Global Sensitivity Analysis of Predictor Models in Software Engineering

Stefan Wagner
Software & Systems Engineering
Technische Universität München
Germany
wagnerst@in.tum.de

May 20, 2007

Introduction

Predictor Models in Software Engineering

- Predictor models describe situations in software engineering
- Various types, aims, ...
- Examples
  - COCOMO (costs)
  - Musa-Okumoto (reliability)
  - Fault-proneness

Problem

- The models themselves can be complex
- Their use needs a lot of effort
- How can I analyse those models themeselves?
- How can I simplify them?
- How do I improve their predictive power?

Example: Fischer-Wagner Model

- Reliability model used at Siemens COM
- Two parameters estimated by failure data
- Failure probability of a fault: $p_a = p_1 \cdot d^{(a-1)}$
- Geometrical distribution : $F_a(t) = 1 - (1 - p_a)^t$
- Cummulated failures up to t:
  $\mu(t) = \sum_{a=1}^{\infty} 1 - (1 - p_1 \cdot d^{(a-1)})$
- Input parameters
  - p_1: Highest failure probability of a fault
  - d: Complexity of the system $]0; 1[$
  - t: Execution time (incidents)
  - *inf*: approximation of ∞
- Output: $\mu(t)$

More details: S. Wagner, H. Fischer. Ada-Europe, 2006

Questions About the Model

- How do the input parameters influence the output?
- Which parameter(s) do I have to estimate best to get a good prediction?
- Are there insignificant parameters that I can remove?

Introduction

Global Sensitivity Analysis

Approach for SE

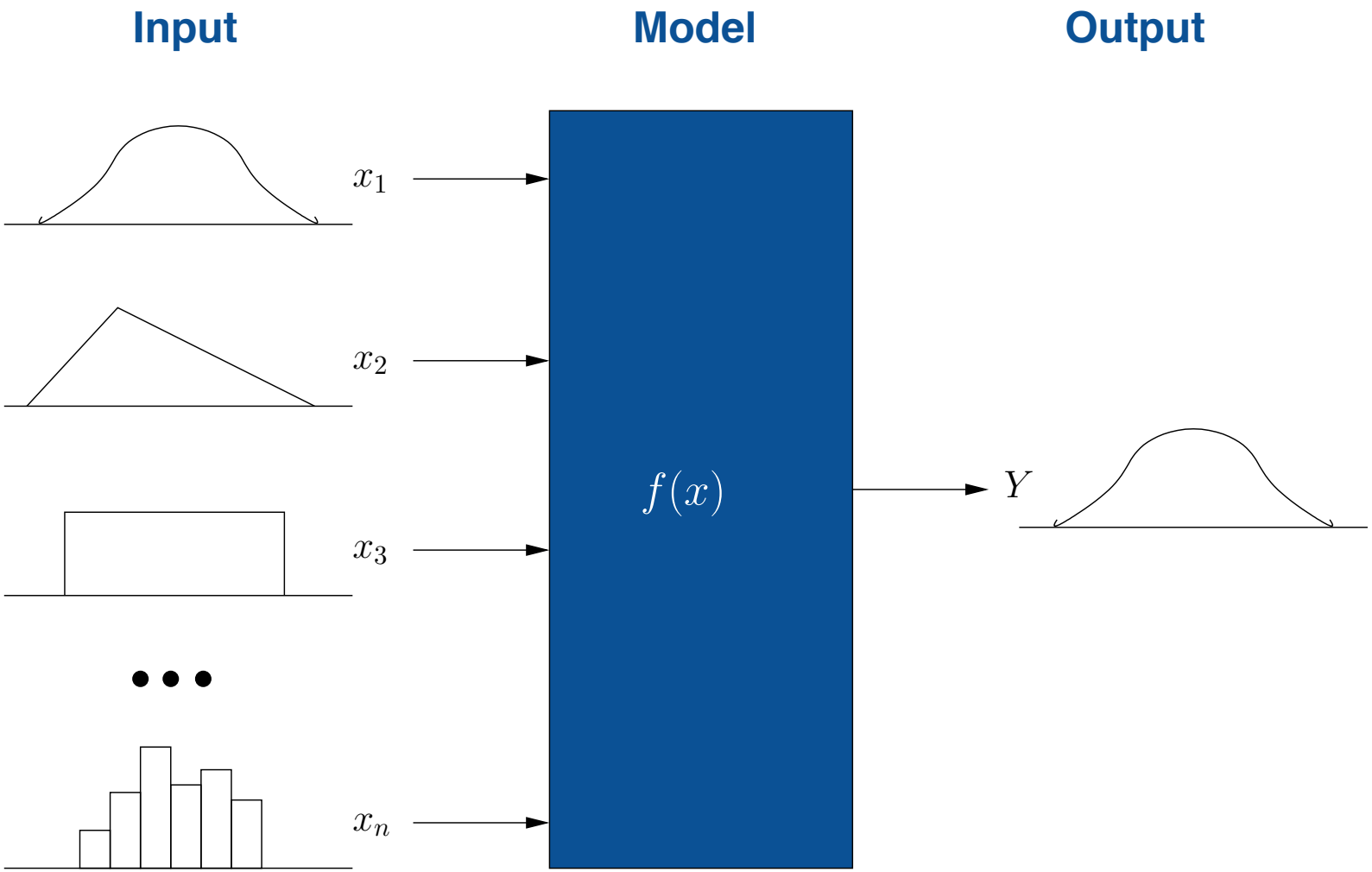
Summary

Global Sensitivity Analysis

Definition

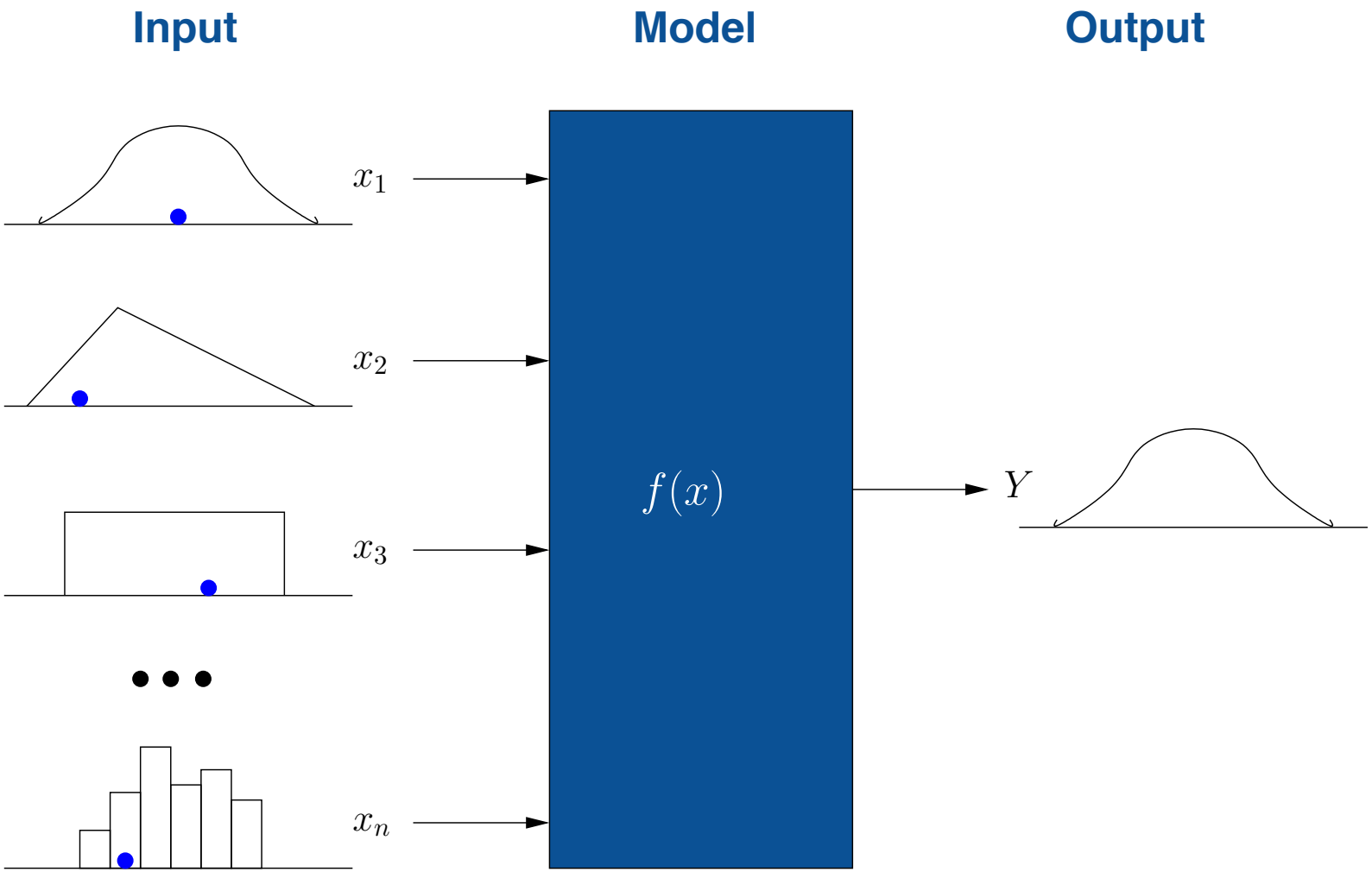
- Saltelli (2000): "Sensitivity analysis studies the relationships between information flowing in and out of the model."
- How do the input parameters influence the output?
- How can this influence by quantified?
- Global properties
  - Inclusion of influence of shape and scale
  - Multidimensional averaging

Method



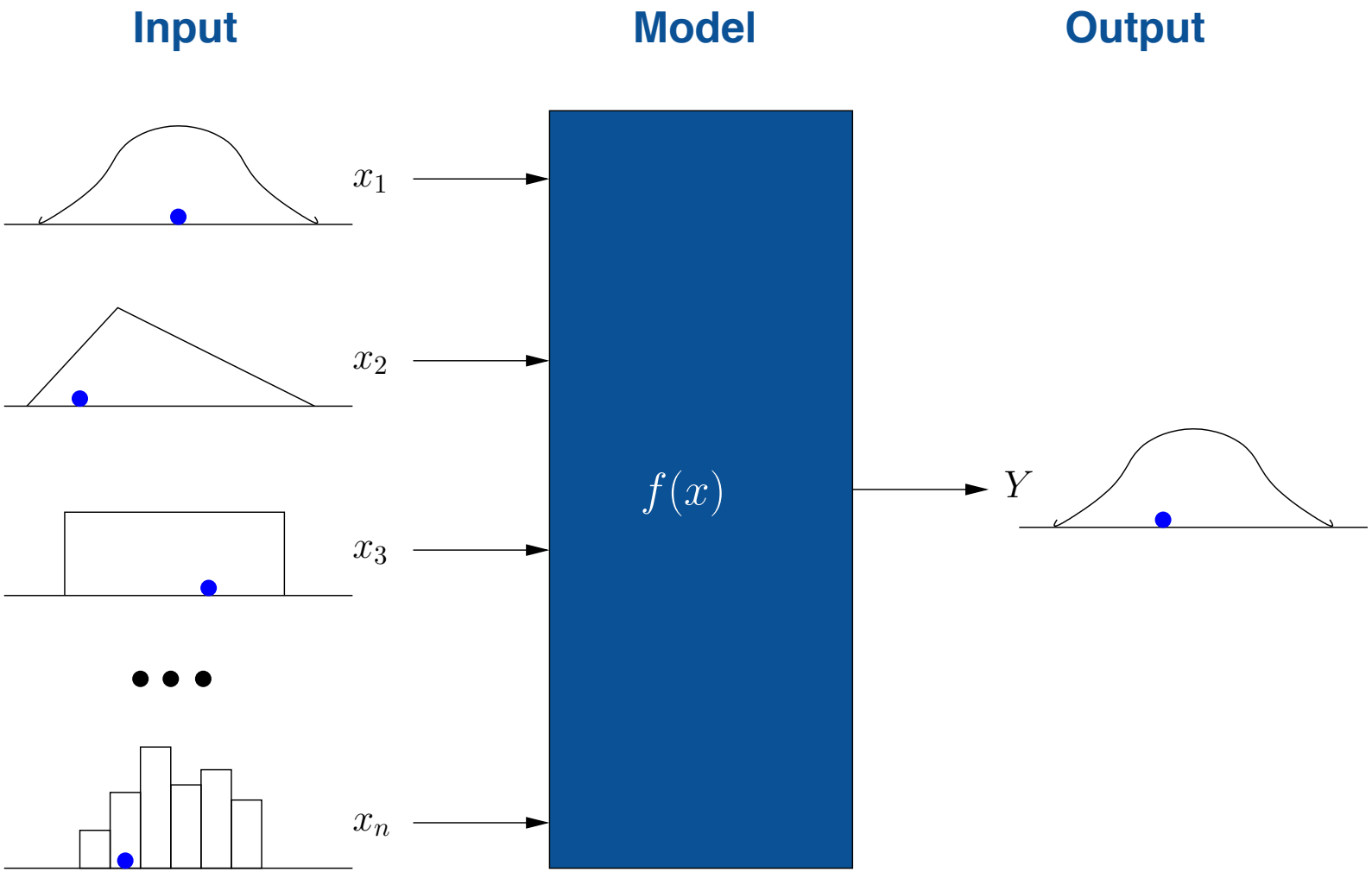
(Based on slide of S. Tarantola)

Method



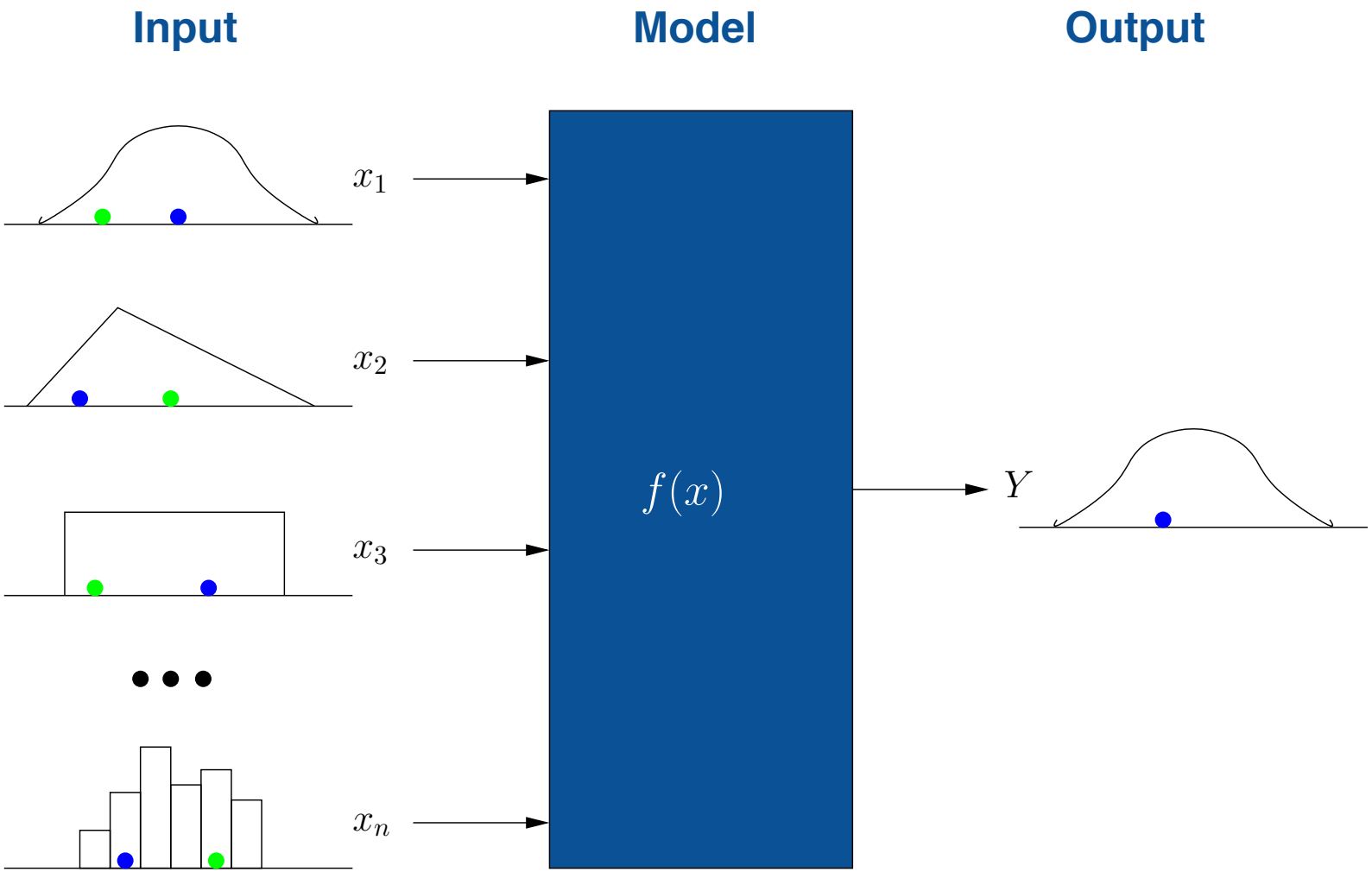
(Based on slide of S. Tarantola)

Method



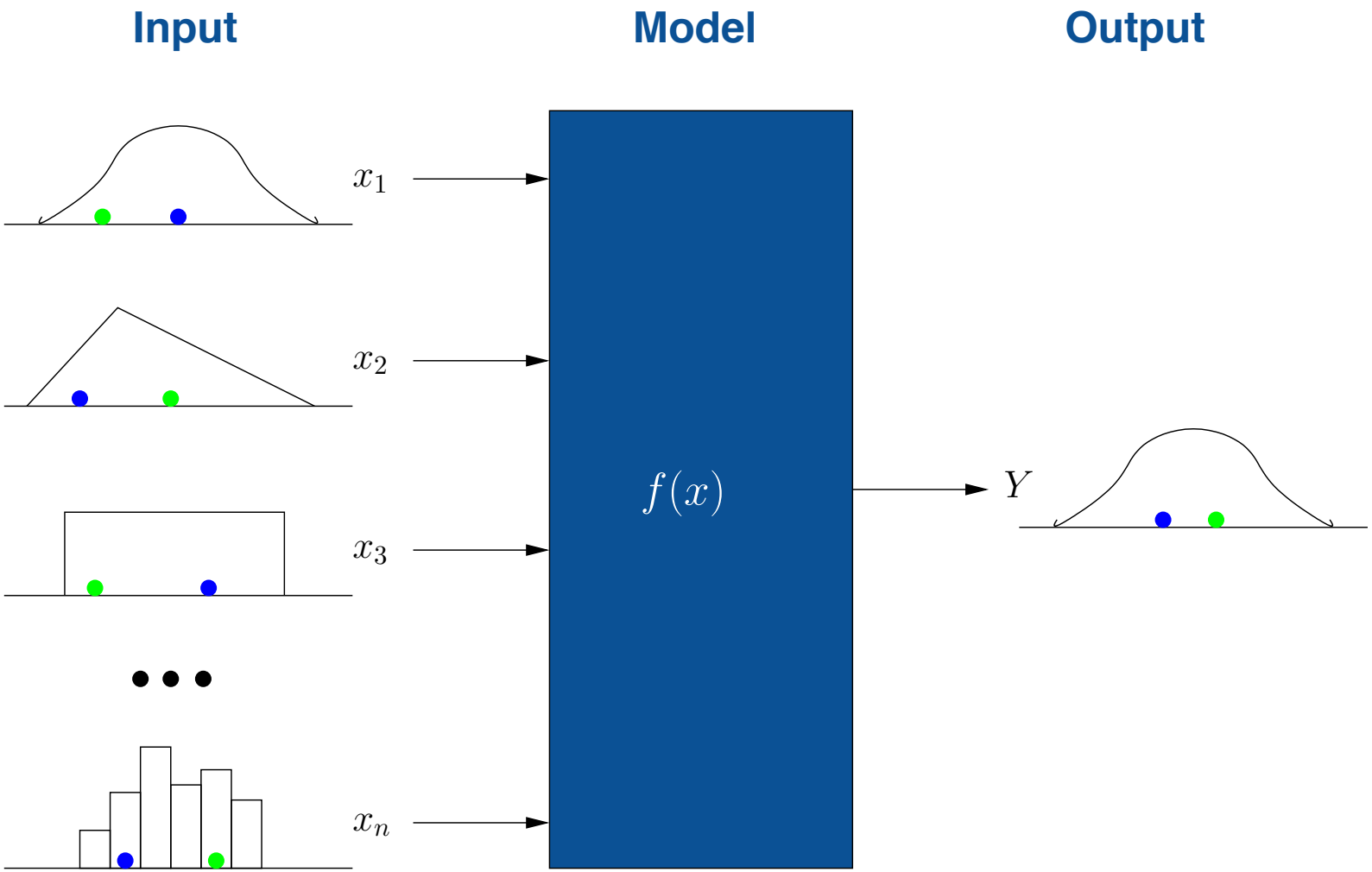
(Based on slide of S. Tarantola)

Method



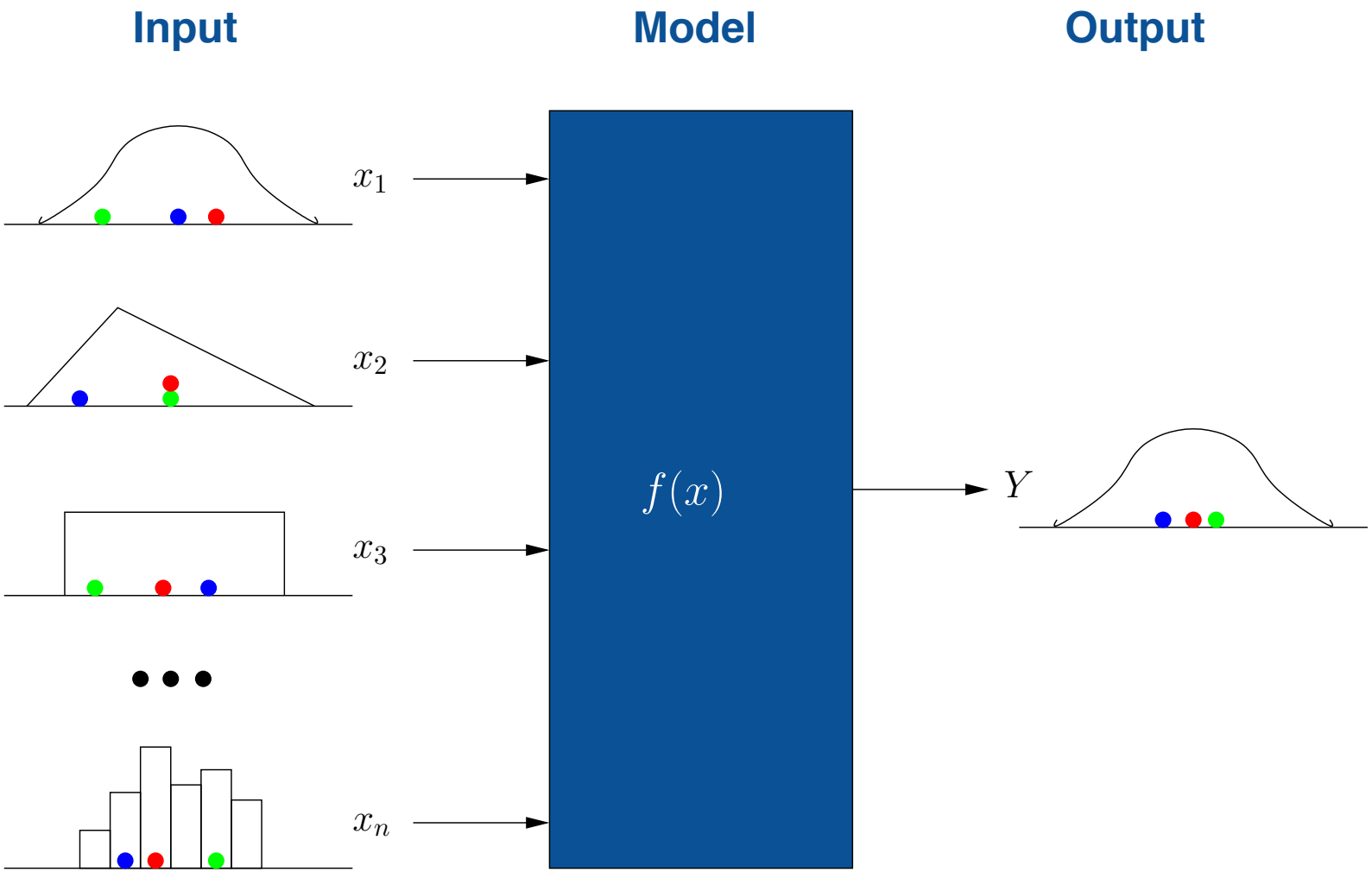
(Based on slide of S. Tarantola)

Method



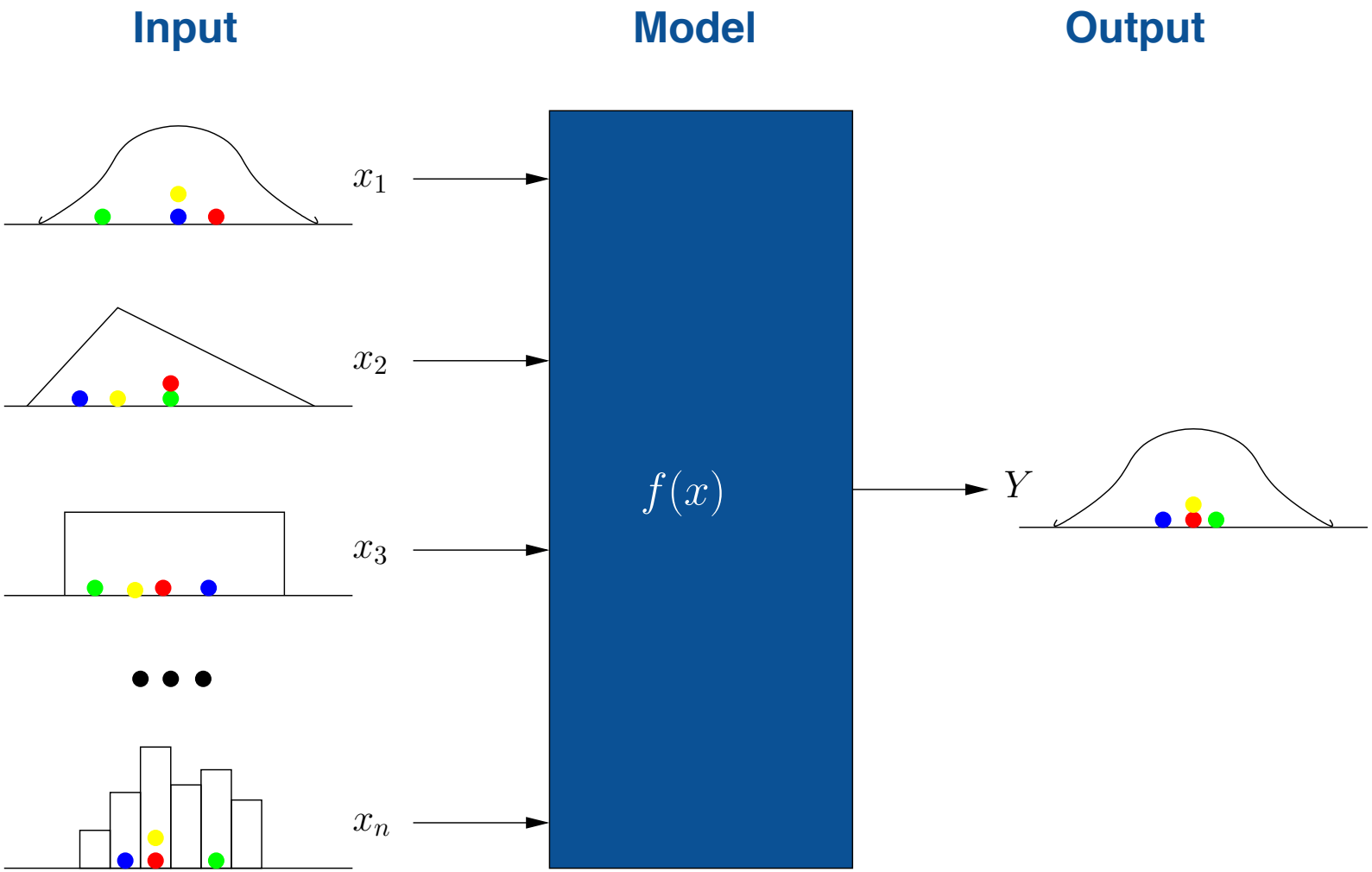
(Based on slide of S. Tarantola)

Method



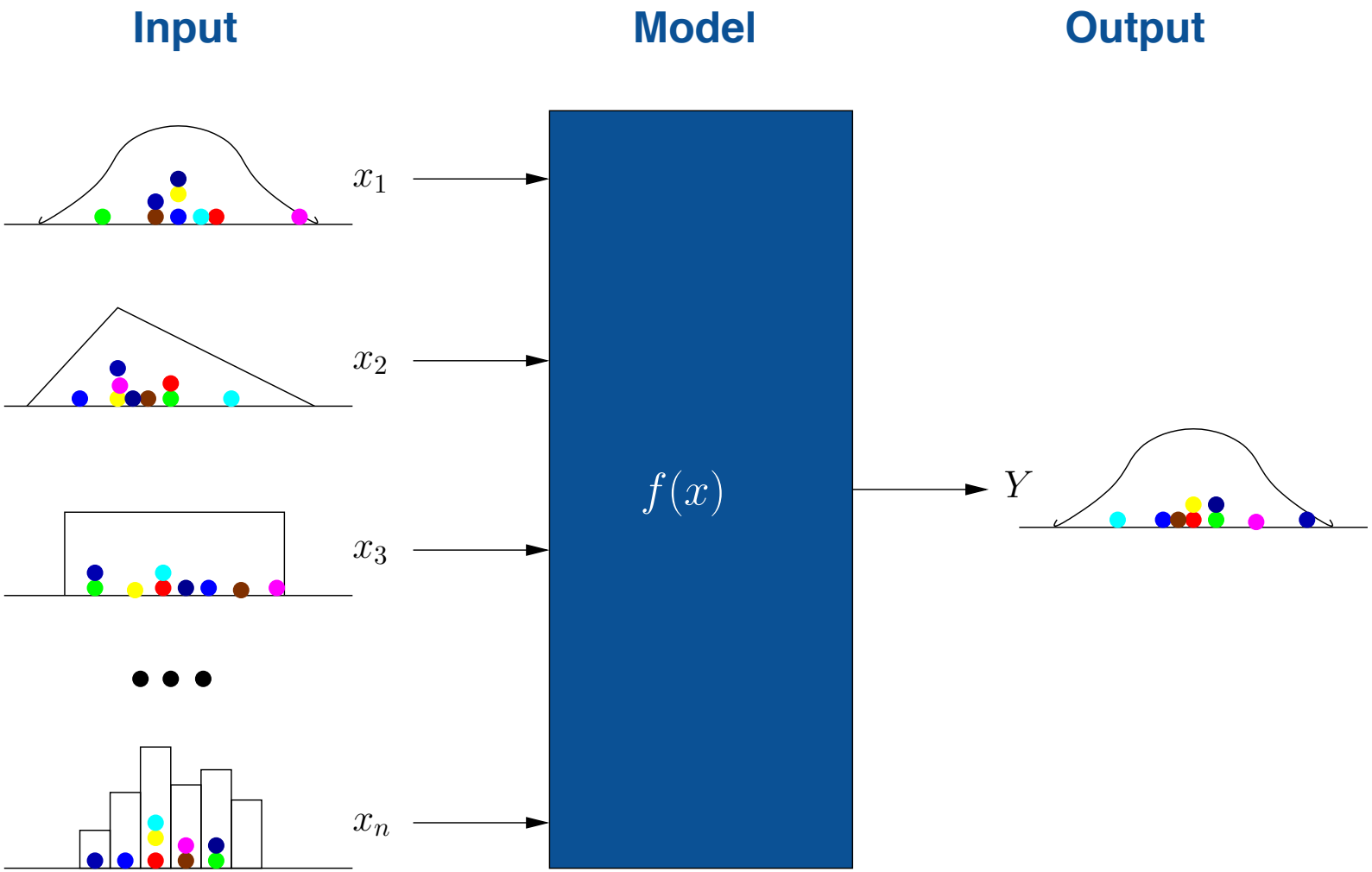
(Based on slide of S. Tarantola)

Method



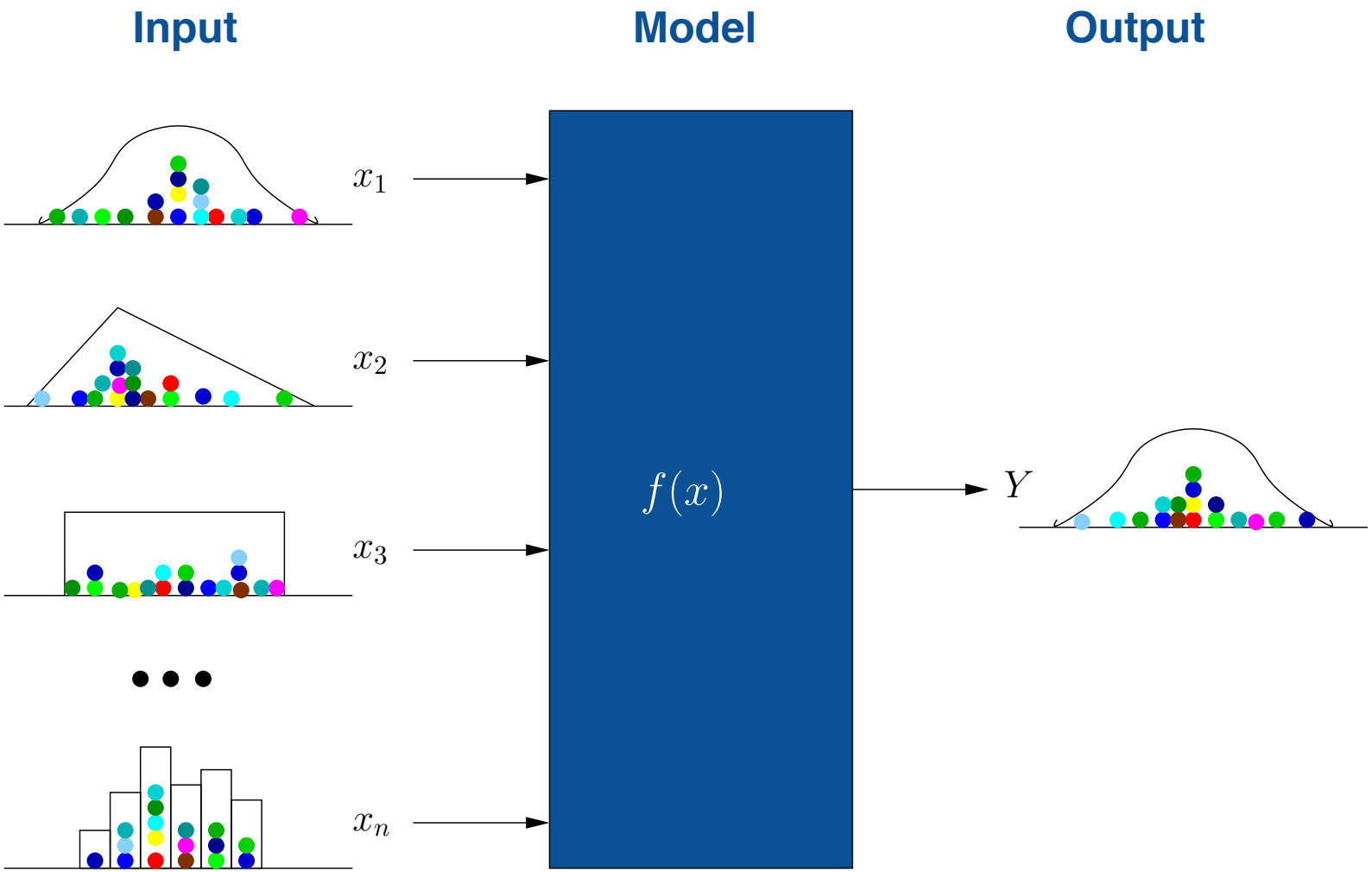
(Based on slide of S. Tarantola)

Method



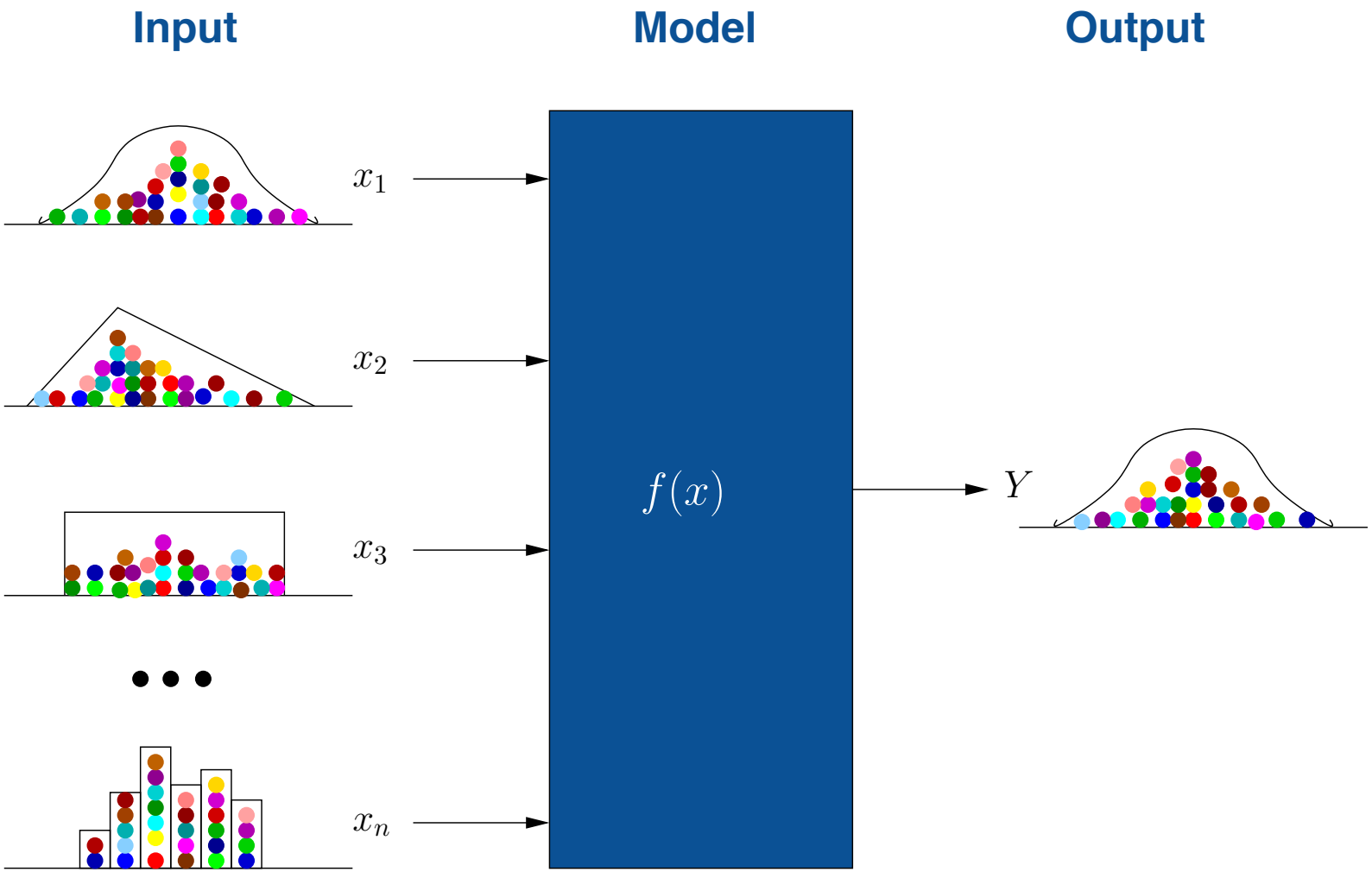
(Based on slide of S. Tarantola)

Method



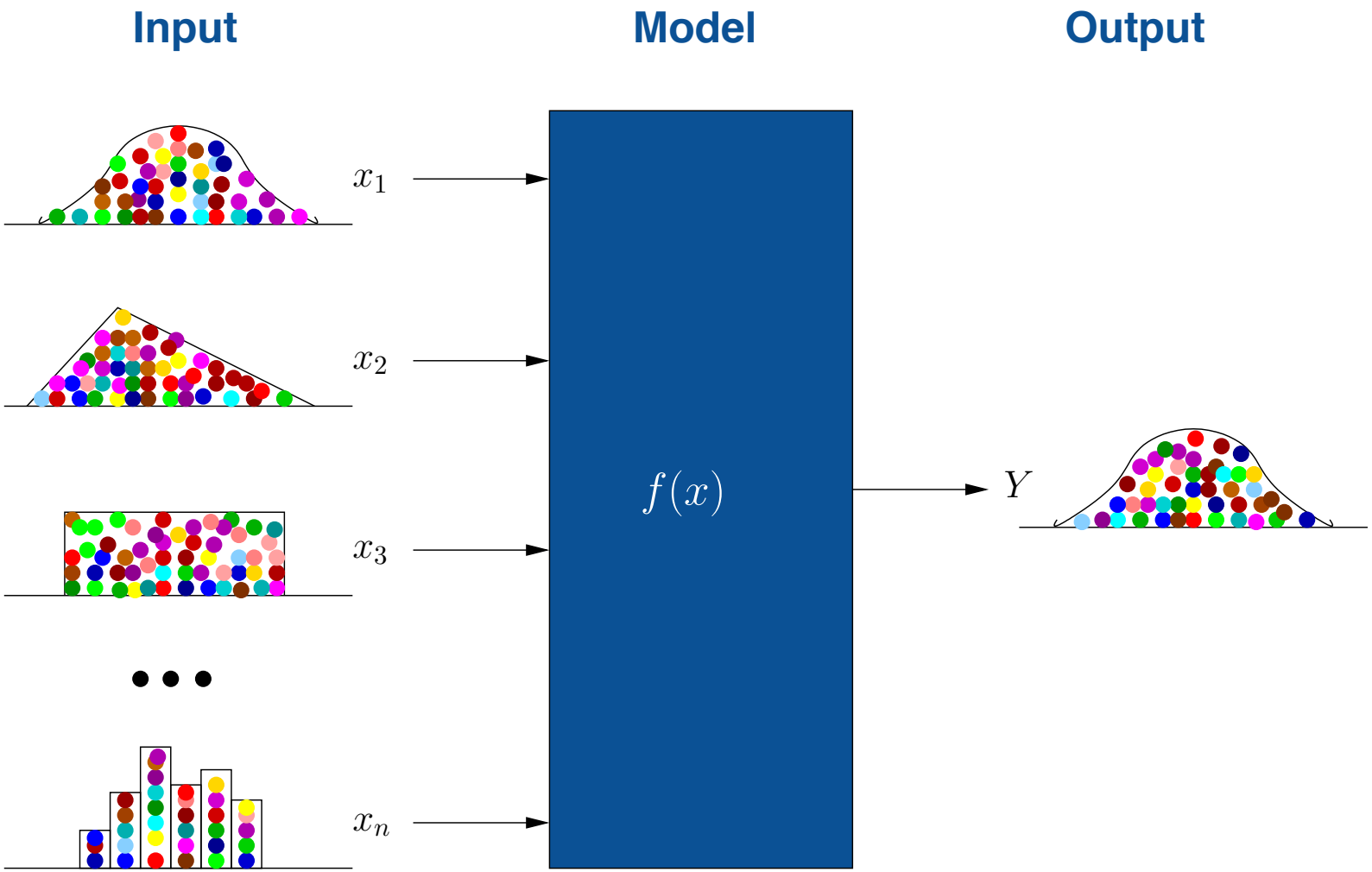
(Based on slide of S. Tarantola)

Method



(Based on slide of S. Tarantola)

Method



(Based on slide of S. Tarantola)

Variance-based Methods

- Global analysis usually variance-based
- "model-free"
- Analysis on the basis of **sensitivity indices**
  - *Main* or *first-order effect*
  - *Higher-order effects*
  - *Total effect*
- Decomposition in *main effects* and *interactions*

$$y = f(x) = f_0 + \sum_{i=1}^{k} f_i(x_i) + \sum_{i} \sum_{j>i} f_{ij}(x_i, x_j) + \ldots + f_{1,2,\ldots,k}(x_1, x_2, \ldots, x_k)$$

Variance-based Methods

- Global analysis usually variance-based
- "model-free"
- Analysis on the basis of **sensitivity indices**
  - *Main* or *first-order effect*
  - *Higher-order effects*
  - *Total effect*
- Decomposition in *main effects* and *interactions*

For example $k = 3$:

$$f(x) = f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2)$$

$$+ f_{13}(x_1, x_3) + f_{23}(x_2, x_3) + f_{123}(x_1, x_2, x_3)$$

Main Effect

Decomposition of the variance V of $f(x)$

$$\mathrm{Var}(y) = V = \sum_{i=1}^{k} V_i + \sum_i \sum_j V_{ij} \sum_i \sum_j \sum_k V_{ijk} + \ldots + V_{1,2,\ldots k}$$

First-order sensitivity coefficient

$$S_i = \frac{V_i}{V}$$

(Factors Priorisation)

Higher-Order Effects

Higher-order indices analog

$$S_{i_1 \ldots i_o} = V_{i_1 \ldots i_o} / V$$

With $k = 4$:

2. order: $S_{12}, S_{13}, S_{14}, S_{23}, S_{24}, S_{34}$
3. order: $S_{123}, S_{124}, S_{134}, S_{234}$
4. order: S_{1234}

Total Effect

- Sum of the first- and higher-order effects of x_i
- Removal of the non-relevant parts

$$S_1 + S_2 + S_3 + S_4 + S_{12} + S_{13} + S_{14} + S_{23} + S_{24} + S_{34} + S_{123} + S_{124} + S_{134} + S_{234} + S_{1234} = 1$$

(Factors Fixing)

Total Effect

- Sum of the first- and higher-order effects of x_i
- Removal of the non-relevant parts

$$S_1 + \cancel{S_2} + S_3 + S_4 + S_{12} + S_{13} + S_{14} + S_{23} + S_{24} + S_{34} + S_{123} + S_{124} + S_{134} + S_{234} + S_{1234} = 1$$

(Factors Fixing)

Total Effect

- Sum of the first- and higher-order effects of x_i
- Removal of the non-relevant parts

$$S_1 + \cancel{S_2} + \cancel{S_3} + S_4 + S_{12} + S_{13} + S_{14} + S_{23} + S_{24} + S_{34} +$$
$$S_{123} + S_{124} + S_{134} + S_{234} + S_{1234} = 1$$

(Factors Fixing)

Total Effect

- Sum of the first- and higher-order effects of x_i
- Removal of the non-relevant parts

$$S_1 + \cancel{S_2} + \cancel{S_3} + \cancel{S_4} + S_{12} + S_{13} + S_{14} + S_{23} + S_{24} + S_{34} +$$
$$S_{123} + S_{124} + S_{134} + S_{234} + S_{1234} = 1$$

(Factors Fixing)

Total Effect

- Sum of the first- and higher-order effects of x_i
- Removal of the non-relevant parts

$$S_1 + \cancel{S_2} + \cancel{S_3} + \cancel{S_4} + S_{12} + S_{13} + S_{14} + \cancel{S_{23}} + \cancel{S_{24}} + \cancel{S_{34}} + S_{123} + S_{124} + S_{134} + S_{234} + S_{1234} = 1$$

(Factors Fixing)

Total Effect

- Sum of the first- and higher-order effects of x_i
- Removal of the non-relevant parts

$$S_1 + \cancel{S_2} + \cancel{S_3} + \cancel{S_4} + S_{12} + S_{13} + S_{14} + \cancel{S_{23}} + \cancel{S_{24}} + \cancel{S_{34}} + S_{123} + S_{124} + S_{134} + \cancel{S_{234}} + S_{1234} = 1$$

(Factors Fixing)

Total Effect

- Sum of the first- and higher-order effects of x_i
- Removal of the non-relevant parts

$$S_1 + \cancel{S_2} + \cancel{S_3} + \cancel{S_4} + S_{12} + S_{13} + S_{14} + \cancel{S_{23}} + \cancel{S_{24}} + \cancel{S_{34}} + S_{123} + S_{124} + S_{134} + \cancel{S_{234}} + S_{1234} = 1$$

$$S_{T1} = S_1 + S_{12} + S_{13} + S_{14} + S_{123} + S_{124} + S_{134} + S_{1234}$$

(Factors Fixing)

Total Effect

- Sum of the first- and higher-order effects of x_i
- Removal of the non-relevant parts

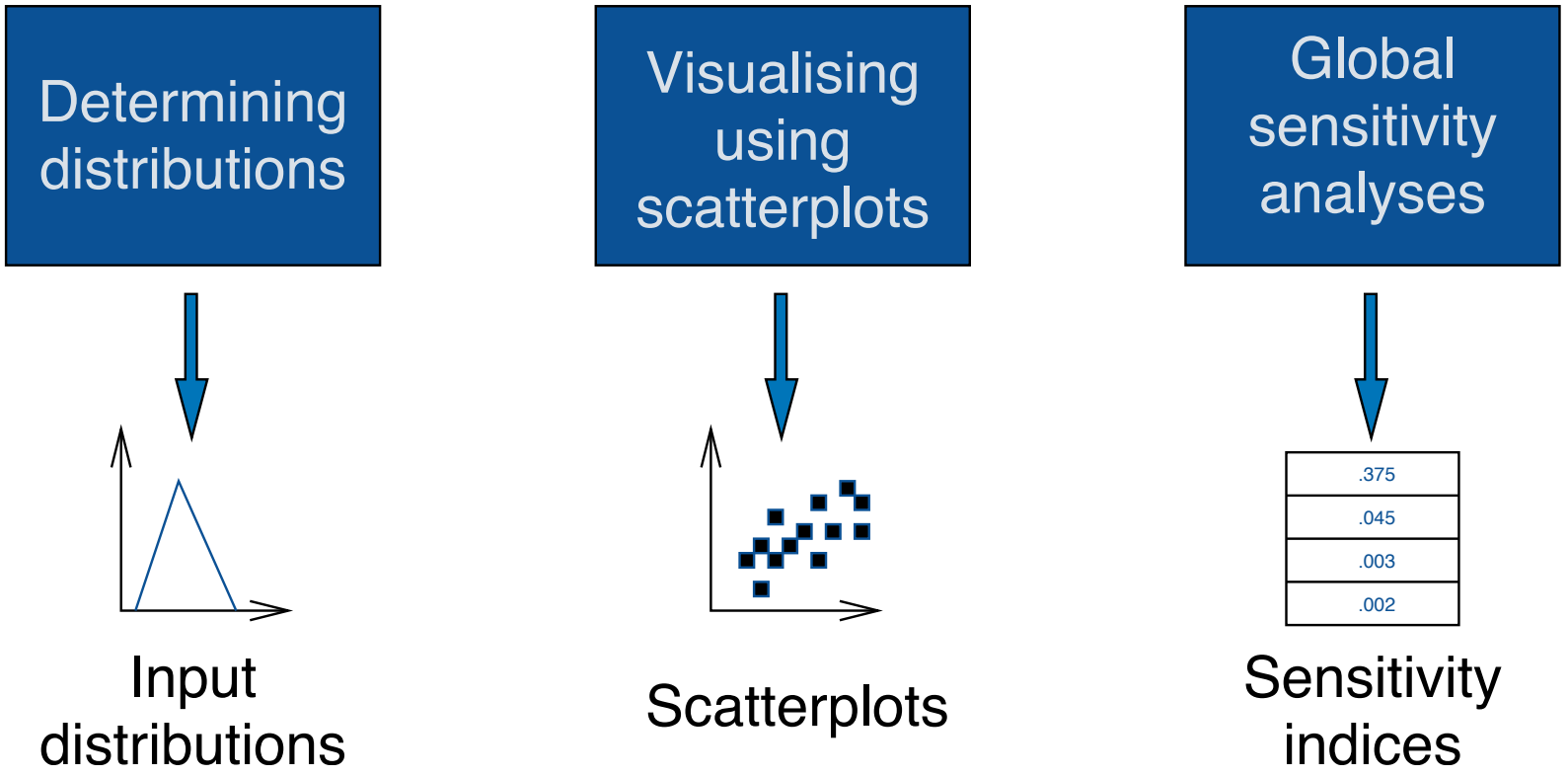
$$S_1 + \cancel{S_2} + \cancel{S_3} + \cancel{S_4} + S_{12} + S_{13} + S_{14} + \cancel{S_{23}} + \cancel{S_{24}} + \cancel{S_{34}} + S_{123} + S_{124} + S_{134} + \cancel{S_{234}} + S_{1234} = 1$$

$$S_{T1} = S_1 + S_{12} + S_{13} + S_{14} + S_{123} + S_{124} + S_{134} + S_{1234}$$
$$S_{T2} = S_2 + S_{12} + S_{23} + S_{24} + S_{123} + S_{124} + S_{234} + S_{1234}$$

(Factors Fixing)

Approach for SE

Overview of the Approach



Determining Distributions

- Distribution of the values of the input parameteters
- Derived from
  - scientific literature
  - physical boundaries
  - expert opinion
  - surveys
  - experiments

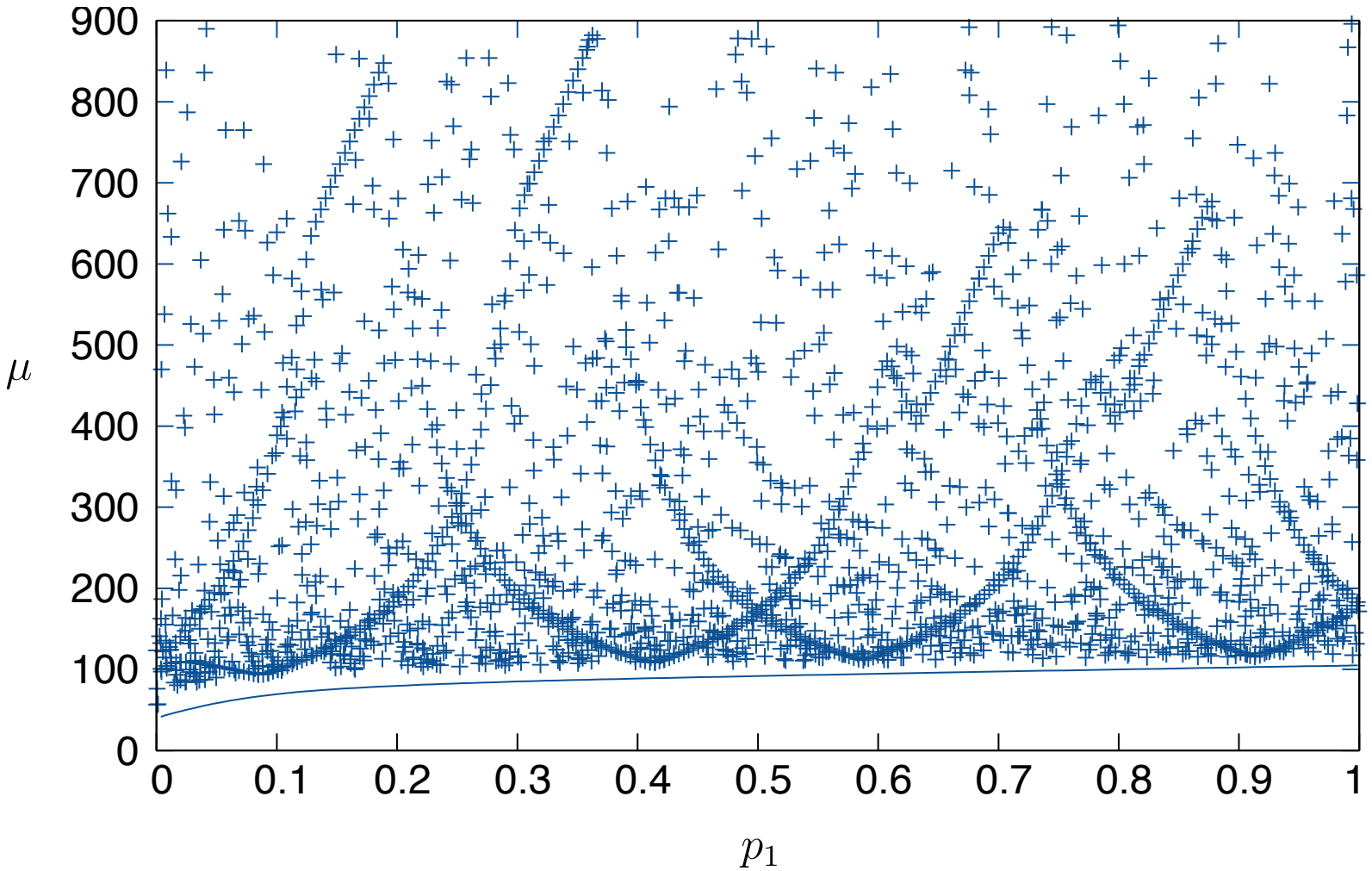
Example: Distributions

- Based on expert opinion
- General parameter
  - p_1 uniformly distributed between 0 and 1
  - d uniformly distributed between 0.9 and 1
- Dependent on application
  - *inf* e.g.. uniformly distributed between 50 and 500
  - For t e.g. $t \sim \mathcal{N}(1000, 200)$

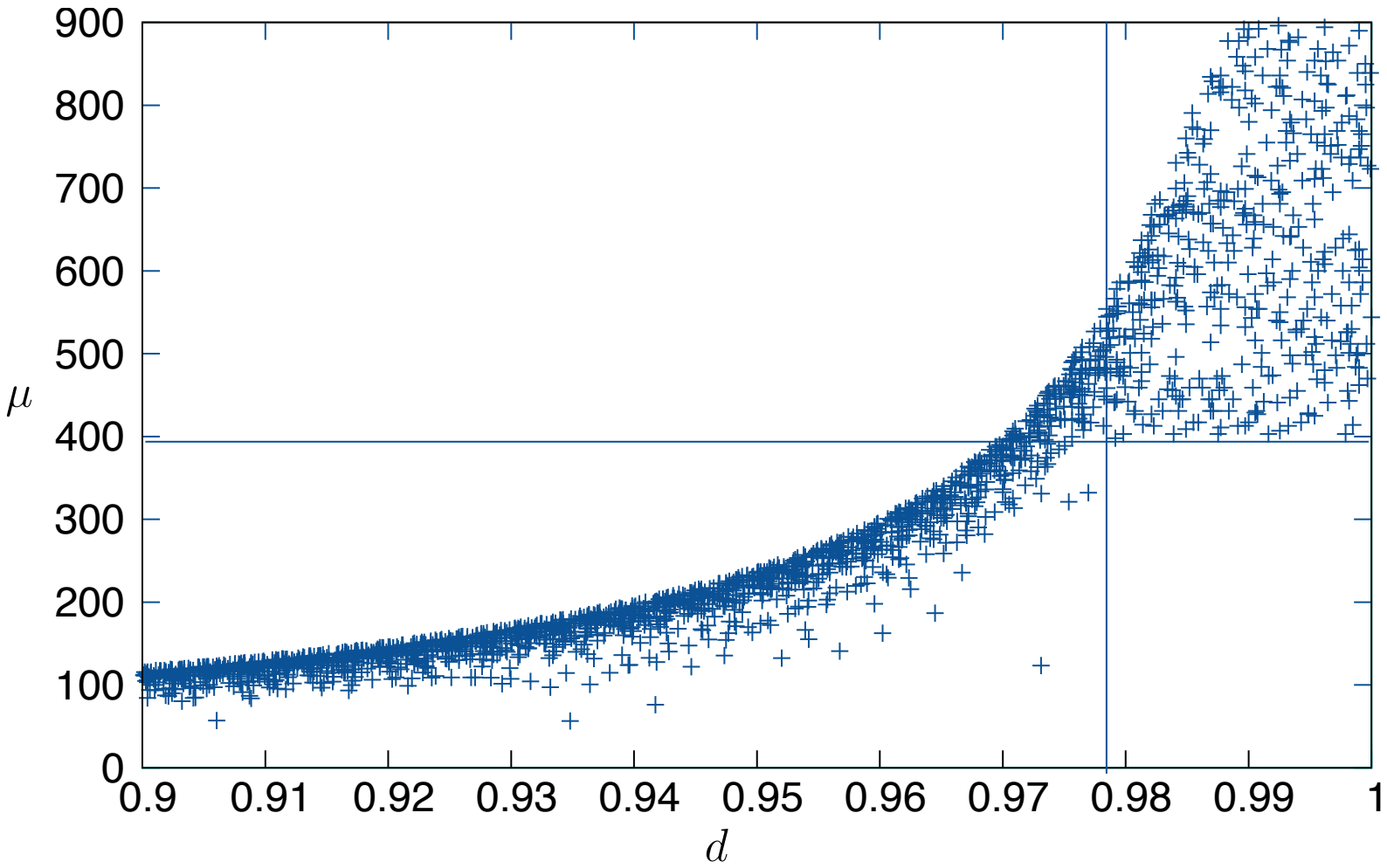
Visualising Using Scatterplots

- Sampling using the distributions
- Pairwise relationship between factors
- Detection of errors in the model
- Indications of influences

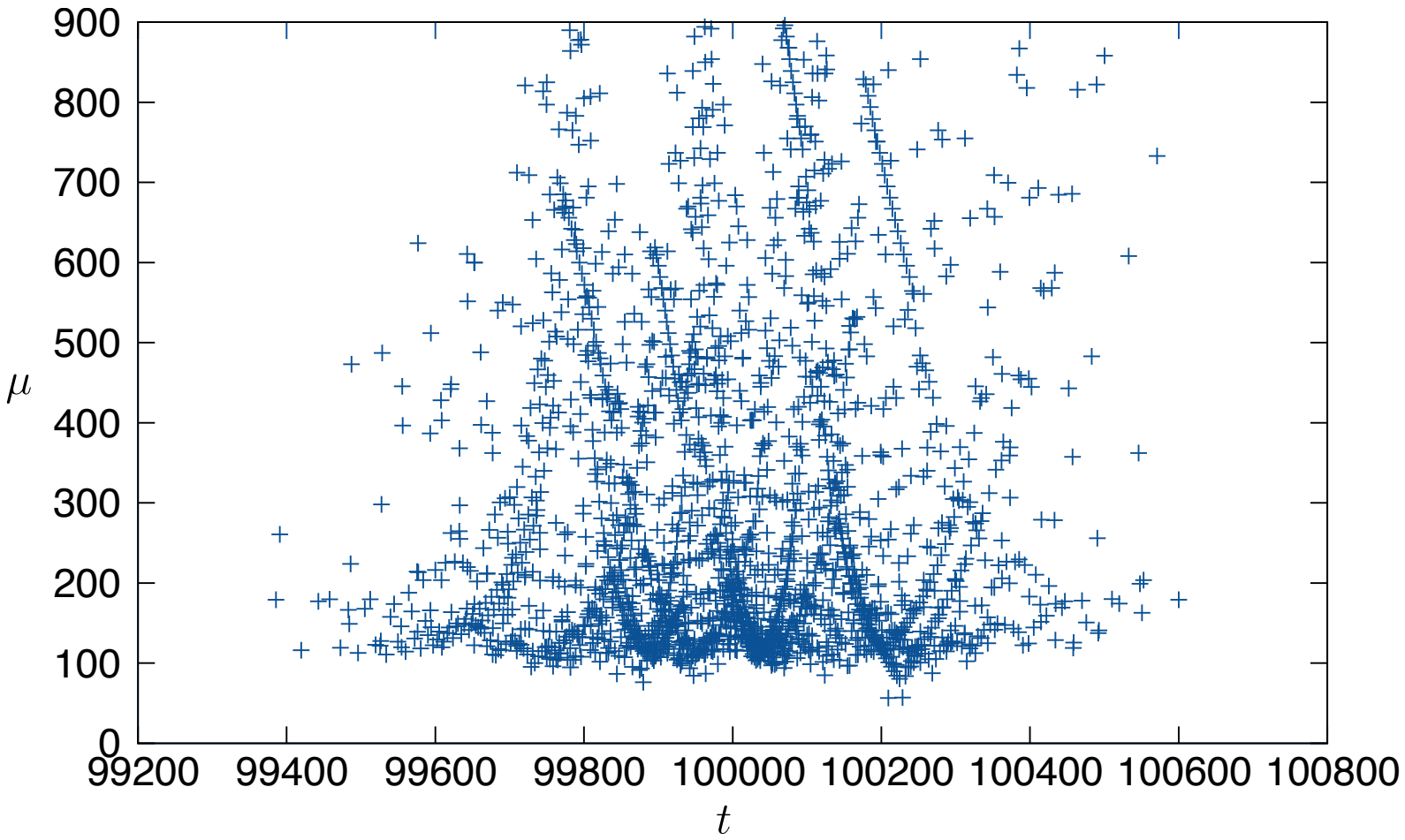
p_1 and μ



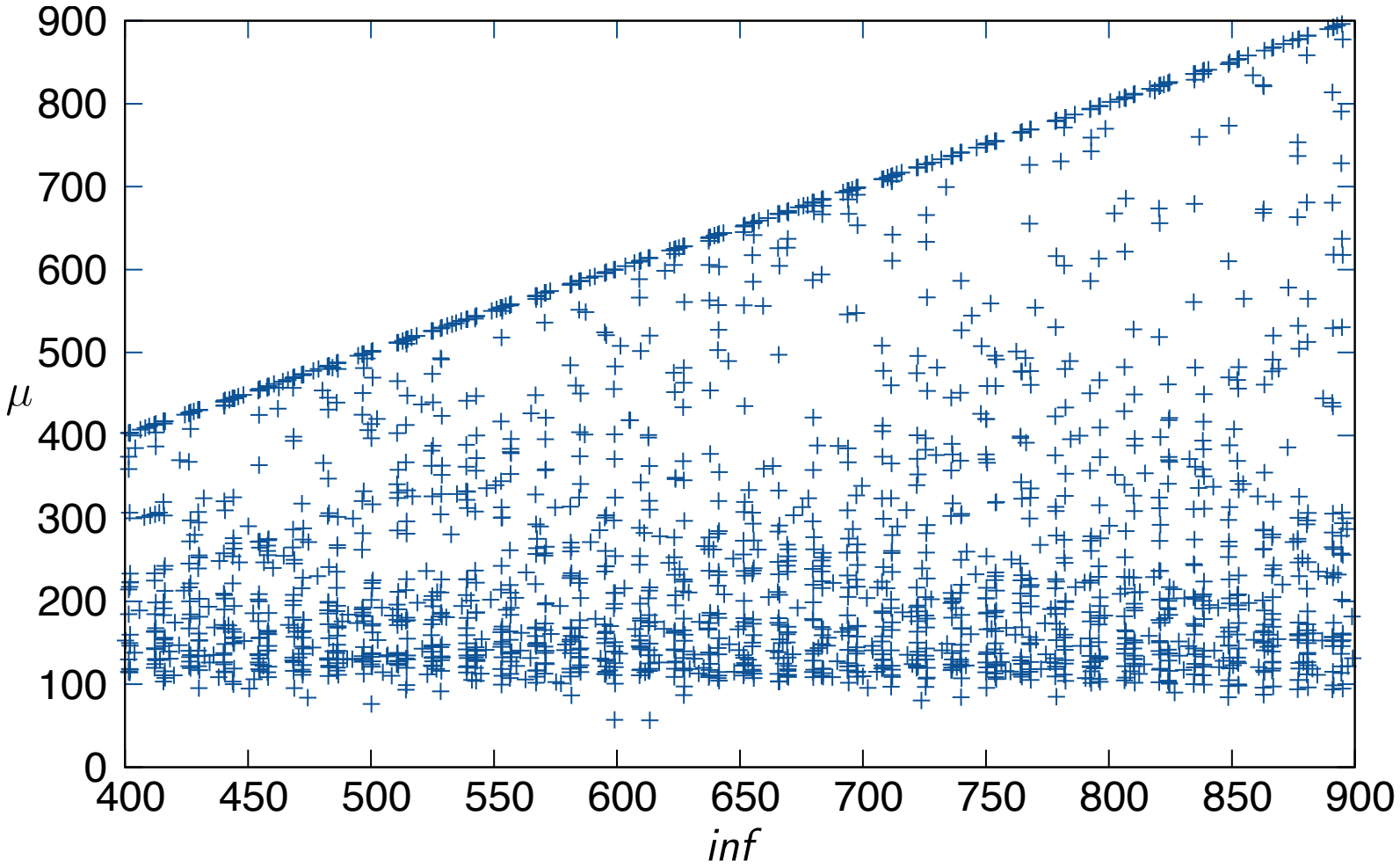
d and μ



t and μ



inf and μ



Applying Global SA

- Various methods for calculating indices
- FAST (Fourier Amplitude Sensitivity Test) gives quantitative results
- Uses sampled data
- Tool-support: Simlab

Main Effect

t	6.88e-5
p_1	0.0090
d	0.9026
inf	0.0158

Indizes

Main Effect

t	6.88e-5
p_1	0.0090
d	0.9026
inf	0.0158

Total Effect

t	0.005788
p_1	0.022365
d	0.975155
inf	0.086735

Summary

Conclusions

- Sensitivity analysis useful tool for predictor models
- SA can help to
  - find errors in the models
  - simplify the models
  - identify interactions between input parameters
  - identify parameters that should be investigated more
  - get more robust predictions
- Experience with
  - reliability model
  - QA economics model
  - process model
  - expert system for IT tools
- Good tool-support available (Simlab: `http://simlab.jrc.cec.eu.int/`)