Scaling out Big Data Missing Value Imputations

(Pythia vs. Godzilla)

Christos Anagnostopoulos & Peter Triantafillou

School of Computing Science, University of Glasgow, UK

20^{th} ACM KDD 2014 New York City

Missing Data

Data quality in Big Data processing:

- **Missing Values (MV)** in multidimensional data.

$$\mathbf{x} = [x_1, x_2, ?, x_4, \ldots, ?, x_d]$$

- **Example**: survey databases; industrial databases; medical databases; gene expression microarray datasets.

…bias is introduced into the induced knowledge.

Missing Data

Common solutions to the MV problem:

- **Ignore** or **exclude** MV data.
- **Fill-in** MVs (*imputation*)
  - **MV (Substitution) Algorithm** replaces MVs with plausible values.
  - **Imputation error:** difference between *actual* (*unknown*) value and *predicted* (*imputed*) value.

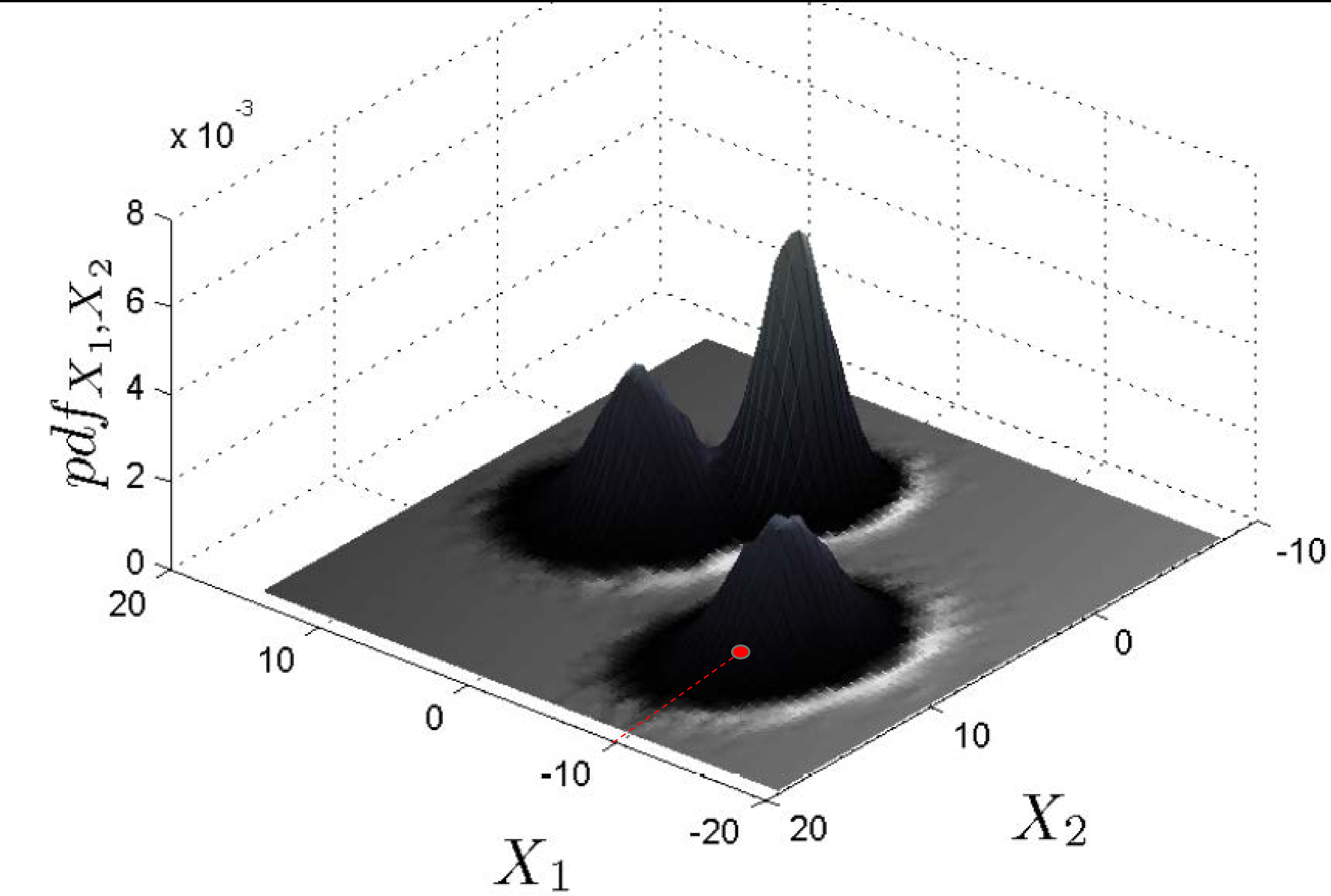
Motivation

- **MV Algorithms** **ensure** low imputation errors but **are** computationally expensive;
  - …performance depends on data size!
- Deal with **large-scale** datasets, which grow significantly with time!
- **User community** can be very large;
  - MV imputation requests' arrival rate becomes high too!

Motivation

- **Not all** MV substitution tasks are ‘**embarrassingly parallelizable**’.

- If so,
  - not all *regions* of a dataset are ‘**relevant**’ for imputation;
  - …some data regions might negatively contribute or even ‘hurt’ the result of the MV Algorithm.


x 10^-3
8
6
4
2
0
pdf_{X_1,X_2}
20
10
0
-10
-20
X_1
-10
0
10
20
X_2

Observation

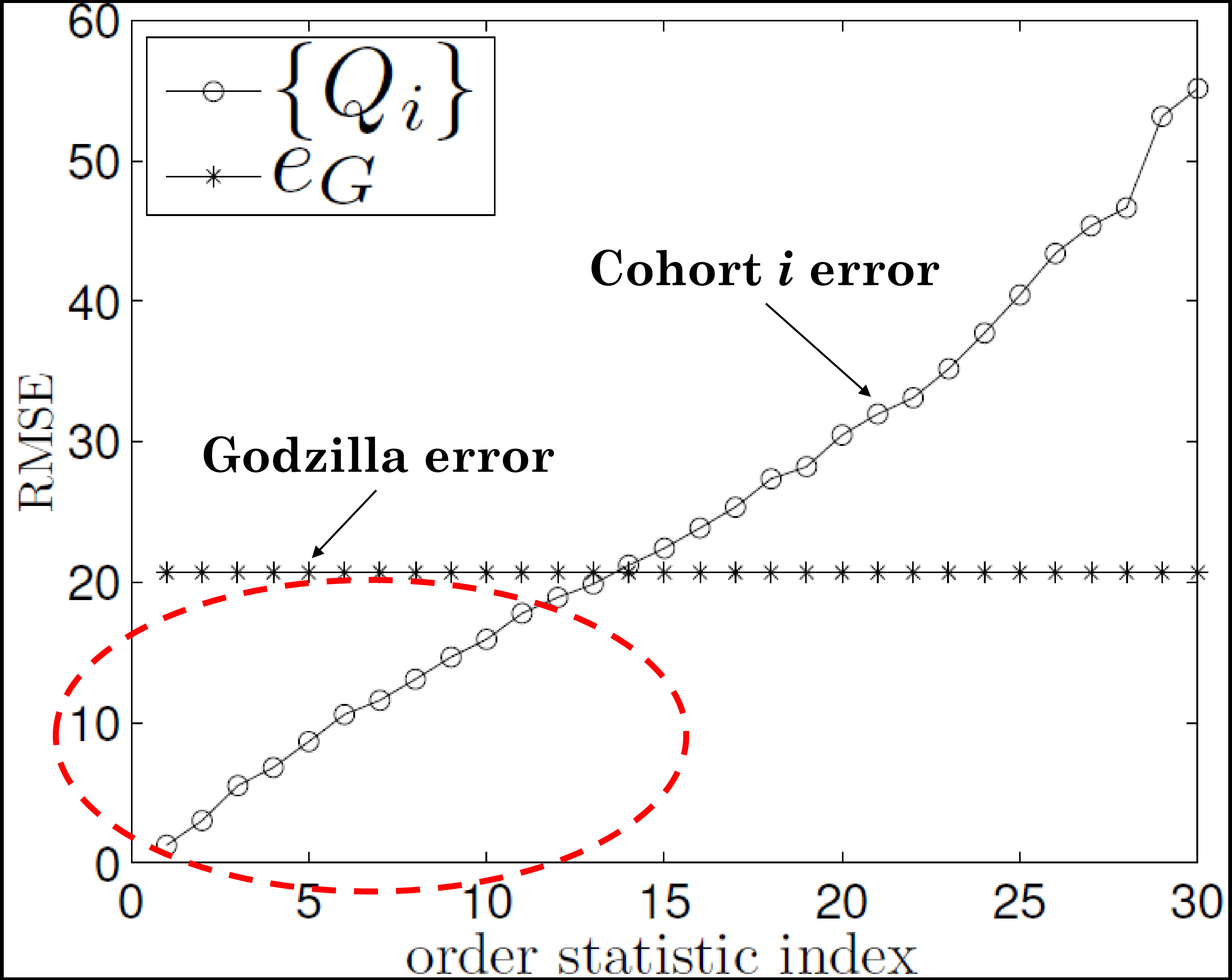
- A **single machine** '**Godzilla**' contains a massive dataset.
- Godzilla serves **MV imputation requests** by performing a MV algorithm.

Idea No 1

Replace Godzilla by a fixed number of commodity machines '**Cohorts**'.

1. **Partition** (randomly) the dataset.
2. Each Cohort contains **a portion of** the dataset.
3. Cohorts **perform** locally a MV Algorithm.
4. **Aggregate** all imputations.

Benefit: We obtain **efficiency** and **scalability**



$\{Q_i\}$
e_G
Cohort i error
Godzilla error
RMSE
order statistic index
0
10
20
30
40
50
60
0
5
10
15
20
25
30

Idea No 2

Pythia predicts the appropriate subset of Cohorts for engaging them in performing MV Algorithm in parallel.

Pythia locally maintains a specific information for each Cohort's dataset: '**Signature**'.

Benefit: **Comparable / better accuracy** instead of

- engaging **all** Cohorts!
- using only Godzilla!

Signature

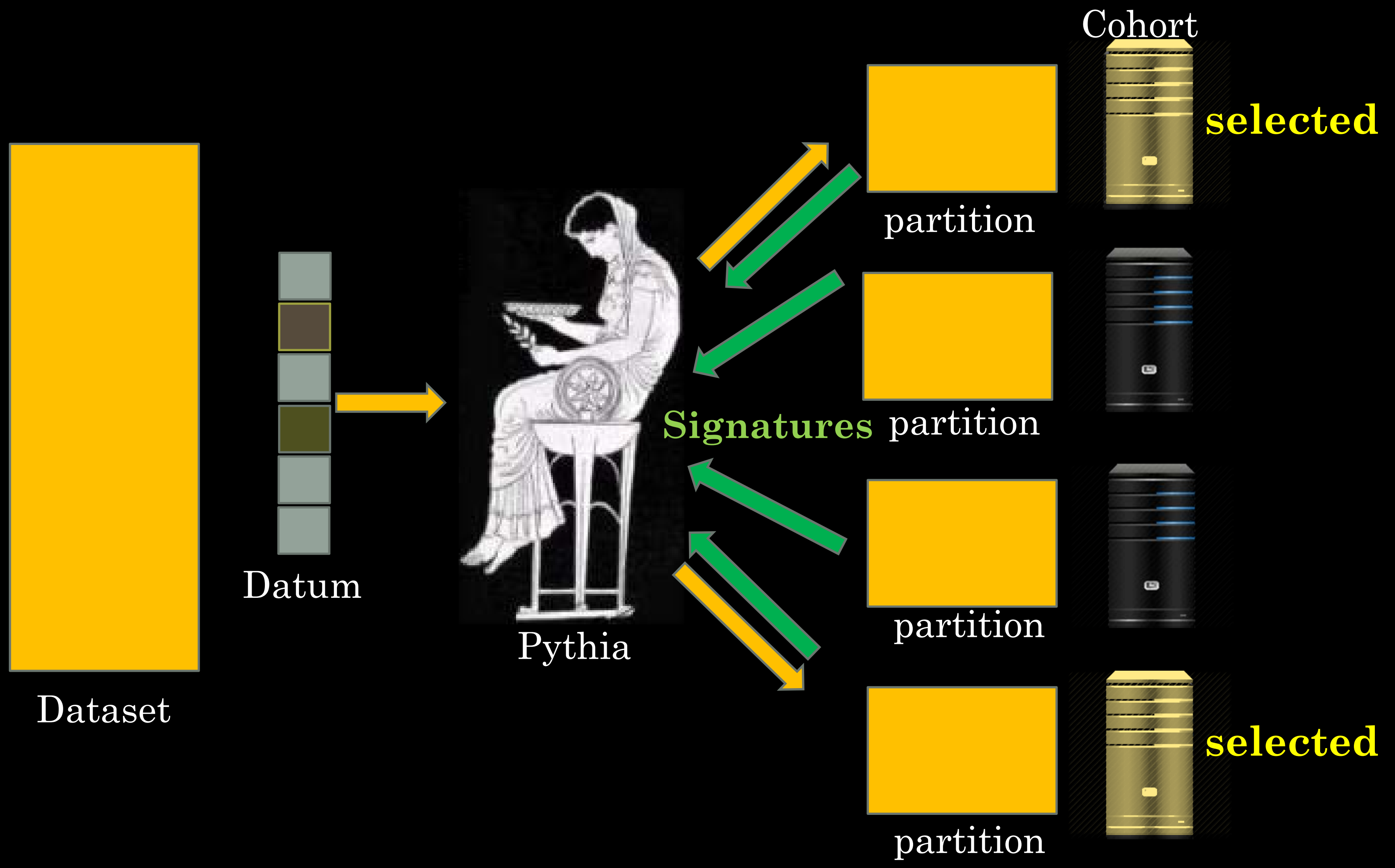
- **Information** used for predicting a subset of Cohorts.
- Each Cohort **incrementally clusters** its data.
  - Adoption of *Adaptive Resonance Theory* (ART).

- Signature is the **set of cluster-heads** of a Cohort's dataset.
- Pythia **collects** all Signatures and stored them **locally**.

Cohort prediction

- Consider an MV imputation request (**input**):
- Pythia **predicts** a Cohort iff the **input** is classified to at least one cluster-head from the Cohort's Signature.
- An **input** is **classified** to a cluster-head iff the Euclidean distance between **non-MVs** is less than a **threshold** (*vigilance* parameter in ART).

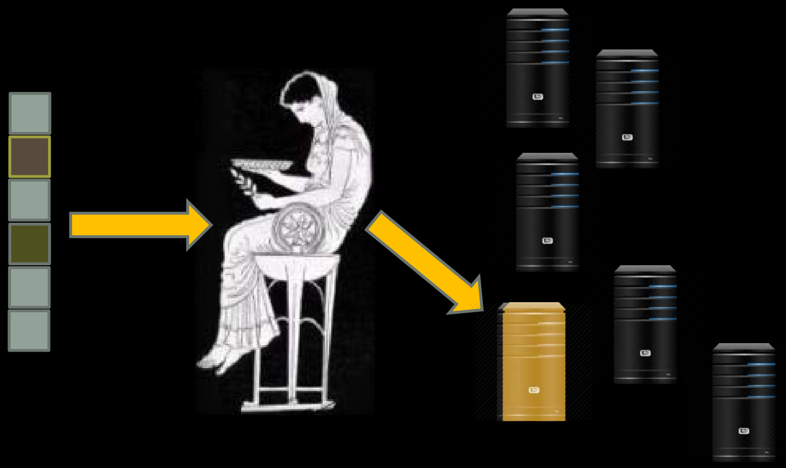
Pythia algorithm



Cost-aware Subset Selection algorithm

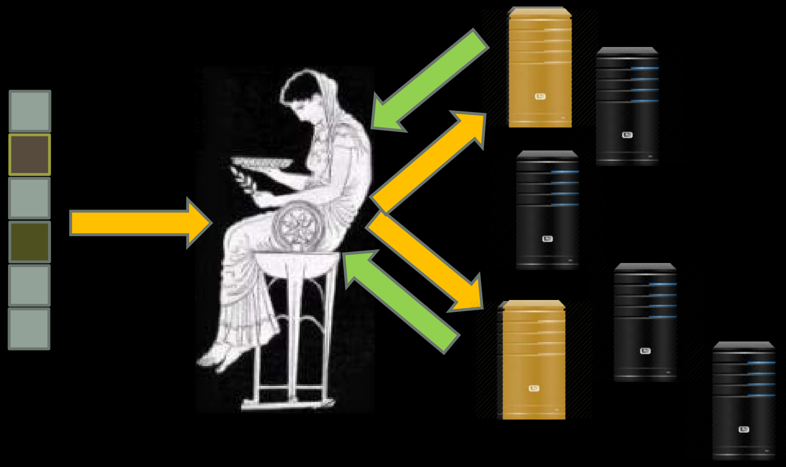
- ...Cohort whose cluster-head is the closest to the input among all **predicted** Cohorts.

- Pythia communicates **only** with this Cohort.



Accuracy-aware Subset Selection algorithm

- Pythia communicates **with each predicted** Cohort.
- Pythia performs a **weighted aggregator operation** over those Cohorts' results which are not assumed as *outliers**.



**outlier* determined by a statistic using the *median* and the *median absolute deviation about the median* of the set of the predicted estimates.

Performance Metrics

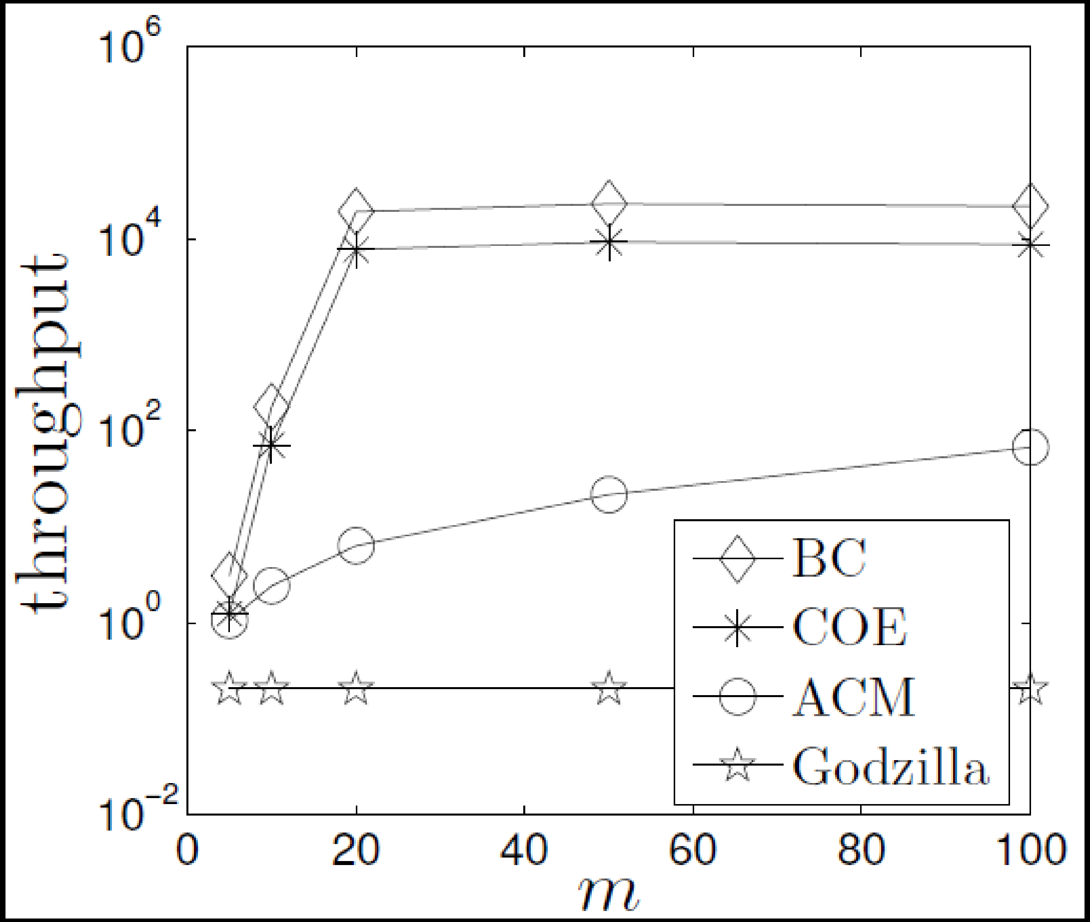
- **Imputation efficiency**
  - **Latency**: the time a system requires to process a MV request.
  - **Speedup**: the ratio of Godzilla latency over Pythia latency.
  - **Throughput**: the rate of imputations delivered by a system.

- **Imputation accuracy**, i.e., RMSE

- **Imputation algorithms** *k*NN (weighted *k*-nearest neighbors)[15]; REG (sequential multivariate regression) [17]

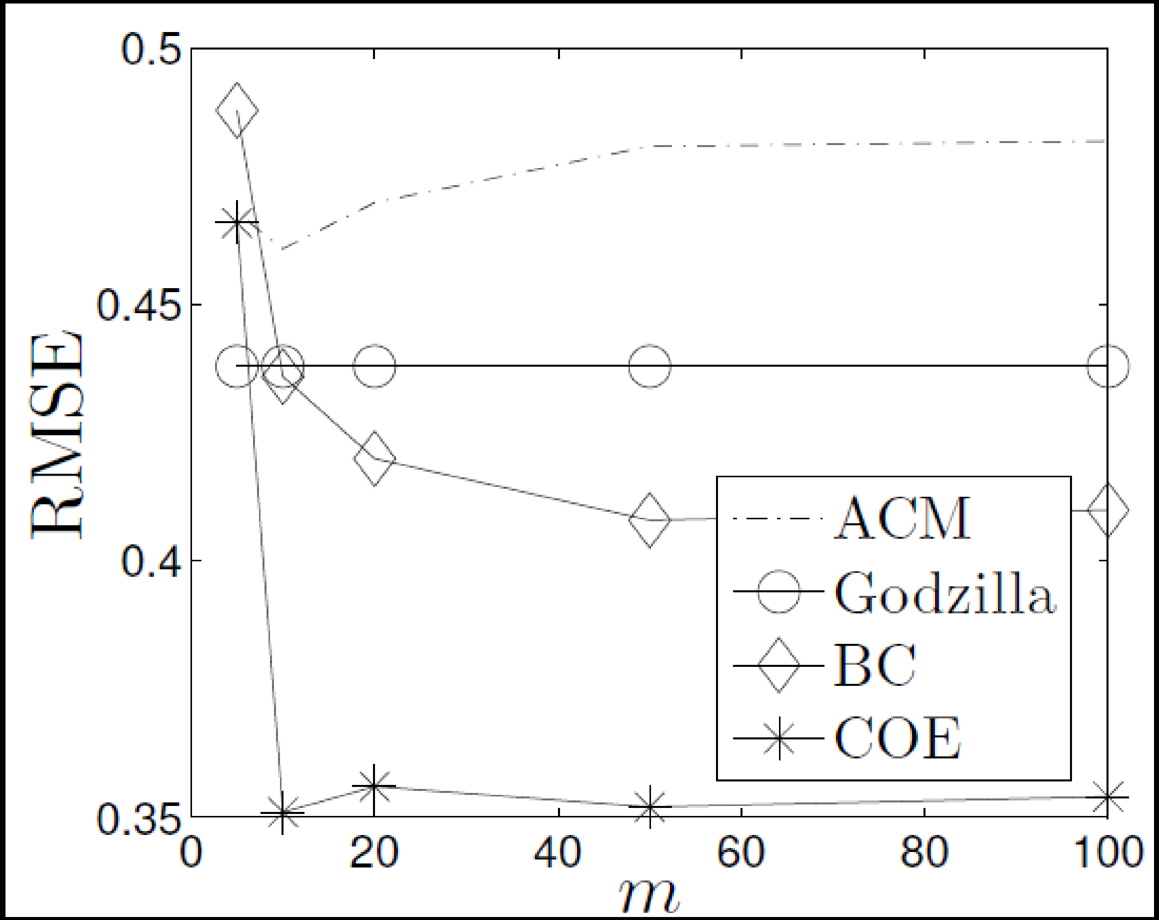
90-dimensional vectors

Rate of imputations



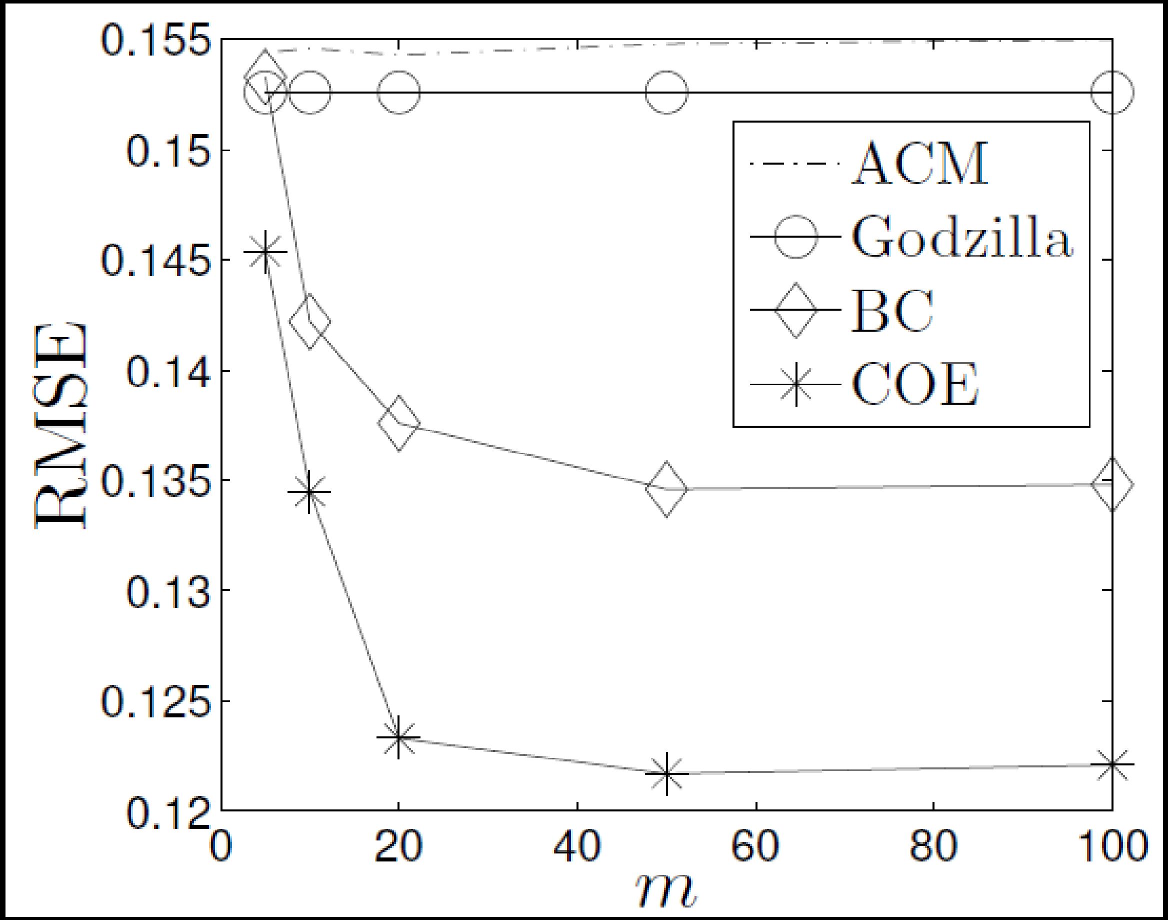
m : number of Cohorts

90-dimensional vectors



m **: number of Cohorts**

384-dimensional vectors

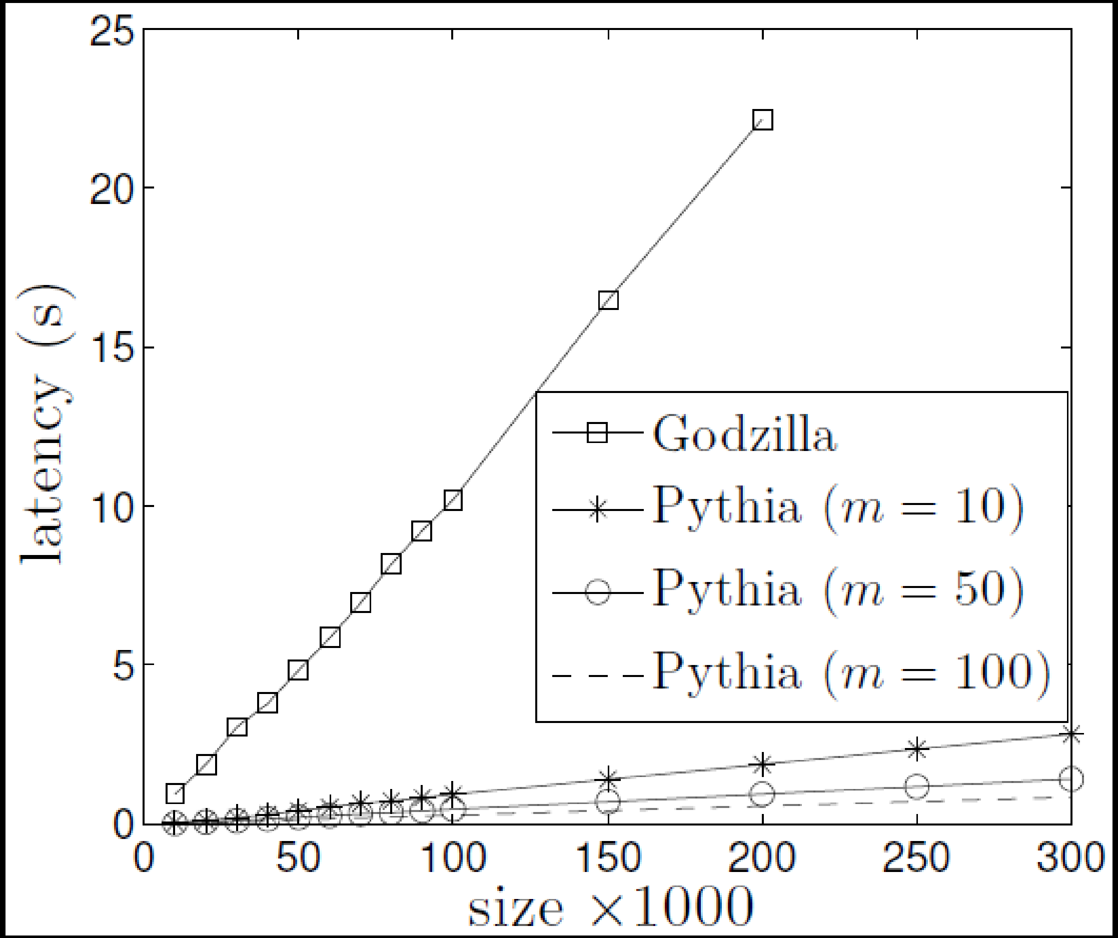


m **: number of Cohorts**

Thank you!

m **: number of Cohorts**

Latency in seconds



Dataset size (90-dimensional vectors)