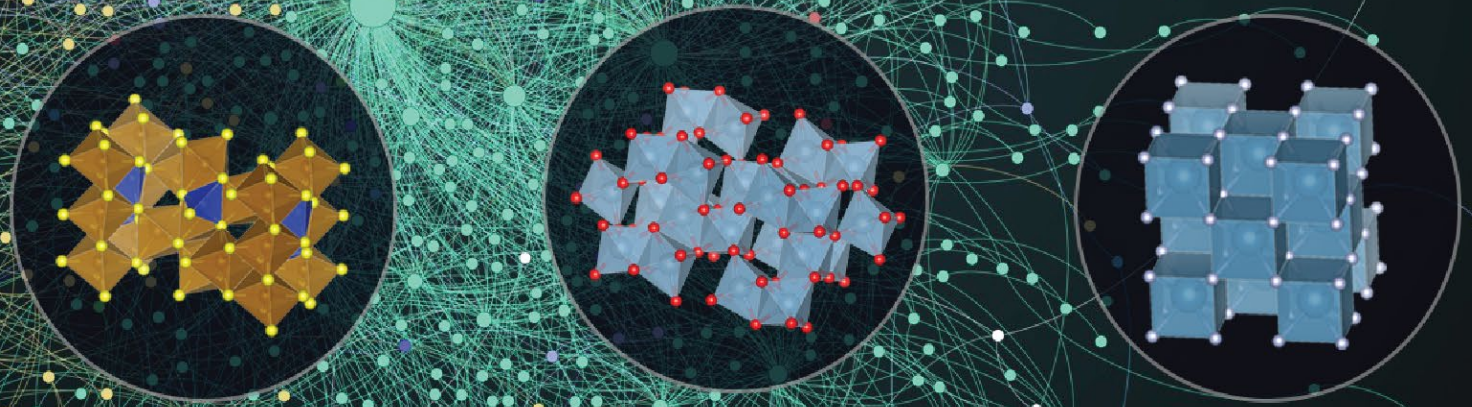
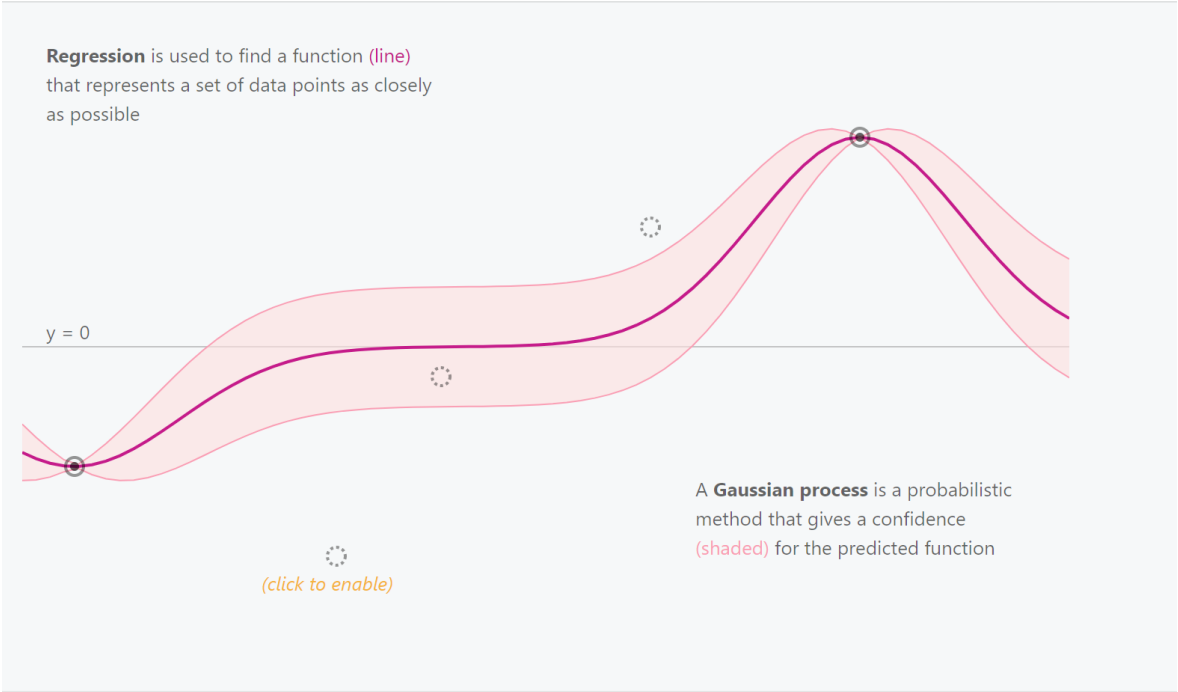
Gaussian processes



Gaussian Processes are a bit tricky to understand!

A Visual Exploration of Gaussian Processes

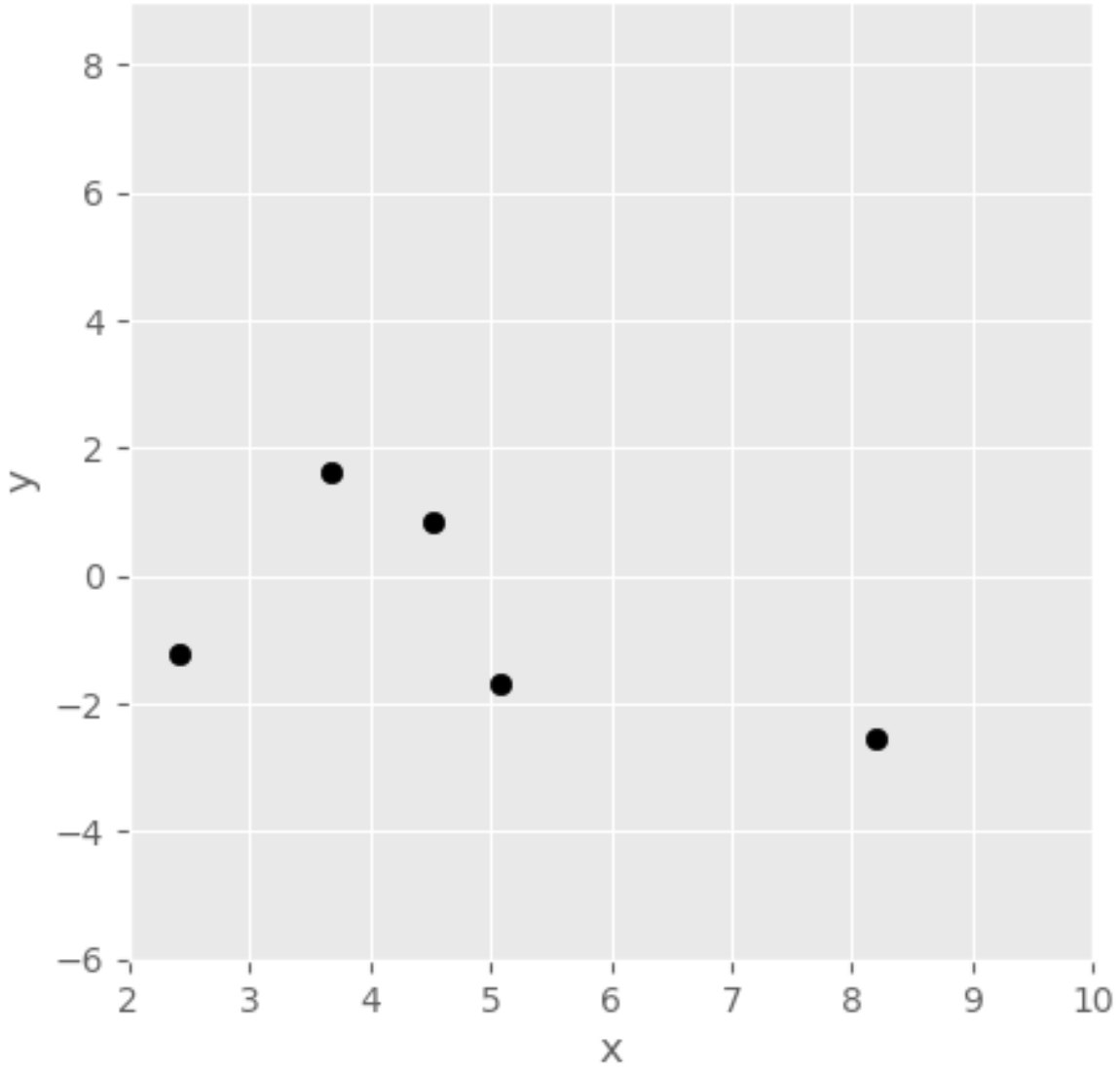
How to turn a collection of small building blocks into a versatile tool for solving regression problems.



AUTHORS	AFFILIATIONS	PUBLISHED	DOI
Jochen Görtler	University of Konstanz	April 2, 2019	10.23915/distill.00017
Rebecca Kehlbeck	University of Konstanz		
Oliver Deussen	University of Konstanz		

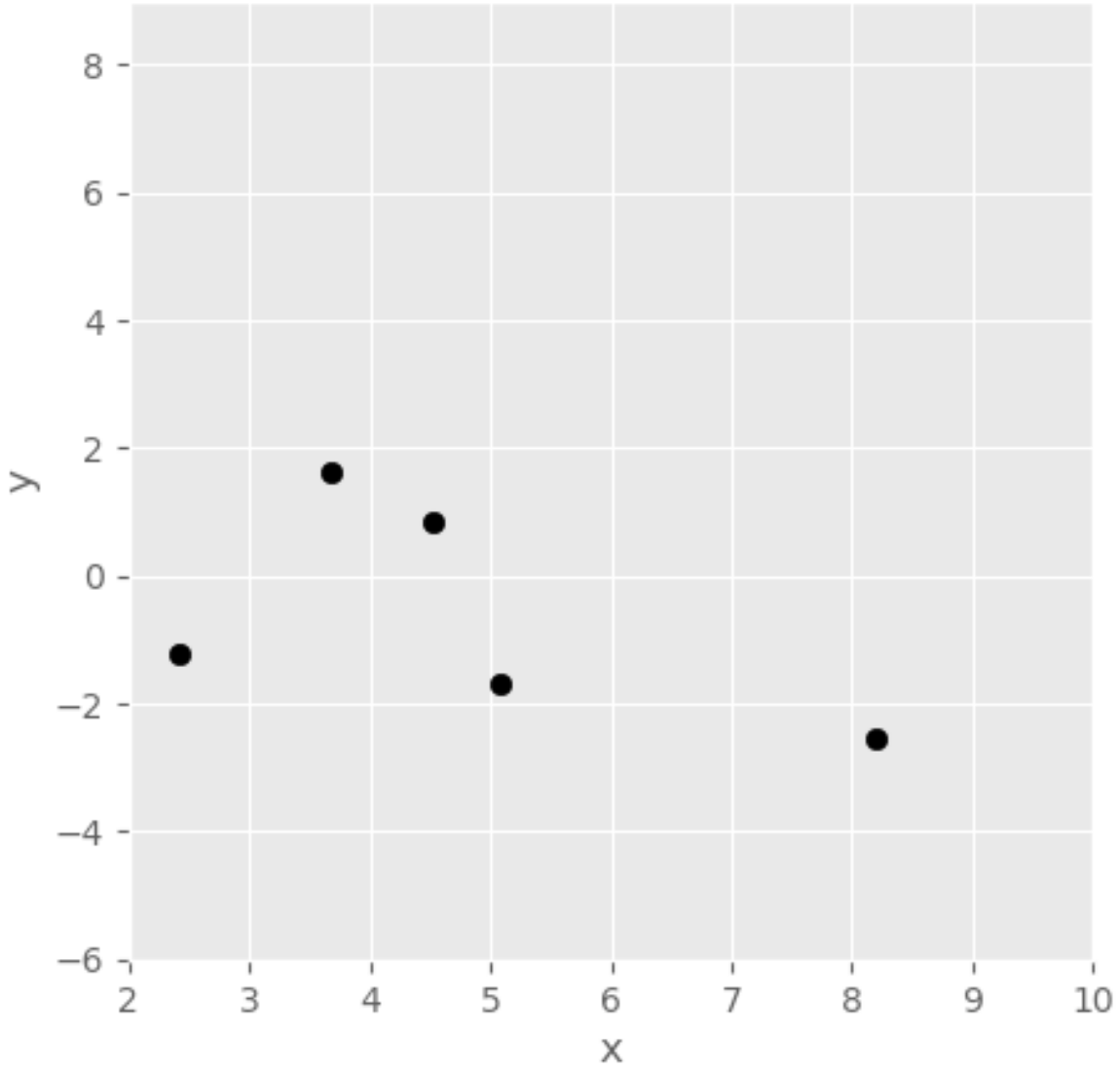
https://distill.pub/2019/visual-exploration-gaussian-processes/

Why should we care about this?

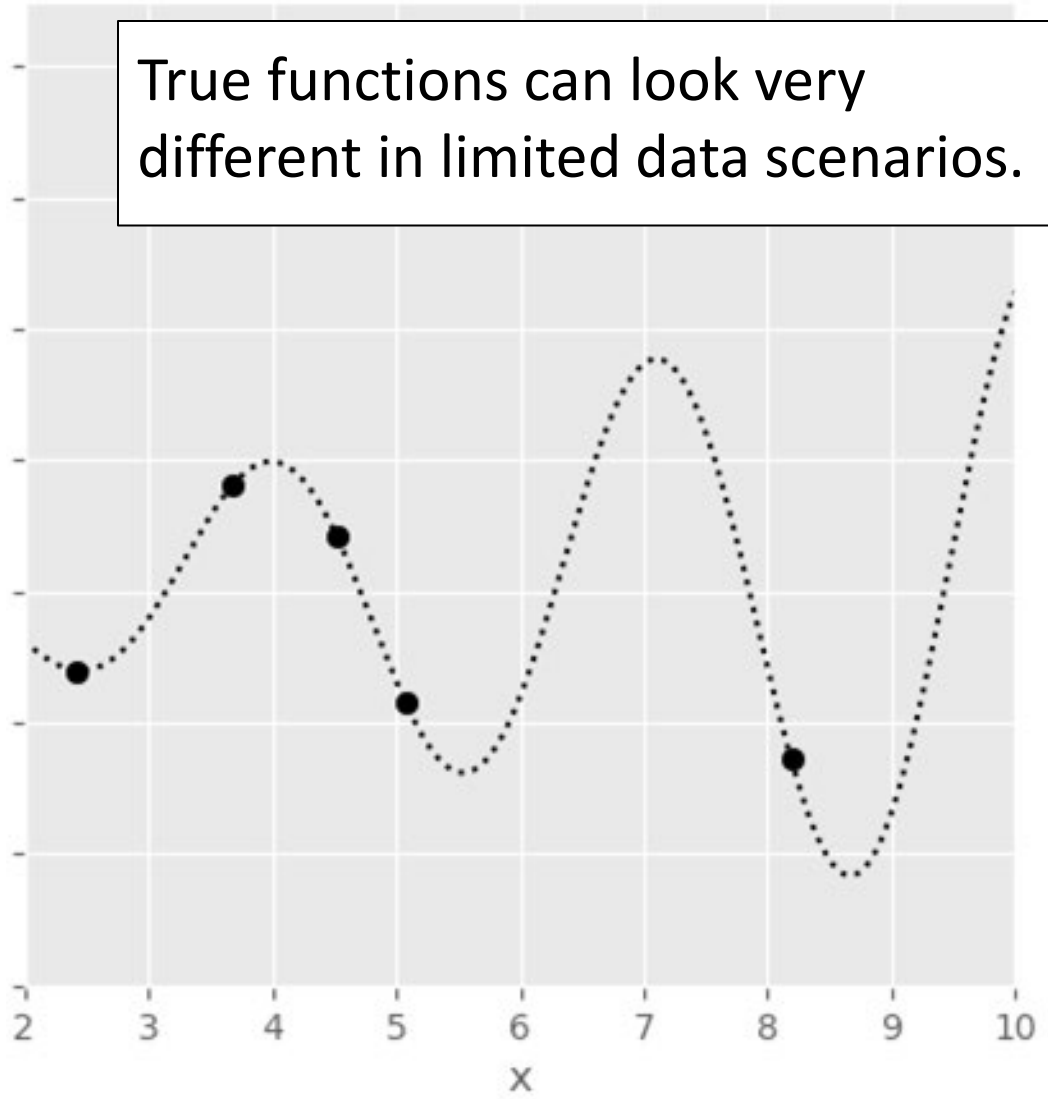


How would you fit this function?

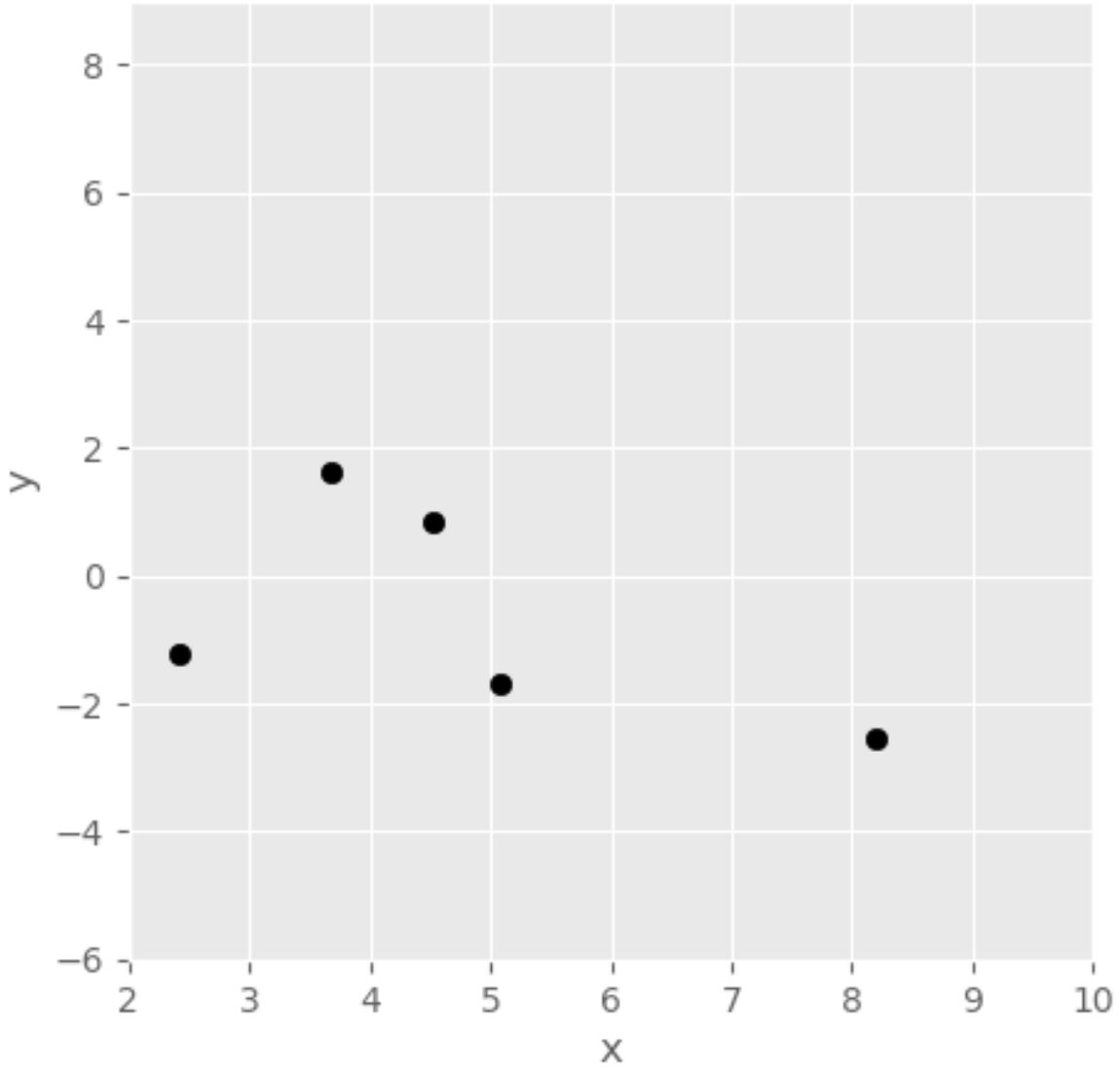
Why should we care about this?



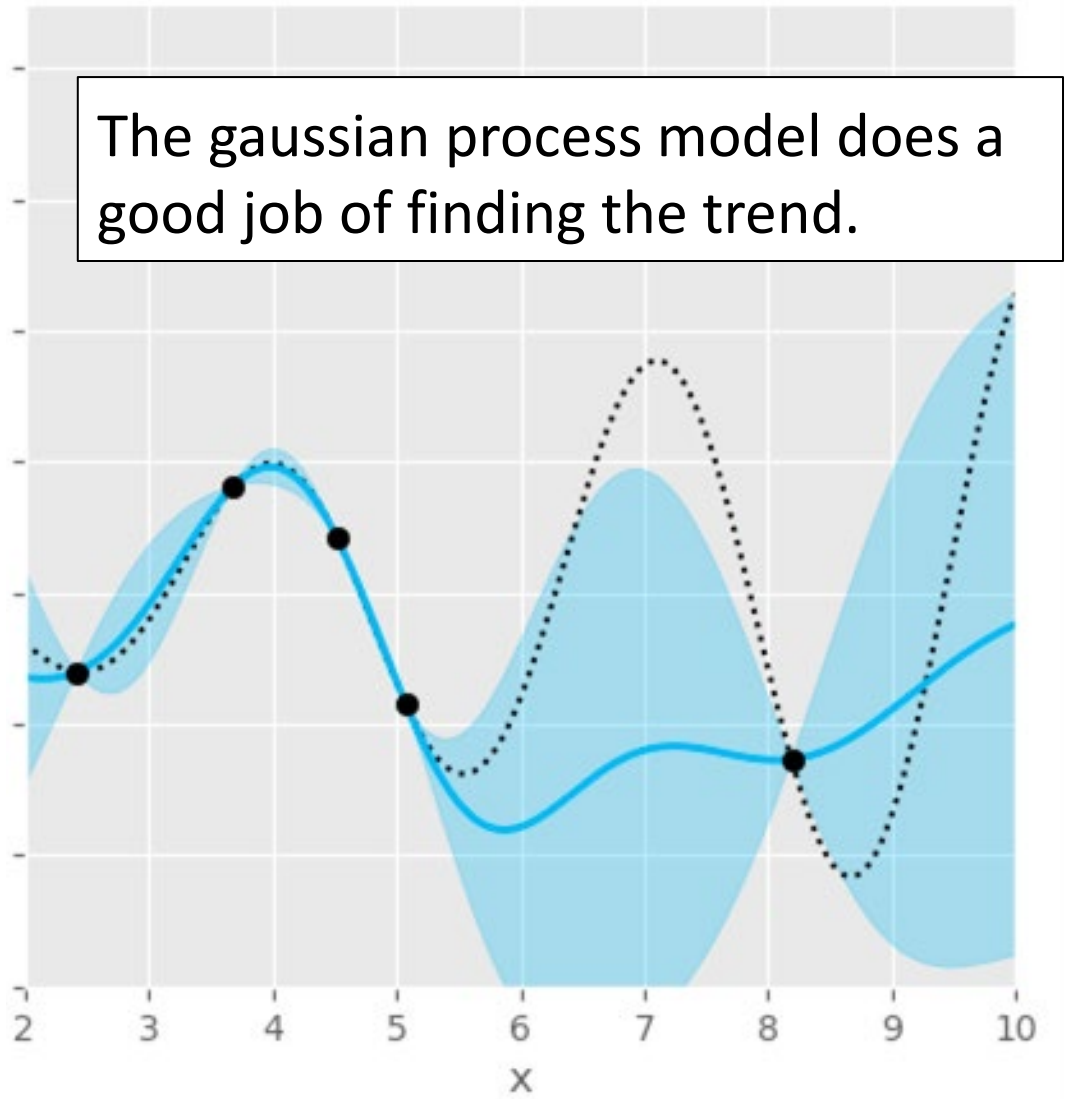
True functions can look very different in limited data scenarios.



Why should we care about this?



The gaussian process model does a good job of finding the trend.



Naïve Bayes had us multiply prior by likelihood of features

Naïve Bayes used prior*likelihoods to make predictions

$$p(y|X) = p(y) * p(y|X_1) * p(y|X_2) \ldots$$

The function could take infinite forms

$$C(y|X) = g(X)$$

Consider linear regression

$$C(y|X) = g(X, \beta) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots$$

Assumptions:

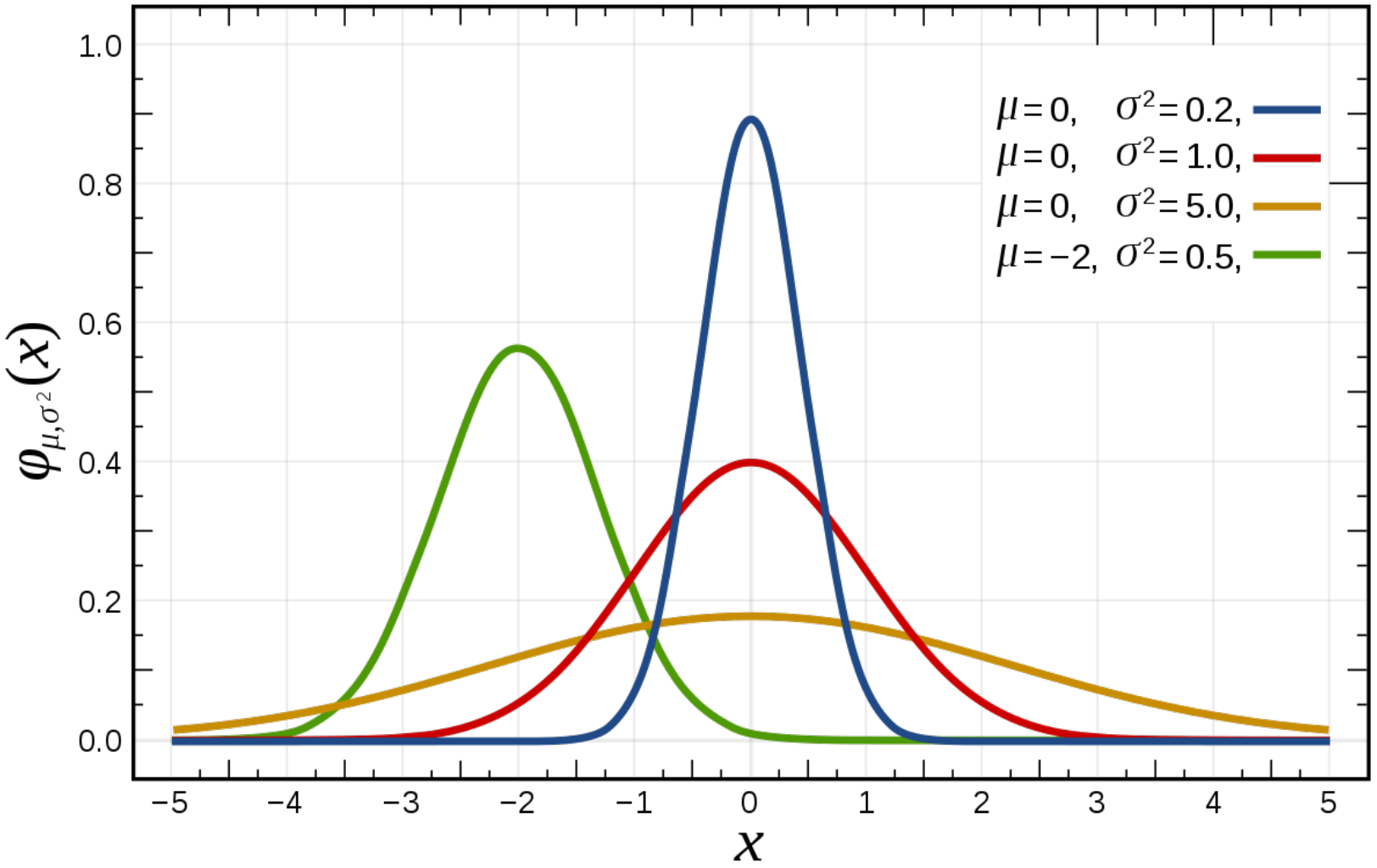
Parametric (data is normally distributed) and linear relationships between variables

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma)$$

Logistic regression also possible

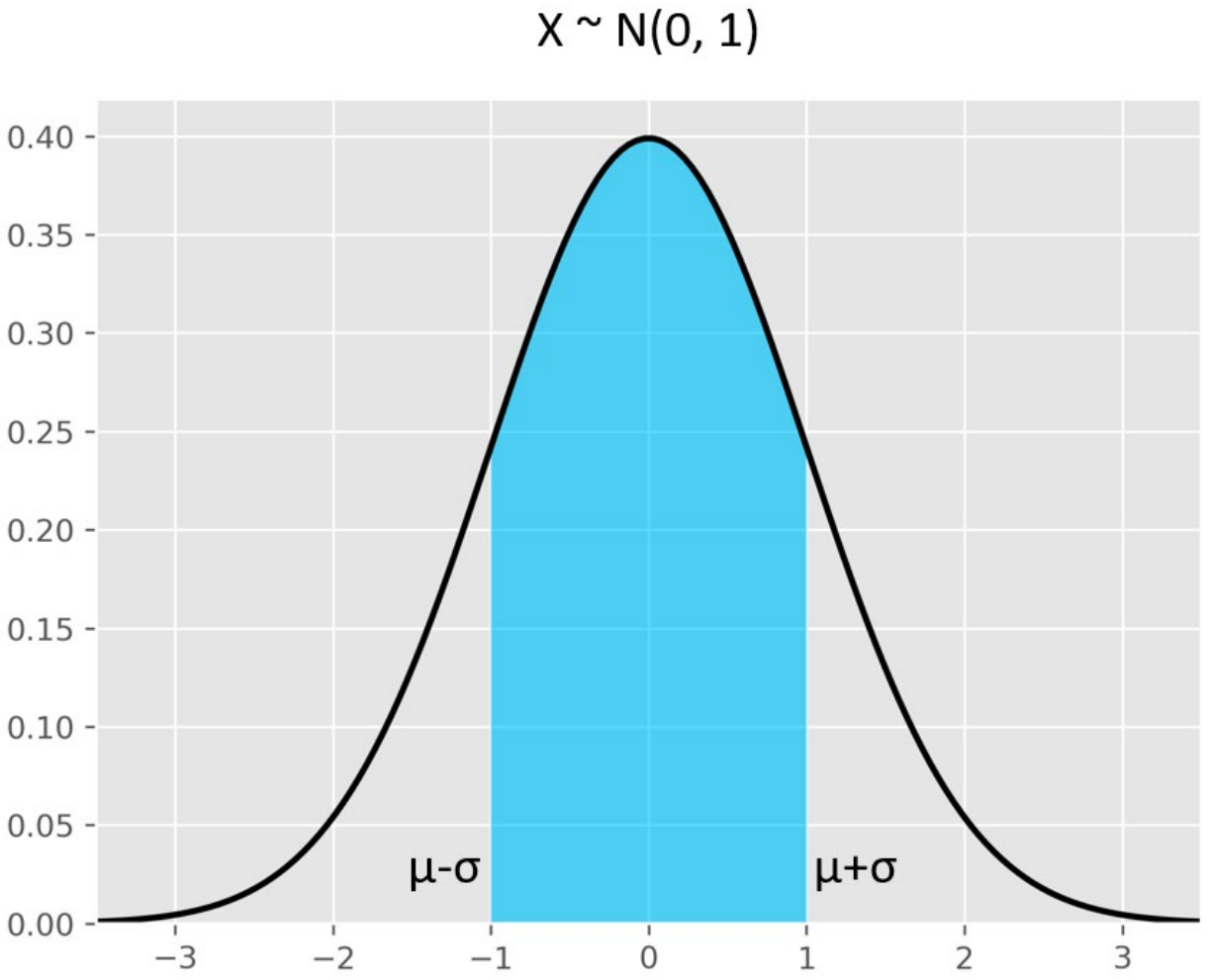
$$C(y|X) = \frac{1}{1 + \exp(\beta X)}$$

Instead of relying on linear or non-linear approaches to fit data, let's rely on gaussians!



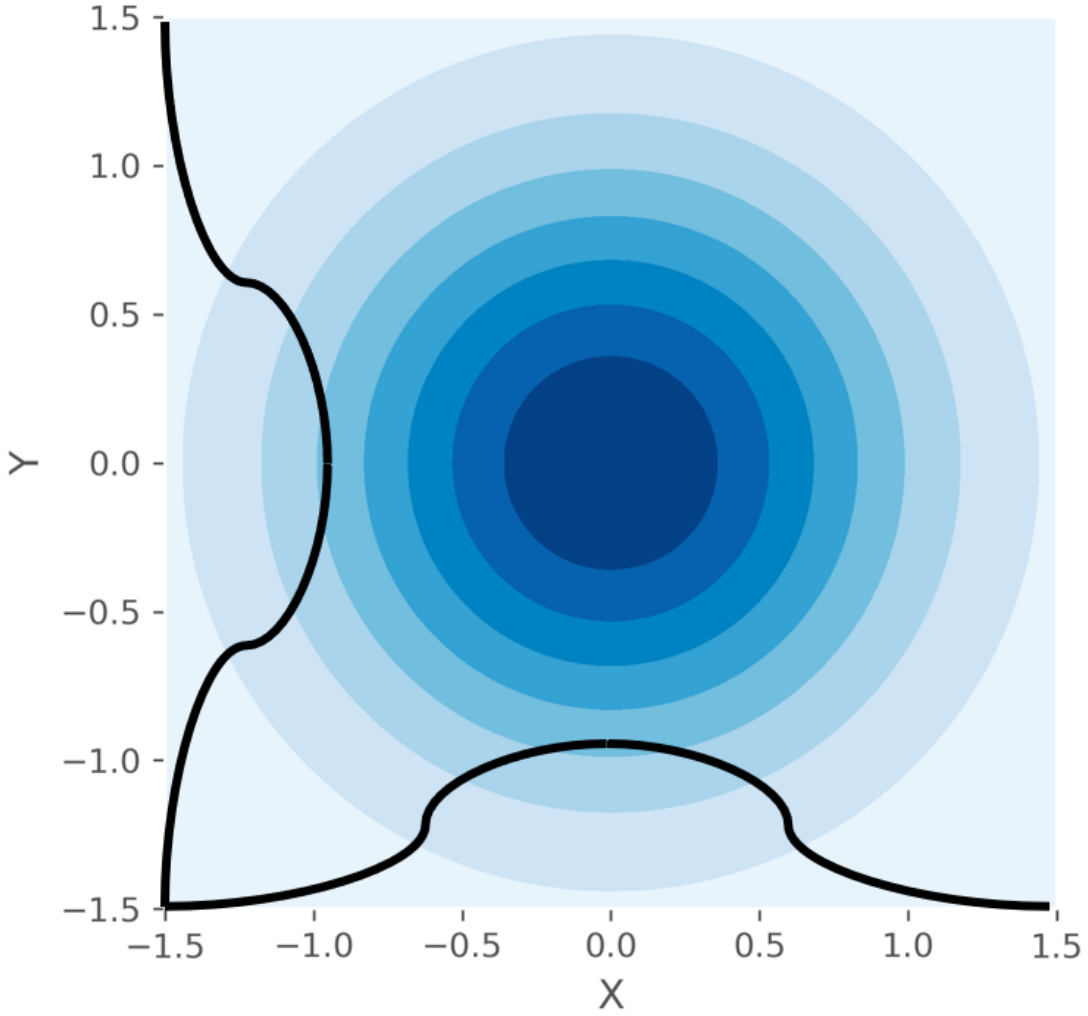
https://en.wikipedia.org/wiki/Normal_distribution

The Gaussian probability function is defined by a mean and a standard deviation



We can have **multivariate** Gaussians distributions

$$\mathcal{N}\left(\begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}\right)$$

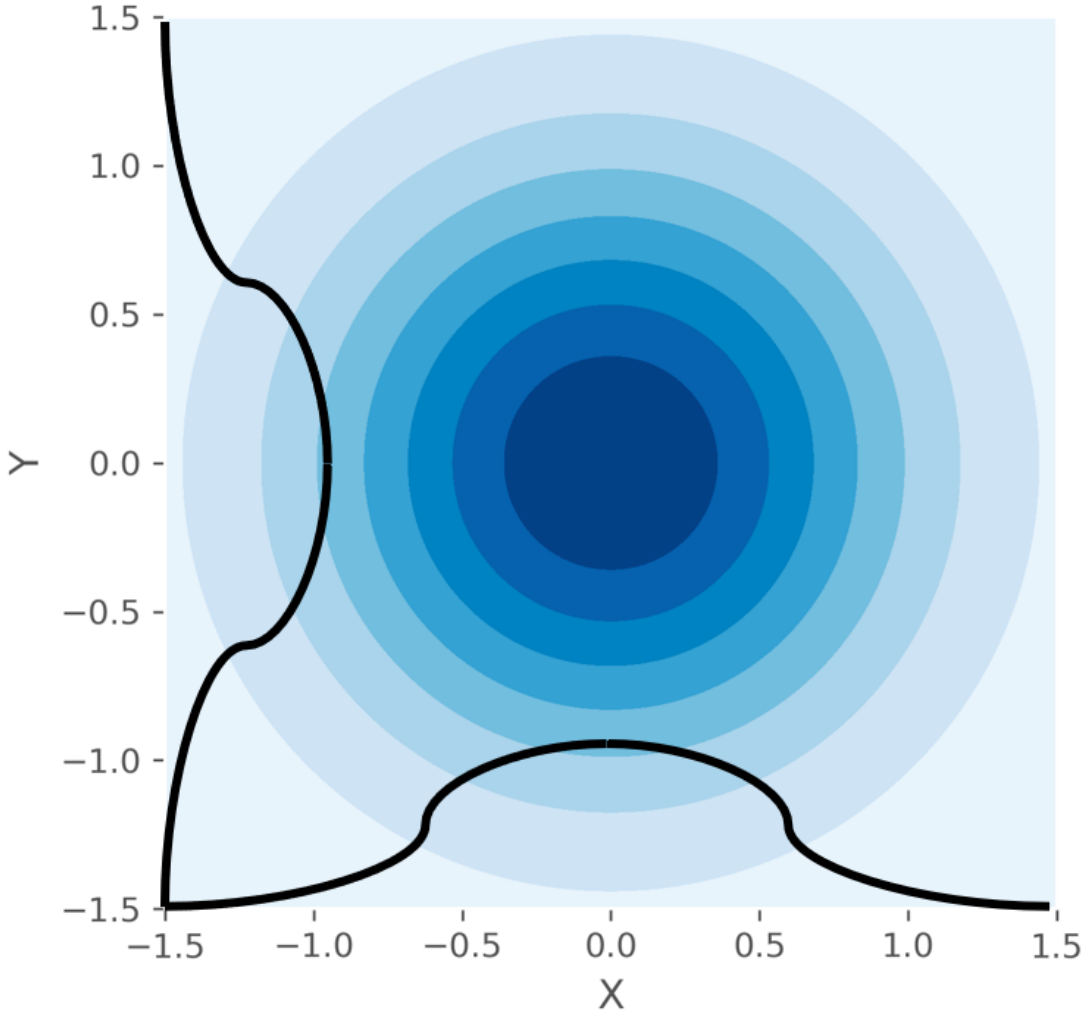


In Naïve Bayes, we assume all features are independent, so covariance matrix off-diagonal elements are all zero

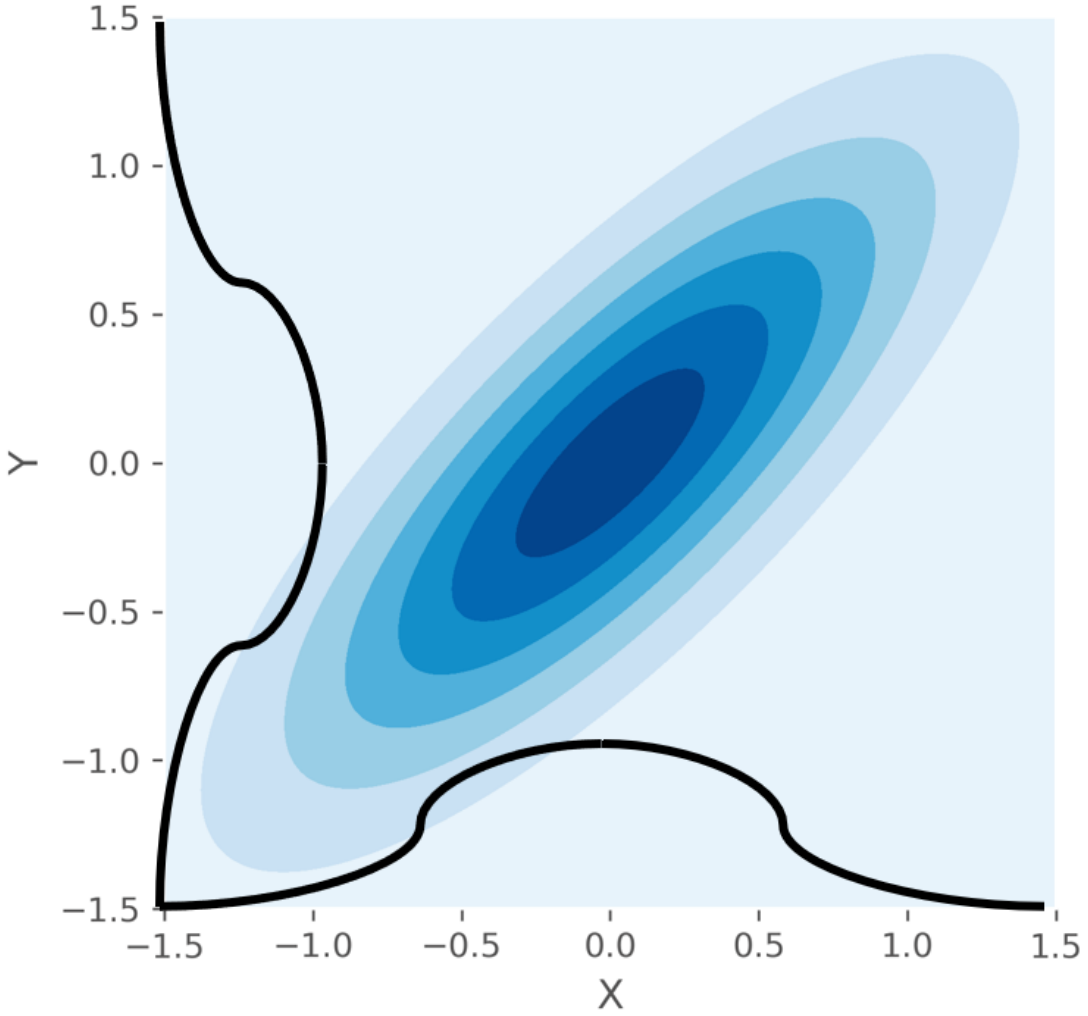
$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

The features could have some correlation!

$$\mathcal{N}\left(\begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}\right)$$



$$\mathcal{N}\left(\begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}\right)$$



Gaussians can be easily marginalized and conditioned

$$p(y_1, y_2 | \mu, \Sigma) = N\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2\end{bmatrix}\right)$$

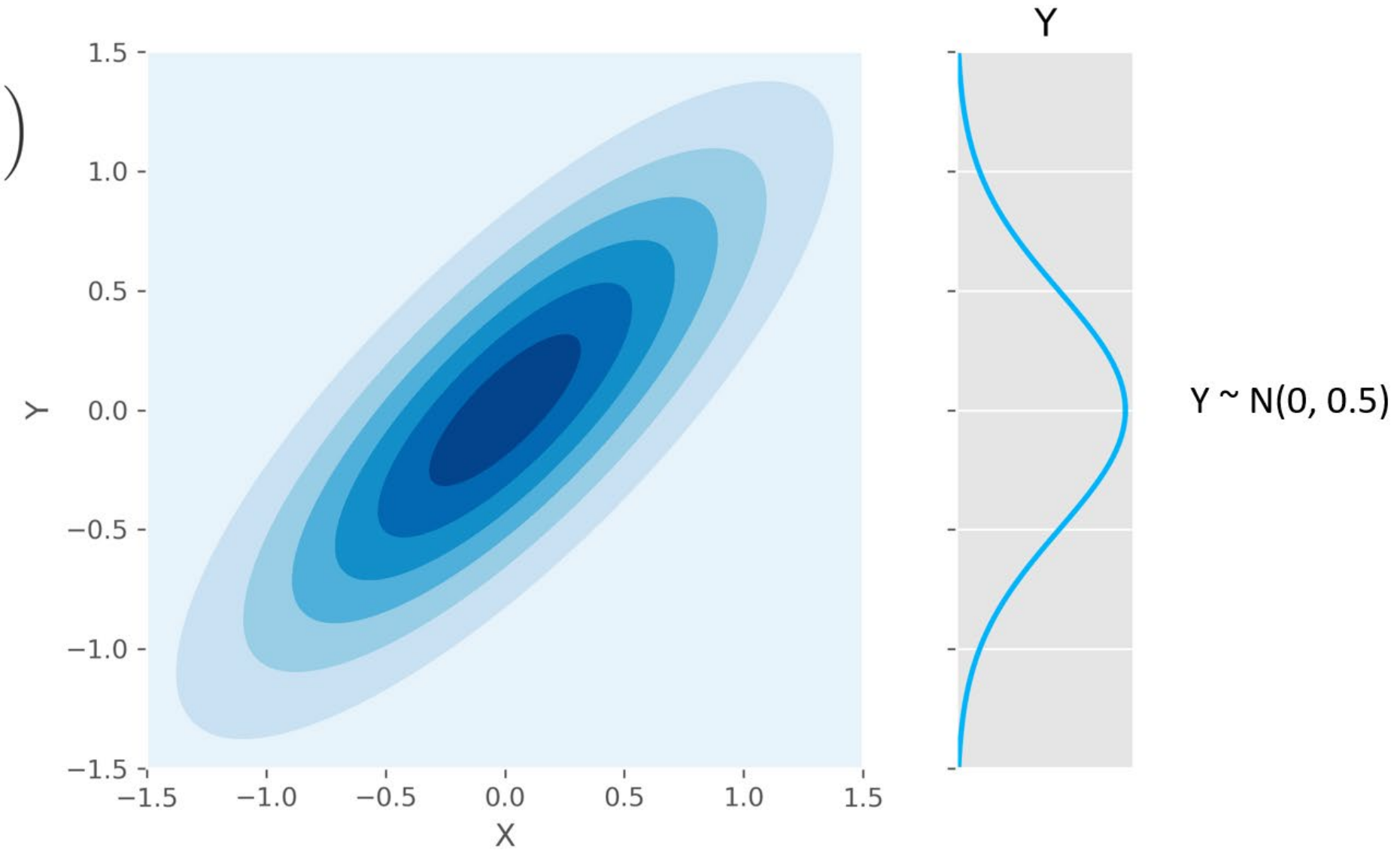
1. They are easy to **marginalize** (remove all parts of covariance matrix except what you want to keep)
   - if you want to ditch y_2, then drop $\mu_2, \sigma_1\sigma_2\rho, and\ \sigma_2^2$

2. They are easy to **condition** (given the value of one value in our multivariate normal distribution, what do we expect some other unobserved value to be?)
   - $p(x|y) = N\left(\mu_x + \Sigma_{xx}\Sigma_y^{-1}\left(y - \mu_y\right)\right), \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^T$

     Conditional mean , conditional covariance
   - Where μ_x is marginal mean (easy) and Σ_x is marginal covariance each minus some easy to calculate terms

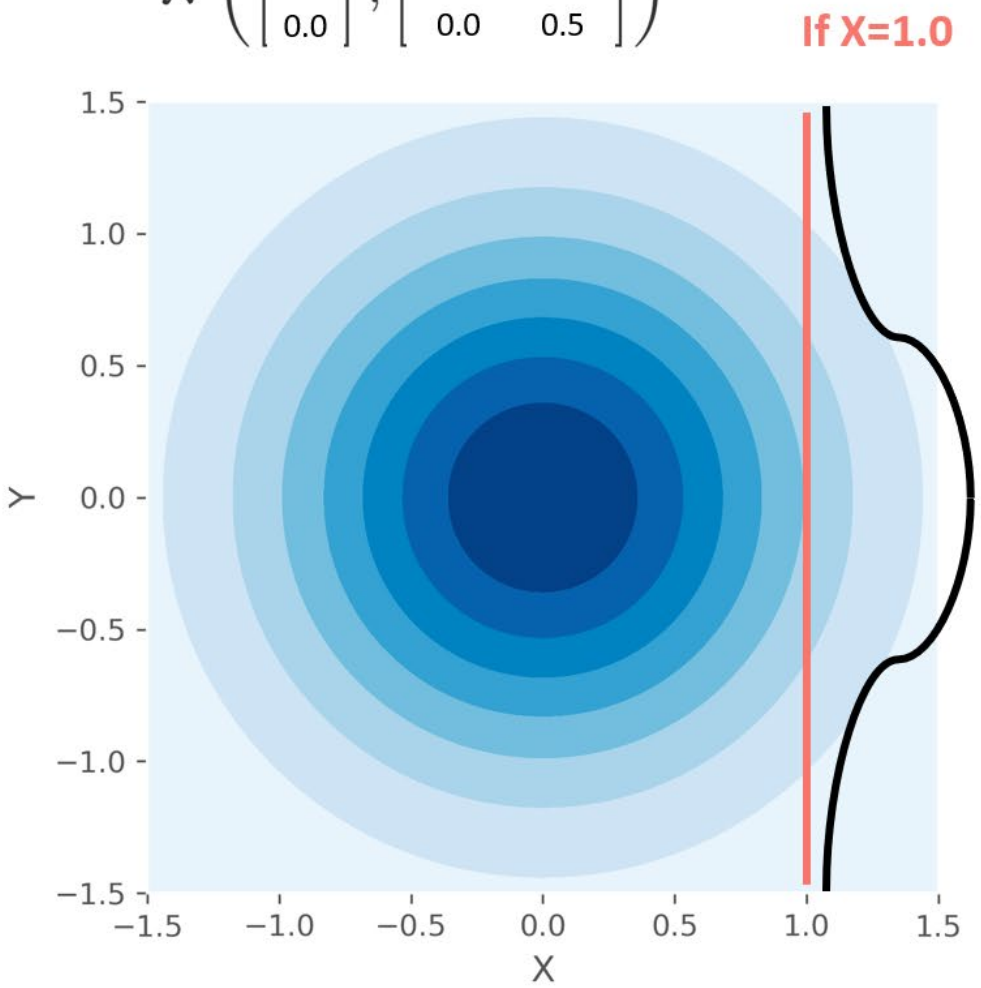
Marginalization allows us to extract individual distributions

$$\mathcal{N}\left(\begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}\right)$$

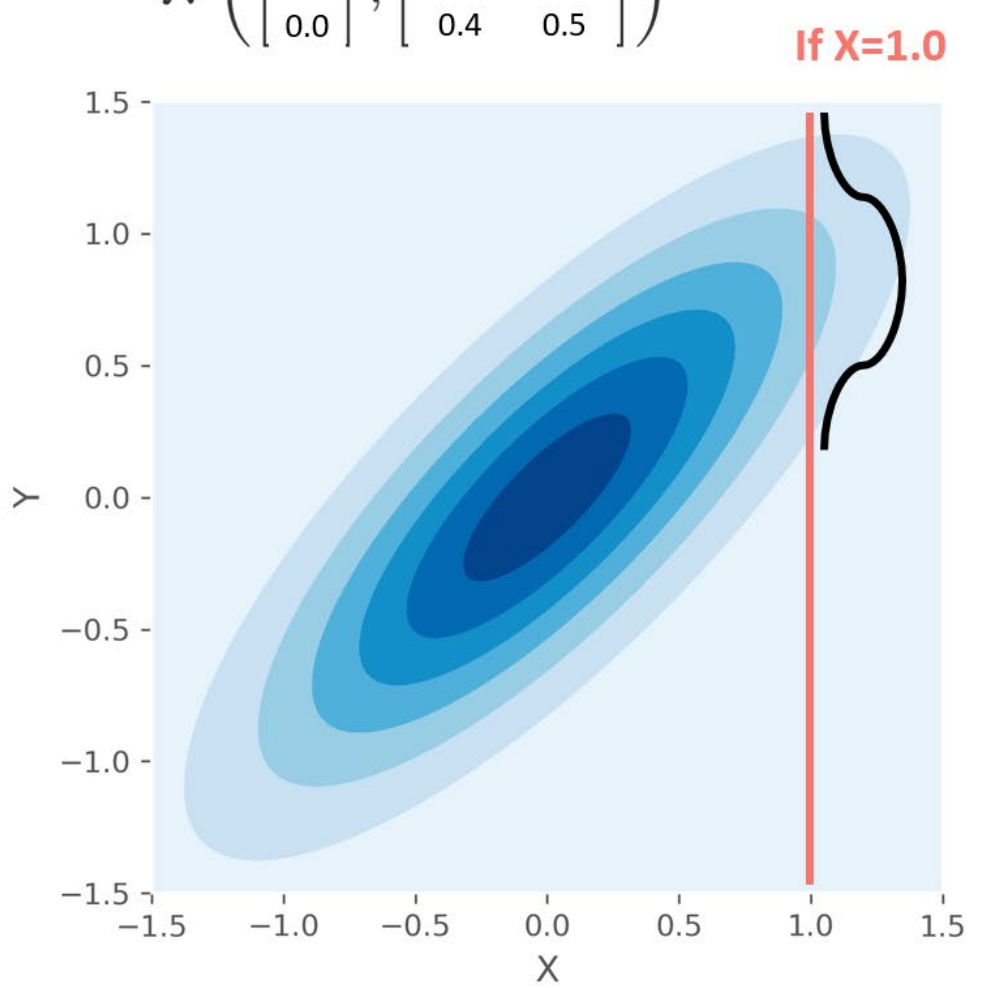


Conditioning tells us how random variables relate to each other

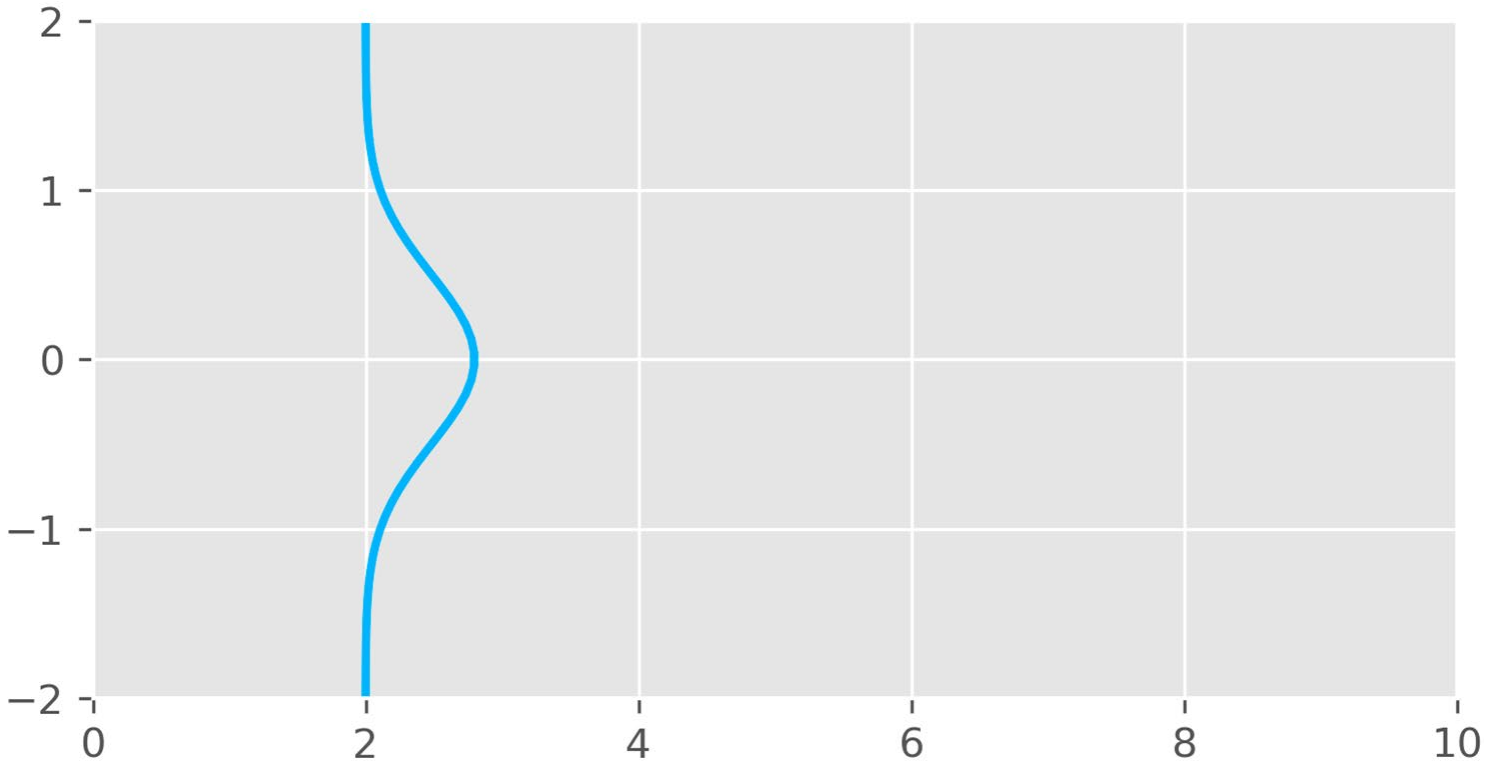
$$\mathcal{N}\left(\begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}\right)$$



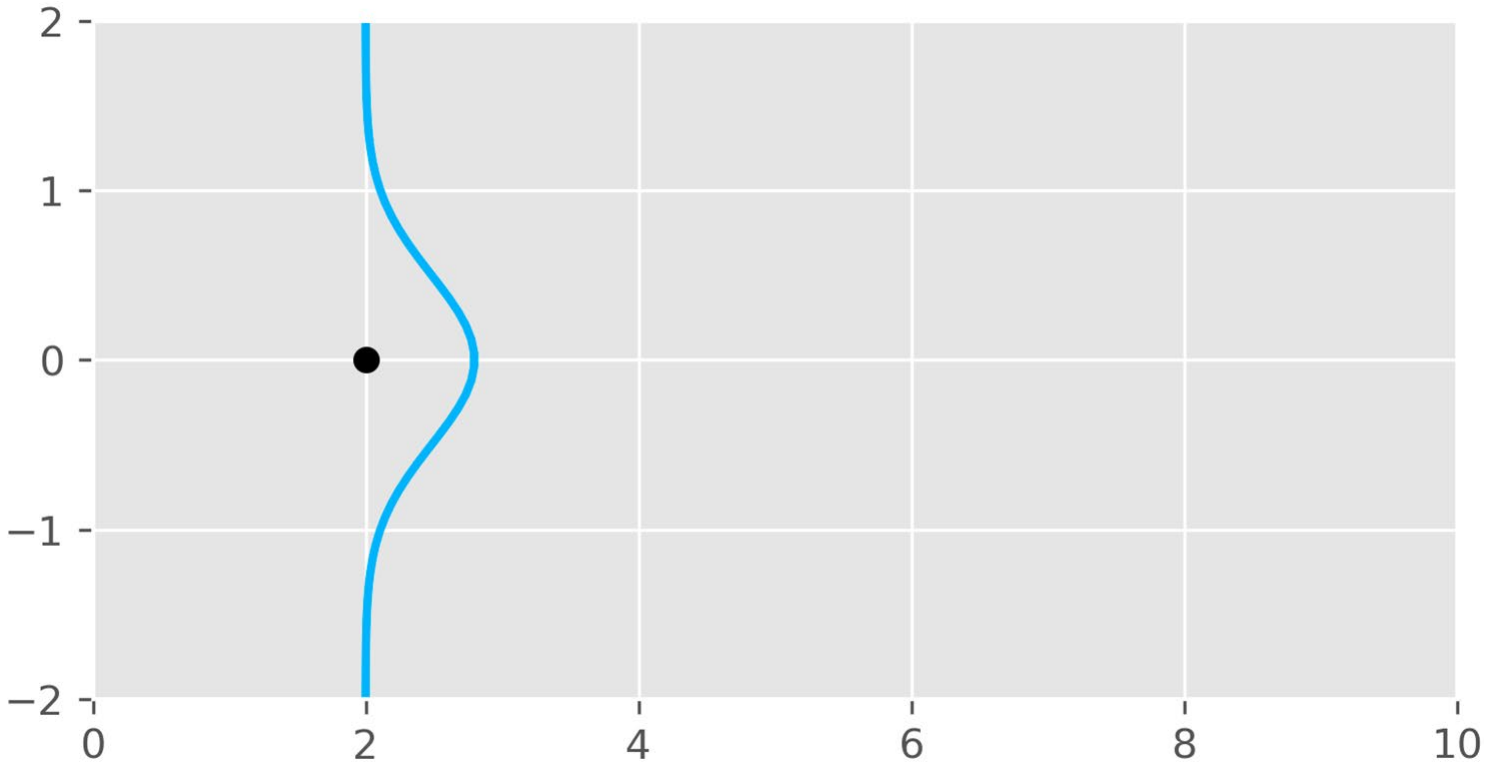
$$\mathcal{N}\left(\begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}\right)$$



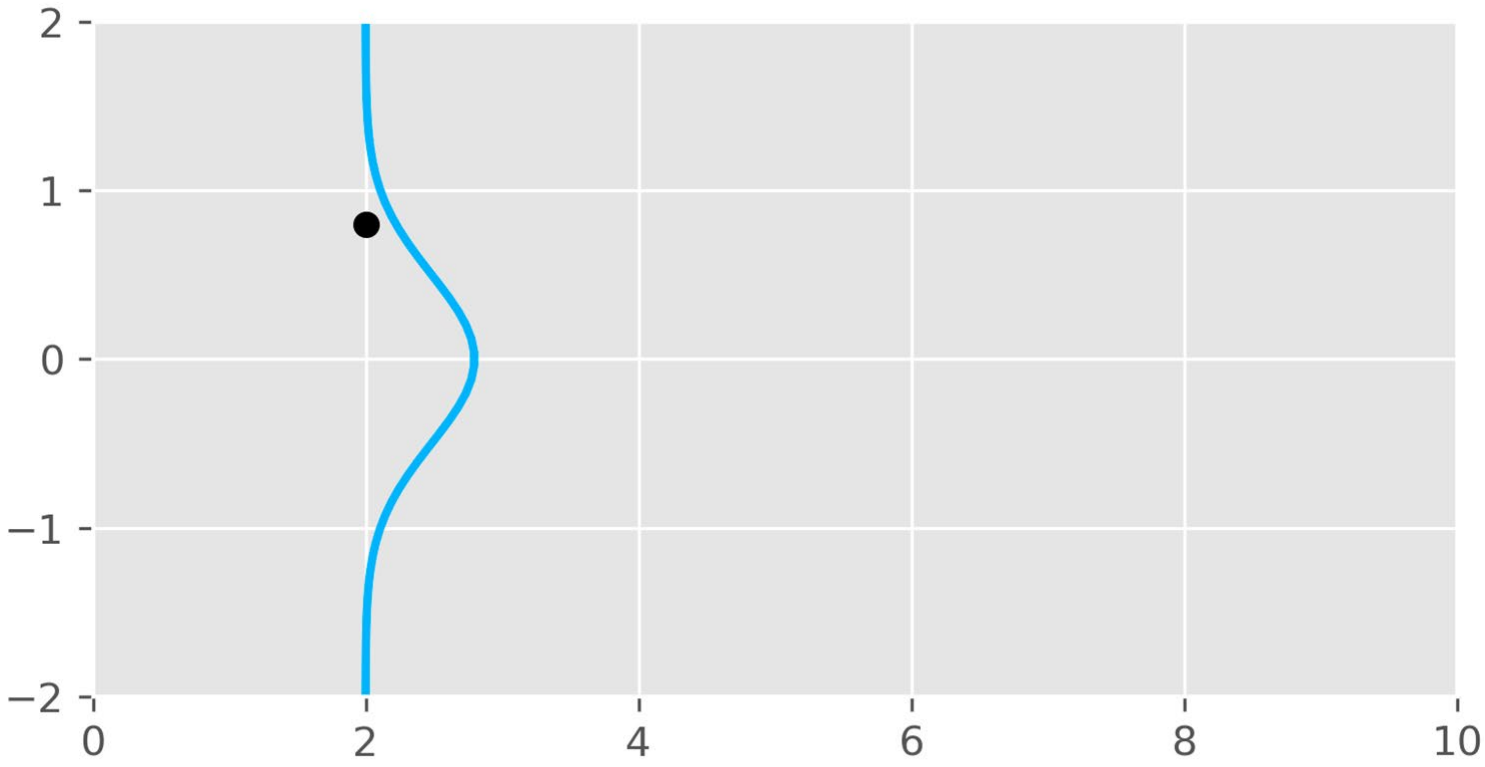
Gaussian Process is a collection of infinitely many Gaussians



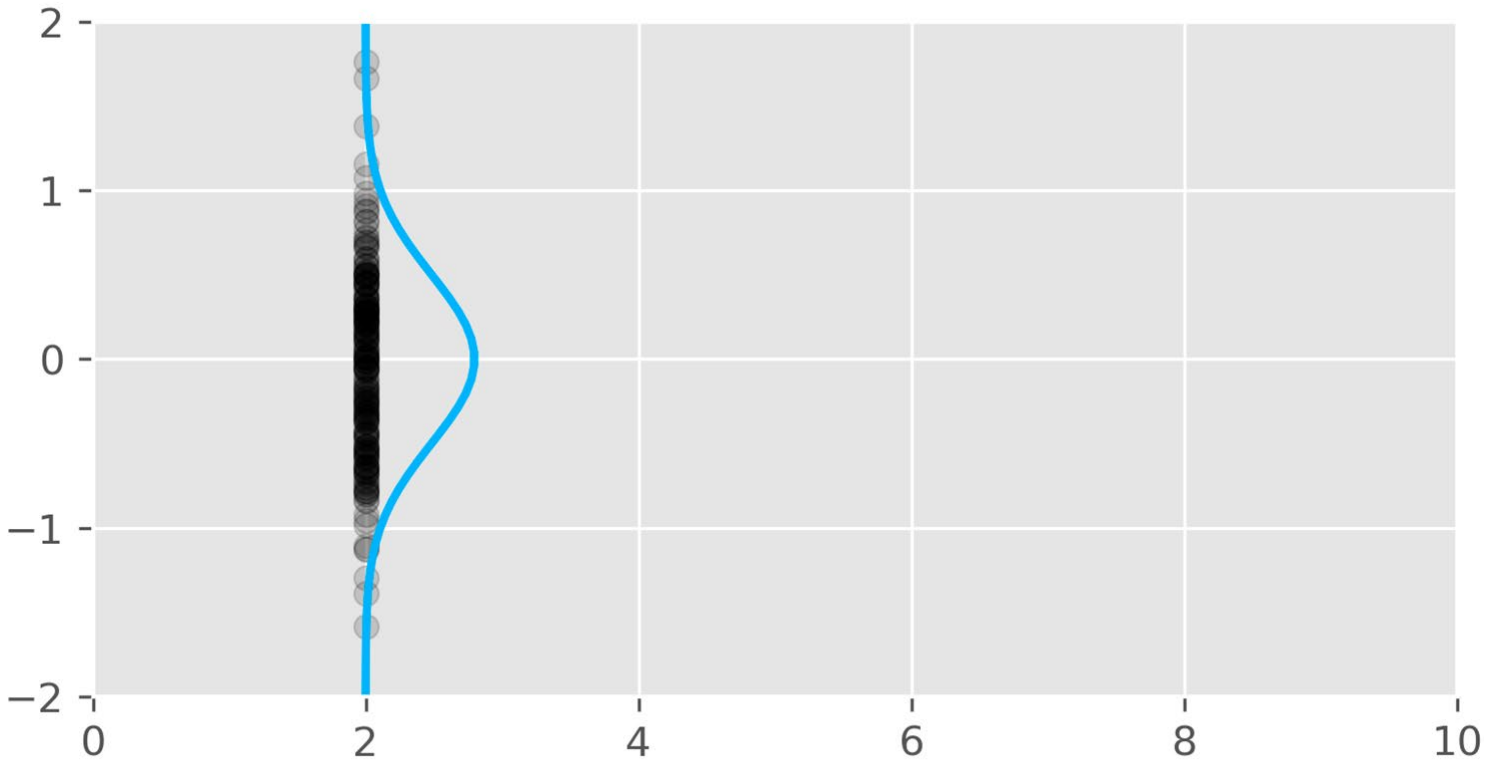
Gaussian Process is a collection of infinitely many Gaussians



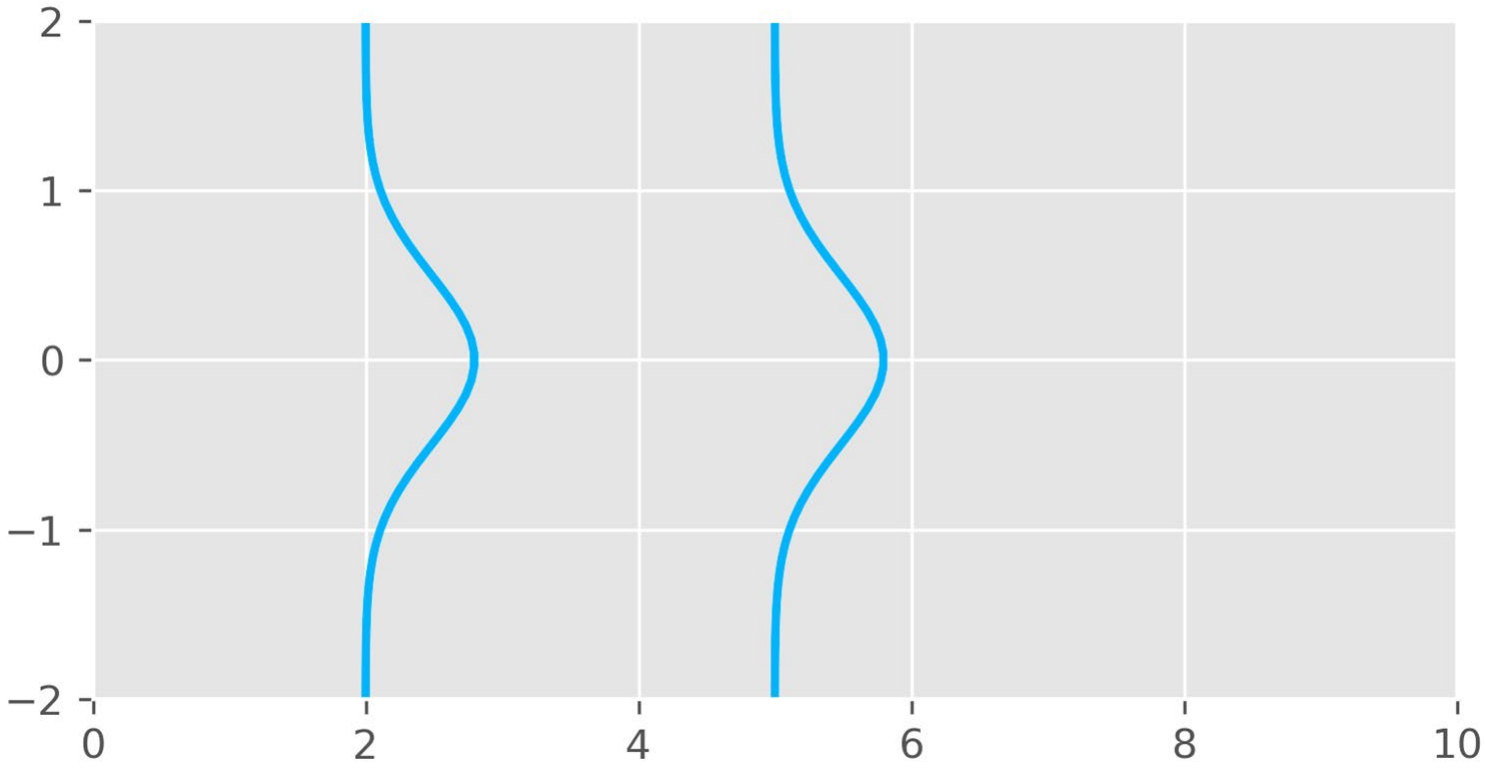
Gaussian Process is a collection of infinitely many Gaussians



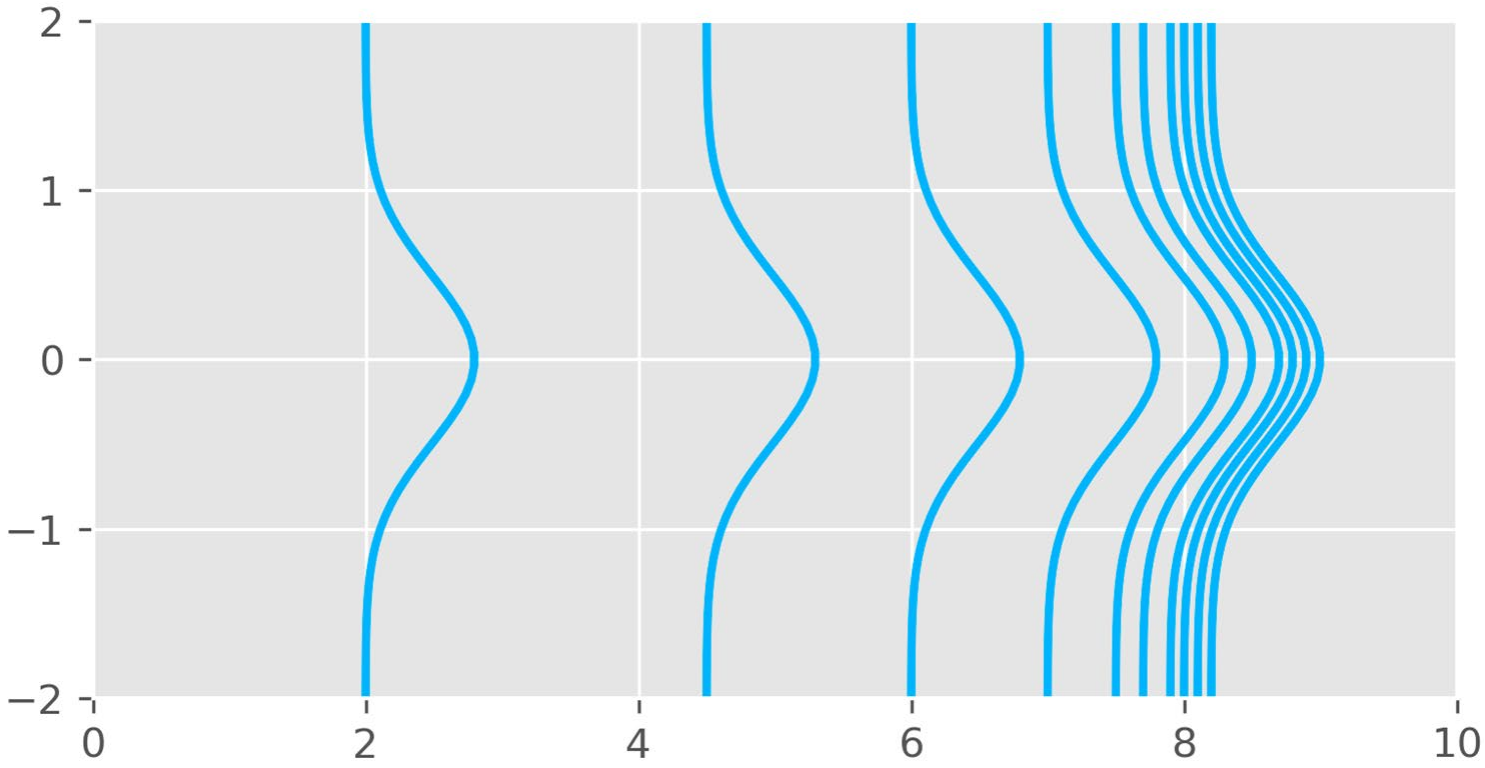
Gaussian Process is a collection of infinitely many Gaussians



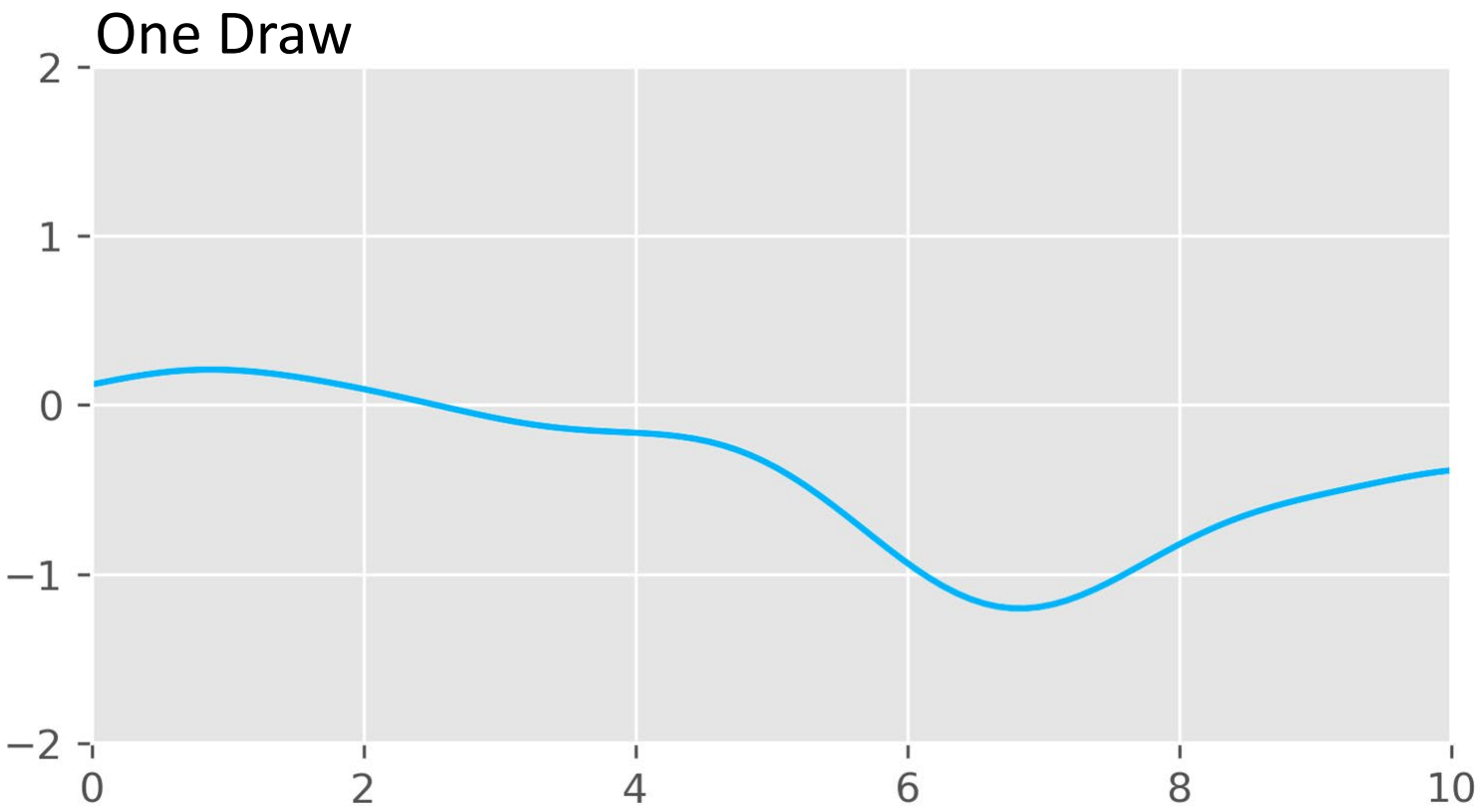
Gaussian Process is a collection of infinitely many Gaussians



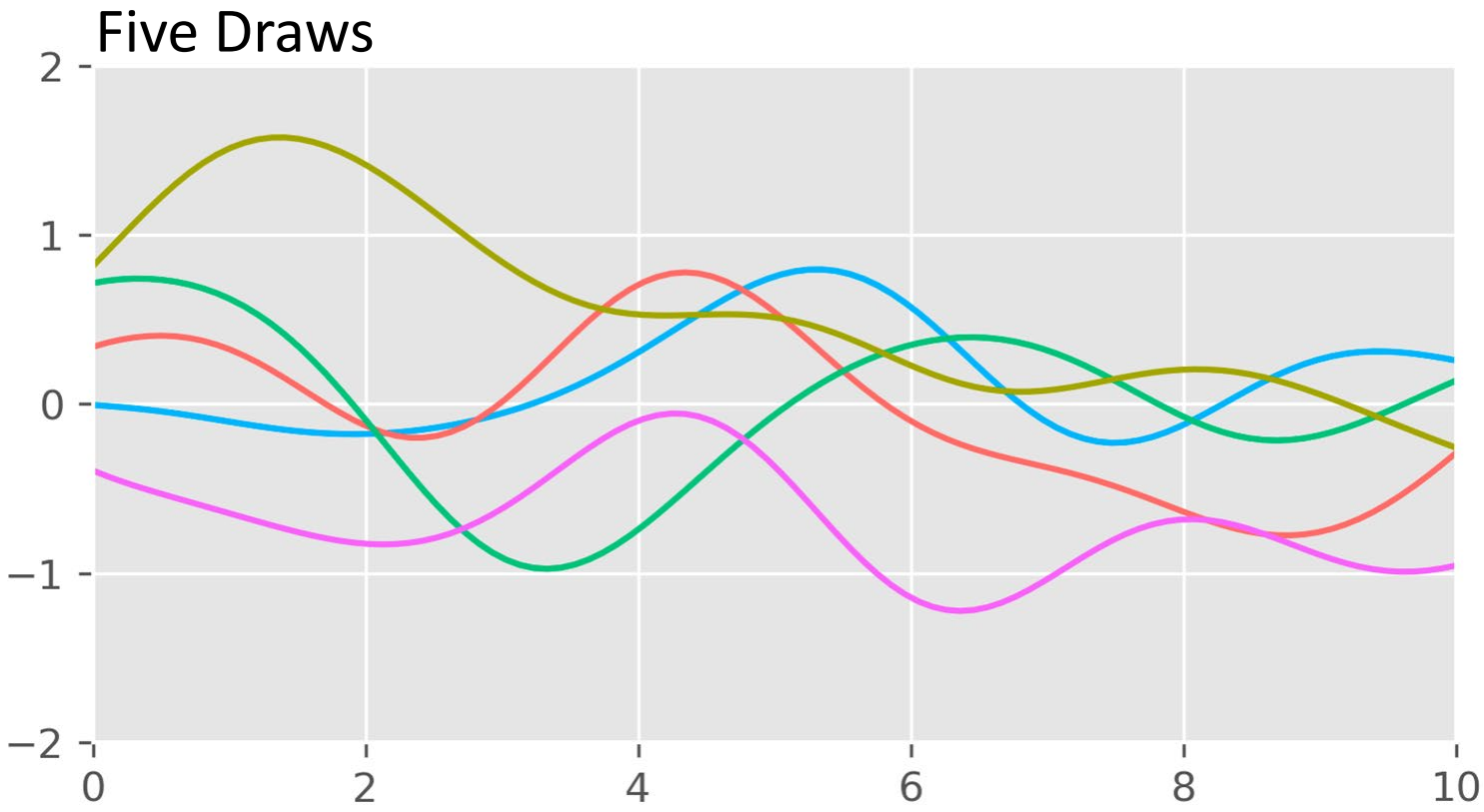
Gaussian Process is a collection of infinitely many Gaussians



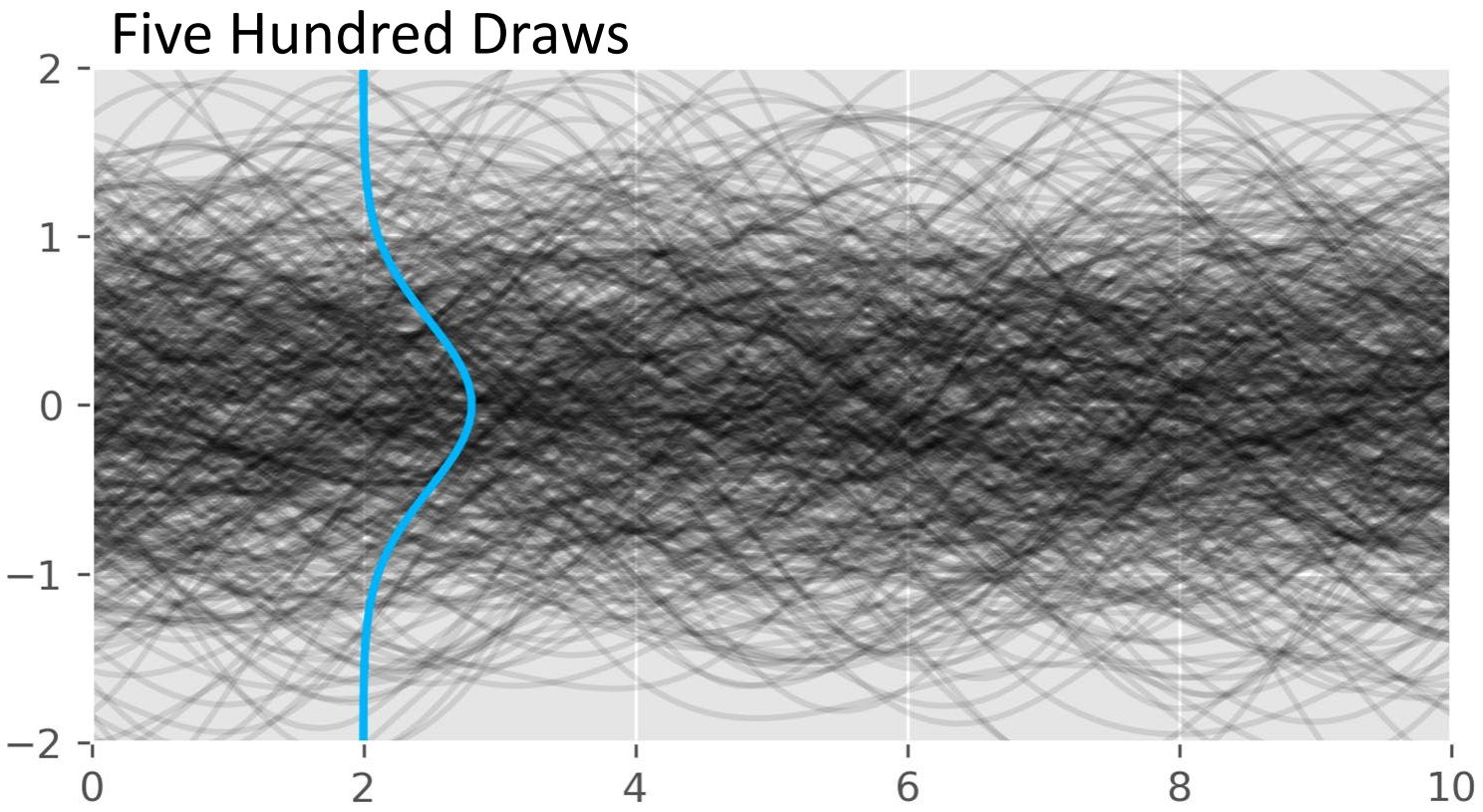
Drawing from Marginalized Gaussians produces a single function



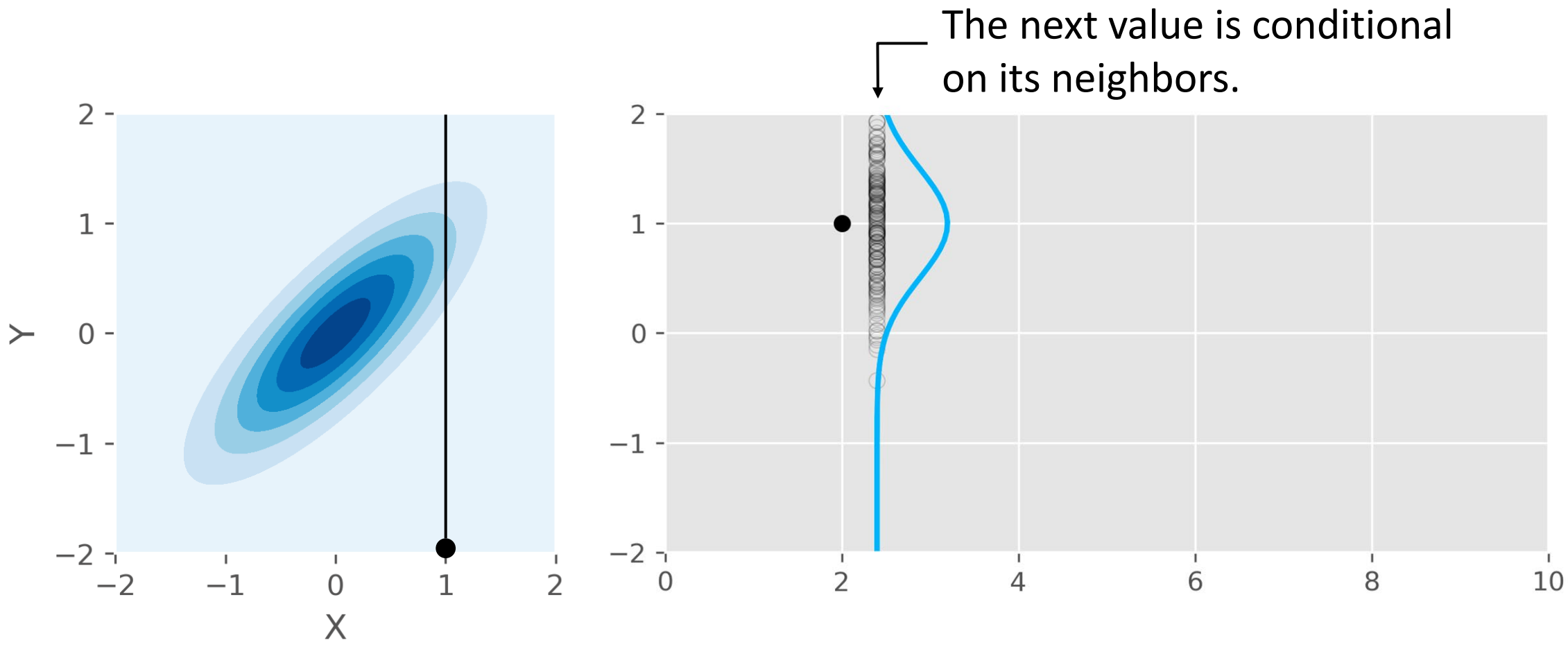
Drawing from Marginalized Gaussians produces many functions



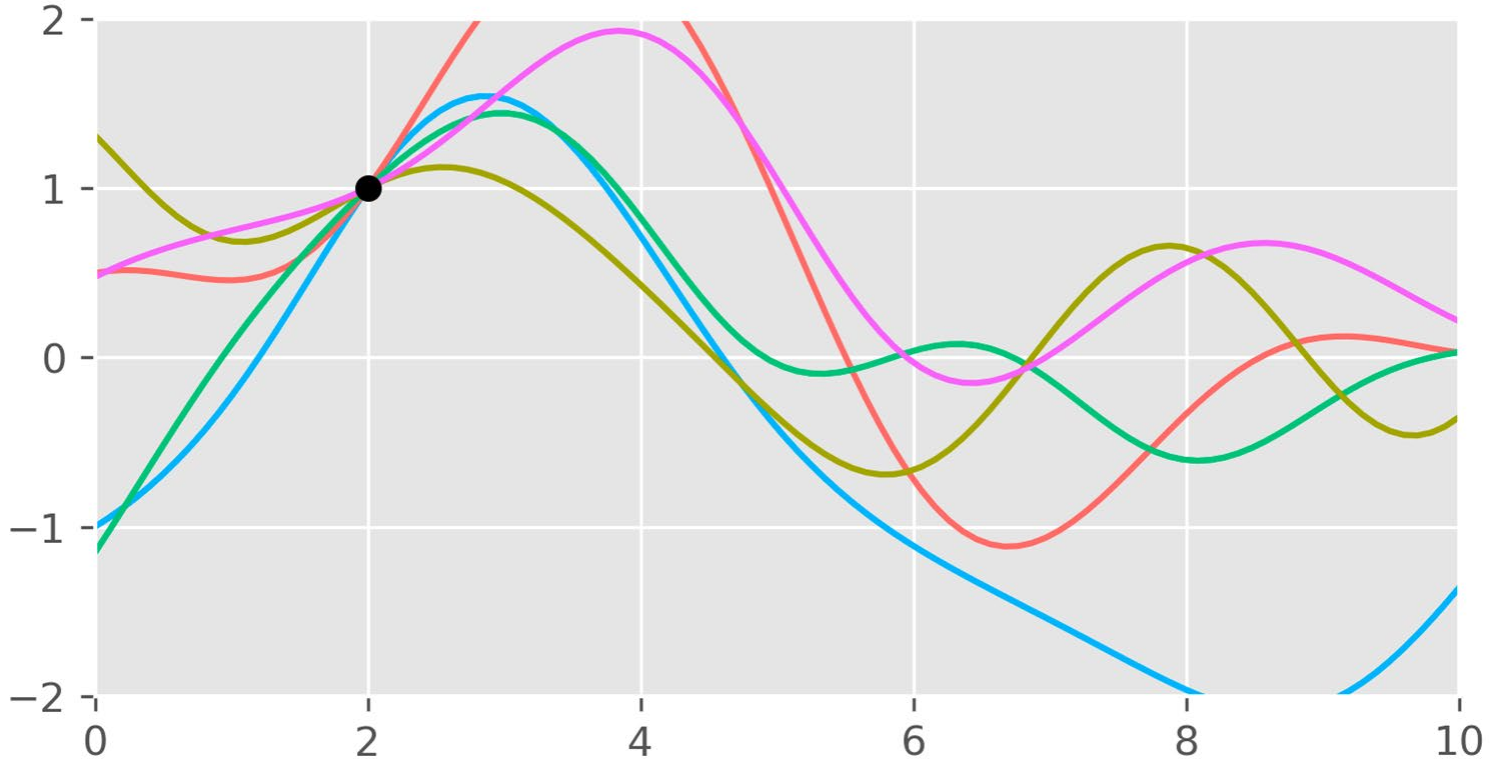
Drawing from Marginalized Gaussians produces infinite functions



Conditioning restricts neighboring points to a subset of values

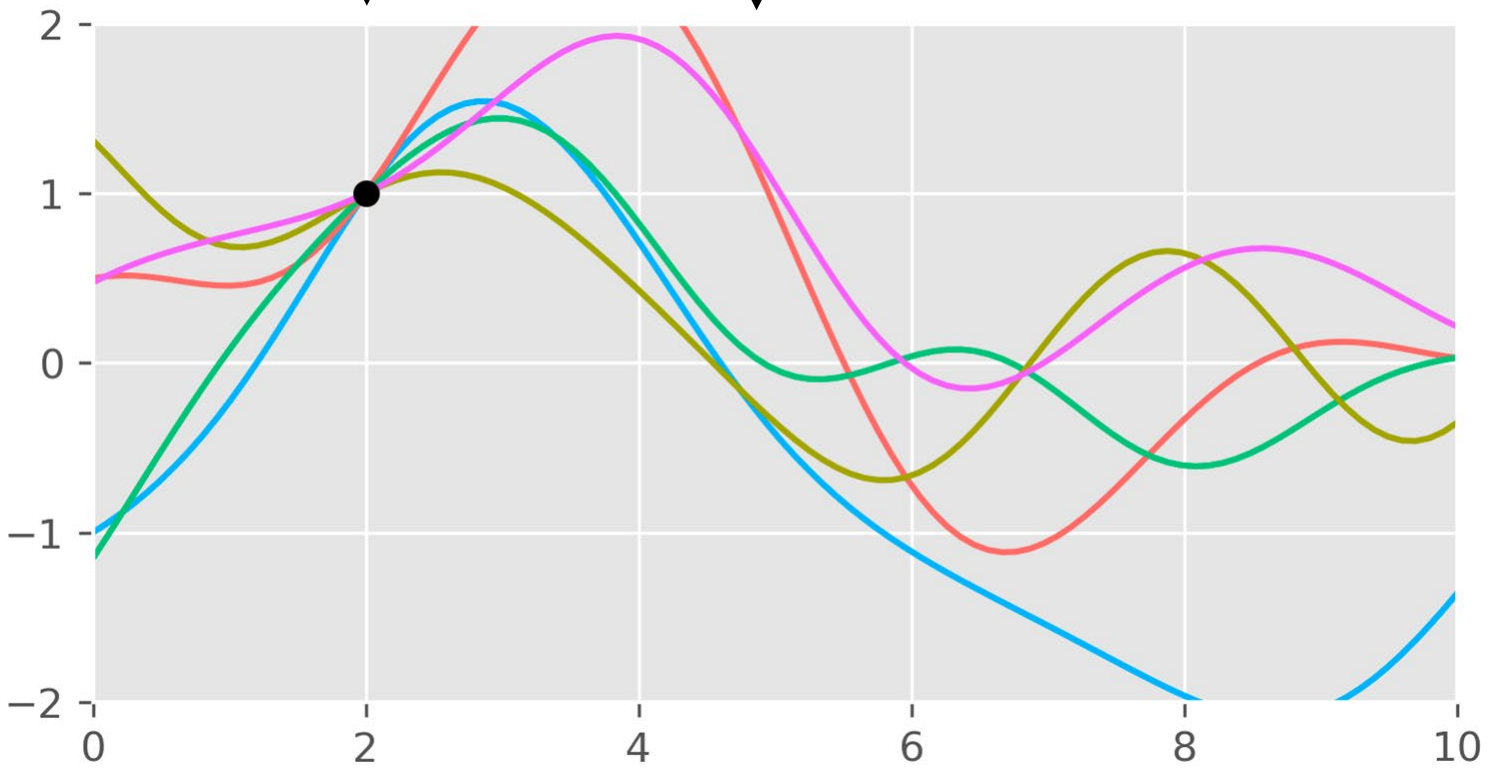


Function draws are now conditioned on observed data



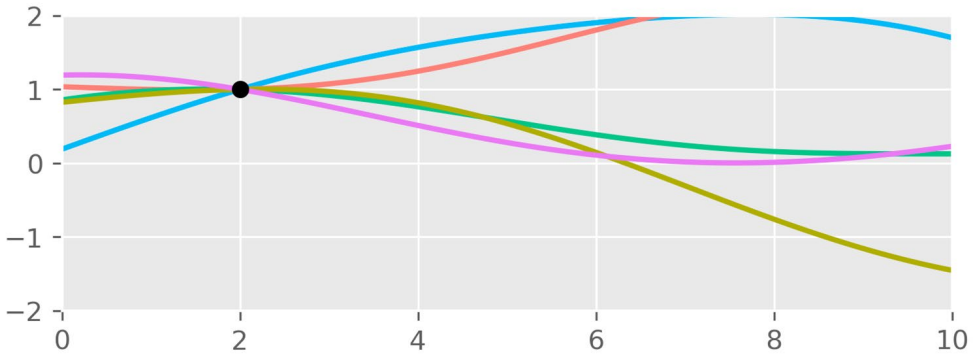
The influence of a point is governed by the length scale parameter

How much should this observation influence this point?

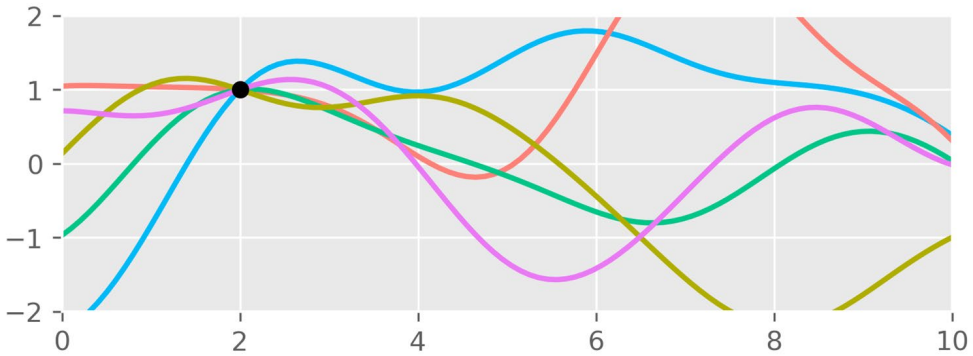


Correlation (lengthscale)

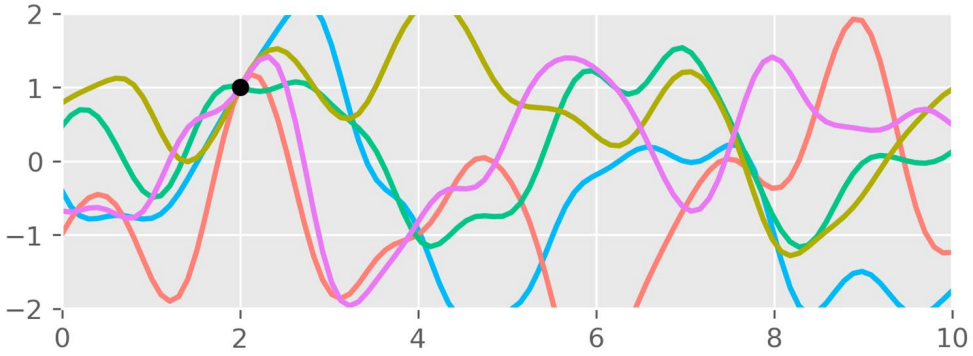
Longer length scales give more influence



L = 5

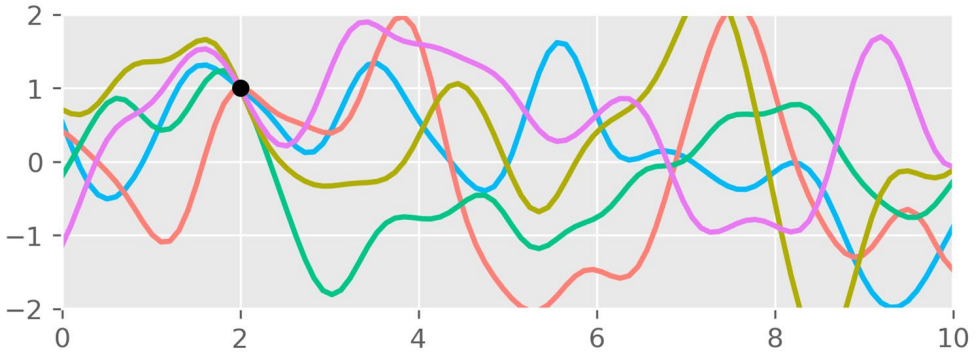


L = 1.5

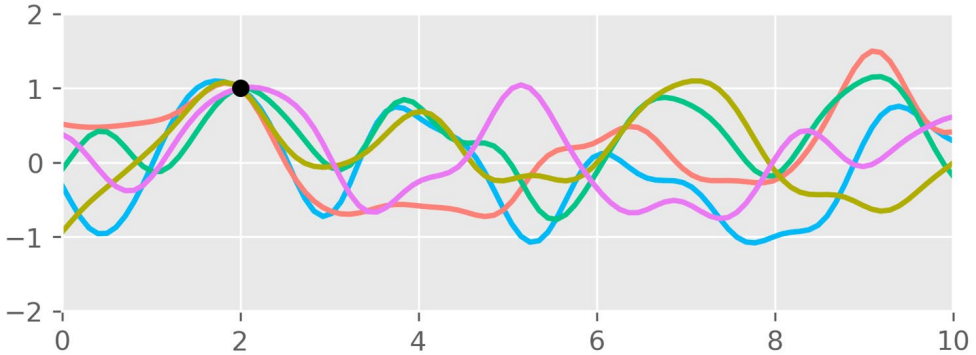


L = 0.5

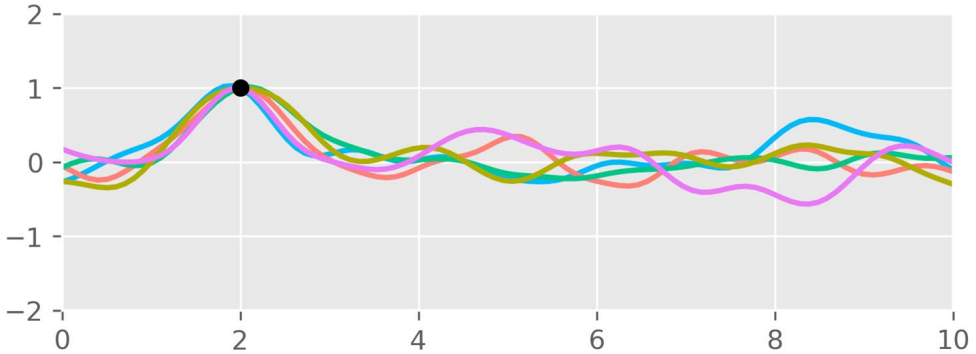
The variance parameter controls the spread of possible functions



$\sigma = 1.0$



$\sigma = 0.25$



$\sigma = 0.05$

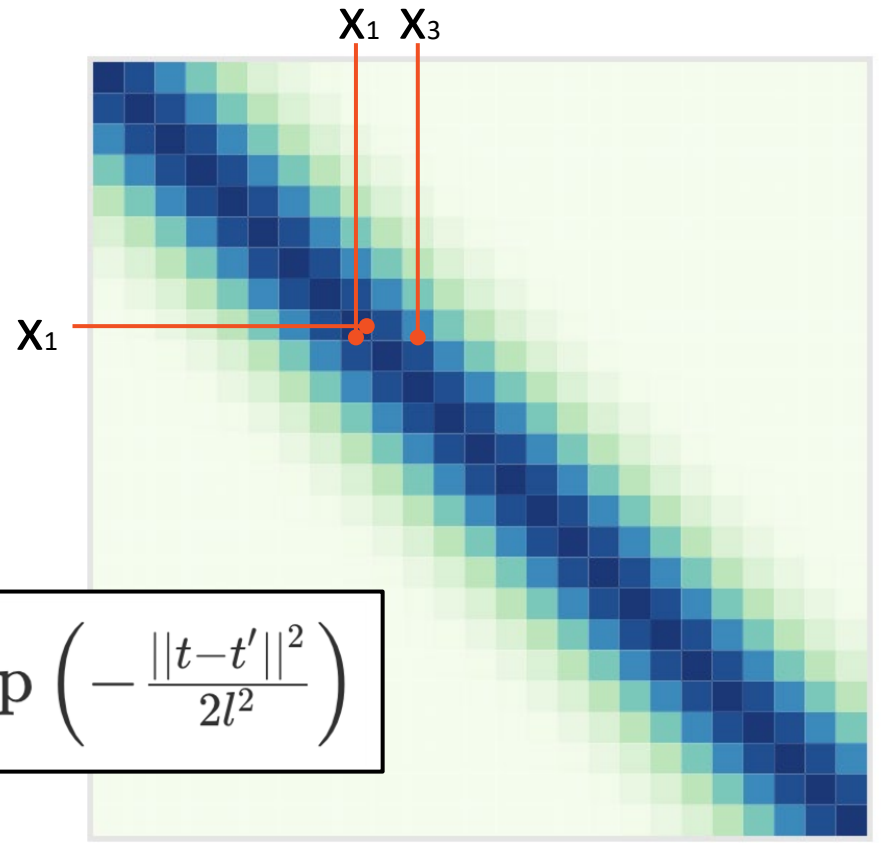
http://www.infinitecuriosity.org/vizgp/

A proper GP is defined by a mean and covariance function

$f(x) \sim GP(m, k)$

Usually a constant, but it doesn't have to be!

Defines how two points in space are correlated.



RBF Kernel

$$k(t,t') = \sigma^2 \exp\left(-\frac{||t-t'||^2}{2l^2}\right)$$

Figure referenced from:
https://distill.pub/2019/visual-exploration-gaussian-processes/

There are tons of different covariance functions

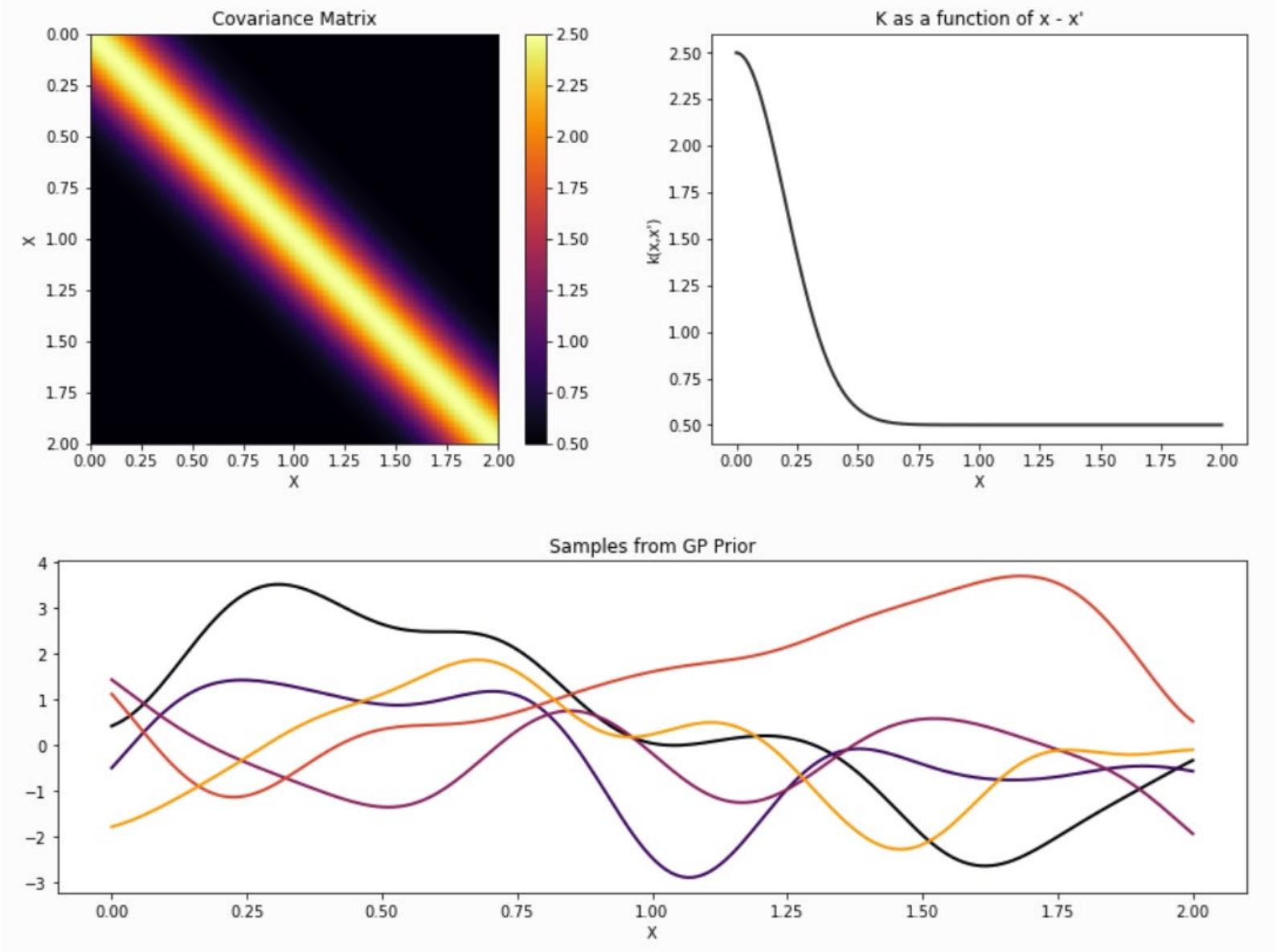


Figure Referenced From
https://www.pymc.io/projects/examples/en/latest/gaussian_processes/GP-MeansAndCovs.html

There are tons of different covariance functions

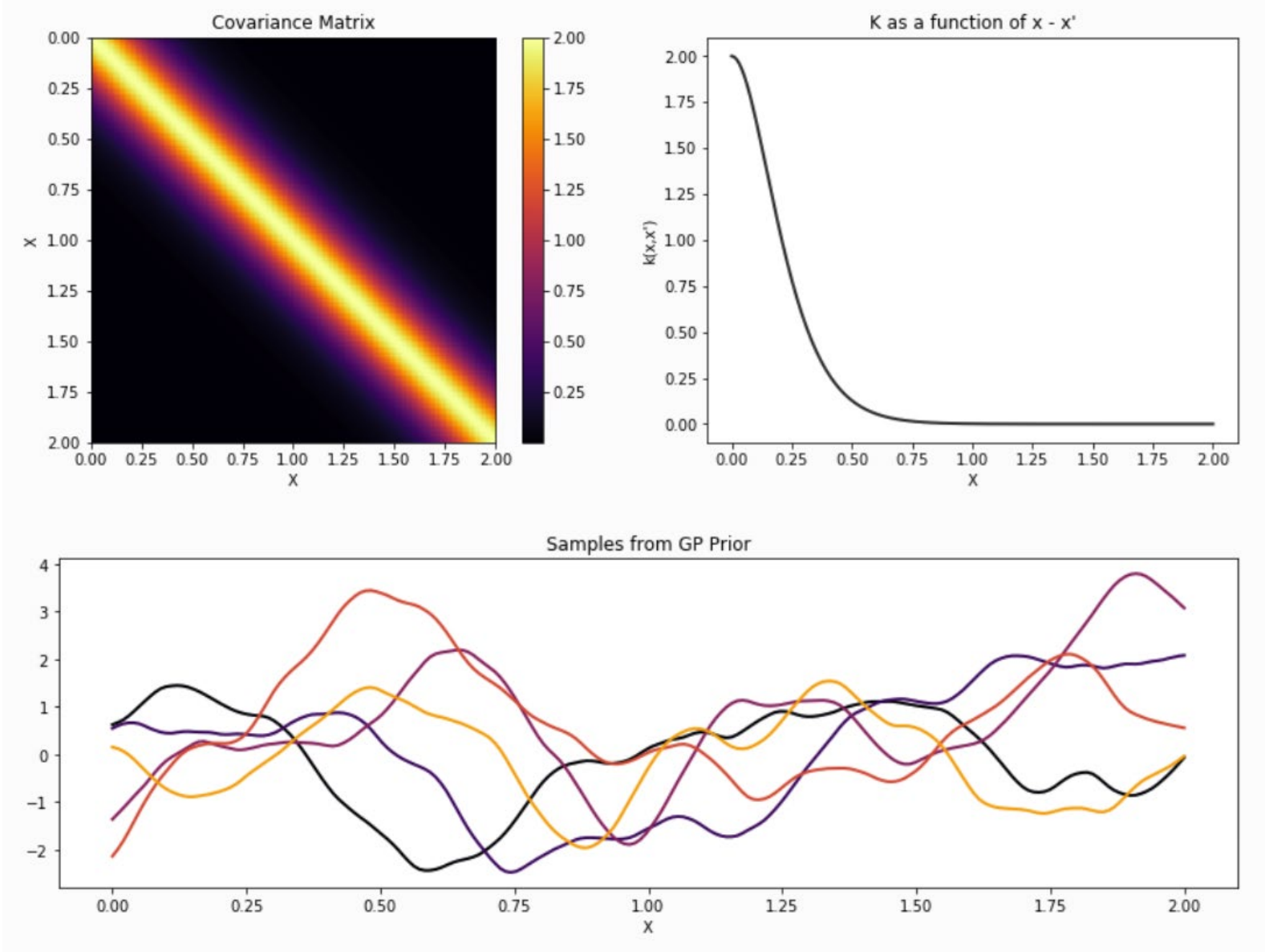


Figure Referenced From
https://www.pymc.io/projects/examples/en/latest/gaussian_processes/GP-MeansAndCovs.html

There are tons of different covariance functions

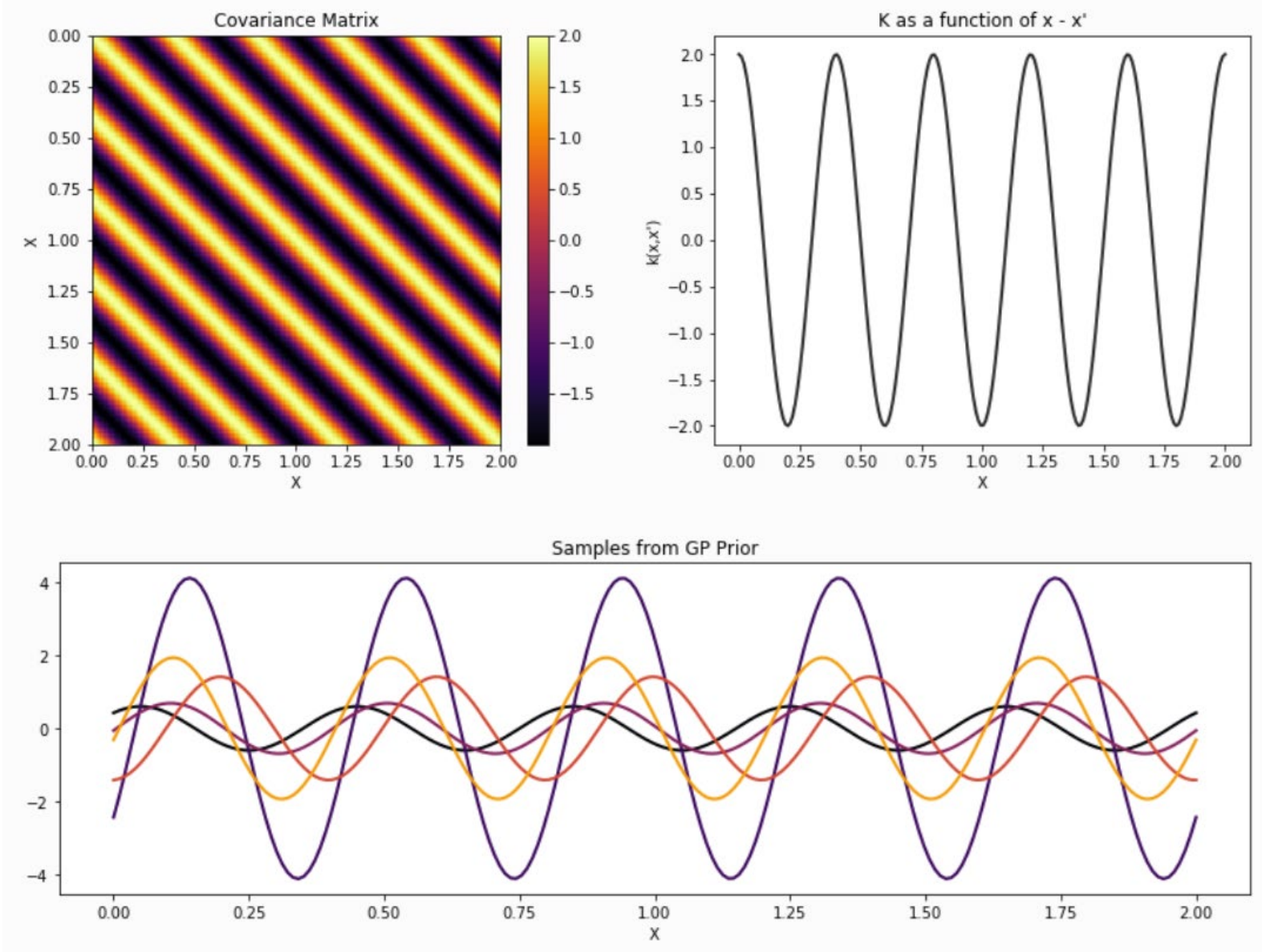


Figure Referenced From
https://www.pymc.io/projects/examples/en/latest/gaussian_processes/GP-MeansAndCovs.html

There are tons of different covariance functions

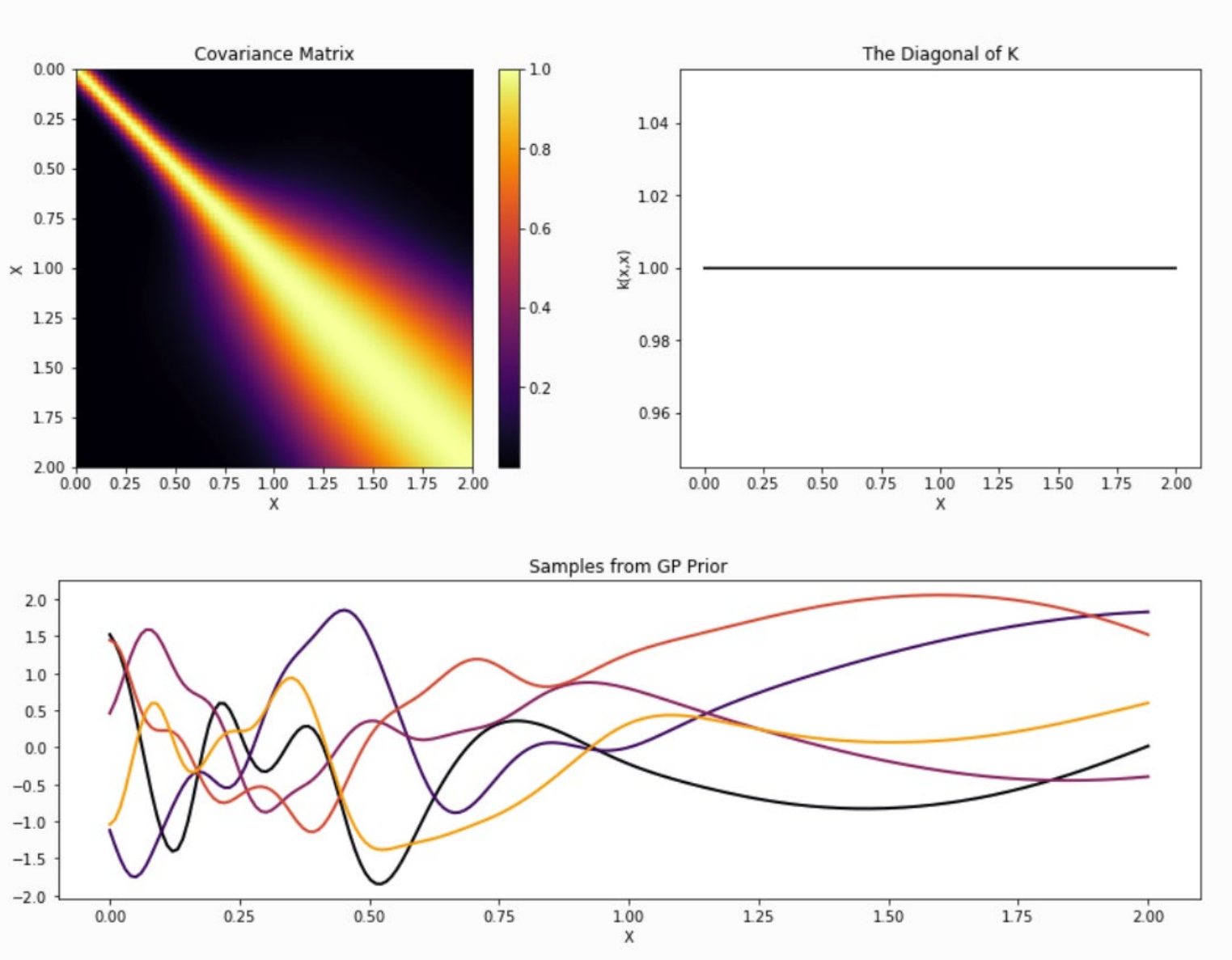
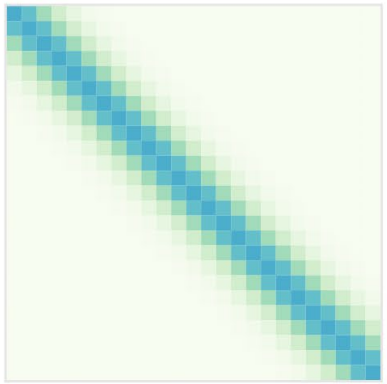


Figure Referenced From
https://www.pymc.io/projects/examples/en/latest/gaussian_processes/GP-MeansAndCovs.html

Tricky part of Gaussian process is figuring out which function to use to model covariance

RBF KERNEL

$$\sigma^2 \exp\left(-\frac{||t-t'||^2}{2l^2}\right)$$

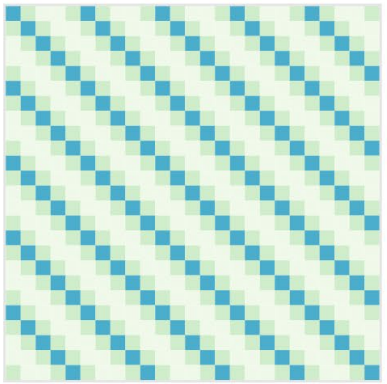


Variance σ = 0.8

Length l = 0.8

PERIODIC

$$\sigma^2 \exp\left(-\frac{2\sin^2(\pi|t-t'|/p)}{l^2}\right)$$



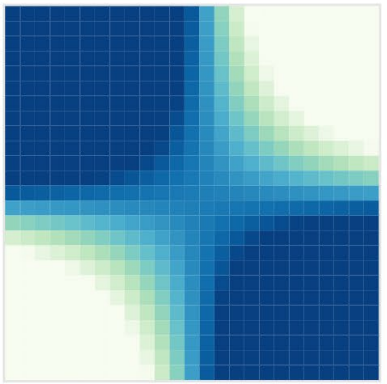
Variance σ = 0.8

Length l = 0.8

Periodicity p = 0.5

LINEAR

$$\sigma_b^2 + \sigma^2(t-c)(t'-c)$$



Variance σ = 0.35

Variance σ_b = 0.8

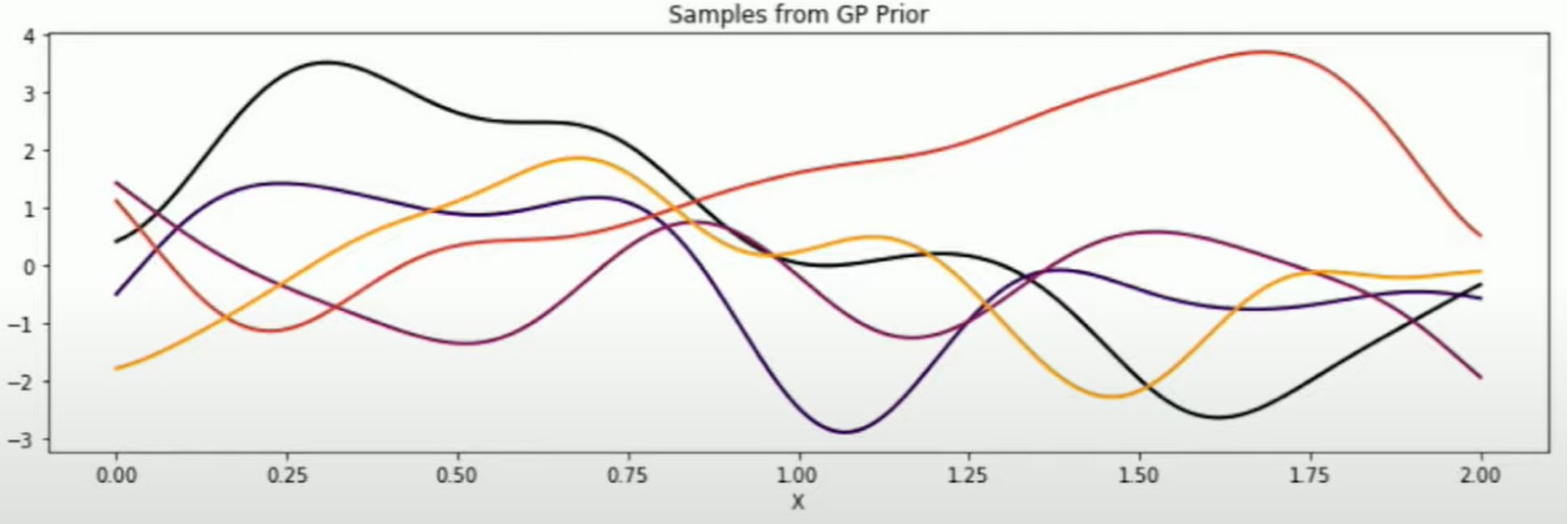
Offset c = 0

For the **Linear** kernel
the parameter **Variance σ** determines
the average distance away from the function's mean.

https://distill.pub/2019/visual-exploration-gaussian-processes/

Consider all the different covariance functions we can get from just one exponential quadratic

$$k(x, x') = \exp\left[-\frac{(x - x')^2}{2\ell^2}\right]$$

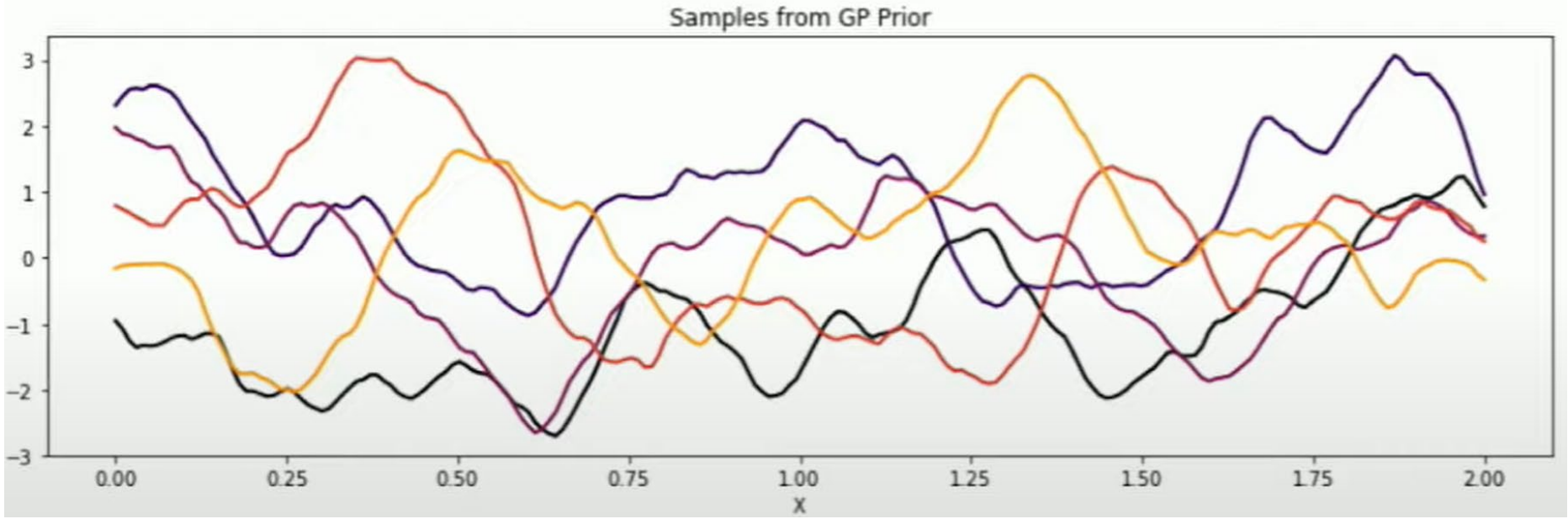


One parameter, length scale, that determines how far away you have to go away from two points to make them independent

Coding Tech (YouTube) A Primer on Gaussian Processes for Regression Analysis || Chris Fonnesbeck

Matern(3/2) is another popular covariance function

$$k(x, x') = \left(1 + \frac{\sqrt{3(x - x')^2}}{\ell}\right) \exp\left[-\frac{\sqrt{3(x - x')^2}}{\ell}\right]$$

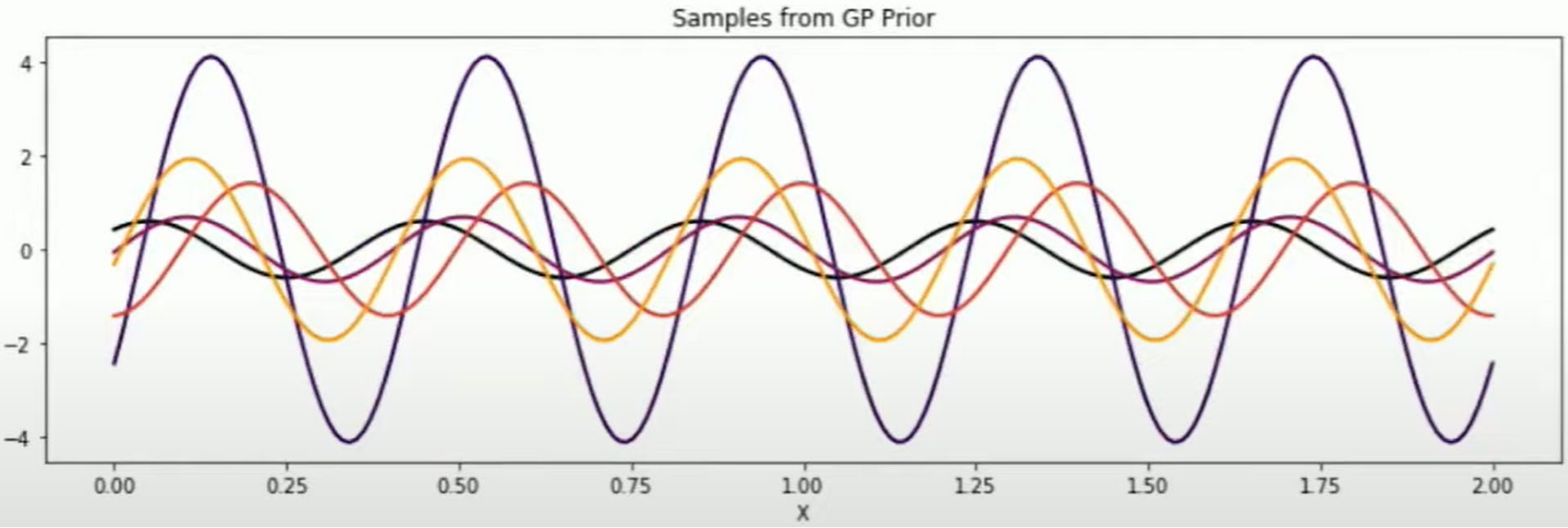


Length scale parameter, plus a scaling term to scale roughness

Coding Tech (YouTube) A Primer on Gaussian Processes for Regression Analysis || Chris Fonnesbeck

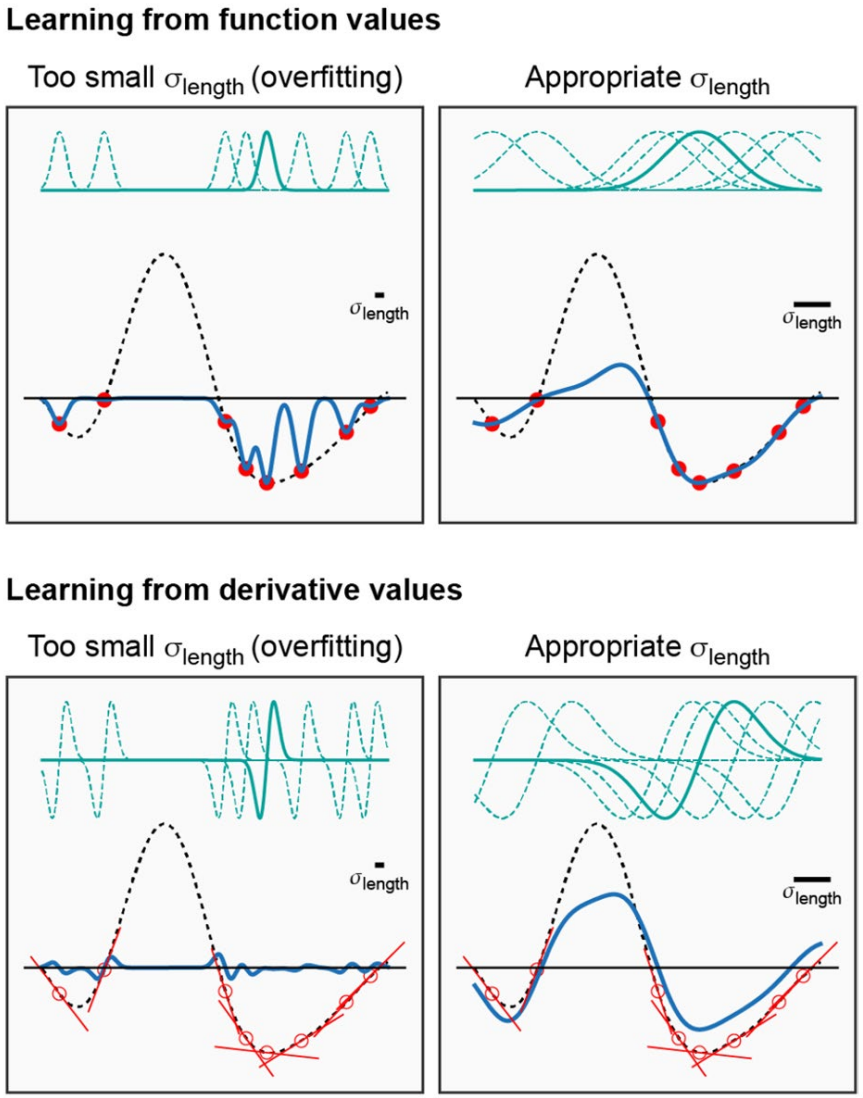
Cosine is another popular covariance function

$$k(x, x') = \cos\left(\frac{||x - x'||}{\ell^2}\right)$$



Coding Tech (YouTube) A Primer on Gaussian Processes for Regression Analysis || Chris Fonnesbeck

Getting the kernel hyperparameters right is important



Deringer et al GaussianProcessRegressionforMaterialsandMolecules, Chemical Reviews 2021

A mean function is also necessary, but not as hard to pick

Mean function generates mean vectors

$$\mu \sim m(x|\phi)$$

Simplest is just zero!

$$m(x) = 0$$

Or constant

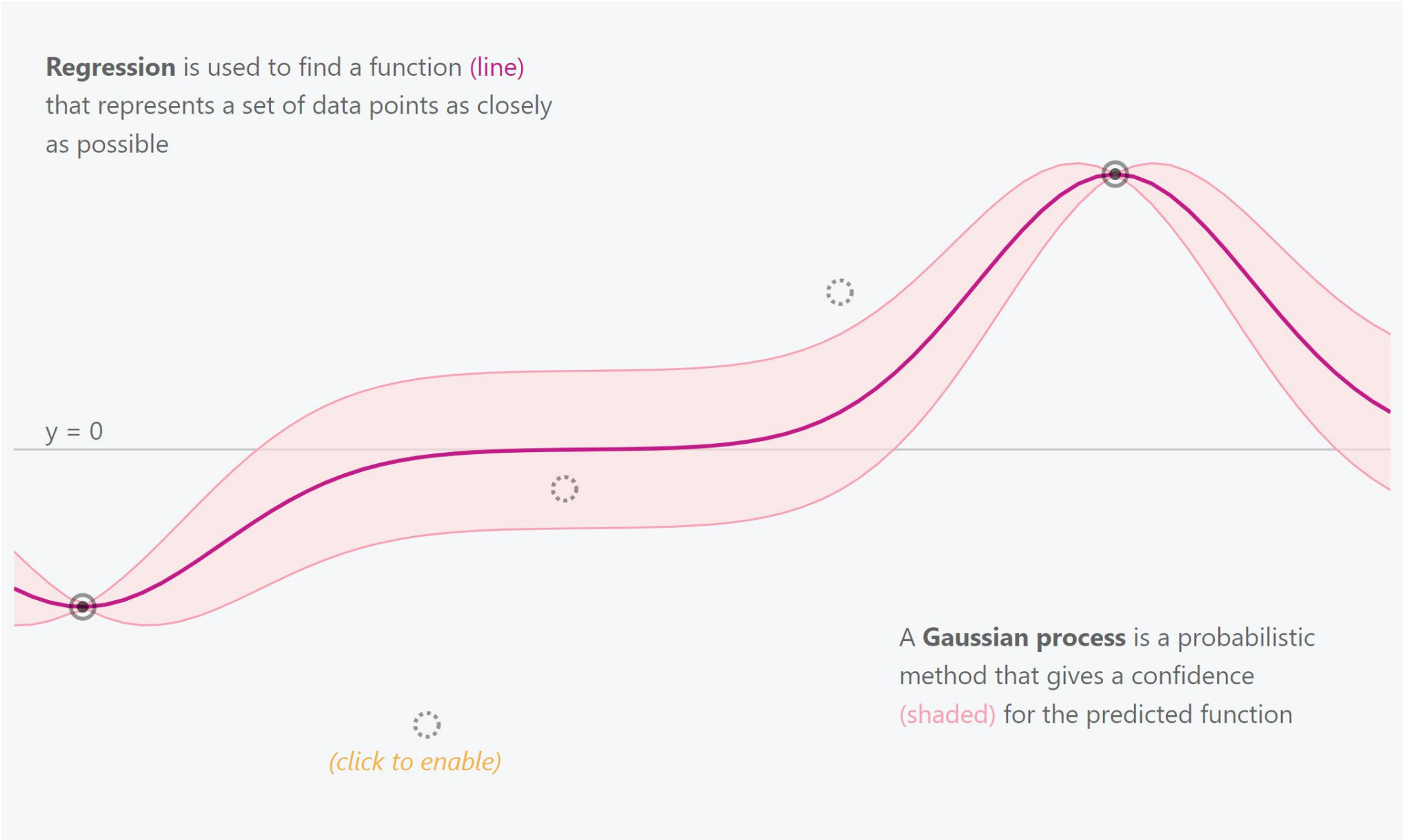
$$m(x) = C$$

Or linear

$$m(x) = Ax + b$$

Posterior does not use the mean, it is just used as a guess for the prior because where you have no data, it trends towards the mean

In the absence of data, the regression trends towards the mean



https://distill.pub/2019/visual-exploration-gaussian-processes/

We now use these priors to calculate posteriors given some data observations

Recall Bayes' formula

$$p(\theta|y) \propto \prod_{i=1}^{N} p(y_i|\theta)\, p(\theta)$$

For Gaussian process, this becomes

$$Closed\ form\ GP = Gaussian\ data * GP\ prior$$

When we fit the model, we are learning the kernel and mean hyperparameters

Predictions are done using Bayes' formula as well

Posterior predictive distribution

$$p(y^{new}|y) \propto \int p(y^{new}|\theta)\, p(\theta|y) d\theta$$

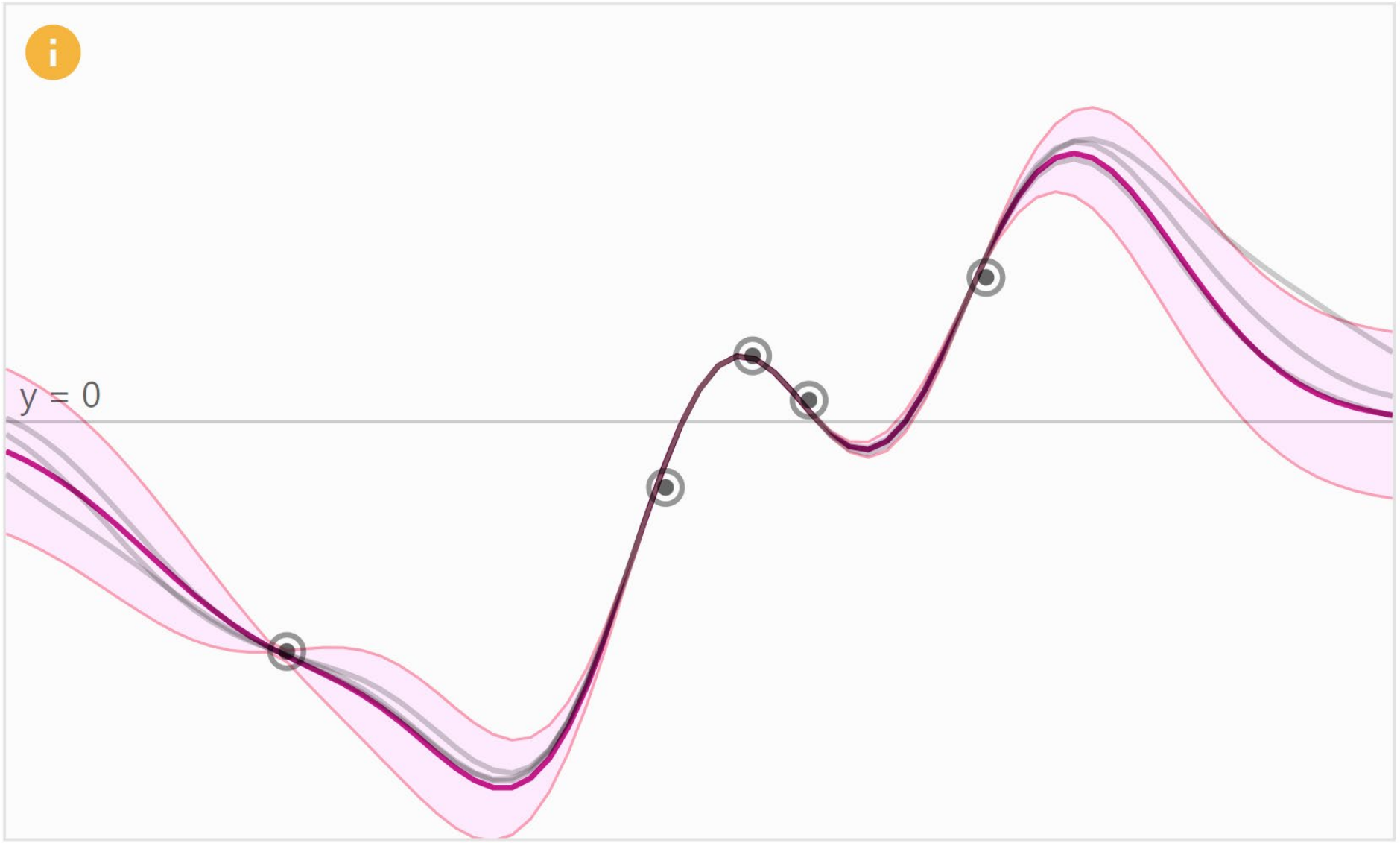
Likelihood is aleatoric uncertainty
Or noise in data itself

Posterior is epistemic uncertainty
Or our unknowns

Predictions are simply:

$$p(y^{new}|x^{new}, y, x) = N(\mu^{new}, \Sigma^{new})$$

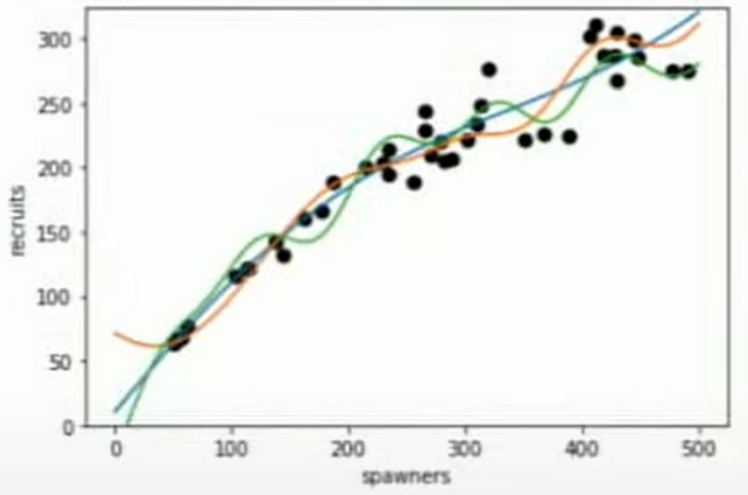
Confidence intervals and predictions come from the distribution statistics μ, Σ



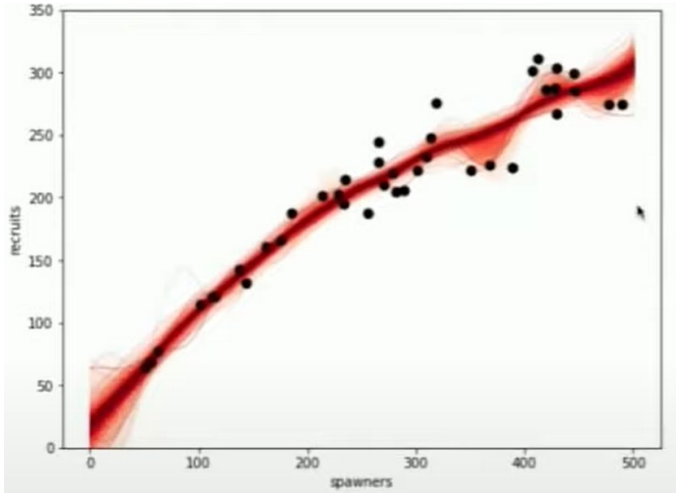
Coding Tech (YouTube) A Primer on Gaussian Processes for Regression Analysis || Chris Fonnesbeck

Confidence intervals and predictions come from the distribution statistics μ, Σ

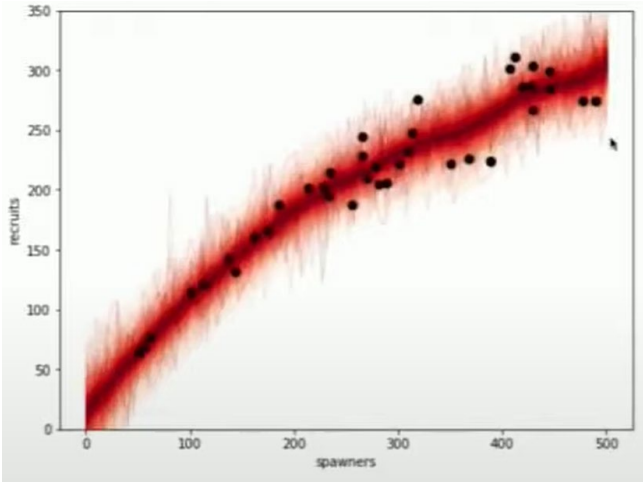
3 attempts at posteriors



100 attempts at posteriors



With prediction noise



Coding Tech (YouTube) A Primer on Gaussian Processes for Regression Analysis || Chris Fonnesbeck

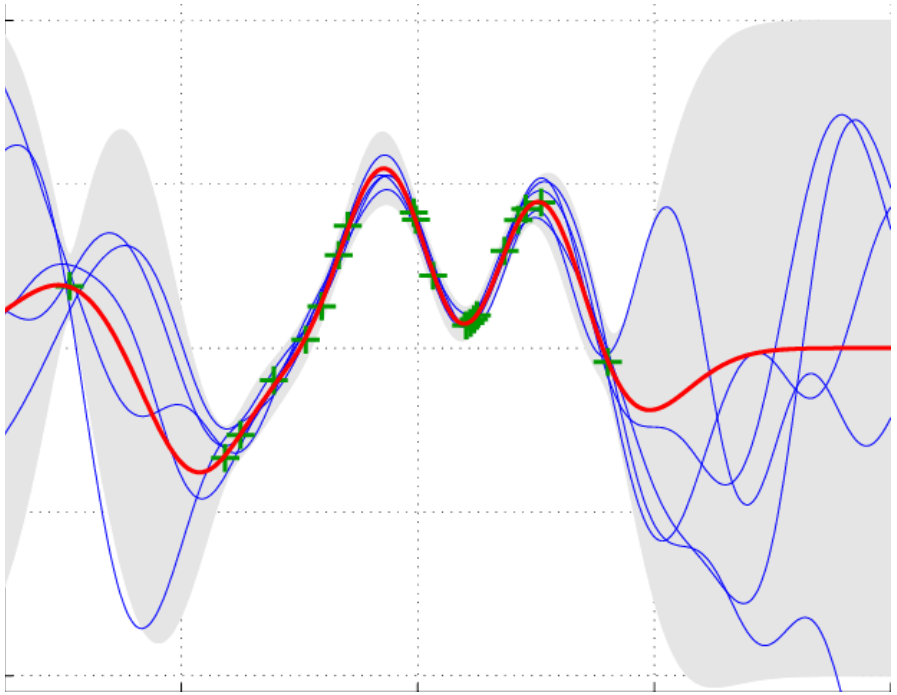
Limitations of Gaussian Processes

Extrapolation does really poorly with Gaussian processes (depends on length scale of kernel)

They are not sparse (they use the whole samples/features information to perform the prediction

They lose efficiency in high-dimensional spaces (features > a few dozen)

Picking and tuning kernel is tricky



Perez-Cruz et al Gaussian Processes for Nonlinear Signal Processing: An Overview of Recent Advances IEEE

Bayesian Optimization