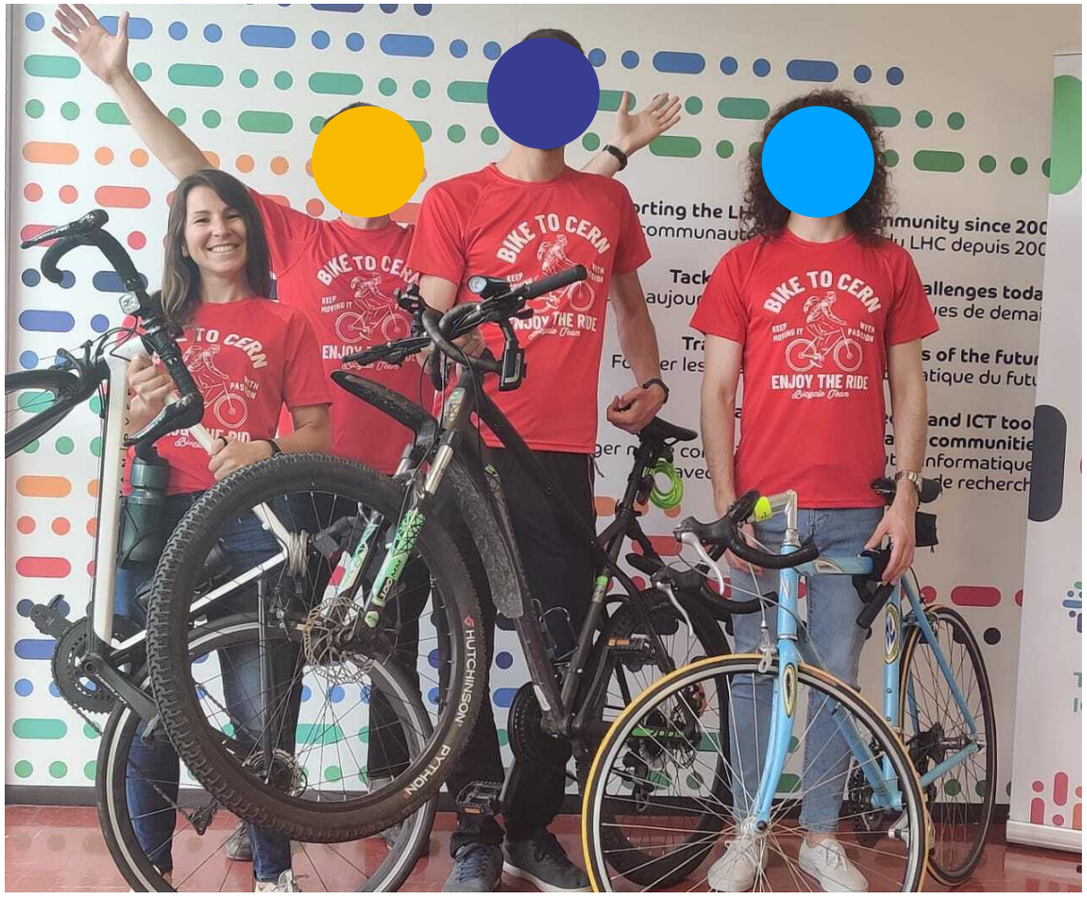
CERN Science for Open Data

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Preliminary Requests

- Raise the hand for any questions
- Pay attention

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

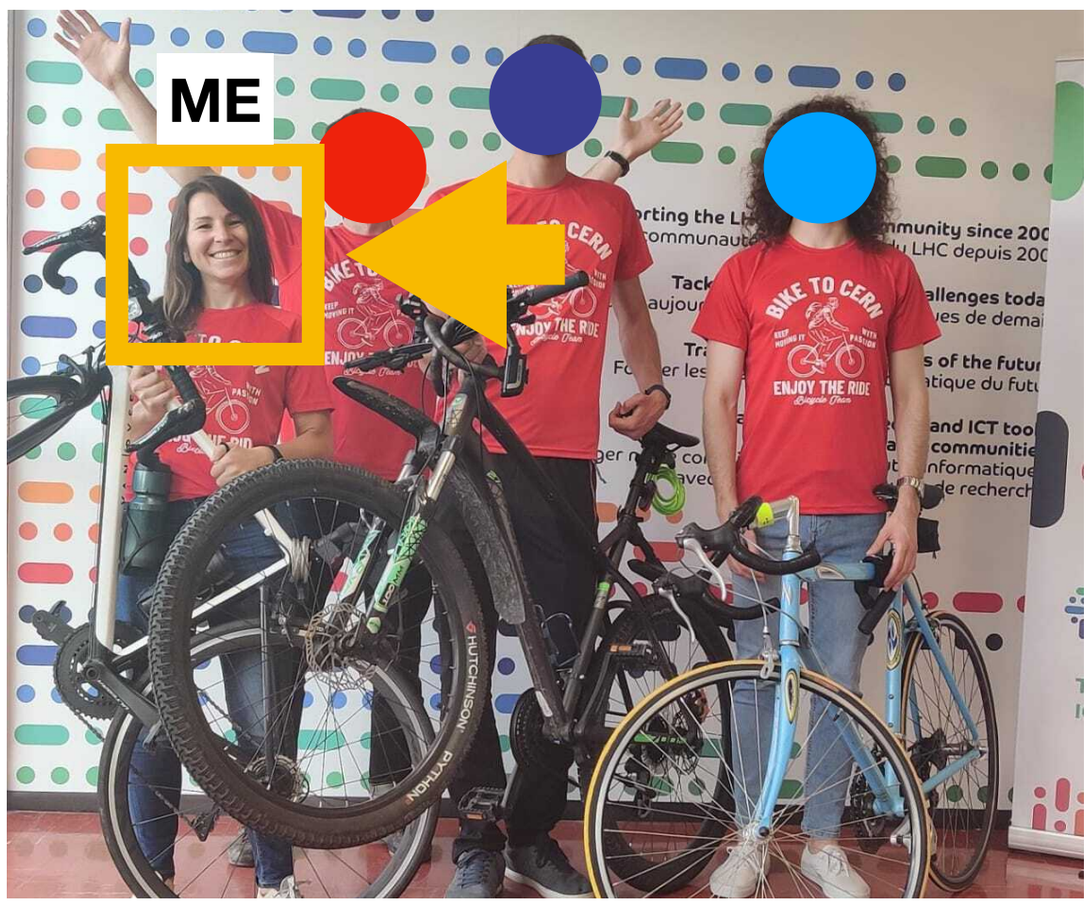


Anna Ferrari

- Graduation in **Statistics & Mathematics**
- PhD in **Machine learning techniques for Human Activity Recognition**
- Working as **Data Scientist at CERN openlab**
- **Keywords:** #CERN, #science, #passion, #data, #datascience, #datapassion
- **Hobbies:** guess!

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch



Anna Ferrari

- Graduation in **Statistics & Mathematics**
- PhD in **Machine learning techniques for Human Activity Recognition**
- Working as **Data Scientist at CERN openlab**
- **Keywords:** #CERN, #science, #passion, #data, #datascience, #datapassion
- **Hobbies:** guess!

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Summary

- **Openness** - definition & values

- **Open Data & Big Data**

- **CERN Science for Open Data**

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Openness

What?

collaboration

sharing

accessible

opportunity and possibilities

transparency

democracy

Why?

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

sharing

What?

Exchanging feelings, ideas

Have or use something at the same time as someone else

Divide and give part of something to someone else

Put something in a webpage so that other people can see it

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

sharing

What?

Exchanging feelings, ideas

Have or use something at the same time as someone else

Divide and give part of something to someone else

Put something in a webpage so that other people can see it

Image source: Future Lab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

sharing

What?

Exchanging feelings, ideas

Have or use something at the same time as someone else

Divide and give part of something to someone else

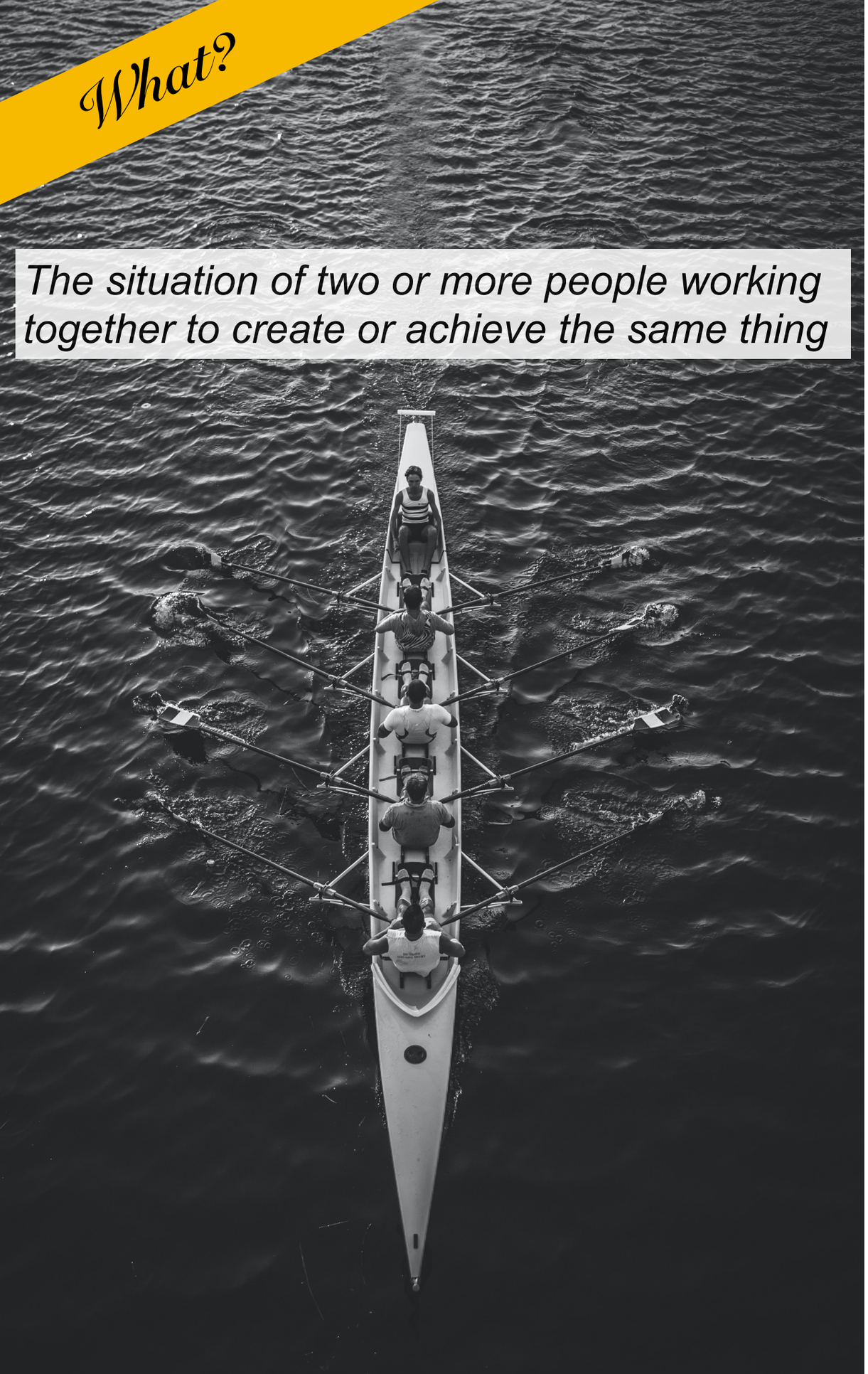
Put something in a webpage so that other people can see it

https://stories.cogdogblog.com/

Blogs
Code sharing
Knowledge sharing

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

collaboration

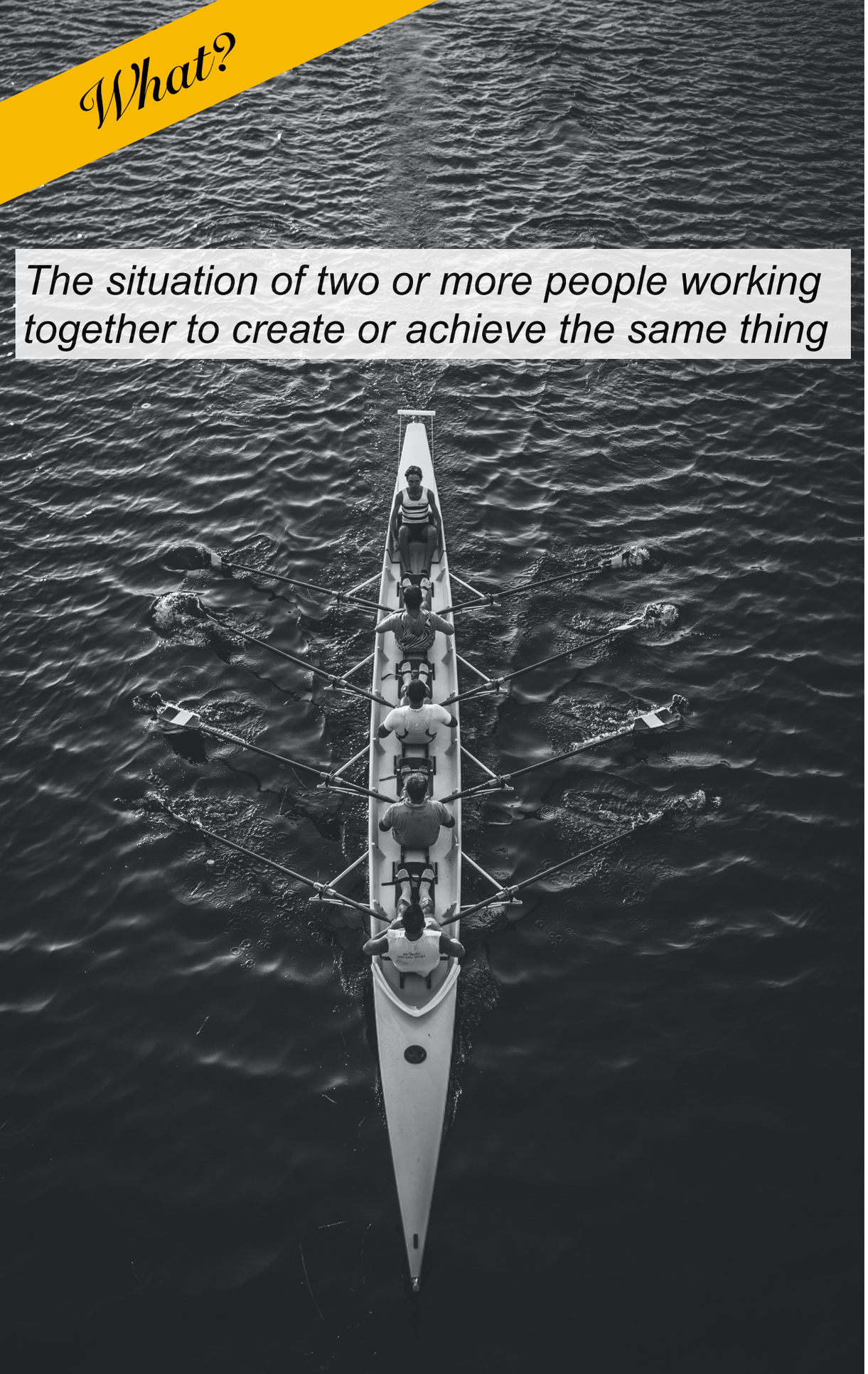


***Exercice**:*
Write in the chat the first example of collaboration that comes to your mind

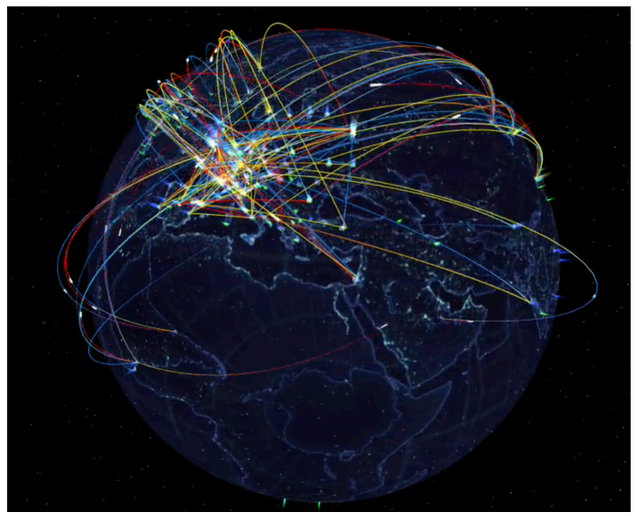
CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

collaboration



WIKIPEDIA
The Free Encyclopedia



The polymath blog

LHC@home

SETI@home

folding@home

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

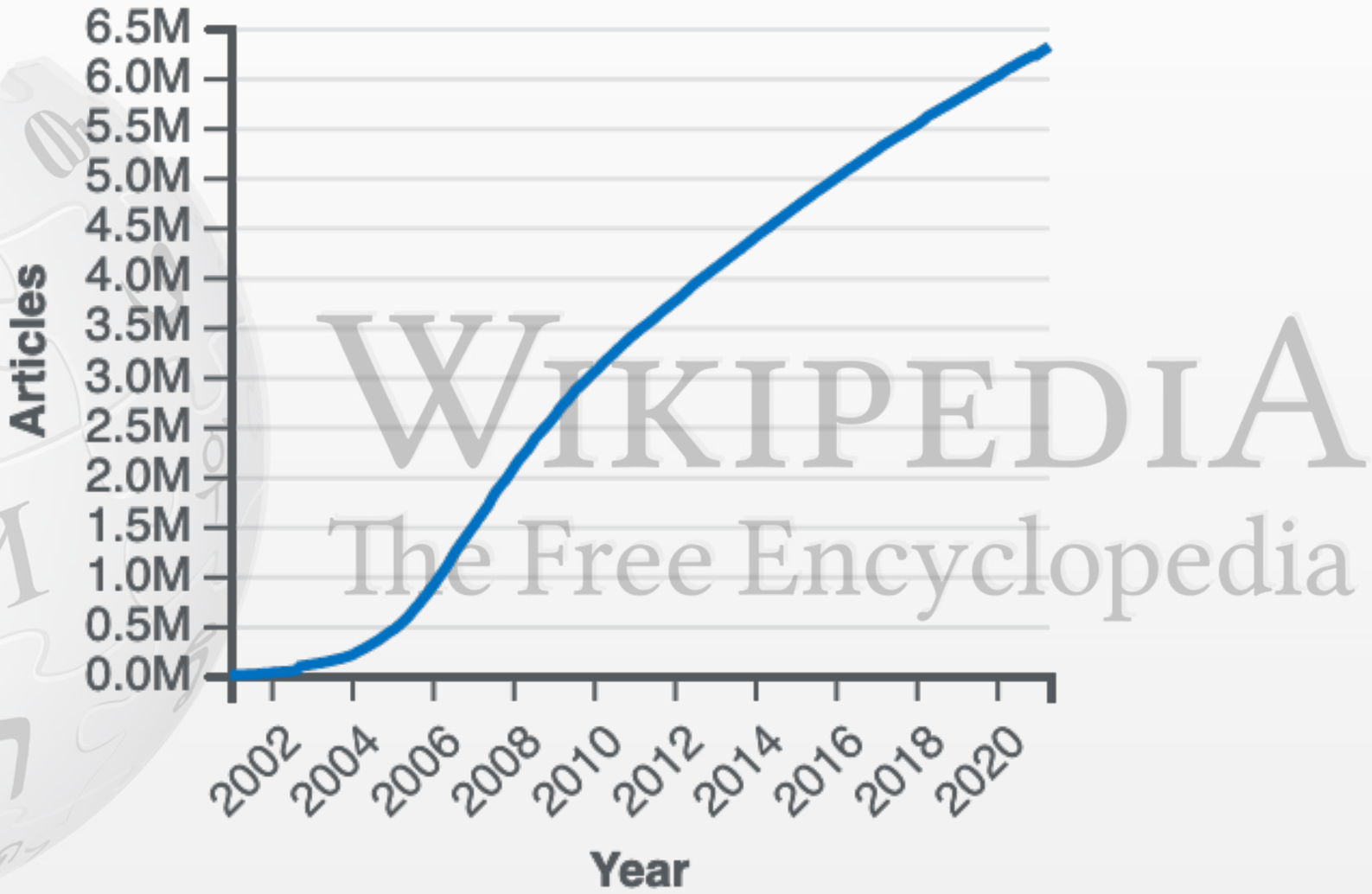
collaboration

What?

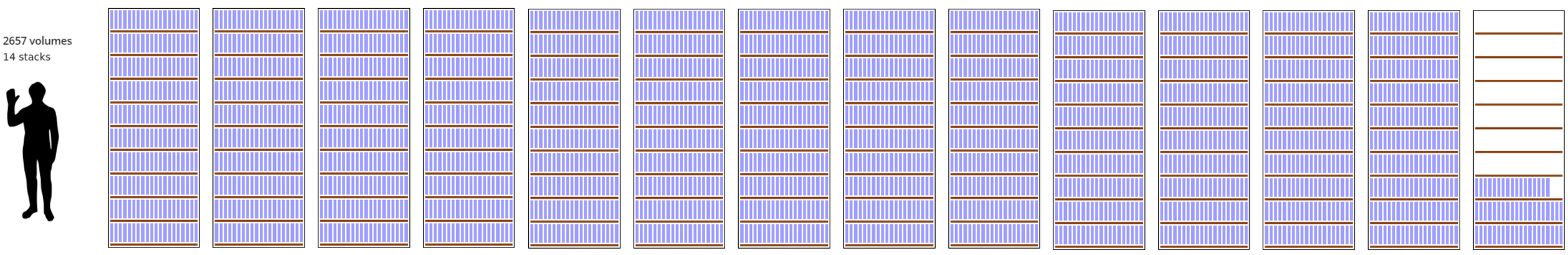
? languages

? articles

? page views



WIKIPEDIA
The Free Encyclopedia



CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

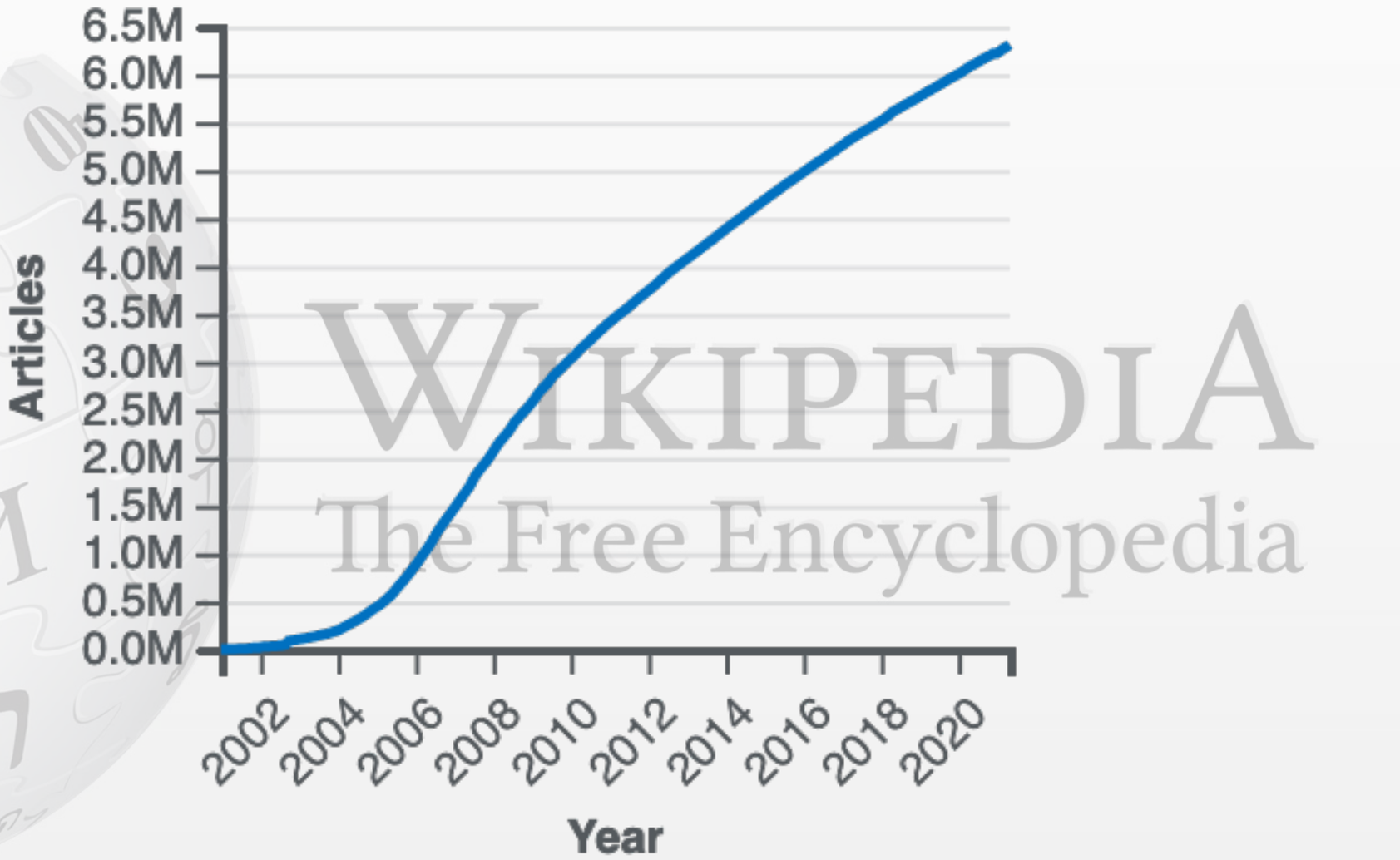
collaboration

What?

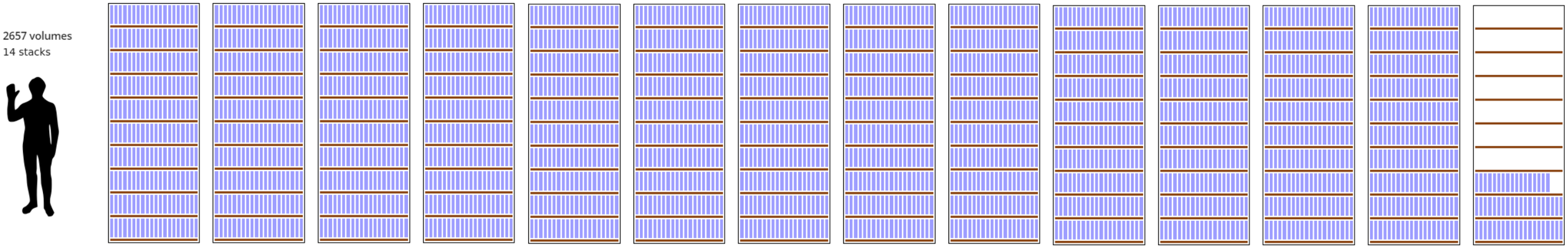
323 languages

6M+ articles

21B+ page views



WIKIPEDIA
The Free Encyclopedia



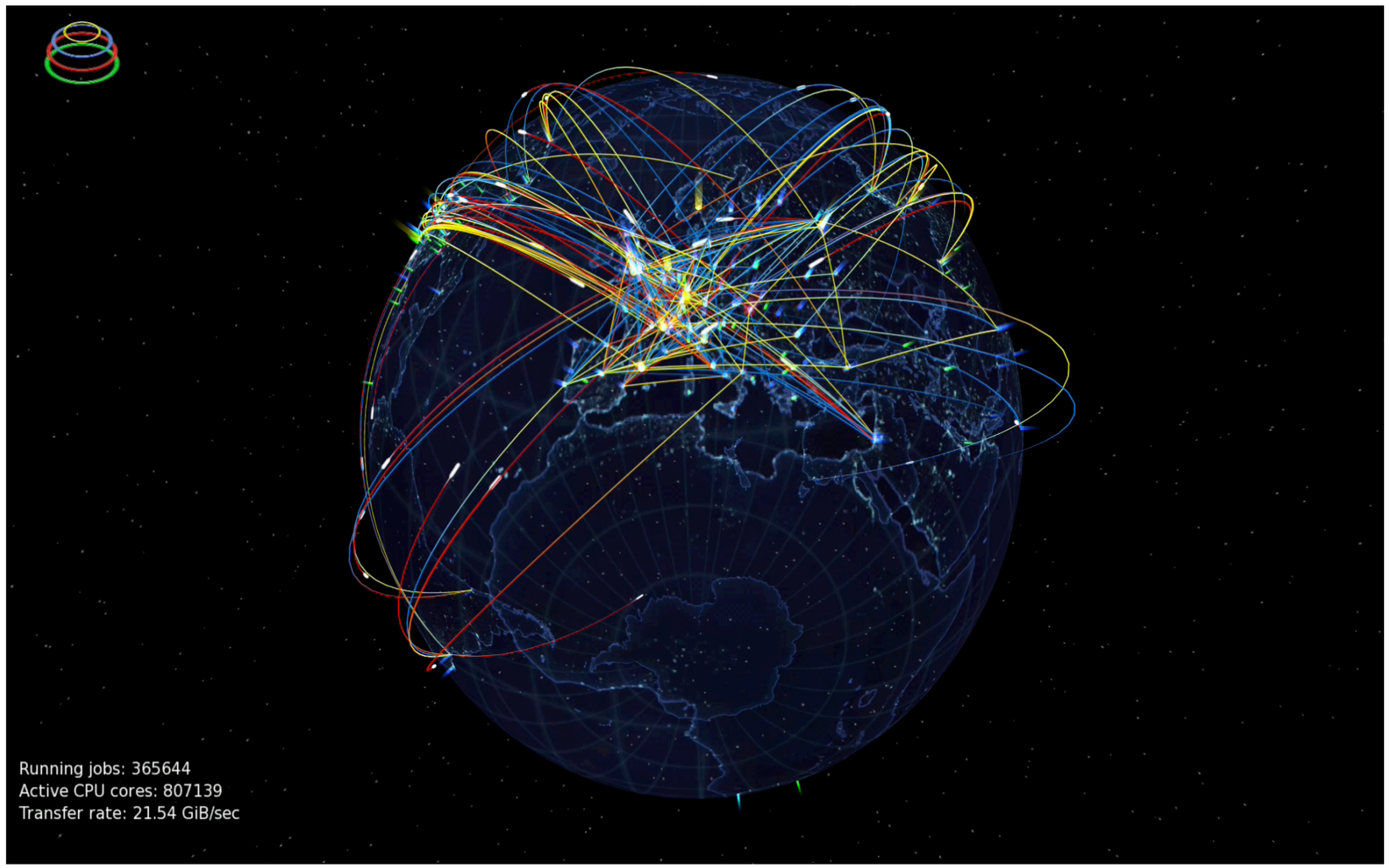
CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

What?

Daily

- 170 collaborating centres
- 42 countries around the world
- more than 2 million ‘tasks’
- 1 million of computer cores



Worldwide LHC Computing Grid

… corresponding to a single computer running for more than ? years.

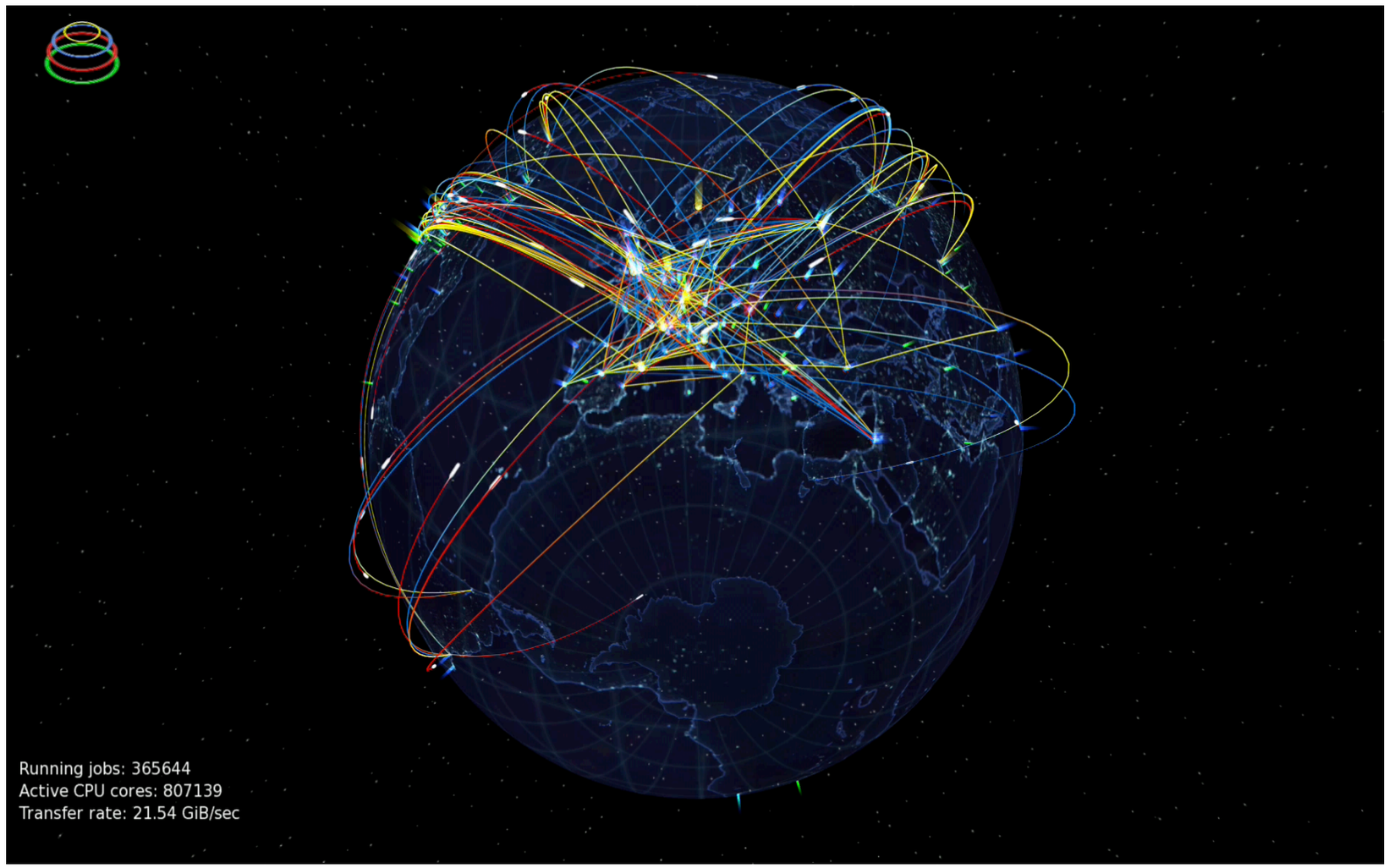
Anna Ferrari
CERN openlab
a.ferrari@cern.ch

collaboration

What?

Daily

- 170 collaborating centres
- 42 countries around the world
- more than 2 million ‘tasks’
- 1 million of computer cores



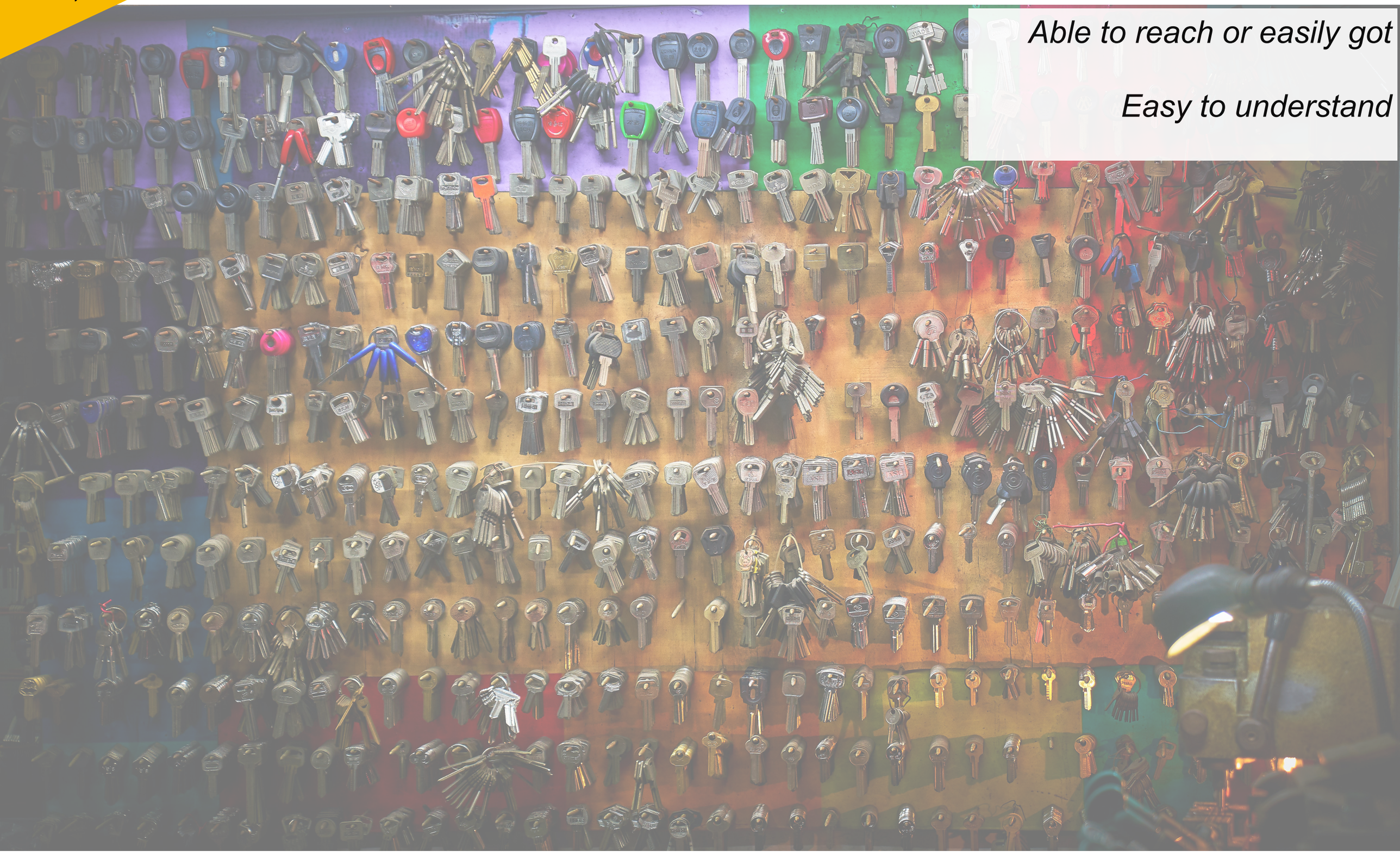
Worldwide LHC Computing Grid

… corresponding to a single computer running for more than **600 years**.

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

What?

accessibility



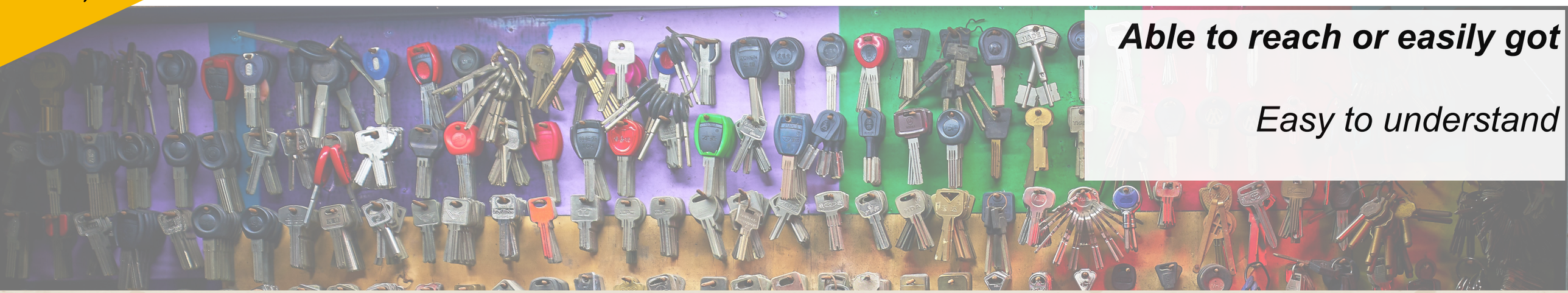
Able to reach or easily got

Easy to understand

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

accessibility

What?



Able to reach or easily got

Easy to understand

THE EVOLUTION OF COMMUNICATION

CAVE PAINTINGS & STONE CARVINGS
30,000 BCE

MARATHON MAN
530 BCE

COMMON ERA

TELEPHONE
1876

TELEGRAPH
1840

1927
TELEVISION
BROADCAST

INSTANT MESSAGING
1997

BLOGGING
1999

INTERNET
1969

CARRIER PIGEONS
776 BCE

FIRST DAILY NEWS PAPER
1650

RADIO SIGNAL
1902

WORLD WIDE WEB
1994

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

What?

CERN openlab

accessibility

Able to reach or easily got

Easy to understand

JAN 2021

DIGITAL AROUND THE WORLD

ESSENTIAL HEADLINES FOR MOBILE, INTERNET, AND SOCIAL MEDIA USE

INTERNET USER NUMBERS NO LONGER INCLUDE DATA SOURCED FROM SOCIAL MEDIA PLATFORMS, SO VALUES ARE NOT COMPARABLE WITH PREVIOUS REPORTS

TOTAL POPULATION	UNIQUE MOBILE PHONE USERS	INTERNET USERS*	ACTIVE SOCIAL MEDIA USERS*
7.83 BILLION	**5.22 BILLION**	**?**	**4.20 BILLION**
URBANISATION: **56.4%**	vs. POPULATION: **66.6%**		vs. POPULATION: **53.6%**

8

SOURCES: THE U.N.; LOCAL GOVERNMENT BODIES; GSMA INTELLIGENCE; ITU; GWI; EUROSTAT; CNNIC; APJII; SOCIAL MEDIA PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; COMPANY EARNINGS REPORTS; MEDIASCOPE. ***ADVISORIES:** INTERNET USER NUMBERS NO LONGER INCLUDE DATA SOURCED FROM SOCIAL MEDIA PLATFORMS, SO VALUES ARE NOT COMPARABLE TO DATA PUBLISHED IN PREVIOUS REPORTS. SOCIAL MEDIA USER NUMBERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. ◈ **COMPARABILITY ADVISORY:** SOURCE AND BASE CHANGES.

we are social · Hootsuite®

What?

CERN openlab

accessibility

Able to reach or easily got

Easy to understand

JAN 2021

DIGITAL AROUND THE WORLD

ESSENTIAL HEADLINES FOR MOBILE, INTERNET, AND SOCIAL MEDIA USE

INTERNET USER NUMBERS NO LONGER INCLUDE DATA SOURCED FROM SOCIAL MEDIA PLATFORMS, SO VALUES ARE NOT COMPARABLE WITH PREVIOUS REPORTS

TOTAL POPULATION	UNIQUE MOBILE PHONE USERS	INTERNET USERS*	ACTIVE SOCIAL MEDIA USERS*
7.83 BILLION	**5.22 BILLION**	**4.66 BILLION**	**4.20 BILLION**
URBANISATION: **56.4%**	vs. POPULATION: **66.6%**	vs. POPULATION: **59.5%**	vs. POPULATION: **53.6%**

8

SOURCES: THE U.N.; LOCAL GOVERNMENT BODIES; GSMA INTELLIGENCE; ITU; GWI; EUROSTAT; CNNIC; APJII; SOCIAL MEDIA PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; COMPANY EARNINGS REPORTS; MEDIASCOPE. ***ADVISORIES:** INTERNET USER NUMBERS NO LONGER INCLUDE DATA SOURCED FROM SOCIAL MEDIA PLATFORMS, SO VALUES ARE NOT COMPARABLE TO DATA PUBLISHED IN PREVIOUS REPORTS. SOCIAL MEDIA USER NUMBERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. ◈ **COMPARABILITY ADVISORY:** SOURCE AND BASE CHANGES.

we are social · Hootsuite®

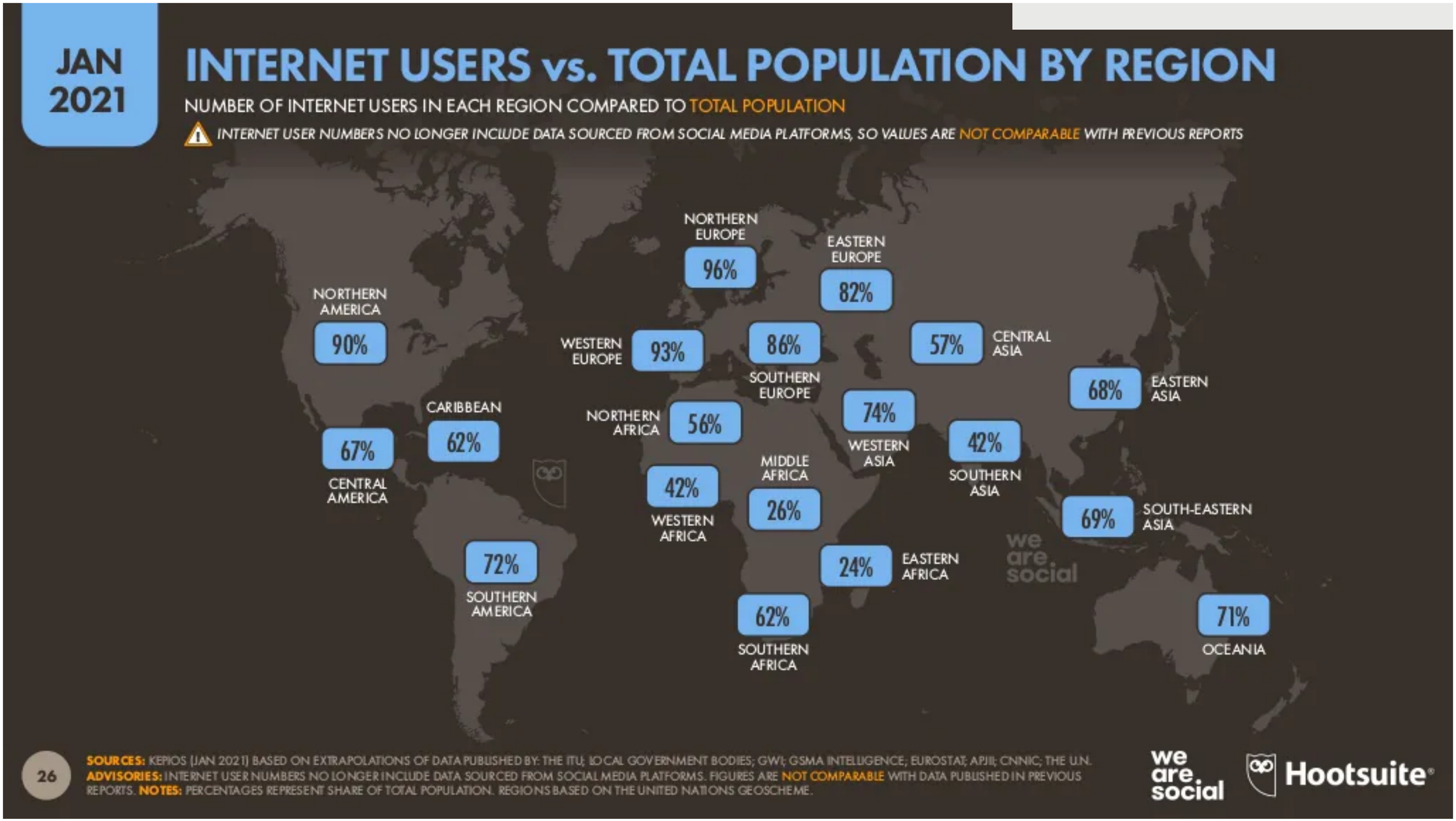
What?

CERN openlab

accessibility

Able to reach or easily got

Easy to understand



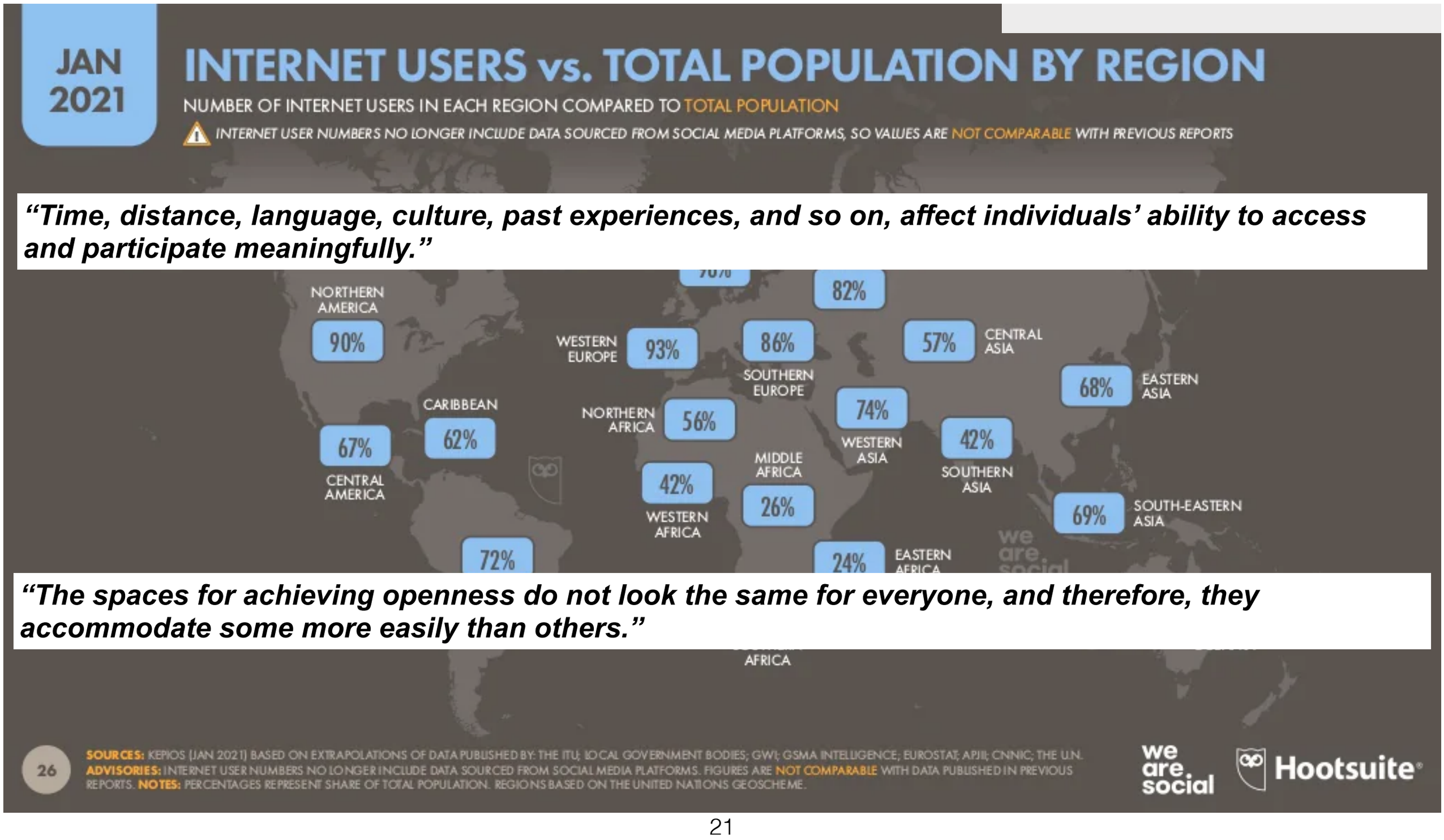
What?

CERN openlab

accessibility

Able to reach or easily got

Easy to understand



"Time, distance, language, culture, past experiences, and so on, affect individuals' ability to access and participate meaningfully."

"The spaces for achieving openness do not look the same for everyone, and therefore, they accommodate some more easily than others."

What?

accessibility

Able to reach or easily got

Easy to understand

"A clearly written treatise is open only to those who are literate and can read the particular language in which it is written."

Long, P. O., 1991

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

What?

accessibility

	EU	
	DESI 2018	DESI 2020
2a1 At least basic digital skills % individuals	**57%** 2017	**58%** 2019
2a2 Above basic digital skills % individuals	**31%** 2017	**33%** 2019
2a3 At least basic software skills % individuals	**60%** 2017	**61%** 2019
2b1 ICT specialists % total employment	**3.7%** 2016	**3.9%** 2018
2b2 Female ICT specialists % female employment	**1.3%** 2016	**1.4%** 2018
2b3 ICT graduates % graduates	**3.5%** 2015	**3.6%** 2017

Source: Digital Economy and Society Index, 2020 European Commission

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

What?

accessibility

	EU	
	DESI 2018	DESI 2020
2a1 At least basic digital skills % individuals	**57%** 2017	**58%** 2019
2a2 Above basic digital skills % individuals	**31%** 2017	**33%** 2019
2a3 At least basic software skills % individuals	**60%** 2017	**61%** 2019
2b1 ICT specialists % total employment	**3.7%** 2016	**3.9%** 2018
2b2 Female ICT specialists % female employment	**1.3%** 2016	**1.4%** 2018
2b3 ICT graduates % graduates	**3.5%** 2015	**3.6%** 2017

Source: Digital Economy and Society Index, 2020 European Commission

"Although already 85% of EU citizens used the internet in 2019, prior to the COVID-19 crisis, only **58%** *possessed at least basic digital skills."*

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Openness

What?

collaboration

sharing

accessible

opportunity and possibilities

transparency

democracy

Why?

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

democracy and transparency

A system that allows the citizens to participate in political decision-making

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

democracy and transparency

A system that allows the citizens to participate in political decision-making

Open business

Open education

Open government

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

democracy and transparency

Open business

Open education

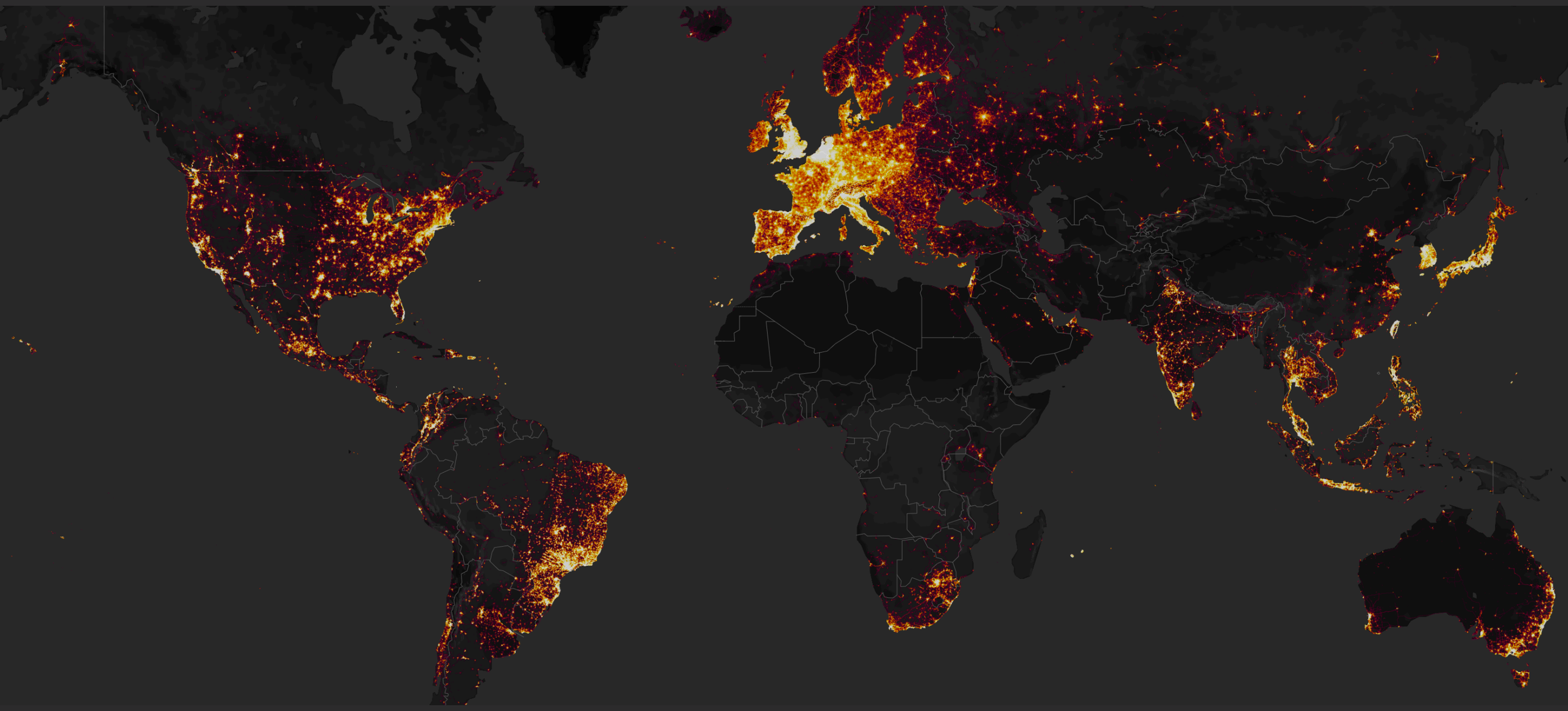
Open government

Data

CERN openlab

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

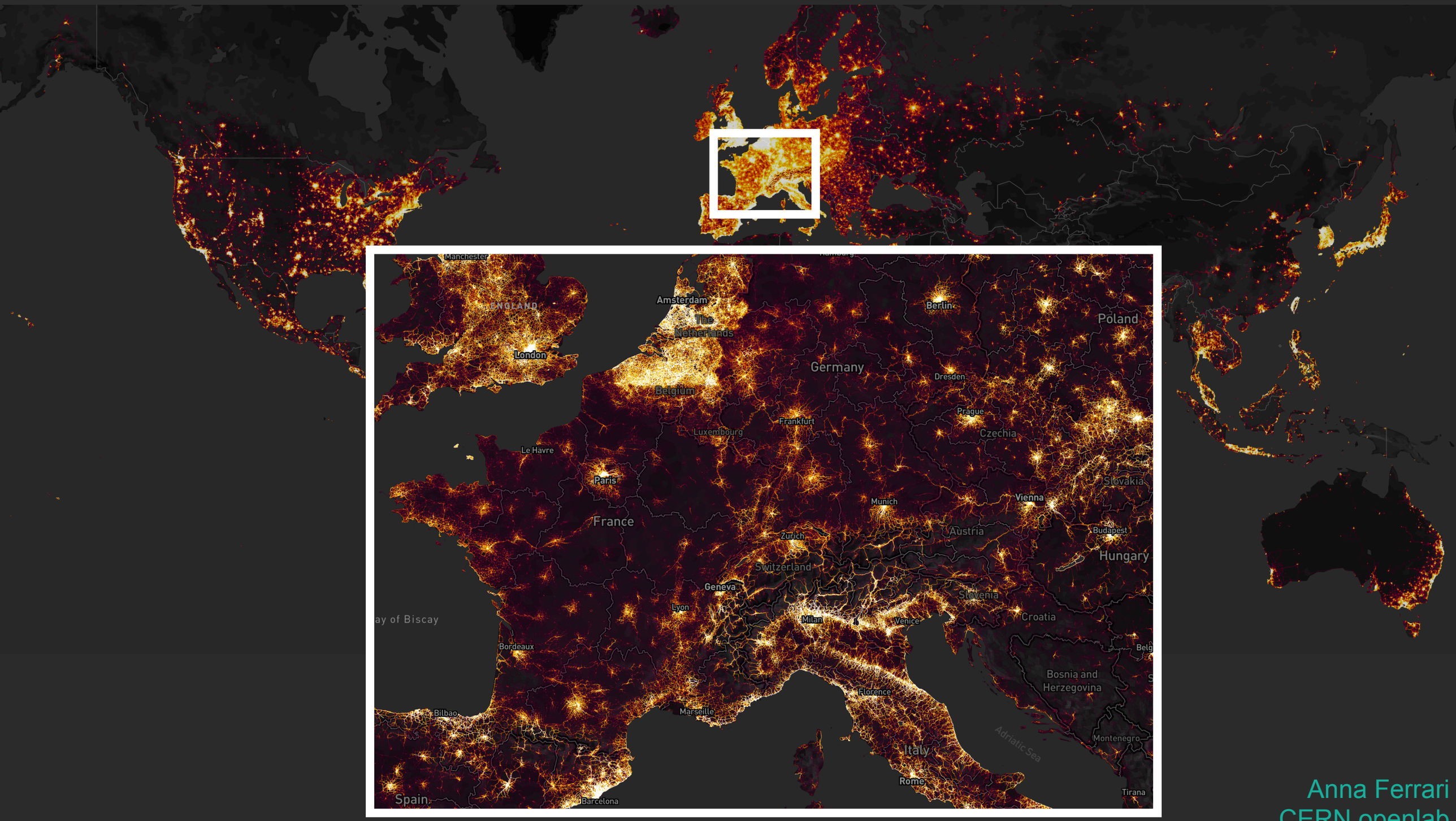
Data & Big Data



Source:strava

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

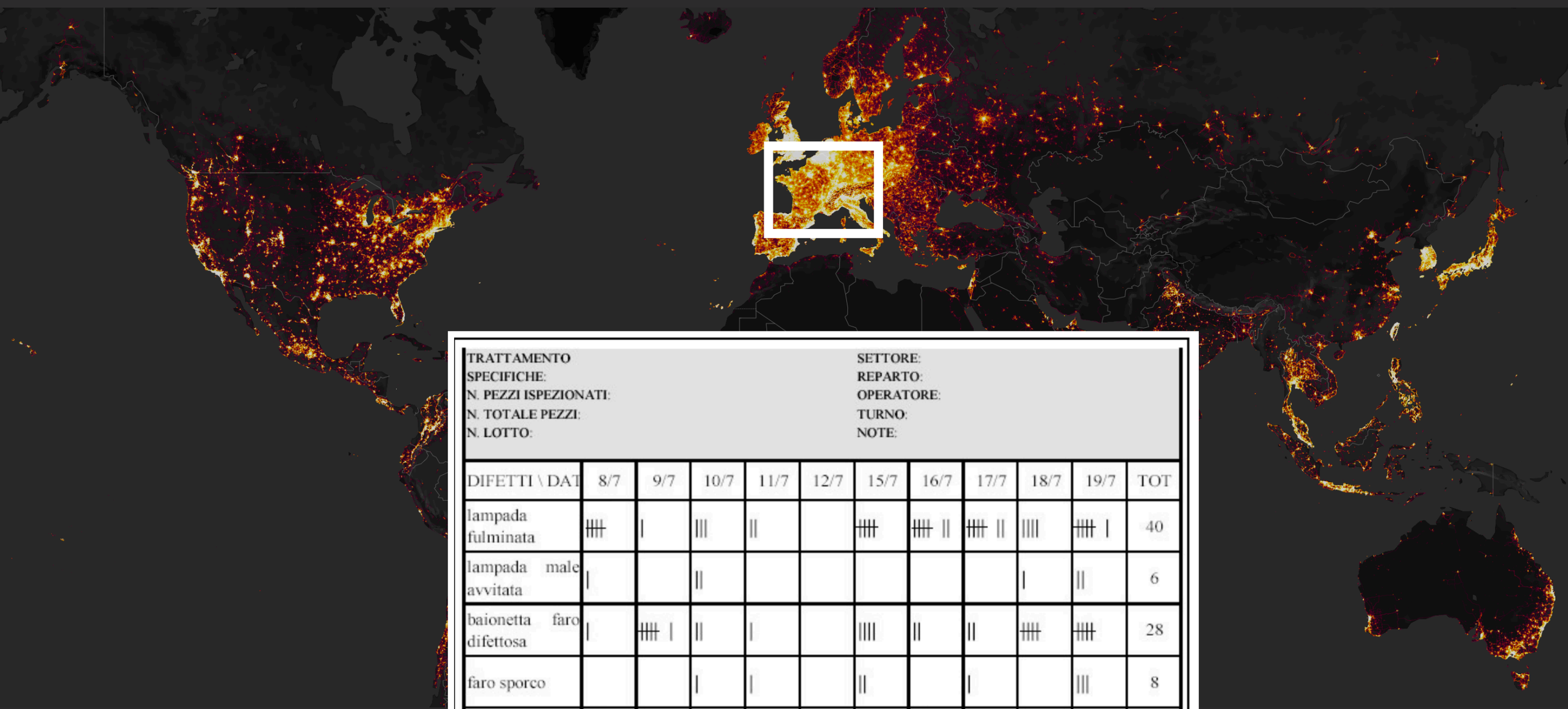
Data & Big Data



Source:strava

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data



TRATTAMENTO SPECIFICHE: N. PEZZI ISPEZIONATI: N. TOTALE PEZZI: N. LOTTO:								SETTORE: REPARTO: OPERATORE: TURNO: NOTE:			
DIFETTI \ DAT	8/7	9/7	10/7	11/7	12/7	15/7	16/7	17/7	18/7	19/7	TOT
lampada fulminata	卌	|	|||	||		卌	卌 ||	卌 ||	||||	卌 |	40
lampada male avvitata	|		||						|	||	6
baionetta faro difettosa	|	卌 |	||	|		||||	||	||	卌	卌	28
faro sporco			|	|		||		|		|||	8
lampada sporca	卌 卌 ||	||	|	卌 ||||	||	|||	卌 卌	|||	卌 |	|||	52
altri	|	|			||		||	||	|||		11
TOTALE	20	10	9	13	4	14	21	15	19	20	145

Source:strava

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Data Size

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Knowledge produced by data

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Quality, availability, completeness

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Data visualisation to better understanding respecting spreadsheets

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Data heterogeneity in format and sources

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Speed in which data growth, changes and become accessible

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Capability of spread of data due to technology and data properties

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Speed of spread and sharing of information across people

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Data Size

Speed of spread and sharing of information across people

Knowledge produced by data

Capability of spread of data due to technology and data properties

Quality, availability, completeness

Speed in which data growth, changes and become accessible

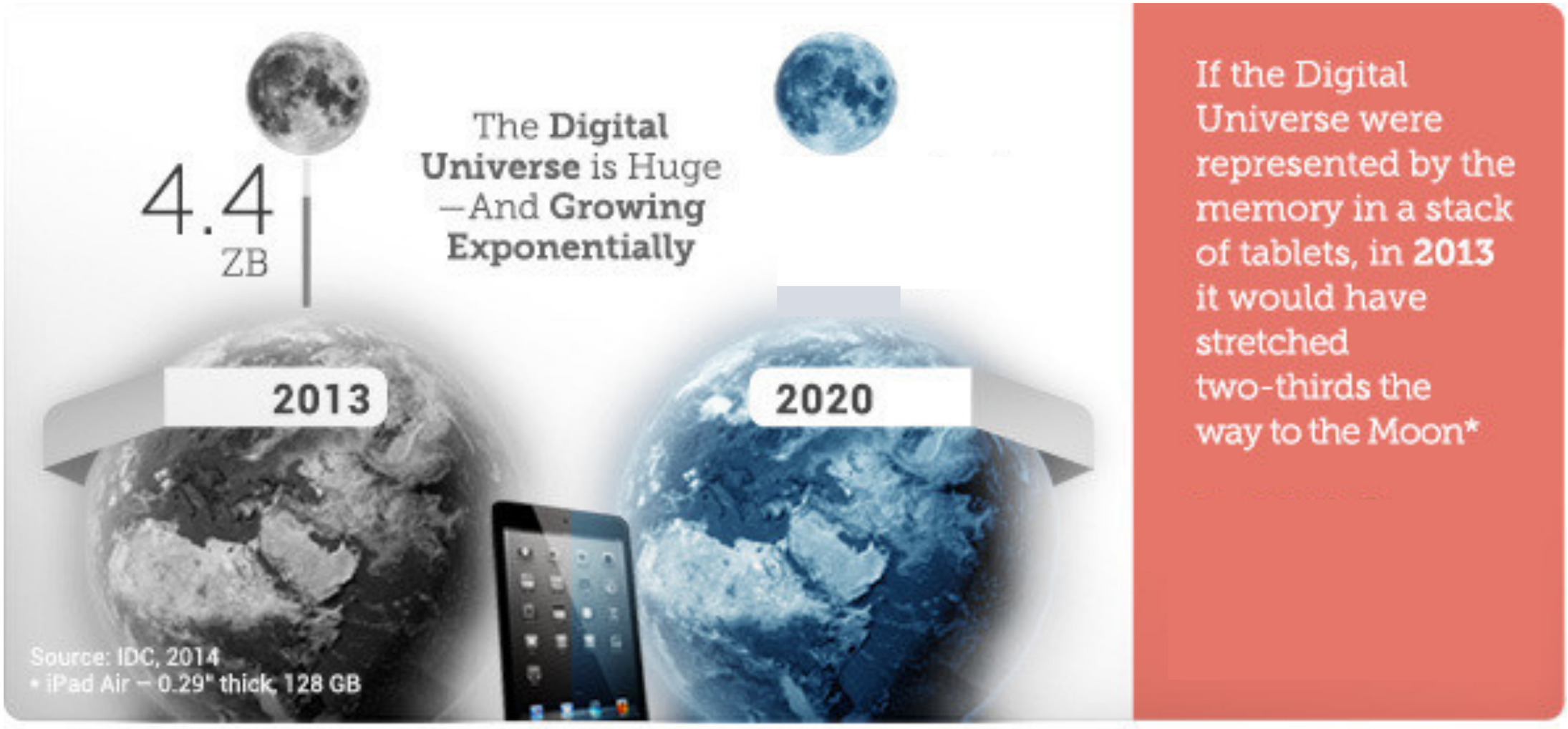
Data visualisation to better understanding respecting spreadsheets

Data heterogeneity in format and sources

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Quiz: Did we reach the moon?



Source: https://www.weforum.org/agenda/2018/01/data-is-not-the-new-oil/

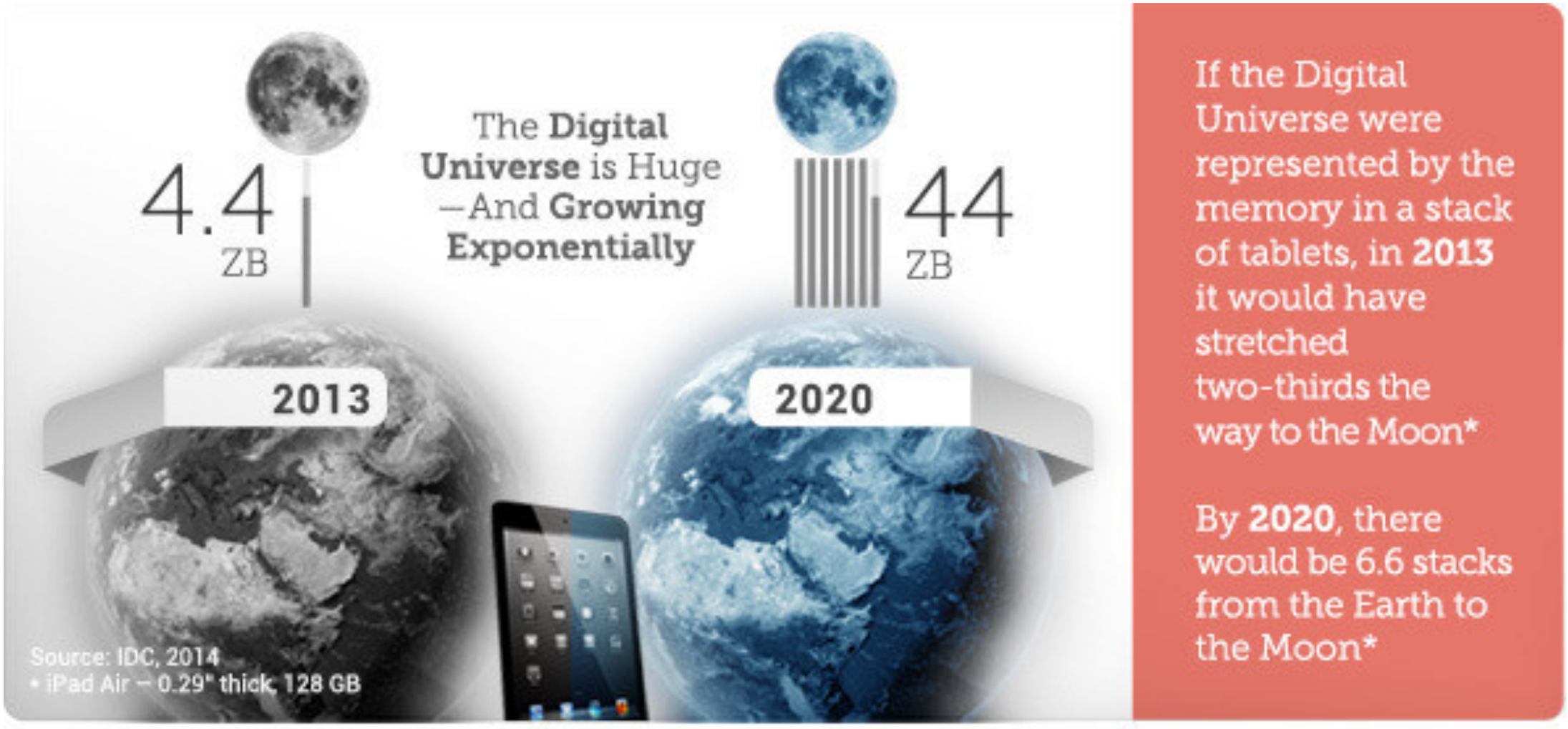
Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Quiz: Did we reach the moon?

Source: https://www.weforum.org/agenda/2018/01/data-is-not-the-new-oil/

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data



Source: https://www.weforum.org/agenda/2018/01/data-is-not-the-new-oil/

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

In 2016, IBM reportedly noted that **90% of the data in the world today had been created in the past two years alone.**

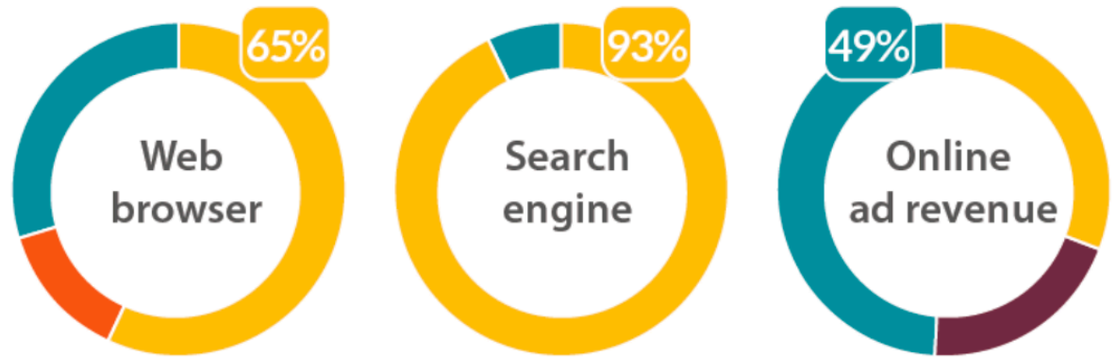
The economy is inevitably moving towards the wider production of new products and services based on availability and access to data.

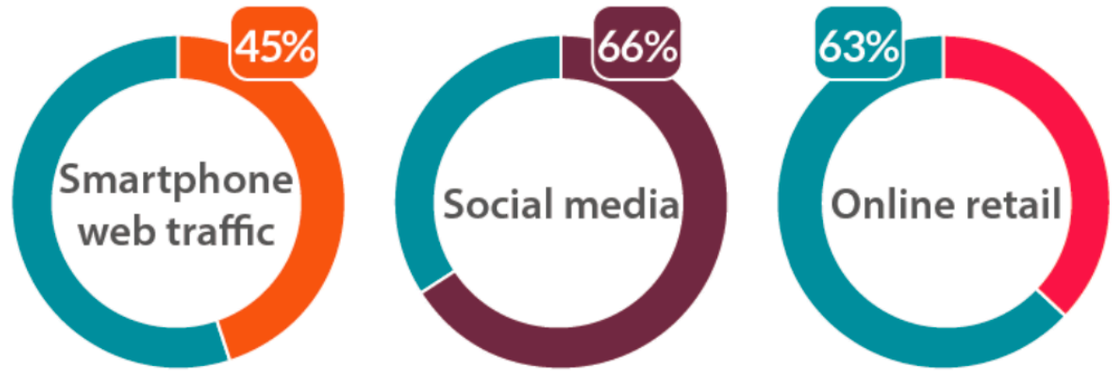
Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Can you associate the colours to the “4 star companies”?





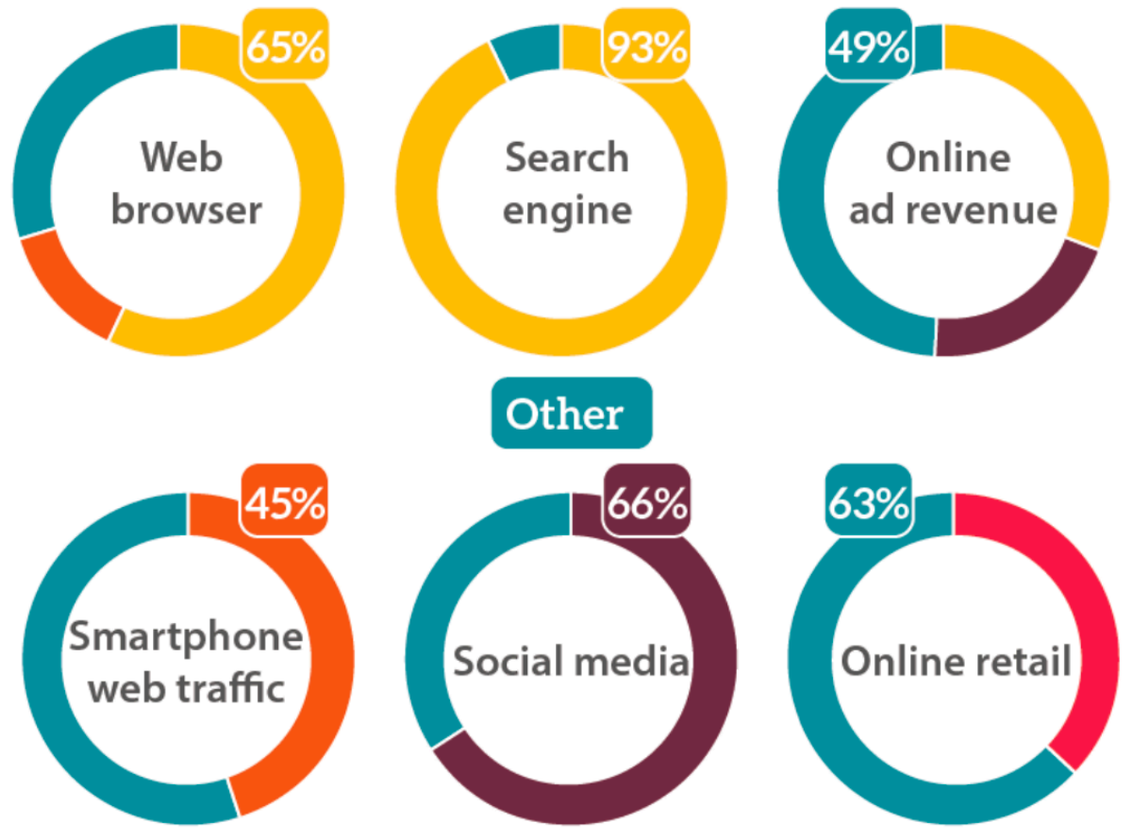
Sources: W3Counter, GSStatCounter, eMarketer.

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data

Can you associate the colours to the “4 star companies”?

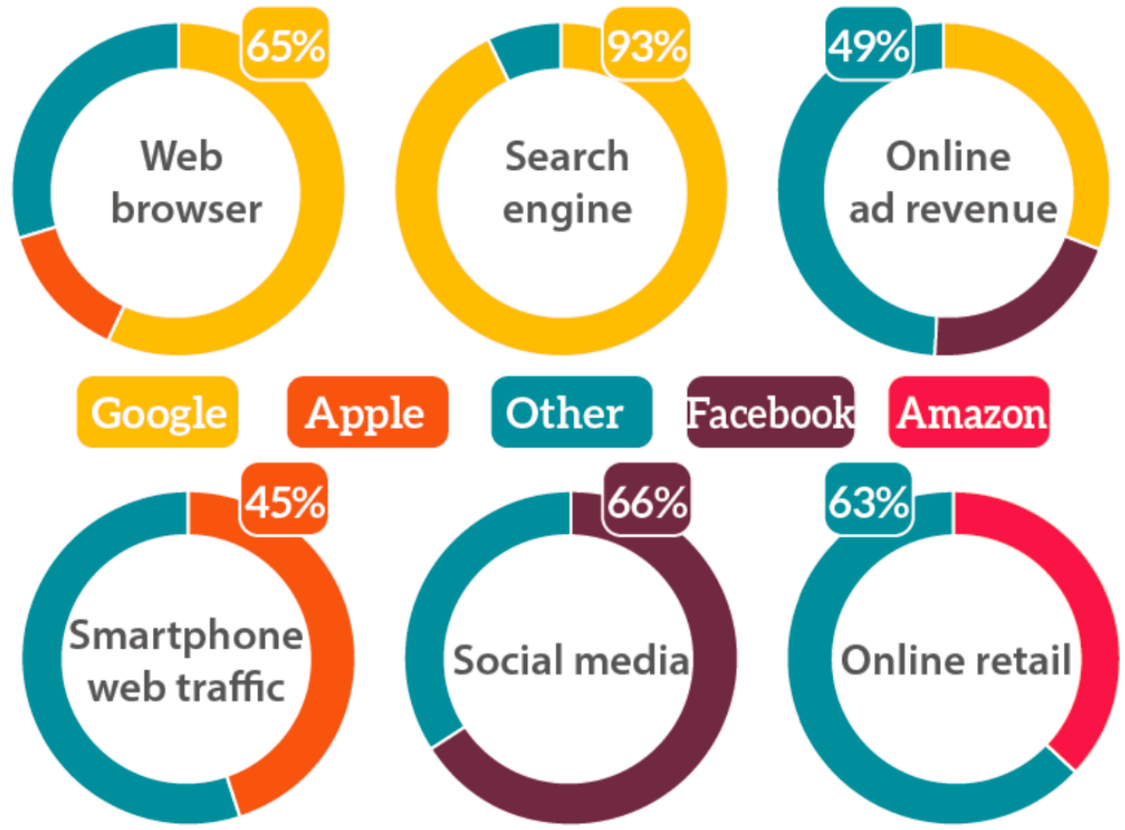


Sources: W3Counter, GSStatCounter, eMarketer.

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data & Big Data



Sources: W3Counter, GSStatCounter, eMarketer.

Source: M-Brain

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

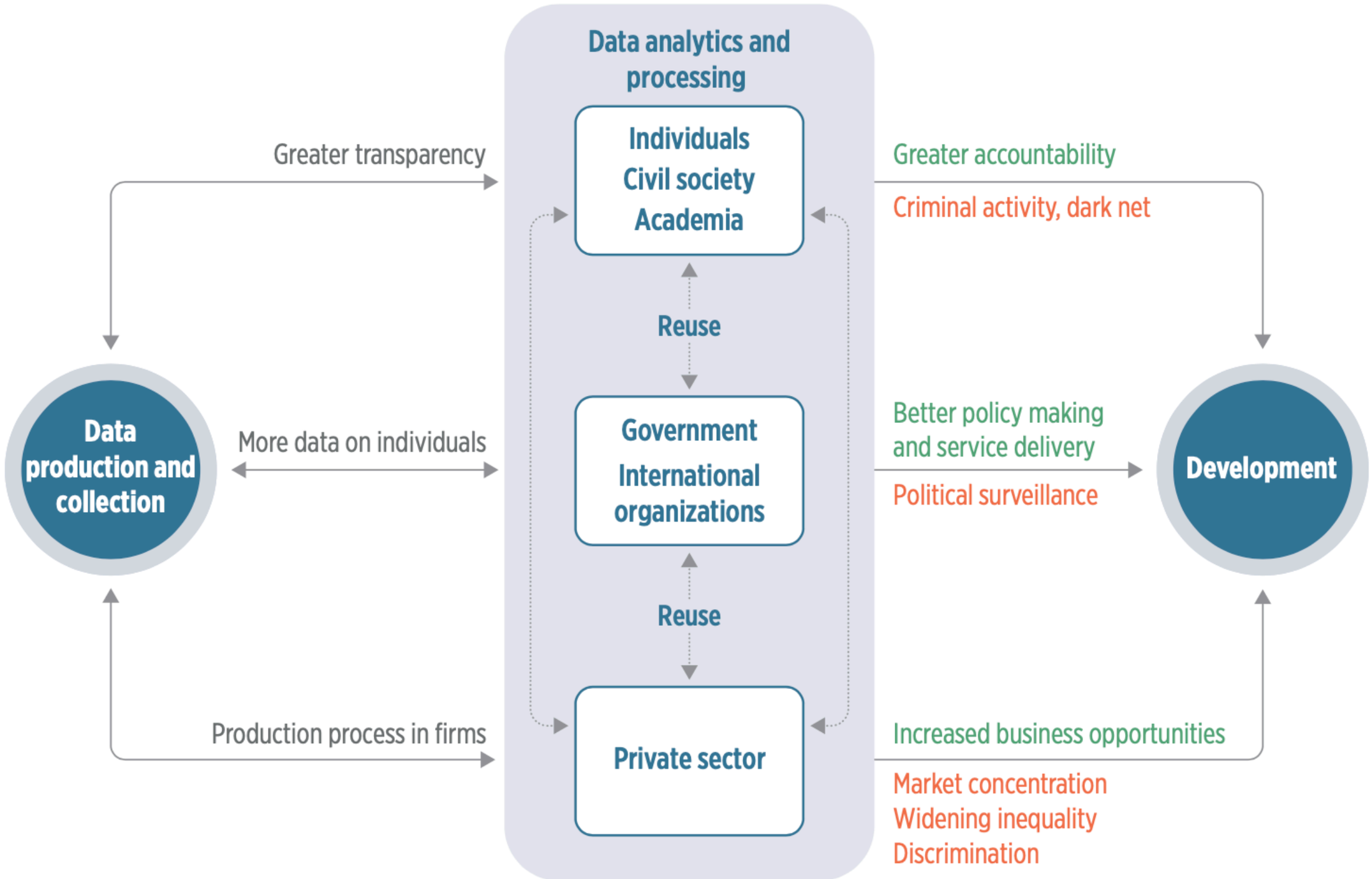
Big Data & Open Data

Source: M-Brain
Illustration: Mijke Coebergh

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Data, Big Data & Open Data

Figure O.1 How data can support development: A theory of change

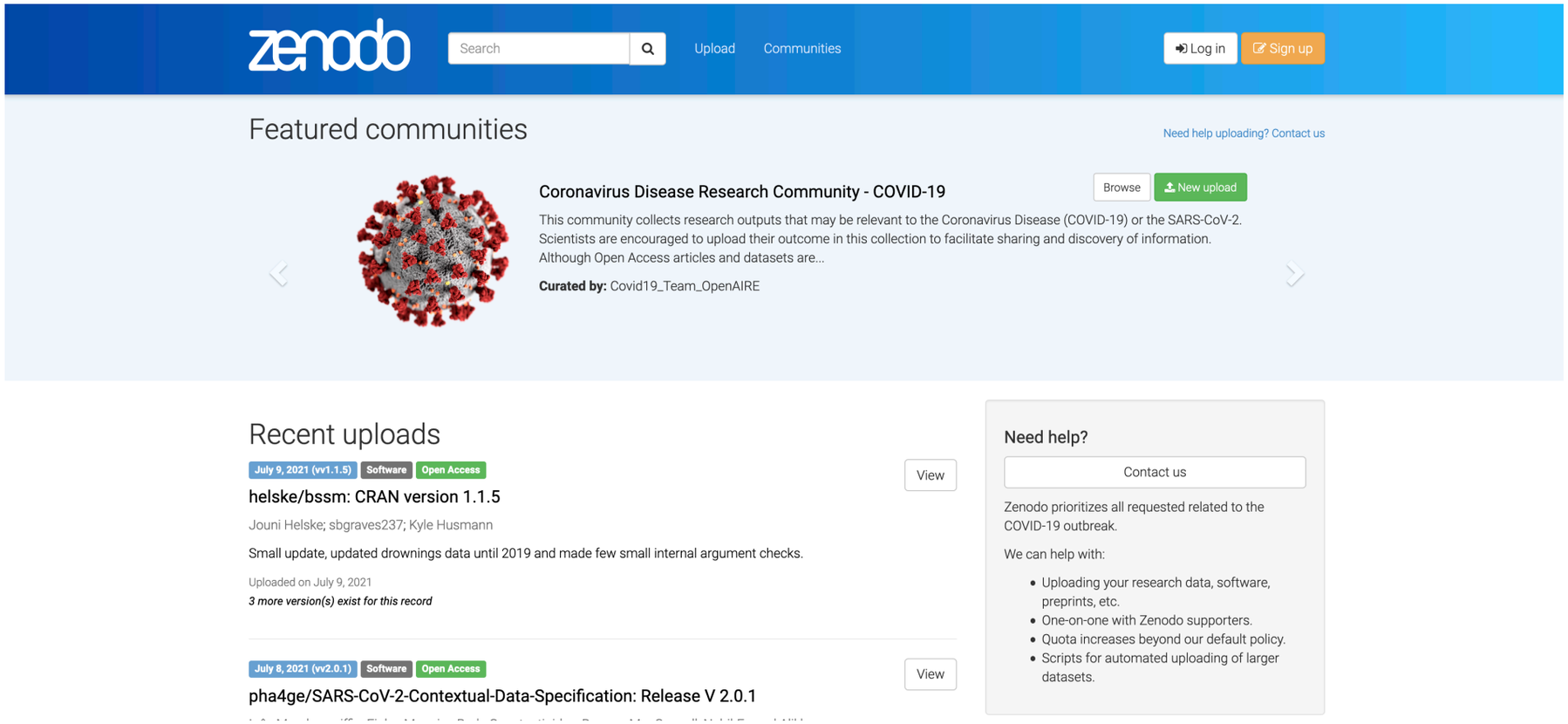


Source: WDR 2021 team.

Note: Positive impacts are shown in green; negative impacts are shown in red.

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Open Data @ CERN



zenodo.org

https://guides.github.com/activities/citable-code/

http://opendata.cern.ch/docs/welcome

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

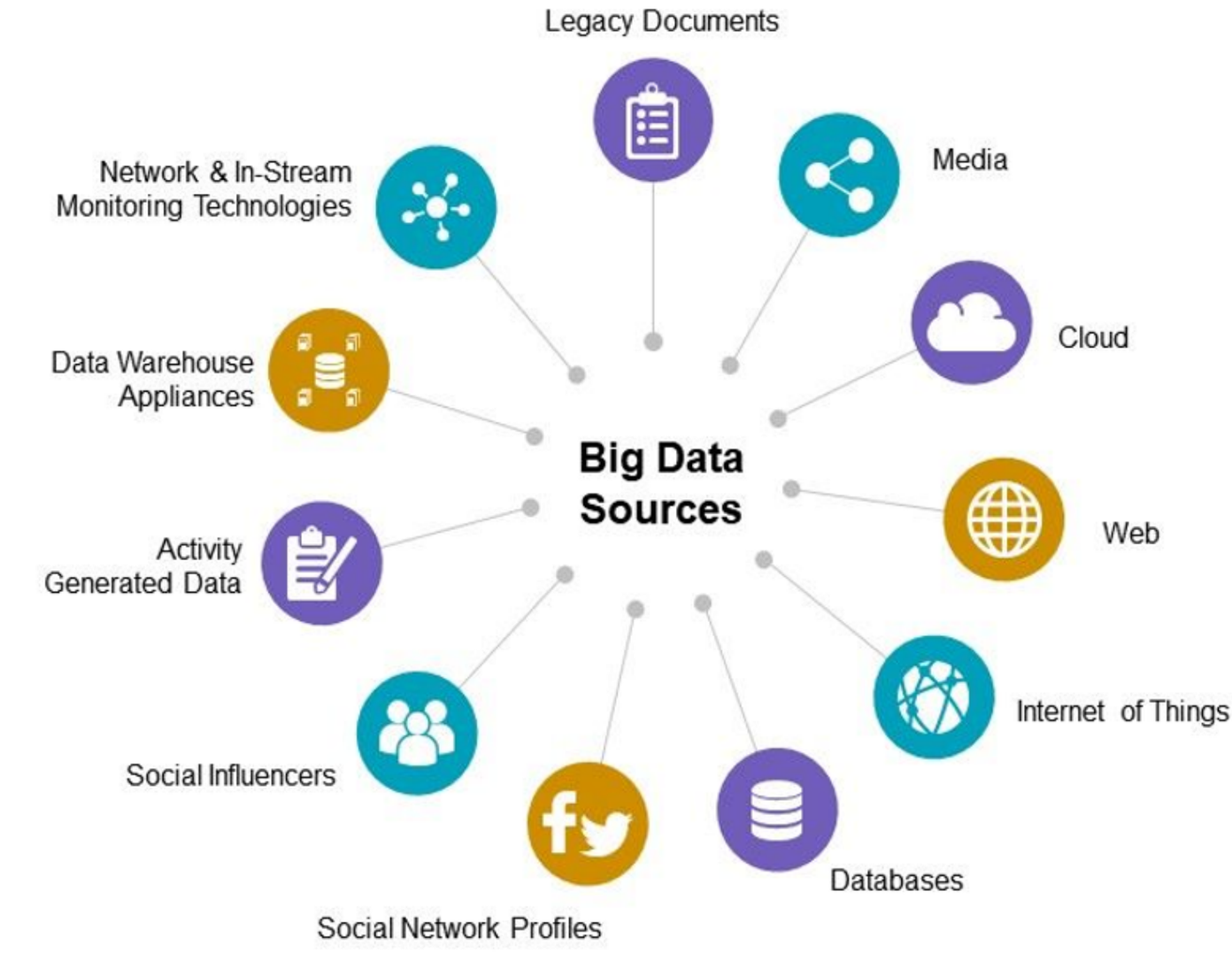
CERN Science for Open Data

Challenges

Data size is huge and of high dimensionality and Data heterogeneity

Data analysis

Data overload



Anna Ferrari CERN openlab
a.ferrari@cern.ch

CERN Science for Open Data

Challenges

Data size is huge and of high dimensionality and Data heterogeneity

Data analysis

Data overload

Anna Ferrari CERN openlab
a.ferrari@cern.ch

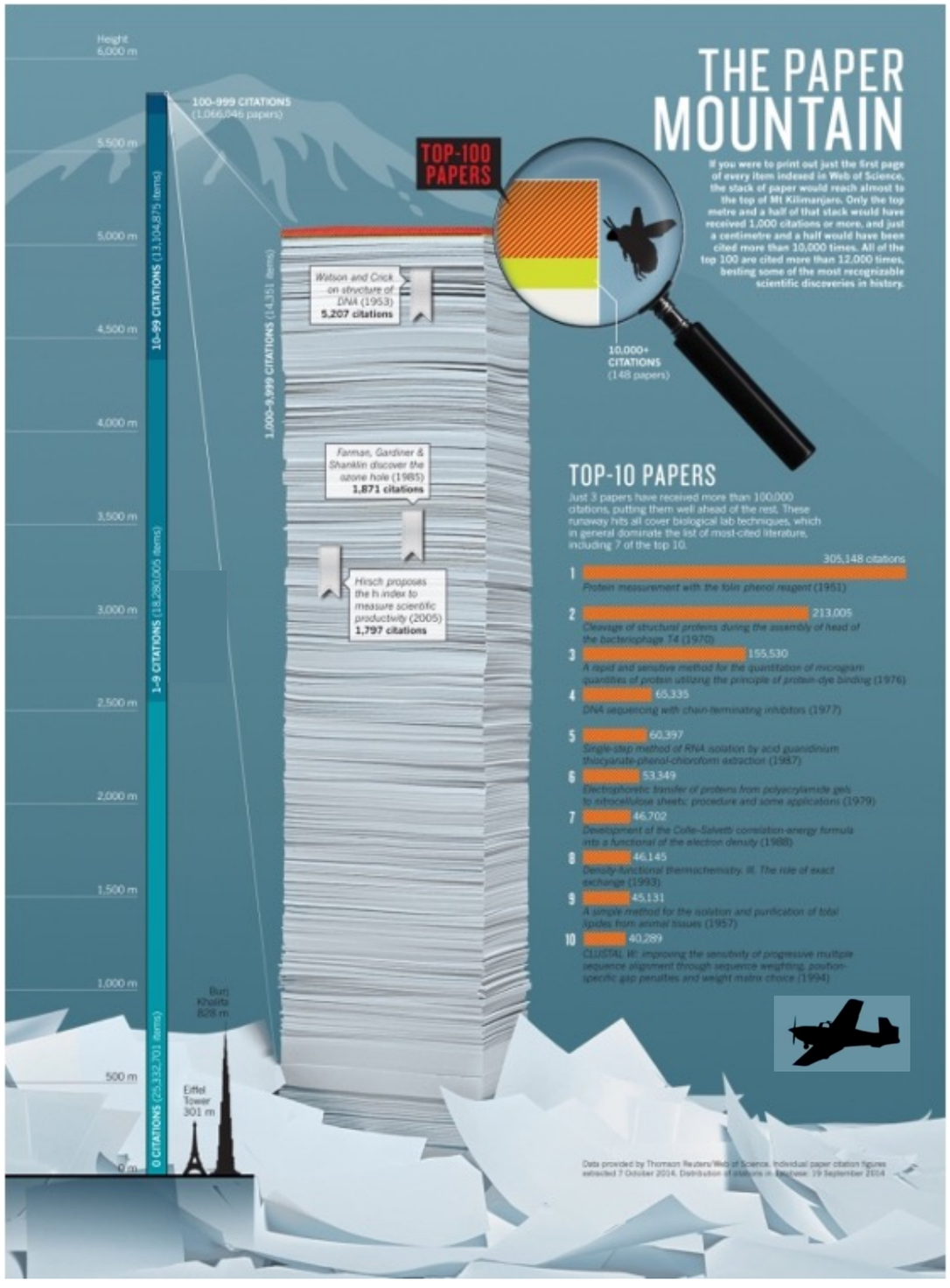
CERN Science for Open Data

Challenges

Data size is huge and of high dimensionality and Data heterogeneity

Data analysis

Data overload



Anna Ferrari CERN openlab
a.ferrari@cern.ch

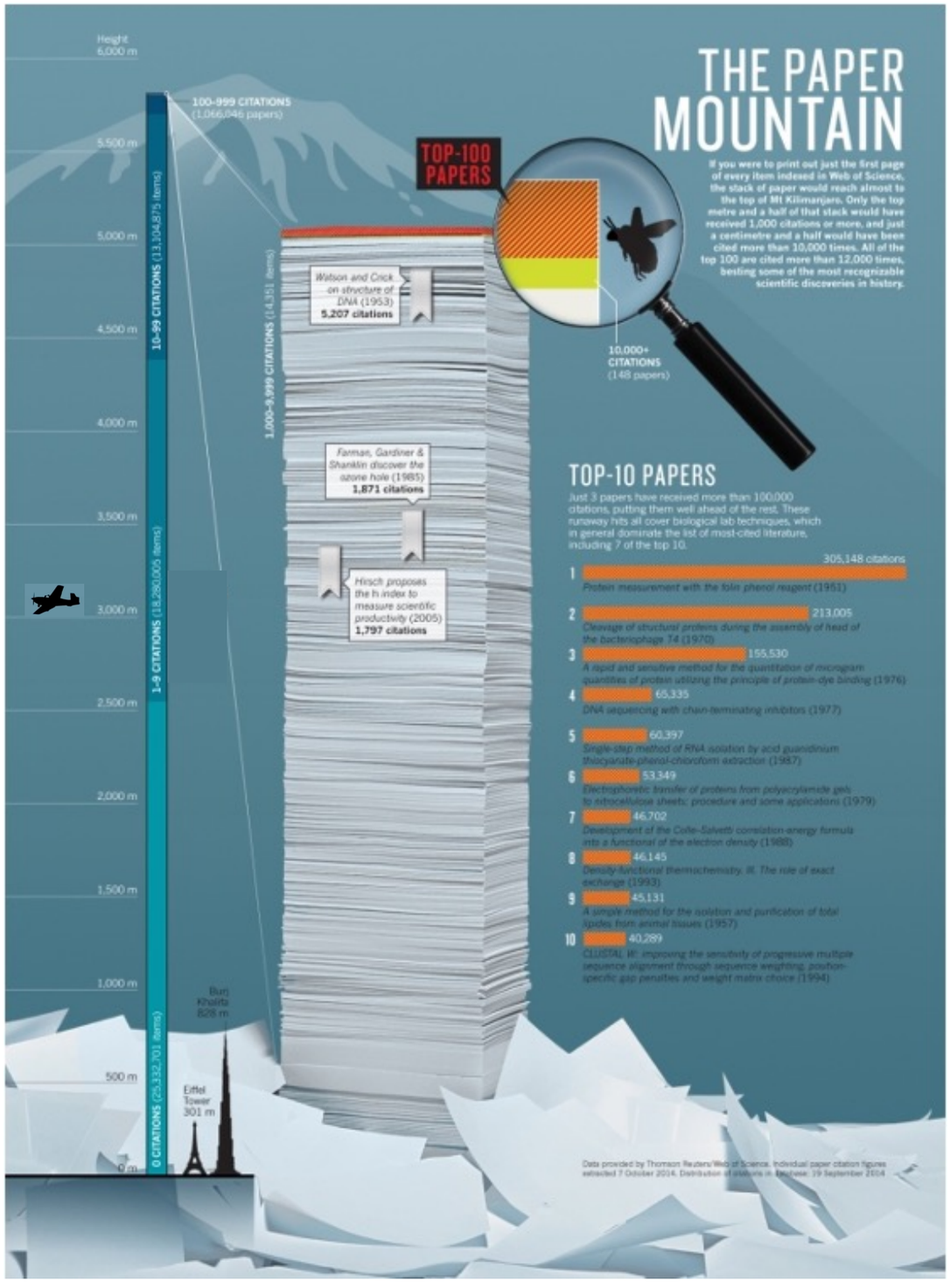
CERN Science for Open Data

Challenges

Data size is huge and of high dimensionality and Data heterogeneity

Data analysis

Data overload



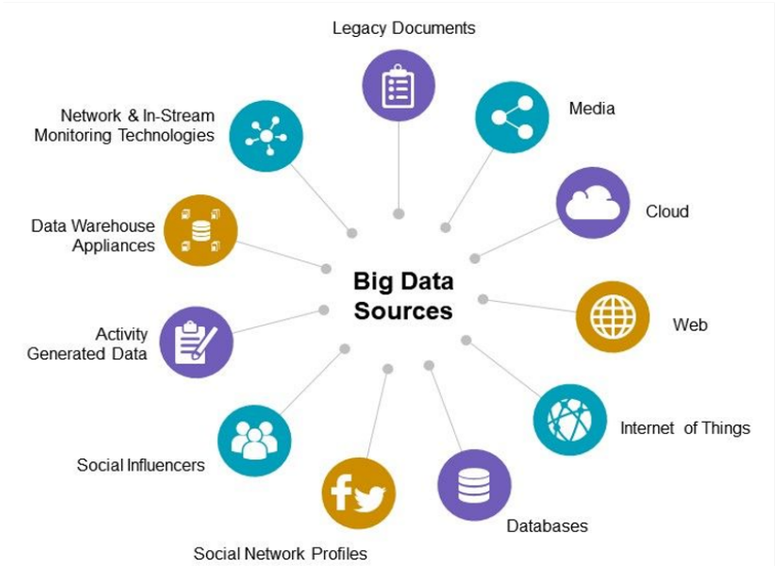
CERN openlab

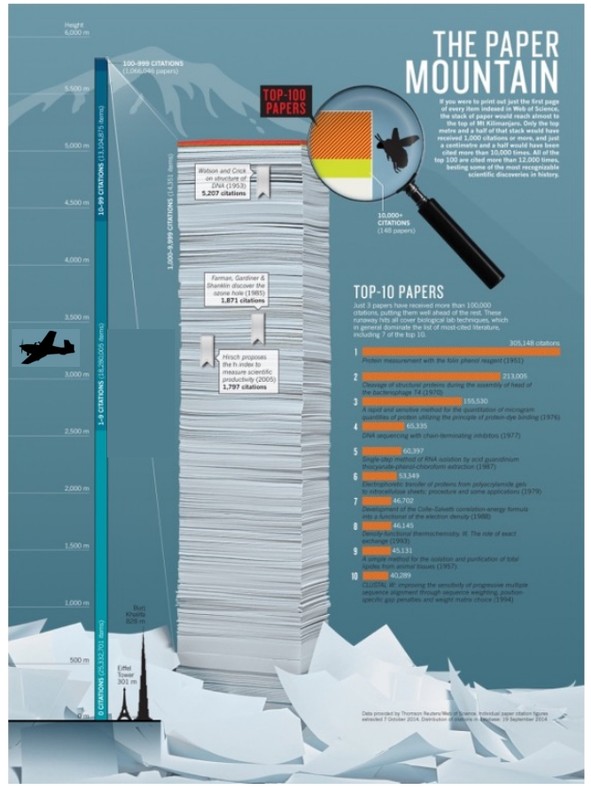
CS4OD
CERN Science for Open Data

Anna Ferrari CERN openlab
a.ferrari@cern.ch

CERN Science for Open Data

Overcome barriers





Data size: overcome barriers related to data governance and storage defining **common principles**

Data heterogeneity: overcome barriers of data access defining a **global coordination** of **open data from multi-domain fields**

Data analysis: overcome barriers of analysis diversity defining **common pipelines and approaches**

Data overload: overcome barriers of excess of information by complying with **results reproducibility and multi-disciplinary expertises exchange**

CERN openlab

CS4OD
CERN Science for Open Data

Anna Ferrari CERN openlab
a.ferrari@cern.ch

CERN technologies

Services and tools

zenodo

reana

SWAN

CERNBox

CERN openlab

CS4OD
CERN Science for Open Data

Anna Ferrari CERN openlab
a.ferrari@cern.ch

Conclusion

closeness

openness

Anna Ferrari
CERN openlab
a.ferrari@cern.ch

Anna Ferrari CERN openlab
a.ferrari@cern.ch

References and sources

- Ted Global, 2012 - *“Four principle for open the world”* - **Don Tapscott** (https://www.ted.com/talks/don_tapscott_four_principles_for_the_open_world?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare)
- Long, P. O., 1991 - *“The openness of knowledge: an ideal and its context in 16th-century writings on mining and metallurgy.*” Technology and Culture, 32(2), 318-355.
- Hootsuite & We Are Social, 2019 - *“Digital 2021 Global Digital Overview,”* retrieved from https://datareportal.com/reports/digital-2021-global-overview-report
- Benkler, Yochai, 2013 - *“Open development: Networked innovations in international development”*. Mit Press.
- By Version 1 by Nohat (concept by Paullusmagnus); Wikimedia. - File:Wikipedia-logo.svg as of 14 May 2010T23:16:42, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=10337301
- https://stats.wikimedia.org/#/all-projects
- https://openknowledge.worldbank.org/handle/10986/35218
- World Bank, 2021 - “*World Development Report 2021: Data for Better Lives.*” Overview booklet. World Bank, Washington, DC. License: Creative Commons Attribution CC BY 3.0 IGO.
- https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2020)646117
- Gurin, J., Bonina, C., & Verhulst, S., 2019 - *“Open data stakeholders: Private sector.”* In T. Davies, S. Walker, M. Rubinstein, & F. Perini (Eds.), *The state of open data: Histories and horizons* (pp. 418–429). Cape Town and Ottawa: African Minds and International Development Research Centre. http://stateofopendata.od4d.net

Image slide 31 is an adaptation of an original work by The World Bank. Views and opinions expressed in the adaptation are the sole responsibility of the author or authors of the adaptation and are not endorsed by The World Bank.