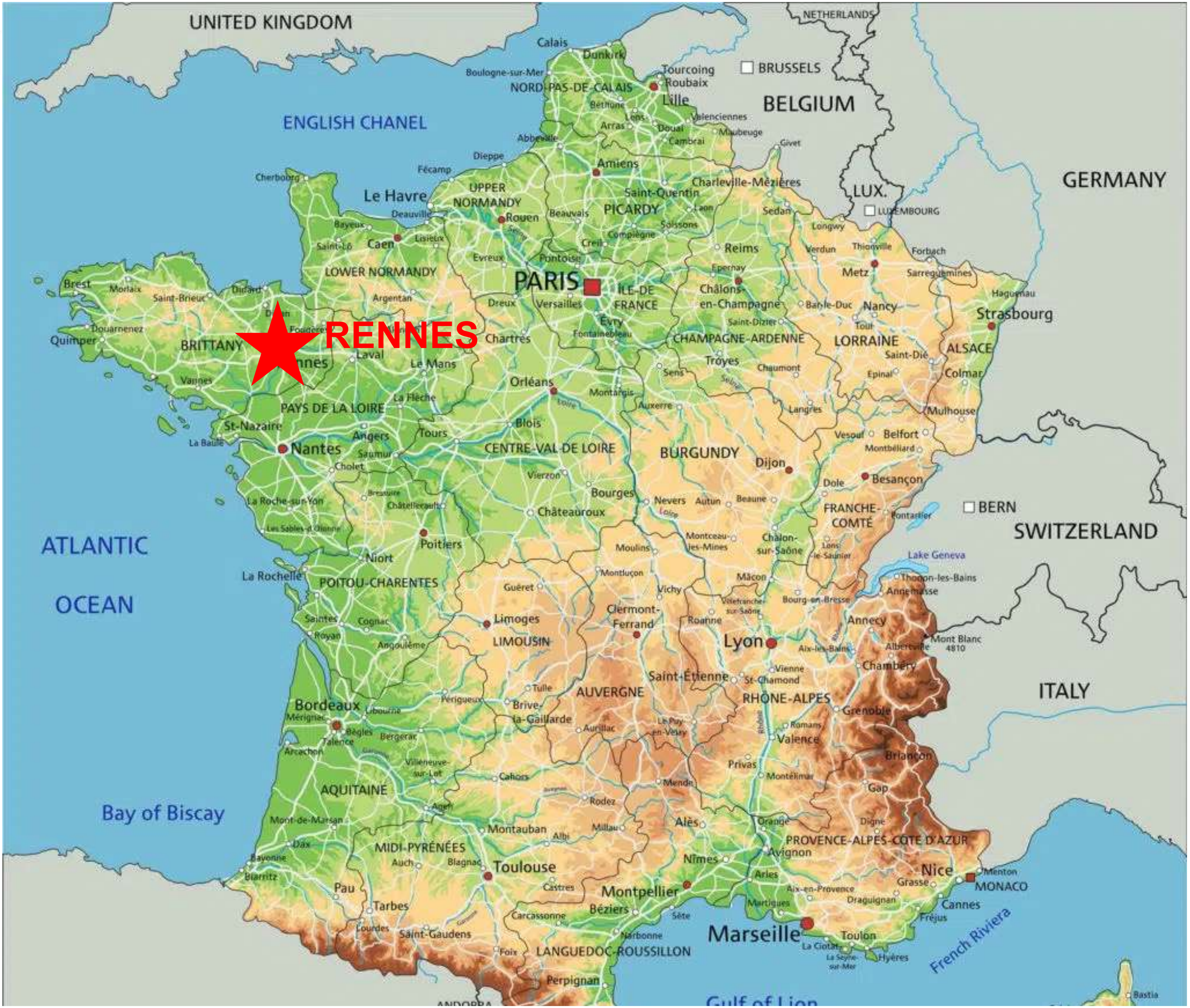
INSTITUT NATIONAL
DES SCIENCES
APPLIQUÉES
RENNES

Introduction to Big Data

Alexandru Costan
alexandru.costan@insa-rennes.fr


RENNES
UNITED KINGDOM
ENGLISH CHANEL
ATLANTIC
OCEAN
Bay of Biscay
NETHERLANDS
BRUSSELS
BELGIUM
LUX.
LUXEMBOURG
GERMANY
BERN
SWITZERLAND
ITALY
MONACO
ANDORRA
Gulf of Lion
French Riviera
Lake Geneva
Calais
Dunkirk
Boulogne-sur-Mer
NORD-PAS-DE-CALAIS
Tourcoing
Roubaix
Lille
Béthune
Lens
Arras
Douai
Valenciennes
Cambrai
Maubeuge
Abbeville
Givet
Dieppe
Amiens
Fécamp
Cherbourg
Le Havre
UPPER NORMANDY
Saint-Quentin
Charleville-Mézières
Sedan
Deauville
Rouen
Beauvais
PICARDY
Laon
Soissons
Bayeux
Compiègne
Longwy
Saint-Lô
Caen
Lisieux
Creil
Reims
Verdun
Thionville
Forbach
Evreux
Pontoise
LOWER NORMANDY
Epernay
Metz
Sarreguemines
Brest
Morlaix
Saint-Brieuc
Dinard
Argentan
PARIS
ILE-DE FRANCE
Châlons-en-Champagne
Bar-le-Duc
Nancy
Haguenau
Dinan
Dreux
Versailles
Evry
Saint-Dizier
Toul
Strasbourg
Douarnenez
Fougères
Quimper
BRITTANY
Laval
Chartres
Fontainebleau
CHAMPAGNE-ARDENNE
LORRAINE
Saint-Dié
ALSACE
Rennes
Le Mans
Troyes
Epinal
Colmar
Vannes
Orléans
Sens
Chaumont
La Flèche
Montargis
Seine
Mulhouse
PAYS DE LA LOIRE
Auxerre
Langres
St-Nazaire
Loire
Blois
Vesoul
Belfort
La Baule
Angers
Tours
Montbéliard
Nantes
Saumur
CENTRE-VAL DE LOIRE
BURGUNDY
Dijon
Cholet
Vierzon
Dole
Besançon
Bressuire
Bourges
Nevers
Autun
Beaune
La Roche-sur-Yon
Châtellerault
FRANCHE-COMTÉ
Pontarlier
Châteauroux
Les Sables-d'Olonne
Poitiers
Montceau-les-Mines
Chalon-sur-Saône
Lons-le-Saunier
Moulins
Niort
La Rochelle
POITOU-CHARENTES
Montluçon
Mâcon
Thonon-les-Bains
Guéret
Vichy
Annemasse
Villefranche-sur-Saône
Bourg-en-Bresse
Clermont-Ferrand
Roanne
Annecy
Saintes
Cognac
Limoges
Royan
Mont Blanc 4810
LIMOUSIN
Lyon
Albertville
Angoulême
Aix-les-Bains
Vienne
Chambéry
Saint-Etienne
St-Chamond
Tulle
Périgueux
AUVERGNE
RHÔNE-ALPES
Bordeaux
Libourne
Brive-la-Gaillarde
Grenoble
Mérignac
Le Puy-en-Velay
Romans
Bègles
Aurillac
Bergerac
Valence
Talence
Arcachon
Privas
Briançon
Villeneuve-sur-Lot
Mende
Montélimar
Cahors
AQUITAINE
Gap
Rodez
Agen
Mont-de-Marsan
Millau
Alès
Orange
Digne
Montauban
Albi
PROVENCE-ALPES-CÔTE D'AZUR
Dax
Avignon
MIDI-PYRÉNÉES
Nîmes
Bayonne
Auch
Blagnac
Toulouse
Nice
Menton
Biarritz
Arles
Grasse
Castres
Montpellier
Aix-en-Provence
Pau
Martigues
Draguignan
Cannes
Tarbes
Béziers
Sète
Fréjus
Carcassonne
Lourdes
Saint-Gaudens
Marseille
Narbonne
Toulon
La Ciotat
Foix
LANGUEDOC-ROUSSILLON
La Seyne-sur-Mer
Hyères
Perpignan
Bastia

Copyright © sabine de villeroy

Rennes capital of Brittany and city of over 400 000 inhabitants

Rennes is the:

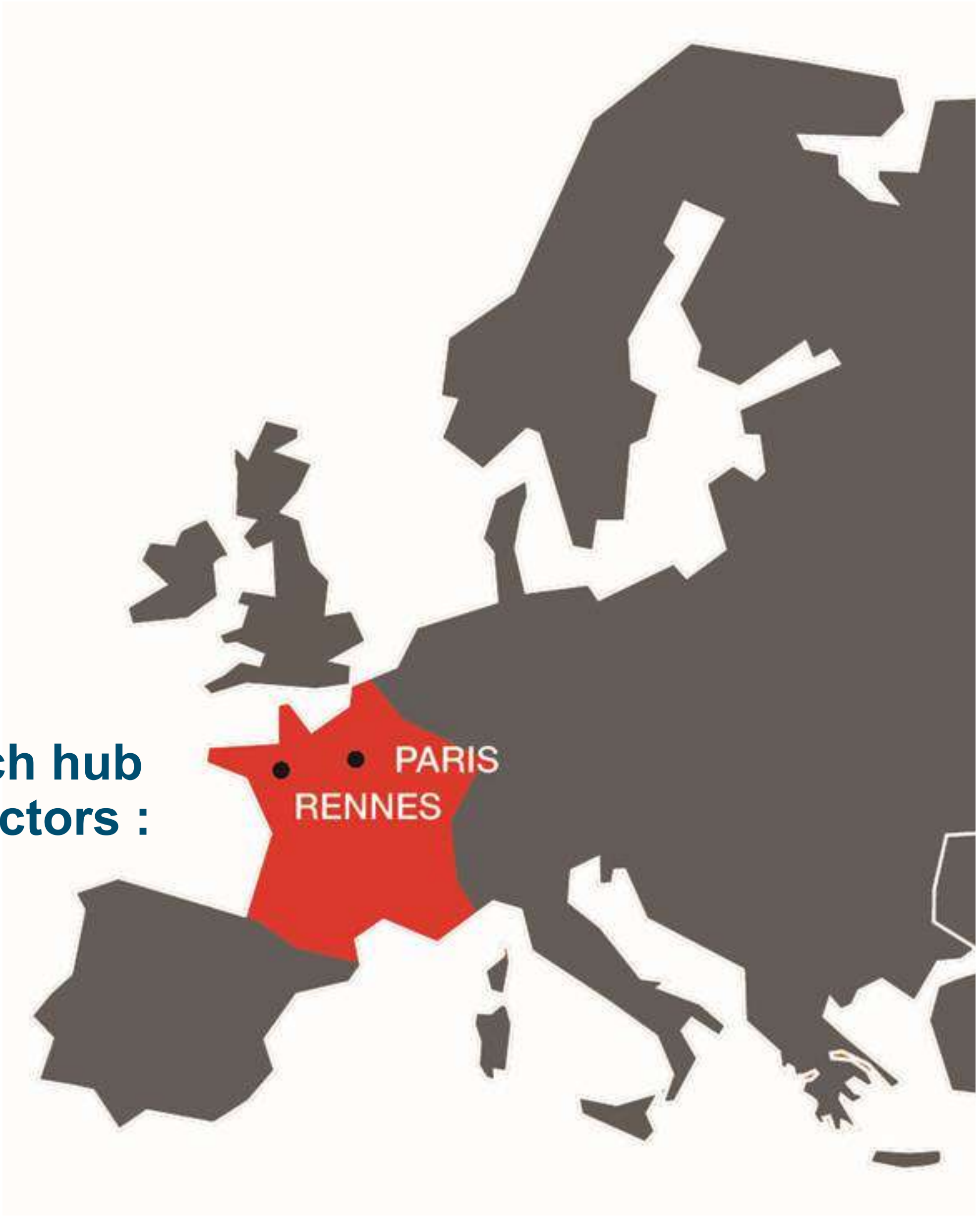
- 11th largest city in France
- 2nd city in France for its student population

Rennes is situated:

- 45 minutes from the sea (Saint-Malo)
- 1,5 hours from Paris

Rennes is a university and research hub of international importance in 3 sectors :

- Health
- Digital Technology
- Eco-activities



INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

INSA Rennes,

a public-funded graduate and post graduate engineering school member of the INSA group

INSA Rennes, public-funded graduate and post graduate engineering school

Founded in 1966, INSA Rennes is classed among **the best graduate and post graduate engineering schools** in France.

In addition to being a research center, INSA Rennes gives professional training in engineering and research in **2 poles of excellence:**

- **Information and Communication Sciences and Technologies** (ICST)
- **Materials, Structures and Mechanics** (MSM)

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION

Liberté Égalité Fraternité

DIPLÔME D'INGÉNIEUR CONTRÔLÉ PAR L'ÉTAT

European Accreditation of Engineering Programmes EUR-ACE®

Cti Commission des Titres d'Ingénieur

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

INSA Rennes, at the heart of a science and technology campus

Equipped 17 hectare campus

- Accommodation and catering in situ
- Pedagogical, scientific and sports equipment

Within the Le POOOL technology cluster

- More than 300 companies
- 80% engineering and technology companies

Within the science campus of Rennes

- 70289 students
- 32 institutes of higher education and graduate schools

This course

Today

What is Big Data?

- Recognize some of the main terminology

What can we do with Big Data?

- Realize the potential of Big Data

Why is it difficult?

- Understand why we need a different paradigm

How to store and process Big Data?

- Know the existing tools

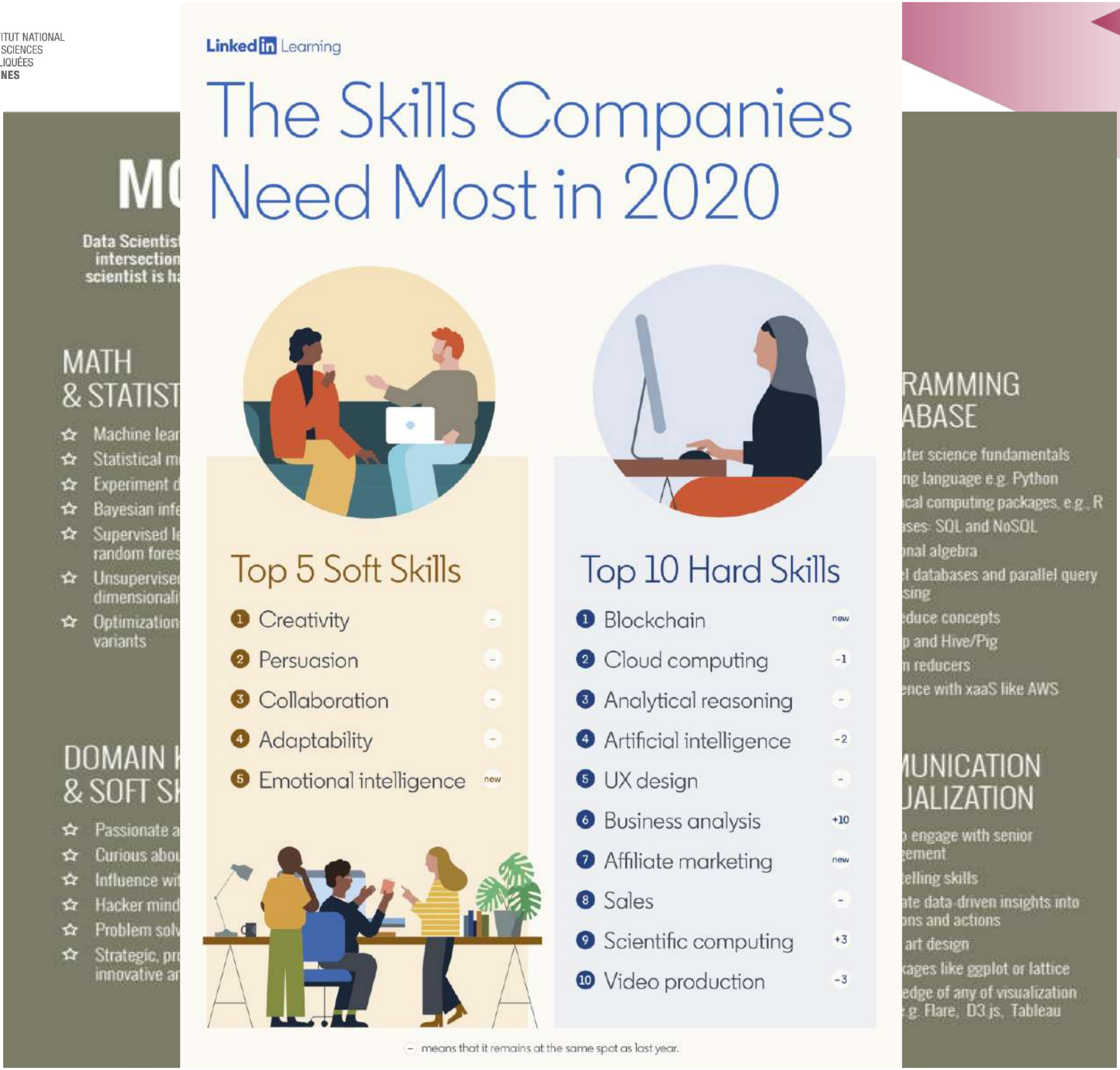
© D.Fletcher for CloudTweaks.com

Not this course

Not focusing on

- machine learning
- data mining
- natural language processing

although we will touch on these



MO

Data Scientist
intersection
scientist is ha

MATH
& STATIST

- ✩ Machine lear
- ✩ Statistical m
- ✩ Experiment d
- ✩ Bayesian infe
- ✩ Supervised le random fores
- ✩ Unsupervised dimensionali
- ✩ Optimization variants

DOMAIN K
& SOFT SK

- ✩ Passionate a
- ✩ Curious abou
- ✩ Influence wit
- ✩ Hacker mind
- ✩ Problem solv
- ✩ Strategic, pr innovative ar

Linked in Learning

The Skills Companies Need Most in 2020

Top 5 Soft Skills

1 Creativity	–
2 Persuasion	–
3 Collaboration	–
4 Adaptability	–
5 Emotional intelligence	new

Top 10 Hard Skills

1 Blockchain	new
2 Cloud computing	-1
3 Analytical reasoning	–
4 Artificial intelligence	-2
5 UX design	–
6 Business analysis	+10
7 Affiliate marketing	new
8 Sales	–
9 Scientific computing	+3
10 Video production	-3

– means that it remains at the same spot as last year.

RAMMING
ABASE

- uter science fundamentals
- ng language e.g. Python
- ical computing packages, e.g., R
- ases: SQL and NoSQL
- onal algebra
- el databases and parallel query sing
- educe concepts
- p and Hive/Pig
- n reducers
- ence with xaaS like AWS

MUNICATION
UALIZATION

- o engage with senior gement
- telling skills
- ate data-driven insights into ons and actions
- art design
- kages like ggplot or lattice
- edge of any of visualization e.g. Flare, D3.js, Tableau

Lectures outline

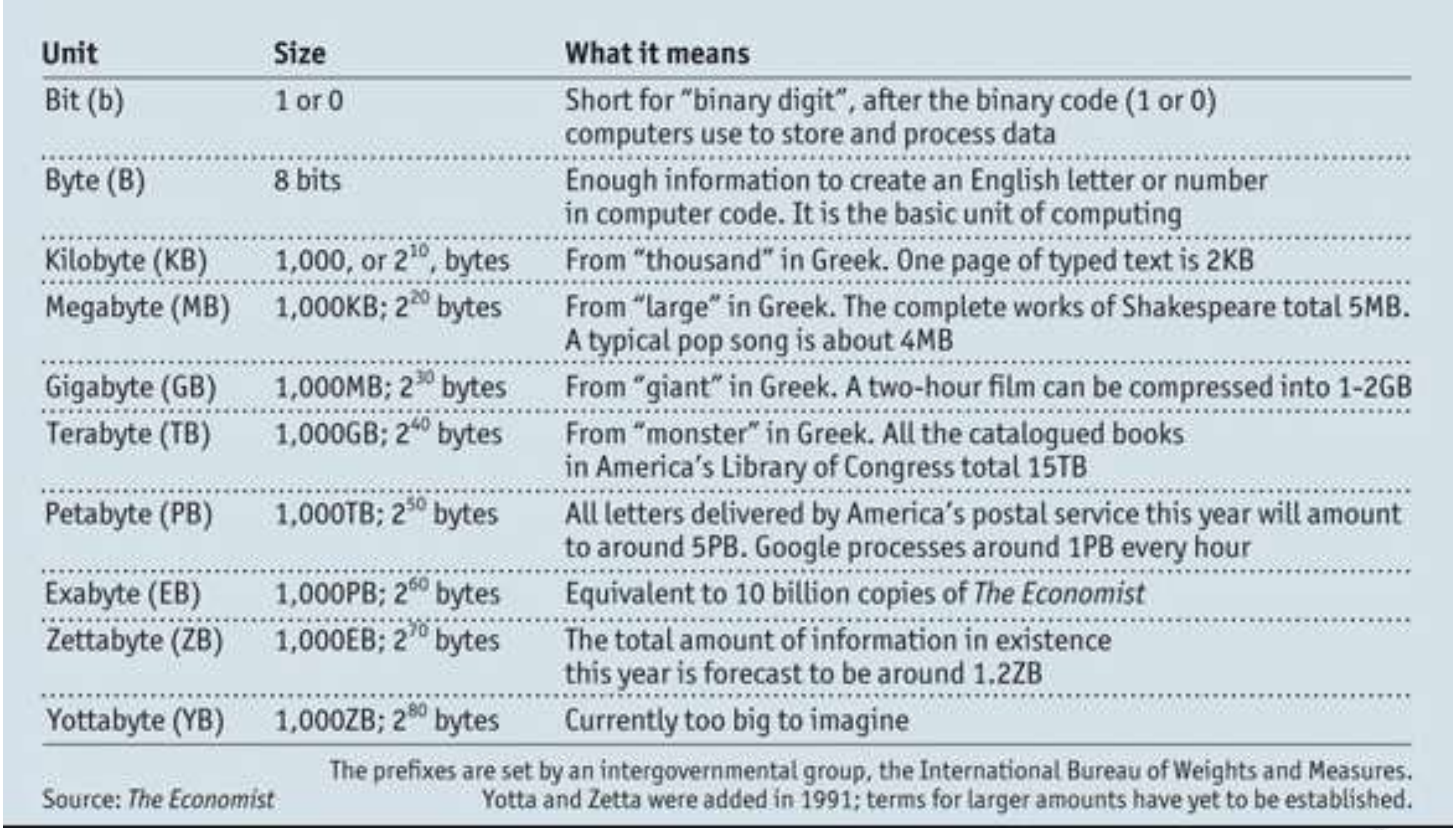
- **Big Data overview**
  - Data and Processing Models
  - Consistency
- **Programming with Big Data**
  - Google MapReduce
  - Apache Hadoop

Readings

- Tony Hey, Stewart Tansley, Kristin Tolle „The Fourth Paradigm: Data-Intensive Scientific Discovery ”, Microsoft Research
- Jeffrey Stanton, “Introduction to Data Science”, Syracuse University Press
- Jeffrey Dean, Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters ”, OSDI 2004
- Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica, “Spark: Cluster Computing with Working Sets”, NSDI 2012
- … additional bibliography specific to each lecture

Sources of Big Data

Big Data Units

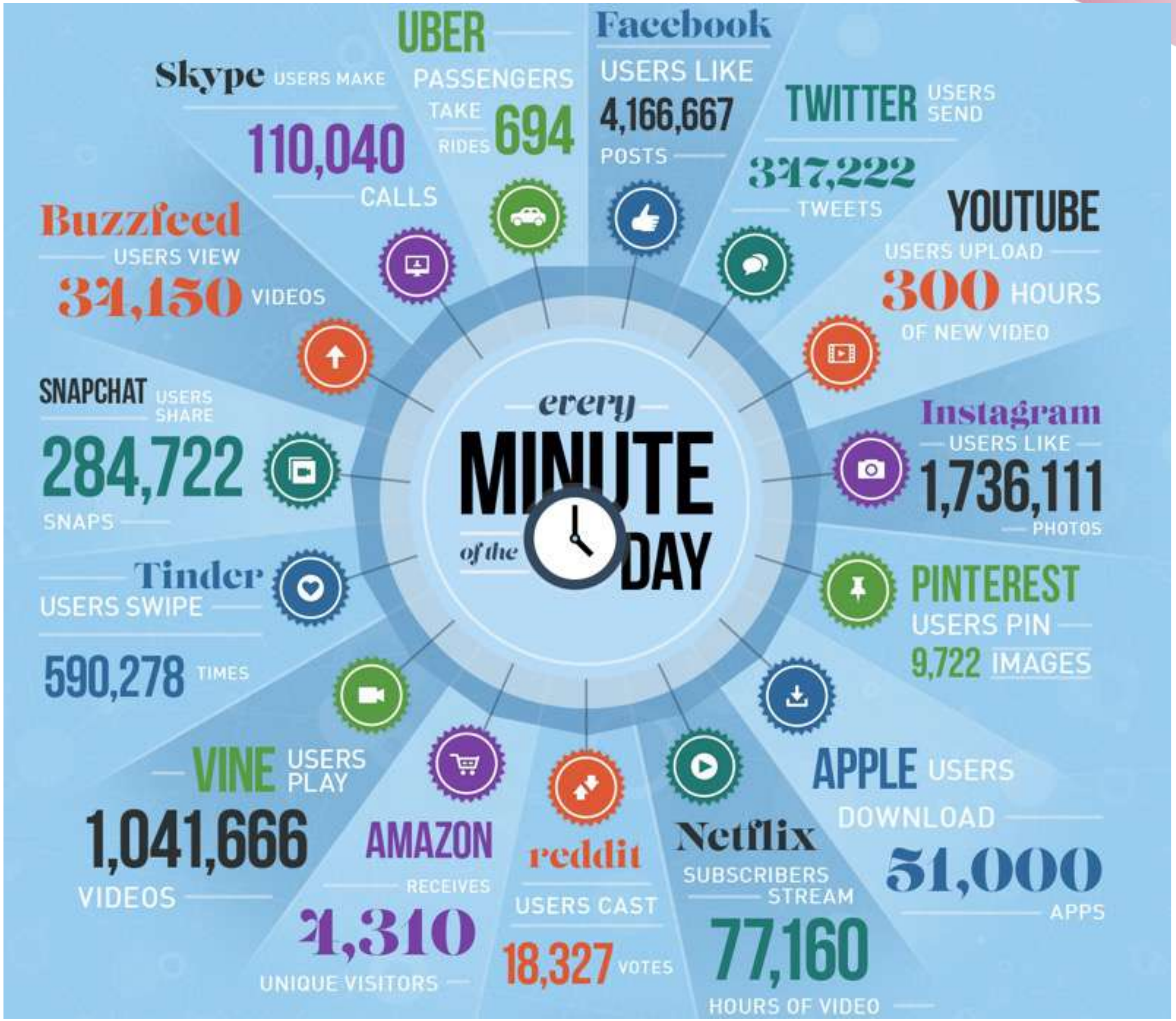


Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Internet


every MINUTE of the DAY
UBER PASSENGERS TAKE 694 RIDES
Facebook USERS LIKE 4,166,667 POSTS
Skype USERS MAKE 110,040 CALLS
TWITTER USERS SEND 347,222 TWEETS
YOUTUBE USERS UPLOAD 300 HOURS OF NEW VIDEO
Buzzfeed USERS VIEW 34,150 VIDEOS
SNAPCHAT USERS SHARE 284,722 SNAPS
Instagram USERS LIKE 1,736,111 PHOTOS
Tinder USERS SWIPE 590,278 TIMES
PINTEREST USERS PIN 9,722 IMAGES
VINE USERS PLAY 1,041,666 VIDEOS
AMAZON RECEIVES 4,310 UNIQUE VISITORS
reddit USERS CAST 18,327 VOTES
Netflix SUBSCRIBERS STREAM 77,160 HOURS OF VIDEO
APPLE USERS DOWNLOAD 51,000 APPS

Digital traces

1 GB of data

/ person

/ day

x 5,000,000,000 users

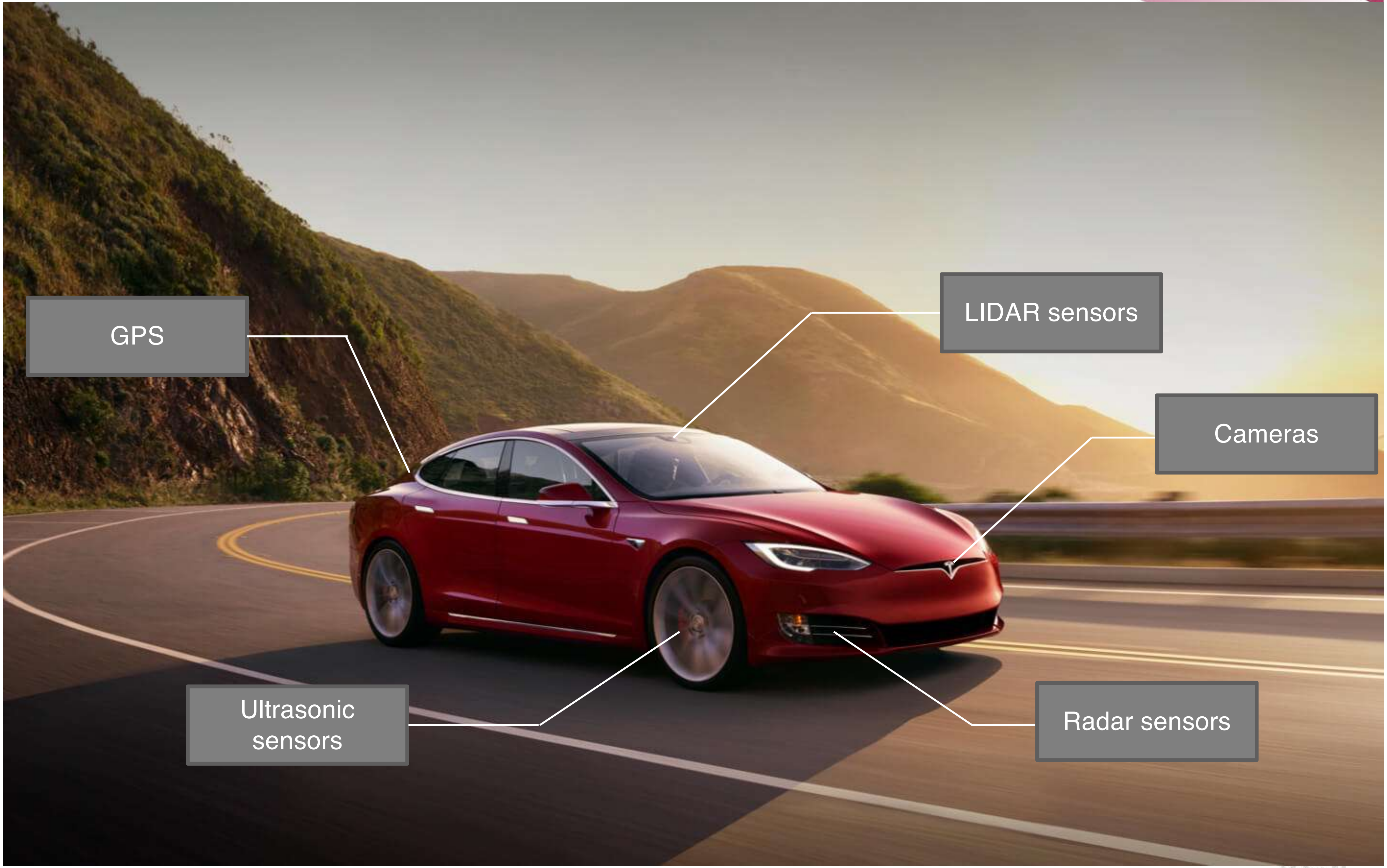
The big picture (in 2020)

- 10PB of Facebook data per day

- 500M tweets per day

- Google processes 100PB a day

Sensors from the Internet of Things



GPS

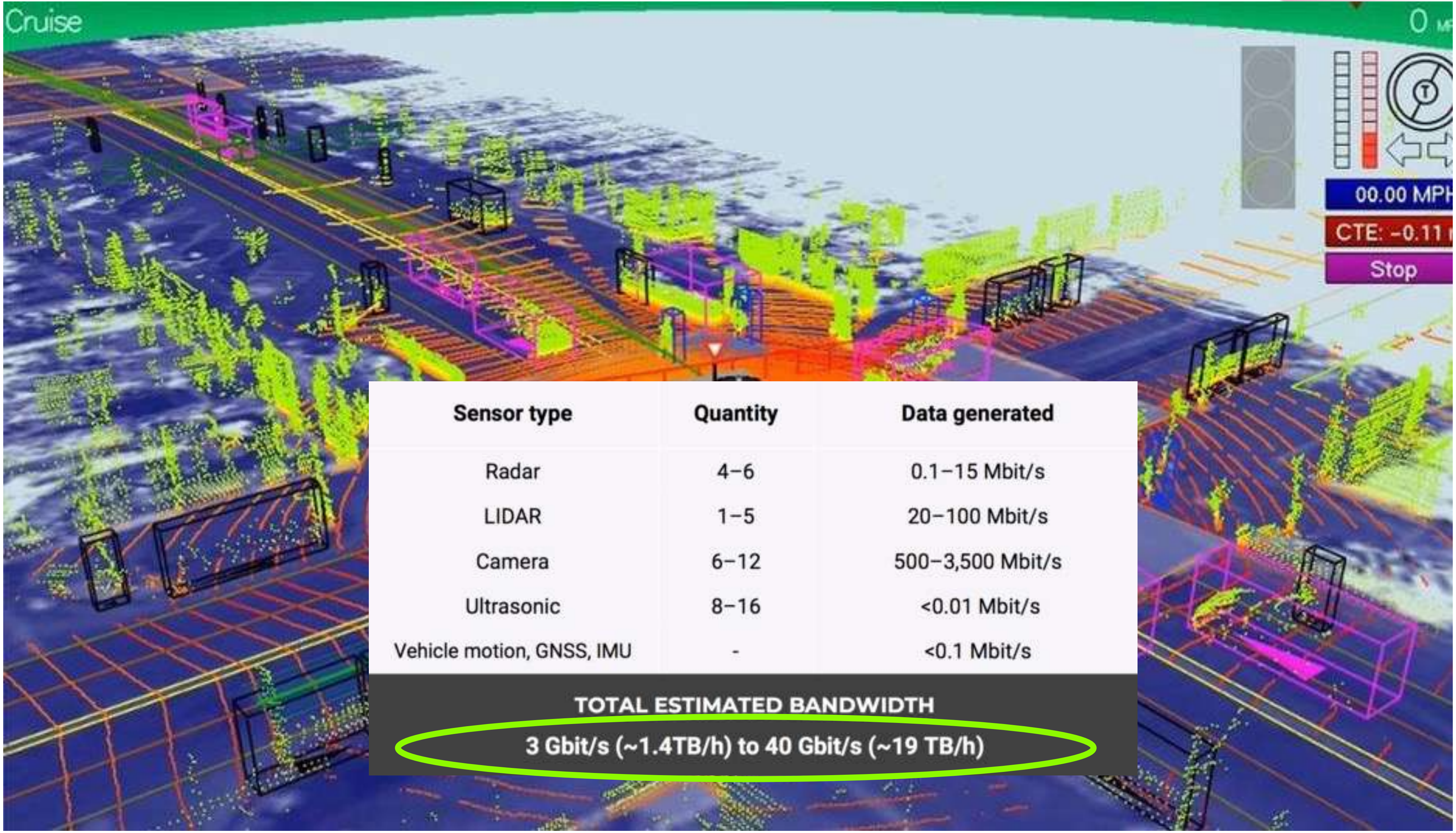
LIDAR sensors

Cameras

Ultrasonic sensors

Radar sensors

Sensors from the Internet of Things

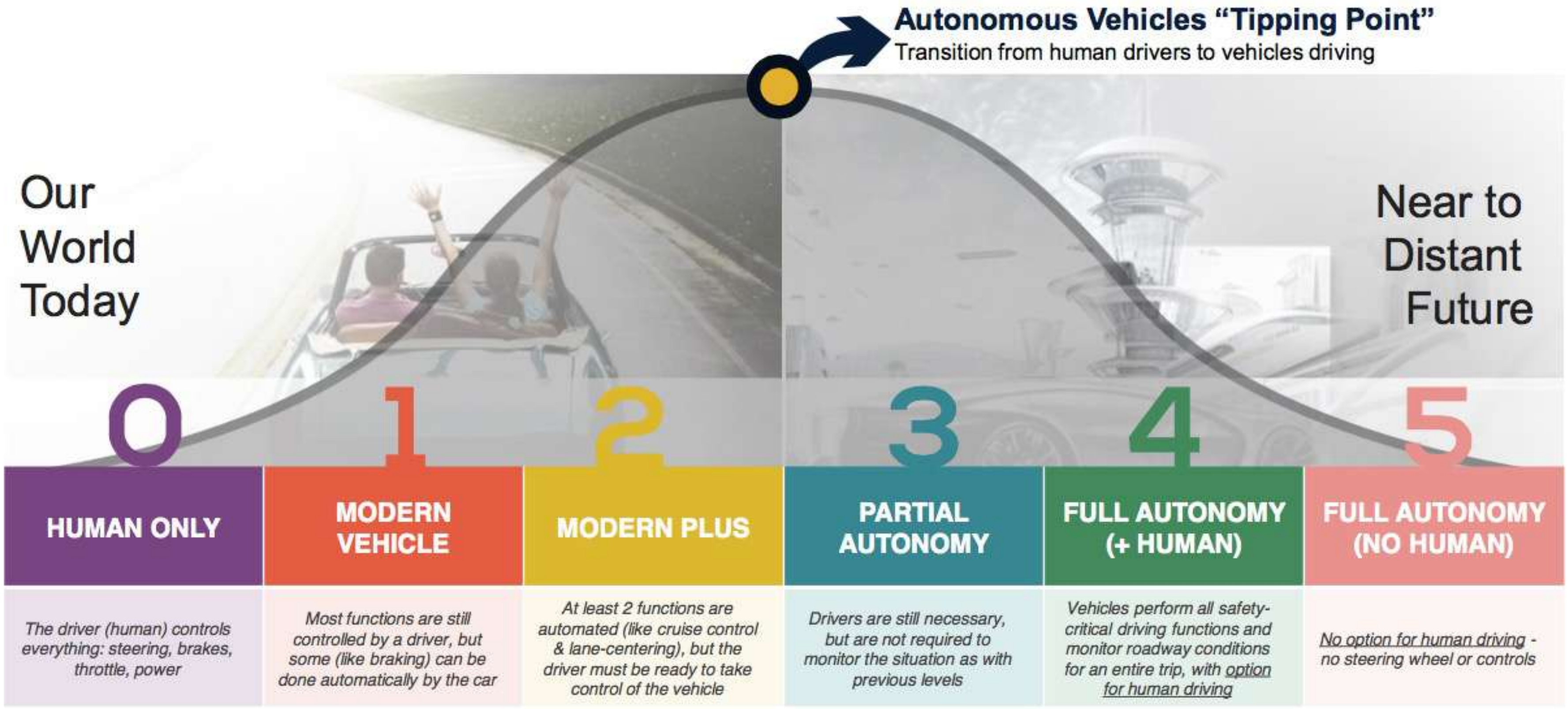


Sensor type	Quantity	Data generated
Radar	4–6	0.1–15 Mbit/s
LIDAR	1–5	20–100 Mbit/s
Camera	6–12	500–3,500 Mbit/s
Ultrasonic	8–16	<0.01 Mbit/s
Vehicle motion, GNSS, IMU	-	<0.1 Mbit/s

TOTAL ESTIMATED BANDWIDTH

3 Gbit/s (~1.4TB/h) to 40 Gbit/s (~19 TB/h)

The levels of Autonomous Vehicles

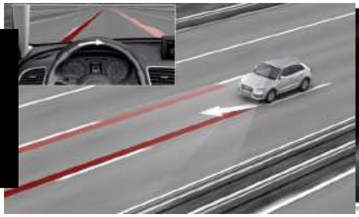


Autonomous Vehicles "Tipping Point"
Transition from human drivers to vehicles driving

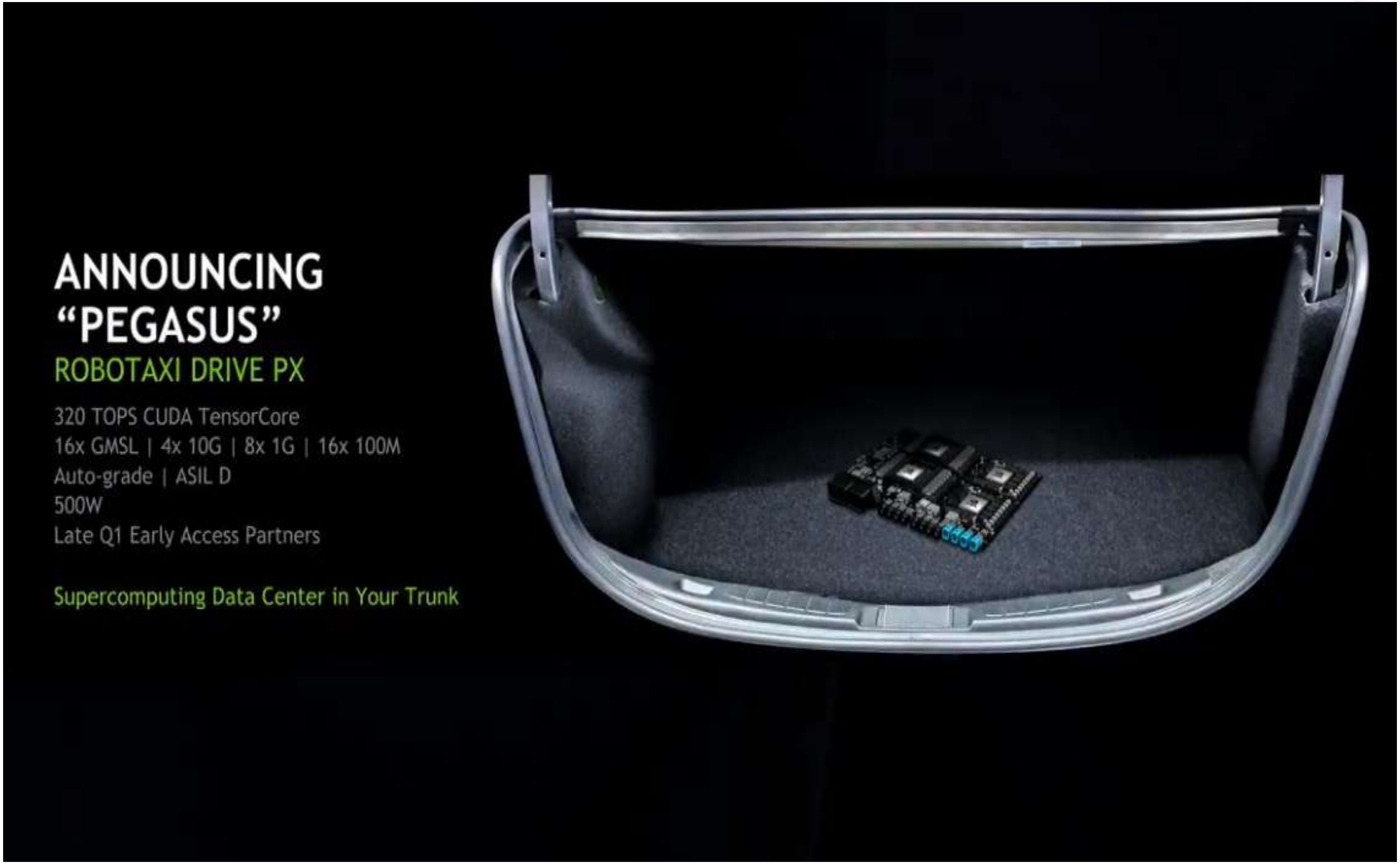
Our World Today

Near to Distant Future

0	1	2	3	4	5
HUMAN ONLY	MODERN VEHICLE	MODERN PLUS	PARTIAL AUTONOMY	FULL AUTONOMY (+ HUMAN)	FULL AUTONOMY (NO HUMAN)
The driver (human) controls everything: steering, brakes, throttle, power	*Most functions are still controlled by a driver, but some (like braking) can be done automatically by the car*	*At least 2 functions are automated (like cruise control & lane-centering), but the driver must be ready to take control of the vehicle*	*Drivers are still necessary, but are not required to monitor the situation as with previous levels*	*Vehicles perform all safety-critical driving functions and monitor roadway conditions for an entire trip, with option for human driving*	*No option for human driving - no steering wheel or controls*



Self-driving cars



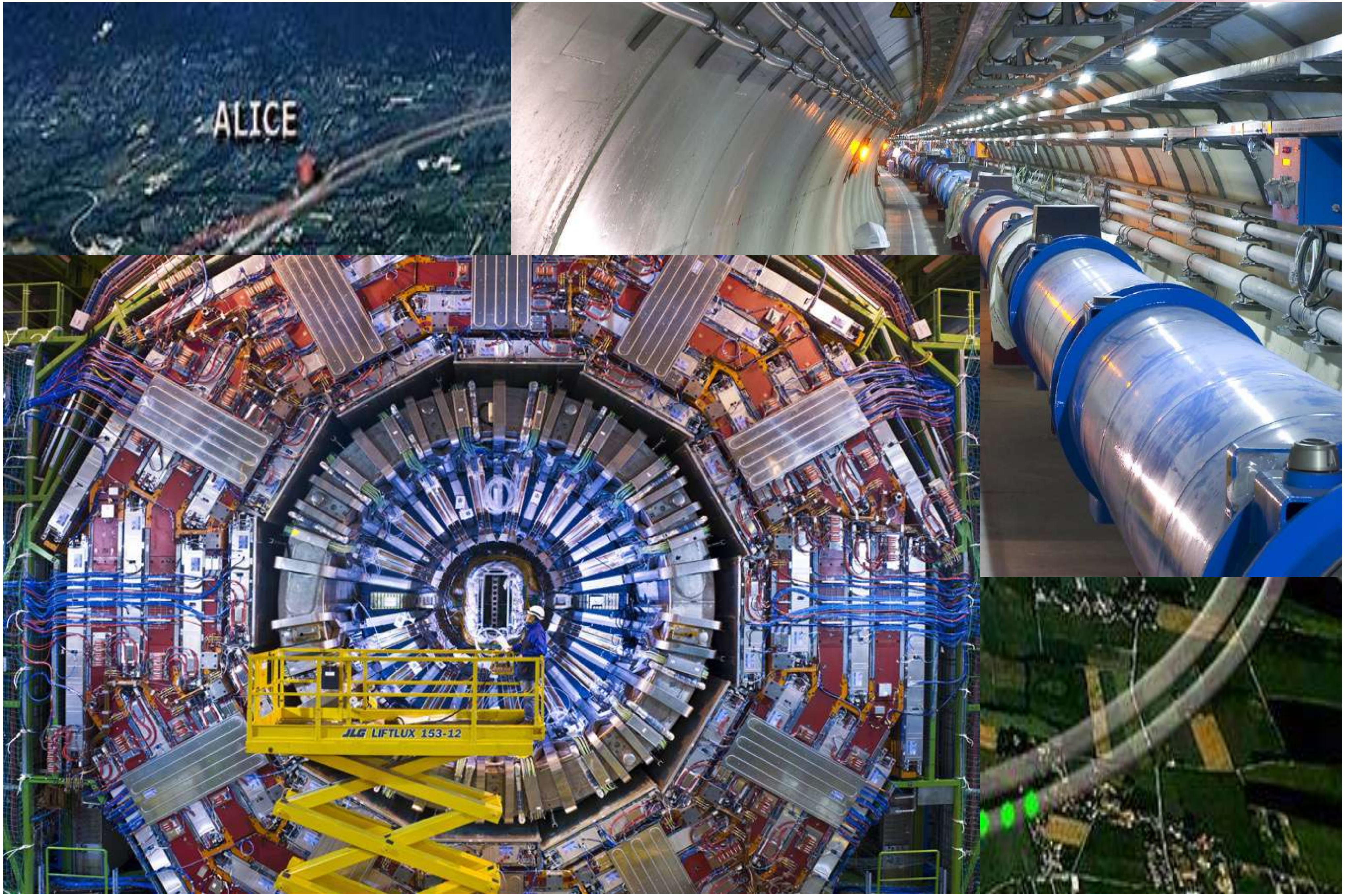
ANNOUNCING
"PEGASUS"

ROBOTAXI DRIVE PX

320 TOPS CUDA TensorCore
16x GMSL | 4x 10G | 8x 1G | 16x 100M
Auto-grade | ASIL D
500W
Late Q1 Early Access Partners

Supercomputing Data Center in Your Trunk

Large Hadron Collider



10 light year away
The SKA will be so sensitive that it will be able to detect an airport radar on a planet at this distance

2'000'000 years
The data collected by the SKA in a single day would take nearly two million years to playback on an ipod

1'000'000+
of 500GB laptops can be filled with SKA data every year

On two sites

South Africa
SKA1-MID

≈200
dishes

5x
more sensitive than any other radio telescope

33'000 m^2
of total collecting area
(=126 tennis courts)

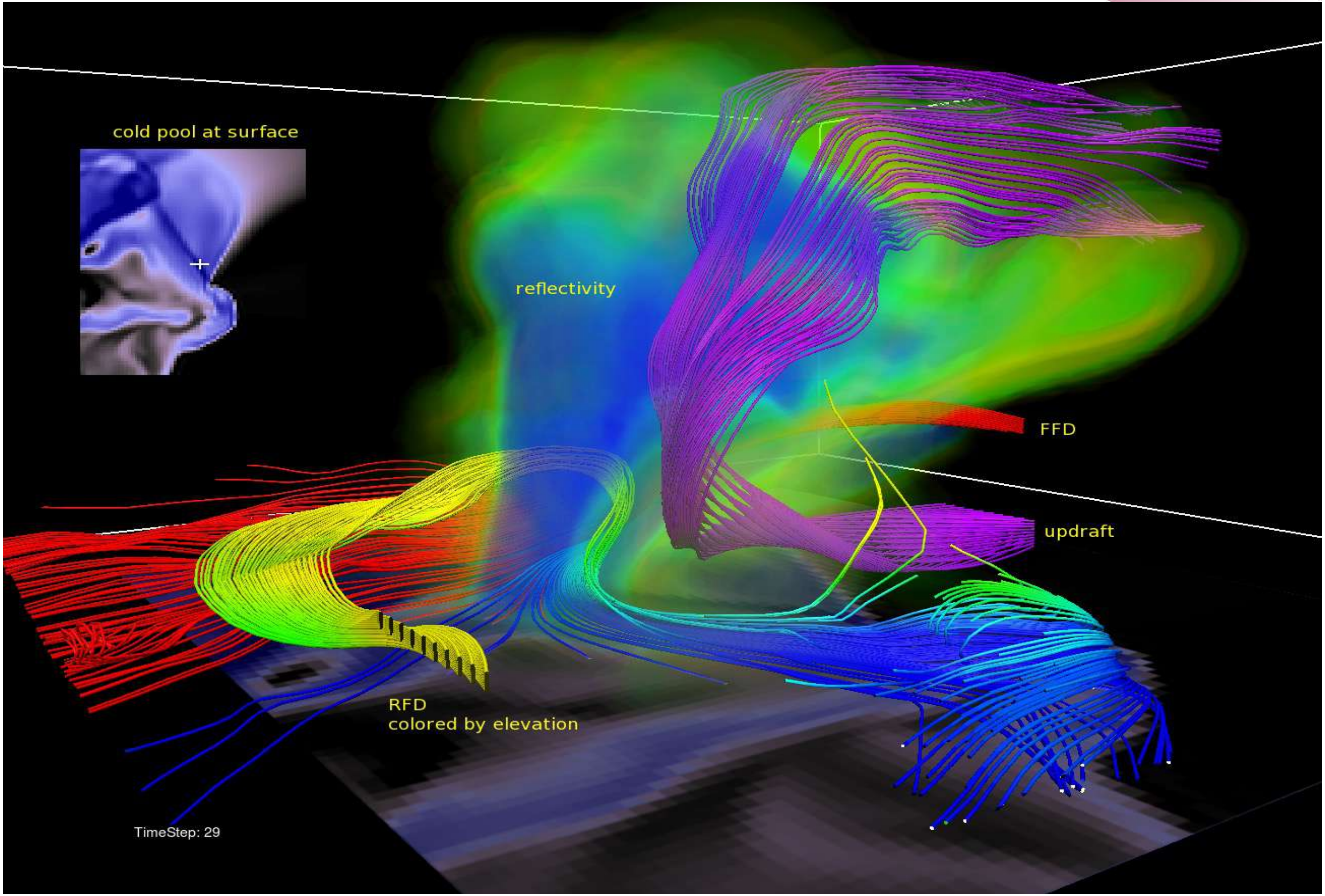
Western Australia
SKA1-LOW

8x
more sensitive than any other radio telescope

≈130'000
antennas spread between 500 stations

420'000 m^2
of total collecting area
(=58 football pitch)

Climate Simulations



3D Rendering for Animations

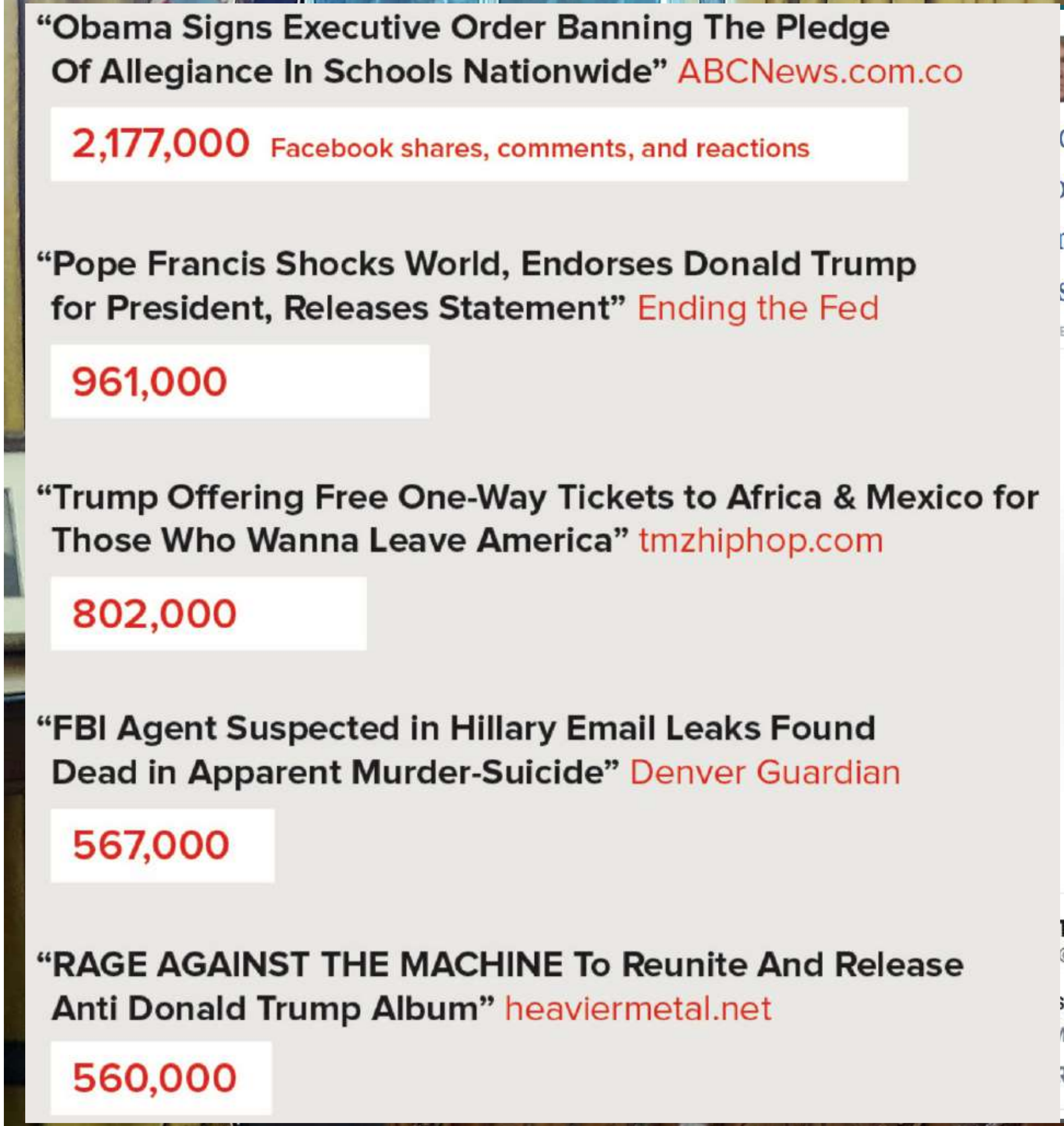
Avatar

- 40,000 processors handling 8 GB of data per second, 24 hours a day
- A final copy equated to 17 GB per minute of storage
- The sum of required computing power for the creation of Avatar reached 205 teraflops

Monsters University

- 2,000 computers with more than 24,000 cores
- Still took 25-30 hours to render a single frame
- All in all, it took over 100 million hours of CPU time to render the entire movie

Fake News



"Obama Signs Executive Order Banning The Pledge Of Allegiance In Schools Nationwide" ABCNews.com.co

2,177,000 Facebook shares, comments, and reactions

"Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement" Ending the Fed

961,000

"Trump Offering Free One-Way Tickets to Africa & Mexico for Those Who Wanna Leave America" tmzhiphop.com

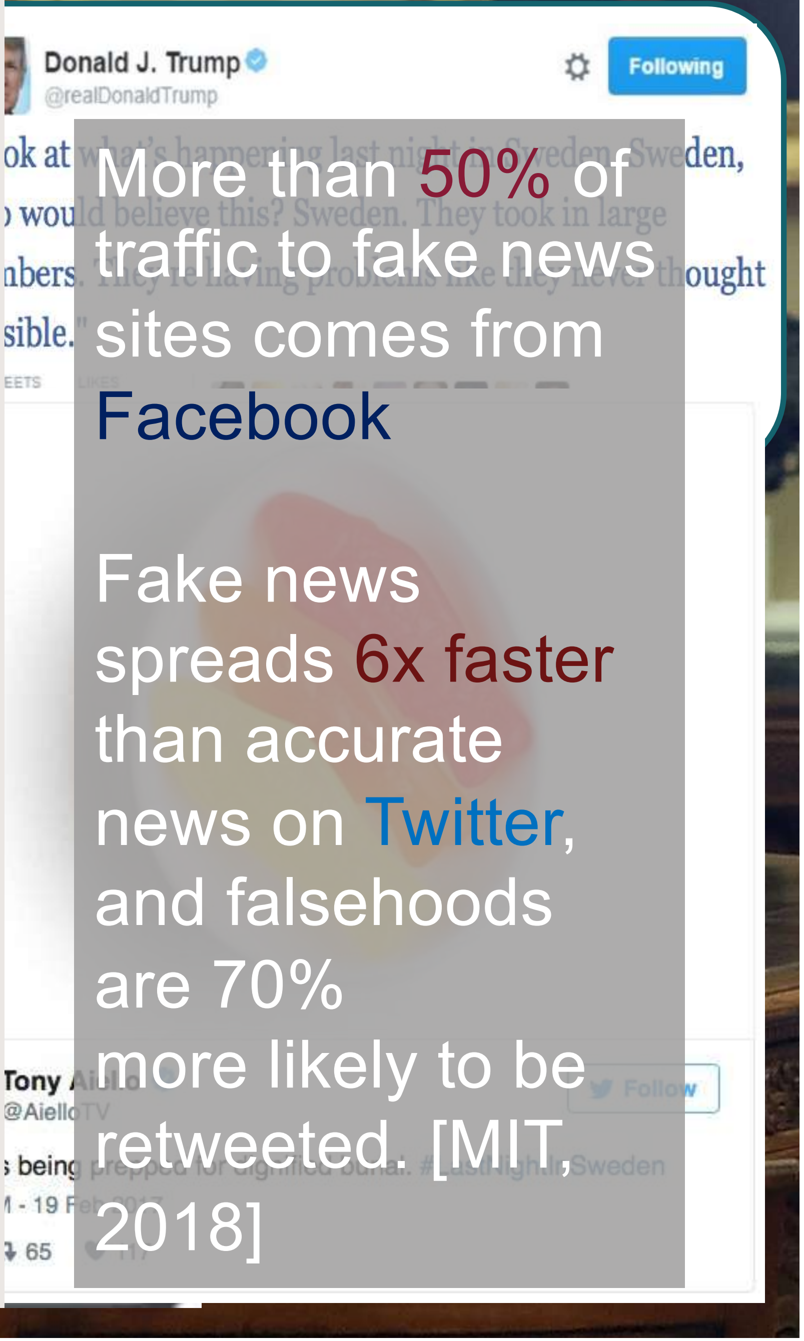
802,000

"FBI Agent Suspected in Hillary Email Leaks Found Dead in Apparent Murder-Suicide" Denver Guardian

567,000

"RAGE AGAINST THE MACHINE To Reunite And Release Anti Donald Trump Album" heaviermetal.net

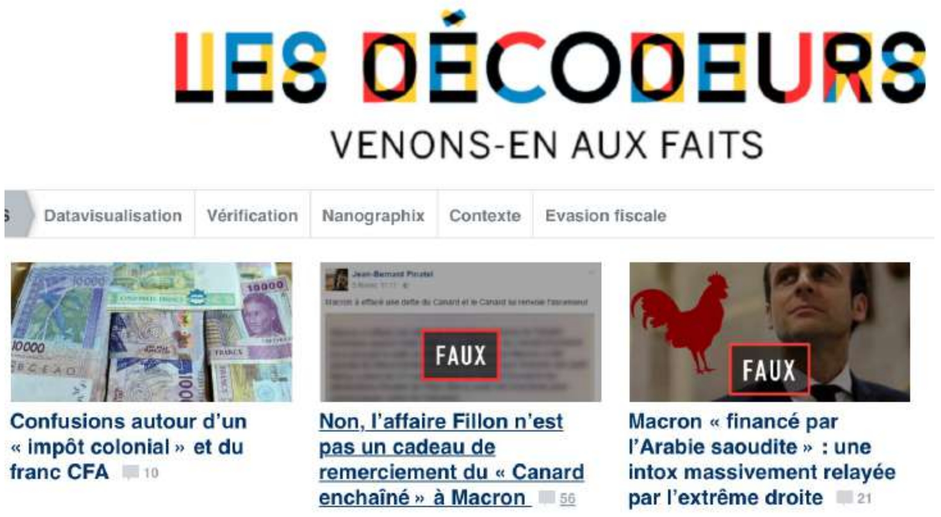
560,000



More than 50% of traffic to fake news sites comes from Facebook

Fake news spreads 6x faster than accurate news on Twitter, and falsehoods are 70% more likely to be retweeted. [MIT, 2018]

Existing tools



LES DÉCODEURS

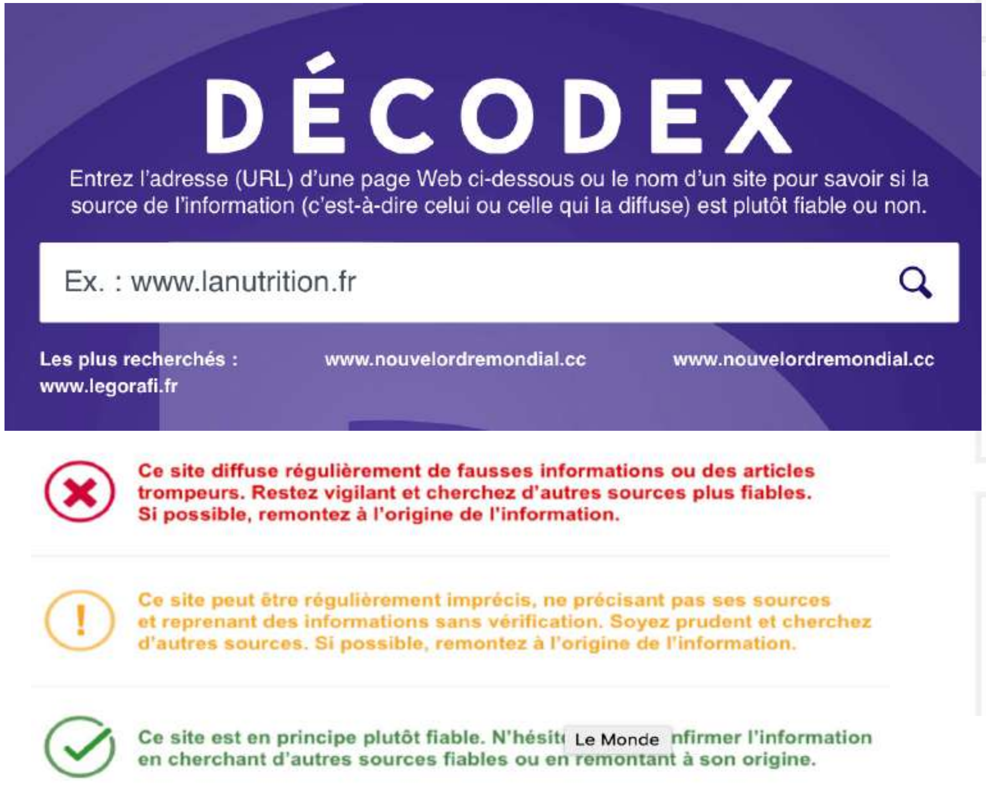
VENONS-EN AUX FAITS

s | Datavisualisation | Vérification | Nanographix | Contexte | Evasion fiscale

Confusions autour d'un « impôt colonial » et du franc CFA 10

Non, l'affaire Fillon n'est pas un cadeau de remerciement du « Canard enchaîné » à Macron 56

Macron « financé par l'Arabie saoudite » : une intox massivement relayée par l'extrême droite 21



DÉCODEX

Entrez l'adresse (URL) d'une page Web ci-dessous ou le nom d'un site pour savoir si la source de l'information (c'est-à-dire celui ou celle qui la diffuse) est plutôt fiable ou non.

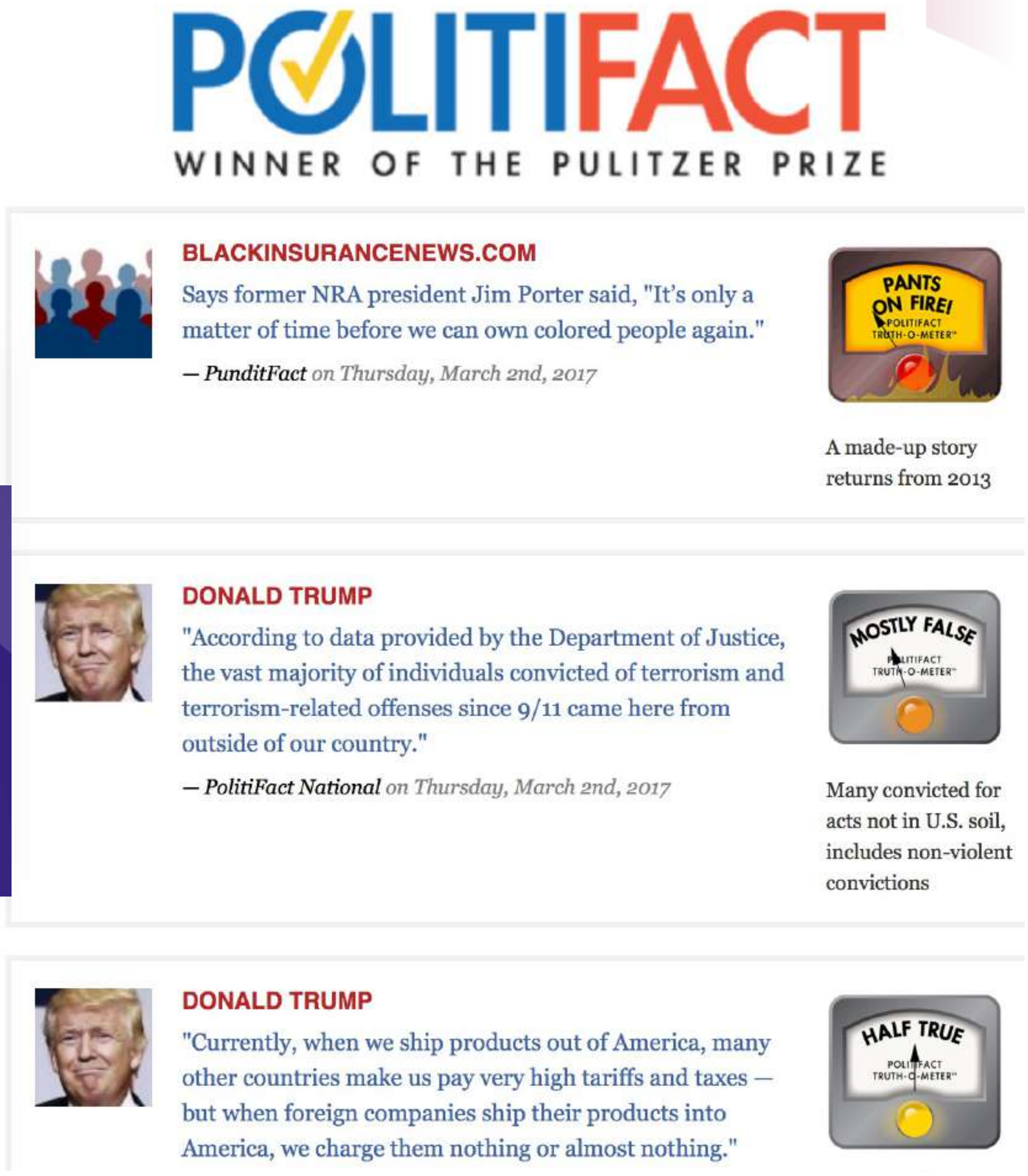
Ex. : www.lanutrition.fr

Les plus recherchés : www.legorafi.fr www.nouvelordremondial.cc www.nouvelordremondial.cc

✗ Ce site diffuse régulièrement de fausses informations ou des articles trompeurs. Restez vigilant et cherchez d'autres sources plus fiables. Si possible, remontez à l'origine de l'information.

! Ce site peut être régulièrement imprécis, ne précisant pas ses sources et reprenant des informations sans vérification. Soyez prudent et cherchez d'autres sources. Si possible, remontez à l'origine de l'information.

✓ Ce site est en principe plutôt fiable. N'hésit Le Monde nfirmer l'information en cherchant d'autres sources fiables ou en remontant à son origine.



POLITIFACT

WINNER OF THE PULITZER PRIZE

BLACKINSURANCENEWS.COM

Says former NRA president Jim Porter said, "It's only a matter of time before we can own colored people again."

— *PunditFact* on Thursday, March 2nd, 2017

A made-up story returns from 2013

DONALD TRUMP

"According to data provided by the Department of Justice, the vast majority of individuals convicted of terrorism and terrorism-related offenses since 9/11 came here from outside of our country."

— *PolitiFact National* on Thursday, March 2nd, 2017

Many convicted for acts not in U.S. soil, includes non-violent convictions

DONALD TRUMP

"Currently, when we ship products out of America, many other countries make us pay very high tariffs and taxes — but when foreign companies ship their products into America, we charge them nothing or almost nothing."

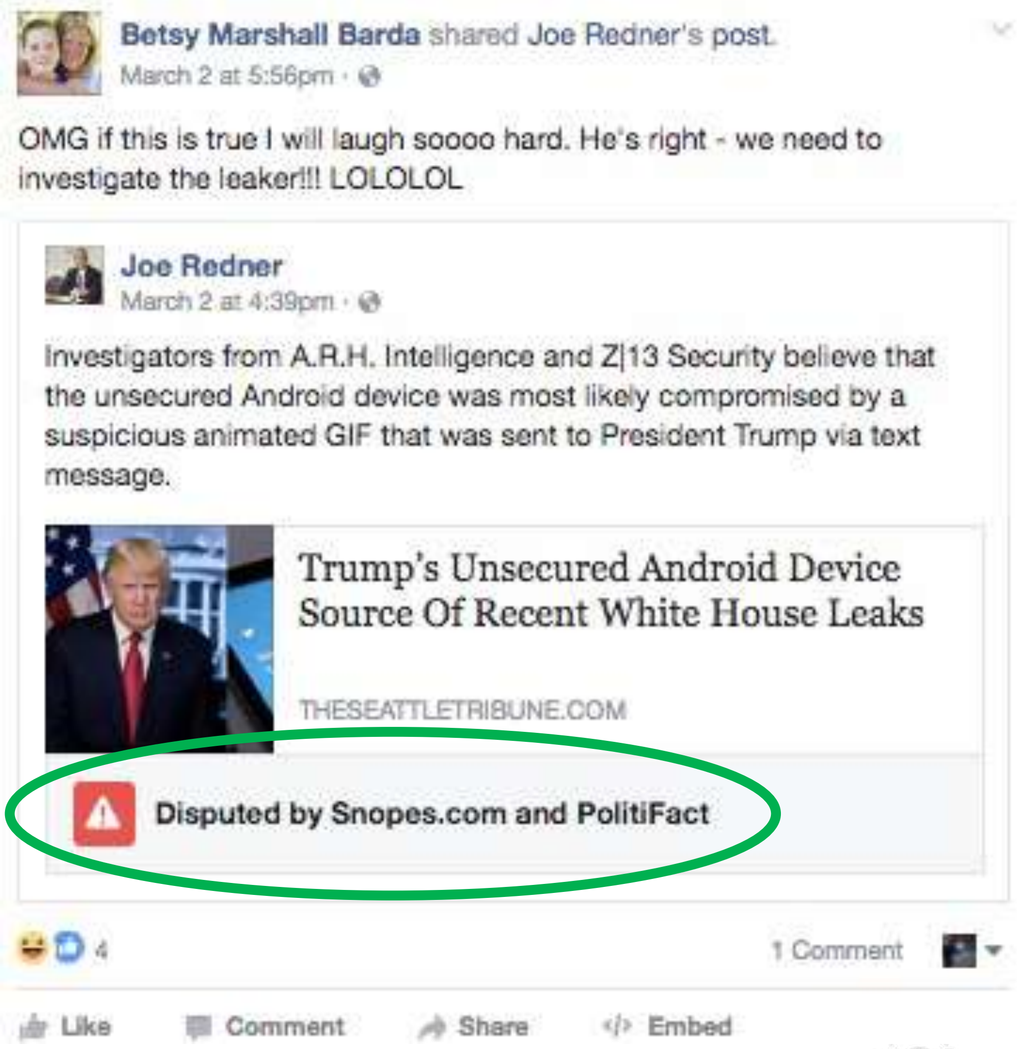
Fighting fake news with Big Data

Facebook to flag fake news stories

- Users report the story as bogus or Facebook's software detects something odd
- Facebook sends the story to some of the organizations that have signed on to provide free fact-checking (e.g. Snopes, Politifact)
- If two of those fact-checkers think it's bogus, the label goes on.

Google presented an algorithm for fake news detection

- Websites are scored according to the accuracy of the of the facts presented



Betsy Marshall Barda shared Joe Redner's post.
March 2 at 5:56pm · 🌐

OMG if this is true I will laugh soooo hard. He's right - we need to investigate the leaker!!! LOLOLOL

Joe Redner
March 2 at 4:39pm · 🌐

Investigators from A.R.H. Intelligence and Z|13 Security believe that the unsecured Android device was most likely compromised by a suspicious animated GIF that was sent to President Trump via text message.

Trump's Unsecured Android Device Source Of Recent White House Leaks

THESEATTLETRIBUNE.COM

⚠ **Disputed by Snopes.com and PolitiFact**

Deep Fake

Machine learning is exceptionally good at learning how to exploit human psychology

- because the internet provides a vast and fast feedback loop to learn what will reinforce and or break beliefs by demographic cohorts
- an AI engine that can generate messages and immediately test if the message is effective

Generative adversarial networks help create AI-generated images and deepfakes

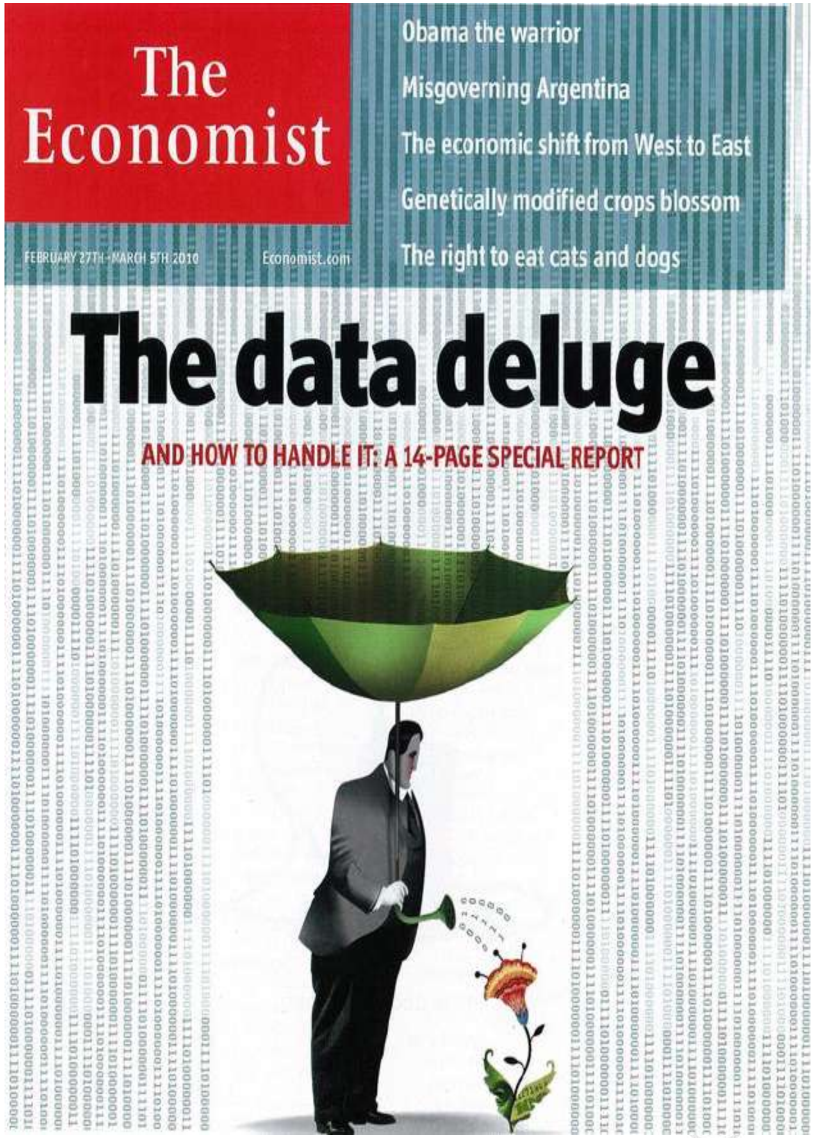
- Two neural networks (a generator and a discriminator) work together to create the fictional image

How *big* is Big Data ?

Total size of the Digital Universe

50 ZetaBytes

in 2021



The Economist

FEBRUARY 27TH–MARCH 5TH 2010

Economist.com

Obama the warrior

Misgoverning Argentina

The economic shift from West to East

Genetically modified crops blossom

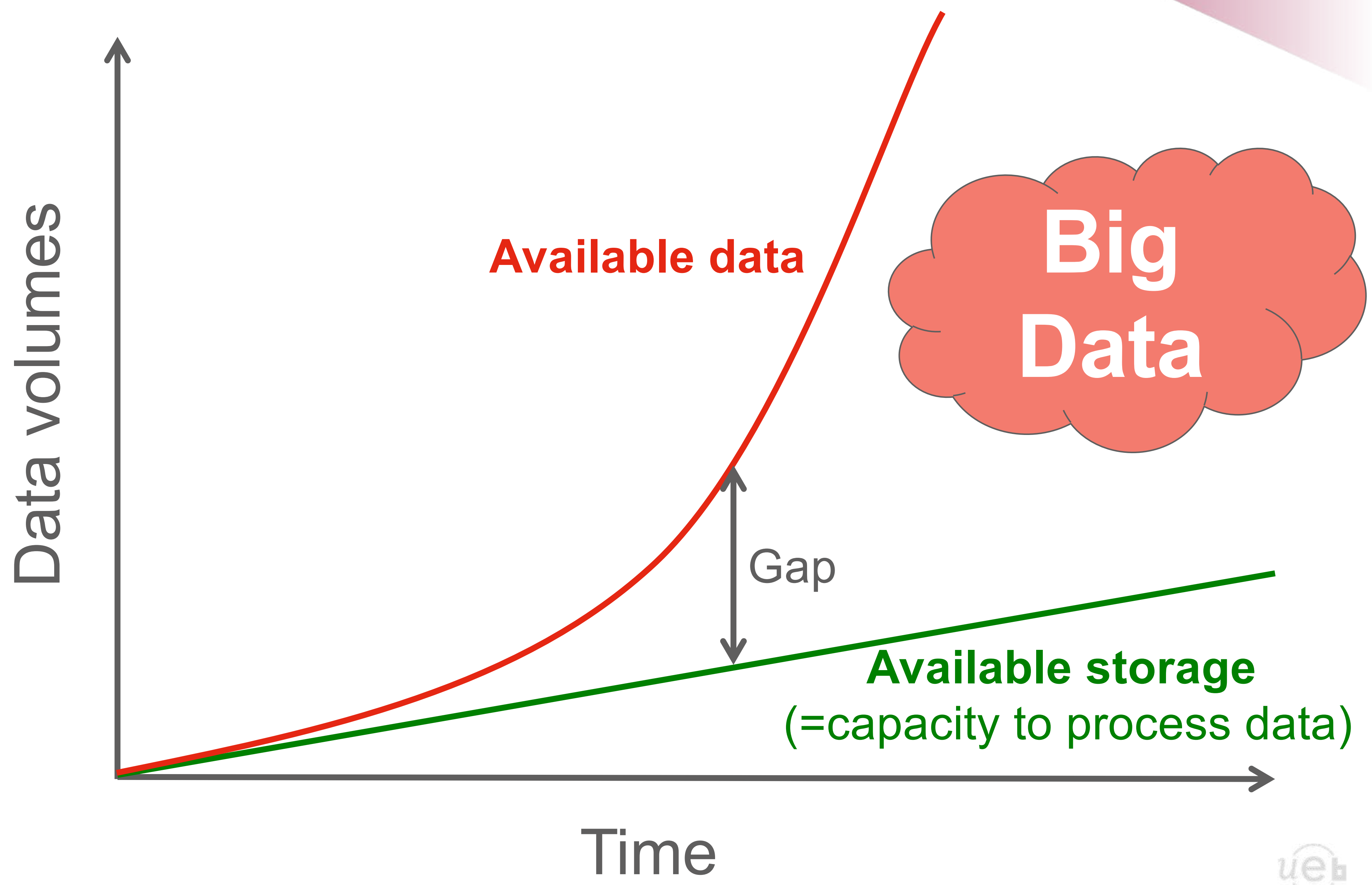
The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

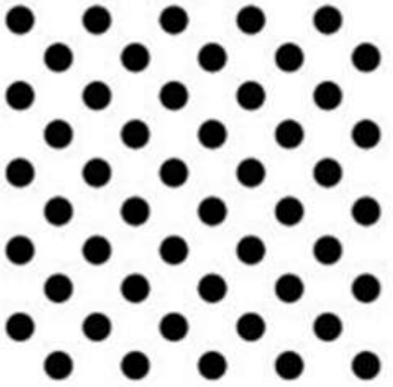
Defining Big Data

What is Big Data ?



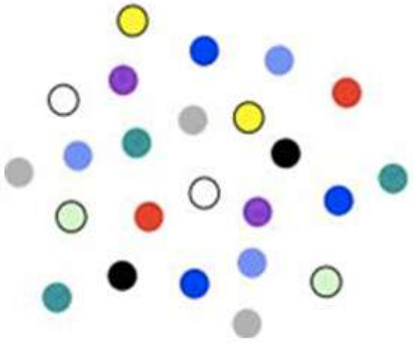
Big Data Features: the 3 Vs

Volume



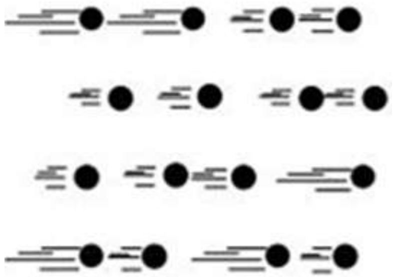
Terabytes to Exabytes of existing data to process

Variety



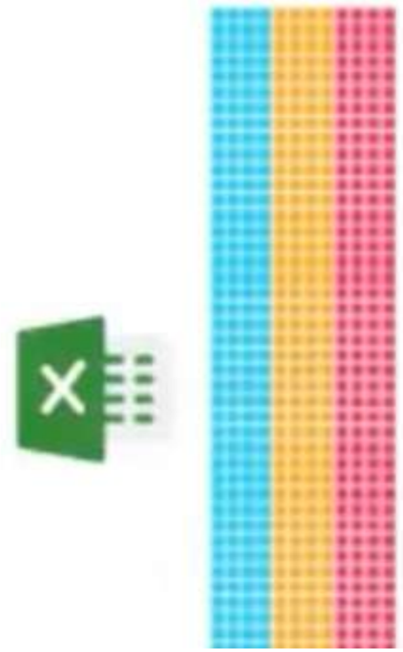
Structured and unstructured data

Velocity

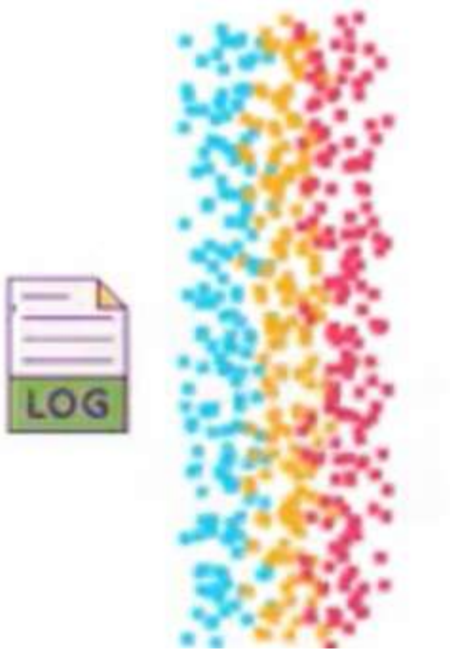


Data requiring milliseconds to seconds to respond

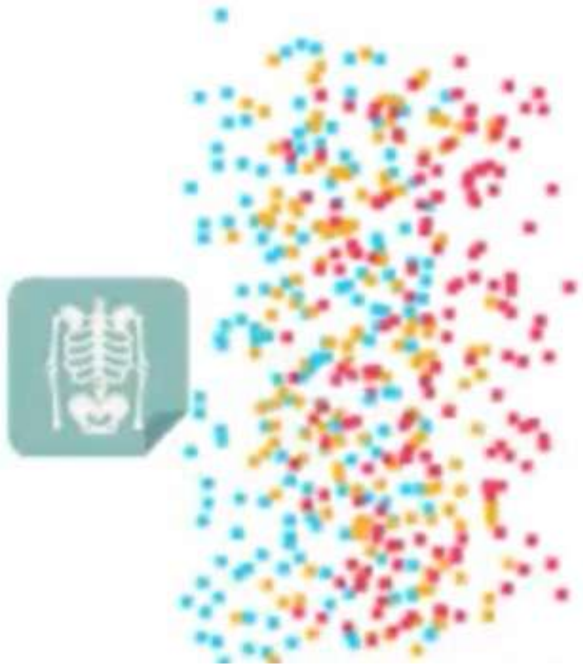
Variety



Structured Data



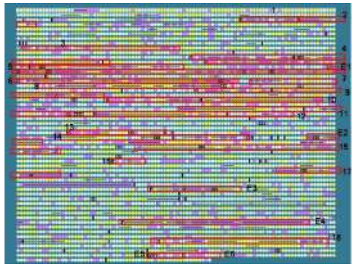
Semi-Structured Data

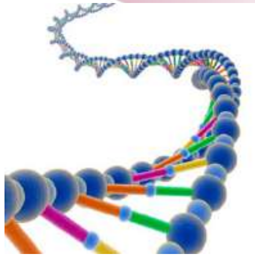


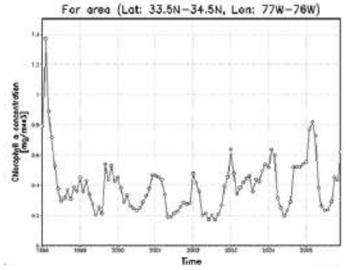
Unstructured Data

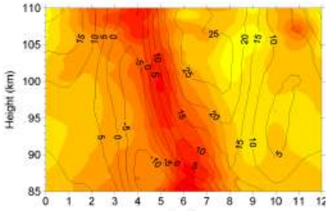
Variety

- **Structured**
  - Relational data: tables, transactions, legacy data
- **Semi-structured**
  - Text: web, documents
  - XML, JSON
  - Large graphs: social networks, semantic web (RDF)
- **Unstructured**
  - Streaming: You can only scan the data once
- A single application can process many types of data





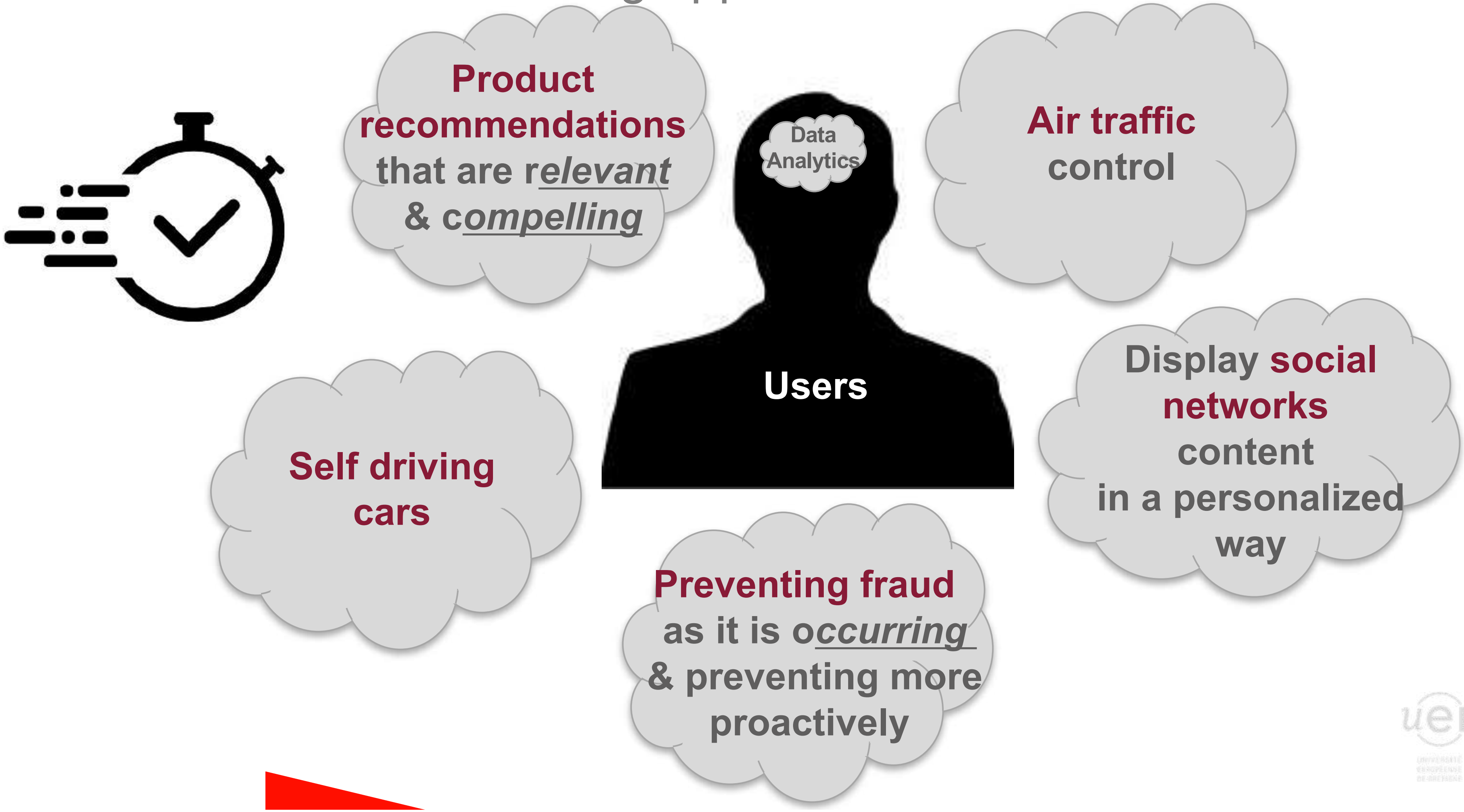

For area (Lat: 33.5N-34.5N, Lon: 77W-76W)
Chlorophyll a concentration [mg/m3]
Time



To extract knowledge→ all these types of data need to be linked together

Velocity

- **Real-time / fast data:** data is being generated fast and need to be processed fast
- Late decisions: missing opportunities



Product recommendations that are *relevant* & *compelling*

Data Analytics

Users

Air traffic control

Self driving cars

Display social networks content in a personalized way

Preventing fraud as it is *occurring* & preventing more proactively

Processing Big Data

What to do with all these data ?

- Stored data = **Costs**
- Information from data = **Profit**

Goal: Extract valuable information / added value from these huge data

What is needed?

Hardware: distributed infrastructures

– Cloud computing allows to lease computing and storage resources

Google Cloud

Azure

amazon web services™

Software: new programming models

– MapReduce: simple yet scalable model for Big Data processing

Google

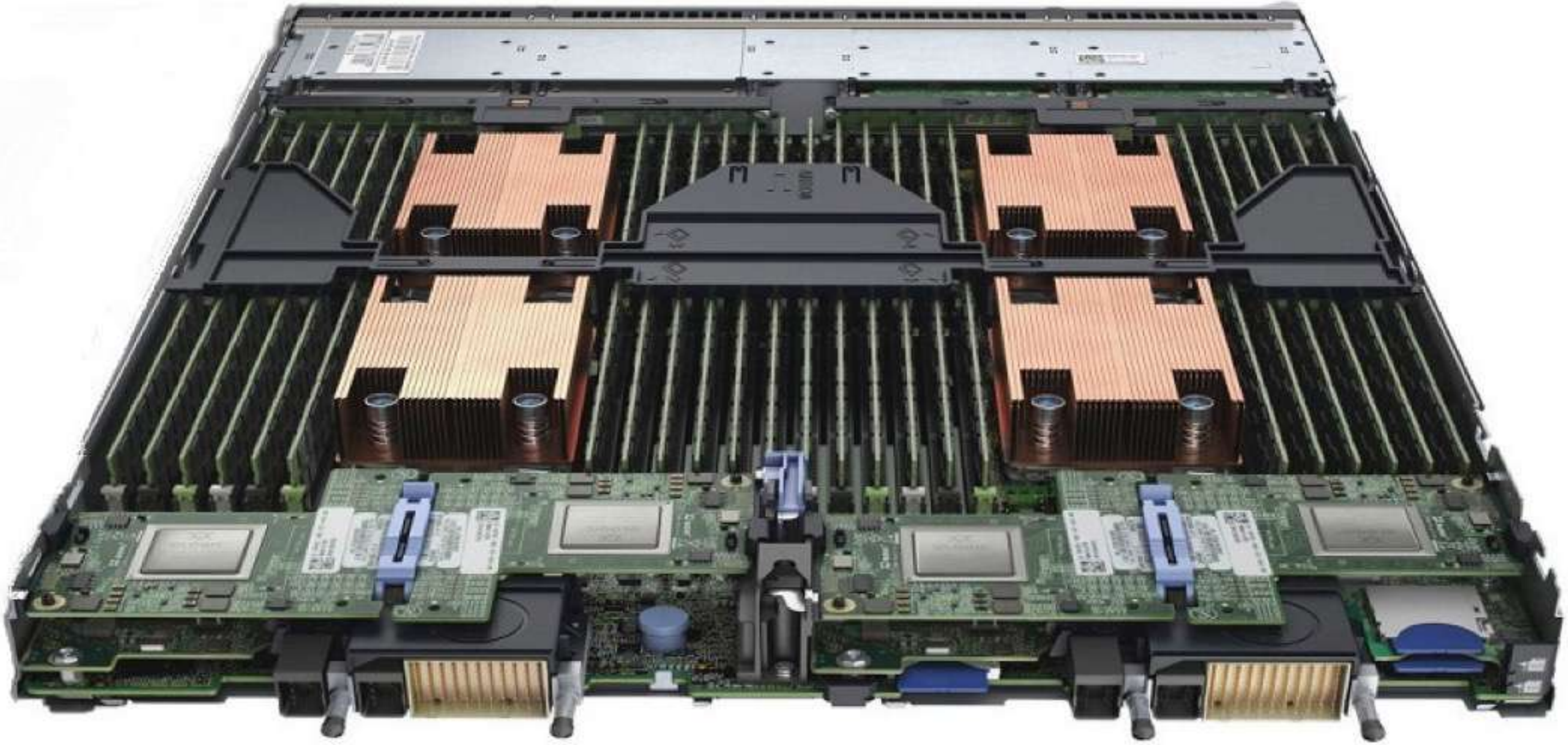
hadoop

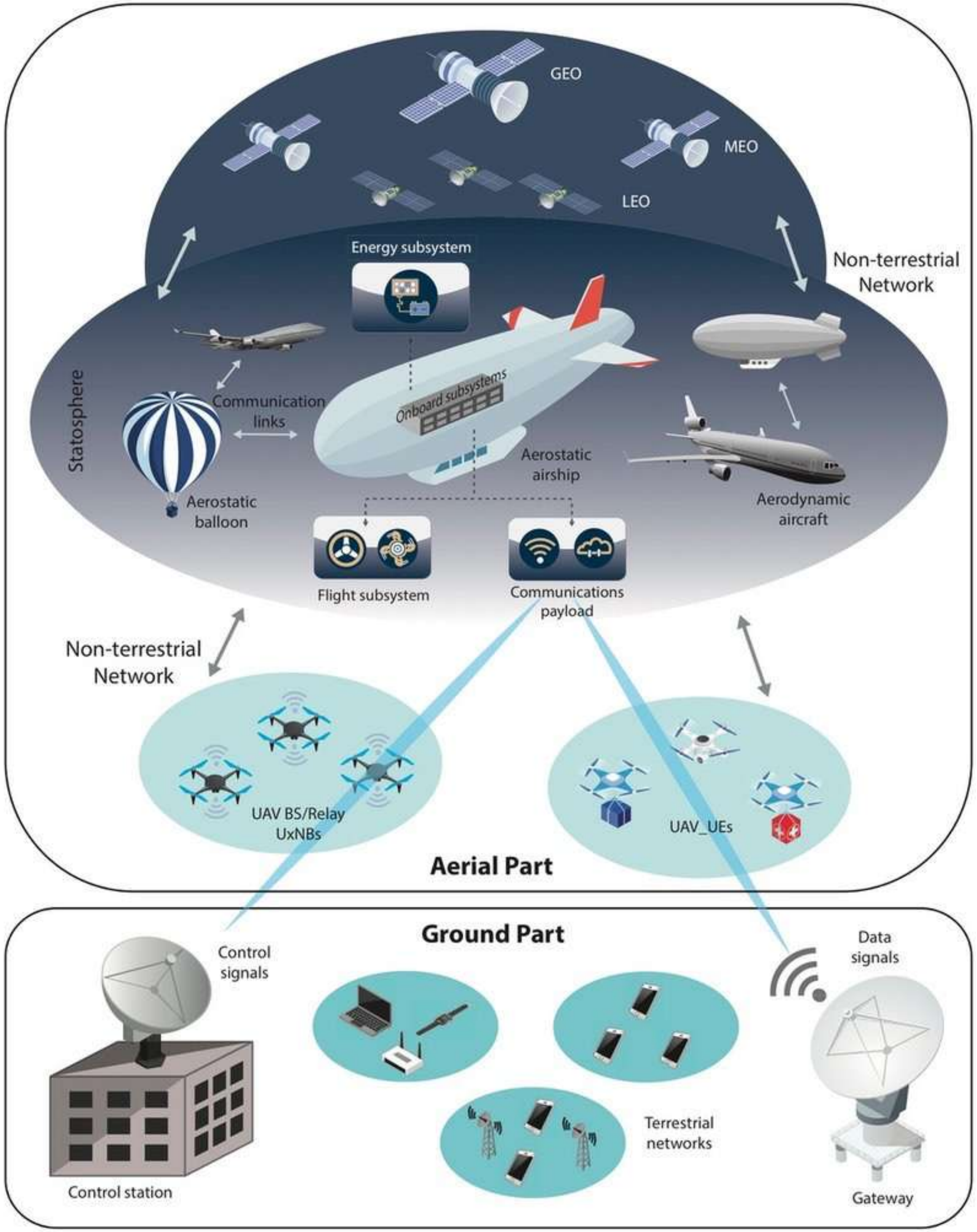
Cloud Computing

- Everything as a Service

- The delivery of computing as a service rather than a product
- Shared resources, software, and information are provided as a metered service over a network (typically the Internet)

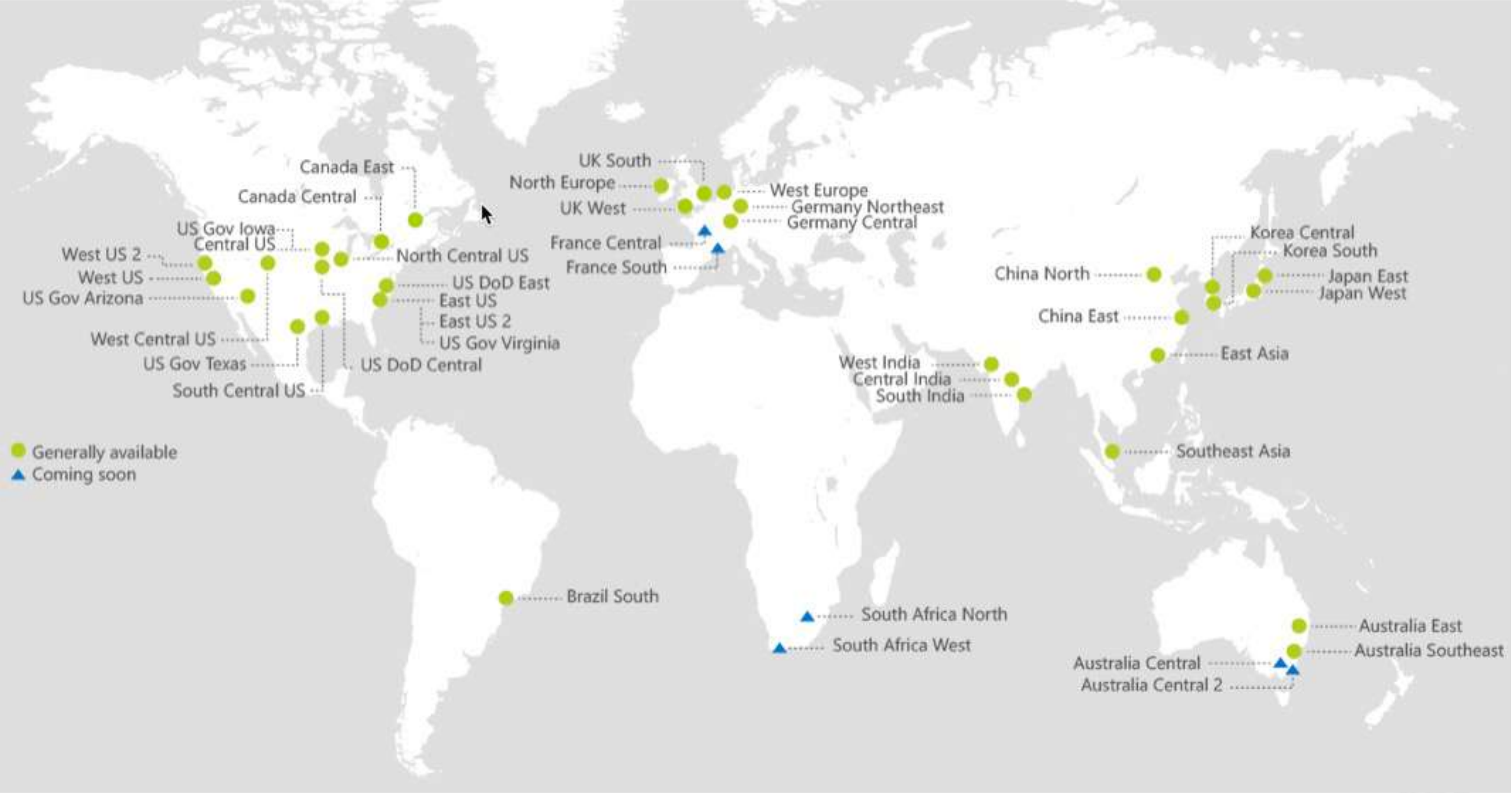
What is the cloud? One datancenter?




GEO
MEO
LEO
Non-terrestrial Network
Energy subsystem
Statosphere
Communication links
Onboard subsystems
Aerostatic airship
Aerostatic balloon
Aerodynamic aircraft
Flight subsystem
Communications payload
Non-terrestrial Network
UAV BS/Relay UxNBs
UAV_UEs
Aerial Part
Ground Part
Control signals
Data signals
Terrestrial networks
Control station
Gateway

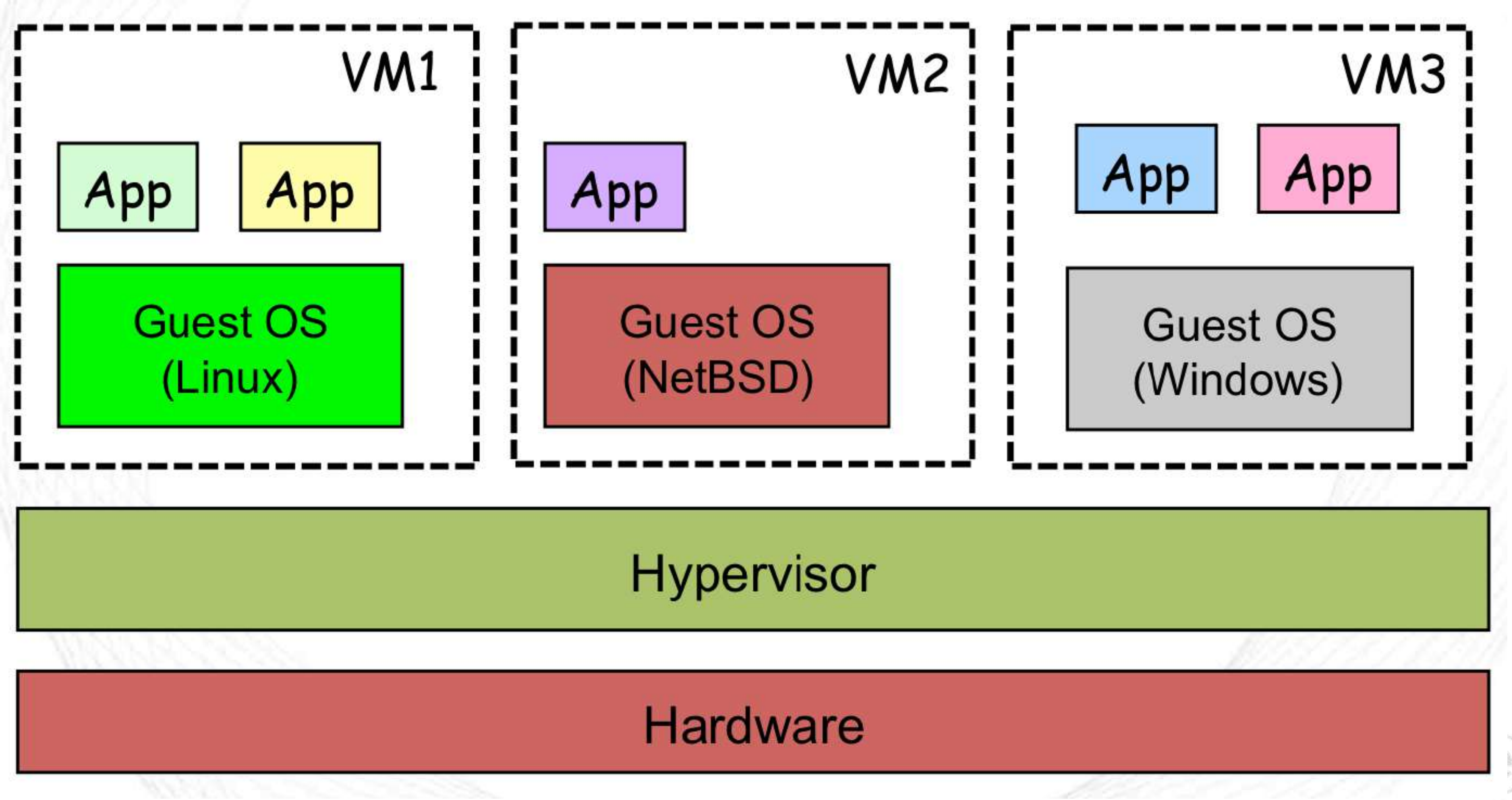
What is the cloud? One datancenter?

Many geographically distributed datacenters

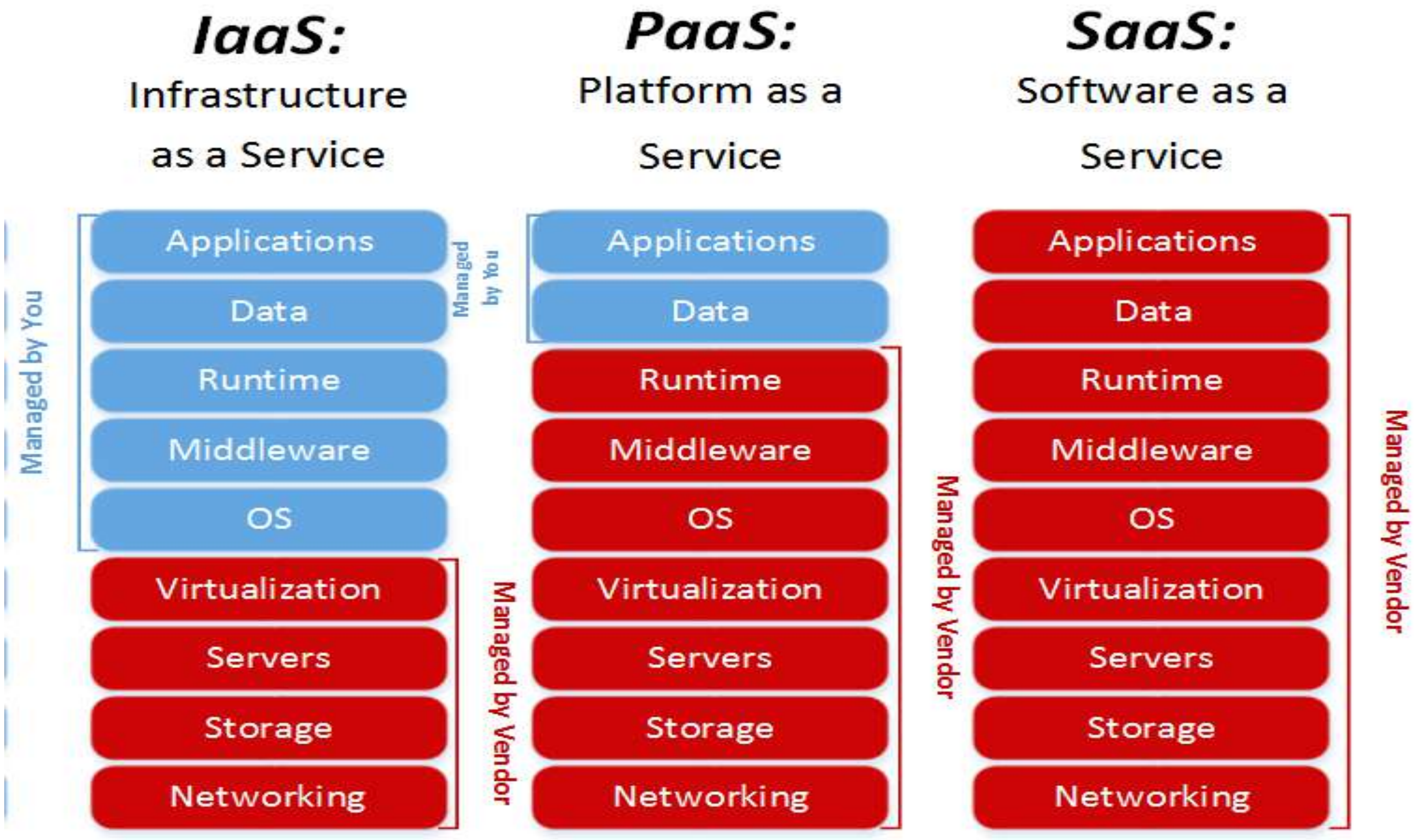


Enabling technology: virtualisation

- Allows multiple virtual machines to run on a single physical machine



Types of clouds



Bare **hardware**: storage and compute

Development **environments** to create applications

Business **applications** (e.g. Dropbox, Office 365)

BIG DATA IS GOING TO IMPACT US SOON!
CALL BRUCE WILLIS!!
DATA
Soulcié
www.nuxeo.com