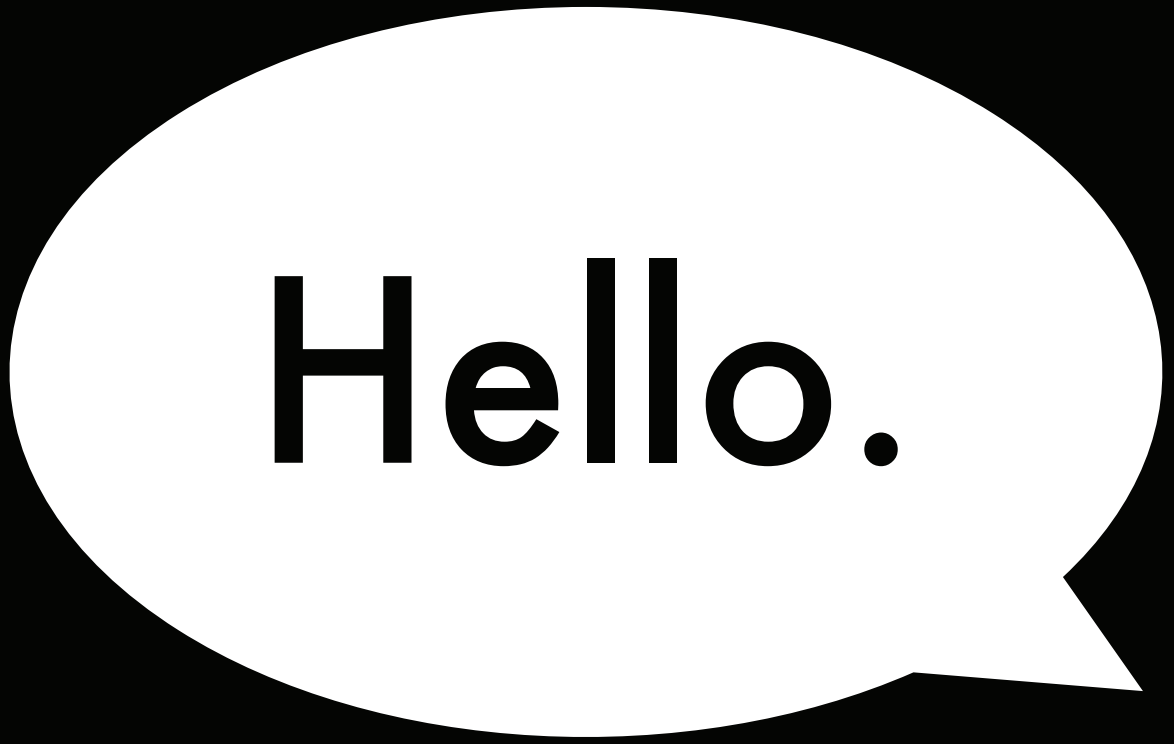
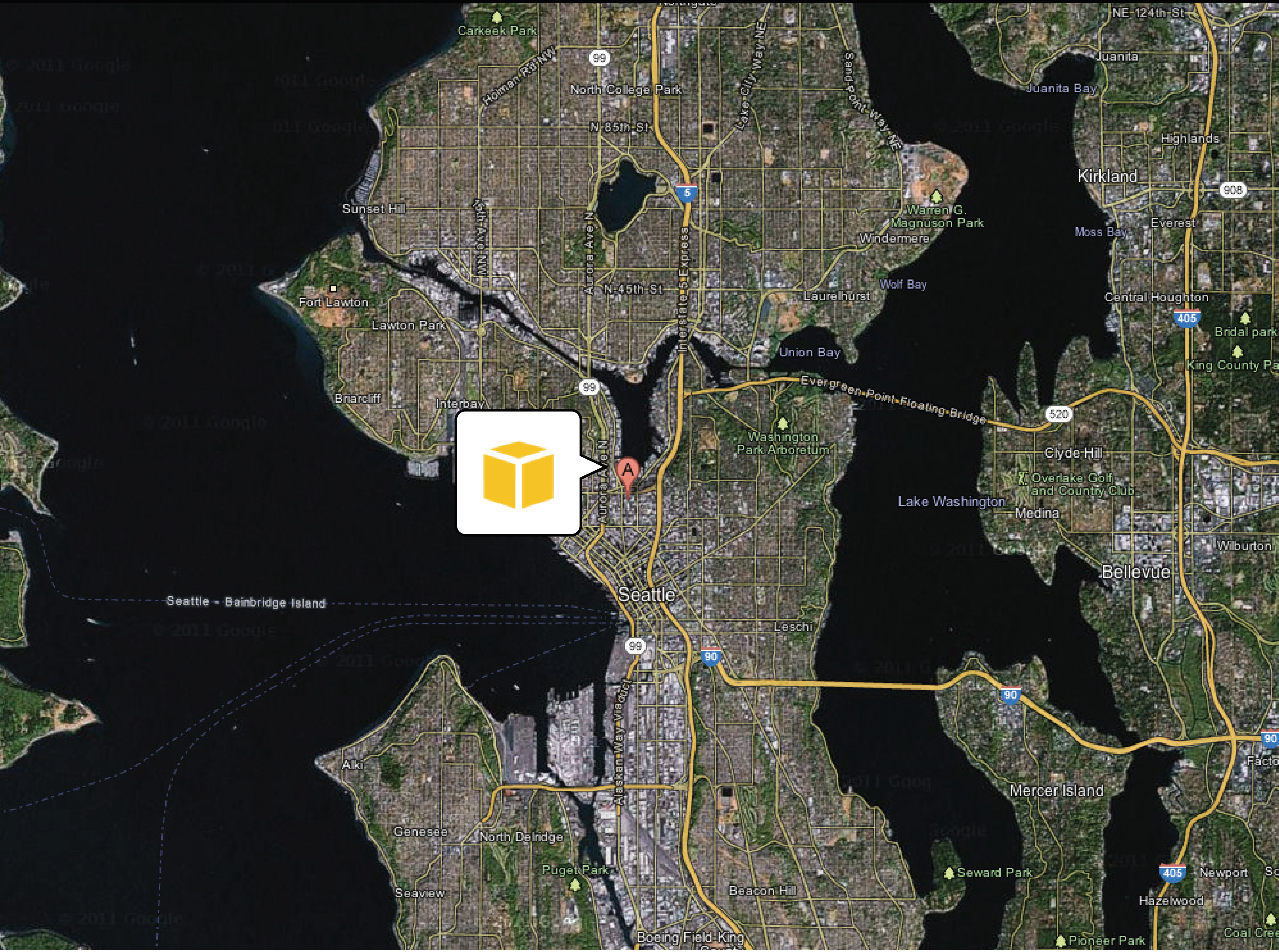
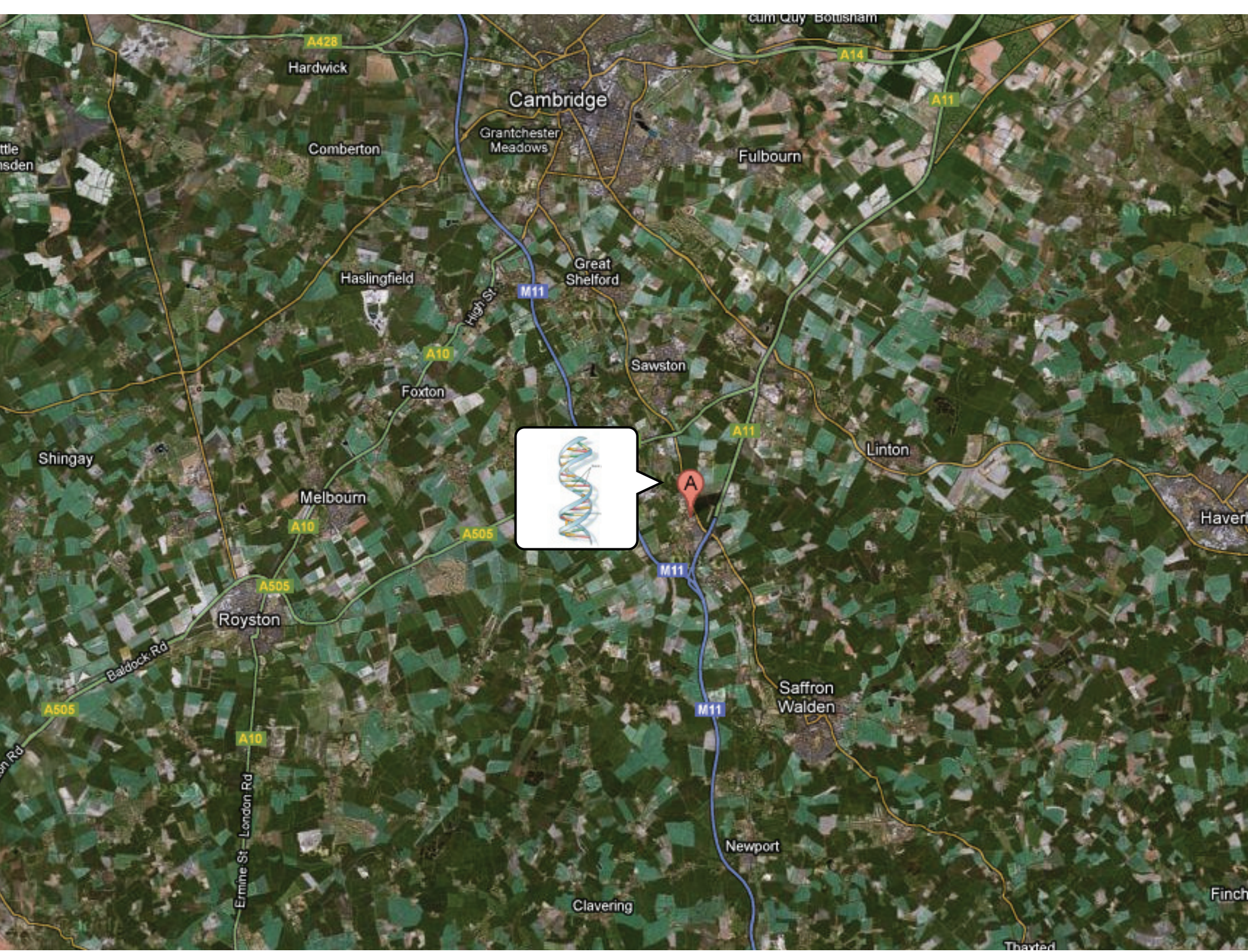
Enabling Research *in the* Cloud

Matt Wood

DATA INTENSIVE & HIGH PERFORMANCE COMPUTING


Hello.


Carkeek Park
Holman Rd NW
99
North College Park
N 85th St
Lake City Way NE
Sand Point Way NE
NE 124th St
Juanita
Juanita Bay
Highlands
Kirkland
908
5
Sunset Hill
15th Ave NW
Aurora Ave N
Interstate 5 Express
Warren G. Magnuson Park
Windermere
Moss Bay
Everest
Fort Lawton
Lawton Park
N 45th St
Laurelhurst
Wolf Bay
Central Houghton
405
Bridal park
King County Pa
Union Bay
Briarcliff
Interbay
Evergreen Point Floating Bridge
520
Washington Park Arboretum
Clyde Hill
Overlake Golf and Country Club
Lake Washington
Medina
A
Wilburton
Bellevue
Seattle
Seattle - Bainbridge Island
Leschi
90
Alaskan Way Viaduct
Alki
Facto
Mercer Island
Genesee
North Delridge
Puget Park
Seward Park
Newport
Seaview
Beacon Hill
Hazelwood
Boeing Field-King
Pioneer Park
Coal Cree


A428
Hardwick
Cambridge
A14
A11
Grantchester Meadows
Comberton
Fulbourn
Haslingfield
Great Shelford
M11
High St
A10
Sawston
Foxton
A11
Linton
Shingay
A
Melbourn
A10
A505
M11
A505
Royston
Baldock Rd
Saffron Walden
A505
M11
A10
London Rd
Ermine St
Newport
Clavering

Thank you

Infrastructure building blocks


amazon.com


amazon.com
Consumer business
Seller business
amazon web services™

Decades of experience

Operations, management and scale

Programmatic access

Unexpected innovation

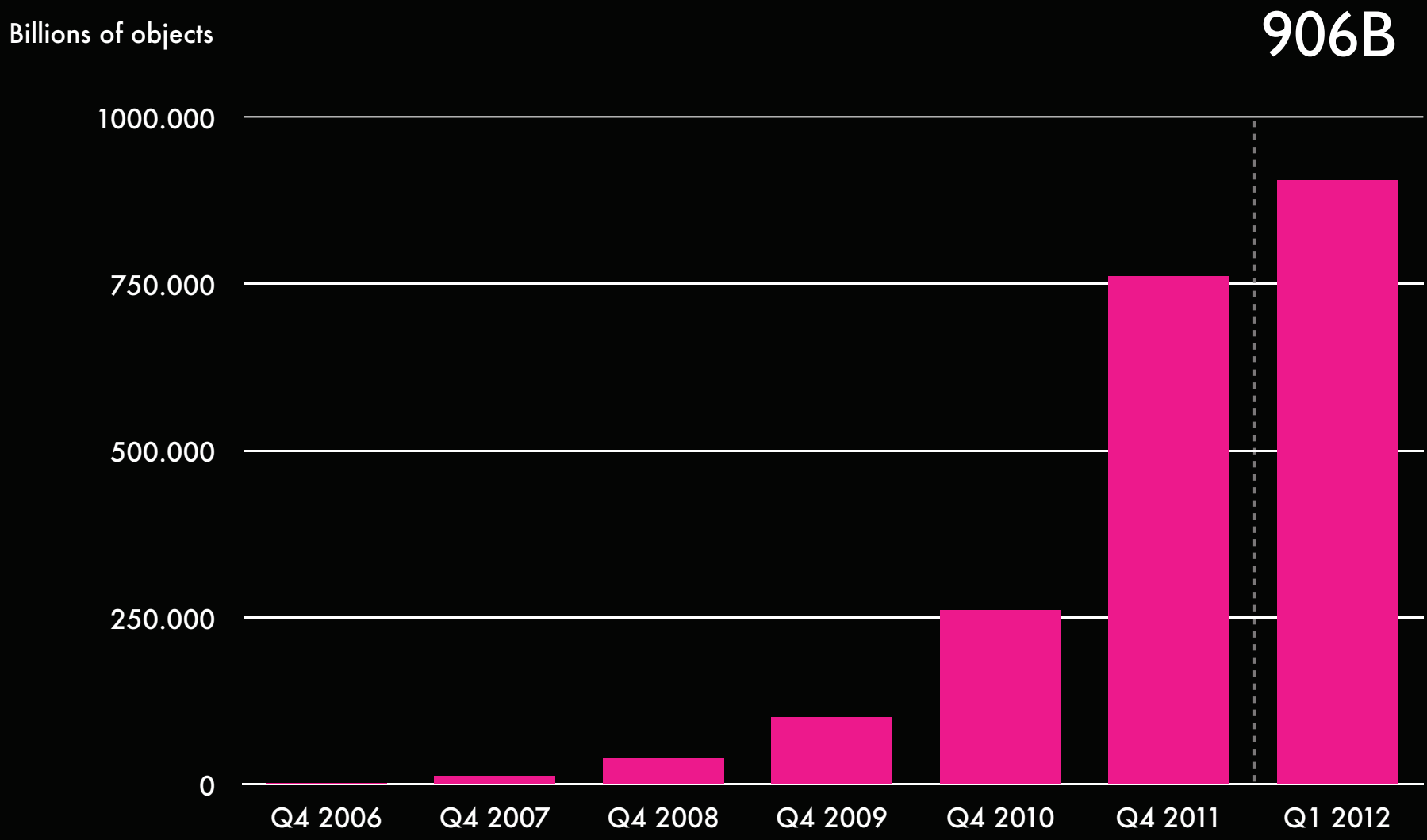
Blinding flash of the obvious


amazon
web services™

6 years young

Amazon S3 launched on March 14th, 2006

Objects in S3

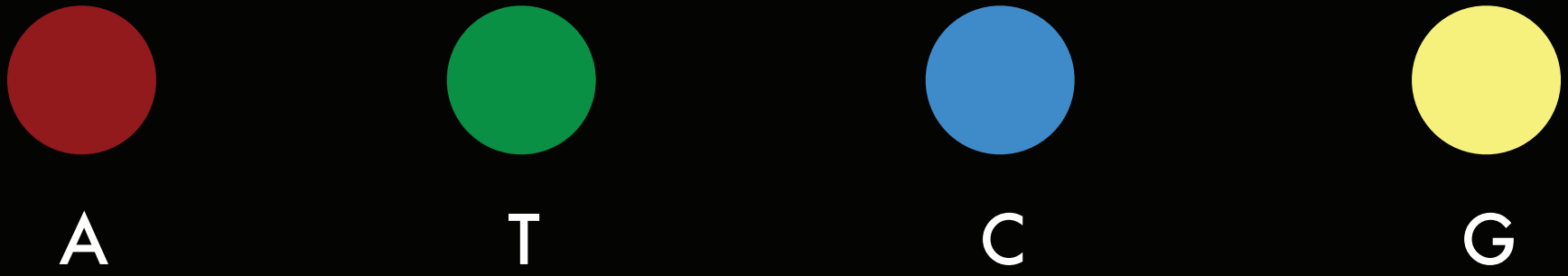


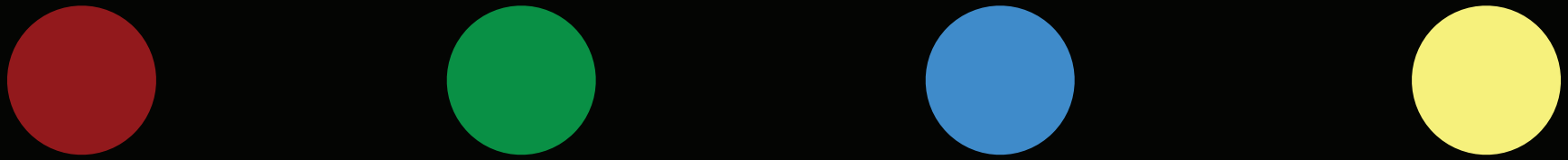
600k+ peak transactions per second

99.999999999%

durability

Life sciences


A
T
C
G


Storage
Compute
Databases
Tools

Collection → Computation → Collaboration


Collection
Computation
Collaboration

Collection  Computation  Collaboration


Collection
Computation
Collaboration

Availability

Programmable

Availability

On-demand

Flexibility

Elasticity

Collection → Computation → Collaboration


Collection
Computation
Collaboration


AWS Regions
Amazon Edge Locations

Data stays local

Availability zones

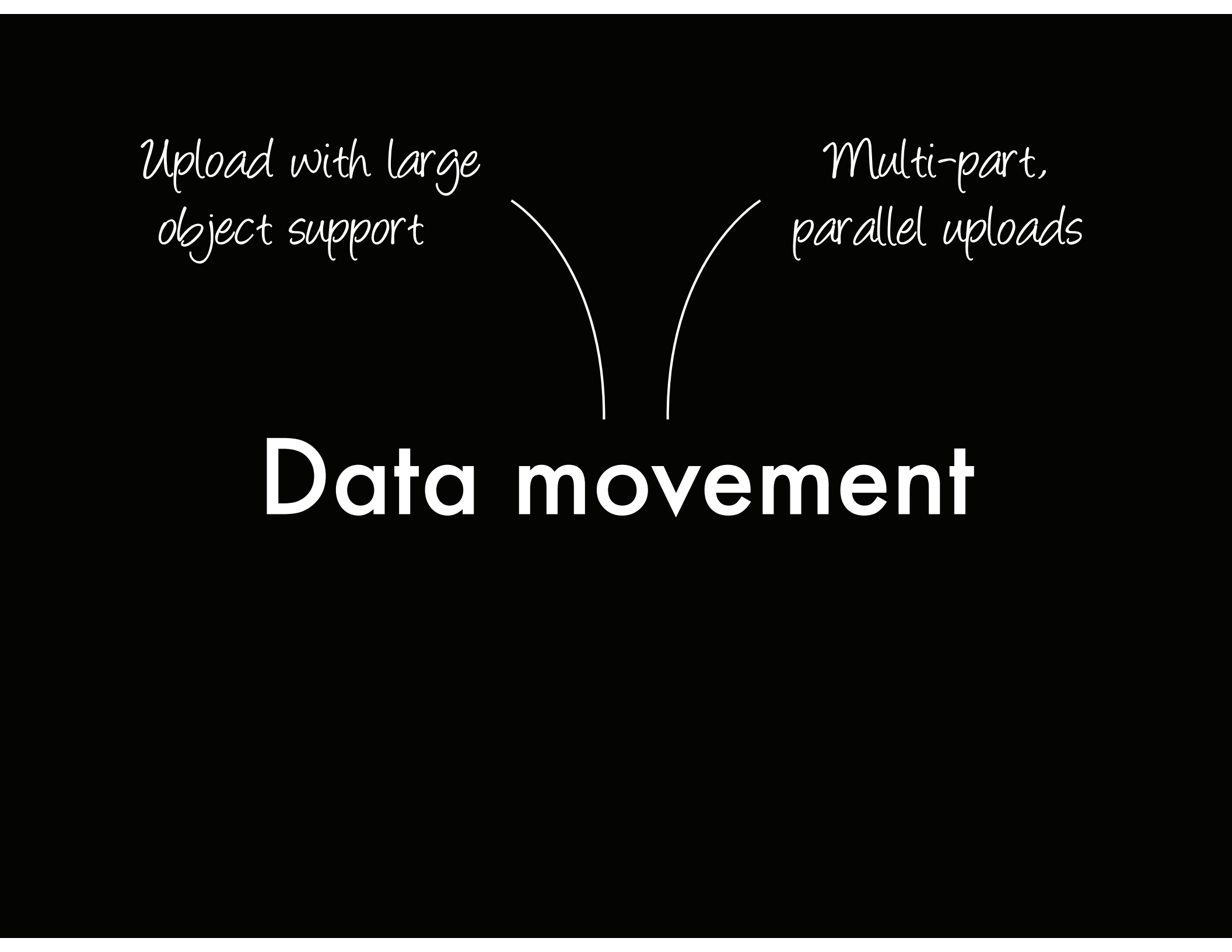
Design for durability

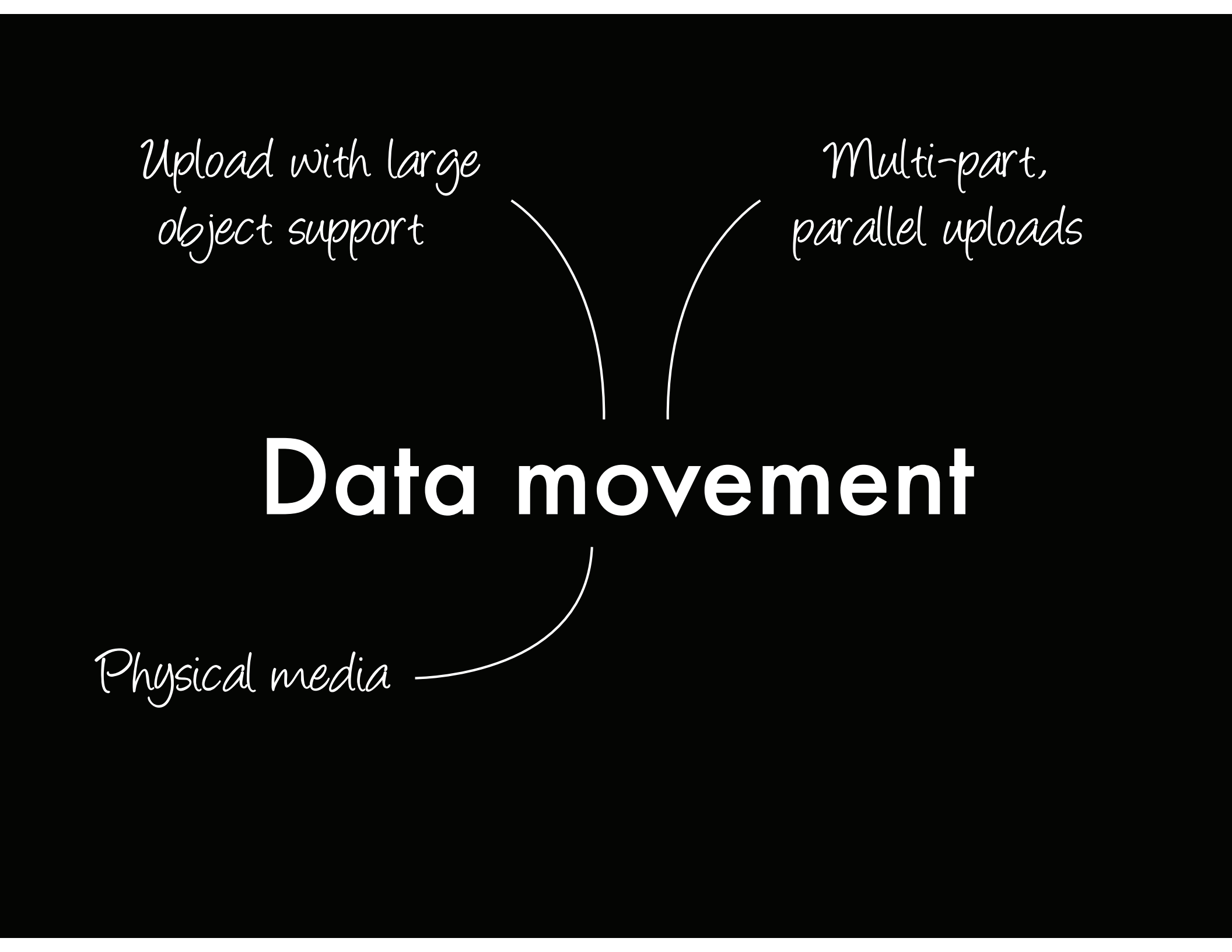
Shared responsibility

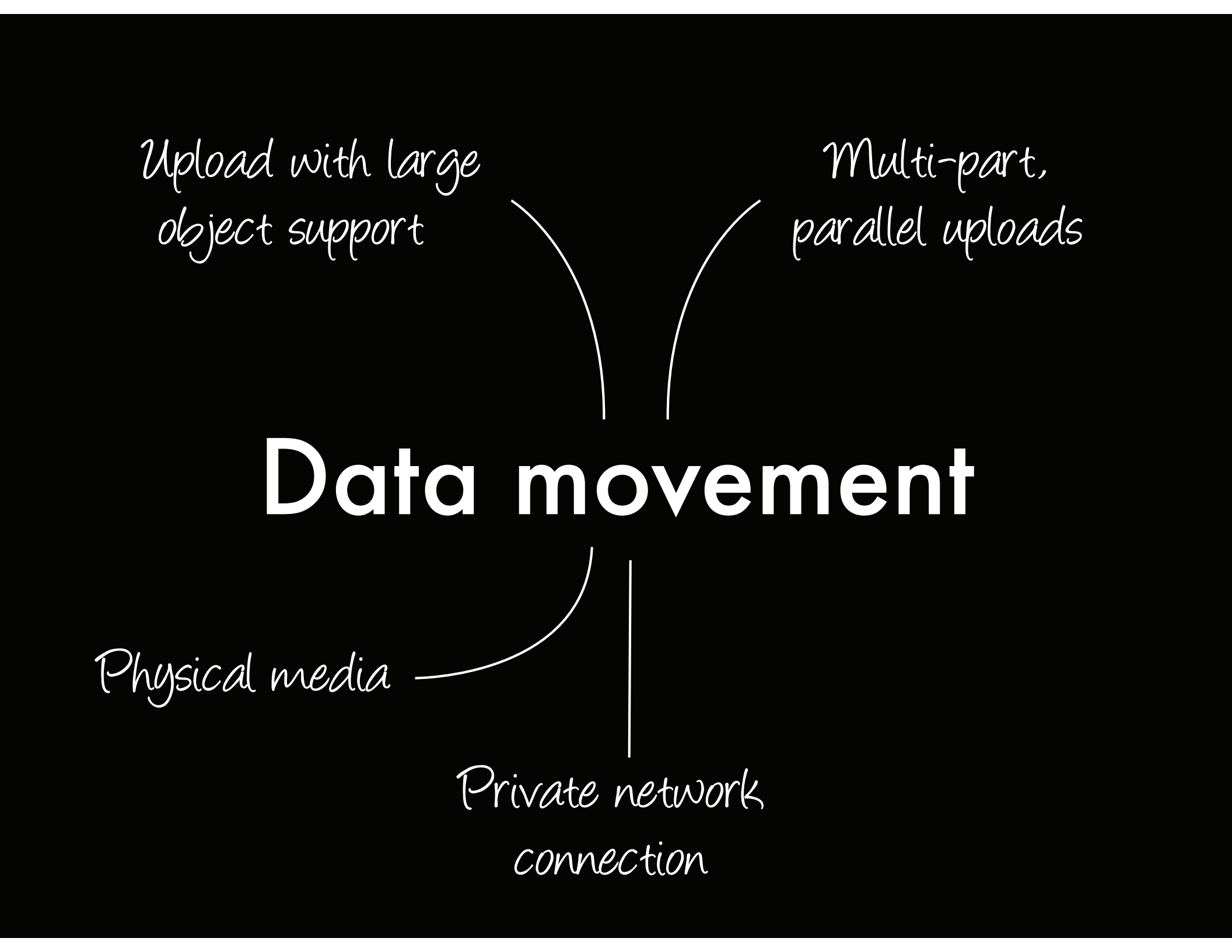
Data movement

Upload with large object support

Data movement


Upload with large object support
Multi-part, parallel uploads
Data movement


Upload with large object support
Multi-part, parallel uploads
Data movement
Physical media


Upload with large object support
Multi-part, parallel uploads
Data movement
Physical media
Private network connection

AWS Direct Connect

Direct connection to AWS regions

Consistent network performance

Private connectivity

Elastic

1Gbps and 10 Gbps

Reduced bandwidth costs

ISP and lower Direct Connect pricing

Globus Online

3.8 PB moved (as of this morning!)

Aspera

Public Datasets

1000 Genomes Project

aws.amazon.com/1000genomes


Collection
Computation
Collaboration

Scale

Scale

How much can I get?
What size will get me time most quickly?

Value

How much do I need?
What value does it have for me?

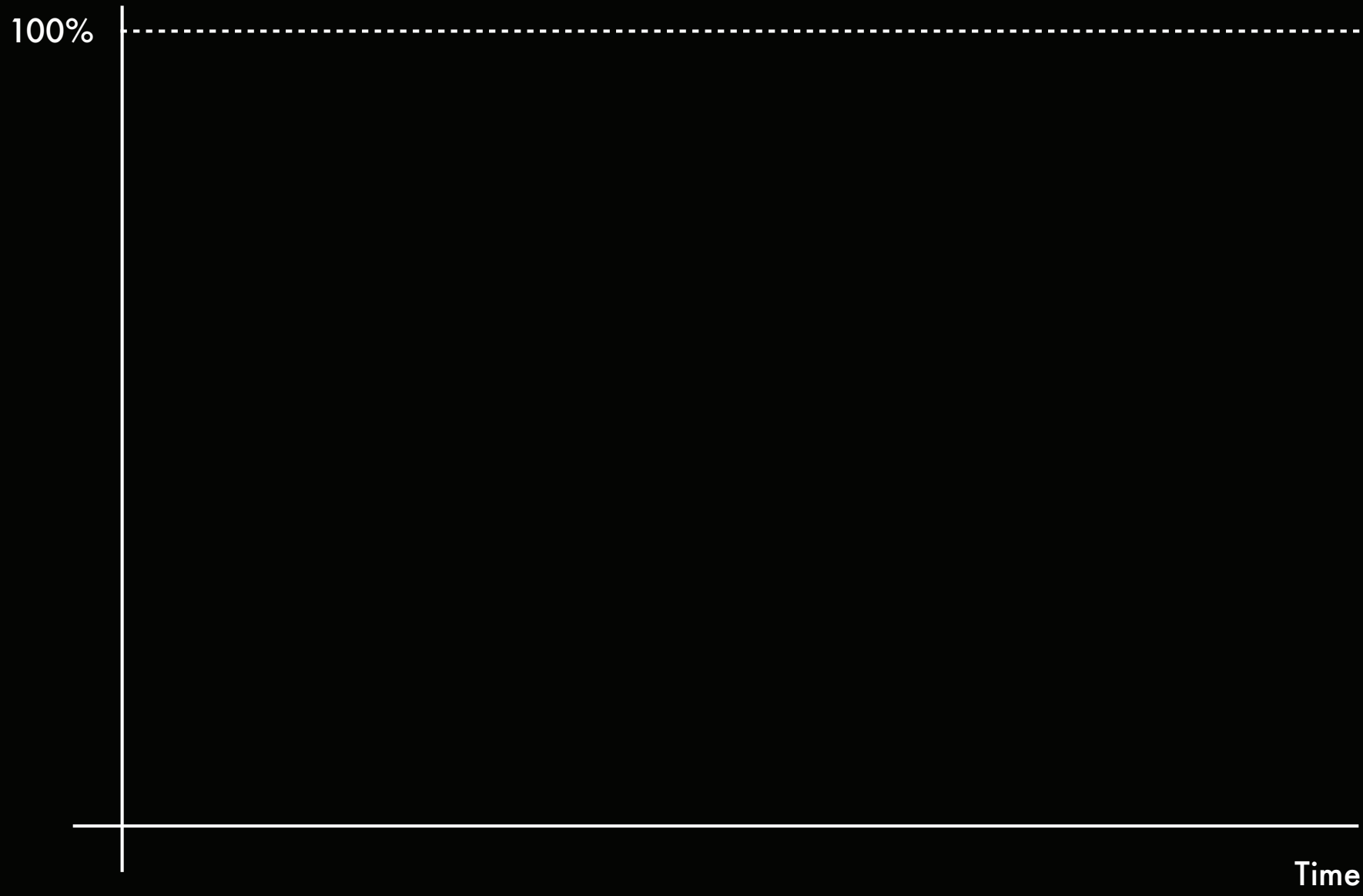
Economies of scale

19 price drops

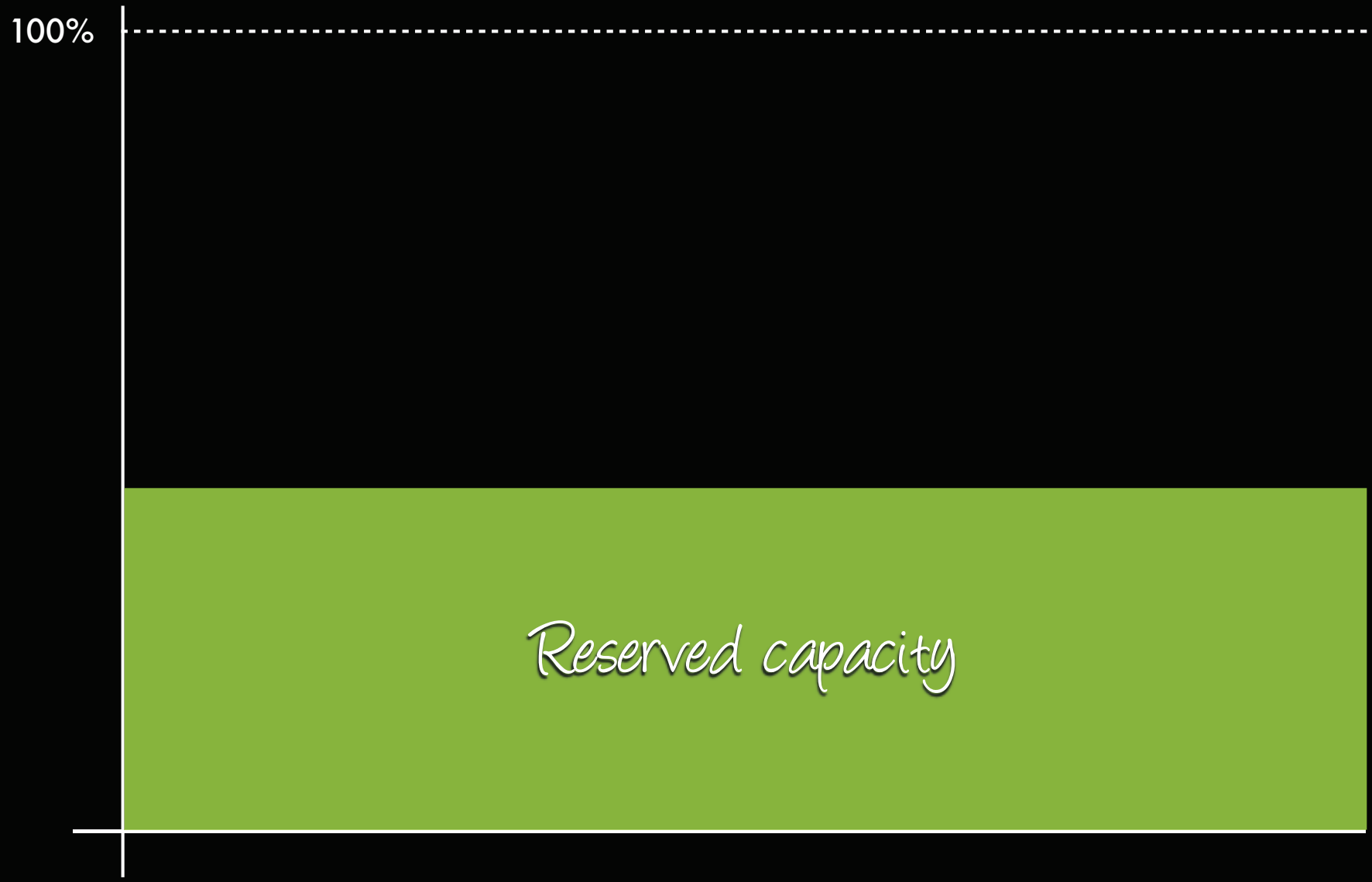
Committed to passing savings to customers

Utilisation

Achieving economies of scale

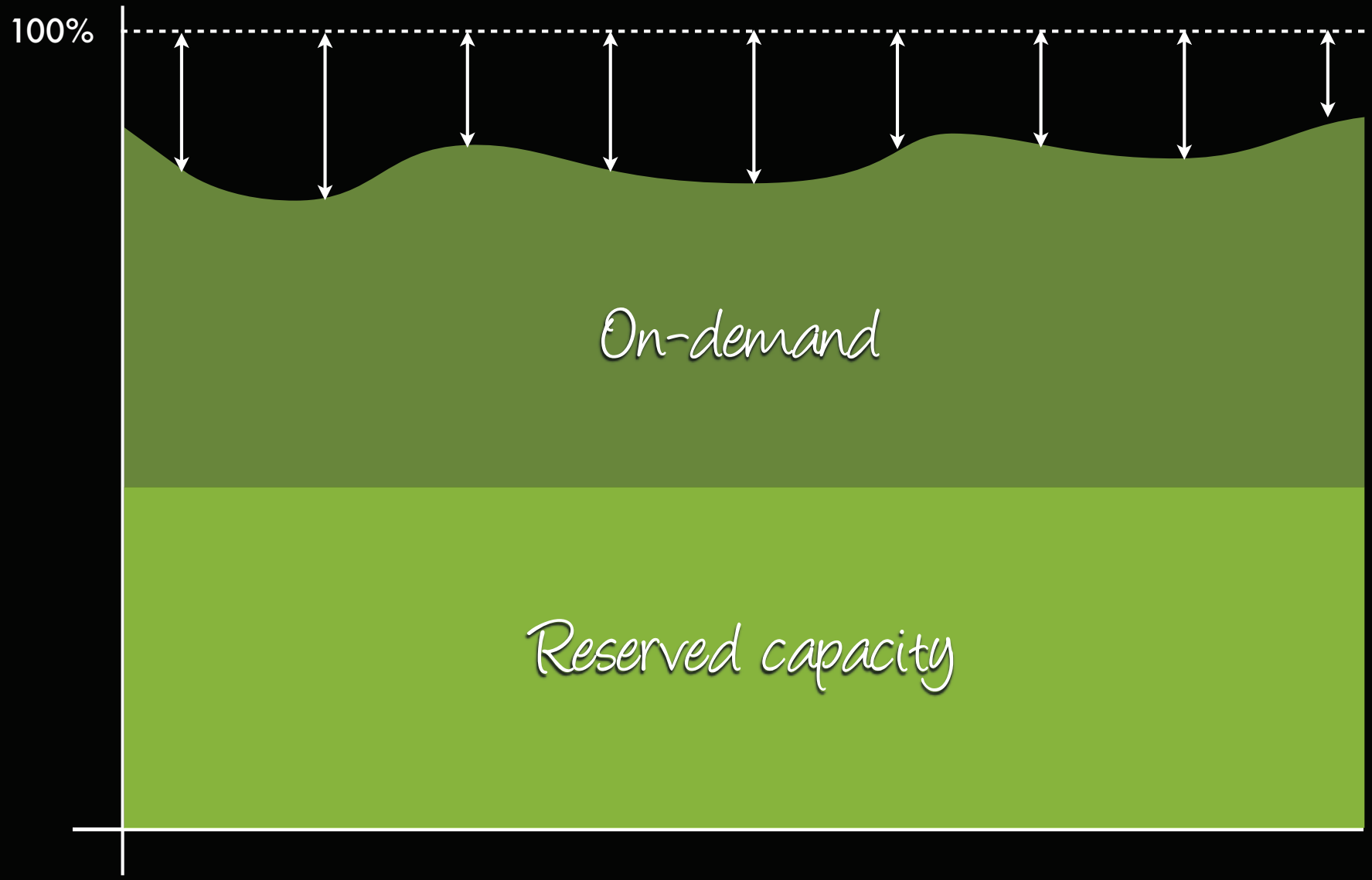


Achieving economies of scale



Achieving economies of scale

Achieving economies of scale



Spot market

Choose your own price for compute

Scale out

30k cores

On the spot market. $1279 per hour.

50k cores

Schrodinger and Cycle Computing

51,132 cores

Schrodinger and Cycle Computing

6742 instances. $4828 per hour

Elastic MapReduce

Myrna. Crossbow.

Scale up

CC2

Tightly coupled workflows

240 TFLOPS

42nd fastest supercomputer

Scale cores

GPU on demand

AMBER

Scale?

Getting stuff done

StarCluster

Cloud BioLinux

Ready to roll with 1000 Genomes data


Collection
Computation
Collaboration

Galaxy

GenomeSpace: Home

www.genomespace.org

GENOMESPACE

What is GenomeSpace? Tools Documentation Developers Support About

GENOMESPACE BETA

Request an invitation to join GenomeSpace

GenomeSpace

GenomeSpace brings together diverse computational tools and enables scientists without programming skills to easily combine the capabilities of these tools. It offers a common space to create, manipulate, and share an ever-growing range of genomic analysis tools.

HIGHLIGHTS

GenomeSpace at BioITWorld

GenomeSpace will be featured in the talk "GenomeSpace: An environment for frictionless bioinformatics" at BioITWorld in Boston on Wednesday, April 25 at 4:45 pm. Please see the schedule for the Bioinformatics track for more information. Also see the GenomeSpace poster in the exhibit hall.

GenomeSpace at ISMB 2012

GenomeSpace will be featured at the Bioinformatics Software Interoperability SIG at ISMB 2012 in Long Beach CA, on July 13. Learn about GenomeSpace and how you can add your genomics tool

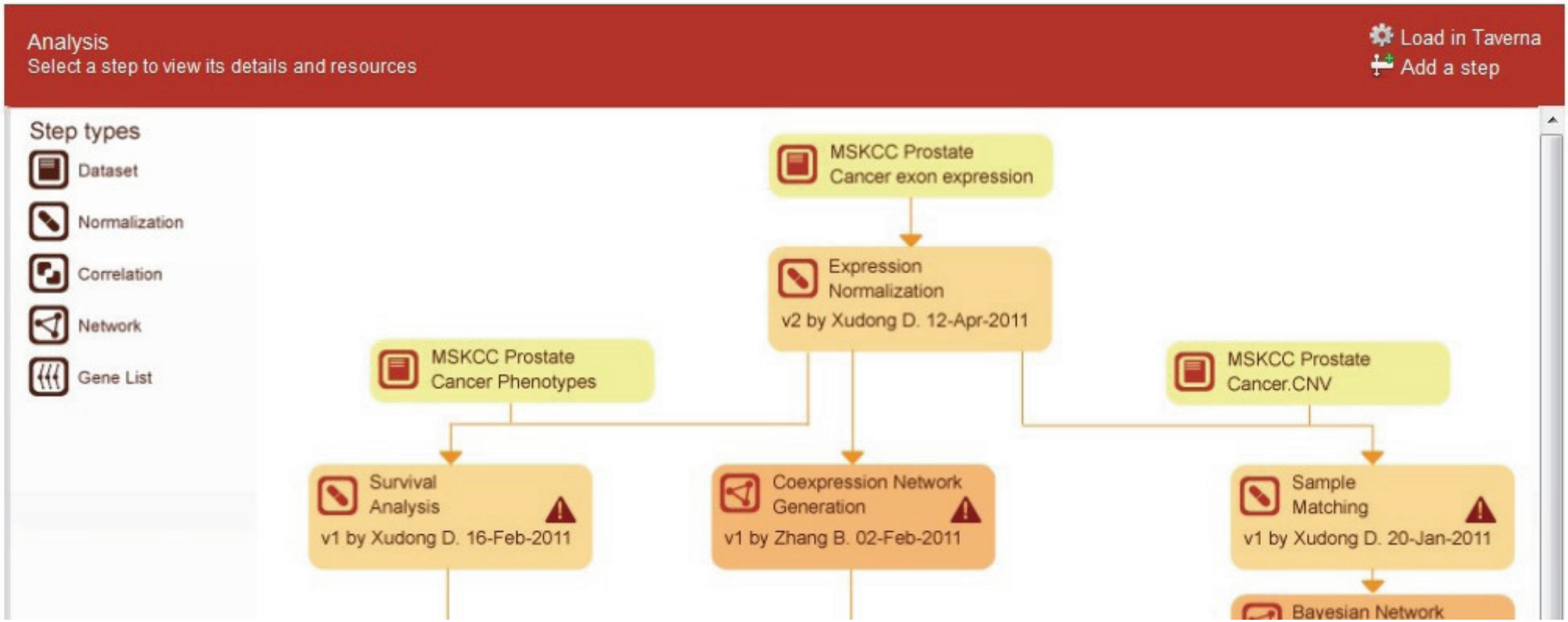
@genomespace

InSilicoDB @InSilicoDB talk starting in 10 minutes at #BioIT12. Learn how to send your genomic datasets to @genomespace, and much more.
16 hours ago · reply · retweet · favorite

Blancahimes JM: Possible solutions for interoperability motivated creation of genomespace #BioIT12
yesterday · reply · retweet · favorite

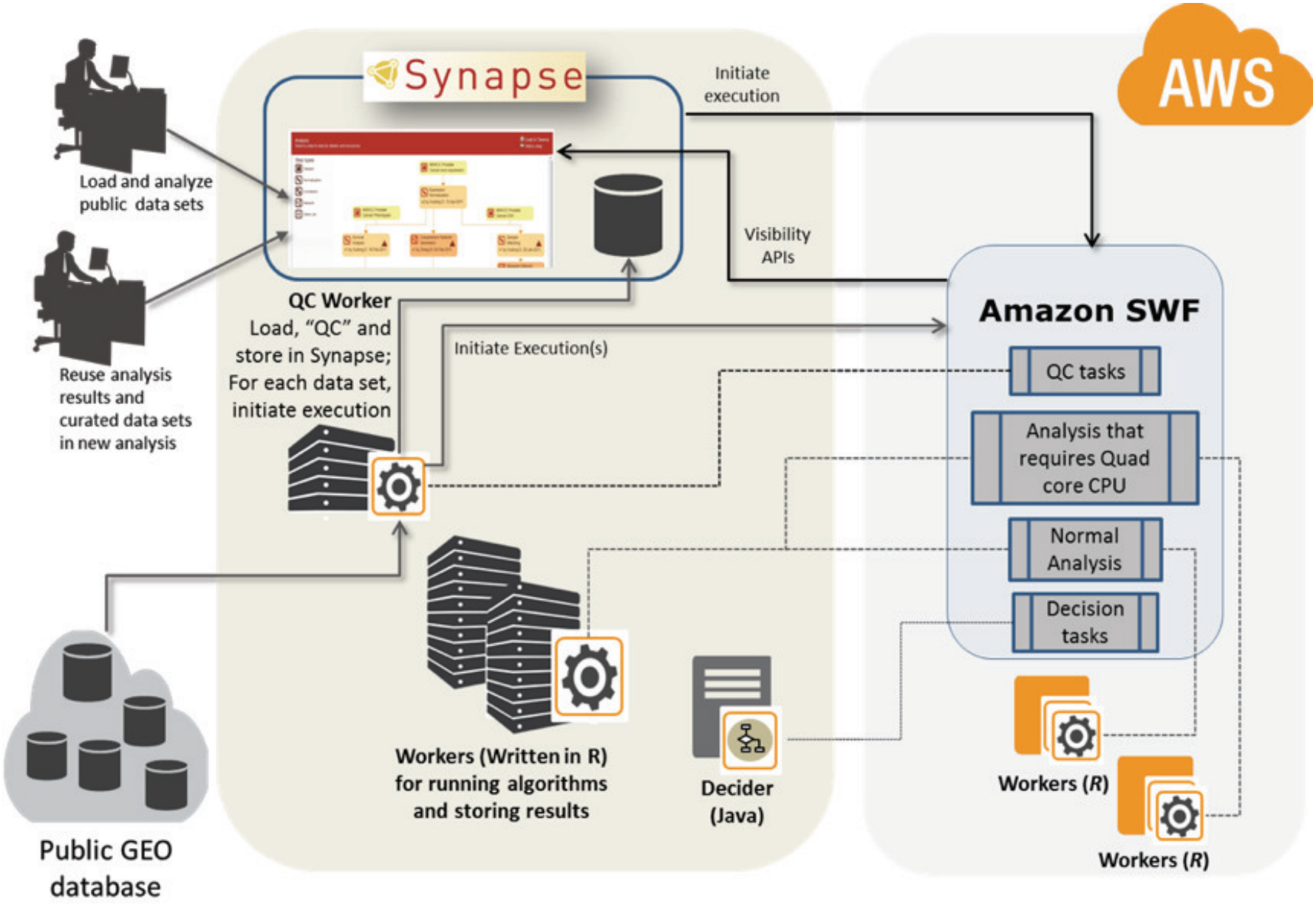
Personal_RX_WI GenomeSpace from Broad. Online community to share

Sage BIONETWORKS



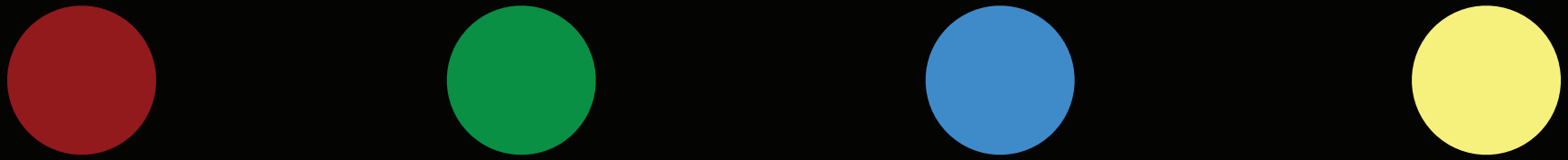
synapse.sagebase.org

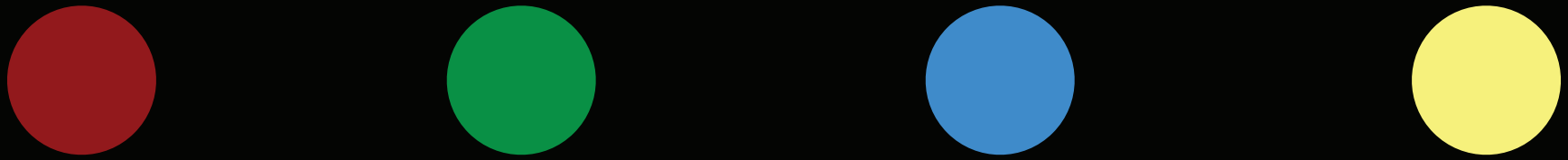
Collaboration platform for clinical genomic datasets


Synapse
Initiate execution
AWS
Load and analyze public data sets
Visibility APIs
Reuse analysis results and curated data sets in new analysis
QC Worker
Load, "QC" and store in Synapse; For each data set, initiate execution
Initiate Execution(s)
Amazon SWF
QC tasks
Analysis that requires Quad core CPU
Normal Analysis
Decision tasks
Workers (Written in R) for running algorithms and storing results
Decider (Java)
Workers (R)
Workers (R)
Public GEO database

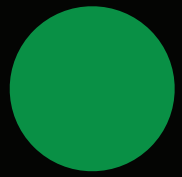
AWS for Education

aws.amazon.com/education

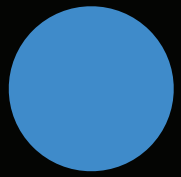

Storage
Compute
Databases
Tools


Materials
Methods
Hypotheses
Results

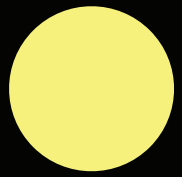
Data



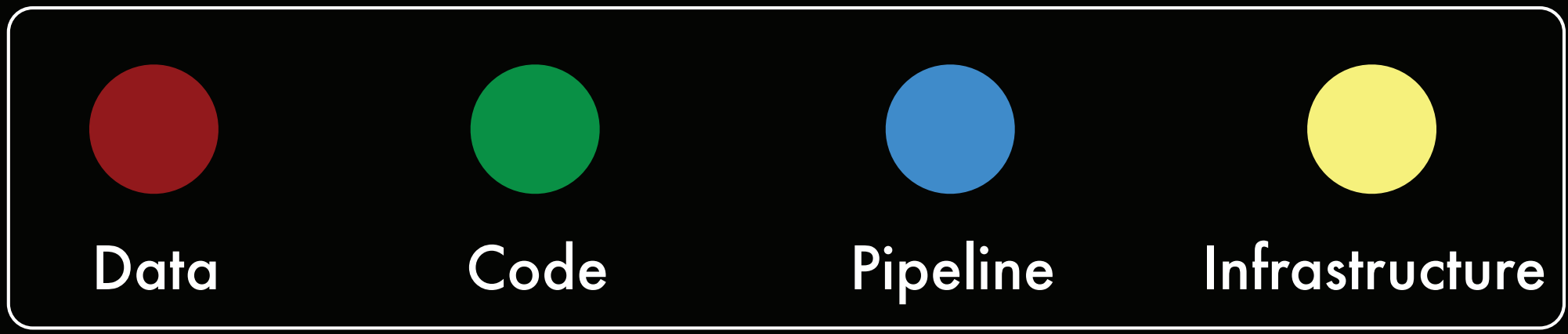
Code



Pipeline

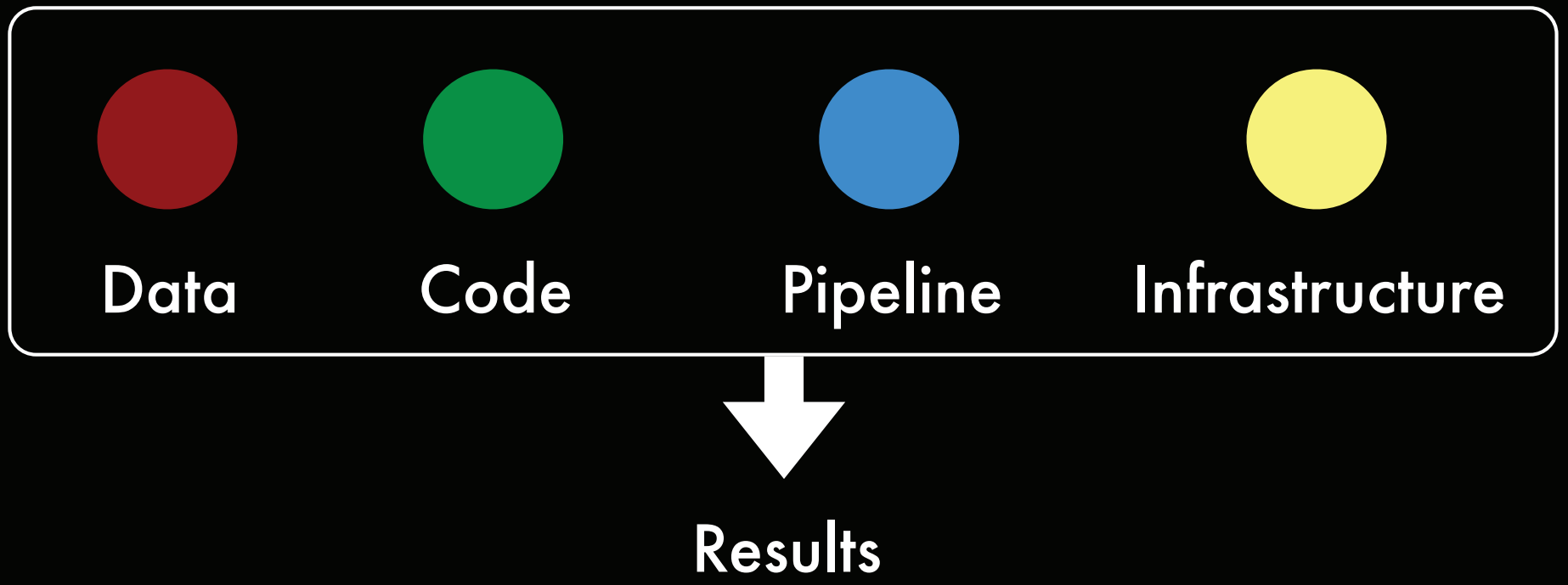


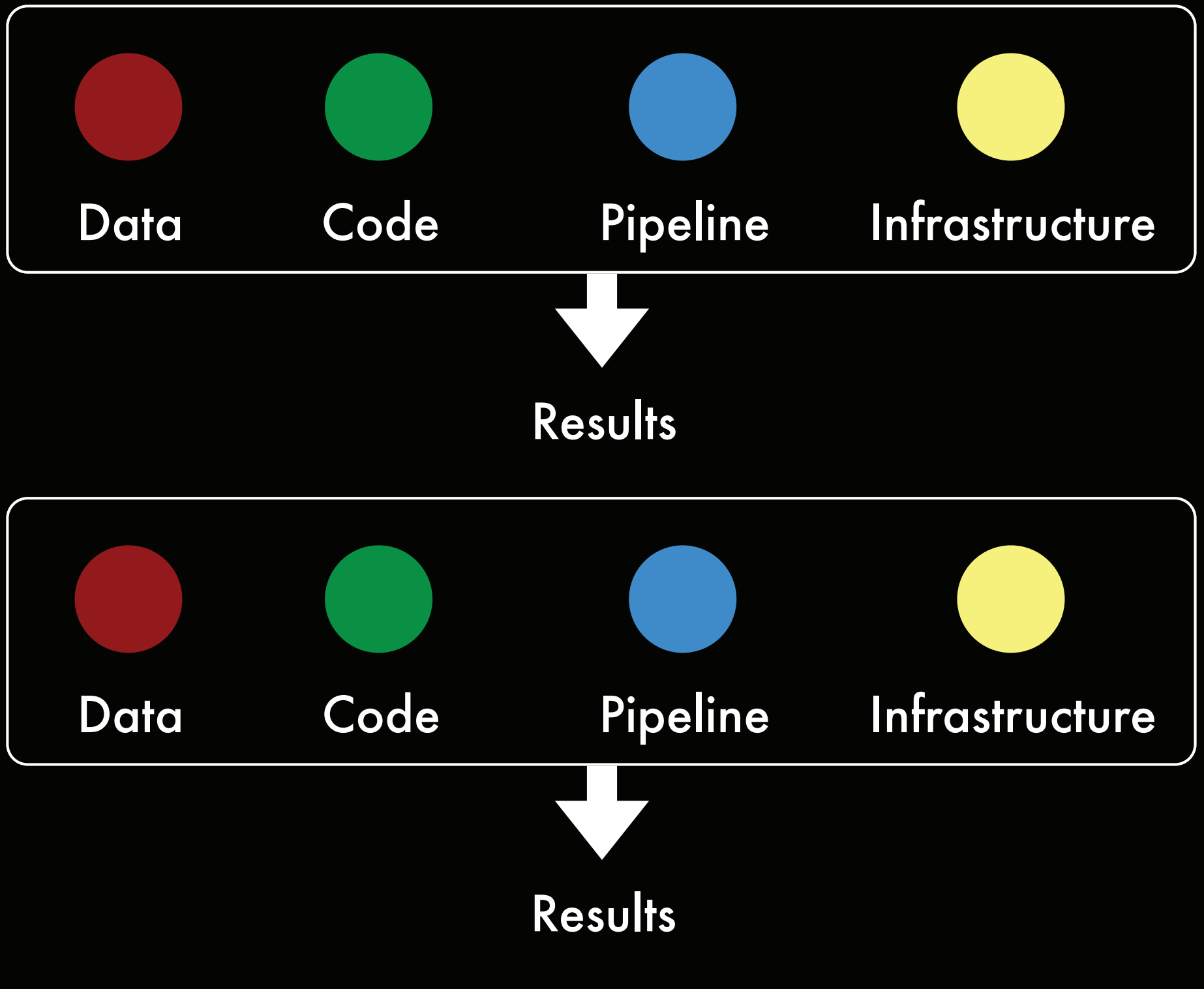
Infrastructure

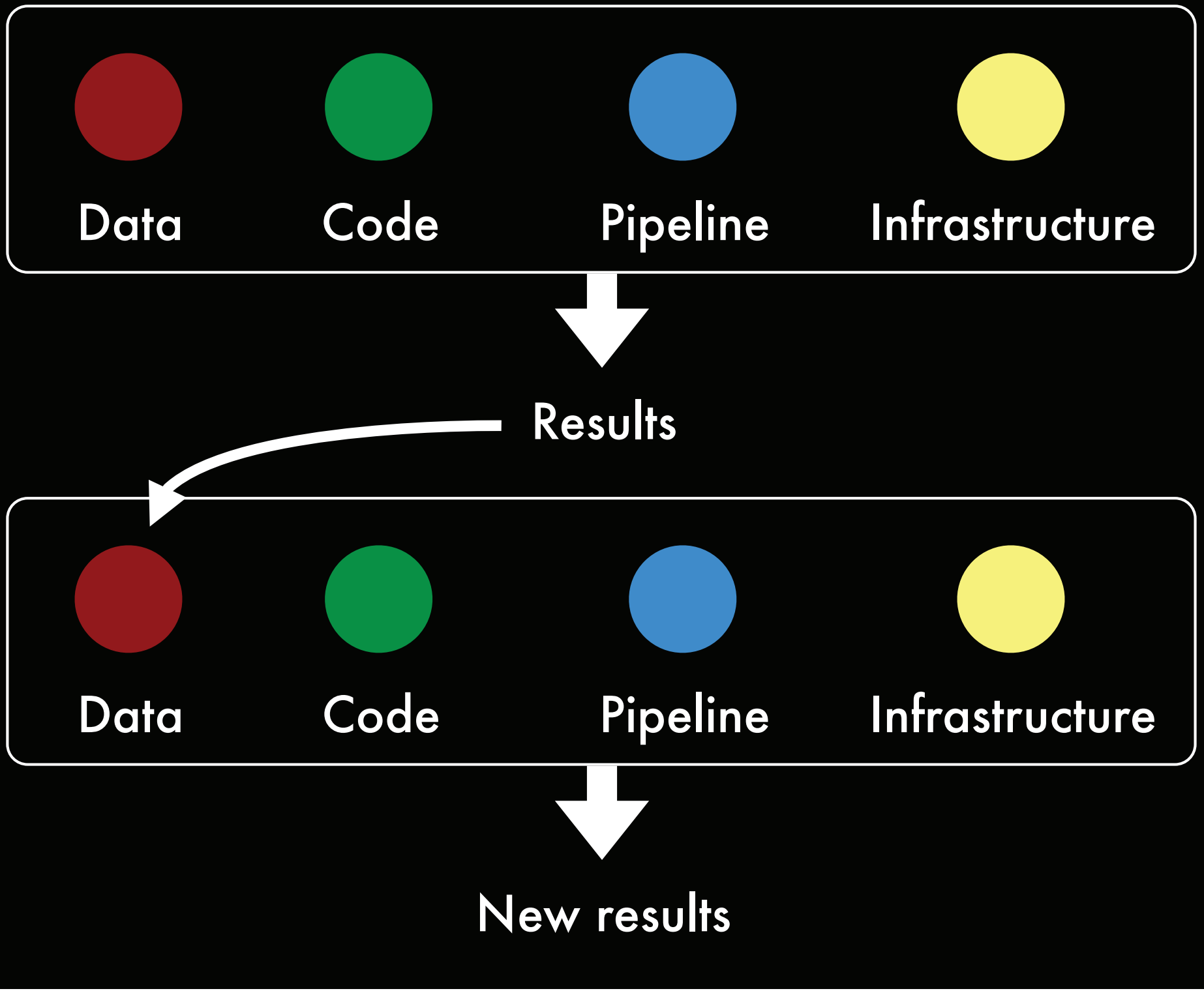

Data
Code
Pipeline
Infrastructure

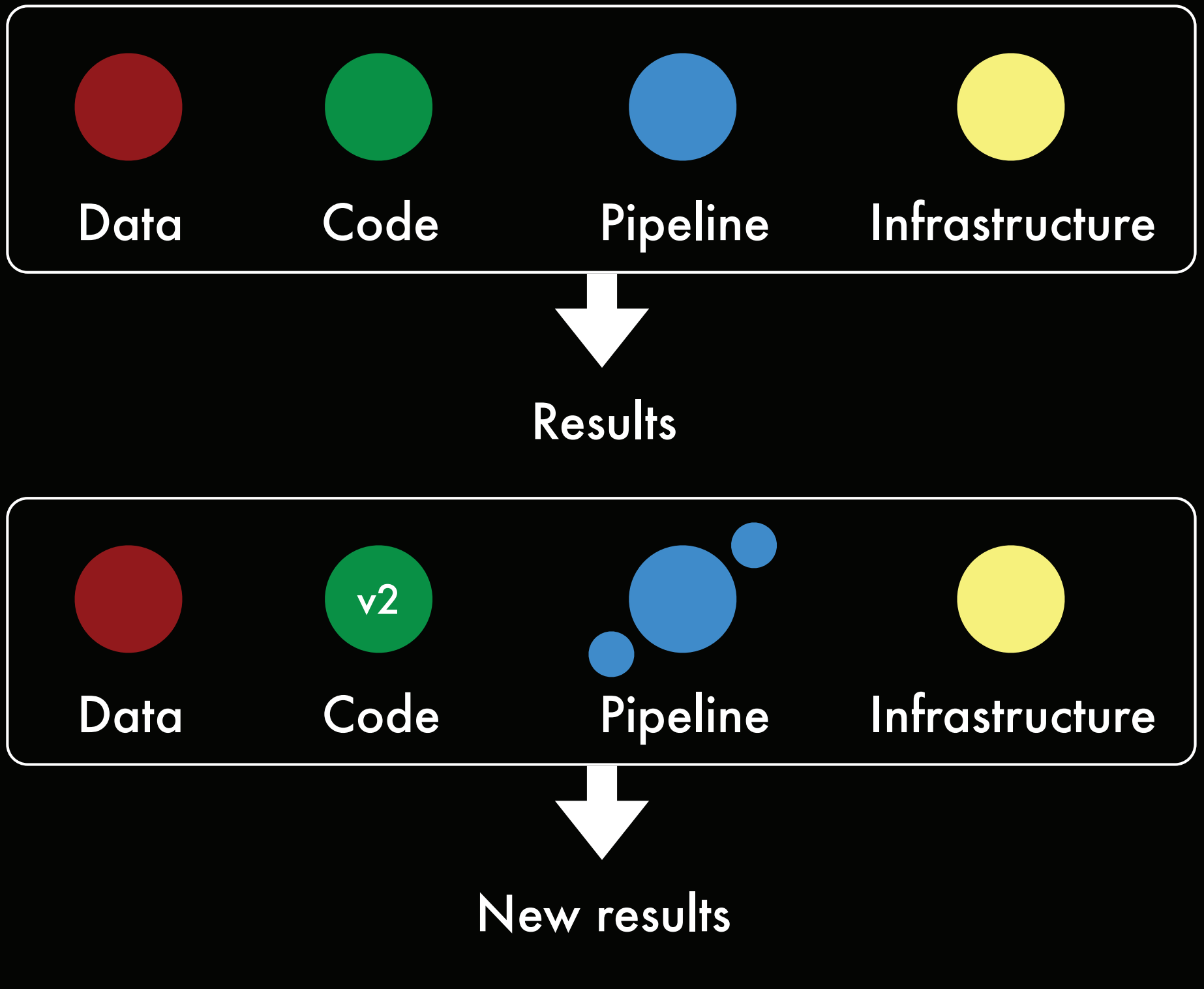
Fully defined

Data sources. Infrastructure stack. Metadata.


Data
Code
Pipeline
Infrastructure
Results


Data
Code
Pipeline
Infrastructure
Results
Data
Code
Pipeline
Infrastructure
Results


Data
Code
Pipeline
Infrastructure
Results
Data
Code
Pipeline
Infrastructure
New results


Data
Code
Pipeline
Infrastructure
Results
v2
Data
Code
Pipeline
Infrastructure
New results

Reproduce.
Remix. Reuse.

Enabled by programmable infrastructure

Enabling science

aws.amazon.com
/genomics

http://aws.amazon.com/big-data-and-hpc-event/boston/

Airbnb, CapitalIQ, Marketshare,
Bioproximity, Schrodinger and MIT

Thank you!

Q & A

matthew@amazon.com

Introducing the panel...

Angel Pizarro - U. Pennsylvania

Anushka Brownley - Complete Genomics

Stephen Litster - Novartis