ANALYZE STOCK MAKET USING CLUSTERING

Group 6

- Huỳnh Thanh Phúc
- Nguyễn Mỹ Khanh
- Lê Khánh Duy

TABLE OF CONTENTS

LITERATURE REVIEW

METHODS

Filtered graph (PMFG, DBHT)

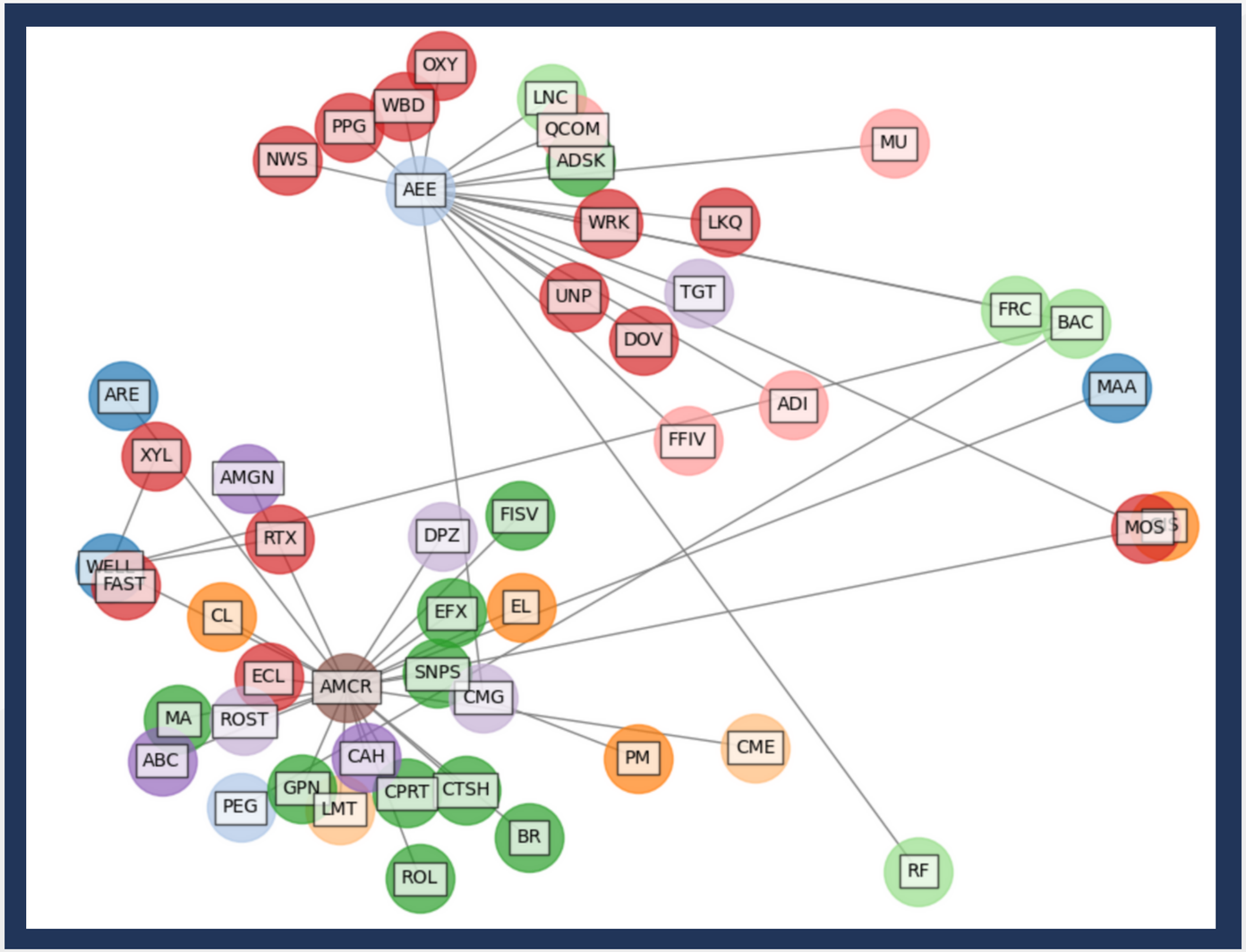
Random matrix theory

Agglomerative hierarchical clustering

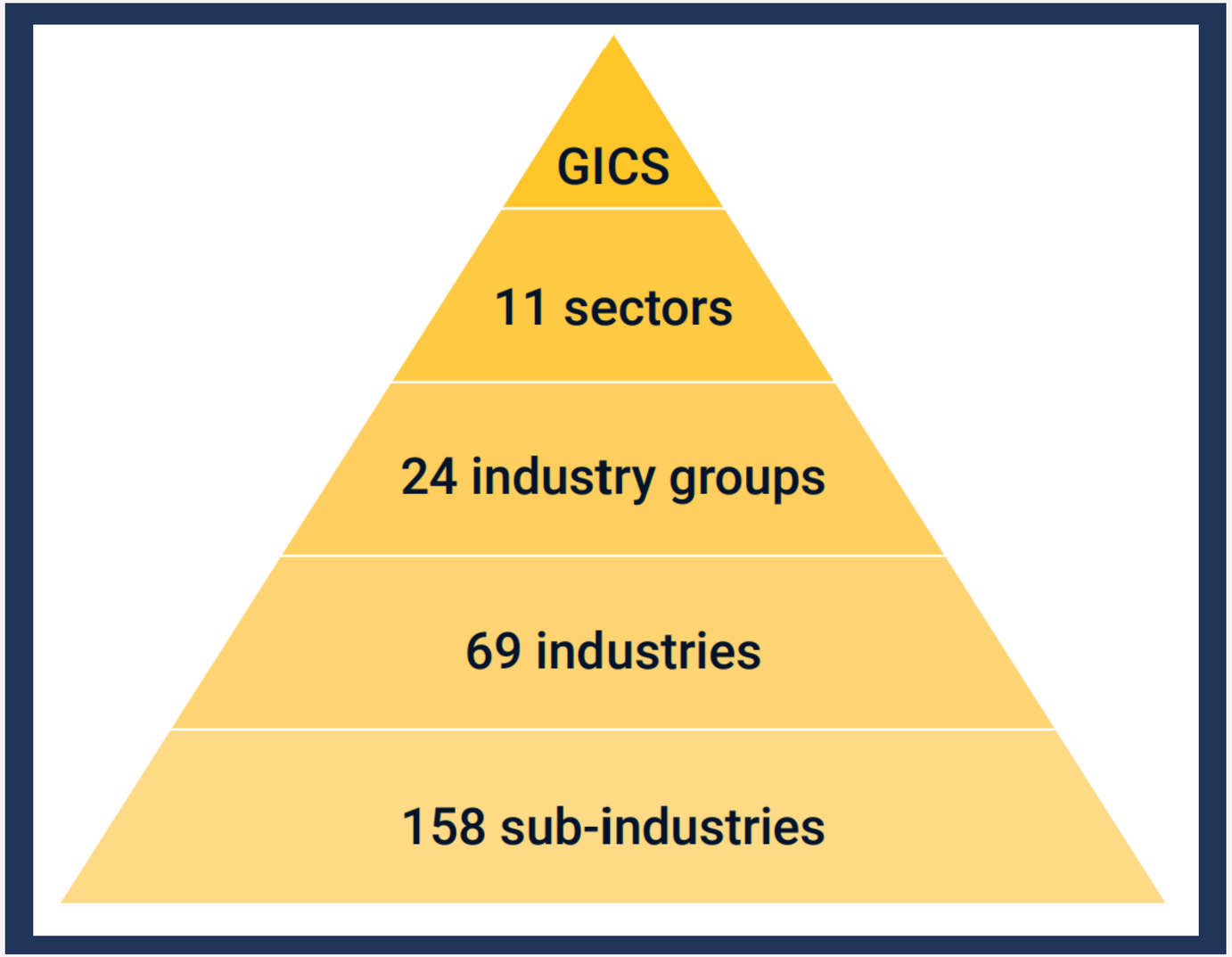
p-median

COMMON POINT

These methods treat stocks like network



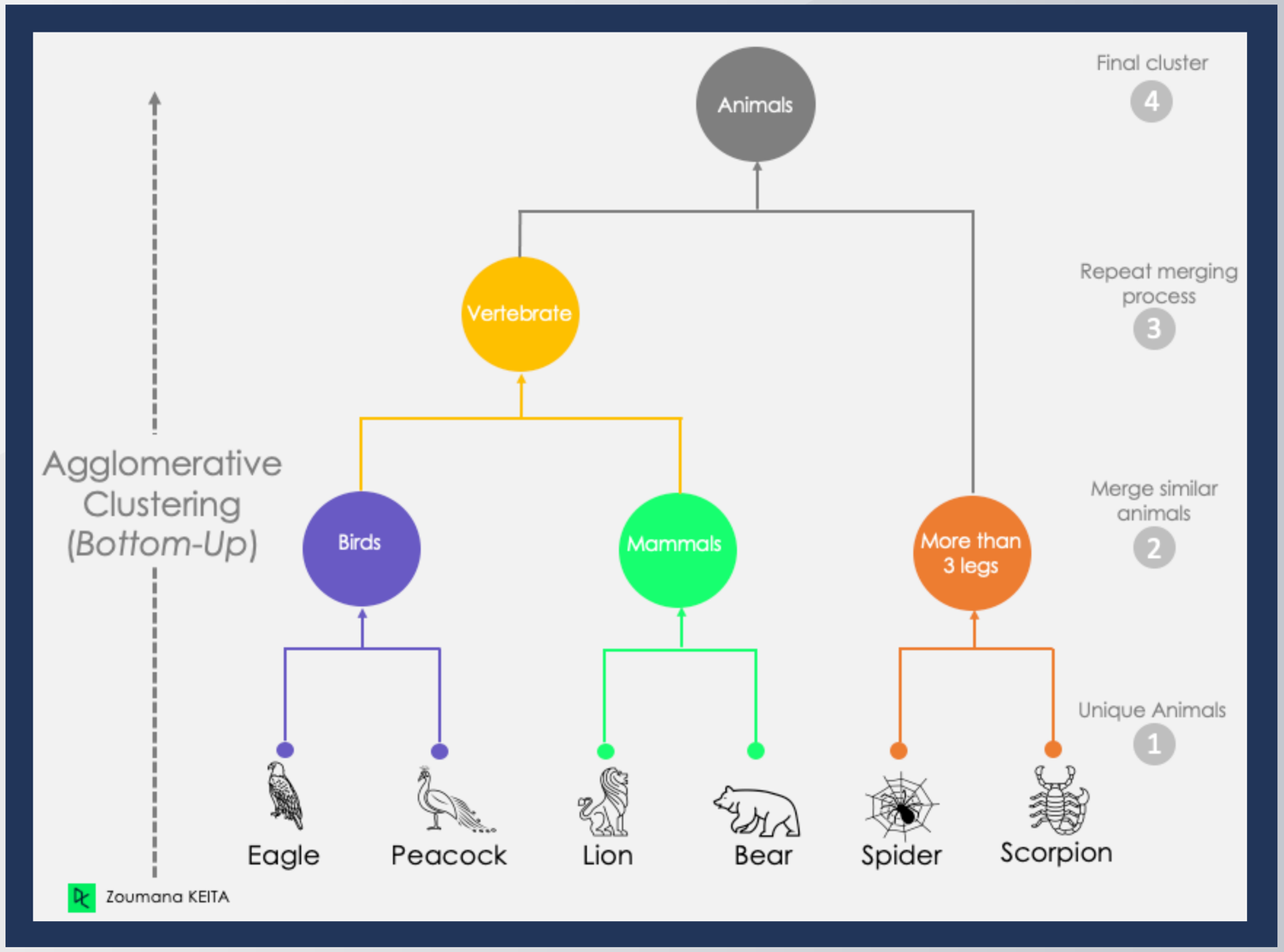
Not only networks, stocks possess hierarchy



HIERARCHICAL CLUSTERING

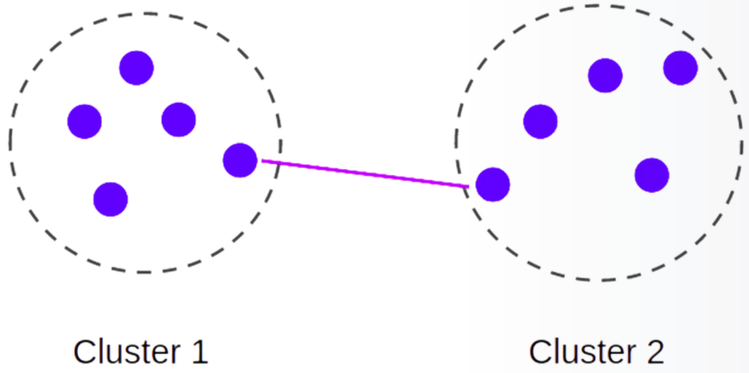
Start from leafnodes, merge

- closest data points to clusters
- closest clusters to bigger clusters



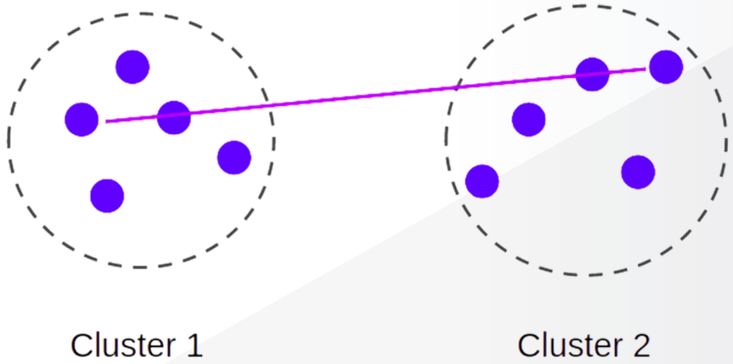
LINKAGE DISTANCE

Single



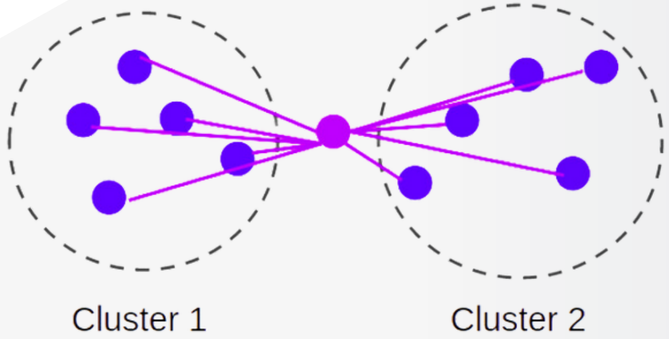
$$d(C_1, C_2) = \min_{x \in C_1, y \in C_2} \{d(x, y)\}$$

Complete



$$d(C_1, C_2) = \max_{x \in C_1, y \in C_2} \{d(x, y)\}$$

Ward



$$d(C_1, C_2) = I(C_1 \cup C_2) - I(C_1) - I(C_2)$$

$$I(C) = \sum_{x \in C} d(x, \mu_C)^2$$

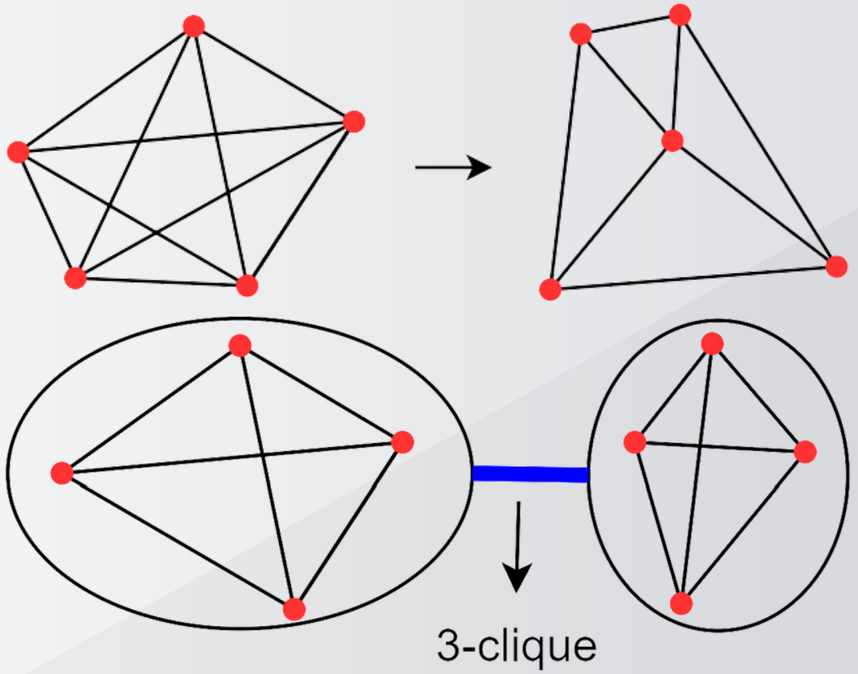
DBHT CLUSTERING

Filter maximal planar graph

Filter bubbles/ low-level clusters

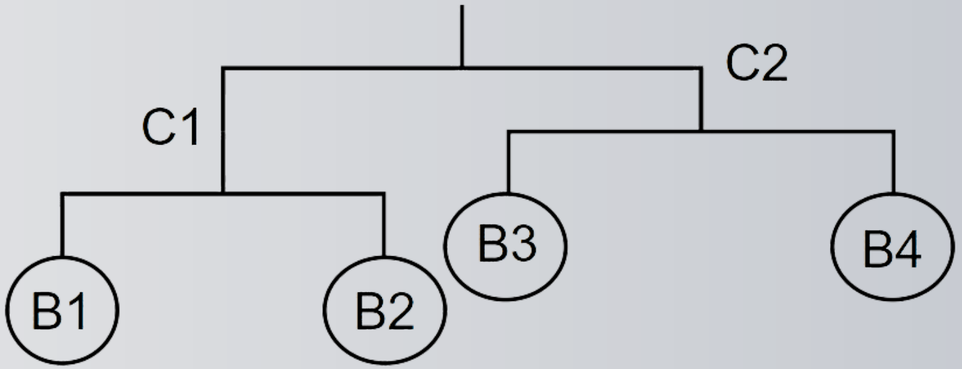
intra-cluster hierarchy

inter-cluster hierarchy



Take important edges

Complete linkage



OUR APPROACH

STANDARD METRICS

Daily relative return of stock i:

$$r_i(t) = \frac{P_i(t)}{P_i(t-1)} - 1$$

Distance between stock i and j:

- **Pearson:**

$$\rho_{ij} = \frac{\mathrm{Cov}[r_i, r_j]}{\sigma(r_i)\sigma(r_j)}$$

- **Spearman's rank:**

$$\rho_{ij} = \frac{\mathrm{Cov}[\mathrm{rank}(r_i), \mathrm{rank}(r_j)]}{\sigma(\mathrm{rank}(r_i))\sigma(\mathrm{rank}(r_j))}$$

$$d_{ij} = \sqrt{2(1 - \rho_{ij})}$$

METHODS

- Linkage algorithms
- Normal DBHT
- Tailored DBHT with ward and single linkage
- Each algorithm tested on both Pearson and Spearman's rank distance.

CONTRIBUTION:

- First comparison between DBHT and Ward
- Improve DBHT

EVALUATION & DISCUSSTION

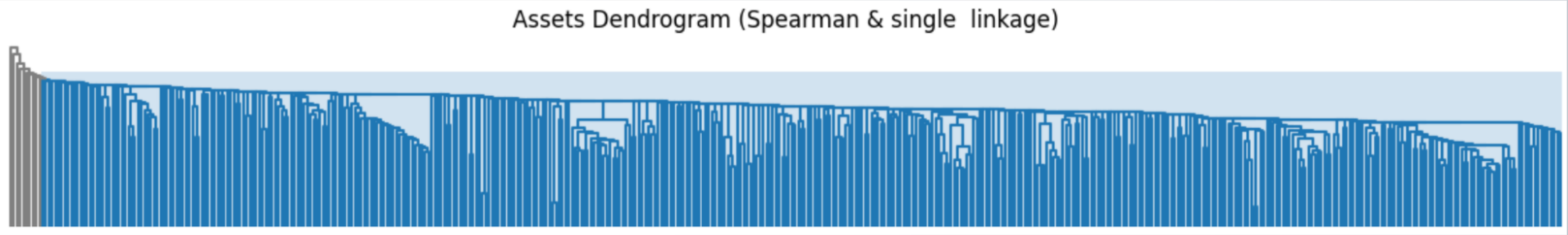
EVALUATION

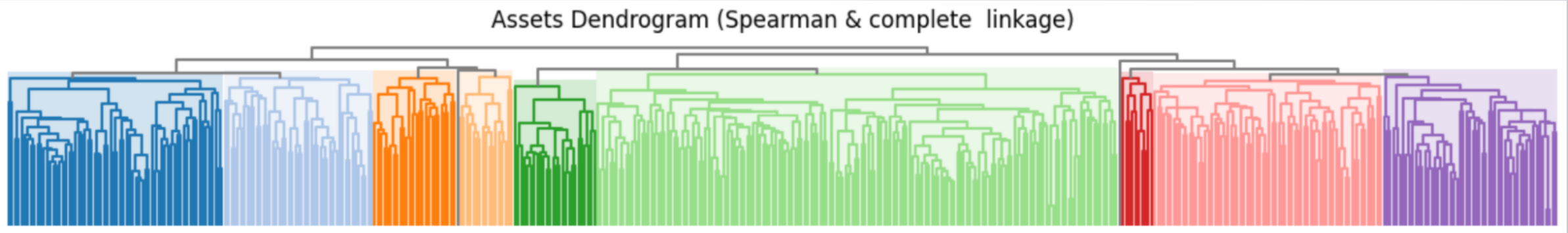
- Technique: Adjusted Rand Index
  - how consistent pairs of datapoints are clustered

- Dataset: SP500 from *2016-01-01* to *2019-12-30*

- Benchmark: GICS taxonomy

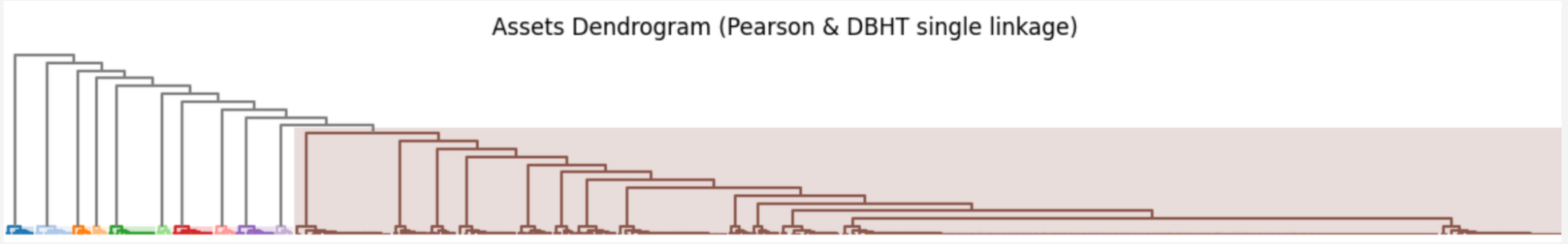
CLUSTERING


Assets Dendrogram (Spearman & single linkage)

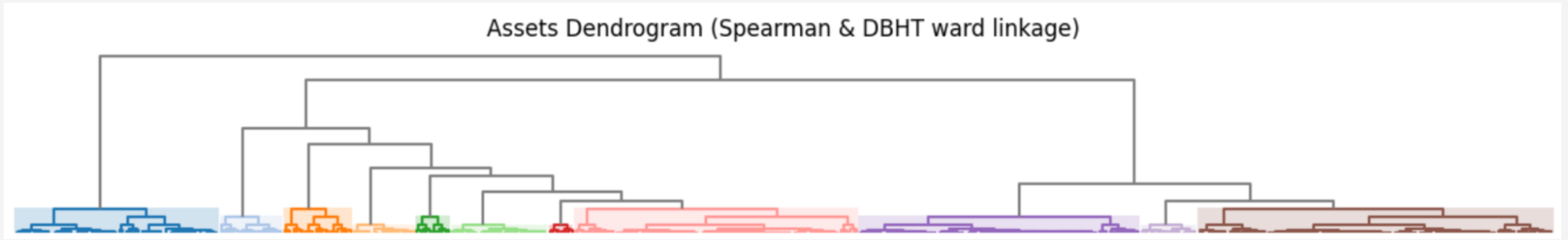

Assets Dendrogram (Spearman & complete linkage)

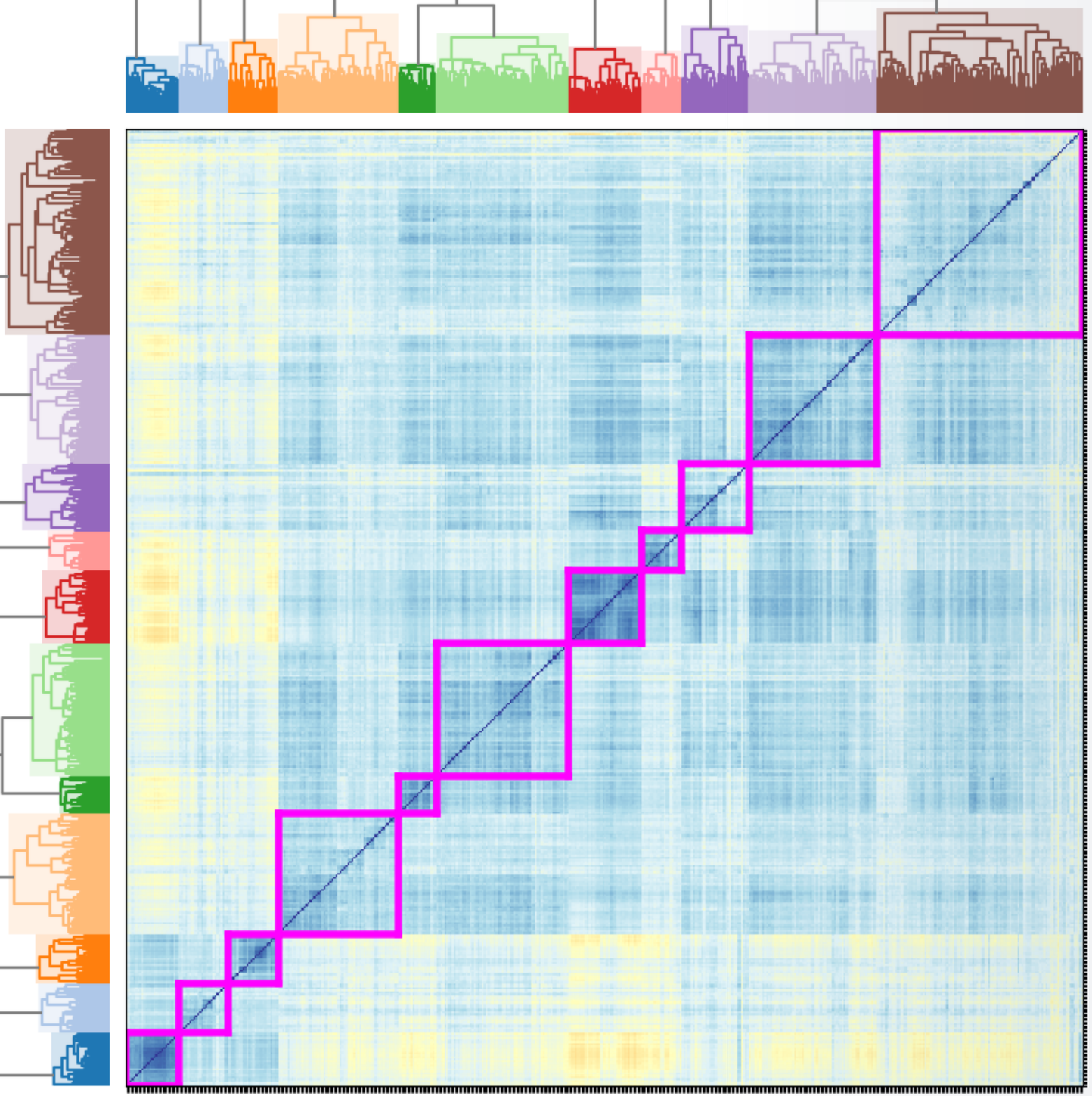

Assets Dendrogram (Spearman & ward linkage)

CLUSTERING

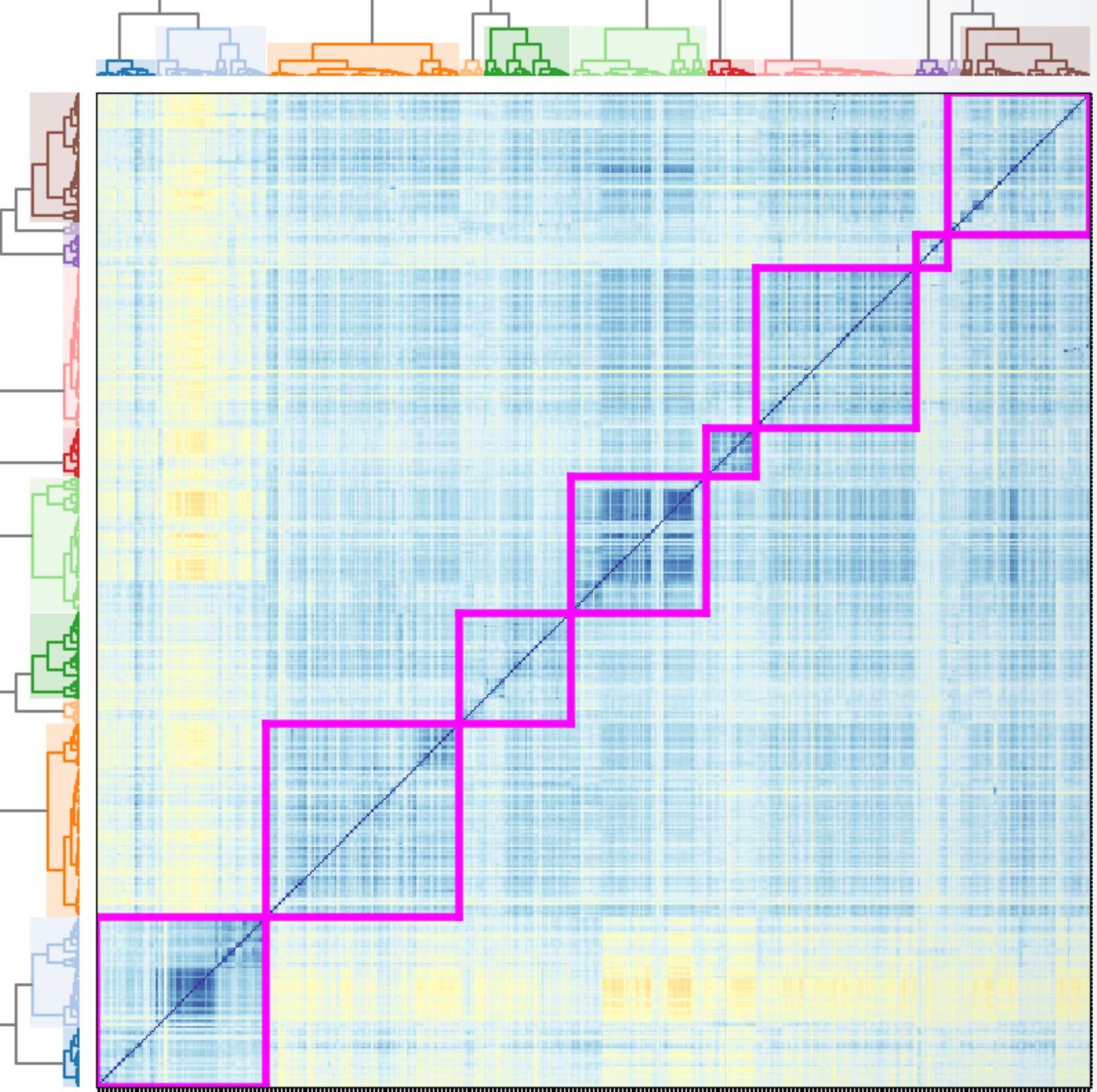

Assets Dendrogram (Pearson & DBHT single linkage)


Assets Dendrogram (Spearman & DBHT complete linkage)


Assets Dendrogram (Spearman & DBHT ward linkage)



Ward Linkage Correlation Matrix



DBHT with Ward Correlation Matrix

RESULT

ARI against GICS	PEARSON	SPEARMAN
Single	0.0016	0.0022
Complete	0.1732	0.2422
Ward	0.4066	0.4460
DBHT single	0.1318	0.1417
DBHT complete	0.3384	0.3701
DBHT ward	0.3969	0.3884

DISCUSSION

- Spearman > Pearson implies the data are less *linear* than they are *monotonic*.
- Single linkage is almost random due to the chaining effect.
- Ward DBHT > Normal DBHT but < Ward Linkage. The output is unstable in Spearman's test.

Speculation: shortest path distance in filtered graph might violate Euclidean assumption.

⟶ Potential room for improvement.

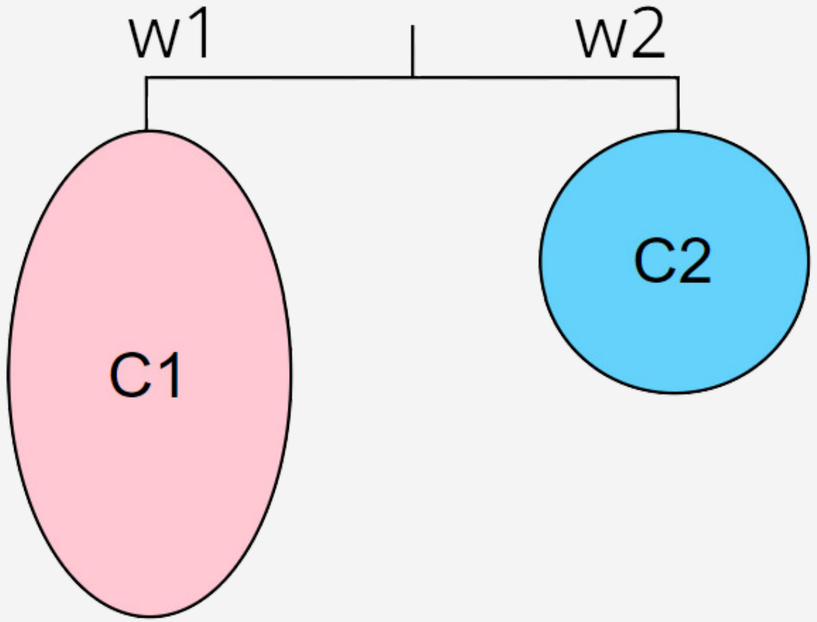
PORTFOLIO CONSTRUCTION

HIERARCHY RISK PARITY

- Allocate assets to minimize risk (variance)

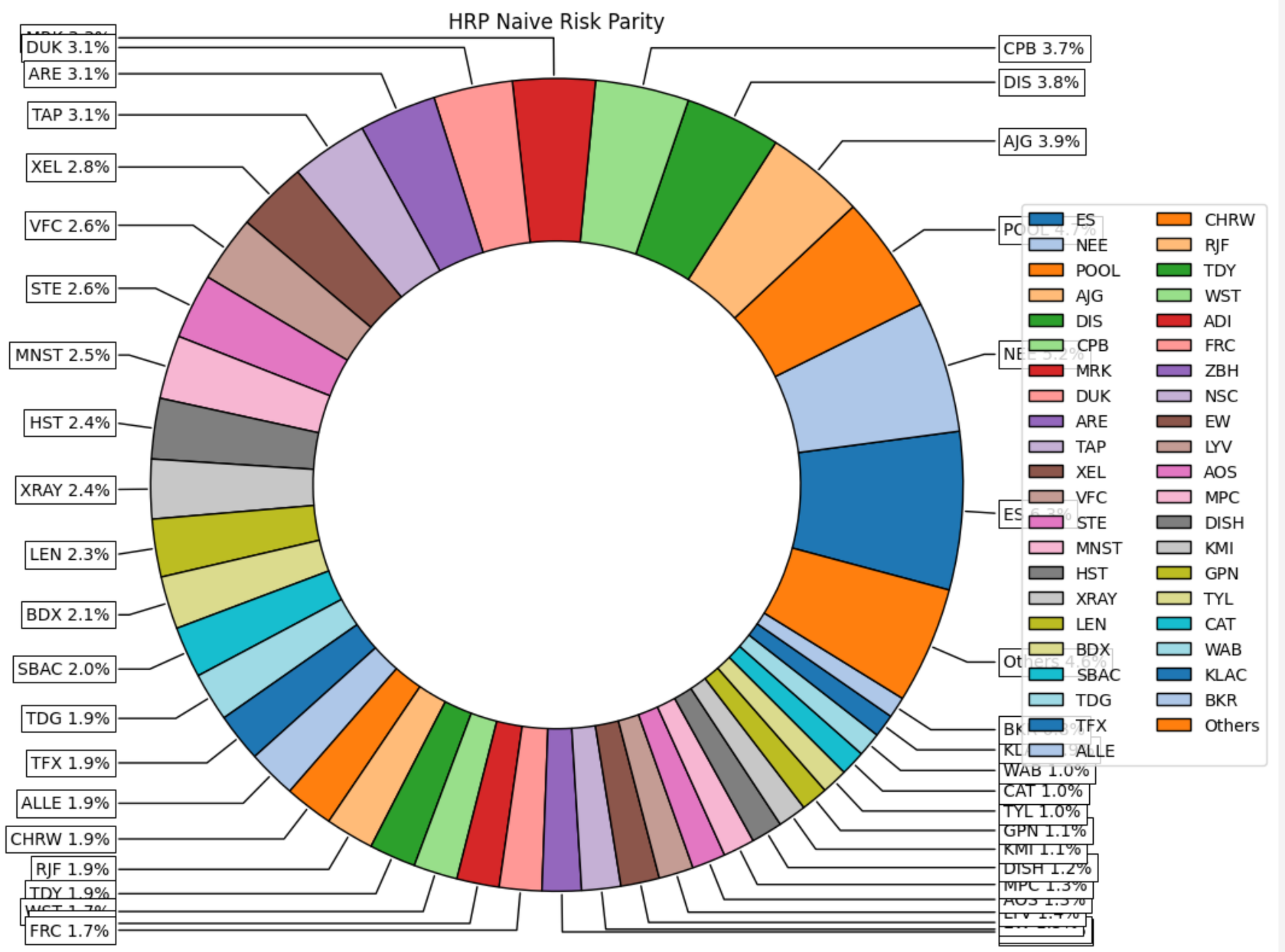
weighted return (wr) = $w_1 r_1 + w_2 r_2$

$$w_1 + w_2 = 1$$



Minimize $\mathrm{Var}[\mathrm{wr}] = w_1^2 \mathrm{Var}[r_1] + w_2^2 \mathrm{Var}[r_2]$

$$\Longrightarrow \begin{cases} w_1 = \dfrac{\mathrm{Var}[r_2]}{\mathrm{Var}[r_1]+\mathrm{Var}[r_2]} \\ w_2 = \dfrac{\mathrm{Var}[r_1]}{\mathrm{Var}[r_1]+\mathrm{Var}[r_2]}. \end{cases}$$


HRP Naive Risk Parity
DUK 3.1%
ARE 3.1%
TAP 3.1%
XEL 2.8%
VFC 2.6%
STE 2.6%
MNST 2.5%
HST 2.4%
XRAY 2.4%
LEN 2.3%
BDX 2.1%
SBAC 2.0%
TDG 1.9%
TFX 1.9%
ALLE 1.9%
CHRW 1.9%
RJF 1.9%
TDY 1.9%
FRC 1.7%
CPB 3.7%
DIS 3.8%
AJG 3.9%
ES
CHRW
NEE
RJF
POOL
TDY
AJG
WST
DIS
ADI
CPB
FRC
MRK
ZBH
DUK
NSC
ARE
EW
TAP
LYV
XEL
AOS
VFC
MPC
STE
DISH
MNST
KMI
HST
GPN
XRAY
TYL
LEN
CAT
BDX
WAB
SBAC
KLAC
TDG
BKR
TFX
Others
ALLE

SUMMARY

SUMMARY

- Compare hierarchical clustering methods
- Discuss some common metrics
- Construct portfolio using stock hierarchy

CITATIONS

[1] Marti, G., Nielsen, F., Bińkowski, M., & Donnat, P. (2021). A review of two decades of correlations, hierarchies, networks and clustering in financial markets. *Progress in information geometry: Theory and applications*, 245-274.

[2] Song, W. M., Di Matteo, T., & Aste, T. (2012). Hierarchical information clustering by means of topologically embedded graphs. *PloS one*, 7(3), e31929.

[3] Lu, Y (2017). Application of Clustering Methods to Trading Strategies in the US Equity Market.

[4] De Prado, M. L. (2016). Building diversified portfolios that outperform out of sample. The Journal of Portfolio Management, 42(4), 59-69.

[5] Randriamihamison, N., Vialaneix, N., & Neuvial, P. (2021). Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. Journal of Classification, 38(2), 363-389.

THANK YOU FOR LISTENING