# VALORACIÓN DE LA INCORPORACIÓN EN LA CURRICULA DE ASIGNATURAS DE IT DE NUEVAS TECNOLOGIAS. EL CASO DE DATOS MASIVOS, Y CIENCIA DE DATOS

**DIAZ, DANIEL J.**
*ddiaz@fcecon.unr.edu.ar*

**Facultad de Ciencias Económicas y Estadísticas. Universidad Nacional de Rosario**

**Palabras claves: Ciencia de Datos, Datos Masivos, Paradigma, Administración, Disrupción**

**ÁREA TEMÁTICA A: DIDACTICA Y CONTENIDOS**

**RESUMEN:**
El presente trabajo tiene como principal objetivo, definir críticamente una apreciación sobre las tecnologías de Datos Masivos y Ciencia de Datos, analizar las mismas como recurso estratégico para las organizaciones, estimar el impacto que las mismas pueden tener sobre modelos, saberes, prácticas de gestión. La intervención académica propuesta toma en consideración la discusión del rol del profesional en ciencias económicas en referencia a estas nuevas tecnologías y aspectos éticos del uso y explotación de Datos Masivos. También se considerará evaluar y abrir el debate acerca de la conveniencia que estos recursos sean incluidos en la currícula de asignaturas de IT en carreras de profesionales en ciencias económicas, cual es el enfoque a darle a esta temática y el modo de enseñarla.

## 1. INTRODUCCIÓN

*Musical recording was a mechanical process until it wasn't,*
*and became a network service.*

Jaron Lanier – Who owns the future

En el ámbito de estudio de la administración de recursos informáticos, resulta casi inevitable hablar de un principio rector, al que se denomina la "Ley de Moore". Este principio permite evaluar la tasa de aceleración con que se incrementan las prestaciones que brindan los componentes de sistemas de cómputos.

De tanto en tanto, esa evolución se ve afectada por lo que denominamos una disrupción. Brom F. (2014) define la cualidad de disrupción como "*innovación radical (no acumulativa ni evolutiva). Es aquella que cambia las reglas de juego de una actividad, sector o industria, por tratarse de una propuesta de valor que cambia*

*radicalmente las preferencias de los clientes, consumidores o usuarios de un producto o servicio*[1]"

En la actualidad, existen dos tecnologías emergentes que están ganando terreno en el ámbito de la Tecnología Informática: Datos Masivos (Big Data) y Ciencia de Datos (Data Science).

Hay coincidencia en caracterizar los desarrollos de Big Data, por medio de tres aspectos destacables conocidos como las 3 V: Volumen, Velocidad, Variedad.


**Figura 1: tres ejes conceptuales de Big Data**

Tal como define Foster Provost (2013), "ciencia de Datos envuelve principios, procesos y técnicas para entender fenómenos por medio de (automatización de) análisis de datos"[2]

En el caso de Datos Masivos, su disrupción implica el replanteo de algunos esquemas y prácticas tan arraigadas en nuestro medio (y en nuestras mentes), como el modelo relacional de Base de Datos.

Si nos enfocamos en la Ciencia de Datos, a juicio del autor, el impacto de la disrupción será aún mayor. Este tipo de soluciones requiere de un enfoque interdisciplinario (Estadísticas, Administración y Tecnología Informática) lo cual, multiplica el efecto sobre modelos, teorías, y prácticas de uso habitual, en la gestión de recursos informáticos.

El presente trabajo tiene como principal objetivo, definir críticamente una apreciación sobre estas nuevas tecnologías, analizar las mismas como recurso estratégico para las organizaciones, estimar el impacto que las mismas pueden tener sobre modelos, saberes, prácticas de gestión. También se considerará evaluar y abrir el debate acerca de la conveniencia que estos recursos sean incluidos en la currícula de asignaturas de IT en carreras de profesionales en ciencias económicas, cual es el enfoque a darle a esta temática y el modo de enseñarla.

El contexto desde el cual se genera la presente propuesta se corresponde con alumnos de tercer año de la carrera de Contador. De acuerdo al plan de estudio

[1] Brom, F (2014)

[2] Traducción del autor. "Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data" Foster Provost (2013)

respectivo, la asignatura en la que se pretende incluir los contenidos en discusión es la primera materia de la carrera en donde se aborda la cuestión de Tecnología Informática. La orientación que se busca a través de la asignatura es el aprendizaje y consideración de los aspectos relevantes de la gestión de recursos informáticos en las organizaciones.

## 2. PLANIFICACIÓN DE LA INTERVENCIÓN/TRANSFERENCIA Y DE SU SEGUIMIENTO

A continuación, se detallarán los aspectos más relevantes de la propuesta de intervención sugerida. Es importante destacar que, a juicio del autor, **el objetivo más destacado** que se busca con esta propuesta es presentar el **debate y evaluación sobre el cambio de paradigma** que se prevé surgirá de la implementación de las 2 tecnologías en cuestión.

### 2.1. Objetivos de la intervención:

- ✓ Lograr que los alumnos realicen una evaluación crítica del cambio de paradigma que provocará la adopción de estas nuevas tecnologías en el ámbito social y empresarial. En especial, en el enfoque de teorías y prácticas de gestión tradicionales y universalmente aceptadas.
- ✓ Reforzar la concepción de los recursos informáticos, como activos estratégicos en el proceso de generación de valor de las organizaciones
- ✓ Transferir los conocimientos básicos sobre las tecnologías de Datos Masivos y Ciencia de Datos, necesarios para comprender su funcionamiento, aspectos relevantes, casos de uso contemplados y tendencia de evolución.

### 2.2. Motivación:

A fin de despertar el interés de los alumnos en la temática a desarrollar, sería conveniente iniciar la intervención proponiendo un debate acerca de la cantidad de información de carácter personal que cedemos a diario a grandes empresas de software como Google, Facebook, Microsoft, etc...

Todas estas fuentes generan un inmenso flujo de datos que transferimos a empresas que se encargan de generar negocios con los mismos.

Es de destacar que Google ha promovido la adopción de desarrollos de código abierto que posibilitan el uso Datos Masivos tales como Apache-Hadoop[3], y también de herramientas de gestión y análisis de los mismos como Spark[4], Pig[5], Hive[6]. También

[3] http://hadoop.apache.org/

[4] http://spark.apache.org/

[5] http://pig.apache.org/

ofrece servicios de Aprendizaje de Máquina (Machine Learning[7]), los cuales se encuentran basados en técnicas de Ciencias de Datos.

El acceso a estas herramientas de código abierto ha brindado un gran impulso al desarrollo y adopción de las 2 tecnologías que estamos tratando. Sin embargo, este beneficio significativo para la sociedad conlleva consigo una evidente estrategia implementada por Google y otras firmas para afianzar su modelo de negocios vinculado con la inmensa cantidad de datos que recibe a diario de sus usuarios.

El debate propuesto tendría como finalidad, el crear conciencia en los alumnos acerca de la cantidad de información vinculada a nuestra privacidad que estamos brindando a diario. Por otro lado, nos brindará el pie para introducir la temática de Datos Masivos y su consecuencia directa de contar con técnicas y herramientas para su análisis (Ciencia de Datos)

Un disparador, a modo de ejemplo, a ser utilizado para abrir este debate con los alumnos, puede ser la noticia de la compra de la aplicación de comunicaciones de uso masivo WhatsApp, por parte de Facebook[8].

En febrero de 2014 Facebook adquirió WhatsApp pagando la increíble suma de 22.000 millones de dólares por la misma. Cuando Mark Zuckerberg (fundador y presidente de Facebook) fue consultado por el monto que se estaba pagando por la adquisición, dijo que la misma había sido una compra “barata”. En ese momento WhatsApp tenía 600 millones de usuarios a nivel global. Zuckerberg dijo que habían pagado 26,67 USD por usuario, mientras que él los valoraba en 100 USD aproximadamente a cada uno.

¿Por qué se está dispuesto a pagar tanto por la información? ¿Cuál es la modalidad y el alcance de estos negocios millonarios? ¿Cuál es el impacto que estos desarrollos y nuevos modelos de negocio están provocando en la sociedad, y en que cuanto nos afecta?. Estas son algunas de las preguntas que pueden motivar el debate a guiar.

.

### 2.3. Nuevos contenidos a incorporar:

A continuación, se enumeran tópicos que se consideran relevantes para el abordaje de la temática.

Los tópicos que se enuncian en esta sección tienen carácter de sugeridos. Es esperable, que cada docente, cátedra o unidad académica decida cuales contenidos incorporar, su enfoque y profundidad.

---

[6] http://hive.apache.org/

[7] https://cloud.google.com/ml-engine/

[8] http://www.lanacion.com.ar/1733420-aprueban-la-compra-de-whatsapp-por-facebook-22000-millones-de-dolares consultado el 13 agosto de 2017

### 2.3.1. Datos Masivos:

- ✓ Revisión de conceptos datos estructurados, semi-estructurados y no-estructurados. Modelo relacional de base de datos
- ✓ Modelo relacional vs. Bases de datos NoSQL
- ✓ Modelos de Bases de Datos NoSQL: Orientadas a columnas. Orientadas a documentos. Clave-valor. Orientadas a Grafos
- ✓ Características de infraestructura de Datos Masivos. Las 3 V: Volumen, Variedad y Velocidad
- ✓ Arquitectura de Datos Masivos. Modelo escalar vertical. Modelo escalar horizontal
- ✓ El caso Hadoop. Experiencia de Google. Ecosistema Hadoop

### 2.3.2. Ciencia de Datos:

- ✓ Revisión de conceptos de DataWarehouse y herramientas de Inteligencia de Negocio.
- ✓ Data Management
- ✓ Análisis Exploratorio
- ✓ Visualización de datos
- ✓ Aprendizaje de Máquinas (Machine Learning)
- ✓ Minería de Datos (Data Mining)
- ✓ Minería de Textos (Text Mining)

### 2.3.3. Casos de uso. Experiencias de implementación:

Solo para mencionar algunos que pueden utilizarse a modo de ejemplo para ejemplificar, se pueden citar:

- ✓ NEGOCIOS: modelos de predictibilidad de quiebras y de detección de fraudes
- ✓ MERCADOS: predicciones de comportamientos de mercados
- ✓ MEDICINA: personalización de tratamientos, predicción y prevención de problemas de salud en pacientes de riesgo
- ✓ CRIMINOLOGIA: generación de modelos espaciales de predictibilidad de mapas de crimen

### 2.3.4. El rol del profesional en ciencias económicas en referencia a estas tecnologías

Esta discusión, se debería alinear con una más genérica, la del rol del profesional en ciencias económicas en referencia a IT (Tecnologías de la Información).

La IFAC (2006), identifica, diferentes roles que puede desempeñar el contador con referencia a la Tecnología Informática (IT):

*"Contadores profesionales generalmente juegan importantes roles como gerentes, consultores y proveedores de aseguramiento en la adopción, despliegue y uso de diversas tecnologías de la información para organizaciones de todo tipo y tamaño"*[9]

Collazo J., Saroka R. (2010) también se refieren al rol del profesional en ciencias económicas respecto a IT. Estos autores lo visualizan, en forma más enfática aún que el documento de IFAC. Ellos expresan:

*"El profesional en Ciencias Económicas que se desempeña como ejecutivo, contador, gerente administrativo, gerente general, asesor administrativo, asesor contable o auditor no tiene otra opción que entender la tecnología informática y participar, …, en las acciones de la organización dirigidas a implantar y explotar los recursos informáticos"*

El documento de IFAC citado, hace también referencia a los conocimientos y habilidades que debe desarrollar el contador a los fines de desenvolverse en el ámbito de gestión estratégica de tecnología de la información:

**Área temática: Estrategia de Tecnología de la Información**[10]

| Conocimientos y habilidades generales en IT | Principal aspecto cubierto |
|---|---|
| **Estrategia corporativa y visión** | Problemática de negocios, interna y externa |
| | Factores que impactan en IT |
| **Evaluación actual y futura del ambiente IT** | Status actual de la entidad en referencia al uso de IT para soportar procesos de negocio |
| | Riesgos y oportunidades de IT |
| **Planeamiento estratégico de IT** | Visión a futuro sobre el estatus de los sistemas de la entidad |
| | Alineación futura de estrategia IT con estrategia de negocio |
| **Gobierno continuo y proceso de monitoreo de resultados** | Marco de trabajo para gobierno de IT |
| | Medición de resultados |

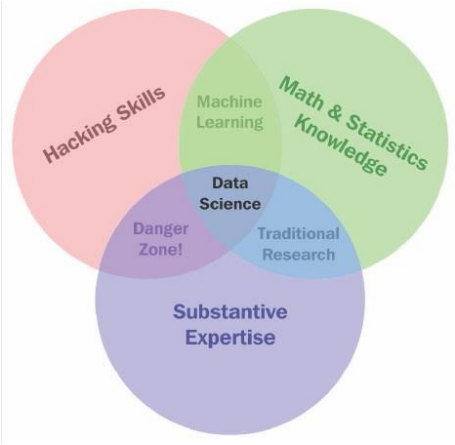
Como se desprende de esto, es preponderante la vinculación del profesional en ciencias económicas con respecto al uso estratégico de los recursos de IT. Este aspecto se verá potenciado en el análisis particular de su rol en referencia a Datos Masivos y Ciencia de Datos.

Un enfoque interesante sobre los conocimientos necesarios para particular de esta cuestión lo da el denominado "diagrama de Venn de Ciencia de Datos", propuesto por Ozdemir (2016).

[9] IFAC – International Federation of Accountants.(2006) Proposed International Education Practice Statement 2.1: Information Technology for Professional Accountants. Punto 7. Página 5 – Traducción del autor "Professional accountants often play important roles as managers, advisors and assurance providers in the adoption, deployment and use of various information technologies by organizations of all types and sises"

[10] IFAC – International Federation of Accountants.(2006) Proposed International Education Practice Statement 2.1: Information Technology for Professional Accountants. Página 23.Traducción del autor – Se suprimió última columna del cuadro en razón de solo tener carácter ilustrativo

**Diagrama de Venn de Ciencia de Datos**[11]




Ozdemir identifica 3 saberes requeridos para desarrollar e implementar soluciones de Ciencia de Datos:

- Conocimientos de Matemáticas y Estadísticas
- Conocimiento sustantivo (lo denominaremos conocimiento del dominio)
- Habilidades de hacking (lo denominaremos conocimientos de IT)

Si analizamos esta visión interdisciplinaria de la Ciencia de Datos, desde la óptica de incumbencias del profesional en ciencias económicas, podemos destacar:

- Básicamente se reconocen 3 ejes disciplinarios: conocimiento del dominio, estadísticas, conocimientos de IT
- En lo que respecta a conocimiento del dominio, tal como se expresa en el documento de IFAC y en la obra de Collazo – Saroka, el profesional en ciencias económicas se encuentra fuertemente involucrado en los procesos de negocio de las organizaciones a las que pertenece o asesora, desde su función de gerente, administrador, consultor, etc…
- En el ámbito del conocimiento de estadísticas, es también un área en donde se encuentra capacitado para actuar. Tal vez no con la profundidad de conocimientos de un profesional en ese tema (Licenciado en Estadística), pero sí con una formación básica en esa disciplina.
- Por último, en el aspecto de IT, la publicación de IFAC citada, destaca los conocimientos requeridos desde la valoración y concepción funcional de tecnologías de IT.

A modo de resumen, podemos considerar que, los saberes del profesional en ciencias económicas, como experto en el campo de conocimiento de dominio, gestor de recursos de IT, y su formación en técnicas estadísticas, abren la posibilidad de detentar una posición preponderante en esta incipiente disciplina de IT.

[11] Ozdemir, S. (2016), página 6

### 2.3.5. Aspectos éticos de la utilización de Datos Masivos

Davis, K (2012) identifica 4 elementos centrales a ser observados al discutir la ética del uso de Datos Masivos:

- ✓ Identidad
- ✓ Privacidad
- ✓ Propiedad
- ✓ Reputación

Basado en estos elementos, algunas de las discusiones planteadas por dicho autor en relación con la ética del manejo de datos personales son:

- ✓ Compra y venta de datos personales
- ✓ Dar accesos a los clientes a poder controlar cuales datos personales pueden ser explotados
- ✓ Alcance y aseguramiento del proceso de "anonimación" de los datos personales
- ✓ Propiedad y alcance de la propiedad de los datos personales
- ✓ Manifestación de valores de las organizaciones que usan los datos personales

Algunos aspectos relevantes de debate planteados por Zook M. y otros (2017) son:

- ✓ Considerar que los datos son "personas" y pueden producir daño
- ✓ Reconocer que la privacidad es más que un valor binario
- ✓ Mantenerse alerta con que se pueda re-identificar tus datos
- ✓ Práctica de la ética en el compartimiento de datos
- ✓ Consideración de la fuerza y limitaciones de tus datos
- ✓ Debatir la dificultad de elecciones éticas
- ✓ Desarrollo de códigos de conducta para las organizaciones, comunidad científica o industria
- ✓ Diseño de datos y sistemas para ser auditables

Sin dudas que el abordaje de una cuestión tan compleja como la ética en el uso de Datos Masivos, es tan amplio que podría derivar en otra propuesta completa de intervención académica independiente. No es el objetivo del presente trabajo, abarcar esa profundidad del tema en cuestión, sino plantearlo para que el/los docentes a cargo de la intervención evalúen su profundidad y alcance.

## 2.4. Cierre:

Desarrollar junto a los alumnos una puesta en común acerca de la irrupción que estas nuevas tecnologías generarán en el ámbito de la administración de organizaciones (en especial) y de la sociedad (en general).

Se propone al docente, conducir a la siguiente reflexión que guie la discusión:

Según la teoría general de sistemas, podemos diferenciar 3 elementos básicos en cualquier sistema: entradas (inputs), procesos (rules) y salidas (outputs)[12]. En la

[12] Von Bertalanffy, L. (1993)

concepción tradicional de la administración, conocemos las entradas de sistema, los procesos que se realizan y desconocemos las salidas.

Las técnicas modernas de Ciencia de Datos, por ej. en Aprendizaje de Máquina, abren un nuevo paradigma según el cual conocemos las entradas, conocemos las salidas, pero lo que desconocemos son los procesos que hacen que las entradas se conviertan en salidas.

La aplicación de estas nuevas técnicas puede brindarnos:

- ✓ Descubrimiento de relaciones de causas-efecto insospechadas que expliquen comportamientos o resultados anómalos de las salidas previstas del sistema,
- ✓ verificar si los resultados a obtener por medio de procesos o reglas a aplicar, se condicen con los resultados de comportamientos ya observados y analizados por técnicas sofisticadas de Ciencia de Datos
- ✓ descubrir desvíos entre modelos de simulación desarrollados con respecto a modelos predictivos obtenidos por la aplicación de Ciencia de Datos.

Dentro de este enfoque es deseable realizar un seguimiento acerca del impacto que estas nuevas tecnologías están generando en diferentes ámbitos, y que algunos autores se encuentran analizando.

Por ejemplo, Jim Gray de la Universidad de Berckley ya sostenía en 2007 que el impacto que estas nuevas tecnologías tienen, nos ubica en la etapa del "cuarto paradigma" de la ciencia, lo que denominó la e-science.

El cuarto paradigma de la ciencia, según Jim Gray[13]

| Paradigma | Concepto | Inicio |
|---|---|---|
| **Primer** | Ciencia empírica – observación de fenómenos | Miles de años atrás |
| **Segundo** | Ciencia abstracta – construcción de modelos teóricos | Cientos de años atrás |
| **Tercer** | Ciencia computacional – Simulación de fenómenos complejos | Últimas décadas |
| **Cuarto** | e-science – unificación de modelos, experimentos y simulación. Exploración de datos | Actualidad |

## 2.5. Evaluación de resultados:

Se propone realizar una evaluación de conocimientos teóricos transferidos por medio de un cuestionario específico, con especial énfasis en la evaluación de relaciones entre conceptos.

Esta evaluación se puede complementar con una breve monografía que gire sobre algunos de los temas debatidos, y sea desarrollada por grupos de alumnos.

[13] Elaboración propia en base a las presentaciones realizadas por Jim Gray en NRC-CSTB1 in Mountain View, CA, 11 Enero de 2007 – Ver: Tansley, S., & Tolle, K. M. (2009).

## 3. CONCLUSIONES

La velocidad de los cambios tecnológicos es incesante y afecta en forma directa nuestra forma de vida, relaciones y el modo en que se desenvuelven las organizaciones. Es lógico suponer que no se adapten los programas de asignaturas de Tecnología Informática, cada vez que un cambio menor, o de poco impacto relativo, se produce.

Sin embargo, hay ocasiones en las que nuevas tecnologías generan una disrupción evidente cuyas consecuencias es necesario estudiar.

Tal como se desarrolla en el presente trabajo, Datos Masivos y Ciencia de Datos, parecen enmarcarse dentro de esta categoría de tecnologías.

Por la relevancia que este cambio se espera produzca en métodos, saberes y técnicas generalmente aceptadas en administración, es que se postula iniciar el debate entre docentes de IT de adopción de los mismos en programas de estudio pertinentes.

## BIBLIOGRAFÍA

Allen, M., & Cervo, D. (2015). *Multi-Domain Master Data Management: Advanced MDM and Data Governance in Practice*. Morgan Kaufmann.

Bell, J., 2015. *Machine Learning: Hands-On for Developers and Technical Professionals.* Willey.

Bhansali, N. (Ed.). (2013). *Data Governance: Creating Value from Information Assets*. CRC Press.

Brom, F (2014). *Innovación estratégica disruptiva – El camino de la innovación en el ecosistema digital*. Edicom – Fondo Editorial Consejo

Campanaro, Rosa S., Diaz, Daniel J., Gardenal, Luciano, Marchese, Alicia G. , (2016), *Análisis de estados contables aplicando XBRL y herramientas de inteligencia de negocios*, DUTI 2016 - Bahía Blanca – Argentina

Colin Ware, E., 2004. *Information Visualization: Perception for Design*. Morgan Kaufmann.

Davis, K. (2012). *Ethics of Big Data: Balancing risk and innovation.* O'Reilly Media, Inc.

Collazo J., Saroka R. (2010). *Informática en las organizaciones*. Edicom – Fondo Editorial Consejo.

Foster Provost, T. F., 2013. *Data Science for Business*. O'Reilly.

Frampton, M. (2014). *Big Data Made Easy: A Working Guide to the Complete Hadoop Toolset.* Apress.

Fryman, L., Lampshire, G., & Meers, D. (2016). *The Data and Analytics Playbook: Proven Methods for Governed Data and Analytic Quality*. Morgan Kaufmann.

Harrington, P., 2012. *Machine Learning in Action*. MANNING.

Harrison, G. (2015). *Next generation databases: NoSQL, NewSQL, and Big Data: what every professional needs to know about the future of databases in a world of NoSQL and Big Data*. Apress (IOUG), New York, The expert's voice in Oracle.

Ian H. Witten, E. F. M. A. H., 2011. *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier / Morgan Kaufmann.

International Federation of Accountants ´(IFAC), 2006. *Proposed International Education Practice Statement 2.1: Information Technology for Professional Accountants*. https://www.ifac.org/system/files/meetings/files/2820.pdf . Observado Setiembre 2017.

KUNCHEVA, L. I., 2014. *COMBINING PATTERN CLASSIFIERS*. Second Edition ed. Wiley.

Kurbanoglu, S., Al, U., Erdogan, P. L., & Ucak, N. (Eds.). (2012). *E-Science and Information Management: Third International Symposium on Information Management*

*in a Changing World*, IMCW 2012, Ankara, Turkey, September 19-21, 2012. Proceedings (Vol. 317). Springer.

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *Big Data Imperatives*. Apress.

Ozdemir,S. (2016). *Principles of Data Science*. Pack Publishing.

Rahman, H. (Ed.). (2009). *Social and Political Implications of Data Mining: Knowledge Management in E-Government: Knowledge Management in E-Government.* IGI Global.

Redmond, E., & Wilson, J. R. (2012). *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf.

Stubbs, E. (2014). *Big Data, Big Innovation: Enabling Competitive Differentiation Through Business Analytics.* John Wiley & Sons.

Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery (Vol. 1).* T. Hey (Ed.). Redmond, WA: Microsoft research.

Tiwari, S. (2011). *Professional NoSQL.* John Wiley & Sons.

Tufte, E. R., 1990. *Envisioning Information.* GRAPHICS PRESS.

Vaish, G. (2013). *Getting started with NoSQL.* Packt Publishing Ltd.

Von Bertalanffy, L. (1993). *Teoría general de los sistemas*. Fondo de cultura económica.

Yu-Wei, Chiu (David Chiu), 2015. *Machine Learning with R Cookbook*. Pack Publishing.

Zook, M., Barocas, S., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., ... & Nelson, A. (2017). *Ten simple rules for responsible big data research*. PLoS computational biology, 13(3), e1005399.