

# DISTRIBUTED COMPUTING AND BIG DATA

**Venkatesh Vinayakarao**

[venkateshv@cmi.ac.in](mailto:venkateshv@cmi.ac.in)

<http://vvtesh.co.in>

---

Chennai Mathematical Institute

---

Data is the new oil. - Clive Humby, 2006.

# Know Your Instructor

BE (Computer Science and Engineering)

Java/J2EE  
Developer

MS (Information Technology)

SDE, Search  
Technologies  
Group, Bing,  
Microsoft

Principal  
Engineer, Cloud  
Platforms Group,  
Yahoo

PhD (Computer Science)

Principal  
Engineer, Search,  
Here  
Technologies

Intern, Porting ML  
Models to Azure,  
Microsoft  
Research

# Agenda

- Introduction to Big Data
- Course Dynamics
- Evolution of Systems and Technologies
  - Data Storage
  - Data Processing

# What Comes Next?

byte

kilobyte

megabyte

gigabyte

??

???

????

?????

# Sizes

Name	Size
Byte	8 bits
Kilobyte	1024 bytes
Megabyte	1024 kilobytes
Gigabyte	1024 megabytes
Terabyte	1024 gigabytes
Petabyte	1024 terabytes
Exabyte	1024 petabytes
Zettabyte	1024 exabytes
Yottabyte	1024 zettabytes

# The Impact of Big Data



## Your train is on time thanks to **big data**

TNW - 31-Dec-2019

Thanks to thousands of sensors and **big data** analytics, train ... It's this data that keeps the Dutch rail network moving, and helps NS deliver a ...



## The power of **data** in smart city developments

Independent Australia - 03-Jan-2020

Other fascinating **big data** developments that were presented included ... led to the production of the Australian **Cancer Atlas** — an interactive, ...



## At HCA Healthcare, Real-Time **Data Saves Lives**

RTInsights (press release) (blog) - 01-Jun-2019

At HCA Healthcare, Real-Time **Data Saves Lives** ... “Our existing **data** infrastructure was designed for **large-scale** business intelligence and ...

# Big Data is Ubiquitous

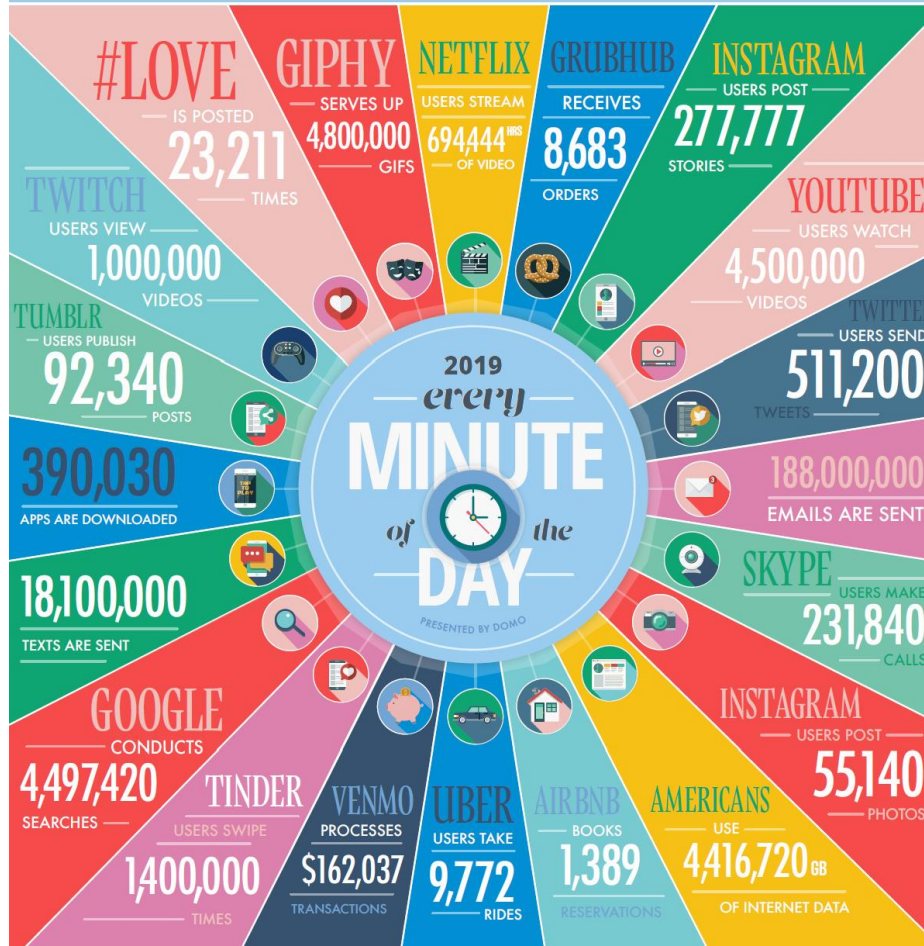
- Facebook (**per day** statistics)
  - 1.5 billion people are active on Facebook **daily!**
  - More than 300 million photos get uploaded **per day!**
  - Totally, more than 2.5 Trillion posts!
- Facebook (per minute statistics)
  - **Every minute** there are 510,000 comments posted and 293,000 statuses updated!
- Youtube (**per minute** statistics)
  - Users watch 4,146,600 YouTube videos!



# DATA NEVER SLEEPS 7.0

How much data is generated *every minute*?

There's no way around it: big data just keeps getting bigger. The numbers are staggering, and they're not slowing down. By 2020, there will be 40x more bytes of data than there are stars in the observable universe. In our 7th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute.



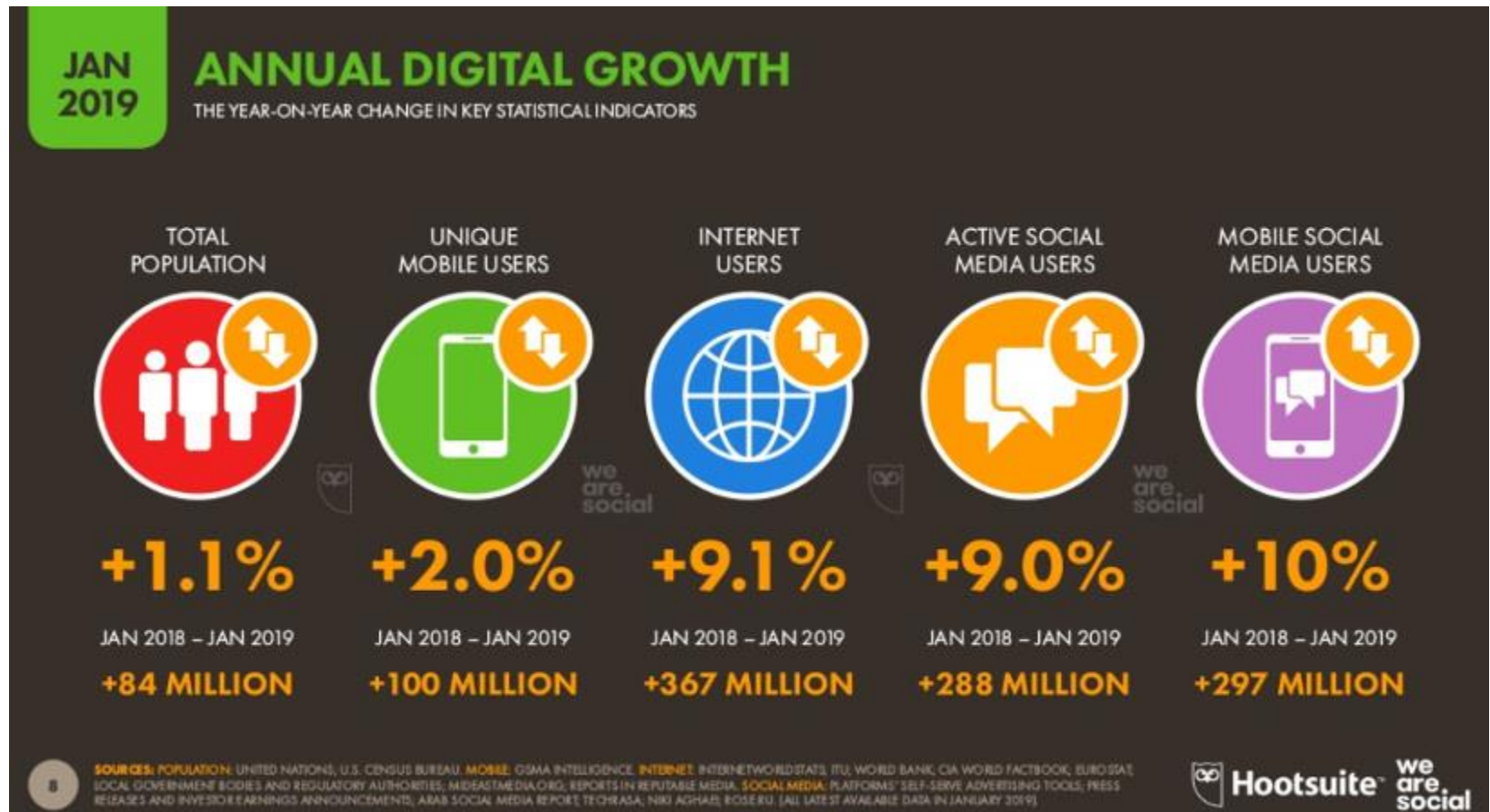
SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRED



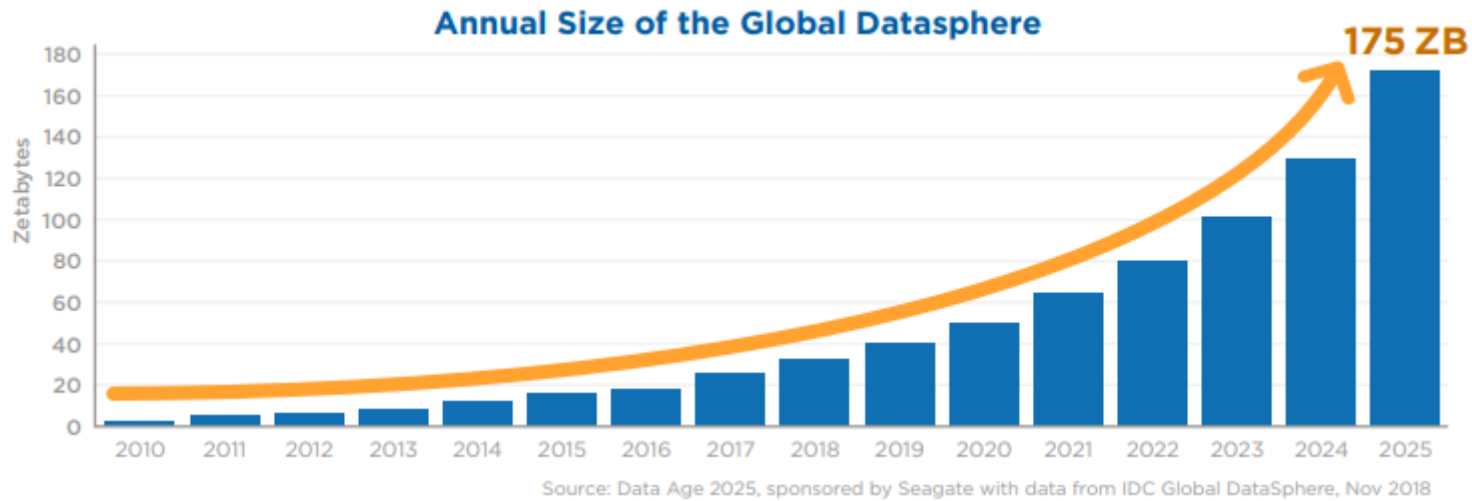
Source: <https://www.visualcapitalist.com/big-data-keeps-getting-bigger/>



# And, It is Growing!



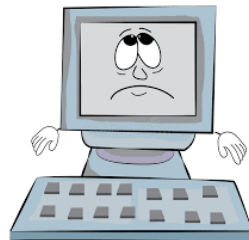
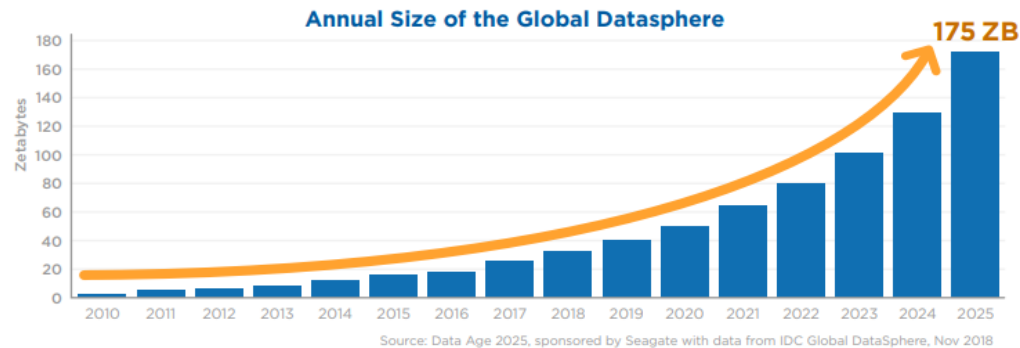
# Data Growth



Mankind's quest to digitize the world!  
33 ZB (2018) → 175 ZB (2025)  
size of global datasphere\*

\*Source: <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>

# Beyond a Single Machine

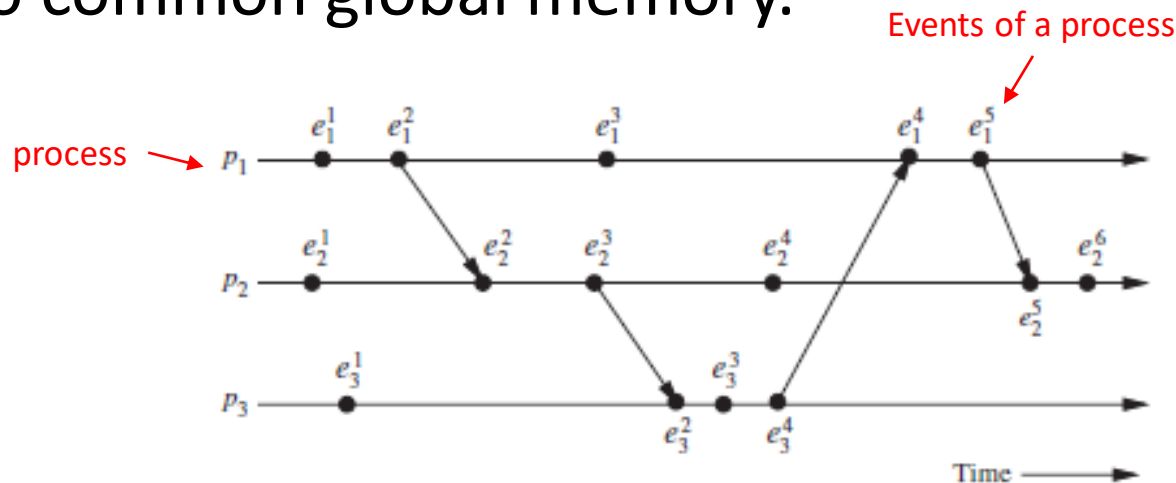


**Global datasphere is growing!**

How has computing evolved to capture, process  
and analyze these data?

# A Model of a Distributed System

- A set of processes connected by a communication network.
- Communication by information exchange.
- No physical global clock.
- No common global memory.



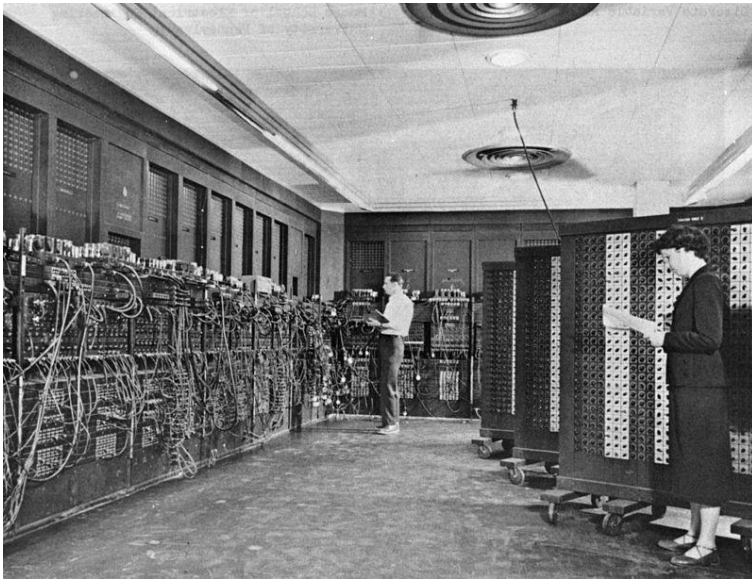
Processors ( $p_i$ ) may fail!

Messages may be delayed or lost!!

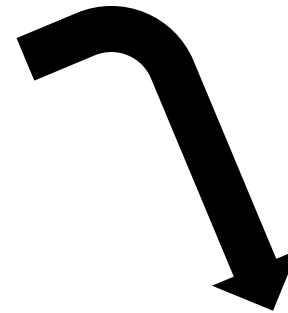
# Course Dynamics

<https://vvtesh.github.io/teaching/bdh-2023.html>

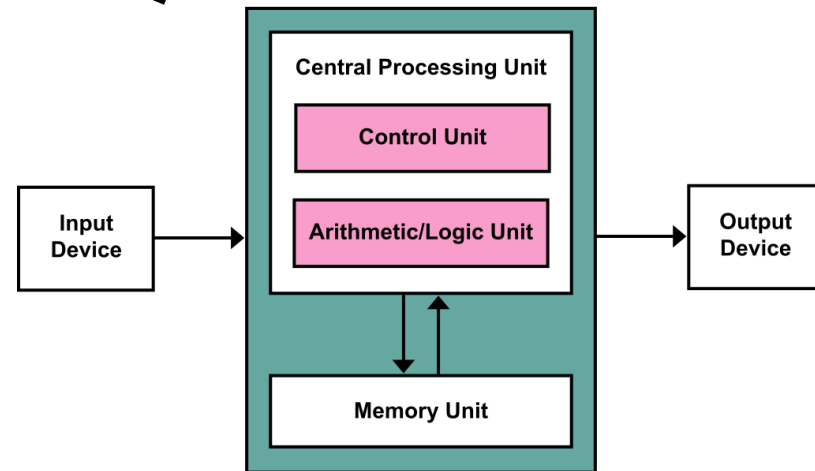
# Evolution of Computers



**ENIAC**  
Early 1900s



**Stored-program  
Von Neumann  
Architecture  
1940**



# Two Kinds of Problems



**Storage**

**Processing**