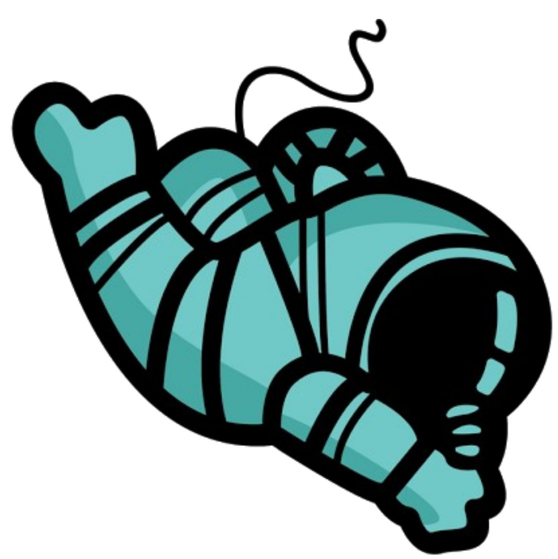
SIG-DL
Hyper parameter optimization

- Florian Goebels

About me

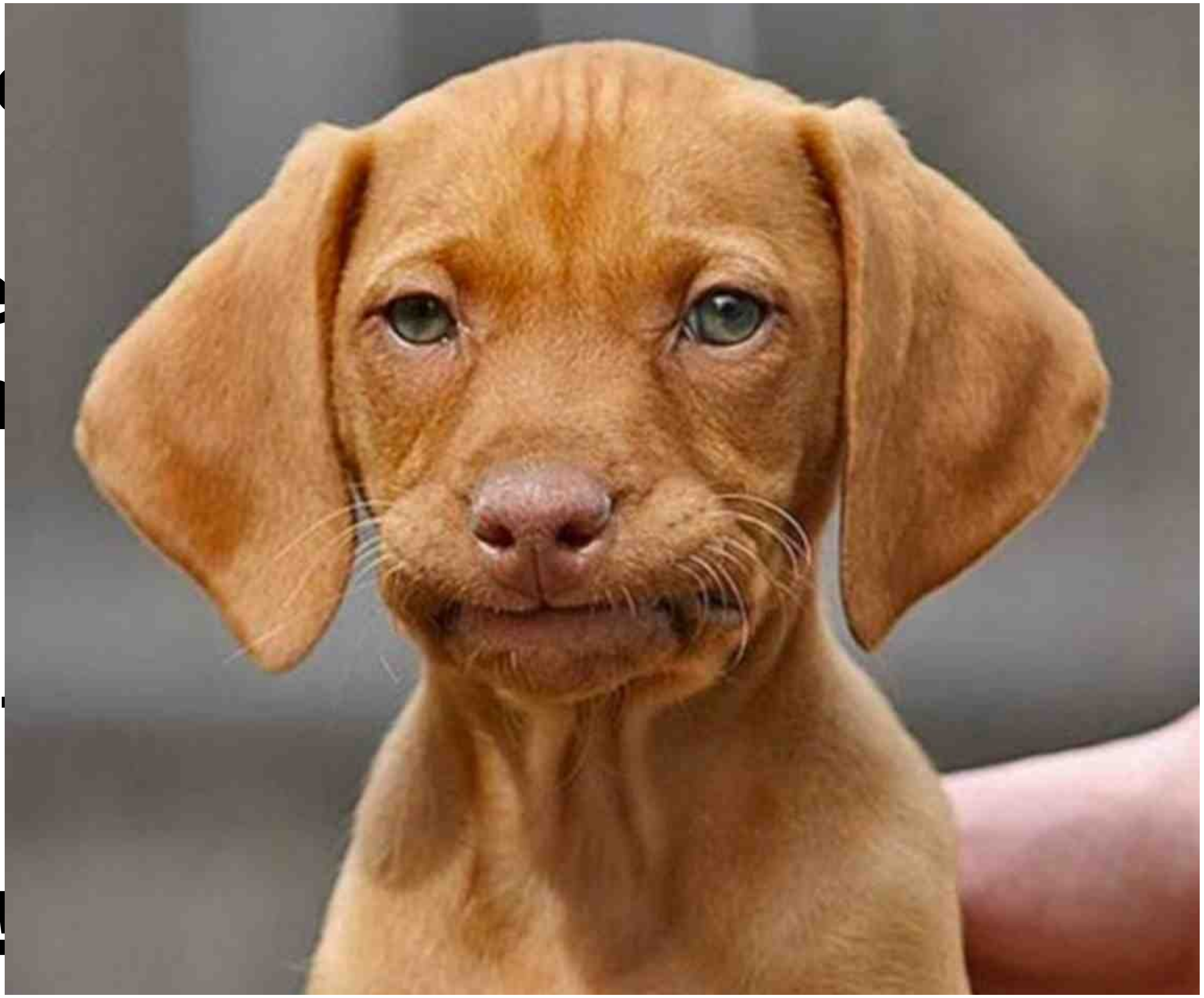


Overview

- **Grid search**
- **Random search**
- **Genetic algorithms**
- **Bayesian hyper parameter optimization**

How do we search

- Define c
- Explore
  - Search
  - Geneti
  - Bayesi
  - Grad s
- Painful!



Grid search

- **Hyper parameters: C, A, B**
- **with values 1, …, n**
- **The Algorithm:**
  - **Test all possible combinations!**

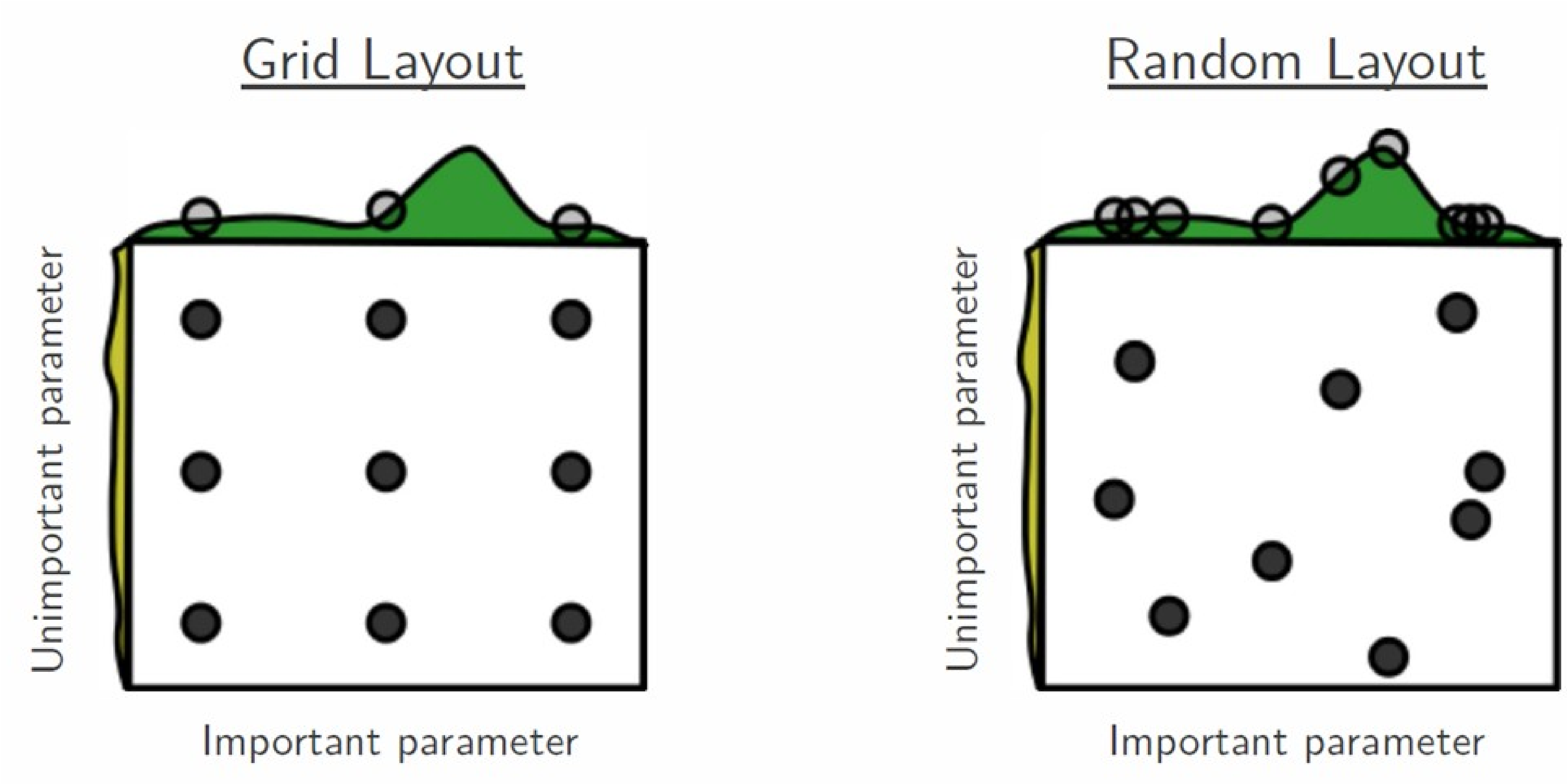
Random search

ameters: C, A, B

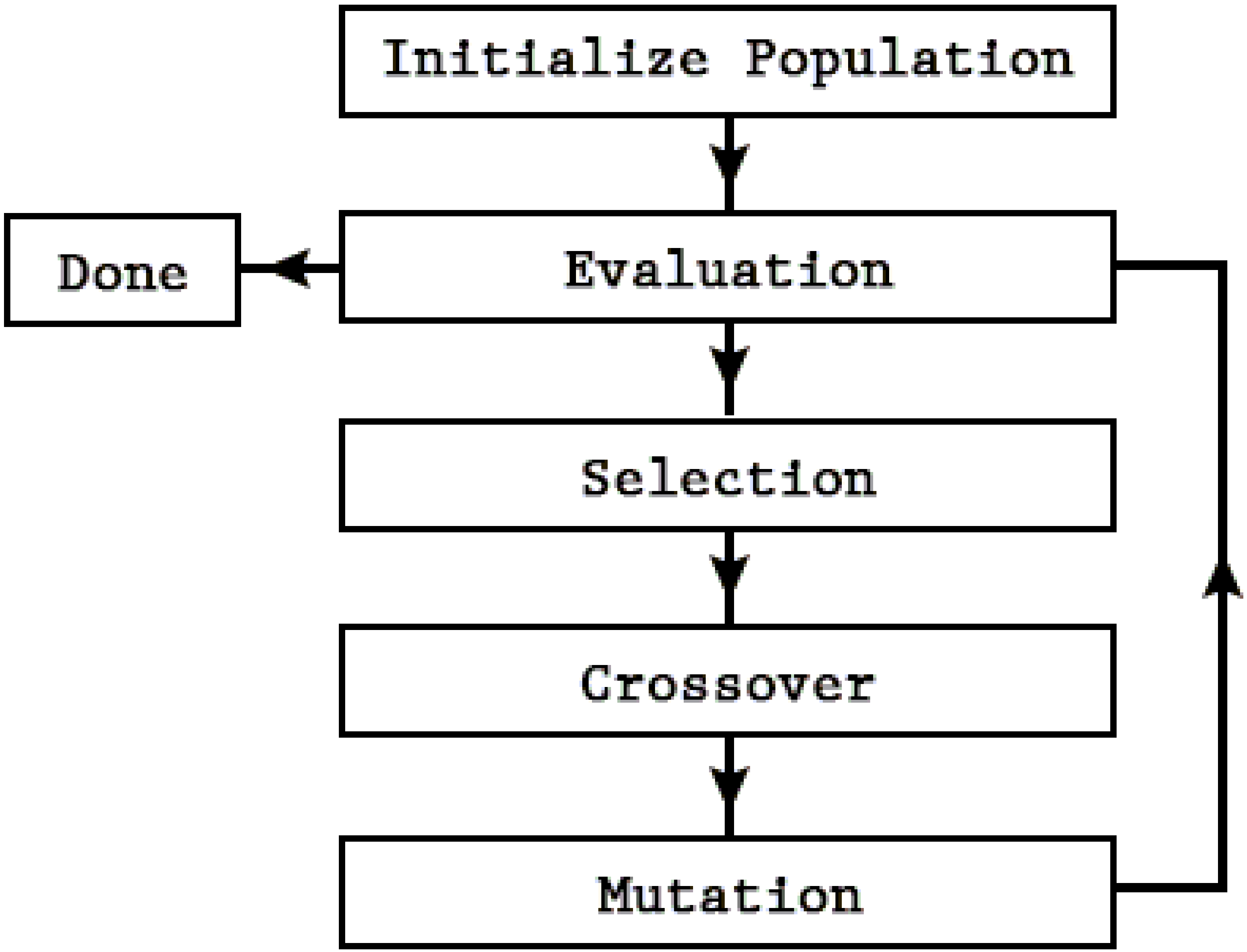
s 1, ..., n

thm:
randomly selected possible combinations at

In comparison

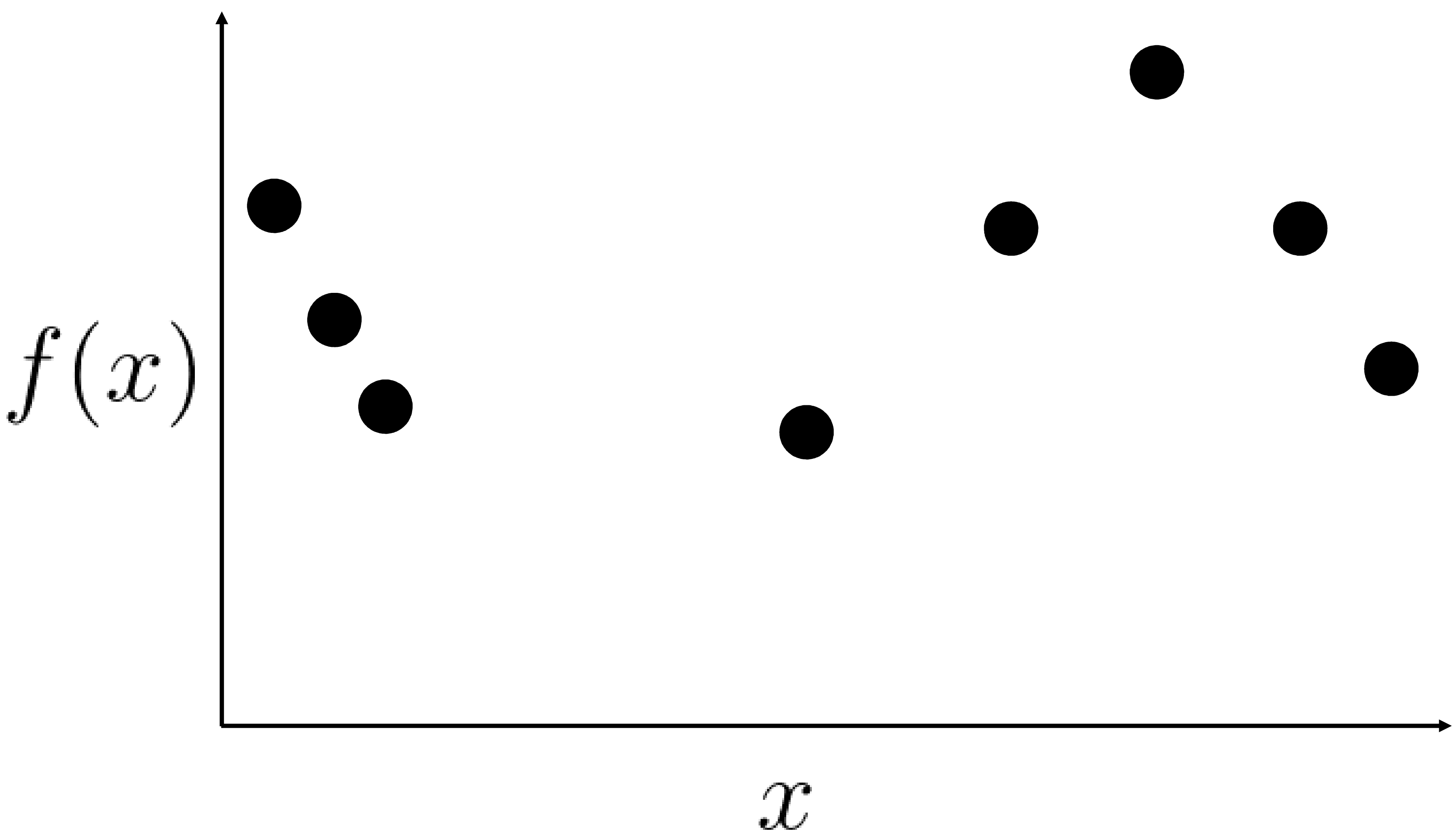


Genetic algorithm

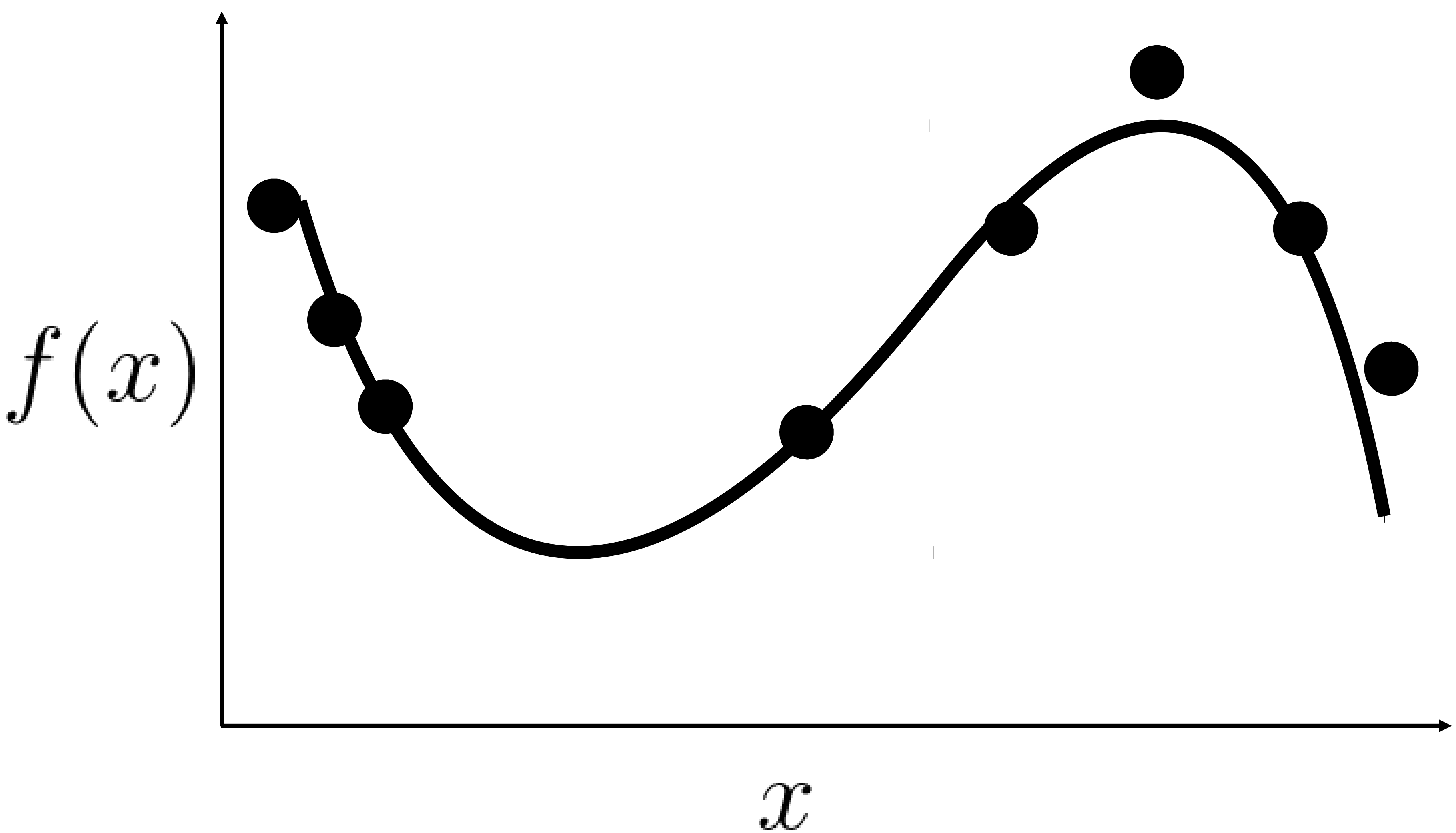


Bayesian Hyperparameter tuning

The problem



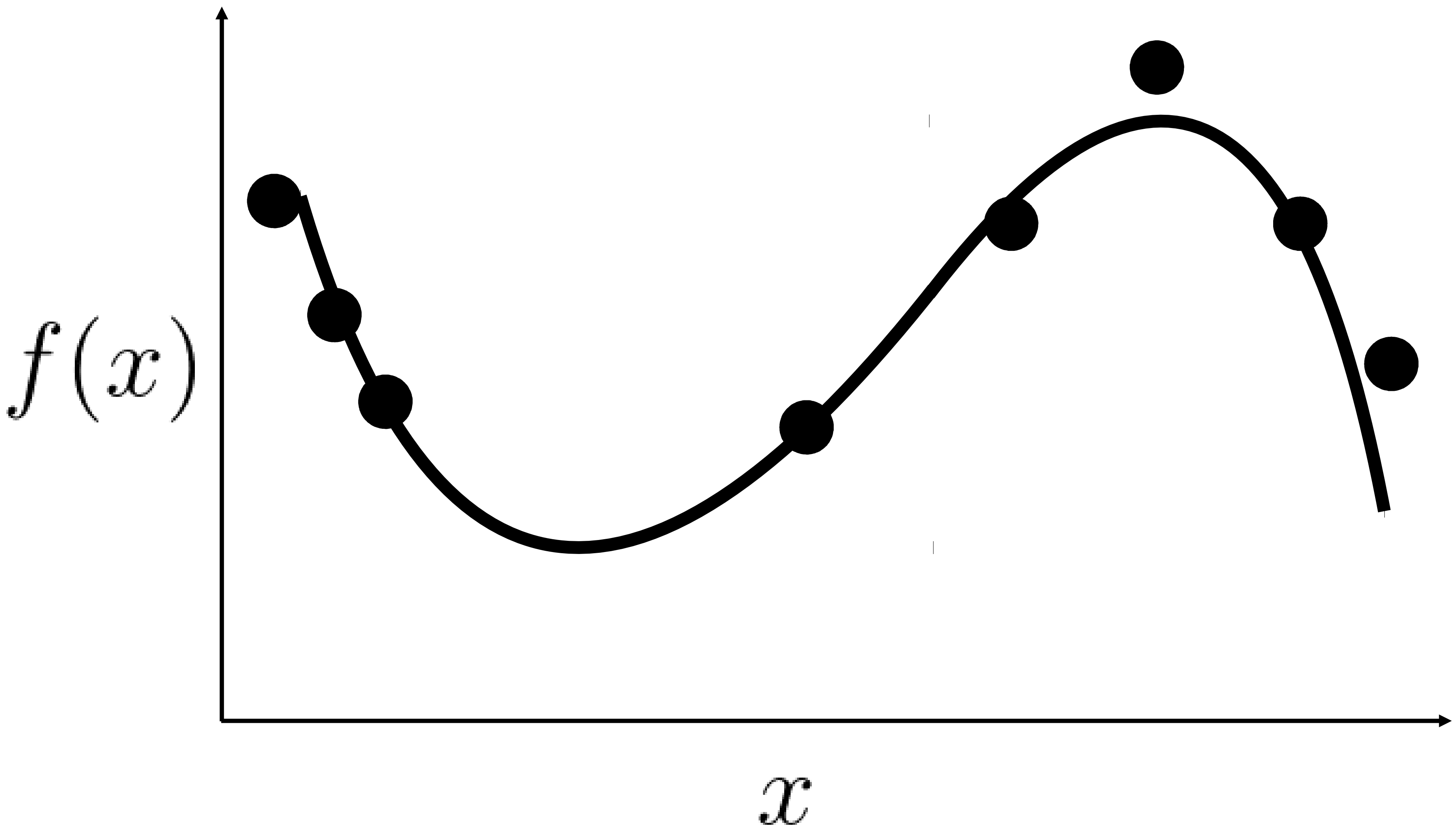
The problem



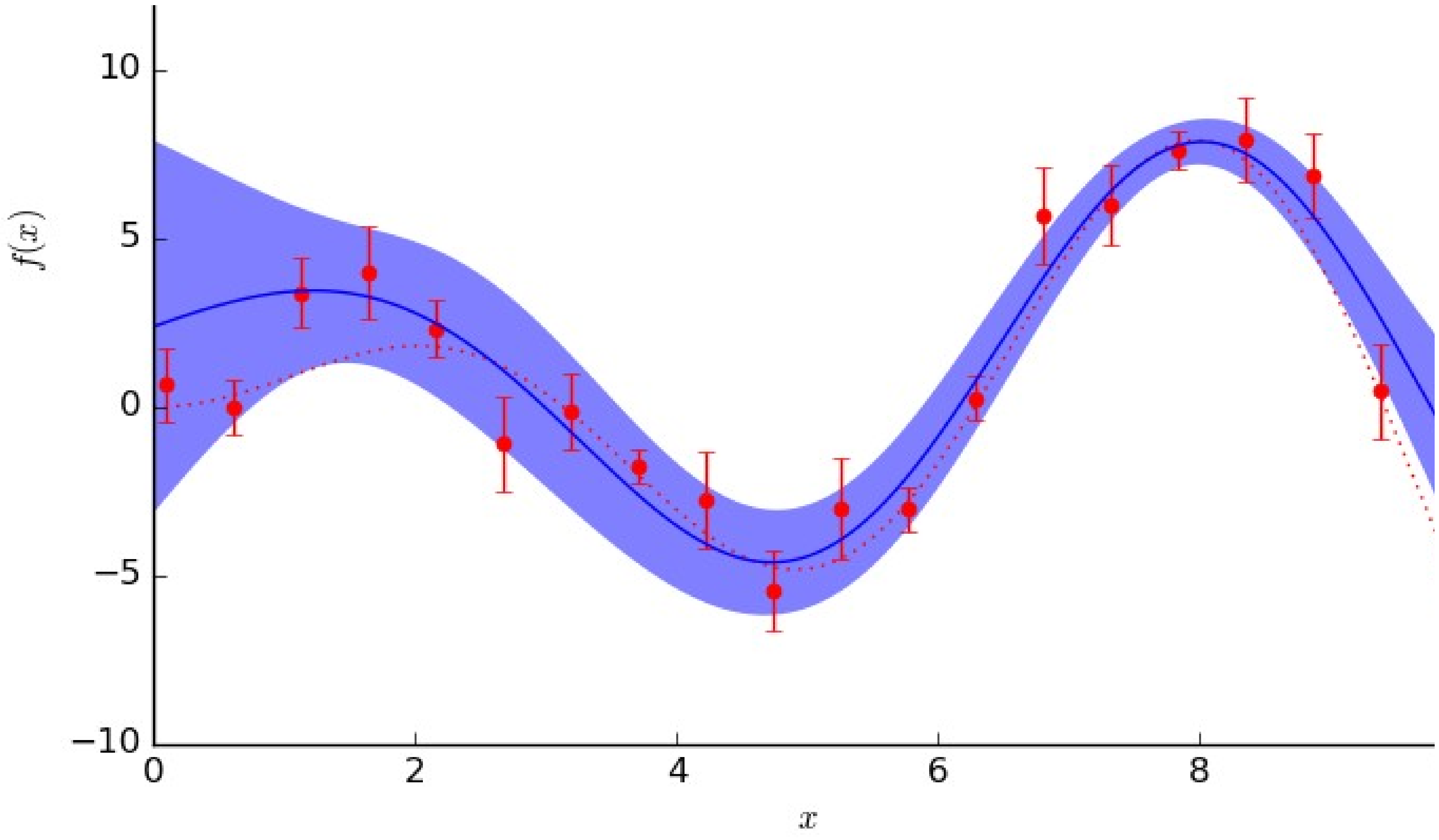
Gaussian process in a nut shell

ample ALL the functio

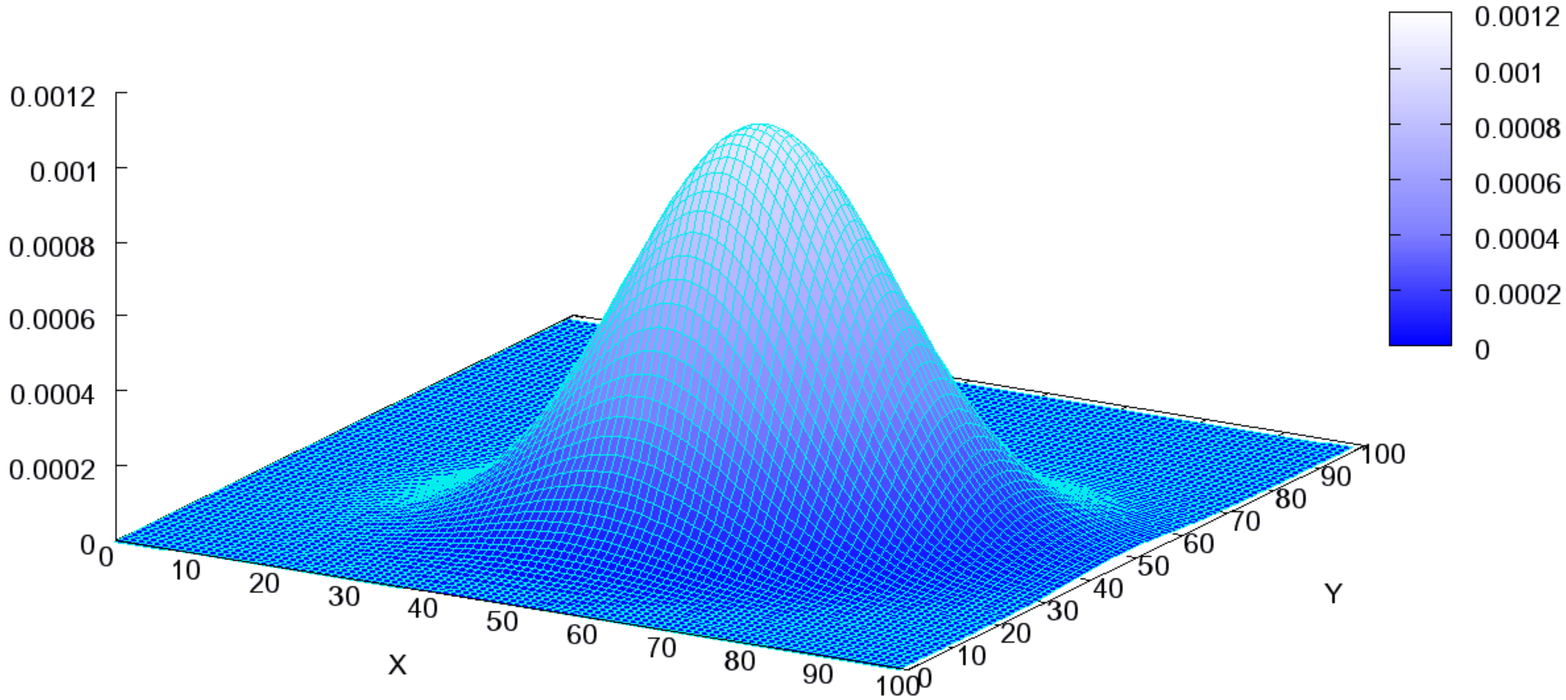
Instead of this



We want this



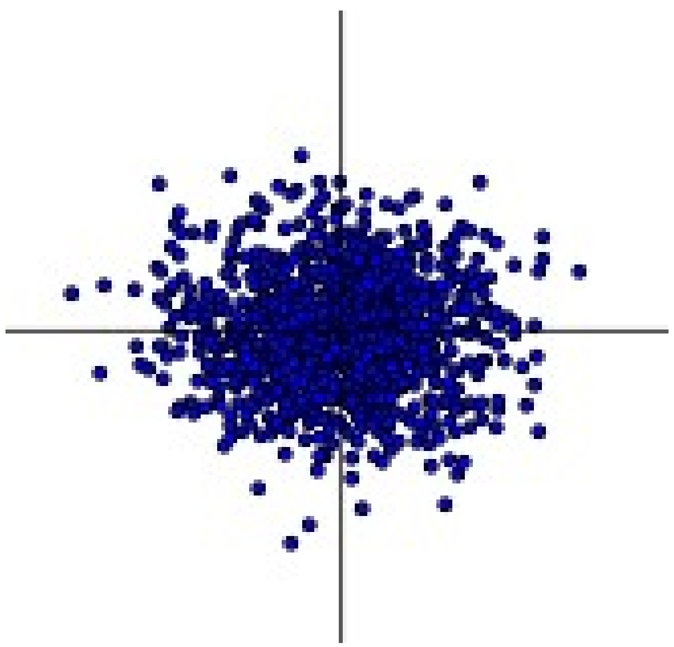
Gaussian distribution



$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right)$$

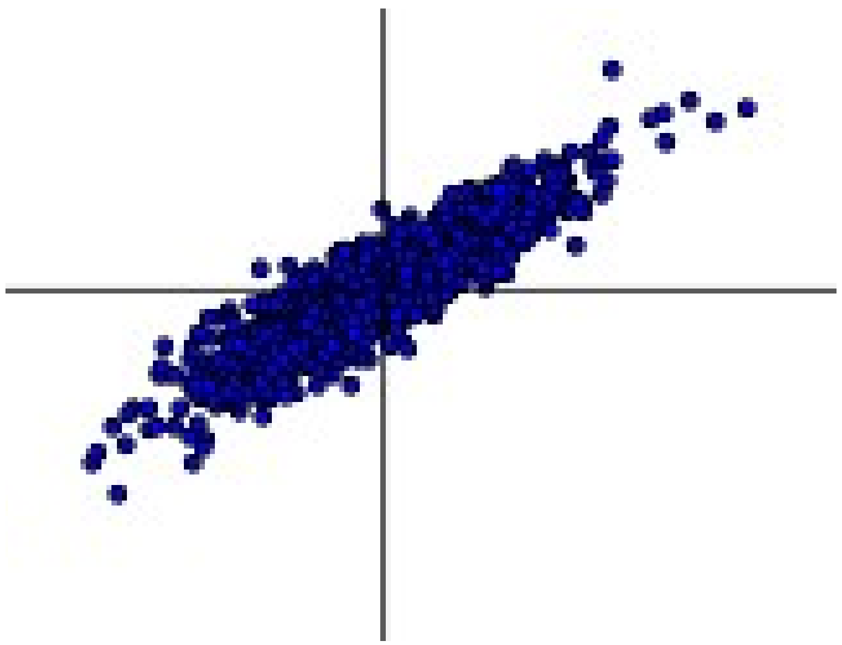
Gaussian distribution

x and y independent



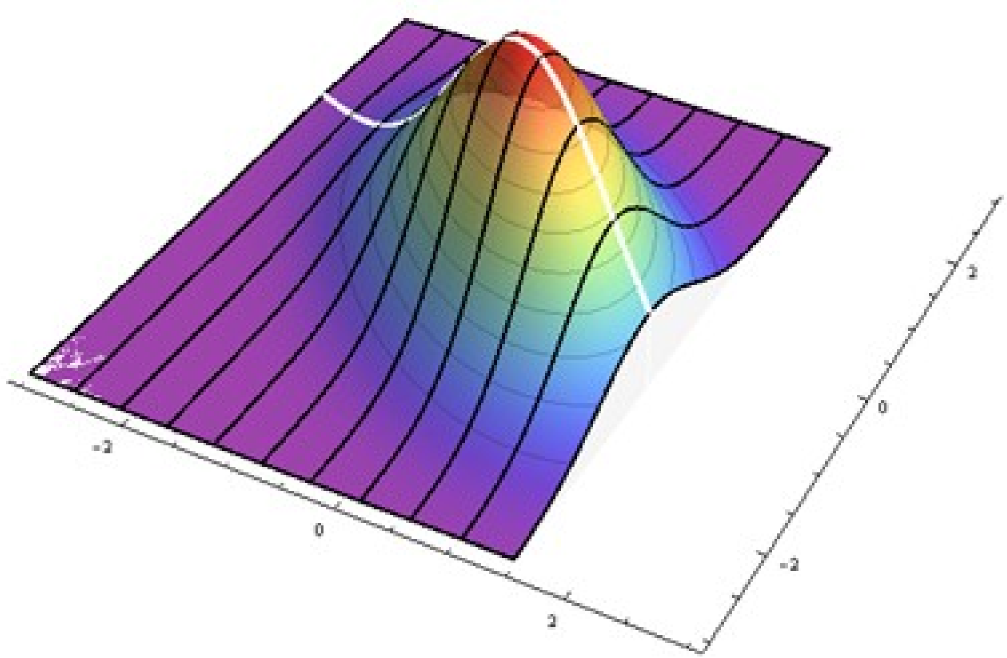
$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

x and y dependent



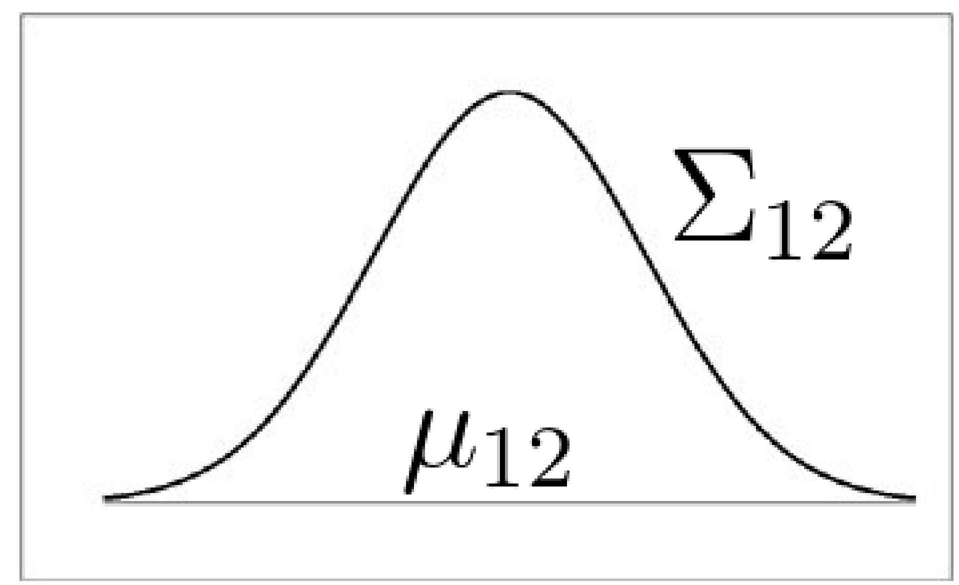
$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Conditional Gaussian



$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right)$$

$$P(x_2|x_1)?$$



$$\Sigma_{12} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\mu_{12} = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

Draw from Gaussian

$$X \sim N(0,1) \qquad X \sim N(\mu, \sigma^2)$$

$$X_i \sim \mu + \sigma N(0,1)$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right)$$

$$X_i \sim \mu + LN(0,1)$$

Gaussian Theorem

$$\mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}\right) \quad \Lambda = \Sigma^{-1} = \begin{bmatrix}\Lambda_{11} & \Lambda_{21}\\ \Lambda_{12} & \Lambda_{22}\end{bmatrix}$$

$$p(x_1) = N(x_1|\mu_1, \Sigma_{11})$$

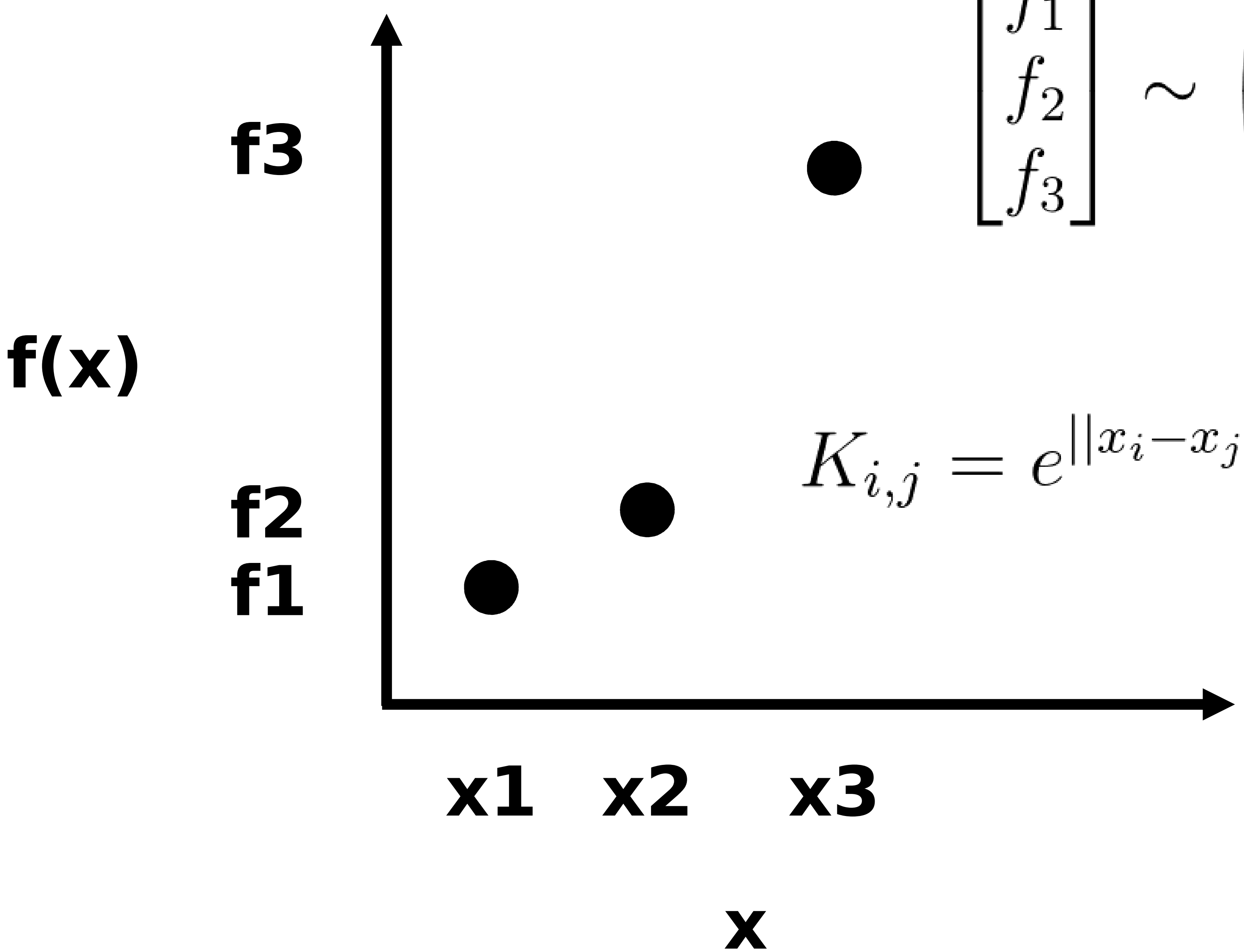
$$\mathrm{p}(\mathrm{x}_2) = N(x_2|\mu_2, \Sigma_{22})$$

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$$

$$\begin{aligned}\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\\ &= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right)\end{aligned}$$

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1}$$

Regression



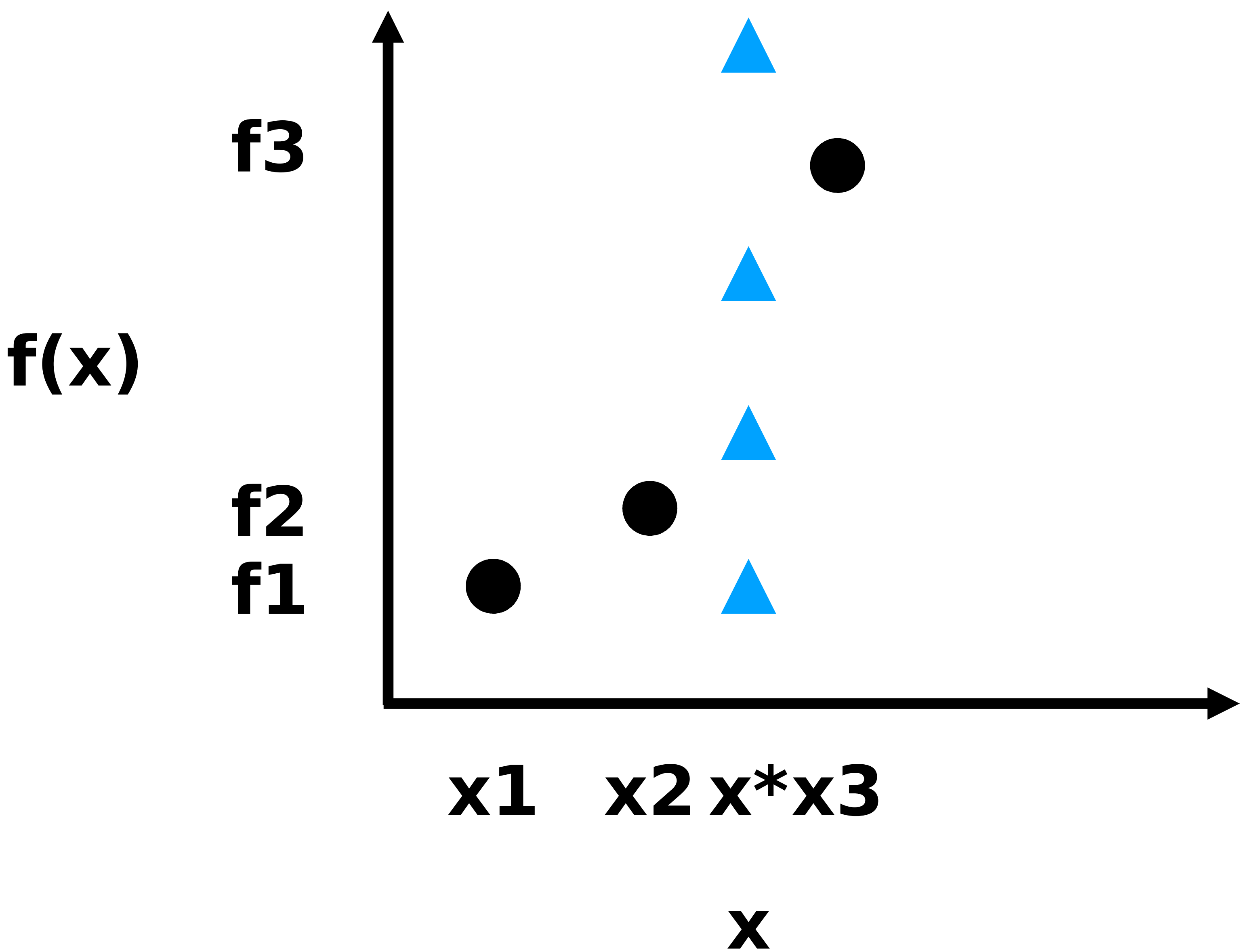
$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{bmatrix} \right)$$

$$K_{i,j} = e^{||x_i - x_j||^2} = \begin{cases} 0 \ if \ ||x_i - x_j|| \to 0 \\ 1 \ if \ x_i = x_j \end{cases}$$

Regression

$$f \sim N(0, K)$$

$$f_* \sim N(0, K_*)$$



Regression

$$f \sim N(0, K) \qquad f_* \sim N(0, K_*)$$

$$\begin{bmatrix} f \\ f_* \end{bmatrix} = N(0, \begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{1*} \\ K_{21} & K_{22} & K_{23} & K_{2*} \\ K_{31} & K_{32} & K_{33} & K_{3*} \\ K_{*1} & K_{*2} & K_{*3} & K_{**} \end{bmatrix})$$

$$p(f_* | f)?$$

Gaussian process

- **If x and x* are similar, thus f(x) and f(x*) are similar**

- **Covariance K(x,x*) return similarity, and thus also encodes similarity between f(x), f(x*)**

- **The describe unknown function f we sample x and convert them to known f(x)**

How do we use this?

- **Sample pints with expected improvement**

- **Normally they are close to already know data point**

- **Seek places for high variance, or low mean**

- **If no improvement we haven't lost anything**

Thank you for your attention

Questions?

References

Gaussian process:

- http://katbailey.github.io/post/gaussian-processes-for-dummies/
- http://www.cs.ubc.ca/~nando/540-2013/lectures/l6.pdf

Bayesian optimization

- http://neupy.com/2016/12/17/hyperparameter_optimization_for_neural_networks.html#gaussian-process
- http://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/slides/lec21.pdf

Paper for GP optimization:

- https://arxiv.org/pdf/1206.2944.pdf