REF.: CREA **LABORATORIO DE INVESTIGACIÓN DENOMINADO "LABORATORIO DE DATA SCIENCE" (DatosLab)** DE LA UNIVERSIDAD DE PLAYA ANCHA DE CIENCIAS DE LA EDUCACIÓN.

DECRETO EXENTO N° 1239 / 2018

VALPARAÍSO, 28 de diciembre de 2018.

VISTOS Y CONSIDERANDO:

1. Que la Universidad de Playa Ancha de Ciencias de la Educación es una institución de educación superior cuyos fines esenciales son el cultivo, transmisión e incremento del saber y su campo esencial de atención es la Docencia, Investigación y Extensión de las disciplinas relacionadas con la Educación y la Cultura.

2. Que siendo uno de sus campos esenciales la Investigación, la Universidad requiere contar con Laboratorios de Investigación, los cuales constituyen un recurso básico para el desarrollo de la investigación en estándares de calidad y alto impacto.

3. Carta de fecha 2 de octubre de 2018 del Profesor Miguel Guevara, a la Decana de la Facultad de Ingeniería (s), doña Verónica Meza Ramírez.

4. Certificado N° 21 de fecha 10 de octubre de 2018, extendido por el Secretario Académico de la Facultad de Ingeniería, don Luis Faúndez Fuentes.

5. Memorándum N° 212/2018 de la Decana (s) de la Facultad de Ingeniería, al Vicerrector de Investigación, Postgrado e Innovación.

6. Memorándum N° 070/2018 del Vicerrector de Investigación, Postgrado e Innovación, al Rector.

7. Lo dispuesto en el inciso 2° artículo 1° de la Ley 18.434, artículo 34 letra c) del D.F.L. N°2 de 1986 y Decreto Supremo N° 269/2018, ambos del Ministerio de Educación.

DECRETO:

1. **CREASE** el Laboratorio de Investigación denominado **"LABORATORIO DE DATA SCIENCE" (DatosLab)**, destinado a desarrollar nuevos métodos y algoritmos de análisis, procesamiento y visualización de altos volúmenes de datos; generar aplicaciones públicas, con orientación social, basadas en datos utilizando APIs (Application Programming Interface) disponibles y métodos de WebScraping; y asesorar a organismos públicos y privados en el análisis de grandes volúmenes de datos, y cuyo proyecto de creación y desarrollo se transcribe a continuación:

PROYECTO:
Creación y desarrollo del "Laboratorio de Data Science" [DatosLab]

Dr. Miguel Guevara
Profesor Titular
Departamento de Computación e Informática
Facultad de Ingeniería.

2 de Octubre de 2018

Resumen Ejecutivo

Hace aproximadamente diez años, con el aparecimiento del primer teléfono celular inteligente, el mejoramiento de dispositivos móviles modernos y la irrupción de aplicaciones sociales en línea, el mundo comienza una era de abundancia de datos disponibles que plantean nuevos desafíos para la investigación y la innovación en el ámbito de la informática y varias disciplinas asociadas, lo que se ha denominado *Data Science* o Ciencia de los Datos.

Estos desafíos se pueden abordar desde cuatro pilares fundamentales: las políticas y filosofía respecto del uso de datos; la extracción, transporte y almacenamiento de datos; los métodos computacionales de análisis descriptivo, inferencial y predictivo; y las aplicaciones o herramientas de software que presentan los datos a través de visualizaciones modernas e interactivas de fácil acceso a la información.

Por otra parte, la Facultad de Ingeniería requiere potenciar la investigación con el objetivo, no solo de producir más publicaciones, sino también de mejorar la formación de sus estudiantes de pregrado y futuros estudiantes de postgrado.

Este documento contiene una propuesta de creación del laboratorio de investigación "Laboratorio de Data Science" (DatosLab) dependiente de la Facultad de Ingeniería de la Universidad de Playa Ancha que realice investigación en la denominada Ciencia de los Datos (*Data Science*) con un enfoque multidisciplinario y orientación social.

Tabla de Contenidos

Introducción

Vivimos una época de sobreabundancia de datos que aún cuando están disponibles, plantean una importante cantidad de desafíos, respecto de su uso, procesamiento e interpretación.

Esta producción de grandes volúmenes de datos o *Big Data*, se define bajo dos características básicas: volumen y velocidad (Vayena, Salathé, Madoff, & Brownstein, 2015). Los datos se van generando a una velocidad sin precedentes en la historia lo que produce volúmenes de datos difíciles de manejar con técnicas tradicionales de ciencias básicas como la estadística u otras. Se requiere de un enfoque multidisciplinario donde la informática lidera los procesos.

Una buena parte de estos datos, son producidos por usuarios, ya sea explícita o implícitamente. De forma explícita, las personas generan datos al ocupar sus redes sociales o aplicaciones que transmiten información públicamente como Twitter, Facebook o su propio correo electrónico. Estos datos, por ejemplo, se han utilizado para mapear los límites de una ciudad (Blanford, Huang, Savelyev, & MacEachren, 2015), detectar información falsa (Mendoza, Poblete, & Castillo, 2010), o intentar predecir elecciones políticas (Tumasjan, Sprenger, Sandner, & Welpe, 2010).

Por otra parte, los datos provistos por los usuarios, de manera implícita (sin que esta sea su intención), pueden proveer grandes volúmenes de información que es posible estudiar en búsqueda de patrones o explicación a ciertos fenómenos. Por ejemplo, el solo hecho de utilizar la señal del celular, se puede traducir en datos de movilidad en base a los cuales se pueden definir patrones de comportamiento (González, Hidalgo, & Barabási, 2008). O el hecho de buscar información en la web,

se puede traducir en tendencias, intereses o –como alguien propusiera–, focos de infección (Ginsberg et al., 2009).

Cuando en otra época, el principal desafío era mejorar el acceso a los datos, en nuestros días, con los datos ya disponibles, la comunidad científica se encuentra abocada a nuevos desafíos, estos se pueden clasificar en cuatro áreas o líneas de investigación: 1) definir, apoyar y difundir políticas claras de calidad, acceso y disponibilidad de datos, 2) mejorar la calidad de datos publicados, su interacción, así como los mecanismos de acceso y almacenamiento, 3) proponer nuevos modelos y algoritmos de análisis de datos, y finalmente, 4) desarrollar estudios y aplicaciones computacionales que permitan utilizar y entender de mejor forma los fenómenos generadores de los datos disponibles.

Es común referirse a este campo de estudio como *Data Science* o Ciencia de los Datos, que es una disciplina emergente y que vincula por un lado la Informática y por otro lado un conjunto de disciplinas, entre ellas la Estadística Computacional, el Aprendizaje Automático, la Visualización de Información y la Minería de Datos, entre otras. Todo esto en un contexto de altos volúmenes de datos o *Big Data*.

Problema y oportunidades

En esta sección se resumen los principales problemas y oportunidades detectadas para un futuro Laboratorio de Data Science.

Políticas de acceso y disponibilidad de datos

El acceso a datos, plantea un equilibro necesario entre transparencia y privacidad. Por un lado, aquellos datos que se consideran públicos, requieren ser liberados (abiertos) con el fin de que los gobiernos entreguen mayor transparencia en su actuar. Por otro lado, se encuentra el derecho a la privacidad de las personas. Es en la búsqueda de este equilibrio, donde aparece la necesidad de definir participativamente políticas y reglamentación de acceso y disponibilidad de datos, que permitan mejorar la transparencia sin atentar contra la privacidad. Avanzar en políticas de acceso a datos (sobre todo públicos) es una discusión necesaria en la que se debe participar desde la academia (Guevara & Pacheco, 2018).

Calidad de datos

En Chile la denominada Ley de Transparencia requiere que las instituciones del estado publiquen en sus sitios Web información institucional básica. Sin embargo, esta reglamentación no indica la calidad con que deben publicarse esos datos. Esta situación es común en gran parte del mundo. La heterogeneidad de formatos en que se publica la información dificulta la realización de estudios globales o de alto impacto en las políticas públicas. Adicionalmente, Chile es miembro de la Alianza para el Gobierno Abierto, lo que implica la adquisición de compromisos para avanzar en Gobierno Abierto desde una mirada de datos abiertos, participación ciudadana y sustentabilidad ("Open Government Partnership - Chile", 2015). Aquí es necesario trabajar en potenciar estándares de datos y en obtener información de la Web de manera automática, técnica conocida como *Web Scrapping* o *Web Crawling*

(Castillo, 2004) y que permite estructurar datos que se publican de manera no estructurada. Por ejemplo, en la investigación de análisis de publicaciones científicas para construir una red compleja de interacciones entre áreas de la ciencia, fue necesario implementar un *Web Crawler* para obtener datos de 12 millones de publicaciones científicas desde Google Scholar (Guevara, Hartmann, Aristarán, Mendoza, & Hidalgo, 2016). Esta técnica también es conocida como Minería de Datos de la Web o *Web Minning*.

Análisis de datos

Los datos individuales, en sí mismos, carecen de valor. Sin embargo, la interrelación, el análisis con modelos estadísticos y el procesamiento con algoritmos computaciones, permiten extraer información valiosa de los fenómenos que producen esos datos. En esta temática, los desafíos se traducen a buscar nuevos algoritmos de Machine Learning (Alpaydin, 2009), aplicar eficientemente modelos estadísticos y la aplicación de métodos modernos que abstraen la complejidad de los datos para simplificar su comprensión. Desde el área de Network Science (Barabási, 2016) se han conseguido importantes aportes en entender fenómenos complejos de los que hoy en día es posible obtener datos. En trabajos previos se han aplicado exitosamente estos métodos, por ejemplo, para determinar la desigualdad de ingreso en industrias productivas (Hartmann, Guevara, Jara-Figueroa, Aristarán, & Hidalgo, 2017; Hartmann, Jara-Figueroa, Guevara, Simoes, & Hidalgo, 2016), para posicionar instituciones y países en las respectivas áreas de la ciencia (Guevara, Hartmann, Aristarán, et al., 2016), o para detectar comunidades en países con

estructuras productivas similares (Guevara & Mendoza, 2016). Esta área, que procesa, analiza, modela y visualiza datos, también se conoce como *Data Analytics*.

Aplicaciones basadas en datos

Existiendo también, datos de buena calidad y fácil acceso, que se obtienen a través de repositorios institucionales o a través de APIs (*Application Programming Interface*) se hace cada vez más necesario contar con aplicaciones informacionales que permitan entender estos datos a través de visualizaciones estáticas o interactivas que entreguen a los usuarios estadísticas de sencilla comprensión y también que alienten la participación ciudadana. En esta área se deberán aplicar técnicas de visualización de datos (Chen, Härdle, & Unwin, 2007) y lenguajes computacionales modernos orientados al desarrollo web o el análisis de datos como R y Python. Aquí el investigador principal ha desarrollado el software **diverse**, que consiste en un *package* para el lenguaje R que facilita el cálculo de métricas en sistemas complejos (Guevara, Hartmann, & Mendoza, 2016). En conjunto con estudiantes, también se ha propiciado la generación de aplicaciones (a nivel prototipo) basadas en datos como la aplicación **Uninews** que rescata desde la Web información de noticias de distintas universidades para mostrarlas en un solo sitio centralizado; o la aplicación PalabraSonLey que construye nubes de texto de las Leyes de Chile, utilizando el API de la Biblioteca del Congreso. La creación de aplicaciones de este estilo, que se consiguen solo con investigación de base en el ámbito de estudio propuesto, son de vital importancia para una mejor vinculación con el medio y mayor participación de la ciudadanía, temática en la que la UPLA tiene larga trayectoria y compromiso.

Objetivos

Objetivo General

- Crear un laboratorio de investigación (DatosLab) en Ciencia de Datos, dependiente de la Facultad de Ingeniería.

Objetivos Especificos

- Propiciar investigación que permita desarrollar nuevos métodos y algoritmos de análisis, procesamiento y visualización de altos volúmenes de datos.
- Generar aplicaciones públicas, con orientación social, basadas en datos utilizando APIs (*Application Programming Interface*) disponibles y métodos de *WebScraping*.
- Asesorar a organismos públicos y privados en el análisis de grandes volúmenes de datos.

Recursos disponibles y financiamiento

Infraestructura

- DatosLab se implementará en la Sala 409 y Sala 601, gestionadas actualmente por la Facultad de Ingeniería. La sala 409 se utilizará para la permanencia de los estudiantes que participen del Laboratorio, mientras que la sala 601 se utilizará para las reuniones y demostraciones del laboratorio.

- Ambas salas cuentan con mobiliario adecuado para el trabajo individual y grupal, tales como cubículos individuales, mesa de reuniones y pizarra móvil.
- En cuanto a implementos computacionales se dispondrá, inicialmente, de aquellos con que ya cuenta el Departamento de Computación e Informática y también de aquellos que previamente se han adquirido por proyectos de investigación por parte del investigador principal. Estos son: Servidor de datos (actualmente en DataCenter Gran Bretaña), Computador iMac de escritorio, Computadores Windows Portátiles, Pantalla de 55".

Recursos humanos

- El laboratorio estará a cargo del Dr. Miguel Guevara A. quien tendría una dedicación de **12 horas.** En caso de tener horas asignadas a investigación por concepto de proyectos u otras tareas relacionadas, estas 12 horas se incluirán dentro de aquellas previamente asignadas.
- Se espera colaboración científica interdisciplinaria con los siguientes académicos y académicas de la Universidad con quienes ya se ha establecido algún nivel de colaboración científica: Dr. Carlos Valle (Facultad de Ingeniería), Dr. José González (Facultad de Ciencias), Mg. Carolina Santielices (Facultad de Ciencias Sociales), Dra. Marcela Prado (Facultad de Humanidades), Dra. Mirta Crovetto (Facultad de Ciencias de la Salud). Siendo esta lista no restrictiva a otras colaboraciones.

- En el ámbito nacional, se espera colaboración científica con los siguientes académicos: Dr. Marcelo Mendoza (UTFSM), Dr. Juan Zamora (PUCV), Dr. Jorge Fábrega (UDD).
- En el ámbito internacional, se mantendrá colaboración con los investigadores Dr. César Hidalgo (MIT, EEUU), Dr. Dominik Hartmann (USP, Brasil), Dr. Rodrigo Costas (CSTS, Países Bajos) y Dr. Héctor Ceballos (TEC, México).

Financiamiento

- Además de la infraestructura y equipamiento disponibles, se espera mejorar el equipamiento tecnológico a través de proyectos de investigación (FONDECYT), fuentes de innovación regional (FIC-R) o financiamiento de equipos a nivel nacional (FONDEQUIP). A través de estos concursos, se espera adquirir computadores modernos potenciados para trabajo diario además de servidores de alto rendimiento para computación paralela y de alto rendimiento. Este equipamiento es fundamental para procesamiento de altos volúmenes de información en tiempos óptimos.
- El Director de la propuesta ya ha postulado a FONDECYT iniciación 2018, con este objetivo. Se espera postular el próximo año a otras fuentes de financiamiento.

Trabajo adelantado

- Se ha construido un prototipo de una página web tipo vitrina, con las publicaciones, software y otras herramientas desarrolladas por el Investigador Principal, con el fin de plasmar la idea central del laboratorio.

Una vez aprobado el proyecto, se afinará esta página para incluir a los otros integrantes además información complementaria. Se puede consultar esta página en el siguiente enlace: http://datoslab.cl/

Plan y metas

- Durante los primeros cinco años, se espera que DatosLab mantenga un promedio de 2 a 3 publicaciones indexadas por Scopus y/o Web Of Science (WoS).
- Cronología anual para los próximos 4 años académicos (marzo-enero):
    - 2018:
        - Creación e implementación de DatosLab.
        - Publicación página web.
    - 2019:
        - Incorporación de 2-3 estudiantes tesistas.
        - Implementación de nuevo equipamiento.
        - Postulación a dos proyectos de Investigación o I+D+I externos.
        - Publicación de 2 artículos indexados.
    - 2020:
        - Incorporación de 1-2 postDocs.
        - Habilitación de espacio exclusivo de uso de DatosLab.
        - Postulación a dos proyectos de Investigación o I+D+I externos. En caso de no haber sido adjudicados en años anteriores.
        - Publicación de 3 artículos indexados.
        - Asistencia congreso nacional o internacional
    - 2021:
        - Incorporación de 1-2 postDocs.

- Postulación a dos proyectos de Investigación o I+D+I externos. En caso de no haber sido adjudicados en año anterior.
- Publicación de 3 artículos indexados.
- Asistencia congreso nacional o internacional

Impacto en Docencia y Vinculación con el Medio

Si bien DatosLab es un laboratorio propuesto en el área investigación, el natural impacto de la investigación que se realice se repercute también en las áreas de Docencia de pre/post grado y de Vinculación con el Medio.

Las capacidades que se desarrollen en el laboratorio requerirán necesariamente de estudiantes de pre y post grado quienes tendrán la oportunidad de trabajar en proyectos de frontera en el área de *Data Science* o Ciencia de los Datos. Se espera iniciar el trabajo científico con el apoyo de estudiantes de las Carreras de Ingeniería Informática e Ingeniería Estadística, expandiendo a futuro esta colaboración a otras carreras de la Facultad y la Universidad. En post-grado, se espera iniciar colaboración con el Magíster en Bibliotecología y el Magíster en Educación mención Nuevas Tecnologías, además de los programas de postgrado que se espera generar en la Facultad de Ingeniería.

En el ámbito de Vinculación con el Medio, se espera que el laboratorio propuesto, participe de concursos públicos con proyectos de investigación aplicada de beneficio para la Región. También, y como se mencionó anteriormente, se espera que DatosLab participe activamente de las políticas públicas y discusiones en el contexto de la temática de Datos además de generar aplicaciones de uso público con orientación social.

Conclusiones

La implementación del Laboratorio de Data Science (DatosLab) en la Universidad de Playa Ancha, permitirá avanzar en un tópico de investigación, que como se ha mostrado, es de gran relevancia a nivel mundial y nacional. De ahí que recientemente se hayan creado, por ejemplo, el Instituto Milenio para investigación en los Fundamentos de los Datos de la Universidad de Chile y el *Institute for Data Science* de la Universidad del Desarrollo. Centros a una escala mayor que la propuesta en este documento, pero que sin embargo, abordan temáticas similares y dan cuenta de una necesidad latente en el mundo académico. Además de los productos de investigación que se han proyectado, la UPLA también se beneficiará a nivel de pregrado y post-grado, por cuanto existirá la posibilidad, por parte de los estudiantes, de acceder a equipamiento moderno y a proyectos de tesis que trabajen con nuevos métodos de procesamiento y análisis sobre altos volúmenes de datos.

Referencias

Alpaydin, E. (2009). *Introduction to Machine Learning* (second edition). The MIT Press.

Barabási, A.-L. (2016). *Network Science*. Cambridge, United Kingdom: Cambridge University Press. Recuperado de http://barabasi.com/networksciencebook/

Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-Located Tweets. Enhancing Mobility Maps and Capturing Cross-Border Movement. *PLOS ONE*, *10*(6), e0129202. https://doi.org/10.1371/journal.pone.0129202

Castillo, C. (2004). *Effective Web Crawling*. Universidad de Chile. Recuperado de http://chato.cl/research/crawling_thesis

Chen, C., Härdle, W. K., & Unwin, A. (Eds.). (2007). *Handbook of Data Visualization*. Springer Science & Business Media.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012. https://doi.org/10.1038/nature07634

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779. https://doi.org/10.1038/nature06958

Guevara, M. R., Hartmann, D., Aristarán, M., Mendoza, M., & Hidalgo, C. A. (2016). The research space: using career paths to predict the evolution of the research output of individuals, institutions, and nations. *Scientometrics*, *109*(3), 1695–1709. https://doi.org/10.1007/s11192-016-2125-9

Guevara, M. R., Hartmann, D., & Mendoza, M. (2016). diverse: an R Package to Measure Diversity in Complex Systems. *The R Journal*, *8*(2), 60–78. https://journal.r-project.org/archive/2016/RJ-2016-033/

Guevara, M. R., & Mendoza, M. (2016). Publishing Patterns in BRIC Countries: A Network Analysis. *Publications*, *4*(3), 20. https://doi.org/10.3390/publications4030020

Guevara, M. R., & Pacheco, C. (2018). Los Datos Abiertos de Chile. *Serie de Documentos Técnicos Facultad de Ingeniería, Aceptado*.

Hartmann, D., Guevara, M. R., Jara-Figueroa, C., Aristarán, M., & Hidalgo, C. A. (2017). Linking Economic Complexity, Institutions, and Income Inequality. *World Development*, *93*, 75–93. https://doi.org/10.1016/j.worlddev.2016.12.020

Hartmann, D., Jara-Figueroa, C., Guevara, M. R., Simoes, A., & Hidalgo, C. A. (2016). Desigualdad del ingreso en América Latina y Asia: Un enfoque desde la estructura productiva. *Integration & Trade journal, Inter-American Development Bank*, *40*, 70–85.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: can we trust what we RT? En *Proceedings of the First Workshop on Social Media Analytics* (pp. 71–79). New York, NY, USA: ACM. http://doi.acm.org/10.1145/1964858.1964869

Open Government Partnership - Chile. (2015). Recuperado 11 de junio de 2015, de http://www.opengovpartnership.org/country/chile

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* Recuperado de http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441

Vayena, E., Salathé, M., Madoff, L. C., & Brownstein, J. S. (2015). Ethical Challenges of Big Data in Public Health. *PLOS Computational Biology, 11*(2), e1003904. https://doi.org/10.1371/journal.pcbi.1003904

2. EL **LABORATORIO DE DATA SCIENCE (DatosLab)**, dependerá de la Facultad de Ingeniería y funcionará en las salas N°s 409 y 601 de dicha Facultad.

3. El Laboratorio estará a cargo del Dr. Miguel Guevara, con dedicación de 12 horas.

REGÍSTRESE POR CONTRALORÍA INTERNA Y COMUNÍQUESE.

UNIVERSIDAD de PLAYA ANCHA de CIENCIAS de la EDUCACIÓN RECTOR VALPARAISO - CHILE

PATRICIO SANHUEZA VIVANCO
RECTOR

REGISTRADO
17 ENE 2019
CONTRALORÍA INTERNA

DISTRIBUCIÓN:

Rectoría
Prorrectoría
Auditoría Interna
Secretaría General
Finanzas
Tesorería
Contabilidad
Presupuesto
Facultad de Ingeniería
Asesoría Jurídica.
Oficinas de la Universidad (30)

ASESORIA JURIDICA VºBº UPLA ASOGADOS

PSV/NJV/mmm.