

UEA Datos a Gran Escala

Introducción a la UEA Datos a Gran Escala



Dr. Pedro Pablo González Pérez

e-mail: pgonzalez@correo.cua.uam.mx

<http://dcni.cua.uam.mx/division/usuario?p=31#>

Departamento de Matemáticas Aplicadas y Sistemas



**UNIVERSIDAD
AUTÓNOMA
METROPOLITANA**
Unidad Cuajimalpa

UEA Datos a Gran Escala

Objetivo general:

- ❑ Al final de la UEA el alumno será capaz de manipular grandes volúmenes de datos, así como los métodos y herramientas para la extracción e inferencia de información a partir de los mismos.



UEA Datos a Gran Escala

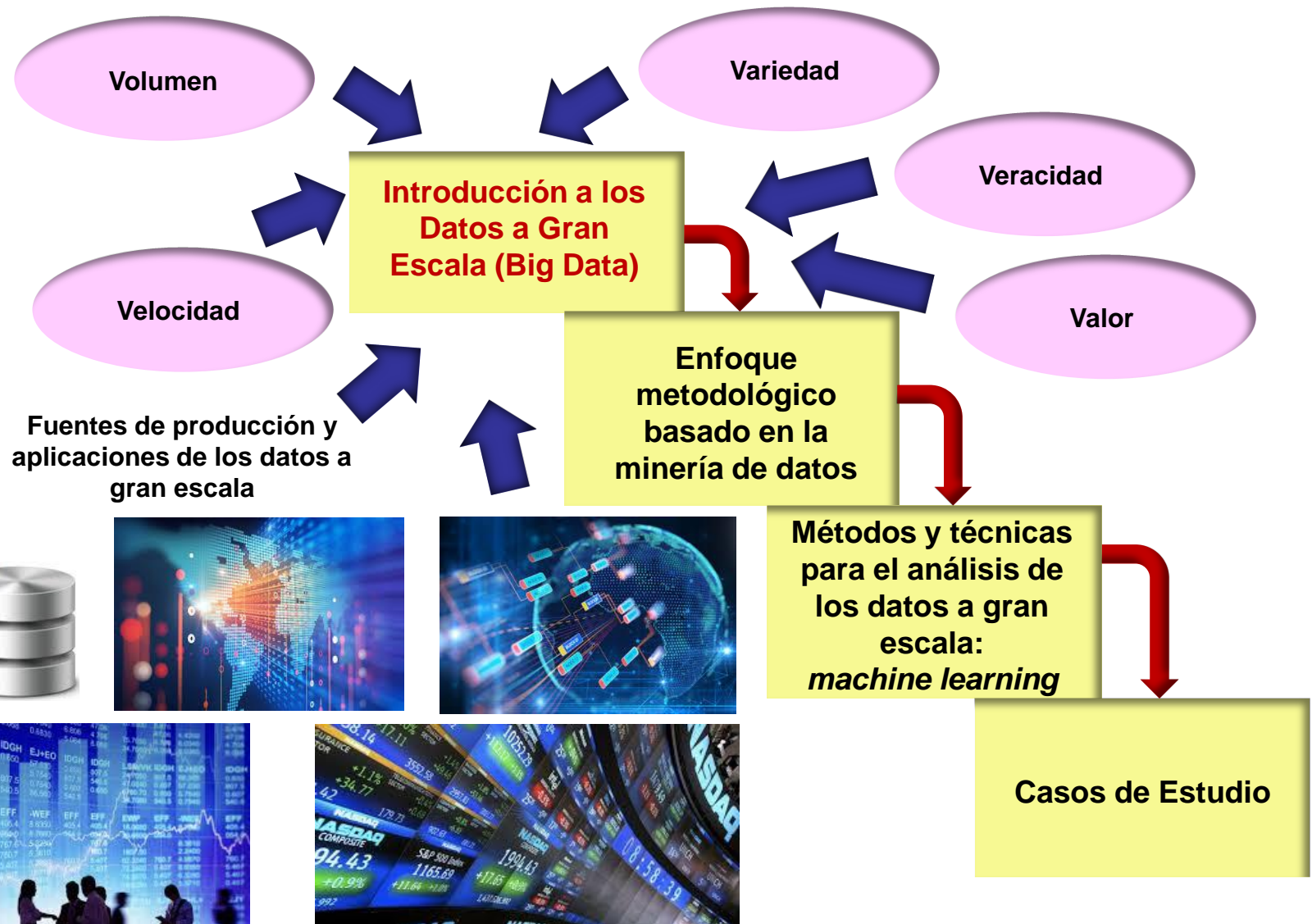
Objetivos parciales:

1. Comprender el alcance de los grandes volúmenes de datos como fuente para la generación de nueva información y conocimiento.
2. Identificar los principales componentes tecnológicos en un sistema de datos a gran escala.
3. Aplicar algunos de los principales métodos y herramientas para la extracción e inferencia de información a partir de grandes volúmenes de datos.



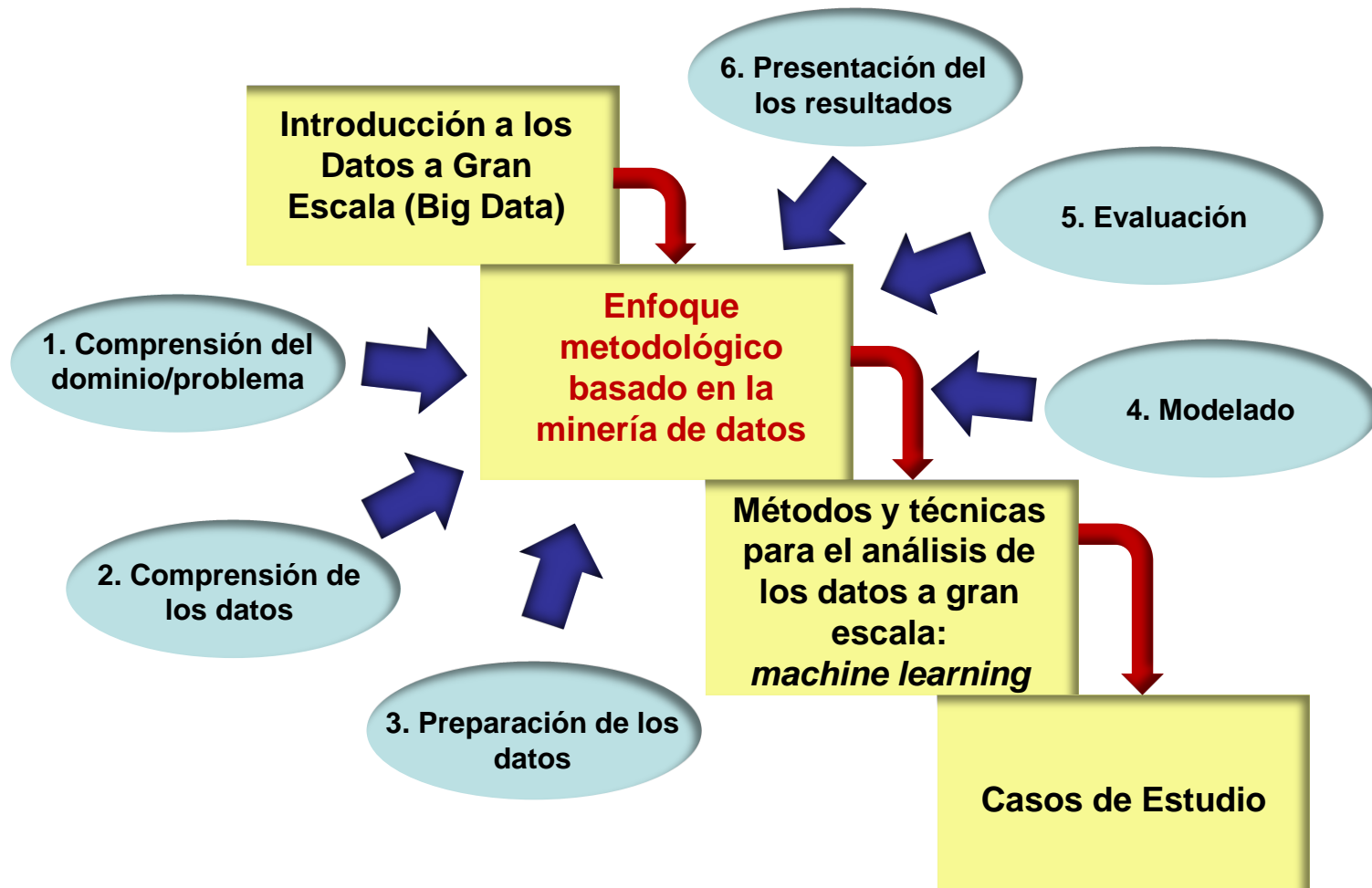
UEA Datos a Gran Escala

Contenido sintético:



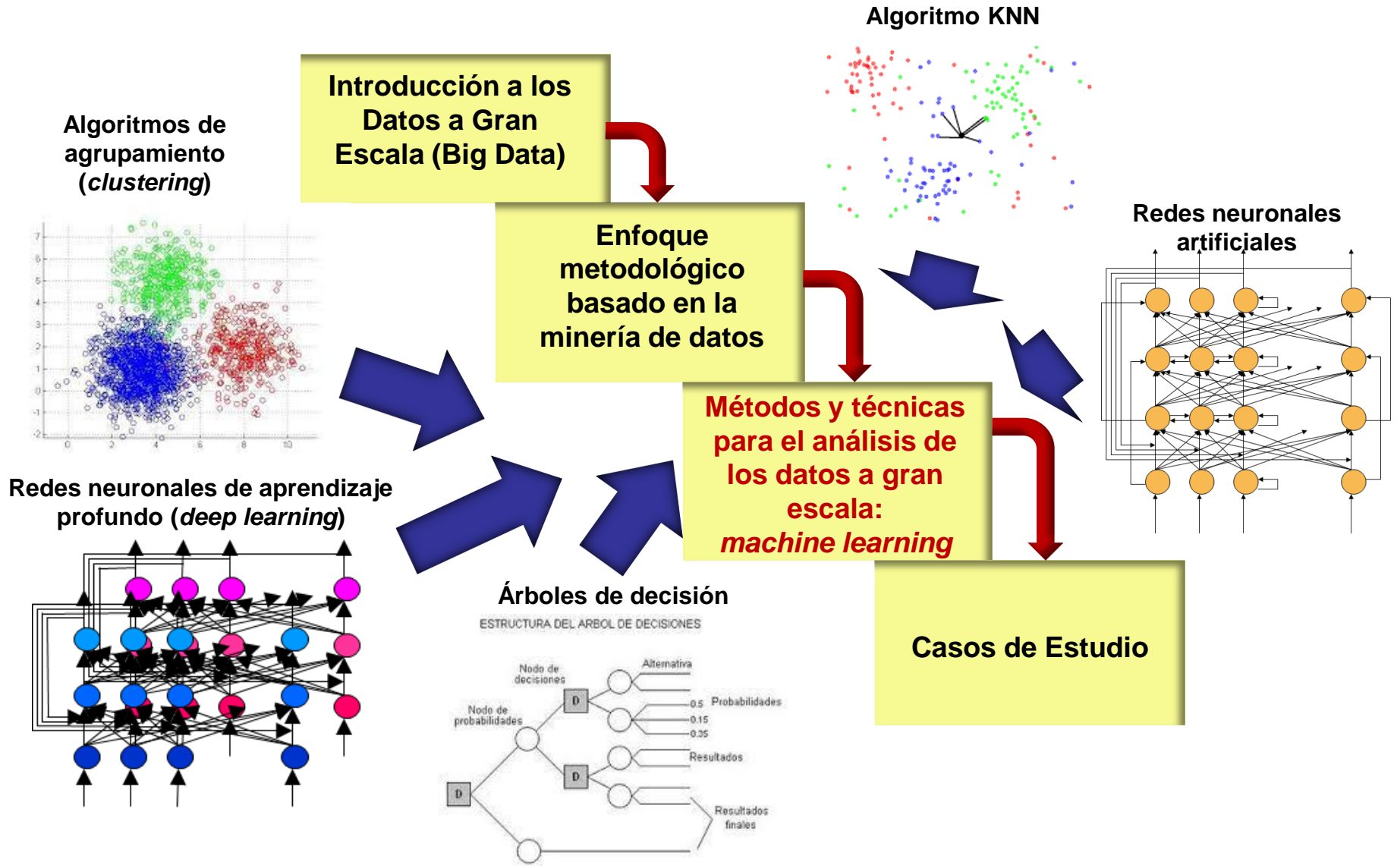
UEA Datos a Gran Escala

Contenido sintético:



UEA Datos a Gran Escala

Contenido sintético:



UEA Datos a Gran Escala

Contenido sintético:

**Introducción a los
Datos a Gran
Escala (Big Data)**

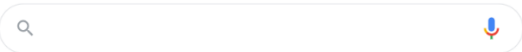
**Enfoque
metodológico
basado en la
minería de datos**

**Métodos y técnicas
para el análisis de
los datos a gran
escala:
*machine learning***

Casos de Estudio

Motores de búsqueda

Google



Buscar con Google

Me siento con suerte

Ofrecido por Google en: [English](#) [Español \(Latinoamérica\)](#)

Plataformas e-commerce



Bases de datos

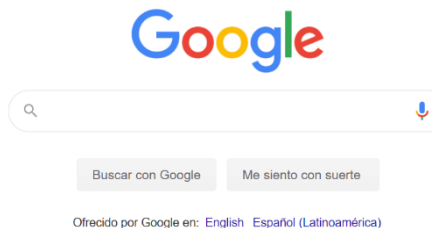


UEA Datos a Gran Escala

Contenido sintético detallado:

1. Introducción a los datos a gran escala (*big data*).

- ☐ Aproximaciones al concepto de datos a gran escala.
- ☐ Características de los datos a gran escala: **volumen**, **velocidad**, **variedad**, incertidumbre en la **veracidad** y **valor**.
- ☐ Fuentes de producción de los datos a gran escala.
- ☐ Principales aplicaciones de los datos a gran escala.

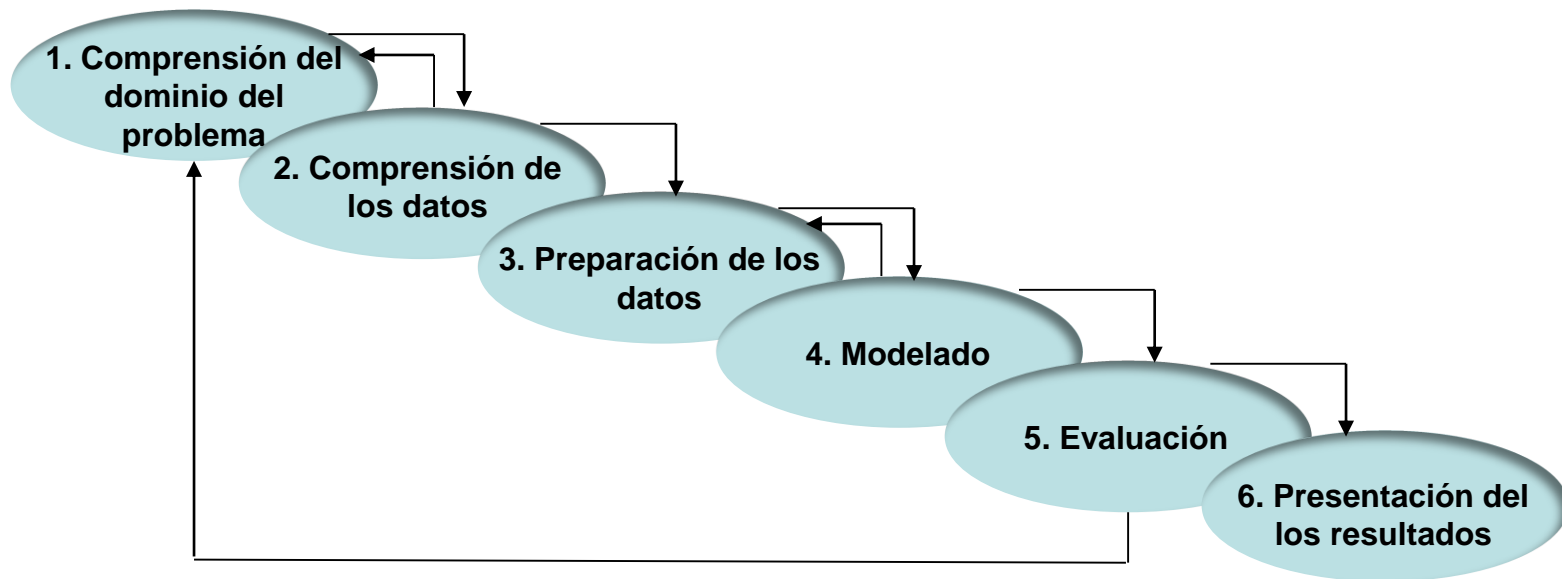


UEA Datos a Gran Escala

Contenido sintético detallado:

2. Enfoque metodológico basado en la minería de datos.

CRISP-DM [Shearer, 2000; IBM Corporation, 2012]

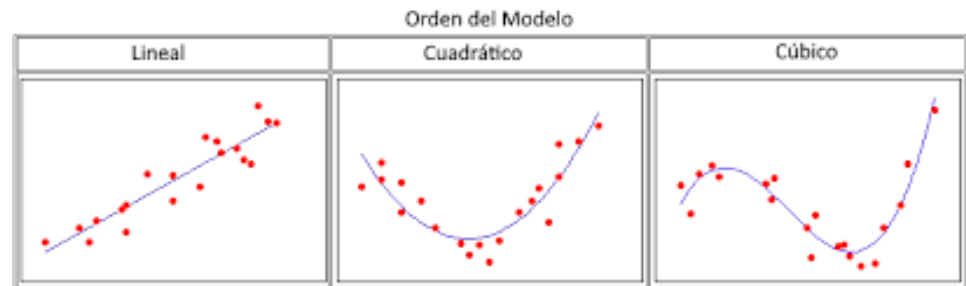
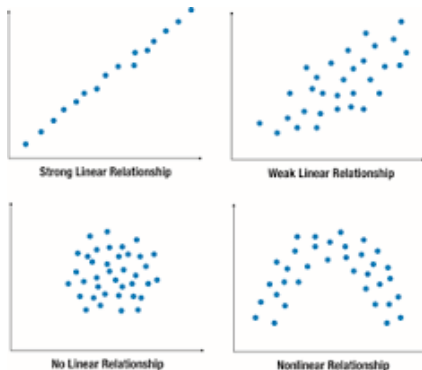
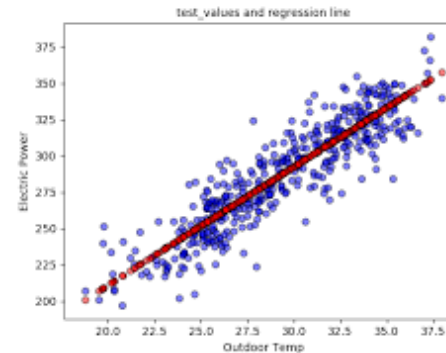
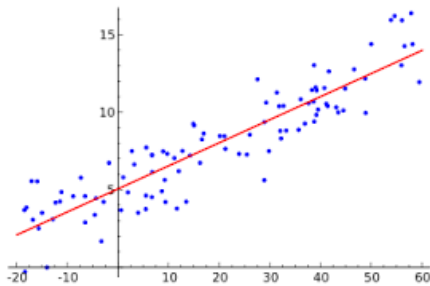


UEA Datos a Gran Escala

Contenido sintético detallado:

3. Métodos y tecnologías para el análisis de los datos a gran escala.

- ❑ **Métodos estadísticos: algoritmos de regresión.**
- ❑ *Aprendizaje automatizado (machine learning).*



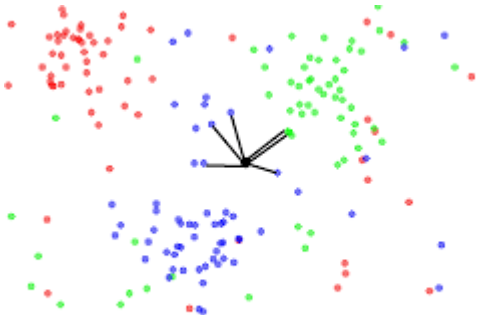
UEA Datos a Gran Escala

Contenido sintético detallado:

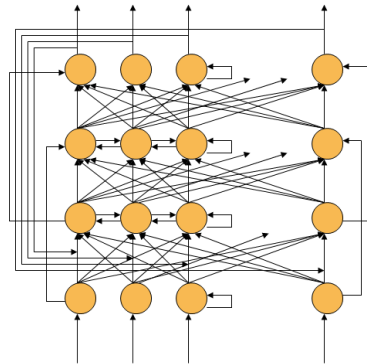
3. Métodos y tecnologías para el análisis de los datos a gran escala.

- ❑ Métodos estadísticos: algoritmos de regresión.
- ❑ **Aprendizaje automatizado (*machine learning*)**.

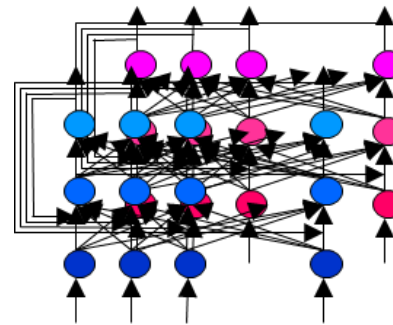
Algoritmo KNN



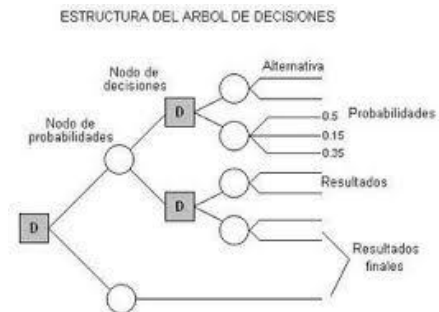
Redes neuronales artificiales



Redes neuronales de aprendizaje profundo (*deep learning*)



Árboles de decisión



UEA Datos a Gran Escala

Contenido sintético detallado:

4. Casos de Estudio

- ☐ Problemas de clasificación.
- ☐ Problemas de predicción.

Problemas de predicción

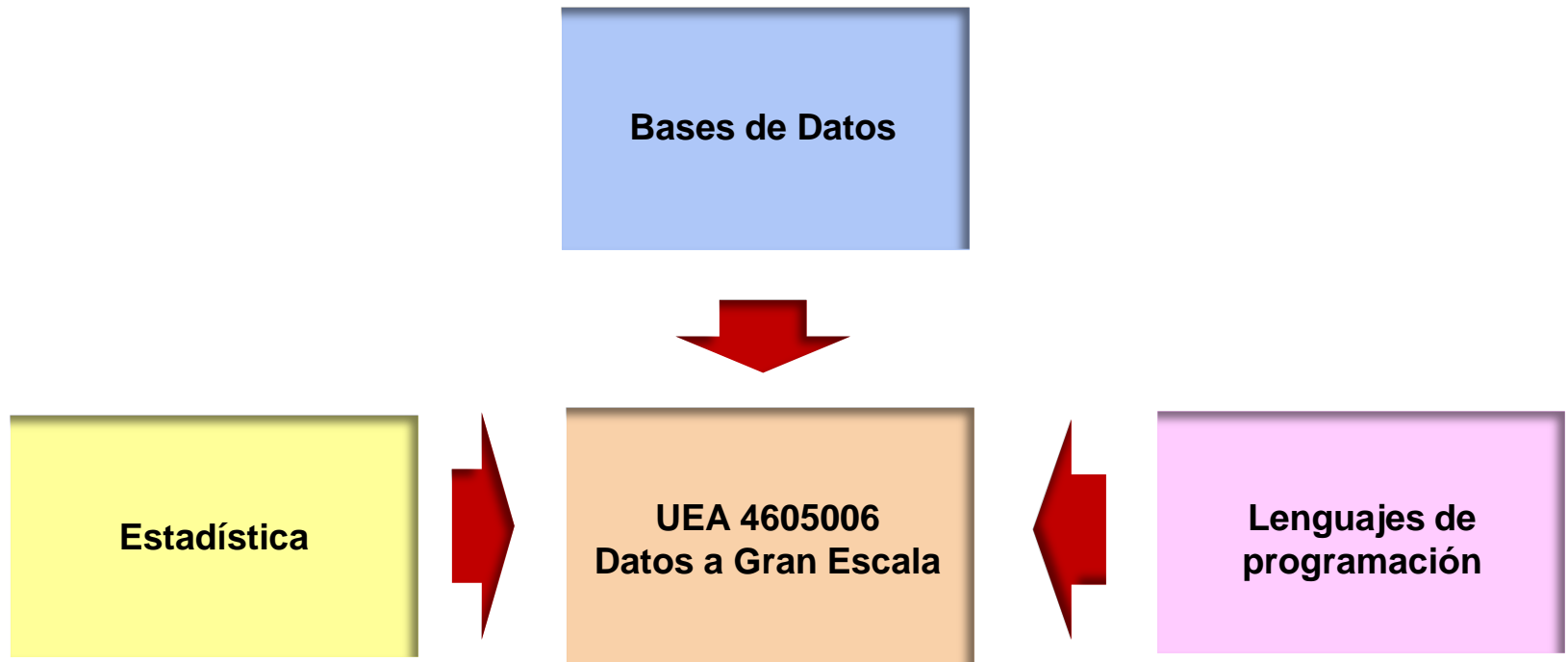


Problemas de clasificación



UEA Datos a Gran Escala

Conocimientos previos deseables



UEA Datos a Gran Escala

Herramientas para el Análisis de Datos a Gran Escala: ***IBM SPSS Modeler***

- Para sacarle el mayor provecho a los datos masivos es imprescindible basarse en una metodología, modelo o enfoque de minería de datos, así como en una fuerte herramienta computacional, que guíen y soporten de una forma altamente organizada y estructurada la preparación, análisis y presentación de estos grandes volúmenes de datos.
- En este sentido, el paquete de cómputo *IBM SPSS Modeler* (<https://www.ibm.com/mx-es/products/spss-modeler>) proporciona un ambiente integrado de trabajo para aplicar múltiples herramientas de minería de datos para el análisis exploratorio de los datos, procesamiento de los datos, modelado, evaluación y presentación de los resultados.

UEA Datos a Gran Escala

Herramientas para el Análisis de Datos a Gran Escala: *IBM SPSS Modeler*



IBM

Buscar

SPSS Modeler Detalles Precios Soporte Recursos

Pruebe sin costo

IBM SPSS Modeler

Impulse el retorno de la inversión y acelere la generación de valor con una herramienta de ciencia de datos intuitiva, con la función de arrastrar y soltar

Pruebe SPSS Modeler sin costo alguno

Planifique una reunión

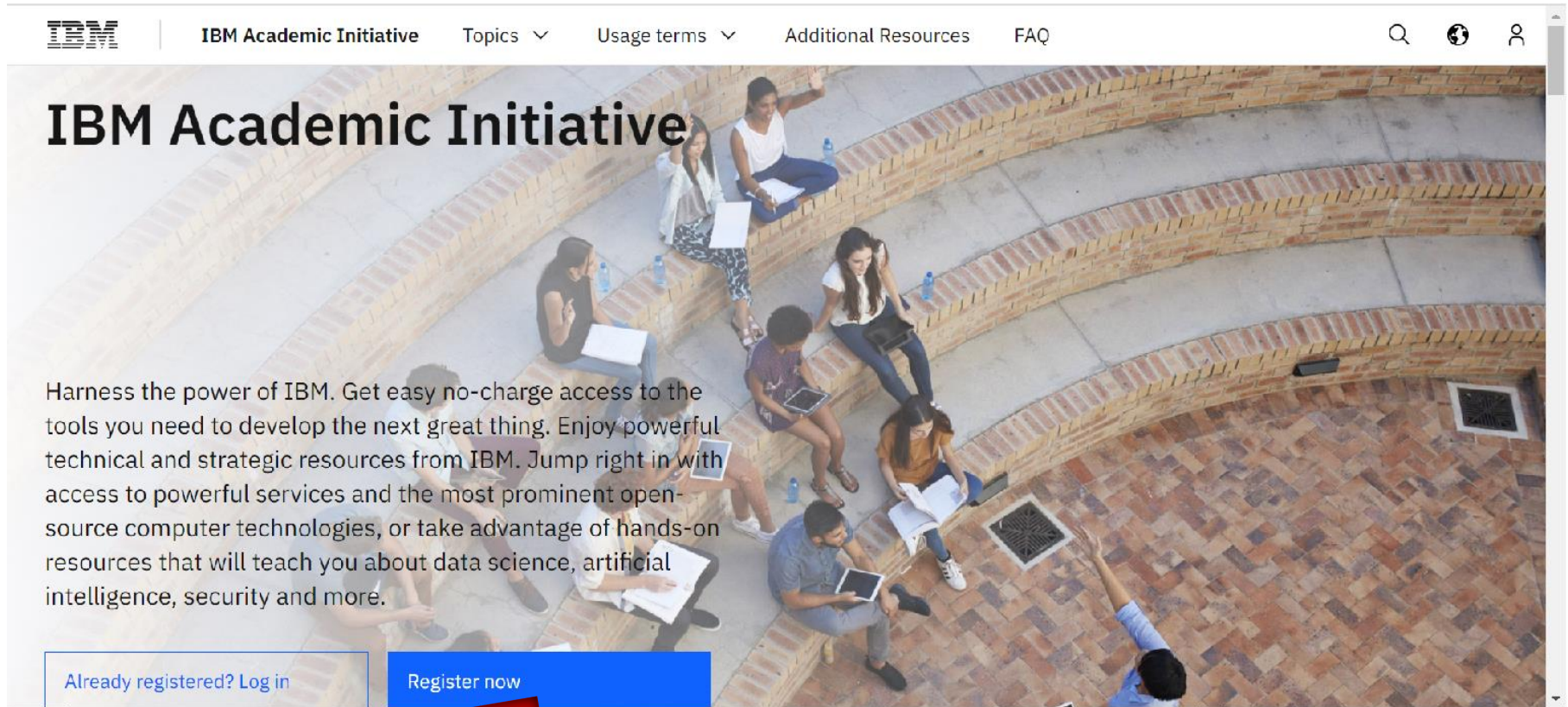
Obtenga la edición para estudiantes

Hablemos

<https://www.ibm.com/mx-es/products/spss-modeler>

UEA Datos a Gran Escala

Herramientas para el Análisis de Datos a Gran Escala: *IBM SPSS Modeler*

The image is a screenshot of the IBM Academic Initiative website banner. At the top, there is a navigation bar with the IBM logo on the left, followed by 'IBM Academic Initiative', 'Topics' with a dropdown arrow, 'Usage terms' with a dropdown arrow, 'Additional Resources', and 'FAQ'. On the right side of the navigation bar are icons for search, a globe, and a user profile. The main banner features a large, curved brick wall with several students sitting on it, some using laptops and tablets. The text 'IBM Academic Initiative' is prominently displayed in the upper left of the banner. Below this, a paragraph describes the initiative: 'Harness the power of IBM. Get easy no-charge access to the tools you need to develop the next great thing. Enjoy powerful technical and strategic resources from IBM. Jump right in with access to powerful services and the most prominent open-source computer technologies, or take advantage of hands-on resources that will teach you about data science, artificial intelligence, security and more.' At the bottom left of the banner, there are two buttons: 'Already registered? Log in' and 'Register now'. A large red arrow points from the 'Register now' button towards the URL at the bottom of the slide.

IBM Academic Initiative

Harness the power of IBM. Get easy no-charge access to the tools you need to develop the next great thing. Enjoy powerful technical and strategic resources from IBM. Jump right in with access to powerful services and the most prominent open-source computer technologies, or take advantage of hands-on resources that will teach you about data science, artificial intelligence, security and more.

Already registered? Log in Register now

<https://www.ibm.com/mx-es/products/spss-modeler>

UEA Datos a Gran Escala

Herramientas para el Análisis de Datos a Gran Escala:
IBM SPSS Modeler

Guía para descargar la herramienta *IBM SPSS Modeller*



<https://www.youtube.com/watch?v=uax3Z2fGnMs>

UEA Datos a Gran Escala

Otras Herramientas que Soportan de Forma Parcial el Análisis de Datos a Gran Escala

- Excel
- IBM SPSS Statistics (<https://www.ibm.com/mx-es/products/spss-statistics>)
- GNU PSPP (<https://www.gnu.org/software/pspp/>)
- TensorFlow (<https://www.tensorflow.org/>)
- Librerías de Python para Machine Learning

UEA Datos a Gran Escala

Repositorios de Datos a Gran Escala

The image is a screenshot of the University of California, Irvine (UCI) website homepage. At the top, there is a dark navigation bar with links for 'About', 'Admissions', 'Academics', 'Research', and 'Community'. To the right of these links are buttons for 'Register to Vote' and 'Give to UCI'. Below this is a blue banner featuring the UCI logo and the text 'University of California, Irvine'. To the right of the logo is a search bar with the placeholder text 'Search...' and two buttons labeled 'Web' and 'People'. Below the blue banner is a dark blue navigation bar with links for 'Top', 'The Buzz', 'News', 'Who We Are', 'Visit', 'Events', 'Arts & Athletics', 'Initiatives', 'Health', 'Alumni & Giving', and 'Resources'. Below this is an orange banner with the text 'Visit [UCI Forward](#) for the latest COVID-19 policies and campus information.' and a yellow button labeled 'UCI Forward'. The main content area features a large blue banner with the text 'Joe C. Wen and family donate \$20 million to support the new UCI Health Center for' and a photograph of the 'UCI Health Joe C. Wen & Family Center for Advanced Care' building.

UCI University of California, Irvine

Search... Web People

Top The Buzz News Who We Are Visit Events Arts & Athletics Initiatives Health Alumni & Giving Resources

Visit [UCI Forward](#) for the latest COVID-19 policies and campus information.

UCI Forward

Joe C. Wen and family donate \$20 million to support the new UCI Health Center for

UCI Health Joe C. Wen & Family Center for Advanced Care

<https://uci.edu/>

UEA Datos a Gran Escala

Repositorios de Datos a Gran Escala

- UCI Machine Learning Repository
(<https://archive.ics.uci.edu/ml/index.php>)



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)

☒ Repository ☐ Web

Search

[View ALL Data Sets](#)

Welcome to the UC Irvine Machine Learning Repository!





We currently maintain 622 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).



Latest News:

09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!
04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010: [Note](#) from donor regarding Netflix data
10-16-2009: Two new data sets have been added.
09-14-2009: Several data sets have been added.
03-24-2008: New data sets have been added!
06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Newest Data Sets:

06-05-2021:  [Average Localization Error \(ALE\) in sensor node localization process in WSNs](#)
05-25-2021:  [9mers from cullpdb](#)
05-18-2021:  [TamilSentiMix](#)
05-02-2021:  [Accelerometer](#)

Most Popular Data Sets (hits since 2007):

4751836:  [Iris](#)
2519466:  [Adult](#)
2041100:  [Dry Bean Dataset](#)
1951123:  [Wine](#)

UEA Datos a Gran Escala

Repositorios de Datos a Gran Escala

- UCI Machine Learning Repository
(<https://archive.ics.uci.edu/ml/index.php>)



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems






[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)

☒ Repository ☐ Web 

[View ALL Data Sets](#)

Browse Through: 622 Data Sets

Table View [List View](#)

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (466) Regression (151) Clustering (121) Other (56)	 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Attribute Type Categorical (38) Numerical (422) Mixed (55)	 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Data Type Multivariate (480) Univariate (30) Sequential (59) Time-Series (126) Text (69) Domain-Theory (23) Other (21)	 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Area Life Sciences (147) Physical Sciences (57)	 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
	 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998

UEA Datos a Gran Escala

Modalidades de conducción:

Clase teórico-práctica a cargo del profesor con participación activa del alumno.

Como estrategia de enseñanza el profesor hará exposiciones de los temas en el aula. Promoverá la aplicación de métodos y herramientas para la extracción e inferencia de información a partir de grandes volúmenes de datos

El profesor diseñará experiencias de aprendizaje por problemas, con nivel de complejidad incremental, tanto en el aula como en el laboratorio. El alumno analizará los problemas planteados y aplicará los conceptos, métodos y tecnologías de los datos a gran escala para darles solución.

UEA Datos a Gran Escala

Habilidades transversales:

- ☐ (Ht0) Lenguaje disciplinar: reforzará y aprenderá nuevos conceptos relacionados con los datos a gran escala.
- ☐ (Ht2) Trabajar armónicamente en equipo: deberán poder transmitir sus ideas para la solución de problemas y recibir retroalimentación a las mismas.
- ☐ (Ht3) Comunicarse eficazmente en forma oral y escrita en español: presentará al profesor, en forma oral y escrita, informes de los trabajos realizados.
- ☐ (Ht4) Comprender perfectamente textos técnicos en español: el profesor deberá proporcionar lecturas sobre temas relacionados con el contenido sintético. Es recomendable que las lecturas ayuden al alumno a encontrar soluciones a los temas tratados en esta UEA.
- ☐ (Ht5) Comprender textos técnicos en inglés: el profesor deberá proporcionar lecturas asociadas al contenido sintético, para que posteriormente el alumno explique en español lo que entendió de dichas lecturas.

UEA Datos a Gran Escala

Habilidades Disciplinarias:

- ❑ (H1) Abstracción, como la habilidad para conceptualizar un problema que permita plantear una solución al mismo: identificar los principales componentes tecnológicos en un sistema de datos a gran escala.
- ❑ (H4) Aplicar los conceptos, métodos y tecnologías en el procesamiento y análisis de los datos a gran escala.
- ❑ (H5) Desarrollar la capacidad para tomar decisiones

UEA Datos a Gran Escala

Actitudes:

- ☐ (A1) Liderazgo en equipos de trabajo multidisciplinario.
- ☐ (A3) Disciplina para aplicar los conocimientos adquiridos.
- ☐ (A5) Voluntad de mantenerse actualizado en su área de trabajo.
- ☐ (A6) Responsabilidad y ética en su desempeño profesional.
- ☐ (A7) Conciencia de la realidad social y responsabilidad ecológica.
- ☐ (A8) Adaptación a nuevos o diferentes entornos tecnológicos.

UEA Datos a Gran Escala

Modalidades de evaluación: Evaluación global

Se ponderarán las siguientes actividades a criterio del profesor:

- ☐ Tareas individuales y en equipo, que incluyen prácticas de laboratorio.
- ☐ Evaluaciones periódicas.
- ☐ Evaluación terminal.
- ☐ Documentación del proyecto.
- ☐ Participación en el proceso de argumentación, tanto en las sesiones de teoría como en las de práctica.
- ☐ Evaluación de las lecturas de textos en inglés, mediante reportes escritos o de forma oral en español.

UEA Datos a Gran Escala

Modalidades de evaluación: Evaluación de recuperación

- ☐ El alumno deberá presentar una evaluación objetiva que contemple los contenidos de la unidad de enseñanza-aprendizaje.
- ☐ No requiere inscripción previa a la UEA.

UEA Datos a Gran Escala

Bibliografía necesaria o recomendable:

1. Joyanes Aguilar, L. Big data. Análisis de grandes volúmenes de datos en organizaciones. Alfaomega, 2013.
2. Marr, B. Big data in practice. How 45 successful companies used big data analytics to deliver extraordinary results. Wiley, 2016.
3. Marz, N., Warren, J. Big data: Principles and best practices of scalable realtime data systems. Manning Publications, 2015.
4. Mayer-Schönberger, V., Cukier, K. Big data. La revolución de los datos masivos. Turner Noema, 2013.
5. Mayer-Schönberger, V., Cukier, K. Big data: A revolution that will transform how we live, work, and think. John Murray, 2017.
6. Sinha, S. Making big data work for your business. Impactt Publishing, 2014.