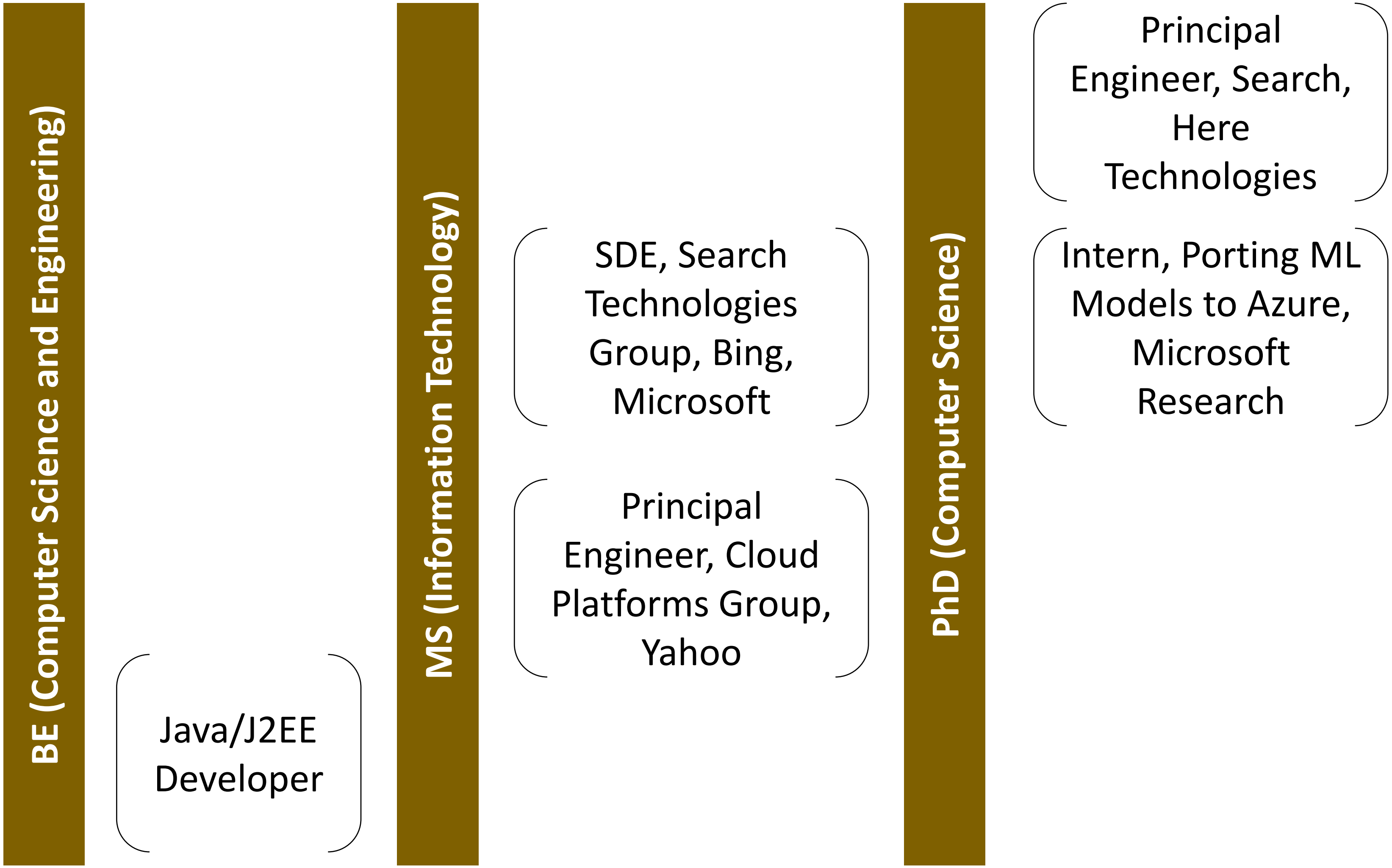
Distributed Computing and Big Data

Venkatesh Vinayakarao
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

??

???

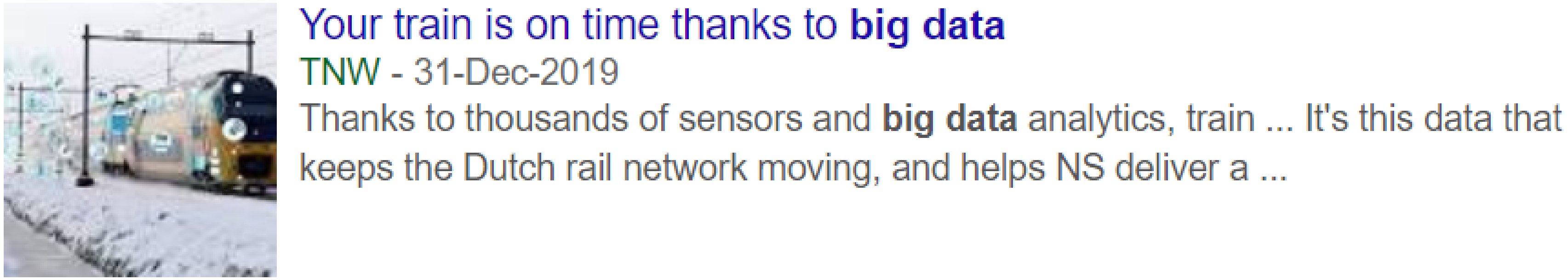
????

?????

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**

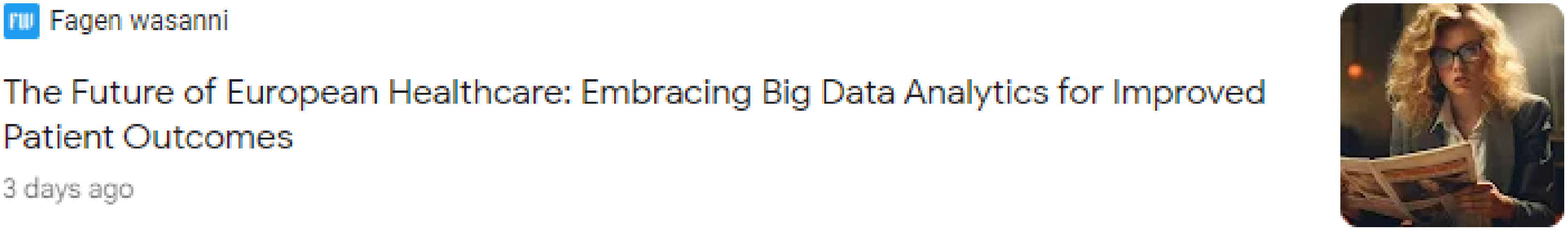
TNW - 31-Dec-2019

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, ...



Fagen wasanni

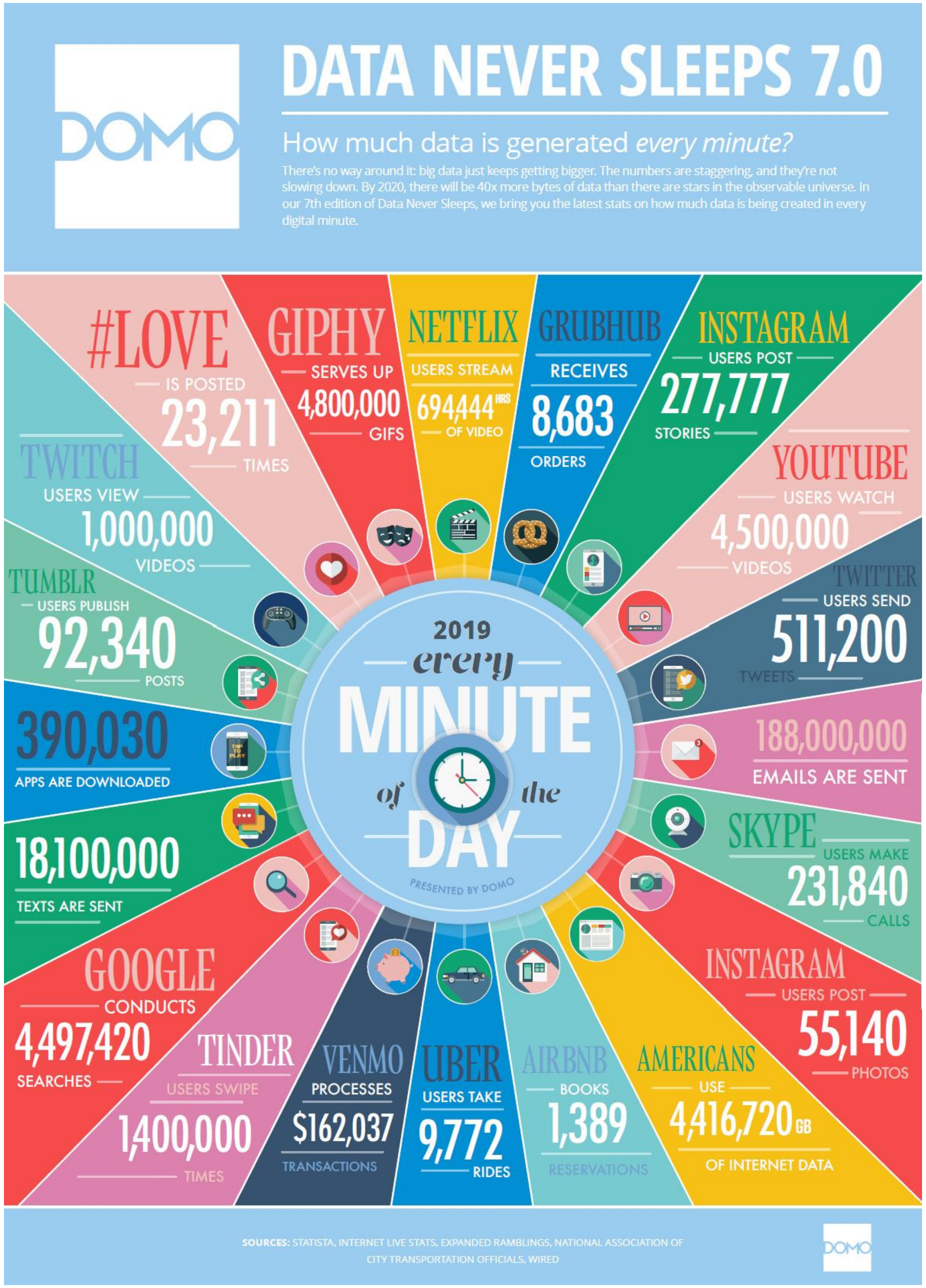
The Future of European Healthcare: Embracing Big Data Analytics for Improved Patient Outcomes

3 days ago

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!

Source: Forbes



DOMO

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.

2019 *every* MINUTE *of the* DAY

PRESENTED BY DOMO

#LOVE IS POSTED 23,211 TIMES

GIPHY SERVES UP 4,800,000 GIFS

NETFLIX USERS STREAM 694444 HRS OF VIDEO

GRUBHUB RECEIVES 8,683 ORDERS

INSTAGRAM USERS POST 277,777 STORIES

YOUTUBE USERS WATCH 4,500,000 VIDEOS

TWITTER USERS SEND 511,200 TWEETS

188,000,000 EMAILS ARE SENT

SKYPE USERS MAKE 231,840 CALLS

INSTAGRAM USERS POST 55,140 PHOTOS

AMERICANS USE 4,416,720 GB OF INTERNET DATA

AIRBNB BOOKS 1,389 RESERVATIONS

UBER USERS TAKE 9,772 RIDES

VENMO PROCESSES $162,037 TRANSACTIONS

TINDER USERS SWIPE 1,400,000 TIMES

GOOGLE CONDUCTS 4,497,420 SEARCHES

18,100,000 TEXTS ARE SENT

390,030 APPS ARE DOWNLOADED

TUMBLR USERS PUBLISH 92,340 POSTS

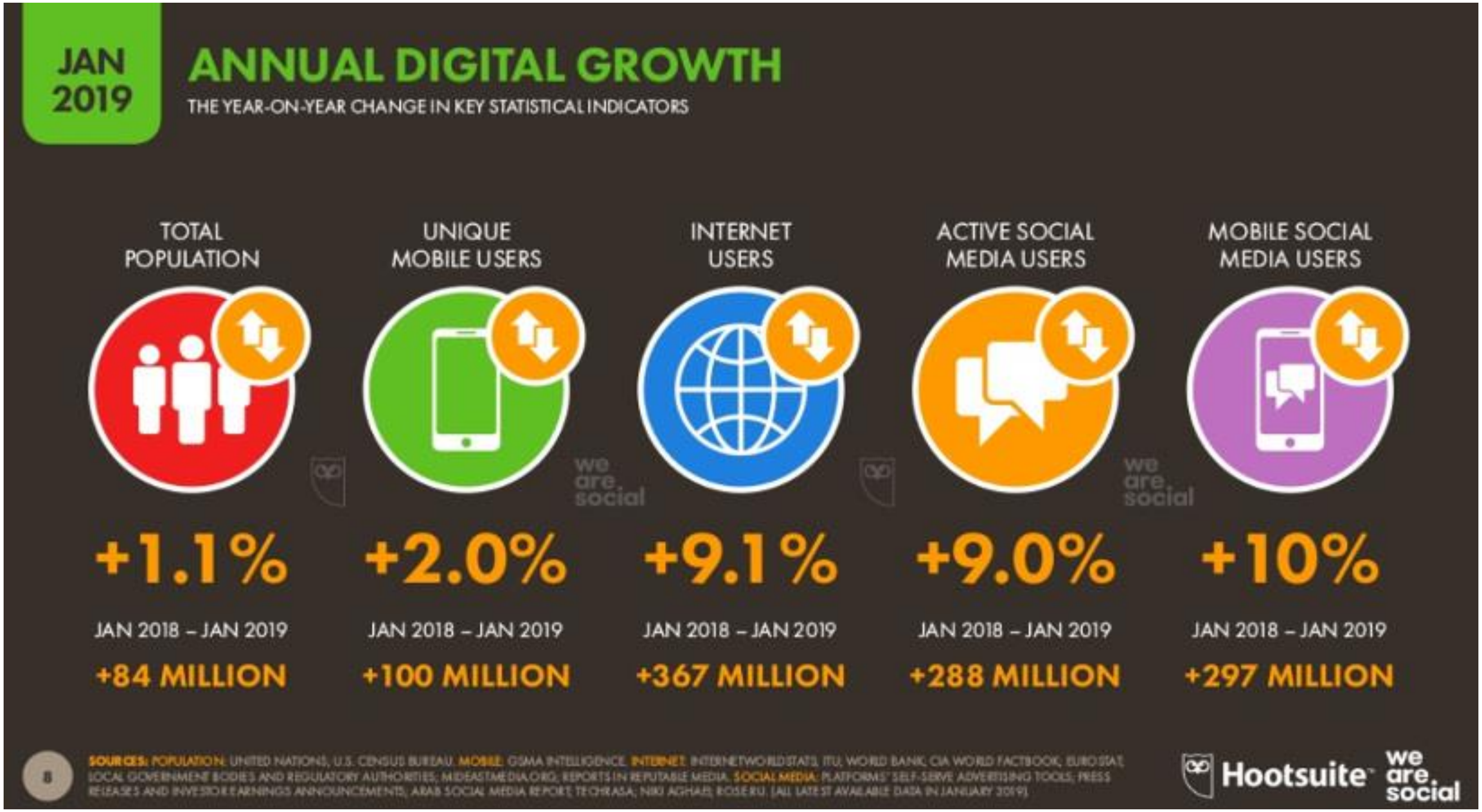
TWITCH USERS VIEW 1,000,000 VIDEOS

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

DOMO

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

And, It is Growing!



JAN 2019

ANNUAL DIGITAL GROWTH

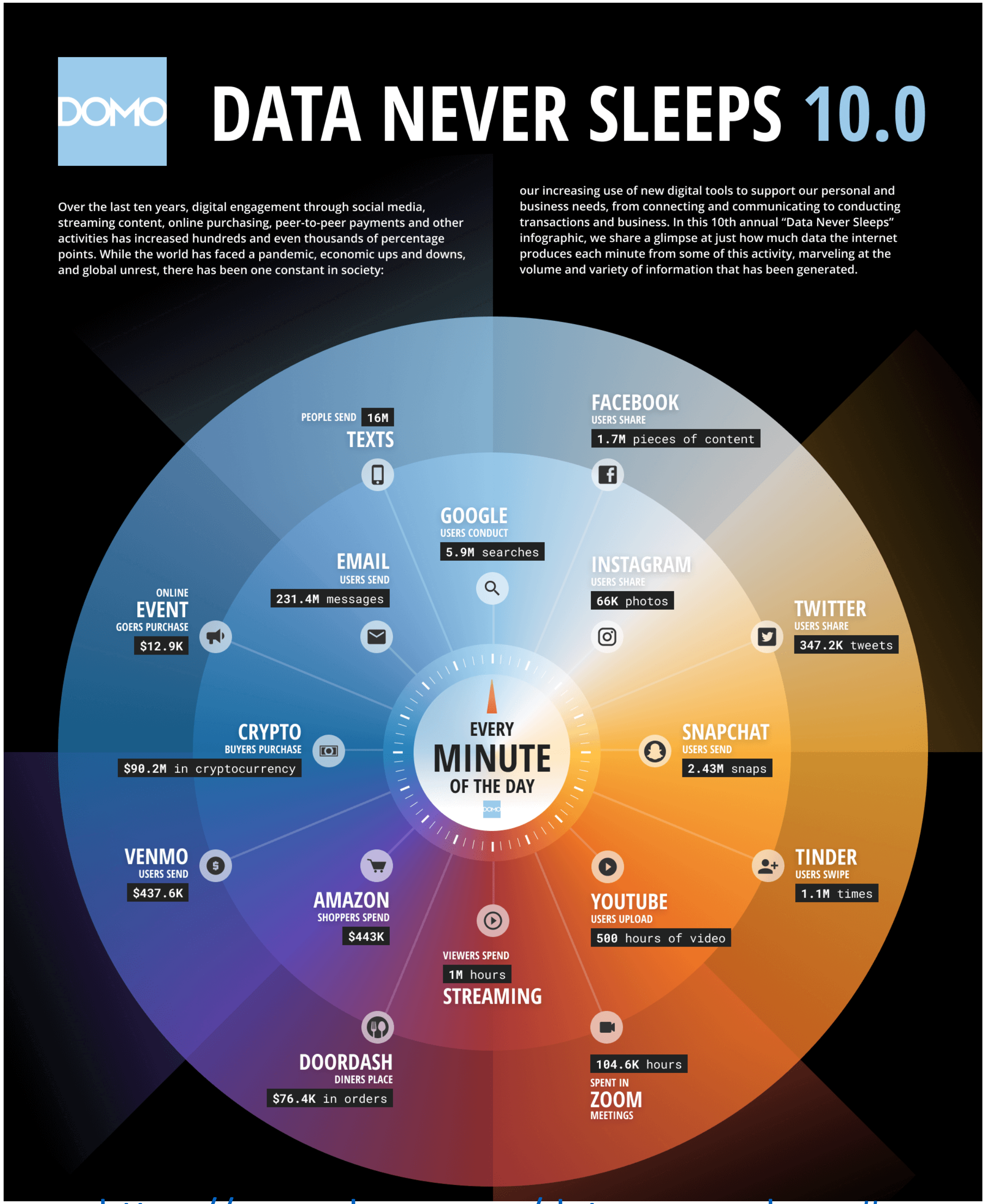
THE YEAR-ON-YEAR CHANGE IN KEY STATISTICAL INDICATORS

TOTAL POPULATION	UNIQUE MOBILE USERS	INTERNET USERS	ACTIVE SOCIAL MEDIA USERS	MOBILE SOCIAL MEDIA USERS
+1.1%	+2.0%	+9.1%	+9.0%	+10%
JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019
+84 MILLION	+100 MILLION	+367 MILLION	+288 MILLION	+297 MILLION

8

SOURCES: POPULATION: UNITED NATIONS, U.S. CENSUS BUREAU. MOBILE: GSMA INTELLIGENCE. INTERNET: INTERNETWORLDSTATS, ITU; WORLD BANK; CIA WORLD FACTBOOK; EUROSTAT; LOCAL GOVERNMENT BODIES AND REGULATORY AUTHORITIES; MIDEASTMEDIA.ORG; REPORTS IN REPUTABLE MEDIA. SOCIAL MEDIA: PLATFORMS' SELF-SERVE ADVERTISING TOOLS; PRESS RELEASES AND INVESTOR EARNINGS ANNOUNCEMENTS; ARAB SOCIAL MEDIA REPORT; TECHRASA; NIKI AGHAEI; ROSE.RU. (ALL LATEST AVAILABLE DATA IN JANUARY 2019).

Hootsuite we are social



DOMO DATA NEVER SLEEPS 10.0

Over the last ten years, digital engagement through social media, streaming content, online purchasing, peer-to-peer payments and other activities has increased hundreds and even thousands of percentage points. While the world has faced a pandemic, economic ups and downs, and global unrest, there has been one constant in society: our increasing use of new digital tools to support our personal and business needs, from connecting and communicating to conducting transactions and business. In this 10th annual "Data Never Sleeps" infographic, we share a glimpse at just how much data the internet produces each minute from some of this activity, marveling at the volume and variety of information that has been generated.

EVERY MINUTE OF THE DAY

PEOPLE SEND 16M TEXTS

FACEBOOK USERS SHARE 1.7M pieces of content

GOOGLE USERS CONDUCT 5.9M searches

EMAIL USERS SEND 231.4M messages

INSTAGRAM USERS SHARE 66K photos

ONLINE EVENT GOERS PURCHASE $12.9K

TWITTER USERS SHARE 347.2K tweets

CRYPTO BUYERS PURCHASE $90.2M in cryptocurrency

SNAPCHAT USERS SEND 2.43M snaps

VENMO USERS SEND $437.6K

AMAZON SHOPPERS SPEND $443K

TINDER USERS SWIPE 1.1M times

YOUTUBE USERS UPLOAD 500 hours of video

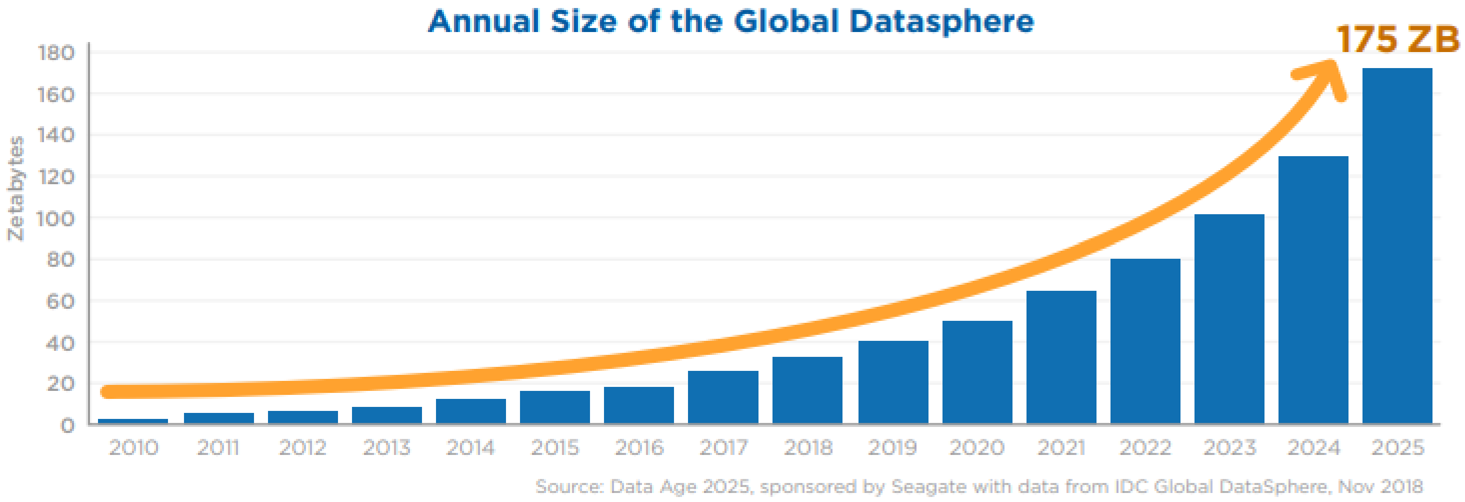
VIEWERS SPEND 1M hours STREAMING

DOORDASH DINERS PLACE $76.4K in orders

104.6K hours SPENT IN ZOOM MEETINGS

https://www.domo.com/data-never-sleeps#

Data Growth

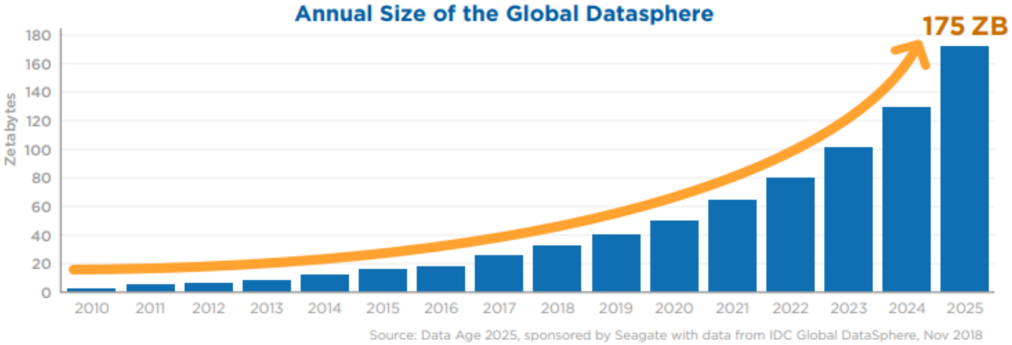


Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

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

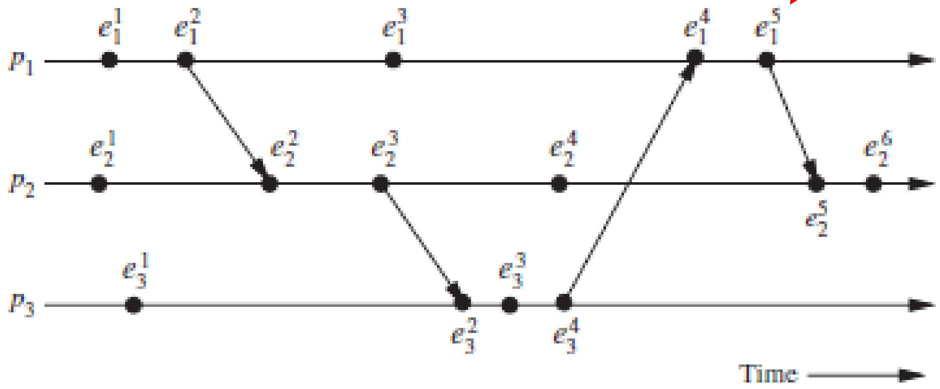


Global datasphere is growing!

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.



process

Events of a process

Processors (p_i) may fail!

Messages may be delayed or lost!!

Course Dynamics

https://vvtesh.github.io/teaching/dcbd-saiu-2023.html