Máster en Ciencia de Datos/ Data Science

Presentación

Jesús Marco de Lucas (CSIC)

Instituto de Física de Cantabria, IFCA

9 Octubre 2017

PRESENTACIÓN DEL MASTER

DESPUES DE UNA “BREVE” INTRODUCCIÓN, DISCUSION SOBRE:

- ¿Un Master *más* en Data Science?
- ¿Qué referencia seguimos?
- Equipo del master
- Alumnado en el curso 17-18
- ¿Qué especialidades se ofertan?
- Prácticas y Data Labs
- TFM: industria versus grupos de investigación
- Herramientas comunes y soporte
- Aproximación Didáctica

PROGRAMA Y DESARROLLO

https://masterdatascience.ifca.es

1.- FUNDAMENTOS

(OCTUBRE-ENERO)

Modulo obligatorio, cinco materias

Panorama de Data Science	Métodos de Data Science	Data Management
Introducción a Big Data y Open Science	Estadística para Data Science	Modelos de datos y sistemas de información
	Data Mining	Ciclo de vida de los datos: de adquisición a presentación.

2.- MÓDULO DE ESPECIALIZACIÓN

(ENERO-ABRIL)

Un área de especialización a elegir:

Data Science Analytics	Data Science Engineering	Open Data Management
Machine Learning I	Sistemas de Computación para Big Data	Acceso a los Datos: Portales y Servicios
Machine Learning II	Cloud para Data Science	Preservación de Datos
Semantics, Linked Data, Text Data Mining	Desarrollo de Proyectos	Repositorios de Datos

3.- MÓDULO PROFESIONAL

(ABRIL-JUNIO)

Este módulo obligatorio incluye:

Seguridad, Privacidad y Aspectos Legales

Nuevos Desarrollos en Data Science (seminarios)

4.- MÓDULO DE ORIENTACIÓN PROFESIONAL

(MAYO-SEPTIEMBRE)

De acuerdo a su interés profesional y cualificación, el/la estudiante podrá optar por realizar prácticas externas y/o "Data Labs" en diferentes áreas

Data Labs		
Biomedicina	Medio Ambiente, Meteorología	Física y Astronomía
Economía y Finanzas	Internet of Things	Ciencias Sociales
+ Prácticas Externas en empresas y grupos de investigación		

5.- TESIS DE MÁSTER

(Comienzo en MAYO, presentación en SEPTIEMBRE)

Un trabajo avanzado desarrollado de forma autónoma por el estudiante bajo la supervisión de un profesor del Master. La temática y orientación dependerán de la especialidad elegida. Supondrá un trabajo de iniciación al contexto profesional que permita unirse en el futuro a una empresa o a un grupo de investigación.

Se promoverá que la tesis se realice asociada a un contrato externo remunerado en las empresas colaboradoras o grupos de investigación.

¿Por qué Open Science y Big Data?

The world is witnessing a dramatic increase in the amount and variety of data being produced. Alongside the data created by billions of people using digital devices and services for personal and professional reasons, and the data generated by the increasing number of connected objects, there is data from research, from digitised literature & archives and from public services such as hospitals and land registries. This Big Data phenomenon creates new possibilities to share knowledge, to carry out research and to develop and implement public policies.
(EU_Commision(2016) 178)

"This exponential growth of data, the availability of increasingly powerful digital technologies, the globalisation of the scientific community, as well as the increasing demand to ***address the societal challenges of our times****, are the bases of an on-going transformation and opening up of science and research, referred to as* ***Open Science****"*

"Open Science has the potential to increase the quality, impact and benefits of science and to accelerate advancement of knowledge by making it more reliable, more efficient and accurate, better understandable by society and responsive to societal challenges, and has the potential to enable growth and innovation through reuse of scientific results by all stakeholders at all levels of society, and ultimately contribute to growth and competitiveness of Europe" (EU_Council **9526/16)**

EJEMPLOS DE APLICACIÓN

- Desarrollo de nuevos medicamentos y tratamientos
- Explotación sostenible de recursos naturales
- Estudio del cosmos (astrofísica y física de partículas)
- Análisis de impacto medioambiental
- Distribución de energía y de agua
- Biodiversidad y ecosistemas
- Modelos socioeconómicos de ciudades y regiones
- Contexto de patrimonio cultural

DEMANDA DE PERFILES PROFESIONALES

- Para:
  - Universidades, centros de investigación de todo el mundo
  - Consultoras
  - Administración pública
  - Empresas de servicios
  - PYMES y start-up
- Tres perfiles de especialización:
  - Data Analytics
  - Data Engineering
  - Open Data Management

EJEMPLOS REALES DE EMPLEOS OFERTADOS

- Data Scientist (>350 demandas)

  The Microsoft Data Insights Global Domain is hiring Data Scientists to join our international team of data science consultants. As a Data Scientist you use statistical, mathematical and predictive models in combination with your business strategy skills to find the right answers to complex business questions. You are able to communicate your findings both orally and visually. You act as a trusted advisor for our enterprise customers. You advise, lead and challenge them when it comes to Data Science.

  The Data Insights Senior Data Scientist must apply advanced analytics skills and experience in principals of the analytics solution engineering to provide detailed reliable solutions for services implementations. A good understanding of competitive and open source technology enables the data scientist to properly assess customer solutions hosted on non-Microsoft platforms.

CAPACITACIÓN

- Titulaciones de entrada:
  - Grados/Licenciaturas en Física, Matemáticas, Ingeniería Informática, Telecomunicaciones, Industriales
  - Otras ingenierías, Economía, Ciencias Sociales, otras titulaciones
- Duración e itinerarios
  - Máster de 60 ECTS
  - Itinerario básico: 1 año académico (ej. Graduados recientes)
  - Itinerario tiempo parcial: 2 años (ej. Profesionales en activo)
  - Itinerario extendido: +30 ECTS formación complementaria +30ECTS optativas/prácticas (ej. Graduados en otros países con 180 ECTS, que quieren acceder al doctorado)
- Oferta de formación inicial complementaria (asociada a los requerimientos)
  - Introducción al análisis de datos, estadística, cálculo, programación, inglés
- Supervisión de la integración del curriculum
  - Herramientas comunes (R, Python; Cloud, HPC…)
  - Claro perfil de especialización (data analytics, open data, data engineering)
- Posibilidad de Prácticas profesionales
  - Acceso a prácticas remuneradas mediante acuerdos con empresas/grupos

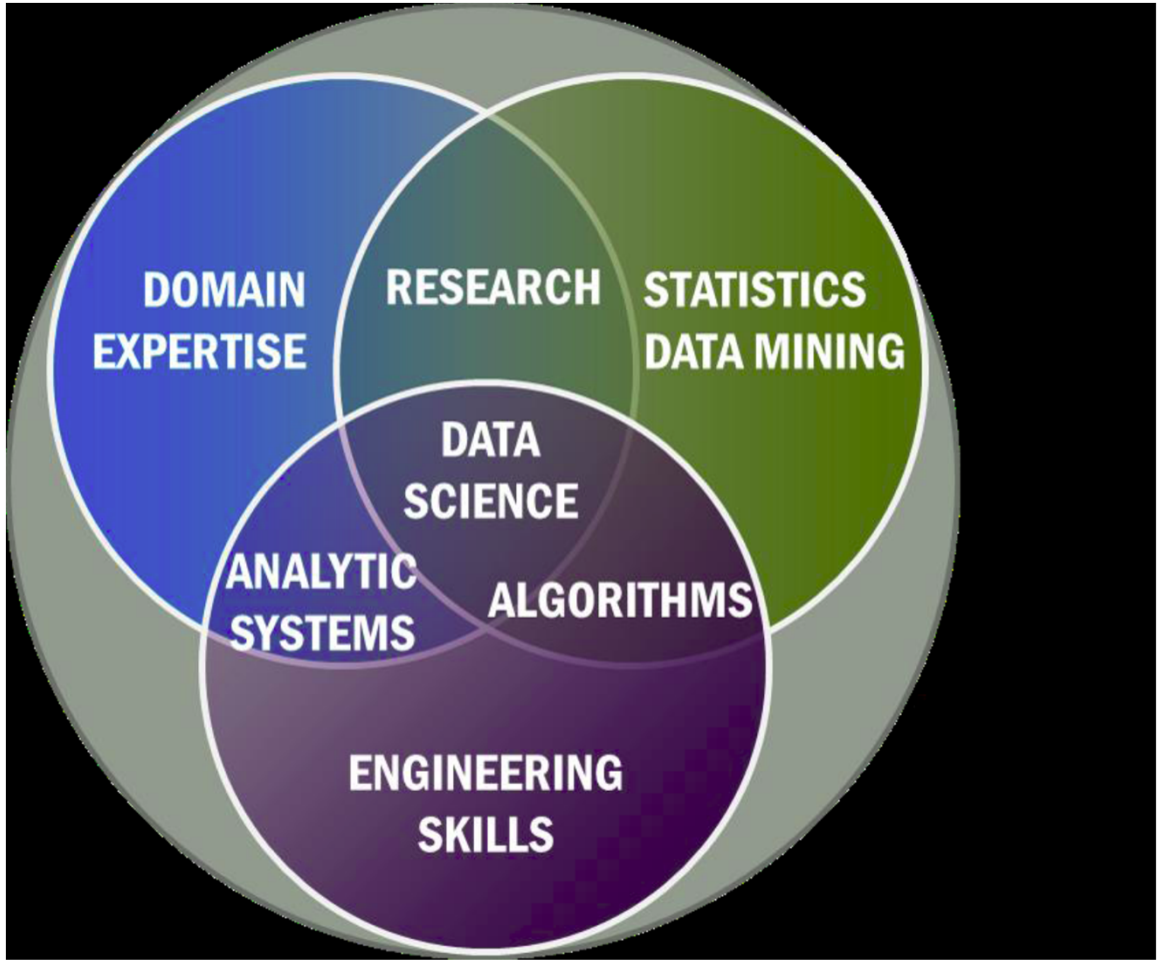
PUNTOS FUERTES:

- Profesorado experto en los diferentes dominios con la visión Open Science
- Recursos necesarios: repositorios “Big Data”, supercomputación, cloud, etc.
- Enfoque “Científico”: Ciclo de vida del dato, simulación, validación, presentación…
- Técnicas punteras aplicadas (ej. Deep NN usando GPUs)
- Integración de los alumnos en problemas reales con salidas profesionales
- Participación de empresas
- Oferta de acceso a prácticas

Qué es un master en Data Science especializado en Open Science y Big Data?

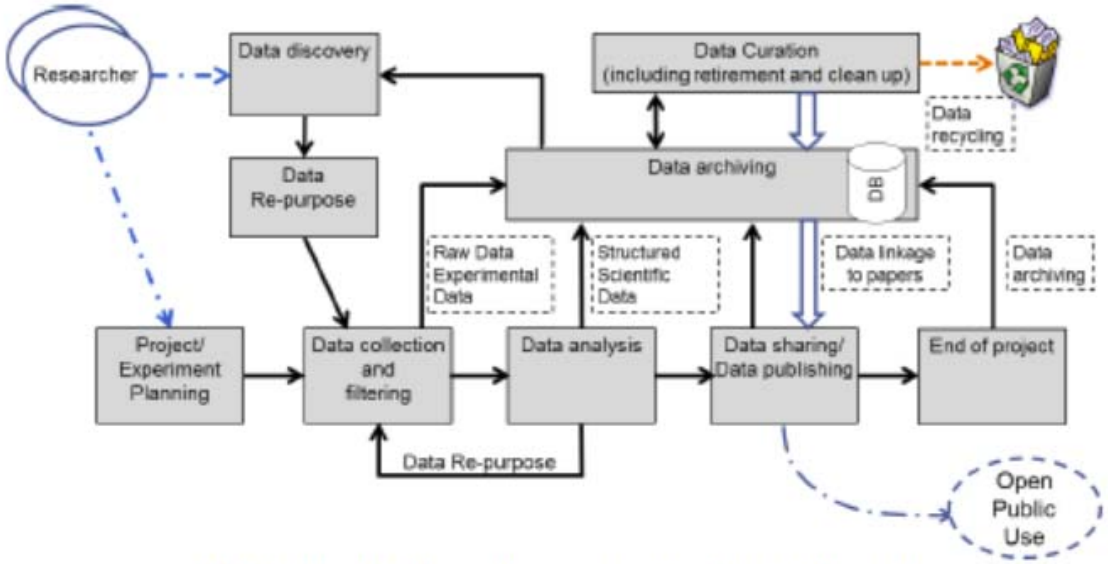
- El perfil en Open Science and Big Data combina un rol técnico de analista de datos con el de la visión que permite acceder a diferentes áreas de conocimiento útiles en múltiples sectores de negocio.
- Este master especializado en Open Science and Big Data va más allá de los masters usuales en Data Science: parte también de una base sólida que incluye estadística, modelización y técnicas avanzadas de análisis de datos. Igualmente proporciona la formación necesaria para hacer de puente entre los responsables de negocio y la división TIC (*person in the middle*), cambiando la forma en que una empresa aborda un reto en Big Data, desde su propia definición hasta su ejecución.
- La especialización en Open Science proporciona además la **visión global de los recursos necesarios en diferentes áreas para abordar un reto complejo desde diferentes ángulos**, y a la vez de cómo conseguirlos y aplicarlos.
- En resumen, este master permite recorrer el camino desde los datos y herramientas en diferentes áreas hasta el nivel de conocimiento final, aplicable en los negocios, siguiendo un proceso común, basado en la forma en que se trabaja actualmente en Ciencia para la explotación de datos.

¿Qué se entiende por Data Science?

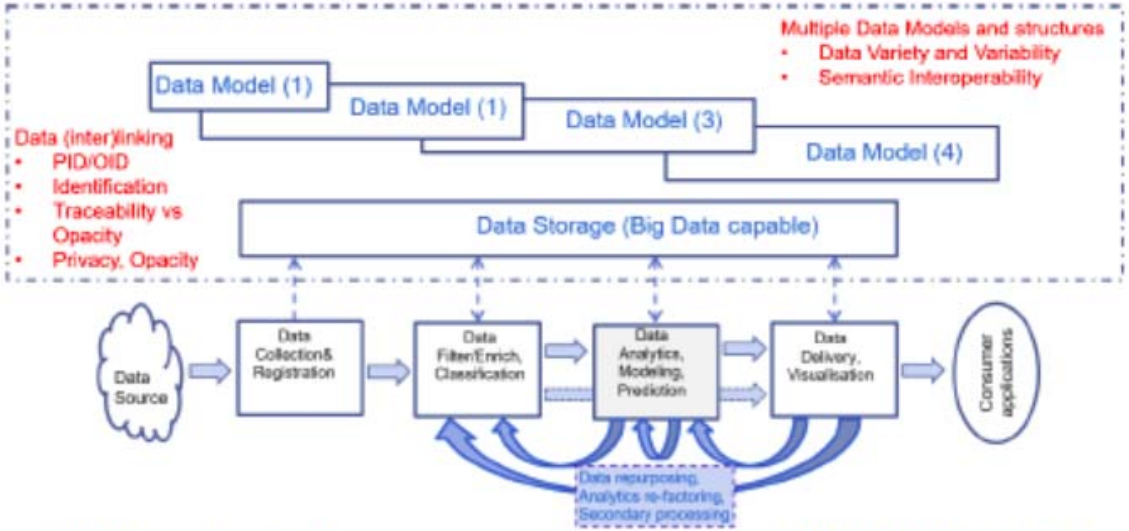


NIST

BIGDATA y Data Life Cycle? (EDISON)

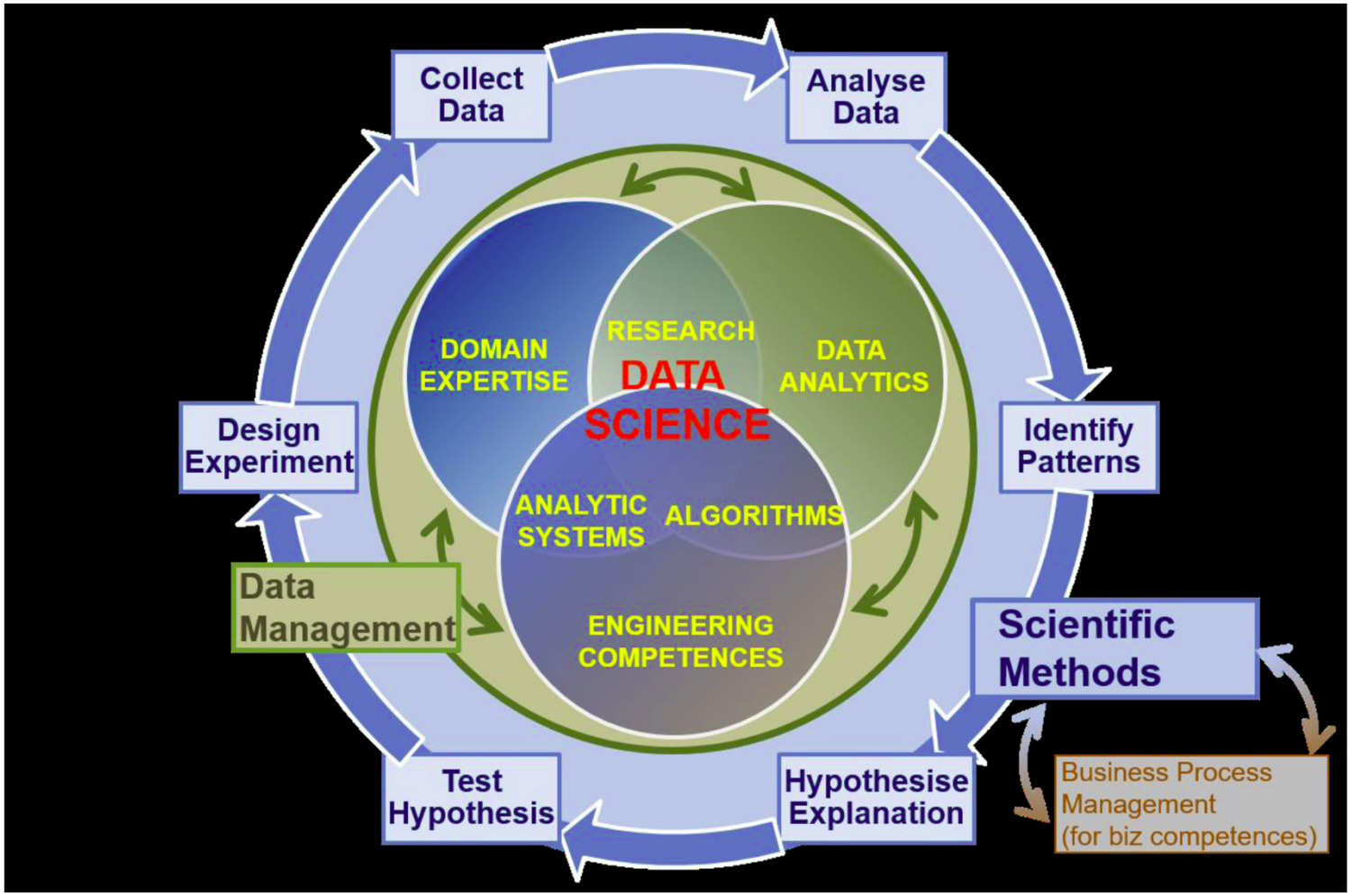


(a) Scientific data lifecycle management - e-Science focused

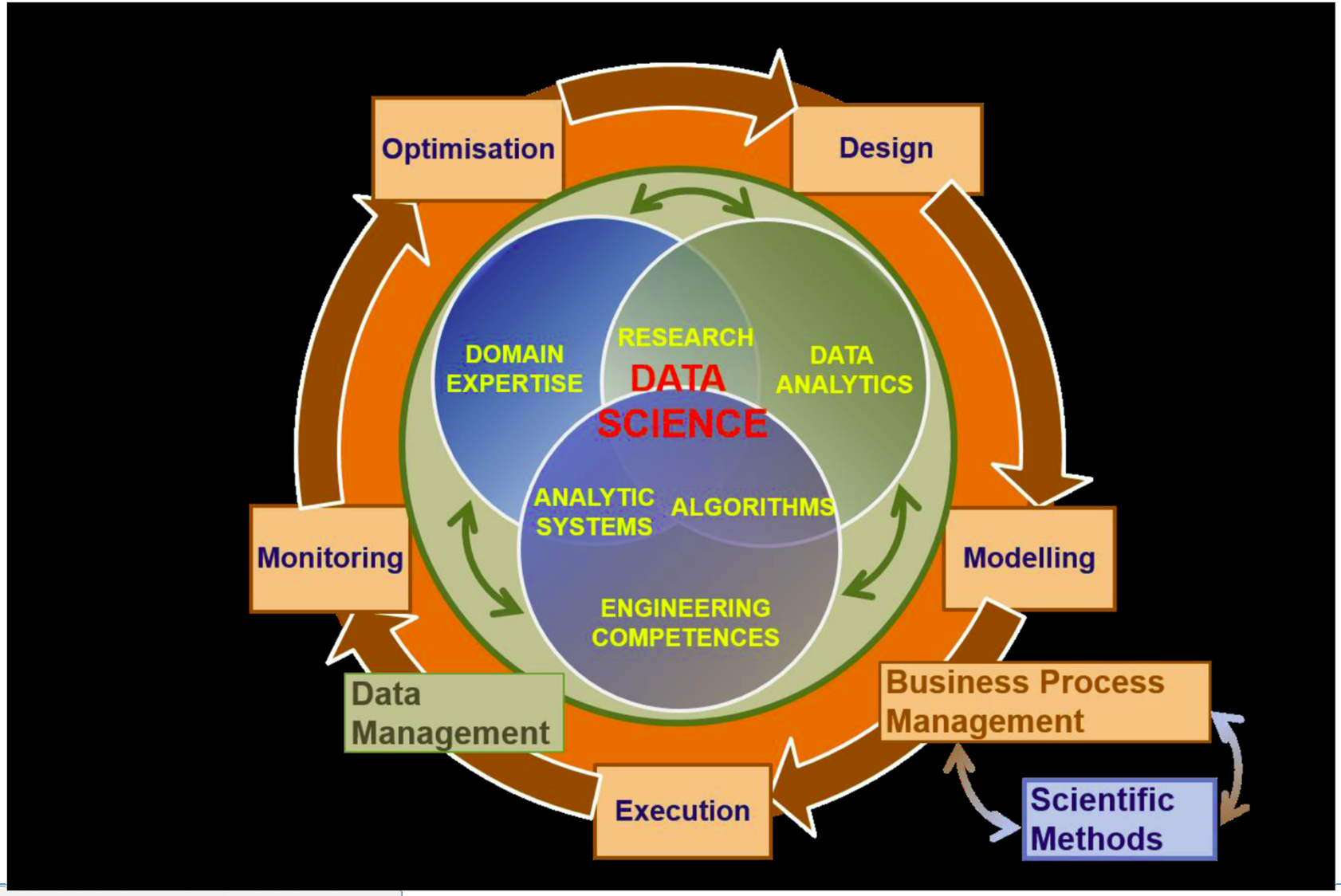


(b) Big Data Lifecycle Management model (compatible with the NIST NBDIF definition)

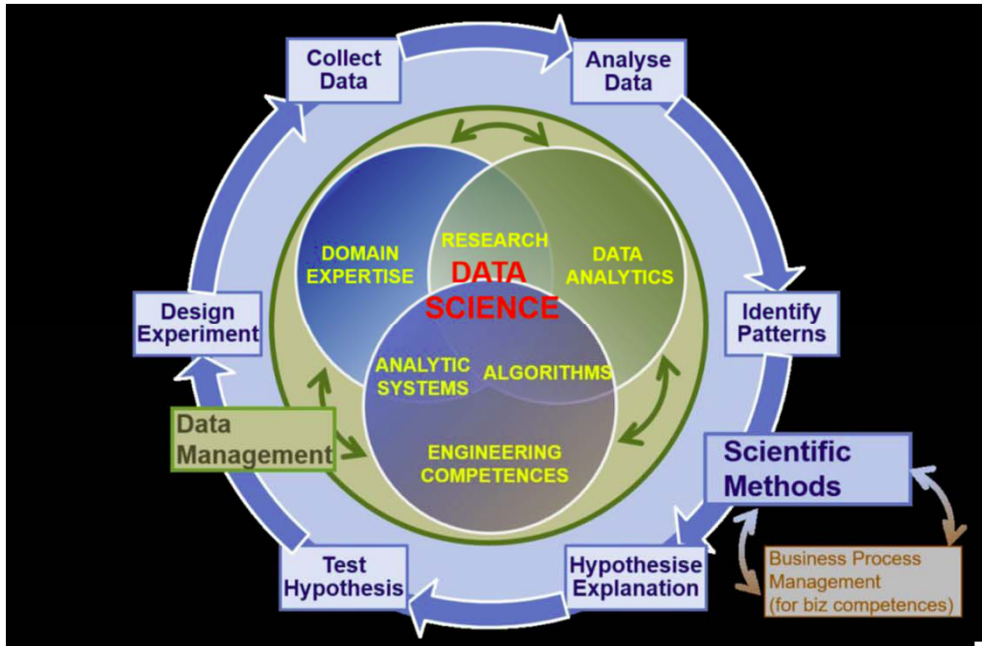
Research cycle versus Business cycle

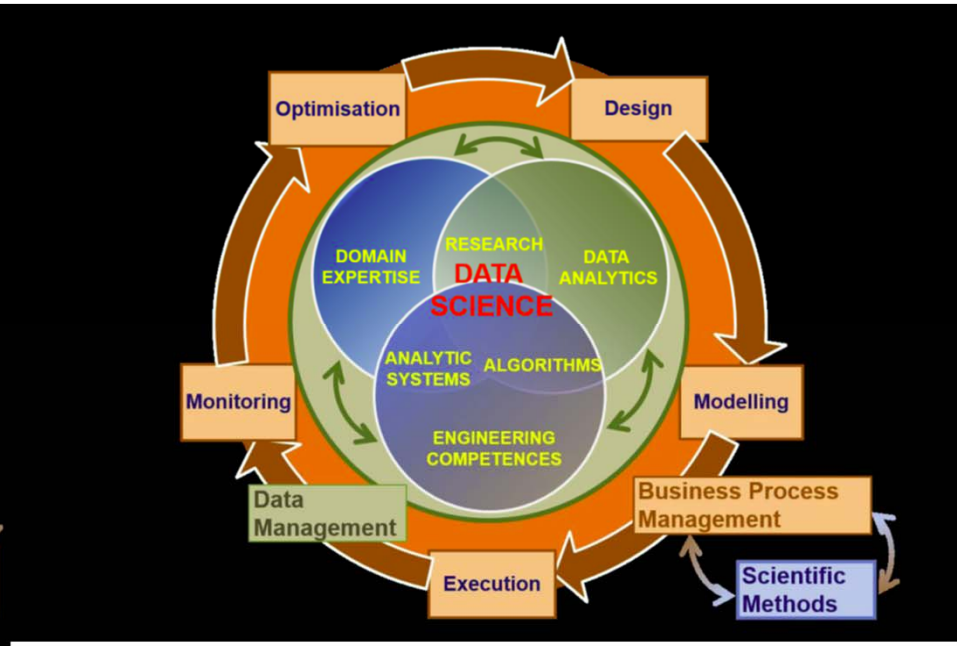


Research cycle versus Business cycle



Research cycle versus Business cycle





Research cycle versus Business cycle

RESEARCH	BUSINESS
Define research questions	Define the business target: both services and customers
Design experiment representing initial model of research object or phenomena	Design the business process Model/Plan
Collect Data	Deploy and Execute
Analyse Data	Monitor and Control
Identify Patterns	
Hypothesise Explanation	Model/Plan
Test Hypothesis	Model/Plan
Refine model and start new experiment cycle	Optimise and Re-design

DATA SCIENCE SKILLS

As outcome of analysis of Data Science job advertisements, three main skills :

- **General Data Science skills** or required (provable) experience
- Knowledge and **experience with Big Data hardware and software platforms**
- **Programming** language (general and those having extended statistics libraries that are generally related to Data Science Engineering skills but in many cases are treated as a separate group of skills).

Can be grouped into:

- **Data Analytics and Machine Learning**
- **Data Management/Curation** (including both general and scientific data management)
- **Data Science Engineering** (hardware and software) skills
- **Scientific/Research** Methods
- Personal, inter-personal communication, **team work** (also called social intelligence or soft skills)
- **Application/subject domain** related (research or business)

CONTENTS

	Big Data Analytics platforms	Math& Stats tools	Databases	Data/ applications visualization	Data Management and Curation platform
1	Big Data Analytics platforms	Advanced analytics tools (R, SPSS, Matlab, etc)	SQL and relational databases	Data visualization Libraries (D3.js, FusionCharts, Chart.js, other)	Data modelling and related technologies (ETL, OLAP, OLTP, etc)
2	Big Data tools (Hadoop, Spark, etc)	Data Mining tools: RapidMiner, others	NoSQL Databases	Visualisation software (D3, Processing. Tableau, Gephi, etc)	Data warehouses platform and related tools
3	Distributed computing tools a plus (Spark, MapReduce, Hadoop, Hive, etc.)	Matlab	NoSQL, Mongo, Redis	Online visualization tools (Datawrapper, Google Charts, Flare, etc)	Data curation platform, metadata management (ETL, Curator's Workbench, DataUp, MIXED, etc)
4	Real time and streaming analytics systems (like Flume, Kafka, Storm)	Python	NoSQL, Teradata		Backup and storage management (iRODS, XArch, Nesstar, others)
5	Hadoop Ecosystem/platform	R, Tableau R	Excel		
6	Spotfire	SAS			
7	Azure Data Analytics platforms (HDInsight, APS and PDW, etc)	Scripting language, e.g. Octave			
8	Amazon Data Analytics platform (Kinesis, EMR, etc)	Statistical tools and data mining techniques			
9	Other cloud based Data Analytics platforms (HortonWorks, Vertica LexisNexis HPCC System, etc)	Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)			

Programming

- Programming languages with extended data analysis and statistics libraries are identified as important for Data Scientist (and typically identified in job descriptions as requiring several years of practical experience):
  - **Python , R**
  - Scala
  - **Pandas** (Python Data Analysis Library)
  - Julia
  - Java and/or C/C++ as general applications programming languages
  - **Git versioning** system as a general platform for software development
  - ***Scrum agile software*** development and management methodology

Perfiles profesionales (ocupación)

(EDISON 3.7, extensión de CWA 16458 ICT Profiles)

- Data Science/Big Data Infrastructure Managers
  - including Research Infrastructures Data Storage Facilities Manager
- Data Science Professionals
  - Data Science researcher
  - (Big) Data Analyst
  - Data Science Application Programmer
  - Business Analyst
  - Large Scale (cloud) Database designer
  - Large Scale (cloud) Database administrator
  - Scientific Database administrator
- Data Science Technology Professionals
  - Big Data Facilities operators
  - Large scale (cloud) data storage operators
  - Scientific Database operator
  - Digital Librarian
  - Data Archivist/ Steward/ Curator

Disciplinas científicas (áreas) (ACM Classification)

1- DATA SCIENCE ANALYTICS (DSA)		
Theory of computation	• Design and analysis of algorithms (incl. Data structures design and analysis) • Theory and algorithms for application domains • Semantics and reasoning	EXTENSION: • Algorithms for Big Data computation
Mathematics of Computing	• Discrete Mathematics (Graph Theory) • Probability and Statistics • Mathematical Software • Information Theory • Mathematical Analysis	EXTENSION: • Mathematical Software for Big Data computation
Computing Methodologies	• Artificial Intelligence • Machine learning	EXTENSION: • New DSA computing (DNN¿?)
Information Systems	• Decision Support Systems • Multimedia Information Systems • Data Mining	EXTENSION: • Big Data systems (cloud based) • Big Data applications • Domain specific data applications

Disciplinas científicas (áreas) (ACM Classification)

2- DATA SCIENCE ENGINEERING (DSE)

Computer systems organization	• Parallel architectures • Distributed architectures	EXTENSION: • Supercomputers architecture • GPUs
Software and its engineering	• Software organization and properties • Software notation and tools • Software creation and management	EXTENSION: • Big Data Application design
Computing Methodologies	• Modelling • Simulation	EXTENSION:
Information Systems	• Information storage systems • Enterprise information systems • Collaborative and social computing systems and tools	EXTENSION: • Big Data and cloud based systems design

Disciplinas científicas (áreas) (ACM Classification)

3- DATA SCIENCE DATA MANAGEMENT (DSDM)		
Data Management systems	• Database design • Data structures • Query languages • Database administration • Middleware for databases	EXTENSION: • Data types and structures description • Metadata standards • Persistent Identifiers (PID) • Data type registries
Information systems applications and information retrieval	• Digital Libraries and Archives • Document representation • Retrieval modes and ranking • Search engines architecture and scalability • Specialized information retrieval	

4- DATA SCIENCE DOMAIN KNOWLEDGE (DSDK)		
Domain Knowledge (EXTENSIONS)	• Physical Sciences and Engineering • Life and Medical Sciences • Law, social and behavioral sciences • Computer forensics • Arts and humanities	• Computers in other domains • Operations research • Education • Document management and text processing

PROGRAMA Y DESARROLLO

1.- FUNDAMENTOS
(OCTUBRE-ENERO)
Modulo obligatorio, cinco materias

Panorama de Data Science	Métodos de Data Science	Data Management
Introducción a Big Data y Open Science	Estadística para Data Science	Modelos de datos y sistemas de información
	Data Mining	Ciclo de vida de los datos: de adquisición a presentación.

2.- MÓDULO DE ESPECIALIZACIÓN
(ENERO-ABRIL)
Un área de especialización a elegir:

Data Science Analytics	Data Science Engineering	Open Data Management
Machine Learning I	Sistemas de Computación para Big Data	Acceso a los Datos: Portales y Servicios
Machine Learning II	Cloud para Data Science	Preservación de Datos
Semantics, Linked Data, Text Data Mining	Desarrollo de Proyectos	Repositorios de Datos

3.- MÓDULO PROFESIONAL
(ABRIL-JUNIO)
Este módulo obligatorio incluye:

Seguridad, Privacidad y Aspectos Legales

Nuevos Desarrollos en Data Science (seminarios)

4.- MÓDULO DE ORIENTACIÓN PROFESIONAL
(MAYO-SEPTIEMBRE)
De acuerdo a su interés profesional y cualificación, el/la estudiante podrá optar por realizar prácticas externas y/o "Data Labs" en diferentes áreas

Data Labs		
Biomedicina	Medio Ambiente, Meteorología	Física y Astronomía
Economía y Finanzas	Internet of Things	Ciencias Sociales
+ Prácticas Externas en empresas y grupos de investigación		

5.- TESIS DE MÁSTER
(Comienzo en MAYO, presentación en SEPTIEMBRE)
Un trabajo avanzado desarrollado de forma autónoma por el estudiante bajo la supervisión de un profesor del Master. La temática y orientación dependerán de la especialidad elegida. Supondrá un trabajo de iniciación al contexto profesional que permita unirse en el futuro a una empresa o a un grupo de investigación.
Se promoverá que la tesis se realice asociada a un contrato externo remunerado en las empresas colaboradoras o grupos de investigación.

PRESENTACIÓN DEL MASTER

DESPUES DE UNA "BREVE" INTRODUCCIÓN, DISCUSION SOBRE:

- ¿Un Master *más* en Data Science?
- ¿Qué referencia seguimos?
- Equipo del master
- Alumnado en el curso 17-18
- ¿Qué especialidades se ofertan?
- Prácticas y Data Labs
- TFM: industria versus grupos de investigación
- Herramientas comunes y soporte
- Aproximación Didáctica