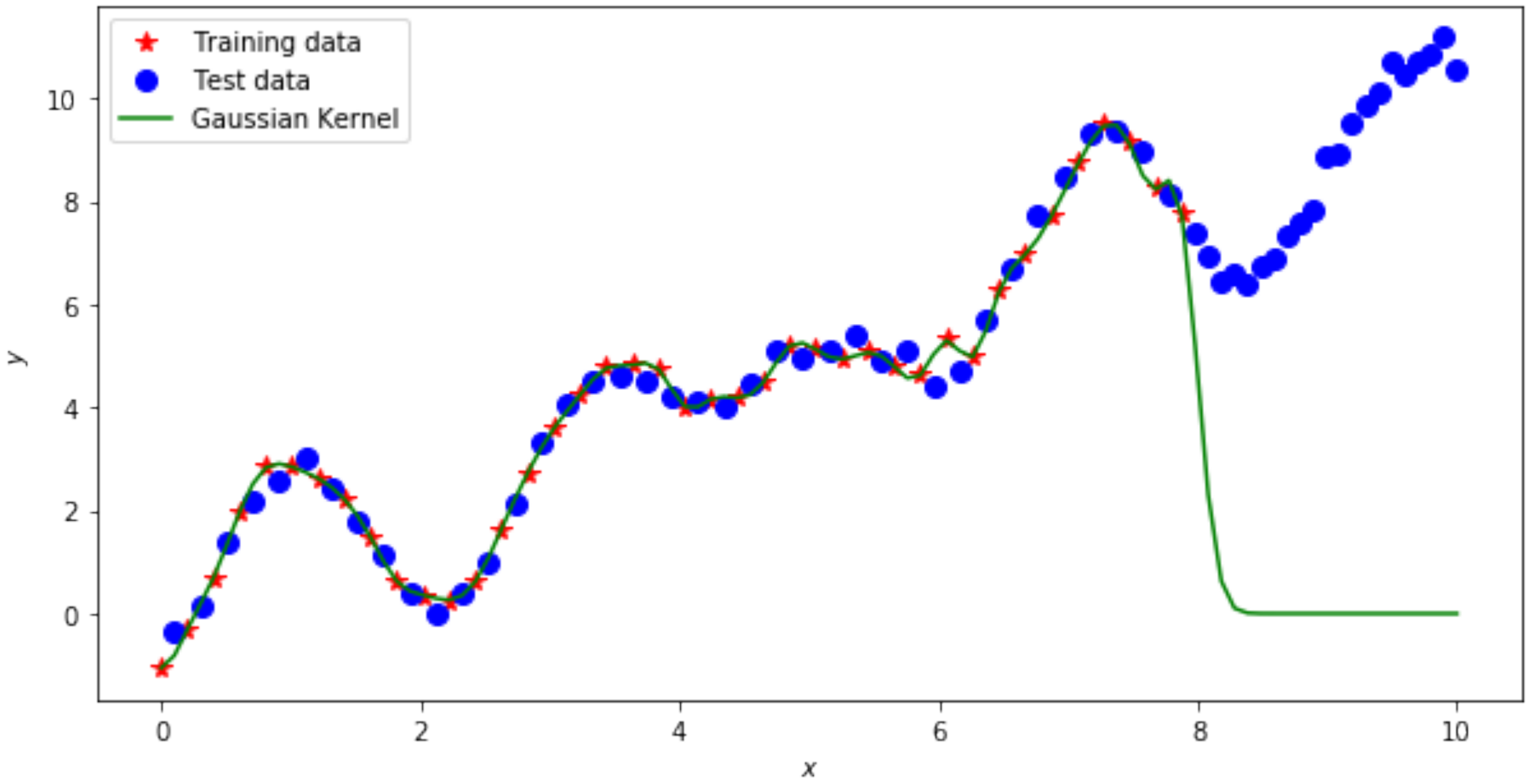
Nonparametric Learning Part II

STAT 37710 / CMSC 35300
Rebecca Willett and Yuxin Chen

Example: kernel ridge regression



How should we model predictive uncertainty?

A Bayesian approach to regression

Goal: find a mapping from input x to output y based on samples

Bayesian approach:

- Model the (x, y) relationship as being of the form

$$y = f(x) + z$$

  where $z \sim \mathcal{N}(0, \sigma^2)$ is random noise
- Place a prior distribution on f to model smoothness

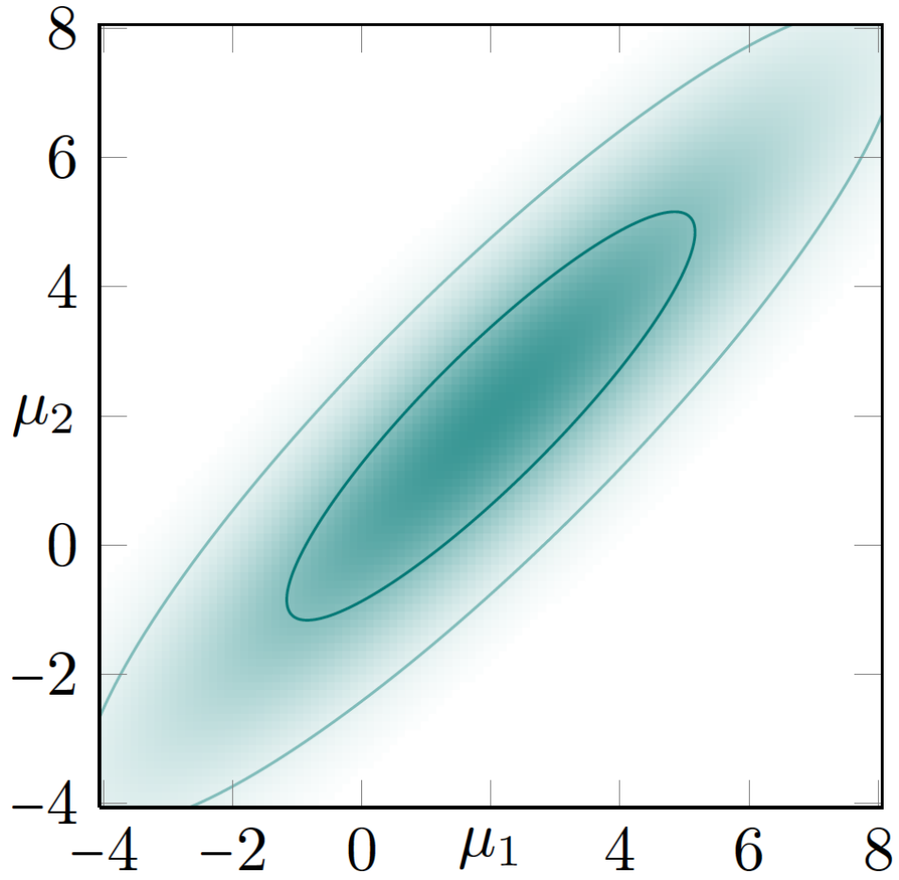
A distribution on f is a distribution over functions (recall stochastic processes)

- For any point x, the function value $f(x)$ is a random variable

The multivariate Gaussian distribution

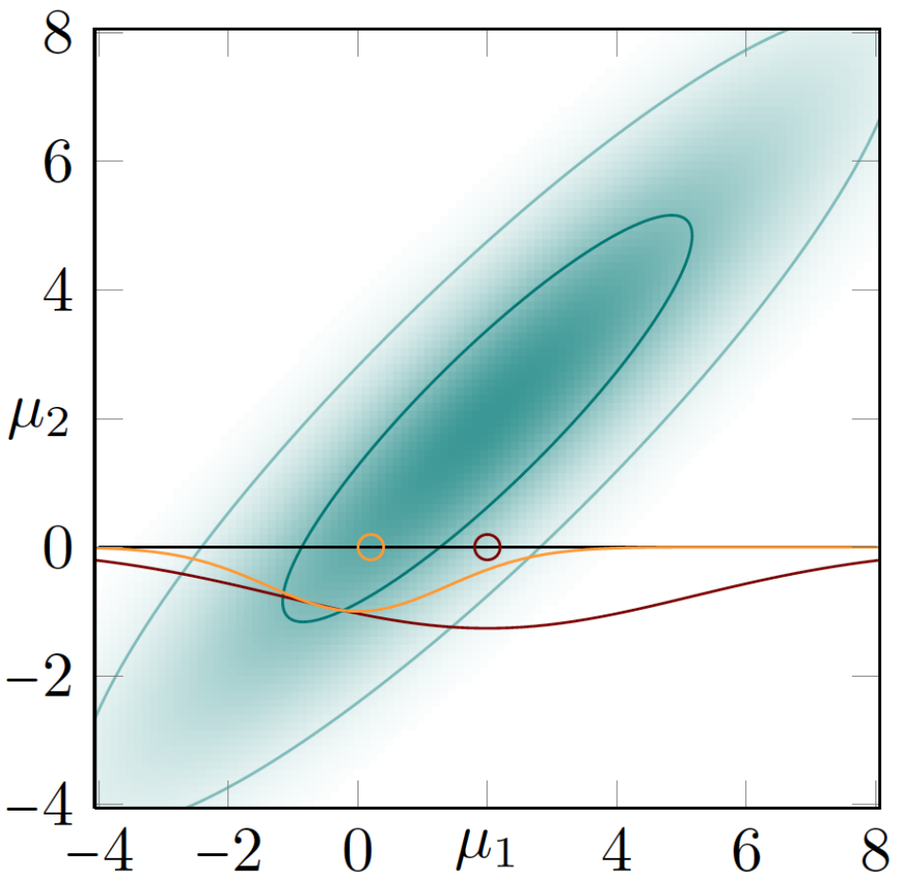
$$x, \mu \in \mathbb{R}^N, \Sigma \in \mathbb{R}^{N \times N}$$

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right]$$



Closure under the rules of probability

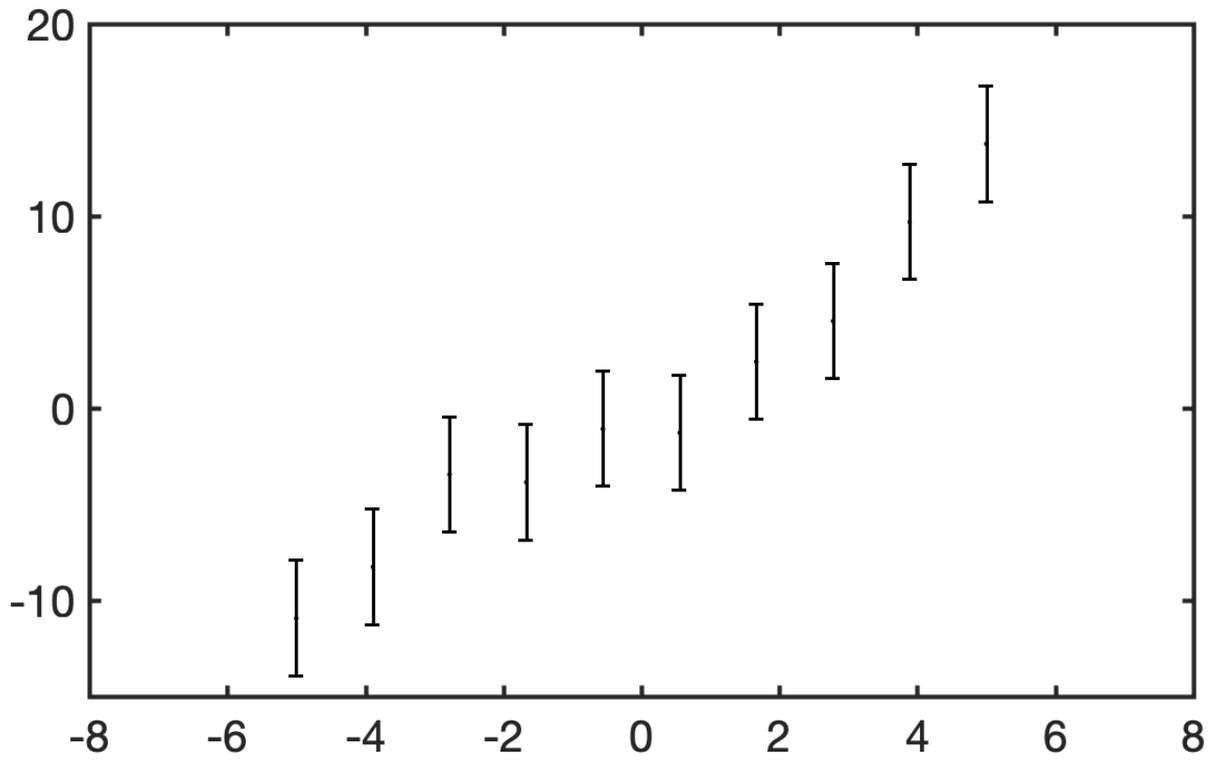
$$p(x \mid y) = \frac{p(x,y)}{p(y)} = \mathcal{N}\left(x; \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right)$$



- Product rule of probability
- Cuts through Gaussians are Gaussians
- Closed under the rules of probability

What can we do with this?

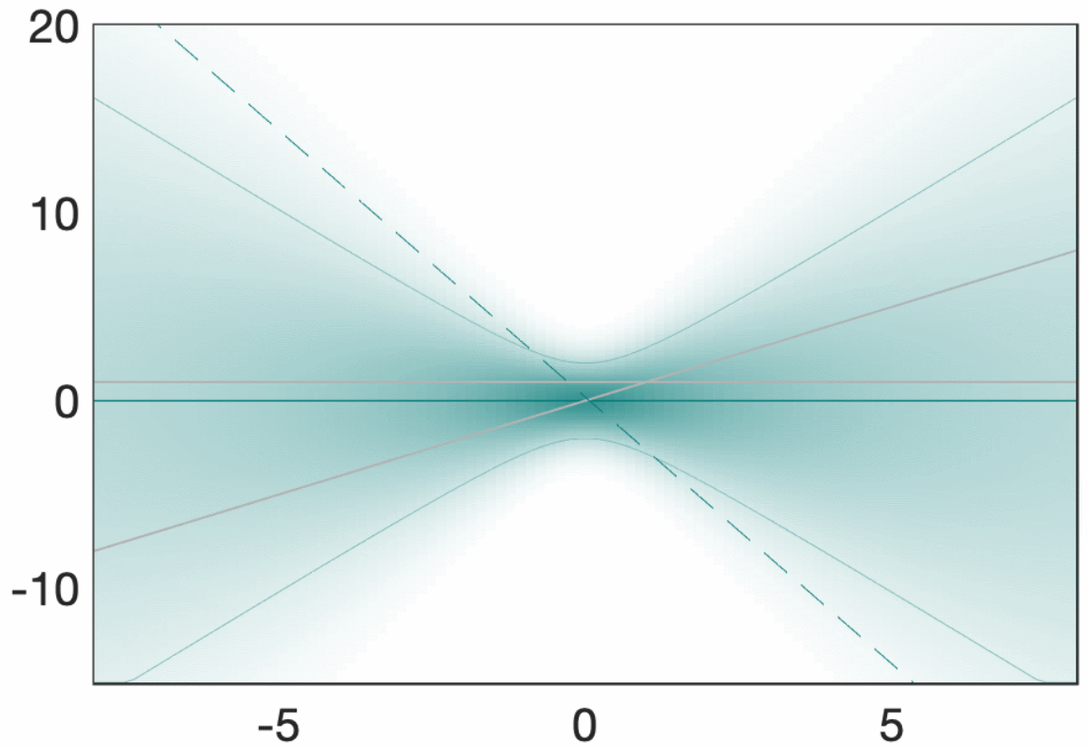
Given $y, p(y \mid f)$, what's f ?



(Parametric) Bayesian linear regression: prior

$$f(x) = w_1 + w_2 x = \phi_x^\top w \qquad p(w) = \mathcal{N}(w; \mu, \Sigma)$$

$$\phi_x = \begin{pmatrix} 1 \\ x \end{pmatrix} \qquad p(f) = \mathcal{N}(f; \phi_x^\top \mu, \phi_x^\top \Sigma \phi_x)$$

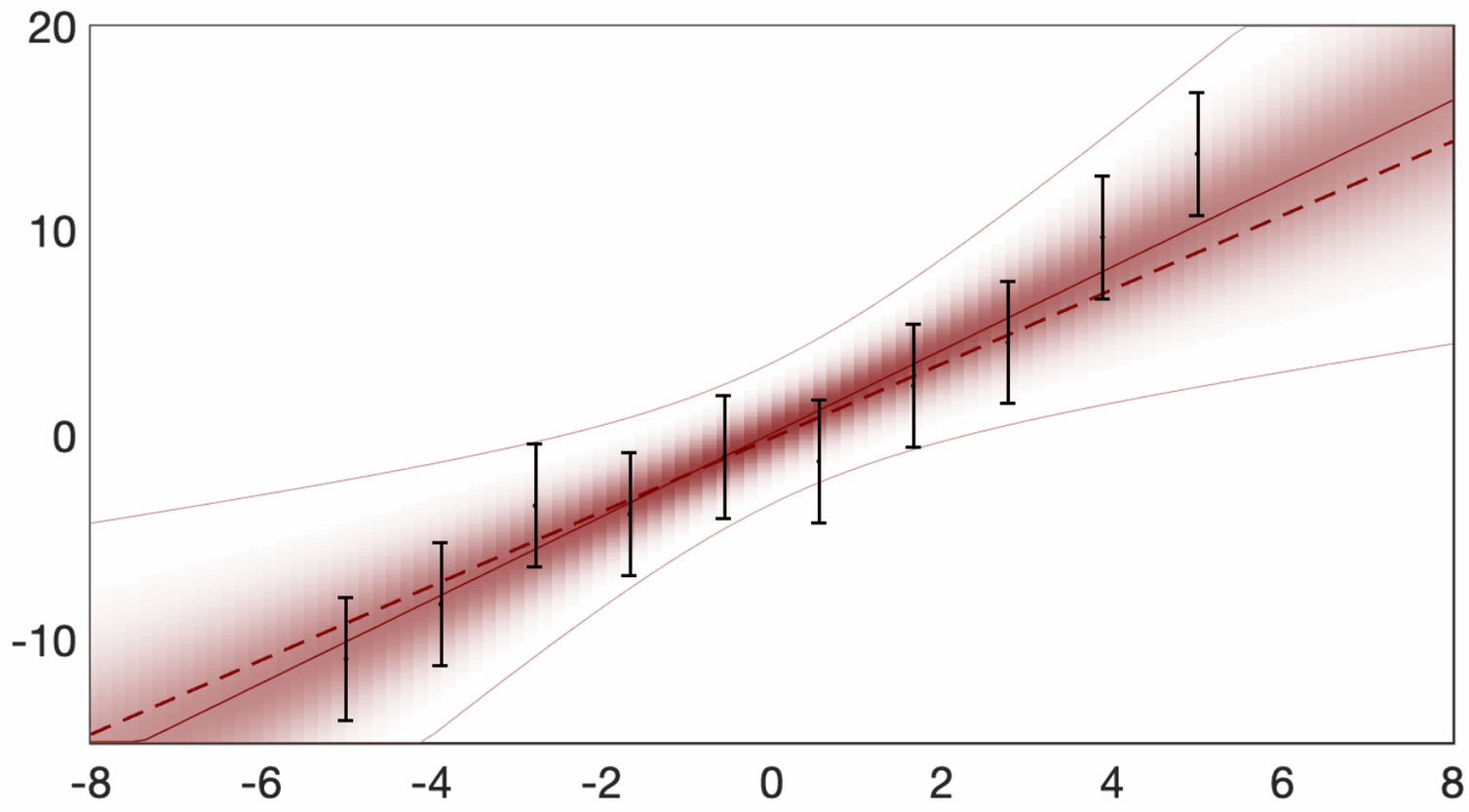


http://mlss.tuebingen.mpg.de/2013/Hennig_2013_Animating_Samples_from_Gaussian_Distributions.pdf

(Parametric) Bayesian linear regression: likelihood and posterior

$$p(y \mid w, \phi_X) = \mathcal{N}(y; \phi_x^\top \mu, \sigma^2 I)$$

$$p(w \mid y, \phi_X) = \mathcal{N}(w; \mu + \Sigma\phi_X(\phi_X{}^\top \Sigma\phi_X + \sigma^2 I)^{-1}(y - \phi_X^\top \mu),$$
$$\Sigma - \Sigma\phi_X(\phi_X{}^\top \Sigma\phi_X + \sigma^2 I)^{-1}\phi_X{}^\top \Sigma\phi_x)$$

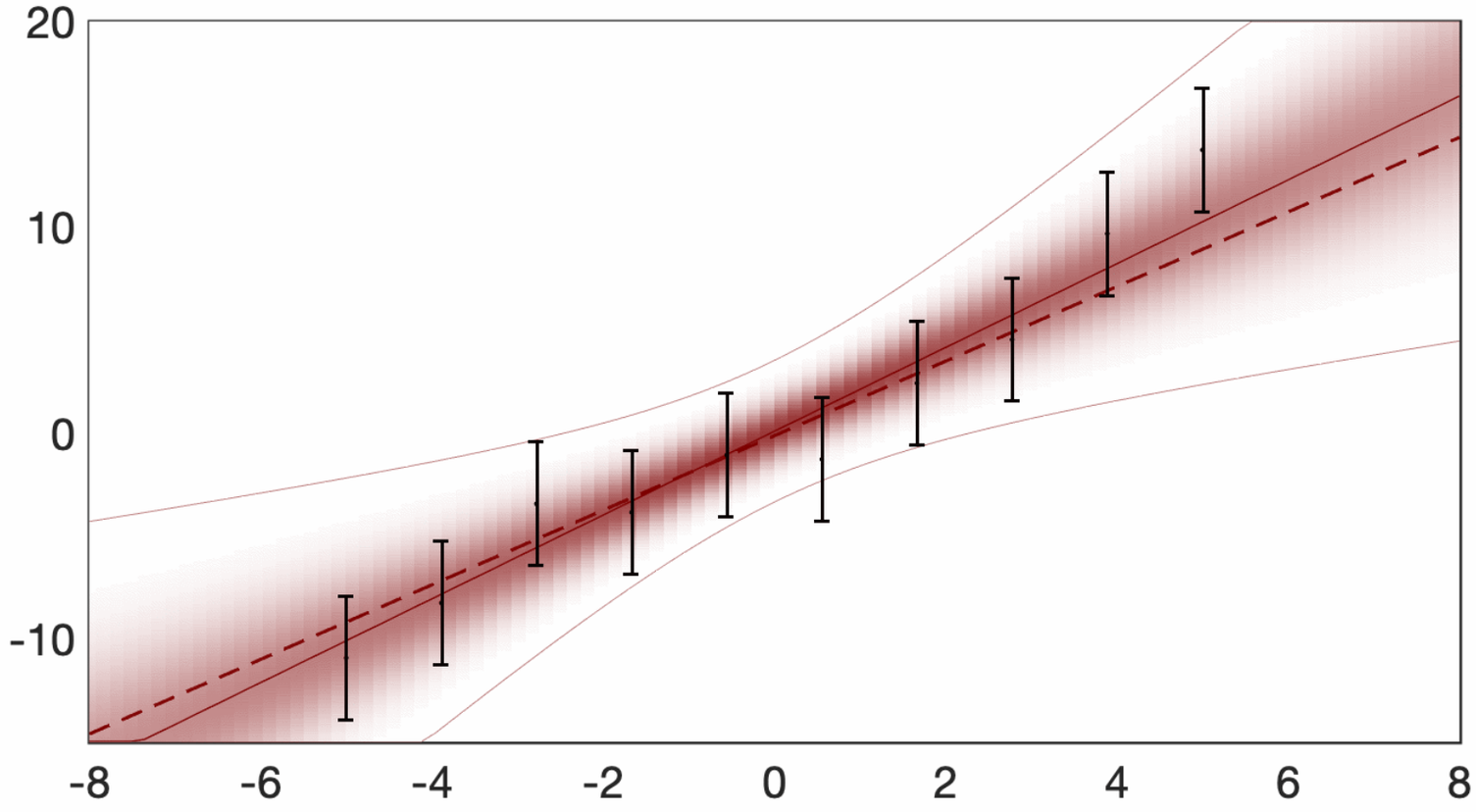


(Parametric) Bayesian linear regression: likelihood and posterior

$$p(y \mid w, \phi_X) = \mathcal{N}(y; \phi_x^\top \mu, \sigma^2 I)$$

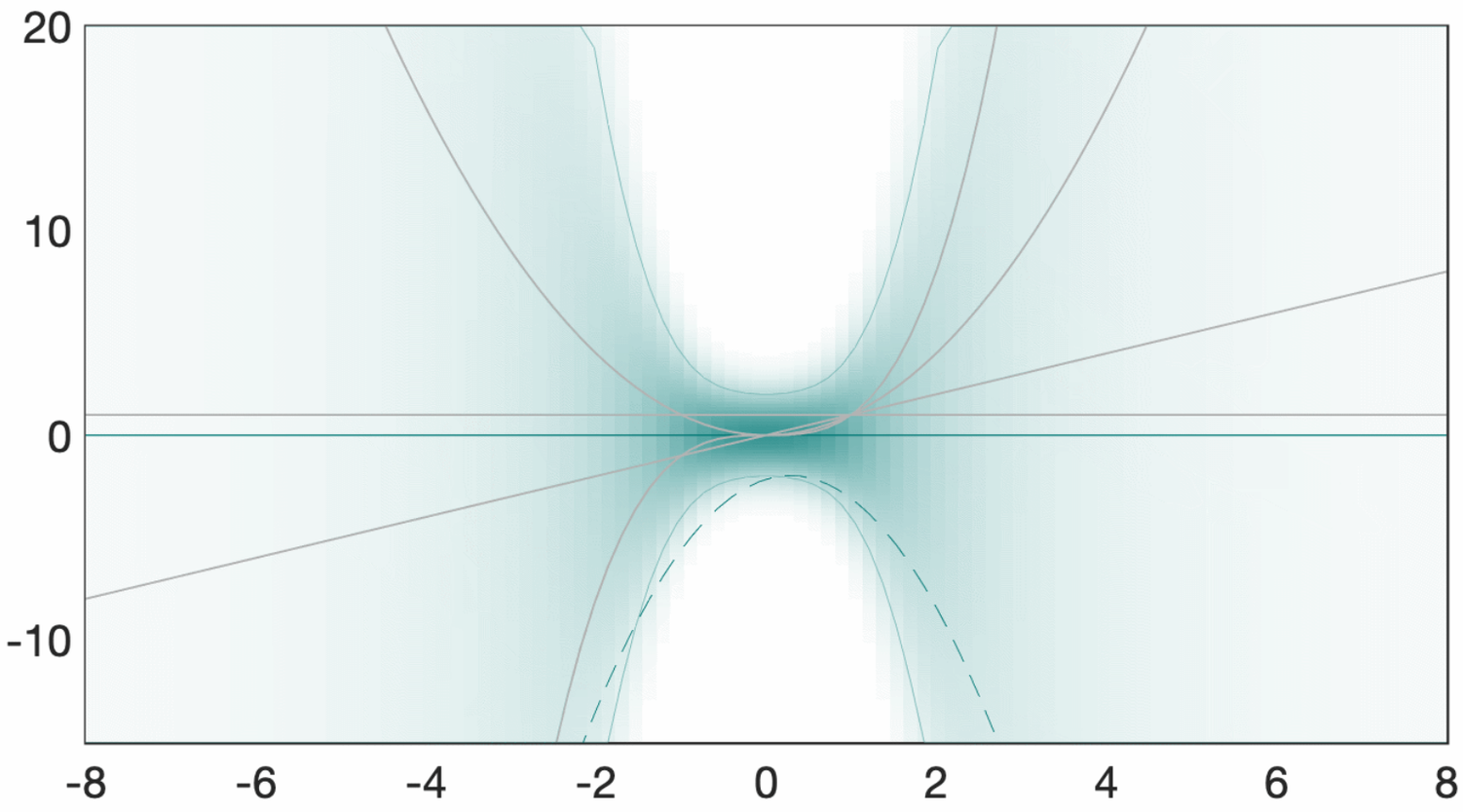
$$p(w \mid y, \phi_X) = \mathcal{N}(w; \mu + \Sigma\phi_X(\phi_X{}^\top\Sigma\phi_X + \sigma^2 I)^{-1}(y - \phi_X^\top\mu),$$
$$\Sigma - \Sigma\phi_X(\phi_X{}^\top\Sigma\phi_X + \sigma^2 I)^{-1}\phi_X{}^\top\Sigma\phi_x)$$

$$p(f_x \mid y, \phi_X) = \mathcal{N}(f_x; \phi_x{}^\top\mu + \phi_x{}^\top\Sigma\phi_X(\phi_X{}^\top\Sigma\phi_X + \sigma^2 I)^{-1}(y - \phi_X^\top\mu),$$
$$\phi_x{}^\top\Sigma\phi_x - \phi_x{}^\top\Sigma\phi_X(\phi_X{}^\top\Sigma\phi_X + \sigma^2 I)^{-1}\phi_X{}^\top\Sigma\phi_x)$$



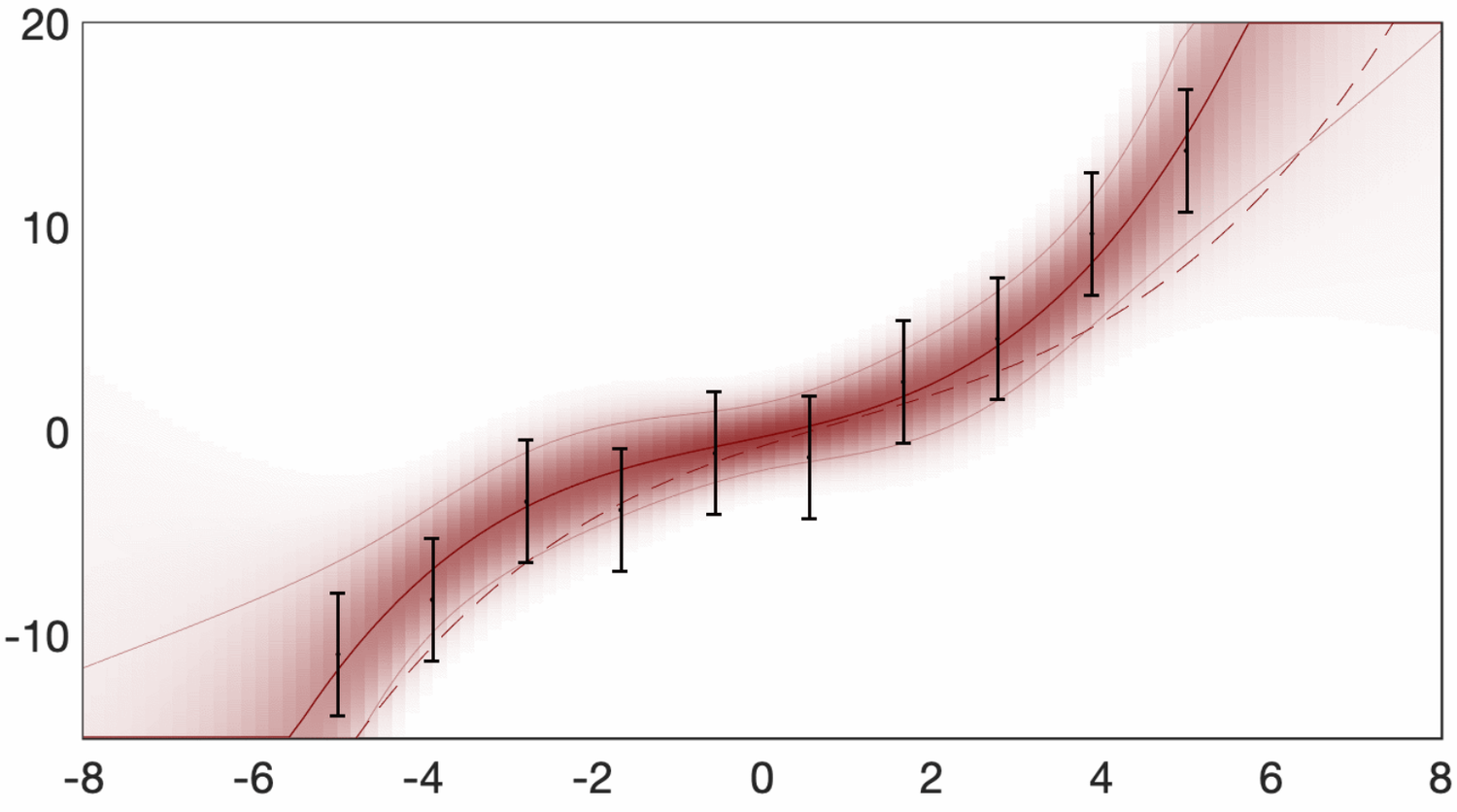
Polynomial (cubic) regression

$$f(x) = \phi(x)^\top w \qquad \phi(x) = \left(1, x, x^2, x^3\right)$$



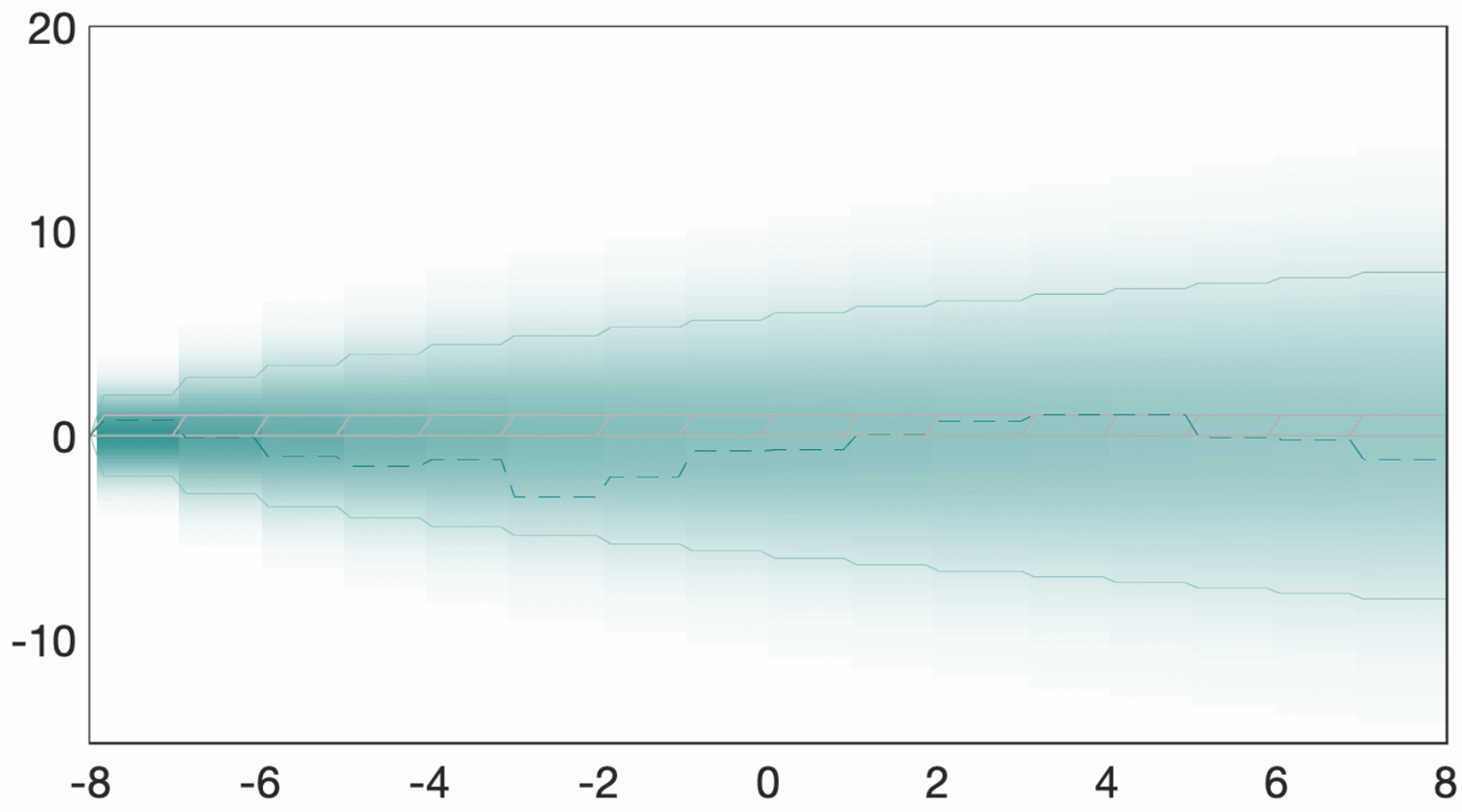
Polynomial (cubic) regression

$$f(x) = \phi(x)^\top w \qquad \phi(x) = \left(1, x, x^2, x^3\right)$$



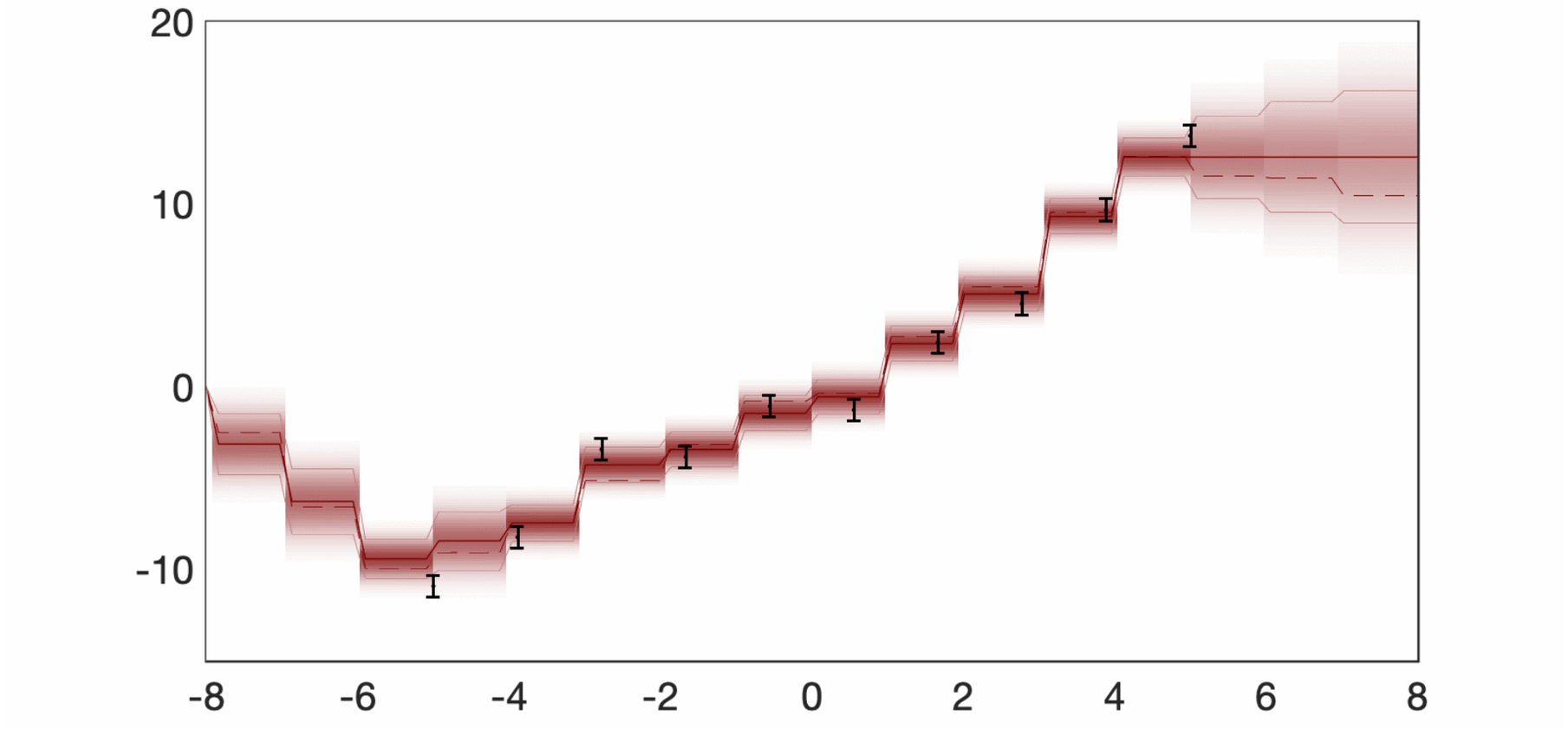
Step regression

$$\phi(x) = (\theta(x-8), \theta(8-x), \theta(x-7), \theta(7-x), \dots)^\top$$



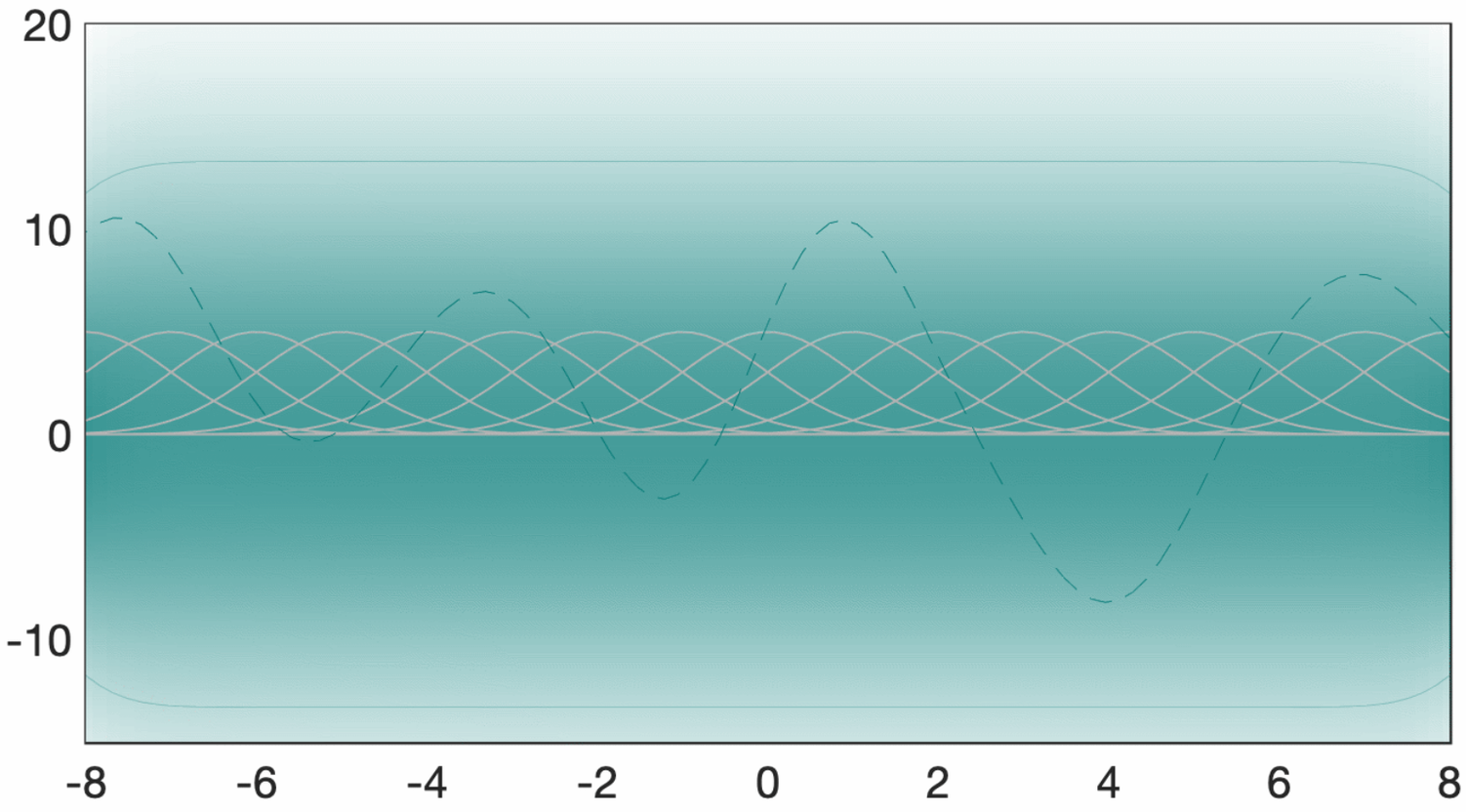
Step regression

$$\phi(x) = (\theta(x-8), \theta(8-x), \theta(x-7), \theta(7-x), \dots)^\top$$



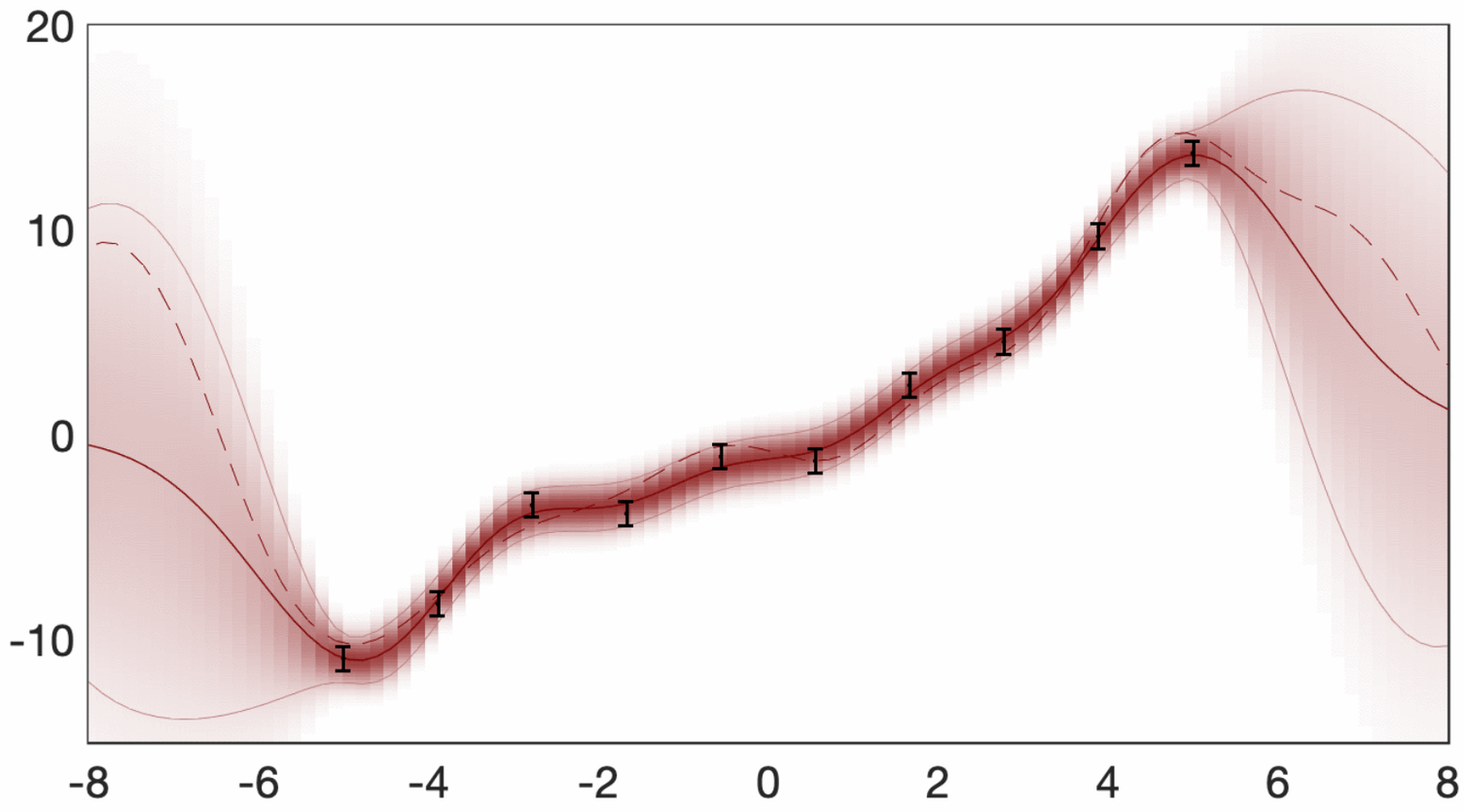
Bell curve regression

$$\phi(x) = \left(e^{-\frac{1}{2}(x-8)^2}, e^{-\frac{1}{2}(x-7)^2}, e^{-\frac{1}{2}(x-6)^2}, \dots \right)^\top$$



Bell curve regression

$$\phi(x) = \left(e^{-\frac{1}{2}(x-8)^2}, e^{-\frac{1}{2}(x-7)^2}, e^{-\frac{1}{2}(x-6)^2}, \dots \right)^\top$$



How many features should we use?

$$p(f_x \mid y, \phi_X) = \mathcal{N}(f_x; \phi_x^\top \mu + \phi_x^\top \Sigma \phi_X (\phi_X^\top \Sigma \phi_X + \sigma^2 I)^{-1}(y - \phi_X^\top \mu),$$
$$\phi_x^\top \Sigma \phi_x - \phi_x^\top \Sigma \phi_X (\phi_X^\top \Sigma \phi_X + \sigma^2 I)^{-1} \phi_X^\top \Sigma \phi_x)$$

All objects involving ϕ are of the form

- $\phi^\top \mu \rightarrow$ the *mean function*
- $\phi^\top \Sigma \phi \rightarrow$ the *kernel*

Once these are known, cost is independent of the number of features

How many features should we use?

Let $\phi_\ell(x) = \exp\left(-\frac{(x - c_\ell)^2}{2\lambda^2}\right)$

$\Sigma = \frac{\sigma^2(c_{\max} - c_{\min})}{F} I$

params/"features"

Now the *kernel* term $\phi^\top \Sigma \phi$ becomes

$$\phi(x_i)^\top \Sigma \phi(x_j)$$

$$= \frac{\sigma^2(c_{\max} - c_{\min})}{F} \sum_{\ell=1}^{F} \exp\left(-\frac{(x_i - c_\ell)^2}{2\lambda^2}\right) \exp\left(-\frac{(x_j - c_\ell)^2}{2\lambda^2}\right)$$

$$= \frac{\sigma^2(c_{\max} - c_{\min})}{F} \exp\left(-\frac{(x_i - x_j)^2}{4\lambda^2}\right) \sum_{\ell}^{F} \exp\left(-\frac{(c_\ell - \frac{1}{2}(x_i + x_j))^2}{\lambda^2}\right)$$

Kernelization to infinitely many features

MacKay, 1998

Let $\phi_\ell(x) = \exp\left(-\frac{(x - c_\ell)^2}{2\lambda^2}\right)$ $\quad$ $\Sigma = \frac{\sigma^2(c_{\max} - c_{\min})}{F} I$

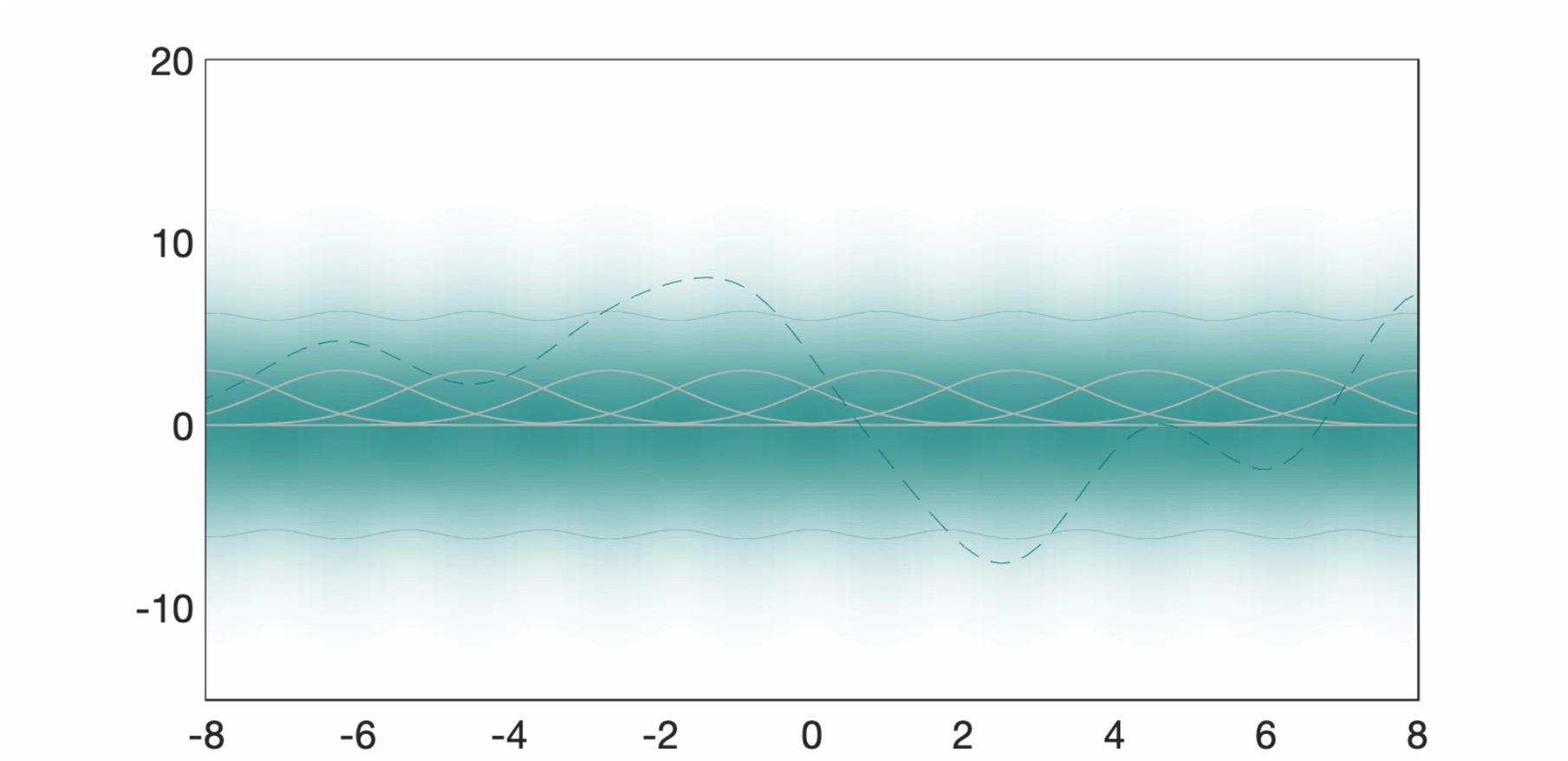
params/“features”

Increase F, and let $c_{\min} \to -\infty$, $c_{\max} \to \infty$

$$\phi(x_i)^\top \Sigma \phi(x_j) \to \sqrt{2\pi}\lambda\sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{4\lambda^2}\right)$$

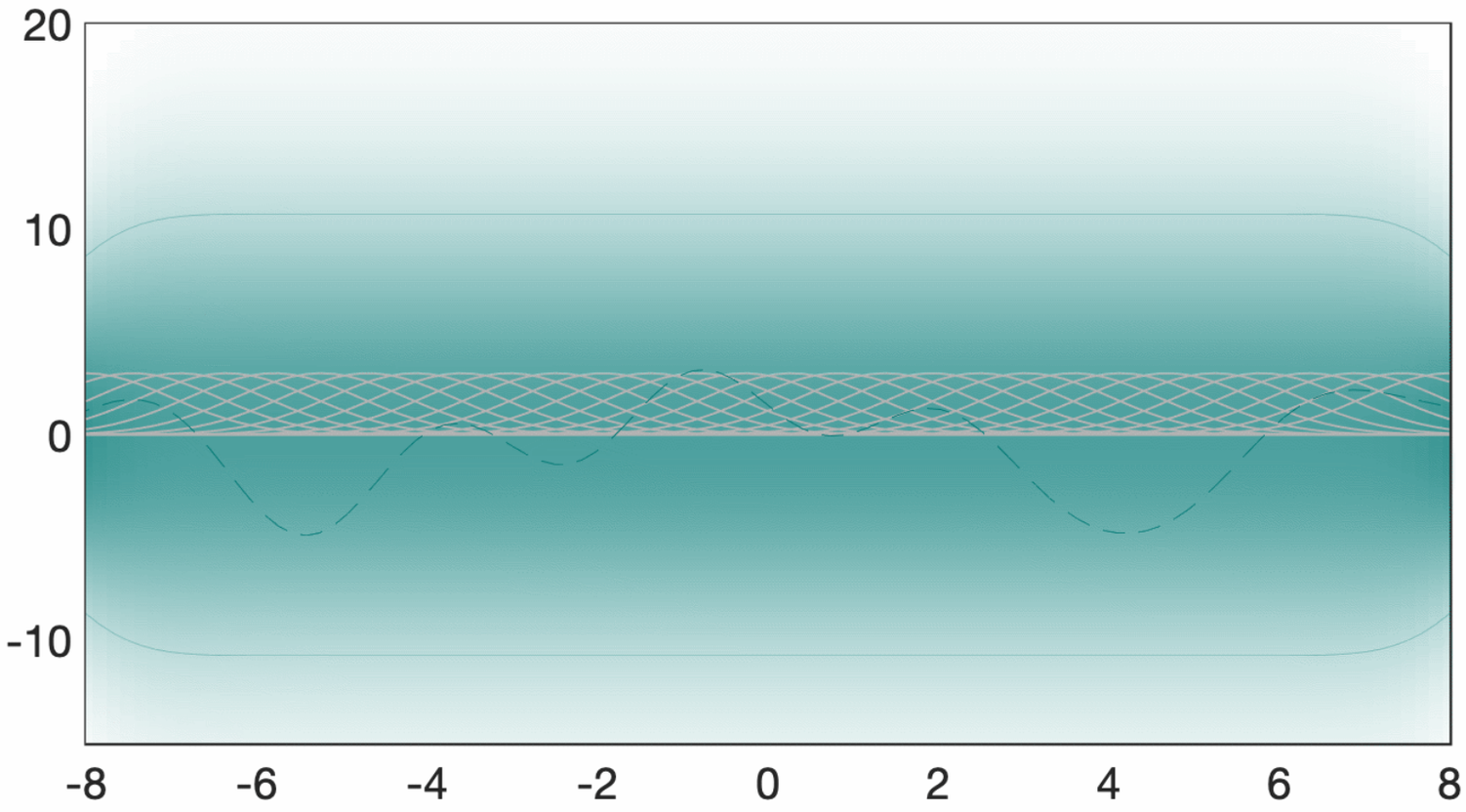
Squared exponential (SE)

$$\phi(x) = \overbrace{\left(e^{-\frac{1}{2}(x-8)^2}, \ldots, e^{-\frac{1}{2}(x+8)^2}\right)}^{10}$$



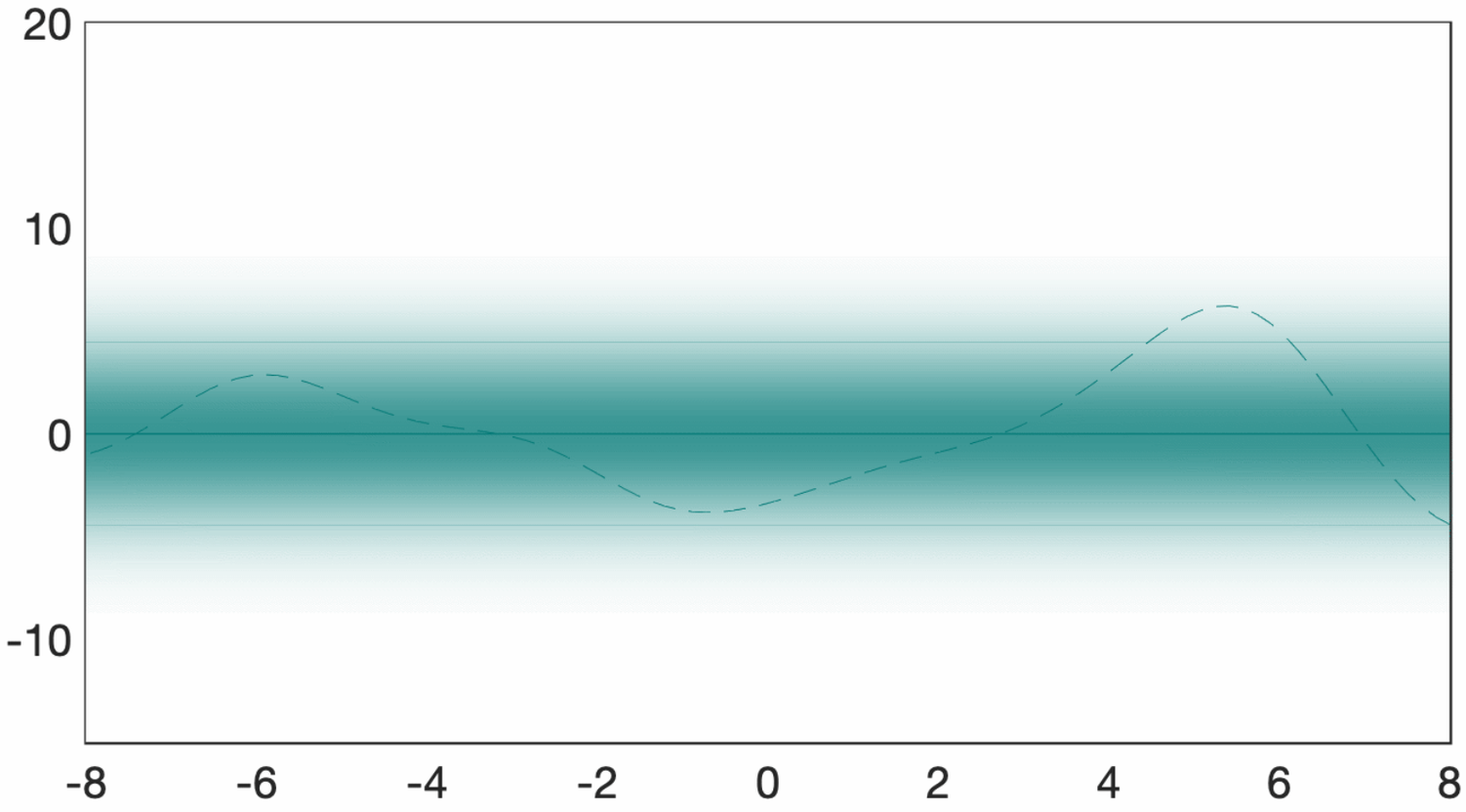
Squared exponential (SE)

$$\phi(x) = \overbrace{\left(e^{-\frac{1}{2}(x-8)^2}, \ldots, e^{-\frac{1}{2}(x+8)^2}\right)}^{30}$$



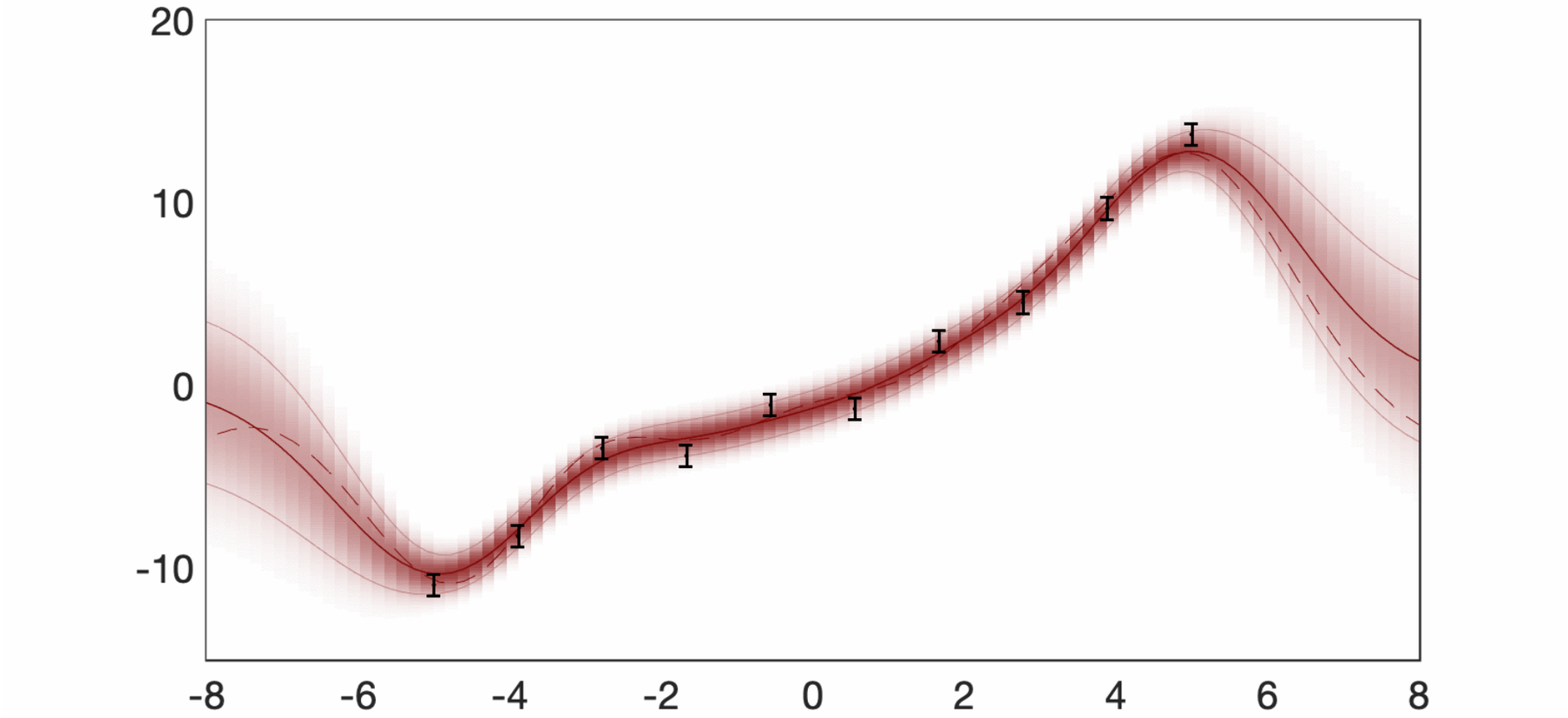
Squared exponential (SE)

$$k(x_i, x_j) = 5 \exp\left(-\frac{(x_i - x_j)^2}{4} \right)$$



Squared exponential (SE)

$$k(x_i, x_j) = 5 \exp\left(-\frac{(x_i - x_j)^2}{4} \right)$$



Gaussian process (GP)

A Gaussian process

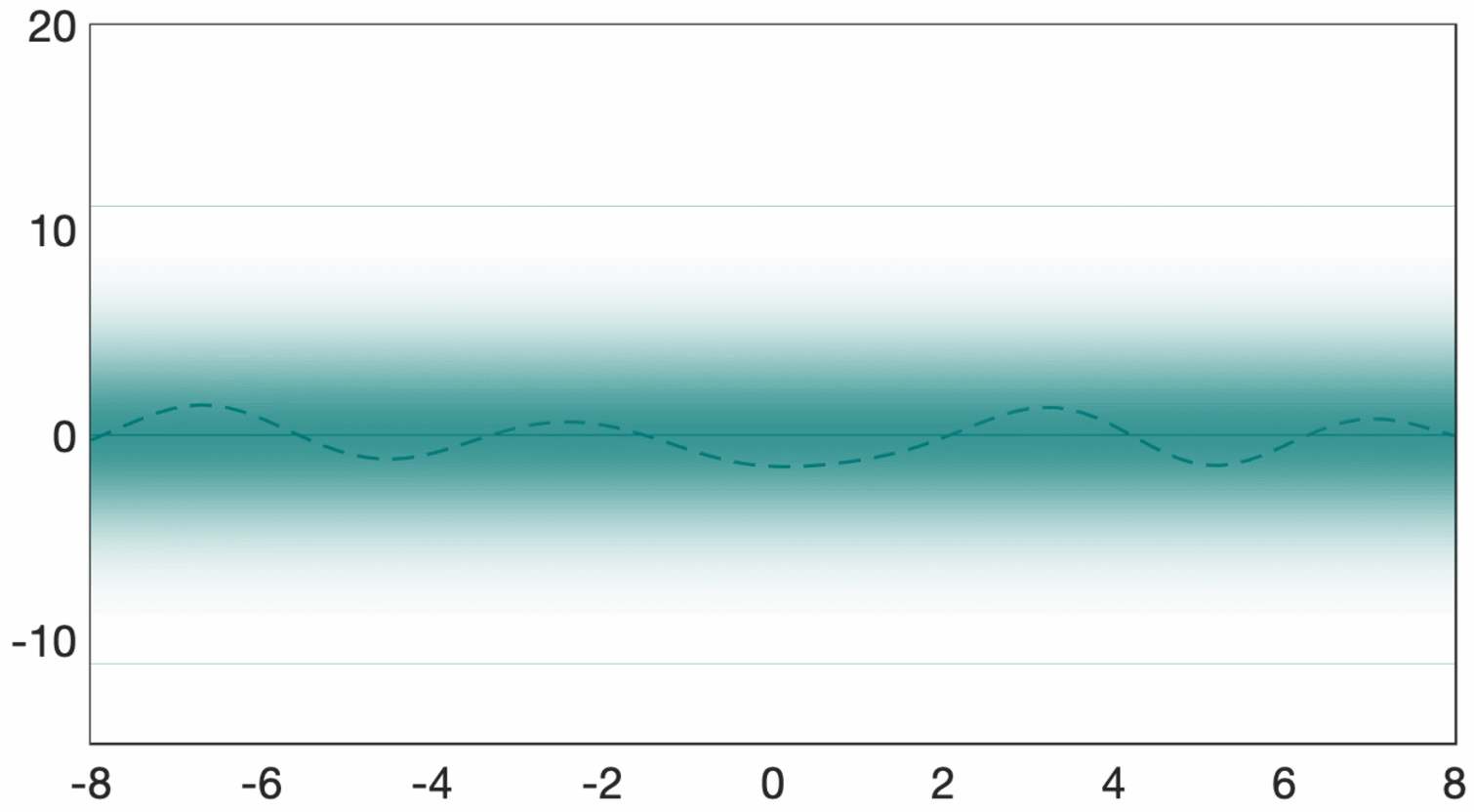
$$p(f) = \mathcal{GP}(f; \mu, k)$$

is a probability distribution over the function *f*, such that every finite restriction to function values $f_X = [f_{x_1}, \ldots, f_{x_N}]$ is a Gaussian distribution $p(f_X) = N(f_X; \mu_X, k_{XX})$.

Gaussian process posterior

$$f \sim \mathcal{GP}(\mu, k(x, x')) \qquad y = f(x) + z, \text{ where } z \sim \mathcal{N}(0, \sigma)$$

$$p(f_x \mid y, X) = \mathcal{N}(f_x; \mu_x + k_{Xx}^\top (k_{XX} + \sigma^2 I)^{-1}(y - \mu_x),$$
$$k_{xx} - k_{Xx}^\top (k_{XX} + \sigma^2 I)^{-1} k_{Xx})$$



Gaussian process regression

Assume $f \sim \mathcal{GP}(\mu, k(x, x')) \quad y = f(x) + z, \text{ where } z \sim \mathcal{N}(0, \sigma)$

Notations $f_x = f(x), \mu_x = \mu(x), k_{X,X} := K = \left[k(x_i, x_j)\right]_{i,j \in \{1,\ldots,N\}}$

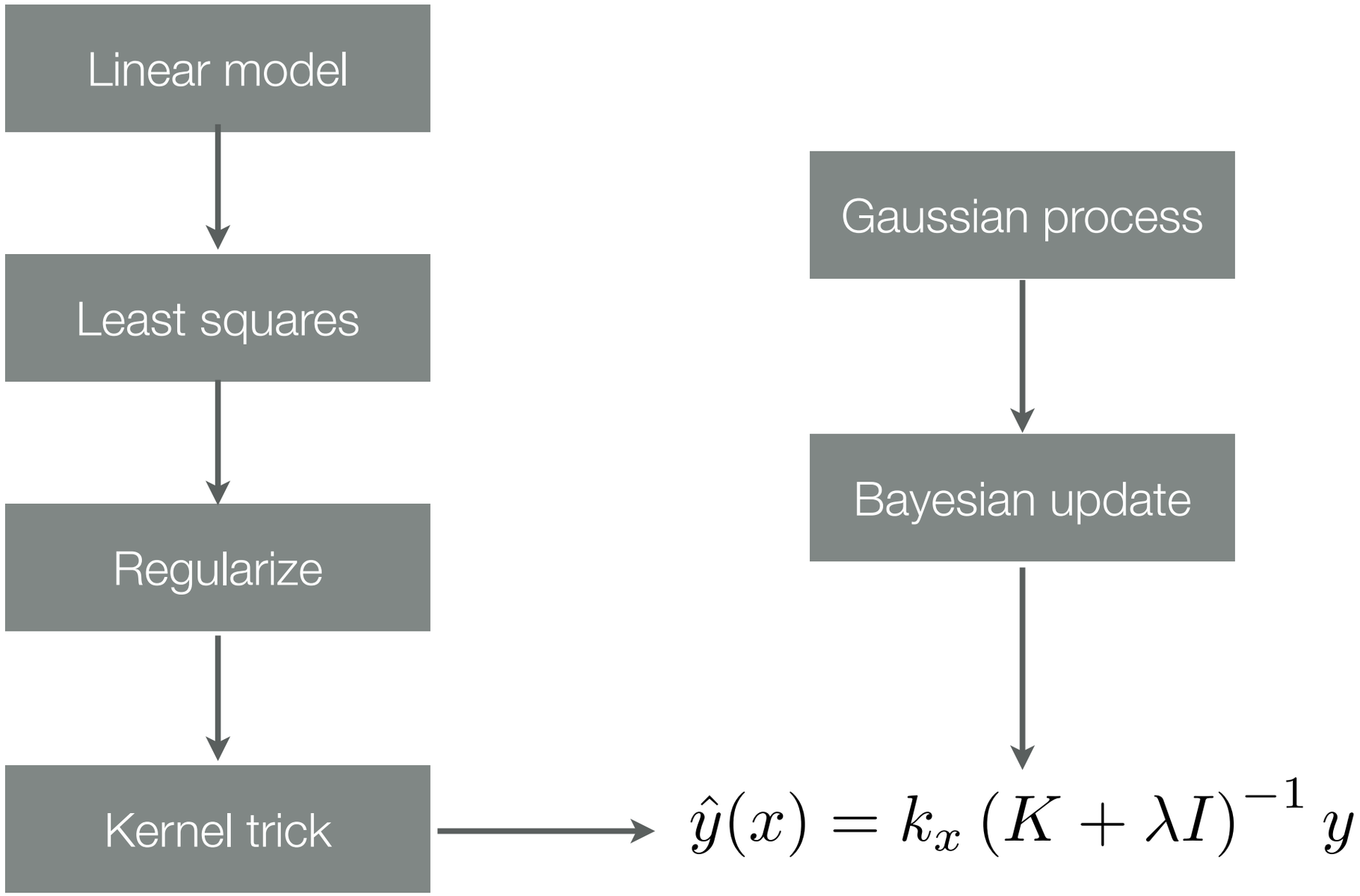
Posterior $p(f_x \mid y, X) = \mathcal{N}(f_x; \mu_x + k_{Xx}^\top (k_{XX} + \sigma^2 I)^{-1}(y - \mu_x),$
$k_{xx} - k_{Xx}^\top (k_{XX} + \sigma^2 I)^{-1} k_{Xx})$

The GP posterior mean is the **regularized least-squares** estimate (a.k.a. kernel ridge regression)

Use the posterior mean $\mu_N(x)$ from N samples for prediction

Moreover, $\sigma^2_N(x)$ gives an **estimate of the uncertainty** in this prediction

Two paths to the same predictor



Advantages and disadvantages for GP

Advantages of GPs:

- Explicit estimates of uncertainty in the prediction
- Wide variety of kernels available for rich modeling of functions
- Often highly effective even when limited data is available

Disadvantages:

- Computing the (exact) posterior takes $O(N^3)$ time
- Choosing a good kernel can be difficult
- Difficulties in scaling to a large input dimension

Examples of commonly used kernels (I)

The choice of kernels is application-dependent

Squared exponential (SE) kernel (aka Gaussian kernel):

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\ell^2}\right)$$

Matérn kernel

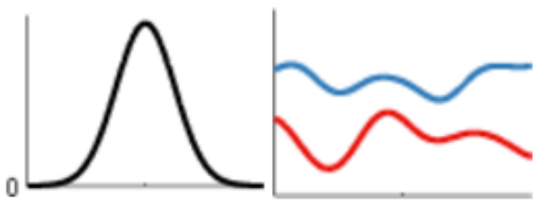
$$k(x_i, x_j) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\,\|x_i - x_j\|_2}{\ell}\right)^{\nu} J_\nu\left(\frac{\sqrt{2\nu}\,\|x_i - x_j\|_2}{\ell}\right)$$

- Γ and J_ν are the Gamma and Bessel functions
- ν is a smoothness parameter (higher = more smooth)
- $\nu \to \infty$ recovers SE kernel

http://www.cs.toronto.edu/~duvenaud/cookbook/index.html

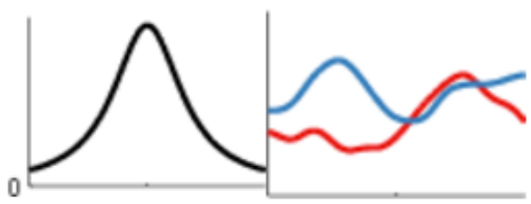
Examples of commonly used kernels (II)

Squared Exponential Kernel



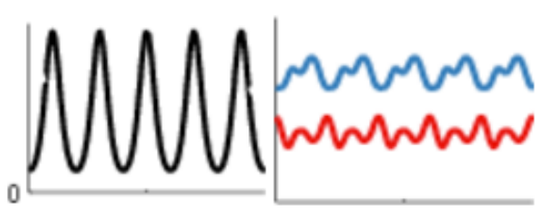
$$k_{\mathrm{SE}}(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$$

Rational Quadratic Kernel



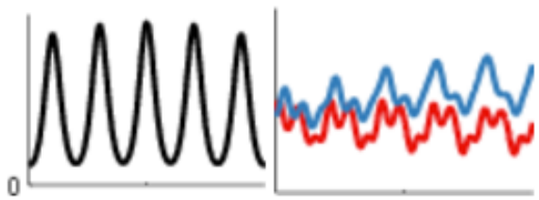
$$k_{\mathrm{RQ}}(x, x') = \sigma^2 \left(1 + \frac{(x-x')^2}{2\alpha\ell^2}\right)^{-\alpha}$$

Periodic Kernel



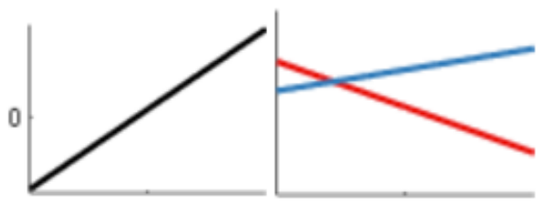
$$k_{\mathrm{Per}}(x, x') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi|x-x'|/p)}{\ell^2}\right)$$

Locally Periodic Kernel



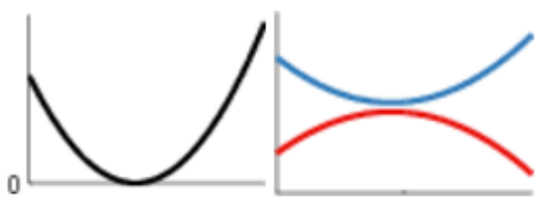
$$k_{\mathrm{LocalPer}}(x, x') = k_{\mathrm{Per}}(x, x') k_{\mathrm{SE}}(x, x')$$

Linear Kernel

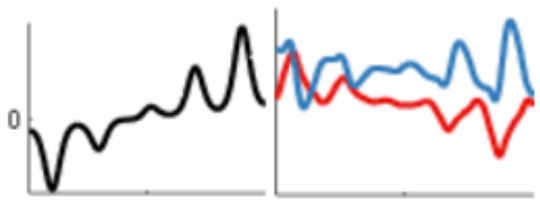


$$k_{\mathrm{Lin}}(x, x') = \sigma_b^2 + \sigma_v^2 (x - c)(x' - c)$$

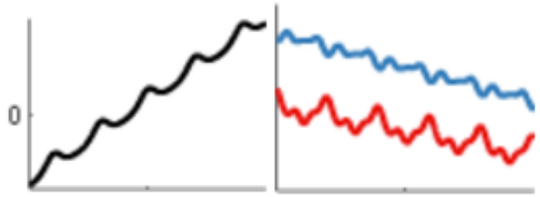
Linear times Linear



Linear times Periodic



Linear plus Periodic



Courtesy [J. Scarlett, 2018]

Principled approach for choosing kernels: Learning from training data

Applications: spatial environmental data

Well-suited to (but not restricted to) modeling "smoothly-varying" data

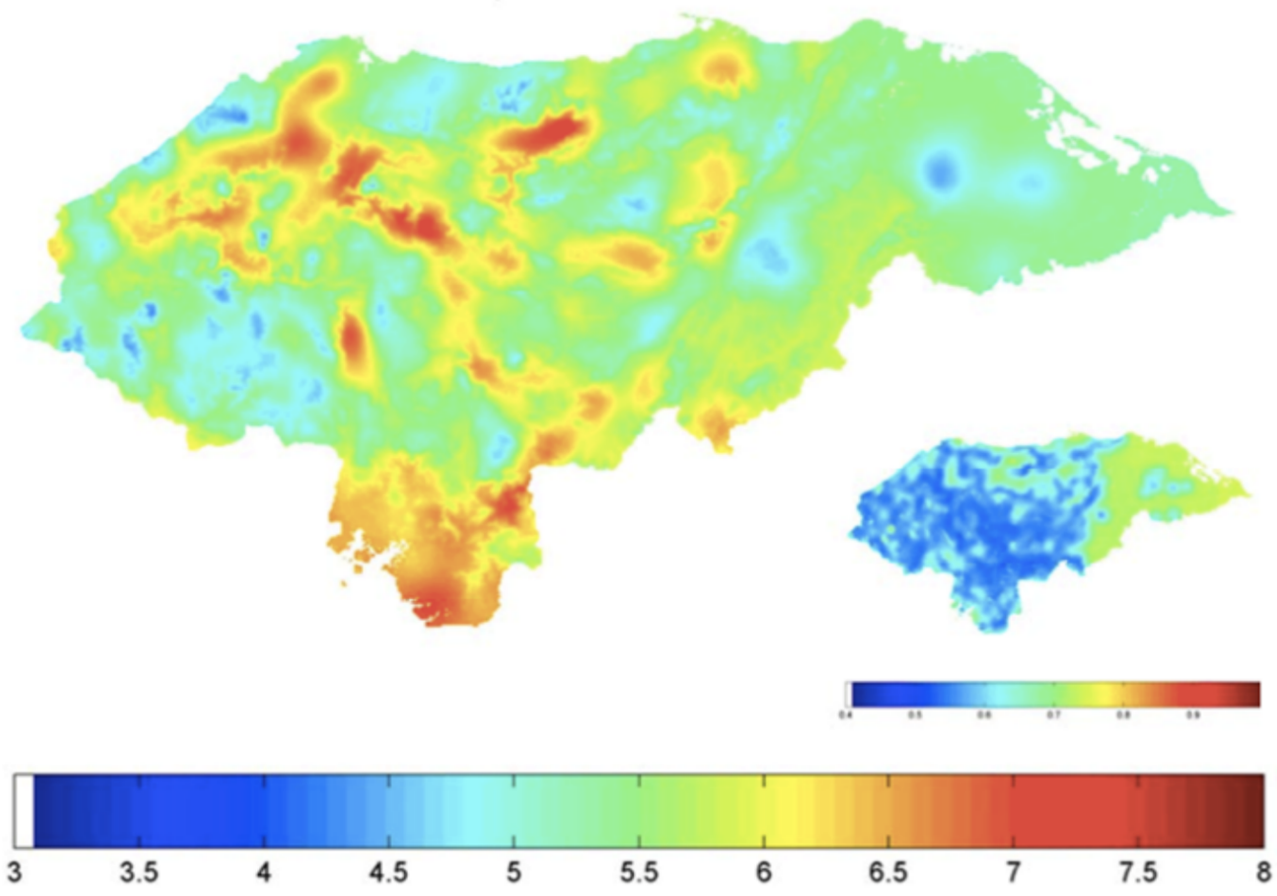
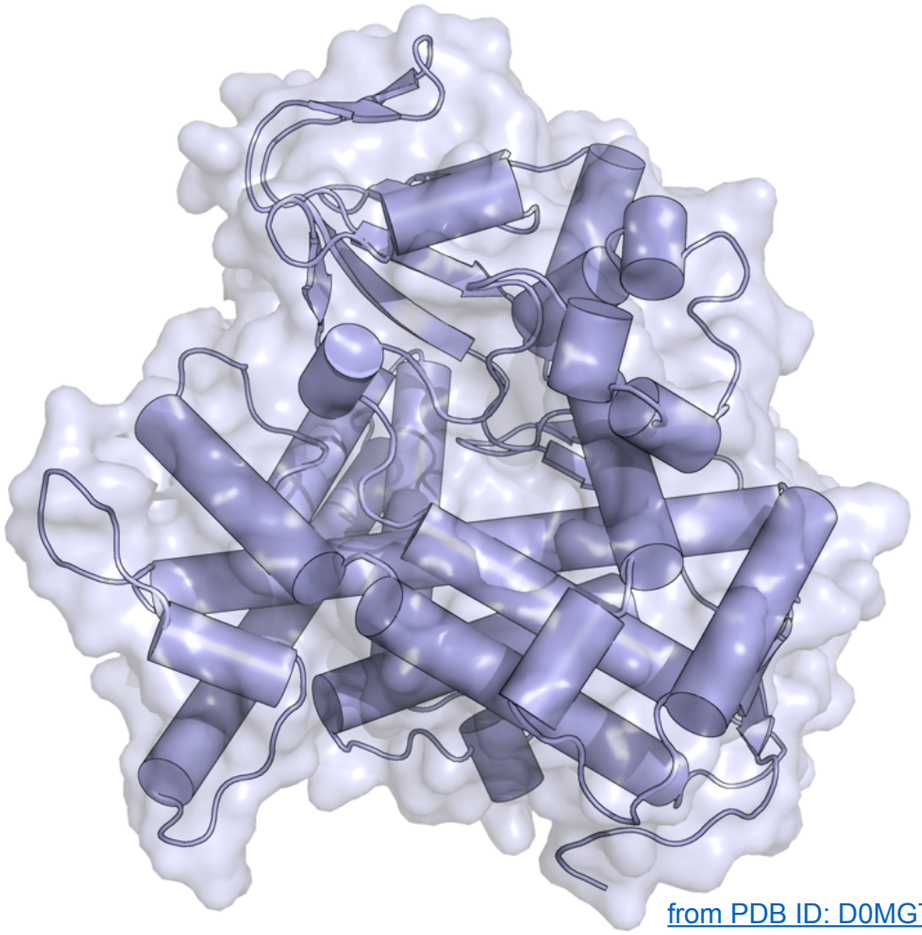


Figure 3. Predicted map of pH in topsoil and 67% confidence interval

[Gonzalez et al., 2007]

Applications: protein engineering



from PDB ID: D0MGT2

Image Credit: Zach Wu
(Arnold Group at Caltech)

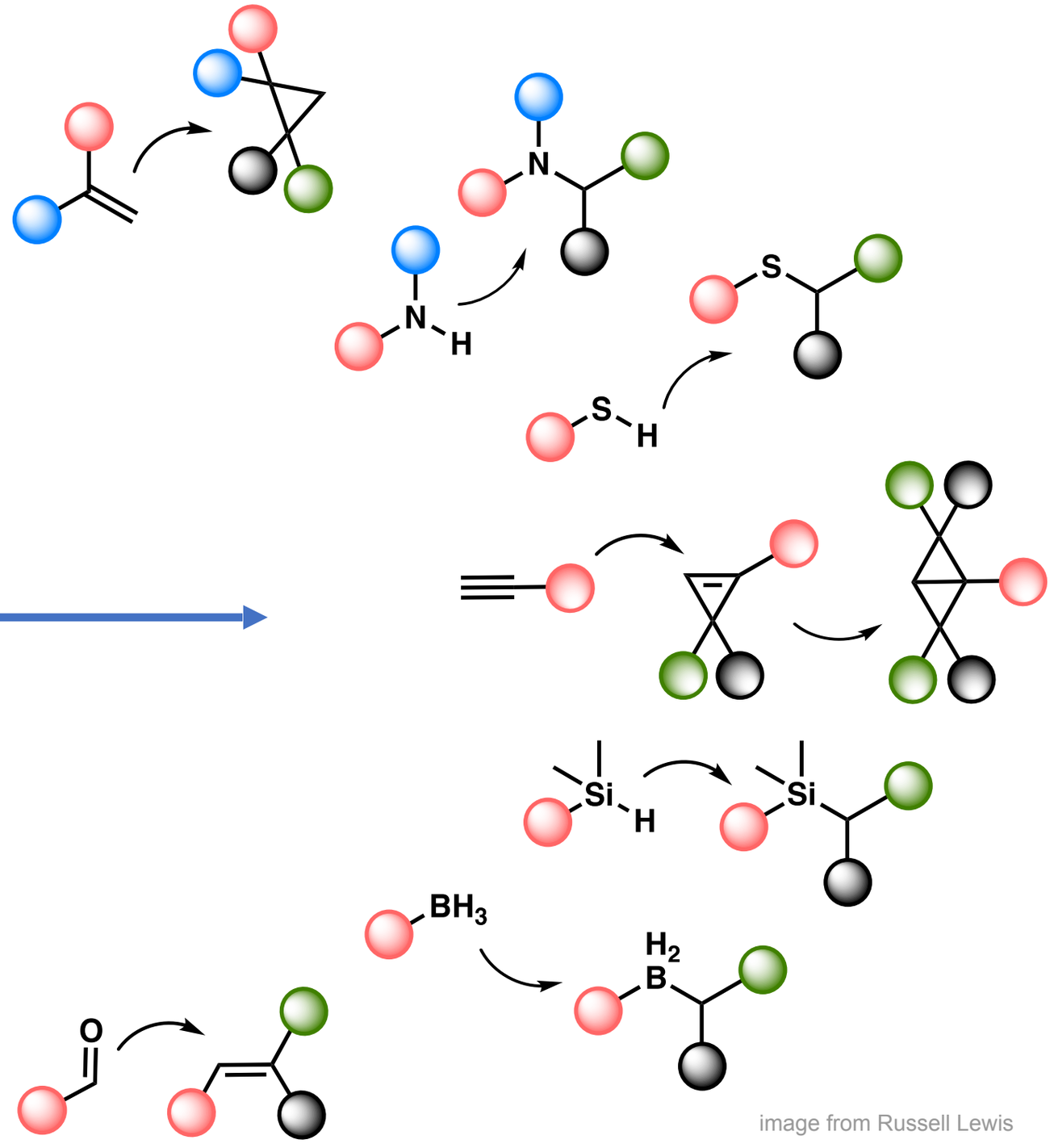
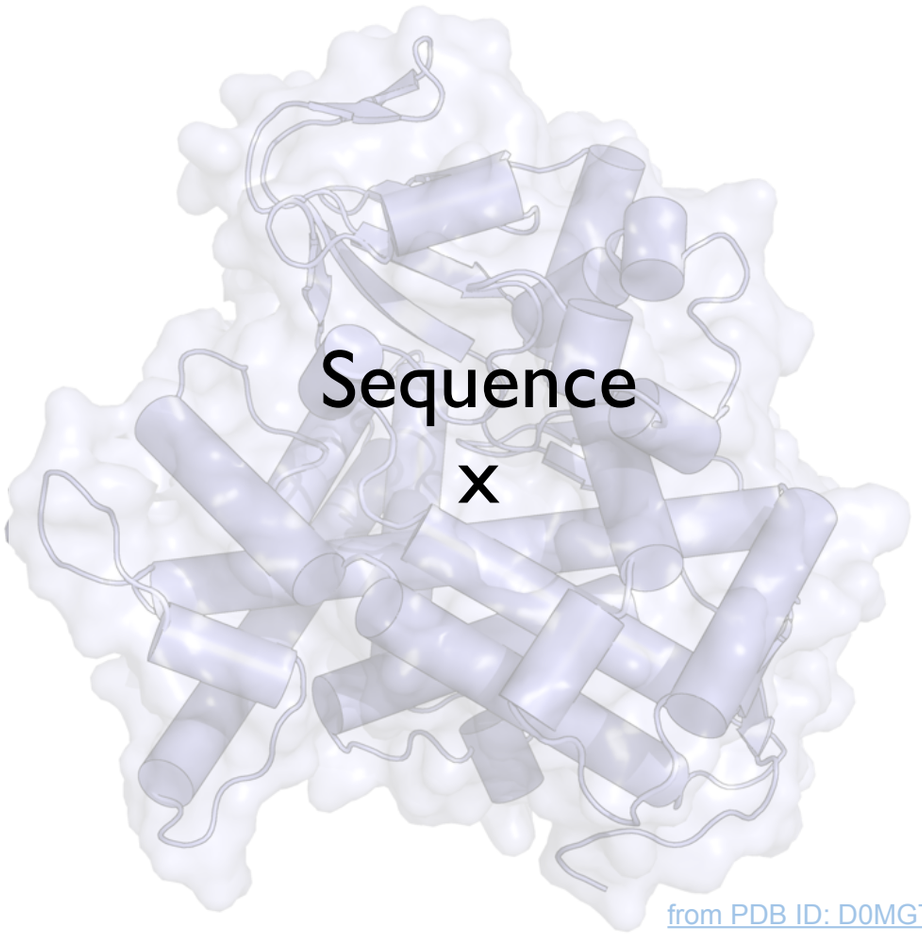


image from Russell Lewis

Application: protein engineering



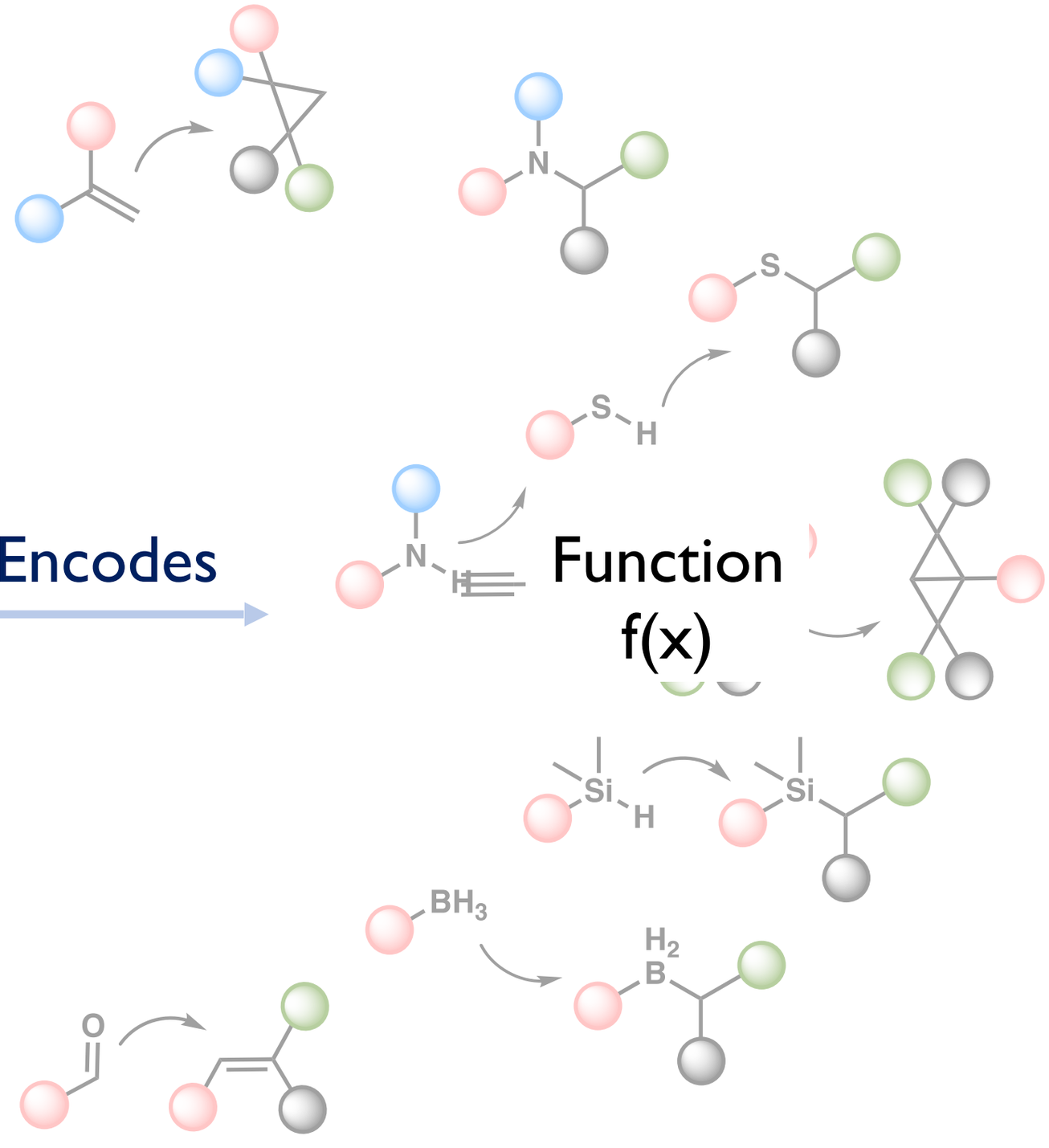
Sequence
x

from PDB ID: D0MGT2

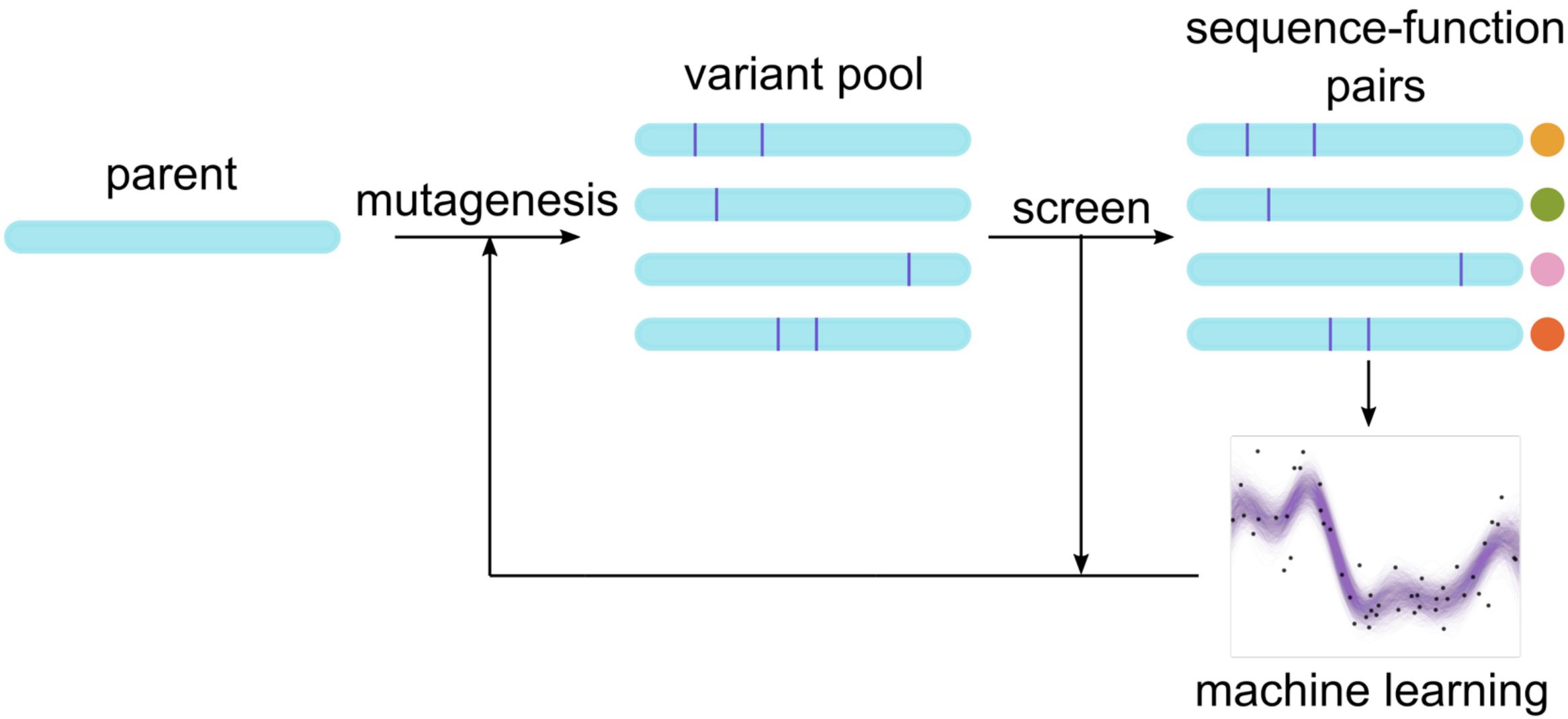
Image Credit: Zach Wu
(Arnold Group at Caltech)

Encodes

Function
f(x)



Directed evolution + Gaussian processes



Yang, Wu, Arnold (2018) bioRxiv

Summary

Bayesian learning

- Bayesian learning with feature maps

Gaussian processes

- Bayesian ridge regression + kernel trick

Reading materials

- Ch 6.4: C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
- Rasmussen & Williams, Gaussian Processes for Machine Learning, 2007
- GP animation based on materials from P. Hennig, Gaussian Processes tutorial `http://mlss.tuebingen.mpg.de/2013/2013/hennig_slides1.pdf`