# DataVinci: Proactive Data Placement for Ad-Hoc Computing

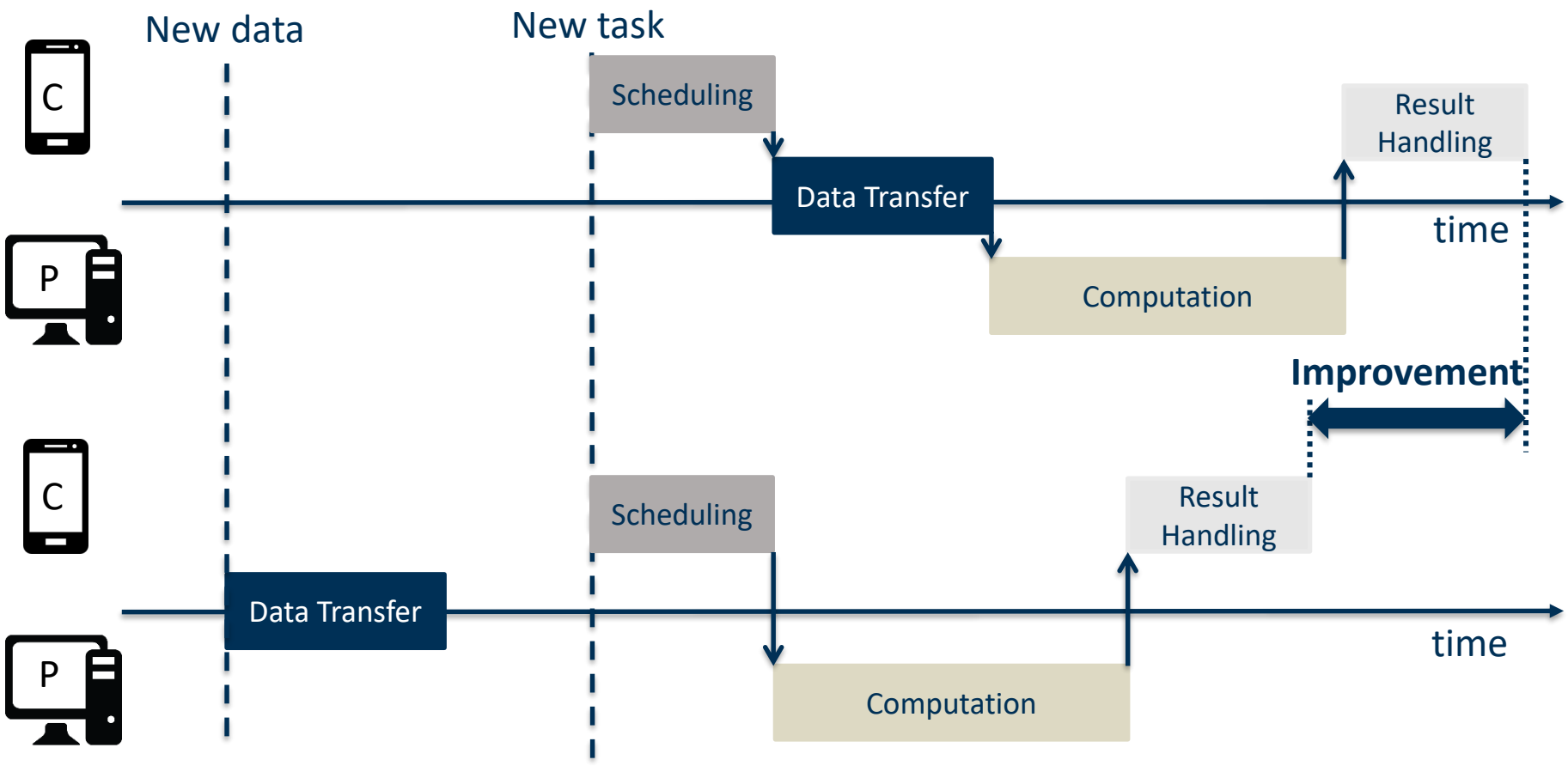
Martin Breitbach, Janick Edinger, Dominik Schäfer, and Christian Becker

UNIVERSITY OF MANNHEIM
Business School

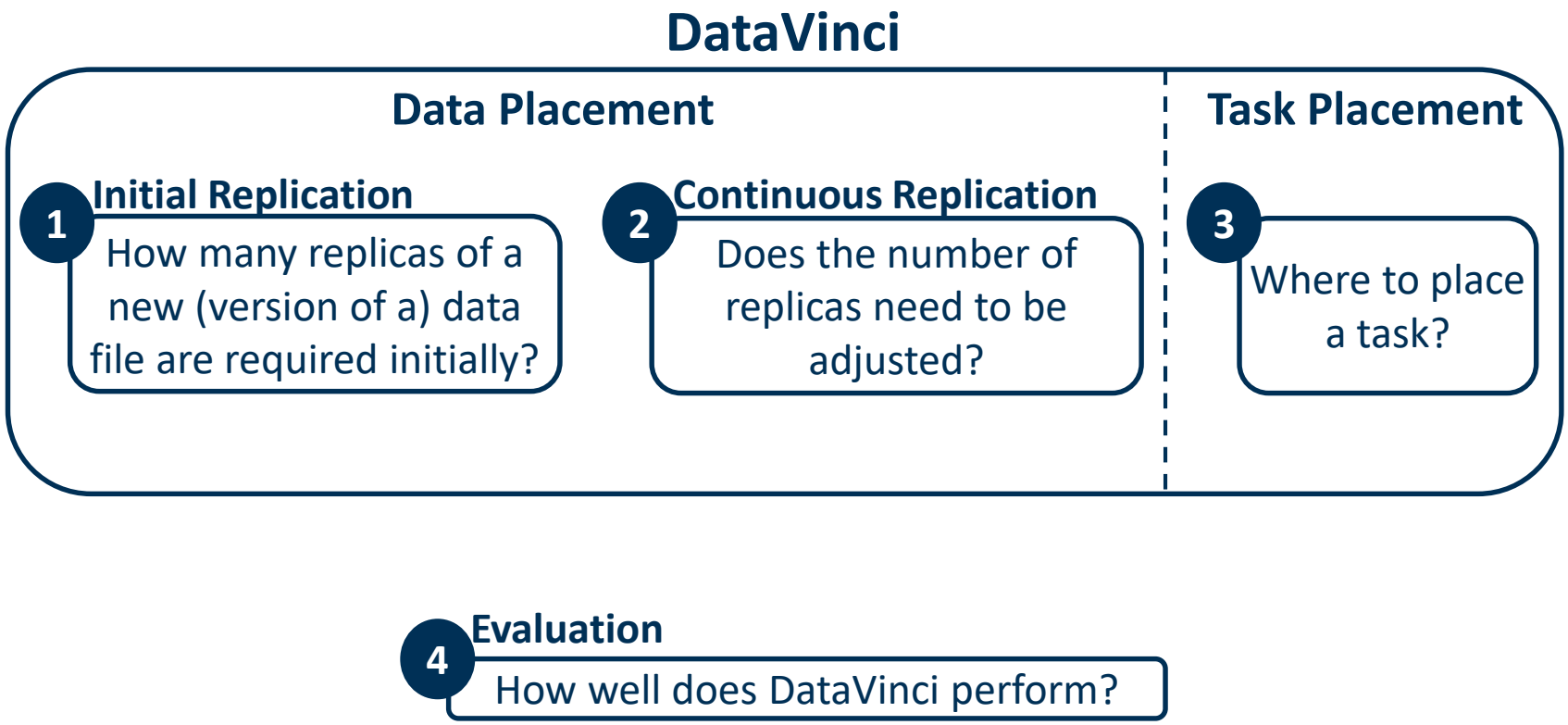
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

May 21, 2021

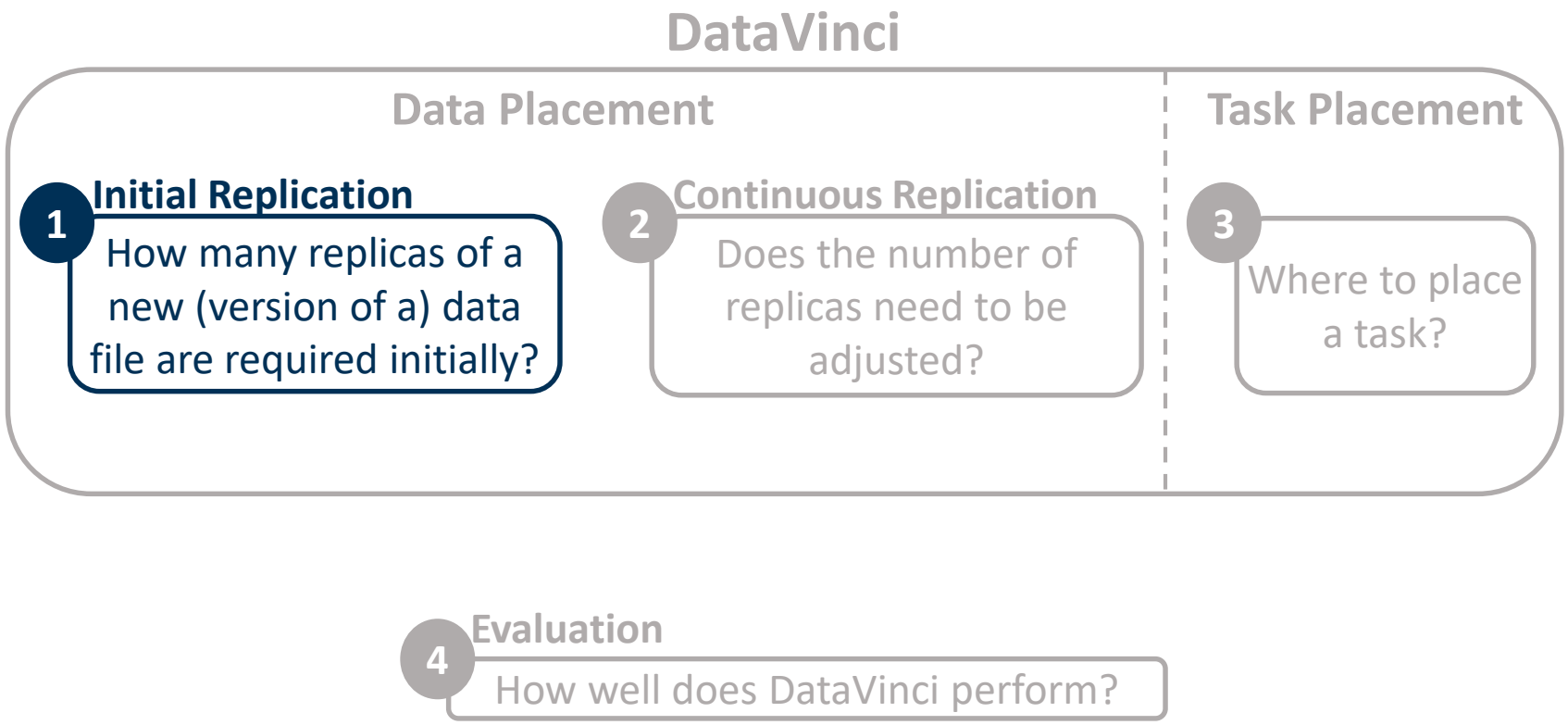
# Motivation


New data
New task
C
Scheduling
Result Handling
Data Transfer
time
P
Computation
Improvement
C
Scheduling
Result Handling
Data Transfer
time
P
Computation

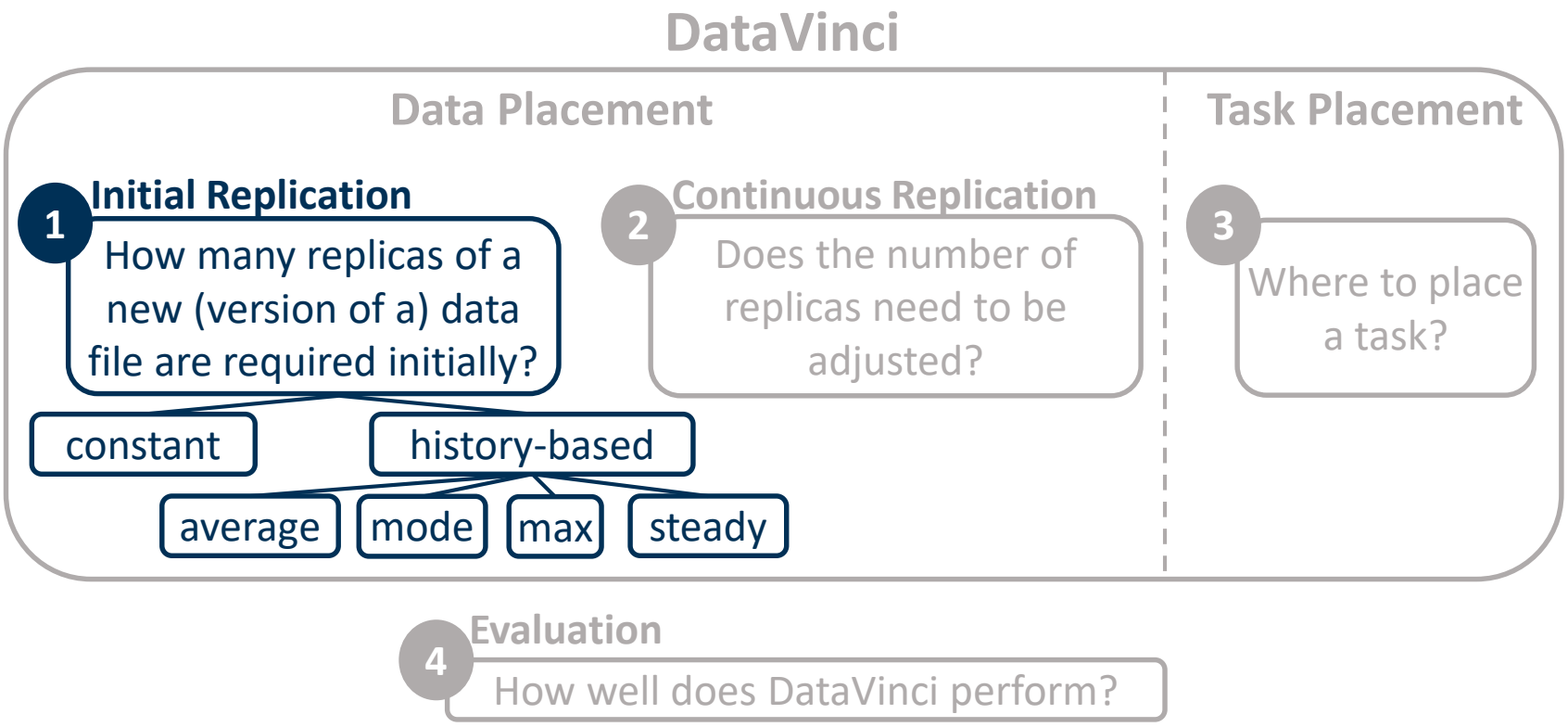
# Agenda


DataVinci
Data Placement
Task Placement
1
Initial Replication
How many replicas of a new (version of a) data file are required initially?
2
Continuous Replication
Does the number of replicas need to be adjusted?
3
Where to place a task?
4
Evaluation
How well does DataVinci perform?

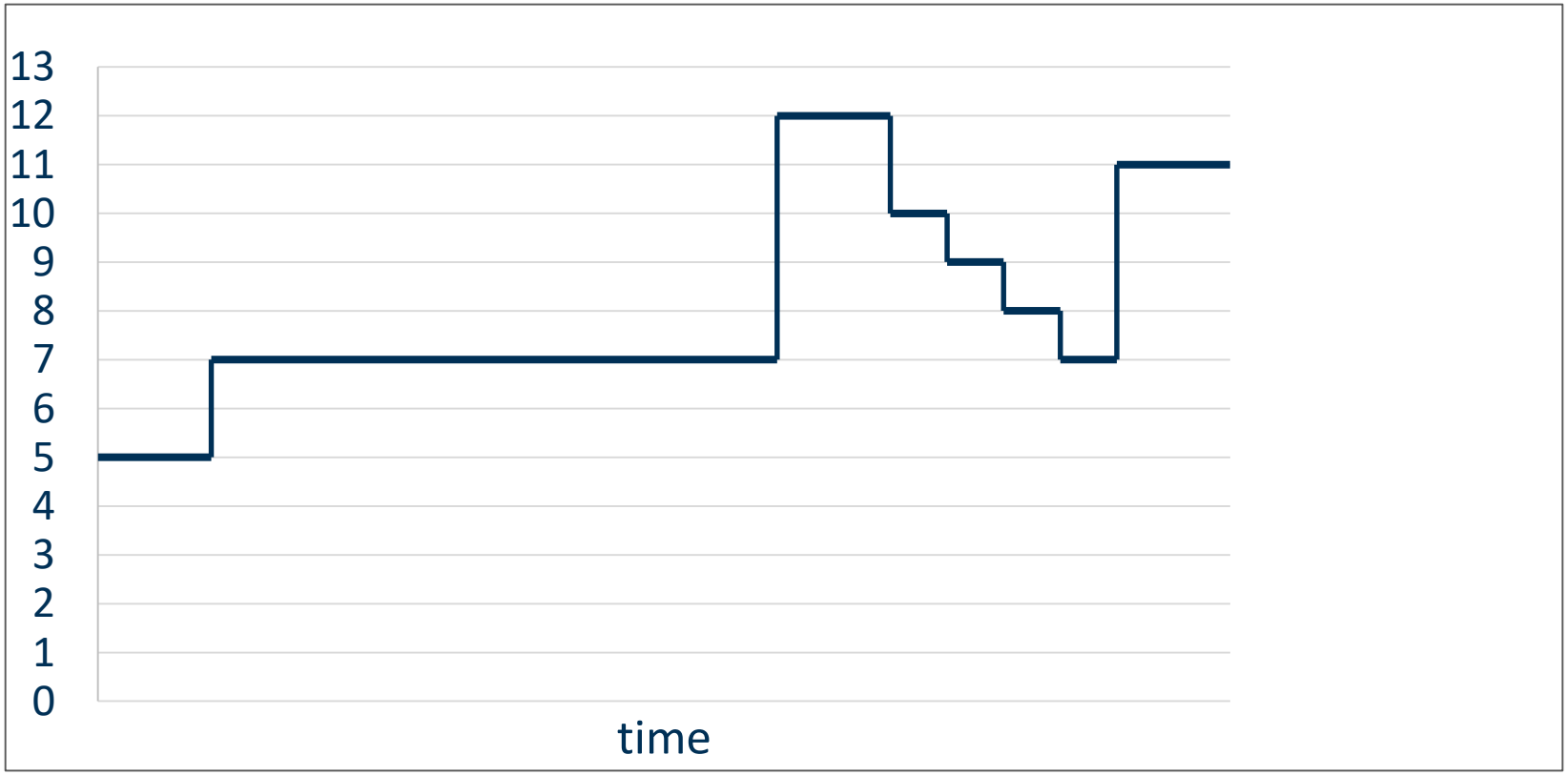
# Initial replication



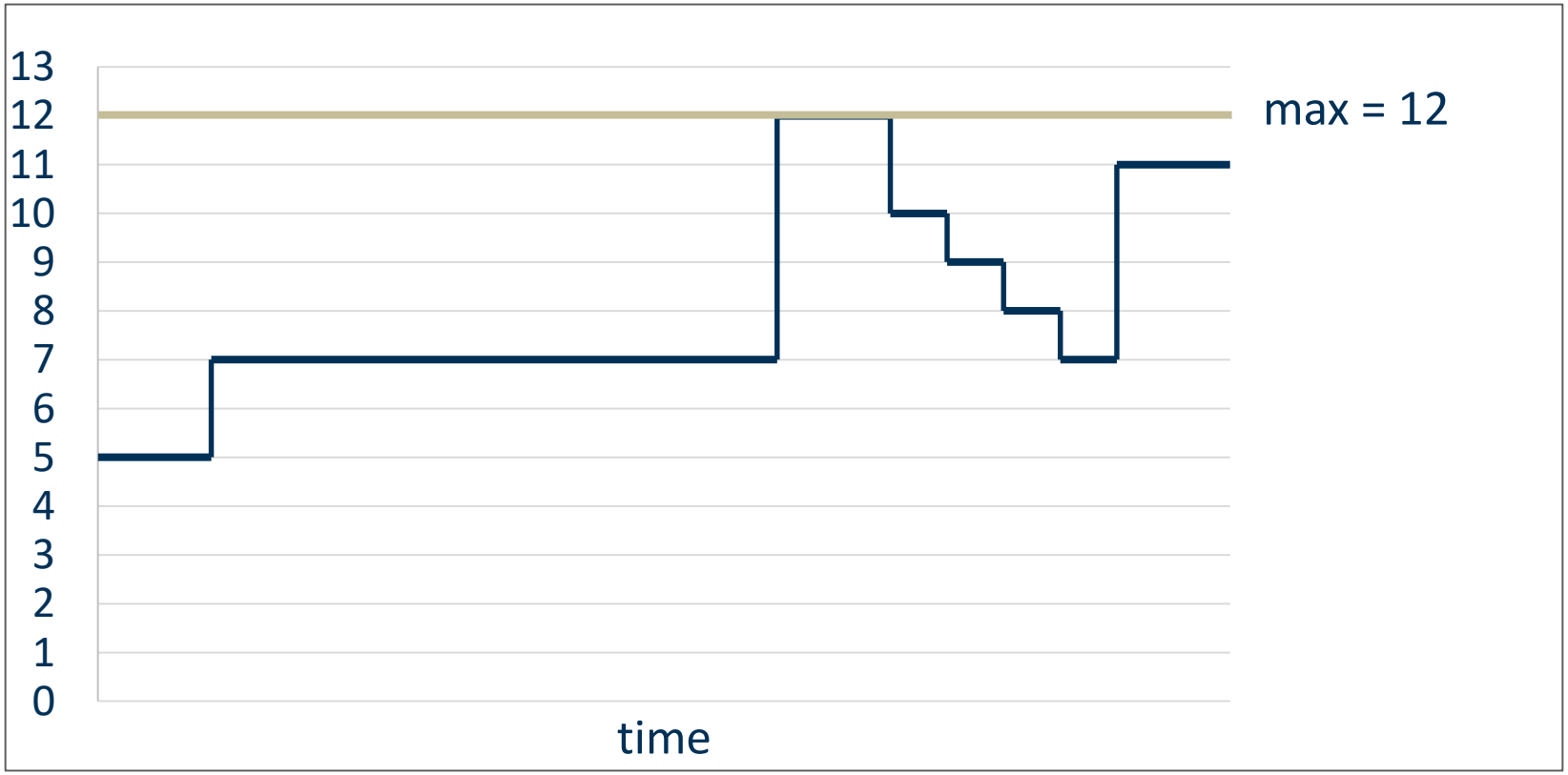
# Initial replication



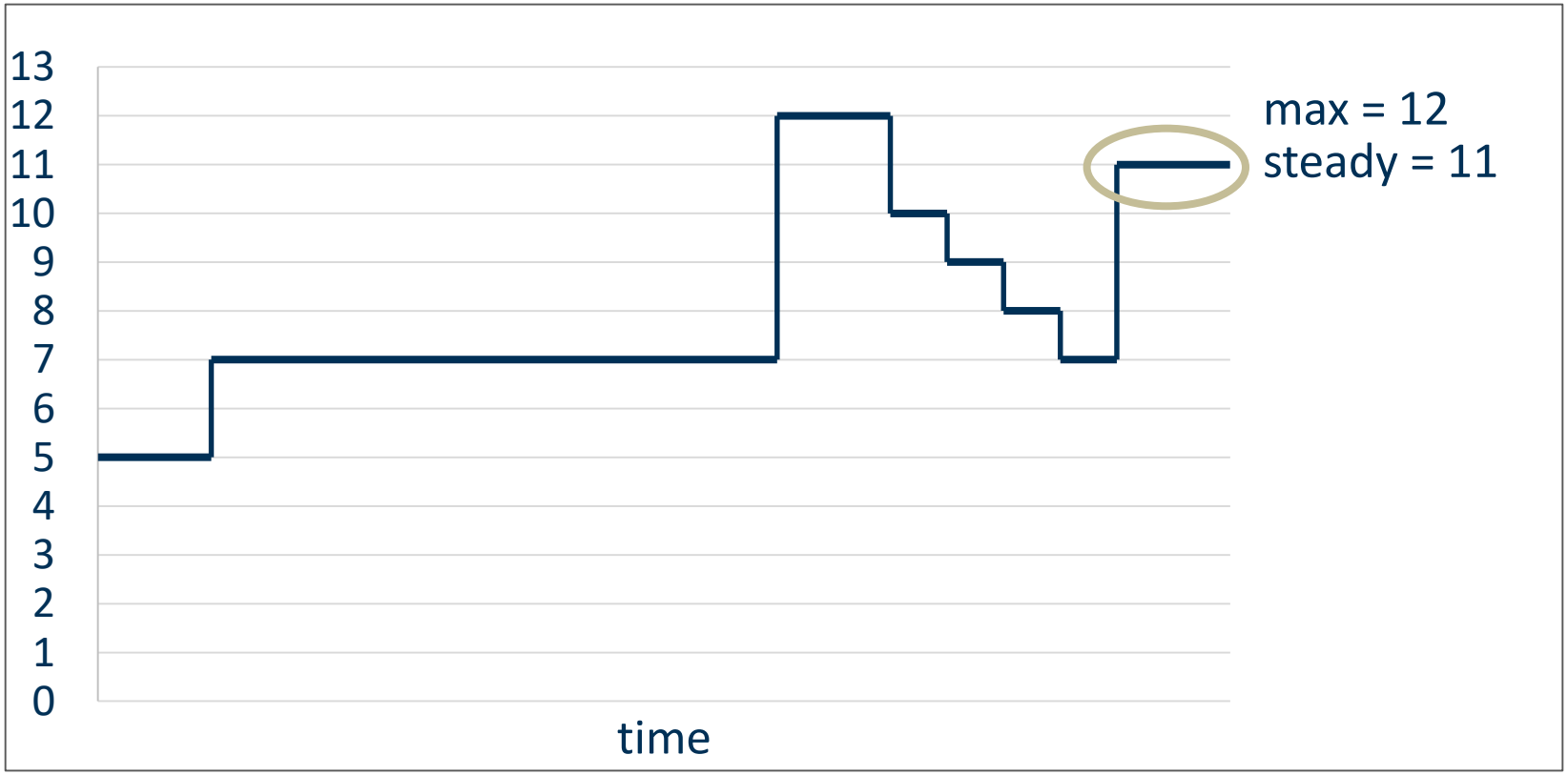
# Initial replication
# Constant vs. history-based replication



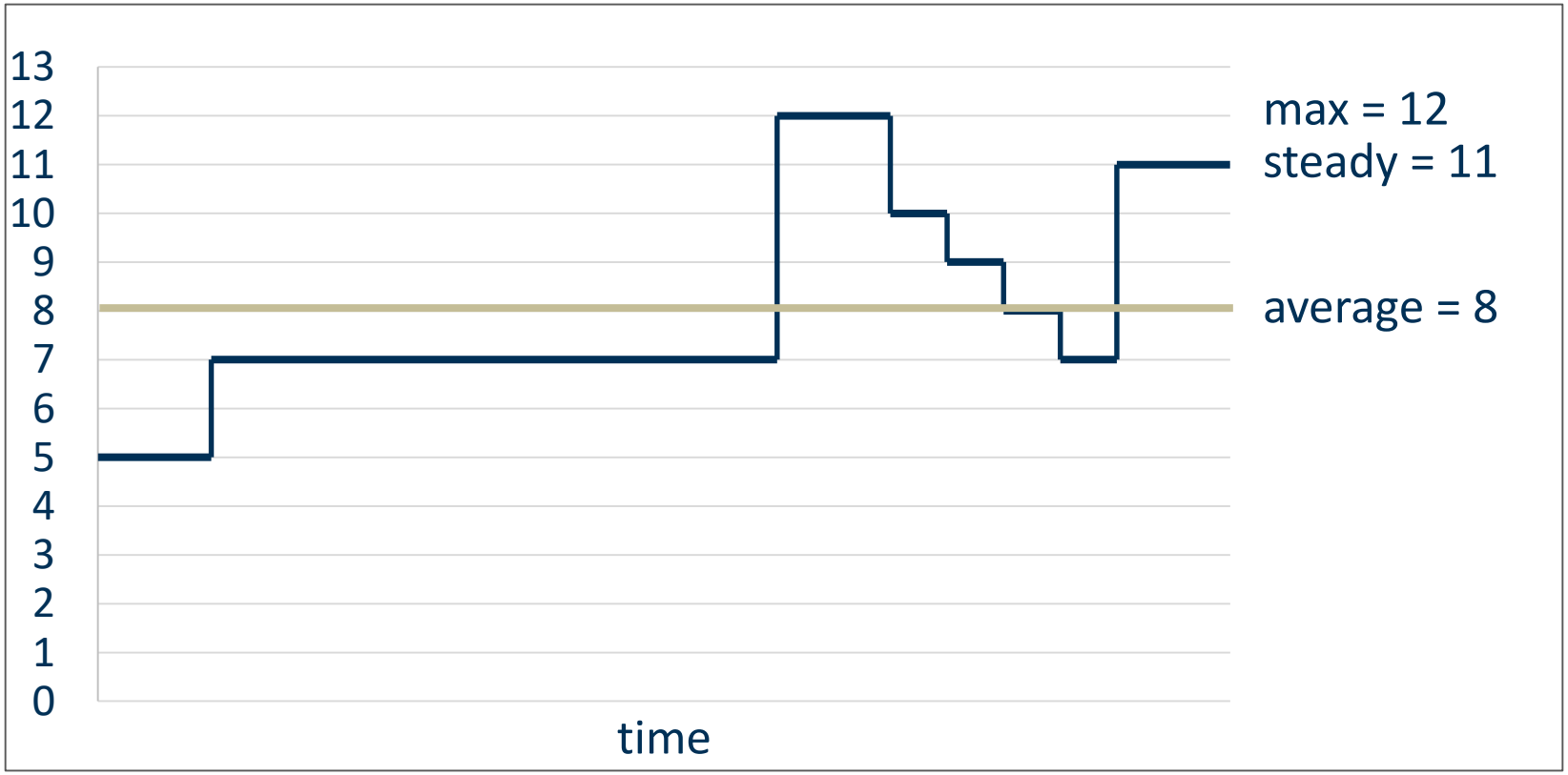
# Initial replication
# Constant vs. history-based replication



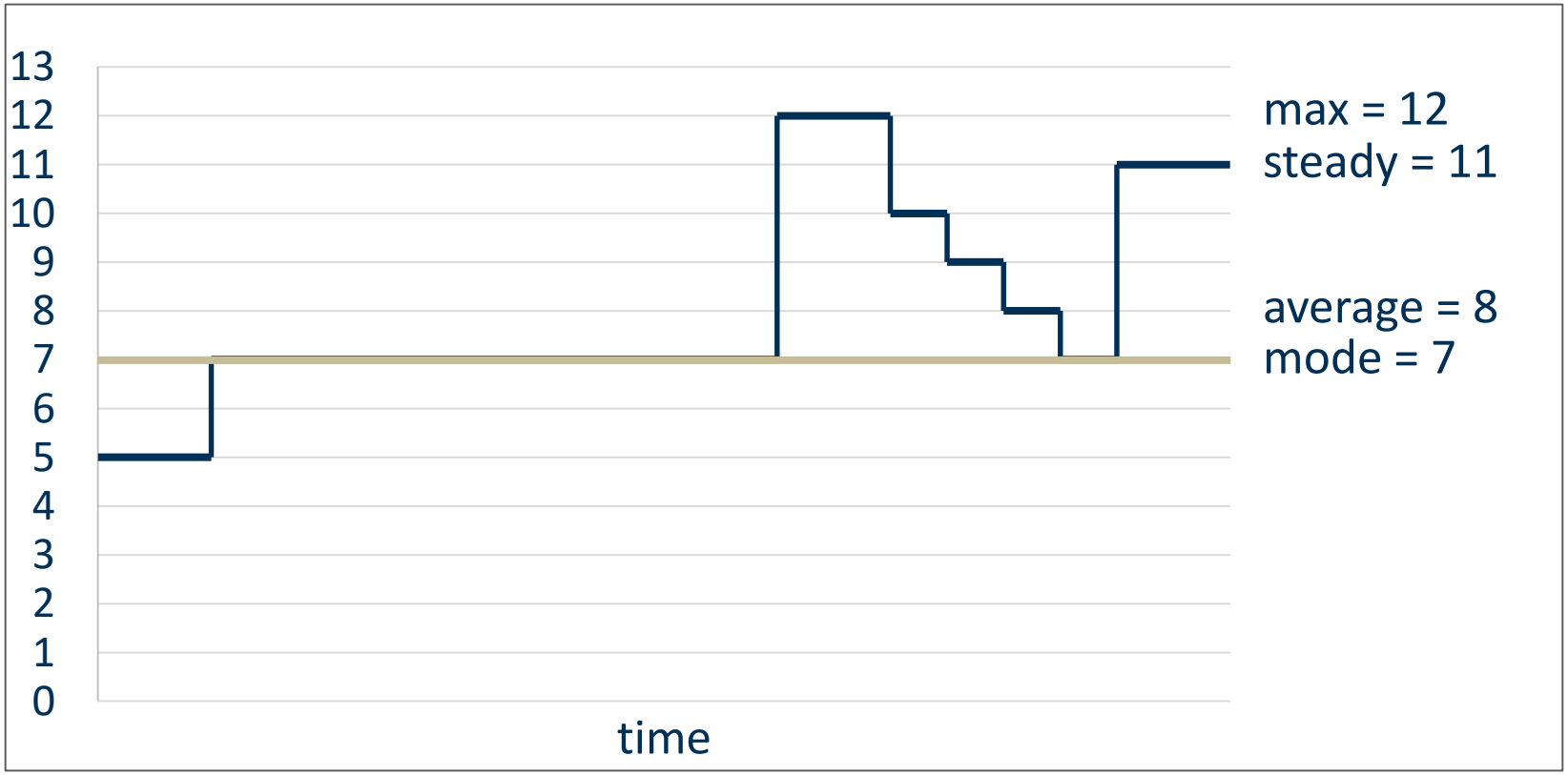
# Initial replication
# Constant vs. history-based replication



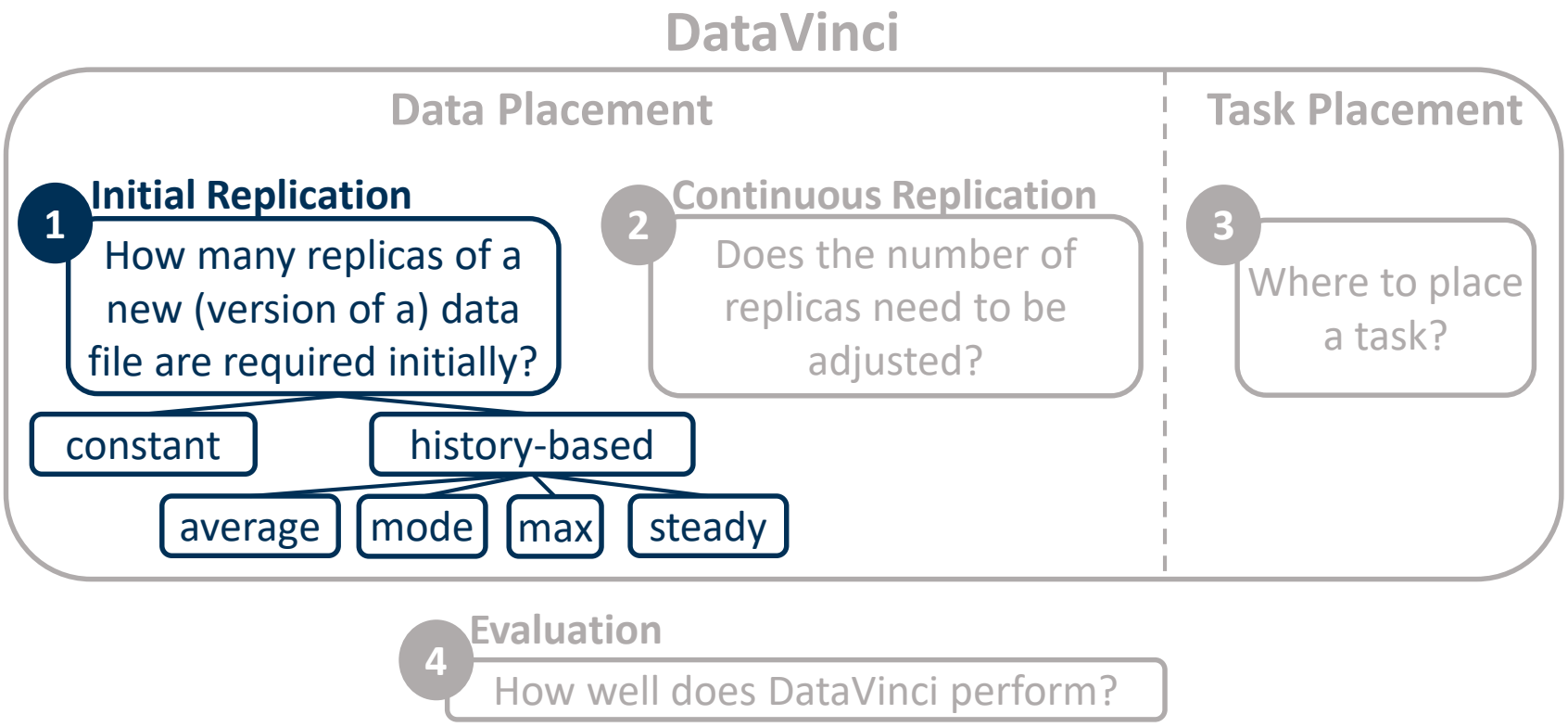
# Initial replication
# Constant vs. history-based replication



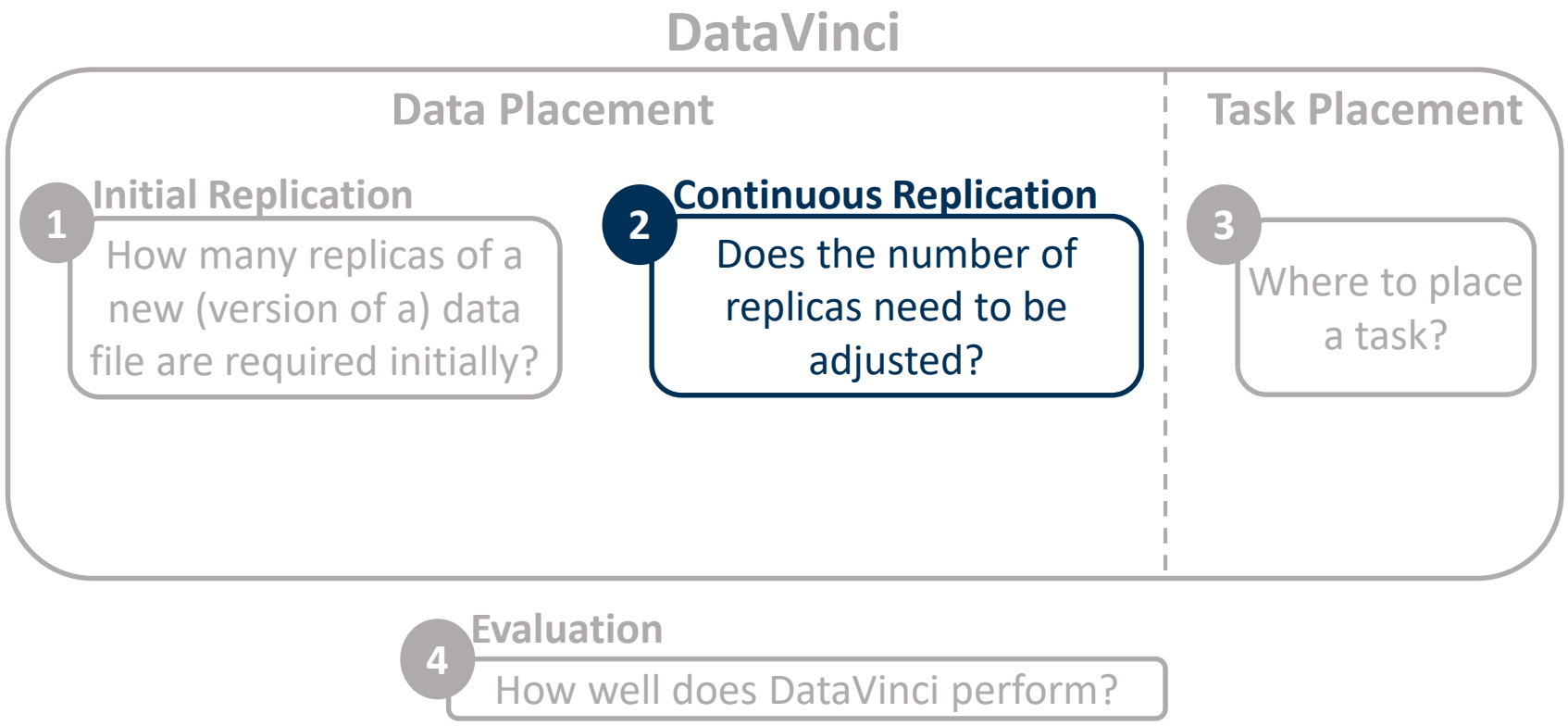
# Initial replication
# Constant vs. history-based replication



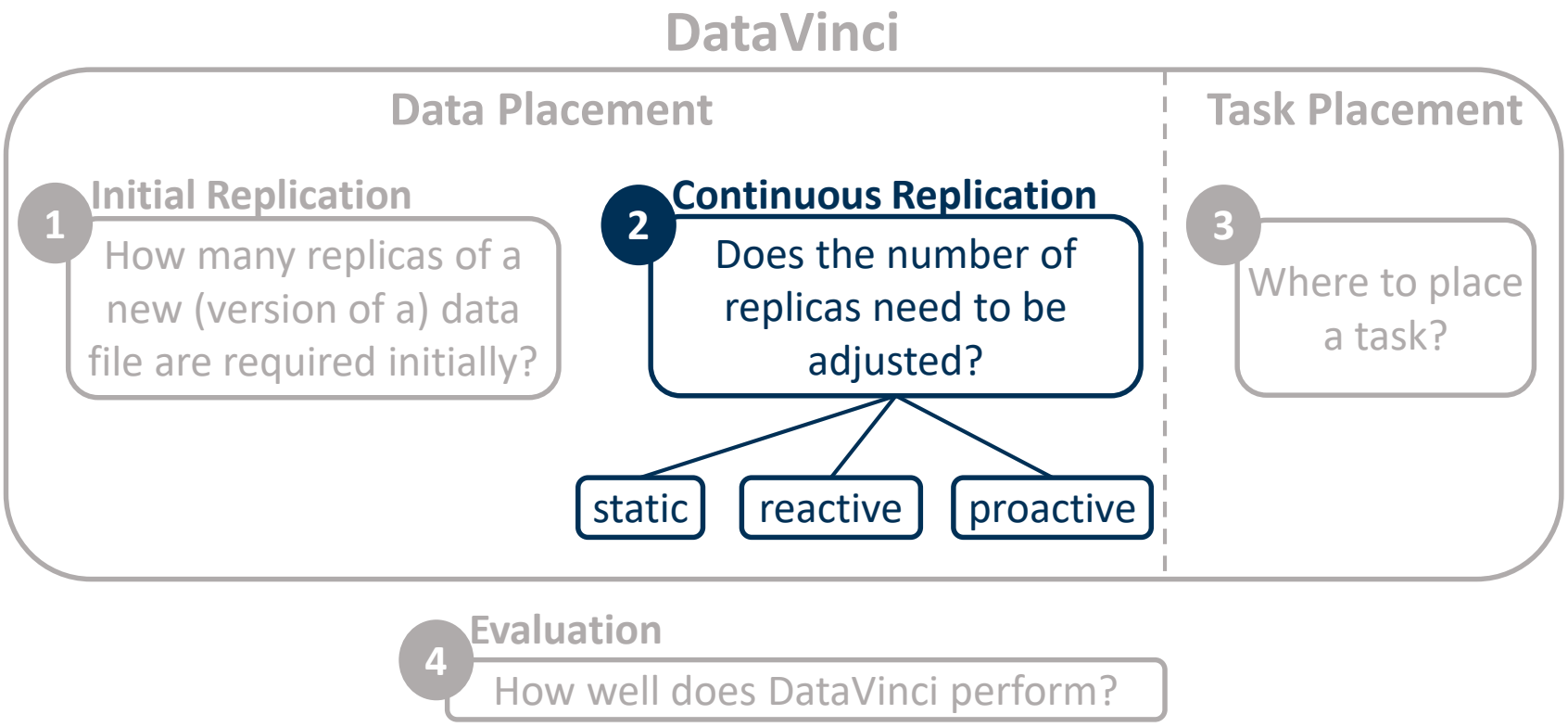
# Initial replication



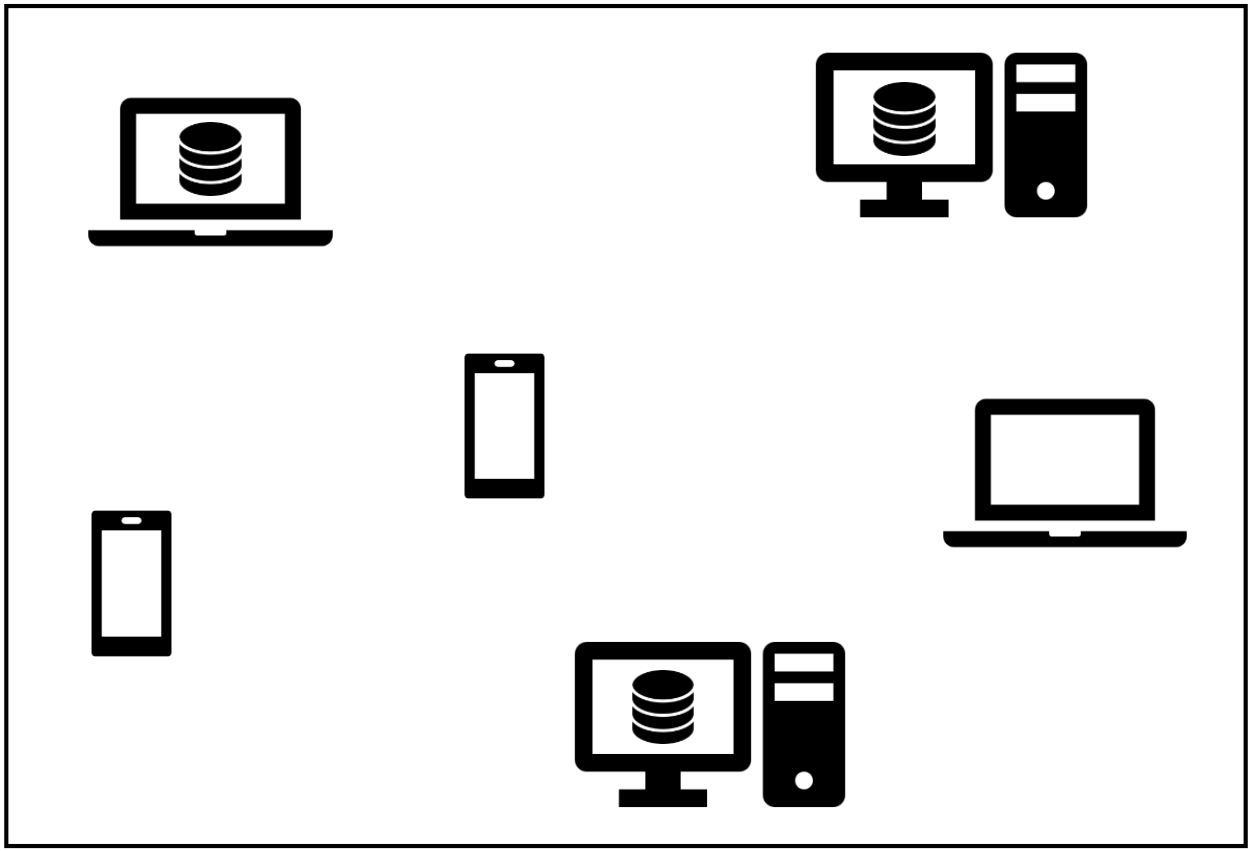
# Continuous replication



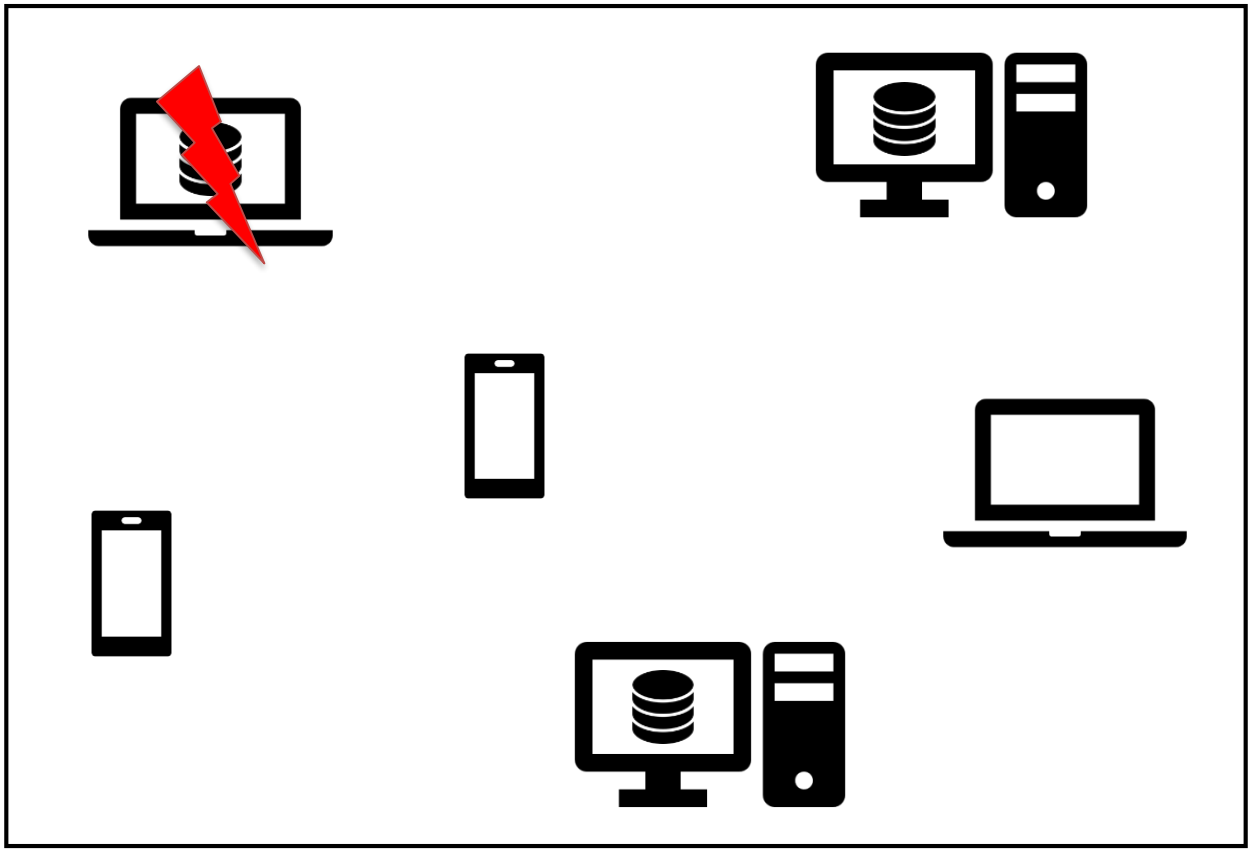
# Continuous replication



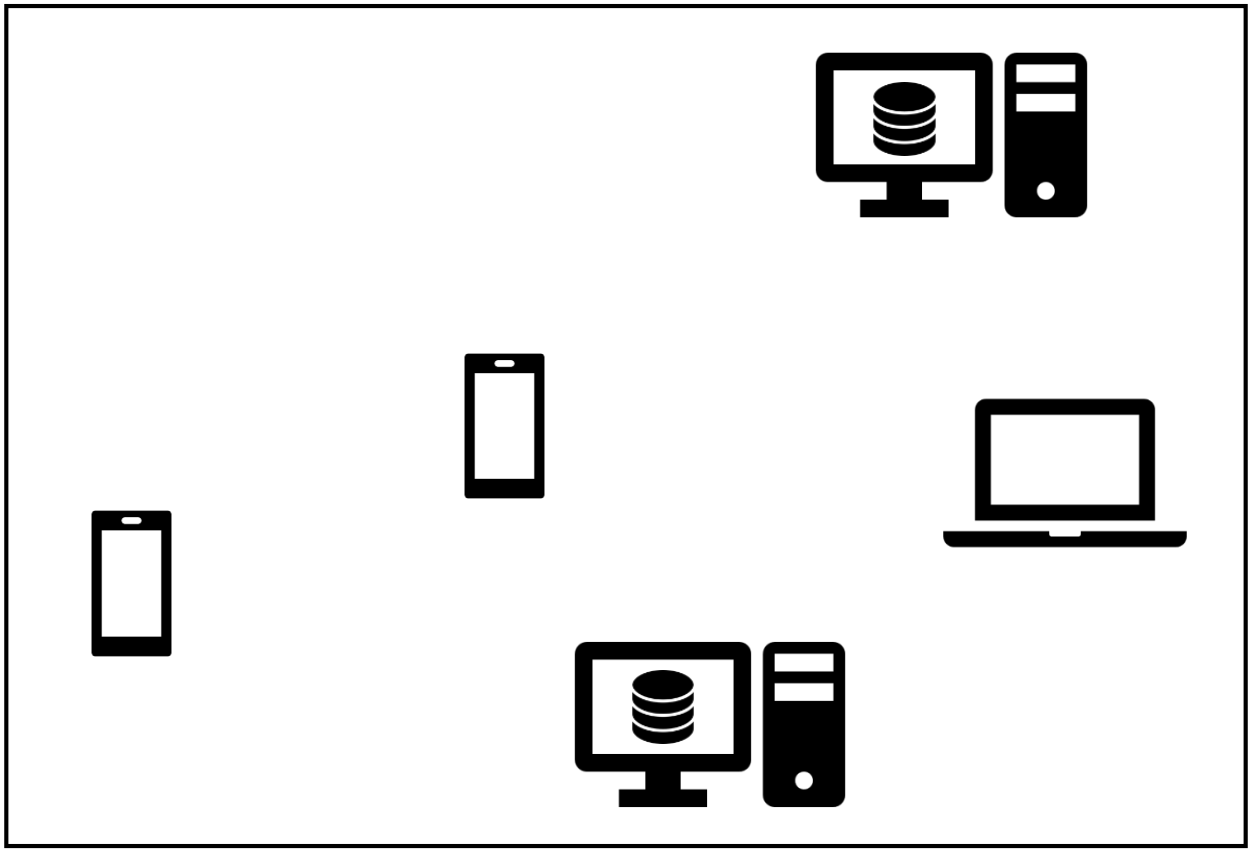
# Continuous replication
# Static



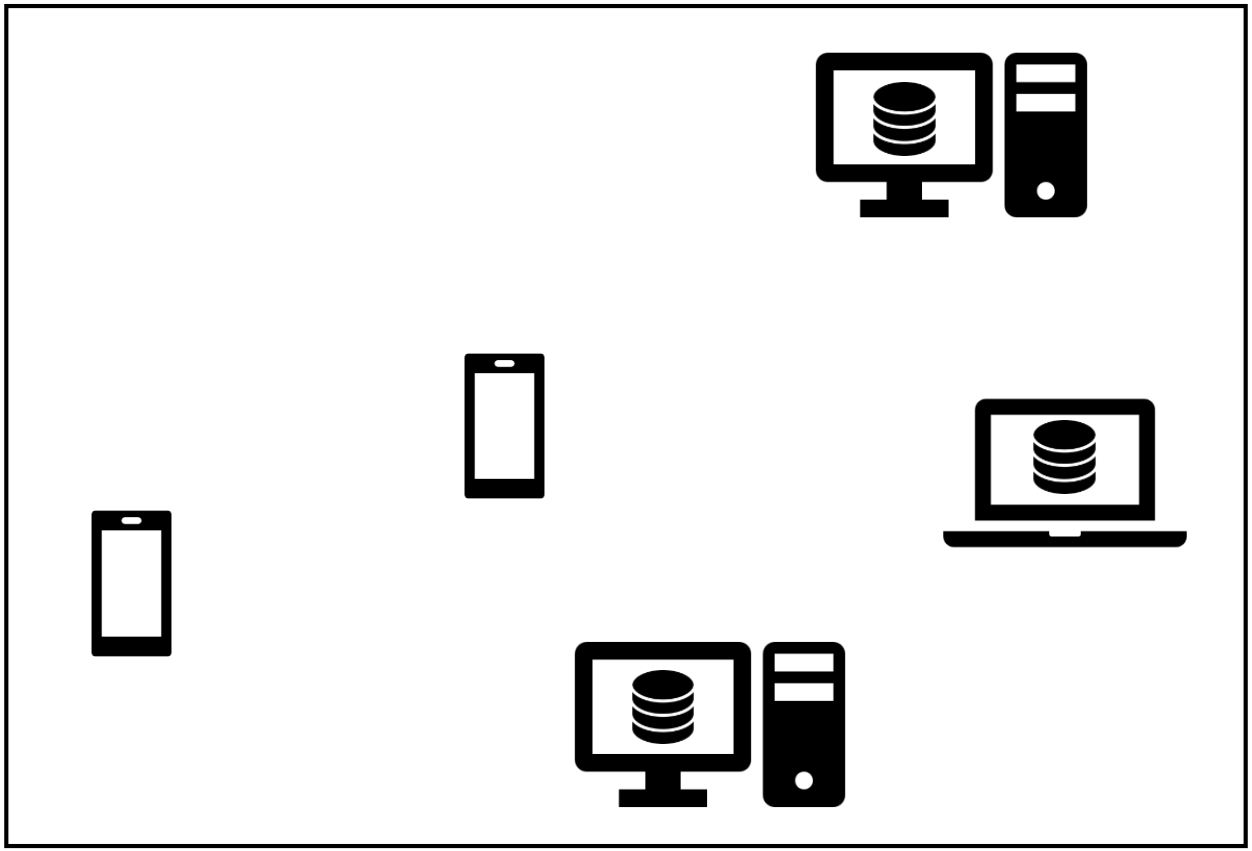
# Continuous replication
# Static



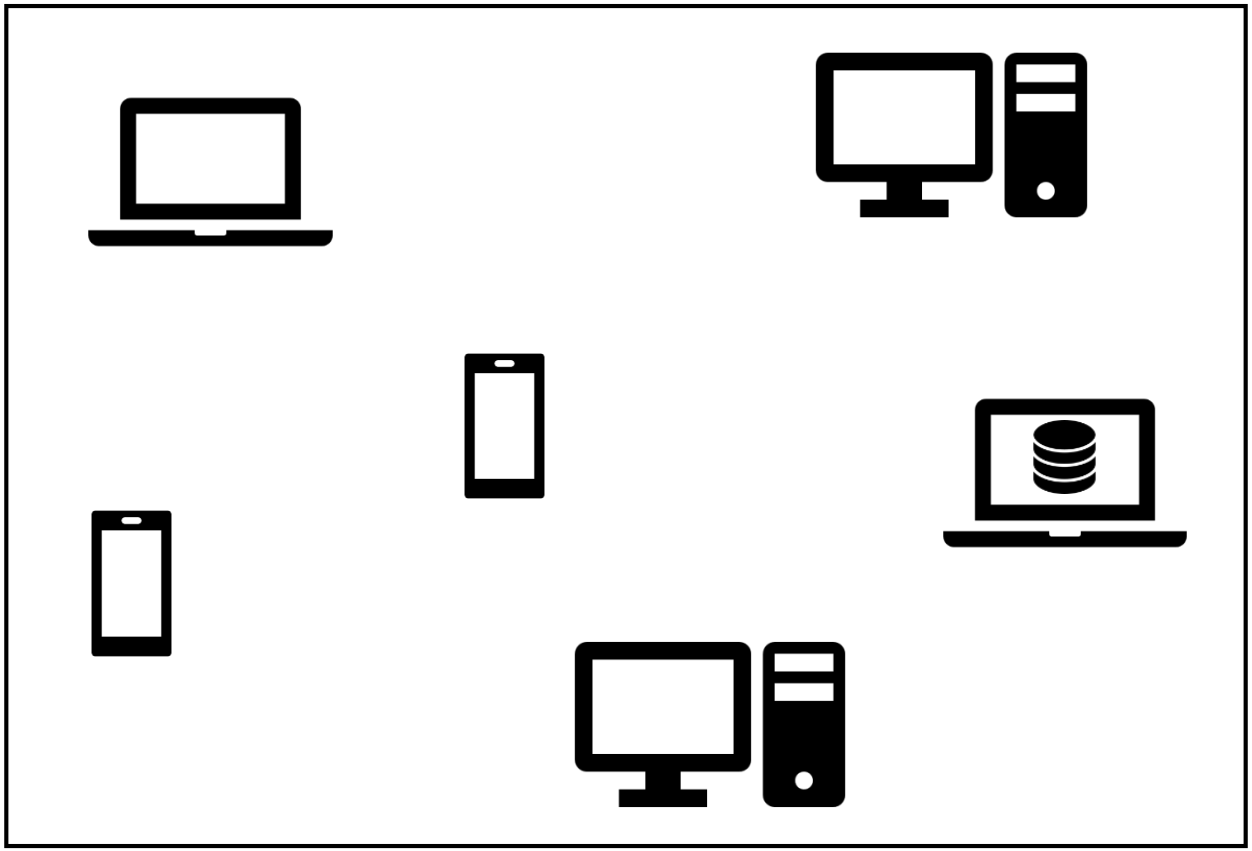
# Continuous replication
# Static



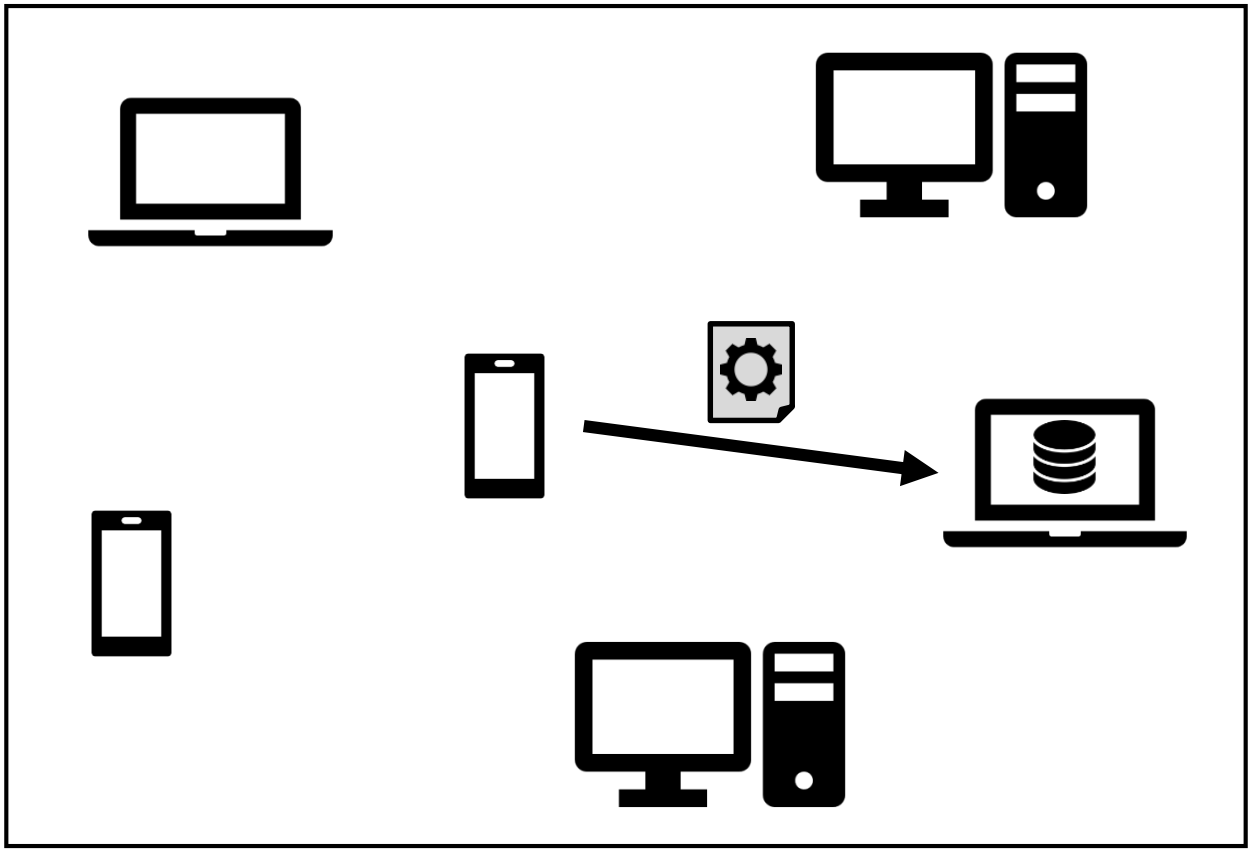
# Continuous replication
# Static



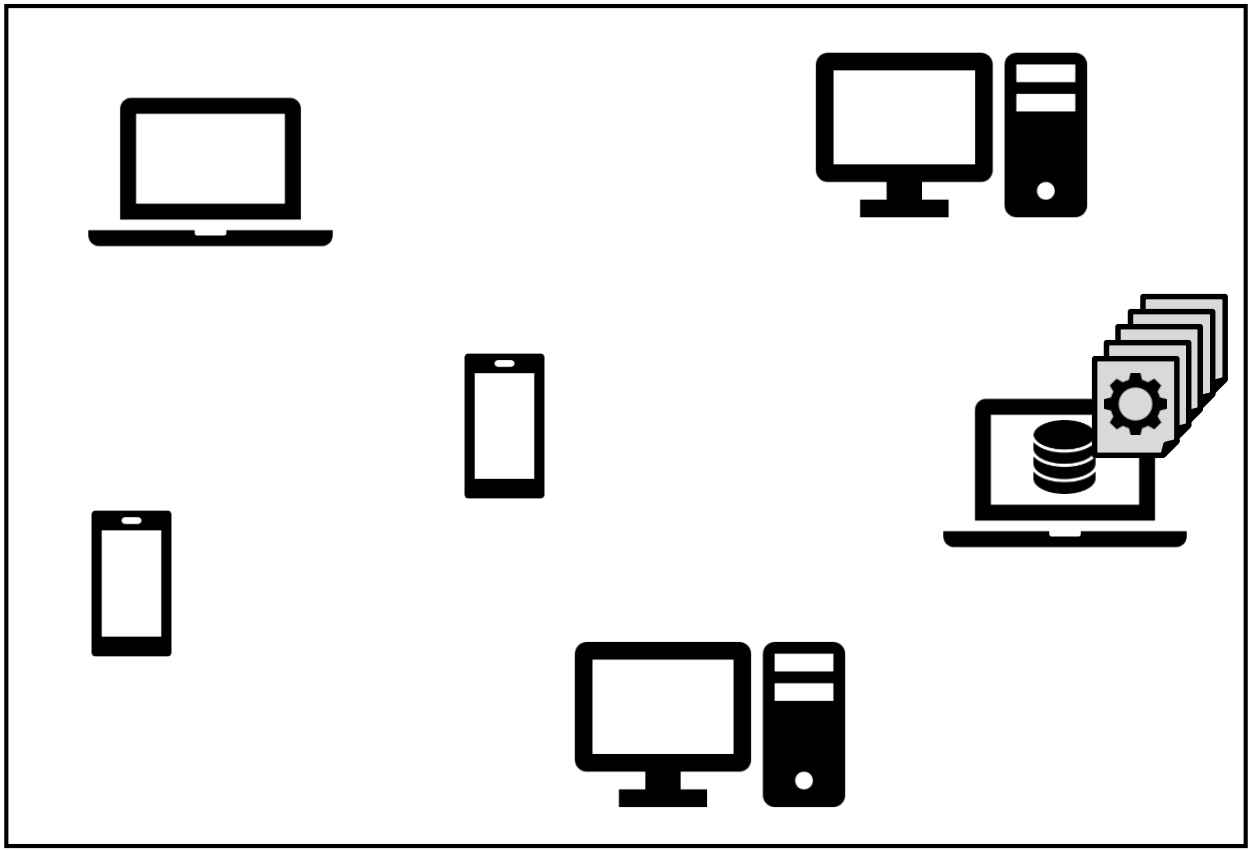
# Continuous replication
# Reactive



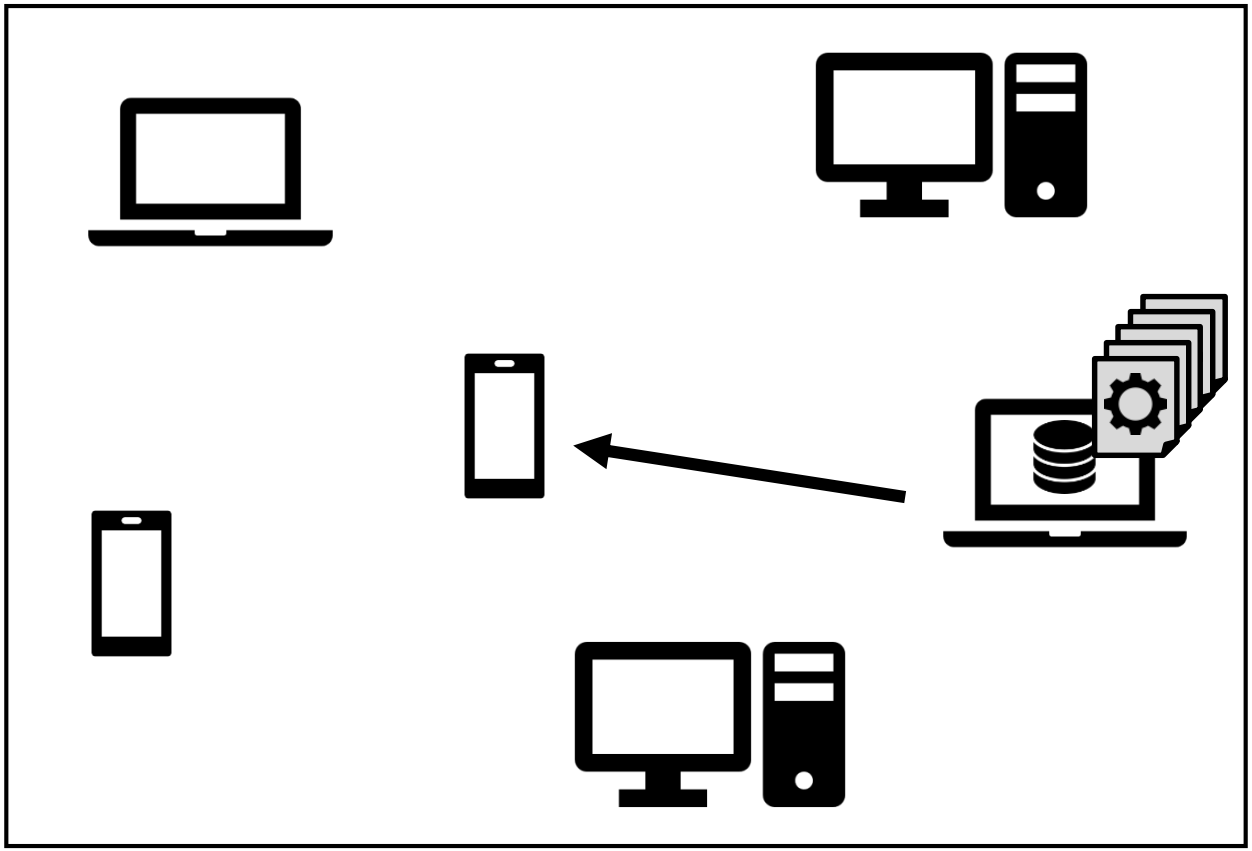
# Continuous replication
# Reactive



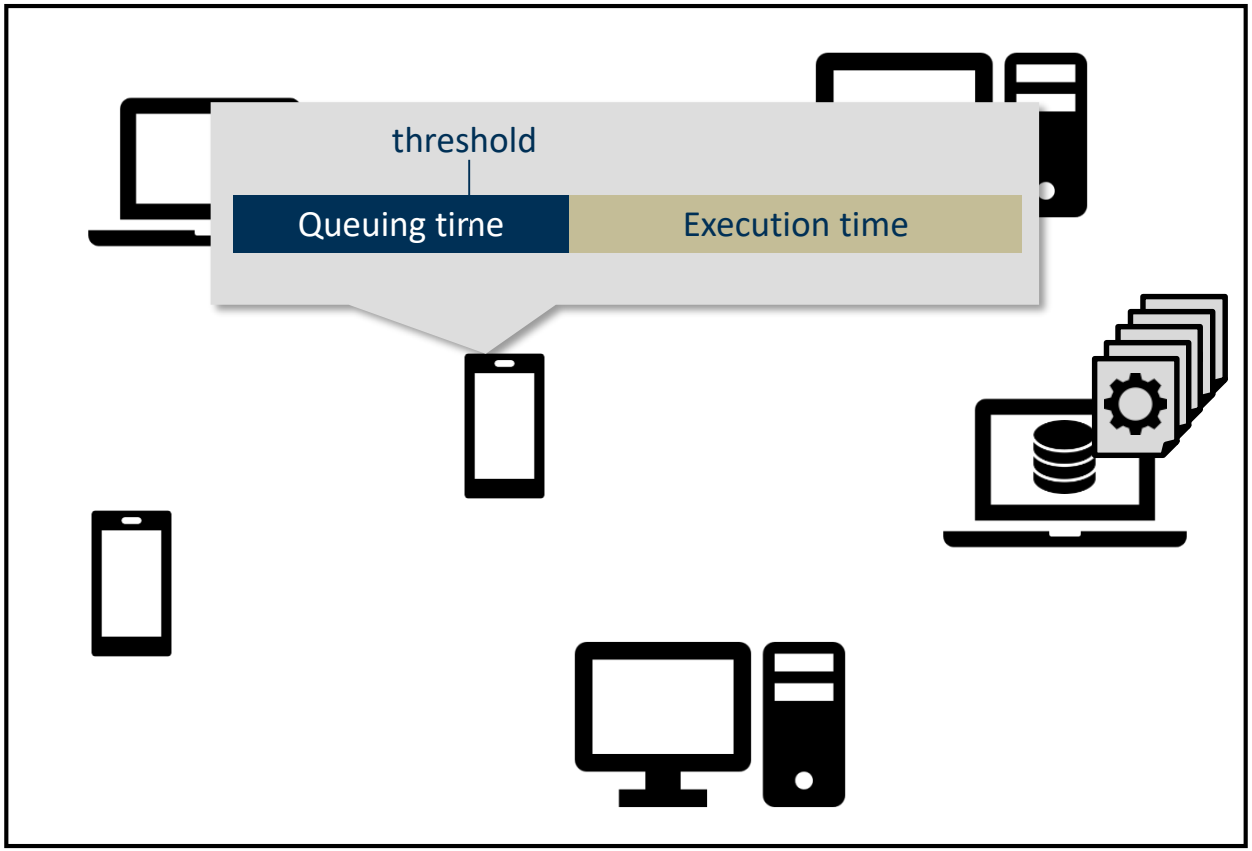
# Continuous replication
# Reactive



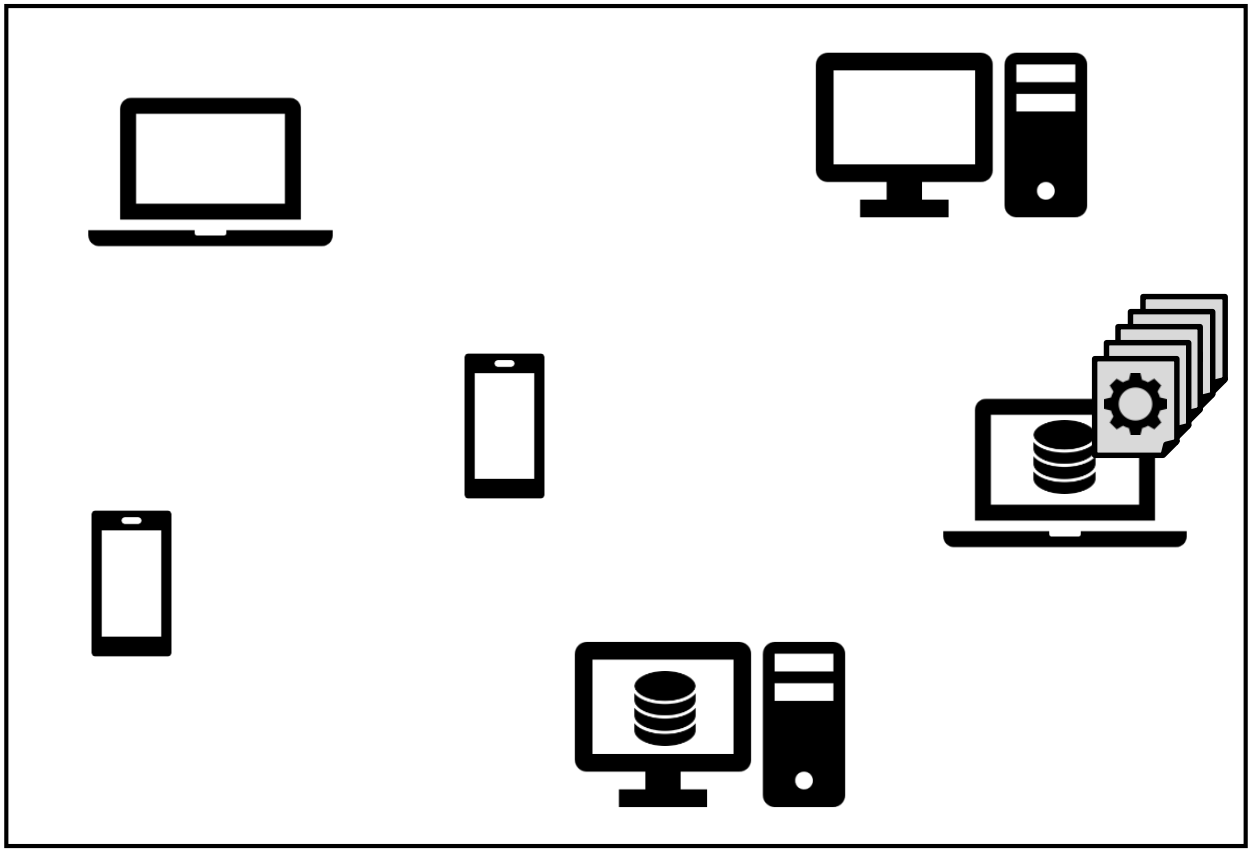
# Continuous replication
# Reactive



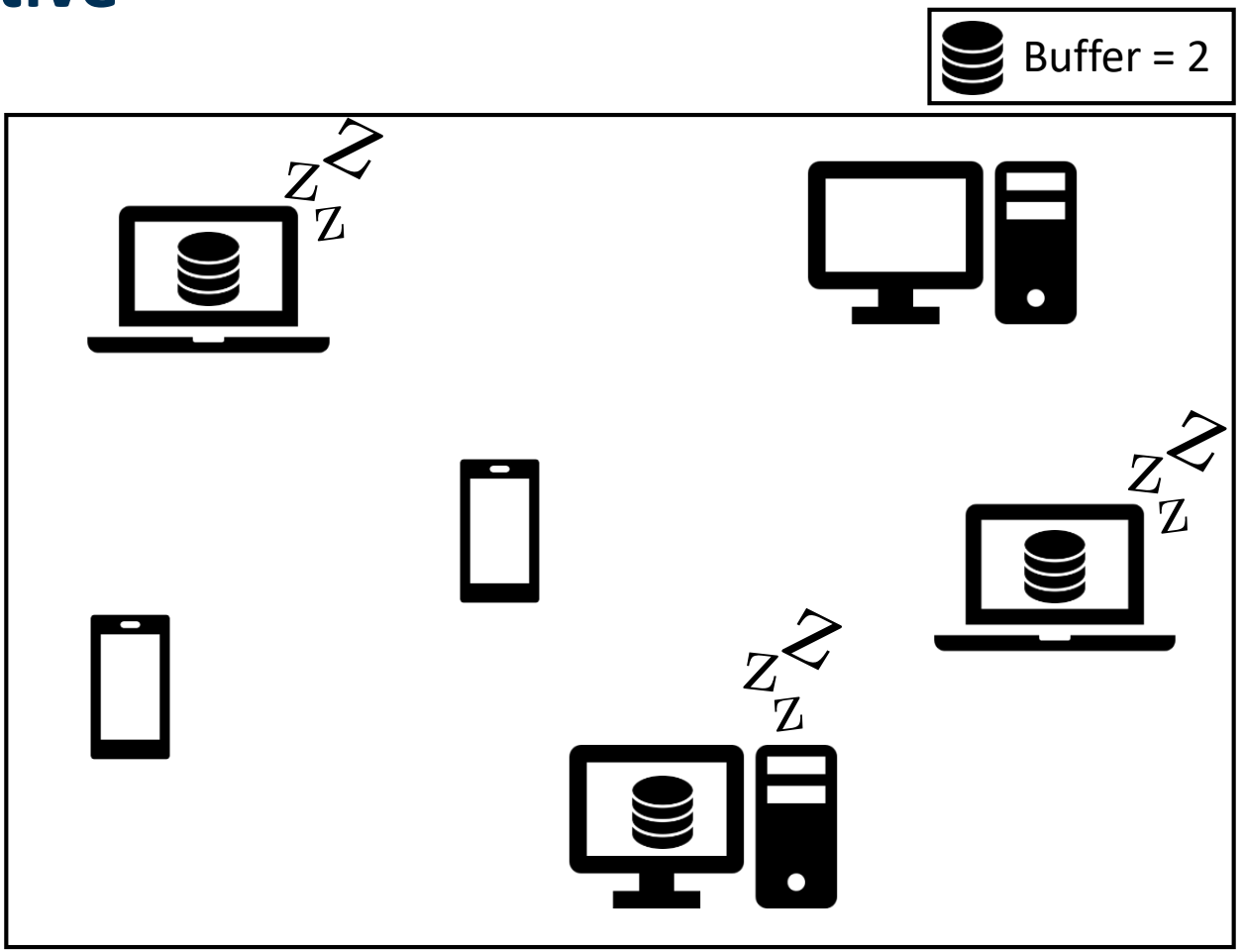
# Continuous replication
# Reactive



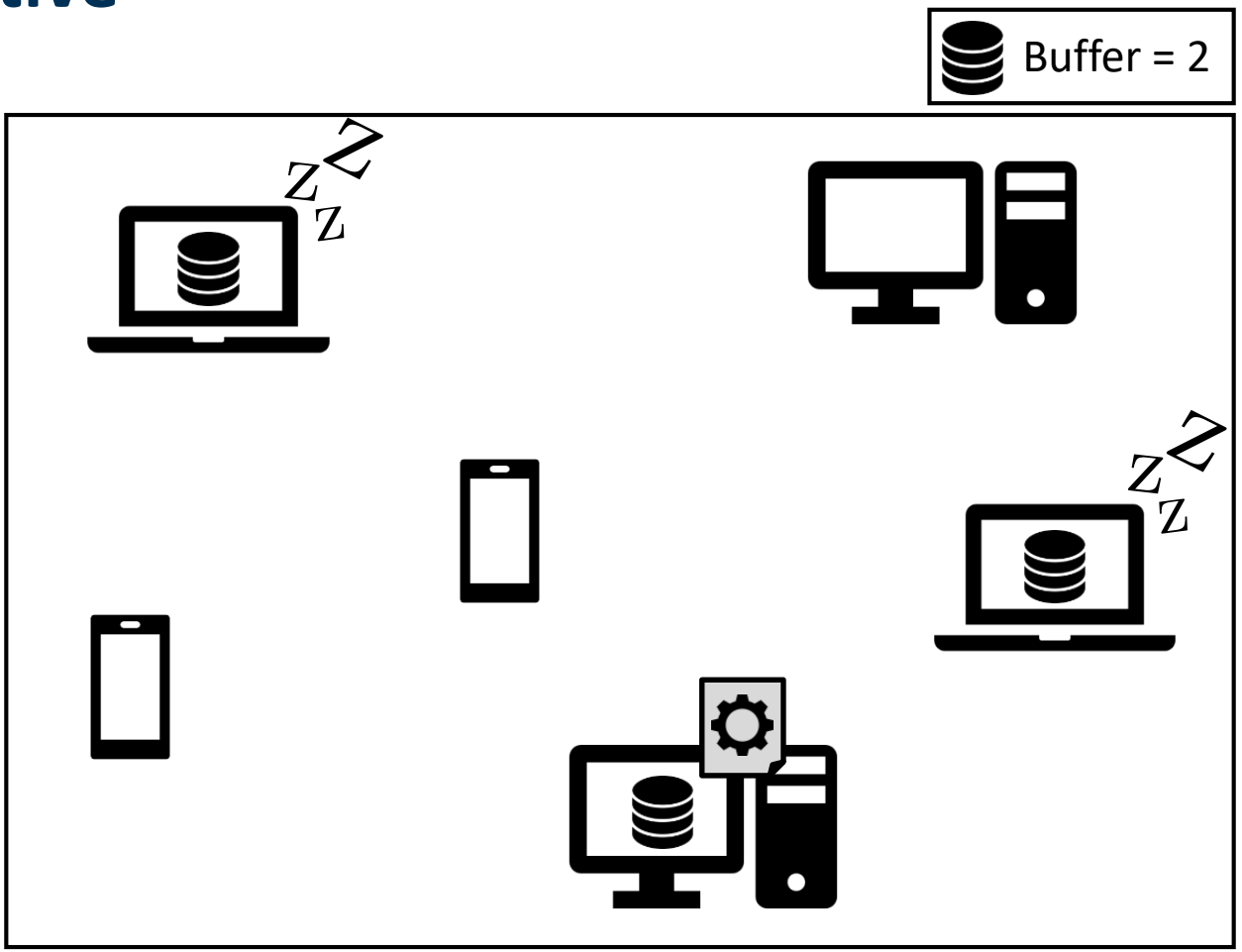
# Continuous replication
# Reactive



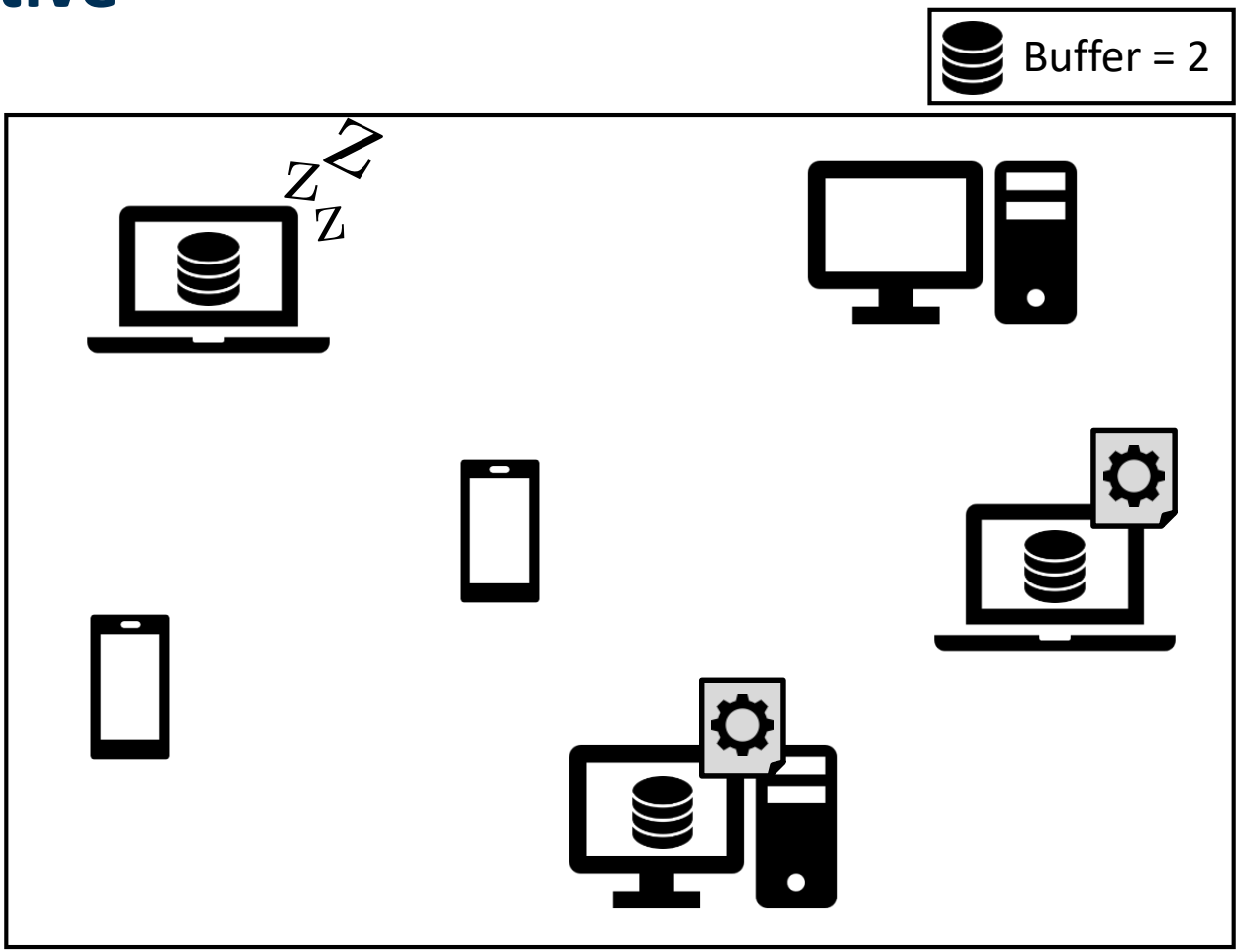
# Continuous replication
# Proactive



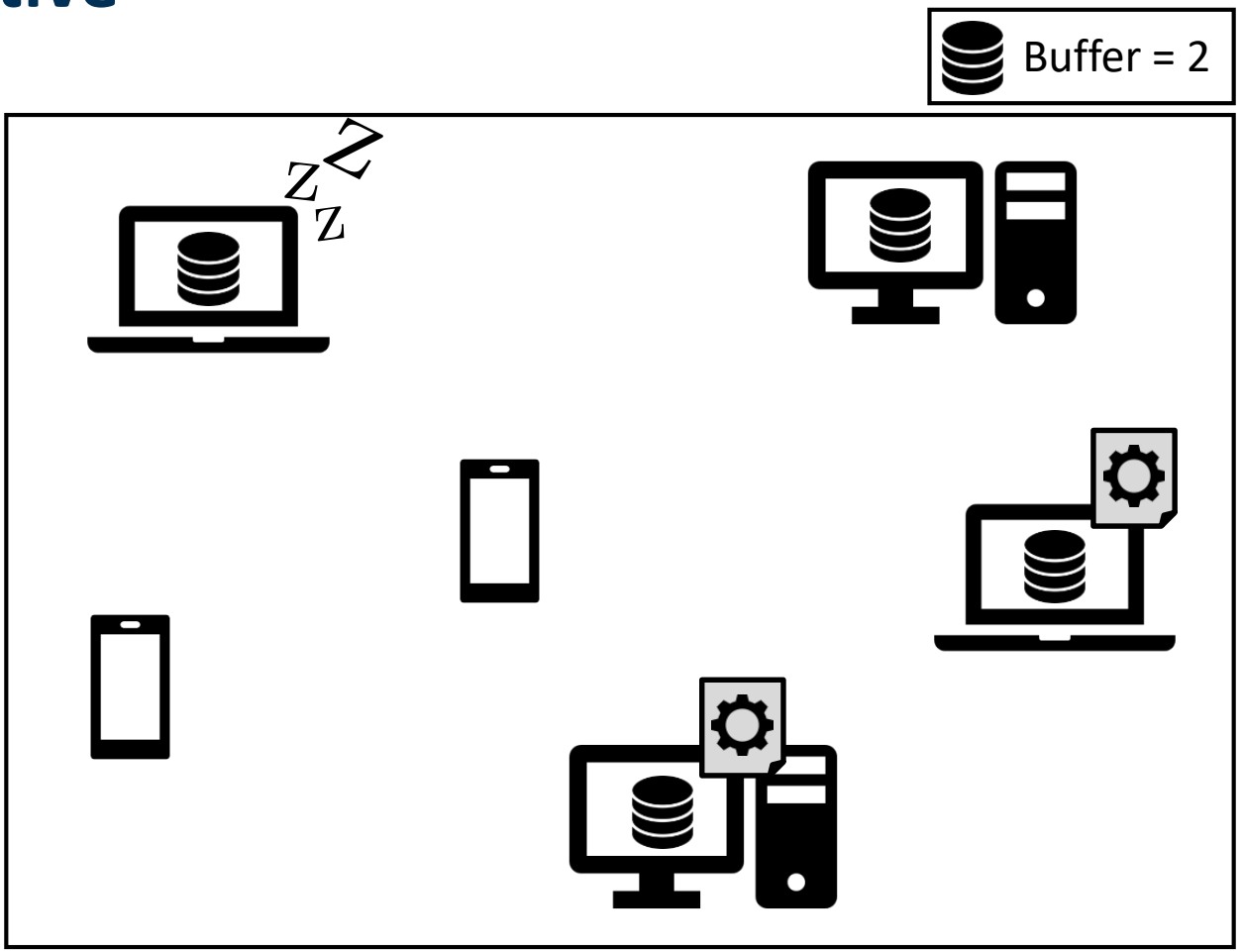
# Continuous replication
# Proactive



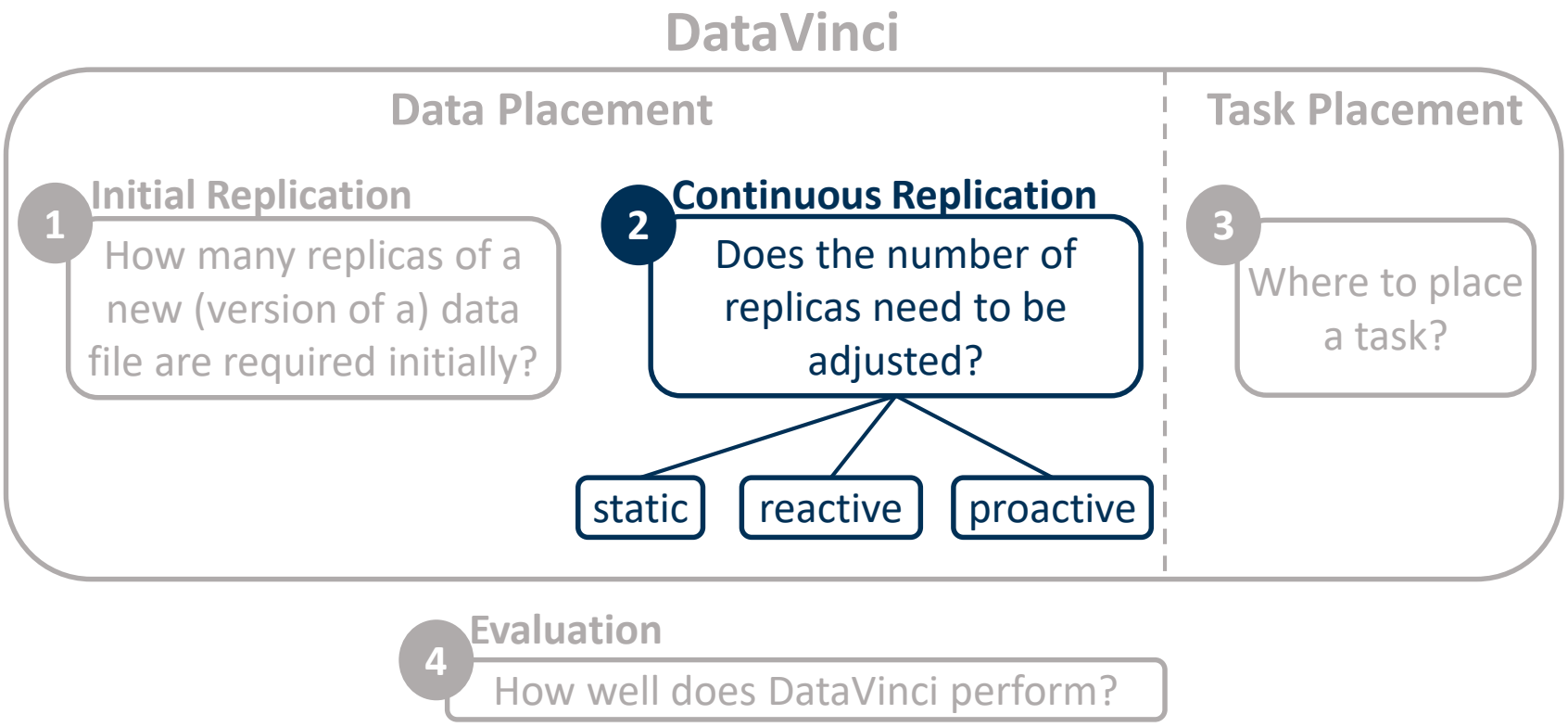
# Continuous replication
# Proactive



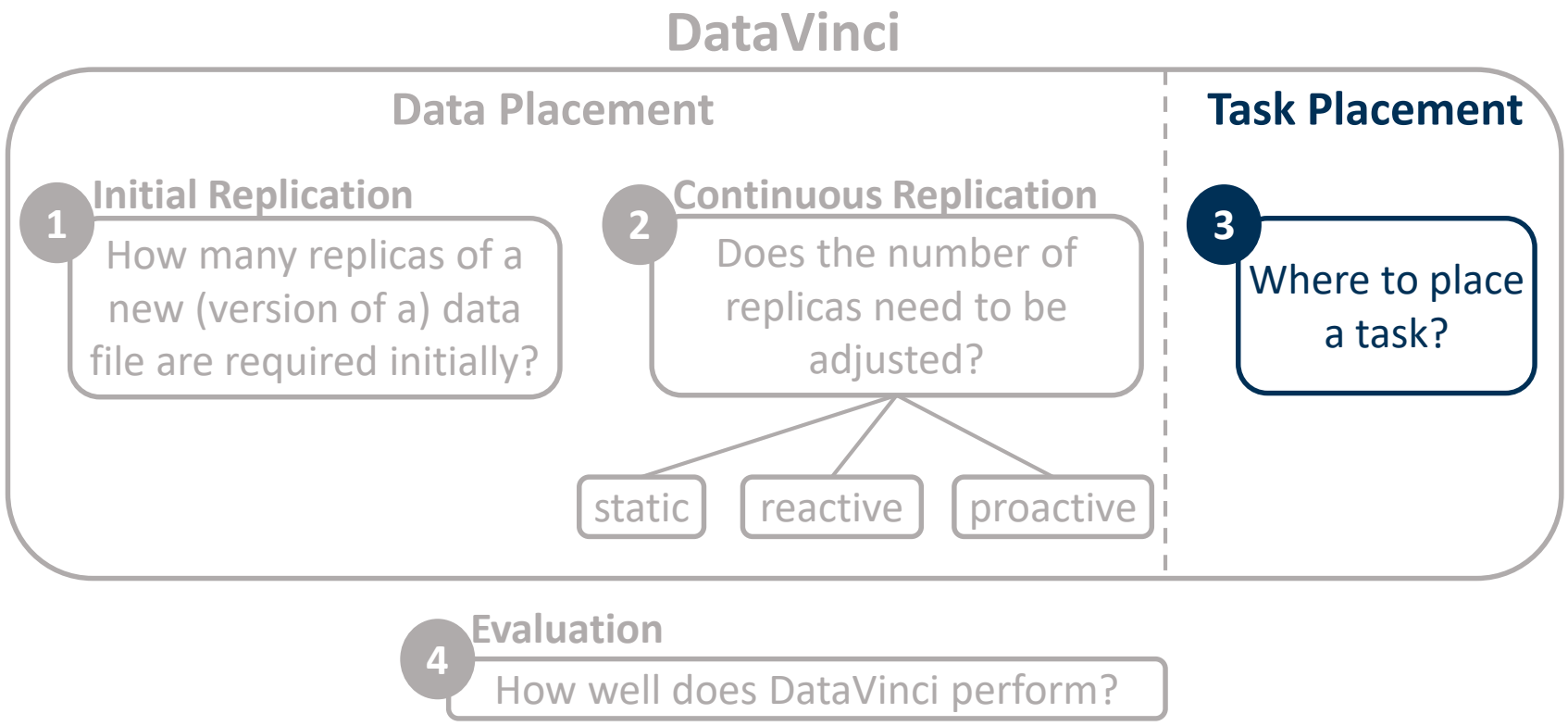
# Continuous replication
# Proactive



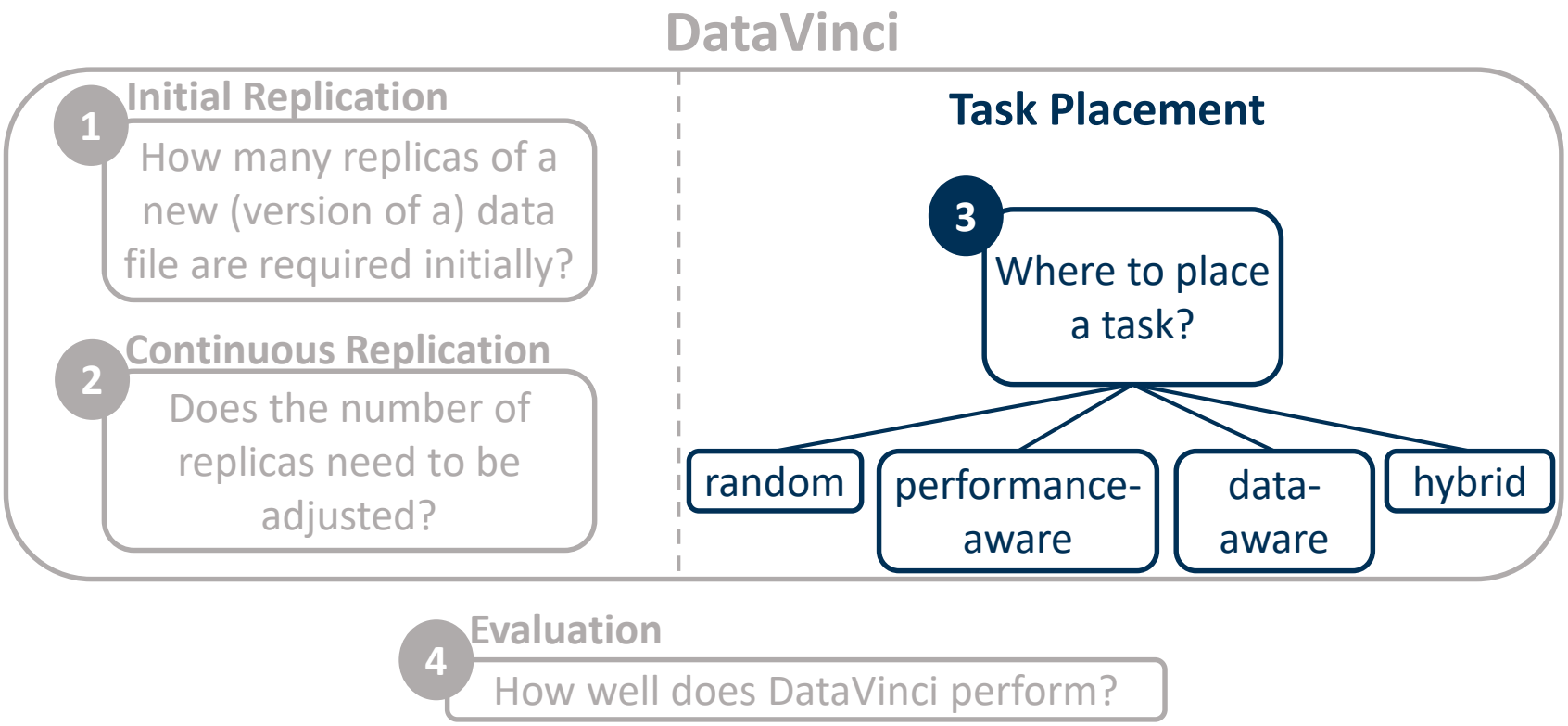
# Continuous replication



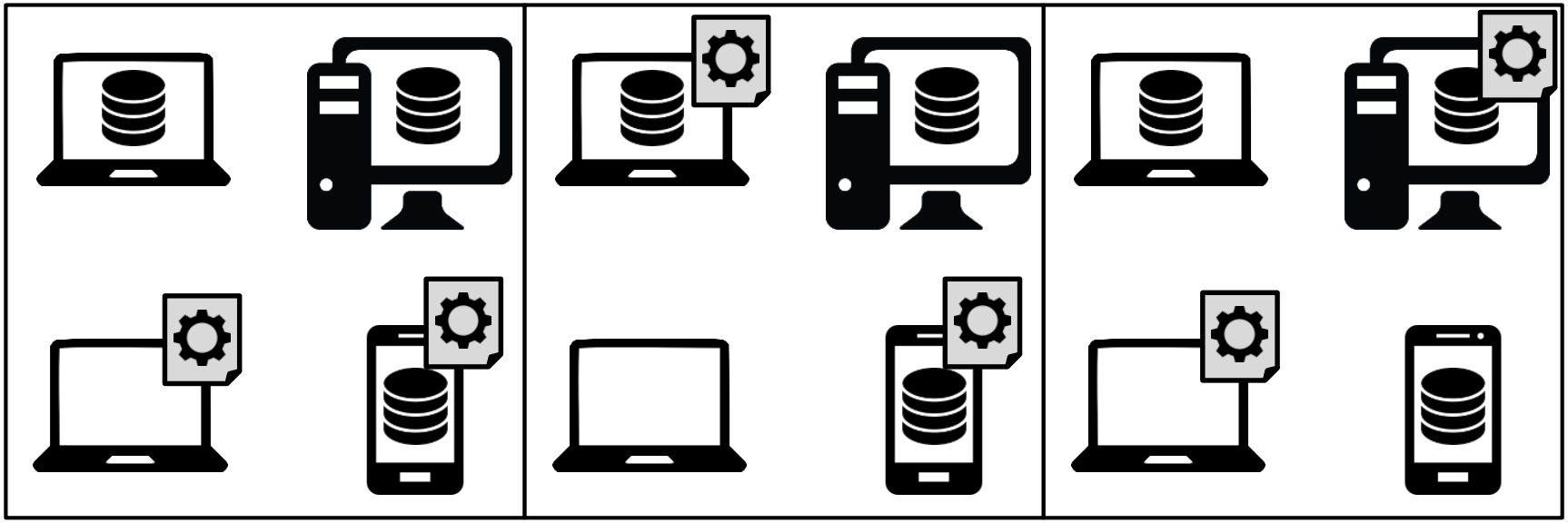
# Task placement



# Task placement

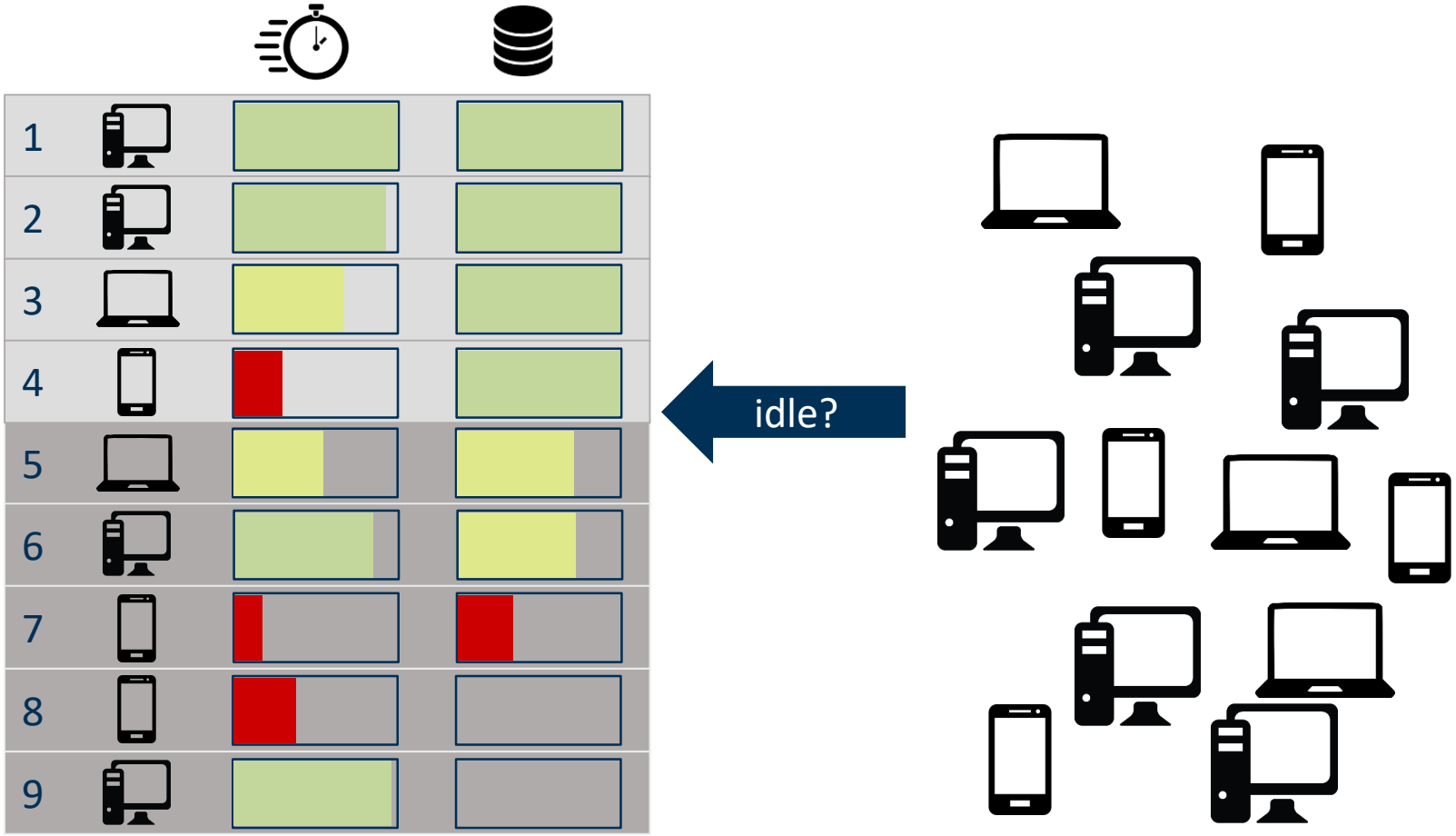


# Task placement

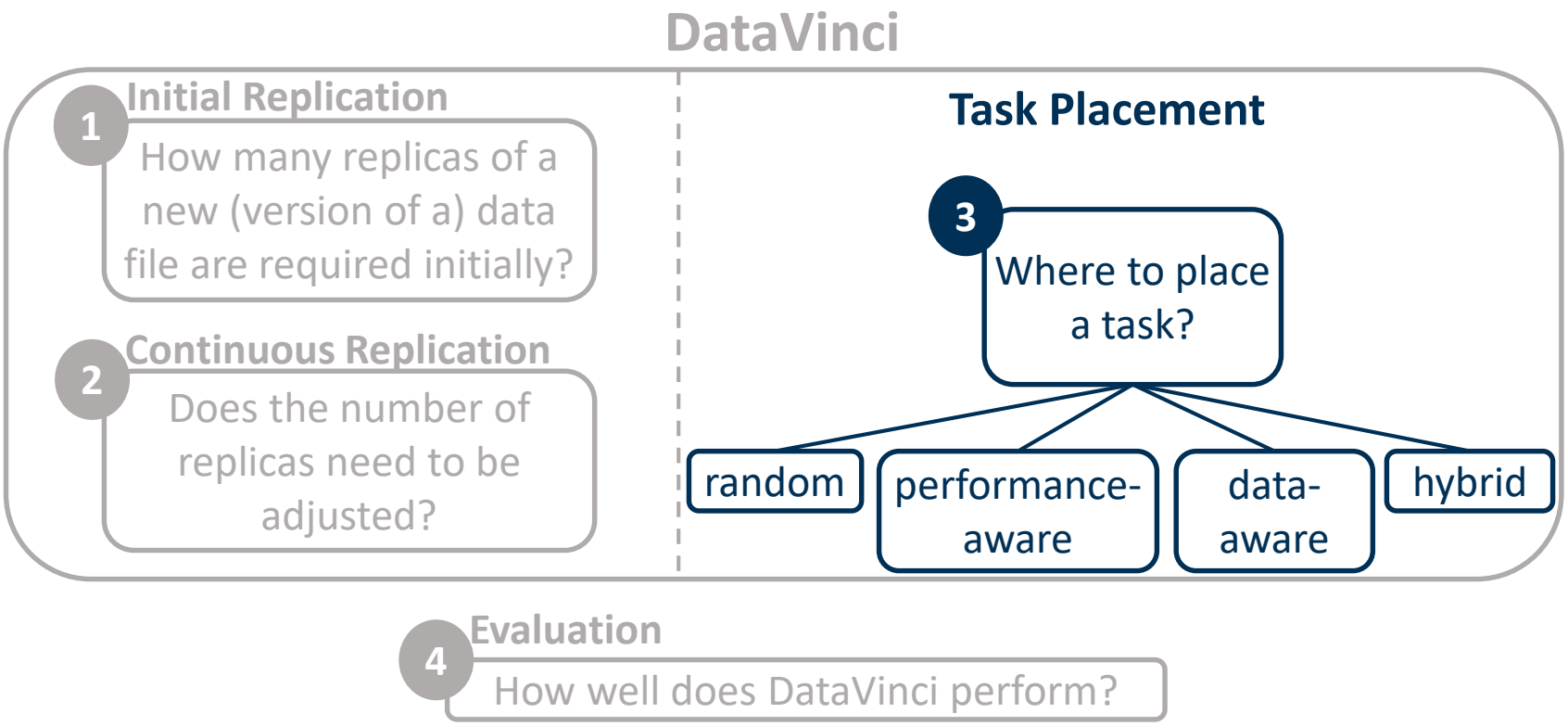


Random | Data-aware | Performance-aware

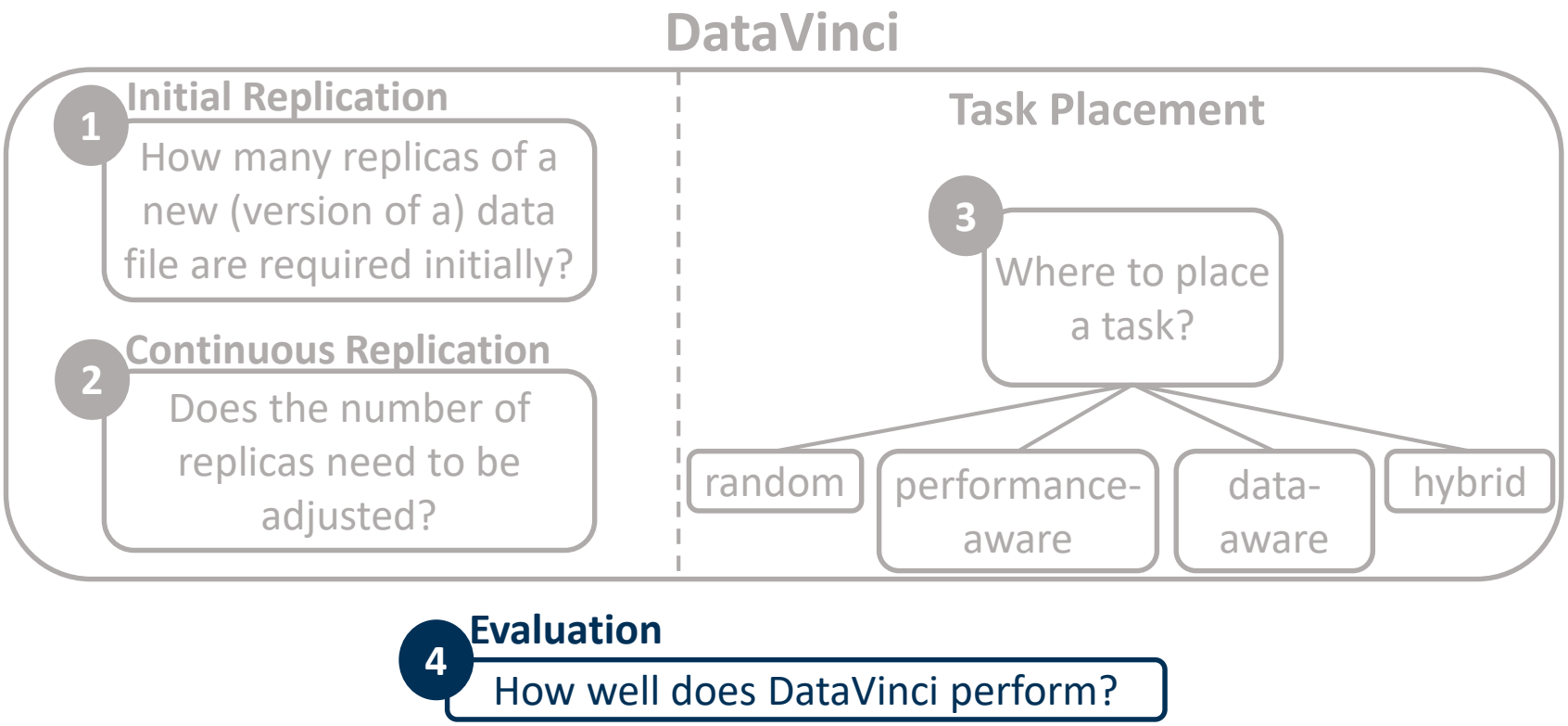
# Hybrid task placement



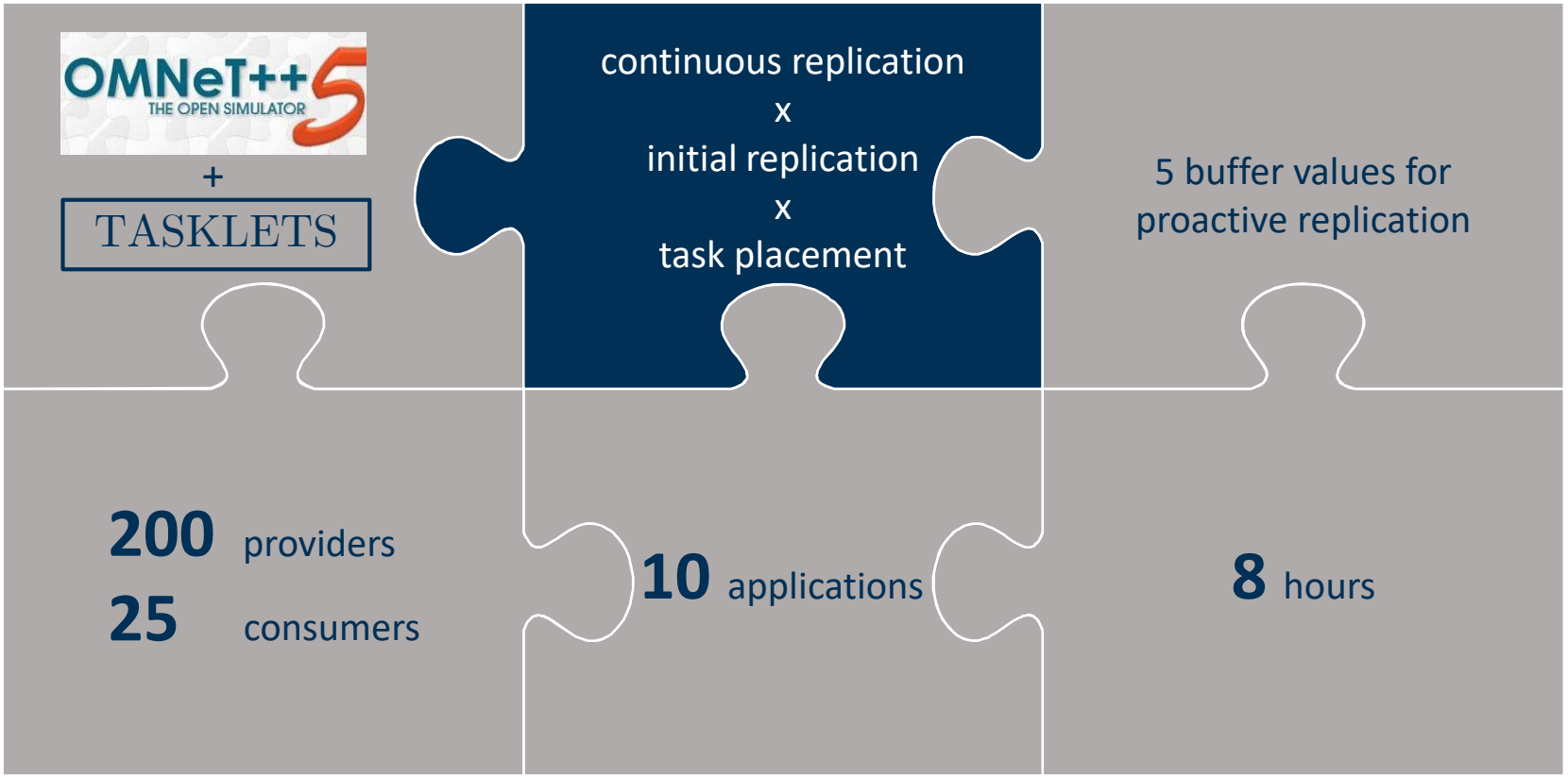
# Task placement



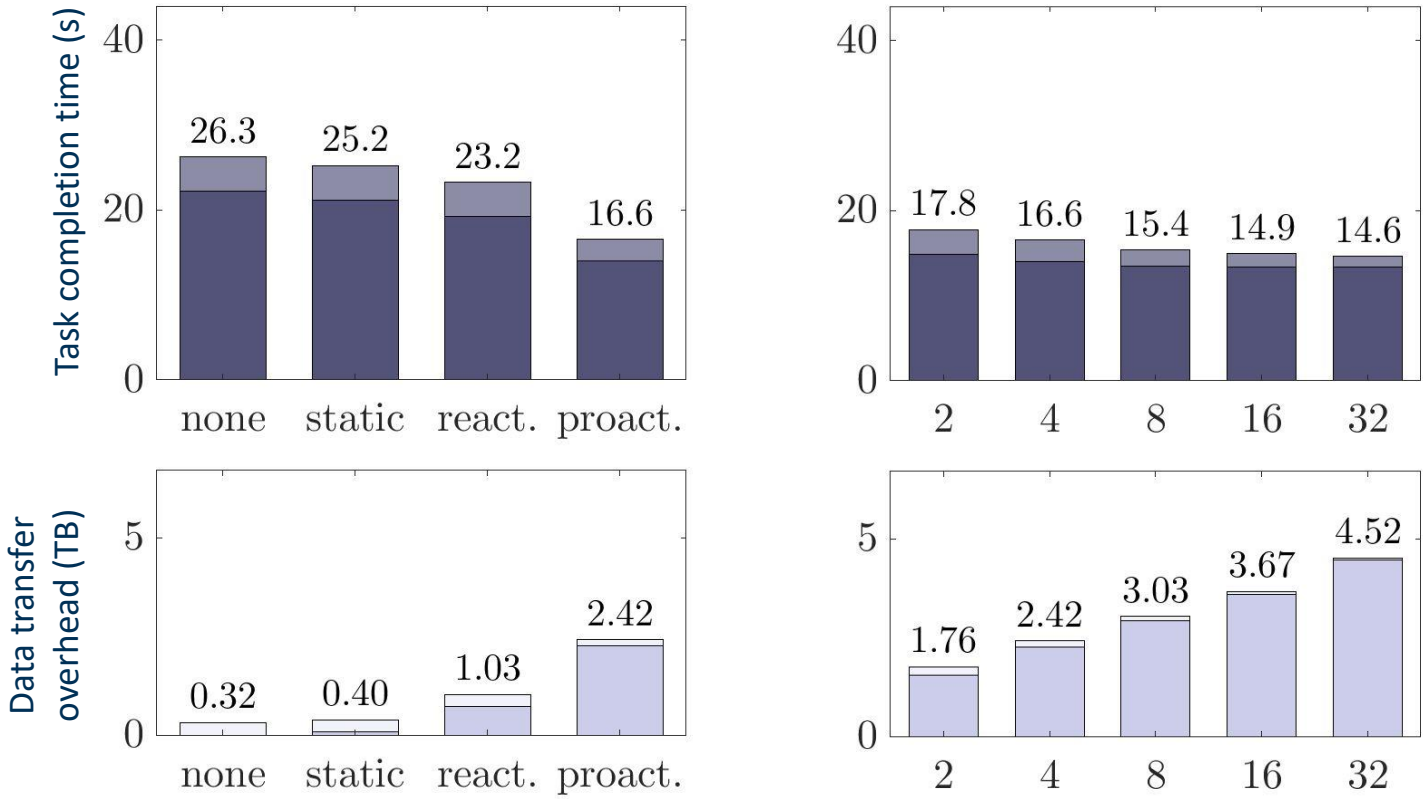
# Evaluation


DataVinci
1
Initial Replication
How many replicas of a new (version of a) data file are required initially?
2
Continuous Replication
Does the number of replicas need to be adjusted?
Task Placement
3
Where to place a task?
random
performance-aware
data-aware
hybrid
4
Evaluation
How well does DataVinci perform?

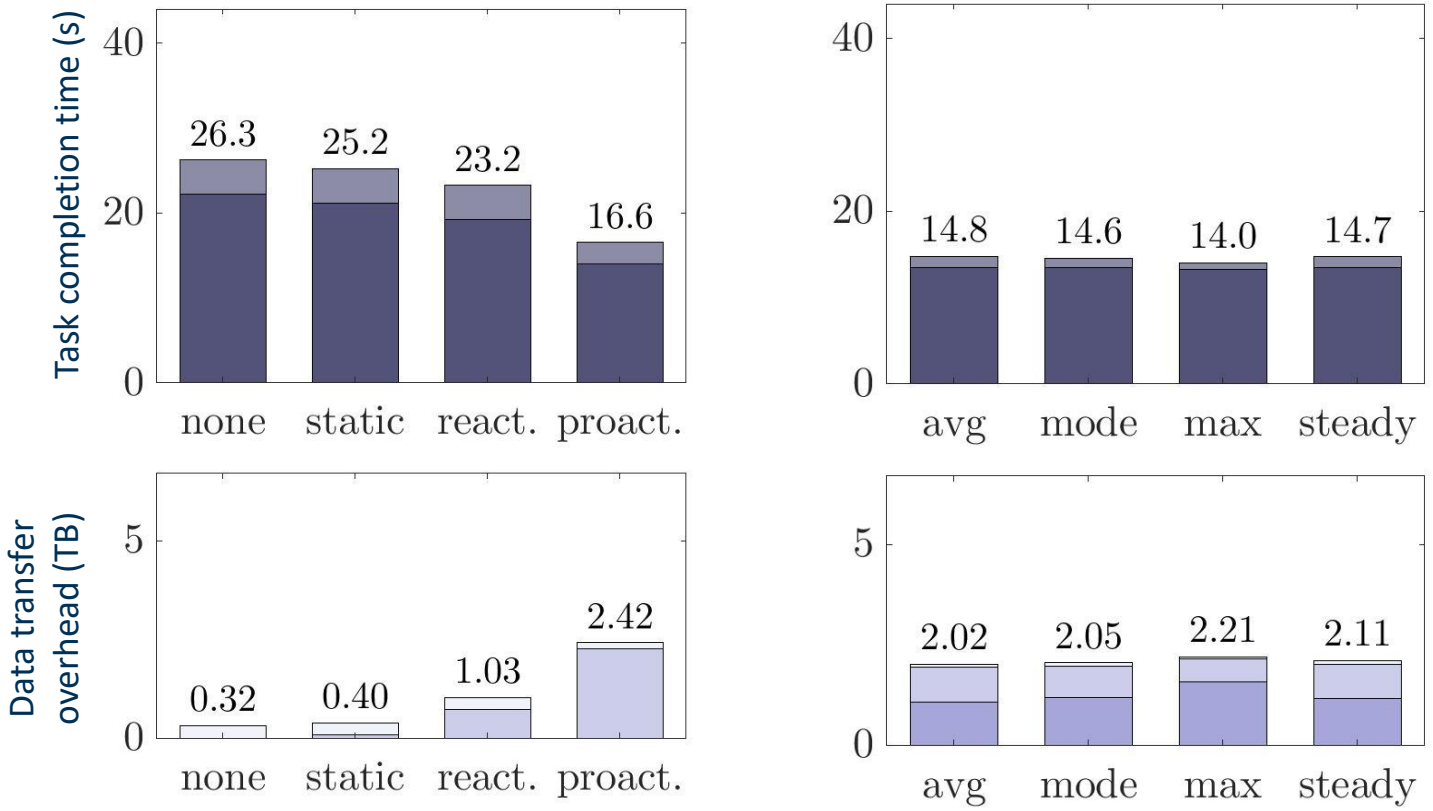
# Experimental setup



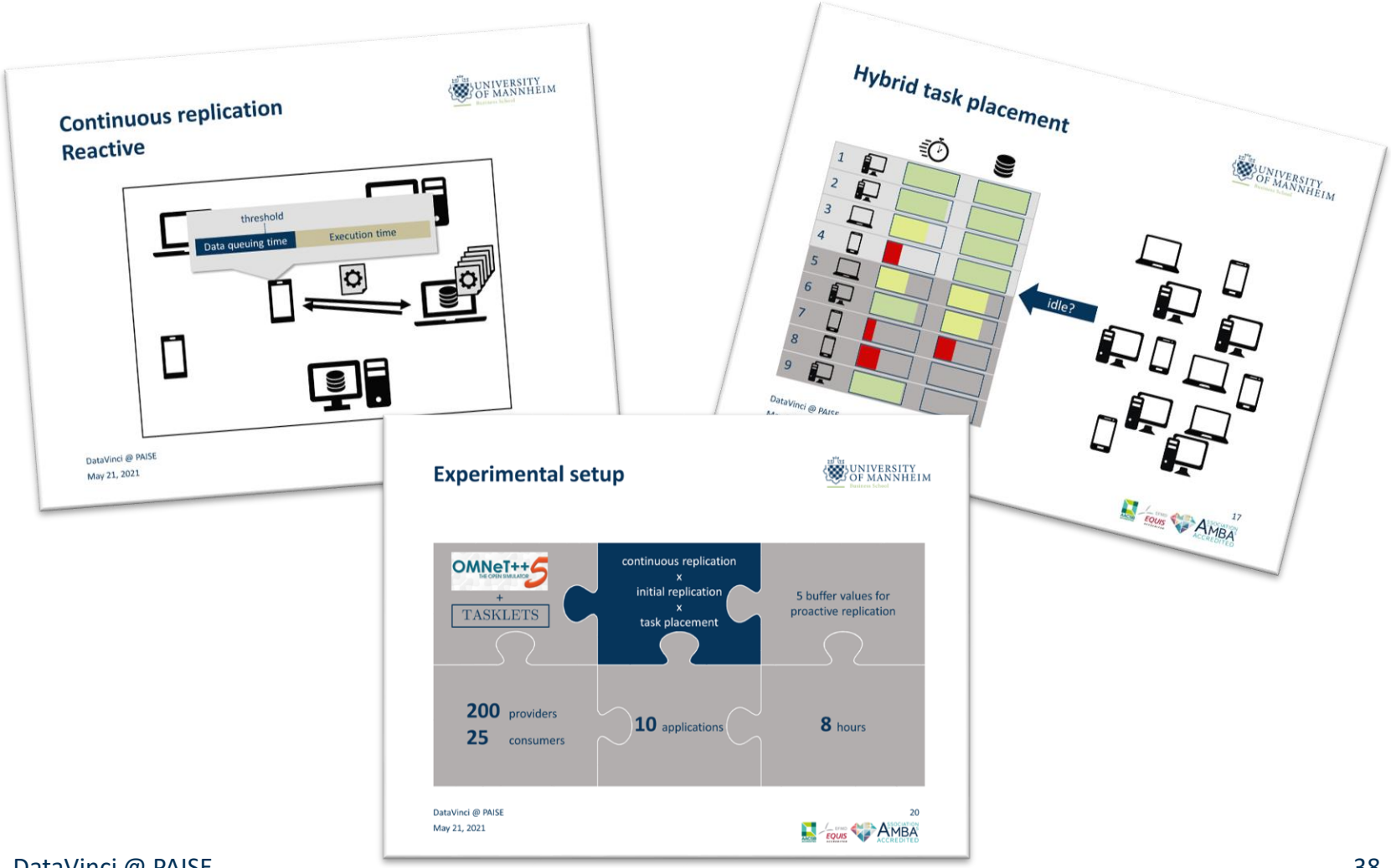
# Results


Task completion time (s)
40
20
0
26.3
25.2
23.2
16.6
none
static
react.
proact.
17.8
16.6
15.4
14.9
14.6
2
4
8
16
32
Data transfer overhead (TB)
5
0
0.32
0.40
1.03
2.42
1.76
2.42
3.03
3.67
4.52

# Results


Task completion time (s)
40
20
0
26.3
25.2
23.2
16.6
none
static
react.
proact.
Data transfer overhead (TB)
5
0
0.32
0.40
1.03
2.42
none
static
react.
proact.
40
20
0
14.8
14.6
14.0
14.7
avg
mode
max
steady
5
0
2.02
2.05
2.21
2.11
avg
mode
max
steady

# Conclusion


Continuous replication
Reactive
threshold
Data queuing time
Execution time
DataVinci @ PAISE
May 21, 2021
Hybrid task placement
idle?
DataVinci @ PAISE
17
Experimental setup
OMNeT++5
+
TASKLETS
continuous replication
x
initial replication
x
task placement
5 buffer values for
proactive replication
200 providers
25 consumers
10 applications
8 hours
DataVinci @ PAISE
May 21, 2021
20

UNIVERSITY OF MANNHEIM
Business School

# Thank you for your attention!

We would like to thank the German Research Foundation DFG.