mdt.

Ex Datis, Claritas

# ¿Qué son los datos sintéticos?

Qué son los datos sintéticos y por qué permiten compartir datos con terceros sin exponer los originales, frente a la anonimización y otros métodos.

Manuel Delgado Tenorio · 26 de diciembre de 2020

Publicado originalmente en www.manueldelgado.com/blog/que-son-los-datos-sinteticos



**Los datos sintéticos (en inglés, "*Synthetic Data*") son datos creados artificialmente que reproducen las propiedades estadísticas de un conjunto de datos real sin contener ninguno de sus registros.**

Uno de sus principales usos es, partiendo de un conjunto de datos que queremos proteger pero que necesitamos compartir con terceros, generar un nuevo conjunto de datos que conserva las características informacionales del conjunto origen pero que no permite recomponer los datos originales a partir de los creados artificialmente.

Los datos sintéticos, usados de esta forma, suponen una evolución de los métodos más comunes para compartir datos con terceros, cada uno de los cuales tiene sus desventajas particulares:

- **La anonimización y pseudonimización,** que consisten en eliminar o enmascarar las partes más sensibles de un conjunto de datos (como mínimo, los datos personales identificables). Hoy en día, está comúnmente aceptado que no son prácticas robustas para compartir datos, ya que es relativamente sencillo inferir los datos eliminados, fundamentalmente mediante el cruce con otros conjuntos de datos.
- **La agregación de datos:** en lugar de compartir los datos detallados, se crean datos "resumidos". Por ejemplo, un negocio retail que necesite compartir sus datos de ventas puede compartir las ventas agregadas por día y categoría de producto, en lugar de compartir un listado detallado de las compras de sus clientes. Si bien este enfoque puede ser válido para algunos usos concretos, no lo es para muchos otros, porque a medida que aumentamos la robustez de los datos compartidos mediante un mayor nivel de agregación, se reduce también su valor informacional (en el ejemplo anterior, perderíamos la posibilidad de analizar el comportamiento de nuestros distintos segmentos de cliente).

Existen otros mecanismos, más avanzados, que aseguran que las actividades de compartición de datos mantienen la privacidad de los usuarios a los que hacen referencia esos datos. Por ejemplo, ya he hablado por aquí de la privacidad diferencial.

Sin embargo, enfoques como la privacidad diferencial no son de aplicación en todos los casos. Es bueno, por tanto, que vayan surgiendo otros métodos que complementen nuestro abanico de posibilidades y, así, podamos elegir el más adecuado a nuestro caso actual. Además, es posible (de hecho, es ideal) combinar ambos métodos: es posible generar datos sintéticos que cumplan con la definición de privacidad diferencial.

## Qué problema resuelven los datos sintéticos

Mediante la generación de datos sintéticos, una organización puede hacer frente de manera más ágil a uno de sus principales retos de datos: compartir información con terceros para extraer valor de sus datos, pero siempre **dentro de un marco de protección de la privacidad y de sus secretos de negocio.**

Son innumerables los distintos casos en los que una organización puede necesitar compartir datos. Destacan las ocasiones en las que un tercero puede ayudarnos mediante el análisis avanzado de esos datos, ya sea porque cuenta con un *expertise* específico o, simplemente, porque se subcontratan esas actividades.

Este tipo de prácticas de compartición de datos suelen verse frenadas por tres motivos que están íntimamente relacionados entre sí:

- El temor de que al compartir **datos insuficientemente anonimizados** con terceros estemos incurriendo en un riesgo relacionado con la privacidad de nuestros clientes o empleados.
- La duda sobre si, al proporcionar **datos que han sido agregados o despojados de información sensible**, esos datos siguen representando la realidad como es debido y, por tanto, si su análisis va a ser certero y relevante.
- La **lentitud y complejidad** para poner en marcha cada proceso de entrega de datos a terceros, que resta agilidad a muchas iniciativas o incluso las vuelve inviables. Estos frenos son tanto procedimentales (procesos de aprobación eternos) como técnicos (sistemas de los que cuesta semanas o meses sacar los datos que necesitas).

La necesidad de este tipo de técnicas es más evidente cuando se comparten datos con terceros. Sin embargo, **también resultan de utilidad cuando los datos se van a tratar internamente:** por mucho control que se intente tener sobre los datos, tras un cierto tiempo es inevitable que haya multitud de ficheros descontrolados y desperdigados por ordenadores, servidores, y hasta servicios cloud de terceros.

## Cómo se generan los datos sintéticos

*Grosso modo*, los datos sintéticos los genera un modelo de *machine learning* que ha sido previamente entrenado sobre los datos originales.

El entrenamiento del modelo tiene como objetivo identificar los patrones de actividad existentes en los datos, para poderlos reproducir. Es decir, el modelo ha de ser capaz de generar un nuevo conjunto de datos cuyas propiedades estadísticas sean iguales a las de los datos originales.

De hecho, el principal control de calidad que se debe realizar sobre un conjunto de datos sintéticos es comparar sus rasgos estadísticos con los del conjunto original: por ejemplo, las tablas de distribución de los valores de las variables más relevantes deben ser iguales o, también, dos *features* que eran codependientes en los datos de origen deben seguir siéndolo en los datos sintéticos. Queremos que los datos se comporten igual al ser analizados: si clusterizamos los datos originales y los sintéticos, los segmentos resultantes deberían ser idénticos, por ejemplo.

Este es un campo relativamente nuevo, así que el futuro nos traerá métodos más avanzados y eficaces de generar datos sintéticos. Sin embargo, ya hay implementaciones prácticas que alcanzan niveles de precisión elevados, como DoppelGANger, que usa redes generativas antagónicas (en inglés, Generative Adversarial Networks -GANs-). Aquí el paper que lo describe y aquí el código en GitHub, con datos de ejemplo.

## Una nota de precaución y otra de optimismo

No hay que subestimar la dificultad de implementar correctamente métodos avanzados y novedosos como este.

Particular atención debemos prestarle a las tecnologías desarrolladas por terceros que plasman estos conceptos. No debemos hacer un ejercicio de confianza ciega, sino poner en duda los resultados hasta estar seguros de que son correctos. La

pega es que, igual que ocurre con el concepto de privacidad diferencial, al generar datos sintéticos es complejo verificar que lo estamos haciendo correctamente. Requiere la participación de personas con una sólida base matemática y buena comprensión de estos métodos.

La sensación de seguridad de usar un método así de avanzado tiene que estar sustentada por una sólida comprensión y confianza en el proceso que se ha seguido, no por el *hype*.

> *No obstante, el valor que se puede lograr al adoptar un método de generación de datos sintéticos es tan grande que la dificultad de la tarea merece la pena.*

Cuando hablamos de generación de datos sintéticos, hablamos de lograr que los científicos de datos puedan empezar a trabajar casi de inmediato, sin burocracias y sin esperas de semanas o meses, con unos datos que representan fielmente la realidad pero que, a su vez, no suponen riesgo para la privacidad o el negocio.

Llegar a ese punto puede cambiar radicalmente la capacidad para aprovechar los datos en la mayoría de organizaciones actuales.

**PARA SEGUIR LEYENDO**

He vuelto a escribir sobre el tema: por qué la industria de la IA ve en los datos sintéticos una salida a la escasez de datos de entrenamiento y por qué no sirven para sustituir las encuestas a clientes reales. Todo, en privacidad y datos sintéticos.




**Sobre el autor**

## Manuel Delgado Tenorio

Manuel Delgado Tenorio es asesor independiente de estrategia de datos e inteligencia artificial, profesor e investigador. Cofundó y vendió una *startup* de inteligencia artificial, fue socio de la práctica de Data & Analytics para servicios financieros en EY y responsable de Data Analytics para el sur de Europa en AWS. Es profesor asociado en IE University, donde imparte la primera asignatura de grado sobre Customer Lifetime Value en España, investiga sobre el CLV en su doctorado en la Universidad Complutense de Madrid y es autor del libro "Customer Lifetime Value (CLV): No todos tus clientes valen lo mismo".

**manueldelgado.com · LinkedIn · X**

Recibe cada artículo en tu correo con la newsletter Ex Datis, Claritas: manueldelgado.com/newsletter/suscribete