Cloud Computing

Module II – Big Data Processing

Nicola Tonellotto
University of Pisa
nicola.tonellotto@unipi.it

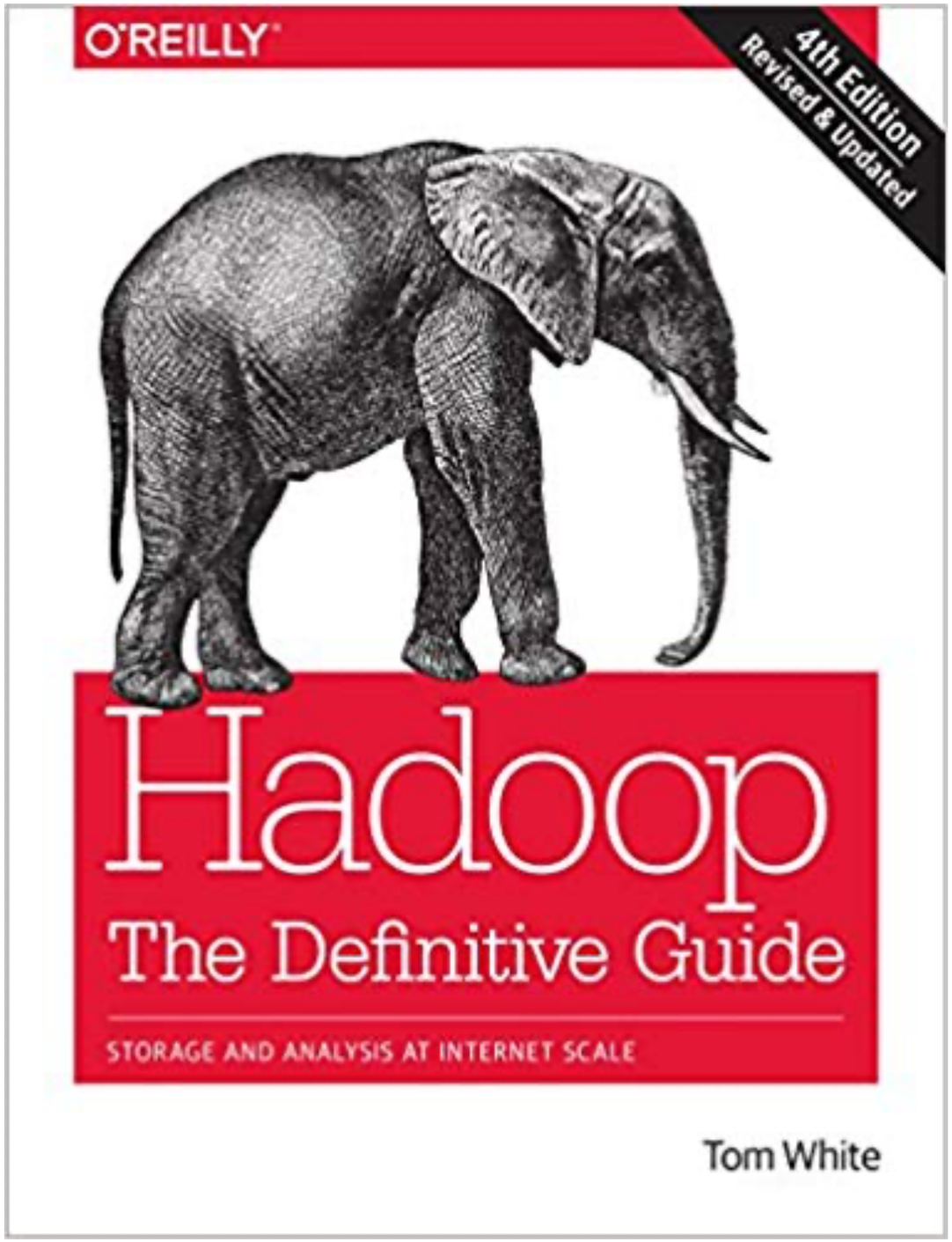
What do I expect from you

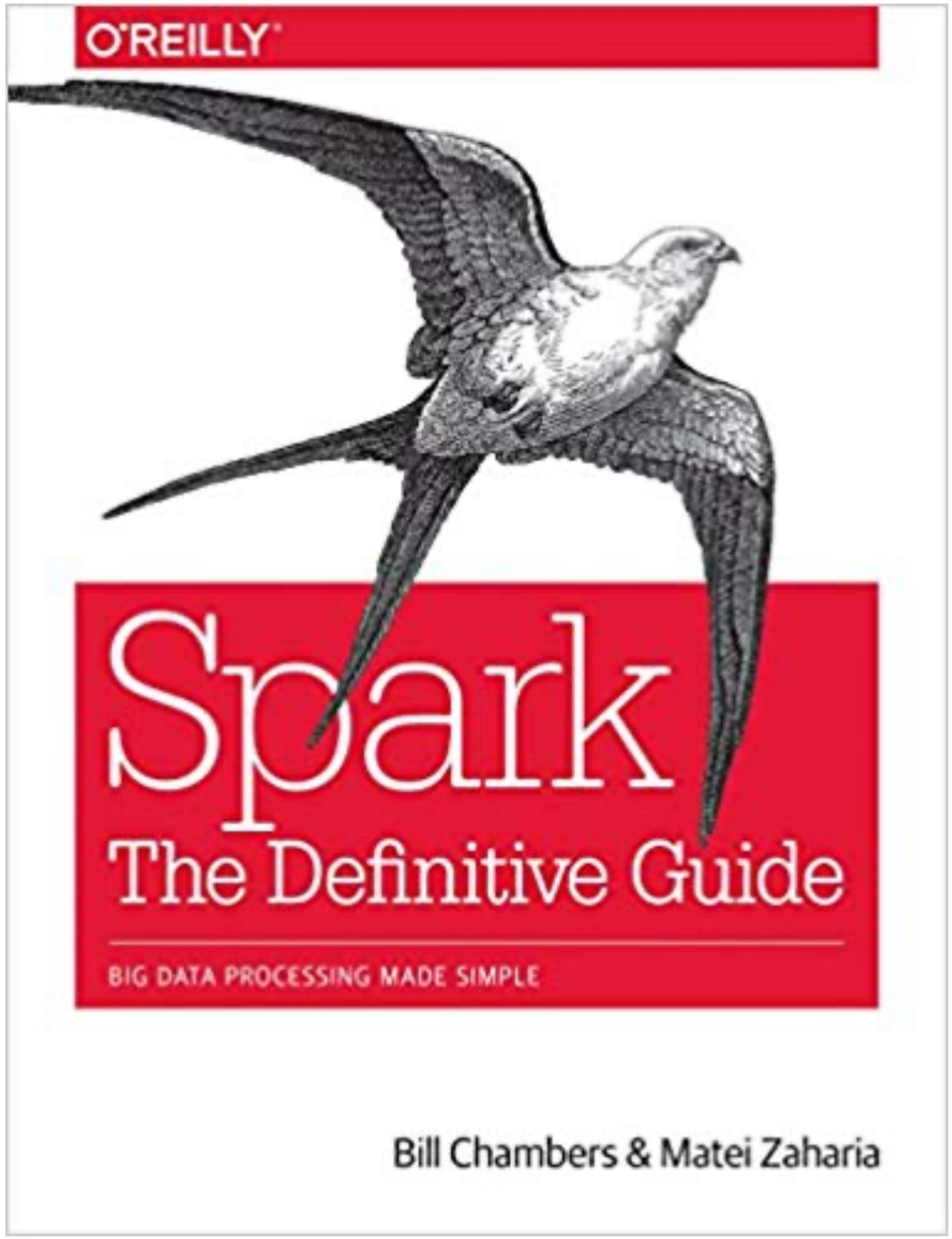
- You are already a good Java programmer
  - This module does not teach programming
  - You are expected to use profitably Java documentation
- You know how to manage a Linux machine
  - You are expected to install software packages
  - Compiling, patching, and installing open source software
- You have basic knowledge of:
  - Probability and statistics
  - Computer architecture
  - Fundamental data structures and algorithms
  - Operating systems
- You have a personal laptop (BYOD lab sessions)

What you will get at the end

- You will:
  - improve your programming skills;
  - be able to process big data to generate new knowledge with the MapReduce paradigm;
  - master tools, best practices and common procedures to design and implement programs on Hadoop;
  - master tools, best practices and common procedures to design and implement program on Spark.

Suggested Textbooks

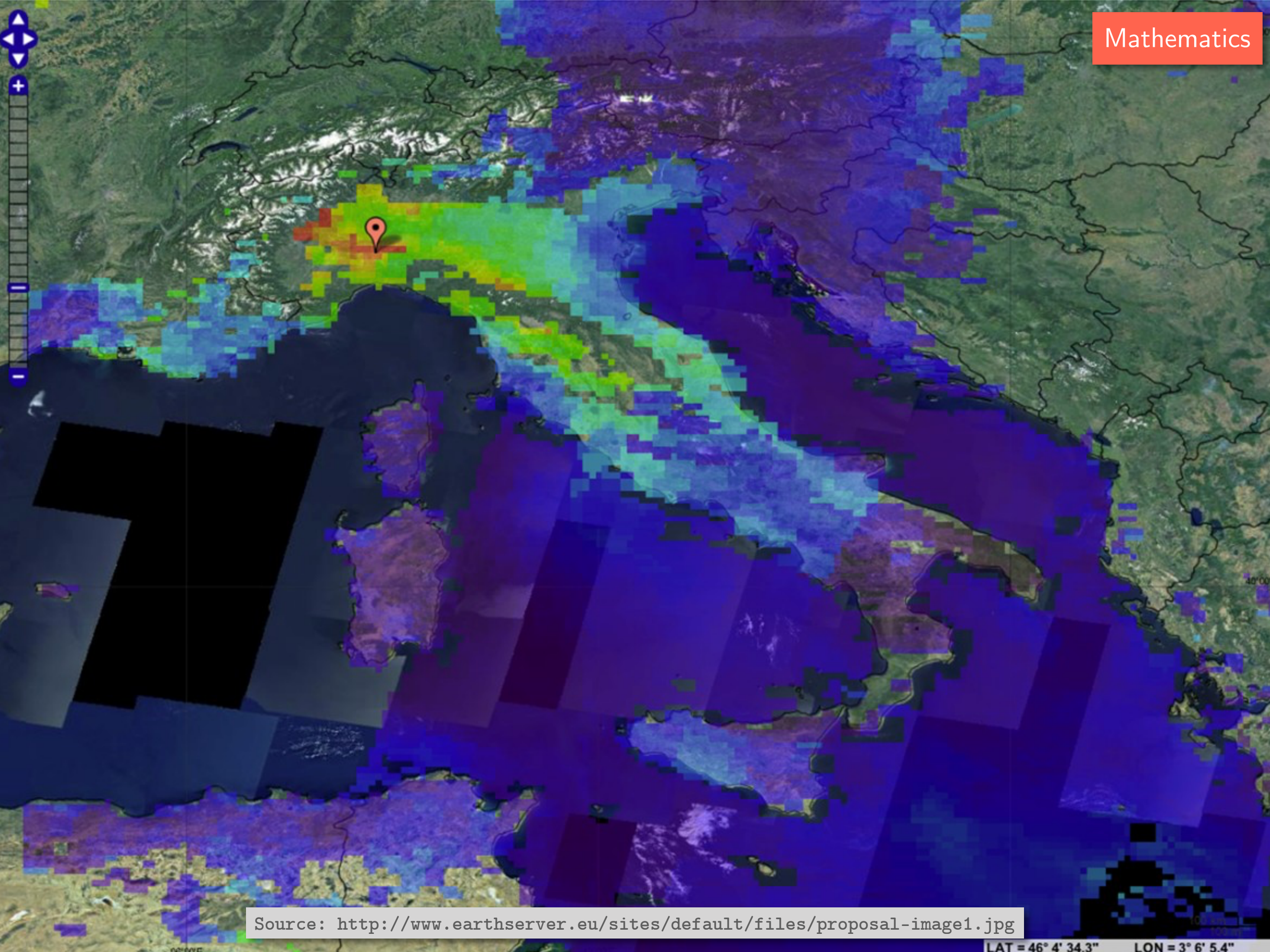




Read the Javadocs!!!

Introduction

Source: https://upload.wikimedia.org/wikipedia/commons/e/eb/Views_of_the_LHC_tunnel_sector_3-4,_tirage_1.jpg



Source: http://www.earthserver.eu/sites/default/files/proposal-image1.jpg

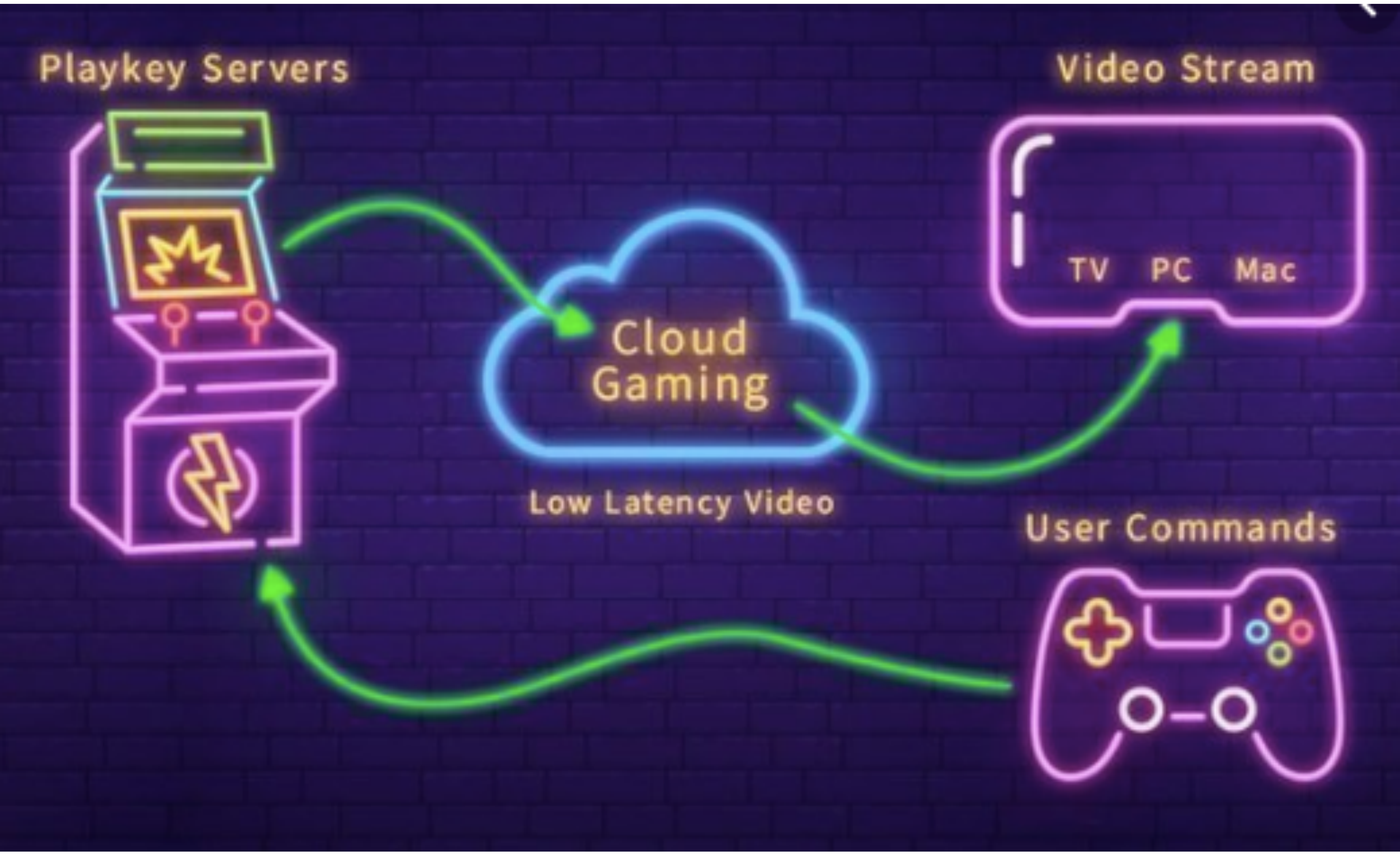
Finance

Source: https://upload.wikimedia.org/wikipedia/commons/1/18/NASDAQ_studio.jpg

Source: https://upload.wikimedia.org/wikipedia/commons/1/19/Times_Square%2C_New_York_City_%28HDR%29.jpg

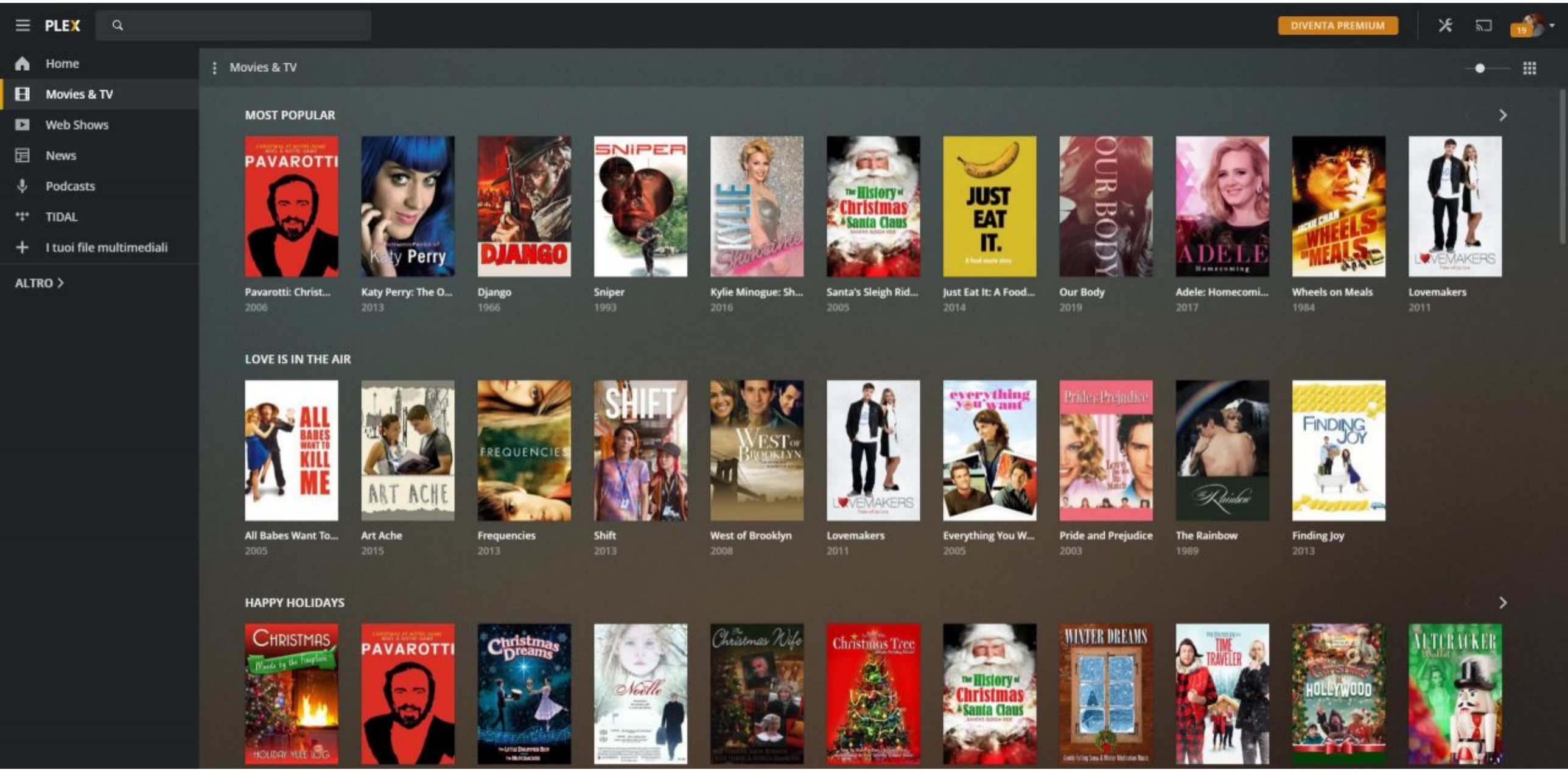
Events

Source: https://upload.wikimedia.org/wikipedia/commons/2/24/Huge_crowd_turns_out_for_MTV_EXIT_concert_against_human_trafficking_and_exploitation.jpg



Source: https://www.sologames.it/wp-content/uploads/2019/11/Cloud-gaming.png

Videos



Source: https://www.smartworld.it/wp-content/uploads/2019/12/plex-movie-e-tv-1280x625.jpg

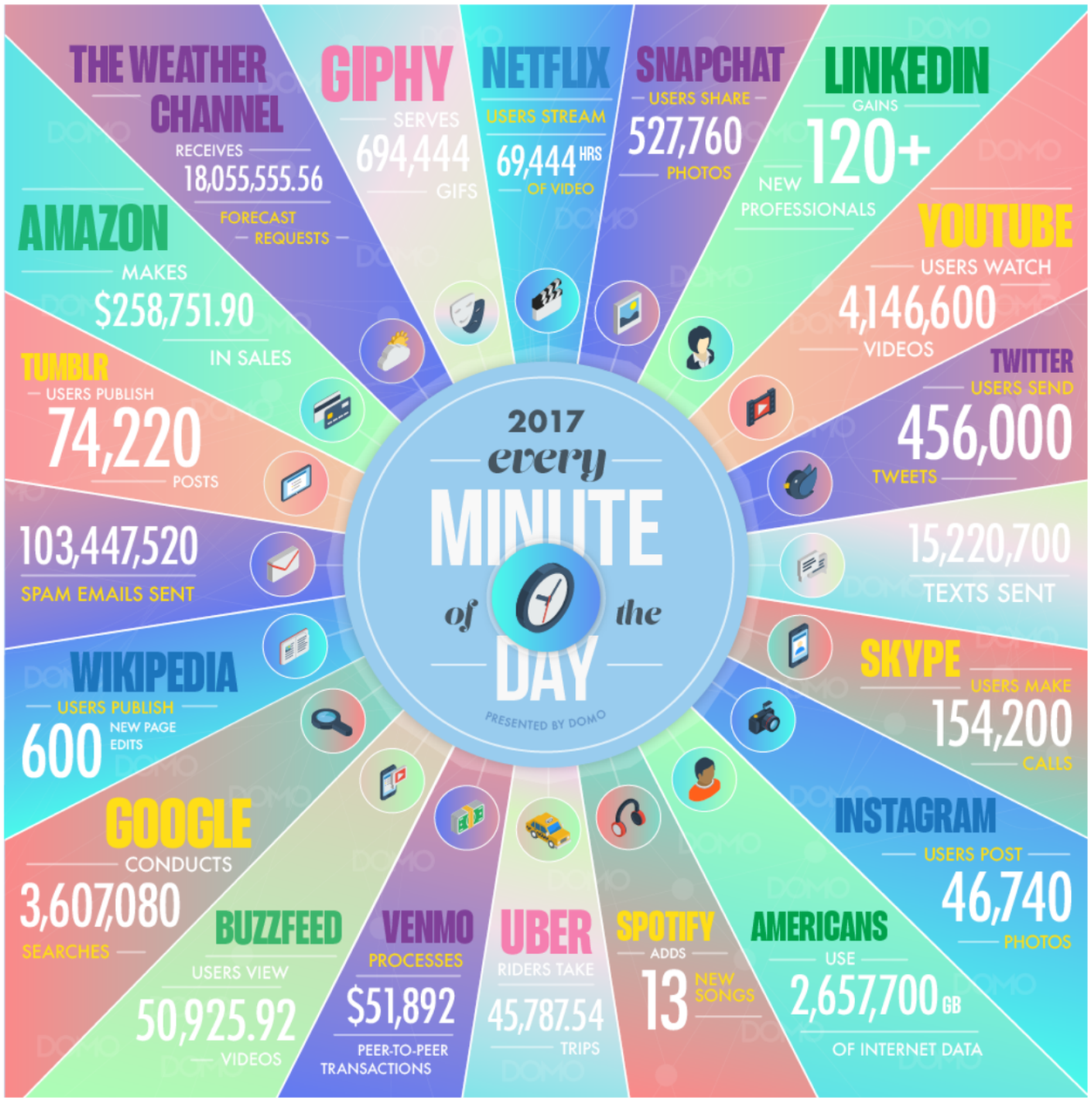
AI

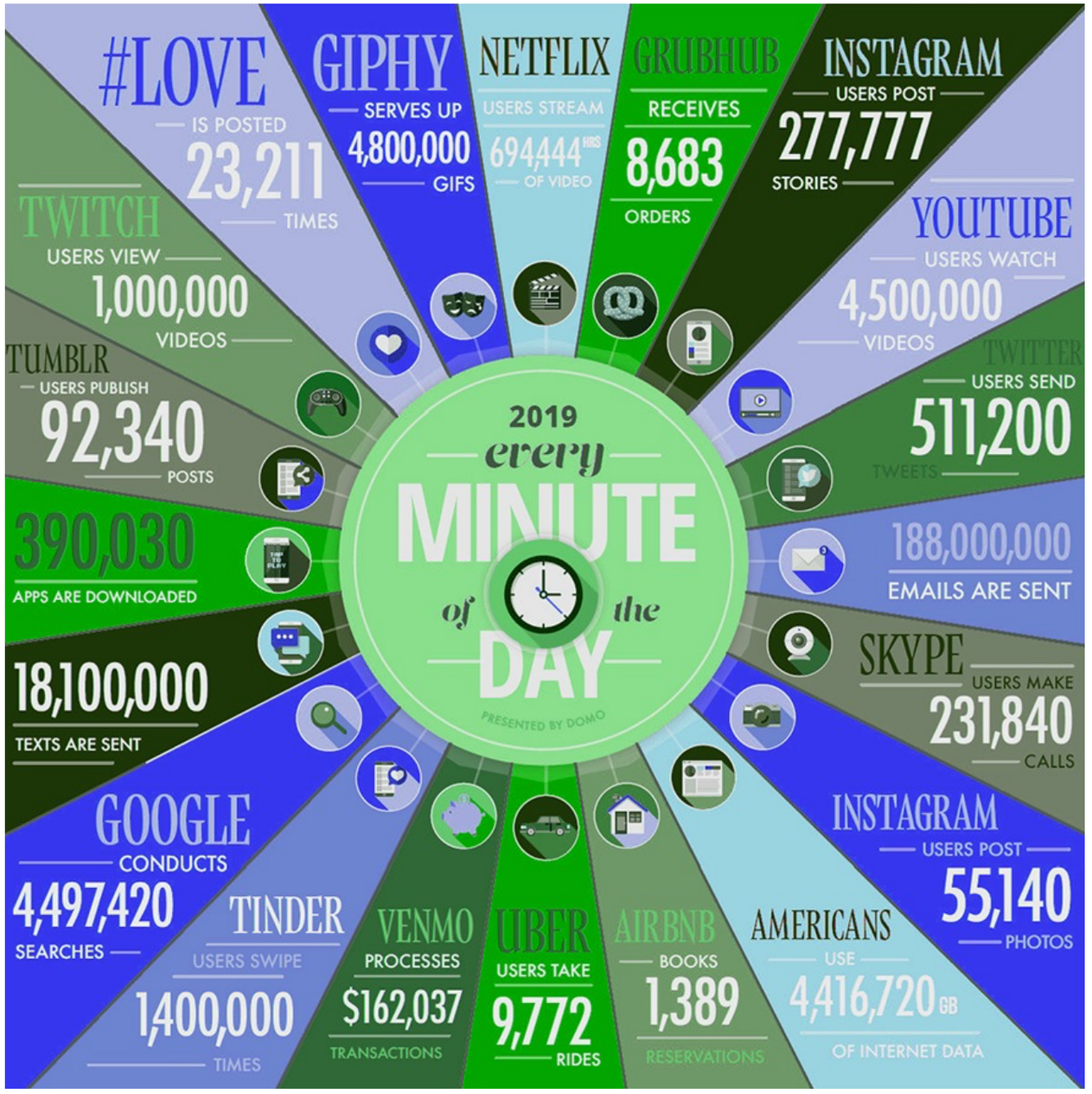
Type of Neural Network	Parameters (MiB)	Training			Inference
		Examples to Convergence	ExaOps to Conv	Ops per Example	Ops per Example
MLP0	225	1 trillion	353	353 Mops	118 Mops
MLP1	40	650 billion	86	133 Mops	44 Mops
LSTM0	498	1.4 billion	42	29 Gops	9.8 Gops
LSTM1	800	656 million	82	126 Gops	42 Gops
CNN0	87	1.64 billion	70	44 Gops	15 Gops
CNN1	104	204 million	7	34 Gops	11 Gops
ResNet	98	114 million	<3	23 Gops	8 Gops

Source: Barroso et al., "The Datacenter as a Computer" 3rd ed, Morgan & Claypool Publishers, 2019

How much data?

Source: https://sparkdatabox.com/blog/wp-content/uploads/2019/10/IMG-20191017-WA0004-1.jpg



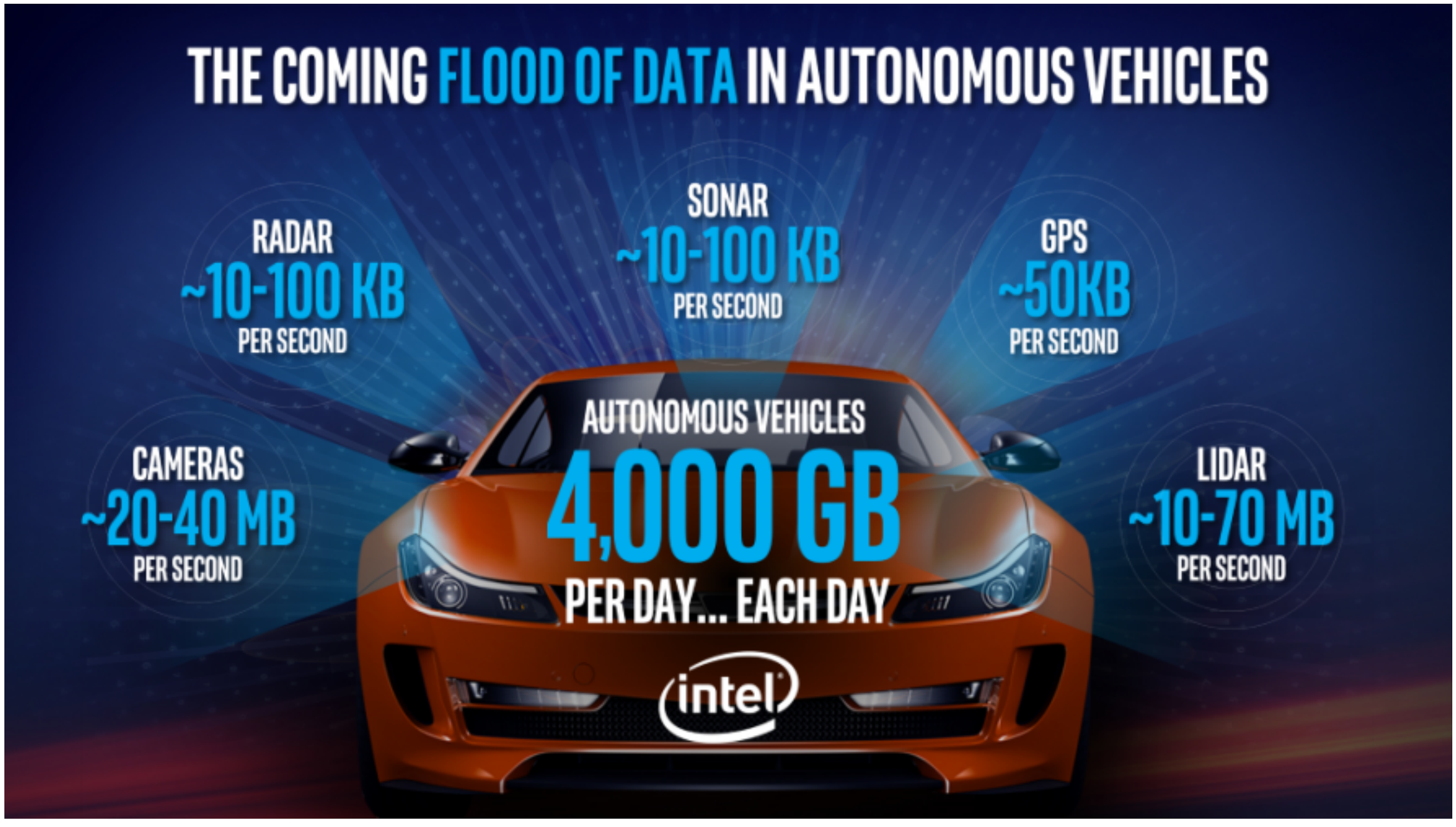


Source: https://web-assets.domo.com/blog/wp-content/uploads/2017/07/17_domo_data-never-sleeps-5-01.png

Connected humans: Web & Social media

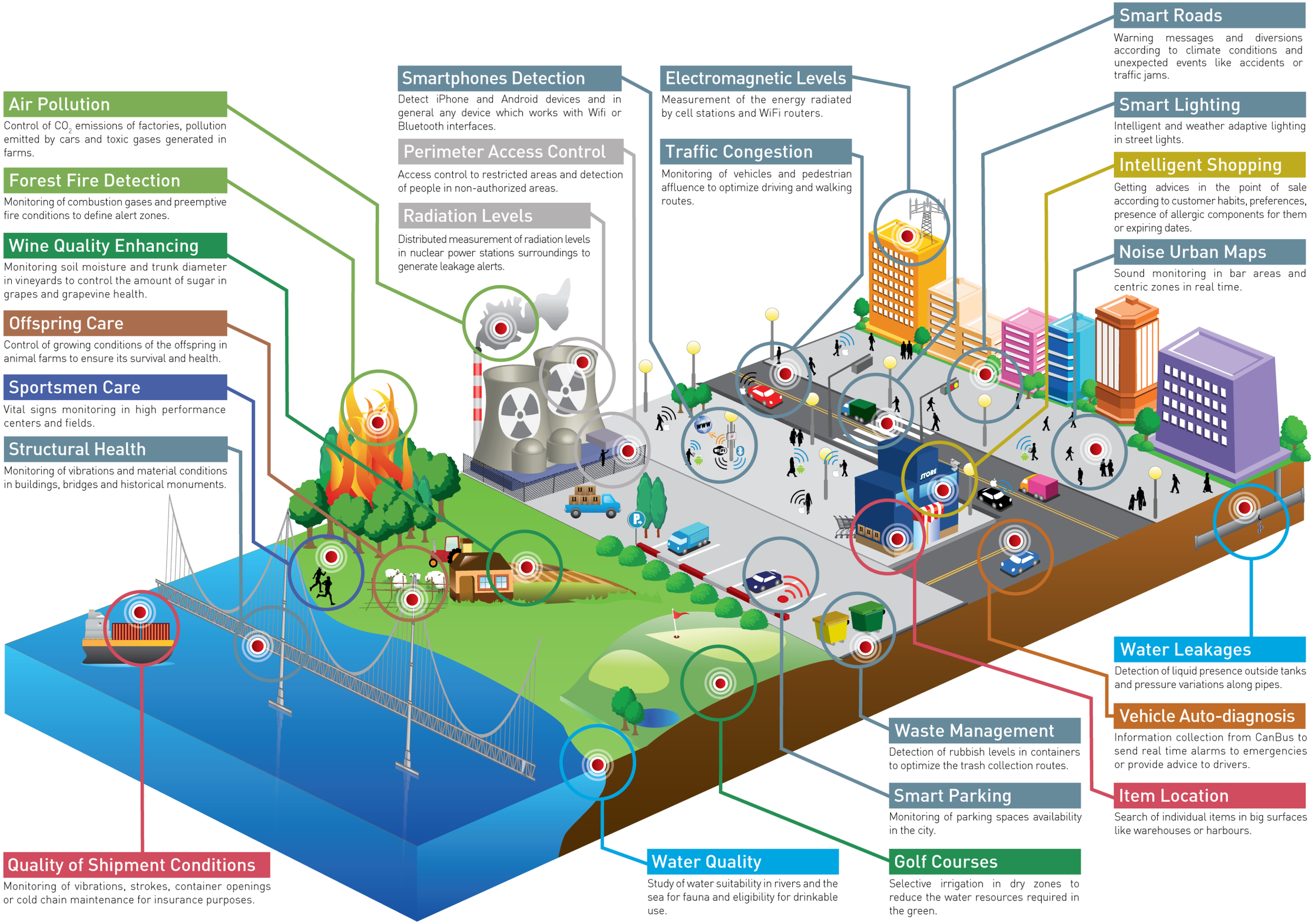
- Google carries 4,497,420 searches
- Youtube users view 4,500,000 videos
- Twitter users post 511,200 tweets
- Instagram users post 277,777 tales
- Skype users make 231,840 calls
- Uber users take 9,772 drives
- Netflix users stream 694,444 hours of video
- Giphy toils up 4,800,000 gifs
- Airbnb books 1,389 reservations
- Tinder users swipe 1,400,000 times
- Twitch users view 1,000,000 videos

Internet of Things



Source: https://newsroom.intel.com/editorials/krzanich-the-future-of-automated-driving/#gs.r10kgf

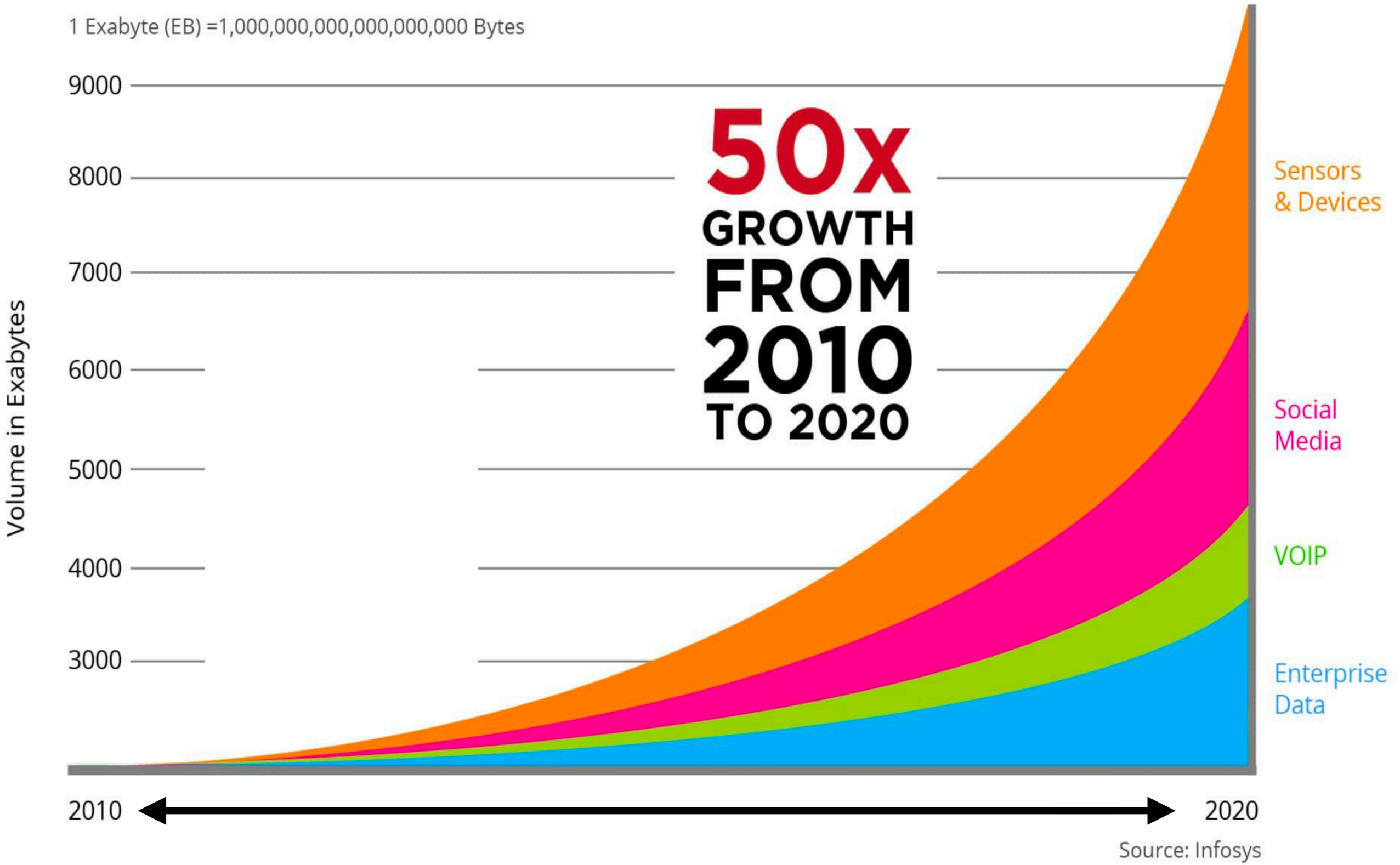
Smart World



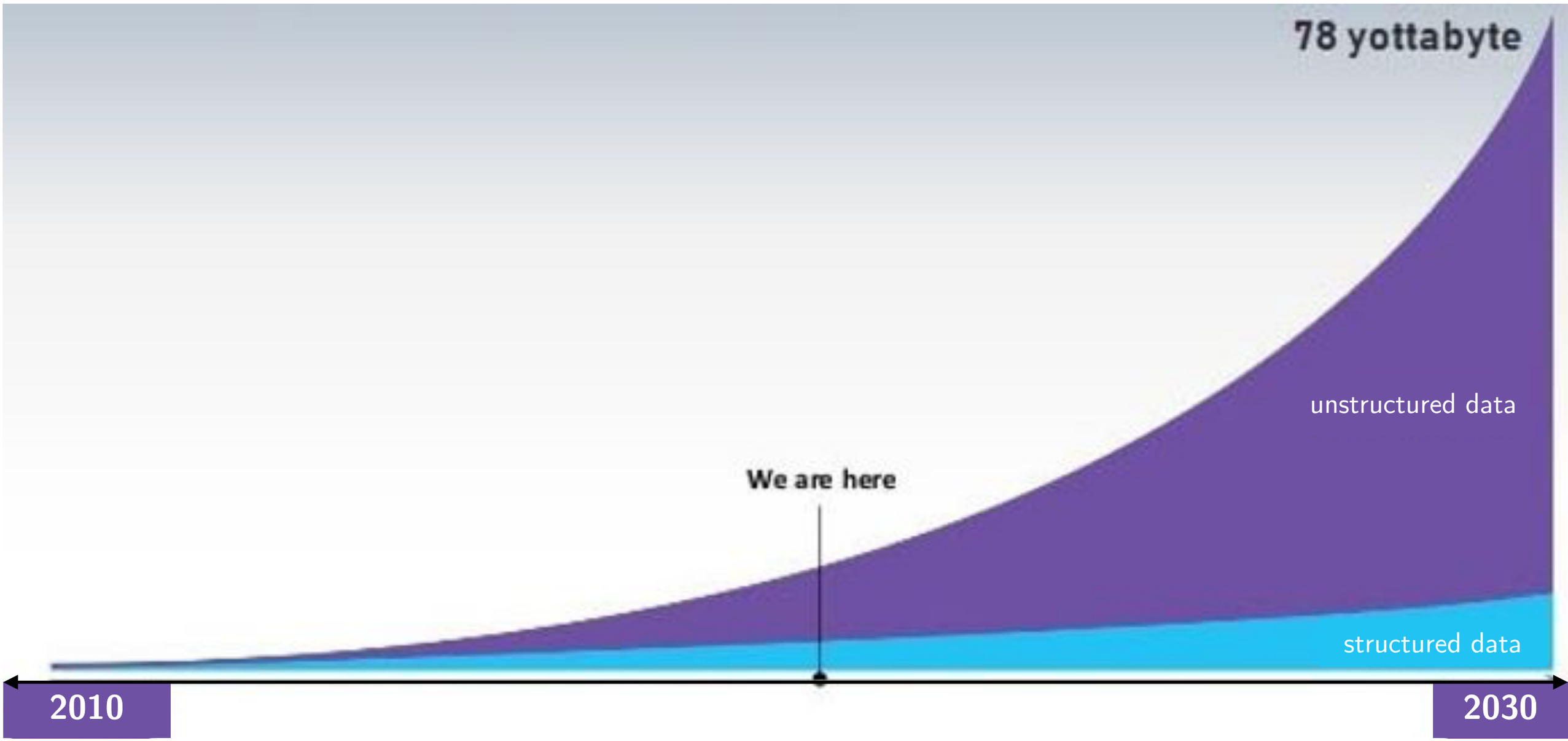
Source: http://www.libelium.com/wp-content/themes/libelium/images/content/applications/libelium_smart_world_infographic_big.png

Connected objects: Internet of Things

- 400 million IoT devices with cellular connections at the end of 2016
- 1.5 billion IoT devices with cellular connections in 2022
- 144.4 million smart homes in the US
- 53 billion US dollars for the global smart home market by 2022
- 79.4 zettabytes of data generated by devices in 2025

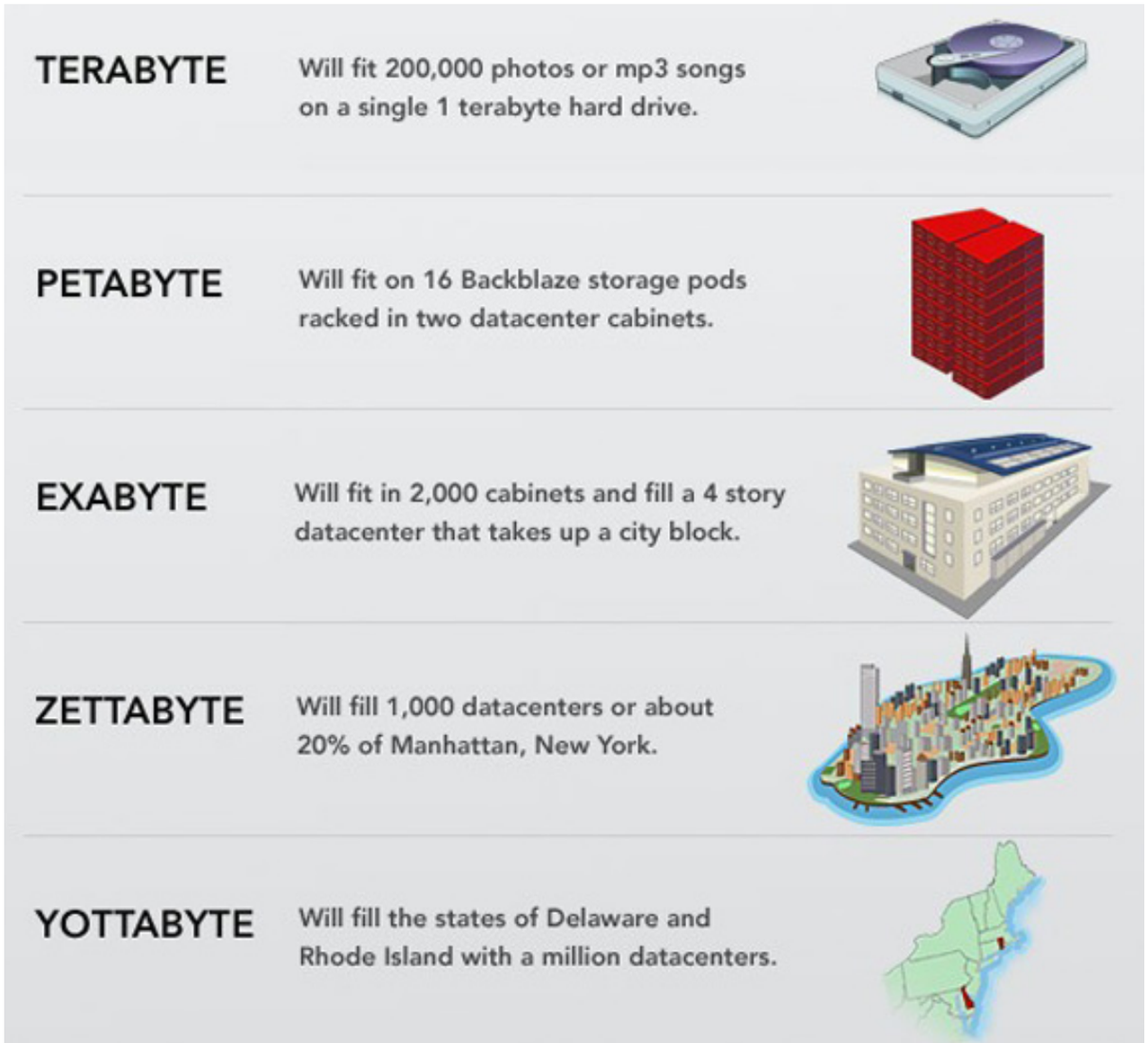


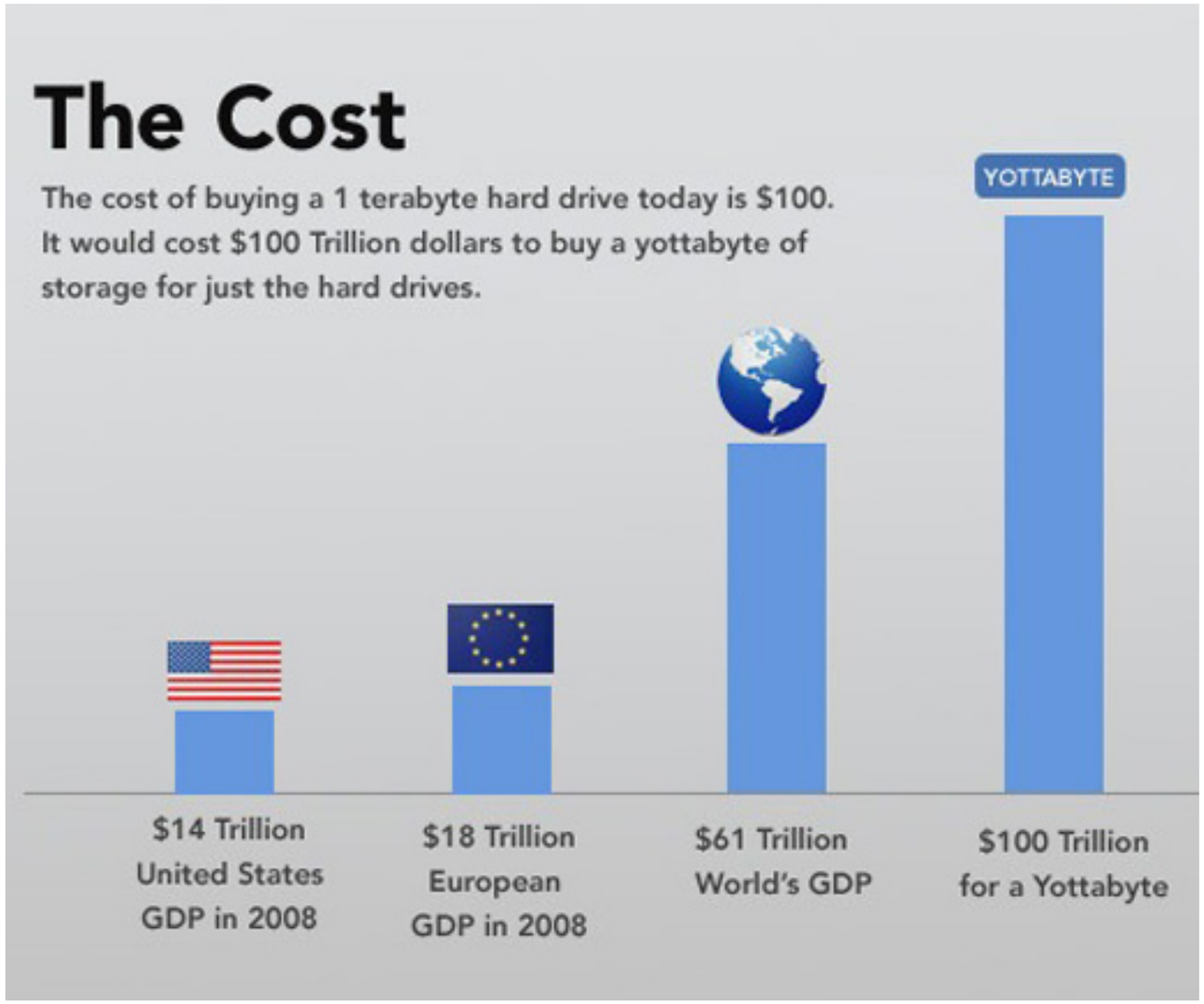
Source: https://therbootcamp.github.io/ML_2019May/_sessions/LookAhead/LookAhead.html#3



Source: https://static.wixstatic.com/media/fd6c49_f00ea2e3f0e24d7e97589d77cd5e3f2a~mv2.jpg/v1/fill/w_689,h_323,al_c,lg_1,q_80/big1.jpg

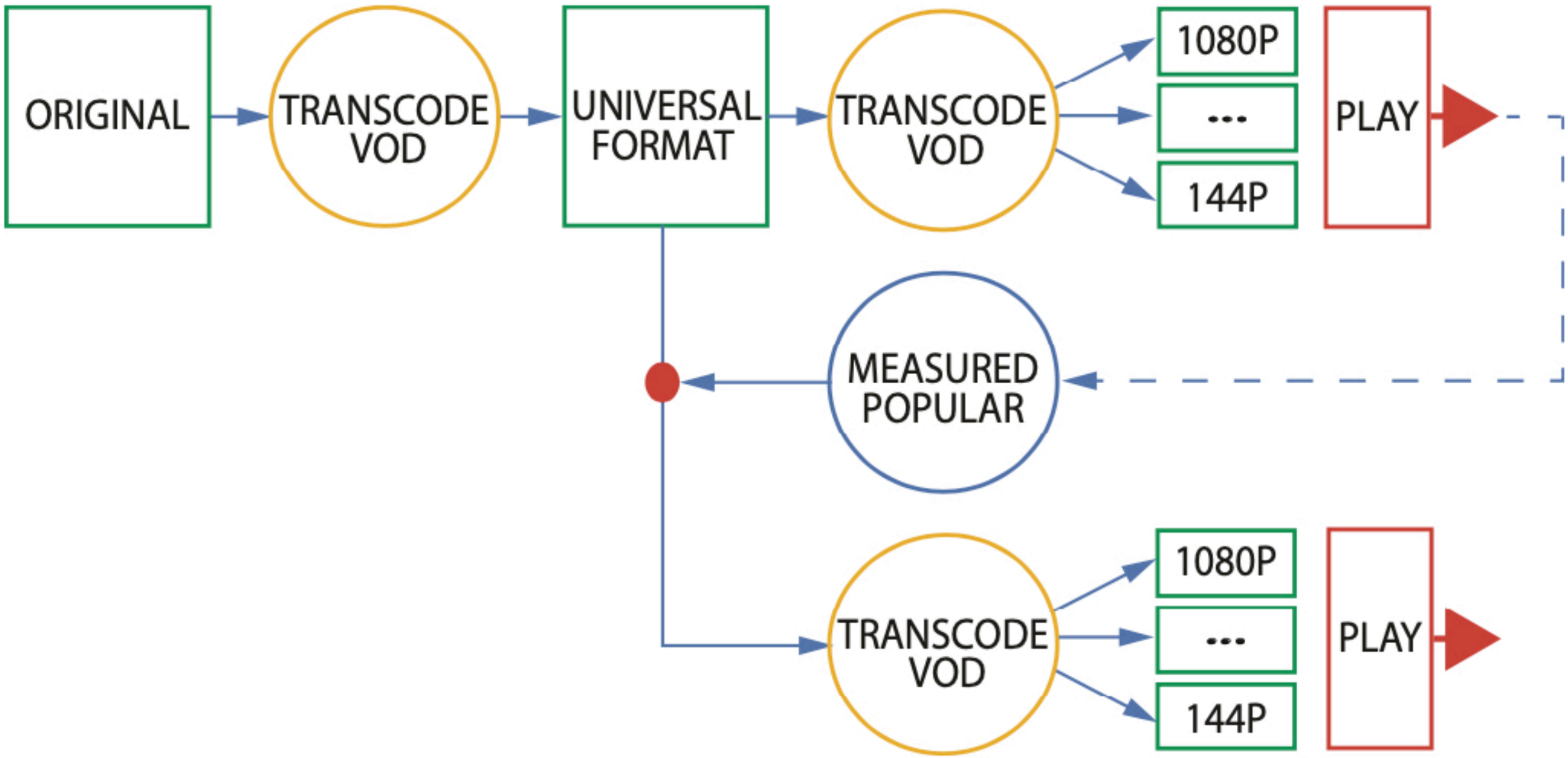
How big is a yottabyte?





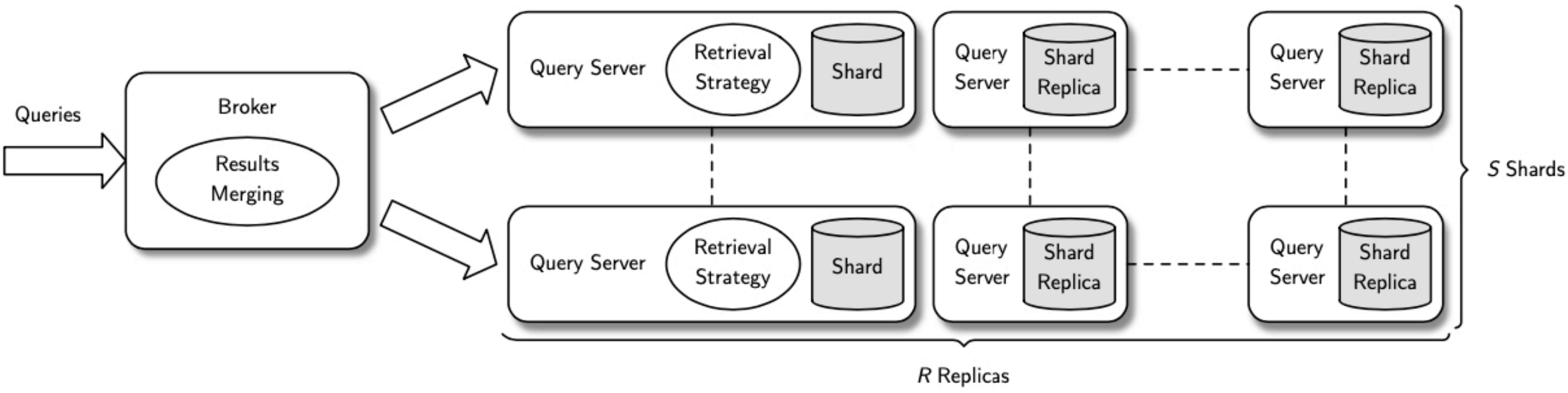
Source: https://cdn.theatlantic.com/static/mt/assets/science/Yottabyte-Post.jpg

The YouTube Video Processing Pipeline



Source: Barroso et al., "The Datacenter as a Computer" 3rd ed, Morgan & Claypool Publishers, 2019

The Web Search Engine Architecture



Source: Tonellotto et al., "Efficient Query Processing for Scalable Web Search", NOW, 2019

Partition/Aggregate Architecture