Caracterización estructural de formaciones astronómicas, utilizando un enfoque de grafos, para reconocimiento de cumulos estelares en galaxias cercanas

Universidad Tecnológica Nacional
Facultad Regional Córdoba

Proyecto de Investigación y Desarrollo (PID)

Tesista: Esp. Ing. Martin Casatti
Director: Dr. Oscar Medina
CoDirectors: Mgr. Cynthia Corso

1. Datos de la investigación

Estado actual de conocimiento del tema

Astroinformática

La astroinformática es una disciplina multidisciplinaria que incluye prácticas que abarcan tanto la astronomía y la astrofísica, como la informática, la ciencia de datos, la estadística y los modelos de simulación[8].
Debido al crecimiento exponencial de la cantidad de datos relevados por observatorios automatizados, reforzada por la aparición de cámaras con cada vez mayores resoluciones y capacidades de almacenamiento de alta velocidad y capacidad, se ha hecho cada vez más necesario el procesamiento automático de dicha información a los fines de obtener datos pre procesados, catalogados y pre analizados, para su uso por parte de los astrónomos.
Considerando que la tasa de relevamiento de datos astronomicos es de tal envergadura que imposibilita el análisis manual de los mismos, la astroinformática ha dejado de ser un mero auxiliar de la astronomía observacional para convertirse en una herramienta crucial en la que descansa la promesa de los próximos grandes descubrimientos en el ámbito astronómico[9].

Clusters estelares

Las agrupaciones estelares, también denominados cúmulos o clusters, han sido objetos reconocidos desde hace tiempo como laboratorios importantes para la investigación astrofísica, siendo muy útiles en varios aspectos, entre los que se pueden destacar los siguientes:

- Contienen muestras estadísticamente significativas de estrellas de aproximadamente la misma edad, con composiciones químicas similares, un amplio rango de masas estelares y localizadas en un volumen relativamente pequeño del espacio, haciéndolas un conjunto ideal para el análisis de características comunes y determinación de los patrones que rigen su surgimiento [2].

- En relación con el proceso de formación estelar, los cúmulos jóvenes permiten esclarecer la forma y las escalas de tiempo en las que estos mecanismos están activos, así como también permiten analizar su dependendencia de los distintos ambientes interestelares de la Vía Láctea o de otras galaxias [3].

Los trabajos mencionados se han focalizado en mejorar el conocimiento de nuestra propia Galaxia (y de las Nubes de Magallanes[5]), pero actualmente hay varios factores que incrementan de forma importante tanto la cantidad de objetos a investigar cómo la metodología para hacerlo.
En la actualidad existe una gran cantidad de información de las galaxias cercanas (a varios Mpc[1]) debido, en gran parte, a que el Telescopio Espacial Hubble (HST)

[1] Megaparsec, medida de distancia, aproximadamente 3.26 millones de años luz

ha permitido obtener datos con alta resolución espacial utilizando varias cámaras de campo amplio (WFPC2; ACS) [4].
Se cuenta con una enorme cantidad de datos proveniente de las varias observaciones continuas que se están realizando y que se proyectan realizar en modo "survey"[2] (p.e. VVV[3] o LSST[4]) que necesitan ser estudiados con métodos automáticos.
En este ámbito, los algoritmos de reconocimiento automático de patrones, están teniendo una importante revisión y desarrollo tal como se puede apreciar en el análisis comparativo de Schmeja (2011)[1].
Tal como se desprende de esa publicación, estos algoritmos se basan en analizar sólo las posiciones espaciales para encontrar a los sistemas estelares por sobre-densidades contra el fondo estelar o por su equivalente relacionado con la distribución de distancias entre estrellas.
Cabe hacer notar que ya se han desarrollado varios algoritmos que han sido aplicados con éxito en otros campos científicos. Entre estos se destacan algoritmos como "K-mean", "Birch", "Spectral Clustering", "Dbscan", etc.[6]

Características estelares

Existen ciertas características que son distintivas para la caracterización de los distintos tipos de cuerpos celestes. Atributos tales como la masa, la distancia a la Tierra, su composición química, velocidad o inclinación con respecto al eje galáctico, entre otros. Gran parte de la astronomía óptica se concentra en aquellos atributos que son comunes a las estrellas que componen cúmulos estelares, tales como su edad, composición química, densidad y propiedades orbitales. Asimismo se incluyen datos que pueden variar de uno a otro componente pero que sirven para definir la estructura del cluster como un todo o para analizar algunas características atípicas de los mismos. En este sentido se puede considerar la masa o el estado evolutivo, también llamado análisis de secuencia principal, el cual utiliza los conocidos diagrams de Herzprung-Russell o diagramas H-R, por ejemplo[10, 11].
Recientemente se han producido grandes avances técnicos que posibilitan una mejor detección de objetos no visibles pero que irradian frecuencias en el espectro de infrarrojos. El más reciente instrumento en utilizar estas capacidades es el observatorio espacial James Webb, el cual está optimizado para la detección de infrarrojos y cuenta con un espejo reflector de 25 m^2, fué lanzado el 25 de diciembre de 2021 y genera 60 GB de datos diariamente, lo que se prolongará durante toda su vida útil[12].
Por otra parte, desde el punto de vista de la captura de información en el espectro visible, el máximo exponente es el instrumento del observatorio Vera C. Rubin, alojado físicamente en Chile. El instrumento del observatorio es una cámara con una lente de 8.4 metros y una resolución de 3.2 Gigapixels, capaz de realizar un mapeo de todo

[2] Técnica que consiste en realizar un mapeo sistemático de una porción determinada de la esfera celeste sin concentrarse de manera puntual en ningún objeto.
[3] https://vvvsurvey.org/
[4] https://www.lsst.org/

el cielo visible cada cuatro noches. Se estima que producirá 20 TB (terabytes) de información cada noche, durante una vida útil de al menos 10 años [7].

Los grafos como mecanismo de representación del conocimiento

Actualmente se deben tener en cuenta al menos cuatro criterios fundamentales a la hora de diseñar un sistema de representación del conocimiento en cualquier dominio dado [13]:

Adecuación Representacional: Habilidad para representar todas las clases de conocimiento que son necesarias en el dominio.

Adecuación Inferencial: Habilidad de manipular estructuras de representación de tal manera que devengan o generen nuevas estructuras que correspondan a nuevos conocimientos inferidos de los anteriores.

Eficiencia Inferencial: Capacidad del sistema para incorporar información adicional a la estructura de representación, llamada metaconocimiento, que puede emplearse para focalizar la atención de los mecanismos de inferencia con el fin de optimizar los cómputos.

Eficiencia en la Adquisición: Capacidad de incorporar fácilmente nueva información. Idealmente el sistema por sí mismo deberá ser capaz de controlar la adquisición de nueva información y su posterior representación.

Para afrontar la cuestión de cuál es el modelo más apropiado para la representación de la información se debe tener en cuenta la natural heterogeneidad de los datos que se pueden extraer de fuentes diversas (catálogos estelares, surveys, observaciones específicas, etc).
Si bien el formato y estructura de dichos documentos está en gran medida estandarizada, también es real que las relaciones entre las entidades que componen un modelo astronómico pueden ser muy variables y variadas.
Existen multitud de características que se pueden asociar entre las observaciones, lo que lleva, en última instancia, a múltiples relaciones N-N (relaciones de todos con todos) lo que puede llevar rápidamente a un crecimiento exponencial de la cantidad de relaciones que un modelo de datos debe manejar.
Si bien en la actualidad el modelo de datos relacional es un modelo probado, conocido y estable, no es el más recomendable en este escenario debido al aumento exponencial de los tiempos de búsqueda al contar con multitud de relaciones de tipo N-N [14].
Desde hace un tiempo y con el advenimiento de disciplinas como el Data Minning, el Data Warehousing y algunas aplicaciones de Inteligencia Artificial y Machine Learning, se están impulsando modelos alternativos que no sufran las limitaciones del modelo relacional y que permitan gestionar eficientemente grandes cantidades de datos heterogéneos [15].

El equipo de investigación ha realizado con anterioridad experiencias con la utilización de bases de datos de grafos y las mismas se consideran, hasta el momento, una de las alternativas más prometedoras.

Definición de grafo: En matemáticas y en ciencias de la computación se define a un grafo como un conjunto de objetos denominados vértices (también pueden mencionarse como nodos), relacionados por enlaces llamados aristas o arcos. Estas relaciones establecen una asociación binaria entre dos nodos, la cual puede ser dirigida, en uno u otro sentido, o no. Los grafos pueden tener información asociada tanto a los nodos como a los arcos, denominándose en este caso grafos etiquetados [13].

Las bases de datos de grafos: Se denomina base de datos de grafos a un sistema de almacenamiento de información que representa de manera eficiente el modelo de grafos, compuesto de nodos y arcos.

Un punto importante a tener en cuenta es el concepto de ‘‘impedancia cognitiva’’, el cual representa el desfasaje conceptual que se produce entre los conceptos modelados y su representación en un formato de almacenamiento determinado [16]. Una gran impedancia hace que sea difícil representar de manera física los conceptos modelados y da como resultado almacenamientos complejos y algoritmos de recuperación de información poco eficientes.
Las bases de datos de grafos tienen una reducida impedancia, lo que permite representar de una manera directa y natural los conceptos modelados, permitiendo relaciones directas e intuitivas entre las entidades que componen la base de datos.

Grado de Avance

Los integrantes del grupo de investigación cuentan con experiencia en el uso de las tecnologías implicadas en el presente proyecto y en las temáticas abordadas.
El director, Dr. Oscar Medina, tiene una amplia trayectoria en el estudio de patrones y modelado conceptual, mientras que la co-directora, Mg. Cynthia Corso, cuenta con experiencia en las áreas de Big Data, Data Mining, reconocimiento de patrones y bases de datos de grafos.
El Esp. Ing. Martin Casatti ha trabajado en múltiples proyectos asociados al modelado de datos utilizando grafos, con aplicaciones en reconocimiento de estructuras semánticas, de relaciones cienciométricas y de modelado de información. El Ing. Casatti ha sido el responsable del diseño conceptual de ambos modelos de datos así como del diseño e implementación de algorimos de análisis de dichas bases de datos. Actualmente se encuentra llevando adelante su tesis doctoral, con el tema ‘‘Detección de cúmulos estelares en galaxias cercanas utilizando técnicas de Machine Learning y algoritmos de aplicación en redes sociales’’ en la que preve utilizar un enfoque basado en grafos para modelar las estructuras de clusters a fin de poder aplicar sobre las mismas los algoritmos de Machine Learning.

El Ing. Federico Benito se ha desempeñado como becario alumno y como becario BINID en proyectos en los que se modela la respuesta de examenes escritos en un modelo de grafos para su comparación las respuestas propuestas por docentes. también modeladas como grafos, para posibilitar la corrección de examenes con respuestas escritas en lenguaje natural, por medio de la comparativa de las estructuras de ambos grafos. El Ing. Benito ha implementado rutinas de acceso y consulta a bases de datos de grafos y otras funcionalides relacionadas.

2. Objetivos

Objetivo principal

1. El presente trabajo tiene como finalidad caracterizar formaciones astronómicas asociadas a clusters estelares desde un punto de vista físico y estructural utilizando grafos como modelo de soporte.

Objetivos secundarios

Se plantean asimismo los siguientes objetivos particulares a alcanzar:

1.1 Determinar los atributos necesarios para poder caracterizar una estructura astronómica, basado en datos obtenidos de los repositorios más utilizados de surveys astronómicos.

1.2 Elaborar un mecanismo confiable de importación de datos que permita plasmar los atributos mencionados en una estructura de gráficos minimizando la pérdida de información.

1.3 Elaborar un algoritmo que permita establecer relaciones entre nodos de forma tal que se expongan las relaciones físicas y astronómicas entre puntos de datos, plasmadas como relaciones en el grafo de soporte.

1.4 Detectar similitudes y diferencias entre las estructuras encontradas y otras estructuras de grafos conocidos y bajo estudio, tanto basados en datos reales como grafos generados de forma artificial para simulación de fenómenos conocidos.

3. Metodología

Para alcanzar los objetivos de la presente tesis, tanto a nivel general general como los objetivos particulares, se realizarán las siguientes actividades:

A1 Para alcanzar el objetivo **1.1** se construirá un entorno de pruebas, con un set de datos acotado y conocido, a partir del repositorio GAIA[5] y similares (ESO[6],

[5] https://gea.esac.esa.int/archive/
[6] http://archive.eso.org/cms.html

NASA[7], etc.), sobre el cual se realizará un estudio exploratorio para determinar el conjunto de atributos comunes que forman parte de las lecturas, a fin de construir un conjunto mínimo, uniforme, de datos a relevar.

A2 Contando con conjuntos de datos de repositorios validados de datos astronómicos, para alcanzar el objetivo **1.2** se desarrollarán rutinas de importación de datos que tengan en cuenta las características particulares de los distintos conjuntos de datos y que posibiliten extraer los atributos comunes a plasmar en el modelo de grafos. Se analizarán distintos algoritmos y lenguajes de implementación para poder determinar la mejor combinación en cuanto a rendimiento y confiabilidad.

A3 Una vez construido el modelo con los datos puntuales (nodos), para alcanzar el objetivo **1.3** se estudiarán distintas técnicas para establecer relaciones entre los mismos, a partir de diferentes atributos. Las mismas se probarán sobre el almacenamiento de datos, teniendo en cuenta la posibilidad de que se puedan volver atrás los cambios para aplicar luego otro criterio.

A4 Para alcanzar el objetivo **1.4** se compararán las estructuras generadas sobre los datos importados de repositorios astronómicos, contra estructuras ya detectadas, pre-existentes, del ámbito astronómico y de otros ámbitos, a fin de detectar posibles coincidencias o similitudes. Esto incluye también análisis comparativos con grafos artificiales creados con la finalidad de realizar simulaciones.

A5 En todo momento, se propone aportar al objetivo general **1** mediante la publicación de los resultados parciales en diversos papers y reuniones científicas, a fin de obtener realimentación y aportes de otros científicos e investigadores, de informática, astronomía y disciplinas afines.

4. Contribuciones del Proyecto

Contribuciones al avance científico, tecnológico, transferencia al medio

El proyecto contribuirá tanto a la comunidad astronómica, brindando un nuevo mecanismo para la caracterización y análisis de estructuras estelares, así como sirviendo de base para la implementación de algoritmos y técnicas de detección de patrones más avanzados y enfocados en características estructurales más que en atributos puntuales de las entidades bajo estudio.
Asimismo contribuirá a la comunidad de científicos de datos y de la computación en general, brindando un nuevo ámbito de aplicación de estructuras y técnicas comunmente aplicadas a otros ámbitos.
Se espera que la interacción de ambas disciplinas permita colaboraciones fructíferas y una realimentación cruzada que favorezca ambos campos de estudio.

[7] https://nssdc.gsfc.nasa.gov/astro/

Contribuciones a la formación de Recursos Humanos

Uno de los miembros del equipo, el Ing. Federico Benito, reforzará su labor en el ámbito de la investigación formal, a la vez que ampliará sus conocimientos ampliando el campo de aplicación a las ciencias astronómicas.
El integrante Tomás Alvarez realizará sus primeras experiencias en investigación formal, lo que será un importante agregado a su formación académica como Ingeniero.
El director y la co-directora, Dr. Oscar Medina y Mgr. Cynthia Corso, respectivamente, ampliarán su experiencia dirigiendo un proyecto en un ámbito interdisciplinario lo que potenciará sus conocimientos y habilidades.
El integrante Esp. Ing. Martín Casatti desarrollará, dentro del ámbito del proyecto, su tesis de Doctorado, titulada ‘‘Detección de cúmulos estelares en galaxias cercanas utilizando técnicas de Machine Learning y algoritmos de aplicación en redes sociales’’ obteniendo de esta forma una valiosa aplicación de los conceptos bajo estudio y pudiendo, a su vez, volcar su experiencia en las labores de investigación del presente proyecto.

5. Cronograma de Actividades 2025 / 2026

Para el período mencionado se propone el siguiente plan de trabajo (Figuras 1 y 2):

1. Mapeo sistemático de literatura relacionada a algoritmos de clustering
2. Construcción de infraestructura de almacenamiento y carga de datos de prueba
3. Aplicación de algoritmos de clustering astronómico y comparación con clusters conocidos
4. Estudio de atributos en grafos de redes sociales para su mapeo como atributos astronómicos
5. Publicación de resultados obtenidos

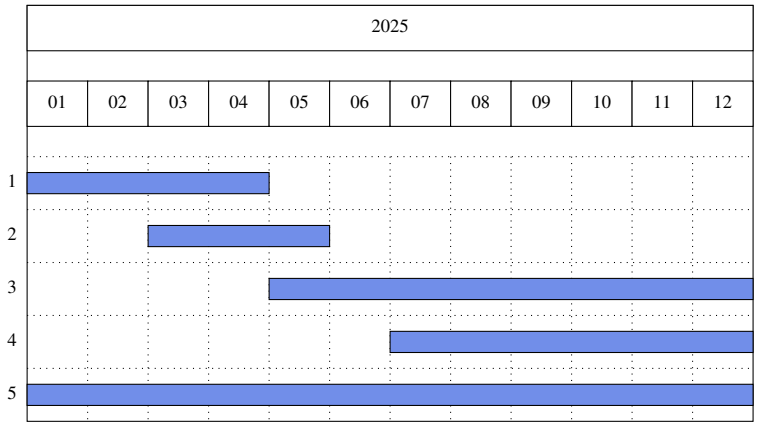

Figura 1: Año 2025

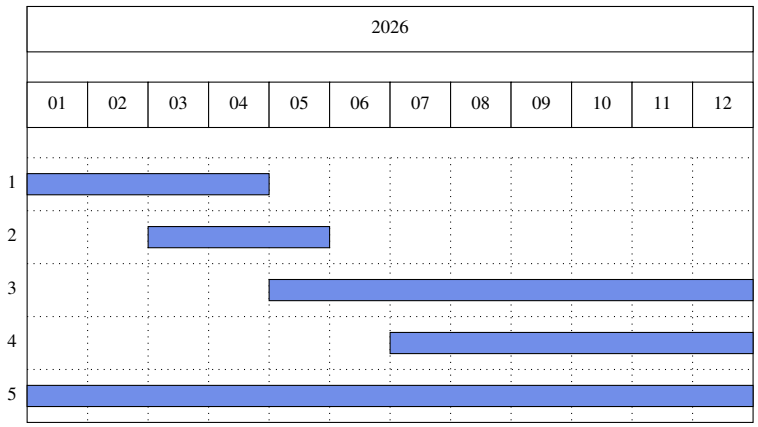

Figura 2: Año 2026

Referencias

1. Schmeja, S. Identifying star clusters in a field: A comparison of different algorithms. *Astronomische Nachrichten* **332,** 172-184 (2011).
2. Klessen, R. S. y Burkert, A. The Formation of Stellar Clusters: Gaussian Cloud Conditions. I. *The Astrophysical Journal Supplement Series* **128,** 287 (2000).
3. Fall, S. M. y Chandar, R. Similarities in populations of star clusters. *The Astrophysical Journal* **752,** 96 (2012).
4. Dalcanton, J. J. *et al.* The ACS nearby galaxy survey treasury. *The Astrophysical Journal Supplement Series* **183,** 67 (2009).
5. Vázquez, R. A. *et al.* Spiral structure in the outer galactic disk. I. The third galactic quadrant. *The Astrophysical Journal* **672,** 930 (2008).

6. Rodriguez, M. Z. *et al.* Clustering algorithms: A comparative approach. *PloS one* **14,** e0210236 (2019).
7. Telescope, L. S. S. Key Numbers. *Rubin Observatory*. https://www.lsst.org/scientists/keynumbers (jul. de 2021).
8. Borne, K. D. Astroinformatics: a 21st century approach to astronomy. *arXiv preprint arXiv:0909.3892* (2009).
9. Mahabal, A. *et al.* AstroInformatics: Recommendations for Global Cooperation. *arXiv preprint arXiv:2401.04623* (2024).
10. Fall, S. M. y Chandar, R. Similarities in populations of star clusters. *The Astrophysical Journal* **752,** 96 (2012).
11. Kalirai, J. S. y Richer, H. B. Star clusters as laboratories for stellar and dynamical evolution. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **368,** 755-782 (2010).
12. Gardner, J. P. *et al.* The james webb space telescope. *Space Science Reviews* **123,** 485-606 (2006).
13. Van Harmelen, F., Lifschitz, V. y Porter, B. *Handbook of knowledge representation* (Elsevier, 2008).
14. Kunii, H. S. *DBMS with graph data model for knowledge handling* en *Proceedings of the 1987 Fall Joint Computer Conference on Exploring technology: today and tomorrow* (1987), 138-142.
15. Vicknair, C. *et al. A comparison of a graph database and a relational database: a data provenance perspective* en *Proceedings of the 48th annual Southeast regional conference* (2010), 1-6.
16. Robinson, I., Webber, J. y Eifrem, E. *Graph databases: new opportunities for connected data* (.º'Reilly Media, Inc.", 2015).

Bibliografía adicional

El presente material bibliográfico se ha utilizado como material de estudio pero no se ha citado directamente en el texto.

. Sung, H., Bessell, M. S. y Lee, S.-W. UBVRI and Hα Photometry of the Young Open Cluster NGC 6231. *The Astronomical Journal* **115,** 734 (1998).
. West, D. B. *et al. Introduction to graph theory* (Prentice hall Upper Saddle River, 2001).
. Barnes, J. A. y Harary, F. Graph theory in network analysis. *Social networks* **5,** 235-244 (1983).
. Alharbi, A. y Alsubhi, K. Botnet detection approach using graph-based machine learning. *IEEE Access* **9,** 99166-99180 (2021).

. Menvielle, M. A. P., Groppo, M. A., Marciszack, M. M. y Casatti, M. *Text format written questions evaluation Methodology* en *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)* (2016), 1-4.

. Paz Menvielle, M. A. *et al. Caso de aplicación de representación del conocimiento utilizando grafos conceptuales en un sistema de corrección automatizado de exámenes* en *XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017).* (2017).

. Menvielle, M. A. P. *et al. Model and evaluation tool using graphs as knowledge base for the automated correction of exams in text format* en *2017 XLIII Latin American Computer Conference (CLEI)* (2017), 1-10.

. Paz Menvielle, M. A., Corso, C. L., Ligorria, K., Guzmán, A. y Casatti, M. *Análisis y detección de patrones en un grafo conceptual construido a partir de respuestas escritas en forma textual a preguntas sobre un tema específico* en *XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste).* (2018).

. Muñoz, R. M. *et al. Análisis cienciométrico de la producción en investigación científica y tecnológica* en *XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz).* (2020).

. Wang, C., Tang, W., Sun, B., Fang, J. y Wang, Y. *Review on community detection algorithms in social networks* en *2015 IEEE international conference on progress in informatics and computing (PIC)* (2015), 551-555.

. Kaur, R. y Singh, S. A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian informatics journal* **17,** 199-216 (2016).

. Boccaletti, S., Ivanchenko, M., Latora, V., Pluchino, A. y Rapisarda, A. Detecting complex network modularity by dynamical clustering. *Physical Review E* **75,** 045102 (2007).

. Kushwah, A. K. S. y Manjhvar, A. K. A review on link prediction in social network. *International Journal of Grid and Distributed Computing* **9,** 43-50 (2016).

. Lancichinetti, A. y Fortunato, S. Community detection algorithms: A comparative analysis. *Phys. Rev. E* **80,** 056117. https://link.aps.org/doi/10.1103/PhysRevE.80.056117 (5 nov. de 2009).

. Sowa, J. F. en *Encyclopedia of Cognitive Science* (John Wiley & Sons, Ltd, 2006). ISBN: 9780470018866. http://dx.doi.org/10.1002/0470018860.s00065.

. Sowa, J. F. en (eds. Nagle, T. E., Nagle, J. A., Gerholz, L. L. y Eklund, P. W.) 3-66 (Ellis Horwood, Upper Saddle River, NJ, USA, 1992). ISBN: 0-13-175878-0. http://dl.acm.org/citation.cfm?id=168857.168864.

. *All Quick Facts* [Online; accessed 4. Feb. 2023]. https://webbtelescope.org/quick-facts/all-quick-facts.

. Rossi, R. A. y Ahmed, N. K. *The Network Data Repository with Interactive Graph Analytics and Visualization* en *AAAI* (2015). https://networkrepository.com.

. Borissova, J. *et al.* New Galactic star clusters discovered in the VVV survey. *Astronomy & Astrophysics* **532,** A131 (2011).

. Tyson, J. A. Large synoptic survey telescope: overview. *Survey and Other Telescope Technologies and Discoveries* **4836,** 10-20 (2002).

. Jurić, M. *et al.* The LSST data management system. *arXiv preprint arXiv:1512.07914* (2015).

. Maia, F. F. *et al.* The VISCACHA survey–I. Overview and first results. *Monthly Notices of the Royal Astronomical Society* **484,** 5702-5722 (2019).

. Yang, J. y Leskovec, J. *Defining and Evaluating Network Communities based on Ground-truth* 2012. arXiv: 1205.6233 [cs.SI].

. *ArcadeDB Manual* [Online; accessed 18. Nov. 2023]. https://docs.arcadedb.com.

. Ahmad, A. y Khan, S. S. Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access* **7,** 31883-31902 (2019).