mdt.

Ex Datis, Claritas

Si te quedas sin datos... invéntatelos

Cuando se acaban los datos para entrenar la IA hay tres salidas: pagar a expertos para crearlos, generarlos con otra IA o, en robótica, teleoperar y simular.

Manuel Delgado Tenorio · 4 de noviembre de 2025

Publicado originalmente en www.manueldelgado.com/blog/si-te-quedas-sin-datos-de-entrenamiento



En mi sección en el programa "No son horas" de Onda Cero, he hablado de los problemas a los que se enfrentan empresas como OpenAI o Anthropic para seguir *entrenando* a las futuras versiones de sus grandes modelos de lenguaje.

También, de cómo otra rama de la inteligencia artificial, la robótica, sufre ese mismo problema de falta de datos con los que entrenar a sus nuevos desarrollos.

Si no pudiste escuchar el programa en directo, puedes oírlo ahora como podcast en Spotify.

El problema: se acaban los datos

Un modelo de lenguaje aprende de texto escrito por personas, y ese texto no es infinito. Los investigadores de Epoch AI calculan que, si se mantiene el ritmo actual, los modelos se entrenarán con conjuntos de datos del tamaño de todo el texto humano público disponible entre 2026 y 2032. Ya hablé de ello en Vas a oír mucho sobre datos sintéticos: **se está acabando la materia prima.**

En robótica, el problema es todavía peor. Para el texto, al menos, existía internet. **No hay un internet de movimientos de robots:** nadie ha ido publicando durante treinta años vídeos de brazos mecánicos cargando lavavajillas. Esos datos, simplemente, no existen, y hay que producirlos uno a uno.

Tres maneras de inventarse los datos

1. Pagar a personas para que los creen

La opción más cara y más lenta, pero también la más fiable. OpenAI, por ejemplo, contrata a antiguos consultores de estrategia para que generen ejemplos de su trabajo con los que entrenar a sus modelos. **La ironía es evidente: esos expertos están enseñando a la máquina que aspira a sustituirlos.**

2. Que los genere otro modelo

Es lo que llamamos datos sintéticos aplicados al entrenamiento: un modelo ya entrenado escribe textos, ejemplos o ejercicios con los que se entrena otro. No es ciencia ficción. Microsoft entrenó phi-1, un modelo pequeño para programar, con textos de la web seleccionados por su calidad y con libros de texto y ejercicios generados por GPT-3.5, y obtuvo resultados comparables a los de modelos mucho más grandes. Hoy, **complementar los datos reales con datos sintéticos es práctica habitual** en toda la industria.

3. En robótica: teleoperar y simular

El robot doméstico Neo, de 1X, es un buen ejemplo de la primera vía, la teleoperación. El vídeo en el que carga un lavavajillas es espectacular, pero, como contaba Forbes, **el robot no es autónomo:** muchas de sus tareas las hace, de momento, una persona que lo maneja a distancia. Lo interesante es que cada una de esas sesiones es, a la vez, un servicio para el cliente y un dato de entrenamiento para la siguiente versión del robot.

La segunda vía es la simulación: plataformas como NVIDIA Cosmos ofrecen modelos del mundo capaces de generar escenas y situaciones físicas con las que entrenar robots sin tener que grabarlas una a una en el mundo real.

El riesgo: entrenar con lo que ya dijo la máquina

Inventarse los datos tiene un precio. Un estudio publicado en Nature en 2024 demostró que entrenar modelos de forma indiscriminada con contenido generado por otros modelos les causa defectos irreversibles: **desaparecen las colas de la distribución,** es decir, lo raro, lo poco frecuente, los matices. Generación tras generación, los modelos se vuelven más uniformes y pierden contacto con la realidad que pretendían describir. Los autores lo llaman *colapso del modelo*.

Por eso, los datos sintéticos funcionan como complemento de los datos reales, bien seleccionados y controlados, y no como sustituto. Es el mismo bucle que critiqué en Me estáis usando los datos sintéticos MAL: si lo único que haces es preguntarle a la máquina, **solo puedes encontrar lo que la máquina ya sabía.**

Qué significa esto para tu empresa

Si los datos reales escasean para los gigantes de la IA, **los tuyos valen más de lo que crees.** Lo que tus clientes hacen, compran y te cuentan es exactamente el tipo de dato que nadie puede descargar de internet. Cuídalo, mejora su calidad y piensa bien antes de cederlo a cambio de un servicio «gratuito».

Y si necesitas compartirlo o ampliarlo, los datos sintéticos tienen usos legítimos y muy potentes: proteger la privacidad al compartir datos con terceros, probar sistemas o aumentar conjuntos de datos pequeños. Lo que no pueden hacer es sustituir a tus clientes de verdad.

Los enlaces que mencioné en la radio

- Vídeo del robot Neo llenando un lavavajillas, en Reddit
- La web de 1X, la empresa fabricante del robot Neo
- Ejemplo de artículo donde se indica que Neo no es autónomo, en Forbes
- OpenAI usa a ex-consultores de estrategia para generar datos de entrenamiento, en Bloomberg



Sobre el autor

Manuel Delgado Tenorio

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

manueldelgado.com · LinkedIn · X

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