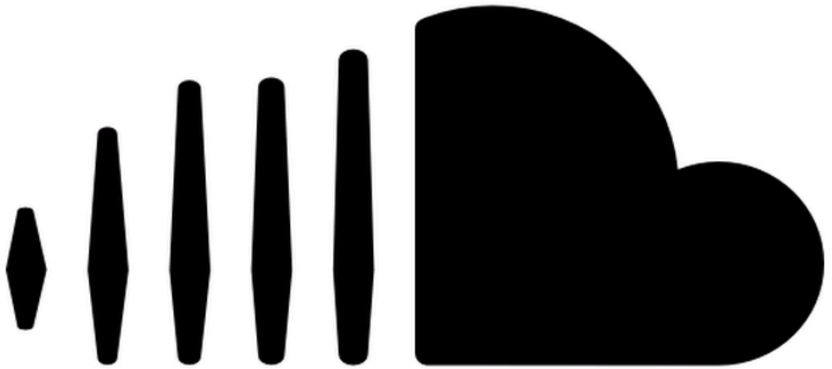
Mechanics of Data Pipelines

Berlin Buzzwords 2017

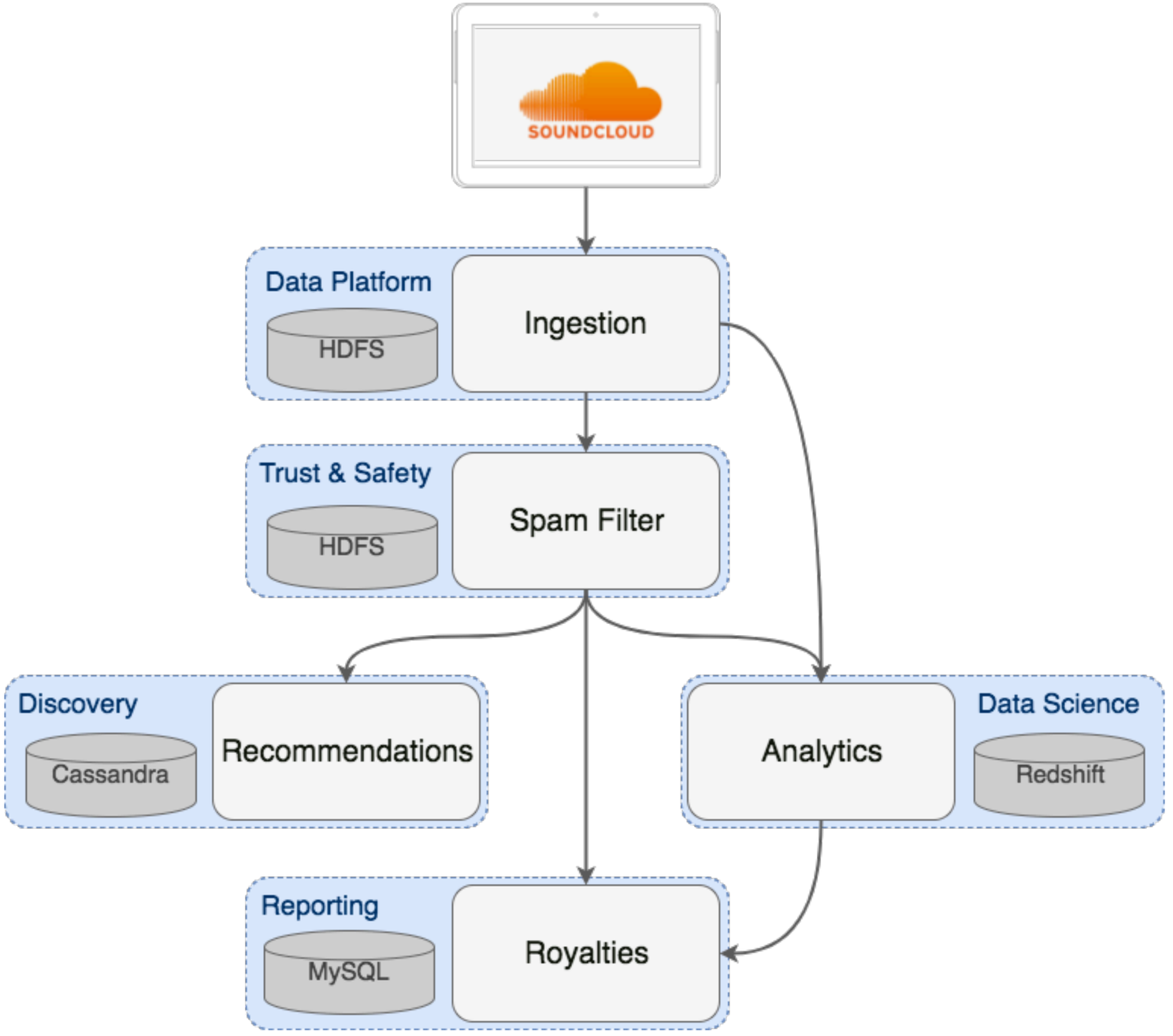
@brapse



Data Pipelines

The software structures which emerge to process and disseminate information.

A connected set of map reduce jobs for loading data into (a) database(s).


SOUNDCLOUD
Data Platform
HDFS
Ingestion
Trust & Safety
HDFS
Spam Filter
Discovery
Cassandra
Recommendations
Analytics
Data Science
Redshift
Reporting
MySQL
Royalties

Why Data Pipelines?

To integrate diverse perspectives.

Enable and empower collaboration between diverse sets of domain experts.

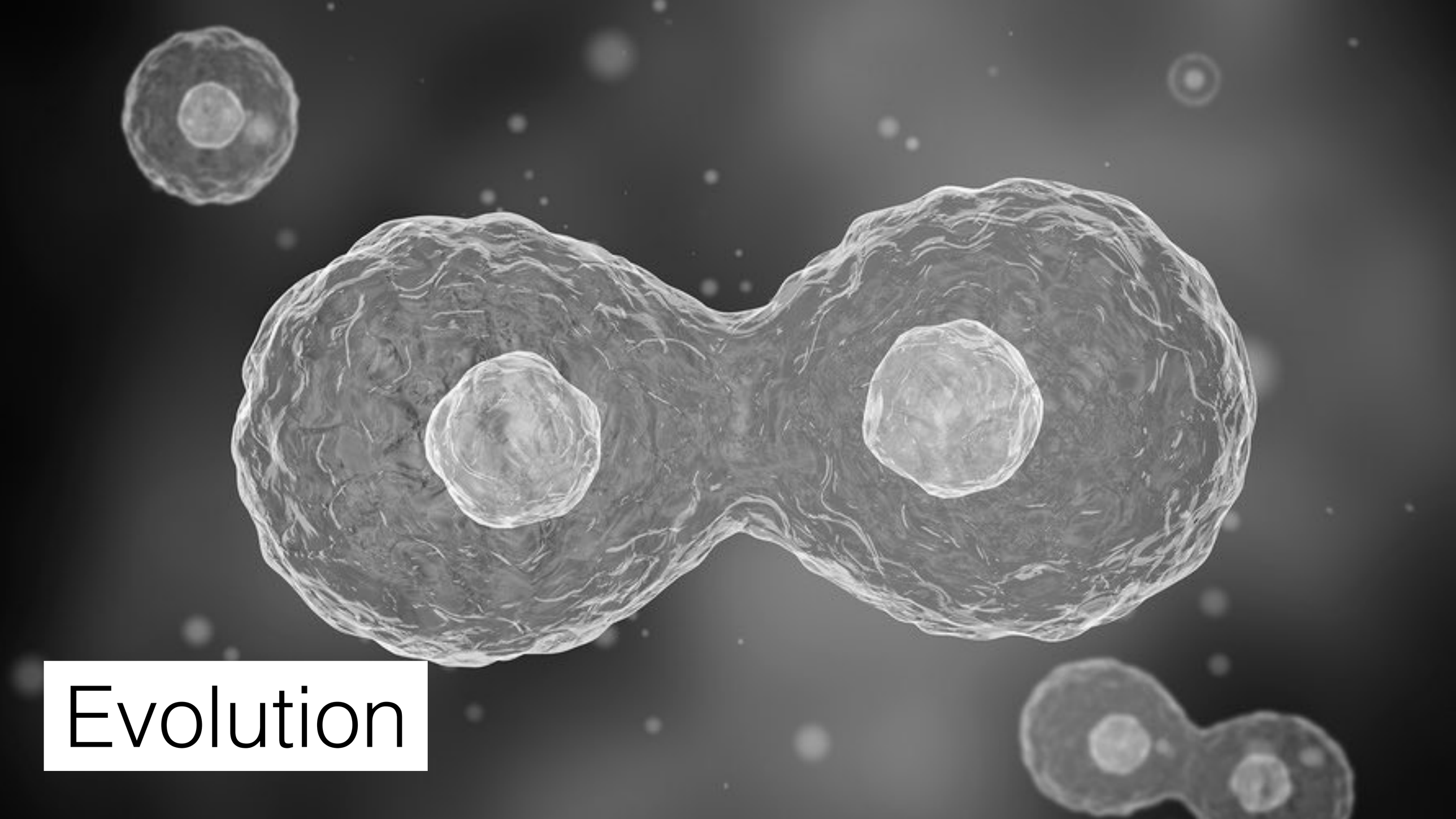
How do we build them?

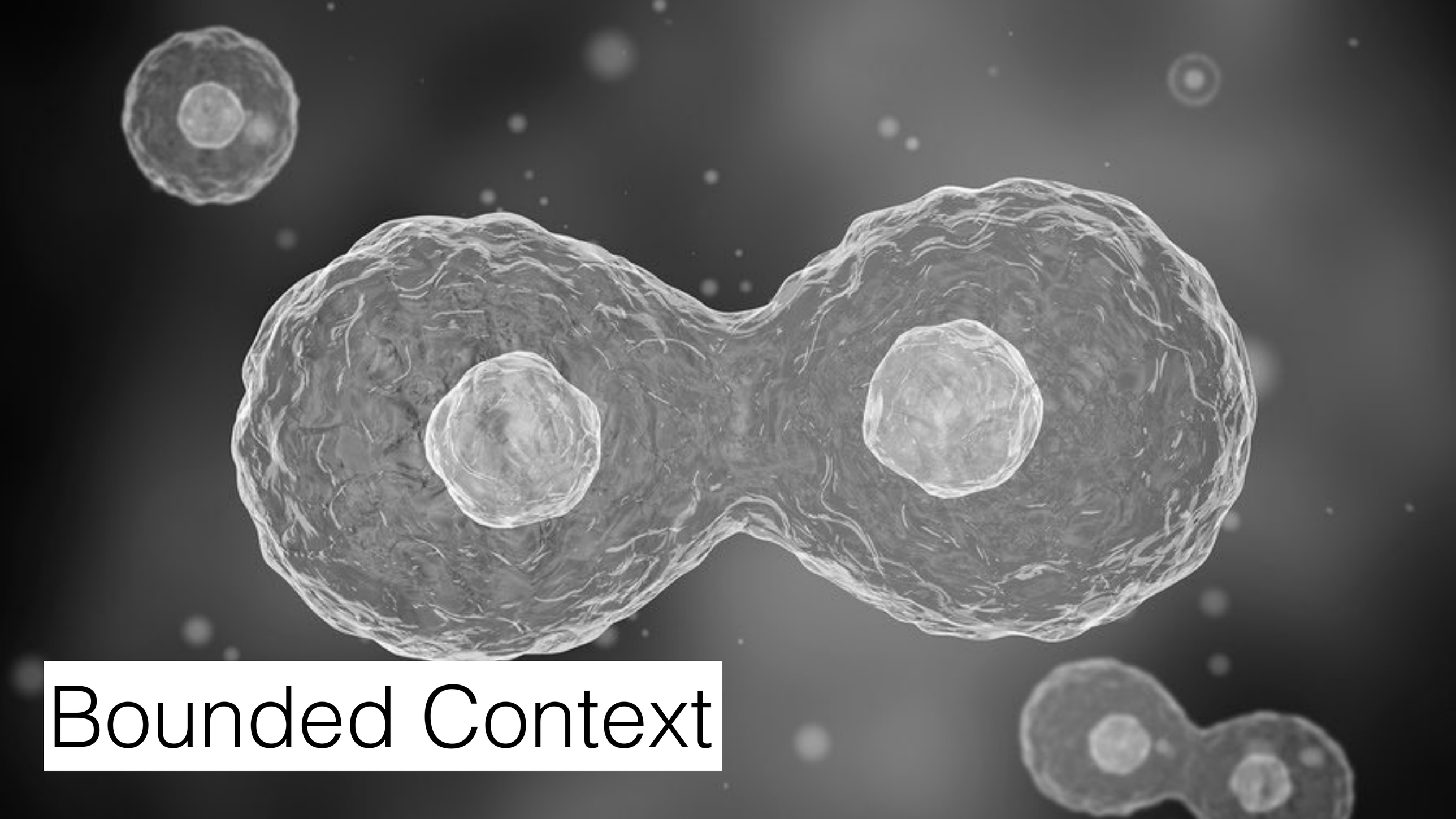
Often Badly.

Misunderstood domain.
Misunderstood integration.
Misunderstood coordination.

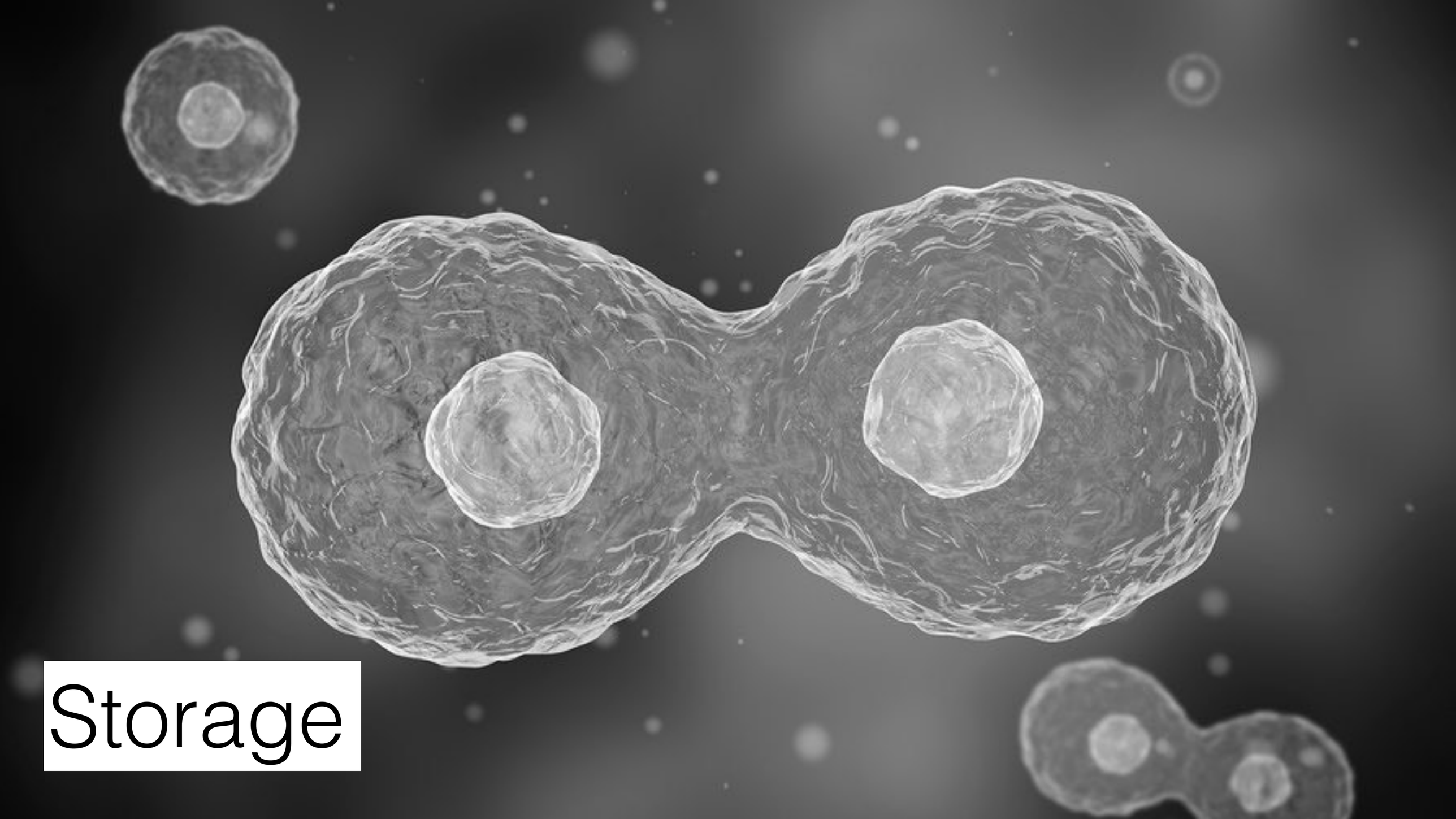
Thesis:
Data Pipelines emerge and grow to reflect collaboration between domains and are impeded by incidental coordination.

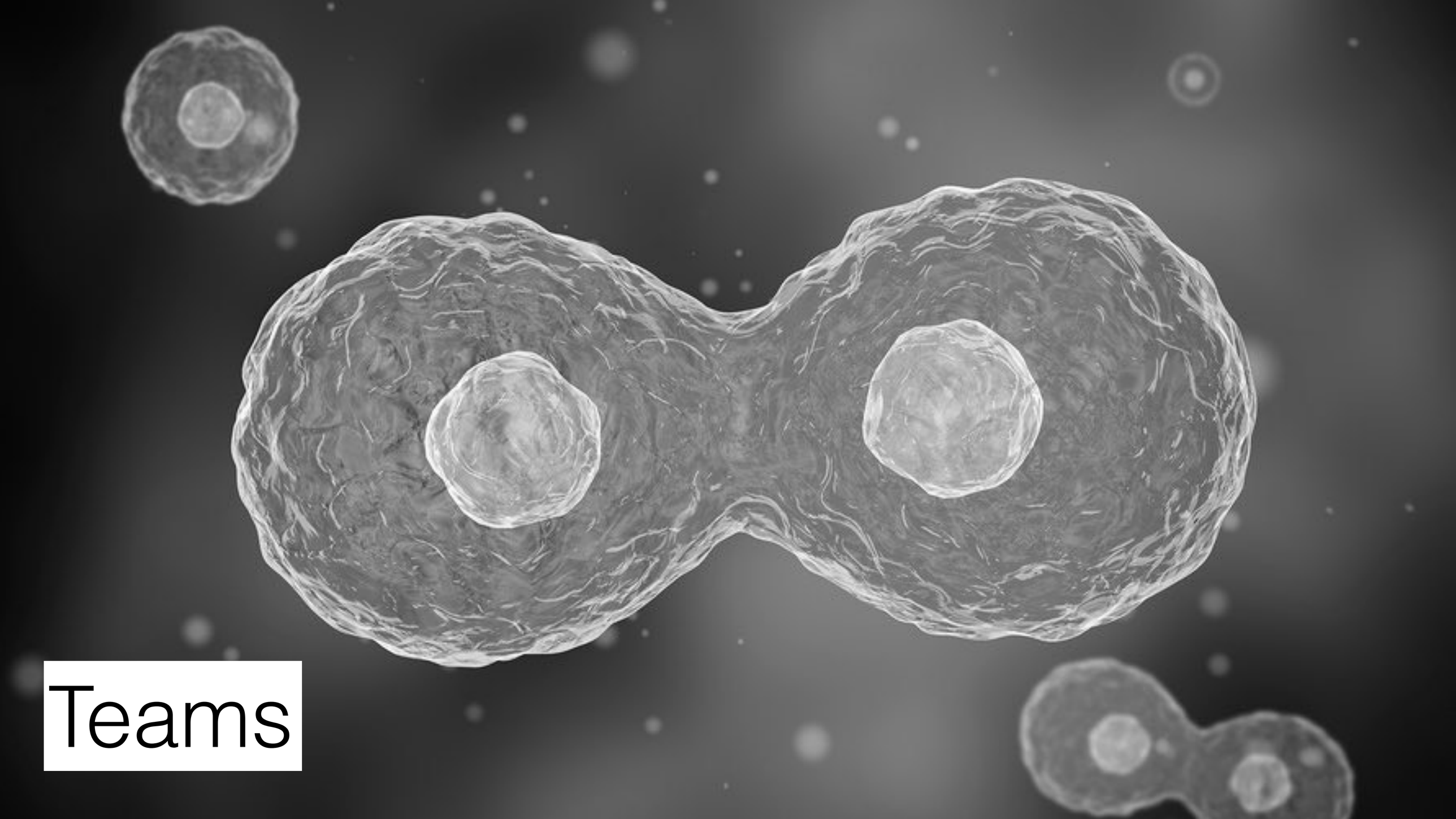
Abstract

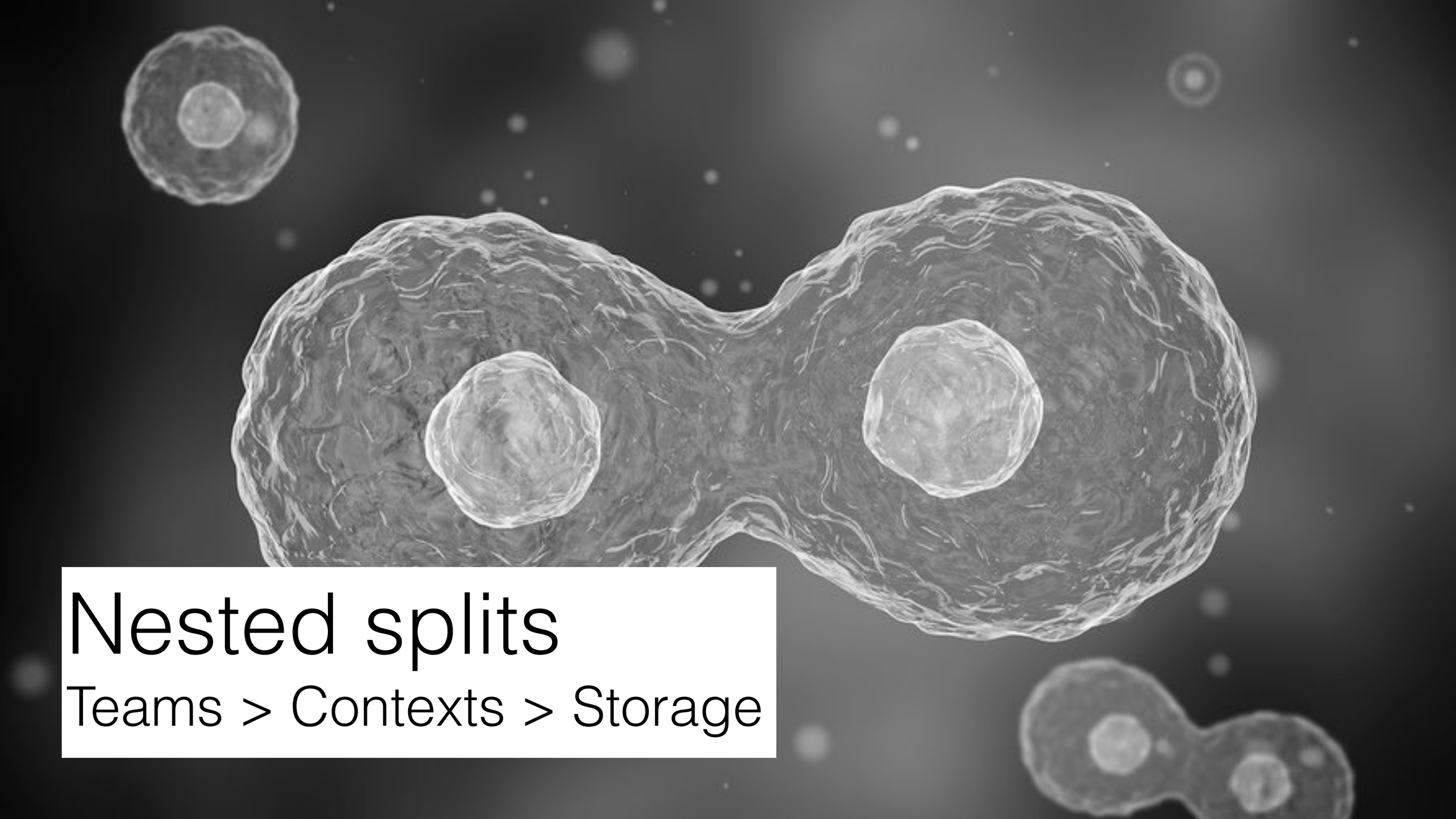

Evolution



Bounded Context


Storage


Teams

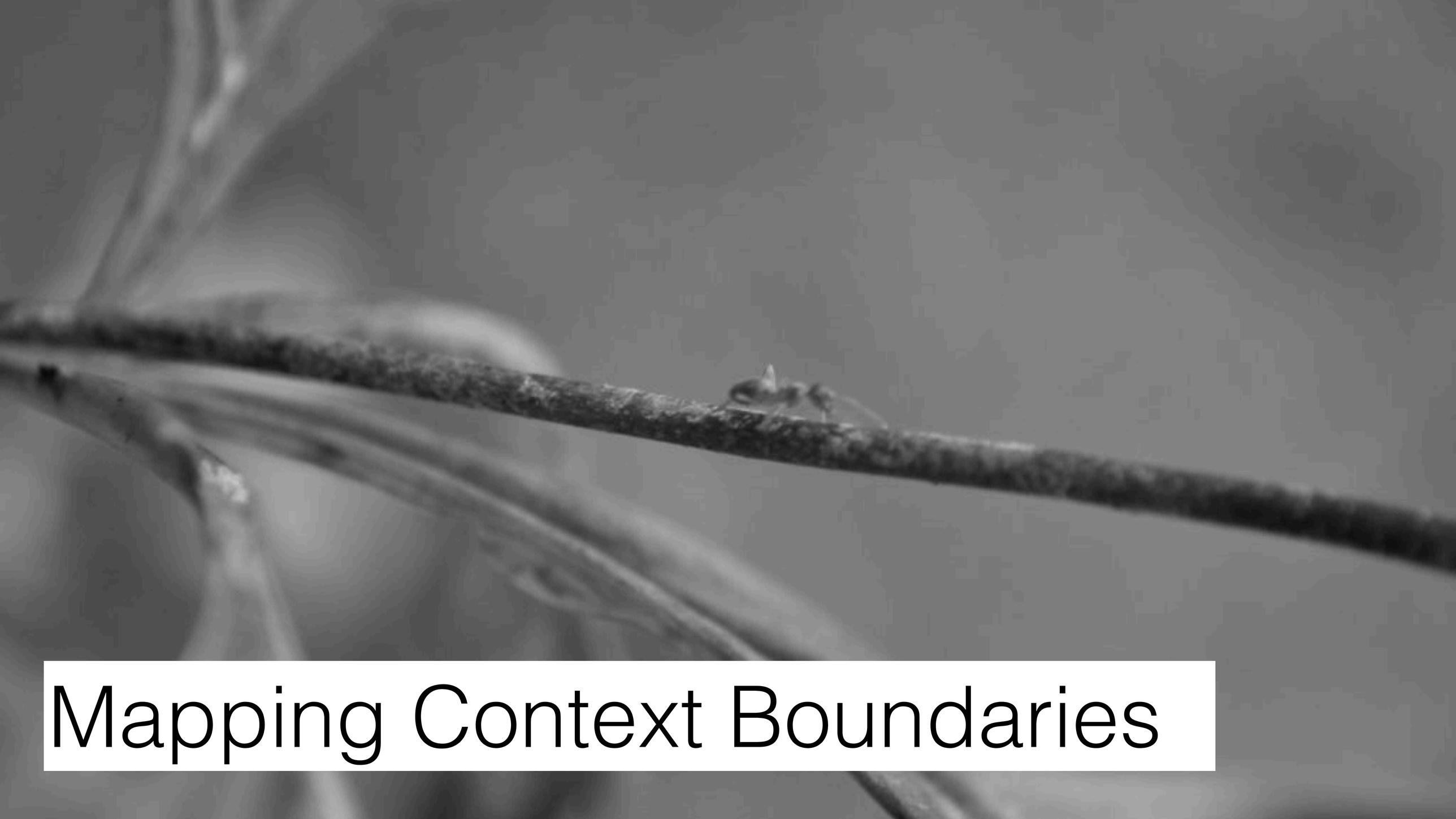


Nested splits

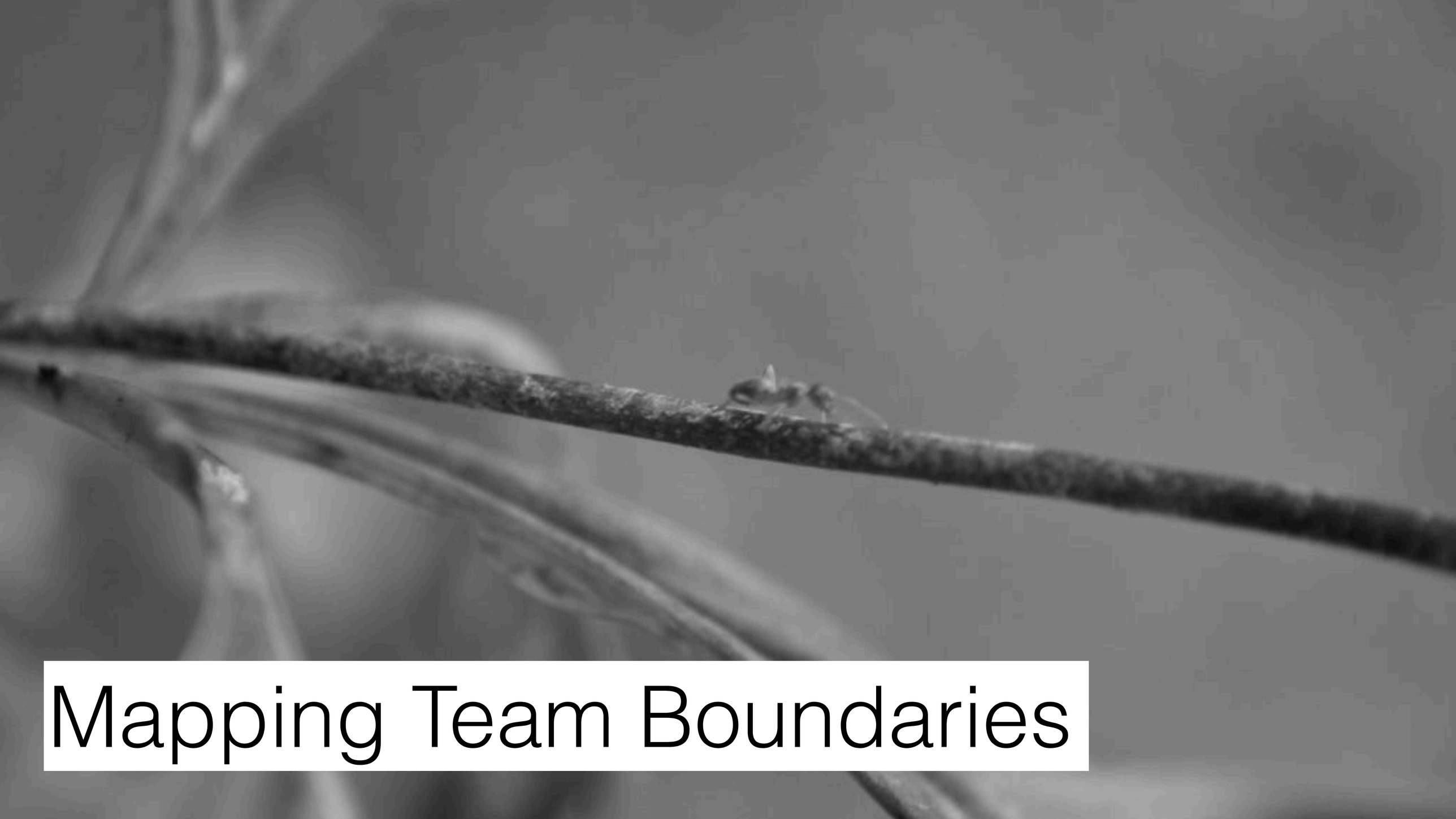
Teams > Contexts > Storage

Mapping

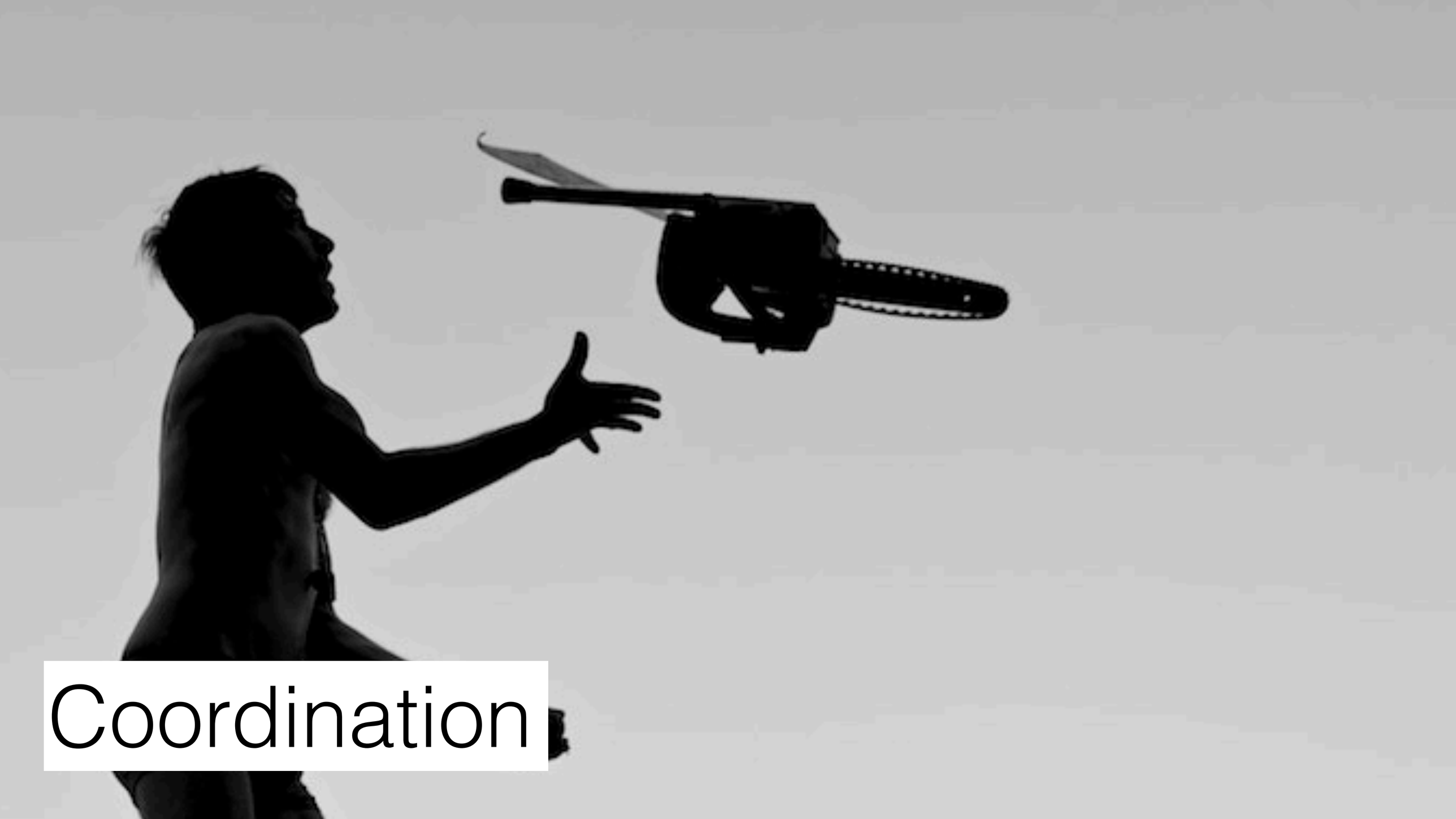
Mapping Storage Boundaries



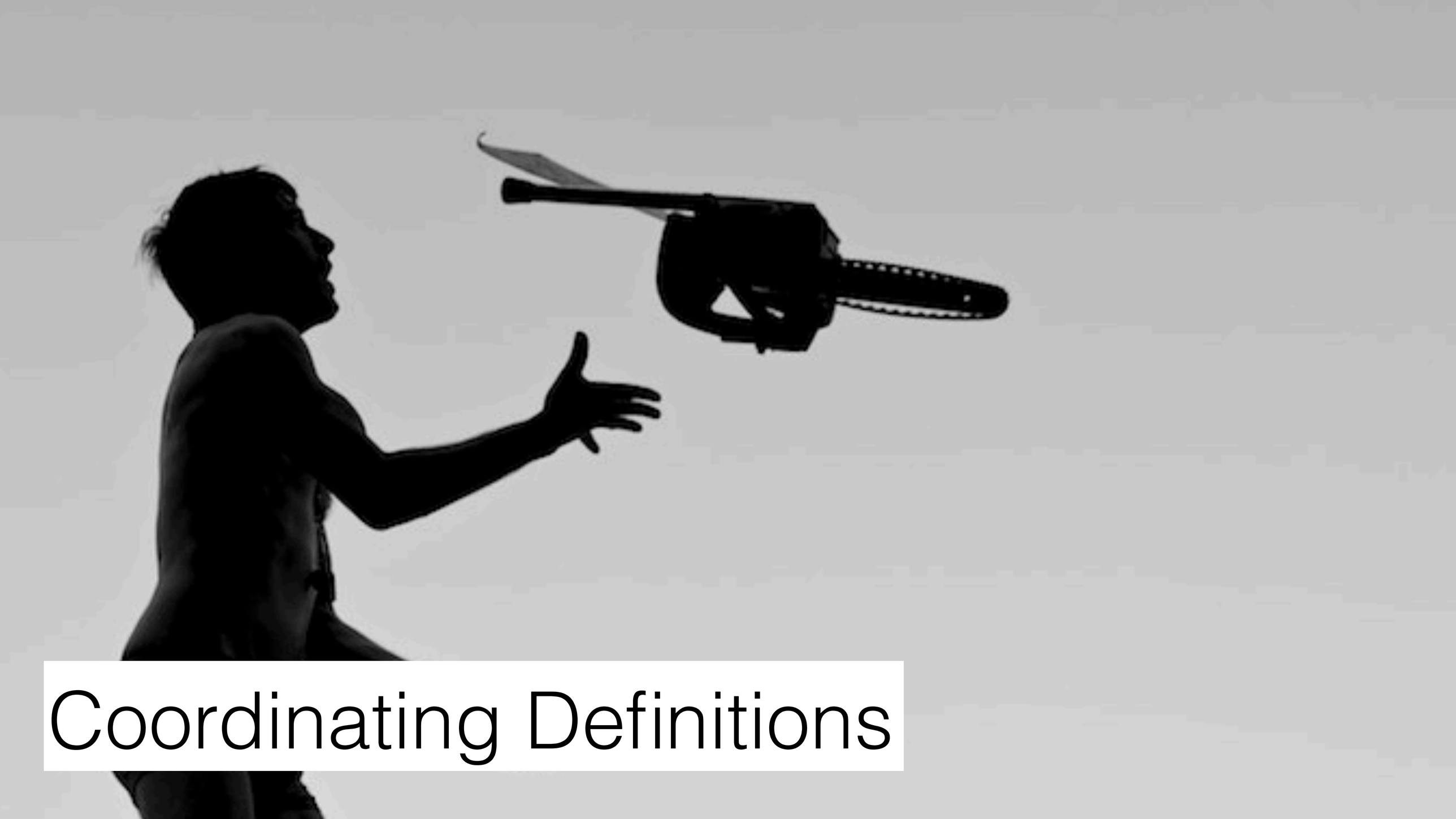
Mapping Context Boundaries



Mapping Team Boundaries



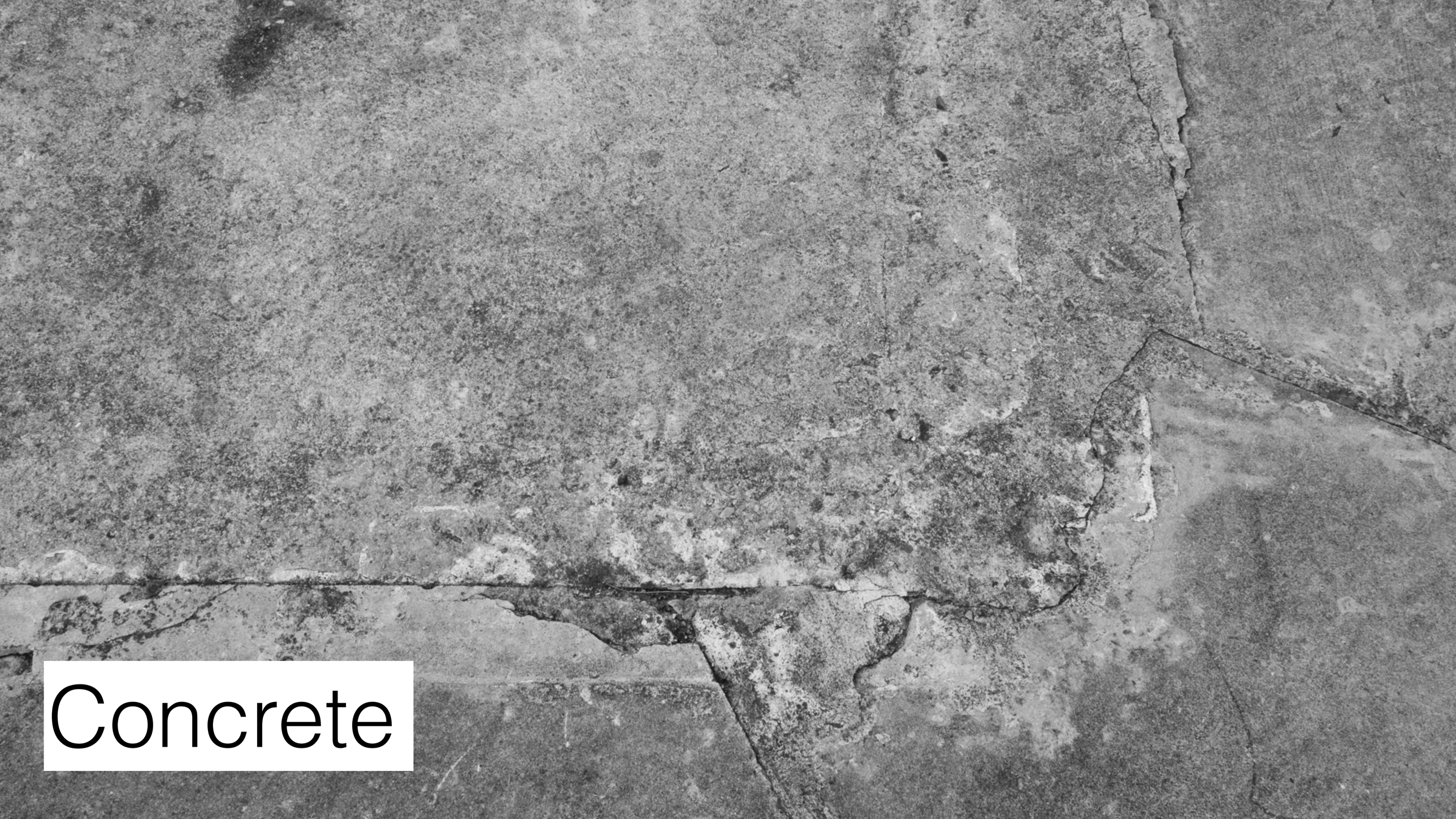
Coordination

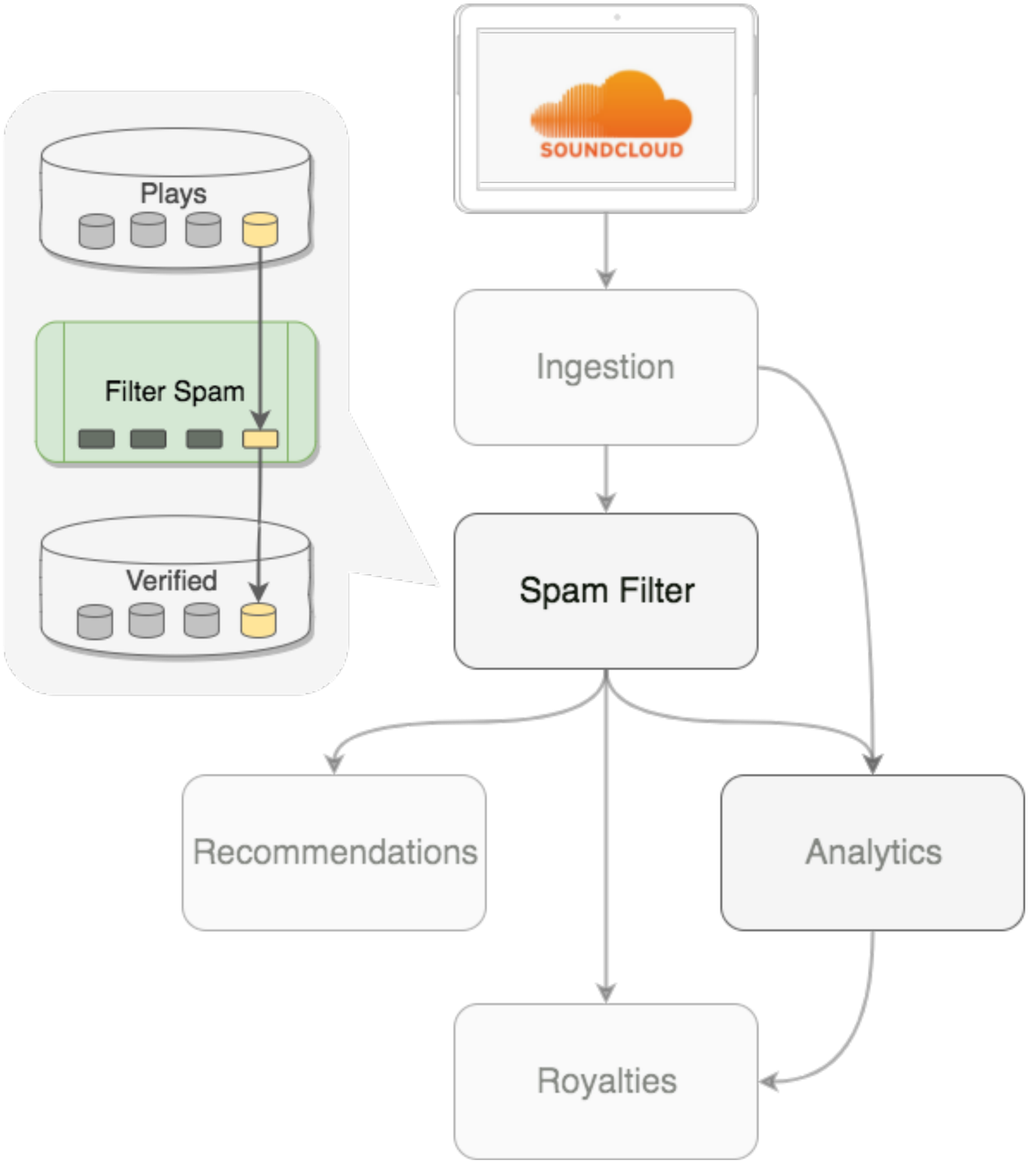


Coordinating Definitions

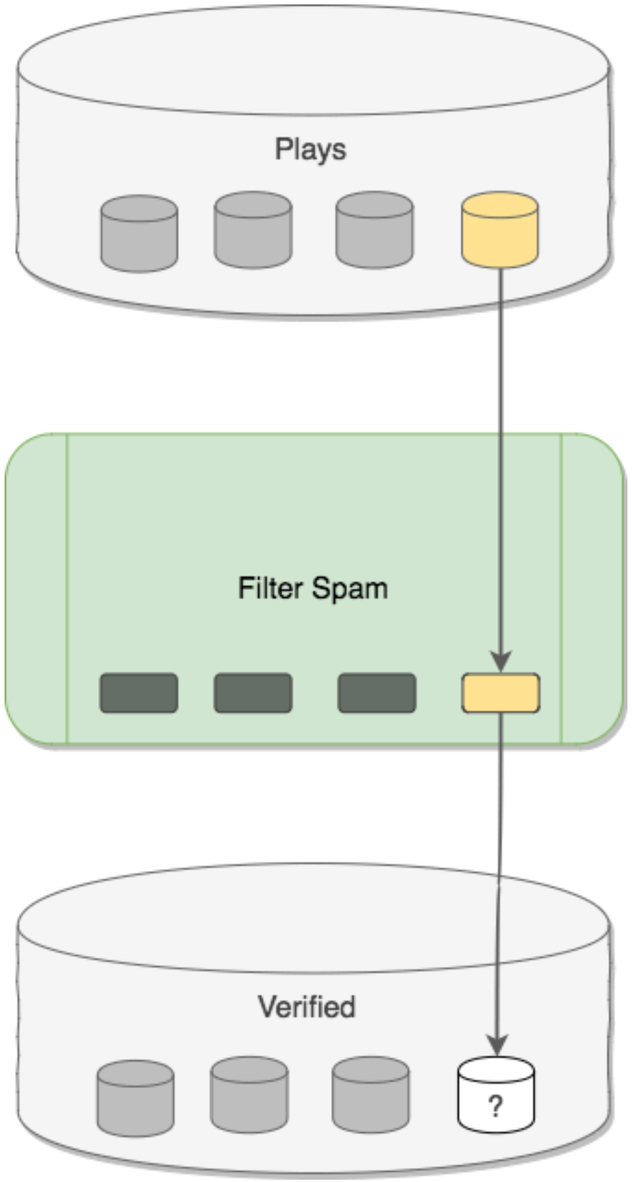
Coordinating Correctness

Coordinating Failure

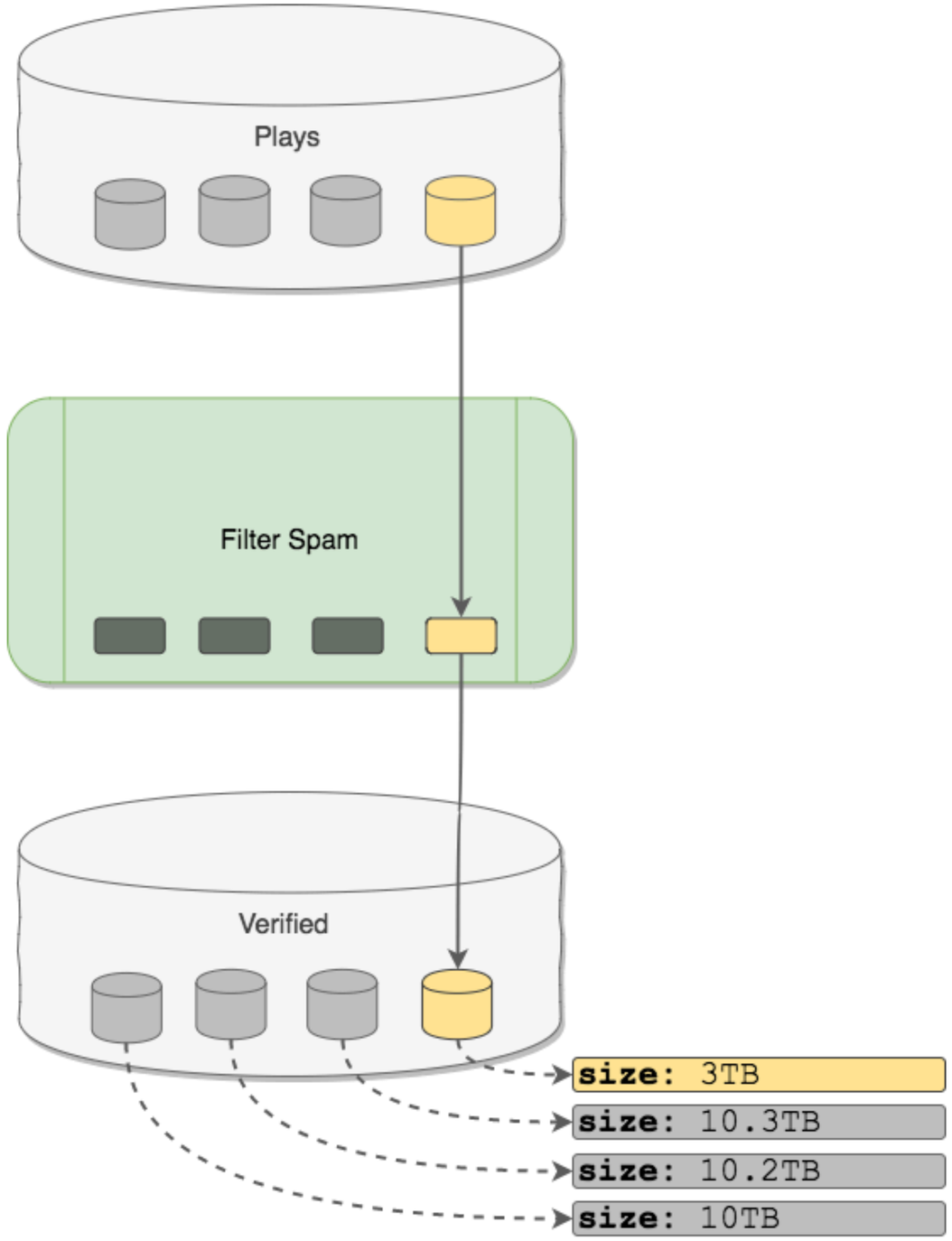

Concrete


SOUNDCLOUD
Plays
Filter Spam
Verified
Ingestion
Spam Filter
Recommendations
Analytics
Royalties

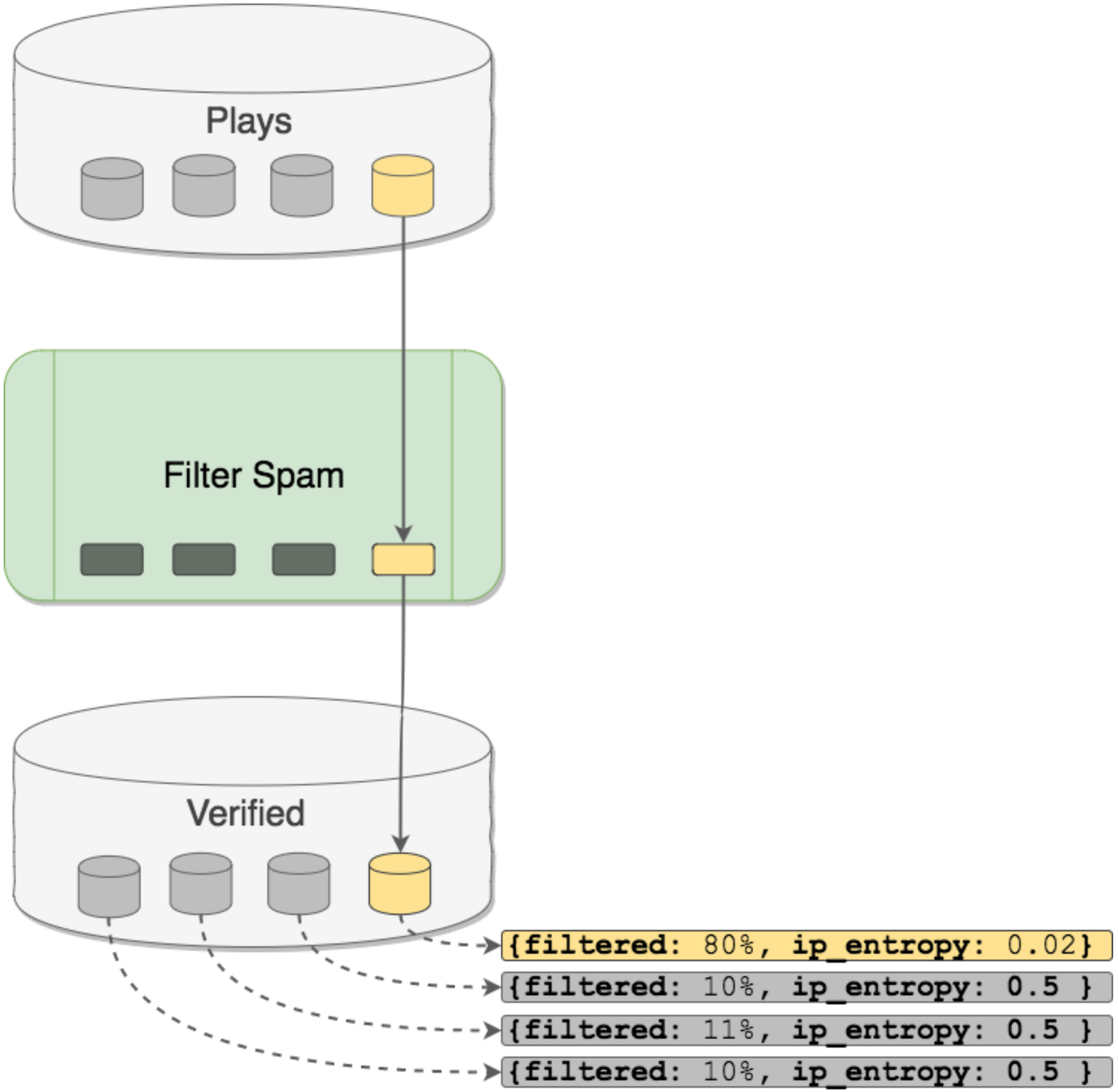
Correctness



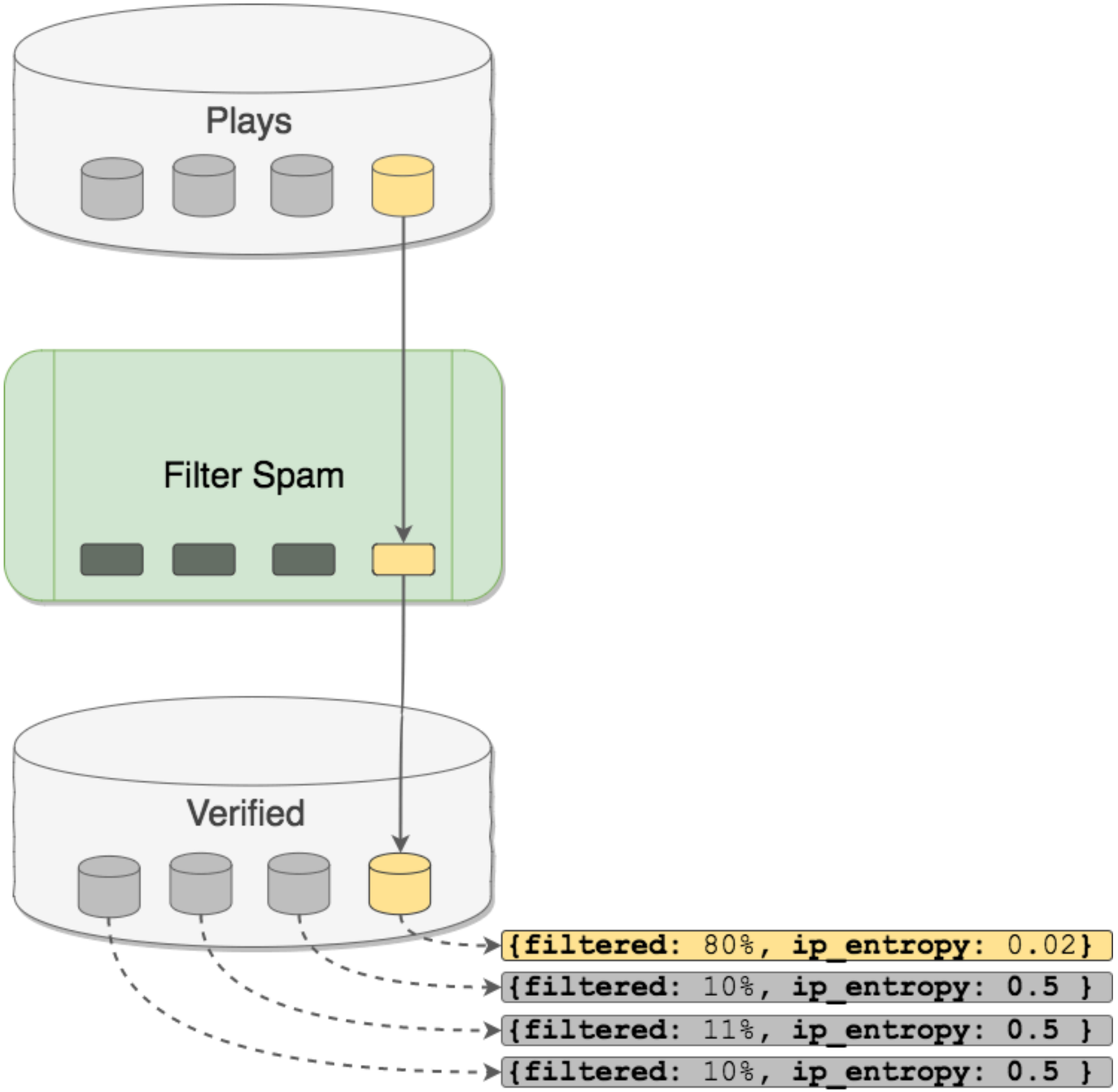
Fingerprints

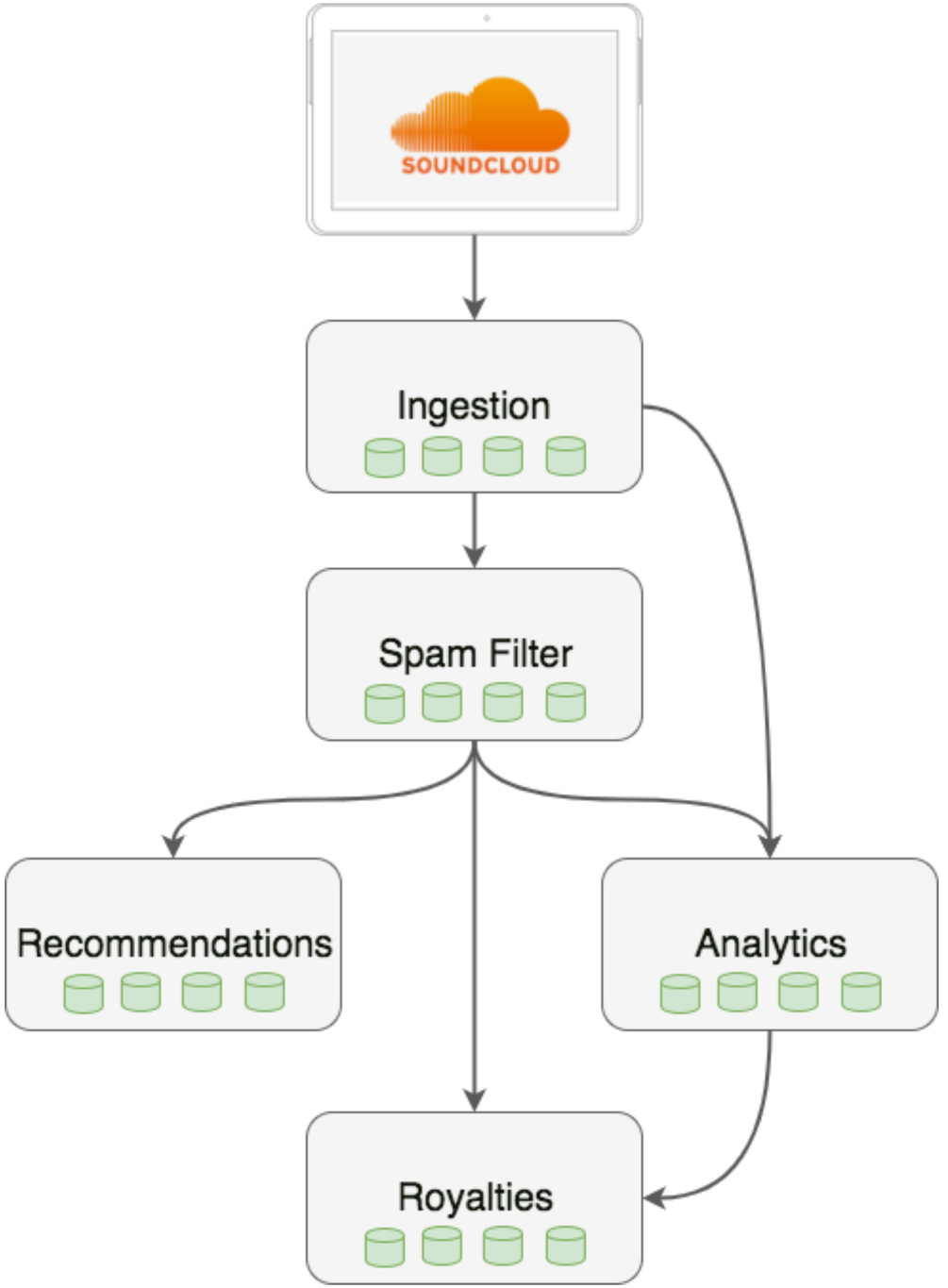


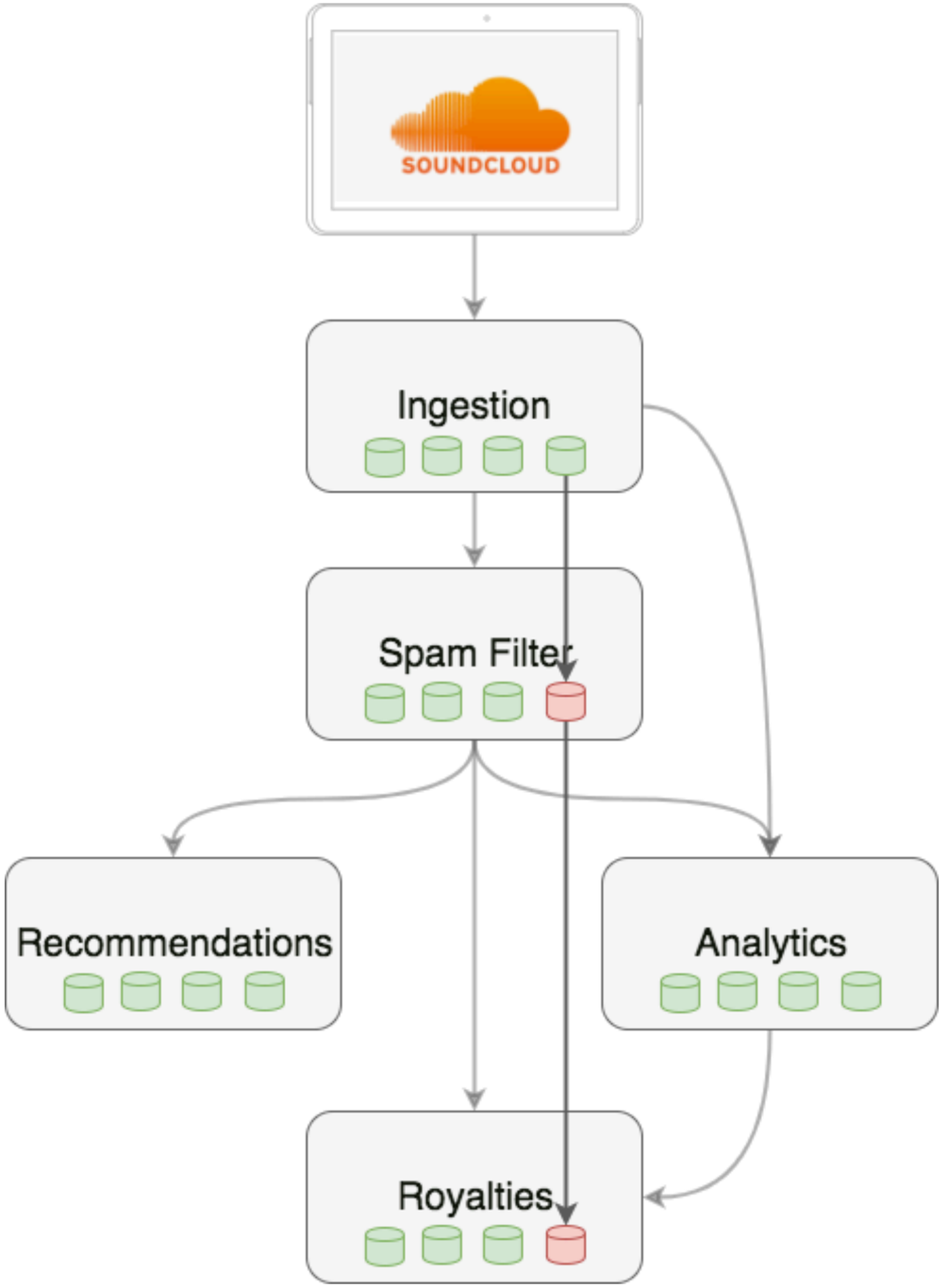
Fingerprints

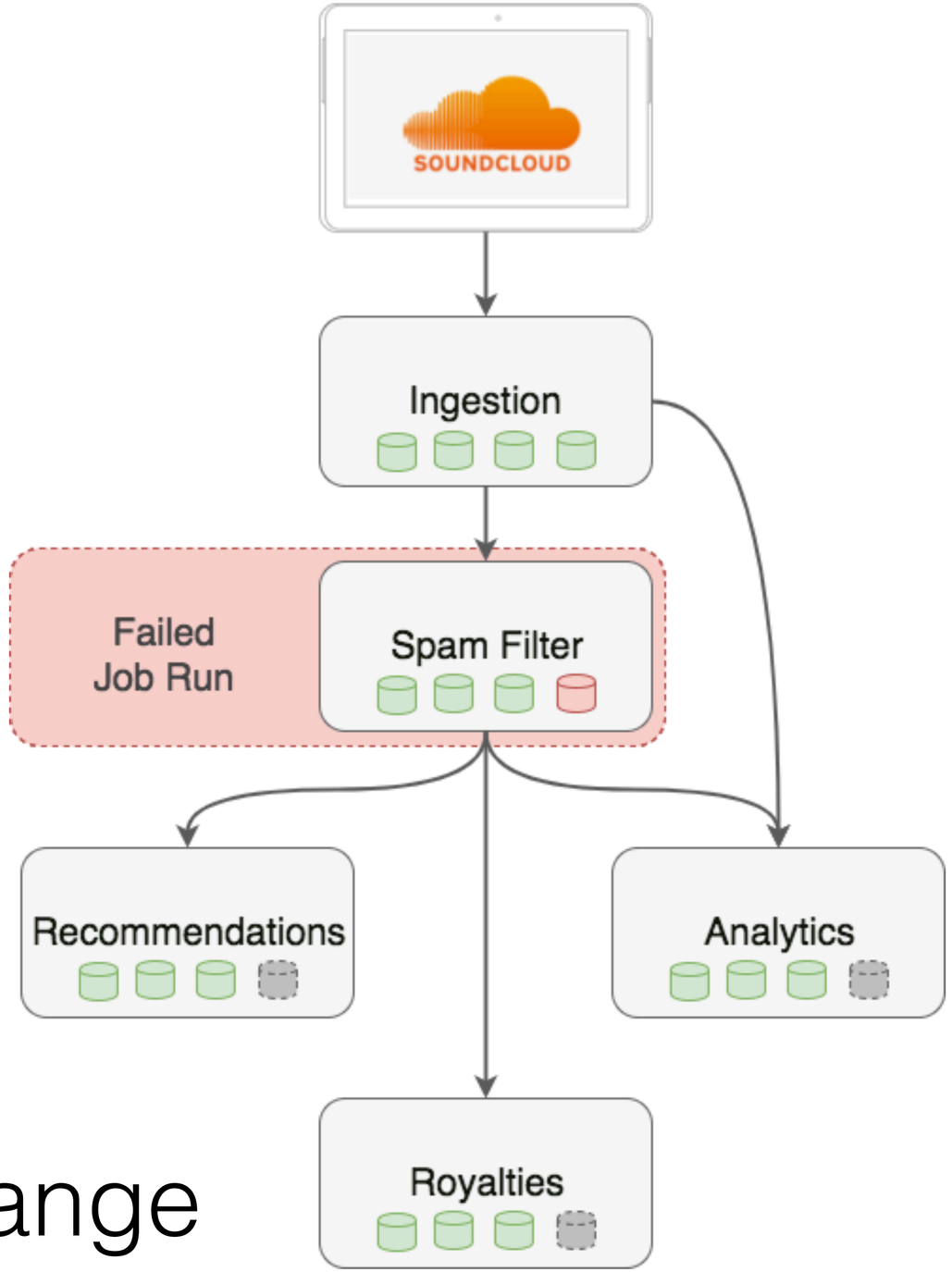


Fingerprints

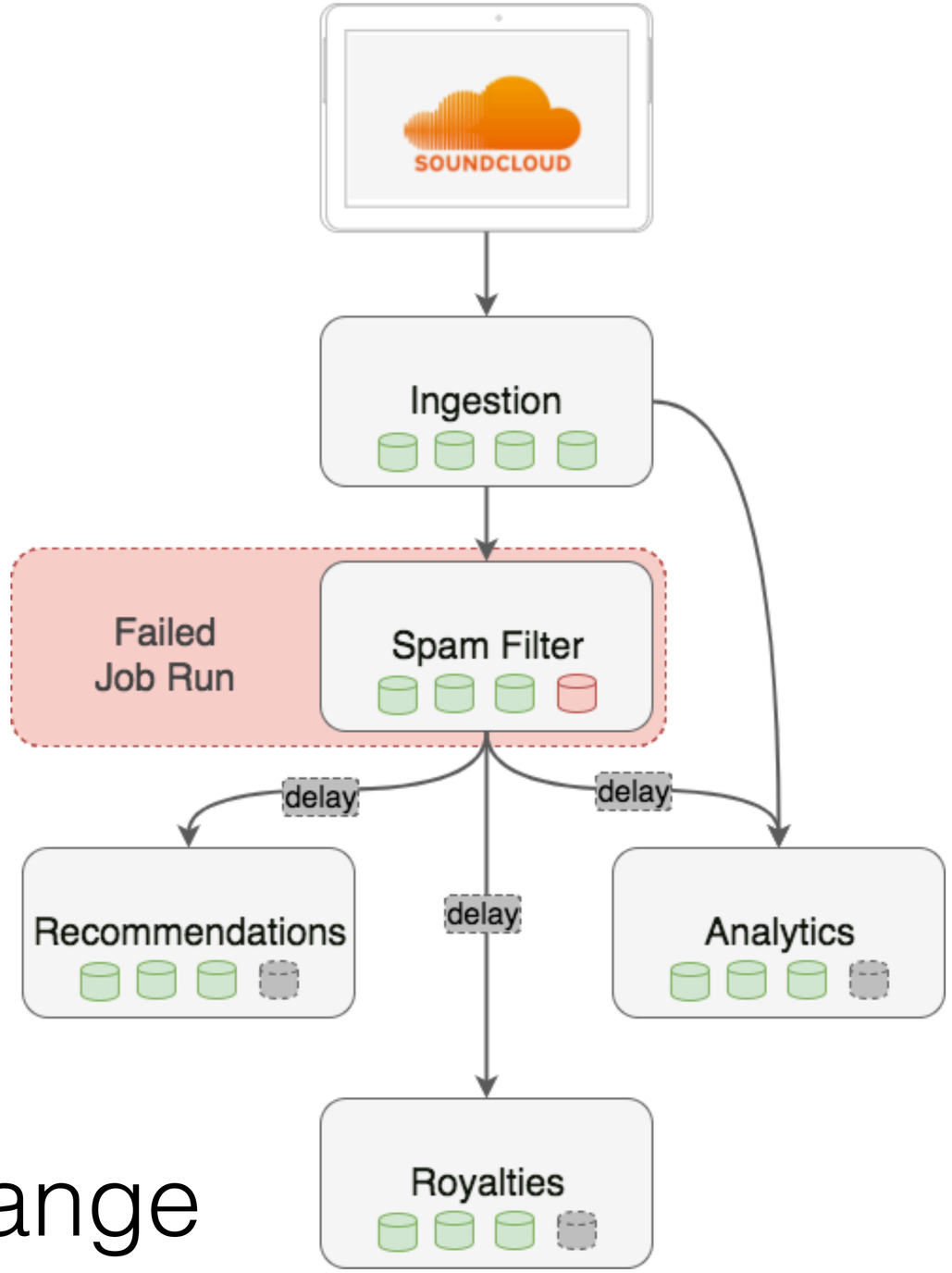



SOUNDCLOUD
Ingestion
Spam Filter
Recommendations
Analytics
Royalties

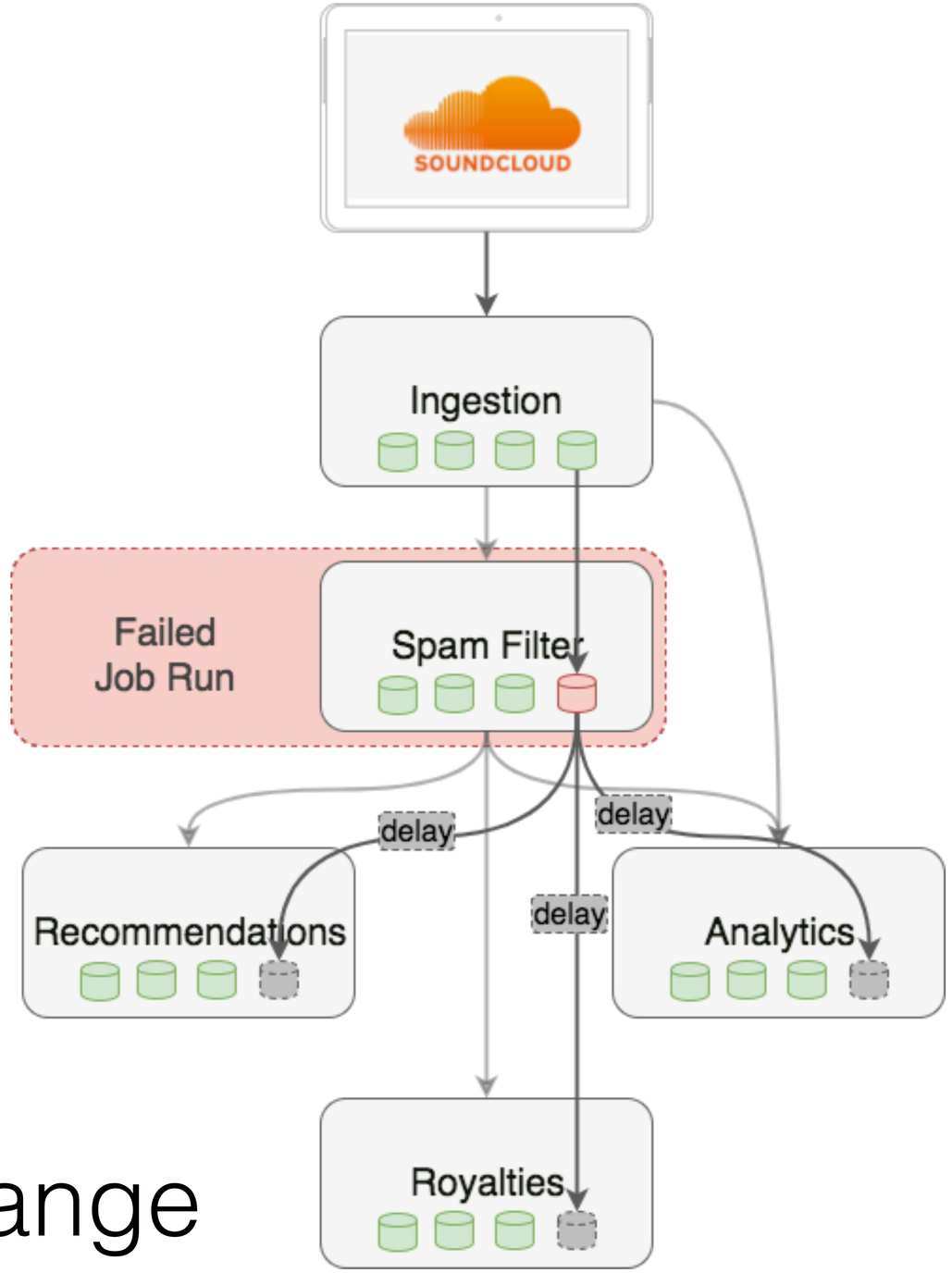

SOUNDCLOUD
Ingestion
Spam Filter
Recommendations
Analytics
Royalties



Coordinate Change

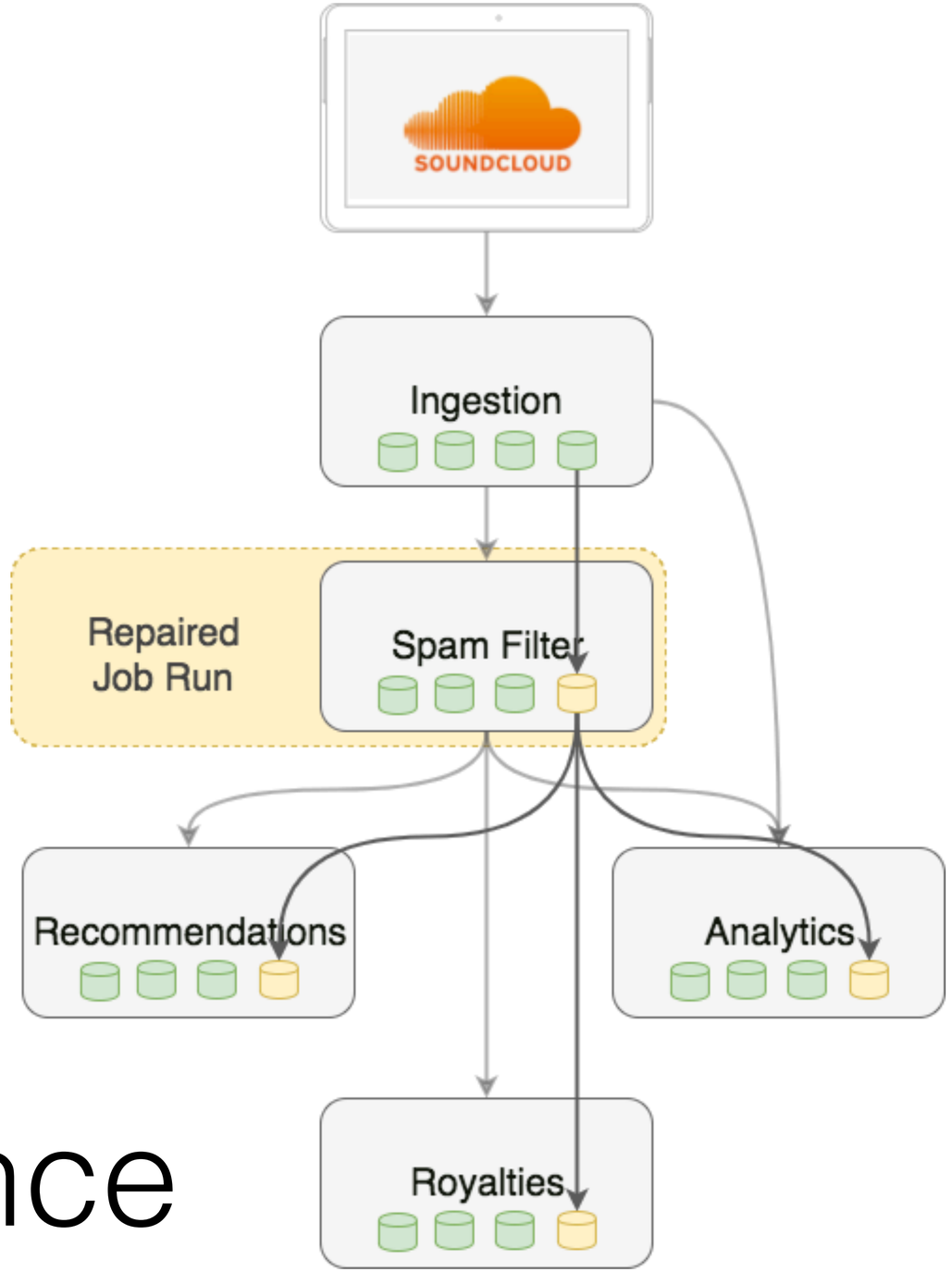


Coordinate Change

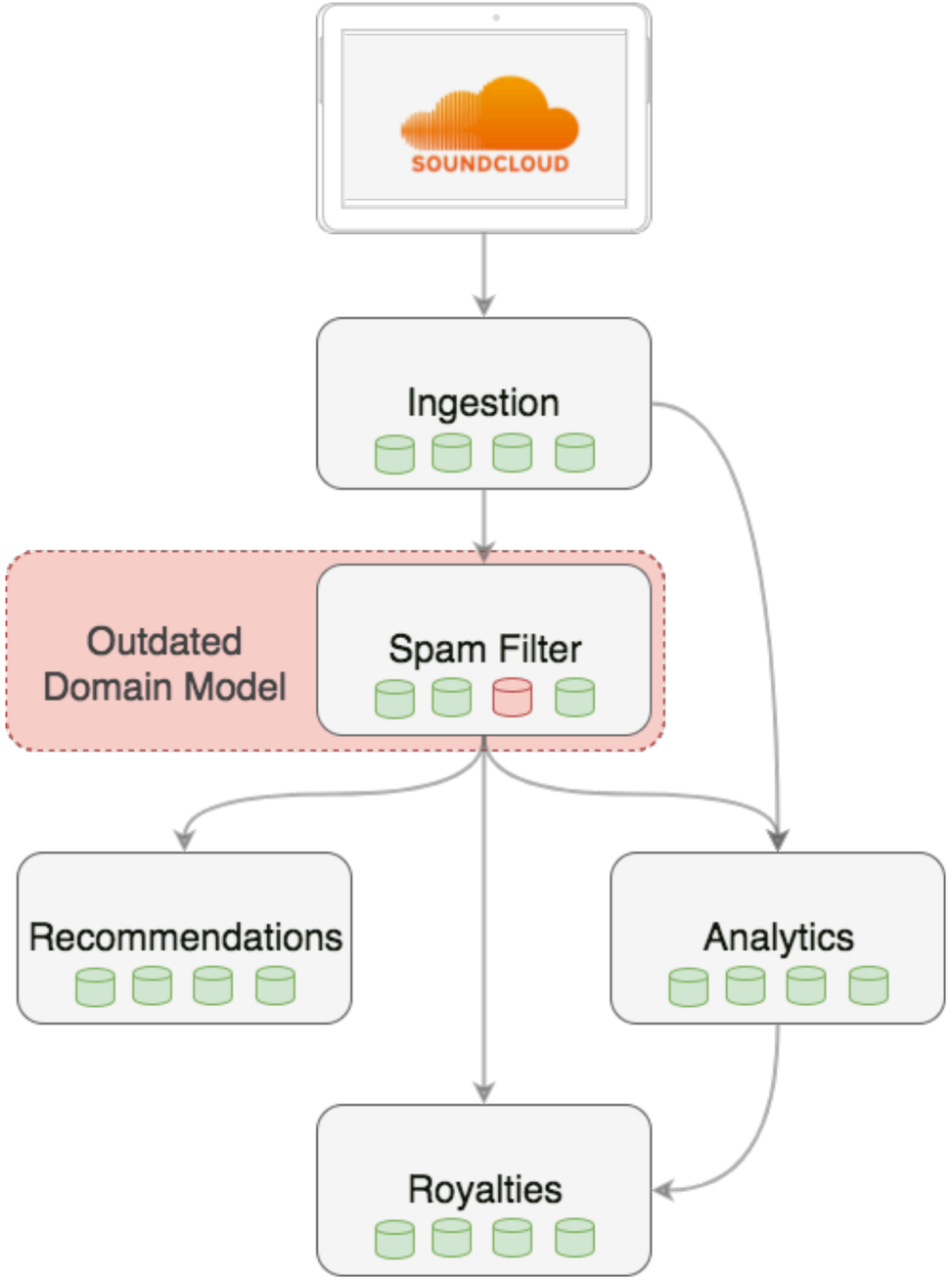


Coordinate Change

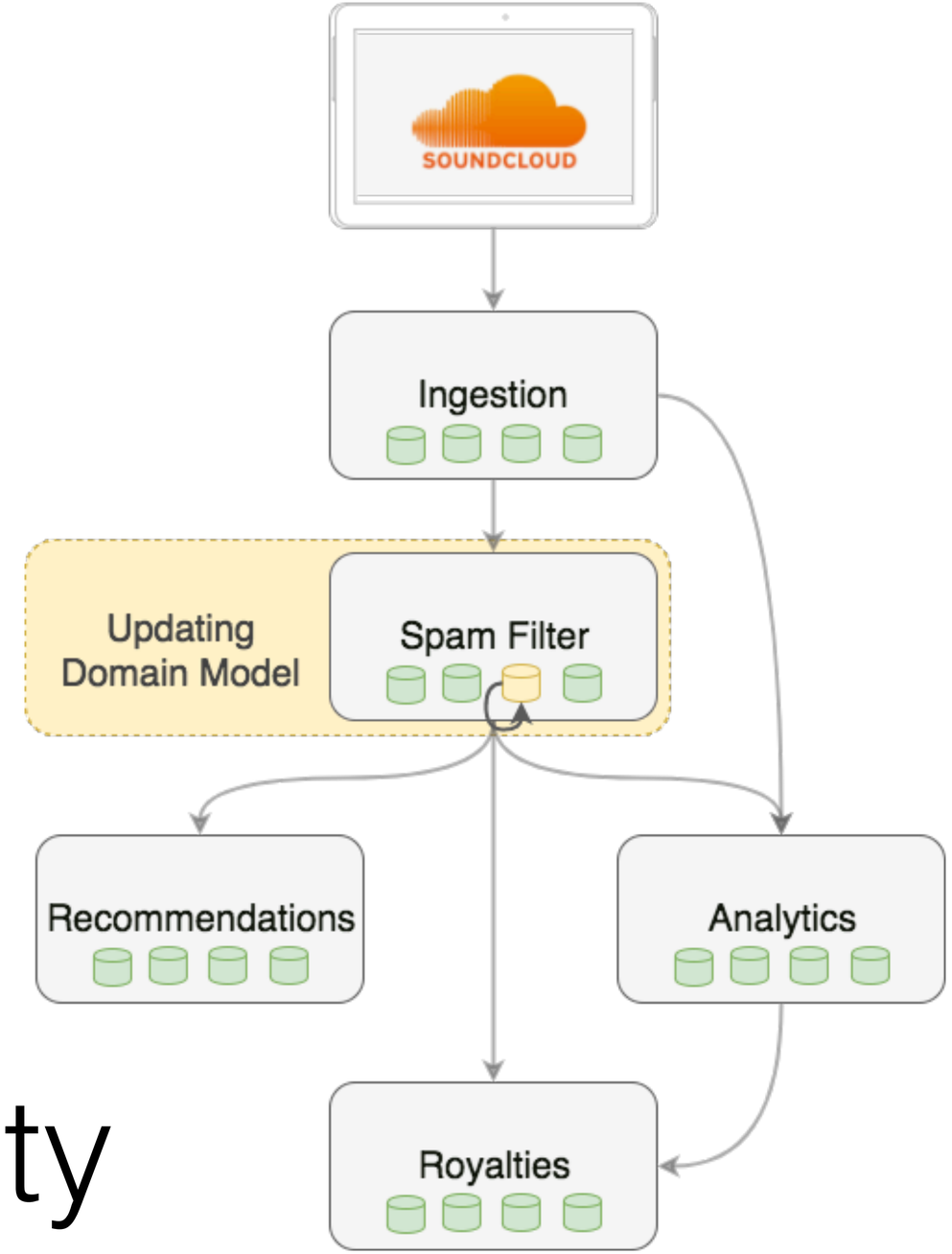
Convergence



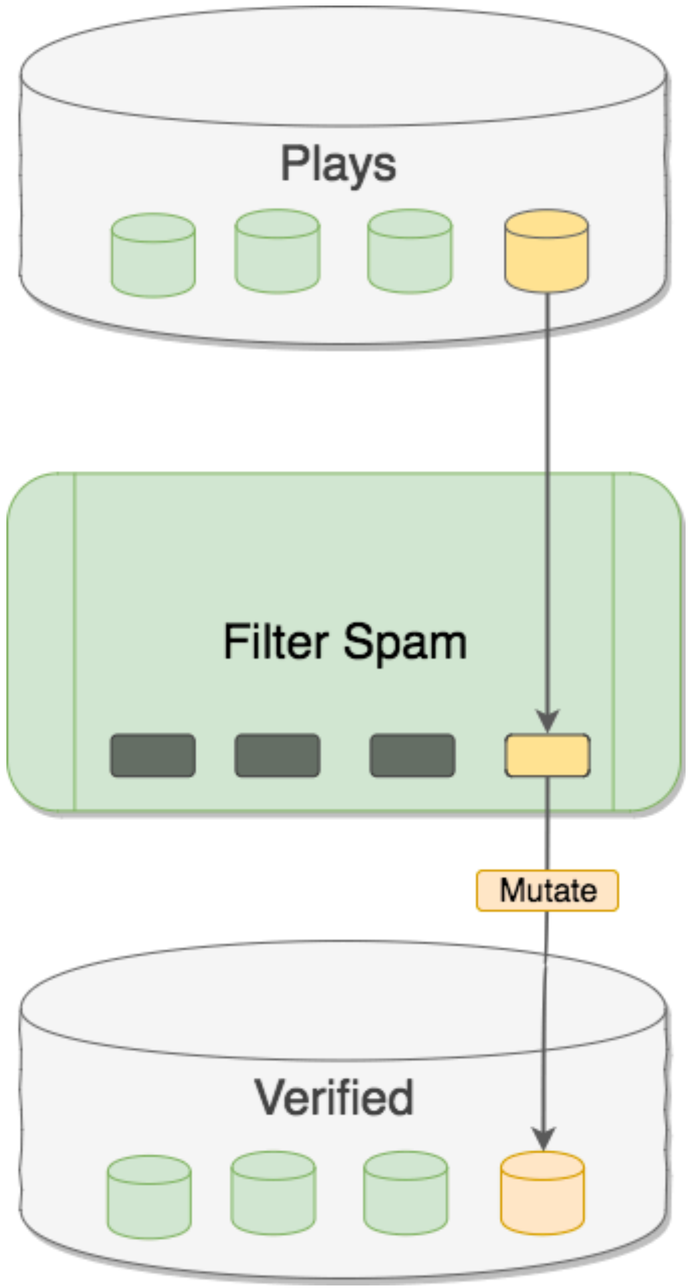
Failure


SOUNDCLOUD
Ingestion
Outdated
Domain Model
Spam Filter
Recommendations
Analytics
Royalties

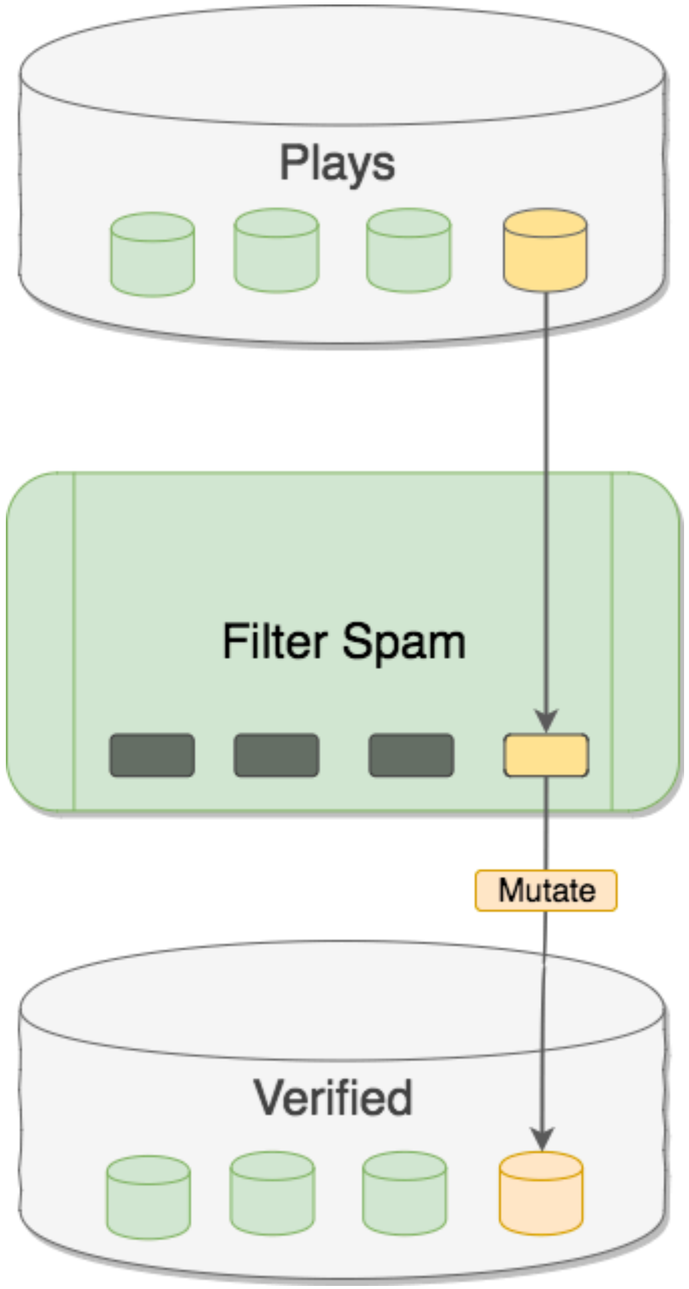
Retroactivity



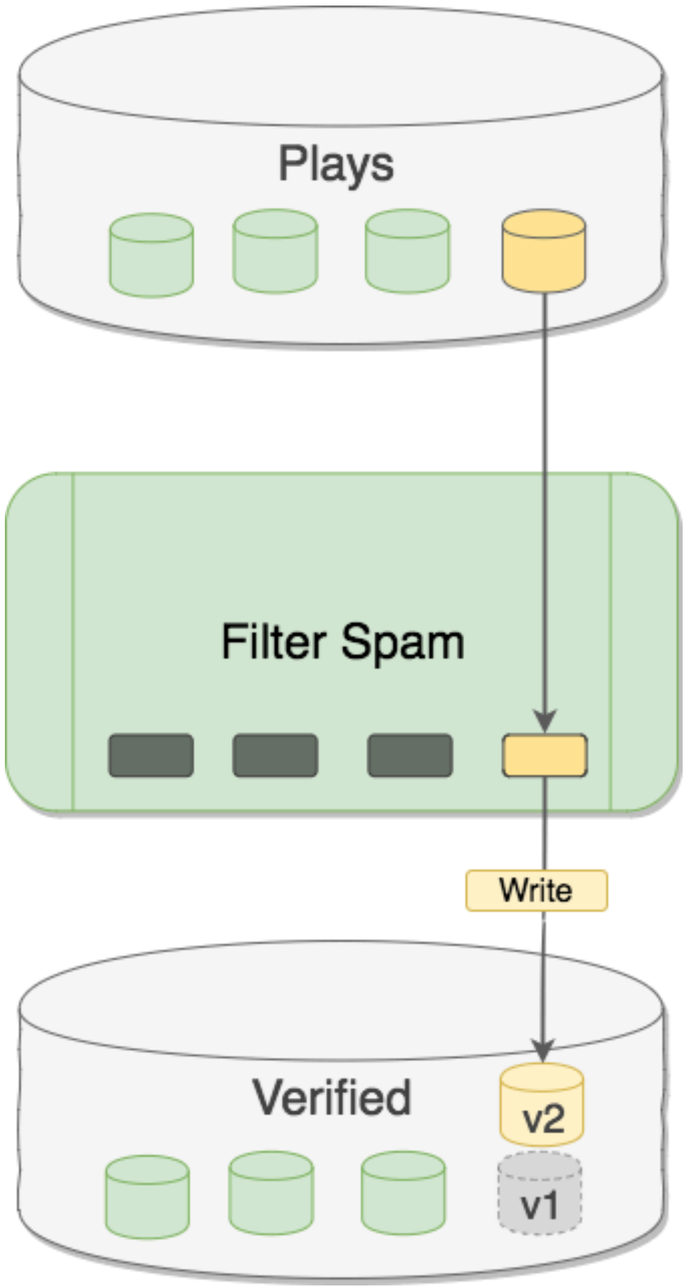
Mutation

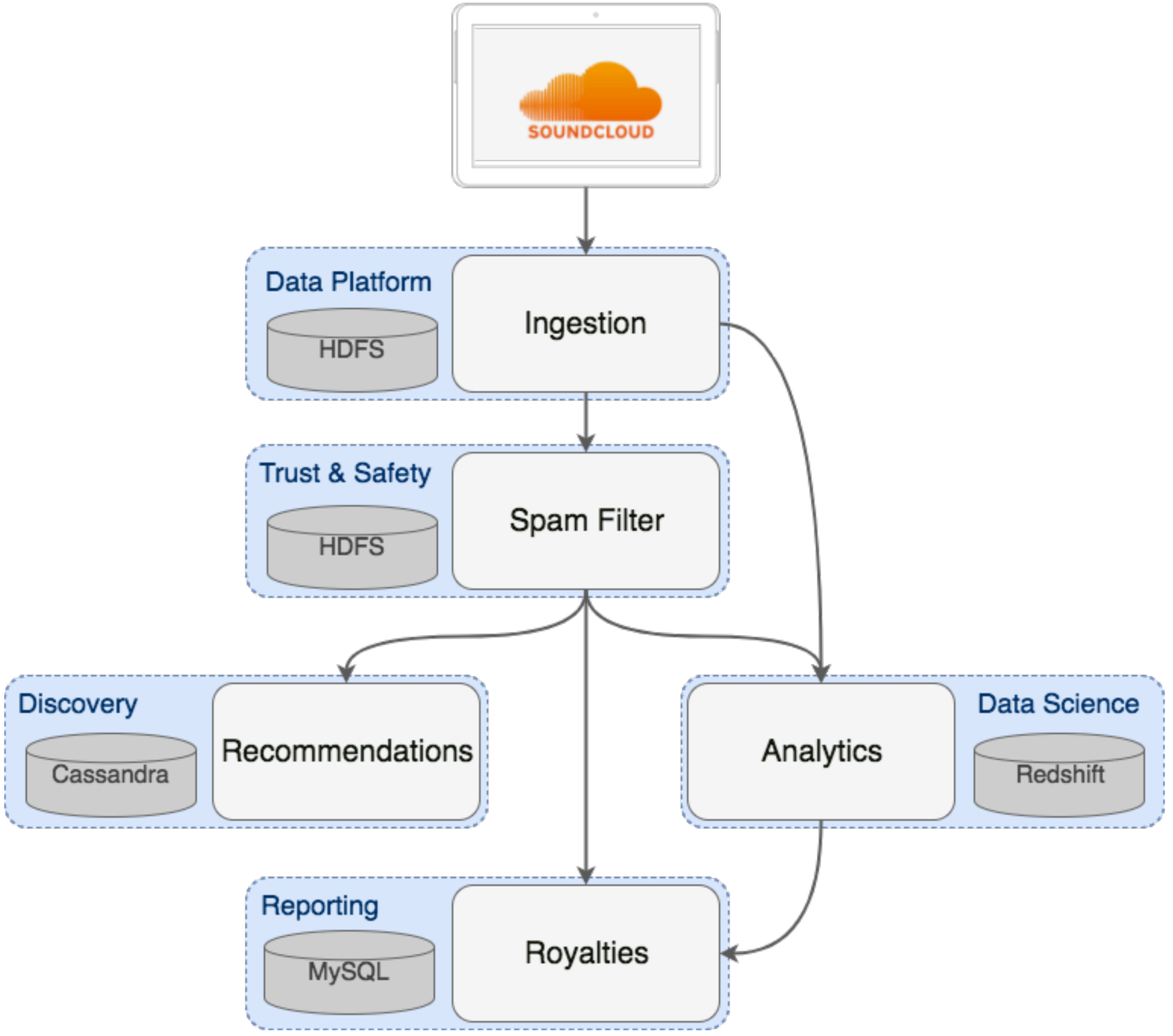


Mutation



Immutability




SOUNDCLOUD
Data Platform
HDFS
Ingestion
Trust & Safety
HDFS
Spam Filter
Discovery
Cassandra
Recommendations
Analytics
Data Science
Redshift
Reporting
MySQL
Royalties

Conclusions

Data pipelines enable coordination but require shared protocols to determine when and how we read and write data.

Emily Green Omid Aladini
Sebastian Ohm Fronx
Wurmus Matthias Georgi
Thank You David Whiting
Lorand Kasler Gavin Bell Jon
Glover Erik Bartels