REVISTA DEL CENTRO DE ESTUDIOS DE SOCIOLOGÍA DEL TRABAJO

Universidad de Buenos Aires
Facultad de Ciencias Económicas

NUEVAS PROFESIONES Y TÉCNICAS DE *WEB DATA MINING* EN ARGENTINA: EL CASO DEL *DATA SCIENTIST*

Autor : Achille Pierre Paliotta

Fuente: Revista del Centro de Estudios de Sociología del Trabajo, Nº 10 (Abril 2018), pp. 95-112

Publicado por: Facultad de Ciencias Económicas de la Universidad de Buenos Aires.

Vínculo:

La Revista del Centro de Estudios de Sociología del Trabajo es una revista académica anual editada por el Centro de Estudios de Sociología del Trabajo (CESOT) perteneciente al Instituto de Investigaciones en Administración, Contabilidad y Métodos Cuantitativos para la Gestión (IADCOM).

Nuevas profesiones y técnicas de web *data mining* en Argentina: el caso del *Data Scientist*

Achille Pierre Paliotta[1]

Resumen

La ciencia de datos es un campo emergente y en crecimiento continuo, que se expande particularmente en el contexto de las organizaciones empresariales a través de la demanda del científico de datos, profesión asociada a ese campo. El artículo presenta un análisis sobre esta nueva profesión, basándose en la extracción de datos no estructurados realizada a partir de un motor de búsqueda vertical. Se trata de un estudio exploratorio focalizado en Argentina, país que carece de una base informativa sobre este perfil profesional. Se trató de identificar y reconstruir el perfil del científico de datos, en lo referido a las habilidades y competencias técnico-profesionales demandadas, utilizando una metodología que demuestra eficacia para extraer información útil de las plataformas virtuales sobre ofertas de empleo.

Palabras clave: Ciencia de datos; científico de datos; nuevas profesiones; *data mining*; ofertas de empleo *on-line*

New professions and data mining technics in Argentina: the case of the Data Scientist

Abstract

Data science is an emerging and continuously growing field, which is particularly expanding in the context of business organizations through the demand of data scientists, professionals who are associated with that field of specialization. The article presents an analysis on this new profession, based on the extraction of unstructured data from a vertical search engine. It is an exploratory study focused on Argentina, where there is no information about this professional profile. The aim was to identify and reconstruct the profile of the data scientist in relation to the skills and technical-professional competencies demanded, using a methodology that demonstrates effectiveness in extracting useful information from virtual platforms on job offers.

Fecha de recepción: 06/11/2017 - Fecha de aceptación 27/02/2018

[1] Master en Data science - Università degli Studi di Roma "Tor Vergata". Investigador Istituto Nazionale per l'Analisi delle Politiche Pubbliche (INAPP). E-mail: a.paliotta@inapp.org.

El autor agradece los comentarios de los evaluadores anónimos de la revista.

Keywords: Data science; Data scientist; new professions; Web data mining; online job ads

> «Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician»,
>
> Josh Wills, @josh_wills, 3 mayo 2012.

Introducción

En octubre de 2012, *Harvard Business Review* publicó un artículo de Thomas Davenport y D.J. Patil, dedicado a la profesión *Data Scientist* (DS). Aunque este perfil fue descrito en correspondencia con los atributos de un profesional de alto rango, con entrenamiento y curiosidad para hacer descubrimientos en el mundo de *big data*, la denominación fue acuñada en 2008 y se considera que J. Hammerbacher y D. Patil fueron promotores de los primeros grupos de trabajo de ciencia de datos en Facebook Inc. y Linkedin Inc., respectivamente.

El título del primer artículo aludía al economista Hal Varian, quien había definido la tarea del DS como «el trabajo más sexy del siglo XXI», expresando sus convicciones sobre el tema y afirmando que aunque los estadísticos representan solo una parte de la contribución a ese quehacer, lo destacado radica en la capacidad necesaria para visualizar los datos, comunicarlos y utilizarlos de manera eficaz, junto con las habilidades para acceder, comprender y comunicar la intuición que se obtiene de los datos.

Al respecto se subrayaba la aptitud y capacidad que debían mostrar los *managers* para acceder y comprender los datos por sí mismos (McKinsey & Company, 2009). Progresivamente la denominación DS fue adoptada por la opinión pública e incorporada a las publicaciones de expertos, constituyéndose rápidamente en un fenómeno emergente, especialmente en los Estados Unidos, junto con el tema de *big data*, para convertirse en una hipérbole comunicativa, en un *hype* (como fue caracterizada por Gartner Inc.). La consecuencia directa de este evento fue la toma de conciencia sobre la insuficiencia de profesionales formados, la carencia de conocimientos en el mercado del trabajo calificado y el inicio de una significativa inversión orientada a

la formación y entrenamiento de este perfil, con el fin de satisfacer las necesidades de las empresas (Provost & Fawcett, 2013).

Para algunos autores (Bughin, 2016, p.12), el mayor impacto en el rendimiento de grandes volúmenes de datos reside en la estrecha complementariedad entre la inversión en tecnología de información y las habilidades laborales. Otros estudios llevados a cabo por empresas de consultoría confirman estos supuestos (McKinsey & Company, 2011).

Con respecto al caso de Argentina, se debe señalar que el interés por la profesión DS es relativamente reciente y ha sido objeto de consideración en algunos artículos periodísticos, para indicar la gran demanda insatisfecha de profesionales de este campo (Garabetyan, 2015). Si bien comienza a considerarse como factor clave que plantea para las empresas un nuevo arco de posibilidades para el desarrollo de sus negocios (Samela, 2017), la nueva figura profesional está emergiendo lentamente. Al respecto se indica que la escasez de profesionales en la industria informática se acentúa en la ciencia de datos y coexiste con la presencia de personas con formación en ciencias aunque carentes de conocimientos de computación, necesarios para el manejo de datos masivos, o bien de aquellas formadas en informática, aunque sin experiencia en análisis de datos, no obstante los programas de formación cortos que tratan de saldar esa brecha (Pernas, 2016).

La difusión de las nuevas tecnologías y el surgimiento de capacidades computacionales sin precedentes ha posibilitado la multiplicación de sitios de encuentro entre oferta y demanda de trabajo/empleo a través de plataformas que proporcionan servicios de información y reclutamiento dirigidos a trabajadores y empleadores (Paliotta, 2015). Estos recursos también favorecen distinto tipo de estudios y permiten aplicar la diversidad de herramientas convencionales utilizadas por el DS.

El artículo se basa específicamente en el análisis de las ofertas de empleo publicadas que hacen uso de la plataforma Indeed. A partir de la extracción y depuración de la información útil publicada, se han identificado los rasgos propios del perfil profesional del DS, en términos de habilidades/capacidades requeridas. La identificación está basada en un modelo tripartito que distingue entre competencias básicas, competencias técnico-profesionales y competencias transversales.

Se trata de un estudio de naturaleza meramente exploratoria que trató de responder a los siguientes interrogantes:

1. ¿Es posible utilizar en forma conveniente y masiva las técnicas automatizadas para la extracción de datos de la *web*?
2. ¿Qué habilidades/competencias buscan las empresas a través de este perfil profesional?
3. ¿Que diferencias sustanciales se presentan entre el perfil de DS y algunos perfiles similares, como analista de negocio o el analista de datos?

Para dar cuenta de los hallazgos, el artículo se estructura en tres secciones. La primera presenta una referencia a la aparición de la profesión DS en Estados Unidos y en Argentina. En la segunda sección se describe la metodología utilizada en el estudio y se caracteriza la unidad de análisis, para considerar en la tercera sección las ofertas de empleo *on-line* y plantear finalmente algunas conclusiones.

1. El desarrollo de la ciencia de datos en los Estados Unidos y la relativa novedad en Argentina

Se puede argumentar que la ciencia de datos constituye esencialmente una derivación de la ciencia de la computación como disciplina académica desarrollada en torno a los años 1960 sobre los lenguajes de programación, de compiladores, de sistemas operativos, etc. Posteriormente, los algoritmos fueron añadidos como un aspecto importante de la teoría, con el objetivo de crear los primeros ordenadores "inteligentes" [2] En los últimos años, sin embargo, la creciente disponibilidad de las enormes bases de datos, principalmente debido a fenómenos como la web 2.0, la aparición de los dispositivos móviles, la computación en nube (*cloud computing*), así como las redes sociales, han impulsado la ciencia de datos como un campo interdisciplinario que se fue configurando lentamente.

[2] Otro tema que concitó la atención en la década de 1990 fue la minería de datos (*Data mining*). La metodología *Knowledge discovery from databases*, KDD surgió en esos años en la comunidad *Machine learning* y el término para el primer taller sobre el mismo tema (KDD-1989) fue acuñado por Gregory Piatetsky-Shapiro al popularizarse en la misma comunidad y luego en los círculos económicos y de prensa. El KDD es la extracción no trivial de información implícita, previamente desconocida, y potencialmente útil a partir de bases de datos; en otras palabras, es el proceso de identificar *patterns* comprensibles de datos (Srikant & Agrawal, 1996).

Originalmente, el término adoptado por algunos precursores era vago y algo indeterminado ya que se utilizaba principalmente para indicar sujetos yuxtapuestos, a menudo diferentes.

John Wilder Tukey, estadístico estadounidense, relevó un área original llamada análisis de datos que rápidamente se convertiría en una nueva ciencia y no simplemente en una rama de la convencional estadística. Al respecto, destacó que durante mucho tiempo había pensado como un estadístico, interesado en inferencias, de lo particular a lo general, aunque la evolución de la matemática estadística y sus motivos de sorpresa y duda lo llevaron a descubrir su interés central por el análisis de datos. Esto lo condujo a incluir, entre otros, procedimientos para el análisis de datos, técnicas para la interpretación de resultados de tales procedimientos, formas de planificación de la gestión de datos para facilitar su análisis, haciéndolo más preciso y exacto, como también técnicas y resultados de las estadísticas (matemática) que se aplican al análisis de los datos (Tukey, 1962,1).

Sin embargo, la semántica del término ciencia de datos se mantuvo fuertemente centrada en el concepto datos, debido a que la cantidad de información disponible para la manipulación es el verdadero valor añadido en comparación con épocas anteriores. Es indudable la existencia actual de un flujo de información sin fin. Los sensores móviles, los dispositivos de internet de los objetos (*Internet of things,* IoT), las "*data-driven apps*" y las tecnologías portátiles (*wereables*) generan abundante información en la red creada por los usuarios, más o menos voluntaria; una especie de *feedback loop* según la opinión de uno de los principales expertos del sector, Mike Loukides.

Básicamente, gracias a la tecnología "inteligente", el mundo resulta progresivamente "mapeado", "medido", "grabado" y "almacenado" en *bits* digitales. Así, crecientes segmentos de la población encuentran su razón de ser en el mundo virtual y la mayoría de los datos disponibles son un subproducto de esta creciente existencia digital. Esa enorme cantidad de datos disponibles será denominada *big data*. En este contexto, las grandes inversiones en las nuevas tecnologías para almacenar, analizar, generar informes y ver datos dependen, en gran medida, de cómo las herramientas son utilizadas por los trabajadores debido a que son los únicos capaces de extraer información útil para finalidades diversas.

Los últimos desarrollos se centran en conseguir en definitiva, soluciones automatizadas que pueden interpretar los datos, que sean "escalables", intentando buscar correspondencias entre los diferentes

fenómenos para ponerlos a disposición de las estrategias empresariales: la inteligencia artificial (IA), el aprendizaje profundo (*deep learning*), la *web* semántica e inteligente (Shroff, 2013, Workman, 2016). El progreso constante en el campo del procesamiento de lenguajes naturales (*natural language processing*, NPL) llevará estas máquinas inteligentes a un nivel cada vez más personalizado (inteligencia similar a la humana).

También en el campo de los negocios, la minería de datos se utiliza para realizar análisis que permiten promover la retención de clientes en cualquier industria en la que se pueden cambiar proveedores a bajo costo y los competidores están interesados en atraerlos. Para los bancos se trata de *attrition*[3] y para las compañías de telefonía de *churn* (Berry & Linoff 2004, 17). El sistema de información bancaria y de seguros contiene grandes volúmenes de datos, tanto operacionales como históricos. La recolección de datos concierne a toda la información del cliente y los bancos y compañías de seguros se interesan en aplicar la minería de datos en sus procesos de toma de decisiones en áreas como el *marketing*, la gestión del riesgo crediticio, la liquidez, la detección de lavado de dinero y de transacciones fraudulentas. Los bancos y compañías de seguros utilizan también sus modelos de riesgo crediticio como herramienta para clasificar a los clientes según las clases de riesgo (Han, Kamber y Pei, 2012; Ye, 2014).

En este contexto, el empleo de la ciencia de datos constituye una realidad significativa en los mercados de Estados Unidos y los principales países desarrollados. En estos, el DS se puede apreciar en términos de profesión, más que como una ocupación (Elliot, 1975).

Las informaciones contextuales acerca de este perfil profesional, se pueden encontrar en las búsquedas de Google Trends de los cinco últimos años. Estas reflejan la tendencia evolutiva y la situación contrastante entre el caso de Estados Unidos y el de Argentina respecto a la denominación "estadístico" y "científico de datos".

Al respecto, un estudio de McKinsey & Company Inc. previó que en 2018 los Estados Unidos se enfrentarán con una falta de personal calificado en habilidades analíticas profundas (*deep analytic skills)* de entre 140.000 y 190.000 profesionales, así como con una carencia de 1,5 millones de *managers* y analistas con las competencias básicas para

[3] El objetivo de *attrition analysis* es identificar clientes que tienen una alta probabilidad de dejar la compañía, y por esta razón la empresa puede llevar a cabo campañas de marketing para cambiar el comportamiento de los usuarios.

utilizar las herramientas de analítica derivadas de *big data* (McKinsey & Company, 2011 p.10).

Por otra parte, algunas estimaciones basadas en la evolución de los anuncios *on-line* de Indeed.com muestran que la profesión DS, que no existía en 2012, constituyó en el curso de tres años una de las mejor remuneradas en los Estados Unidos (Sinclair, 2015). Este dato se confirma también en un estudio relativo al salario promedio de este perfil que indica que se encuentra entre los más altos de la industria del *software* (King & Magoulas, 2015 p. 6). La encuesta, llevada a cabo a partir de los datos de Indeed.com muestra que el DS ocupa la primera posición junto con el *Software architect*, seguido del *Software engineer*, *Mobile engineer* y *Mobile developer*. Las primeras diez posiciones, de la lista, se completan con *UI/UX Developer*, *Software developer* y *Front-end developer*, *Web developer* y por último, aunque con cierto distanciamiento la de *Data analyst*[4].

Mientras que esta tendencia prevalece en Estados Unidos, donde el mercado de trabajo está caracterizado por una vasta información sobre este perfil profesional, no es posible realizar un análisis específico ni establecer comparaciones en el caso de Argentina debido a la ausencia de datos sobre remuneraciones y características de los potenciales postulantes, ya sea demográficas, de difusión del perfil en el mercado de trabajo, de competencias/habilidades, etc.

En el análisis que se presenta, la búsqueda del perfil profesional DS se focalizó en un marco general adaptado para contextualizar el estudio. Se utilizaron los datos disponibles en la red profesional Linkedin, de más de 546 millones de usuarios en más de 200 países y territorios, y de Indeed, importante motor de búsqueda vertical que registra más de 200 millones de usuarios únicos por mes y está disponible en más de 60 países y en 28 idiomas, abarcando el 94% del PIB mundial.

[4] En la lista existe duplicación de algunos profesionales en cuyo perfil aparece la palabra clave *software*, ya sean arquitectos, ingenieros o desarrolladores. La desigualdad salarial sugiere algunas distinciones entre ellos. Las principales diferencias se encuentran en que los arquitectos están más frecuentemente involucrados en actividades de liderazgo, gestión y administración general del proyecto y esto puede explicar su salario más alto, con respecto a los ingenieros y los desarrolladores, que contribuyen solo en partes del proyecto. Por otra parte existen diferencias significativas entre la tarea *mobile engineer* que requiere un mayor uso de habilidades técnicas y profesionales, como los lenguajes de programación y base de datos y *mobile developer*, quien elabora sistemas operativos móviles como iOS, Android y otras aunque las retribuciones son similares.

Sobre esta base, la búsqueda de profesionales considerados como DS el 26 junio de 2017, en Linkedin permitió encontrar 32.704 casos (Tabla 1).

En esa distribución, sobresale la posición de los Estados Unidos (56,6%), la mayoría proveniente de demandas de San Francisco Bay Area, Nueva York y Boston. En un grado de importancia relativa menor se encuentra la India (10,6%) y le sigue, en Europa continental, Francia (9,0%), el Reino Unido (8,2%), Alemania (4,9%), España (3,5%) e Italia (2,3%). Comparativamente, los registros para los países del Cono Sur son escasamente significativos. Sobresale Brasil (1,5%) al que sigue Argentina (0,4%) y Chile (0,2%).

Tabla 1. DS, clasificados *on-line* en Linkedin e Indeed

Países \ Plataformas	Linkedin	Indeed
Estados Unidos	18.525	2.662
India	3.472	470
Francia	2.948	380
Reino Unido	2.679	875
Alemania	1.602	465
España	1.146	104
Australia	897	86
Italia	761	69
Brasil	485	27
Argentina	124	17
Chile	65	4

Fuente: elaborado en base a datos de Linkedin e Indeed (06-2017)

Cabe destacar que en Indeed, para la misma fecha, se pueden identificar 5.159 empleos demandados para la posición de DS y la distribución proporcional es similar a la de Linkedin acentuándose la distancia entre los Estados Unidos y el resto de los países.

A la luz de estos registros se aprecia la distancia entre los mercados de trabajo nacionales y se pone de manifiesto la posición de Argentina, como un país rezagado en lo que se refiere a un perfil que "habla" sobre todo en inglés.

Para complementar estos datos que caracterizan la situación general, el estudio de caso se focalizó en el tipo de competencias/habilidades requeridas a los postulantes para la posición de DS.

2. Los anuncios clasificados on-line y la metodología de investigación

El estudio se basa en la recopilación de los anuncios disponibles en las plataformas especializadas en la búsqueda de empleo. La unidad de análisis bajo examen corresponde a los anuncios de empleo *on-line* publicados en la red con la descripción de los conocimientos requeridos a los postulantes. Estos anuncios pueden ser considerados un tipo de «comunicación organizacional» (Yates y Orlikowski, 1992:323) y pueden calificarse como «artefactos organizacionales» y un medio que conecta a individuos, grupos, ocupaciones y organizaciones (Rafaeli & Oliver, 1998 p. 343). Dado que los métodos tradicionales de recolección son insuficientes para un análisis exhaustivo de los datos, en las últimas dos décadas se ha desarrollado en forma creciente el área de búsqueda del *web data mining*. Los métodos se han extendido rápidamente entre los expertos y en el ámbito más general de la *comunidad de negocios,* lo que lleva a la extensión de las técnicas que se pueden relacionar actualmente con los motores de búsqueda *ad hoc* (*crawler*), la clasificación automática de documentos *web*, el análisis de los *web logs*, el desarrollo de *query* inteligentes, la implementación de modelos predictivos, etc. En este artículo se hará uso de una forma peculiar de esta técnica, definida como *web data mining* (Liu, 2011), para extraer informaciones en forma no estructurada, utilizando los anuncios clasificados de empleo.

La información que proviene de esa fuente contiene el nombre del perfil profesional (*job title*), la descripción del puesto (*job description),* el nivel jerárquico y las diversas características requeridas en términos de calificaciones, tipo de contrato, experiencia profesional, nivel de graduación, etc.

De todos estos aspectos los más importantes para el análisis son relativos a las habilidades y conocimientos (competencias). En términos generales, la competencia puede ser entendida como una capacidad demostrada para utilizar conocimientos y habilidades personales en situaciones de trabajo. En este análisis se utiliza una versión simplificada del modelo de competencia (ISFOL, 1994; Di Francesco, 1998) que distingue entre:

- competencia *básica* (conocimientos generales y habilidades técnicas básicas para la empleabilidad y ciudadanía; estas habilidades se relacionan con la dimensión cultural general de un individuo);
- competencia *técnica y profesional* (altamente específica, conocimientos relacionados con un contenido de trabajo e

identificación con los oficios y disciplinas. Incluye el conocimiento y las técnicas operativas específicas de alguna actividad que la persona debe dominar a fin de "actuar con competencia");

– competencia *transversal* (habilidades transversales, no conectadas a un estado de actividad o empleo específico que se pueden aplicar en las áreas de trabajo y en la vida, en general). Estas habilidades aparecen como estrategias generales, relacionadas con el medio ambiente y son flexibles y modificables.

El modelo ha sido aplicado al corpus de inserciones, como una grilla de análisis que permite identificar algunas características básicas, especialmente en la fase de procesamiento de datos y permite dar cuenta de los resultados del estudio. La metodología se desarrolló en cuatro etapas. A partir de la selección de diferentes plataformas de búsqueda de empleo como los sitios *web* corporativos, *job boards*, etc. (Paliotta, 2015) se utilizó el motor de búsqueda especializado o vertical. Las plataformas recogen los anuncios de empleo de miles de sitios *web*, incluyendo bolsas de trabajo, sitios de periódicos, anuncios de empresas de intermediación y selectoras de personal, sitios de empresas privadas etc. Entre los *job websites* se ha elegido el más importante a nivel mundial, Indeed.com, refiriendo a su versión argentina (*www.ar.indeed.com*).

En la segunda etapa se recopiló información relacionada con los anuncios que contenían la figura profesional del DS. En general, la actividad de *web data mining* se enfrenta con diferentes formatos tecnológicos (XML, JSON, HTML5, AJAX) y requiere varias operaciones de extracción y preparación en un formato adecuado para su sucesivo análisis. En consecuencia, se deben utilizar bibliotecas *software* para analizar documentos HTML. También se llevó a cabo una búsqueda restringida a la figura del DS, para reducir el campo a los anuncios que contenían solo ese perfil, con la finalidad de extraer, desde el sitio argentino de Indeed.com, todos los registros en un momento dado.

La tercera etapa de depuración de datos permitió, por las técnicas de *scraping,* abordar el tema de la calidad. En efecto, los anuncios de empleo publicados en la red, son, de hecho, un *corpus*, fácilmente utilizable sólo en apariencia. En realidad, no se pueden utilizar directamente para el análisis porque los datos de la red presentan una alta diversidad de contenido y un bajo nivel de estructuración. Existe también el grave problema de duplicación del mismo anuncio de empleo de un sitio a otro. Los problemas están relacionados con los formatos tecnológicos y de otro tipo. Además, la mayoría de los

clasificados está escrito en lenguaje natural y, por lo tanto, los textos no estructurados requieren un pre-procesamiento antes de ser tratados. En esta fase, se deben aplicar las técnicas de minería de textos (*text mining*) como la eliminación de las palabras vacías (*stop-words)* por carecer de significado propio y los errores tipográficos de re-escritura (*spell-errors*). En este análisis, el exceso de espacio en blanco, los caracteres especiales y los signos de puntuación fueron eliminados.

La cuarta y última etapa fue el tratamiento de los datos por medio de diversos *packages:* 'stringi' (Gagolewski, 2017) y 'tm' (Feinerer, 2017), escritos en R, para el *text mining*.

3. Detección on-line de las ofertas de empleo: skills y job description

Desde el sitio argentino Indeed.com (*www.ar.indeed.com*), se extrajeron en el mes de mayo de 2017 28 anuncios bajo el perfil DS y uno bajo el de "analista científico de datos". Luego de la depuración, por razones de duplicación del mismo anuncio (en número de 5), por estar redactado en otros idiomas (holandés e inglés) y porque no son del perfil DS, permanecieron diecisiete anuncios con los perfiles *DevOps Java Ssr*; *Software engineer*, *Product manager*, *Data engineer*, *Medical science liason* y *Fellow consultant*.

No obstante el empleo formal que pueden generar las grandes empresas en Argentina y la posición de las empresas exportadoras (Ministerio de Producción, 2017) la importancia que se puede suponer con respecto a la demanda laboral en este campo se refleja en un número de avisos reducido que indica la poca importancia del perfil de DS aún en comparación con otros países latinoamericanos como Chile o Brasil.

En lo que concierne a la lengua escrita de los clasificados, prevalecen los textos redactados en español (11), le siguen en inglés (4) y bilingües (español-inglés). Al respecto se debe señalar que la figura profesional del DS y su trabajo se caracteriza estructuralmente por el vocabulario técnico inglés de origen, de modo que varios anuncios se publican en inglés, aún cuando se trata de ofertas de empleo de empresas con sede en Argentina. Esto da cuenta de la creciente internacionalización de los mercados de trabajo nacionales y de la operatoria de las empresas, multinacionales o no, en todos los mercados, utilizando la lengua inglesa.

En cuanto a la ubicación geográfica de los anuncios 13 corresponden a Ciudad Autónoma de Buenos Aires y el resto a San Juan, San Luis y Córdoba, cada provincia con un solo anuncio, y uno sin

explicitar la ubicación específica. Esta tendencia se corresponde con la localización del segmento de empresas innovadoras del país ya que la demanda de este perfil profesional, se localiza casi exclusivamente en Buenos Aires. La mayoría de los anuncios se refieren a más de 30 días del mes de mayo, periodo de tiempo suficiente para un estudio exploratorio.

Las competencias básicas principalmente demandadas son el conocimiento de idiomas, en el caso de los anuncios escritos en español. El conocimiento de inglés es requisito en 8 casos de 11. En cuanto a la calificación se indica la graduación en disciplinas como estadística, economía, matemáticas, sistemas, física, actuario, ciencias exactas, ingenierías, administración de empresas, computación, informática, finanzas y química.

Las competencias técnicas y profesionales incluyen como áreas específicas el dominio de herramientas para el manejo de *big data* y *cloud computing*; *databases*; paquetes de análisis estadístico; sistemas operativos y lenguajes de programación y otros programas. Debido a su importancia en la vida cotidiana de trabajo del DS se le ha dado un espacio más destacado a estos tipos de competencia que a las otras (básicas y transversales). Como se destaca en los resultados de las búsquedas, el DS debe conocer necesariamente la forma de manipular los datos. Además de saber cómo trabajar con un paquete de análisis estadístico, debe ser capaz de programar (Python, Java (Javascript), C ++, Scala) utilizando un lenguaje de procesamiento de datos conectado a un *database* y utilizar algunos software de *cloud computing*, hacer uso de técnicas de *machine learning* y *data mining* además de los paquetes de Office tradicionales. Debe enfrentarse además, con frecuencia, a los problemas relacionados con *big data* y, por último, debe "dar voz" a los datos utilizando herramientas de visualización de datos.

Es muy difícil encontrar y reclutar un perfil que comprenda el conjunto de estas habilidades, requiere la flexibilidad y capacidad para moverse "ágilmente" entre ellas más que una alta especialización en todas ellas. No se trata por ejemplo de un *Database administrator* altamente calificado, sino de una persona que conozca las funciones básicas del lenguaje SQL; tampoco se exige la competencia de un programador experto sino la eficacia para recuperar en la red un par de *scripts* adecuados y el dominio de las técnicas estadísticas relativas al fenómeno analizado.

Se ha registrado al respecto que en ausencia de este perfil profesional se viene desarrollando una oferta de postgrado y una formación *ad hoc* debido a que aún se está definiendo esta disciplina, y

la propia naturaleza interdisciplinaria de los perfiles supone la convergencia de numerosas especialidades (Pernas, 2016).

En Argentina, la formación de este perfil se realiza a nivel postgrado y la formación de base proviene de la ingeniería, matemática, física y química a la que se agrega computación, *software* y arquitectura de base de datos (Samela, 2017). Luego de adquirir un título de educación en este campo -condición que puede resultar decisiva para acceder rápidamente a un puesto de trabajo- habrá posibilidades para el desarrollo de una carrera que otorgue una credencial educativa en un mercado de trabajo cada vez más global.

En materia de competencias técnicas y profesionales, el DS debe ser capaz de concatenar varias competencias/habilidades en un conjunto de idiomas diferentes para entender y manejar, con destreza, cargas de trabajo flexibles y complejos flujos de trabajo requeridos en las fases de funcionamiento de un proyecto. Su capacidad principal debe radicar en poder manipular e interactuar con plataformas de información innovadoras (*information platforms or dataspace*s según la definición de Hammerbacher). La gran cantidad de datos requiere también otras tecnologías de *big data*, tales como Hadoop y Spark, que pueden proporcionar un almacenamiento masivo para cualquier tipo de datos, mostrar un enorme poder de procesamiento y capacidad para manejar tareas o trabajos prácticamente ilimitados. La creación de modulos de *cloud computing* ha permitido, por último, en un tiempo de cálculo relativamente corto, la manipulación de múltiples datos y diferentes algoritmos favoreciendo así la difusión de las técnicas de *machine learning* más allá de los especialistas.

En cuanto a las competencias *transversales* se ha registrado que se requieren capacidad de análisis, solución de problemas, trabajo en equipo, predisposición al cambio, innovación y descubrimiento de nuevas tecnologías, habilidades de liderazgo, coordinación de equipos y áreas; proactividad; habilidades de comunicación e interpersonales, gestión de proyectos complejos, etc.

En síntesis, las diferencias del perfil DS con aquel similar, el analista de negocios (Paliotta & Lovergine, 2017), revela para el DS una gestión y una mayor orientación estratégica, como la presencia de numerosas habilidades técnicas-profesionales o conjunto de competencias como, por ejemplo, algoritmos, *data mining*, lenguajes de programación, *text mining*, *neural network*, inteligencia artificial, o *recommendation engine*s, lo que parece confirmar esa hipótesis. Por el contrario, el perfil de analista de negocios requiere el conocimiento de técnicas y

conocimientos propios de la gestión de empresas, y los negocios, más que la manipulación de complejos conjuntos de datos.

Por último, la *inteligencia empresarial* (BI) se basa principalmente en el uso de una base de datos corporativa que integra y depura información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta (*data warehouse)* y alguna forma de informes o representaciones gráficas de los principales indicadores que intervienen en la consecución de los objetivos de negocio y se orientan a la toma de decisiones para optimizar la estrategia empresarial. En este sentido, los datos se utilizan para alimentar continuamente la base de datos de la empresa y para responder a interrogantes específicos, por lo general relacionados con un contexto típico de la gestión y administración. Además, como parte de la BI, los analistas de negocios no utilizan, a nivel personal, sus propios productos de análisis, ya que éstos siempre están relacionados con un contexto de negocios.

Conclusiones

Los resultados del estudio exploratorio presentado en el artículo permitieron esbozar el alcance de la problemática sobre ofertas de empleo vinculadas con el perfil DS. A modo de conclusión se puede indicar, entre algunos aspectos destacados, la eficacia de la técnica utilizada de extracción de datos no estructurados de la *web* y su potencialidad para aplicarse a una gran cantidad de informaciones. A ello se añade la conveniencia de utilizar datos provenientes de la red y la posibilidad de tener en cuenta la velocidad del proceso evolutivo del campo de las profesiones, tanto en relación con el número total - aparición casi diaria de nuevas ocupaciones - como respecto a sus características esenciales en términos de competencias/habilidades. Ambos aspectos son posibles, porque se renuncia a una clasificación previa y se decide, en cambio, recoger datos relativos a la demanda de las empresas.

En cuanto a la reconstrucción del perfil DS, se ha identificado que se caracteriza por requerir un conjunto innovador de habilidades en distintas áreas.

Es evidente que la ciencia de datos en el contexto argentino, a nivel general de difusión e importancia, y en cuanto al perfil profesional asociado se encuentra en una fase incipiente, comparada con el mercado de trabajo en los Estados Unidos en que la profesión del DS es creciente, incluso en lo relativo a remuneraciones del puesto de trabajo.

Sin embargo, el perfil comienza a considerarse a partir de un enfoque sistémico, debido a la creciente utilización de datos a todos los niveles por parte de las empresas y el consecuente acceso a todas las áreas (*marketing*, finanzas, recursos humanos, áreas de producción, etc.). Esto puede generar la necesidad de incorporar personal calificado con los conocimientos básicos en este campo y conducir eventualmente no sólo al reclutamiento de DS sino, también a la contratación de personal con una cultura básica orientada por el enfoque de toma de decisiones estratégicas basadas en análisis de datos e interpretación (*data-driven),* sin requerir conocimientos especializados. En este caso se demandan habilidades básicas en el manejo y análisis de los datos relativos al objeto del trabajo profesional y cierto nivel de competencia y capacidad para dar respuestas adecuadas a los objetivos de la empresa.

Una pequeña muestra de la red profesional Linkedin pone en evidencia el surgimiento de nuevos profesionales que se autocalifican como DS y muestra que la profesión, una de las más requeridas y mejor remunerada en los Estados Unidos, reemplaza en parte, otros perfiles de gran notoriedad como los de ingeniero de *software* o analista de datos. En ese contexto perduran los perfiles del analista de negocios debido a que sus competencias encuentran mayor aplicación operativa en los procesos de gestión y de negocios de las empresas.

Desde las referencias anteriores se pone de manifiesto que en el caso argentino la brecha tecnológica con respecto a los países más desarrollados parece ser cada vez mayor y se mantiene constante en este campo. En caso de permanecer en un futuro próximo puede representar la pérdida de una oportunidad para contar con mano de obra altamente calificada y capaz de hacer frente a los temas más innovadores, lo que significa desaprovechar nuevas oportunidades para el desarrollo en un segmento del mercado de trabajo, ahora de dimensiones globales, y altamente competitivo.

En síntesis, los datos relativos a la Argentina revelan que el DS más que constituir una profesión, aún se encuentra en el estadio de "ocupación". Los resultados aquí obtenidos son apenas un antecedente y pueden enriquecerse mediante un estudio longitudinal que posibilite seguir la evolución del fenómeno y establecer comparaciones con otros países latinoamericanos donde la temática tiene mayor alcance y envergadura, como por ejemplo, Brasil.

Referencias

Berry M.J.A., Linoff G.S. (2004) *Data Mining Techniques. For Marketing, Sales, and Customer Relationship Management*. Wiley Publishing. Indianapolis (IN).

Bughin J. (2016) Big Data, Big Bang?. *Journal of Big Data*. 3, 2, 14.

Davenport T.H., Patil D.J. (2012) Data Scientist. The Sexiest Job of the 21st Century. *Harvard Business Review*, pp. 70-76.

Di Francesco G. (a cura di) (1998) *Unità Capitalizzabili e crediti formativi. Metodologie e strumenti di lavoro*. ISFOL. Franco Angeli. Milano (IT).

Elliot P. (1975) *La sociología de las profesiones*. Tecnos. Madrid (ES).

Fayyad U., Piatetsky-Shapiro G., Smyth P. (1996) From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. American Association for Artificial Intelligence (AAAI). a. XVII, n. 3, pp. 37-54.

Feinerer I. (2017) *Package 'tm'*. package description, pp. 64.

Gagolewski M. (2017) *Package 'stringi'*. package description, pp. 133.

Garabetyan E. (2015) *'Científicos de datos': la nueva estrella entre las profesiones tech*. Recuperado de: http://www.perfil.com/ciencia/cientificos-de-datos-la-nueva-estrella-entre-las-profesiones-tech-1017-0184.phtml.

Han J., Kamber M., Pei J. (2012) *Data Mining. Concepts and Techniques*. Morgan Kaufmann. Waltham (MA).

ISFOL (1994) *Competenze trasversali e comportamento organizzativo. Le abilità di base per il lavoro che cambia*. Franco Angeli. Milano (IT).

King J., Magoulas R. (2015) *2015 Data Science Salary Survey. Tools, Trends, What Pays (and What Doesn't) for Data Professionals*. O'Reilly. Sebastopol (CA).

Liu B. (2011) *Web Data Mining. Exploring Hyperlinks Contents and Usage Data*. Springer. Berlin (DE).

McKinsey & Company. (2009). *Hal Varian on how the Web challenges managers.* Recuperado de: http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers.

McKinsey & Company (2011) *Big Data. The Next Frontier for Innovation, Competition, and Productivity*. report, pp. 156.

Ministerio de Producción. Argentina (2017) "Una radiografía de nuestro sistema productivo". Recuperado de http://gpsemp.produccion.gob.ar/index.php/datos_internacionalizacion/.

Paliotta A.P. (2015) Where the Jobs Are. Diffusione, tipologie e caratteristiche dei job websites negli USA e in Italia. *Osservatorio ISFOL*. n.s., a. V, n. 4, pp. 133-153.

Paliotta A.P., Lovergine S. (2017) Web data mining e costruzione di profili professionali. Il Business analyst nelle inserzioni di lavoro on-line. *SINAPPSI*. a. VII, n. 2-3, pp. 1-18.

Pernas M. (2016) *Científicos de datos: un perfil entre la ciencia y el negocio*. Recuperado de: https://www.clarin.com/economia/Cientificos-datos-perfil-ciencia-negocio_0_HJrchn_PXx.html.

Provost F., Fawcett T. (2013) *Data Science for Business. What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly. Sebastopol (CA).

Rafaeli A., Oliver A.L. (1998) Employment Ads. A Configurational Research Agenda. *Journal of Management Inquiry*. v. VII, n. 4, pp. 342-358.

Samela G. (2017) *Las empresas demandan científicos de datos*. Recuperado de: https://www.clarin.com/ieco/campus/empresas-demandan-cientificos-datos_0_ByCxMDfe-.html.

Shroff G. (2013) *The Intelligent Web. Search, Smart Algorithms, and Big Data*. Oxford University Press. Oxford (UK).

Sinclair T. (2015) *Beyond the Talent Shortage. How Tech Candidates Search For Jobs*, Indeed Report. Recuperado de: http://blog.indeed.com/hiring-lab/beyond-the-global-talent-shortage/.

Srikant R., Agrawal R. (1996) *Mining Sequential Patterns. Generalizations and Performance Improvements*. Proceedings of the 5th International Conference on Extending Database Technology. Advances in Database Technology. Avignon (FR), pp. 3-17.

Tukey J.W. (1962) The Future of Data Analysis. *The Annals of Mathematical Statistics*, pp. 67.

Workman M. (ed.) (2016) *Semantic Web. Implications for Technologies and Business Practices*. Springer International Publishing. Cham (CH).

Yates J., Orlikowski W.J. (1992) Genres of Organizational Communication. A Structurational Approach to Studying Communication and Media. *The Academy of Management Review*. v. XVII, n. 2, pp. 299-326.

Ye N. (2014) *Data Mining. Theories, Algorithms, and Examples*. CRC Press. Boca Raton (FL).