BID
Banco Interamericano de Desarrollo

El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe

Patricio Rodríguez
Norma Palomino
Javier Mondaca

Sector de Conocimiento y Aprendizaje (KNL)
Biblioteca Felipe Herrera (FHL)

DOCUMENTO PARA DISCUSIÓN Nº
IDB-DP-514

Mayo de 2017

El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe

Patricio Rodríguez
Norma Palomino
Javier Mondaca

Mayo de 2017

BID
Banco Interamericano de Desarrollo

Norma Palomino (npalomino@iadb.org), Banco Interamericano de Desarrollo, Departamento de Conocimiento y Aprendizaje
Patricio Rodriguez y Javier Mondaca, Centro de Investigación Avanzada en Educación, Universidad de Chile

Resumen

El presente trabajo comenta la definición de "datos masivos", en evolución permanente, y describe un panorama de las técnicas analíticas más utilizadas en el contexto de la formulación de políticas públicas en Latinoamérica y el Caribe. Asimismo, se presentan las conclusiones de tres estudios exploratorios realizados por equipos sectoriales del BID en las áreas de productividad a nivel de firma, movilidad urbana sostenible y ciudades inteligentes. A partir de los cuales se analizan aspectos sensibles del uso de “datos masivos” en el marco de políticas públicas, tales como seguridad y pertenencia de datos, privacidad, marco ético de uso, entre otros. Finalmente, se ofrecen recomendaciones para la adopción de la inteligencia de valor público por parte de las agencias de gobierno y una rúbrica de competencias de los consumidores inteligentes. La audiencia de este documento de discusión es, principalmente, tomadores de decisiones en distintas áreas de gobiernos en la región, profesionales del sector público y especialistas en desarrollo social y económico.

Palabras clave: políticas públicas, toma de decisiones, inteligencia de valor público, datos masivos, analítica de datos, gestión de datos, ciencia de datos.

Índice

1. Introducción

Actualmente, en la actividad económica moderna, los datos se constituyen como un factor esencial para la producción, tal como los activos fijos y el capital humano [1]. Con el advenimiento de las tecnologías de información, los datos han pasado de ser escasos a ser superabundantes [1]–[3]. Si bien, cada vez se dispone de mayor capacidad de almacenamiento y cómputo, se estima que actualmente la cantidad de datos que se generan sobrepasa la capacidad de almacenarlos físicamente: sólo entre 1986 y 2007 la capacidad de cómputo creció a una tasa anual del 58% [4]. Estos factores indican claramente que estamos en la era del *'Big Data'* o de los vastos conjuntos de datos.

El avenimiento de este fenómeno ha permitido el desarrollo una serie de tecnologías y conocimientos que dependen de estas capacidades [5]. Los posibles impactos positivos abarcan diversos sectores, desde el *retail* y la manufacturación de productos hasta la salud o la administración pública [1].

De hecho, en muchas áreas, el uso de datos es comparado con el avenimiento de la electricidad: ya no se puede trabajar sin ellos [6]. Es más, se ha llegado a plantear un escenario en que la producción de conocimiento es radicalmente distinta; los métodos científicos tradicionales, la teoría y la experticia profesional no serían necesarias, ya que los datos "hablarían por sí mismos"[7].

Muy por el contrario, la aplicación fallida del *'Big Data'* en la predicción de los niveles de infección de la influenza [8], [9] o la mejora educativa [10] (ver sección 4) han arrojado un manto de dudas sobre los logros del *'Big Data'*; pasando del desmedido optimismo a un también exagerado pesimismo. En consecuencia, es sumamente relevante comprender el fenómeno del *'Big Data'* para así poder evaluarlo en su justa medida, entendiendo su potencial y sobre todo sus limitaciones. En el documento se expondrá que, contrario a lo que se piensa, el uso de analítica avanzada requiere de decisiones expertas en todo momento; desde la selección de las fuentes de datos, las metodologías de análisis y —sobre todo— la interpretación y comunicación de los resultados. Esta serie de decisiones son las que finalmente determinan el éxito o fracaso de un proyecto de analítica de datos.

Finalmente, si bien sin las tecnologías de información y comunicación (TIC) no sería posible llevar a cabo cualquier esfuerzo de analítica que procese datos masivos, es sumamente importante no caer en el error de considerar un proyecto de este tipo simplemente como otro proyecto informático de implementación de infraestructura tecnológica, porque su naturaleza y potencial impacto son distintos: son estratégicos para sustentar la toma de decisiones basada en evidencia dentro de las organizaciones.

El objetivo del presente documento es revisitar el concepto de datos masivos y sus técnicas analíticas en el contexto de la formulación de políticas públicas, con enfoque en América Latina y el Caribe. Asimismo, se analizarán tres casos que aportarán a la comprensión de los desafíos de la puesta en marcha de proyectos de analítica en el sector público de la región, entregando recomendaciones y sugerencias para su implementación exitosa. Por consiguiente, el público objetivo del presente documento son principalmente los tomadores de decisiones, profesionales del sector público, autoridades de gobierno a nivel micro, meso y macro de la región.

El documento se estructura de la siguiente manera; en la sección dos se revisará brevemente el concepto de *'Big Data'*, dando cuenta de las metodologías que utiliza para gestionar los datos, analizarlos, y de las tecnologías empleadas para eso. Luego, en la sección tres, se discutirá, a través de casos exploratorios, cómo aplicar concretamente la analítica sobre datos masivos para generar evidencia que sirva a la toma de decisiones de política pública en América Latina y el Caribe.

Finalmente, en la sección cuatro, se entregan conclusiones y recomendaciones para la adopción de la inteligencia de valor público por parte de las agencias de gobierno, además de la presentación de una rúbrica para autoevaluación de competencias en los tomadores de decisiones.

2. Marco conceptual

En primer lugar, es necesario establecer un marco conceptual que permita delimitar el fenómeno del *'Big Data'*, debido a que este término tiene múltiples acepciones. En el imaginario colectivo se combinan conceptos relativos tanto a los datos propiamente tales, su manipulación, las técnicas y tecnologías para su análisis, y los profesionales y capacidades necesarias para realizar esta tarea.

2.1. ¿Qué se entiende hasta ahora por *'Big data'*?

El término *'Big Data'* proviene originalmente del ámbito de las ciencias de la computación y ha sido típicamente empleado para referirse a sets de datos cuyo tamaño excede al que puede manejar el software y hardware estándares disponible para capturar, almacenar y analizarlos [1], [5], [11]–[15]. En un principio, muchos autores tomaron las llamadas "**Tres Vs**" como características que definen **qué** es *'Big Data'*. Estas son [5], [16]–[20]:

(1) el **volumen**, esto es, la enorme cantidad de datos existentes. El volumen se asocia a los recursos requeridos por los datos, tanto de almacenamiento como de capacidad de cómputo. Si bien, inicialmente esta fue su característica más notoria, dándole su nombre, el continuo avance de las capacidades de hardware y software hacen que esta dimensión deje de ser determinante para caracterizar este fenómeno.

(2) la **velocidad** en que estos datos se producen y son analizados, en otras palabras, al ritmo en que estos se crean, procesan, analizan y almacenan [14], [19]. Sin ir más lejos, las plataformas y dispositivos de comunicación actuales permiten crear y/o compartir información con facilidad, lo cual genera grandes cantidades de información que requiere ser almacenada y procesada en tiempo real [18].

(3) la **variedad** de fuentes y tipos de datos. Estos últimos corresponden a la estructura de los datos, vale decir, estructurados, semiestructurados e inestructurados [5], [16].

Aunque la definición de datos masivos en términos de "las tres Vs" resultó ser instrumental para modelar problemas en el ámbito de las ciencias de la computación e informática, el hecho de que se base en características técnicas implica que tenga que ser revisitada continuamente [21]. En consecuencia, otros autores incorporan dimensiones más cualitativas respecto a los datos [16], [19], [22], tales como:

(4) **Variabilidad:** cuando el volumen de datos es reducido, debido a la dispersión (estadística) de los mismos, aparecen observaciones que presentan anomalías (usualmente llamados *'outliers'*) respecto de patrones prominentes [23]. Sin embargo, en los datos masivos, la cantidad de dichas "anomalías" es tan abundante que pierden dicha condición, volviéndose parte integrante del fenómeno a analizar. Un ejemplo de esto son los fenómenos virales en internet [24].

(5) **Complejidad**: se explica por la múltiple y variada cantidad de fuentes de datos existentes, causada por la proliferación de diferentes dispositivos conectados en línea. Ejemplos de estos son los dispositivos de seguimiento satelital (GPS), los sensores utilizados para crear el internet de las cosas (IOT por sus siglas en inglés), la generación espontánea de datos por ciudadanos, y otros fenómenos de la sociedad digital. Asimismo, se distinguen dos tipos de fuentes: inter-sujetos e intra-sujeto [25]. La primera se relaciona con la capacidad de recabar datos de muchos sujetos en un instante, mientras que el segundo tipo refiere la capacidad de recabar continuamente datos de un mismo sujeto (por ejemplo, datos biométricos de un sensor de

ejercicio). Además, los datos masivos también son exhaustivos respecto a su alcance, ya que un set de datos puede comprender todas las observaciones de una muestra determinada, lo cual nos permiten analizarla en su totalidad [18]. Más aún, muchos de estos datos están disponibles en tiempo real o al momento de ser generados, lo cual permite realizar predicciones casi inmediatas o '*nowcasting*' [26]. En términos de profundidad, los datos masivos poseen la cualidad de brindar máxima resolución, ya que alcanzan un nivel de detalle profundo, como en el caso de transacciones online que permiten registrar cada minúsculo detalle de cada operación [18]. Otra característica de los datos masivos es su flexibilidad tanto en términos de su extensibilidad, esto es, la capacidad de agregar nuevos tipos de datos fácilmente, como de su escalabilidad, es decir, su propiedad de expandirse en tamaño rápidamente [18].

(6) **Veracidad:** entendida como la calidad, confiabilidad y la certeza asociada a los datos, especialmente en relación a su origen y construcción. Por ejemplo, el análisis de datos masivos a partir de mensajes de redes sociales puede estar minado de información falsa o estar basada en percepciones subjetivas, que son imprecisas y engañosas [27]. En la misma línea, datos de transacciones online pueden sufrir cortes o pérdidas de segmentos de datos por problemas tecnológicos, problema que se exacerba cuando varios set de datos se usan en conjunto [28].

(7) **Representatividad**: cuestiona si los datos masivos representan adecuadamente las poblaciones analizadas, por la naturaleza propia de los datos o los medios establecidos para obtenerlos. Por ejemplo, en el caso de los datos producidos por medios sociales, estos presentan problemas de baja representación (por falta de participación o acceso a las mismas), sobre representación (cuentas y perfiles personas fallecidas), y de "multiplicidad" (múltiples apariciones del mismo individuo) [29].

2.2. ¿Cómo se procesa y analiza el *'Big Data'*?

Los datos masivos sin procesar tienen poco valor por sí mismos, ya que éste sólo se obtiene luego de pasar por un procesamiento completo [16]. Dicho valor se relaciona tanto con el retorno a la inversión, como con la posibilidad de construir conocimiento valioso, de mejorar procesos, y contribuir a la toma de decisiones disminuyendo la incerteza, entre otros.

Para el procesamiento y análisis del *'Big Data'* ha surgido una disciplina denominada **Ciencia de Datos** [15], [19]. La Ciencia de Datos (del inglés *'Data Science*') combina un conjunto amplio de técnicas provenientes de múltiples disciplinas tales como Ciencias de la Computación, Matemáticas, Estadística, Econometría e Investigación Operativa [30], [31]. El ciclo de vida del análisis de datos considera a lo menos seis pasos, que se detallan en la

Figura **1**.

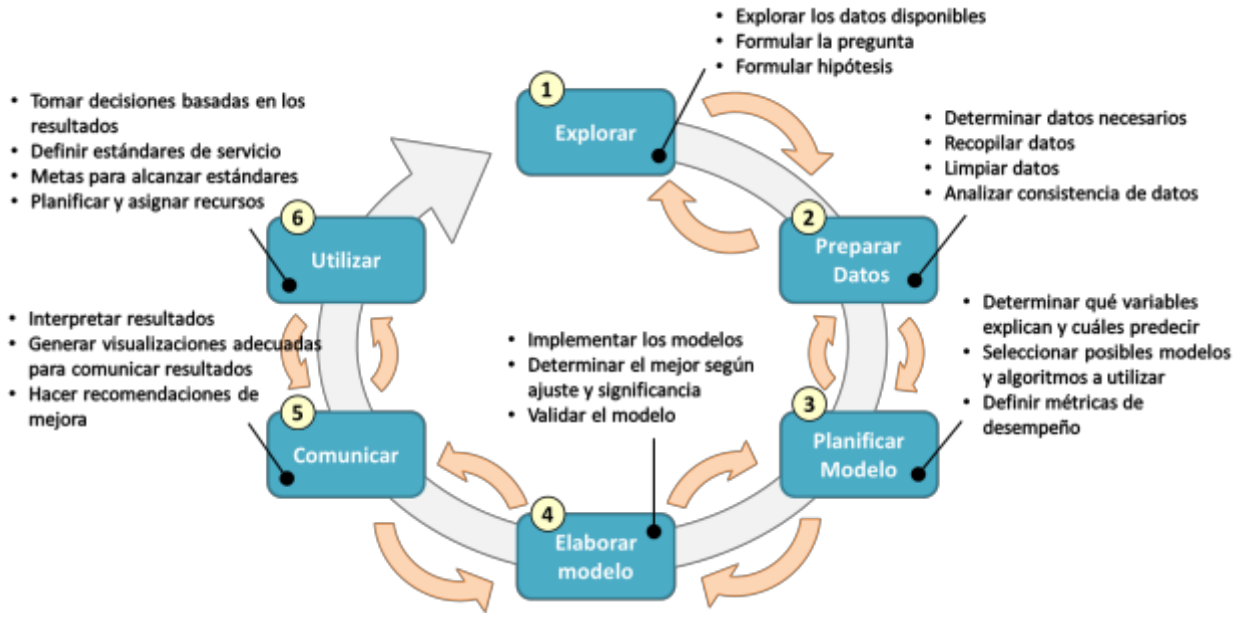

Figura 1: Ciclo de vida del análisis de datos [32].

Como se puede evidenciar en la Figura 1, el ciclo de vida del análisis de datos no es lineal; muchas veces se requiere reformular las preguntas en función de la disponibilidad de datos, o reinterpretar los resultados a la luz de nueva evidencia. Por ende, es un proceso iterativo en que se puede retroceder a etapas previas. De todas maneras, se puede describir el procesamiento de los datos masivos a partir de dos etapas principales: la gestión de los datos y la analítica de datos [16], [33].

La **gestión de los datos** se compone de tres aspectos: (1) adquisición y almacenamiento de los datos, (2) limpieza y depuración de los datos y, (3) la preparación para su análisis. La **analítica de datos**, por su lado, se refiere a la respuesta de preguntas y/o hipótesis formuladas a partir de técnicas de modelamiento y análisis. Como se puede constatar, este proceso no es particularmente distinto del proceso de indagación científica presente en cualquier disciplina; la principal diferencia subyace en las características generales de los datos que se utilizan —revisadas en la sección previa— y los desafíos que significan el acceso y manipulación de los datos.

Respecto a la **gestión de datos**, existe una amplia gama de metodologías que permiten cumplir los aspectos descritos en el párrafo anterior. Una posible clasificación se expone en la Tabla 1.

Tabla 1: **Metodologías asociadas a la gestión de datos para el procesamiento de *'Big Data'*. Construcción propia en base a Gandomi & Haider [16].**

Tipo de datos	Ejemplos de técnicas de procesamiento según el tipo de datos
Texto	Extracción de la información: obtener datos estructurados de un texto reconociendo entidades y relaciones entre ellas. Resumen de texto: construir resúmenes de uno o múltiples documentos usando procesamiento de lenguaje natural. Respuesta a la pregunta: responde a preguntas formuladas en lenguaje natural usando procesamiento desarrollado para tales fines. Análisis de sentimiento: analiza un texto de opinión y genera una respuesta negativa o positiva.

Audio	Enfoque basado en transcripción: se genera una transcripción textual de contenido del audio a través de reconocimiento automático del habla con grandes diccionarios. El resultado se analiza con las técnicas analíticas de texto. Enfoque basado en la fonética: trabaja a partir de sonidos o fonemas que se traducen a una secuencia a partir del habla. La representación fonética de un término se busca en la secuencia.
Video	Arquitectura basada en el servidor: servidor dedicado al análisis de videos. Arquitectura basada en el borde: el video es analizado en forma local y sobre el video sin compresión de datos.
Redes sociales	Analítica basada en contenido: se enfoca en los datos posteados por los usuarios, que son analizados posteriormente con técnicas descritas anteriormente, sea texto, audio o video. Analítica basada en la estructura: sintetiza los atributos estructurales de la red social y extrae inteligencia a partir de las relaciones de las entidades participantes. Las técnicas incluyen detección de comunidades, análisis de influencia social y predicción de enlaces.

Respecto a la **analítica de datos**, existen distintas metodologías, esto es, los métodos científicos que se utilizan para los análisis (Tabla 2) que se pueden tener diferentes implementaciones tecnológicas en forma de productos y servicios.

Tabla 2: Ejemplos de metodologías para modelar y analizar grandes volúmenes de información.

Metodología	Descripción	Aplicaciones / Ejemplos
Análisis espacial	Conjunto de técnicas que analizan las propiedades geométricas, topológicas y geográficas de un conjunto de datos [34]–[38].	• Regresiones espaciales (consumo vs. distancia a centros comerciales), simulaciones (desempeño de una cadena de distribución con bodegas en distintos lugares).
Análisis de redes	Conjunto de técnicas que caracterizan relaciones entre nodos discretos en un grafo o una red [39].	• Identificación de líderes de opinión para focalizar campañas de marketing. • Identificar cuellos de botella en flujos de información de una empresa. • Modelamiento de redes de transporte y predicción del tiempo de desplazamiento de un punto a otro.
Aprendizaje automático (*Machine Learning*)	Subespecialidad de la Ciencia de la Computación (denominada históricamente "Inteligencia Artificial") que se ocupa del diseño y desarrollo de algoritmos que permiten inferir comportamientos basados en datos empíricos. El aprendizaje automático puede ser de dos tipos: supervisado y sin supervisión. En el caso del aprendizaje supervisado se debe inferir una función a partir de un conjunto de ejemplos de entrenamiento. Estos consisten en un conjunto de entradas (en forma de vector) y un conjunto de salidas que son casos exitosos (satisfacen la función). Los casos exitosos permiten	• Predicción de fenómenos como crimen, deserción escolar y universitaria, esperanza de vida post-operatoria, ventas. • Sugerencias y recomendaciones de productos en función de historial pasado. • Procesamiento de lenguaje natural: reconocimiento de voz y lenguaje para interacción humano computador (ej: Siri, Cortana, Alexa), y análisis de

Metodología	Descripción	Aplicaciones / Ejemplos
	generar una medida de error respecto a las predicciones que se quieren hacer.[1] Ahora, en el caso del aprendizaje no supervisado estos "casos exitosos" no se conocen (o no se necesitan) y, por lo tanto, no existe retroalimentación para ajustar una función. El objetivo del algoritmo es organizar los datos o describir su estructura.	sentimientos en textos y redes sociales. • Reconocimiento de patrones: texto manuscrito, procesamiento de imágenes y reconocimiento de caras para buscar sospechosos de crímenes. • Detección de anomalías: detección de fraudes bancarios en base a actividad inusual en compras usando tarjeta de crédito.
Inteligencia Territorial	Metodologías de análisis espacial que, a través de tecnologías de información, combinan enfoques cualitativos, cuantitativos y espaciales, respetando además los enfoques de participación y aproximación global multidisciplinaria y multisectorial. Por ejemplo, análisis de agrupamiento y valores atípicos [40], [41] en conjunto con análisis multicriterio que combinan diferentes características espaciales [42].	• Indicadores espaciales de nivel de servicios públicos y privados. • Análisis de brechas y crecimiento espacial de la oferta y la demanda de servicios. • Accesibilidad urbana y rural, caracterización de los territorios en diferentes dimensiones geográficas, sociodemográficas, económicas y desarrollo humano.
Optimización	Técnicas numéricas de modelamiento para rediseñar y mejorar tanto procesos como sistemas complejos en múltiples dimensiones.	• Localización óptima de recursos en hospitales, escuelas, centros productivos, bodegas. • Producción: programación de maquinarias para la fabricación, manejo de inventarios.
Pruebas A/B	Una técnica en la que un grupo de control se compara con una variedad de grupos de prueba con el fin de determinar qué cambios mejorarán una variable objetivo dada. Por esto, esta técnica también se conoce como *Split testing* o *Buckettesting*. El gran volumen de datos permite ejecutar y analizar un gran número de pruebas, asegurando que los grupos son de tamaño suficiente para detectar diferencias estadísticamente significativas entre los grupos de control y tratamiento).	• Probar la efectividad de distintas campañas de marketing. • Probar la efectividad de un tratamiento médico, o un tipo de educación a través de experimentos naturales donde algunos sujetos participan de una intervención y otros no, por diferentes circunstancias. Se busca además que estos sujetos sean lo más parecido en sus características (pareo) para controlar la mayor cantidad de variables posible.

1 Como se puede notar, muchas de las técnicas de análisis estadísticos más tradicionales, como la regresión lineal múltiple, pueden entenderse como modelos de aprendizaje automático donde las medidas de error entregan retroalimentación sobre el ajuste del modelo. De hecho, hay modelos como el multinivel, que van dando cuenta de la anidación de los datos y requieren una escala mínima de estos para funcionar.

Metodología	Descripción	Aplicaciones / Ejemplos
Simulación	Modelación del comportamiento de sistemas complejo para pronóstico, predicción y planificación de escenarios.	• Pronóstico de los resultados financieros de una empresa dado circunstancias de incertidumbre. • Pronóstico del clima.
Visualización analítica de datos	Forma de descubrir y entender patrones en grandes conjuntos de datos vía interpretación visual, para que así los usuarios pueden navegar y explorar los datos.	• Análisis visual interactivo de componentes principales [43], [44].
Visualización de datos	Comunicación de información en forma clara y efectiva a través de distintas formas de representación gráfica interactiva [45]–[48].	• Infografías. • Tableros de mando (*Dashboards*), para seguimiento y síntesis de ciertos fenómenos.

Finalmente hay productos y servicios tecnológicos (y muchas veces, software) que permiten gestionar y analizar los datos, algunos de los cuales se ejemplifican en la Tabla 3.

Tabla 3: Servicios tecnológicos utilizados para manipular y analizar grandes volúmenes de información.

Servicio	Descripción	Ejemplos
Nuevos frameworks analíticos	Son entornos de trabajo que contienen o pueden contener una serie de paquetes y librerías que permiten reutilizar código, facilitando tareas comunes.	Hadoop (Google, Apache), Spark.
Almacén de datos y Lagos de datos (*Data warehouse & Data lakes*)	Son repositorios de datos a los que sólo se insertan nuevos datos y, a través de los llamados *data marts*, permiten la generación de bases de datos. La diferencia entre *data warehouse* y *data lake* radica en la estructura de los datos. Los *data warehouses* almacenan datos estructurados (vale decir, tablas con datos ordenados en filas y columnas), mientras que los *data lakes* pueden almacenar datos estructurados, semi-estructurados e inestructurados	SQL Server, Azure SQL, NoSQL.
Base de datos relacional (*relational database*)	Base de datos compuesta de una colección de tablas (relaciones). Los sistemas de gestión de bases de datos relacionales (RDBMS) almacenan un tipo de datos estructurados. SQL es el lenguaje más utilizado para la gestión de bases de datos relacionales (véase el tema más adelante).	MySQL, PostegreSQL, Oracle, SparkSQL.
Base de datos no relacional (*non-relational database*)	Base de datos que no almacena los datos en formato tabular (filas y columnas).	MongoDB, Cassandra
Visualización de datos	Herramientas que permiten la visualización de datos. Se diferencian comúnmente en la flexibilidad y versatilidad que ofrecen para crear visualizaciones personalizadas.	D3js, Google Charts, Tableau, Vega
Herramientas/*Plugins* estadísticos	Refiere a paquetes estadísticos o extensiones, que permitan realizar análisis estadísticos sobre los datos. Los más sencillos usualmente permiten realizar todas las tareas a través de una interfaz gráfica, mientras que los más complejos requieren conocimientos de algún lenguaje de programación.	SAS, Stata, SPSS, Matlab, R, Python Pandas

Sistema de información geográfica (***Geographic information system***)	Son sistemas diseñados para manipular, almacenar, analizar y visualizar datos geográficos.	Leaflet, PostGIS, Esri ArcGIS, CartoDB
Servicios en la nube (***Cloud Computing***)	Es un modelo que permite acceso *on-demand* a una serie de recursos computacionales configurables. Estos recursos pueden ser utilizados o liberados sin necesidad de interacción con el proveedor y con pocos recursos a nivel de gestión [1]. Cinco se reconocen como las características esenciales del Cloud Computing: auto servicio *on-demand*, mayor acceso a red (*broad network access*), conjunto de recursos (*resource pooling*), rápida elasticidad y capacidad para controlar y optimizas el uso de recursos de forma automática.	Amazon Web Services, Google Cloud Platform, Microsoft Azure, Digital Ocean: Cloud Services for Developers.

Este ciclo de análisis que supone la Ciencia de Datos requiere de profesionales especialistas con una formación sólida en alguna de las Ciencias de la Computación, uso y desarrollo de aplicaciones, modelamiento, estadística, analítica y matemáticas. A estos profesionales se les denomina **Científicos de Datos** (del inglés *Data Scientists*), quienes exploran, generan preguntas, realizan análisis de escenarios ("¿qué pasa si?") y cuestionan los supuestos y procesos existentes utilizando múltiples fuentes de datos de diferentes orígenes [19], [49].

3. Uso de analítica avanzada para la toma de decisiones, el diseño, implementación y evaluación de políticas públicas

En el caso de la toma de decisiones, el diseño, implementación y evaluación de políticas públicas el objetivo del uso de la Ciencia de Datos es producir evidencia que sea pertinente, de calidad y oportuna, para así fundamentar y orientar decisiones. Esto significa diagnosticar problemas que pasan inadvertidos o desapercibidos y, por lo tanto, son imposibles de accionar [43]. Este proceso se denomina "toma de decisiones guiadas por datos" (del inglés *'Data-driven decision making'*) [15].

Existe fuerte evidencia de que las aplicaciones de *'Big Data'* pueden desempeñar un rol importante para beneficiar no sólo a las empresas privadas, sino a las economías de cada país y sus ciudadanos [1]. Estos beneficios se producen a través de la creación de valor para la economía mundial, mejorando la productividad y competitividad privada y públicas, y creando excedentes económicos para los consumidores [1].

En las empresas privadas, el análisis de datos para tomar decisiones ha sido utilizado desde hace bastante tiempo. Estas empresas realizan cálculos complejos sobre los datos de sus clientes, usando técnicas de análisis provenientes del *'Big Data'* denominadas "Inteligencia de negocios" para —por ejemplo—descubrir patrones y tendencias que predigan el comportamiento futuro de los consumidores, evaluar el impacto de la segmentación de una campaña de *marketing*, o recomendar productos y servicios en base al historial pasado de compras [1], [50].

Por su parte, el análisis de *'Big Data'* también puede ser utilizado para la mejora de la administración pública, a través de la generación de más y mejores soluciones que satisfagan necesidades de salud, educación, transporte, vivienda, atención e inclusión de grupos desaventajados, entre otras, a partir de contextos sociales, demográficos y territoriales particulares. Este aspecto es especialmente relevante en el sector público, en el cual prevalece una cultura de tratar a todos los ciudadanos de la misma forma, independientes de sus características personales y grupales [1].

De esta manera, con el acceso a datos masivos y el uso de técnicas analíticas adecuadas, se pueden identificar y medir problemáticas que habían permanecido invisibles, lo que, obviamente, las hacía imposibles de gestionar. De esta forma, es posible desarrollar una "**inteligencia de valor público**" (un equivalente social de la "inteligencia de negocios"), que tiene la potencialidad de ser un componente estratégico para la toma de decisiones y el diseño, implementación y evaluación de políticas públicas dentro de los gobiernos de América Latina y el Caribe.

En concreto, en el contexto del primer *'Big Data Innovation Challenge'* del Banco Mundial [51], se premió una serie de iniciativas al servicio de la generación de valor público con *'Big Data'* en distintos ámbitos; entre los que se encuentran:

- **Pobreza**: en India se utilizaron imágenes satelitales nocturnas para analizar la cobertura eléctrica en las más de 600.000 aldeas del país para así dar cuenta de sus necesidades. Imágenes similares se utilizaron en Sri Lanka y Pakistán para evaluar otras variables (autos, áreas construidas, sombras, tipos de techo, tipos de calles, entre otras), generando indicadores más baratos y con una precisión similar (y a veces mejor) que los tradicionalmente utilizados. Por su parte, en Nigeria, se evaluó la relación entre pobreza y baja eficacia de mercado analizando y combinando los precios mensuales de cientos de *'commodities'* con imágenes satelitales nocturnas (que dan cuenta de sectores con electricidad).
- **Crimen y seguridad**: En Bogotá se estudió la asociación entre crimen e infraestructura urbana. Para esto se utiliza la información de las rutas del *'Bus rapid transit'* (BRT), un sistema de buses, en conjunto con el modelamiento de terreno riesgoso *('risk terra in modeling'* en inglés) para analizar los datos. De esta manera, se asociaron ciertas zonas cerca de hospitales, colegios, farmacias y estaciones de buses a asaltos y asesinatos, así como también se identificaron horas de mayor criminalidad, y predijeron las áreas de la ciudad que sufrirán crímenes en el futuro.
- **Transporte**: en Filipinas se desarrollaron las aplicaciones *OpenRoad* y *Open Traffic*. La primera es un portal interactivo que permite a los usuarios hacer seguimiento a proyectos viales con financiamiento público, y entregar *feedback* por proyecto o localidad. La segunda, permite visualizar y analizar información de la velocidad del tráfico, utilizando información de los GPS instalados en taxis, así como datos recabados por los teléfonos celulares de los taxistas. Otra iniciativa, con apoyo del gobierno de Bielorrusia, desarrolló una aplicación llamada *RoadLab*, que permite evaluar la calidad de la superficie de las calles y caminos utilizando el acelerómetro de los teléfonos celulares y entregando su posición a través del GPS. Para esto, se dividieron las calles en segmentos de 100 metros, georreferenciando el comienzo y el final. Los datos recabados se utilizan para estimar un valor del '*International Roughness Index*'.
- **Salud**: en Sudáfrica se generaron algoritmos que permiten unificar las bases de datos de distintas instituciones públicas que manejan información respecto a pacientes con SIDA, centros de salud que ofrecen asistencia en esta materia, cantidad de pruebas de laboratorio realizadas para detectar esta enfermedad, entre otra información de la materia. De este modo, se puede identificar a nivel nacional, provincial, distrital o de dependencia de salud, los lugares en que se atienden las mayores proporciones de pacientes con SIDA [52].

Otras experiencias relevantes, pero que no participaron del *'Big Data Innovation Challenge'* del Banco Mundial, son dos casos en salud pública en los estados de Chicago e Indiana, en EEUU [53]. Chicago mejoró la desratización de la ciudad a partir del trabajo en conjunto con el organismo encargado de desratizar y el uso de datos abiertos. Por su lado, Indiana, con el objetivo de disminuir la mortalidad infantil, creó un centro de datos de alto nivel, a través de la inversión en personal y tecnología de punta, así como en la seguridad de la información.

A continuación, se revisarán en profundidad tres casos de implementación de analítica basada en *'Big Data'* en América Latina y el Caribe. Estos casos sirven para ilustrar concretamente el uso e impacto de la analítica avanzada en estas regiones, sistematizando los principales desafíos, oportunidades y aprendizajes de estos proyectos. El primer caso se relaciona con aportar información para la política pública desde la ejecución de un proyecto de intervención específica en la región. El segundo también pretende colaborar con la política pública, pero desde la evaluación de experiencias realizadas por terceros. El último caso presentado se distancia de los dos anteriores, presentando un ejercicio más académico, pero que entrega valiosa información desde esta perspectiva y desde la metodológica.

Para facilitar la comprensión de los casos, cada uno se presenta separadamente, estructurando su revisión y análisis de la siguiente manera:

(1) Se introduce brevemente el caso, identificando las necesidades/problema a abordar, así como las posibles preguntas o hipótesis que puedan haberse desarrollado.
(2) Se examinan los datos que fue necesario recolectar o que estaban disponibles para contestar las preguntas o evaluar las hipótesis planteadas.
(3) Se revisan las metodologías utilizadas para obtener y procesar los datos, así como para generar los resultados obtenidos.
(4) Se sintetizan las conclusiones relevantes para la política pública o el área específica del caso.

3.1. Movilidad Urbana Sostenible, Datos Masivos y Políticas Públicas: estudio de la movilidad de los ciclistas en la Ciudad de Rosario, Argentina

Descripción del caso, necesidades y/o problemáticas detectadas

El primer caso a revisar se realizó en Rosario, Argentina. En esta ciudad se llevó a cabo un estudio que dio cuenta de la movilidad de los ciclistas de la ciudad, utilizando dispositivos de georreferenciación. Los objetivos de la investigación fueron entender 1) los patrones de movilidad de los ciclistas en relación con la infraestructura vial existente (sean ciclovías o no), 2) la relación de estos patrones y los accidentes de tránsito, y 3) las posibles mejoras de la infraestructura vial [54].

El estudio tuvo dos fases. En la primera, se entregaron e instalaron dispositivos de localización GPS a 40 usuarios voluntarios privados (que tienen sus propias bicicletas). En la segunda, se pusieron a disposición 150 bicicletas en un sistema público llamado "Mi bici tu bici". Al momento del estudio se contaba con 173 bicicletas activas en el sistema público, por lo que el 85% de la flota estuvo cubierta por los GPS. El motivo de la inclusión de usuarios privados fue su patrón de movilidad, ya que no están geográficamente limitados por la ubicación de los terminales de bicicletas públicas; de esta manera, se pueden evaluar otro tipo de recorridos.

Cada dispositivo instalado, tanto en las bicicletas públicas como las privadas, contaba con un acelerómetro y un GPS, lo que permitió recabar datos respecto a la velocidad y el recorrido de los viajes. Por su parte, para el análisis de los accidentes, se recopilaron datos de distintas agencias públicas. En el caso de los siniestros, lesionados y muertes, desde el Observatorio Vial de la Agencia Provincial de Seguridad Vial de Santa Fe; para la relación entre el uso de las ciclovías y los choques en la ciudad, los datos se obtuvieron desde el Ente de la Movilidad de Rosario por último, se realizaron entrevistas focales a ciclistas privados para validar la información cuantitativa obtenida.

Metodologías para analizar los datos

Los datos obtenidos desde fuentes privadas fueron recolectados en un periodo de dos semanas. En el caso de los datos correspondientes al sistema público, estos fueron recolectados en un periodo de

seis semanas. En ambos casos, sólo se recabaron datos correspondientes a días hábiles (esto es, de lunes a viernes). A partir de estos, se sistematizó la cantidad de viajes, los tiempos invertidos en los mismos, las distancias o la velocidad promedio, y los ejes viales más utilizados. Esta información pudo ser agregada a nivel de mes, día, horario o individuo y diferenciada según rango etario y género.

Respecto al procesamiento de los datos, en un primer momento, estos se exploraron visualmente a través de la creación de mapas que mostraban los principales corredores utilizados por los ciclistas, la velocidad del tránsito de bicicletas, los focos de accidentes y la gravedad de dichos siniestros. Esto permitió identificar vías, cruces o zonas específicas que, por la frecuencia y gravedad de los accidentes, requerían mayor atención. En un segundo momento, se utilizaron entrevistas para profundizar en estos focos, así como en problemas generales.

Conclusiones para la política pública

El caso revisado evidencia que, efectivamente, es posible utilizar análisis de datos masivos para generar diagnósticos y acciones que mejoren el bienestar ciudadano. Los resultados del estudio detectaron que, en términos generales, los ciclistas prefieren las calles que cuentan con ciclovías, por ser más rápidas y seguras. De la misma forma, los datos permitieron reconocer focos de accidentes en las calles sin ciclovías, pero con altos volúmenes de ciclistas. Adicionalmente, se identificaron varios ejes viales sin ciclovías que son altamente utilizados por los ciclistas en sus desplazamientos hacia y desde entidades de gobierno, servicios públicos, universidades y centros educativos.

A partir de esta información, se pueden tomar decisiones informadas para mejorar los servicios entregados a la ciudadanía. En este caso específico, el estudio da cuenta de que la calle Bulevar Oroño, que cumple una función de eje articulador de la ciudad (concentra distintos servicios públicos, educacionales, de salud y culturales), es ampliamente utilizada por ciclistas, pese a no contar ni con infraestructura *ad hoc* ni estar autorizada la circulación en ese medio de trasporte. Al conocer la causa de este movimiento, se plantean diferentes opciones de modificación arquitectónica de dicho bulevar que consideren la convivencia con el peatón, además de mantener el atractivo del barrio.

En términos de la adopción de la inteligencia de valor público, se observan un desafío en particular. El procesamiento y análisis de los datos fue realizado externamente, lo que indica que aún no existen ni la infraestructura y/o el capital humano en las agencias gubernamentales requeridos para afrontar este tipo de proyectos.

Finalmente, las mismas técnicas de recolección de datos y análisis podrían utilizarse para la evaluación de impacto del proyecto. Por ejemplo, observar el cambio en el uso de la infraestructura vial (antigua y nueva) para determinar si se cumplieron los objetivos de mejoramiento de la política pública.

3.2. Computando una nueva trayectoria para la gobernanza: innovaciones en datos masivos en América Latina y el Caribe

Descripción del caso, necesidades y/o problemáticas detectadas

La segunda experiencia es un análisis de cuatro ciudades inteligentes (‘*smart cities’* en inglés) en Latinoamérica. Al hablar de ciudades inteligentes, se hace referencia a tres aspectos relativos al gobierno de una ciudad: aumentar la transparencia, mejorar la eficiencia y conseguir innovación continua [55].

Los casos analizados se enmarcan en un estudio del BID [55] cuyo objetivo fue evaluar la capacidad de las ciudades para: apoyar iniciativas innovadoras de análisis de datos masivos, entender las particularidades latinoamericanas en dichas ciudades para mejorar la gobernanza, e identificar los

desafíos específicos, soluciones e innovaciones que surgen desde esta zona. A continuación, se describe brevemente cada caso:

- **Bahía Blanca (Argentina)**: la iniciativa '¿Qué pasa Bahía Blanca?' (QPBB) surge de la tensión entre activistas medioambientales y la industria petroquímica de la zona. La respuesta gubernamental fue instalar sensores de distintas variables ambientales en sectores estratégicos y compartir la información producida por estos a través una plataforma y aplicación móvil. A partir de estas, era posible realizar un seguimiento en tiempo real de la contaminación del aire y acústica producida por dichas plantas. La información producida quedó disponible en la plataforma de forma abierta.
- **Córdoba (Argentina)**: se desarrolló un sistema de seguimiento de la flota del transporte público, particularmente en el centro de la ciudad. Aunque uno de los operadores ya entregaba su propia información respecto a tiempos de viaje, recorridos e ingresos, se estimó necesario hacer un seguimiento desde el sector público por dos motivos. Primero, existía preocupación respecto a la exactitud de la información entregada, sobre todo en lo relativo a ingresos de estas empresas; además de la dificultad para integrar los datos provenientes del operador. Segundo, existía una pretensión de parte de la ciudad de "recuperar" su legado pionero en relación al transporte público, que en los últimos años quedó bajo la operación privada.
- **São Bernardo do Campo (Brasil)**: En respuesta a los problemas que plantea el crecimiento de la ciudad, el gobierno federal realizó un esfuerzo progresivo por mejorar la infraestructura y la logística detrás de diversos servicios públicos. Para esto creó *Você SBC*, una aplicación móvil que permite recolectar quejas de los ciudadanos y sugerencias relacionadas con una amplia gama de servicios no-urgentes (por ejemplo, problemas de basura, ruidos, hoyos en las calles, entre muchos otros). Esta aplicación permite conocer y hacer seguimiento de las necesidades de la ciudad y sus habitantes.
- **Fortaleza (Brasil)**: el proyecto 'Fortaleza Inteligente' tiene su génesis principalmente en dos esfuerzos para mejorar la gobernanza de la ciudad, que comenzaron el 2013. El primero fue la creación de CITINOVA, una fundación pública cuya misión es promover el uso de la ciencia, tecnología e innovación en el gobierno para mejorar los servicios entregados a los ciudadanos. El segundo esfuerzo fue el establecimiento de "Plan de acciones inmediatas de transporte y tránsito" (PAITT en portugués), un plan maestro de iniciativas para mejorar el transporte público y tráfico de la ciudad. De esta manera, 'Fortaleza Inteligente' genera tres proyectos pilotos. El primero utiliza información del sistema GPS en los buses del transporte público, para evitar retrasos y sobrepaso del límite de su capacidad. El segundo, utiliza datos del sistema de bicicletas públicas de la ciudad para analizar su uso dentro la ciudad y así generar evidencia para su expansión. Finalmente, se creó un *dashboard* que unifica indicadores del sistema de transporte en su conjunto, entregando visualizaciones a través de la web.

Datos necesarios o disponibles

La variedad de las iniciativas hizo necesario una amplia diversidad de datos requeridos. Por una parte, el proyecto en **Bahía Blanca** (Argentina) requirió la instalación de diversos sensores que miden continuamente: la calidad del aire[2], efluentes líquidos y la contaminación acústica de origen industrial en las vecindades de las plantas petroquímicas; así como también la geolocalización de las plantas. Para complementar esta información se dispuso de cámaras que monitorean en vivo algunas plantas.

[2] Presencia de material particulado, dióxido de azufre, ozono, óxidos de nitrógeno y monóxido de carbono, así como la dirección y velocidad del viento y la presión atmosférica. Fuente: http://www.quepasabahiablanca.gov.ar.

En el caso de **Córdoba** (Argentina), se debieron instalar dispositivos GPS en todos los buses del sistema público[3]. Además, se recolectó información del pago a través del terminal utilizado por 'tarjeta única'-sistema de pago unificado para el sistema de transporte público de la ciudad-, tal como datos sobre la hora, tarifa y localización (a través del GPS) de cada transacción realizada a través de la tarjeta.

En el caso de **São Bernardo do Campo (SBC)**, la plataforma 'Você SBC' recopila datos enviados desde los dispositivos de los propios ciudadanos de la ciudad, que deben estar registrados previamente. Permite subir solicitudes de diversos tipos de carácter no urgente. Asimismo, la aplicación permite también georreferenciar estas solicitudes, pudiendo identificar zonas donde se está acumulando basura o donde se cayó un árbol que debe ser retirado, como también locaciones de interés público, como podrían ser ferias libres.

Finalmente, en el caso de **Fortaleza,** dado que consta de tres experiencias distintas, se recolectó mayor variedad de datos. Al igual que en Córdoba, se instalaron dispositivos GPS en el transporte público para medir la velocidad y posición de los cerca de 2.000 buses de la flota. También, se cuenta con una tarjeta de pago única para todo el servicio (incluyendo las bicicletas públicas), llamado el '*bilhete único'*. Si bien se recabó información similar entre Córdoba y Fortaleza, en esta última también se recopilaron datos de sensores dispuestos en intersecciones, radares (para medir los niveles de congestión vehicular); además de fuentes complementarias, como estadísticas de crimen y accidentes del transporte público.

Desde la perspectiva de esta evaluación se recopilaron datos englobados en tres grupos: **1)** aspectos específicos de cada proyecto (orígenes, etapas, tomadores de decisiones clave, proceso de diseño, modelo de negocios, entre otros); **2)** información relativa al estado de madurez del proyecto (que se abordará en mayor detalle en el análisis) y **3)** variables contextuales fuera del control de los agentes involucrados (como aspectos políticos, tecnológicos y sociales a nivel macro).

Metodologías usadas para analizar los datos

Para el análisis de las experiencias señaladas, se generaron reportes que dan cuenta del diseño y evolución cronológica de los casos; así como también de sus obstáculos, resultados e impactos. La principal herramienta de análisis utilizada fue el 'modelo de maduración de datos masivos urbanos' propuesto por el BID [55]; ésta consiste en una rúbrica de cinco dimensiones, con cinco niveles (ver Anexo A). Esta rúbrica permite estimar en qué nivel de desarrollo (o madurez) se encuentra la iniciativa de datos masivos urbanos evaluada. A continuación, se presentan las dimensiones de la rúbrica, que serán utilizadas al momento de describir y analizar el caso.

1. **Datos abiertos**: guarda relación con proveer los datos necesarios, así como con generar la demanda por gobernanza basada en datos.
2. **Cultivar ecosistemas de datos**: se relaciona con la creación de comunidades, mecanismos para compartir datos y una cultura de uso de datos, sobre todo a nivel de tomadores de decisiones y partes interesadas.
3. **Analítica**: refiere a las técnicas utilizadas para analizar, resumir y visualizar la información.
4. **Toma de decisiones basada en datos**: son las habilidades individuales, prácticas institucionales y culturales necesarias para poder utilizar los datos para mejorar la política pública.

[3] Incluyendo los que ya contaban con otro GPS instalado por el operador respectivo.

5. **Participación y servicios públicos**: guarda relación con la utilización de los datos para la innovación de la relación entre los gobiernos y ciudadanos.

Cada una de estas cinco dimensiones tiene una rúbrica de adopción que va desde resolver problemas específicos del momento, hasta un nivel en que la apropiación está completa y la mejora continua se releva como importante. El detalle de cada nivel de adopción se describe en la tabla del Anexo A. Los cuatro casos analizados fueron evaluados a partir de estas dimensiones; el nivel de madurez final para los distintos casos se observa en la Tabla 4.

Tabla 4: **Nivel de madurez inicial y final alcanzada por los proyectos según el modelo de adopción de *Big Data* en problemas urbanos en Zambrano [55].**

Casos	Periodo de implementación	Nivel inicial — final en cada dimensión en la rúbrica				
		Datos abiertos	Cultivar ecosistemas de datos	Analítica	Toma de decisiones basada en datos	Participación y servicios públicos
Bahía Blanca	2012—2016	1,0 —2,0	1,0 —2,0	1,0 —1,0	1,0 —1,0	1,0 —2,0
Córdoba	2012—2016	2,5—2,5	1,0 —2,0	1,5 —2,5	1,5 —2,5	2,0—2,0
São Bernardo do Campo	2014—2016	2,0—2,0	2,0—2,0	2,0—2,0	2,0—3,5	2,0—3,0
Fortaleza	2013—2015	2,5 —2,5	2,0 —3,0	2,0—3,0	2,0—3,0	2,0—2,0

Conclusiones para las políticas públicas

A partir de las experiencias evaluadas y las herramientas de análisis propuestas, existen varias consideraciones al momento de utilizar datos masivos al servicio de la política pública. Los objetivos y necesidades diversas, como también distintos niveles de madurez, dan cuenta de distintos desafíos y tensiones que proveen lecciones para el futuro.

Si bien los procedimientos técnicos y de análisis son relevantes, se hará énfasis en las características, capacidades y potencialidades de las instituciones analizadas dado que son requisitos para el éxito de implementación de iniciativas de inteligencia de valor público dentro de los gobiernos. En términos generales, se reconocen las siguientes dimensiones para dicho éxito:

(1) La **construcción de una institucionalidad** que permita generar, administrar y dar continuidad a los recursos para la infraestructura y el personal dedicado a la inteligencia de valor público. Consecuentemente, se establece un terreno para fijar y discutir los objetivos y alcances de dichas innovaciones como ocurrió, por ejemplo, en el caso de Fortaleza. Asimismo, la institucionalidad juega un rol relevante al momento de establecer la propiedad o el acceso a los datos; las instituciones deben generar tempranamente lineamientos claros, sobre todo cuando colaboran con terceros no gubernamentales.

(2) Lograr una **comunicación transparente y fluida con otras entidades externas** (públicos o privadas) y con la ciudadanía. En etapas tempranas de un proyecto de inteligencia de valor público, la capacidad de adaptar la información a las necesidades de los ciudadanos —y no de las entidades que los procesan— es fundamental para cimentar lazos con la ciudadanía. Un ejemplo de esto son los casos de Bahía Blanca (entregando información a otras entidades) y de São Bernardo do Campo, en que la aplicación *Você SBC*—pensada para tener un vínculo con la ciudadanía—fue desarrollada en una hackatón. De esta manera, se promueve una doble participación ciudadana: agentes de innovación y productores de datos. La interacción con otros organismos públicos también es relevante, ya que el análisis de datos tiene uso potencial en otras agencias y se requiere una perspectiva multilateral para maximizar el impacto. Esta interacción se facilita con la implementación de políticas e infraestructura de datos abiertos, o

a partir de la creación de un ecosistema de datos para que diversas agencias públicas puedan compartir sus datos a otros agentes claves del sistema. Finalmente, la academia puede aportar con sus propios saberes y experiencia, como en el caso de Córdoba, en que una institución de educación superior y una institución pública realizaron el trabajo de la validación datos en conjunto, promoviendo la innovación tecnológica en el sector privado para la consecución de objetivos públicos.

(3) La **disponibilidad del capital humano** necesario, ya sean profesionales que trabajen directamente en los análisis, tomadores de decisiones o actores claves que definan preguntas y objetivos. Por su parte, los científicos de datos tienen los conocimientos y competencias necesarios para mejorar la infraestructura y el ecosistema de los datos abiertos, así como también son capaces de proponer la analítica pertinente para su análisis. Respecto a los tomadores de decisiones y agentes claves, estos deben ser capaces de guiar el trabajo de los profesionales de los datos, para así orientarlo adecuadamente a las necesidades ciudadanas. Por cierto, el capital humano necesario puede encontrarse fuera del sector público; por ejemplo, en Córdoba, se recurrió a una empresa privada para recopilar, limpiar y analizar sus datos[4]. De cualquier modo, para la sustentabilidad en el largo plazo, se requiere desarrollar estas capacidades al interior del sector público.

Los aspectos anteriormente descritos están interrelacionados, y si bien apelan al modelo de madurez, presentan similitudes con algunas capacidades institucionales definidas para la adopción de datos masivos en el sector público [30], [43], [56]. Estas capacidades se discutirán desde una perspectiva más general en la sección 4.

De esta manera, el modelo de madurez no solo permite adquirir lecciones para la política pública, sino que ofrece posibilidades no exploradas que pueden ser provechosas. Utilizando este modelo es posible realizar evaluaciones *ex ante*, *a posteriori,* e incluso auto-evaluativas las que permitirían identificar los desafíos y fortalezas de una organización embarcada en un proyecto de inteligencia de valor público [55]. Sin ir más lejos, en los proyectos se realizan dos evaluaciones, en distintos momentos temporales, con el fin de dar cuenta del avance madurativo del proyecto.

3.3. Utilizando datos a nivel de empresa para estudiar el crecimiento y dispersión en el factor de productividad total

Descripción del caso, necesidades y/o problemáticas detectadas

La tercera experiencia es un estudio que, utilizando datos a nivel de empresa, estima el crecimiento y dispersión en la productividad total de los factores (*total factor productivity* en inglés, o TFP). La TFP es la proporción de la producción que no está explicada por las cantidades de insumos que necesita para ser producida; y su nivel estará determinado por cuán eficiente e intensivamente se usan los insumos en la producción [57].

En la literatura, los esfuerzos por estimar la TFP de más de un país en forma simultánea son escasos. En cambio, el estudio analizado calculó la TFP para cerca de 20 millones de empresas, en alrededor de 30 países durante ocho años, entregando un panorama más general respecto al estado y evolución de dicho indicador. Dos antecedentes son relevantes para contextualizar este estudio: (1) antes de la crisis financiera del 2008 hubo un crecimiento sostenido de los TFP y, desde el 2011, esta tendencia se invirtió declinando de forma continua. (2) La dispersión entre países ha aumentado desde el 2010. Se considera que esta desaceleración y aumento de la dispersión deberían reflejarse en el comportamiento de las empresas.

[4]El diseño, implementación, así como la recopilación y análisis de los datos fue dejado en manos de la empresa ATOS SIEMENS.

Datos necesarios o disponibles

Para dar cuenta del amplio escenario económico, se requirió recolectar una cantidad importante de datos de empresas. Estos datos fueron obtenidos desde la base de datos Orbis [58] que dispone de toda la información necesaria para computar las TFP, incluyendo balances financieros y estimaciones de las medidas de productividad.[5]

Metodologías para analizar los datos

En términos metodológicos, el proyecto contó con dos grandes fases: la preparación de los datos y el cálculo de funciones de productividad. El gran desafío fue la limpieza y preparación de los datos, la que requirió de una cantidad considerable de tiempo, además del capital humano adecuado para realizarla. Para la limpieza de los datos fue necesaria la imputación de datos faltantes (posibles de calcular a partir de la información disponible), así como la eliminación de empresas que no reportaron valores clave. Por su parte, la preparación de los datos implicó el cálculo de nuevas variables que describen a los trabajadores, materiales y máquinas de las empresas; a partir de supuestos respecto a cómo se utilizan los insumos en distintos rubros de negocios.

Posteriormente, se calcularon las funciones de producción, a partir modelos de regresión que determinan cuánto se produjo a partir de los insumos disponibles. Para esto se emplearon cuatro metodologías distintas, todas basadas en mínimos cuadrados ordinarios, para salvaguardar la robustez de los resultados [59]. En términos computacionales, esta fase fue la parte más intensiva del proceso, tanto por la cantidad de datos, como por la variedad de metodologías utilizadas en los cálculos. Luego se calcularon las elasticidades de los factores de producción, que son el peso de cada factor de producción en su industria, país respectivo, e incluso a nivel de firma.

Conclusiones para la política pública

En términos de resultados, se observa que la dispersión identificada no tiene una relación clara con la combinación entre TFP promedio (considerando los distintos rubros) y países. Respecto al crecimiento, al controlar por los niveles de línea base de los TFP, se observa que casi en todas las medidas existe una relación negativa con el crecimiento futuro y las TFP.

Este proyecto utilizó una mixtura de elementos relativos a datos masivos y estadística tradicional. Por un lado, los datos están montados y fueron analizados en un servidor que permitió una capacidad de cómputo muy por sobre los dispositivos disponibles para uso doméstico (o incluso profesional en algunos casos). Por otro lado, los análisis se corresponden con técnicas de estadística inferencial tradicional, basadas en métodos de mínimos cuadrados ordinarios.

Ahora, este caso abre una discusión muy relevante para la academia y para el sector público respecto a la **representatividad**. Esto porque, pese a que se cuenta con una fuente de datos muy abundante, el continente europeo está mucho más presente en los datos.

Además, al ser una experiencia que utiliza estadística tradicional en muestras de datos que podrían considerarse masivos (debido a su volumen) existe un problema -al que ya se ha visto enfrentada la analítica de datos masivos en general- relativo a la validez de los resultados; este puede sintetizarse en la siguiente interrogante: **¿qué es lo que se pretende, predecir un fenómeno o comprender su causalidad?** En efecto, la analítica de los datos masivos se utiliza típicamente al servicio de la

[5] De esta fuente se obtuvieron también 479 códigos NAICS *North American Industry Classification System'* (‘), que son códigos para clasificar los distintos tipos de negocios, utilizada como estándar en las agencias federales de estadísticas en Estados Unidos (http://www.census.gov/eos/www/naics/).

predicción (pese a que puede utilizarse para estudiar causalidad), pues entrega validez externa a los resultados [60]. Por su parte, los métodos de mínimos cuadrados ordinarios son utilizados para ayudar a modelar explicaciones causales, por su aporte a la validez interna. De hecho, para efectos de los objetivos de la investigación, lo que se precisaba era lo segundo: indagar las causas de una posible variabilidad y desaceleración de la TFP; por lo cual las dificultades ligadas a la representatividad no representarían un problema importante. Sin embargo, en caso de investigar empresas en otros países o predecir su comportamiento a partir de la información disponible, será necesario evaluar la representatividad de la muestra y los métodos utilizados, para asegurar una validez externa apropiada en los resultados.

4. Discusión

Como se pudo constatar en la sección anterior, la analítica de datos masivos permite generar evidencia para el diseño, desarrollo y evaluación de políticas públicas. Esto tiene la potencialidad de mejorar la toma de decisiones y, de esta manera, generar una administración pública que preste mejores servicios a sus ciudadanos.

De esta manera, la analítica avanzada sobre datos masivos se constituye como una herramienta y no un fin en sí mismo. Asimismo, la tecnología es una condición necesaria para desarrollar este proceso, aunque no suficiente. El desarrollo de una inteligencia de valor público en los distintos niveles de gobierno (micro, meso y macro) es estratégico para instalar una cultura de toma de decisiones basada en evidencia. Como toda nueva herramienta tiene ciertas limitaciones metodológicas, cuestionamientos relativos a la privacidad, usos éticos, legales, intelectuales y de seguridad que es necesario tomar en consideración.

En esta sección se discutirán las limitaciones de la analítica avanzada, a la luz del marco conceptual de *'Big Data'* (sección 2) y los casos de estudio (sección 3). Además, se entregarán recomendaciones para la implementación en las agencias públicas y se discutirán algunas oportunidades que se generan, como por ejemplo, la capacidad de compartir y diseminar datos a través de distintas entidades públicas.

4.1. Desafíos y limitaciones

Análisis de datos, metodologías y tecnologías

En primer lugar, se ha igualado la superabundancia de los datos con la representatividad de los mismos. De esta manera, tanto los aspectos metodológicos como la confiabilidad de las fuentes, sean más relevantes que nunca [5], [28]. En concreto, los datos recopilados por canales digitales sólo son representativos de ciertos usuarios más activos y, en el mejor de los casos, sólo de aquellos que tienen acceso a tecnologías de información y comunicación, cuya tasa de penetración en Latinoamérica y el Caribe está lejos del 100% [61]. De esta manera, los fenómenos de baja y sobre-representación, y multiplicidad presentan claros desafíos a la capacidad de realizar inferencias generalizables, ya que cuestionan si los datos masivos representan la diversidad de la población bajo estudio [29].

En segundo lugar, faltan técnicas estocásticas probadas y rigurosas de compensación estadística de errores, sesgos o desviación [29]. Finalmente, los datos masivos han desafiado la definición académica clásica del concepto mismo de "dato". En el marco del trabajo analítico econométrico tradicional, por ejemplo, un dato se define como el valor de una variable parte de un modelo metodológico que apunta

a responder una pregunta de investigación.[6] Ahora bien, en el caso de datos masivos provenientes de transacciones, el concepto de "dato" no está vinculado a un modelo econométrico particular, sino al registro de transacciones y sus "huellas" concomitantes en plataformas o sistemas electrónicos de información [29], [62]. Así, los datos masivos se acercan en definición y tipo a los datos administrativos, ya que han sido creados para fines distintos a los de investigación, generándose de un modo más "orgánico" [29].

Adicionalmente, si bien gran parte del procesamiento de los datos puede automatizarse gracias a la existencia de diversas tecnologías (ver Tabla 1, Tabla **2** y Tabla 3 en la sección 2.2), esto no significa que los científicos de datos no deban tomar una serie de decisiones. Por ejemplo, a la hora de implementar procesos de extracción y limpieza de datos, en el ciclo de vida del análisis de datos (ver

Figura **1** en sección 2.2), esto es, preferencia de ciertos tipos de análisis en desmedro de otros y, finalmente, en la interpretación de los resultados, que no son auto-explicativos [28]. Esto significa que la analítica de datos masivos no es una disciplina enteramente **objetiva**, si no, por el contrario, tiene un componente importante de **subjetividad** [28]. De esta manera, los **10 puntos críticos** [63] para evitar problemas durante los procesos de analítica ligados a datos masivos son seleccionar: **1)** el problema analítico correcto, **2)** la población de datos correcta, **3)** las fuentes de datos correctas, **4)** las muestras correctas, **5)** las versiones adecuadas del modelo, **6)** las variables adecuadas, **7)** los algoritmos y modelos que sean aptos según la naturaleza de las variables, **8)** la frecuencia de validaciones del modelo, **9)** las validaciones y ajustes apropiados para determinar significancia y **10)** las visualizaciones adecuadas.

Aun así, se pueden generar algunos errores de predicción, como en el caso de Google FluTrends – predicción de casos de contagio de influenza- que presentó una gruesa falla a principios de 2013. Se especula que, dado que el algoritmo predice las tendencias de contagio basándose en las búsquedas que se ingresan al sitio de Google, la amplia cobertura de prensa a finales de 2012 gatilló búsquedas relacionadas con la enfermedad de gente que no necesariamente estaba contagiada [8]. Otras explicaciones de la falla apuntan a las sugerencias de diagnóstico de enfermedades entregadas por Google, a partir de síntomas ingresados por los usuarios [9]. Cualquiera sea la explicación, los algoritmos de predicción deben ser constantemente ajustados y además validados con otras fuentes de datos, dado que las búsquedas en la red poseen los sesgos poblacionales previamente discutidos.

En otros casos, el fracaso de la predicción no se debe a los datos mismos, sino a los errores que se comenten en el análisis e interpretación [5]. Es altamente posible que el gran volumen de los datos produzca correlaciones espurias entre variables y alta significancia estadística de los resultados [5], [15], [64]. Siendo un potencial problema el "sobreajuste"[7] de los modelos, y su potencial generalización cuando los datos y, por lo tanto su significado, son absolutamente dependientes del contexto [28].

Un ejemplo de esto es el análisis de los resultados de PISA, que muestran que aquellos sistemas educativos más exitosos son los que combinan calidad con equidad [65], en una relación que no es necesariamente causal. Sin embargo, se implementan reformas basadas en interpretaciones de grandes tendencias de evaluaciones nacionales e internacionales, sin un profundo entendimiento de los detalles que hacen la diferencia en las escuelas [10].

6 Por ejemplo, la variable "salario" puede tener valores discretos dentro de un rango esperado, y dicho número es lo que en econometría se denomina dato.

7Del inglés "*overfitting*", corresponde cuando un modelo estadístico se ajusta demasiado a los datos de entrenamiento, reduciendo su validez predictiva fuera de dicho conjunto de datos.

En conclusión, no se puede esperar que el uso de analítica avanzada de datos masivos sustituya por sí misma a métodos de investigación y análisis más tradicionales, sino que, por el contrario, debe y puede servir de complemento para otros [66], especialmente de índole cualitativa.

Privacidad, aspectos éticos y legales, seguridad y pertenencia

La cantidad de variables personales que contienen los datos masivos que se analizan tanto para el uso público como privado levantan inevitablemente consideraciones éticas y legales respecto a:

(1) La **protección de la información privada de las personas**, es decir, mantener el anonimato de las personas cuyos datos se están analizando.

(2) El **análisis de los datos privados**, es decir, el cuestionamiento acerca de la inequidad o perjuicio que genera la intromisión en la vida privada[8] de las personas [67], [68].

(3) La **propiedad de los datos masivos** y los correspondientes derechos y licencias para su administración, mantenimiento, explotación y uso.

Con respecto a la **protección de la información privada de las personas,** a partir de la combinación de distintas fuentes permite conocer mucha información de individuos que puede llevar a su identificación. Por ejemplo, Sweeney [69] estima que tan solo conociendo el código postal, género, y fecha de nacimiento se puede identificar al 87% de la población de los EE.UU. Del mismo modo, un estudio de Bahamonde et al[70] evidencia cuán fácil es obtener y calcular la dirección de la residencia utilizando la información almacenada en las tarjetas de prepago (bip!) del transporte público de Santiago de Chile.

En el caso del **análisis de los datos privados**, la intromisión en la vida privada de las personas puede dar pie a prácticas discriminatorias como la elegibilidad a empleos o acceso a servicios.

Ligado a lo anterior, el asunto de la seguridad de la información se refiere a la protección de la privacidad de los datos durante su captura, gestión y análisis. Un recurso que ha venido a hacerse cargo de esto, ha sido el cifrado de los datos, tanto en su almacenamiento como en los canales de distribución.

Otro aspecto que ha generado interrogantes se relaciona con la **propiedad de los datos**: ¿Quién es el dueño de un conjunto de datos?, ¿los propietarios de plataformas tecnológicas (como Facebook) o quienes generan las transacciones en dichas plataformas (o sea, los individuos que crean perfiles y pueblan sus muros virtuales de datos masivos)?, ¿qué derechos o licencias de uso tiene asociada a dicha pertenencia?, ¿qué define un uso “justo” y “seguro”, o sea, que respete la integridad personal de los individuos? En resumidas cuentas, no porque los datos estén disponibles públicamente significa que sea ético utilizarlos; es necesario que exista un uso consciente como mecanismos de rendición de cuentas por parte de quienes analizan la información [28], especialmente cuando se trata de servidores públicos. En este contexto, es importante que, cuando entidades externas (universidades o empresas) realicen análisis, se resguarde adecuadamente la propiedad de los datos, se establezcan mecanismos para su protección, y una prohibición de uso posterior para otros fines. Ese tipo de temas legales necesitan de pronta clarificación [1], y estos aspectos deben ser cuidadosamente ponderados; especialmente cuando se externalizan los análisis a terceros o emplean productos que estipulan en sus términos y condiciones que se conceden permisos para el acceso a los datos analizados o se transfiere la propiedad de los mismos al fabricante del producto o proveedor del servicio. Por ejemplo,

[8] Un ejemplo de esto último es el del supermercado estadounidense Target, que, en función de la compra de ciertos productos, predice qué clientes están embarazadas y les envía promociones acordes. El problema ocurrió cuando una adolescente embarazada, que no lo había hablado con sus padres, recibió una de estas promociones [50].

¿los datos originados por actividad de teléfonos celulares le pertenecen a las compañías telefónicas o al cliente del servicio?, o ¿realmente el manejo de niveles de agregación permite el absoluto anonimato de los individuos?, son parte ineludible de una definición de grandes datos para la formulación de políticas públicas.

Finalmente, la **responsabilidad legal** identifica quién es responsable cuando las acciones de análisis de datos masivos generan consecuencias negativas: aspectos de pertenencia y protección de datos, privacidad de las personas y protección del consumidor, problemas con la seguridad de los datos entre otros [71].

4.2. Recomendaciones

A continuación, tanto de la revisión de los casos como de la discusión del marco conceptual propuesto, se proponen una serie de recomendaciones.

Sobre la adopción de la inteligencia de valor público en las agencias de gobierno

En primer lugar, para implementar proyectos de inteligencia de valor público se requiere de una serie de **capacidades institucionales** dentro del gobierno. Algunos autores identifican a lo menos tres dimensiones: capital humano, tecnología y desarrollo de estrategias [30], [43], [56]. Estas se detallan a continuación.

- **Capital humano**: para cubrir tareas como estudiar y pensar respecto de la información disponible; limpiar, preparar, formatear y asegurar la confiabilidad de los datos y realizar capacitación específica en el análisis de datos y soluciones basadas en ellas. Además, hay una carencia de consumidores inteligentes que analicen críticamente la información; es necesario un liderazgo para la mejora educativa y el uso de datos, y una cultura organizacional orientada al uso de datos.
- **Tecnología**: existe una carencia de recursos tecnológicos para el uso de grandes conjuntos de datos y los servicios de software y almacenamiento asociados a ellos. Igualmente, falta interoperabilidad entre sistemas de distintas agencias y/o departamentos, y herramientas para generar acciones a partir de los datos.
- **Desarrollo de estrategias**: es necesario un plan que determine qué preguntas son urgente contestar, qué datos recopilar y con qué técnicas analizarlas. Además de alianzas estratégicas con organizaciones cuya misión es apoyar el uso de datos, recursos calidad y confiabilidad de la información disponible.

En consecuencia, establecer una **institucionalidad** es importante para mantener las iniciativas en el tiempo. [55]) describen cómo Bahía Blanca fue paulatinamente cerrando su iniciativa. Al contrario, el caso de Fortaleza da cuenta cómo el establecimiento de la institucionalidad aseguró la continuidad, los recursos, las líneas de trabajo e incluso el acceso a los datos. Además, la institucionalidad debe hacerse cargo de lograr una comunicación **transparente y fluida con otras entidades externas** (ver sección 3.2), enfrentando el desafío básico de organizar y compartir los datos que serán insumo de los análisis. Esto significa, compartir los datos entre sus distintas agencias [20] y desarrollar un liderazgo dentro del gobierno para establecer cómo se usarán los datos masivos y para qué. También es relevante que los propios tomadores de decisiones se involucren, permitiendo un acceso oportuno a los datos (sobre todo a aquellos con una utilidad limitada a ciertos periodos de tiempo), y facilitando el cambio cultural hacia procesos de toma de decisiones basadas en estos. En el caso de Córdoba, se crearon vínculos entre una universidad y una institución pública, además de generar incentivos para promover la inversión privada en tecnología al servicio del cumplimiento de objetivos públicos.

Otro aspecto necesario de tomar en cuenta es la **comunicación clara y transparente con la ciudadanía** (ver sección 3.2). Ejemplos de estos son los casos de Bahía Blanca y São Bernardo do Campo. En el primero, se logró establecer una buena comunicación con la ciudadanía, adaptando la información a sus necesidades. En el segundo, se logró hacer partícipe a la ciudadanía en el desarrollo de la aplicación (a través de una hackatón) y en la producción de los datos.

Respecto al **capital humano,** son necesarios tanto los profesionales que llevan a cabo el proceso de análisis como los "**consumidores inteligentes**" de la evidencia producida. La labor de estos últimos es la formulación de preguntas y el análisis crítico de la información recibida, cuestionando las fuentes, supuestos y metodología utilizados para producirlos [43]. En este documento, se propone una rúbrica de competencias de los consumidores inteligentes (**Anexo B**).

Respecto a los **científicos de datos**, se requieren profesionales con competencias bastante específicas para llevar a cabo análisis que produzcan información valiosa y que alimente el ciclo de decisiones guiadas por datos. Sin embargo, estos profesionales son escasos. Se estima que sólo en EE.UU. para el año 2018, habrá un déficit de entre 140.000 y 190.000 profesionales que puedan emprender este desafío y un millón y medio de gerentes y analistas con conocimientos suficientes como para poder generar las preguntas adecuadas y poder comprender los resultados objetivos [1]. Por lo tanto, se requiere formar y capacitar profesionales para cerrar esta brecha.

Lamentablemente, las dificultades enfrentadas en torno a la capacidad profesional no se acaban en la potencial escasez. La utilización progresiva de datos relativos a la conducta humana, se alejan del área de conocimiento dentro de la ingeniería o las llamadas ciencias duras. Esta situación demanda de un análisis trabajo multidisciplinario y multisectorial, en diferentes contextos sociales, demográficos y geográficos. En otras palabras, se necesitan profesionales que puedan desenvolverse competentemente en esa diversidad.

Transparentar la analítica utilizada para generar la evidencia

Como se mencionó anteriormente, la Ciencia de Datos requiere tomar una serie de decisiones en los análisis, y lidiar con supuestos que son discutibles e impactan en la evidencia generada para la toma de decisiones y la generación de políticas públicas. Por lo tanto, es sumamente importante **documentar y transparentar los procesos de análisis llevados a cabo**, para que sean auditables y respondan a los mecanismos de rendición de cuentas. De esta forma, existe la oportunidad para la mejora continua de los análisis y sus resultados, la diseminación dentro del sector público de las metodologías empleadas y—sobre todo— la posibilidad de corregir errores a tiempo. Especialmente cuando existen filtraciones de la información privada de las personas, o se producen inequidades producto de las recomendaciones erradas de un algoritmo. Al respecto, se ha creado una corriente de investigación denominada *E-science* que hace énfasis en la trazabilidad y reproducibilidad de los resultados de la investigación científica, especialmente la que proviene de usos masivos mediante el uso de tecnologías de información [72].

4.3. Oportunidades

Finalmente, a partir de los desafíos y recomendaciones, y del análisis de los casos detectamos algunas oportunidades que sintetizaremos en esta sección a continuación.

Nivel de desarrollo (o madurez) de proyectos de datos masivos y de los "consumidores inteligentes" de evidencia basada en análisis de datos masivos

Una primera oportunidad es la utilización de la rúbrica desarrollada por Townsend & Zambrano-Barragan [55] para evaluar las iniciativas de relacionadas con masivos urbanos (**Anexo A**). Este es un instrumento que, con ligeras adaptaciones, puede utilizarse para evaluar la madurez general de

cualquier proyecto de análisis de datos masivos dentro del sector público. En este caso, la dimensión de "Participación y servicios públicos" puede adaptarse rápidamente al contexto del proyecto en evaluación. Del mismo modo, se debe tener en consideración la rúbrica sobre "consumidores inteligentes" (**Anexo B**) para asegurar un nivel mínimo de competencias para lograr hacer interpretación y uso de la evidencia producida a partir del análisis de datos masivos.

Compartir y diseminar datos dentro del sistema público

Una segunda oportunidad que se vislumbra es que los proyectos de analítica pública generen una serie de datos con un uso potencial más allá de la iniciativa que los produjo. Por ejemplo, en el caso de la aplicación móvil *Você SB* desarrollada en São Bernardo do Campo (Brasil), los datos que producen los ciudadanos pueden servir para diferentes agencias de gobierno en temas de seguridad, medioambiente entre otros.

Por lo tanto, existen claras oportunidades de sinergia entre distintas agencias de gobierno con el potencial de desarrollar análisis que den cuenta de la necesidad de toma de decisiones y desarrollo de políticas multisectoriales, como por ejemplo el caso del transporte, la contaminación ambiental y la concentración de centros educativos [73]. Esto requiere avanzar en políticas de creación y curación continua de los datos generados, y el desarrollo de una institucionalidad que: **1)** lidere el uso de análisis sobre datos masivos para generar una cultura de toma de decisiones basadas en evidencia; **2)** se haga cargo de la sustentabilidad de mantener y administrar los datos masivos manteniendo todos los resguardos necesarios (ver sección 4.1), y **3)** promueva la comunicación clara y fluida con otras agencias gubernamentales y entidades externas (como universidades y centros de investigación).

Tipos de problemática a abordar

Una tercera familia de oportunidades para el uso de analítica sobre datos masivos es un tipo específico de problemas de "predicción pura" [60]. Estos problemas son aquellos que no necesitan establecer una causalidad para apoyar la toma de decisiones, en el escenario típico de evaluación con o sin la implementación de una política pública. Kleinberg et al. [60] denominan estos problemas del "tipo paraguas", esto es, que tienen que ver con una decisión a tomar. Por ejemplo: ¿es la posibilidad de lluvia lo suficientemente alta para salir con paraguas?: en dicho caso, no se necesita saber qué causa la lluvia, solo estimar si se producirá o no.

De esta manera, para apoyar la toma de decisiones, se pueden emplear técnicas de aprendizaje automático supervisado (ver **Tabla 2**), utilizando datos históricos para entrenar algoritmo que entreguen un pronóstico de mejor calidad y más oportuno que el que podría realizar un experto humano. Así, por ejemplo, se han generado aplicaciones para:

- Estimar automáticamente el nivel socioeconómico de una cierta zona geográfica en base a información satelital para establecer políticas de ayuda social en un territorio [74].
- Estimar el riesgo de deserción escolar de un alumno [75], [76] y escoger qué intervenciones es más costo-efectiva para retenerlo [77].
- Mejora de las políticas de fiscalización usando inspecciones predictivas basadas en reseñas en línea de clientes [78].

Referencias bibliográficas

[1] J. Manyika *et al.*, "Big Data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, 2011.

[2] Staff Science, "Challenges and Opportunities," *Science*, vol. 331, no. 6018, pp. 692–693, Nov. 2011.

[3] The Economist, "Data, data everywhere," *The Economist*, Feb-2010.

[4] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, Apr. 2011.

[5] S. T. McAbee, R. S. Landis, and M. I. Burke, "Inductive reasoning: The promise of Big Data," *Hum. Resour. Manag. Rev.*, 2016.

[6] J. Bertolucci, "Big Data's new buzzword: datafication," *InformationWeek*, 2013. [Online]. Available: http://www.informationweek.com/big-data/big-data-analytics/big-datas-new-buzzword-datafication/d/d-id/1108797?print=yes. [Accessed: 19-Dec-2016].

[7] C. Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *WIRED*, 23-Jun-2008. [Online]. Available: https://www.wired.com/2008/06/pb-theory/. [Accessed: 18-Dec-2016].

[8] D. Butler, "When Google got flu wrong," *Nat. News*, vol. 494, no. 7436, p. 155, Feb. 2013.

[9] T. Harford, "Big data: are we making a big mistake?," *Financial Times*, 28-Mar-2014. [Online]. Available: https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0. [Accessed: 19-Dec-2016].

[10] V. Strauss, "'Big data' was supposed to fix education. It didn't. It's time for 'small data.,'" *The Washington Post*, 09-May-2016. [Online]. Available: https://www.washingtonpost.com/news/answer-sheet/wp/2016/05/09/big-data-was-supposed-to-fix-education-it-didnt-its-time-for-small-data/. [Accessed: 19-Dec-2016].

[11] J. Chen *et al.*, "Big Data challenge: a data management perspective," *Front. Comput. Sci.*, vol. 7, no. 2, pp. 157–164, Apr. 2013.

[12] L. Manovich, "Trending: The promises and the challenges of big social data," *Debates Digit. Humanit.*, vol. 2, pp. 460–475, 2011.

[13] M. R. Parks, "Big Data in communication research: Its contents and discontents," *J. Commun.*, vol. 64, no. 2, pp. 355–360, Abril 2014.

[14] D. J. Power, "Using 'Big Data' for analytics and decision support," *J. Decis. Syst.*, vol. 23, no. 2, pp. 222–228, Abril 2014.

[15] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[16] A. Gandomi and M. Haider, "Beyond the hype: Big Data concepts, methods, and analytics," *Int. J. Inf. Manag.*, vol. 35, no. 2, pp. 137–144, Abril 2015.

[17] R. Kitchin, "Big data and human geography: Opportunities, challenges and risks," *Dialogues Hum. Geogr.*, vol. 3, no. 3, pp. 262–267, 2013.

[18] R. Kitchin, “Big Data, new epistemologies and paradigm shifts,” *Big Data Soc.*, vol. 1, no. 1, pp. 1–12, 2014.

[19] I.-Y. Song and Y. Zhu, “Big Data and data science: what should we teach?,” *Expert Syst.*, vol. 33, no. 4, pp. 364–373, Agosto 2016.

[20] L. Tomar, W. Guicheney, H. Kyarisiima, and T. Zimani, “Big Data in the public sector: Selected applications and lessons learned,” Inter-American Development Bank, 2016.

[21] L. Hill, F. Levy, V. Kundra, B. Laki, and J. Smith, *Data-Driven Innovation for Growth and Well-being*. OECD, 2014.

[22] L. Rodríguez-Mazahua, C.-A. Rodríguez-Enríquez, J. L. Sánchez-Cervantes, J. Cervantes, J. L. García-Alcaraz, and G. Alor-Hernández, “A general perspective of Big Data: applications, tools, challenges and trends,” *J. Supercomput.*, vol. 72, pp. 3073–3113, 2015.

[23] B. F. Welles, “On minorities and outliers: The case for making Big Data small,” *Big Data Soc.*, vol. 1, no. 1, pp. 1–2, 2014.

[24] K. Nahon and J. Hemsley, *Going viral*, 1st ed. Polity Press, 2013.

[25] J. D. Morrison and J. D. Abraham, “Reasons for enthusiasm and caution regarding Big Data in applied selection research,” *Ind. Psychol.*, vol. 52, no. 3, pp. 134–139, 2015.

[26] L. Taylor, R. Schroeder, and E. Meyer, “Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?,” *Big Data Soc.*, vol. 1, no. 2, p. 2053951714536877, 2014.

[27] United Nations Global Pulse, “Big Data for development: opportunities & challenges”: A Global Pulse White Paper,” United Nations Global Pulse, 2012.

[28] D. Boyd and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” *Inf. Commun. Soc.*, vol. 15, no. 5, pp. 662–679, 2012.

[29] F. Kreuter and R. D. Peng, “Extracting Information from Big Data: Issues of Measurement, Inference and Linkage,” in *Privacy, Big Data, and the Public Good*, J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, Eds. Cambridge University Press, 2014, pp. 257–275.

[30] J. A. Marsh, J. F. Pane, and L. S. Hamilton, “Making Sense of Data-Driven Decision Making in Education,” 2006. [Online]. Available: http://www.rand.org/pubs/occasional_papers/OP170.html. [Accessed: 28-Jan-2017].

[31] UNESCO, “Policy brief - Learning Analytics.” UNESCO Institute for Information Technologies in Education, 2012.

[32] B. Schmarzo, *Big Data: Understanding How Data Powers Big Business*. Wiley, 2013.

[33] A. Labrinidis and H. V. Jagadish, “Challenges and Opportunities with Big Data,” *Proc. VLDB Endow.*, vol. 5, no. 12, pp. 2032–2033, 2012.

[34] A. E. Joseph and P. R. Bantock, “Measuring potential physical accessibility to general practitioners in rural areas: a method and case study,” *Soc. Sci. Med.*, vol. 16, pp. 85–90, 1982.

[35] W. Luo and Y. Qi, “An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians,” *Health Place*, vol. 15, no. 4, pp. 1100–1107, Diciembre 2009.

[36] W. Luo and F. Wang, “Measures of spatial accessibility to health care in a GIS environment: synthesis and a case study in the Chicago region,” *Environ. Plan. B Plan. Des.*, vol. 30, no. 6, pp. 865 – 884, 2003.

[37] J. Radke and L. Mu, “Spatial Decompositions, Modeling and Mapping Service Regions to Predict Access to Social Programs,” *Geogr. Inf. Sci.*, vol. 6, no. 2, pp. 105–112, Diciembre 2000.

[38] Z. Wei, “A study of accessibility to health facilities for elderly population in metro Atlanta using a categorical multi-step floating catchment area method,” Thesis, uga, 2013.

[39] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

[40] L. Anselin, “Local Indicators of Spatial Association—LISA,” *Geogr. Anal.*, vol. 27, no. 2, pp. 93–115, 1995.

[41] H. J. Miller, “Tobler’s First Law and Spatial Analysis,” *Ann. Assoc. Am. Geogr.*, vol. 94, no. 2, pp. 284–289, Jun. 2004.

[42] J. I. Barredo, M. Kasanko, N. McCormick, and C. Lavalle, “Modelling dynamic spatial processes: simulation of urban future scenarios through cellular automata,” *Landsc. Urban Plan.*, vol. 64, no. 3, pp. 145–160, Jul. 2003.

[43] M. Bienkowski, M. Feng, and B. Means, “Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics.” U.S. Department of Education, Office of Educational Technology, 2012.

[44] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag, 2002.

[45] V. Friedman, “Data Visualization and Infographics | Smashing Magazine,” 2008. [Online]. Available: http://www.smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics/. [Accessed: 18-Jun-2014].

[46] M. Lima, *Visual Complexity: Mapping Patterns of Information*. Princeton Architectural Press, 2011.

[47] J. Steele and N. Iliinsky, Eds., *Beautiful Visualization: Looking at Data through the Eyes of Experts*, 1st ed. O’Reilly Media, 2010.

[48] N. Yau, *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*, 1st ed. Wiley, 2011.

[49] IBM, “What is a Data Scientist? – Bringing big data to the enterprise,” 2015. [Online]. Available: http://www-01.ibm.com/software/data/infosphere/data-scientist/?cm_mc_uid=23497733502814413056500&cm_mc_sid_50200000=1441305650. [Accessed: 07-Sep-2015].

[50] L. Floridi, “Big data and their epistemological challenge,” *Philos. Technol.*, pp. 1–3, 2012.

[51] K. M. Kelm *et al.*, “Big data innovation challenge : pioneering approaches to data-driven development,” The World Bank, 107751, Jan. 2016.

[52] W. MacLeod, J. Bor, K. Crawford, and S. Carmona, “Analysis of Big Data for better targeting of ART adherence strategies : spatial clustering analysis of viral load suppression by South African province, district, sub-district and facility (April 2014-March 2015),” The World Bank, 2015.

[53] S. Goldsmith, S. Crawford, and B. Weinryb Grohsgal, “Innovations in public service delivery - Issue No. 4: predictive analytics: driving improvements using data,” IDB, Documento para discusión IDB-DP-440, Jul. 2016.

[54] H. Terraza, P. Deregibus, C. Galeota, and M. Ponce de León, “Movilidad urbana sostenible, datos masivos y políticas públicas: estudio de la movilidad de los ciclistas en la ciudad de Rosario (Argentina) a través del uso de dispositivos de geo-referenciación,” Banco Interamericano de Desarrollo, 2016.

[55] A. Townsend and P. Zambrano-Barragan, “Computing a new trajectory for urban gobernance: Urban Big Data innovation in Latin America and the Caribbean,” Banco Interamericano de Desarrollo, 2016.

[56] B. Means, C. Padilla, A. DeBarger, and M. Bakia, *Implementing Data-Informed Decision Making in Schools: Teacher Access, Supports and Use*. US Department of Education, 2009.

[57] D. Comin, “Economic Growth,” in *Economic Growth*, S. N. Durlauf and L. E. Blume, Eds. Palgrave Macmillan UK, 2010, pp. 260–263.

[58] Bureau van Dijk, “Orbis | Detailed global private company information,” 2017. [Online]. Available: http://www.bvdinfo.com/en-gb/our-products/company-information/international-products/orbis. [Accessed: 12-Jan-2017].

[59] D. Bahar, “Using firm-level data to study growth and dispersion in total factor productivity,” The Brookings Institution Harvard Center for International Development, 2016.

[60] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, “Prediction policy problems,” *Am. Econ. Rev.*, vol. 105, no. 5, pp. 491–495, 2015.

[61] CEPAL, “Estado de la banda ancha en América Latina y el Caribe 2016,” CEPAL, Sep. 2016.

[62] J. Howison, A. Wiggins, and K. Crowston, “Validity Issues in the Use of Social Network Analysis with Digital Trace Data,” *J. Assoc. Inf. Syst.*, vol. 12, no. 12, Dec. 2011.

[63] J. Kobielus, “Data Scientist: Master the Basics, Avoid The Most Common Mistakes,” *IBM Data&Analytics Hub*, 02-Jul-2013. .

[64] K. Crawford, “The hidden biases in Big Data,” *Harvard Business Review*, 01-Apr-2013.

[65] OECD, *Equity and Quality in Education*. Paris: Organisation for Economic Co-operation and Development, 2012.

[66] D. Lazer, R. Kennedy, G. King, and A. Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis,” *Science*, vol. 343, no. 6176, pp. 1203–1205, Mar. 2014.

[67] C. Miller, “When Algorithms Discriminate,” *The New York Times*, 09-Jul-2015.

[68] Nature, “More accountability for big-data algorithms,” *Nat. News*, vol. 537, no. 7621, p. 449, Sep. 2016.

[69] L. Sweeney, “Simple Demographics Often Identify People Uniquely,” Pittsburgh, 2000.

[70] J. Bahamonde, A. Hevia, G. Font, J. Bustos-Jiménez, and C. Montero, “Mining Private Information from Public Data: The Transantiago Case,” *IEEE Pervasive Comput.*, vol. 13, no. 2, pp. 37–43, Abril 2014.

[71] European Big Data Value Partnership, “European Big Data value strategic research & innovation agenda: Version 0.99,” Jul. 2014.

[72] T. Hey, S. Tansley, and K. Tolle, “Jim Gray on eScience: A transformed scientific method,” *Fourth Paradigm Data-Intensive Sci. Discov.*, vol. 1, 2009.

[73] P. Rodríguez *et al.*, “Apoyando la formulación de políticas públicas y toma de decisiones en educación utilizando técnicas de análisis de datos masivos: el caso de Chile,” 2016.

[74] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, “Combining satellite imagery and machine learning to predict poverty,” *Science*, vol. 353, no. 6301, pp. 790–794, Aug. 2016.

[75] C. Escobar and F. Lolas, “Desarrollo de un sistema prototipo para la detección temprana de la deserción escolar en escuelas públicas chilenas",” Memoria de Título, Universidad Adolfo Ibáñez, Santiago de Chile, 2015.

[76] Mineduc, “Informe de Piloto de Modelo Predictivo, Seguimiento de Estrategias de Apoyo (Sistema de Alerta Temprana).” Jul-2015.

[77] Microsoft, “Predicting student dropout risks, increasing graduation rates with cloud analytics,” Aug-2016. [Online]. Available: https://customers.microsoft.com/en-us/story/tacomapublicschoolsstory. [Accessed: 14-Dec-2016].

[78] J. S. Kang, P. Kuznetsova, M. Luca, and Y. Choi, “Where not to eat? Improving public policy by predicting hygiene inspections using online reviews.,” in *EMNLP*, 2013, pp. 1443–1448.

[79] P. Rodríguez and J. Mondaca, “Rúbrica de competencias de los consumidores inteligentes de inteligencia de valor público.” 2016.

[80] G. Bellinger, D. Castro, and A. Mills, “Data, information, knowledge, and wisdom,” 2004.

[81] M. O. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications*, 1 edition. Natick, Mass: A K Peters/CRC Press, 2010.

[82] P. R. Keller and M. M. Keller, *Visual cues: practical data visualization*, vol. 2. IEEE Computer Society Press Los Alamitos, CA, 1993.

Anexo A: Rúbrica del modelo de madurez de uso *Big Data* en problemas urbanos

Nivel de madurez	Datos abiertos	Cultivando ecosistemas de datos	Analítica	Toma de decisiones basada en datos	Participación y servicios públicos
Nivel 5 — Optimizando (Colaboración urbana inteligente)	Plataformas de lectura-escritura facultan a las comunidades usuarias para curación y la extensión de los datos. Los protocolos de gobernabilidad incorporados en el software permiten compartir de manera responsable.	Industria, academia, gobierno y ciudadanos comparten datos confiables. Los mercados de datos crean una plataforma segura y segura para el intercambio muchos-a-muchos de *Big Data* urbanos.	Plataformas analíticas abiertas permiten una rápida innovación en algoritmos. Auto-optimización de operaciones a través de la automatización extensiva de los análisis.	Organización y sus operaciones se adaptan y mejoran continuamente utilizando una visión analítica acorde con los objetivos estratégicos de la política. Los procesos que requieren un modesto juicio humano están sujetos a una automatización potencial.	Visión basada en ciudadanía y programa de innovación en gobernabilidad. Plataformas abiertas de innovación para servicios públicos basados en datos, gobernanza compartida de datos.
Nivel 4 — Avanzado (Administración inteligente)	Todos los datos no confidenciales son publicados abiertamente con robusto apoyo a la comunidad de usuarios de datos y los procesos existentes de solicitud de datos.	Mayoría de los datos útiles son 'grandes'; recopilación de datos colaborativa y abierta generalizada. Intercambio externo de datos con el sector privado; son comunes los incentivos para el intercambio de datos.	Analítica predictiva se utiliza ampliamente para optimizar la toma de decisiones de la organización de modo que se tomen las mejores medidas para maximizar la eficacia operacional y lograr resultados de políticas.	Tomadores de decisiones están bien informados con el conocimiento de la analítica y la organización es capaz de actuar para maximizar los indicadores clave de rendimiento. Procesos que requieren poco juicio humano son automatizados.	Completa integración de servicio en toda la ciudad con potenciales consumidores ciudadanos impulsando la innovación de servicios. Sólida gestión de innovación entre departamentos.
Nivel 3 — Intermedio (Toma de decisiones inteligente)	La política y la regulación de datos abiertos establecen un calendario para la divulgación completa de datos, sujeto a la revisión de seguridad y privacidad. Datos en tiempo real publicados cuando sea factible.	Redes de sensores integradas que soportan múltiples usuarios. Las plataformas de datos permiten el intercambio automatizado. Integración de datos de diversas fuentes.	Los análisis predictivos proporcionan información sobre la probabilidad de cambios importantes en los patrones de actividad que afectan las operaciones o políticas de la organización. Acelerando las mejoras a través	La organización es capaz de tomar decisiones de negocio limitadas usando la visión analítica para mejorar la eficiencia operativa y generar más valor. Los paneles de datos soportan una cultura basada en datos.	Visión, estrategia e implementación iniciadas para la participación y los servicios públicos basados en datos. La plataforma integrada de entrega incorpora ciclos de retroalimentación ciudadana.

			del aprendizaje automático y otras técnicas avanzadas de Inteligencia Artificial		
Nivel 2 — Básico (Gobierno de una ciudad inteligente)	El portal de datos abiertos agrega los conjuntos de datos públicos publicados.	Las redes de sensores específicas de la aplicación recopilan datos relevantes. Políticas para la privacidad de los datos, la seguridad y el intercambio establecidos. La calidad de los datos es deficiente. La interconexión requiere una integración manual que consume tiempo.	Los análisis se utilizan para informar a los responsables de la toma de decisiones sobre las causas y los factores que contribuyen a los procesos y eventos clave en las operaciones de la organización.	La organización entiende las causas detrás de lo que observan, pero su cultura es en gran medida resistente a la adaptación para aprovechar dicha visión.	Nichos de innovación en los servicios públicos, con cierta integración e intercambio de datos entre distintos departamentos. Participación ciudadana limitada.
Nivel 1 — Ad Hoc	El intercambio de datos se habilita a través de las regulaciones especiales y la política departamental.	Las agencias se basan en datos históricos de escape de operaciones. Los datos están en silos con poco intercambio.	Los análisis se limitan a describir lo que ha sucedido.	La aplicación de la visión analítica es la elección del individuo y tiene poco efecto en cómo funciona la organización.	Poca participación o uso de datos en la gestión de gobierno o la prestación de servicios. Las estrategias de servicios públicos digitales no existen o no están aisladas.

Fuente: Townsend & Zambrano-Barragan[55].

Anexo B: Rúbrica de competencias de los consumidores inteligentes de inteligencia de valor público [79]

Objetivos

El objetivo de la rúbrica es presentar criterios generales para que el usuario final de los resultados de proyectos de inteligencia de valor público pueda autoevaluar su capacidad para aproximarse, comprender y, en último término, tomar decisiones a partir de la información presentada.

Marco conceptual

Si bien la literatura diferencia distintos tipos de alfabetismo ("*literacies*" en inglés), para efectos de la rúbrica propuesta se hablará de alfabetismo de la información. Entendiendo información desde las ciencias de la computación [80], que diferencia entre datos e información. En este contexto, los datos hacen referencia a valores "en bruto" (que pueden estar organizados o no en una matriz o base de datos) y cuya principal característica es no haber sido sometidos a ningún tipo de procedimiento o análisis. De esta forma, aún no constituyen un sustrato a partir del cual puede generarse alguna conclusión o tomarse una decisión. Por su parte, la información ya ha sido sometida a algún tipo de procesamiento, y puede ser utilizada para generar decisiones, juicios o conclusiones. Dado que el usuario final será expuesto a información y no a datos, la rúbrica versará sobre el alfabetismo de la información.

Se utilizan dos insumos para la construcción de la rúbrica; primero, los componentes, conceptos y habilidades descritas por Means et al. [56] en el contexto de la evaluación de la capacidad de tomar decisiones a partir de datos, y, segundo, las características y tareas que deben definirse para el usuario de visualizaciones, formuladas por Ward et al [81].

El instrumento construido por Means et al. [56] contempla cinco componentes[9]:

1. **Formulación de pregunta**: generar una pregunta que pueda ser respondida desde los datos que están disponibles. Esto contempla una relación semántica entre la pregunta propuesta y la estructura de los datos.
2. **Localización**: identificar los datos relevantes para responder la pregunta.
3. **Comprensión**: conocimiento del medio a través del que se presentan los datos (por ejemplo, una tabla o gráfico), permitiendo la estructuración de una respuesta coherente a la pregunta formulada.
4. **Interpretación**: el conocimiento, al menos cualitativo, de nociones estadísticas que permitan entender el comportamiento de los datos. Por ejemplo, uno o varios valores atípicos pueden sesgar el promedio de una medición determinada.
5. **Uso**: posibilidad de asociar los datos con los contextos específicos en que deben ser usados. Entendimiento de cómo los datos reflejan o pretenden reflejar un fenómeno particular.

Si bien, Means et al. [56] hablan de datos cuando se refieren a lo que mide el instrumento, abarcan desde el procesamiento de los datos hasta la consecución de información.

En términos generales, la rúbrica pretende evaluar información y no a datos, pues los productos entregados ya vendrán procesados y, por lo mismo, limitados. Por un lado, esto implica que la manipulación de los datos por parte del usuario final es mínima o nula (o cosmética, en el mejor de los

[9] Para ser coherentes con las definiciones de Means et al. [56], también se hablará de datos y no de información.

casos), y por otro, que la formulación de preguntas se restringiría a los límites de la información entregada.

En relación a las visualizaciones, Ward et al [81] definen tareas y características de usuario. Las tareas refieren a acciones necesarias para generar un juicio válido a partir de la visualización presentada. Para esto toman una lista de acciones desarrollada por [82]:

1. **Identificar**: reconocer los elementos relevantes en el medio presentado.
2. **Localizar**: establecer la posición de un objeto en el medio presentado.
3. **Distinguir**: determinar que un elemento es distinto de otro.
4. **Categorizar**: clasificar elementos en distintos tipos.
5. **Agrupar**: agrupar elementos basado en algún tipo de relación o elemento común.
6. **Priorizar**: organizar un conjunto de elementos en un determinado orden.
7. **Comparar**: examinar las similitudes y diferencias entre dos o más elementos.
8. **Asociar**: generar una relación entre dos o más elementos.
9. **Correlacionar**: establecer una relación recíproca entre dos o más elementos.

Estas tareas se utilizarán para definir los distintos niveles de experticia que requiere el usuario para operar con visualizaciones, y serán operacionalizadas a través de los descriptores de la rúbrica vinculados con visualizaciones.

Las características del usuario descritas por Ward et al [81] son cinco y guardan relación con conocimientos y habilidades de los datos[10] que debe tener un usuario para una comprensión adecuada de la información. Las características propuestas por los autores son:

1. **Familiaridad con el dominio**: La experticia y conocimiento que el usuario posee respecto del dominio o contexto en que opera los datos con los que se está trabajando.
2. **Familiaridad con la tarea**: La experiencia del usuario en torno a la tarea asignada.
3. **Familiaridad con los datos**: La experiencia que tenga el usuario con los datos a partir de los cuales se obtuvo la información; y si ha tenido la oportunidad de generar algún modelo mental para darles coherencia.
4. **Familiaridad con la técnica de visualización**: La experiencia y experticia que tenga el usuario utilizando técnicas específicas de visualización.
5. **Familiaridad con el entorno de visualización**: El conocimiento y experticia que tenga el usuario con la herramienta(s) de visualización con las que se esté presentando la información.

Componentes

Algunos de los componentes propuestos por [56] se superpondrían al ser utilizados en el contexto específico de la rúbrica. Sin ir más lejos, la formulación de preguntas se traslaparía con el uso, esto porque el usuario debe conocer y comprender los contextos de uso de la información para generar una pregunta que sea coherente y que pueda responderse. Por su parte, la localización y la comprensión también tendrían elementos comunes, ya que, al no poder manipular los datos, el usuario tampoco podrá manipular las formas de presentar la información. Esto implica que la localización relevante quedaría supeditada a un correcto conocimiento de la forma en que la información es presentada. Por lo mismo, se proponen tres componentes para la construcción de la rúbrica, a partir de los propuestos por [56].

[10] Al igual que Means et al. [56], Ward et al. [81] hablan de datos.

Identificación de la información: conocimiento de la información, independiente de cómo se presente. Implica reconocer el tipo de información presentada y conocer los contextos en los que opera, y en los que no.

1. **Comprensión de la información**: conocimiento relativo a la forma en que se presenta la información y a las habilidades para comprenderla. Implica conocer los medios a través de los cuales se presenta la información, localizar información relevante, comparar sus elementos, entre otros.
2. **Interpretación de la información**: uso de información externa a la presentada, con el objetivo de profundizar en el conocimiento de la información disponible. Implica conocimientos conceptuales, metodológicos y estadísticos que permitan ir más allá de la comprensión literal, y poder explorar la "validez" de las conclusiones sacadas a partir de la información.

Descriptores, habilidades y conocimientos

En esta sección definiremos dos elementos; en primer lugar, el continuo sobre el que versarán los descriptores para las habilidades y conocimientos de cada componente, y en segundo, las habilidades y conocimientos incluidas dentro de cada componente. Ambos serán construidos a partir de las propuestas de Ward et al. (2010).

1. **Identificación de la información**: en términos de descriptores, estos abarcan desde una nula familiaridad con los conocimientos requeridos, hasta una amplia comprensión de los mismos, esto es, el usuario puede asociarlos entre sí de manera coherente y con otros conocimientos pertinentes. En términos de conocimientos requeridos, estos son una selección de las características de los usuarios, definidas por Ward et al. (2010). No se tomaron todas las características, ya que algunas se traslapan con elementos relativos a la comprensión de la información, y en pos del orden se prefirió mantenerlos separados.
    a. **Familiaridad con el dominio**: guarda relación con la experticia y conocimiento que el usuario posee respecto del dominio o contexto en que opera la información recibida.
    b. **Familiaridad con los datos**: se vincula con la experiencia que tenga el usuario con los datos a partir de los cuales se obtuvo la información, y si ha tenido la oportunidad de generar algún modelo mental para darles coherencia.
    c. **Familiaridad con el entorno de visualización**: conocimiento y experticia que tenga el usuario con la herramienta(s) de visualización con las que se esté presentando la información.
2. **Comprensión de la información**: En caso de la comprensión de la información, los descriptores se construirán a partir de las tareas descritas por Ward et al. (2010). Para efectos de rúbrica, estos se dividirán en tres grupos, el primero incluye la capacidad de identificar, localizar y distinguir información. El segundo, de categorizar, agrupar y priorizar. El último, de comparar, asociar y correlacionar.
    a. **Visualización básica**: son el tipo más común; por ejemplo, tablas de contingencia, gráficos de barra, torta, entre otros.
    b. **Visualización multidimensional**: similares a las visualizaciones básicas, pero incluyen más de tres dimensiones. Por este motivo, con el fin de asegurar su comprensión, requieren de ciertas consideraciones especiales al momento de ser diseñadas.
    c. **Visualización geoespacial**: son visualizaciones que hacen referencia a territorios. Requieren tanto de conocimiento de mapas como de los territorios que pretenden representar.

3. **Interpretación de la información**: al igual que con la identificación de la información, los descriptores en este caso abarcan desde un conocimiento familiar o nominal de los conceptos, a un manejo que permita manipular y evaluar críticamente –y a la luz de información externa a la visualización- la información contenida en la visualización.
   a. **Conceptos**: son todas aquellas nociones que el usuario debe manejar, en algún grado de experticia, para tener una correcta aproximación a la información presentada.
   b. **Estadística**: refiere a nociones básicas de estadística descriptiva, tales como medidas de tendencia central, medidas de posición y medidas de dispersión, que el usuario debe manejar para ampliar la comprensión de la información, así como para mesurar su validez.

Rúbrica

A continuación, se presenta la rúbrica para la dimensión **identificación de la información**. En **gris** se resalta el nivel esperado por el usuario final, en términos de habilidades o conocimientos.

Tabla B1: Rúbrica para la dimensión identificación de la información

Habilidad o conocimiento	Descriptores			
	Nulo	**Básico**	**Medio**	**Avanzado**
Familiaridad con el dominio	Las nociones que posee no alcanzan el umbral mínimo para comenzar a comprender el dominio de análisis.	Posee un conocimiento general y poco profundo del dominio. No conoce en detalle aspectos funcionales y estructurales que permitan un acercamiento crítico a la información de la visualización.	Posee un conocimiento suficiente del dominio. Conoce aspectos funcionales y estructurales en un detalle suficiente como para comprender la información de las visualizaciones.	Posee un conocimiento experto del dominio. Conoce en amplio detalle distintos aspectos funcionales y estructurales, lo cual le permite un acercamiento crítico a la información presentada en las distintas visualizaciones.
Familiaridad con los datos	No tiene conocimientos relativos a los datos, no conoce su origen, ni sus contextos válidos de aplicación.	Conocimiento mínimo de las fuentes de los datos, de sus contextos de aplicación y sus implicancias. No logra evaluar su validez o pertinencia en las visualizaciones.	Está familiarizado con algunos datos, así como fuentes que los producen. Conoce los objetivos por los cuales fueron creados los datos, sus contextos de aplicación e implicancias en el sistema. Puede evaluar la pertinencia de algunos de ellos dentro de las visualizaciones.	Posee un conocimiento experto de distintos tipos de datos, así como las fuentes que los producen. Conoce y comprende los objetivos por los que fueron creados los datos, sus contextos de aplicación y las implicancias que tienen para el sistema educativo. Es capaz de evaluar críticamente su pertinencia dentro de las visualizaciones.
Familiaridad con el entorno de visualización	No conoce el entorno de visualización	Conoce entornos similares al utilizado, lo que permite una manipulación mínima	Conoce el entorno de visualización, pero tiene dificultades para	Está familiarizado con el entorno de visualización utilizado, pudiendo

	utilizado ni entornos similares.	de información presentada.	manipular la información que en él se presentan.	manipular la información de forma efectiva.

A continuación, se presenta la rúbrica para la dimensión **comprensión de la información.** En **gris** se resalta en nivel esperado por el usuario final, en términos de habilidades o conocimientos.

Tabla B2: Rúbrica para la dimensión comprensión de la información

Habilidad o conocimiento	Descriptores			
	Nulo	**Básico**	**Medio**	**Avanzado**
Visualización básica	No logra distinguir las diferencias entre distintas visualizaciones básicas, ni con otras formas de visualizar la información.	Es capaz de identificar algunas visualizaciones básicas. De la misma forma es capaz de identificar y localizar los elementos relevantes dentro de cada una de ellas, pudiendo hacer distinciones entre dichos elementos.	Puede categorizar los elementos de una visualización básica en distintos tipos, organizarlos de acuerdo a un criterio común y/u ordenarlos de acuerdo a algún criterio. De la misma forma, es capaz de discernir cuando es apropiado utilizar alguno de estos procedimientos.	Posee un conocimiento avanzado, que le permite comparar, asociar y/o correlacionar distintos elementos de una visualización básica, así como realizar estos procedimientos entre visualizaciones de características similares.
Visualización multidimensional	No logra distinguir las diferencias entre distintas visualizaciones multidimensionales, ni con otras formas de visualizar la información.	Es capaz de identificar algunas visualizaciones multidimensionales, pudiendo distinguirlas de las localizaciones básicas. De la misma forma es capaz de identificar y localizar los elementos relevantes dentro de cada una de ellas, pudiendo hacer distinciones entre dichos elementos.	Puede categorizar los elementos de una visualización multidimensional en distintos tipos, organizarlos de acuerdo a un criterio común y/u ordenarlos de acuerdo a algún criterio. De la misma forma, es capaz de discernir cuando es apropiado utilizar alguno de estos procedimientos.	Posee un conocimiento avanzado, que le permite comparar, asociar y/o correlacionar distintos elementos de una visualización multidimensional, así como realizar estos procedimientos entre visualizaciones de características similares. Es capaz de proponer ajustes a las visualizaciones en función de necesidades o requerimientos para la toma de decisiones.
Visualización geoespacial	No logra distinguir las diferencias entre distintas visualizaciones geoespaciales, ni con otras formas de	Es capaz de identificar algunas visualizaciones geoespaciales y sus elementos (tales como las “manzanas”). De la	Puede categorizar los elementos de una visualización geoespacial en distintos tipos, organizarlos de acuerdo a un criterio	Posee un conocimiento avanzado (por ejemplo, de aplicación de modelos matemáticos

Habilidad o conocimiento	Descriptores			
	Nulo	Básico	Medio	Avanzado
	visualizar la información.	misma forma es capaz de identificar y localizar los elementos relevantes dentro de cada una de ellas, pudiendo hacer distinciones entre dichos elementos.	común y/u ordenarlos de acuerdo a algún criterio. De la misma forma, es capaz de comprender visualmente indicadores geográficos, además de comprender cuándo es apropiado utilizar estos procedimientos de visualización.	geoespaciales), que le permite comparar, asociar y/o correlacionar distintos elementos de una visualización geoespacial, así como relacionar esta visualización con otras. Es capaz de proponer ajustes a las visualizaciones en función de necesidades o requerimientos para la toma de decisiones.

A continuación, se incluye la rúbrica para la dimensión interpretación de la información. En **gris** se resalta el nivel esperado por el usuario final, en términos de habilidades o conocimientos.

Tabla B3: Rúbrica para la dimensión interpretación de la información

Habilidad o conocimiento	Descriptores			
	Nulo	Básico	Medio	Avanzado
Conceptos	No maneja los conceptos mínimos para dar sentido a la información expuesta en las visualizaciones.	Posee un conocimiento conceptual mínimo, que le permiten comprender la información de las visualizaciones.	Posee un conocimiento conceptual suficiente, que complementa su comprensión de las visualizaciones.	Posee un conocimiento de temas no necesariamente vinculados, pero que le permita enriquecer su comprensión desde áreas distintas de su especialidad.
Estadística	No maneja nociones mínimas de estadística descriptiva, por lo que no puede interpretar la información de las visualizaciones.	Maneja nociones muy básicas, como el promedio, que no son suficientes para comprender la información de las visualizaciones.	Maneja nociones de estadística descriptiva, tales como medidas de tendencia central, de dispersión y posición, como también nociones básicas de probabilidad, lo que le permite comprender y enriquecer la información de las visualizaciones.	Manejas nociones y técnicas estadísticas avanzadas y de minería de datos, lo que le permite no sólo comprender la información, sino evaluarla críticamente desde una perspectiva estadística.