# Introducción a la Ciencia de los Datos

Una breve aproximación a Data Science

# ¿Qué es Data Science? (I)

- No existe una definición de consenso, sino que difiere según las fuentes.
- Podemos decir que trata del estudio de la **extracción generalizada de conocimiento a partir de información, de datos**.

- ¿Esto es algo nuevo? ¿No se parece a alguna ciencia con la que ya estamos familiarizados?

# Data Science y Estadística (I)

- La Estadística consiste en el estudio de la recolección, **análisis**, **interpretación**, **presentación** y organización de datos.
- La Ciencia de Datos trata del estudio de la **extracción generalizada de conocimiento a partir de información, de datos**.

- ¿Estamos hablando de lo mismo? Veamos la opinión de Jeff Wu, de la University of Michigan:
  - http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

# Data Science y Estadística (II)

- Tras consultar varias opiniones y diversas fuentes, podemos llegar a la conclusión de que existen diferencias:
  - El **enfoque** de Data Science es más **holístico**, más global, para partiendo de grandes volúmenes de datos poder extraer conocimiento que aporte valor a una determinada organización del tipo que sea.
  - El foco principal se sitúa en la **extracción de conocimiento**, empleando para ello las herramientas que estén al alcance.
- Veamos en qué se traduce lo anterior, mediante una definición más completa.

# ¿Qué es Data Science? (II)

- Ya hemos podido intuir que se trata de algo más que la Estadística. Veamos qué más… Wikipedia recopila muchos de los principals campos implicados en Data Science, indicando que emplea:
  - Técnicas y teorías de muchos campos dentro de amplias áreas como la **Matemática**, la **Estadística** y las **Tecnologías de la Información**, incluyendo: procesamiento de señales, modelos probabilísticos, machine learning, aprendizaje estadístico, programación, ingeniería de datos, reconocimiento de patrones, visualización, modelización de la incertidumbre, data warehousing, and computación de altas prestaciones.
- Podemos encontrar un análisis de algunas fuentes relacionadas en este enlace:
  - http://www.sorayapaniagua.com/2011/11/01/la-ciencia-de-los-datos-bdii/

# ¿Por qué ahora?

- Con la emergencia los últimos años del **Big Data**, existe una disponibilidad enorme de datos tanto a nivel de Internet como en las organizaciones.
- Además, existe una importante apuesta por los datos abiertos, Open Data, y muchos organismos se están adhiriendo a esta iniciativa, como podemos leer en este enlace: http://www.sorayapaniagua.com/2011/11/01/big-data-o-la-nueva-ciencia-de-los-datos-bgi/.
- Gobiernos y compañías han puesto énfasis en el valor de la cantidad de datos disponibles y la posibilidad de extraer conocimiento de ellos, relevante para su operativa y la ayuda a una mejor toma de decisiones.

# Sitios de interés

- Common Crawl, que pone a disposición un gran número de datos de miles de millones de webs:
  - http://www.sorayapaniagua.com/2013/01/28/common-crawl-datos-gratuitos-de-cinco-mil-millones-de-paginas-web/
- Un listado de lugares en los que podemos encontrar datos abiertos:
  - http://blog.visual.ly/data-sources/
- Kaggle, un sitio interesante en el que existen competiciones públicas extracción de conocimiento y predicción a partir de datos:
  - http://www.kaggle.com/

# El científico de datos (data scientist)

- Hoy día un perfil muy demandado a nivel internacional es el de Data Scientist o científico de datos (http://www.norbertogallego.com/data-scientist-empleo-indefinido-sin-reforma/2014/03/05/), capaz de estudiar las diversas fuentes de información disponibles en una organización, extraer datos a partir de diversos formatos tanto de bases de datos relacionales y no relacionales como de muchos otros tipos, depurarlos, analizarlos, idear y desarrollar algoritmos, en algunos casos paralelos, realizar inferencias, preparar y comunicar los resultados de dichos análisis y ser capaz de transmitir conclusiones acerca de los estudios para finalmente repercutir en un mayor conocimiento que ayude a la Gerencia del organismo o compañía a tomar mejores decisiones.
- Este perfil fue catalogado en octubre de 2012 como el trabajo más sexy del siglo 21:
  - https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1

# Nuestra caja de herramientas

- Muchos de los conocimientos tienen que ver con la Matemática y la Estadística, pero además se precisa de diversos conocimientos tecnológicos:
  - Bases de datos relacionales, SQL
  - Bases de datos no relacionales, Big Data, NoSQL…
  - Lenguajes de programación: R, Python
  - Machine Learning
  - Programación de altas prestaciones, programación distribuida, Hadoop

# Una oferta de trabajo actual…

estamos buscando Data Scientists con experiencia en análisis de datos, mineria de datos, cálculo, algoritmos, etc.

**Habilidades**: Imprescindible conocimientos de estadística y experiencia en tecnologías Big Data como SPSS, R, Machine Learning, etc.

**Estudios mínimos**:

Licenciatura o ingeniería de especialidad científica o tecnológica: Ingeniería, matemáticas, estadística, etc.

Nivel Alto de Inglés, se valorará muy positivamente francés y alemán

**Requisitos mínimos**:

- Amplia experiencia en la resolución de problemas analíticos mediante enfoques cuantitativos
- Amplia experiencia en data mining y Machine Learning
- Experiencia en uso de algoritmos estadísticos (K-Means Clustering, Decission Trees, Random Forest, Naive Bayes, Lineal Regression, etc)
- Experiencia en manipulación de grandes volúmenes de datos, así como enanálisis de datos complejos desde diferentes fuentes
- Conocimiento experto de herramientas de análisis estadístico con R en entornos Big Data, así como lenguajes de programación como Python
- Experiencia con tecnologías Big Data (MapReduce, Hadoop, Hive, MongoDB, Cassandra, etc), Bases de Datos y SQL
- Fuerte pasión por la investigación empírica y por responder a preguntas difíciles con los datos
- Enfoque analítico flexible que permite obtener resultados a diferentes niveles de precisión.

# Algunos recursos disponibles

- Curso MOOC gratuito de especialización en Data Science. En realidad se trata de una especialización que consta de 9 cursos: http://aspgems.com/blog/fernando-calle/cursos-gratuitos-de-especializacion-en-data-science
- Recopilación de cursos presenciales y online: https://josejuliolopezsantos.wordpress.com/2014/06/14/los-recursos-mas-destacados-para-la-formacion-en-big-data-en-espana-y-online/comment-page-1/
- Libro gratuito "Big Data Now": http://www.oreilly.com/data/free/files/bigdatanow2013.pdf