Encuentro Internacional de Educación en Ingeniería ACOFI
2º Congreso Latinoamericano de Ingeniería
2019 10 al 13 de septiembre - Cartagena de Indias, Colombia
RETOS EN LA FORMACIÓN DE INGENIEROS EN LA ERA DIGITAL

EXPLORACIÓN DE INFORMACIÓN HETEROGÉNEA CON TÉCNICAS DE ANÁLISIS VISUAL E INGENIERÍA DE CARACTERÍSTICAS: APLICACIÓN AL ANÁLISIS EXPLORATORIO DE DATOS DEL CEREBRO HUMANO

Duván Alberto Gómez Betancur, José Tiberio Hernández Peñaloza

Universidad de los Andes
Bogotá, Colombia

Resumen

Descifrar el alcance de la relación entre las características anatómicas del cerebro y su influencia en el funcionamiento de este, es uno de los campos de investigación más activos dentro de las Neurociencias. Actualmente, debido al incremento en las capacidades computacionales de almacenamiento y procesamiento y ante la cantidad y variedad de información disponible, una de las opciones para abordar el estudio de la relación anatómico-funcional del cerebro es mediante el Análisis Exploratorio de Datos (EDA), que consiste en ir sobre los datos y mediante el uso de herramientas de análisis visual y técnicas automáticas o semiautomáticas encontrar patrones que conduzcan a la generación de nuevo conocimiento. Después de una revisión de la literatura, se encontró que uno de los limitantes de las herramientas propuestas hasta ahora para análisis exploratorio de datos del cerebro humano, es la subespecialización de cada una de ellas a un único tipo de información y la consecuente necesidad de los especialistas para utilizar más de una herramienta si quieren analizar diferentes tipos de datos de un mismo sujeto. Este proyecto de tesis busca proponer un modelo que permita el análisis exploratorio de información heterogénea combinando la ingeniería de características con técnicas de análisis visual, manteniendo siempre a los especialistas como eje centralizador y permitiendo la interacción de profesionales de distintas especialidades.

Uno de los casos de estudio que se tiene previsto abordar es el análisis de una cohorte de jóvenes de 20 años nacidos prematuros. Dicho estudio comprende información de estructuras cerebrales y

de funcionamiento cerebral ante paradigmas por ejemplo de miedo y coordinación. La información está registrada en imágenes MRI, fMRI y DTI.

Palabras clave: análisis visual; estudios de cohortes; análisis exploratorio de datos; ingeniería de características

Abstract

Understanding the scope of the relationship between the anatomical characteristics of the brain and its influence on the functioning of the brain is one of the most active fields of research within the Neurosciences. Currently, due to the increase in computational storage and processing capabilities and the amount and variety of information available, one of the options to approach the study of the anatomical-functional relationship of the brain is through the Exploratory Data Analysis (EDA), which refers on going over the data using visual analysis tools and automatic or semiautomatic techniques in order to find patterns that lead to the generation of new knowledge. After a literature review, it was found that one of the limitations of the tools proposed so far for exploratory analysis of human brain data is the subspecialization of each one of them to a single type of information and the consequent need for specialists to use more than one tool if they want to analyze different types of data from the same subject. This thesis project seeks to propose a model that allows the exploratory analysis of heterogeneous information by combining feature engineering with visual analysis techniques, allowing the interaction of professionals from different specialties and always keeping specialists as the centralizing axis of the task.

Keywords*: visual analytics; cohort studies; exploratory data analysis (eda); feature engineering*

1. Introducción

El aumento en las capacidades de recolección, almacenamiento y procesamiento de información, acompañado de una tendencia que viene en alza entre los grupos de investigación y que hace referencia a poner disponibles públicamente los datos recolectados en los diferentes proyectos de investigación han apalancado lo que actualmente se conoce como investigación orientada por los datos.

En este tipo de investigación, los datos se encuentran disponibles y con técnicas de análisis automático o semiautomático se busca identificar estructuras o patrones en ellos, llevando a la formulación de nuevas preguntas y/o hipótesis que una vez validadas o rechazadas permitan la generación de nuevo conocimiento alrededor del sujeto o fenómeno estudiado.

La investigación desarrollada de esa manera se conoce también como análisis exploratorio de datos (EDA por sus siglas en inglés) y se apoya en la mayoría de los casos en herramientas de análisis visual. En el EDA, los investigadores exploran los datos para encontrar una estructura, un patrón o un comportamiento común (Tukey, 1980). En pocas palabras, los investigadores quieren

descubrir qué pueden decirles los datos sobre los fenómenos que se están estudiando y a partir de ahí generar nuevos conocimientos.

La transformación de datos, el uso de estadísticas sólidas, la visualización de datos para encontrar valores atípicos o patrones, y la generación de modelos son tareas involucradas en el análisis exploratorio de datos. El investigador comienza con una idea vaga y desarrolla su investigación de manera iterativa entre hacer preguntas y crear diseños o modelos.

En muchas ocasiones, la única forma de encontrar esos diseños o modelos y descubrir patrones en los datos es a través de transformaciones visuales, lo que hace que las visualizaciones científicas y estadísticas sean una parte fundamental del análisis exploratorio de datos. Asimismo, cada vez es más común encontrar trabajos que adoptan herramientas de aprendizaje automático o machine learning, para explorar los datos y descubrir patrones en estos (Thiagarajan, Kailkhura, Anirudh, Jain, & Islam, 2017).

En áreas del conocimiento como la economía y la inteligencia de negocios, en las cuales abundan los datos, se ha demostrado la utilidad del análisis exploratorio de datos (Thomas & Cook, 2006). Esto ha despertado el interés de la comunidad de investigadores alrededor de las neurociencias, la cual incluye no sólo médicos sino también ingenieros, psicólogos, pedagogos, y antropólogos entre otros, pues el cerebro humano desde siempre ha representado un enigma no solo para los neurólogos sino también para los investigadores de muchas áreas del conocimiento.

Debido a este involucramiento de especialistas de distintas áreas, en el estudio del cerebro humano, actualmente se tienen disponibles bases de información que incluyen datos de distinta naturaleza, incluyendo exámenes clínicos, conductuales, de laboratorio e imágenes, entre otros. Este proyecto de tesis doctoral busca proponer un modelo híbrido que combinando la ingeniería de características (Feature Engineering) con técnicas de análisis visual (Visual Analytics) permita el análisis exploratorio de toda esa información heterogénea.

En las siguientes secciones se describe el contexto del problema de investigación, un compendio de trabajos relacionados y recientes en el área y la propuesta metodológica que se está construyendo.

2. Contexto del problema de investigación

El análisis del contexto médico realizado hasta ahora dentro del proyecto, parte de la revisión de trabajos investigativos previos y que se encuentran tanto entre los proyectos realizados por el grupo Imagine de la Universidad de los Andes, así como en la lista de trabajos relacionados construida a partir de la revisión bibliográfica realizada.

El proyecto de investigación que motivó la propuesta de investigación presentada en este documento, corresponde a una tesis doctoral desarrollada dentro del grupo Imagine de la Universidad de los Andes y presentada en (Angulo, Schneider, Oliver, Charpak, & Hernandez, 2016). En este trabajo los autores utilizan una metodología centrada en el usuario y a partir de las

etapas iniciales de la misma, establecen que los investigadores del cerebro humano necesitan correlacionar medidas asociadas a conexiones neuronales, patrones o mapas de activación neuronal y modelos neurofisiológicos, con datos de una naturaleza diferente, como son el rendimiento neuropsicológico del paciente, sus comportamientos y otros datos clínicos.

Sin embargo, para hacer esa correlación, los especialistas usualmente cuentan con herramientas que son generalmente específicas para un solo tipo de datos y están optimizadas para soportar flujos de trabajo lineales. De ello se deduce que los expertos a menudo deben cambiar entre herramientas para integrar y analizar datos de diferentes dominios llegando incluso en el peor de los casos, a tener que cambiarse a otra computadora para poder hacerlo. Este proceso es lento y repetitivo, haciendo que el analista centre la atención en el "cómo" en lugar del "qué" y, por lo tanto, el análisis exploratorio de información sigue siendo un desafío que no todos los especialistas están dispuestos a asumir.

En (Keiriz, Zhan, Ajilore, Leow, & Forbes, 2018; Keiriz et al., 2017) identificaron algunas tareas de análisis visual y que son comunes entre los neurocientíficos. Entre las tareas identificadas se encuentran (Keiriz, Zhan, Ajilore, Leow, & Forbes, 2018):

T1: Identificar las regiones del cerebro que son responsables de funciones cognitivas específicas y estudiar sus interacciones con otras regiones.
T2: Comparar las redes neuronales de cada individuo contra la “red promedio” de un grupo definido. En los estudios grupales, se estudian las variaciones individuales y las características de la red conjunta para identificar puntos en común o diferencias.
T3: Identificar el efecto de la conectividad estructural en la actividad funcional del cerebro. Comparando a la vez estructural y funcional para revelar las complejas dependencias entre ellos.
T4: Identificar los cambios individuales o grupales que ocurren en las conexiones estructurales o funcionales debido a la aparición de una enfermedad o el envejecimiento, así como las debidas a las diferencias de género.
T5: Identificar cambios dinámicos en la conectividad estructural y funcional a lo largo del tiempo, tanto a nivel intra sujeto como inter sujeto.

Estas cinco tareas constituyen el marco contextual médico de esta investigación y lo que se busca es desarrollar una metodología de análisis exploratorio de datos que se pueda materializar en una aplicación web para dar soporte a los especialistas en la ejecución de las mismas.

Finalmente, vale la pena mencionar que, si bien existen técnicas de análisis automático que pueden aplicarse en el análisis exploratorio de datos del cerebro humano y en el soporte de las tareas arriba mencionadas, se debe considerar que sin duda alguna el análisis de imágenes médicas requiere de una experiencia sustancial en el dominio, y esto si bien ha representado una barrera para que los investigadores en informática y estadística apliquen de manera inmediata muchas de las innovaciones originadas en sus campos al análisis de datos del cerebro humano, abre al mismo tiempo una oportunidad para la integración del análisis automático y la visualización de información.

3. Trabajos relacionados

Como se mencionó anteriormente, nuestra referencia principal es uno de los trabajos de tesis doctoral realizados al interior de Imagine y que se describe en (Angulo et al., 2016). En este trabajo, Angulo propuso una herramienta de visualización llamada BRAVIZ en la cual se puede visualizar información de diferente naturaleza, a saber: datos espaciales representados en diferentes tipos de imágenes, y datos no espaciales como por ejemplo información clínica o psicosocial de los sujetos.

En BRAVIZ, los usuarios pueden ver y analizar diferentes tipos de imágenes del cerebro, por ejemplo, MRI, fMRI y DWI entre otras. Además, es posible visualizar el análisis estadístico de información extraída de las imágenes en conjunto con información contextual del individuo o del grupo de individuos, lo que le permite al especialista recopilar más información sobre el cerebro y facilitar la interacción y el trabajo colaborativo de expertos de diferentes áreas.

Uno de los principales inconvenientes de BRAVIZ está relacionado con el conjunto de ventanas independientes que lo componen y específicamente con la oclusión involuntaria que se puede presentar debido a que los usuarios pueden cambiar de forma arbitraria el tamaño y la posición de dichas ventanas y podrían terminar perdiéndose en el análisis exploratorio debido al desorden producido.

Otra desventaja de BRAVIZ es que en su concepción inicial presupone que será utilizado por un equipo de profesionales multidisciplinarios incluyendo ingenieros expertos, y por lo tanto, la configuración de BRAVIZ para trabajar con un nuevo conjunto de datos requiere la experiencia de ingenieros con habilidades específicas de programación.

Recientemente, se han propuesto otros enfoques, por ejemplo, en (Keiriz et al., 2018, 2017) presentan NeuroCave, una herramienta de software que facilita la exploración simultánea de conectomas en una variedad de configuraciones.

Aunque NeuroCave permite a los investigadores realizar comparaciones significativas entre los conectomas, uno de sus inconvenientes es que funciona solamente con imágenes tipo fMRI (resonancia magnética funcional) y DWI (difusión por resonancia magnética).

Yeatman (Yeatman, Richie-Halford, Smith, Keshavan, & Rokem, 2018) desarrolló una herramienta de software que visualiza los resultados del análisis cuantitativo en tractografía, facilitando el análisis exploratorio de los datos a través de la implementación de vistas enlazadas de los mismos.

En (de Ridder, Klein, & Kim, 2018) presentan una revisión sobre el análisis visual de las incertidumbres en el análisis de fMRI. De Ridder dividió el análisis fMRI en tres fases: adquisición y procesamiento, análisis de imágenes y visualización. De Ridder también identificó numerosas incertidumbres asociadas a cada una de las fases y enfatizó la forma en que se combinan a lo largo del proceso.

Los trabajos anteriores, aunque buscan facilitar la exploración de datos del cerebro, presentan soluciones que usualmente consideran sólo un tipo de datos, ya sean imágenes o datos tabulados con imágenes como contexto. Además, siguen dejando una carga cognitiva muy alta al especialista.

Dicha carga se puede reducir usando técnicas de análisis semiautomático que, con una intervención mínima por parte del usuario, sea capaz de extraer y construir visualizaciones con las características más relevantes de los datos. En este apartado, en (Liu & Su, 2004) aplicaron aprendizaje por refuerzo para la selección de características que permitieran identificar objetos en una imagen, y además agregaron la interacción hombre-computadora (HCI por sus siglas en ingles) para reducir la complejidad del aprendizaje y acelerar la tasa de convergencia del algoritmo propuesto.

Un trabajo reciente en ingeniería de características es (Cao et al., 2014) en el cual se usan tensores para reunir diferentes vistas en un espacio conjunto y se propone un método dual de selección de características de vista múltiple (DUAL-TMFS).

A partir de los trabajos identificados hasta ahora, se dilucida una alta oportunidad de cooperación entre la visualización de información y el análisis automático para crear una metodología de análisis exploratorio de datos del cerebro humano.

4. Propuesta

La analítica visual que integra técnicas de aprendizaje autónomo, estadística y visualización interactiva de información, favorece la creación de herramientas que les permiten a los especialistas una correcta aproximación a los datos y un uso óptimo de los mismos en el marco de la investigación exploratoria.

Se considera que el campo de las neurociencias y específicamente el estudio de la relación entre la estructura física del cerebro y el funcionamiento de este, puede beneficiarse significativamente de herramientas de análisis exploratorio.

En ese orden de ideas, esta propuesta busca estructurar una metodología híbrida que permita el análisis exploratorio de información heterogénea combinando la ingeniería de características (Feature Engineering) con técnicas de análisis visual (Visual Analytics). Se busca que la metodología propuesta tenga como eje central a los especialistas, permitiendo la interacción de profesionales de distintas especialidades y procurando que el analista centre la atención en el “qué" en lugar del "cómo".

La propuesta contempla las siguientes etapas:

4.1. Caracterización de datos disponibles y posibles "usuarios" de la herramienta

El objetivo principal de esta etapa es caracterizar los datos disponibles y las fuentes de estos, los usuarios potenciales, las tareas de análisis y los escenarios de análisis. En esta tesis se buscará caracterizar diferentes perfiles asociados a especialistas de diferentes áreas, para detectar puntos comunes y diferencias entre los flujos de trabajo de cada uno.

La información para esta etapa del proyecto se recopilará a partir de revisiones bibliográficas e interacción con expertos.

4.2. Revisión de trabajos relacionados

En esta etapa se buscarán los trabajos recientes en el área para identificar las fortalezas y las debilidades de los mismos. En la Sección 3 del presente documento se presenta una revisión inicial de literatura relacionada. En dicha revisión se encontró que uno de los limitantes de las herramientas propuestas hasta ahora es la especialización de cada una de ellas a un tipo de información y la consecuente necesidad de los especialistas de utilizar más de una herramienta si quieren analizar diferentes tipos de datos de un mismo sujeto.

Asimismo, otro limitante identificado en las herramientas propuestas, es que usualmente seleccionan técnicas puramente de visualización o técnicas puramente de aprendizaje automático y suelen orientarlas únicamente a un tipo de especialistas, ya sean neurólogos, psicólogos, antropólogos, etc. dificultando así el trabajo colaborativo.

4.3. Propuesta

Este proyecto de tesis doctoral concibe un escenario de exploración datos espaciales y no-espaciales del cerebro humano, es decir, un escenario con múltiples vistas del mismo sujeto (Figura 1).

4.3.1. Representaciones conjuntas

En esta era en la que es cada vez más común hablar de datos masivos, es posible acceder fácilmente a la información desde múltiples vistas que pueden obtenerse de diferentes fuentes (Figura 1).

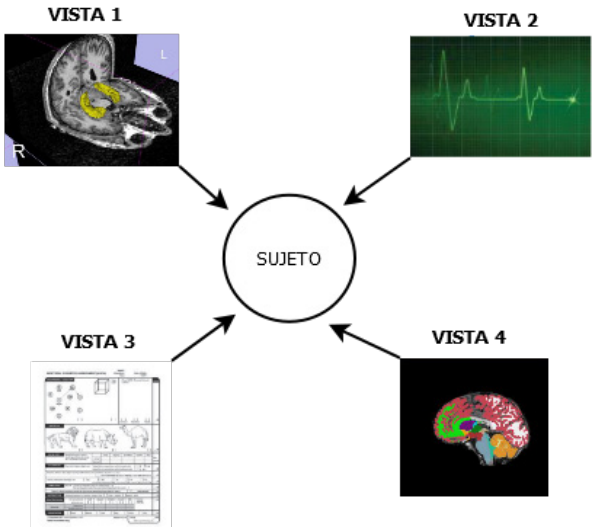

Figura 1 - Ejemplo de información multivista en estudios médicos.

En general, las diferentes vistas que se pueden tener de un sujeto o fenómeno proporcionan información complementaria para las tareas de análisis. Por lo tanto, la consideración e inclusión de esas múltiples vistas puede facilitar el proceso de análisis, exploración y aprendizaje alrededor de un conjunto de datos.

Como se observa en la Figura 1, la medicina es uno de los campos donde las mediciones desde múltiples vistas son más comunes, pues por ejemplo para cada sujeto se puede tener una serie de exámenes médicos, incluidas las medidas clínicas, las imágenes diagnósticas, exámenes de laboratorio y hasta pruebas cognitivas que se obtienen de múltiples fuentes.

Específicamente, para el diagnóstico cerebral, se pueden tener diferentes análisis cuantitativos que pueden verse como diferentes vistas o diferentes subconjuntos de características de un sujeto. Este hecho se evidencia en la base información usada en este proyecto y proveída por la Fundación Canguro.

Es deseable combinar todas estas características de una manera efectiva para el diagnóstico de enfermedades, por esa razón se buscará una manera de representar de manera conjunta información espacial con información no espacial. Una de las pistas que se está siguiendo en este momento es la planteada en (Cao et al., 2014) donde utilizan tensores para representar de manera conjunta información proveniente de diferentes fuentes. Se espera que esta subetapa conduzca a la formulación de una contribución en esta área.

4.3.2. Ingeniería de Características

Como se mencionó en el apartado anterior, contar con información de diferentes vistas o fuentes permite enriquecer el análisis exploratorio de los datos. Sin embargo, deben considerarse dos aspectos. El primero es que algunas mediciones entregadas por exámenes médicos pueden ser irrelevantes o ruidosas para algunos de los especialistas según su área de interés y producir efectos

indeseados tanto al construir las representaciones conjuntas antes descritas como en el flujo de trabajo de los usuarios.

Por lo tanto, se considera que la selección de características debe incorporarse en el proceso de análisis de múltiples vistas, de tal manera que se evite presentarle al especialista información errática o que pueda distraer su análisis.

Una correcta ingeniería de características permitirá no solo filtrar información irrelevante, sino un trabajo colaborativo y constructivo en la búsqueda de nuevo conocimiento por parte de especialistas de diferentes áreas, trabajando cada uno sobre el mismo conjunto de datos, pero con la comodidad que brinda un ambiente de trabajo personalizado a sus intereses.

4.3.3. Visualizaciones interactivas y enlazadas.

Herramientas como BRAVIZ (Angulo et al., 2016) demuestran que la visualización desempeña un papel integral en la exploración de datos y la investigación científica, incluso más allá de la comunicación de los resultados.

Campos del conocimiento como la astronomía y la genómica que empezaron a adoptar Big Data en sus investigaciones, han apalancado sus investigaciones en el desarrollo de herramientas que implementan vistas enlazadas o vinculadas de un conjunto de datos, donde la interacción con una visualización de una dimensión evoca un cambio en otra visualización de los mismos datos.

En esta investigación se quiere materializar la metodología propuesta en una aplicación web con riqueza visual de los datos, para facilitar el trabajo de los especialistas.

4.4. Evaluación de la herramienta en un caso real

El método de análisis exploratorio de datos heterogéneos que resulte del desarrollo de esta tesis doctoral, así como la aplicación web asociada al mismo, se evaluarán abordando el análisis de una cohorte de jóvenes de 20 años y que fueron nacidos prematuramente. El estudio está centrado en la información de estructuras cerebrales y paradigmas por ejemplo de miedo y coordinación registrados en imágenes MRI, fMRI y DTI.

Para el caso de evaluación mencionado, se cuenta con el acompañamiento de expertos de la Fundación Canguro con quienes se ha establecido un acuerdo para el análisis conjunto de los datos asociados.

5. Conclusiones

A partir de la revisión de literatura realizada hasta ahora nos motiva la idea de que la ingeniería de características de la misma manera que ayuda a los algoritmos de aprendizaje automático para lograr mejores resultados y en un período de tiempo más corto, puede usarse en conjunto con

visualizaciones interactivas para mejorar el análisis de datos exploratorios, permitiendo que el analista centre la atención en el "qué" de su investigación en lugar del "cómo".

6. Referencias

- Angulo, D. A., Schneider, C., Oliver, J. H., Charpak, N., & Hernandez, J. T. (2016). A Multi-facetted Visual Analytics Tool for Exploratory Analysis of Human Brain and Function Datasets. *Frontiers in Neuroinformatics*, *10*, 36. https://doi.org/10.3389/fninf.2016.00036
- Cao, B., He, L., Kong, X., Yu, P. S., Hao, Z., & Ragin, A. B. (2014). Tensor-Based Multi-view Feature Selection with Applications to Brain Diseases. In *2014 IEEE International Conference on Data Mining* (Vol. 2014, pp. 40–49). IEEE. https://doi.org/10.1109/ICDM.2014.26
- de Ridder, M., Klein, K., & Kim, J. (2018). A review and outlook on visual analytics for uncertainties in functional magnetic resonance imaging. *Brain Informatics*, *5*(2), 5. https://doi.org/10.1186/s40708-018-0083-0
- Keiriz, J. J. G., Zhan, L., Ajilore, O., Leow, A. D., & Forbes, A. G. (2018). NeuroCave: A Web-based Immersive Visualization Platform for Exploring Connectome Datasets. *Network Neuroscience*, 1–19. https://doi.org/10.1162/NETN_a_00044
- Keiriz, J. J. G., Zhan, L., Chukhman, M., Ajilore, O., Leow, A. D., & Forbes, A. G. (2017). Exploring the Human Connectome Topology in Group Studies. Retrieved from http://arxiv.org/abs/1706.10297
- Liu, F., & Su, J. (2004). An Online Feature Learning Algorithm Using HCI-Based Reinforcement Learning (pp. 293–298). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28647-9_50
- Thiagarajan, J., Kailkhura, B., Anirudh, R., Jain, N., & Islam, T. (2017). PADDLE: Performance Analysis using a Data-driven Learning Environment. Retrieved from https://www.osti.gov/servlets/purl/1455398
- Thomas, J. J., & Cook, K. A. (2006). A visual analytics agenda. *IEEE Computer Graphics and Applications*, *26*(1), 10–13. https://doi.org/10.1109/MCG.2006.5
- Tukey, J. W. (1980). We Need Both Exploratory and Confirmatory. *The American Statistician*, *34*(1), 23. https://doi.org/10.2307/2682991
- Yeatman, J. D., Richie-Halford, A., Smith, J. K., Keshavan, A., & Rokem, A. (2018). A browser-based tool for visualization and analysis of diffusion MRI data. *Nature Communications*, *9*(1), 940. https://doi.org/10.1038/s41467-018-03297-7

Sobre los autores

- **Duván Alberto Gómez Betancur**: Ingeniero Electrónico, Especialista en Gerencia de Sistemas y Tecnología, MSc. Ing. de Sistemas. Candidato a Doctor en Ingeniería de la Universidad de los Andes. Docente de Tiempo Completo en el Tecnológico de Antioquia – Institución Universitaria. da.gomez16@uniandes.edu.co - duvan.gomez34@tdea.edu.co

- **José Tiberio Hernandez Peñaloza**: Ingeniero de Sistemas y Computación, MSc Ing. de Sistemas y Computación, DEA Informatique Appliquée, Docteur Ingénieur, ENSTA-Paris. Profesor Asociado, director grupo IMAGINE. jhernand@uniandes.edu.co

---