Benchmarking In The Dark: On The Absence of Comprehensive Edge Datasets

Oleg Kolosov, Gala Yadgar
Technion

Sumit Maheshwari, Emina Soljanin
Rutgers University

TECHNION
Israel Institute
of Technology

RUTGERS

MOTIVATION



Edge

Local services	Susceptive to fluctuations
Distributed	• Multiple providers • Considerable heterogeneity **Optimization not trivial**

Use case: Design and evaluation of an edge-based storage service

- Need a workload
- Important for system research, design, and optimization
  - Define system design objectives
  - Identify optimization goals
  - Make appropriate tradeoffs
  - Evaluate and compare

EXISTING WORKLOADS

- Existing data center workloads rarely reflect
  - Edge infrastructure
  - Edge application requirements
- In existing edge papers :
  - Some aspects are irrelevant
  - Some aspects can be modeled by general datasets
  - Some examples:

Applications	App data is easy to obtain (HotEdge '18, HotEdge '19)
Security & Privacy	System (SEC '16, GLOBECOM '17) and data (IEEE IRI '14, GLOBECOM '16) are trivial
Mobility	Geolocation data is easy to obtain (TON Vol.25, SEC '17)
Infrastructure	System dataset is trivial, synthetic workloads are used (ICDCS '17, MECOMM '17)

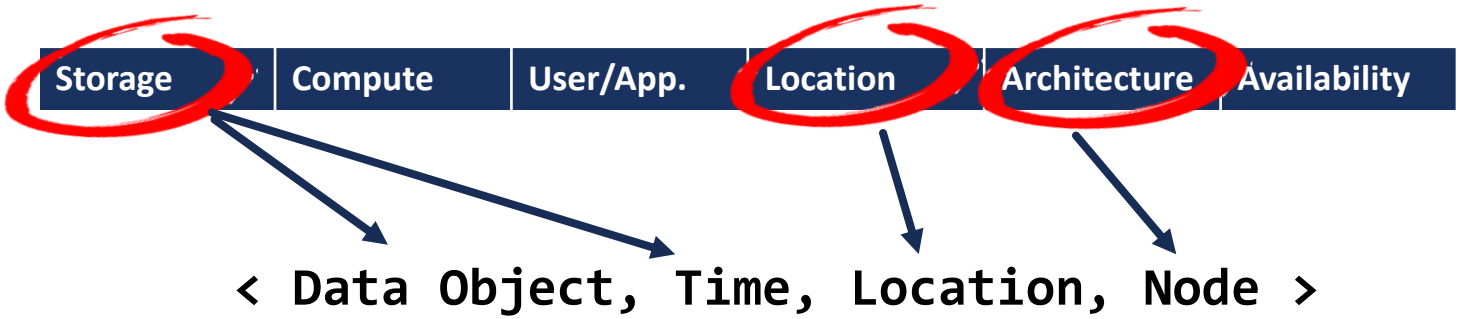
Our use case is focused on storage
Key aspects aren't trivial

There are no operational edge systems that can provide the desired workload

<u>Small number of deployed real edge systems</u>

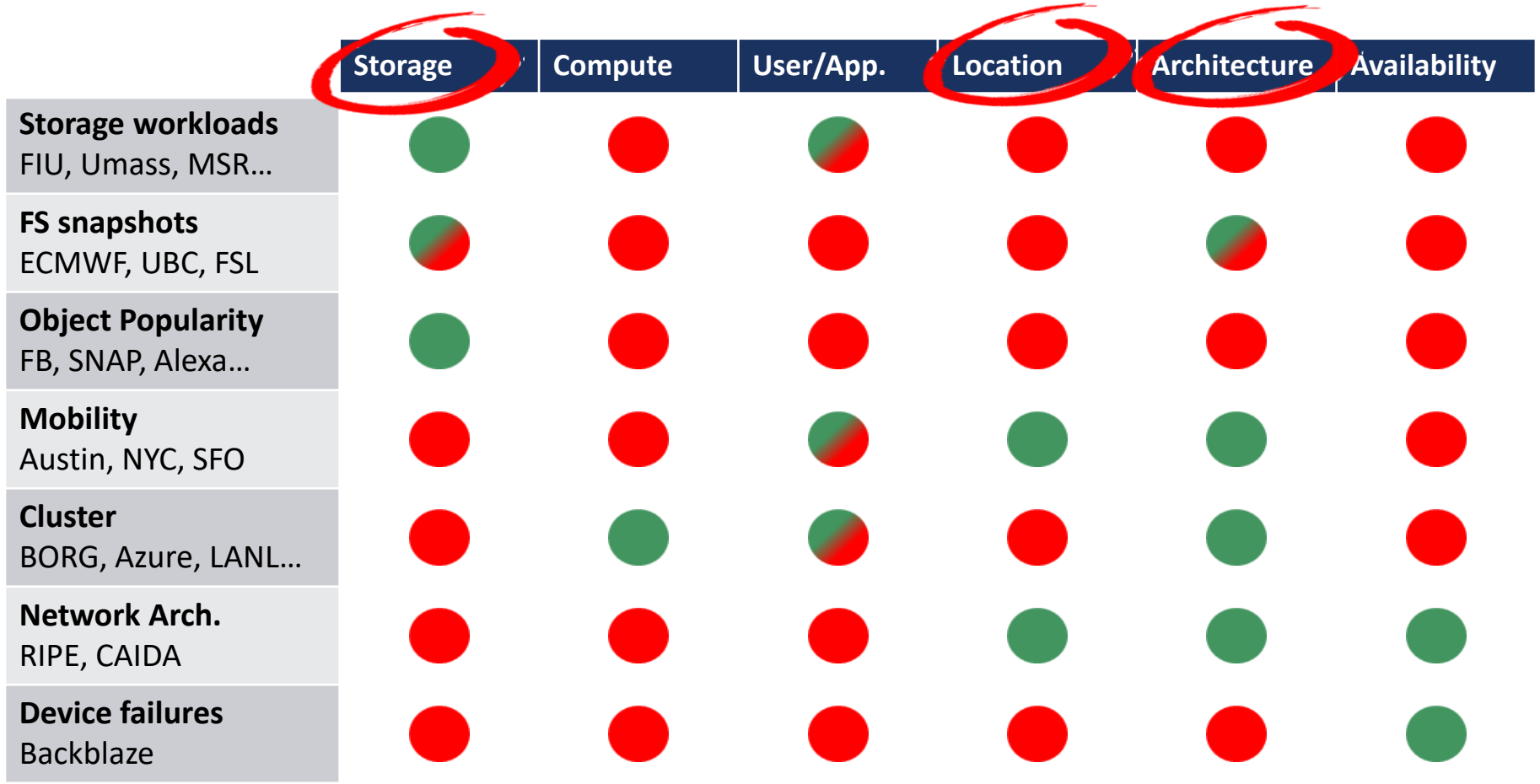
DATASETS AND ATTRIBUTES

The datasets we need:



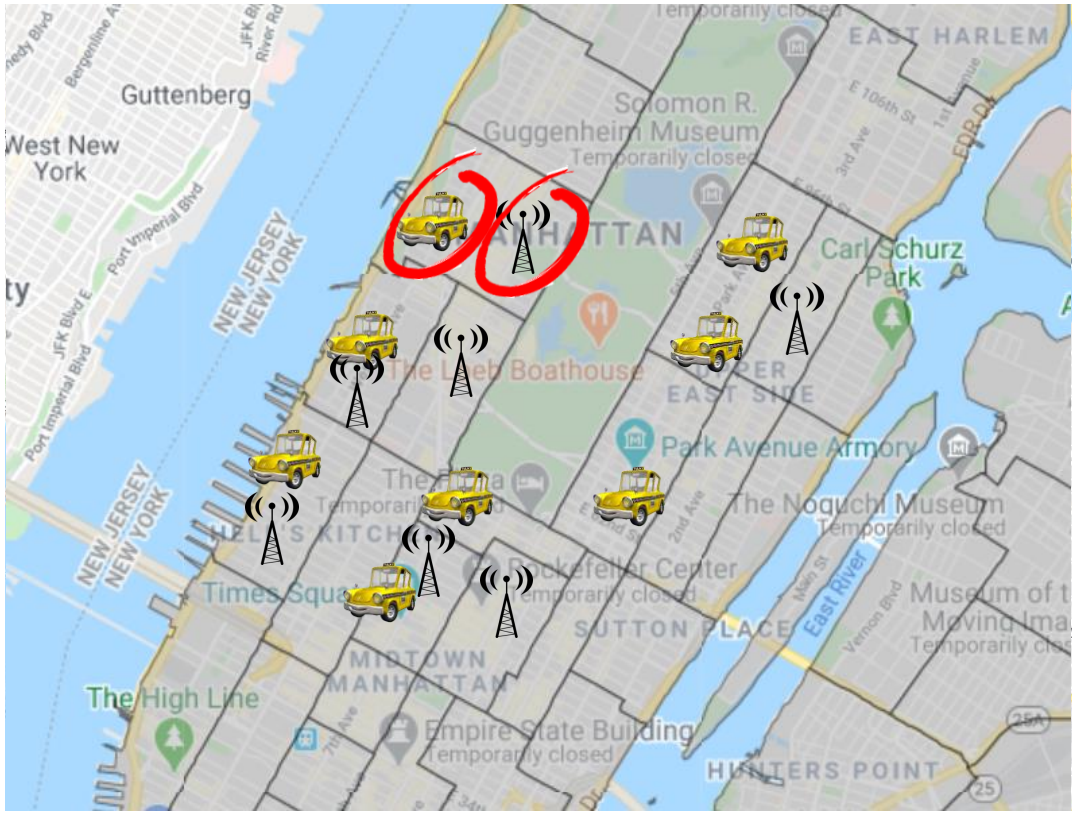
DATASETS AND ATTRIBUTES

The datasets we have:



WORKLOAD COMPOSITION

- How to bridge the gap?
  - Join attributes from several available datasets



NYC Hotspots

NYC Taxi Zones

NYC Yellow Taxis Trip Data

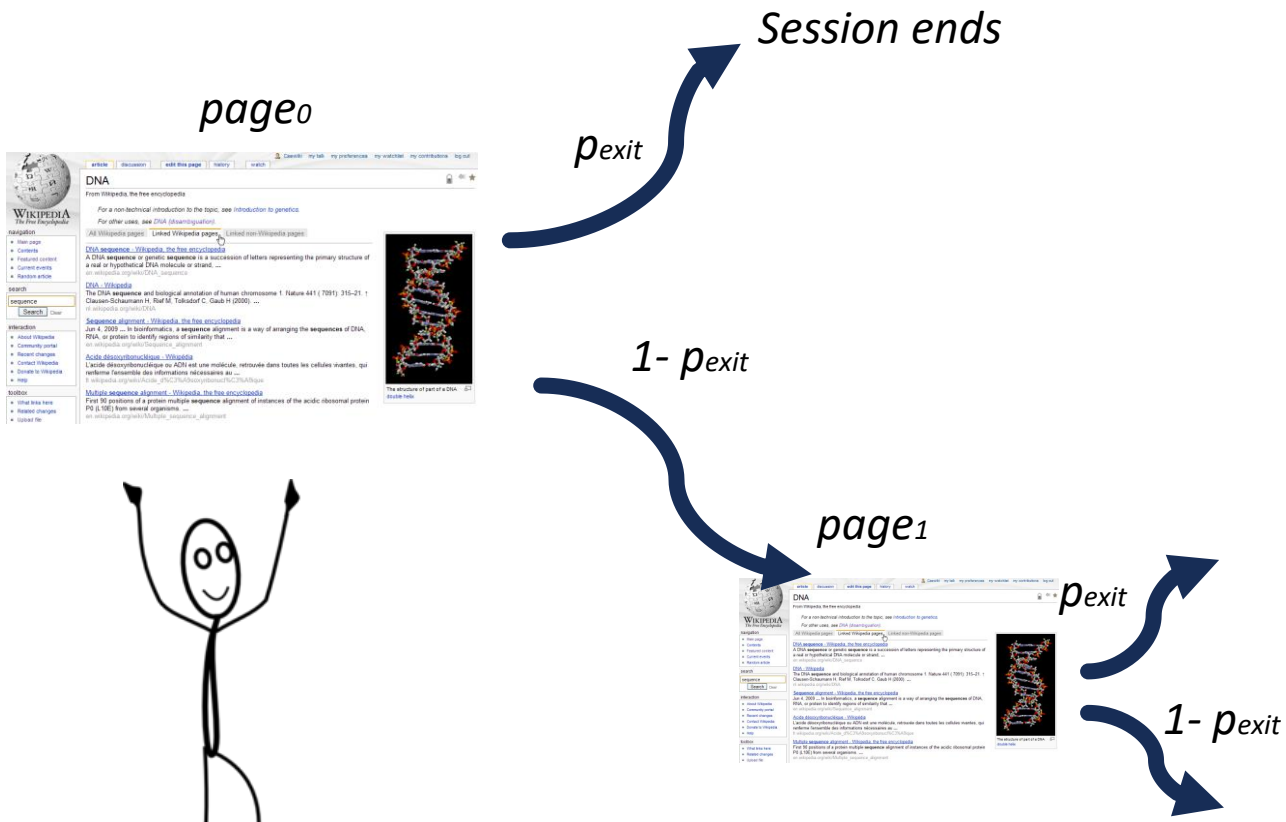
< Data Object, Node, Location, Time >

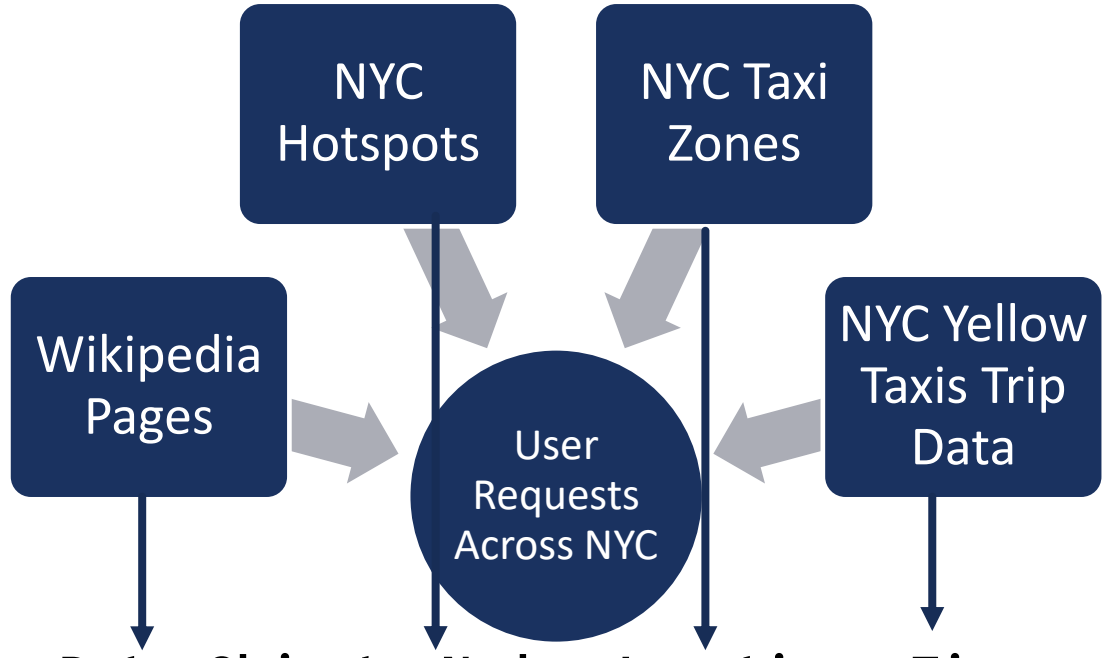
- Taxi drop-offs represent demand in a zone
  - A 'browsing session' starts at a drop-off time and zone
  - Starts at drop-off *node* - Random hotspot from the drop-off zone

Use case: Design and evaluation of an edge-based storage service

WORKLOAD COMPOSITION

The ‘browsing session’





< Data Object, Node, Location, Time >

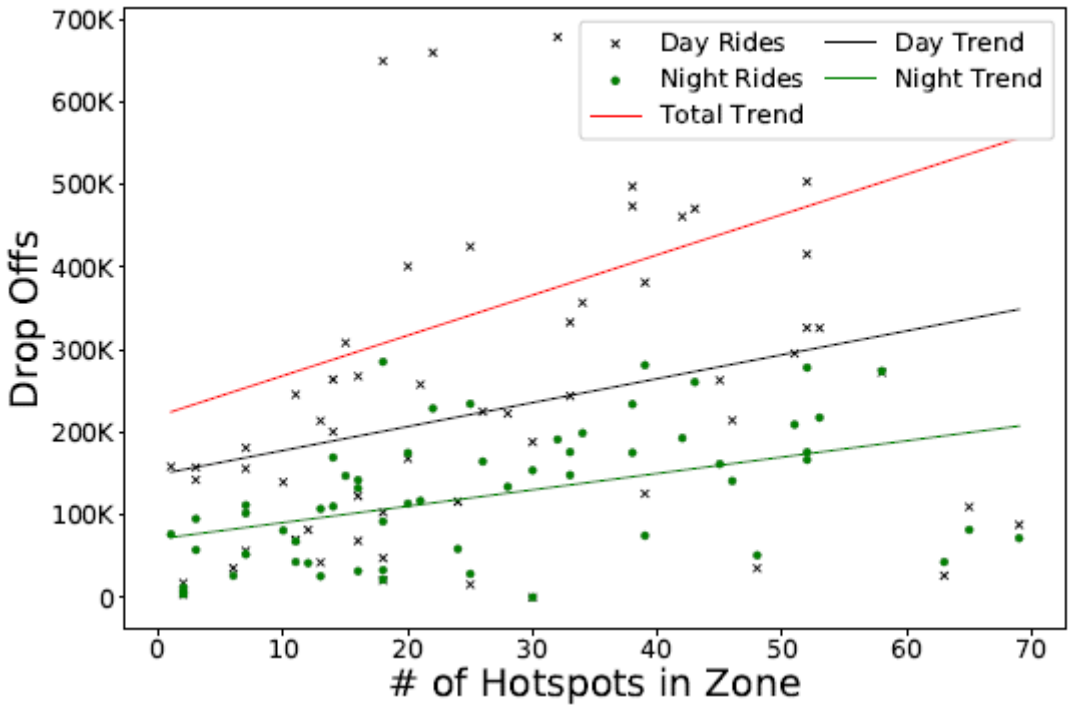
- Session of *n* pages, Drop-off at time *T*

Trace of GET requests:

< $page_i$, $node_h$, $location_j$, $T+i\times\varepsilon$ >

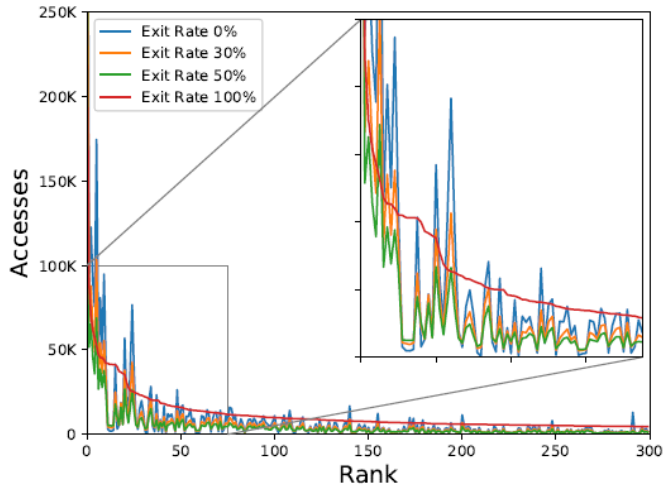
for $0 \le i < n$. ε – request rate within a session.

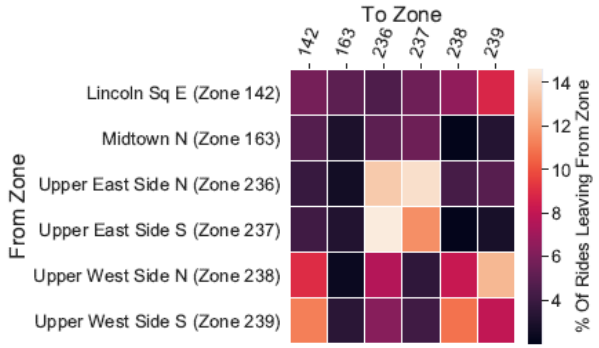
CHARACTERIZING THE SYSTEM AND ITS USERS



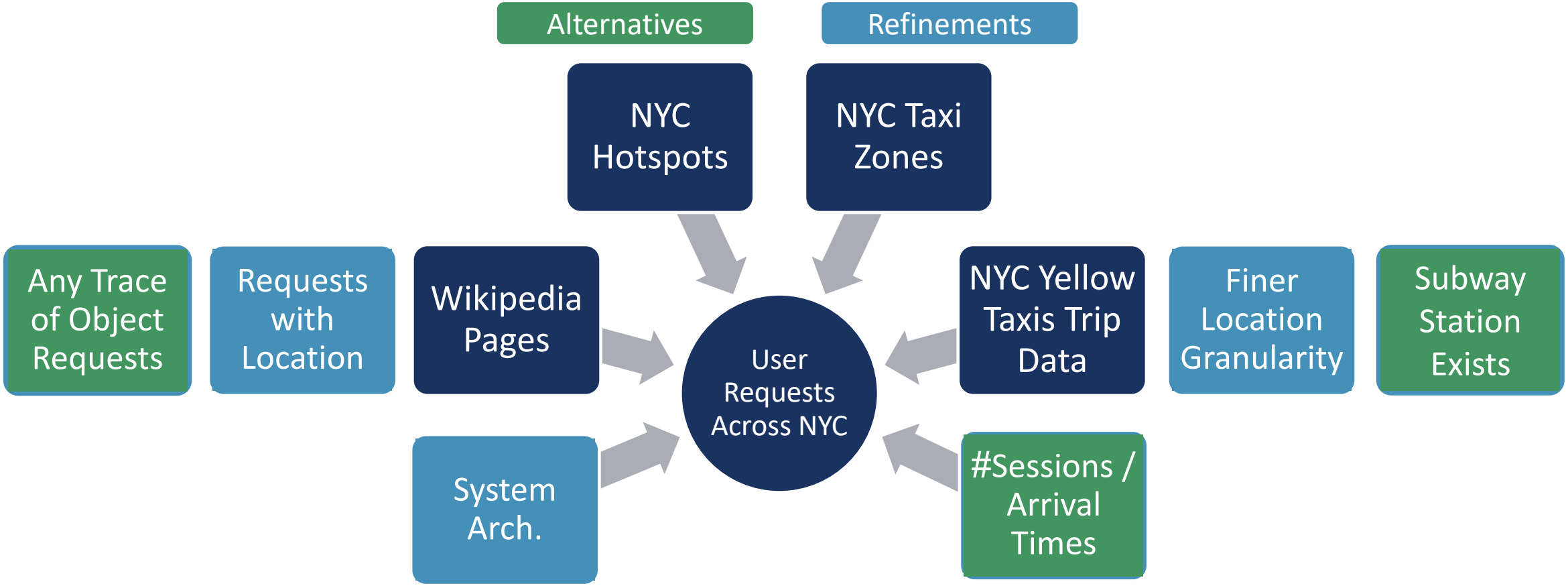
- The workloads are lightly correlated
- The workload composition is not random

Additional characterizations



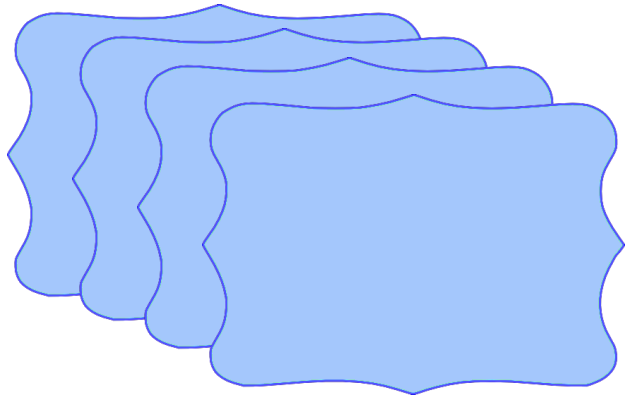


GENERALIZATION



SUMMARY

- Conclusions
  - The problem is not unique for this specific case (general problem)
  - Described important categories of attributes
  - Showed how partial datasets can be used to compose a workload



- Discussion
  - Is the absence of datasets really temporary?
  - Which basic workloads to use?
  - How can we leverage synthetic distributions?
  - How to generate realistic and useful compositions?

Thank you