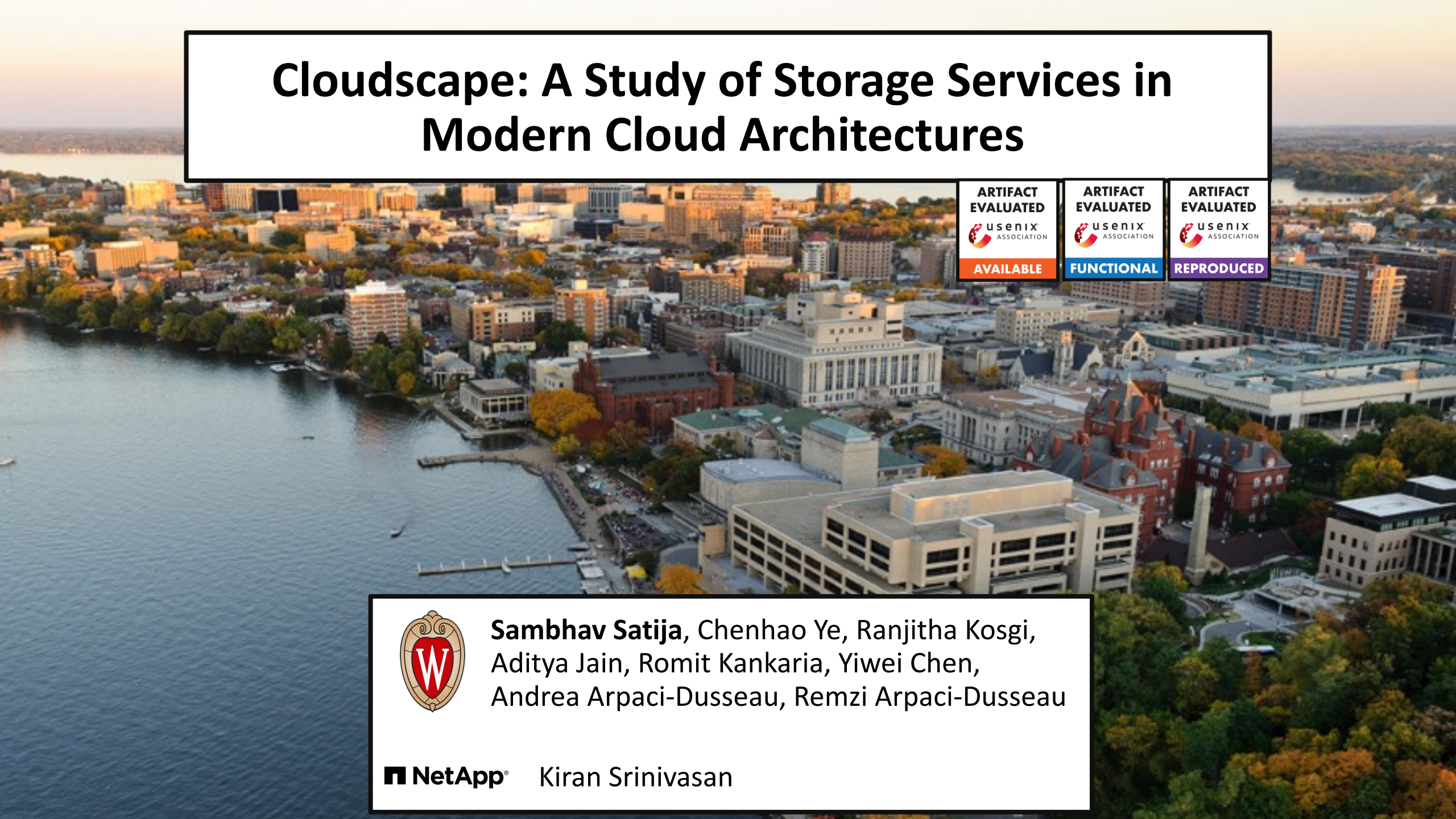
Cloudscape: A Study of Storage Services in Modern Cloud Architectures



Sambhav Satija, Chenhao Ye, Ranjitha Kosgi, Aditya Jain, Romit Kankaria, Yiwei Chen, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau

NetApp® Kiran Srinivasan

Motivation: Usage of Cloud Services

Most companies build on cloud[1]

$600 Billion dollar market size[2]

4 million companies use AWS[3]


Netflix

SAMSUNG


Snapchat

No dataset captures real-world architectures

[1] StackOverflow: 2024 Developer Survey
[2] Fortune Business Insights: Cloud Computing 2024
[3] HGInsights: AWS Market Share 2025

Motivation: Usage of Cloud Services

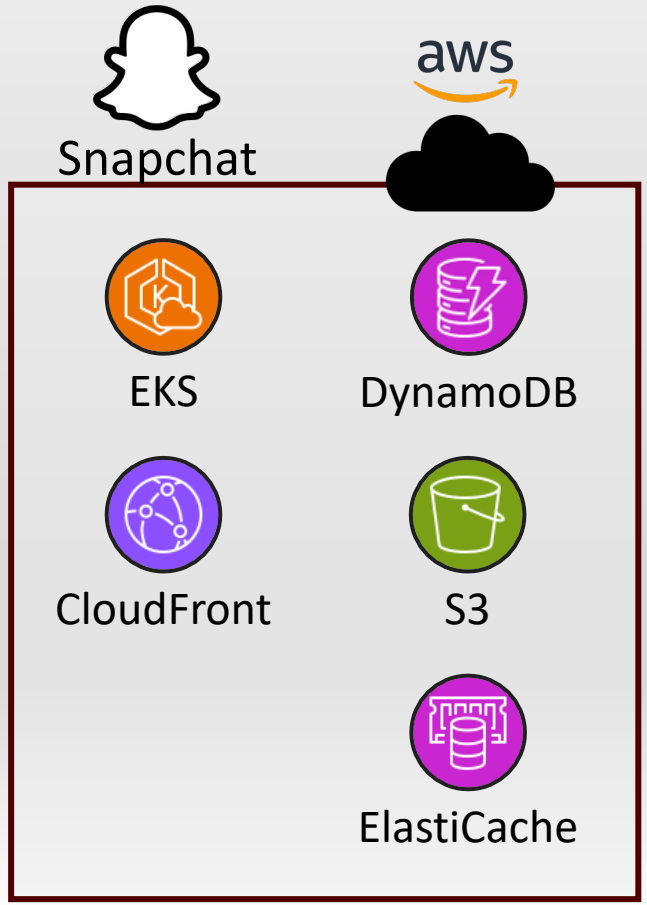
Most companies build on cloud[1]

$600 Billion dollar market size[2]

4 million companies use AWS[3]

No dataset captures real-world architectures

Cloud architectures are heterogenous



[1] StackOverflow: 2024 Developer Survey
[2] Fortune Business Insights: Cloud Computing 2024
[3] HGInsights: AWS Market Share 2025

Motivation: Usage of Cloud Services

Most companies build on cloud[1]

$600 Billion dollar market size[2]

4 million companies use AWS[3]

No dataset captures real-world architectures

Cloud architectures are heterogenous

Motivation: Need dataset about usage of cloud services to inform research directions

Goal: Build this dataset

[1] StackOverflow: 2024 Developer Survey
[2] Fortune Business Insights: Cloud Computing 2024
[3] HGInsights: AWS Market Share 2025

Data Source: YouTube Videos

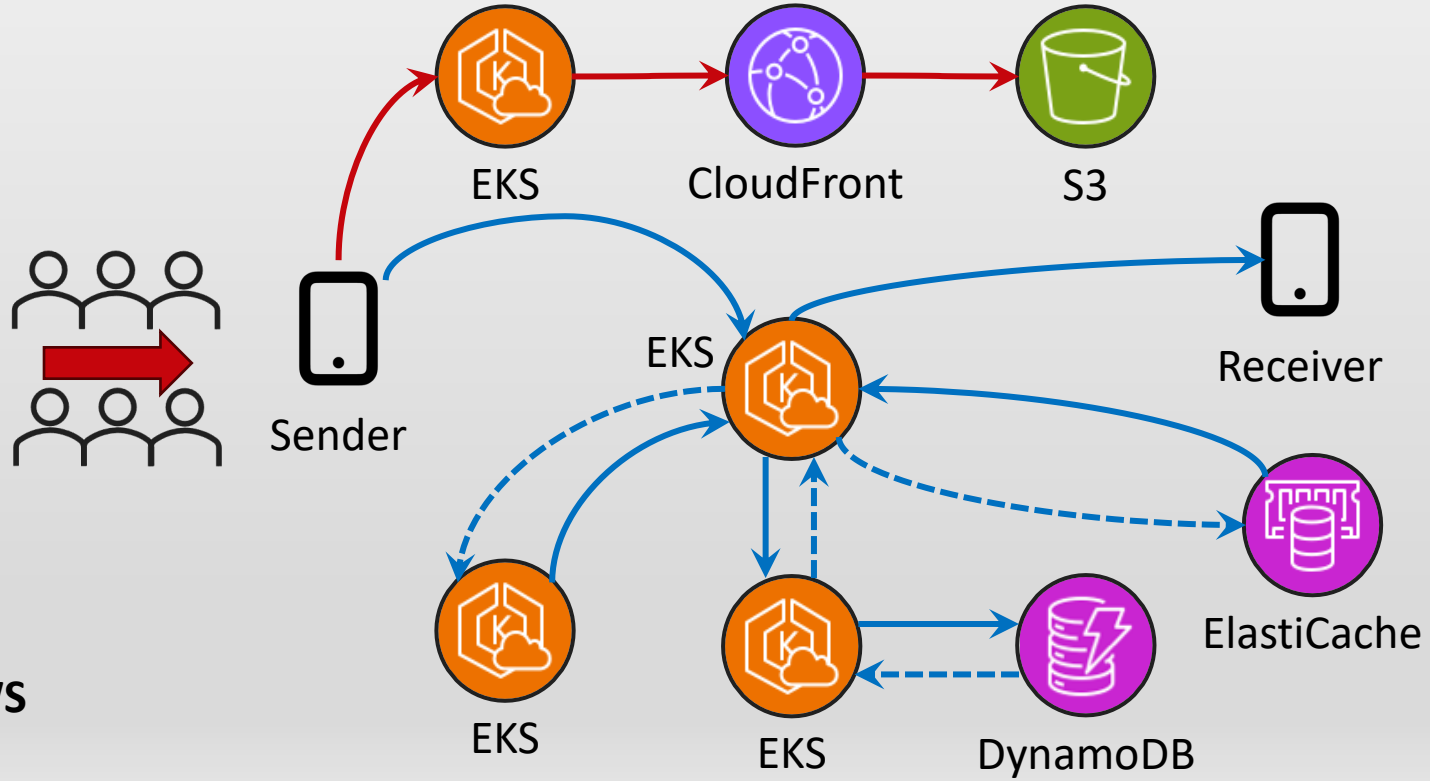
Playlist from AWS: “This is My Architecture”

Snap: Journey of a Snap on Snapchat Using AWS

Methodology: Unstructured to Structured

Playlist from AWS: “This is My Architecture”

Snap: Journey of a Snap on Snapchat Using AWS



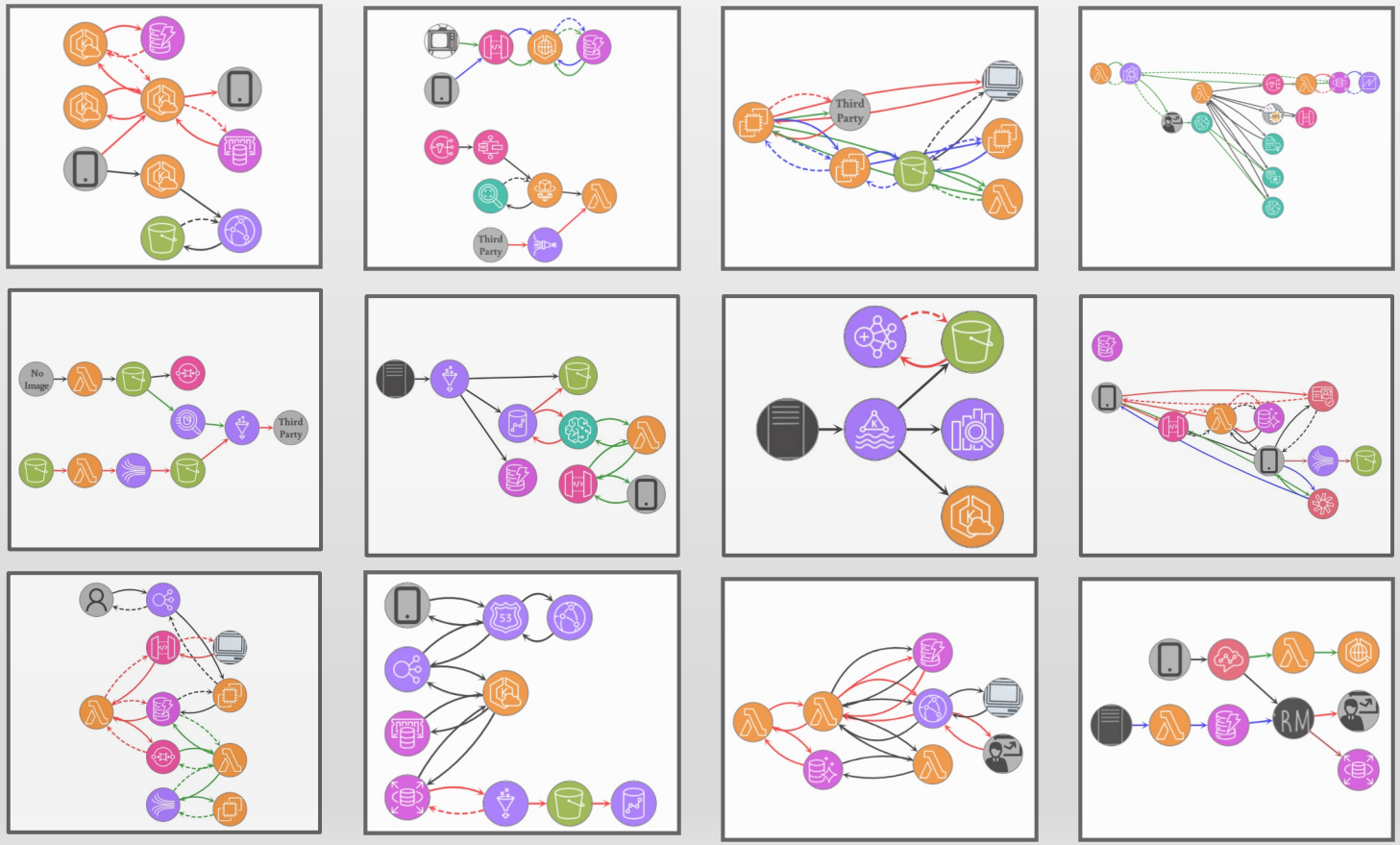
Cloudscape: Dataset of Real-World Architectures

Nearly 400 cloud architectures

380 different companies

Architectures from 2019-2023

134 unique AWS services



Results at a Glance

- **The storage layer of cloud architectures is**
  - **diverse:** 14 different storage services
  - **heterogeneous:** majority architectures use 2+ storage services
- **Popularity of storage services**
  - S3 is most popular (68%); distributed file systems are not popular (4%)
- **What data do services store?**
  - **S3:** data with different availability requirements
  - **DynamoDB:** data extracted from richer S3 object; cross-referenced data
- **Other services interacting with storage**
  - Mostly compute services; chiefly Lambda (40%)

Many more in the paper!

Outline

- Introduction
- **Methodology**
- Example architecture in Cloudscape
- Four findings about the storage layer
  - Composition of the storage layer
  - Popularity of storage services
  - Content of storage services
  - Services interacting with storage

Data Source

- Videos are unstructured
- Dense 6 minute videos (on average)
- Information spread across audio / visual
- Diverse contextual information

Goal: Reliably extract common types of data

The Washington Post

Nielsen

Methodology: Iterative Coding

Coding: ~~programming~~ “process of assigning a code for classification and identification”

Derived from Collaborative Qualitative Coding techniques[1]

1. Coders independently code to determine granularity of information to extract

3 coders independently coded the same 5 videos

[1] Collaborative Qualitative Coding, University of Colorado Boulder

Methodology: Iterative Coding

Coding: ~~programming~~ “process of assigning a code for classification and identification”

Derived from Collaborative Qualitative Coding techniques[1]

1. Coders independently code to determine granularity of information to extract
2. Discuss and repeat for larger set of architectures

Achieve consensus of rules for extracting information.

Independently code 15 architectures.

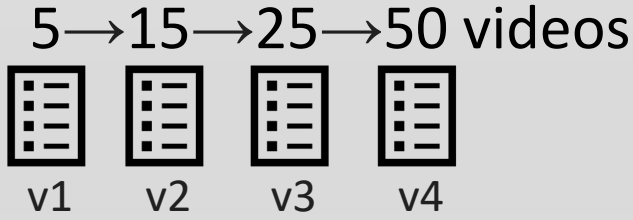
[1] Collaborative Qualitative Coding, University of Colorado Boulder

Methodology: Iterative Coding

Coding: ~~programming~~ “process of assigning a code for classification and identification”

Derived from Collaborative Qualitative Coding techniques[1]

1. Coders independently code to determine granularity of information to extract
2. Discuss and repeat for larger set of architectures
3. Repeat 1,2 for more architectures



[1] Collaborative Qualitative Coding, University of Colorado Boulder

Methodology: Iterative Coding

Coding: ~~programming~~ “process of assigning a code for classification and identification”

Derived from Collaborative Qualitative Coding techniques[1]

1. Coders independently code to determine granularity of information to extract
2. Discuss and repeat for larger set of architectures
3. Repeat 1,2 for more architectures
4. Total 6 coders. Each new coder independently coded few videos and helped solidify annotations. Initial additions were verified by multiple coders.

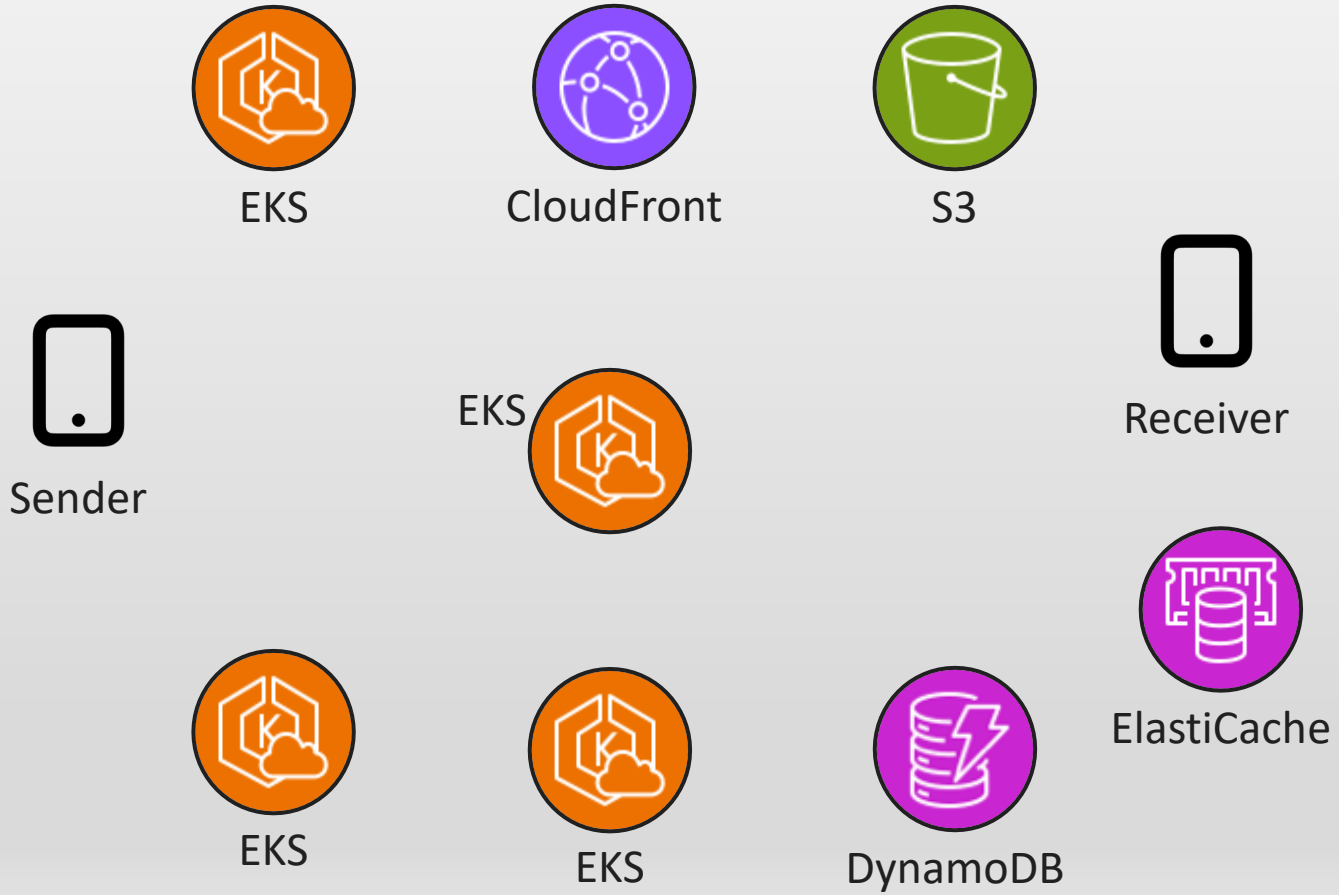
[1] Collaborative Qualitative Coding, University of Colorado Boulder

Outline

- Introduction
- Methodology
- **Example architecture in Cloudscape**
- Four findings about the storage layer
  - Composition of the storage layer
  - Popularity of storage services
  - Content of storage services
  - Services interacting with storage

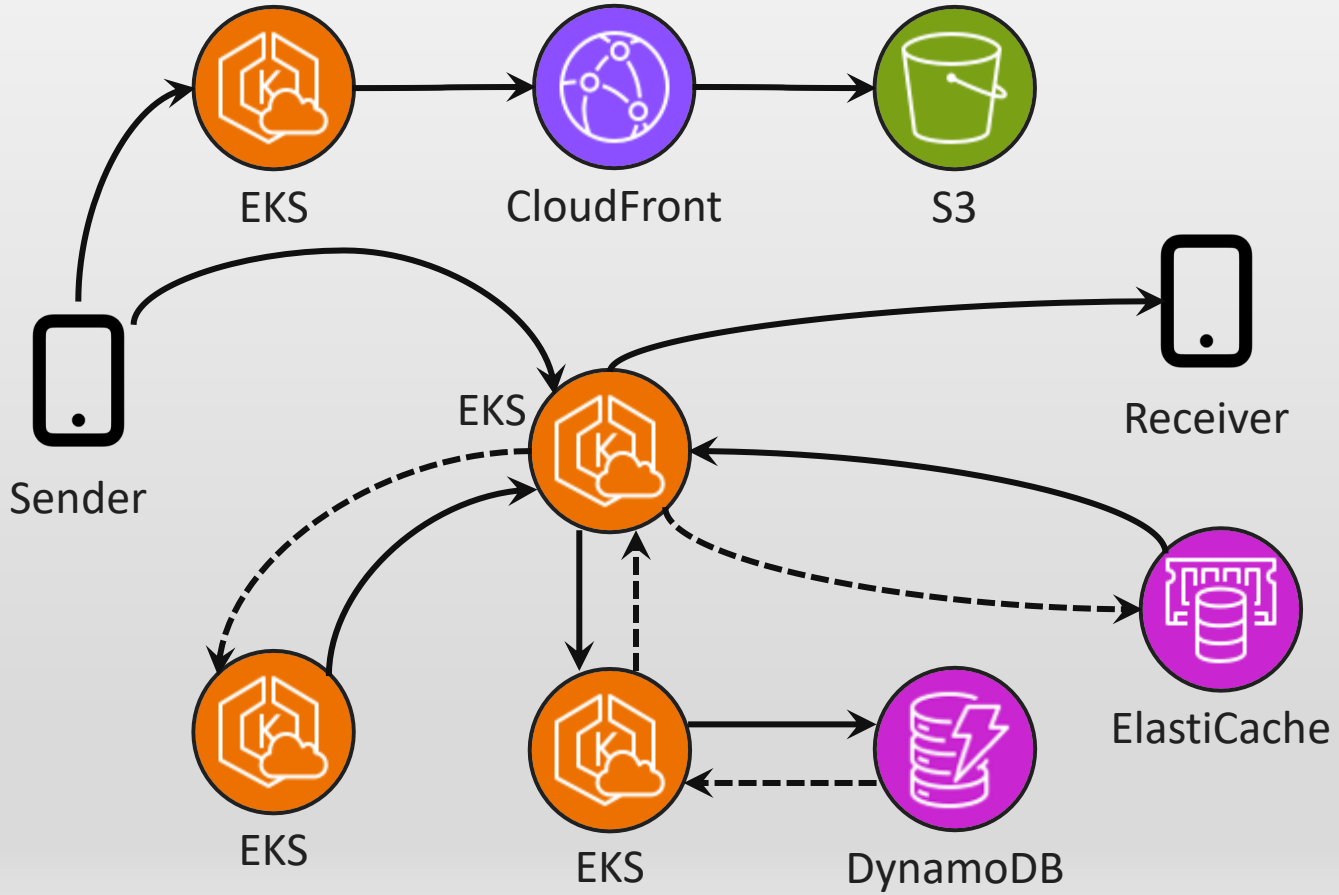
Example Annotated Architecture: Snapchat

- Services *(nodes)*



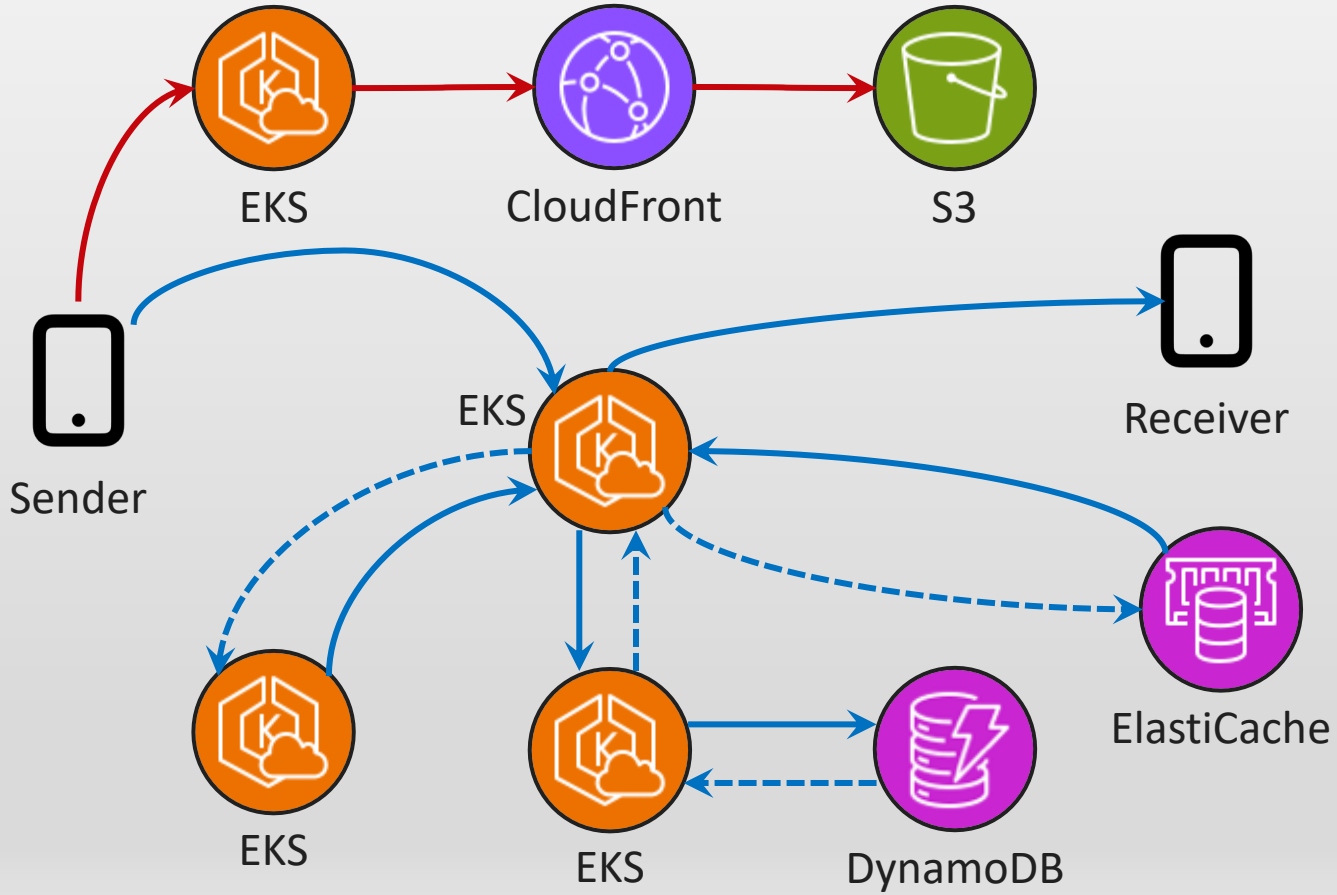
Example Annotated Architecture: Snapchat

- Services *(nodes)*
- Interactions *(edges)*
  - Data edges
  - Meta edges



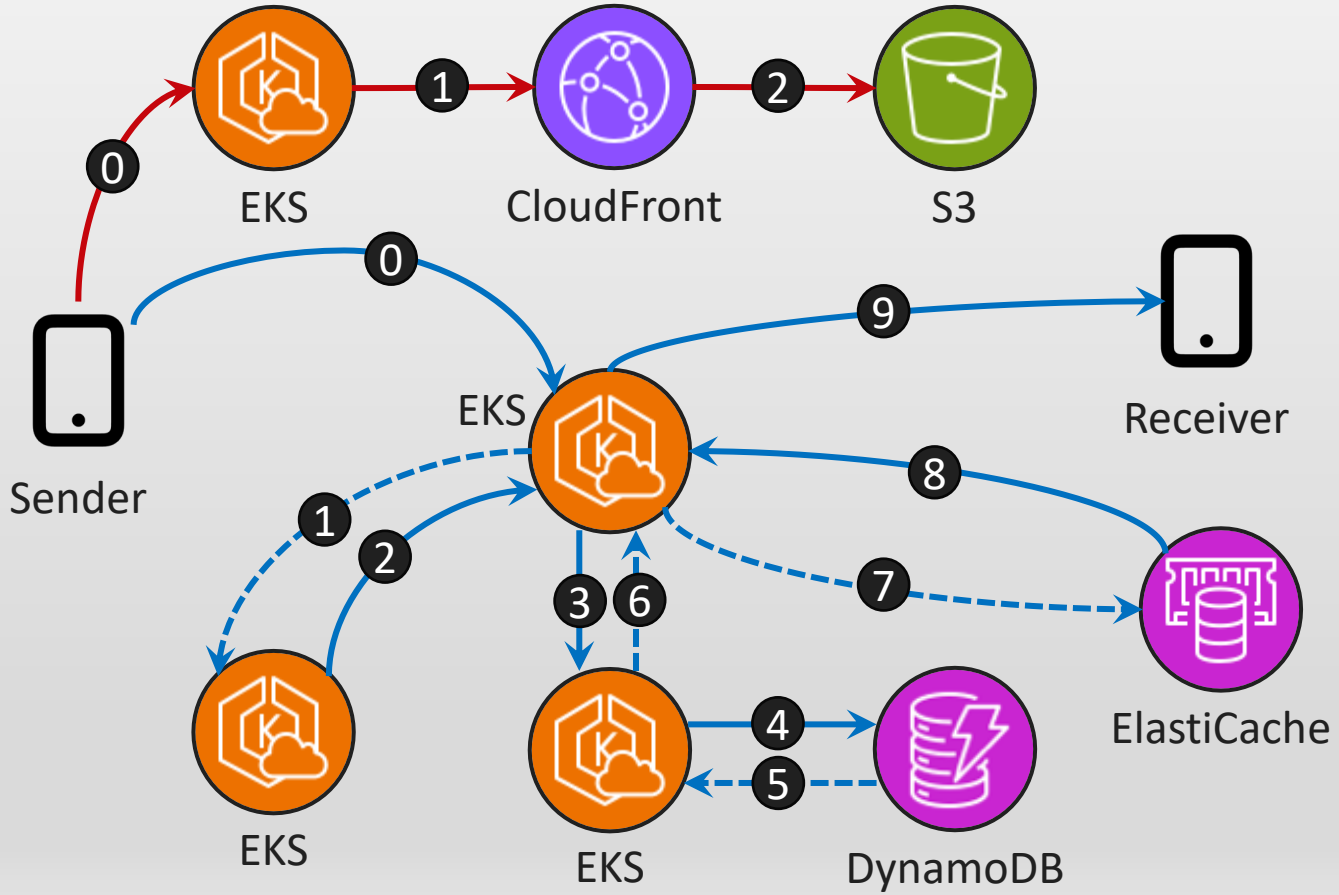
Example Annotated Architecture: Snapchat

- Services *(nodes)*
- Interactions *(edges)*
  - Data edges
  - Meta edges
- Workflows (*edge colors*)



Example Annotated Architecture: Snapchat

- Services *(nodes)*
- Interactions *(edges)*
  - Data edges
  - Meta edges
- Workflows (*edge colors*)
- Sequence ⓿❶❷



Outline

- Introduction
- Methodology
- Dataset description
- Four findings about the storage layer
  - **Composition of the storage layer**
  - Popularity of storage services
  - Content of storage services
  - Services interacting with storage

Composition of Storage Layer

- Storage layer is **diverse:** 14 different storage services

1. Aurora
2. DocumentDB
3. DynamoDB
4. EBS
5. EFS
6. ElastiCache
7. FSX
8. MemoryDB
9. MediaStore
10. Neptune
11. RDS
12. RedShift
13. S3
14. Timestream

Composition of Storage Layer

- Storage layer is **diverse:** 14 different storage services

- 10% architectures do not use any storage service
- 38% architectures use only 1 storage service

1. Aurora
2. DocumentDB
3. DynamoDB
4. EBS
5. EFS
6. ElastiCache
7. FSX
8. MemoryDB
9. MediaStore
10. Neptune
11. RDS
12. RedShift
13. S3
14. Timestream

Number of storage services used	Percentage of architectures
0	**10%**
1	**38%**
2	39%
3	11%
4	1%
5	1%

Composition of Storage Layer

- Storage layer is **diverse:** 14 different storage services

- 10% architectures do not use any storage service
- 38% architectures use only 1 storage service

- Storage layer is **heterogeneous**: **52%** architectures use 2 or more storage services

1. Aurora
2. DocumentDB
3. DynamoDB
4. EBS
5. EFS
6. ElastiCache
7. FSX
8. MemoryDB
9. MediaStore
10. Neptune
11. RDS
12. RedShift
13. S3
14. Timestream

Number of storage services used	Percentage of architectures
0	10%
1	38%
2	**39%**
3	**11%**
4	**1%**
5	**1%**

Composition of Storage Layer

Summary: Cloud architectures offload storage needs to specialized services

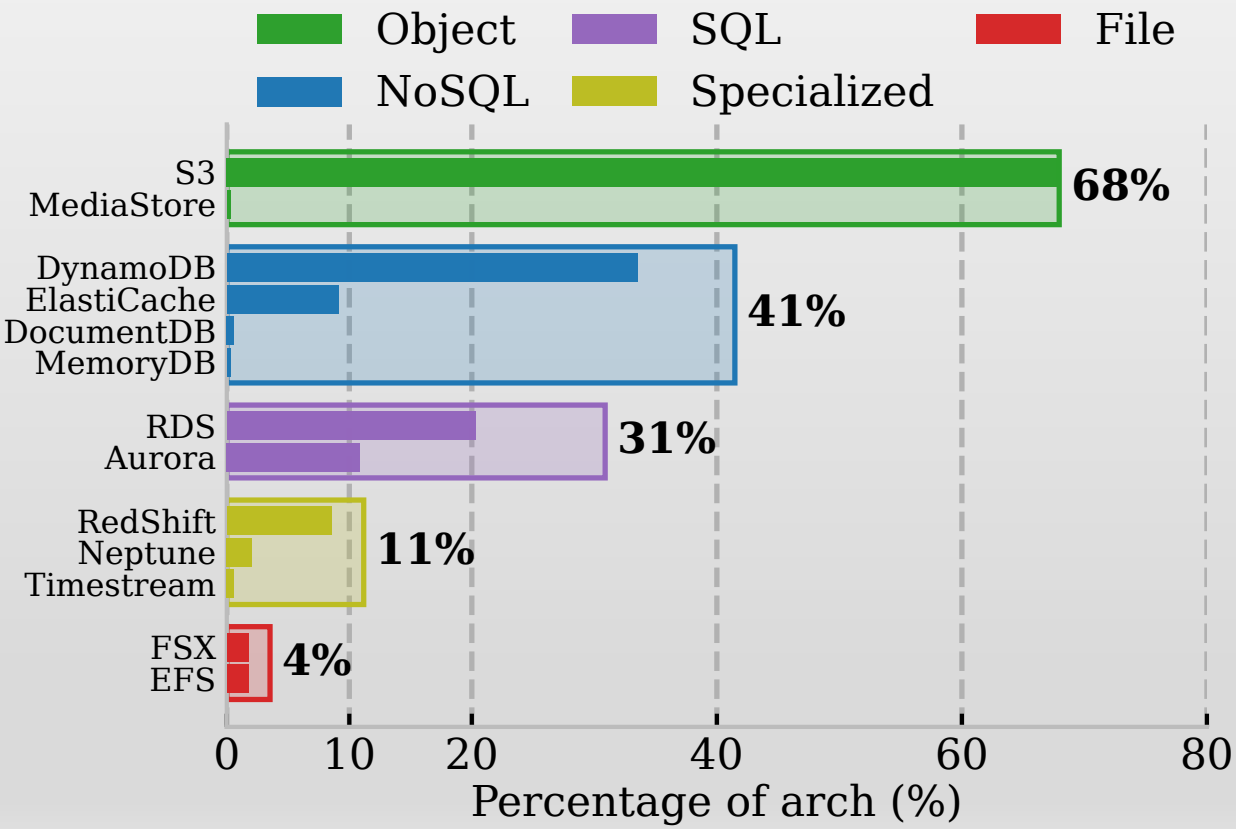
Implication: Focus on cloud-native and multi-tenant storage services

Outline

- Introduction
- Methodology
- Dataset description
- Four findings about the storage layer
  - Composition of the storage layer
  - **Popularity of storage services**
  - Content of storage services
  - Services interacting with storage

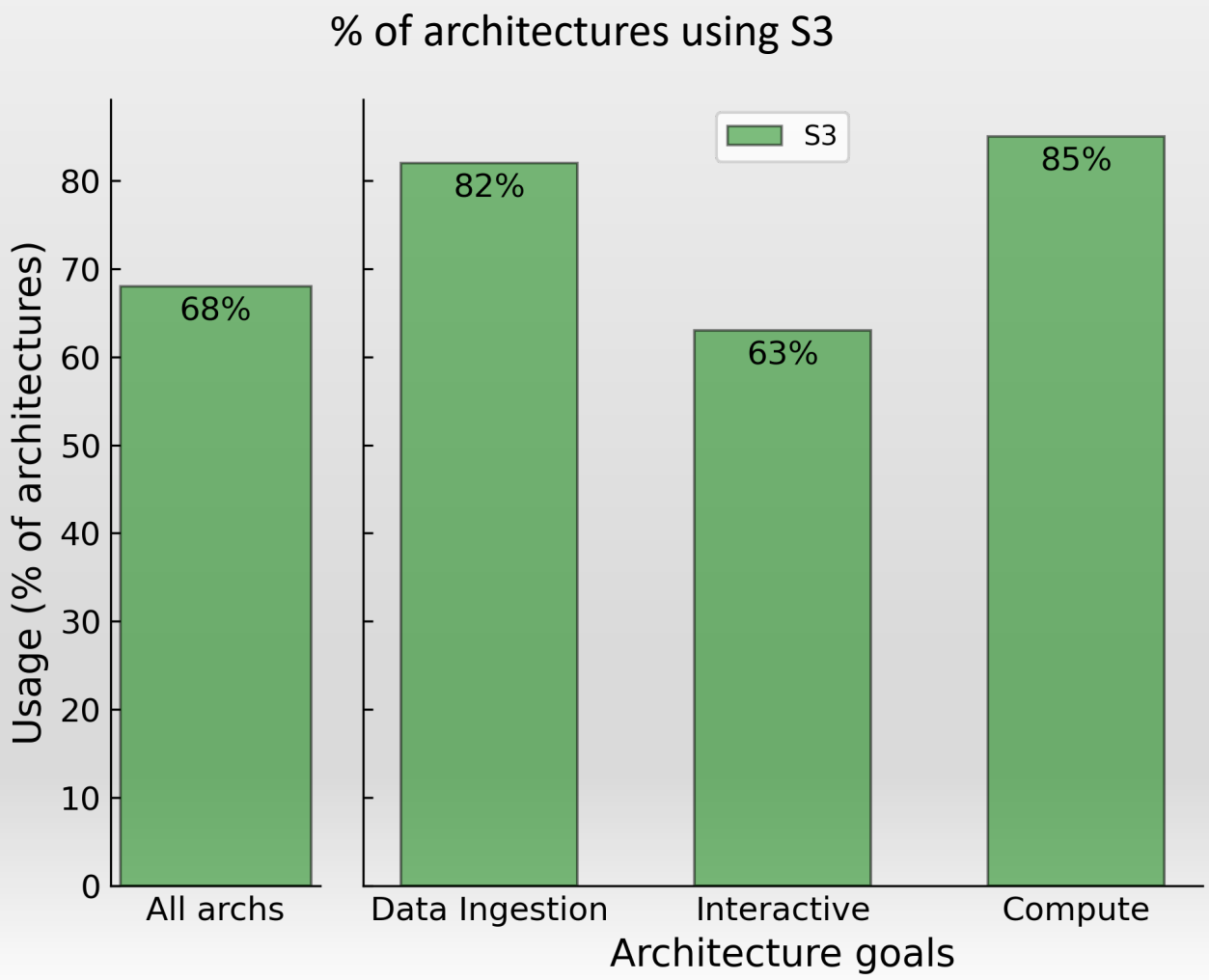
Popularity of Storage Services

- Group services based on schema
- S3 is used in 68% architectures
- S3 is twice as popular as next service, DynamoDB
- Distributed filesystems are not important in this context



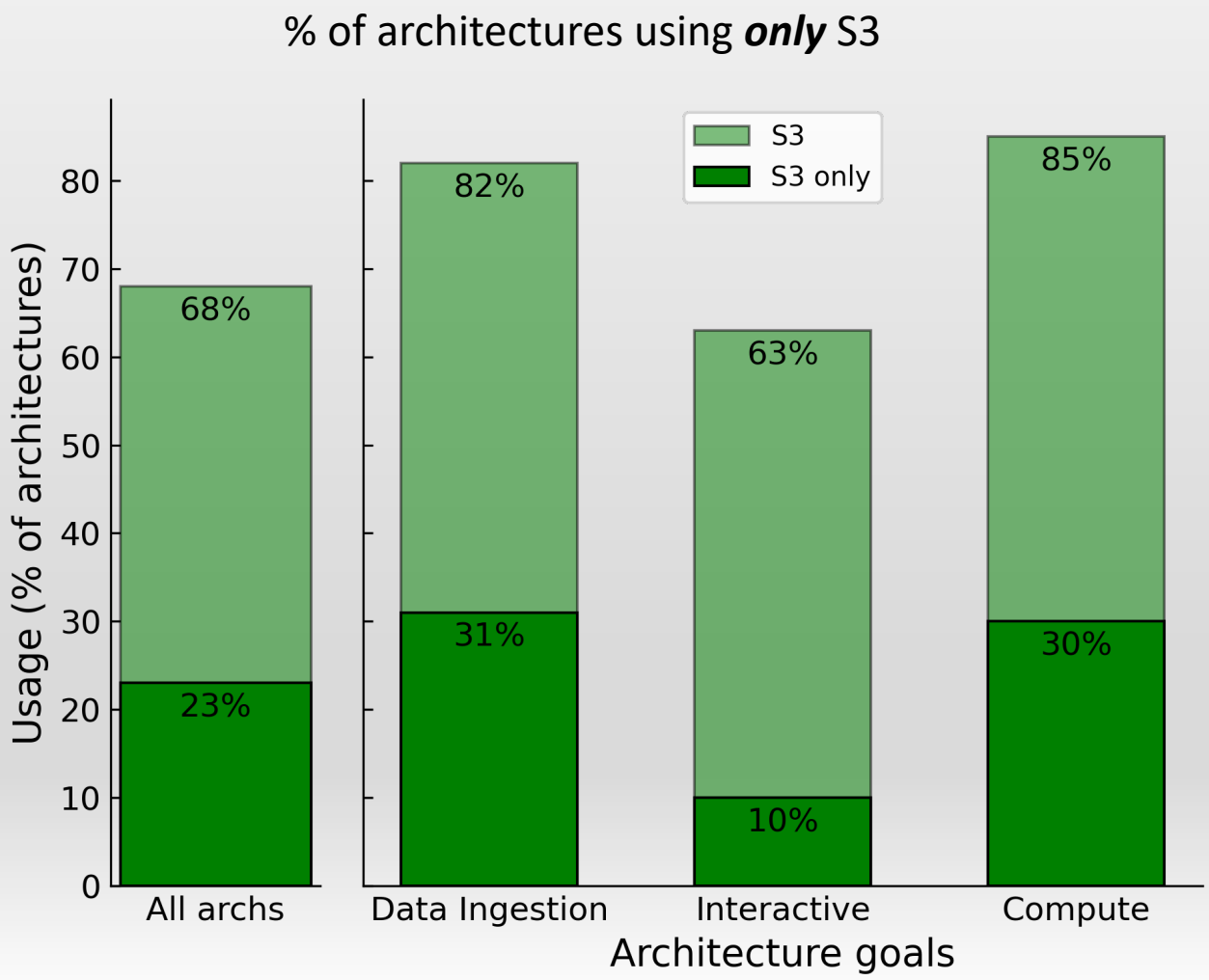
Popularity of Storage Services | Understanding S3

- S3 appears in 68% architectures
- Adopted across all architecture *goals*



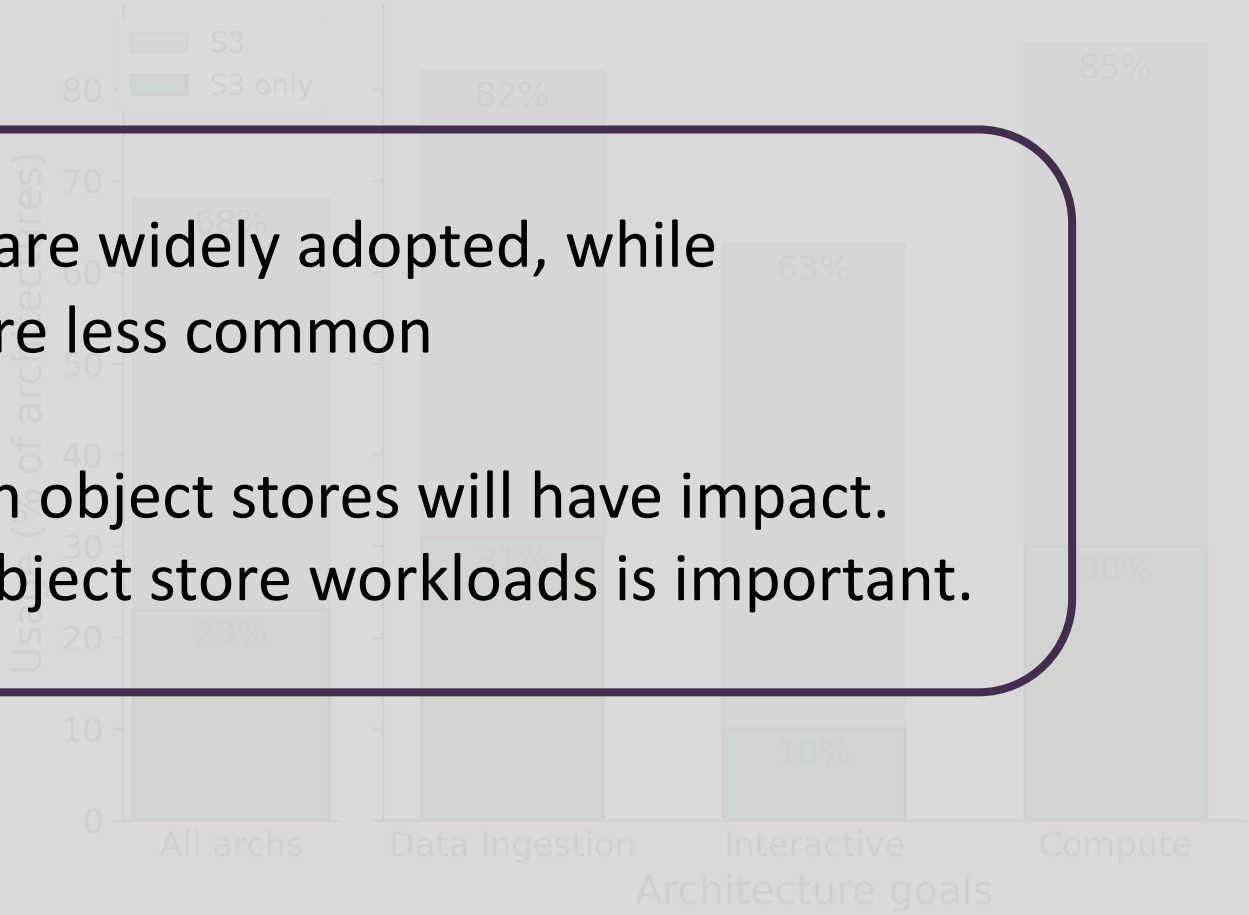
Popularity of Storage Services | Understanding S3

- S3 appears in 68% architectures
- Adopted across all architecture *goals*
- Suffices as the only storage service for 23% architectures



Popularity of Storage Services

- S3 appears in 68% architectures
- Adopted across all architecture *goals*
- Suffices as only storage service for 23% architectures



Summary: Object stores are widely adopted, while distributed file systems are less common

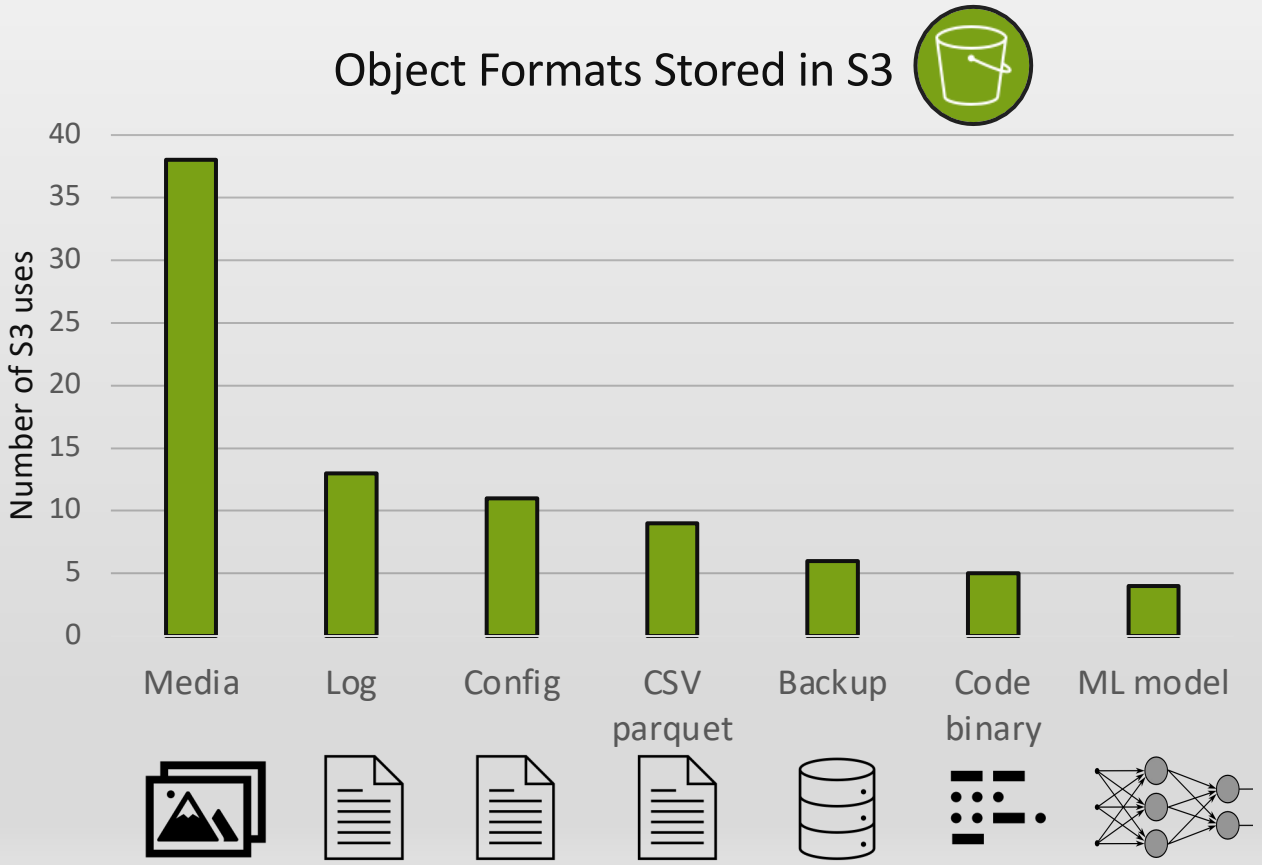
Implications: Research on object stores will have impact. Understanding realistic object store workloads is important.

Outline

- Introduction
- Methodology
- Dataset description
- Four findings about the storage layer
  - Composition of the storage layer
  - Popularity of storage services
  - **Content of storage services**
  - Services interacting with storage

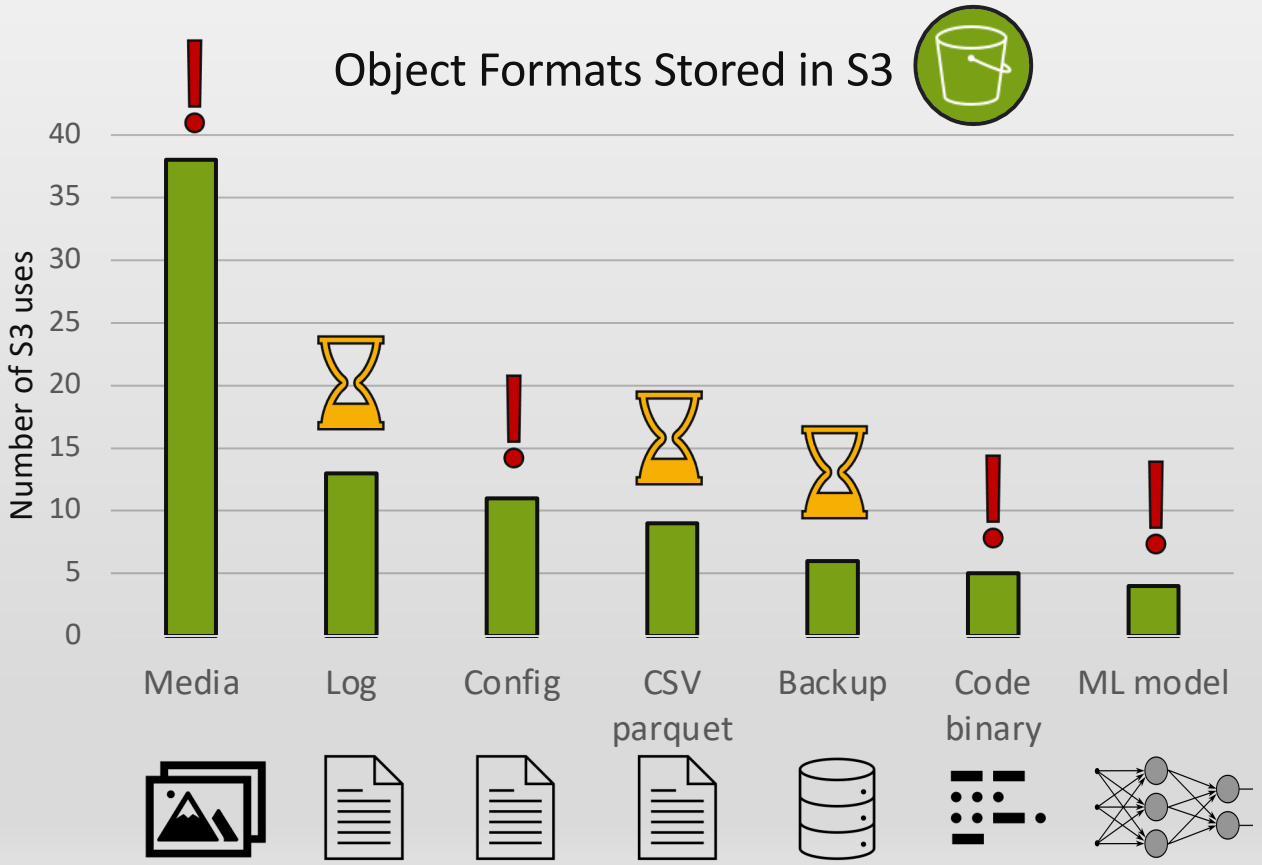
Data Formats Stored Across Services | S3

- Wide variety of data
- Semantically different data appear the same



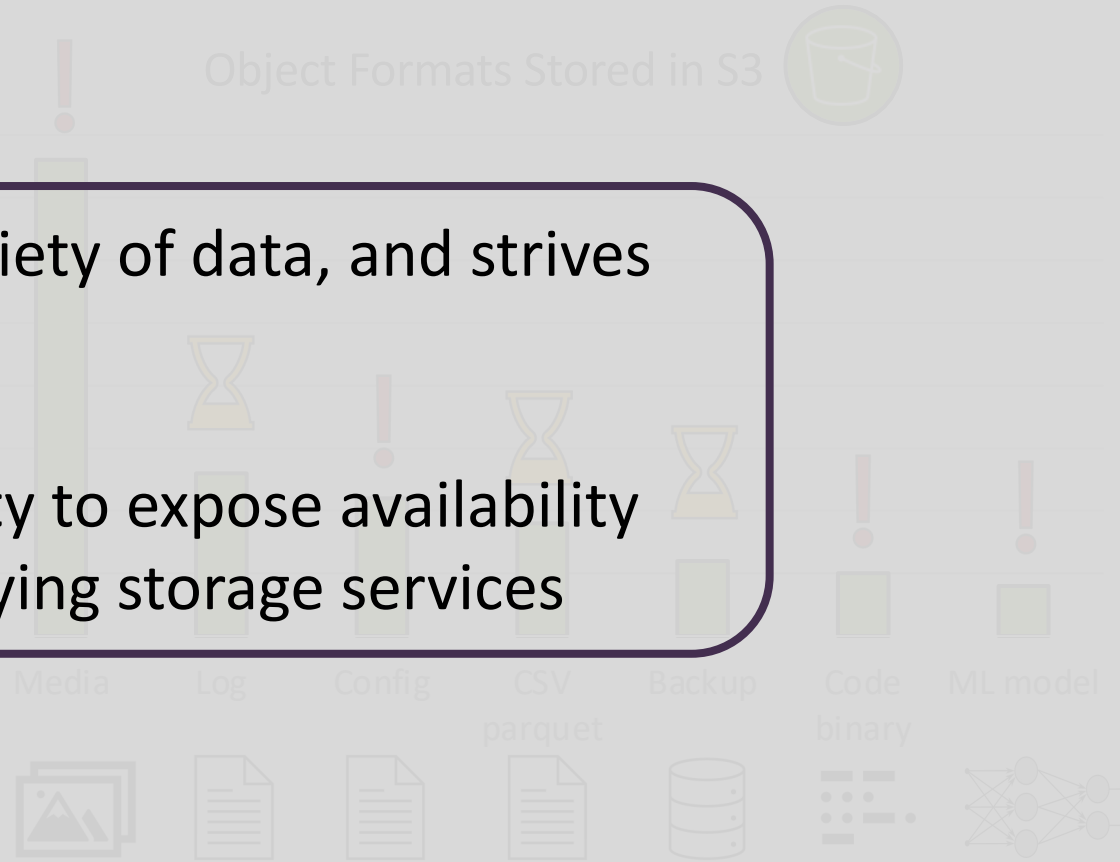
Data Formats Stored Across Services | S3

- Wide variety of data
- Semantically different data appear the same
- In practice, *availability requirements* for data differ



Data Formats Stored Across Services | S3

- Wide variety of data
- Semantically different data appear the same
- In practice, *availability requirements* for data differ

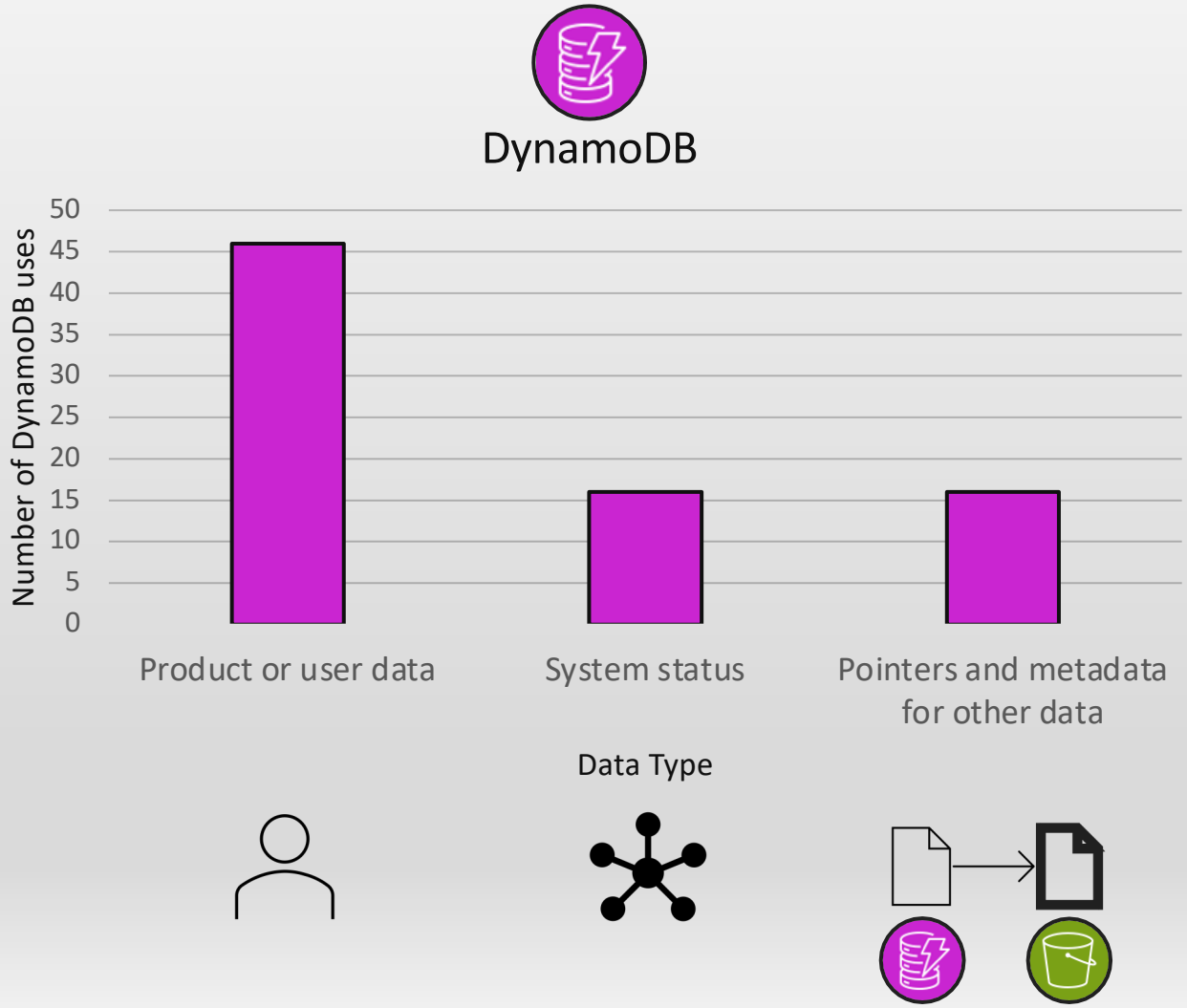


Summary: S3 stores variety of data, and strives for high availability

Implication: Opportunity to expose availability requirements to underlying storage services

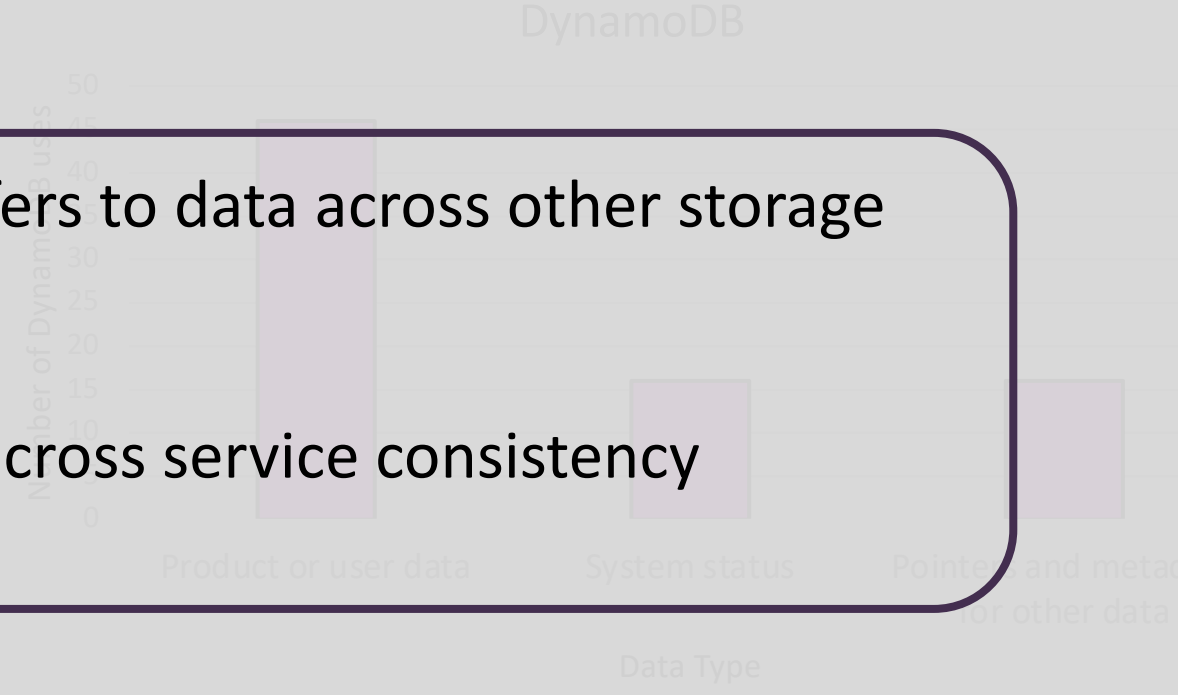
Data Formats Stored Across Services | DynamoDB

- Primarily stores user info, financial info, metrics etc.
- Orchestrator for long-running cross-service workflows
- Partial/extracted data linked to complete data



Data Formats Stored Across Services | DynamoDB

- Primarily stores user info, financial info, metrics etc.
- Orchestrator for long-running cross-service workflows
- Partial/extracted data linked to complete data



Summary: Storage refers to data across other storage services

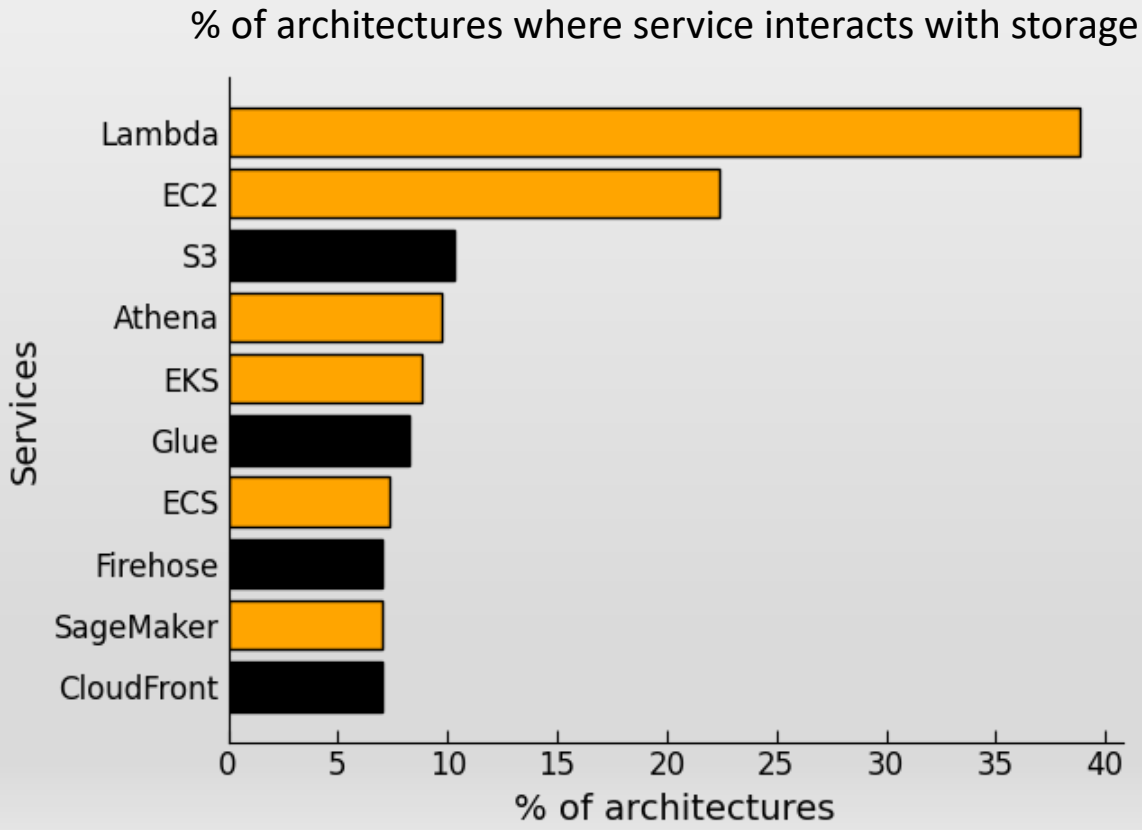
Implication: Need for cross service consistency mechanisms

Outline

- Introduction
- Methodology
- Dataset description
- Four findings about the storage layer
  - Composition of the storage layer
  - Popularity of storage services
  - Content of storage services
  - **Services interacting with storage**

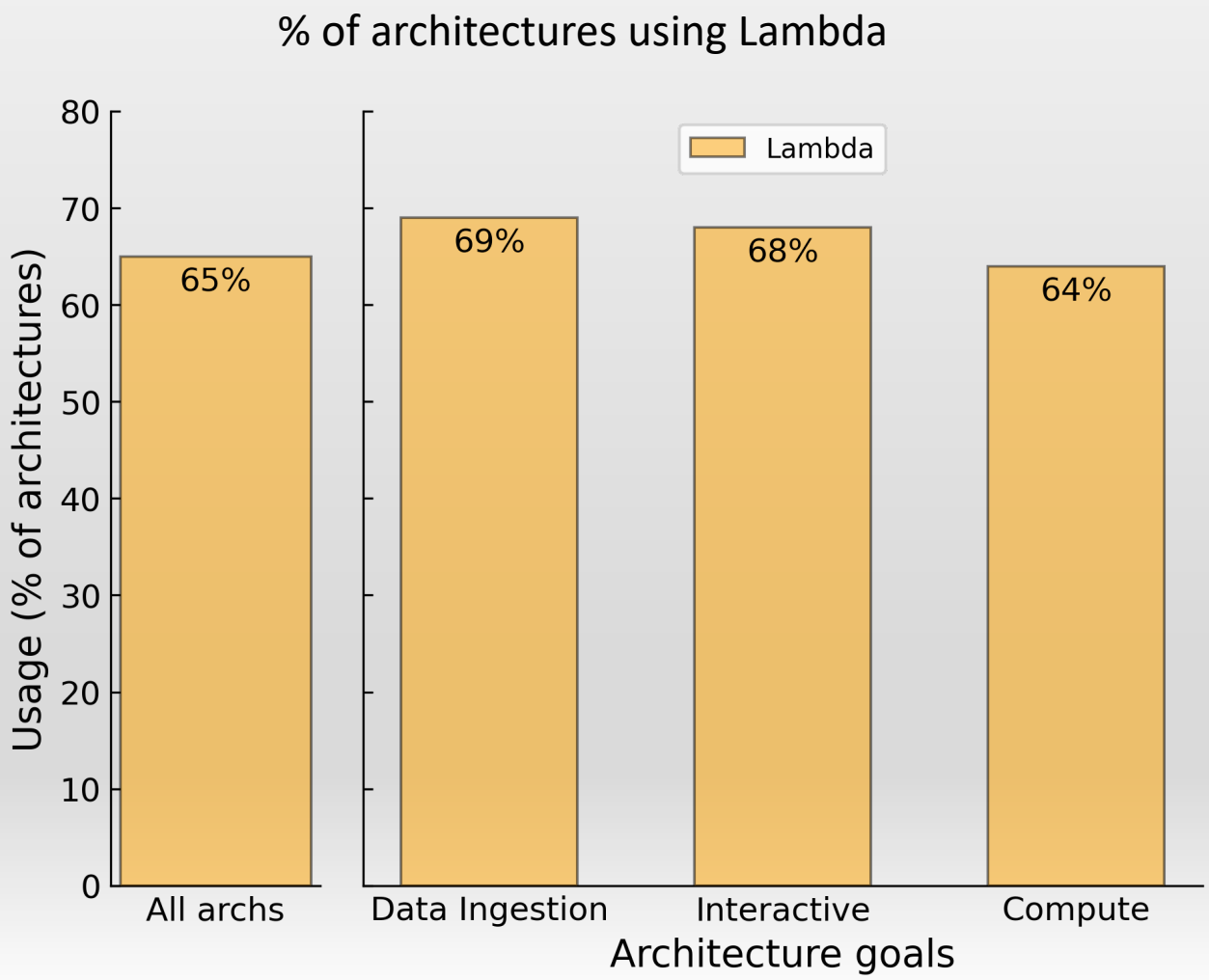
Services Interacting With Storage

- Compute services like Lambda, EC2 interact with storage services chiefly
- Lambda is the most common user of storage



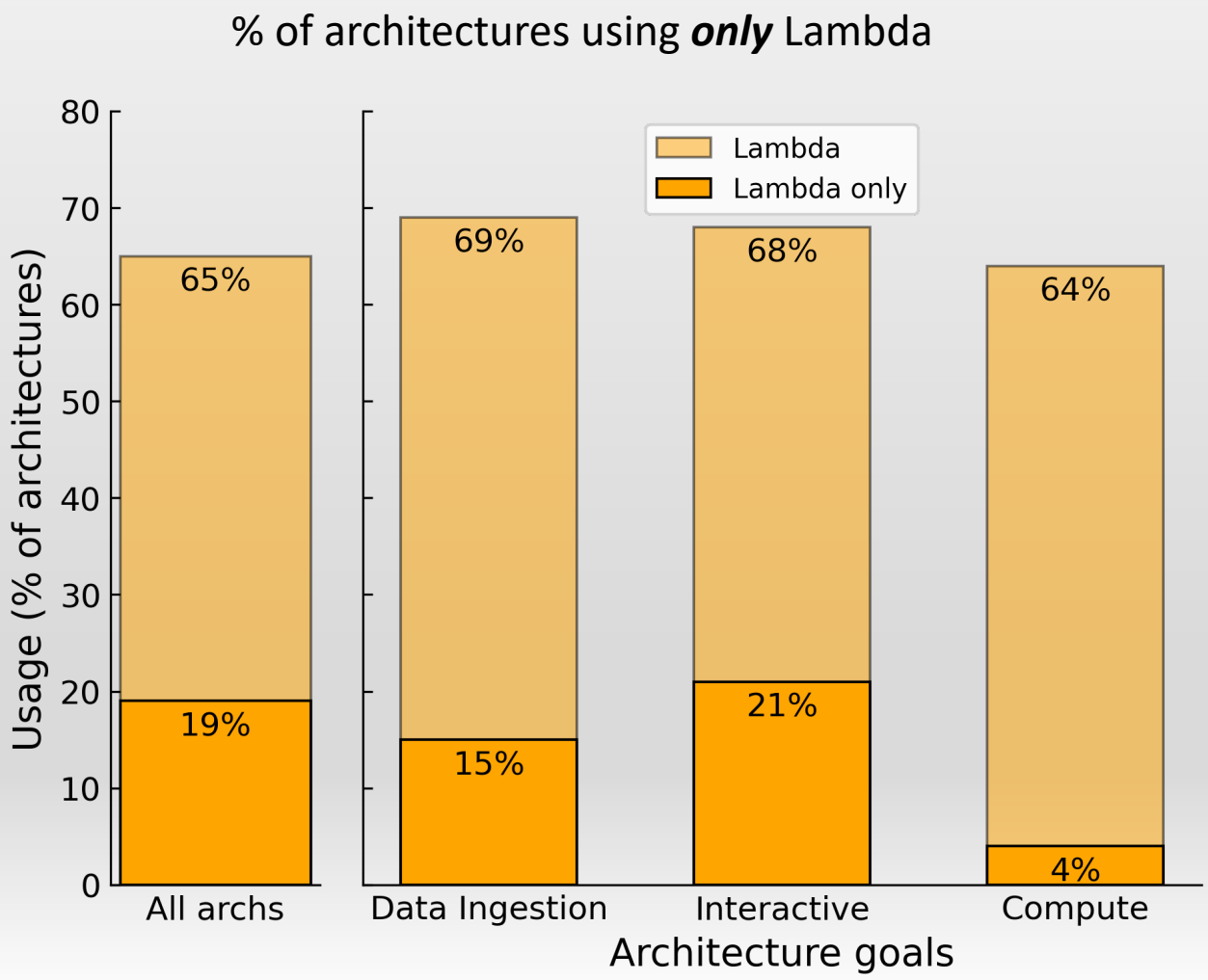
Uncovering Lambda Adoption

- Lambda is used in 65% architectures
- Used across *all goals*



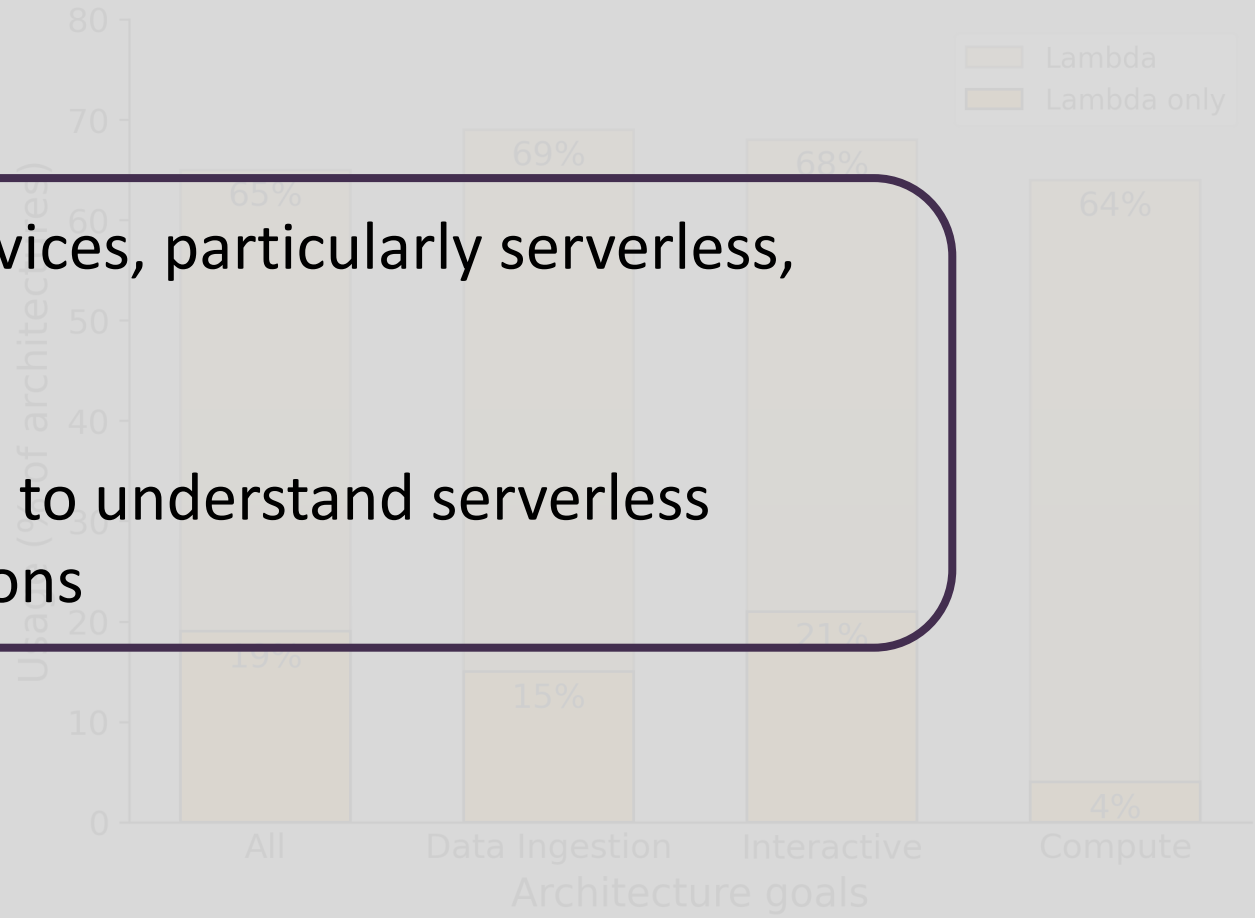
Uncovering Lambda Adoption

- Lambda is used in 65% architectures
- Used across *all goals*
- Lambda powers 20% architectures



Uncovering Lambda Adoption

- Lambda is used in 65% architectures
- Lambda powers 20% architectures



Summary: Compute services, particularly serverless, stress storage services

Implication: Rising need to understand serverless workloads and interactions

Conclusion

Cloudscape is the first large-scale dataset of real-world cloud architectures

We found:

- the storage layer is diverse and heterogeneous
- S3 is the most popular storage service, while distributed filesystems are rare
- storage services should consider data semantics, not just format
- it is important to understand object store and serverless function workloads
- *… and more in the paper!*

We need more such datasets!

Dataset

github.com/WiscAdsl/Cloudscape

Explorer

cloudscape.cs.wisc.edu