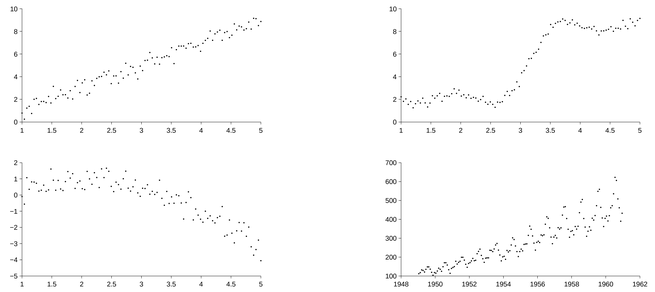
Automating pattern discovery and the statistical process for regression

David Duvenaud[1], **James Robert Lloyd**[1], Roger Grosse[2],
Joshua B. Tenenbaum[2], Zoubin Ghahramani[1]

1: Department of Engineering, University of Cambridge, UK 2: Massachusetts Institute of Technology, USA

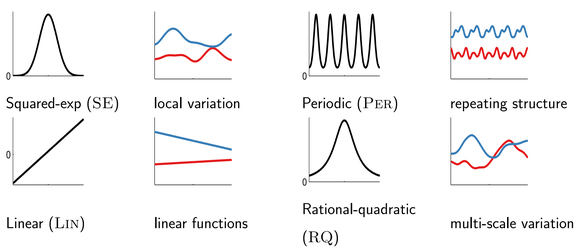
Data may often exhibit high level structure e.g. linearity, periodicity etc.



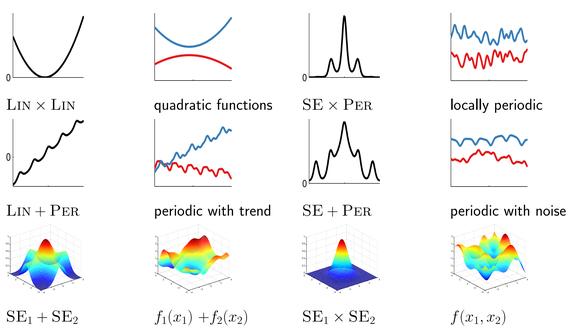
- Smoothing methods (e.g. local linear regression) would all produce good estimates of the regression function within the data for the above...
- ...but this would ignore any high level structure
- Traditionally, different statistical models would be required to produce a parsimonious fit to the above and enabel extrapolation

Gaussian process regression can model many structures with an appropriately chosen kernel

- The kernel encodes the inductive bias of the model i.e. the types of functions the model 'believes in'
- Below we depict standard base kernels, and examples of functions the model believes in (samples from the prior)



- Base kernels can be combined to create more complicated structural assumptions



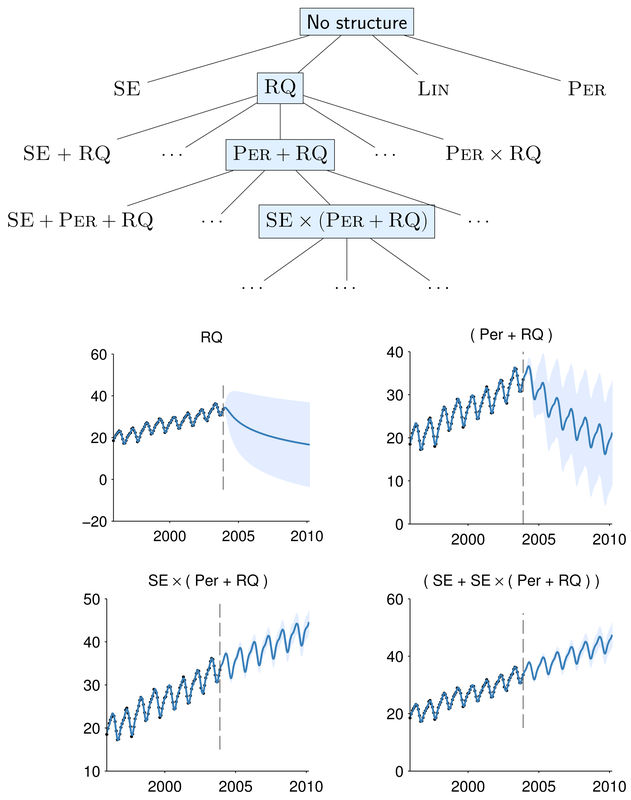
We consider all kernel expressions derived from a generative grammar...

- Constructing appropriate composite kernels has previously been the domain of Gaussian process experts
- We consider all algebraic expressions involving a small number of base kernels and the operations '+' and '×', which includes

Bayesian linear regression	LIN
Bayesian polynomial regression	$\mathrm{LIN} \times \mathrm{LIN} \times \ldots$
Generalized Fourier decomposition	$\mathrm{PER} + \mathrm{PER} + \ldots$
Generalized additive models	$\sum_{d=1}^{D} \mathrm{SE}_d$
Automatic relevance determination	$\prod_{d=1}^{D} \mathrm{SE}_d$
Linear trend with deviations	$\mathrm{LIN} + \mathrm{SE}$
Linearly growing amplitude	$\mathrm{LIN} \times \mathrm{SE}$

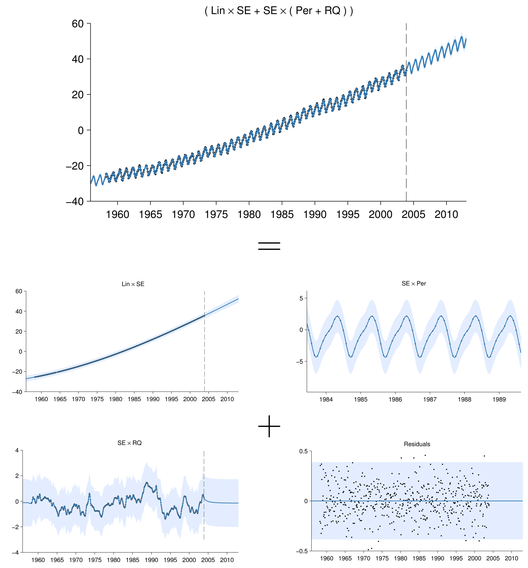
...which we search greedily, producing progressively better statistical models

- We try all base kernels, selecting the one with the highest (approximate) marginal likelihood which balances data fit and model complexity
- The search continues by adding an extra term to the current best kernel, stopping when marginal likelihood no longer improves



Example: Mauna Loa CO_2 concentration

- By automatically inferring an appropriate kernel, we can also automatically decompose functions into additive components



Example: International airline passengers