

Cloud Computing

Module II – Big Data Processing

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

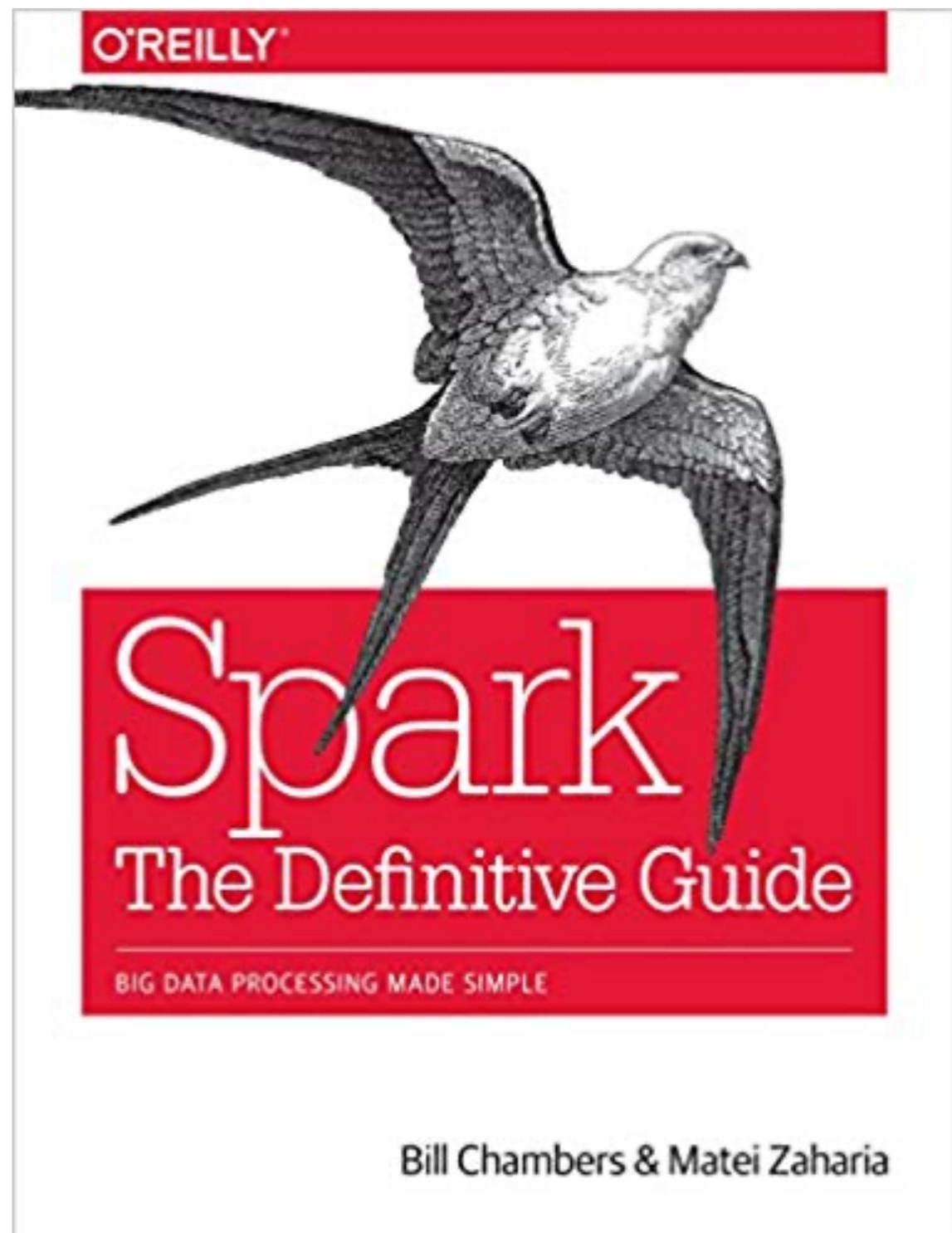
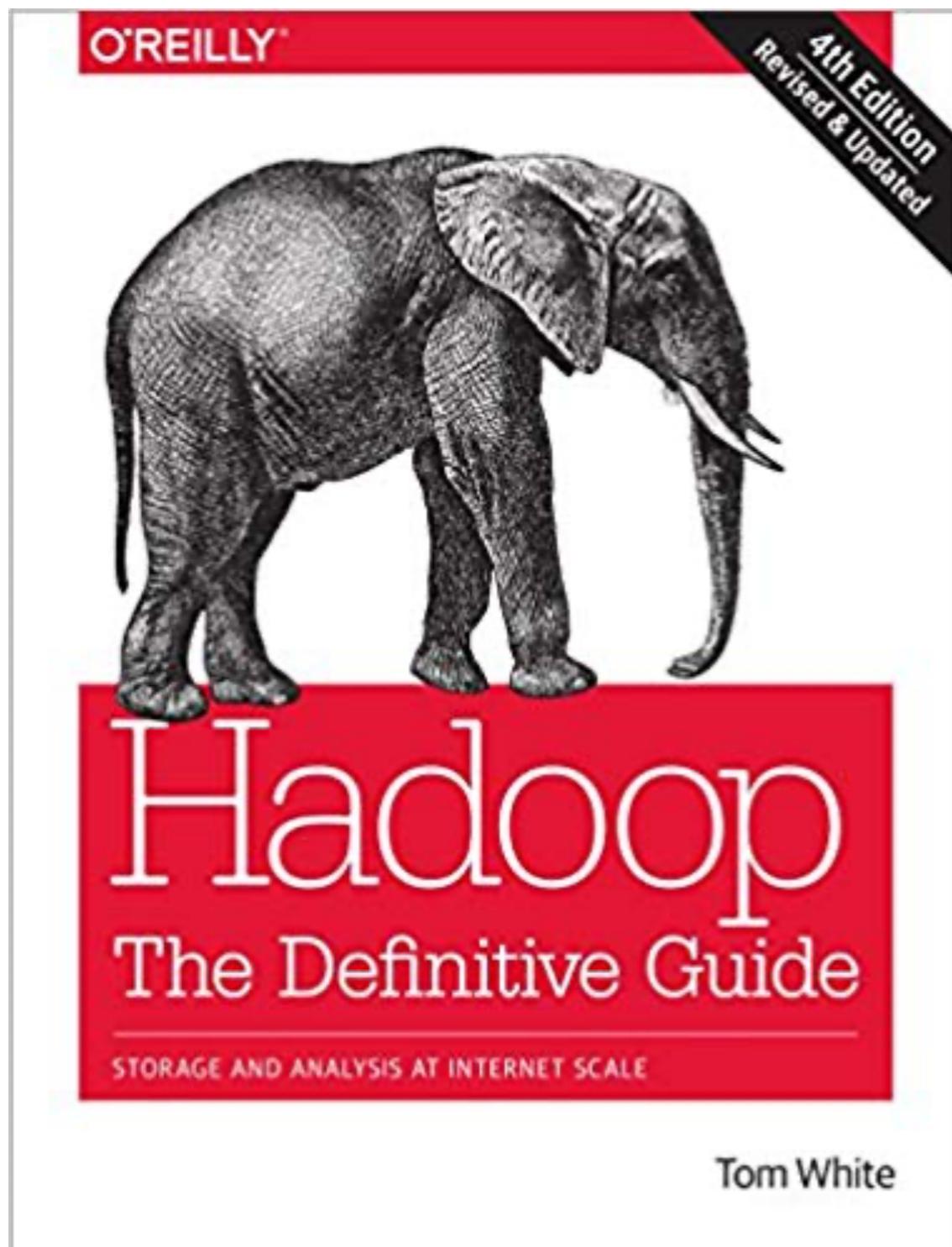
What do I expect from you

- 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

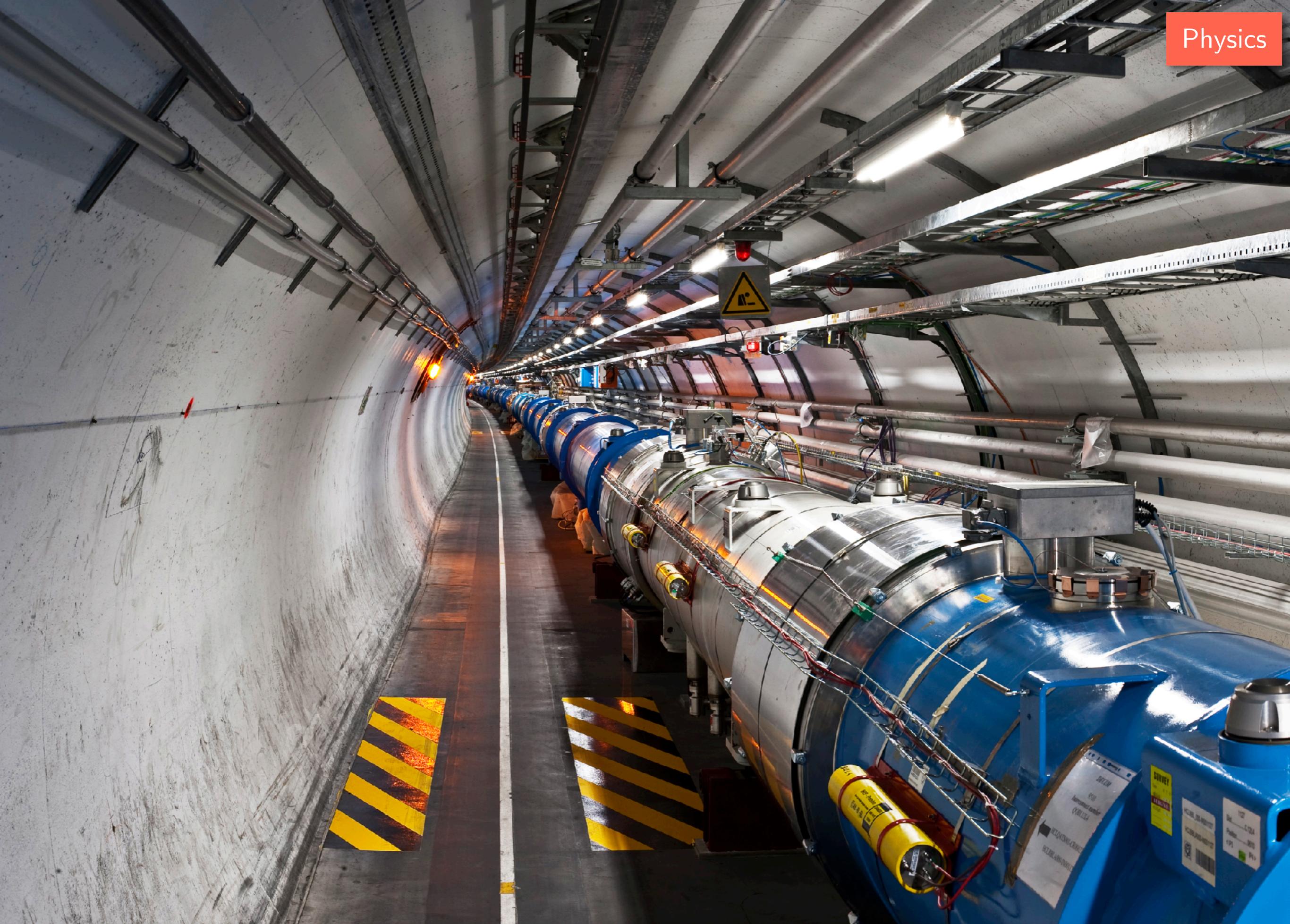
- You will:
 - improve your programming skills;
 - be able to process big data to generate new knowledge with the MapReduce paradigm;
 - master tools, best practices and common procedures to design and implement programs on Hadoop;
 - master tools, best practices and common procedures to design and implement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

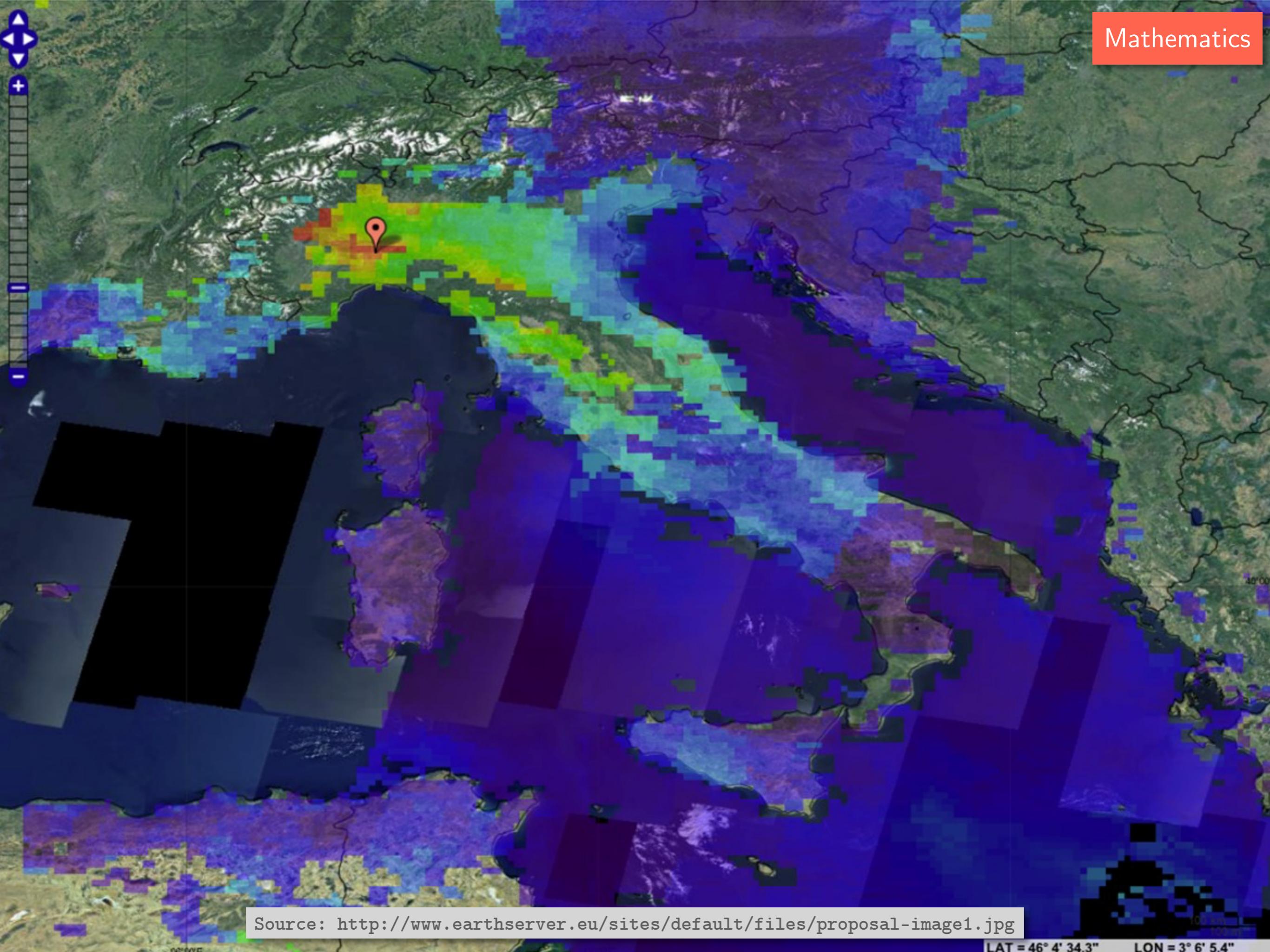
Suggested Textbooks



Read the Javadocs!!!

Introduction



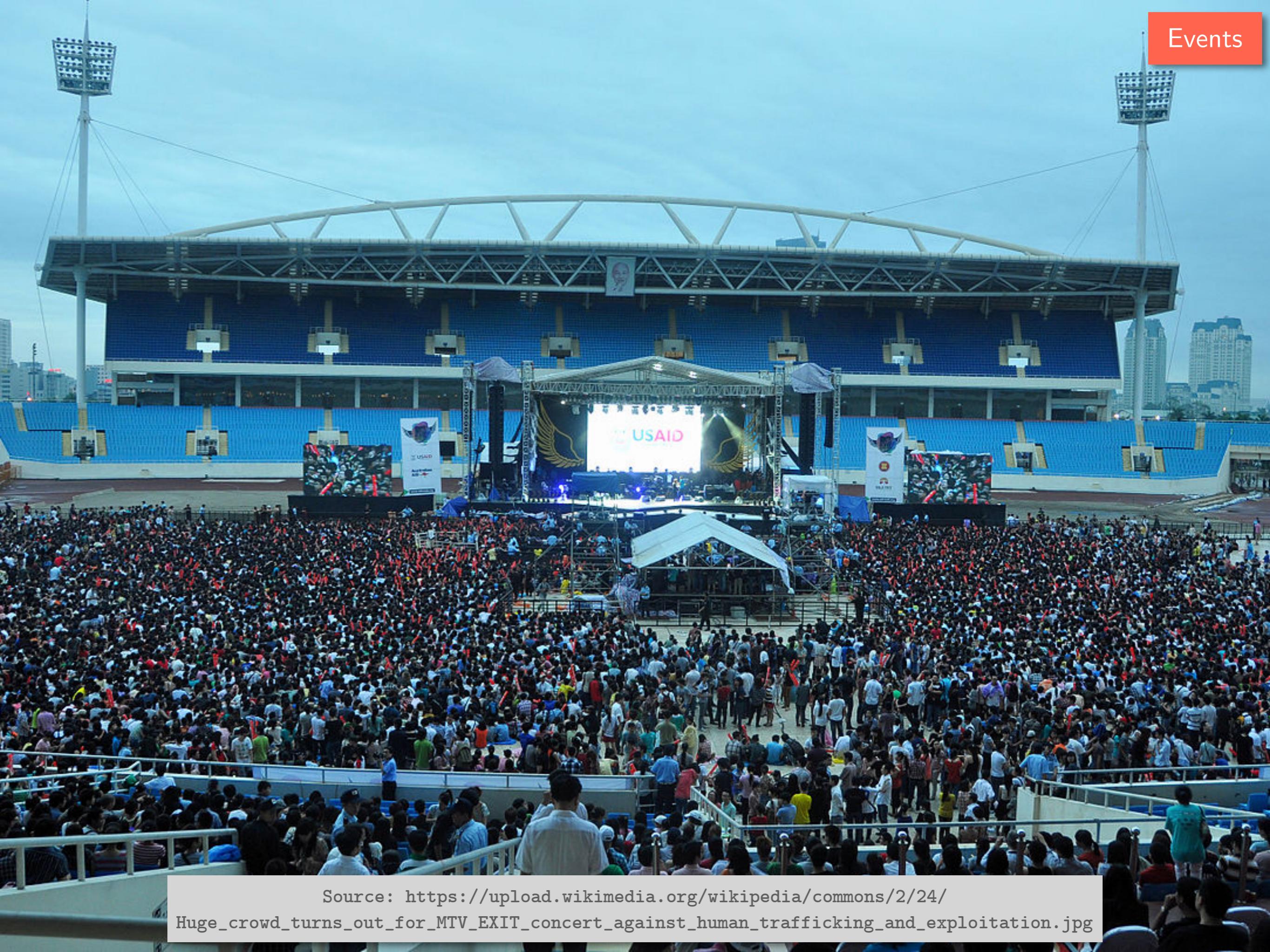


Source: <http://www.earthserver.eu/sites/default/files/proposal-image1.jpg>

LAT = 46° 4' 34.3" LON = 3° 6' 5.4"





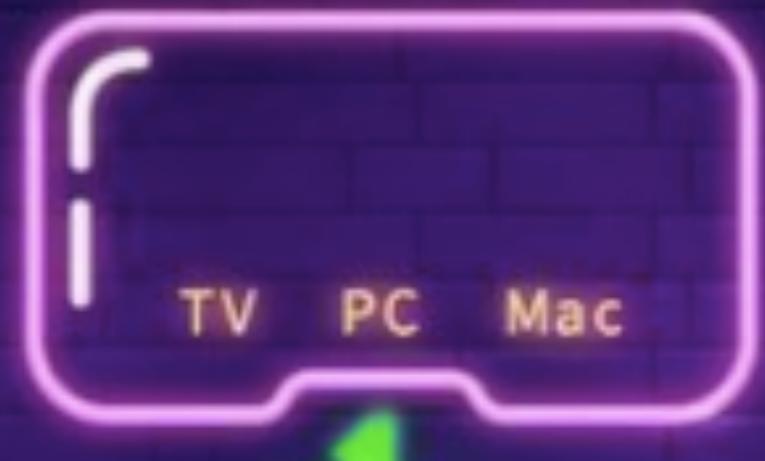


Source: https://upload.wikimedia.org/wikipedia/commons/2/24/Huge_crowd_turns_out_for_MTV_EXIT_concert_against_human_trafficking_and_exploitation.jpg

Playkey Servers



Video Stream



Cloud
Gaming

Low Latency Video

User Commands



PLEX

Movies & TV

DIVENTA PREMIUM

MOST POPULAR

Pavarotti: Christ...	Katy Perry: The O...	Django	Sniper	Kylie Minogue: Sh...	Santa's Sleigh Rid...	Just Eat It: A Food...	Our Body	Adele: Homecomi...	Wheels on Meals	Lovemakers
2006	2013	1966	1993	2016	2005	2014	2019	2017	1984	2011

LOVE IS IN THE AIR

All Babes Want To...	Art Ache	Frequencies	Shift	West of Brooklyn	Lovemakers	everything you want	Pride and Prejudice	The Rainbow	Finding Joy
2005	2015	2013	2013	2008	2011	2005	2003	1989	2013

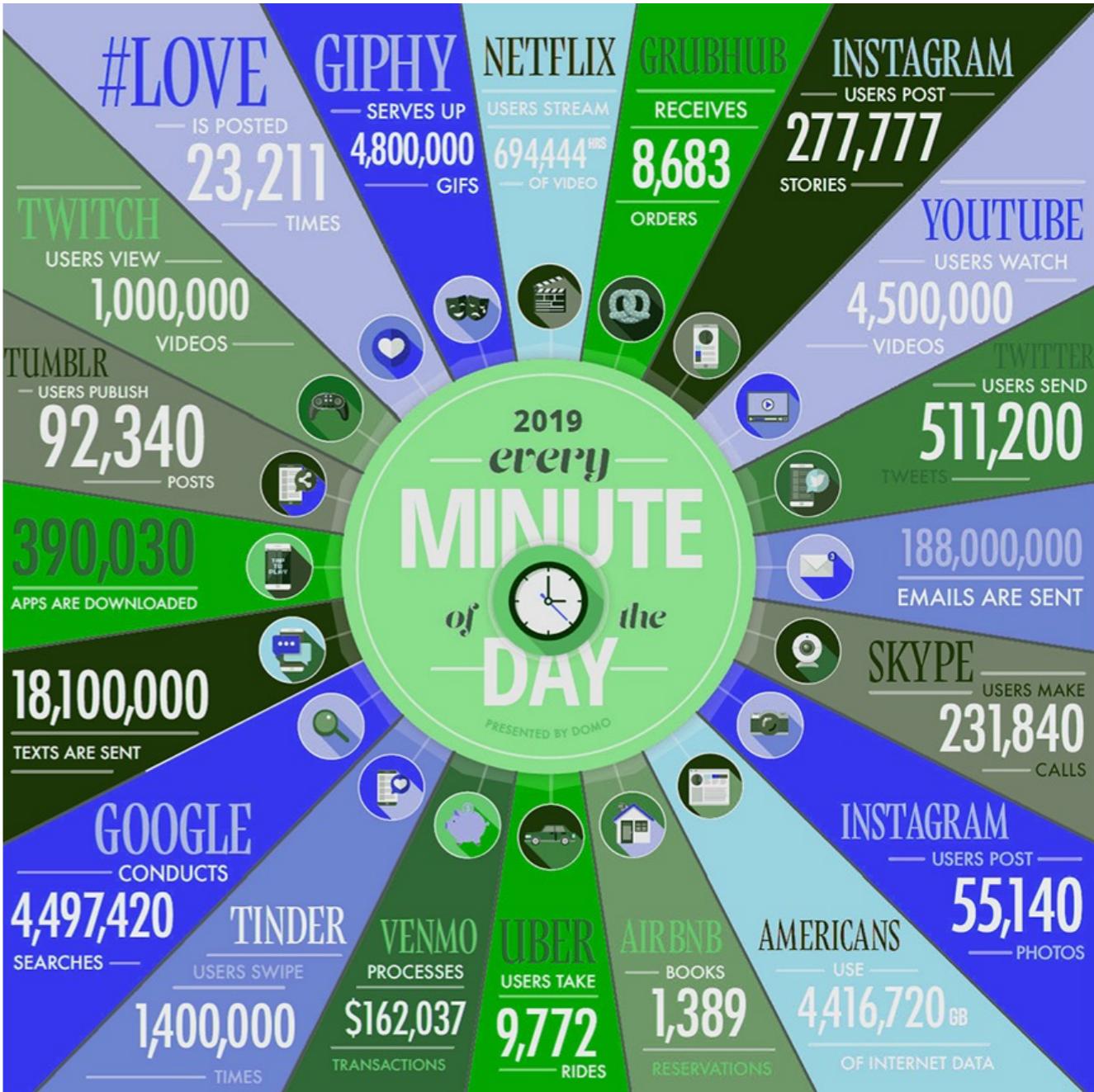
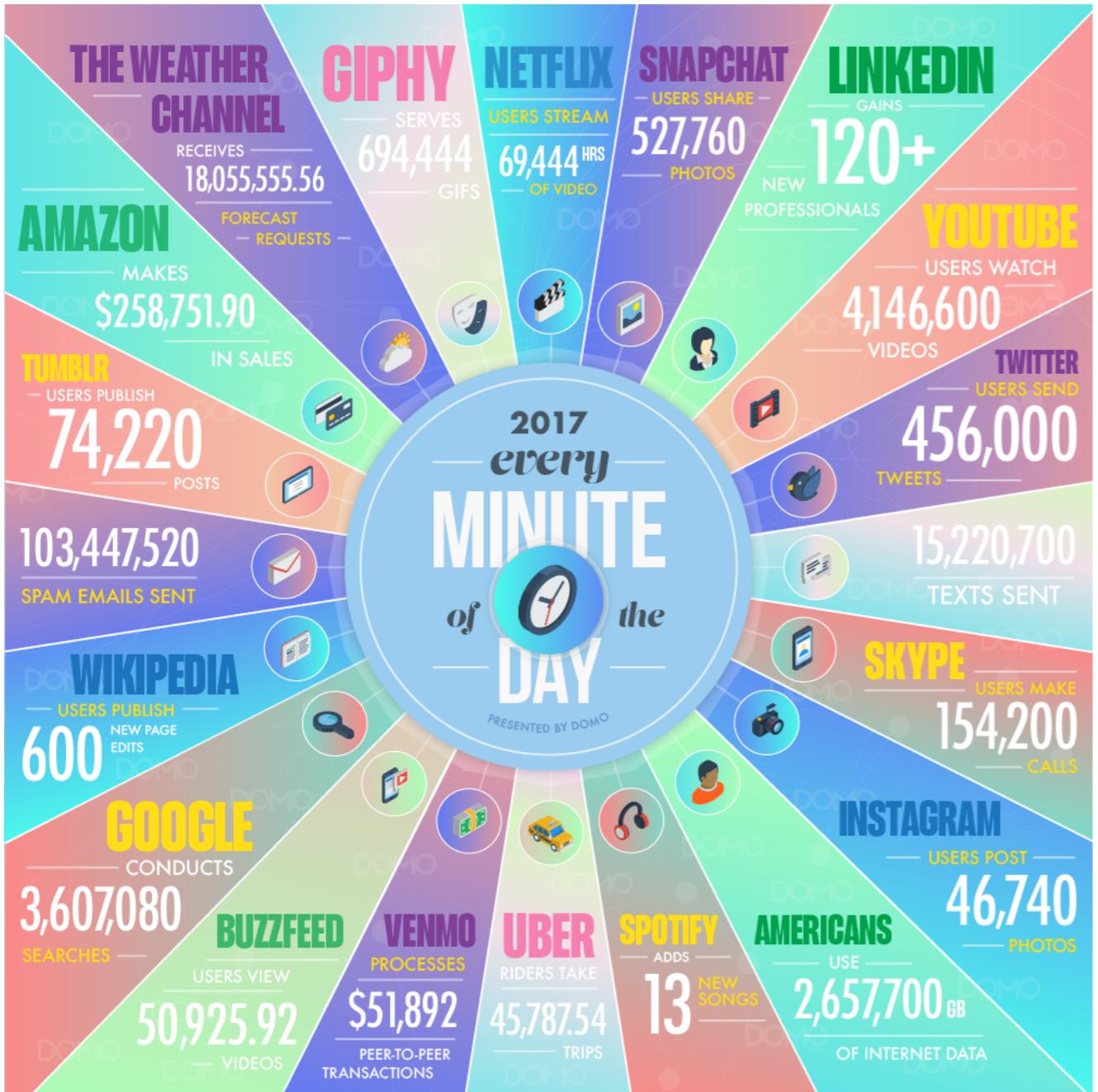
HAPPY HOLIDAYS

Christmas Woods by the Fireplace	PAVAROTTI	Christmas Dreams	Noelle	The Christmas Wife	The Christmas Tree	The History of Christmas Santa Claus	Winter Dreams	My Fair Time Traveler	It's Christmas in Hollywood	The Nutcracker Ballet
Holiday Yule Log	1994	1994	1994	1994	1994	1994	1994	1994	1994	1994

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

How much data?

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

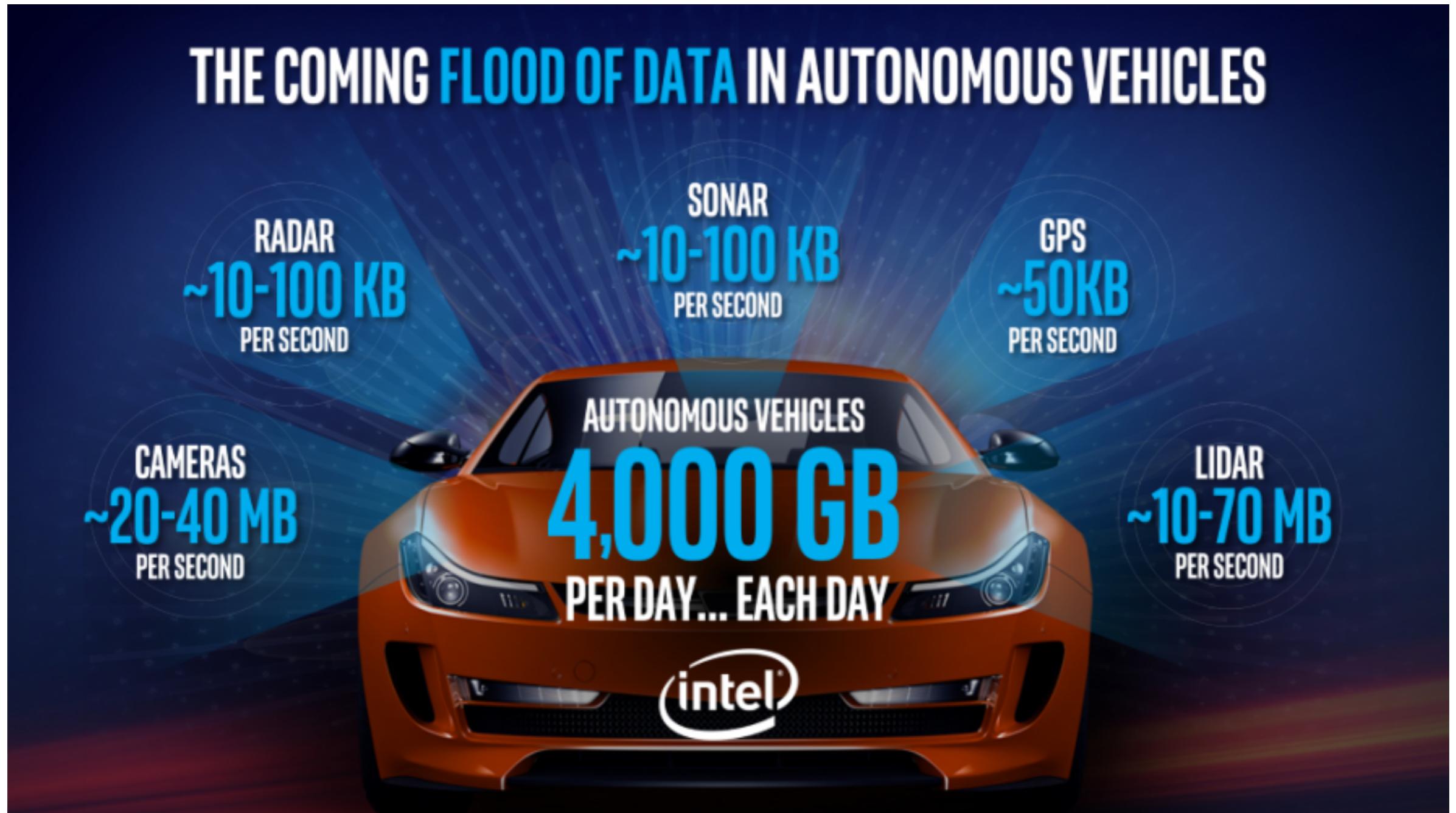


Source: https://web-assets.domo.com/blog/wp-content/uploads/2017/07/17_domo_data-never-sleeps-5-01.png

Connected humans: Web & Social media

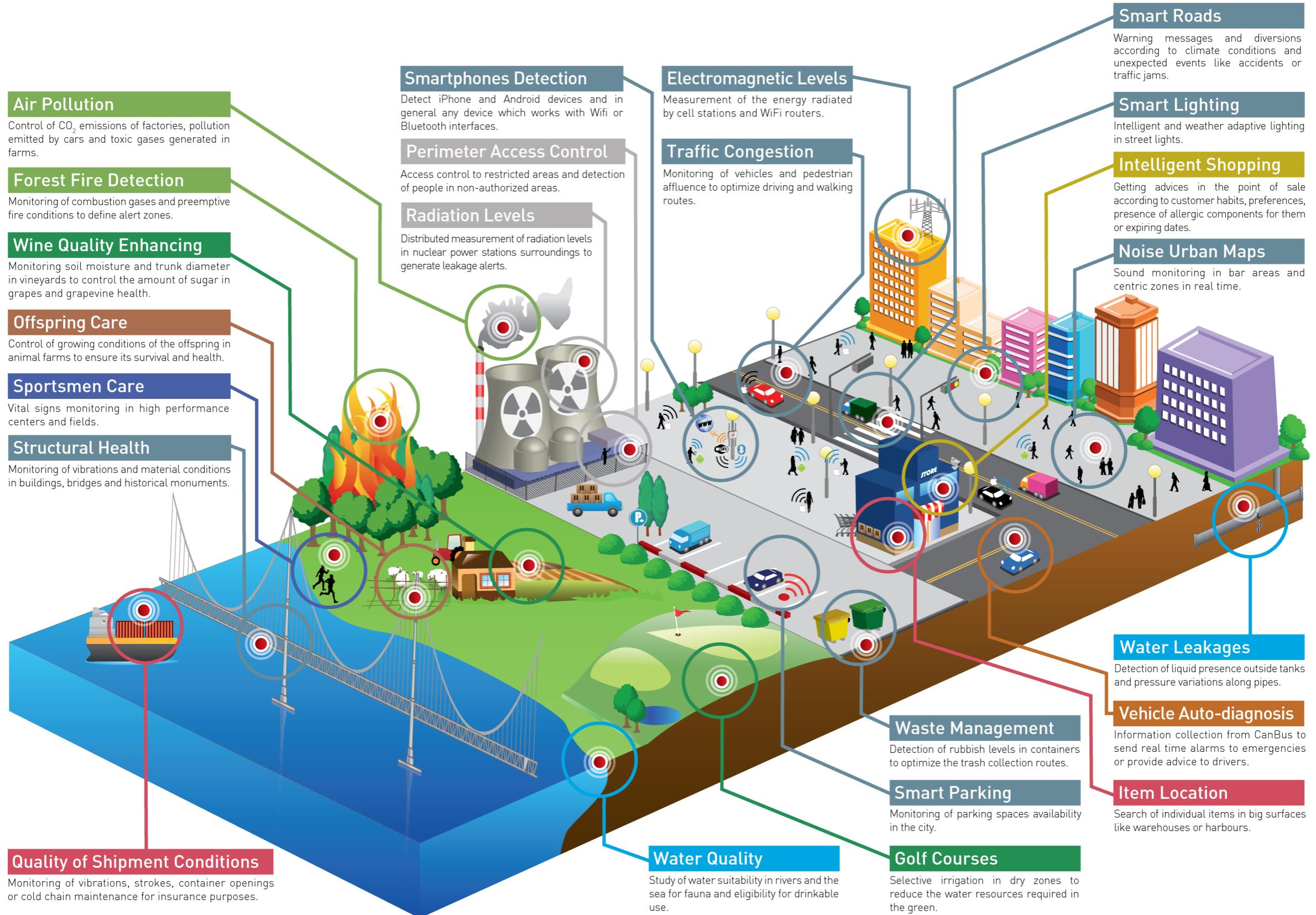
- Google carries 4,497,420 searches
- Youtube users view 4,500,000 videos
- Twitter users post 511,200 tweets
- Instagram users post 277,777 tales
- Skype users make 231,840 calls
- Uber users take 9,772 drives
- Netflix users stream 694,444 hours of video
- Giphy toils up 4,800,000 gifs
- Airbnb books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



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

1 Exabyte (EB) = 1,000,000,000,000,000,000 Bytes

9000

8000

7000

6000

5000

4000

3000

2010

2020

**50X
GROWTH
FROM
2010
TO 2020**

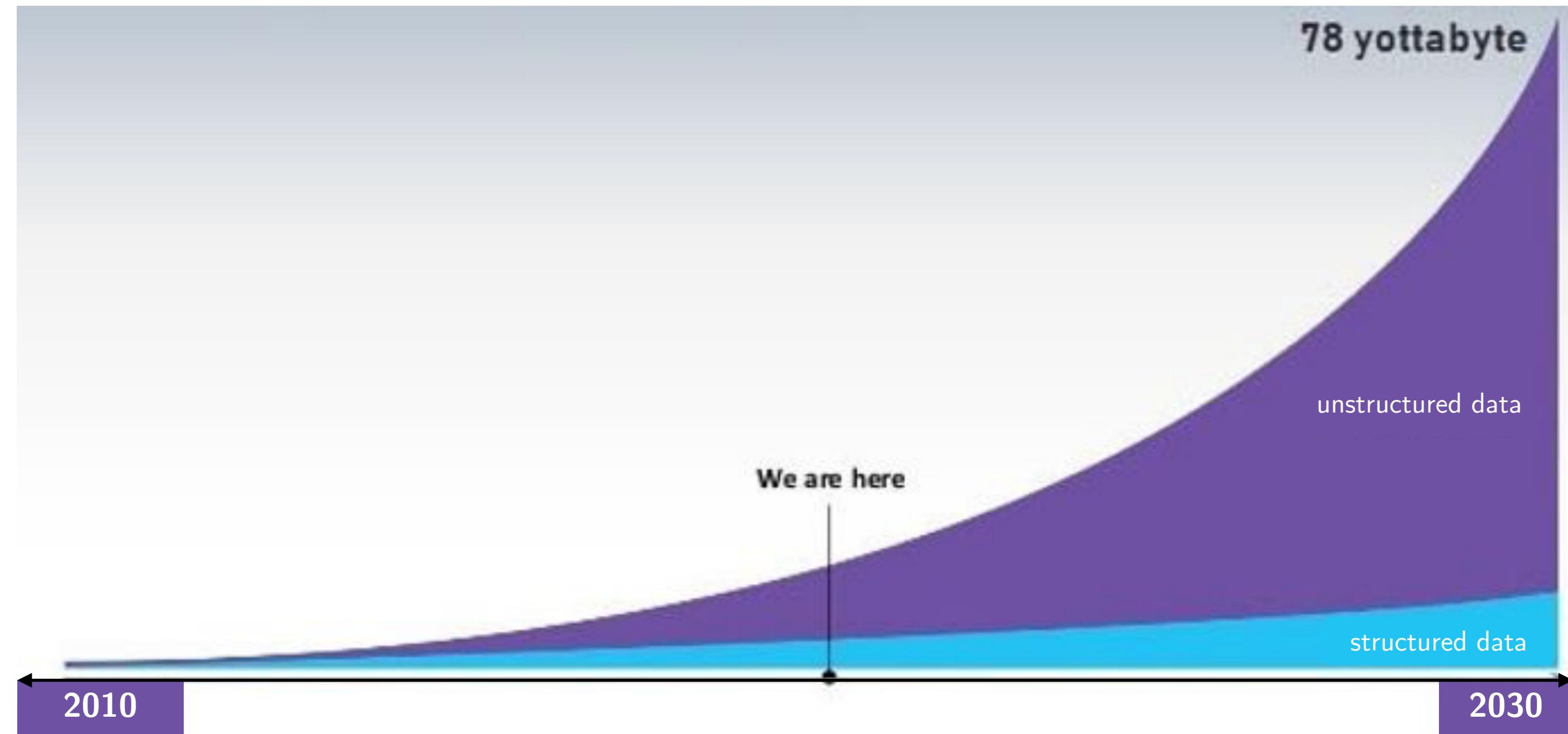
Sensors
& Devices

Social
Media

VOIP

Enterprise
Data

Source: Infosys



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?

TERABYTE

Will fit 200,000 photos or mp3 songs on a single 1 terabyte hard drive.



PETABYTE

Will fit on 16 Backblaze storage pods racked in two datacenter cabinets.



EXABYTE

Will fit in 2,000 cabinets and fill a 4 story datacenter that takes up a city block.



ZETTABYTE

Will fill 1,000 datacenters or about 20% of Manhattan, New York.



YOTTABYTE

Will fill the states of Delaware and Rhode Island with a million datacenters.



The Cost

The cost of buying a 1 terabyte hard drive today is \$100. It would cost \$100 Trillion dollars to buy a yottabyte of storage for just the hard drives.

YOTTABYTE



\$14 Trillion
United States
GDP in 2008



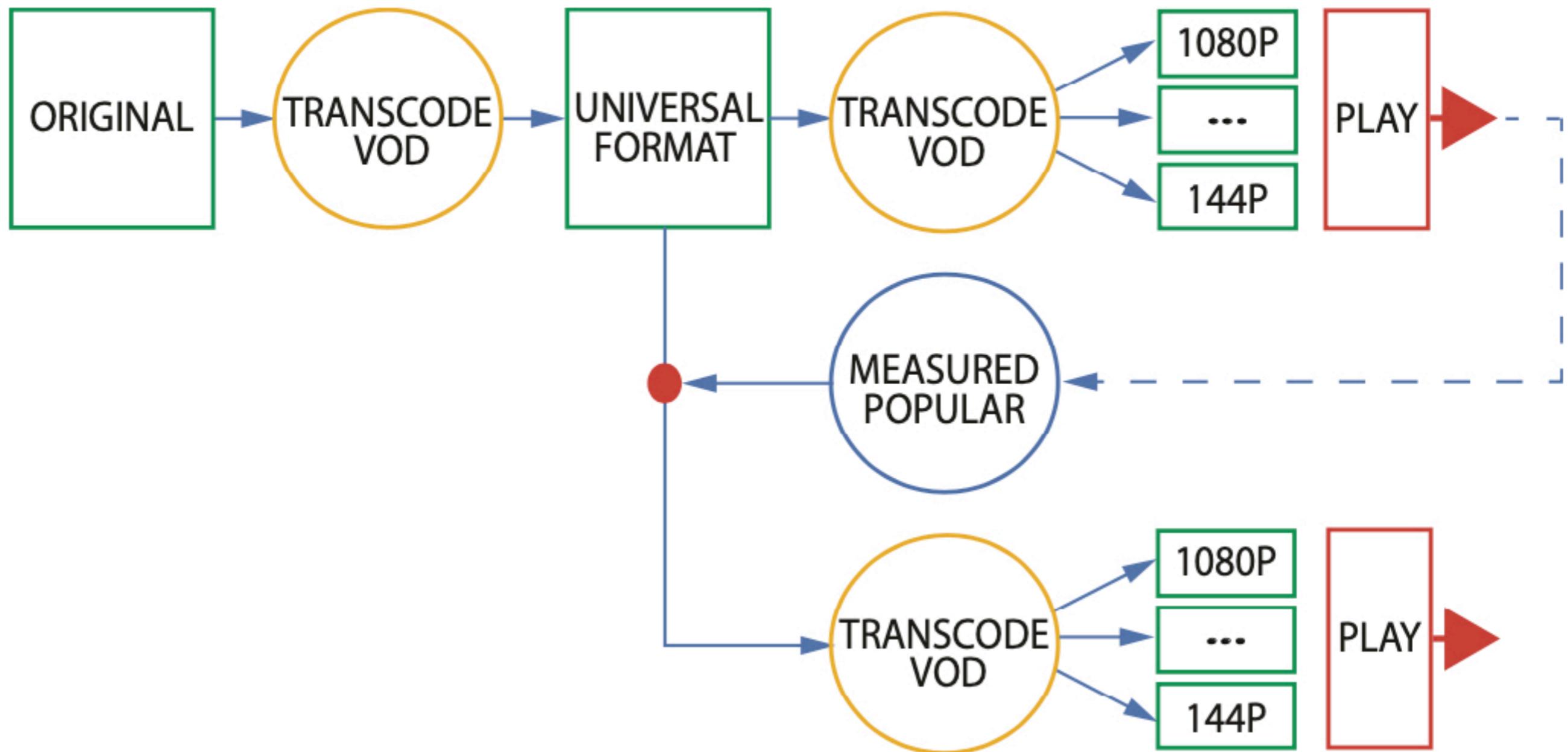
\$18 Trillion
European
GDP in 2008



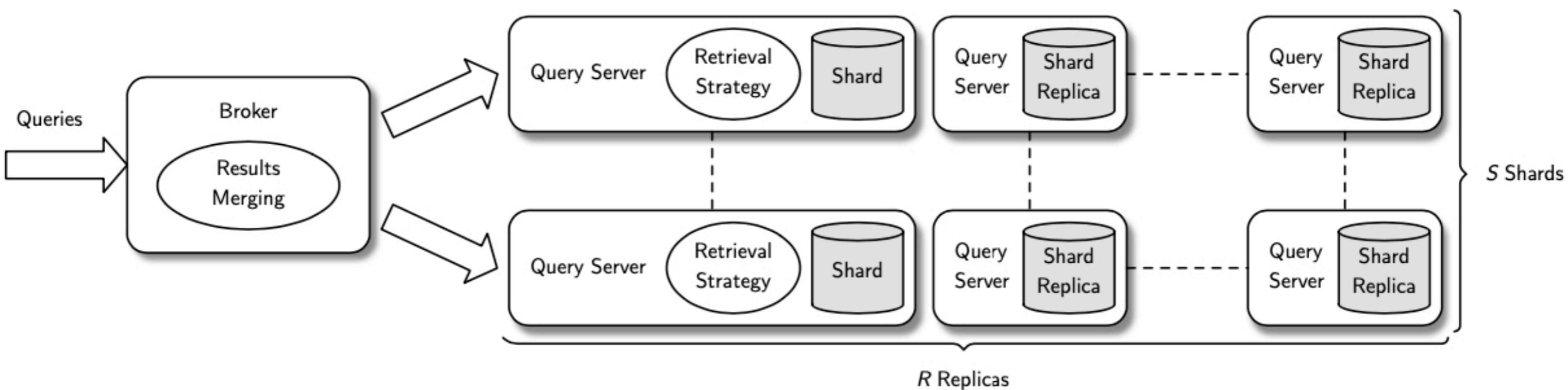
\$61 Trillion
World's GDP

\$100 Trillion
for a Yottabyte

The YouTube Video Processing Pipeline



The Web Search Engine Architecture



Partition/Aggregate Architecture

Top-level
Aggregator

Mid-level
Aggregators

Workers

