Manual sobre utilidades del **big data** para bienes públicos

Entinema

Capítulo 8

Innovación basada en ciencia de datos, modelos y tecnologías

Diego May*, Frans van Dunné**

› Introducción

En la última década se fue haciendo cada vez factible obtener valor de los datos ya que era más barato generarlos, almacenarlos y procesarlos. Hace solo veinte años (1996) se hizo más rentable guardar los datos en computadoras que en papel (Morris y Truskowski, 2003). Las grandes empresas de tecnología están creando importantes ventajas competitivas a través del uso intensivo de datos. Los gobiernos están abriendo *datasets* relevantes que los desarrolladores y las empresas utilizan cada vez más. Y hay cada vez más soluciones y plataformas para procesar y crear valor a partir de los datos.

Las oportunidades de innovación a través de los datos se hacen cada vez más claras y los modelos, *frameworks* y prácticas van madurando en tanto estos se aplican a empresas de distintos tamaños en diferentes segmentos de mercado.

Esta tendencia que comenzó con las grandes empresas de tecnología como Google, Facebook, Twitter, AirBnB (por nombrar algunas) está llegando ya a otros niveles de empresas y en los próximos diez años será masivo. Todas las organizaciones contarán con infraestructura y tecnología para recolectar y almacenar datos, y deberán entender y aplicar ciencia de datos para sacar provecho de forma adecuada.

Además de nuevas oportunidades, este desarrollo también presenta nuevas necesidades ya que va a requerir profesionales que a todos niveles, desde los más técnicos hasta gerentes y líderes, puedan no solo entender el valor de los datos sino también interpretar los resultados y saber cómo actuar de acuerdo con los mismos.

* Cofundador de ixpantia (Ciencia de Datos) y de Junar (Datos Abiertos). MBA por el MIT Sloan School of Management.
** Cofundador de ixpantia. PhD en Biología de la Universidad de Amsterdam.

El objetivo de este capítulo es cubrir algunos de estos aspectos básicos sobre infraestructura y ciencia de datos. Los temas que se abordarán son los siguientes:

- Modelos generales sobre innovación basada en datos.
- El ciclo de innovación en ciencia de datos y un modelo que describe puntos de entrada.
- Un modelo para caracterizar los datos y según esto definir cómo afrontar distintos tipos de problemas y preguntas en ciencia de datos.
- Algunos comentarios acerca de plataformas, tecnologías y herramientas.

Este capítulo debiera dar una buena introducción a la temática y ofrecer algunas herramientas para aquellas organizaciones que están buscando desarrollar iniciativas de innovación basada en datos, así como también a profesionales que están interesados en esta temática.

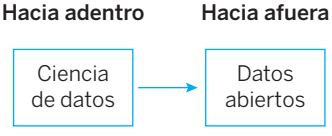
Dos aproximaciones a la innovación basada en datos

Cada vez es más fácil guardar mayores volúmenes de datos y hacerlos disponibles de formas adecuadas para que estos puedan transformarse en información relevante. Pero también es cierto que con el mayor volumen de datos existentes el análisis de los mismos se hace más complejo y es importante encontrar las formas adecuadas para generar dicha innovación a partir de los datos.

En esta sección hablaremos de dos grandes paradigmas a considerar a la hora de evaluar alternativas de innovación en base a datos. Las mismas son:

1. Innovación hacia fuera de la organización o alineada con el movimiento de *datos abiertos*.
2. Innovación hacia dentro de la organización o alineada con la analítica, *big data* y ciencia de datos.

Imagen 1. Modelos de innovación



Hacia afuera, innovación basada en datos abiertos

La innovación basada en datos abiertos se ve en general en el sector. En estos casos, los gobiernos (nacionales, regionales o municipales) abren datos a través de portales con la motivación de generar: transparencia, colaboración, participación, eficiencias e innovación.

Los gobiernos cuentan con muchos datos valiosos que podrían generar un alto impacto para los ciudadanos. Por otro lado, estos datos que tienen los gobiernos son recolectados y mantenidos gracias a los impuestos que paga la ciudadanía, por lo cual se considera que estos datos deberían estar disponibles de manera abierta, siempre que no se comprometa la privacidad o seguridad de los individuos.

Imagen 2. Etapas *open data*

El paradigma de colaboración e innovación en estos casos es simple: los gobiernos tienen los datos, los ciudadanos y el sector privado en general pueden contribuir a generar soluciones innovadoras estos datos. Y estas soluciones pueden suponer un impacto directo en la ciudadanía (*apps*, nuevos servicios) y en el gobierno (*apps*, mayor eficiencia), o contribuir al desarrollo de nuevas empresas o a la mejora de productos que tal vez más indirectamente impactan positivamente a la sociedad.

Para implementar estos programas se suelen seguir los siguientes pasos:

1. **Mapeo de los datos.** Típicamente existe una unidad (tecnologías de la información o desarrollo económico por dar un par de ejemplos) con la responsabilidad de realizar esta apertura de datos en el gobierno. Esta unidad trabaja con el resto de la organización para entender cuáles son los conjuntos de datos que los diferentes departamentos generan y mantienen.

Según este mapeo y al entender los pedidos de información y datos de la ciudadanía se genera una lista completa de la información que podría abrirse y después de pasar los filtros legales generarse un plan gradual de apertura de datos (*open data roadmap*).

2. **Creación de un portal de datos abiertos.** Existen diferentes formas de implementar un portal de datos abiertos. Son muy pocos los gobiernos que actualmente optan por desarrollar y mantener una solución pero existen algunos de estos casos. Típicamente los gobiernos optan por implementar una solución de *software* en la nube (Socrata, Junar, *open data soft*) o implementar alguna solución de código abierto que luego la organización puede mantener (*ckan, dkan, junar*). Cualquiera sea el portal de datos abiertos desarrollado por la organización, el mismo deberá permitir el acceso a los datos publicados considerando las características solicitadas por las distintas audiencias: ciudadanos (encontrar y entender datos a través de visualizaciones), academia (poder acceder a los datos crudos), desarrolladores y periodistas (poder tener acceso sistemático vía APIs).
3. **Generar programas de comunicación, promoción y utilización de los datos.** No basta con haber dado los pasos 1 y 2 para tener éxito en generar innovación. Es clave que las organizaciones que publican los datos, tanto en el sector público como privado, hagan esfuerzos para promover iniciativas que permitan la difusión y que incentiven la utilización de estos datos. Más allá de hacer prensa y difusión sobre los programas, se logra mucho cuando los gobiernos generan *hackathones*. Estos eventos que atraen a desarrolladores, diseñadores, *hackers* cívicos y al sector privado en general buscan que en periodos cortos de tiempo se generen algunas soluciones a problemas existentes en la ciudad, de acuerdo con desarrollos de aplicaciones fundadas en los datos que abren las instituciones de gobierno.

Son muchos los ejemplos de ciudades que han logrado generar impacto e innovación a través de programas de datos abiertos. En EE. UU. grandes ciudades como Chicago, San Francisco y New York y ciudades más pequeñas como Mesa (en Arizona) y Palo Alto (en California) han desarrollado programas muy completos. En Latinoamérica hay casos relevantes en Chile (Providencia, Puente Alto), en Perú (Miraflores, San Isidro) y en Argentina (Ciudad de Buenos Aires, Ciudad de Bahía Blanca) como también en México, Colombia y Brasil.

Los casos de éxito tienen en común:

- Realizan una apertura inicial bien promocionada.
- Generan algún evento de innovación o *hackathon*.

- Dan seguimiento a iniciativas interesantes y apoyan a emprendedores e innovadores.
- Tienen un *roadmap* de apertura datos que incluye datos en tiempo real (valiosos para innovar).

Lo que ha generado impacto en el sector gobierno ya está permeando al sector académico y al sector de impacto social (ONG, fundaciones) y existen casos de apertura de datos para generar innovación en el sector privado (competencias Kaggle, Netflix Prize). Y esto es solo el comienzo.

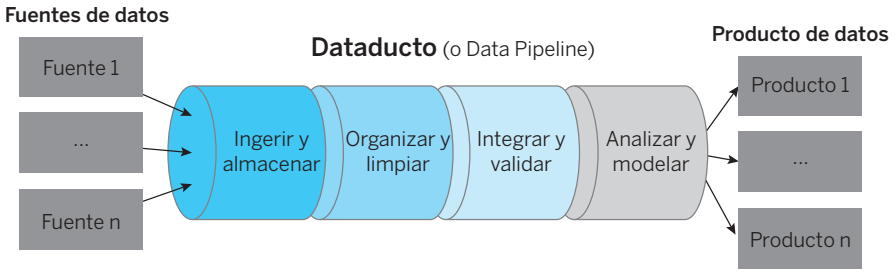
Hacia adentro, innovación basada en ciencia de datos

En las próximas secciones de este capítulo hablaremos en mayor detalle acerca de esta innovación basada en ciencia de datos. Lo que nos interesa ahora es poder dejar claro tanto la distinción como la relación entre innovación hacia dentro e innovación hacia fuera.

Cuando las organizaciones están generando innovación hacia dentro, deben antes que nada contar con recursos humanos que sepan utilizar técnicas, procesos y metodologías para interpretar y sacar valor de los datos que idealmente se deben encontrar alojados en una infraestructura que permita el fácil acceso a los mismos.

El científico de datos tendrá preguntas específicas que contestar y según esto podrá tener los *datasets* en una infraestructura de fácil acceso (un *lake* o lago de datos) y los hará interoperables a través de procesos de integración. También podrá ver qué datos externos a la organización pueden enriquecer el análisis con el objetivo de responder las preguntas. Finalmente podrá hacer el análisis y generar resultados, reportes y productos de datos. Todo esto se verá en próximas secciones y se puede ver esquematizado en la imagen 3.

Imagen 3. El proceso de ciencia de datos

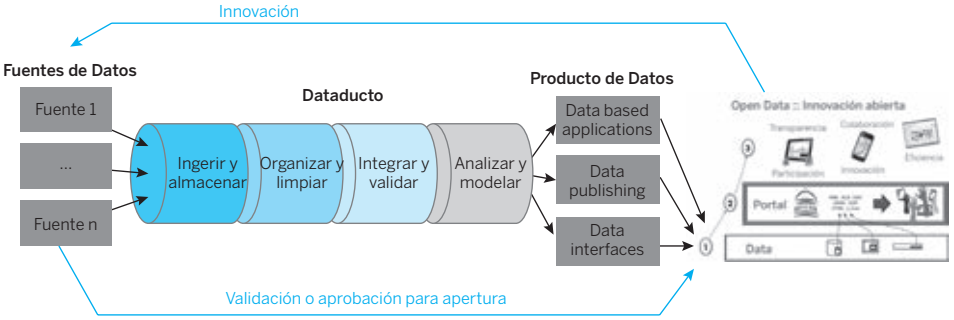


Los equipos que suelen trabajar en prácticas de ciencia de datos en las organizaciones son diferentes a los tradicionalmente conocidos como BI (*business intelligence*) o *inteligencia de negocios*. Principalmente porque en ciencia de datos se buscan equipos que operen con un método científico para tratar de explorar caminos no transitados y responder a preguntas nuevas, mientras que tradicionalmente el equipo de BI se dedicaba a interactuar con las bases de datos y sistemas para generar reportes con resultados específicos esperados y mayormente conocidos.

Por este motivo, según el tipo de organización y de la complejidad de los datos, así como de la importancia que se da a la innovación a través de estos, se definen distintos esquemas de trabajo. Los tres esquemas más conocidos son: equipos centralizados, equipos distribuidos o equipos híbridos. No es objetivo de este capítulo entrar en detalle en estos modelos, pero sí al menos describirlos para que los lectores puedan luego profundizar. Algunas empresas optan por tener un equipo de ciencia de datos central que atiende las distintas unidades de negocios. Otras organizaciones prefieren un equipo distribuido con científicos de datos operando en las distintas unidades de negocios. Finalmente, existen modelos híbridos donde se busca obtener los beneficios de un equipo centralizado de ciencia de datos que se retroalimenta y que permite alta innovación, junto con la ventaja que tienen los equipos distribuidos de operar de forma cercana a las unidades de negocios y entender con mayor profundidad las problemáticas que tienen que resolver estos grupos.

Es interesante ver que ambos modelos de innovación basada en datos (hacia dentro y hacia fuera) pueden interactuar y complementarse. Existirán "preguntas" que deberán responderse internamente o que al menos requerirán un preproceso a lo interno de la organización aplicando metodologías de ciencia de datos. Como resultado se pueden generar nuevas preguntas que podrán ser adecuadas para programas de innovación de base a datos abiertos llegando a poblaciones externas a la organización (ciudadanía, empresas, academia, empresarios, innovadores). Esta interacción está esquematizada en la imagen 4.

Imagen 4. Interacción entre innovación basada en datos hacia adentro (ciencia de datos) y hacia afuera (datos abiertos)



El ciclo de innovación basada en ciencia de datos

Diferentes modelos de innovación se han descrito con diversos modelos tales como lineales, cadenas interconectadas o circulares (Kline y Rosenberg, 1986). También existen modelos con diferentes enfoques como gestión (Bassiti y Ajhoun, 2013) o el proceso creativo (Buchanan, 1992). En esta sección miramos un ciclo de innovación que propone integrar el método científico en la aproximación. Esto no solo sirve para asegurar una metodología replicable y eficiente en producir preguntas que se dejan responder con los datos disponibles, sino también para plantear las pautas bajo las cuales podemos identificar prioridades en las propuestas de innovación.

El ciclo que proponemos para la innovación basada en datos se puede ver en la imagen 5. Como modelo para empezar a generar ideas y para identificar oportunidades de innovación se puede comenzar desde cualquier punto del ciclo. Comúnmente empezamos desde la búsqueda de una ventaja competitiva identificada o desde observaciones. Daremos ejemplos al describir los pasos uno por uno.

Pasos del ciclo de innovación basada en ciencia de datos

Cada uno de los pasos descritos a continuación puede ser el puntapié inicial para un proyecto de innovación a través de datos o simplemente una parte del proceso para gestionar dicha innovación.

Imagen 5. Ciclo de Innovación basado en datos

Búsqueda de ventajas competitivas

Mucha de la innovación que se genera en las organizaciones se basa en esta búsqueda constante de diferenciación. Tener una ventaja competitiva permite a las organizaciones ofrecer mayor valor al mercado a un menor costo, o cuando podemos pedir un mayor precio porque nos diferenciamos favorablemente de nuestros competidores. Los datos ofrecen múltiples formas para encontrar ventajas ya que permiten incrementar:

- la velocidad de las decisiones,
- la calidad de las decisiones,
- la velocidad de respuesta a clientes y a desarrollos en el mercado,
- la eficiencia de nuestros procesos de negocio o de producción.

Según estas búsquedas de ventajas competitivas se puede comenzar el proceso de innovación en base a datos para posteriomente generar las preguntas que serán respondidas con datos.

A partir de observaciones

En muchos casos el puntapié inicial para generación de innovación es el contar con una observación que proviene de alguna de las áreas de negocios. Aquí un par de ejemplos de este tipo de observaciones:

- "integrando estos procesos seríamos más eficientes",
- "segmentando los mercados seríamos más efectivos con nuestras ofertas".

Según estas observaciones se puede comenzar un proceso que nos lleve a buscar los datos que permitan validar (o no) la observación y así innovar. Por ejemplo, si hemos decidido que necesitamos incrementar la velocidad de responder a quejas de clientes, necesitamos observar dónde o cuándo se quejan los clientes. El mejor caso es cuando llaman al *call center* o a nuestros representantes de ventas. Pero muchas empresas han visto que frecuentemente las quejas se hacen de forma indirecta, a través de medios de comunicación sociales como Facebook o Twitter.

El ejemplo anterior también pudo haber sido nuestro punto de comienzo: alguien en la organización observa que las quejas no van a través de canales ya identificados, sino a través de medios sociales.

Preguntas

Las preguntas que se hace la organización pueden también ser una forma de comenzar estos procesos de innovación en base a datos. Debajo algunos ejemplos de preguntas:

Las observaciones nos llevan a formular preguntas que nos indican dónde, cuáles y qué tipo de datos necesitamos colectar. Siguiendo el ejemplo de arriba te podrías imaginar preguntas como:

- ¿Qué medios reciben más quejas?
- ¿Qué medios reciben las peores quejas?
- ¿Dónde se ven más quejas?
- ¿Dónde se ve más discusión sobre quejas?
- ¿Cómo influye una queja el sentimiento sobre nuestra marca?

Estas preguntas nos llevan luego a explorar los datos que nos llevarían a las respuestas y así comenzar un nuevo proyecto de ciencia de datos en la organización.

Datos

En muchos casos las organizaciones llegan a la conclusión de que cuentan con muchos datos valiosos y se dan cuenta de que es posible que estos datos, bien utilizados, permitan responder preguntas y generar mayores eficiencias y ventajas competitivas. Este *approach* es cada vez más relevante dado que son muchas las organizaciones que al ver el advenimiento de *big data* y ciencia de datos sienten que tienen que hacer algo.

El primer paso en estos casos es entender los datos con que se cuenta y la arquitectura de la información en la organización. De acuerdo con esto, después tiene

sentido tratar definir alguna pregunta puntual que pueda permitir el desarrollo de un primer proyecto de ciencia de datos.

Traducir preguntas en la necesidad de tener datos ya empieza a explicar por qué hablamos de ciencia de datos. Se acerca más al diseño experimental porque las decisiones sobre cuáles, cuántos y qué tipos de datos recogemos tienen un efecto directo sobre las decisiones a las cuales podemos dar soporte.

Al trabajar los datos también hay que tener en cuenta los posibles análisis y formas que estos pueden tener.

Productos de datos

Aunque los productos de datos suelen ser un resultado de un proyecto de innovación a través de datos, dada la experiencia previa existente en inteligencia de negocios en las organizaciones suele suceder que un resultado de un cubo o un reporte puede generar un nuevo proyecto de ciencia de datos.

En muchos casos, el proyecto de innovación en base a datos habrá comenzado desde alguna pregunta u observación y esto concluirá en el desarrollo de un producto de datos. Para mejorar una parte de la organización, por ejemplo, el proceso de negocio llamado "atención al cliente", necesitamos traducir lo que aprendimos de los datos en un producto. Hablamos de un producto en un sentido muy amplio porque puede incluir cosas como:

- Una gráfica o series de gráficas que nos ayudan a ejecutar un trabajo o tomar decisiones.
- Una fuente de datos trabajados que son leídos por otro proceso automático para mejorarlo.
- Un informe que dirige acciones entre personas.
- Una alarma que le indica a una persona o a una máquina que ha de tomar acción.

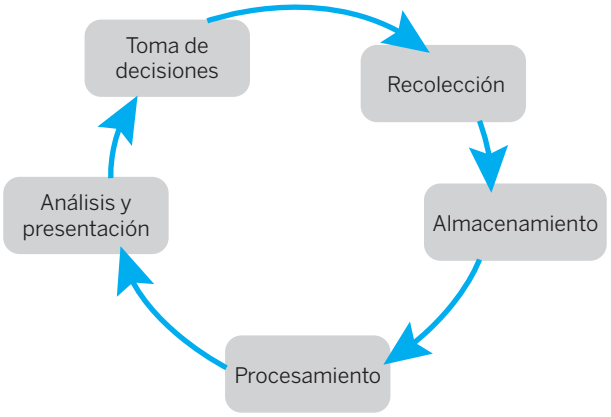
No podríamos dar una lista completa de ejemplos, ya que muchos de estos aún están por descubrirse. La innovación basada en datos está en movimiento.

Ciclo de valor IBD (innovación basada en datos)

En el ejemplo anterior llegamos a la siguiente iteración del ciclo cuando seguimos buscando dónde y cómo innovar con base en datos. Por otro lado, los productos que hemos generado entran al ciclo de valor de innovación basada en datos (*DDI Value Cycle*) como se presenta en la imagen 6. En este ciclo mejoramos

el producto de datos de forma continua y tenemos un marco de referencia para medir la calidad de cada paso.

Imagen 6. Ciclo de valor de datos (basado en Hayden *et al*., 2015)



Recolección

La recolección de datos, una vez que se ha llevado el producto a producción (implica que ya está operativo y utilizado por usuarios reales), incluye todos los pasos desde la fuente (que puede ser un cliente, un agente de ventas o una máquina) hasta el momento en que lo almacenamos. Esto incluye los posibles controles de disponibilidad y calidad que hayamos definido.

Almacenamiento

El almacenamiento de datos tiene que cumplir varias funciones importantes para asegurar que tenemos los datos disponibles en el plazo necesario:

- La forma de acceso, al nivel de agregación necesario para el consumidor.
- El gobierno del acceso (*data governance*) para que solo los indicados tengan acceso a los datos.
- La calidad del almacenamiento (si lo necesitamos en 30 años, ¿cómo nos aseguramos de que no hay cambios o borrado de datos inesperados?).
- La calidad de *retrieval*, incluyendo la gestión de metadatos, la auditoría de uso y aplicación.

Procesamiento de datos

Los datos crudos (no procesados), si bien tienen mucho valor latente, difícilmente están listos para ser consumidos y generar valor directamente. El valor se

añade en el momento de procesarlos, y la ventaja competitiva de nuestras empresas de cara al futuro está en gran parte en nuestros activos de procesamiento de datos: algoritmos y dataductos (del inglés *data pipelines*).

Es muy importante que en el procesamiento de datos se apliquen las reglas de gobierno de datos también, para poder validar que en el momento de procesamiento no se rompen las reglas.

Análisis y presentación

En este paso se puede ver cómo empacar los datos o presentarlos para que sean consumidos fácilmente por los distintos actores o *stakeholders*.

Toma de decisiones

La toma de decisiones es quizá el momento que se deja medir más fácilmente de cara al negocio. Si la calidad se incrementa porque se mide mejor, hay mejor conversión en ventas o una mejora en indicadores operativos, sabemos que nuestros productos de datos están cumpliendo con sus objetivos.

No basta con medir la mejora de toma de decisiones una sola vez y declarar un producto de datos como un éxito. Hay muchos puntos dentro del ciclo de valor DDI que pueden cambiar la calidad de la toma de decisiones. Quizá la decisión más importante es ver si aún estamos evaluando las preguntas correctas, para lo cual hay que regresar al ciclo de innovación basada en datos.

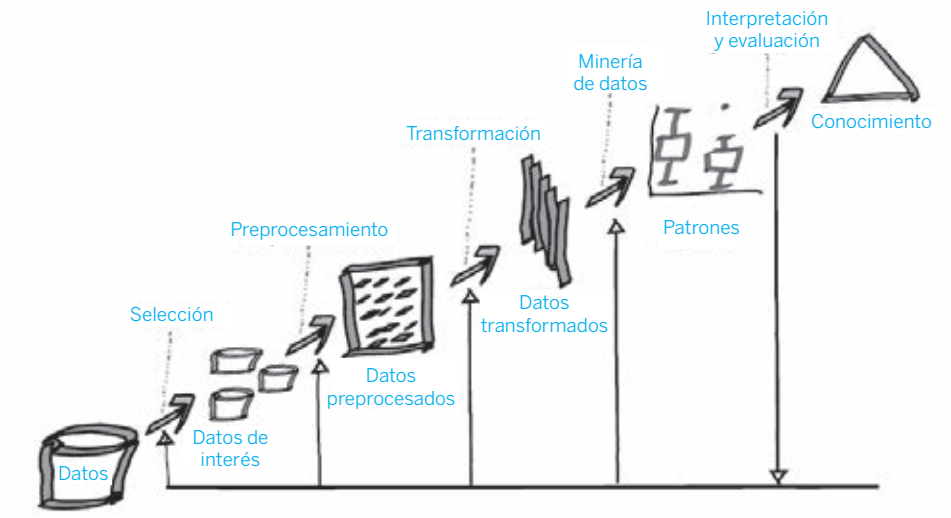
Es importante aclarar a estas alturas que en algunas empresas (StitchFix es un buen ejemplo) los productos de datos pueden ser composiciones sofisticadas de algoritmos y procesos. En estos casos, la mejora continua de estos productos puede implicar la interacción entre algoritmos y el criterio experto de los humanos (en el caso de estas empresas, los diseñadores que toman las recomendaciones de las máquinas o incluso las devoluciones de los mismos clientes).

› ¿Tengo datos, ahora qué?

En la práctica descrita arriba operan científicos de datos y consultores de negocio mano a mano para poder identificar las prioridades de innovación. Pero un momento clave en todo el proceso es el momento en que hay datos disponibles y en coro suena la pregunta: "¿Tengo datos, ahora qué?".

Quizá el modelo más conocido para llegar desde datos (crudos) a conocimiento (el producto final de datos) es el de Fayyad *et al.* (1996) llamado "descubrimiento de conocimiento en bases a datos". Los pasos que describe este modelo se pueden aplicar en casi todas las circunstancias y ayudan a entender el proceso (imagen 7).

Imagen 7. El proceso de descubrimiento de conocimiento en bases de datos (basado en Fayyad *et al*,. 1996)



Lo primero que necesitamos identificar es dónde están los datos. En un contexto organizativo esto puede ser un ejercicio de mapear dónde hay datos, para lo cual nos debemos asegurar de incluirlos todos y no solamente aquellos que están mapeados y se manejan a través de un protocolo o método de gestión establecido. En otras palabras, podemos tener acceso a datos en bases de datos "oficiales" dentro de la organización, pero probablemente hay datos en bases de datos no oficiales o semioficiales. Hay datos en diferentes estados de validez en hojas Excel y en documentos. Además, y cada vez con más relevancia, hay bases de datos externas que son de interés, ya sean datos abiertos o datos relevantes para la industria en la cual estamos.

De todos estos datos necesitamos identificar cuáles son los de mayor interés, por ejemplo, usando el ciclo de innovación basada en datos descrito anteriormente. Al identificarlos ponemos en marcha un proceso técnico para seleccionarlos

(con SQL, R-dplyr, MapReduce, etc.). Esto no da la selección de datos, probablemente tenemos que procesarlos para que se puedan usar para análisis. Por ejemplo, necesitamos incrementar el valor de los datos sacando todos los errores de ortografía en una categorización de interés. O tenemos datos en dos o más bases de datos separadas que queremos integrar.

Estos pasos del preprocesamiento también se traducen a código y pasos operativos en tecnología. Y con ellos estamos construyendo el siguiente paso en nuestro dataducto, a través del cual vamos a asegurarnos de que podemos repetir todos los pasos. Esto lo hacemos no solamente para protocolizar de forma automática los pasos para ingerir datos nuevos de forma rápida, sino también para poder implementar correcciones y mejoras en el proceso a medida que vamos aprendiendo más de cada iteración de nuestro ciclo de innovación y ciclo de valor.

Los datos preprocesados sirven para un análisis exploratorio que nos va a indicar qué transformaciones tenemos que hacer en los datos para que se puedan analizar para reconocer patrones y construir modelos predictivos. Por ejemplo, necesitamos convertir pies a metros para tener una sola dimensión de distancia en el conjunto de datos completo. O tenemos datos de interés sobre una cantidad que aún están incluidos como texto.

Es laborioso llegar al punto en que los datos están listos para aplicarles métodos de estadísticas y *machine learning*. Una forma de evitar malgastar recursos en limpiar datos que no sirven tanto como pensamos es definir pruebas de conceptos sobre un subconjunto de los datos para valorar el resultado y el impacto que tendrá sobre nuestro negocio.

El paso que Fayyad llama "minería de datos" (imagen 7) es descrita por Wickham y Grolemund (2016) como un ciclo que requiere que sepamos cómo tratar los datos de interés. Aquí entra la necesidad de saber sobre estadísticas y *machine learning* para poder definir qué métodos se pueden aplicar, tanto de acuerdo a lo que los datos permiten, como a lo que necesitamos para obtener respuesta a la pregunta que hemos formulado. Los pasos de transformar, modelar y visualizar datos en un ciclo estrecho e iterativo lleva cada vez a mayor conocimiento sobre el fenómeno bajo estudio.

El objetivo es poder comunicar de forma efectiva, y esta comunicación es lo que anteriormente hemos llamado producto de datos. El objetivo del proceso total es identificar la metodología, el modelo, el algoritmo o la visualización que puede ser la base para el producto de datos que estamos construyendo.

Esto nos lleva a las tres preguntas esenciales que hay que considerar para definir el modelo, algoritmo o visualización que va a formar la base para el producto de datos. Las presentamos en la imagen 8. En resumen son:

1. ¿Qué quiero con mis datos?
2. ¿Qué características tienen los conjuntos de datos?
3. ¿Cómo son las variables de mis datos?

Imagen 8. Un esquema para marcar las preguntas y respuestas que determinan qué métodos se pueden usar para analizar datos

¿Qué quiero con mis datos?

Lo principal es saber qué es lo que quieres o en otras palabras: ¿a qué pregunta quieres dar respuesta? A grandes rasgos vemos que comúnmente lo que queremos hacer con los datos es comparar, agrupar, predecir, reconocer o asociar.

- **Comparar.** Cuando queremos comparar datos necesitamos saber de antemano qué grupos tenemos. Esto parece lógico, pero muchas veces pasa que queremos comparar los tres grupos meta más importantes cuando lo que realmente queremos es "identificar nuestros tres grupos meta más importantes". Casos más claros son, por ejemplo, la comparación del desempeño de mercados regionales. Es importante notar que la decisión de qué método aplicar para poder comparar grupos es en la mayor parte una razón estadística (hay métodos paramétricos, no paramétricos y de rangos por dar algunos ejemplos) que depende de la tercera categorización que se describe más adelante, como son las variables de mis datos.

- **Agrupar.** Otro popular objetivo es agrupar. Es una forma natural de enfrentar la diversidad en nuestro entorno, donde nuestro cerebro busca rasgos comunes para categorizar objetos. Así reconocemos un tomate naranja, aunque solamente hemos estado expuestos a tomates rojos y verdes antes. Al introducir formas de agrupar en nuestros negocios es importante saber qué pasos en el tiempo queremos revisar en la agrupación que tenemos. ¿Es algo para revisar de forma continua?, ¿con qué periodicidad?, ¿cada segundo, cada mes, cada año? La respuesta tendrá poco que ver con datos y mucho que ver con objetivos de negocio.
- **Predecir.** Con predicción nos referimos a la construcción de modelos con datos que tenemos disponibles, para poder decir algo sobre situaciones de los cuales no tenemos (todos) los datos disponibles aún. Y predicción se puede hacer a diferentes niveles de exactitud y de precisión según el contexto. Si vendo helados quisiera saber "más o menos" qué caluroso va a ser el año entrante. Pero si tengo un modelo para predecir cuándo reemplazar un componente de un motor de avión, el "más o menos" ya no es suficiente, tiene que ser con un muy alto grado de precisión.
- **Reconocer**. Cuando queremos reconocer patrones queremos contar con la ayuda de un computador para, por ejemplo, reconocer la cara de las personas al entrar a un estadio o estación de metro. O queremos reconocer el tamaño y la forma de una piña para identificar estos datos como indicadores calidad de exportación. Hay un sinnúmero de oportunidades de negocio en la industria en esta temática, sobre todo, en el control de calidad de sistemas de producción.
- **Asociar**. Como último caso están las asociaciones, como, por ejemplo, las correlaciones. Son relaciones numéricas entre variables que no necesariamente indican una causa y efecto, pero indican una tendencia. Pueden ser una buena base para una observación con la cual arrancar nuestro ciclo de innovación basado en datos, por ejemplo. Pero también se usan con frecuencia para dar soporte a decisiones sobre mercadeo y optimización de procesos.

¿Qué características tienen los conjuntos de datos?

Sabiendo lo que queremos con nuestros datos, el siguiente paso es identificar las características principales de los datos. ¿Cuál es el volumen de los datos? ¿Qué grande es la variedad de los datos? ¿Se producen o hay que ingerirlos a cierta velocidad? ¿Qué veraces son y cuál es su valor?

- **Volumen**. Cuando hablamos del volumen de los datos, buscamos indicar cuántos puntos de medición hay o cuántos *bytes* si se tratan de texto o

imágenes. Los problemas se dan cuando hay más datos de lo que cabe en la memoria del computador que estás usando, o peor, más de lo que cabe en el disco duro de tu servidor. Por el otro lado, también hay problemas cuando el volumen es muy bajo y tienes pocos datos y muchas preguntas.

- **Variedad**. Cuanto más homogéneas sean las variables, más fácil será trabajar los conjuntos de datos, y esto marcará la variedad de los datos. Por ejemplo, si solo tengo datos numéricos, o mejor aún solo datos numéricos continuos, tengo más posibles metodologías que si tengo 100 GB de texto con comentarios en palabras en 40 lenguajes. O si tengo una mezcla de datos numéricos, texto, vídeo e imágenes. Cuanto más diferentes son los tipos de variables en mi conjunto de datos, más alta es la variedad.
- **Velocidad**. Cuando los datos se generan continuamente y fluyen con cierta velocidad (*streaming data*) se requieren soluciones especiales. Cuando la velocidad es baja, podríamos, por ejemplo, recalcular el promedio de una variable cada ciertos intervalos de tiempo. Pero cuando la velocidad del flujo de información incrementa muy pronto, ya no tenemos la velocidad de cálculo disponible para recalcular con todos los datos en tiempo real.
- **Veracidad**. También es importante saber qué datos reflejan hechos reproducibles. En otras palabras, qué veraces son los datos que tenemos. Por ejemplo, si uso datos experimentales, voy a tener mayor confianza en mis datos que si hago el experimento yo mismo. Si les pido a 20 personas al azar repetir el experimento, es posible que la confiabilidad (a priori) de los datos sea más baja. Algunas personas quizá no entiendan el experimento o se inventen resultados para poder entregar más rápido. Igual pasa en ciencia de datos, sobre todo en datos de empresas grandes, donde a veces ya nadie sabe de dónde vienen los datos, o se trata de encuestas donde no todos los que responden tienen interés en responder con precisión.
- **Valor y costo**. Por último, vale la pena pensar en el costo o valor de los datos que tengo disponibles o que necesito conseguir. Hay conjuntos de datos que han costado millones de dólares reunir (valor en dinero). O donde un solo punto extra, cuesta días, meses o más en conseguir (valor en tiempo). También hay análisis donde un error puede suponer perder millones.

¿Cómo son las variables de mis datos?

Por último, y ya a un nivel de detalle del conjunto de datos específico que vamos a incluir en un modelo, algoritmo o visualización, es si los datos son continuos, ordinales, nominales, texto o imágenes. Por lo general este es el primer capítulo de todo libro de introducción a estadísticas porque tiene impacto directo en qué

métodos puedo aplicar y cuáles no. Cada tipo de datos se comporta de una forma propia al someterlos a análisis.

- **Continuos**. Datos continuos son los que se pueden dividir sin límite. Por ejemplo, en una escala de temperatura se puede medir en grados Celsius o fracciones hasta tal punto que pueda obtener precisión de mi instrumento de medir.
- **Ordinales**. Datos ordinales tienen un orden, por ejemplo, grande, pequeño y mediano, pero no los puedo fraccionar. No hay una medida entre grande y pequeño.
- **Nominales**. Datos nominales son los que tienen un nombre, pero no un orden predeterminado. Son datos que podríamos tener en el CRM de la empresa, tales como el oficio de las personas, el género o el tipo de dispositivo móvil que usan.
- **Palabras**. Cada vez más se ven palabras como variable. El contenido de documentos, de e-mails, de mensajes en Twitter u otros medios sociales. Carecen de una estructura previa y tienen su propia familia de metodologías para analizarlos.
- **Imágenes**. Por último, tenemos imágenes, ya sean estáticas (fotos) y dinámicas (vídeos). Estos también son relativamente nuevos y tienen menos historia de análisis estadístico que los demás. Pero quizá tienen más historia de análisis en *machine learning* e inteligencia artificial.

Resumiendo

Obtener datos automáticamente nos obliga a pensar en qué hacer con ellos. Aun cuando hemos pensado en cómo analizarlos antes de recolectarlos es probable que encontremos sorpresas que nos permitan dar respuestas a preguntas que no habíamos anticipado, o que encontremos límites a la aproximación que habíamos planteado. Es indispensable estar claros sobre lo que se quiere obtener, cómo se caracterizan los conjuntos de datos e identificar los tipos de variables que nuestros conjuntos de datos contienen.

Plataformas, tecnologías y herramientas

En gran parte el auge de la ciencia de datos se debe no solo a la mayor disponibilidad de datos sino también a la disponibilidad de infraestructura a bajo costo que nos permite almacenar, procesar, analizar los datos y presentar los resultados de tal forma que humanos y máquinas pueden tomar mejores decisiones más rápidamente.

Es un mundo en sí hablar del desarrollo de tecnología. De entre un tema tan amplio vamos a destacar los siguientes aspectos: infraestructura, lenguajes y algoritmos, y Data Ops. Esta sección busca dar una visión amplia para tener una base que permita entender lo que se encuentra en el mercado, tanto a nivel de necesidades como de soluciones.

Infraestructura

No es muy difícil recolectar tantos datos que ya no caben en una hoja de cálculo. Para algunos, ¡eso ya es suficiente razón para pensar que tienen un *big data*! En la práctica, definir *big data* en términos de volumen no es muy preciso. ¿Es lo que no cabe en un solo disco duro o es lo que requiere almacenamiento en clústeres de servidores? ¿O quizá es lo que no cabe en una cantidad pagable de memoria RAM, digamos para hoy en día más de 512 GB, y nos obliga a paralelizar nuestros algoritmos de análisis?

Los datos vienen en tres formas: estructurados, semiestructurados y no estructurados. Es más, una buena definición de *big data* no toma como base el volumen (la cantidad), sino la gran diversidad y falta de estructura (la calidad) de los datos para dar una definición (Letouzé, 2015). Este cambio de énfasis de cantidad a calidad se refleja en las soluciones tecnológicas que se han desarrollado tanto a nivel de almacenamiento y gestión de los datos, como en el procesamiento de datos y la gestión de dataductos.

Gestión de datos en big data

Quizá cuando pensamos en datos lo primero que viene a la mente son sistemas de archivos (las carpetas y las archivos con los que trabajamos a diario en nuestros ordenadores) y bases de datos. Ambas formas de almacenar información llegan a un límite cuando estamos usando un solo servidor, y para ambas hay métodos para extender el almacenaje a un conjunto (un clúster) de servidores, de tal forma que podemos acceder a los datos como si estuvieran en un solo sistema. Hay coherencia.

Un ejemplo conocido de sistemas de archivos distribuidos es el sistema de archivos Hadoop (HDFS), el componente de almacenaje por defecto de Hadoop. Encima de esto, Hadoop da varias formas de obtener acceso, y el más conocido es MapReduce. Mapear y reducir datos es necesario cuando hemos distribuido los archivos en múltiples servidores, porque cualquier búsqueda que hago en los datos, lo tengo que repetir en todos los servidores que son parte de mi clúster en forma paralela. Y el resultado tiene que ser un resultado agregado de los resultados de cada servidor.

HDFS y MapReduce representan una forma de gestionar muchos datos de gran diversidad. Pero la desventaja es que es costoso de implementar y relativamente lento. Cualquier búsqueda se necesita repartir sobre múltiples servidores. Otros sistemas de almacenaje distribuido que solo mencionamos son GFS (Google File System), MapR FS, CEPH, Lustre y Terragrid, entre otros.

Por otro lado está el desarrollo de las bases de datos NoSQL. A diferencia de bases de datos como Microsoft SQL, MySQL y PostgreSQL, estas no consisten de tablas de datos relacionadas entre sí con columnas llave. Las bases de datos NoSQL consisten de objetos con una llave única para identificar el objeto. Esto puede ser un documento JSON, en los *document stores.* Puede identificar un dato único, como en los *key-value stores,* puede ser una fila de un *column store* o un nodo único que está conectado con otros nodos a través de bordes en un *graph database*.

La gran ventaja que tienen las bases de datos NoSQL es que brindan más velocidad de acceso a los datos (dada una implementación correcta) y soluciones especializadas para aplicaciones geográficas. Además muchos tienen una mayor tolerancia a errores, aunque esa tolerancia tiene un precio en términos de consistencia. Casi todas las bases de datos NoSQL permite despliegue distribuido, lo que da la posibilidad de trabajar con datos a gran escala.

Por último, las base de datos SQL, que no han perdido su popularidad por varias razones. Porque se trabaja con ellas desde hace décadas y, por lo tanto, son fáciles de implementar y dan valor a un costo conocido y bajo. Además, por su estructura relacional imponen un orden, una calidad de los datos, lo cual tiene beneficios a largo plazo en términos de gestión de calidad y a corto plazo porque es más fácil acceder los datos para finalmente procesarlos y analizarlos.

Como regla simple es bueno tratar de buscar almacenar datos estructurados en forma estructurada y si no podemos evaluar trabajar con datos semiestructurados en una base de datos NoSQL apropiada. Los datos no estructurados (por ejemplo, el contenido de libros, grabaciones de vídeo, imágenes, contenido de e-mails, etc.) tienen retos más grandes para hacer preguntas que añaden valor.

Para todas estas soluciones tenemos servicios disponibles en plataformas de IAAS (infraestructura como servicio) de proveedores como Amazon, IBM, Microsoft y Google.

Los lenguajes de datos

Hace diez años la lengua franca de datos era SQL como se refleja en su nombre: es la abreviación de *standard query language*. Los analistas eran estadísticos que usaban aplicaciones y entornos de análisis especializados y costosos como SAS y SPSS. Pero en los últimos diez años hemos visto un proceso impresionante de democratización de análisis impulsado por dos lenguajes R y Python. Tanto que recientemente R superó a SPSS en su uso en publicaciones académicas (Muenchen, 2016).

Para tratar de explicar la popularidad de R y Python, y poner a ambos en el contexto de *big data* vamos a tratarlos por separado. Y al hacerlo somos muy conscientes de que cualquier organización que usa Hadoop va a tener Java como lenguaje de desarrollo principal. Para optimizar algoritmos en producción vamos a necesitar C++ o Julia. Pero el objetivo aquí es no entrar en detalles especializados pero describir el desarrollo a grandes rasgos. Y el punto de entrada para análisis para la mayoría de nosotros es R.

El desarrollo de R comenzó en 1997 cuando John Chambers comenzó a automatizar el uso de unos algoritmos escritos en FORTRAN. Como punto de referencia usó la descripción del lenguaje S, del cual R es una implementación. Hasta el día de hoy vemos que R es un lenguaje para agilizar el uso de algoritmos ya escritos por otros. Muchos de los algoritmos usados en R hasta el día de hoy están escritos en FORTRAN (muchos otros en C++). Y esa también es parte de la razón de su popularidad.

R es un lenguaje específico para estadísticas (*domain specific language*) y está enfocado en el uso interactivo para trabajar con conjuntos de datos. Esto es de enorme ventaja para el que no tiene un enfoque de programación, porque la sintaxis refleja la forma de trabajar y hacerse preguntas de alguien interesado en datos y no en programación. Es por esto que muchas empresas (AirBnB, StackOverflow) usan R para diseminar algoritmos dentro de sus empresas.

Por otro lado esta Python, un lenguaje general que tiene su lugar y fama dentro del mundo de la ciencia de datos por unas bibliotecas reconocidas como NumPy, SciPy, Pandas y SciKit. La razón principal para escocer Python para análisis es porque ya se conocía el lenguaje previamente, o porque hay razones puntuales de integración o de la disponibilidad de un algoritmo.

Más común es la posición de Python como el lenguaje para la construcción de dataductos. Por ejemplo, para construir flujos de trabajo automatizados tenemos

varios paquetes en Python disponible que son capaces de orquestar y monitorear tareas de ETL y análisis. Proyectos como Dask, Airflow, Pinball y Luigi son conocidos en este contexto.

DataOps

Hay otro desarrollo importante, que trae su nombre del desarrollo de *software*. DataOps es un término nuevo que busca describir el conjunto de mejores prácticas para mantener el desarrollo, testeo y despliegue de productos de datos dentro del margen de responsabilidad del autor. En otras palabras, a la hora de querer poner un producto de datos en producción no queremos tener que pasar por múltiples capas de protocolos para realizar el testeo y despliegue. Ser flexible y rápido en este desarrollo es un indicador importante de éxito.

Una parte importante del mundo de DataOps es trabajar bajo una arquitectura de microservicios. Esto significa que permitimos que la solución que tenemos operando en cualquier momento sea una colección fluida y mutable de soluciones puntuales. Todos los servicios están unidos de una forma que se denomina acoplamiento ligero. Si uno de los servicios deja de funcionar genera un aviso de que hay un error, pero el servicio como total sigue funcionando.

Una forma de imaginar esto es en un dataducto donde tenemos múltiples fuentes de datos, que cada uno pasa por diferentes pasos de preprocesamiento y transformación antes de alimentar un modelo predictivo. Podemos identificar la carga computacional de cada paso para asignar infraestructura a medida, y de esta forma optimizar el uso. Todas las plataformas IAAS de IBM, Microsoft y Amazon brindan servicios ya prefabricados que toman parte del dataducto y lo ejecutan. Sobre todo en el área de análisis hay una diversidad creciente de servicios disponibles, ya sean generales como *machine learning* (por ejemplo, Azure ML, Bluemix Predictive Analytics, Amazon ML) o específicos como análisis de sentimientos o reconocimiento visual.

Estos servicios los podemos integrar dentro de nuestro dataducto con servicios propios, por ejemplo, desplegados en contenedores Docker. Un contenedor Docker contiene el mínimo absoluto de carga de sistema operativo y permite crear servicios donde no instalamos todo un servidor en una máquina virtual, pero solo el *software* mínimo para ejecutar una tarea. Esto nos ahorra infraestructura en términos de tiempo de computación y uso de memoria. Además, nos permite crear dataductos resilientes de los cuales podemos incrementar la escala de aquellos servicios que requieren incremento.

Conclusiones y recomendaciones

En la última década hemos visto cómo los titanes de tecnología como Google, Facebook, Twitter, AirBnB y otras de este calibre han ido incorporando el análisis de datos de forma masiva para generar ventajas competitivas o incluso nuevos negocios.

También se están dando modelos como el de la empresa Stitch Fix (por poner un ejemplo), donde se parte de un modelo de negocios en el cual el uso de datos y la aplicación de algoritmos es el diferencial principal y la base sobre la cual se construye el negocio. En este caso, los datos permiten hacer recomendaciones que finalmente son curadas por humanos en un proceso iterativo hombre-máquina.

Además hemos visto cómo las grandes empresas y bancos han comenzado a incorporar a sus tradicionales equipos de *business intelligence* algunos procesos y unidades para descubrir cosas nuevas en base a datos, y así generar innovación y diferenciadores.

Esta es una tendencia que comenzó con las grandes empresas de tecnología y los *startups* tecnológicos enfocados en la temática. Ahora está siendo también incorporada por los grandes bancos y conglomerados. Pero está muy claro que en breve esta tendencia y forma de trabajo con los datos irá permeando hacia otros actores del sector privado incorporando empresas medianas, luego pequeñas, ONG y otros.

Las empresas que no comiencen a evaluar formas de generar valor en base a sus datos no serán competitivas en 2025. Nuevamente por si no quedó claro: ¡las empresas que no comiencen a evaluar formas de generar valor en base a sus datos no serán competitivas en 2025!

Por todo lo mencionado, los profesionales deben comenzar a educarse en datos. La tecnología es accesible y es clave que los profesionales puedan definir estrategias basadas en datos, implementarlas, operar equipos que cuentan con analistas, y asegurarse de que sus unidades interactúan con los equipos de ciencia de datos que si no existen ya en sus organizaciones, definitivamente se harán presentes en próximos años.

Comenzar no es complicado, es cosa de revisar temas de estadística, jugar con Excel y luego pasar a entender un poco de R. Esto además de prestar atención a lo que está pasando en distintos verticales y mercados con un ojo estadístico puede ser un buen primer paso.

Referencias bibliográficas

El Bassiti, L., R. Ajhoun (2013). "Towards an Innovation Management Framework". *International Journal of Innovation, Management and Technology,* 4(6): 551-559.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-54. Retrieved from http: //dx.doi.org/10.1609/aimag.v17i3.1230

Glass, H., Livesay, A., Preston, D. (2015). "Data driven innovation in new zealand". Innovation Partnership.

Kline, Stephen J., Rosenberg, N. (1986). "An Overview of Innovation". *The Positive Sum Strategy: Harnessing Technology for Economic Growth*. National Academies Press.

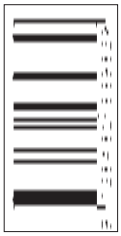
Morris, R. J. T., Truskowski, B. J. (2003). "The Evolution of Storage Systems". *IBM Systems Journal,* 42(2): 205-217.

Muenchen, B. (2016). "R Passes SAS in Scholarly Use (Finally)". Blog. *r4stats*, June 8.

Wickham, H., Grolemund, G. (2016). *R for Data Science*. O'Reilly.

Si existe un fenómeno tecnológico que ha inundado de manera masiva los medios generalistas (y también los especializados) en los últimos años, ese es el *big data*. Sin embargo, como ocurre con muchas tendencias tecnológicas, hay cierta confusión en su definición que genera incertidumbre y dudas cuando se trata de entender cómo esta tecnología puede ser usada desde el sector público para mejorar la forma en la que se toman decisiones o se prestan bienes y servicios a la ciudadanía.

A través de la lectura de los veinte capítulos de este manual, el lector podrá descubrir y descifrar los diferentes aspectos que son necesarios analizar antes de formular y desarrollar políticas o proyectos públicos en los que el uso de herramientas de ciencias de datos o vinculadas al *big data* sean la clave del éxito.



Con la colaboración de:

Telefónica