# A gentle introduction to Gaussian process regression

Natalie Klein
CMU Statistics and Machine Learning

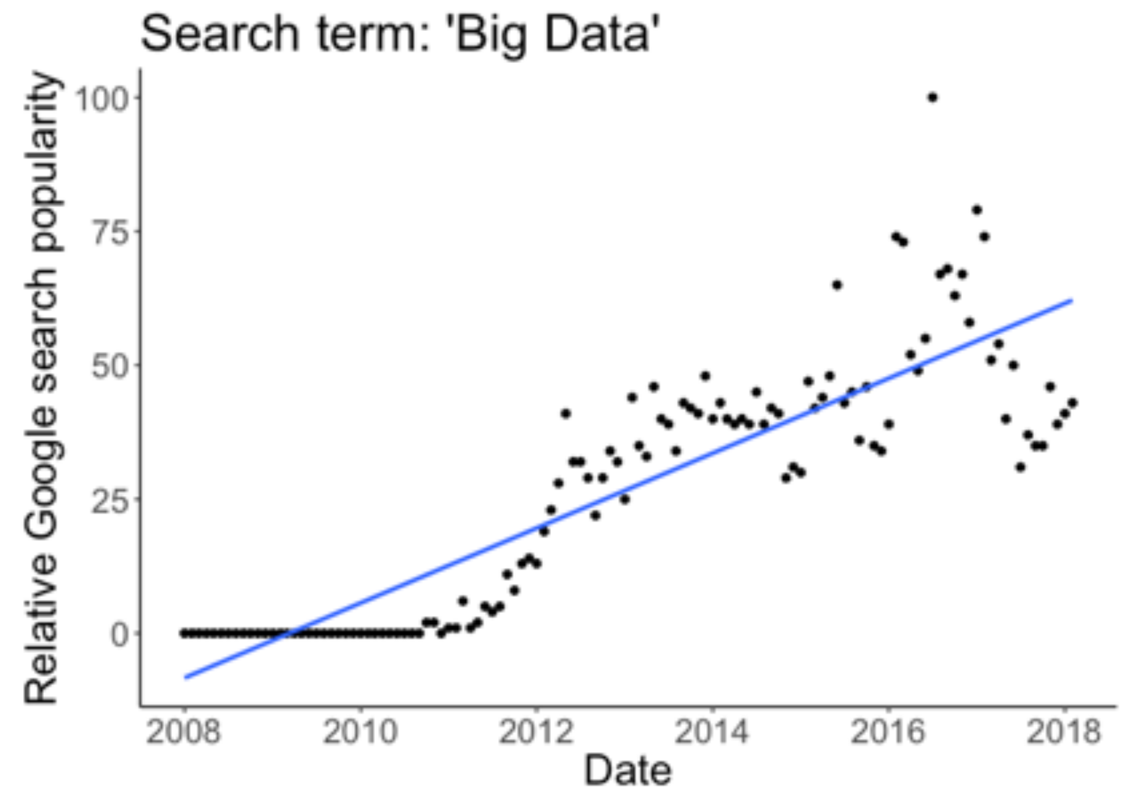
CMU Data Science Club 2/15/18

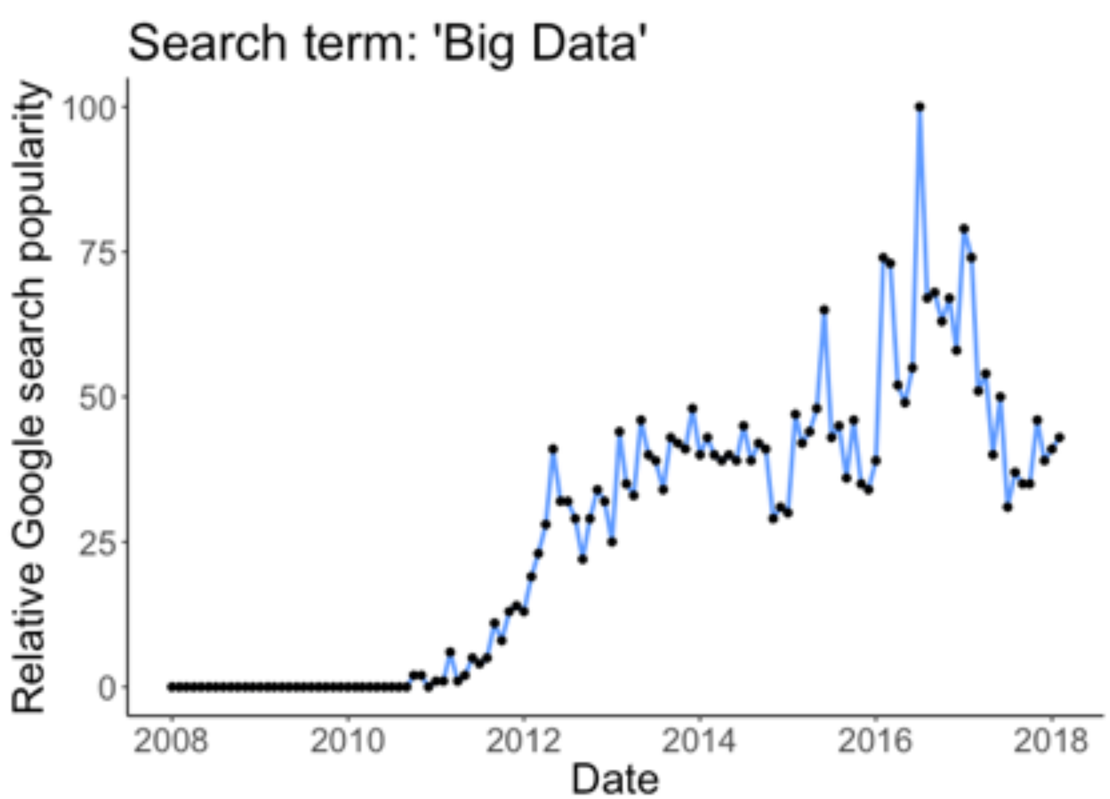
# Motivation: R demo

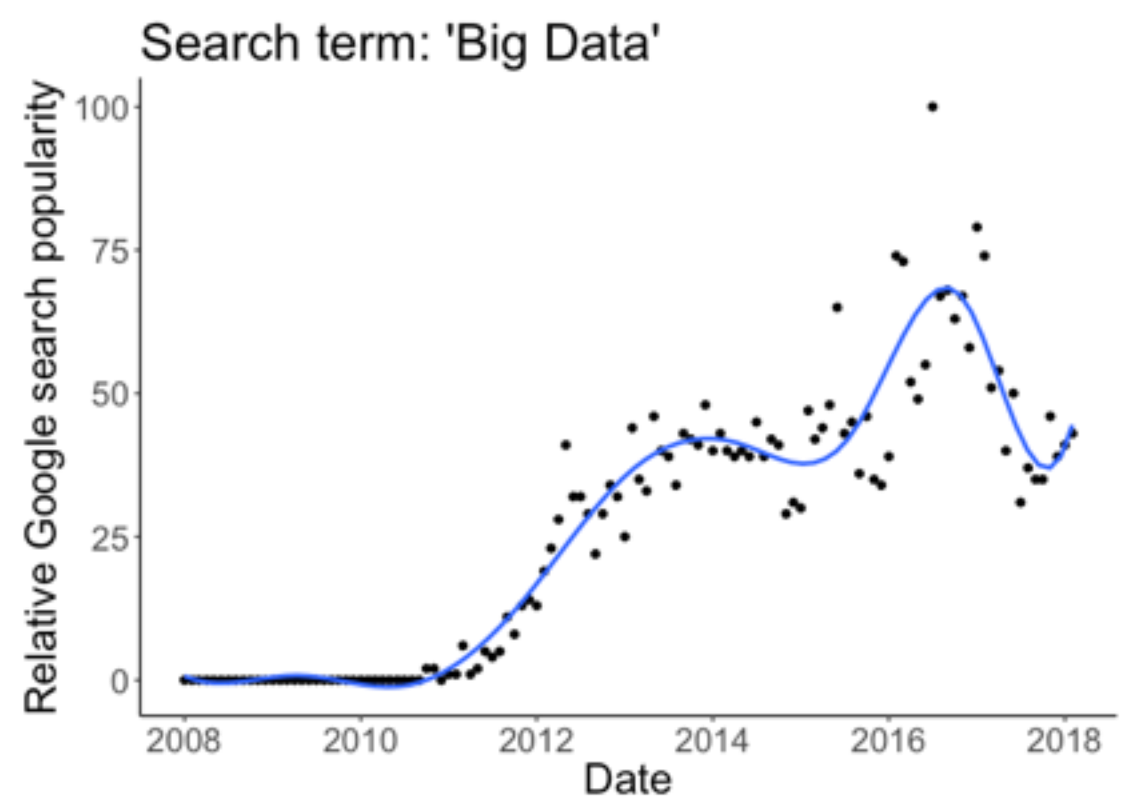
# Goals of regression

- Prediction
- Feature learning

How to handle bias/variance tradeoff?

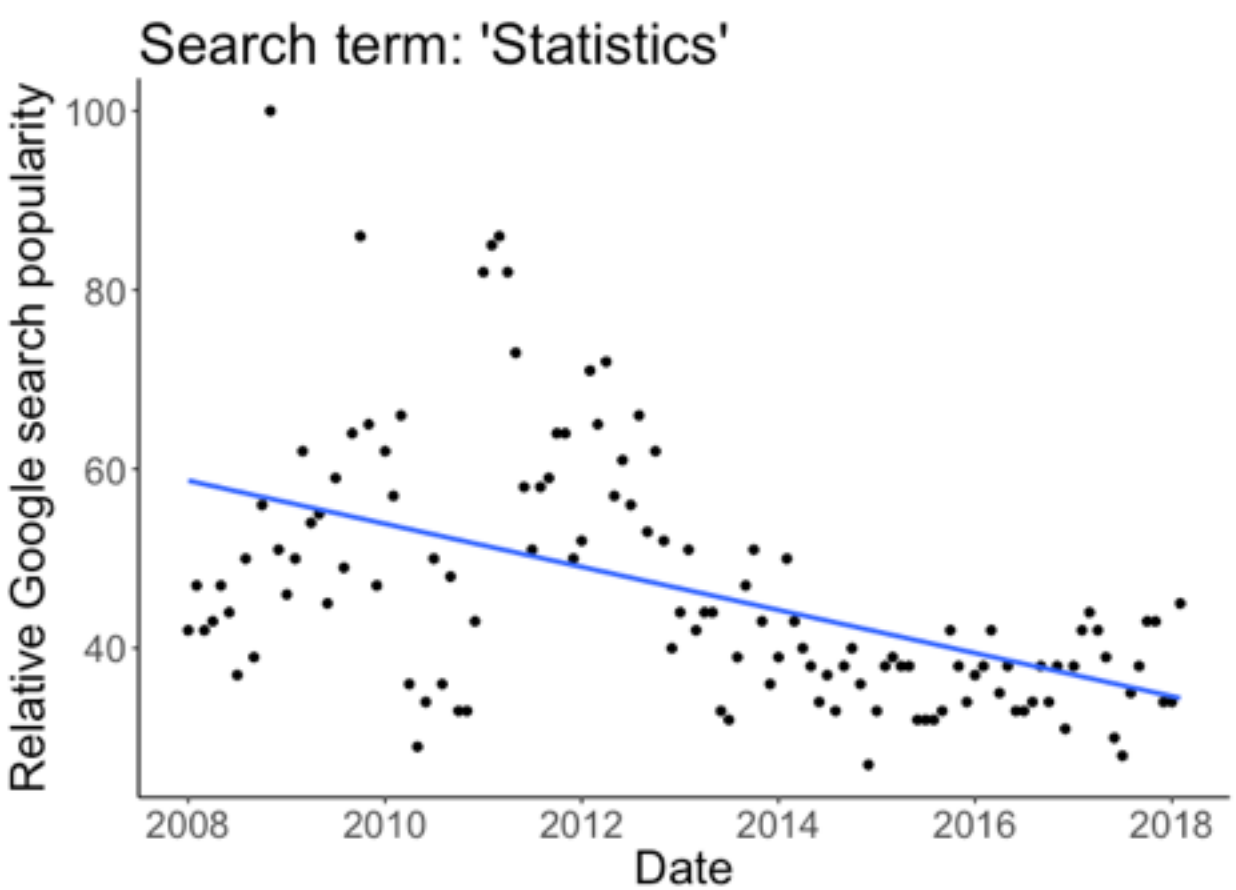


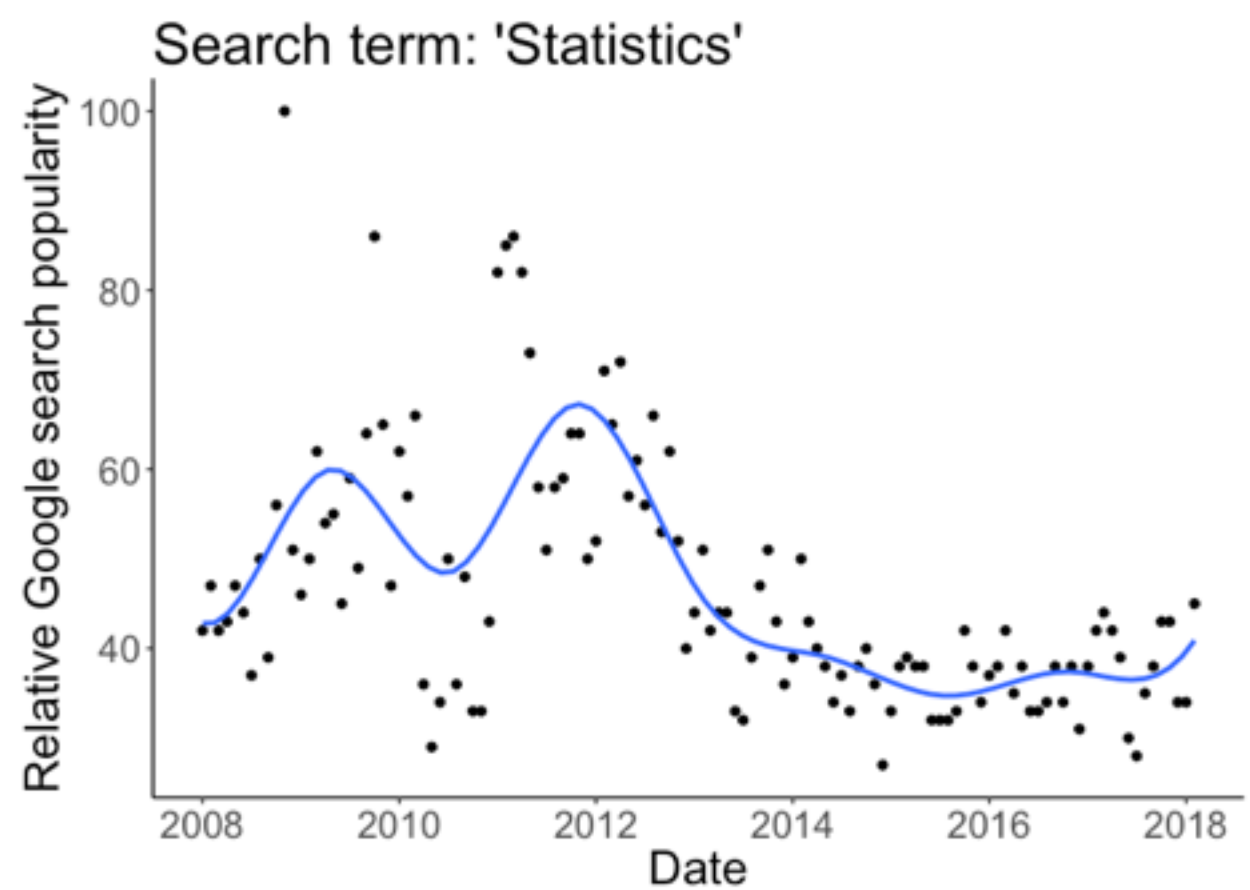



# Basic regression model

Model: $y = f(\mathbf{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0,\ \sigma_n^2)$



# Estimation: toy example

Given a random sample from a Gaussian distribution, how would you estimate the population mean?

$$Y_1, ..., Y_n \sim N(\mu, \sigma^2)$$

$$\hat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

If population mean is some fixed but unknown value…

sample average is unbiased, consistent, and efficient!

# The Bayesian approach

- Treat $\mu$ as another random variable

- Goal: get $p(\mu|y_1, ..., y_n)$ (called the *posterior*)

- How? Recall: $p(\mu|y_1, ..., y_n) = \dfrac{p(y_1, ..., y_n, \mu)}{p(y_1, ..., y_n)}$

- Also… $p(y_1, ..., y_n|\mu) = \dfrac{p(y_1, ..., y_n, \mu)}{p(\mu)}$

Posterior distribution:

$$p(\mu|y_1, ..., y_n) = \frac{p(y_1, ..., y_n|\mu)p(\mu)}{p(y_1, ..., y_n)} \propto p(y_1, ..., y_n|\mu)p(\mu)$$

# The Bayesian approach

Bayes' Theorem!

$$p(\mu|y_1, ..., y_n) \propto p(y_1, ..., y_n|\mu)p(\mu)$$

Likelihood

Prior

- From posterior we can infer:
  - the likely values of mu (most likely: the mode)
  - the uncertainty in mu (posterior variance)

# The likelihood function

We know $$p(y_i|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\}$$

Ignoring normalization constants:

$$\begin{aligned}
p(y_1, ..., y_n|\mu) &\propto \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2(\sigma^2/n)}(\bar{y}-\mu)^2\right\} = p(\bar{y}|\mu)
\end{aligned}$$

still Normal…

# Bayesian estimator

- Prior distribution for the mean: $p(\mu) = N(m, s^2)$

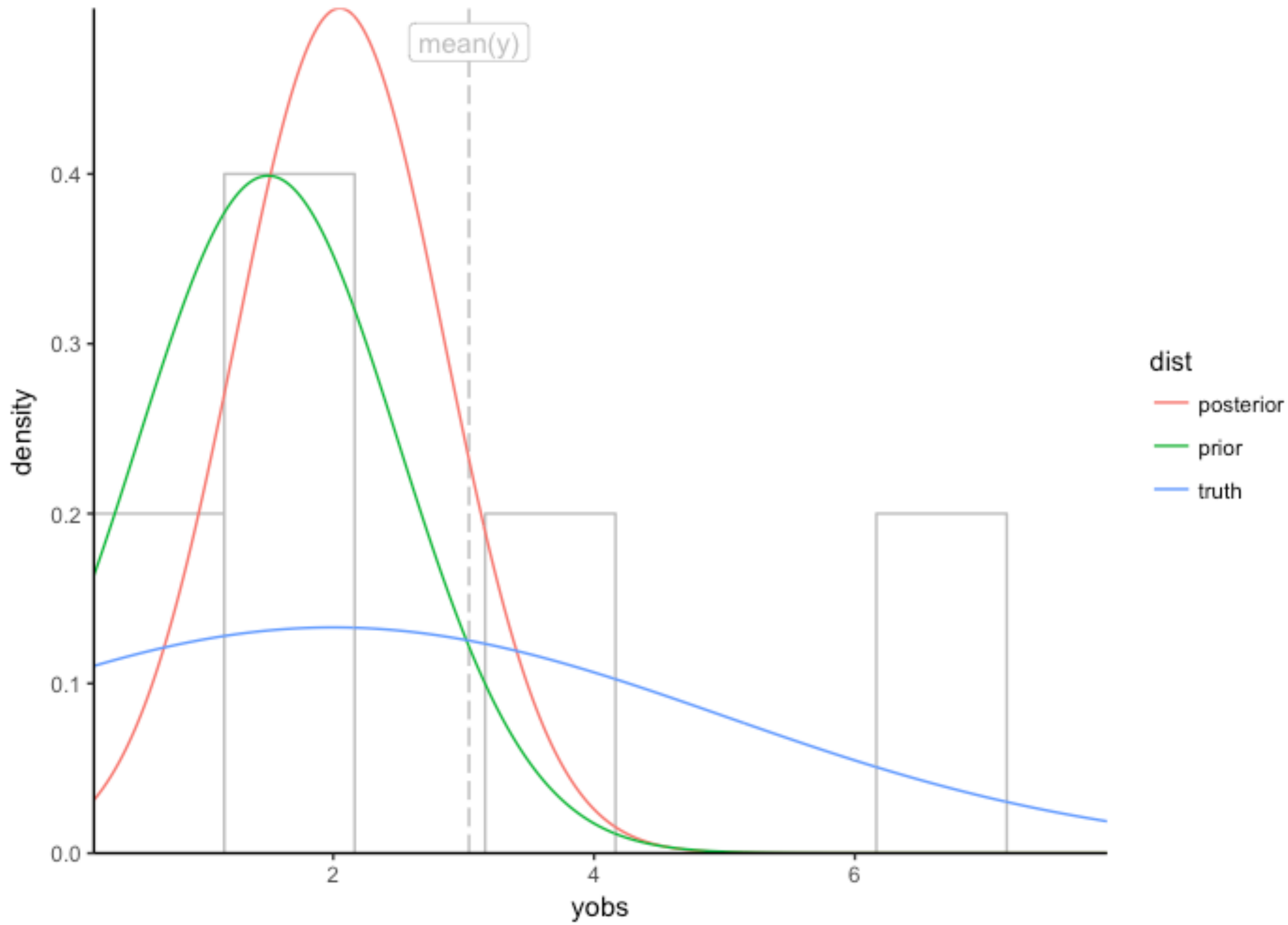
- Then

$$\underbrace{p(\mu|\bar{y})}_{\text{Posterior}} \propto \exp\left\{-\underbrace{\frac{1}{2(\sigma^2/n)}}_{\text{Likelihood}}(\bar{y}-\mu)^2\right\} \cdot \exp\left\{-\underbrace{\frac{1}{2s^2}}_{\text{Prior}}(\mu - m)^2\right\}$$

After some algebra…

$$p(\mu|\bar{y}) \propto \exp\left\{-\frac{1}{2\underbrace{\boxed{(\sigma^2/n)s^2/(\sigma^2/n+s^2)}}_{\text{Posterior variance}}}\left(\mu - \underbrace{\boxed{\frac{(\sigma^2/n)m + s^2\bar{y}}{\sigma^2/n+s^2}}}_{\text{Posterior mean}}\right)^2\right\}$$

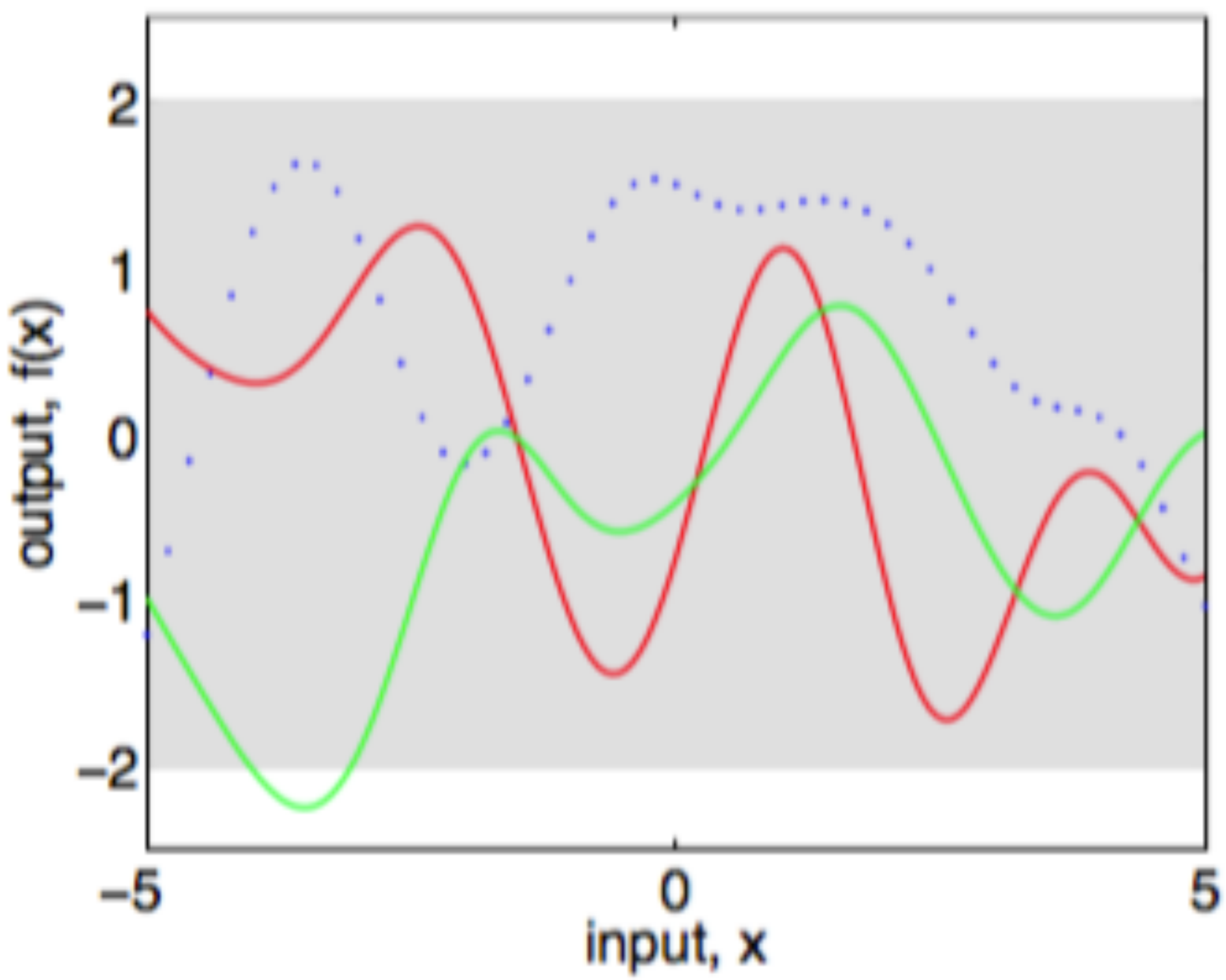
# Bayesian toy example in R



# Gaussian process regression

Model: $y = f(\mathbf{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0,\, \sigma_n^2)$

Prior: $f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big)$




[Rasmussen, Williams 2006]

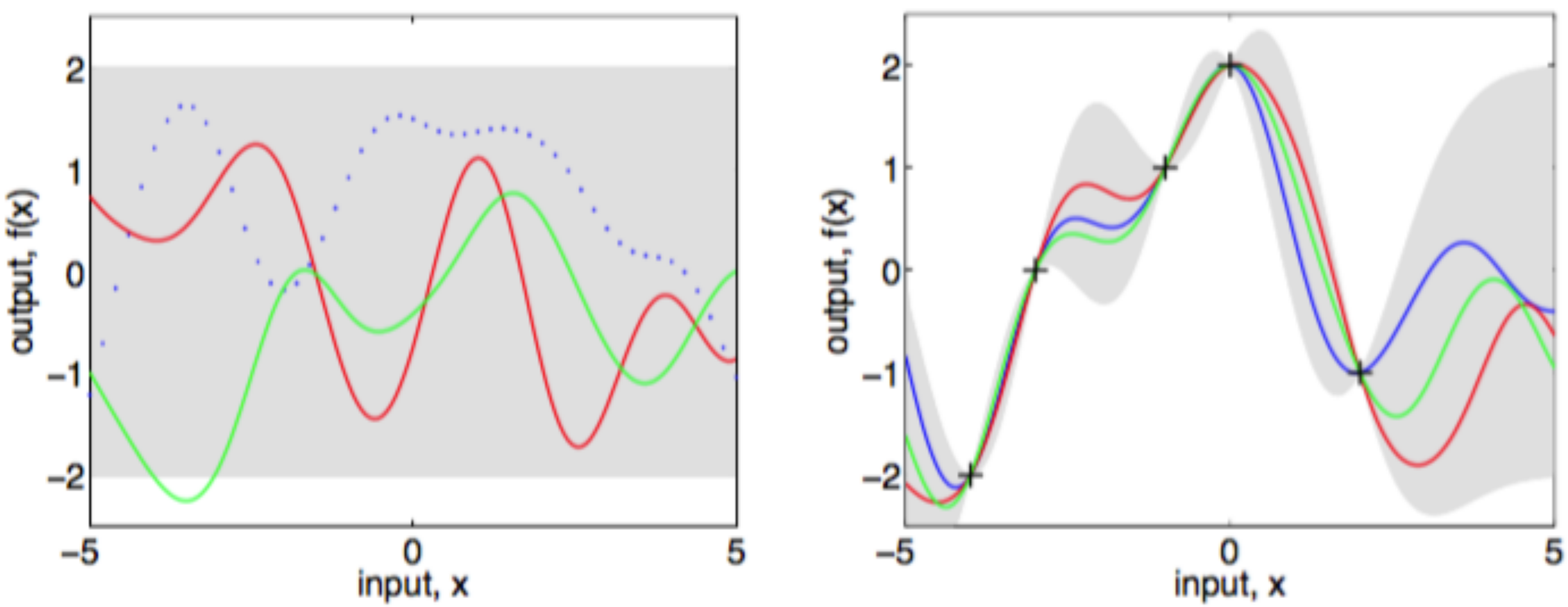
# Gaussian process regression

Posterior:

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*, \operatorname{cov}(\mathbf{f}_*)\big), \text{ where}$$

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_* | X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y},$$

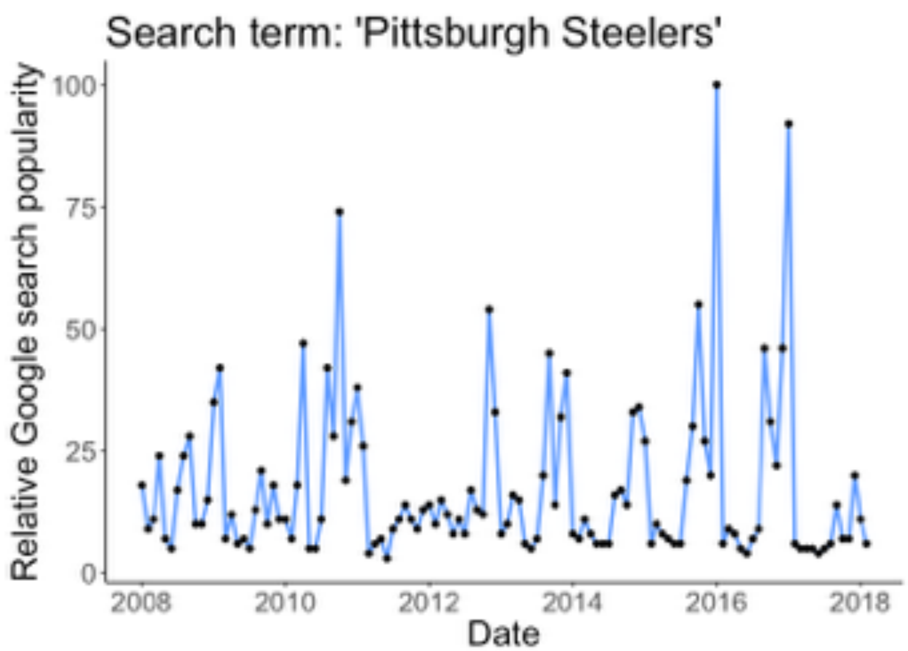
$$\operatorname{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*)$$

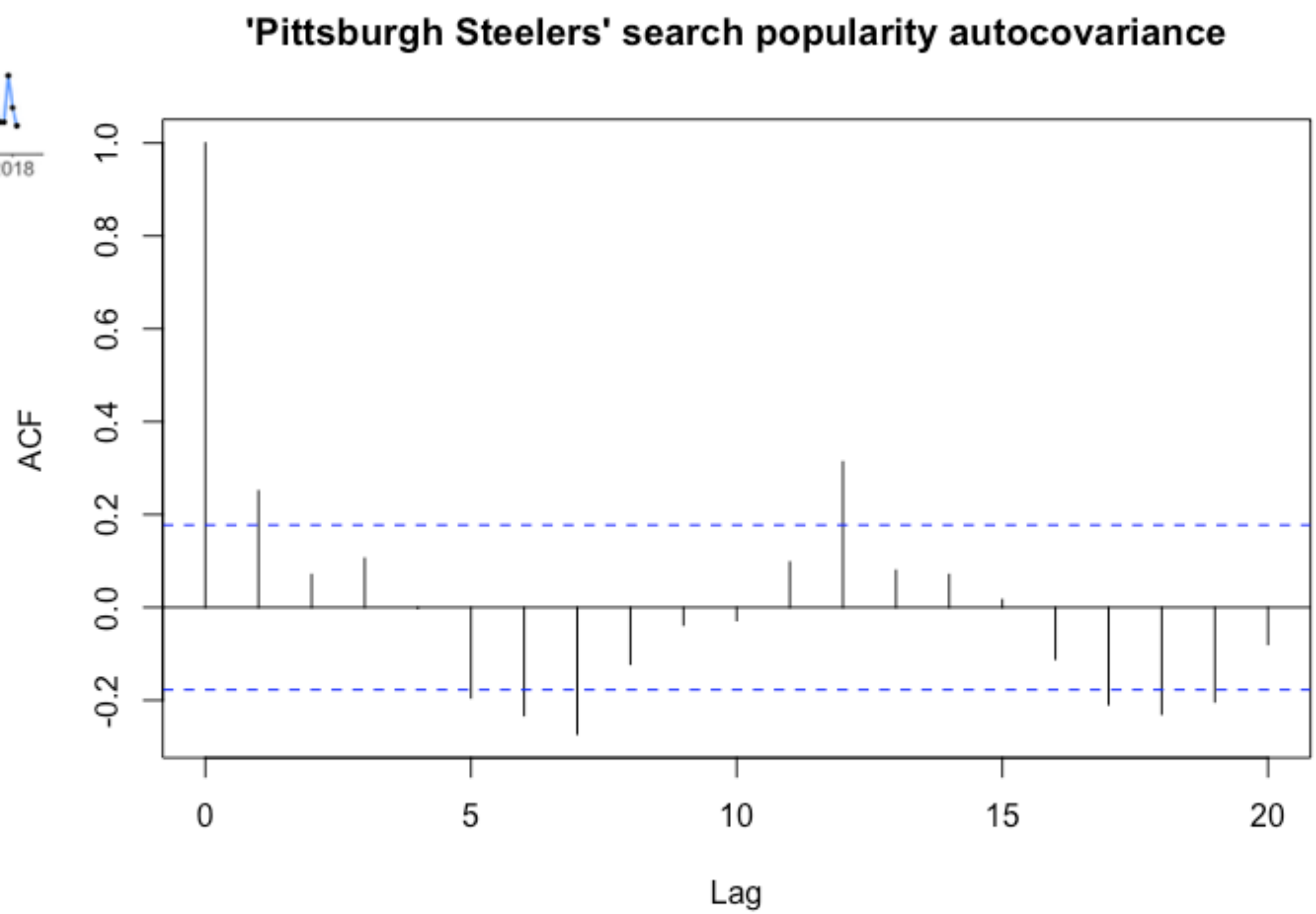



(a), prior (b), posterior

[Rasmussen, Williams 2006]

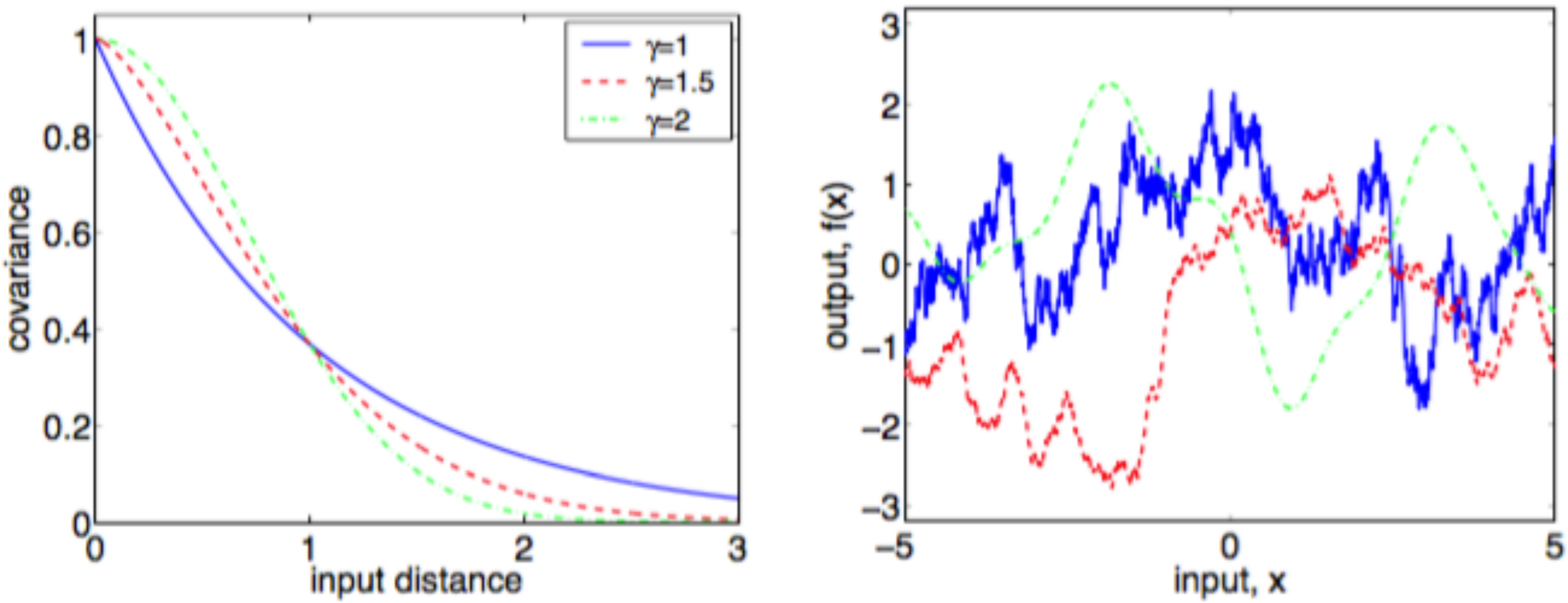
# Autocovariance



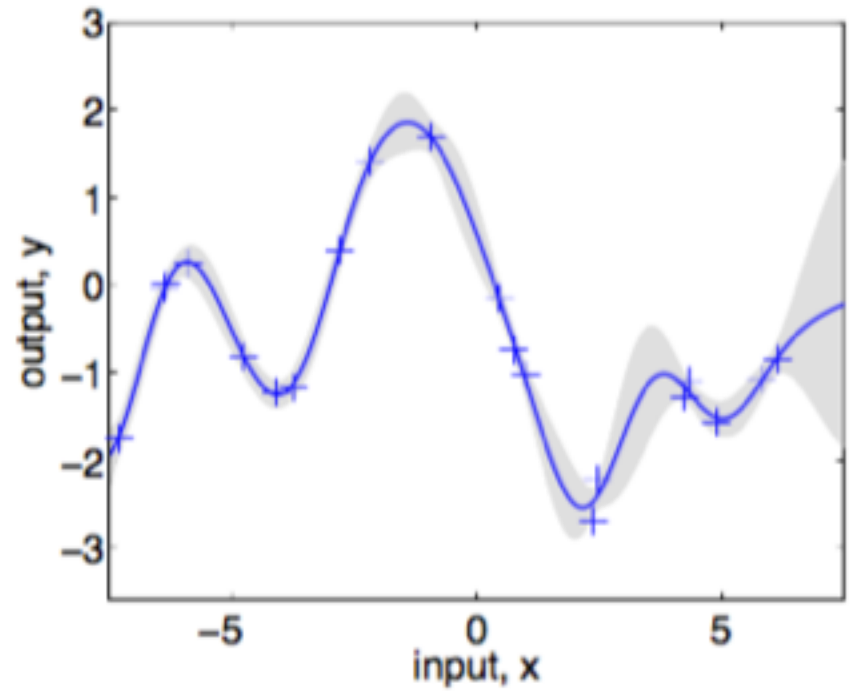
# Covariance functions

Example: $k(r) = \exp\left(-(r/\ell)^\gamma\right)$ for $0 < \gamma \leq 2$.

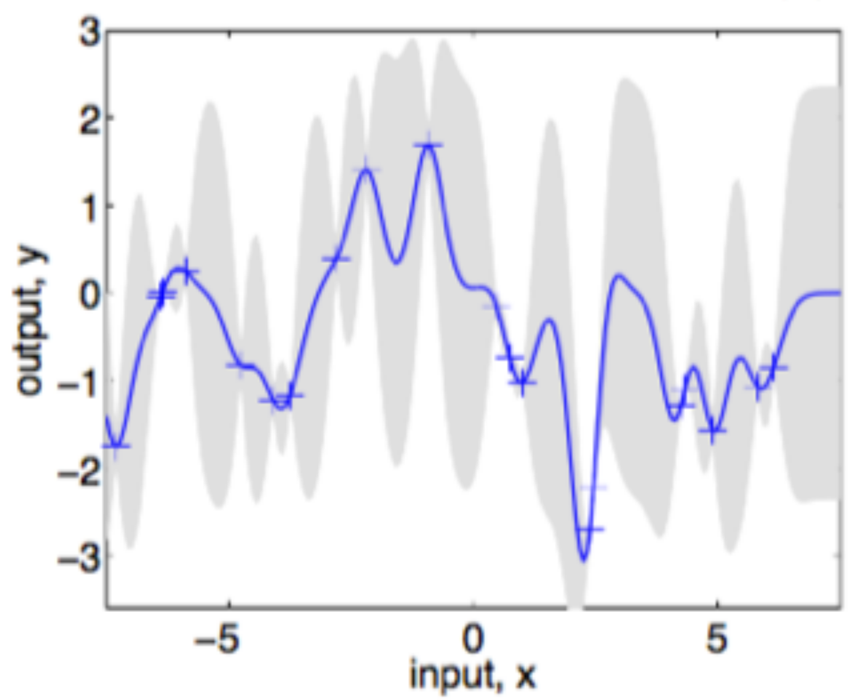



[Rasmussen, Williams 2006]

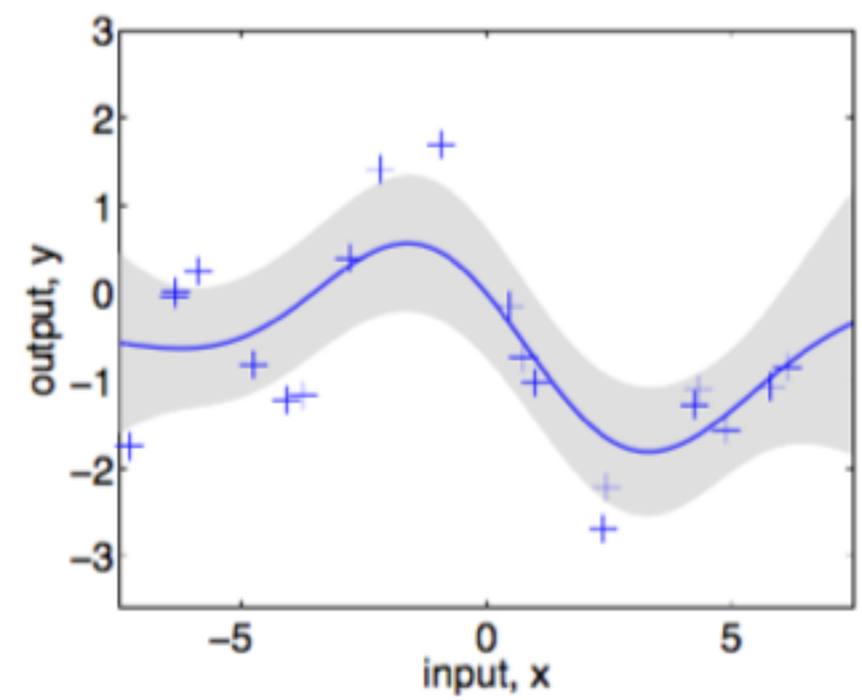
# Covariance hyperparameters




(a), $\ell = 1$




(b), $\ell = 0.3$




(c), $\ell = 3$

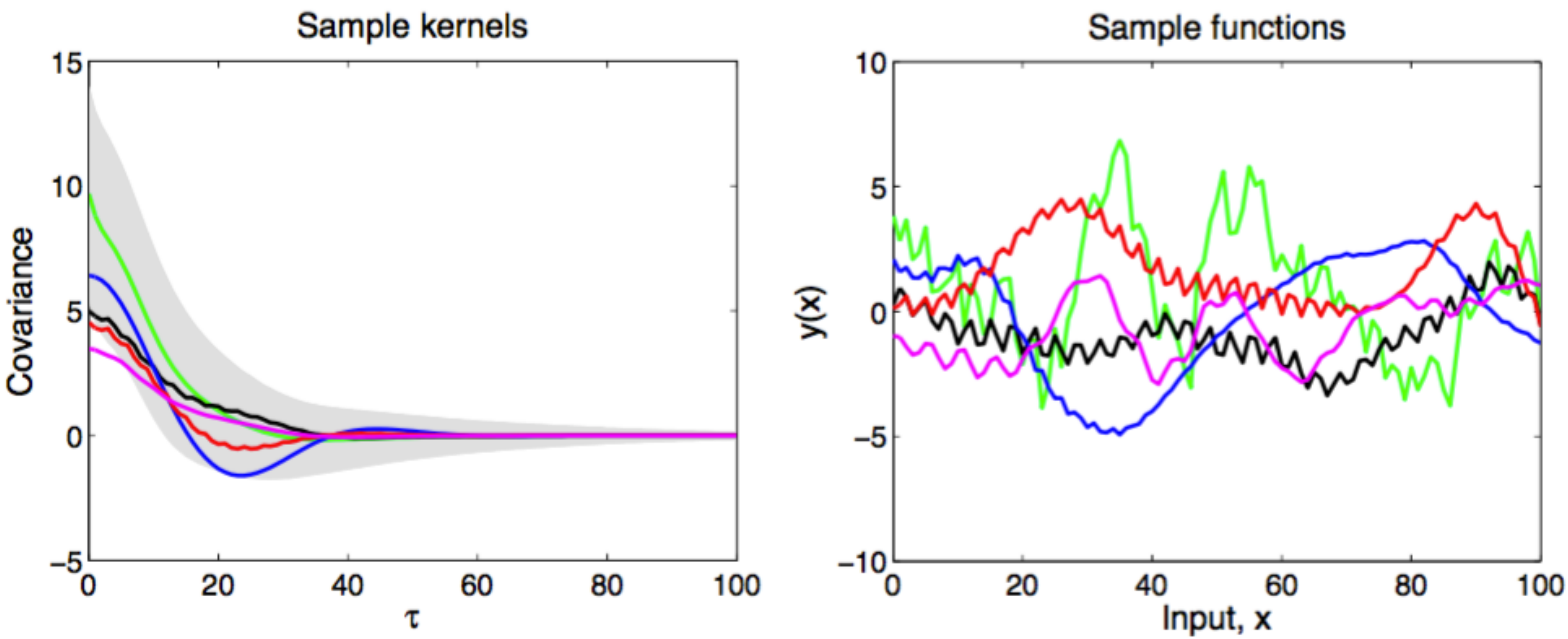
[Rasmussen, Williams 2006]

# Recap

- Regression: fit $y = f(\mathbf{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0,\ \sigma_n^2)$
  - Can predict new values using $\boldsymbol{f}$
  - Can learn features of data
- Have to constrain $\boldsymbol{f}$
- Gaussian process regression puts a prior $\boldsymbol{f}$ on which is controlled by a covariance function

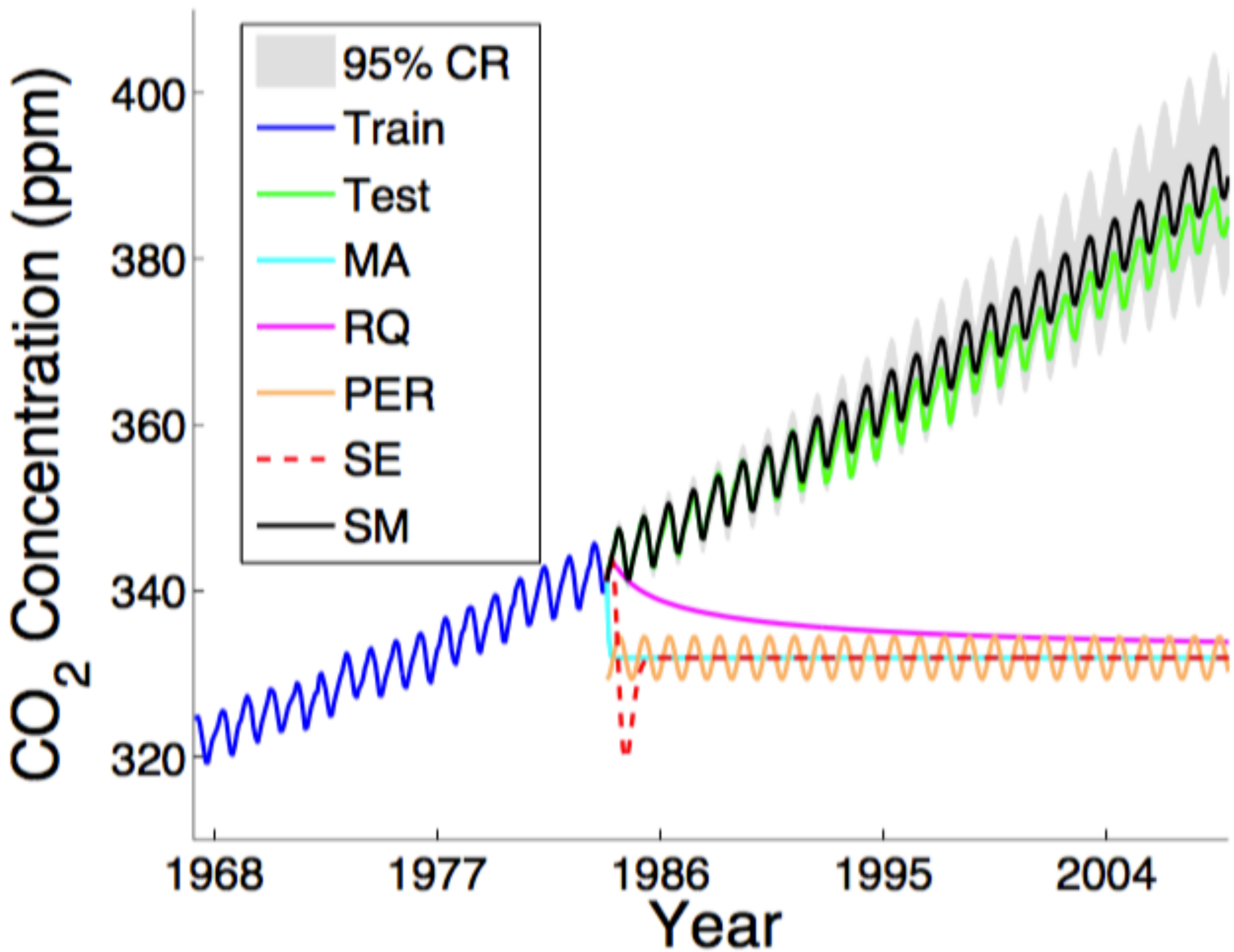
# GP examples in R

# Fancier covariance functions



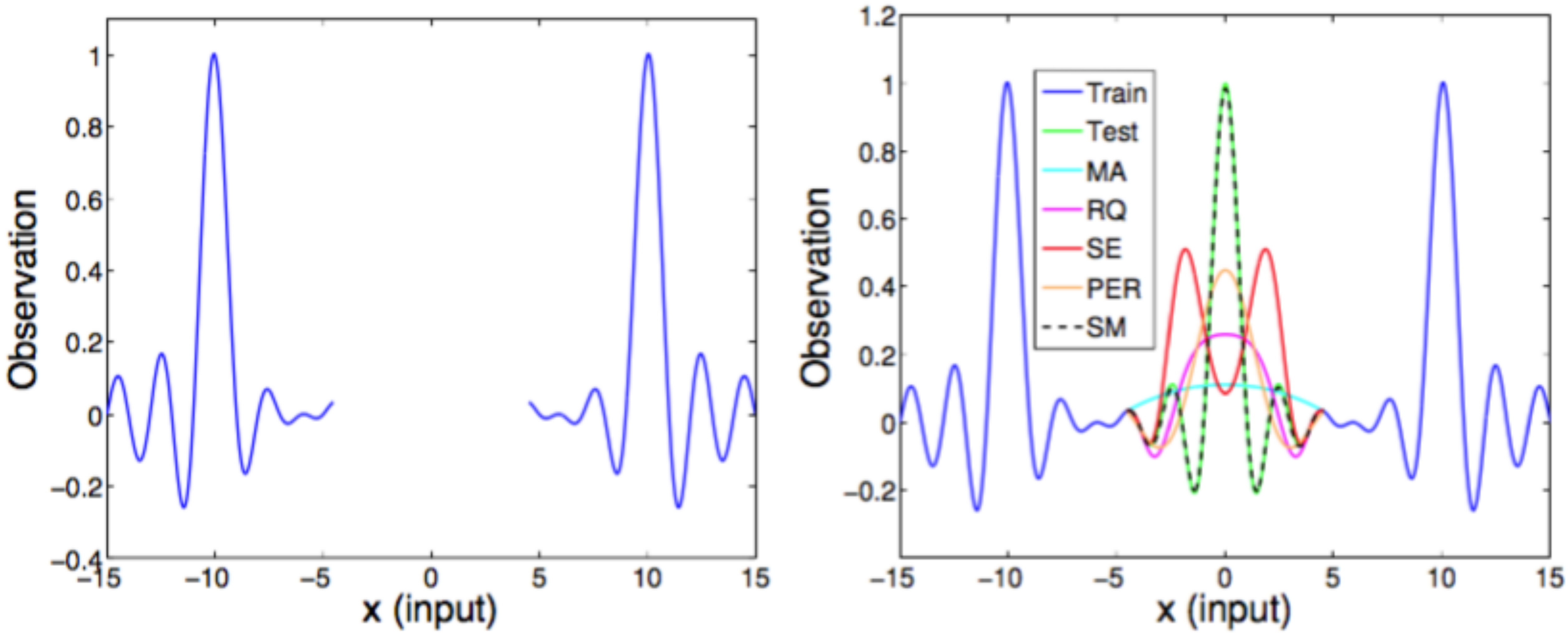

[Wilson 2013]

# Fancier covariance functions



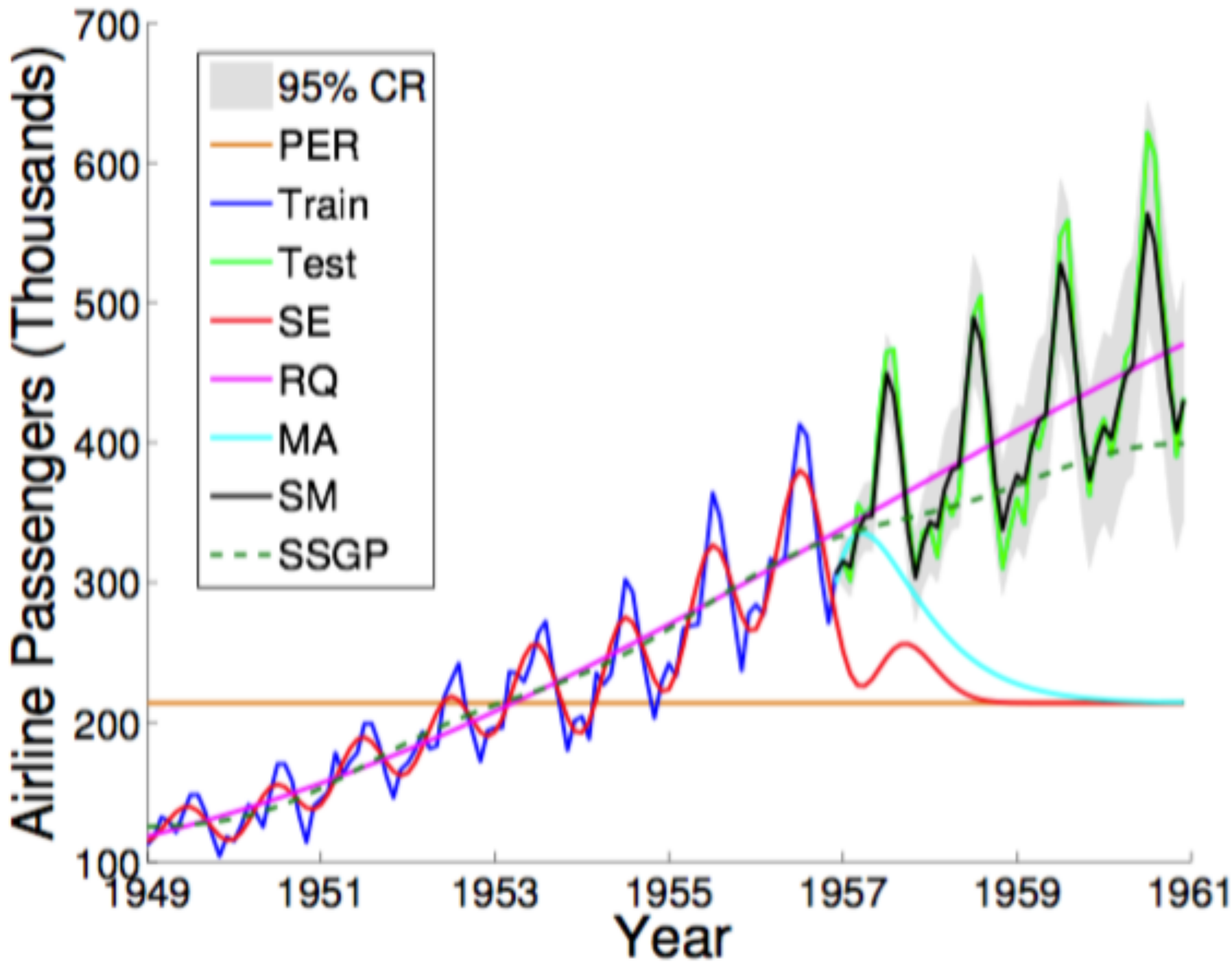

[Wilson, Adams 2013]

# Fancier covariance functions




[Wilson, Adams 2013]

# Fancier covariance functions




[Wilson, Adams 2013]

# References


- Rasmussen, C. E., & Williams, C. K. (2006). Gaussian process for machine learning. MIT press.

- Wilson, A., & Adams, R. (2013, February). Gaussian process kernels for pattern discovery and extrapolation. In International Conference on Machine Learning (pp. 1067-1075).


# A quick plug…

**Women in Statistics and Data Science Conference at CMU: March 9**

Info and registration: http://stat.cmu.edu/wids/