Automatically constructing models, and automatically explaining them, too



David Duvenaud[1], James Robert Lloyd[2], Roger Grosse[3],

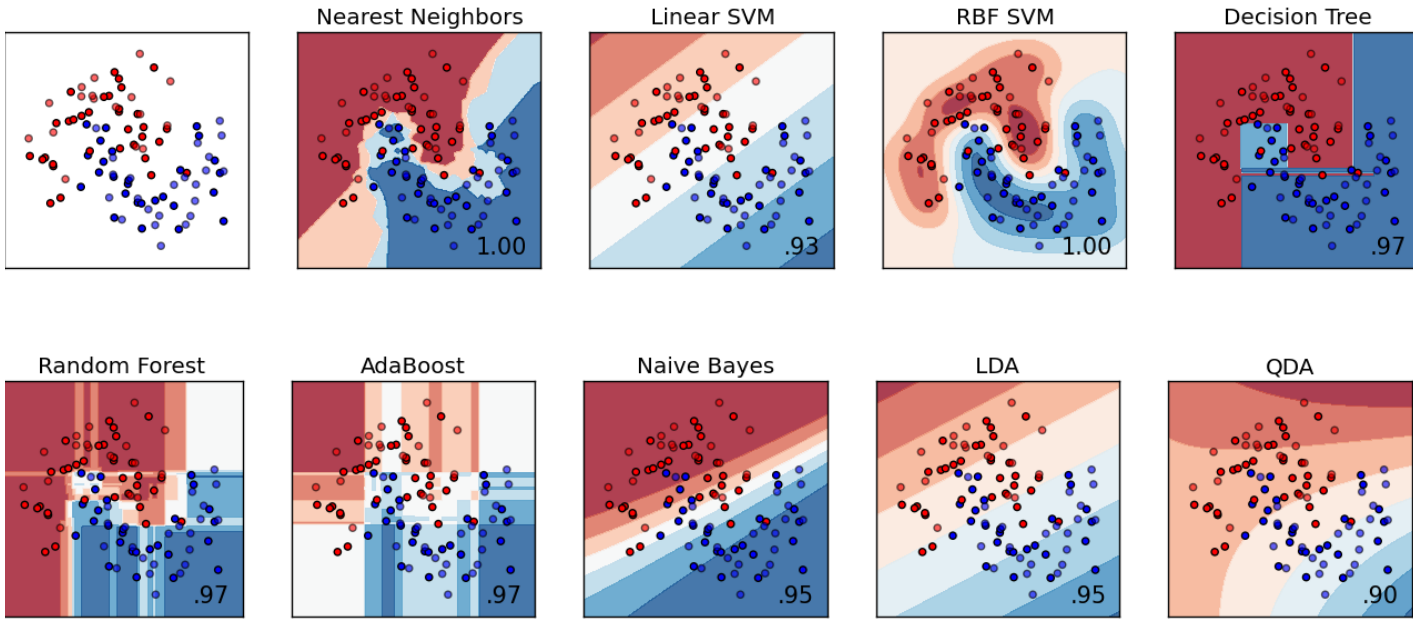

Joshua Tenenbaum[4], Zoubin Ghahramani[2]

2: University of Cambridge, 1: Harvard University
3: University of Toronto, 4: Massachusetts Institute of Technology

March 4, 2015

TYPICAL STATISTICAL MODELLING

- Models typically built by hand, or chosen from a fixed set

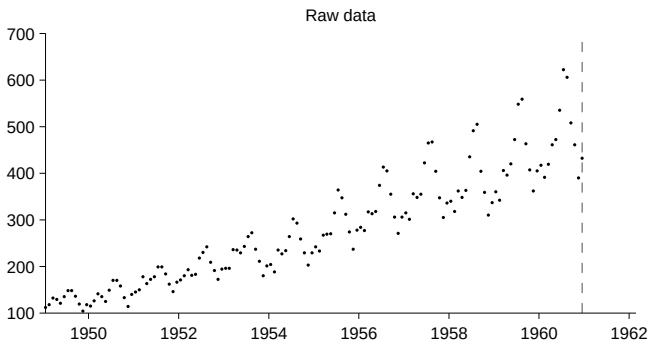


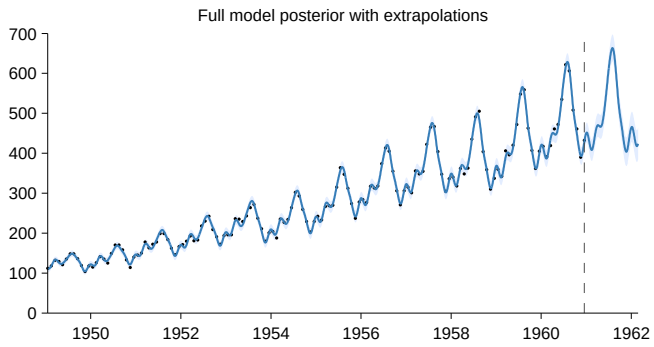
- Building by hand requires considerable expertise
- Just being nonparametric isn't good enough
  - Nonparametric does not mean assumption-free!
- Can silently fail
  - How to tell if none of the models fit the data well?

CAN WE DO BETTER?

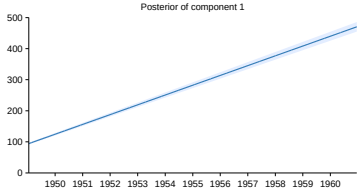
- How could an AI do modeling, forecasting, and statistics?
- An artificial statistician would need:
  - a language that could describe arbitrarily complicated models
  - a method of searching over those models
  - a procedure to check model fit
- We construct such a language over regression models, a procedure to search over it, and a method to describe in natural language the properties of the resulting models

An entirely automatic analysis

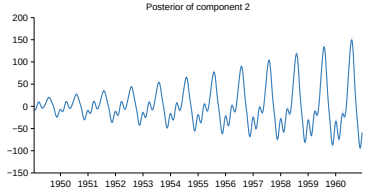




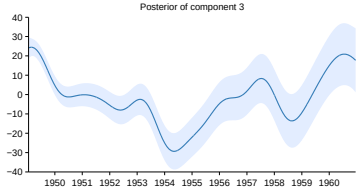
Four additive components have been identified in the data



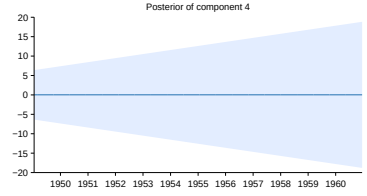
A linearly increasing function



An approximately periodic function with a period of 1.0 years with linearly increasing amplitude



A smooth function

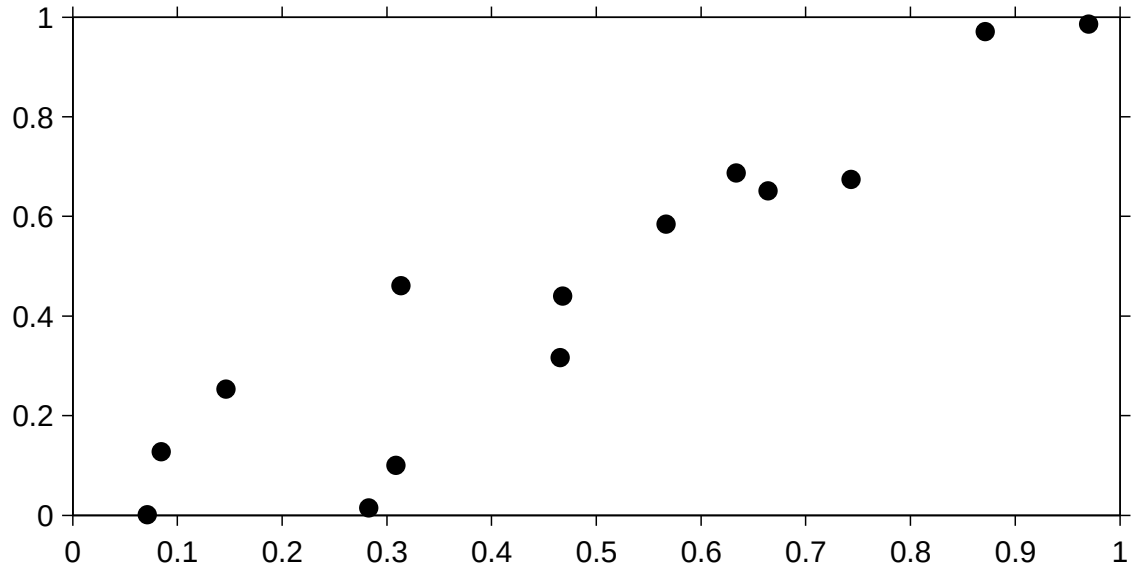


Uncorelated noise with linearly increasing standard deviation

Agenda

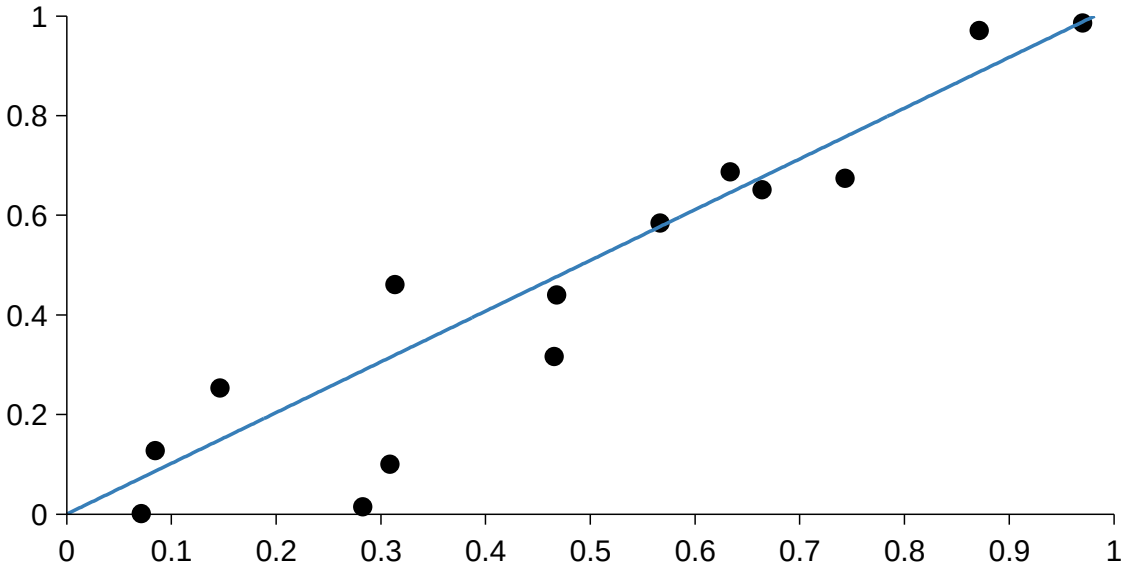
- Introduction to Bayesian models and Gaussian processes
- Description of automatic statistician
- Examples of automatically generated models and descriptions

How to do linear regression



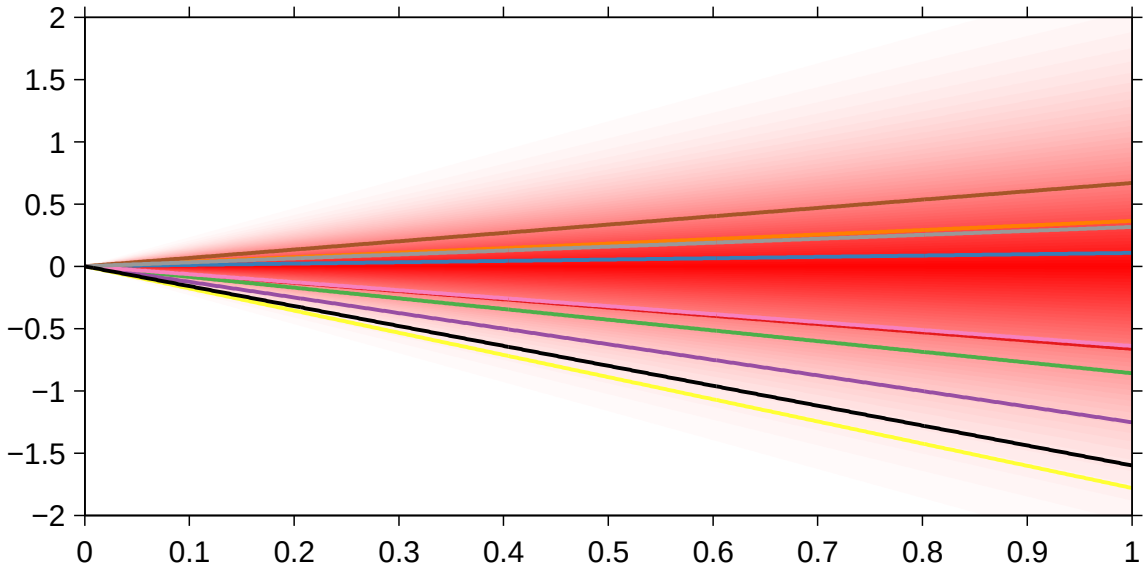
- Linear regression: $y_i = mx_i + \varepsilon_i$ where x are inputs, and y outputs, and ε are errors or noise

How to do linear regression



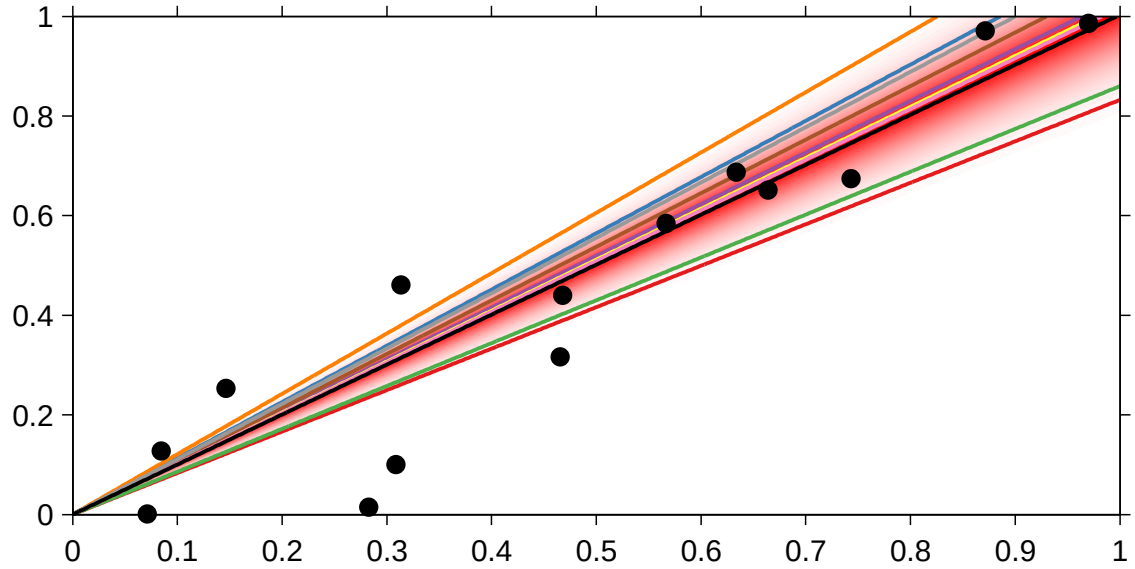
- Common approach: Estimate m by least squares

How to do Bayesian linear regression



- Bayesian approach:
  - Specify beliefs about m using probability distributions

How to do Bayesian linear regression



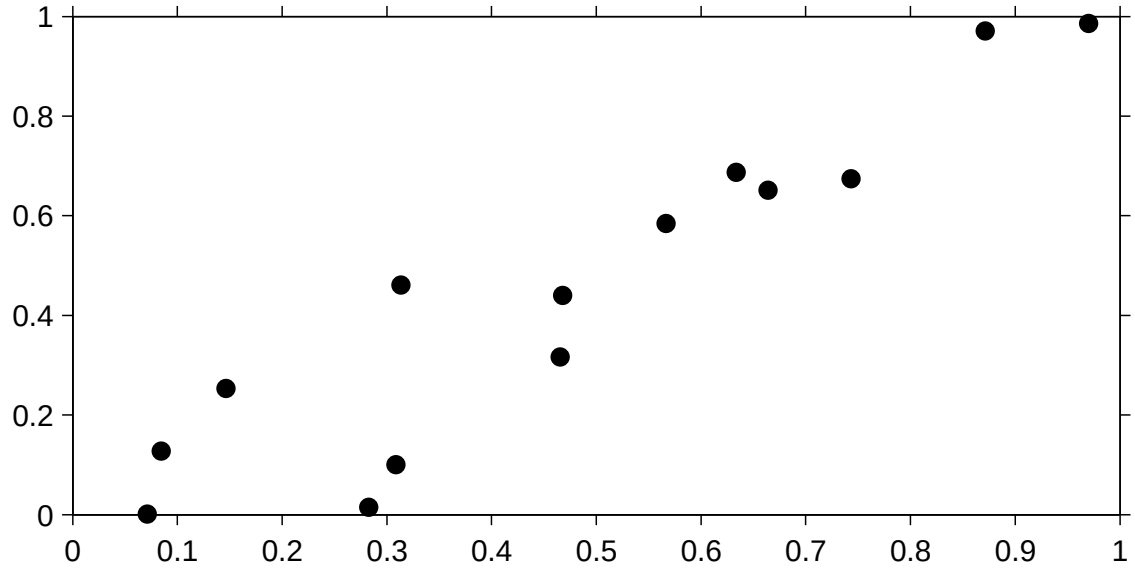
- Bayesian approach:
  - Specify prior beliefs about m using probability distributions
  - Follow rules of probability to update beliefs after observing data

BAYES RULE

- Suppose we wish to fit a model M with parameters θ to data D
  - M defined by $p(D \mid \theta, M)$ - the likelihood
- We represent our uncertainty about θ with a probability distribution
  - $p(\theta \mid M)$ - the prior
- We then derive $p(\theta \mid D, M)$ - the posterior

$$\underbrace{p(\theta \mid D, M)}_{\text{posterior}} = \frac{\overbrace{p(D \mid \theta, M)}^{\text{likelihood}} \overbrace{p(\theta \mid M)}^{\text{prior}}}{\underbrace{\int p(D \mid \theta, M) p(\theta \mid M) \mathrm{d}\theta}_{\text{marginal likelihood}}}$$

BAYESIAN LINEAR REGRESSION

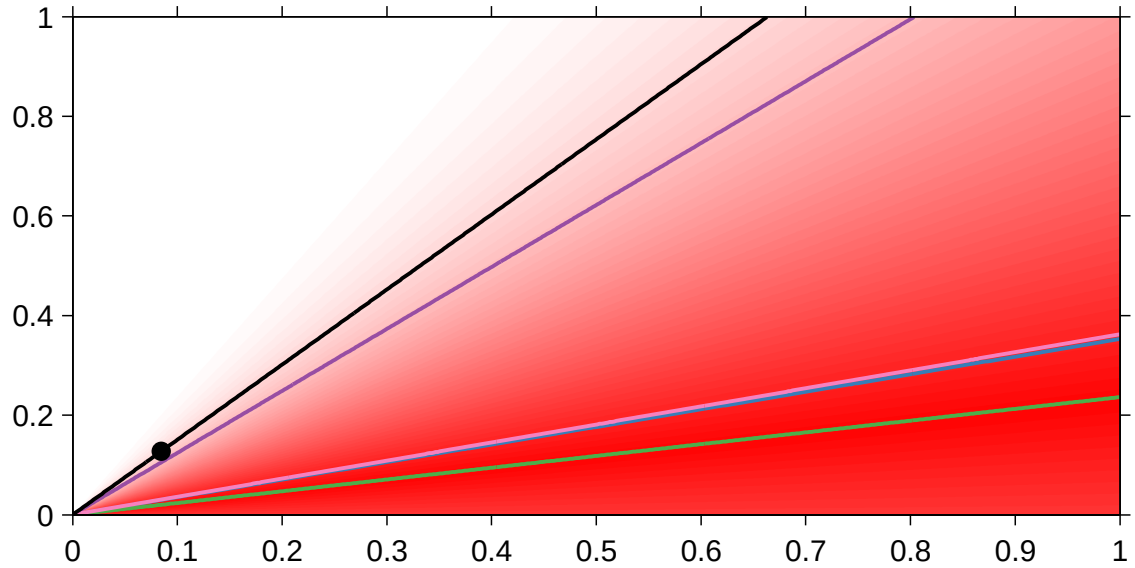


- Linear regression starts by assuming a model of the form $y_i = mx_i + \varepsilon_i$ where x and y are inputs and outputs respectively and ε are errors or noise

BAYES RULE APPLIED TO LINEAR REGRESSION

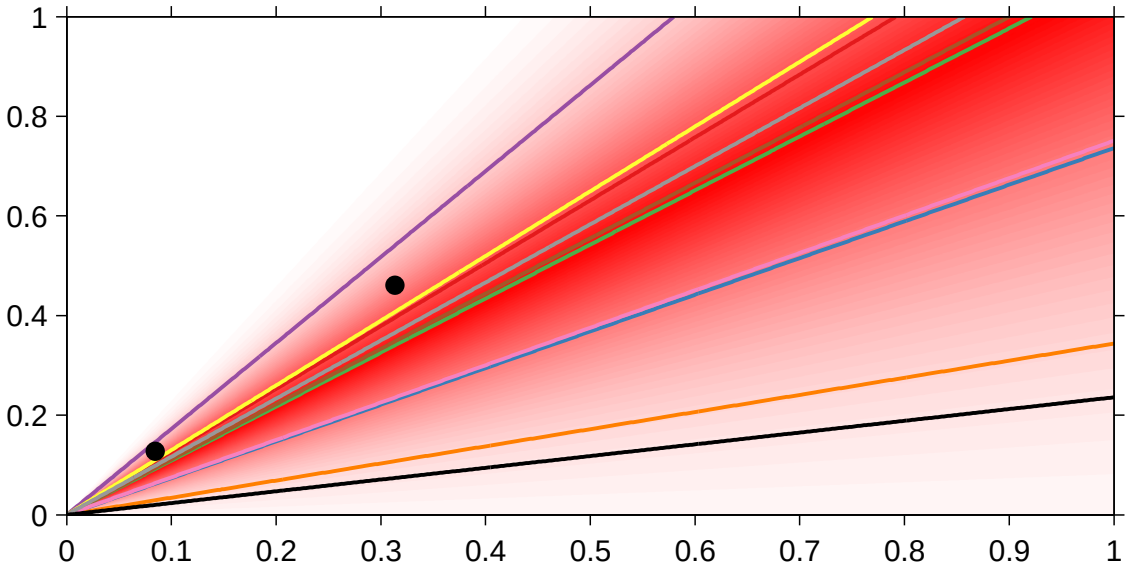
$$
\begin{aligned}
p(m \mid y, x) &= \frac{p(y \mid m, x)\, p(m)}{\int p(y \mid m, x) p(m)\, \mathrm{d}m} \\
&\propto p(y \mid m, x)\, p(m) \\
&\propto \left(\prod_i \frac{1}{2\pi\sigma_\varepsilon} e^{-(y_i - m x_i)^2/(2\sigma_\varepsilon^2)} \right) \frac{1}{2\pi} e^{-m^2/2} \\
&= \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2} \right)
\end{aligned}
$$

Bayesian linear regression



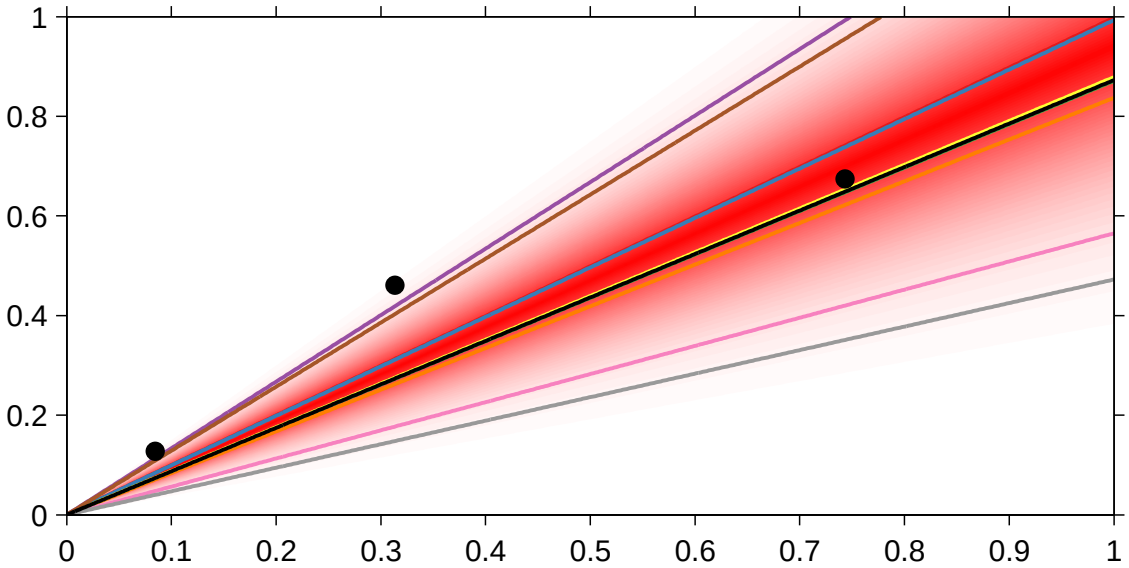
$$p(m \mid y, x) = \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2} \right)$$

Bayesian linear regression



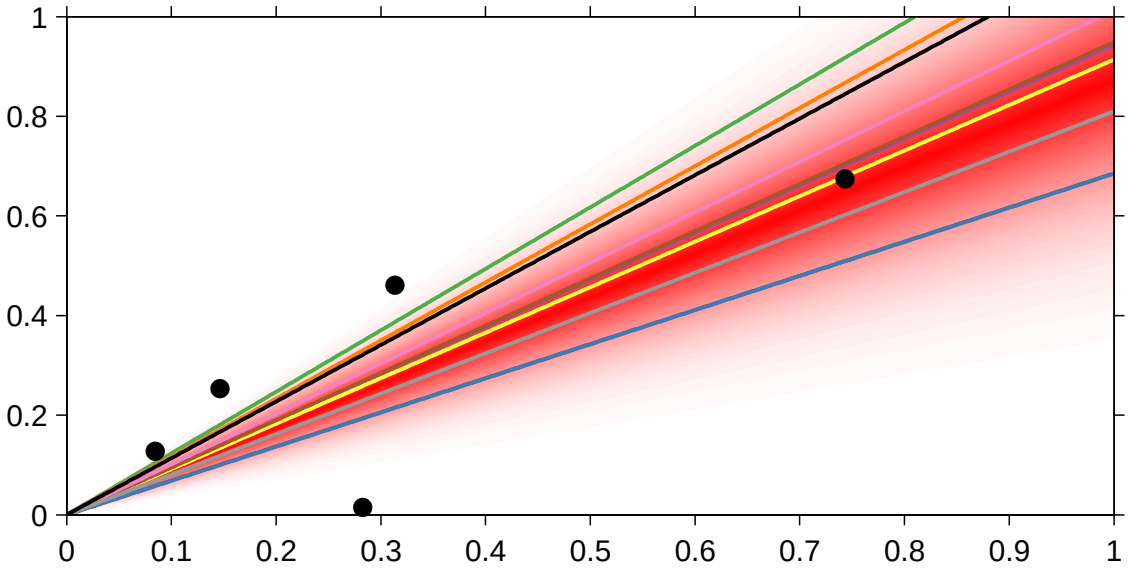
$$p(m \mid y, x) = \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2} \right)$$

Bayesian linear regression



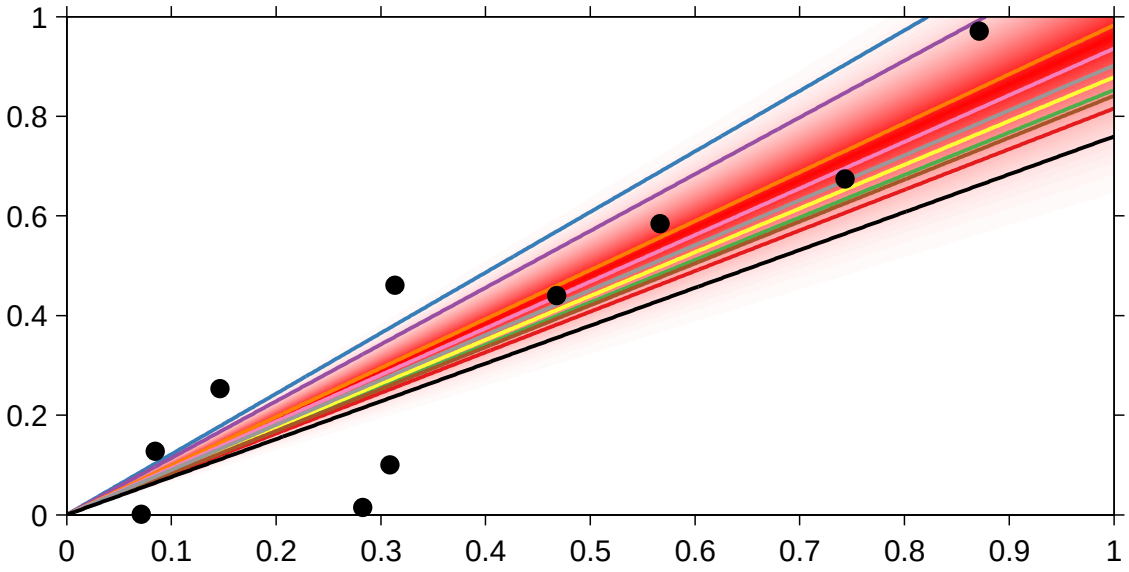
$$p(m \mid y, x) = \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2} \right)$$

Bayesian linear regression



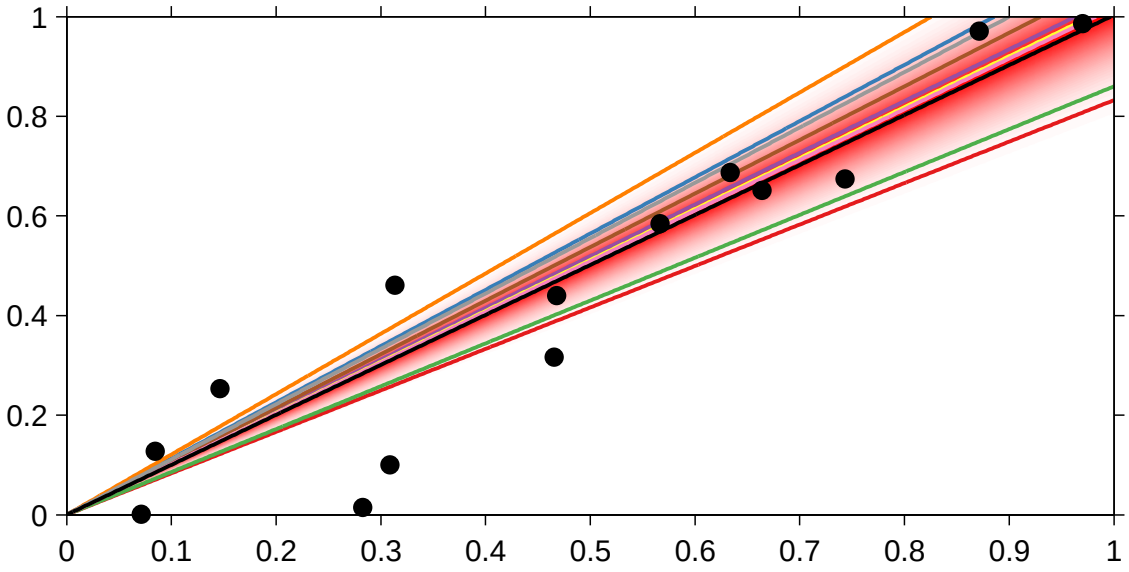
$$p(m \mid y, x) = \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2} \right)$$

Bayesian linear regression



$$p(m \mid y, x) = \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2} \right)$$

Bayesian linear regression



$$p(m \mid y, x) = \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2} \right)$$

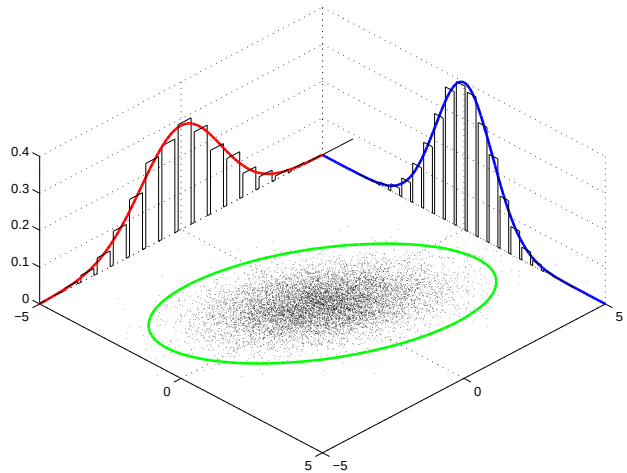
BAYESIAN LINEAR REGRESSION REWRITTEN

- We defined a Bayesian linear regression model by specifying priors on m and ε_i
  - $m \sim \mathcal{N}(0, 1)$, $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$
  - $y_i \,|\, m, \varepsilon_i = m x_i + \varepsilon_i$

- This implicitly defined a joint prior on $\{y_i \,:\, i = 1, \ldots, n\}$
  - $y_i \sim \mathcal{N}(0, x_i^2 + \sigma_\varepsilon^2)$ (sum of two normals)
  - $\text{cov}(y_i, y_j) = x_i x_j \quad \forall\, i \neq j$

$\{y_i \,:\, i = 1, \ldots, n\}$ has a multivariate normal distribution

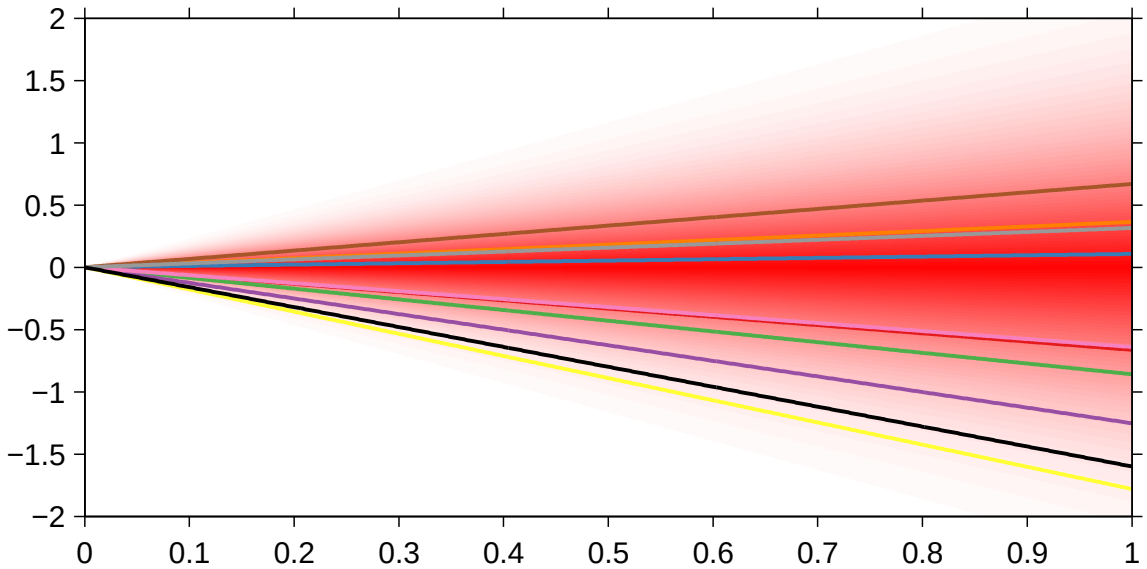
- $y \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$
- where $k_{ij} = x_i x_j + \delta_{ij} \sigma_\varepsilon^2$



GAUSSIAN PROCESSES

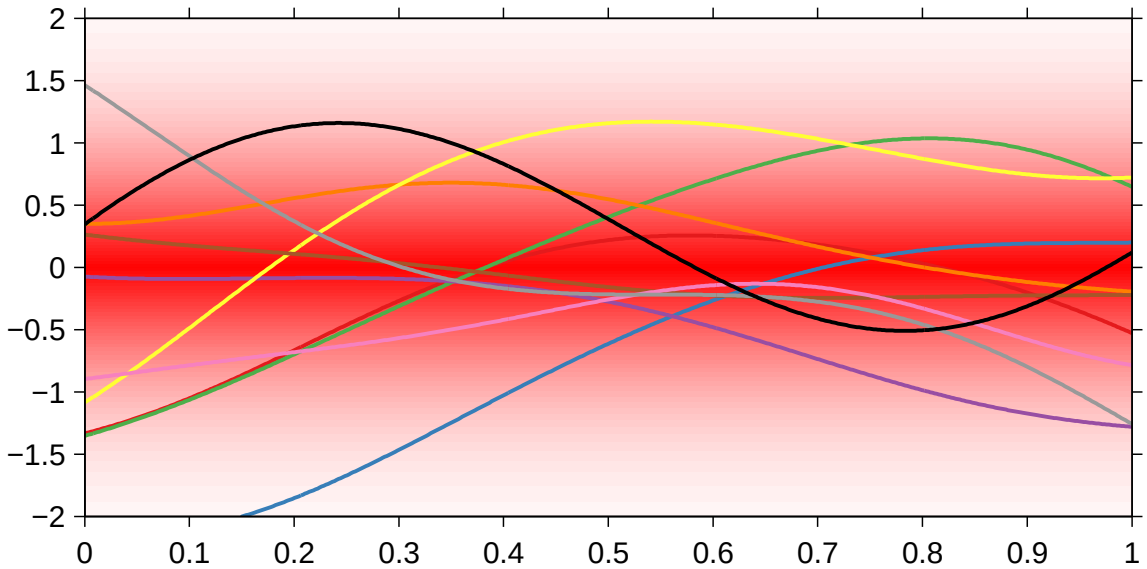
- A Gaussian process is collection of random variables, any finite number of which have a joint Gaussian distribution
- We can write this collection of random variables as $\{f(x) : x \in \mathcal{X}\}$ i.e. a function f evaluated at inputs x
- A GP is completely specified by
  - Mean function, $\mu(x) = \mathbb{E}(f(x))$
  - Covariance / kernel function, $k(x, x') = \text{Cov}(f(x), f(x'))$
  - Denoted $f \sim \text{GP}(\mu, k)$
- Can be thought of as a probability distribution on functions

LINEAR KERNEL = LINEAR FUNCTIONS



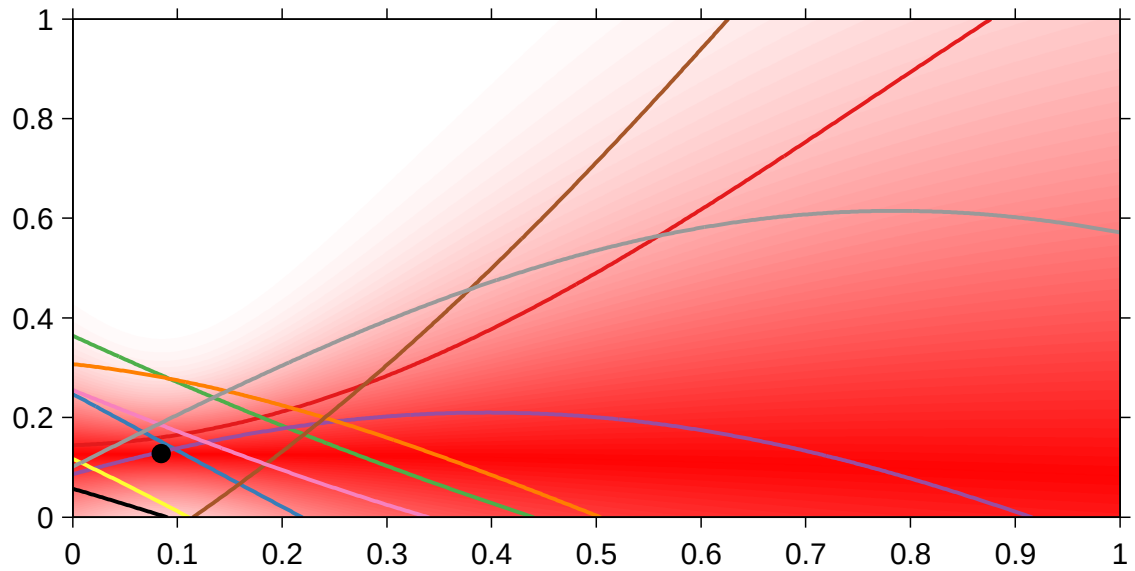
$$k(x, x') = xx'$$

What about other kernels?

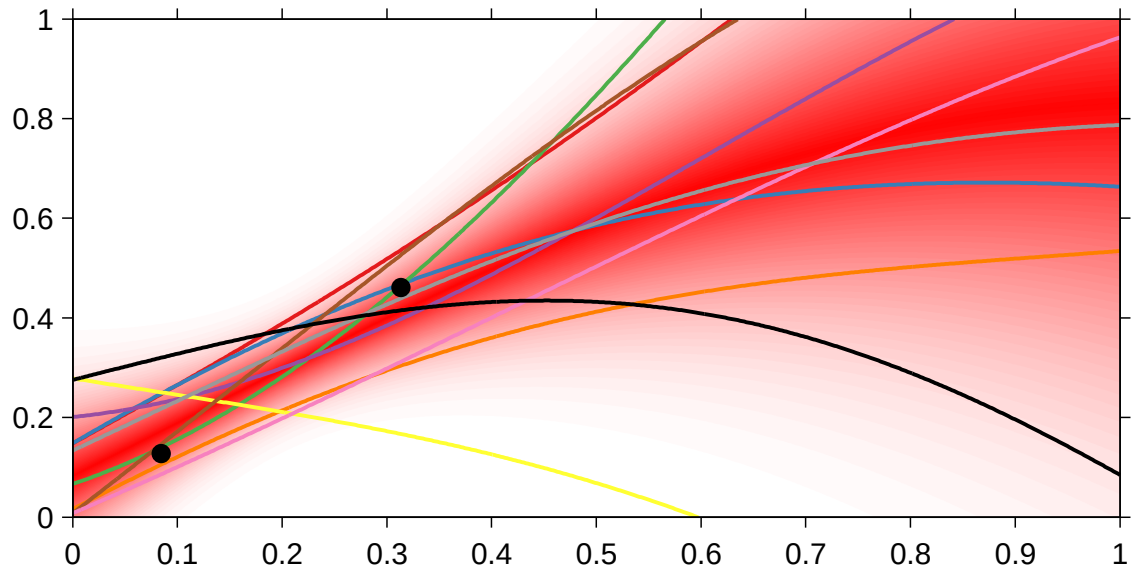


$$k(x, x') = e^{-(x-x')^2}$$

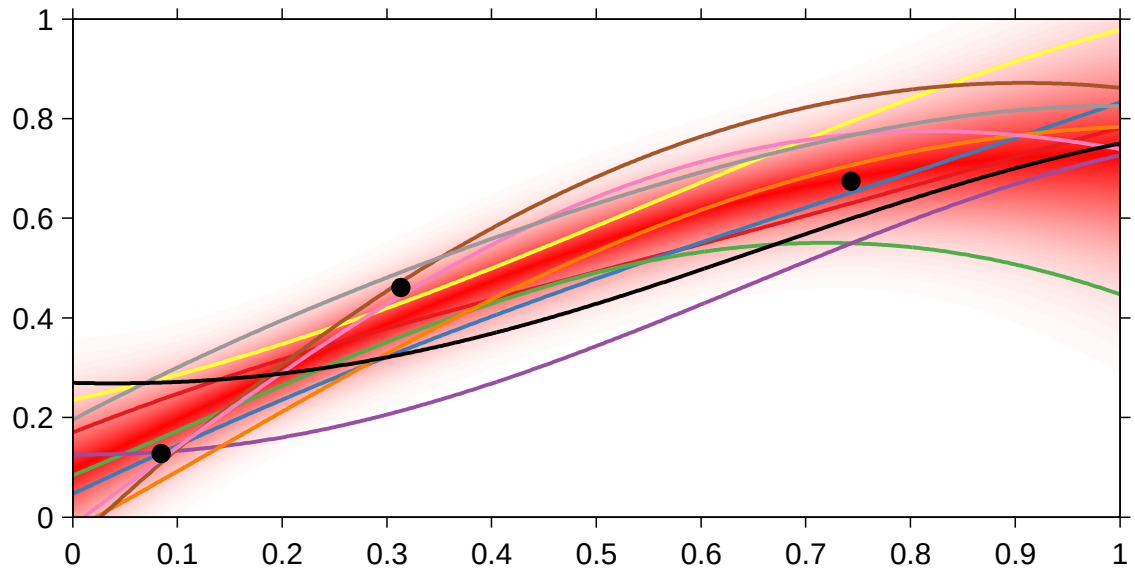
Bayesian non-linear regression



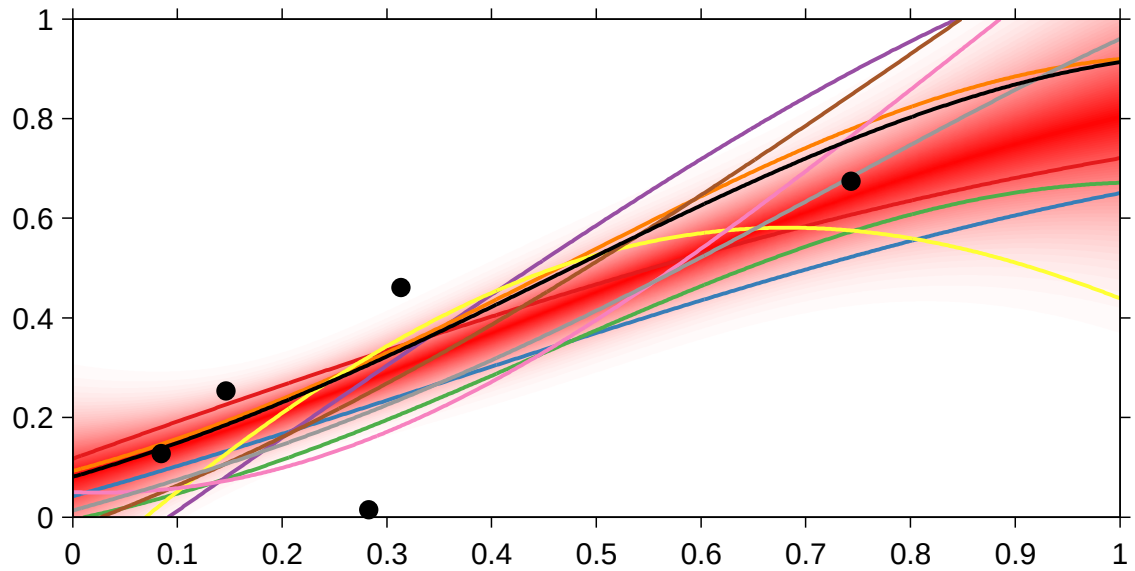
Bayesian non-linear regression



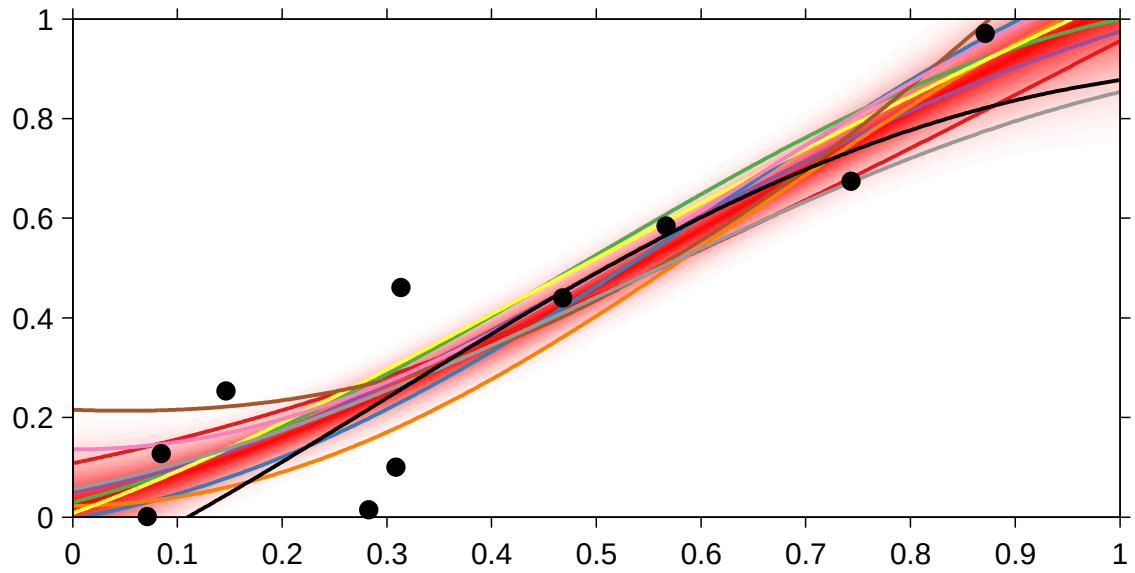
Bayesian non-linear regression



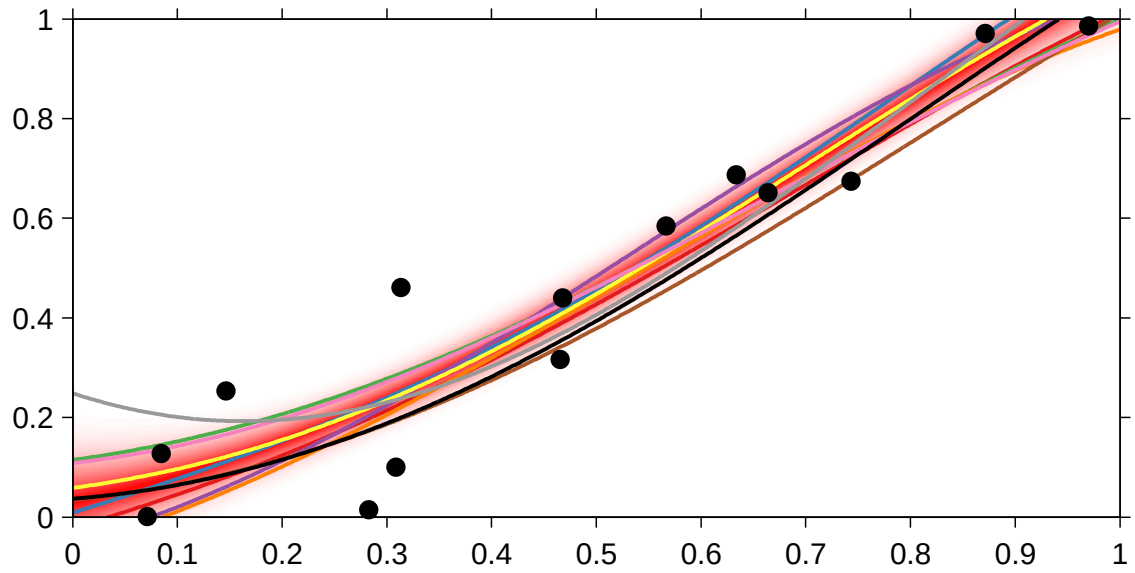
Bayesian non-linear regression



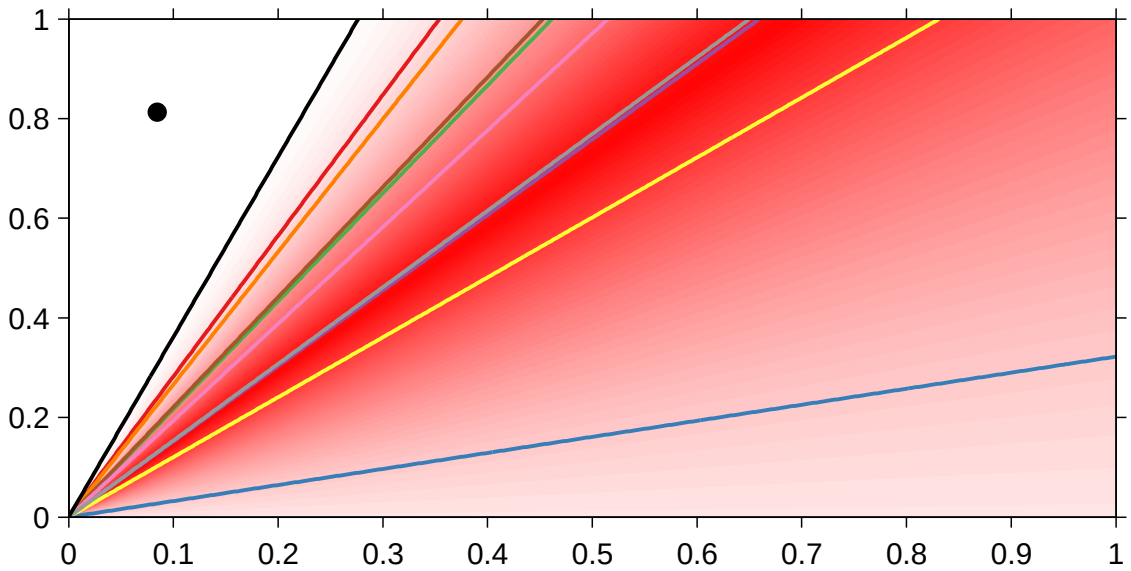
Bayesian non-linear regression



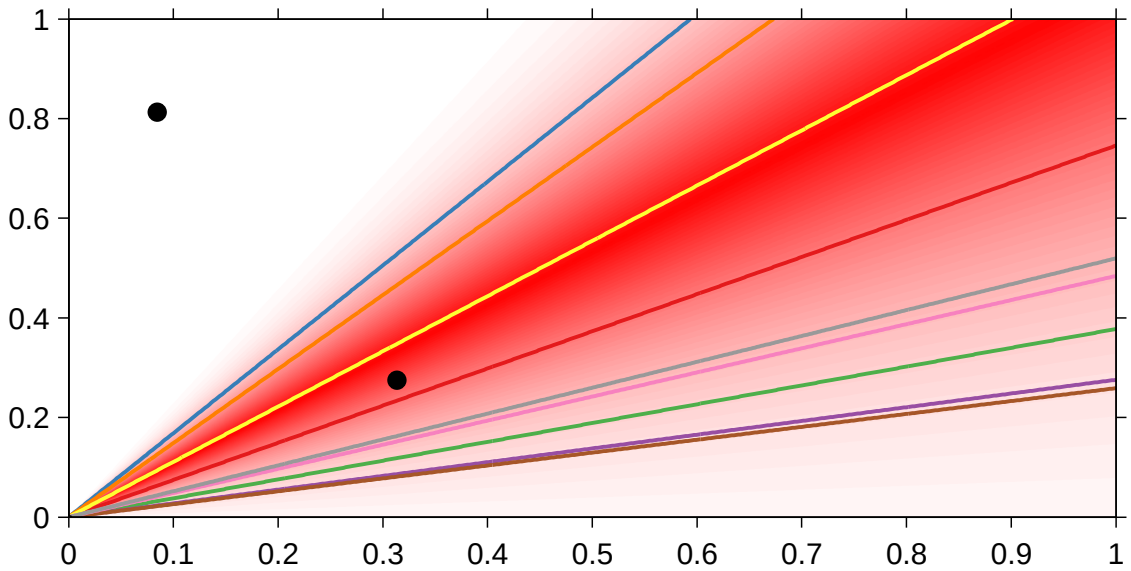
Bayesian non-linear regression



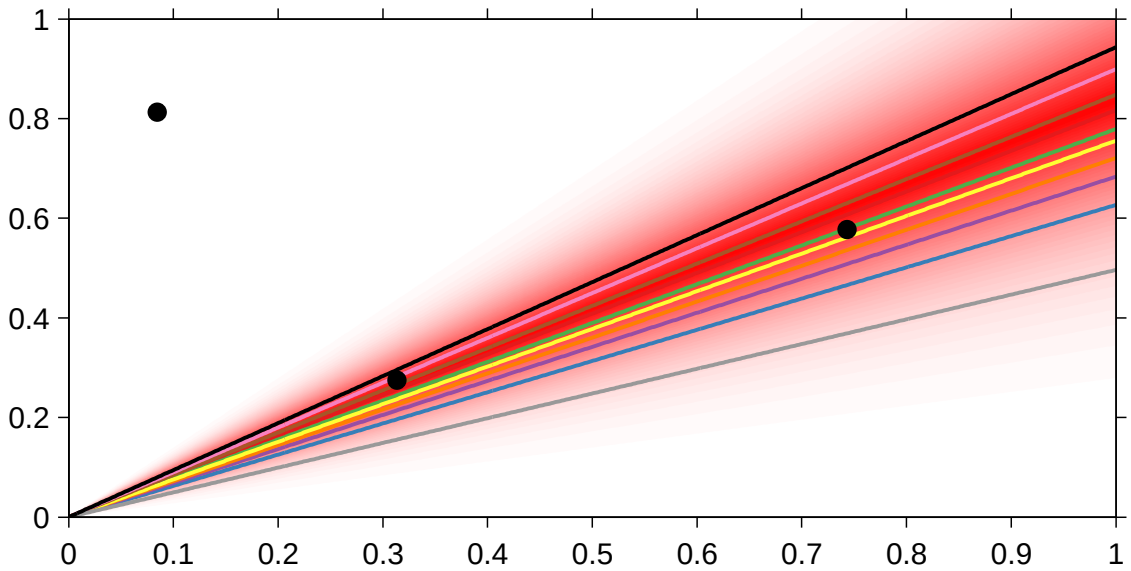
Bayesian linear regression gone wrong



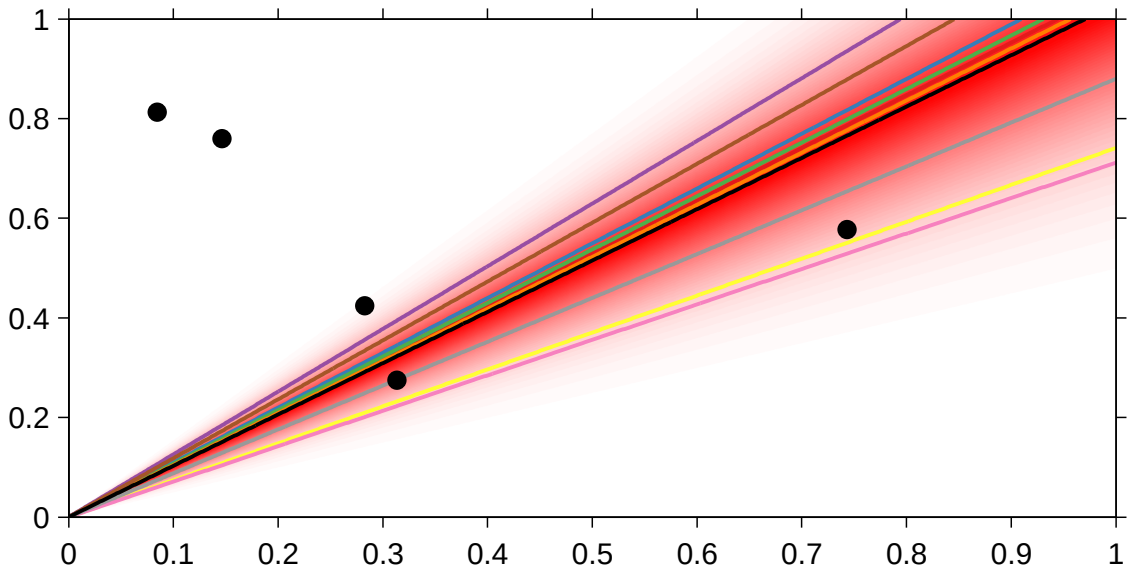
Bayesian linear regression gone wrong



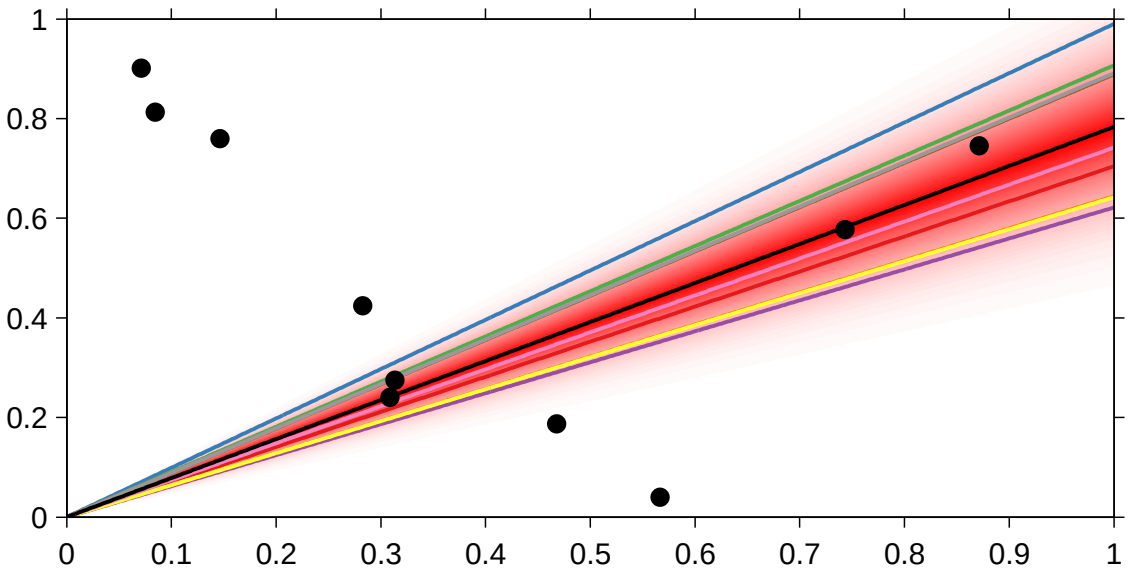
Bayesian linear regression gone wrong



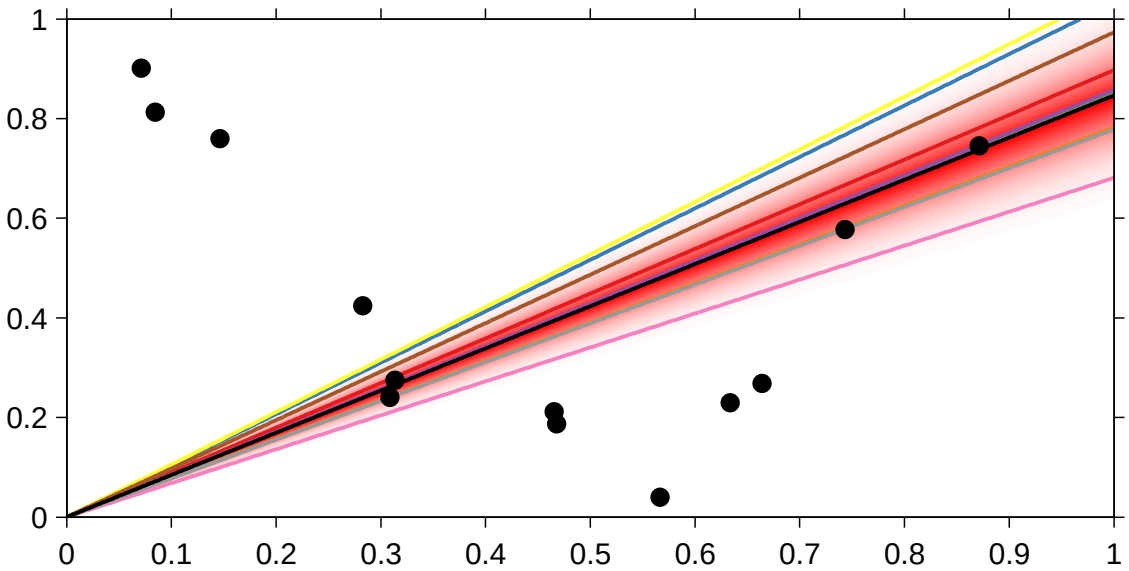
Bayesian linear regression gone wrong



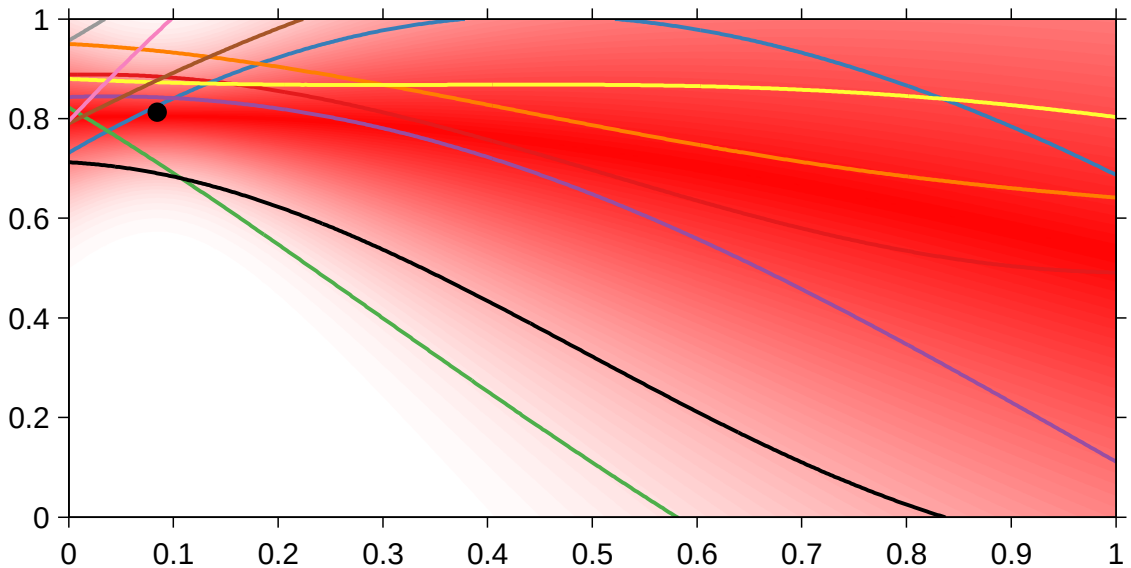
Bayesian linear regression gone wrong



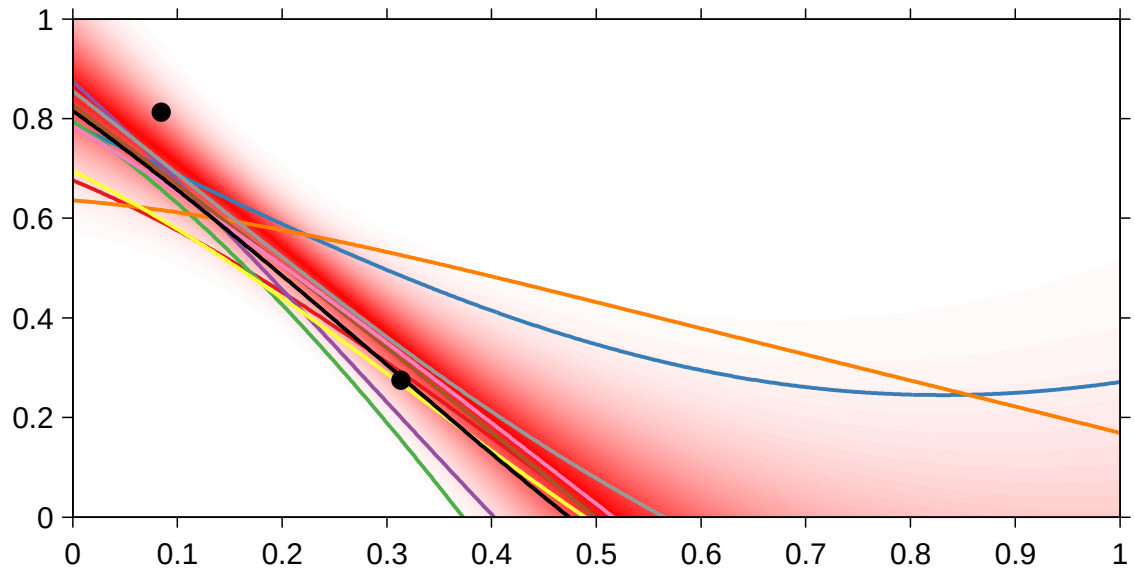
Bayesian linear regression gone wrong



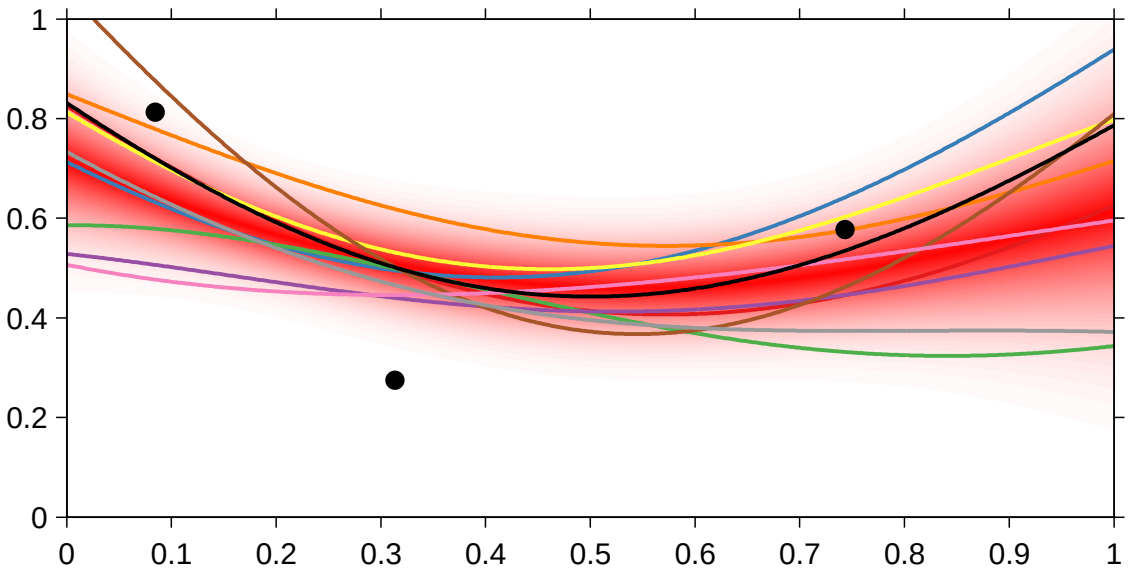
Non-linearity to the rescue



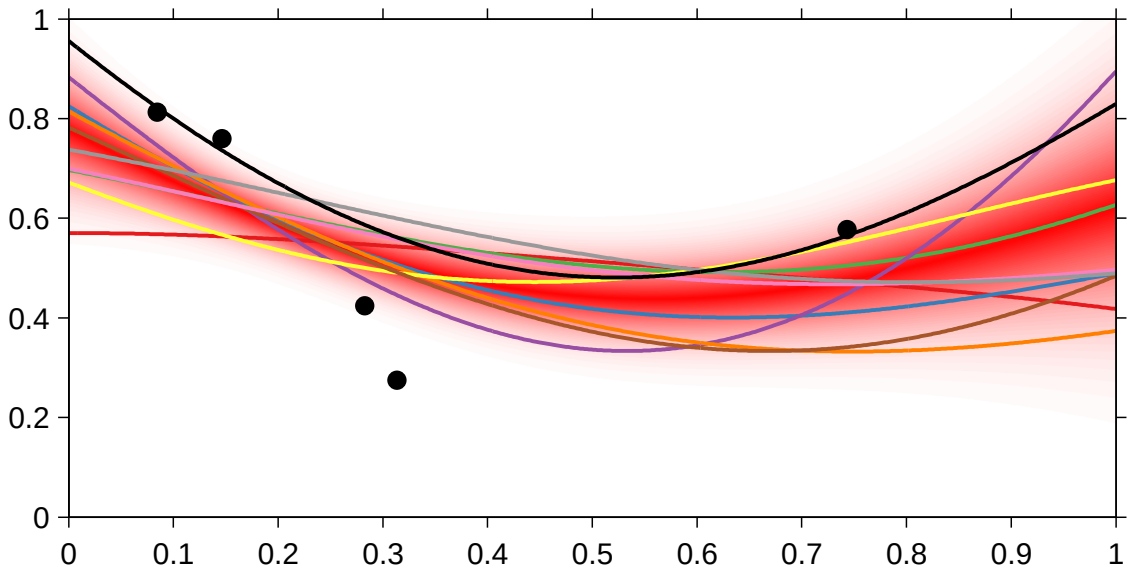
Non-linearity to the rescue



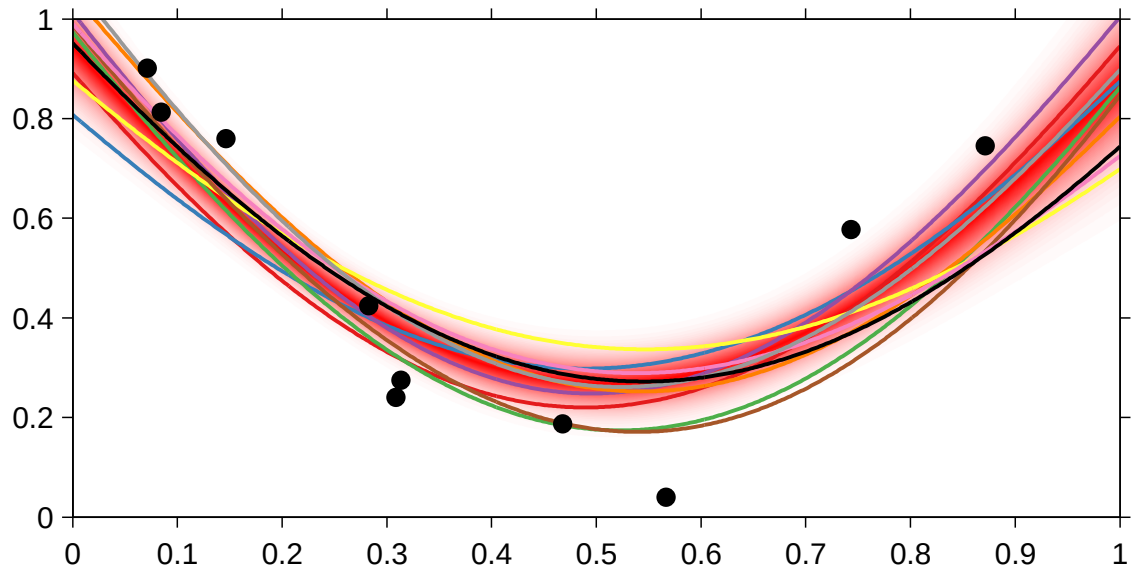
Non-linearity to the rescue



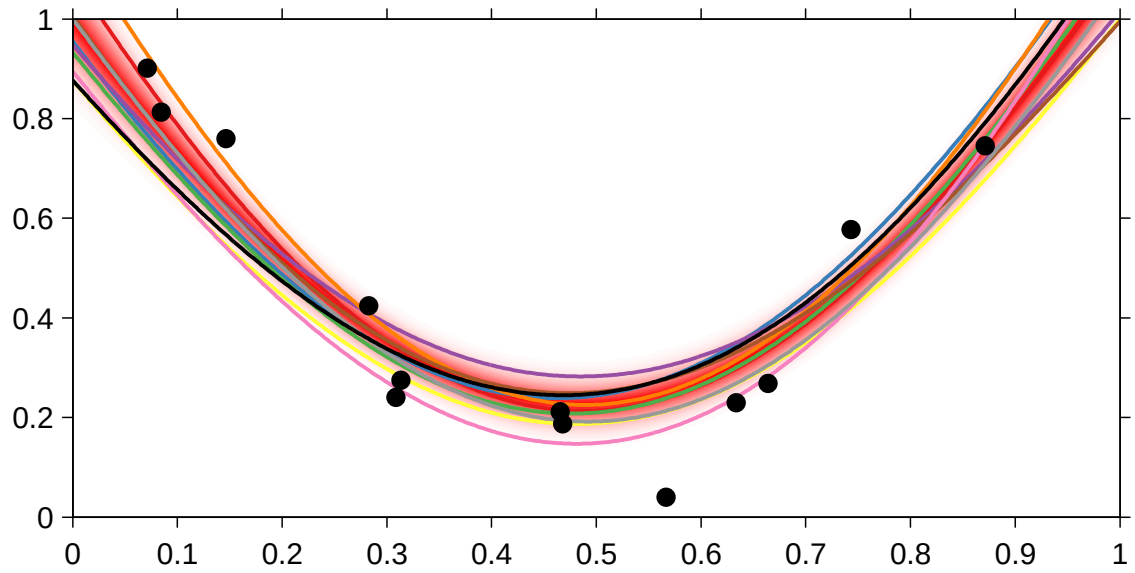
Non-linearity to the rescue



Non-linearity to the rescue



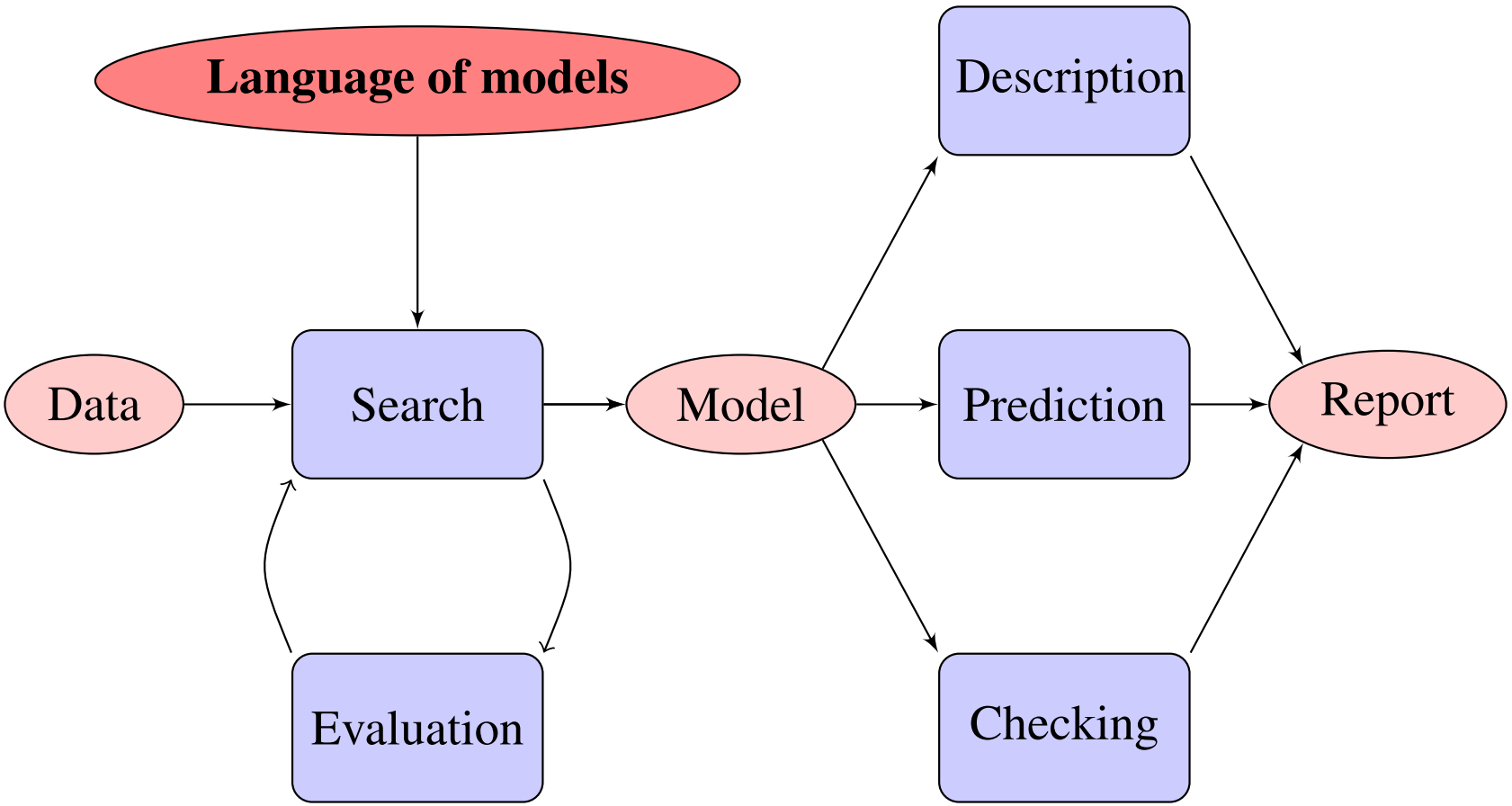
Non-linearity to the rescue



WHICH KERNEL FUNCTION?

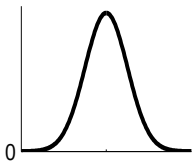
How far can we go with kernel functions?

Defining a language of models

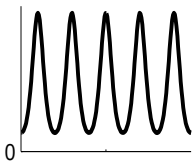


THE ATOMS OF OUR LANGUAGE

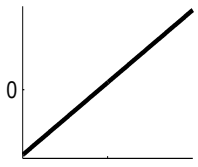
Five base kernels...



Squared exp. (SE)

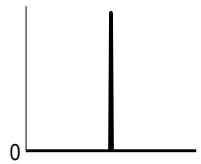


Periodic (PER)



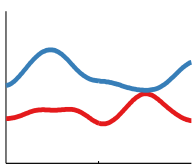
Linear (LIN)

Constant (C)

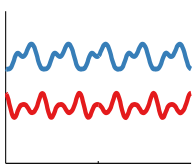


White noise (WN)

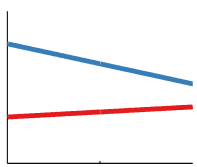
...encoding for the following types of functions



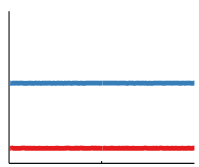
Smooth functions



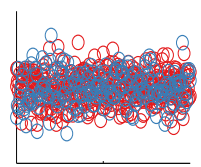
Periodic functions



Linear functions



Constant functions

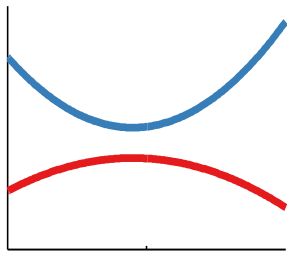


Gaussian noise

THE COMPOSITION RULES OF OUR LANGUAGE

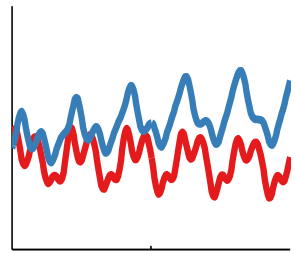
- Two main operations: addition, multiplication

LIN × LIN



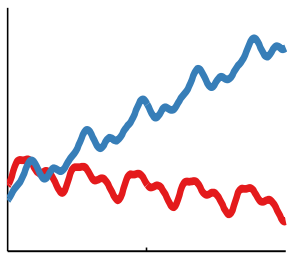
quadratic functions

SE × PER



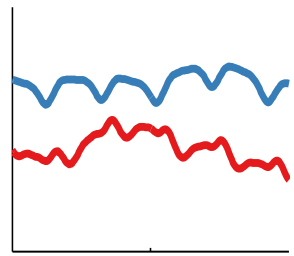
locally periodic

LIN + PER



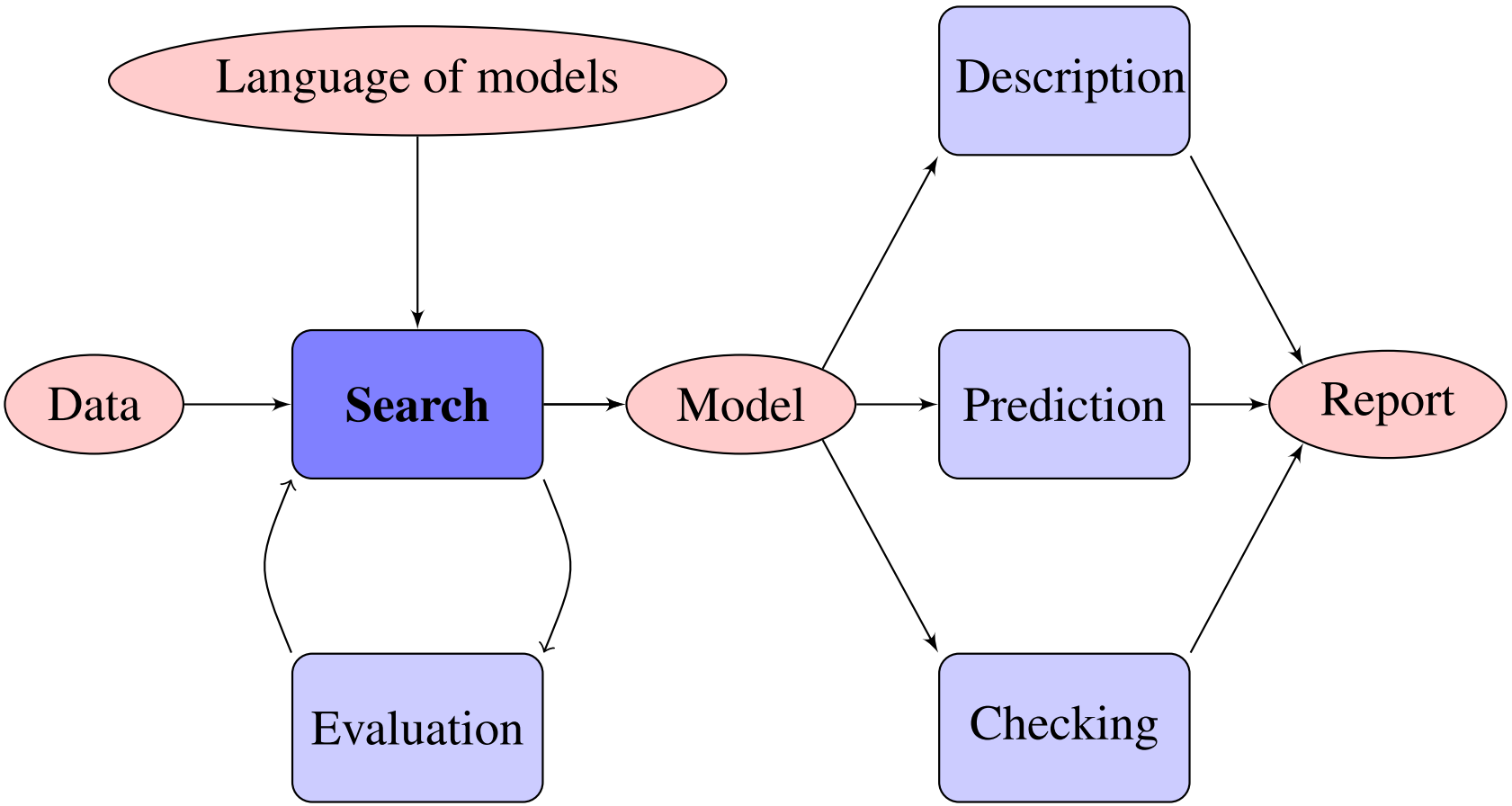
periodic plus linear trend

SE + PER



periodic plus smooth trend

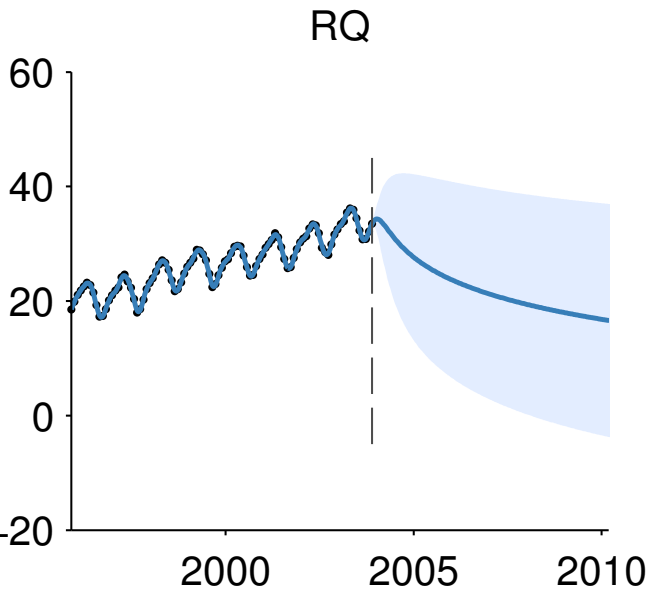
Discovering a good model via search

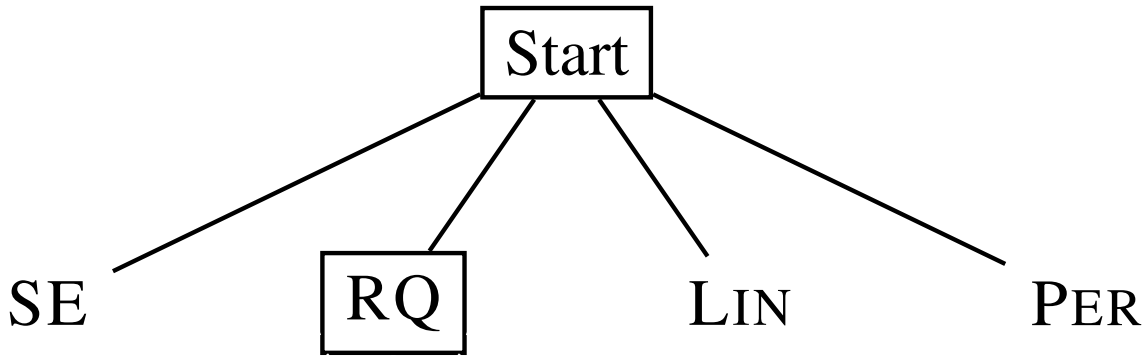


DISCOVERING A GOOD MODEL VIA SEARCH

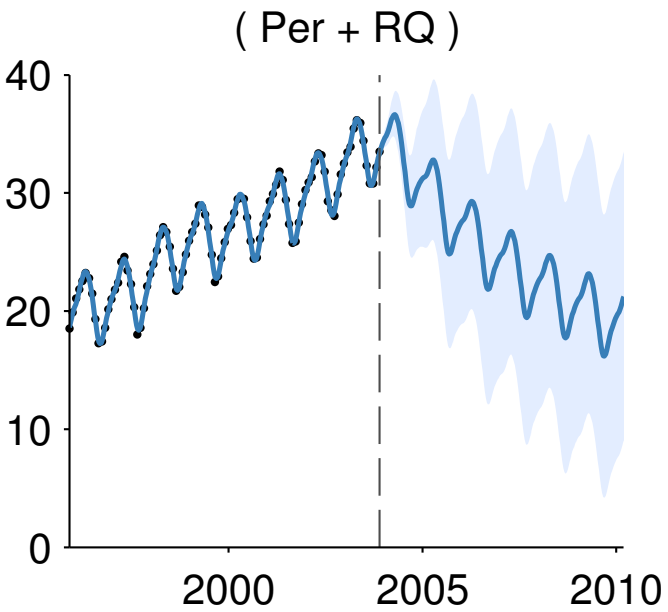
- Language defined as the arbitrary composition of five base kernels (WN, C, LIN, SE, PER) via three operators ($+$, $\times$, CP).
- The space spanned by this language is open-ended
- We use a greedy search - simple and similar to human model-building

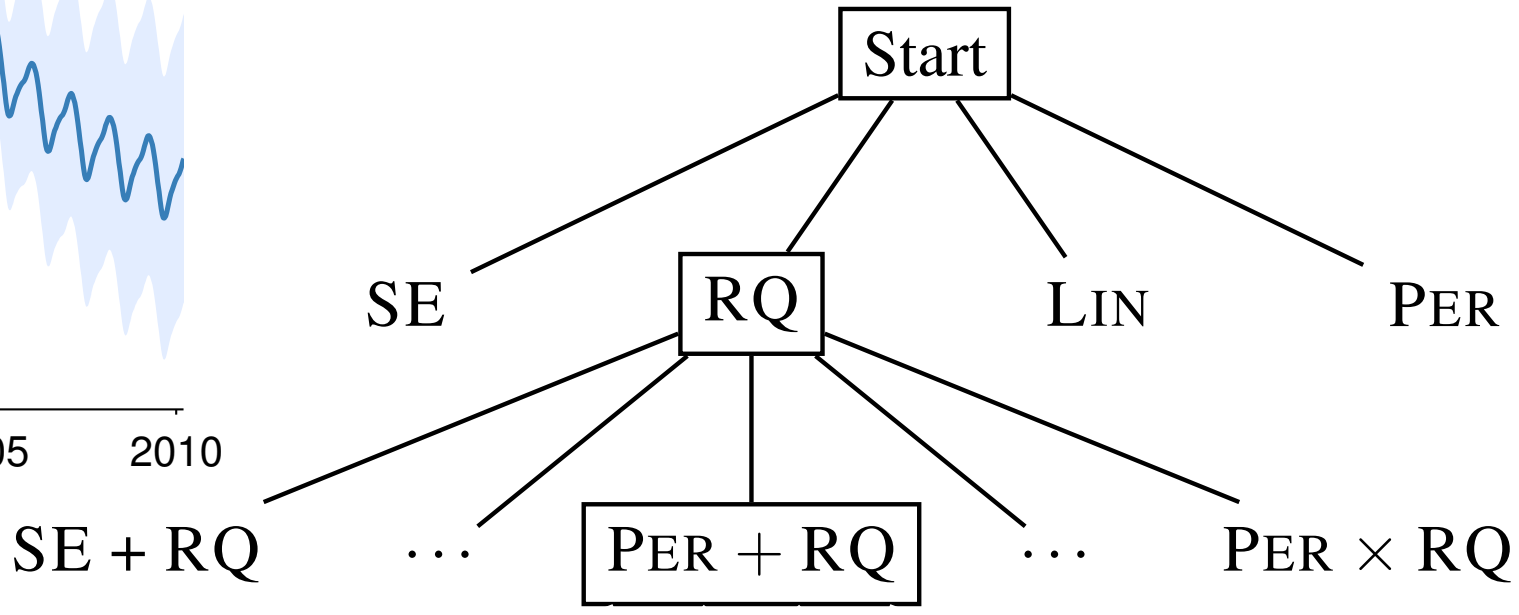
EXAMPLE: MAUNA LOA CO_2 CURVE



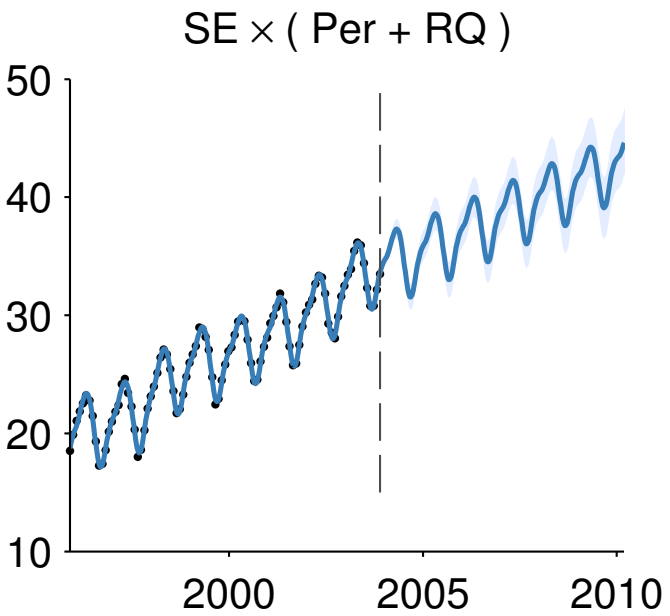


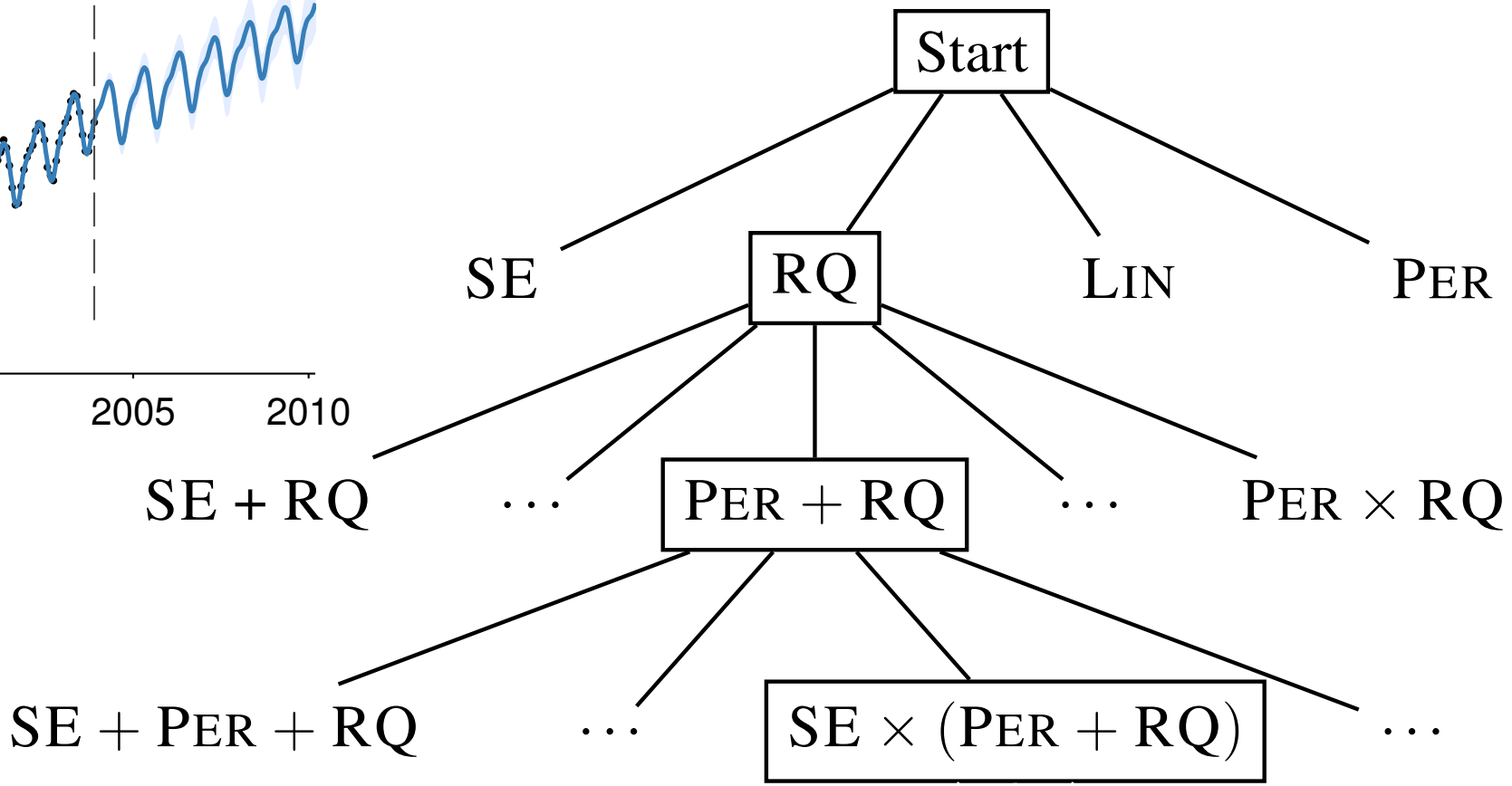
EXAMPLE: MAUNA LOA CO_2 CURVE



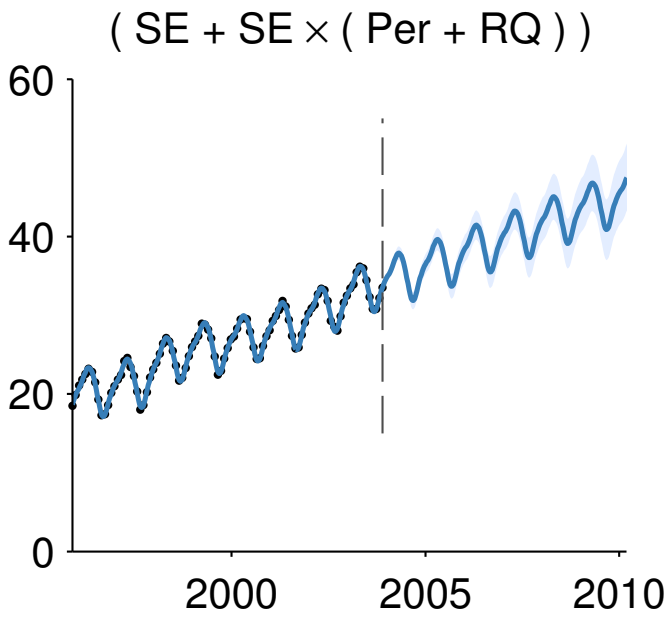


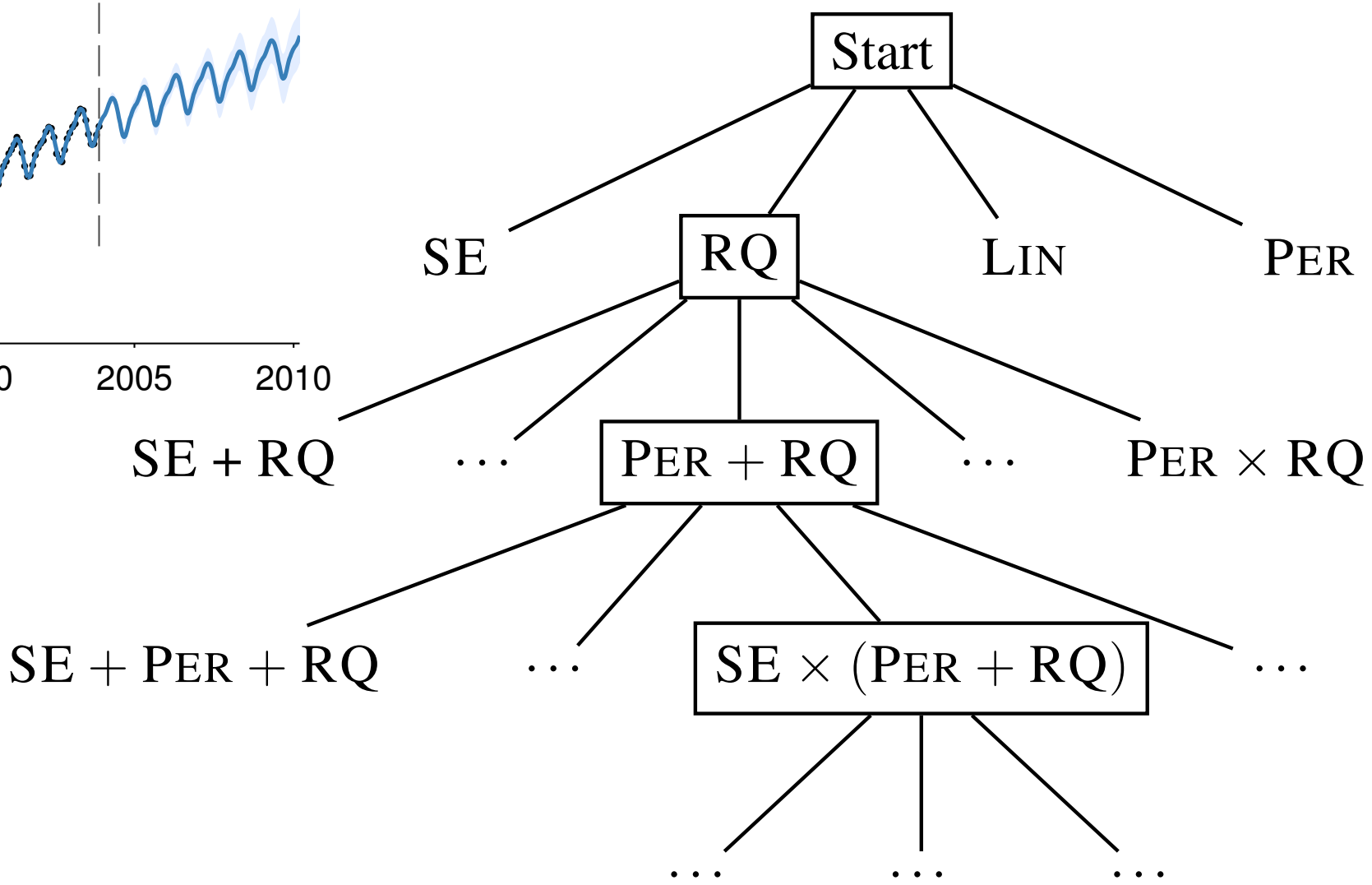
EXAMPLE: MAUNA LOA CO_2 CURVE



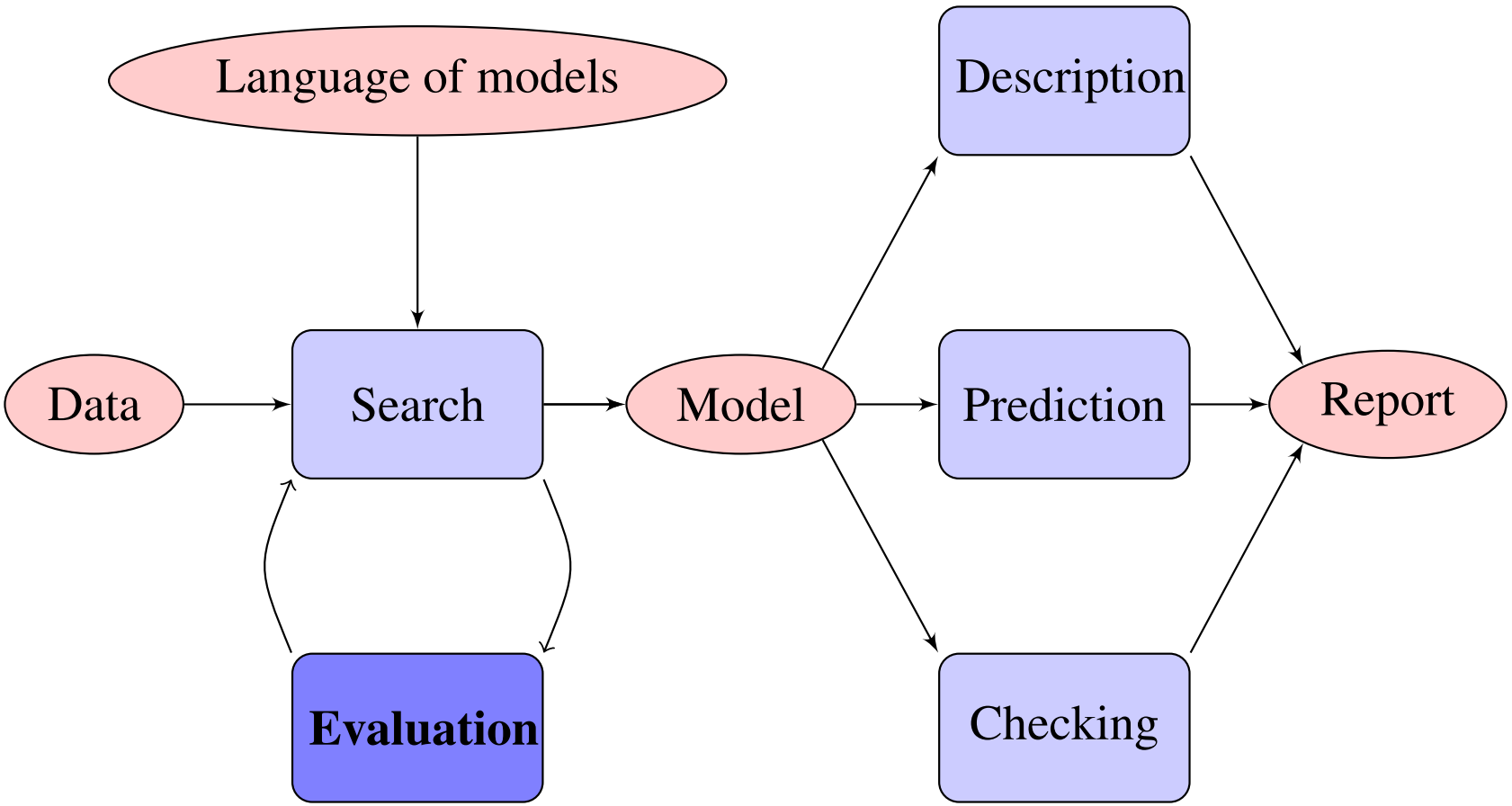


EXAMPLE: MAUNA LOA CO_2 CURVE





Model evaluation



BAYESIAN MODEL SELECTION

Suppose we have a collection of models $\{M_i\}$ and some data D

Bayes rule tells us

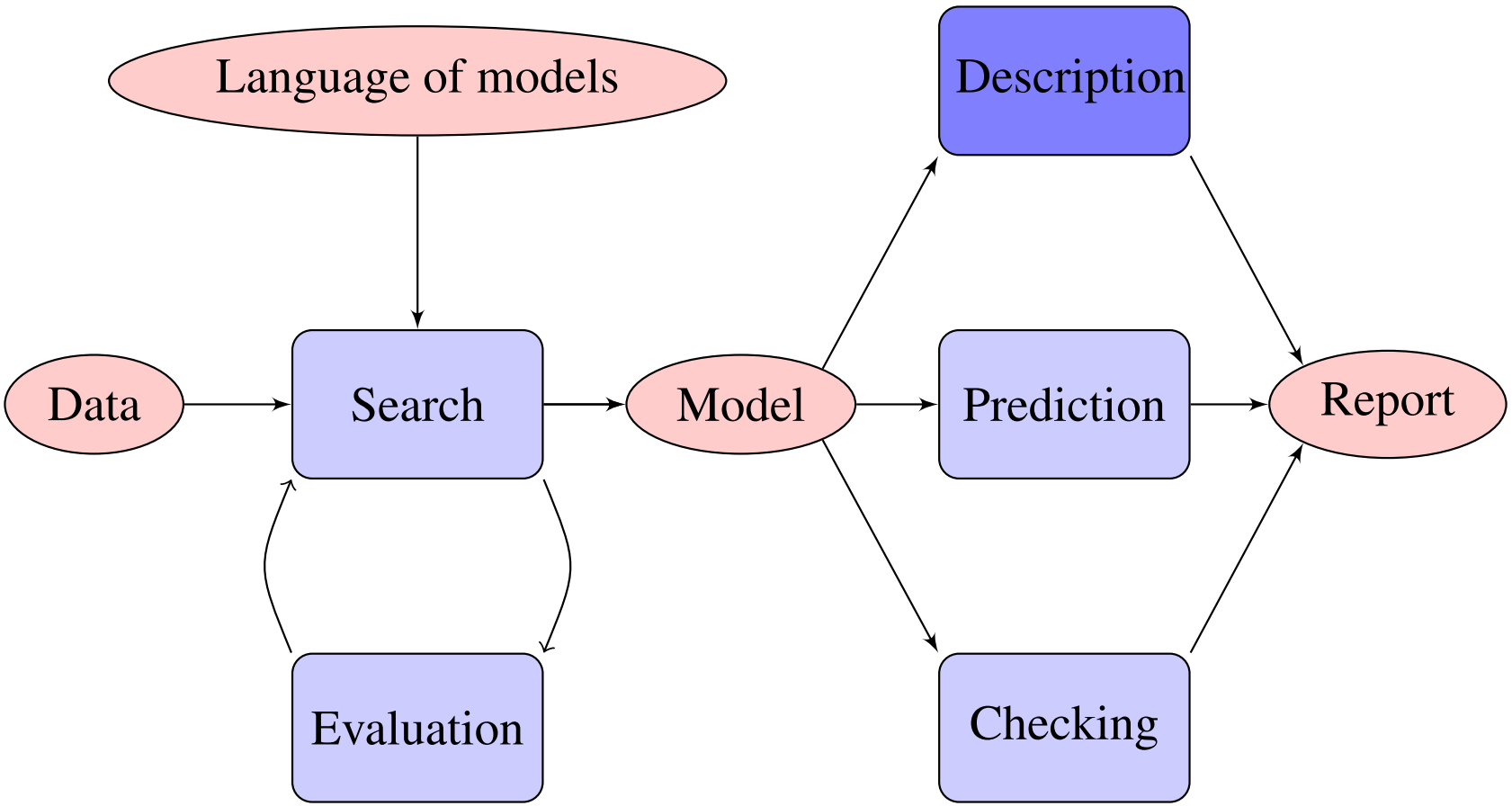
$$p(M_i \mid D) = \frac{p(D \mid M_i)p(M_i)}{p(D)}$$

If $p(M_i)$ is equal for all i (prior ignorance) then

$$p(M_i \mid D) \propto p(D \mid M_i) = \int p(D \mid \theta_i, M_i)p(\theta_i \mid M_i)\mathrm{d}\theta_i$$

i.e. The most likely model has the highest marginal likelihood

AUTOMATIC TRANSLATION OF MODELS



AUTOMATIC TRANSLATION OF MODELS

- Search can produce arbitrarily complicated models from open-ended language but two main properties allow description to be automated
- Kernels can be decomposed into a sum of products
  - A sum of kernels corresponds to a sum of functions
  - Therefore, we can describe each product of kernels separately
- Each kernel in a product modifies a model in a consistent way
  - Each kernel roughly corresponds to an adjective

SIMPLIFYING KERNELS

Suppose the search finds the following kernel

$$\text{SE} \times (\text{PER} + \text{RQ})$$

Multiplication can be distributed over addition

$$\text{SE} \times \text{PER} + \text{SE} \times \text{RQ}$$

SUMS OF KERNELS ARE SUMS OF FUNCTIONS

If $f_1 \sim \mathrm{GP}(0, k_1)$ and independently $f_2 \sim \mathrm{GP}(0, k_2)$ then

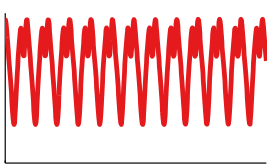
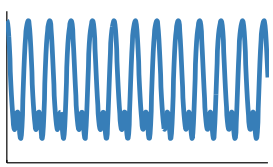
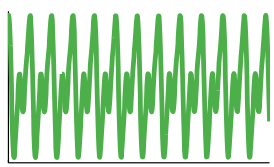
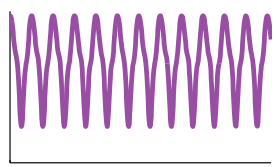
$$f_1 + f_2 \sim \mathrm{GP}(0, k_1 + k_2)$$

We can therefore describe each component in a sum separately

PRODUCTS OF KERNELS

$$\underbrace{\text{PER}}_{\text{periodic function}}$$

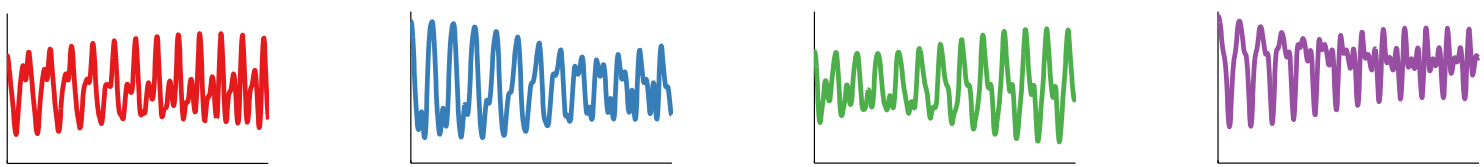
On their own, each kernel is described by a standard noun phrase

PRODUCTS OF KERNELS - SE

$$\underbrace{\text{SE}}_{\text{approximately}} \times \underbrace{\text{PER}}_{\text{periodic function}}$$

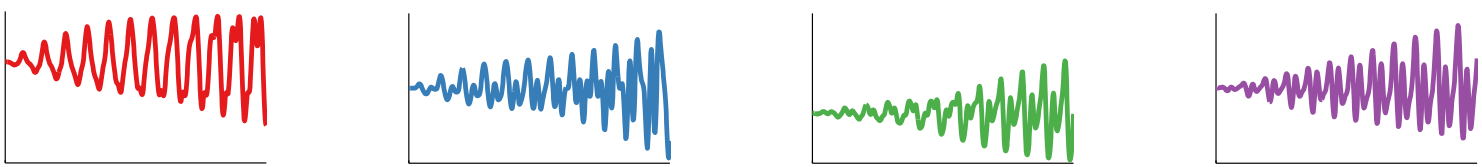
Multiplication by SE removes long range correlations from a model since $\text{SE}(x, x')$ decreases monotonically to 0 as $|x - x'|$ increases.



PRODUCTS OF KERNELS - LIN

$$\underbrace{\text{SE}}_{\text{approximately}} \times \underbrace{\text{PER}}_{\text{periodic function}} \times \underbrace{\text{LIN}}_{\text{with linearly growing amplitude}}$$

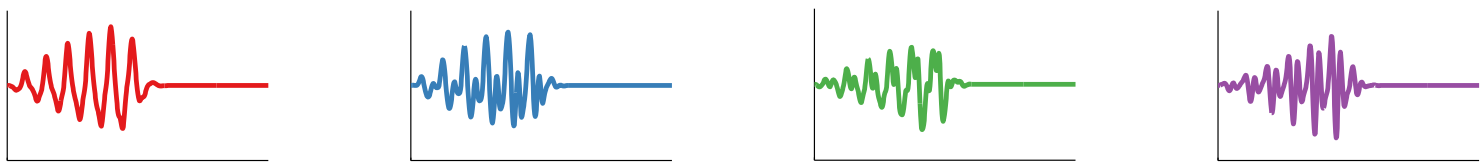
Multiplication by LIN is equivalent to multiplying the function being modeled by a linear function. If $f(x) \sim \text{GP}(0, k)$, then $xf(x) \sim \text{GP}(0, k \times \text{LIN})$. This causes the standard deviation of the model to vary linearly without affecting the correlation.



PRODUCTS OF KERNELS - CHANGEPOINTS

$$\underbrace{\text{SE}}_{\text{approximately}} \times \underbrace{\text{PER}}_{\text{periodic function}} \times \underbrace{\text{LIN}}_{\text{with linearly growing amplitude}} \times \underbrace{\sigma}_{\text{until 1700}}$$

Multiplication by σ is equivalent to multiplying the function being modeled by a sigmoid.



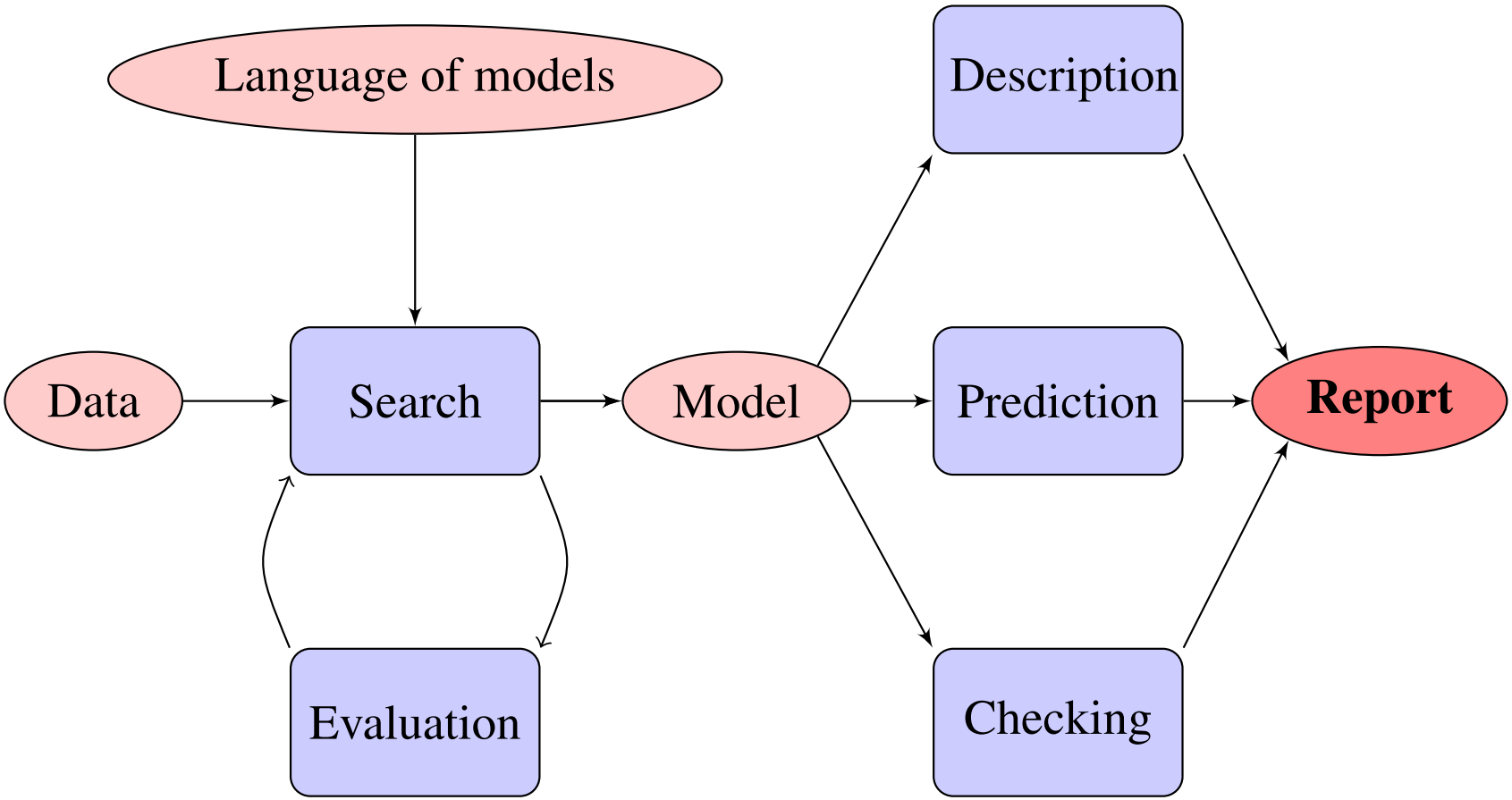
NOUN PHRASE FORMS OF KERNELS

Kernel	Noun phrase
WN	uncorrelated noise
C	constant
SE	smooth function
PER	periodic function
LIN	linear function
$\prod_k \text{LIN}^{(k)}$	polynomial

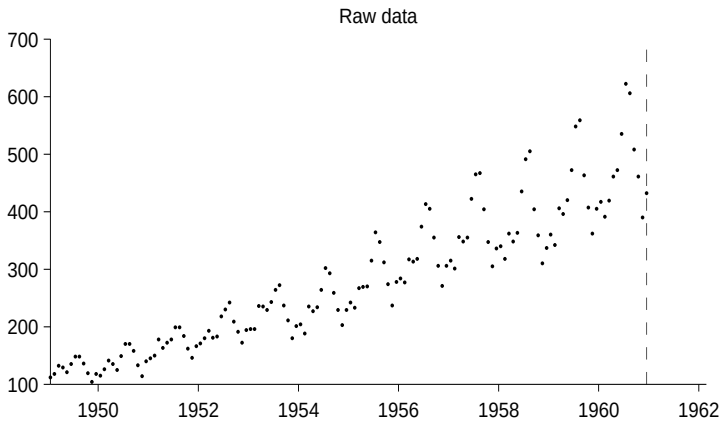
POSTMODIFIER FORM OF KERNELS

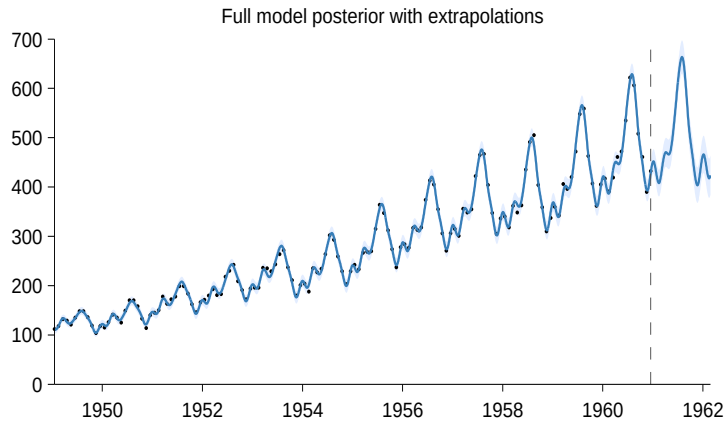
Kernel	Postmodifier phrase
SE	whose shape changes smoothly
PER	modulated by a periodic function
LIN	with linearly varying amplitude
$\prod_k \text{LIN}^{(k)}$	with polynomially varying amplitude
$\prod_k \boldsymbol{\sigma}^{(k)}$	which applies until / from [changepoint]

AUTOMATICALLY GENERATED REPORTS



EXAMPLE: AIRLINE PASSENGER VOLUME



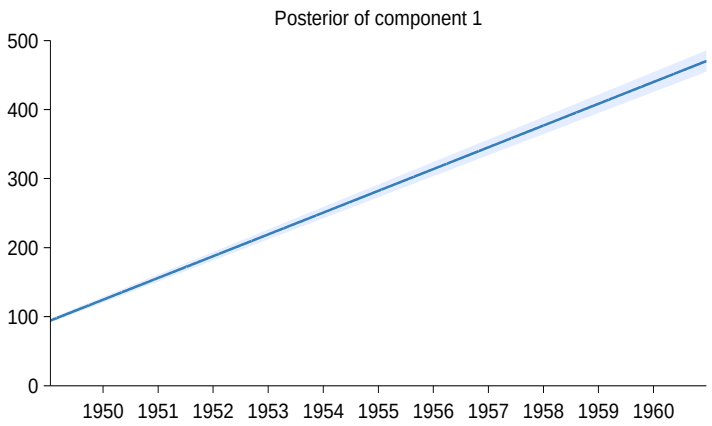


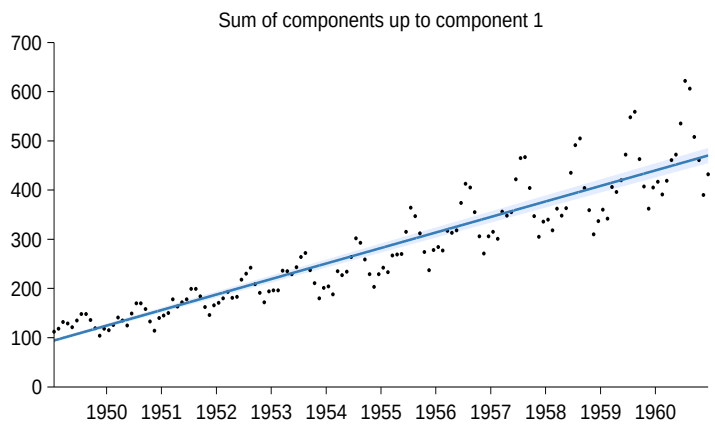
Four additive components have been identified in the data

- A linearly increasing function.
- An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.

EXAMPLE: AIRLINE PASSENGER VOLUME

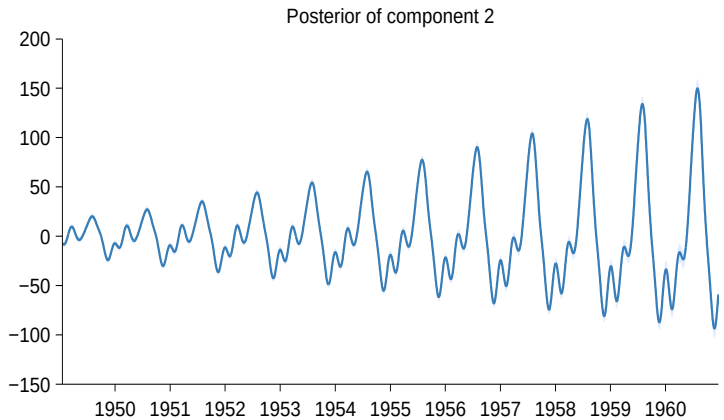
This component is linearly increasing.

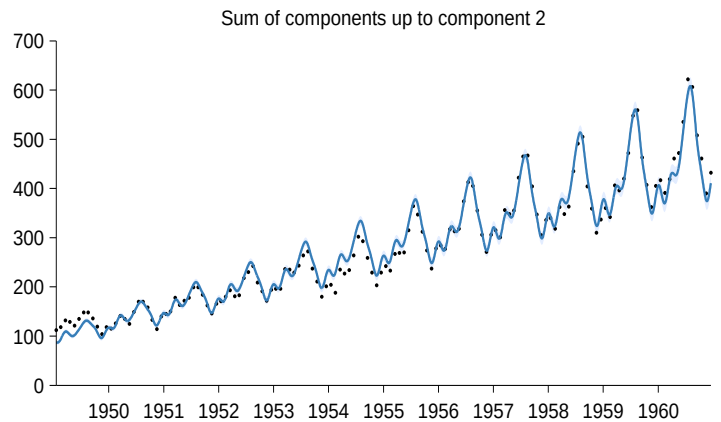




Example: Airline passenger volume

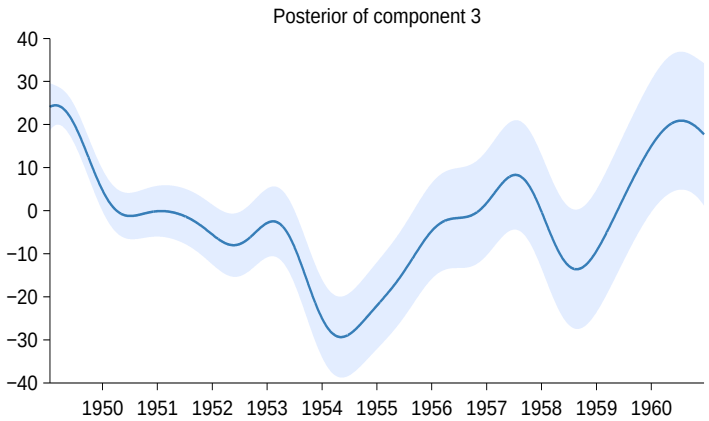
This component is approximately periodic with a period of 1.0 years and varying amplitude. Across periods the shape of this function varies very smoothly. The amplitude of the function increases linearly. The shape of this function within each period has a typical lengthscale of 6.0 weeks.

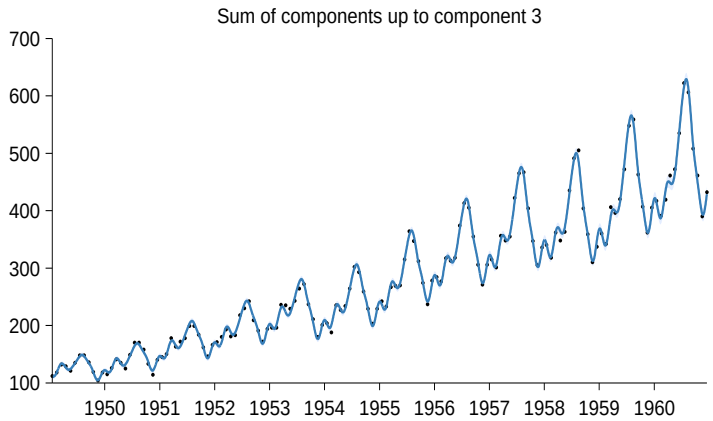




Example: Airline passenger volume

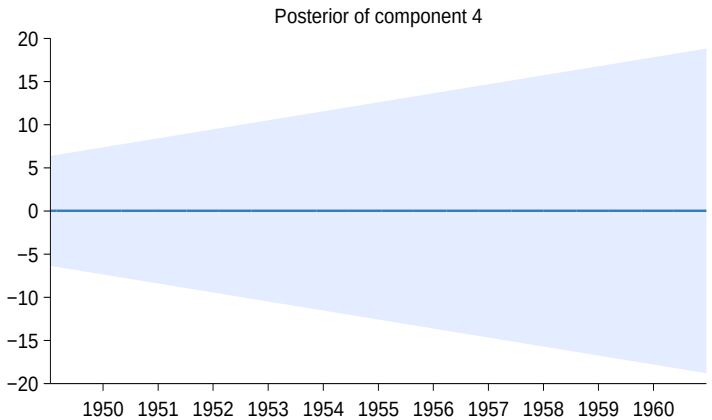
This component is a smooth function with a typical lengthscale of 8.1 months.

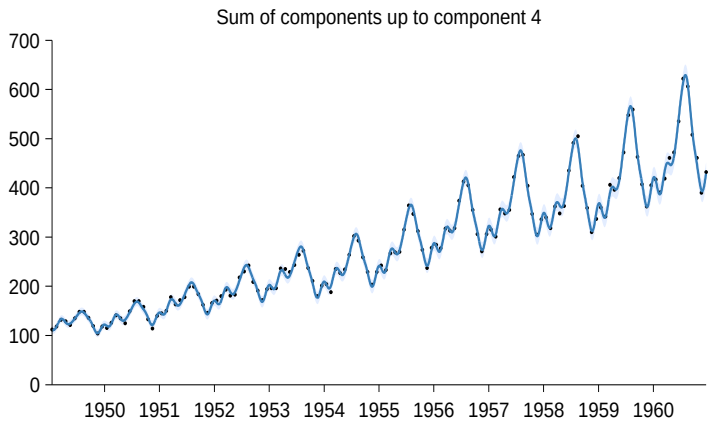




Example: Airline passenger volume

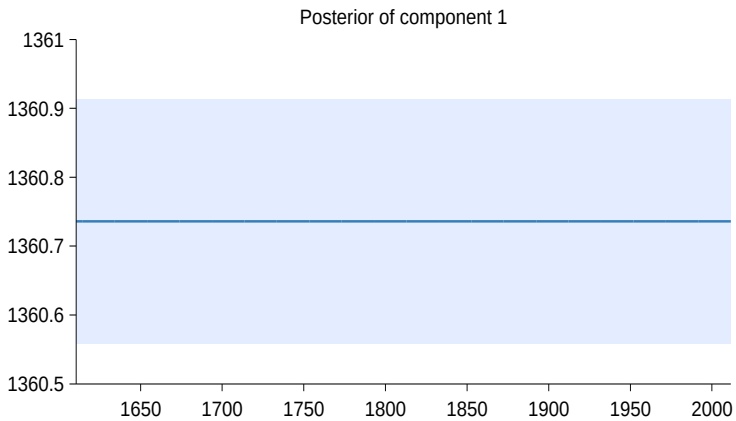
This component models uncorrelated noise. The standard deviation of the noise increases linearly.

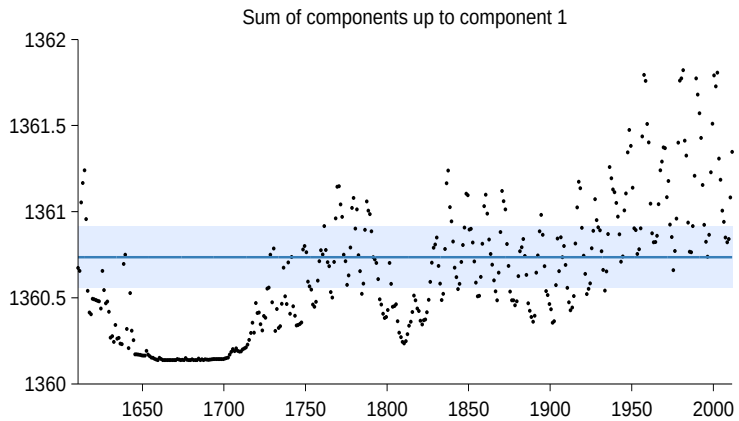




EXAMPLE: SOLAR IRRADIANCE

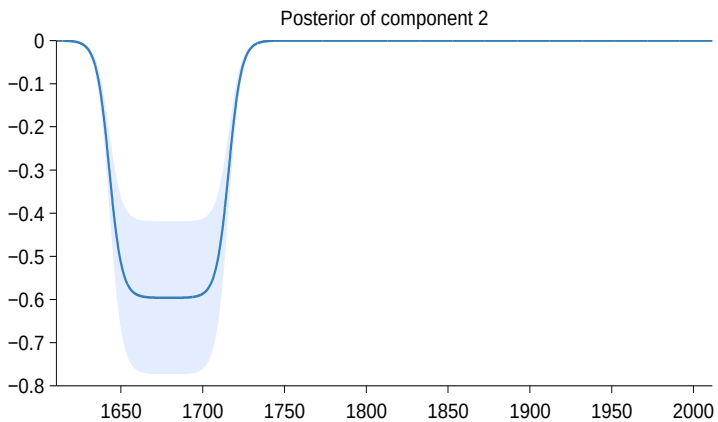
This component is constant.

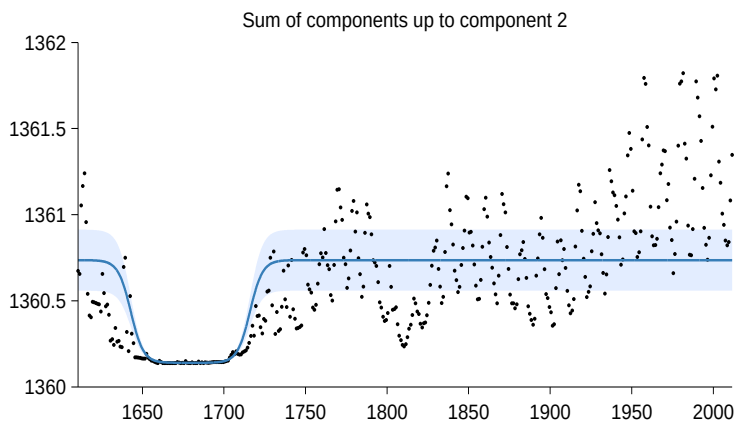




EXAMPLE: SOLAR IRRADIANCE

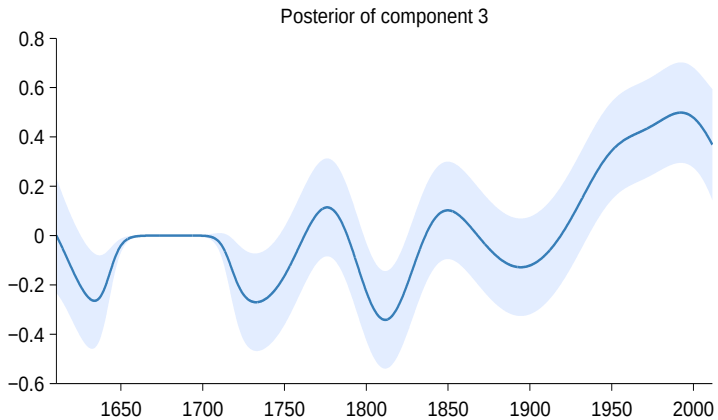
This component is constant. This component applies from 1643 until 1716.

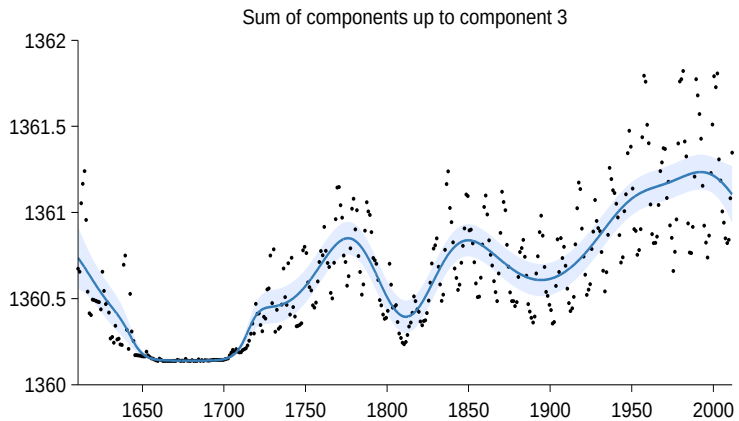




Example: Solar irradiance

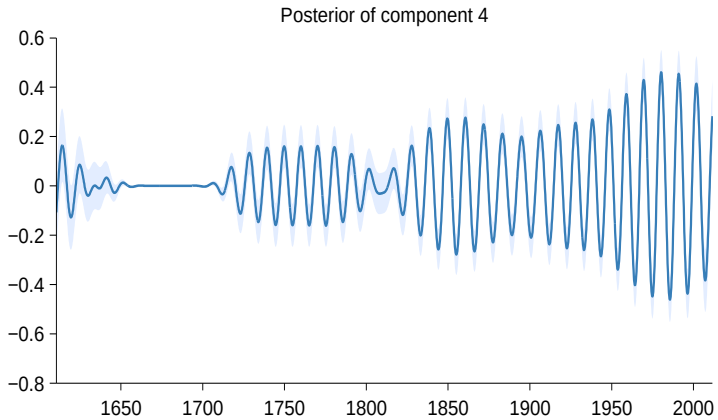
This component is a smooth function with a typical lengthscale of 23.1 years. This component applies until 1643 and from 1716 onwards.

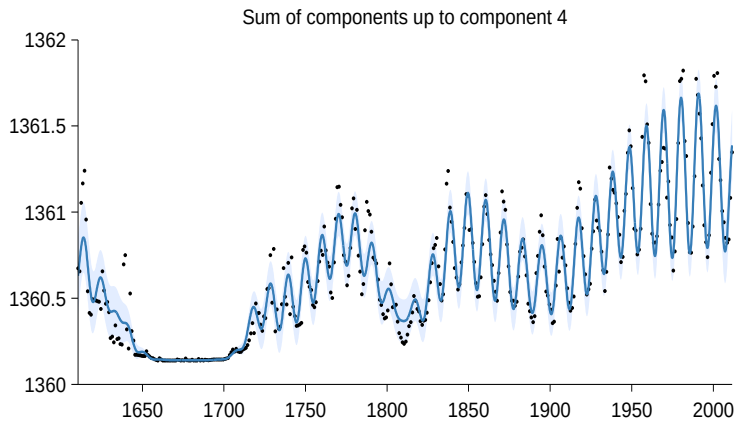




EXAMPLE: SOLAR IRRADIANCE

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.



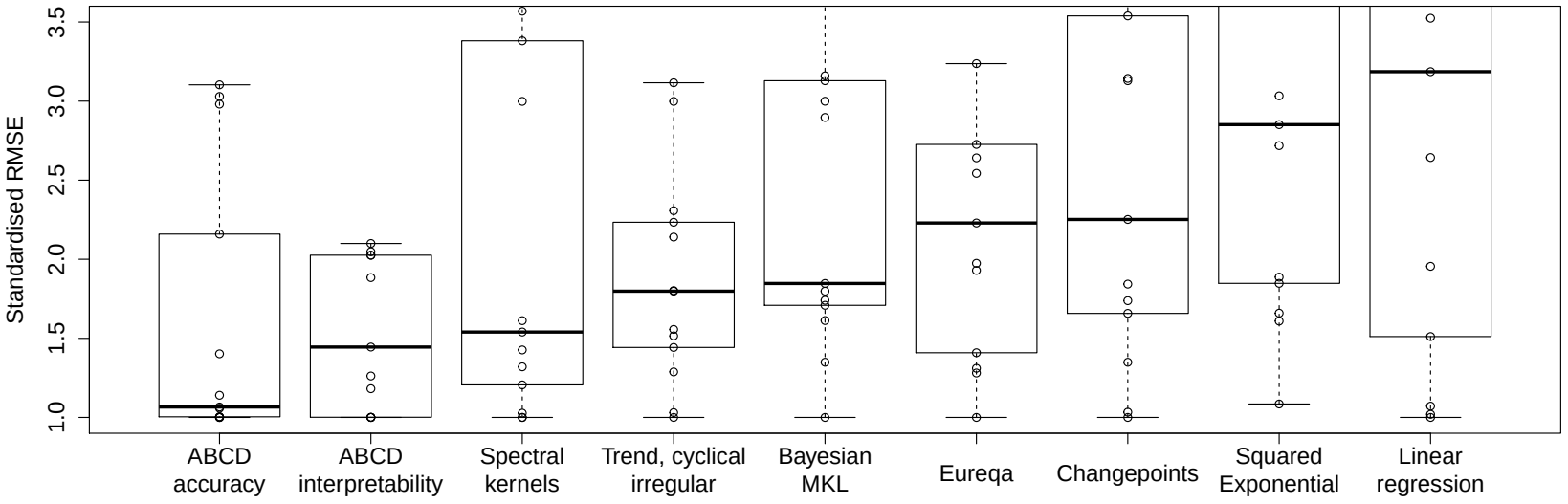


Example: Full report

See pdf

GOOD PREDICTIVE PERFORMANCE AS WELL

Standardised RMSE over 13 data sets



- This method is slow but contains most common methods as a special case

SUMMARY

How could an AI do statistics?

- Grammars over composite structures are a simple way to specify open-ended model classes.
- Composite structures sometimes give interpretable decompositions.
- Searching over these model classes is a step towards automating statistical analysis.

Thanks