# Generate Synthetic Data with Archetypal Analysis

GRBIO RETREAT
17/07/2025

IRIS Institute for Research and Innovation in Health

GRBIO

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Liukuan Yu

SUPERVISORS： Jordi Cortés Martínez and Daniel Fernández Martínez

# CONTENTS

01 Introduction

02 Proposed Approach

03 Results

04 Summary and Future Research Directions

PART 01

# Introduction

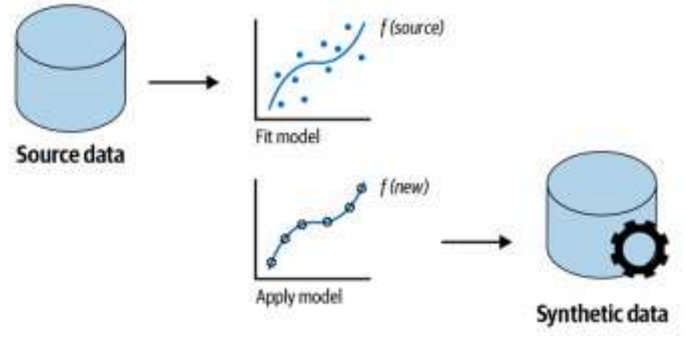
# What Is Synthetic Data?

Synthetic Data

Synthetic data is not real data , but it has the same statistical properties

1

## Synthetic data (from real data)

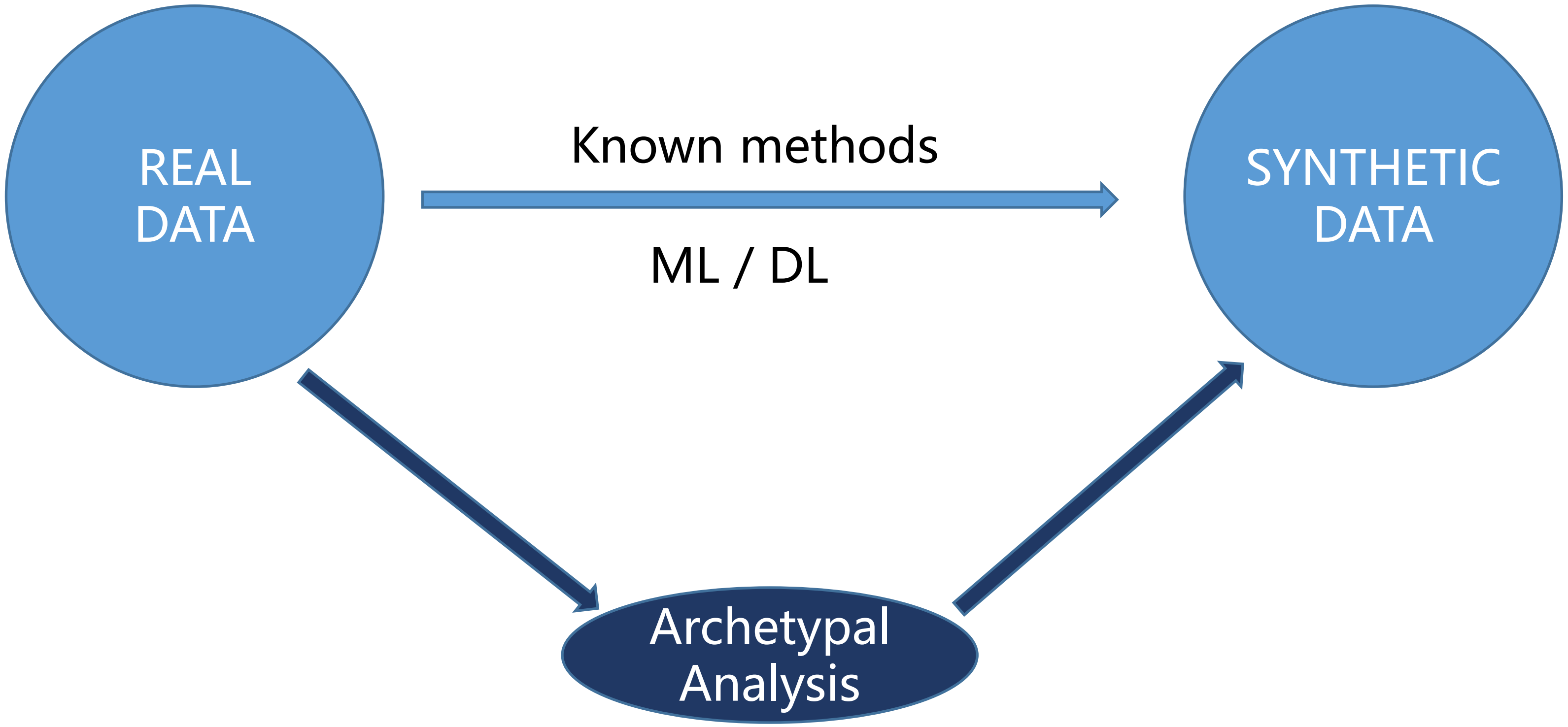


2

## Simulation (without real data)

It is created by using existing models or the background knowledge of the analysts.

# Objective-for Generate Synthetic Data with Real Data



REAL DATA

Known methods

ML / DL

SYNTHETIC DATA

Archetypal Analysis

# What Is the Archetypes?

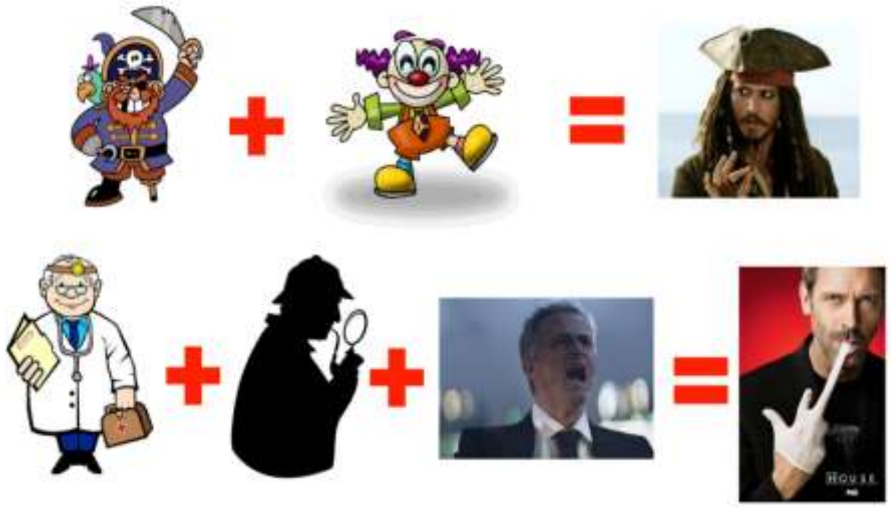
Archetype (Wikipedia): from Greek:

- *archē*: “beginning” , “origin” .
- *tupos*: “pattern” , “model” , or “type” .

**Original pattern from which copies are made.**

Archetypes in everyday language:

- Jack Sparrow: 40% pirate and 60% clown.
- Dr. House: 20% doctor, 30% detective, and 50% bad temper.



In Statistics, the **concept of archetypes** is the same as in common life.

# Archetypal Analysis (AA)

**AA (Cutler and Breiman, 1994) aims to find extreme cases:**

1. **Archetypes** are convex combinations of the **Observations**

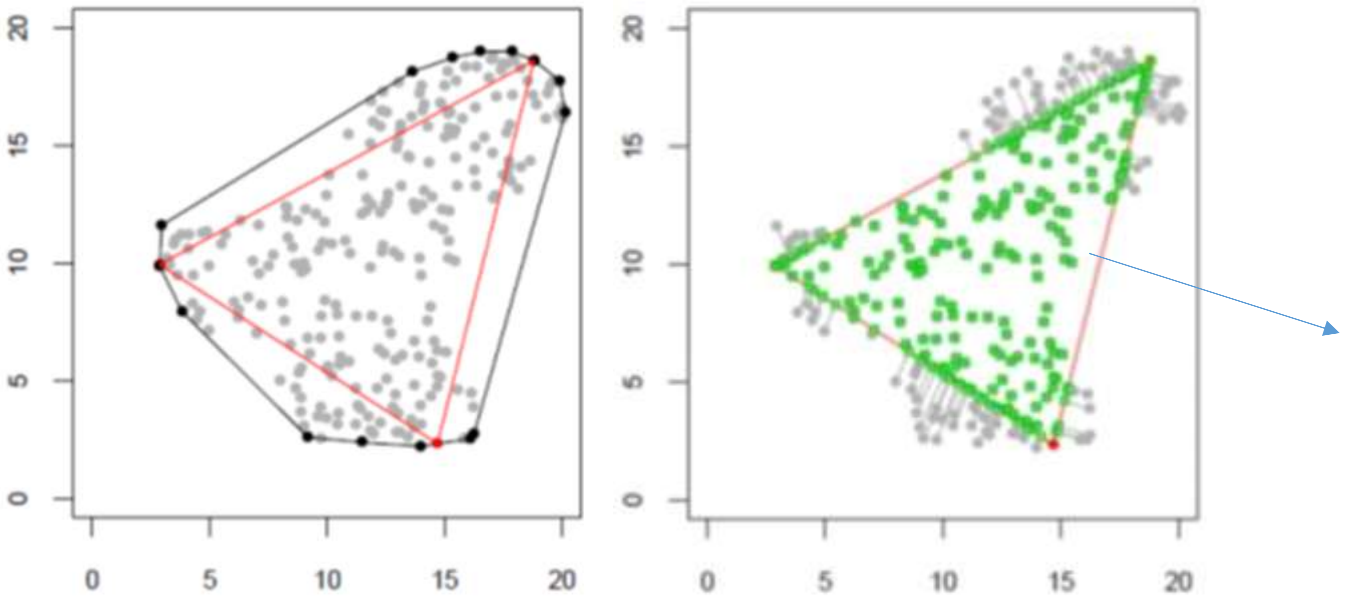
2. **Observations** are convex combinations of the **Archetypes**

Minmize $RSS\ \sum_{i=1}^{n} ||\boldsymbol{x_i} - \sum_{j=1}^{k} \alpha_{ij}\boldsymbol{z_j}||^2 = \sum_{i=1}^{n} ||\boldsymbol{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\boldsymbol{x}_l||^2$

Under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $\mathrm{j} = 1, \ldots, k$



convex combination of archetypes

PART 02

# Proposed Approach

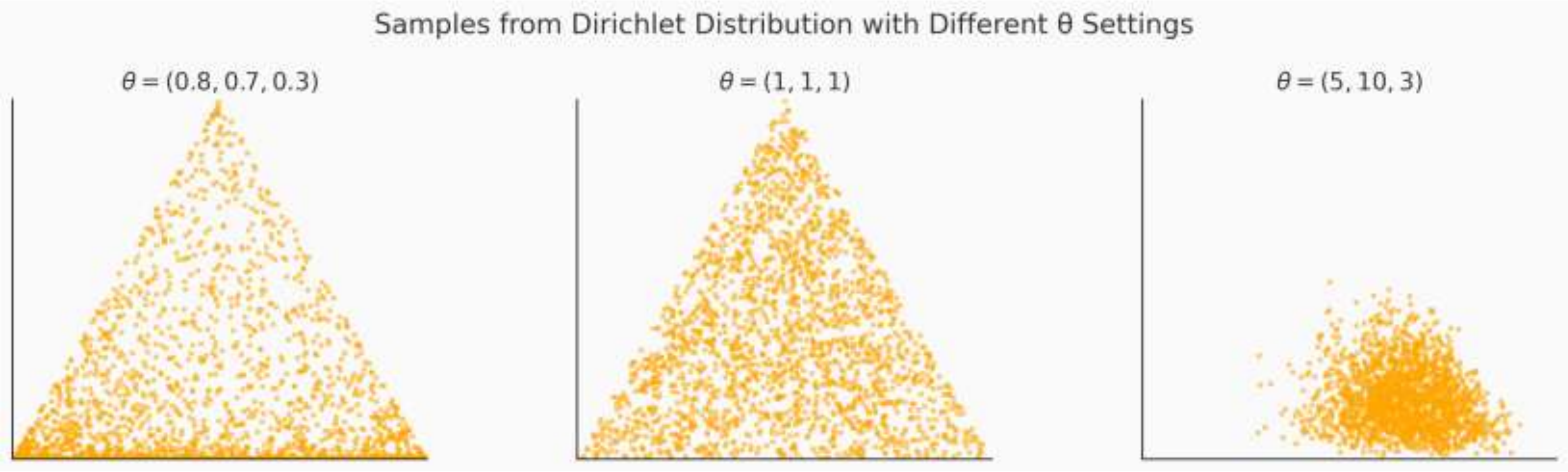
# Dirichlet Distribution

Key Properties:

a) It generates random probability vectors $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\mathrm{k})$ with $\alpha_i \geq 0$ and $\sum \alpha_i = 1$

b) The $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_\mathrm{k})$ is the vector of parameters of the Dirichlet distribution being each $\theta_\mathrm{i} > 0$

c) Behavior depends on the $\theta$ values:
If all $\theta_\mathrm{i} = 1$ → uniform distribution over the simplex
Larger $\theta_\mathrm{i}$ → samples are concentrated around the center
Smaller $\theta_\mathrm{i}$ → samples are clustered near the corners of the simplex

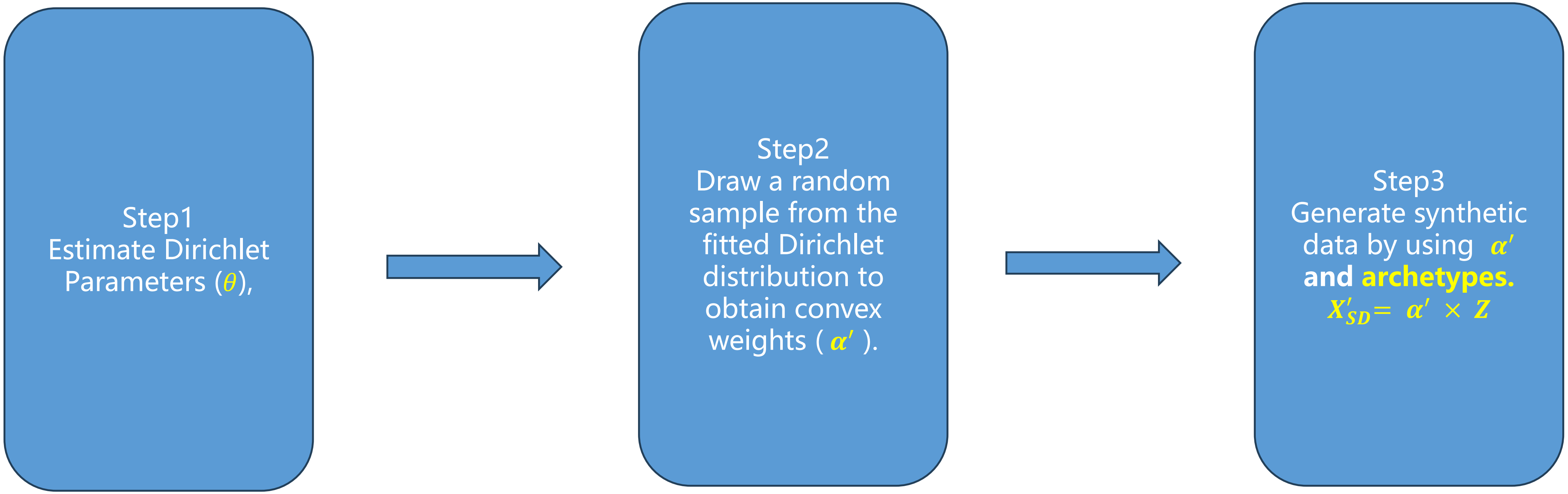


# Proposed Approach

Minimize $RSS \;\; \sum_{i=1}^{n} ||\boldsymbol{x_i} - \sum_{j=1}^{k} \alpha_{ij}\boldsymbol{z_j}||^2$

One of the constraints

$\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1 , \ldots , n$



Step1
Estimate Dirichlet Parameters ($\theta$),

Step2
Draw a random sample from the fitted Dirichlet distribution to obtain convex weights ( $\alpha'$ ).

Step3
Generate synthetic data by using $\alpha'$ **and archetypes.**
$X'_{SD} = \alpha' \times Z$

**Output:**
**Dirichlet parameters derived from $\theta$**

**Output:**
**New convex weight**

**Output:**
**Synthetic data**

PART 03

# Results

# IRIS Dataset

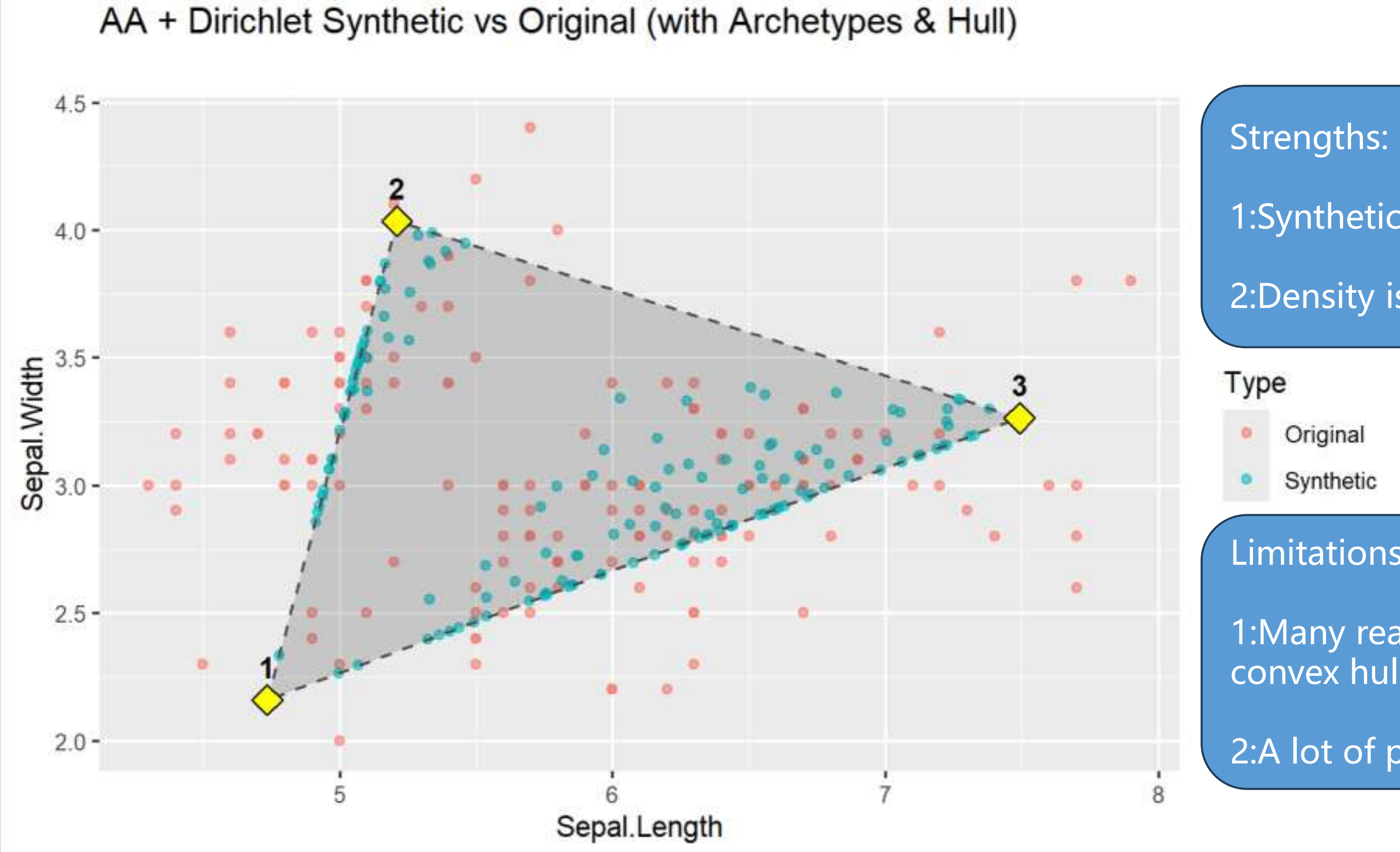
A classic Dataset in statistics and machine learning

- 150 iris flowers, 3 species: Setosa, Versicolor, Virginica

- 4 features: Sepal length, Sepal width, Petal length, Petal width

- Widely used for clustering, classification

# Result：



Strengths:

1:Synthetic data shows general structure.

2:Density is relatively same.

Limitations:

1:Many real data points lie outside the convex hull (74.7% inside).

2:A lot of points are on the boundary.

PART 04

# Summary and Future Research Directions

# Summary

## Methodology

- It has been defined using AA & Dirichlet

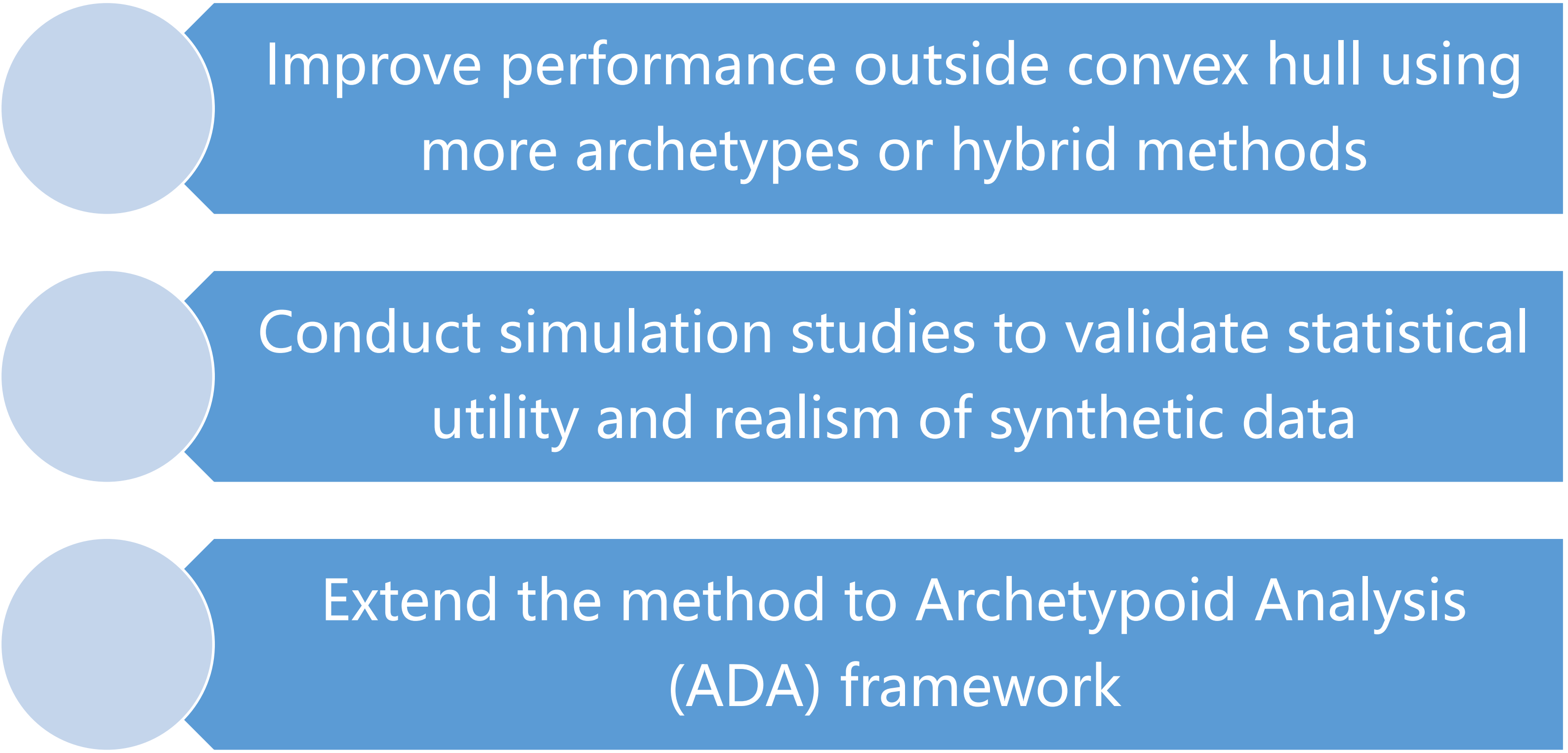
## Implementation

- Preliminary application using Iris dataset

## Results

- Initial results show not that good similarity between real and synthetic data

# Future Research Directions



- Improve performance outside convex hull using more archetypes or hybrid methods
- Conduct simulation studies to validate statistical utility and realism of synthetic data
- Extend the method to Archetypoid Analysis (ADA) framework

# Acknowledgements

This research was funded by the MICIU / AEI / 10.13039/501100011033(Spain) and by FEDER(EU) [PID2023-138033OB-C21] & and by grant 2021 SGR 01421 (GRIBO) administrated by Departament de Recerca i Universitats de la Generalitat de Catalunya (Spain).

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

UNIÓN EUROPEA
FONDO EUROPEO DE
DESARROLLO REGIONAL
"Una manera de hacer Europa"

AGENCIA
ESTATAL DE
INVESTIGACIÓN

# Reference

1：**Cutler, A., & Breiman, L. (1994).** Archetypal analysis. *Technometrics*, *36*(4), 338–347.

2：**Vinué, G., Epifanio, I., & Alemany, S. (2015).** Archetypoids: A new approach to define representative archetypal data. *Computational Statistics & Data Analysis*, *87*, 102–115.

3：**El Emam, K., Mosquera, L., & Hoptroff, R. (2020).** *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O'　Reilly Media.

4：**Fernández, D., Epifanio, I., & McMillan, L. F. (2021).** Archetypal Analysis for Ordinal Data. *Information Sciences*, *579*, 281–292.

# Gracias

谢谢