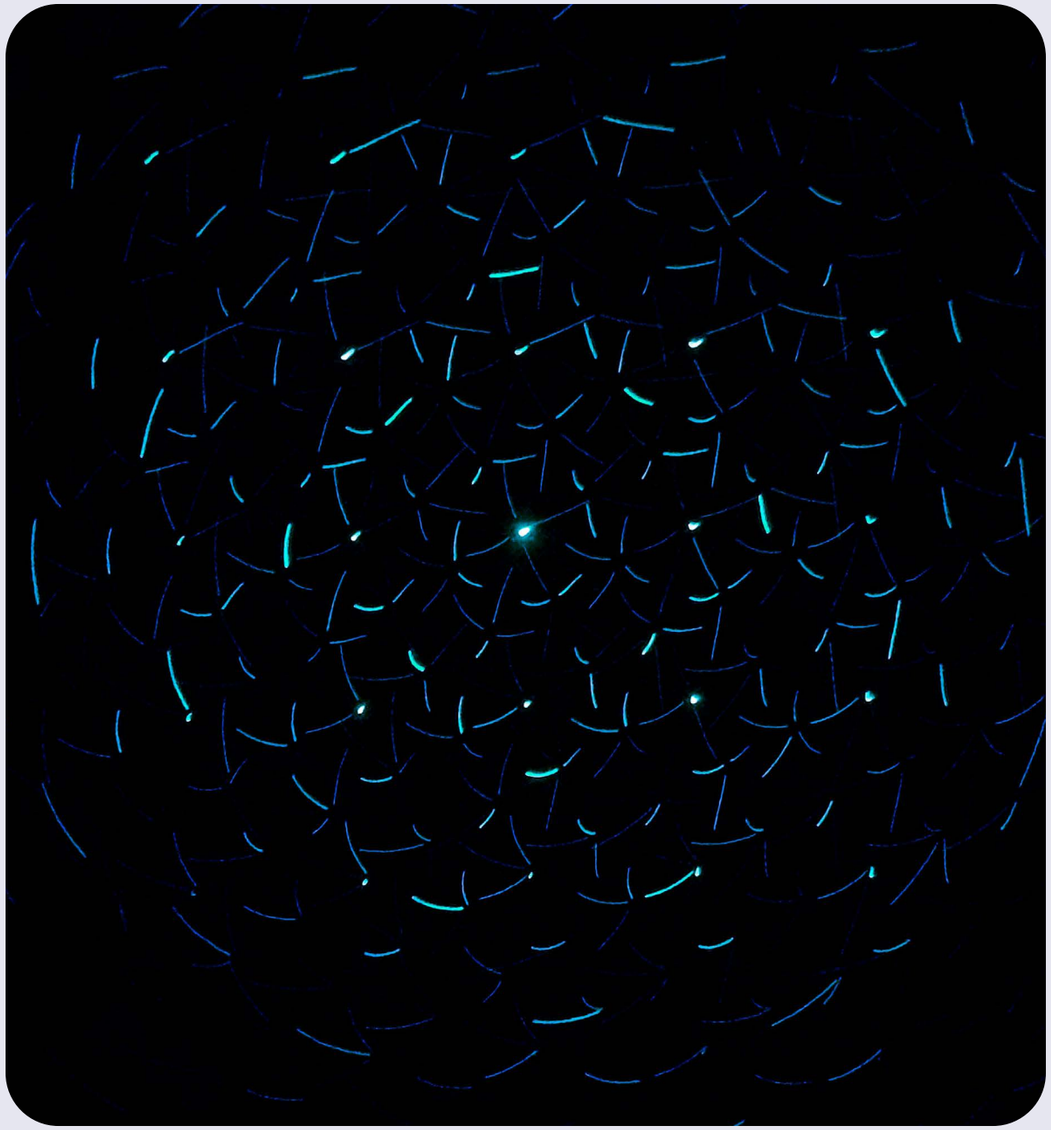
Softtek®

The rise of Synthetic Data: data without borders

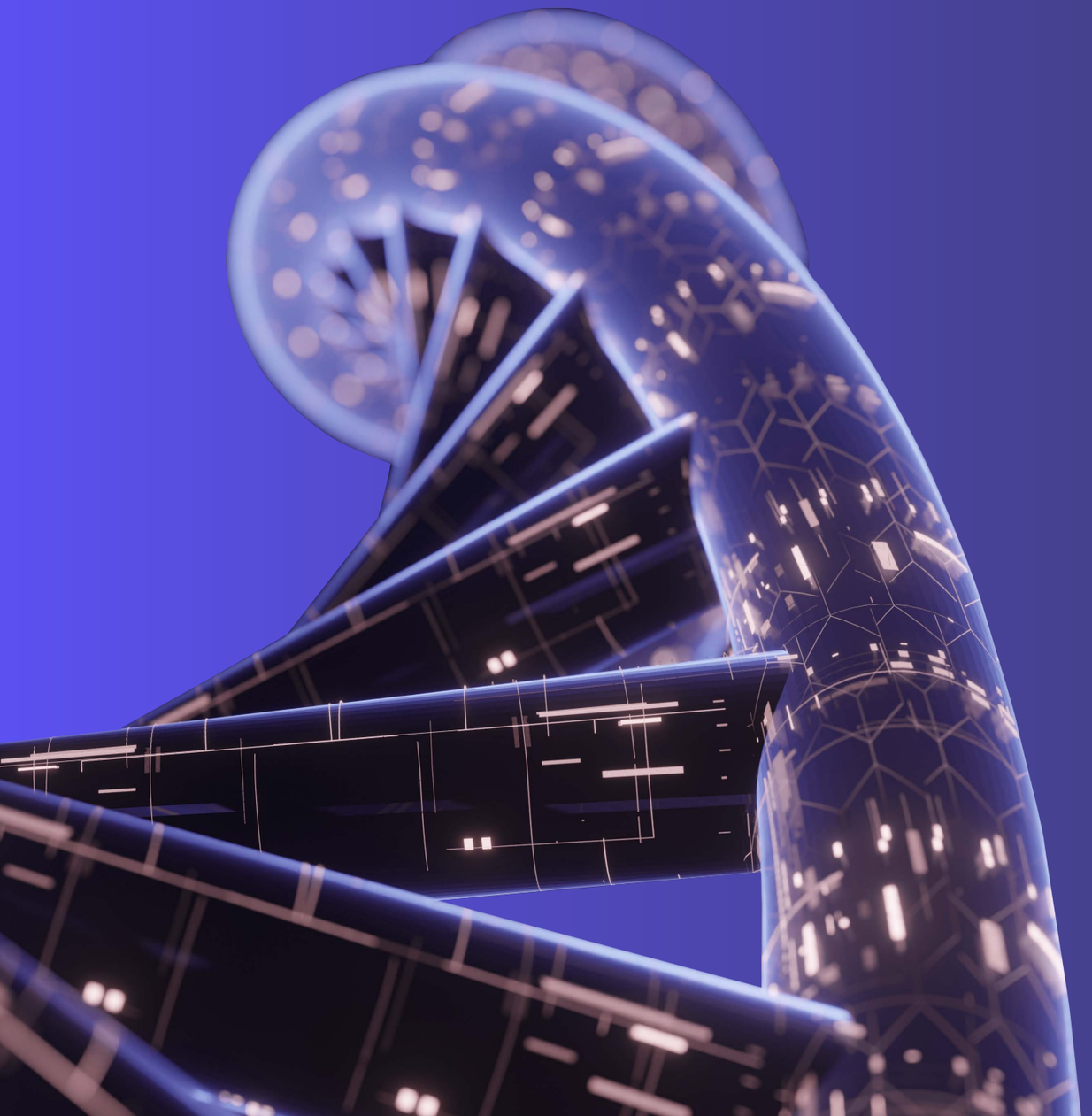
Los datos sintéticos son la clave para impulsar la innovación y garantizar la privacidad en la nueva era



Los datos sintéticos han emergido como una solución transformadora en la industria tecnológica, especialmente en un mundo donde el acceso a datos reales está cada vez más limitado y protegido por regulaciones, leyes y preocupaciones de privacidad. Esta tecnología permite generar datos artificiales que imitan con precisión las propiedades estadísticas de los datos reales, a través de técnicas avanzadas como el *deep learning* y la IA generativa. Desde sus inicios en los años 30, con experimentos de síntesis de audio, hasta su aplicación moderna en la creación de conjuntos de datos para entrenar modelos de IA en sectores como la sanidad, las finanzas y los vehículos autónomos o el entretenimiento, los datos sintéticos han evolucionado para convertirse en un pilar clave de la innovación. Según Gartner, **se espera que para 2026, el 75% de las empresas empleen IA generativa para crear datos sintéticos de clientes.**

Este avance tiene implicaciones directas en el desarrollo de software, la evolución de la nueva era del código y la analítica avanzada. Los datos sintéticos **permiten superar obstáculos como la escasez de datos, los altos costes de recolección y las barreras de privacidad**, permitiendo que las empresas creen modelos de IA más robustos, inclusivos y éticos. A medida que las tecnologías de generación de datos continúan mejorando, estos datos no solo complementan a los reales, sino que también posibilitan la creación de sistemas de IA más precisos y eficientes.

Gracias a ellos, **las empresas pueden alimentar sus algoritmos con océanos de datos sin naufragar en el mar de las restricciones legales**, conciliando así el hambre de datos de la IA con las expectativas de clientes, reguladores y sociedad en cuanto a privacidad.



Un nuevo paradigma: privacidad e innovación en tiempos de datos artificiales

Los datos se consolidan como activos clave, mientras su uso enfrenta regulaciones más estrictas y protocolos cada vez más rigurosos

Los datos evolucionan como motor clave de transformación empresarial

El mercado de datos ha experimentado una **transformación notable en las últimas décadas**, impulsada por avances tecnológicos como el *Cloud Computing* y los microservicios o la implementación de la inteligencia artificial en muchos de sus ámbitos. En el pasado, la recopilación y el uso de datos dependían en gran medida de las infraestructuras físicas, como las líneas arrendadas para la transmisión de datos entre proveedores y consumidores.

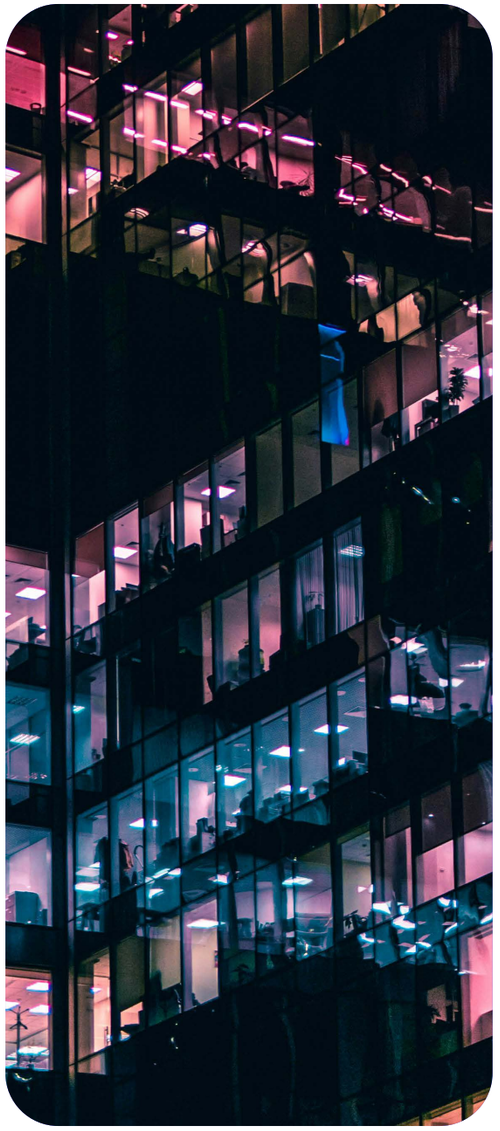
El valor estratégico de los datos se ha vuelto más evidente gracias a que las organizaciones han reconocido que no son simplemente un subproducto de sus operaciones, sino un recurso esencial para la toma de decisiones, la innovación y el diseño de nuevos modelos de negocio. Hoy, **los datos se consideran un activo corporativo** que puede generar valor directo o habilitar soluciones más sofisticadas y personalizadas.

Grandes corporaciones han demostrado que los datos **pueden ser el núcleo de su modelo de ingresos, monetizando la información de los usuarios a través de servicios como la publicidad dirigida o las recomendaciones personalizadas.** Además de ser una posible fuente de ingresos, los datos impulsan la innovación mediante el desarrollo de productos personalizados y el uso de inteligencia artificial y aprendizaje automático.

Fuente: Synthesis

180 zettabytes

es el volumen de datos generados, consumidos, copiados y almacenados que se proyecta superar para 2025



Pero la escasez y exclusividad de datos limita innovación y colaboración empresarial

Pero aun con esta concienciación sobre la importancia de los datos, **las empresas hoy enfrentan una creciente escasez de datos reales adecuados para entrenar modelos de IA.**

Obtener estos datos no solo es complicado y costoso, también implica procesos largos y complejos como el uso de encuestas, experimentos controlados o la compra de licencias para acceder a bases de datos privadas, que requieren inversiones significativas. A estos costes se suman los gastos en equipos especializados para recolectar y procesar la información.

Incluso cuando las organizaciones logran superar estas barreras económicas, **la calidad de los datos sigue siendo un problema crítico.** Muchos conjuntos de datos contienen sesgos o están incompletos, lo que afecta directamente la precisión y efectividad de los modelos.

Y, además, **el panorama actual se caracteriza por un ambiente competitivo** en el que las organizaciones se vuelven cada vez más protectoras con su información, dificultando la colaboración y el intercambio de datos, lo que obstaculiza el desarrollo de modelos colaborativos e innovadores. La exclusividad de los datos se ha convertido en un factor limitante, cerrando el acceso a recursos clave para muchas empresas, especialmente a las más pequeñas que no tienen acceso a bases de datos masivas.

La evolución hacia los datos sintéticos refleja una respuesta estratégica a las crecientes demandas de privacidad y al cumplimiento normativo

A este entorno, se suman las normativas de privacidad, como el GDPR en Europa y la CCPA en California, **que imponen restricciones cada vez más estrictas sobre el uso de datos personales.** Estas regulaciones exigen transparencia, el consentimiento informado del usuario, y el derecho de acceso y eliminación de datos personales, transformando la gestión de la información en las empresas.

Las legislaciones están llevando a las organizaciones a repensar sus prácticas de recopilación y uso de datos, obligándolas a encontrar alternativas que respeten tanto la privacidad de los usuarios como la necesidad de innovar.

De hecho, el **85% de los consumidores son plenamente conscientes del valor y la importancia de sus datos** y, es por ello, que la protección de la información de los clientes es una prioridad para las organizaciones, con un 96% de ellas reconociendo su relevancia mucho más allá de lo que exigen las normativas regulatorias.

Fuentes: Martech I Nymiz I Synthesis

Cómo la evolución regulatoria impulsa la innovación en protección de la privacidad

Esta evolución regulatoria también impulsa la innovación en tecnologías de protección de la privacidad, **o *Privacy-Enhancing Technologies* (PET).** Estas tecnologías permiten **extraer el valor de los datos sensibles sin poner en riesgo la privacidad de los individuos**. Métodos como la anonimización avanzada, aseguran que los datos se protejan durante su procesamiento, permitiendo su uso para análisis colaborativos y respetando al mismo tiempo las estrictas normativas de privacidad.

La anonimización elimina identificadores directos como nombres, direcciones o números de identificación. Sin embargo, con el tiempo se ha demostrado que, aunque útil, **no siempre es suficiente.** En muchos casos, combinaciones de datos aparentemente anónimos pueden ser reidentificadas, especialmente cuando se cruzan con otras fuentes disponibles públicamente.

Ante estas limitaciones y las presiones regulatorias hiperprotectoras actuales, **las organizaciones han comenzado a adoptar enfoques más robustos, como la privacidad diferencial y, más recientemente, los datos sintéticos.** Estos son conjuntos de información generados de manera artificial a través de algoritmos, simulaciones o modelos de IA.

82%

de las compañías admiten ponerse en riesgo cuando recopilan datos reales

91%

afirman que dirigen más esfuerzos a tranquilizar a sus clientes sobre cómo se utilizan sus datos con IA

94%

Concuerdan en que sus clientes les abandonarán si sus datos no están bien protegidos

Frente a riesgos legales y de privacidad, los datos sintéticos emergen como la opción más robusta y escalable disponible

Tabla comparativa: datos reales vs. anonimizados vs. sintéticos

Criterio	Datos reales	Datos anonimizados	Datos sintéticos
Privacidad	**Baja** Contienen información identificable	Media se eliminan identificadores directos	Alta no contienen datos reales
Riesgo de reidentificación	**Alto**	**Medio** riesgo si se combinan con fuentes externas	**Muy bajo** si están bien generados
Utilidad analítica	**Muy alta**	**Alta** pero puede perder precisión	**Baja** Contienen información identificable
Cumplimiento normativo	**Baja** requiere justificación legal	**Medio** pero aún sujeto a restricciones	**Alto** si no es posible revertir a datos reales
Escalabilidad	**Limitada** por disponibilidad y regulaciones	**Limitada** por calidad y esfuerzo manual	**Alta** se pueden generar grandes volúmenes
Coste de obtención/uso	**Alto** requiere permisos, infraestructura segura	**Medio** requiere procesos manuales o semi-automáticos	**Bajo o medio** depende de la tecnología usada
Tiempo de acceso	**Bajo** revisiones legales y éticas	**Medio** necesita validación de anonimización	**Alto** se generan bajo demanda

El mercado de datos sintéticos crecerá rápidamente, impulsado por IA, aprendizaje automático, IoT y tecnologías conectadas emergentes

Crecimiento global de datos sintéticos: impulso por IA, regulaciones y expansión regional

De hecho, se proyecta que el **mercado global de generación de datos sintéticos llegará a $1.788,1 millones para 2030, con una tasa de crecimiento anual compuesta (CAGR)** del 35,3% entre 2024 y 2030, impulsado principalmente por la creciente adopción de tecnologías emergentes como la IA, el ML y el IoT, junto con un aumento en el uso de tecnologías de dispositivos conectados. De hecho, Forbes ya la ha nombrado una de las "*5 Biggest Data Science Trends in 2022*".

Aunque América del Norte domina el mercado desde 2023, Asia-Pacífico es la región con el crecimiento más rápido y escalado. En Norteamérica, los impulsores del crecimiento en este mercado se centran en varios factores.

Sus ecosistemas avanzados de investigación en inteligencia artificial, las fuertes regulaciones en privacidad y la adopción temprana por parte de sectores como el financiero, tecnológico y de servicios.

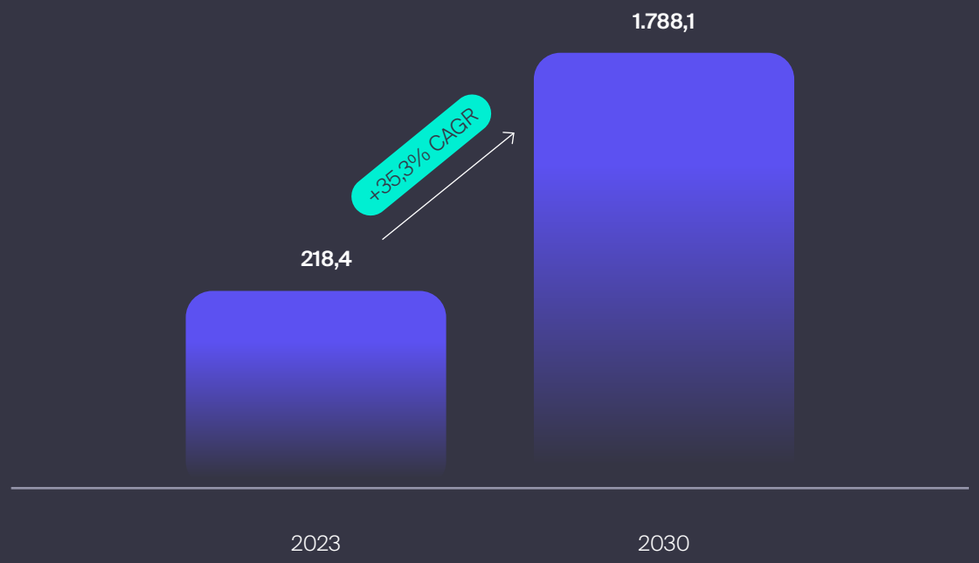
60%

de los **datos utilizados para la IA serán sintéticos a finales de 2024**, frente al 1 % en 2021, lo que destaca su papel clave en la simulación, el modelado y la reducción de riesgos.

Fuentes: Gartner I Grand View Research I Nature

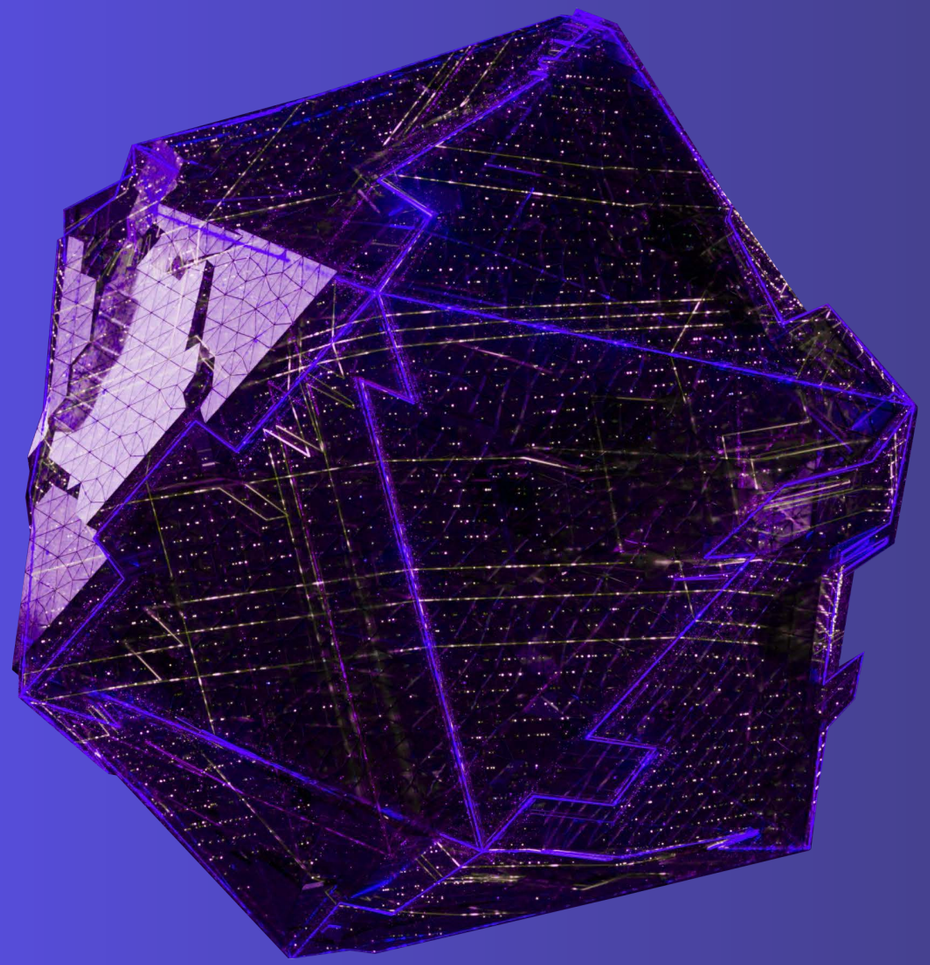
Tamaño del mercado global de generación de datos sintéticos

(en millones de $)



En Europa, **el cumplimiento con el Reglamento General de Protección de Datos es un impulsor fundamental,** además, sectores industriales clave como la automoción y la manufactura están impulsando la demanda de datos sintéticos. Más allá, la existencia de *sandboxes* regulatorios en algunos países europeos también facilita la experimentación con ellos.

En la región de Asia-Pacífico, se está invirtiendo en infraestructuras de inteligencia artificial y existe una rápida adopción de tecnologías digitales en países como China, India y Japón gracias a la expansión de la digitalización y el aumento de dispositivos conectados. Además, **el apoyo gubernamental a la innovación en IA está impulsando aún más la adopción de estas tecnologías en la región.**



Datos sintéticos: reinventando la realidad digital

Los datos sintéticos pueden generarse en diversas formas según el grado en que son artificiales y cómo se integran con datos reales

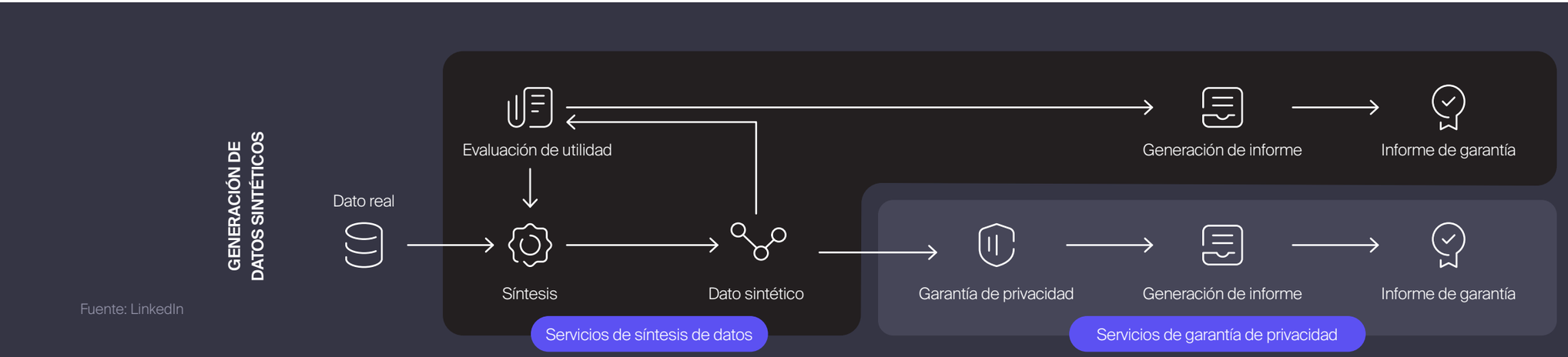
Más allá de la réplica real, una solución para la creación y validación de modelos

Los datos sintéticos no son simplemente copias de información real, sino que **son creaciones nuevas basadas en las distribuciones y relaciones estadísticas de los datos originales.** Actualmente, tienen el potencial de cubrir una amplia gama de aplicaciones, incluyendo pruebas, desarrollo de nuevos modelos, entrenamiento de algoritmos de *Machine Learning* y validación de modelos predictivos.

La generación de datos sintéticos **implica el uso de algoritmos y modelos estadísticos para producir datos que no han sido recolectados de fuentes del mundo real.** Estos datos pueden adoptar diversas formas. Las técnicas utilizadas para la creación de datos sintéticos se basan en el análisis de distribuciones estadísticas subyacentes, modelos de aprendizaje automático y aprendizaje profundo.

Sus distintas formas son:

- **Datos sintéticos parciales:** aquellos que **sustituyen solo una parte de un conjunto de datos real por información sintética.** Son útiles para proteger información sensible, como los nombres o detalles personales. Esta técnica preserva la privacidad de la información y ayuda a proteger los datos personales manteniendo las relevantes características de los datos reales.
- **Datos sintéticos híbridos:** este tipo de datos **combinan datos reales con datos completamente artificiales**, es decir, se toma un conjunto de datos real y se mezcla aleatoriamente con registros sintéticos. Este enfoque es muy útil para simular escenarios más completos sin utilizar datos sensibles directamente, lo que reduce el riesgo de reidentificación.
- **Datos sintéticos completos:** son aquellos que no **contienen datos reales en absoluto.** Se generan datos completamente nuevos que siguen las mismas relaciones y propiedades estadísticas de los datos reales, pero no están vinculados a personas ni eventos reales. Este enfoque es particularmente útil cuando no se dispone de datos suficientes para entrenar modelos de *Machine Learning* o para simulaciones de procesos.



Fuente: LinkedIn

Los datos sintéticos ofrecen mayor seguridad, eliminando riesgos de reidentificación que persisten en técnicas tradicionales

Anonimización: la solución del pasado; datos sintéticos: la apuesta del futuro

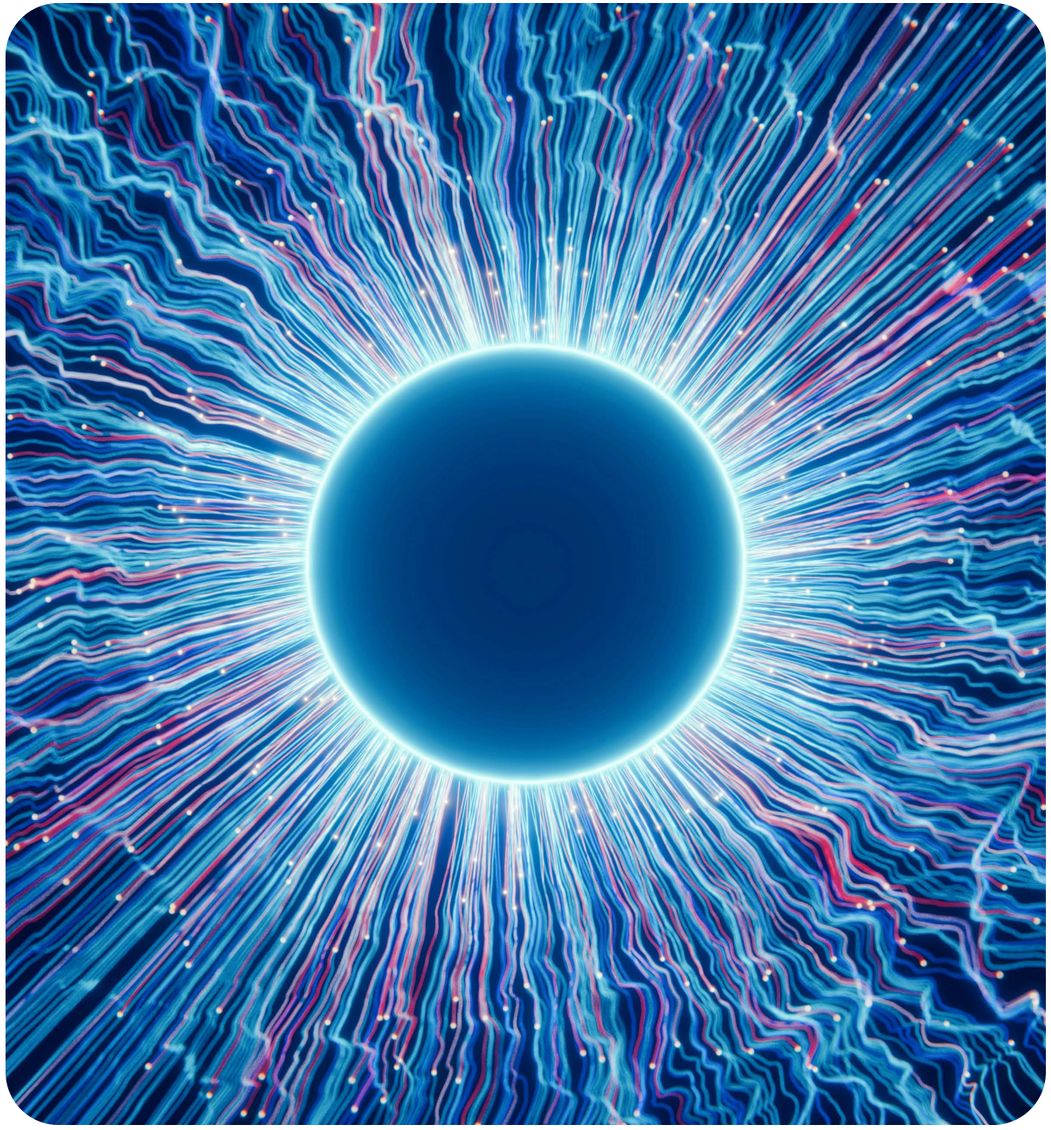
Tradicionalmente, **las técnicas de anonimización, seudonimización y enmascaramiento se han utilizado para proteger la información personal.** Sin embargo, con la creciente complejidad y demanda de IA, ML y otros análisis avanzados, estas técnicas presentan limitaciones evidentes, lo que ha dado paso a nuevas alternativas como los datos sintéticos.

80%

de **reducción en multas potenciales**, ya que los datos sintéticos disminuyen los riesgos de filtración frente a los datos reales

25%

más rápido en la creación de prototipos de IA, lo que brinda una ventaja competitiva en sectores como finanzas y salud



Fuentes: LinkedIn | Syntho | Geeks for Geeks | Better Data

Diferentes técnicas de protección de la privacidad

Anonimización

Es el proceso de **eliminar o transformar datos personales** para que no puedan vincularse directamente a una persona, haciendo que no sea posible identificarla ni revertir los datos a su estado original.

Desventaja

- Dado que las técnicas son estáticas, **pueden no ser suficientes frente a nuevos métodos de re-identificación.**
- En muchos casos, **reduce la granularidad de la información,** eliminando detalles esenciales que pueden ser cruciales para análisis avanzados.

Seudonimización

Es un proceso en el que los **identificadores personales directos son reemplazado**s por seudónimos. Aunque es una mejora respecto a la anonimización, no elimina por completo el riesgo de reidentificación, ya que los seudónimos pueden ser revertidos.

- **Su misma naturaleza reversible la hace vulnerable** y un riesgo significativo en términos de privacidad.
- Si se puede acceder a las claves o si se combinan los datos con otras bases de datos, **es posible reidentificar** a los individuos, por lo que debe ser especialmente controlada.

Enmascaramiento

Este proceso **altera los valores de los datos originales para que estos no sean reconocibles.** Por ejemplo, nombres y direcciones pueden ser sustituidos por valores ficticios, mientras que la estructura y el formato de los datos se mantienen intactos.

- Si se tiene acceso a las claves , a su forma original. **Los datos pueden ser restaurados resultando** en vulnerabilidades si los mecanismos de protección no son robustos.
- Aunque se mantiene la estructura, **la calidad de los datos enmascarados a menudo se ve comprometida**, lo que puede limitar su utilidad en modelos de IA y análisis predictivos.

Datos sintéticos

Esta técnica **genera artificialmente datos mediante algoritmos de IA que imitan las características y patrones de los datos reales**, pero sin contener ninguna información personal o identificable. A diferencia de otras técnicas, no dependen de datos originales y eliminan el riesgo de reidentificación.

- Generar datos de alta calidad puede ser **un proceso computacionalmente costoso**, especialmente si se usan modelos avanzados.
- Su creación **requiere el uso de modelos de IA avanzados y experiencia** en la gestión de datos.

Cada organización debe evaluar su caso, aunque los datos sintéticos se posicionan como solución superior en privacidad y utilidad

Métodos de anonimización

Métodos	Dato real	Dato anonimizado
Enmascaramiento	EFGH3456	EFGH**** | ****3456
Seudonimización	David	a0001
Generalización	Edad: 24	20-25 | 20-30
Intercambio de datos	David Lorenzo John Carlos	David Carlos John Lorenzo

Datos sintéticos

	Nombre	Correo	Número de ID	Edad
Dato Real	David	david@gmail.com	1234	36
Datos totalmente sintéticos	James	james@gmail.com	8456	36
Datos sintéticos parciales	David	David@outlook.com	4312	36

Aunque las técnicas de anonimización, seudonimización y enmascaramiento han sido útiles en el pasado, **las limitaciones en términos de utilidad de los datos y riesgo de reidentificación han llevado a la adopción de datos sintéticos** como la solución más robusta.

Ofrecen mayor privacidad, mantienen la utilidad de los datos y son fácilmente escalables, lo que los convierte en la opción preferida en aplicaciones modernas, especialmente cuando se manejan grandes volúmenes de datos o situaciones raras que los métodos tradicionales no pueden cubrir de manera efectiva.

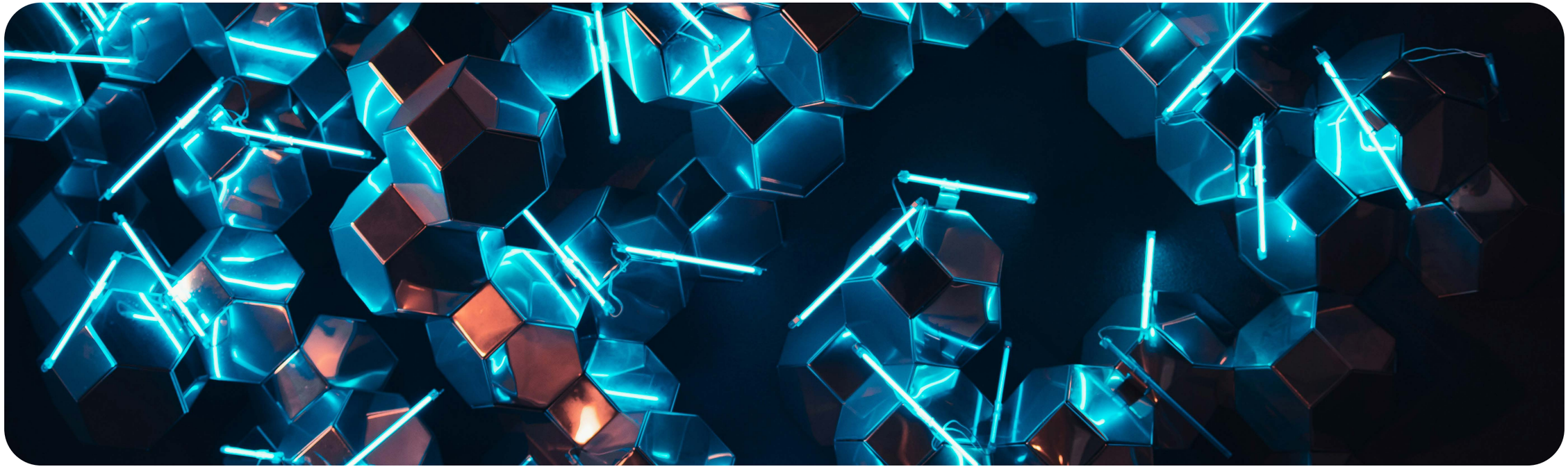
Aun así, **la elección entre estos métodos deberá depender de las necesidades específicas de cada organización,** el tipo de datos con los que trabaja y las regulaciones de privacidad que debe cumplir. Sin embargo, los datos sintéticos se están consolidando como la mejor alternativa para proteger la privacidad sin sacrificar la utilidad de los datos, convirtiéndolos en una herramienta indispensable en el mundo moderno de *Big Data* e inteligencia artificial.

Gestionar datos implica *trade-offs* constantes, pero los datos sintéticos permiten avanzar sin comprometer privacidad, utilidad ni costes

Trade-offs entre privacidad, utilidad y coste

Cualquier decisión sobre datos conlleva un compromiso entre maximizar la privacidad y maximizar la utilidad, además del factor coste o esfuerzo involucrado.

Los datos sintéticos se convierten una solución que equilibra estos tres ejes de manera novedosa, pero no son una bala de plata absoluta.



Privacidad vs. utilidad

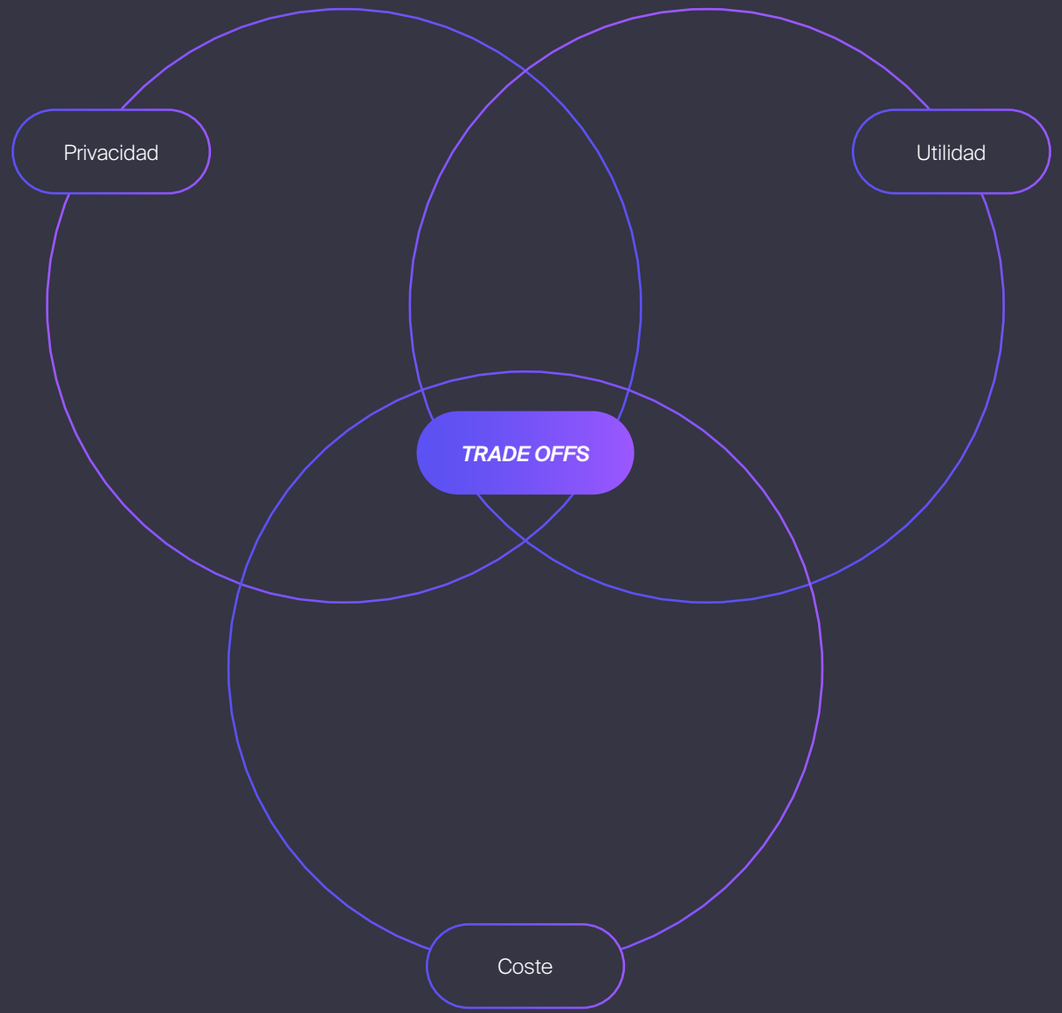
En los métodos tradicionales este *trade-off* era severo ya que eliminar o enmascarar campos reduce la información explotable y agregar mucho ruido para anonimizar arruina análisis finos. Con los datos sintéticos, el objetivo es lograr una alta privacidad con alta utilidad, pero para garantizar privacidad a veces se debe sacrificar algo de precisión, por ejemplo, introduciendo ligeras inexactitudes para evitar coincidencias con datos reales. El truco está en calibrar ese balance, hay técnicas como privacidad diferencial que permiten ajustar un parámetro *epsilon* (un *epsilon* más pequeño da más privacidad, pero más distorsión, y viceversa). Además, los **datos sintéticos pueden añadir utilidad en ciertos casos al generar más volumen de datos para algoritmos (mejoran generalización) o eliminar errores de datos reales** (los datos reales tienen ruido de medición y erratas que los datos sintéticos pueden obviar). Por tanto, el *trade-off* privacidad/utilidad con sintéticos se ha estrechado notablemente en comparación con métodos previos.

Utilidad vs. coste

Tradicionalmente, recolectar muchos datos reales de calidad es costoso (en tiempo, dinero y esfuerzo humano). Además, limpiarlos y etiquetarlos encarece los proyectos de IA. **Los datos sintéticos pueden reducir estos costes**, es más, está confirmado que generar datos sintéticos suele ser más barato que recopilar y etiquetar datos reales y, además, es potencialmente escalable a bajo coste. Por lo general, los sintéticos tienden a mejorar la utilidad disponible dentro de un presupuesto dado.

Privacidad vs. coste

En mecanismos clásicos, añadir privacidad solía implicar coste extra (consultores para anonimizar, tiempo en preparar datos *safe*). Con datos sintéticos, una vez que el proceso está implementado, **la privacidad viene integrada en los datos generados.** Esto puede simplificar flujos de trabajo y ahorrar costes operativos en *compliance*. Desde la perspectiva regulatoria, empresas señalan que proyectos que antes requerían complejos acuerdos de uso de datos (costosos en tiempo legal) ahora con sintéticos se aceleran, reduciendo coste de oportunidad.



Fuentes: LinkedIn I Aetion

Además, lograr el equilibrio entre fidelidad y generalización en datos sintéticos es clave para evitar reidentificación o pérdida de valor

El dilema de la fidelidad: ¿falsos positivos o generalización útil?

Uno de los mayores desafíos es **encontrar el equilibrio adecuado entre fidelidad y generalización.** Este dilema implica decidir hasta qué punto los datos generados deben replicar de manera fiel los patrones y relaciones presentes en los datos reales, ya que un nivel demasiado alto de fidelidad puede provocar la introducción de falsos positivos o riesgos de reidentificación, mientras que una fidelidad demasiado baja puede llevar a una generalización excesiva, perdiendo información crítica y útil para los modelos.

Fidelidad alta:
riesgo de falsos positivos

Cuando los datos sintéticos buscan replicar casi idénticamente los datos reales, **existe el riesgo de que se reproduzcan patrones irrelevantes o ruido estadístico presente en los datos originales**, conocido como el sobreajuste a la realidad. Los modelos entrenados con estos datos sintéticos de alta fidelidad pueden aprender correlaciones espurias o patrones que no tienen una base causal real.

Si los datos sintéticos replican demasiado la realidad, **podrían arriesgar la privacidad de los individuos al memorizar información confidencial (como *outliers*).** Además, una fidelidad excesiva limita la capacidad de generar datos diversos que puedan hacer que el modelo sea más robusto a situaciones no vistas.

Fidelidad baja:
generación útil, pero a qué precio

Por otro lado, una fidelidad demasiado baja implica que los datos sintéticos se generalicen excesivamente. En este caso, se evita el sobreajuste, pero a costa de perder detalles importantes que podrían ser cruciales para un modelo robusto. Un modelo generativo que suaviza demasiado las distribuciones de los datos **podría no capturar comportamientos críticos de ciertos subgrupos o patrones de interés.**

En términos estadísticos, esto se refiere al clásico dilema de sesgo versus varianza: **demasiada suavización introduce sesgo, mientras que no suficiente suavización puede llevar a una alta varianza.** La clave está en encontrar un equilibrio donde los datos sintéticos no omitan señales útiles.

Tecnologías clave como GAN, *transformers* y simuladores impulsan los datos sintéticos, equilibrando realismo, control y aplicabilidad práctica

Los modelos generativos pueden dividirse en explícitos e implícitos

Los modelos explícitos son aquellos que modelan directamente y de manera transparente la distribución subyacente de los datos. Esto significa que puedes observar y entender cómo se generan los resultados, ya que el modelo utiliza funciones matemáticas bien definidas para crear los datos. Son fáciles de interpretar y aplicar cuando se requiere una alta precisión y trazabilidad en los datos generados. Estos modelos están más enfocados en replicar datos con una alta fidelidad a las distribuciones estadísticas originales.

Los implícitos en cambio, no modelan directamente la distribución de los datos. En lugar de eso, aprenden a aproximar esta distribución a través del entrenamiento con datos. Estos modelos son más poderosos en términos de capacidad de generar datos realistas, pero son más difíciles de interpretar y controlar, ya que no existe un mapeo explícito entre la distribución de datos y los resultados generados.

Modelos explícitos

Autocodificadores Variacionales (VAE): aprenden representaciones latentes de los datos a través de un proceso de compresión y reconstrucción. Un codificador comprime los datos en un espacio de menor dimensión, mientras que un decodificador reconstruye los datos sintéticos a partir de esta representación. A través de un proceso probabilístico, generan nuevas instancias de datos que son diversas y realistas, pero sin replicar puntos de datos específicos.

Modelado Probabilístico: implica el uso de distribuciones estadísticas y simulaciones (como la simulación de Montecarlo) para generar datos sintéticos. Este enfoque se basa en la observación de patrones en los datos reales y la creación de muestras sintéticas que respeten esas distribuciones.

Simuladores: utilizan modelos físicos, matemáticos o lógicos para crear datos sintéticos a través de la simulación de sistemas o procesos reales. Estas tecnologías permiten explorar escenarios hipotéticos controlando variables específicas, lo que resulta útil en la experimentación sin los riesgos asociados con los datos reales.

Modelos implícitos

Redes Generativas Antagónicas (GAN): modelos de aprendizaje profundo compuestos por dos redes neuronales: un generador que crea datos sintéticos y un discriminador que evalúa la calidad de esos datos comparándolos con datos reales. Ambas redes compiten entre sí, mejorando mutuamente sus capacidades y, a medida que las redes se entrenan, el generador mejora hasta que el discriminador no puede distinguir entre los datos reales y los sintéticos.

Modelos de Transformadores (GPT, BERT): modelos de aprendizaje profundo altamente eficientes en el procesamiento de secuencias, como el lenguaje natural. Modelos como GPT o BERT aprenden la estructura y las relaciones lingüísticas a partir de grandes volúmenes de datos textuales. Los codificadores transforman los datos de entrada en representaciones numéricas llamadas *embeddings*, mientras que el decodificador genera nuevas secuencias de salida utilizando mecanismos de autoatención, lo que permite que el modelo se concentre en los *tokens* relevantes.

Modelos de difusión: modelos generativos que crean datos, especialmente imágenes, mediante un proceso de difusión progresiva. En este proceso, agrega ruido aleatorio a los datos hasta que se vuelven irreconocibles y luego aprende a revertir ese proceso.

Además, elegir entre modelos generativos o simulaciones depende del equilibrio entre realismo, privacidad, explicabilidad y contexto de uso

Aplicaciones comunes de los modelos generativos

Crear datos realistas

Generación de imágenes y videos

Generación de texto

Entrenamiento de modelos de IA

Aplicaciones comunes de la simulación

Explorar diferentes escenarios

Optimización de procesos industriales

Simulación de sistemas financieros

Simulación en la salud

Modelos generativos vs. simulación

Tratando de la generación de datos sintéticos, es importante comprender la distinción entre dos enfoques diferentes, **los modelos generativos que incluirían las GAN, VAE o Transformadores, y los modelos basados en simulación.**

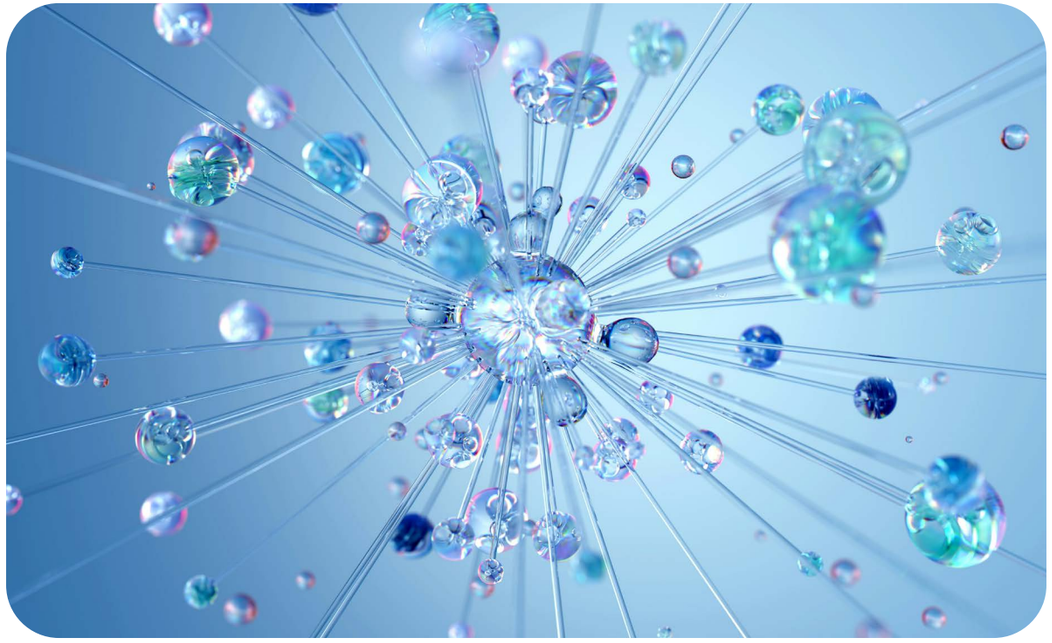
Los modelos generativos están diseñados para crear datos sintéticos que imitan las características estadísticas de los datos reales. A través de algoritmos de aprendizaje automático, **los modelos generativos aprenden las distribuciones de los datos originales y luego generan nuevas muestras de datos que siguen estas mismas distribuciones.** Esta técnica de generación de datos se caracteriza por la generación realista de los mismos tanto por su estructura como por sus características estadísticas y los hace ideales para imitar datos no estructurados, como imágenes, texto y audio.

Además, para funcionar correctamente, estos modelos **requieren ser entrenados con datos reales**, permitiéndole replicar comportamientos similares en los datos generados. Estos modelos se utilizan en diversos sectores, desde la creación de contenido (imágenes, música o texto) hasta el entrenamiento de IA y la protección de datos sensibles, ya que pueden generar datos de alta calidad sin comprometer la privacidad de los individuos.

A diferencia de los modelos generativos, los modelos de simulación no se centran en crear datos nuevos a partir de un conjunto de datos existente, sino en **replicar el comportamiento de sistemas reales bajo condiciones controladas.** Estos modelos utilizan enfoques matemáticos, físicos o lógicos para modelar fenómenos y eventos del mundo real. Se caracteriza por su eficacia en el análisis de sistemas dinámicos y complejos, como los procesos económicos, de salud o industriales, donde las interacciones entre las variables no siempre son observables directamente.

También son ideales cuando se desea estudiar el comportamiento de un sistema en diferentes condiciones, sin necesidad de utilizar datos reales. Pueden implicar la simulación de flujos de trabajo, interacciones de agentes o procesos físicos. A través de la simulación, **es posible crear una variedad de escenarios "qué pasaría si", donde se pueden ajustar variables y observar los resultados sin comprometer recursos ni riesgo en situaciones reales.** Y, a diferencia de los modelos generativos, los modelos de simulación permiten un control completo sobre las variables involucradas en el modelo, lo que resulta útil para probar hipótesis o hacer predicciones en entornos controlados.

El uso de datos sintéticos revoluciona el intercambio interno y colaboración externa, asegurando privacidad, agilidad y confianza total



Cuándo y cómo utilizar los datos sintéticos: desde el intercambio interno hasta la colaboración externa

Su flexibilidad y capacidad para representar escenarios diversos los convierten en una herramienta clave en distintas fases del ciclo de vida de los modelos de inteligencia artificial y *Machine Learning.* Para maximizar su valor, es esencial comprender en qué contextos y procesos resultan más adecuados, **abordando su implementación con un análisis estratégico que considere objetivos, necesidades y entorno.**

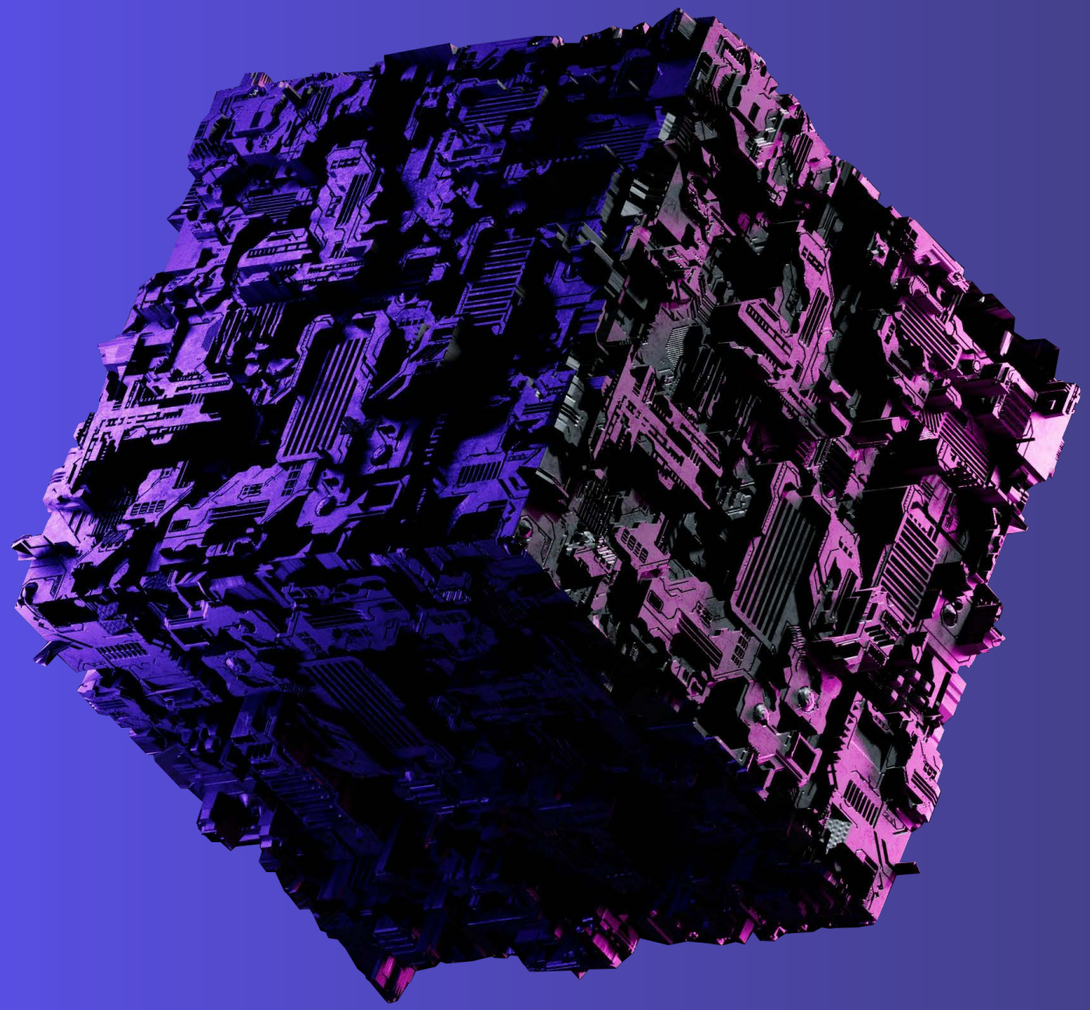
Antes de decidir su uso, debe realizarse un **estudio de las necesidades organizacionales y de las limitaciones existentes en el manejo de datos reales:** restricciones de acceso, coste de adquisición o calidad. Este análisis permitirá determinar si los datos sintéticos ofrecen una alternativa viable.

Una vez evaluados estos aspectos, es necesario **definir cómo y cuándo aplicarlos en las diferentes fases del ciclo de vida de los modelos de IA** y en las operaciones empresariales. Algunos factores estratégicos clave son:

- **Etapa de desarrollo del producto o servicio:** permiten probar y validar sin comprometer datos reales.
- **Entrenamiento de modelos de IA y ML:** resultan útiles cuando los datos reales son insuficientes, incompletos o sesgados.
- **Simulaciones y escenarios hiperrealistas:** posibilitan recrear situaciones de riesgo o condiciones extremas difíciles de obtener.
- **Colaboración interna y externa:** facilitan el intercambio de datos representativos sin afectar la privacidad ni la seguridad.

Al identificar las necesidades y barreras de la organización, se puede definir de forma efectiva cuándo y cómo implementarlos. No obstante, su adopción requiere una **planificación cuidadosa que alinee los objetivos tecnológicos con la estrategia empresarial.**

Los datos sintéticos fortalecen la colaboración interna y externa en investigación, innovación y formación



Del laboratorio al mercado: oportunidades y desafíos

Con datos sintéticos, las empresas transforman restricciones en ventajas, impulsan innovación rápida y colaboran sin riesgos

Desbloqueando nuevas oportunidades con datos sintéticos

Los datos sintéticos se posicionan como una **nueva técnica para resolver una variedad de problemas en la gestión y análisis de datos**, particularmente en sectores que enfrentan limitaciones por regulaciones de privacidad, escasez de datos y costes elevados. Sin embargo, su implementación trae consigo tanto oportunidades como desafíos que deben ser comprendidos y gestionados cuidadosamente.

Protección de la privacidad sin limitar la innovación

Integran la privacidad proactivamente, eliminando información personal identificable, reduciendo riesgos y facilitando auditorías, **convirtiéndola en una ventaja estratégica que desbloquea el acceso a datos y mantiene la confianza de reguladores y clientes.**

Eficiencia operativa y velocidad

Eliminan las ineficiencias de los datos reales, y proporcionan un conjunto de datos listos sin necesidad de aprobaciones legales, formularios o limpiezas. Esto agiliza el desarrollo de IA, pruebas y diseño de productos, resultando en una **experimentación más rápida, ciclos de desarrollo reducidos.**

Colaboración segura y crecimiento del ecosistema

Transforman el problema de la escasez de datos permitiendo compartir conjuntos de datos de manera segura y legal. Esto se traduce en **una oportunidad para escalar la innovación y la I+D a través de colaboraciones externas, *joint ventures* o asociaciones de datos.**

Diferenciación e innovación a futuro

Permiten modelar eventos raros, simular el comportamiento del cliente y diseñar para casos límite, todo sin esperar ejemplos del mundo real. También posibilitan la **creación de modelos de IA más justos y menos sesgados**, alineándose con agendas de sostenibilidad y tecnología responsable.

Avances éticos en IA

Al generar datos que simulan condiciones del mundo real, pero sin los prejuicios inherentes a los datos reales, los modelos entrenados con datos sintéticos pueden ayudar a **evitar discriminaciones en áreas críticas** como el reconocimiento facial, el crédito y la salud. Lo que permite a las empresas cumplir con sus compromisos éticos y generar confianza.

Nuevas fuentes de ingresos

Implementando esta técnica de generación de datos, se **abre una variedad de nuevas fuentes de ingresos**, como la monetización de datos, asociaciones R&D, *marketplaces* o catálogos de datos sintéticos, entre otras.

Superar los desafíos técnicos de los datos sintéticos es clave para garantizar su valor, utilidad y adopción generalizada

Los desafíos están relacionados principalmente con la calidad, el realismo y la representatividad de los datos generados, aspectos cruciales para garantizar que los modelos de inteligencia artificial funcionen correctamente cuando se entrenan con ellos.

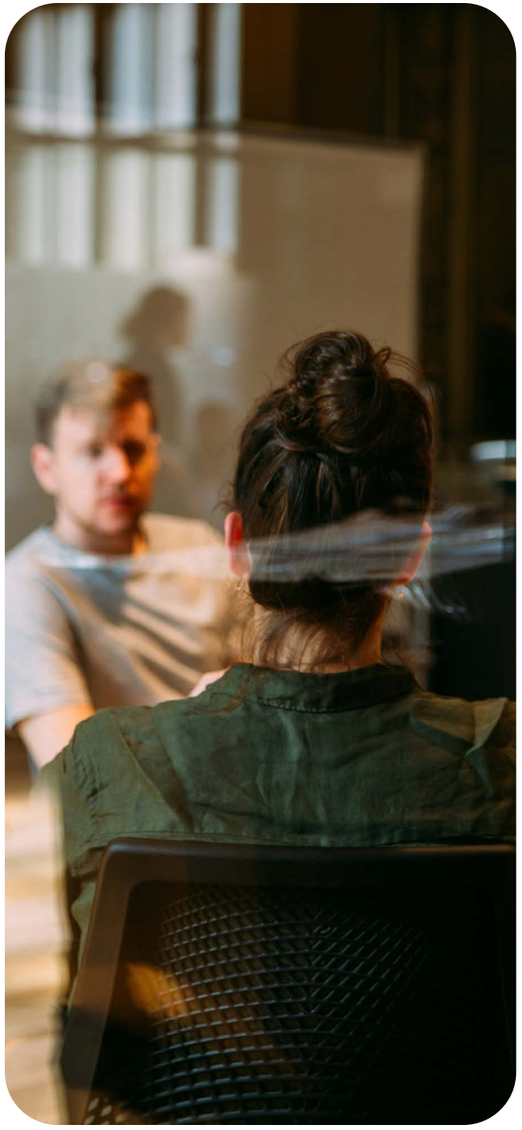
1. **Calidad y realismo**
   Uno de los desafíos más críticos de los datos sintéticos es garantizar su calidad y realismo. Los **modelos de IA dependen en gran medida de la calidad de los datos para aprender patrones correctos y hacer predicciones precisas.** Si los datos sintéticos no reflejan con exactitud las complejidades del mundo real, los modelos de IA pueden tener dificultades para generalizar o incluso producir resultados erróneos.

   La generación de datos sintéticos que sean estadísticamente similares a los datos reales es un proceso complejo, **los modelos deben ser capaces de capturar variaciones sutiles y comportamientos inusuales para evitar predicciones imprecisas.** Los métodos de generación han mejorado considerablemente la generación de datos realistas, pero siguen siendo susceptibles a generar información sesgada o no representativa si los modelos no son entrenados adecuadamente.

2. **Representatividad**
   Otro desafío es la representatividad, para que los datos sintéticos sean útiles, **deben ser capaces de simular no solo la mayoría de los casos comunes, sino también aquellos que son poco frecuentes o extremos** y que también son esenciales en la toma de decisiones. Estos casos son cruciales, como las transacciones fraudulentas en el sector financiero o los diagnósticos raros en el sector sanitario. Además, a medida que se entrenan modelos de IA con datos sintéticos, para que los datos sean representativos, los algoritmos deben aprender las relaciones y correlaciones de las variables de manera precisa, lo cual puede ser difícil cuando las características de los datos reales son complejas o altamente no lineales.

3. **Equilibrio**
   La combinación de datos reales y sintéticos es clave para crear modelos robustos y precisos, los sintéticos pueden llenar vacíos de datos, especialmente en situaciones donde los datos reales son escasos, difíciles de acceder o sujeto a restricciones de privacidad. Sin embargo, **depender únicamente de los datos sintéticos podría comprometer la exactitud de los modelos y limitar su aplicabilidad en escenarios del mundo real.**

4. **Validación**
   La falta de métricas y estándares claros para evaluar la calidad de los datos sintéticos ha sido una de las principales barreras para su adopción generalizada. Actualmente, falta un marco unificado para medir la calidad de los datos sintéticos, lo que dificulta la tarea de comparar y validar la calidad de los datos generados por distintos algoritmos y modelos, pudiendo dar lugar a la adopción de datos sintéticos que, aunque puedan parecer representativos, no lo sean en la práctica.



Los datos sintéticos pueden amplificar sesgos y riesgos de privacidad si sus procesos generativos no son controlados cuidadosamente

Riegos y sesgos: ¿se puede auditar lo que no ocurrió?

También se plantean una serie de riesgos éticos y técnicos que deben ser gestionados cuidadosamente, estos incluyen los **sesgos inherentes a los datos originales, los problemas relacionados con la auditoría y trazabilidad, y las cuestiones legales y regulatorias** vinculadas al uso de estos datos. Además, existe un dilema fundamental: ¿cómo auditar un conjunto de datos que, por definición, no proviene de hechos ocurridos en el mundo real?

1. Si los datos originales contienen sesgos, como en el caso de las decisiones crediticias sesgadas racialmente, **los datos sintéticos pueden reproducir o incluso agravar estos sesgos**, afectando la equidad de los modelos de IA que se entrenan con ellos. Este problema puede ser aún más grave si el modelo generativo no maneja adecuadamente las distribuciones de los datos, lo que puede dar lugar a sesgos reciclados que refuercen desigualdades ya presentes.

2. Además, la auditoría de los datos sintéticos se convierte en un reto debido a que, al no provenir de hechos reales, **no se pueden verificar de la misma manera que los datos tradicionales.** En lugar de verificar en contraste a la realidad, los auditores deben centrarse en las propiedades estadísticas del modelo generativo y en los procedimientos utilizados, lo que requiere una nueva forma de validación. En lugar de buscar correspondencias directas con eventos reales, se debe asegurar que los datos sintéticos mantengan una coherencia estadística y no introduzcan sesgos inadvertidos.

3. Otro riesgo es la reidentificación, aunque los datos sintéticos están diseñados para proteger la privacidad, **si el modelo generativo no está bien regulado, podría memorizar detalles de los datos reales y recrearlos en los datos sintéticos,** lo que pondría en riesgo la misma privacidad. Este problema refleja la importancia de aplicar técnicas de protección diferencial, que añaden ruido controlado para proteger la identidad de los individuos.

4. El bucle autoreferente es otro problema, que ocurre cuando un modelo es entrenado con datos sintéticos generados por otro modelo sintético. En este ciclo, **los errores o sesgos del modelo original pueden amplificarse en los modelos sucesivos**, lo que puede alejar a los datos generados de los patrones reales. Para mitigar este riesgo, es fundamental reintroducir periódicamente datos reales o combinar datos reales con sintéticos en el proceso.

5. Finalmente, los **dilemas éticos sobre la responsabilidad** surgen cuando los modelos entrenados con datos sintéticos fallan o causan daño. En esos casos, surge la pregunta de quién es el responsable; el creador del modelo o el usuario que implementó el modelo. Es especialmente relevante en áreas como la toma de decisiones automatizadas, donde los modelos pueden afectar directamente a las personas. La solución requiere una auditoría más técnica y transparente, centrada en la validación de los procesos y las propiedades estadísticas de los datos sintéticos, en lugar de una verificación directa con los datos reales.

El impacto y valor de los datos sintéticos reside en redefinir la privacidad, potenciar la innovación y democratizar el acceso a los datos

Impacto de los datos sintéticos: impulsando una economía digital más eficiente y democrática

La adopción de los datos sintéticos por múltiples actores del sector representa una **completa transformación en la manera en que los distintos agentes abordan la privacidad, la sostenibilidad y la innovación.**

1. **Agilidad, eficiencia y nuevos ingresos**

   Permiten a las empresas moverse rápido sin comprometer la seguridad ni el cumplimiento. Al eliminar las restricciones de los datos personales, se reducen los tiempos de espera, se acelera el desarrollo de productos y se disminuyen los gastos operativos. Además, abren acceso a nuevos mercados y segmentos previamente inaccesibles debido a barreras regulatorias, lo que se traduce en mayor velocidad, menor fricción y mayor capacidad de monetización de datos.

2. **Innovación sin comprometer privacidad**

   Su capacidad de generar datos ilimitados **permite simular riesgos, validar ideas y entrenar algoritmos en situaciones no cubiertas** por datos reales. En un entorno donde innovación, ética y regulación deben ir de la mano, los datos sintéticos representan una estrategia de crecimiento responsable, que equilibra agilidad tecnológica y protección de la privacidad.

3. **Escalabilidad y resiliencia**

   Su adaptación no solo cumple con la ley, sino que se adelantan a ella. Esta tecnología permite operar y escalar productos en varias jurisdicciones sin los riesgos legales de los datos reales, convirtiéndola en una palanca para la expansión, y la reducción de riesgos reputacionales.

4. **Sostenibilidad y ética**

   Permiten reducir la huella de datos reales, al generar datos artificiales en lugar de depender de información personal, aliviando la presión sobre las infraestructuras digitales, reduciendo almacenamiento y procesamiento. Además, las empresas adoptan un modelo más ético y responsable, integrando la privacidad desde el diseño y reforzando la confianza del usuario.

5. **Democratización del acceso**

   Su rápido acceso y recopilación permite a las pequeñas y medianas empresas acceder a *datasets* valiosos sin los riesgos asociados, nivelando la competencia con los gigantes tecnológicos. Esto abre nuevas oportunidades para la colaboración y cocreación de valor.

6. **Diferenciación competitiva y posicionamiento**

   Su uso muestra su compromiso con la privacidad, la equidad algorítmica y la innovación responsable. Integrarlos optimiza procesos y refuerza el posicionamiento como líder digital confiable y ético. **La confianza del usuario y la transparencia son más clave que nunca.**

Calcular el ROI de datos sintéticos permite demostrar beneficios inmediatos en agilidad, ahorro, cumplimiento y ventaja competitiva

Cómo calcular y visualizar rápidamente el retorno inmediato de la inversión al implementar datos sintéticos

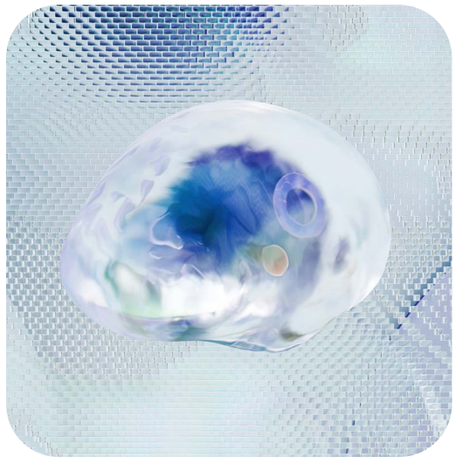
De la mano del impacto que ejerce esta tendencia en el sector, también destaca por el ROI que puede generar en su uso. La clave del ROI de los datos sintéticos **radica en la reducción de costes y la mitigación de riesgos previamente comentadas.**

Calcular el ROI de los datos sintéticos se basa en tres pasos:

1. **Identificación de beneficios tangibles** como la reducción de costes de cumplimiento, el ahorro de tiempo en el procesamiento de datos y la mejora en la eficiencia operativa.
2. **Cuantificación de los beneficios** asignando un valor monetario, como la reducción del tiempo del desarrollo o ahorros.
3. **Evaluación de los costes** de la inversión inicial en herramientas de software, recursos informáticos y validación.

Al final, no adoptar datos sintéticos expone a riesgos, frena la innovación y deja a las empresas fuera de la competencia futura

¿Por qué el momento es ahora? Factores que exigen la adopción de datos sintéticos



La creciente adopción de datos sintéticos está siendo impulsada por una convergencia de factores tecnológicos, regulatorios y éticos que reflejan tanto la demanda urgente de datos de alta calidad como la necesidad de cumplir con estrictas normativas de privacidad. El retraso en la adopción de datos sintéticos puede acarrear consecuencias significativas en varios frentes, desde los costes regulatorios hasta las desventajas competitivas y la pérdida de agilidad en el desarrollo de productos. Existen riesgos para las empresas que no integren rápidamente los datos sintéticos en sus operaciones:

Mayor riesgo regulatorio y exposición legal

A medida que las regulaciones de privacidad se endurecen, depender exclusivamente de datos reales expone a las empresas a sanciones, brechas de seguridad y demandas. Sin datos sintéticos, es difícil cumplir principios como la minimización de datos o el consentimiento explícito. Las consecuencias son **multas millonarias, perdidas de confianza en clientes y bloqueos operativos.**

Acceso limitado a datos críticos para la innovación

Los modelos de IA y análisis avanzados necesitan grandes volúmenes de datos diversos y las restricciones legales reducen la disponibilidad de datos reales. Sin datos sintéticos, las empresas enfrentan escasez de datos y sesgos en los modelos. Esto impacta en **modelos incompletos, innovación ralentizada y falta de capacidad para simular escenarios o anticipar riesgos.**

Tiempos más lentos y mayores costos operativos

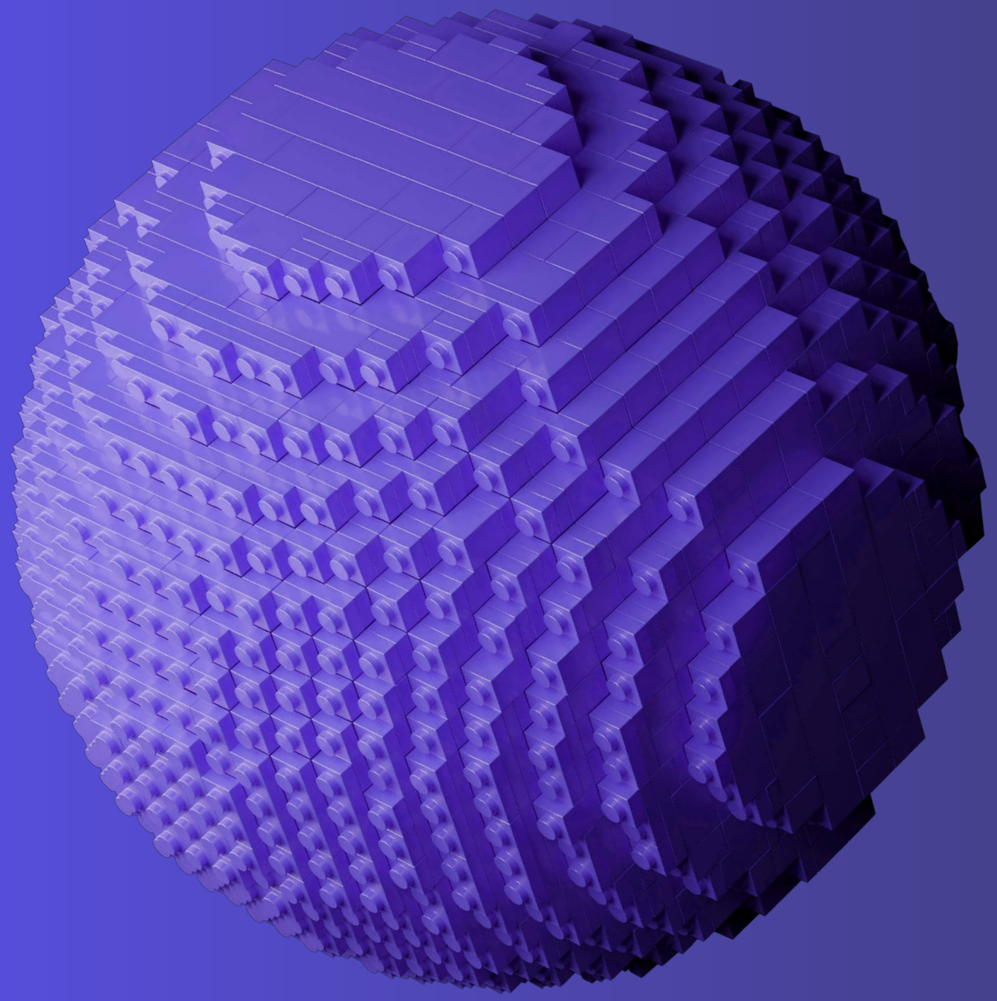
El uso de datos reales implica ciclos largos de aprobación, anonimización y revisión legal, los datos sintéticos permiten prototipar, probar y escalar más rápido y sin fricciones legales. Las consecuencias de no usarlos son un **mayor *time-to-market*, costes operativos elevados y lentitud frente a competidores más ágiles.**

Desventaja competitiva en IA y estrategia de datos

Líderes en salud, banca, seguros y tecnología ya usan datos sintéticos para entrenar mejores modelos y reducir sesgos. No adoptarlos implica quedar rezagado tecnológicamente frente a empresas que priorizan privacidad e innovación. El impacto se refleja en **pérdidas de posicionamiento, menor capacidad de personalización e imposibilidad de competir en mercados regulados.**

Dificultades para colaborar y escalar datos

Compartir datos reales entre equipos o con socios externos es cada vez más difícil y arriesgado.Los datos sintéticos permite una colaboración segura, interoperabilidad y resiliencia en entornos complejos. Las consecuencias de no usarlos son **silos de datos internos, obstáculos para proyectos conjuntos y transformación digital ralentizada.**

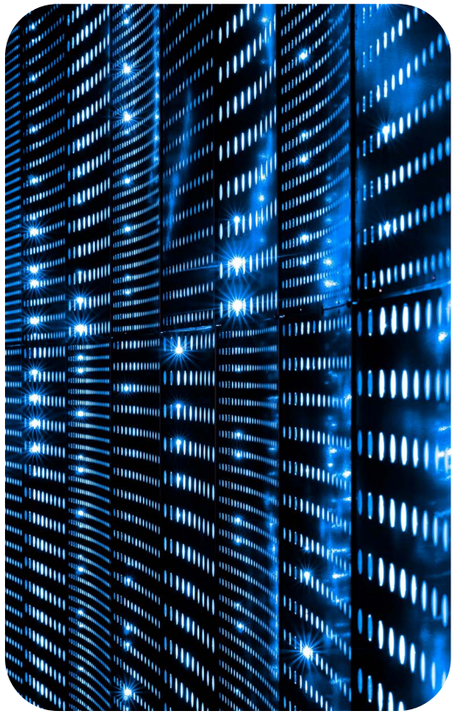


Regulación, gobernanza y estándares: entre barrera y catalizador

Sin ética ni regulación estricta, el uso de datos sintéticos puede violar derechos y generar sanciones millonarias

La necesidad de equilibrar la protección de derechos, gobernanza algorítmica y fomento de la economía digital requiere marcos normativos robustos. El GDPR, el AI Act y otras regulaciones emergentes son fundamentales para guiar el uso responsable de los datos sintéticos. A medida que su adopción crece, **se deben establecer principios éticos claros, así como mecanismos de certificación y auditoría para asegurar la transparencia y la explicabilidad en su uso.**

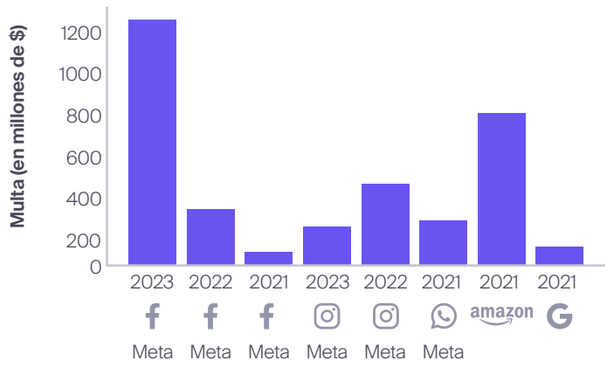
Las sanciones a gigantes tecnológicos subrayan la importancia de cumplir con estas normativas. Meta, por ejemplo, fue multada con 1.200 millones de euros por violar el GDPR en la transferencia de datos internacionales, mientras que Amazon recibió una multa de 746 millones de euros por no obtener el consentimiento adecuado para el seguimiento de datos. Todo ello resaltando la **necesidad de una regulación clara para proteger la privacidad y asegurar que la tecnología se use de forma ética y responsable.**



Fuente: Statista

Las mayores multas por infringir uno o más artículos del reglamento general de protección de datos

(en millones de $)



El rol del GDPR, AI Act y otras normativas emergentes

El Reglamento General de Protección de Datos de la UE ha sido probablemente el mayor impulsor indirecto de la tecnología de datos sintéticos, restringiendo enormemente el tratamiento de datos personales identificables (PII), requiriendo bases legales, consentimiento, minimización y severas sanciones si se reidentifica a alguien. Esta dureza regulatoria es lo que llevó a muchas organizaciones a buscar soluciones como los sintéticos para poder seguir usando datos sin infringir la ley. El GDPR no regula un *database* que no incluya datos personales, según este, si no es atribuible a persona identificada o identificable, entonces está fuera del alcance del reglamento.

Es más, el GDPR en su artículo 89 **fomenta el uso de técnicas para proteger datos y permitir investigación**, y menciona explícitamente la "seudonimización" y otras técnicas. Aunque no nombra lo sintético, este espíritu de *privacy by design* encaja con la filosofía de generar datos no reales para proteger privacidad.

Por su parte, la propuesta de **Ley de Inteligencia Artificial de la UE (AI Act), que entró en vigor el 1 de agosto de 2024**, empieza a reconocer los datos sintéticos. Menciona que el uso de datos sintéticos puede ser una medida para garantizar la calidad de datos de entrenamiento y la protección de derechos.

Regulaciones como GDPR, AI Act, CCPA/CPRA, HIPAA y APACs aseguran protección y ética en el tratamiento de datos sensibles

En EE.UU., leyes estatales como CCPA/CPRA, también se enfocan en datos personales. Si una empresa usa sintéticos en lugar de reales, posiblemente reduce su exposición, cabe destacar que no hay mención explícita en estas leyes, pero el principio es similar a GDPR. Un caso especial es el sector salud con HIPAA en EE.UU.

El HIPAA define metodologías de des-identificación (lista de supresión de identificadores o certificación de experto), los datos sintéticos no aparecen explícitamente, pero pueden encajar en una certificación de experto que avale que los datos no identifican a nadie. De hecho, la FDA y organismos de salud están explorando su uso para ensayos clínicos simulados, etc., lo que sugiere adaptaciones futuras en regulaciones sanitarias.

En el ámbito internacional, organismos como la OCDE y el G7 han destacado PETs. La OCDE ya publicó información sobre tecnologías emergentes de privacidad e incluyó a los datos sintéticos, discutiendo su potencial y la necesidad de marcos legales. Finalmente, el G7 en su fórum de IA mencionó la importancia de fomentar las PETs.

En cierto sentido, estas regulaciones actuaron como una barrera inicial, pero a la vez como un catalizador para innovar en PETs. Es más, algunas de ellas podrían requerir a ciertos proveedores de IA que apliquen PETs en sus procesos, con lo cual **los datos sintéticos pasarían de ser opcionales a casi obligatorios en entornos regulados.**

GDPR

Transparencia
Consentimiento informado
Derecho al olvido
Minimización de datos
Anonimización
Responsabilidad y gobernanza

AI ACT

Regulación según riesgo
Ética y derechos fundamentales
Transparencia
Evaluación continua

CCPA (EE.UU.)

Derecho a la información
Derecho a la eliminación
Opt-out de la venta de datos
No discriminación

HIPAA (EE.UU.)

Protección de datos de salud
Acceso y corrección
Confidencialidad
Notificación de violaciones

APACs (ASIA Y PACÍFICO)

Consentimiento
Acceso y corrección
Transparencia
Gestión de datos
Protección de datos sensibles

Pero existe ambigüedad y vacíos legales y se necesitan normas claras que alineen generadores, usuarios y reguladores

Navegando el vacío legal para una innovación responsable

A pesar de los esfuerzos, como los informes técnicos de ISO/IEC JTC1 SC42, que proporcionan una visión general de los datos sintéticos y sus aplicaciones, actualmente no existe un **estándar universalmente aceptado para certificarlos.** Esta falta de estandarización genera incertidumbre tanto para las organizaciones que utilizan datos sintéticos como para los reguladores que buscan proteger los derechos de los usuarios.

En este contexto, algunas empresas, como Gretel Labs, están implementando herramientas como "*Privacy Reports*" y "*Utility Reports*" para evaluar la calidad y la privacidad de los datos generados, lo cual es un primer paso hacia la normalización.

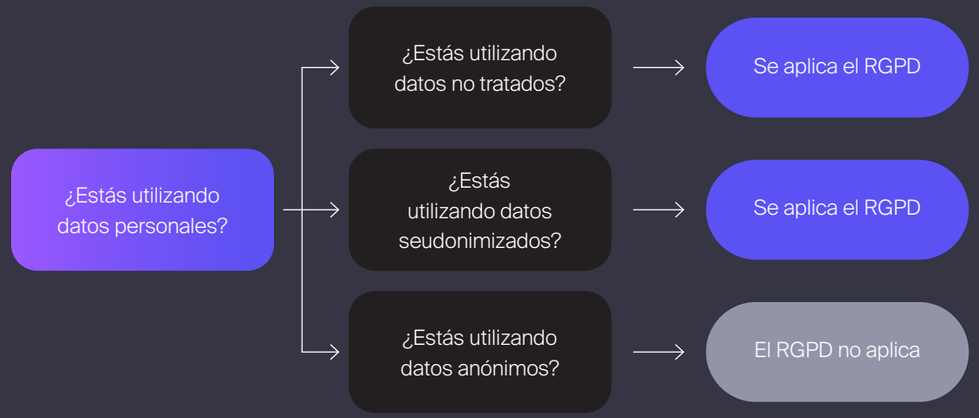
Sin embargo, aún se requieren normas claras para garantizar que los generadores, usuarios y reguladores estén alineados. Entre los problemas más destacados, encontramos:

- **Definición y clasificación.** La pregunta fundamental es si los datos sintéticos son siempre "anónimos". A pesar de que muchos marcos legales, como el GDPR y el AI Act, aceptan que los datos sintéticos no deberían estar sujetos a regulaciones de privacidad si no pueden vincularse a individuos, aún no existe un test técnico o legal universalmente aceptado para determinar cuándo cumplen con los requisitos de anonimización.
- **Solapamientos y vacíos entre leyes.** Algunas normativas hacen referencia directa a los datos sintéticos, las condiciones de validación, transparencia y los riesgos asociados a la ingeniería inversa siguen siendo ambiguos. Las leyes de gobernanza de datos se diseñaron para datos recogidos, no para datos generados.
- **Requisitos de responsabilidad y auditoría.** El AI Act propone etiquetar los contenidos sintéticos y demostrar que no provienen de datos reales, pero aún no existe un proceso de auditoría estandarizado para verificar su privacidad o representatividad. Tampoco se han definido estándares claros sobre la documentación necesaria ni se han establecido normativas que determinen las responsabilidades en caso de sesgo o representaciones inexactas.
- **Riesgos legales de reidentificación y calidad.** Los ataques que podrían desanonimizar los conjuntos de datos sintéticos están en aumento. Los datos sintéticos de alta fidelidad pueden inadvertidamente reproducir patrones o información personal que puede ser rastreada hasta individuos, generando una zona gris legal.
- **Implicaciones para el intercambio de datos internacional y competencia.** Los sintéticos pueden reducir las barreras de acceso a los datos en mercados competitivos, lo que plantea nuevos desafíos en términos de leyes antimonopolio. Además, las normativas sobre el intercambio transfronterizo de datos no son claras sobre cómo deben tratarse los conjuntos de datos sintéticos generados a partir de datos de individuos de diferentes jurisdicciones.

Las reformas legales y la clarificación de los estándares técnicos son ampliamente reconocidas como necesarias para cerrar **estas brechas y asegurar el uso seguro y ético de las nuevas tendencias.**

El futuro regulatorio exigirá mayor responsabilidad, transparencia y controles claros en el uso de datos sintéticos empresariales

Aplicación del Reglamento General de Protección de Datos (RGPD)



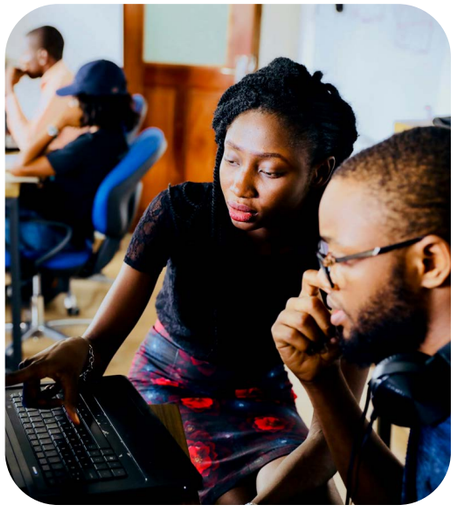
Tendencias futuras y recomendaciones para las empresas

Para el año 2026, el AI Act tendrá establecido un marco más riguroso, clasificando los usos de la IA según el nivel de riesgo e imponiendo obligaciones específicas para sistemas entrenados con datos sintéticos. Además, **impulsará una mayor exigencia de transparencia en cuanto a la generación y el uso de estos datos, exigiendo que las empresas demuestren cómo se protege la privacidad y la equidad en los modelos de IA.**

Es más, el uso estos datos sintéticos basados en los reales podría **obligar a las empresas a realizar evaluaciones de impacto en privacidad** (EIPD), auditar los riesgos de reidentificación y documentar controles de mitigación. Es decir, la trazabilidad y la explicabilidad de los procesos de generación serán claves para garantizar la conformidad regulatoria.

Además, el EU Data Act, el cual establece **reglas para el acceso y uso de los datos**, será aplicable a partir de finales del año 2025. Esto indica un entorno de datos más abierto, transparente y competitivo, favoreciendo la innovación y la protección de los derechos de los usuarios.

Para prepararse ante un entorno en creciente regulación, las **empresas deberán asegurarse de que se adaptan a las restricciones mediante la implementación de evaluaciones de impacto en privacidad y establecer un marco de gobernanza sólido**, con controles de acceso, gestión de riesgos y etiquetado claro de los conjuntos de datos generados. Las pruebas de validación, como las de reidentificación o calidad, serán las mejores aliadas de las organizaciones generadoras



Con esta *checklist*, las organizaciones pueden preparar su uso de datos sintéticos evitando sanciones y mostrando compromiso ético

Checklist estratégico-regulatorio para preparar el uso de los datos sintéticos

Se deben tomar ciertas acciones ahora para poder minimizar riesgos a sanciones, incidencias o reputaciones en el futuro, además de demostrar un **liderazgo ético y tecnológico en la economía digital.**

Evaluación previa y análisis de riesgos

Identificar fuentes de datos: clasificar datos de origen y justificar su uso.

Análisis de privacidad y riesgos: realizar una DPIA (*data protection impact assessment*) inicial y periódica.

Consentimiento y bases legales: verificar que la recopilación y uso cumplan con la base legal.

Diseño y configuración de modelos

Selección de técnicas apropiadas: escoger modelos generativos según el caso de uso y el tipo de dato.

Configuración de controles de privacidad: integrar técnicas de anonimización para reducir riesgos de reidentificación.

Definir objetivos de negocio y *compliance*: alinear la generación de datos sintéticos con objetivos estratégicos y de cumplimiento corporativo.

Validación y aseguramiento de calidad

Control de fidelidad y utilidad: realizar pruebas estadísticas, controles de calidad y auditorías.

Evaluación de riesgos de reidentificación: aplicar *benchmarks* y ataques simulados.

Evaluación de equidad y sesgos: realizar análisis de sesgo y representatividad para evitar la amplificación de desigualdades.

Gobernanza, documentación y auditoría

Gobernanza clara: definir roles y responsabilidades, así como políticas para su ciclo de vida.

Documentación exhaustiva: mantener documentos actualizados preparados.

Estándares y certificaciones: adoptar marcos reconocidos para demostrar *compliance* y fortalecer la confianza de clientes, reguladores y *partners*.

Monitoreo y mejora continua

Seguimiento regulatorio: monitorizar cambios en leyes (GDPR, AI Act, etc.) Y adaptarse ante nuevas obligaciones.

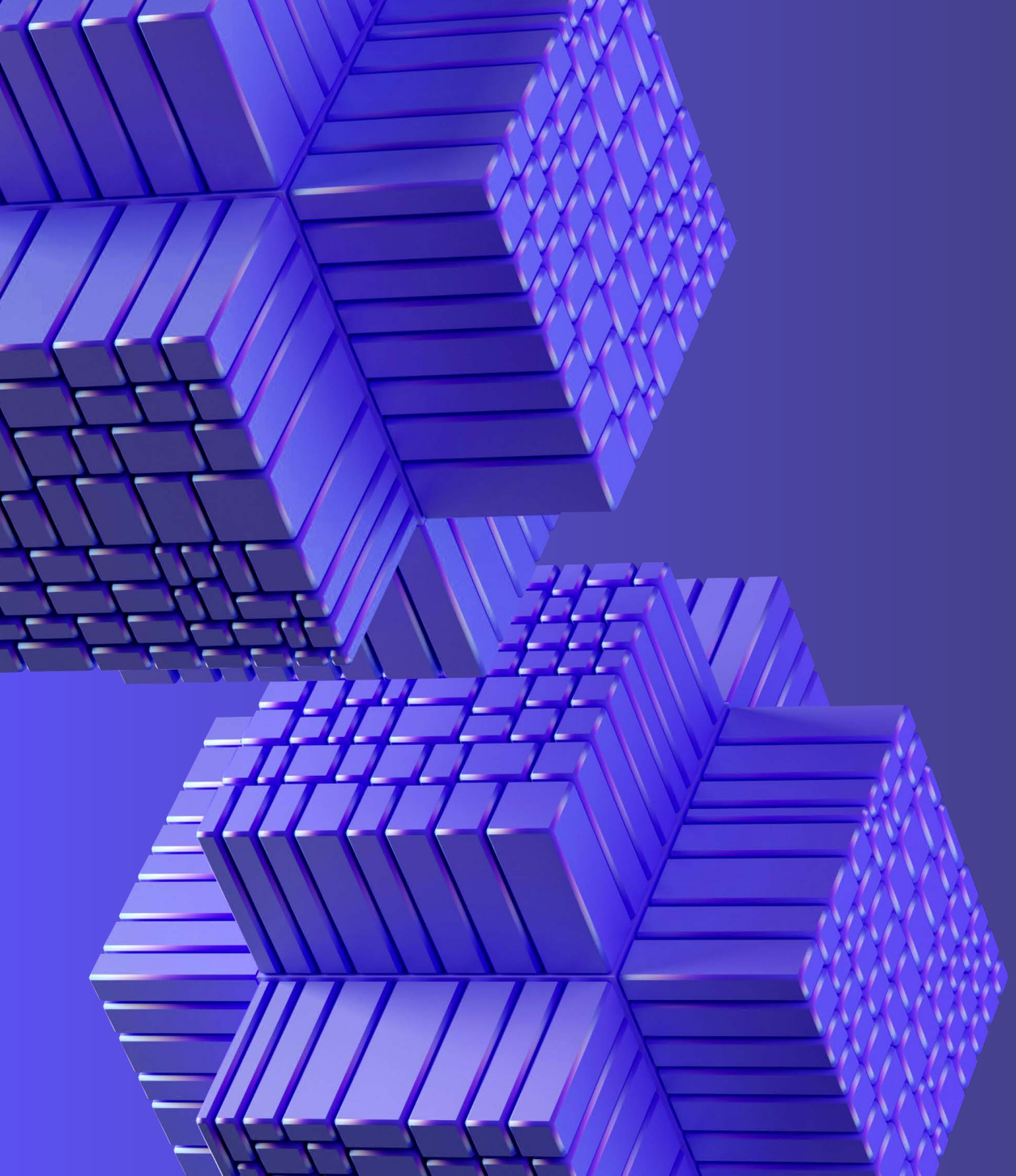
Control de calidad automatizado: usar herramientas *AI-driven* para revisar la calidad, privacidad y utilidad de los datos.

Auditoría y respuesta a incidentes: establecer sistemas de alerta y protocolos de actuación rápida ante vulnerabilidades.

Actualización y formación

Capacitación del personal: invertir en formación sobre nuevas tecnologías y mejores prácticas.

Comunicación y transparencia: informar a consumidores y reguladores sobre el uso, beneficios y salvaguardas de los datos sintéticos.



Los *stakeholders* y las percepciones, prioridades y resistencias

La percepción general de los líderes de la industria es positiva, pero persisten barreras de conocimiento, calidad e implementación efectiva

Opiniones de los líderes de la industria sobre la aplicación de datos sintéticos

Actualmente, **un 89% de los ejecutivos de grandes corporaciones consideran el uso de datos sintéticos como un elemento esencial** para mantener su competitividad en el mercado.

Los líderes de la industria con conocimiento sobre las tecnologías de datos sintéticos de **vanguardia han expresado su confianza en la capacidad de esta tecnología para abordar problemas críticos utilizando datos del mundo real.** A pesar de reconocer la importancia de mejorar los datos, más de la mitad (51%) de los encuestados no están alineados con la definición técnica explícita de los enfoques avanzados de datos sintéticos, lo que revela una brecha de conocimiento crítica.

De los que comprendieron la definición correcta, el 50% mencionó que uno de los principales beneficios de los datos sintéticos es superar la limitación de *labels* proporcionadas a través del aprendizaje supervisado y la anotación humana.

Aunque muchos están convencidos de los beneficios potenciales de estos datos, **existen barreras significativas para su adopción**, es más, el 67% de los líderes de la industria coinciden en que su organización carece del conocimiento necesario para implementar datos sintéticos de manera efectiva. Asimismo, el 67% reconocen que los usuarios de su industria no aceptarán los datos sintéticos hasta que vean beneficios claros por sí mismos.

Fuente: Synthesis

Entre los aspectos más desafiantes de su utilización dentro de las organizaciones:

46%

de los encuestados se preocupan por la calidad de los modelos creados con datos sintéticos, temiendo que no sean tan buenos como los generados con datos "reales".

45%

mencionaron las dificultades para crear datos sintéticos de alta calidad para sistemas complejos.

42%

indicaron que los costes de integración e implementación representan un desafío significativo.

A pesar de estos desafíos, el **59% de los responsables de la toma de decisiones creen que su industria utilizará datos sintéticos, ya sea de manera independiente o combinados con datos "reales", en el futuro.** Esto sugiere que muchas organizaciones están comenzando a experimentar con esta tecnología, pero aún queda camino por recorrer. Es decir, en general, aunque hay un gran interés en los datos sintéticos, aún se necesita una mayor infraestructura y capacitación para su implementación generalizada.

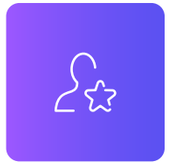
Además, la adopción exitosa de datos sintéticos requiere la alineación entre CIO, CDO y DPO para integrar intereses y gestionar el cambio

CIO, CDO, DPO: el tablero de decisión empresarial

La introducción de **datos sintéticos en una organización involucra a diversos *stakeholders*** o partes interesadas, cada uno con perspectivas y preocupaciones particulares. Comprender estas percepciones es vital para gestionar el cambio hacia la adopción de datos sintéticos.

La adopción de esta tecnología no solo depende de la innovación tecnológica, sino también de la integración de intereses. En este contexto, **los roles de los CIO, CDO y DPO juegan el papel de decisores** en la toma de decisiones que impactan directamente la adopción y el éxito de los datos sintéticos dentro de una organización. Cada uno de estos actores aporta un enfoque distintivo a la mesa de decisión, siendo su alineación estratégica crucial para avanzar en la adopción.

Los datos sintéticos pueden mejorar significativamente el rendimiento del modelo y pueden reemplazar eficazmente los datos reales en un 60% a 80% de los casos sin perder rendimiento



CIO (*Chief Information Officer*): como **responsables de la estrategia digital y la infraestructura tecnológica**, los CIO juegan uno de los principales papeles en la adopción de los datos sintéticos. Para ellos, esta tecnología representa una oportunidad estratégica para generar ventajas competitivas mediante la creación de modelos de IA personalizados entrenados con datos exclusivos. Su enfoque se centra en la aceleración de la innovación y la mitigación de riesgos tecnológicos. Su decisión de adoptar estos datos, sin embargo, depende de la demostración de un retorno de inversión claro, lo que a menudo se logra mediante la mejora en la eficiencia de los procesos de desarrollo.

CDO (*Chief Data Officer*): el CDO es el **principal defensor de la calidad de los datos dentro de la organización** y ve en los datos sintéticos una solución vital para resolver el problema del acceso a datos de calidad. Están interesados en cómo los datos sintéticos pueden desbloquear activos de datos restringidos por normativas de privacidad, permitiendo su uso más amplio y facilitando el intercambio de datos dentro y fuera de la empresa. Además, buscan en la monetización de datos sintéticos una oportunidad para crear productos de datos que no infringen la privacidad.

DPO (*Data Protection Officer*): los DPO, **encargados de asegurar el cumplimiento de las normativas de privacidad**, tienen una perspectiva más crítica en cuanto al uso de estos datos. Para ellos, esta tecnología puede ser una herramienta clave para cumplir con el principio de minimización de datos establecido por el GDPR. Su enfoque se centra en garantizar que los datos sean realmente irreversibles e inidentificables, evitando cualquier posibilidad de reidentificación. Son responsables de implementar políticas internas que definan cuándo los datos sintéticos cumplen con los requisitos de anonimización, evitando riesgos legales.

CIO, CDO y DPO deben colaborar integrando tecnología, gobernanza y privacidad para adoptar datos sintéticos exitosamente

	CIO: impulsor de la innovación y la competitividad	**CDO:** guardián del valor de los datos	**DPO:** asegurando la privacidad y cumplimiento regulatorio
Preocupaciones y necesidades	• **Modernizar la infraestructura** tecnológica para adoptar tecnologías emergentes ágilmente. • **Acelerar la integración de soluciones de vanguardia** garantizando la diferenciación en el mercado. • **Proteger el ecosistema integrando IA y ML** para la prevención de amenazas e integridad de la data. • **Abordar brecha de habilidades** en los equipos.	• **Garantizar auditabilidad y cumplimento** de estándares internos y regulatorios. • **Asegurar el mantenimiento de su valor analítico** sin comprometer la calidad ni la representatividad. • **Ofrecer una solución para el intercambio** seguro de datos. • **Gestionar la adaptación rápida** ante nuevas regulaciones y asegurar validación de datos.	• **Asegurar los datos sintéticos sean irreversibles e inidentificables**, cumpliendo el principio de minimización. • **Garantizar el cumplimiento** de leyes de protección de datos y evitar cualquier violación. • **Documentar procesos** de generación con controles robustos y acceso a auditorias.
Pasos a seguir	1. **Alinear la estrategia** de la empresa con el uso de datos sintéticos. 2. **Implementar** soluciones de datos sintéticos para la mejora de la **agilidad operativa y la reducción de costes.** 3. **Asegurarse de que la infraestructura pueda manejar tanto los datos sintéticos como los reales** de manera eficiente y escalable.	1. **Desarrollar estrategia de gobernanza clara alineada** con los estándares internos y regulatorios. 2. **Evaluar calidad y representatividad** para garantizar su utilidad en los modelos de IA y analítica avanzada. 3. **Facilitar integración en los procesos** de colaboración de datos sintéticos tanto internos como externos.	1. **Verificar** que los datos sintéticos sean generados **siguiendo los principios normativos aplicables.** 2. **Implementar mecanismos de auditoría y transparencia** para supervisar el proceso de generación de datos. 3. **Colaborar internamente** para evitar riesgos de reidentificación ni violaciones la privacidad de los usuarios.

Para los consumidores, los datos sintéticos combinan seguridad y progreso, pero requieren transparencia para asegurar aceptación y confianza

¿Qué piensan los consumidores?

A medida que la tecnología de datos sintéticos sigue ganando terreno, la confianza del consumidor se presenta como un factor fundamental para su adopción y ejecución. El 94% de las organizaciones reconoce que, si no protegen adecuadamente los datos, los consumidores dejarán de comprar sus productos. Sin embargo, **un 87% de los consumidores no confían en que las empresas manejen sus datos de manera responsable,** lo que resalta la importancia de generar confianza.

Son una solución realmente prometedora, especialmente cuando se comunican adecuadamente como una medida para proteger la privacidad y reducir riesgos de exposición, ya que la gran **mayoría de consumidores priorizan a las empresas que valoran su privacidad, aumentando su lealtad.** No obstante, la falta de transparencia o el uso no autorizado de datos puede generar desconfianza. El 91% de los consumidores exige que las empresas protejan más sus datos personales, especialmente en el uso de tecnologías como IA y datos sintéticos.

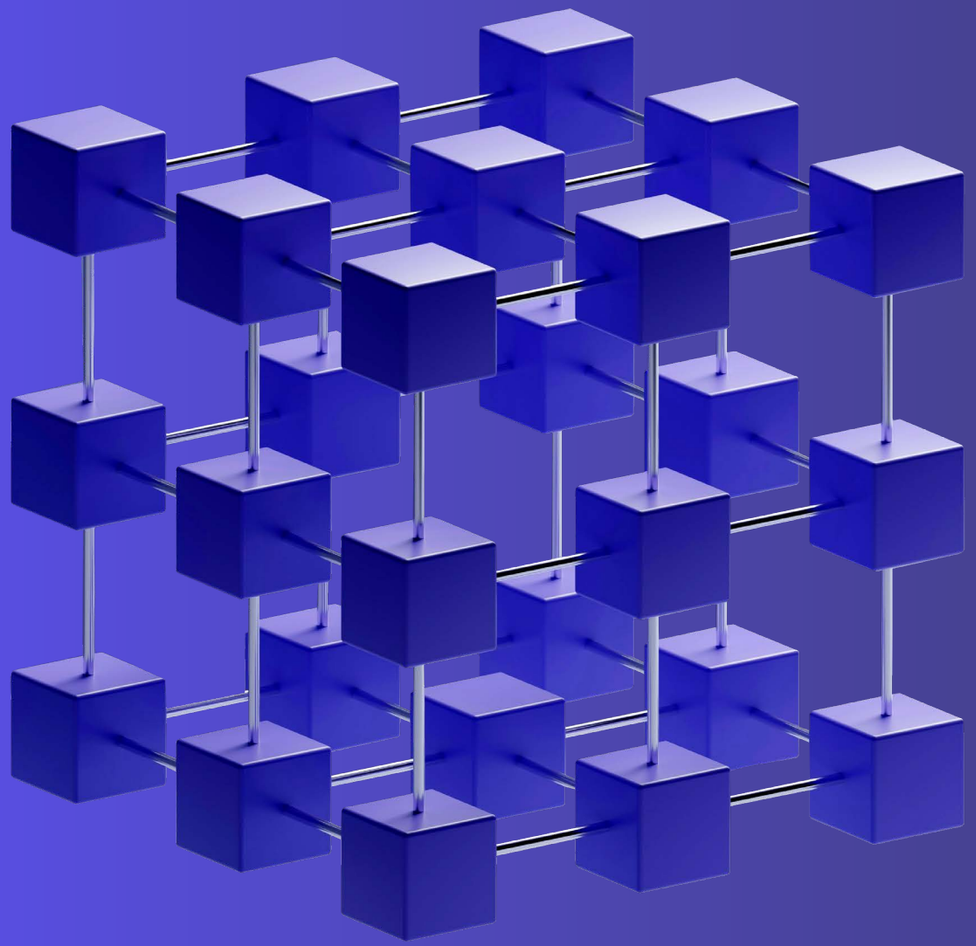
En este sentido, **el 80% cree que las políticas de privacidad son positivas**, pero la confianza depende de una gestión ética, de la transparencia y de ir más allá de las regulaciones para asegurar la protección efectiva de los datos.

Impacto de los datos sintéticos en la privacidad y confianza del consumidor

En los entornos entre las empresas y los consumidores finales es fundamental proteger la privacidad y los datos sintéticos lo consiguen **eliminando la necesidad de utilizar información personal real.** Esto permite a las empresas innovar en productos y servicios, y ajustar estrategias como precios, sin comprometer la seguridad de los datos.

Los beneficios directos para los consumidores son la mayor protección de la privacidad, la reducción de riesgos asociados con brechas de seguridad y una mayor transparencia en las prácticas de privacidad. Sin embargo, **la falta de comprensión pública sobre qué son los datos sintéticos puede generar desconfianza** si no se comunica adecuadamente, especialmente con decisiones algorítmicas de generación de datos.

Por ello, es esencial que las empresas **expliquen claramente cómo se generan y utilizan estos datos** para asegurar la confianza de los consumidores y garantizar que las decisiones basadas en estos datos sean transparentes y responsables.



Impacto sectorial: datos sintéticos como motor de transformación

La adopción temprana se concentra en los sectores donde la presión por innovar, escalar y proteger datos es crítica para la operación diaria

La adopción de los datos sintéticos sigue una curva típica de difusión tecnológica, donde algunos sectores y empresas están liderando su implementación, mientras que otros todavía se encuentran en una fase más rezagada. Aparecen los siguientes perfiles en la curva de adopción.

Innovators

Al igual que con otras innovaciones tecnológicas, los primeros en adoptar estas soluciones, los llamados innovators, están formados por sectores de alta inversión en I+D y una gran cultura de la experimentación, como startups de IA, Big Tech, Insurtechs, sector automotriz y las Fintechs emergentes.

Early adopters: los líderes en la adopción de datos sintéticos

Varios sectores son líderes en capitalizar el uso de estos datos. Estos comprenden las ventajas estratégicas que ofrecen los datos sintéticos, especialmente en términos de agilidad, cumplimiento normativo, y privacidad.

- **Sector financiero (banca, seguros).** La banca y las aseguradoras han sido de los primeros en adoptarlos, impulsados por regulaciones estrictas de privacidad y una gran necesidad de manejar grandes volúmenes de datos de manera eficiente.
- **Salud y ciencias de la vida.** Empresas farmacéuticas, así como centros de investigación, han comenzado a implementar datos sintéticos en ensayos clínicos y estudios médicos. En este sector, la privacidad y la precisión son extremadamente importante, lo que ha impulsado su adopción temprana. La investigación médica se está beneficiando de los datos sintéticos para realizar estudios sin exponer información sensible.
- **Tecnológicas.** Las grandes empresas tecnológicas han sido pioneras en el uso de datos sintéticos, principalmente para entrenar modelos de IA en entornos de simulación. También en el sector automotriz, están utilizando los datos sintéticos para entrenar sistemas de conducción autónoma.
- **Sector público innovador.** Algunos gobiernos, como el del Reino Unido, Corea del Sur, y Singapur, han comenzado a experimentar con datos sintéticos para mejorar la transparencia y la compartición de datos sin comprometer la privacidad de los ciudadanos.

Early majority: adoptando la tecnología tras ver resultados claros

Estas industrias ahora están comenzando a ver los beneficios claros de los datos sintéticos, como la mejora en la eficiencia operativa y la optimización de recursos.

- **Industria 4.0.** Empresas de manufacturación están utilizando datos sintéticos para simular fallos de máquinas y entrenar sistemas de monitoreo predictivo.
- **Telecomunicaciones.** Las empresas de telecomunicaciones están utilizándolos para probar redes y simular tráfico de datos sin comprometer la privacidad de los usuarios.
- **Energía y servicios públicos.** Las empresas de energía están empezando a generar datos sintéticos para simular redes eléctricas y predecir demandas futuras.

Sectores tradicionales avanzan con cautela, limitados por barreras estructurales y tecnológicas y menor urgencia digital

Late majority

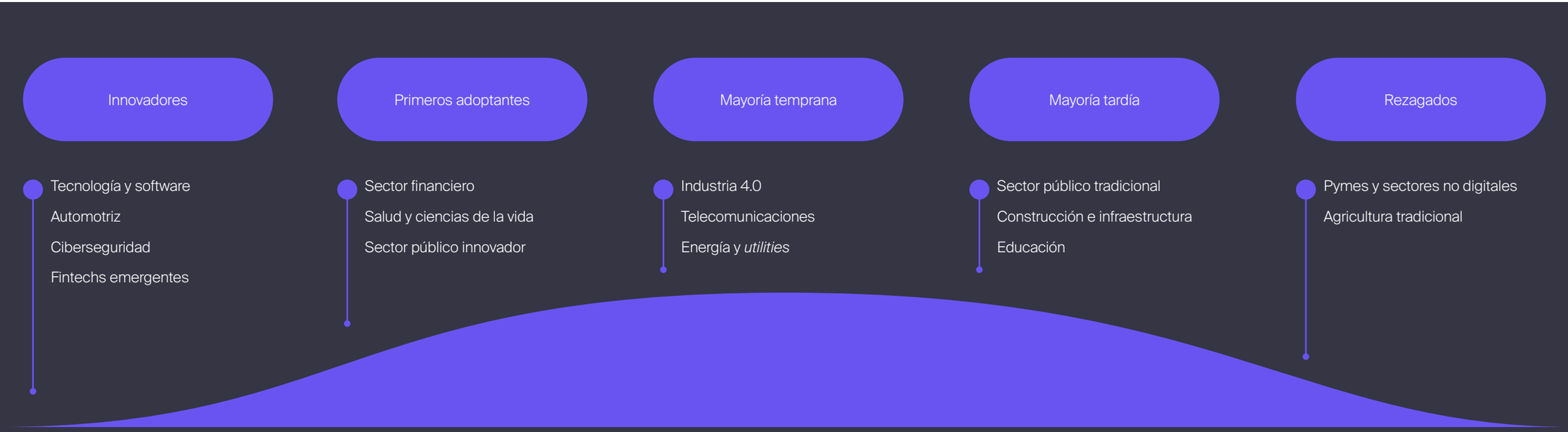
El sector público tradicional y la educación son ejemplos de sectores que adoptarán esta tecnología más tarde, **debido a su naturaleza más conservadora y la preferencia por esperar** que las soluciones sean más maduras y estándar. Estos sectores se centran en la regulación y la seguridad, por lo que la adopción de los datos sintéticos será más lenta, pero se espera que ocurra a medida que la tecnología madure y se convierta en una práctica más común.

Laggards

Por último, **los sectores tradicionales son los más rezagados** en la adopción de los datos sintéticos. Muchas pequeñas empresas no ven la necesidad inmediata de adoptar estos datos, ya que no cuentan con grandes volúmenes de datos para analizar ni con equipos internos especializados en IA. Sin embargo, a medida que la tecnología se convierta en una solución más accesible y se integren más herramientas de datos sintéticos en las plataformas tradicionales, estos sectores adoptarán la tecnología de manera más generalizada.

Barreras para los sectores rezagados

La falta de estándares para la creación de datos sintéticos, la incertidumbre regulatoria, y la resistencia cultural al cambio tecnológico como obstáculos importantes. **Las organizaciones que no han tenido una experiencia directa con las ventajas de los datos sintéticos suelen ser más cautelosas**, esperando que la tecnología se consolide aún más y que sus proveedores tradicionales la integren de forma más sencilla en sus soluciones.



En el sector tecnológico, los datos sintéticos aceleran innovación y pruebas, eliminando barreras legales y riesgos de privacidad reales

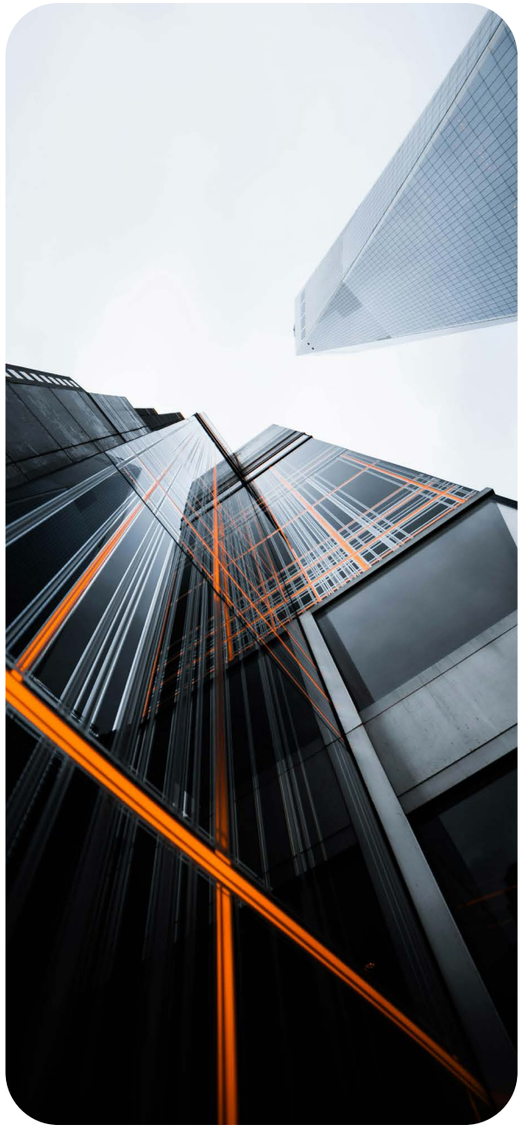
Los datos sintéticos están transformando el sector *tech*, ofreciendo soluciones innovadoras en áreas como análisis avanzado, IA/ML, desarrollo de software, ciberseguridad y simulación. A medida que las empresas enfrentan mayores desafíos en cuanto a la privacidad y la seguridad de **los datos, los datos sintéticos permiten mantener la utilidad de la información sin comprometer la protección de los individuos.**

Las empresas *tech* operan con ciclos rápidos de lanzamiento (CI/CD). Para ello, necesitan entornos de prueba robustos. Los datos sintéticos han revolucionado las prácticas de *testing* y QA al ofrecer *datasets* de prueba masivos y realistas bajo demanda. Antes, un equipo podía estar limitado, ahora puede generar cientos de miles de casos diversos, cubriendo condiciones raras. Mejorando así la calidad de los productos, reduciendo *bugs* que solo aparecerían con *inputs* poco comunes.

En análisis avanzado, los sintéticos permiten obtener información valiosa sin exponerse a riesgos de privacidad, ayudando a las organizaciones a obtener *insights* sin violar regulaciones. En IA y ML, **facilitan la creación de grandes volúmenes de datos de alta calidad para entrenar modelos más precisos y mejorar el rendimiento,** reduciendo costes y tiempo. En desarrollo de software, permiten realizar pruebas más realistas y eficaces, acelerando el ciclo de desarrollo y mejorando la fiabilidad de las aplicaciones.

En ciberseguridad, **los datos sintéticos se utilizan para entrenar sistemas de detección de amenazas sin exponer datos sensibles,** permitiendo crear simulaciones de ataques cibernéticos realistas. Además, en simulación y modelado, los datos sintéticos crean entornos virtuales realistas para realidad virtual, mejorando la experiencia del usuario y las *tech* (especialmente las que desarrollan IA) encuentran en ellos un gran aliado para entrenar modelos innovadores. Por ejemplo, empresas de software de conducción autónoma o empresas de desarrollo de software asistido por IA.

Curiosamente, dentro del sector *tech* han **surgido nuevas empresas cuyo propio producto son los datos sintéticos** (Mostly AI, Gretel, Hazy o Synthetaic). Estas ofrecen plataformas y APIs para que terceros generen datos.



En el sector salud, permiten entrenar IA médica y compartir datos clínicos sin comprometer la privacidad de los pacientes

La industria de **la salud está adoptando los datos sintéticos como una herramienta clave para transformar áreas como simulación de ensayos clínicos, desarrollo de medicamentos y diagnóstico asistido por IA.** Estos datos permiten trabajar con conjuntos representativos sin comprometer la privacidad del paciente.

En ensayos clínicos, los datos sintéticos simulan **poblaciones de pacientes, optimizando los diseños de los ensayos, reduciendo el tiempo de espera y disminuyendo la presión financiera de estos servicios**, por primera vez se pueden compartir datos de pacientes para entrenar IA sin revelar ninguna información de un paciente real. Por ejemplo, una empresa de análisis de bienestar ha usado cohortes sintéticas de control y en lugar de reclutar un grupo de control, generaron pacientes sintéticos con características similares a los del grupo de tratamiento para comparar resultados. También se utilizan para modelos de predicción de salud, simulando la progresión de enfermedades y mejorando los sistemas de monitoreo de pacientes mediante IA.

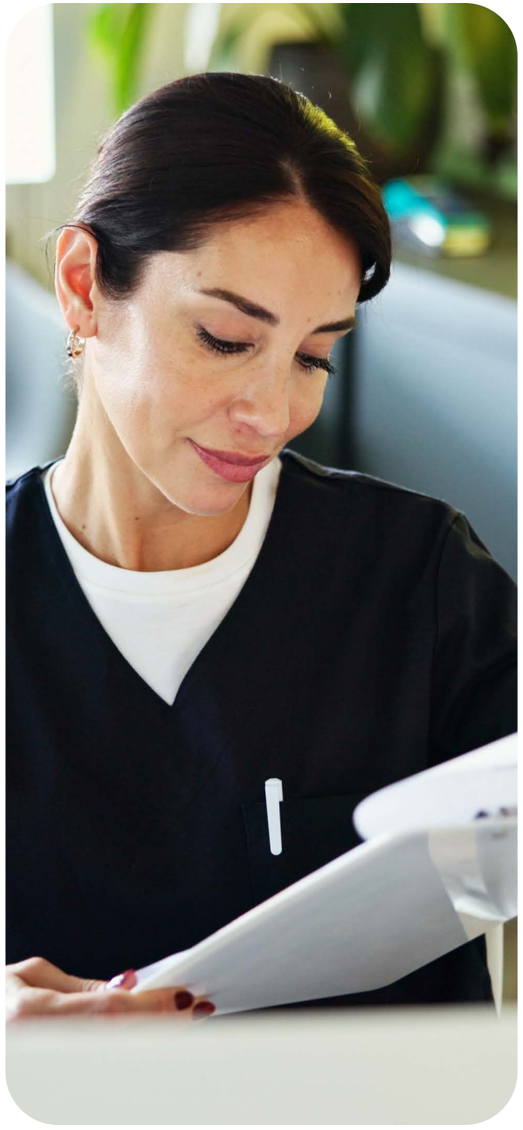
La comunidad médica académica sufre porque datos valiosos (expedientes de hospitales, registros epidemiológicos) a menudo no pueden ser compartidos libremente para análisis multicentro. Los datos sintéticos ofrecen la posibilidad de crear repositorios públicos sintéticos de datos de salud, un ejemplo es el UK Biobank, que tiene datos genéticos y clínicos de 500.000 personas y está explorando la posibilidad de generar un banco de datos sintéticos mostrando las mismas correlaciones genotipo-fenotipo, pero sin exponer a ningún individuo real.

En el área de imágenes médicas, **proporcionan imágenes y contenido multimedia realista para entrenar modelos diagnósticos y de detección temprana sin usar datos personales**, garantizando privacidad y mejorando la precisión de los diagnósticos. Además, estos datos facilitan la creación de modelos de riesgo de enfermedades, incluyendo poblaciones subrepresentadas y condiciones raras, lo que permite identificar pacientes en riesgo más temprano.

En desarrollo de medicamentos, **aceleran la prueba de tratamientos y simulan respuestas de pacientes a nuevas terapias,** reduciendo los gastos asociados con la investigación. En el ámbito de la salud pública, estos datos se utilizan para simular brotes de enfermedades y ayudar a la planificación de políticas sanitarias.

Un punto fundamental es asegurarse de que los datos sintéticos **no comprometen la precisión clínica.** Es decir, cualquier modelo o conclusión derivada de sintéticos debe reflejar la realidad médica. Hasta ahora, la evidencia sugiere que con técnicas adecuadas sí se logra.

El **30%** de todos los datos del mundo son datos de atención sanitaria y la cifra va en aumento.



Fuentes: LeewayHertz I Reply I Syntheticus I Syntho I V7

En el sector financiero, permiten simular fraudes, evaluar riesgos y testear modelos sin usar información confidencial de clientes reales

El sector financiero (bancos, aseguradoras, mercados de capitales) **es uno de los que más se está transformando con el uso de datos sintéticos,** aplicándolos a casos de detección de fraude, pruebas de estrés financiero y modelado de riesgos.

En la detección de fraudes, los datos sintéticos simulan transacciones fraudulentas, lo que mejora la precisión de los modelos de IA sin comprometer la privacidad del cliente. Visa, por ejemplo, realizó un estudio en 2024 donde generó datos sintéticos de comerciantes con comportamientos anómalos (como comercios ficticios que intentan estafar) y entrenó un modelo de detección. El resultado fue una mejora del 15% en la tasa de detección de fraudes sin utilizar ninguna transacción real de titulares de tarjeta. De forma similar, en la evaluación del crédito, se utilizan datos sintéticos para entrenar modelos de puntaje crediticio sin acceder a datos financieros reales, lo que permite una mayor equidad y precisión.

Otro ámbito es AML (*Anti-Money Laundering*). **Los bancos han empezado a usar datos sintéticos para simular redes de transacciones de lavado de dinero y así probar sus sistemas de monitoreo de transacciones.** Bancos globales mencionan que han podido mejorar la precisión de sus alertas AML ajustando parámetros sobre estos conjuntos sintéticos que replican patrones de estructuras de lavado (como "*smurfing*" o *transfers* fraccionadas), probando sin riesgo y afinando sin esperar a un caso real.

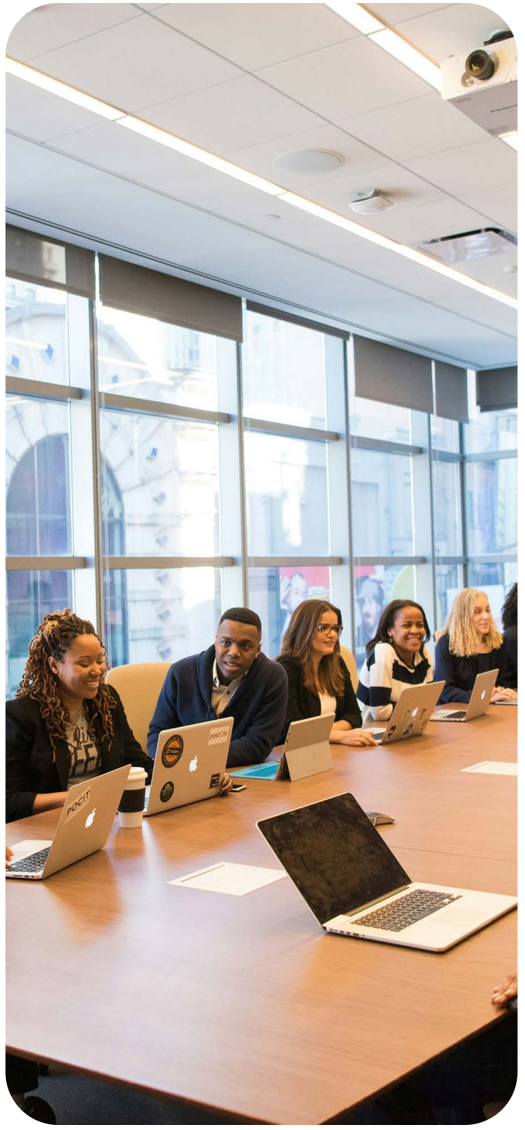
Los modelos de estrés y el análisis de escenarios son otras áreas donde los datos sintéticos resultan extremadamente útiles, permitiendo a las instituciones financieras generar datos que simulan fluctuaciones del mercado y escenarios adversos, ayudando a las organizaciones a probar la resistencia de sus carteras ante la volatilidad del mercado y el cumplimiento de regulaciones. Por ejemplo, generar un millón de clientes sintéticos con sus préstamos y simular que la tasa de desempleo sube al 15%, ¿cuántos impagarían?

Esto es diferente de solo hacer proyecciones agregadas, se puede modelar heterogeneidad individual, lo que da resultados más realistas.

Al crear modelos de *scoring* o de provisiones, a veces faltan suficientes datos de comportamiento en crisis. Los sintéticos rellenan ese hueco y **mejoran la robustez del modelo a eventos no vistos.**

Además, los datos sintéticos **facilitan la gestión de pruebas de datos en entornos de desarrollo y aceptación**, lo que reduce significativamente los tiempos de prueba y la carga operativa. También mejoran la innovación de procesos al permitir pruebas rápidas de nuevos flujos de trabajo sin comprometer datos sensibles.

50% de reducción en el esfuerzo del equipo de QA gracias a la generación de datos de prueba impulsada por IA, lo que permite ciclos de prueba más rápidos y menos defectos.

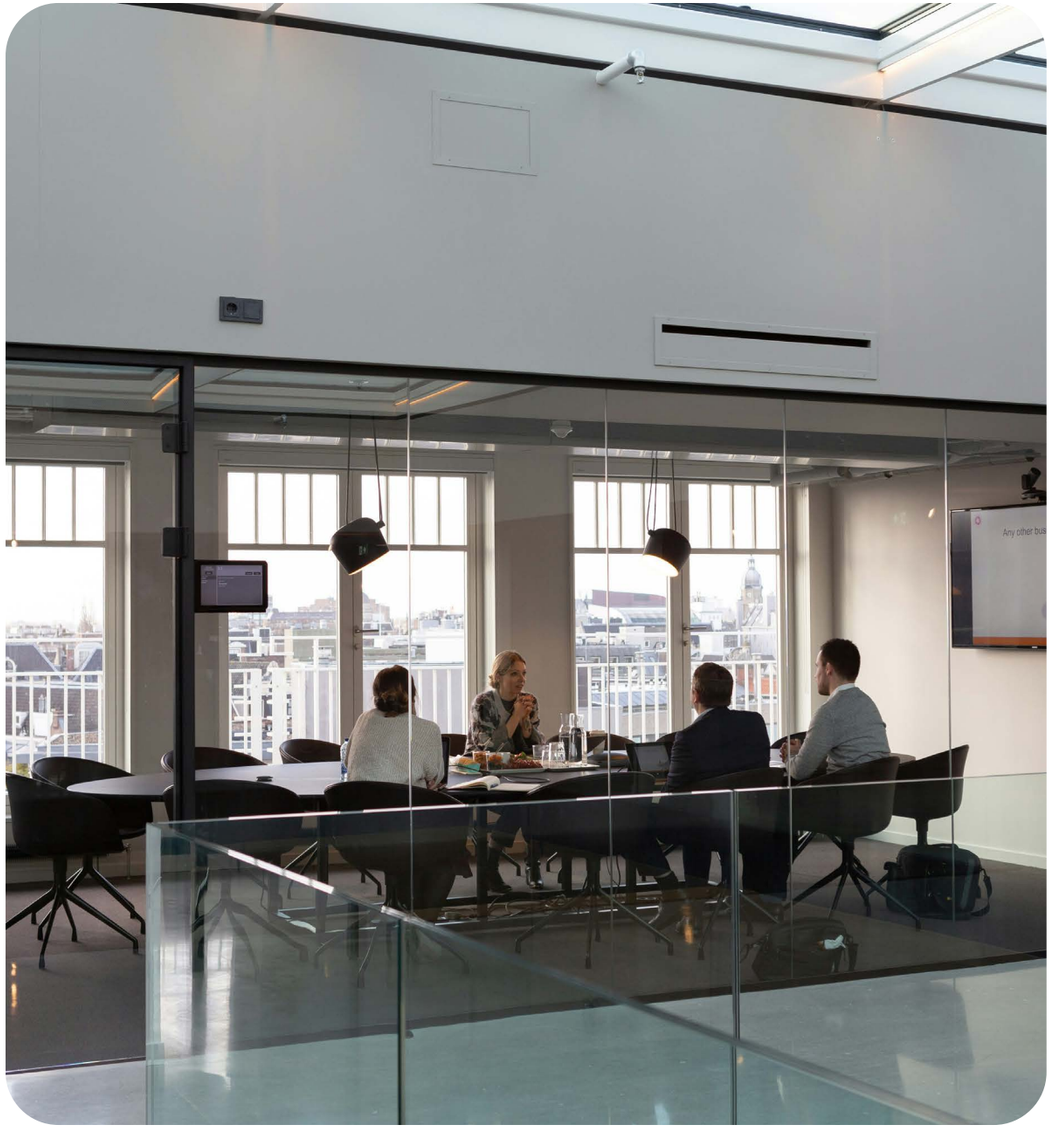


Fuente: Reply

En el sector asegurador, ayudan a evaluar siniestros, modelar escenarios extremos y personalizar ofertas sin exponer datos sensibles

El sector asegurador se está beneficiando enormemente del uso de datos sintéticos para optimizar sus operaciones, desde la evaluación de riesgos hasta la detección de fraudes y la gestión de reclamaciones. Estos **datos permiten a las aseguradoras crear modelos predictivos más precisos, simular eventos de alto riesgo, y cumplir con las regulaciones.** Uno de los mayores desafíos para las aseguradoras es la evaluación de riesgos y la tarificación de pólizas, áreas que se benefician enormemente esta data. Permiten simular perfiles de clientes, lo que mejora la precisión de los modelos de riesgo. Además, al simular escenarios catastróficos o de alto riesgo, las aseguradoras pueden anticipar y planificar mejor ante situaciones imprevistas, reduciendo su exposición a riesgos.

La detección de fraudes es otra área clave. Los datos sintéticos permiten crear **grandes volúmenes de datos de transacciones fraudulentas simuladas**, lo que mejora la eficacia de los modelos de detección sin exponer la información sensible de los clientes. Y, en cuanto a la gestión de reclamaciones, los datos sintéticos permiten realizar pruebas de escenarios sin la necesidad de utilizar datos reales. Las aseguradoras también pueden utilizar estos datos para mejorar la personalización de la experiencia del cliente, desarrollando productos y servicios que se adapten mejor a sus necesidades.



En *gaming* y entretenimiento, permiten generar perfiles y escenarios realistas, mejorando el diseño sin comprometer privacidad de usuarios

Dentro de la industria del videojuego y el entretenimiento, las tecnologías basadas en IA siempre han sido buenos impulsores. En el caso de los datos sintéticos, existen varias áreas donde es conveniente implementar una tecnología que **no solo mejora la eficiencia en los procesos de creación, sino que también potencia la experiencia del jugador y optimiza la creación de contenidos.** Uno de los desafíos más grandes en el desarrollo de videojuegos es crear personajes detallados y entornos realistas que mantengan la inmersión y la creatividad. Los sintéticos permiten a los desarrolladores generar personajes y escenarios artificiales que replican de manera precisa los aspectos deseados del mundo sin la necesidad de diseñarlos manualmente desde cero, lo que acelera considerablemente los tiempos de producción. Otro aspecto de la experiencia de juego es la interacción con oponentes inteligentes y desafiantes. El uso de estos datos para simular comportamientos de jugadores y oponentes permite entrenar a IA de oponentes sin necesidad de utilizar datos difíciles de recopilar. Todo ello facilita la creación de oponentes más realistas, que responden de manera dinámica y adaptable a las estrategias del jugador. El desarrollo de videojuegos también requiere de pruebas exhaustivas para detectar errores y mejorar el equilibrio del juego. Esta tecnología eleva y simplifica el proceso al simular escenarios de juego y generar comportamientos de jugadores sintéticos que permiten detectar problemas como fallos en el servidor o dificultades de jugabilidad. Por ejemplo, en un juego multijugador, miles de jugadores sintéticos pueden ser simulados para probar la carga del servidor, identificar cuellos de botella y ajustar el balance del juego.

Finalmente, **los niveles dinámicos y la personalización de la experiencia del juego** son cruciales en muchos títulos modernos y, mediante la generación de contenido procedimental con los sintéticos, los desarrolladores pueden crear niveles de juego adaptativos que cambian según la habilidad del juego, ofreciendo una experiencia personalizada y envolvente. Los datos sintéticos permiten crear una enorme variedad de entornos y situaciones que mantienen el interés del juego sin necesidad del diseño manual.

En el sector educativo, los datos sintéticos mejoran aprendizaje y en transporte, planifican movilidad respetando privacidad ciudadana

En el ámbito educativo los datos sintéticos pueden servir varios propósitos. Desde la investigación educativa, académicos que estudian rendimiento estudiantil, abandono escolar u otros factores, a menudo tienen datos limitados o dificultades para acceder a registros reales por privacidad. Con sintéticos, podrían **obtener *datasets* artificiales que replican características de alumnos, calificaciones o contextos socioeconómicos**, permitiendo probar hipótesis sobre, por ejemplo, cómo ciertas intervenciones afectarían diferentes tipos de alumnos.

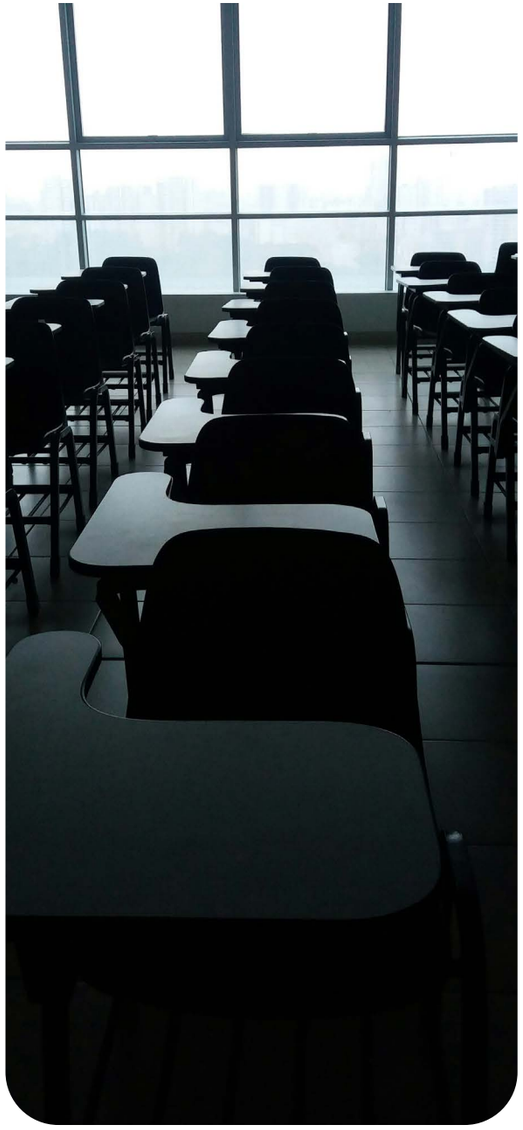
Empresas de tecnología educativa (plataformas de aprendizaje, sistemas de tutoría inteligente) pueden usar **datos sintéticos de interacción de estudiantes para entrenar sus algoritmos de recomendación o detección de dificultades**, sin necesidad de esperar a recopilar miles de horas de uso real. Por ejemplo, generar secuencias sintéticas de respuestas de alumnos a ejercicios basadas en patrones observados para preentrenar un sistema que identifica cuándo un alumno está estancado. Esto mejora estos sistemas antes de lanzarlos, con menos datos reales.

En el sector del transporte y la movilidad urbana los sintéticos se utilizan para generar datos sintéticos de movilidad, calibrados con conteos reales, para **probar cambios en infraestructura y planificar *smart cities*.** Por ejemplo, una ciudad puede crear 100.000 ciudadanos sintéticos con patrones de desplazamiento basados en encuestas de movilidad, luego simular que se cierra una vía principal o se introduce una nueva línea de metro, y observar cómo se rerutean los viajes sintéticos.

Esto mismo ayuda a planificadores **a predecir congestión o impacto en transporte público** sin tener que experimentar en la vida real primero. *Smart cities* como Singapur ya usan gemelos digitales con agentes sintéticos para tales fines.

50% de adopción de datos sintéticos en Finanzas y *Smart Cities* proyectado para 2026

Los organismos públicos de transporte pueden usar sintéticos **para evaluar políticas de seguridad**, por ejemplo, simular millones de viajes sintéticos de vehículos en una zona con distintos límites de velocidad para estimar accidentes, en vez de implementar y esperar años, o el transporte público puede simular pasajeros sintéticos en una red de buses para optimizar horarios o rutas sin afectar a pasajeros reales durante la experimentación.



Fuente: LinkedIn

En administración pública, mejoran la planificación y colaboración interinstitucional sin exponer datos personales protegidos

La implementación de datos sintéticos en el sector público representa una oportunidad para **mejorar la eficiencia, acelerar la toma de decisiones y garantizar la protección de datos sensibles.** Los gobiernos gestionan grandes volúmenes de datos confidenciales, y los datos sintéticos permiten crear conjuntos que conservan las características de los datos reales sin comprometer la privacidad.

También facilitan la **colaboración entre organismos públicos y privados**, al eliminar restricciones legales y de privacidad en el intercambio de datos. Esto acelera la innovación en políticas públicas, investigación y proyectos interinstitucionales. Un ejemplo claro es el uso de datos sintéticos en la detección de fraudes, donde se pueden generar datos sin exponer información real.

Además, los datos sintéticos permiten simular políticas públicas, como reasignación de presupuestos, proporcionando un análisis más preciso que las suposiciones tradicionales y pueden ser utilizados para crear entornos de formación segura, donde los funcionarios públicos puedan practicar con datos sin riesgos de exposición.
Es más, en situaciones de crisis o desastres, los gobiernos pueden **simular diferentes escenarios para optimizar su respuesta ante emergencias**, desde epidemias hasta desastres naturales.

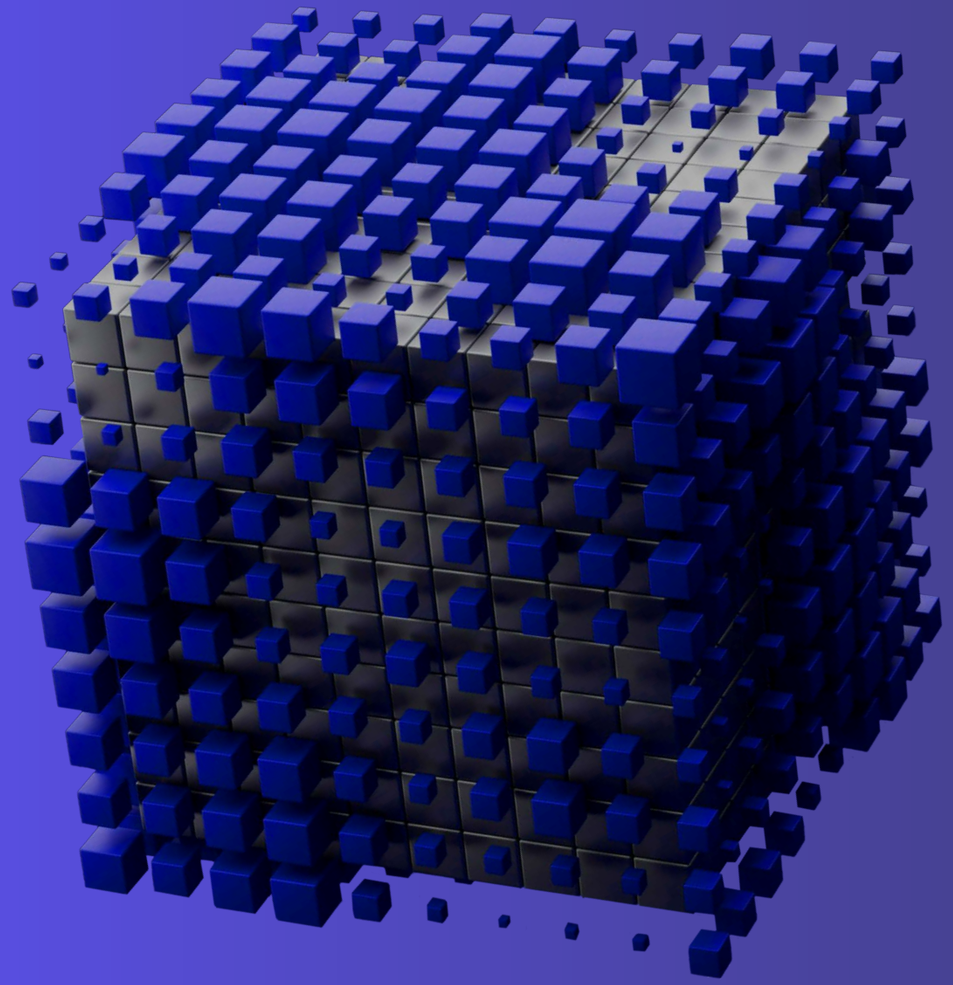
45% los lideres de TI del gobierno señalan la infraestructura de datos como una barrera para la digitalización

56% de las organizaciones públicas mencionaron el intercambio de datos y la privacidad como un desafío

Fuente: Reply

Todos los sectores están impulsados por privacidad, cumplimiento, innovación, reducción de costes, escalabilidad y mitigación de riesgos

Industria	Casos de uso	Impulsores
Sector tecnológico	Optimización de redes, ciberseguridad, pruebas de software, entrenamiento de IA	Implementación de 5G, *edge computing*, refuerzo de la seguridad
Salud	Ensayos clínicos, diagnóstico por imágenes, protección de datos del paciente, investigación farmacéutica	Cumplimiento de regulaciones, escasez de datos, aceleración en la investigación
Finanzas, seguros y reguladores financieros	Prevención de fraudes, modelado de riesgos, comercio algorítmico, auditorías de cumplimiento	Exigencias regulatorias, evaluación de riesgos en tiempo real
Gaming y entretenimiento	Desarrollo de videojuegos, experiencias inmersivas, simulación en AR y VR, optimización de gráficos, análisis de comportamiento de jugadores	Demanda de contenido interactivo, adopción de tecnologías inmersivas, plataformas de *streaming*, monetización de experiencias en línea
Educación	Aprendizaje adaptativo, plataformas de educación online, simulaciones educativas, evaluación en tiempo real, herramientas de colaboración virtual	Personalización de la educación, expansión de la enseñanza online, formación continua, tecnologías emergentes
Transporte y movilidad urbana	Simulación de vehículos autónomos, pruebas de escenarios extremos, calibración de sensores	Avances en conducción autónoma, validación de seguridad, optimización de costes
Gobierno y administración pública	Entrenamiento en simulaciones, análisis de inteligencia, seguridad pública, ciudades inteligentes	Protección nacional, privacidad de los ciudadanos, continuidad operativa
Retail* y *e-commerce	Modelado del comportamiento del consumidor, personalización de la experiencia, gestión de inventarios	Protección de la privacidad del cliente, análisis competitivo, expansión de mercado
Industria manufacturera	Gemelos digitales, control de calidad, mantenimiento predictivo, optimización de procesos	Industria 4.0, integración del Internet de las Cosas (IoT), mejora de la eficiencia operativa

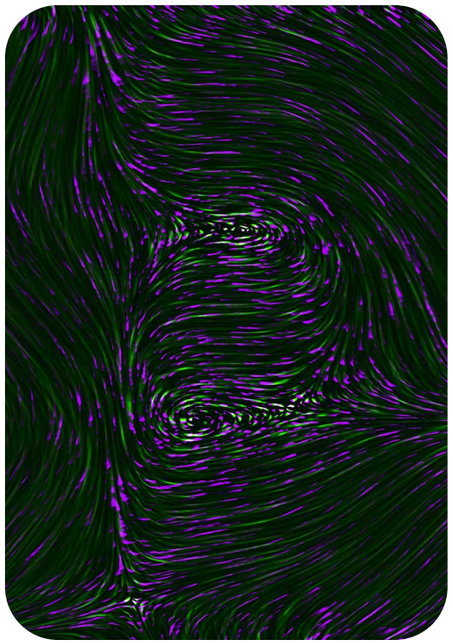


Conclusiones y predicciones: el futuro de los datos sintéticos

Los datos sintéticos son respuesta inmediata a los límites del dato real y, en los próximos años, dominarán la estrategia de datos corporativa

La promesa emergente de los datos sintéticos ya se ha convertido en **herramienta crítica en las estrategias de datos empresariales.** En los próximos 5 a 10 años, se prevé una adopción exponencial impulsada por las tendencias regulatorias que exigen protección de datos desde el diseño, la creciente necesidad de datos seguros y escalables para la IA y la analítica avanzada y las limitaciones prácticas del uso de datos reales y anonimización tradicional.

Las tendencias futuras en la adopción de datos sintéticos apuntan a un **panorama en el que las empresas y gobiernos se beneficien de la creación de modelos de inteligencia artificial más robustos, competitivos y ágiles.**



Qué esperar los próximos años

Ventaja competitiva empresarial
donde las organizaciones que dominan la generación de datos sintéticos innovan más rápido que las que dependen de datos reales

Soberanías y competitividad nacional,
con países como Corea del Sur y Singapur que utilizan estos datos para impulsar su competitividad global

Creación de mercados de datos sintéticos,
con empresas ofreciendo *data packs*, abriendo la posibilidad de intercambios globales ante situaciones de interés mundial

Resiliencia y continuidad de negocio,
facilitando la operación frente a disrupciones o ataques y agilizando la adaptación general al cambio

Innovación acelerada
en sectores como salud, tecnología y ciberseguridad o finanzas, donde pueden reducir los tiempos de experimentación, acelerando el desarrollo de nuevos productos

Posible nueva brecha
en el acceso a los mejores generadores de datos, lo cual podría estar dominado por las grandes corporaciones tecnológicas

Para transformar riesgos en oportunidades, las empresas deben fortalecer controles, calidad y talento en datos sintéticos

Riesgos asociados a los datos sintéticos y acciones prioritarias para las empresas

Aunque su adopción ofrece múltiples oportunidades, también presentan **riesgos que las empresas deben conocer y abordar** para maximizar los beneficios de esta tecnología sin comprometer su integridad.

Las empresas deben establecer **un enfoque proactivo para la adopción de datos sintéticos**, lo que incluye la inversión en tecnologías y modelos generativos, la colaboración con socios tecnológicos especializados y la creación de marcos internos de ética y gobernanza para el uso responsable de estos datos. Además, la creación de un entorno de pruebas controladas y simuladas permitirá una transición más fluida sin comprometer la privacidad ni la seguridad.

La desinformación y el mal uso de datos sintéticos pueden llevar a manipulaciones que generen desinformación o sesguen los resultados.	**Establecer marcos éticos claros** para su creación y uso y contar con procedimientos rigurosos de monitoreo de la calidad de los datos.
Las brechas en la calidad de los datos pueden llevar a que las empresas tomen decisiones equivocadas, lo que afectaría tanto sus operaciones como la experiencia del cliente.	**Regularizar monitoreos de calidad de datos** para que los modelos entrenados sobre estos no sean defectuosos y establecer un marco normativo interno, que regule cómo generar y evaluar los datos.
Los riesgos de seguridad y privacidad surgen porque tanto la generación como el procesamiento de estos datos pueden ser vulnerables a ciberataques o permitir la inferencia de información.	**Invertir en infraestructura tecnológica de vanguardia** para garantizar que las plataformas de datos sintéticos estén protegidas contra amenazas e implementación de políticas *privacy by design*.
La concentración de poder en grandes corporaciones tecnológicas puede llevarlas a dominar el acceso a esta tecnología, generando una nueva brecha.	**Fomentar la colaboración** con ecosistemas abiertos y promover la creación de plataformas accesibles donde todas las partes interesadas puedan participar.

Para 2030, la protección de datos exigirá gobernanza adaptativa, usando datos sintéticos para garantizar cumplimiento y confianza

El futuro de la privacidad: un cambio inminente

Estamos al borde de un cambio trascendental en la forma en que gestionamos la privacidad. A medida que el panorama regulatorio de la protección de datos continúa evolucionando, las leyes se están volviendo más estrictas, complejas y con mayores implicaciones. Mientras avanzamos hacia 2030, **el futuro de la privacidad se perfila como un espacio lleno de desafíos y cambios, pero también de oportunidades estratégicas.** Ya pudimos ver los primeros indicios de este cambio, el GDPR ha establecido un fuerte precedente, imponiendo sanciones millonarias a empresas que no cumplen con sus estrictos requisitos. Sin embargo, esto no es más que el comienzo.

Este nuevo enfoque no se detendrá en la protección de los datos personales, **las futuras normativas abarcarán áreas más amplias, incluyendo la ética de la IA y el uso responsable de tecnologías emergentes.** La privacidad no solo será una cuestión de datos personales, sino de cómo las organizaciones los usan para analizar y tomar decisiones. En el futuro, seremos testigos de un cambio hacia una “hiperregulación”, donde las organizaciones deberán estar más preparadas que nunca para cumplir con requisitos normativos cada vez más estrictos y complejos.

A medida que la regulación se vuelve **más global y más interconectada**, la gestión de datos no será solo una cuestión de cumplir con una única ley nacional, sino de navegar por un complejo laberinto de normativas multijurisdiccionales. Pero, aquellas que logren adaptarse de manera ágil a las regulaciones podrán aprovechar la transparencia y el cumplimiento como ventajas competitivas.

La soberanía de los datos se convierte en otro cambio cada vez más importante. La fragmentación de la información puede poner en peligro la innovación, ya que las empresas se verán limitadas al no poder acceder o compartir datos entre regiones. Pero no todo es negativo. **Las oportunidades surgen cuando las empresas se adaptan a este cambio con una estrategia localizada en cada jurisdicción**, demostrando su compromiso con las regulaciones nacionales.

Aquí es donde los datos sintéticos entran en juego como una **herramienta habilitadora para las empresas del futuro**, permiten a las organizaciones cumplir con los requisitos regulatorios y al mismo tiempo seguir innovando. Las entidades que los adopten estarán mejor posicionadas para navegar el, cada vez más, complejo mar de regulaciones y aprovechar las oportunidades que vienen con el cumplimiento.

5 pasos para comenzar con fuerza; el liderazgo exige institucionalizar los datos sintéticos para innovar con responsabilidad

5 Acciones inmediatas para prepararse

1. **Auditar el uso actual de datos sensibles**
   Identificar procesos críticos que dependan de datos personales y detectar riesgos legales
2. **Formar un equipo transversal de datos sintéticos**
   Incluir perfiles de IT, legal, *compliance* y analítica para coordinar la adopción responsable y efectiva
3. **Seleccionar un primer caso de uso piloto**
   Elegir un área de alto impacto y bajo riesgo (*testing*, I+D o analítica interna) para generar datos sintéticos
4. **Evaluar proveedores o tecnologías de generación sintética**
   Para reducir la presión sobre los recursos propios y acelerar tiempos
5. **Diseñar un marco ético y de gobernanza de datos sintéticos**
   Establecer principios base sobre transparencia, validación, trazabilidad y límites de uso

5 Acciones para liderar la tendencia

1. **Integrar datos sintéticos como parte de la estrategia de datos e IA**
   Considerarlos un pilar de la innovación, el cumplimiento y la escalabilidad
2. **Desarrollar capacidades internas de generación y validación avanzada**
   Invertir en talento, I+D y herramientas propias para evitar la dependencia a terceros
3. **Impulsar políticas públicas y estándares industriales**
   Participar activamente en foros y *sandboxes* regulatorios para dar forma a las normativas del sector
4. **Crear nuevas soluciones en base a datos sintéticos**
   Desarrollar simuladores, entrenar IA y plataformas de *testing* seguras para generar ventaja competitiva
5. **Convertirse en un caso de éxito de uso responsable**
   Publicar resultados y compartir aprendizajes, posicionarse como referente en privacidad, innovación y uso ético

Softtek®

softtek.com