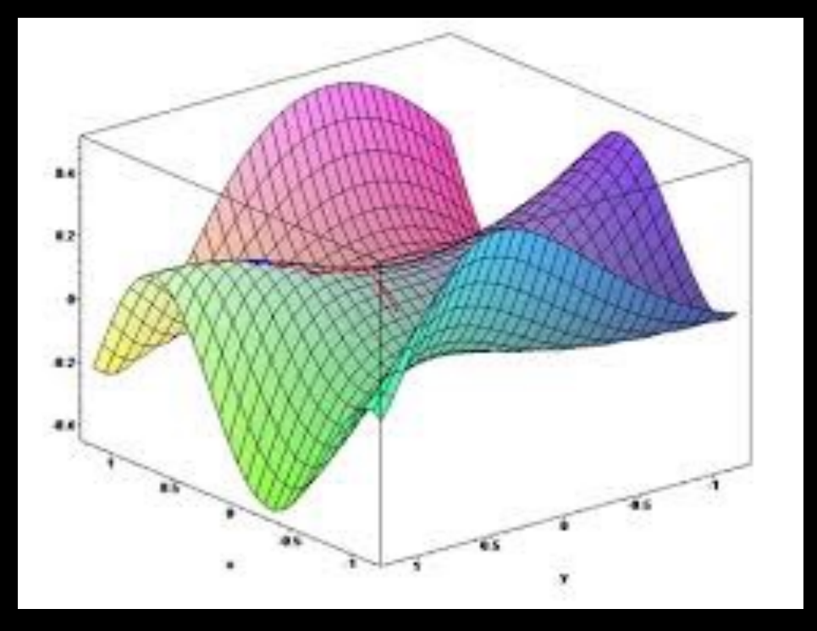
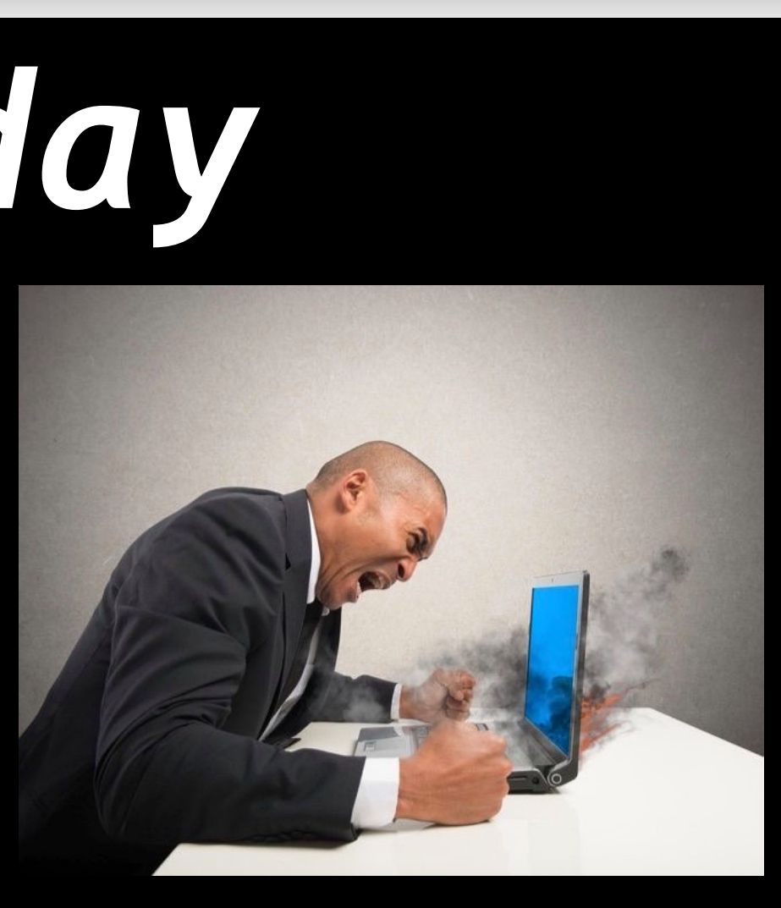
Improving the life of a data scientist

Data scientist today

What people think he does

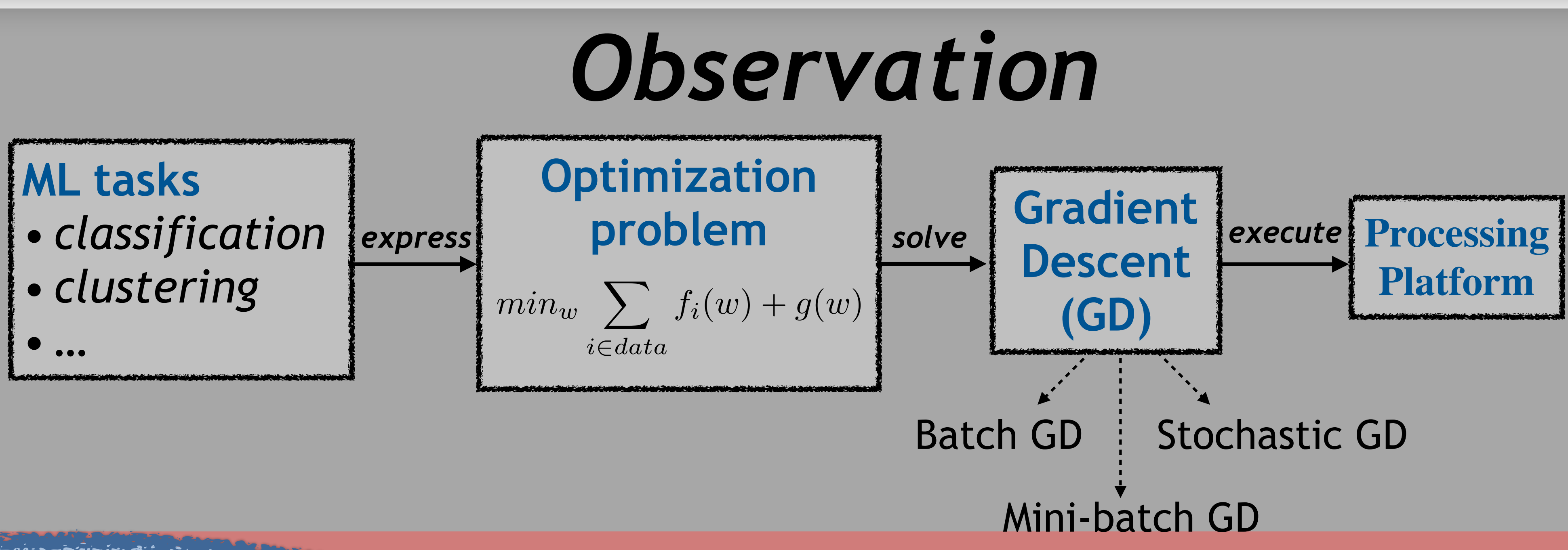


What he thinks he does



What he actually does

Observation

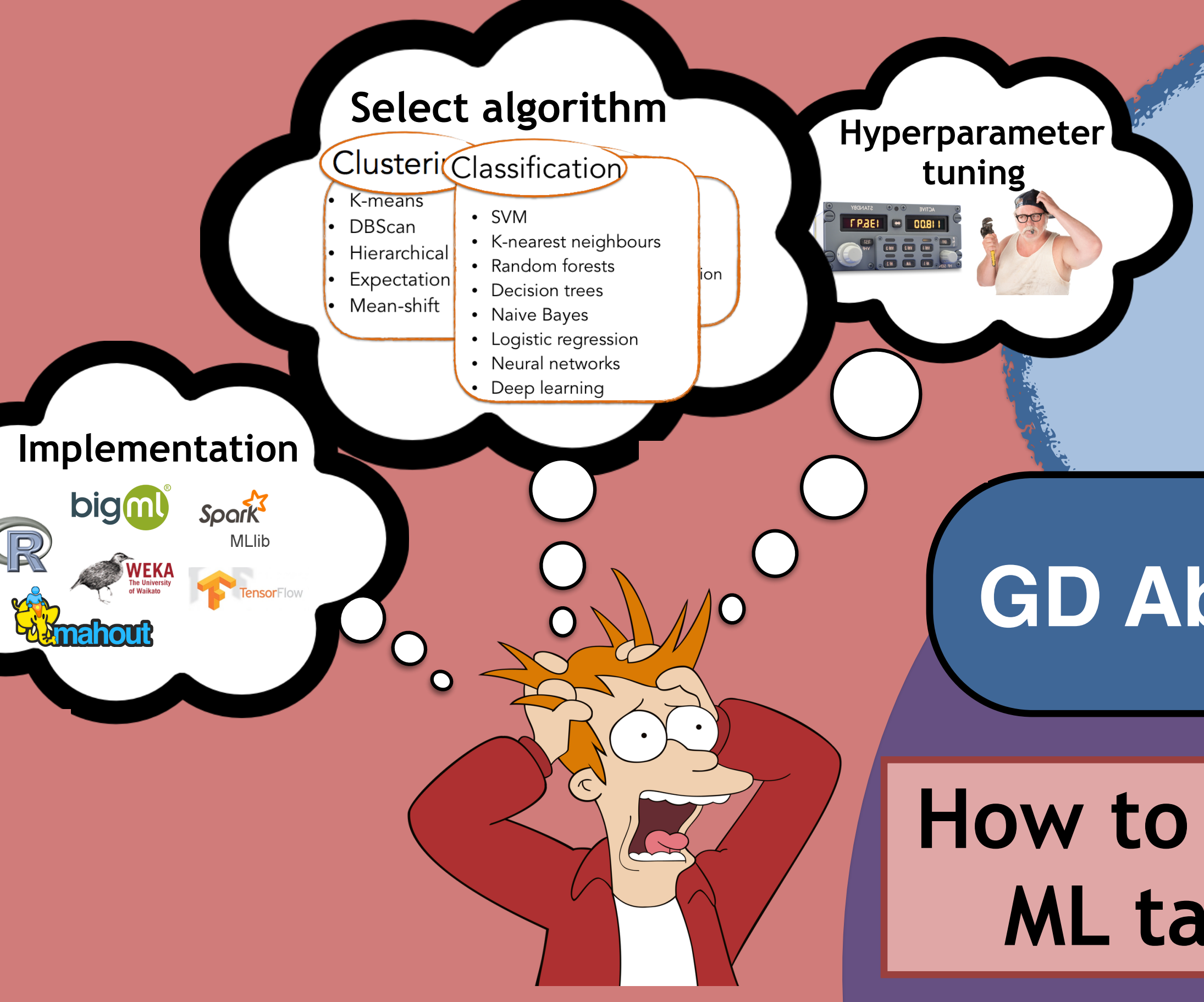


ML tasks
- classification
- clustering
- …

express → Optimization problem $min_w \sum_{i \in data} f_i(w) + g(w)$ → solve → Gradient Descent (GD) → execute → Processing Platform

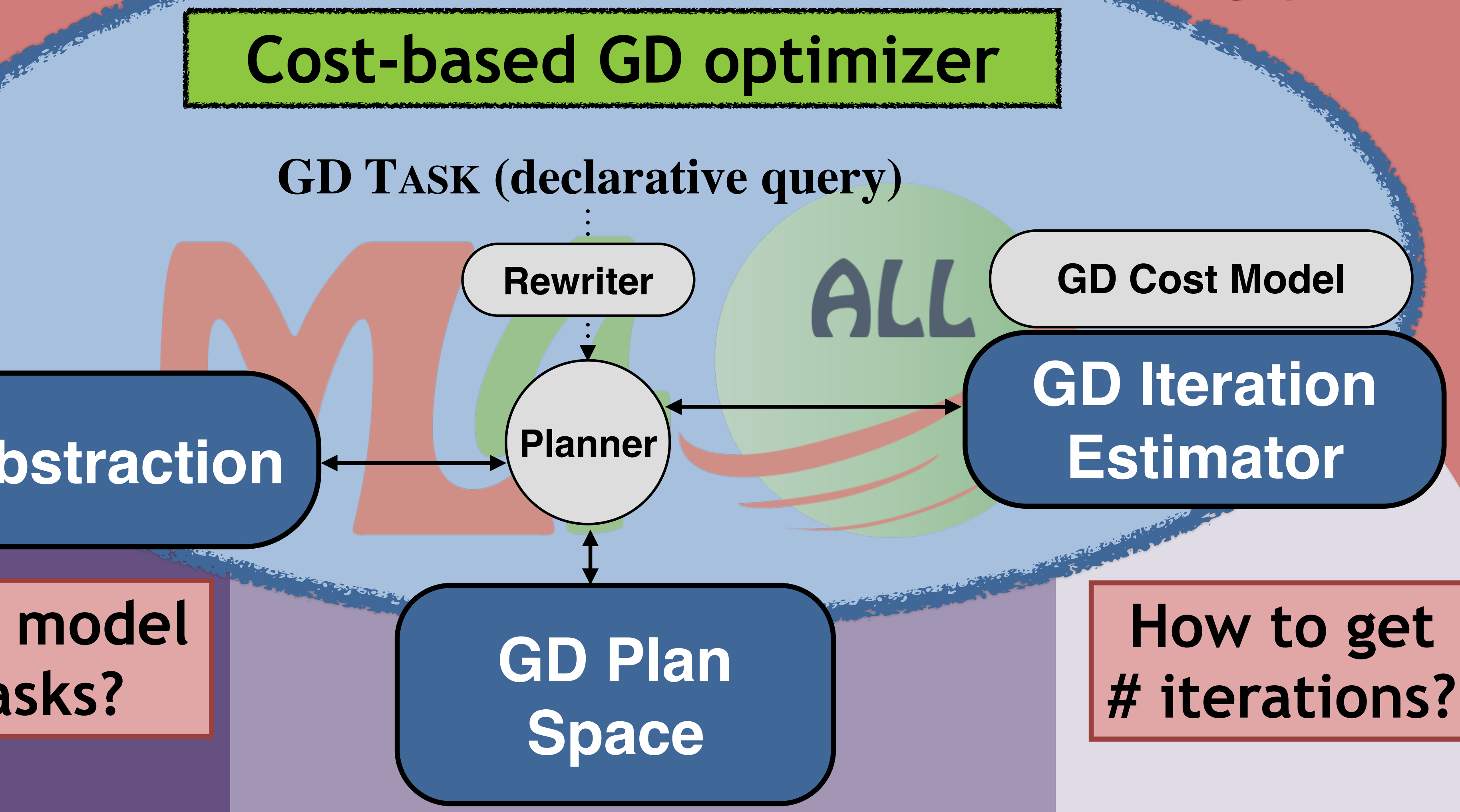
Batch GD · Stochastic GD · Mini-batch GD

General Problem



Solution

Cost-based GD optimizer



GD TASK (declarative query)

Rewriter

GD Cost Model

Planner

GD Abstraction

GD Iteration Estimator

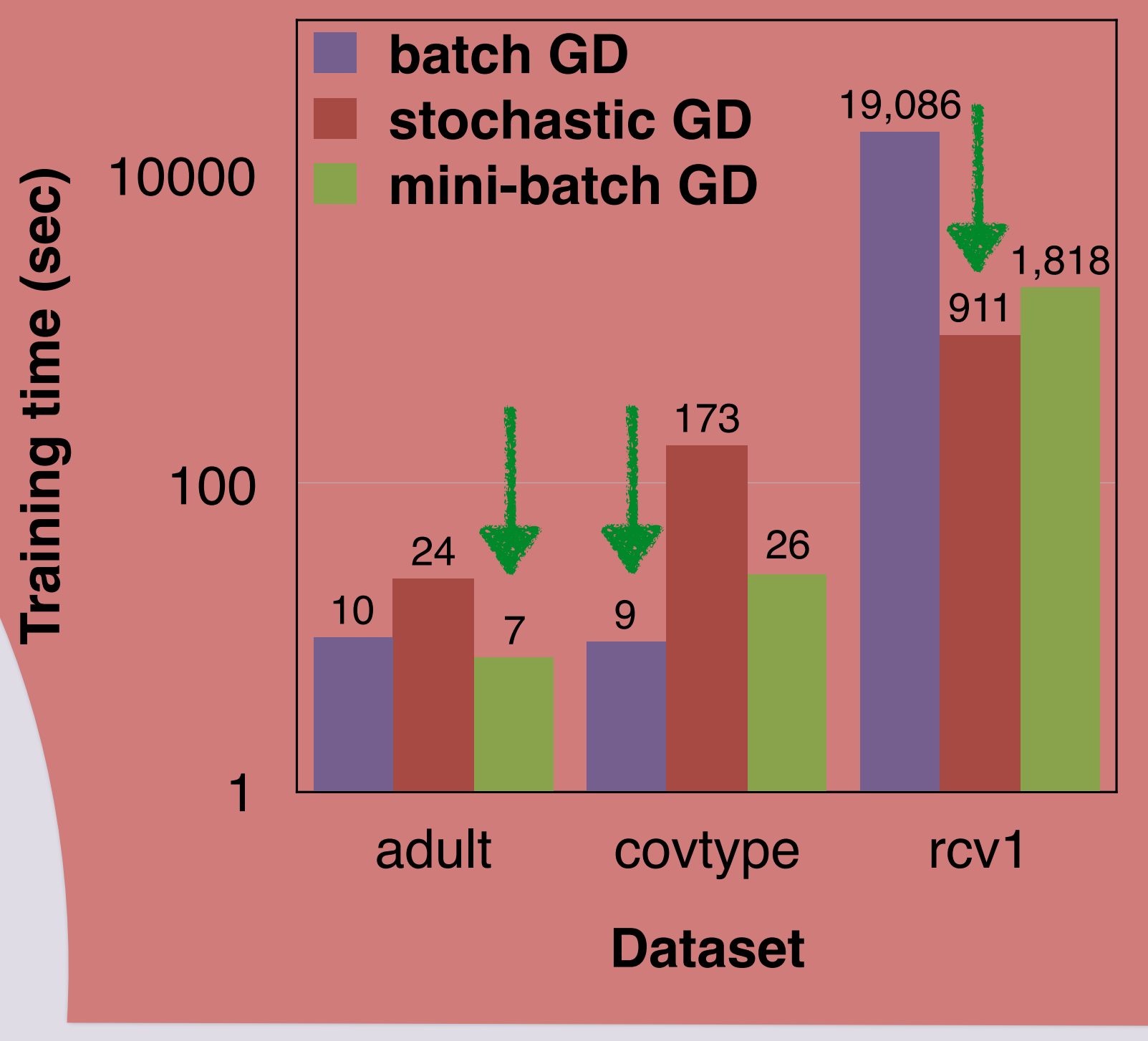
GD Plan Space

How to model ML tasks?

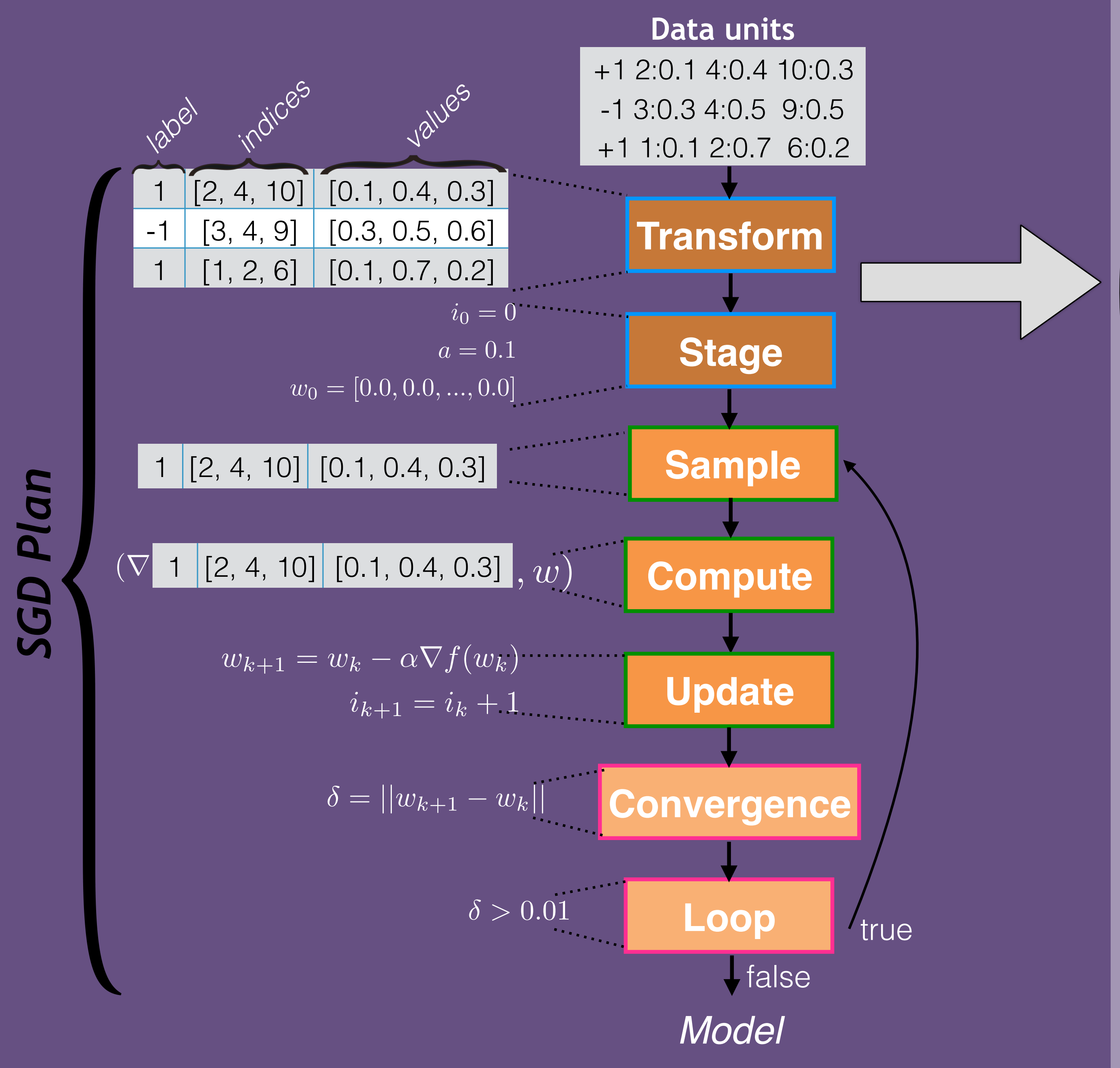
How to get # iterations?

Our Focus

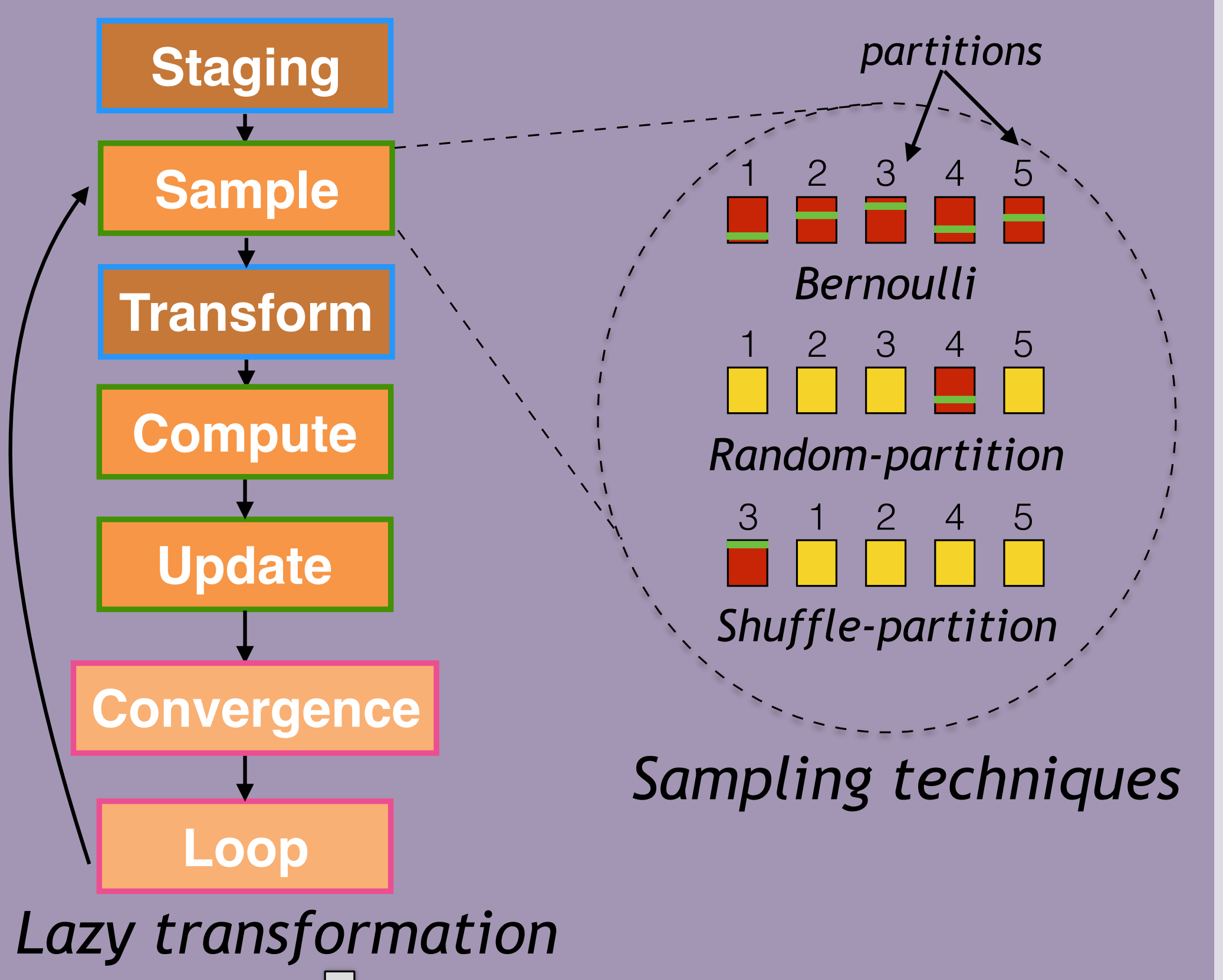
Not all-times winner



1. Preparation phase
2. Processing phase
3. Convergence phase

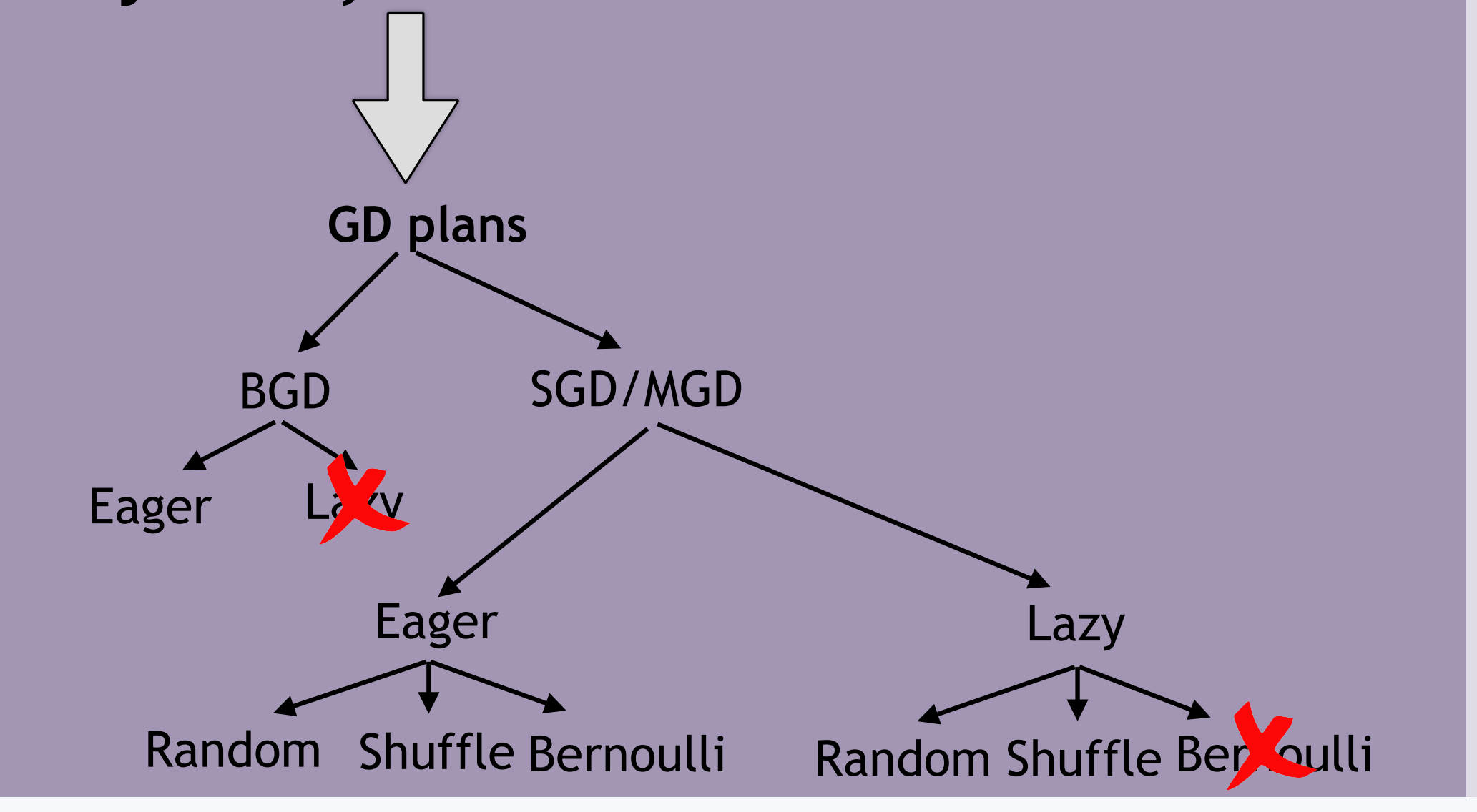


How to optimize GD plans?



Lazy transformation

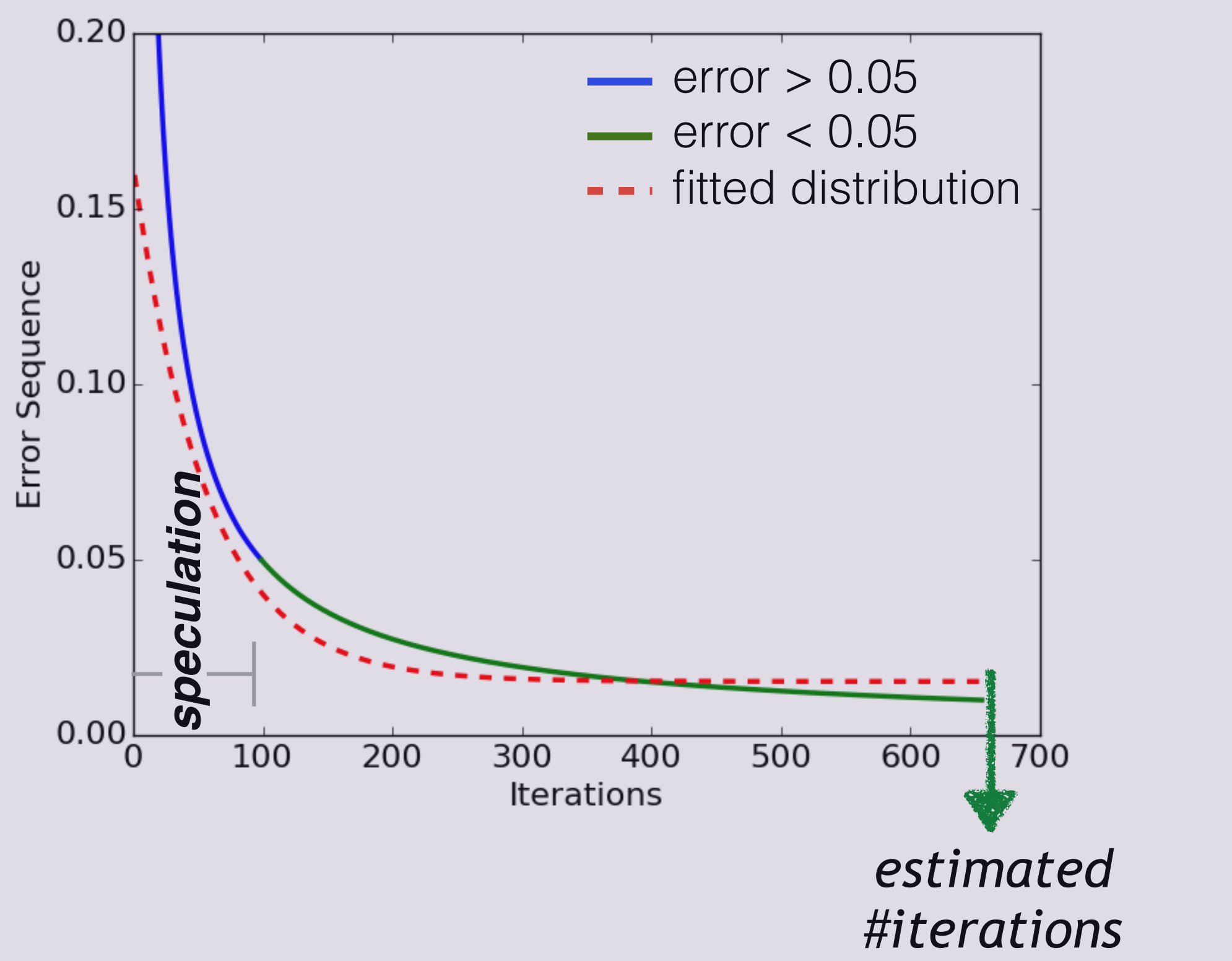
Sampling techniques



1. Error sequence follows known distribution
2. Shape of error sequence on sample D’ << D $\simeq$ Shape of error sequence over D

Speculative approach

1. Take a sample D’ << D
2. Run GD for a larger error
3. Fit the distribution



More info

A DB-like Machine Learning System

Source code of abstraction available at https://github.com/rheem-ecosystem/rheem

A Cross-Platform System

Visit our web page: http://da.qcri.org/rheem

Source code: https://github.com/rheem-ecosystem/rheem

- **Rheem: Enabling Multi-Platform Task Execution**
  SIGMOD 2016, San Francisco, USA (demo paper)
- **Road to Freedom in Data Analytics**
  EDBT 2016, Bordeaux, France (vision paper)

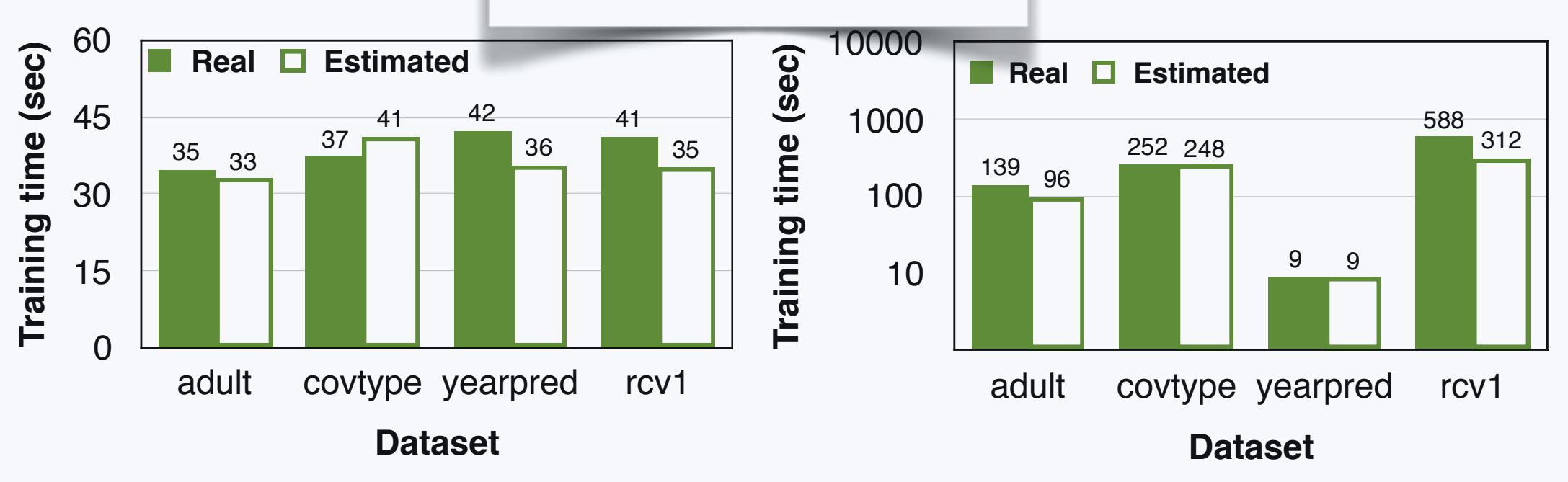
Join us at the
Spark Summit 2017
San Francisco, USA

ML4all

Rheem

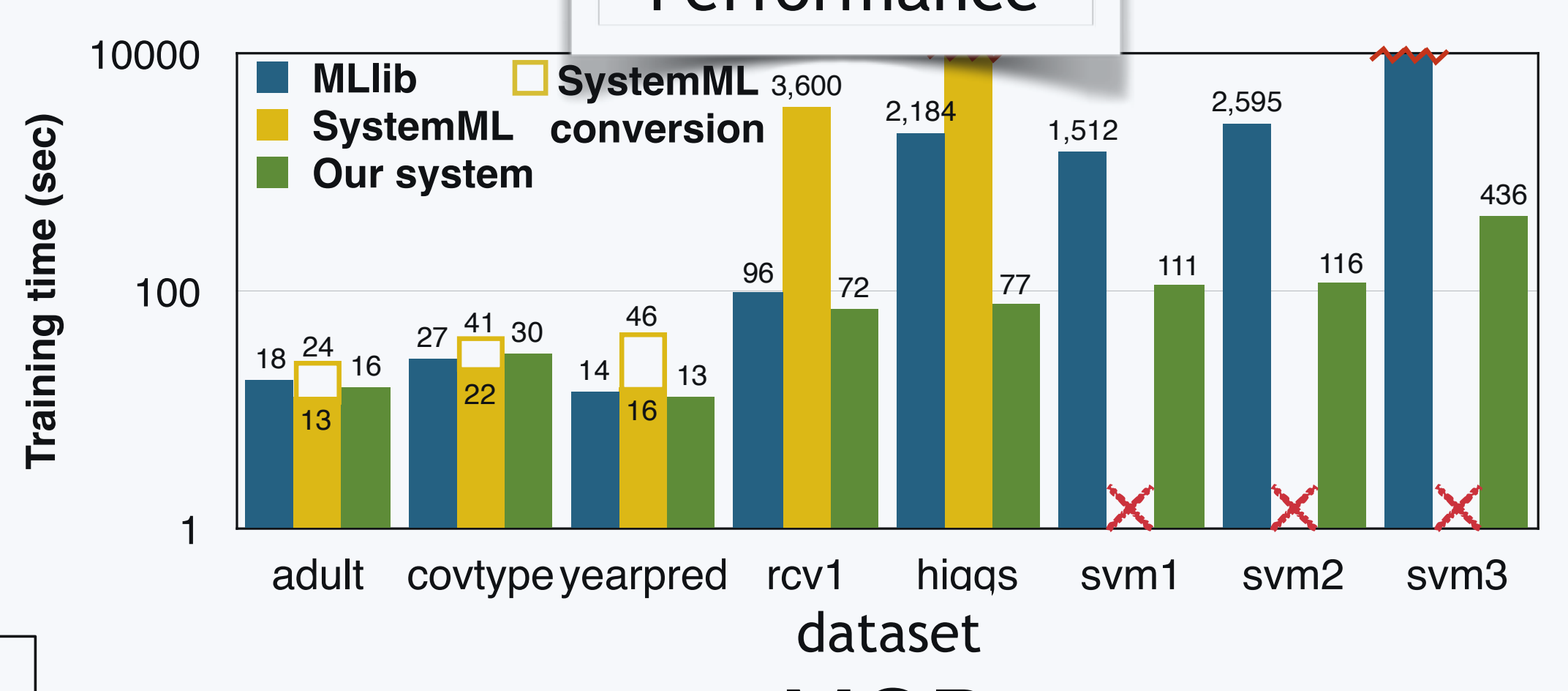
Results

Time estimates

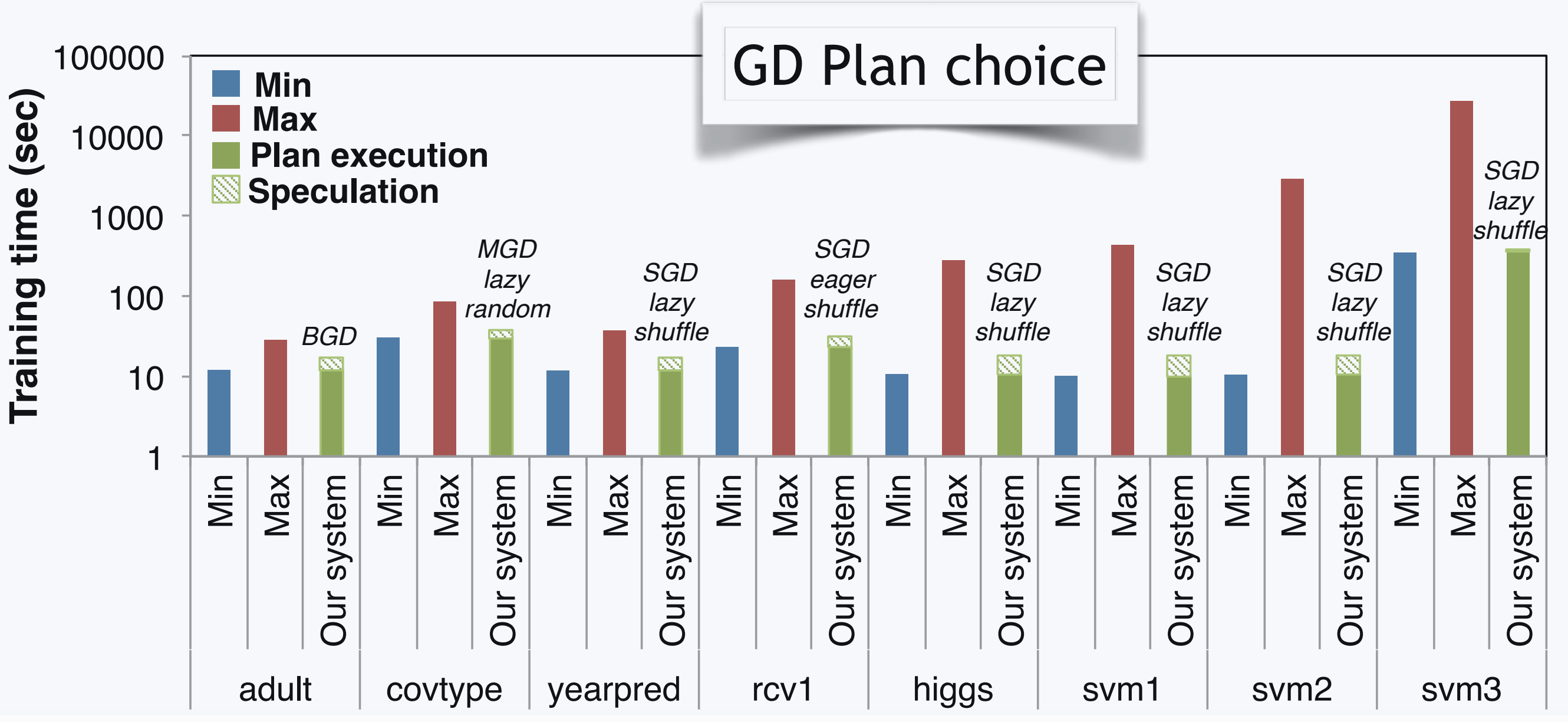


(a) Run of 1,000 iterations (b) Run to convergence

Performance

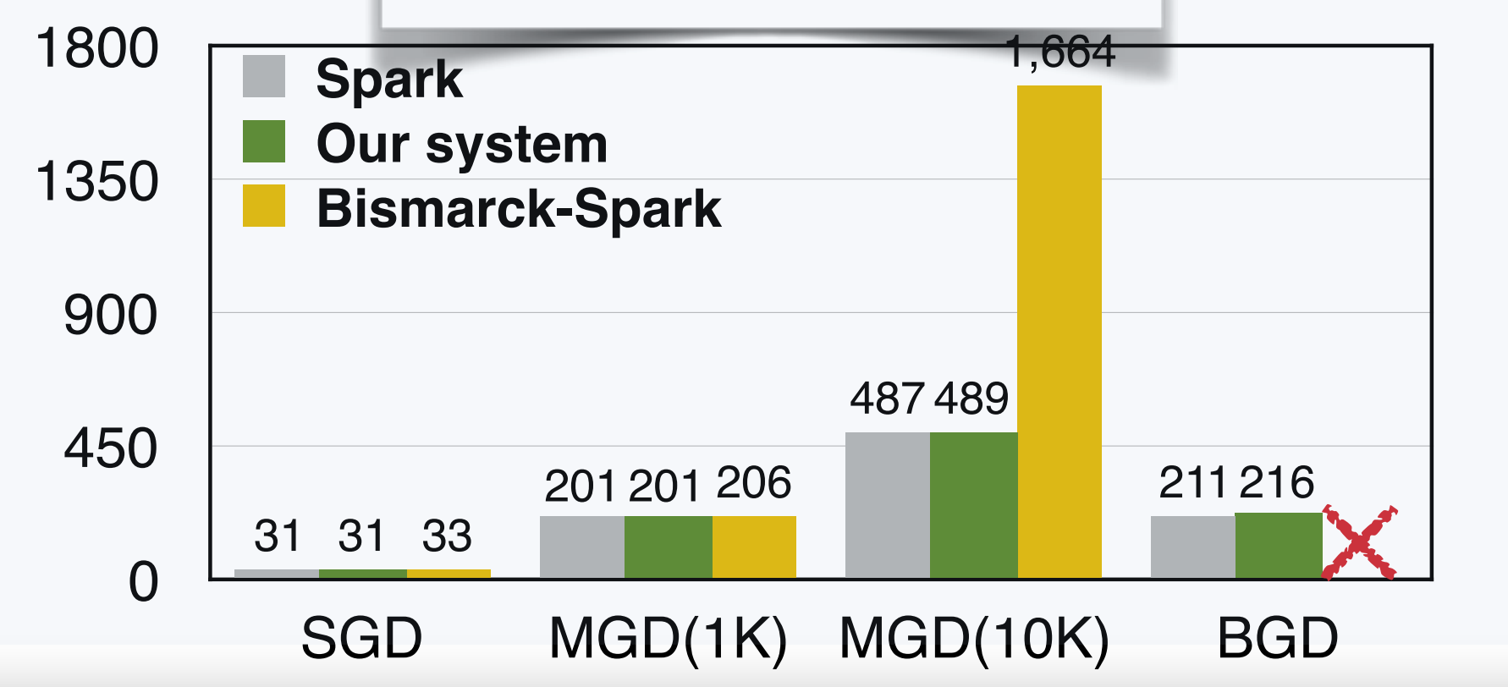


GD Plan choice



MGD

Abstraction benefit



A Cost-based Optimizer for Gradient Descent Optimization

Zoi Kaoudi Jorge Quianè-Ruiz Saravanan Thirumuruganathan Sanjay Chawla Divy Agrawal

UCSB
QCRI
Qatar Computing Research Institute