FACULTY OF ENGINEERING
AND ARCHITECTURE

DEPARTMENT INFORMATION TECHNOLOGY
RESEARCH GROUP IDLAB

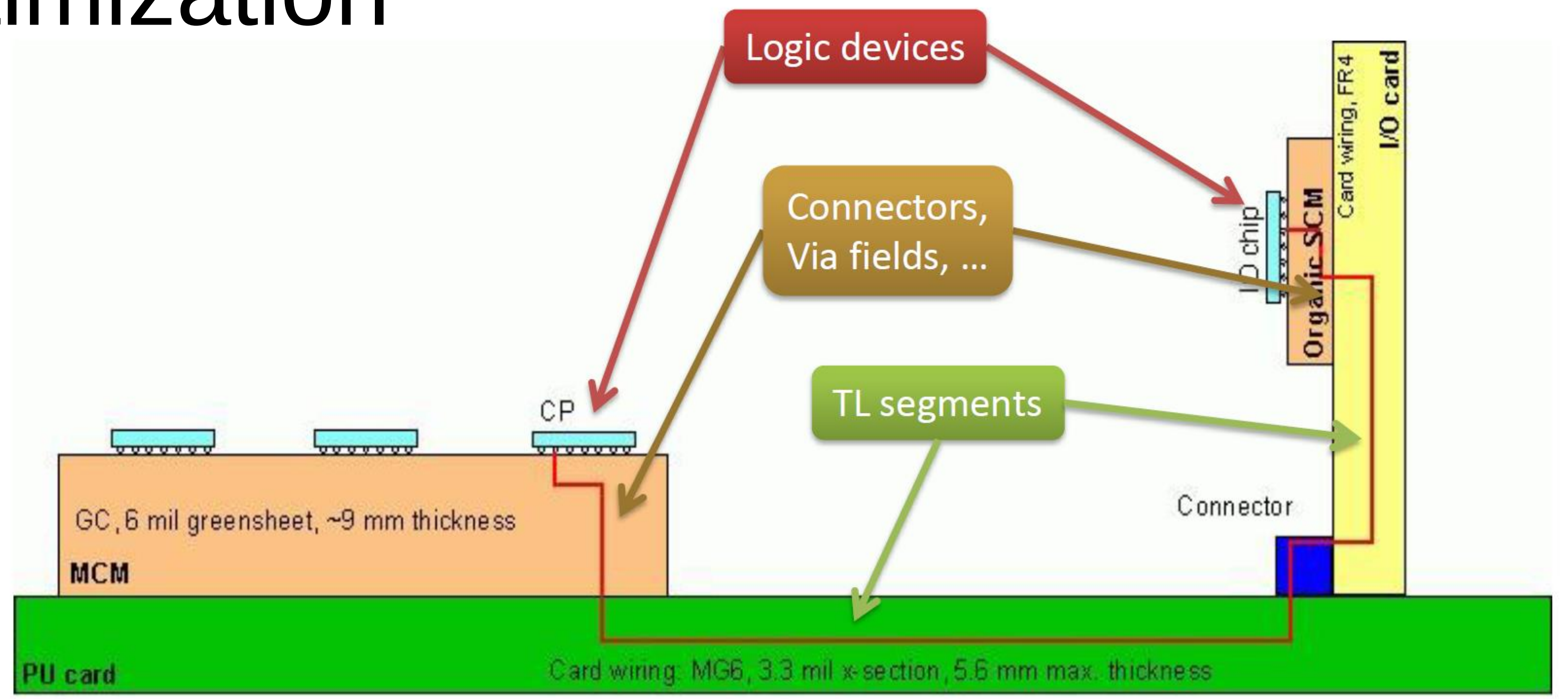
BAYESIAN OPTIMIZATION

Automated machine learning – Ivo Couckuyt

GHENT
UNIVERSITY

INTRODUCTION

- **Example**: Computer-aided design (CAD)
  - Easy prototyping
  - Design space exploration and optimization

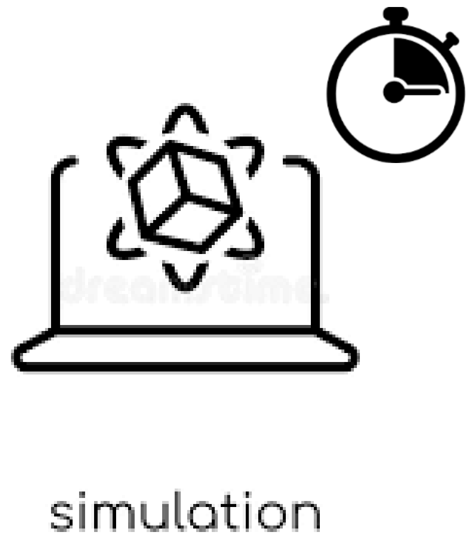


Design variables

- **But**…complex simulations
  - Many design requirements
  - Large-scale
  - …
- Difficult to design and characterize

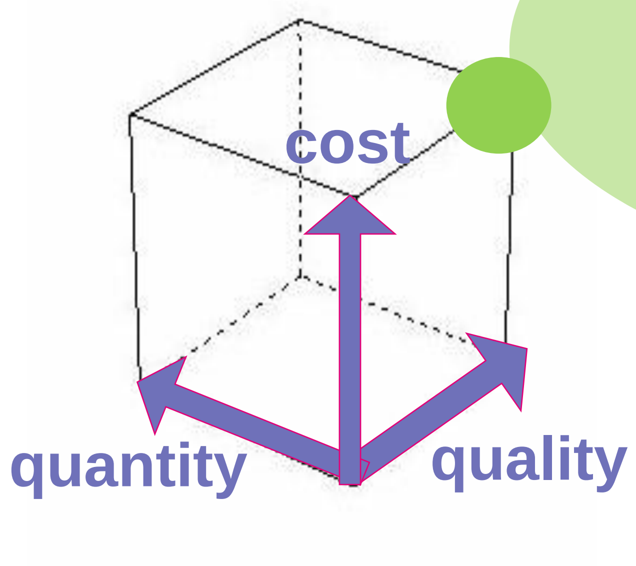
GHENT UNIVERSITY

INTRODUCTION



Small data

- **Physics-based** simulations
  - Finite elements, fluid dynamics, etc.
  - Time-consuming
    - Ford: *"36-160 hours for 1 crash simulation"*

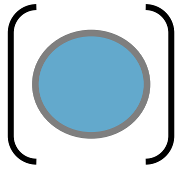


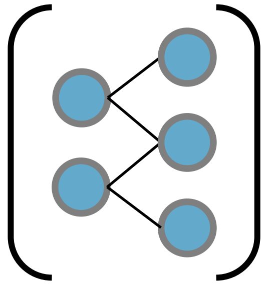
- Quantity = small
- Quality = high
- Cost = high

INTRODUCTION

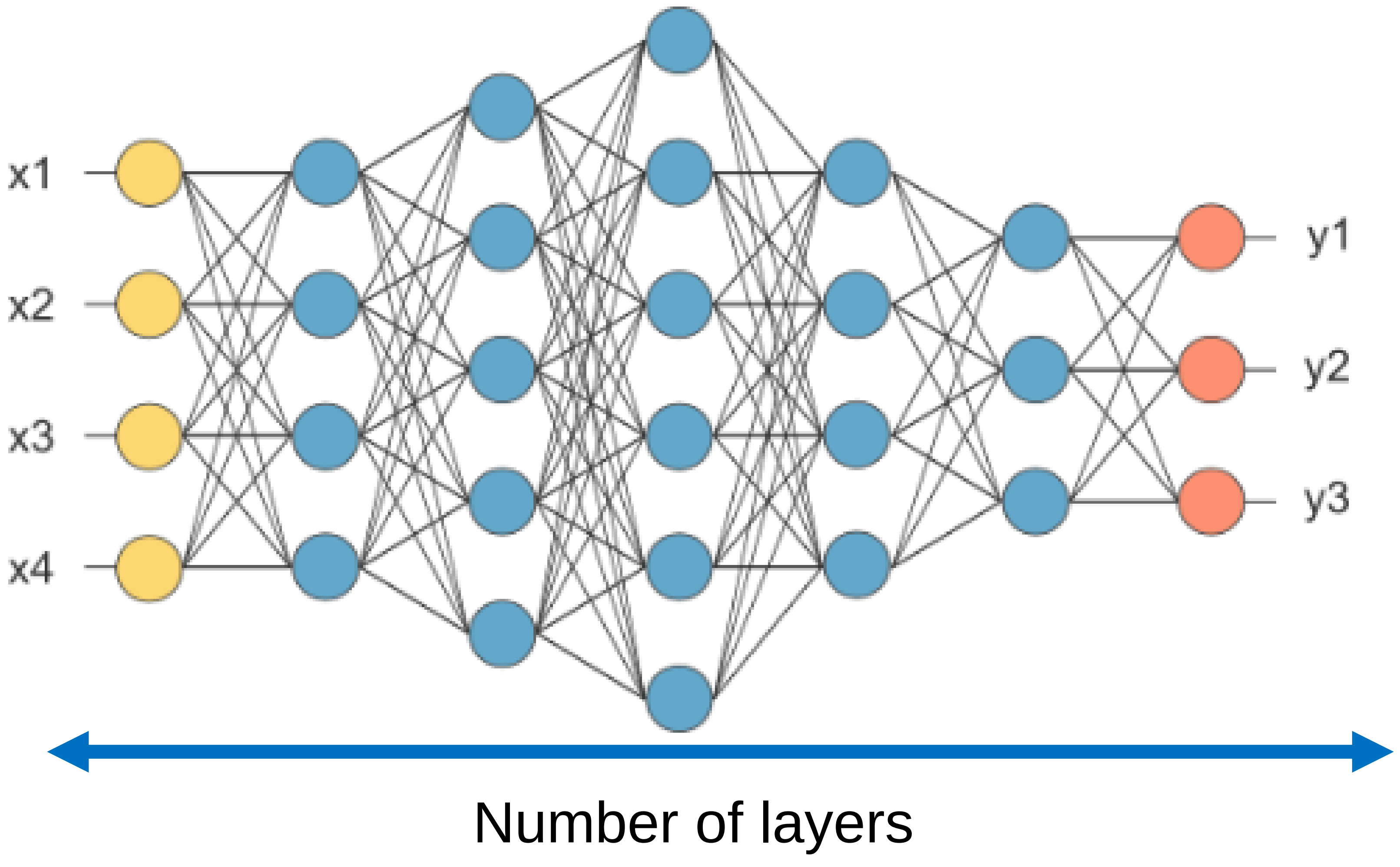
- **Example**: Large neural networks
  - Very successful
  - Visual object detection, speech recognition,…

- **But**… expensive to train
- Many choices
  - Number of neurons
  - Number of layers
  - Learning rate
  - …



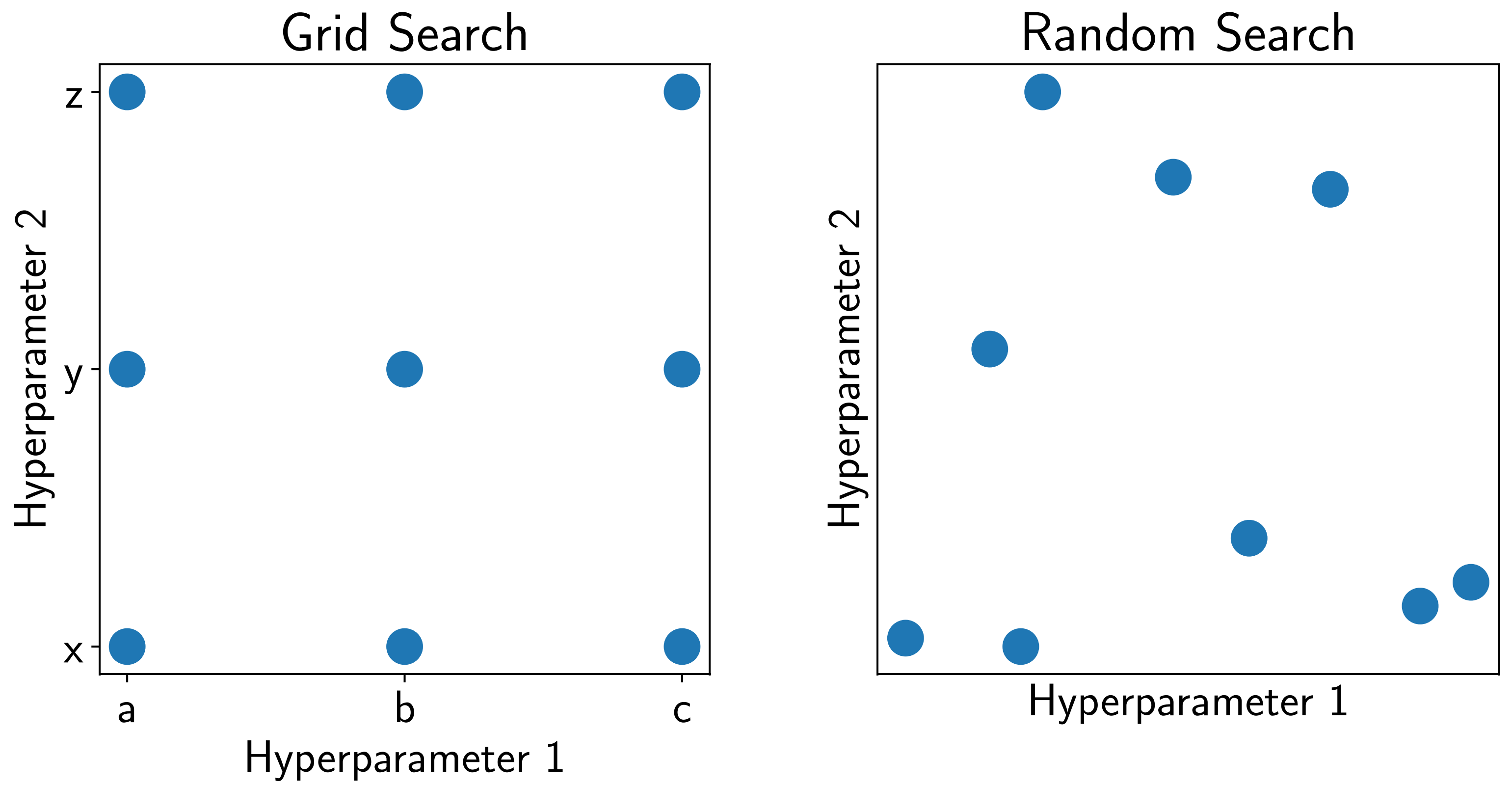


Hyperparameters



SEARCH FOR HYPERPARAMETERS

- **How do people currently search?**
  - Trial-and-error
  - Grid search
  - Random search
- **Painful!** Requires many training cycles
  - Exponential increase for grid search



GHENT UNIVERSITY

GLOBAL OPTIMIZATION

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathcal{X}}{\operatorname{argmin}} f(\boldsymbol{x})$$

x : Variables of interest

f(**x**): Objective function

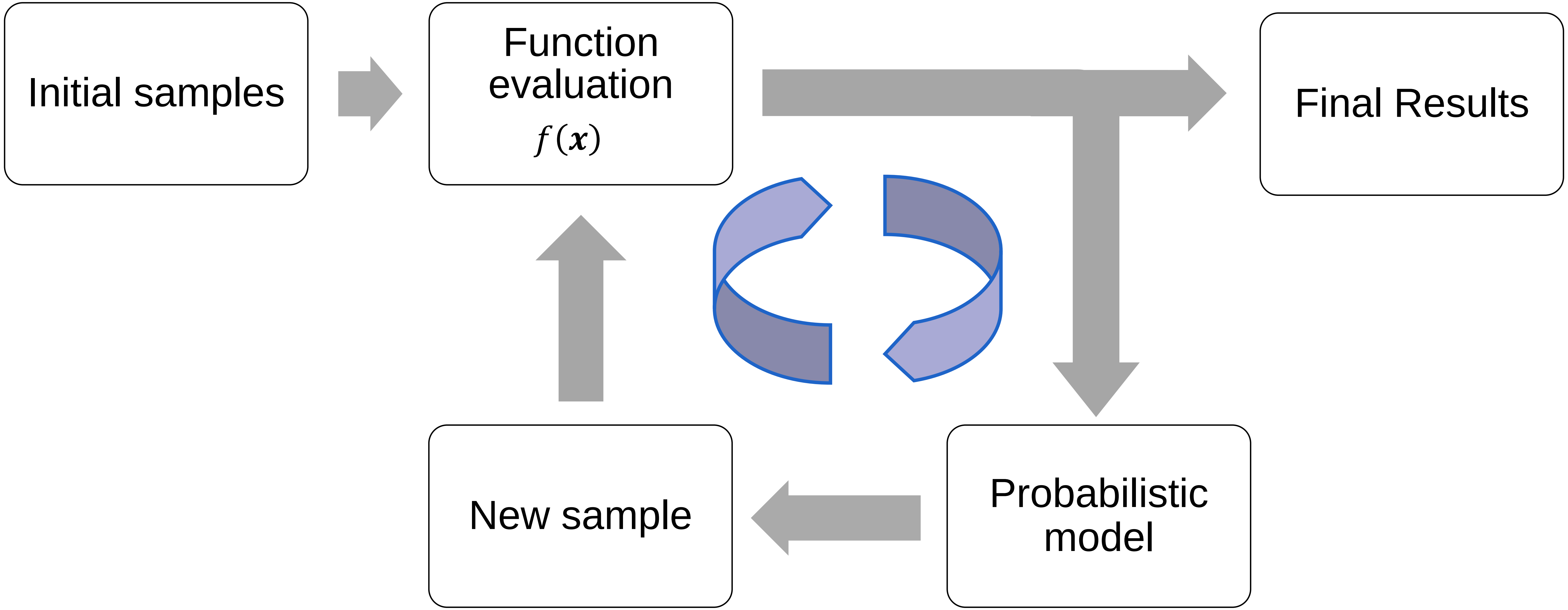
– Behavior unknown
– Time-consuming

Bayesian optimization

- A probabilistic method for data-efficient global optimization
- Minimizes $f(\boldsymbol{x})$ **and** the number of evaluations

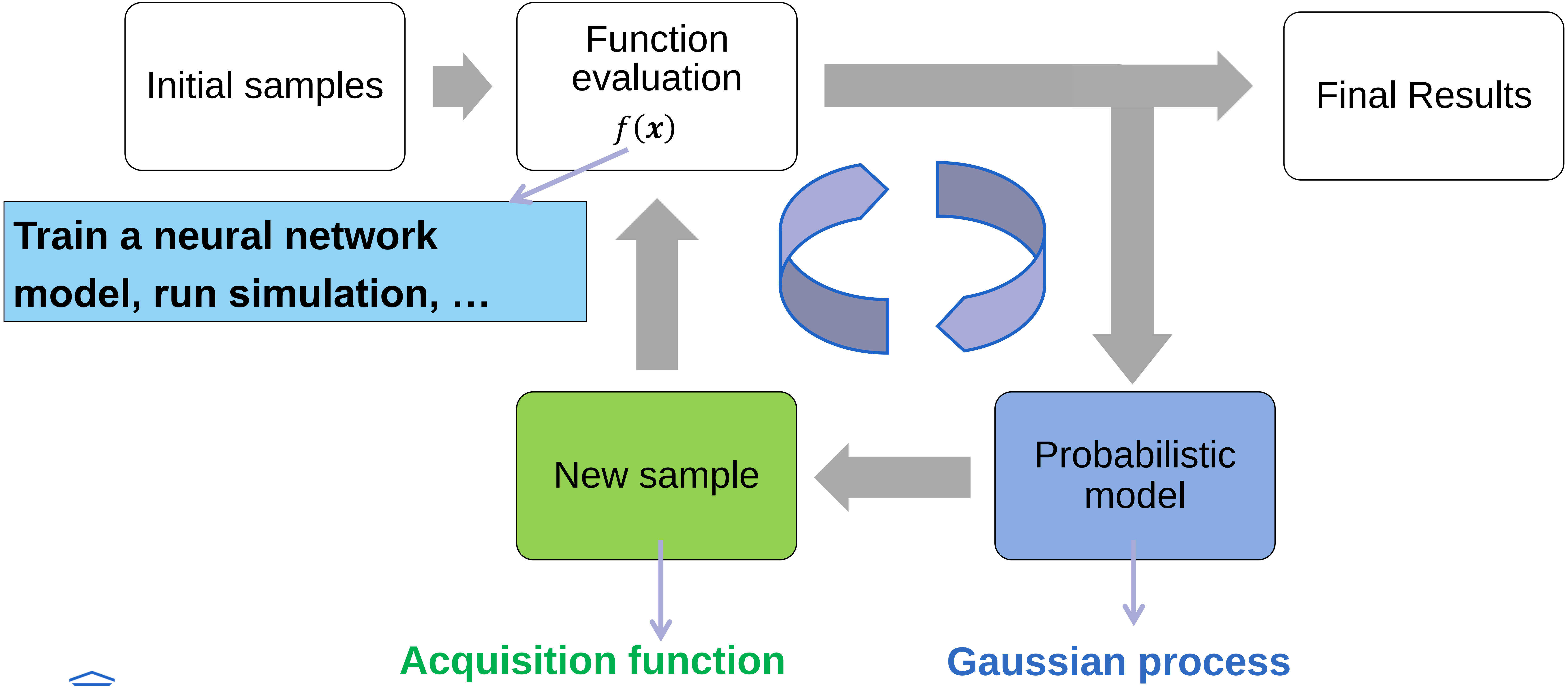
GHENT UNIVERSITY

BAYESIAN OPTIMIZATION



GHENT UNIVERSITY

BAYESIAN OPTIMIZATION

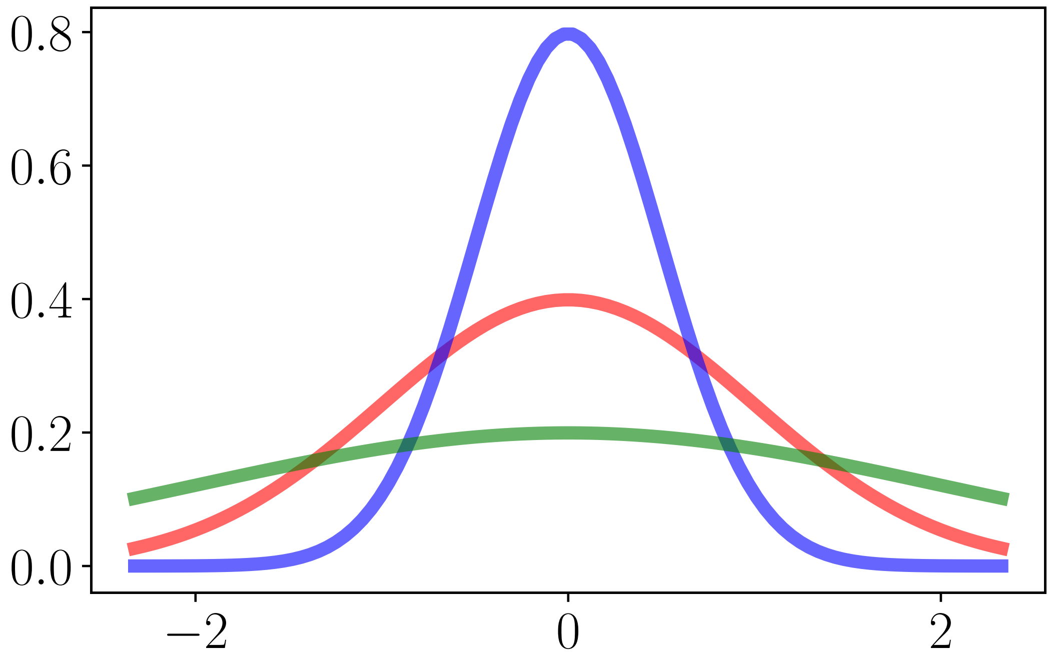


NORMAL DISTRIBUTION

– Gaussian (normal) distribution

$$y \sim \mathcal{N}(0, \sigma^2)$$

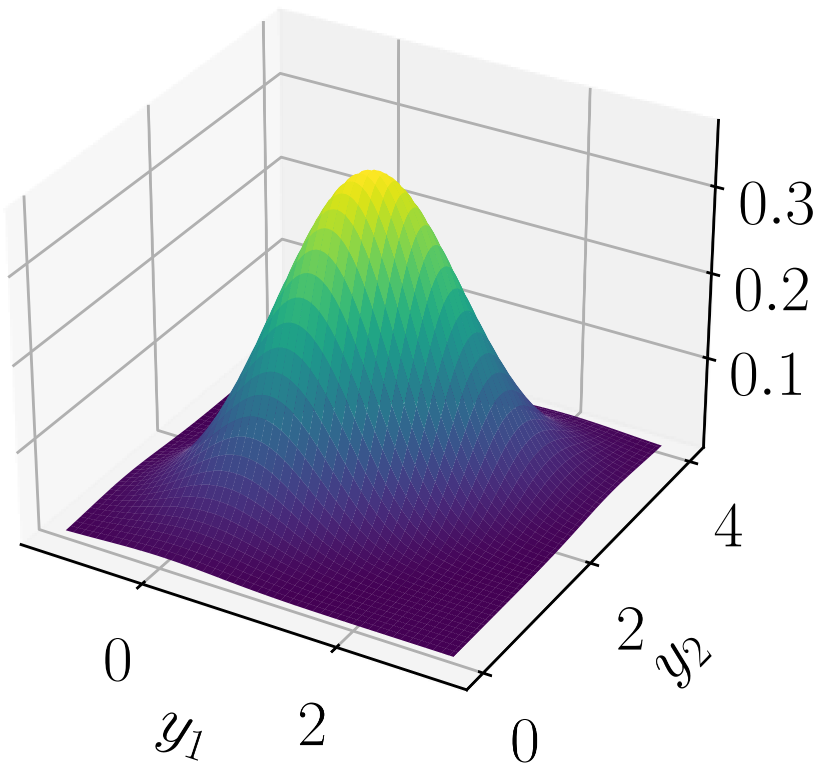
Mean (often **0**)   Variance



– **Multivariate** Gaussian (normal) distribution

$$\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \mathrm{K})$$

Mean (often **0**)   Kernel (or covariance) matrix



GAUSSIAN PROCESS

Gaussian distribution

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathrm{K})$$

distribution over **vectors**

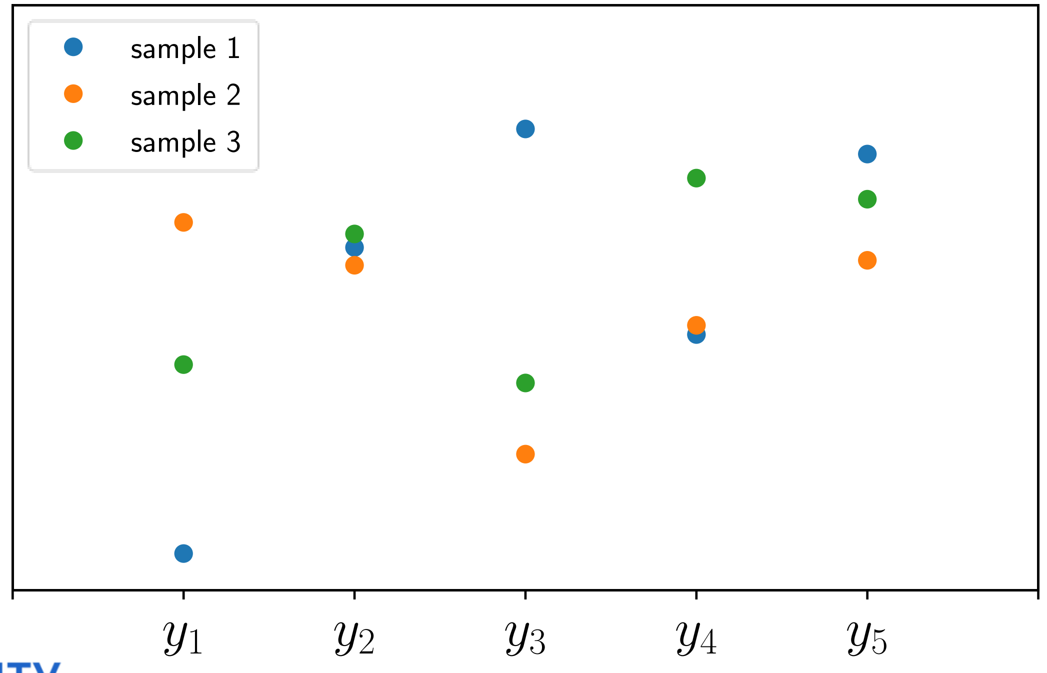
fully specified by a mean & **covariance**

Gaussian Process

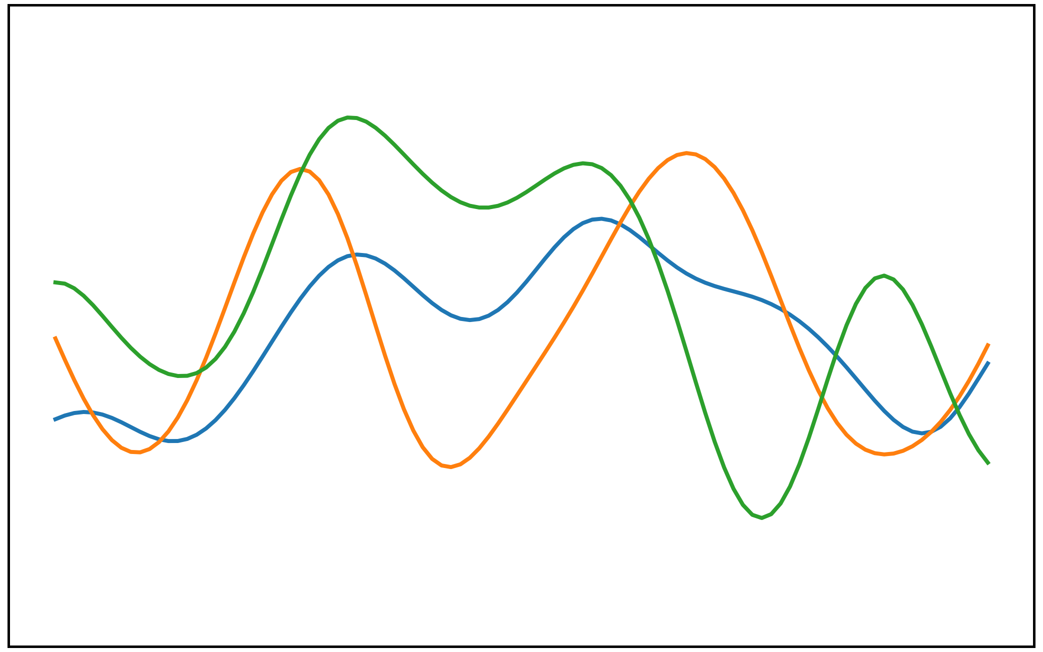
$$f \sim \mathcal{GP}\left(\mathbf{0}, k\left(\boldsymbol{x}^i, \boldsymbol{x}^j\right)\right)$$

distribution over **functions**

fully specified by a mean function & **kernel function** (or covariance)

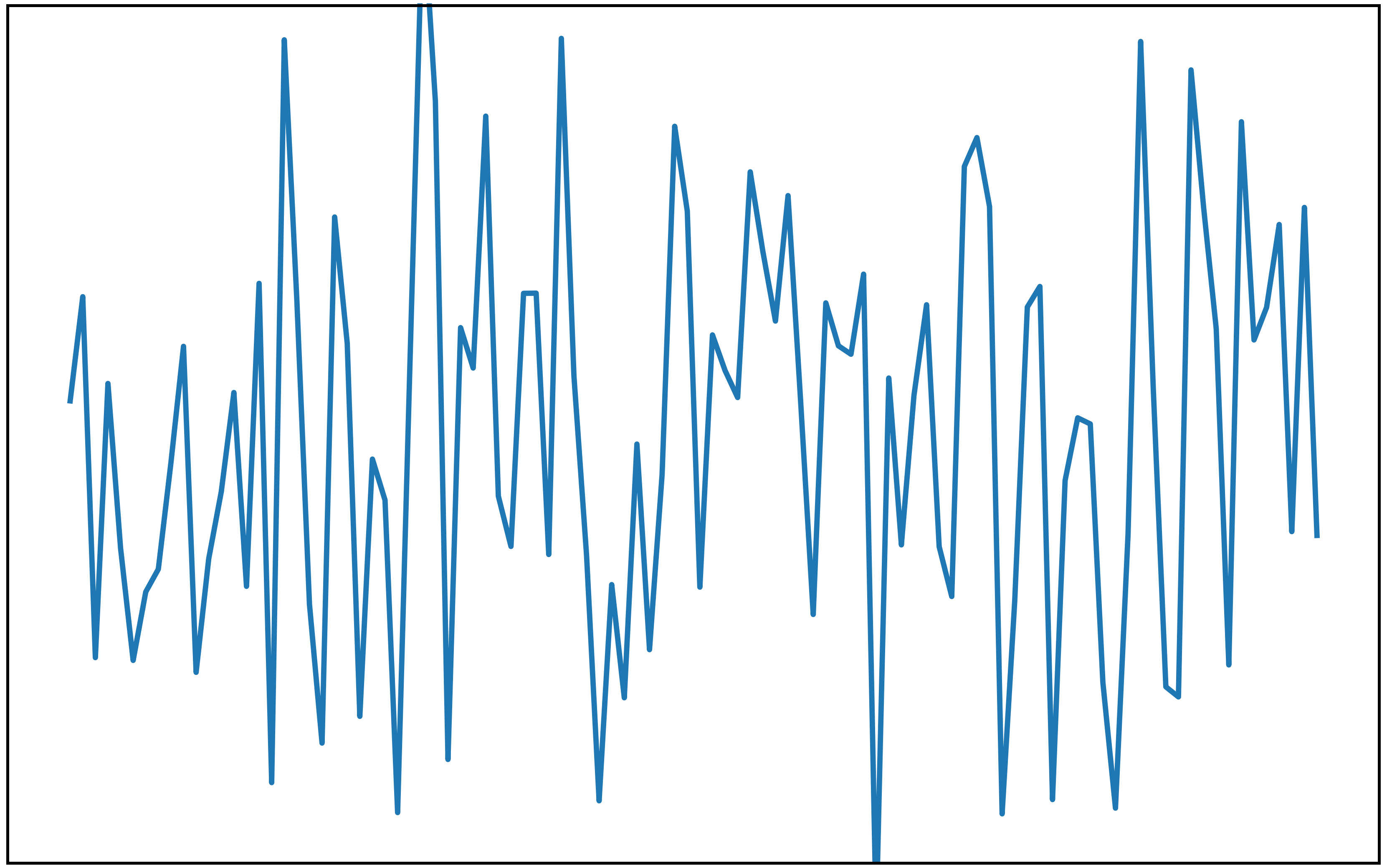


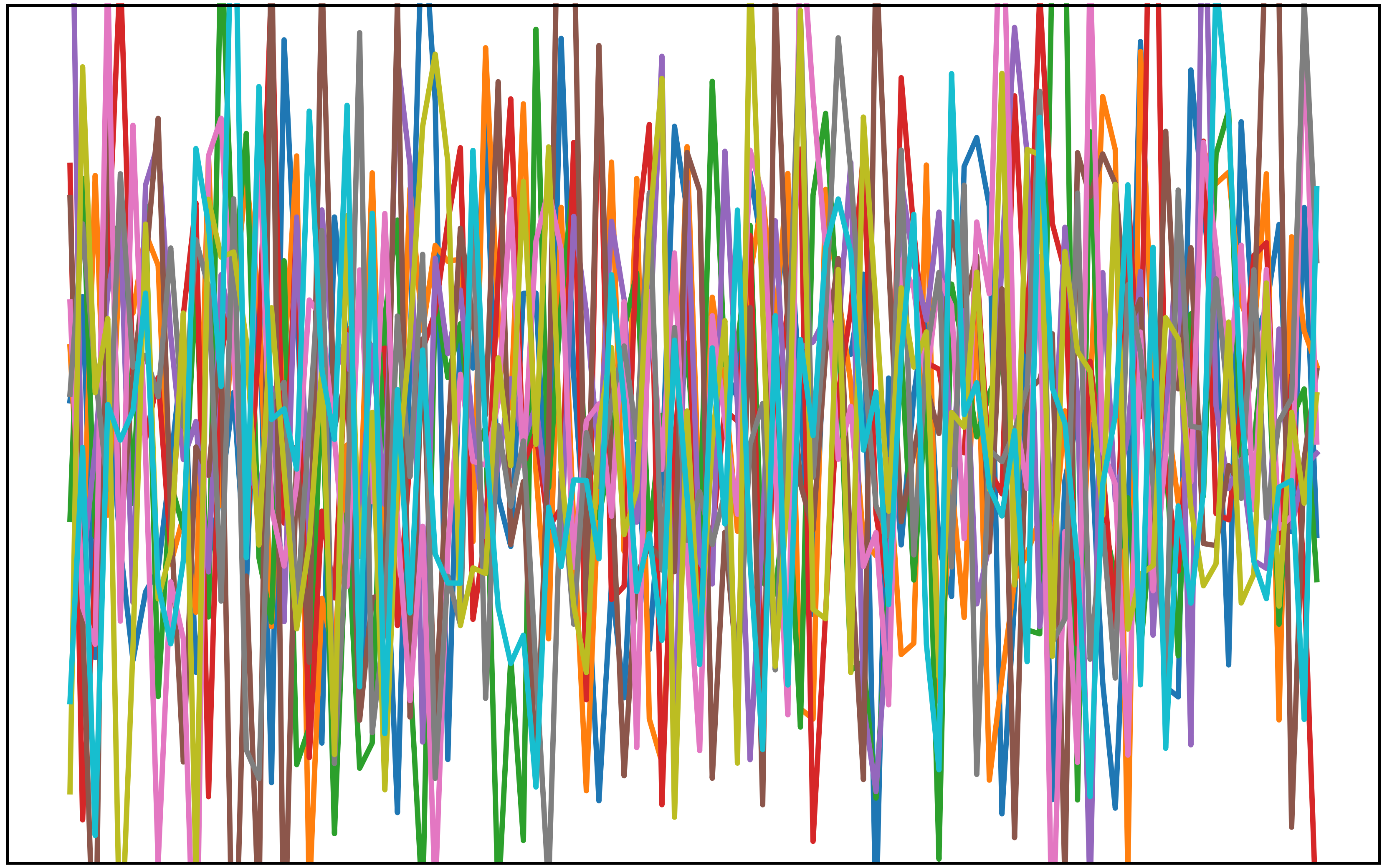
infinite vector **y**



GAUSSIAN PROCESS

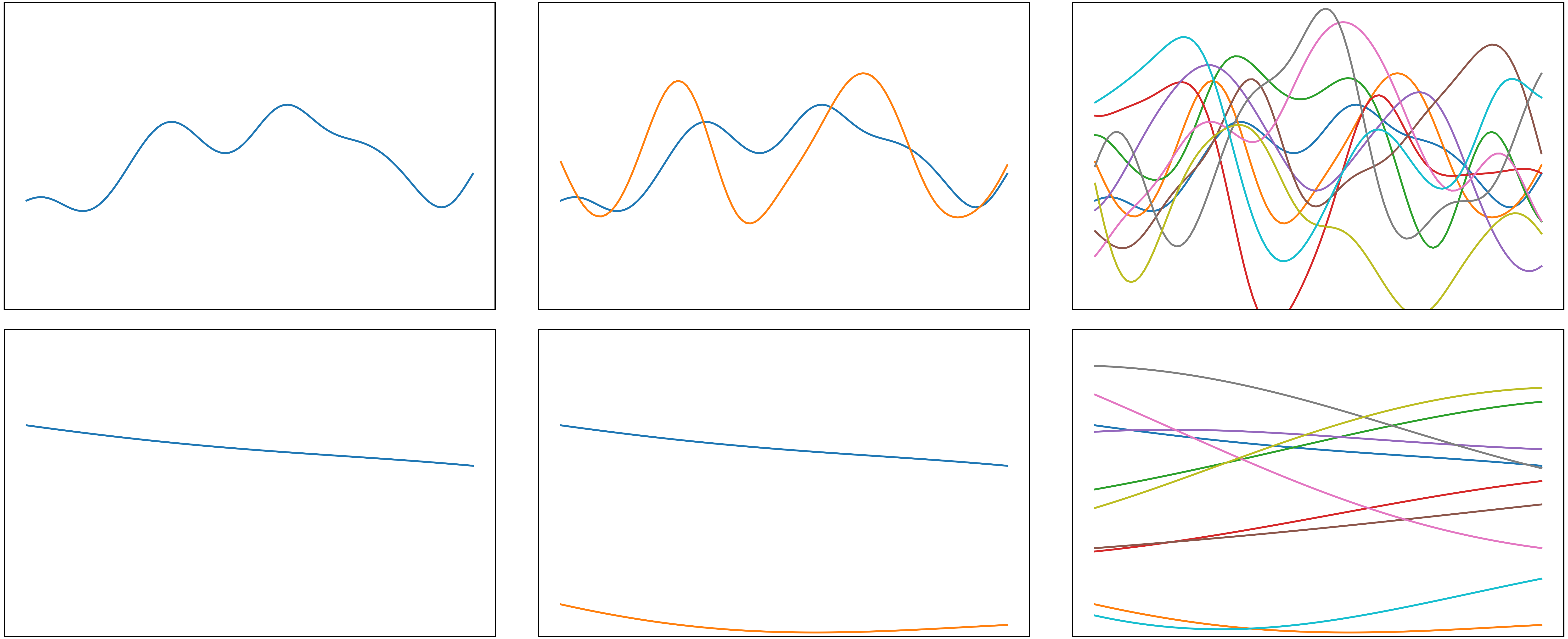
- Sample from $\mathcal{N}(\boldsymbol{y}|\mathbf{0}, K)$
- Let $K = I$ (identity matrix)
  - Independent normal distributions





GAUSSIAN PROCESS

- Sample from $\mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, K)$
- Let $K_{i,j} = k(x^i, x^j)$ (squared exponential)



KERNEL FUNCTION

- Kernel: How similar are two points?
- Example: The **Squared Exponential** (**SE**) kernel
  - Weighted distance

$$k(\boldsymbol{x}^i, \boldsymbol{x}^j) = \sigma_f^2 \exp(-\frac{1}{2l^2}(\boldsymbol{x}^i - \boldsymbol{x}^j)^2)$$

signal variance

lengthscale

Hyperparameters θ

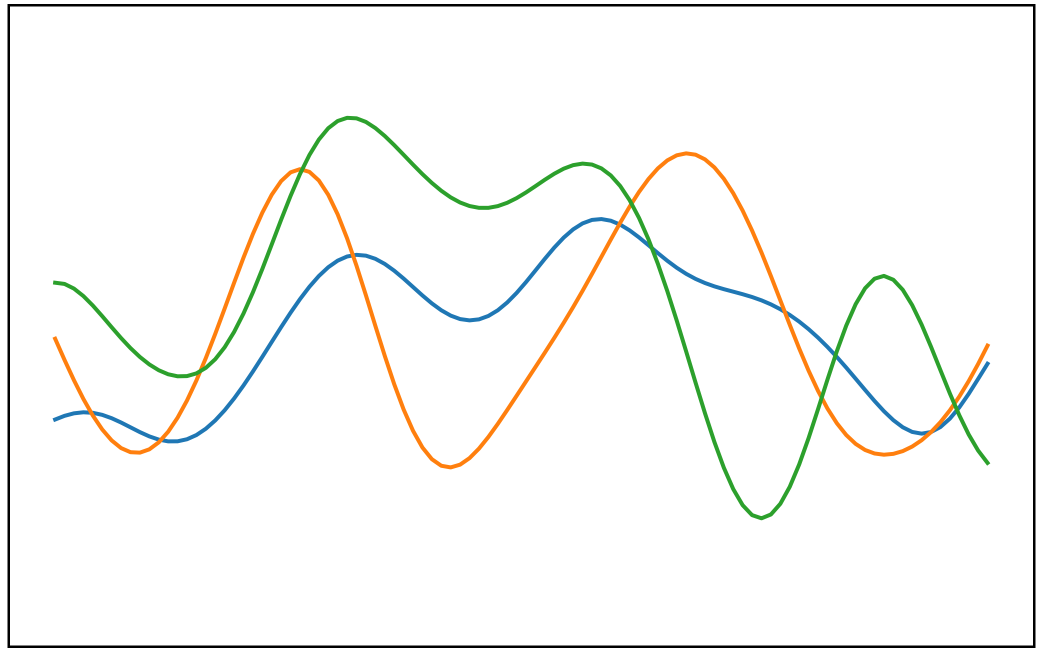
14

GHENT UNIVERSITY

GAUSSIAN PROCESS

- Prior (no data)
  - Assumptions about $f(\boldsymbol{x})$

$$f \sim \mathcal{GP}(\mathbf{0}, k(\boldsymbol{x}^i, \boldsymbol{x}^j))$$



- Posterior (training of model)
  - Updated belief based on the data set
  - Uses Bayes theorem!

$$\mathrm{f}(\mathbf{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$$

$\mu(\boldsymbol{x})$ → Prediction

$\sigma^2(\boldsymbol{x})$ → Uncertainty (confidence)

GHENT UNIVERSITY

LIKELIHOOD

- Likelihood: training model
  - Hyperparameters θ

$$\mathcal{L}(\theta) = -\log \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, K_\theta) = \frac{1}{2}\log|2\pi K_\theta| + \frac{1}{2}\boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}$$

Cubic

Capacity control
Regularization

Data-fit term

- **Needs to be optimized**
  - E.g., gradient descent
  - Expensive (but not for small data)

GAUSSIAN PROCESS TIME COMPLEXITY

– Mean: model prediction

$$\mu(\boldsymbol{x}_{1:m}) = k(\boldsymbol{x}_{1:m}, \boldsymbol{x}_{1:n})^T \underbrace{(K + \sigma_n^2 I)^{-1} f_{1:n}}_{\text{Constant: } \gamma}$$

$m \times n$ $n \times 1$

Linear

– Variance: model uncertainty

$$\sigma^2(\boldsymbol{x}_{1:m}) = k(\boldsymbol{x}_{1:m}, \boldsymbol{x}_{1:m}) - k(\boldsymbol{x}_{1:m}, \boldsymbol{x}_{1:n})^T \underbrace{(K + \sigma_n^2 I)^{-1}}_{\text{Constant}} k(\boldsymbol{x}_{1:m}, \boldsymbol{x}_{1:n})$$

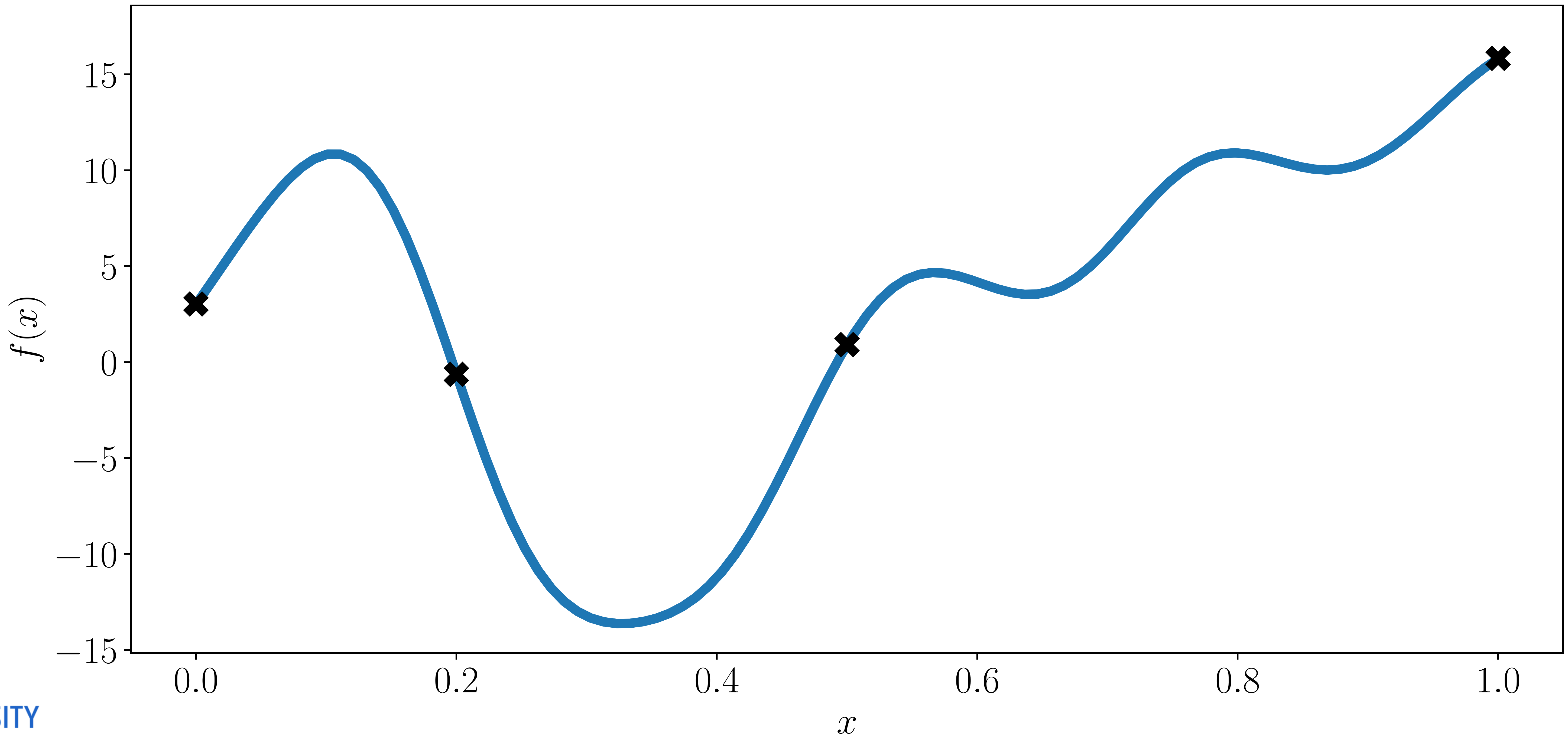
$m \times m$ $m \times n$ $n \times n$ $n \times m$

Quadratic

GHENT UNIVERSITY

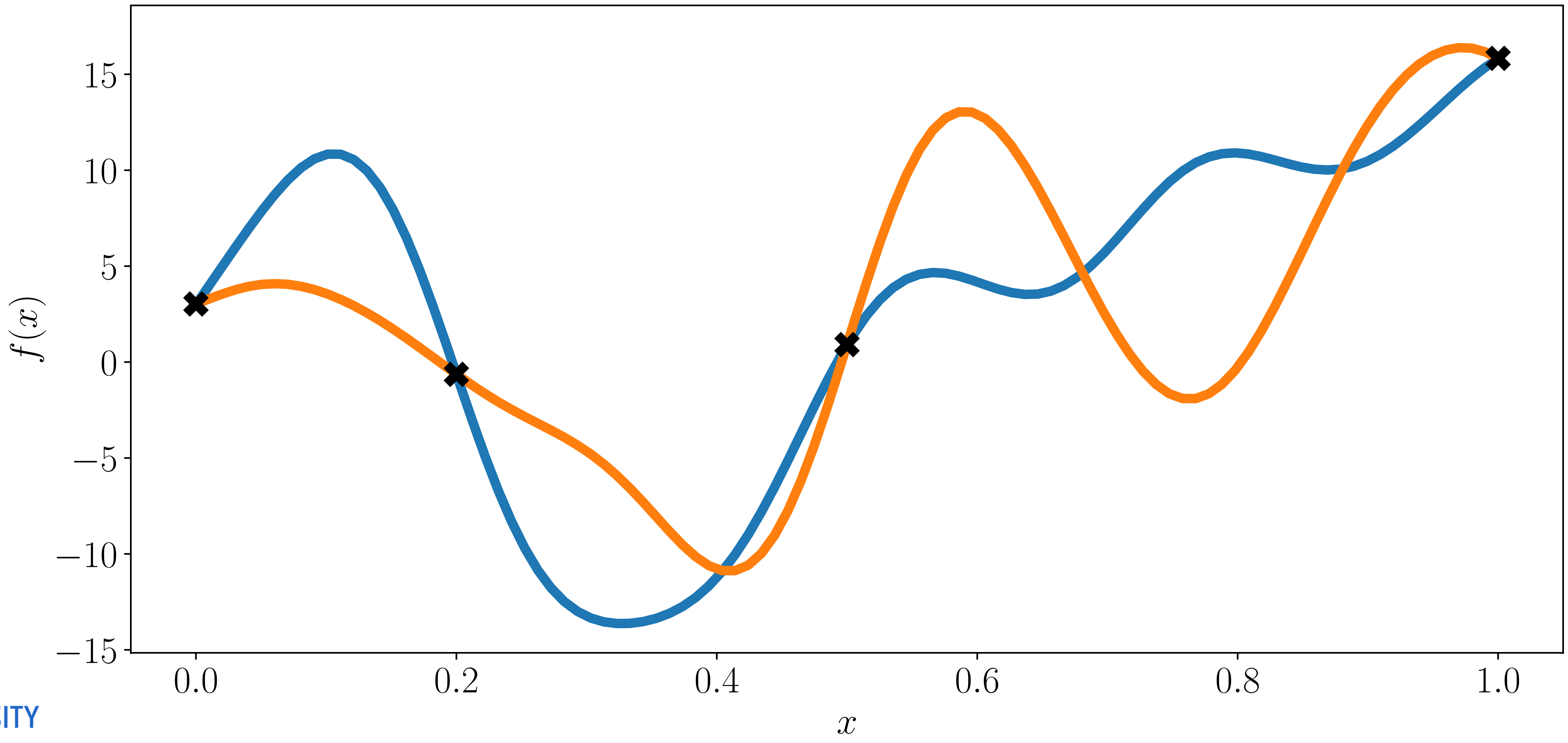
GAUSSIAN PROCESS POSTERIOR

Sample from $\mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$



GAUSSIAN PROCESS POSTERIOR

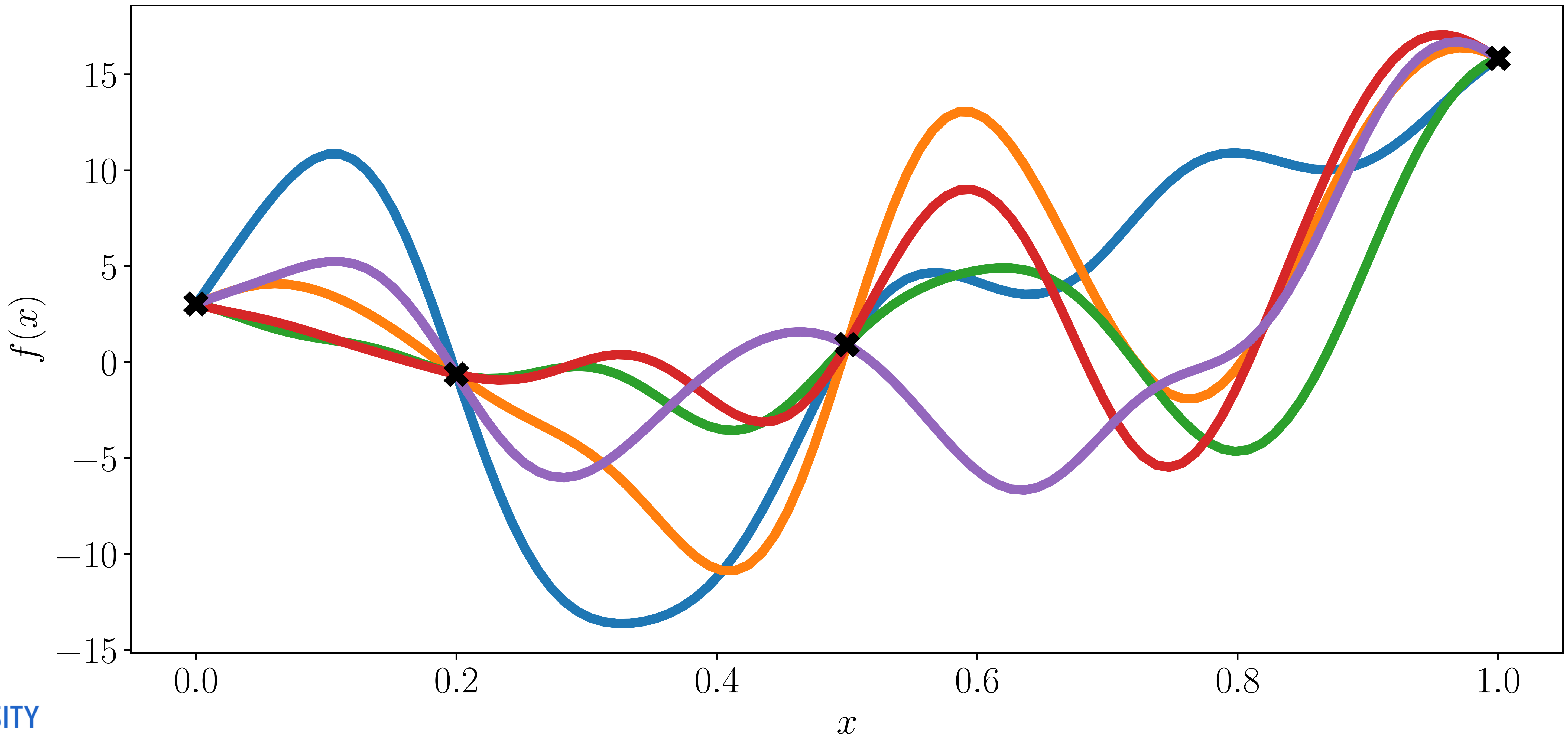
Samples from $\mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$



GHENT UNIVERSITY

GAUSSIAN PROCESS POSTERIOR

Samples from $\mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$

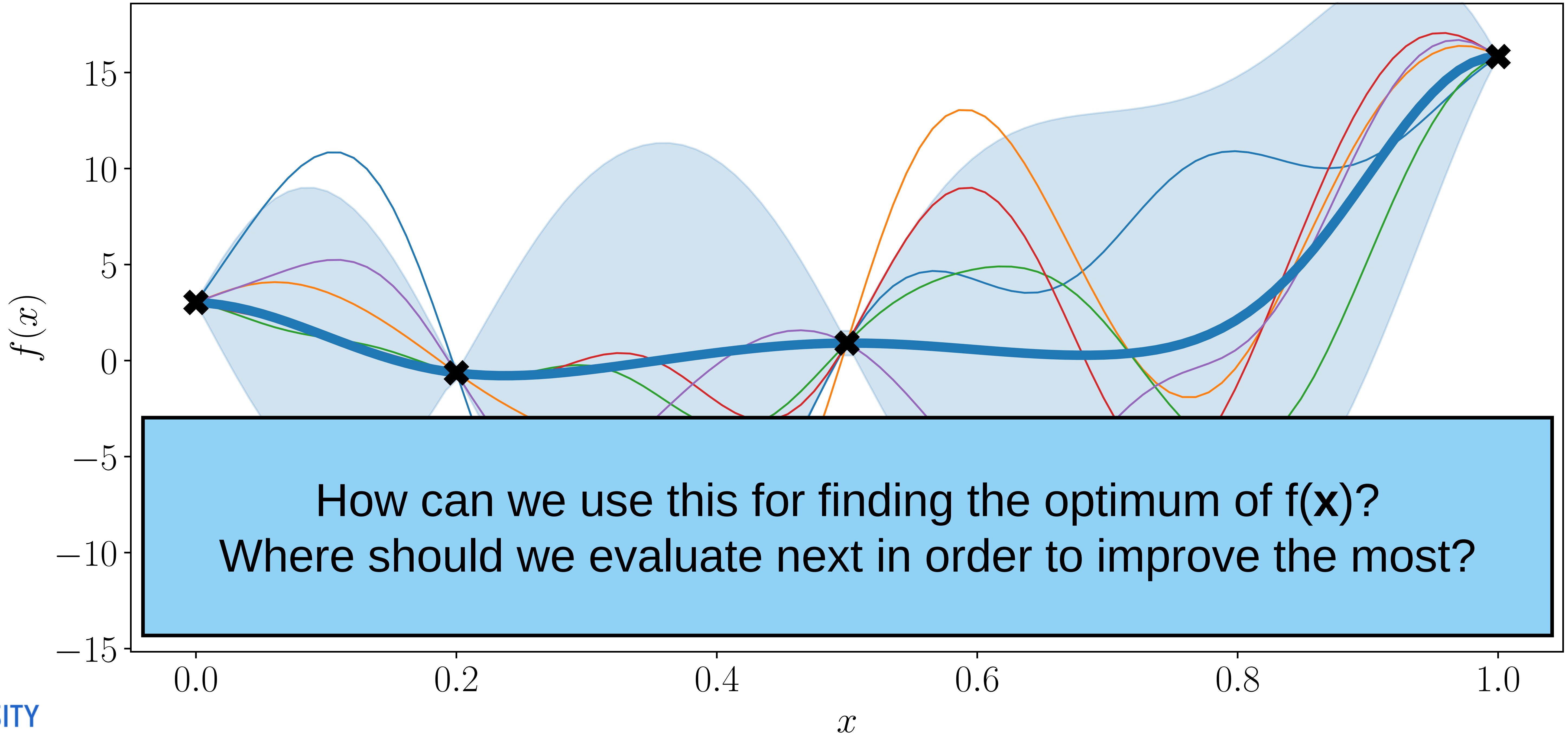


GHENT UNIVERSITY

GAUSSIAN PROCESS POSTERIOR

$\mathrm{f}(\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$

Gaussian Processes *know what they don't know*



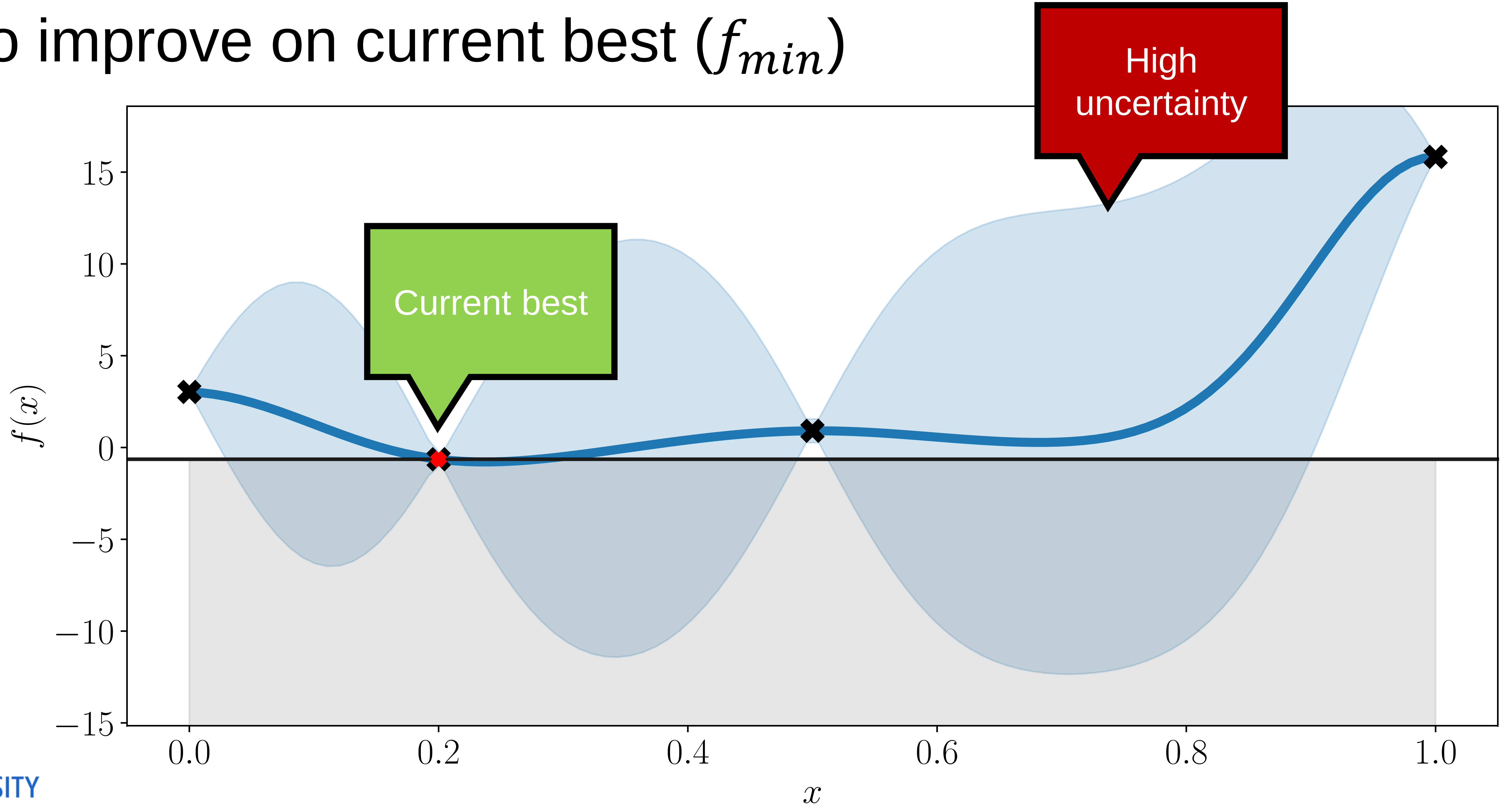
How can we use this for finding the optimum of f(**x**)?
Where should we evaluate next in order to improve the most?

GHENT UNIVERSITY

ACQUISITION FUNCTION

Where to evaluate next?

– to improve on current best (f_{min})



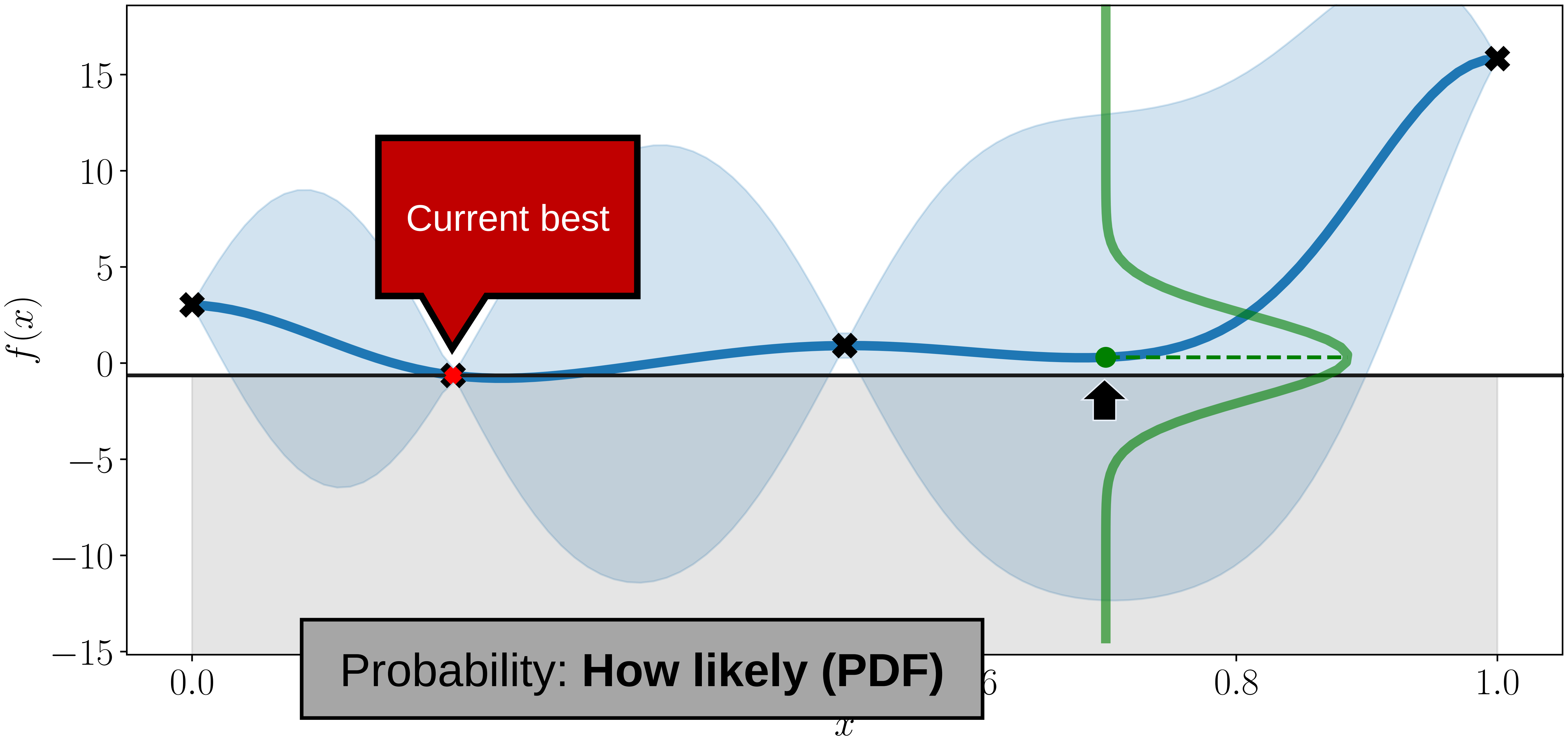
GHENT UNIVERSITY

ACQUISITION FUNCTION

- **Definition**: Acquisition function $\alpha(\boldsymbol{x})$
  - Measures how interesting a location $\boldsymbol{x}$ is
  - Higher the better (more ‘*interesting*’)
- **Balance**
  - Exploitation → Finding a more accurate neural network
    - Seek places with low prediction mean
  - Exploration → Improving the accuracy of the Gaussian process
    - Seek places with high uncertainty
- **Example**: Expected improvement

GHENT UNIVERSITY

ACQUISITION FUNCTION



$$P[I] = \int_{-\infty}^{f_{min}} \overbrace{\psi\left(y|\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\right)} dy$$

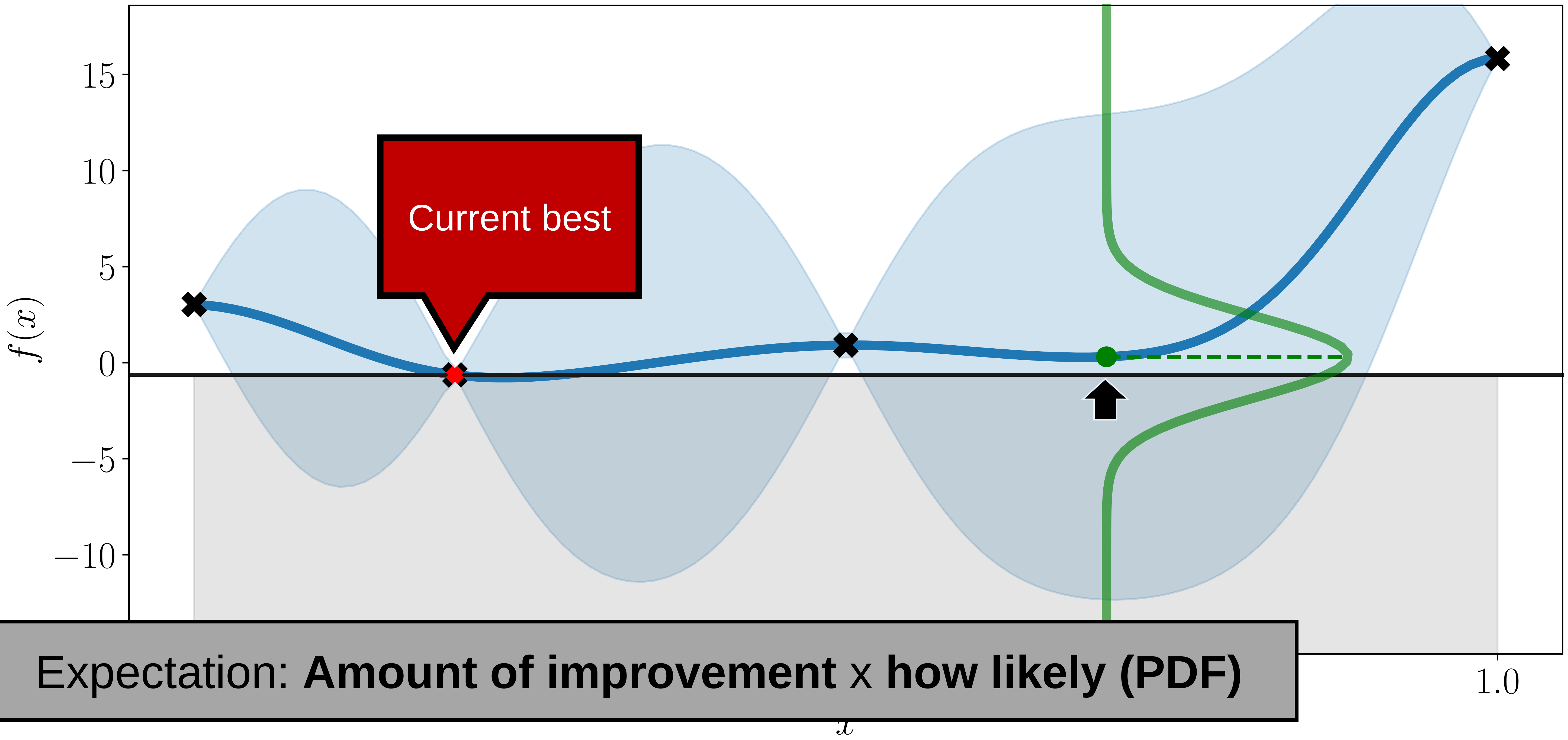
ACQUISITION FUNCTION

– **Example**: Probability of improvement

$$\alpha(x) = \phi\left(\frac{f_{min} - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)$$

– Already very useful, but …
– Does not specify the amount of improvement

ACQUISITION FUNCTION



Expectation: **Amount of improvement** x **how likely (PDF)**

$$E[I] = \int_{-\infty}^{f_{min}} \overbrace{(f_{min} - y)} \overbrace{\psi\left(y|\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\right)} dy$$

GHENT UNIVERSITY

ACQUISITION FUNCTION

– **Example**: Expected Improvement

$$\alpha(\boldsymbol{x}) = \left(f_{min} - \mu(\boldsymbol{x})\right)\Phi\left(\frac{f_{min} - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right) + \sigma(\boldsymbol{x})\psi\left(\frac{f_{min} - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)$$

Normal cumulative density function (CDF)

Normal probability density function (PDF)

Exploitation ⟺ Exploration

ACQUISITION FUNCTION

- **Example**: Lower confidence bound (LCB)

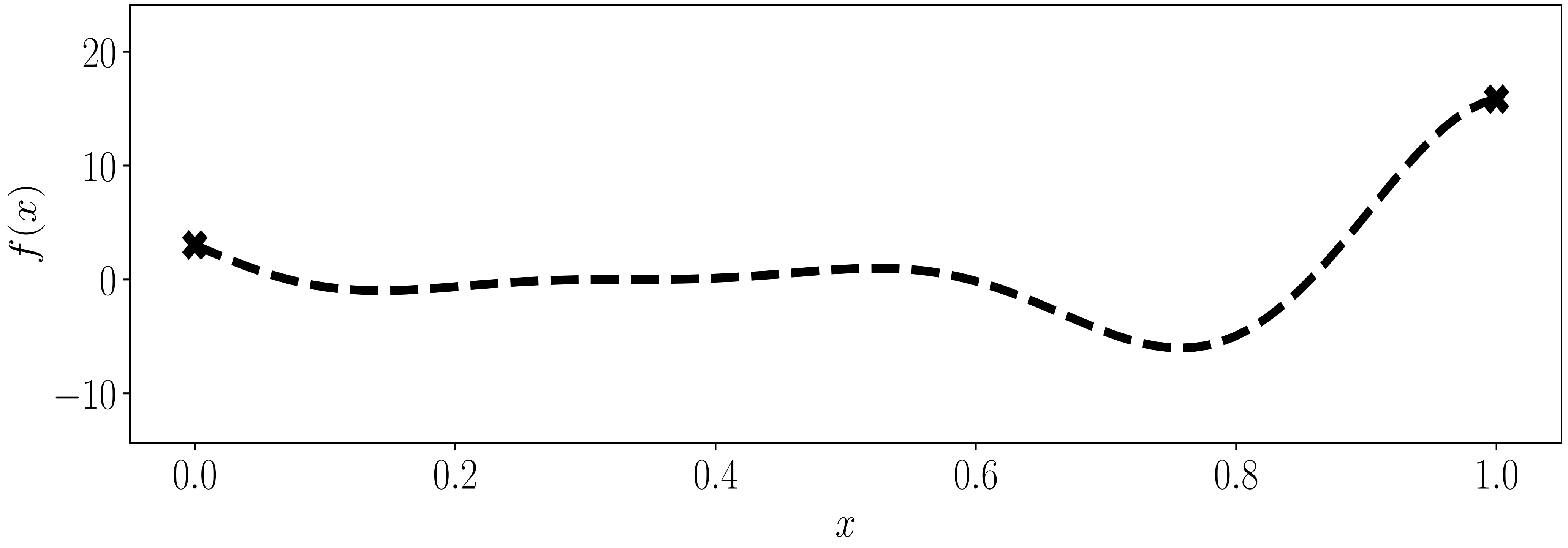
$$\alpha(\boldsymbol{x}) = \mu(\boldsymbol{x}) - \beta\sigma(\boldsymbol{x})$$

- β user-defined parameter
- No uncertainty => minimizes prediction
- If uncertainty is high enough => exploration

BAYESIAN OPTIMIZATION EXAMPLE

- **Problem:**
  - Discrete small dataset

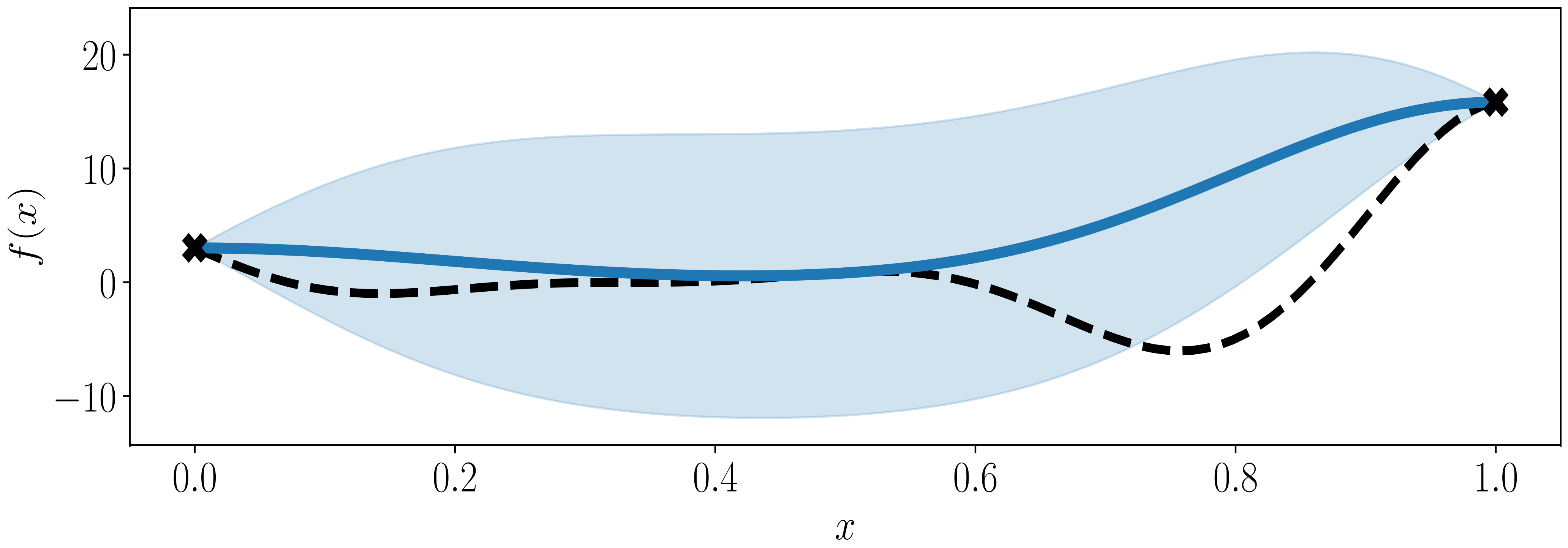
- **Goal:**
  - Minimize



GHENT UNIVERSITY

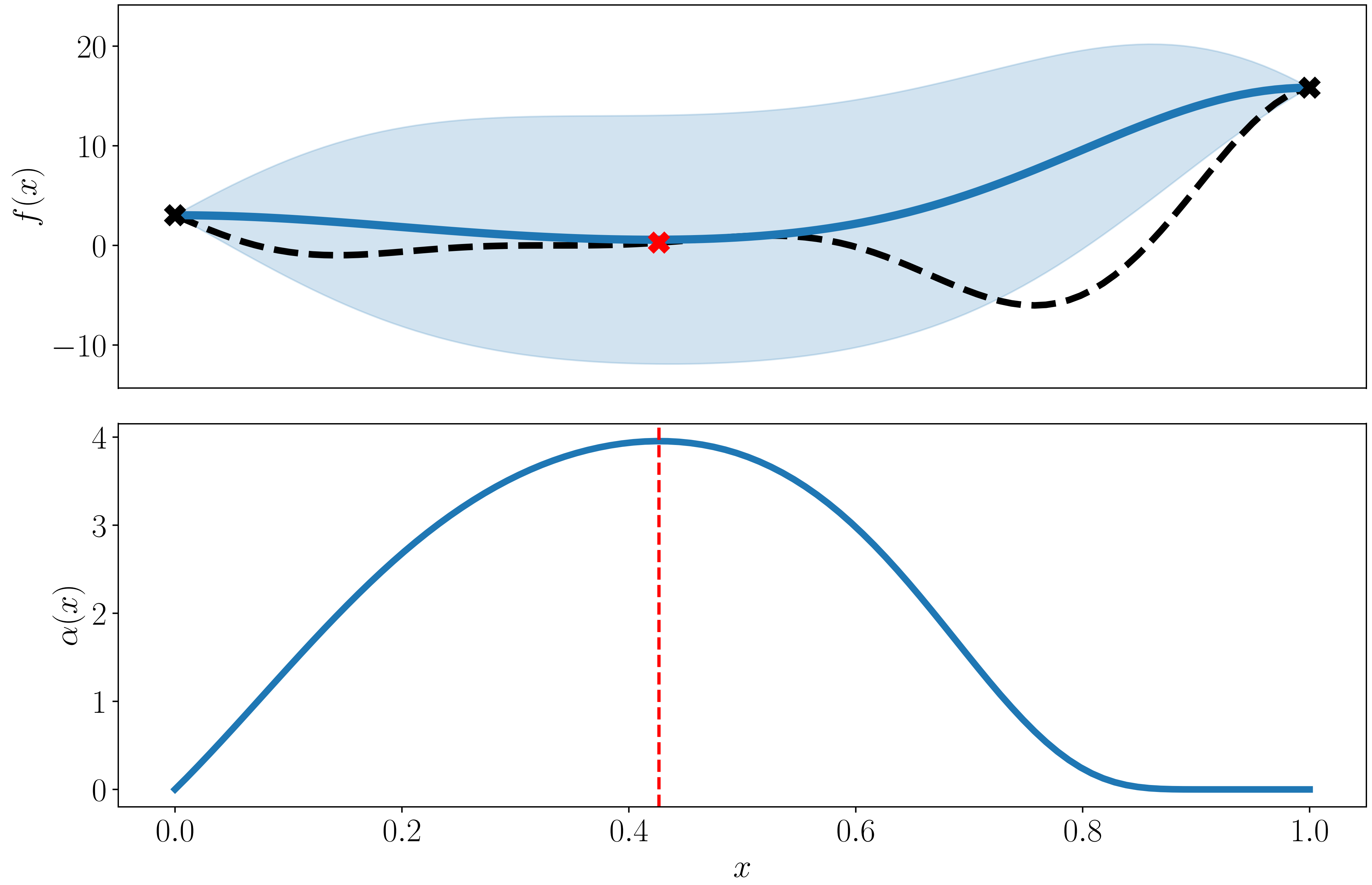
BAYESIAN OPTIMIZATION EXAMPLE

- **Problem:**
  - Discrete small dataset
- **Goal:**
  - Minimize
- **Approach:**
  - Build GP model



BAYESIAN OPTIMIZATION EXAMPLE

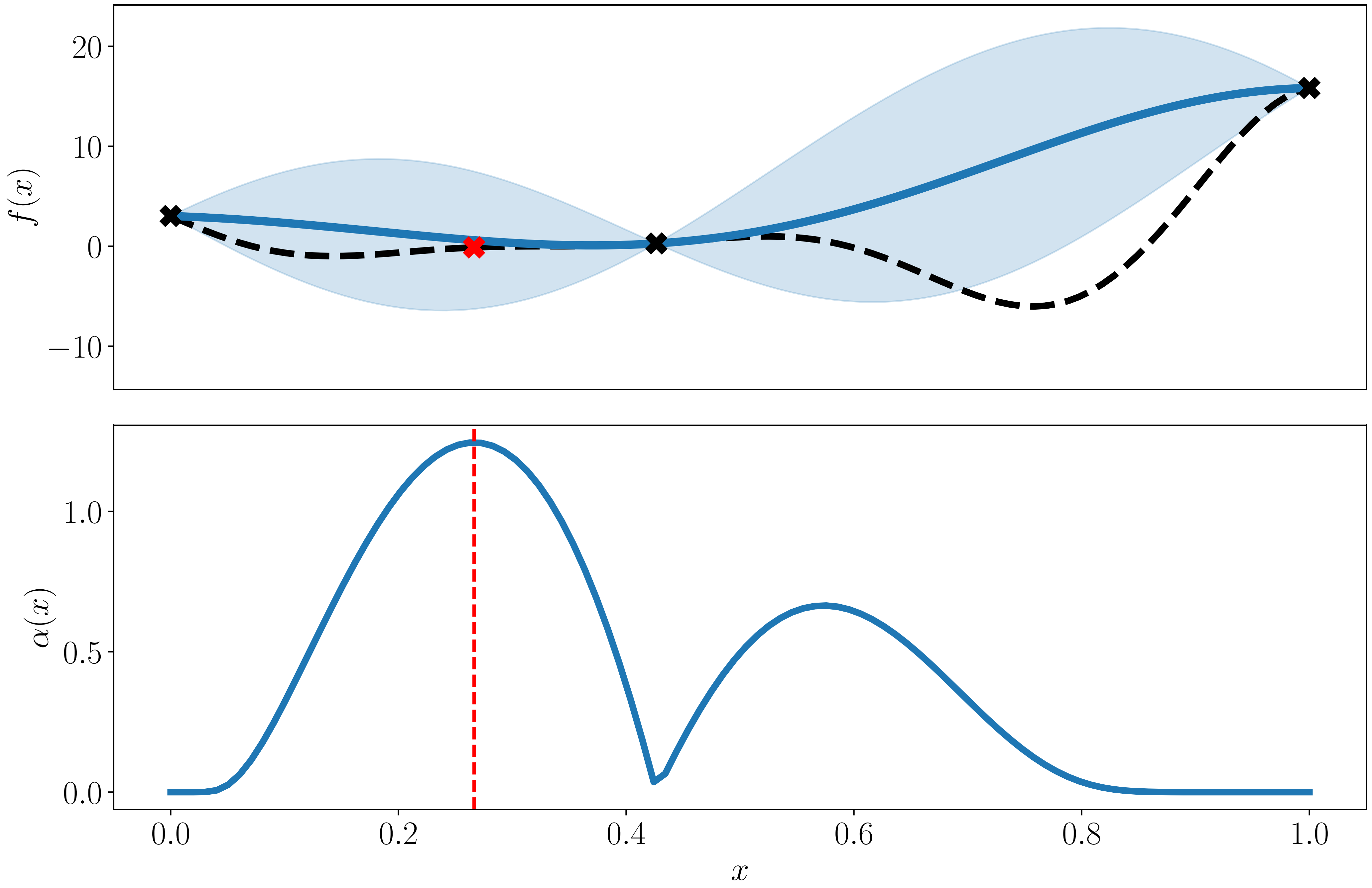
- **Problem:**
  - Discrete small dataset
- **Goal:**
  - Minimize
- **Approach:**
  - Build GP model
  - Calc. acquisition function
  - Add sample…



GHENT UNIVERSITY

BAYESIAN OPTIMIZATION EXAMPLE

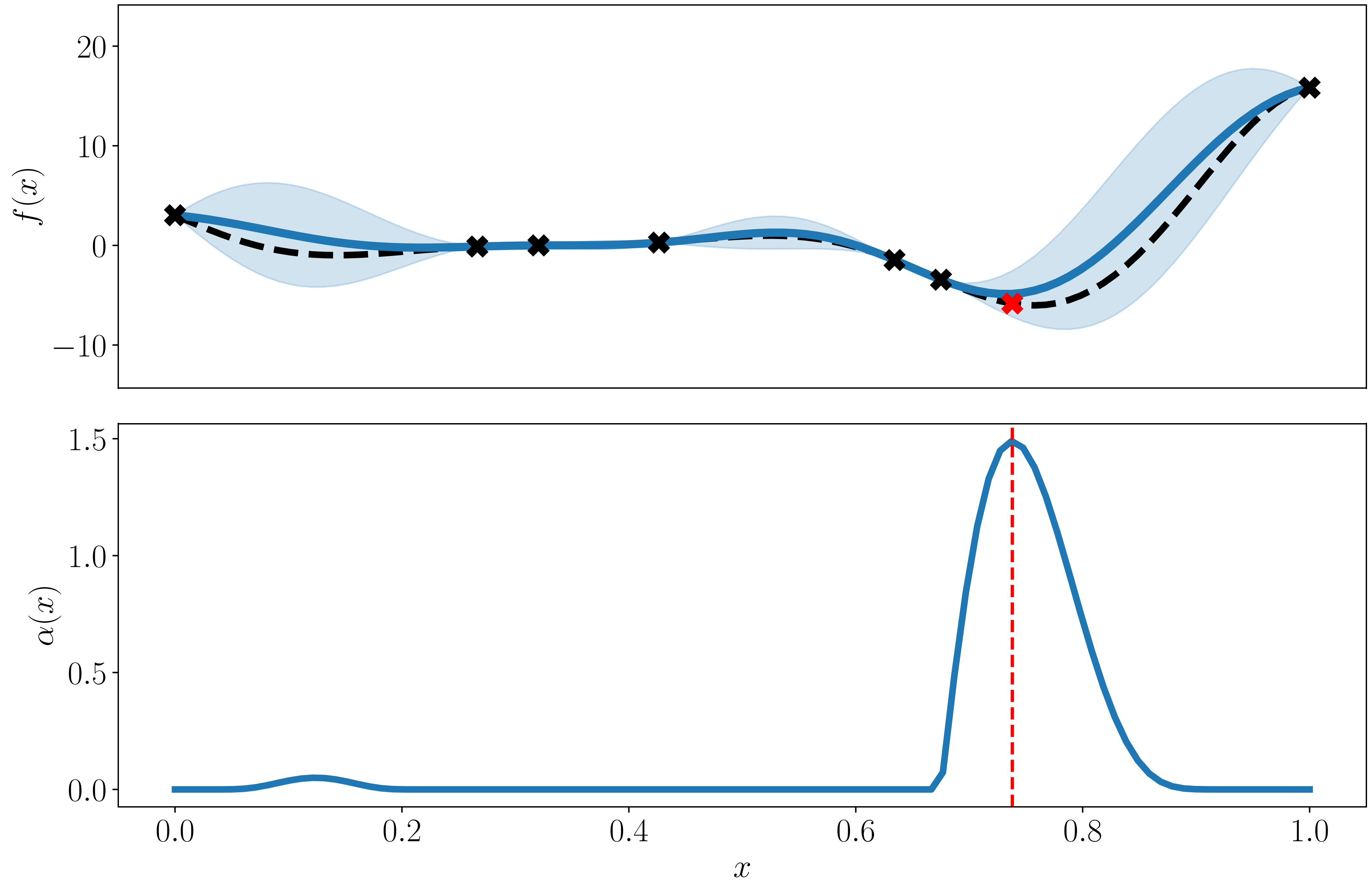
- **Problem:**
  - Discrete small dataset
- **Goal:**
  - Minimize
- **Approach:**
  - Build GP model
  - Calc. acquisition function
  - Add sample…



GHENT UNIVERSITY

BAYESIAN OPTIMIZATION EXAMPLE

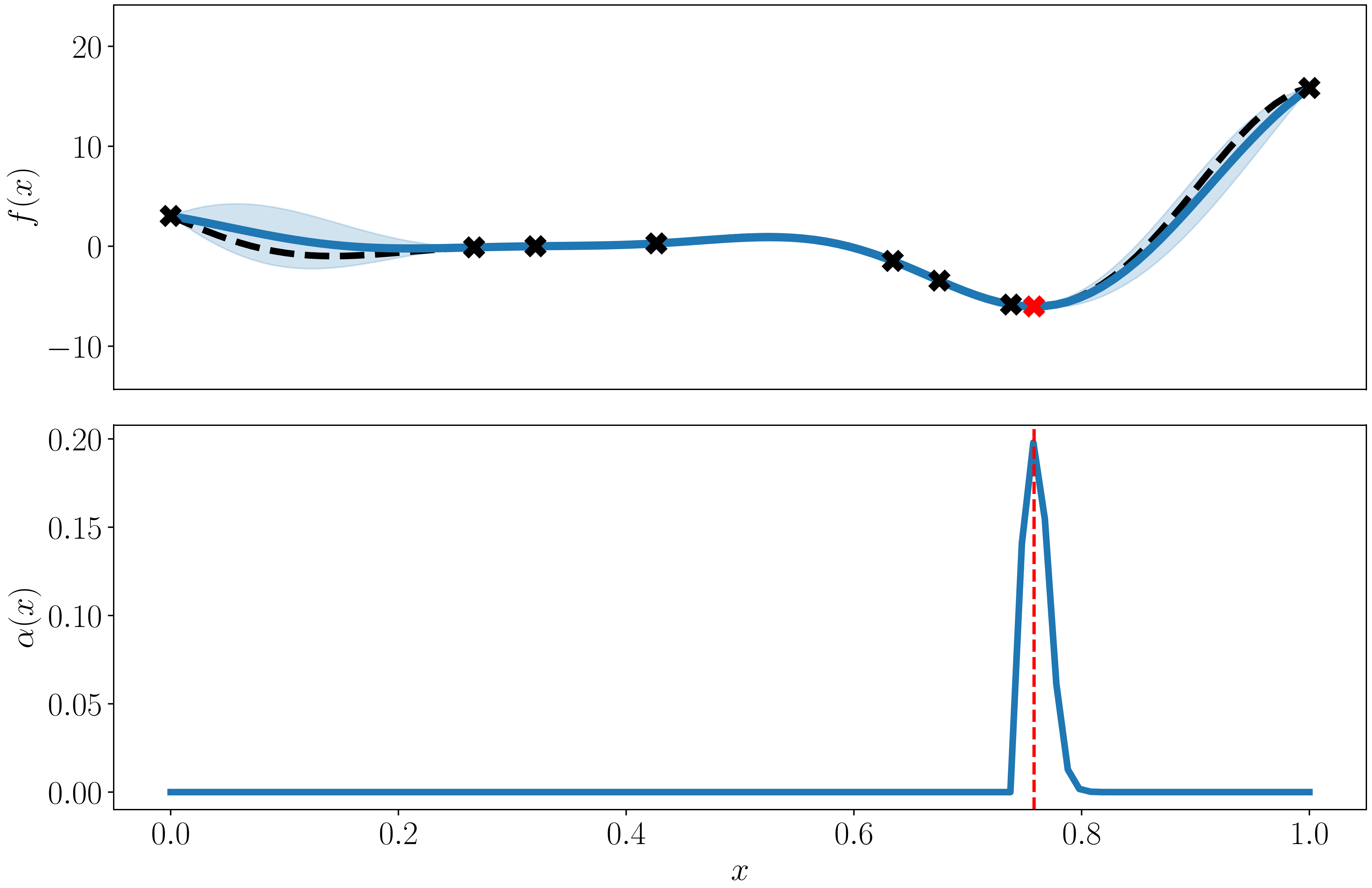
- **Problem:**
  - Discrete small dataset
- **Goal:**
  - Minimize
- **Approach:**
  - Build GP model
  - Calc. acquisition function
  - Continue…



GHENT UNIVERSITY

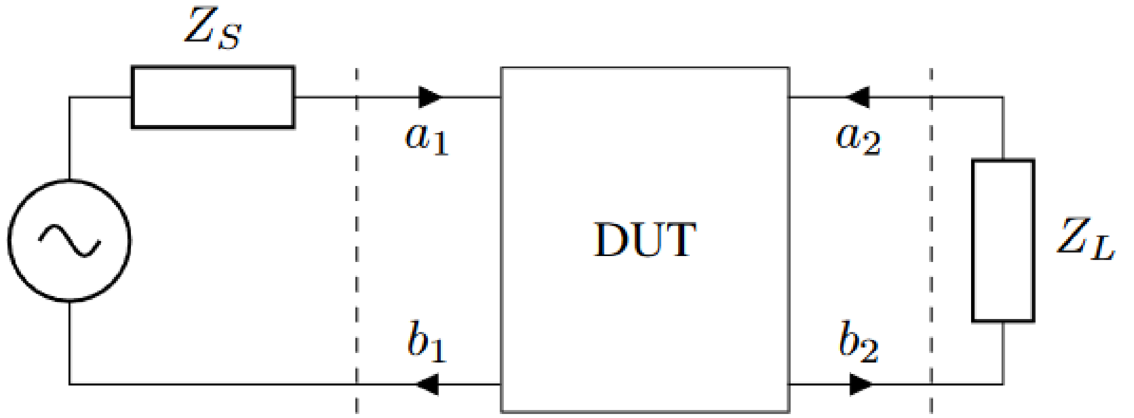
BAYESIAN OPTIMIZATION EXAMPLE

- **Problem:**
  - Discrete small dataset
- **Goal:**
  - Minimize
- **Approach:**
  - Build GP model
  - Calc. acquisition function
  - … until convergence

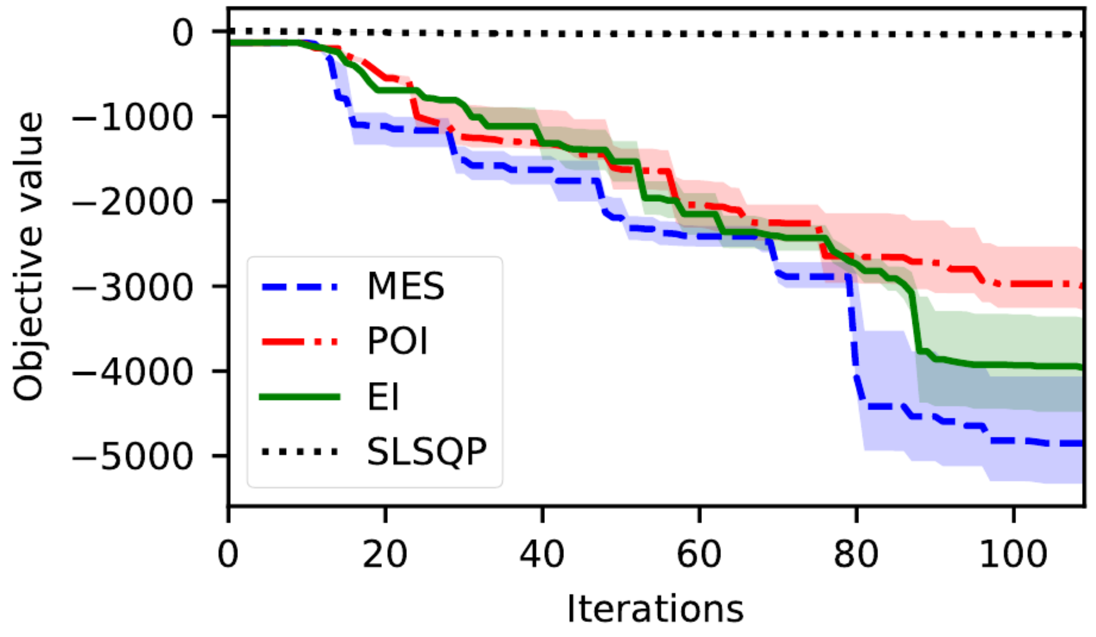


GHENT UNIVERSITY

BAYESIAN OPTIMIZATION



- **Example**: Power amplifier
- **Problem**: design of a power amplifier
  - Simulated in Keysight ADS
- **Goal**: optimize gain for 4 design variables
- **Results**: a better design in less simulations
  - vs traditional methods (no feasible design found)



GHENT UNIVERSITY

BAYESIAN OPTIMIZATION IN A NUTSHELL

- Strategy to transform

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathcal{X}}{\operatorname{argmin}}\, f(\boldsymbol{x})$$

unsolvable

- Into a series of problems

$$\boldsymbol{x}_{i+1} = \underset{\boldsymbol{x} \in \mathcal{X}}{\operatorname{argmax}}\, \alpha(\boldsymbol{x})$$

solvable

GHENT UNIVERSITY

CONCLUSION

- Bayesian optimization
  - A **probabilistic**, **data-efficient** optimization method
  - Used when the objective is **time-consuming**
- Applications
  - **Hyperparameter tuning** of neural networks
  - **Design optimization** in engineering
- Software
  - Trieste / GPFlowOpt (python)
    - https://github.com/secondmind-labs/trieste
    - https://github.com/GPflow/GPflowOpt
  - SUMO toolbox (Matlab)
    - http://sumo.intec.ugent.be/SUMO_toolbox

GHENT UNIVERSITY

FACULTY OF ENGINEERING
AND ARCHITECTURE

Ivo Couckuyt

Department of Information Technology
SUMO research cluster - IDlab

E ivo.couckuyt@ugent.be
T +32 9 331 49 91

sumo.intec.ugent.be
www.ugent.be

Universiteit Gent
@ugent
@ugent
Ghent University

GHENT
UNIVERSITY