Synthetic Tabular Data and Goodness Measure

Jayanthi Suryanarayana May 2023

Reduce privacy risk, Accelerate AI Innovation, Improve software quality

Bio

Name : Jayanthi Suryanarayana

Experience : Tech, Retail and healthcare

Title: Developer, Architect, Senior Principal Engineer

What do I do: Build data platforms for enterprise, Passionate about the use of synthetic data to enable people who work with data (analysts, engineers, scientists anyone who wants to get value out of data)

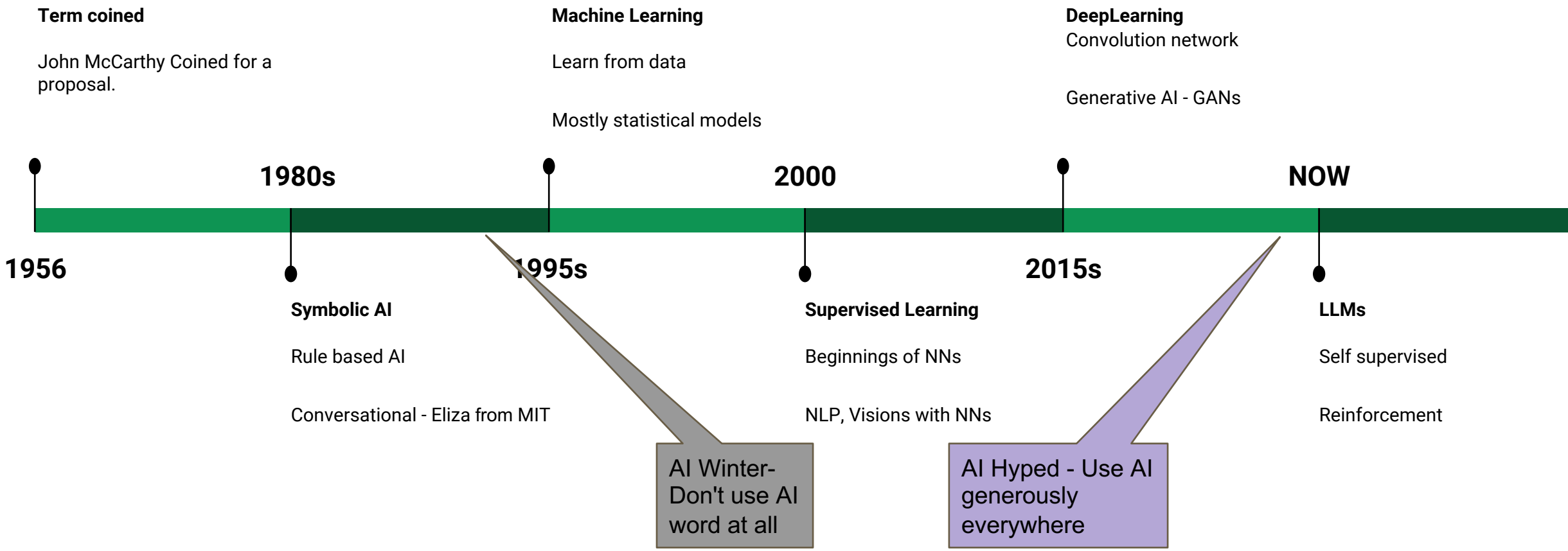
Journey:
Electronics engineer by training, pivoted to software engineering. Understand the nuances of app, data and ML engineering. Been in data/ML space for about 7 years now and discovered synthetic data in my DL/ML Journey primarily in healthcare

Outside of work:
Enjoy time with family, cook and do yoga

Agenda

1. AI and Generative Technologies - a Timeline
2. Synthetic Tabular Data - Context and Intuition
3. Synthetic data – Use cases
4. How to - Tools, technology and platforms
5. Goodness Measure - Concepts
6. Q & A

AI - and its implied meaning - a timeline view



Synthetic data generation Techniques

I

Statistical Models

Probability Distribution Function

Copulas

II

GANs

Generative Adversarial Networks

III

LLMs

Pre trained Transformers

GPT + Reinforcement

Synthetic Tabular Data – Intuition and Context

Tabular data

Data that is organized in rows and columns, similar to a spreadsheet or database table. Examples of tabular data include sales records, customer information, health data, financial data, and scientific measurements. . Most **enterprise data** in this form

Conceptually

Synthetic data is data generated from real data that is Structurally and Statistically similar.

For an intuition:

An analyst running queries against the original data should get comparable results when running it on synthetic data

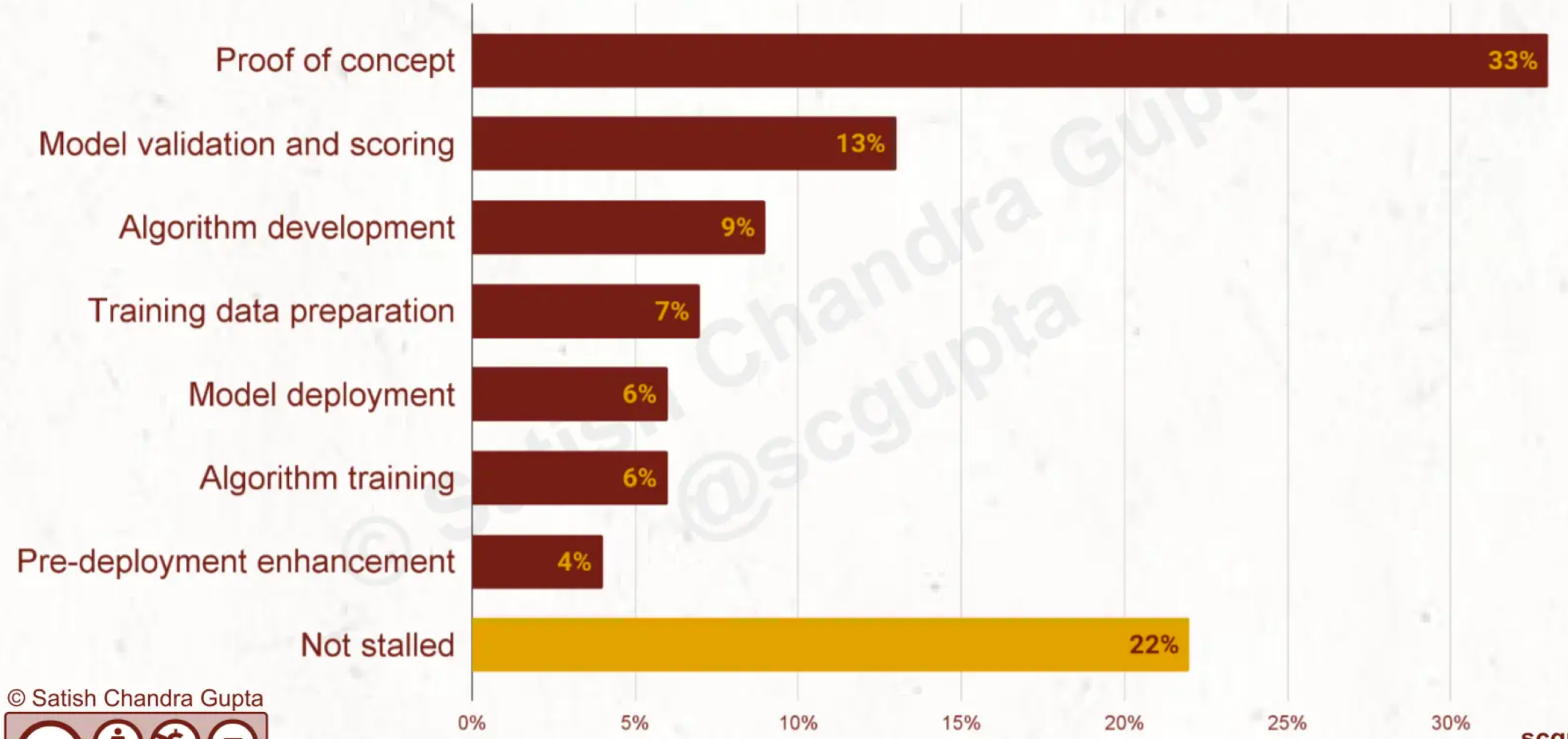
.

Utility Metrics

A set of measurements run against real and synthetic data to understand how good the synthetic data is in comparison to real data. e.g., Similar meta data for synthetic data,Distance measures, Goodness of fit tests, Privacy metrics

78% of AI or ML Projects Stall at Some Stage Before Deployment

Source: Dimensional Research - Alegion Survey. https://content.alegion.com/dimensional-researchs-survey



scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta

Data Breach statistics - health care

As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals. The following breaches have been reported to the Secretary:

https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

Average Healthcare Data Breach Costs Surpass $10M, IBM Finds

https://healthitsecurity.com/news/average-healthcare-data-breach-costs-surpass-10m-ibm-finds

Synthetic Data – Top 3 Use cases

Reduce Privacy Risk

Reduce the likelihood that individuals will experience problems from data processing and the impact should they occur. It is a risk mitigation approach.

Accelerate Enterprise AI

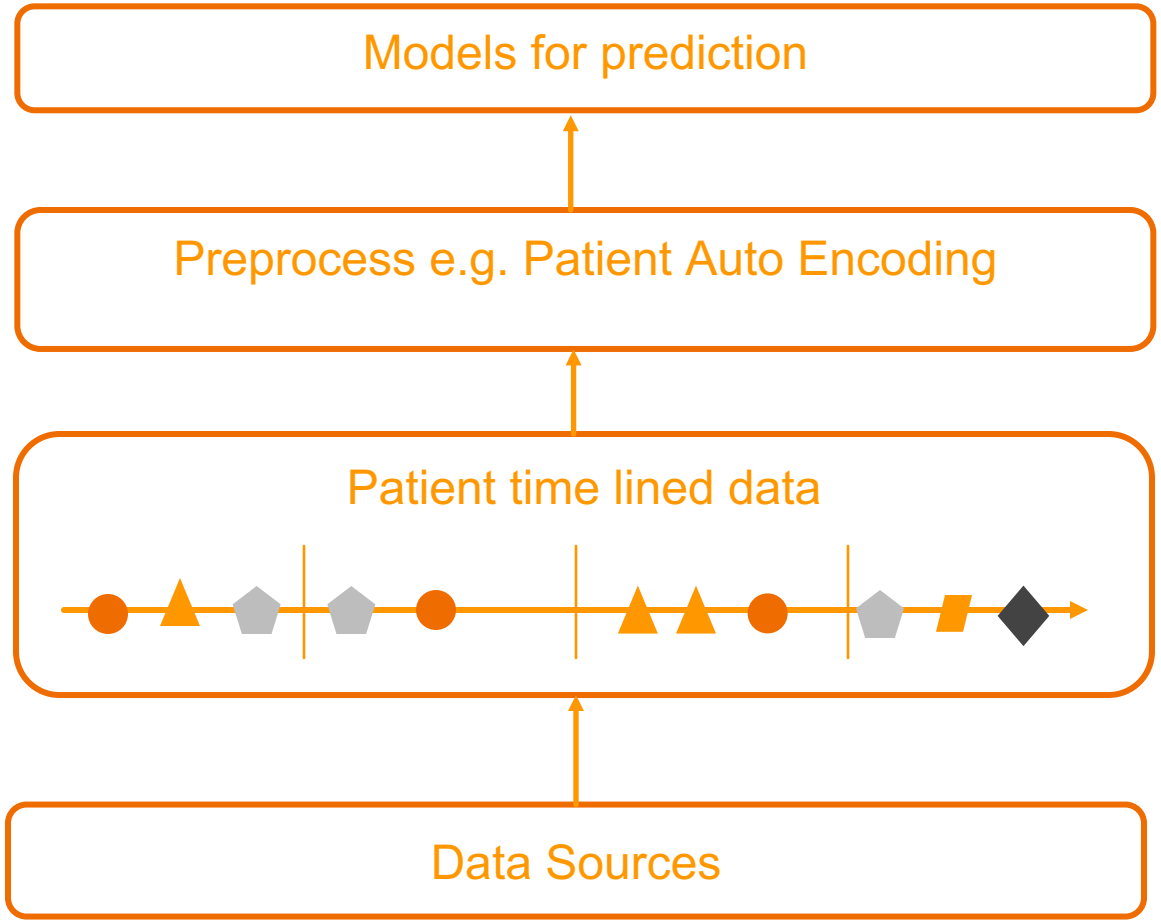
Enable acceleration for data discovery, talent development and data augmentation for model development.

Improve Software Quality

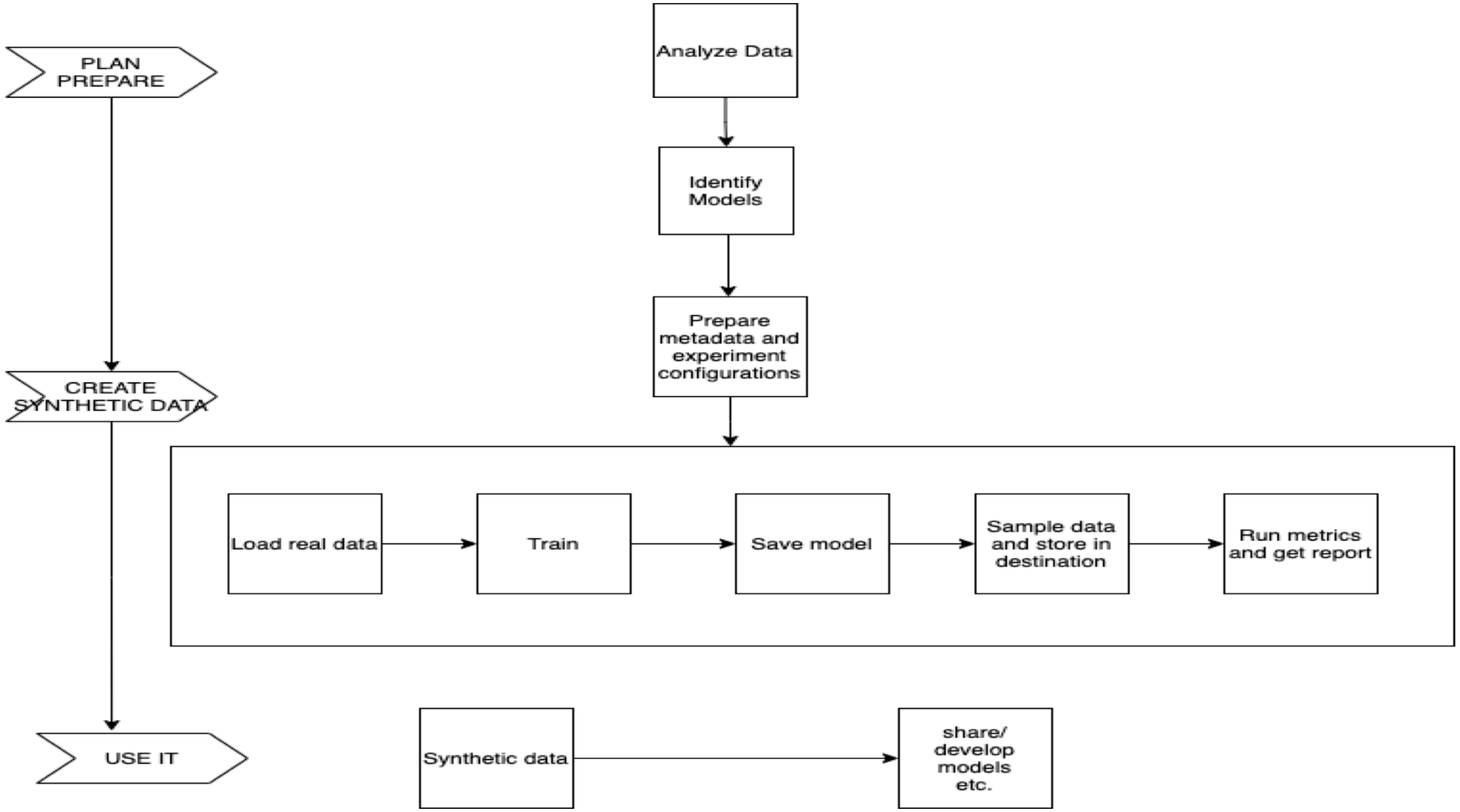
Raise the quality in the products and increase developer productivity by making this a test data generation strategy

Synthetic Data – A differentiator in Enterprise Data strategy

Typical Solution Framework – longitudinal patient data



Methodology



Framework for an implementation approach

DIY

Do it Yourself

- Use a open source framework
- Find your compute, Find your resource
- Great for learning, workshops

Vendor

SaaS products

- Many vendor products on SaaS
- Industry specific vendos
- Cloud complexities managed by them
- Specific projects, low hanging fruit to begin with

Data strategy

Make it part of your data strategy

- Make it part of your data strategy
- Integrate it with your data platform
- Treat this as a part of data curation component

Structure synthetic data vendors

https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784

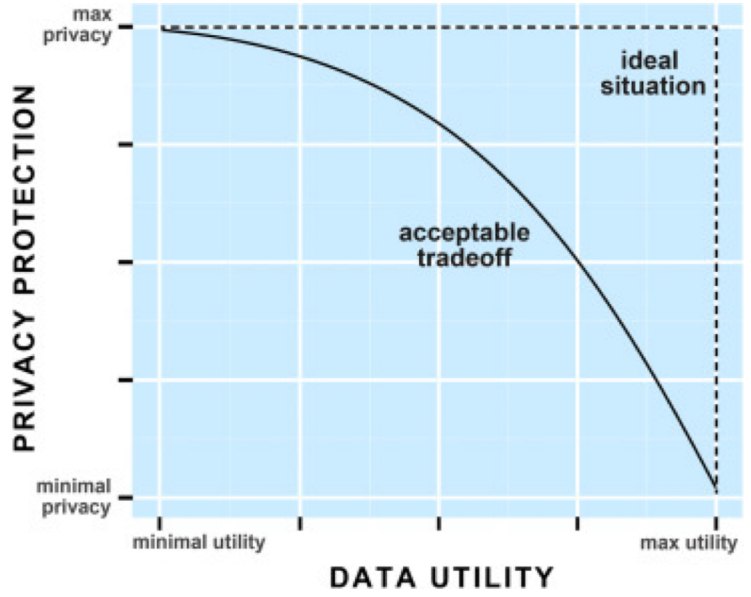
So how good is your generated synthetic data?

Goodness Measures - Why

Challenge:
Articulate the utility of Synthetic data

Technology exists to create synthetic data, but how do you know how good is the synthetic data and quantify for the purpose for which it is used. What is the evidence for demonstrating the utility of data?

Utility / privacy tradeoff



Goodness Measures - Approaches

Known by different names:

Evaluation framework, utility assessment etc.

Under the covers uses the very proven statistical and practical methods to get to a quantitative measure so confidence can be gained for using synthetic data

Brings data science dividing cultures together - statistics and linear algebra

Quantification for Utility - considerations

- Structural
- Subjective
- Correlations
- Relationships
- Privacy

Goodness Measures – Types of Metrics

Domain Experts – Subjective Assessment

Utility Metrics – Use statistical methods e.g. descriptive, correlation : irrespective of the domain
Column level, Table level, row levels, Multiple Table levels

https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary

Comparison with publicly available summary statistics if applicable

An Enterprise PETs to address privacy that enhance data sharing

PET (Privacy enhancing technologies) is an ecosystem of technologies that help to enhance the privacy of datasets. There are many PETs such as

Synthetic Data

Data generated from real data that has similar characteristics as original data

Tokenization: A piece of sensitive data such as credit card number is replaced by a surrogate value called token

Differential privacy

Privacy guarantee technique by adding noise to results of aggregate queries of the dataset

Homomorphic Encryption

Set of algorithm on encrypted data

Anonymization

Subtracting and transformation techniques that are irreversible. Masking and de-identification

Secure Enclave and more

Code execution in trusted hardware environments

Disclosure Analysis - (From CDC site)

Identity Disclosure

Identity Disclosure occurs if a third party can identify a subject or respondent from the released data

Attribute Disclosure

Occurs when confidential information about a person or a facility's operations is revealed, or can be closely estimated.

Inference Disclosure

Occurs when individual information can be inferred with high confidence from statistical properties of the released data.

Disclosure Analysis – For synthetic data

Identity / Attribute Disclosure

Unless overfitted, the risk is very low.

Inference Disclosure

Fit a model on synthetic data and try to infer from the fitted model and compare with original data

Demos

Synthetic data generation and evaluation:

https://github.com/jaynetra/WiDS2023JS/blob/main/SDV_Synthesize_a_table_(CTGAN).ipynb

GANS:
https://poloclub.github.io/ganlab/

Inspiration - Women in this field

Dr. Latanya Sweeney — Work on re-identification

Cynthia Dwork — Work on differential privacy

Top 30 List:

https://medium.com/@elise-deux/20-women-experts-on-the-topic-of-synthetic-data-60b7118f833

Closing thoughts

Synthetic Data - a key differentiator in enterprise data strategy.

From Gartner report -

By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated

Synthetic Data will play a important role in "**Data Centric AI"**

Q &A