Curso Big Data

Taller de Empleo Camargo X

Biuse Casaponsa
deducedatasolutions.com

“Big data” hace referencia a conjuntos de datos tan grandes y complejos como para que hagan falta aplicaciones informáticas no tradicionales de procesamiento de datos para tratarlos adecuadamente. *Wikipedia*

"DATA IS THE NEW OIL."

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered **data** to be an economic **asset**, like oil.

From the beginning of recorded time until 2003, we created

5 exabytes (5 billion gigabytes) **of data.**

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

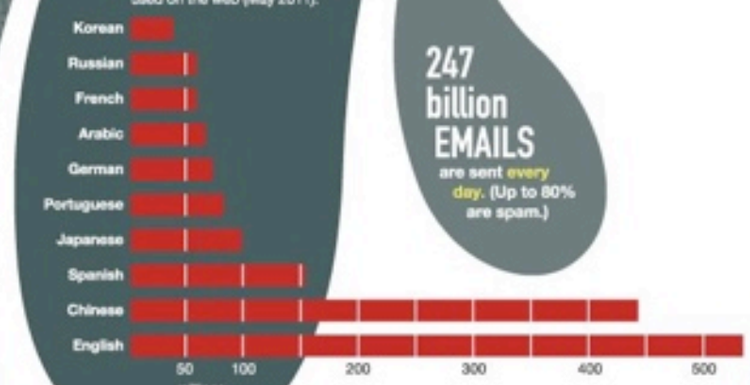
There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

There are 133 million BLOGS on the web.

As of August 2012, there were just over 4 million articles in the English Wikipedia.

English is the dominant language of the web. But by 2014 it will be Chinese, if its current rate of increase continues.

Top languages used on the web (May 2011):



80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

247 billion EMAILS are sent every day. (Up to 80% are spam.)

Every hour, we create enough Internet traffic to fill 7 billion DVDs. Side by side, that's that's seven times the height of Everest.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, high-frequency traders, with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a comm
New cable being laid under the Atlantic will sl

5 milliseconds from the current 65
it takes for trac
to travel b
New

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 millisecon

The depth of the Atlantic Ocean va

The new cable will lie on areas of floor that are up to 1,000 feet sh than the current fastest cable. a different route, the new cab shorter, meaning that the tim takes for messages to tra along it is shortened.

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

10% of all photos ever taken were taken in 2011.

50% of 5-year-old kids in the U.S. are given access to a smartphone.

"DATA IS THE NEW OIL."

Coined in 2006 by Clive Humby, a British data commerce entrepren this now famous phrase wa embraced by the World Economic Forum in a 2011 report, which considered **data** to be an economic **asset**, like oil.

From the beginning of recorded time until 2003, we created **5 exabytes** (5 billion gigabytes) **of data.**

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

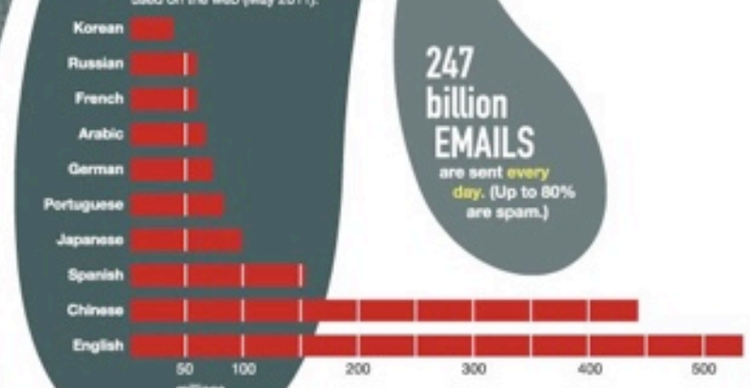
1 exabyte =
1000 millones de GB

Every hour, we create enough Internet traffic to fill **7 billion DVDs.**

Side by side, that's that's seven times the height of Everest.

English is the dominant language of the web. But by 2014 it will be **Chinese,** if its current rate of increase continues.

Top languages used on the web (May 2011):



247 billion EMAILS are sent every day. (Up to 80% are spam.)

80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

10% of all photos ever taken were taken in 2011.

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders,** with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a comm
New cable being laid under the Atlantic will sl

5 milliseconds
from the current 65
it takes for trac
to travel b
New

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 millisecon

The depth of the Atlantic Ocean va

The new cable will lie on areas of floor that are up to 1,000 feet sh than the current fastest cable. a different route, the new cab shorter, meaning that the tim takes for messages to tra along it is shortened.

50% of 5-year-old kids in the U.S. are given access to a smartphone.

1 exabyte =
1000 millones de
GB

1000 millones de
GB = 2 millones
de ordenadores

"DATA IS THE NEW OIL."

From the beginning of recorded time until 2003, we created
5 exabytes (5 billion gigabytes) of data.

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Coined in 2006 by Clive Humby, a British data commerce entrepren this now famous phrase wa embraced by the World Economic Forum in a 2011 report, which considere data to be an economi asset, like oil.

Every hour, we create enough Internet traffic to fill
7 billion DVDs.
Side by side, that's that's seven times the height of Everest.

80%

Japanese
Spanish
Chinese
English
50 100 200 300 400 500
millions of users

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan,
high-frequency traders,
with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms
make split-second decisions to buy or sell a comm
New cable being laid under the Atlantic will sl
5 milliseconds
from the current 65
it takes for trad
to travel b
New

With new fiber-optic cable,
the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 millisecon

The depth of the Atlantic Ocean va

The new cable will lie on areas of floor that are up to 1,000 feet sh than the current fastest cable. a different route, the new cab shorter, meaning that the tim takes for messages to tra along it is shortened.

The new cable takes a shallower, therefore shorter route.
USA

10% of all photos ever taken were taken in 2011.

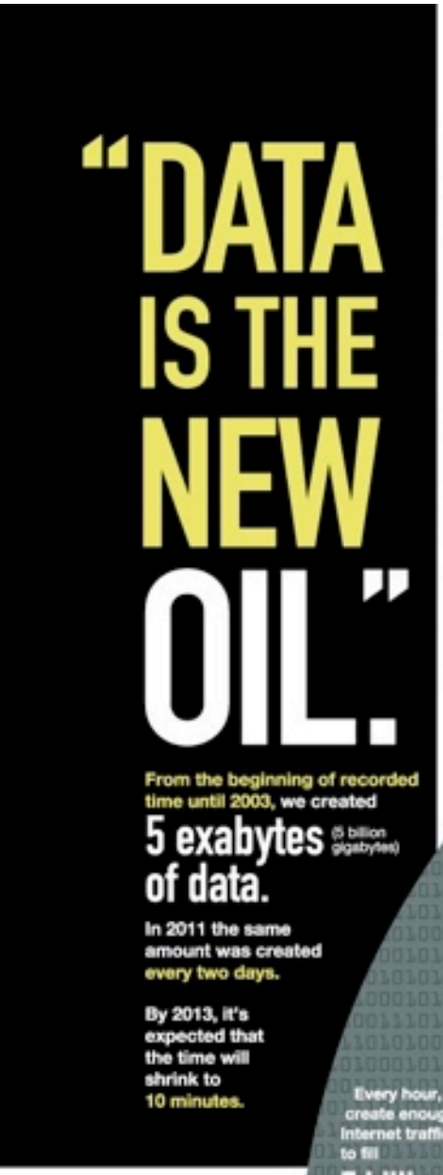
60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

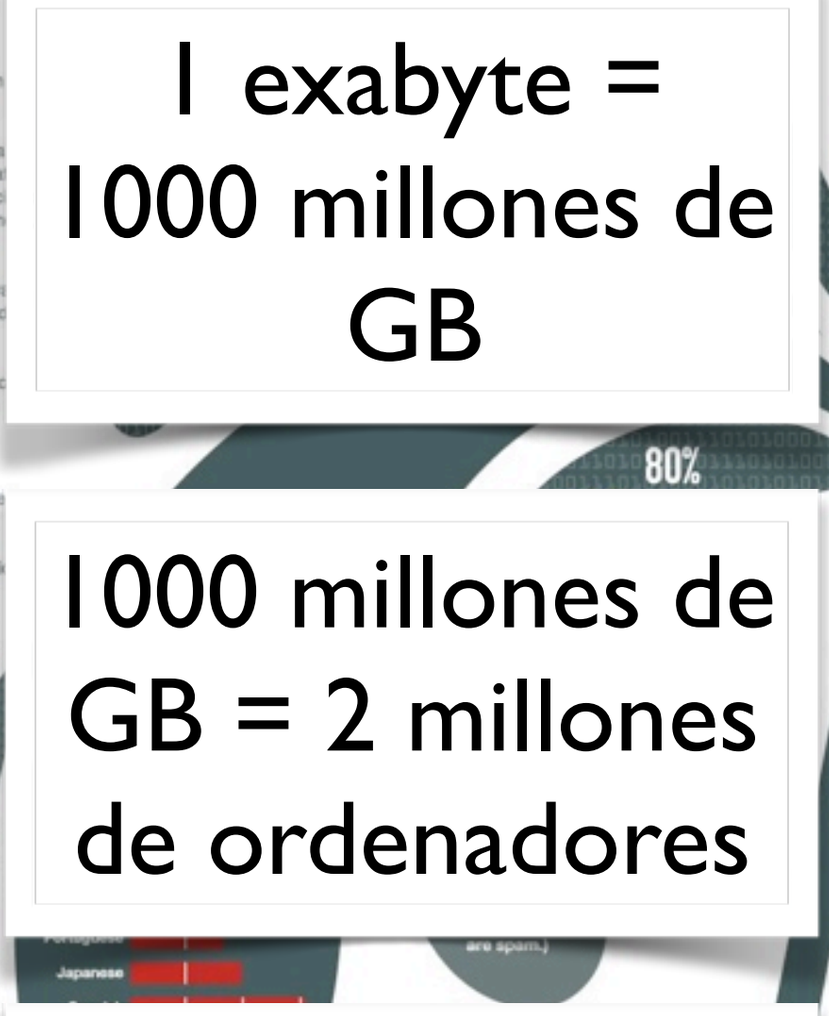
50% of 5-year-old kids in the U.S. are given access to a smartphone.

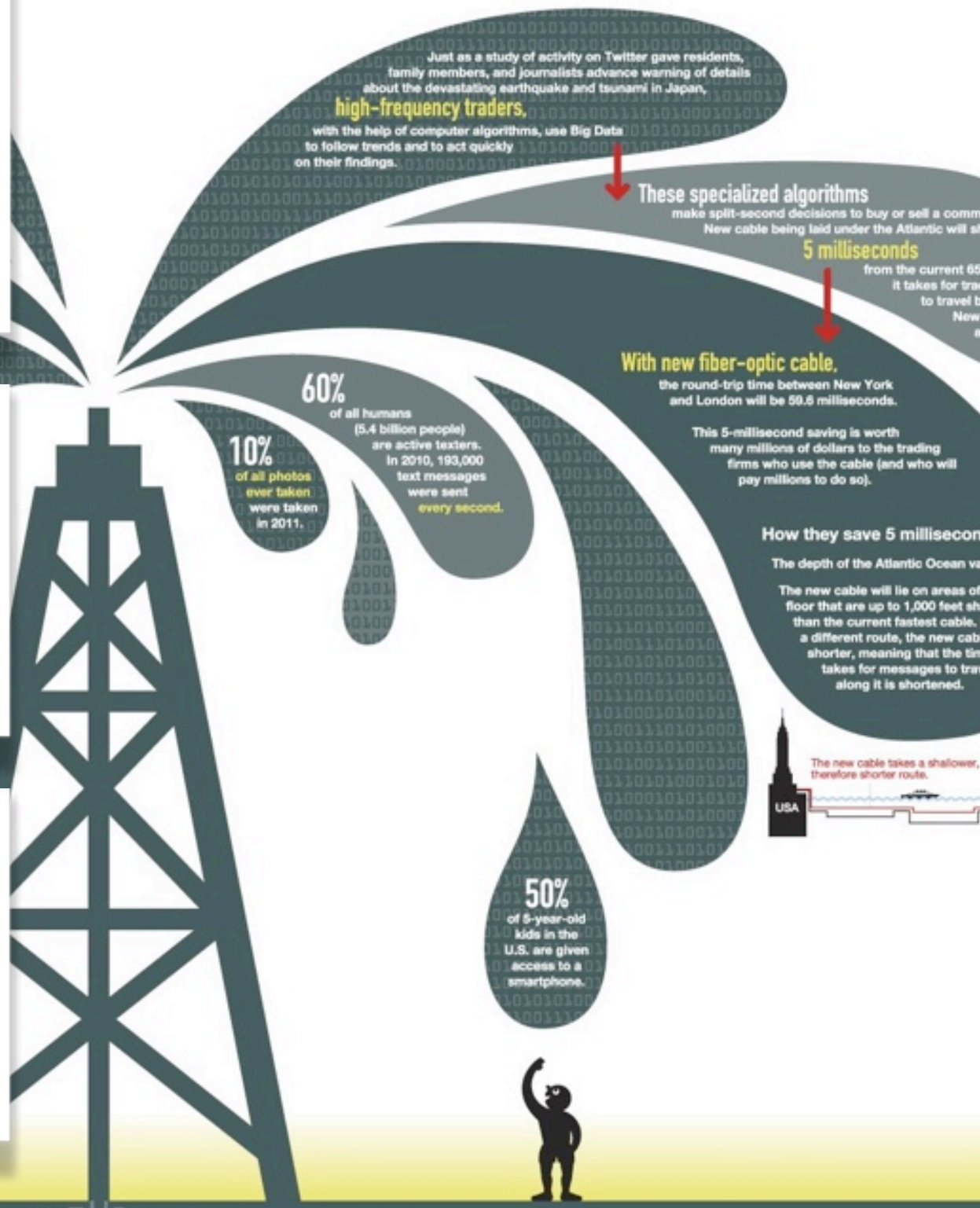
1 exabyte =
1000 millones de
GB

1000 millones de
GB = 2 millones
de ordenadores

generados 5
exabytes hasta
2003





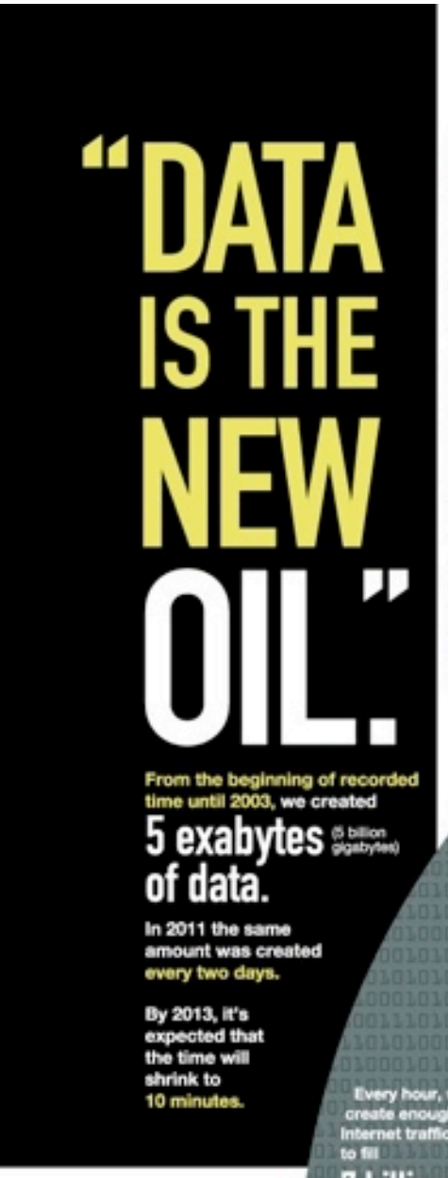


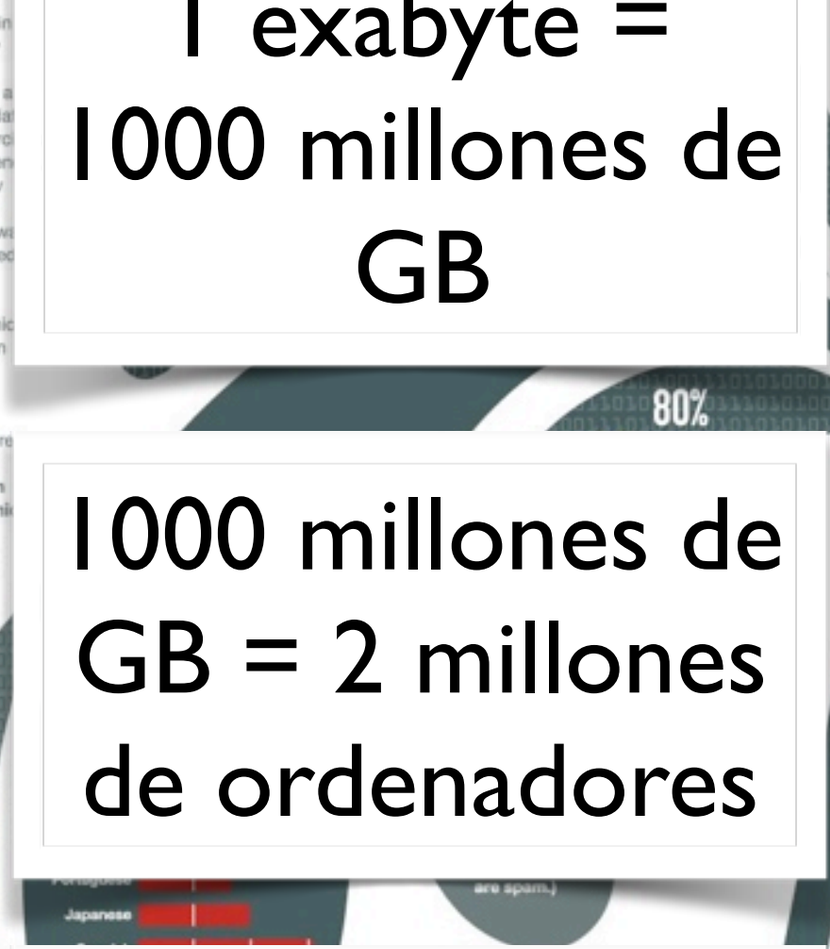
1 exabyte = 1000 millones de GB

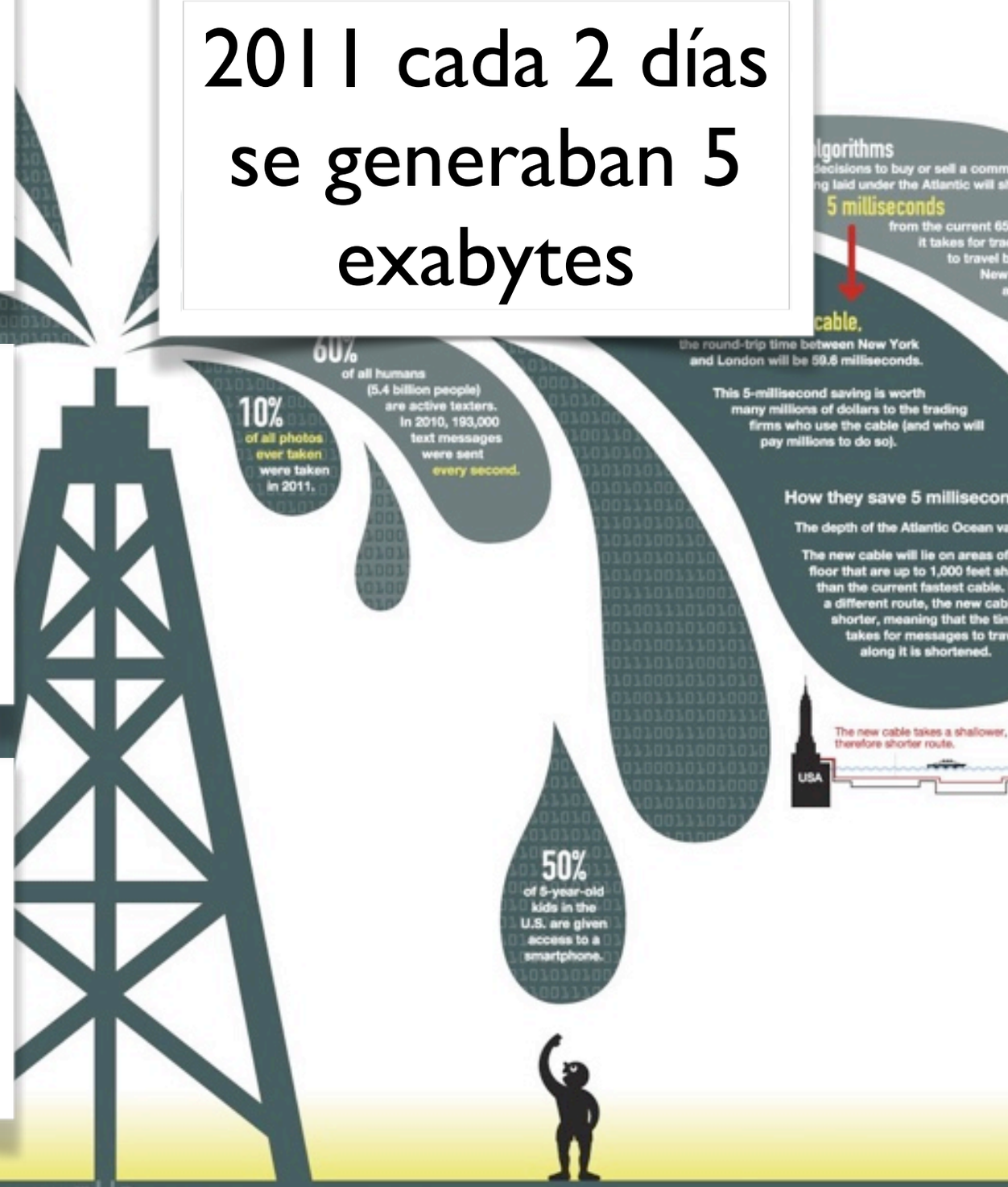
1000 millones de GB = 2 millones de ordenadores

generados 5 exabytes hasta 2003

2011 cada 2 días se generaban 5 exabytes







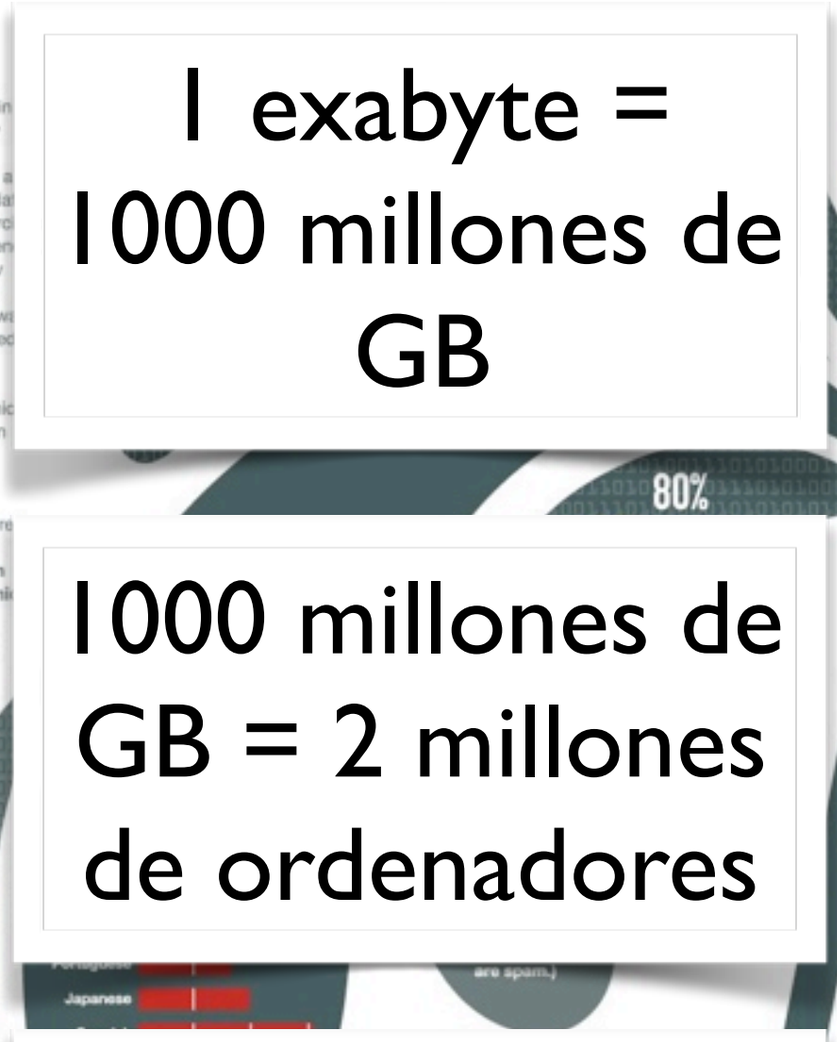
1 exabyte =
1000 millones de
GB

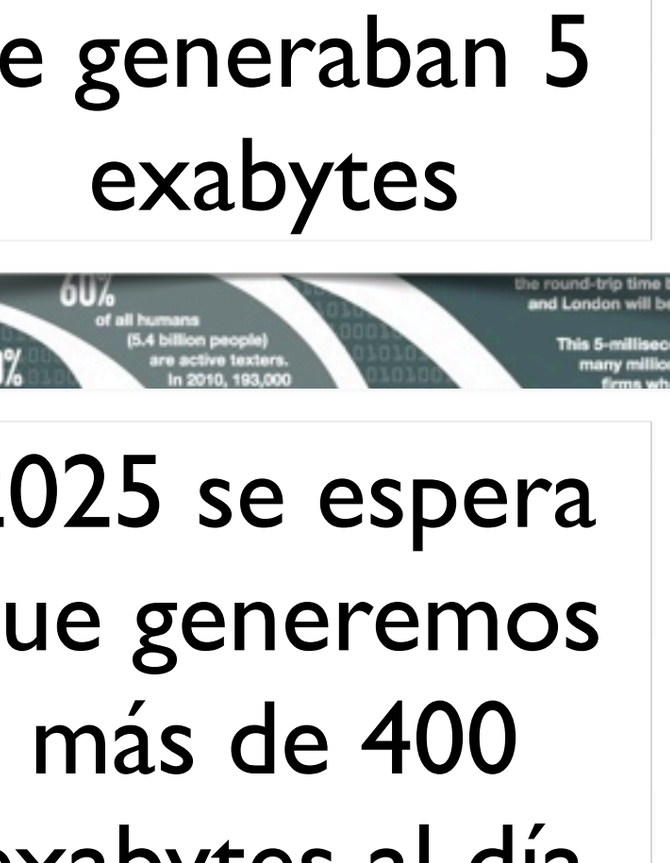
1000 millones de
GB = 2 millones
de ordenadores

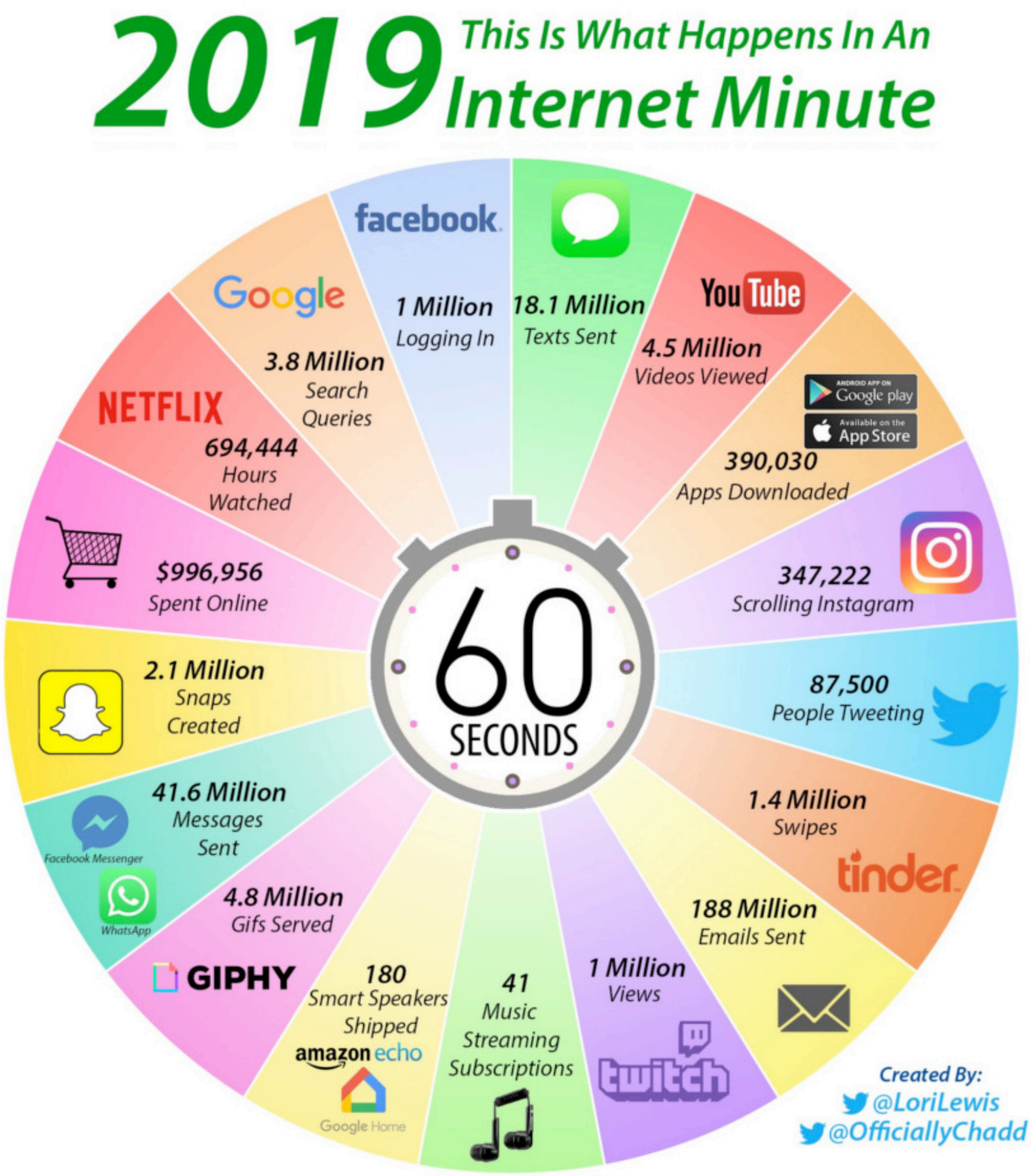
generados 5
exabytes hasta
2003

2011 cada 2 días
se generaban 5
exabytes

2025 se espera
que generemos
más de 400
exabytes al día







Todo lo que hacemos en internet se guarda (o se puede guardar)

Todo lo que hacemos en internet se guarda
o se puede guardar)

- Cada vez recogemos y almacenamos más datos pero *"solo unos pocos % se analizan correctamente"* (Cunef)

- Ej. "las empresas que utilizan sus datos para tomar decisiones tienen 23 veces más posibilidades de conseguir nuevos clientes y 6 veces menos de perderlos" (Mckinsey Business Technology Office)

- También tiene interés en otros campos no económicos: predicción **de crisis climáticas,** gestión de **catástrofes**, **seguridad,** ciudades **inteligentes,** internet de las cosas

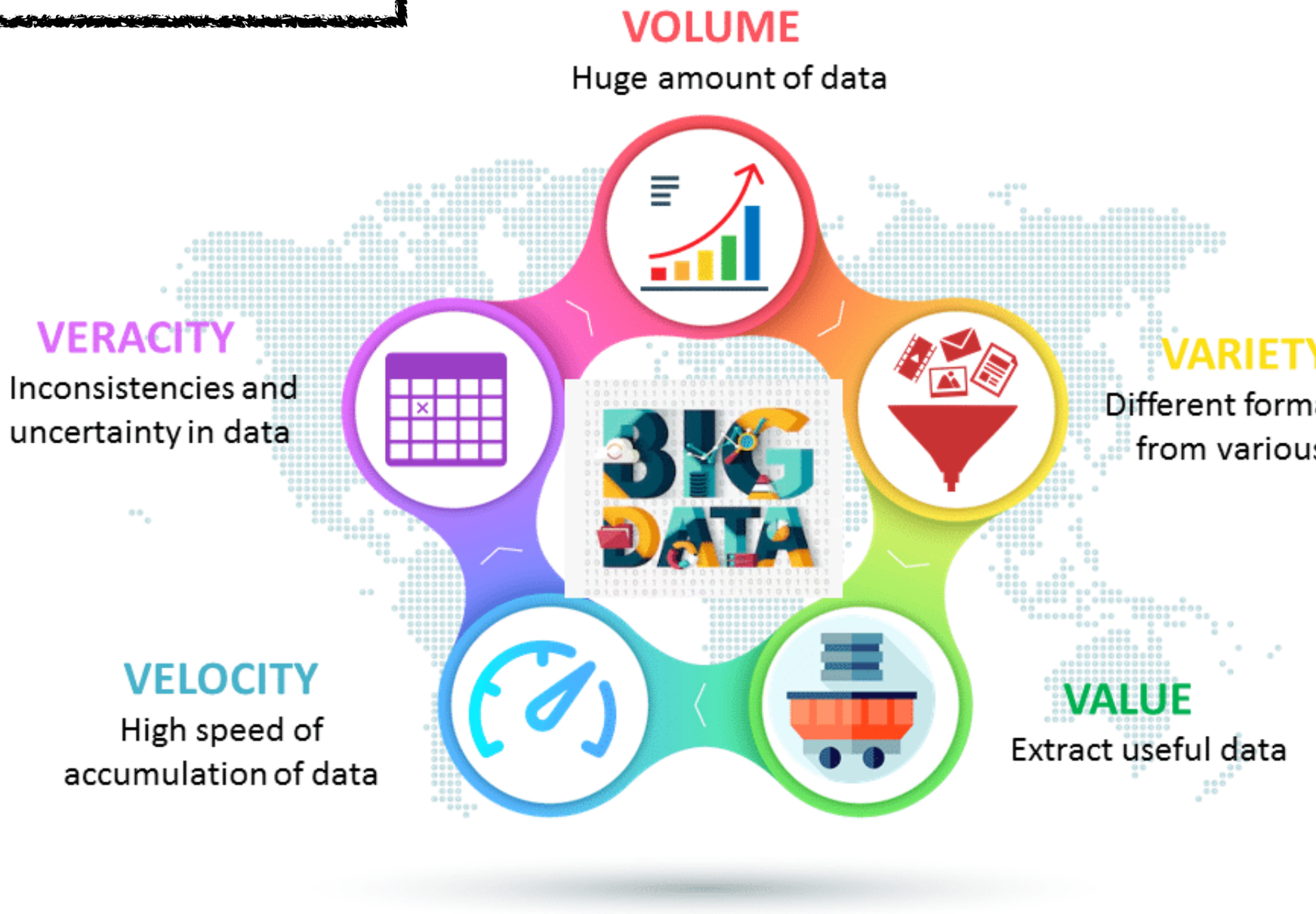
- Cada vez recogemos y almacenamos más datos pero *"solo unos pocos % se analizan correctamente"* (Cunef)

- Ej. "las empresas que utilizan sus datos para tomar decisiones tienen 23 veces más posibilidades de conseguir nuevos clientes y 6 veces menos de perderlos" (Mckinsey Business Technology Office)

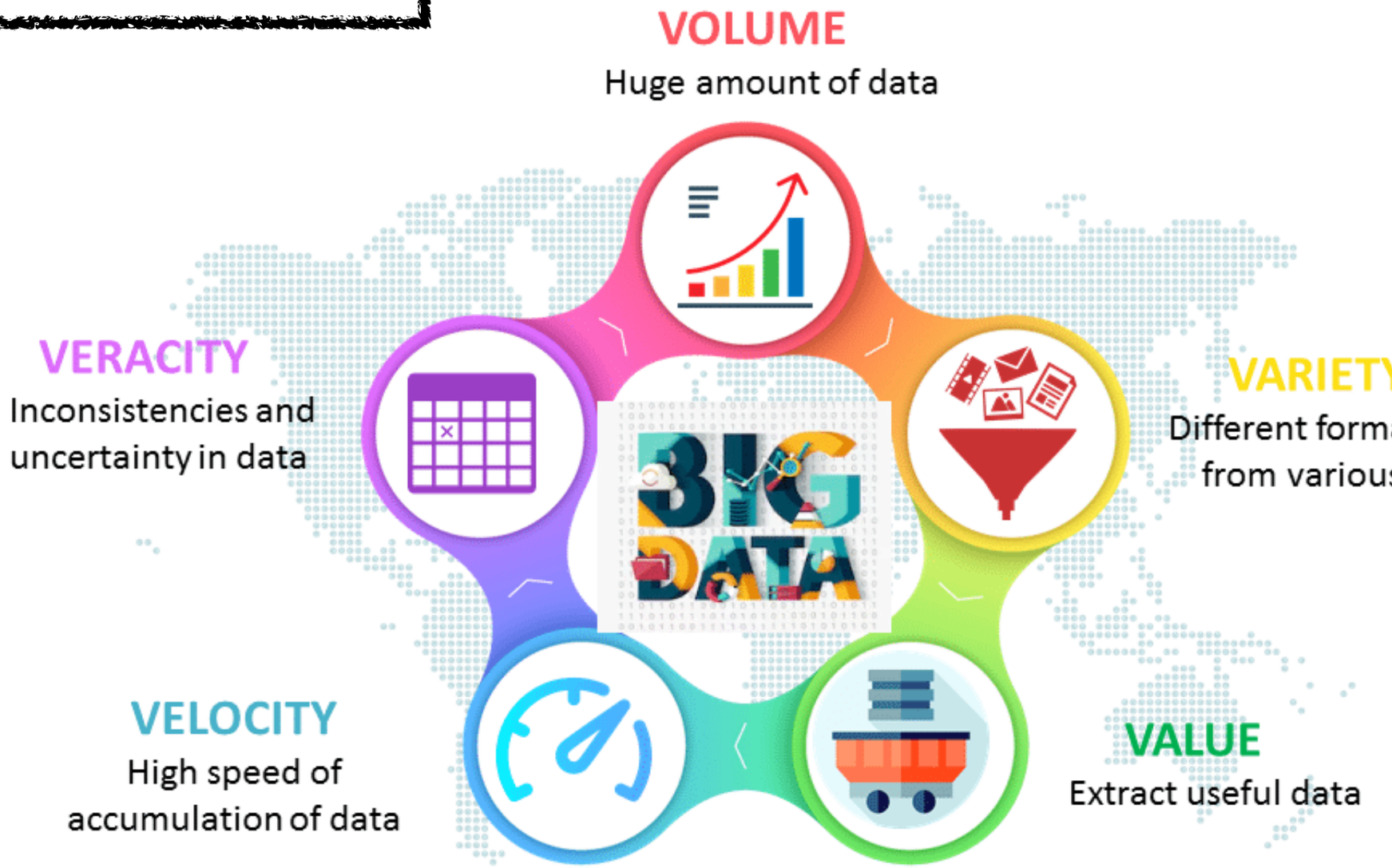
- También tiene interés en otros campos no económicos: predicción **de crisis climáticas,** gestión de **catástrofes**, **seguridad,** ciudades **inteligentes,** internet de las cosas

La Casa Blanca ha invertido más de 200,000 millones en proyectos de Big Data

5V BIG DATA

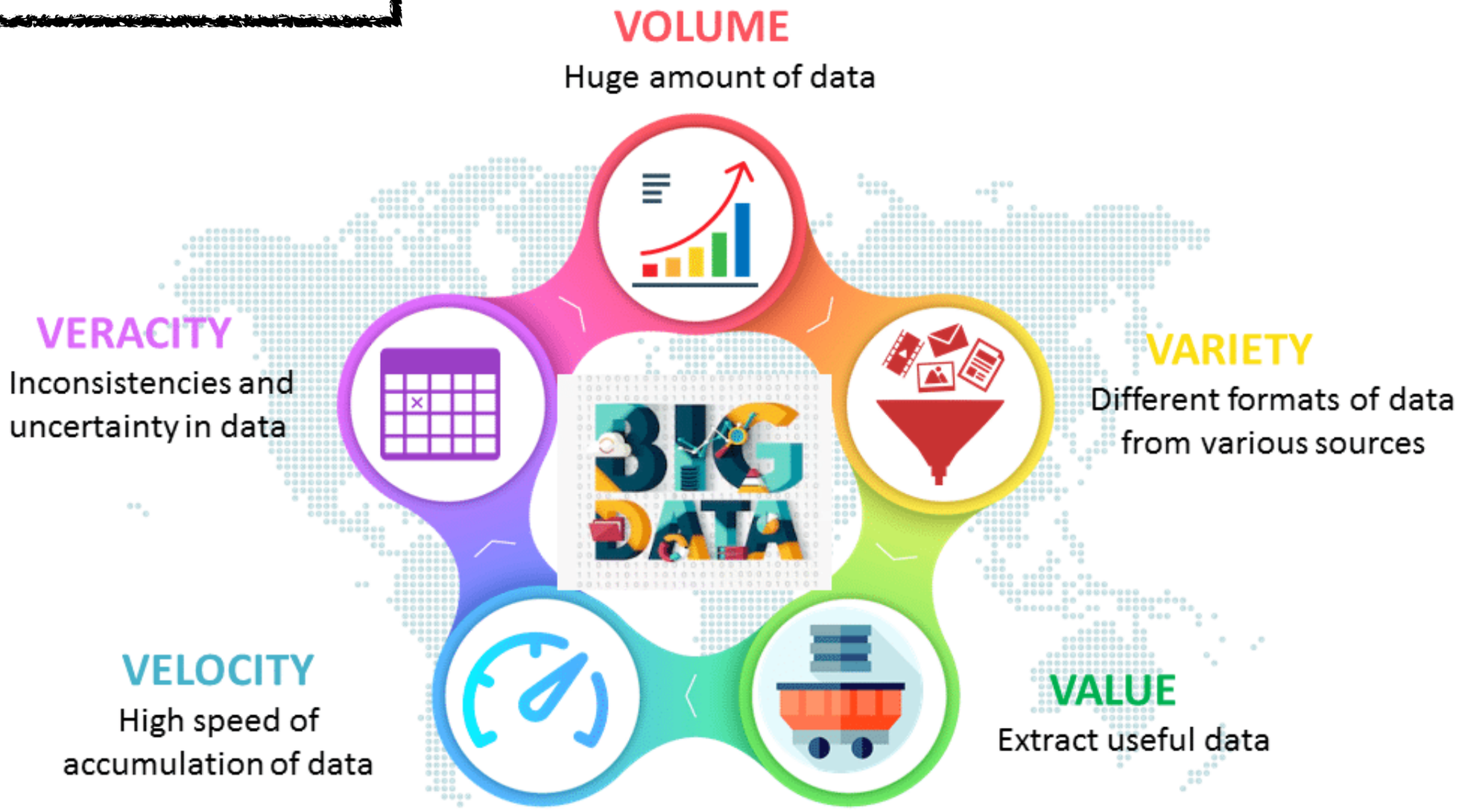

VOLUME
Huge amount of data
VERACITY
Inconsistencies and uncertainty in data
VARIETY
Different formats of data from various sources
BIG DATA
VELOCITY
High speed of accumulation of data
VALUE
Extract useful data

5V BIG DATA



Informáticos, ingernieros de sistemas,...

5V BIG DATA

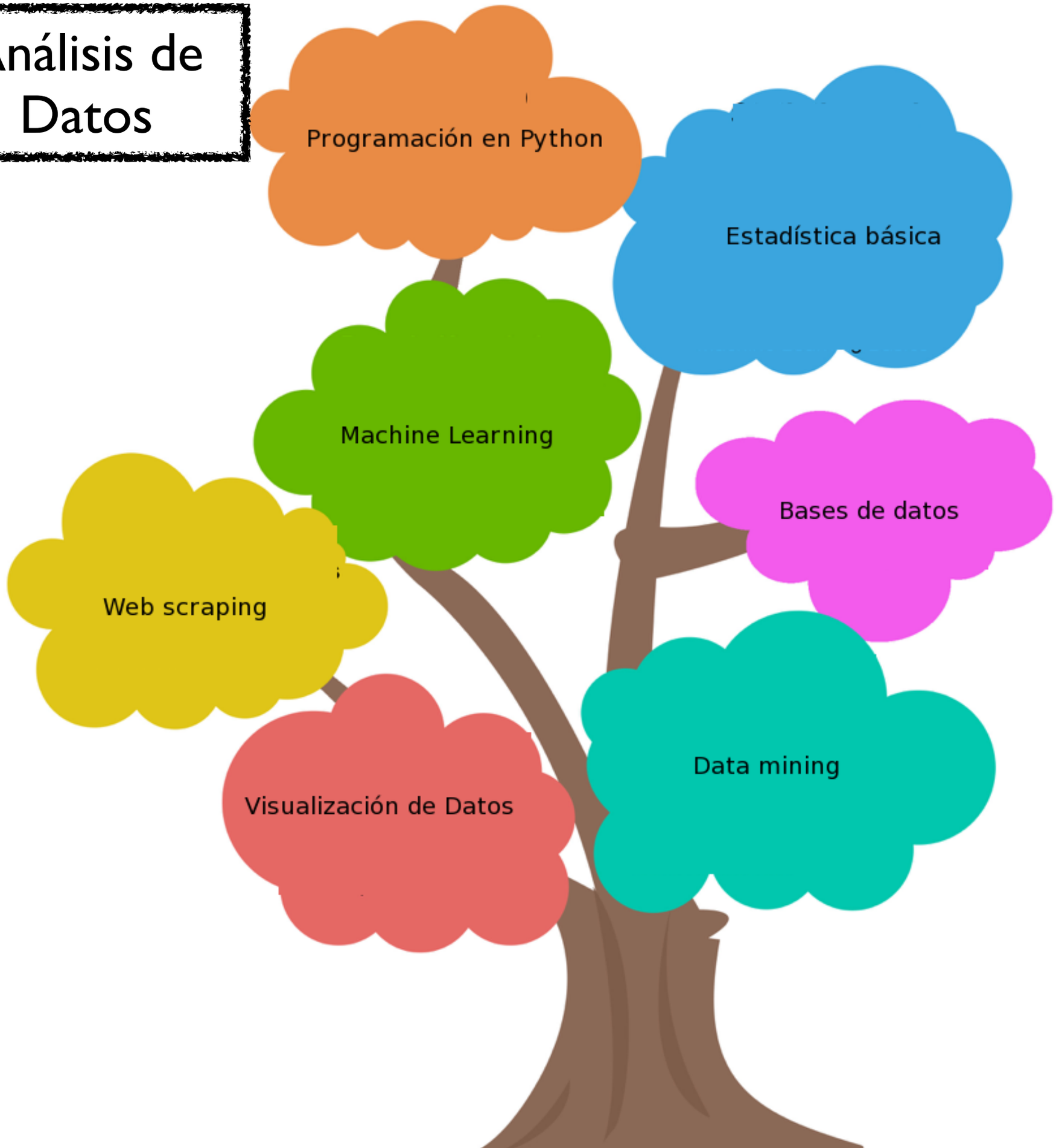


Informáticos, ingernieros de sistemas,...

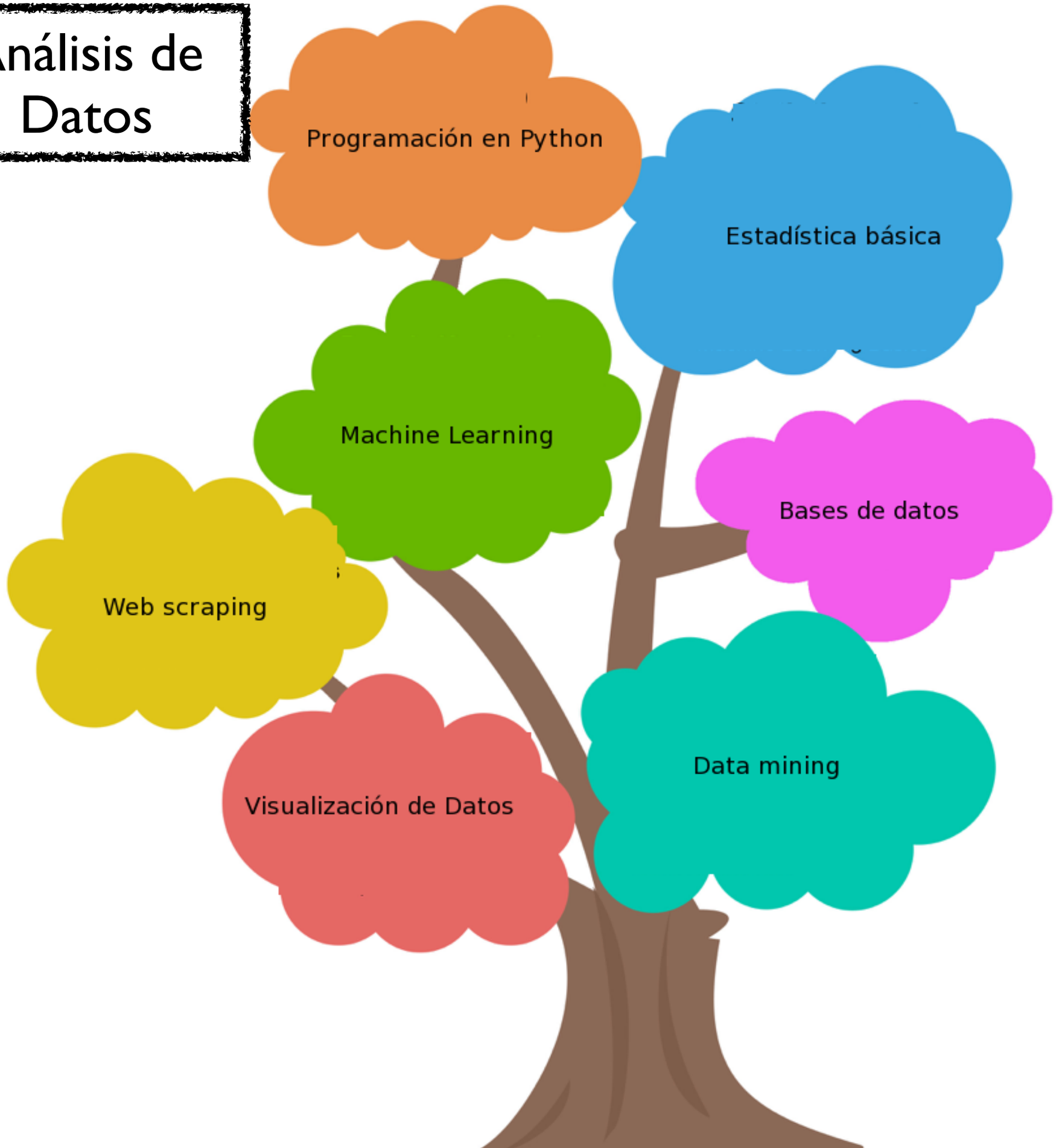
Analistas de Datos (matemáticos, físicos,…)

Análisis de Datos

Análisis de Datos



Análisis de Datos



Análisis de Datos

Programación en Python

Estadística básica

Machine Learning

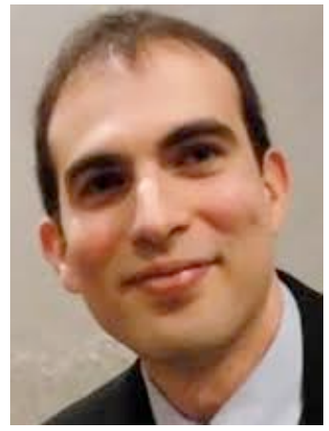
Bases de datos

Web scraping

Data mining

Visualización de Datos

Marcos. Estadística



Biuse. Programación y redes neuronales

Carlos. Bases de datos y automatización de sistemas

Diego. Bases de datos, web scraping, machine learning

Steven. Visualización

Let's talk about you

Hemos asumido:

Al empezar:

* nivel de matemáticas bajo

* nivel programación zero

* cultura digital **alta**

Let's talk about you

Hemos asumido:

Al empezar:

* nivel de matemáticas bajo

* nivel programación zero

* cultura digital **alta**

Al acabar:

* nivel de matemáticas un poco más alto

* conocimientos de programación

* cultura big data

* Soltura en herramientas para interaccionar con datos

Programación en Python

Programación

Programar : Hablar con cualquier dispositivo (móvil, ordenador,…) en un leguaje que el ordenador entienda (java, C++, python,...)

Órdenes sencillas. Ej. suma, ve al link, etc

O más complejas: Secuencia de órdenes.
script, código, programa

Programación

Programar : Hablar con el ordenador (móvil, ordenador,…) en un lenguaje que el ordenador entienda (java, C++, …)

Órdenes sencillas. Ej. suma, ve al link, etc

O más complejas: Secuencia de órdenes.
script, código, programa

TODO lo que hacemos con el ordenador ha estado programado por alguien. *Apps, E-mail, búsquedas Google,...*

Y hacen **exactamente** eso.

Programación

Un código no tiene porque hacer cálculos:

ej. 1

1. Abre un archivo
2. Busca la palabra árbol
3. Cambiala por cerezo
4. Guarda archivo

ej.2

1. Si presiono flecha mueve bruja
2. Mueve bruja hasta que presione flecha
3. Si bruja toca dragón, bruja muere

4. Si toca enter bruja hace hechizo
5

Programación

Un código no tiene porque hacer cálculos:

ej. 1

1. Abre un archivo
2. Busca la palabra árbol
3. Cambiala por cerezo
4. Guarda archivo

ej.2

1. Si presiono flecha mueve bruja
2. Mueve bruja hasta que presione flecha
3. Si bruja toca dragón, bruja muere
4. Si toca enter bruja hace hechizo
5.

Scratch

Programación en Python

Python :

- Sencillo
- Mucha información On-line
- Muchos páquetes ya programados para usar
- Interactúa bien con otros programs (excel, VLC, etc...)
- Gratis !!

¿Quién usa python?:

- Google
- NASA
- Youtube
- IFCA
- .....

Programación en Python

Python :

- funciones básicas 3+2, type(a), print()
- y paquetes numpy, scipy, matplotlib, …

Anaconda :

- Plataforma de python que te permite instalar paquetes fácilmente

- No es estrictamente necesario

Programación en Python

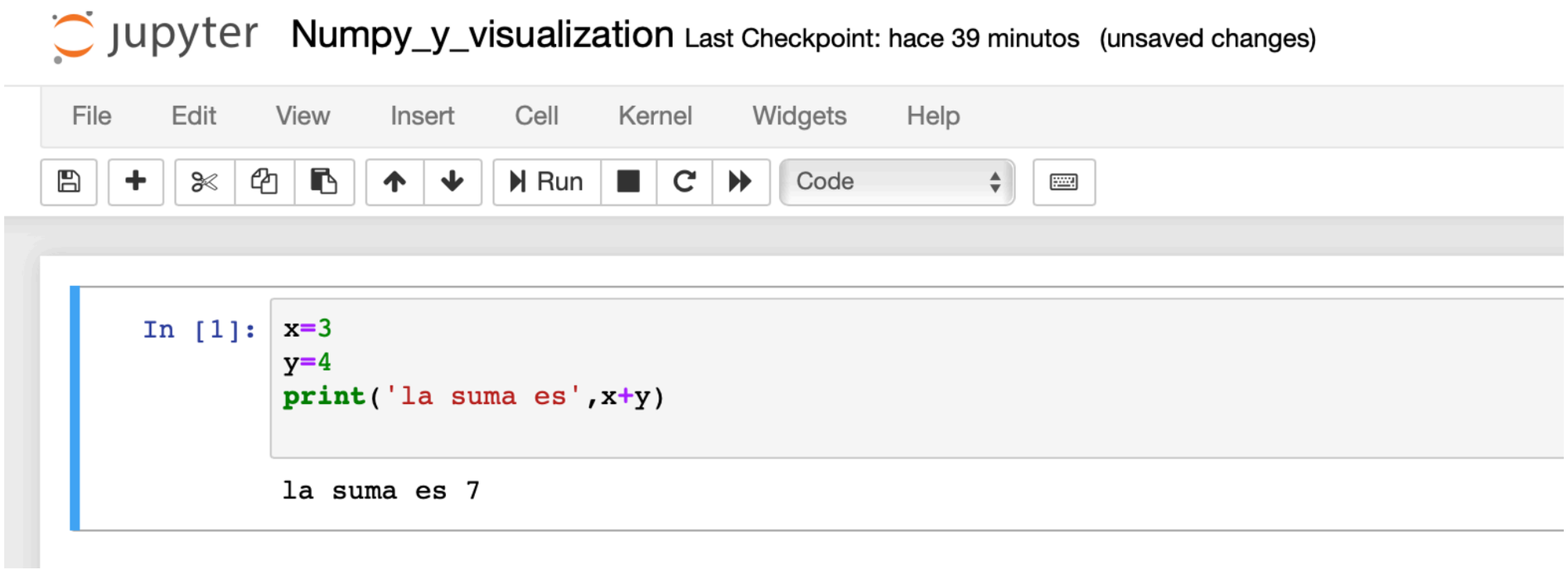
Documento de texto .py :

codigo.py

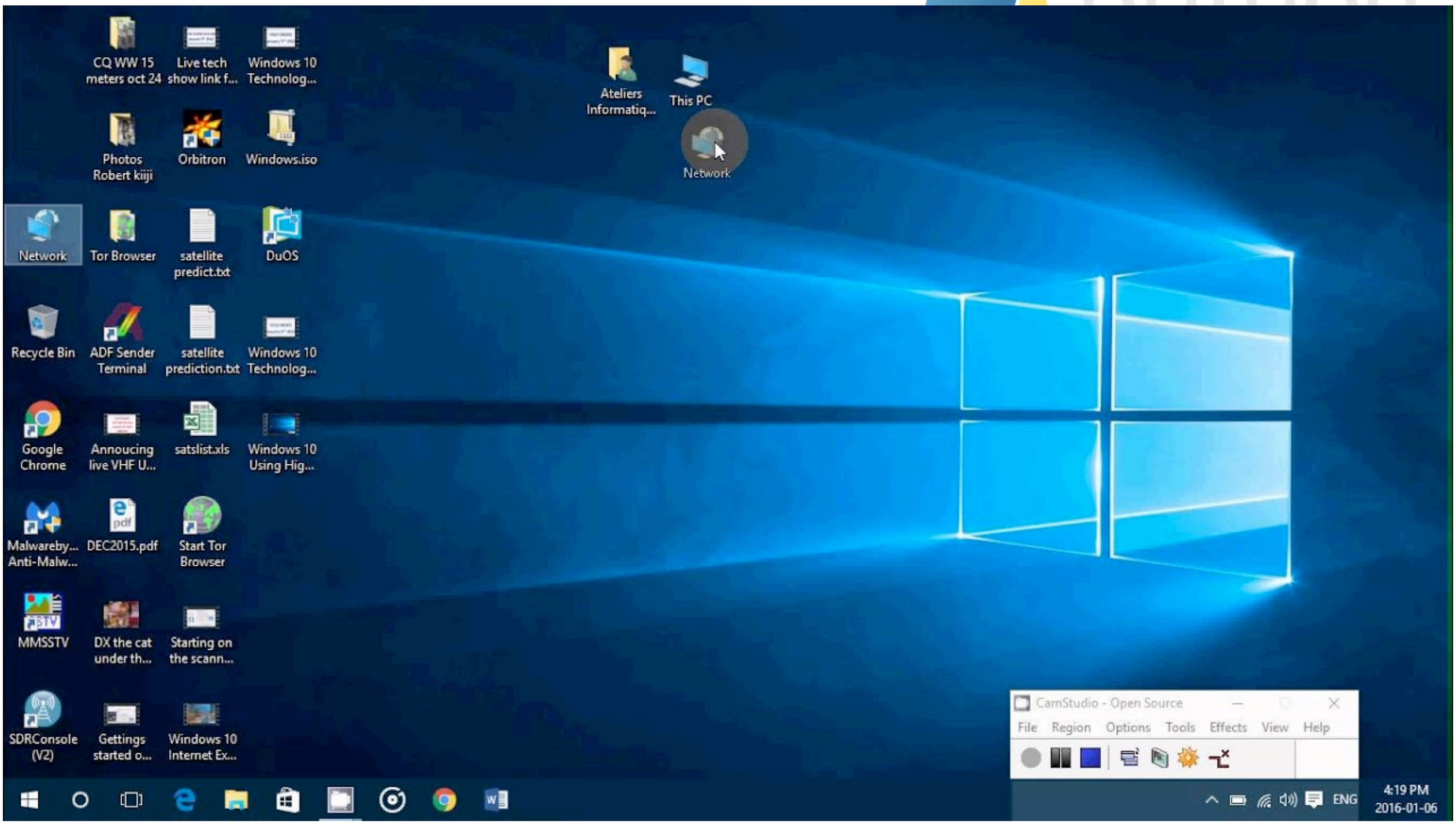
```
# programa python
x=3
y=4
print('la suma es',x+y)
```

```
>> python codigo.py
>> la suma es 7
```

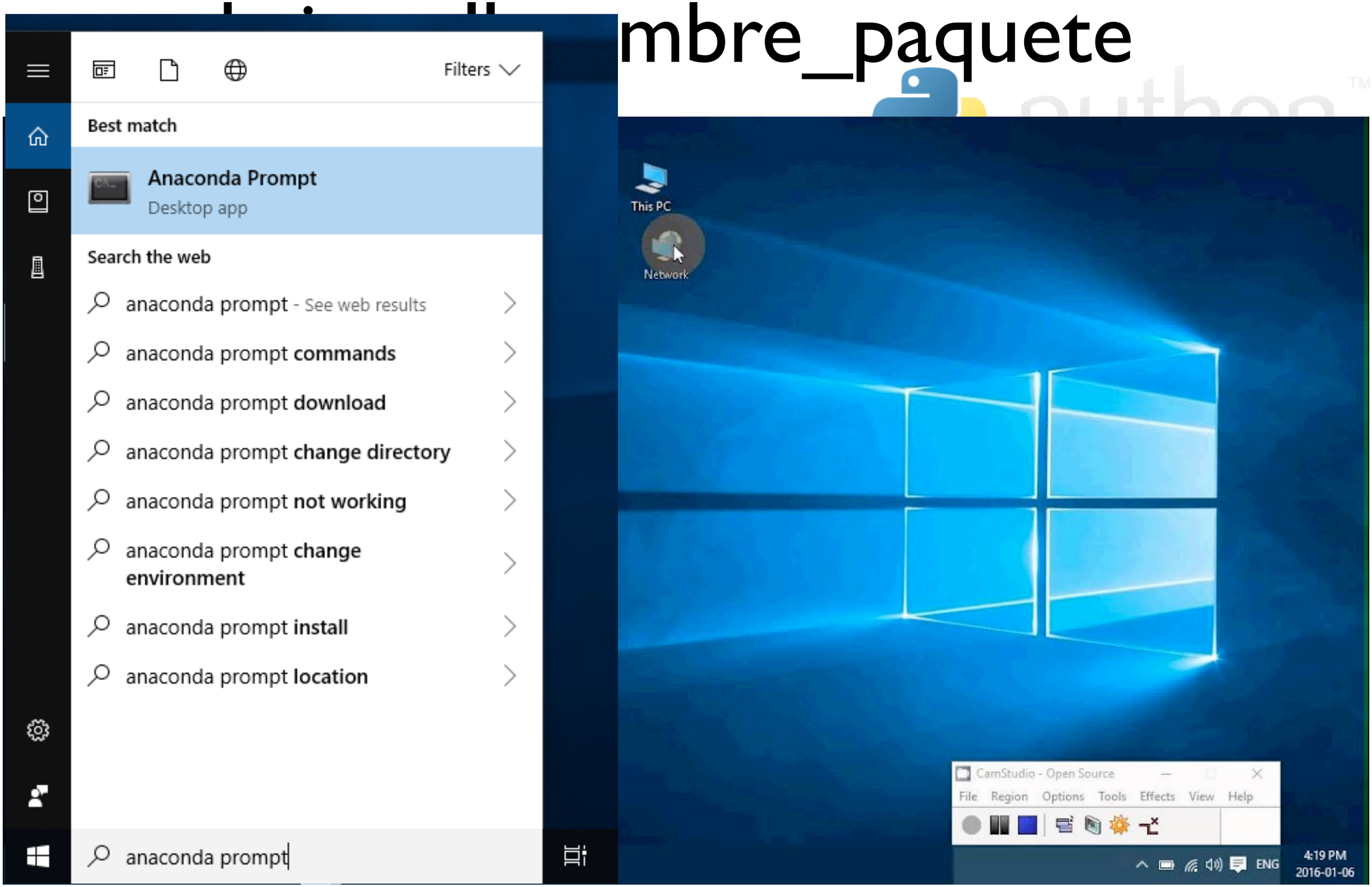
Jupyter Notebook:



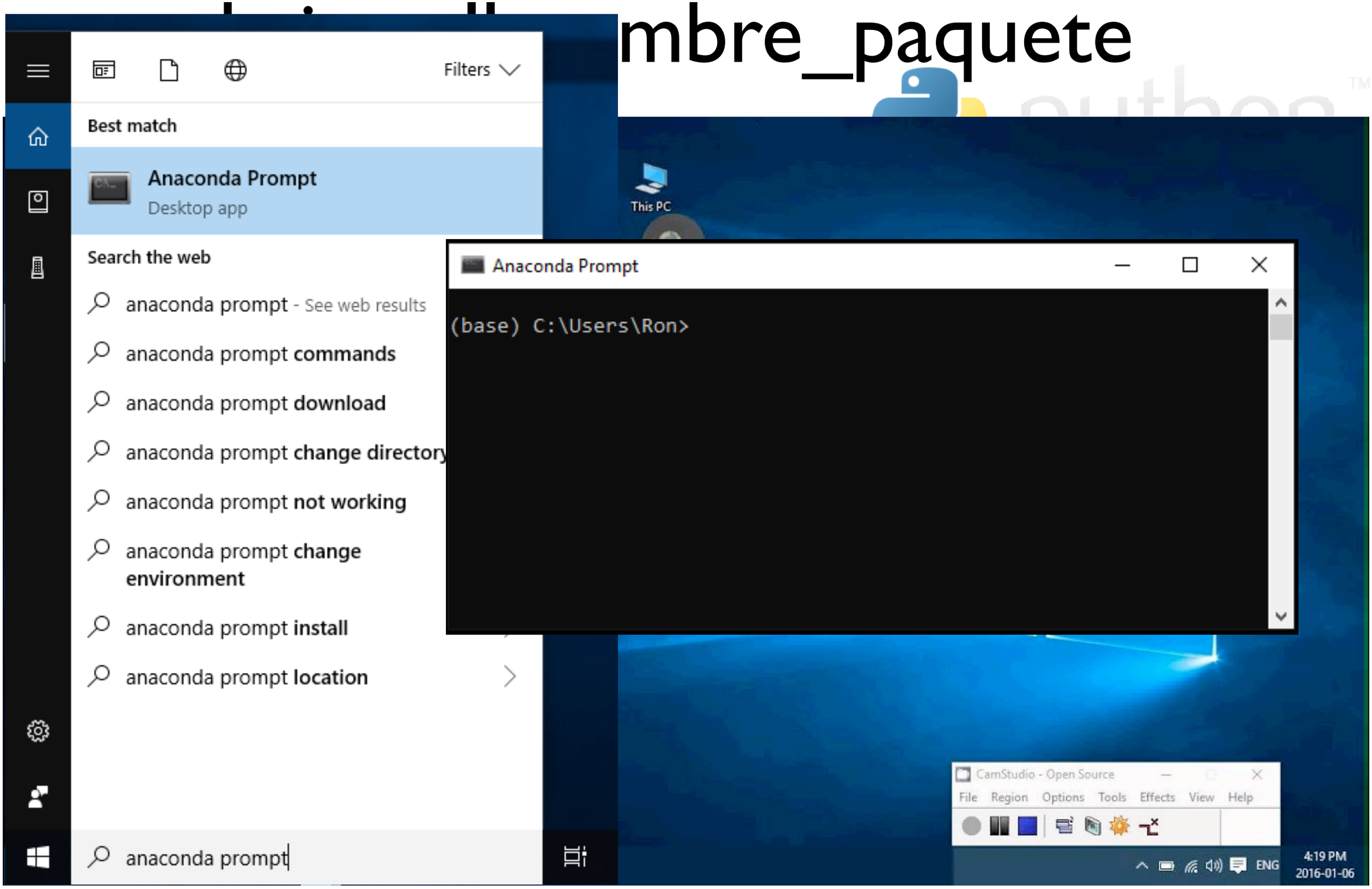
Instalación de paquetes:
conda install nombre_paquete



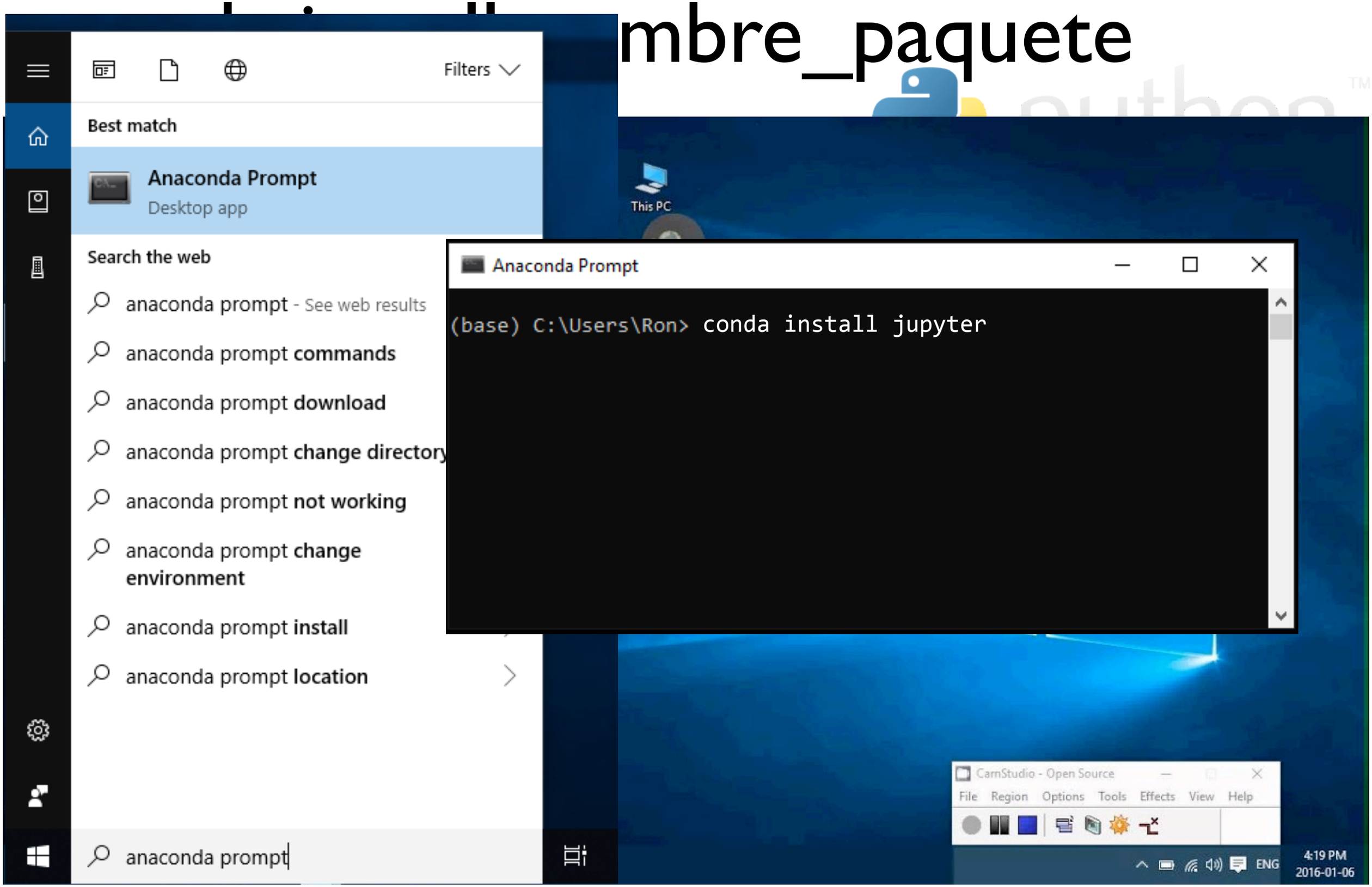
Instalación de paquetes: mbre_paquete



Instalación de paquetes: mbre_paquete



Instalación de paquetes: mbre_paquete



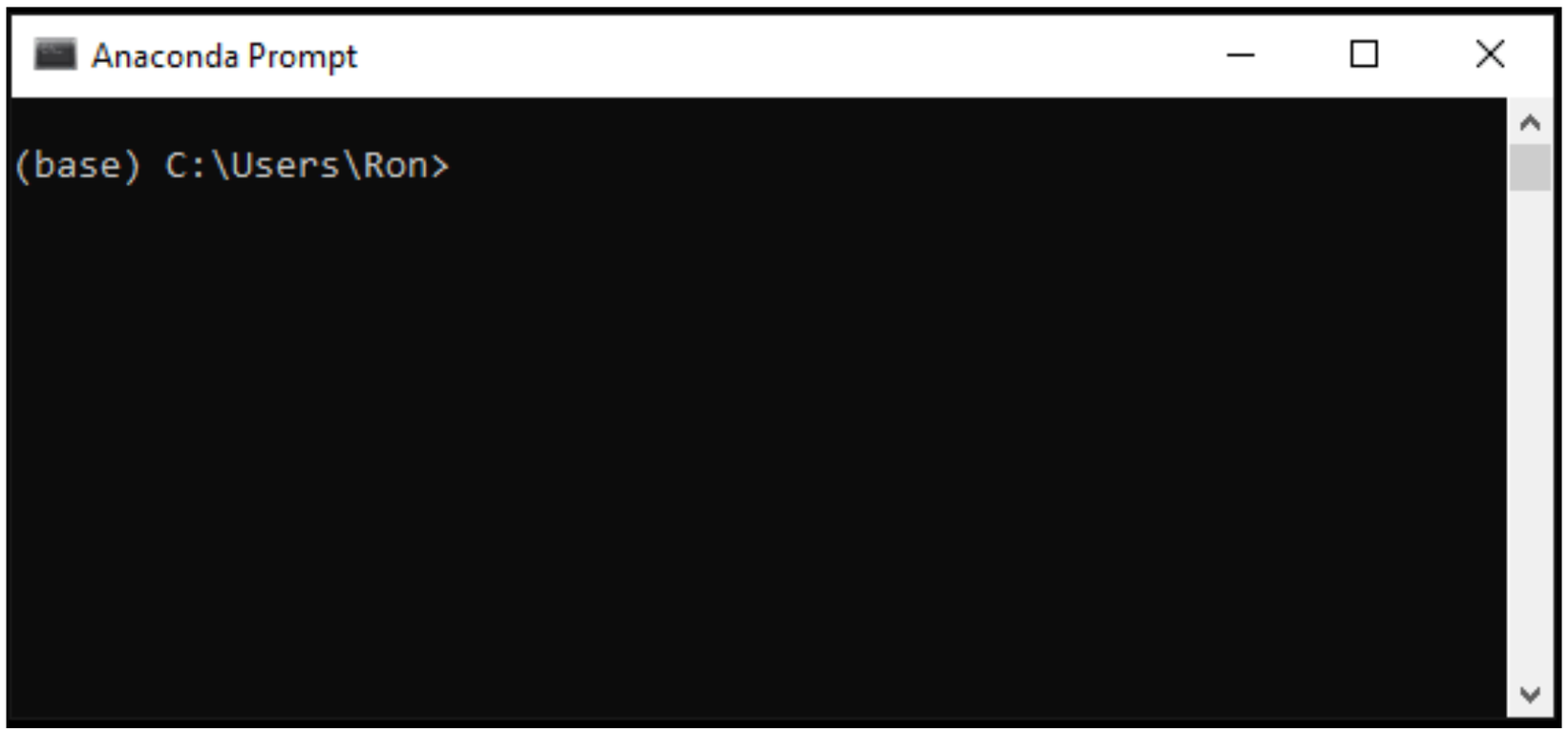
Programación en Python

Abrimos la notebook (libreta)
de Jupyter

Programación en Python

python™

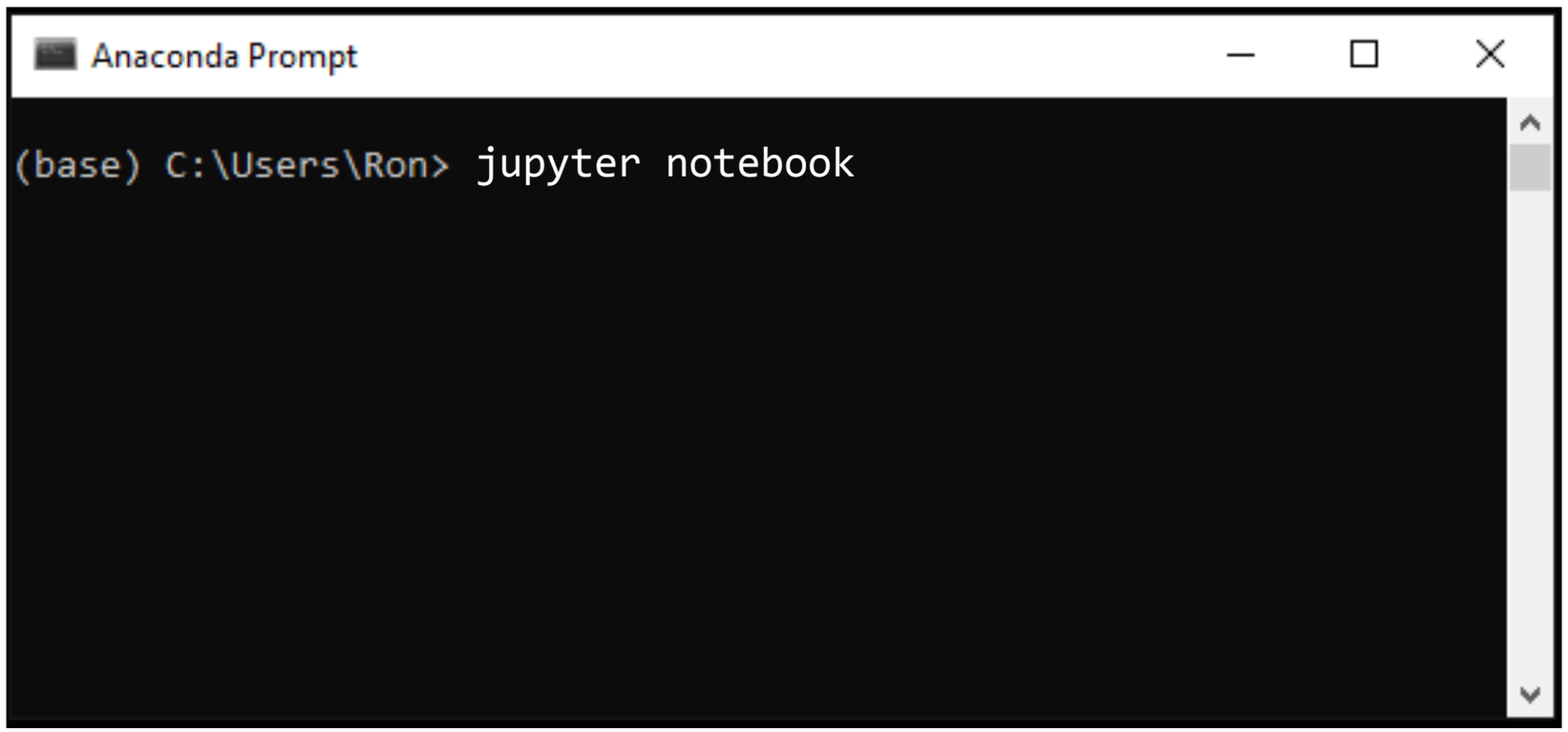
Abrimos la notebook (libreta) de Jupyter



Programación en Python

python™

Abrimos la notebook (libreta) de Jupyter



Programación en Python

https://github.com/biuse/TallerEmpleo