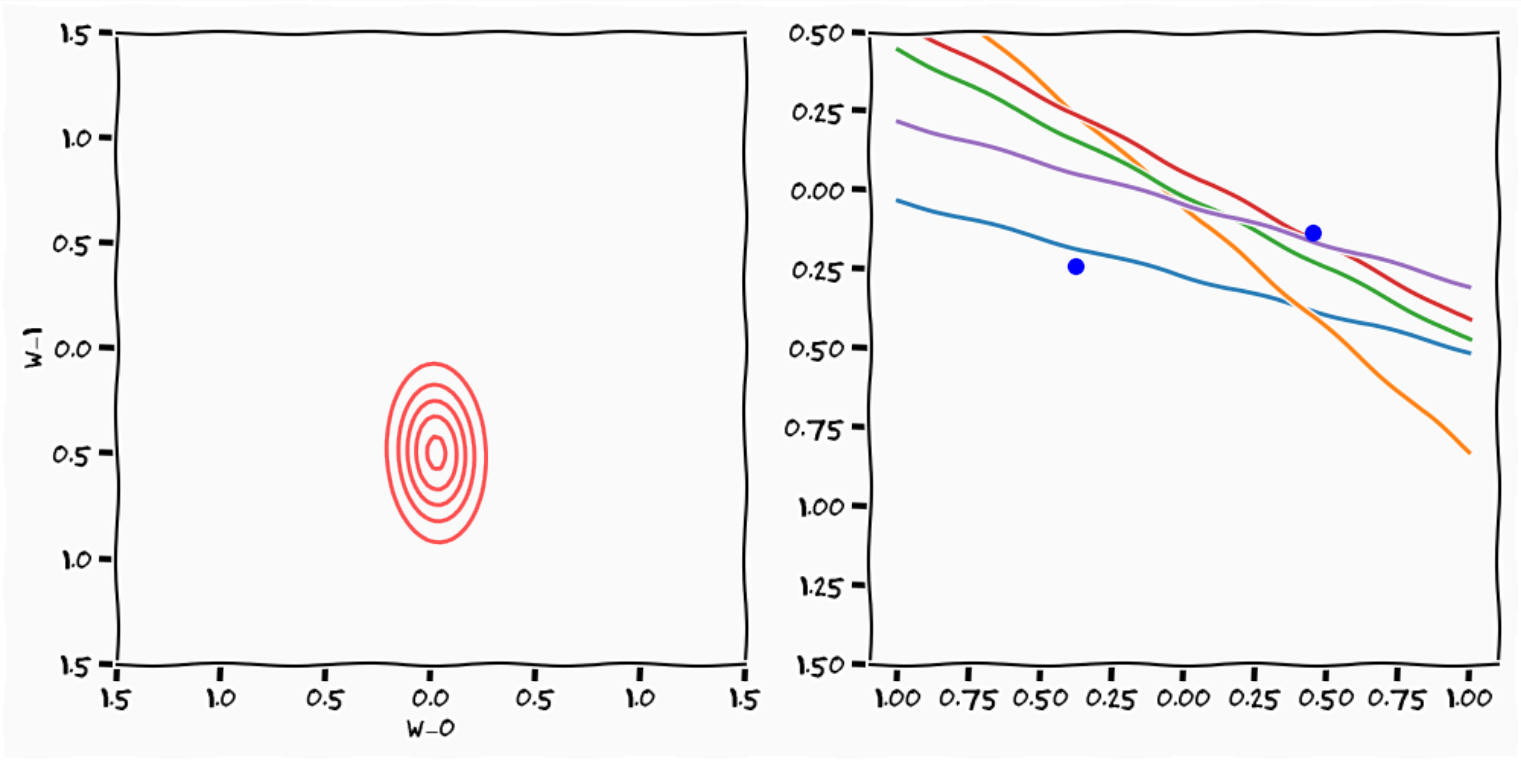
Linear Regression Example



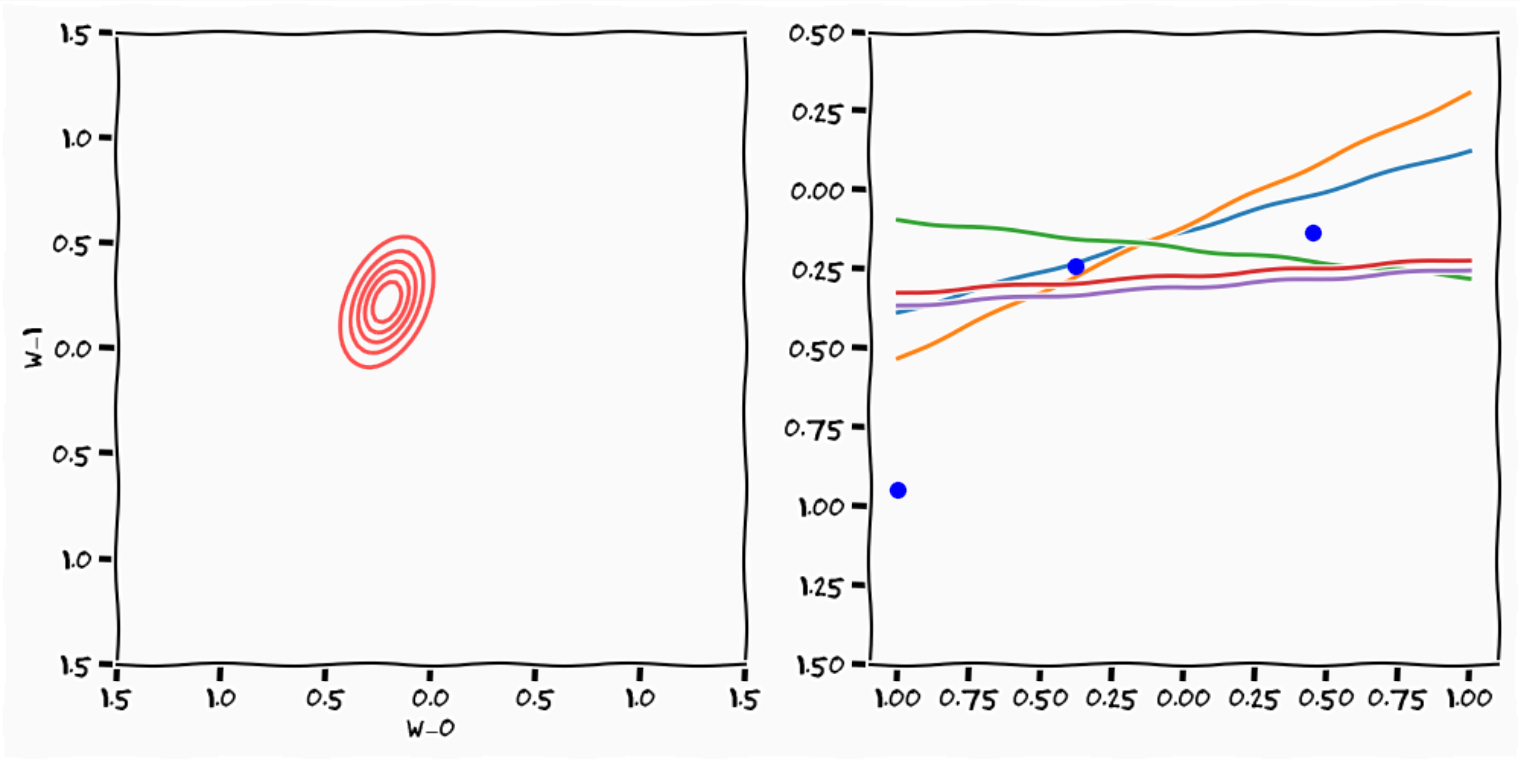
Linear Regression Example



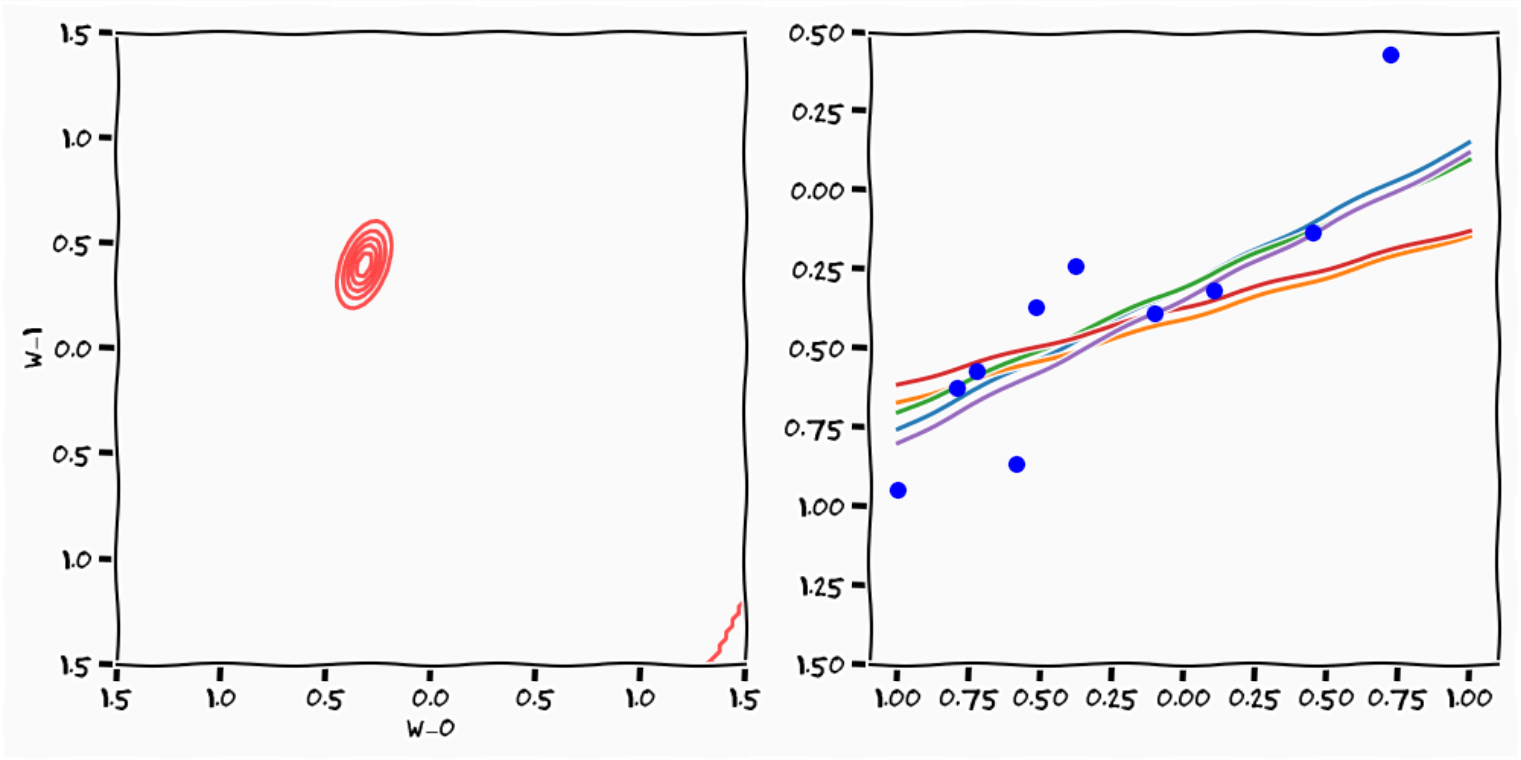
Linear Regression Example



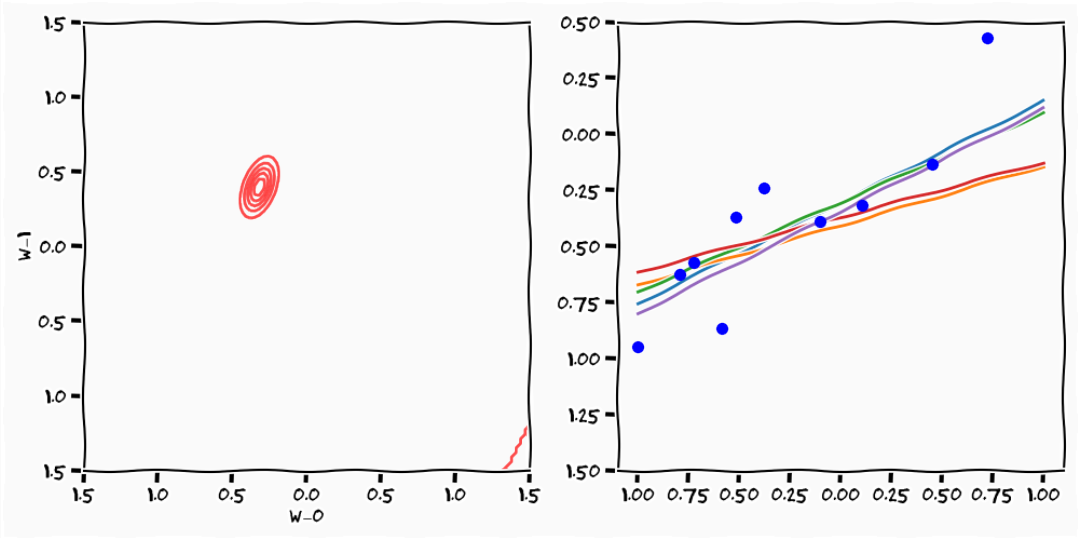
Linear Regression Example

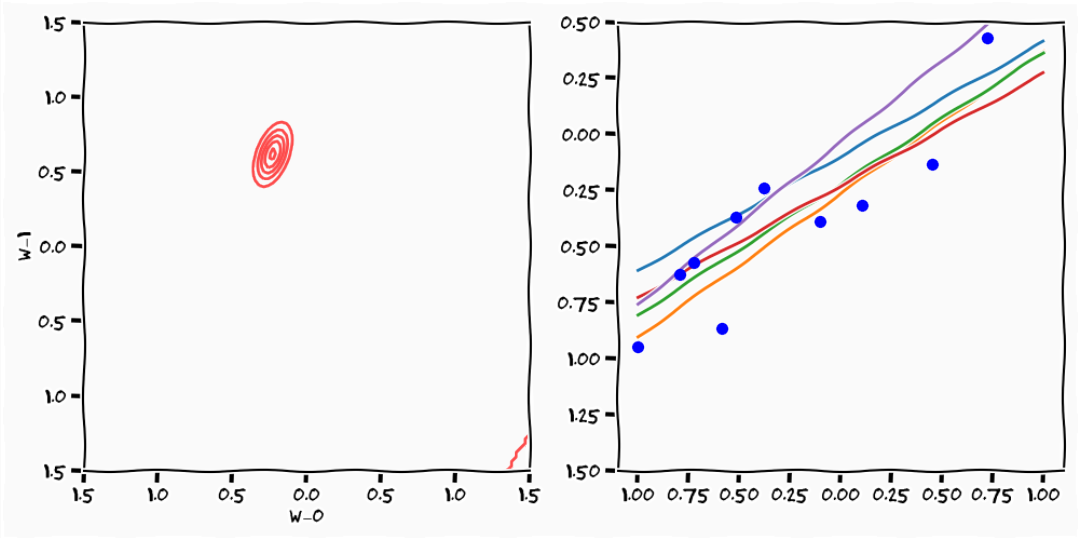


Linear Regression Example

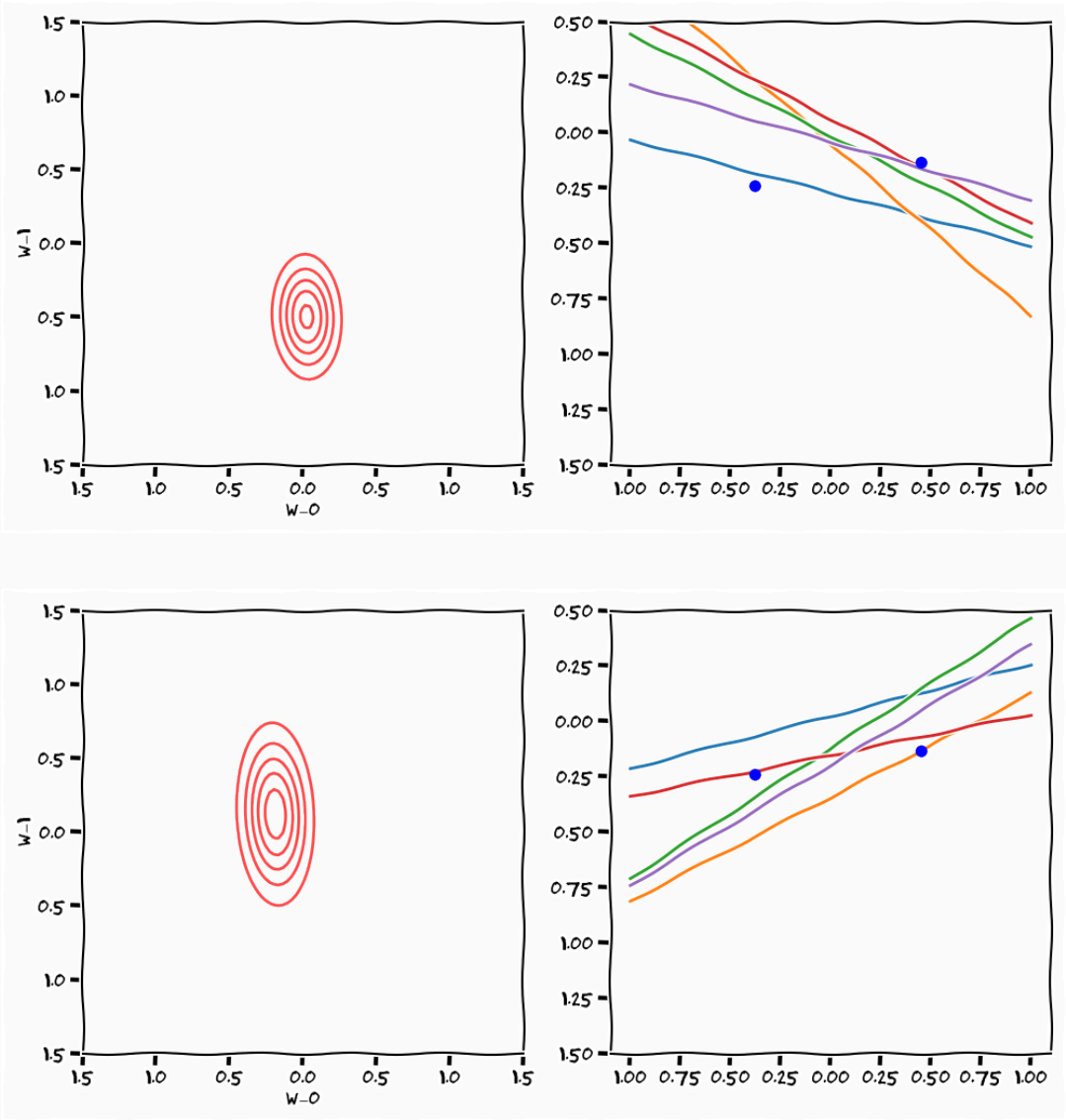


There is overwhelming evidence





Knowledge is Relative



Algorithmic Learning

- Do not underestimate what we just did

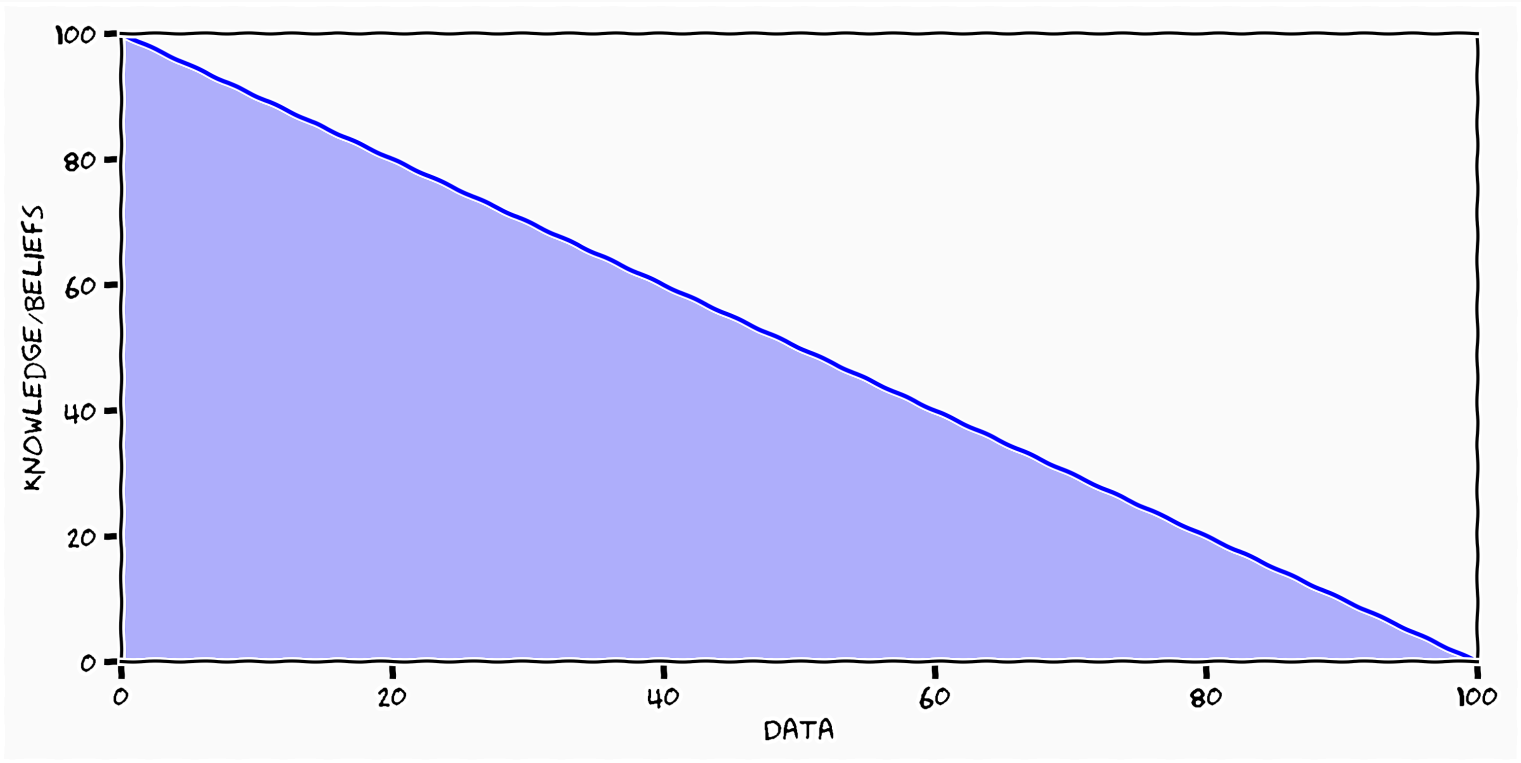
Algorithmic Learning

- Do not underestimate what we just did
  - We took knowledge
  - We took data
  - We created new knowledge

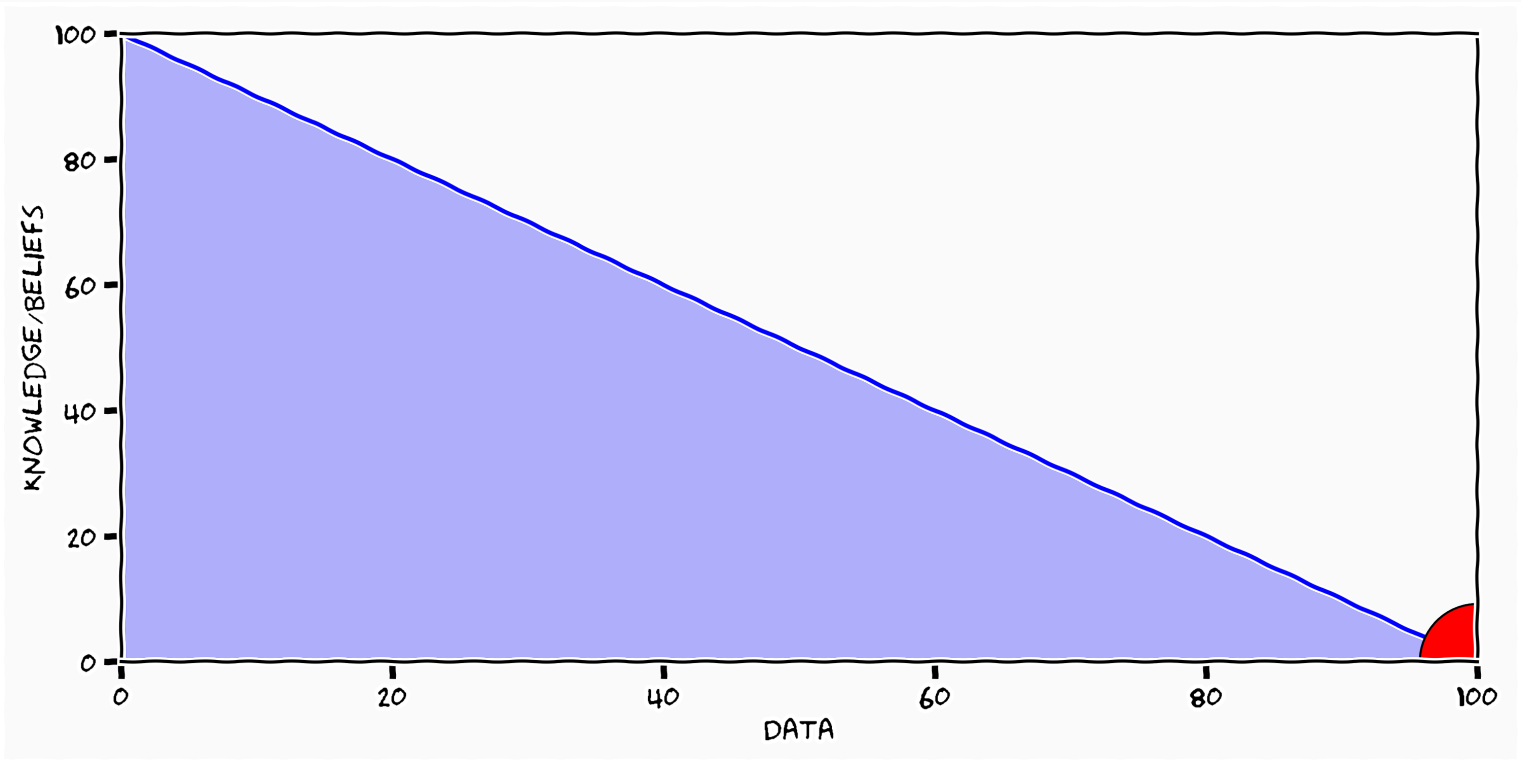
Algorithmic Learning

- Do not underestimate what we just did
  - We took knowledge
  - We took data
  - We created new knowledge
- This is it
  - everything from now on is just semantics

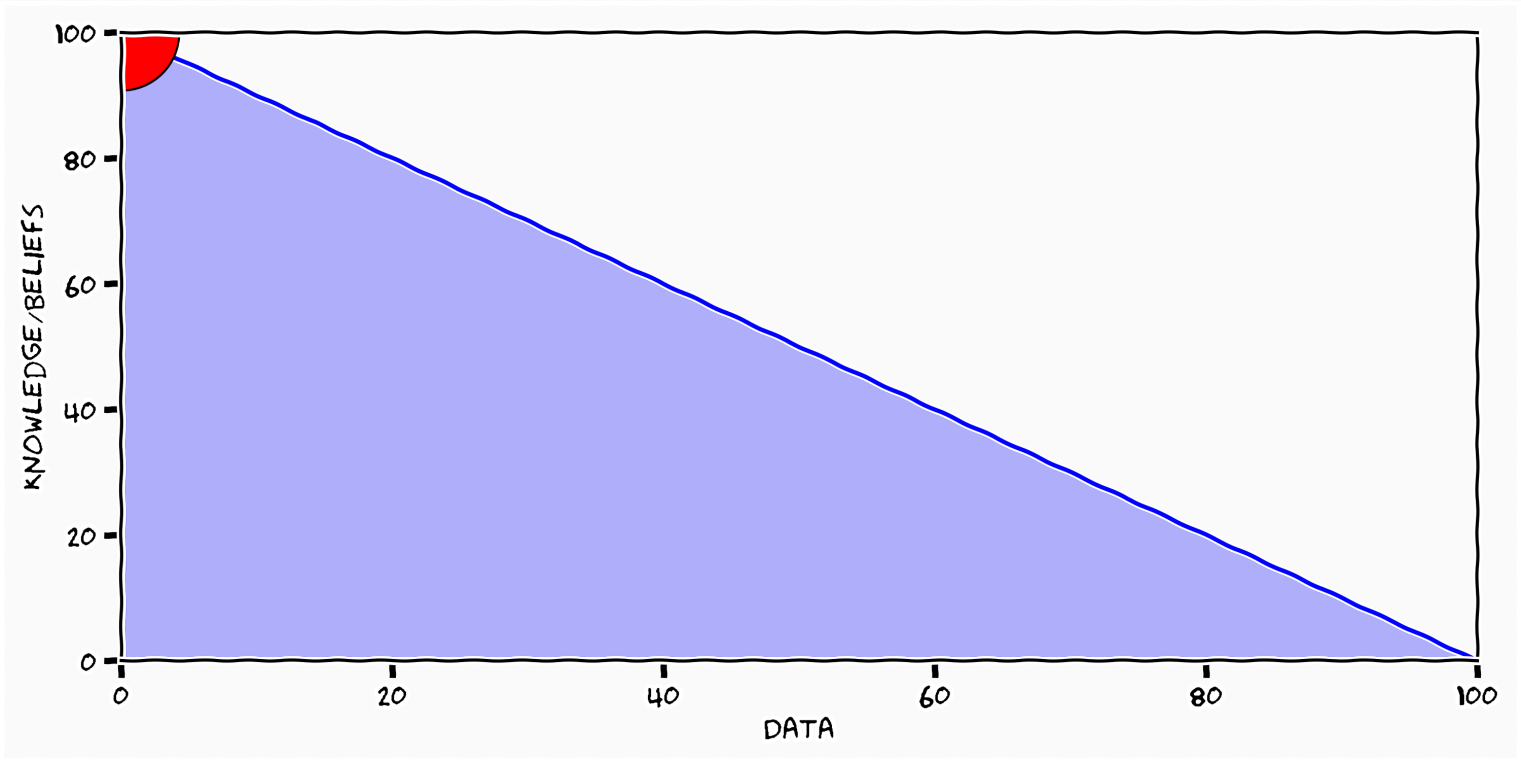
Data and Knowledge



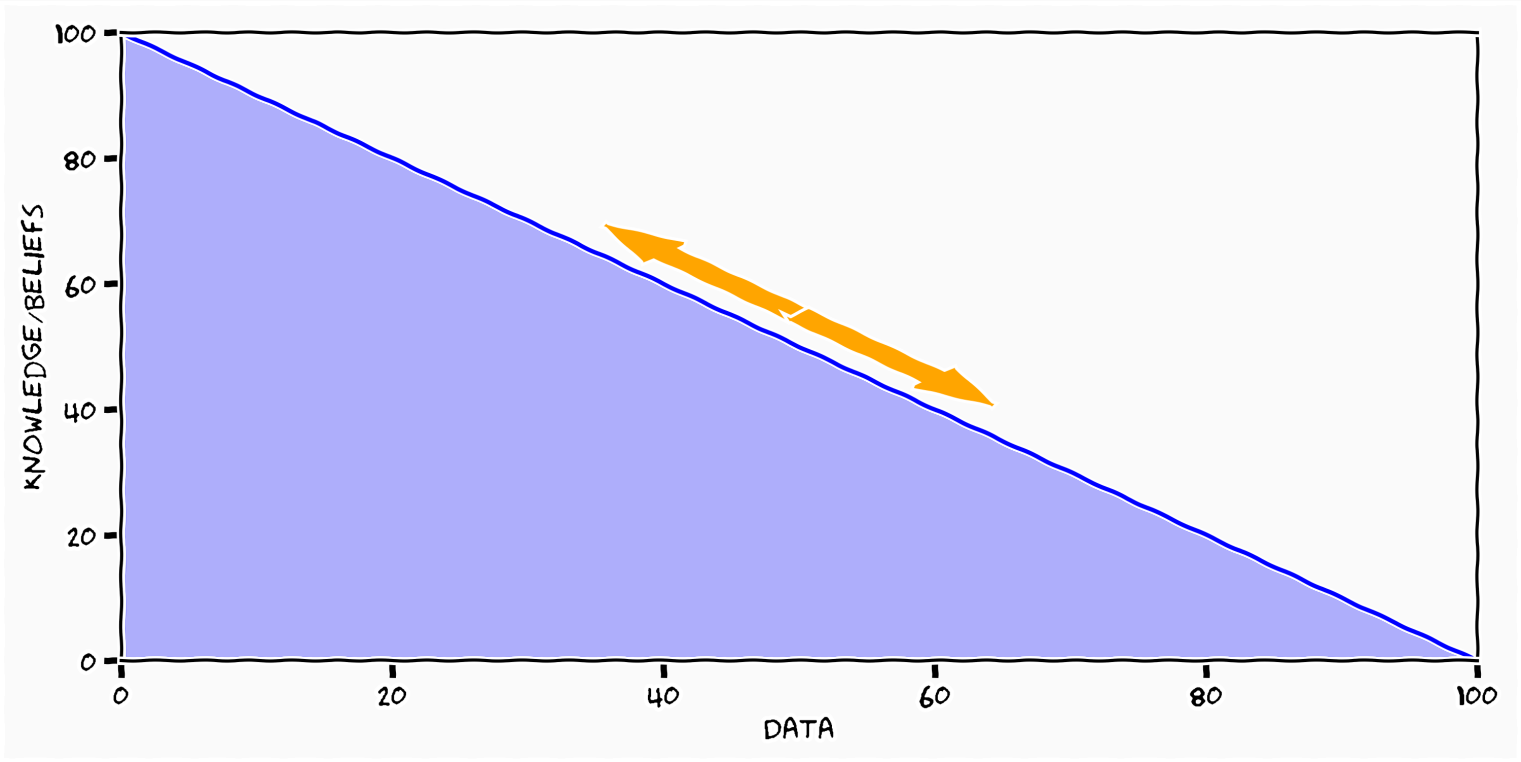
Data and Knowledge



Data and Knowledge



Data and Knowledge



Are beliefs objective?

NO of course not and therefore we all learn different things from the same data

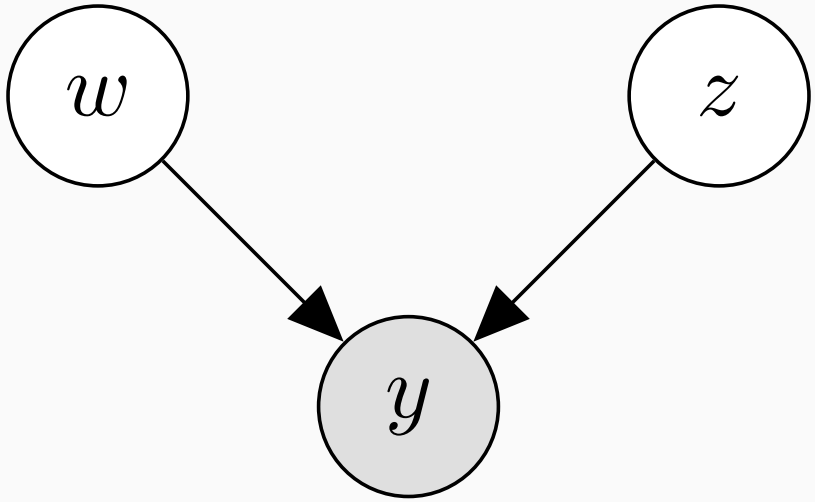
Are beliefs objective?

NO of course not and therefore we all learn different things from the same data

YES if two exact copies of the same "person" have different beliefs they cannot be the same person, therefore beliefs are objective, its only different amount of data/knowledge that generates differences

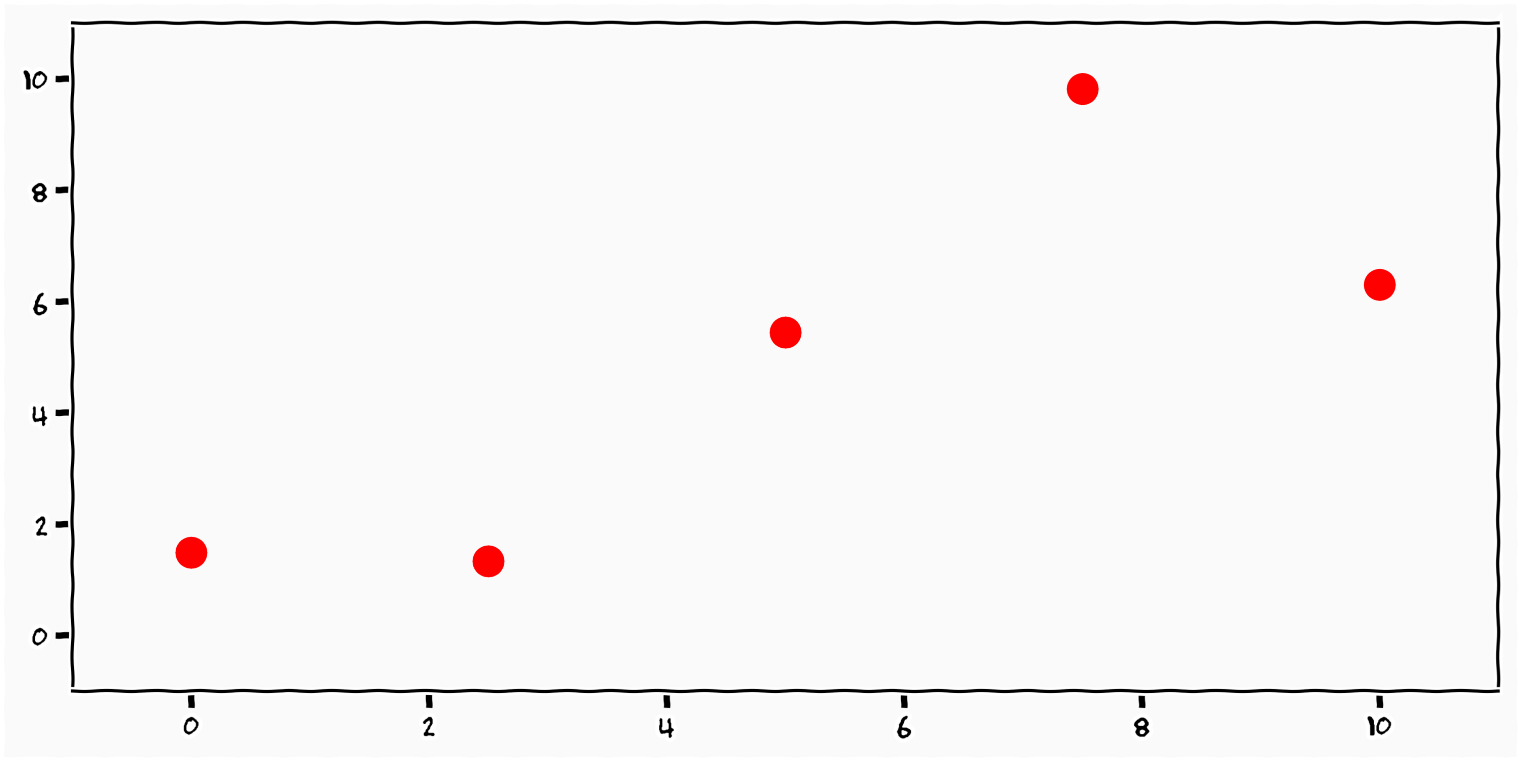
Building Models

Explaining Away



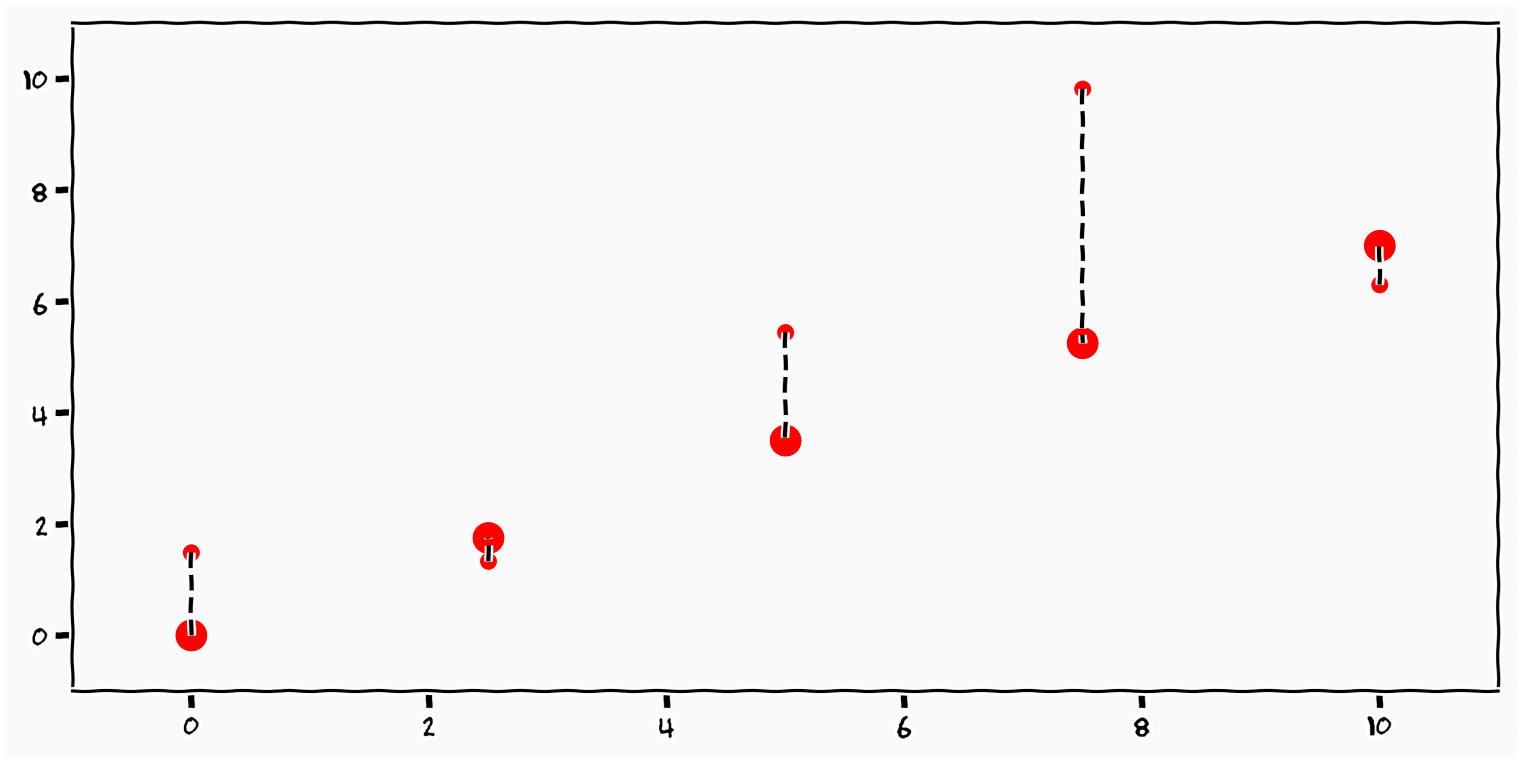
$$y = w \cdot x + z$$

Explaining Away



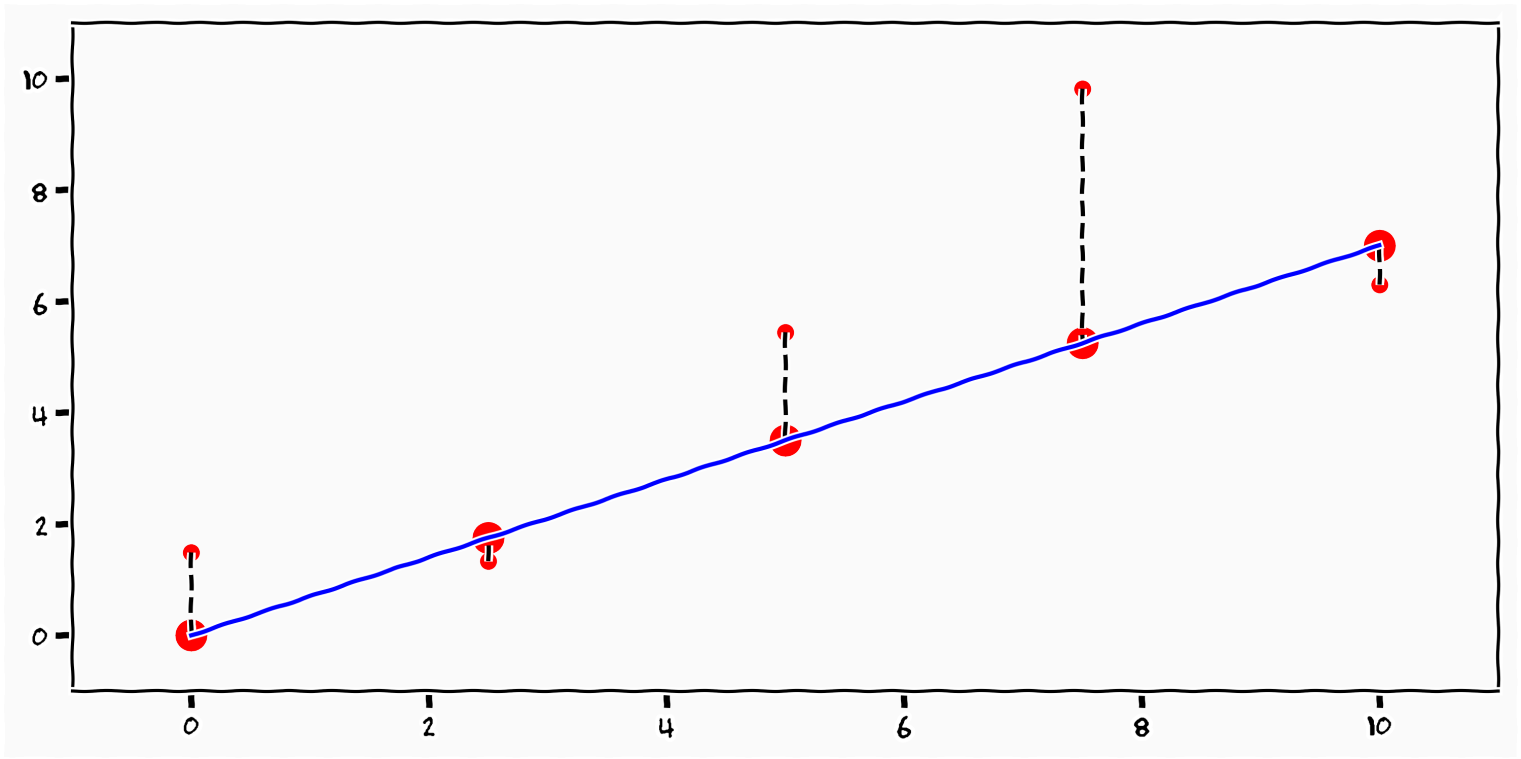
$$y = w \cdot x + z$$

Explaining Away



$$y - z = w \cdot x$$

Explaining Away



$$\tilde{y} = w \cdot x$$

Explaining Away

- When we build models we need to explain all variations in the data

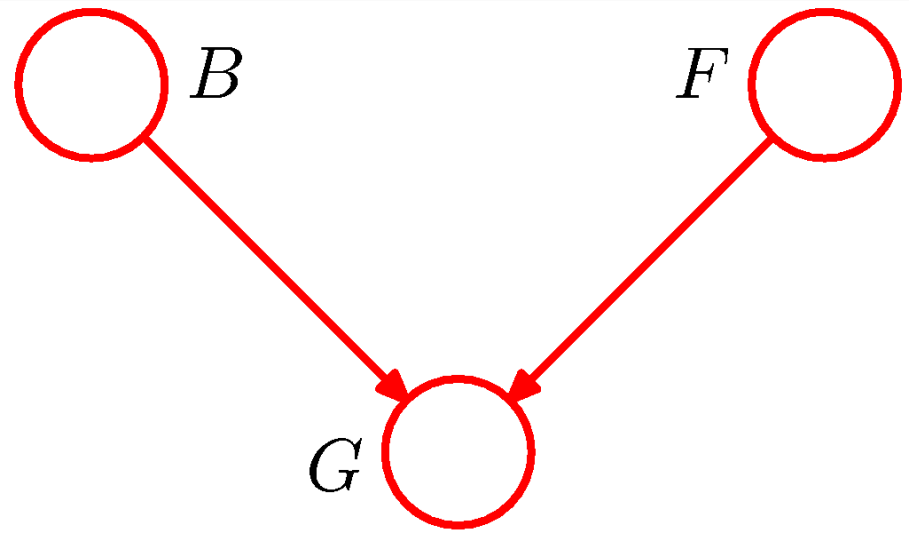
Explaining Away

- When we build models we need to explain all variations in the data
- We wish to explain away variations with variables to other give variables "meaning"

Explaining Away

- When we build models we need to explain all variations in the data
- We wish to explain away variations with variables to other give variables "meaning"
- Example to describe the orbit of a planet I need to *explain away* the variations due to measurement noise

Example p. 377 Bishop, 2006

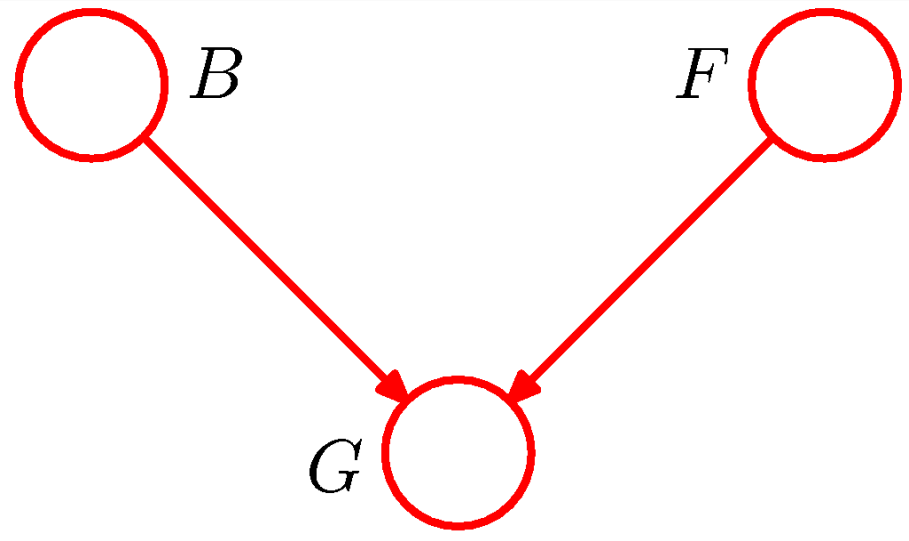


B Battery: $1 \rightarrow$ Full, $0 \rightarrow$ Empty

F Fuel Tank: $1 \rightarrow$ Full, $0 \rightarrow$ Empty

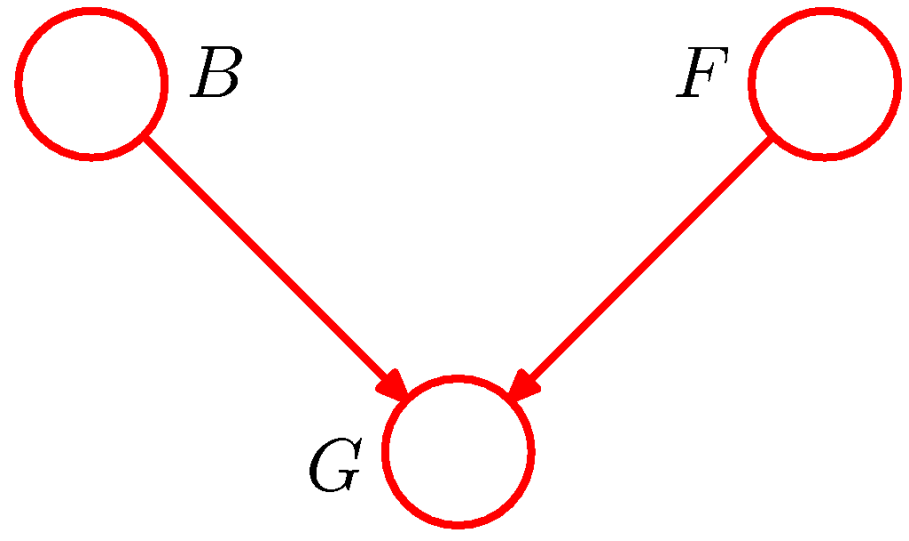
G Fuel Gauge: $1 \rightarrow$ Indicates Full, $0 \rightarrow$ Indicates Empty

Example p. 377 Bishop, 2006



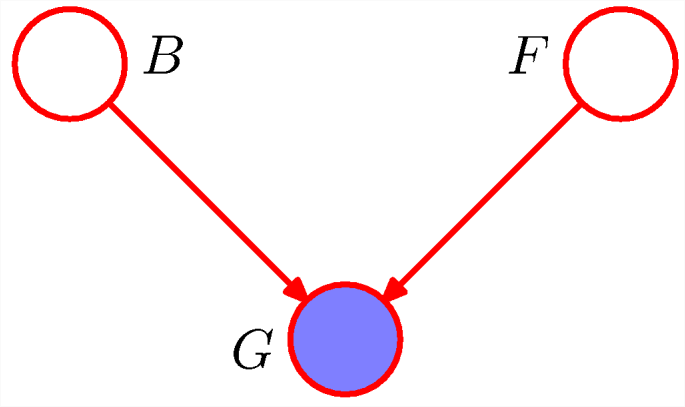
$$p(B = 1) = 0.9$$
$$p(F = 1) = 0.9$$

Example p. 377 Bishop, 2006



$$p(G = 1|B = 1, F = 1) = 0.8$$
$$p(G = 1|B = 1, F = 0) = 0.2$$
$$p(G = 1|B = 0, F = 1) = 0.2$$
$$p(G = 1|B = 0, F = 0) = 0.1$$

Example p. 377 Bishop, 2006



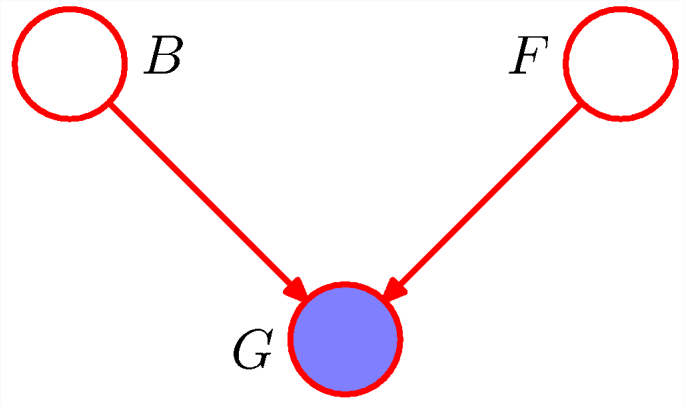
We observe an empty fuel tank $G = 0$

$$p(G = 0) = \sum_{B \in \{0,1\}} \sum_{F \in \{0,1\}} p(G = 0|B, F)p(B)p(F) = 0.315$$

$$p(G = 0|F = 0) = \sum_{B \in \{0,1\}} p(G = 0|B, F = 0)p(B) = 0.81$$

$$p(F = 0|G = 0) = \frac{p(G = 0|F = 0)p(F = 0)}{p(G = 0)} \approx 0.257$$

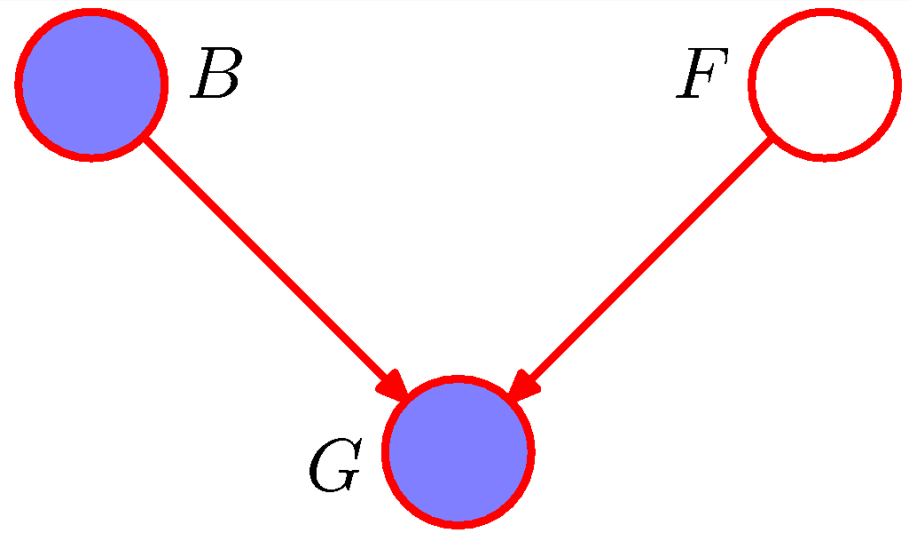
Example p. 377 Bishop, 2006



$$p(F = 0|G = 0) > p(F = 0)$$

The gauge does provide information about the tank

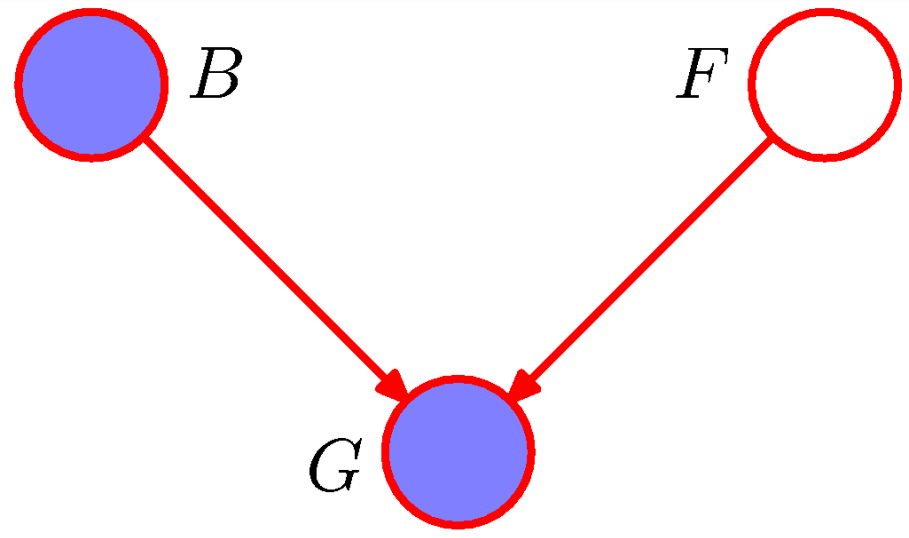
Example p. 377 Bishop, 2006



We observe an empty fuel tank $G = 0$ and Battery empty $B = 0$

$$p(F = 0|G = 0, B = 0) = \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F \in \{0,1\}} p(G = 0|B = 0, F)p(F)} \approx 0.111$$

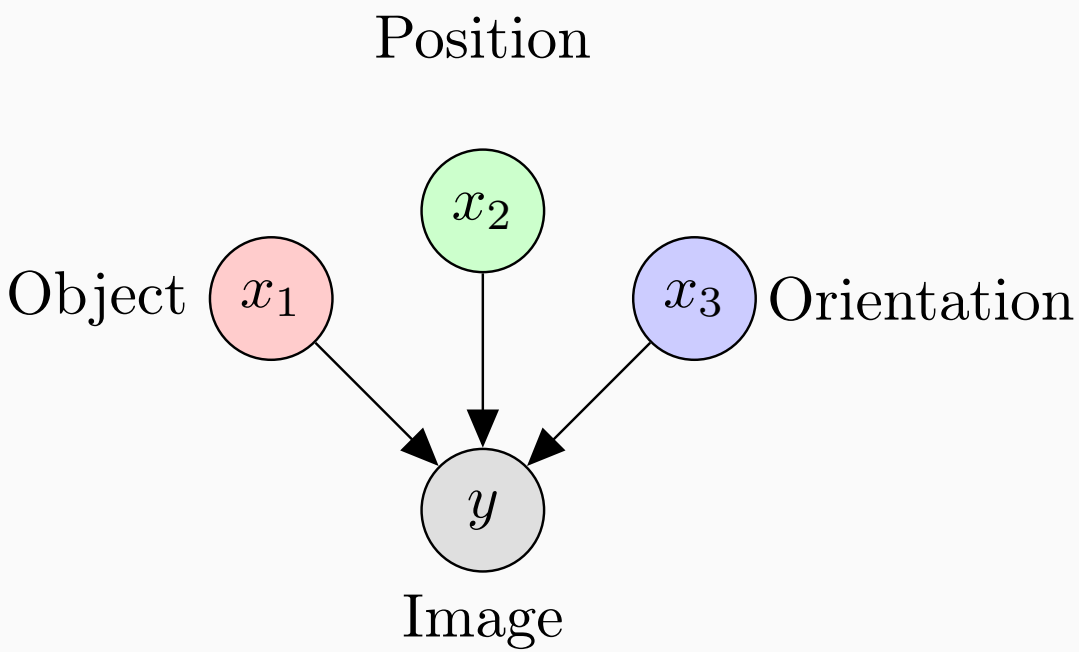
Example p. 377 Bishop, 2006



$$p(F = 0|G = 0) > p(F = 0|G = 0, B = 0) > P(F = 0)$$

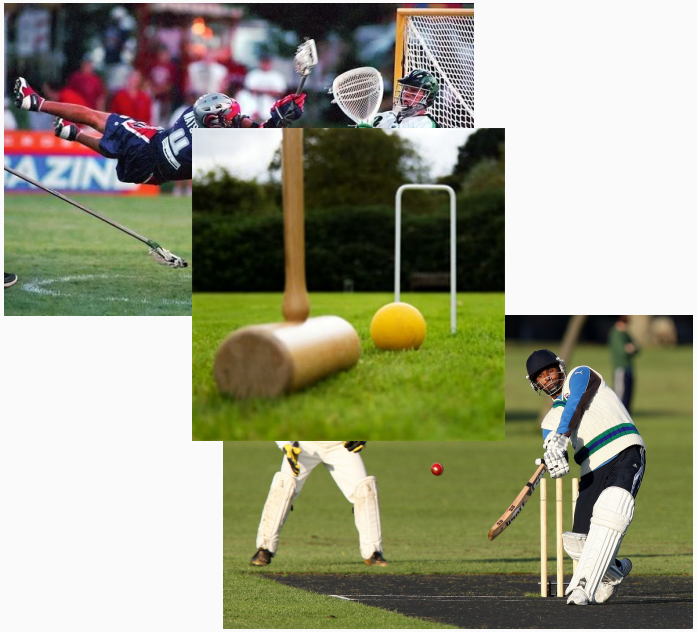
Knowing that the battery is empty explains away the information about the Gauge indicating empty

Generative Model

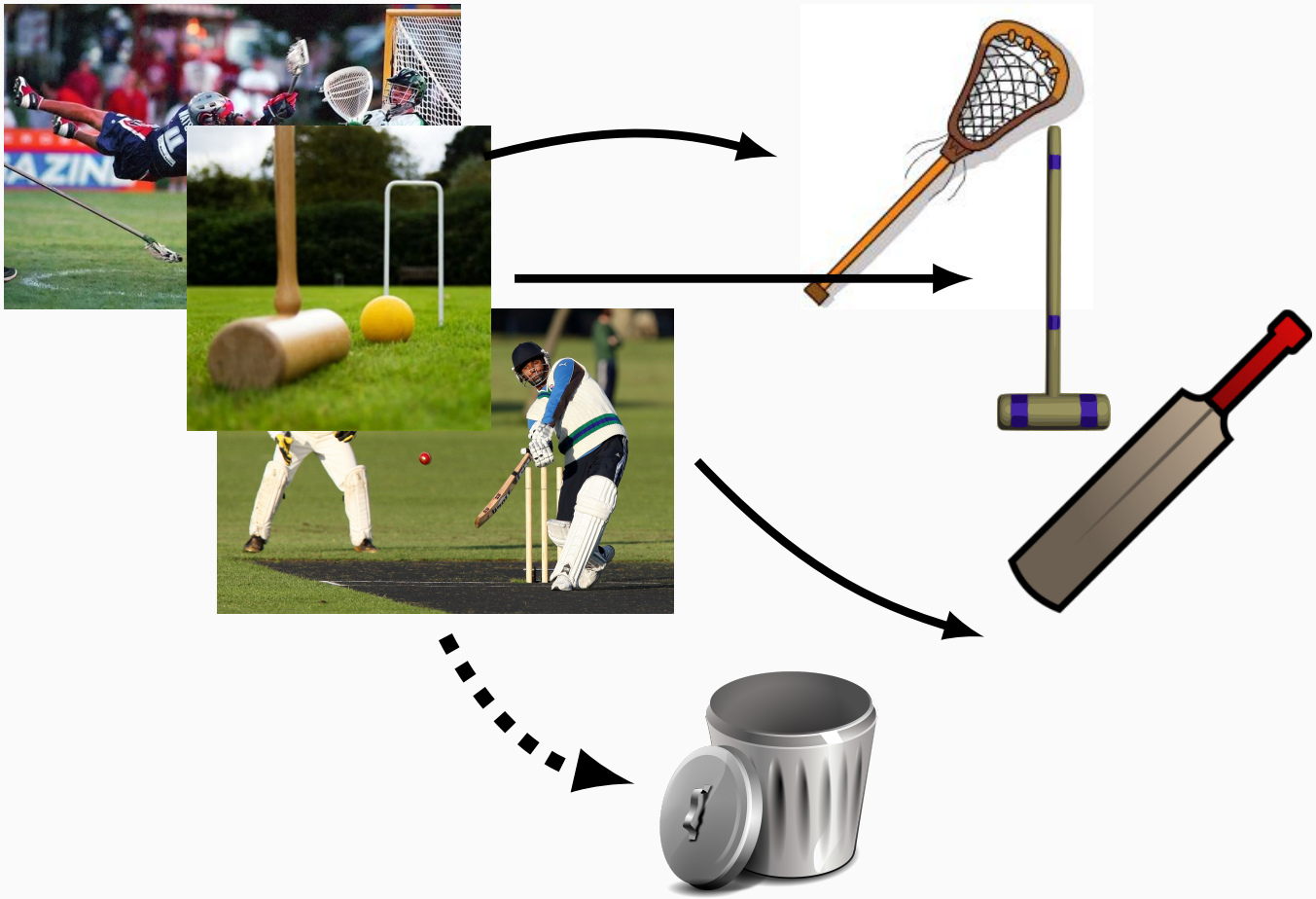


- The Object variable *explains away* variance associated with objects from the image
  - $\rightarrow$ position won't contain object variations
  - $\rightarrow$ orientation won't contain object variations

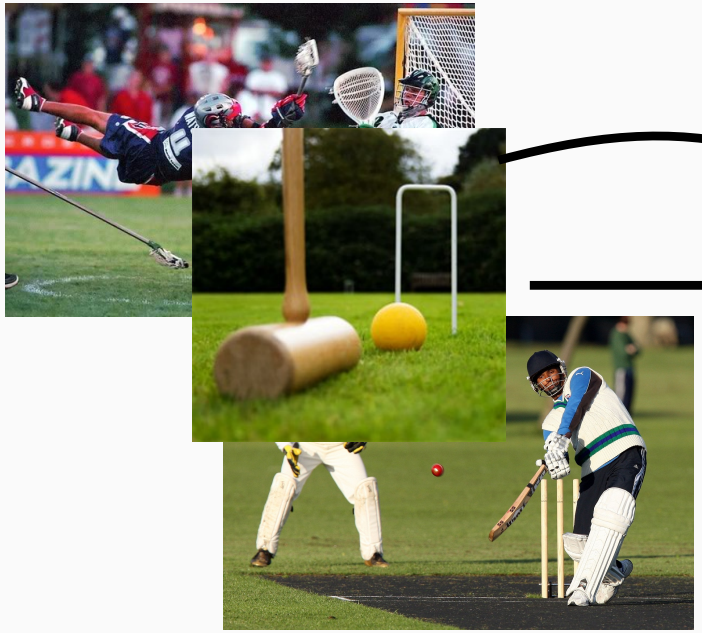
Why we build models

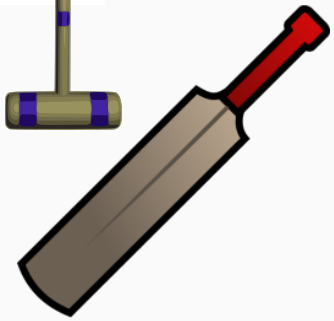


Why we build models



Why we build models



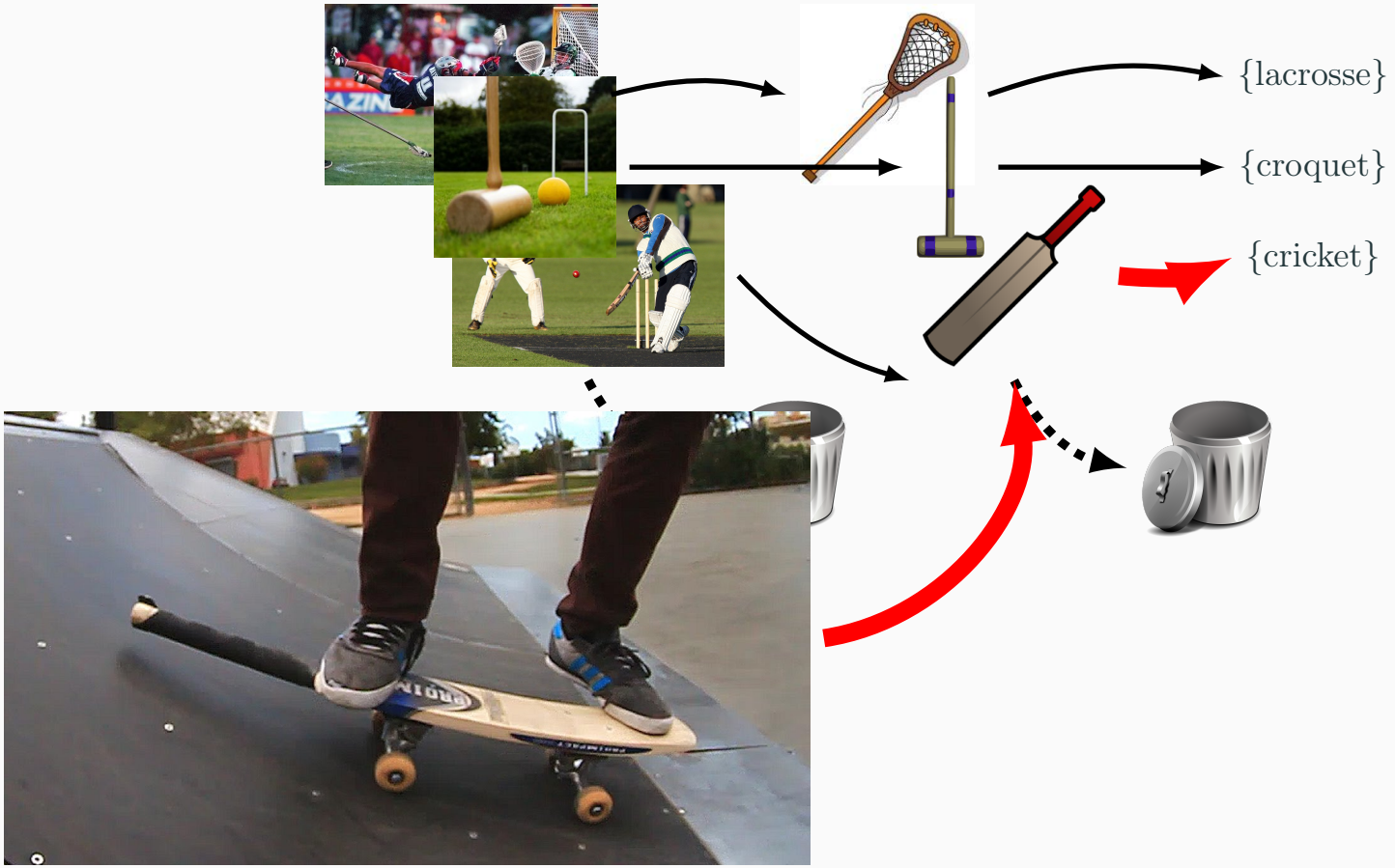


{lacrosse}

{croquet}

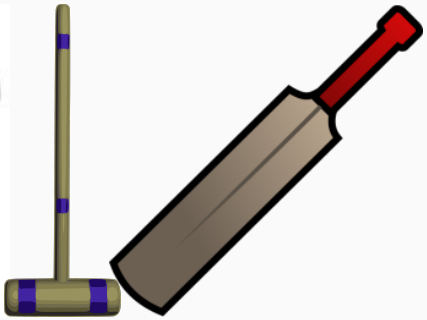
{cricket}

Why we build models



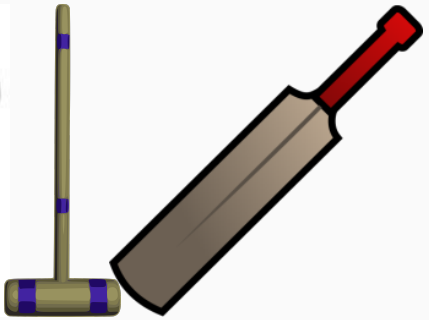
Why we build models

{lacrosse}
{croquet}
{cricket}



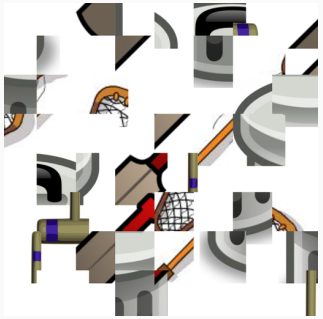
Why we build models

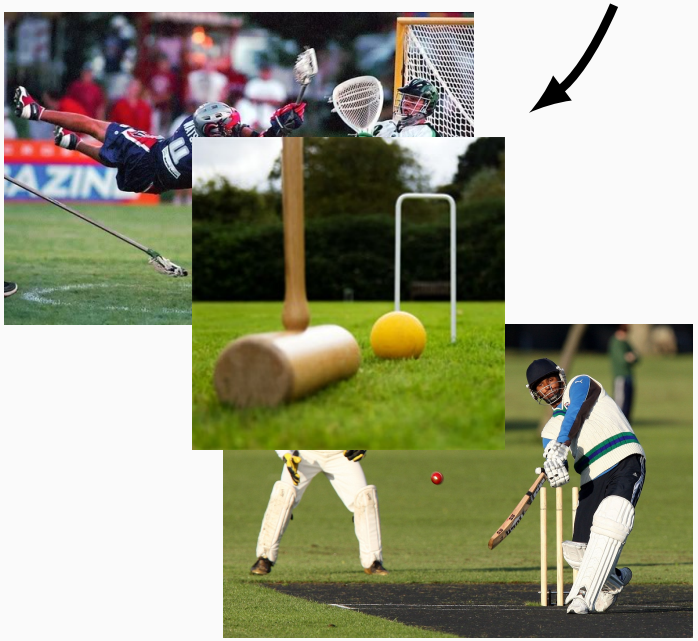
{lacrosse}
{croquet}
{cricket}



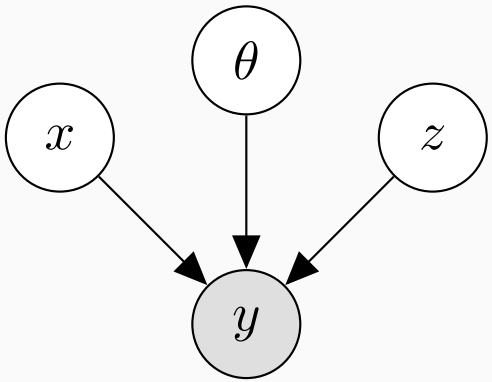
Why we build models

{lacrosse}
{croquet}
{cricket}

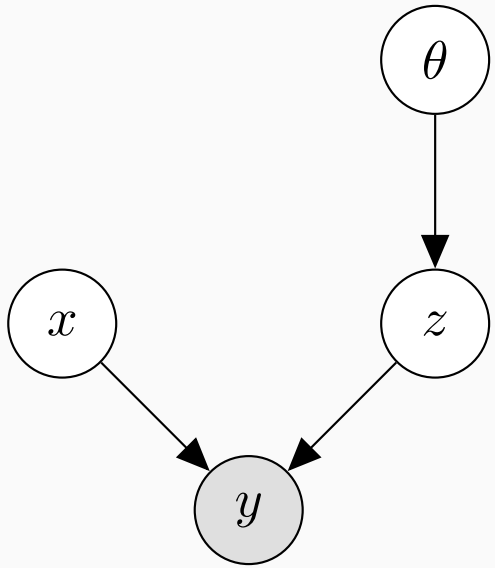




How to choose model?



$p(y \mid x, \theta, z)p(x)p(\theta)p(z)$



$p(y \mid x, \theta, z)p(x)p(z \mid \theta)p(\theta)$

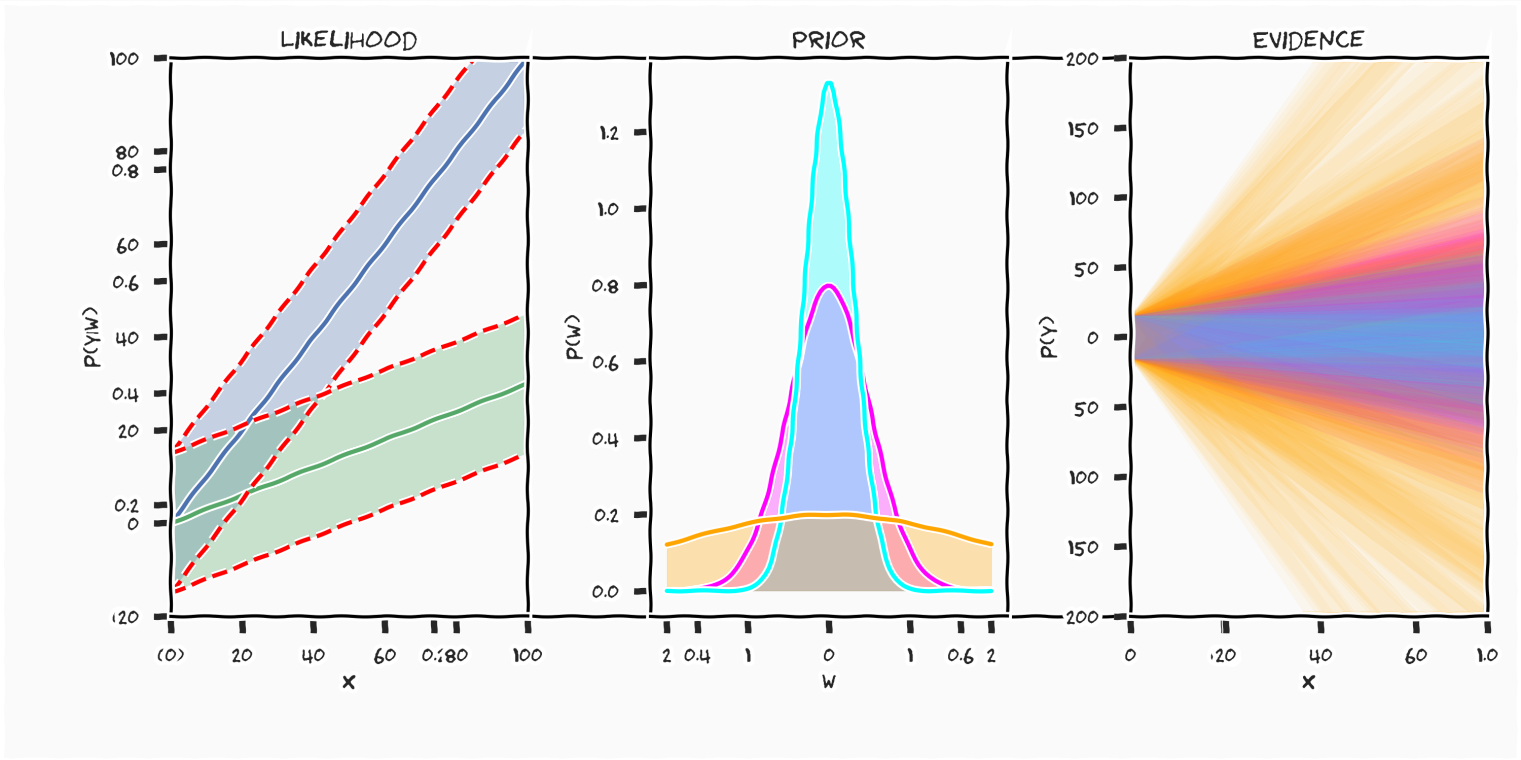
Model Selection

Distributions

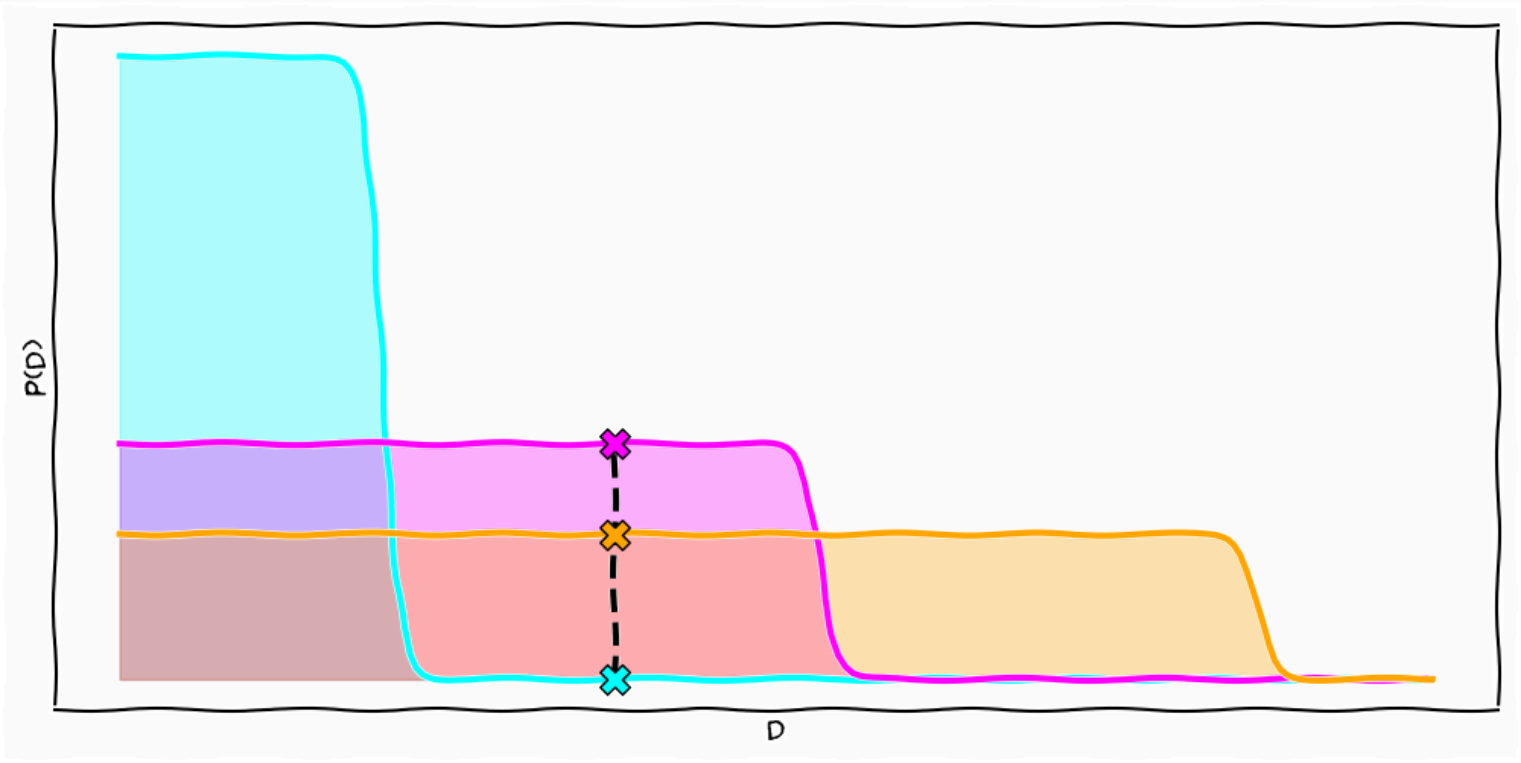
Model Evidence and Occams' Razor

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

What is can be falsified?



The MacKay Plot Mackay, 1991



Is Machine Learning a Science?

- How to build mathematical models of hypothesis

$$\text{hypothesis} \approx p(w)$$

Is Machine Learning a Science?

- How to build mathematical models of hypothesis

$$\text{hypothesis} \approx p(w)$$

- How to mathematically update our knowledge with data

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

Unsupervised Learning

Supervised Learning

$$y_i = f(x_i)$$

- learn relationship $f(\cdot)$ between pairs of data x_i and y_i

Machine Learning

Supervised Learning

$$y_i = f(x_i)$$

- learn relationship $f(\cdot)$ between pairs of data x_i and y_i

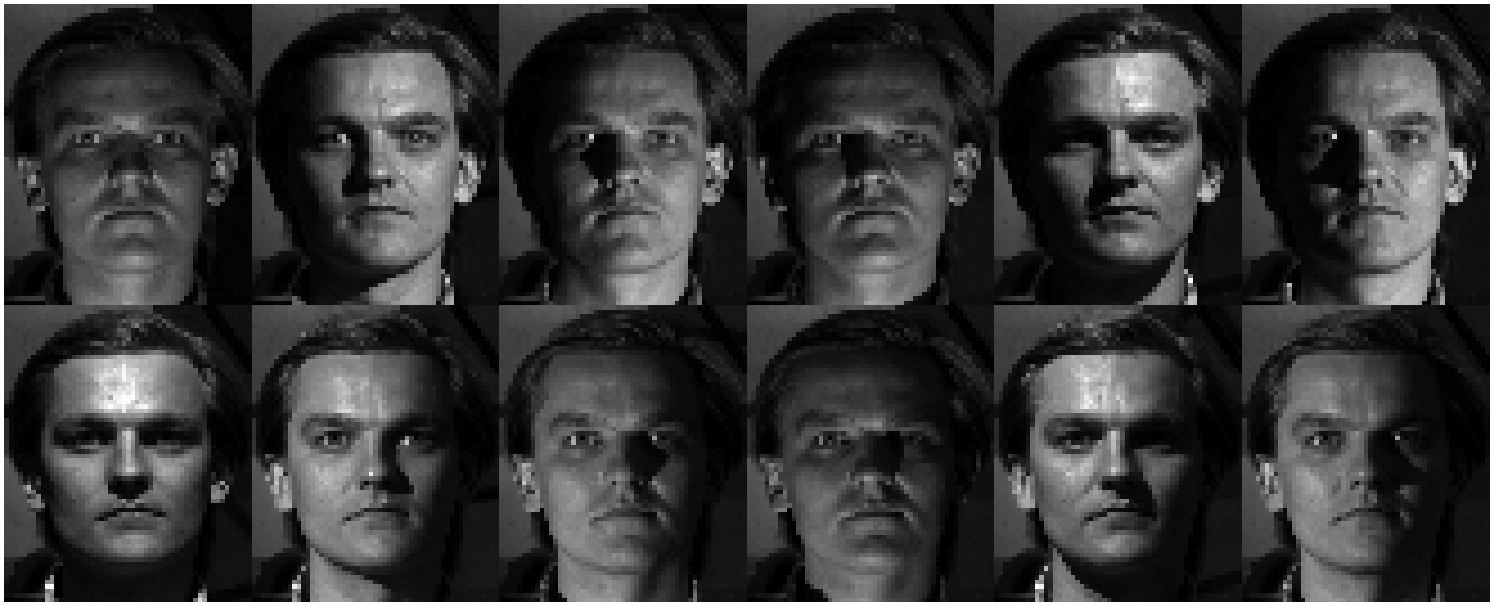
Unsupervised Learning

$$y_i = f(x_i)$$

- learn a representation $\mathbf{X}$ from data $\mathbf{Y}$

Latent Variable Models

Latent Variable Models



output data $\mathbf{y} \in \mathbb{R}^{256\times256} \rightarrow 65536$ dimensions

input location on sphere $\rightarrow 3$ dimensions

manifold images lie on a 3 dimensional surface in 65536 dimensions

Strength of Priors

$$y = f(x)$$

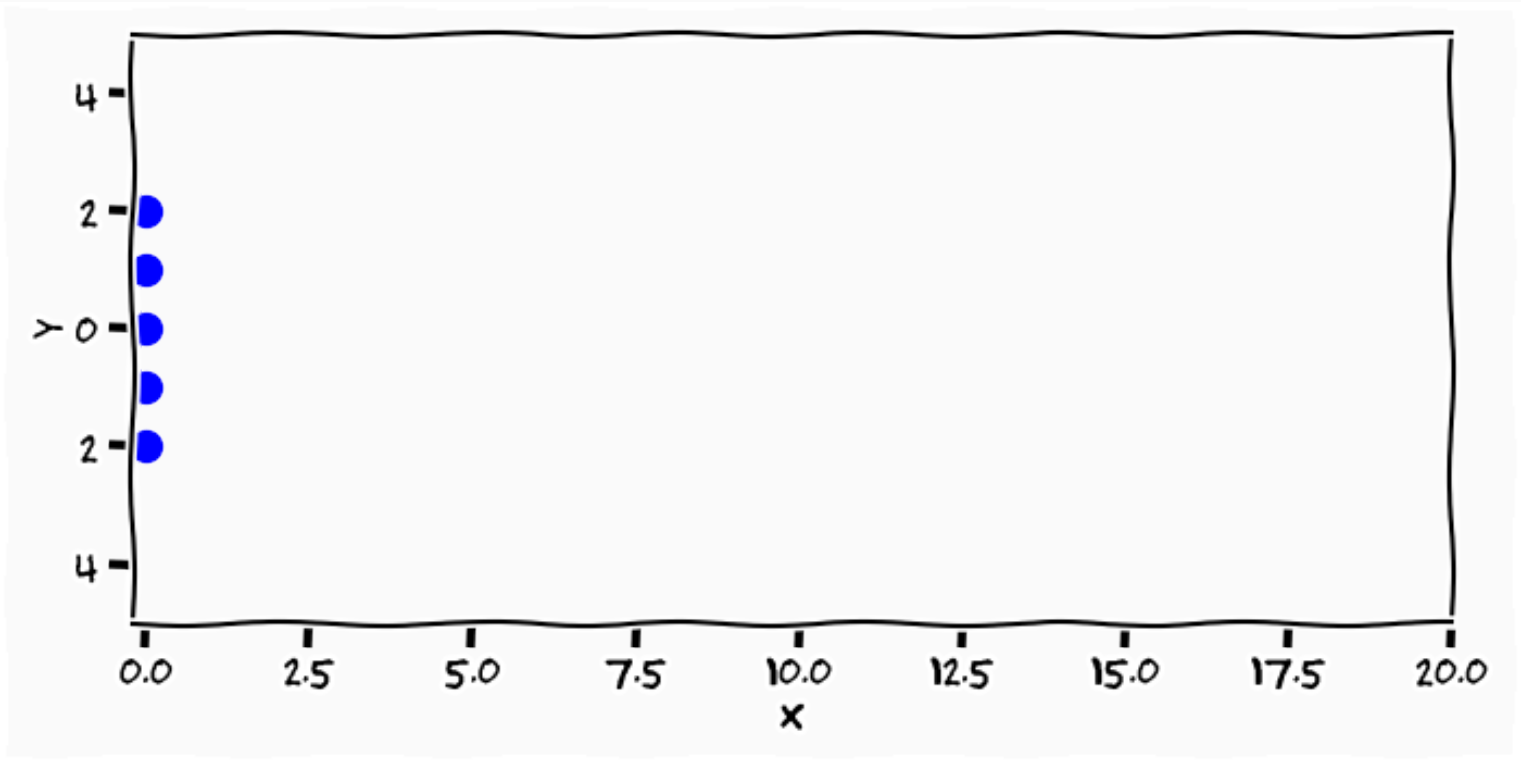
- given input output pairs we have made assumptions about f
- from data we can update our assumption
- can we push this further?

Unsupervised learning

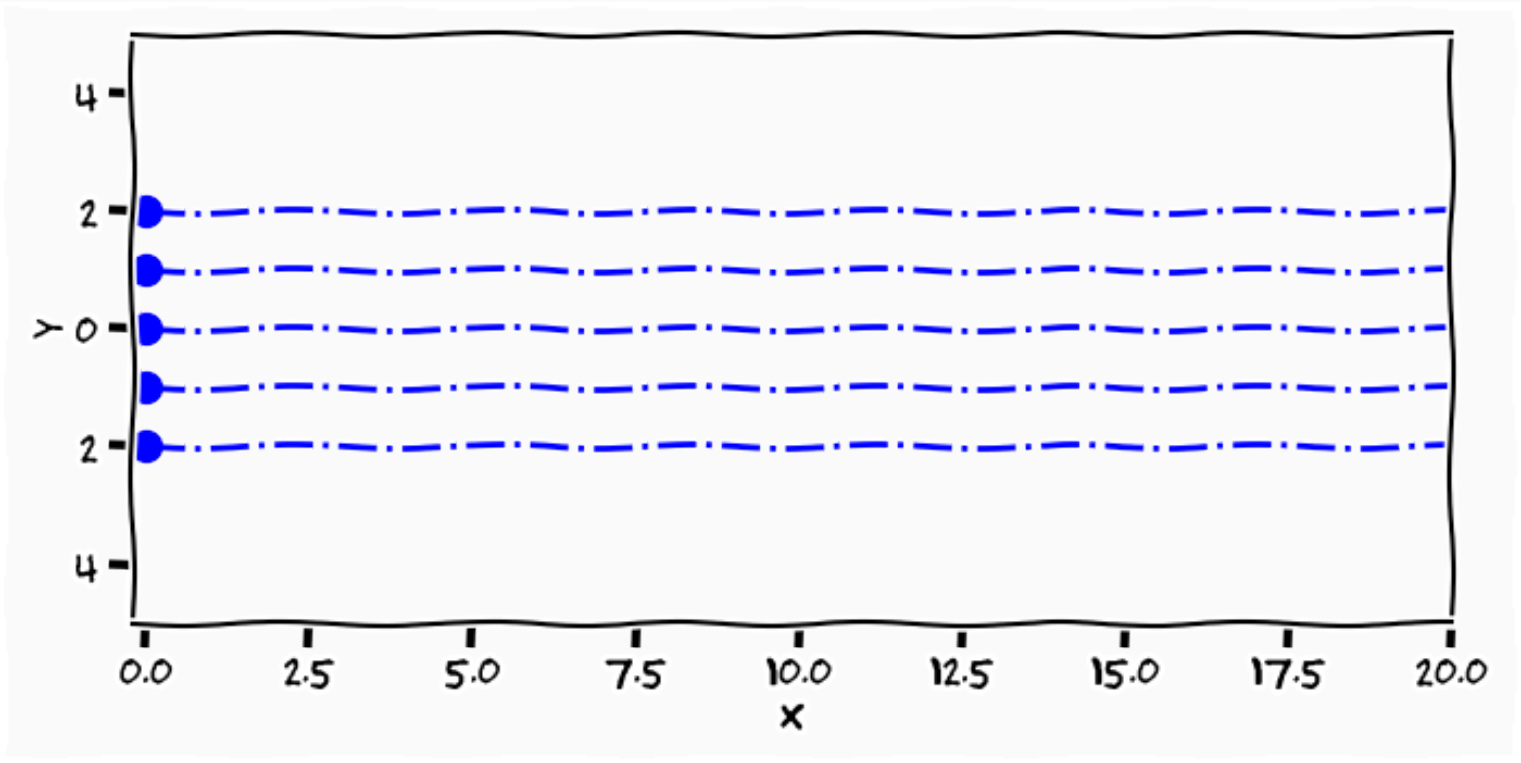
$$y = f(x)$$

- In unsupervised learning we are given only output
- Input is *latent*
- Task: recover both f and x

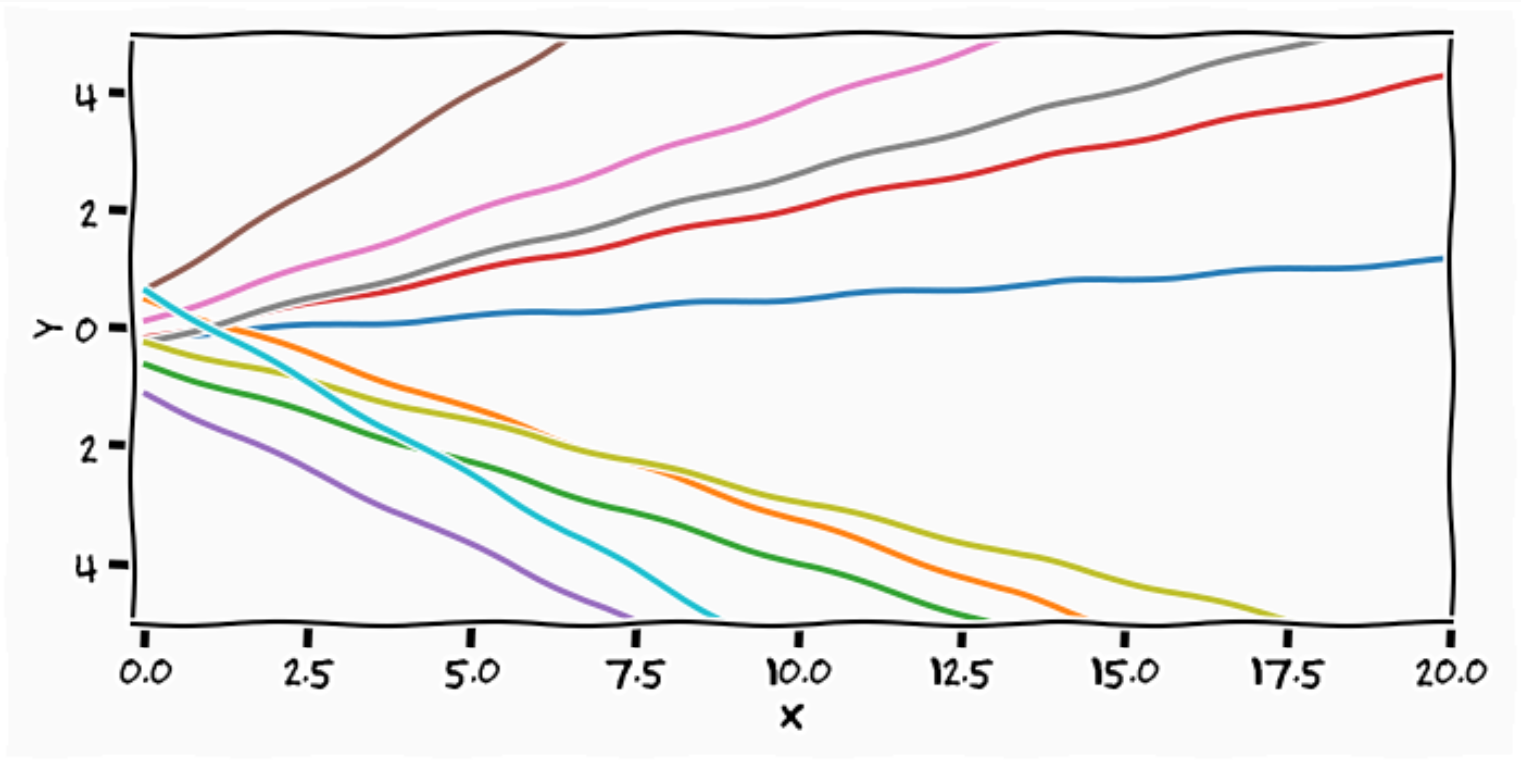
Unsupervised Learning



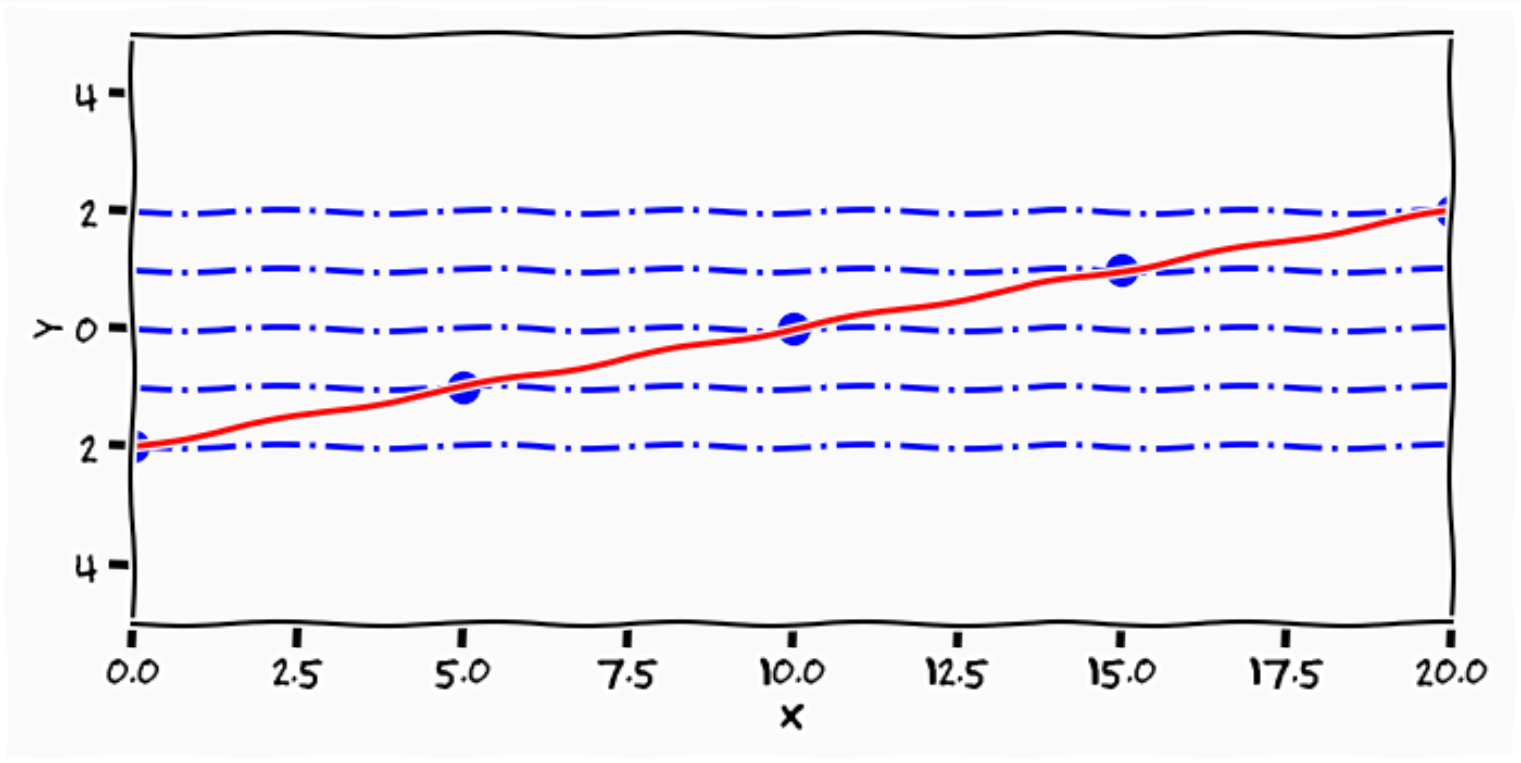
Unsupervised Learning



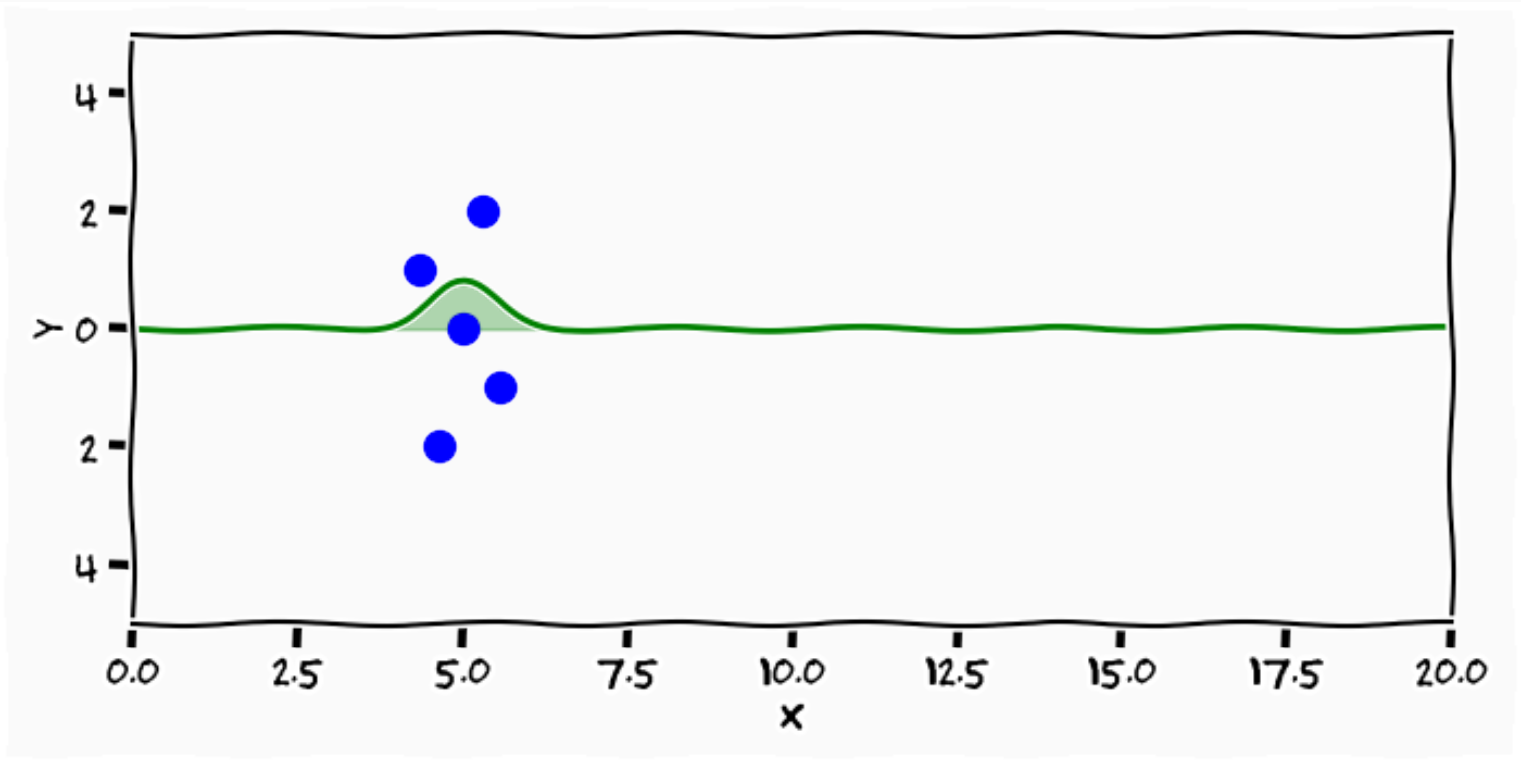
Unsupervised Learning



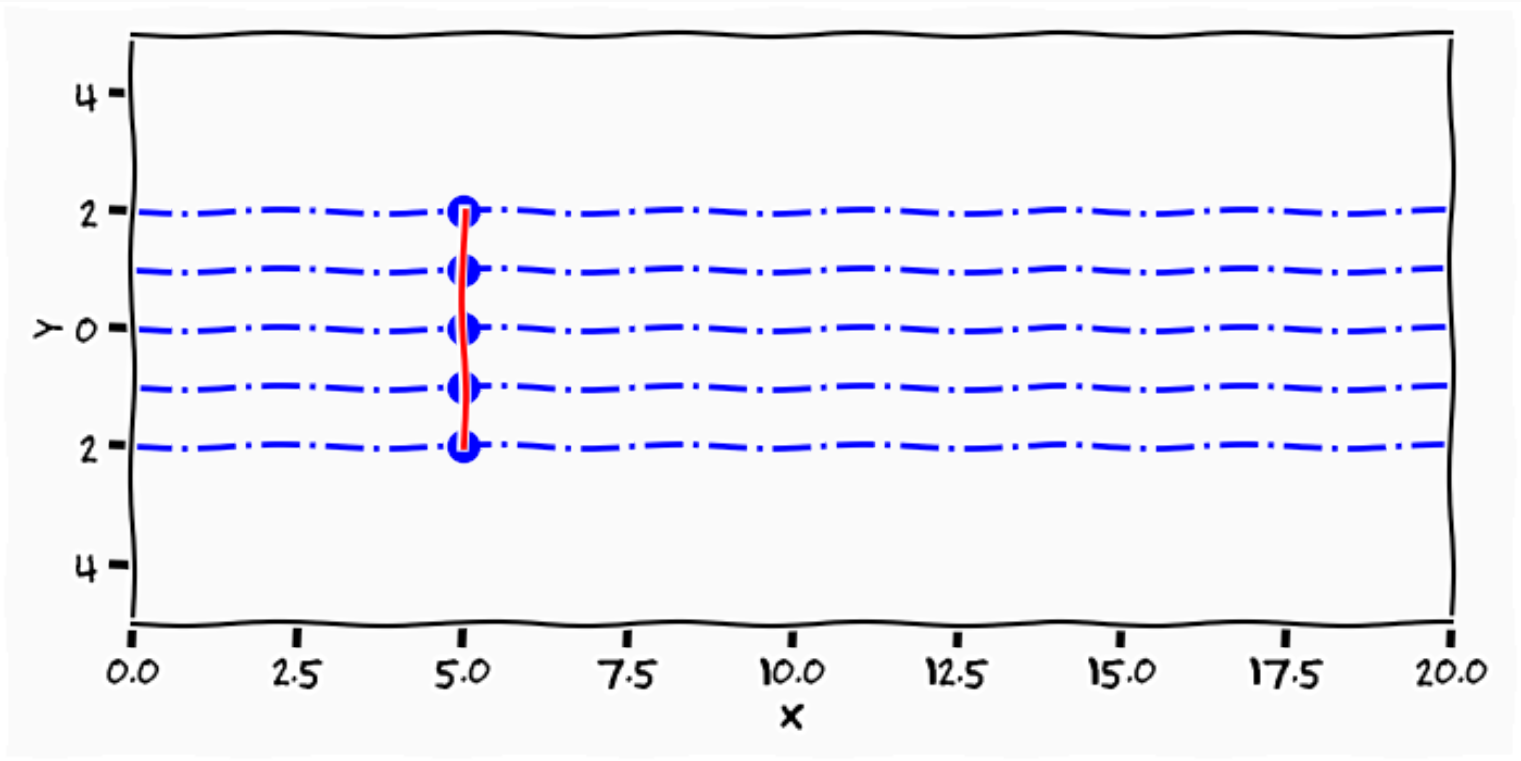
Unsupervised Learning



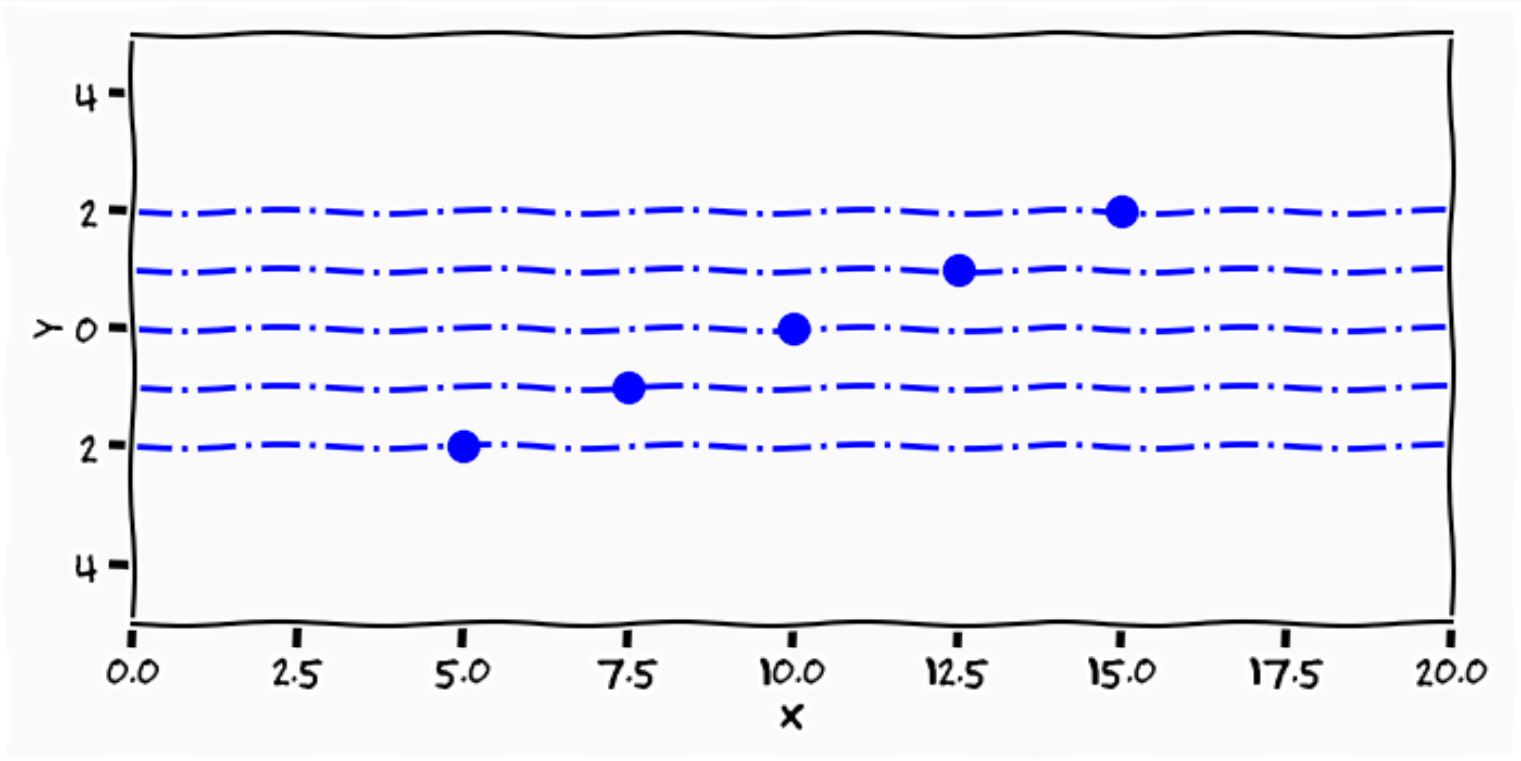
Unsupervised Learning



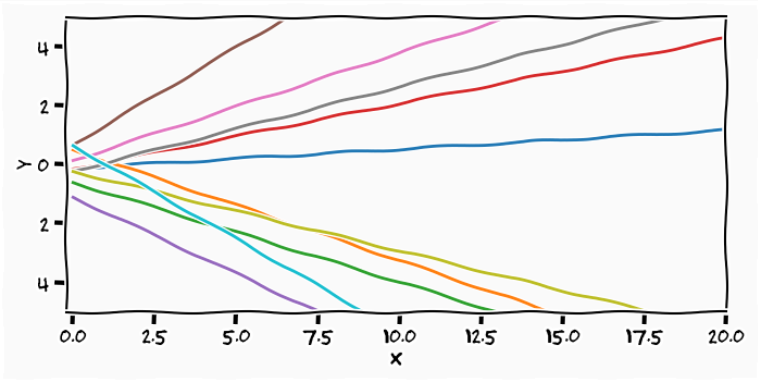
Unsupervised Learning

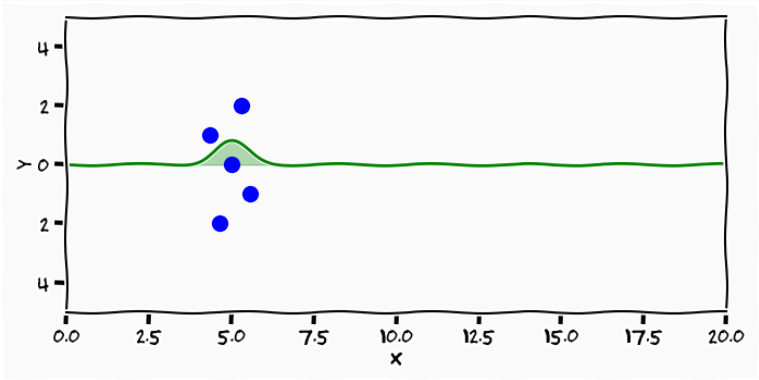


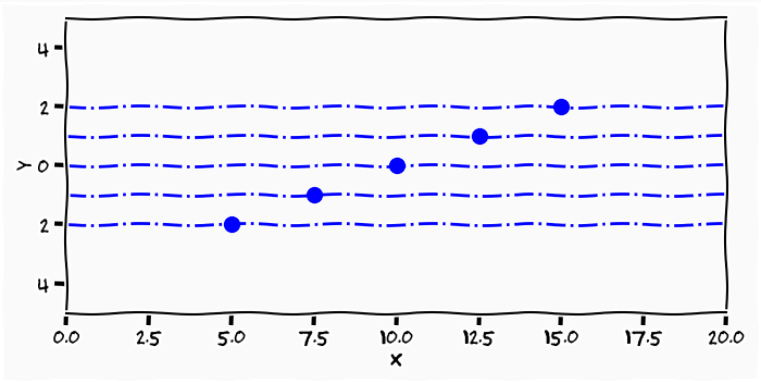
Unsupervised Learning



Unsupervised Learning







Linear Latent Variable Models

- Linear Regression

$$p(\mathbf{W}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

Linear Latent Variable Models

- Linear Regression

$$p(\mathbf{W}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

- Linear Unsupervised Learning

$$p(\mathbf{W}, \mathbf{X}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{W}, \mathbf{X})p(\mathbf{W})p(\mathbf{X})$$

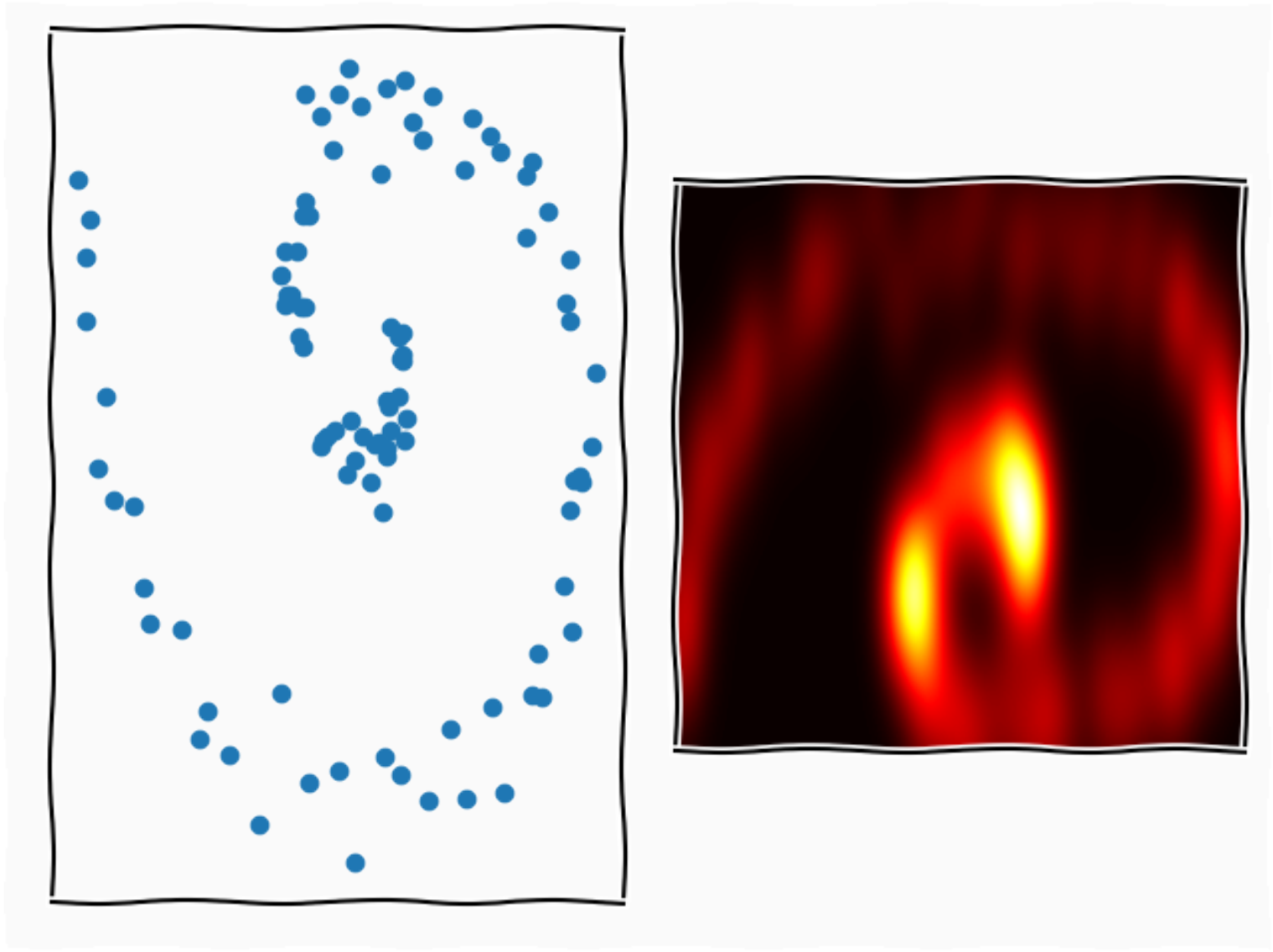
Linear Latent Variable Model

$$p(\mathbf{x}, \mathbf{W}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}, \mathbf{W})p(\mathbf{w})p(\mathbf{x})}{p(\mathbf{y})}$$

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{W}, \mathbf{x})p(\mathbf{W})p(\mathbf{x})\mathrm{d}\mathbf{W}\mathrm{d}\mathbf{x}$$

- Intractable to reach posterior distribution of both variables

Example



Principal Component Analysis [3]

$$\mathbf{V}\Lambda\mathbf{V}^{\mathrm{T}} = \mathbf{y}^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{x} = \sum_{i}^{d} \mathbf{y}\mathbf{V}_i$$

- The above is the solution if $\sigma^2 \to 0$

[3] Spearman, 1904

Principal Component Analysis

- You might have seen this explained in a different way
  - *Retain variance*
  - *Error minimisation*
- These provides the same solution as the maximum likelihood but solved by an eigenvalue problem
- Do not provide intuition as it doesn't state assumptions

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

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

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

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

2. Formulate your likelihood

$$p(\mathcal{D}|\theta = \theta_i)$$

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

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

2. Formulate your likelihood

$$p(\mathcal{D}|\theta = \theta_i)$$

3. Formluate your belief in the "setting" of the model

$$p(\theta)$$

Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

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

2. Formulate your likelihood

$$p(\mathcal{D}|\theta = \theta_i)$$

3. Formluate your belief in the "setting" of the model

$$p(\theta)$$

4. Aquire data

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

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

2. Formulate your likelihood

$$p(\mathcal{D}|\theta = \theta_i)$$

3. Formluate your belief in the "setting" of the model

$$p(\theta)$$

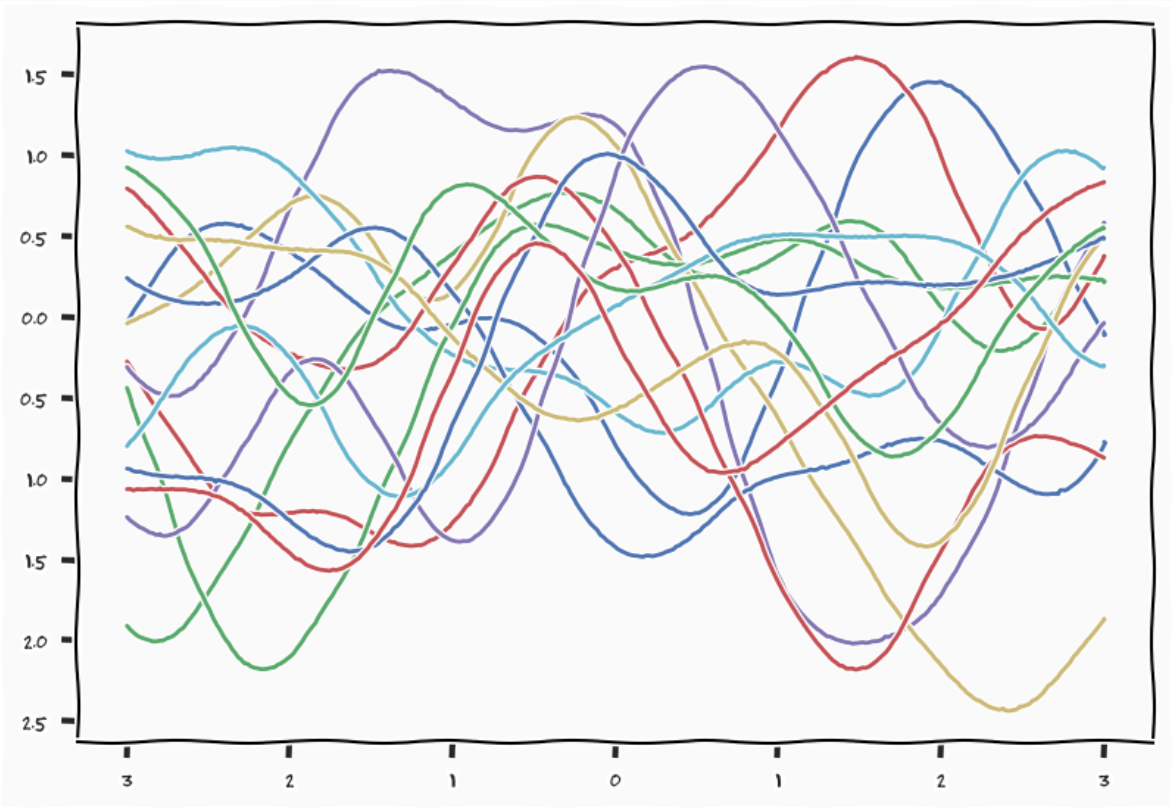
4. Aquire data
5. Derive your updated belief, derive knowledge from data

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)}$$

Unsupervised Learning

- Unsupervised learning is a misnomer, there is no such thing, you have to have beliefs in order to learn.
- Think about unsupervised learning as "more supervised" learning, you have to have stronger beliefs

More Interesting Priors



Demo

Font Demo

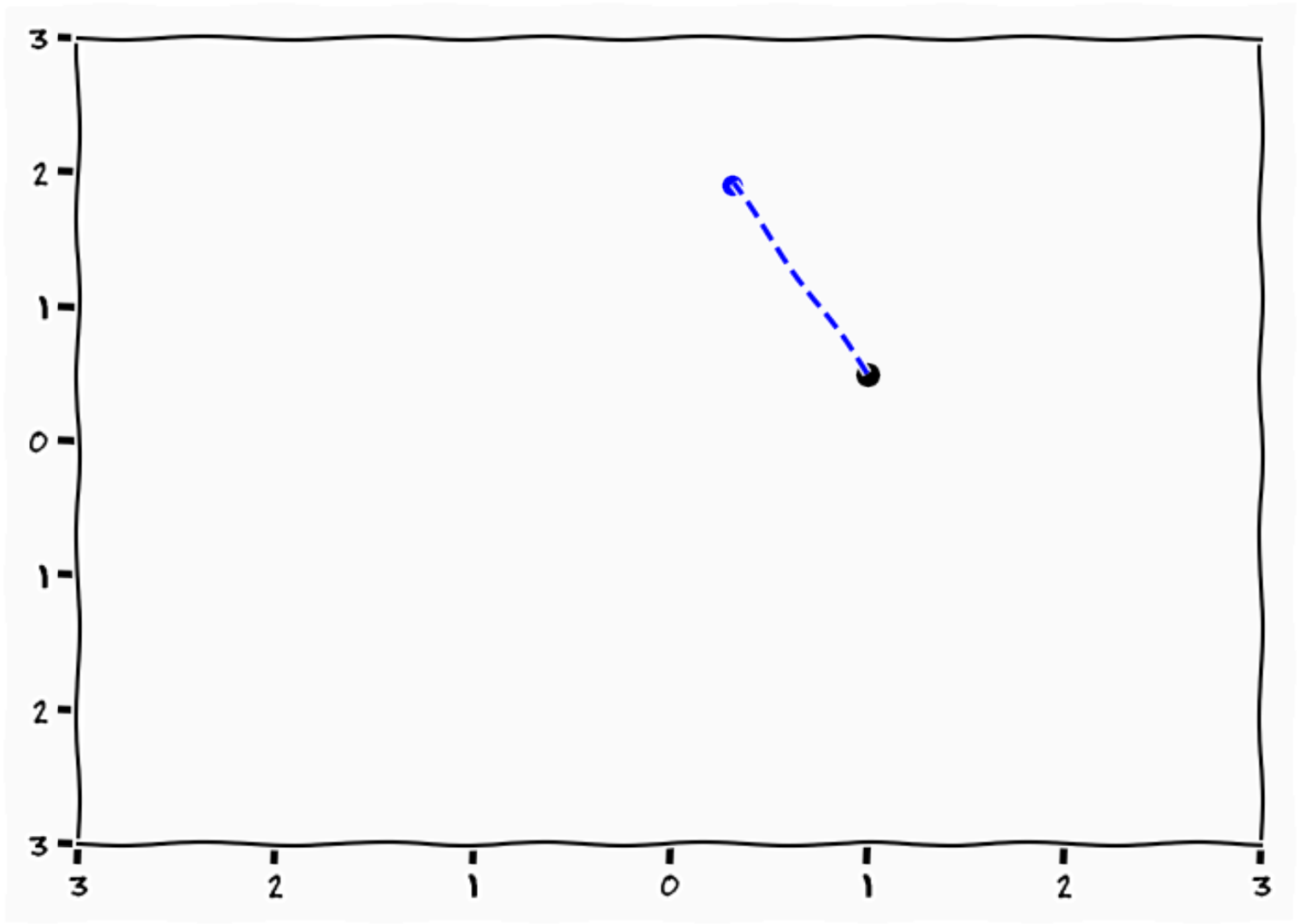
Non-parametric Methods

Non-parametrics

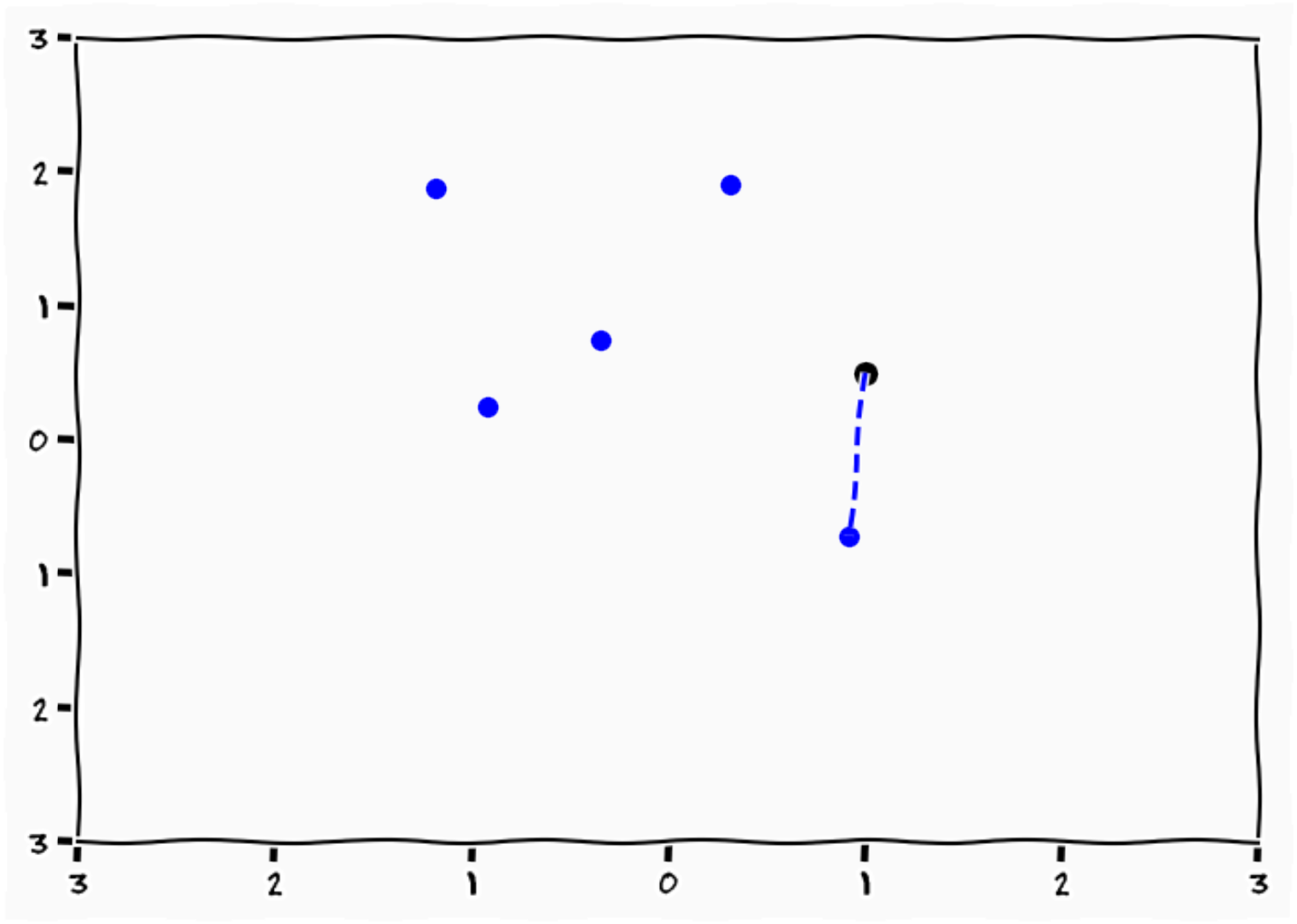
Non-parametrics

Non-parametrics

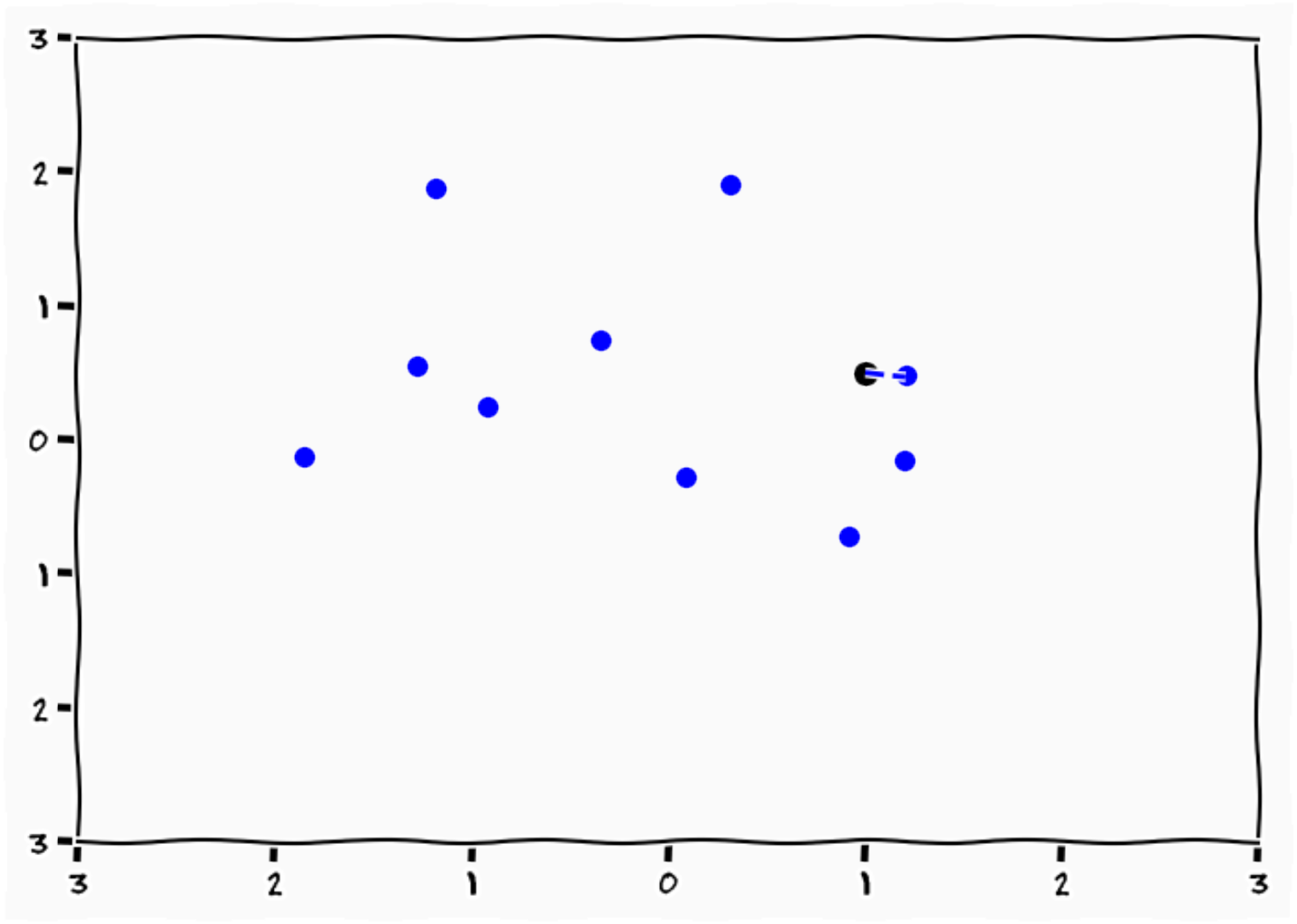
Nearest Neighbour



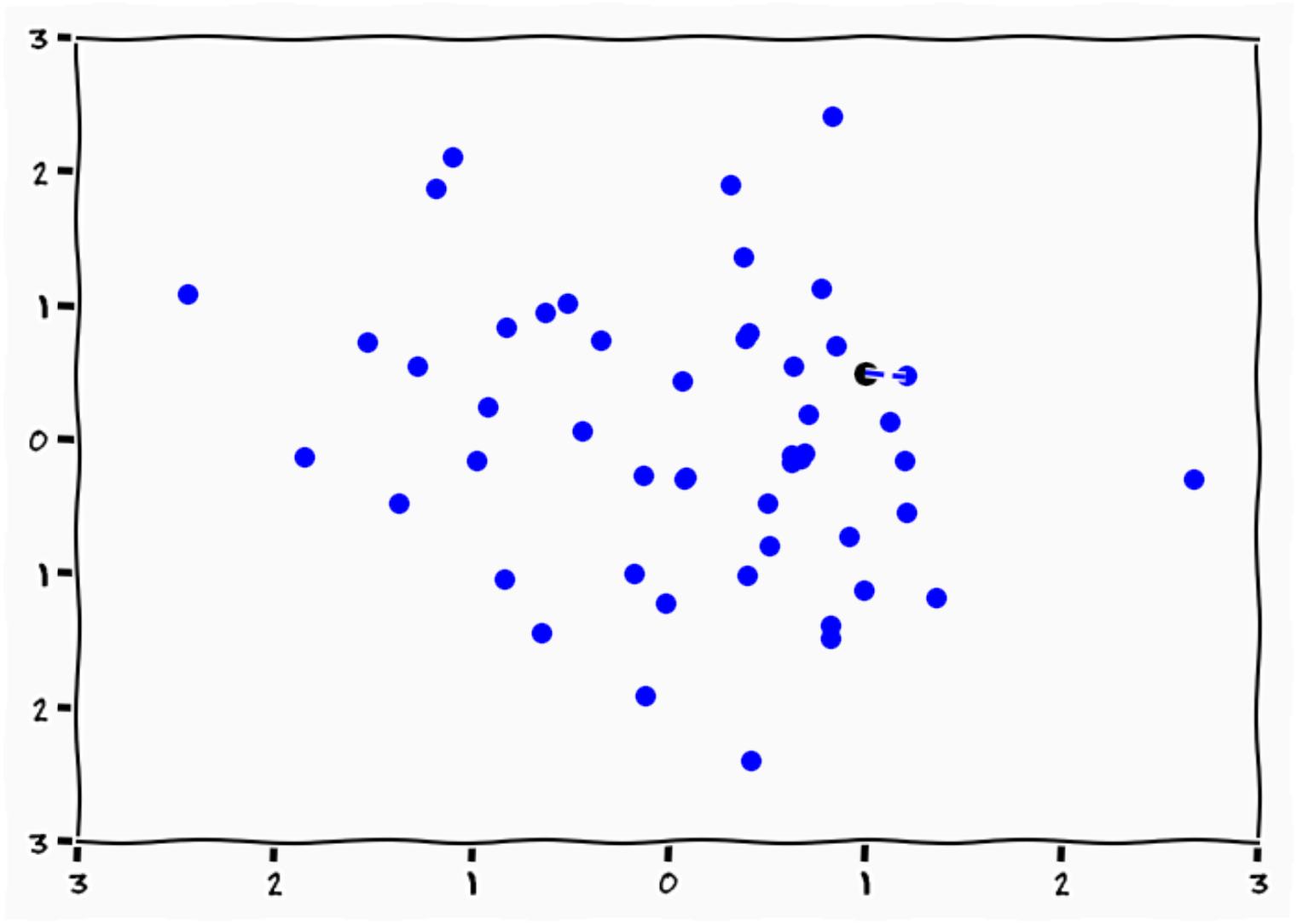
Nearest Neighbour



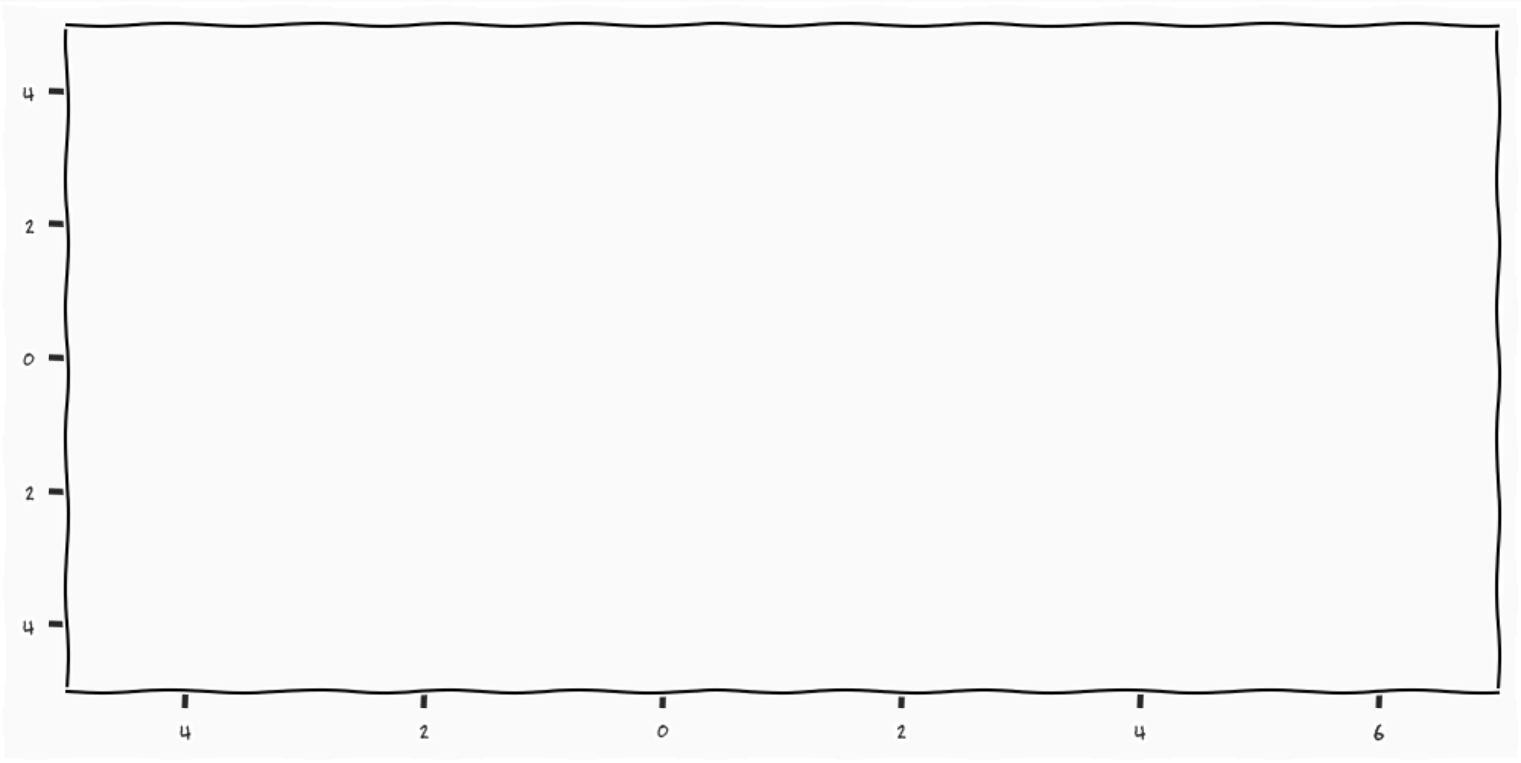
Nearest Neighbour



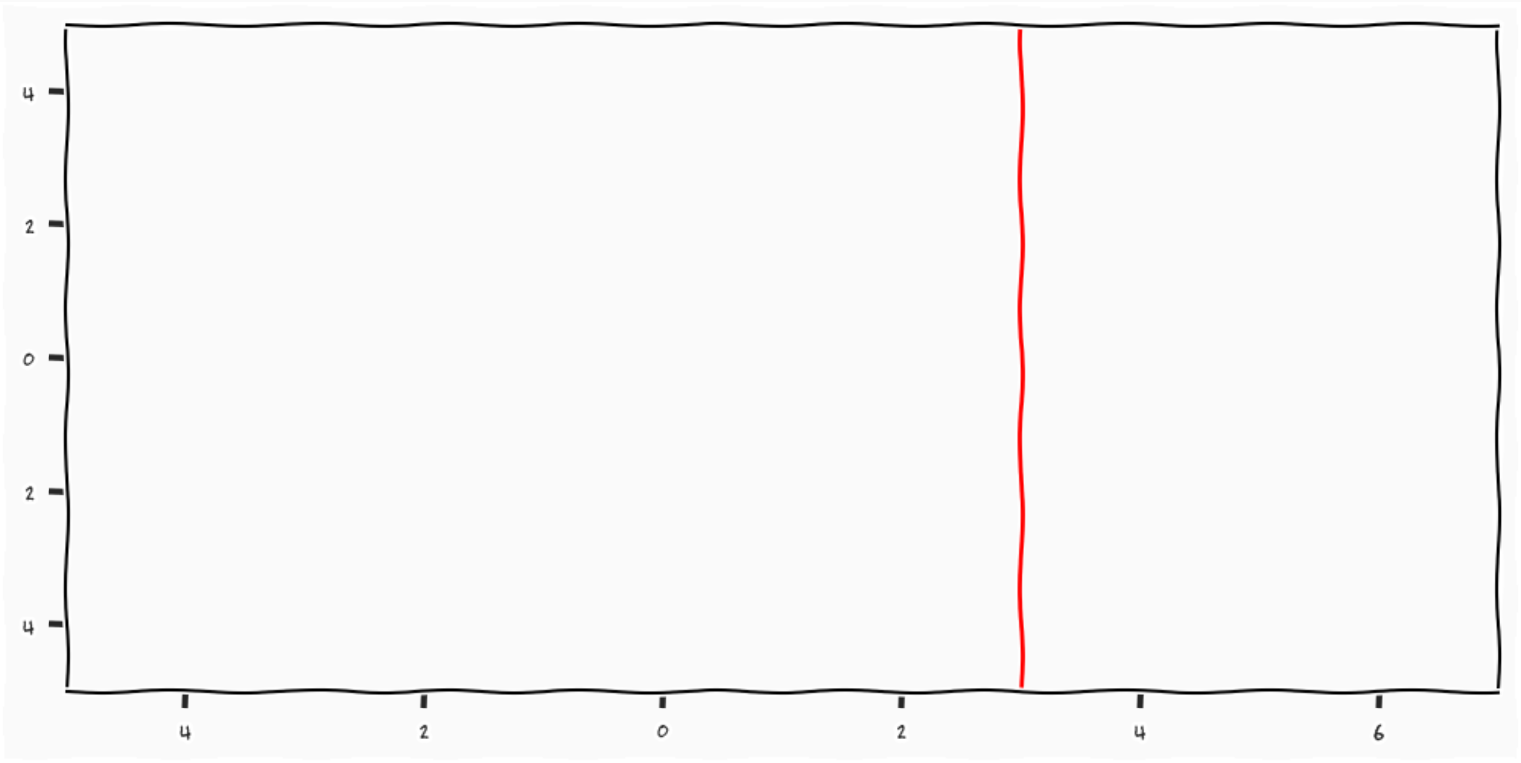
Nearest Neighbour



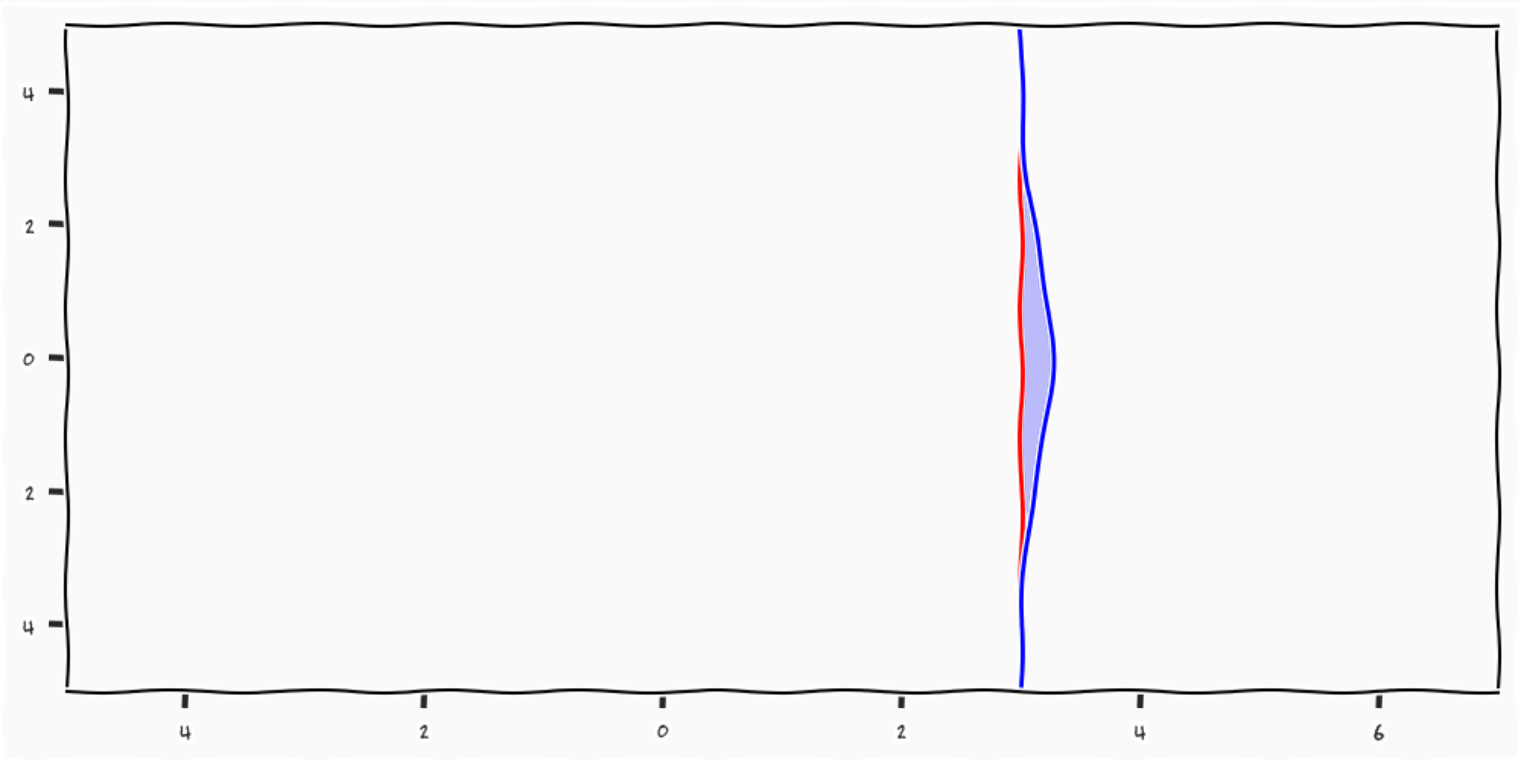
Gaussian Processes



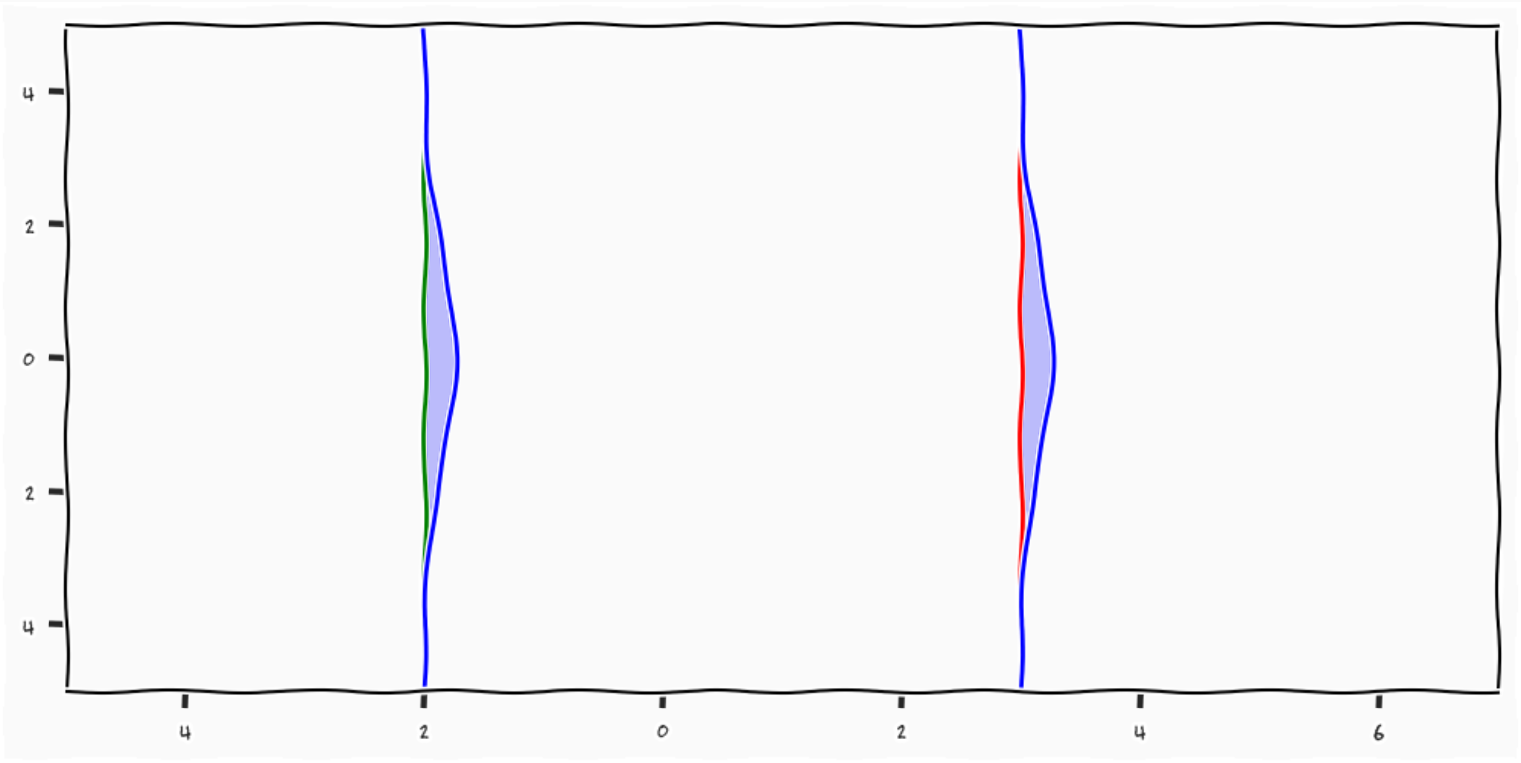
Gaussian Processes



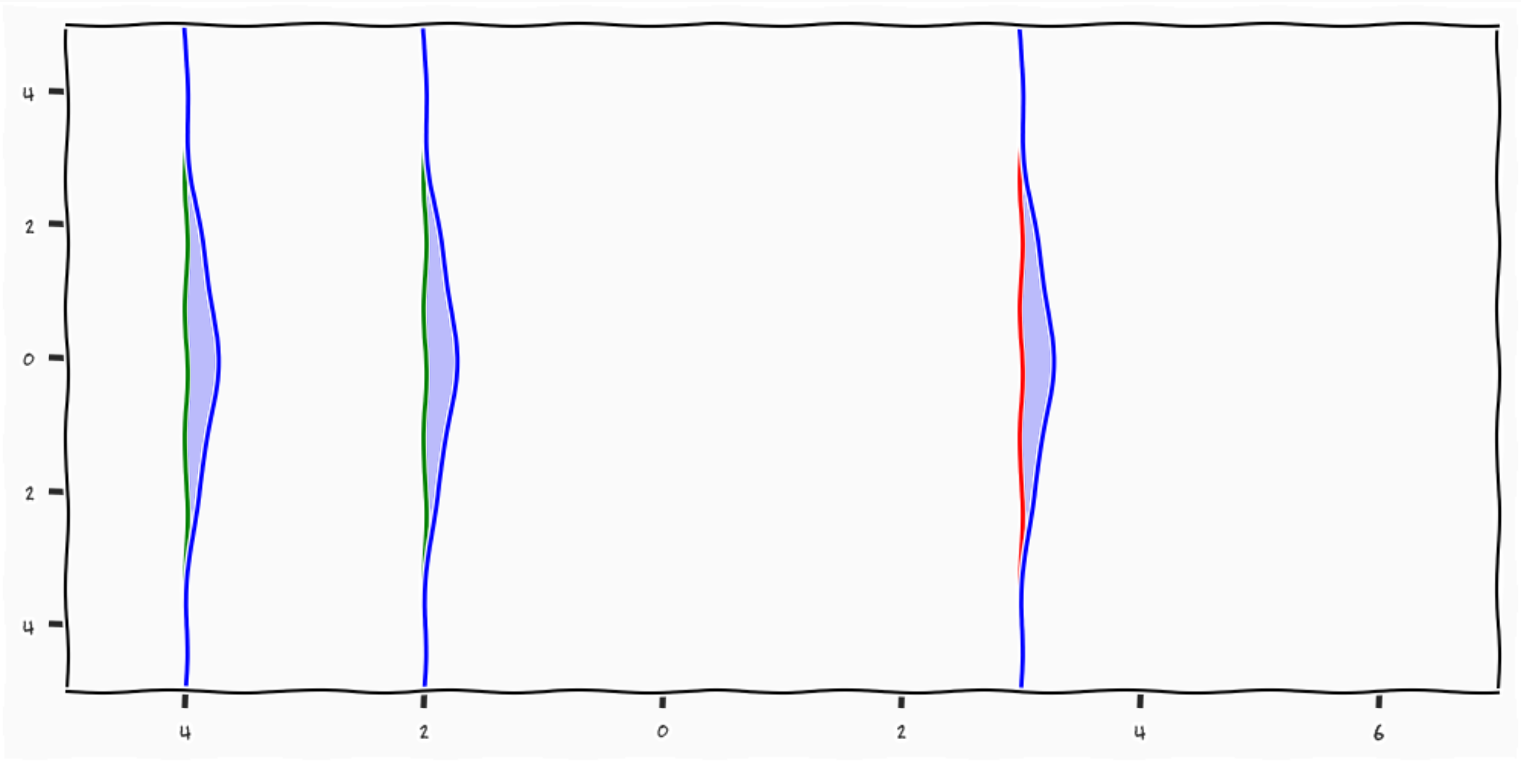
Gaussian Processes



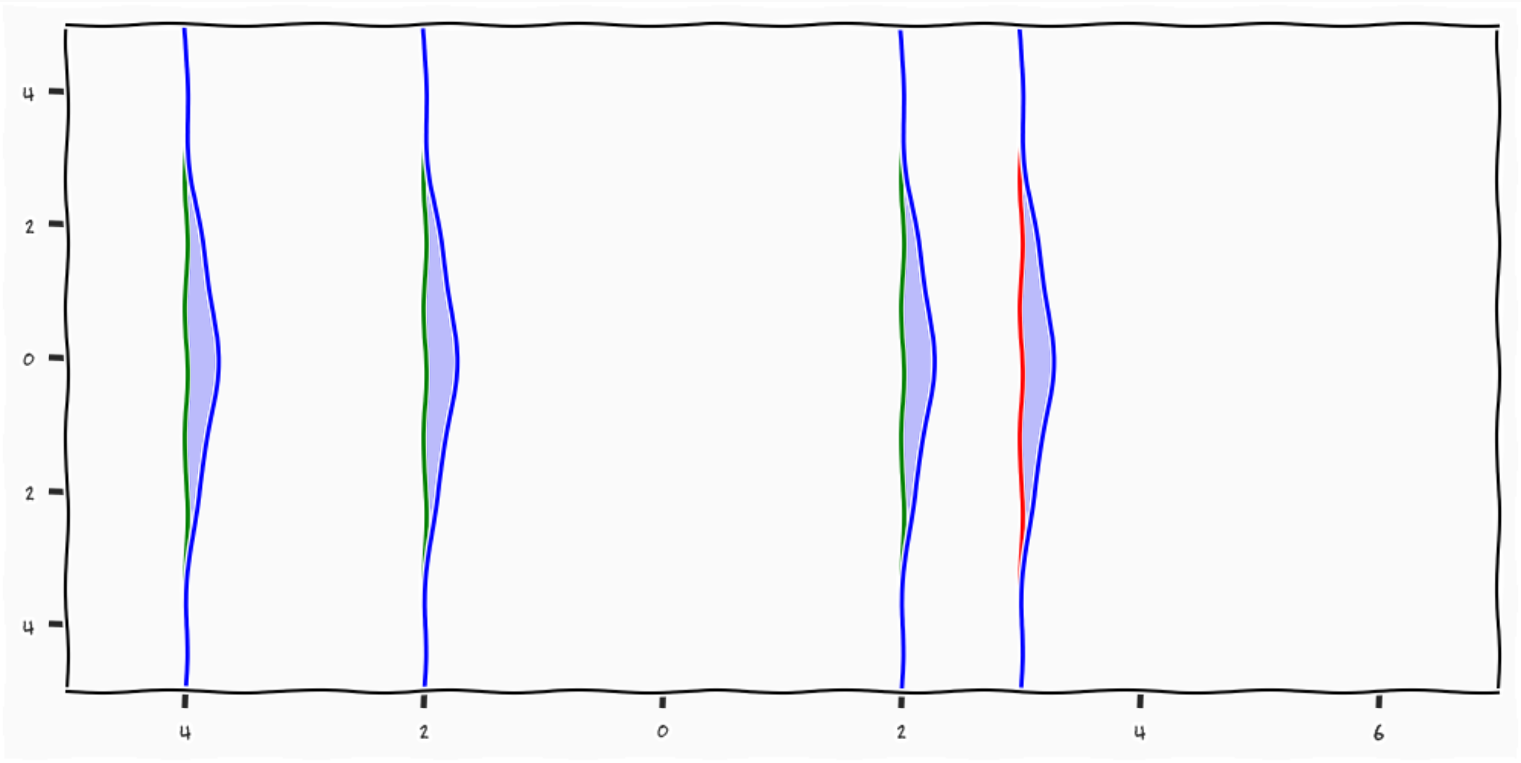
Gaussian Processes



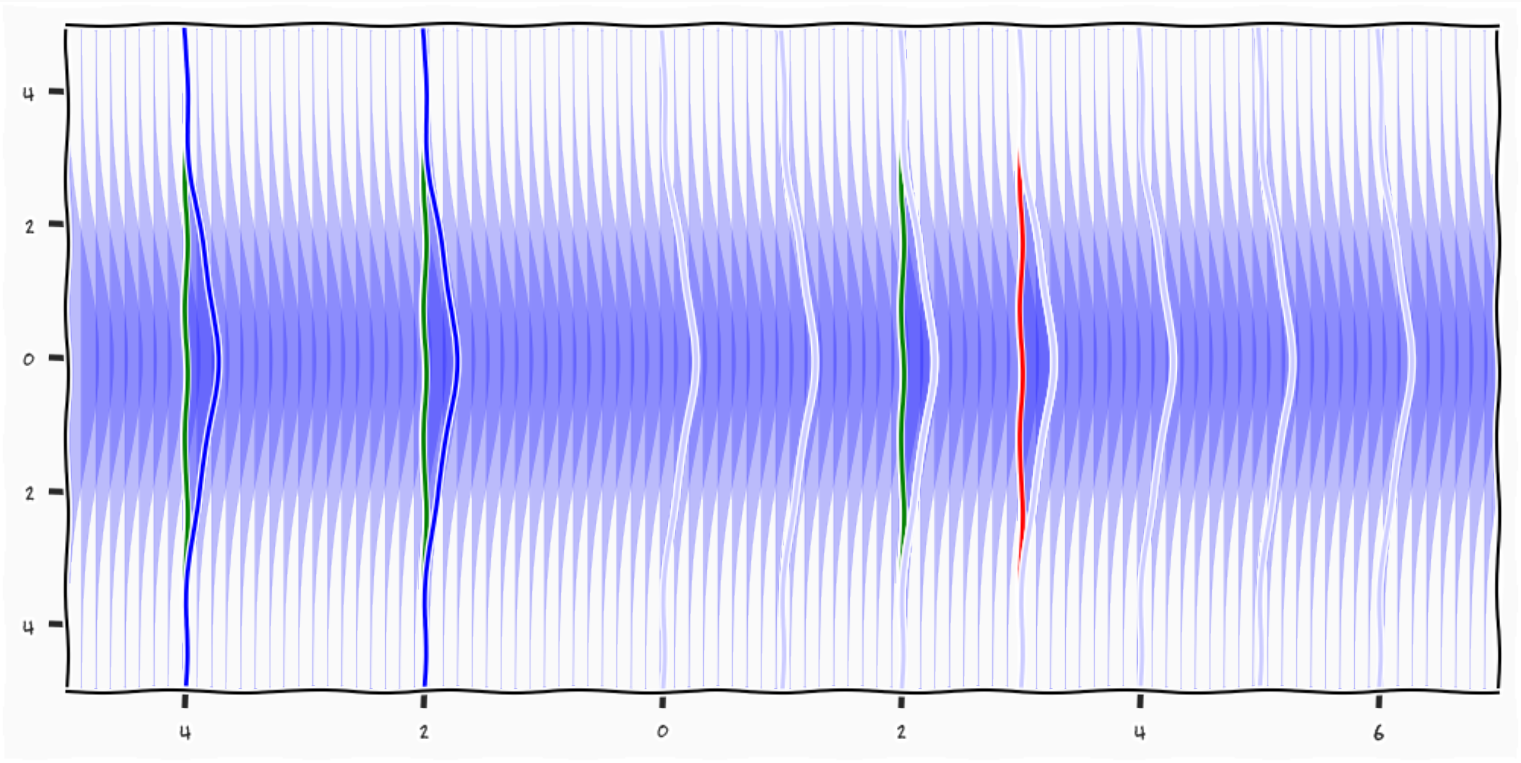
Gaussian Processes



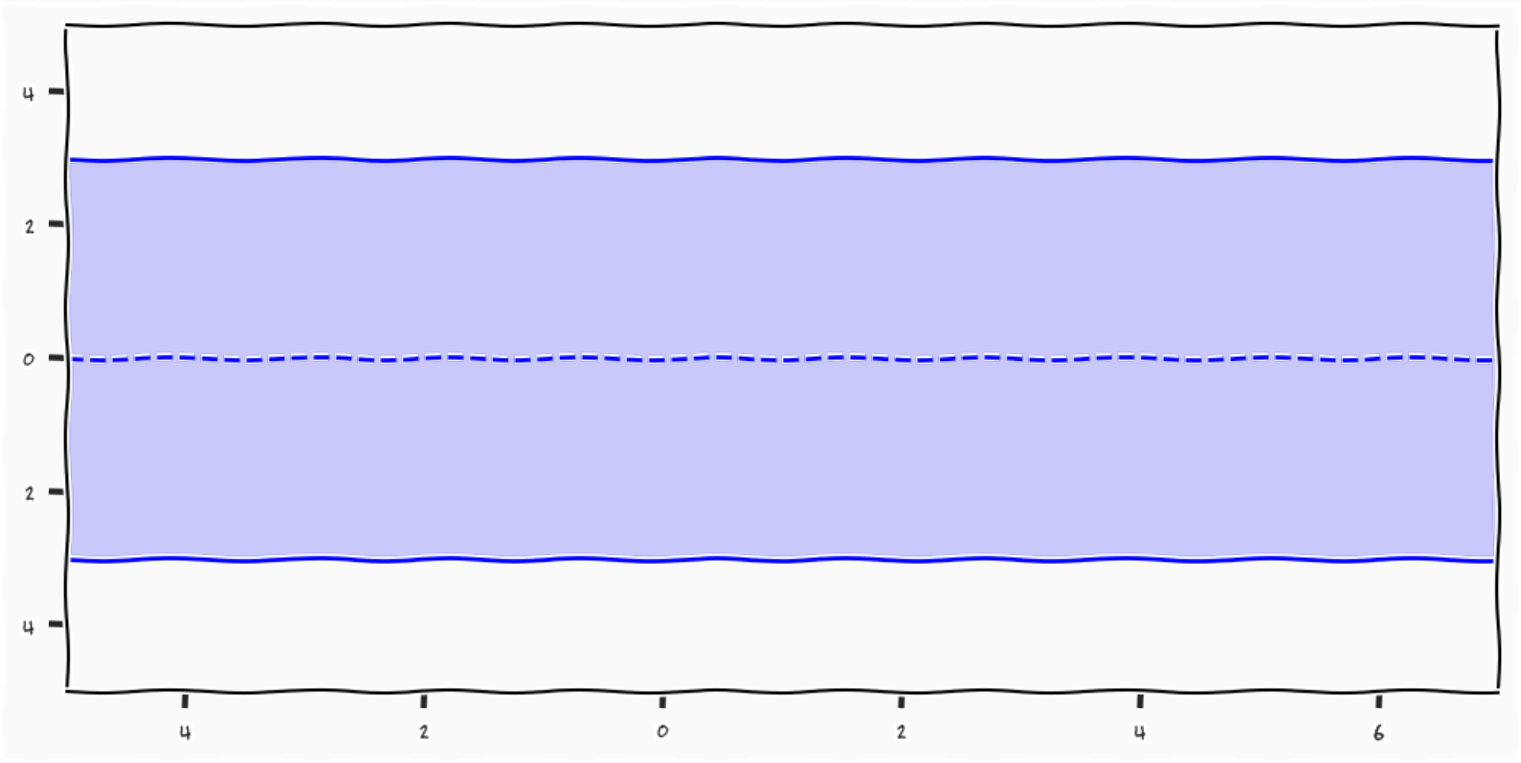
Gaussian Processes



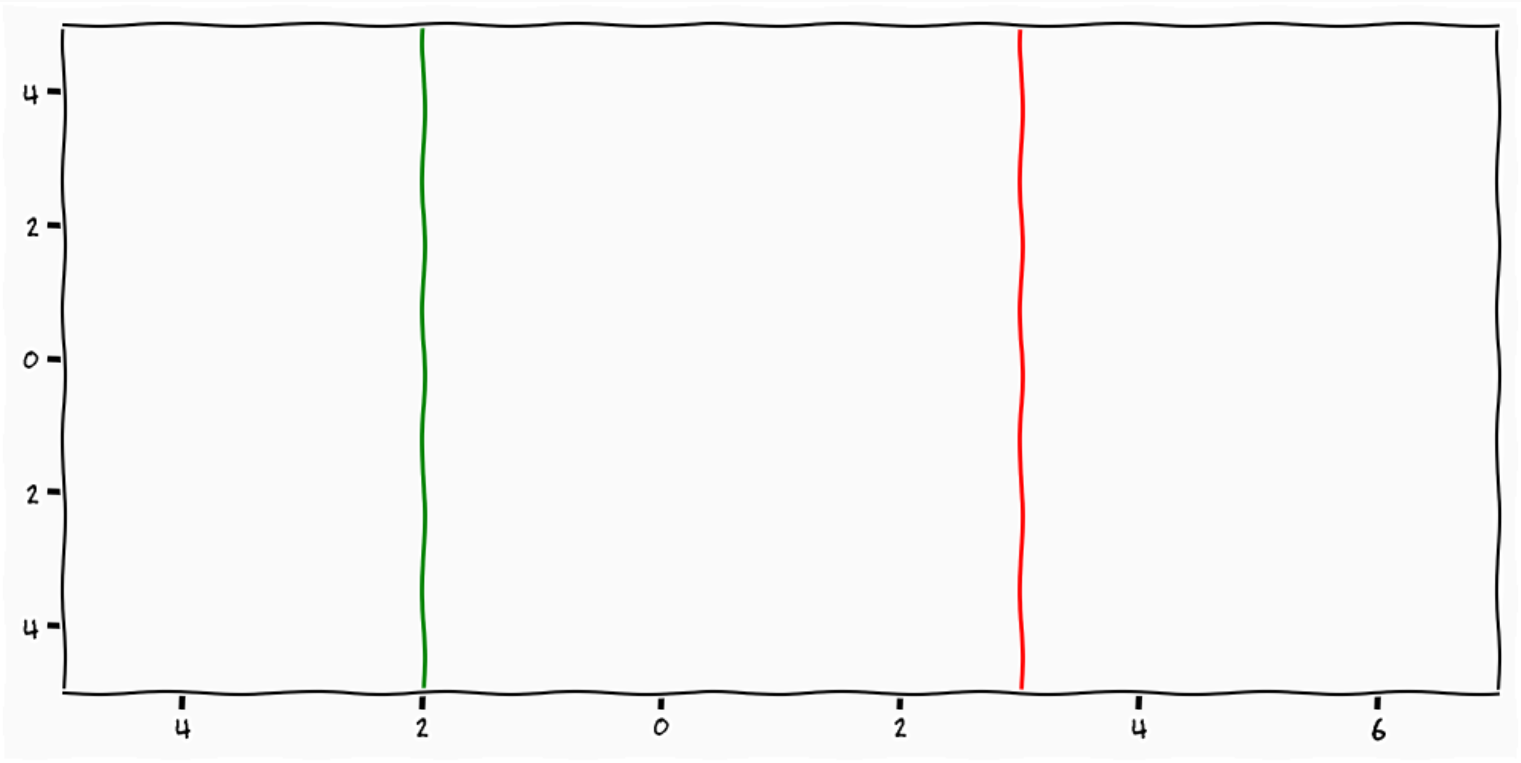
Gaussian Processes



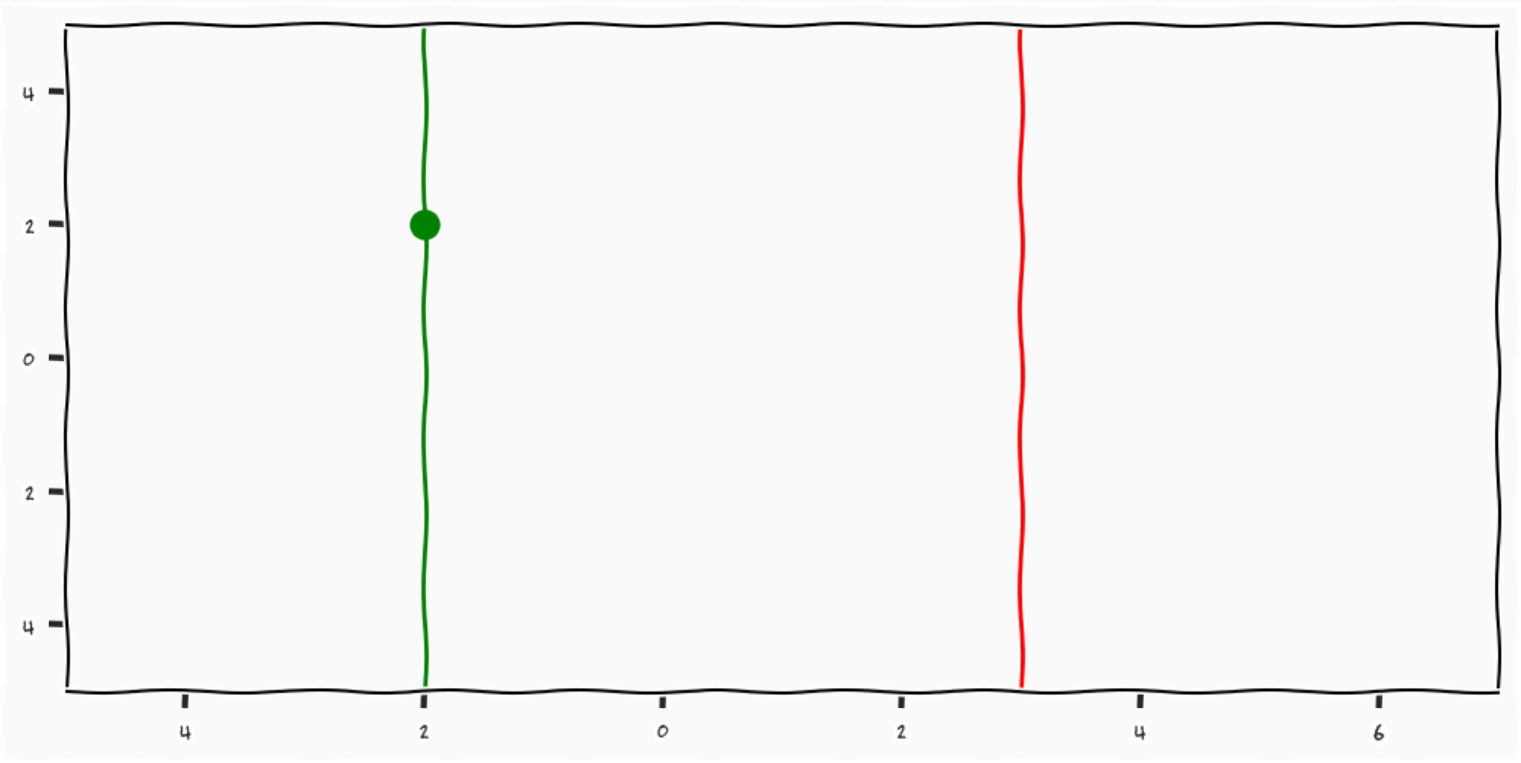
Gaussian Processes



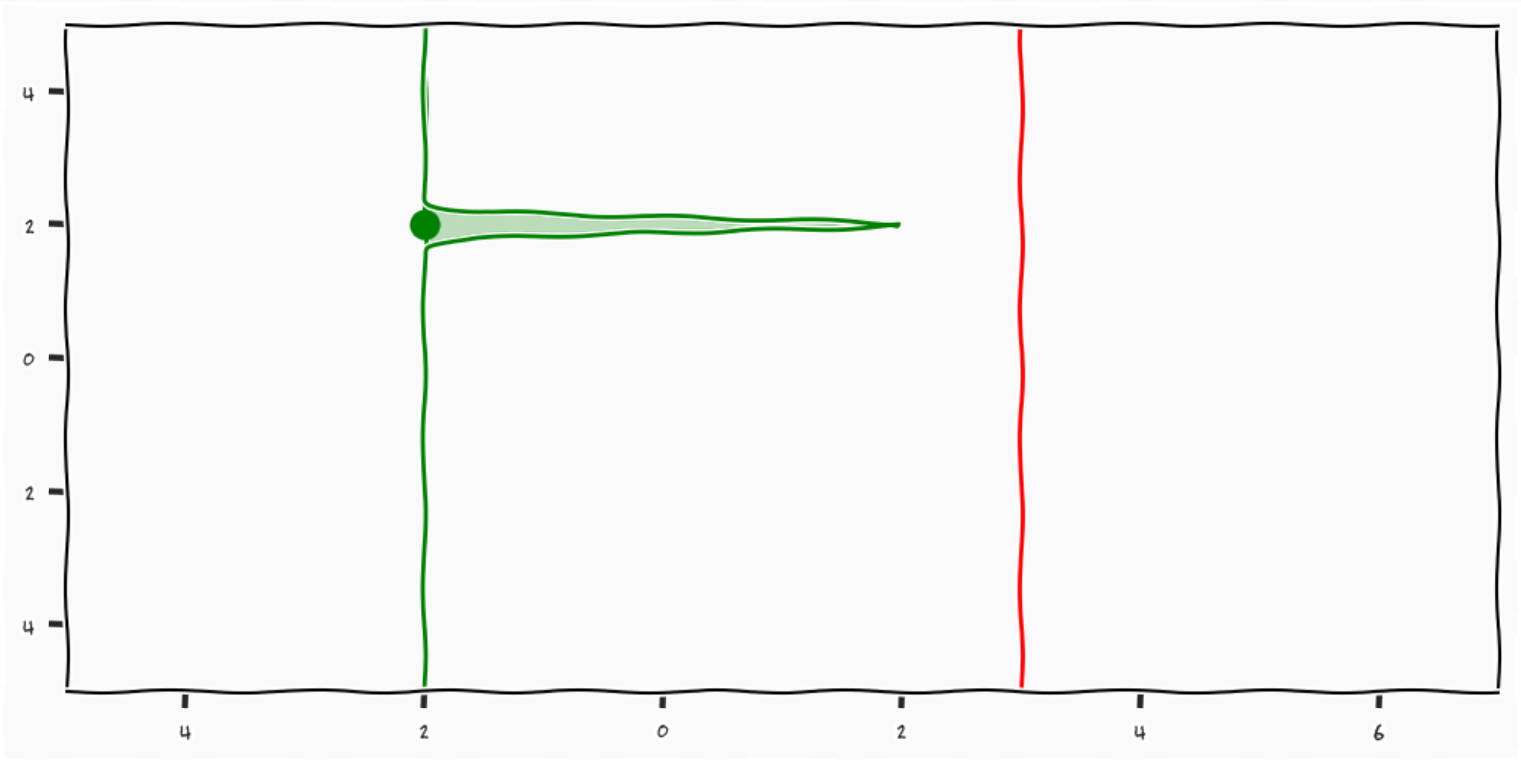
Gaussian Processes



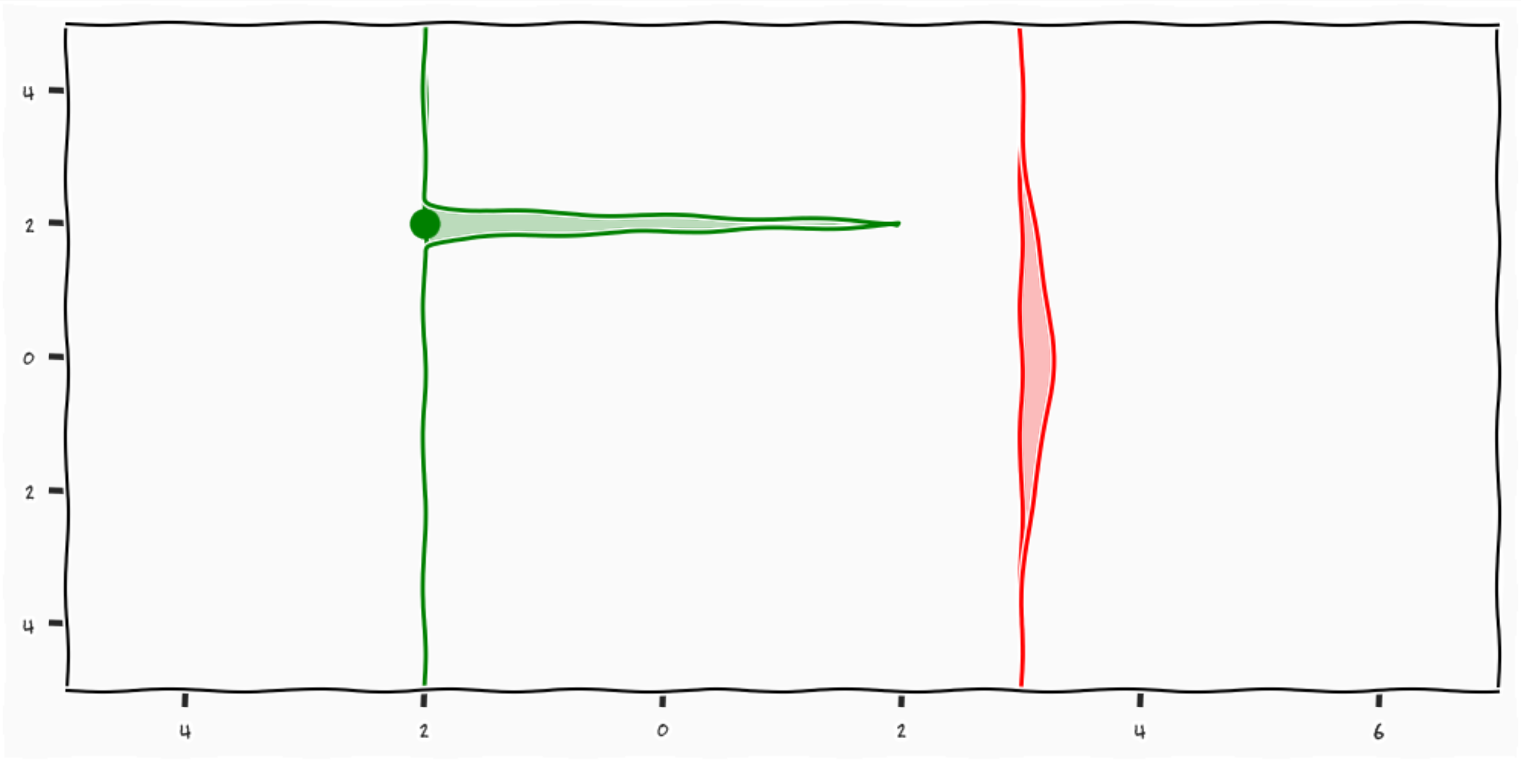
Gaussian Processes



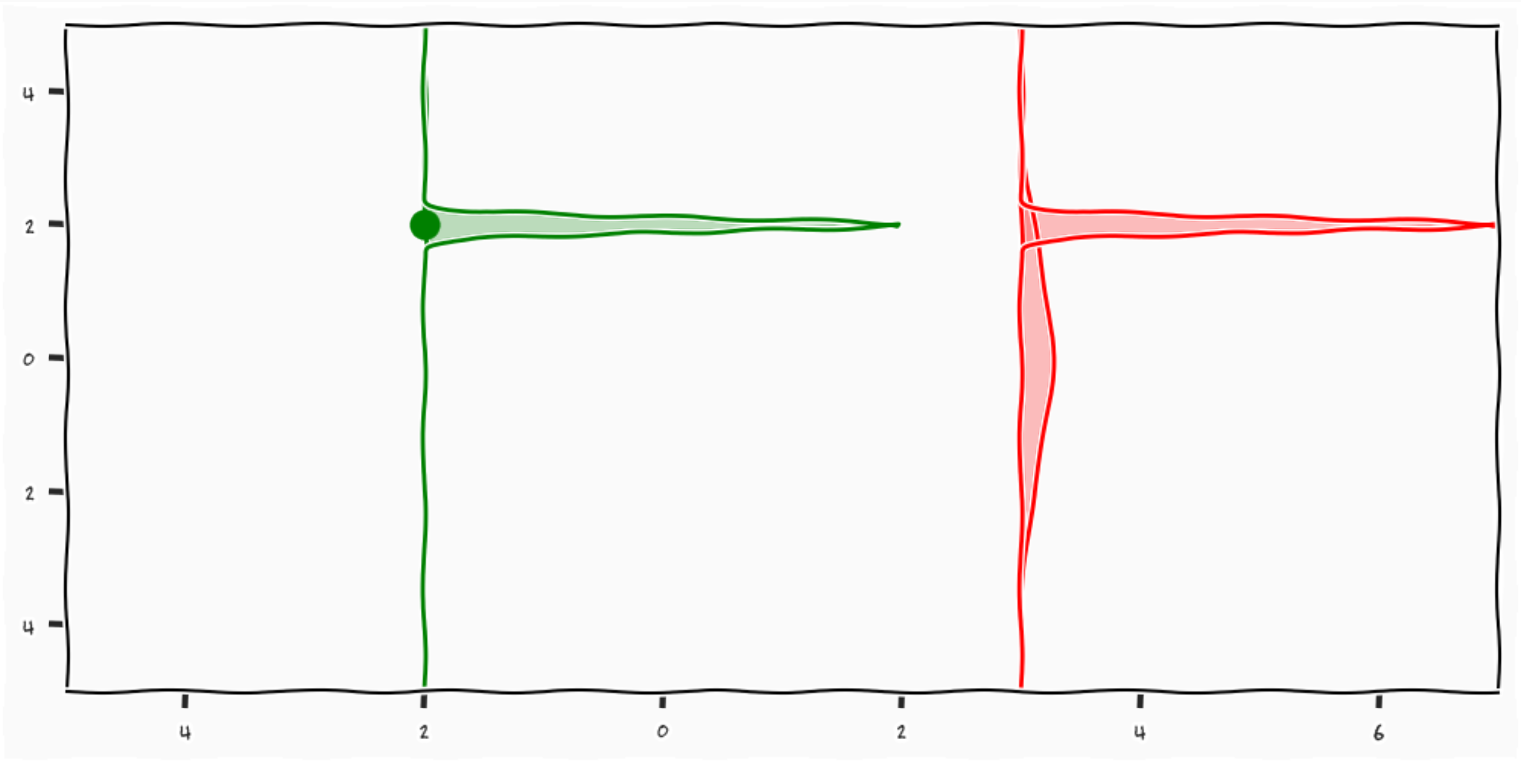
Gaussian Processes



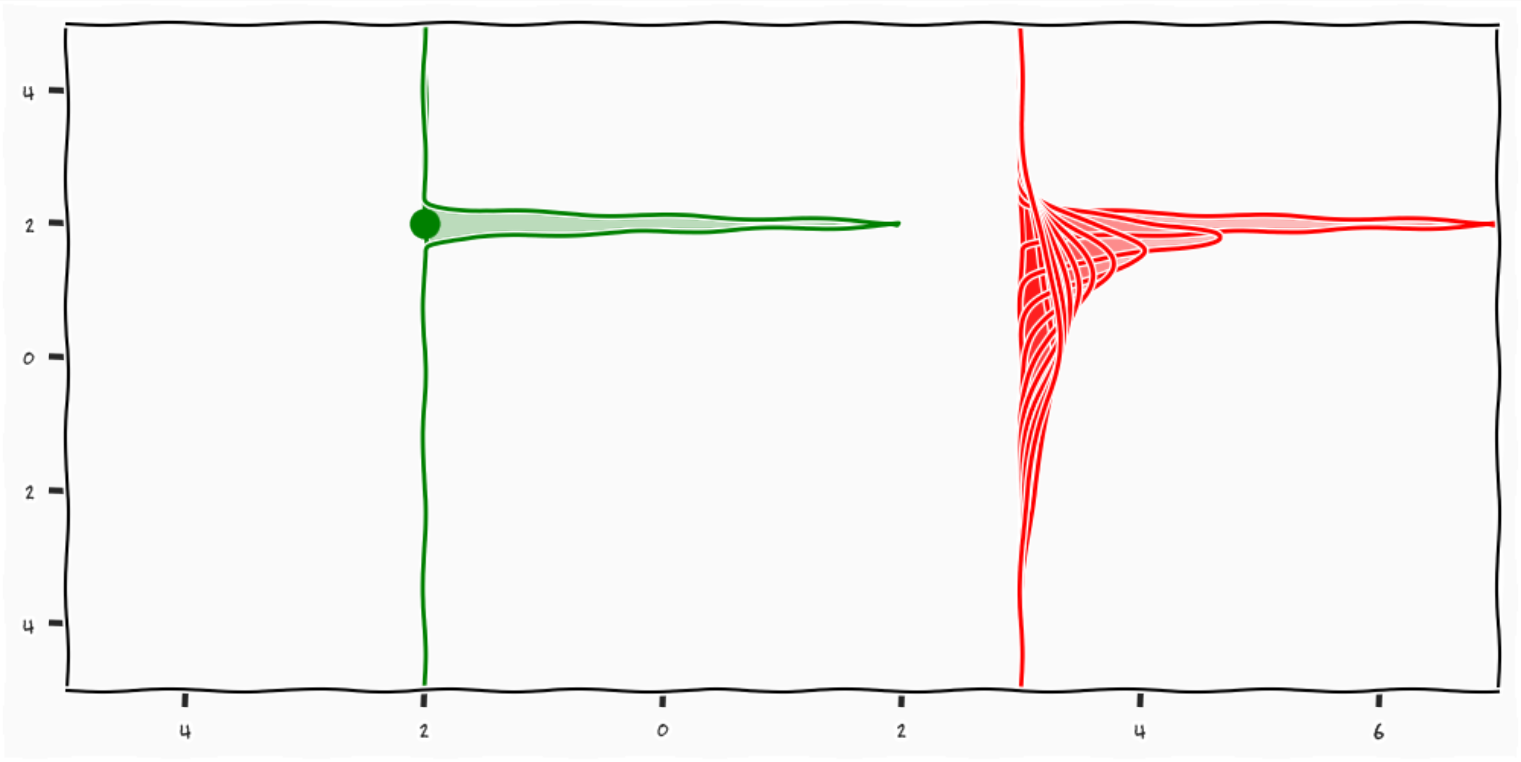
Gaussian Processes



Gaussian Processes



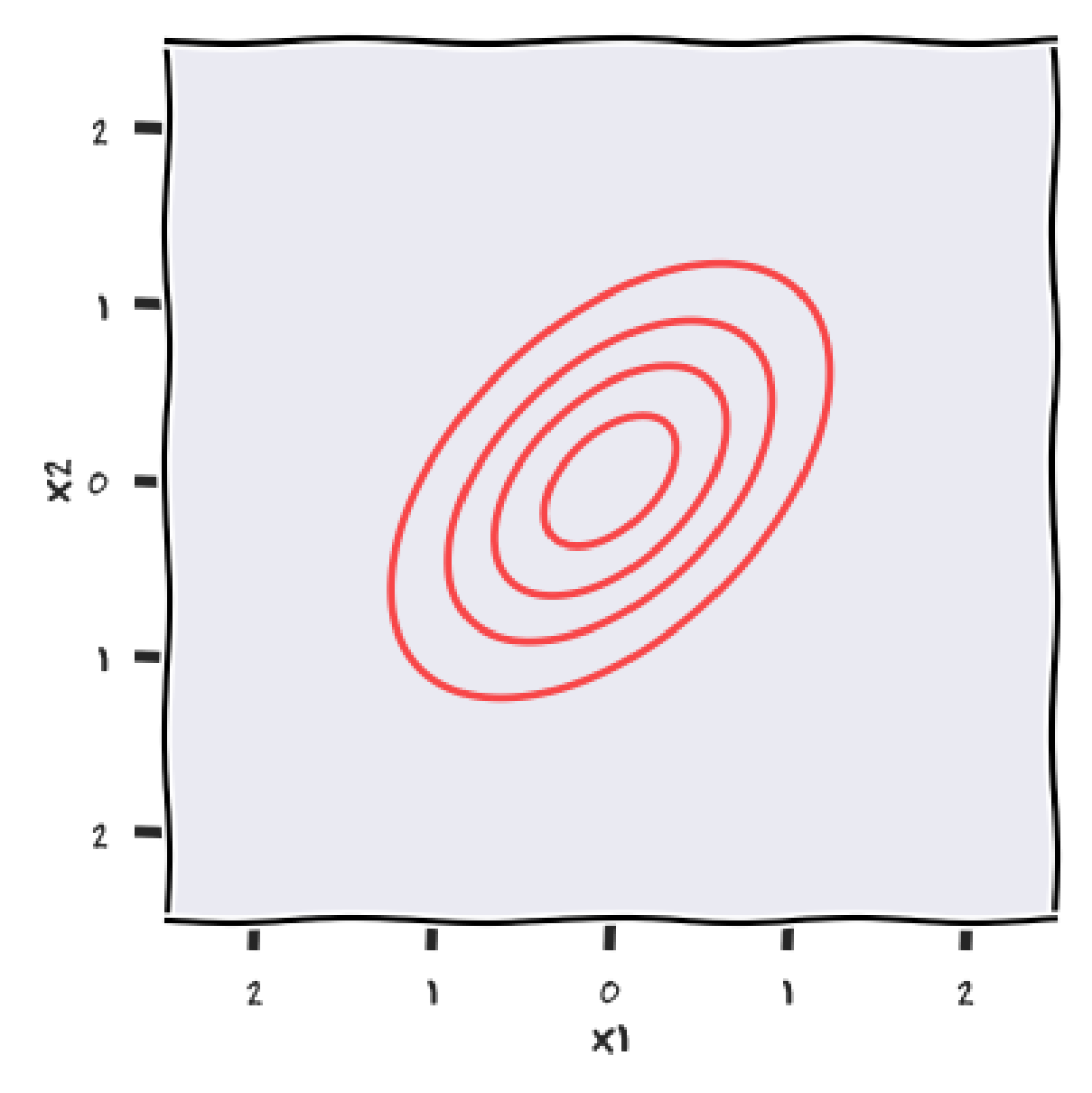
Gaussian Processes



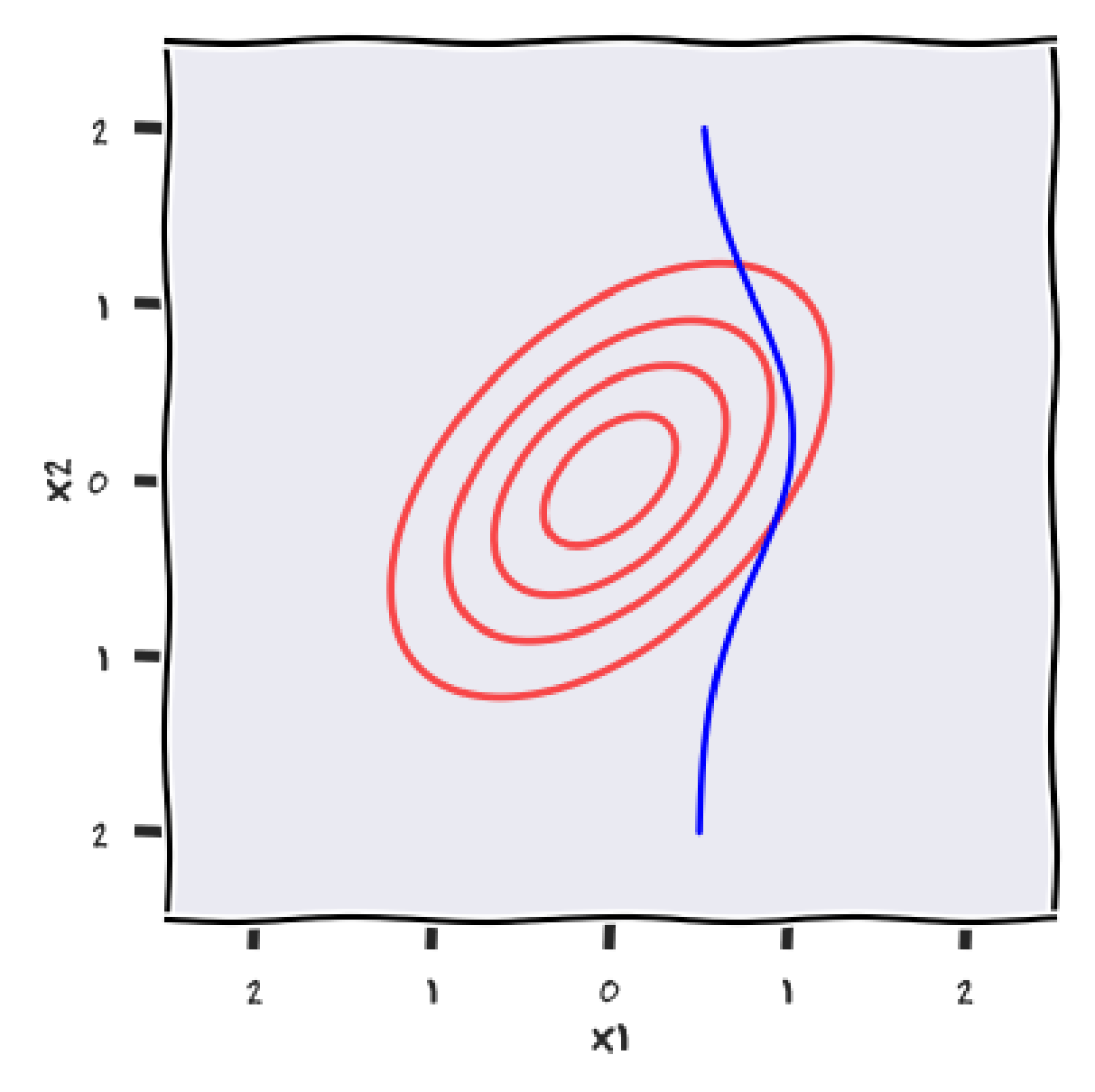
Conditional Gaussians

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

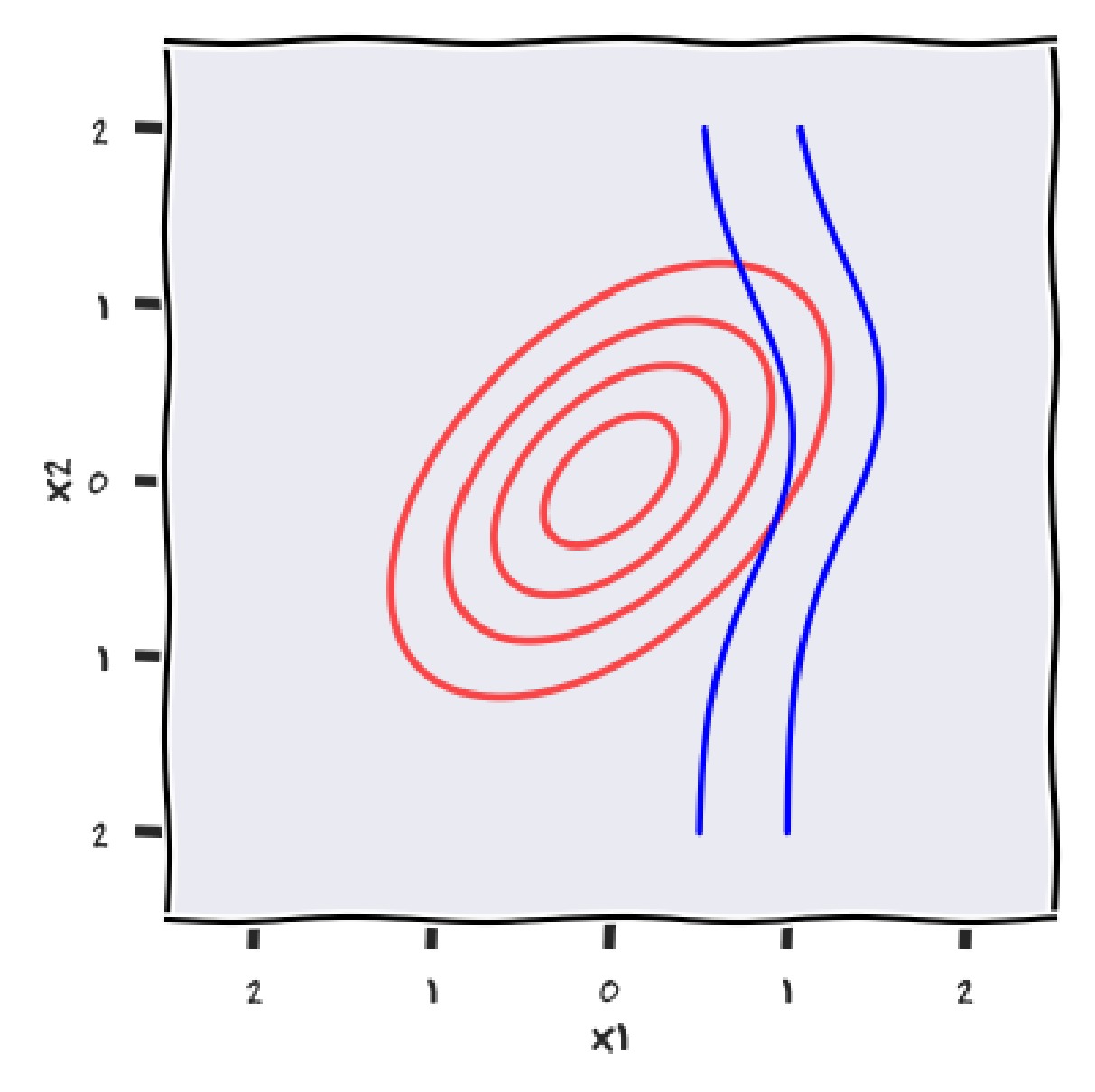
Conditional Gaussians



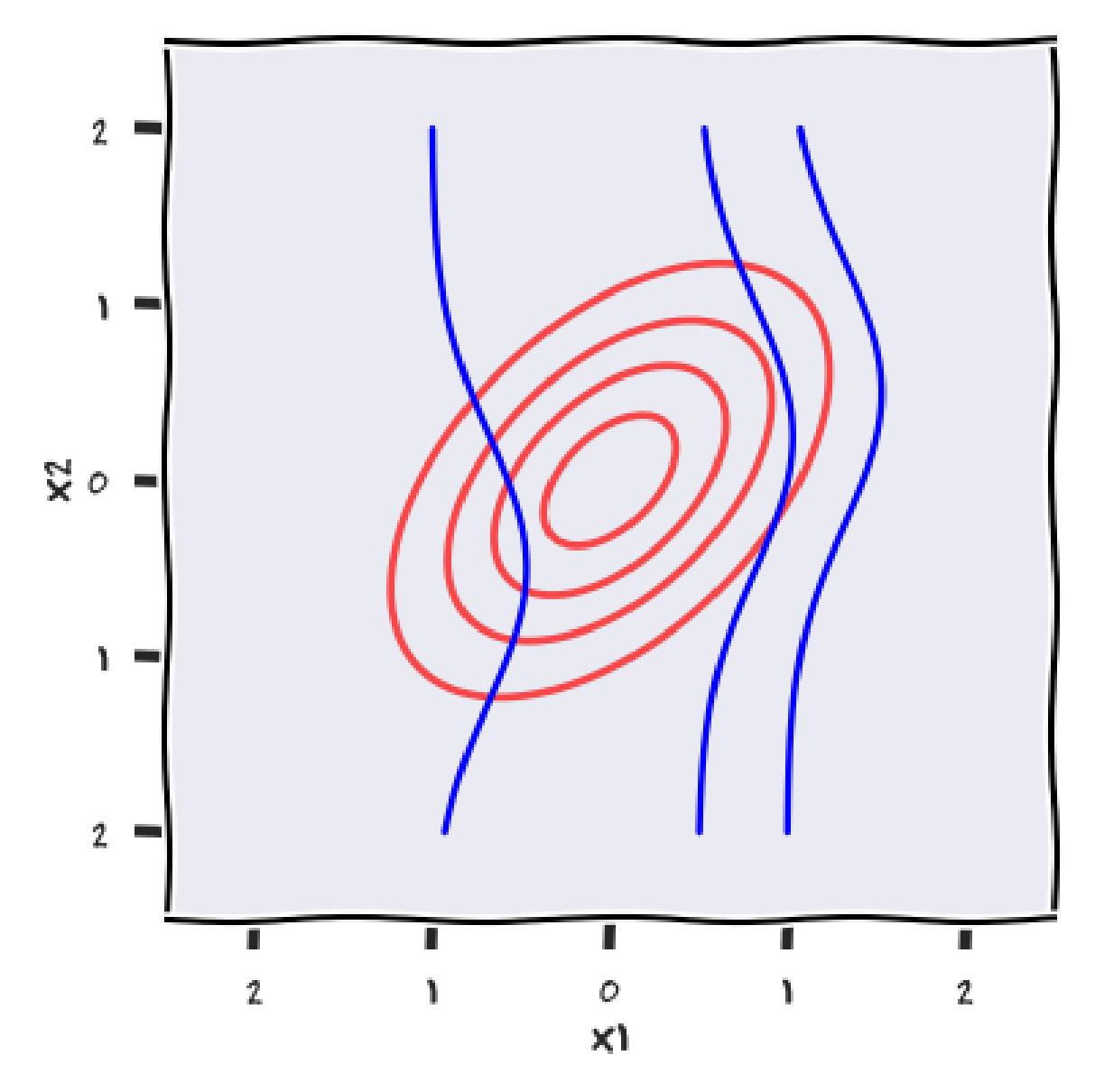
Conditional Gaussians



Conditional Gaussians



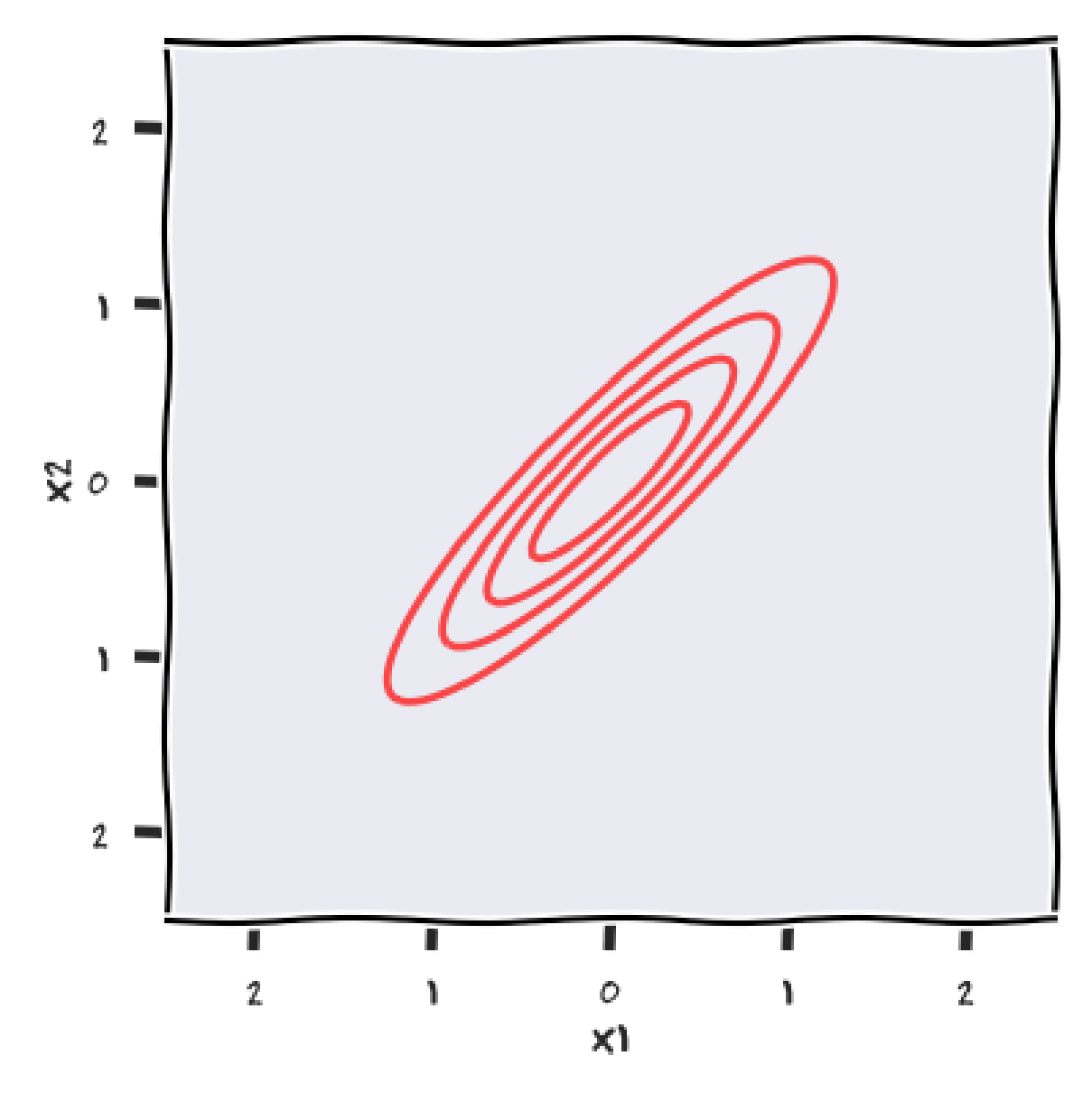
Conditional Gaussians



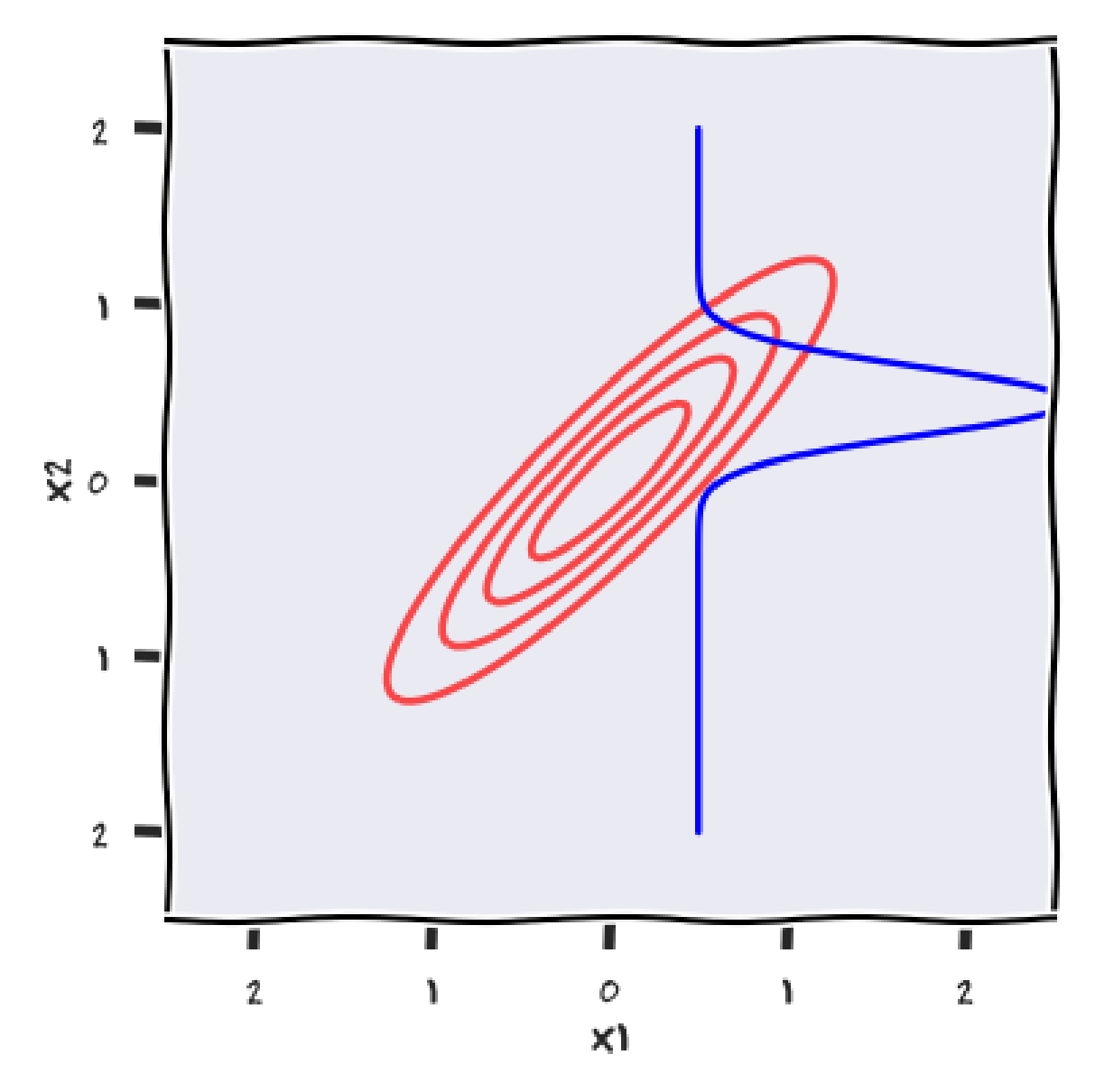
Conditional Gaussians

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$

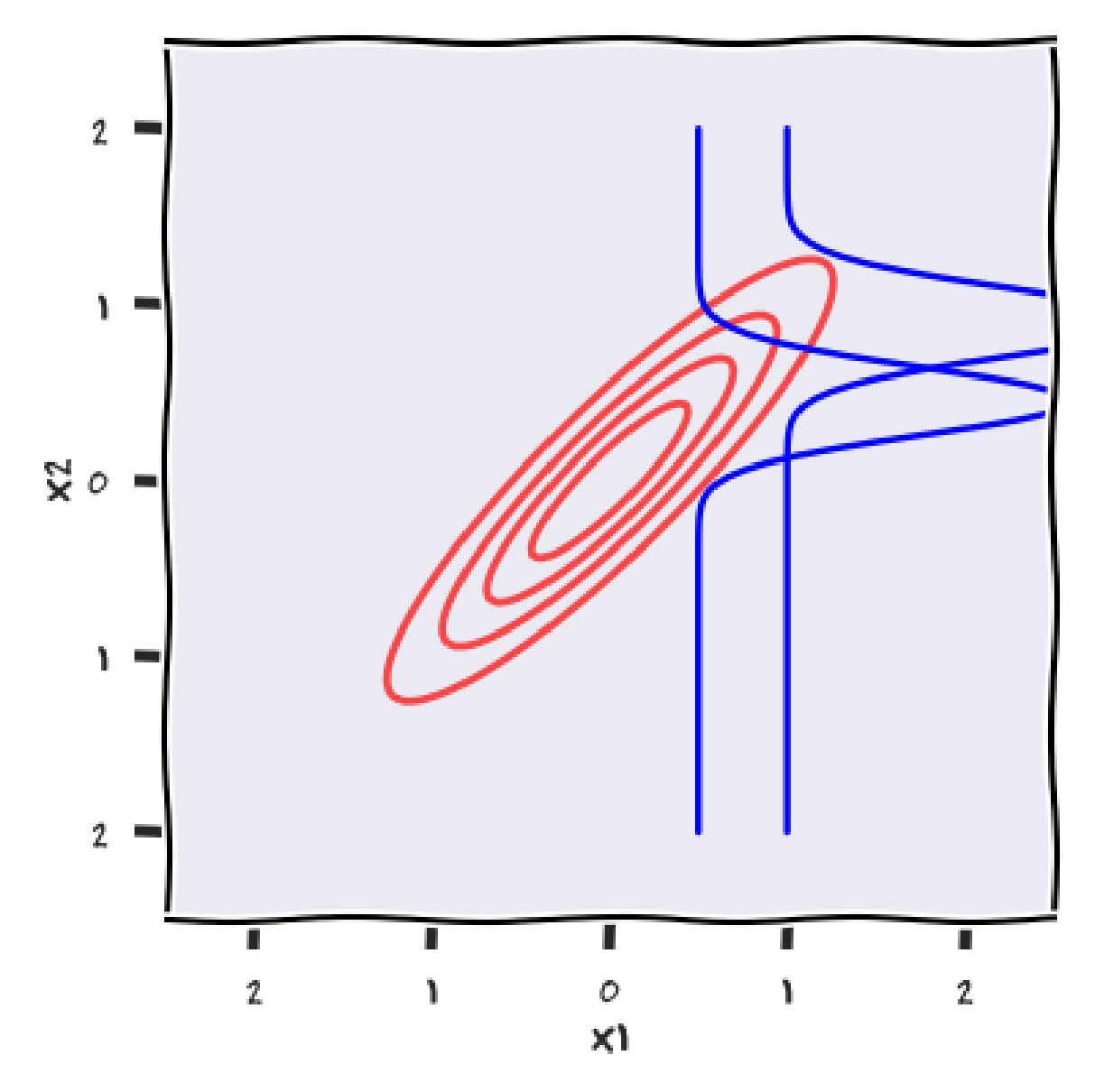
Conditional Gaussians



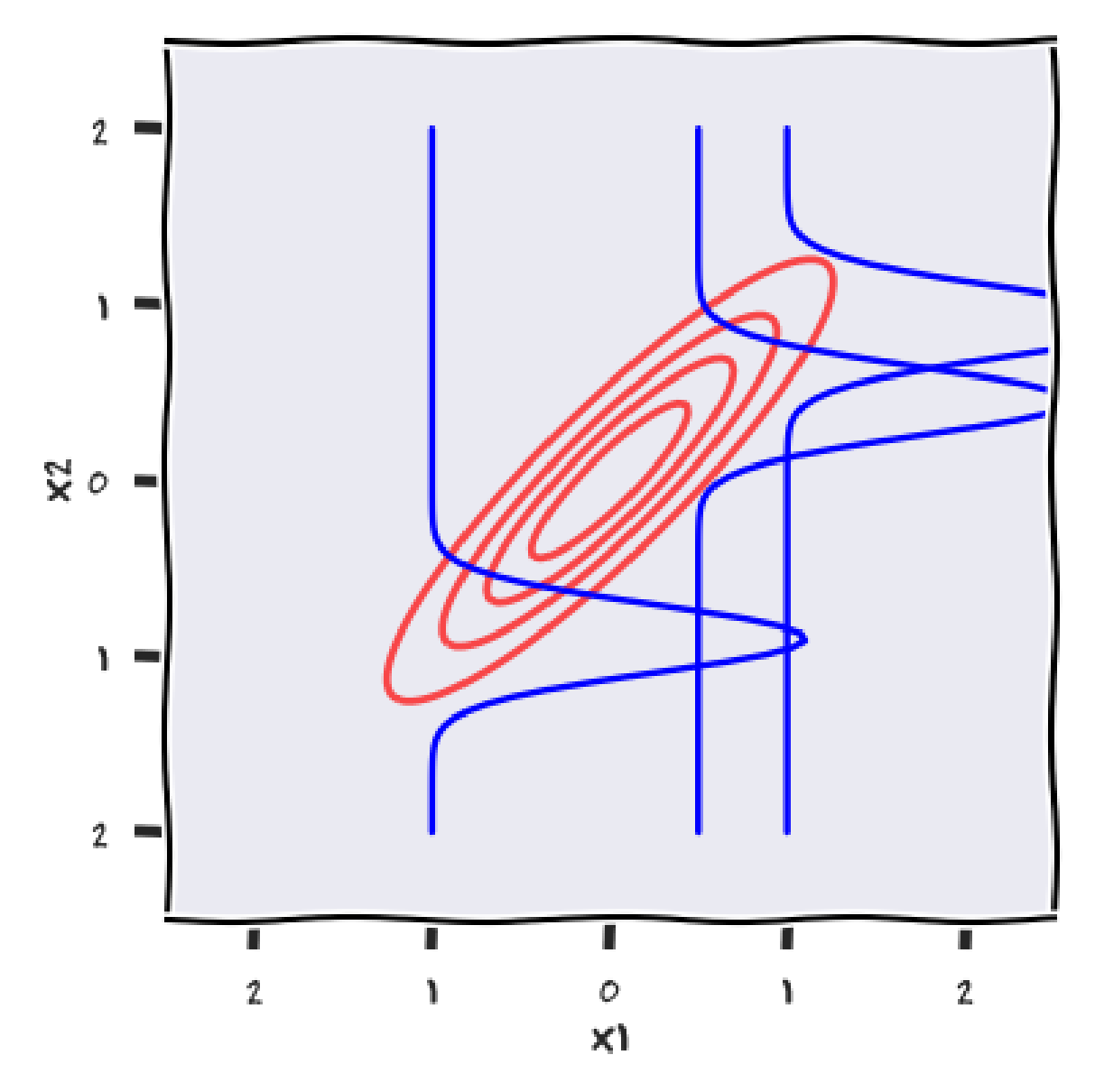
Conditional Gaussians



Conditional Gaussians



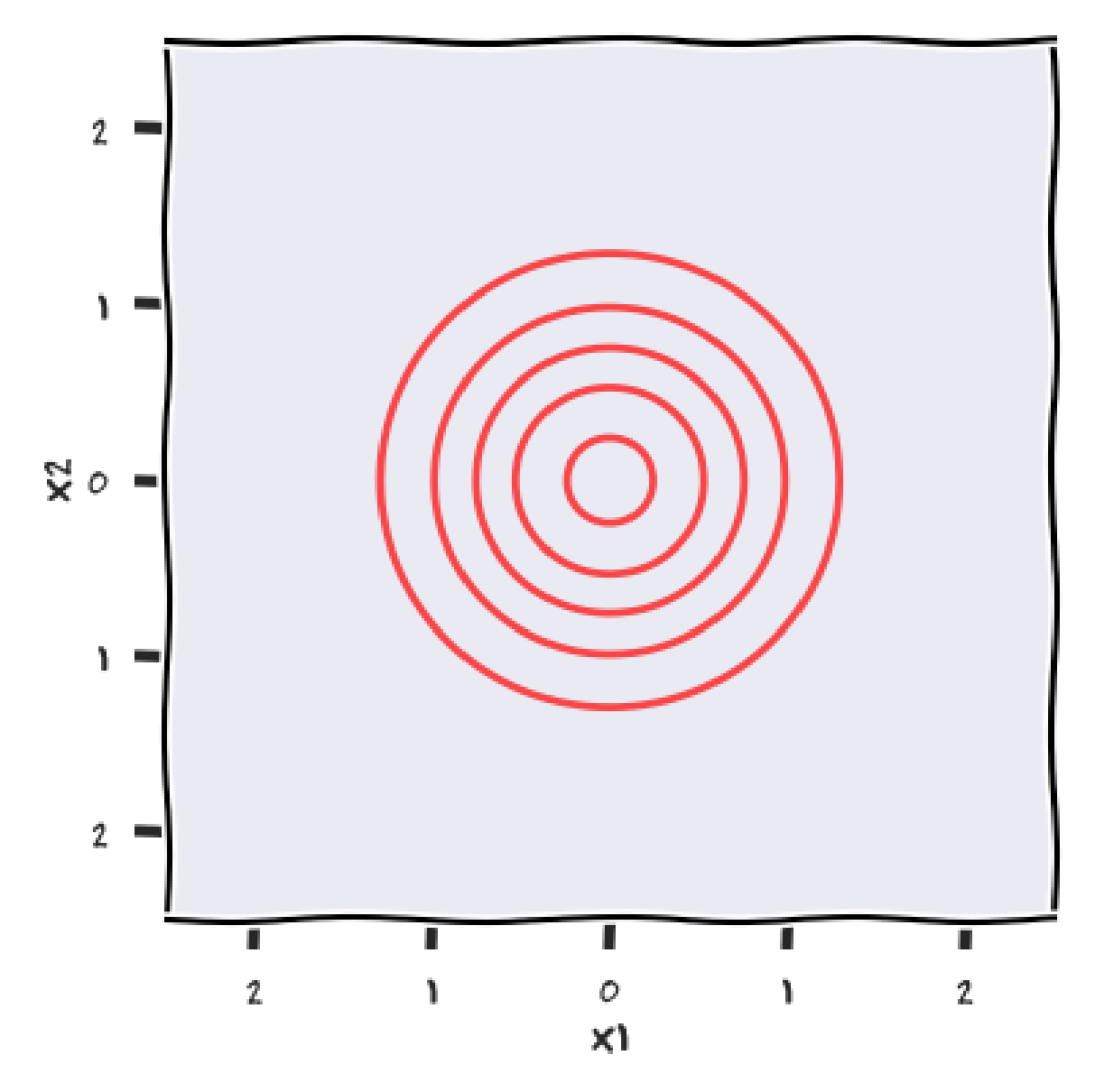
Conditional Gaussians



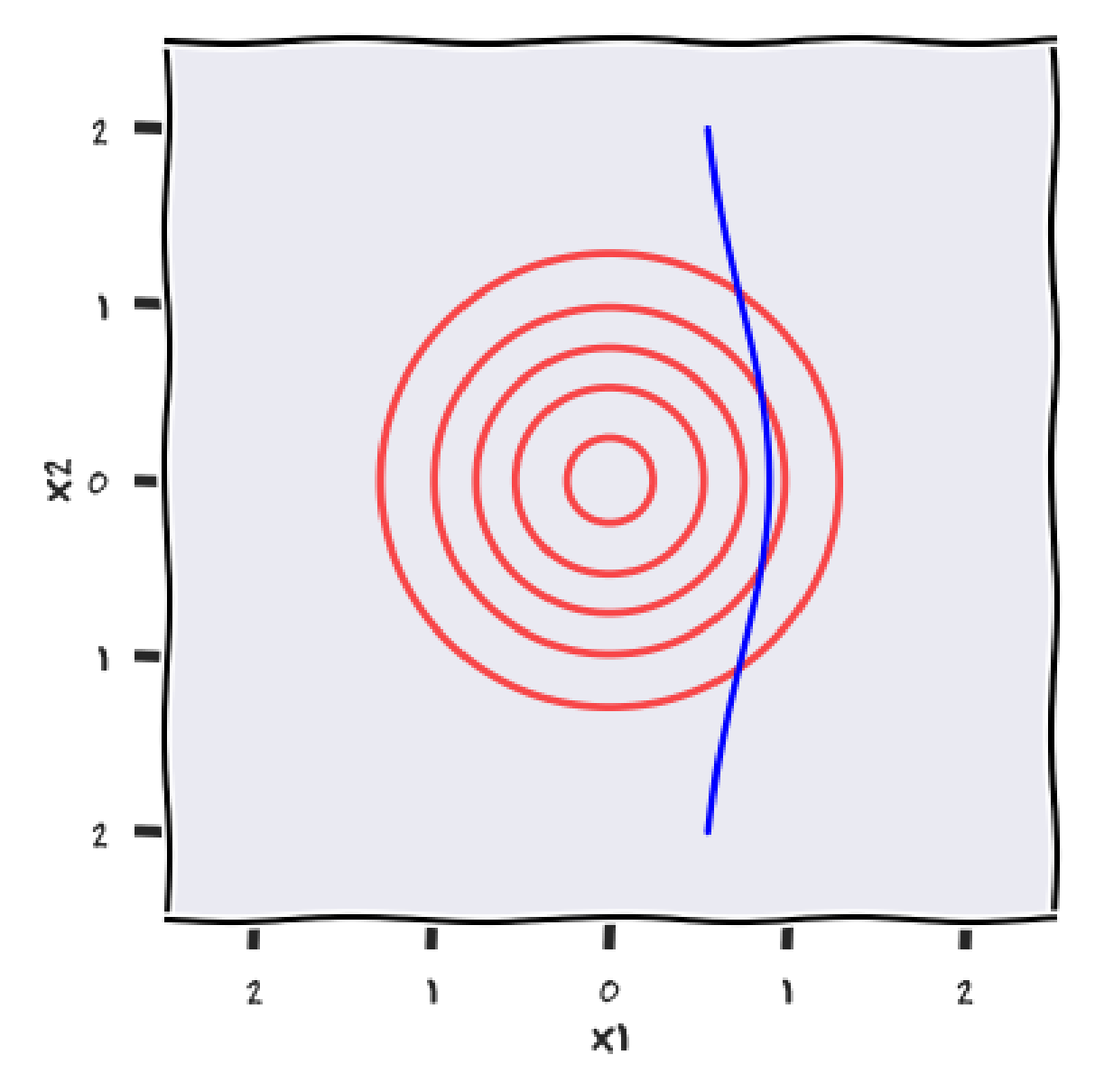
Conditional Gaussians

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

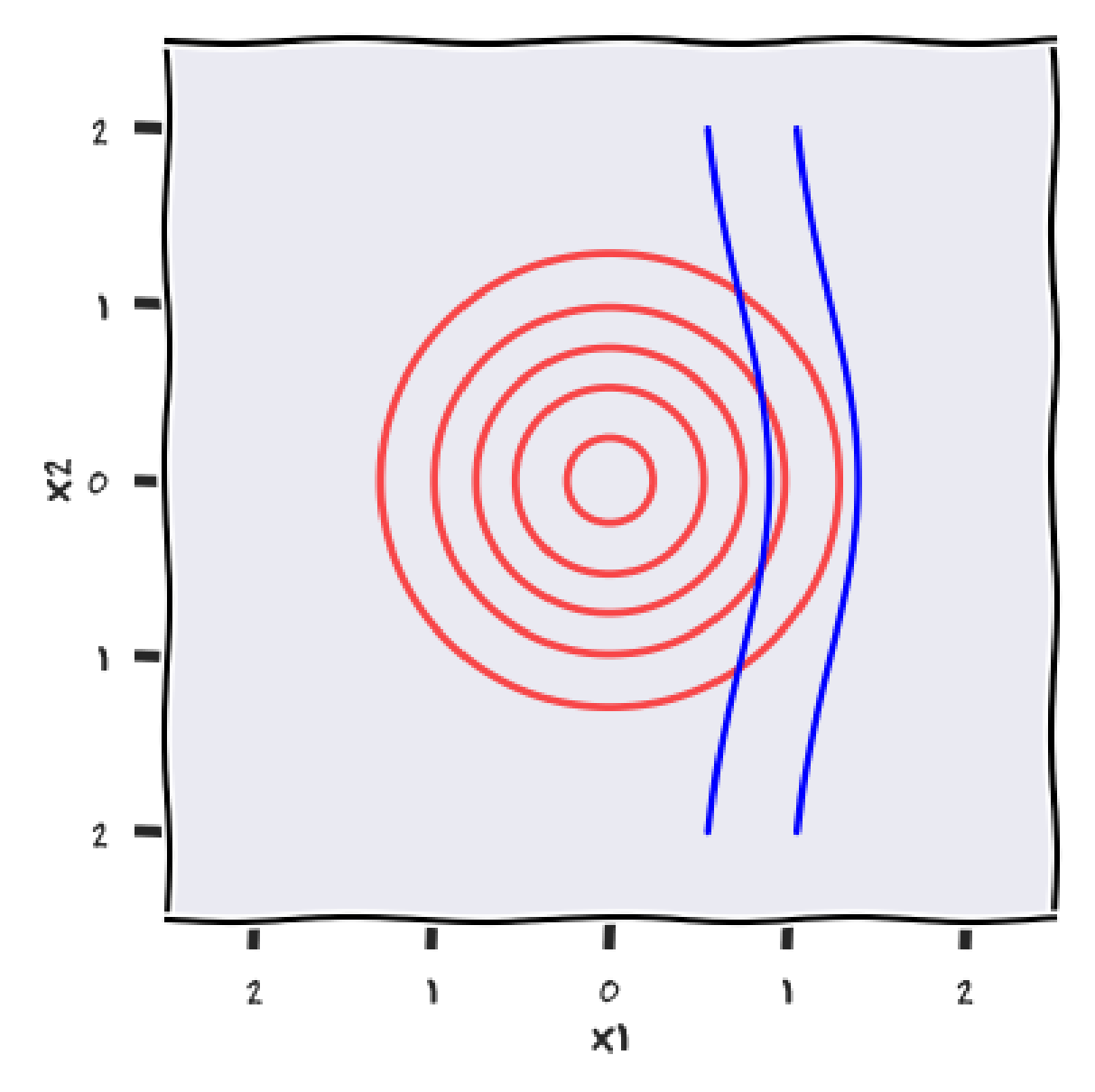
Conditional Gaussians



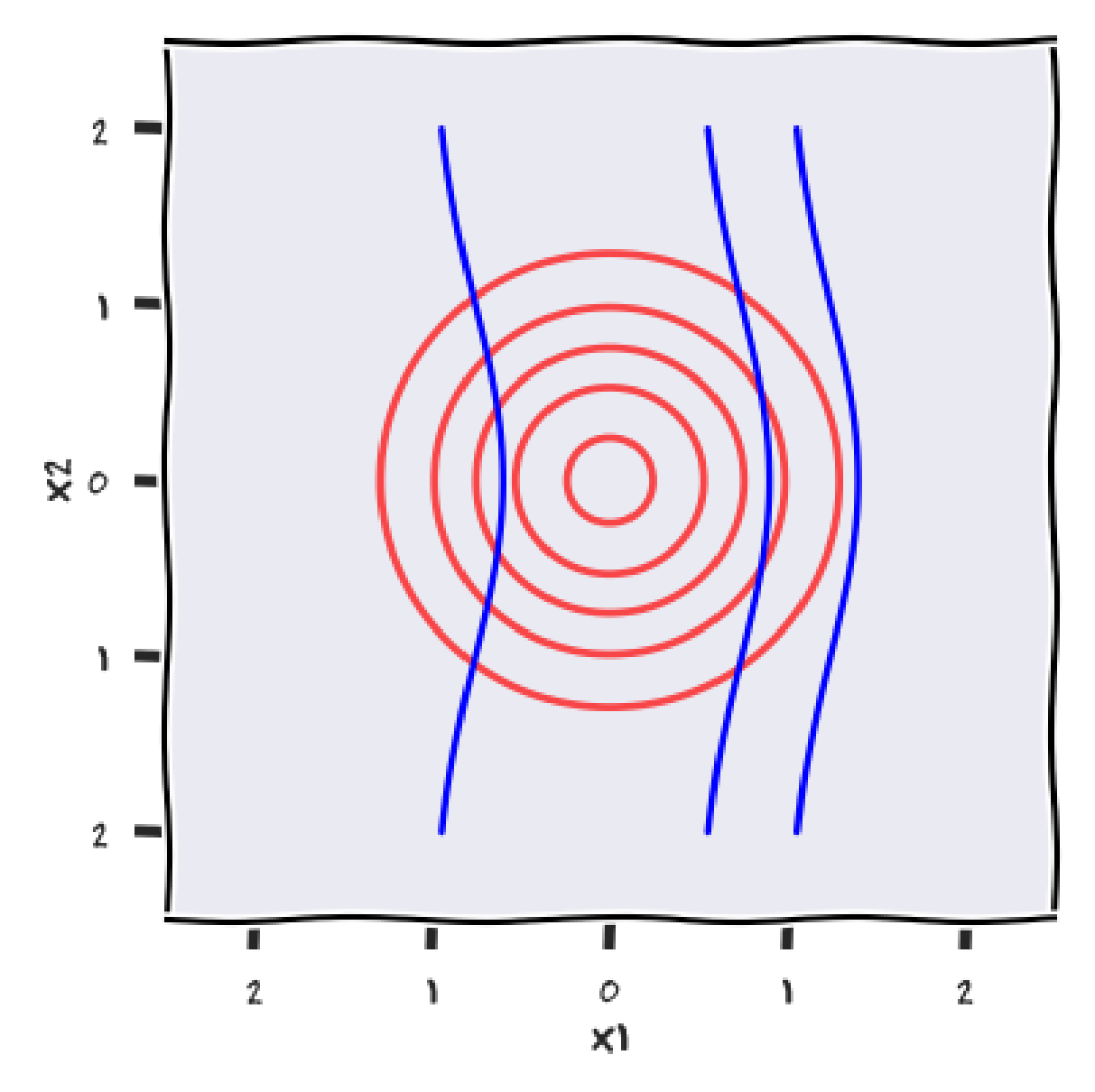
Conditional Gaussians



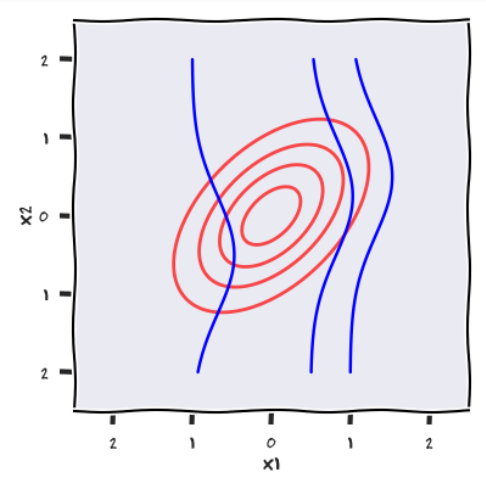
Conditional Gaussians



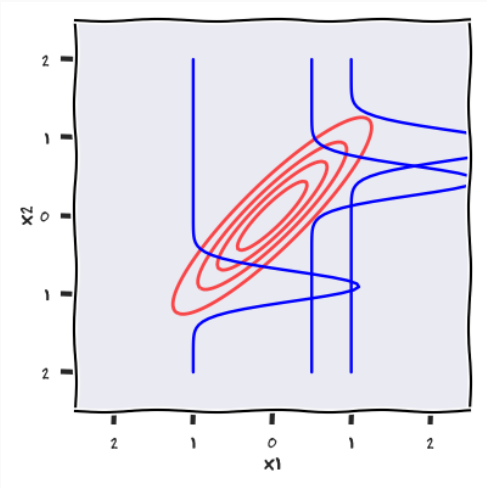
Conditional Gaussians



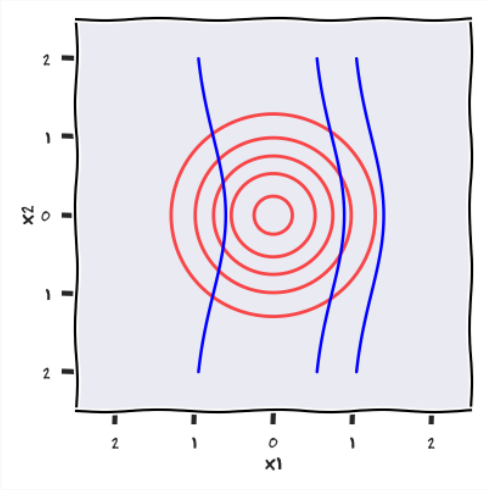
Conditional Gaussians



$$\mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

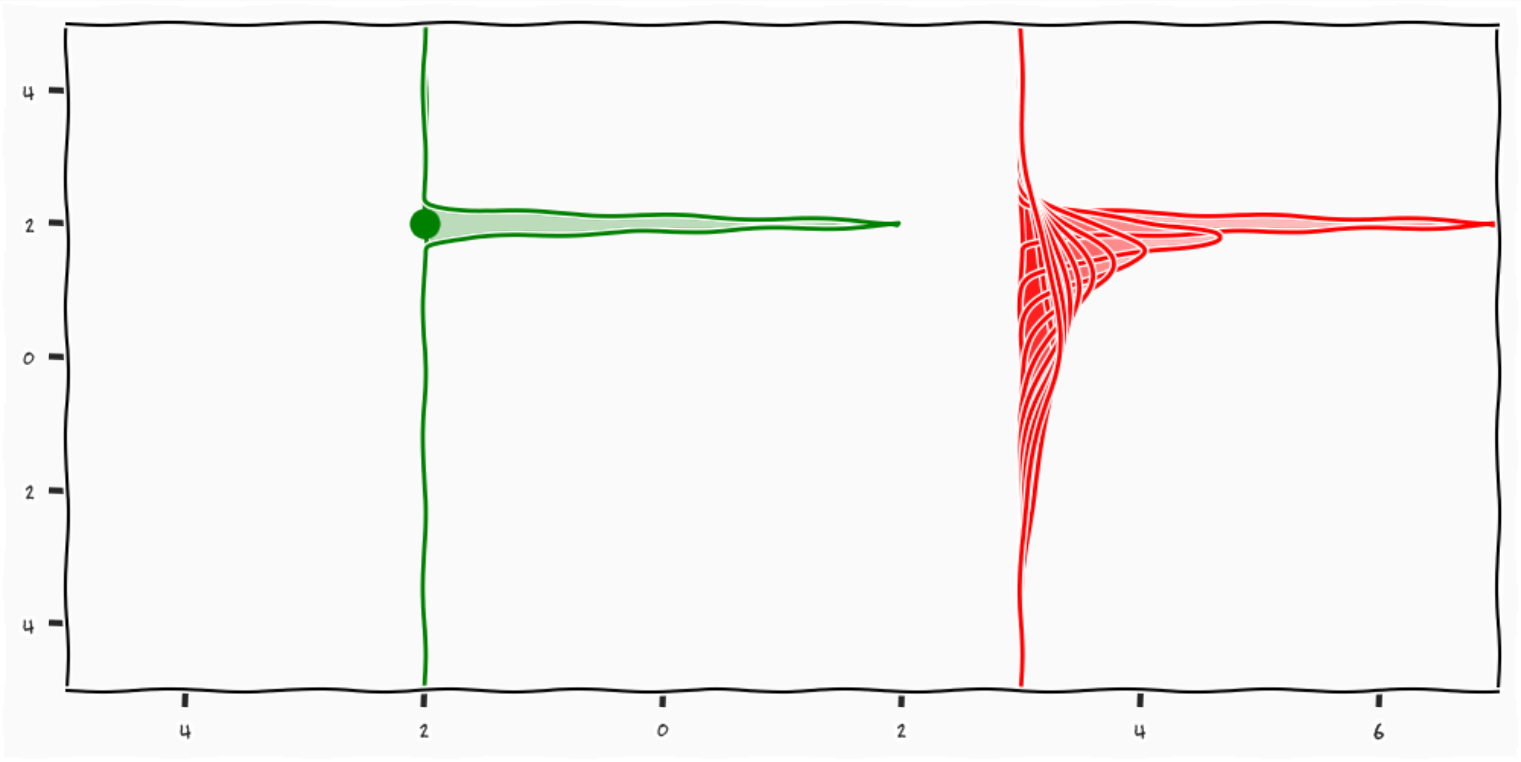


$$\mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$



$$\mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

Gaussian Processes



Gaussian Processes

$$p(\mathbf{f}|\mathbf{x}, \boldsymbol{\theta}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

Definition (Gaussian Process)

A Gaussian Process is an infinite collection of random variables who any subset is jointly gaussian. The process is specified by a mean function $\mu(\cdot)$ and a co-variance function $k(\cdot, \cdot)$

Gaussian Marginal

$$p(f_1, f_2, \ldots, f_N, \ldots | \mathbf{x}, \boldsymbol{\theta}) = \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

$$= \mathcal{N}\left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) & \cdots \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \right)$$

Gaussian Processes

$$
\begin{aligned}
p(\mathbf{f}|\mathbf{x}, \boldsymbol{\theta}) &\sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x})) \\
\mathbf{y}_i &= f_i + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \\
p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) &= \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{x}, \boldsymbol{\theta}) \mathrm{d}f
\end{aligned}
$$

$\mathcal{GP}$ is infinite, but we only observe finite amount of data. This means conditioning on a subset of the data, the $\mathcal{GP}$ is a just a Gaussian distribution

Uncertainty over functions

- Regression model,

$$\mathbf{y}_i = f(\mathbf{x}_i) + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

- Introduce f_i as *instantiation* of function at location x_i

$$f_i = f(\mathbf{x}_i),$$

- as a new random variable.
- now we have a "handle" to specify our assumptions over

Uncertainty over functions

Joint

$$p(\mathbf{y}, \mathbf{f}|\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x}, \boldsymbol{\theta})$$

Prior

$$p(\mathbf{f}|\mathbf{x}, \boldsymbol{\theta}),$$

Likelihood

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i}^{N} p(y_i|f_i)$$

Gaussian Processes

The Mean Function

- Function of only the input location
- What do I expect the function value to be only accounting for the input location

The Covariance Function

- Function of two input locations
- How should the information from other locations with known function value observations effect my estimate

Gaussian Processes

The Mean Function

- Function of only the input location
- What do I expect the function value to be only accounting for the input location
- We will assume this to be constant

The Covariance Function

- Function of two input locations
- How should the information from other locations with known function value observations effect my estimate

Gaussian Processes

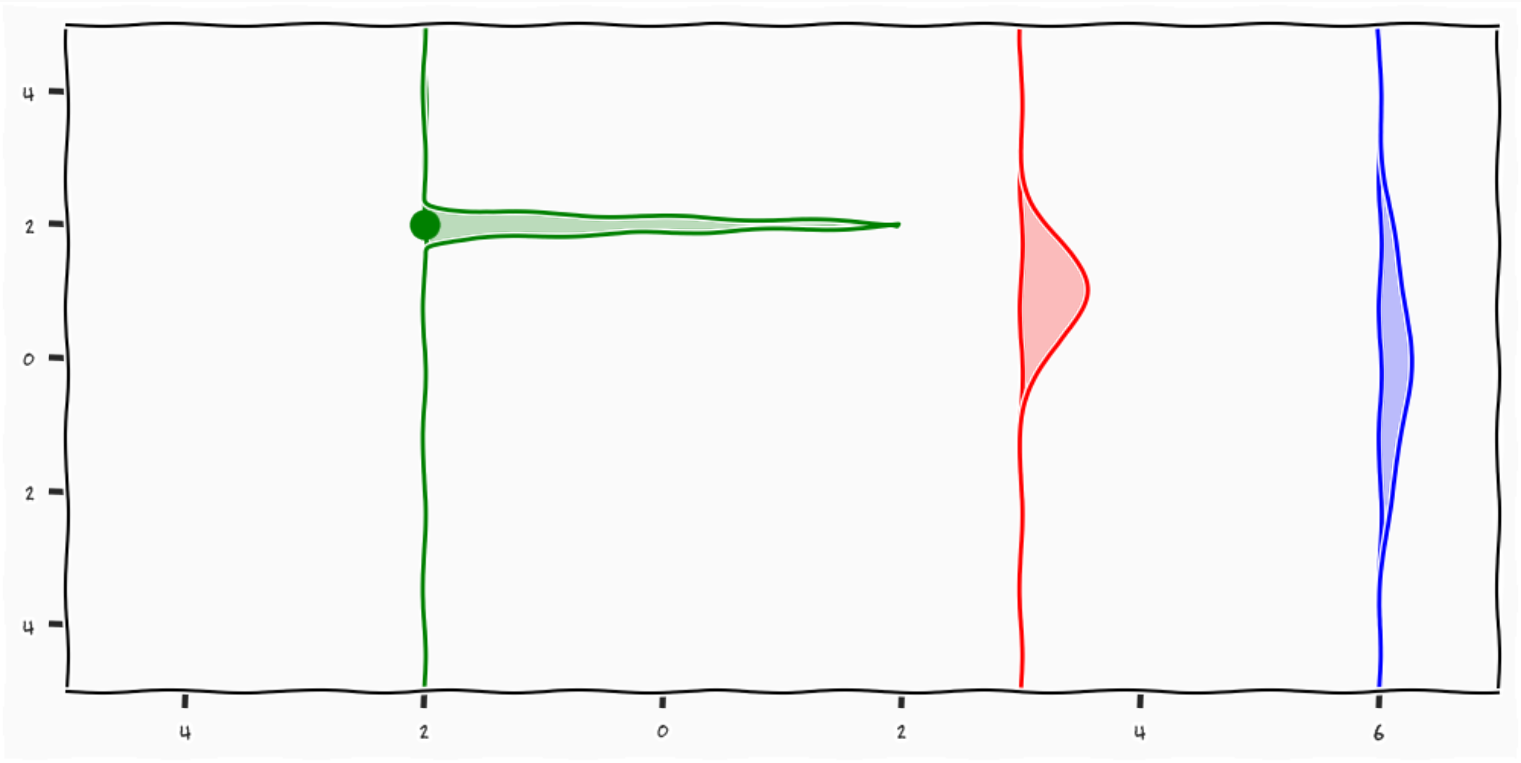
The Mean Function

- Function of only the input location
- What do I expect the function value to be only accounting for the input location
- We will assume this to be constant

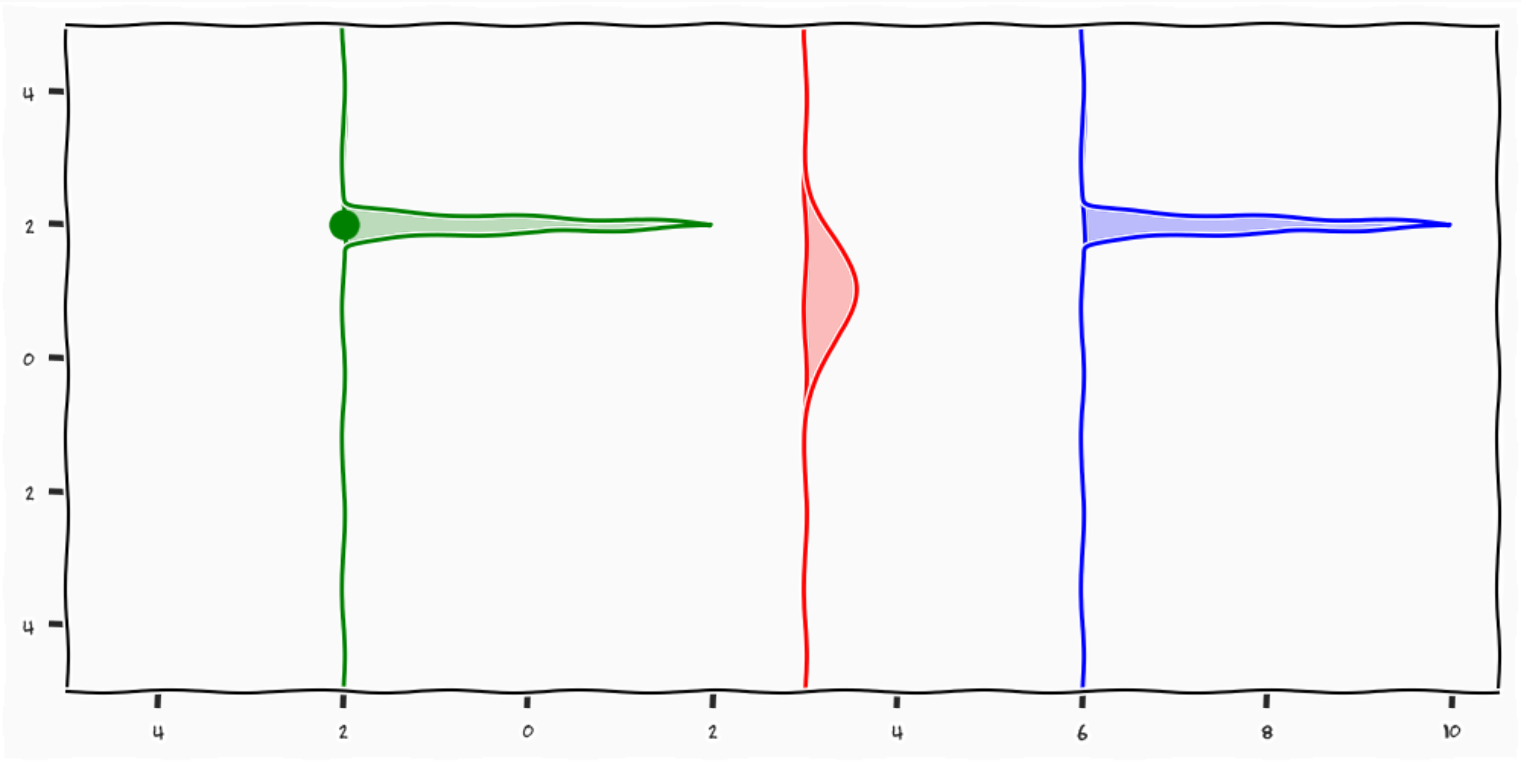
The Covariance Function

- Function of two input locations
- How should the information from other locations with known function value observations effect my estimate
- Encodes the behavior of the function

Gaussian Processes



Gaussian Processes

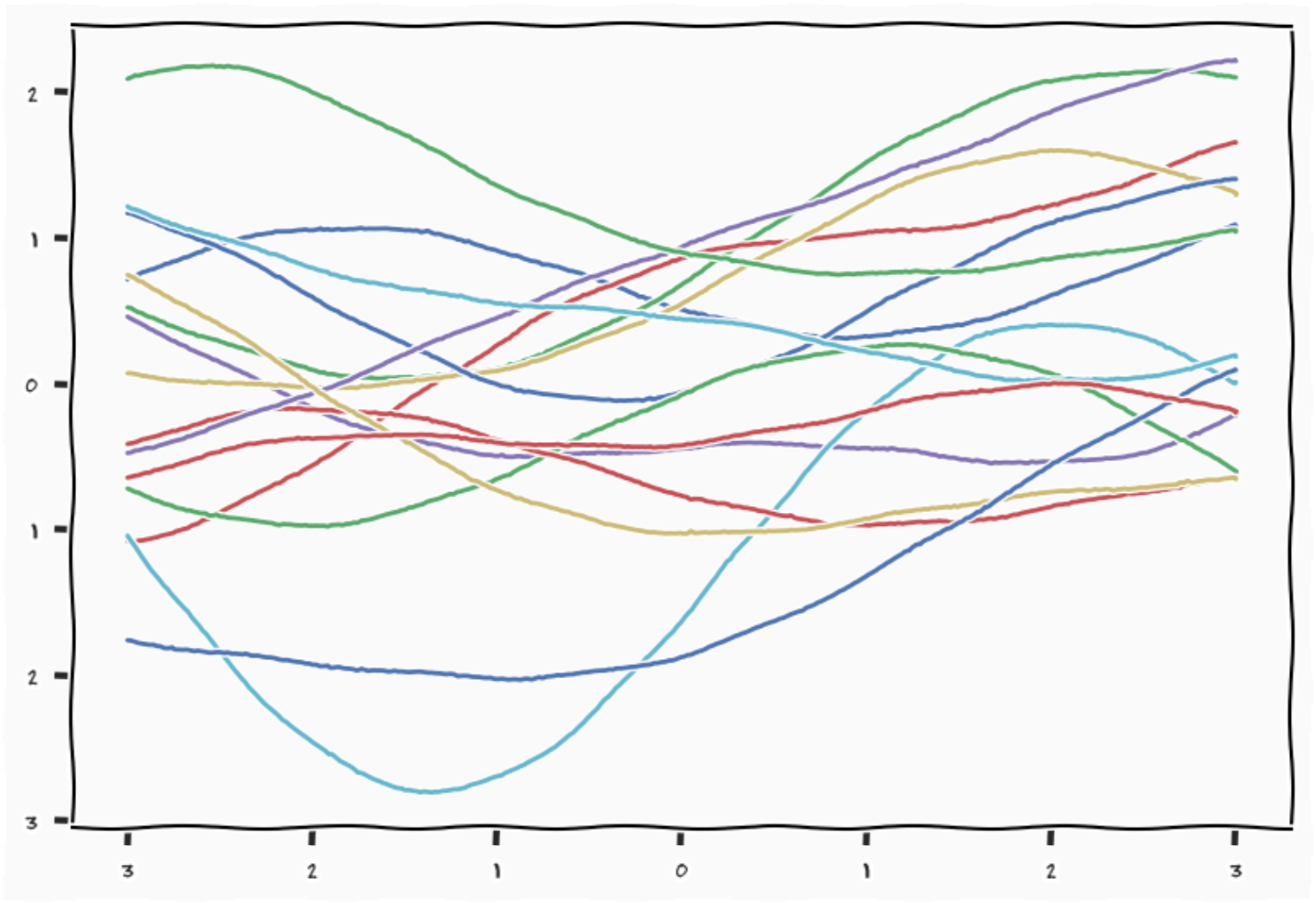


Gaussian Process: Samples

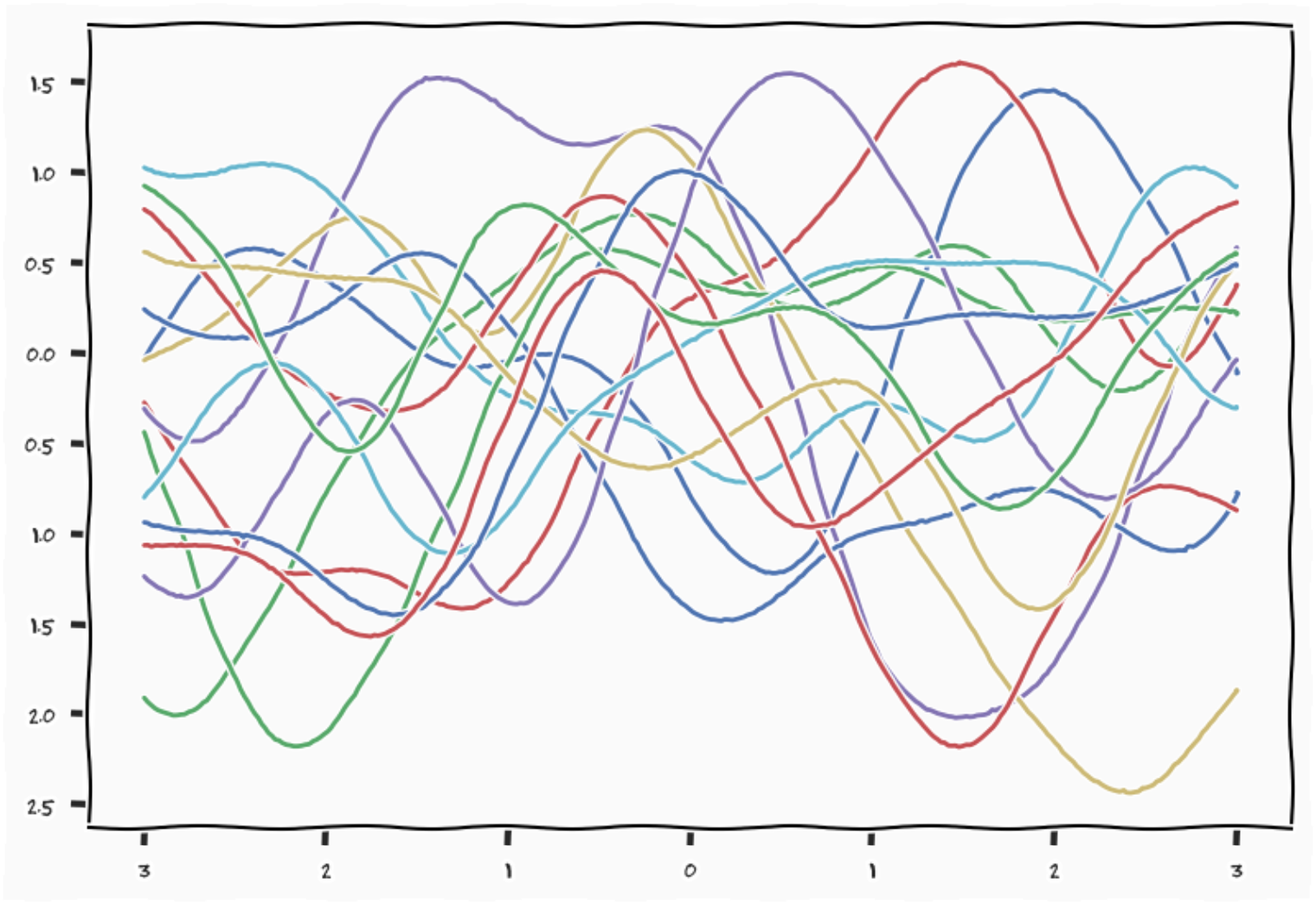
$$p(f_1, f_2, \ldots, f_N, \ldots | \mathbf{x}, \boldsymbol{\theta}) = \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

$$= \mathcal{N}\left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ t \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) & \cdots \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) & \cdots \end{bmatrix} \right)$$

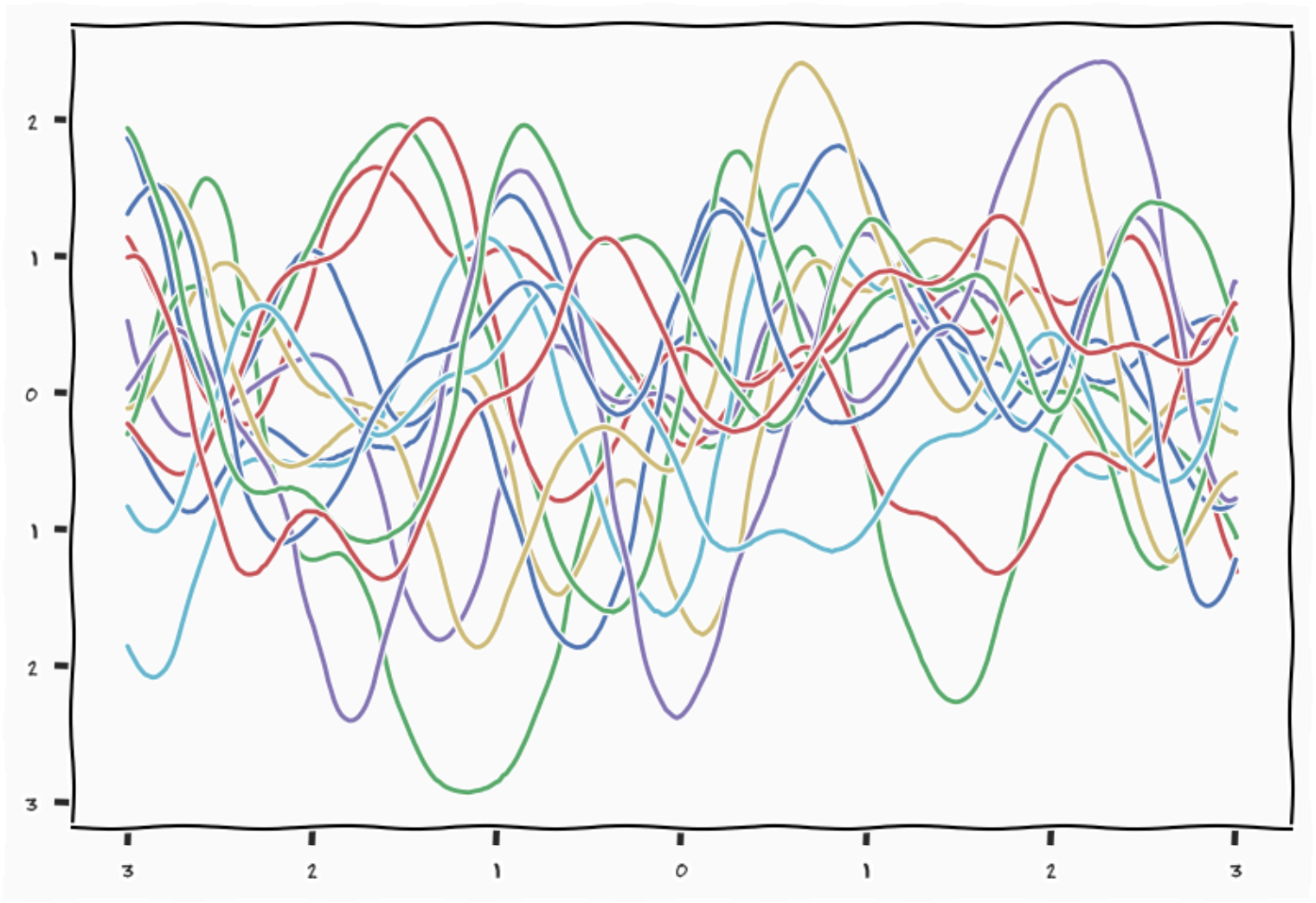
Gaussian Processes: Samples



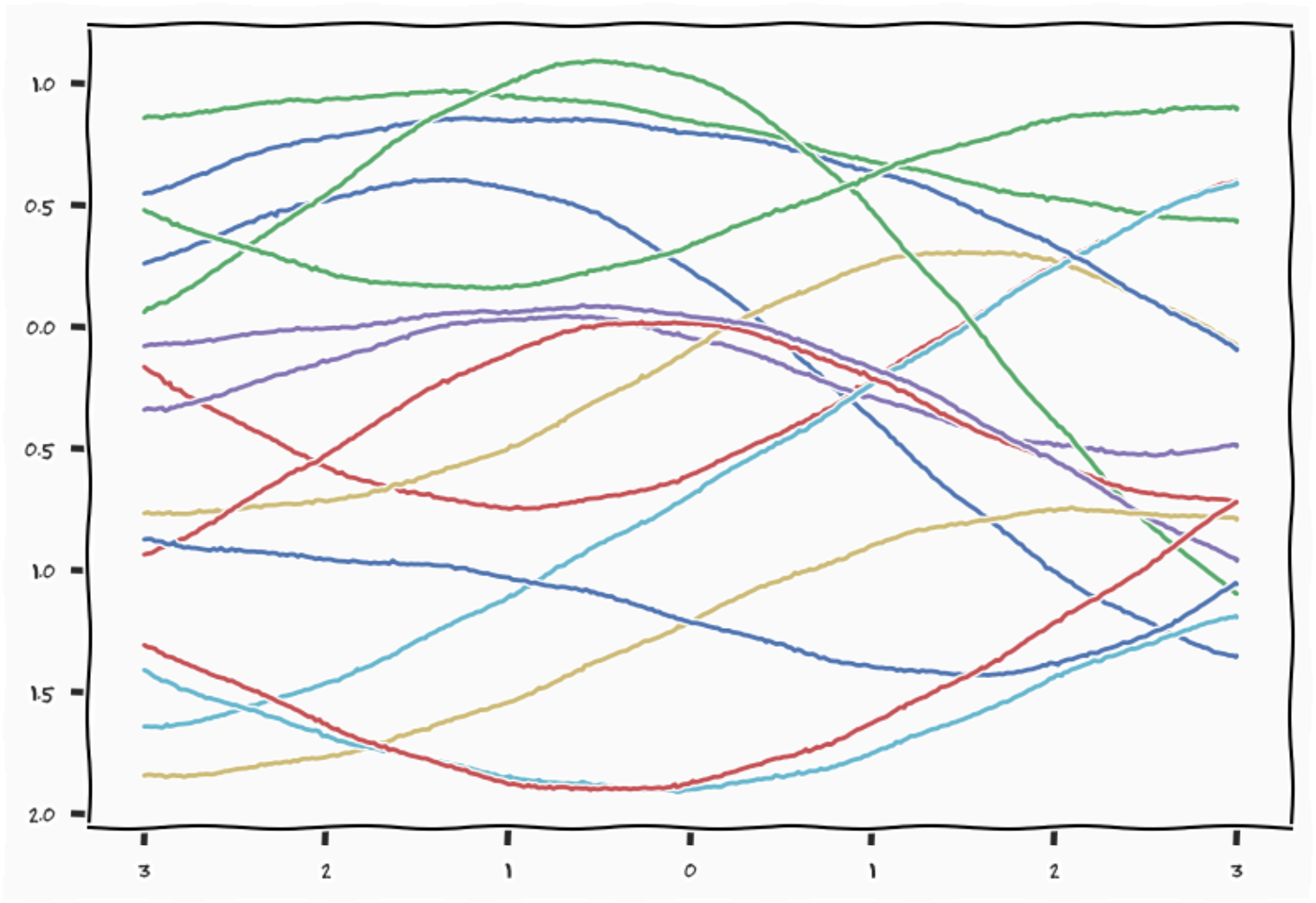
Gaussian Processes: Samples



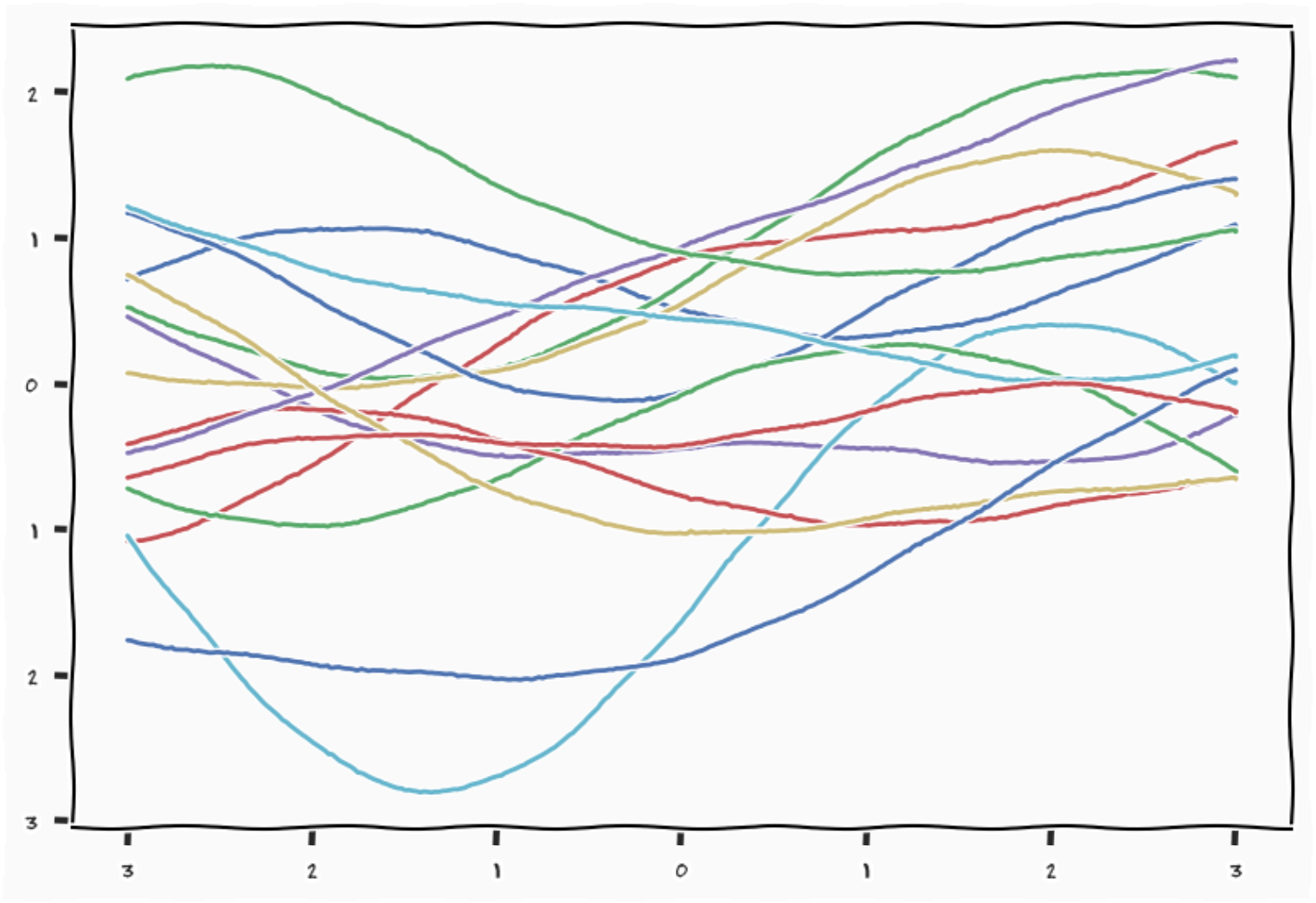
Gaussian Processes: Samples



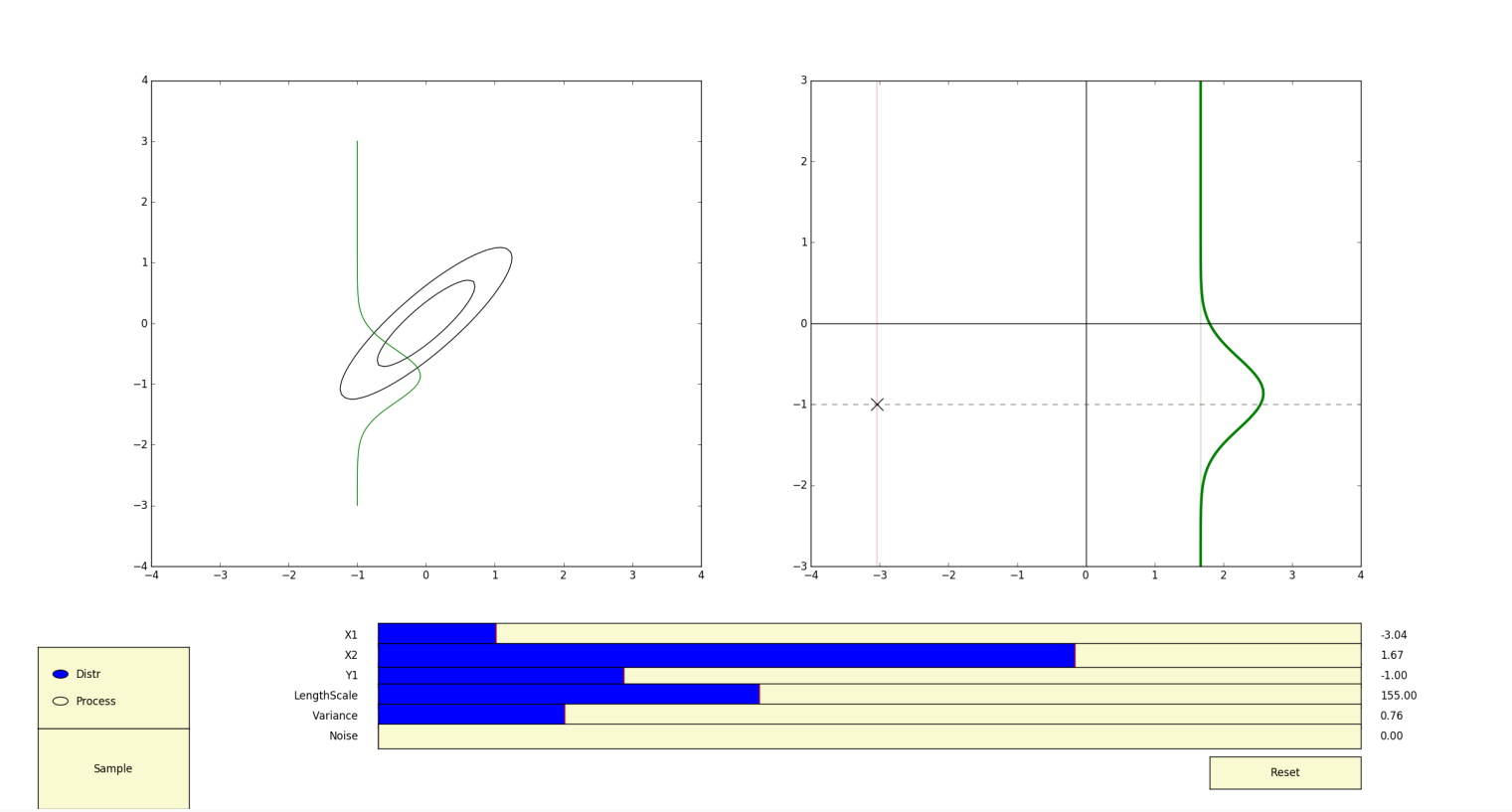
Gaussian Processes: Samples



Gaussian Processes: Samples



Demo


Distr
Process
Sample
X1
X2
Y1
LengthScale
Variance
Noise
-3.04
1.67
-1.00
155.00
0.76
0.00
Reset

Gaussian Process: Posterior

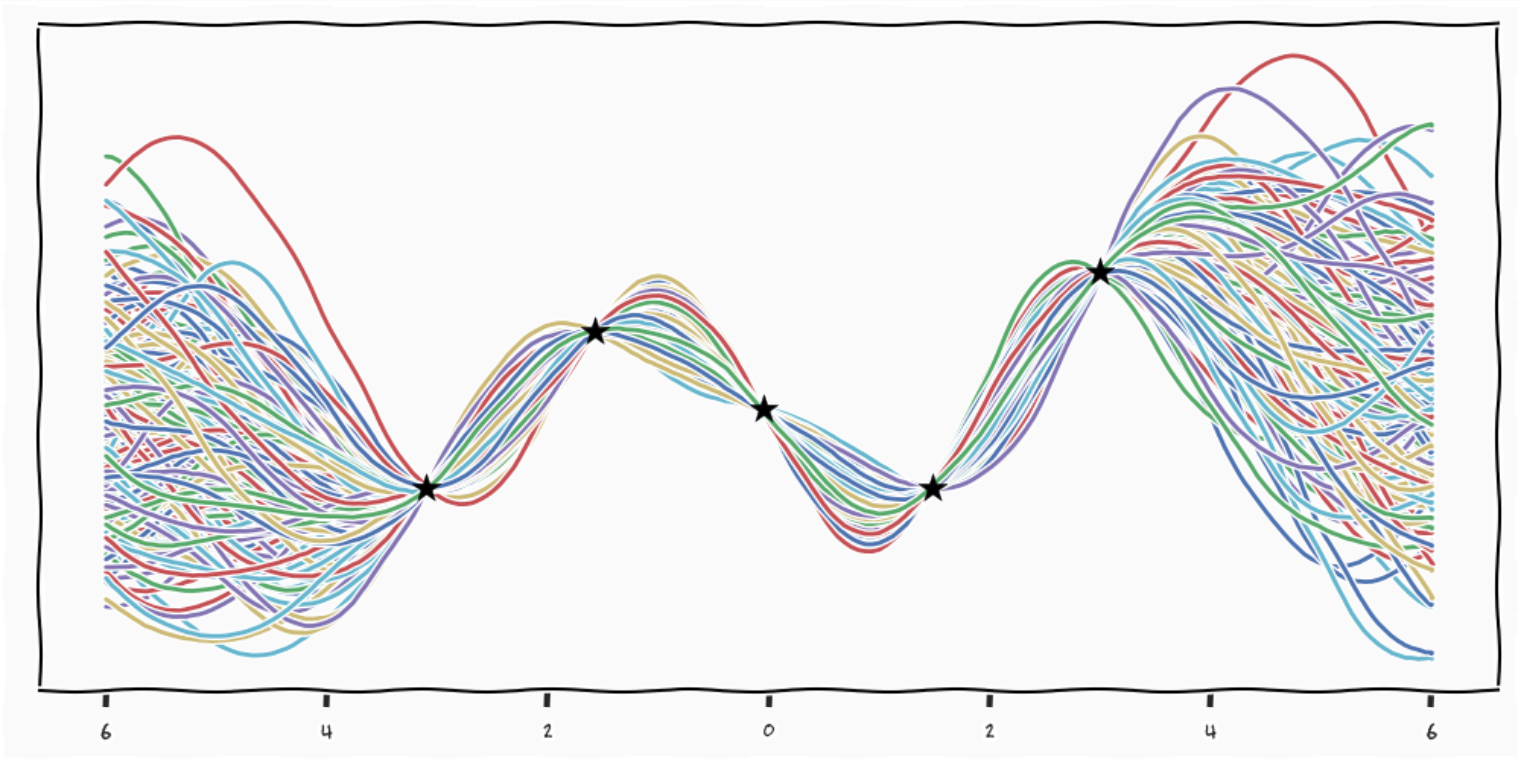
- All instantiations are jointly Gaussian

$$\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{X}, \mathbf{X}) & k(\mathbf{X}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{X}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right)$$

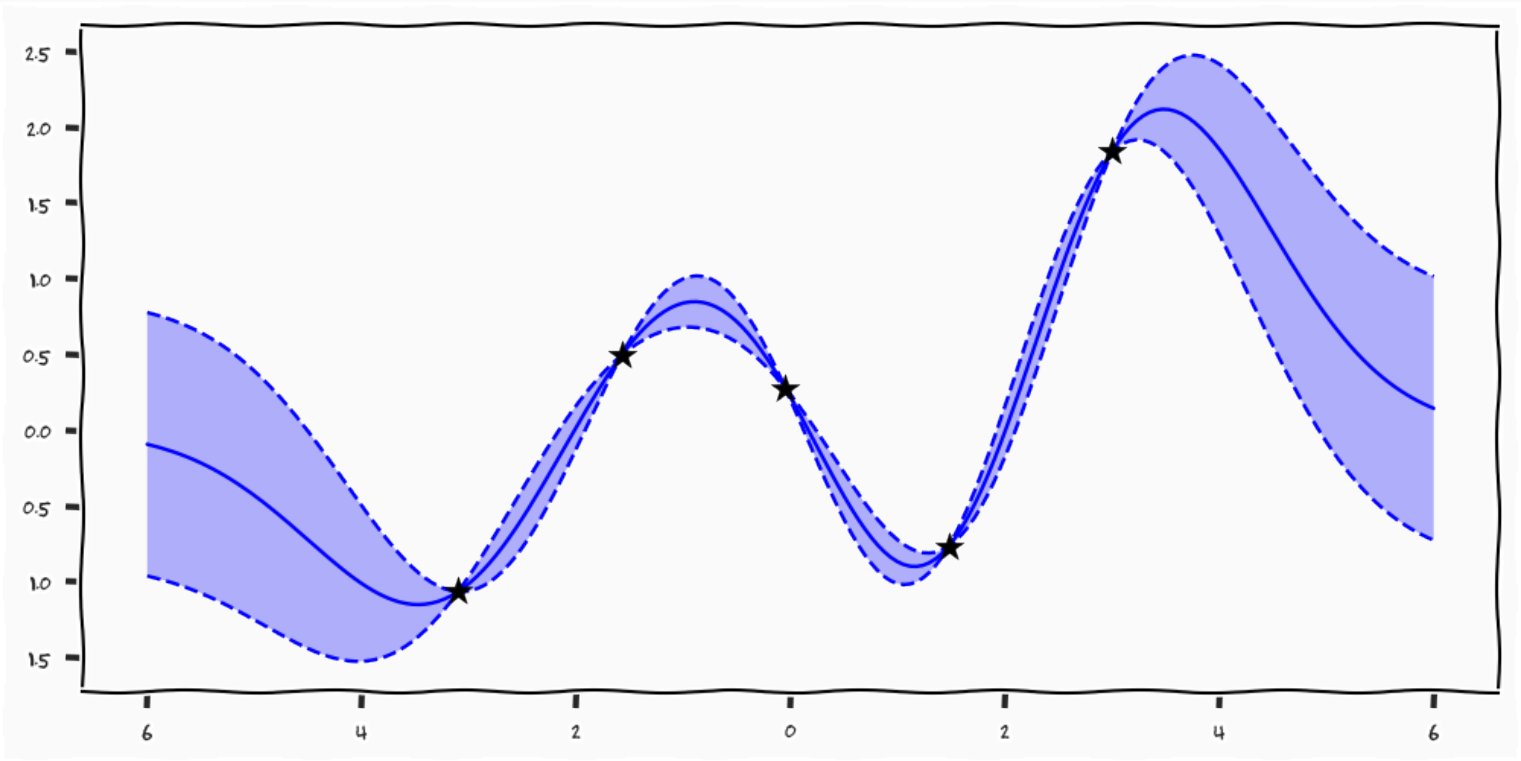
- Conditional Gaussian (same as always)

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}) = \mathcal{N}(k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} k(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}, \\ k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} k(\mathbf{X}, \mathbf{X})^{-1} k(\mathbf{X}, \mathbf{x}_*)$$

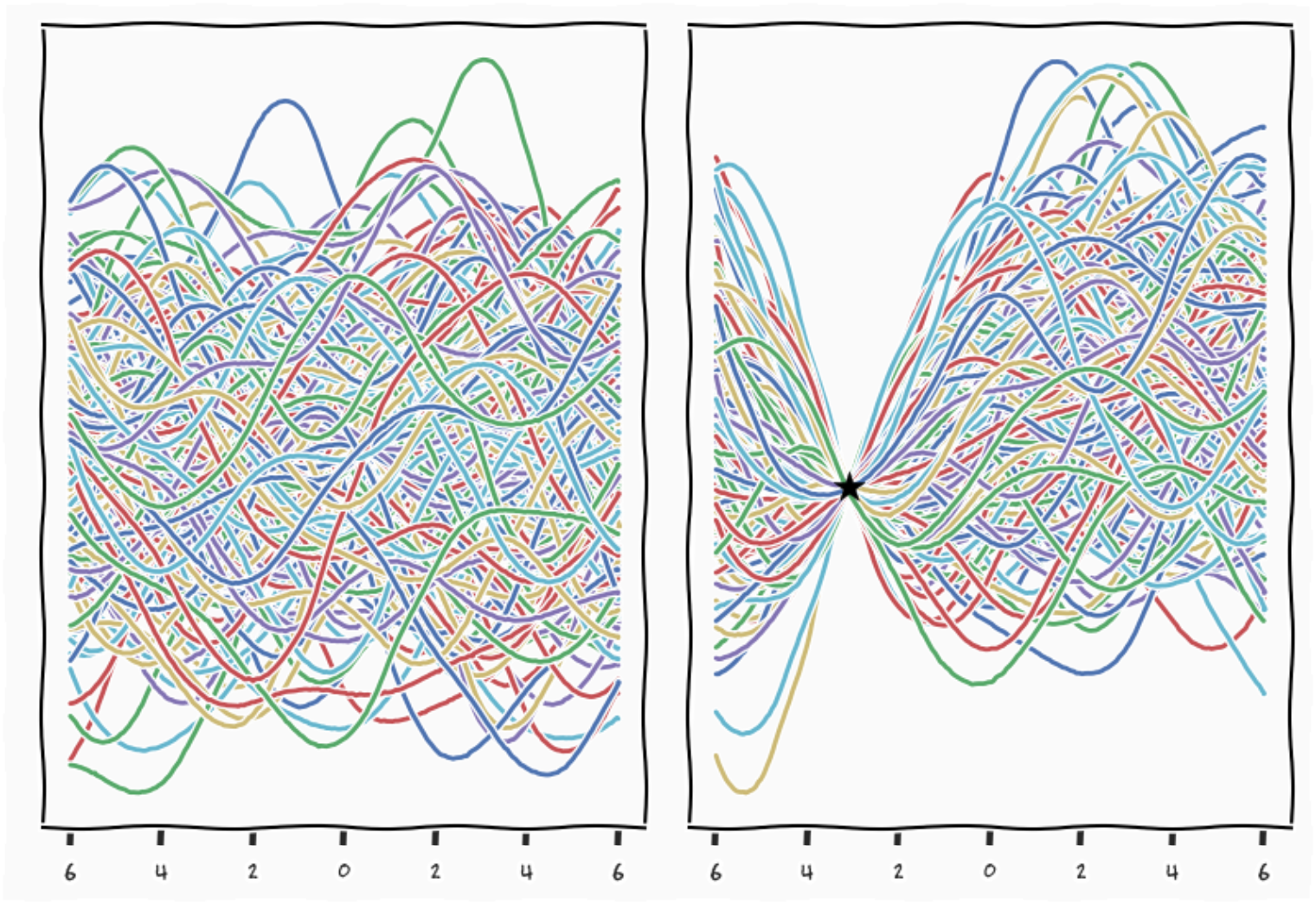
Gaussian Processes Posterior



Gaussian Processes Posterior



Gaussian Processes: Posterior Samples



Gaussian Processes: Noisy observations

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I} & k(\mathbf{X}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{X}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right)$$

$$p(f_*|\mathbf{x}_*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}}(K(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I}))^{-1}\mathbf{y}, \\ k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}}(K(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1} K(\mathbf{x}, \mathbf{x}_*))$$

- Add noise to observations

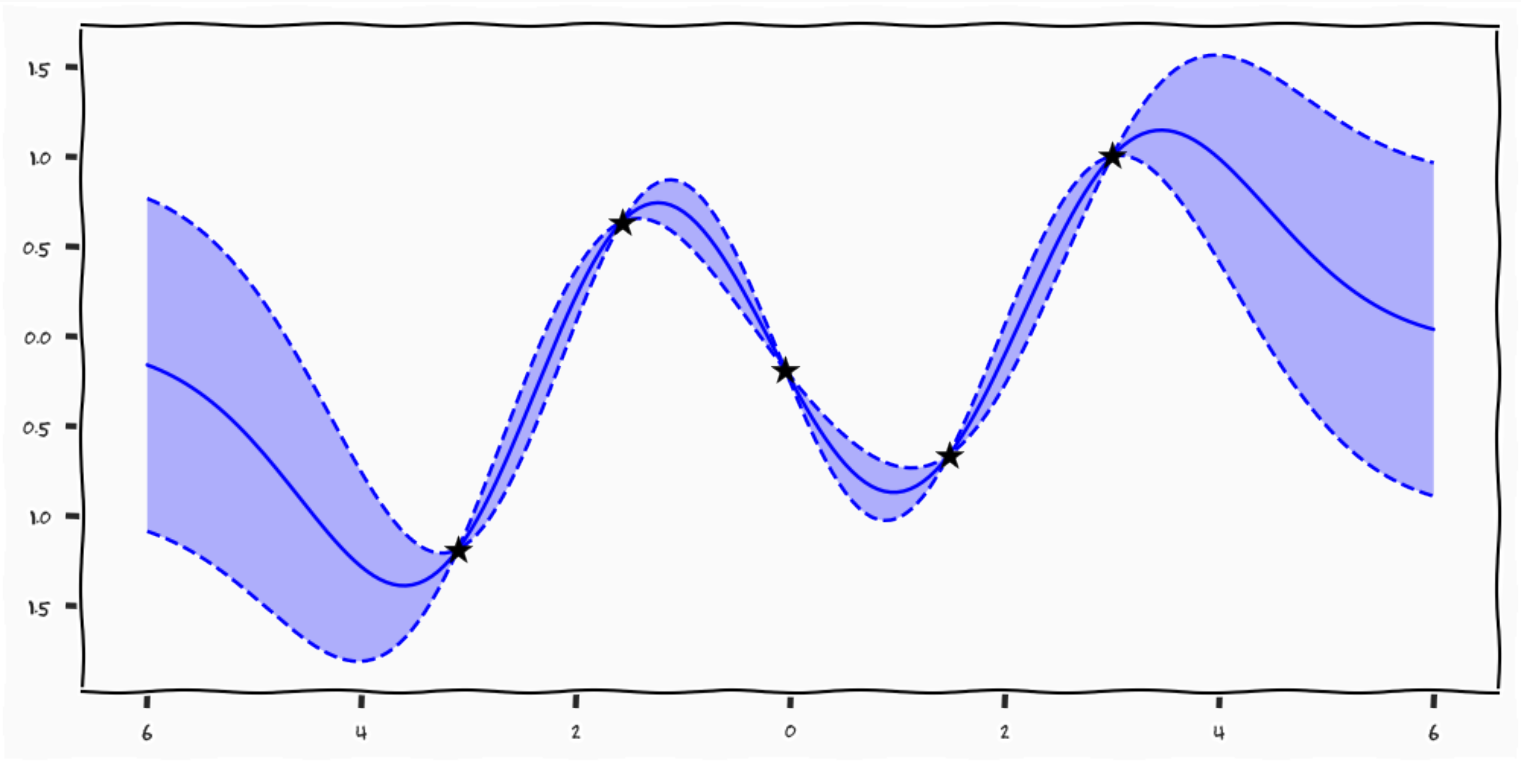
Gaussian Processes: Noisy observations

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I} & k(\mathbf{X}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{X}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right)$$

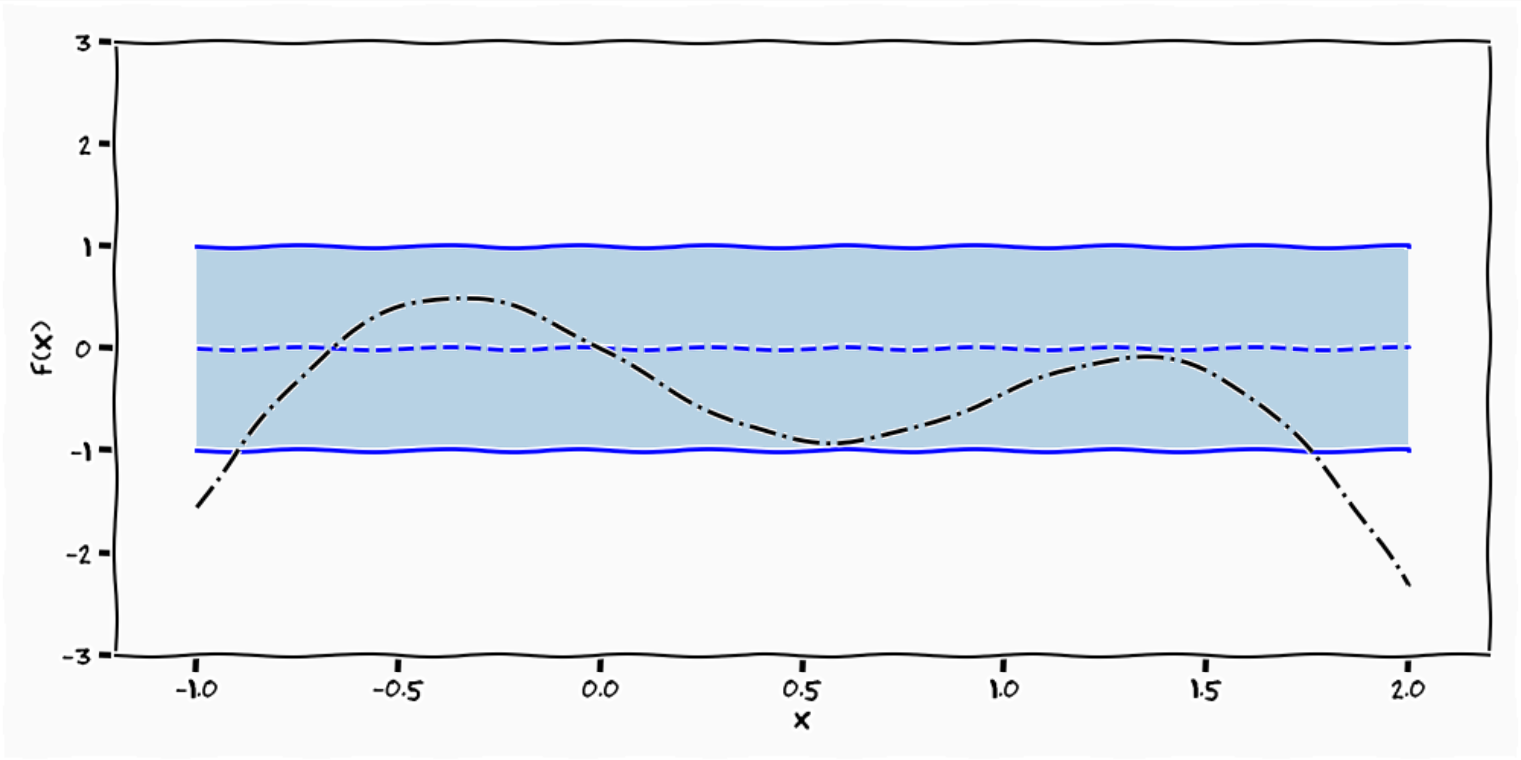
$$p(f_*|\mathbf{x}_*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}}(K(\mathbf{x}, \mathbf{x}) + \sigma^2\mathbf{I}))^{-1}\mathbf{y}, \\ k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}}(K(\mathbf{x}, \mathbf{x}) + \sigma^2\mathbf{I})^{-1}K(\mathbf{x}, \mathbf{x}_*))$$

- Add noise to observations
- *Do you recognise the mean?*

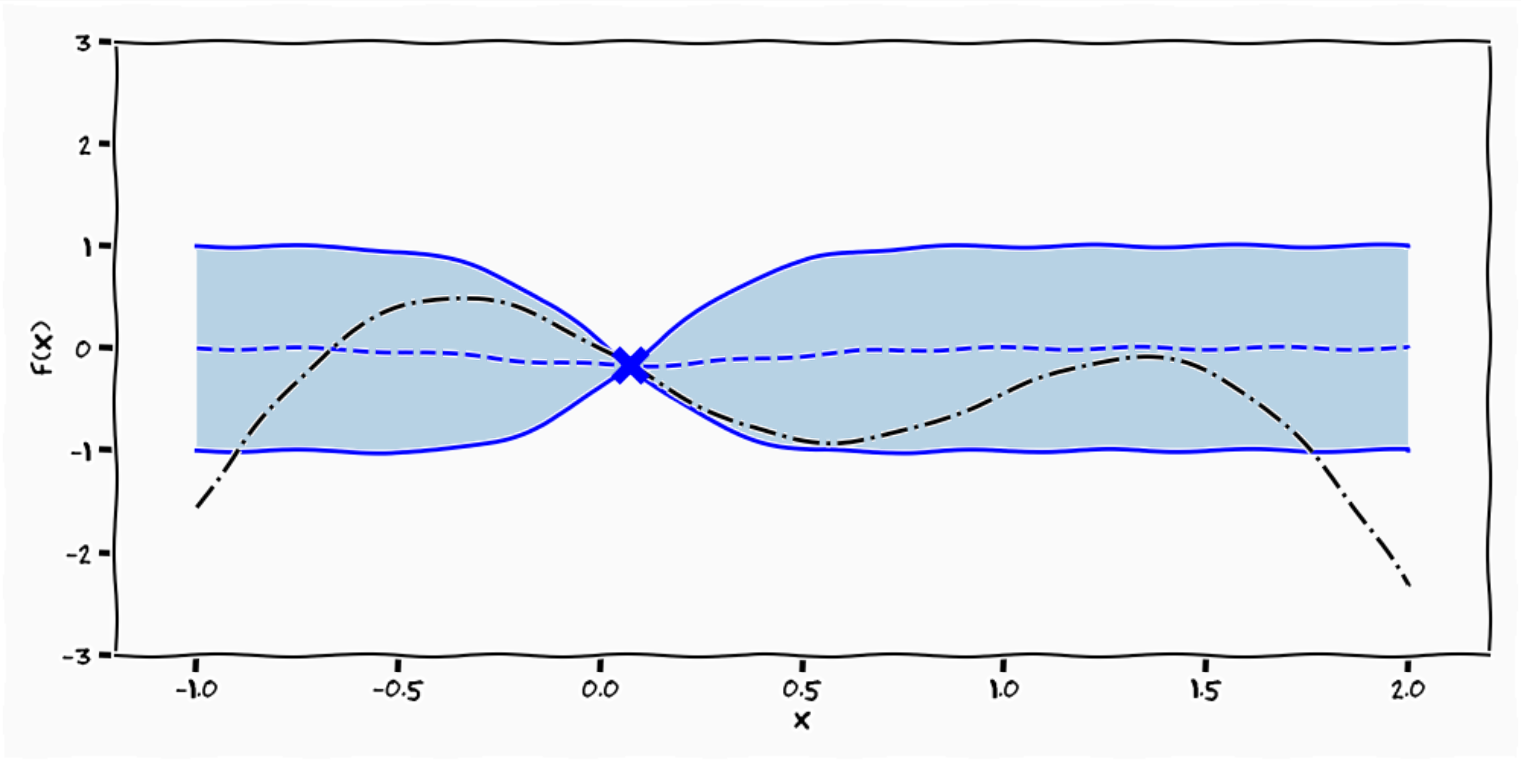
Gaussian Processes



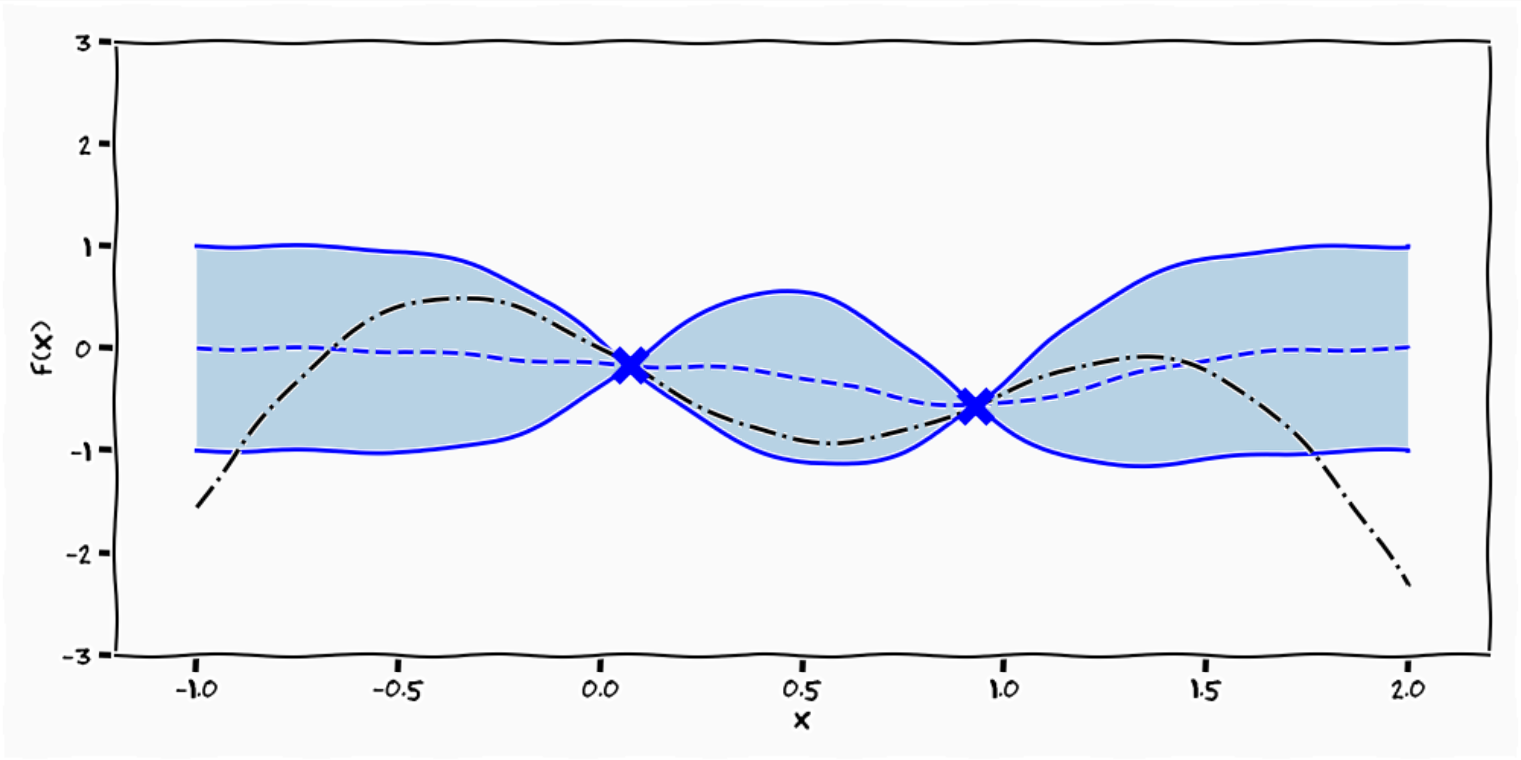
Posterior Processes



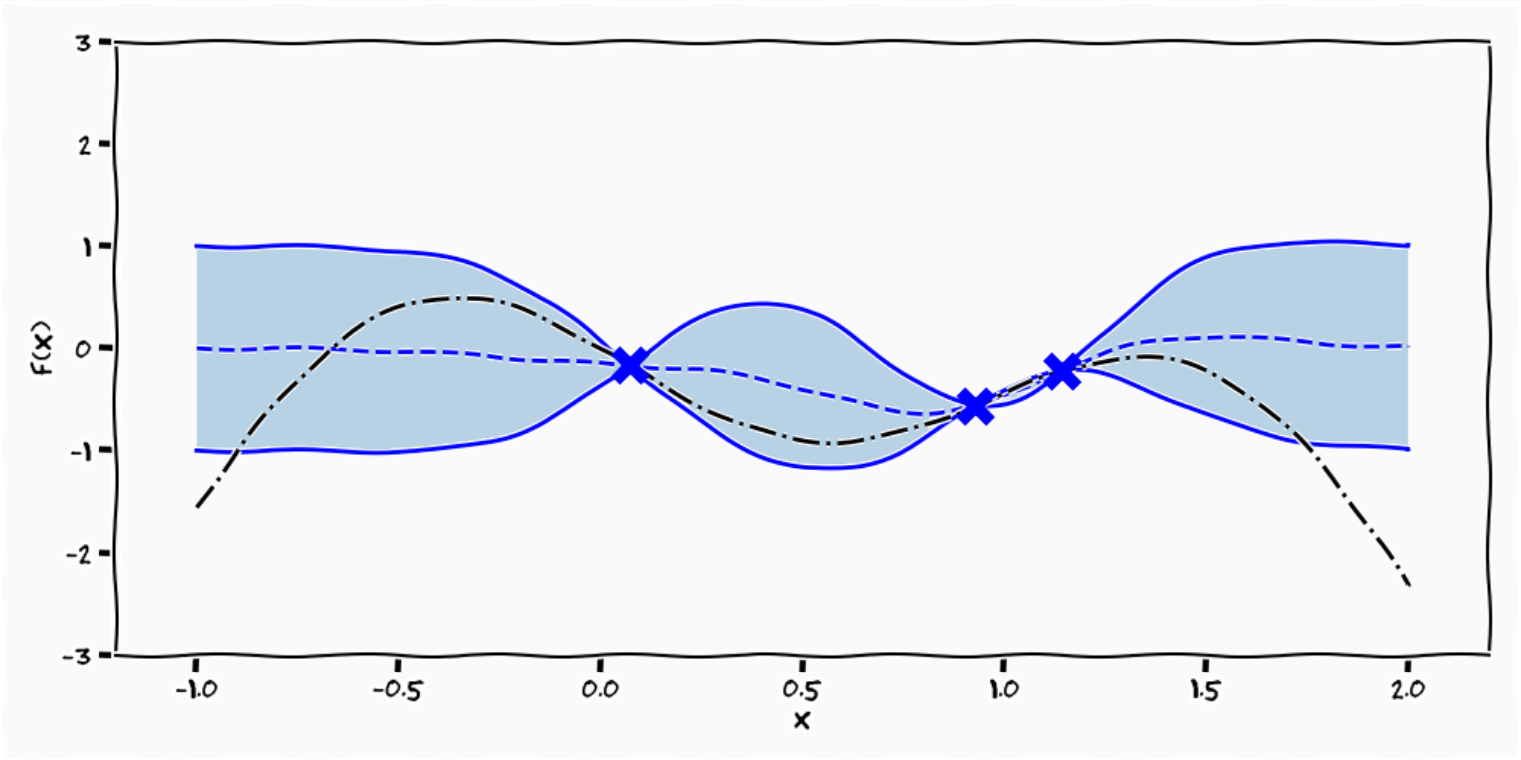
Posterior Processes



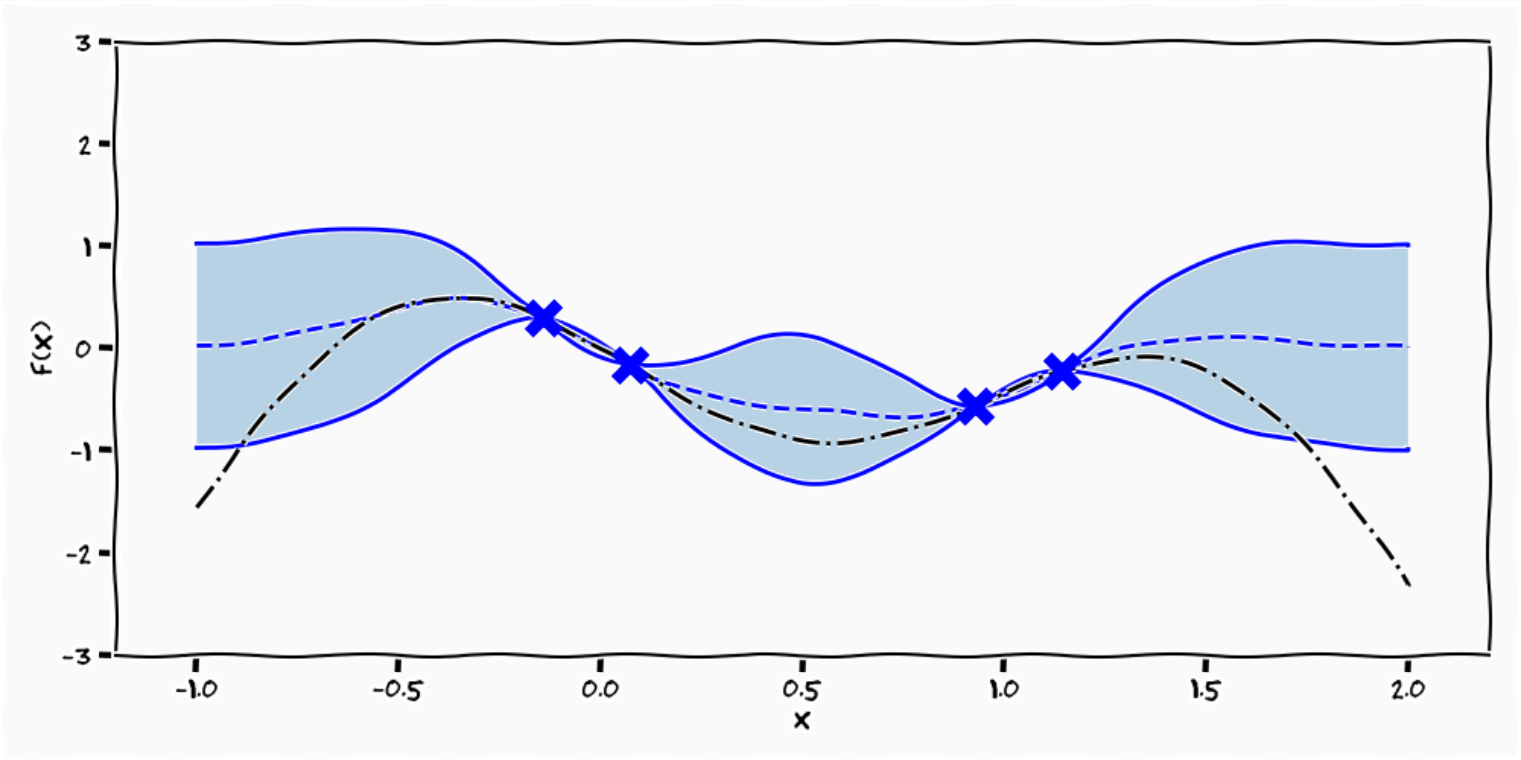
Posterior Processes



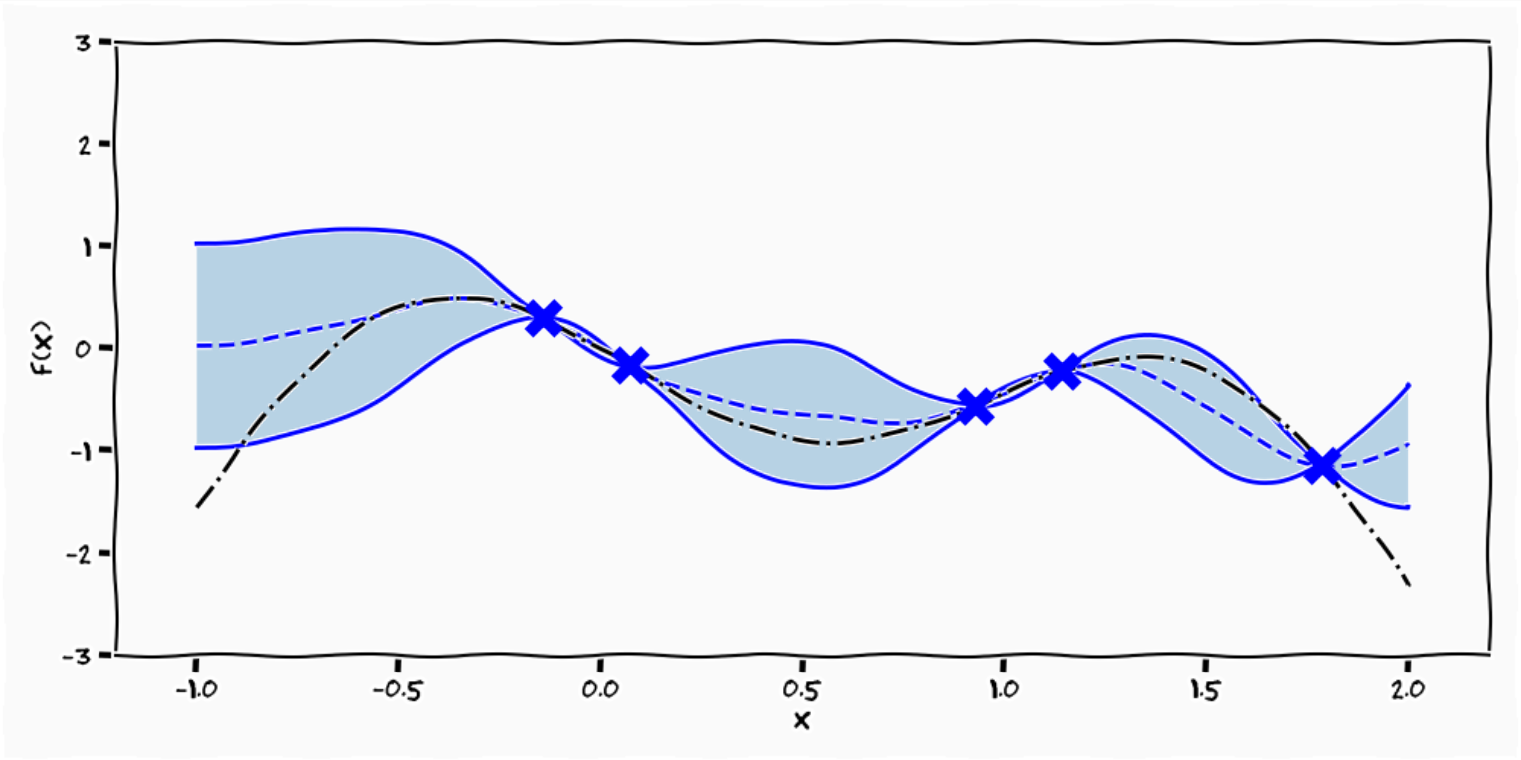
Posterior Processes



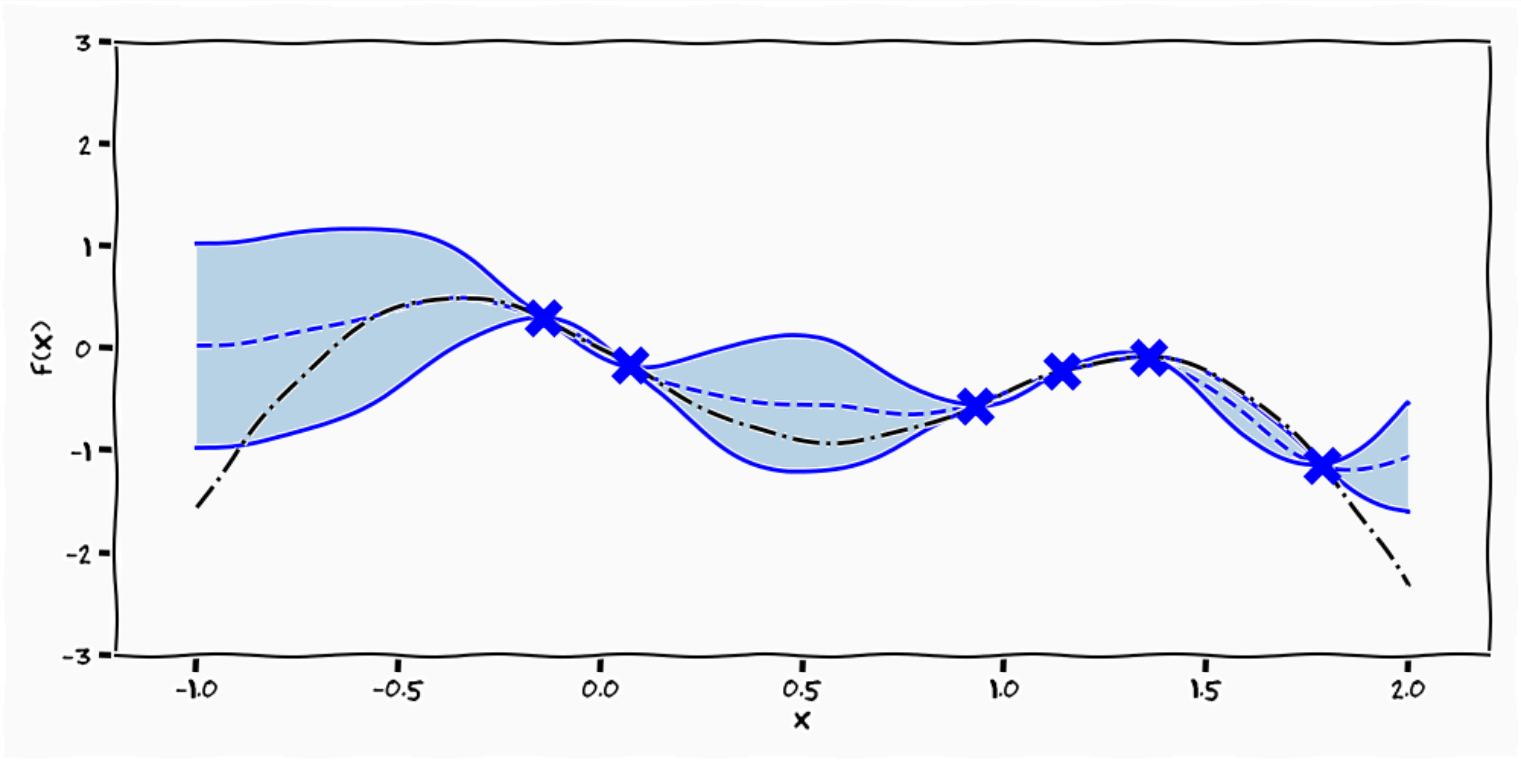
Posterior Processes



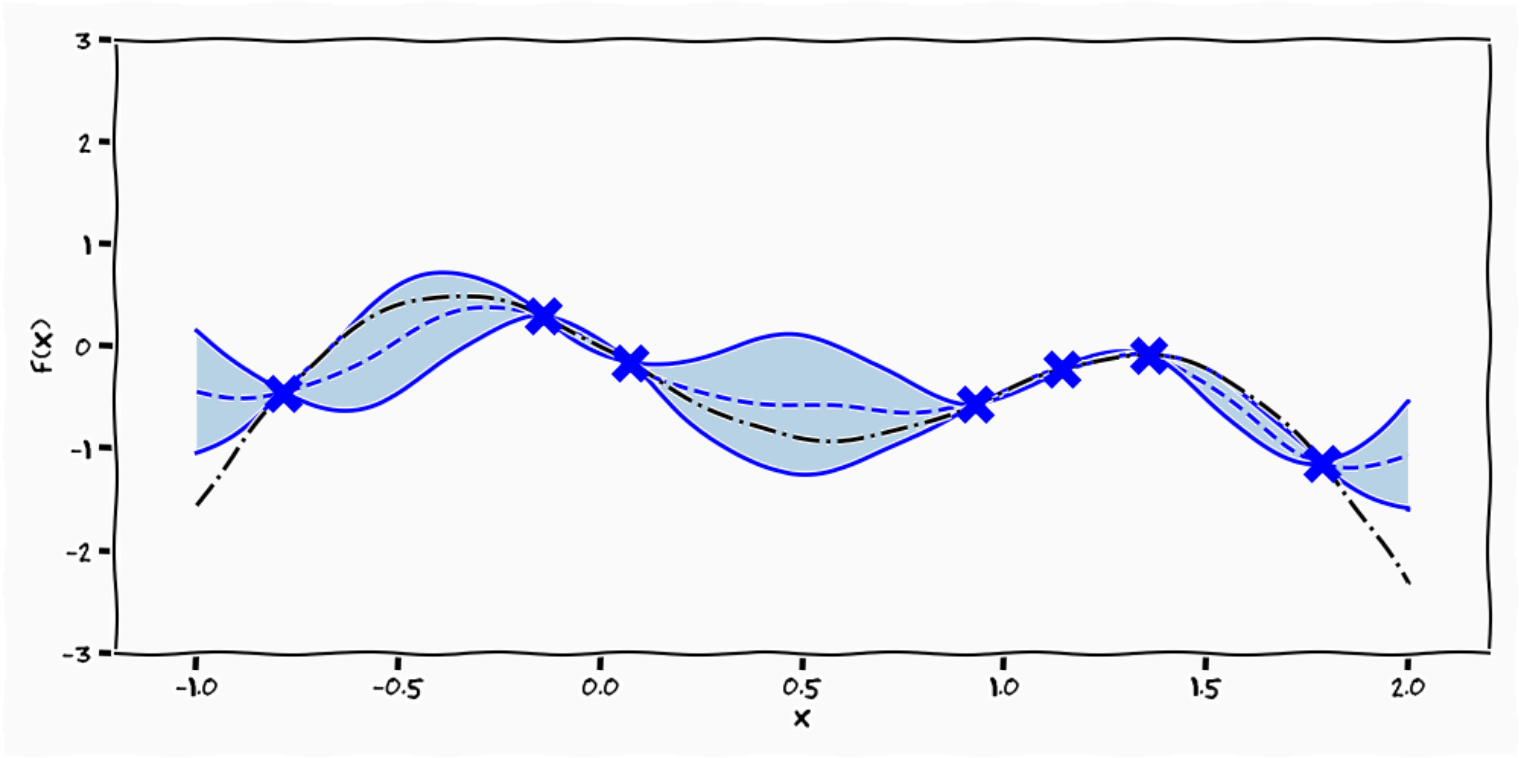
Posterior Processes



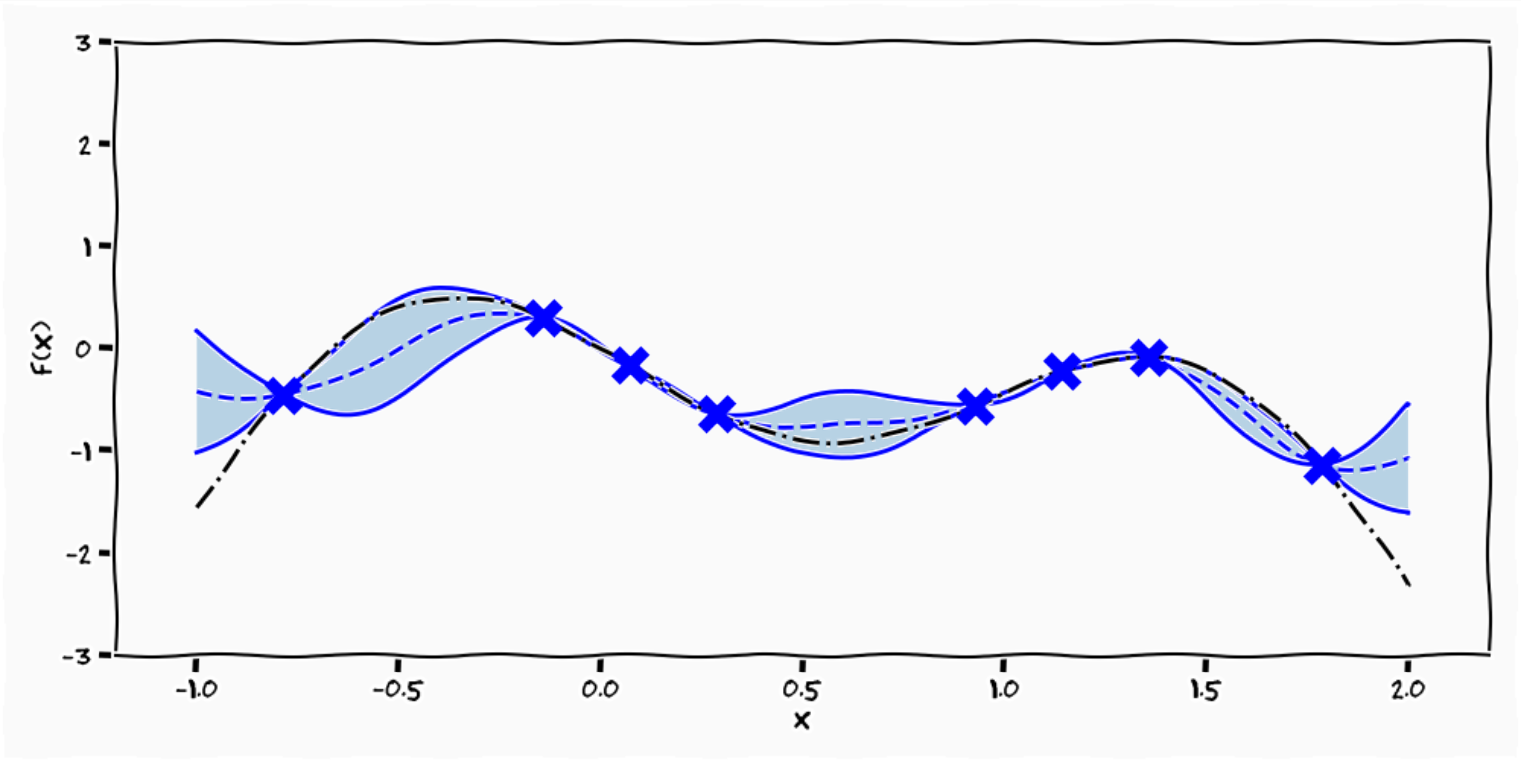
Posterior Processes



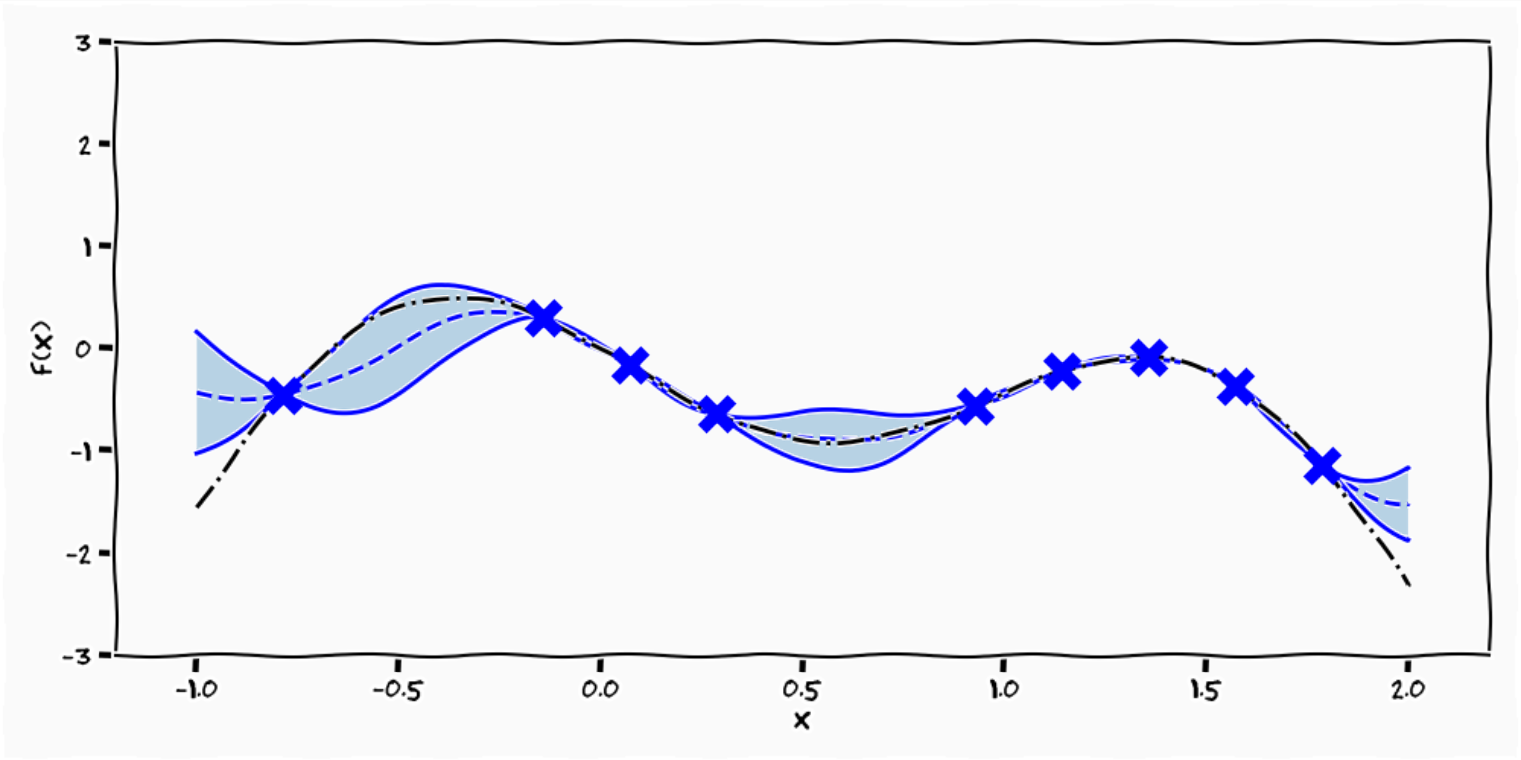
Posterior Processes



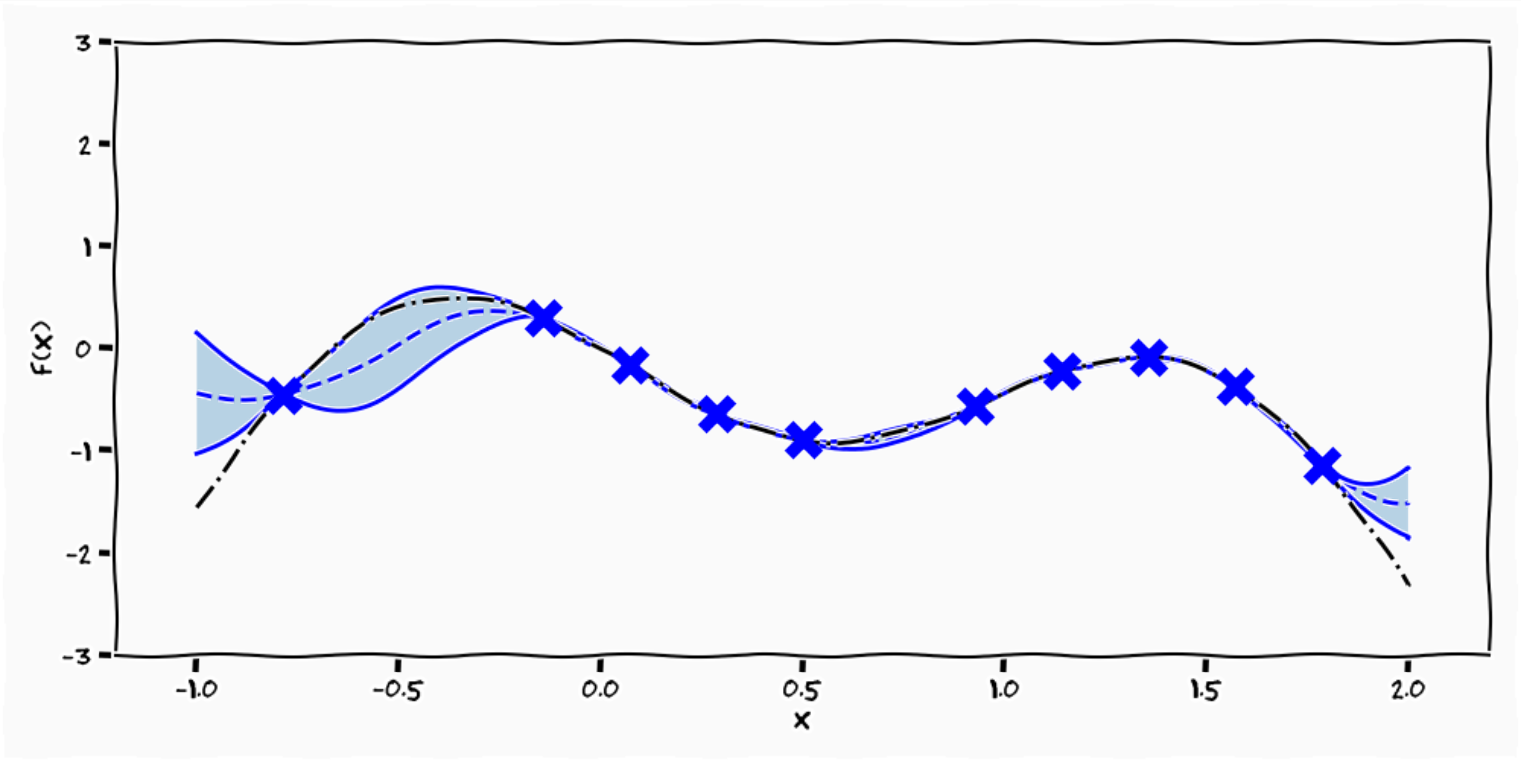
Posterior Processes



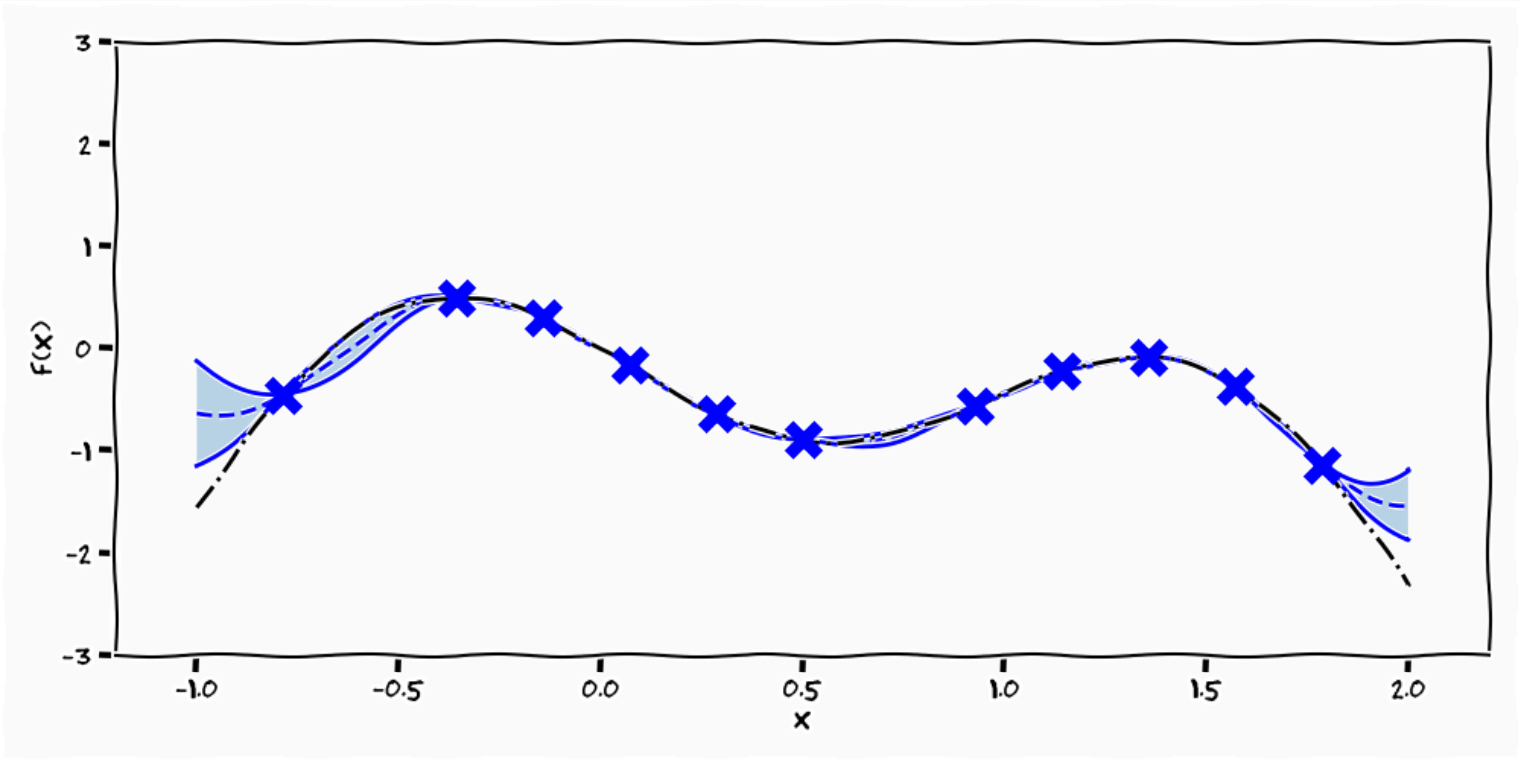
Posterior Processes



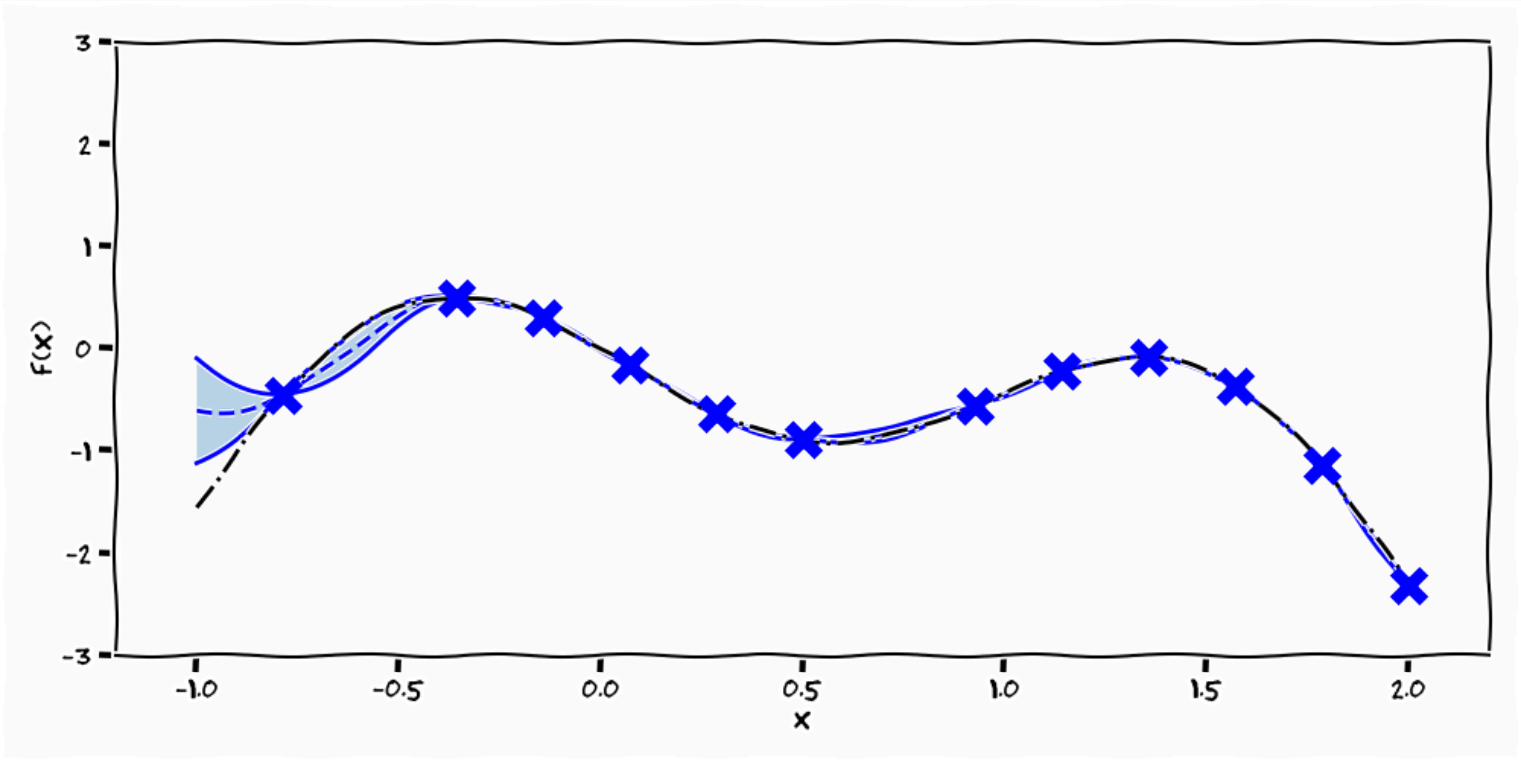
Posterior Processes



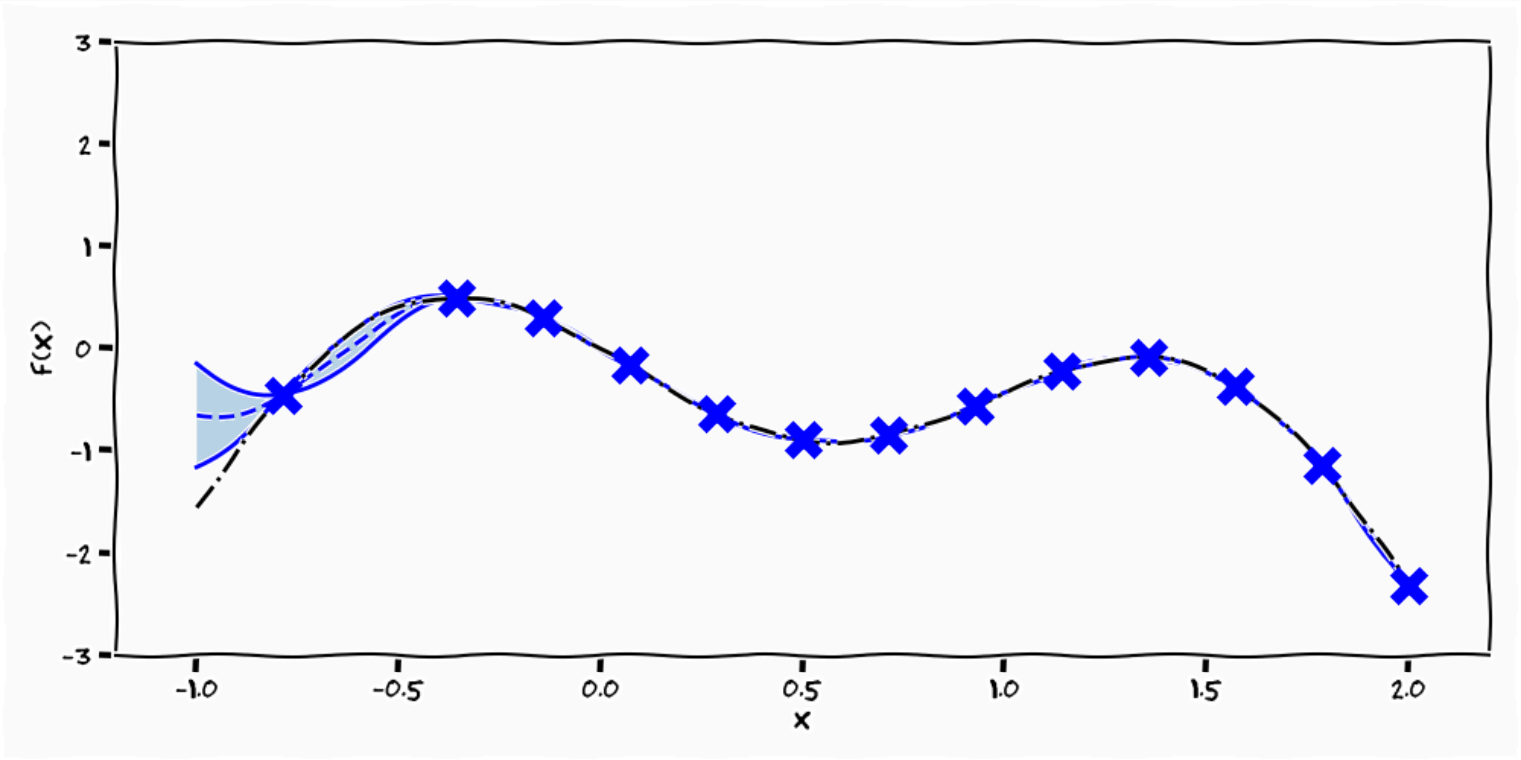
Posterior Processes



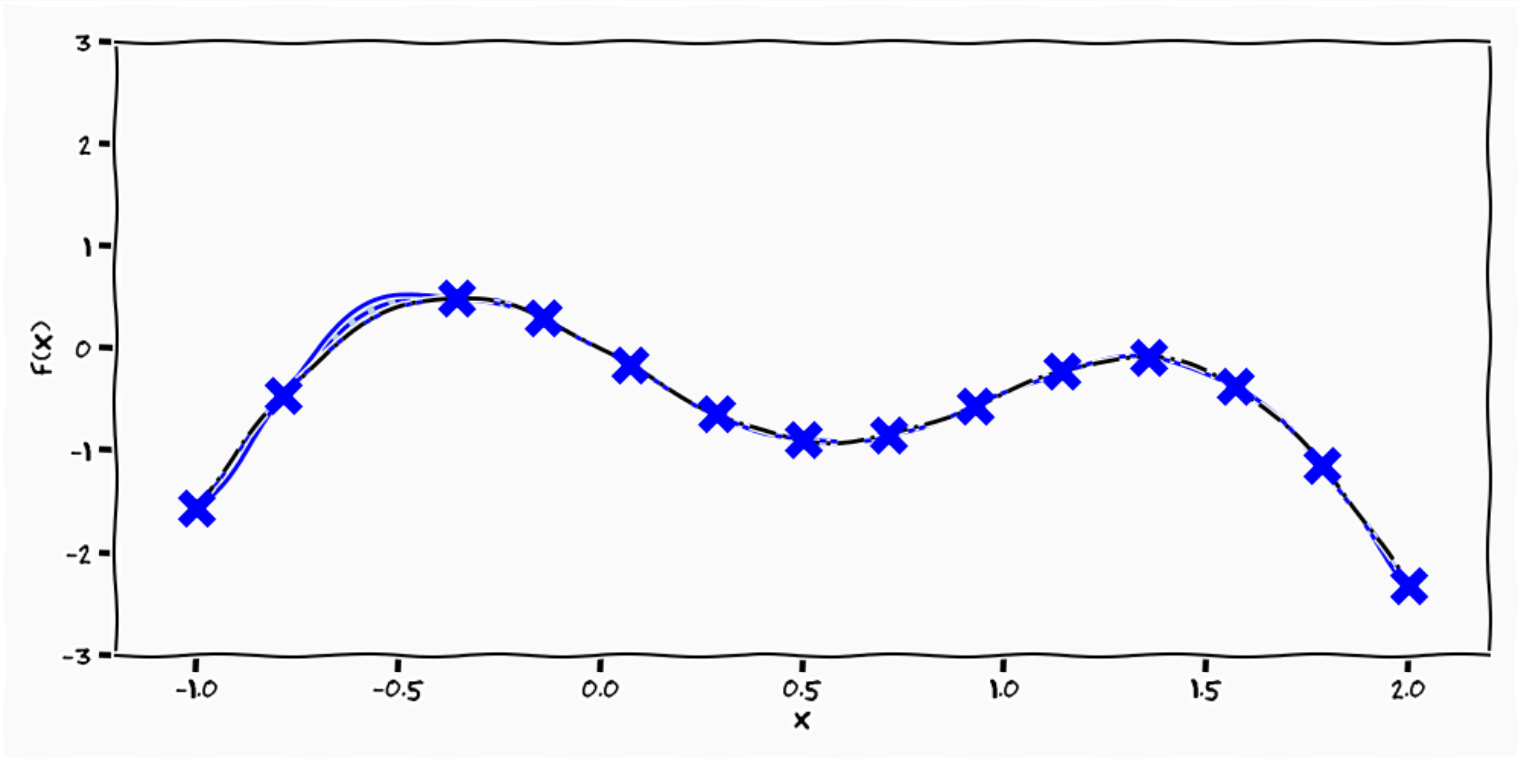
Posterior Processes



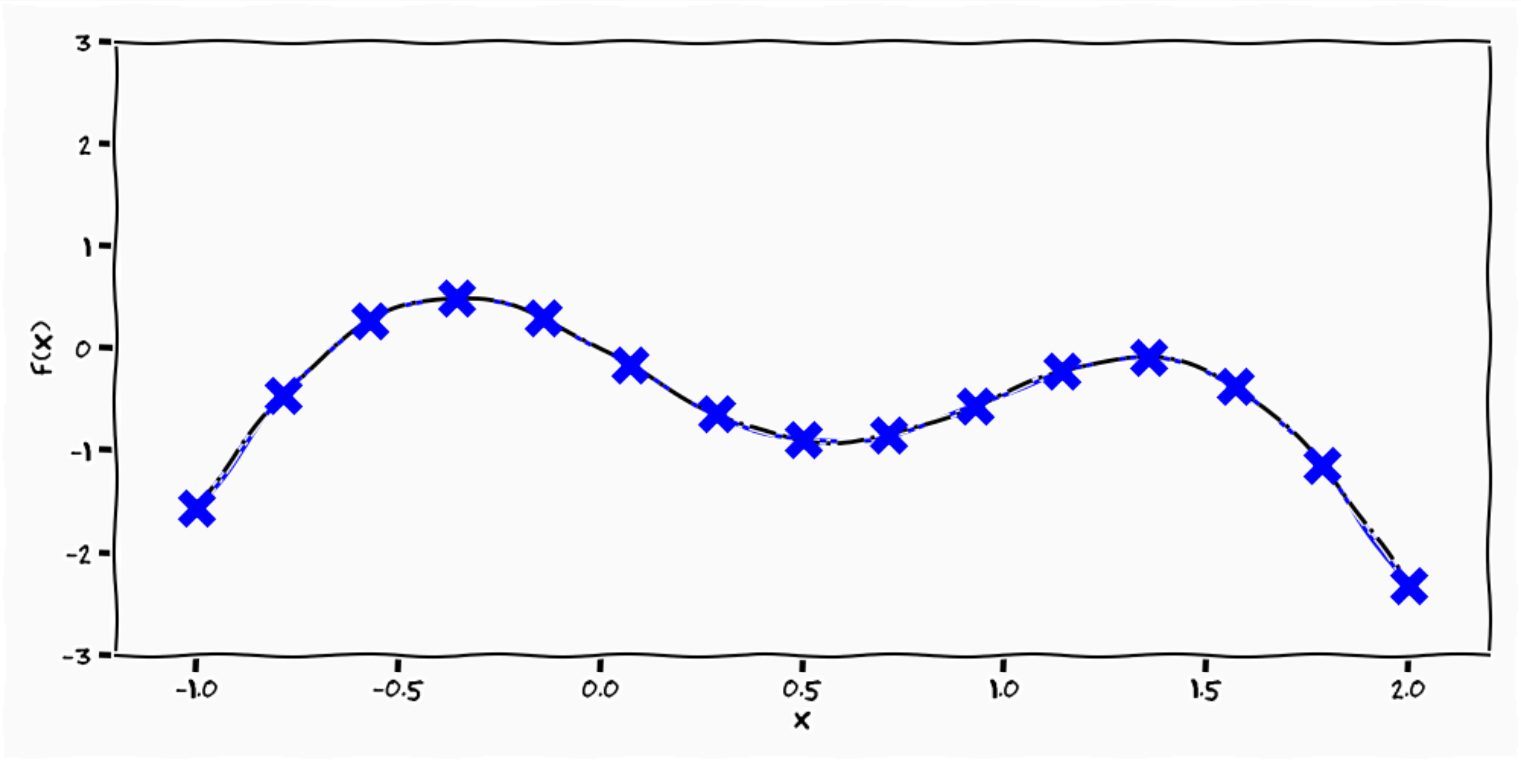
Posterior Processes



Posterior Processes



Posterior Processes



Summary

Summary

- Graphical models shows a factorisation of a probability distribution
- Marginal Likelihood can be done to select explanations
- Occam's razor implies a notion of simple defined by your prior

References

References


Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.


Mackay, David J C (Dec. 1991). "Bayesian methods for adaptive models ". PhD thesis. California Institute of Technology: California Institute of Technology.


Spearman, Charles (1904). """ General Intelligence," Objectively Determined and Measured". In: *The American Journal of Psychology* 15.2, pp. 201–292.