

SYNTHETIC DATA

ข้อมูลสังเคราะห์ที่กำลัง เปลี่ยนวงการวิจัย

คู่มือฉบับปี 2026 สำหรับนักวิจัยและนักสร้างนวัตกรรม



Paphichaya Amphawanwong
Demand Generation Specialist
Questionpro Thailand

Synthetic Data: ข้อมูลสังเคราะห์ที่กำลังเปลี่ยนวงการวิจัย

สิ่งที่นักวิจัยและนักธุรกิจต้องรู้ก่อนตัดสินใจใช้

ข้อมูลคือทรัพยากรพื้นฐานของการวิจัยและนวัตกรรมในปัจจุบัน แต่เมื่อมูลค่าของข้อมูลเพิ่มขึ้น ประเด็นด้านความเป็นส่วนตัวและความปลอดภัยก็ยิ่งสำคัญตามไปด้วย Synthetic Data หรือข้อมูลสังเคราะห์ จึงกลายเป็นหัวข้อที่ถูกพูดถึงมากที่สุดในการวิจัยตลาด เมื่อ AI เข้ามามีบทบาทในกระบวนการวิจัยมากขึ้น ทีมงานต่างมองหาวิธีค้นหาอินไซต์ได้เร็วขึ้นและลดต้นทุนในระยะสำรวจ

บทความนี้รวบรวมสาระสำคัญเกี่ยวกับ Synthetic Data ไว้อย่างครบถ้วนและเข้าใจง่าย

1. Synthetic Data คืออะไร

Synthetic Data คือข้อมูลที่ถูกสร้างขึ้นโดยอัลกอริทึม ไม่ใช่เก็บมาจากบุคคลหรือเหตุการณ์จริง แต่ถูกออกแบบให้มีรูปแบบทางสถิติและความสัมพันธ์ใกล้เคียงกับข้อมูลจริง โดยไม่มีข้อมูลส่วนบุคคลจริงหลงเหลืออยู่เลย

จุดที่ต้องแยกให้ออกคือ Synthetic Data **ไม่ใช่** การนำข้อมูลจริงมาดัดแปลงหรือปิดบัง (Data Masking) แต่ถูก **"สร้างขึ้นใหม่ทั้งหมด"** ด้วยเทคนิคอย่าง Machine Learning หรืออัลกอริทึมที่อิงกฎเกณฑ์

คุณลักษณะสำคัญ 3 ประการ

1. **สร้างขึ้นด้วยอัลกอริทึม** — สืบย้อนกลับไปหาตัวบุคคลไม่ได้
2. **คล้ายคลึงเชิงสถิติ** — คงแนวโน้มเดียวกับข้อมูลต้นฉบับ จึงยังใช้วิเคราะห์ได้จริง
3. **ปกป้องความเป็นส่วนตัว** — ลดความเสี่ยงทั้งด้านกฎหมายและจริยธรรม



ตัวอย่าง: โรงพยาบาลต้องการแบ่งปันข้อมูลผู้ป่วยให้ทีมวิจัยภายนอก แต่ติดกฎหมายความเป็นส่วนตัว การสร้างระเบียบผู้ป่วยสังเคราะห์ที่สะท้อนแนวโน้มเดียวกับข้อมูลจริง จึงช่วยให้วิเคราะห์และทดสอบสมมติฐานได้ โดยไม่กระทบความเป็นส่วนตัวของผู้ป่วย

2. บทบาทในกระบวนการวิจัย และการเปรียบเทียบ

Synthetic Data ไม่ได้มีเจตนาแทนที่งานวิจัยแบบดั้งเดิม แต่นักวิจัยบางกลุ่มมองว่าเป็น **"มิติที่สาม"** ของงานวิจัย เพิ่มจากเชิงคุณภาพและเชิงปริมาณ ทำหน้าที่เป็นขั้นตอนสำรวจก่อน ลงทุนกับการศึกษาเต็มรูปแบบ ช่วยทดสอบว่าสมมติฐานคุ่มค่าศึกษาต่อหรือไม่ จำลองการ สัมภาษณ์ผู้เชี่ยวชาญเบื้องต้น และปรับปรุงคำถามแบบสำรวจก่อนใช้จริง

ข้อควรจำ: ผลลัพธ์จาก Synthetic Data ควรมองเป็น **ทิศทางบ่งชี้ ไม่ใช่หลักฐานสุดท้าย**

นอกจากนี้ยังช่วยจำลองสถานการณ์ที่หายากหรืออ่อนไหว ซึ่งเก็บข้อมูลจริงได้ยาก เช่น ทีมวิจัย มหาวิทยาลัยที่ศึกษาสุขภาพจิต ใช้ข้อมูลสังเคราะห์ปกป้องตัวตนผู้เข้าร่วม ทำให้แชร์ชุดข้อมูลกับ นักวิจัยทั่วโลกได้โดยไม่เสี่ยงละเมิดความเป็นส่วนตัว

เทียบกับข้อมูลจริงและ Data Masking

เกณฑ์	ข้อมูลจริง	Data Masking	Synthetic Data
แหล่งที่มา	เหตุการณ์/ ผู้ใช้จริง	ข้อมูลจริงที่ถูกดัดแปลง	สร้างขึ้นใหม่ทั้งหมด
ความเสี่ยงความเป็นส่วนตัว	สูง	ลดลง แต่ยังมีช่องโหว่	ต่ำมาก
ความยืดหยุ่นในการใช้งาน	จำกัด	จำกัดตามวิธีปิดบัง	สูง

สรุป: งานที่ต้องการความแท้จริงระดับสมบูรณ์ เช่น รายงานตามกฎระเบียบ ยังต้องใช้ข้อมูลจริง แต่เมื่อความเป็นส่วนตัวเป็นอุปสรรค Synthetic Data คือทางเลือกที่ยืดหยุ่นกว่ามาก

3. สร้างขึ้นมาได้อย่างไร

แบ่งเป็น 3 แนวทางหลัก

- **อิงกฎเกณฑ์ (Rule-based):** ใช้สูตรที่กำหนดไว้ล่วงหน้า ตรงไปตรงมาแต่ขาดความซับซ้อนแบบข้อมูลจริง
- **แบบจำลองทางสถิติ:** วิเคราะห์ค่าเฉลี่ยและความสัมพันธ์ของข้อมูลจริง แล้วสร้างข้อมูลใหม่ให้ตรงสถิตินั้น
- **Machine Learning ขั้นสูง:** ใช้ **GANs** (โครงข่ายสองส่วนแข่งกันเรียนรู้จนข้อมูลสมจริงมาก) และ **VAEs** (เข้ารหัส-ถอดรหัสข้อมูลเพื่อสร้างตัวอย่างใหม่ที่ยังคงความซับซ้อนไว้)

กระบวนการ 5 ขั้นตอน: วิเคราะห์ข้อมูลจริง → ฝึกโมเดล → สร้างข้อมูลใหม่ → ตรวจสอบ เทียบข้อมูลจริง → นำไปใช้งาน

ความท้าทาย: ต้องหาสมดุลระหว่างความสมจริงกับความเป็นส่วนตัว ข้อมูลที่คล้ายต้นฉบับมากเกินไปอาจเผยข้อมูลอ่อนไหวโดยไม่ตั้งใจ (Overfitting) ส่วนข้อมูลซับซ้อนอย่างภาพ เสียง หรือข้อความ ก็ยากต่อการสังเคราะห์และใช้ทรัพยากรมาก

4. ข้อจำกัด และเหตุผลทางธุรกิจ

ข้อจำกัดที่ต้องรู้ก่อนใช้งาน

- อาจสูญเสียรายละเอียดปลีกย่อยที่หายาก ซึ่งบางครั้งคือสิ่งมีค่าที่สุดในงานวิจัย
- เสี่ยงสืบทอดอคติจากข้อมูลต้นฉบับ หากไม่มีการตรวจสอบระหว่างสร้าง
- ไม่เหมาะกับงานที่ต้องการความแม่นยำระดับสมบูรณ์ เช่น หลักฐานทางกฎหมาย หรือการวัดผลขั้นสุดท้าย

ข้อเสนอแนะ: ใช้เป็น "ตัวเสริม" คู่กับการตรวจสอบเทียบข้อมูลจริงอยู่เสมอ ไม่ใช่ตัวแทนทั้งหมด

เหตุผลทางธุรกิจ 4 ข้อ

1. **แฮร์ข้อมูลข้ามองค์กร/ข้ามประเทศได้** โดยไม่ติดกำแพงกฎหมายความเป็นส่วนตัว
2. **เข้าถึงกลุ่มตัวอย่างหรือเหตุการณ์ที่หายากได้** โดยไม่ต้องรอเหตุการณ์จริงหรือเสี่ยงผิดจริยธรรม
3. **ลดต้นทุนและเวลา** ในการทดสอบสมมติฐานเบื้องต้น ก่อนลงทุนเก็บข้อมูลจริงเต็มรูปแบบ
4. **ลดความเสี่ยงด้านกฎหมาย** ในยุคที่ GDPR, PDPA เข้มงวดขึ้นทุกปี

สรุปบทความ

Synthetic Data คือข้อมูลที่สร้างขึ้นด้วยอัลกอริทึมให้เลียนแบบข้อมูลจริงทางสถิติ โดยไม่มีข้อมูลส่วนบุคคลจริงหลงเหลืออยู่ ช่วยให้องค์กรแฮร์ข้อมูลข้ามพรมแดน ศึกษาสถานการณ์ที่หายาก ลดต้นทุน และลดความเสี่ยงทางกฎหมาย แต่ยังมีข้อจำกัดชัดเจนทั้งเรื่องอคติและความไม่เหมาะกับงานที่ต้องการความแม่นยำสมบูรณ์ ทางที่ดีที่สุดคือใช้เป็นตัวเสริมงานวิจัยเชิงคุณภาพและเชิงปริมาณ ไม่ใช่สิ่งทดแทนทั้งหมด