# Ciencia de Datos: Un Enfoque Práctico en la Era del Big Data

## Introducción

Ciencia de Datos es un área de trabajo interdisciplinar que incluye procesos para recopilar, preparar, analizar, visualizar y modelizar datos que permitan generar conocimiento útil para comprender problemas complejos y ayudar en la toma de decisiones. Estos datos con frecuencia son no estructurados y heterogéneos. En muchas ocasiones, se trata de grandes volúmenes de datos que por su complejidad y diversidad requiere de arquitecturas y técnicas innovadoras para extraer conocimiento relevante: es el conocido *big data*. Ciencia de Datos es un campo emergente con una alta aplicabilidad en ciencias de la salud, marketing, negocios, mercados financieros, transporte, comunicaciones, redes sociales, etc.

Como indica la consultora Gartner (la más prestigiosa en tecnologías de la información), los científicos de datos no son analistas de negocio tradicionales, son profesionales con la rara capacidad para obtener modelos matemáticos a partir de datos que generan beneficios empresariales claros y contundentes. Así, cada vez más se exigen profesionales con habilidades en campos como informática, matemáticas, estadística o negocios que dominen las nuevas tecnologías y sepan gestionar datos. Las empresas de todos los sectores están adoptando cada vez más la Ciencia de Datos, de modo que la demanda de expertos en este sector es enorme; así lo refleja un estudio del MIT Sloan Management Review (2015). Está considerada como una de las mejores oportunidades laborales de los próximos años. Catalogada por el Harvard Business Review como la profesión 'más sexy del siglo XXI' (2012). Según un estudio de LinkedIn (2015), el número de profesionales en Ciencia de Datos se ha duplicado en los últimos cuatro años. Otro estudio de Burtch Works (2015) reconoce el impacto positivo en el salario al incluir conocimientos de Ciencia de Datos.

## Objetivos

Los planes de formación universitaria reglada difícilmente reaccionan a oportunidades laborales emergentes. Además, se tiende a delimitar fronteras que dificultan el desarrollo de especialidades híbridas. Este curso pretende iniciar al alumno en el campo de Ciencia de Datos, sirviendo así de puente entre diversas disciplinas y ayudando a completar la formación universitaria con una orientación eminentemente práctica. El curso se compone de 30 horas lectivas presenciales repartidas en 15 horas de conceptos teóricos y fundamentos y otras 15 horas de prácticas con *software* especializado y datos de casos reales.

La teoría incluye visualización de datos, técnicas de clasificación básicas (árboles de decisión, redes neuronales…) y avanzadas (máquinas de soporte vectorial, *ensemble learning, deep learning*…), preprocesado (eliminación de ruido, imputación de valores perdidos, reducción de datos…), aprendizaje no supervisado (agrupamiento y reglas de asociación), aprendizaje incremental y minería de flujo de datos, *big data* y sus paradigmas y, finalmente, experiencias reales de Ciencia de Datos en la empresa. La práctica introduce al alumno en herramientas *software* tales como KNIME y R y arquitecturas como Hadoop y Spark. También se adquirirá experiencia en la plataforma Kaggle para competiciones en problemas reales.

## A quién va dirigido

Las personas que se dedican a la Ciencia de Datos se conocen como científicos de datos, que no es más que una mezcla de matemáticos, estadísticos, informáticos y creativos con habilidades para recopilar, procesar y extraer valor de las diversas y extensas bases de datos; imaginación para comprender, visualizar y comunicar sus conclusiones a los no científicos de datos; y capacidad para crear soluciones basadas en datos que aumentan los beneficios, reducen los costos y ayudan a construir un mundo mejor.

El curso se orienta a estudiantes de grado, máster y profesionales con formación previa principalmente en informática, matemáticas, estadística, física, ingeniería o empresariales que busquen completar su formación como científico de datos. La presentación de los fundamentos teóricos y el uso de *software* especializado se impartirán de forma apropiada para atender a las diferentes necesidades del alumnado. Ciencia de Datos es una disciplina que se nutre de experiencias y formaciones diversas, de forma que el curso aprovechará la variedad de necesidades y capacidades del alumnado.

## Equipo docente

El profesorado lo componen docentes e investigadores universitarios senior y jóvenes del área de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada y de Jaén. Se trata de personal altamente especializado en Ciencia de Datos con excelentes trayectorias en investigación. El área de Ciencias de la Computación de la Universidad de Granada está considerada según el prestigioso ranking ARWU 2015 de Shanghái como la 42 mejor del mundo, sexta de Europa y primera de España. En esta misma área, la Universidad de Jaén se encuentra en el puesto 51-75, siendo la segunda mejor de España.

Jorge Casillas (Universidad de Granada) - coordinador
Jesús Alcalá (Universidad de Granada)
Francisco Charte (Universidad de Jaén)
Alberto Fernández (Universidad de Jaén)
Salvador García (Universidad de Granada)
Sara del Río (Universidad de Granada)

## Contenidos de Teoría (15h)

1. Ciencia de Datos, analítica avanzada y *big data* (1h) – **Jorge Casillas**
2. Análisis exploratorio de datos: visualización (1h) – **Jorge Casillas**
3. Fundamentos de clasificación: árboles de decisión, *lazy*, RNA, bayesianos, evaluación (2h) – **Salvador García**
4. Preprocesamiento: selección y procesado de instancias y características, tratamiento del ruido (2h) – **Salvador García**
5. Clasificación avanzada: problemas no balanceados, SVM, *ensemble learning*, *deep learning* (2h) – **Alberto Fernández**
6. Segmentación y relaciones: *clustering* y reglas de asociación (2h) – **Jorge Casillas**
7. Aprendizaje incremental y *data stream mining* (1h) – **Jorge Casillas**
8. *Big data*: fundamentos y paradigmas (2h) – **Alberto Fernández**
9. Ciencia de Datos en acción: experiencias de empresa (2h) – ***Invitado***

**Contenidos de Prácticas** (15h)

1. KNIME (5h): predicción fundamental – **Jesús Alcalá**
2. R para Ciencia de Datos (ggplo2, caret, rattle, neuralnet, e1071, randomForest, gbm, h2o, autoencoder, SAENET…) (5h): visualización y predicción avanzada – **Francisco Charte**
3. Hadoop + Mahout, Spark + MLLib (5h): big data – **Sara del Río**

**Asistencia**: obligatorio asistir al 80% de las clases

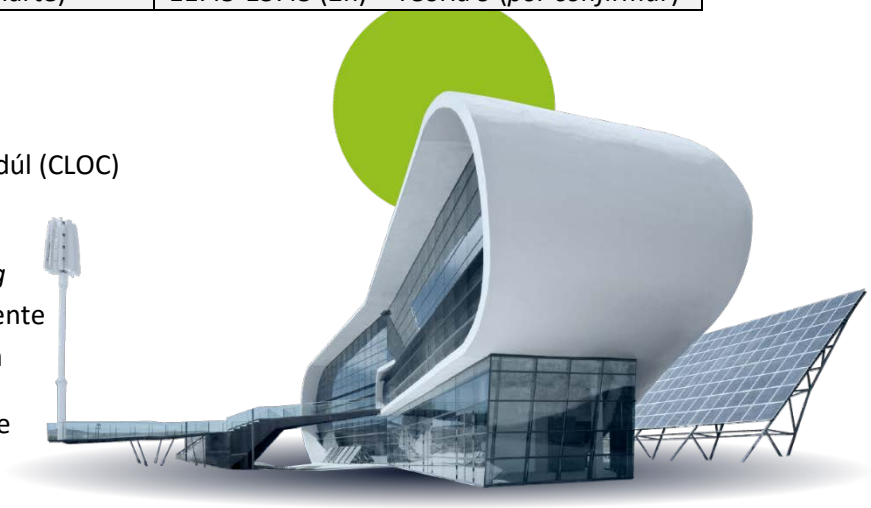
**Evaluación**: Elaboración de resúmenes de los contenidos y competición en Kaggle (http://www.kaggle.com) en equipos interdisciplinares.

**Planificación**:

| | |
|---|---|
| **Sesión 1 - sábado, 05/03/2016**<br>08:45-09:00 – Presentación del curso<br>09:00-10:00 (1h) – Teoría 1 (J. Casillas)<br>10:05-11:00 (1h) – Teoría 2 (J. Casillas)<br>11:05-12:00 (1h) – Teoría 3 (S. García)<br>12:15-14:15 (2h) – Prácticas 1 (J. Alcalá) | **Sesión 4 - sábado, 02/04/2016**<br>09:00-11:00 (2h) – Teoría 6 (J. Casillas)<br>11:05-12:00 (1h) – Teoría 7 (J. Casillas)<br>12:15-14:15 (2h) – Prácticas 2 (F. Charte) |
| **Sesión 2 - sábado, 12/03/2016**<br>08:30-09:30 (1h) – Teoría 3 (S. García)<br>09:35-11:30 (2h) – Teoría 4 (S. García)<br>11:45-14:45 (3h) – Prácticas 1 (J. Alcalá) | **Sesión 5 - sábado, 09/04/2016**<br>09:00-11:00 (2h) – Teoría 8 (A. Fernández)<br>11:15-14:15 (3h) – Prácticas 3 (S. del Río) |
| **Sesión 3 - sábado, 19/03/2016**<br>09:00-11:00 (2h) – Teoría 5 (A. Fernández)<br>11:15-14:15 (3h) – Prácticas 2 (F. Charte) | **Sesión 6 - sábado, 16/04/2016**<br>09:30-11:30 (2h) – Prácticas 3 (S. del Río)<br>11:45-13:45 (2h) – Teoría 9 (*por confirmar*) |

**Lugar de celebración:**

Centro de la Construcción Sostenible de Padúl (CLOC)
http://www.cloc.es



Vivero de empresas, formación y *coworking* en un edificio vanguardista altamente eficiente ubicado en Padúl, a 20 minutos de Granada

Las prácticas se desarrollarán en una sala de 25 ordenadores de última generación

Ubicación: https://goo.gl/maps/ZoqN63ZnqJt
Horario de autobuses: http://lahuertadelcura.com/wp-content/uploads/LINEA_360.pdf

**Número de plazas:** 25

**Precio:** 100 € (10% de becas)

**Más información:** Centro Mediterráneo: http://www.ugr.es/~cm/accesos/p1.html

Pendiente de reconocimiento de 3 créditos ECTS optativos para Grados en Ing. Informática, Ing. de Telecomunicaciones, Matemáticas, Estadística, Física, Ing. Electrónica Industrial, CC Económicas y Empresariales…