Hewlett Packard
Enterprise

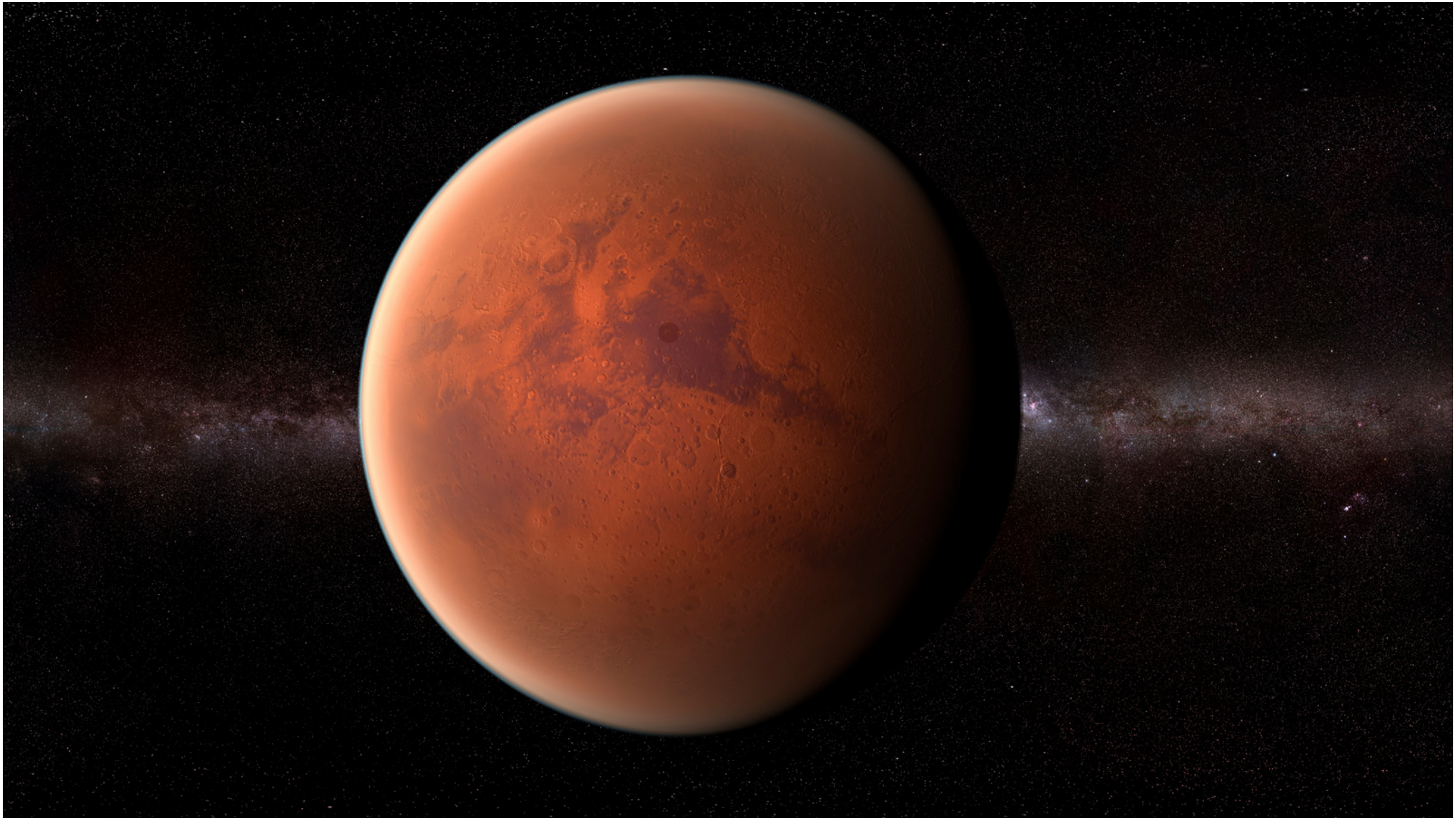
LEARNING FROM EXTREMES

Ted Dunning

September, 2020

Ted Dunning
CTO for Data Fabric, HPE
@ted_dunning
ted.dunning@hpe.com

Stick around for Q&A after this session



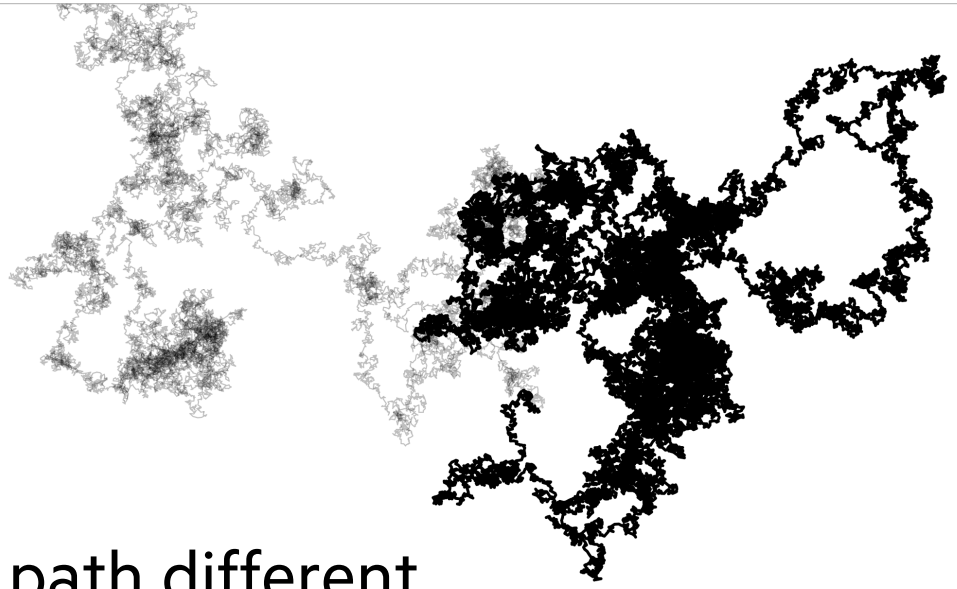
Why do these things fascinate us?

Let’s get a bit geeky

BROWNIAN MOTION

Random motion like small particles in a fluid

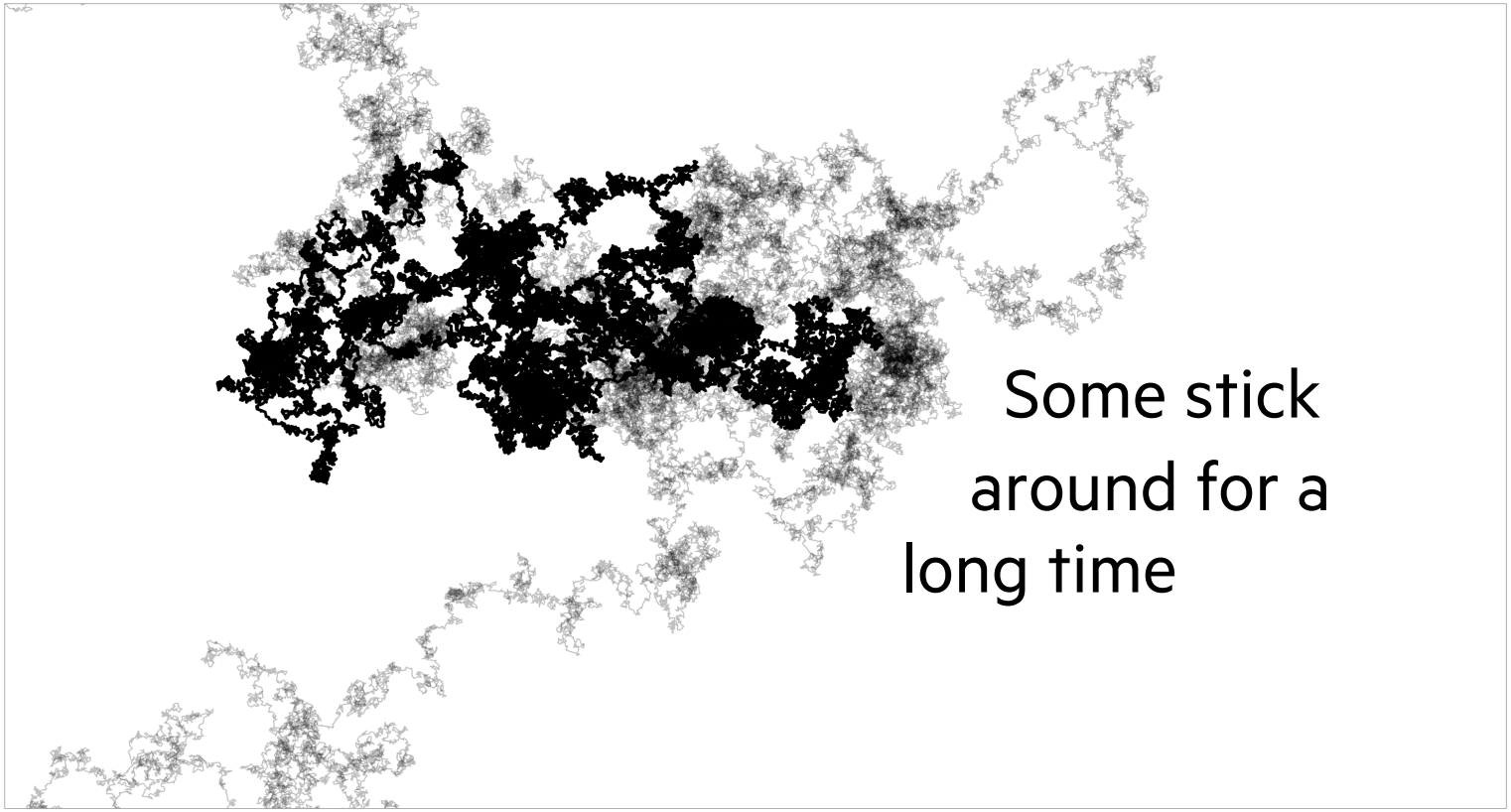
BROWNIAN MOTION



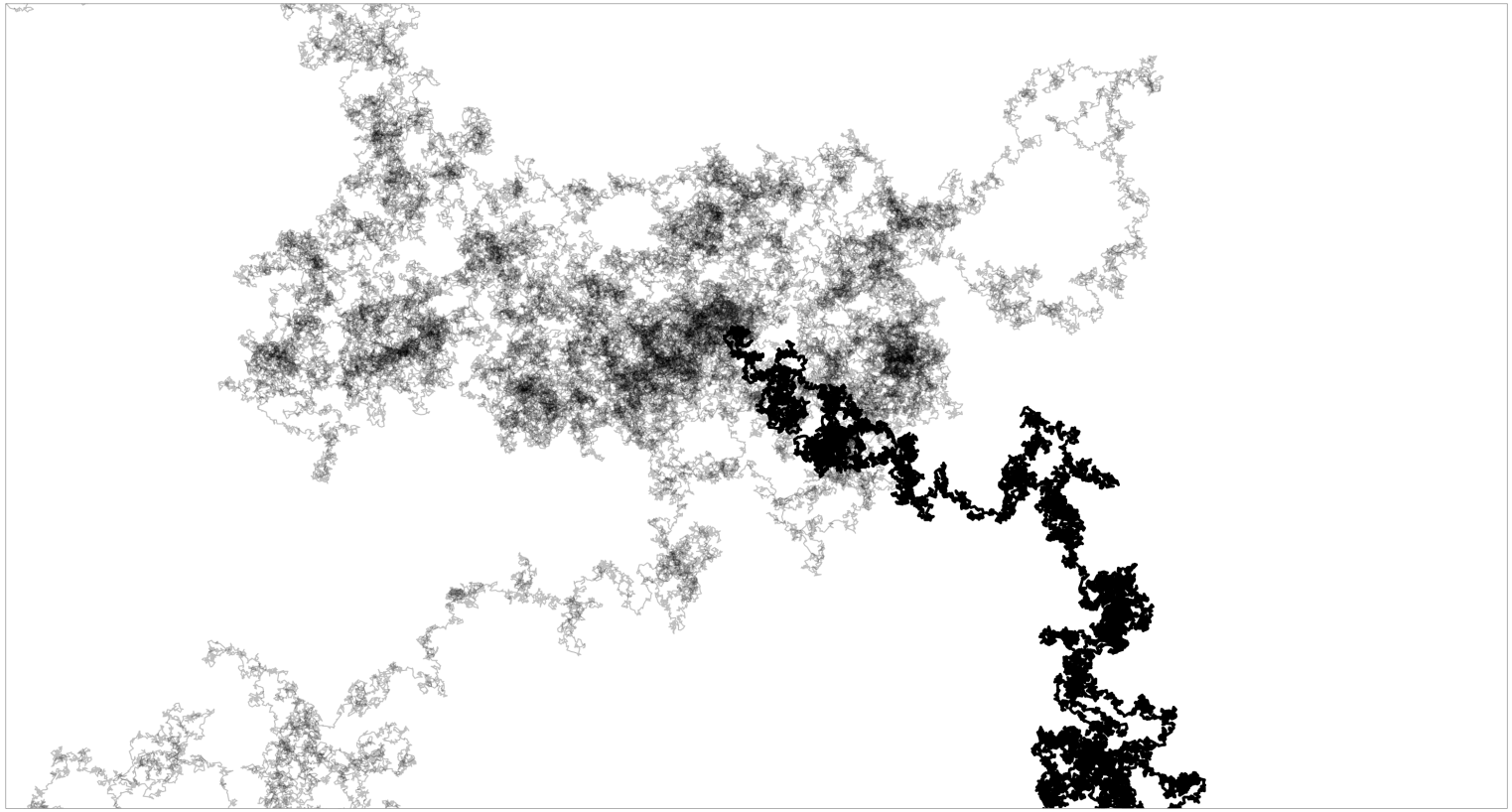
Each path different,
but qualitatively similar

BROWNIAN MOTION

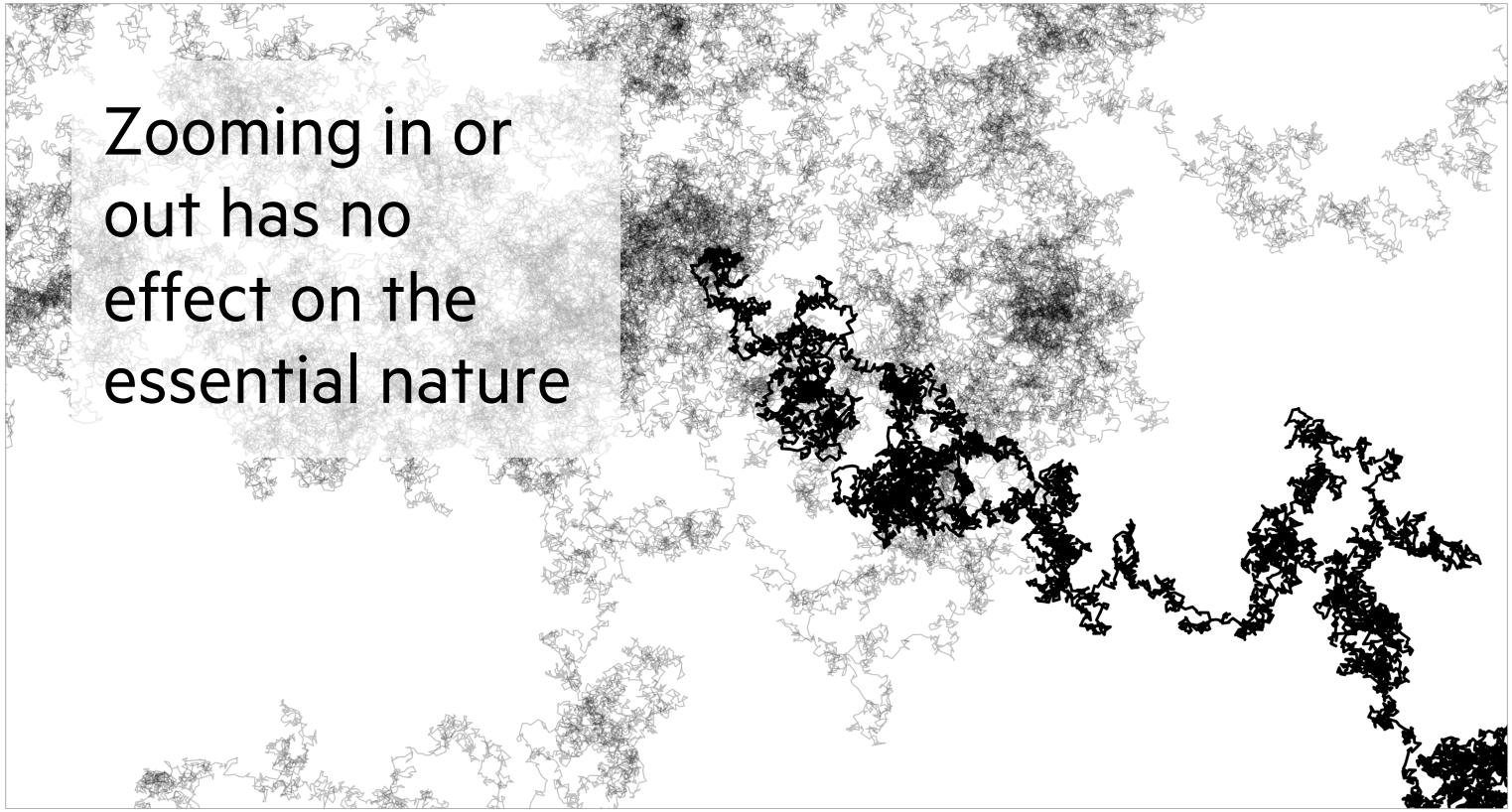
BROWNIAN MOTION



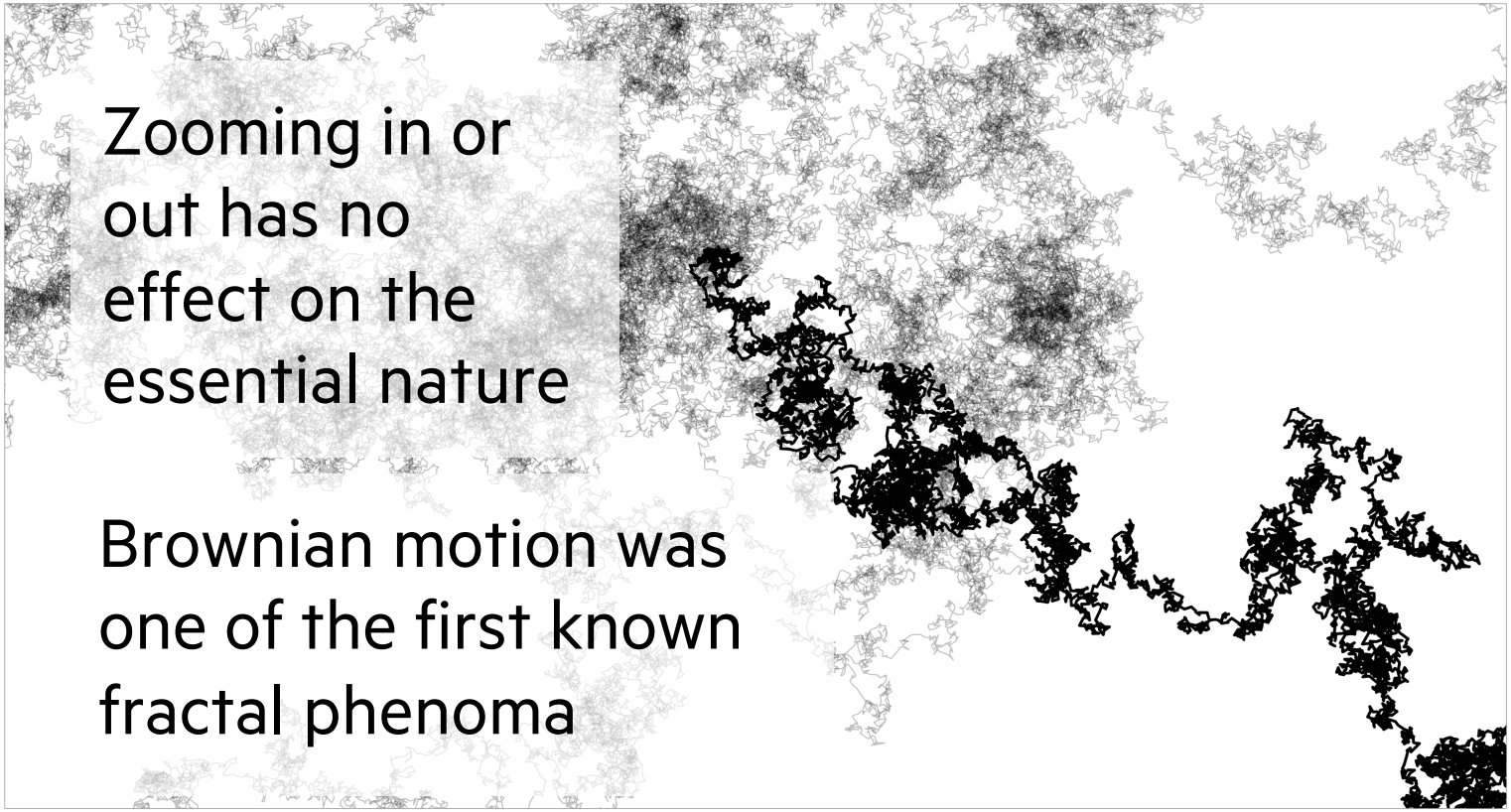
BROWNIAN MOTION



BROWNIAN MOTION – 2X MAGNIFICATION



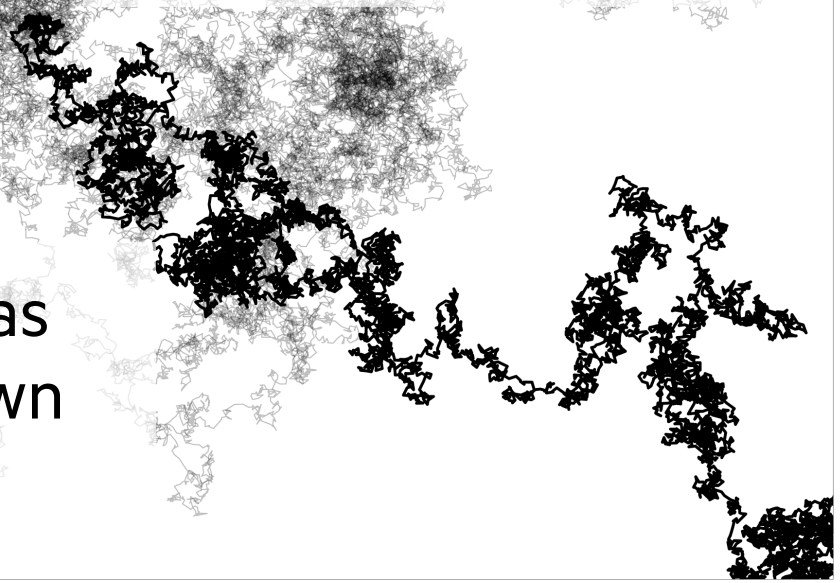
BROWNIAN MOTION – 2X MAGNIFICATION



BROWNIAN MOTION – 2X MAGNIFICATION

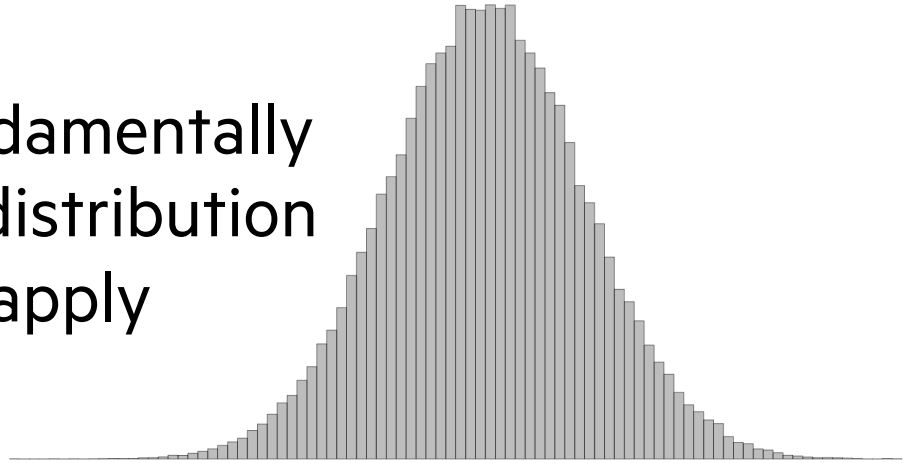
Zooming in or out has no effect on the essential nature

It underpins much of physics and economics



Brownian motion was one of the first known fractal phenoma

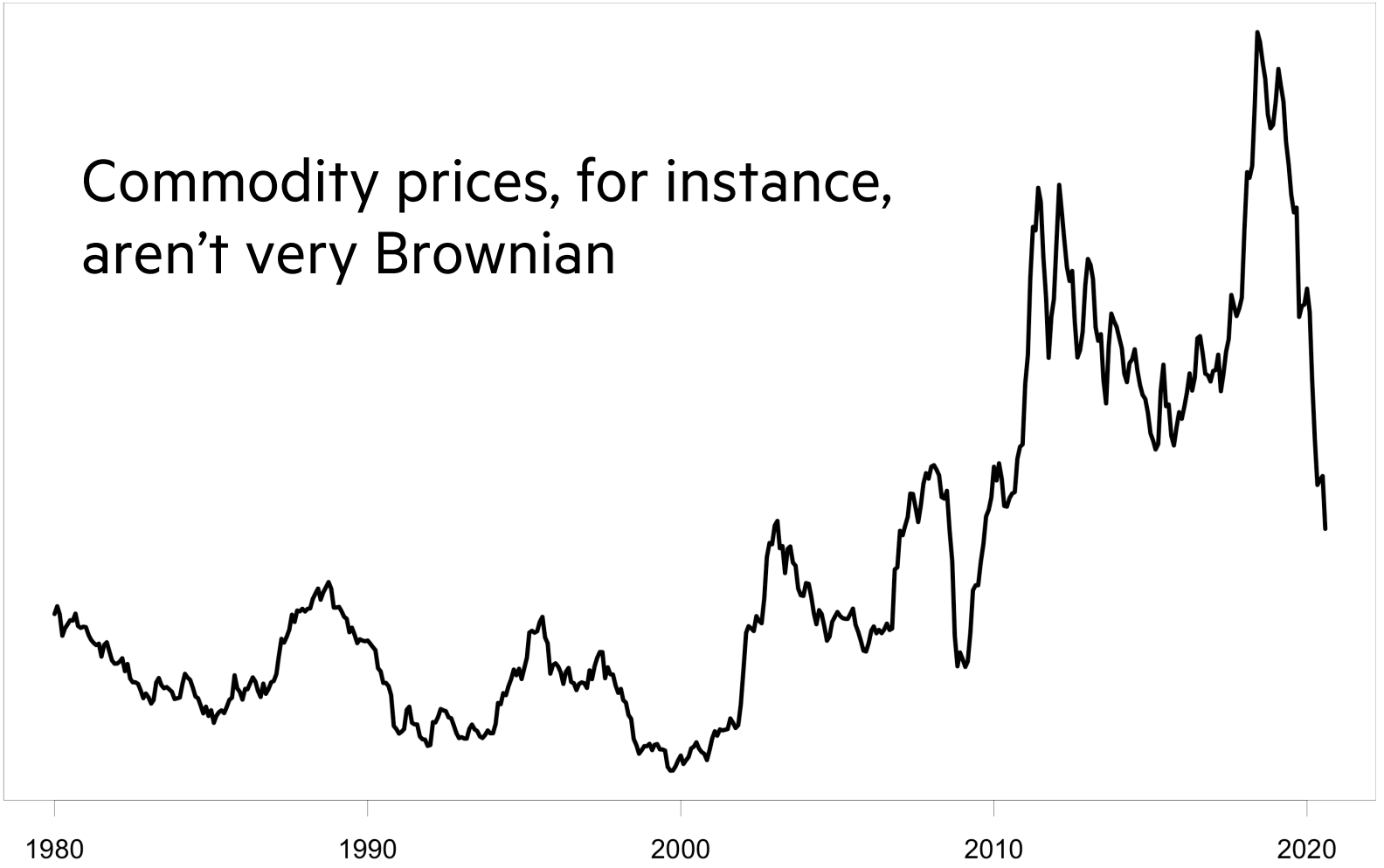
Brownian motion is based fundamentally on assuming that the normal distribution and the law of large numbers apply



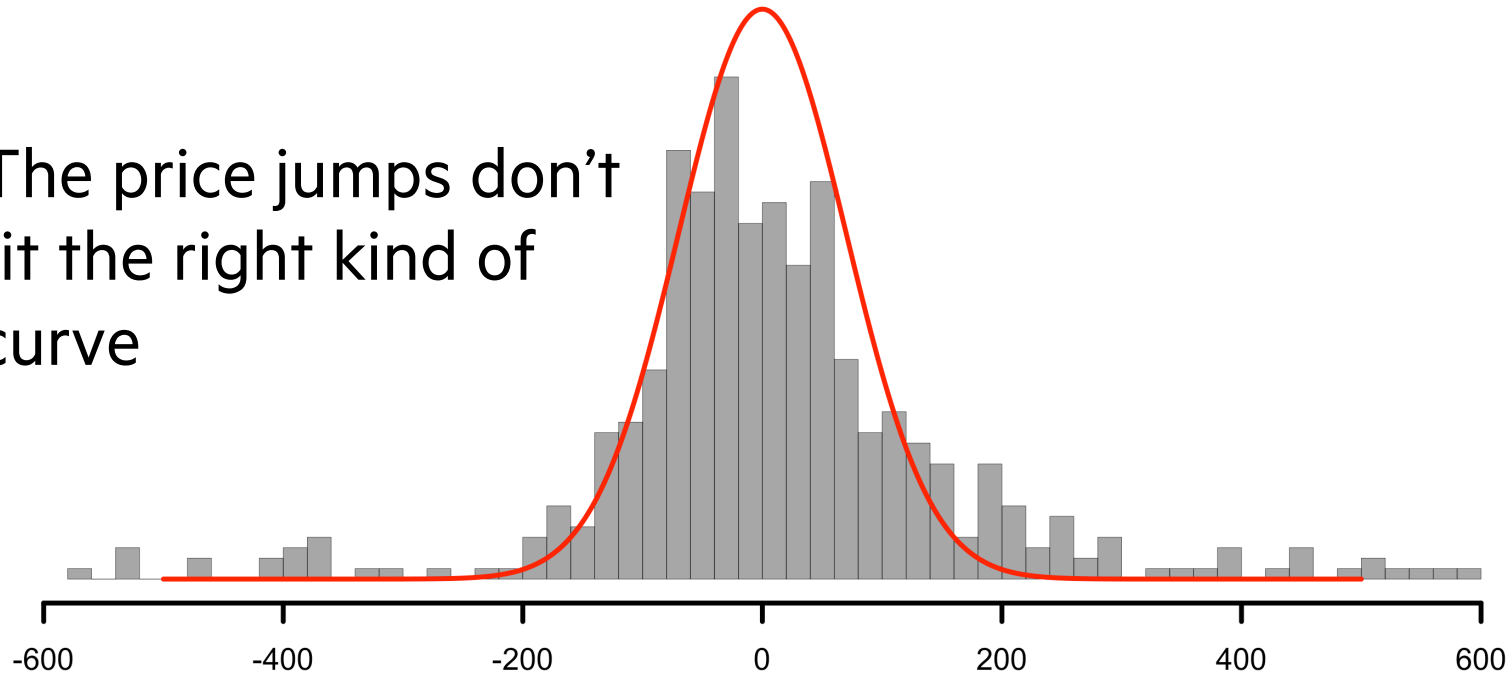
In fact, this assumption appears to be incorrect for many phenomena involving humans or turbulence (or both)

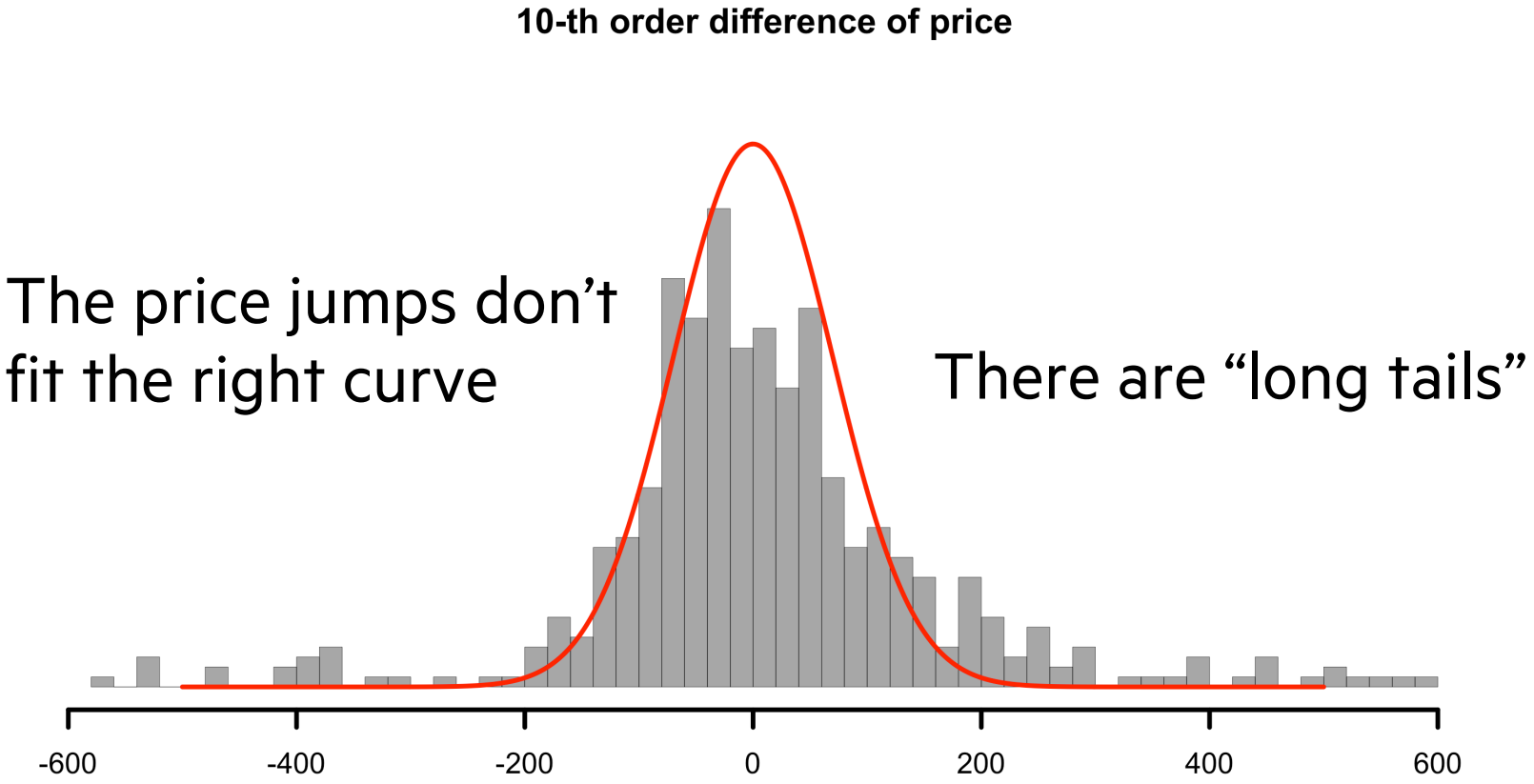
PWOOLC

Commodity prices, for instance, aren’t very Brownian



The price jumps don’t fit the right kind of curve




10-th order difference of price
The price jumps don’t fit the right curve
There are “long tails”
-600
-400
-200
0
200
400
600

But let’s get back to mountaineering and space travel

By nature, these long-tail events are quite extreme and rare

By nature, these long-tail events are quite extreme and rare

Learning about them first-hand is almost impossible to do (and can be fatal)

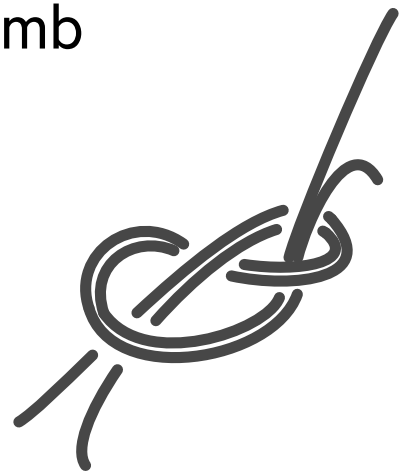
By nature, these long-tail events are quite extreme and rare

Learning about them first-hand is almost impossible to do (and can be fatal)

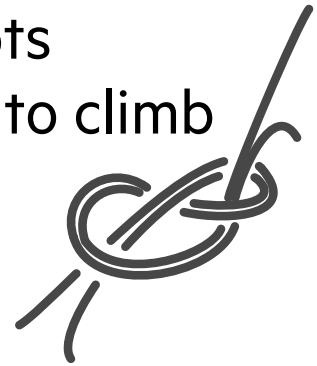
Learning about them from others is a *survival* advantage

That only works if the lessons are general and don’t have big adoption costs

Check your knots
before starting to climb



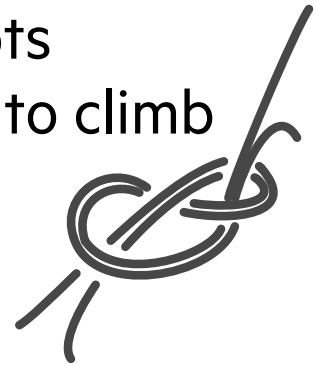
Check your knots
before starting to climb



Stop and think
before hitting return

```
# rm -rf foo /*
```

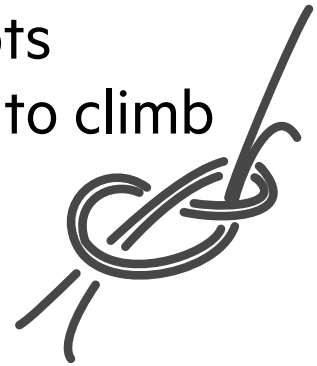
Check your knots
before starting to climb



Stop and think
before hitting return

```
# rm -rf foo /*
```

Check your knots
before starting to climb



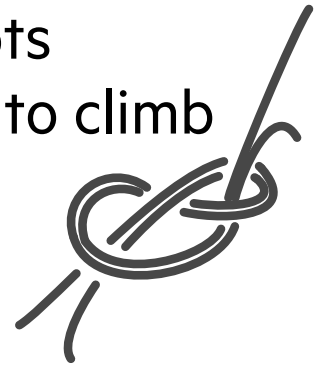
Stop and think
before hitting return

```
# rm -rf foo /*
```

If screwup == dying,
checklists are a good idea

APOLLO 11	
LM LUNAR SURFACE CHECKLIST	
PART NO	S/N
SKB32100074-363	1002

Check your knots
before starting to climb



If screwup == dying,
checklists are a good idea

APOLLO 11	
LM LUNAR SURFACE CHECKLIST	
PART NO	S/N
SKB32100074-363	1002

Stop and think
before hitting return

```
# rm -rf foo /*
```

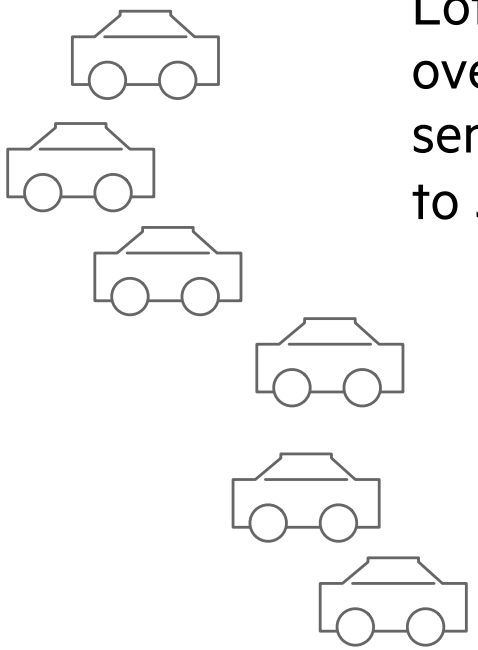
Actually, checklists are
just a good idea generally

1. Notify stakeholders of the impending upgrade, stop accepting new jobs and applications.
2. Disconnect NFS mounts.
3. Stop ecosystem services on the nodes.
4. Locate CLDB and ZooKeeper services.
5. Stop Warden on CLDB nodes first, then others.
6. Stop ZooKeeper on all nodes where installed.

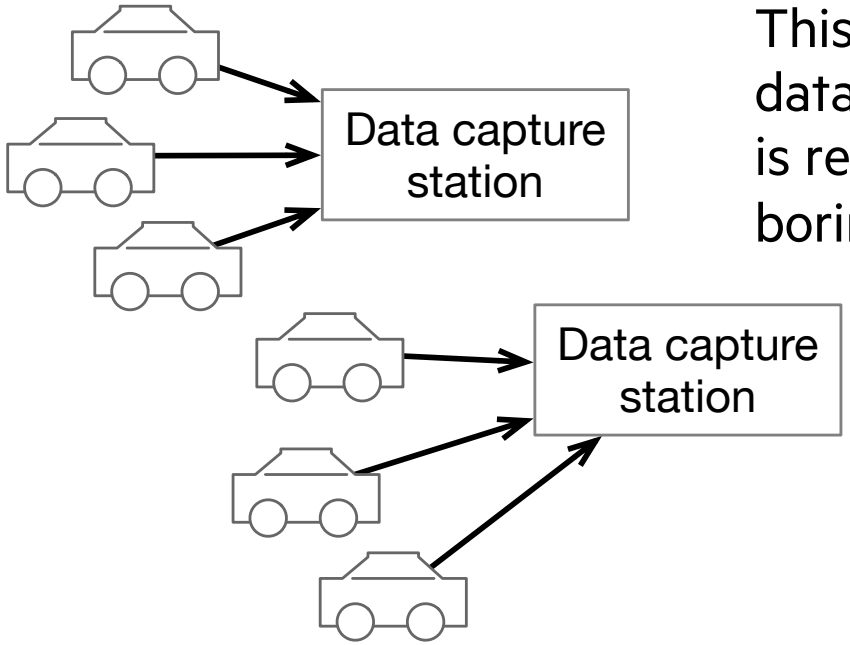
This same process lets us learn within the field as well

LET'S LOOK AT AUTONOMOUS CARS

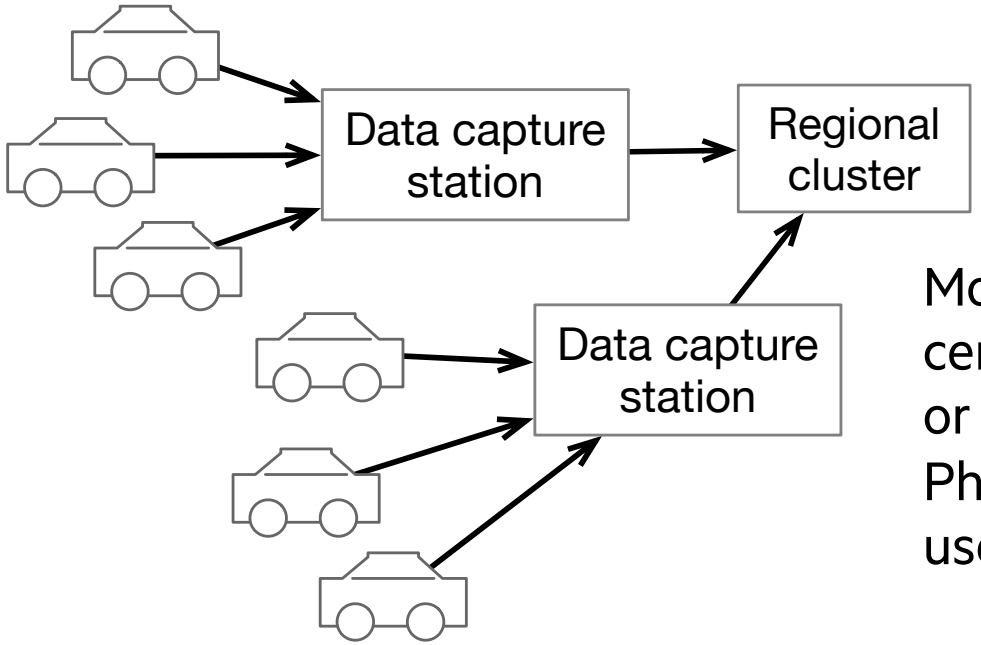
IT equivalent of the Eiger Nordwand



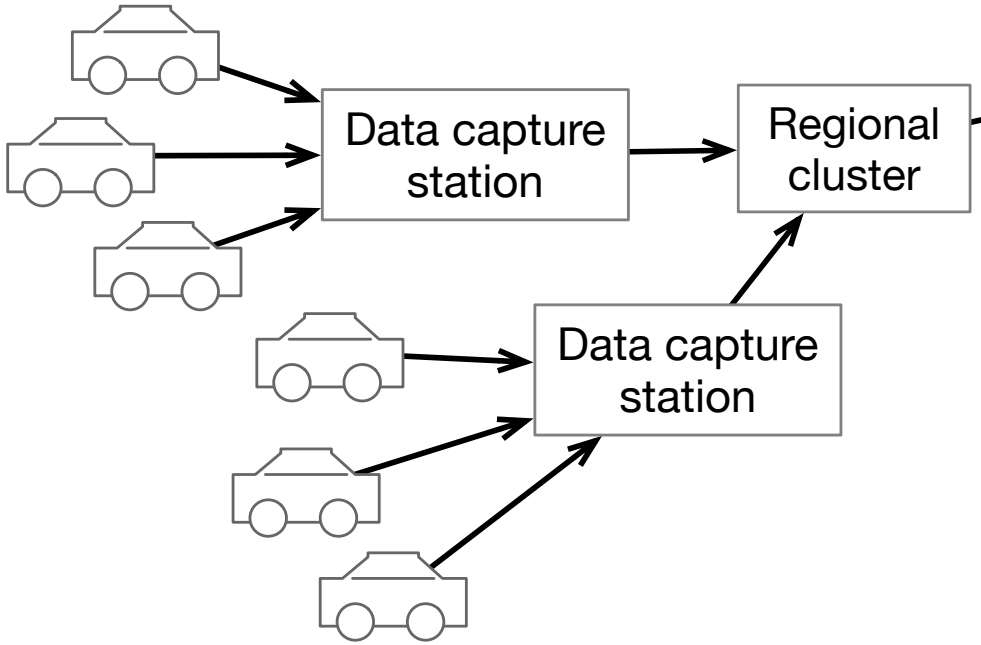
Lots of cars drive around all over the world with all kinds of sensors on them gathering up to 5GB/s each



This data is transferred to local data capture stations where it is reduced (largely by deleting boring bits).

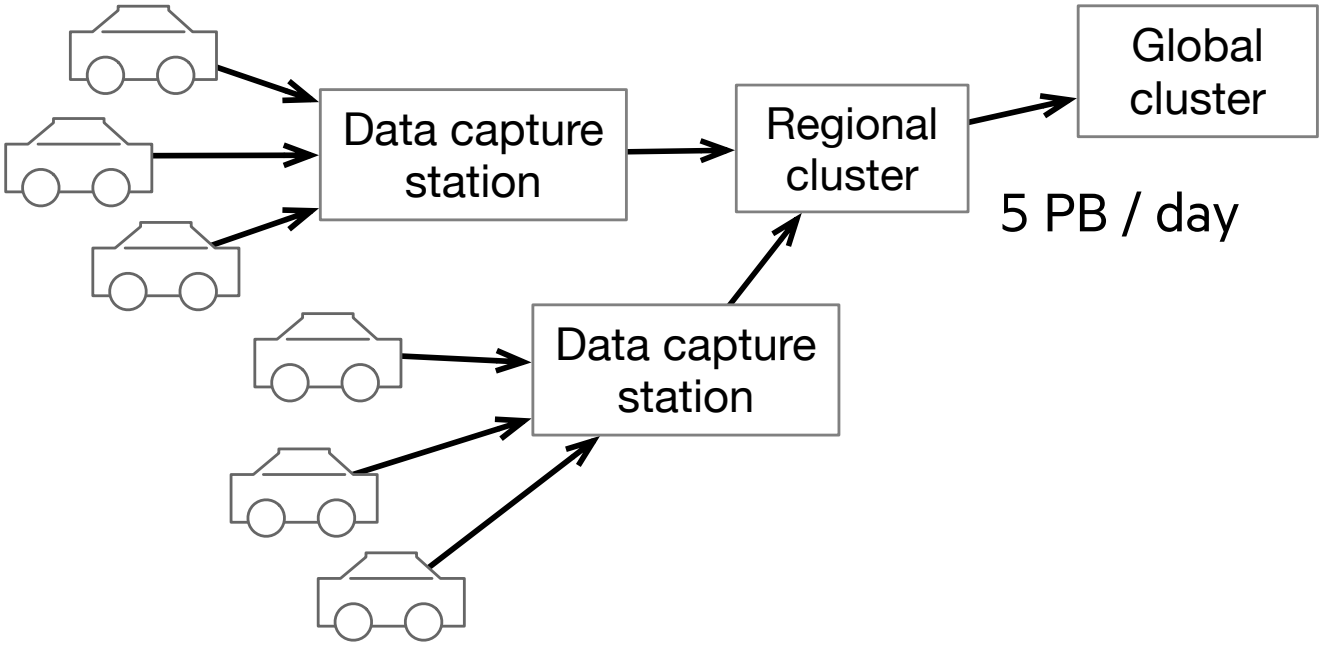


Movement to regional data centers is either by sneakernet or by networked transfers. Physical transfer is surprisingly useful here.



Moving data back to the galactic headquarters is almost always done using network transfers. The geo-stability of the regional cluster makes this easier than physical transfers.

>400 PB total


Data capture station
Regional cluster
Global cluster
5 PB / day
Data capture station

KEY TASKS

- Need to transfer data to central site
  - This is staged with lots of edge compute

- Feature extraction workload dominates raw transfer in central cluster
  - Must scan >100PB in less than a day
  - Some worker nodes can process in excess of 1GB/s
  - Key design constraint here favors density over speed (surprisingly)

- Main learning workload is heavily compute bound with “small” training data

- Many process meta-data workloads as well
  - How much data? What quality? How many modeling runs? How full is the infrastructure?

HOW MUCH DATA COULD A WOOD CHUCK CHUCK?

- Data is stored on Apollo 4510
  - 60 x 10TB drives for older machines
  - 60 x 16TB drives for newer machines

- Controller can sustain 10GB/s, (assume 5GB/s)
  - Can read 80% of all data on machine in
    - 11 hours for triplicated data
    - 22 hours for ECC data

  - Cluster read speeds are independent of cluster size

  - HDDs are fast enough to give aggregate I/O >10TB/s

That seems kind of outlandish. 50PB ingested, 5 retained per day?

But this design can scale down 300x and retains operational simplicity.

That seems kind of outlandish. 50PB ingested, 5 retained per day?

I could just script these copies

But this design can scale down 300x and retains operational simplicity.

Even at a few TB, platform level data motion is far more robust than hand-coding.

That seems kind of outlandish. 50PB ingested, 5 retained per day?

I could just script these copies

Why run K8s and Spark at the edge?

But this design can scale down 300x and retains operational simplicity.

Even at a few TB, platform level data motion is far more robust than hand-coding.

Simple anomaly detection can save 90% on transfer size

Even with a large specialist team, the 90% budget share of logistics can easily grow to 99% or 110%.

Downscaling the problem makes this more true

Without platform efficiencies, many AI projects become infeasible.

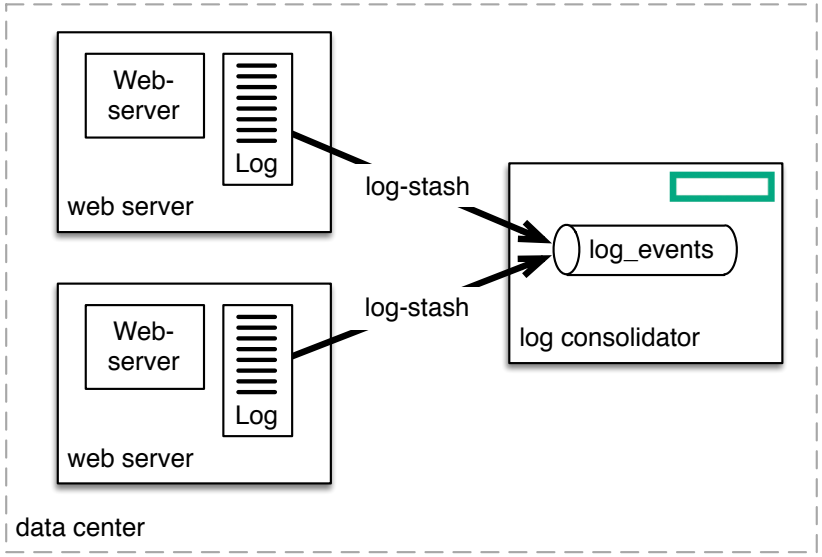
TELEMETRY AND SCALING CARDINALITY

IT equivalent of the Dragon capsule

STREAMING TELEMETRY

- Data fabric level data motion simplifies telemetry tasks in real-time or batchwise
- This allows very simple and scalable scaling both in terms of data size, but also in terms of system diversity and complexity
- Edge analytics not critical here, but replication to a single stream better with very large number of topics

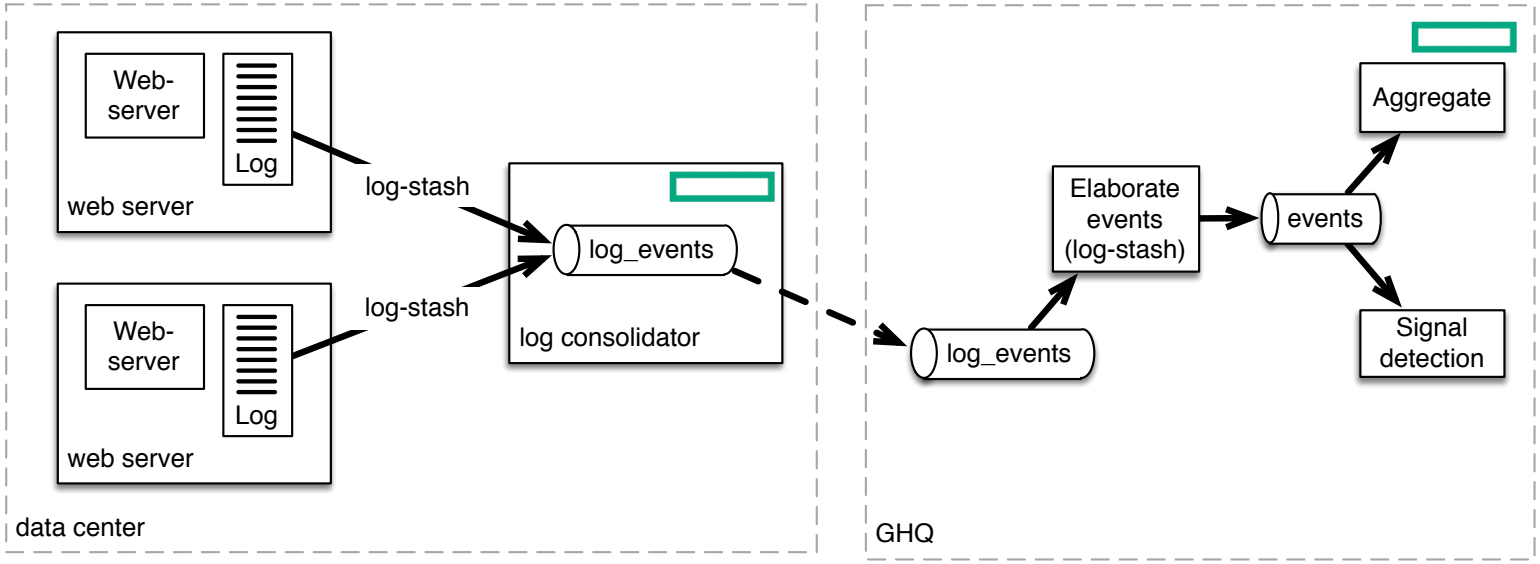
COLLECT DATA



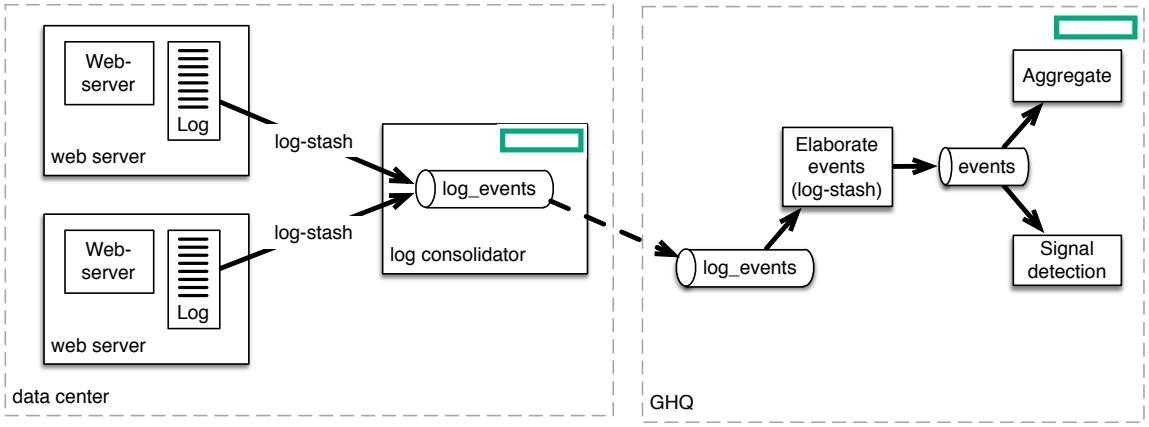
Data acquisition locally is very simple and can be done using simple open-source tools

AND TRANSPORT TO GLOBAL ANALYTICS

Data motion to the core is handled at the fabric level with no explicit action required other than initial configuration

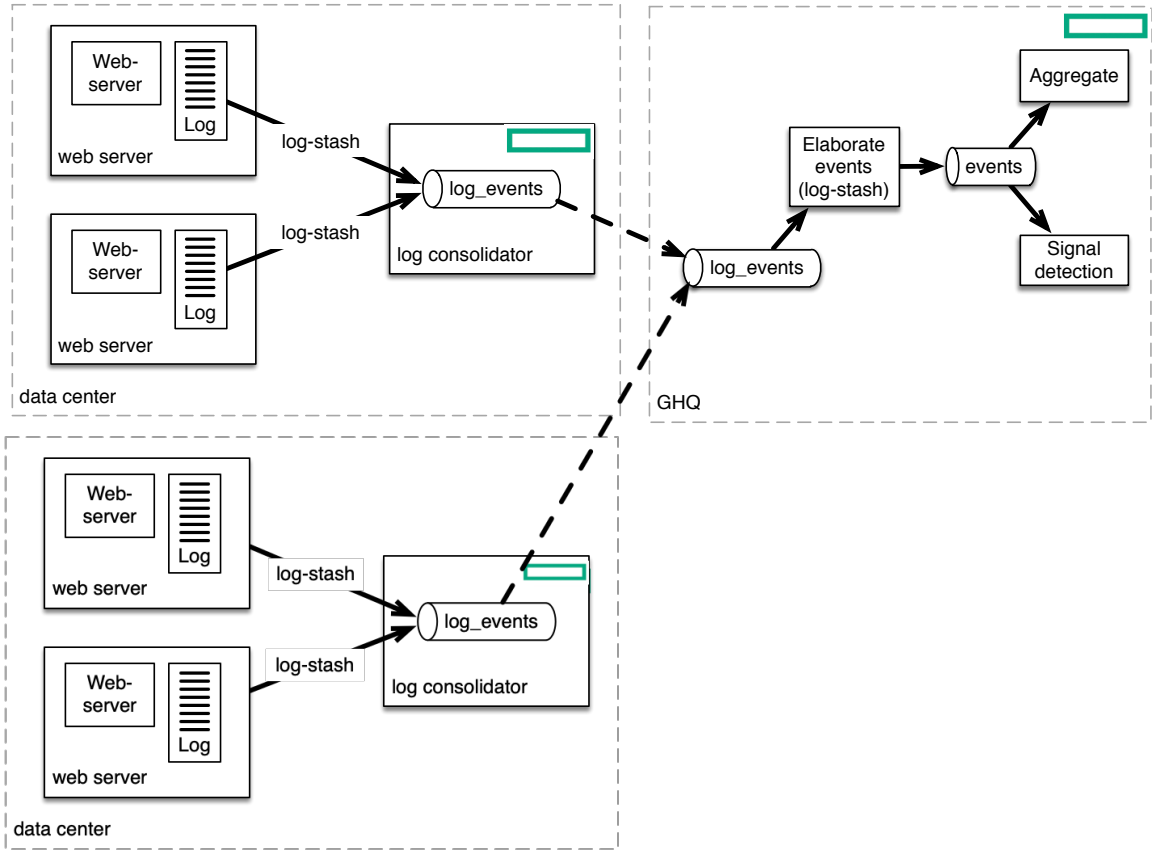


WITH MANY SOURCES

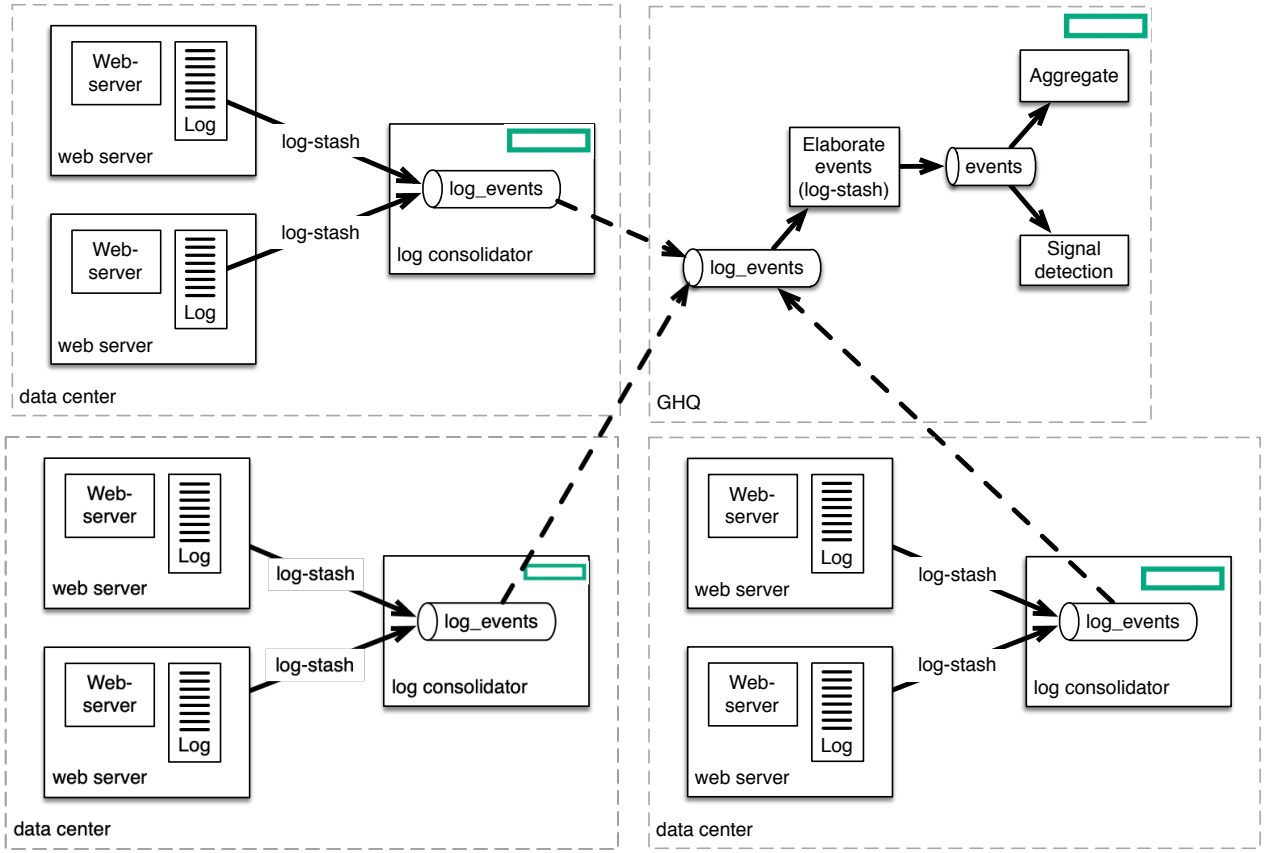


This scales easily in terms of more data sources

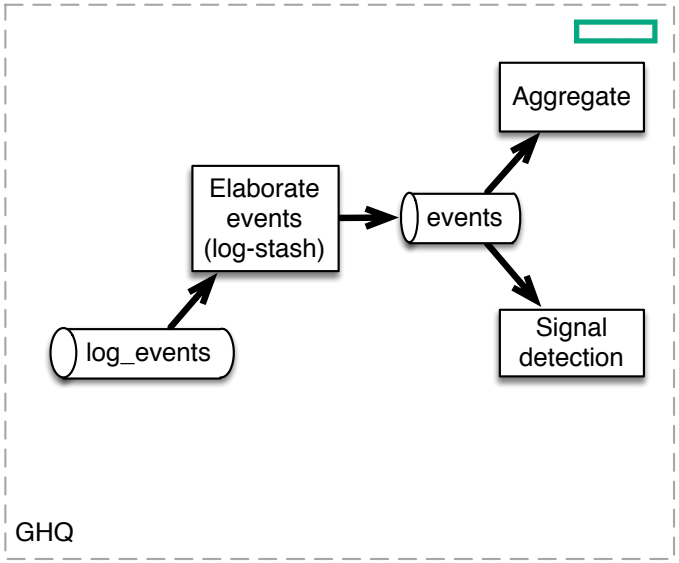
WITH MANY SOURCES



WITH MANY SOURCES



ANALYTICS SHOULDN'T CARE ABOUT LOCATION



The core analytics can be done entirely without reference to any of the source sites (the data includes that information, of course). This allows complete separation of concerns.

Shouldn’t big data be measured in thousands of kilometers?

If you have a single data structure that spans a data fabric, it can be

OTHER KINDS OF CARDINALITY

IT equivalent of the soloing El Cap

A note about this use case. It’s real. A large fraction of the audience uses it every day.

It’s just invisible. Because the customer likes it that way.

What's the simplest way to start storing message attachments?

(naively, a file for each attachments)

What happens when you have 10k attachments or more?

(a directory for each 10k or 100k attachments)

What happens when you have 10M to 100M attachments or more?

(a data fabric volume for every 50M attachments)

What happens when you have 10B to 100B attachments or more?

(nothing … you just have 1000 volumes)

What happens when you have 1T attachments or more?

(nothing … you just have 100,000 volumes)

What happens when you have 10T to 100T attachments or more?

(you add a cluster in the same name space every few trillion files)

What happens when you have 10T to 100T attachments or more?

(you add a cluster in the same name space every few trillion files)

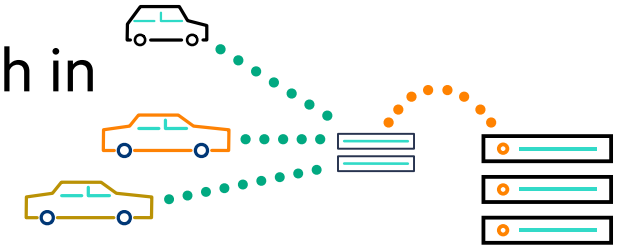
What happens when you have 10^15 files or more

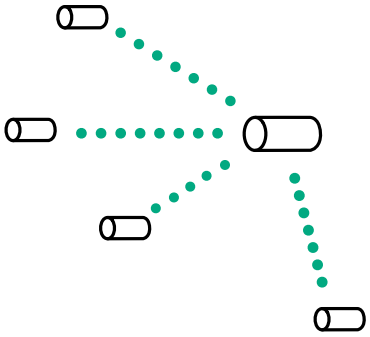
(nobody's asked yet … that a *lot* of files)

The point is that climbing the rock can be roughly the same as walking to the rock.

You mostly just need some better shoes. (ish)

The edge is very much in reach, even at scale





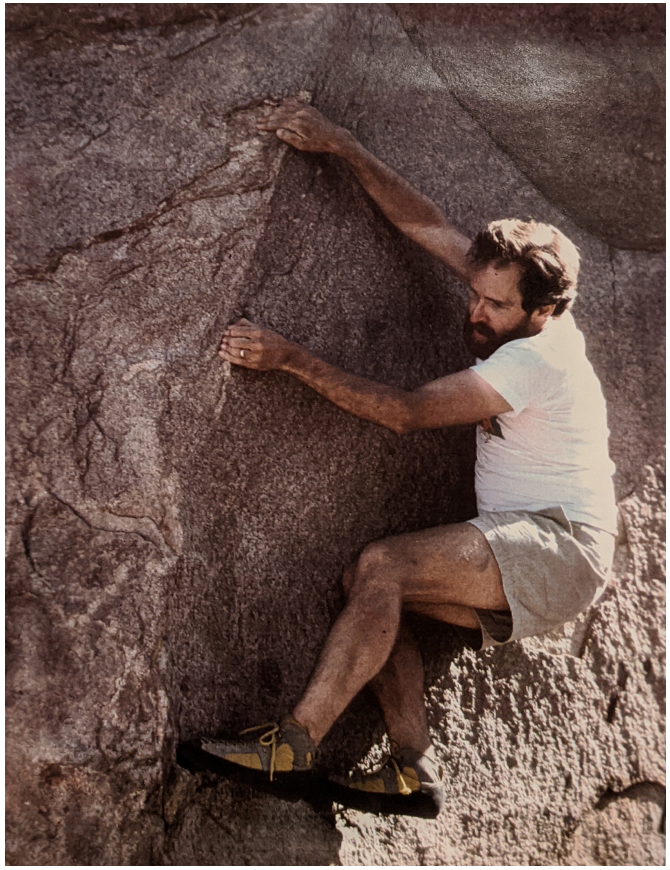
Scale can mean physical diameter or file count or row count

Extreme cases teach us how to get beyond our comfort zone

THANK YOU

Ted Dunning
CTO for Data Fabric, HPE
@ted_dunning
ted.dunning@hpe.com

Stick around for Q&A after this session



NOTES ABOUT CONTENT

- Do not remove this slide from the presentation
- HPE does not warrant or represent that it will introduce any product to which the information relates
- The information contained herein is subject to change without notice
- HPE makes no warranties regarding the accuracy of this information
- The only warranties for HPE products and services are set forth in the express warranty statements accompanying such products and services
- Nothing herein should be construed as constituting an additional warranty
- HPE shall not be liable for technical or editorial errors or omissions contained herein
- Strict adherence to the HPE Standards of Business Conduct regarding this classification level is critical