Raymond A. Mason
School of Business
WILLIAM & MARY

Synthetic data

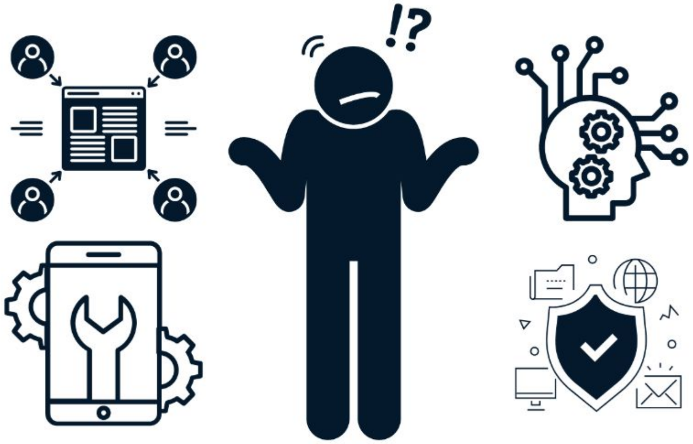
Team 8: Luke Smith, Garrison Chura, Sofiya Kuzina, and Cynthia Marquez

Problem Statement

What is synthetic data, what is its purpose, and how effective is its use in AI modeling?

We aim to explore the effectiveness, challenges, and applications of using synthetic data to overcome the limitations of using real-world data to train AI/ML models.

Context



Real data:

Difficult to access/laborious

Expensive

Constrained by regulations – privacy concerns

Synthetic data:

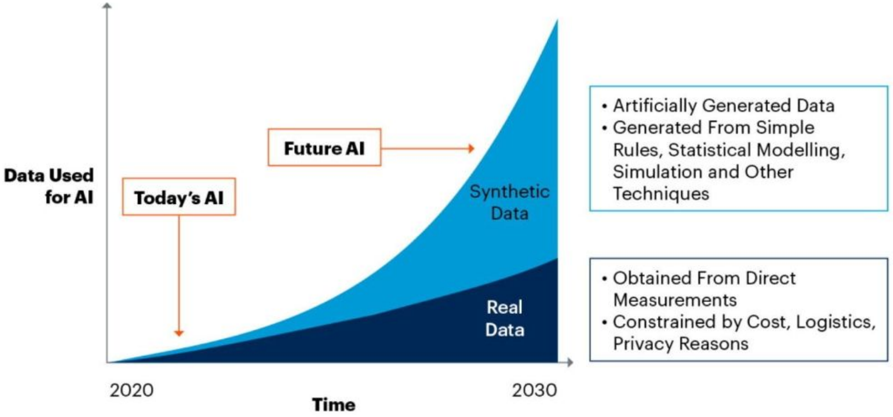
Computer-generated data that is similar to real-world data

Primary purpose: increase the privacy and integrity of systems

Source: Creating Synthetic Data with Python Faker Tutorial | DataCamp (https://www.datacamp.com/tutorial/creating)

Importance of Synthetic Data

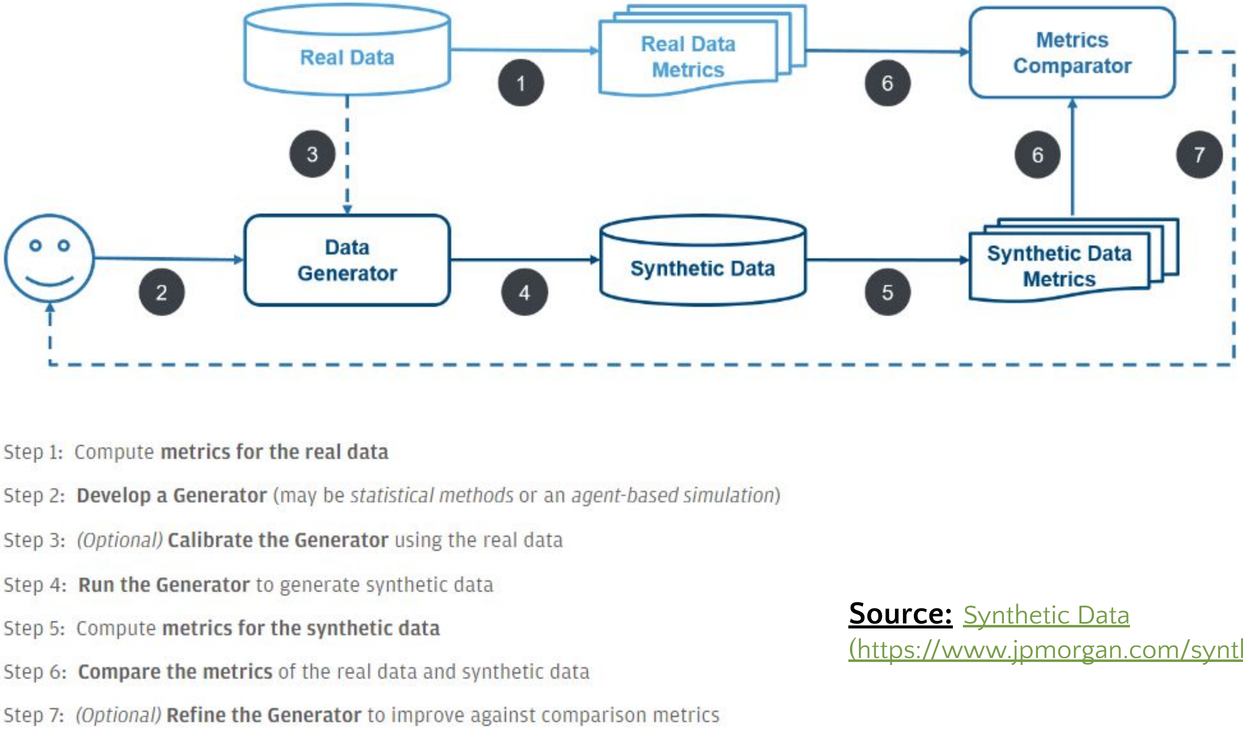
- Privacy
- Cost
- Diversity
- Control
- Scalability



Source: Gartner
750175_C

Source: Creating Synthetic Data with Python Faker Tutorial | DataCamp (https://www.datacamp.com/tutorial/creating)

Process to Generate Synthetic Financial Datasets



Step 1: Compute **metrics for the real data**

Step 2: **Develop a Generator** (may be *statistical methods* or an *agent-based simulation*)

Step 3: *(Optional)* **Calibrate the Generator** using the real data

Step 4: **Run the Generator** to generate synthetic data

Step 5: Compute **metrics for the synthetic data**

Step 6: **Compare the metrics** of the real data and synthetic data

Step 7: *(Optional)* **Refine the Generator** to improve against comparison metrics

Source: Synthetic Data (https://www.jpmorgan.com/synthetic-data)

Raymond A. Mason
School of Business
WILLIAM & MARY

Literature Review

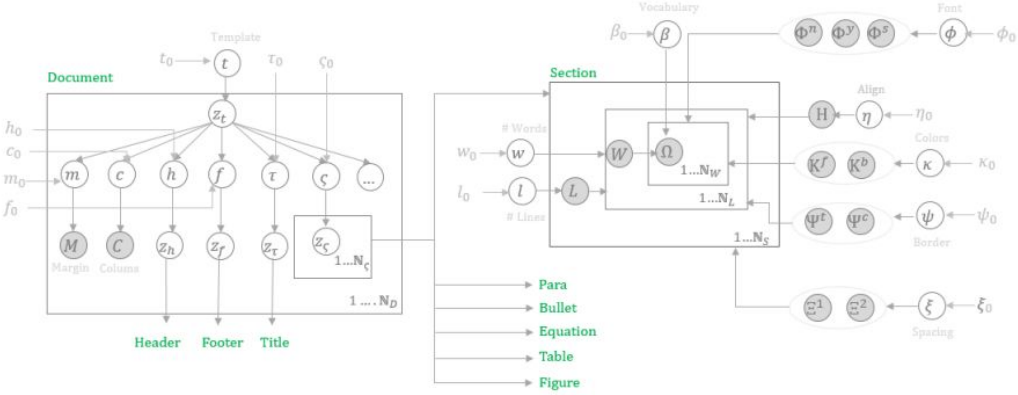
"Synthetic Document Generator For Annotation-Free Layout Recognition"

Natraj Raman, Sameena Shah and Manuela Veloso

JPMorgan AI Research Lab

“Synthetic Document Generator for Annotation-free Layout Recognition”

- Analyze document layout
- Use Bayesian Network to make synthetic documents
- Train object detection model to predict labels for each part of a document’s layout



Source: https://doi.org/10.1016/j.patcog.2022.108660

“Synthetic Document Generator for Annotation-free Layout Recognition”

- Titles, sections, headers/footers,, tables, figures etc. help understand document content
- **Layout recognition**
  - Use object detection model
  - Input images → scaled feature maps → identify layout elements and boundaries
- **Synthetic document generator** → produces realistic documents that have labeled layout elements/spatial positions

“Synthetic Document Generator for Annotation-free Layout Recognition”

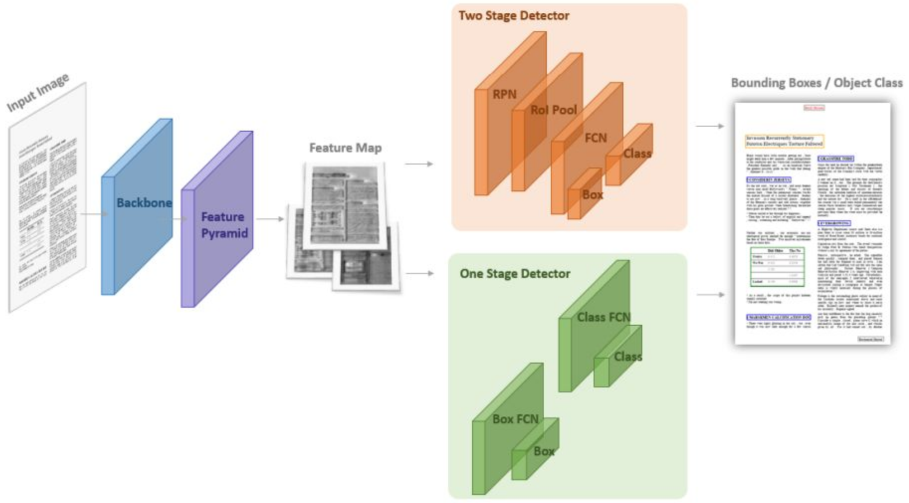
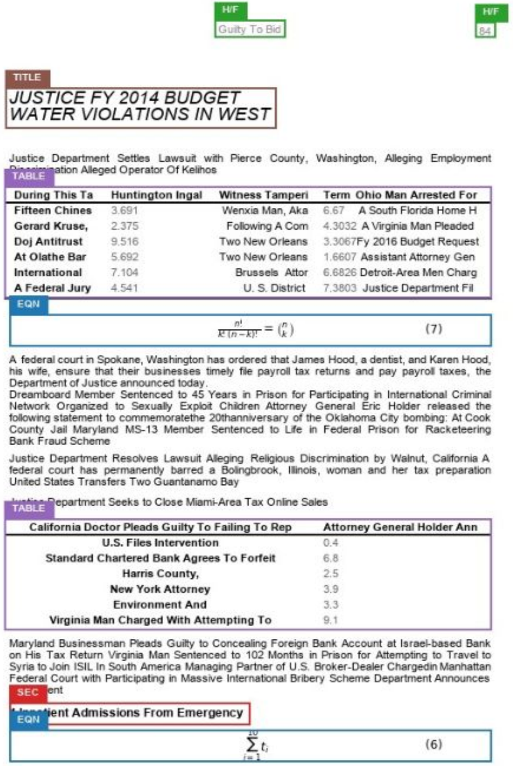


Figure 3: Layout Recognition Model Architecture. A feature extraction network takes an image of arbitrary size as input and produces feature maps at different scales. An object detector network determines the categories and bounding boxes of the layout elements.


H/F
Guilty To Bid
H/F
84
TITLE
TABLE
EQN
TABLE
SEC
EQN

JUSTICE FY 2014 BUDGET WATER VIOLATIONS IN WEST

Justice Department Settles Lawsuit with Pierce County, Washington, Alleging Employment Discrimination Alleged Operator Of Kelihos

During This Ta	Huntington Ingal	Witness Tamperi	Term	Ohio Man Arrested For
Fifteen Chines	3.691	Wenxia Man, Aka	6.67	A South Florida Home H
Gerard Kruse,	2.375	Following A Com	4.3032	A Virginia Man Pleaded
Doj Antitrust	9.516	Two New Orleans	3.3067	Fy 2016 Budget Request
At Olathe Bar	5.692	Two New Orleans	1.6607	Assistant Attorney Gen
International	7.104	Brussels Attor	6.6826	Detroit-Area Men Charg
A Federal Jury	4.541	U. S. District	7.3803	Justice Department Fil

$$\frac{n!}{k!\,(n-k)!} = \binom{n}{k} \qquad (7)$$

A federal court in Spokane, Washington has ordered that James Hood, a dentist, and Karen Hood, his wife, ensure that their businesses timely file payroll tax returns and pay payroll taxes, the Department of Justice announced today.
Dreamboard Member Sentenced to 45 Years in Prison for Participating in International Criminal Network Organized to Sexually Exploit Children Attorney General Eric Holder released the following statement to commemoratethe 20thanniversary of the Oklahoma City bombing: At Cook County Jail Maryland MS-13 Member Sentenced to Life in Federal Prison for Racketeering Bank Fraud Scheme

Justice Department Resolves Lawsuit Alleging Religious Discrimination by Walnut, California A federal court has permanently barred a Bolingbrook, Illinois, woman and her tax preparation United States Transfers Two Guantanamo Bay

Justice Department Seeks to Close Miami-Area Tax Online Sales

California Doctor Pleads Guilty To Failing To Rep	Attorney General Holder Ann
U.S. Files Intervention	0.4
Standard Chartered Bank Agrees To Forfeit	6.8
Harris County,	2.5
New York Attorney	3.9
Environment And	3.3
Virginia Man Charged With Attempting To	9.1

Maryland Businessman Pleads Guilty to Concealing Foreign Bank Account at Israel-based Bank on His Tax Return Virginia Man Sentenced to 102 Months in Prison for Attempting to Travel to Syria to Join ISIL In South America Managing Partner of U.S. Broker-Dealer Chargedin Manhattan Federal Court with Participating in Massive International Bribery Scheme Department Announces ent

Inpatient Admissions From Emergency

$$\sum_{i=1}^{10} t_i \qquad (6)$$

Source:
https://doi.org/10.1016/j.patcog.2022.108660

“Synthetic Document Generator for Annotation-free Layout Recognition”

Results:

- Train layout detectors on synthetic data → as good as real documents
- Increase number of synthetic documents → performance of real and synthetic documents converge
- “Granularity” of the layout categories could impact recognition quality

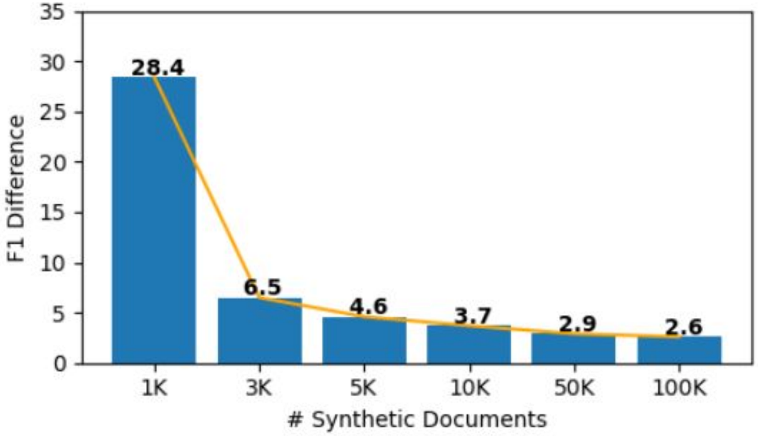
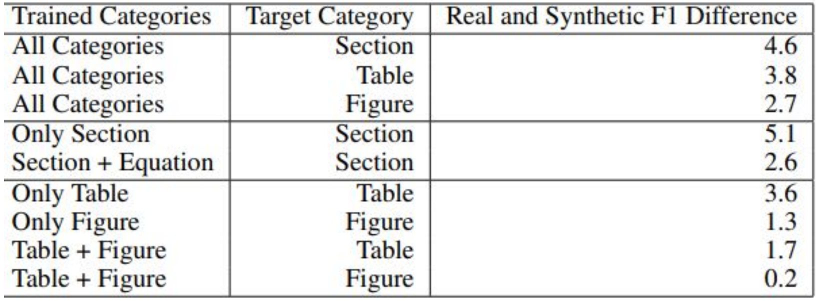


Table 8: Impact of training with a subset of layout categories.



Trained Categories	Target Category	Real and Synthetic F1 Difference
All Categories	Section	4.6
All Categories	Table	3.8
All Categories	Figure	2.7
Only Section	Section	5.1
Section + Equation	Section	2.6
Only Table	Table	3.6
Only Figure	Figure	1.3
Table + Figure	Table	1.7
Table + Figure	Figure	0.2

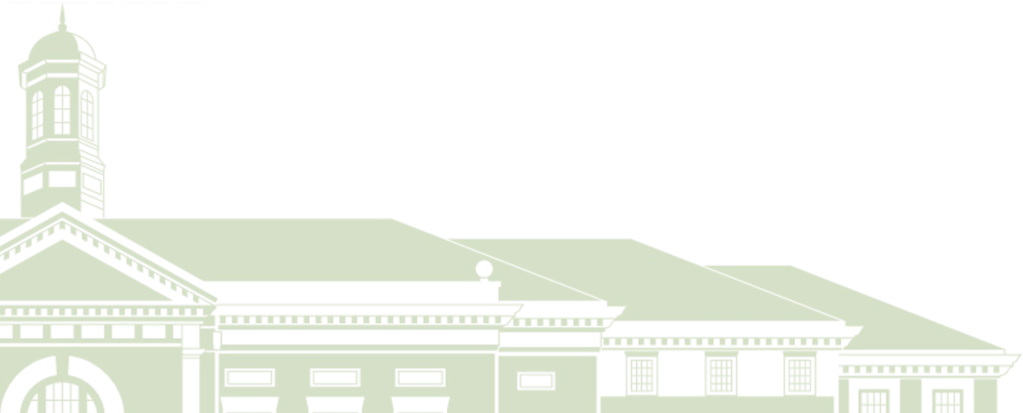
Source: https://doi.org/10.1016/j.patcog.2022.108660

Code Demo

Make_classification (sklearn.datasets)

Function used to generate synthetic datasets for classification tasks (random n-class)

- Testing Machine Learning Algorithms
- Creating Imbalanced Datasets and Data
- Synthetic Data Augmentation



Explainability, Challenges, & Ethical Concerns

Synthetic data can…

- Amplify biases that exist in real-world data
- Miss complexities (such as outliers) of real data
- Not be sensitive to real-time changes
- Ethics:
  - Ownership of synthetic data from publically available data
  - Privacy → can help ensure privacy, but…
    - Data leakage – risk of individuals being identified from real data
    - Web scraping, using real data without consent

However…

- Synthetic datasets can as accurate than real-world data

  (Raman, S., Shah, S., & Veloso, M. (2022))

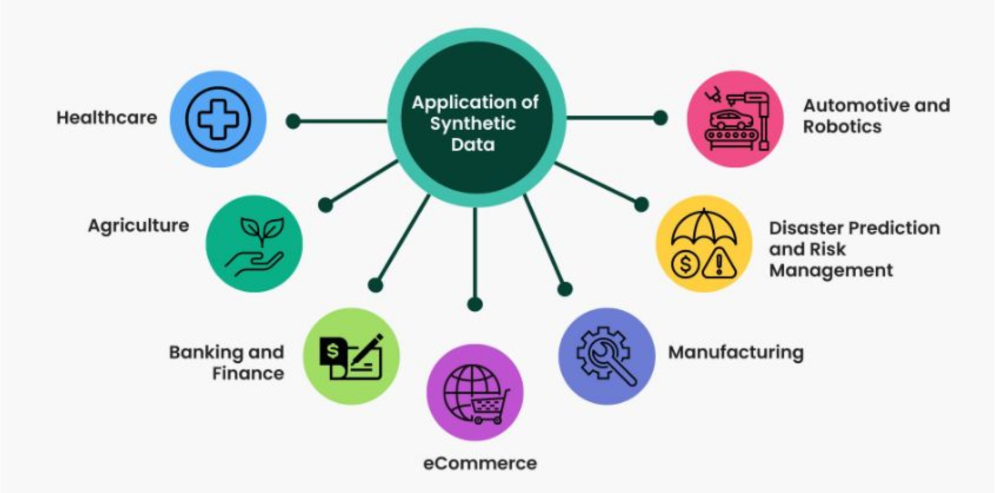
What's Next?

Future Developments

- Industry-specific applications
- Techniques to capture data from different sources
- New idea → spread awareness

Concerns:

- Address potential biases
- Ensure synthetic data represents diversity of real-world data
- Develop standard to measure synthetic data accuracy (validation)



Source: Synthetic Data Generation: Definition, Types, Techniques, & Tools (https://www.turing.com/kb/synthetic-data-generati)

Raymond A. Mason School of Business
WILLIAM & MARY

References

Awan, A. (2022, August). Creating Synthetic Data with Python Faker Tutorial. DataCamp. https://www.datacamp.com/tutorial/creating-synthetic-data-with-python-faker-tutorial

J.P. Morgan. (n.d.). Synthetic Data. https://www.jpmorgan.com/synthetic-data

Radecic, D. (2021, January 10). How to Make Synthetic Datasets with Python: A Complete Guide for Machine Learning. Better Data Science. https://betterdatascience.com/python-synthetic-datasets/

Raman, S., Shah, S., & Veloso, M. (2022). Synthetic document generator for annotation-free layout recognition. Pattern Recognition, 128. https://doi.org/10.1016/j.patcog.2022.108660

Turing.com. (n.d.). Synthetic Data Generation: Definition, Types, Techniques, and Tools. https://www.turing.com/kb/synthetic-data-generation-techniques

Zewe, A. MIT News. (2022, November 3). In machine learning, synthetic data can offer real performance improvements. https://news.mit.edu/2022/synthetic-data-ai-improvements-1103