Erik Isakson - #533764978

Designing Storage Architectures for Digital Collections

Juan Rivera - Getty Images

Key Points

- Data tends to be unstructured
- Backup strategies tend to be weak and untested
- Geographic distances matter and latency hurts
- Replication is needed
- Keep it simple

Dimitri Otis - #151081265

Unstructured Data is Normal

- Digital collections are organic in nature
- Hard to predict size and timing of growth
- Enormous variation in file size and performance requirements
- Randomized access patterns
- Data never shrinks

Jorg Greuel - #463344123

You Can’t Recover from Disaster Quickly

John Lund - #166346390

Tape Backup Doesn’t Work at Scale

A single LTO-6 drive can recover one petabyte in 72 days.

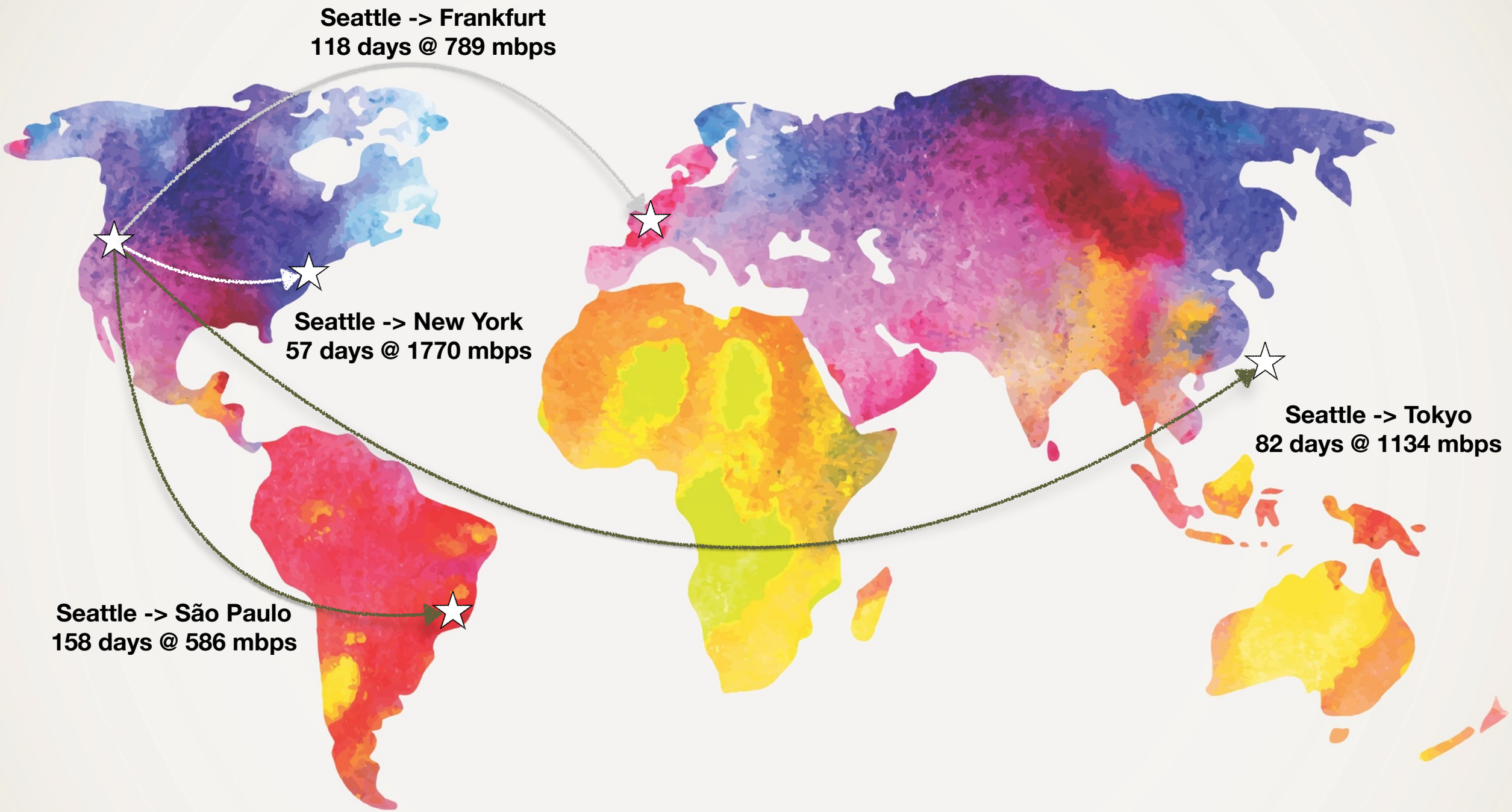
10 drives = 7 days

20 drives = 3.5 days

Don’t Underestimate the Effects of Latency

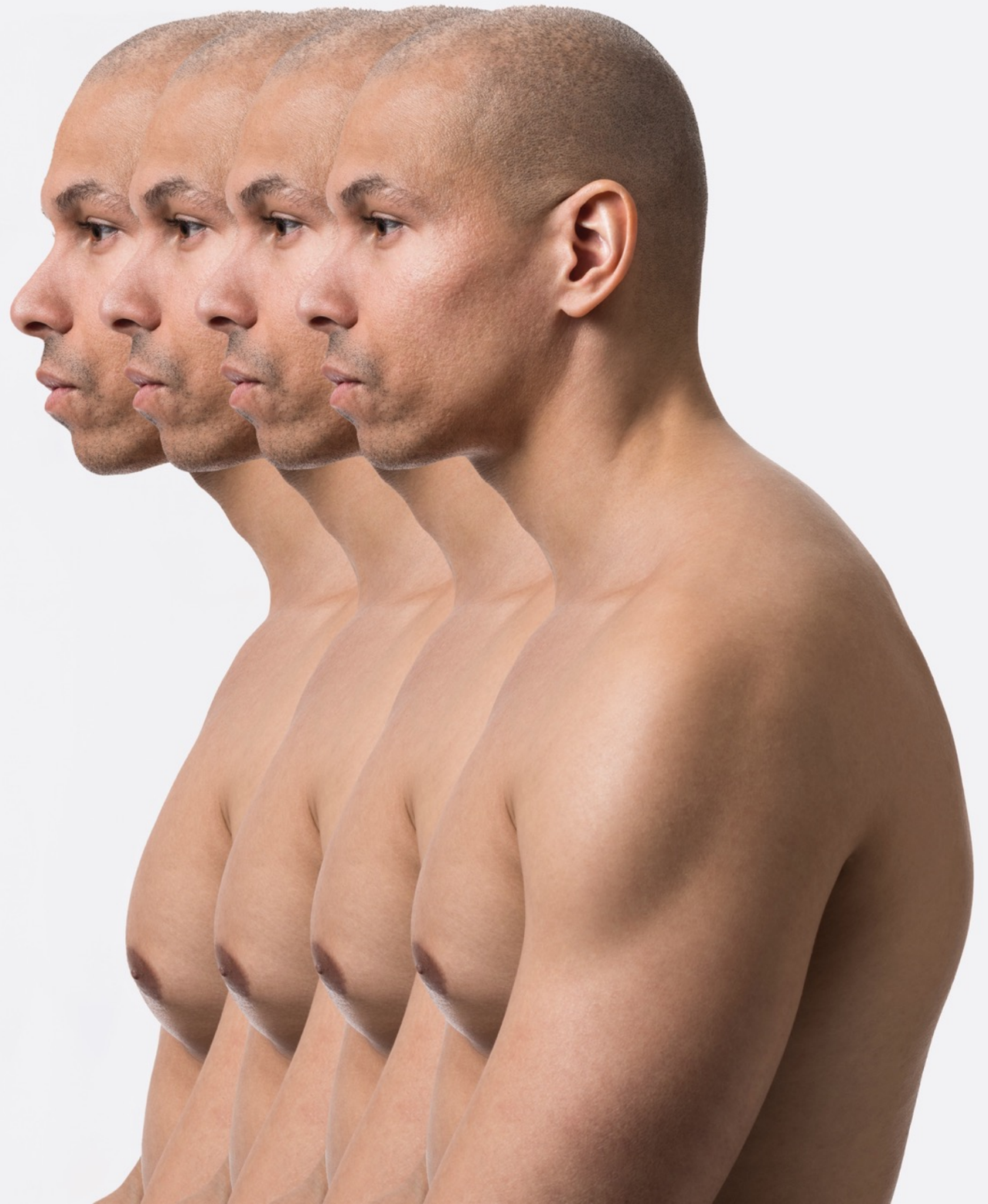
- Customers notice and care about download speeds and durations
- Data transfers across wide area networks will take more time than expected
- Ideal conditions don’t last
- Data transfers are labour intensive and need to be heavily coordinated

Time to Transfer One Petabyte Worldwide



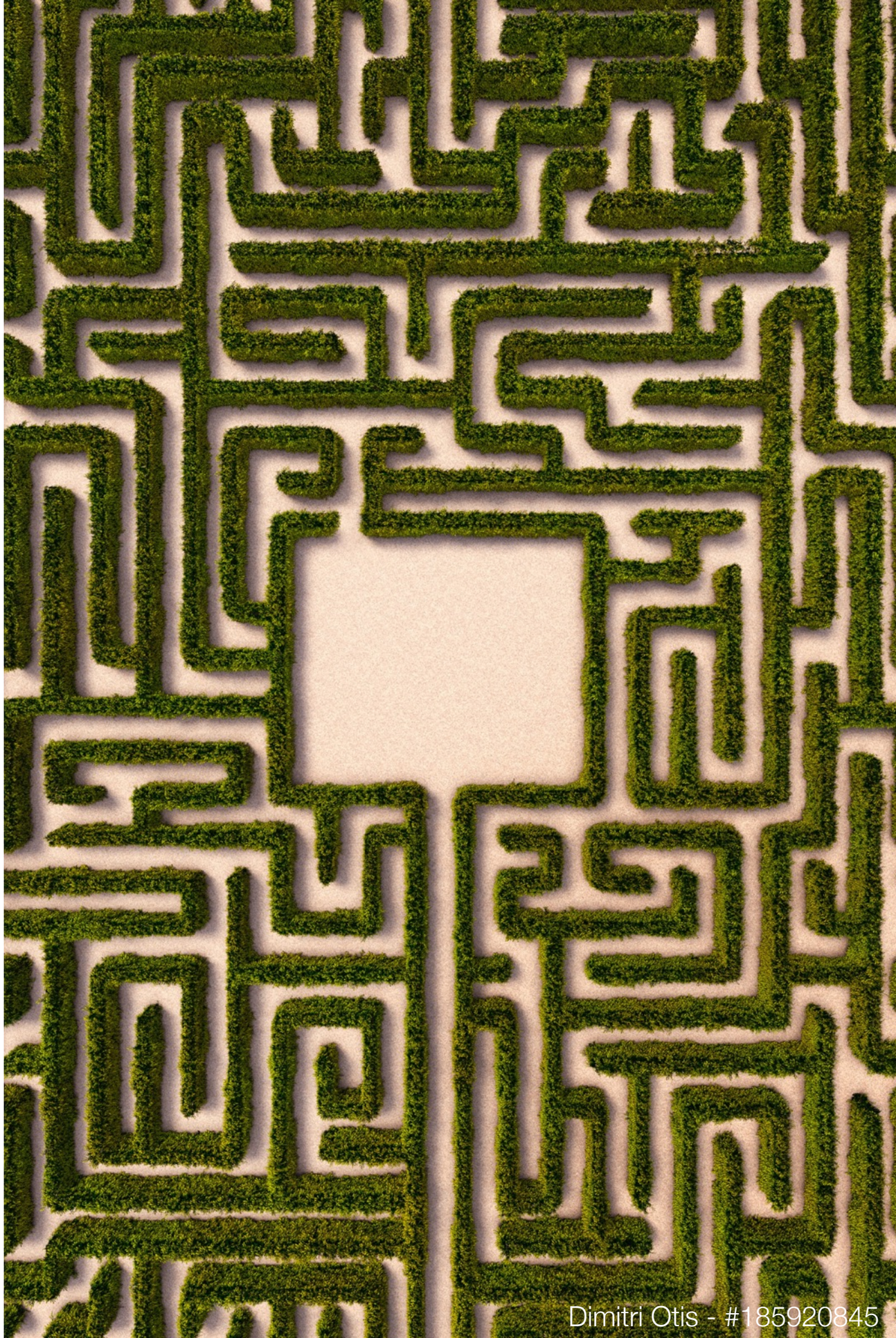
Replication is Key

- Routine or realtime data replication is required based on RTO/RPO
- Datasets are so large that active/active or active/failover systems are required
- Offline data recovery is the only viable alternative
- Geographical separation is highly desirable for disaster recovery



Simplicity Matters

- Object storage systems work well in this space
- Converge down to a single protocol or API to maximize integration
- Deterministic asset lookups centralize business logic
- Develop location aware applications that assume failure will occur



Catherine Ledner - #200497095-001

Questions?

juan.rivera@gettyimages.com