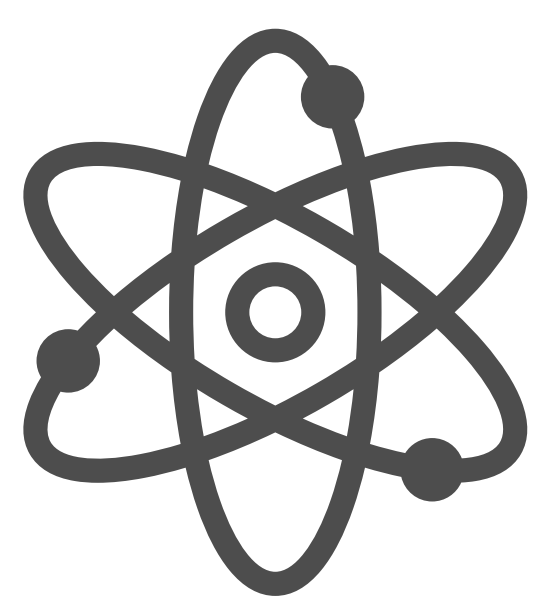



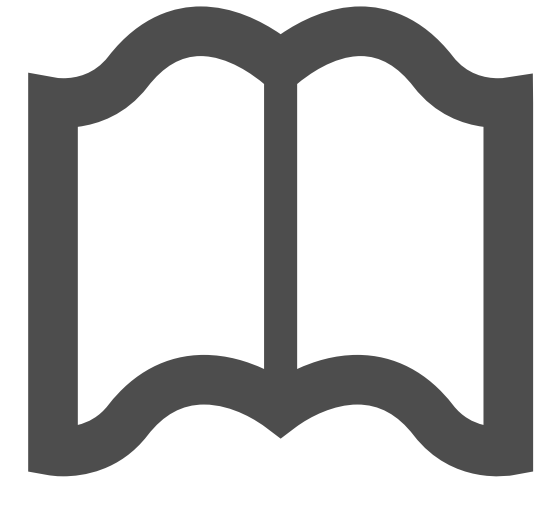
Motivation



Scientific discovery

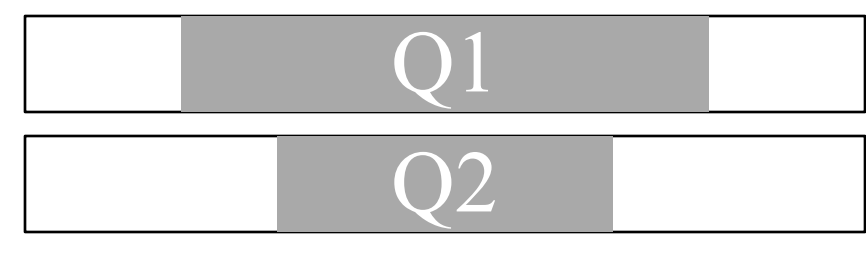


Data science

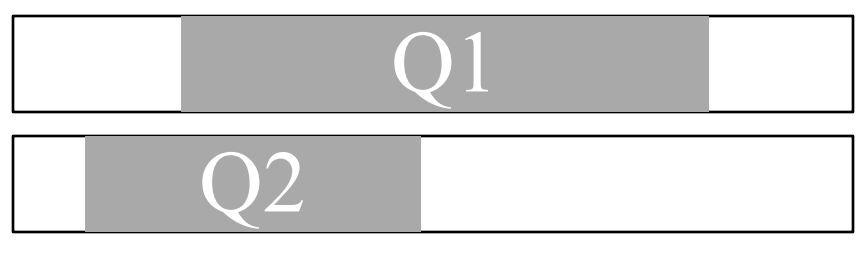


Machine learning

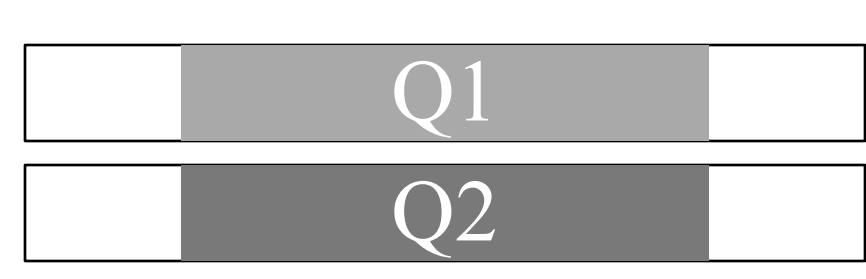
Statistics are everywhere!



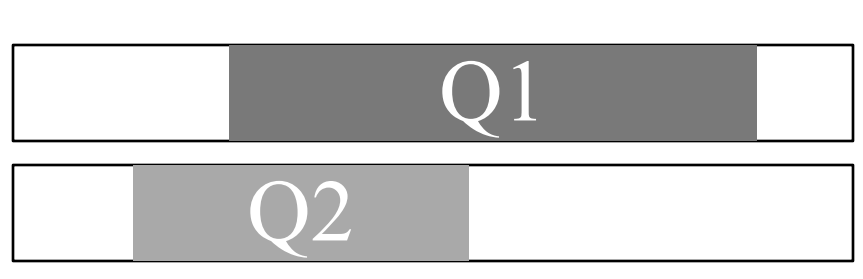
Sub-ranges



Overlapping ranges




Different statistics




Mixed

Exploratory workloads are repetitive



Repeated data access

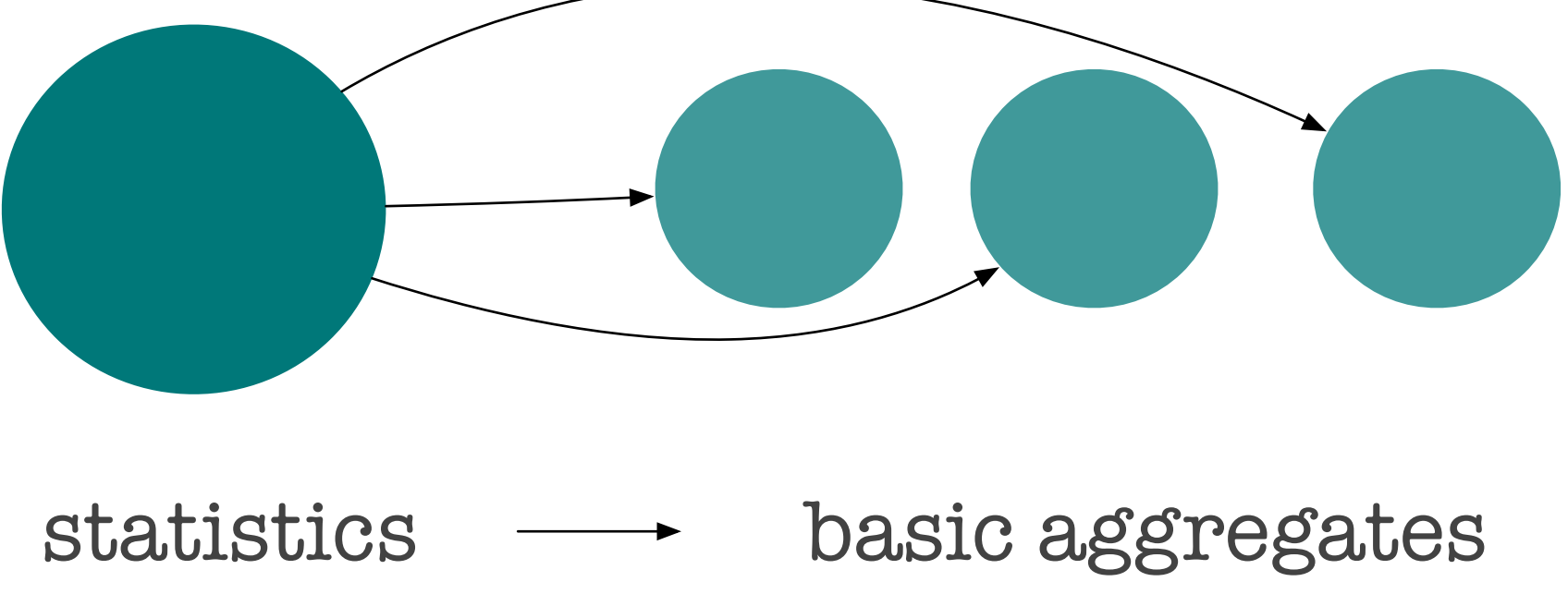


Cumulative data access

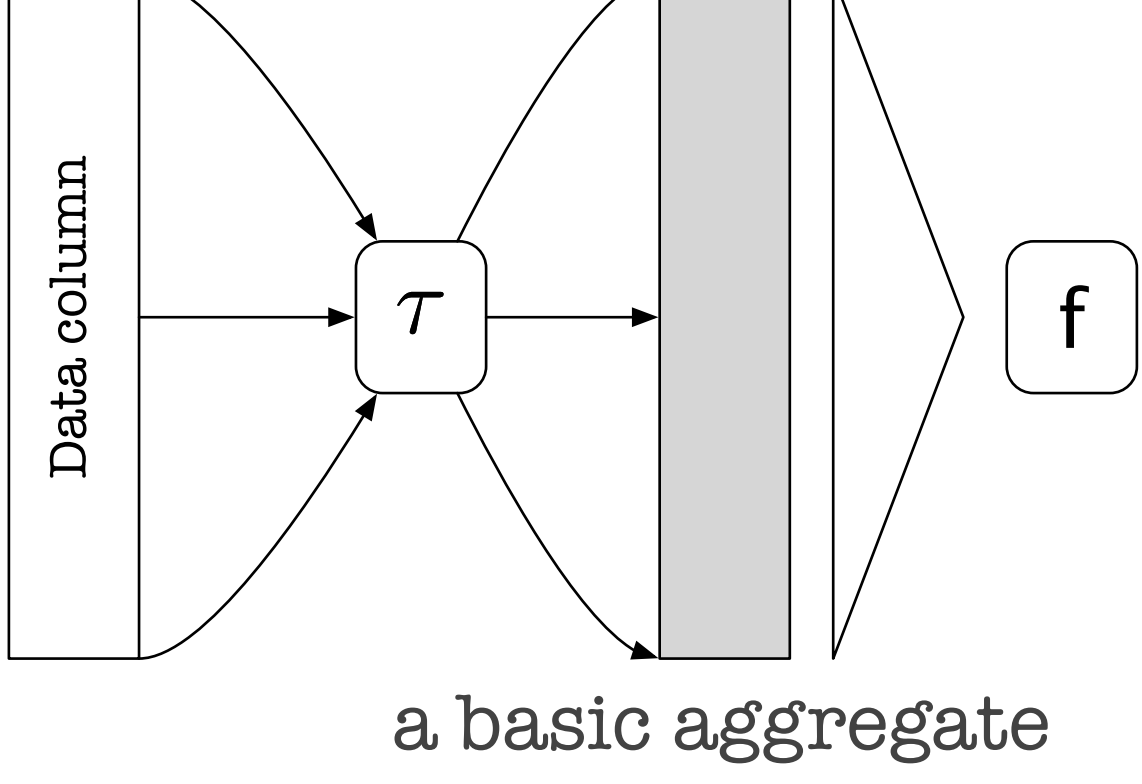
Statistical queries

Existing tools compute from scratch

Data Canopy



statistics → basic aggregates

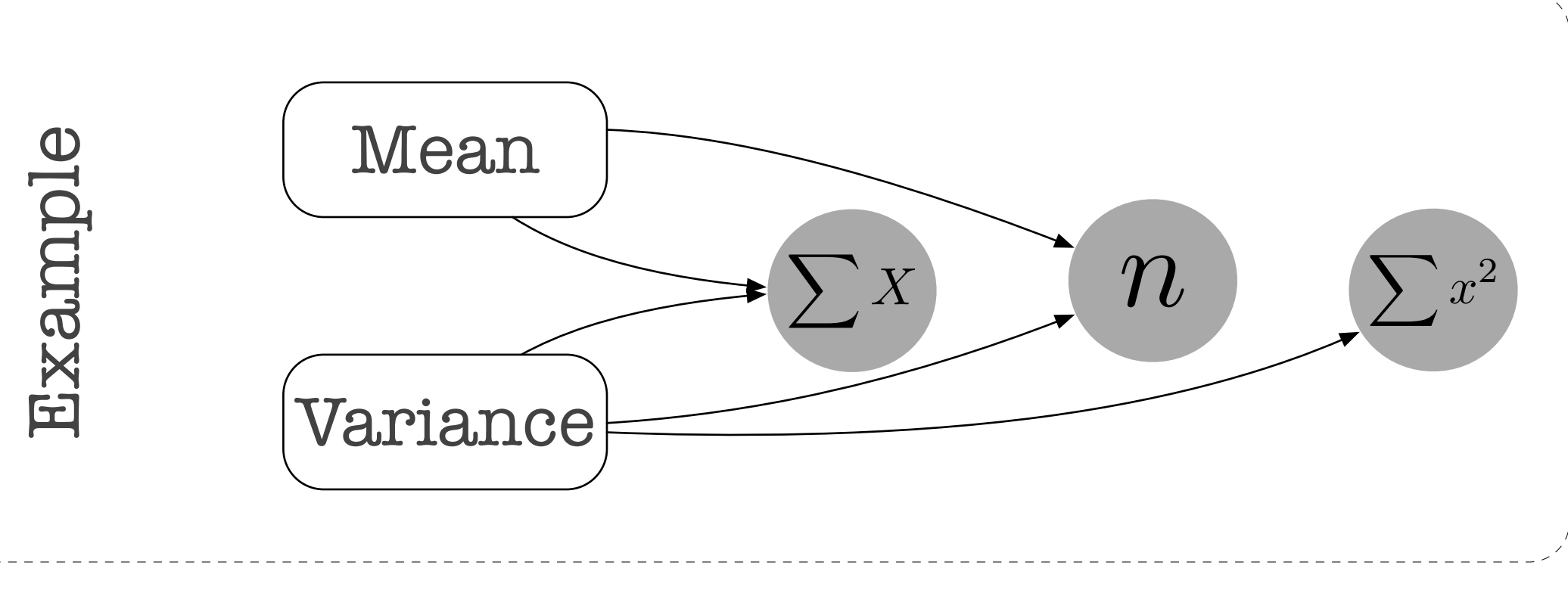


Data column

τ

a basic aggregate

f



Example

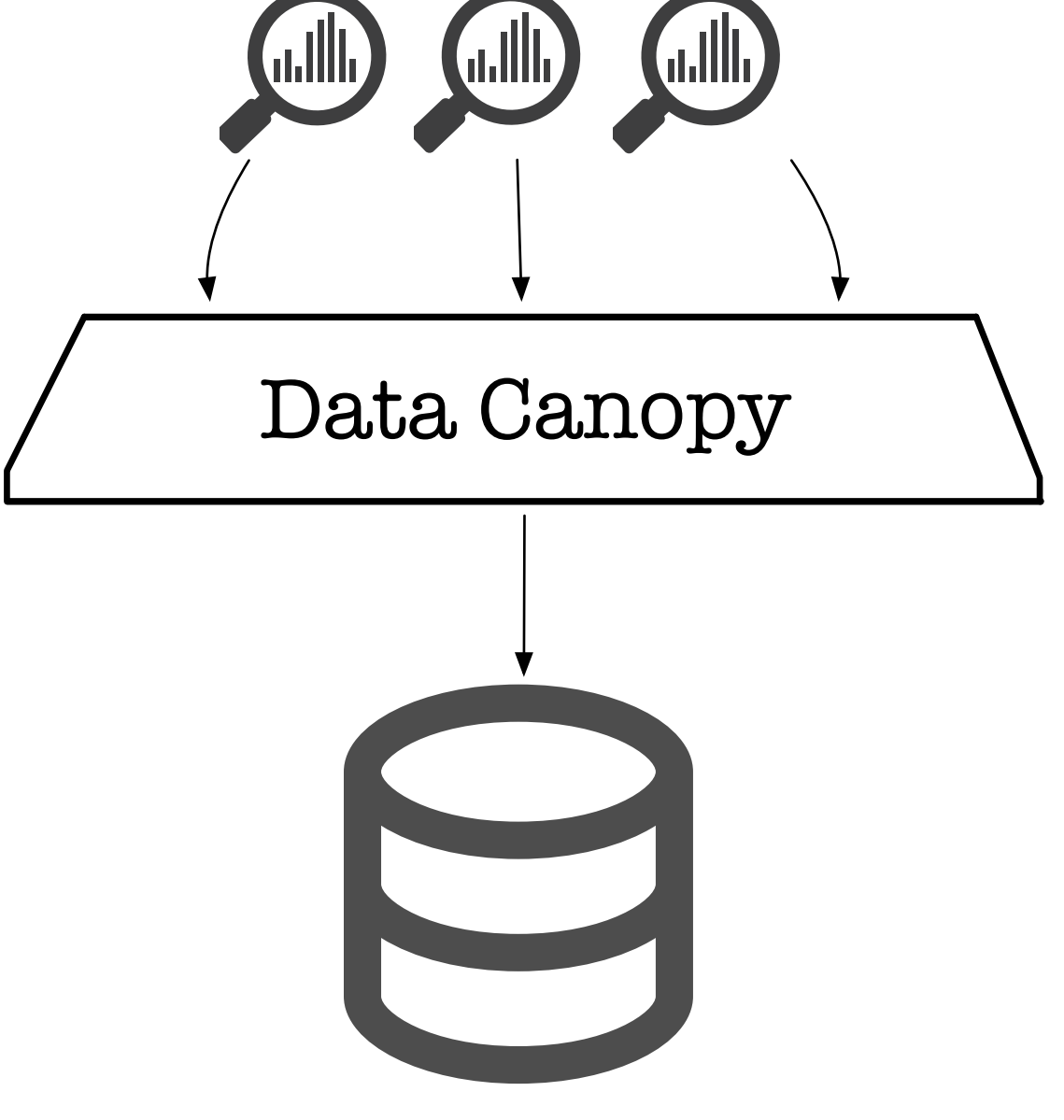
Mean

Variance

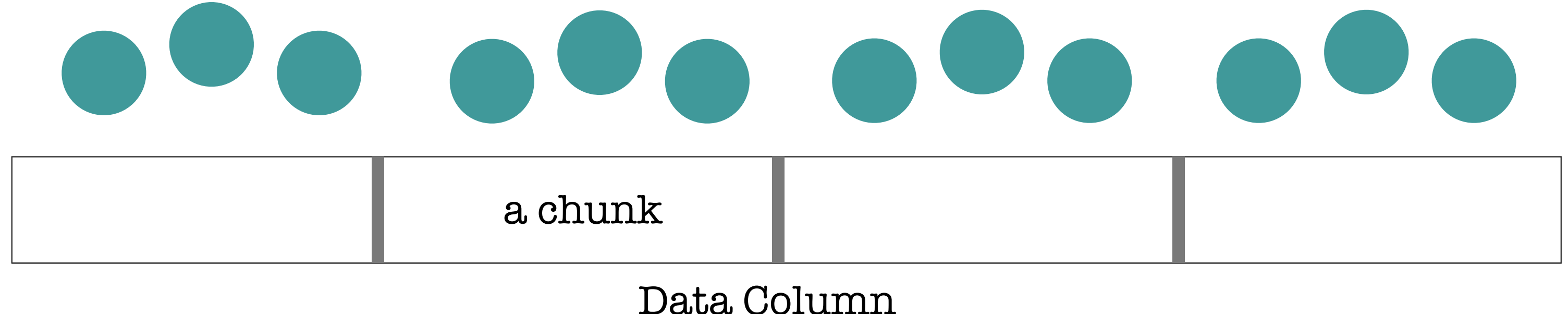
$\sum x$

n

$\sum x^2$

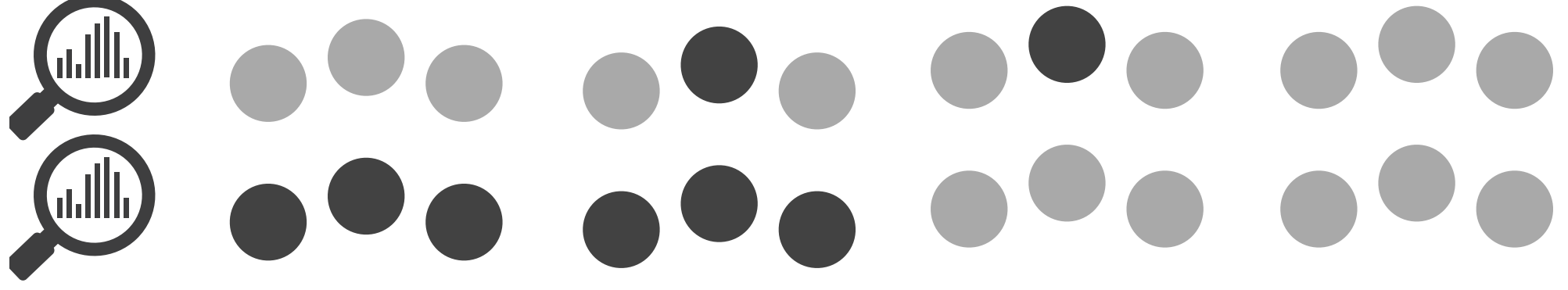


Data Canopy




a chunk

Data Column



Use across statistics and ranges

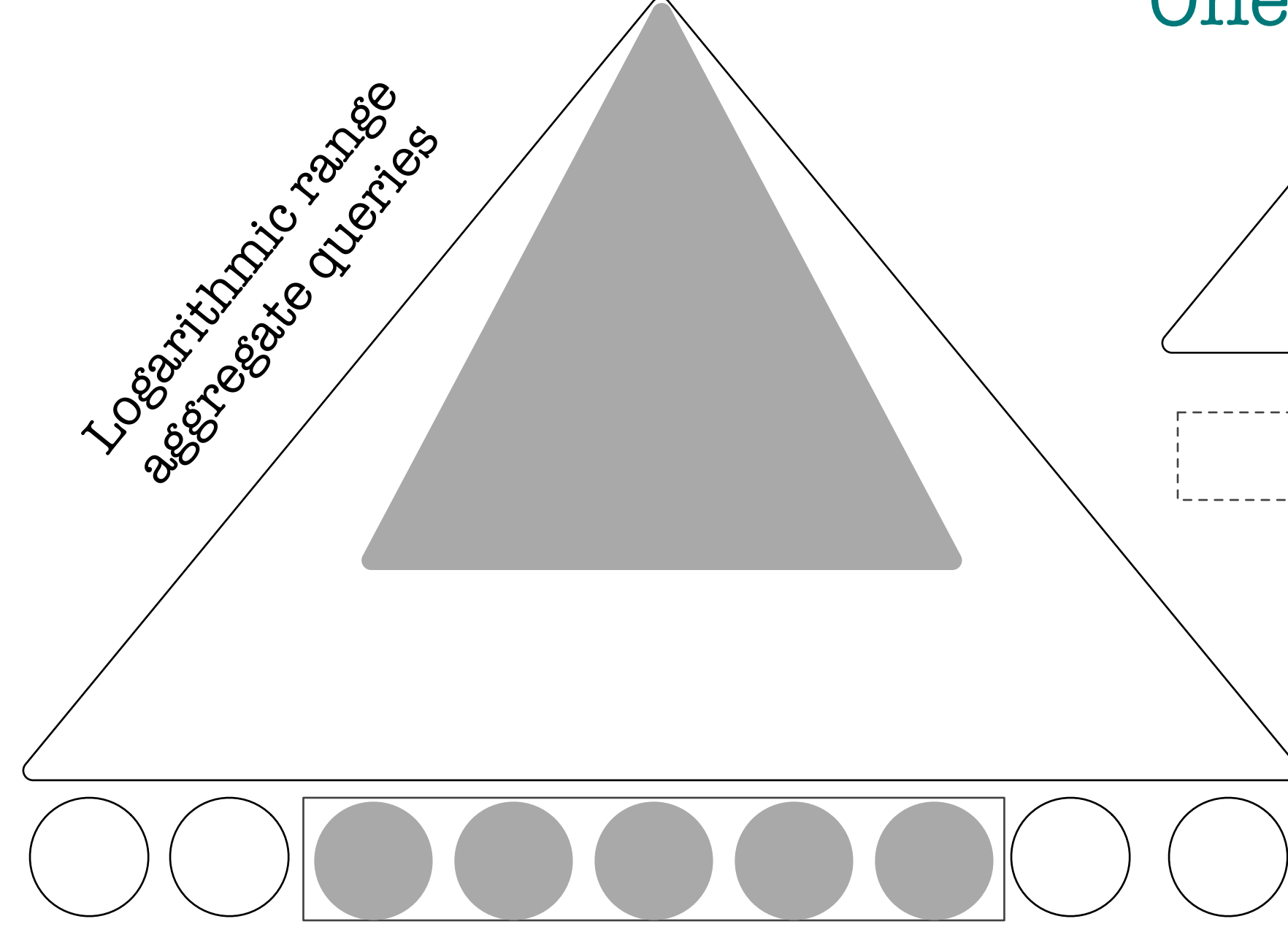


Avoid redundant data access

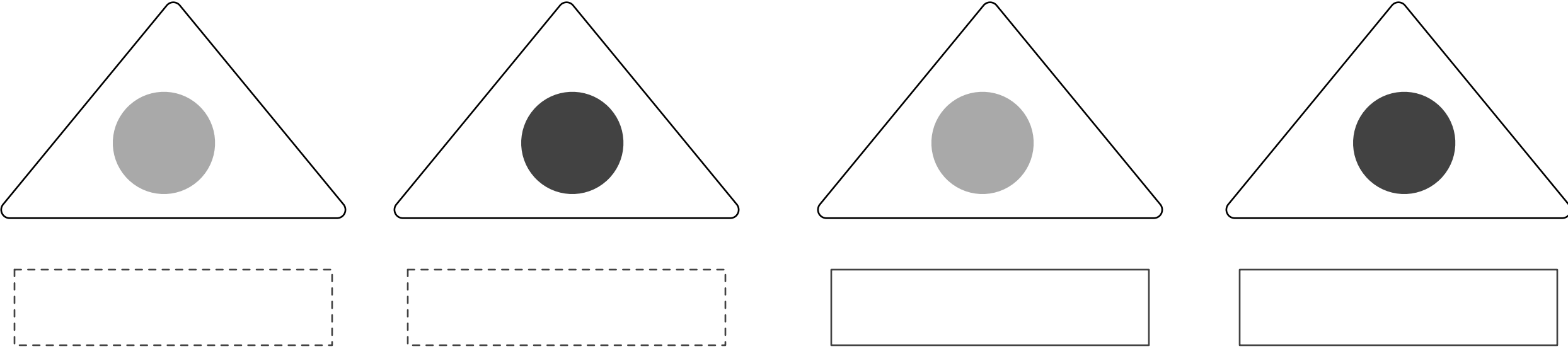
Maintain basic aggregates for sub-ranges (chunks)

Synthesize statistics from primitives


Layout



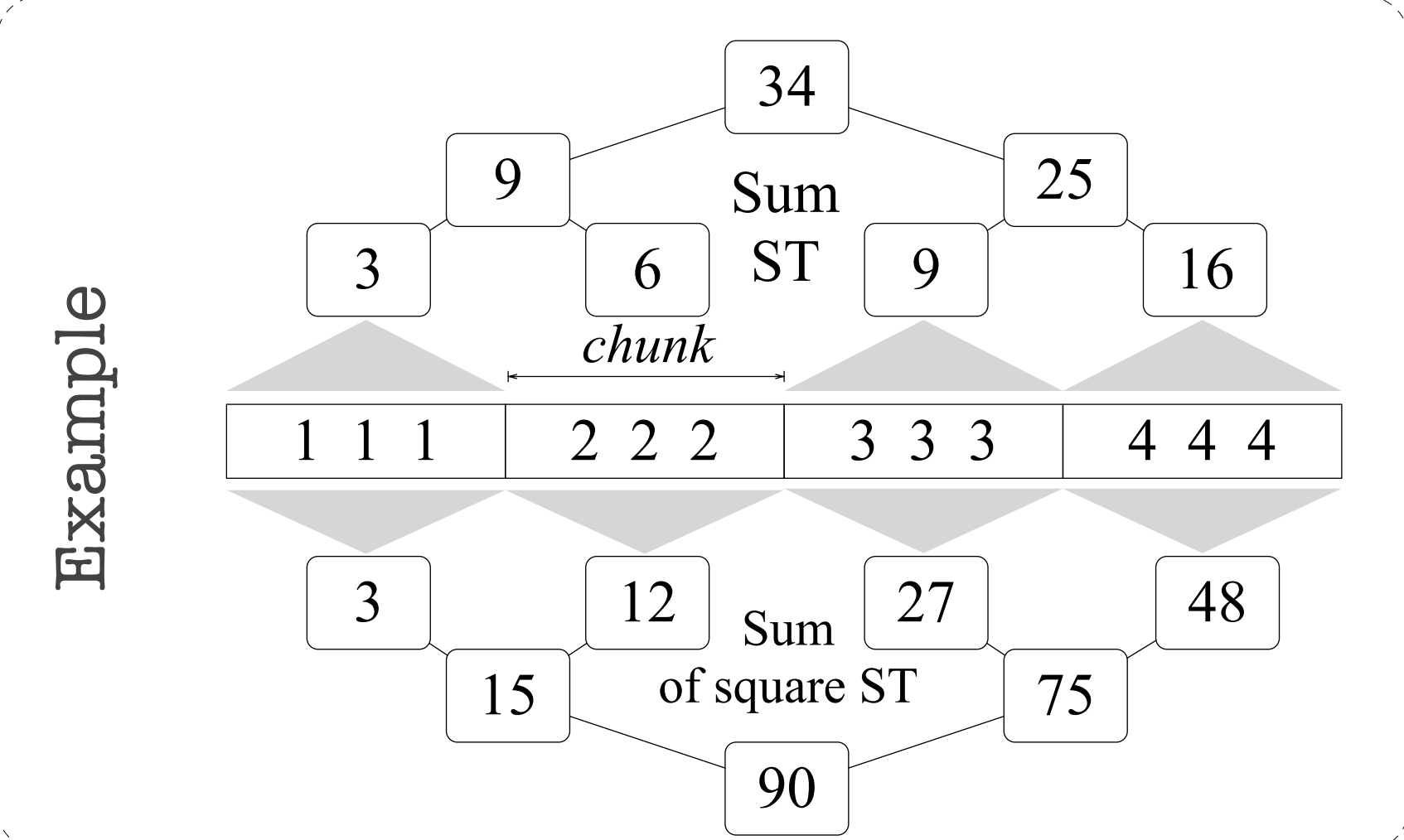
Logarithmic range aggregate queries



One segment tree per basic aggregate per column



Segment trees store basic aggregates



Example

34

9 6 25

Sum ST

3 9 16

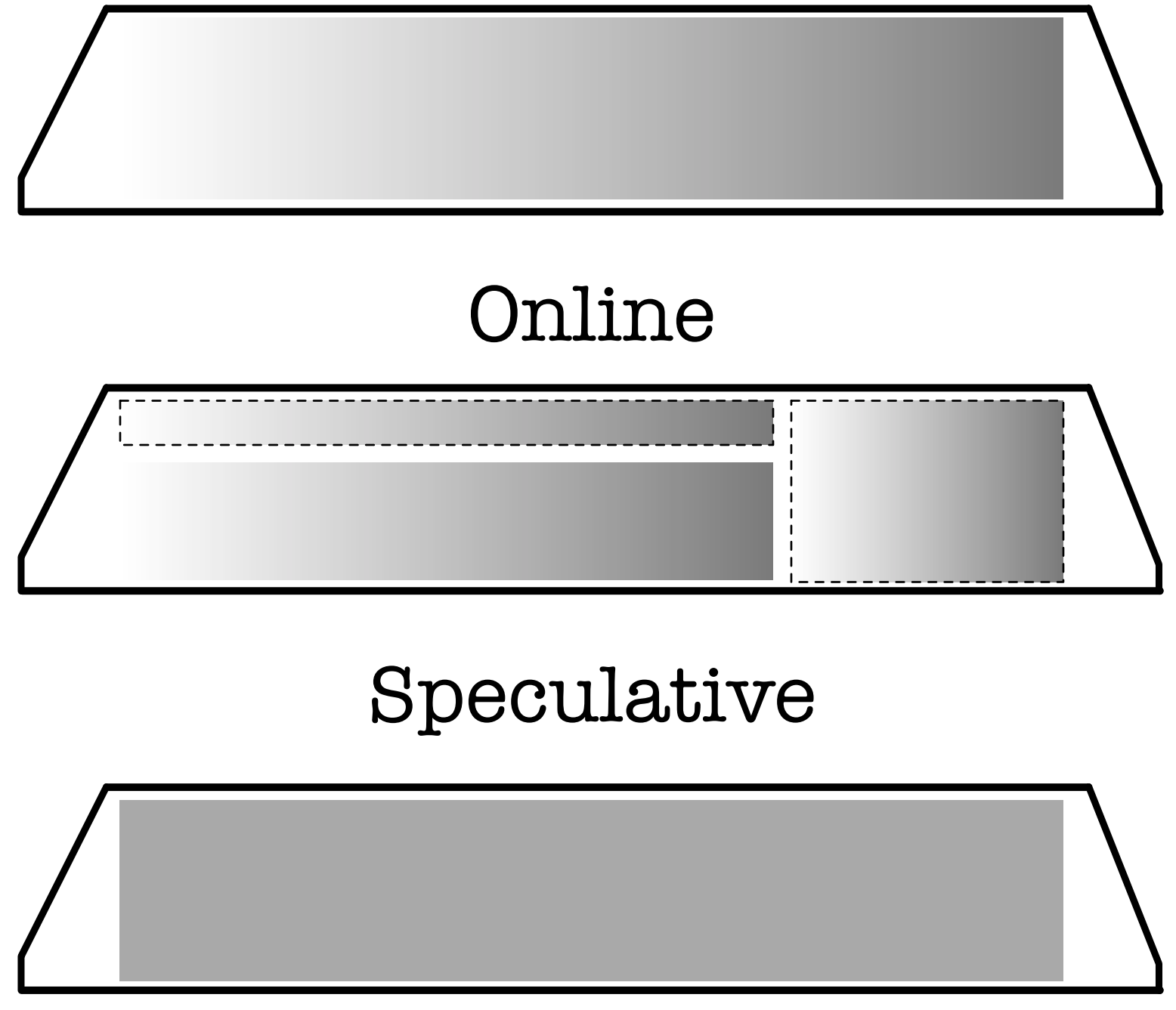
1 1 1 2 2 2 3 3 3 4 4 4

chunk

3 12 27 48

Sum of square ST

15 75 90



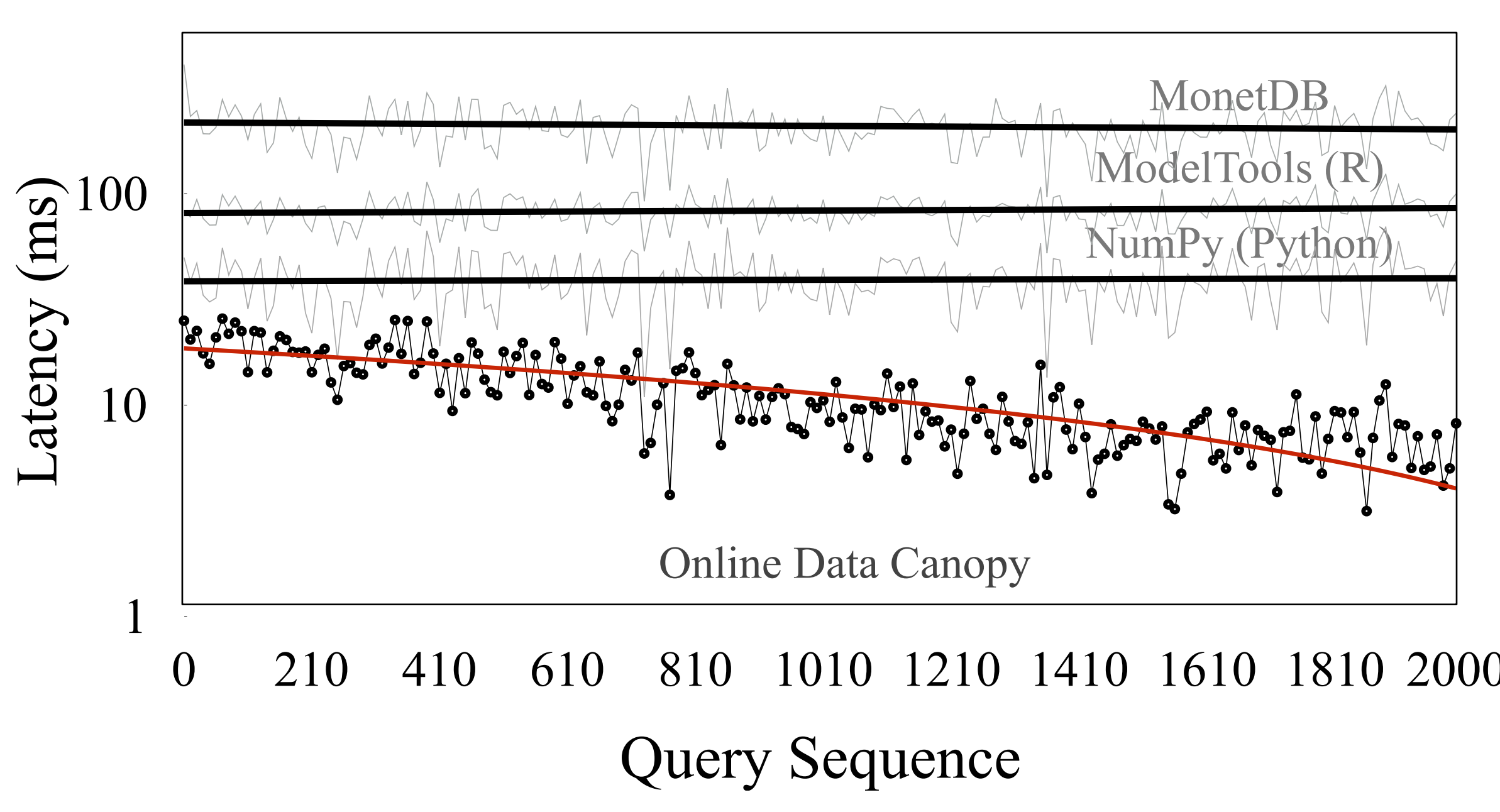
Online

Speculative

Offline

Operate in multiple scenarios

Evaluation



Latency (ms)

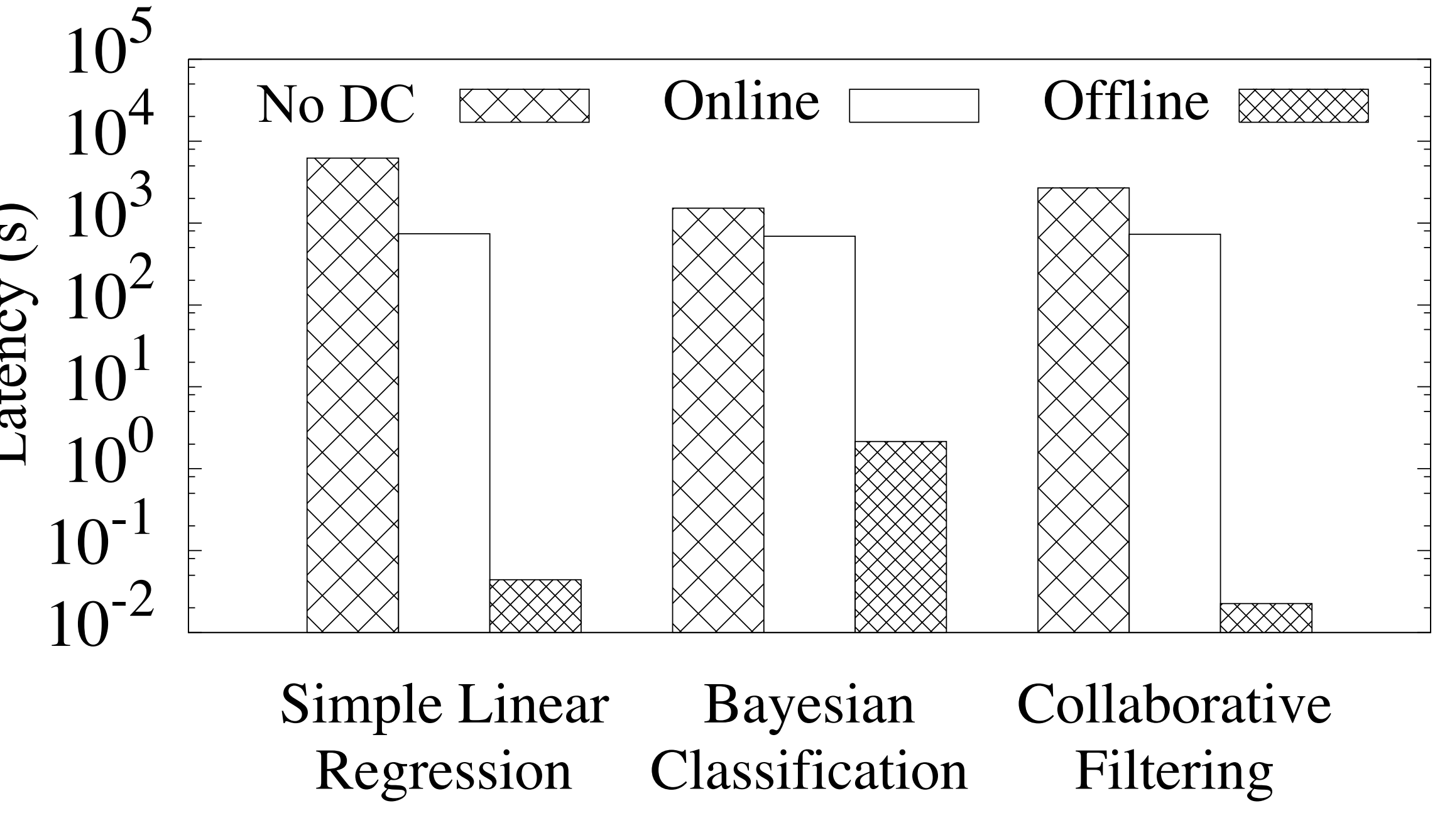
Query Sequence

MonetDB

ModelTools (R)

NumPy (Python)

Online Data Canopy



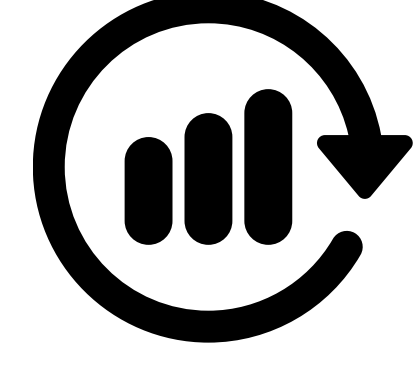
Latency (s)

No DC Online Offline

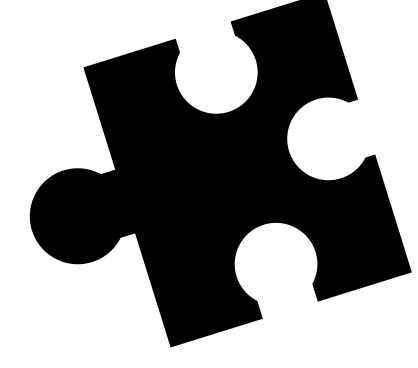
Simple Linear Regression

Bayesian Classification

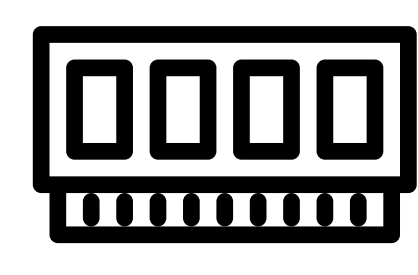
Collaborative Filtering



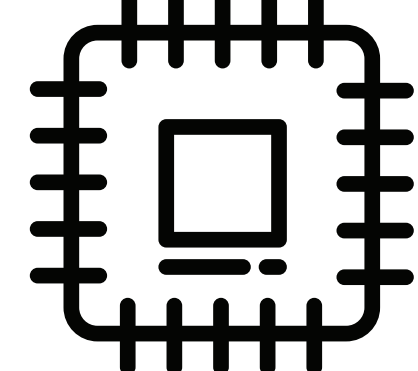
Updates



Chunk size



Memory pressure



Multicore

Accelerating exploratory statistical analysis

Accelerating machine learning algorithms

More in Paper