REVOLUTIONIZING DATA DRIVEN INNOVATION: SYNTHETIC DATA

THE BENEFITS AND CHALLENGES OF SYNTHETIC DATA

Cecilia Dones :: October 2024

3 Standard Deviations

BOTTOM LINE UP FRONT

EXPLORE SYNTHETIC DATA TO DRIVE INNOVATION WHILE CAREFULLY BALANCING PRIVACY, COMPLIANCE, AND DATA INTEGRITY.

WHAT YOU WILL TAKE WITH YOU TODAY

Definition of Synthetic Data

Pros/Cons + Trade-Offs

Industry Applications

CECILIA DONES :: SHORT INTRODUCTION

Founder, 3 Standard Deviations

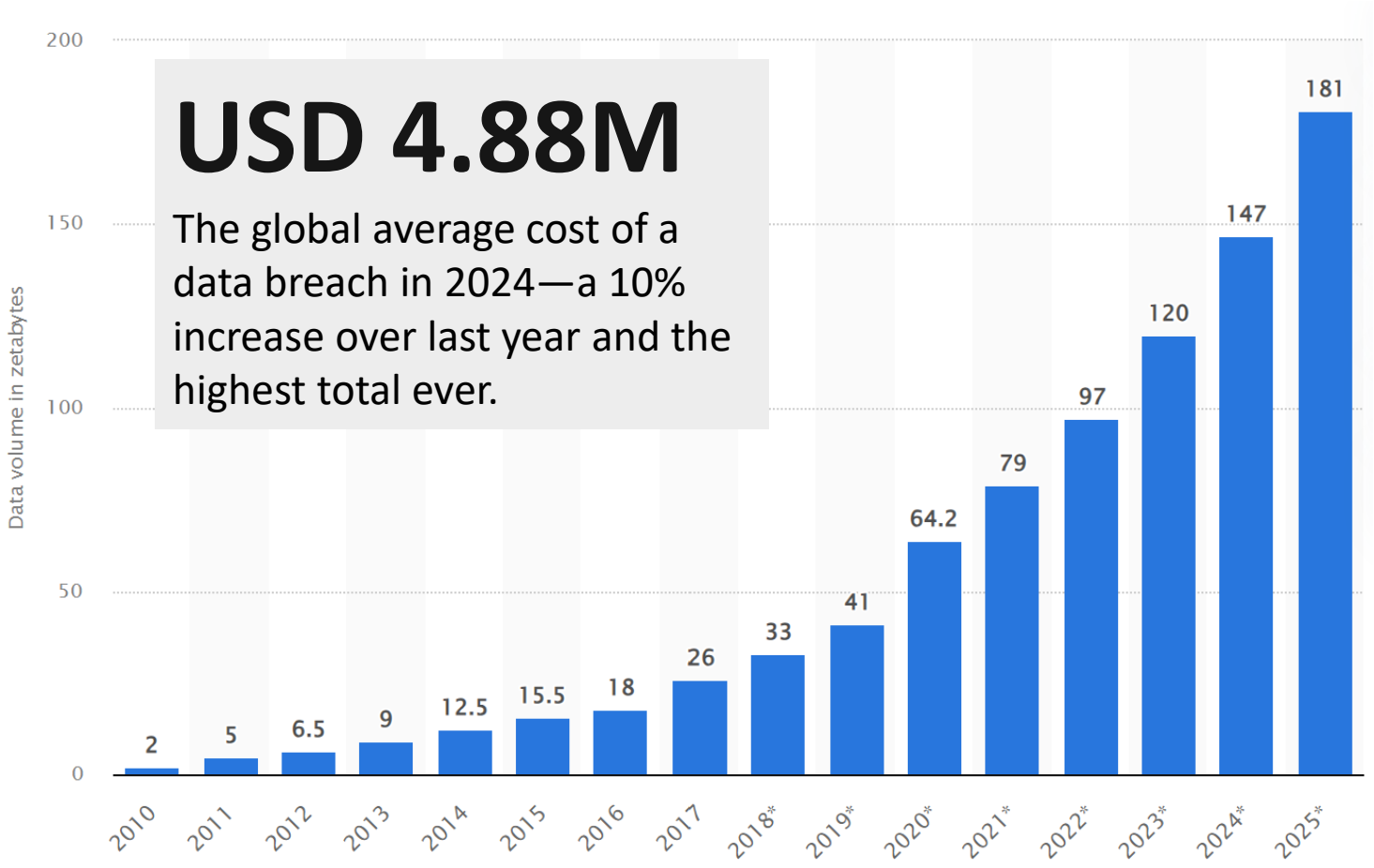
15+ Years in Industry, Researcher, Practitioner, Educator, Academic

Specialties: Qualitative, Quantitative Methods, Marketing, Interdisciplinary, Pragmatic Solutions For Businesses Trying To Leverage Data & AI


Coca-Cola
classic


Coca-Cola
zero
REAL
Coca-Cola
TASTE
AND ZERO
CALORIES

DATA IS COMPLEX AND CONSEQUENTIAL



USD 4.88M

The global average cost of a data breach in 2024—a 10% increase over last year and the highest total ever.

CHALLENGES

- Increasing demand for insights
- Stricter privacy regulations
- Limitations of traditional anonymization

LEARN MORE: Statista. (2021). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/; IBM. (2023). Cost of a data breach report 2023. https://www.ibm.com/reports/data-breach


Coca-Cola
zero
REAL
Coca-Cola
TASTE
AND ZERO
CALORIES


ZERO SUGAR
ZERO CALORIE COLA


ZERO SUGAR
ZERO CAFFEINE
ZERO CALORIE COLA


Cherry Vanilla
ZERO SUGAR
FLAVORS
ZERO CALORIE COLA
Cherry Vanilla Flavored
& Other Natural Flavors


Cherry
ZERO SUGAR
FLAVORS
ZERO CALORIE COLA
Cherry Flavored
& Other Natural Flavors


Vanilla
ZERO SUGAR
FLAVORS
ZERO CALORIE COLA
Vanilla Flavored
& Other Natural Flavors

Fully Synthetic Data

Data created entirely by algorithms, without direct one-to-one mapping to real-world data points.

Example: Generating a complete set of fictional customer profiles for a non-existent e-commerce platform, including demographics, purchase history, and browsing behavior.

Partially Synthetic Data

Real dataset where some variables are replaced with synthetic values.

Example: A hospital dataset where patient names and addresses are replaced with synthetic ones, but medical conditions and treatments remain real.

Hybrid Synthetic Data

A mix of real and synthetic data points, where synthetic data is used to augment or expand real data.

Example: A financial dataset that contains real transaction data for existing products, combined with synthetic data for hypothetical new products or market scenarios.

Simulated Data

Data generated from a model designed to replicate specific real-world processes or systems.

Example: Climate model outputs that simulate future weather patterns based on current climate data and physical models.

TRADEOFFS IN REAL VS. SYNTHETIC DATA

	Real Data	Synthetic Data
Privacy Risk	High	Low
Data Availability	Limited	Abundant
Compliance Complexity	High	Low
AI/ML Model Training Efficiency	Constrained by data limits	Enhanced by data generation
Data Authenticity	High	Variable
Anomaly Representation	Naturally occurring	May be underrepresented
Unforeseen Insights	Possible	Limited
Real-world Accuracy	High	Dependent on generation quality

STRATEGIC IMPLICATIONS FOR BUSINESS

- Accelerating Innovation Cycles
- Enabling Cross-Border Data Sharing
- Improving Data Democratization
- Enhancing AI/ML Development

IMPLEMENTATION CHALLENGES AND SOLUTIONS

Challenge	Solution	Key Considerations
Data quality concerns	Advanced validation techniques	- Implement statistical validation methods - Use machine learning for anomaly detection - Establish data quality metrics and thresholds - Conduct regular data audits
Infrastructure integration	Phased approach and APIs	- Start with non-critical systems - Develop robust API architecture - Ensure scalability and compatibility - Implement strong security measures
ROI measurement	Defined KPIs and benchmarking	- Establish clear baseline metrics - Define both short-term and long-term KPIs - Conduct regular performance reviews - Use A/B testing for comparative analysis
Organizational resistance	Education and pilot programs	- Develop comprehensive training programs - Start with small-scale pilot projects - Showcase early wins and success stories - Foster a data-driven culture

INDUSTRY APPLICATIONS

Financial Services

Fraud detection models

AI-powered anomaly detection in transactions

Healthcare

Clinical trial simulations

Drug development through virtual patient populations

Retail

Customer Behavior Models

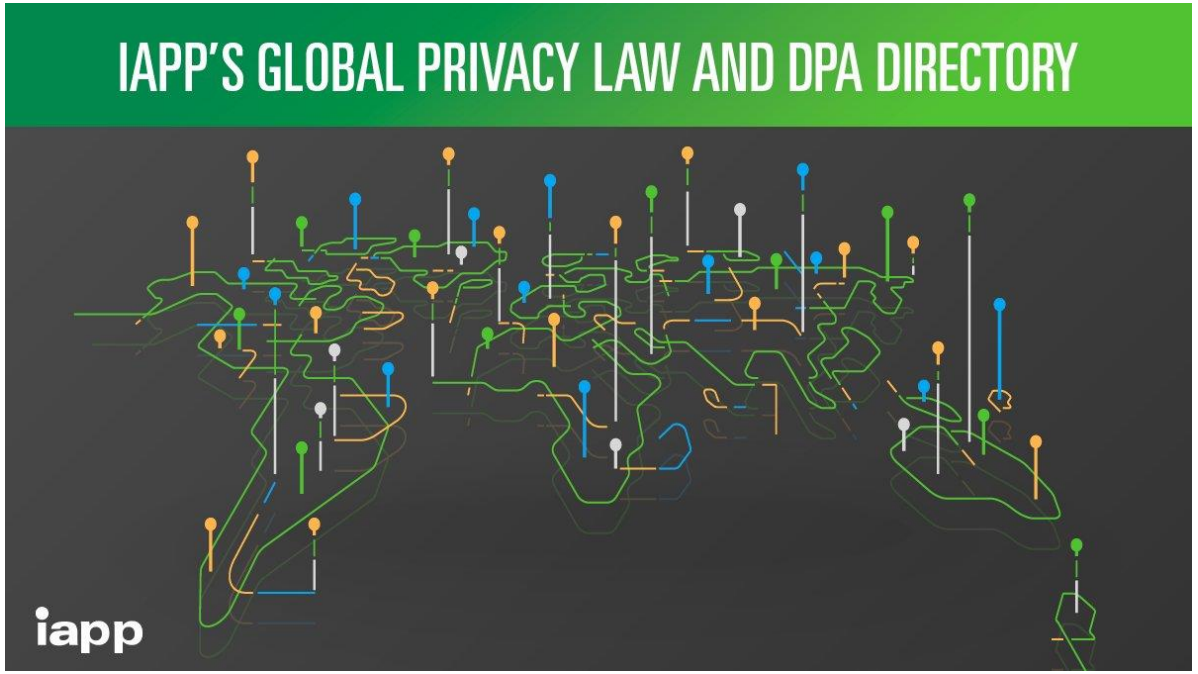
Predicting trends and personalizing experiences

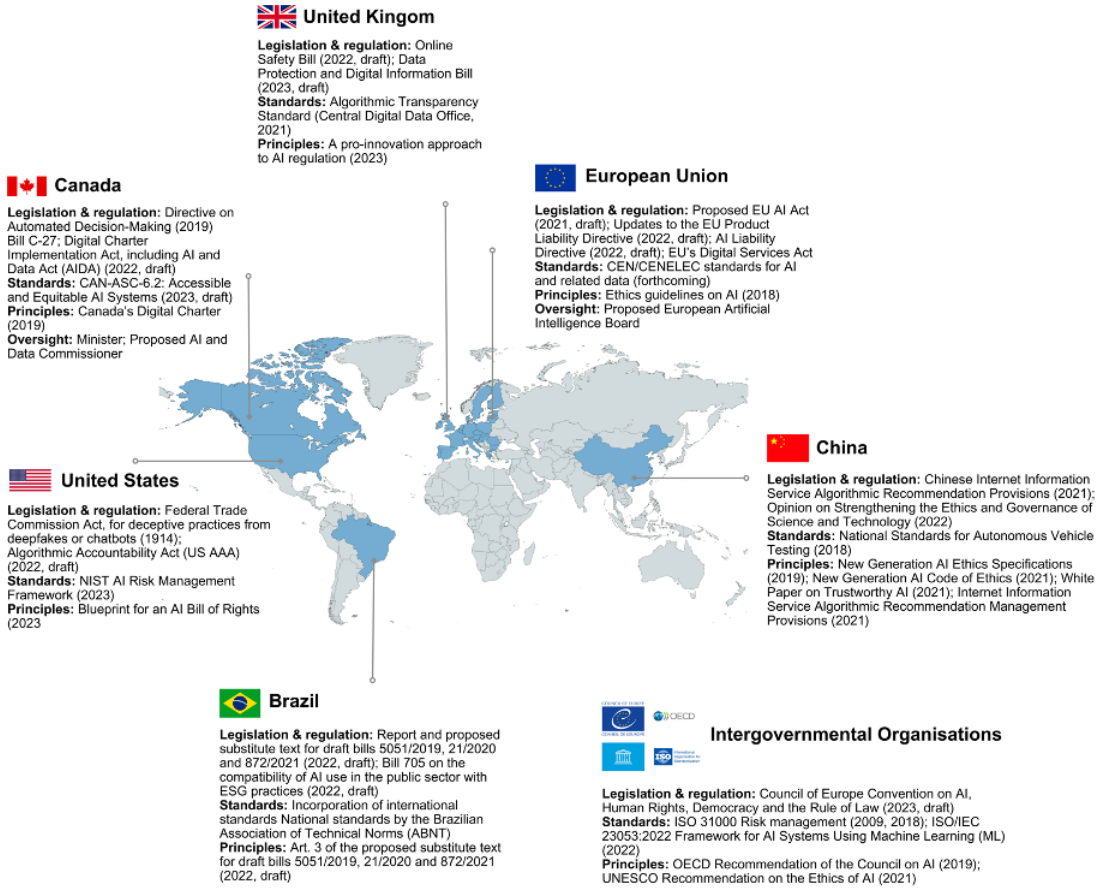
Manufacturing

Supply Chain Optimization

AI-driven efficiency in production and logistics

PRIVACY REGULATION & TRUSTWORTHY AI





LEARN MORE: International Association of Privacy Professionals. (2024). Global privacy law and DPA directory. https://iapp.org/resources/global-privacy-directory/; Organisation for Economic Co-operation and Development. (2023). National policies. https://oecd.ai/en/wonk/national-policies-2

THE FUTURE OF SYNTHETIC DATA

Integration with Federated Learning

- Decentralized AI training with privacy-preserving synthetic data

Advancements in Generative AI for Data Creation

- Next-gen AI creating hyper-realistic synthetic datasets

Synthetic Data Marketplaces

- Democratizing access to high-quality, diverse datasets

WHAT YOU CAN DO TODAY

Evaluate potential in your organization

- Identify high-impact use cases for synthetic data

Start with a pilot project

- Test the waters with a low-risk, high-reward initiative

Collaborate with synthetic data experts

- Leverage expertise to accelerate your journey

CAN OR FOUNTAIN?

WHAT YOU WILL TAKE WITH YOU TODAY

Definition of Synthetic Data

Pros/Cons + Trade-Offs

Industry Applications

EXPLORE SYNTHETIC DATA TO DRIVE INNOVATION WHILE CAREFULLY BALANCING PRIVACY, COMPLIANCE, AND DATA INTEGRITY.

THANK YOU

IF YOU WOULD LIKE TO COLLABORATE WITH ME, PLEASE FEEL FREE TO FIND ME ON LINKEDIN OR EMAIL ME:

CECILIA.DONES@3STANDARDDEVIATIONS.COM