PROWLER.io®
the decision company

Kernel Selection Methods

Fergus Simpson

20/02/2020

Overview

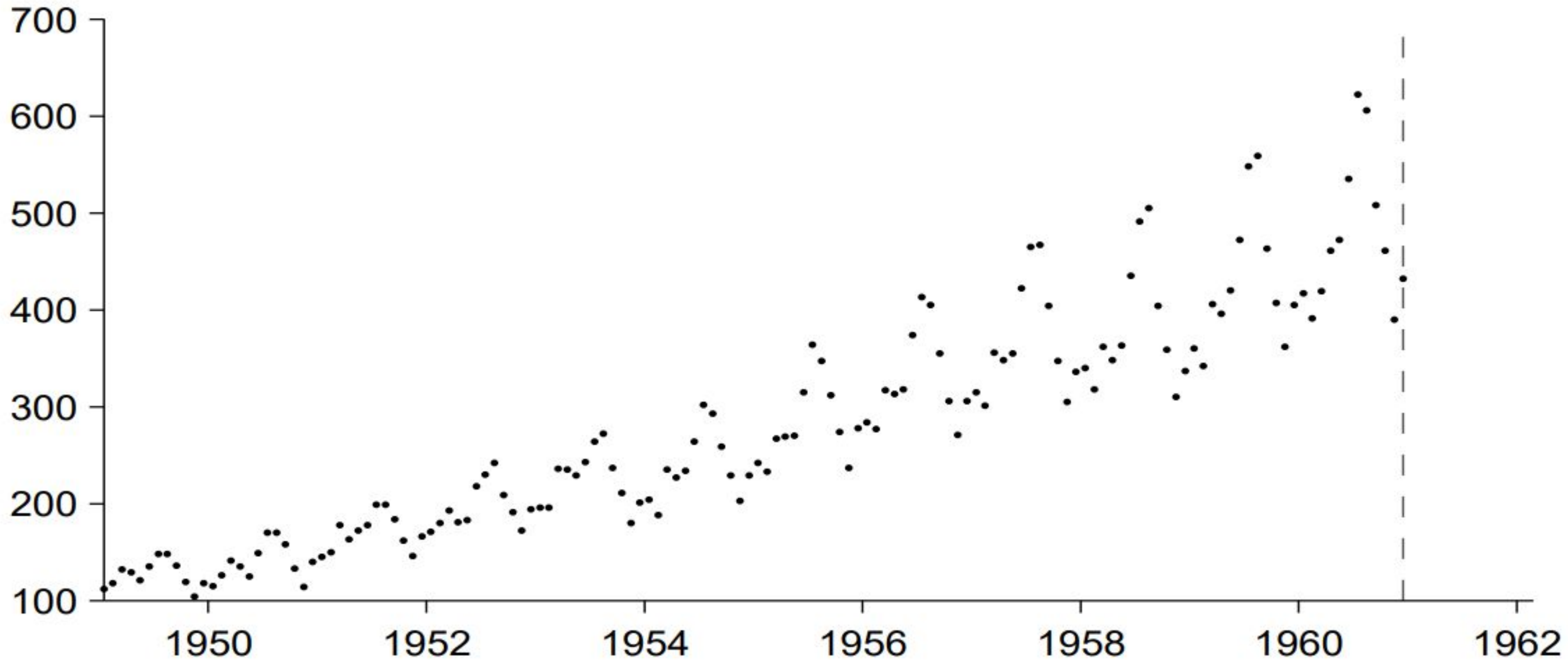
What is a kernel?

- Primitive kernels
- How to combine kernels

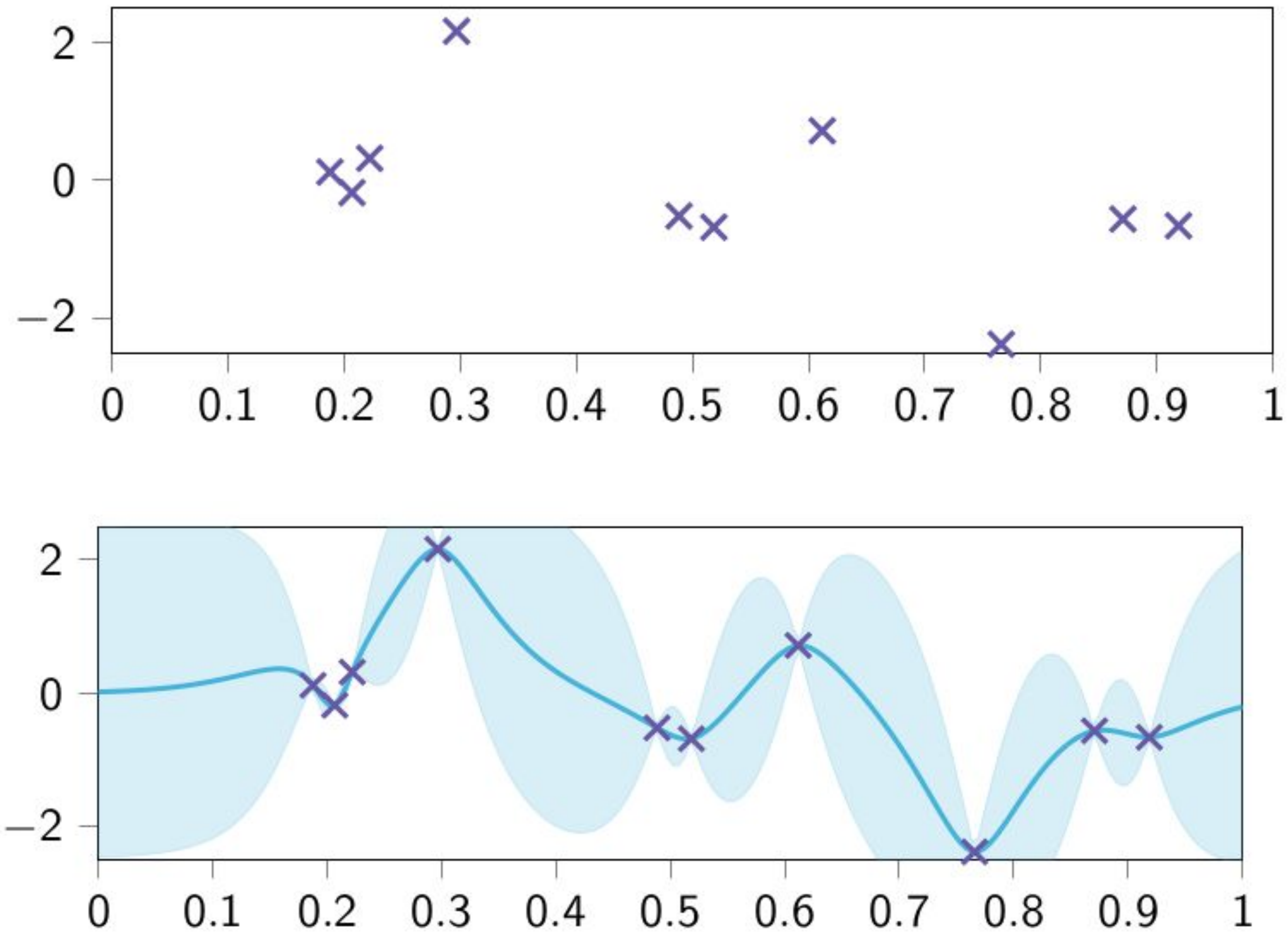
Searching for the right kernel

- Hand crafting
- The Automatic Statistician
- Spectral Kernels
- The Neural Kernel Network

Objective



Gaussian Process Regression



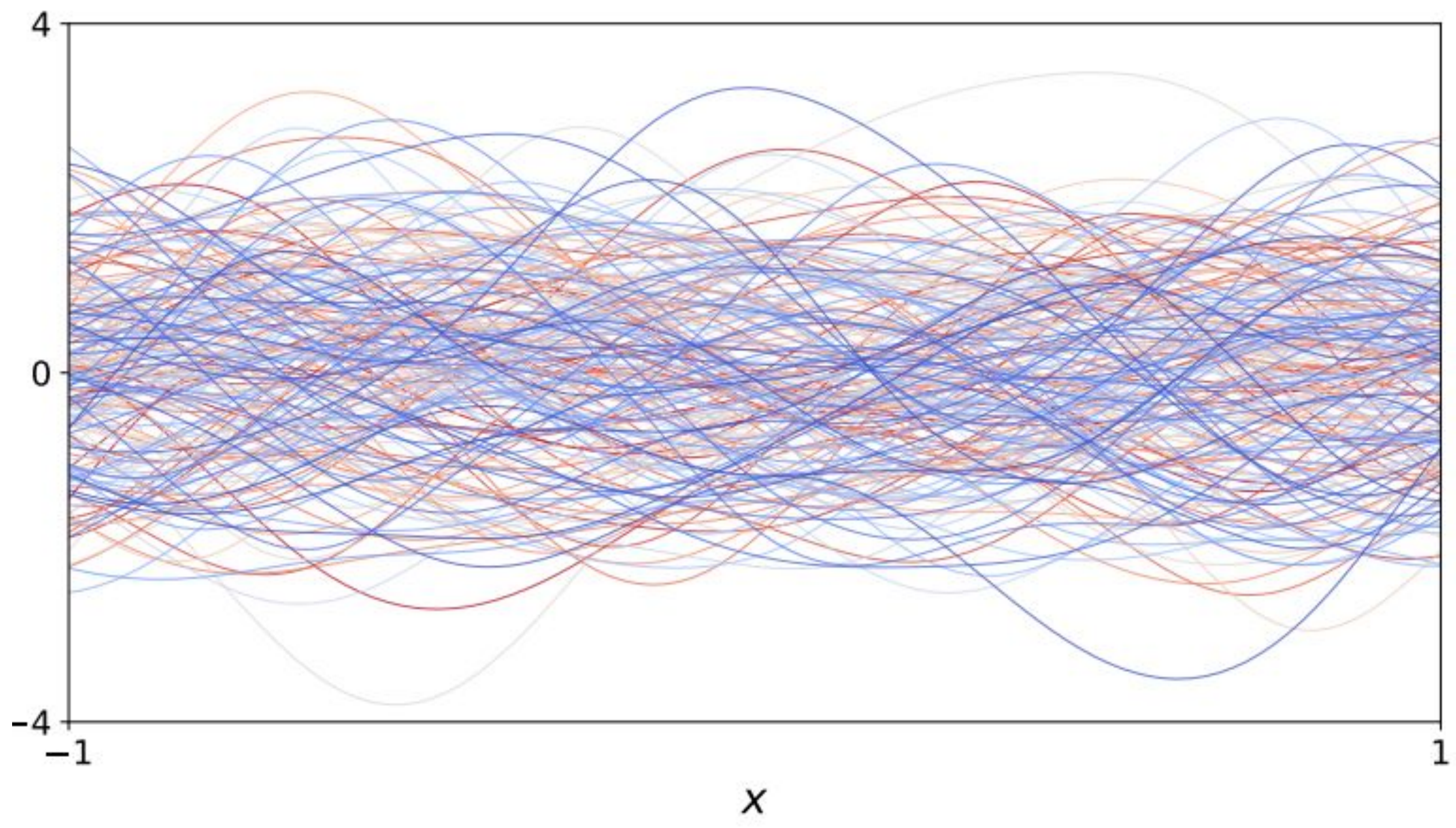
$$p(f(\cdot)|\mathcal{D}) = \frac{p(\mathcal{D}|f(\cdot))p(f(\cdot))}{\int p(f(\cdot)|\mathcal{D})p(f(\cdot))}$$

What is a kernel?

What is a kernel?

The distribution of a GP is fully characterised by:

- its mean function m defined over D
- its covariance function (or kernel) k defined over $D \times D$:
  $k(x, x') = \mathrm{cov}(f(x), f(x'))$



Limitations of kernels

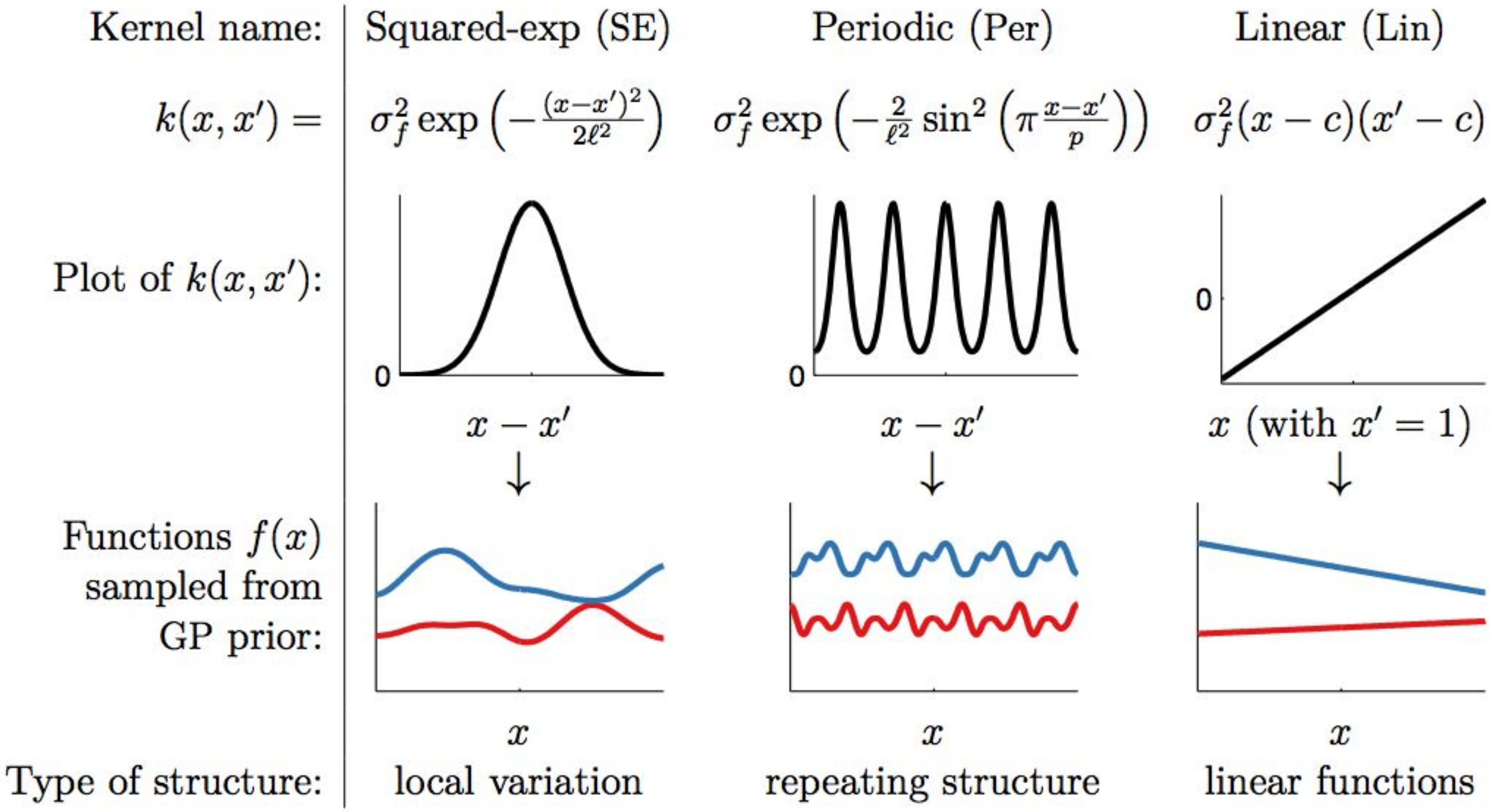
A kernel satisfies the following properties:

- It is symmetric: $k(x, x') = k(x', x)$
- It is positive semi-definite (psd): $$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

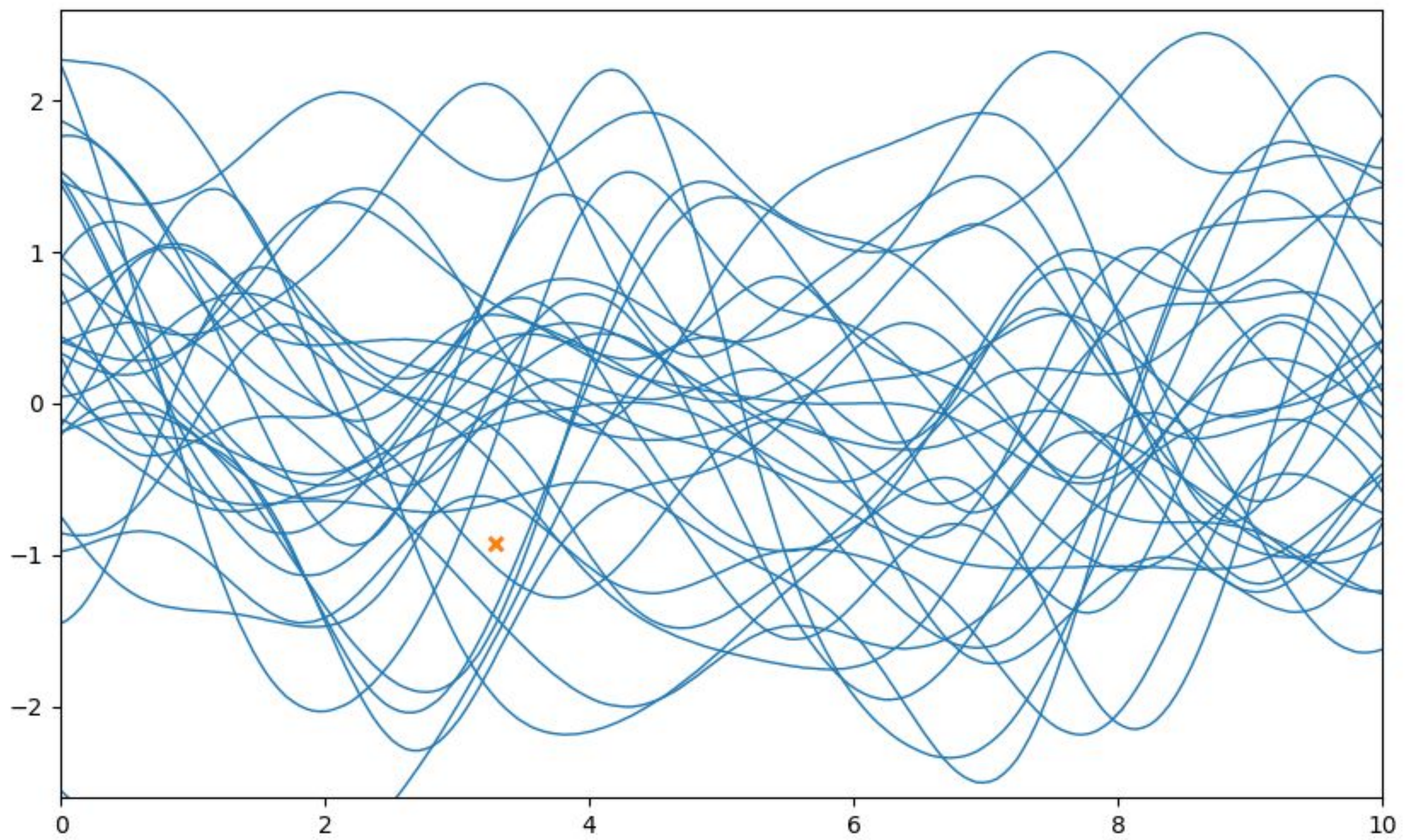
For example

$k(x1, x1) \geq 0$

$|k(x1, x2)|^2 \leq k(x1, x1)\, k(x2, x2)$



Kernel name:
Squared-exp (SE)
Periodic (Per)
Linear (Lin)
$k(x, x') =$
$\sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$
$\sigma_f^2 \exp\left(-\frac{2}{\ell^2}\sin^2\left(\pi\frac{x-x'}{p}\right)\right)$
$\sigma_f^2(x-c)(x'-c)$
Plot of $k(x, x')$:
0
0
0
$x - x'$
$x - x'$
x (with $x' = 1$)
Functions $f(x)$ sampled from GP prior:
x
x
x
Type of structure:
local variation
repeating structure
linear functions



2
1
0
−1
−2
0
2
4
6
8
10

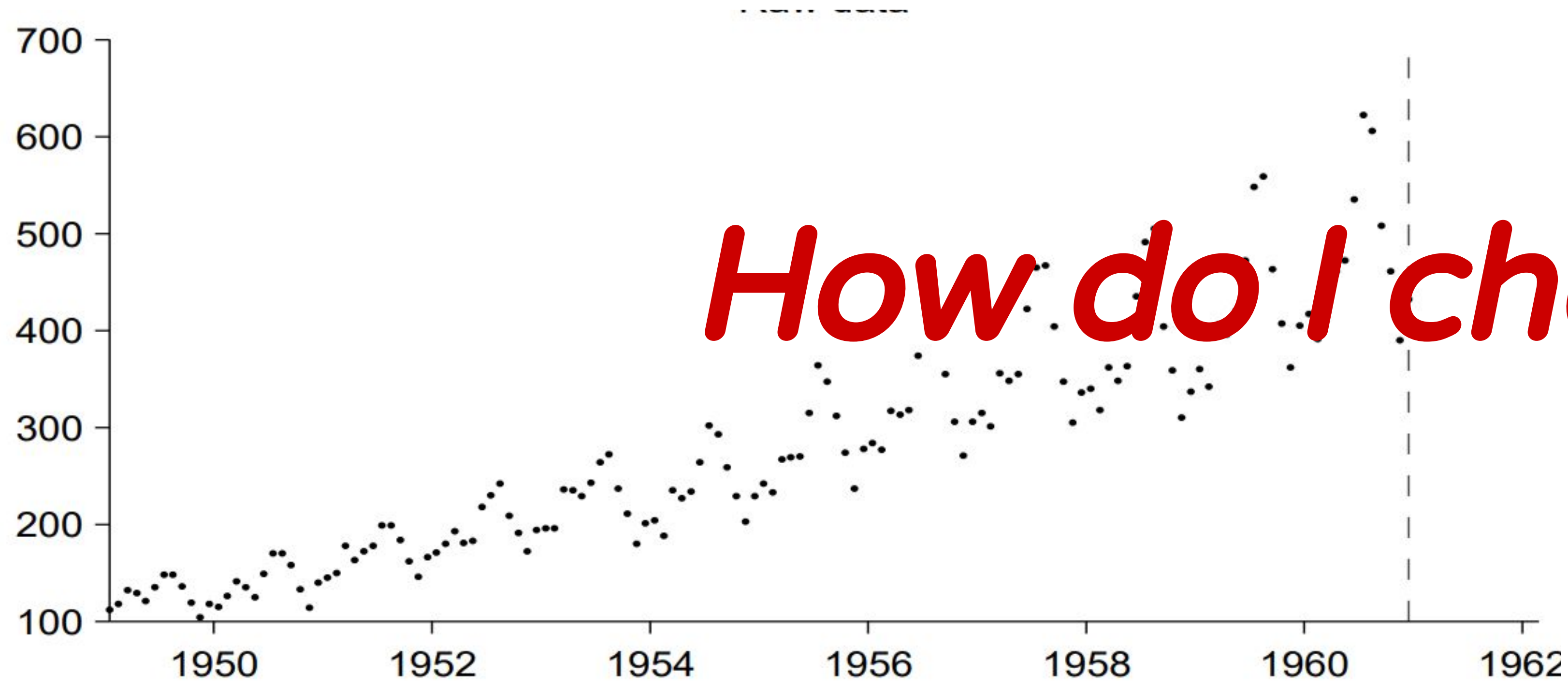
Gaussian Processes

Examples of kernels

constant	$k(x, x') = \sigma^2$
white noise	$k(x, x') = \sigma^2 \delta_{x,x'}$
Brownian	$k(x, x') = \sigma^2 \min(x, x')$
exponential	$k(x, x') = \sigma^2 \exp\left(-\lvert x - x' \rvert / \theta\right)$
Matérn 3/2	$k(x, x') = \sigma^2 \left(1 + \lvert x - x' \rvert\right) \exp\left(-\lvert x - x' \rvert / \theta\right)$
Matérn 5/2	$k(x, x') = \sigma^2 \left(1 + \lvert x - x' \rvert / \theta + 1/3 \lvert x - x' \rvert^2 / \theta^2\right) \exp\left(-\lvert x - x' \rvert / \theta\right)$
squared exponential	$k(x, x') = \sigma^2 \exp\left(-(x - x')^2 / \theta^2\right)$
linear	$k(x, x') = \sigma^2 xy$

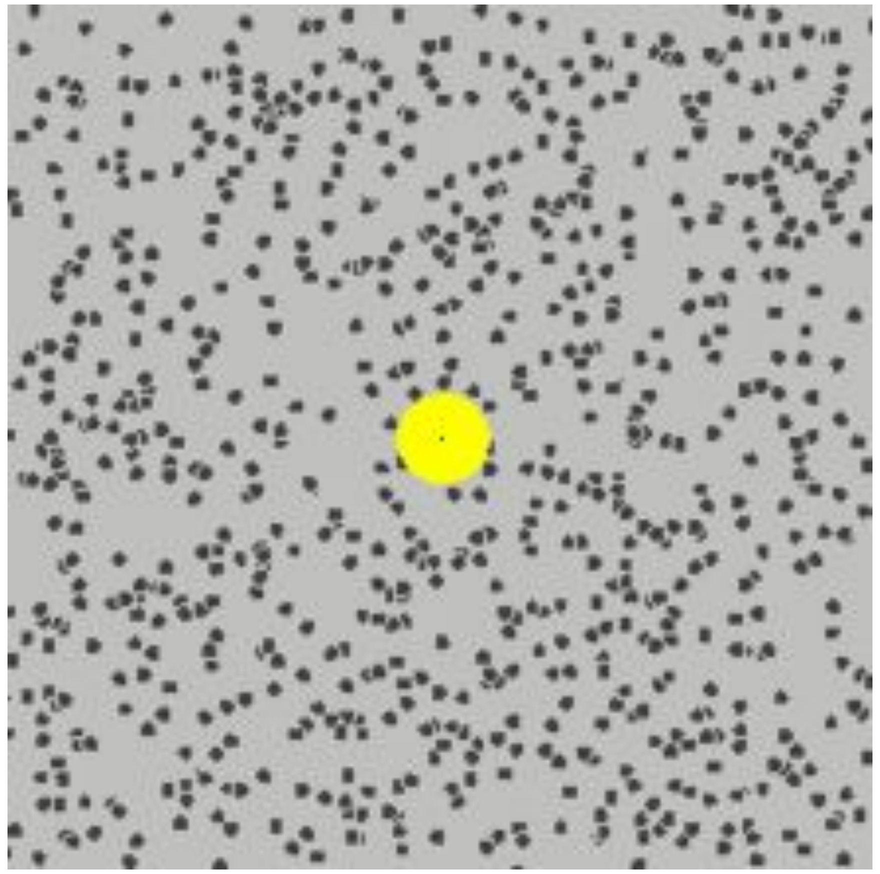
Gaussian Processes

Random samples

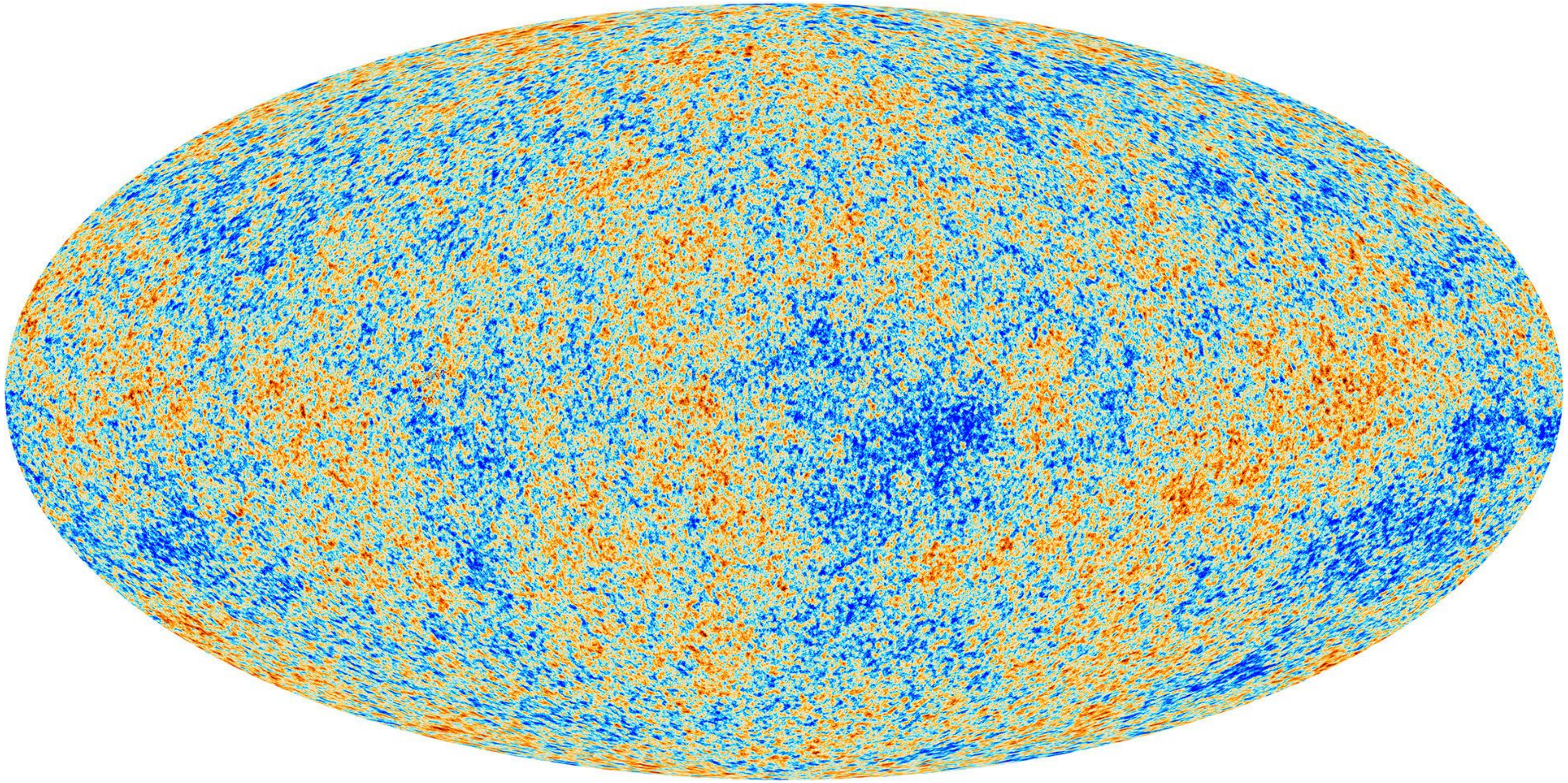


How do I choose the right one?

Physically motivated kernels

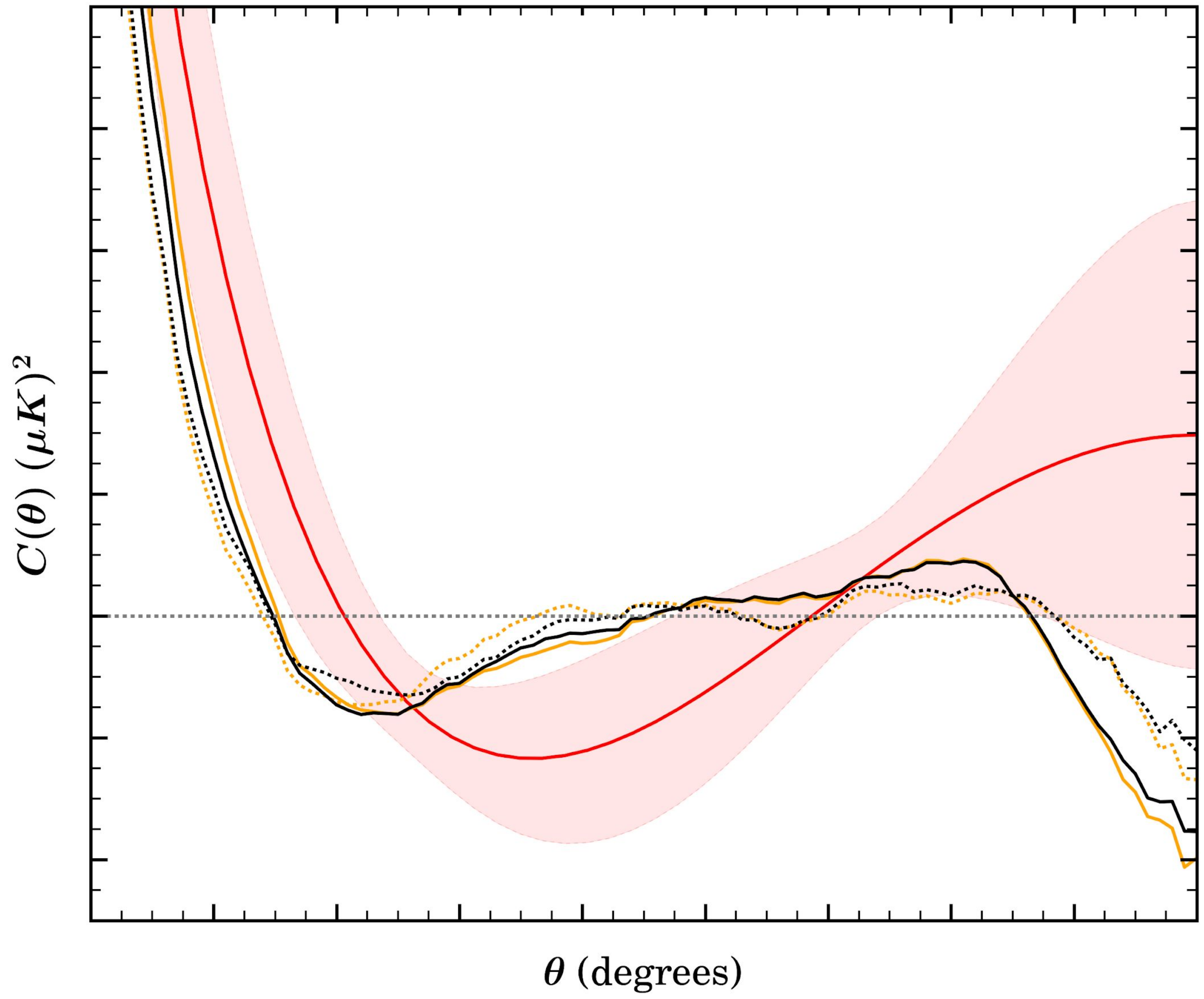


Physically motivated kernels



ESA and the Planck Collaboration

Physically motivated kernels

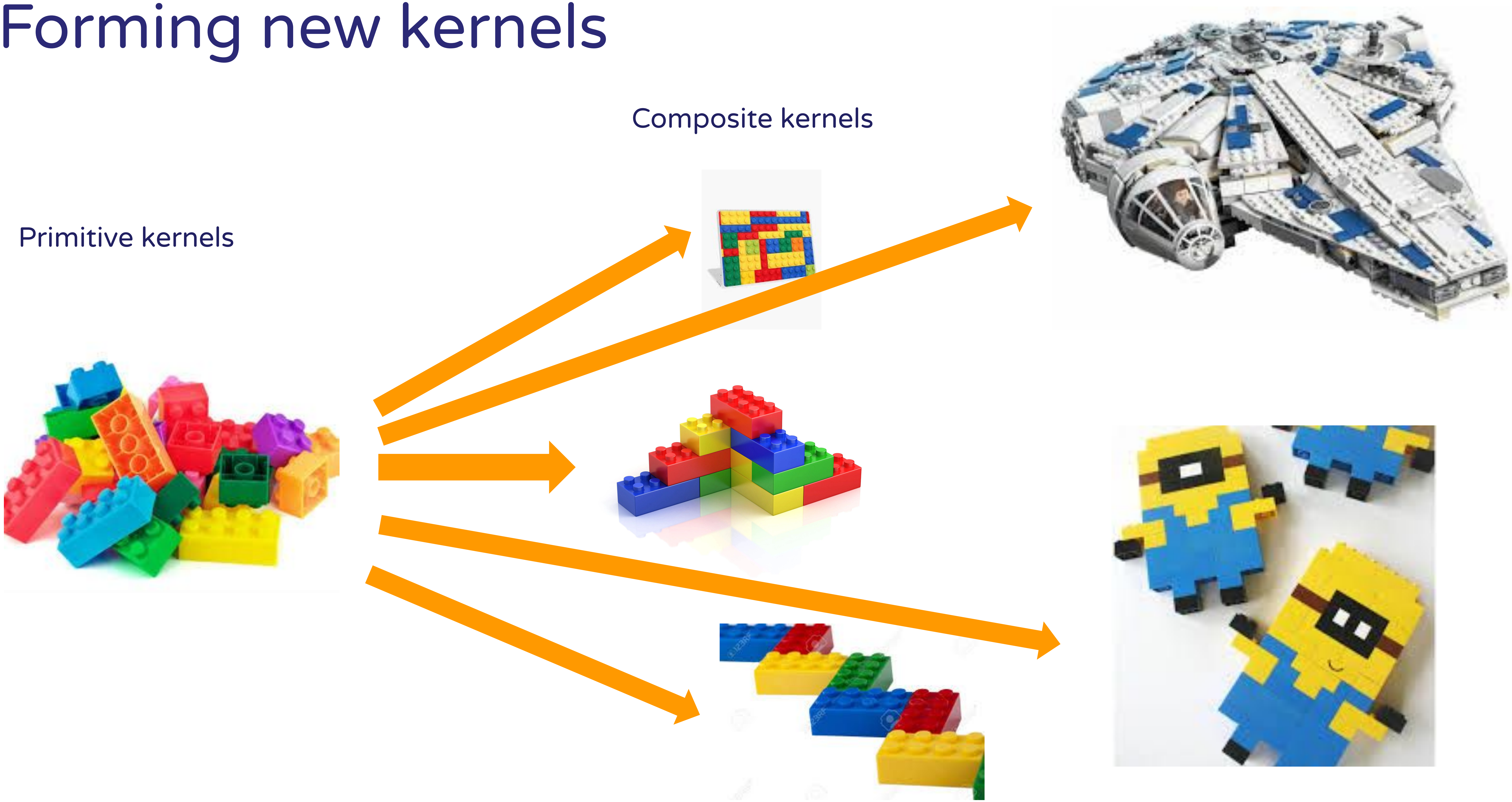


Hand crafted kernels

Forming new kernels

Composite kernels

Primitive kernels



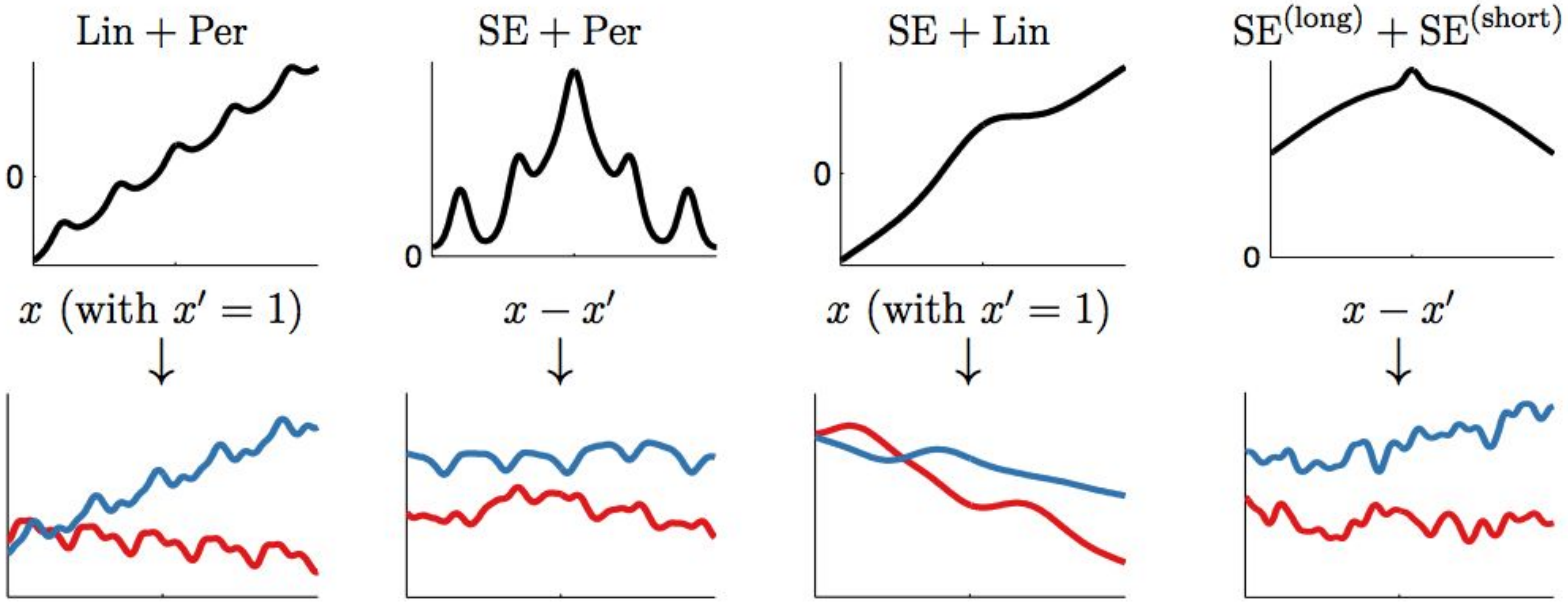
Forming new kernels

Kernel addition

Adding two kernels generates a new kernel

Forming new kernels

Kernel addition



Duvenaud (2014)

Forming new kernels

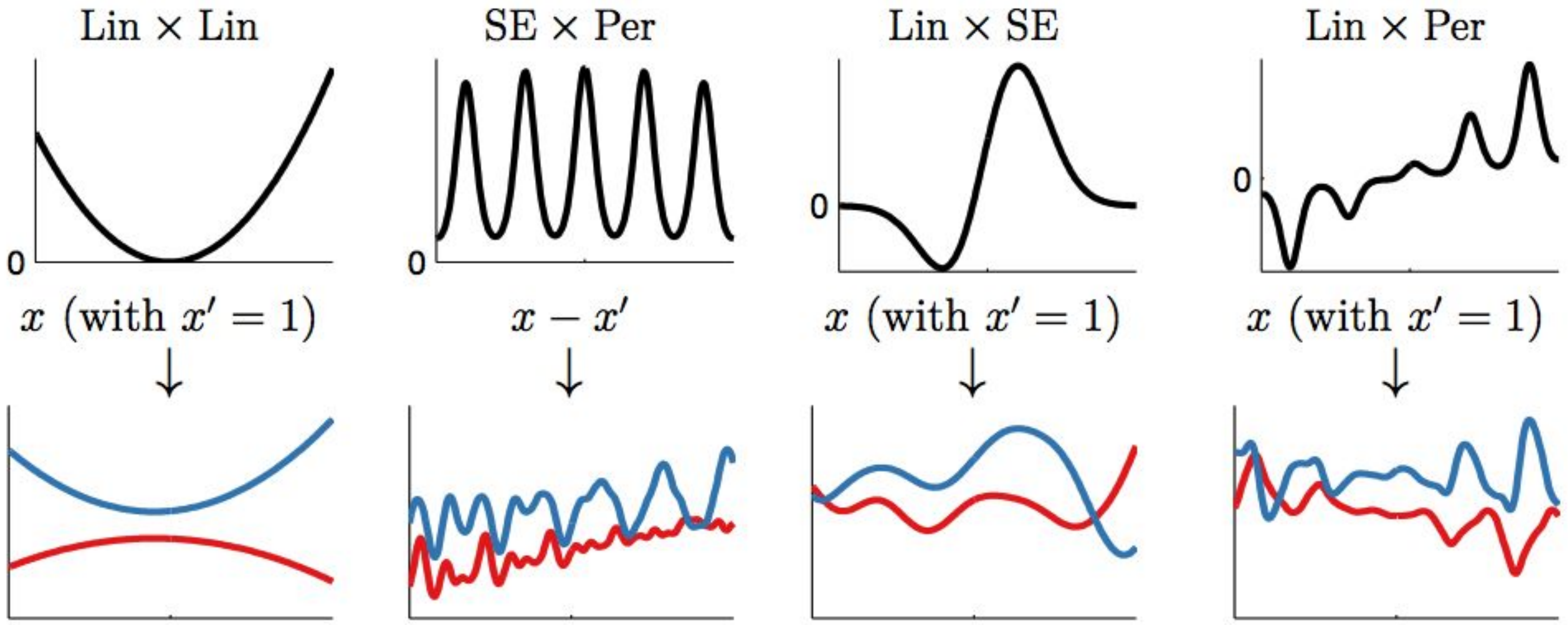
Kernel products

Adding two kernels generates a new kernel

Multiplying two kernels generates a new kernel

Forming new kernels

Kernel products



Duvenaud (2014)

Forming new kernels

Kernel changepoints

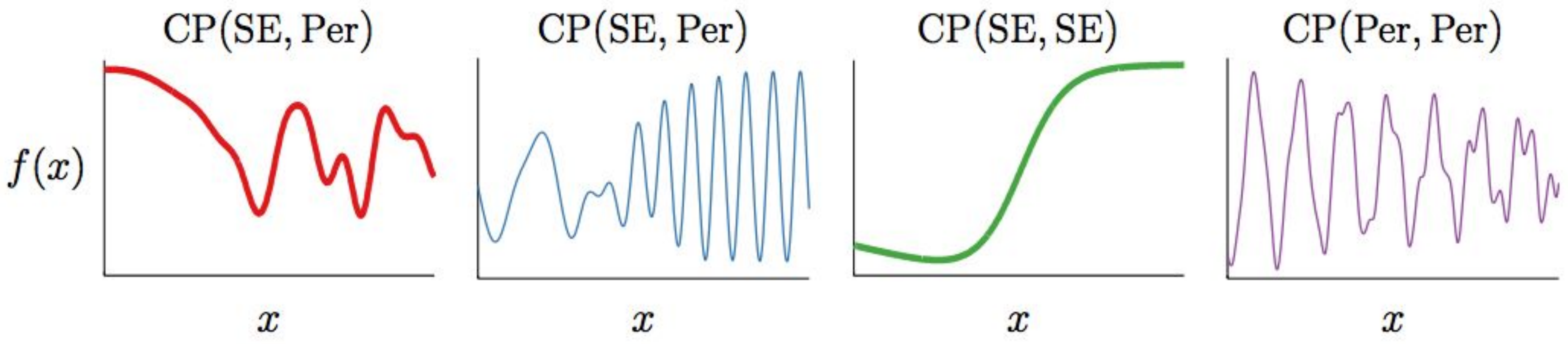
Adding two kernels generates a new kernel

Multiplying two kernels generates a new kernel

Stitching two kernels together generates a new kernel

Forming new kernels

Kernel changepoints

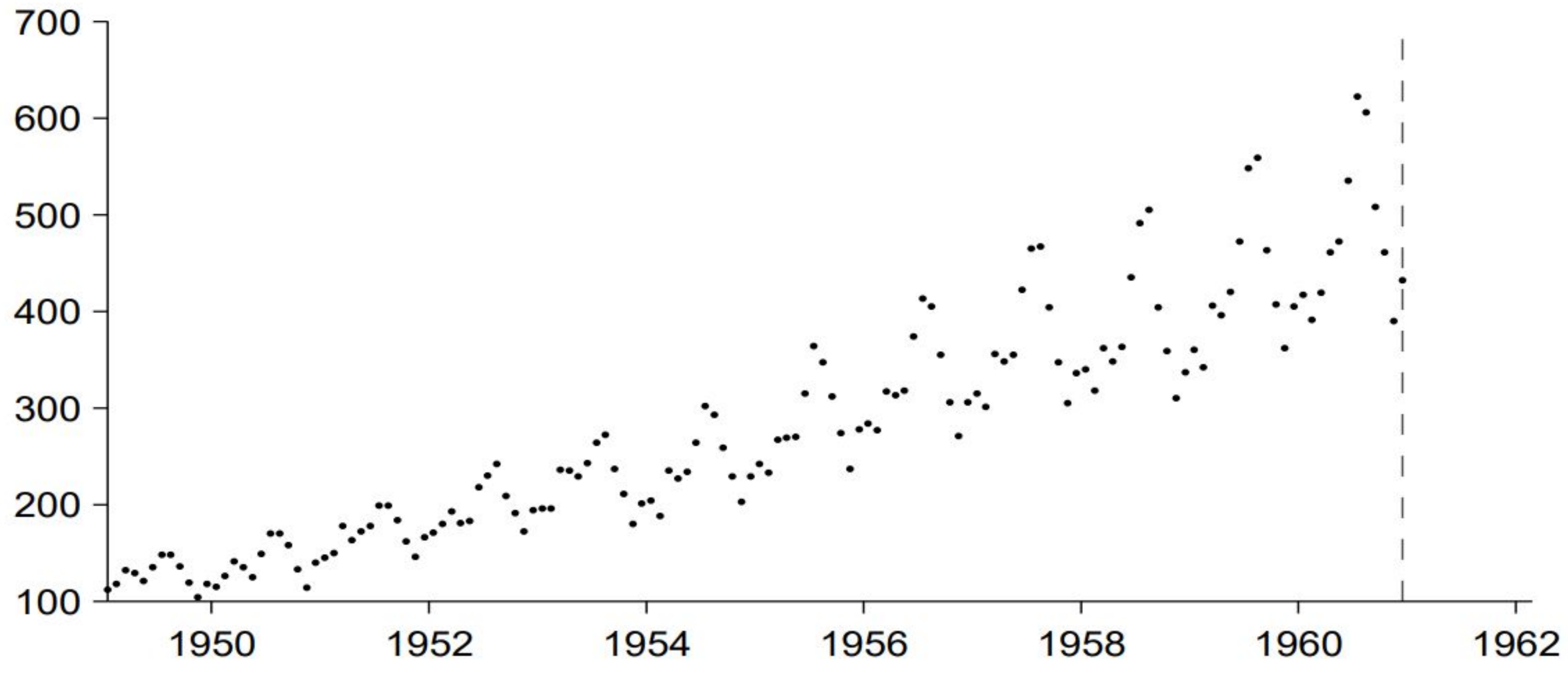


$$\mathrm{CP}(k_1, k_2) = k_1 \times \boldsymbol{\sigma} + k_2 \times \bar{\boldsymbol{\sigma}}$$

where $\boldsymbol{\sigma} = \sigma(x)\sigma(x')$ and $\bar{\boldsymbol{\sigma}} = (1 - \sigma(x))(1 - \sigma(x'))$.

Duvenaud (2014)

Manual approach



1) Try k_{SE}

2) Try $k_{SE} \times k_{Per} + k_{SE} + k_{Noise}$

3) Try $k_{SE} + k_{SE} \times k_{Per} + k_{RQ} + k_{SE} + k_{Noise}$

4) Try ....

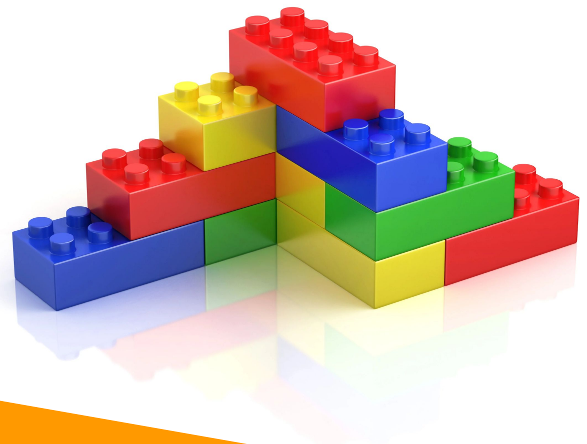
Building bespoke models by hand is very time-consuming

The Automatic Statistician

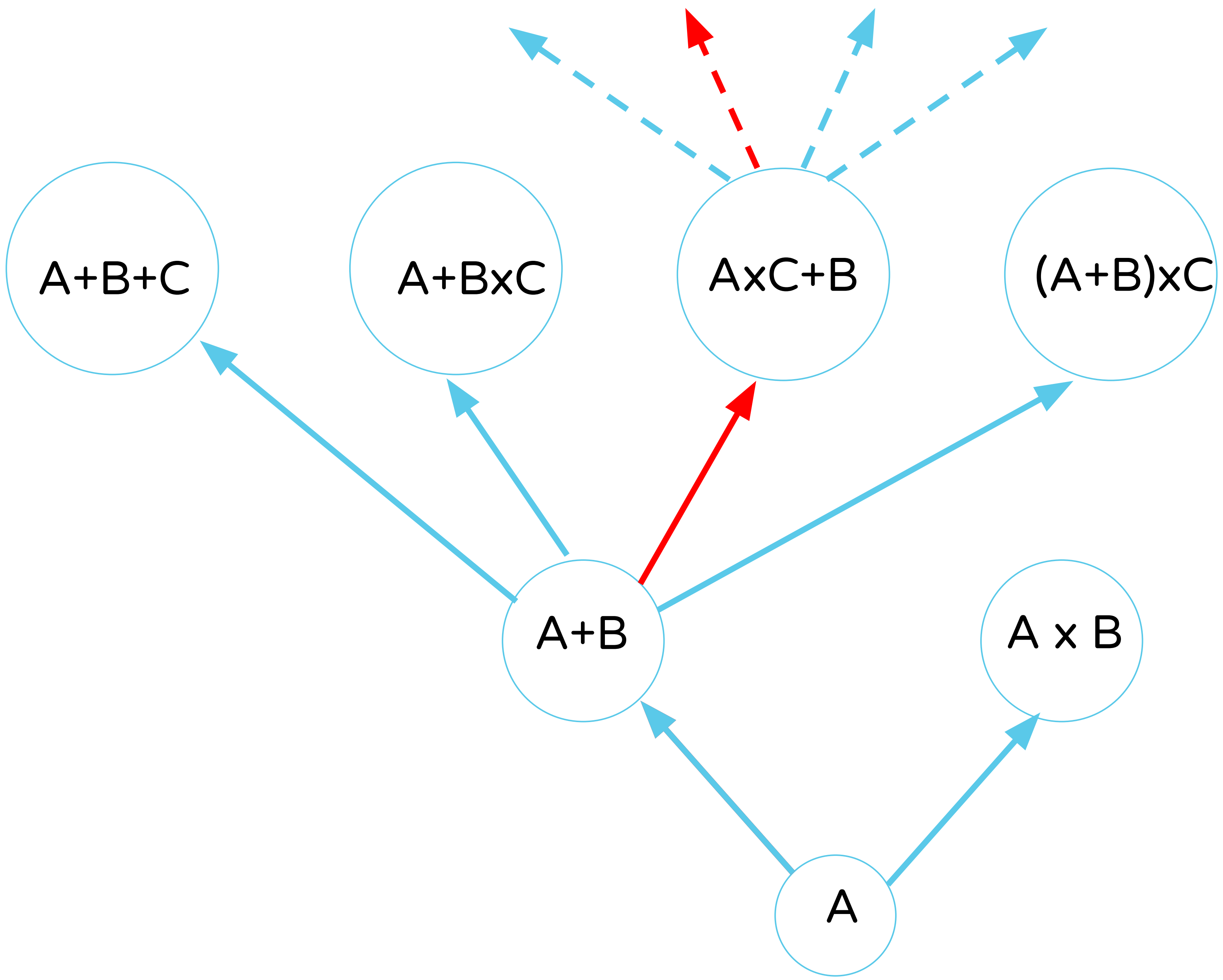
Many possible combinations

Composite kernels

Primitive kernels



Tree of kernels

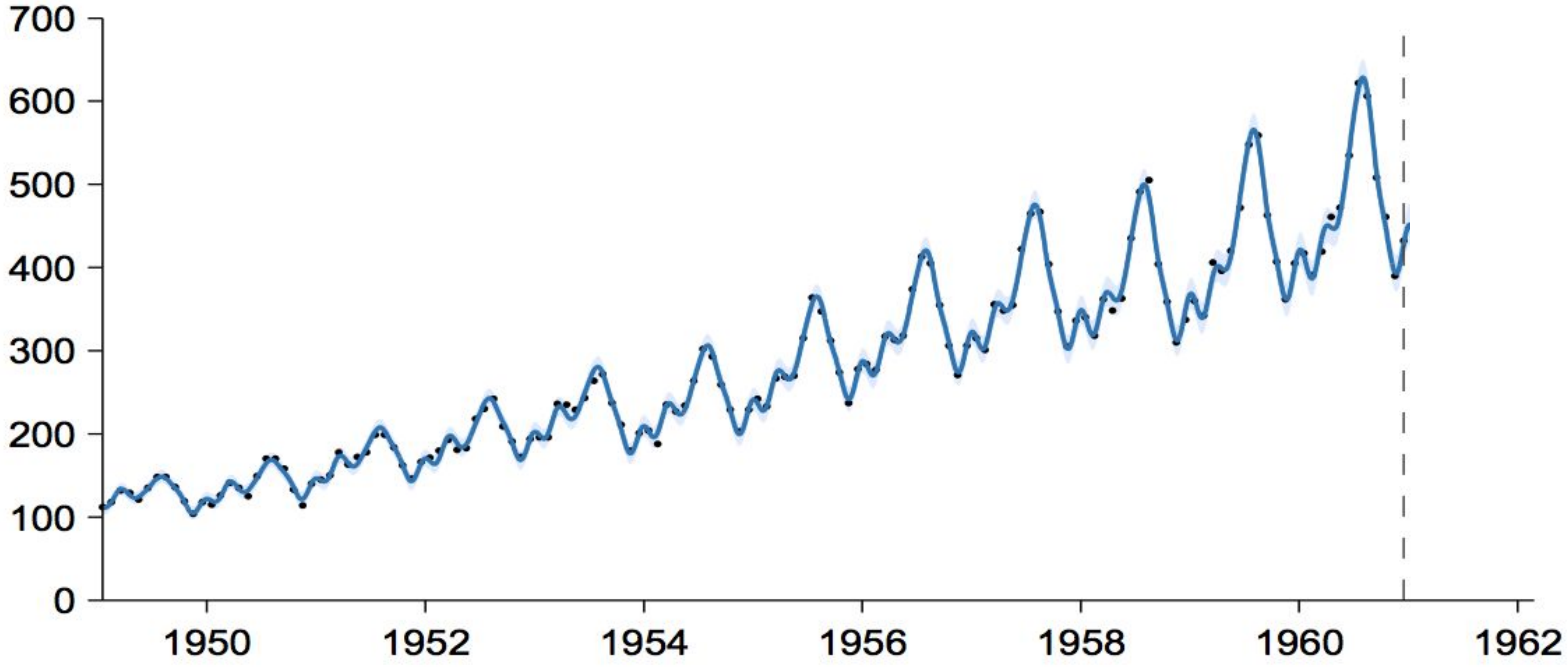


where B can be:

Linear
Periodic
Squared exponential
Constant
Noise

Does it work?

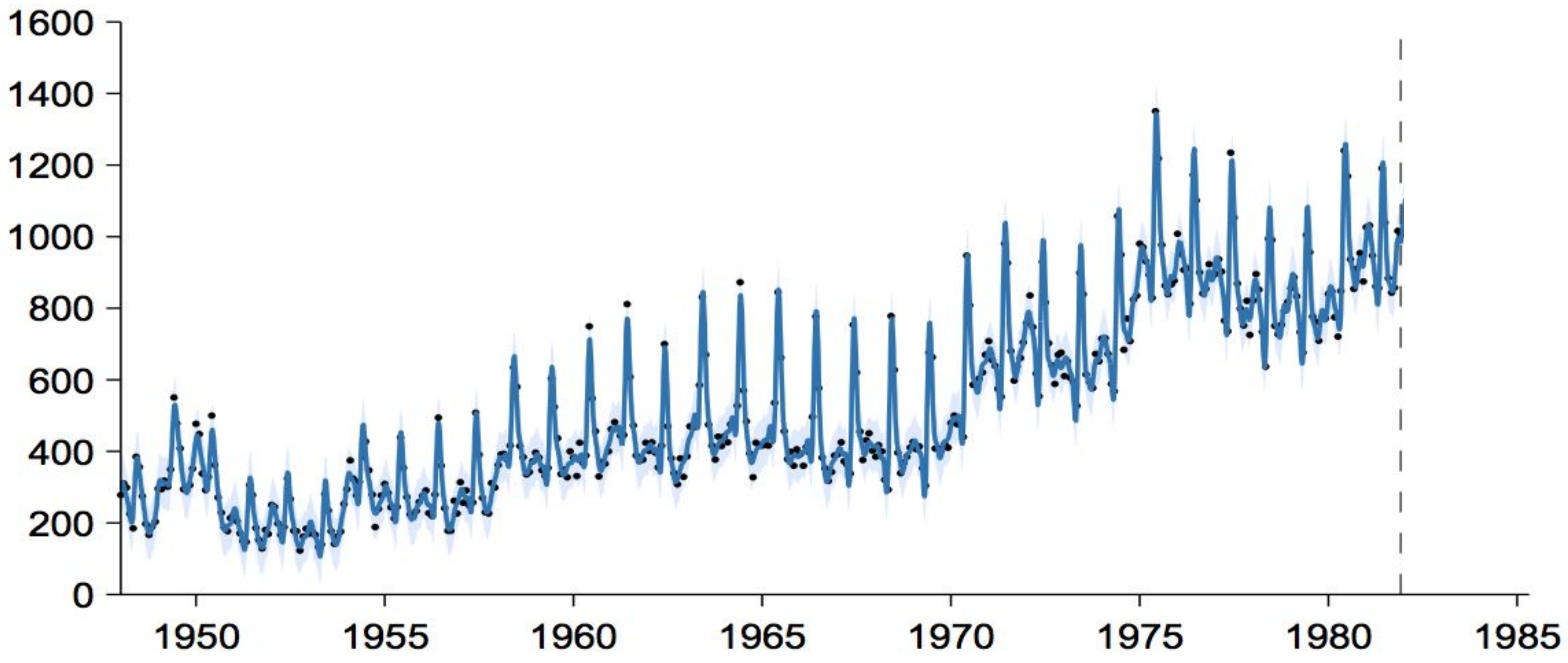
Predicting airline passengers



From www.automaticstatistician.com

Does it work?

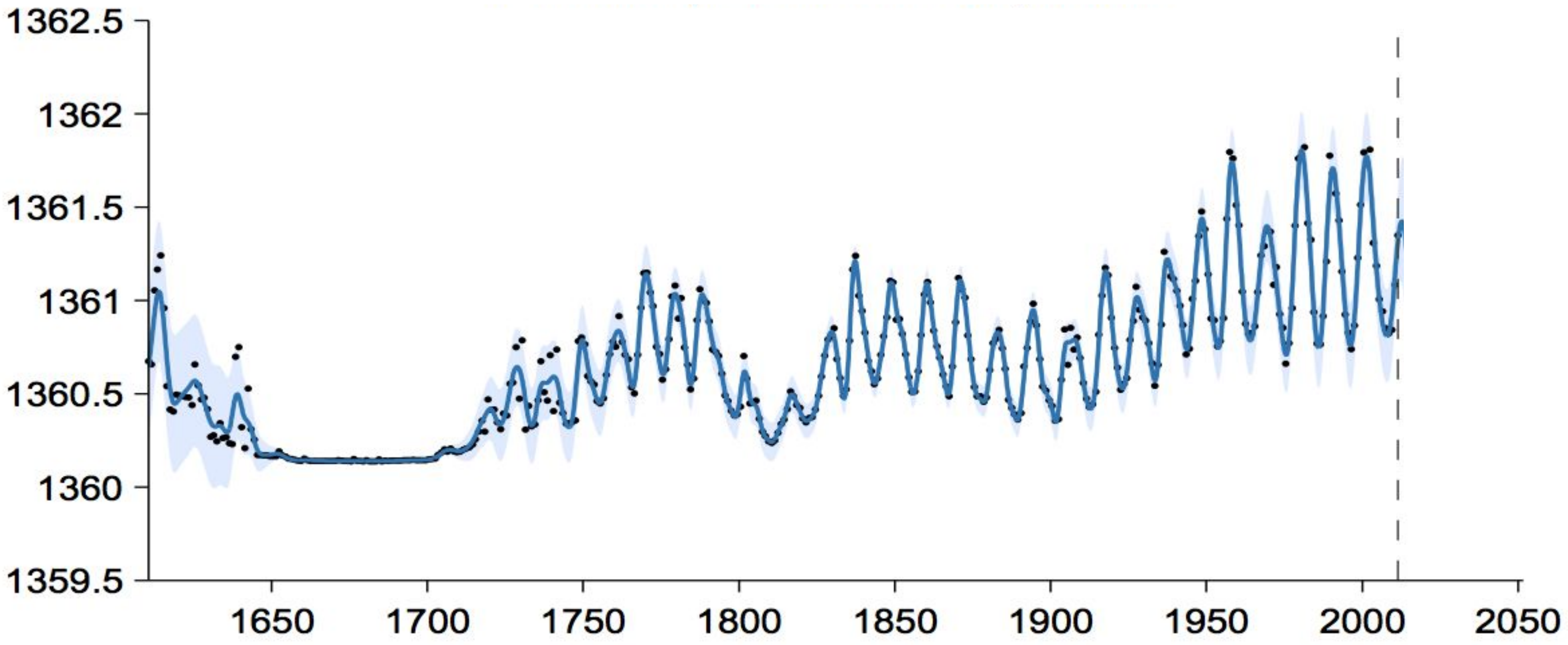
Forecasting unemployment levels



From www.automaticstatistician.com

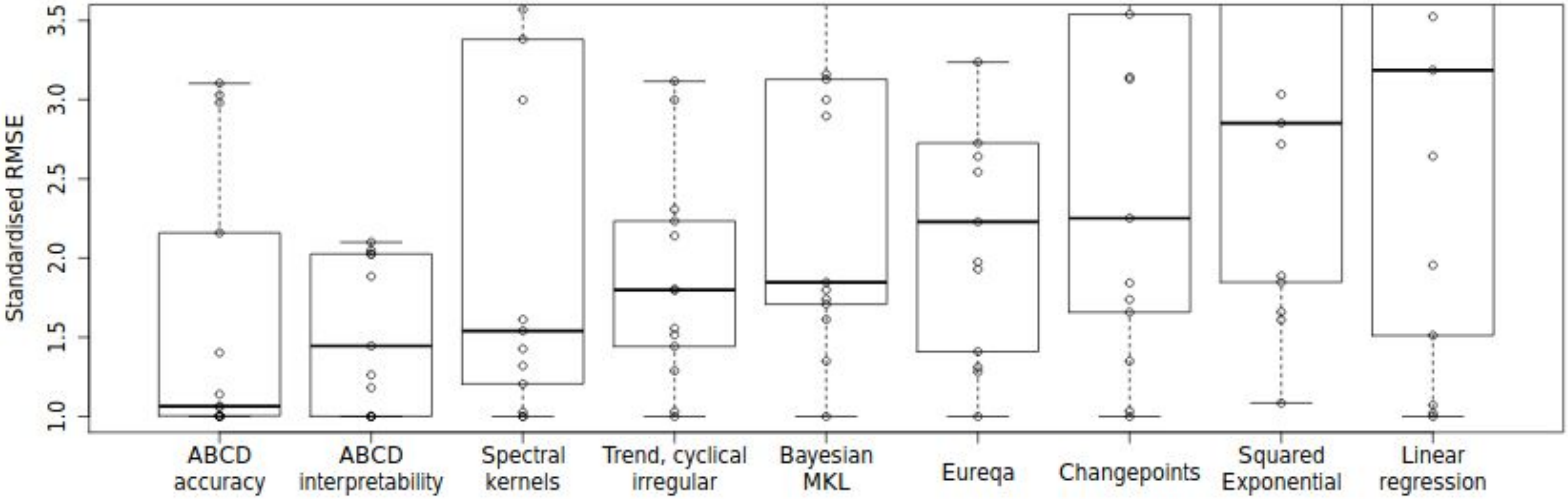
Does it work?

Forecasting solar irradiance



Lloyd et al (2014)

Does it work?



Lloyd et al 2014

But…

Searching deep into the tree is slow

Difficult to decide when to stop

Ad hoc selection of primitive kernels

Group → / Period ↓	1	2	3		4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																		2 He
2	3 Li	4 Be												5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg												13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc		22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y		40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57 La	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89 Ac	**	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og
				*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu	
				**	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr	

Spectral Kernels

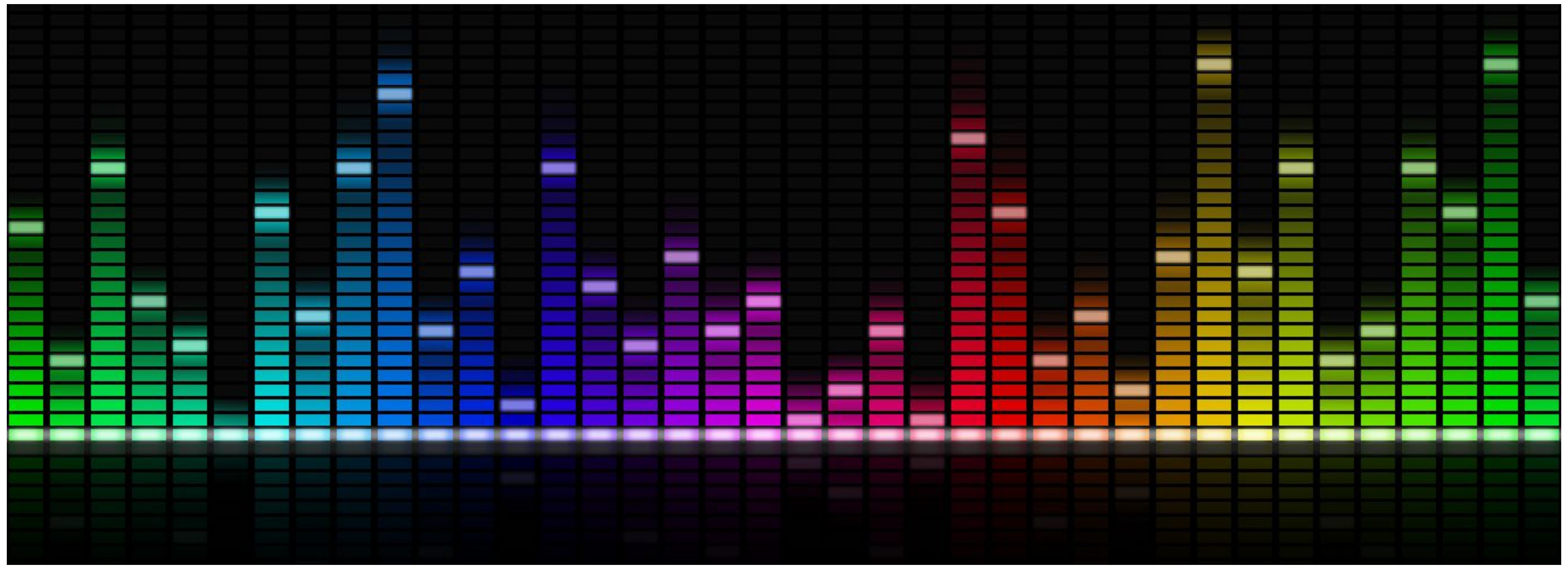
Spectral Kernel

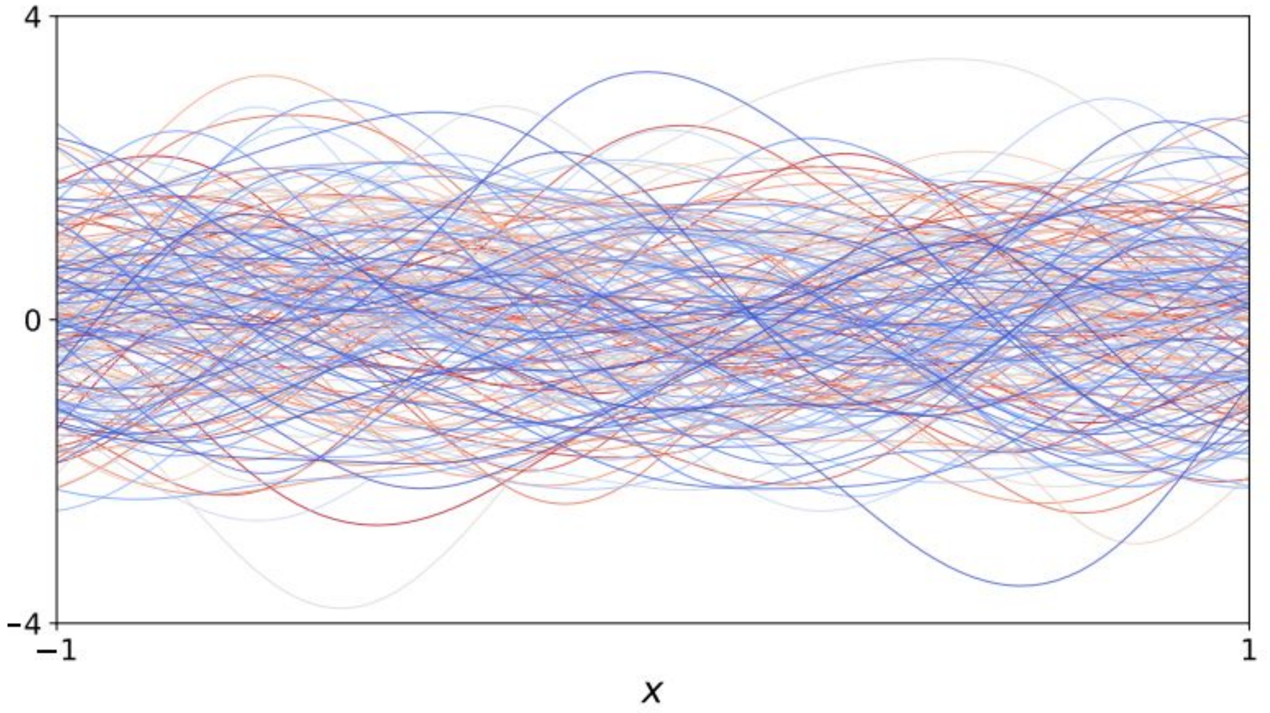
$$A = k_{SE} \times k_{COS}$$

$$K = A + A + A + A + A + A + A + A + A + A$$

Spectral representation

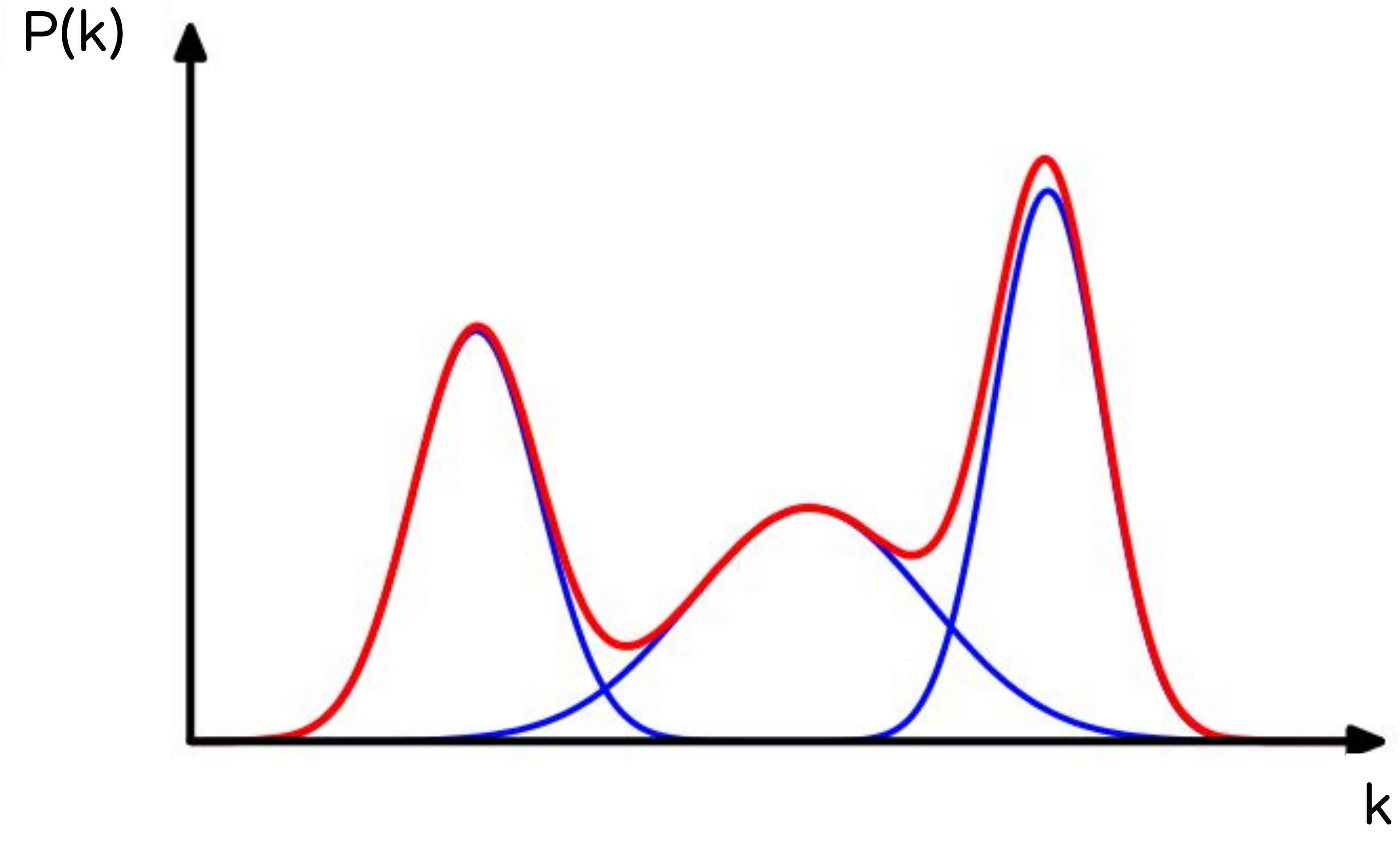
Any stationary kernel can be represented by a spectral density





Spectral Mixture Kernel

A mixture of Gaussians can approximate any stationary Gaussian process



Spectral Mixture Kernel

Power spectrum described by a single Gaussian:

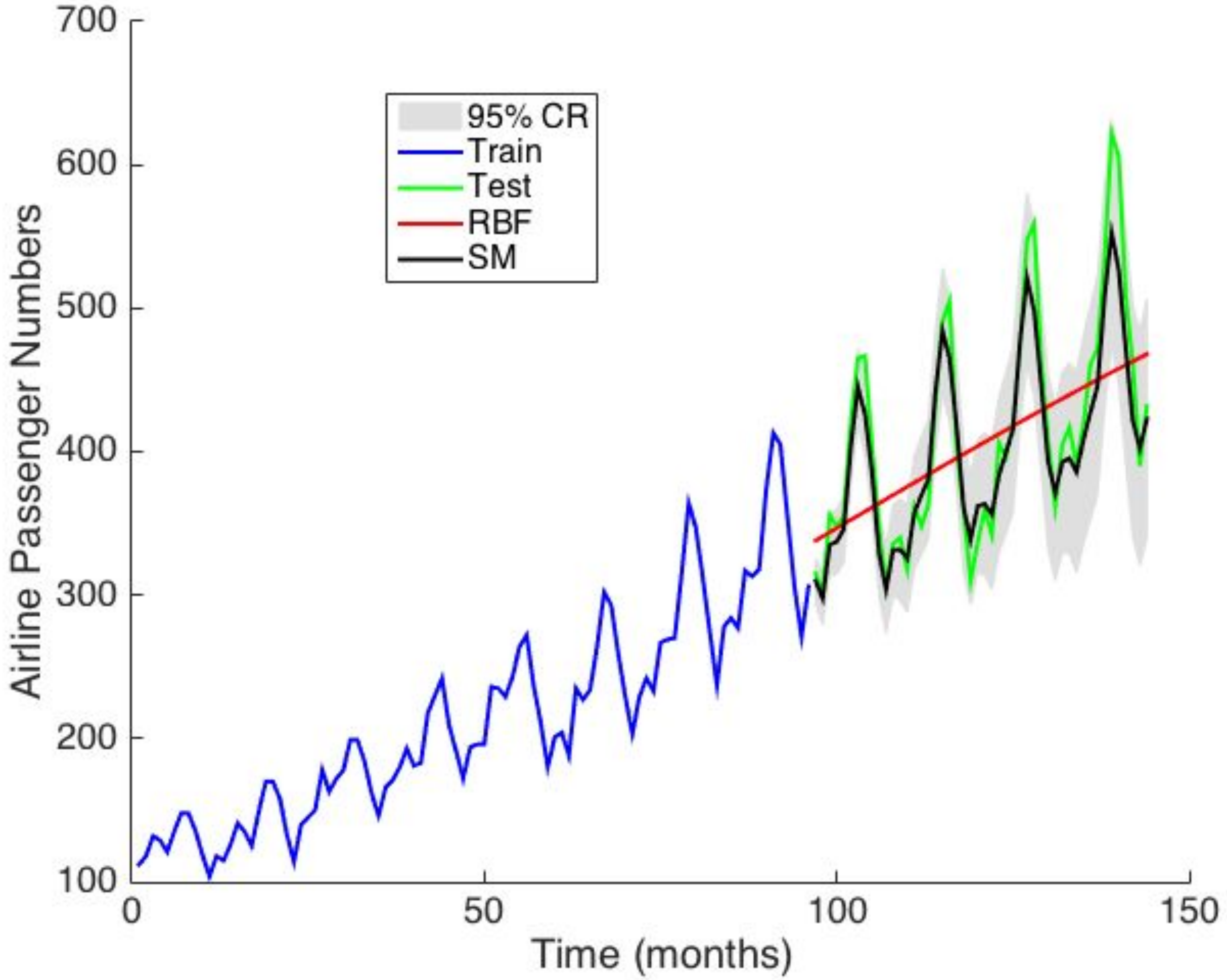
$$P(k) = G(\mu, \sigma^2)$$

$$K(r) = \exp(-2\pi^2 r^2 \sigma^2) \cos(2\pi r \mu))$$

A mixture of Gaussians can approximate **any** stationary Gaussian process:

$$K(r) = \sum_i w_i K_i(r)$$

Does it work?



Wilson & Adams (2013)

Spectral kernels

Pros:

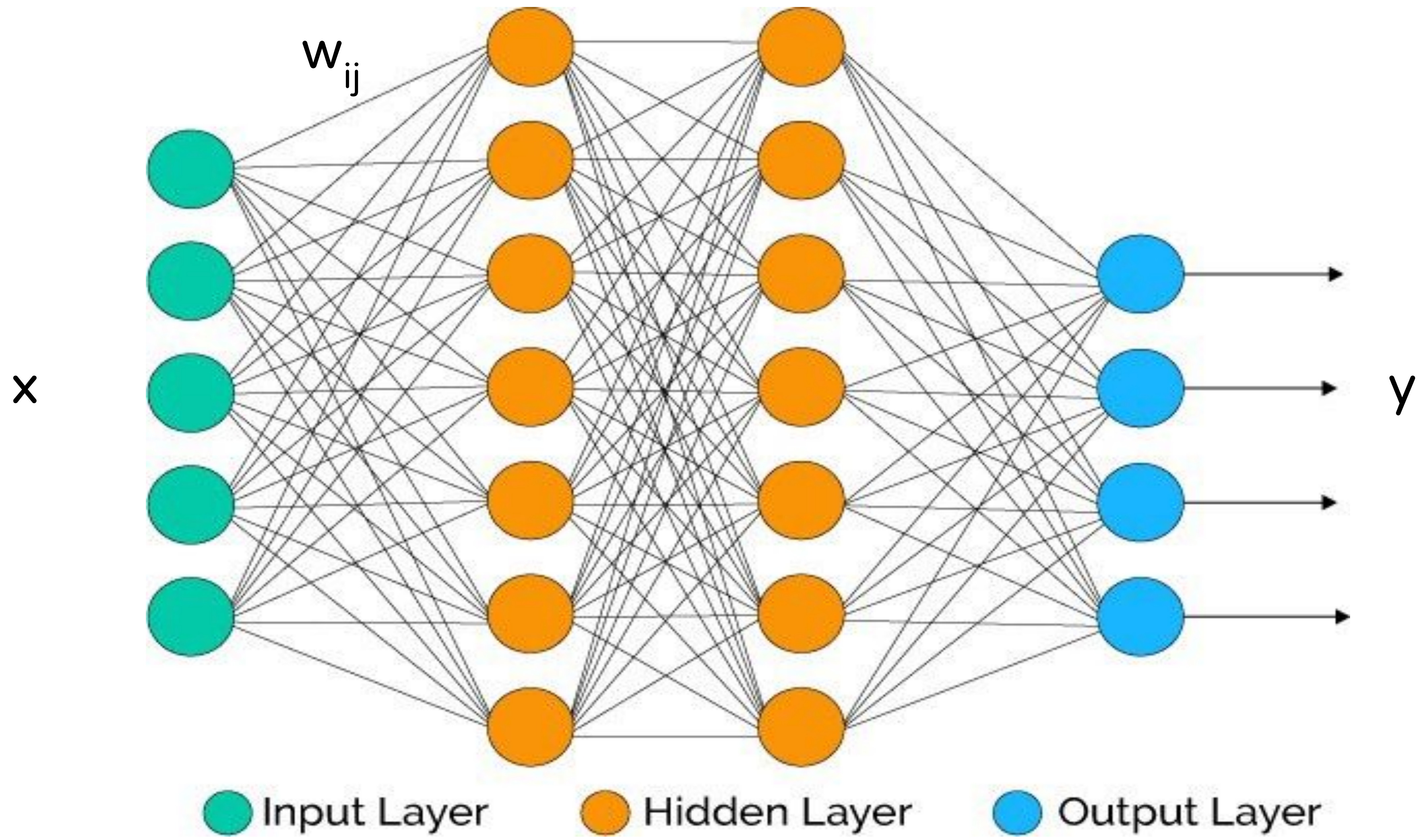
- Spans all stationary kernels
- Avoids complex search

Cons:

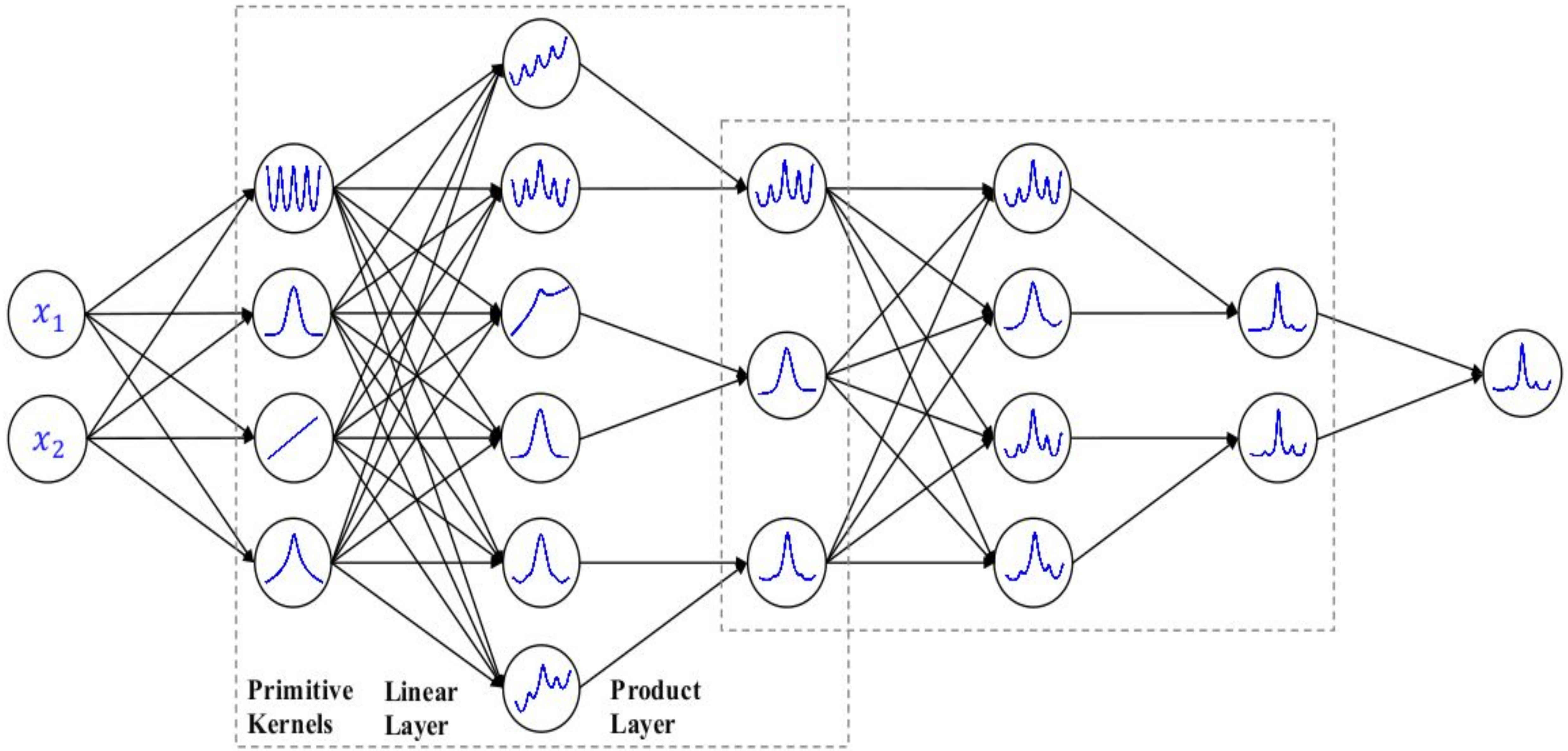
- Challenging to optimize
- Conventional approach cannot generate non-stationary kernels

Neural Kernel Network

Neural Network



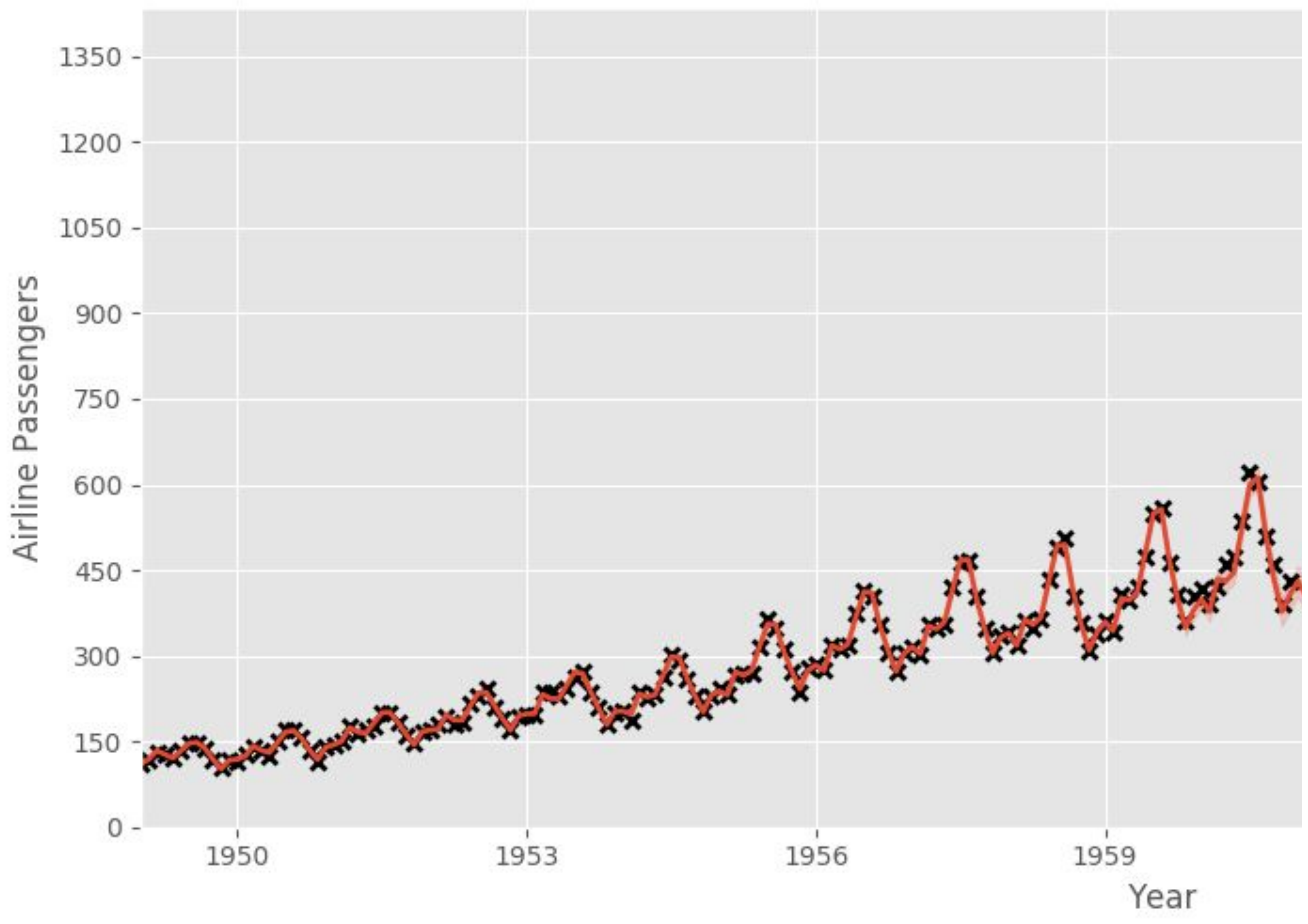
Neural Kernel Network



Sun et al (2018)

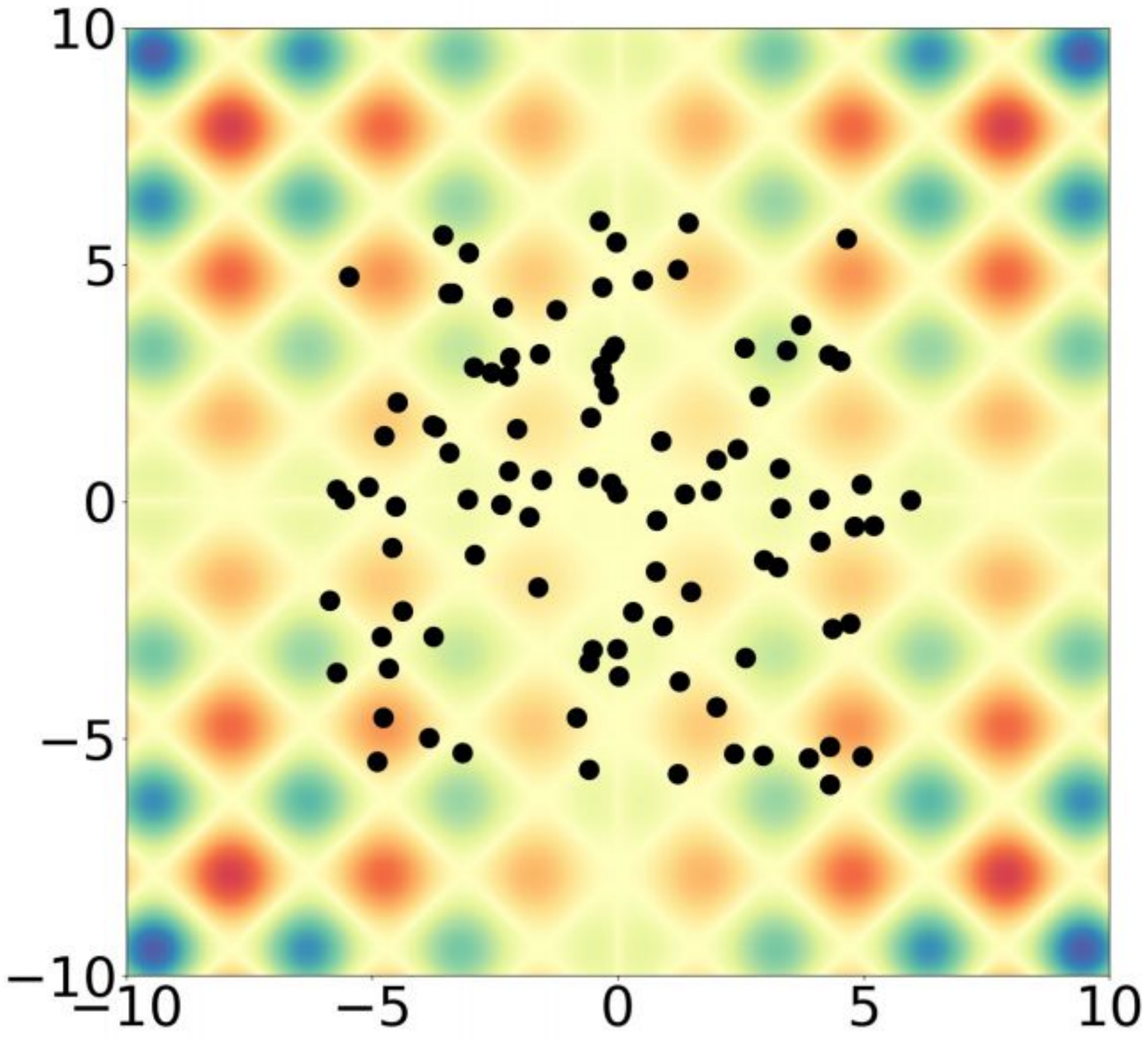
Does it work?

Predicting airline passengers



Does it work?

Extrapolating patterns



Sun et al (2018)

Summary

Kernel Selection Methods

Automatic Statistician

www.automaticstatistician.com

Spectral Kernels

Generalised Spectral Mixture https://github.com/sremes/nssm-gp

Multi-output https://github.com/GAMES-UChile/mogptk

Neural Kernel Network

Sun et al 2018: https://arxiv.org/abs/1806.04326

GPflow implementation https://github.com/frgsimpson/kernel_learning