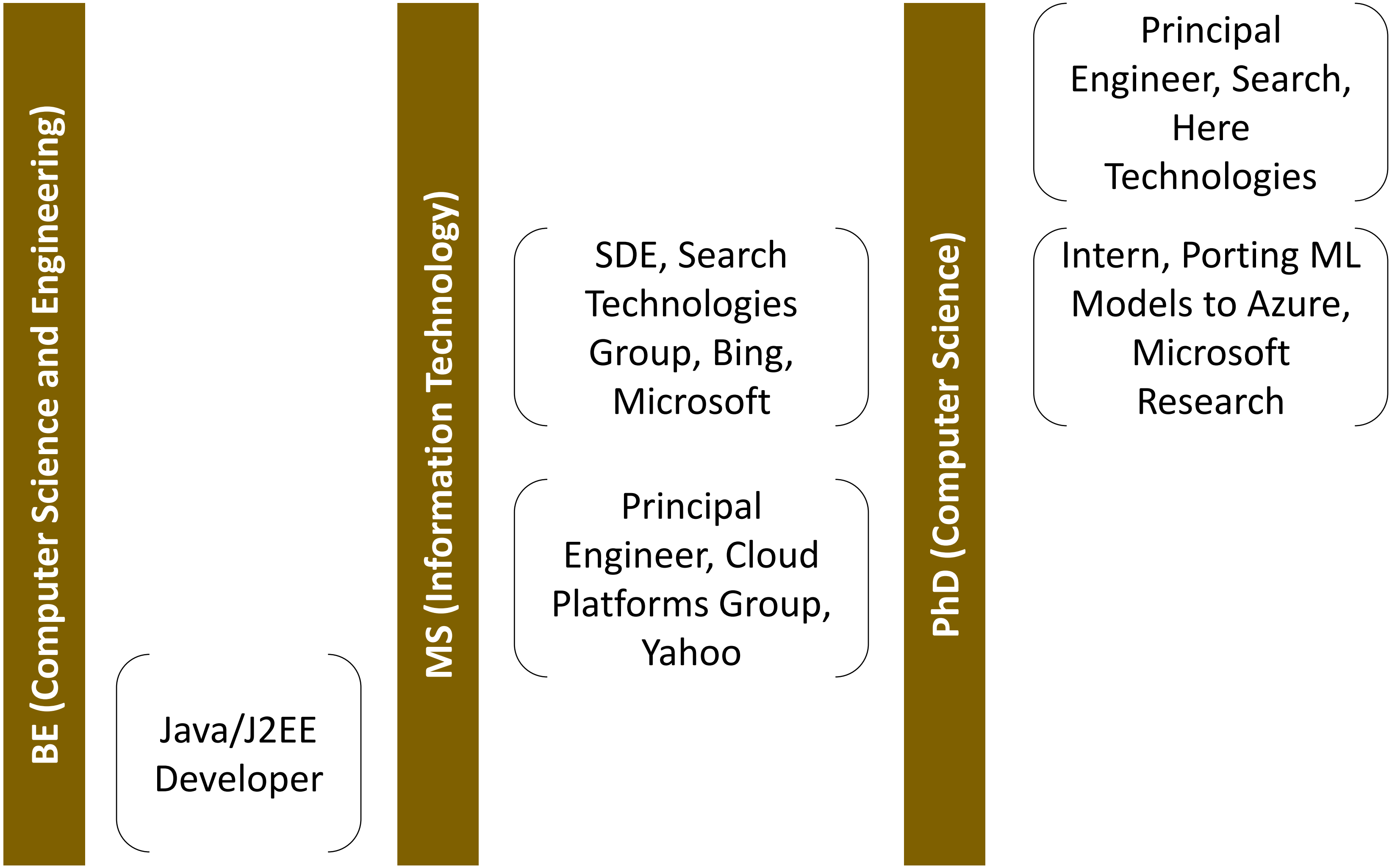
Distributed Computing and Big Data

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)

Intern, Porting ML Models to Azure, Microsoft Research

Principal Engineer, Search, Here Technologies

Things to Remember

- Lectures usually start on time and end on time.
- All exams are closed book exams.
- Visit your course page for slides, readings and more.
- Discussion during lecture sessions are encouraged. Do not wait till the end if you have questions/comments.
- Exams will be in-person (not online).
- All assignment submissions will be on moodle. Assignment deadlines penalties are very strict.
    - 1 min delay: 0.5 marks
    - 1 hour delay: 1 marks
    - 12 hours delay: 2 marks
    - 24 hours delay: 3 marks
    - Beyond 24 hours, a new make-up assignment might be given.
- Please keep your video “on” if possible so that I can assign faces to names.
- I provide recommendation letters only when you have worked with me beyond the course work (some research project, internships, etc).

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

Trivia

- How many websites are there in the world wide web?
- If each website has 100 pages on average, how many pages would we have?
- If each page consumes 1 MB (on average), how much space is required to store all the pages? (Use 1000 instead of 1024 for ease of calculation)

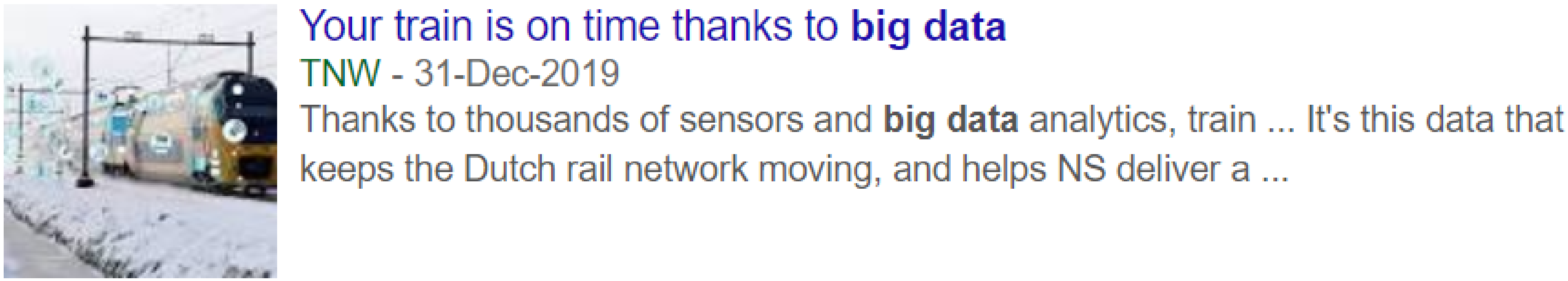
Trivia

- How many websites are there in the world wide web?
  - Approximately 1.1 Billion
- If each website has 100 pages on average, how many pages would we have?
  - 110 Billion
- If each page consumes 1 MB (on average), how much space is consumed?
  - 110 * 10^9 * 1 MB = 110 * 10^9 * 10^6 bytes = 110 PB

Activity

- Visit https://news.google.com/
- Search for “Big data”
- Paste the most interesting news title that you saw which got published in the recent one year.

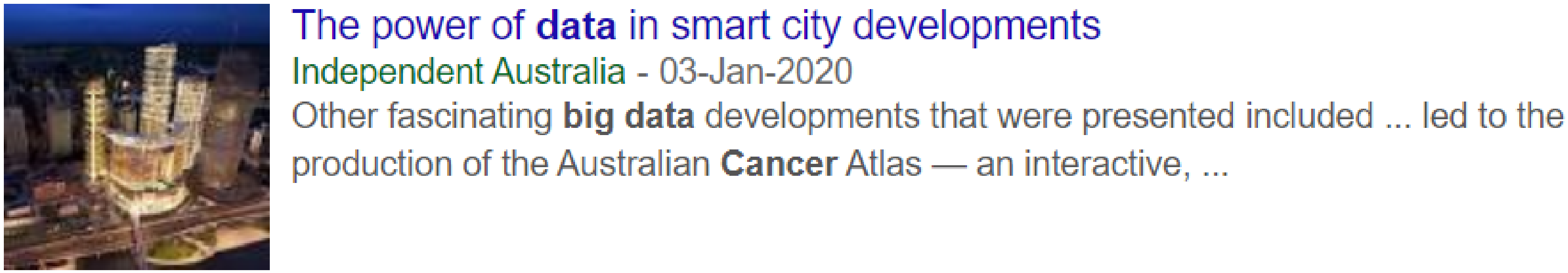
The Impact of Big Data



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

TNW - 31-Dec-2019

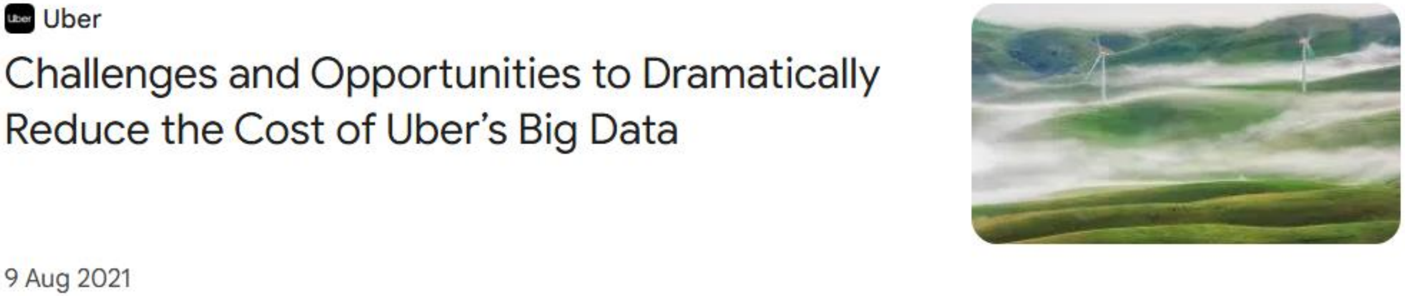
Thanks to thousands of sensors and **big data** analytics, train ... It's this data that keeps the Dutch rail network moving, and helps NS deliver a ...



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, ...

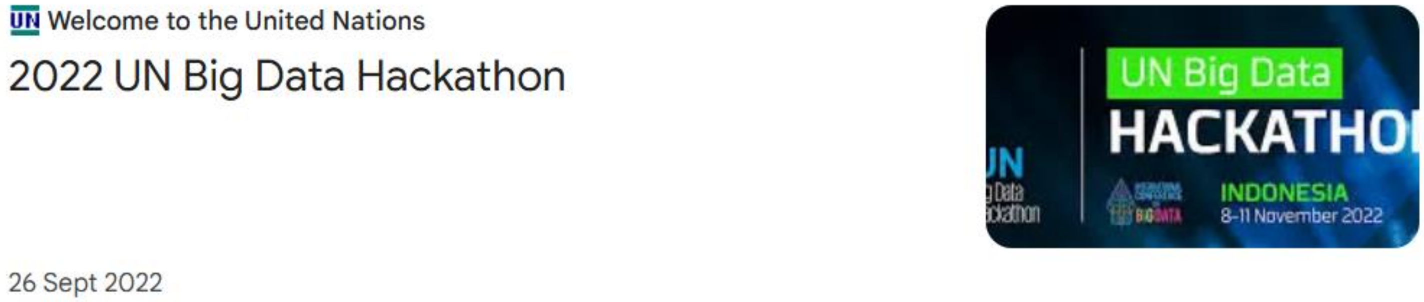


Uber

Challenges and Opportunities to Dramatically Reduce the Cost of Uber's Big Data

9 Aug 2021

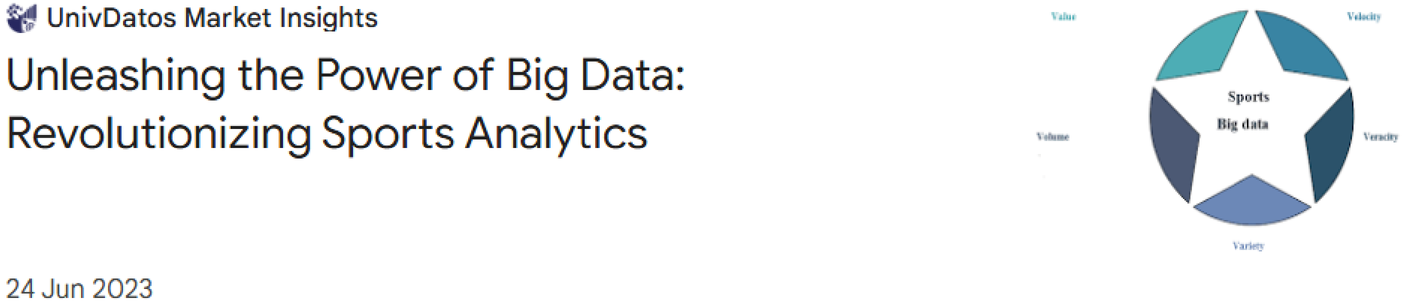
The Impact of Big Data



Welcome to the United Nations

2022 UN Big Data Hackathon

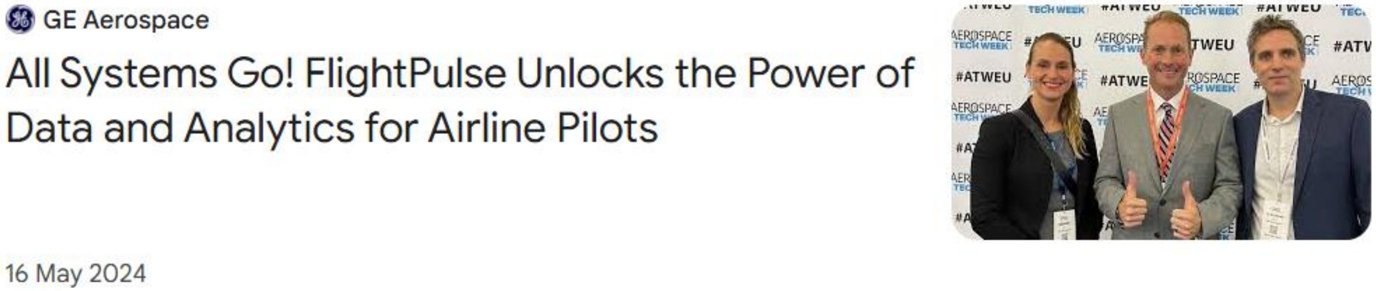
26 Sept 2022



UnivDatos Market Insights

Unleashing the Power of Big Data: Revolutionizing Sports Analytics

24 Jun 2023



GE Aerospace

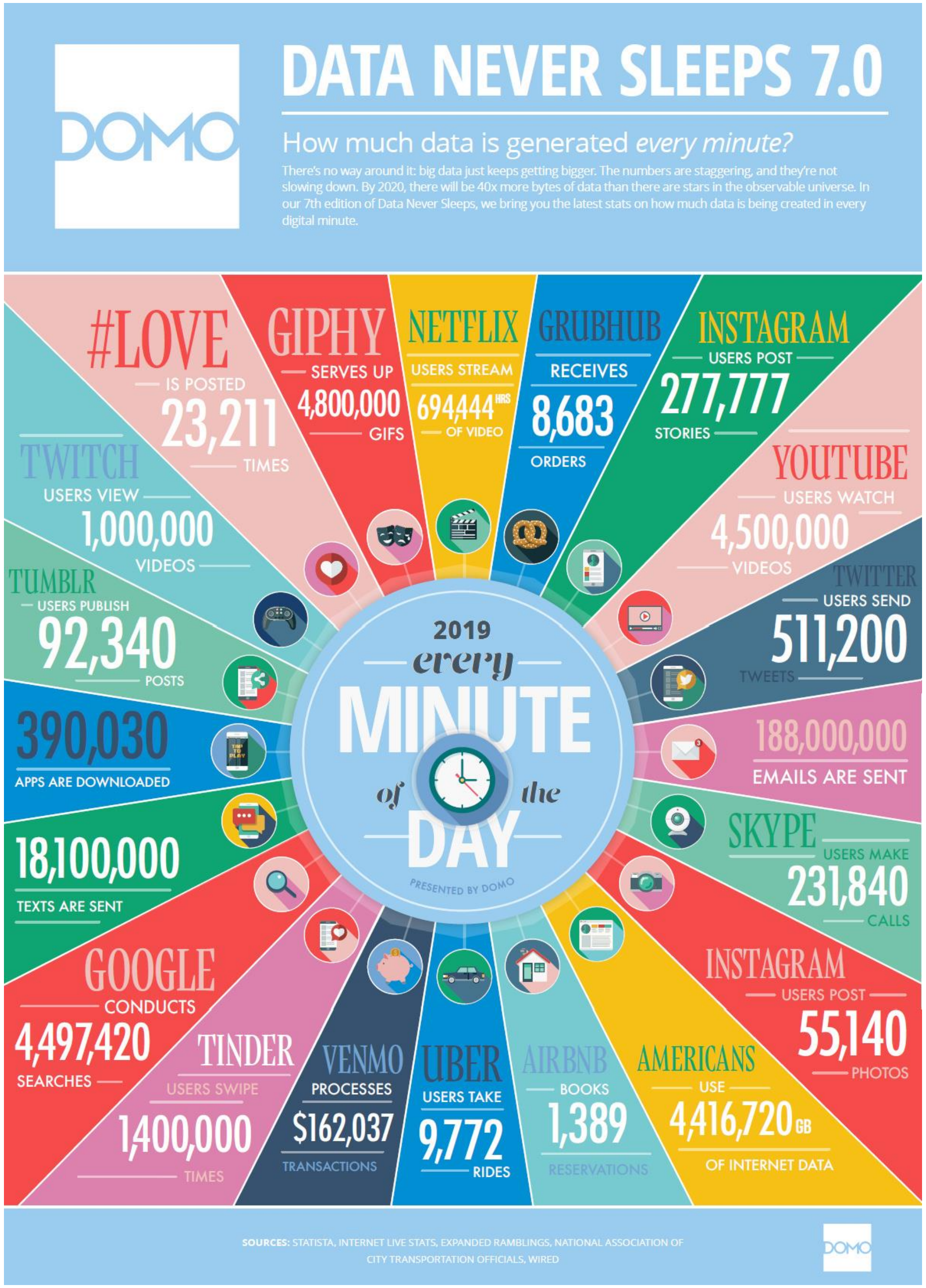
All Systems Go! FlightPulse Unlocks the Power of Data and Analytics for Airline Pilots

16 May 2024

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

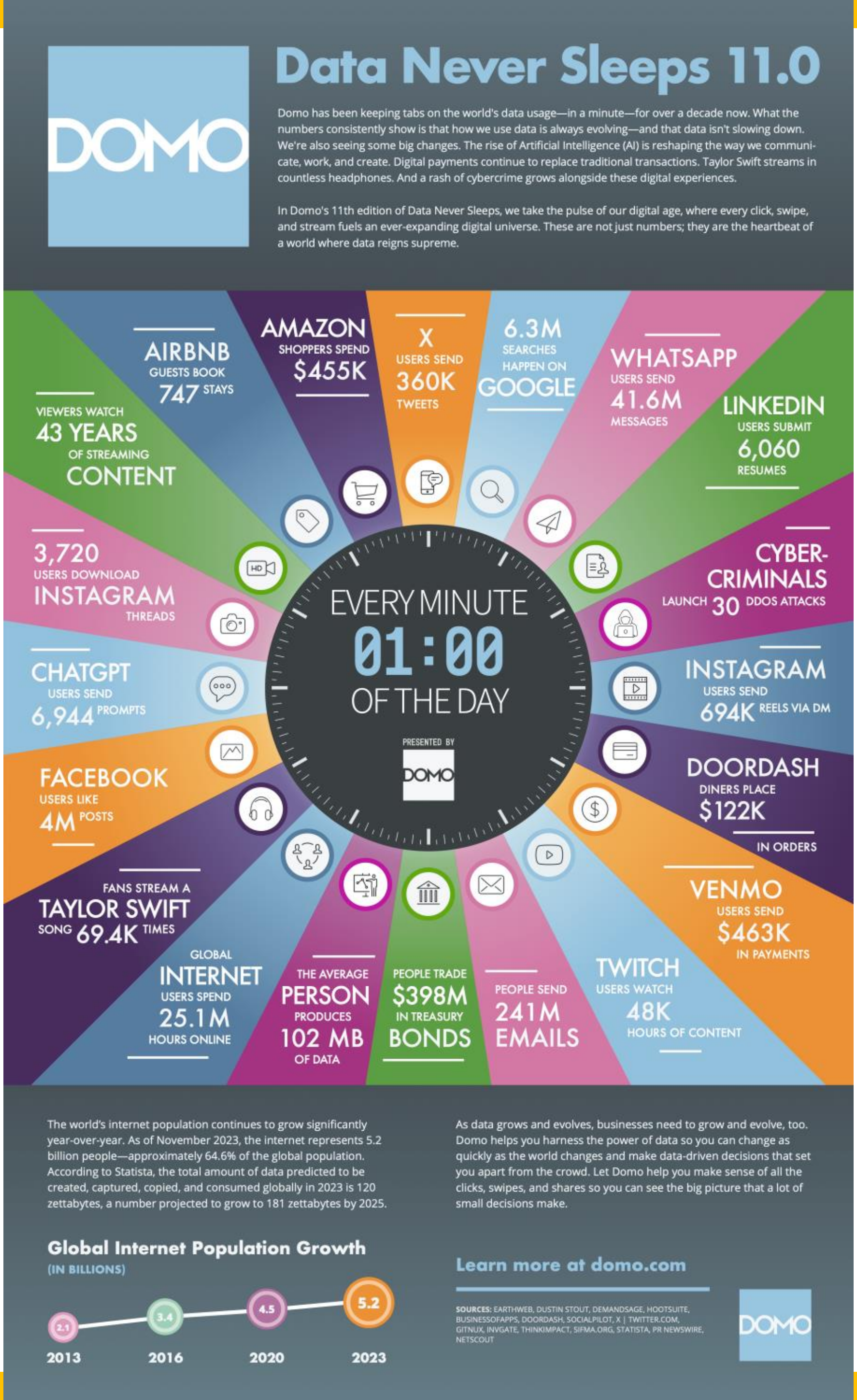
TUMBLR USERS PUBLISH 92,340 POSTS

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/



Data Never Sleeps 11.0

Domo has been keeping tabs on the world's data usage—in a minute—for over a decade now. What the numbers consistently show is that how we use data is always evolving—and that data isn't slowing down. We're also seeing some big changes. The rise of Artificial Intelligence (AI) is reshaping the way we communicate, work, and create. Digital payments continue to replace traditional transactions. Taylor Swift streams in countless headphones. And a rash of cybercrime grows alongside these digital experiences.

In Domo's 11th edition of Data Never Sleeps, we take the pulse of our digital age, where every click, swipe, and stream fuels an ever-expanding digital universe. These are not just numbers; they are the heartbeat of a world where data reigns supreme.

EVERY MINUTE 01:00 OF THE DAY

PRESENTED BY DOMO

AIRBNB GUESTS BOOK 747 STAYS

AMAZON SHOPPERS SPEND $455K

X USERS SEND 360K TWEETS

6.3M SEARCHES HAPPEN ON GOOGLE

WHATSAPP USERS SEND 41.6M MESSAGES

LINKEDIN USERS SUBMIT 6,060 RESUMES

CYBER-CRIMINALS LAUNCH 30 DDOS ATTACKS

INSTAGRAM USERS SEND 694K REELS VIA DM

DOORDASH DINERS PLACE $122K IN ORDERS

VENMO USERS SEND $463K IN PAYMENTS

TWITCH USERS WATCH 48K HOURS OF CONTENT

PEOPLE SEND 241M EMAILS

PEOPLE TRADE $398M IN TREASURY BONDS

THE AVERAGE PERSON PRODUCES 102 MB OF DATA

GLOBAL INTERNET USERS SPEND 25.1M HOURS ONLINE

FANS STREAM A TAYLOR SWIFT SONG 69.4K TIMES

FACEBOOK USERS LIKE 4M POSTS

CHATGPT USERS SEND 6,944 PROMPTS

3,720 USERS DOWNLOAD INSTAGRAM THREADS

VIEWERS WATCH 43 YEARS OF STREAMING CONTENT

The world's internet population continues to grow significantly year-over-year. As of November 2023, the internet represents 5.2 billion people—approximately 64.6% of the global population. According to Statista, the total amount of data predicted to be created, captured, copied, and consumed globally in 2023 is 120 zettabytes, a number projected to grow to 181 zettabytes by 2025.

As data grows and evolves, businesses need to grow and evolve, too. Domo helps you harness the power of data so you can change as quickly as the world changes and make data-driven decisions that set you apart from the crowd. Let Domo help you make sense of all the clicks, swipes, and shares so you can see the big picture that a lot of small decisions make.

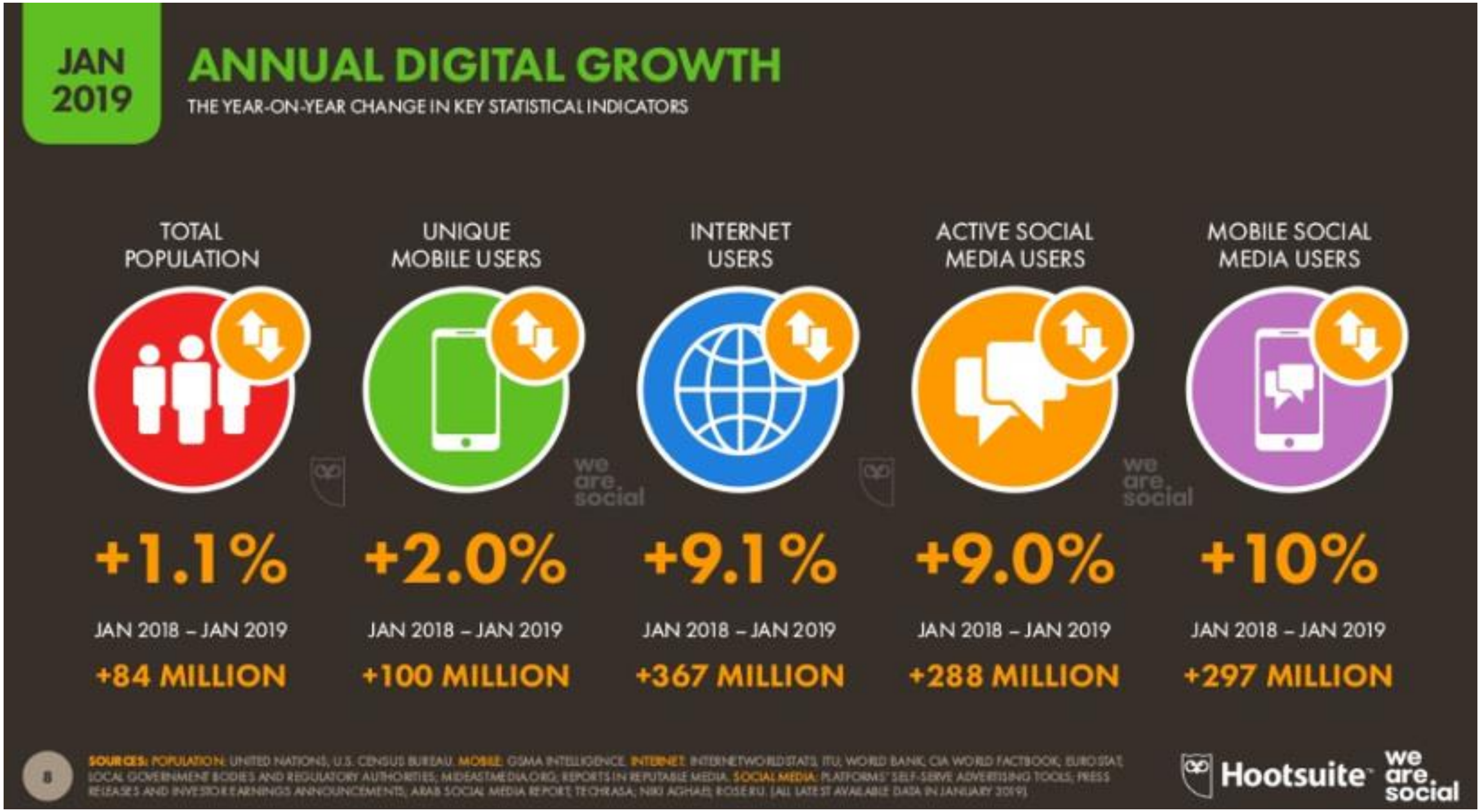
Global Internet Population Growth

(IN BILLIONS)

Learn more at domo.com

SOURCES: EARTHWEB, DUSTIN STOUT, DEMANDSAGE, HOOTSUITE, BUSINESSOFAPPS, DOORDASH, SOCIALPILOT, X | TWITTER.COM, GITNUX, INVGATE, THINKIMPACT, SIFMA.ORG, STATISTA, PR NEWSWIRE, NETSCOUT

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

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).

JAN
2024

DIGITAL GROWTH

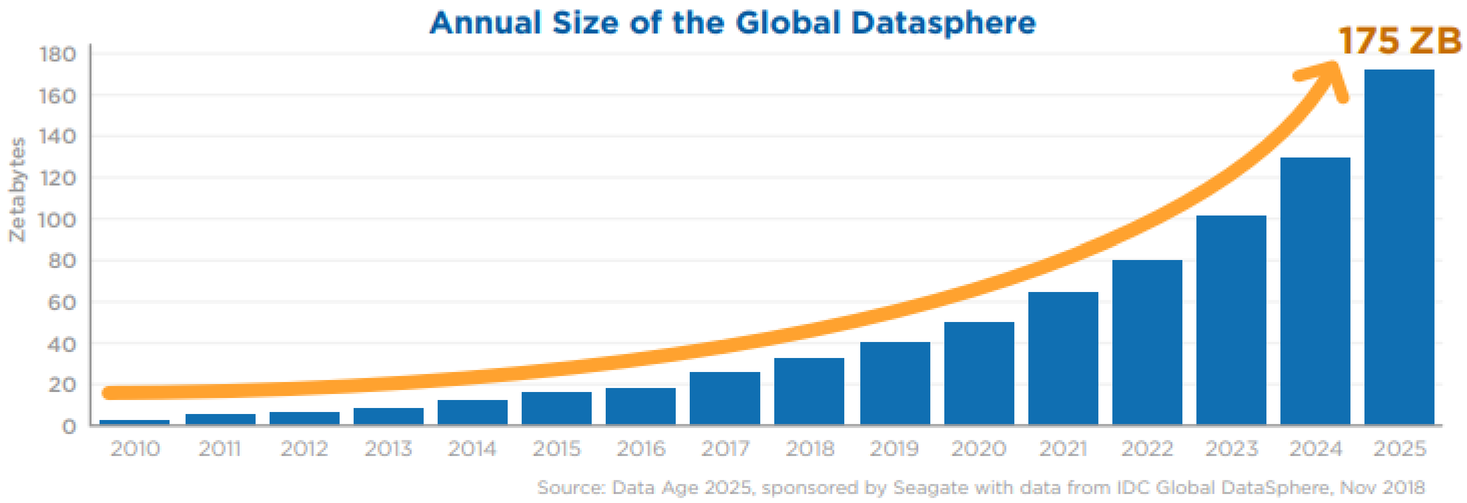
CHANGE IN THE USE OF CONNECTED DEVICES AND SERVICES OVER TIME

GLOBAL OVERVIEW

TOTAL POPULATION	UNIQUE MOBILE PHONE SUBSCRIBERS	INDIVIDUALS USING THE INTERNET	SOCIAL MEDIA USER IDENTITIES
+0.9%	**+2.5%**	**+1.8%**	**+5.6%**
YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE
+74 MILLION	**+138 MILLION**	**+97 MILLION**	**+266 MILLION**

SOURCES: KEPIOS ANALYSIS; UNITED NATIONS; GOVERNMENT RESOURCES; GSMA INTELLIGENCE; ITU; EUROSTAT; CNNIC; KANTAR & IAMAI; PLATFORM RESOURCES; COMPANY EARNINGS REPORTS; OCDH; BETA RESEARCH CENTER. **ADVISORY:** SOCIAL MEDIA USER IDENTITIES MAY NOT REPRESENT UNIQUE INDIVIDUALS. **COMPARABILITY:** BASE REVISIONS; SOURCE CHANGES. SEE NOTES ON DATA.

Data Growth

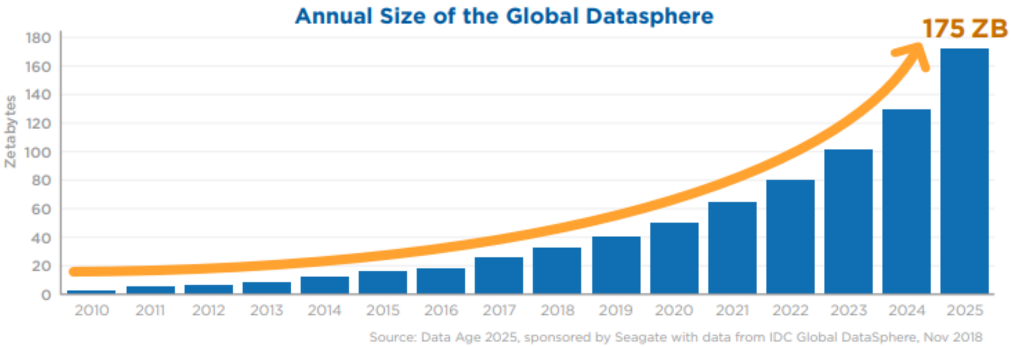


Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

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

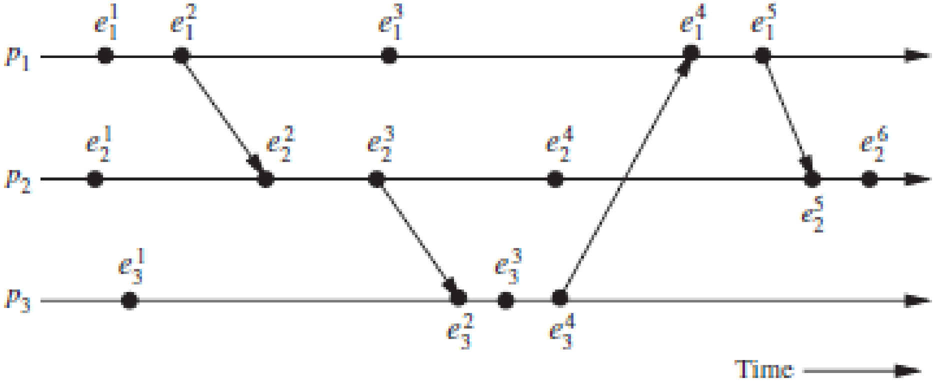


Global datasphere is growing!

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.



Events of a process

process

Processors (p_i) may fail!

Messages may be delayed or lost!!

Course Dynamics

Visit the course website. Bookmark it.