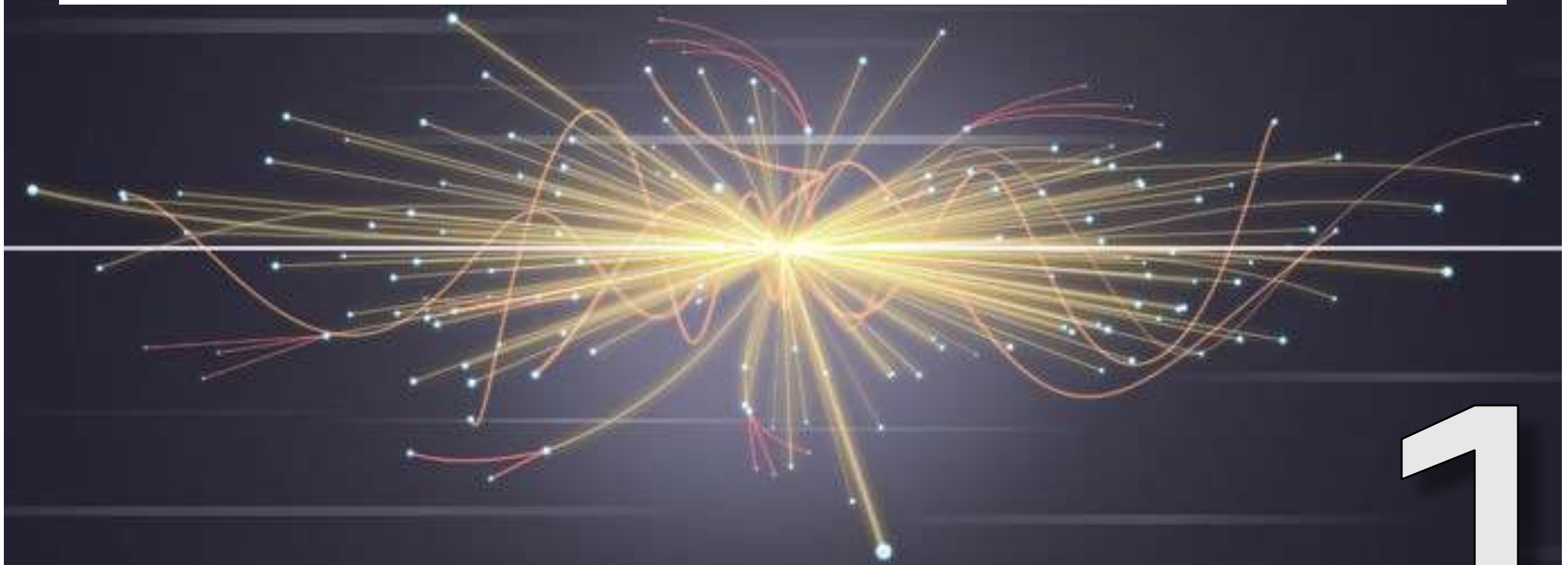
ETH zürich

Ghislain Fourny

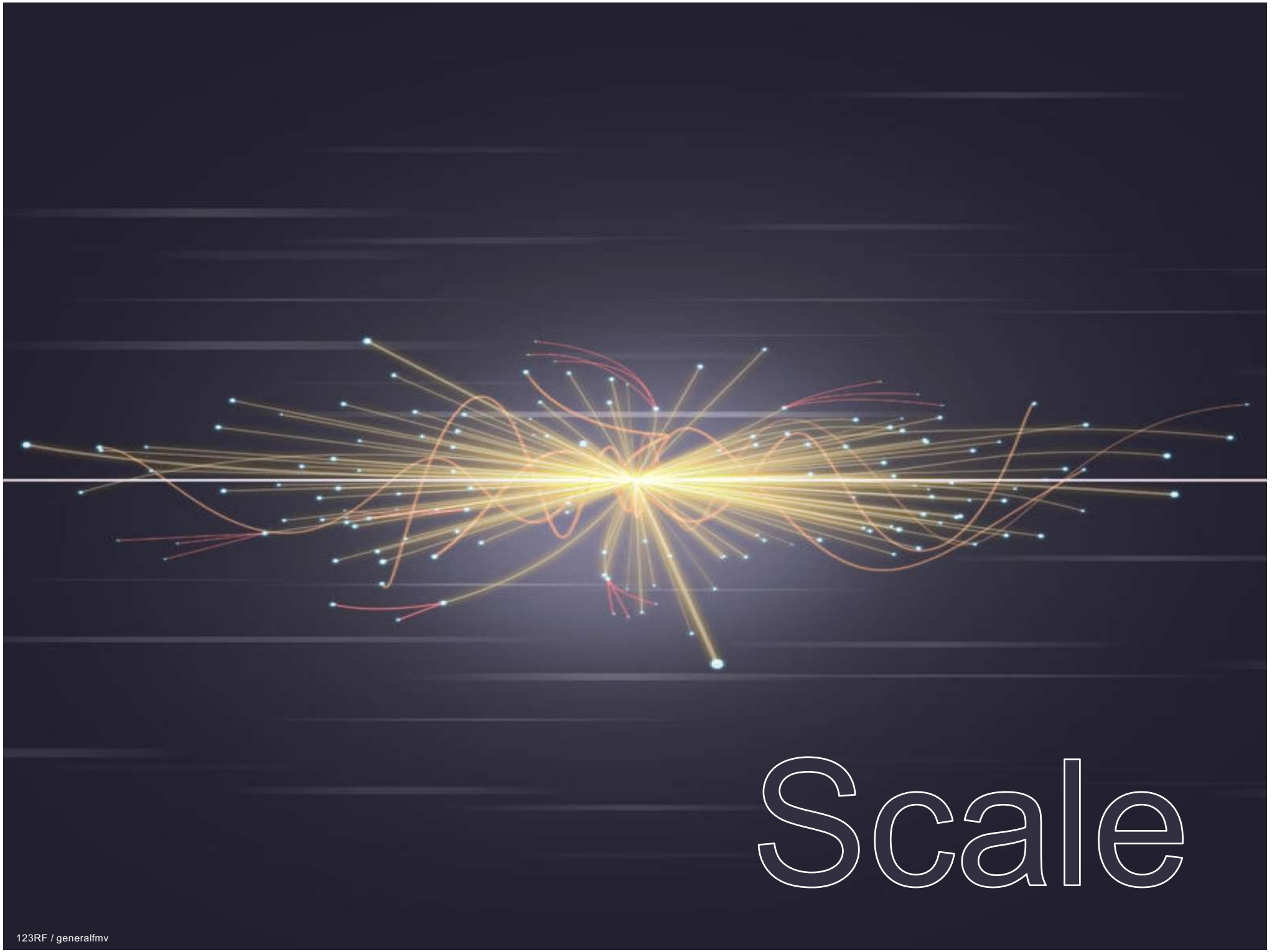
Big Data for Engineers Spring 2020

1. Introduction



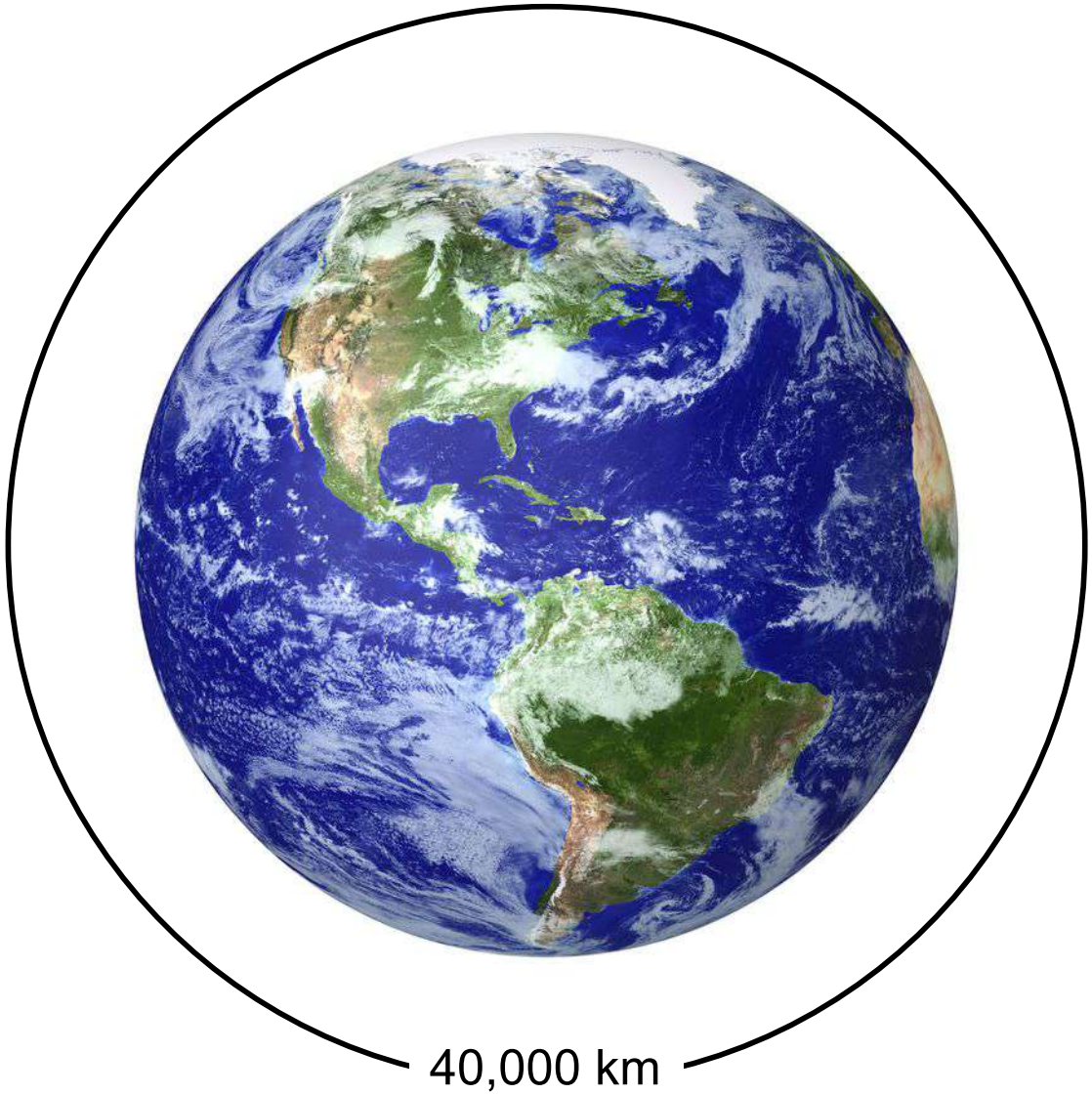
1

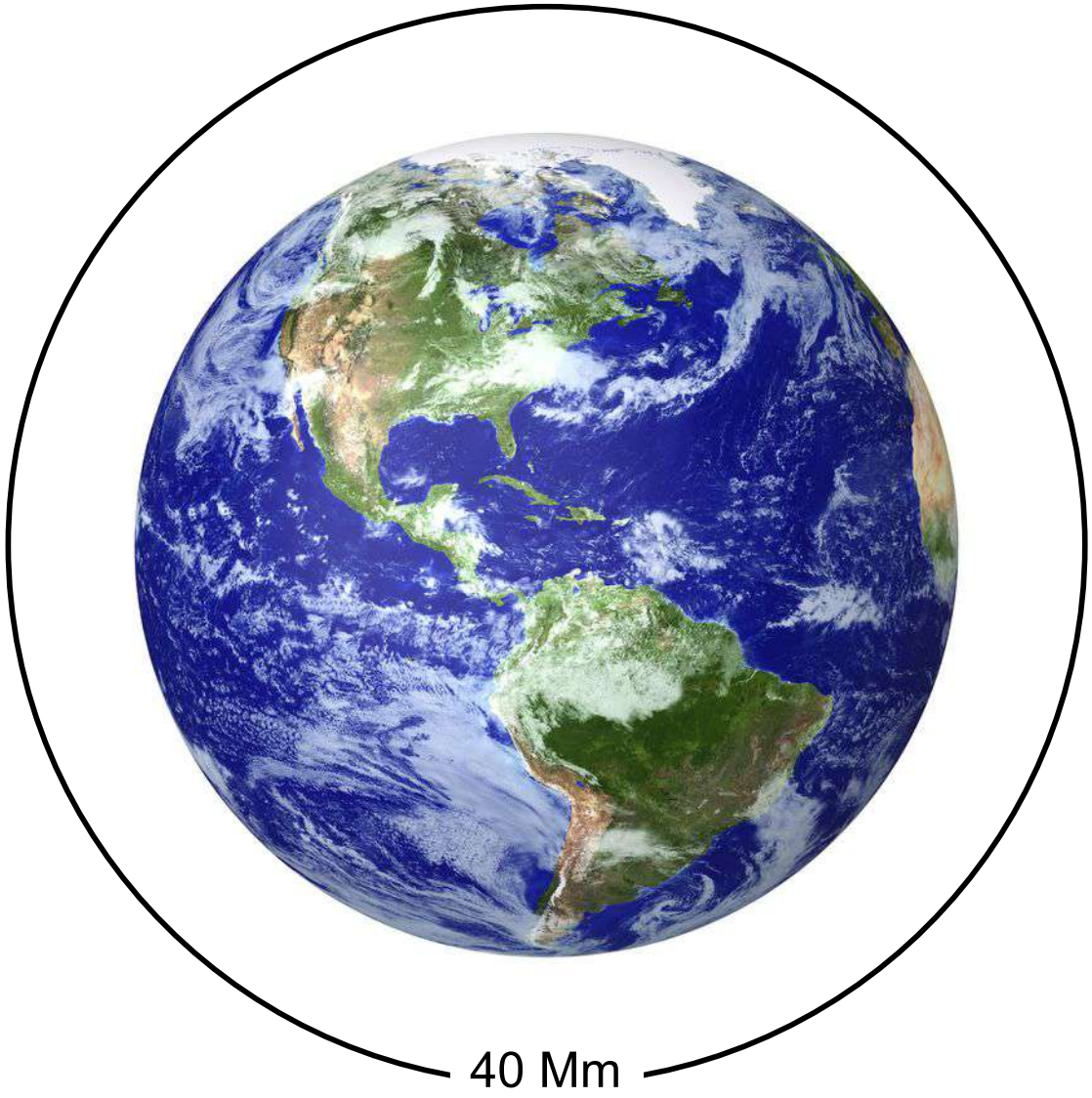
123RF / generalfmv

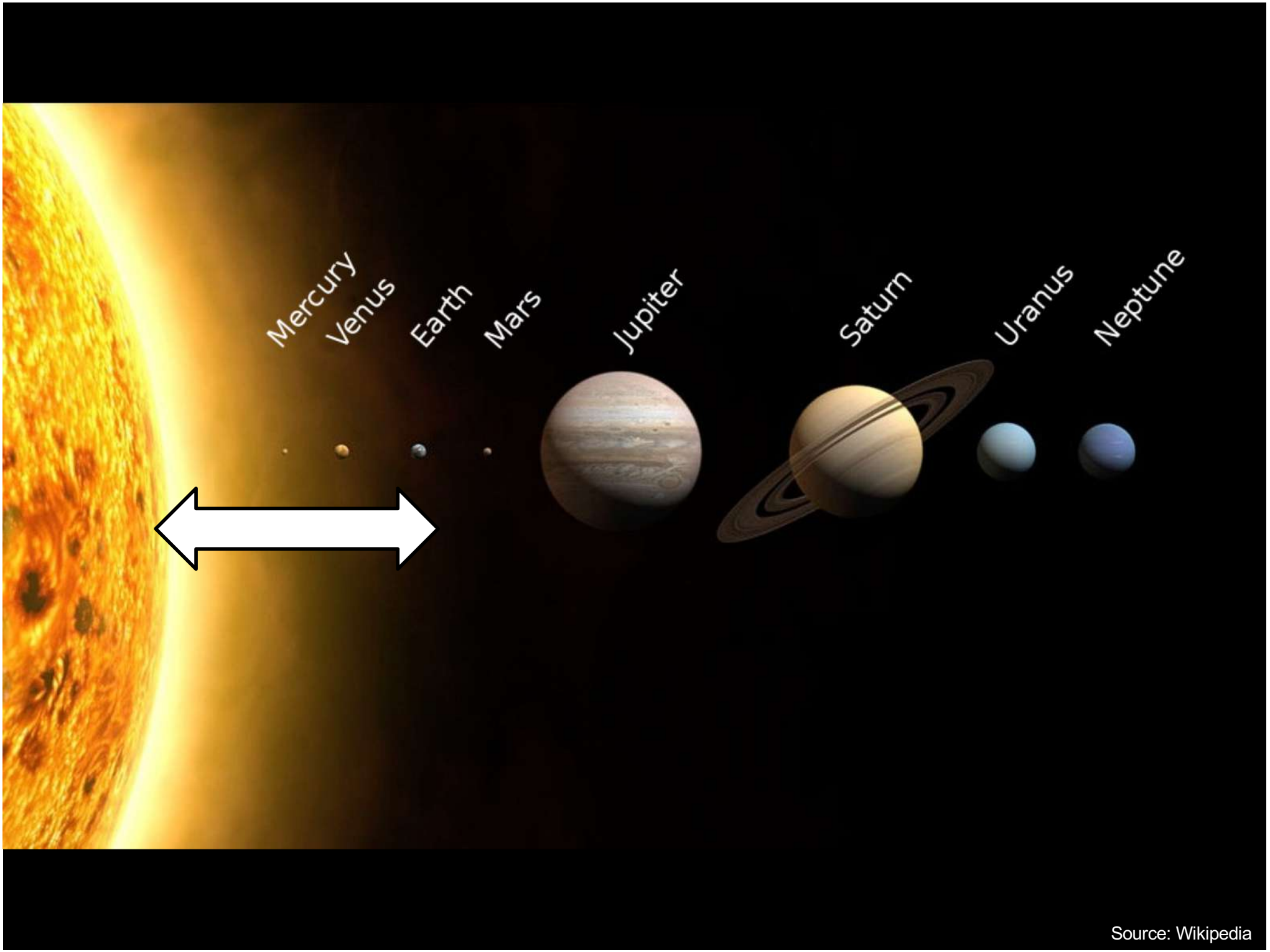

Scale
123RF / generalfmv

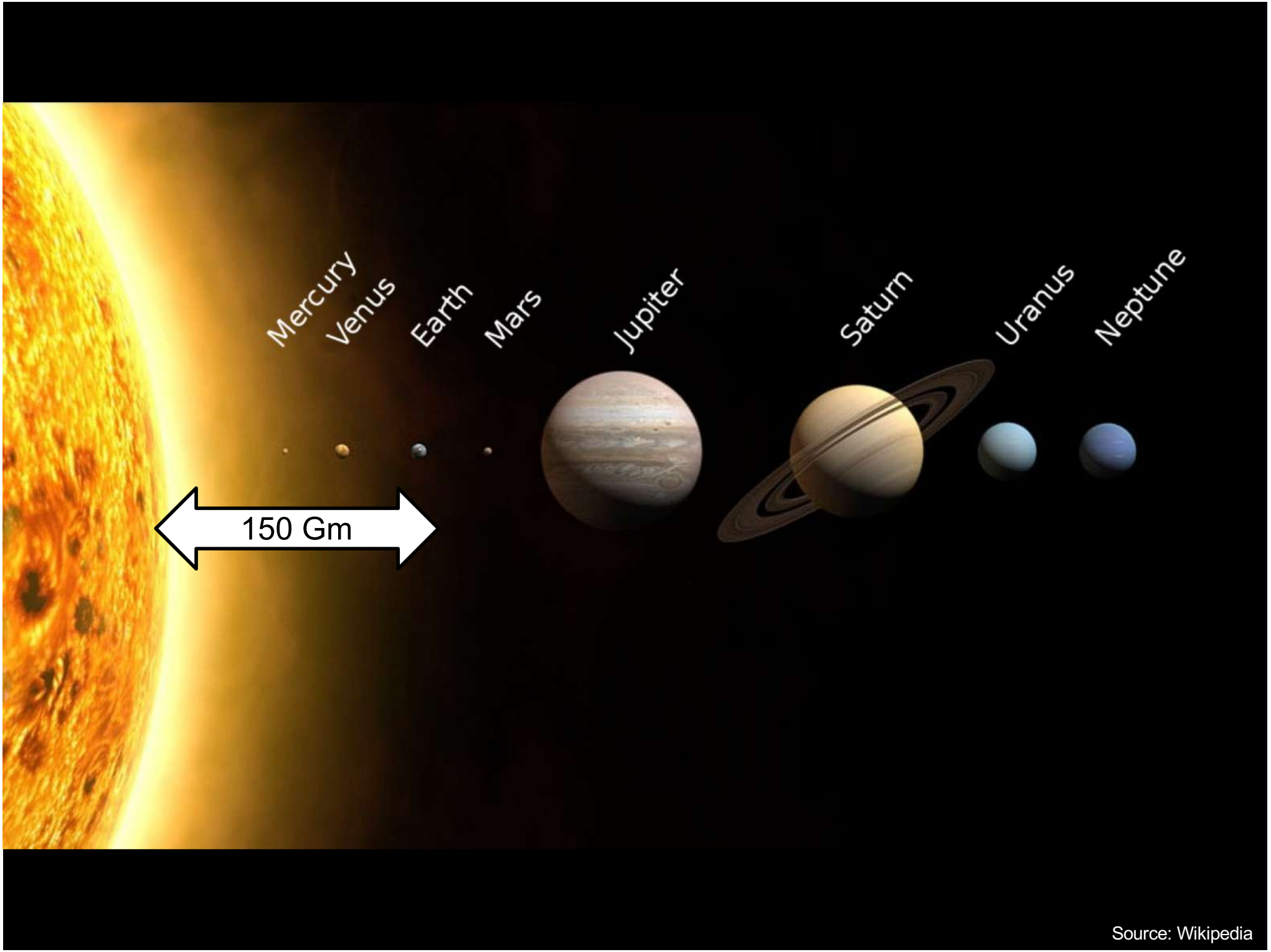
1.85m

123RF / shtanzman

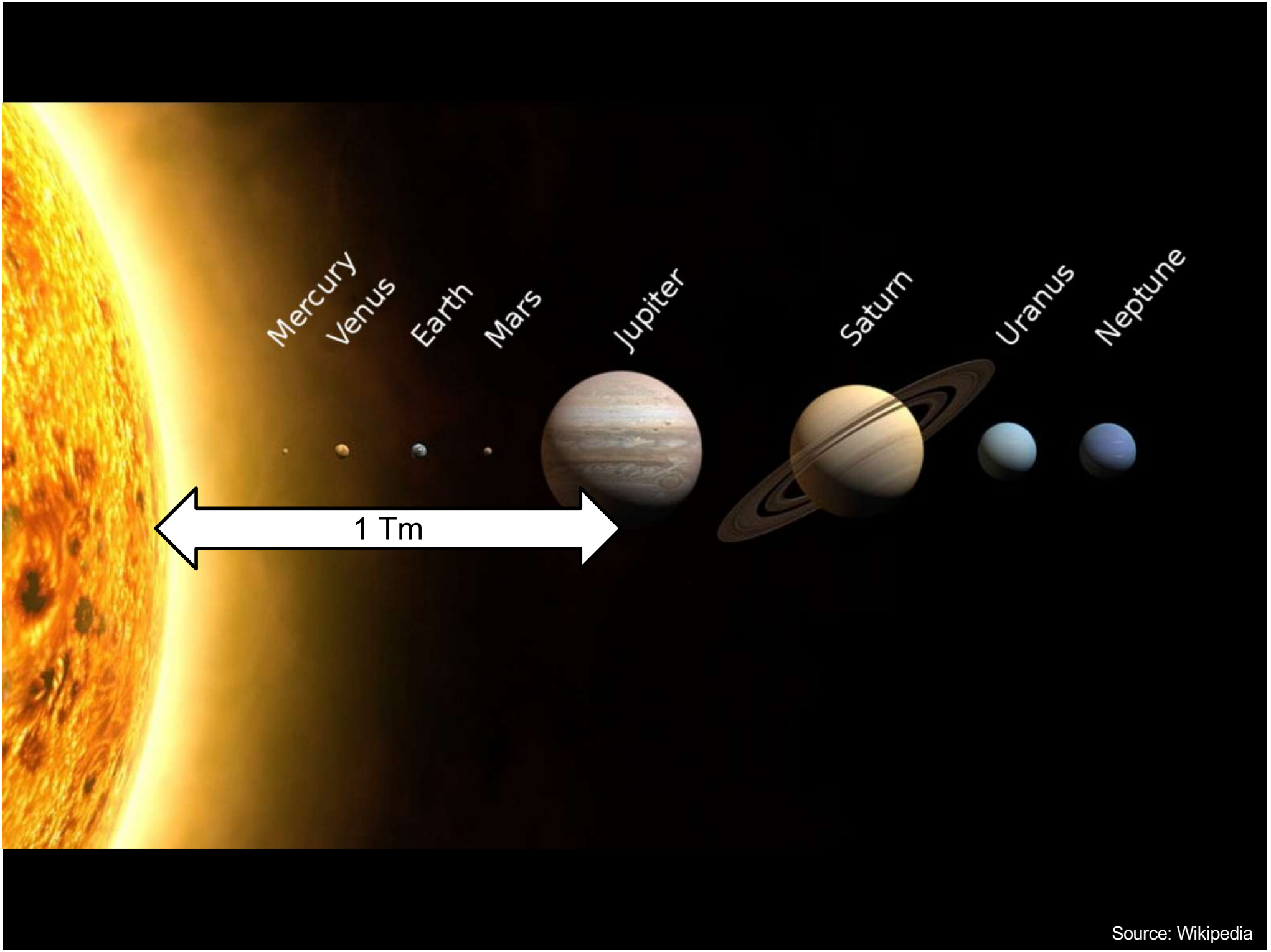

40,000 km


40 Mm

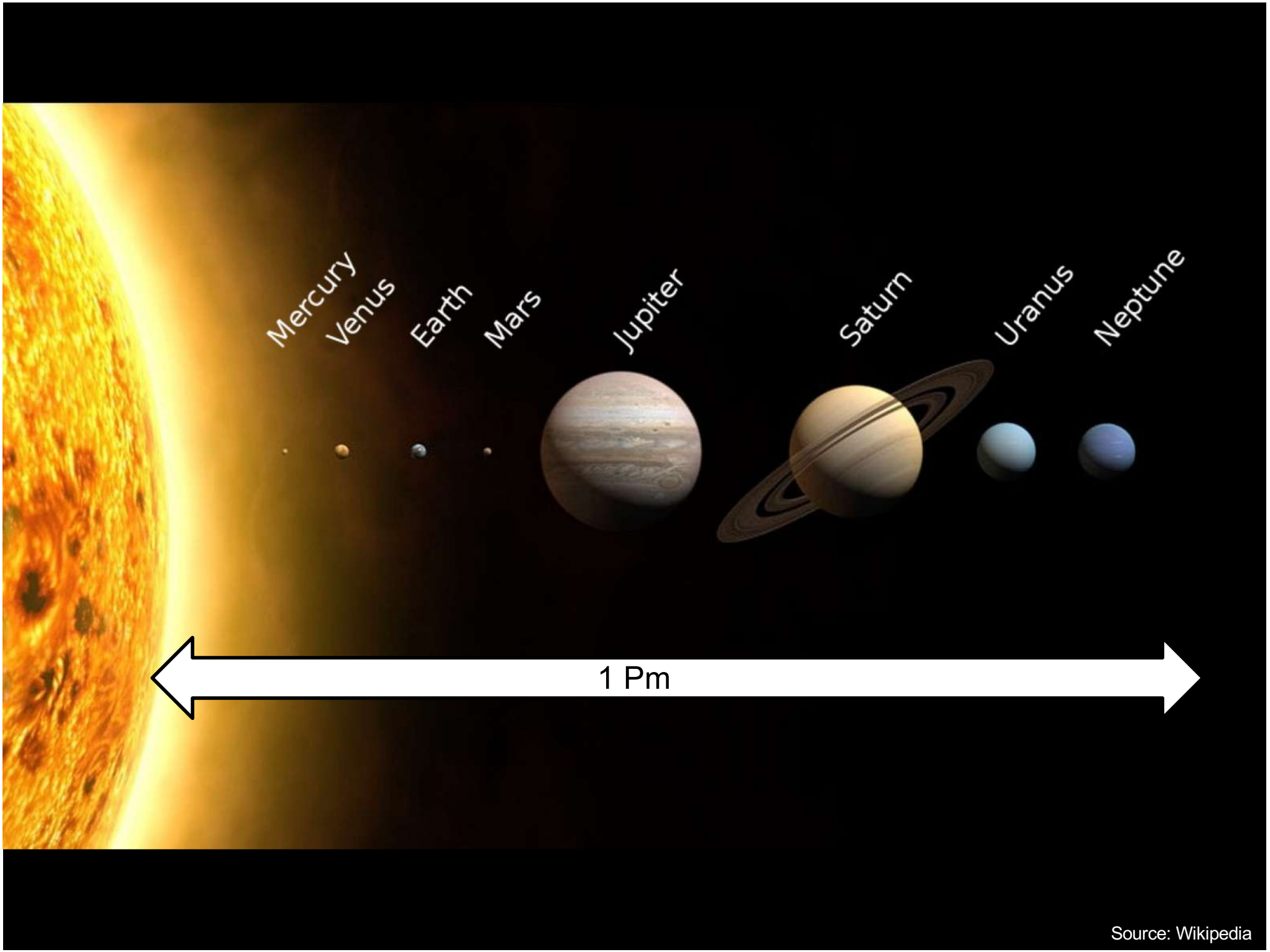

Mercury
Venus
Earth
Mars
Jupiter
Saturn
Uranus
Neptune
Source: Wikipedia

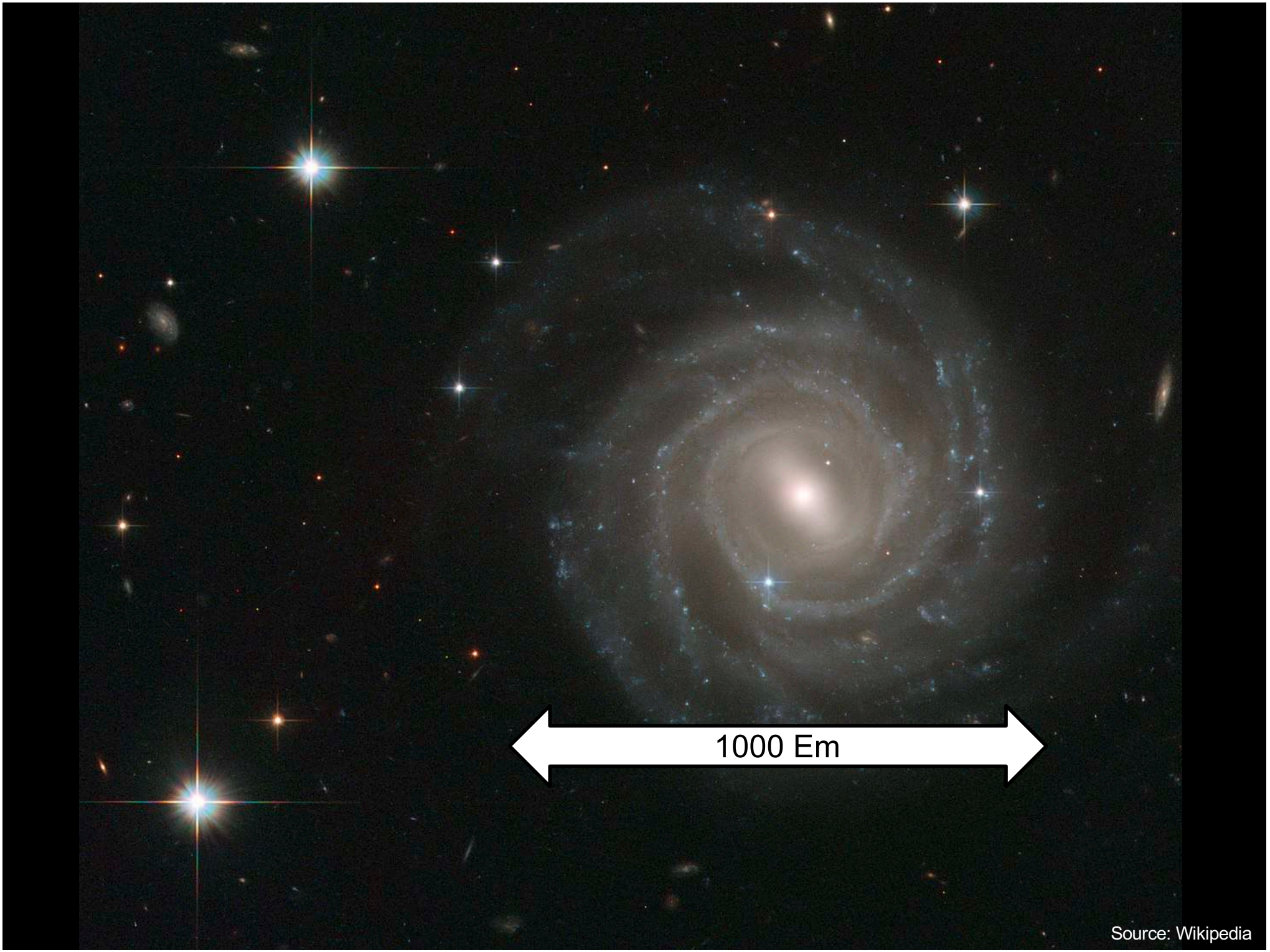


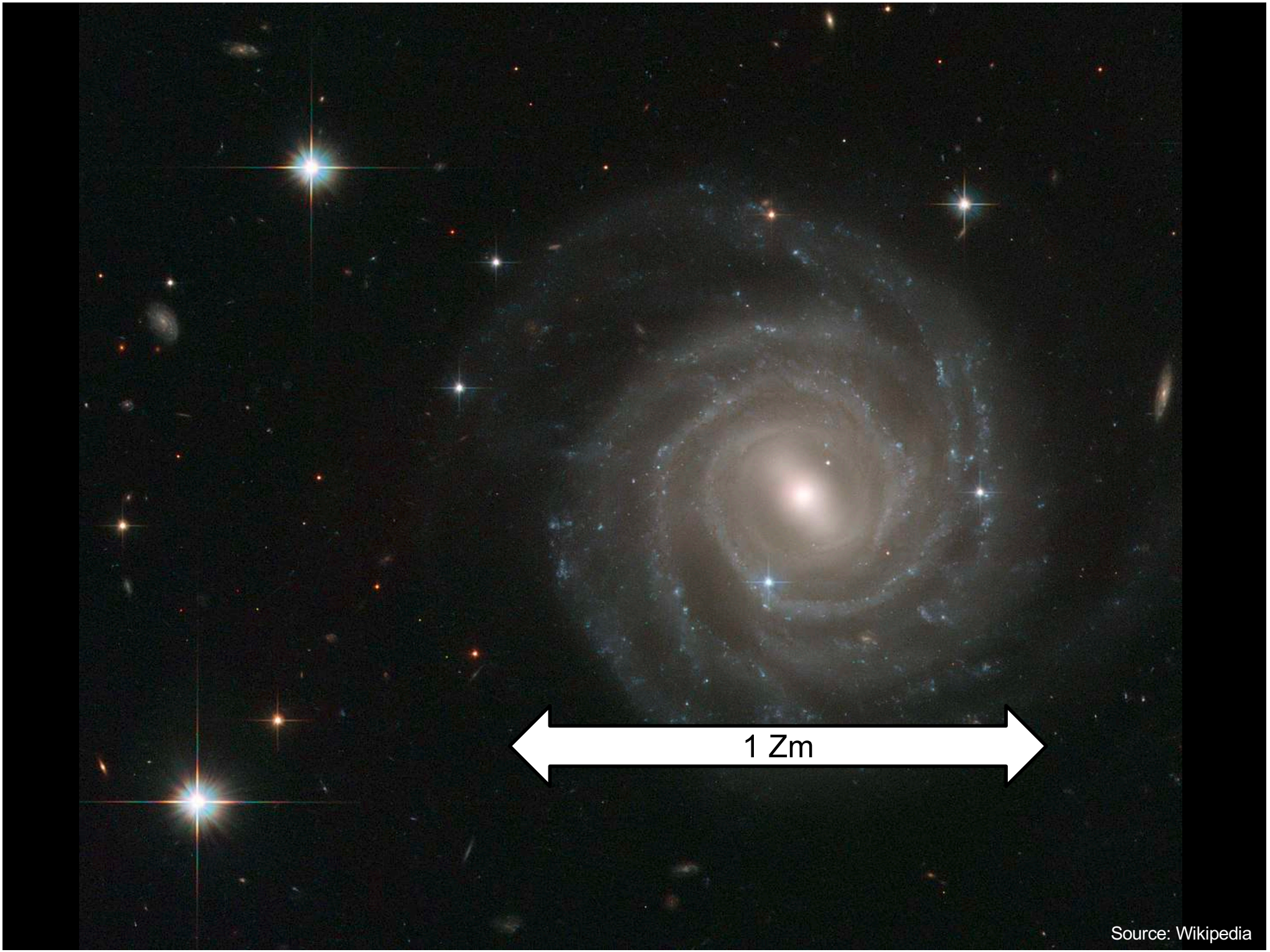
Source: Wikipedia

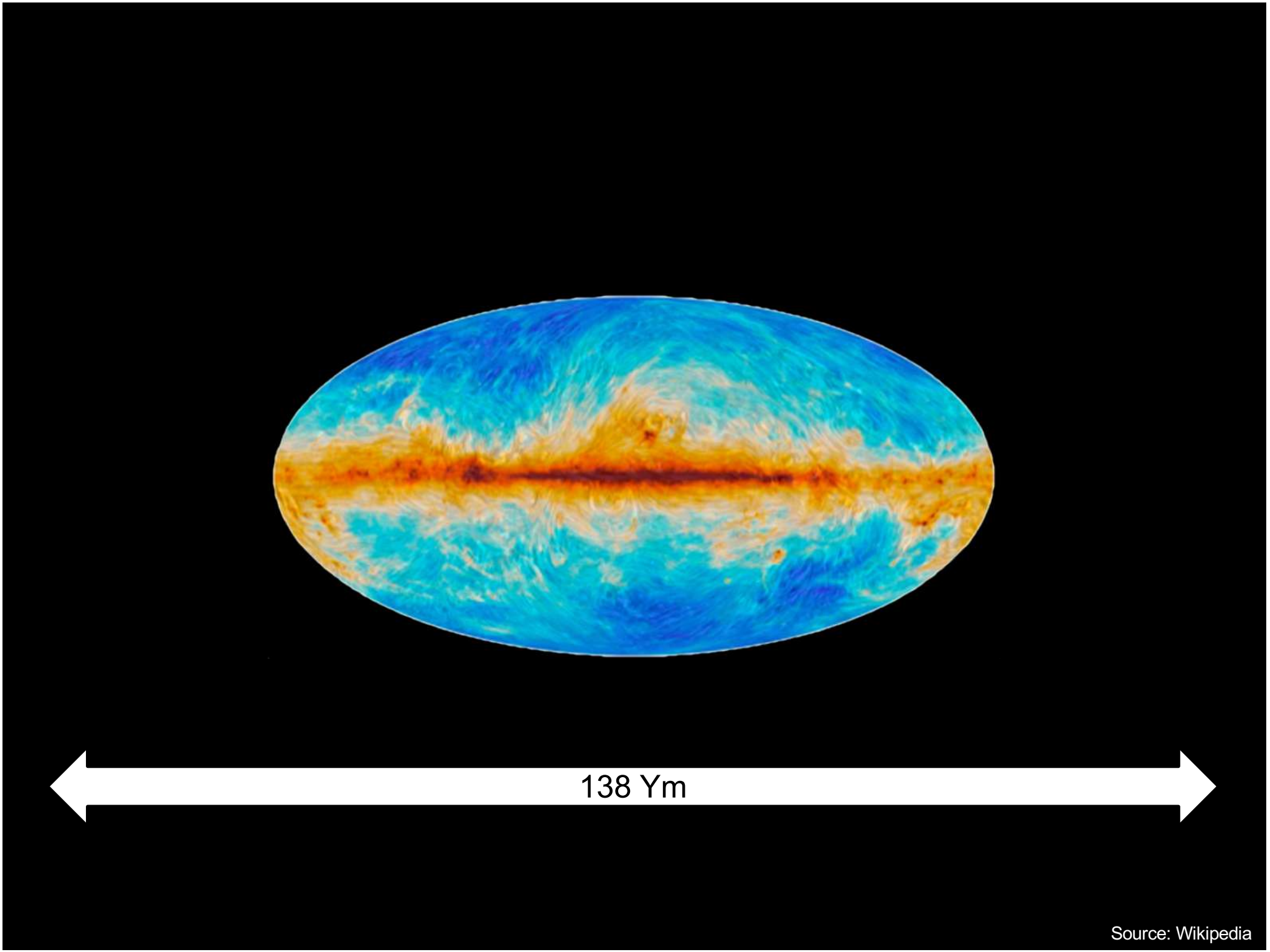


Source: Wikipedia

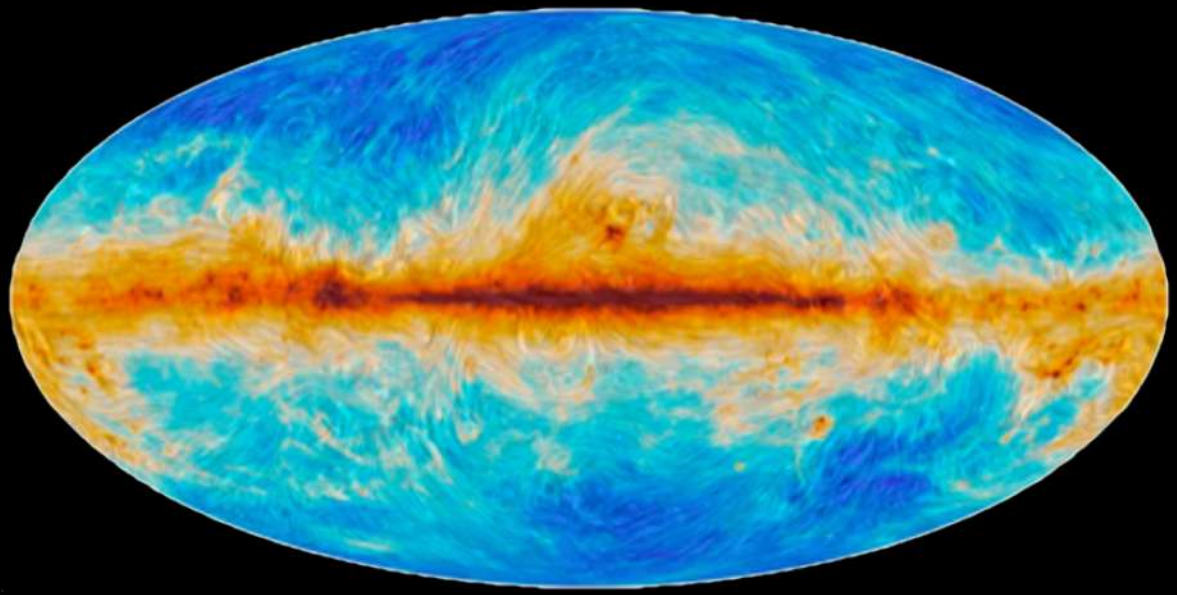

Mercury
Venus
Earth
Mars
Jupiter
Saturn
Uranus
Neptune
1 Pm
Source: Wikipedia


1000 Em
Source: Wikipedia


1 Zm
Source: Wikipedia

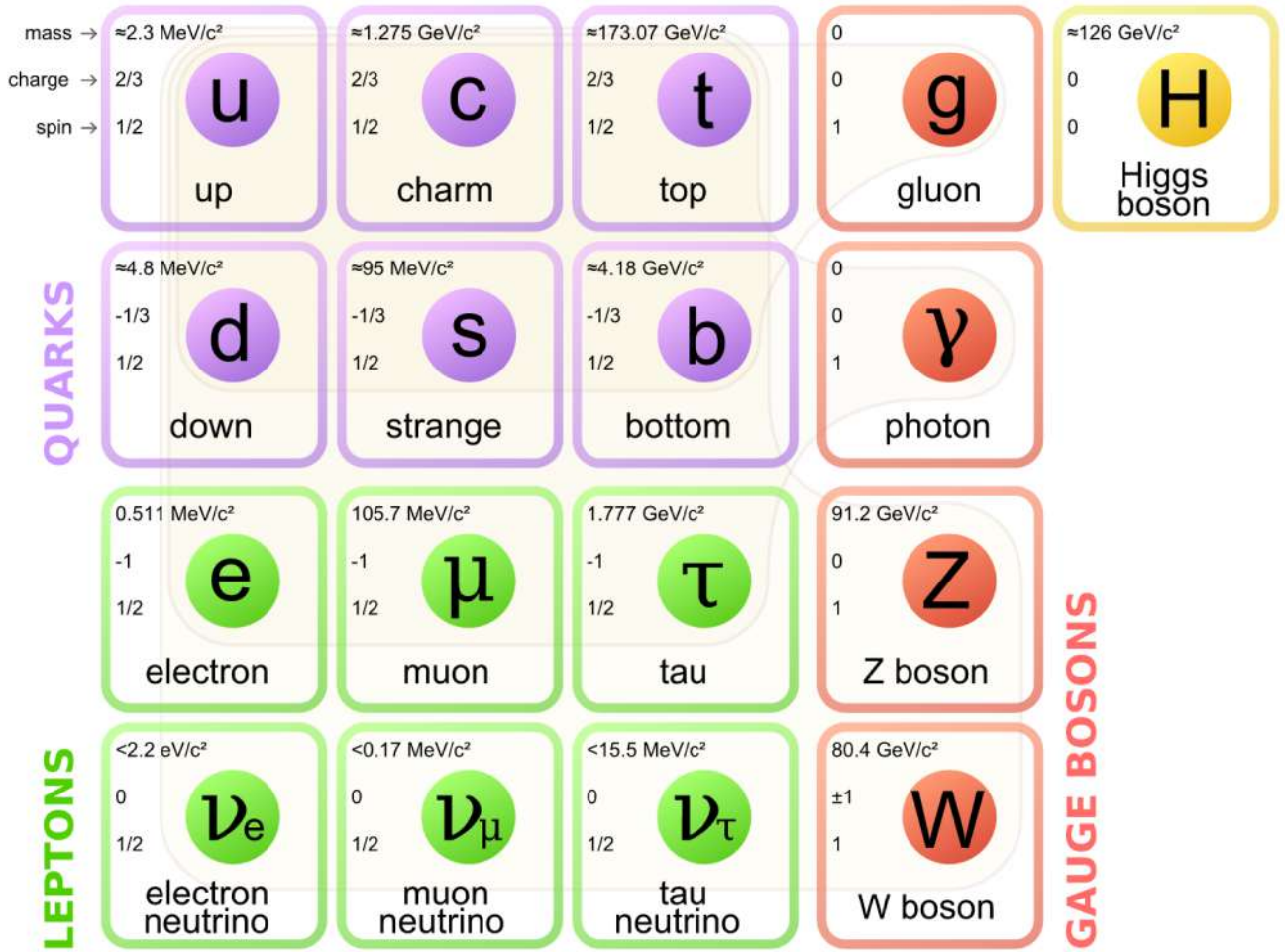

138 Ym
Source: Wikipedia

Exploring the infinitely big...



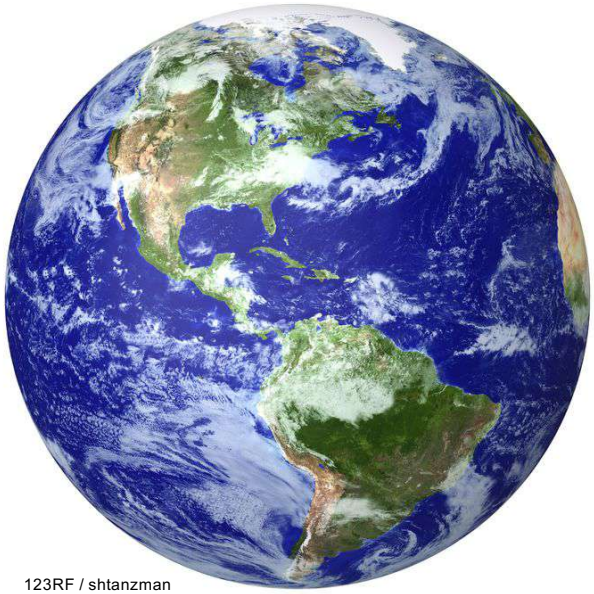
Source: Wikipedia

... means exploring the infinitely small



Source: Wikipedia

Data is like matter



123RF / shtanzman

Study of the real world

Study of the data world

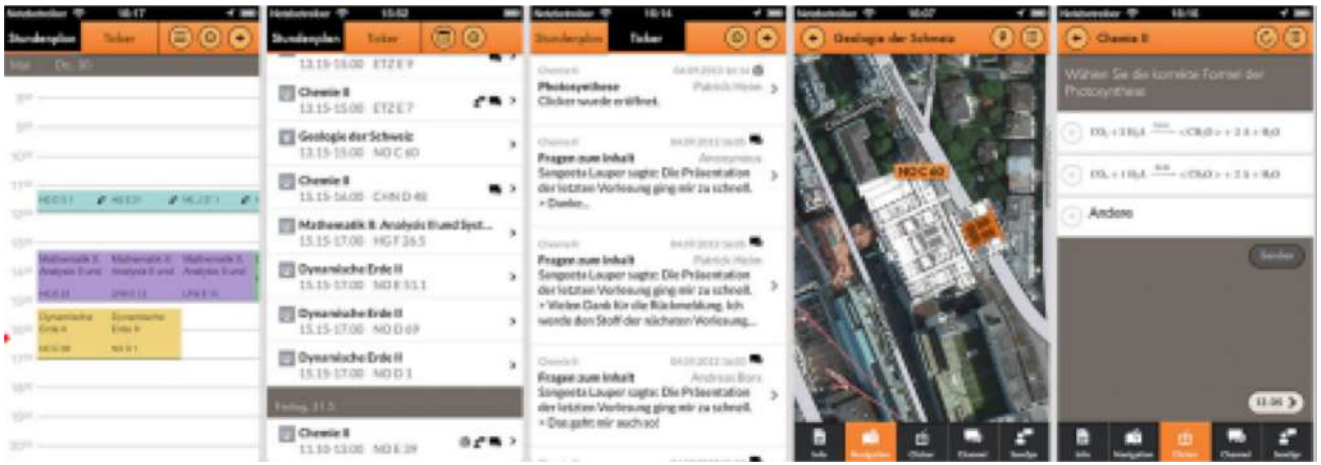
Physics Data Science

Poll

Go *now* to:

https://eduapp-app1.ethz.ch/

Available on the App Store or install EduApp 3.x ANDROID APP ON Google play



My How-We-Do-Science Matrix

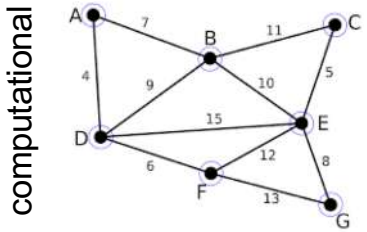
Mathematics

Physics

My How-We-Do-Science Matrix

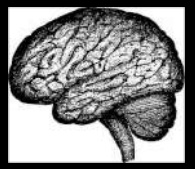
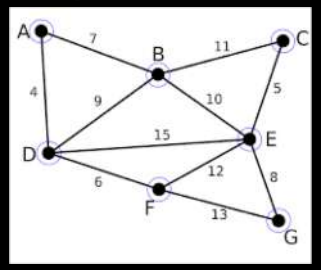
		Ontological	***Epistemic***
		The world as it must be (necessary)	The world as it is (contingent)
Thinking	With our brain (natural) 	Mathematics  theoretical	Physics  empirical

My How-We-Do-Science Matrix

		Ontological	***Epistemic***
		The world as it must be (necessary)	The world as it is (contingent)
Thinking	With our brain (natural) 	Mathematics (theoretical) 	Physics (empirical) 
Computing	With a machine (artificial) 	Computer Science (computational) 	

My How-We-Do-Science Matrix

The four paradigms

		Ontological	***Epistemic***
		The world as it must be (necessary)	The world as it is (contingent)
Thinking	With our brain (natural) 	Mathematics 	Physics 
Computing	With a machine (artificial) 	Computer Science 	**Data Science** 

theoretical

empirical

computational

data driven

The Physics of CS

A good decision is based on knowledge, not on numbers.

- Plato

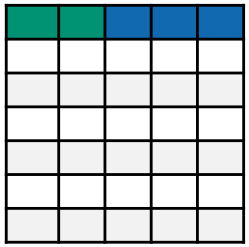
Data Management Lectures at ETH

Computer Science
Data Science
CBB MSc with CS background

Big Data
(Fall)

This lecture

Other departments



Information Systems for Engineers

Spring 2020

Big Data for Engineers

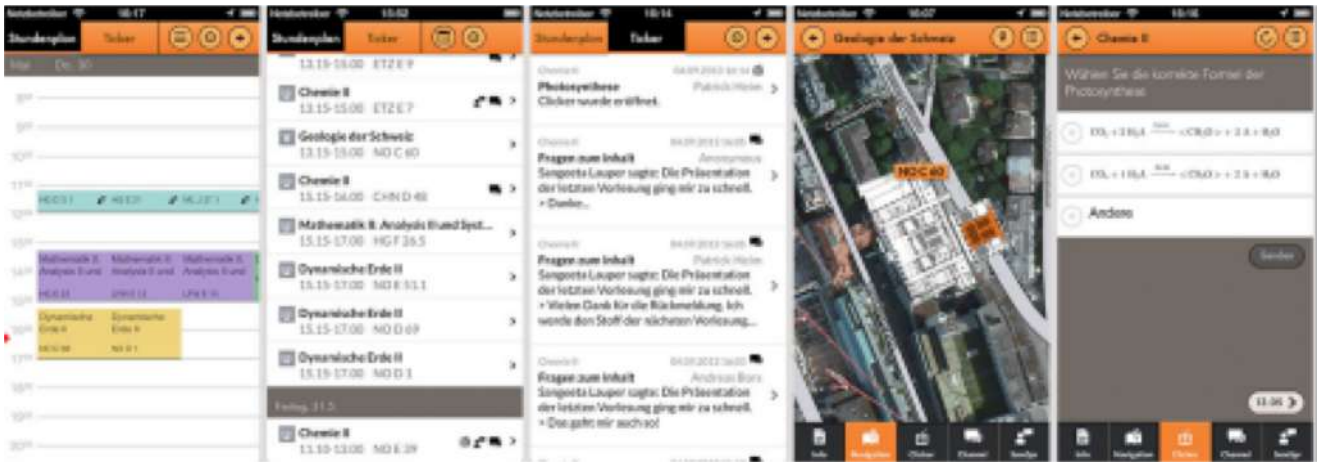
Spring 2020

Poll

Go *now* to:

https://eduapp-app1.ethz.ch/

Available on the App Store or install EduApp 3.x ANDROID APP ON Google play

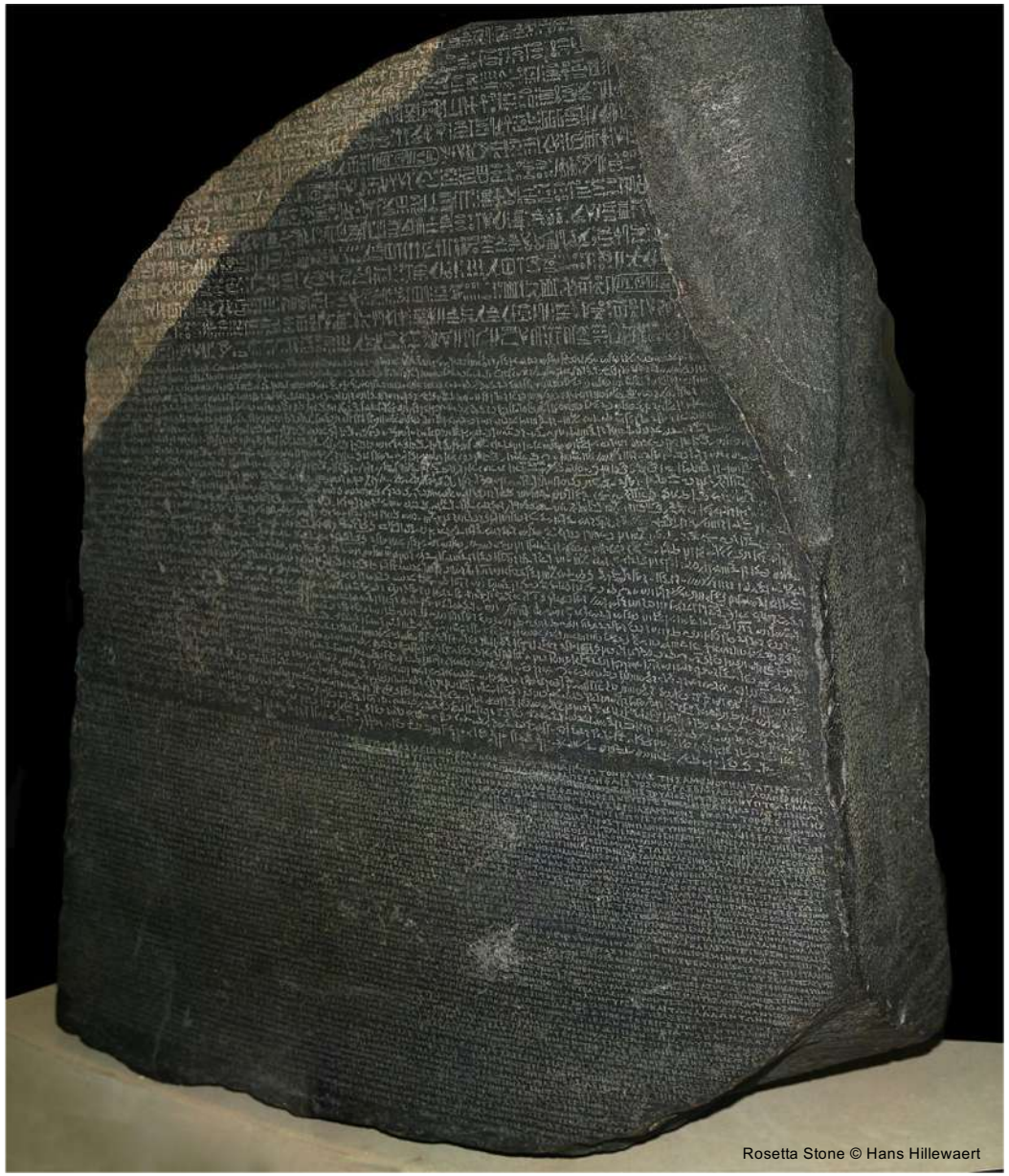


123RF / Samantha Craddock

A Short History of Databases

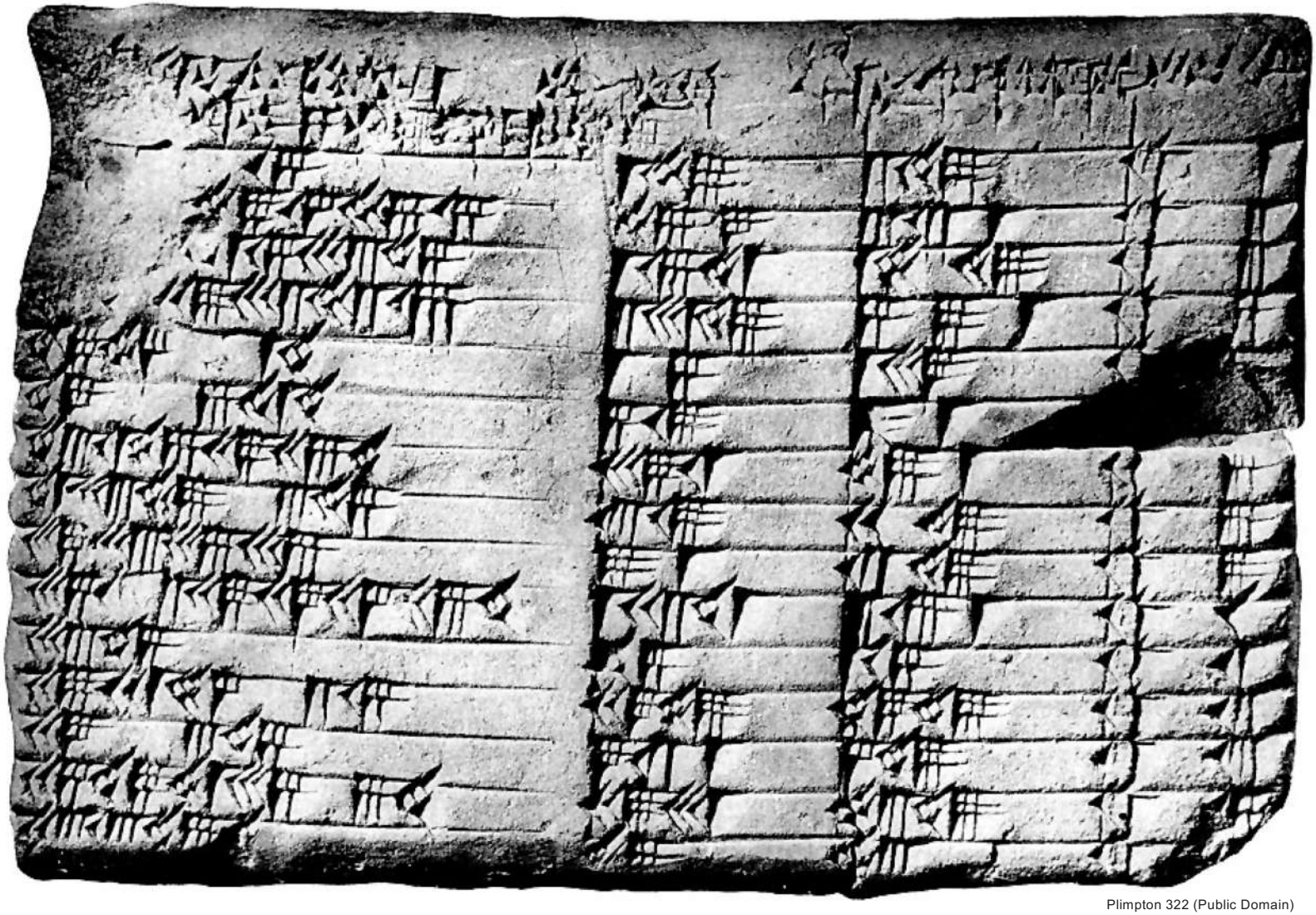

Database Prehistory
123RF / andreykuzmin

Speaking/Singing

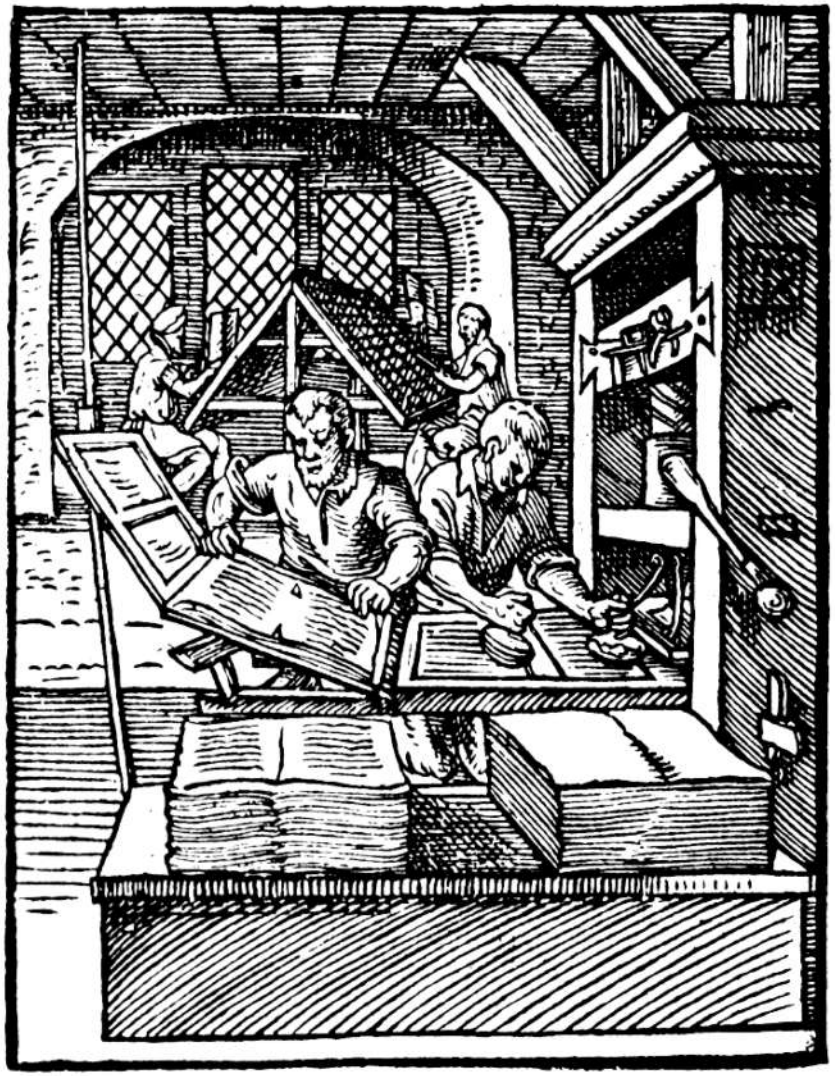

Rosetta Stone © Hans Hillewaert

Writing

Accounting

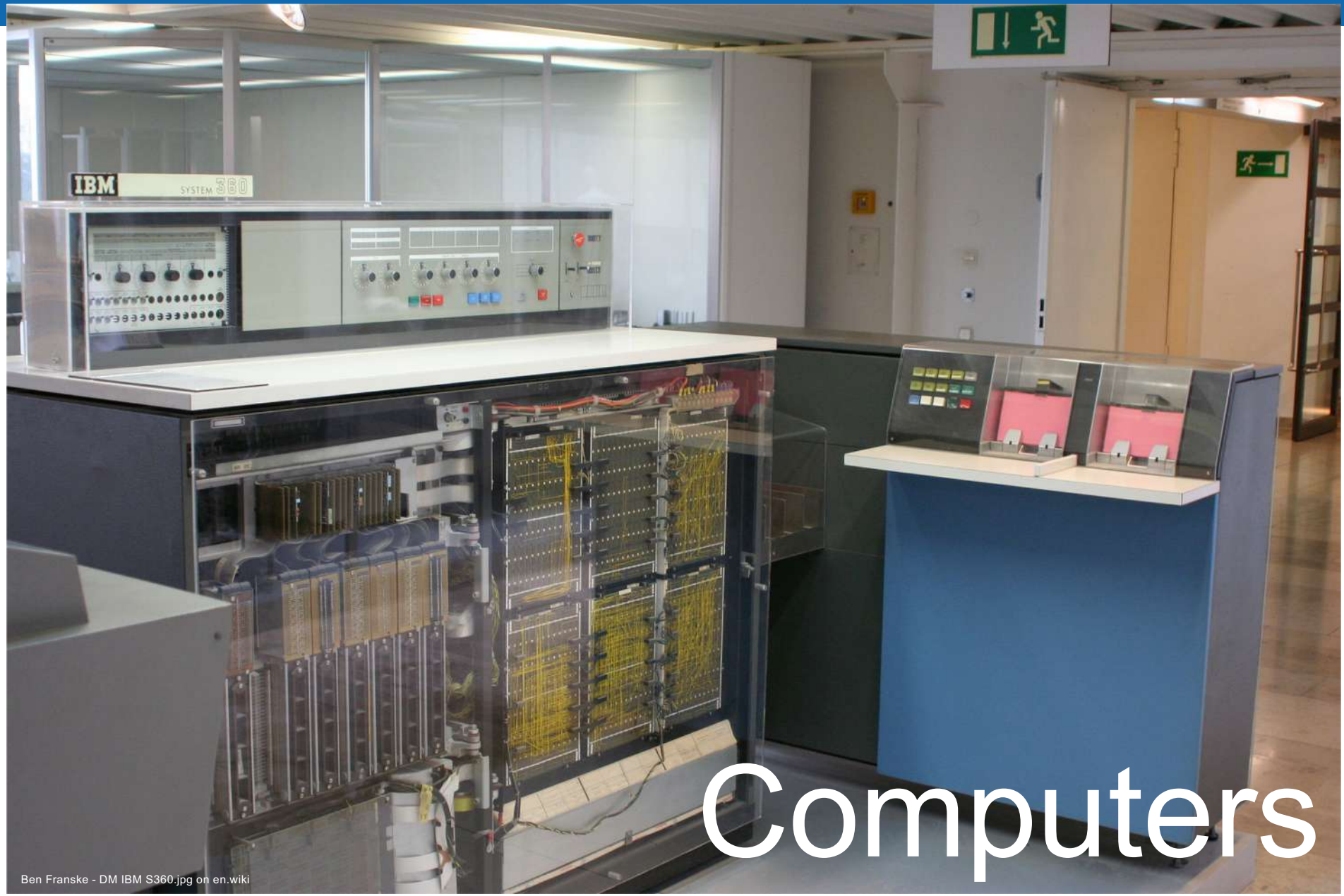


Plimpton 322 (Public Domain)



Printing

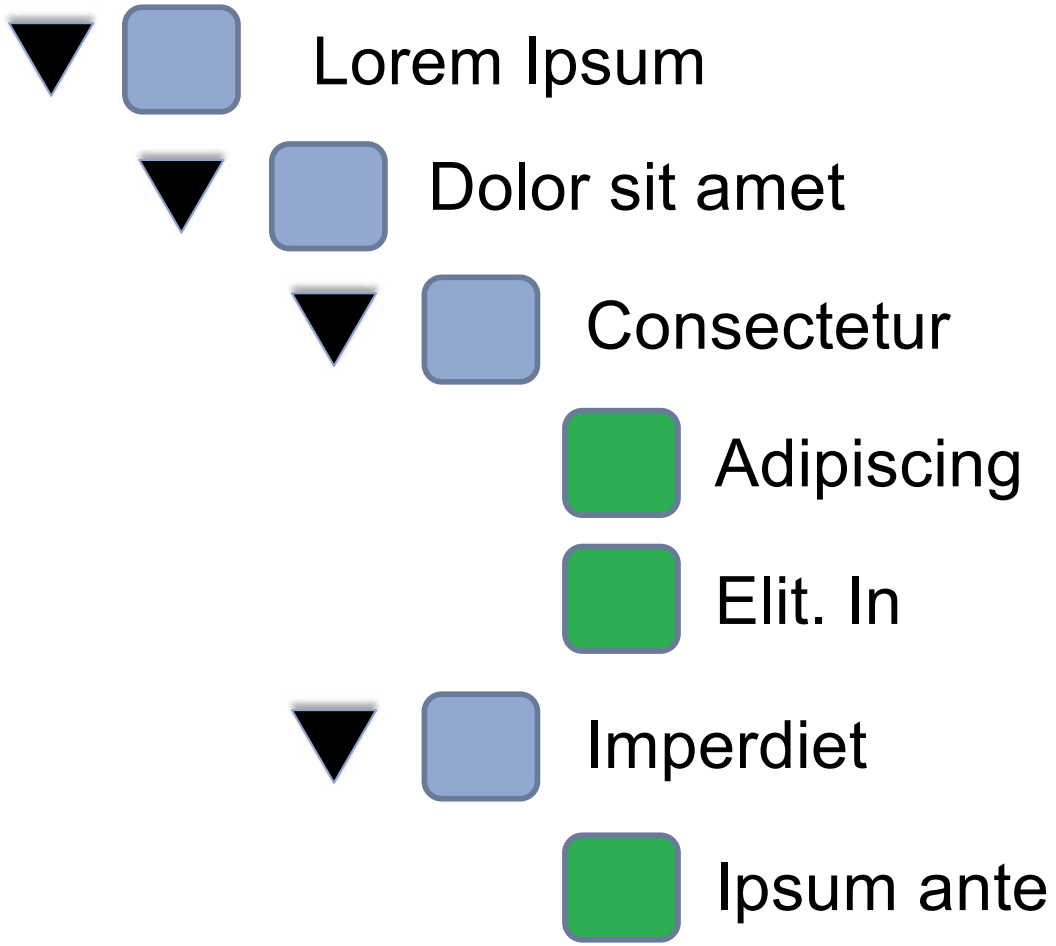
Willi Heidelbach



Computers

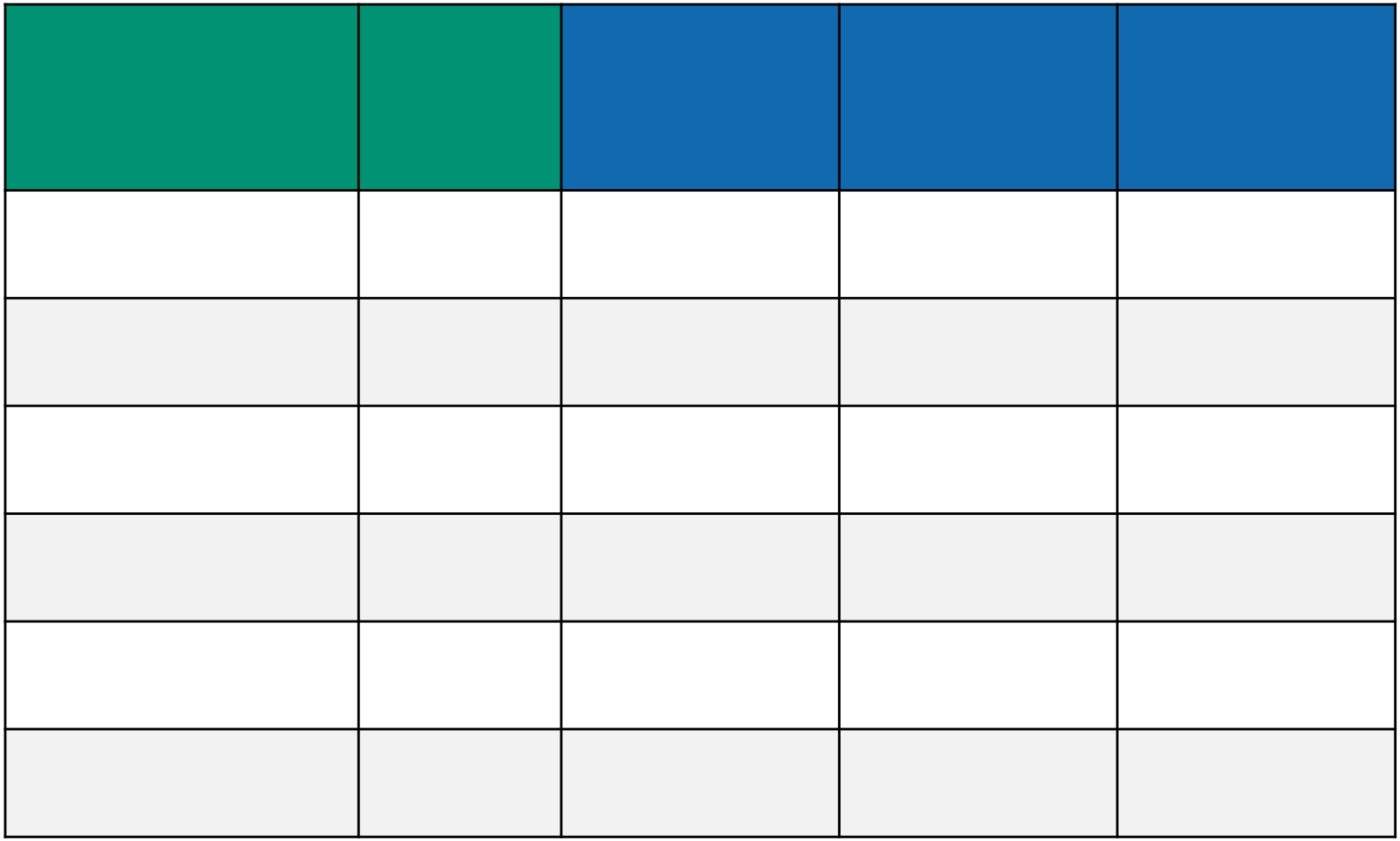
Ben Franske - DM IBM S360.jpg on en.wiki

1960s: File Systems



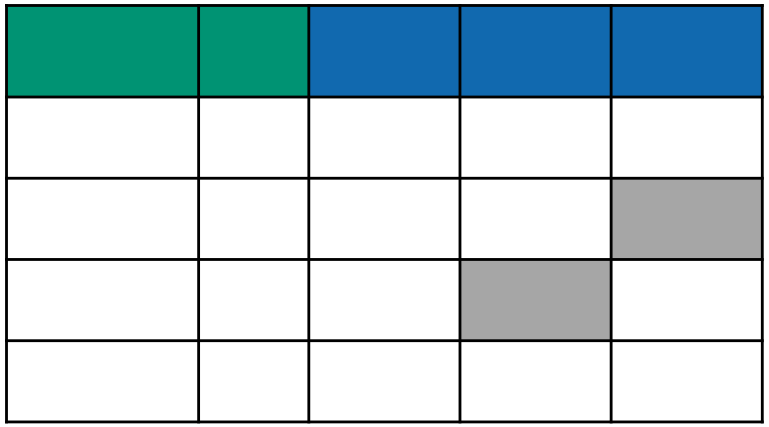
123RF / andreykuzmin

1970s: The Relational Era



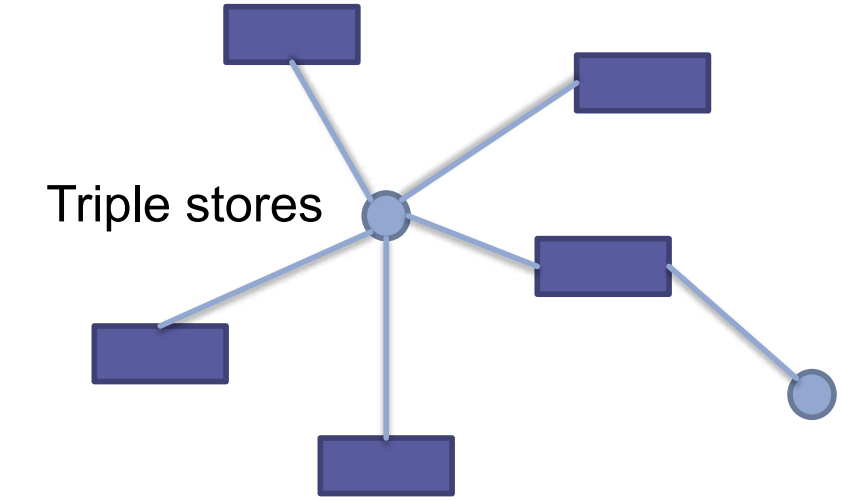
2000s: The NoSQL Era

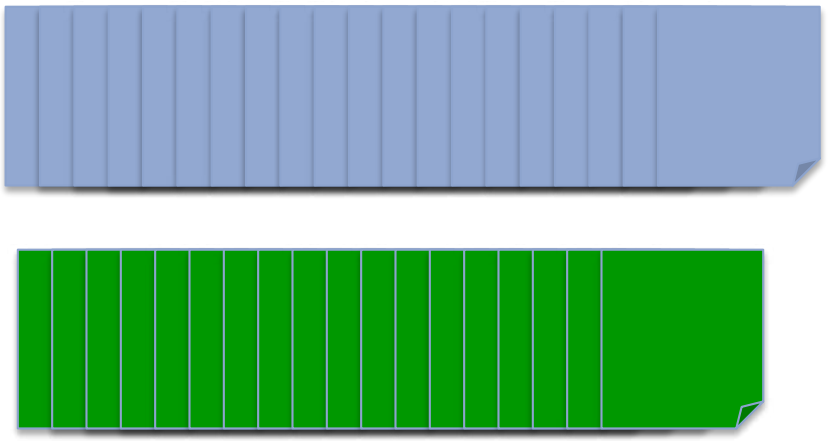
Key-value stores



Column stores

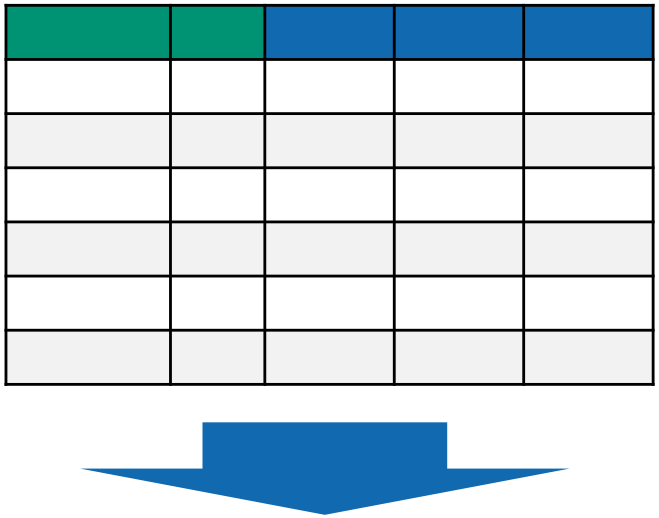
Triple stores





Document stores

In short?



We threw data at computers.

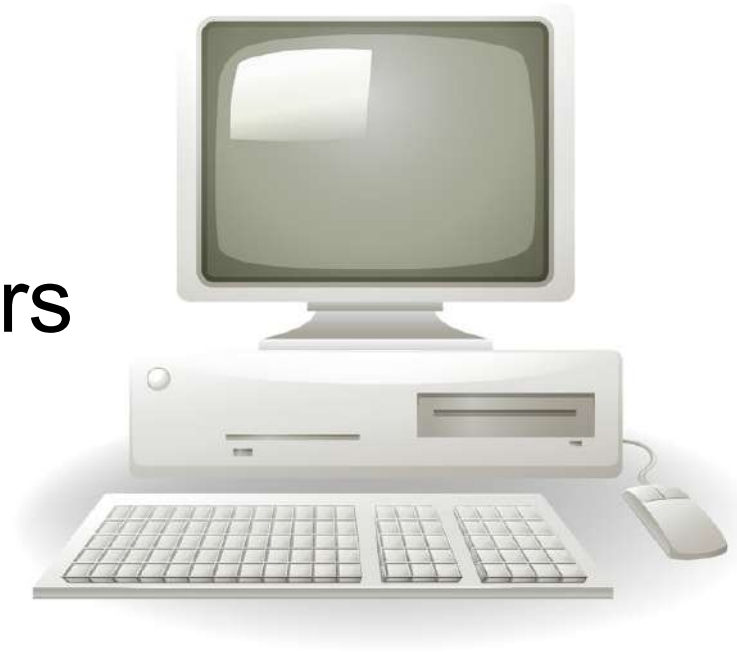
1970

123RF / blueringmedia

In short?

We threw computers at computers.

1990



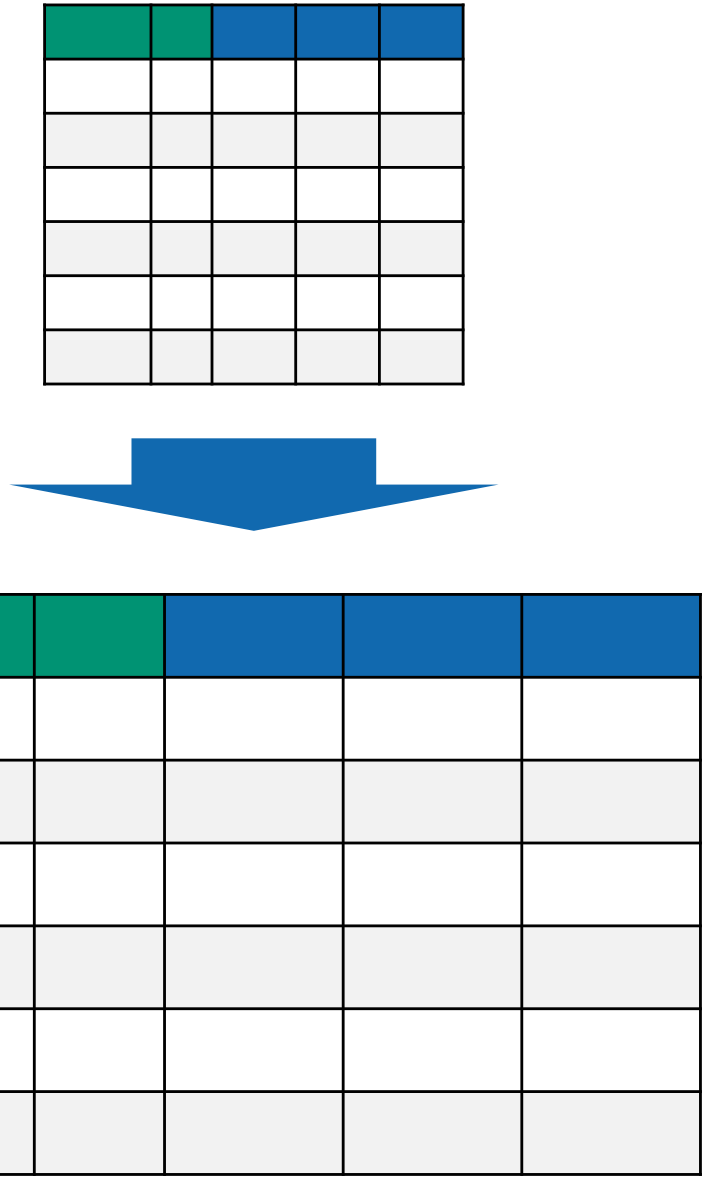
123RF / blueringmedia

In short?

We threw computers at data.

2000

In short?

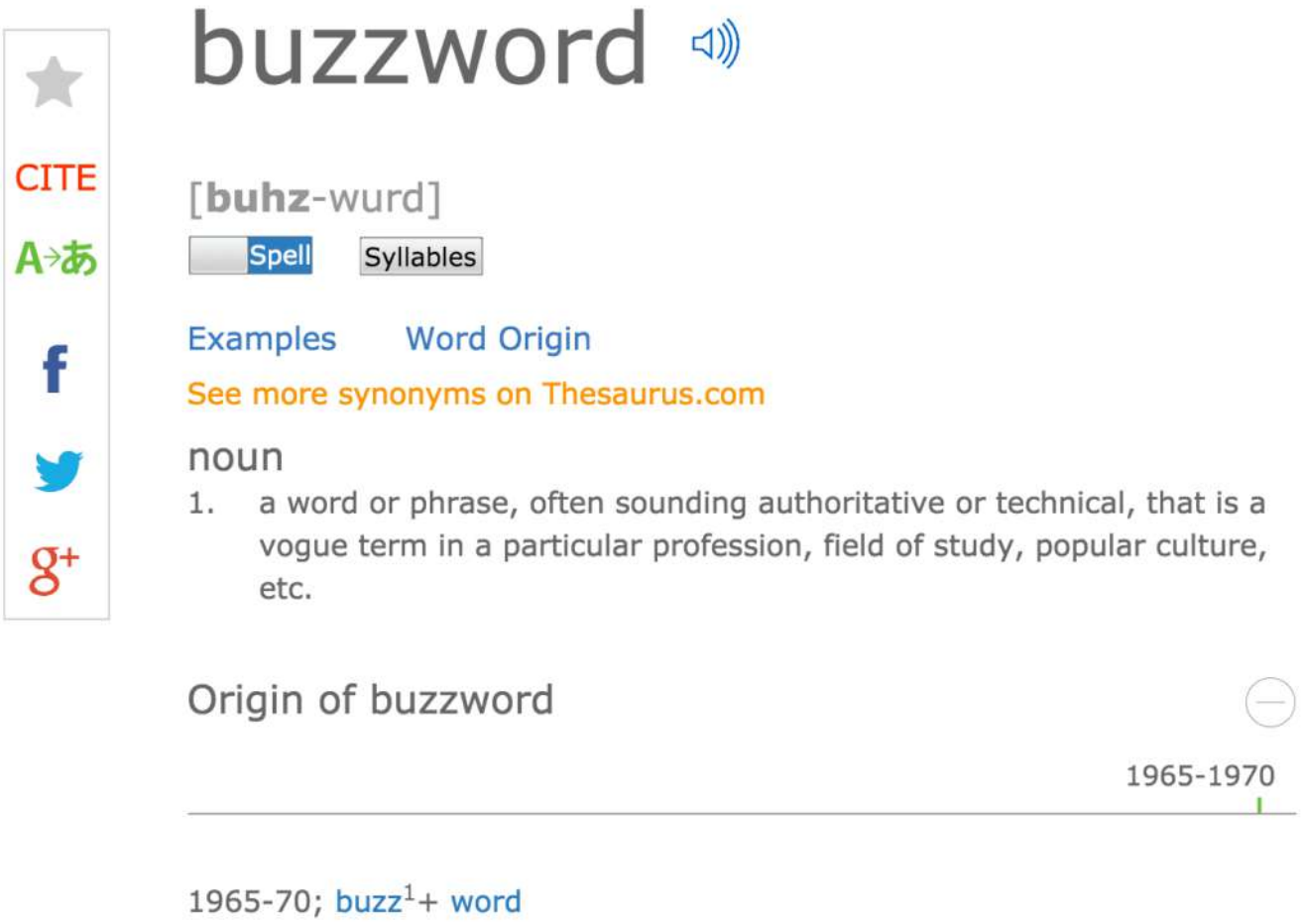


We are throwing data at data.

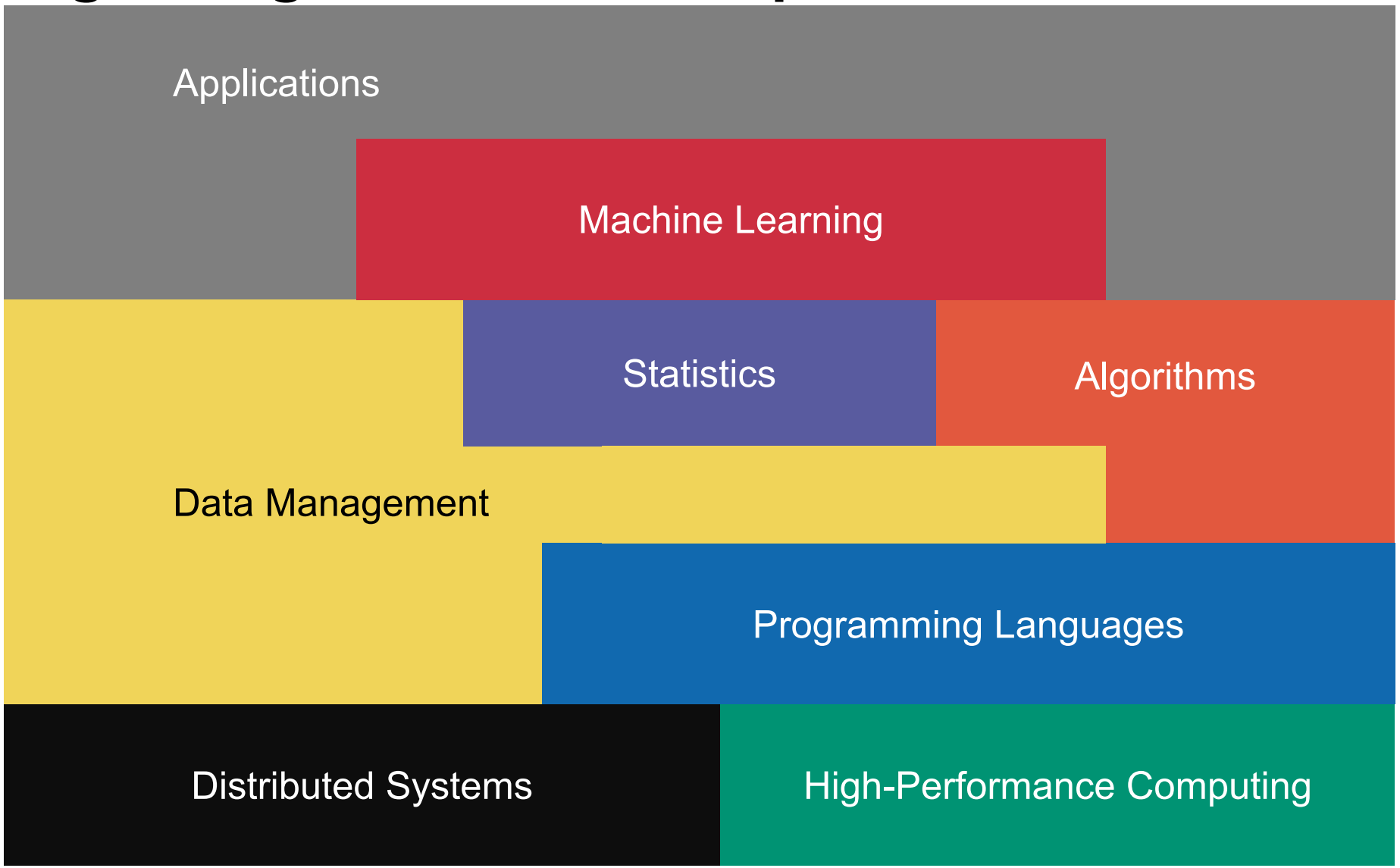
now

Big Data

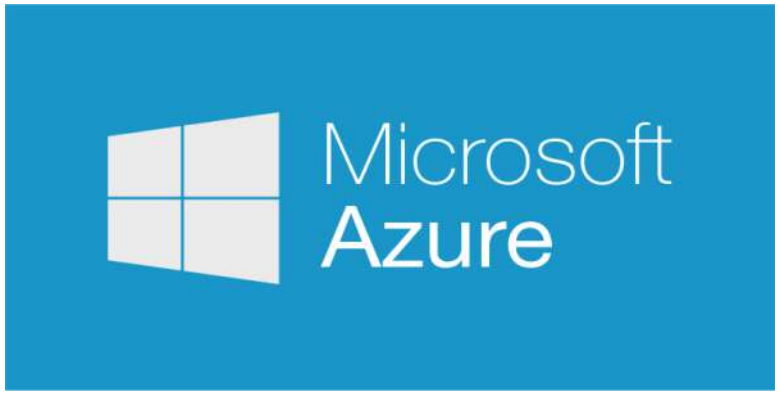
It's a buzzword!

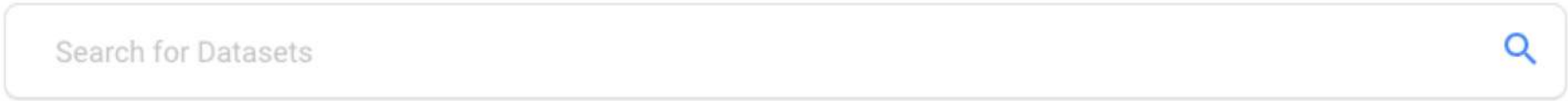


Big Data goes across disciplines



Big Data involves a lot of proprietary technology

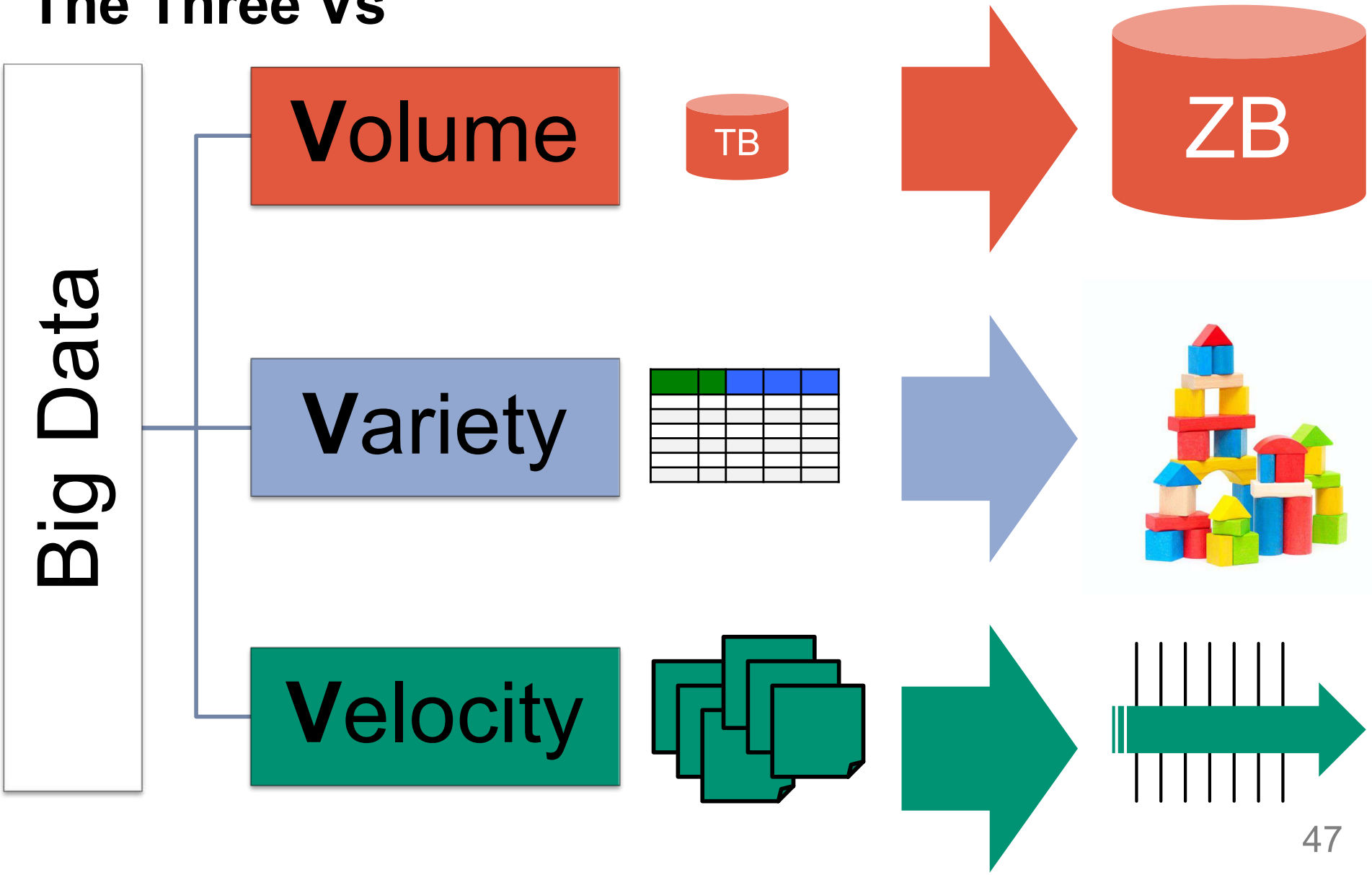




Try boston education data or weather site:noaa.gov

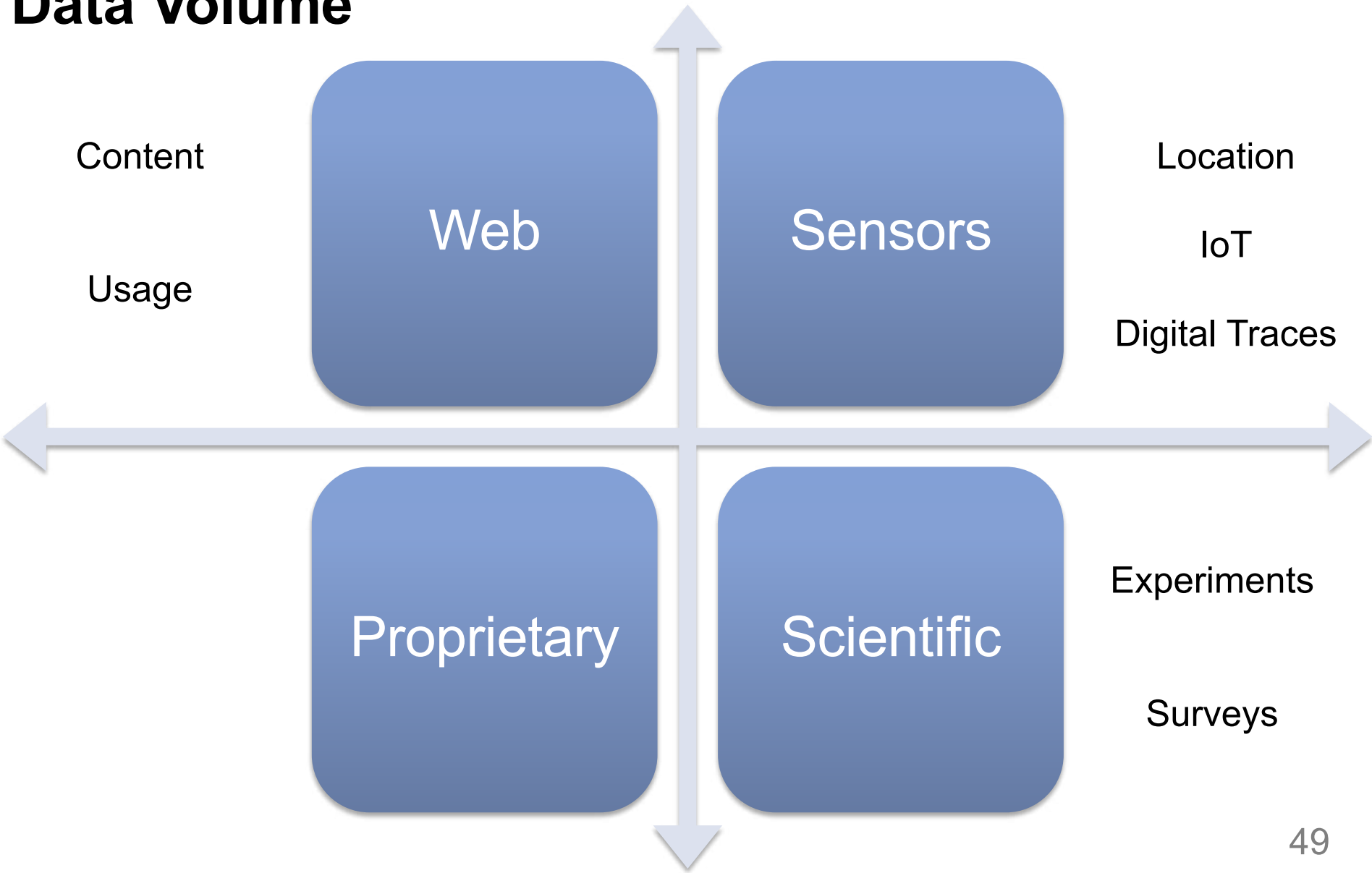
The Big in Big Data

The Three Vs



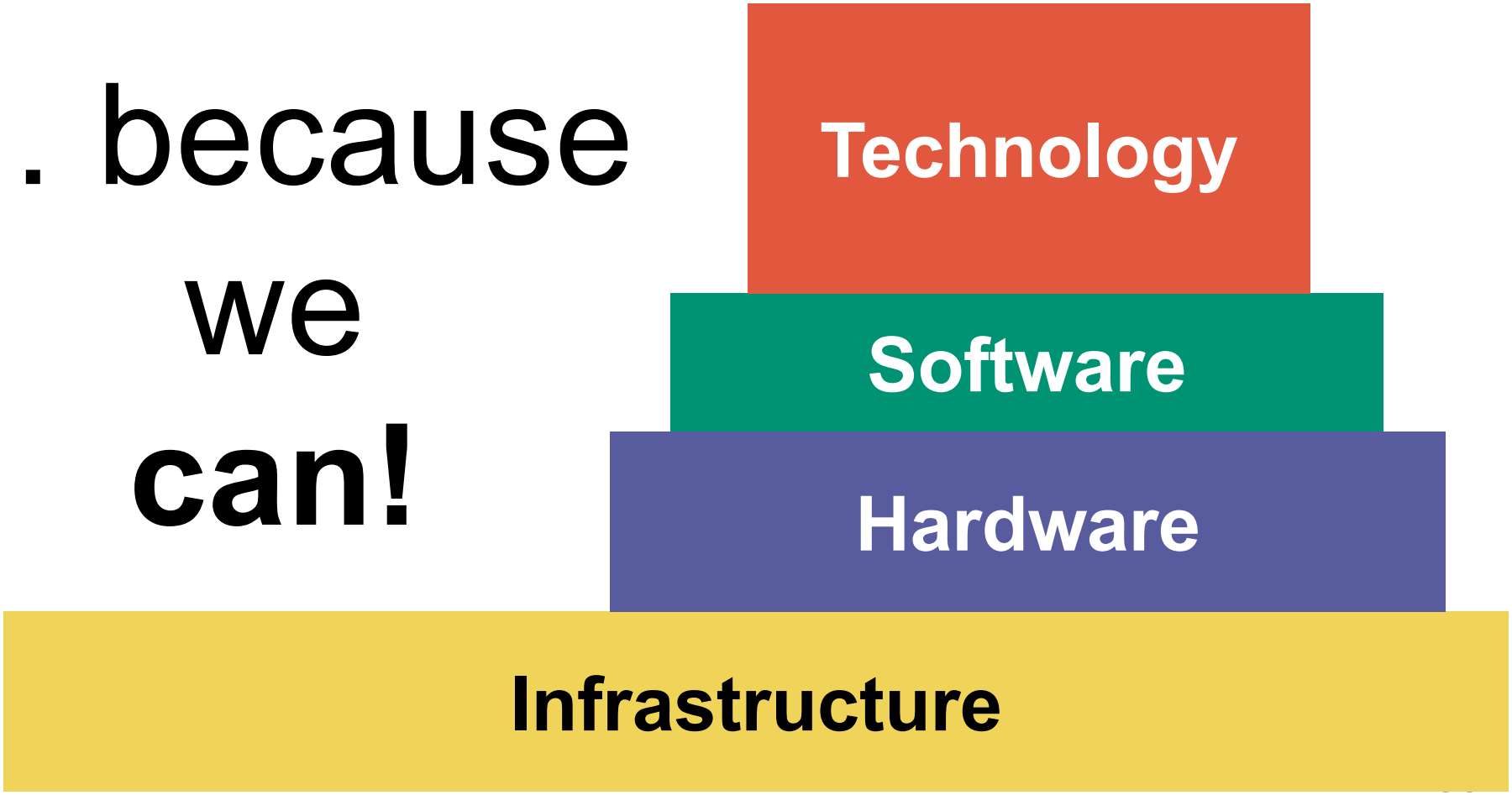
Data Volume

Data Volume


Content
Usage
Web
Sensors
Location
IoT
Digital Traces
Proprietary
Scientific
Experiments
Surveys

Data Volume

… because we **can!**



Data Volume

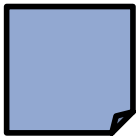
… because

data

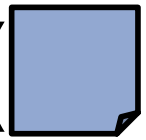
carries

value

Data is worth more than the sum of its parts

Utility( + )

>

Utility() + Utility()

Data totality: one must have complete data

All flights
All hotels
All shops
…

Prefixes (International System of Units)

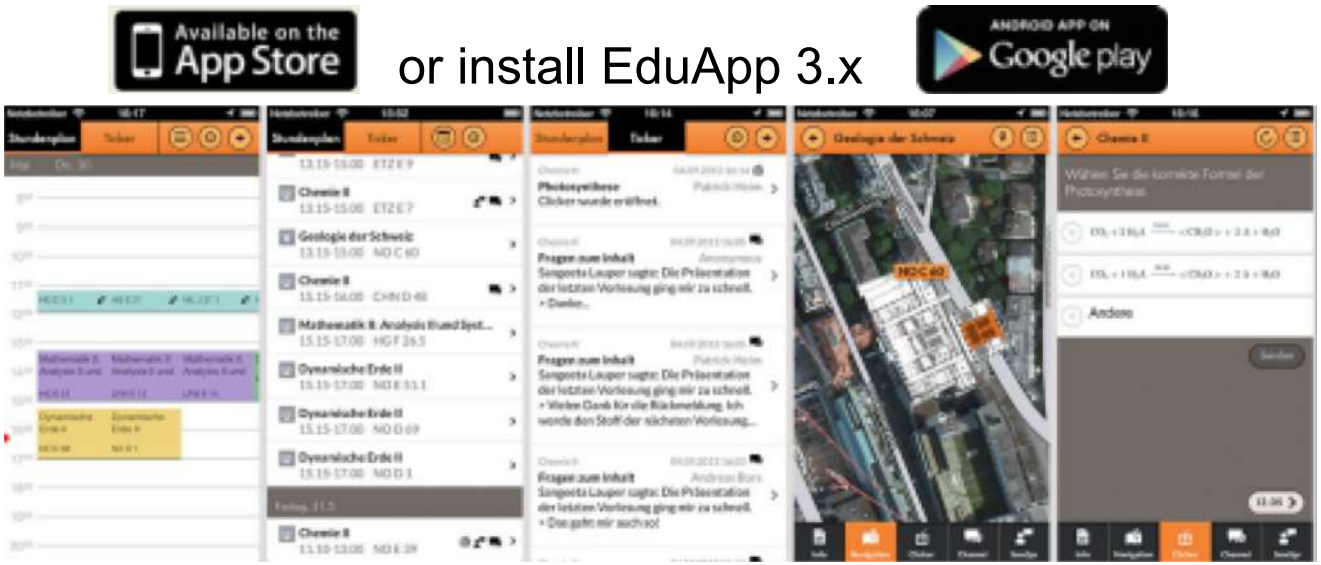
kilo (k)	1,000 (3 zeros)
Mega (M)	1,000,000 (6 zeros)
Giga (G)	1,000,000,000 (9 zeros)
Tera (T)	1,000,000,000,000 (12 zeros)
Peta (P)	1,000,000,000,000,000 (15 zeros)
Exa (E)	1,000,000,000,000,000,000 (18 zeros)
Zetta (Z)	1,000,000,000,000,000,000,000 (21 zeros)
Yotta (Y)	1,000,000,000,000,000,000,000,000 (24 zeros)

You must know this by ❤!

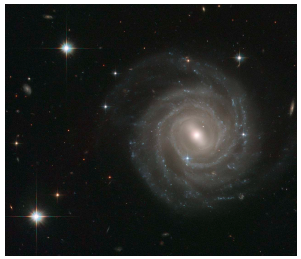
Clicker question

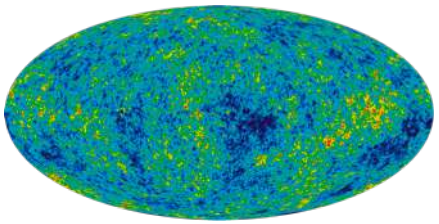
Go *now* to:

https://eduapp-app1.ethz.ch/



Prefixes (International System of Units)





kilo (k)
Mega (M)
Giga (G)
Tera (T)
Peta (P)
Exa (E)
Zetta (Z)
Yotta (Y)

Prefixes (International System of Units)

kibi (ki)	1,024 (2^{10})
Mebi (Mi)	1,048,576 (2^{20})
Gibi (Gi)	1,073,741,824 (2^{30})
Tebi (Ti)	1,099,511,627,776 (2^{40})
Pebi (Pi)	1,125,899,906,842,624 (2^{50})
Exbi (Ei)	1,152,921,504,606,846,976 (2^{60})
Zebi (Zi)	1,180,591,620,717,411,303,424 (2^{70})
Yobi (Yi)	1,208,925,819,614,629,174,706,176 (2^{80})

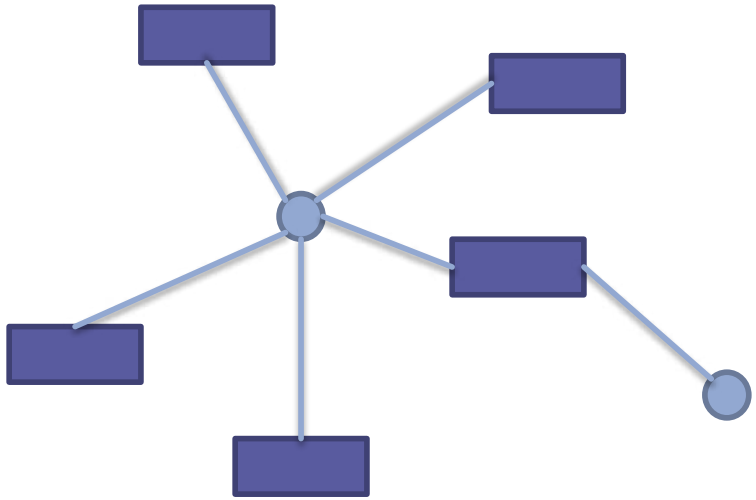
You must NOT know this by ❤ !

Data Variety

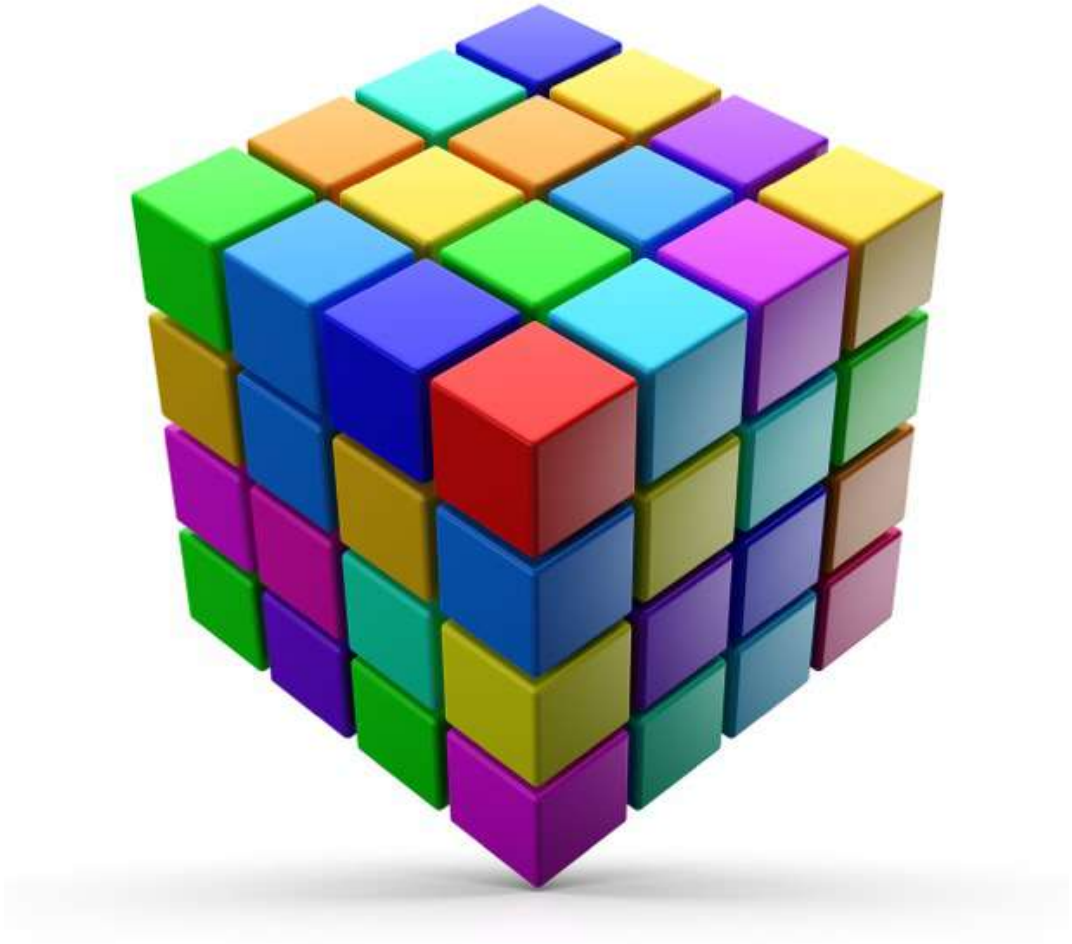
Data Shapes: Tables

Data Shapes: Trees

Data Shapes: Graphs



Data Shapes: Cubes



Data Shapes: Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vel erat nec dui aliquet vulputate sed quis nulla. Donec eget ultricies magna, eu dignissim elit. Nullam sed urna nec nisl rhoncus ullamcorper placerat et enim. Integer varius ornare libero quis consequat. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean eu efficitur orci. Aenean ac posuere tellus. Ut id commodo turpis.

Praesent nec libero metus. Praesent at turpis placerat, congue ipsum eget, scelerisque justo. Ut volutpat, massa ac lacinia cursus, nisl dui volutpat arcu, quis interdum sapien turpis in tellus. Suspendisse potenti. Vestibulum pharetra justo massa, ac venenatis mi condimentum nec. Proin viverra tortor non orci suscipit rutrum. Phasellus sit amet euismod diam. Nullam convallis nunc sit amet diam suscipit dapibus. Integer porta hendrerit nunc. Quisque pharetra congue porta. Suspendisse vestibulum sed mi in euismod. Etiam a purus suscipit, accumsan nibh vel, posuere ipsum. Nulla nec tempor nibh, id venenatis lectus. Duis lobortis id urna eget tincidunt.

Data Velocity

Data is generated automatically

Picture: Vladimir Voronin/123RF

Data is a realtime byproduct of human activity

Picture: Ash Waechter/123RF

Three paramount factors

Capacity

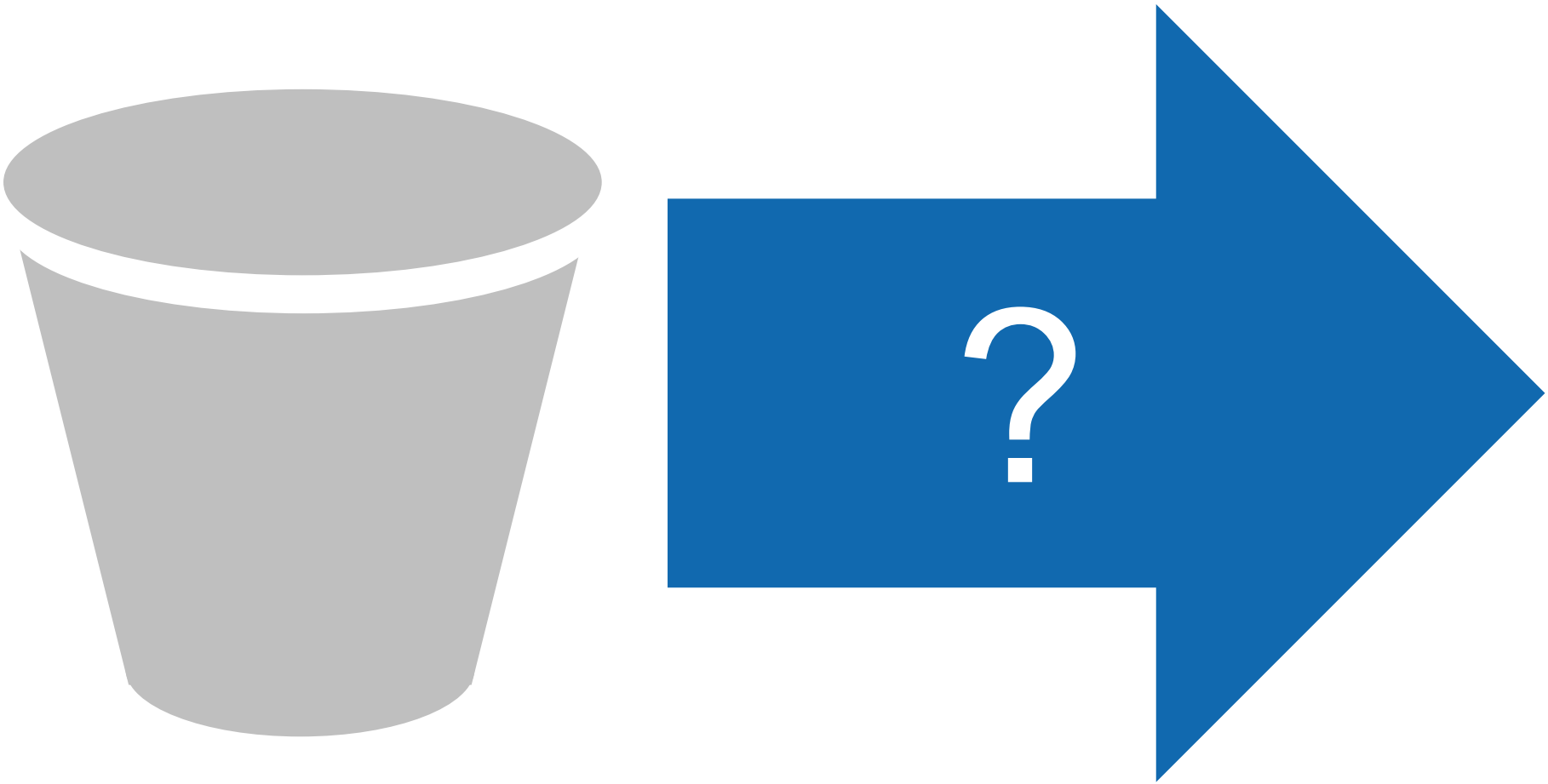
Throughput

Latency

Capacity

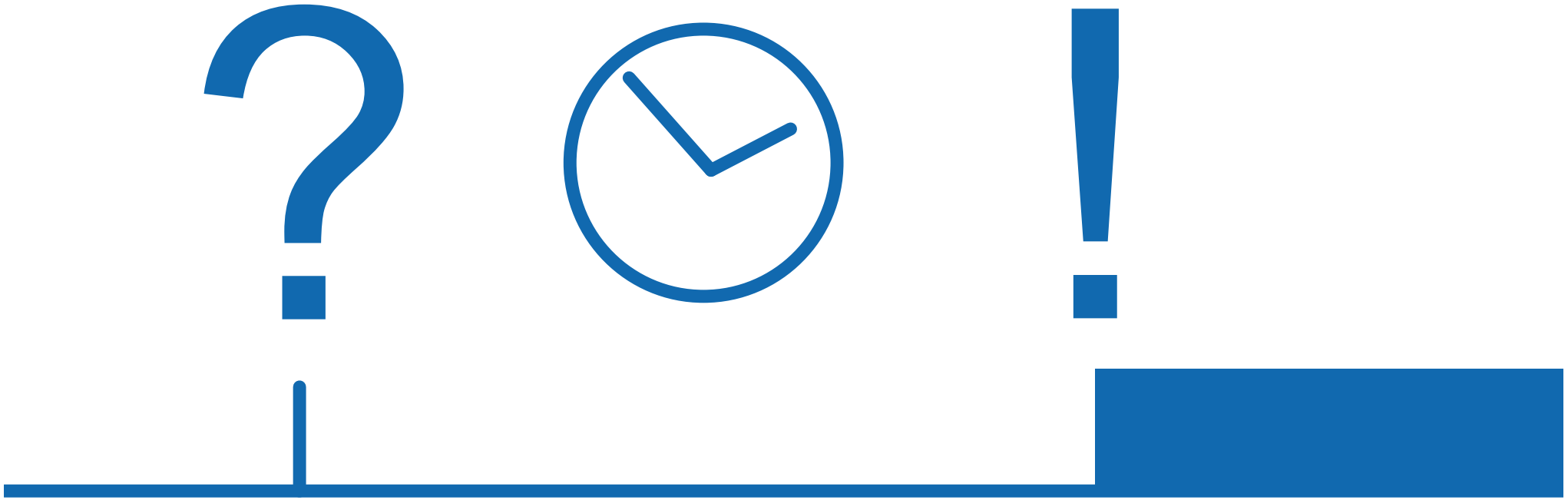
"How much data can we store?"

Throughput



"How fast can we transmit data?"

Latency



"When do I start receiving data?"

1956: IBM RAMAC 350

Computer history museum
computerhistory.org

1956: IBM RAMAC 350

Wikipedia
Deep silence

2020: Western Digital Ultrastar DC HC650

2020: Western Digital Ultrastar DC HC650

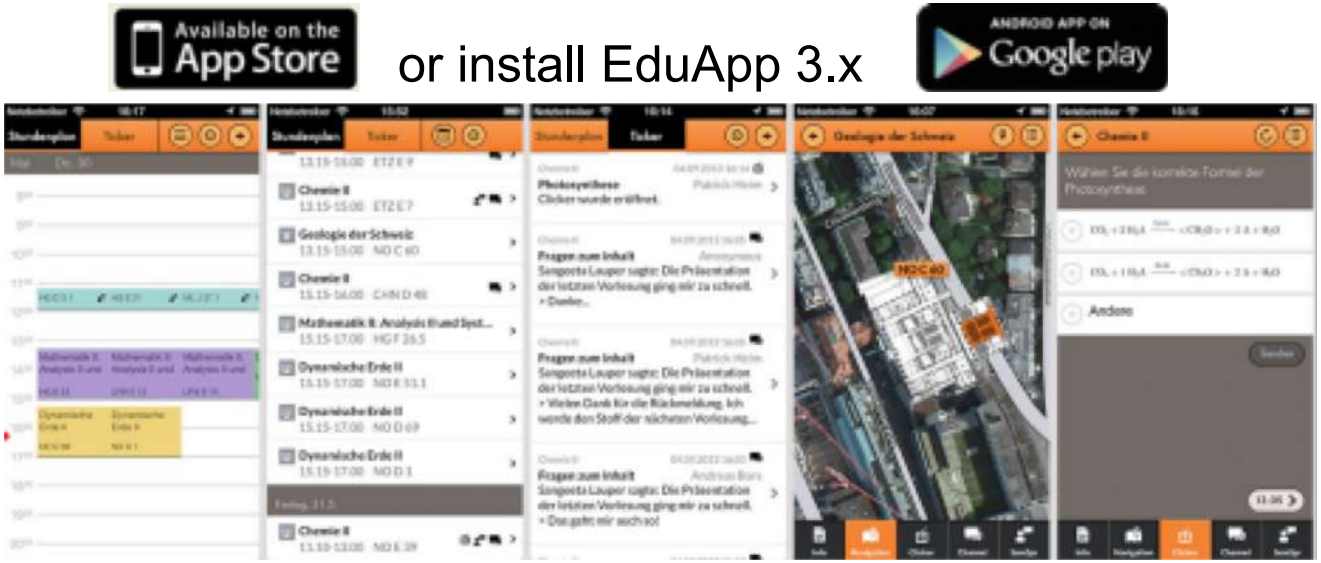
2020: Western Digital Ultrastar DC HC650

2020: Western Digital Ultrastar DC HC650

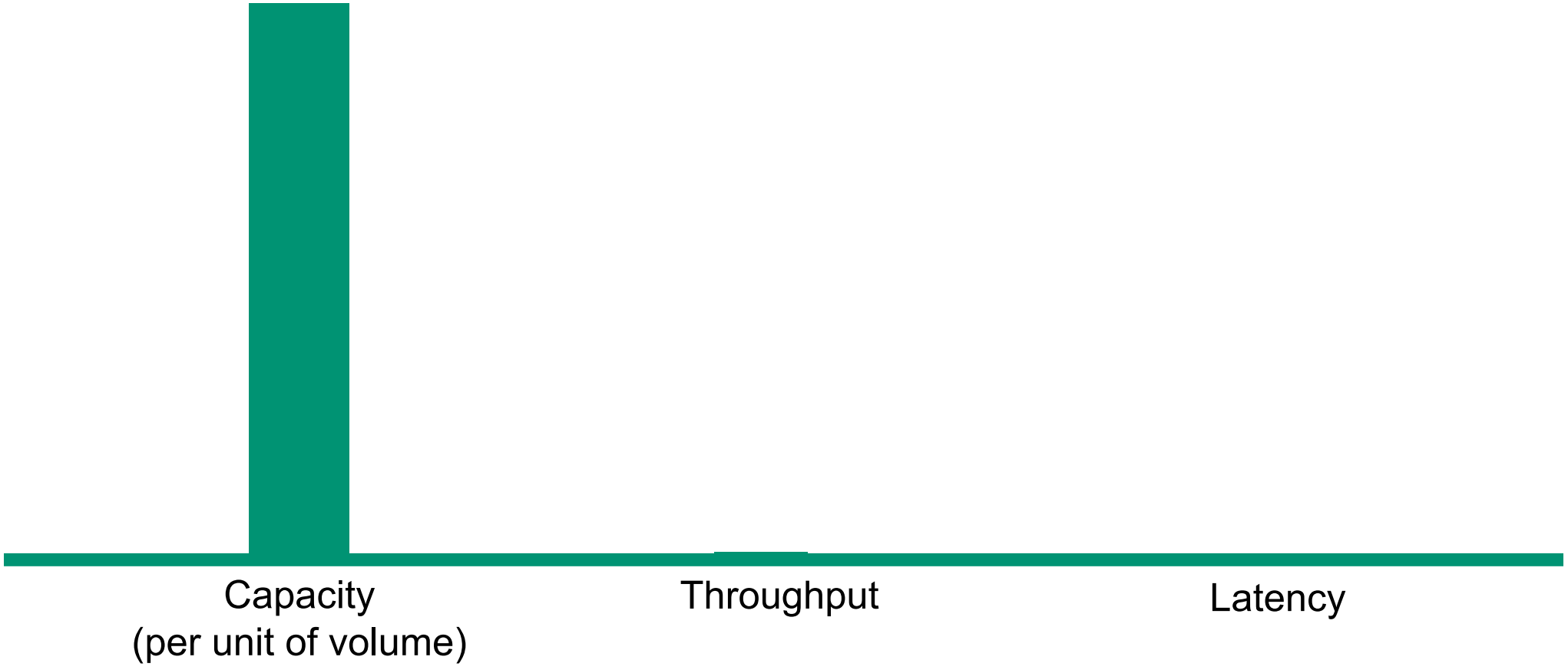
Clicker question

Go *now* to:

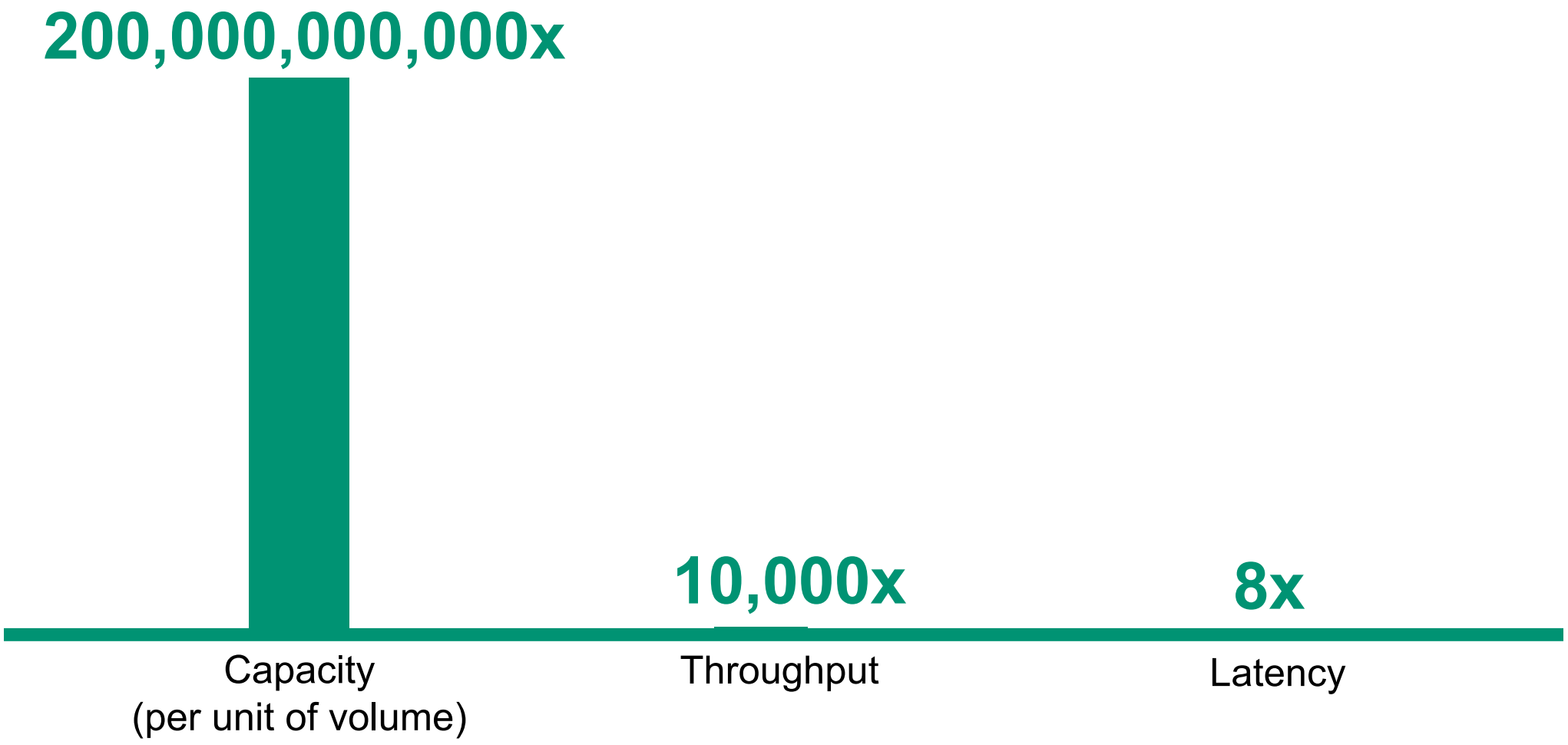
https://eduapp-app1.ethz.ch/



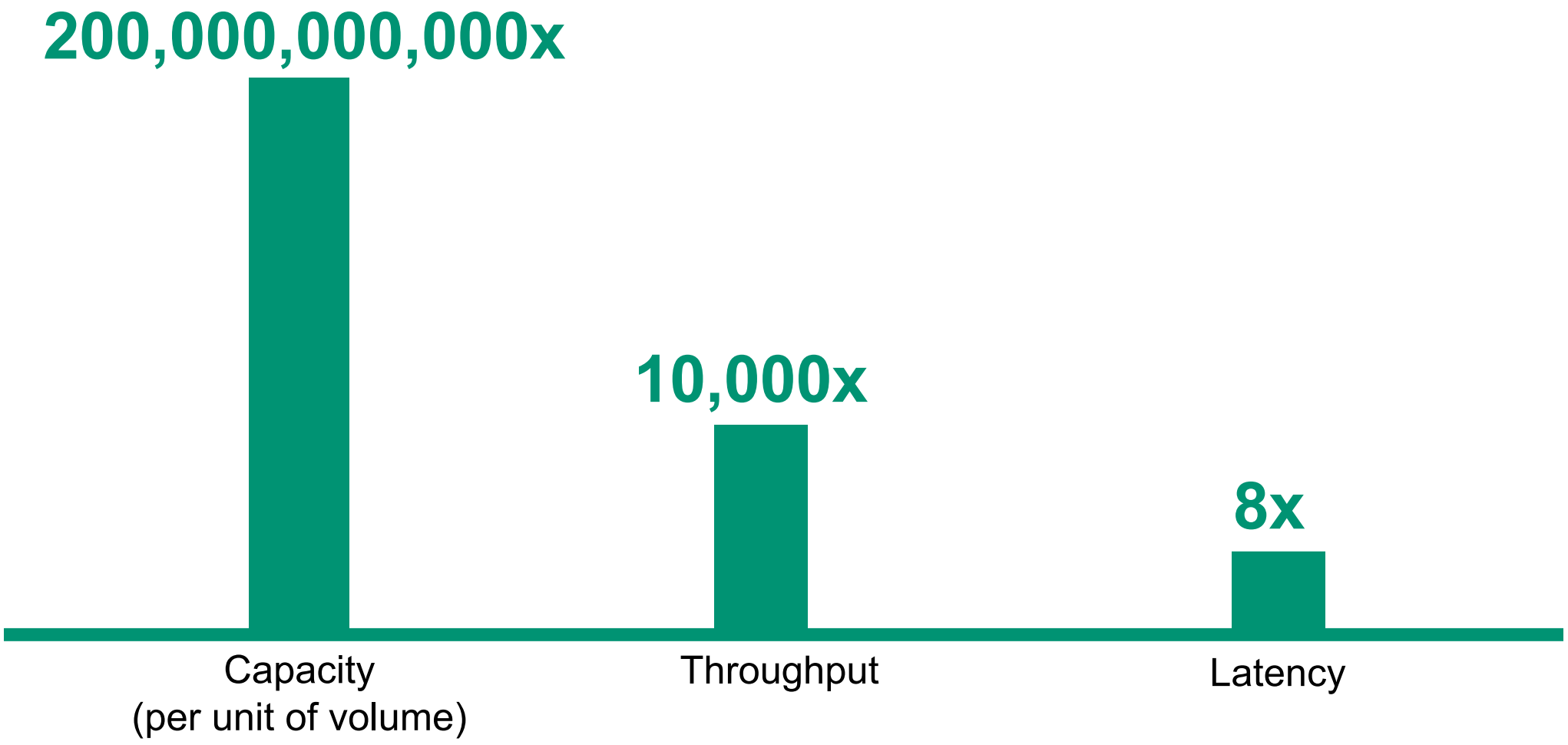
The progress made (1956-2020)



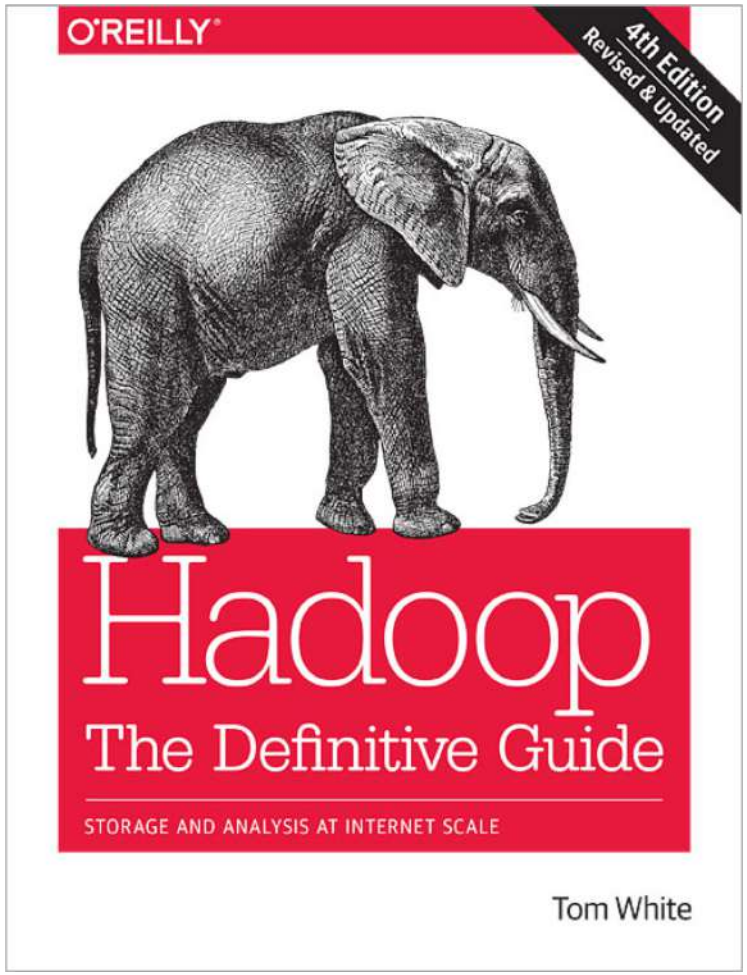
The progress made (1956-2020)



The progress made (1956-2020): Logarithmic

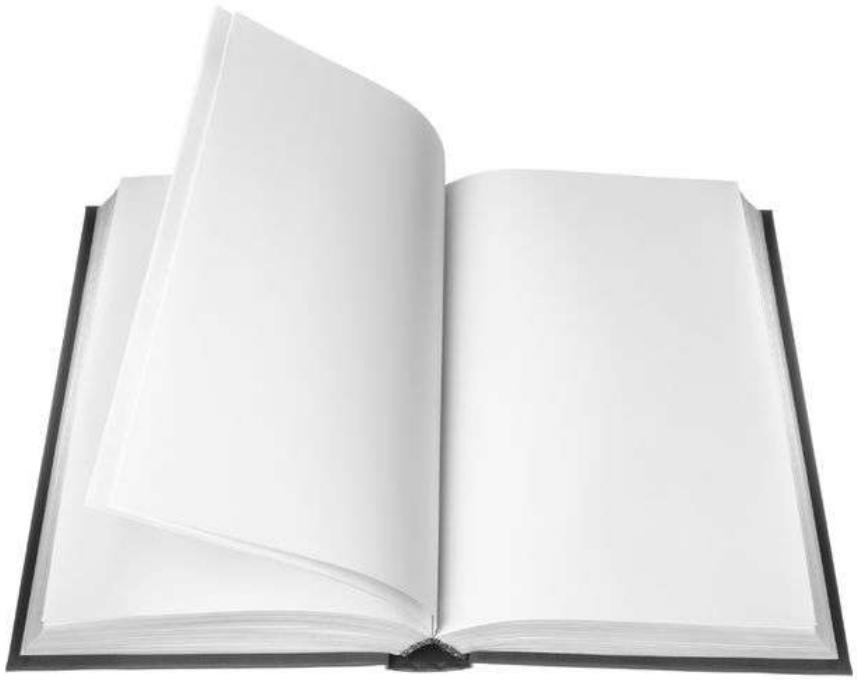


Capacity: Example



~ 600,000 words.

Throughput: Example



~1,000 word per minute

Latency: Example

~ 1 minute to

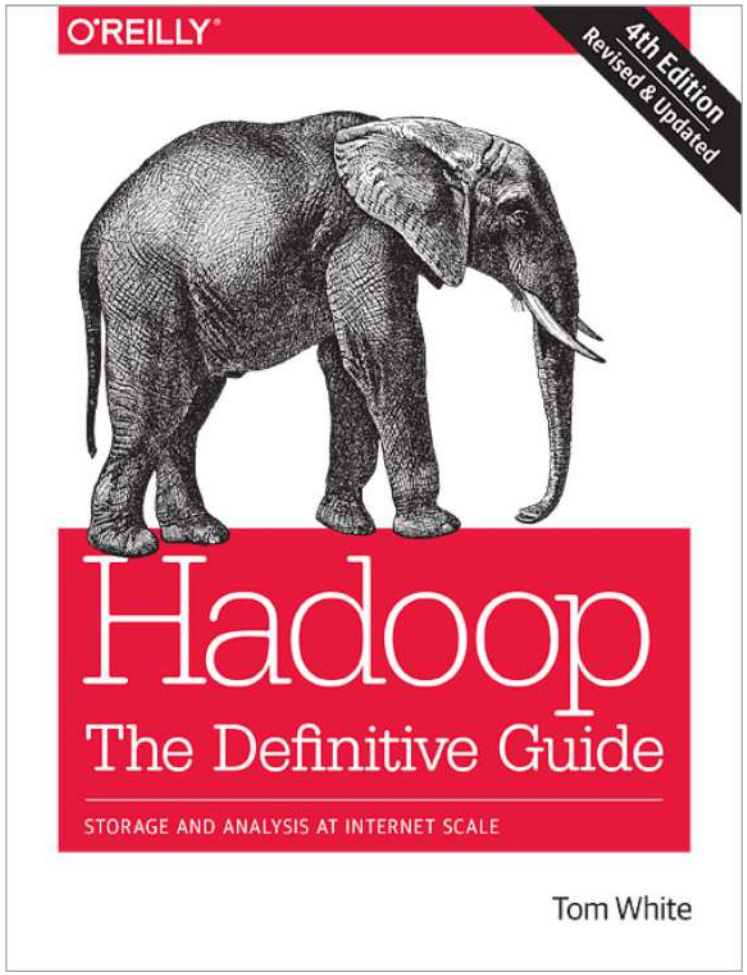
stand up,

go to the shelf,

pick the book,

find the page.

2020 – Analogy with a book

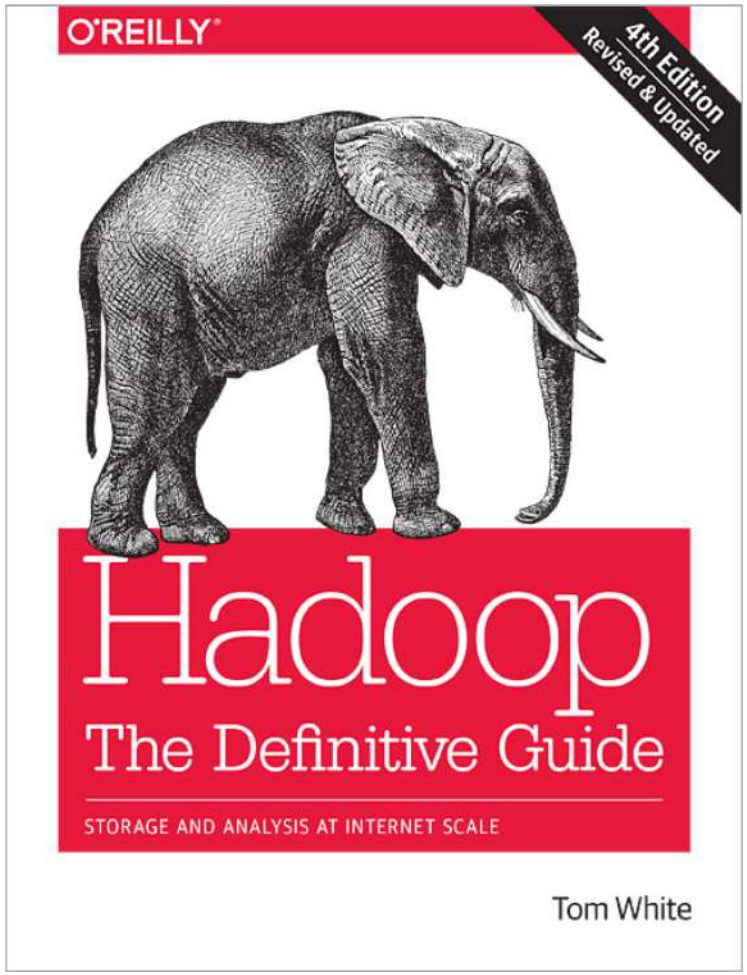


600,000 words

1,000 words per minute

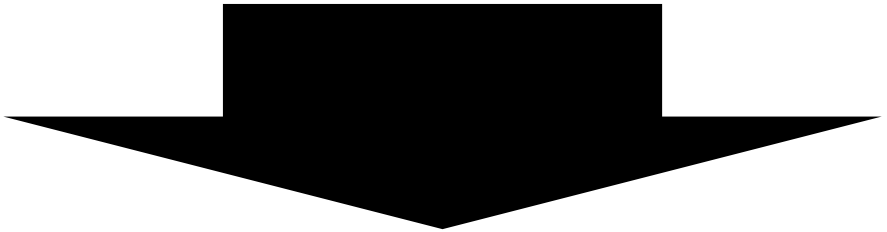
10 hours

2220 – Analogy with a book



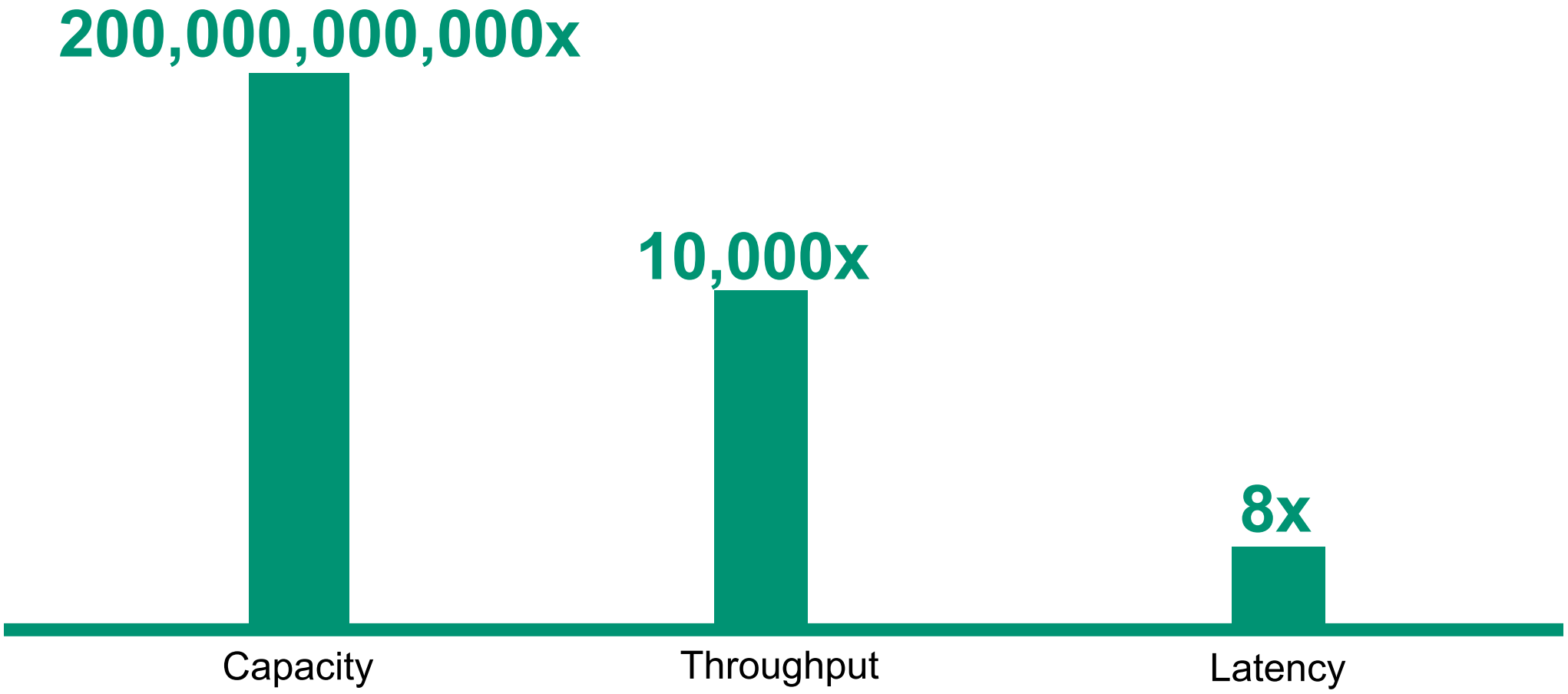
120,000,000,000,000,000 words

10,000,000 words per minute

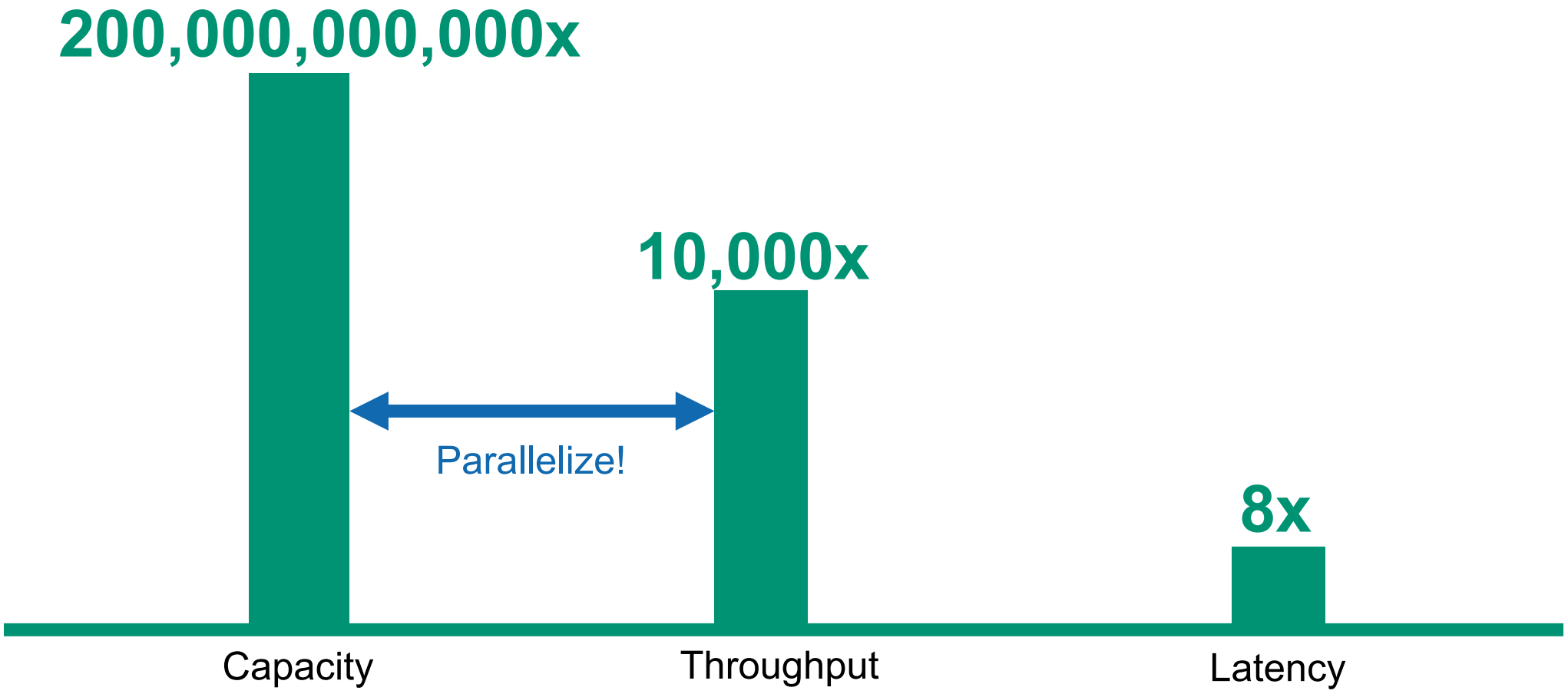


22,800 years

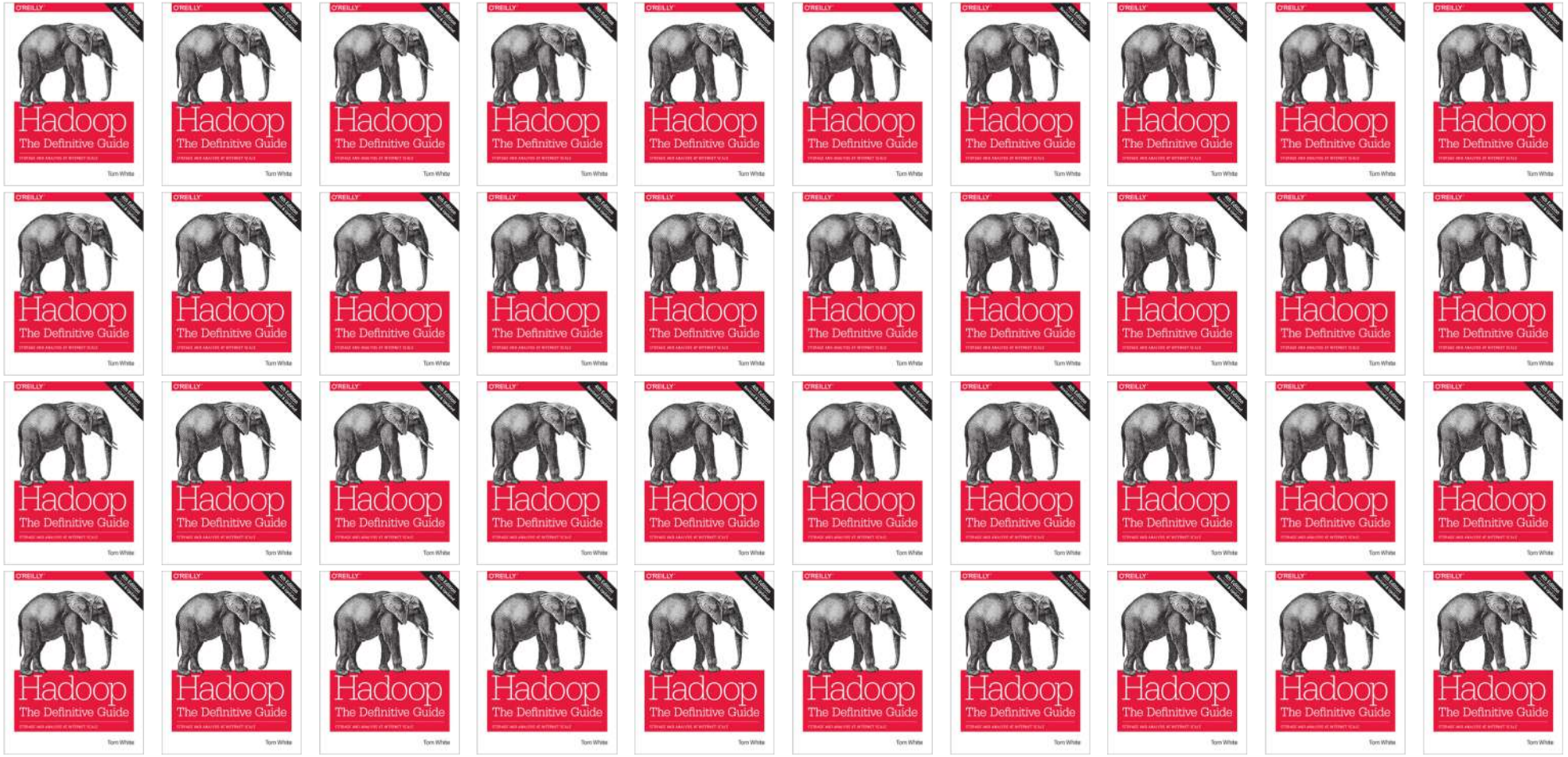
The progress made (1956-2020): Logarithmic



The progress made (1956-2020): Logarithmic



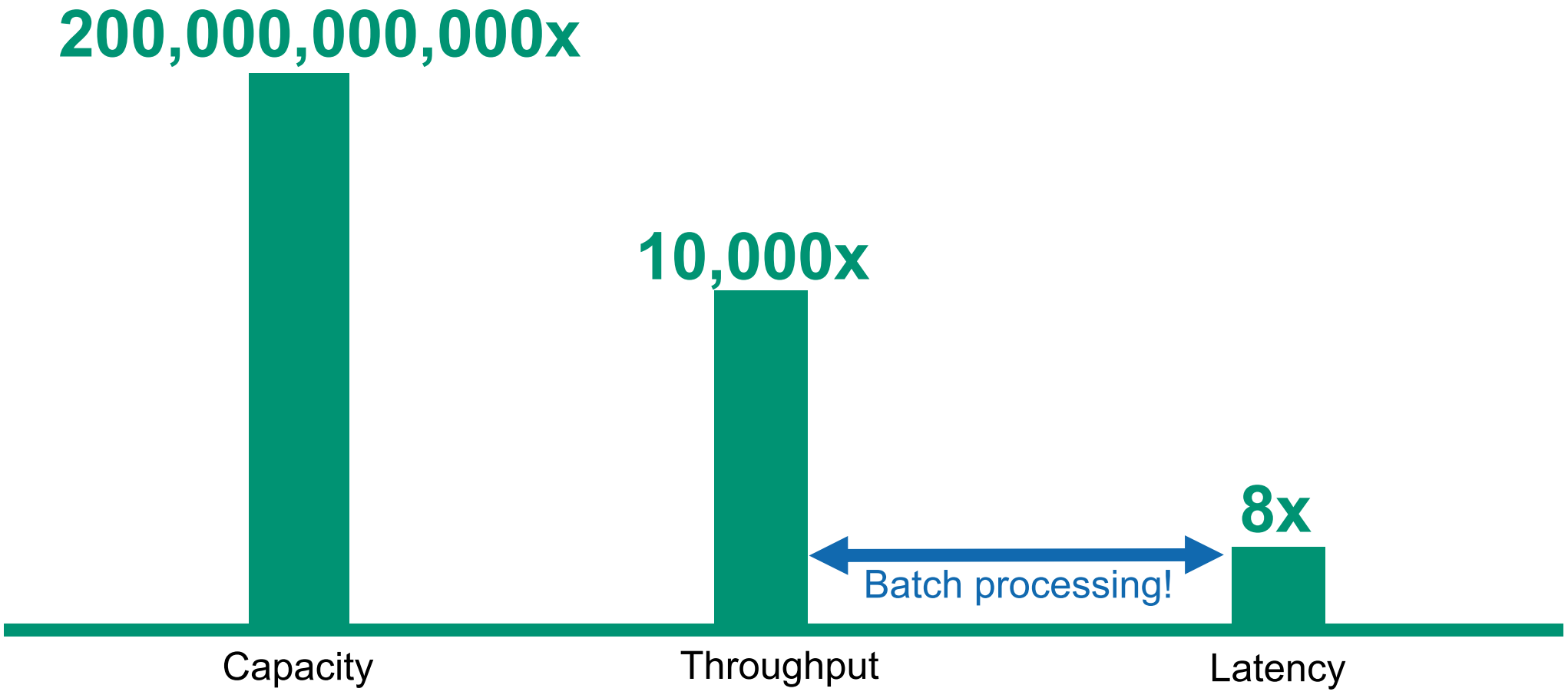
2220: 200,000,000 persons could read it all in 10 hours.



Data centers: clusters of machines (10,000s)

123RF / blueringmedia

The progress made (1956-2020): Logarithmic



What is Big Data (my definition)?

Big Data is a portfolio of technologies that were designed to

store, **manage** and **analyze data** that is too **large** to fit on a single machine

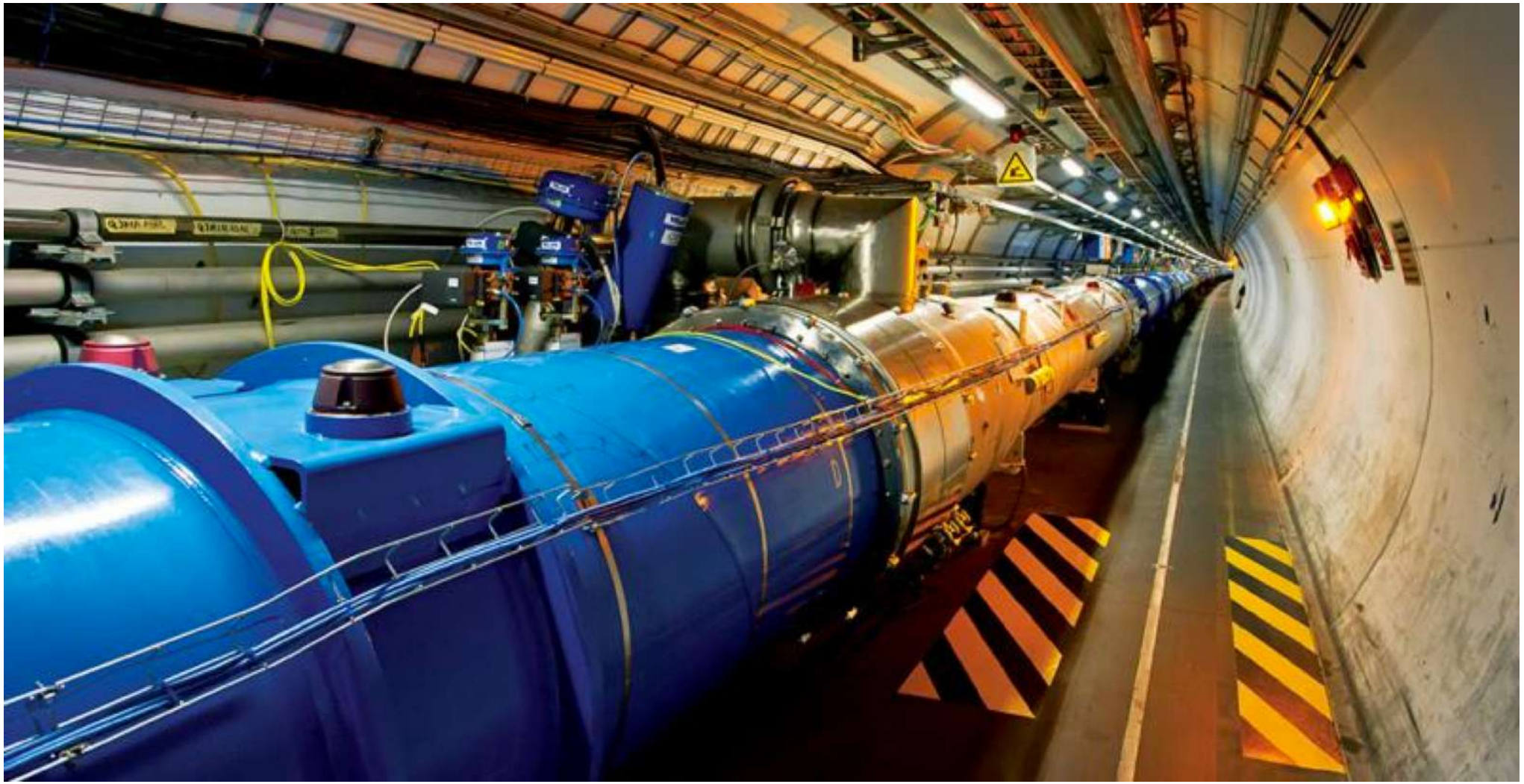
while accommodating for the issue of

growing **discrepancy** between **capacity**, **throughput** and **latency**.

Big Data in the Sciences

Physics: CERN pioneers, produces 30 PB/year

Picture: CERN



Wait!
Actually,
that was
three years
ago!

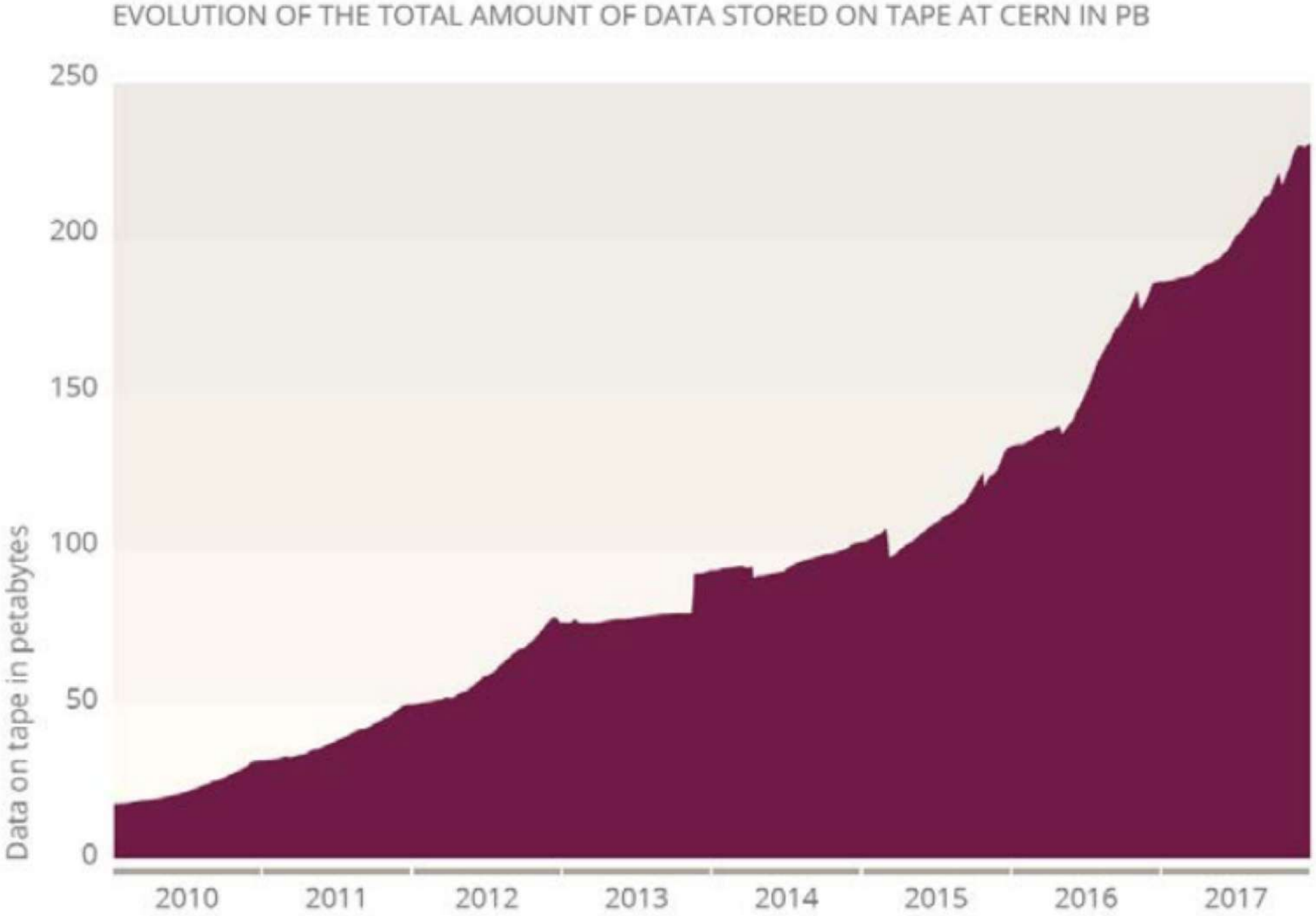
Physics: CERN pioneers, produces 50 PB/year

1,000,000,000 collisions/second

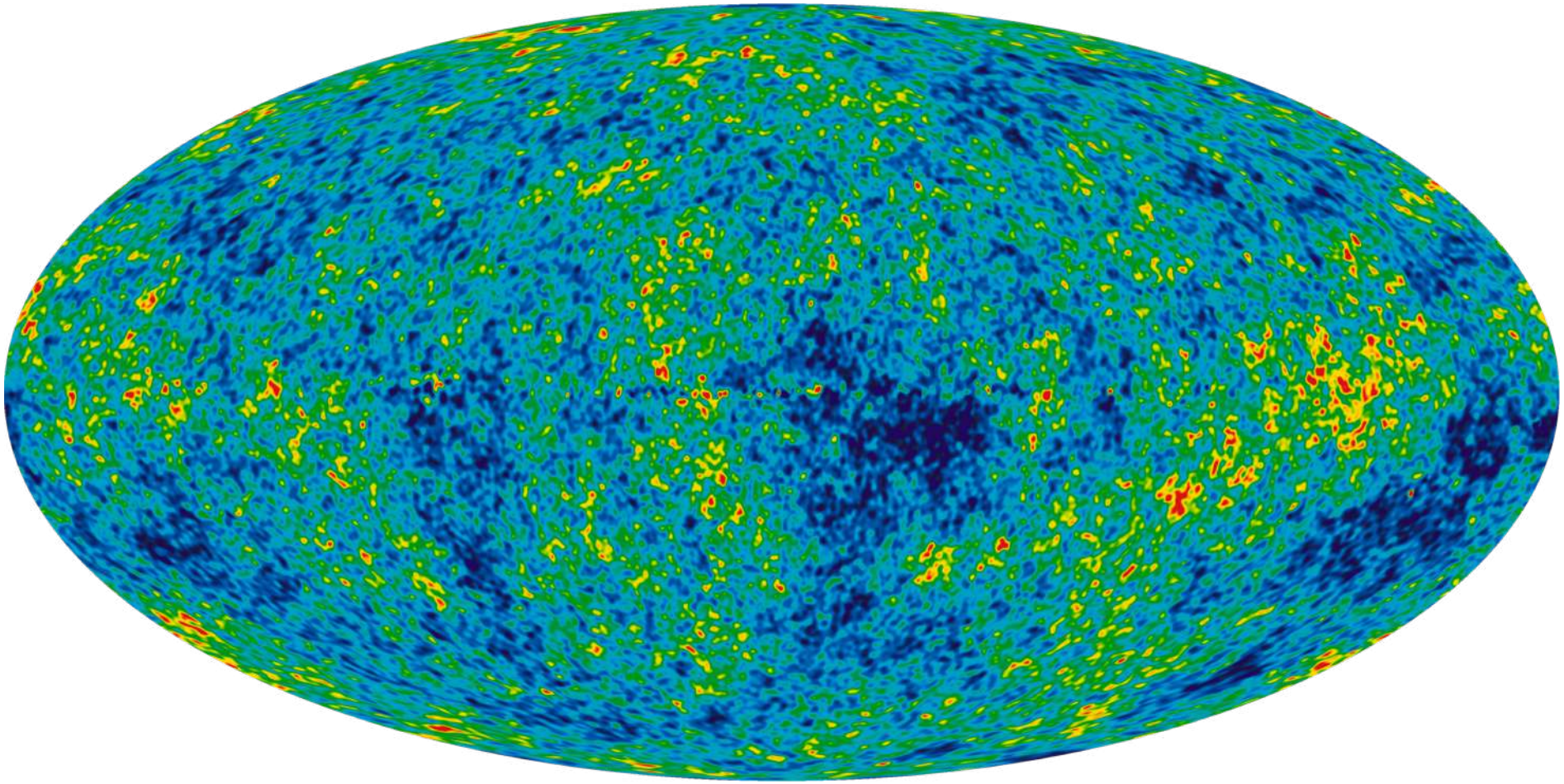
15,000 servers

230,000+ cores

More on http://monit-grafana-open.cern.ch/d/000000884/it-overview?orgId=16


EVOLUTION OF THE TOTAL AMOUNT OF DATA STORED ON TAPE AT CERN IN PB
Data on tape in petabytes
250
200
150
100
50
0
2010
2011
2012
2013
2014
2015
2016
2017

Astronomy: Sloan Digital Sky Survey



Picture: NASA / WMAP Science Team

Astronomy: Sloan Digital Sky Survey

Since 2000, now in phase IV till 2020

The **most detailed**
3D maps of the Universe
ever made

35% covered so far

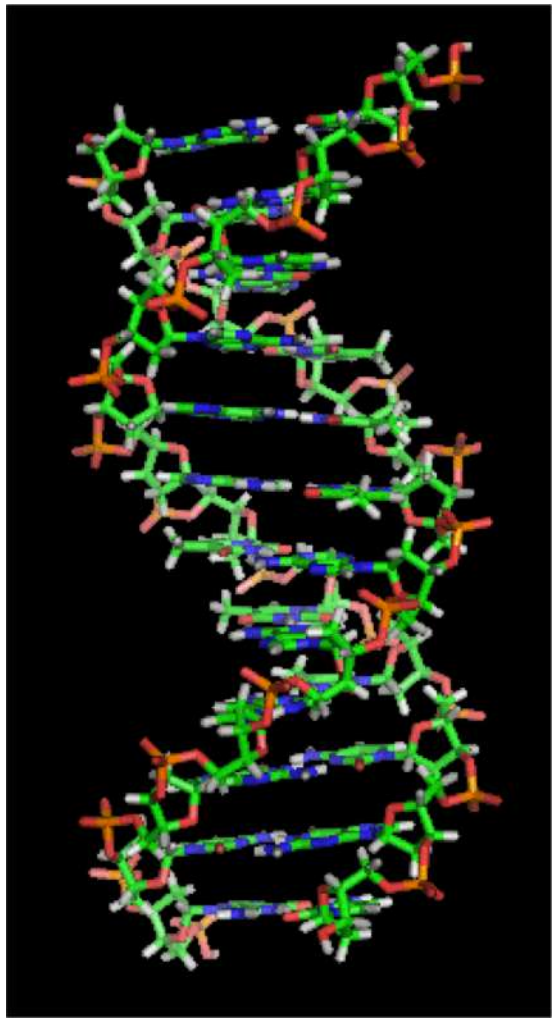
1G objects, 4 spectra

Astronomy: Sloan Digital Sky Survey

200 GB/night

Picture: Wikipedia/EdPost

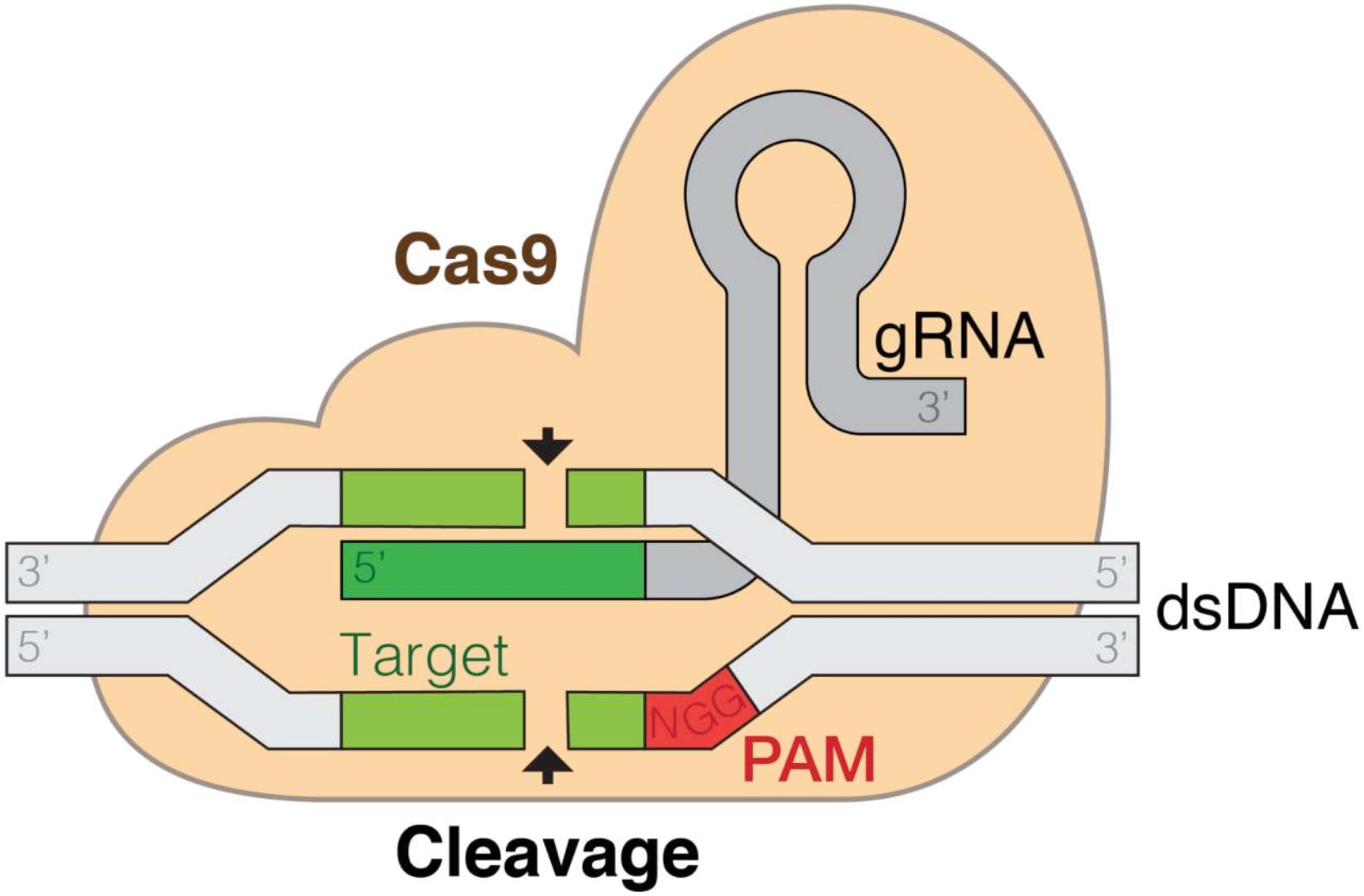
Genomics: the complete human genome



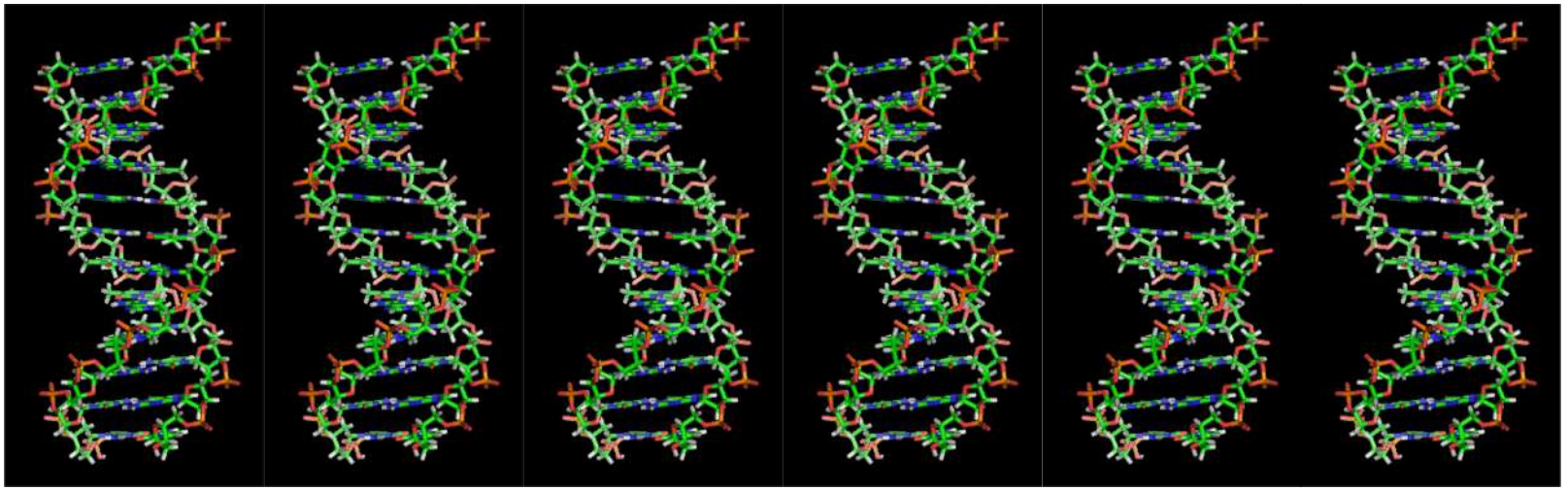
Picture: Wikipedia/Zephyris

3B base pairs

Genomics: CRISPIR-Cas9

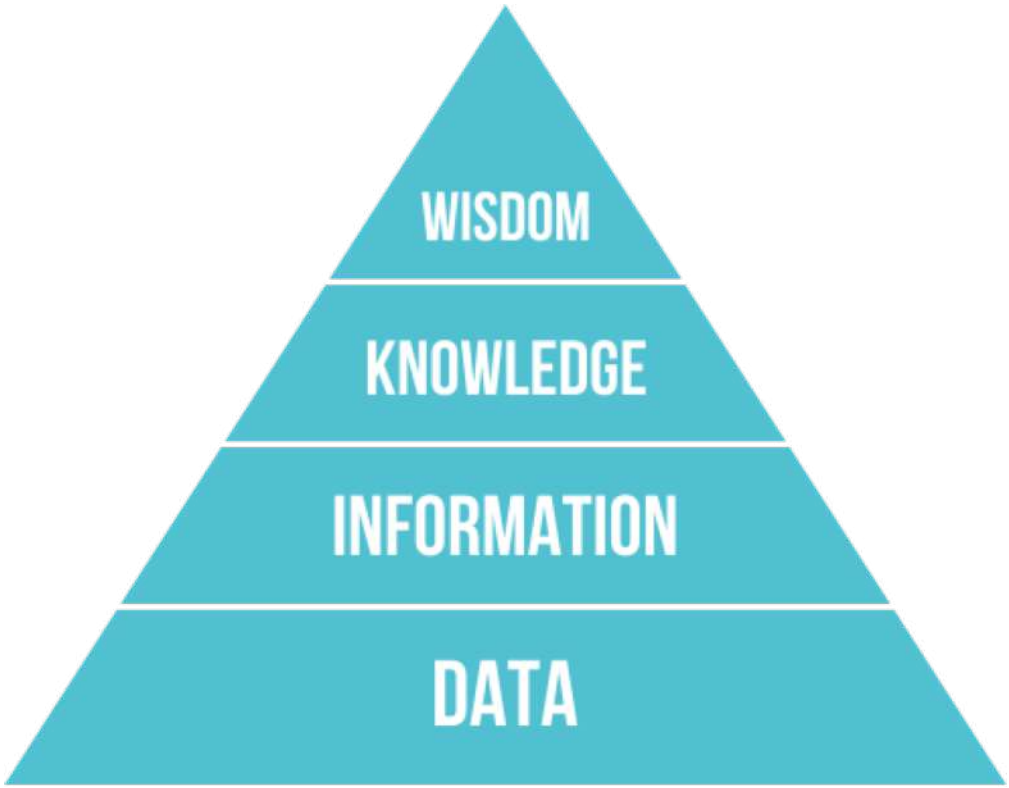


New (2018): DNA as a storage layer

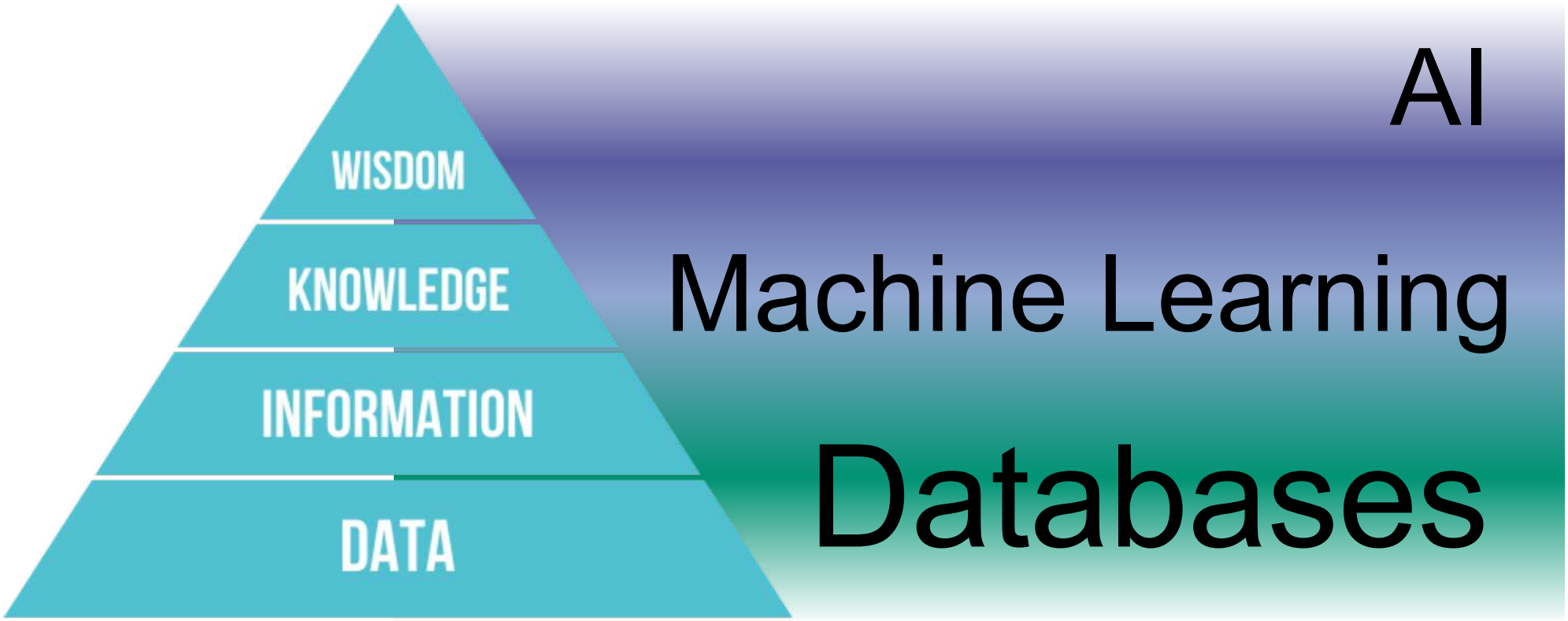


Lecture Scope

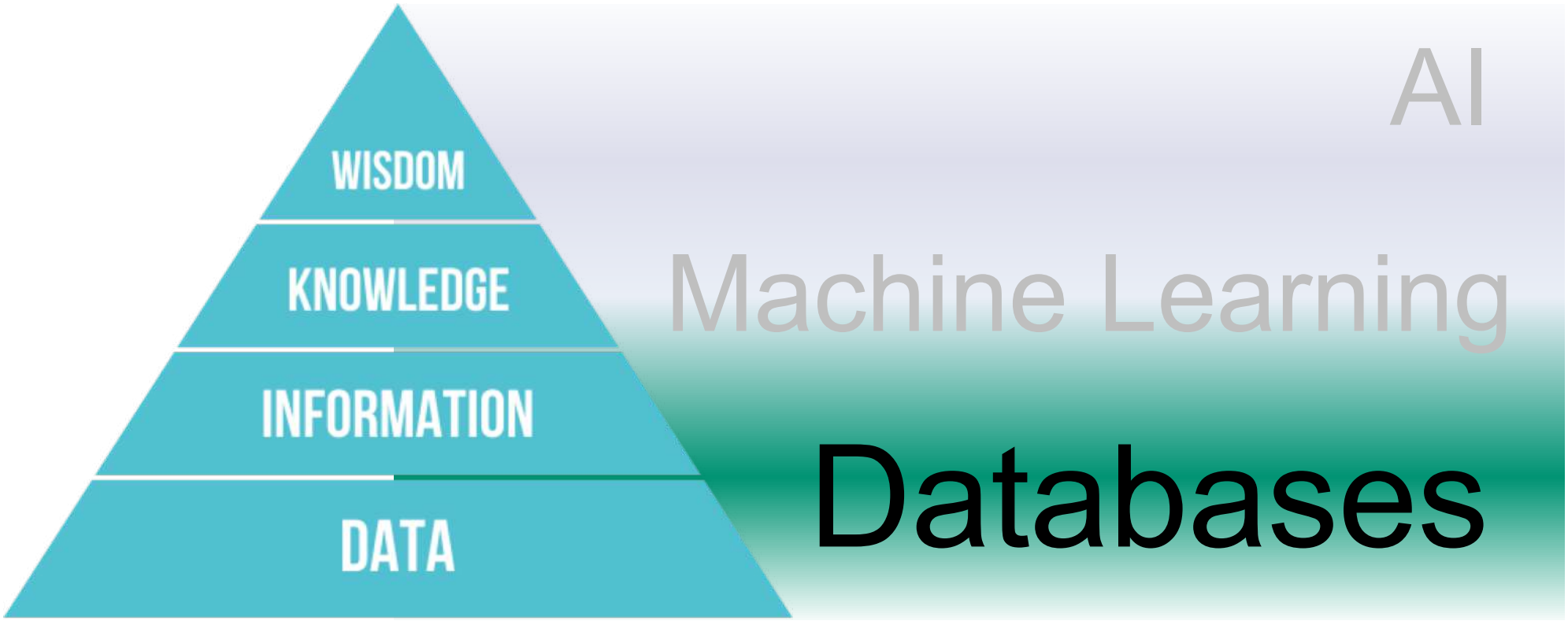
Lecture scope



Lecture scope



Lecture scope: databases only



Lecture Team

Ghislain Fourny

Rodrigo Bruno
(Head TA)

Damien Desfontaines
(TA)

Niels Gleinig
(TA)

Dan Graur
(TA)

Giulia Lanzillotta
(TA)

Lecture Overview

	Concepts	Technologies
Storage	Object storage	S3, Azure Blog Storage
	Distributed file systems	HDFS
	Syntax	XML, JSON
Models	Wide column stores	HBase
	Data models and schemas	XML/JSON Schema
Processing	2-step distributed query processing	Hadoop MapReduce
	Resource management	YARN
	DAG-based distributed query processing	Spark
Management	Document storage	MongoDB
	Query languages	JSONiq

What is expected

Attendance of the **weekly lecture**
(2 hours/w Tuesdays 10-12)

What is expected

Attendance of the **weekly lecture**
(2 hours/w Tuesdays 10-12)

Attendance of the **exercise session**
(2 hours/w Wednesdays/Fridays)

What is expected

Attendance of the **weekly lecture**
(2 hours/w Tuesdays 10-12)

Attendance of the **exercise session**
(2 hours/w Wednesdays/Fridays)

Hands-on self-study, **read the books,**
play with technology (1-2 hours/w)

What is expected

Attendance of the **weekly lecture**
(2 hours/w Tuesdays 10-12)

Attendance of the **exercise session**
(2 hours/w Wednesdays/Fridays)

Hands-on self-study, **read the books,**
play with technology (1-2 hours/w)

Passing the **written exam**
(150 minutes, Summer session)

What is expected

Attendance of the **weekly lecture**
(2 hours/w Tuesdays 10-12)

Attendance of the **exercise session**
(2 hours/w Wednesdays/Fridays)

Hands-on self-study, **read the books,**
play with technology (1-2 hours/w)

Passing the **written exam**
(150 minutes, Summer session)

Bonus points!

Bonus points!

Bonus points!

You SHOULD solve the weekly exercise sheets (theoretical, practical)

Bonus points!

You SHOULD solve the weekly exercise sheets (theoretical, practical)

We will grade the exercises marked as such (25 of them)

Bonus points!

You SHOULD solve the weekly exercise sheets (theoretical, practical)

We will grade the exercises marked as such (25 of them)

You get 0.01 extra point per passed assignment: at most 0.25

Bonus points!

You SHOULD solve the weekly exercise sheets (theoretical, practical)

We will grade the exercises marked as such (25 of them)

You get 0.01 extra point per passed assignment: at most 0.25

"0.n" will thus be added to your exam grade before rounding

Self-study: Azure