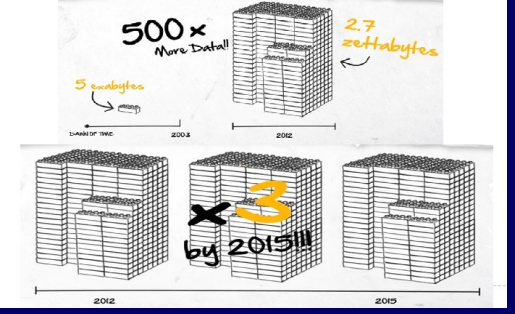
BIG DATA

- WHAT IS BEHIND THE TERM? -



BIG DATA

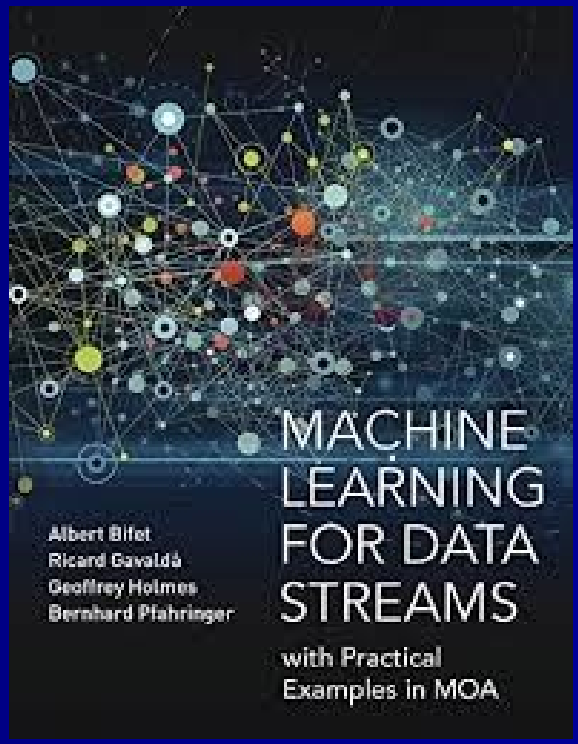
- New technological concept, new generation of technologies
- In part, "inheritor" of data mining
- The term "was coined by 2015-2016"

- Related to the challenges exposed to manipulate massive datasets (petabytes, exabytes):
  - Capture and storage
  - Processing and computing
  - Analysis and mining

- New data-types: Social networks, Electronical purchases, financial companies, GPS systems, weather, sensors, images…

BEFORE THE BIG DATA "FASHION": DATA STREAMING

- "Pioneers" of massive data analysis
- Big data: save all data
- Data streaming: no
- Online, continuous adapting learning of the model
- Concept drift: detect changes in training samples



BIG DATA

- "Save all the data"
- Traceability
- Data storage and computing challenge

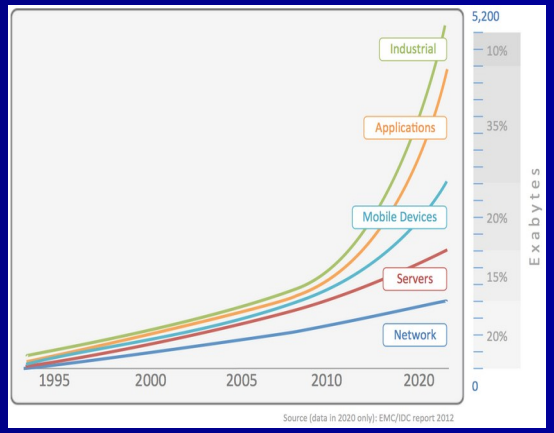
- Business oportunity
- Oportunity for humanitarian action

Big Data, Big Impact:
New Possibilities for International Development

UNITED NATIONS GLOBAL PULSE
Harnessing big data for development and humanitarian action

WHY "BIG"?

- Megabytes, Gigabytes… → Terabytes, Exabytes
- Classic numerical matrices … → images, text, links, localizations
- PC's … → advanced parallel computing platforms
- Excel sheets, databases … → more advanced hosting database systems (e.g. MongoDB)



BIG DATA

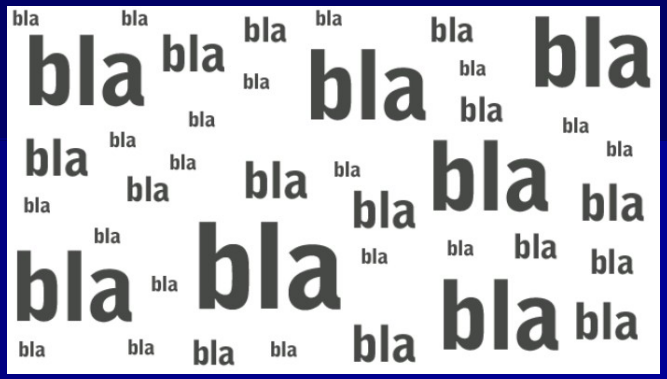
- Parallel computation
- Massive database storage

- Grid and cloud computing, GPUs, CPUs...
- NoSQL, MongoDB non-relational...

- Application fields → credit card transactions, telecommunication companies, web processing and traffic, industry, social networks...

"Big data is like teenage sex;
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it".

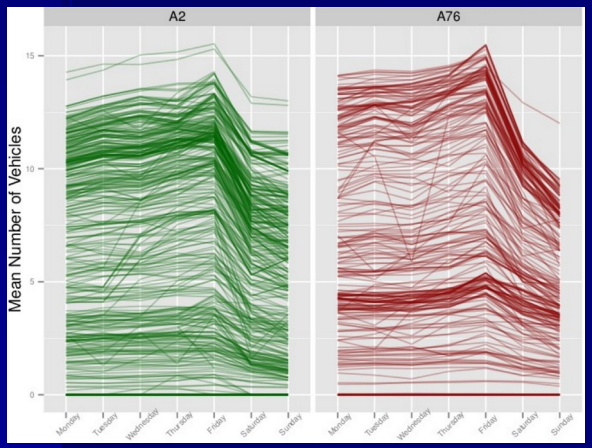
Dan Ariely, Duke University

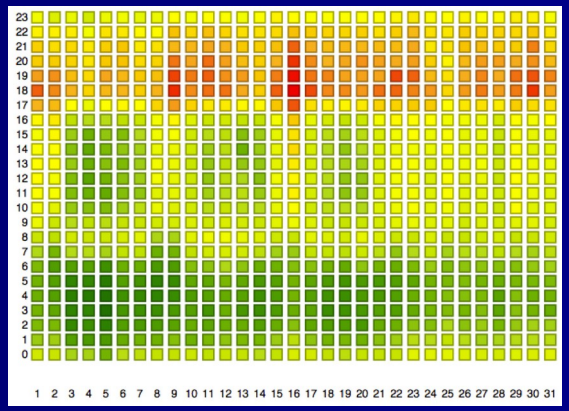


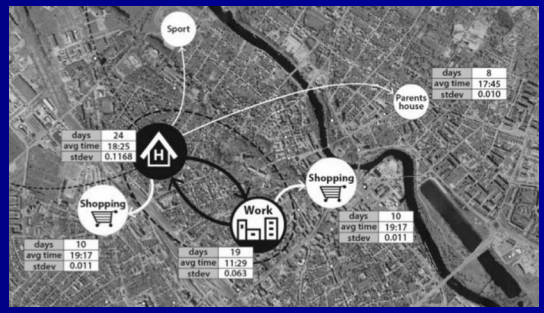
APPLICATIONS (I) - EuroStat

- ESSnet (European Statistical System) Big Data project [link]

- Traffic loops: visualizing hourly traffic intensity
- Electricity consumption visualization
- Phones' usage: people movement in a city and tourism indicators







HERALDO
Viernes, 2 de diciembre de 2016, actualizado a las 19:08 horas

'Big data' para saber cómo se desplazan los aragoneses

El Gobierno de Aragón ha adjudicado un contrato pionero a nivel autonómico para, a través del posicionamiento de los teléfonos, conocer detalles de la movilidad en la región.

smartphone. **Un teléfono conectado a internet que deja un 'rastro' considerable de datos a todas horas:** páginas webs consultadas, mensajes recibidos, posicionamiento del terminal... El análisis de estos registros es lo

estudio del posicionamiento de los teléfonos. **La información recogida por los teléfonos móviles, una vez analizada pormenorizadamente, servirá para revisar de la organización autonómica del transporte públic**o en la

El estudio que ha lanzado la DGA utilizará información **de 600.000 usuarios recopilados entre el 5 de noviembre y el 2 de diciembre de 2015**. "Se han seleccionado estas fechas porque suponen un

EL PAÍS

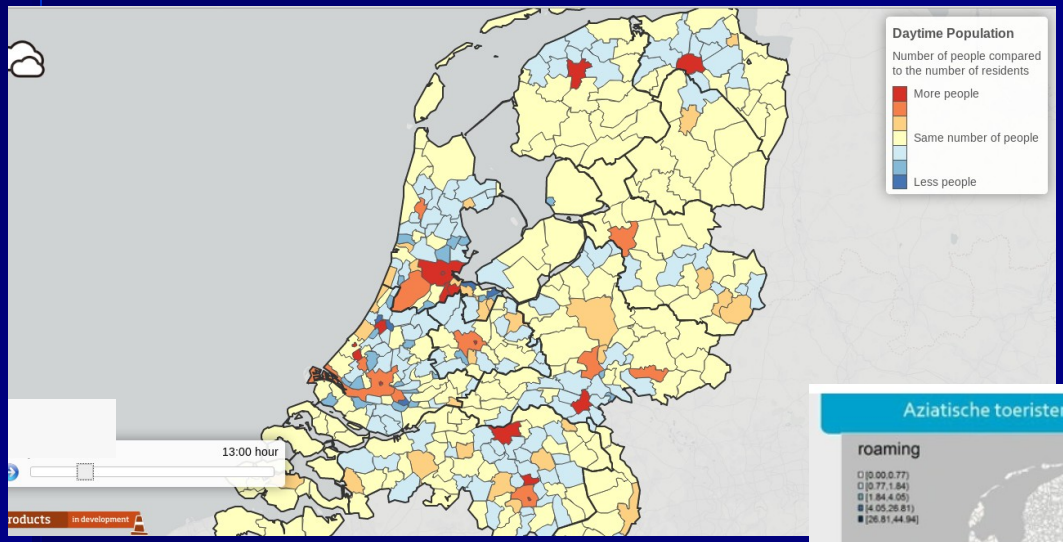
El INE seguirá la pista de los móviles de toda España durante ocho días

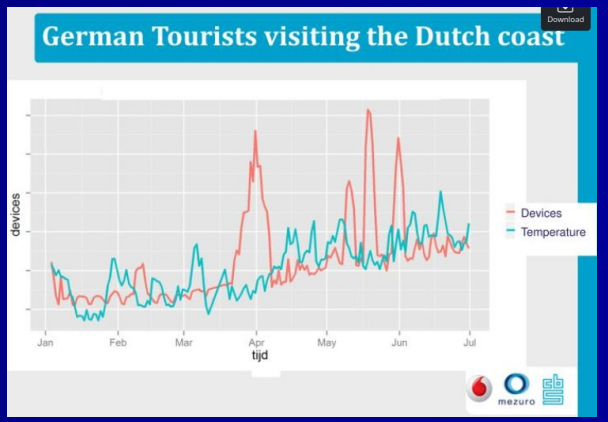
El Instituto Nacional de Estadística pacta con las operadoras realizar un estudio sobre movilidad empleando información anónima

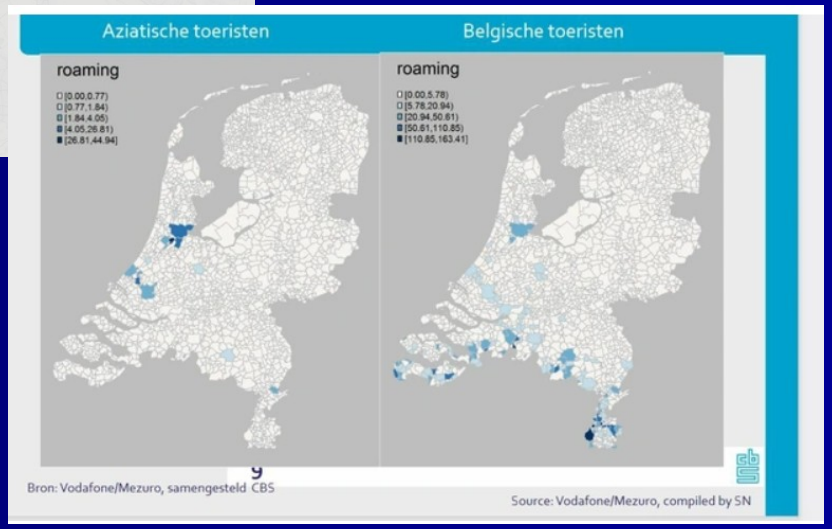
El Instituto Nacional de Estadística (INE) conocerá cómo se mueven los españoles gracias a sus teléfonos móviles. Durante cuatro días laborables de noviembre, un domingo y tres días de vacaciones seguirá los movimientos de los terminales, según confirman fuentes del organismo. Eso sí, los datos serán completamente anónimos: merced a un acuerdo pionero en Europa con las tres principales operadoras, el instituto estadístico recibirá las posiciones agregadas de los números, pero no los titulares de las líneas. La información es relevante para averiguar cuáles son los desplazamientos habituales de la población y, por tanto, dónde se deben prestar los servicios públicos y reforzar las infraestructuras. También se sabrá adónde van los españoles de vacaciones dentro del territorio nacional. La operación empezará en tres semanas.

APROVECHAR MÁS EL 'BIG DATA' Y MENOS LAS ENCUESTAS

CBS – Holland Statistics Institute
In collaboration with Vodafone and mezuro.com

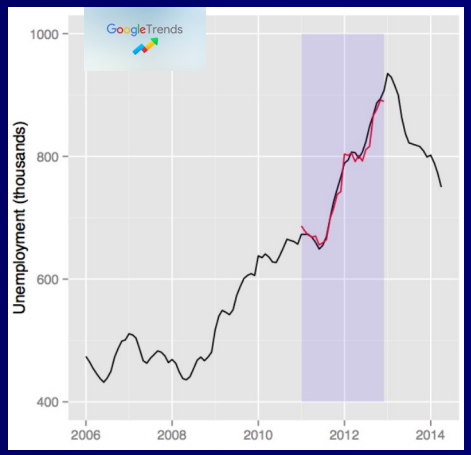






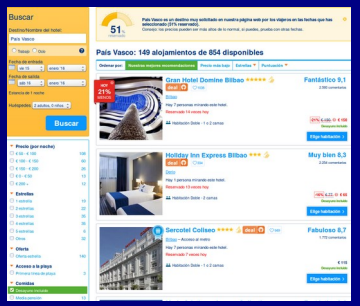
APPLICATIONS (II) - EuroStat

- GoogleTrends searched terms → prediction of future unemployment
- Wikipedia pages-views' logs → relation with tourism
- Webscraping job offer portals



APPLICATIONS (III) - EuroStat

- Webscraping real state market → prices and new dwelling
- Webscraping daily hotel prices in websites
- Webscraping online supermarket prices
  - (Alternative) Consumer price index calculation



APPLICATIONS (III) - EuroStat

La tarifa media diaria (ADR) y, por lo tanto, también los ingresos por habitación disponible (RevPar), se obtienen a partir de octubre de 2020 con técnicas de Big Data

Se informa que en octubre de 2020 Eustat ha introducido en sus procesos de producción de la ETR técnicas avanzadas de ***Big Data*** para la obtención de la tarifa media diaria (ADR) y, como variable derivada, también de los ingresos por habitación disponible (RevPar). Para cada establecimiento hotelero se captura de las plataformas de reserva *online* el precio para cada día los 120 días precedentes utilizando tecnología de *web scraping*. Esta información se combina con la del directorio de establecimientos turísticos de la encuesta mediante modelos estadísticos validados con información muestreada en los últimos años. Para más información, se puede consultar el *documento metodológico* publicado en la web de Eustat.

APPLICATIONS (IV) - EuroStat

- Satelite images
- EU Copernicus project → free images

- Agriculture → Prediction of crop type
- Detection of solar panels in roofs
- Green areas in a city

LOCAL APPLICATION BASQUE INDUSTRY 4.0

- 'Basque Industry 4.0'
- Machine Tool sector

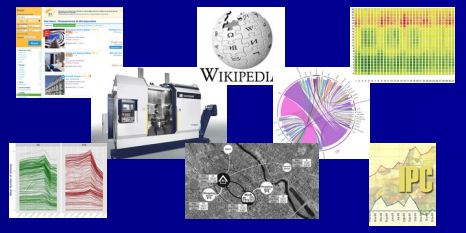
- Sensors in manufacturing machines

- Early fault detection
- Prediction of appropriate maintenance events
- Streaming visualization of sensors' values

DANOBAT
FAGOR
FAGOR INDUSTRIAL
Pasaban
LOIRE SAFE
ETXE-TAR
group

APPLICATIONS
- SUMMARY -

- WHERE are big data sources-torrents?
  - Internet → APIs [list of public APIs]
  - Internet → web scraping
    - Python “BeautifulSoup”, “Rvest”
  - Log analysis
  - Private companies
  - Sensors
  - Open government data portal...

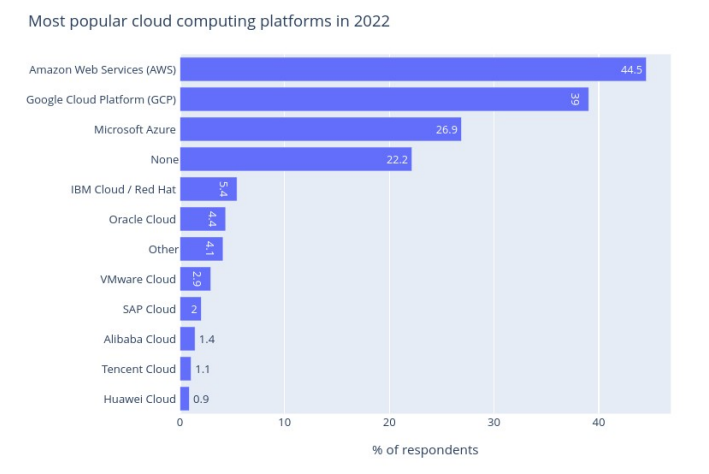


- TYPE of analysis:
  - Alternative calculation of indexes
  - Visualization: maps, evolution, time series...
  - When a classification problem exists: ML
  - Alternative correlations...

PLATFORMS FOR BIG DATA

- Free-software Initiatives → h2o AI, Apache Mahout, etc.

- Private services → AmazonWS, IBM Watson, Microsoft Traffic, Google Cloud AI...
- Machine and Deep Learning and Visualization as a Service (MLaaS)
- Storage as a Service
- Computing as a Service



BIG DATA: IS THE HYPE GOING?

- An established technology in companies
- Technology problem currently “solved”
- No “research current challenge”

- Challenge → on the application

- Is my application 'BigData'?
  - Depending on the type of data torrent-source

- Definition of technologies: afterwards