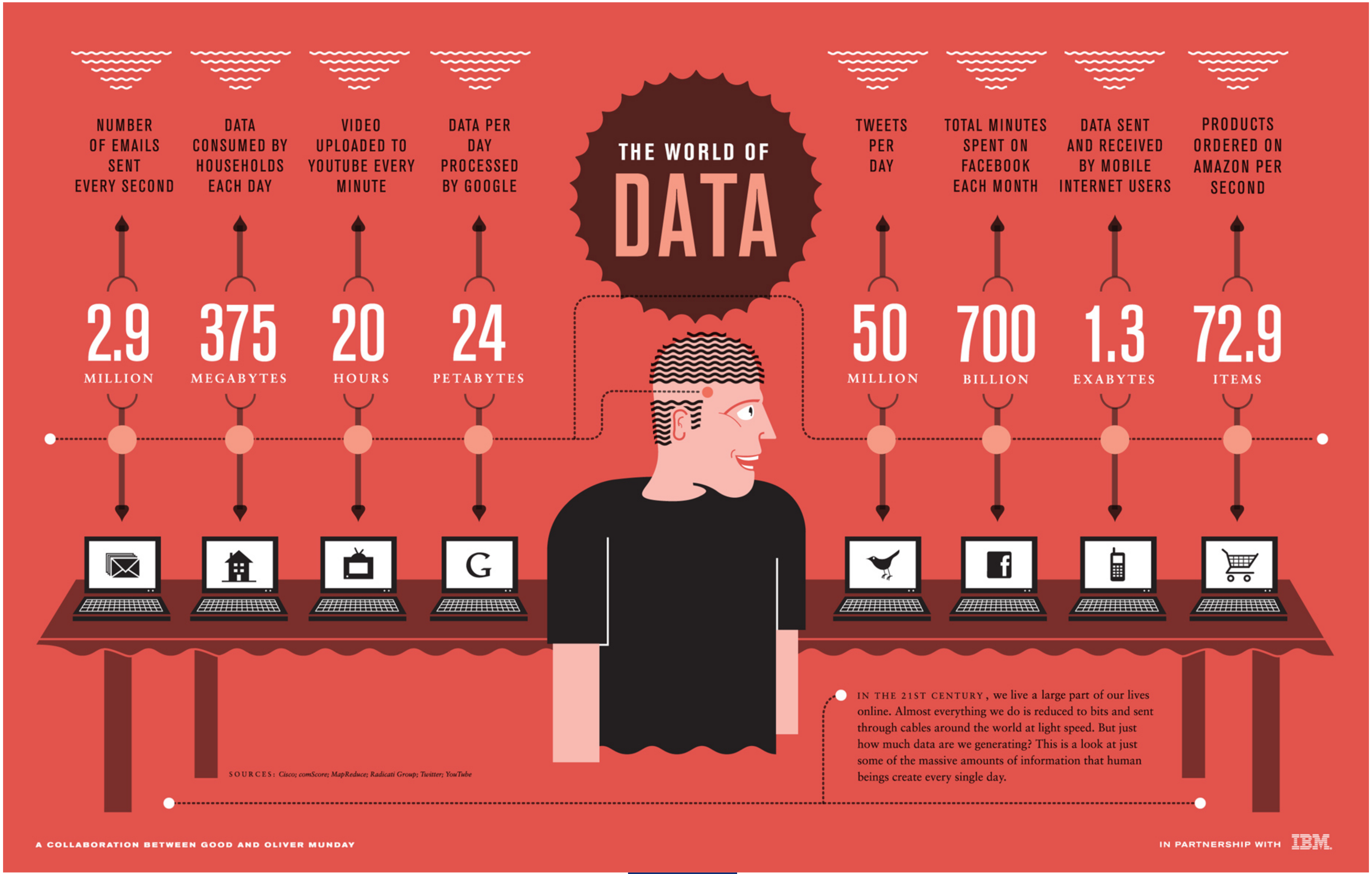
CSPA

Common Statistical Production Architecture

International activities on Big Data in Official Statistics

Carlo Vaccari

Istat (vaccari@istat.it)



THE WORLD OF DATA

NUMBER OF EMAILS SENT EVERY SECOND: 2.9 MILLION

DATA CONSUMED BY HOUSEHOLDS EACH DAY: 375 MEGABYTES

VIDEO UPLOADED TO YOUTUBE EVERY MINUTE: 20 HOURS

DATA PER DAY PROCESSED BY GOOGLE: 24 PETABYTES

TWEETS PER DAY: 50 MILLION

TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH: 700 BILLION

DATA SENT AND RECEIVED BY MOBILE INTERNET USERS: 1.3 EXABYTES

PRODUCTS ORDERED ON AMAZON PER SECOND: 72.9 ITEMS

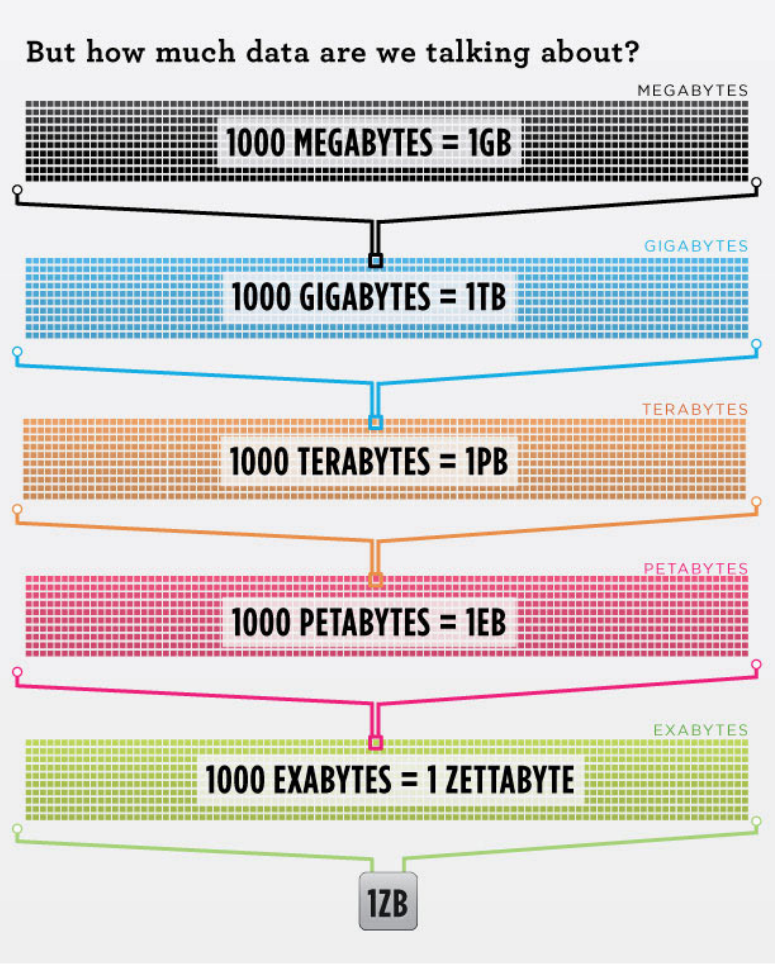
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

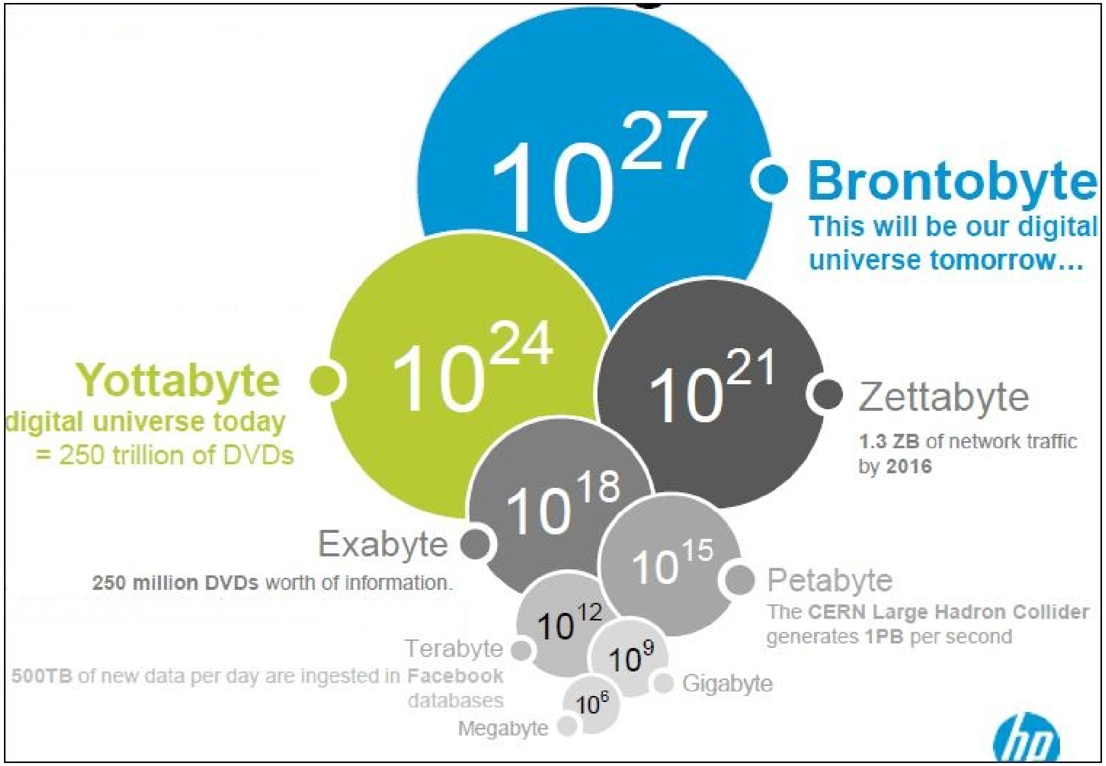
SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH IBM

Data deluge


But how much data are we talking about?
MEGABYTES
1000 MEGABYTES = 1GB
GIGABYTES
1000 GIGABYTES = 1TB
TERABYTES
1000 TERABYTES = 1PB
PETABYTES
1000 PETABYTES = 1EB
EXABYTES
1000 EXABYTES = 1 ZETTABYTE
1ZB


10^{27}
Brontobyte
This will be our digital universe tomorrow…
10^{24}
Yottabyte
digital universe today = 250 trillion of DVDs
10^{21}
Zettabyte
1.3 ZB of network traffic by 2016
10^{18}
Exabyte
250 million DVDs worth of information.
10^{15}
Petabyte
The CERN Large Hadron Collider generates 1PB per second
10^{12}
Terabyte
500TB of new data per day are ingested in Facebook databases
10^{9}
Gigabyte
10^{6}
Megabyte
hp

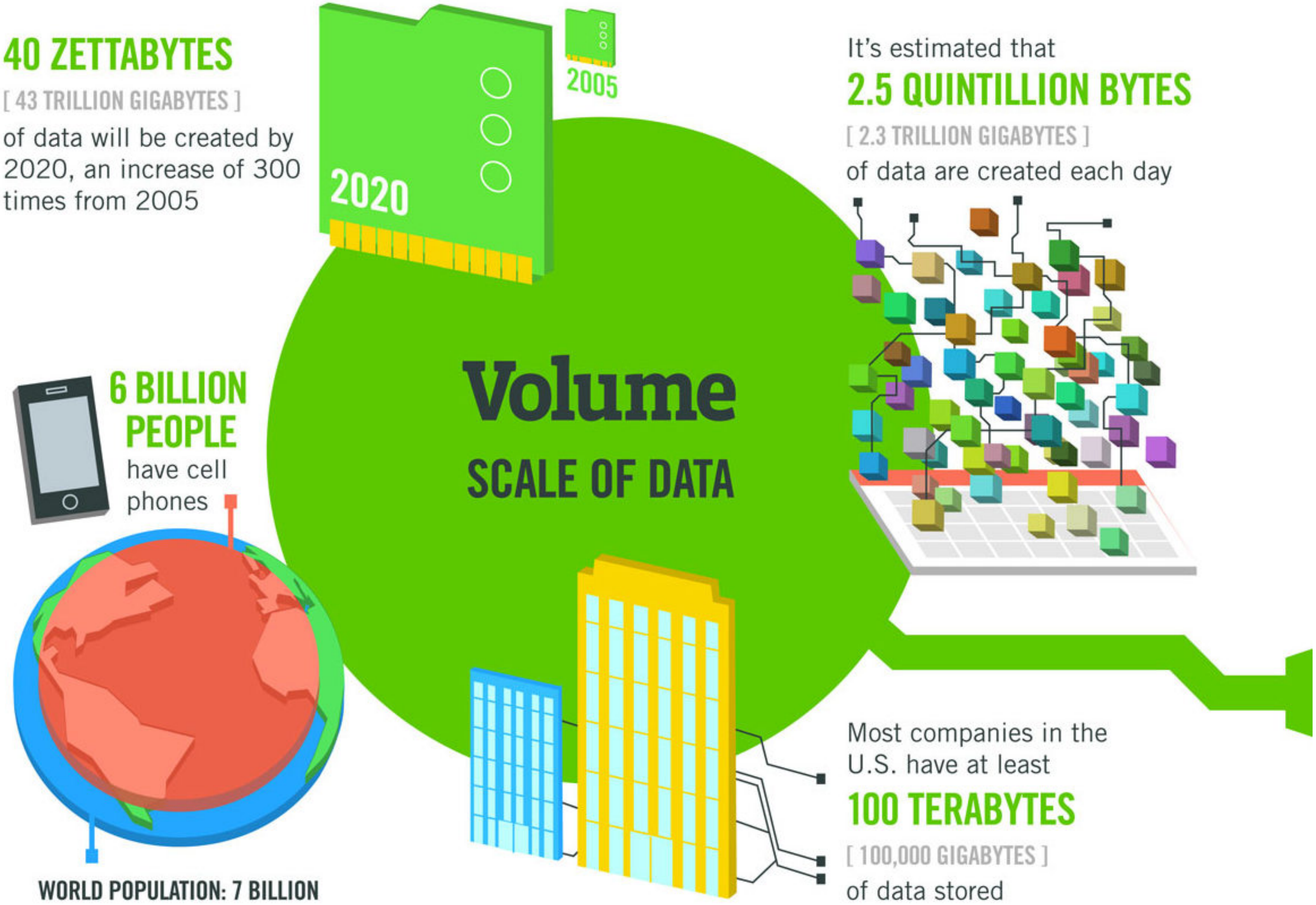
Big Data definitions

Data Characteristics:
3 Vs: Volume Velocity Variety
(but also Veracity Value Viability)

Data sources: many taxonomies presented

- Human sources information
- Traditional Business Systems
- Internet of Things

European Commission



Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

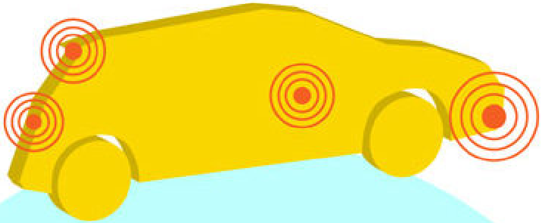
[100,000 GIGABYTES]

of data stored

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

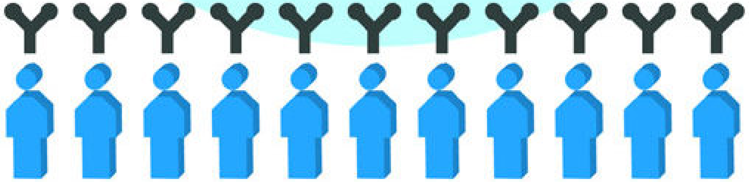
Velocity

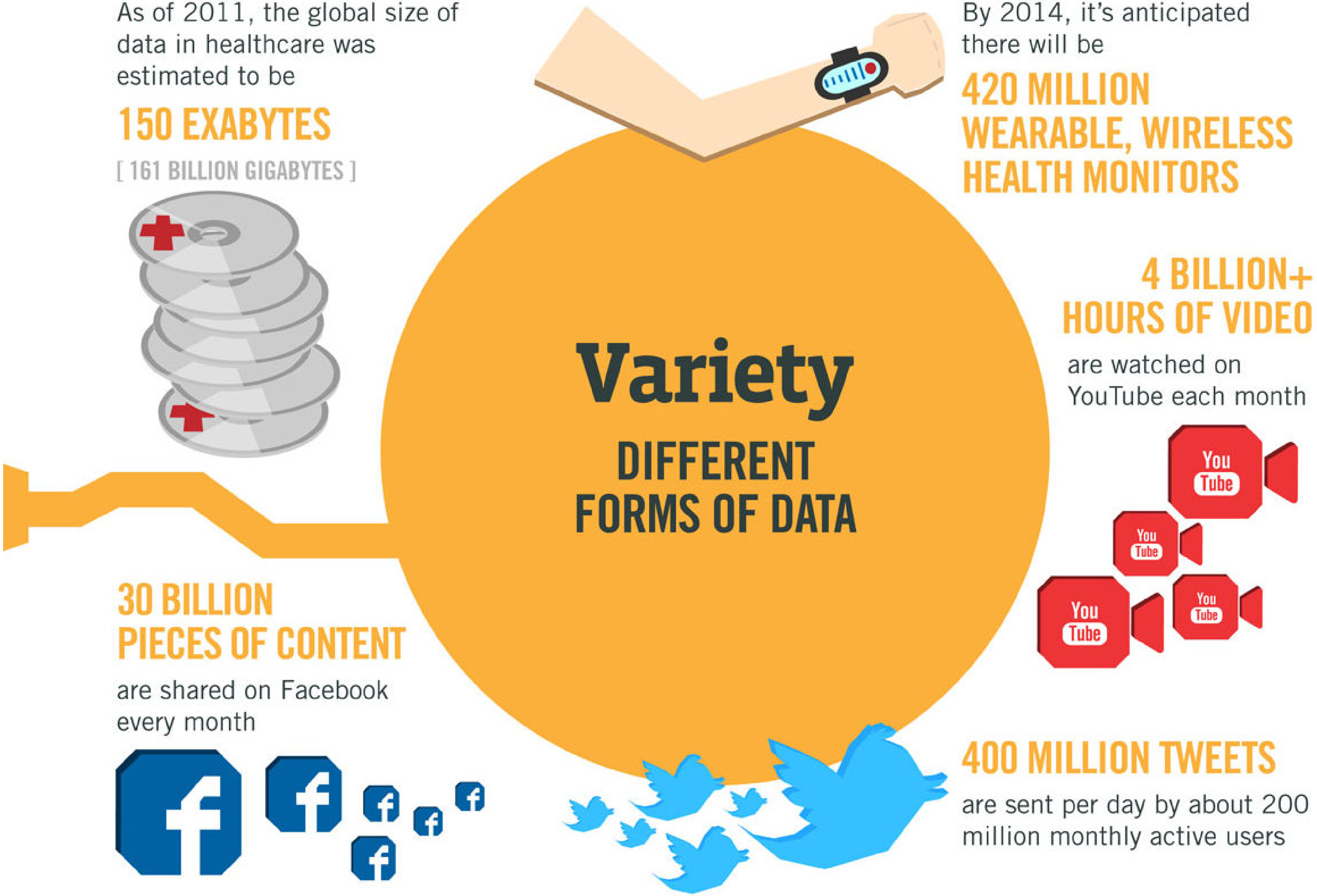
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth





Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Big data challenges for NSIs

1. Legislative: are Big Data accessible to Statistical Organisations and at what conditions?

2. Privacy: in accessing and processing Big Data, what assurances exist on the protection of the confidentiality?

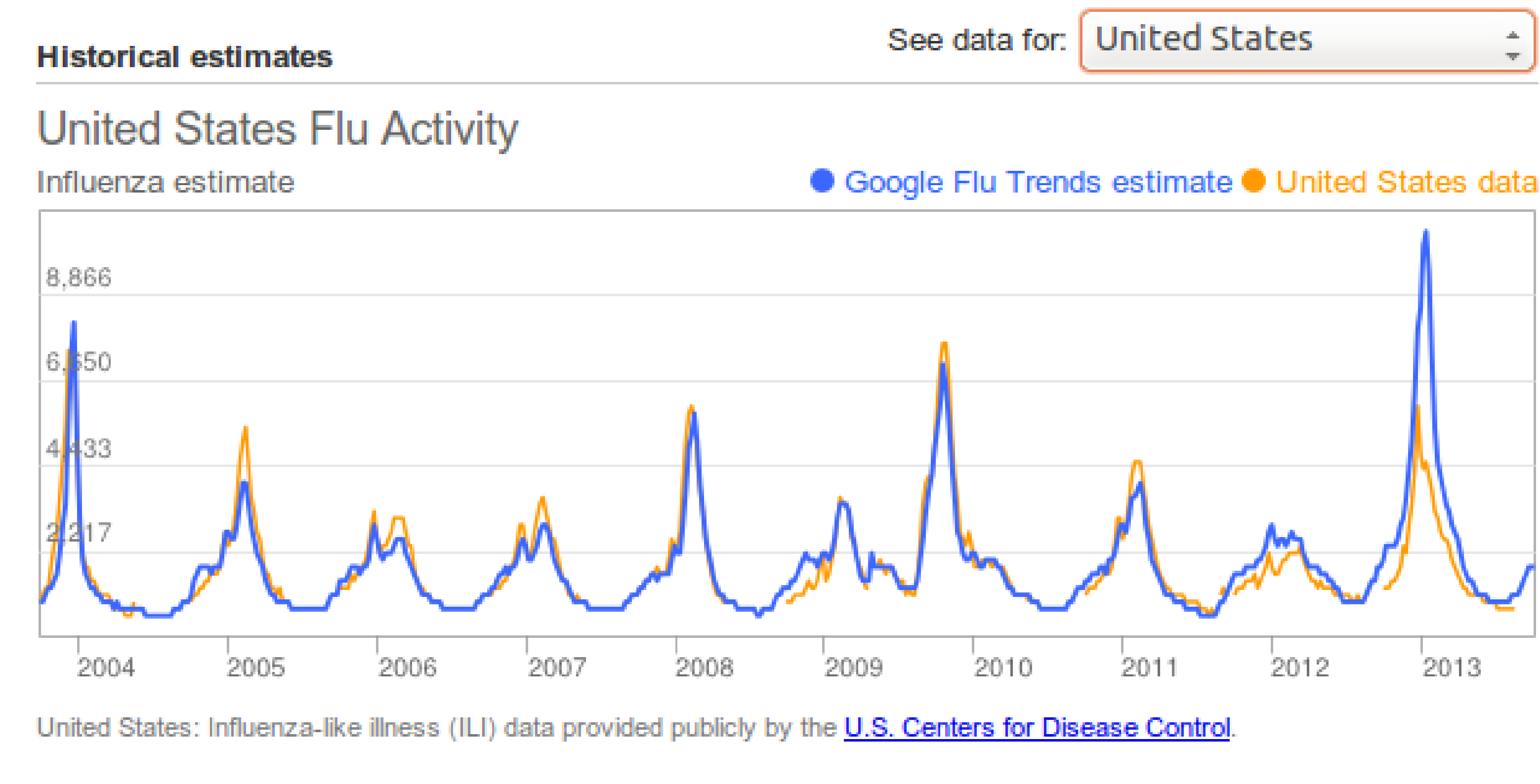
3. Financial: access to Big Data often has a cost, maybe lower then statistical data, but sometimes considerable

4. Management: what is the impact on the organization of a NSI when Big Data become an important source of data?

5. Technological: what paradigm shift is required in Information Technology in order to start using Big Data?

6. Methodological: what is the impact of the use of Big Data (in combination or in substitution of statistical data) on methods of data collection, processing and dissemination?

Big data first experiences



Historical estimates

See data for: United States

United States Flu Activity

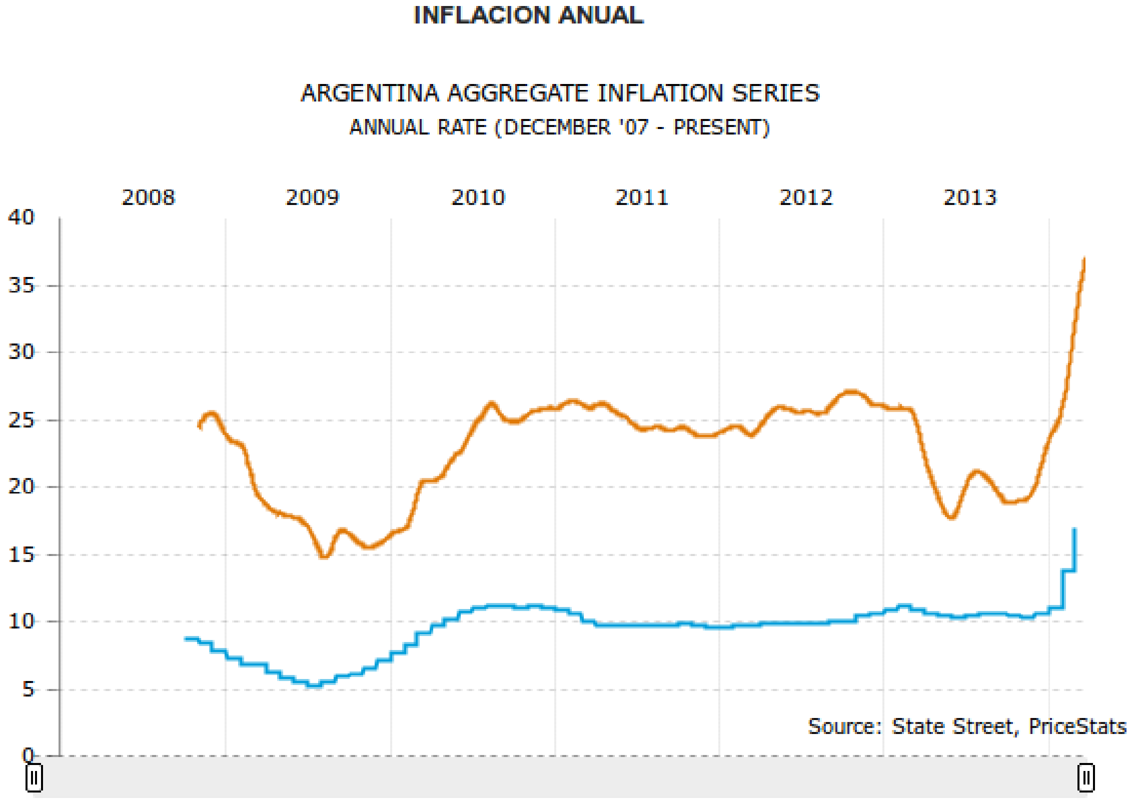
Influenza estimate

● Google Flu Trends estimate ● United States data

United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.

Data from Google Search anticipate data from CDC

Big data first experiences in statistics



For the first time Big Data denies Official Statistics

Big data first experiences in NSIs

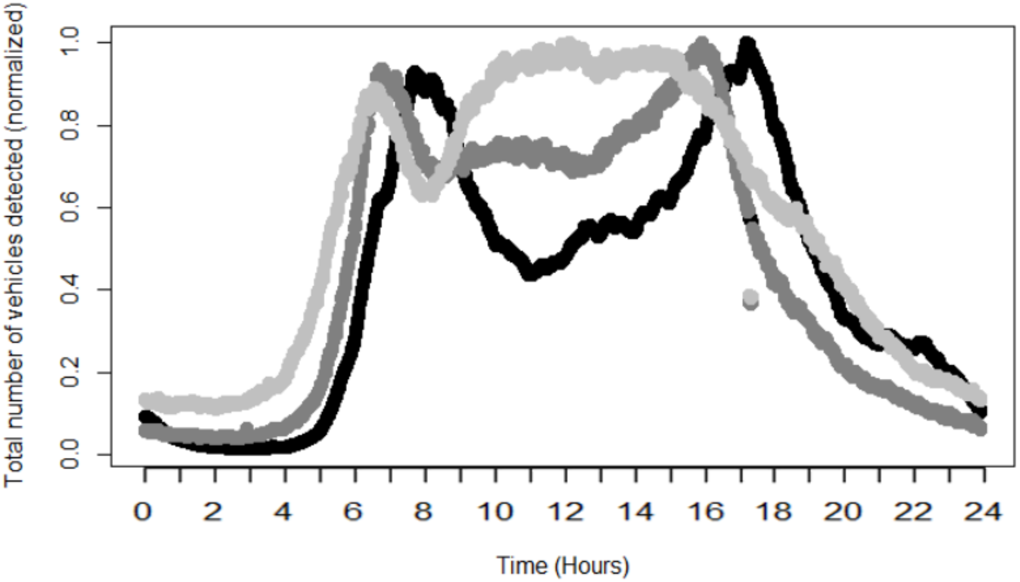


Figure 2. Normalized number of vehicles detected in three length categories on December 1st, 2011 after correcting for missing data. Small (<= 5.6 meter), medium-sized (>5.6 and <= 12.2 meter) and large vehicles (> 12.2 meter) are shown in black, dark grey and grey, respectively. Profiles are normalized to clearly reveal the differences in driving behaviour.

Statistical Organizations start using Big Data for old and new statistics

HLG Vision

The product challenge

Changeover from a society with little or no data available to one that has an abundance of data.

In this light we have to rethink our traditional business values and the reasons of our existence.

Now other organisations developing statistics similar to ours but much, much quicker (e.g. Google) and on an almost global scale.

We also see other very interesting uses of statistics, prompted by the availability of so much data.

HLG Vision

Vision on products:

Data are everywhere and are much cheaper than they used to be. Slowly people are beginning to understand the real value of this fact.

Social networking sites and search engines are now perceived as data collection devices. NSIs are in a unique position to connect to the data of the emerging information society and transform them into something useful.

As the global dimension of events gains importance we can no longer work on a national level only. We need to expand our work and deliver products that explain what is happening on a multinational level.

HLG Vision

Vision on products:

In the chapter dedicated to “Rationalising processes” there is a proposal

“(d) Develop new methodologies to reflect the changes in data acquisition and the dramatic increase of the volume of data available, for example, on topics such as noise and error reduction in large data sets, pattern recognition and other methodological tools appropriate for "Big Data"”.

in March 2013 the HLG published the paper “What does Big Data mean for Official Statistics”

MSIS 2013

In Meeting on Statistical Information Systems (MSIS 2013) many points of discussion focused on Big Data. In the summary of the session on Collaboration, the Session Organizer asked “How can we collaborate on Big Data? Many initiatives and groups are involved in collaboration. How can the MSIS group contribute most effectively in this new structure?”

Also a special panel session was organized with the title: “Plug and play architecture and collaborative development will allow us to accelerate our Big Data programs by ...” The panel discussion followed a Pecha Kucha format: 20 slides for 20 seconds each.

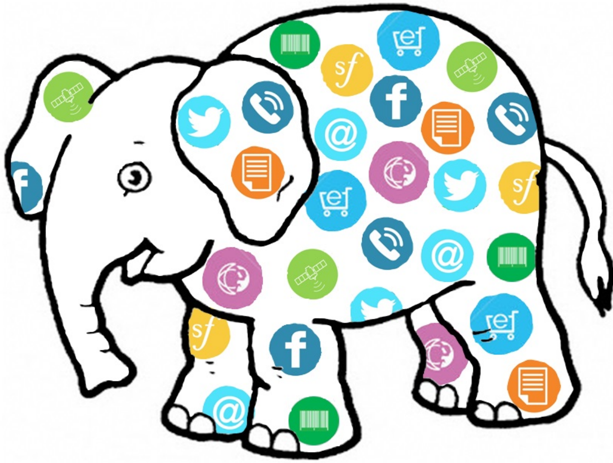
Big Data Task Team

In May 2013 a temporary task team was set up, composed by members coming from thirteen organizations. The task team, working virtually through teleconferences and sharing documents on the wiki, started to define the key issues with using Big Data for official statistics, identify priority actions and formulate a project proposal.

Two preliminary activities:

- a classification scheme for Big Data sources
- the development of an inventory containing structured and searchable information about actual and planned use of Big Data in statistical organizations.

Big Data Project



Project presented to HLG and CES

Task team composed by people from 13 organisations

The project consists of four Work-packages:

- WP1 - Issues and Methodology
- WP2 - Shared computing environment ('sandbox')
- WP3 - Training and Dissemination
- WP4 - Project Management

Big Data Project: the sandbox

Sandbox will evaluate the feasibility of the following propositions:

- 'Big Data' sources can be obtained (in a stable and replicable way), and manipulated with relative ease and efficiency on the chosen platform, within realistic technological and financial constraints;
- The chosen sources can be processed to produce statistics which conform to the usual quality criteria used to assess official statistics;
- The resulting statistics correspond in a predictable way with existing mainstream products, such as price statistics, HBS indicators, etc.;
- Platforms, tools, methods and datasets can be used in similar ways to produce analogous statistics in different countries;
- The different participating countries can share tools, methods, datasets and results operating on the principles established in CSPA.

Big Data Project: the sandbox

Sandbox: web-accessible environment where researchers coming from different institutions will be able to explore tools and methods needed for statistical production and the feasibility of producing Big Data-derived statistics.

List of tools chosen:

Hadoop, Hortonworks,

Pentaho suite, R, ...

Big Data Project: the sandbox

The sandbox will be hosted at the Irish Center for High-End Computing (ICHEC) which has the mission to provide High-Performance Computing resources, support, education and training for researchers in third-level institutions.

ICHEC will assist the task team to implement the Big Data environment for the testing and evaluation of Hadoop work-flows and associated data analysis application software Pentaho and R

High Performance Computing Linux cluster composed of 60 compute nodes each of which has two quad-core processors, 48GB of RAM and a 1TB local disk. A 20TB shared filesystem is available to all nodes.

Big Data Project: recent steps

Virtual Sprint (March 2014) → new document

Workshop in Roma (April 2014)

Sandbox installation and verification

Testing scenarios for BD usage in Official Statistics:

- use as auxiliary information to improve an existing survey
- replacing all or part of an existing survey with Big Data
- producing a predefined statistical output either with or without supplementation of survey data
- producing a statistical output guided by findings from the data