Optare Solutions
NATURAL BORN TELCO

Máster en Técnicas Estadísticas
10 AÑOS

Data Science: del Data Mining al Deep Learning

Irene Castro Conde[1,2]
icastro@optaresolutions.com

[1]Optare Solutions S.L.
[2]Grupo SiDOR, Universidad de Vigo

I Jornada de Orientación Profesional del MTE

23 Junio 2016

Esquema

Esquema

Mi perfil

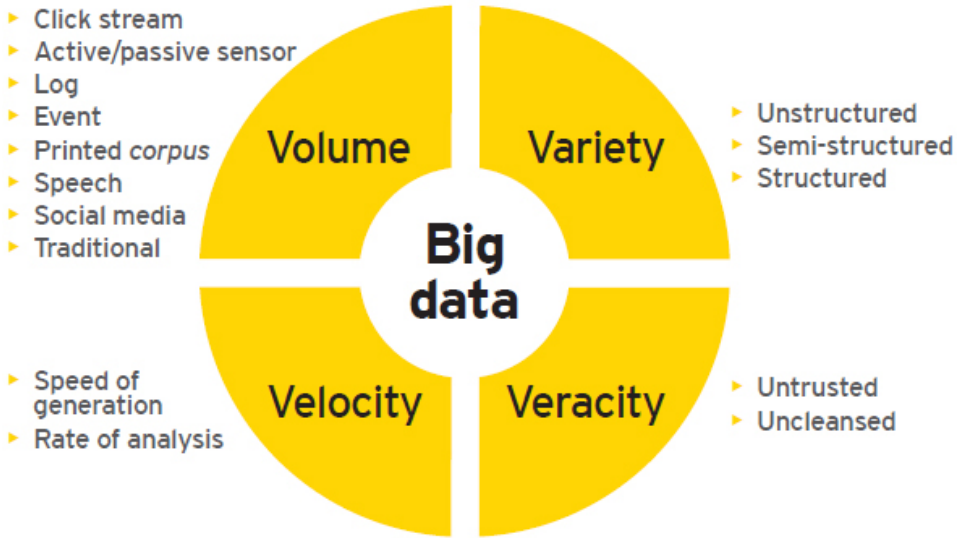
- 2006-2011 Licenciatura en **Matemáticas**: Especialidad en Estadística e IO
- 2011-2013 **Máster** en Técnicas Estadísticas: TFM académico (Modalidad A)
- 2012-2014 **Doctorado** en Estadística e IO (Lectura tesis 18/12/2014)
- 01/2015- **Optare Solutions**
  - Proyecto de Investigación Optare Solutions - UVigo
  - Contrato Torres Quevedo

Optare Solutions

- **Consultora tecnológica** especializada en el desarrollo y la integración de sistemas de soporte a las operaciones (OSS) y al negocio (BSS) para operadoras de telecomunicaciones.
- Nació en el año **2002** con el objetivo de proporcionar consultoría técnica para la provisión de servicios complejos a una operadora emergente del mercado español.
- Formada por más de 50 personas, con sede en Vigo (Parque Tecnológico y Logístico de Valladares) y filial en México DF, más del 80 % de titulados superiores en informática y telecomunicaciones, **2 estadísticos**.
- Más del 15 % de la facturación dedicada a la realización de proyectos **I+D** que incluyen proyectos regionales (Incite, Peme I+D), nacionales (Fondo Tecnológico, Innterconecta, AEESD) y europeos (FP7, H2020).
- **Proyecto BDA4T** (Big Data Analytics for Telecoms): proyecto de I+D en Cooperación Nacional, financiado por el CDTI (Centro para el Desarrollo Tecnológico Industrial) y que cuenta con el apoyo de la Universidad de Vigo como organismo de investigación.

Motivación del Proyecto BDA4T

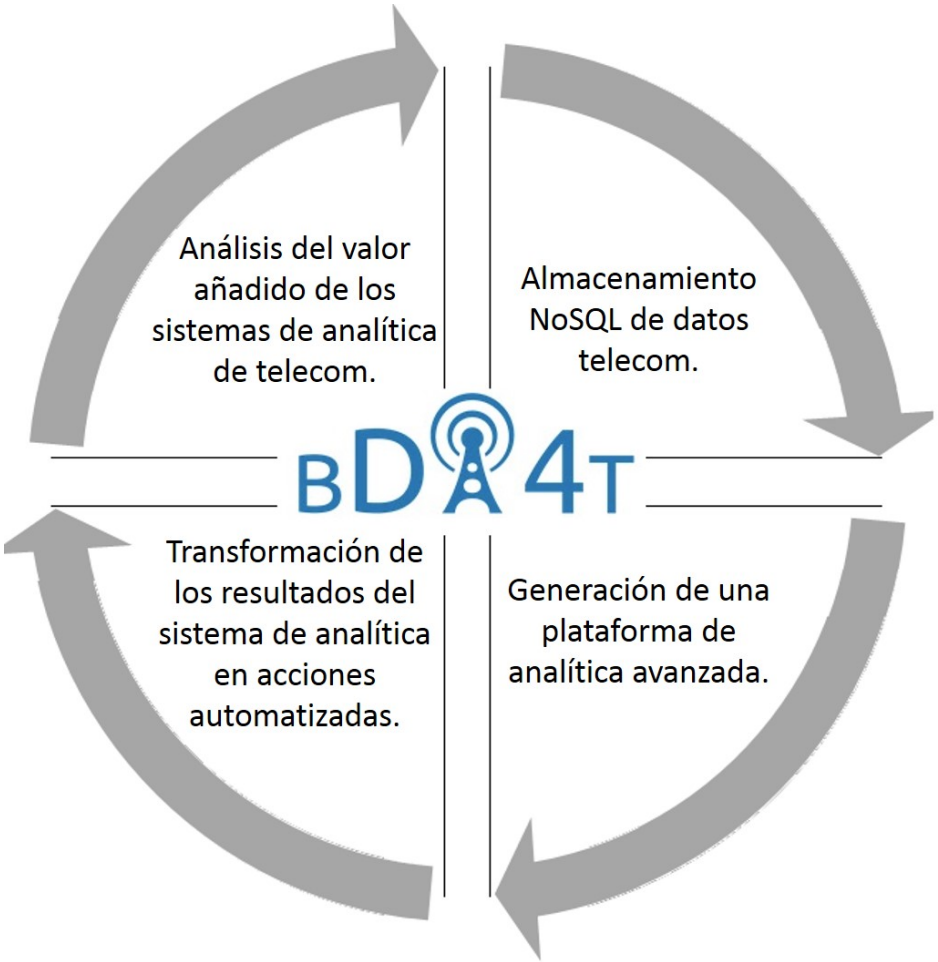
- Los operadores disponen de grandes volúmenes de datos que no aprovechan completamente que provienen de:



1. Ámbito de cliente
2. Ámbito de red
3. Ámbito TIC

⟵ **Esos datos son: BIG DATA**

- Los mercados maduros necesitan el análisis avanzado de los datos disponibles para incrementar el valor de la información obtenida: incrementar los ingresos, gestionar la base de clientes, optimizar los procesos y generar nuevos modelos de negocio.



- **Objetivo:** desarrollo de un sistema de análisis predictivo de bajas en los operadores de telecomunicaciones, con datos proporcionados por sus sistemas de gestión relacionados con el negocio e interacción con el cliente.

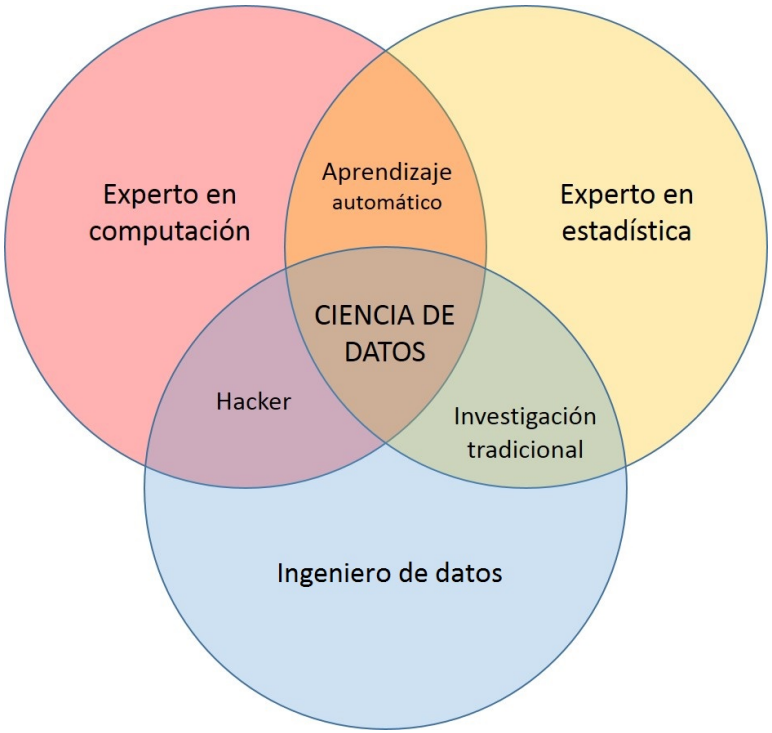
- **Casos de Uso:**

  - **Customer Lifetime Value**: estimar el Tiempo de Vida y los Ingresos futuros de los clientes que permitan conocer su valor.

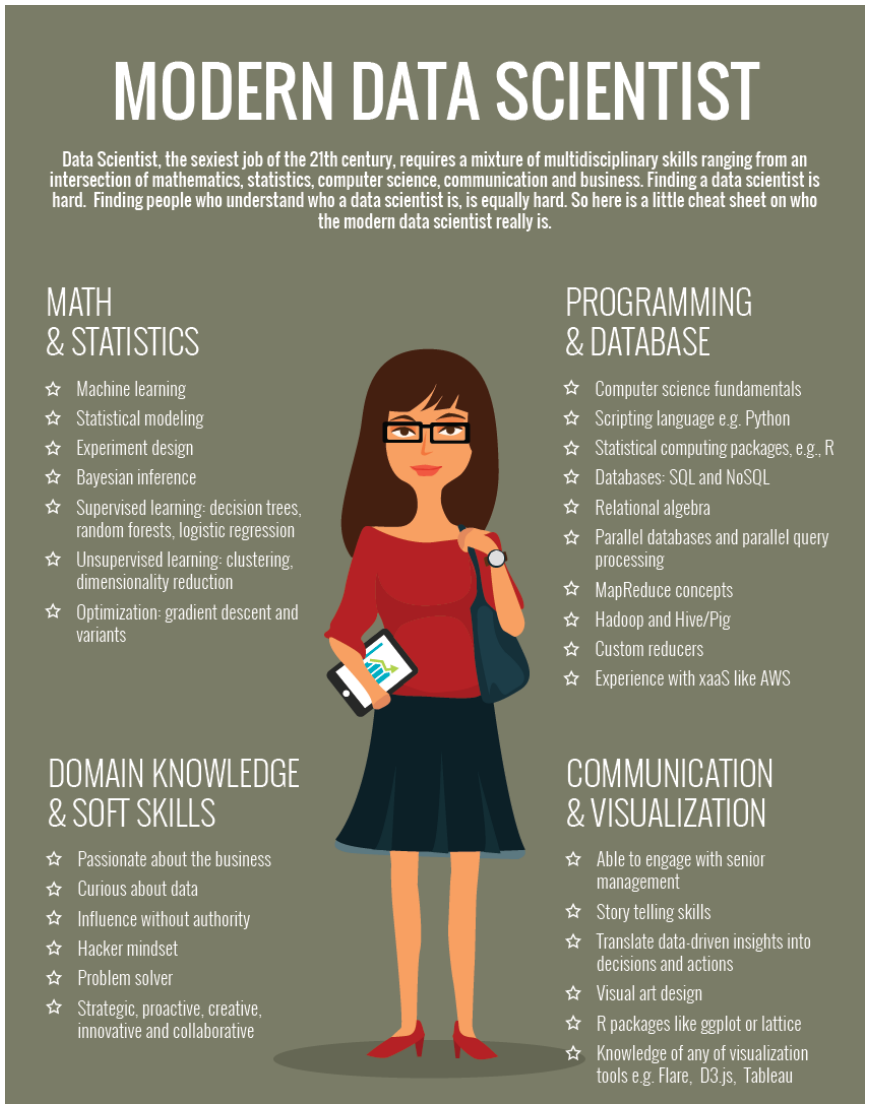
  - **Modelado predictivo de bajas**: predicción de la probabilidad de baja de un cliente según su perfil comercial permitiendo a los operadores centrar sus esfuerzos de fidelización en los clientes con mayor probabilidad de abandonar la compañía.

Esquema

Data Science: la Ciencia de los Datos



La **Ciencia de Datos** es un campo interdisciplinar que involucra los procesos y sistemas para extraer conocimiento o un mejor entendimiento de grandes volúmenes de datos en sus diferentes formatos.



Data Mining

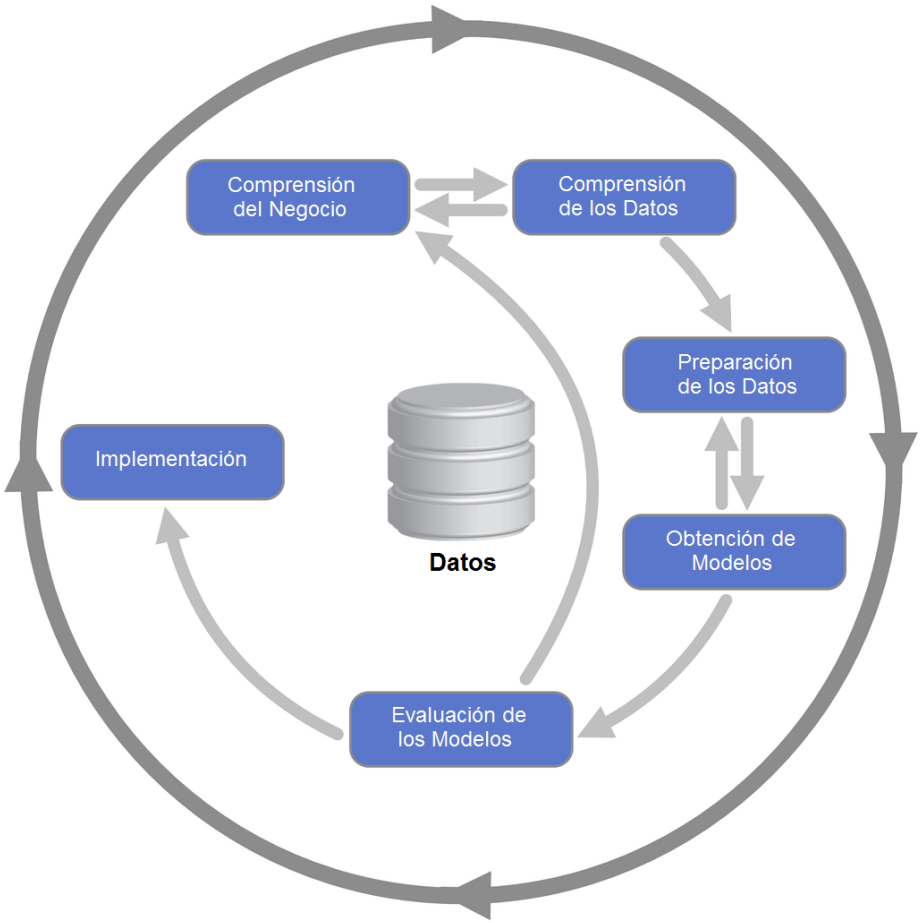
Definición: campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

Utiliza los **métodos** de inteligencia artificial, aprendizaje automático, **estadística** y sistemas de bases de datos, ...

Objetivo: el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos usando técnicas de analítica para optimizar la toma de decisiones:

- **Inteligencia de Negocios:** proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos.

El modelo CRISP-DM
(Cross Industry Standard Process for Data Mining)



- Modelo del proceso de minería de datos, que describe el enfoque utilizado comúnmente.
- Consta de 6 fases.
- Las flechas indican las dependencias más frecuentes entre fases.
- El círculo exterior simboliza la naturaleza cíclica de la minería de datos: continúa después del despliegue porque las lecciones aprendidas pueden provocar nuevas preguntas de negocio, más centradas, y posteriores procesos de minería de datos se beneficiarán de la experiencia de los anteriores.

Analítica Predictiva

Definición: aplicación de técnicas estadísticas avanzadas para predecir el comportamiento futuro teniendo en cuenta lo que ocurrió en el pasado. También se conoce como Aprendizaje Supervisado.

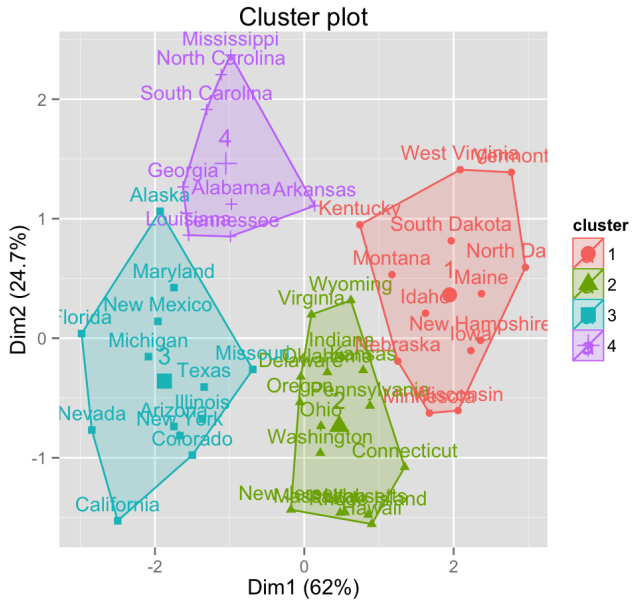
Diferentes Tipos de Técnicas

- **Clasificación:** predecir una variable discreta basándose en otros atributos del conjunto de datos. Ej.: Churn: baja/no baja. Campañas: compra/no compra, ...
- **Regresión:** predecir una variable continua basándose en otros atributos del conjunto de datos. Ej.: nº de bajas, beneficios, ...
- **Series de Tiempo:** predecir valores futuros de una variable temporal. Ej.: nº reproducciones diarias de un servicio de TV a la carta, ...

Materias del Máster: Series de Tiempo, Modelos de Regresión

Aprendizaje No Supervisado

Definición: consiste en ajustar un modelo a los datos sin conocimiento a priori.



Diferentes Tipos de Técnicas:

- **Segmentación (Análisis Cluster):** dividir los datos en grupos de elementos que tienen propiedades similares entre sí y diferentes a las de los otros grupos. Ej.: clientes con un perfil de consumo similar, ...
- **Detección de Anomalías:** identificar casos inusuales para investigarlos más a fondo (casos interesantes, errores). Ej.: fraude, caídas sistema, ...
- **Minería de Texto:** obtener información interesante de texto plano, como patrones que se repiten o Análisis de Sentimientos. Ej.: usar las redes sociales para conocer la satisfacción de los clientes, ...

Materias del Máster: Análisis Exploratorio de Datos, Análisis Multivariante

Deep Learning

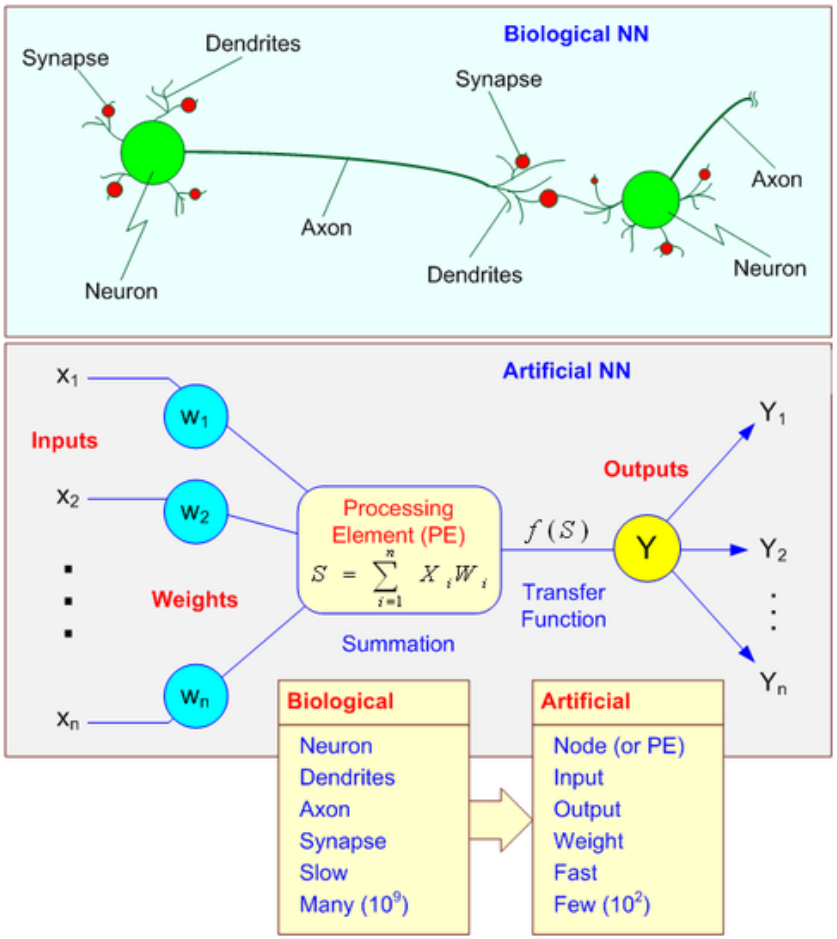
Qué es?

Definición: Conjunto de algoritmos que intentan modelar abstracciones de alto nivel en datos usando arquitecturas compuestas de varias capas de transformaciones no-lineales. Las arquitecturas más extendidas son las redes neuronales profundas (Deep Neural Networks).

Ventajas

- Modelo simple conceptualmente.
- Modelo no lineal altamente flexible.
- Eficiente para múltiples tipos de problemas.

La teoría de **redes neuronales** está inspirada en la estructura y funcionamiento de los sistemas nerviosos, donde la neurona es el elemento fundamental.



Deep Learning

Casos de Uso

- **Búsqueda por voz:** Siri de Apple (2011) es el más conocido, Google Now, el asistente por voz para Android, o Microsoft Cortana introducido en abril de 2014 en Windows Phone 8.1.

- **Recomendadores:** empresas como Netflix, Amazon, Google, Facebook y Twitter tienen acceso a una enorme cantidad de datos generados por los usuarios que les ha permitido implementar sistemas de recomendación que se han vuelto mucho más inteligentes gracias a la utilización del aprendizaje profundo para predecir las preferencias del usuario y proporcionarle recomendaciones precisas.

- **Reconocimiento de Imágenes:** El objetivo es reconocer e identificar objetos en imágenes, así como entender el contenido y el contexto. CamFind es una aplicación móvil que reconoce e identifica los objetos en las fotografías.

- **Etiquetado/Búsqueda de Imágenes:** Google utiliza esta tecnología para permitir a los usuarios de Google+ buscar sus fotos por el contenido sin tener que etiquetar las fotos de antemano. Facebook está usando el etiquetado de imagen para mejorar la experiencia de intercambio de fotos de los usuarios.

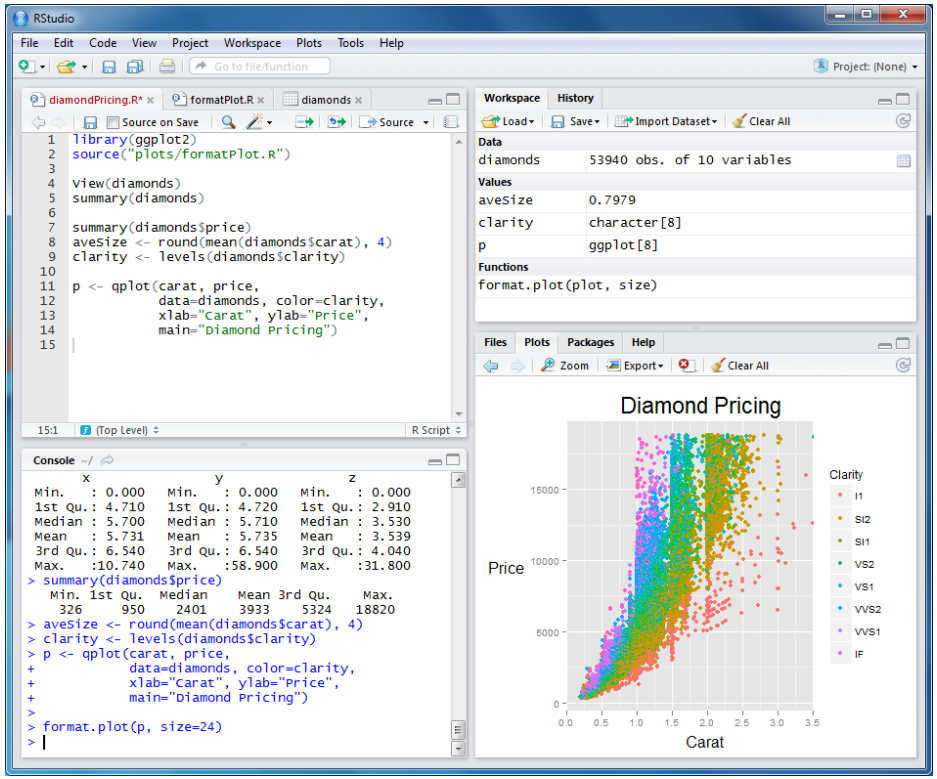
- **Otros ejemplos:** El coche autónomo de Google.

Esquema

Lenguaje de Programación **R**

Es un lenguaje y entorno de programación para análisis estadístico y gráfico. Es un dialecto libre del lenguaje S y fue desarrollado en la Universidad de Auckland en 1993.

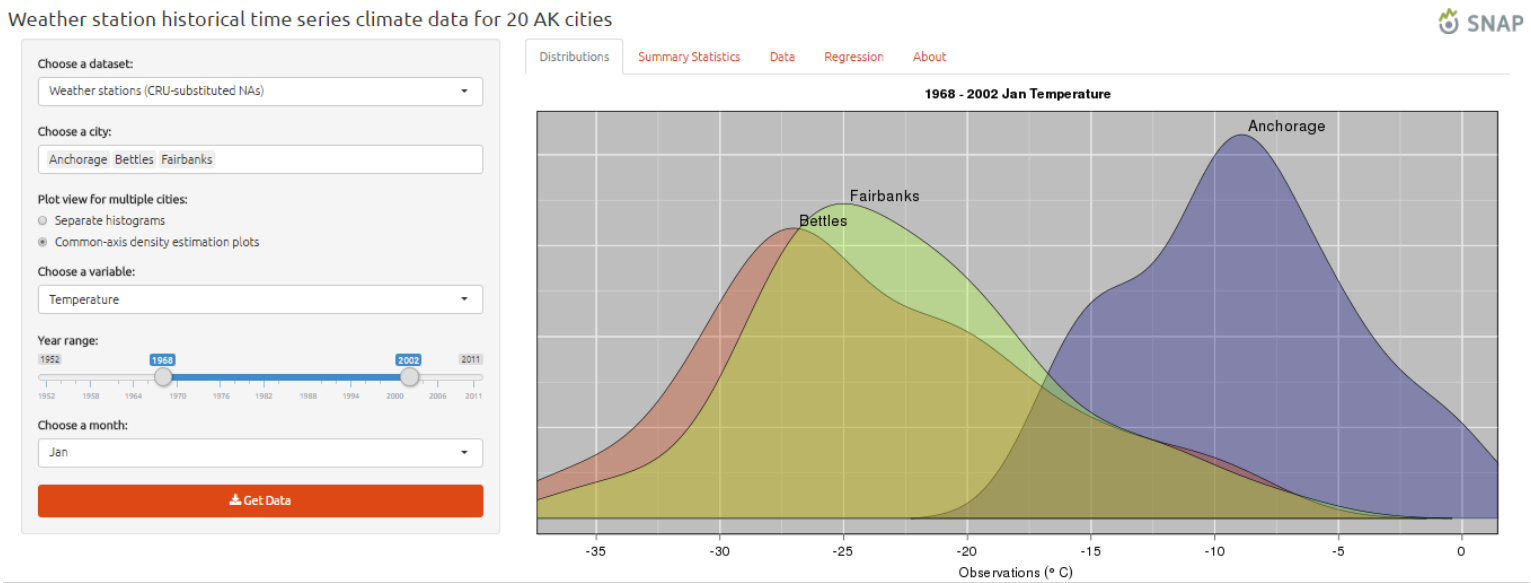
- Está **orientado a la estadística**, proporcionando un amplio abanico de herramientas (paquetes).
- Permite generar **gráficos de alta calidad** de manera sencilla.
- También puede usarse como herramienta de cálculo numérico y a la vez ser útil para la **minería de datos**.
- Puede integrarse con distintas **bases de datos** y existen bibliotecas que facilitan su utilización desde **lenguajes de programación** interpretados como Perl.
- Cuenta con un entorno de desarrollo muy productivo llamado **RStudio** que se puede descargar de forma gratuita.



Shiny by RStudio

Es un paquete en R creado en 2012 por Rstudio para desarrollar aplicaciones Web utilizando R con el objetivo de compartir los desarrollos realizados en R de una manera rápida y flexible. Permiten a los usuarios interactuar con los datos sin tener que manipular el código.

Proporciona un análisis básico de exploración de datos de la temperatura y las precipitaciones de Alaska.



Esquema

Casos de Uso en el Sector Telco

- **Predicción del churn (tasa de abandono)**: conocer de forma temprana qué clientes abandonarán una compañía o un servicio con el fin de construir una estrategia sólida de retención. Es uno de los problemas que resulta de mayor interés dentro del sector de las Telecomunicaciones puesto que resulta entre cinco y quince veces más caro captar nuevos clientes que retener a los actuales.

- **Campañas:** predecir qué clientes de tu cartera son más propensos a aceptar una campaña concreta.

- **Perfilado de clientes:** detectar grupos de clientes con gustos o comportamientos similares para mejorar su experiencia, recomendaciones, campañas, ...

- **Detección de Anomalías:** detectar comportamientos poco usuales o fraudulentos.

- **Series de Tiempo:** predicción de variables temporales: nº de reproducciones semanales de un servicio de TV, nº de altas mensuales, ...

Ayuda Torres Quevedo

- **¿Qué?** Ayudas de una duración de tres años.
- **¿A Quién?** a empresas, centros tecnológicos, etc.
- **¿Para Qué?** Para la contratación laboral de doctores que desarrollen proyectos de investigación a fin de favorecer su carrera profesional y ayudar a la consolidación de empresas tecnológicas de reciente creación.

Evaluación de técnicas Deep Learning en entornos telecom para cálculo de Churn:

- Estudio de viabilidad: estudio destinado a la evaluación del potencial de un proyecto, descubriendo sus puntos fuertes y débiles, así como de determinar los recursos necesarios para llevarlo a cabo y sus perspectivas de éxito.
- Descripción Científico - Técnica:
  - Análisis del estado del arte de las técnicas Deep Learning.
  - Análisis de la aplicabilidad de estas tecnologías al sector telecom.
  - Realizar un diagnóstico de las necesidades tecnológicas.
  - Seguimiento de la evolución de las técnicas Deep Learning.

Máster en Técnicas Estadísticas

TFM Modalidad B

Objetivo: que el alumno analice y estudie problemas del área de la estadística o la investigación operativa en los que estén interesadas las empresas colaboradoras y que estén relacionados con algunas de las técnicas estudiadas en el máster. **Tiempo:** al menos 225 horas.

- Modelos predictivos del churn (tasa de abandono) para operadores de telecomunicaciones. David Lozano Núñez. Optare Solutions. Verano 2014
- Predicción de series temporales. Comparación de modelos ARIMA vs modelos GAM. Marta Cousido Rocha. Optare Solutions. Verano 2015
- Aprendizaje estadístico para sistemas de recomendación. Plain Concepts.
- Caracterización de perfiles de glucosa en población no diabética. Unidad de epidemiología clínica del Hospital Universitario de Santiago de Compostela.
- Validación inicial de la capacidad predictiva de los estudios genéticos en la evaluación del riesgo de muerte súbita en pacientes con Miocardiopatía Hipertrófica. Health in Code, S.L.
- Desarrollo de un algoritmo de optimización de citas para pacientes del Hospital de Día. CHUS.
- Formulación de un modelo de scoring para solicitantes de créditos, no clientes de una entidad financiera + Modelo de estimación de ingresos para no clientes. ABANCA.
- Optimización bajo incertidumbre en redes de gas. Instituto Tecnológico de Matemática Industrial.

Esquema

Empleabilidad - Big Data Analytics

"Data scientists are like yetis - there aren't many of them"
(Big Data Summit, HP, 2014)

"Why Data Science Jobs Are in High Demand?"
(Harvard Extension School, 2015)

"Career of the Future: Data Scientist Study"
(EMC, 2015)

"Big data: The next frontier for innovation, competition, and productivity"
(McKinsey Global Institute, 2011)

"Big Data Creates Big Jobs: 4·4 Million IT Jobs Globally to Support Big Data By 2015"
(Gartner, 2012)

"Data Scientist: The Sexiest Job of the 21st Century"
(Harvard Business Review, 2012)

"Big y Open Data como motores de crecimiento en la Europa de 2020"
(OSIMGA, 2014)

Noticias en Blogs Tecnológicos

- `http://www.teknlife.com/`: Según previsiones del MBIT School (Madrid Business Intelligence Technology), en los próximos años el 20 % de las pymes, el 80 % en la mediana empresa y del 100 % en las grandes empresas de nuestro país contarán con más de un Data Scientist, y consideran que esta es actualmente una de las profesiones más demandadas.

- `http://www.t-systemsblog.es/`: Esta es una profesión que se encuentra, actualmente, entre las más demandadas y cuenta con grandes expectativas de futuro en el mercado laboral.

- `http://www.datahack.es//`: LA PROFESIÓN DEL FUTURO, DATA SCIENTIST. El nivel de estudios en estos profesionales suele ser bastante alto: el 46 % de ellos tiene un doctorado y el 42 % un máster.

Optare Solutions
NATURAL BORN TELCO

Máster en Técnicas Estadísticas

10 AÑOS

Data Science: del Data Mining al Deep Learning

Irene Castro Conde[1,2]
icastro@optaresolutions.com

[1]Optare Solutions S.L.
[2]Grupo SiDOR, Universidad de Vigo

I Jornada de Orientación Profesional del MTE

23 Junio 2016