

Estadística oficial y big data

*II Jornadas Estadísticas Oficiales e Indicadores Sociales
y Económicos.* Madrid, 5 de abril de 2017



Índice

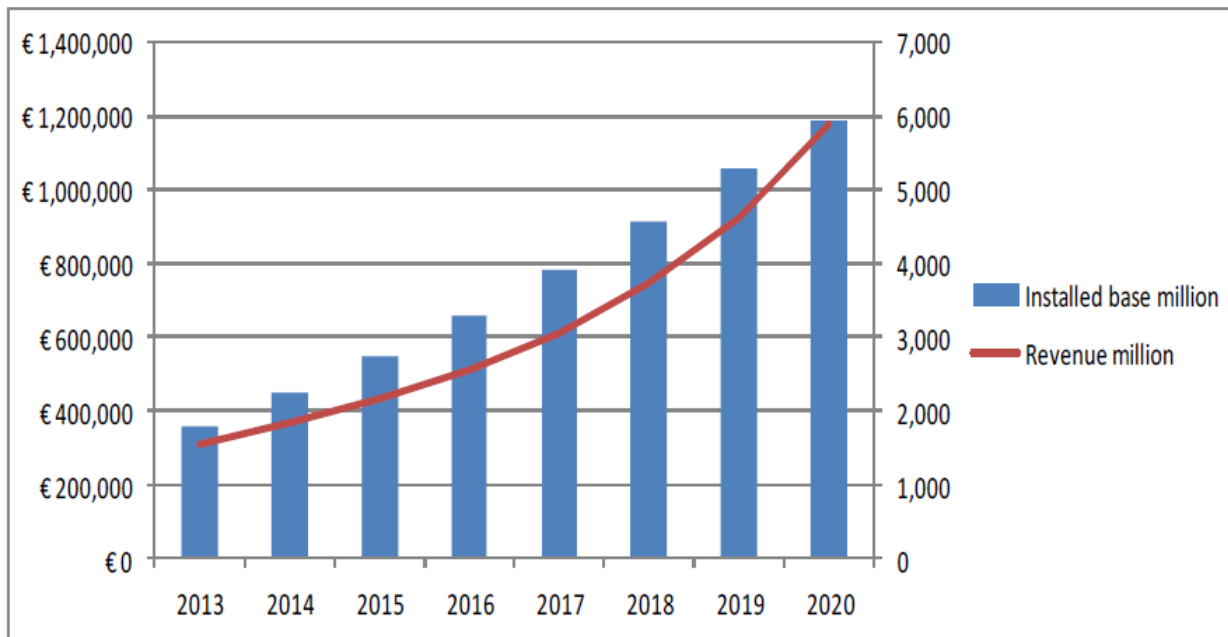
- 1.- ¿Necesidad de utilizar big data en la estadística oficial?
- 2.- Cuestiones a resolver
- 3.- El proyecto sobre big data del Sistema Estadístico Europeo
- 4.- Algunos ejemplos



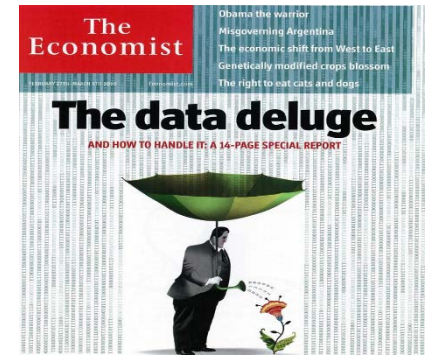
1.- ¿Necesidad de utilizar big data en la estadística oficial?

- Hay una ingente cantidad de información: ¿la despreciamos?

Figure 4 IoT Installed Base and Revenues in EU 28 2013-2018



Source: IDC 2014



The number of IoT connections within the EU28 will increase from approximately 1.8 billion in 2013 (the base year) to almost 6 billion in 2020

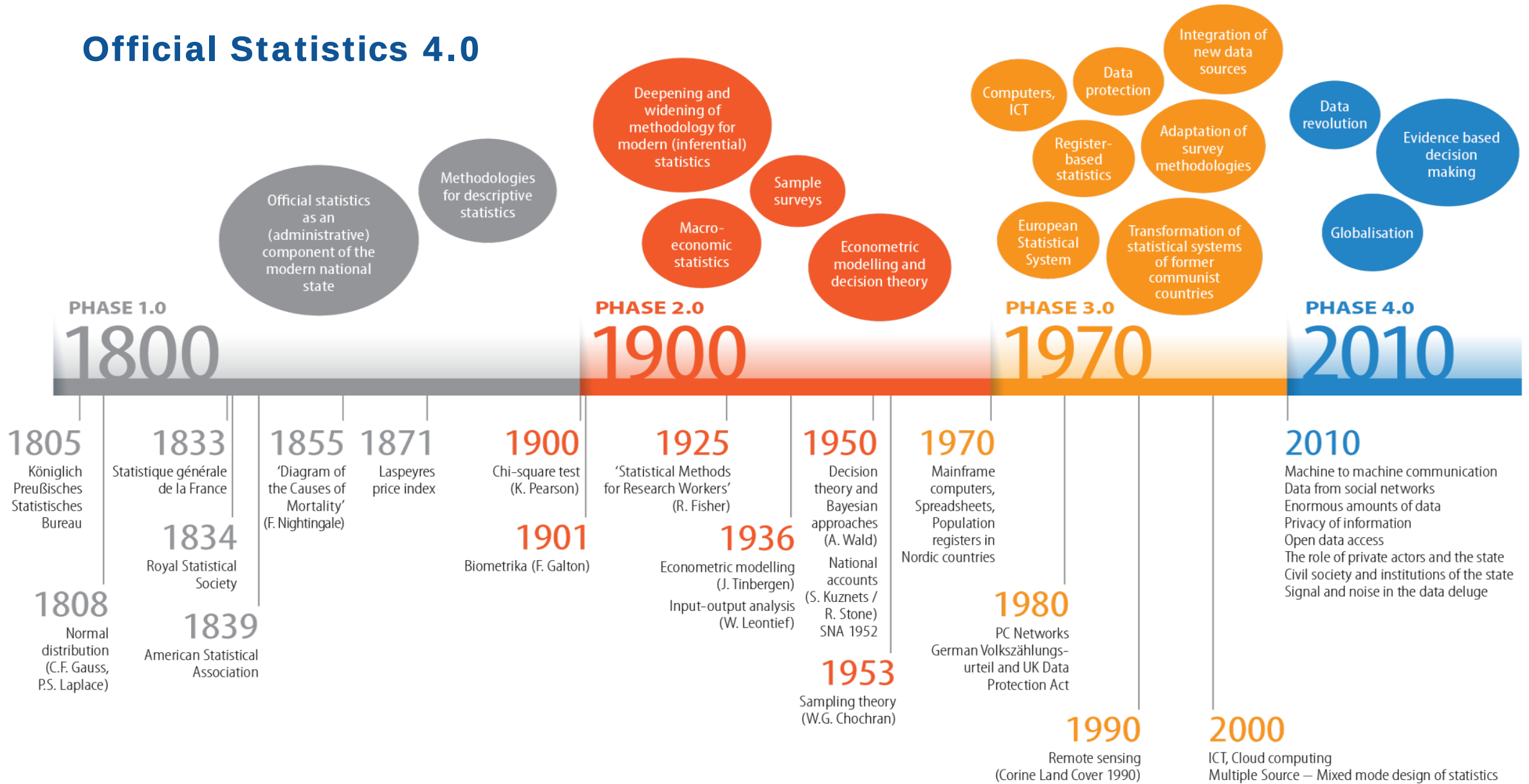
Innovation Policy Leveraging Cloud Computing and IoT Combination. A study prepared for the European Commission. DG Communications Networks, Content & Technology

1.- ¿Necesidad de utilizar big data en la estadística oficial?

- Reducir carga de trabajo a los informantes
- Mejorar nuestra eficiencia
 - Reducir costes
 - Mejorar la puntualidad
- Mejorar la información que producimos; generar nuevos productos: más datos quiere decir más conocimiento

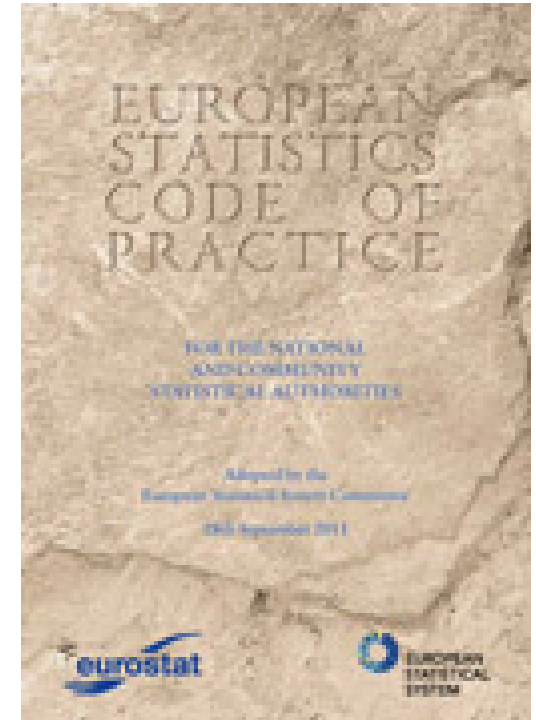


Official Statistics 4.0



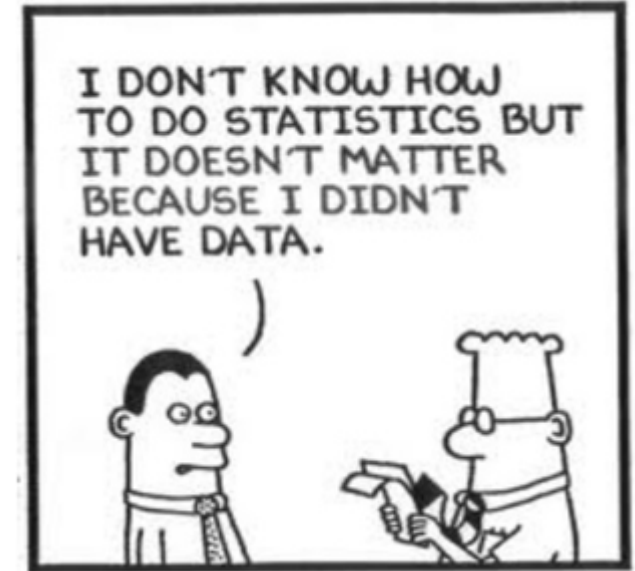
1.- ¿Necesidad de utilizar big data en la estadística oficial?

- Se dan las condiciones para que surjan productores de estadísticas relevantes y puntuales
- La estadística oficial aporta a la sociedad:
 - Independencia
 - Transparencia



2.- Cuestiones a resolver

- Acceso a las grandes acumulaciones de datos:
 - Legislación: Estadística/LOPD/ Telecomunicaciones
 - Preocupaciones de los propietarios de los datos
 - A nuestra competencia
 - A la competencia en general
 - A peticiones múltiples desde las Administraciones Públicas
 - Costes de extracción (datos brutos/datos pre-procesados) y lugar de procesamiento <http://www.opalproject.org/>
 - Repercusión reputacional de la cesión de datos



2.- Cuestiones a resolver

- Acceso a las grandes acumulaciones de datos:
 - Valor de los datos:

- *Este año se gastarán 150.800 millones de dólares en soluciones de análisis y Big Data. Para 2020 serán ya 210.000 millones.*

<http://www.silicon.es/gasto-analisis-big-data-2017-2331184>

- Bien social: <http://data-speaks.luca-d3.com/2016/10/the-6-challenges-of-big-data-for-social.html#more>

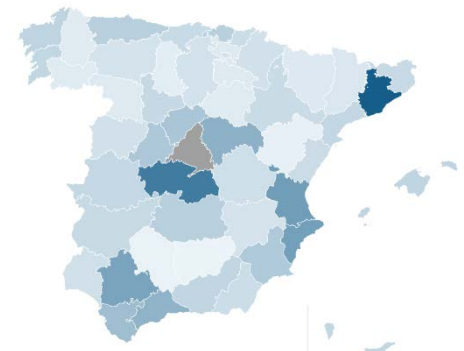
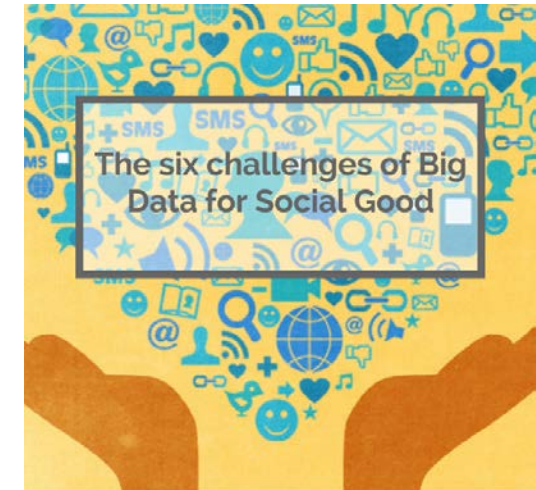


Figura 2: Distribución de los destinos elegidos por los habitantes de Madrid en el puente de diciembre.
<http://data-speaks.luca-d3.com/2016/12/donde-van-los-madrilenos-en-el-puente.html>



2.- Cuestiones a resolver

- Acceso a las grandes acumulaciones de datos:

¿Obligar o colaborar?



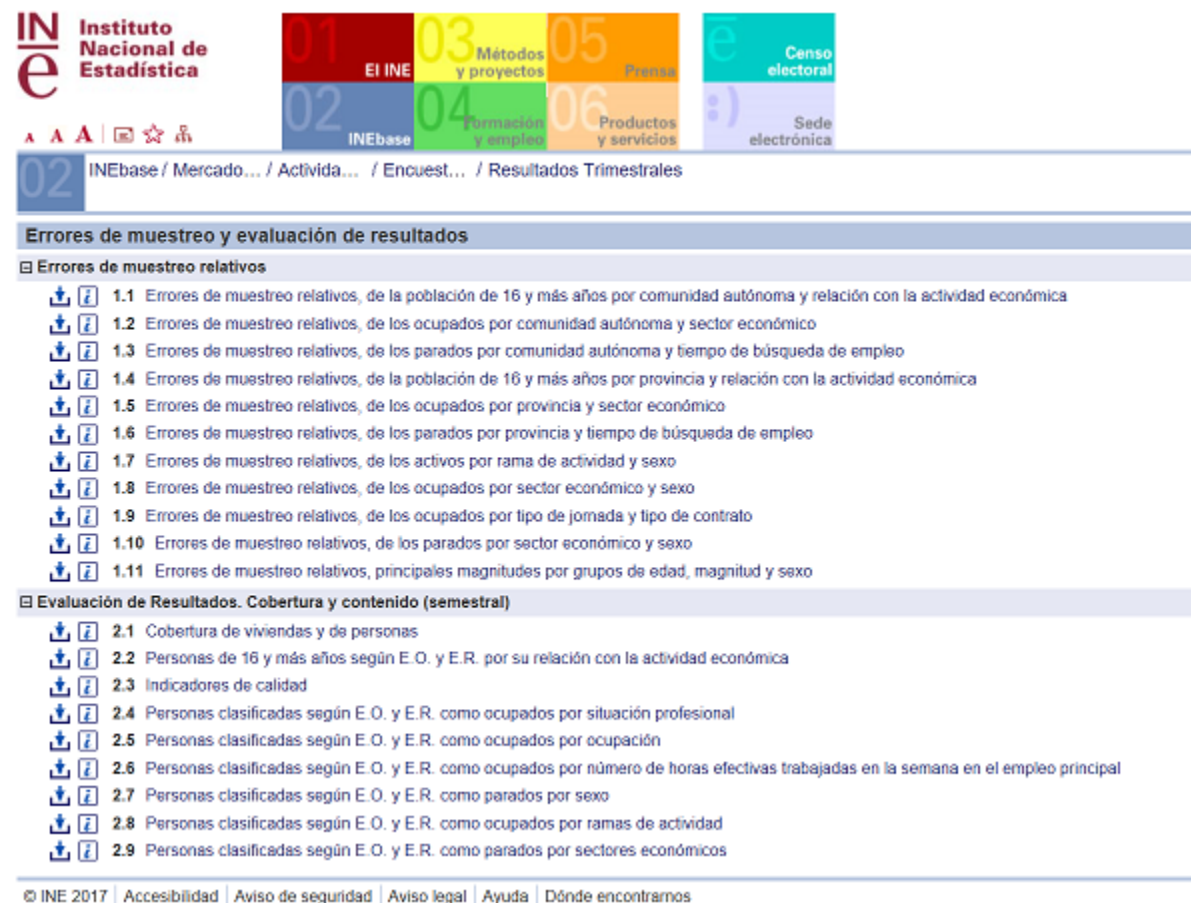
2.- Cuestiones a resolver

- Más temas a resolver
 - Seguridad/ Confidencialidad/ Ética/ Repercusiones reputacionales
 - Sostenibilidad de la producción a lo largo del tiempo
 - Nuevas competencias de los recursos humanos
 - Competir por esos recursos humanos cualificados
 - Infraestructura tecnológica



2.- Cuestiones a resolver

- Y sobre todo:
 - Metodología: De la inferencia clásica a bayesiana y modelos. Aprendizaje (no) supervisado; Algoritmos de reducción, de clasificación; Regresión; Visualización
 - Calidad



The screenshot shows the INE website interface. At the top, there's a navigation bar with the INE logo and a menu with items: 01 EI INE, 02 INEbase, 03 Métodos y proyectos, 04 Formación y empleo, 05 Prensa, 06 Productos y servicios, Censo electoral, and Sede electrónica. Below the navigation bar, there's a breadcrumb trail: 02 INEbase / Mercado... / Activida... / Encuest... / Resultados Trimestrales. The main content area is titled 'Errores de muestreo y evaluación de resultados'. It contains two sections: 'Errores de muestreo relativos' and 'Evaluación de Resultados. Cobertura y contenido (semestral)'. Each section lists 11 items, each with a download icon and a question mark icon.

Errores de muestreo relativos

- 1.1 Errores de muestreo relativos, de la población de 16 y más años por comunidad autónoma y relación con la actividad económica
- 1.2 Errores de muestreo relativos, de los ocupados por comunidad autónoma y sector económico
- 1.3 Errores de muestreo relativos, de los parados por comunidad autónoma y tiempo de búsqueda de empleo
- 1.4 Errores de muestreo relativos, de la población de 16 y más años por provincia y relación con la actividad económica
- 1.5 Errores de muestreo relativos, de los ocupados por provincia y sector económico
- 1.6 Errores de muestreo relativos, de los parados por provincia y tiempo de búsqueda de empleo
- 1.7 Errores de muestreo relativos, de los activos por rama de actividad y sexo
- 1.8 Errores de muestreo relativos, de los ocupados por sector económico y sexo
- 1.9 Errores de muestreo relativos, de los ocupados por tipo de jornada y tipo de contrato
- 1.10 Errores de muestreo relativos, de los parados por sector económico y sexo
- 1.11 Errores de muestreo relativos, principales magnitudes por grupos de edad, magnitud y sexo

Evaluación de Resultados. Cobertura y contenido (semestral)

- 2.1 Cobertura de viviendas y de personas
- 2.2 Personas de 16 y más años según E.O. y E.R. por su relación con la actividad económica
- 2.3 Indicadores de calidad
- 2.4 Personas clasificadas según E.O. y E.R. como ocupados por situación profesional
- 2.5 Personas clasificadas según E.O. y E.R. como ocupados por ocupación
- 2.6 Personas clasificadas según E.O. y E.R. como ocupados por número de horas efectivas trabajadas en la semana en el empleo principal
- 2.7 Personas clasificadas según E.O. y E.R. como parados por sexo
- 2.8 Personas clasificadas según E.O. y E.R. como ocupados por ramas de actividad
- 2.9 Personas clasificadas según E.O. y E.R. como parados por sectores económicos

© INE 2017 | Accesibilidad | Aviso de seguridad | Aviso legal | Ayuda | Dónde encontramos

2.- El proyecto sobre big data del Sistema Estadístico Europeo

- Objetivos:
 - Identificación y análisis del portfolio de fuentes de big data
 - La cuestión del acceso a las fuentes de big data
 - Identificación de nuevas competencias profesionales
 - Revisión de los marcos metodológicos y de calidad
 - Análisis de de la infraestructura TIC



3.- El proyecto de big data del Sistema Estadístico Europeo

- Estrategia:
 - Empezar por lo particular, mostrar primeros resultados y luego pasar a lo general
 - Pilotos por fuente de información usable en múltiples países
 - Subproyectos horizontales: generalizar las conclusiones de los pilotos, revisar y enriquecer la literatura existente, elaborar recomendaciones a los INEs



3.- El proyecto de big data del Sistema Estadístico Europeo

- Organización del trabajo:
 - WP1/2: Proyectos piloto sobre el uso de websites (web scraping): ofertas de trabajo y características de empresas
 - WP3: Proyectos piloto sobre el uso de “contadores inteligentes”
 - WP4: Proyectos piloto sobre sistemas de identificación de buques
 - WP5: Proyectos piloto sobre el uso de datos de posicionamiento de móviles
 - WP6/7: Integración de datos, (combinación de diversas fuentes de Big Data)
 - WP8: Metodología y temas horizontales
 - WP9: Comunicación y difusión



4.- Algunos ejemplos

- Holanda: <https://www.cbs.nl/en-gb/our-services/innovation/project/social-tension-indicator-based-on-social-media>
- Eurostat: <http://ec.europa.eu/eurostat/documents/3888793/7753027/KS-TC-16-024-EN-N.pdf/50a4e9ea-d88e-4d59-bf37-2c6200f757b0>
- Canadá: <http://www.statcan.gc.ca/eng/crowdsourcing?HPA=1>



Gracias por la atención

miguelangel.martinez.vidal@ine.es
Departamento de Metodología y
Desarrollo de la Producción
Instituto Nacional de Estadística

