Centrum Kompetencji
w Zakresie Rozproszonych Infrastruktur
Obliczeniowych Typu Gridowrgo – PLGrid Core

Parameter Sweep and Resources Scaling Automation in Scalarm Data Farming Platform

J. Liput, M. Paciorek, M. Wrona, M. Orzechowski, R. Slota, and J. Kitowski

ACC Cyfronet AGH
Department of Computer Science, AGH UST

SCALARM

CGW'15 Krakow, Poland
26-28 October 2015

INNOWACYJNA GOSPODARKA
NARODOWA STRATEGIA SPÓJNOŚCI

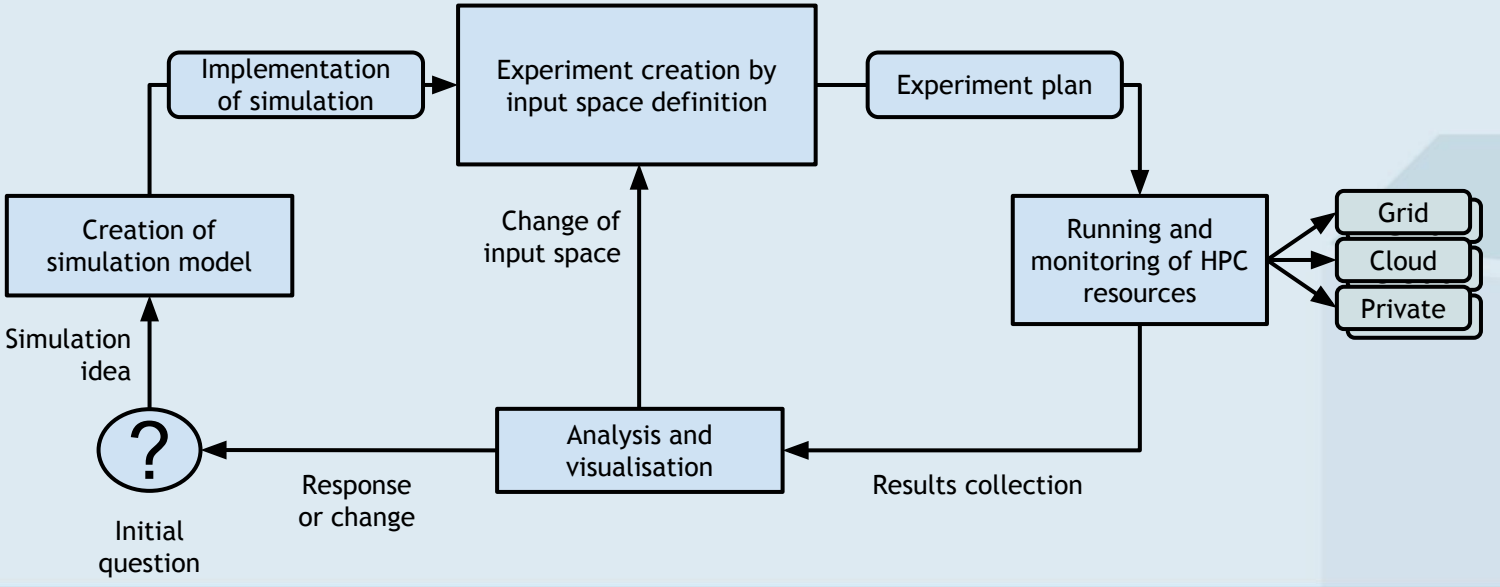
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ
ROZWOJU REGIONALNEGO

Agenda

- Data processing in modern science
- Problem description
- Scalarm overview
- Scalarm approach to automation
- Results
- Conclusions and future work

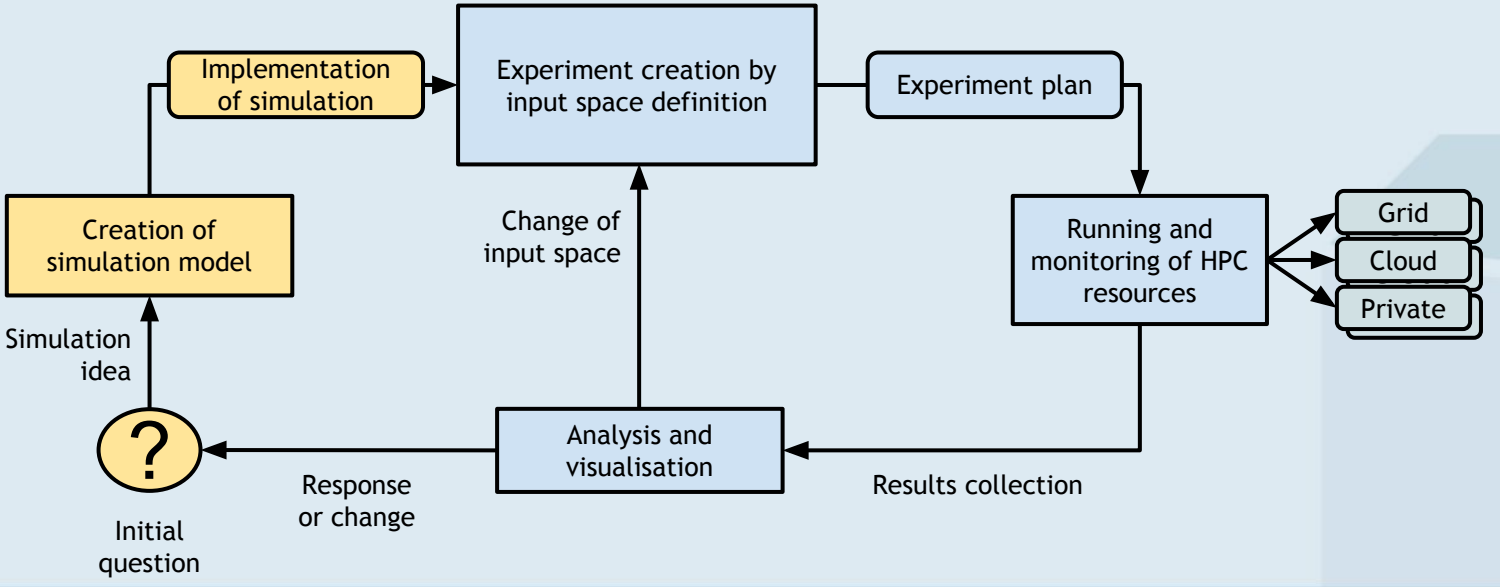
Data processing in modern science

- Scientific research methods often rely on executing numerous simulations each with different input parameter values
- One such approach is called data farming



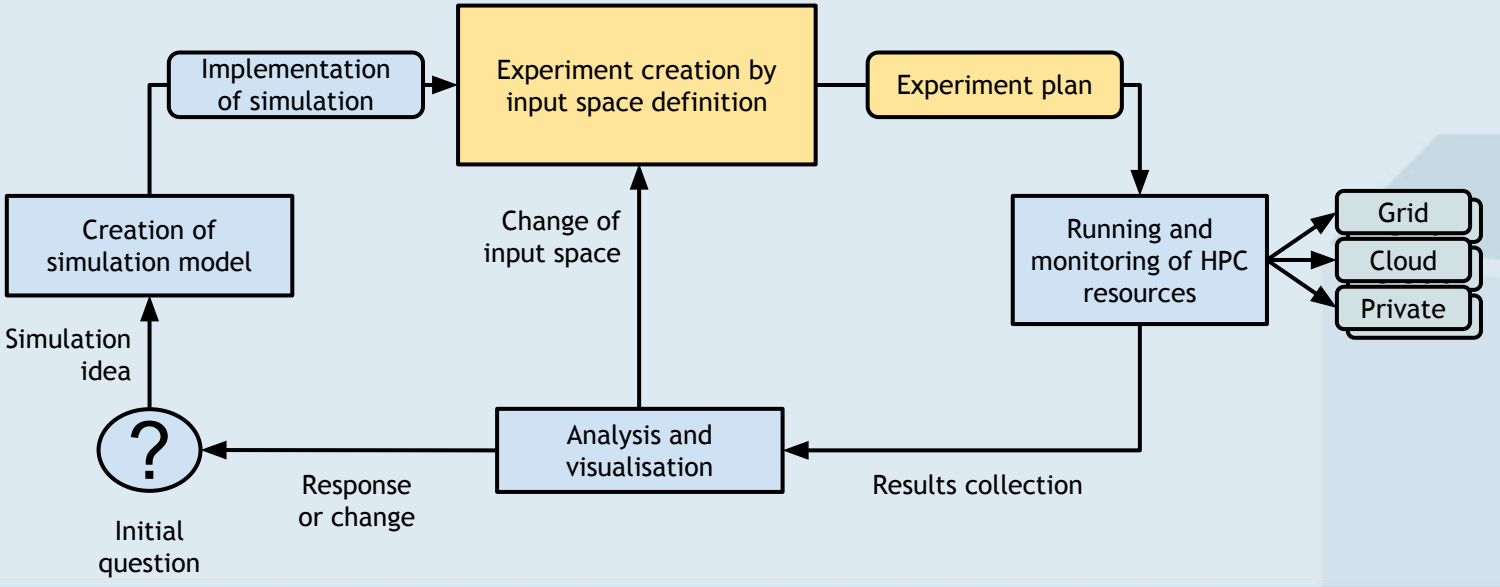
Data processing in modern science

- Scientific research methods often rely on executing numerous simulations each with different input parameter values
- One such approach is called data farming



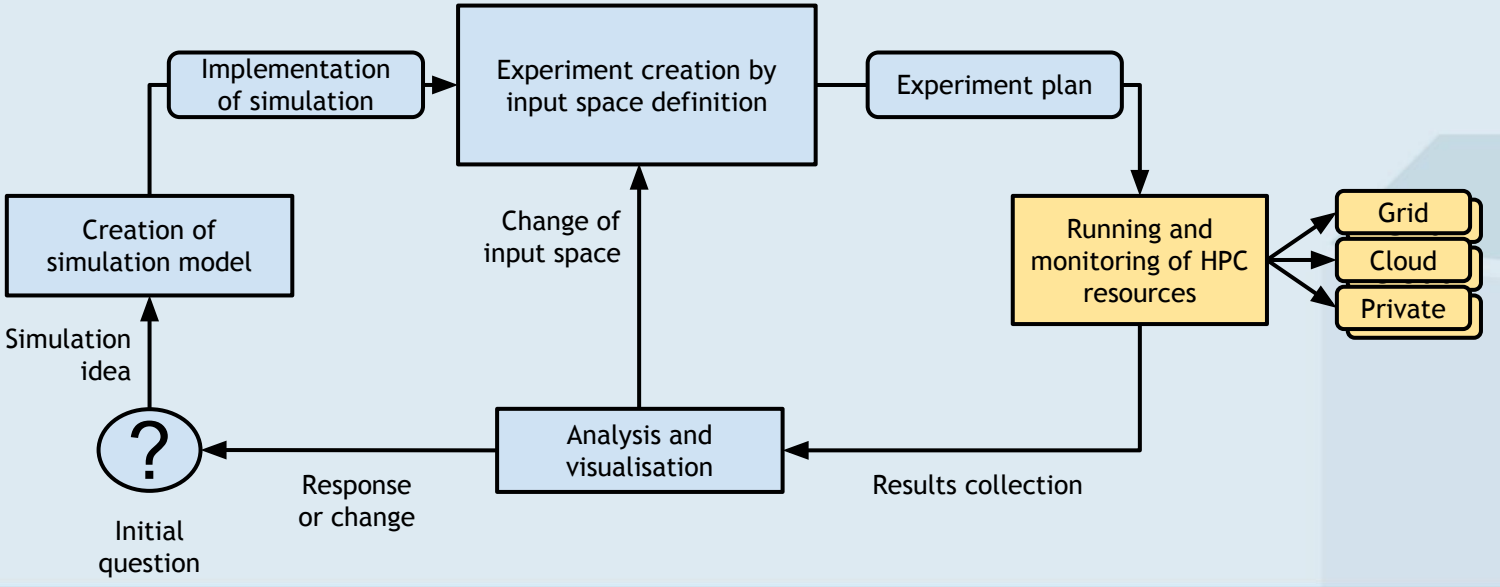
Data processing in modern science

- Scientific research methods often rely on executing numerous simulations each with different input parameter values
- One such approach is called data farming



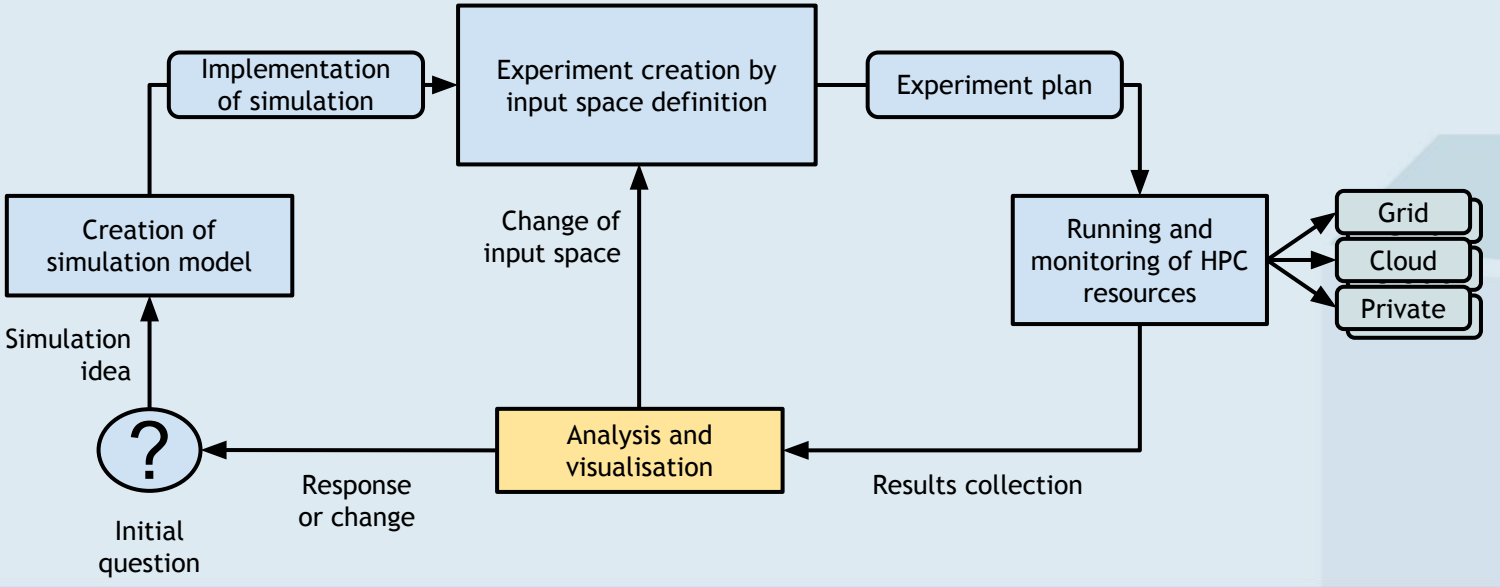
Data processing in modern science

- Scientific research methods often rely on executing numerous simulations each with different input parameter values
- One such approach is called data farming



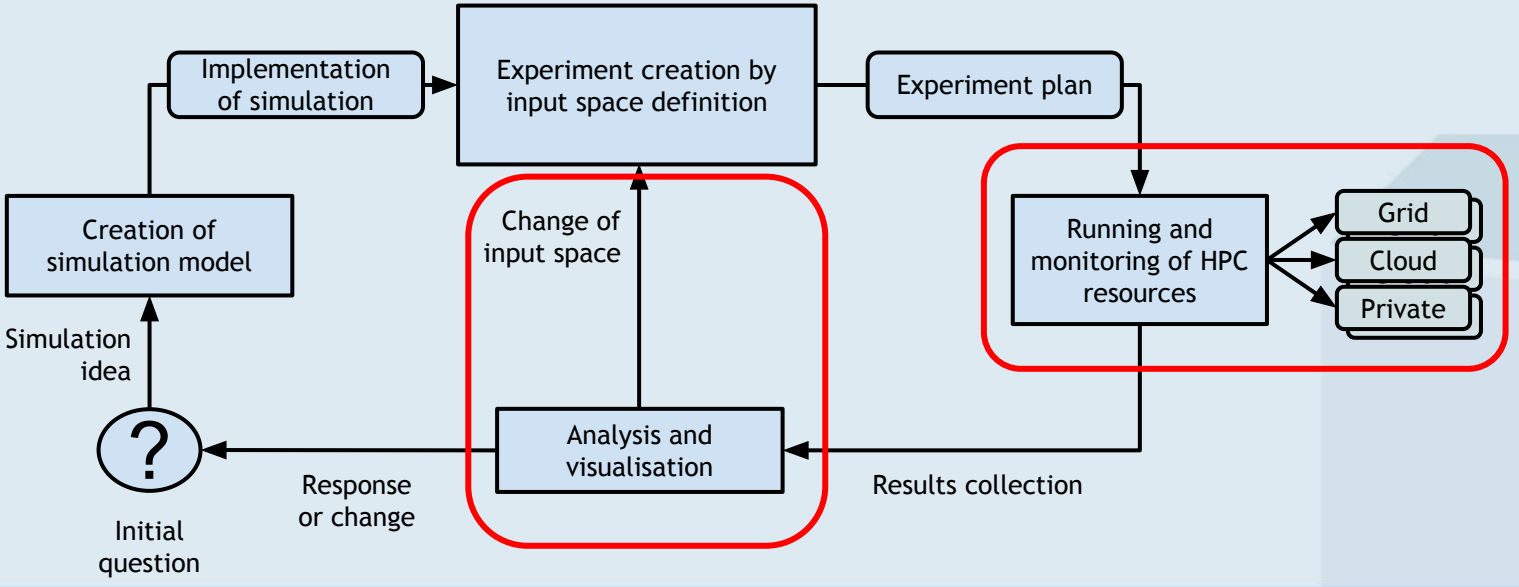
Data processing in modern science

- Scientific research methods often rely on executing numerous simulations each with different input parameter values
- One such approach is called data farming



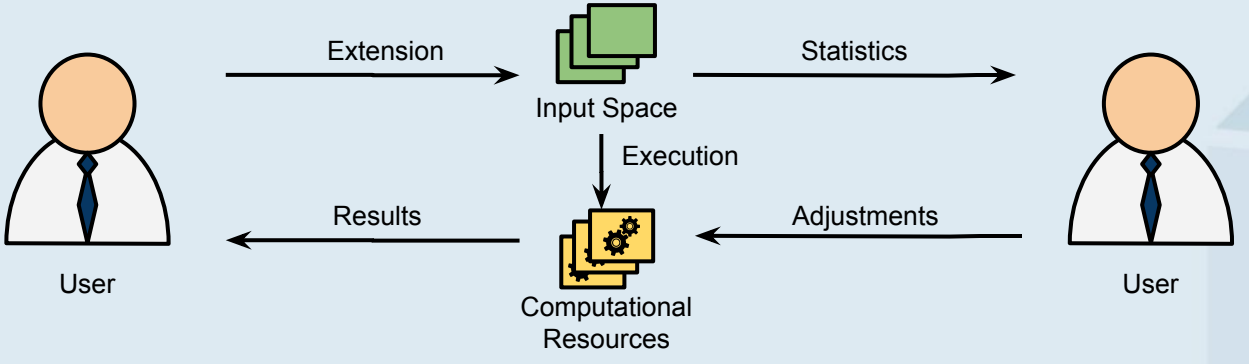
Problem description

- Data science computation often requires input space adjustments according to collected partial results
- Need of resources management according to changing computational power requirements



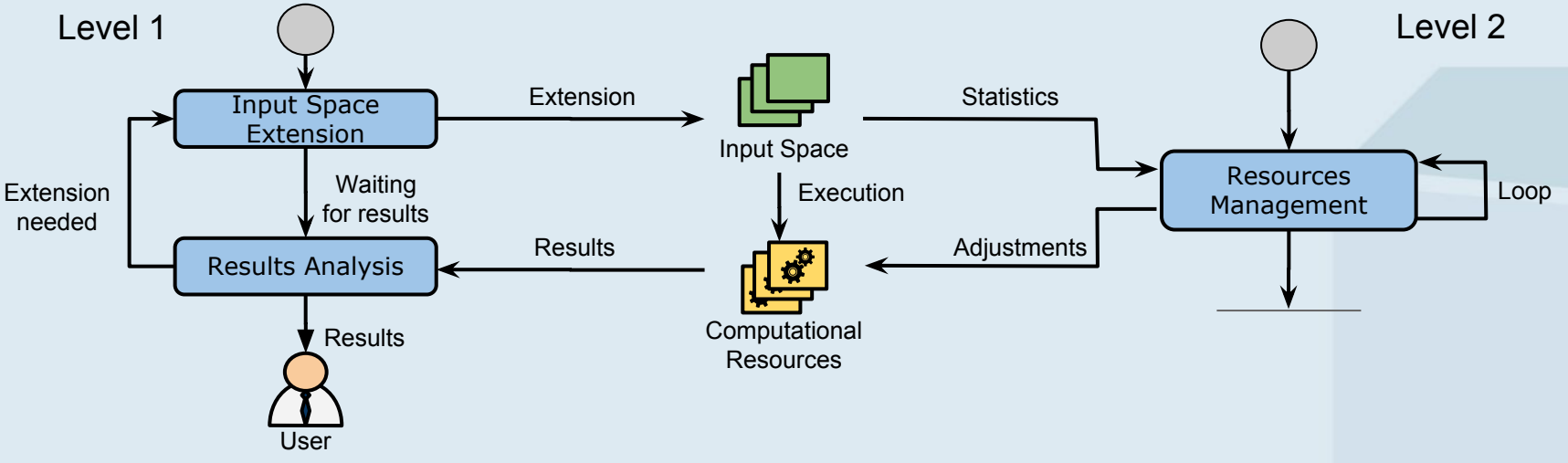
Scalarm overview

- Scalarm - a platform for data farming, allows user to execute experiments in convenient way
- Unified management of heterogeneous resources
- Partial results analysis with numerous methods
- Input space extension during experiment



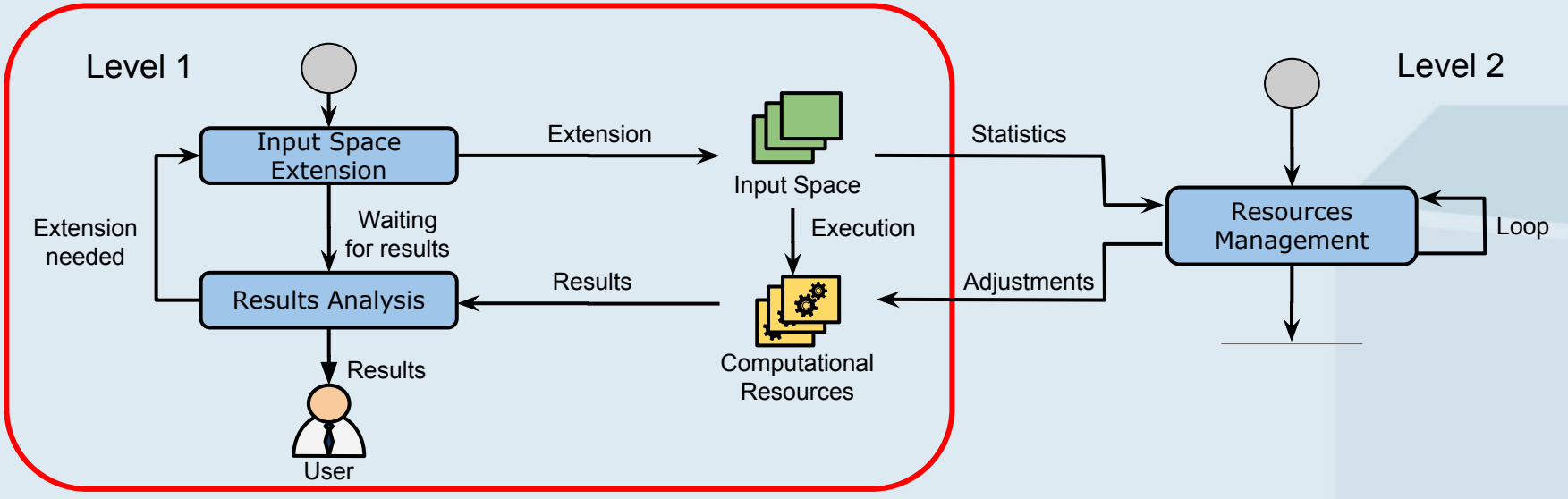
Scalarm approach to automation

- Two levels of experiment automation:
  - Input space adjustment
  - Resources management



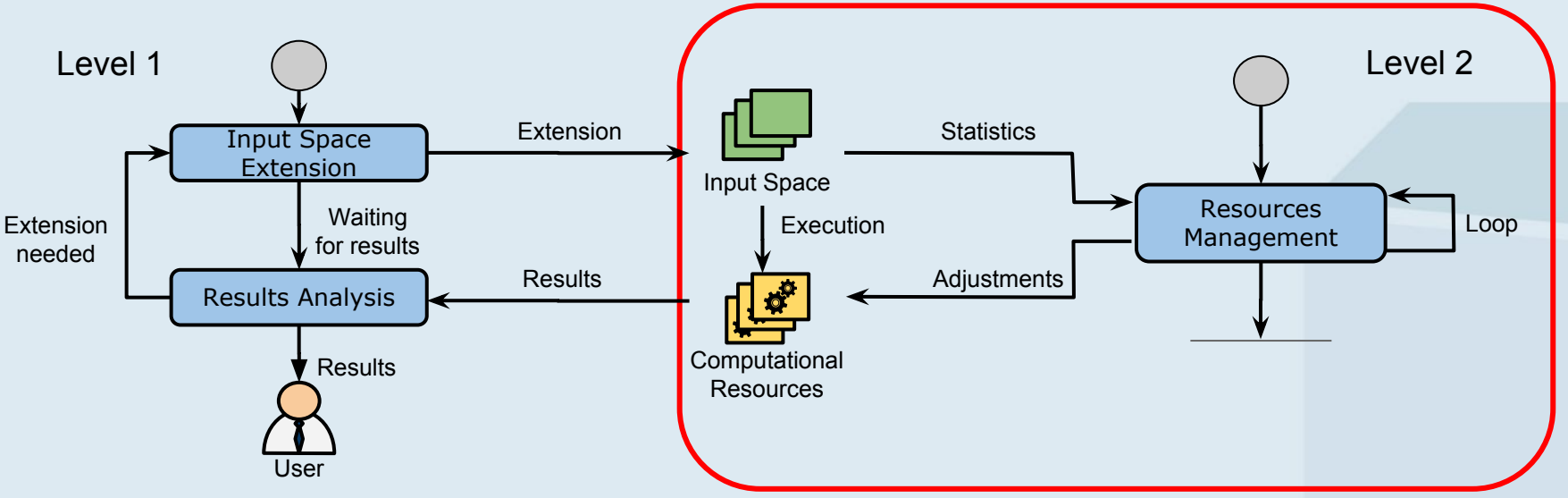
Input space adjustment

- Input space management algorithm:
  - Input space extension
  - Analysis of requested data
- Dedicated simulated annealing algorithm



Resources management

- Resources management algorithm:
  - Pulling metrics about current state
  - Metrics analysis
  - Increasing or decreasing amount of workers



Resources management metrics

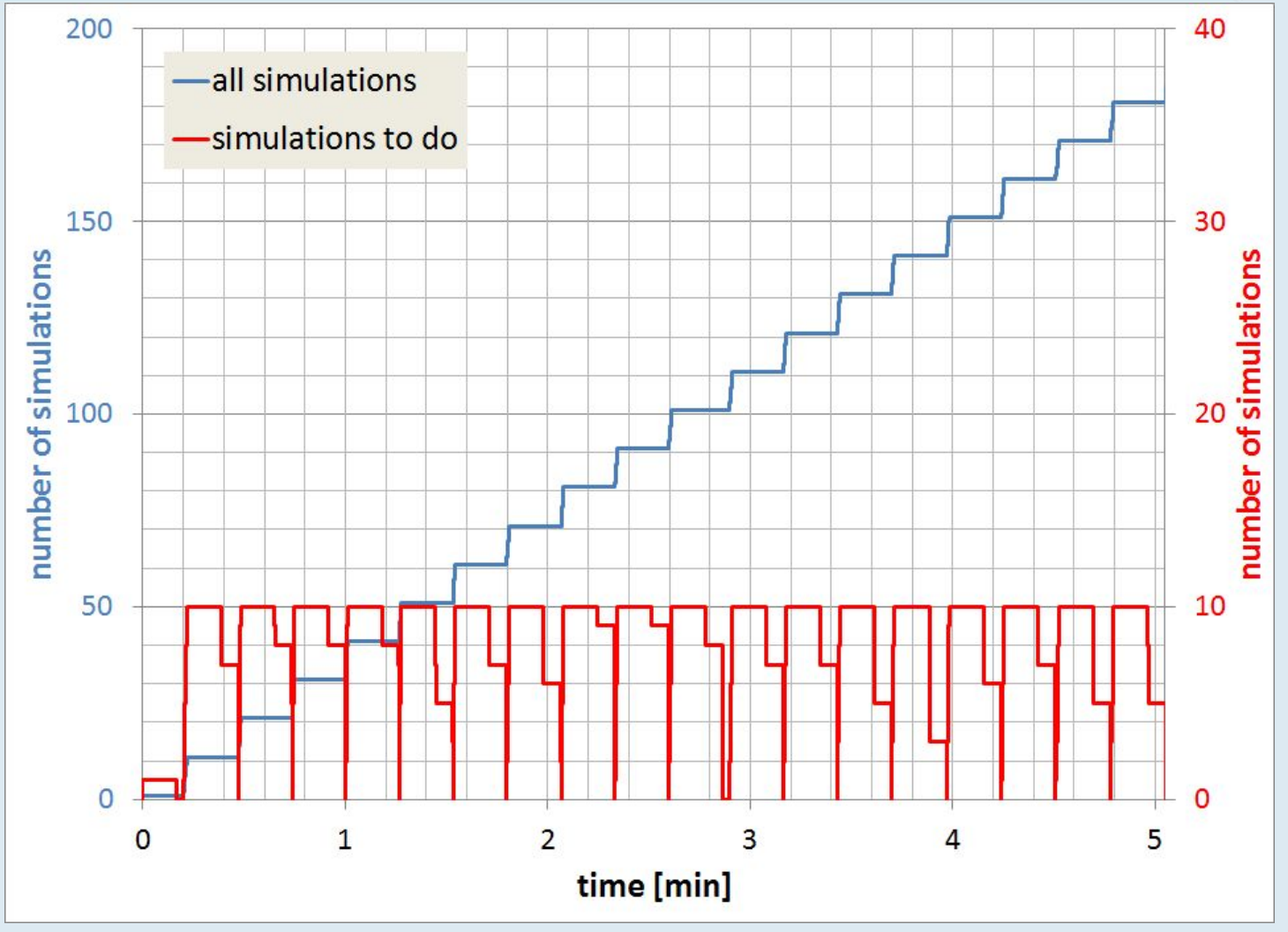
- Metrics used during resources management
  - workers throughput $T_W = \frac{done\ simulations}{execution\ time}$
  - system throughput $T_S = \sum T_W$
  - target throughput $T_T = \frac{simulations\ to\ run}{time\ left}$
  - makespan [time] $M = \frac{simulations\ to\ run}{T_S}$

Evaluation

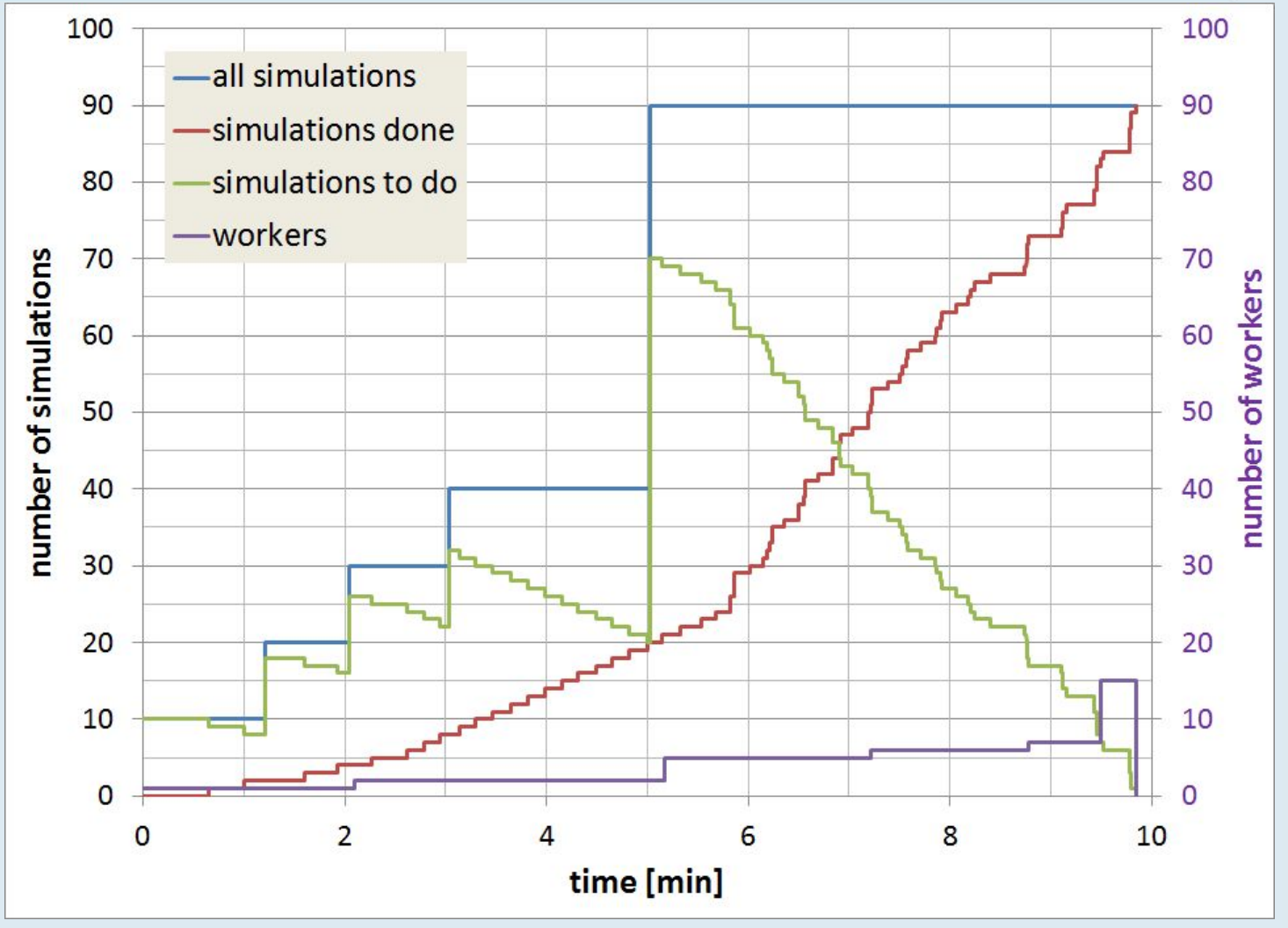
Test 1: Automated input space extension evaluation
- Input space controlled by simulated annealing algorithm
- Fixed time of simulation execution - 10 seconds
- Fixed number of workers - 10

Test 2: Automated resources management evaluation
- Resources controlled by our resources management algorithm
- Fixed time of simulation execution - 20 seconds
- Manual input space extensions
- Experiment execution time constrained to 10 minutes

Automated input space extensions



Automated resources management



Conclusions

- Two levels of automation - input space adjustments and resources management
- Plugin-based architecture allows an easy extension with new algorithms
- Integration of these levels of automation is challenging
  - Automated input space extension requires calculation of all simulation from ‘bundle’ before scheduling next one
  - Resources management algorithm must take into account input space extension by bundle of simulations

Future Work

- Resources management algorithm better suited to data farming experiments
  - Predicting amount of simulation yet to be scheduled based on available data
  - Metrics extension
- Additional dedicated input space management algorithms, e.g. genetic algorithm