JOIN LIVE

Synthetic data for AI Live demo

LEARN TO GENERATE SYNTHETIC DATA AND USE THEM FOR YOUR NEXT ML PROJECT.

LIVE ONLINE SESSION

ON 20-AUG-2022@ 7.00 - 08.30 PM IST (SATURDAY)
@ 9.30 - 11.00 AM EST
@ 2.30 - 04.00 PM LONDON

REGISTER FREE

REGISTRATION LINK:
HTTPS://FORMS.GLE/3NJRRMCWNHGUKYYE7

BY PRASANNA VENKATESH J

AI & CLOUD EVANGELIST

HTTPS://WWW.LINKEDIN.COM/IN/PRASANNA-VENKATESH-JAYAPRAKASH/

PRASANNAVJ@GMAIL.COM

+91-9840337443

AGENDA

- Introduction to Synthetic Data
- When and where synthetic data is used
  - Approaches to generate Synthetic data
  - Types of Synthetic data – framework,tools and models
  - Applications of Synthetic data
- LIVE DEMO

prasannavj@gmail.com

INTRODUCTION

- AI Data Market worth USD 4.8 Billion by 2027*
- **Problems faced by companies**
  - Collecting real data constraint
    - Time consuming process
      - Capture historic deails.
    - Privacy / sensitivity information (GDPR etc)
    - Processing time for cleaning and making it ready.
    - Limited information – cancer patients
    - Very high - class imbalance on real time data

* By Grand view research inc

prasannavj@gmail.com

SYNTHETIC DATA - INTRODUCTION

- Synthetic data mimic the real world observation used to train machine learning models when there is no real data is available.

- **Synthetic data is manuafctured information rather than recorded from real world events**

* Synthetic Data Generation in Machine Learning

prasannavj@gmail.com

DATA AUGMENTATION VS **DATA RANDAMIZING** VS **DATA SYNTHESIS**

- Understanding the difference
  - **Data Augmentation** vs **Data randamizing** vs **Data Synthesis**
- E.g
  - Face generation example
    - Modify the color of the eye – data augmentation
    - Swap the nose and position of nose from one image to another image – data ranfamizing
    - Generate new faces which are not real – data synthesis

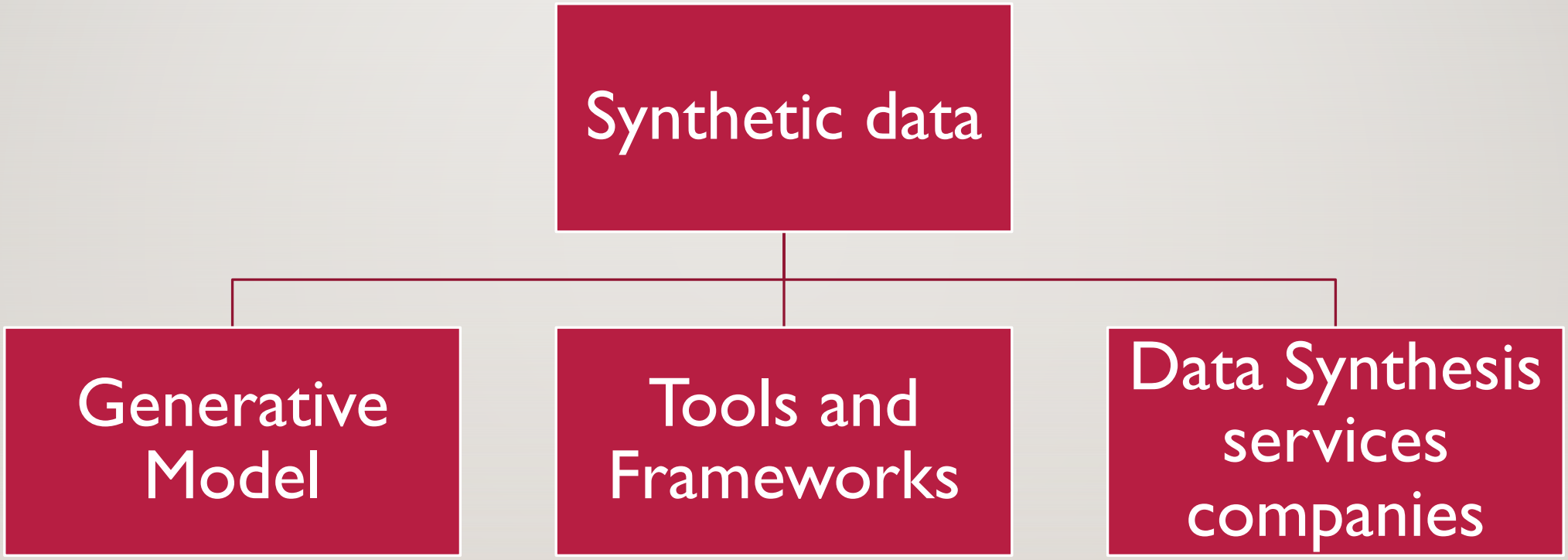
prasannavj@gmail.com

DATA TYPES

- Tabular Data
- Time Series Data
- Image Data
- Text Data
- Audio Data

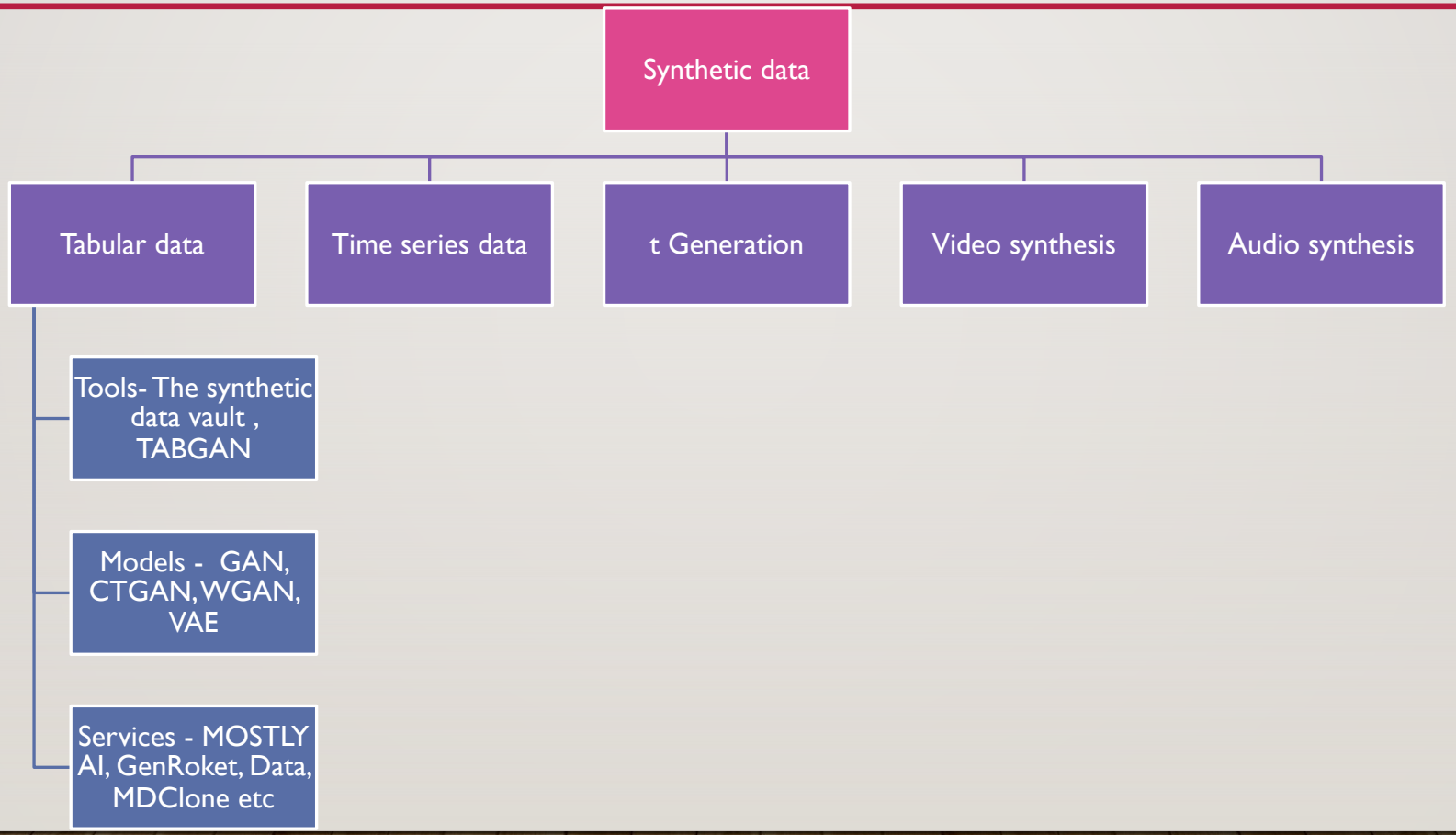
prasannavj@gmail.com

APPROACH TO GENERATE SYNTHERIC DATA



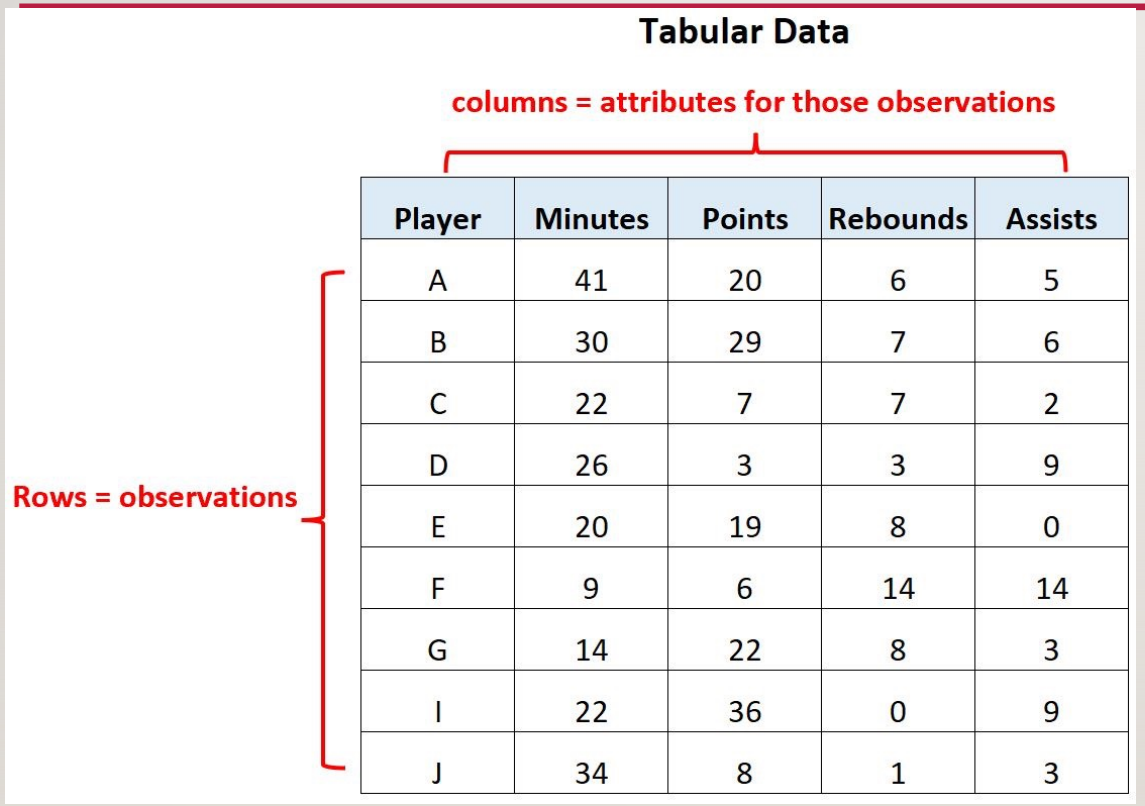
prasannavj@gmail.com

APPROACH TO GENERATE SYNTHERIC DATA



prasannavj@gmail.com

TABLULER SYNTHETIC DATA GENERATION



Player	Minutes	Points	Rebounds	Assists
A	41	20	6	5
B	30	29	7	6
C	22	7	7	2
D	26	3	3	9
E	20	19	8	0
F	9	6	14	14
G	14	22	8	3
I	22	36	0	9
J	34	8	1	3

Sensitive invormation should be handled;

- PII data (Personal identifiable information)

- Address
- Name
- Sex

- <u>Focus on:</u>
  - Single table
  - Multiple table
    - With maintaining relationships

prasannavj@gmail.com

UNDERSTANDING GAN – GENERATIVE ADVERSARIAL NETWORKS

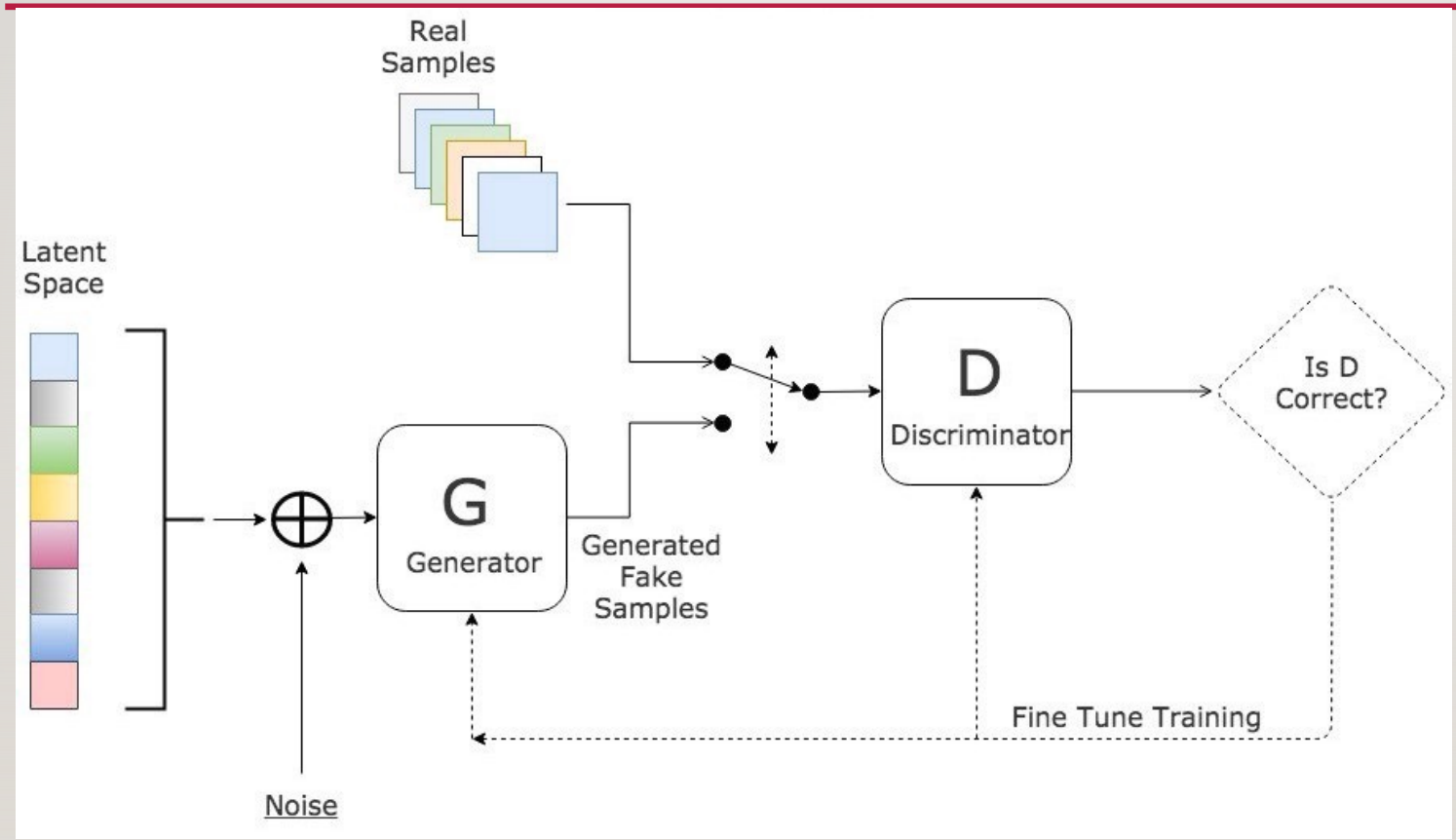


Image - GAN training pipeline. By Jonathan Hui — What is Generative Adversarial Networks GAN? [1]

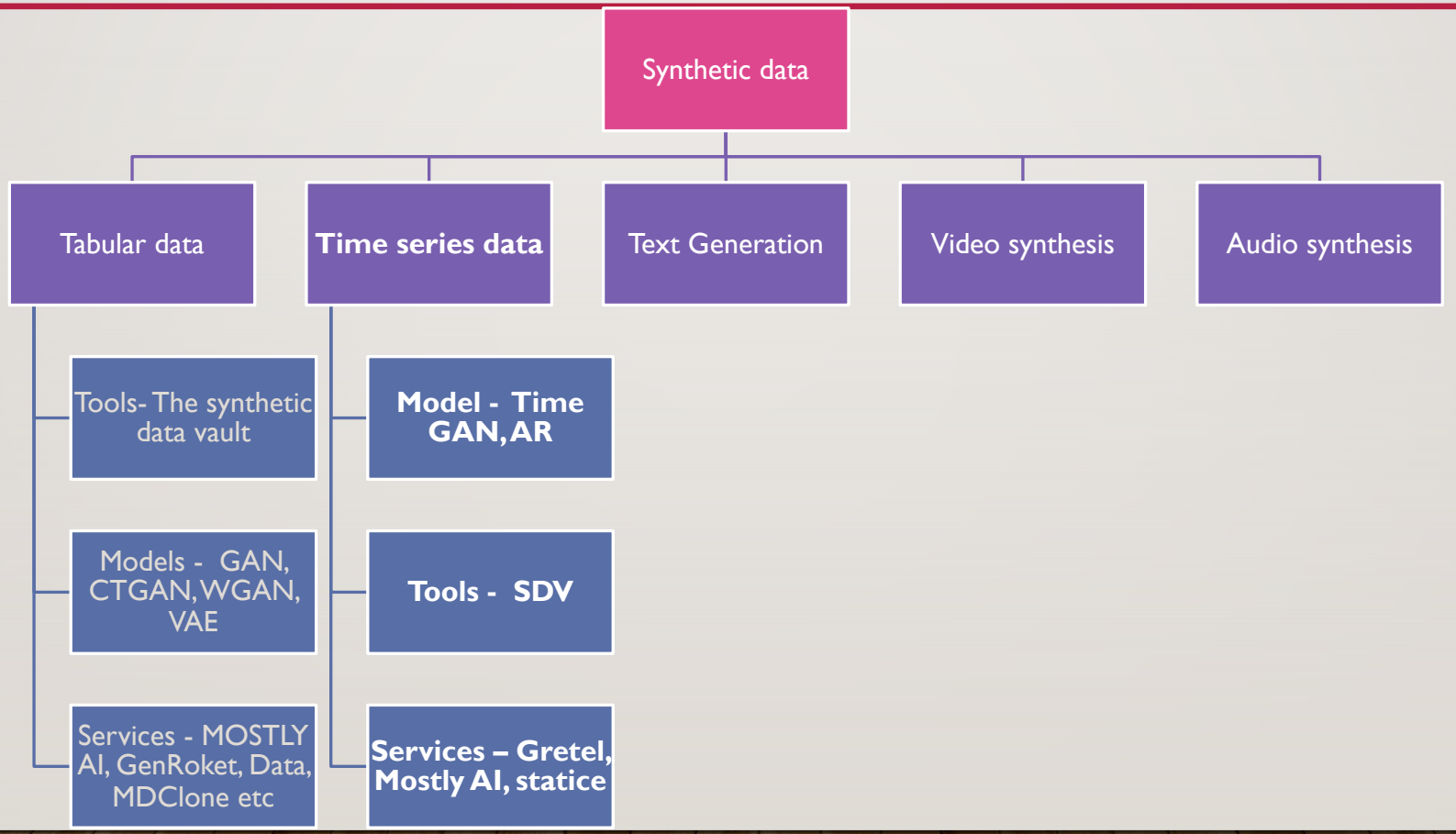
prasannavj@gmail.com

DEMO – SDV & TABGAN

- Link to download the source code: https://github.com/prasannavj/synthericdataai

- Original Source code path - https://github.com/sdv-dev/SDV

APPROACH TO GENERATE SYNTHERIC DATA



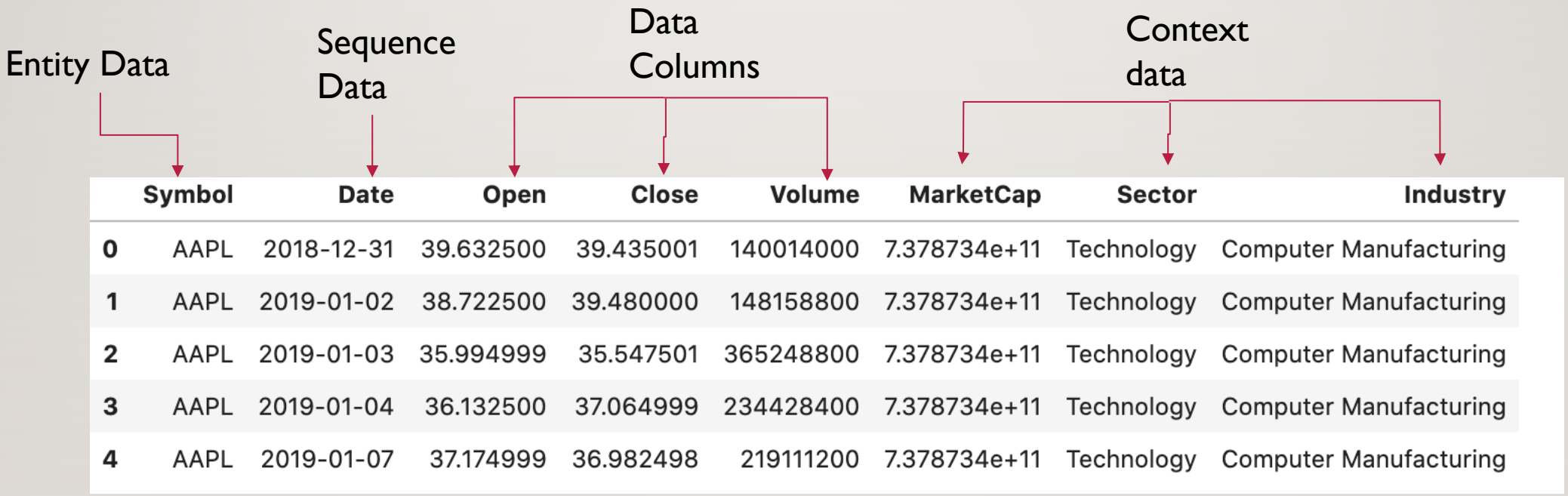
prasannavj@gmail.com

UNDERSTANDING TIME SERIES DATA – STOCK MARKET DATA

- Example Time series data: Stock data for all stocks in NYSE – for 2018-2019.

	Symbol	Date	Open	Close	Volume	MarketCap	Sector	Industry
0	AAPL	2018-12-31	39.632500	39.435001	140014000	7.378734e+11	Technology	Computer Manufacturing
1	AAPL	2019-01-02	38.722500	39.480000	148158800	7.378734e+11	Technology	Computer Manufacturing
2	AAPL	2019-01-03	35.994999	35.547501	365248800	7.378734e+11	Technology	Computer Manufacturing
3	AAPL	2019-01-04	36.132500	37.064999	234428400	7.378734e+11	Technology	Computer Manufacturing
4	AAPL	2019-01-07	37.174999	36.982498	219111200	7.378734e+11	Technology	Computer Manufacturing

TIME SERIES DATA INTRO



	Symbol	Date	Open	Close	Volume	MarketCap	Sector	Industry
0	AAPL	2018-12-31	39.632500	39.435001	140014000	7.378734e+11	Technology	Computer Manufacturing
1	AAPL	2019-01-02	38.722500	39.480000	148158800	7.378734e+11	Technology	Computer Manufacturing
2	AAPL	2019-01-03	35.994999	35.547501	365248800	7.378734e+11	Technology	Computer Manufacturing
3	AAPL	2019-01-04	36.132500	37.064999	234428400	7.378734e+11	Technology	Computer Manufacturing
4	AAPL	2019-01-07	37.174999	36.982498	219111200	7.378734e+11	Technology	Computer Manufacturing

prasannavj@gmail.com

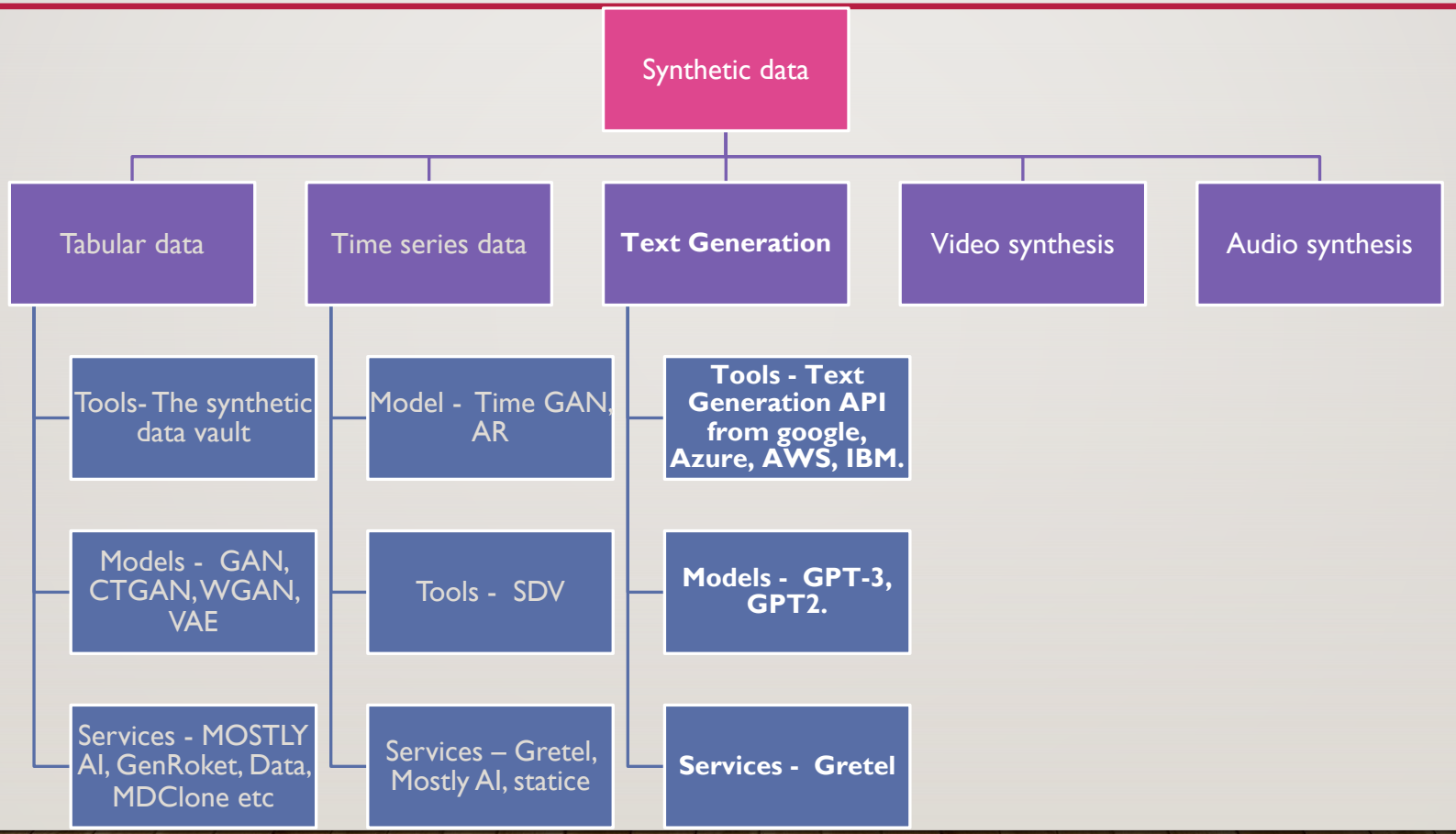
SYNTHETIC DATA GENERATION OF TIMESERIES DATA DEMO

- Link to download the source code:

- Original Source code path - https://github.com/sdv-dev/SDV

prasannavj@gmail.com

APPROACH TO GENERATE SYNTHERIC DATA

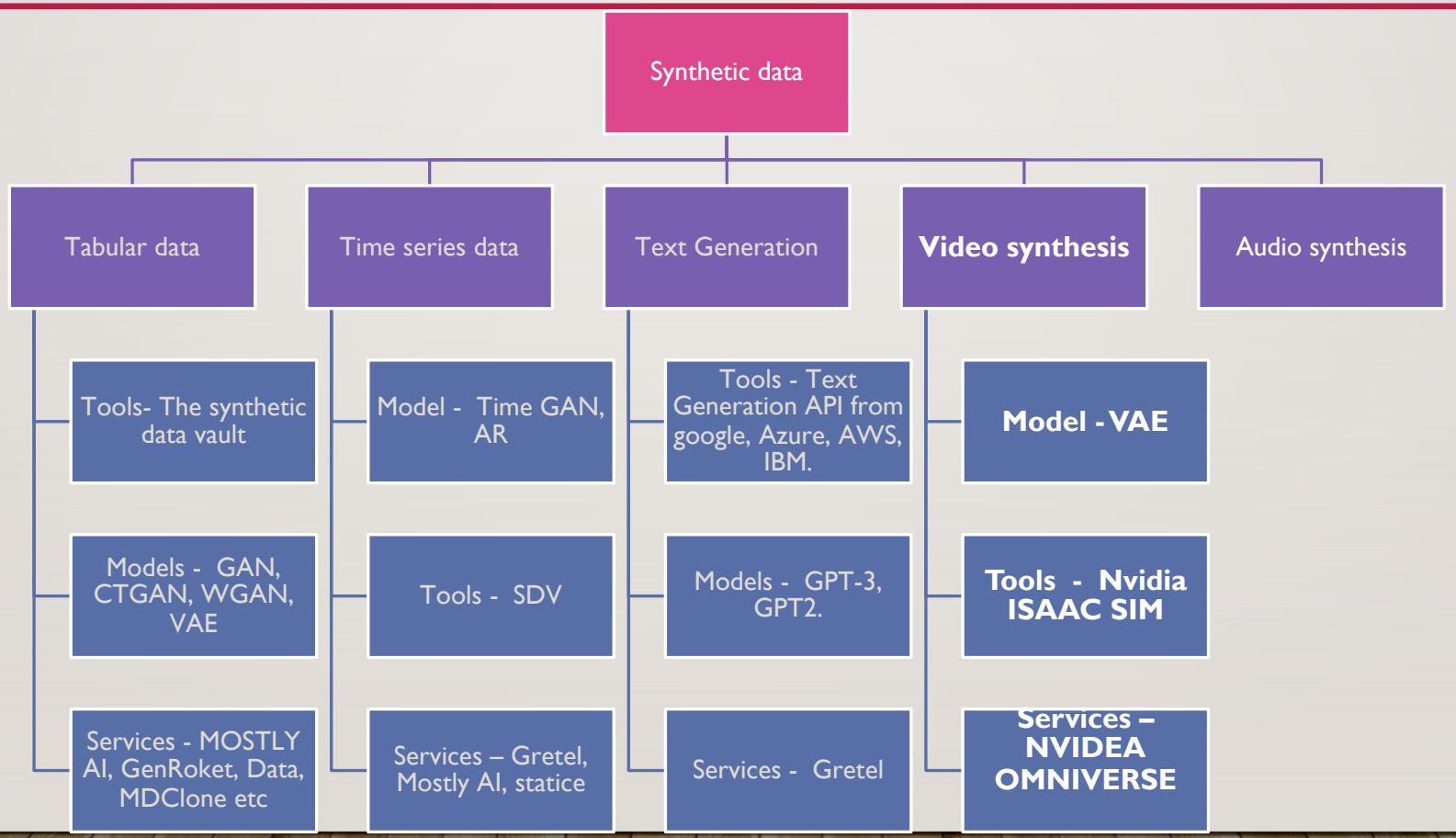


prasannavj@gmail.com

TEXT GENERATION - LIVE DEMO

- DeepAI – Text Generation Model available for consumption.
- https://deepai.org/machine-learning-model/text-generator
- Also provided by following providers: Hugging faces, google, AWS, Microsoft Azure, IBM etc.

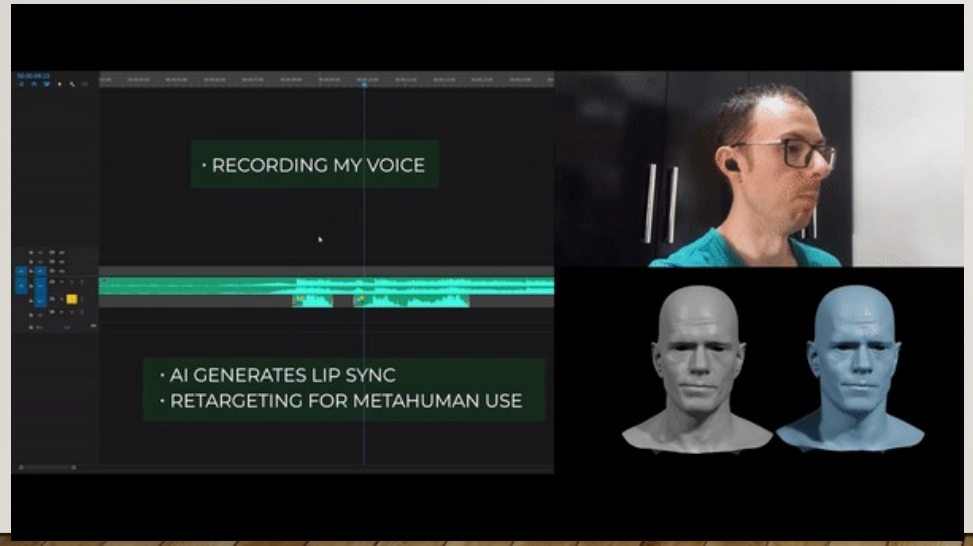
VIDEO / IMAGE SYNTHESIS

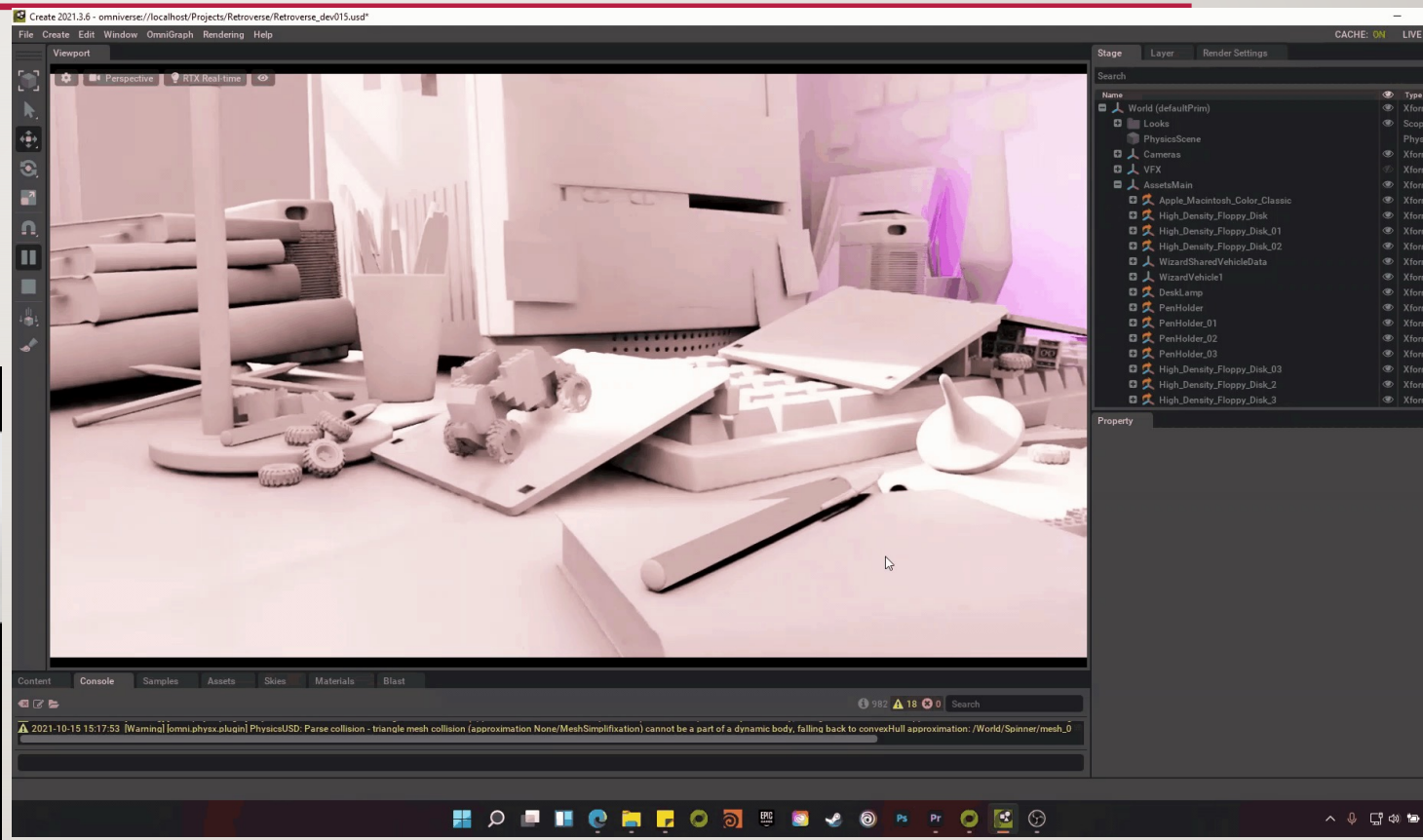


prasannavj@gmail.com

VIDEO / IMAGE SYNTHESIS

- NVIDIA – Omniverse and ISAAC SIM
- Option for simulate realworld environment and interactions





prasannavj@gmail.com

MICROSOFT SYNTHETIC FACE GENERATION AND TRAINING - FAKE IT TILL YOU MAKE IT.

- https://microsoft.github.io/FaceSynthetics/
- https://www.youtube.com/watch?v=jSsMqjMcRAg&feature=emb_rel_end
- More than 10,000 synthetic face images generated.

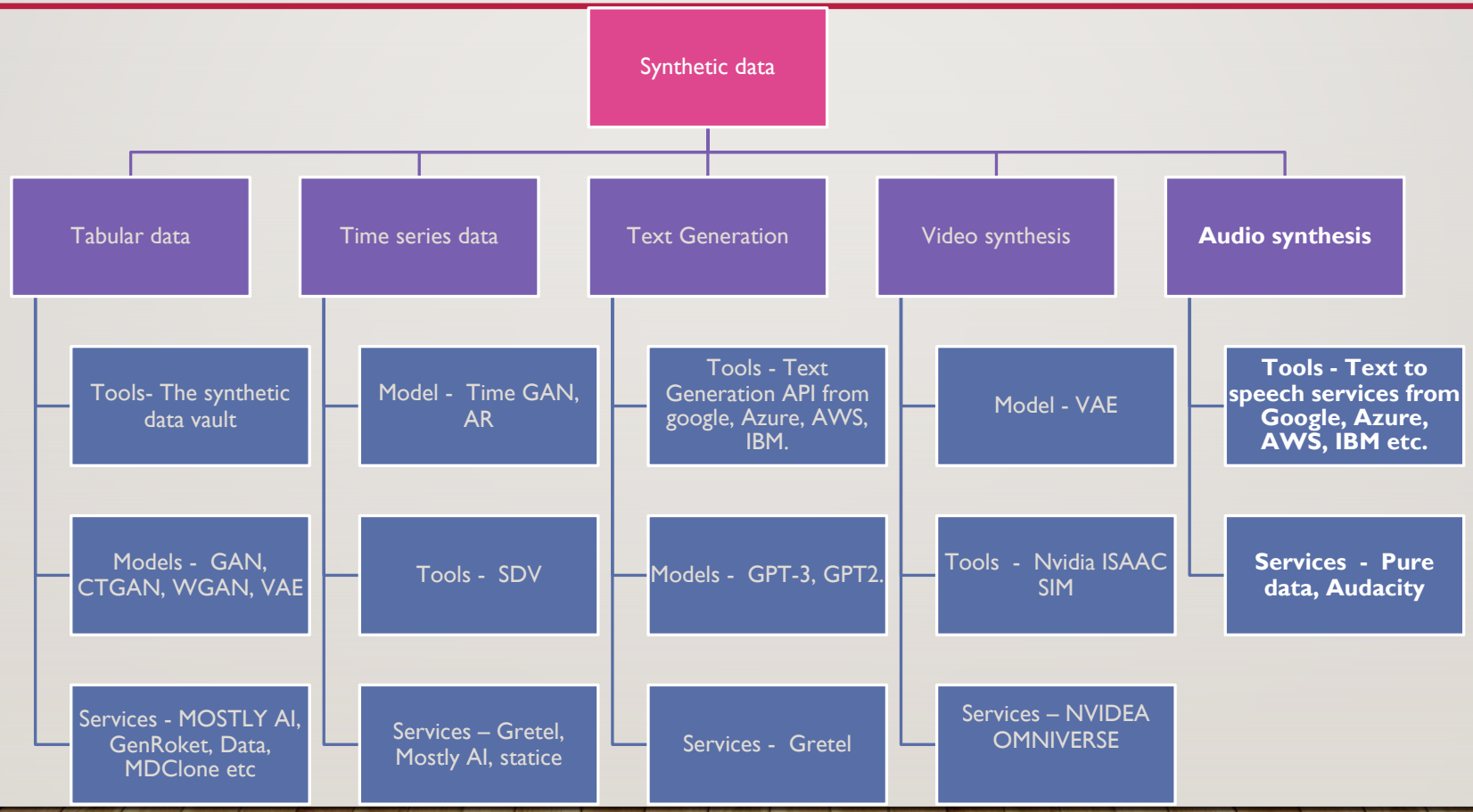
prasannavj@gmail.com

REFERENCE LINKS

- NVIDIA ISAAC SIM – https://www.youtube.com/watch?v=-VQLqs6s9y0
- This person does not exist: https://this-person-does-not-exist.com/en
- EveryBody Dance Now - https://github.com/carolineec/EverybodyDanceNow

prasannavj@gmail.com

AUDIO SYNTHESIS



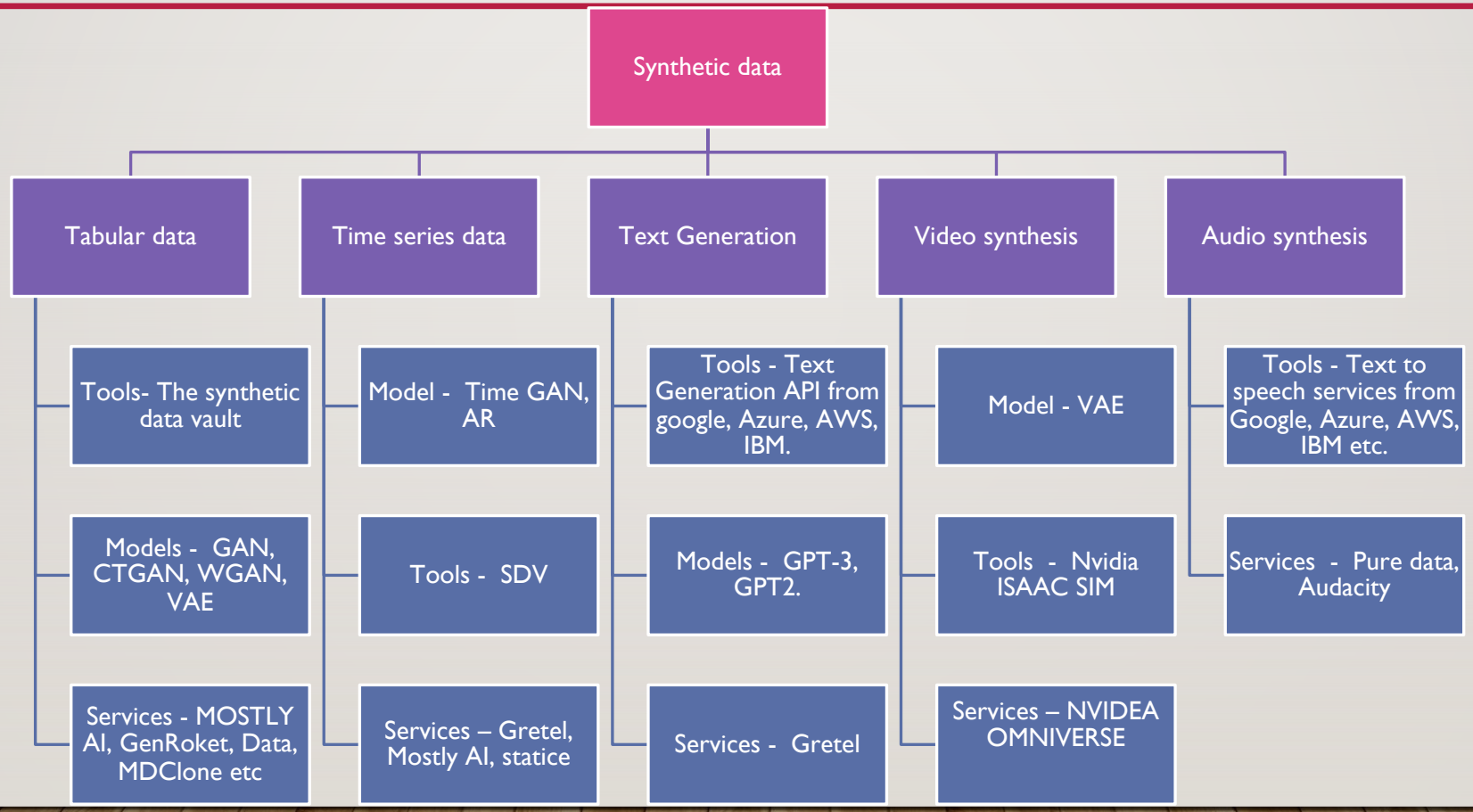
prasannavj@gmail.com

TEXT TO SPEECH GENERATION API AND TOOLS

- Google Text to speech generation.
- AWS, Azure, IBM etc all provide Text to speech generation services.

- Demo Link - https://aws.amazon.com/polly/features/#Brand_Voice

APPROACH TO GENERATE SYNTHERIC DATA



prasannavj@gmail.com

SYNTHETIC DATA GENERATION - USECASES

prasannavj@gmail.com

SYNTHETIC DATA GENERATION - USECASES

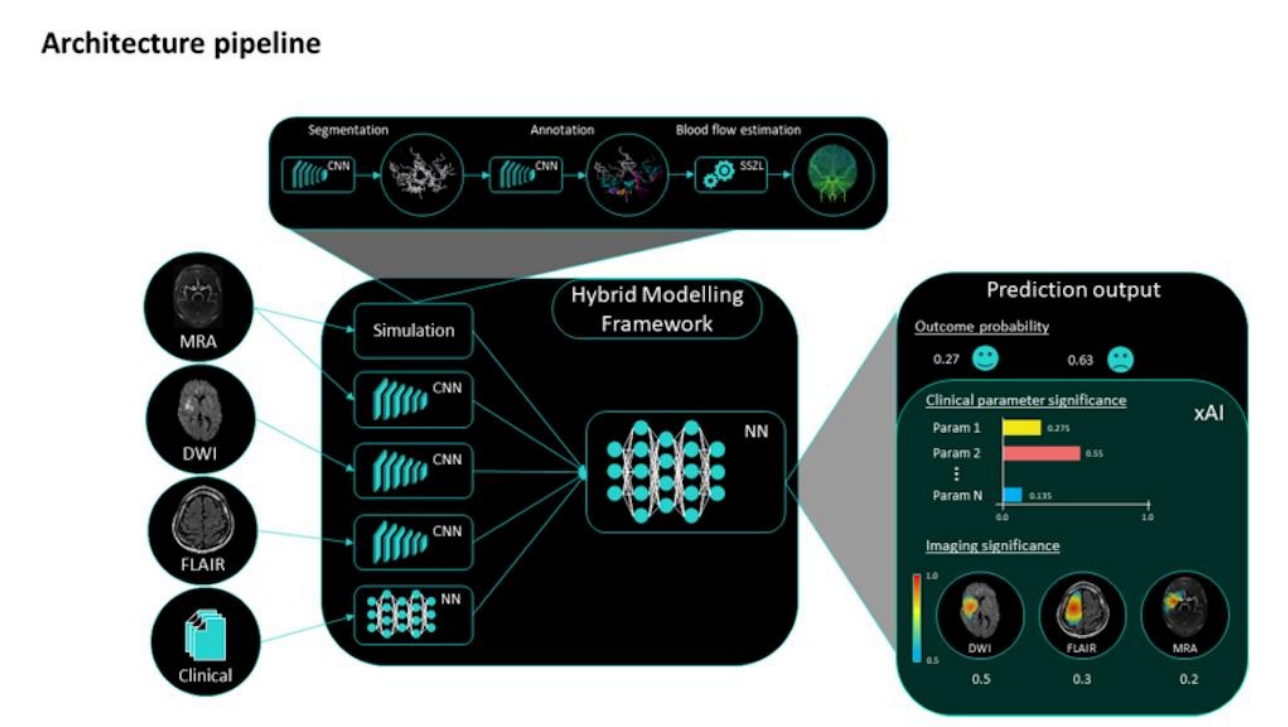
American express fraud detection.

- As they do not have enough fraud data – synthesize various scenario and generate data for fraud detection.

https://www.bangkokpost.com/business/2158831/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data

prasannavj@gmail.com

MEDICAL SIMULATIONS



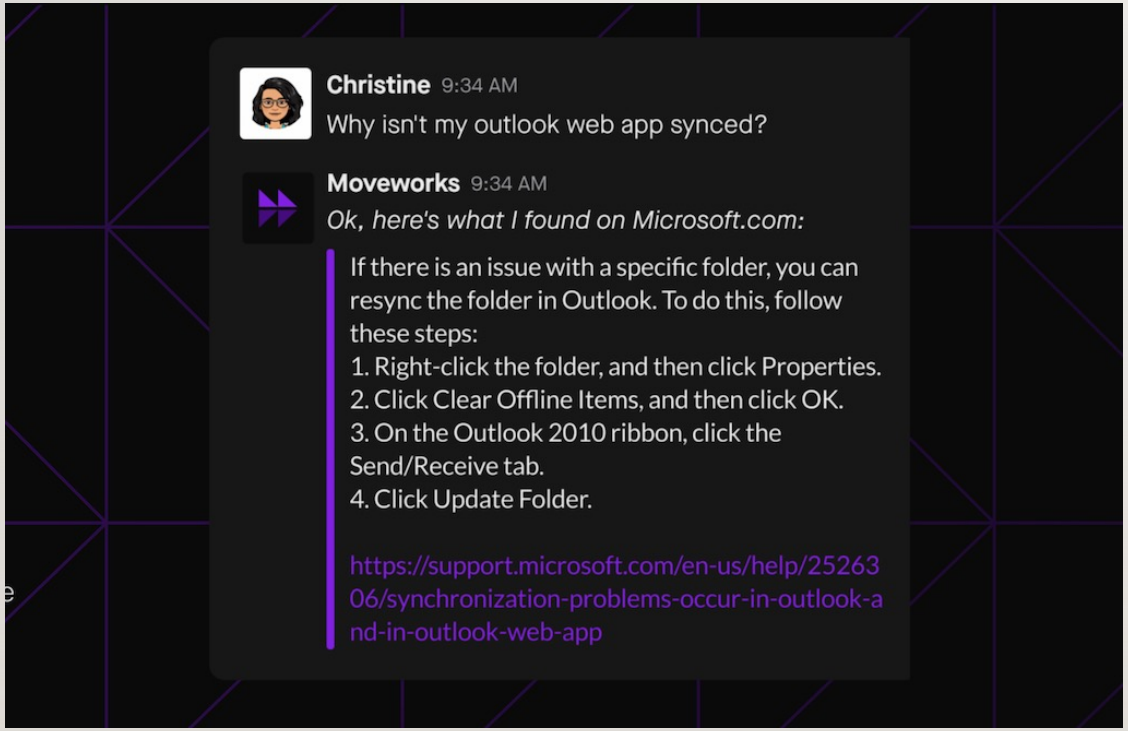
Charite - **PREDICTioN2020**

- Medical simulation platform for stroke prediction.
- Uses the sysnthetic medical and image data for platform building.

- https://claim.charite.de/en/projects/prediction2020/

prasannavj@gmail.com

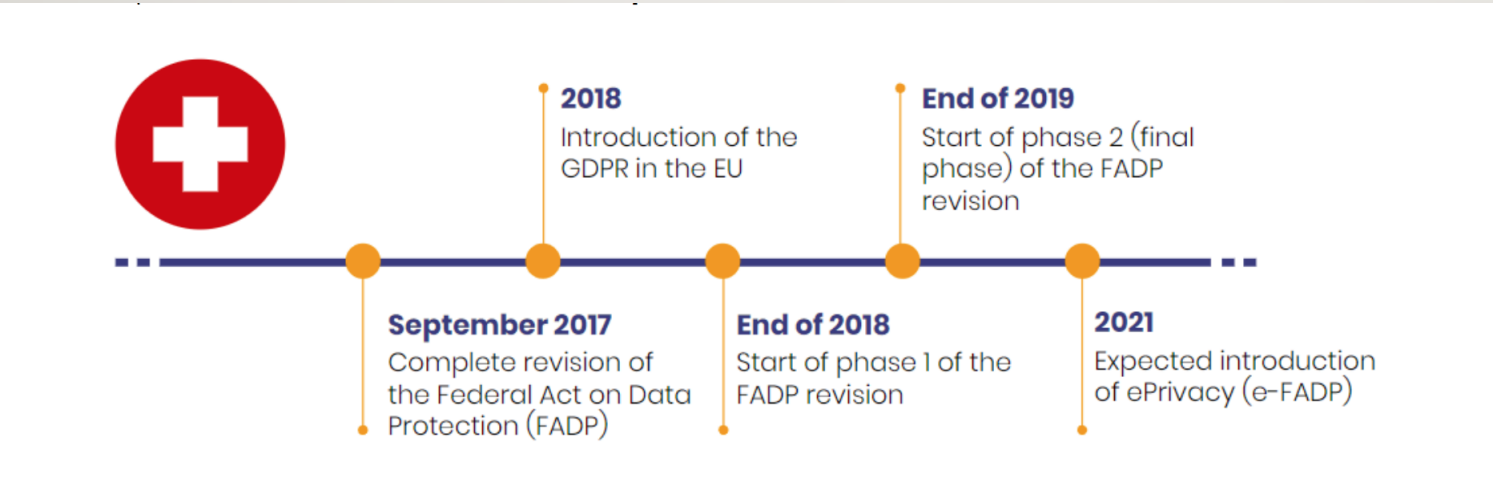
CHATBOT DEVELOPMENT



Moveworks chatbot was trained based on synthetic data generated.

prasannavj@gmail.com

CHURN MODEL SIMULATION

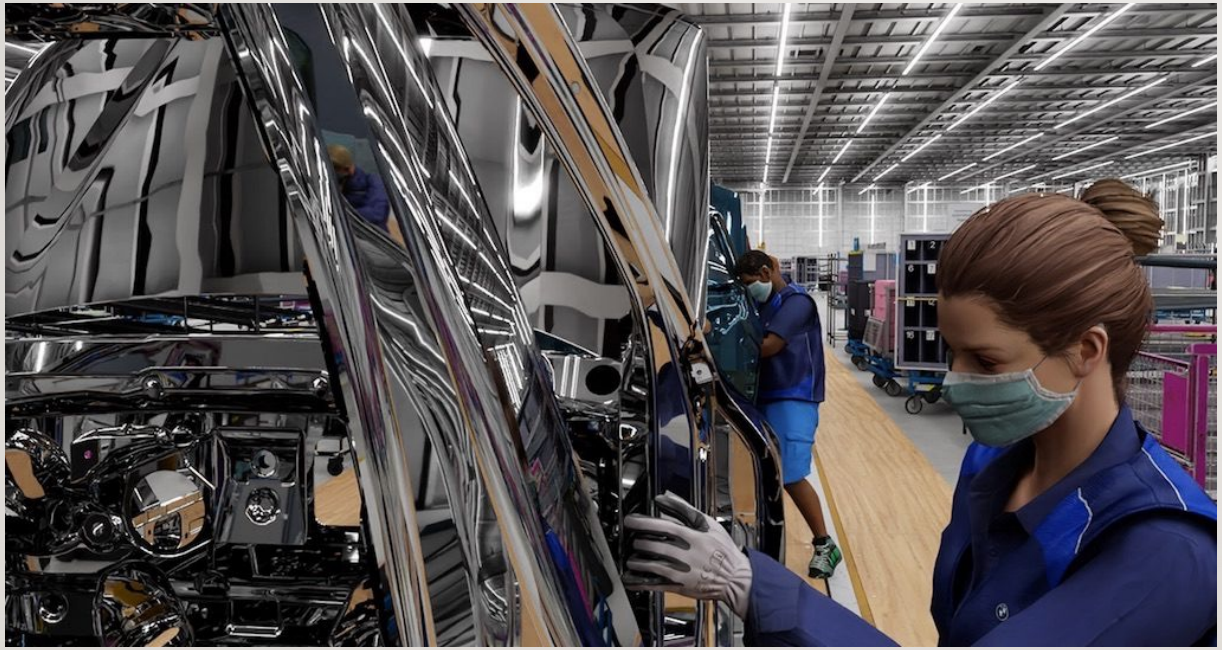


Swiss **insurance company La Mobilière uses synthetic data to predict churn prediction.**

https://f.hubspotusercontent10.net/hubfs/3832818/Resources/case_studies/CS_Mobiliere_x_Statice.pdf?hsCtaTracking=de21e86a-a06f-4e3a-b218-01515973014 9%257C471d7f33-9783-4c15-9607-b2c6fc0ecb3a

prasannavj@gmail.com

VIRTUAL FACTORY FACILITY



BMW along with NVIDIA build a virtual factory using synthetic data for effective robotic operations.

Creating digital twin.

https://blogs.nvidia.com/blog/2021/04/13/nvidia-bmw-factory-future/

prasannavj@gmail.com

DIALOG PROCESSING

Amazon alexa was trained on synthetic data for better the performance of the conversational bot.

https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases

prasannavj@gmail.com

REFERENCES

- https://www.altexsoft.com/blog/synthetic-data-generation/
- https://deepai.org/machine-learning-model/text-generator
- https://sdv.dev/
- https://dai.lids.mit.edu/contact/
- https://github.com/carolineec/EverybodyDanceNow
- https://docs.aws.amazon.com/polly/latest/dg/voicelist.html
- https://microsoft.github.io/FaceSynthetics/