

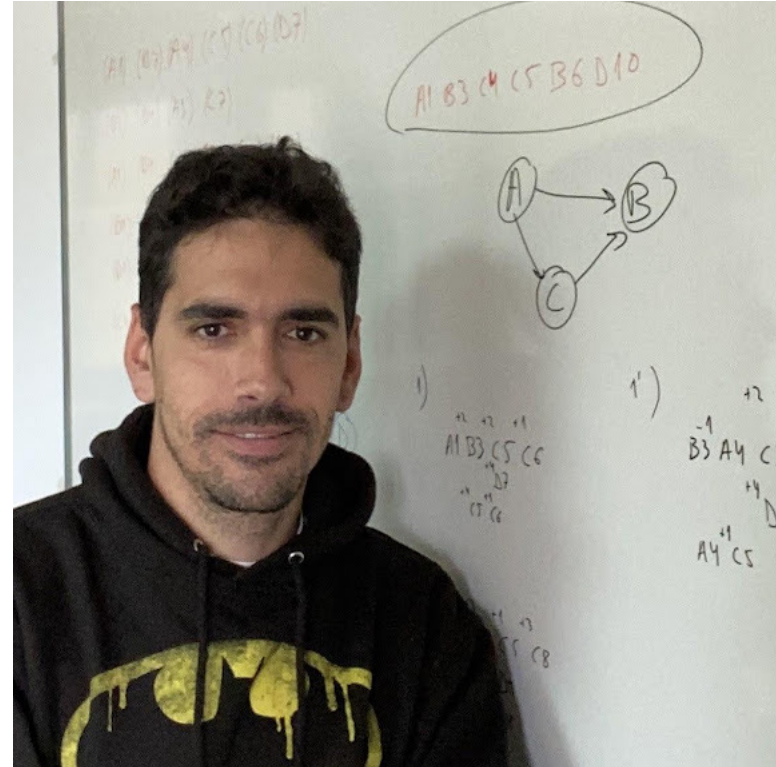
A stylized sunburst graphic in shades of purple and blue, located in the top-left corner of the slide.

# Métodos descriptivos supervisados

Máster en Ciencia de Datos

## José María Luna

recibió el título de Doctor en Ciencias de la Computación en 2014, por la Universidad de Granada. Su carrera investigadora fue inicialmente subvencionada por el Ministerio de Educación de España bajo el programa FPU (predoctoral) y el programa Juan de la Cierva (postdoctoral). Actualmente es profesor de la Universidad de Córdoba en el departamento de Informática y Análisis Numérico. Dr. Luna ha sido autor de los libros monográficos "Pattern Mining with Evolutionary Algorithms" y "Supervised Descriptive Pattern Mining", ambos publicados por la editorial Springer. Además, ha publicado más de 30 artículos en revistas científicas de alto impacto. Actualmente tiene un total de 2260 citas en Google Scholar y un índice H de 24. Su investigación es llevada a cabo en el grupo de investigación *Knowledge Discovery and Intelligent Systems*, donde investiga temas relativos a computación evolutiva, minería de patrones, reglas de asociación y sus aplicaciones.



# Métodos descriptivos supervisados (Parte 2)

*Descubrimiento de subgrupos*

# Descubrimiento de subgrupos

- Introducción

- *Subgroup Discovery* es una técnica de minería de datos cuyo objetivo es el descubrimiento de relaciones de interés entre diferentes objetos respecto a una propiedad de interés para el usuario
- Los patrones extraídos se representan en forma de reglas llamadas subgrupos
- A diferencia de técnicas como Class Association Rules:
  - El descubrimiento de subgrupos no se centra únicamente en soporte y confianza
  - Las técnicas de reglas asociación se focalizan más en extraer cualquier regla que exista, mientras que el descubrimiento de subgrupos proporciona un subconjunto reducido que es de interés para el usuario en base a unas métricas

# Descubrimiento de subgrupos

- Definición formal

*En subgroup Discovery, asumimos que tenemos una población de individuos (objetos, clientes, etc) y una propiedad de esos individuos sobre la que estamos interesados en analizar. La tarea de subgroup Discovery consiste, por tanto, en descubrir subgrupos de esa población que son estadísticamente más interesantes. Es decir, subconjuntos que son lo más grandes posibles y que tienen una serie de características estadísticas respecto a la propiedad de interés.*

$$Cond \rightarrow Target_{value}$$

# Descubrimiento de subgrupos

- Ejemplos

- *Edad < 25, Pais = Alemania -> Rico*

Representa un subgrupo de personas que viven en Alemania, que tienen menos de 25 años y que tienen una probabilidad de ser ricos inusualmente alta respecto al resto de la población

- *Edad > 60, Sexo = Mujer -> Normal*

Representa un subgrupo de mujeres que tienen más de 60 años y que tienen una probabilidad de tener una economía normal inusualmente alta respecto al resto de la población

# *Descubrimiento de subgrupos*

- Tipos de variables objetivo
  - Binaria
    - La variable objetivo sólo tiene dos valores. La tarea consiste en obtener subgrupos de interés respecto a cada uno de los dos posibles valores
  - Nominal
    - La variable objetivo puede tomar un número indeterminado de valores. El análisis es similar al anterior, obtener subgrupos de interés respecto a cada uno de los posibles valores
  - Numérica
    - Es el tipo más complejo ya que la variable objetivo tiene que ser estudiada de diferentes formas: dividir la variable objetivo en rangos respecto a la media, discretizar la variable en intervalos, búsqueda de desviaciones respecto a la media, etc.

# *Descubrimiento de subgrupos*

- Medidas de calidad
  - Complejidad
    - Relativas a la interpretabilidad de los subgrupos. Trata de simplificar el conocimiento extraído por los subgrupos
  - Generalidad
    - Cuantifican la calidad de las reglas individuales de acuerdo con los ejemplos de interés cubiertos por la regla
  - Precisión
    - Muestran la exactitud de los subgrupos
  - Interés
    - Tratar de seleccionar y ordenar los resultados en base al interés de los mismos respecto al usuario



# *Descubrimiento de subgrupos*

- Medidas de calidad
  - Complejidad
    - Relativas a la interpretabilidad de los subgrupos. Trata de simplificar el conocimiento extraído por los subgrupos
      - Número de reglas
      - Número de variables. Para un conjunto de reglas, se calcula como la media de variables

Las fórmulas pueden ser consultadas en: [https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324\\_2011-Herrera-KAIS.pdf](https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324_2011-Herrera-KAIS.pdf)

# *Descubrimiento de subgrupos*

- Medidas de calidad
  - Generalidad
    - Cuantifican la calidad de las reglas individuales de acuerdo con los ejemplos de interés cubiertos por la regla
      - Coverage. Mide el porcentaje de ejemplos cubiertos (condición de la regla) respecto al total
      - Support. Mide la frecuencia de ejemplos correctamente clasificados (condición de la regla y variable objetivo)

Las fórmulas pueden ser consultadas en: [https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324\\_2011-Herrera-KAIS.pdf](https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324_2011-Herrera-KAIS.pdf)

# *Descubrimiento de subgrupos*

- Medidas de calidad
  - Precisión
    - Muestran la exactitud de los subgrupos
      - Confidence. Mide la frecuencia relativa de los ejemplos que satisface la regla completa respecto a los que satisfacen sólo la condición
      - Precision measure Qc. Función lineal entre los verdaderos y falsos positivos
      - Precision measure Qg. Mide la relación entre el numero de ejemplos clasificados correctamente y la inusualidad de la distribución del subgrupo.

Las fórmulas pueden ser consultadas en: [https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324\\_2011-Herrera-KAIS.pdf](https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324_2011-Herrera-KAIS.pdf)

# *Descubrimiento de subgrupos*

- Medidas de calidad

- Interés

- Tratar de seleccionar y ordenar los resultados en base al interés de los mismos respecto al usuario
      - Interes. Mide el interés de una regla dado por el antecedente y el consecuente usando medidas de ganancia de información
      - Novelty. Permite detectar subgrupos inusuales
      - Significance. Mide la significancia del descubrimiento obtenido por la regla

Las fórmulas pueden ser consultadas en: [https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324\\_2011-Herrera-KAIS.pdf](https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324_2011-Herrera-KAIS.pdf)

# *Descubrimiento de subgrupos*

- Otras medidas de calidad
  - Este grupo de medidas combinan otras medidas como generalidad, interés y precisión
    - Sensitivity. Combina precisión y generalidad. Se calcula como la proporción de ejemplos que han sido correctamente clasificados
    - Specificity. Combina precisión y generalidad. Se calcula como la proporción de ejemplos negativos incorrectamente clasificados
    - Unusualness. Combina precisión, interés y generalidad Se calcula como un balance entre las medidas de Coverage y Confidence

Las fórmulas pueden ser consultadas en: [https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324\\_2011-Herrera-KAIS.pdf](https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1324_2011-Herrera-KAIS.pdf)

# Descubrimiento de subgrupos

- Clasificación de algoritmos de *subgroup discovery*
  - Extensiones de algoritmos de clasificación (primeros algoritmos)
  - Extensiones de algoritmos de reglas de asociación
  - Algoritmos evolutivos

# Descubrimiento de subgrupos

- Clasificación de algoritmos de *subgroup discovery*
  - Extensiones de algoritmos de clasificación
    - EXPLORA (variable objetivo categórica)
    - MIDOS (variable objetivo binaria)
    - SubgroupMiner (variable objetivo categórica)
    - SD (variable objetivo categórica)
    - CN2-SD (variable objetivo categórica)
    - RSD (variable objetivo categórica)

# Descubrimiento de subgrupos

- Clasificación de algoritmos de *subgroup discovery*
  - Extensiones de algoritmos de reglas de asociación
    - Apriori-SD (variable objetivo categórica)
    - SD4TS (variable objetivo categórica)
    - SD-MAP (variable objetivo binaria)
    - SD-MAP\* (variable objetivo numérica)
    - DpSubgroup (variable objetivo categórica y binaria)
    - MergeSD (variable objetivo numérica)
    - IMR (variable objetivo categórica)



# Descubrimiento de subgrupos

- Clasificación de algoritmos de *subgroup discovery*
  - Algoritmos evolutivos
    - SDIGA (variable objetivo categórica)
    - MESDIF (variable objetivo categórica)
    - NMEEF-SD (variable objetivo categórica)
    - G3P-SD (variable objetivo categórica)

A stylized sunburst graphic in shades of purple and blue, located in the top-left corner of the slide. It features a semi-circle on the left with several rays extending outwards to the right.

¡Gracias!

**UCO**  
ONLINE

A decorative horizontal bar at the bottom of the slide, consisting of alternating yellow and red rectangular segments.