Una mirada a la Astroinformática

Marìa Alejandra Malberti Riveros, Raul Oscar Klenzi

Departamento de Informática – Instituto de Informática – Universidad Nacional de San Juan (UNSJ) – San Juan, Argentina

{amalberti,rauloscarklenzi}@gmail.com

Abstract. *Astroinformatic is presented as a new scientific discipline / sub-discipline. Based on Data Science, some features or aspects, which deserve further analysis by experts in Astronomy and Data Mining, are observed. To illustrate this, a set of data captured by the Sloan Digital Sky Survey Telescope, has been processed in Weka and Rapidminer environments by using K-means algorithm.*

Resumen. *En el trabajo se presenta a la Astroinformática como nueva disciplina/subdisciplina científica. A la luz de la Ciencia de Datos, se reconocen características o aspectos a ser ahondados en próximas investigaciones por expertos en Astronomía y por expertos en Minería de Datos-MD. Con el propósito de ejemplificar lo analizado se procesa un conjunto de datos, captados por el telescopio de exploración del espacio digital Sloan, en los entornos Weka y Rapidminer, haciendo uso del algoritmo K-means.*

1. Introducción

La disponibilidad de grandes volúmenes de datos provenientes de diferentes disciplinas científicas, ha llevado a que el descubrimiento científico orientado por los datos sea nombrado como una nueva ciencia X-Informática, donde X refiere a cualquier ciencia (por ejemplo Bio, Geo o Astro), e informática alude a la disciplina que organiza, describe, accede, integra, realiza minería y analiza diversas fuentes de datos con fines de descubrimiento científico. Es así que muchas disciplinas científicas están desarrollando sub-disciplinas formales que son ricas en información y basadas en datos. Borne, en el articulo Astroinformatics: A 21st Century Approach to Astronomy recomienda la creación de una nueva disciplina denominada Astroinformática, que incluye "*a set of naturally-related specialties including data organization, data description, astronomical classification taxonomies, astronomical concept ontologies, data mining, visualization, and statistics* **...** *Now is the time for the recognition of Astroinformatics as an essential methodology of astronomical research*.**"** [Borne, 2009a: 1]**.** En Astroinformatics: data-oriented astronomy research and education el mismo autor se hace eco de otros científicos al plantear una serie de cuestionamientos o áreas de interés en investigación, que podrían surgir de las problemáticas "“*while data doubles every year, useful information seems to be decreasing*” ..., and “*there is a growing gap between the generation of data and our understanding of it*”. " [Borne, 2010: 6]

Surgen entonces nuevos modos de descubrimiento científico a raíz del crecimiento de los datos y de los recursos computacionales disponibles y emergentes. Esta infraestructura incluye bases de datos, observatorios virtuales, computación de alto rendimiento (clusters y máquinas petaescala), computación distribuida (Grid, la nube, redes peer to peer), herramientas inteligentes de búsqueda y descubrimiento, e innovadores entornos de visualización. Asimismo se propone un enfoque de investigación científica de cuatro componentes en el que los datos, sensores, modelos y simulaciones actúan de forma sinérgica para impulsar nuevos descubrimientos, con datos como el eje central de las actividades científicas. Se perpetúa así, en Astronomía, la sinergia Datos-Sensores-Computación-Modelo.

Con el fin de aumentar el descubrimiento científico y la generación de conocimiento a partir de colecciones de datos masivos, Borne propone que la Astroinformática se transforme en una disciplina de investigación independiente con características como las siguientes: E-ciencia en Astronomía, áreas de investigación clave de una disciplina Astroinformática, minería de datos en Astronomía, y Astroinformática como el nuevo paradigma en la Astronomía del siglo 21.

El llamado Observatorio Virtual-VO (Fuentes heterogéneas distribuidas, federadas, accesibles a través de múltiples interfaces) es la ciberinfraestructura fundacional de la Astroinformática. Por medio de él los astrónomos realizan minería de datos tendiente al descubrimiento científico sobre recursos heterogéneos.

En Astroinformática el objeto básico debe ser el objeto astronómico (la estrella, la galaxia, la supernova, el asteroide, el quasar, etc.) y las herramientas de MD podrían incluir por ejemplo algoritmos de clasificación y agrupación, para caracterizar, clasificar, organizar y resumir el conjunto de objetos que pueblan el Universo.

Respecto a las áreas claves, el reporte Towards 2020 Science [Emmott and Rison 2008] identifica las categorías: (a) interacción inteligente y descubrimiento de información y (b) la transformación de la comunicación científica.

Por su parte en los artículos Extracting knowledge from massive astronomical data sets [Brescia, Cavuoti, Djorgovski, Donalek, Longo, and Paolillo 2012] y Astroinformatics, data mining and the future of astronomical research [Brescia and Longo 2013], también se expresa que hay razones para considerar a la Astroinformática como una nueva disciplina, pues los problemas que la misma ataca difieren en tipo y tamaño de los tratados por otras disciplinas. Es así que la Astroinformática está desencadenando un verdadero cambio metodológico dentro de la comunidad astronómica, que puede entenderse mejor teniendo en cuenta la cantidad y complejidad de los datos. La nueva generación de telescopios, instrumentos y sensores producen grandes cantidades de datos que no pueden ser eficazmente exploradas con las herramientas de hardware y software tradicionales, pues se requieren mayores facilidades de procesamiento, de reducción y de análisis. Sin embargo se expresa que mas desafiante es el aspecto de complejidad, y se tratan de mostrar sus implicaciones a partir de introducir el concepto de "*astronomical parameter space*". La historia completa de los descubrimientos astronómicos podría volver a leerse como la historia de un espacio de parámetros en constante evolución, por ejemplo "*quasars were disentangled from stars when the radio flux dimension was added; pulsars when the sampling of the radio flux was increased down to a few milliseconds, etc.*" [Brescia and Longo 2013: 2]

Un aspecto que en esos artículos se menciona es que los algoritmos de minería de datos no son robustos con datos faltantes y tampoco son efectivos al momento de tratar límites superiores. En otras palabras, si por ejemplo en un registro con datos de personas falta el valor del campo edad, puede significar que el dato se ha perdido, pero si un objeto astronómico carece de una magnitud en una banda fotométrica específica puede que el dato se haya extraviado o que sea tan débil que no puede ser detectado por la banda. En el caso de dato perdido, este podría ser recuperado usando algún modelo de datos previamente elaborado y ya adoptado por la comunidad astronómica por ejemplo en la separación de estrellas y galaxias, o en la evaluación de "*photometric redshifts*". El otro caso abarcaría aplicaciones astronómicas más interesantes, por ejemplo "*the search for obscured or high redshift quasars*". Situaciones como estas llevarían entonces a adaptar métodos existentes o a desarrollar nuevos métodos de aprendizaje automático.

Otro problema serio lo constituye la escalabilidad de los algoritmos y de los métodos existentes. Se expresa que la mayoría de los métodos de minería de datos escalan mal ya sea con número creciente de registros y/o de características. Es así que cuando se trabaja con datos masivos, el problema generalmente se trata extrayendo subconjuntos de datos y realizando con ellos entrenamiento y validación de los métodos, para luego extrapolar los resultados al conjunto completo. Esta forma de trabajar produce cierto sesgo y a la vez perjudica su adopción por parte de la comunidad astronómica.

La práctica en minería de datos requiere a veces de cientos de experimentos para identificar el método más adecuado, o dentro de un mismo método la mejor arquitectura o combinación de parámetros. Cada funcionalidad puede implementarse con una variedad de modelos (por ejemplo, redes neuronales, máquinas de vectores soporte, redes bayesianas, etc.), y los casos de uso son definidos por una asociación de la funcionalidad y el modelo.

Hay que tener en cuenta que la minería de datos astronómicos es extremadamente heterogénea, pues los conjuntos de datos masivos necesitan ser accedidos y utilizados por una amplia comunidad de diferentes usuarios con diversos objetivos, intereses y métodos científicos.

A la vez, para que la minería de datos sobre conjuntos masivos de datos astronómicos sea eficaz y fácil de usar habría que proporcionar a los usuarios un acceso más fácil tanto a métodos como a potencia de cálculo; identificar e implementar algoritmos mejores y más rápidos que exploten el paradigma HPC- High-Performance Computing; minimizar o reducir la transferencia de datos por ejemplo moviendo programas en lugar de los datos, entre otros aspectos.

2. Caso de estudio: Agrupamiento (Clustering) sobre datos astronómicos

Diversos entornos de Minería de Datos, como Weka (*Waikato Environment for Knowledge Analysis*, entorno para análisis de conocimiento de la Universidad de Waikato, Nueva Zelanda) y Rapidminer (originalmente llamado YALE- *Yet Another Learning Environment,* desarrollado inicialmente por el Departamento de Inteligencia Artificial de la Universidad de Dortmund, Alemania), ofrecen numerosos algoritmos para las diferentes funcionalidades, entre las que se encuentra el agrupamiento o

clustering. Clustering consiste en particionar los datos en grupos de objetos similares entre sí, lo que causa que de ellos se pierdan ciertos detalles finos a cambio de simplificación. Desde una perspectiva de aprendizaje automático los grupos corresponden a patrones ocultos y la búsqueda de los grupos es de aprendizaje no supervisado.

En el caso de Weka, version 3.7.10, los algoritmos disponibles son: Cobweb, EM, FarthesFirst, FilteredClustered, HierarchicalClusterer, MakeDensityBasedClusterer, SimpleKMeans. Por su parte, en Rapidminer version 5.3.015 los algoritmos para clustering son: K-Means, DBSCAN, k-Medoids, Expectation Maximization Clustering, Support Vector Clustering, Random Clustering, Agglomerative Clustering, etc.

Con la intención de introducir la problemática que se presenta al trabajar con datos astronómicos, se usó el algoritmo SimpleKMeans de Weka y el operador K-Means de Rapidminer. K-means es un método de agrupamiento que tiene como objetivo la partición de un conjunto de n observaciones en k grupos. El nombre de K-means surge porque se representa a cada uno de los grupos por la media (o media ponderada) de sus puntos, es decir, por su centroide. Es así que cada observación pertenece al grupo más cercano a la media.

Los datos utilizados corresponden a una porción de los datos captados por el telescopio de exploración del espacio digital Sloan -conocido como Sloan Digital Sky Survey, SDSS-, el cual constituye una inmensa base de datos con información de objetos extragalácticos, obtenida mediante imágenes captadas por medio de un sistema fotométrico de 5 filtros. Este catálogo se actualiza cada cuatro años y tiene 357 millones de objetos con gran cantidad de parámetros, conformando un total de 18TB.

El archivo de trabajo consta de un catálogo organizado como una matriz en la que cada fila representa una galaxia, 721336 en el experimento realizado, y las columnas refieren a sus características, siendo estas: posición espacial del objeto en el cielo (ra,dec), indicador de la distancia (zfinal), magnitudes absolutas en diferentes filtros indicadores de la luminosidad de cada galaxia (absu, absg, absr, absi, absz). Si estas magnitudes se restan entre si dan información del color que indica el grado de evolución de las galaxias (si son más rojas significa que son más viejas, probablemente más grandes y tienen poca formación estelar). Otras características son la Masa en estrellas de las galaxias (M*), la Tasa específica de formación estelar en escala logarítmica (SFR/M*), que proporciona una estimación de la formación estelar que tiene la galaxia en unidades de masas solares, un índice (n_sersic) que indica si la galaxia tiende a ser tipo disco, o tipo esferoidal, un indicador de la edad (Dn4000) (a valores más grandes, galaxias más viejas) y su error estimado (errDn4000), y finalmente un Indicador de Metalicidad (12+log(O/H)) relacionado con la composición química.

En la tabla 1 se resumen los principales resultados surgidos de la ejecución del algoritmo K-means, en los entornos mencionados. Si bien en ambas situaciones se usó el valor k=2 (2 grupos), difiere significativamente la cantidad de registros que contienen los grupos generados. Esto es, mientras en Weka el cluster 0 contiene el 23% de registros y el cluster 1 el restante 77% de los datos de entrada, en Rapidminer el cluster 0 contiene el 50,8% y el cluster 1 contiene el 49,2%. Otra diferencia se manifiesta en los valores de los centroides para los atributos correspondientes a cada cluster.

Tabla 1 - Resultados de ejecución de K-means en Weka y Rapidminer

	Weka		Rapidminer	
	SimpleKmeans		Operador K-means	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Cantidad de registros	166832	554504	366501	354835
Atributos	**Centroides**			
ra	183.9387	185.6705	228.247	140.880
dec	24.9017	24.8843	24.252	25.546
zfinal	0.0766	0.1234	0.112	0.113
absu	-18.3995	-18.8025	-18.700	-18.719
absg	-19.5373	-20.4492	-20.238	-20.239
absr	-20.04	-21.2848	-20.999	-20.995
absi	-20.3396	-21.6627	-21.358	-21.355
absz	-20.4557	-21.9227	-21.586	-21.581
M*	9.7575	10.4185	10.237	10.295
SFR/M*	-9.7921	-12.8871	-12.238	-12.102
n_sersic	1.3257	2.735	2.482	2.334
Dn4000	1.3664	1.6653	1.596	1.597
errDn4000	0.021	0.2177	0.179	0.166
12+log(O/H)	8.8826	-99.9	-75.761	-73.686

A la búsqueda de conocimiento en datos astronómicos, por medio de minería de datos, se le agrega la dificultad de tener que tratar grandes conjuntos de datos con numerosos atributos de diferentes tipos (alta dimensionalidad), lo que lleva a que en los algoritmos se deban considerar nuevos requerimientos computacionales. Sin ahondar en detalles vinculados a los entornos de software libre de MD, y a la aplicación de diferentes algoritmos provistos por los mismos para cada una de las estrategias, de las cuales se ha elegido clusterización, las diferencias en los resultados presentados en la tabla 1 sacan a la luz la necesidad de tener en cuenta diversos aspectos cuando se trabaja con MD. Respecto a los algoritmos de clustering, en Astroinformatics, data mining and the future of astronomical research [Brescia and Longo 2013] se expresa que los algoritmos de clustering deberían, entre otras, tener las siguientes características:

• Escalabilidad para grandes conjuntos de datos;

• Capacidad para trabajar con datos de muchas dimensiones (espacio de parámetros multi-Dimensional, múltiples longitudes de onda, etc.);

• Capacidad para encontrar grupos de forma irregular;

• Manejo de los valores atípicos;

• Complejidad de tiempo;

• Dependencia del Orden de los datos;

• Etiquetado o asignación (hard or strict vs. soft of fuzzy);

• Confianza en un conocimiento a priori y en parámetros definidos por el usuario;

• Interpretación de los resultados;

La minería de datos es un proceso complejo, y en general los mejores resultados suelen encontrarse en base a prueba y error mediante la comparación de diferentes métodos o de diferentes implementaciones del mismo método. La aplicación de MD requiere una buena comprensión de las matemáticas que subyacen a los métodos, de la infraestructura informática y de los flujos de trabajo que deben ponerse en práctica. Hasta ahora, la mayoría de los expertos en diferentes dominios de la comunidad científica no están dispuestos a hacer el esfuerzo necesario para entender los detalles finos del proceso. Sin embargo esta situación es insostenible en el tiempo, pues estarán disponibles conjuntos de datos masivos cada vez de mayor tamaño, y una alternativa viable para su explotación es la MD. Desde la perspectiva de los autores del trabajo, el sólo hito del lanzamiento del telescopio espacial James Webb (JWST) y el constante incremento de puntos de observación hará que la magnificencia de datos generada por la Astronomía sea de tal envergadura que cobren mayor auge cada uno de los aspectos en el marco de la Astroinformática y de la Ciencia de los Datos. Estos abarcan desde las posibilidades que deberá brindar el hardware así como las nuevas potencialidades que debería ofrecer el software en lo referente a lenguajes, representación de los datos, algoritmos, formas de visualizar conclusiones, entre otros, tornando a la Astroinformática un área de permanente interés por mucho tiempo.

3. Conclusiones

La Astroinformática ha surgido en respuesta a la necesidad de vincular convenientemente la Informática con la Astronomía. Se trata de dos disciplinas que han evolucionado a pasos agigantados a la luz de diversos avances tecnológicos, especialmente de aquellos directamente relacionados con la recopilación y el procesamiento de los datos entre los que encuentran los Observatorios Virtuales y la Minería de Datos.

Si bien hay diferentes posturas epistemológicas respecto a cómo se construye y evoluciona el conocimiento científico, lo cierto es que en este proceso se debería incorporar lo relacionado con el tratamiento de datos captados desde diferentes fuentes. Es en este punto que comienza a adquirir mayor importancia la Ciencia de Datos, entendiendo a la Ciencia de Datos como un área de trabajo emergente relacionada con la recopilación, elaboración, análisis, visualización, gestión y conservación de grandes colecciones de datos. Consiste en la aplicación, especialmente a grandes cantidades de datos, de técnicas avanzadas de minería de datos y estadísticas, y de métodos y algoritmos provenientes de la ciencia y de otros campos.- Figura 1

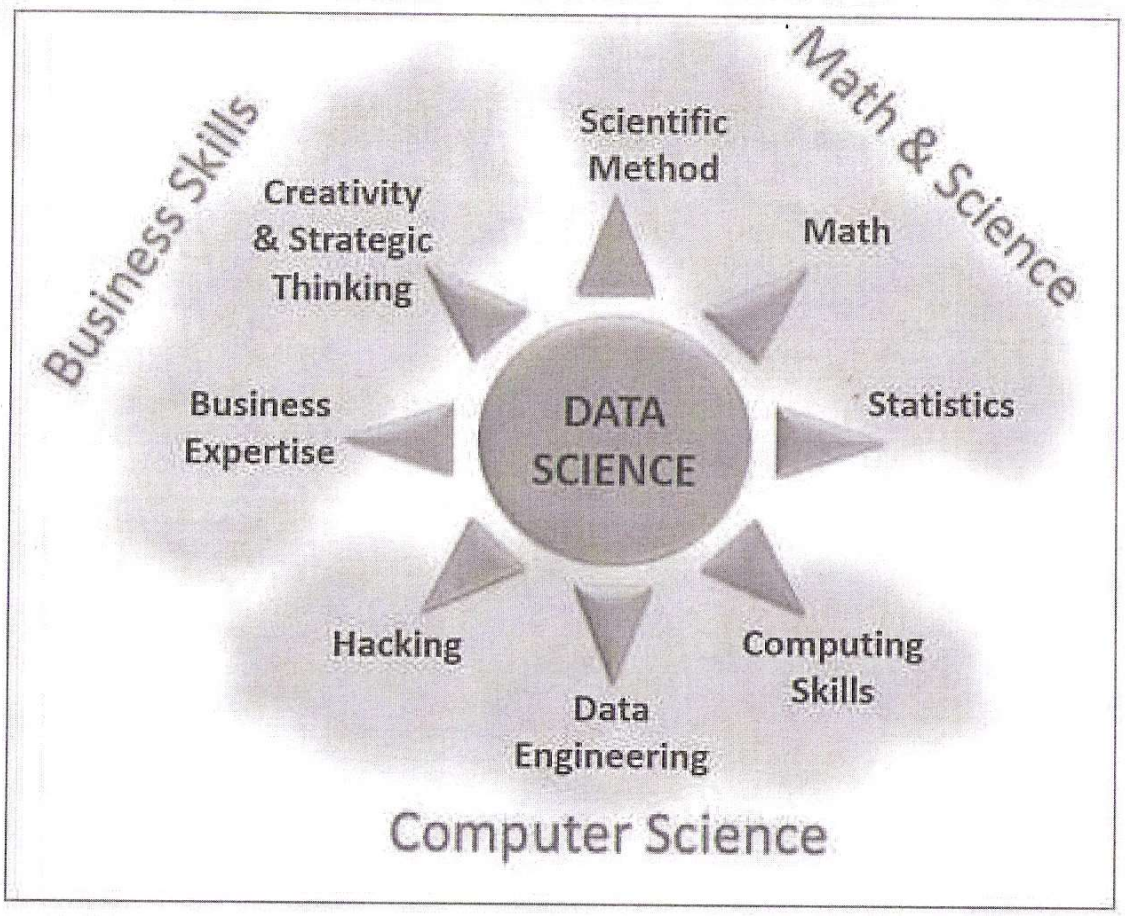

Figura 1- http://www.datascienceguide.com/data-science.html

Surge entonces la necesidad imperiosa de considerar en el proceso de desarrollo de conocimiento, aspectos relacionados con la disponibilidad y tratamiento de grandes volúmenes de datos; en otras palabras involucrar en estos procesos la Ciencia de Datos.

En particular, entendiendo a la Astroinformática como una emergente disciplina/subdisciplina científica, y desde lo propuesto por los autores tratados, adquieren entonces mayor importancia una serie de cuestiones:

En lo que refiere a los datos recolectados por los observatorios virtuales, existe una necesidad inminente de establecer estándares, tarea a la que se encuentran abocada diversas organizaciones.

En cuanto a aspectos relacionados con la búsqueda de conocimiento en los datos astronómicos, además de proporcionar a los usuarios un fácil acceso tanto a métodos como a potencia de cálculo se encuentran:

- Analizar las funcionalidades a ser aplicadas en los casos de uso de astrofísica. Un punto de partida podría ser la taxonomía propuesta en Extracting knowledge from massive astronomical data sets [Brescia, Cavuoti, Djorgovski, Donalek, Longo, and Paolillo 2012].
- Establecer claramente el tratamiento a realizar con datos faltantes o anómalos.

- Proponer criterios para efectuar análisis comparativo de comportamiento de algoritmos disponibles en el marco de diferentes funcionalidades. El caso de estudio tratado saca a la luz la problemática de producir diferentes resultados, que suele darse al usar distintos entornos (Weka y Rapidminer) y aplicar un mismo algoritmo, sobre el mismo conjunto de datos. Esta problemática se agudiza si se amplía el espectro de algoritmos utilizados.

Asimismo, la avalancha de datos ha promovido que varios grupos de estudio de Estados Unidos trabajen en temas relacionados con acciones a tomar en ciencia, en ingeniería y en el entorno académico. Por ejemplo el reporte Bits of Power de la Academia Nacional de Ciencias (National Academy of Sciences-NAS) lista recomendaciones que incluyen "*Improve science education in the area of scientific data management*" [Borne 2009: 4]. Atkins, en el reporte Revolutionizing science and engineering through cyberinfrastructure de la Fundación Nacional de Ciencia (National Science Foundation - NSF) establece qué destrezas en "D*igital libraries, metadata standards, digital classification, and data mining are critical*" [Atkins 2003: 20]. En The Revolution in Astronomy Education: Data Science for the Masses se expresa que en el workshop de la NSF del año 2007 se estableció explícitamente que "*Data-driven science is becoming a new scientific paradigm—ranking with theory, experimentation, and computational science*" [Borne, Jacoby, Carney, Connolly, Eastman, Raddick and Hamilton 2009: 3] . Es así que se plantea la necesidad de tener fuerzas de trabajo entrenadas en la disciplina de ciencia de datos, tal como también se expresa en el reporte del año 2009, Harnessing the Power of Digital Data for Science and Society [Council 2009], del grupo de trabajo en datos digitales del Consejo Nacional de Ciencia y Tecnología (National Science and Technology Council- NSTC).

4. Agradecimientos

Agradecemos la colaboración prestada por las docentes-investigadoras de la Universidad Nacional de San Juan, Prof. María José Marcovecchio y Dra. Georgina Codwell, por su aporte en la revisión y confección del abstract, y tratamiento de datos astronómicos respectivamente.

5. Referencias

Atkins, D. (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure.

Borne, K. D. (2010). Astroinformatics: data-oriented astronomy research and education. *Earth Science Informatics*, *3*(1-2), 5-17.

Borne, K. D. (2009). Astroinformatics: a 21st century approach to astronomy.*arXiv preprint arXiv:0909.3892*.

Borne, K. D., Jacoby, S., Carney, K., Connolly, A., Eastman, T., Raddick, M. J., ... & Hamilton, T. (2009). The Revolution in Astronomy Education: Data Science for the Masses. In astro2010: The Astronomy and Astrophysics Decadal Survey (Vol. 2010).

Brescia, M., & Longo, G. (2013). Astroinformatics, data mining and the future of astronomical research. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *720*, 92-94. http://arxiv.org/pdf/1201.1867

Brescia, M., Cavuoti, S., Djorgovski, G. S., Donalek, C., Longo, G., & Paolillo, M. (2012). Extracting knowledge from massive astronomical data sets. InAstrostatistics and Data Mining (pp. 31-45). Springer New York.

Brescia, M., Cavuoti, S., Esposito, F., Fiore, M., Garofalo, M., Guglielmo, M., ... & Vellucci, C. (2016). DAMEWARE-Data Mining & Exploration Web Application Resource. arXiv preprint arXiv:1603.00720. http://arxiv.org/pdf/1603.00720

Emmott, S., & Rison, S. (2008). Towards 2020 science. Science in Parliament,65(4), 31-33. http://www.scienceinparliament.org.uk/wp-content/uploads/2013/09/sip65-4-17.pdf

Feigelson, E. D., Ivezić, Ž., Hilbe, J., & Borne, K. D. (2013). New Organizations to Support Astroinformatics and Astrostatistics. arXiv preprint arXiv:1301.3069.

Interagency Working Group on Digital Data to the National Science and Technology Council.(2009). Harnessing the Power of Digital Data for Science and Society. https://www.nitrd.gov/About/Harnessing_Power.aspx

National Research Council^ dCommittee on Issues in the Transborder Flow of Scientific Data. (1997). Bits of power: issues in global access to scientific data. http://www.nap.edu/read/5504/chapter/1