DASCI

Instituto Andaluz Interuniversitario
en Ciencia de Datos e
Inteligencia Computacional

# Ciencia de Datos a través del Big Data

Diego García (djgarcia@ugr.es)
Isaac Triguero (isaaktriguero@ugr.es)

Financiado por
la Unión Europea
NextGenerationEU

GOBIERNO
DE ESPAÑA

MINISTERIO
PARA LA TRANSFORMACIÓN DIGITAL
Y DE LA FUNCIÓN PÚBLICA

SECRETARÍA DE ESTADO
DE DIGITALIZACIÓN
E INTELIGENCIA ARTIFICIAL

Plan de
Recuperación,
Transformación
y Resiliencia

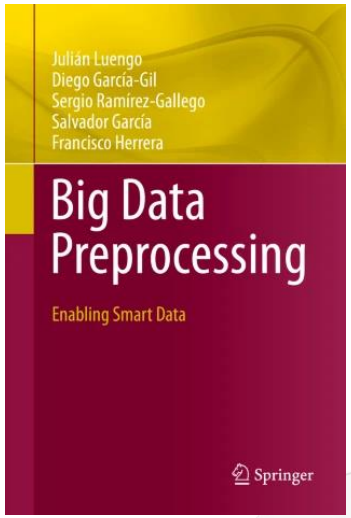
UNIVERSIDAD
DE GRANADA

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

REPSOL

# Introducción: quiénes somos

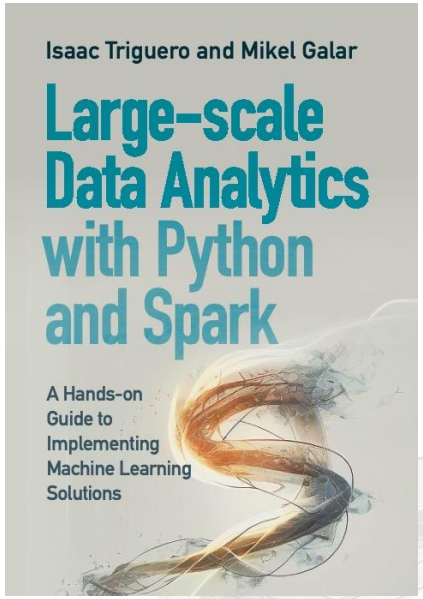
- **Diego García** (djgarcia@ugr.es)
- Profesor Ayudante Doctor del departamento de Lenguajes y Sistemas Informáticos (LSI)
- Investigador en Big Data
  - Ensembles
  - Preprocesamiento de datos
  - Detección de anomalías
- Autor del libro "Big Data Preprocessing - Enabling Smart Data"




https://scholar.google.com/citations?user=cYtfjosAAAAJ

# Introducción: quiénes somos

- **Isaac Triguero** (isaaktriguero@ugr.es)
- Investigador Senior Distinguido (María Zambrano)
- Investigación en Big Data
  - Preprocesamiento de datos
  - Aplicaciones en Energía, Transporte, Medicina
- +10 años impartiendo cursos de Big Data
  - Autor de un libro docente en la temática
- Colaboraciones con empresas en Big Data
  - Unilever
  - E.ON



# Descripción del curso

- Este curso ofrece una inmersión en el mundo del Big Data, desde los conceptos fundamentales hasta las últimas herramientas de análisis.
- Exploraréis los principios esenciales de Big Data, aprendiendo a gestionar grandes volúmenes de datos.
- Además, nos sumergiremos en el ecosistema de Hadoop, dominando el arte del procesamiento y almacenamiento distribuido en el Hadoop Distributed File System (HDFS).
- Una vez establecidas las bases, el curso se adentrará en Apache Spark, ofreciendo una visión detallada de esta potente y popular plataforma de procesamiento en memoria y su papel fundamental en el análisis de Big Data.
- Desde la manipulación de datos hasta aplicaciones reales, adquiriréis habilidades prácticas para abordar los desafíos más complejos en el ámbito del análisis de datos.

# Programa y evaluación

- Semana 1:
  - Introducción al Big Data
  - Fundamentos de MapReduce
  - Fundamentos de Hadoop
- Semana 2:
  - Programación de altas prestaciones con Spark
  - Spark SQL – sesión práctica
  - Aprendizaje automático con Spark

- Al final de cada bloque se realizarán cuestionarios tipo test

# ¿Conocimientos previos?

- Infraestructuras
  - Sistemas de ficheros
  - HPC
- Programación
  - Python
  - Estructuras de datos como DataFrames
  - Pipelines
  - Programación en paralelo
- Aprendizaje automático
  - Tipos de aprendizaje (supervisado vs no supervisado)
  - Algoritmos clásicos (árboles de decision, k-means)
  - Métricas de evaluación

CAMBRIDGE UNIVERSITY PRESS

# Chapter 1

# Introduction to big data

# Learning Outcomes

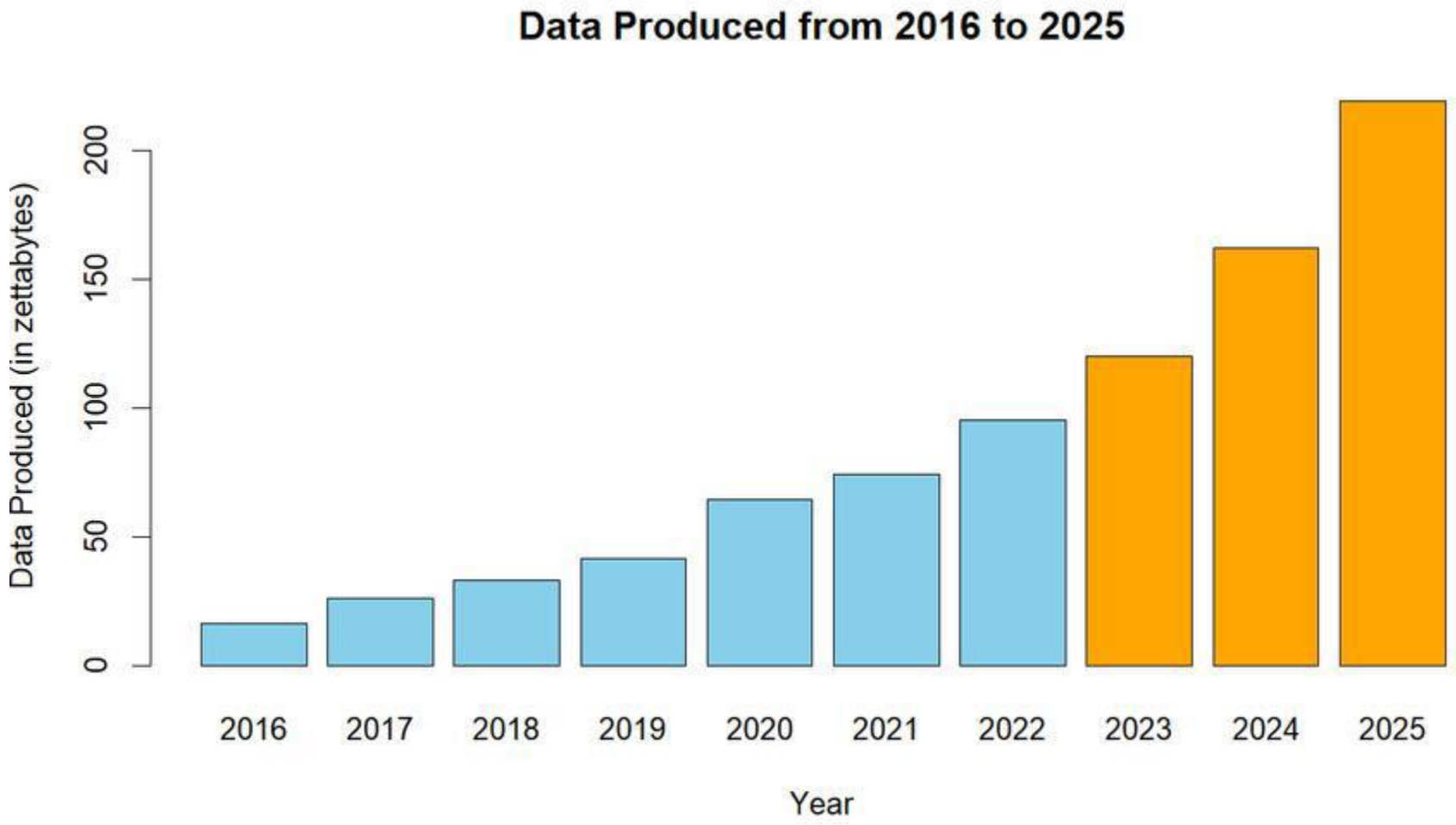
- Learning the principles of Big Data and its importance [KU]
- Understanding that scale-out is the way to address Big Data problems [KU]
- Understanding that Big Data processing requires data locality and a new big data computing paradigm [KU]

# Outline

- What is Big Data?
- Big data evolution and definition
- How to deal with big data?
  - Scale-up vs. scale-out
  - Traditional Distributed Computing vs Big Data Computing
  - The principle of Data Locality

# What is Big Data?

- Tons of data!



# What is Big Data?

There is not a standard definition!

"**Big Data**" involves data whose volume, diversity and complexity **requires new techniques, algorithms and analyses** to extract valuable (hidden) knowledge

**Data-Intensive applications**

**Large-Scale Data Processing**

# Example: Target

- Target registers everything their clients buy
- They started looking for patterns of known pregnant women
  - Lotions! Pregnant women buy unscented lotion around the beginning of their second trimester
  - They identified 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score
  - They could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy
- But there is a problem…

# Example: Google Flu - 2009

- 2009, new strand of the flu - H1N1 – They were scared of a pandemic
- Medical Data is slow to collect (2 weeks delay)
- They established a correlation of keywords people were using in their search to identify it!
  - There were >3,000 million searches every day
  - They compared most common 50 million search terms against a database of flu propagation from 2003 to 2008
  - They found 45 terms that provided a high correlation between real data from Statistical Centres and Search, but in real time!

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature*, **457**(7232), 1012–1014. Link

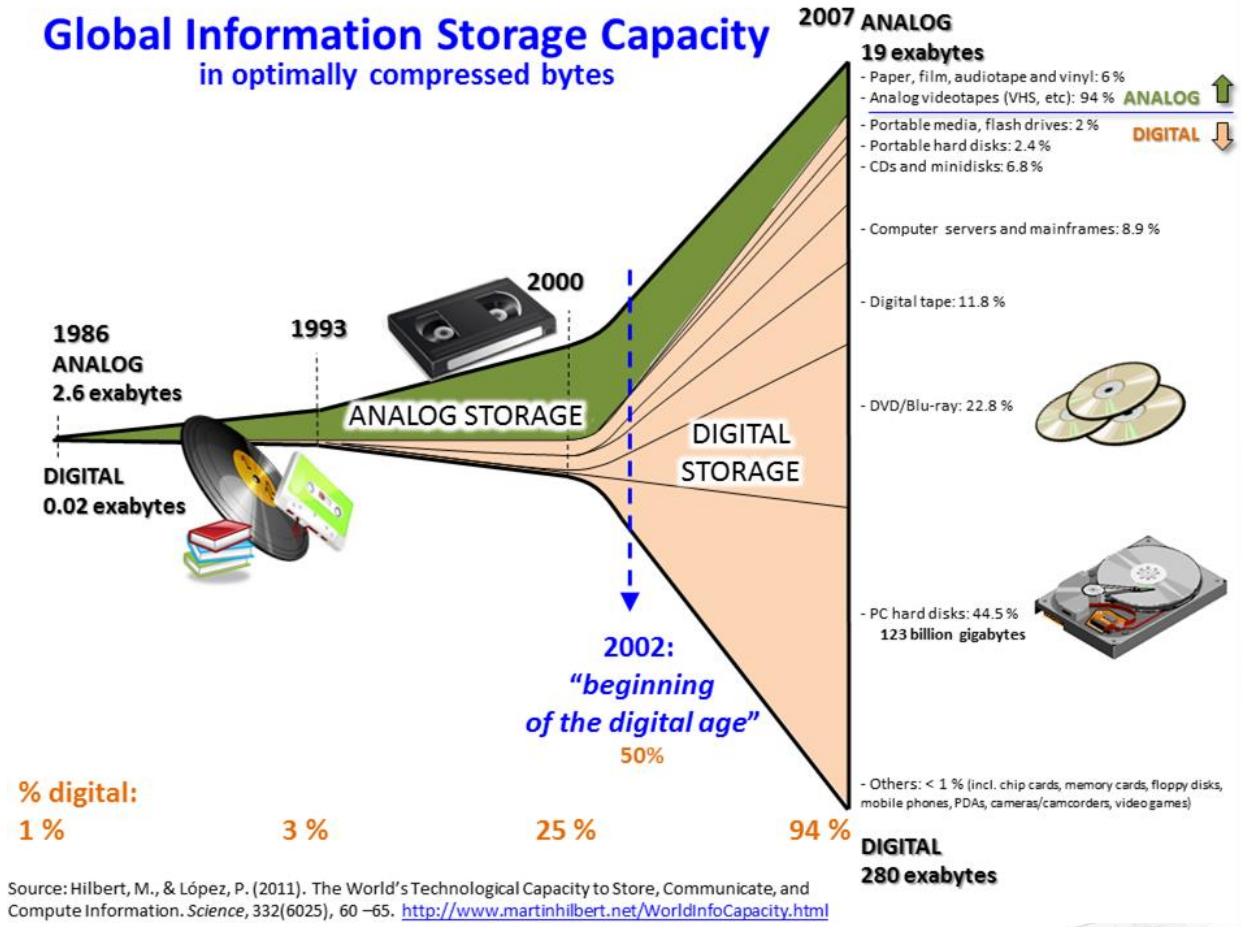
# Example: Google Flu - 2009

- But why is this Big Data?
  - They needed to parallelize their data analytics!
  - They used Hadoop MapReduce
- From success to failure:
  - Google Flu Trends (GFTs) later overestimated the flu levels in 2013
  - Changes in people behavior
  - Models need updating with new data (data streams!)
  - GFT stopped in 2014

Lazer, D., Kennedy, R., King, G., and Vespignani, A. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, **343**(14 March), 1203–1205. Link

# Data Evolution

- Our world revolves around data
- Who generates that data?
  - Science, Business, Entertaining
  - Social media, Industry, Energy,…
- At the beginning of 2020, the digital universe was estimated to consist of 44 zettabytes of data (seedscientific.com)
- But is all about volume?

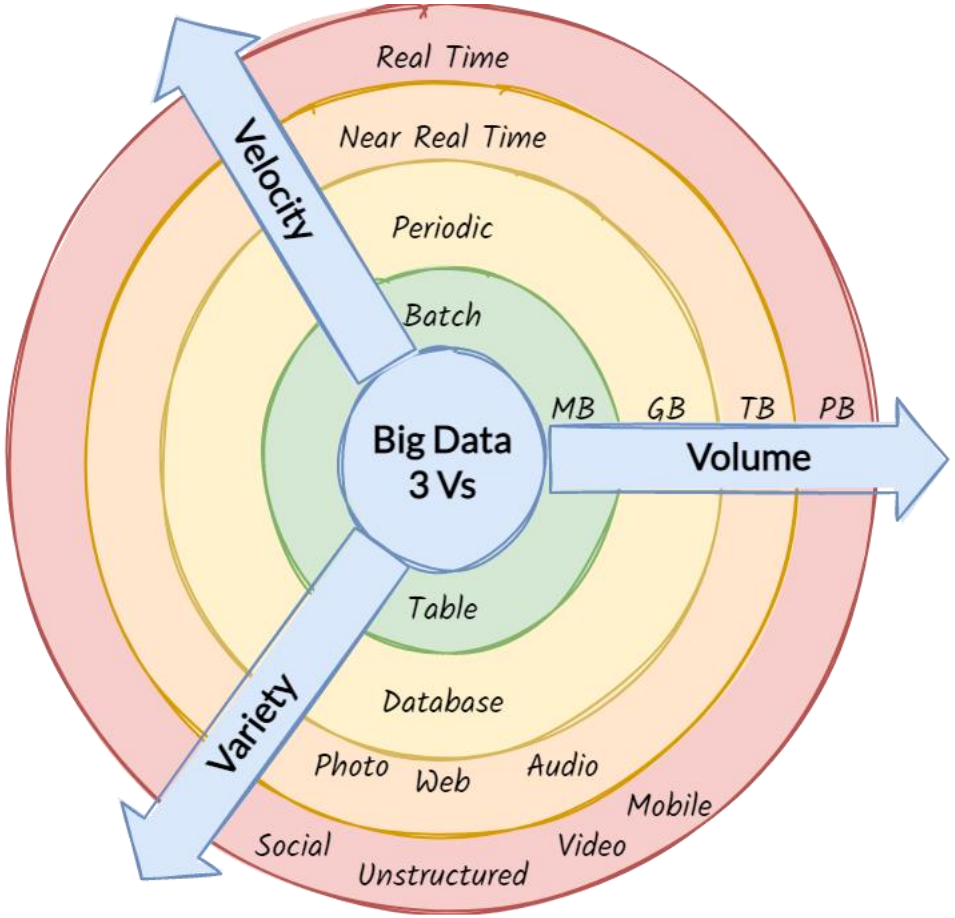



Hilbert, M. and López, P. 2011. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, **332**(6025), 60 –65. Link to paper and Link to webpage

# Data Evolution

- The Large Hadron Collider (LHC) experiments generate about 600 petabytes of data per year
- Twitter: 6,000 tweets every second
- YouTube: 500 hours of video uploaded every minute
- Spotify: 60,000 tracks uploaded everyday
- More than 300 billion emails are sent each day
- And this is just the private companies!
- The Sentinel 2 satellite, put into orbit by the European Space Agency, scans the whole world every five days, generating one terabyte of data every day

# What is Big Data? The 3V's definition




Laney, Douglas. 2001 (February). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rept. META Group.

# What is Big Data? The 3V's definition

**Volume: data at rest**

- Vast amounts of data generated every second
- Data sets are becoming too large to store using traditional database technology
- Big data technology stores these data sets using distributed systems

# What is Big Data? The 3V's definition



**Velocity: data in motion**

- Speed at which:
  - Data is generated
  - Data needs to be analysed.
- Continuous data streams are being captured (e.g. from sensors or mobile devices) and produced
- Late decisions imply missed opportunities

# What is Big Data? The 3V's definition



**Variety: data in many forms**

- One application may generate/use several formats and structures:
  - **Structured data** (we know the schema)
    - Tables, relational databases
  - **Semi-Structured data** (we can infer the schema)
    - XML files, Tagged text
  - **Unstructured data** (we don't know the schema)
    - Text, images, audio, video

# What is Big Data? The 5V's definition

CAMBRIDGE UNIVERSITY PRESS

**Veracity: data in doubt**

- Uncertainty about the quality of the data.
  - E.g., natural language processing on social media: typos, abbreviations, colloquial speech, sarcasm.
- Data may be missing, ambiguous, or even completely wrong.

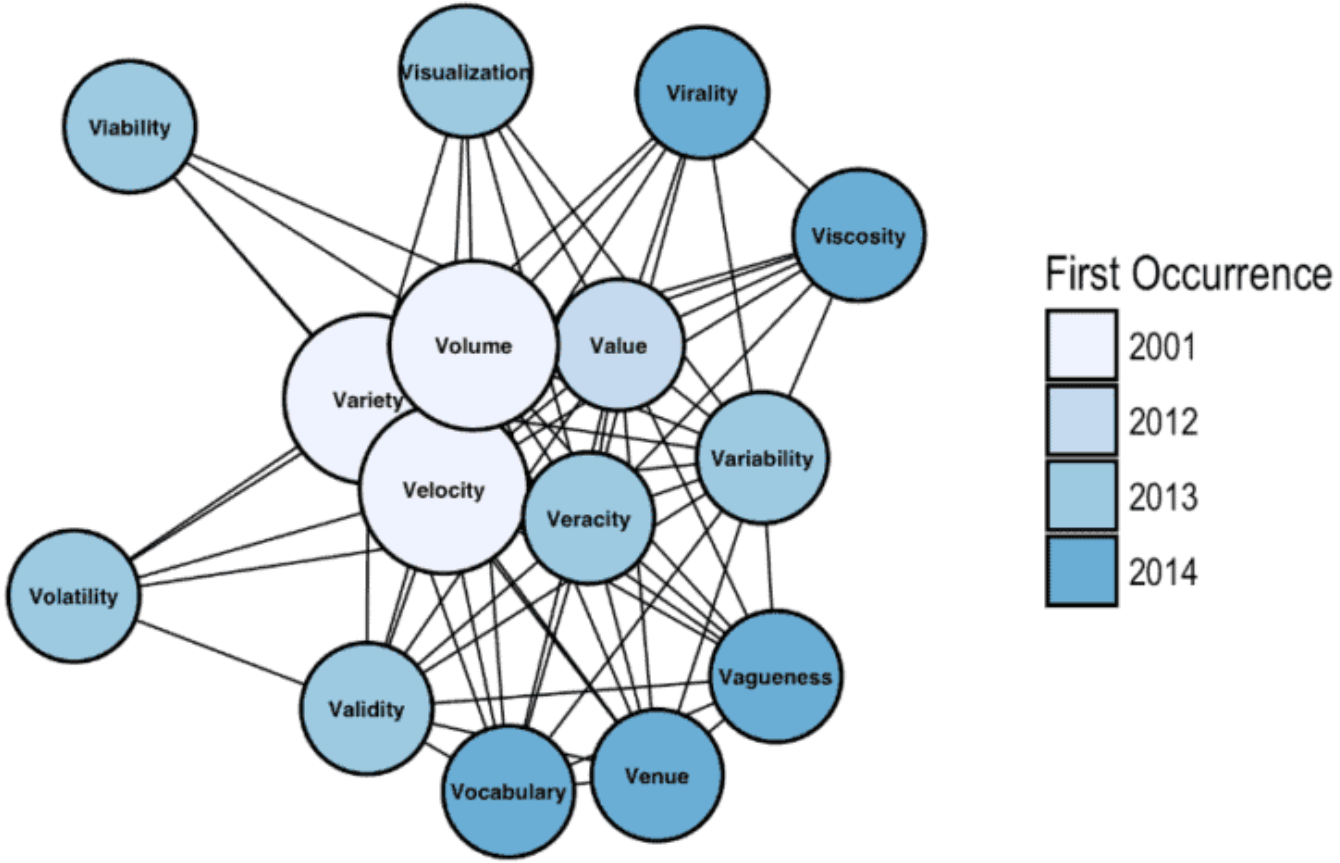
# What is Big Data? The 5V's definition

**Value: data in use**

- Most important motivation for big data
- Big data may (*more like should*) result in:
  - Better statistics/models
  - Novel insights
  - New opportunities for research and industry

# 42 V's of Big Data

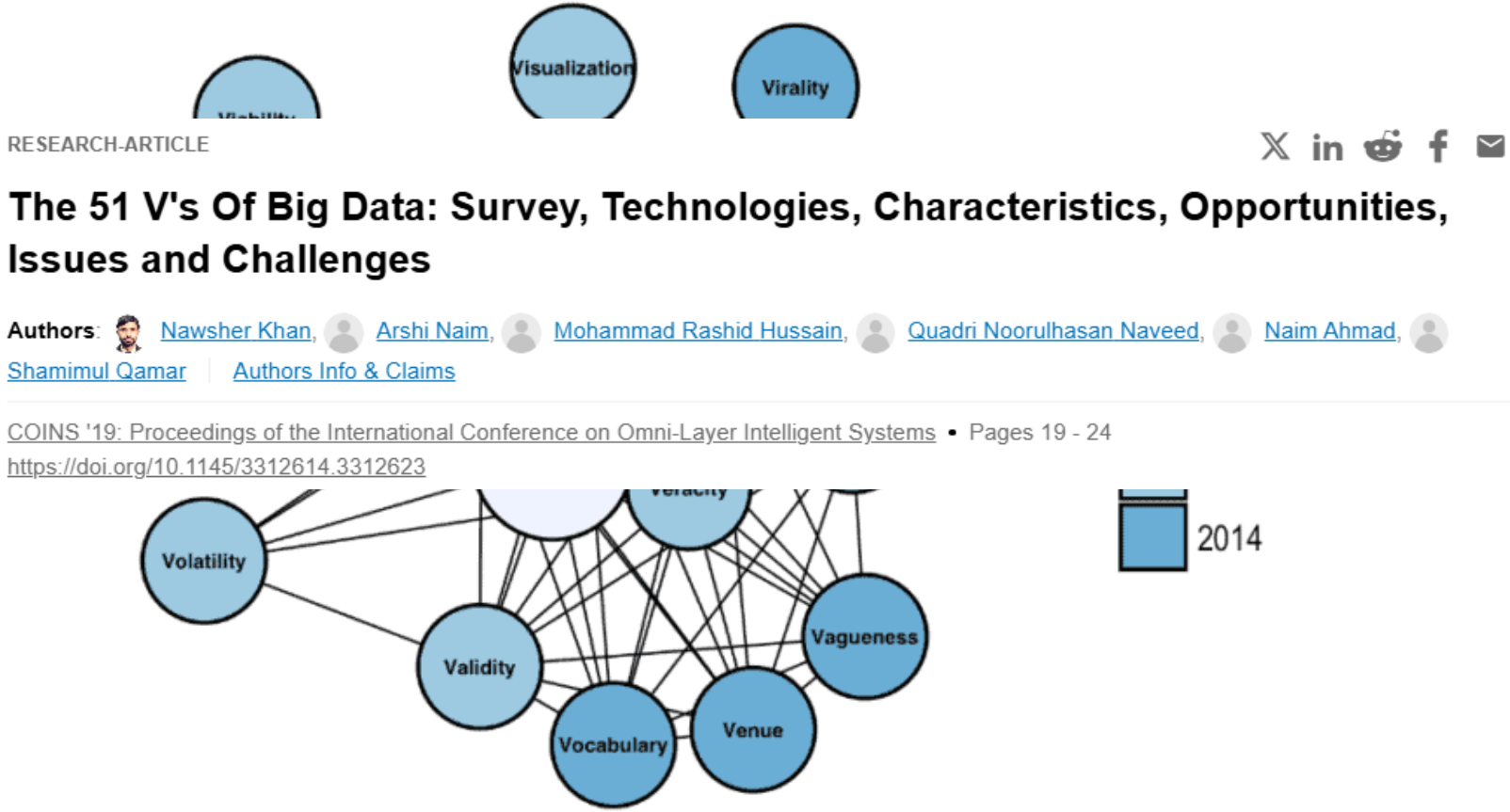
- But wait, there's more!




https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/

# 42 V's of Big Data

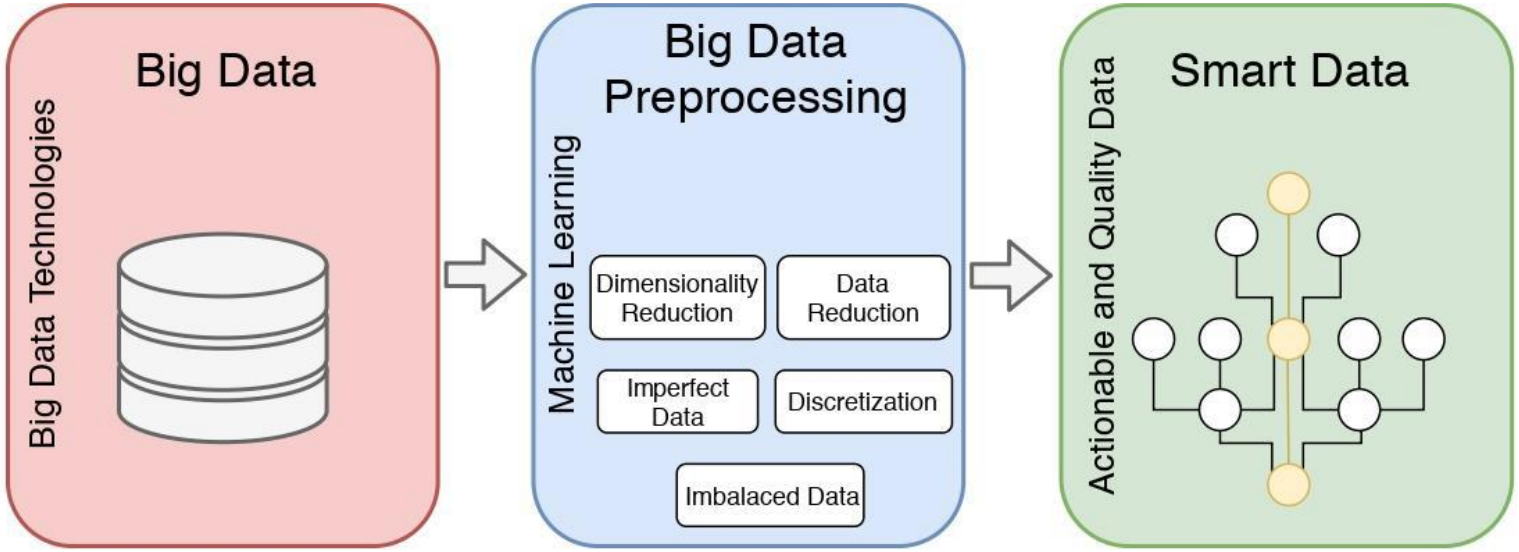
- … and more!




https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/

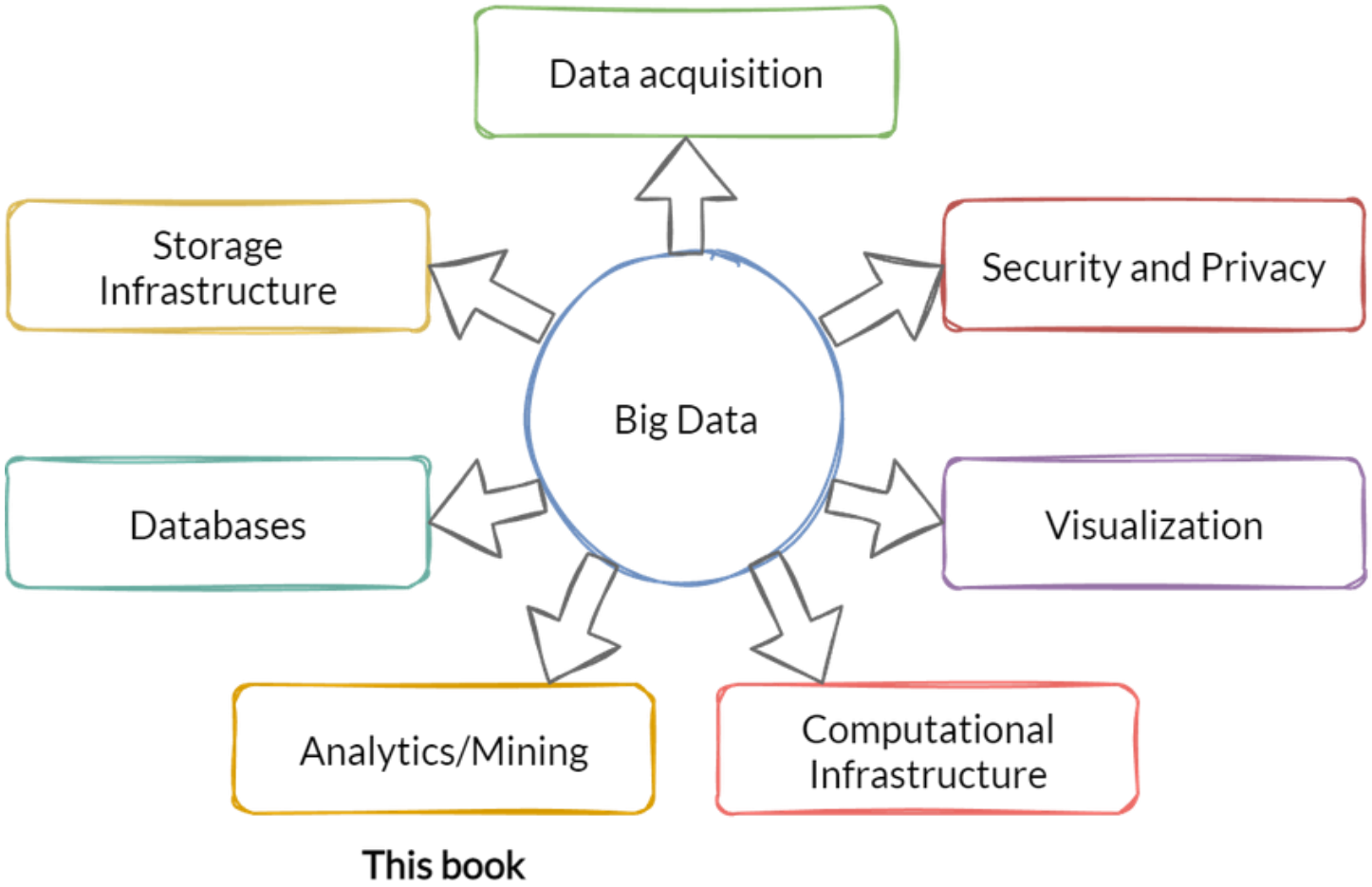
# Smart Data

- Big Data is just a collection of Big and raw data
- Smart Data separates the physical part of the data (Volume, Velocity, Variety), from the Smart part of it (Veracity, Value)




Luengo, J., García Gil, D., Ramírez Gallego , S., García , S., & Herrera, F. (2020). Big Data Preprocessing: Enabling Smart Data. Springer Nature.

# The Faces of Big Data



# How to deal with Big Data?

**Problem statement**: scalability to big data sets.

**Example**:

- Explore 100 TB by 1 node @ 50 MB/sec = 23 days
- Exploration with a cluster of 1000 nodes = 33 minutes

**Solution → Divide-And-Conquer**

A single machine cannot efficiently manage high volumes of data.

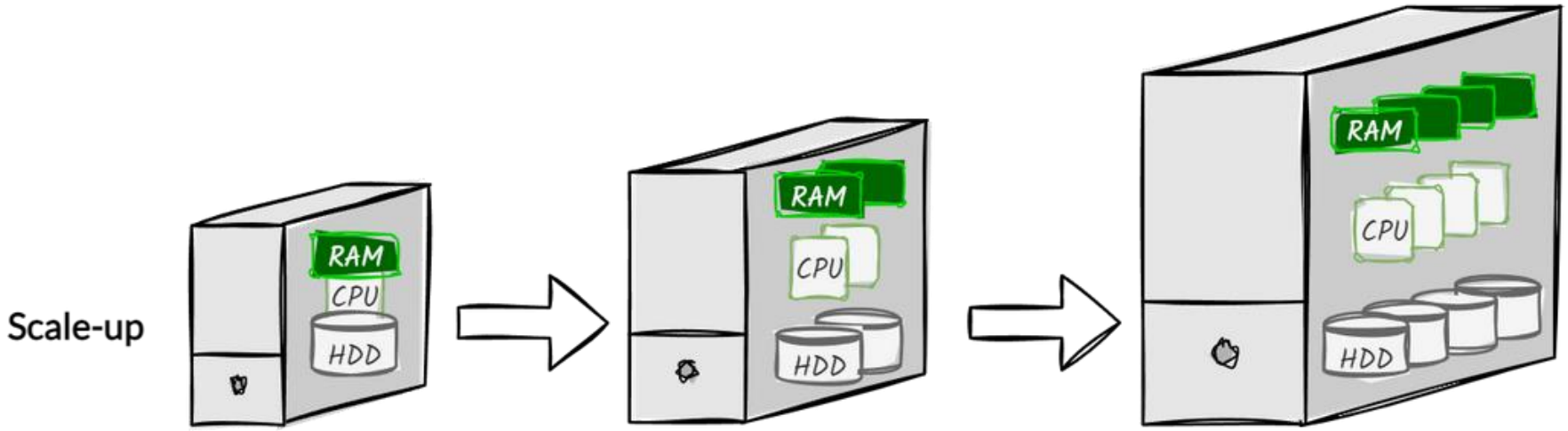
- Vertical scalability → **Horizontal scalability**

Cluster of Computers

# Scale-up vs. Scale-out

**Scale-up (Vertical scalability)**

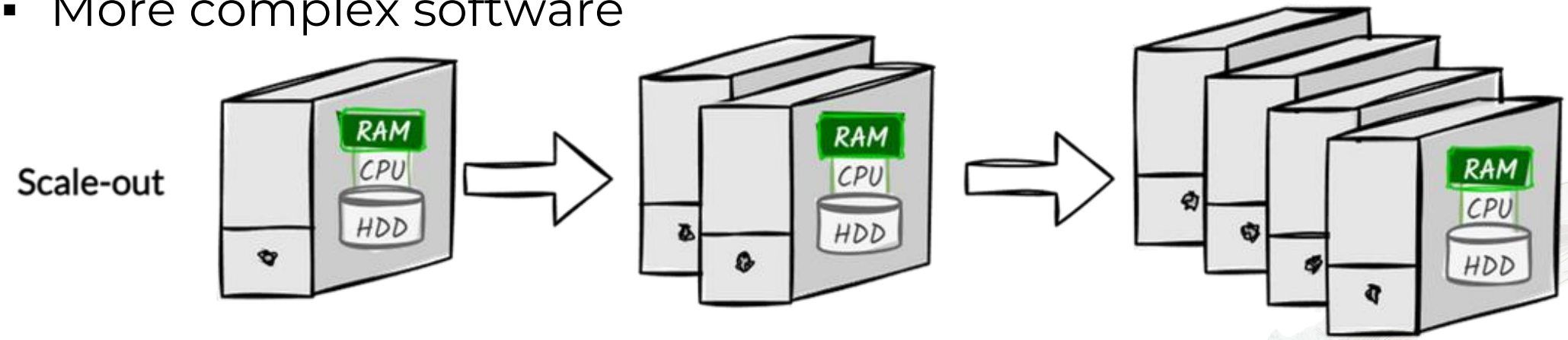
- Use of traditional tools (e.g. matlab or excel) in a single machine
- If we have big datasets, you will need more powerful machines
  - Way more expensive
  - There are limits we can't overcome



# Scale-up vs. Scale-out

**Scale-out (horizontal scalability)**

- Multiple machines connected by a network – **distributed computing**
- No need of high-end hardware to tackle Big Data
  - Each computing node will be cheaper
  - We can add more of those relatively easily
- **Problems**
  - Network bottleneck
  - More complex software



# Scale-up vs. Scale-out: Summary

CAMBRIDGE UNIVERSITY PRESS

**Scale-up**

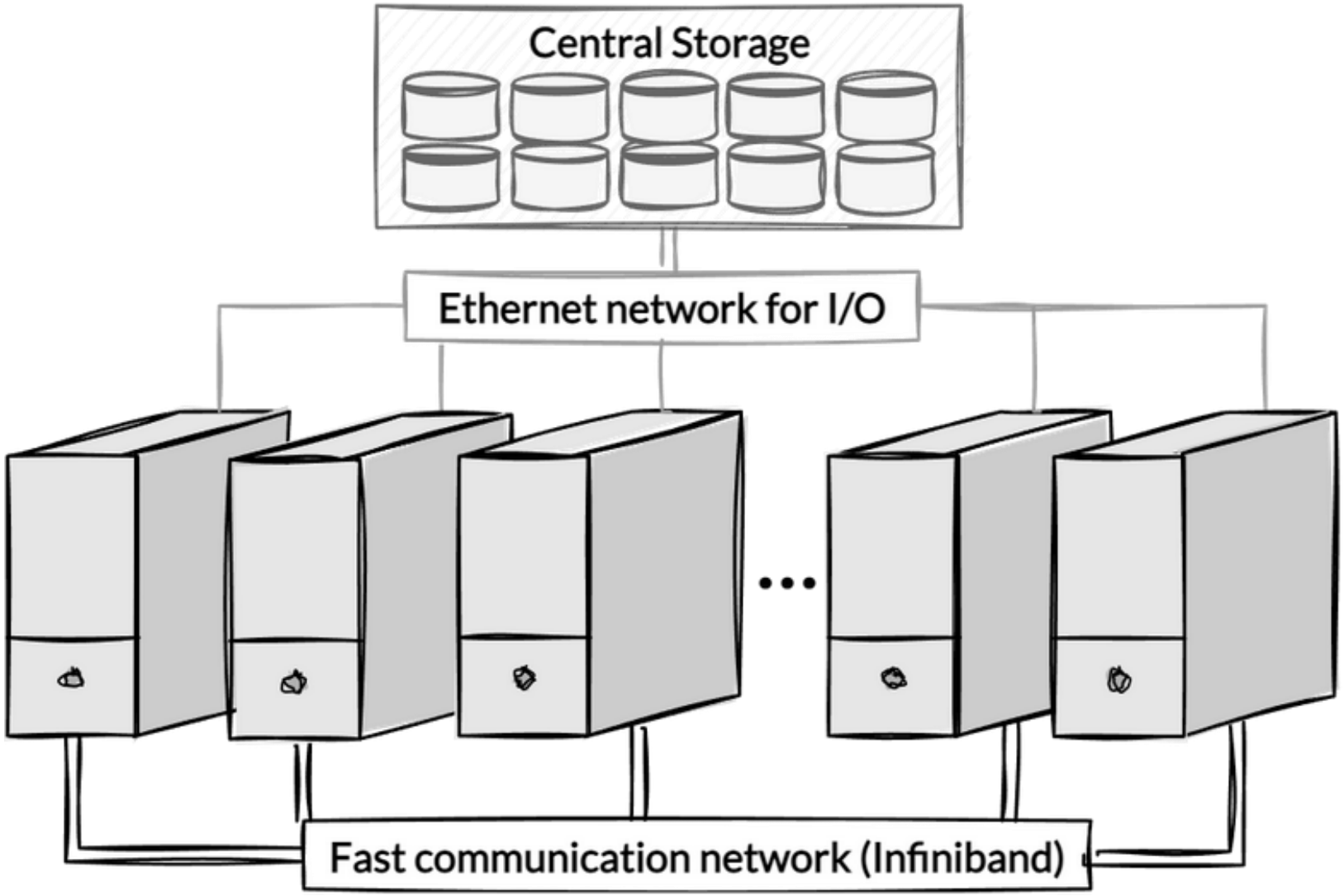
- Add more memory, processors, ...
- Advantages
  - Less energy consumption
  - Less cost in cooling systems
  - Easier to implement solutions
- Disadvantages
  - Price
  - No fault-tolerant
  - Limited upgrades

**Scale out**

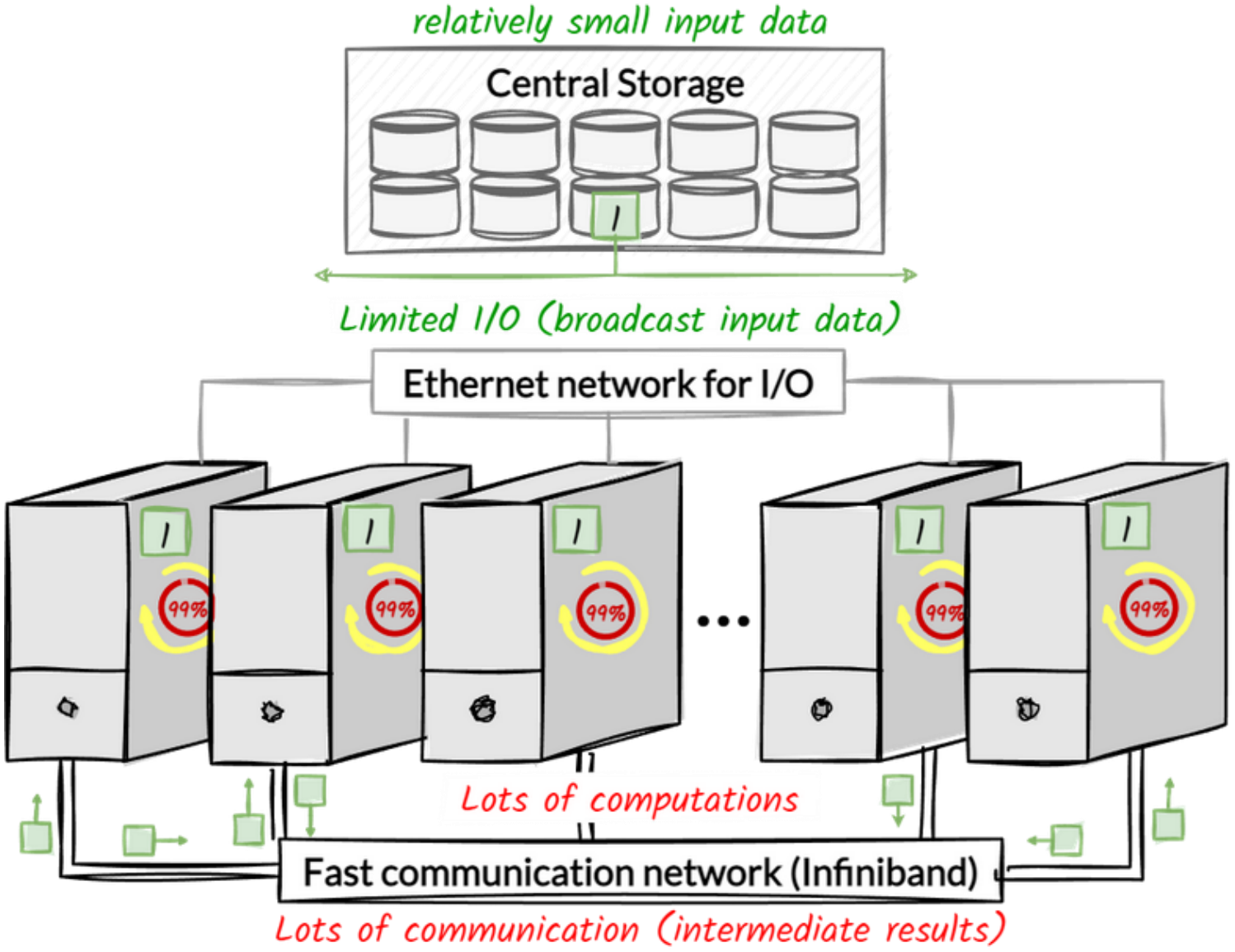
- Add more (cheaper) nodes
- Advantages
  - Cheaper
  - Fault-tolerant is possible
  - Easy to grow
- Disadvantages
  - More physical space
  - Energy costs (elec. and cooling)
  - Network equipment

**Big Data technologies are based on Scale-out**

# What is an HPC?



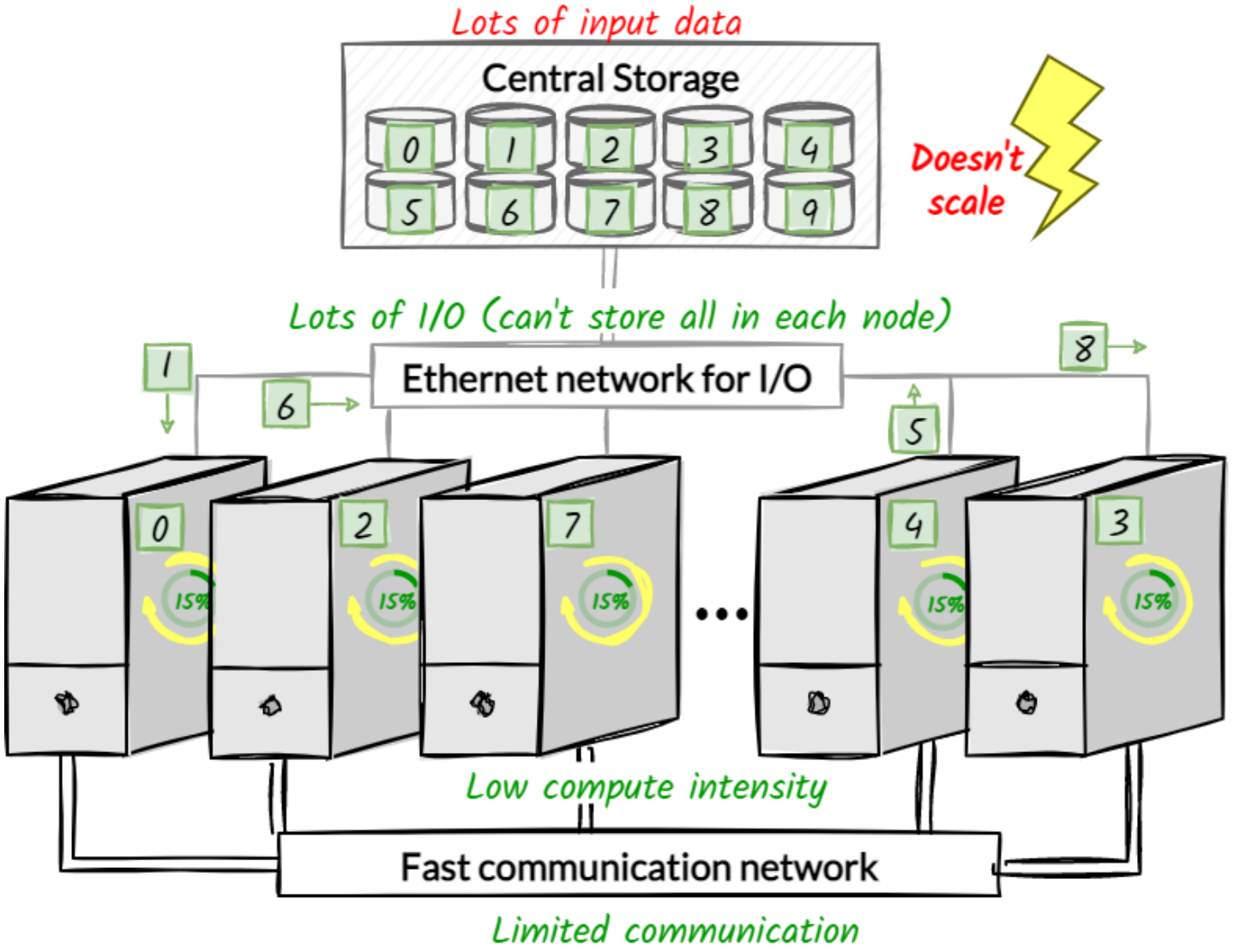
# Traditional HPC way of doing things



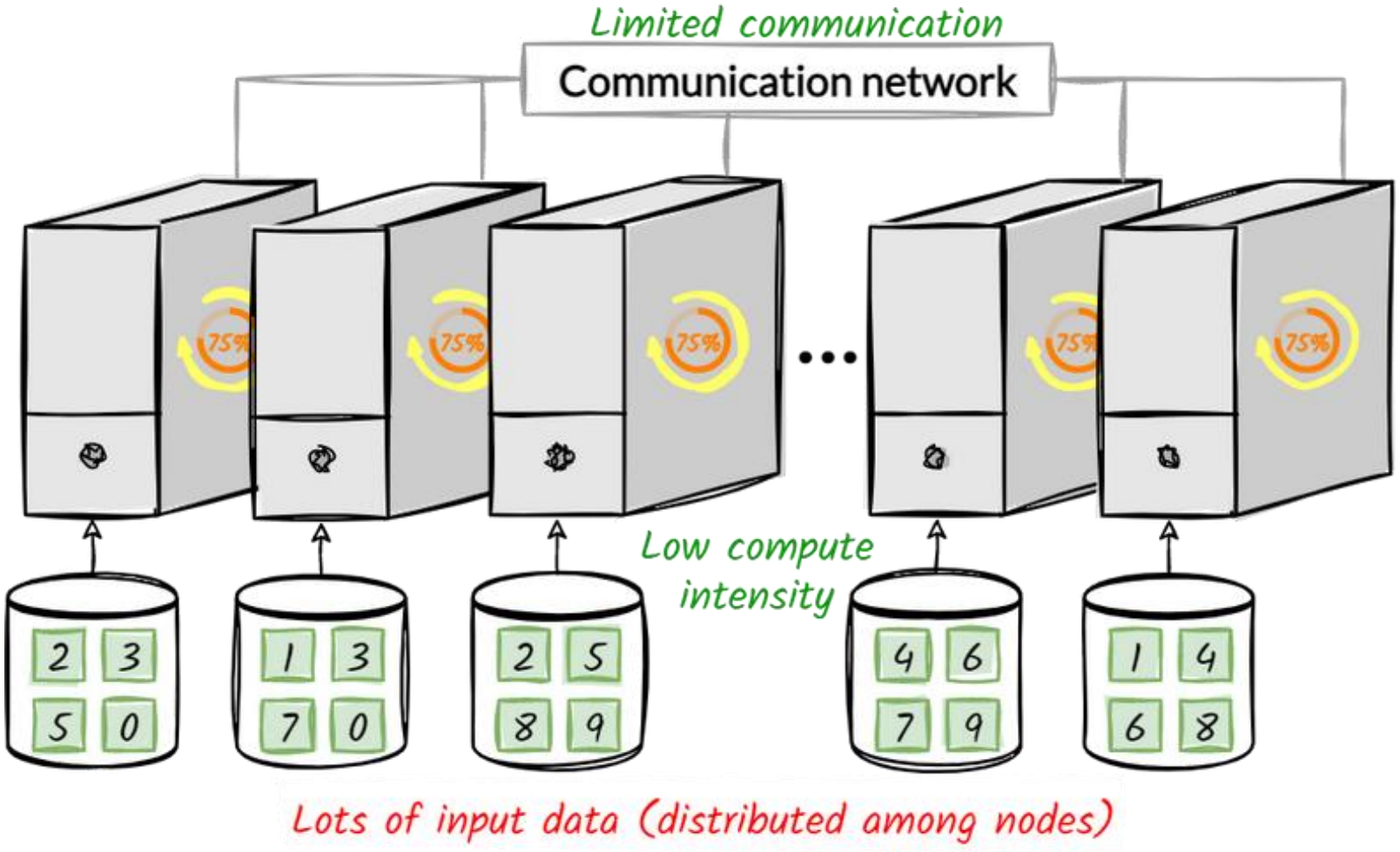
# How to deal with Big Data?


- How do we program that distributed computation?
  - We normally have a driver node, and workers that do the computation
- A classic example for distributed computing is **MPI** (Message Passing Interface)
  - **Scatter**: Send a message to all computing nodes that perform an operation and return a response
  - **Gather**: Receive/collect compute from workers
- Disadvantages:
  - Not *transparent* and it is not *fault-tolerant*!
  - But apart from that why is it not appropriate for Big Data?

# Data-intensive jobs



# Data-intensive jobs: Solution



# Summary: Big data is not HPC

CAMBRIDGE UNIVERSITY PRESS

## Big data

- Focus on data-intensive jobs
- H/W failure common
- Code: graphs, data mining
- Usually mix CPU and data
- Job is moved to where the data is located
- SIMD model: Data parallelism
- Okay with commodity hardware

## HPC

- Focus on computation-intensive jobs
- Surprised by H/W failure
- Code: simulation, optimization
- Mix CPU/GPU
- Data is moved to where it will be processed.
- SIMD/MIMD model (more general parallelism)
- Need specialized hardware

# Distributed Systems in Big Data

**Objective**: To apply an operation to all data

- *Assumption*: One machine cannot process or store all data
  - Data is distributed in a cluster of computing nodes
  - It does not matter which machine executes the operation
  - It does not matter if it is run twice in different nodes (due to failures or straggler nodes)
  - We look for an abstraction of the complexity behind distributed systems

**DATA LOCALITY** is crucial

- Avoid data transfers between machines as much as possible

# Distributed Systems in Big Data

- We will need a new programming model: **MapReduce**
  - *"Moving computation is cheaper than moving computation and data at the same time"*
- **Idea**
  - Data is distributed among nodes (distributed file system)
  - Functions/operations to process data are distributed to all the computing nodes
  - Each computing node works with the data stored in it
  - Only the necessary data is moved across the network

# Take-home Message

- Big Data is not just volume
- Big Data has multiple faces and challenges
- Scale-up vs. scale-out
- Characteristics of a Big Data Cluster
- The principle of Data locality

Next Lecture:

- MapReduce!

# To Learn More

CAMBRIDGE UNIVERSITY PRESS

- **Big Data and large-scale data analytics**
  - Big Data: A Revolution, book (Mayer-Schönberger and Cukier, 2013)
  - The Big Challenges of Big Data, article (Marx, 2013)
  - Big Data and Philosophy, book (Pietsch, 2021)
- **HPC and MPI**
  - High Performance Computing, book (Severance and Dowd, 2010)
  - MPI, book (Gropp et al., 2014)

CAMBRIDGE UNIVERSITY PRESS

# Chapter 1

# Introduction to big data