Universidad Tecnológica de Pereira

MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA

Multivariate statistics

Introductory session

May 17, 2024

Orlando Sabogal-Cardona
@Antonio Sabogal
orlando.sabogal.20@ucl.ac.uk

Who

What

How

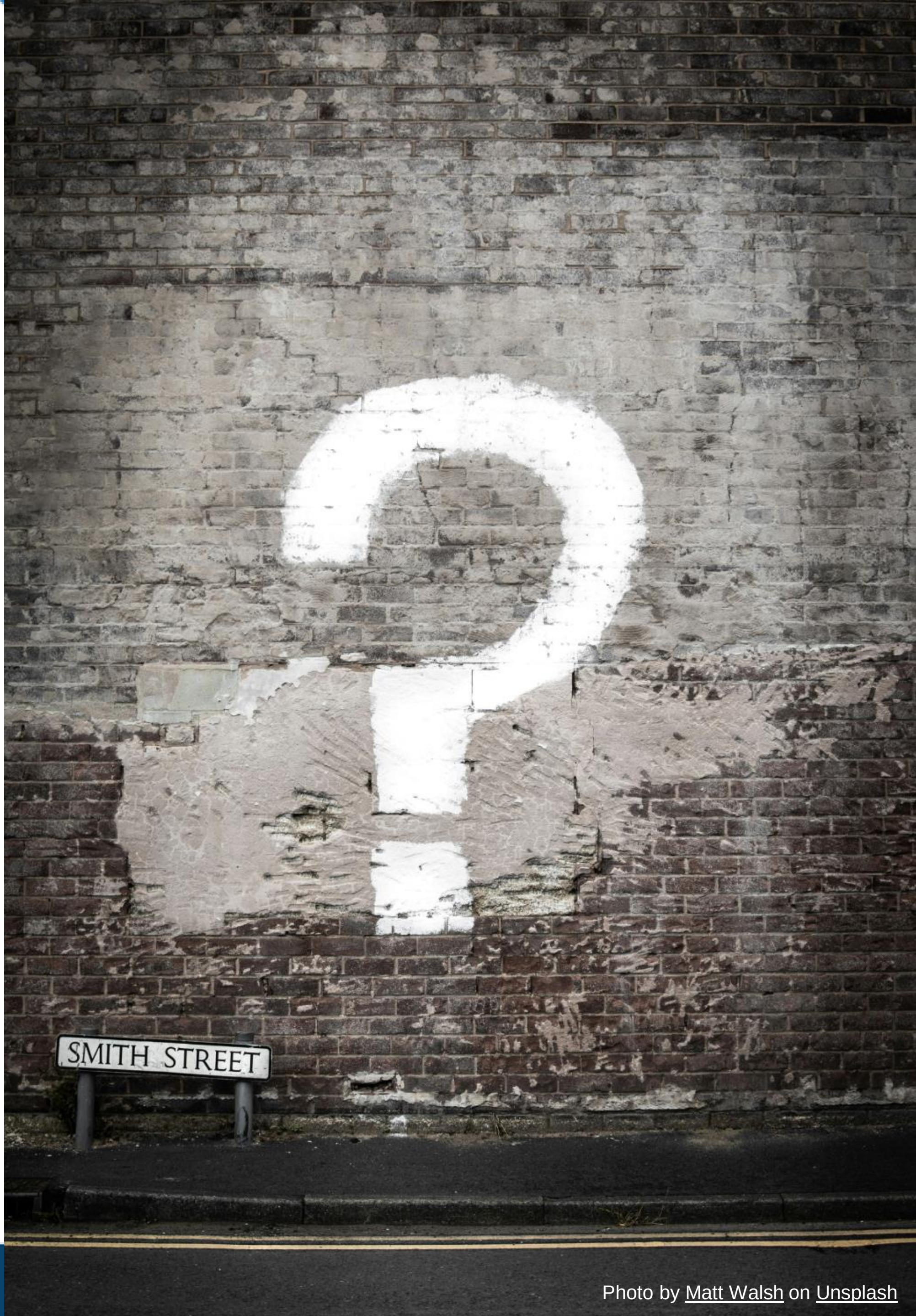


Photo by Matt Walsh on Unsplash

Icebreaker: round of introductions

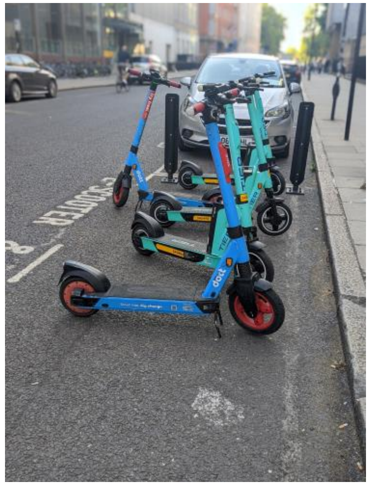
About me:

- UTP Alumni
- IDB Consultant
- UCL student (Development Planning)
- R user, advocate, and activist
- Transport

IDB dpu Development Planning Unit UCL THE BARTLETT

My current research agenda

APP-based mobility ABM



Leisure cycling

Walkability

Sustainability and climate change

Social inclusion

Well-being

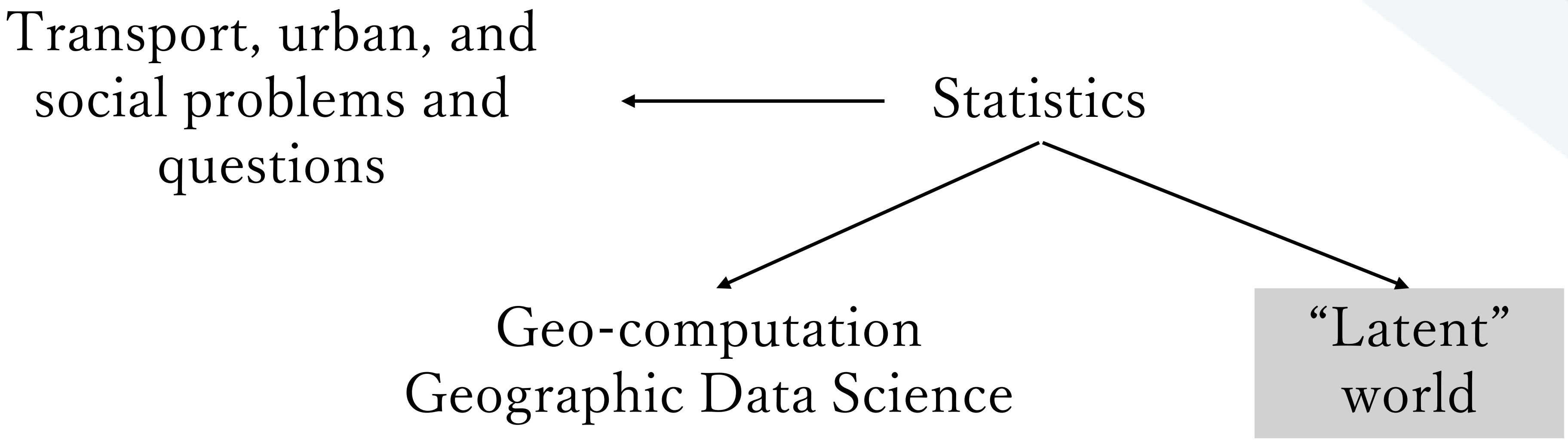
Statistics is not really about statistics.

Statistics is not really about statistics.
But it is all about statistics.

"Strong familiarity with the theoretical and empirical literature in your research area is the single most important thing you could bring to SEM. This is because everything— from the specification of your initial model to modification of that model in subsequent reanalyses to interpretation of the results— must be guided by your domain knowledge. So, you need, first and foremost, to be a researcher , not a statistician or a computer nerd. This is true for most kinds of statistical analysis in that the value of the product (numerical results) depends on the quality of the ideas (your hypotheses) on which the analysis is based.“

Kline, Rex B.. Principles and Practice of Structural Equation Modeling (2023), Guilford Publications
(Please ignore “computer nerd”)

What

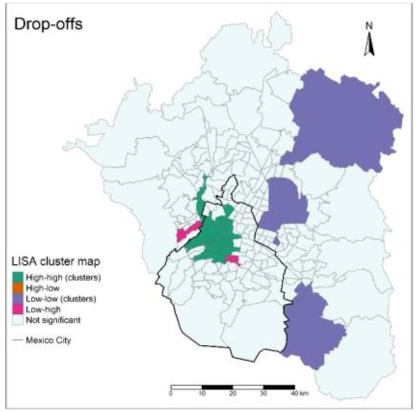


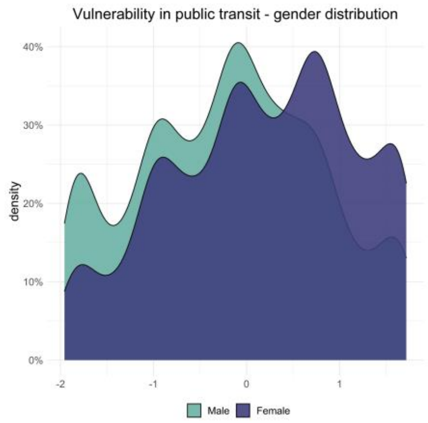
Transport, urban, and social problems and questions

Statistics

Geo-computation
Geographic Data Science

“Latent” world



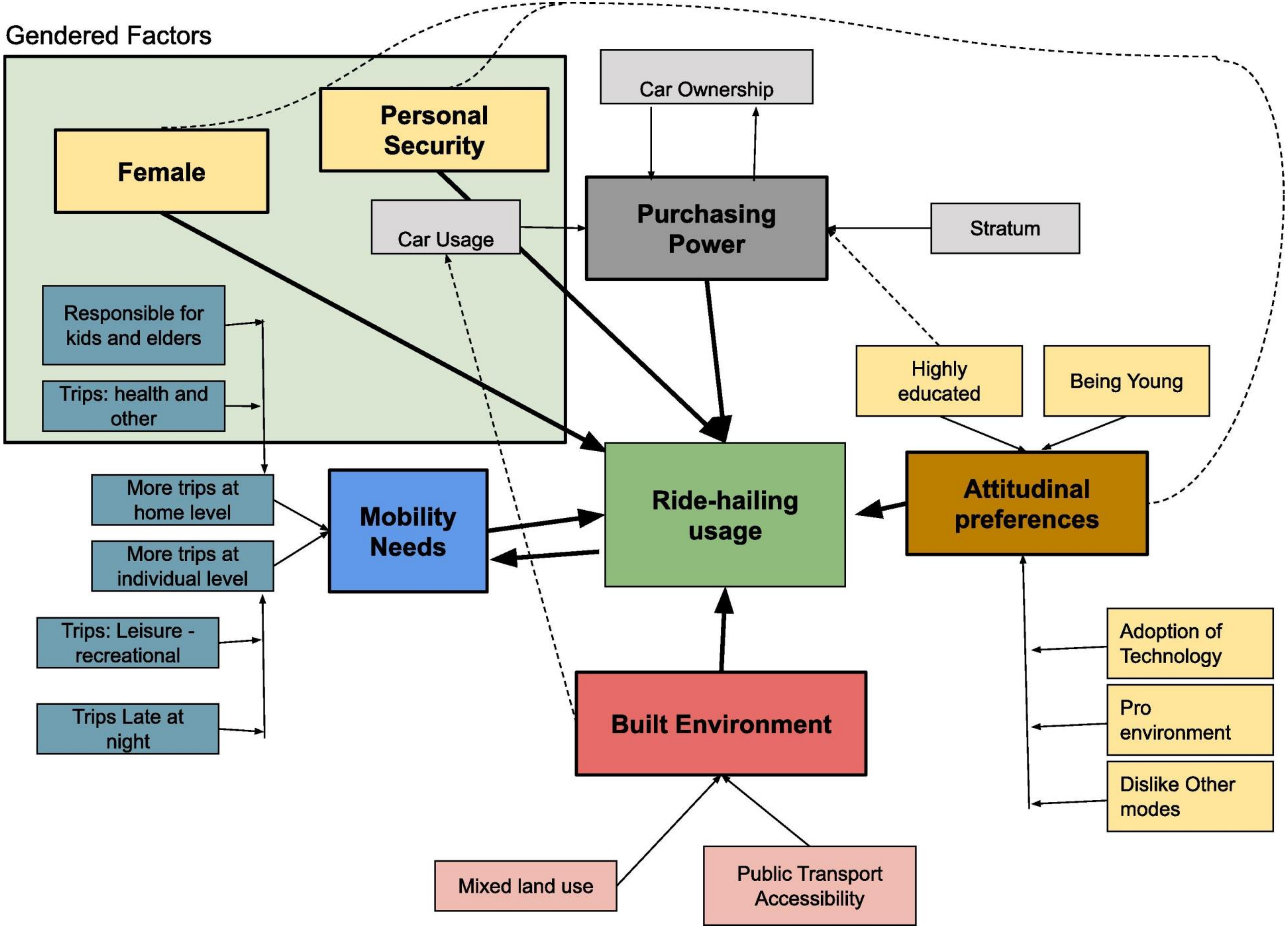


Multivariate Analysis

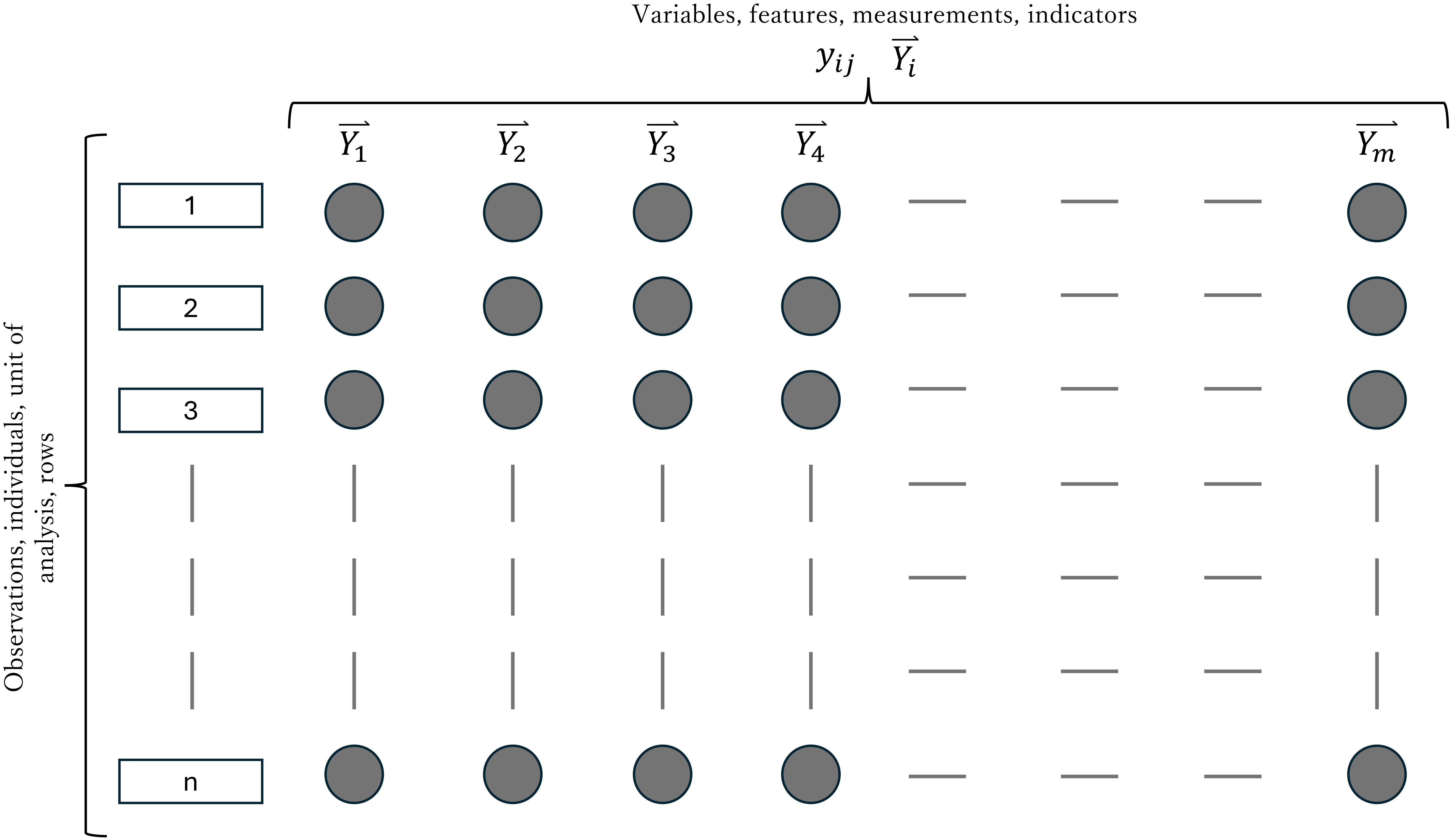
"Multivariate analysis refers to all statistical techniques that simultaneously analyze multiple measurements on individuals or objects under investigation. Thus, any simultaneous analysis of more than two variables can be loosely considered multivariate analysis" (Multivariate Data Analysis, Hair et al., 2019,page 9)

Multivariate Analysis

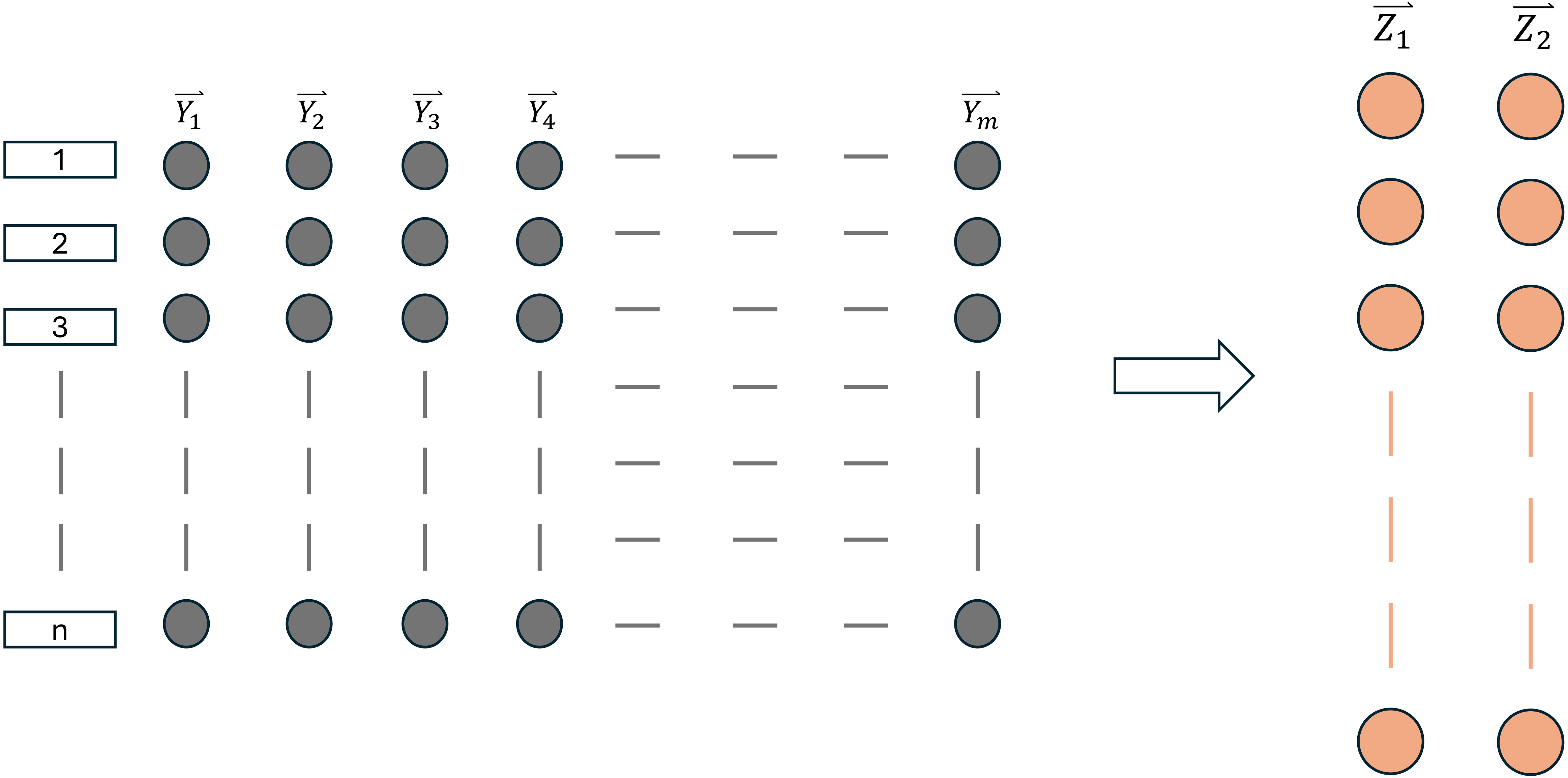
"Some authors state that the purpose of multivariate analysis is to measure, explain, and predict the degree of relationship among variates (weighted combinations of variables). Thus, the multivariate character lies in the multiple variates (multiple combinations of variables), and not only in the number of variables or observations. For the purposes of this book, we do not insist on a rigid definition of multivariate analysis"(Multivariate Data Analysis, Hair et al., 2019,page 10)



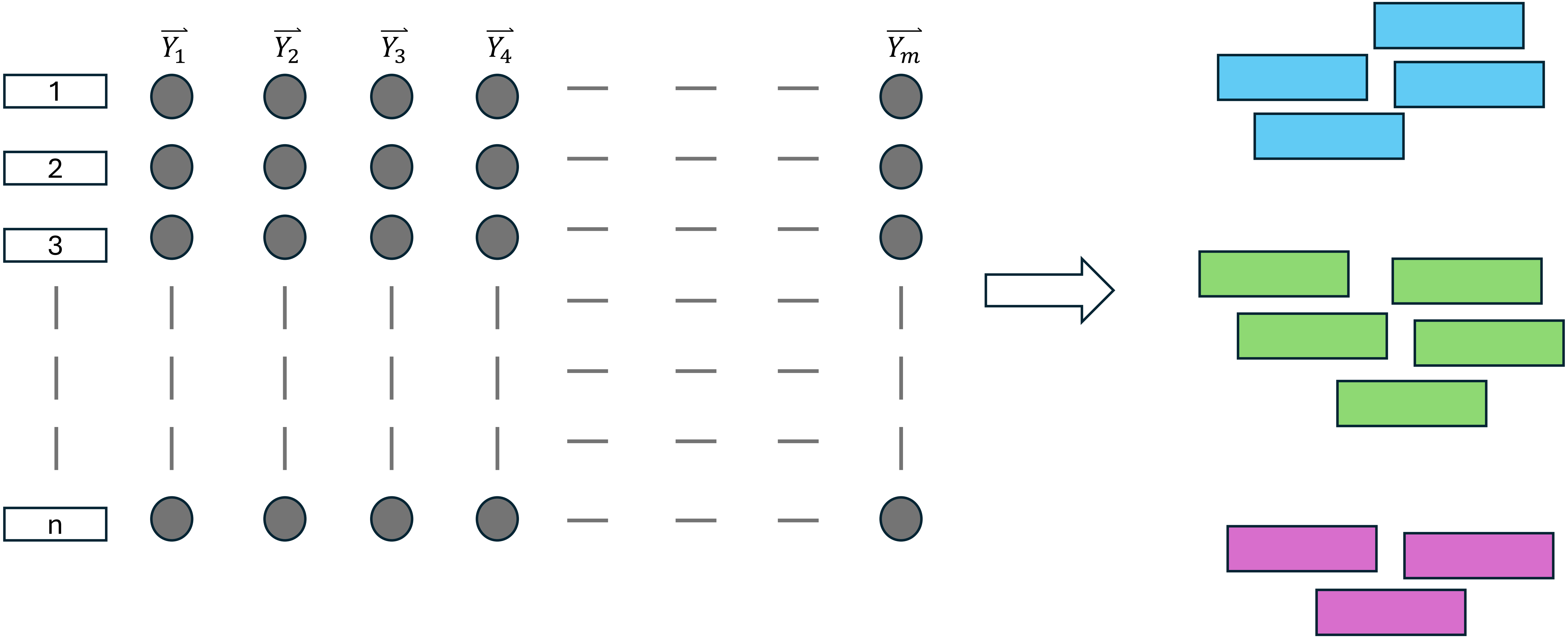
“Not my usual trip” Sabogal-Cardona et al., (2021)


Variables, features, measurements, indicators
y_{ij}
$\overrightarrow{Y_i}$
$\overrightarrow{Y_1}$
$\overrightarrow{Y_2}$
$\overrightarrow{Y_3}$
$\overrightarrow{Y_4}$
$\overrightarrow{Y_m}$
Observations, individuals, unit of analysis, rows
1
2
3
n

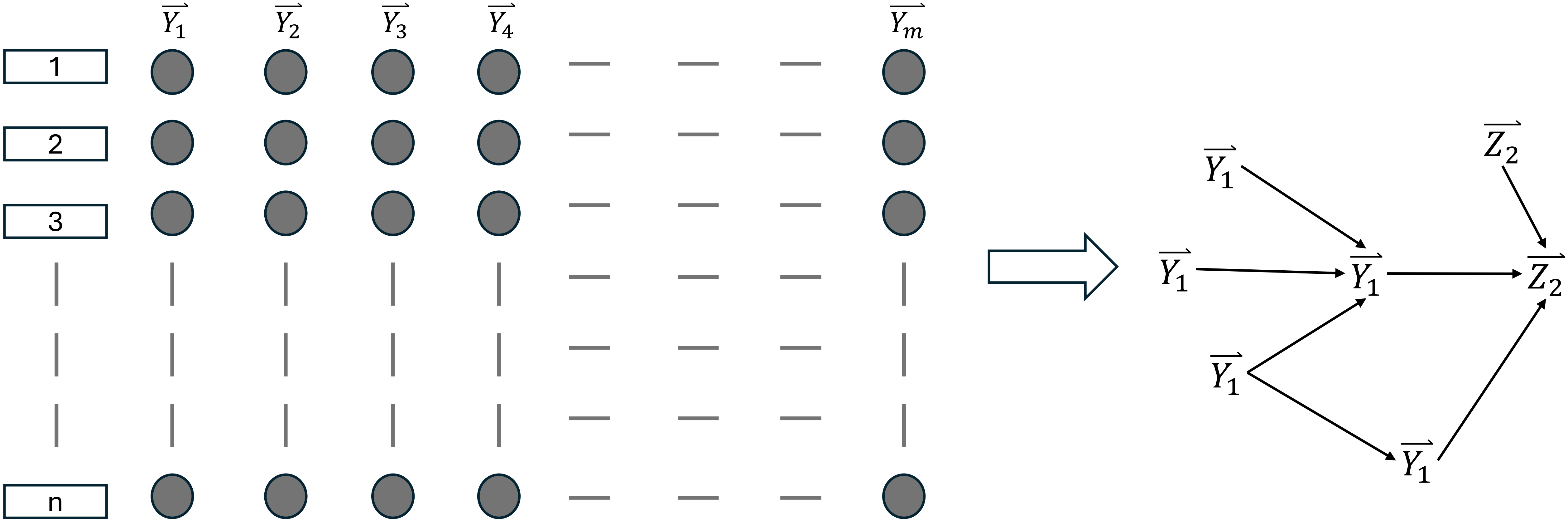
Dimensionality reduction (not really)



Clustering


$\overrightarrow{Y_1}$
$\overrightarrow{Y_2}$
$\overrightarrow{Y_3}$
$\overrightarrow{Y_4}$
$\overrightarrow{Y_m}$
1
2
3
n

Association and causality (?)



Multivariate Normal Distribution

Generalization of the one-dimensional normal distribution to higher dimensions.

$$X \sim N\left(\mu_X, \sigma_X^2\right) \longrightarrow \boldsymbol{X} \sim N(\boldsymbol{M}, \boldsymbol{\Sigma}) \; ; \; \begin{Bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{Bmatrix} \sim N\left(\begin{Bmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \vdots \\ \mu_{x_n} \end{Bmatrix}, \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{X_1,X_2} & \cdots & \sigma_{X_1,X_n} \\ \sigma_{X_2,X_1} & \sigma_{x_2}^2 & & \vdots \\ \vdots & & \ddots & \\ \sigma_{X_n,X_1} & \dots & & \sigma_{x_n}^2 \end{bmatrix} \right)$$

Variance-Covariance matrix

$$S_{x_i}^2 = \frac{1}{(n - 1)} \sum_{j=1}^{n} \left(X_{i_i} - \overline{X}_i\right)^2 \quad ; \quad \sigma_{x_i}^2 = \frac{1}{m} \sum_{j=1}^{m} \left(X_{i_j} - \mu_i\right)^2$$

$$\begin{bmatrix} \sigma_{x_1}^2 & \sigma_{X_1,X_2} & \cdots & \sigma_{X_1,X_n} \\ \sigma_{X_2,X_1} & \sigma_{x_2}^2 & & \vdots \\ \vdots & & \ddots & \\ \sigma_{X_n,X_1} & \ldots & & \sigma_{x_n}^2 \end{bmatrix}$$

$$S_{X_i,X_k} = \frac{1}{n - 1} \sum_{j=n}^{n} \left(X_{i_j} - \overline{X}_i\right)\left(X_{k_i} - \overline{X_k}\right)$$

$$r_{X_i,X_k} = \frac{S_{X_i,X_k}}{S_{x_i}^2 S_{x_k}^2}$$

Variance-Covariance matrix

$$S^2_{x_i} = \frac{1}{(n\ -1)}\sum_{j=1}^{n}\left(X_{i_i}\ -\ \bar{X}_i\right)^2 \quad ; \quad \sigma^2_{x_i} = \frac{1}{m}\sum_{j=1}^{m}\left(X_{i_j}\ -\ \mu_i\right)^2$$

$$\begin{bmatrix} \sigma^2_{x_1} & \sigma_{X_1,X_2} & \cdots & \sigma_{X_1,X_n} \\ \sigma_{X_2,X_1} & \sigma^2_{x_2} & & \vdots \\ \vdots & & \ddots & \\ \sigma_{X_n,X_1} & \cdots & & \sigma^2_{x_n} \end{bmatrix}$$

$$S_{X_i,X_k} = \frac{1}{n\ -1}\sum_{j=n}^{n}\left(X_{i_j}\ -\ \bar{X}_i\right)\left(X_{k_i}\ -\ \overline{X_k}\right)$$

$$r_{X_i,X_k} = \frac{S_{X_i,X_k}}{S^2_{x_i}S^2_{x_k}}$$

For the purpose of this module: the holy grail!

Normal Multivariate

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma_x}}\, e^{\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right)} \longrightarrow f_X(\boldsymbol{X}) = \frac{1}{(2\pi)^{\frac{n}{2}}(det\boldsymbol{\Sigma})^{\frac{1}{2}}}\, exp\left[-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right]$$

Normal Multivariate

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma_x}} e^{\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right)} \longrightarrow f_X(\boldsymbol{X}) = \frac{1}{(2\pi)^{\frac{n}{2}}(det\boldsymbol{\Sigma})^{\frac{1}{2}}} \, exp\left[-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{M})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{M})\right]$$

Normality:	Shapiro-Wilk Kolmogorov-Smirnov Anderson-Darling	Mardia Henze-Zirkler Royston Doornik-Hansen Energy
Groups:	ANOVA t-test	MANOVA Hotelling’s T2
Distances:	Euclidean	Mahallanobis

Old content

- Review: probability, calculus, linear algebra, linear regression, statistics
- Normal multivariate
- Hotelling's T-square
- Logistic regression and GLM
- Missing data
- Repeated measures
- Linear discriminant
- Canonical correlation
- Correspondence Analysis
- Conjoint analysis
- MDS
- MANOVA
- Decision trees
- PCA
- EFA
- CFA
- SEM
- And more

Old content

- Review: probability, calculus, linear algebra, linear regression, statistics
- Normal multivariate
- Hotelling's T-square
- Logistic regression and GLM
- Missing data
- Repeated measures
- Linear discriminant
- Canonical correlation
- Correspondence Analysis
- Conjoint analysis
- MDS
- MANOVA
- Decision trees
- PCA
- EFA
- CFA
- SEM
- And more

- ➢ I am assuming you have some specific background (calculus, linear algebra, regression, probability, ancient Greek)
- ➢ Covered in other modules
- ➢ Absorbed by modern data science
- ➢ Not that popular anymore
- ➢ Very popular and colonizing the field

Primary topics

- ➢ PCA
- ➢ Factor Analysis (EFA + CFA)
- ➢ SEM

The "how"

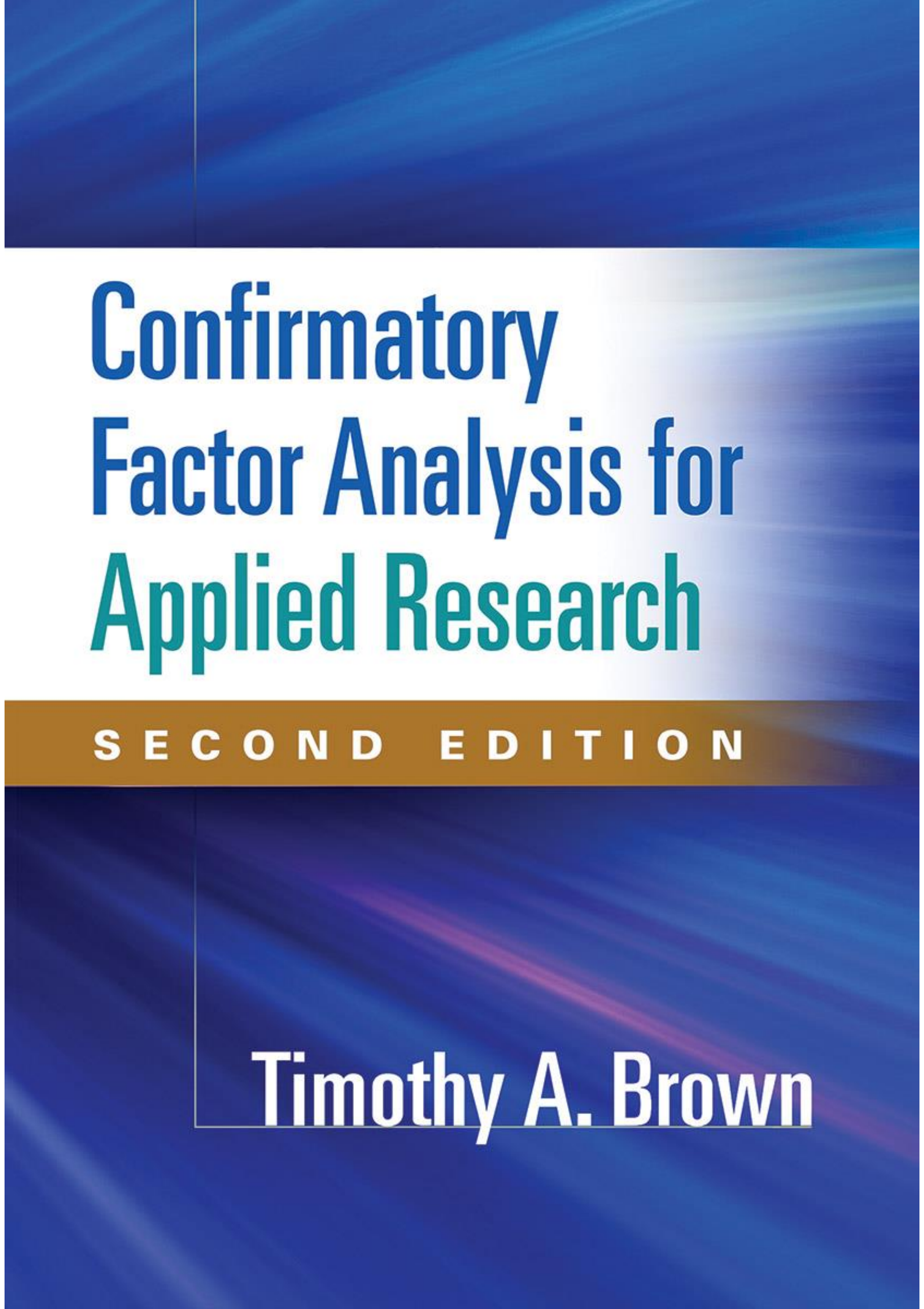
- Lectures
- Concepts and intuition over math
- You will have to read
- R workshops

Relevant material

Multivariate Data Analysis
(Joseph F. Hair Jr., William C. Black, Barry J. Babin, Rolph E. Anderson)

Exploratory Factor Analysis
(W. Holmes Finch)

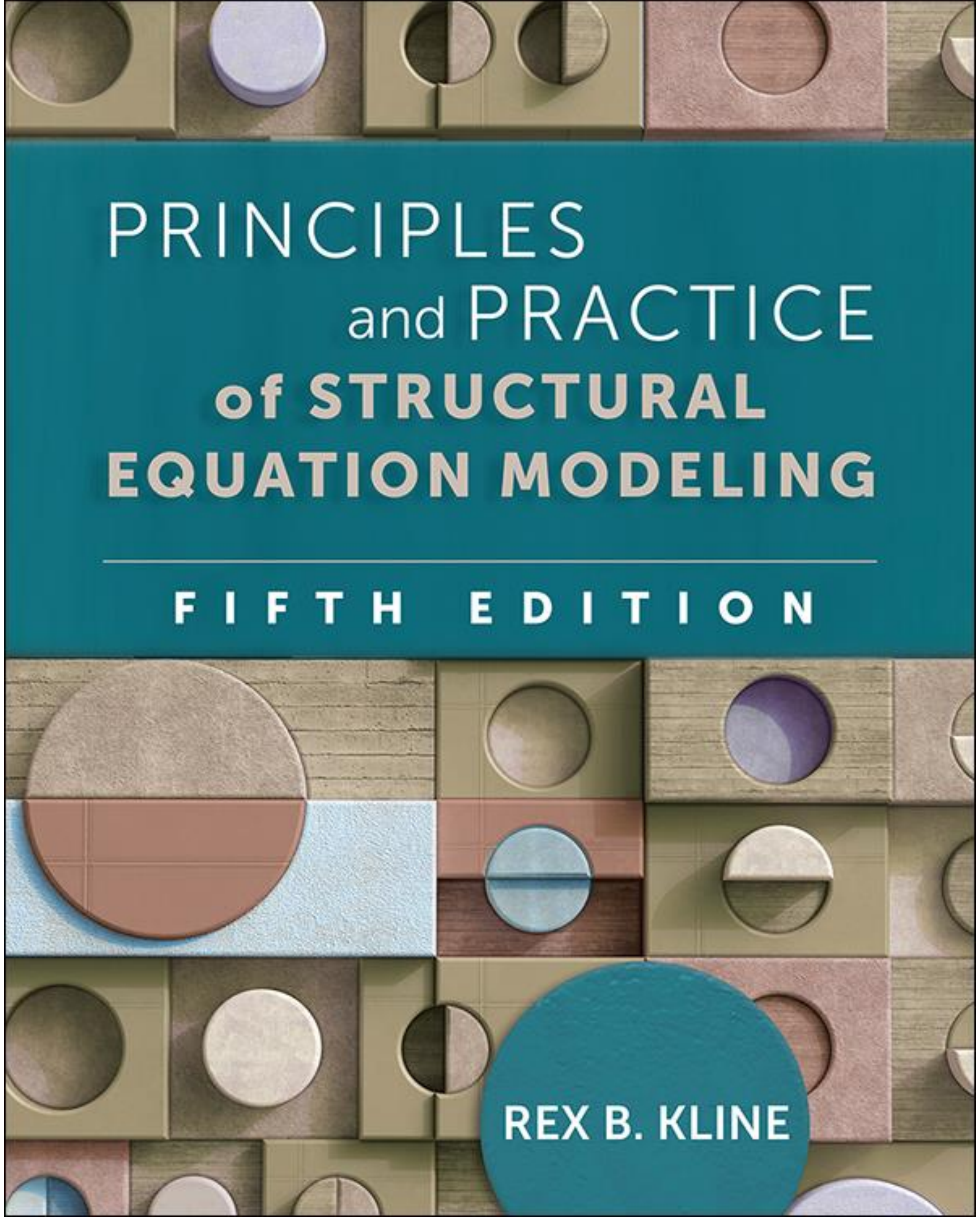
Structural Equation Modeling with lavaan
(Kamel Gana and Guillaume Broc)



Confirmatory Factor Analysis for Applied Research

SECOND EDITION

Timothy A. Brown

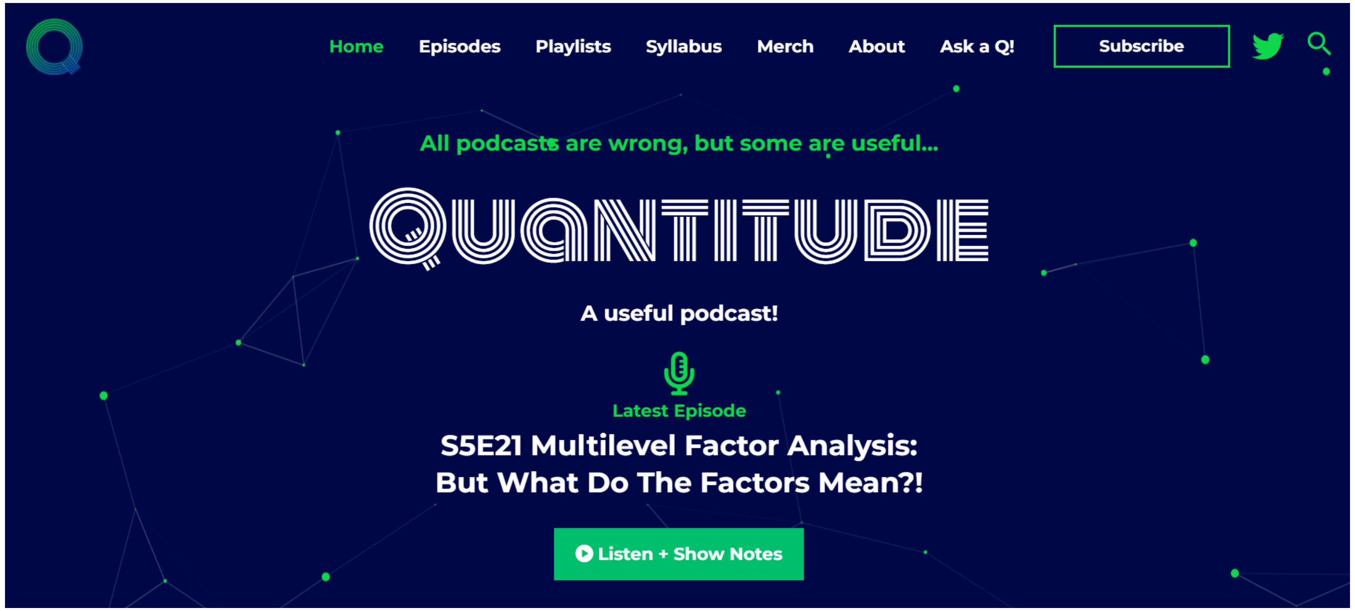


PRINCIPLES and PRACTICE of STRUCTURAL EQUATION MODELING

FIFTH EDITION

REX B. KLINE

https://quantitudepod.org/



All podcasts are wrong, but some are useful...

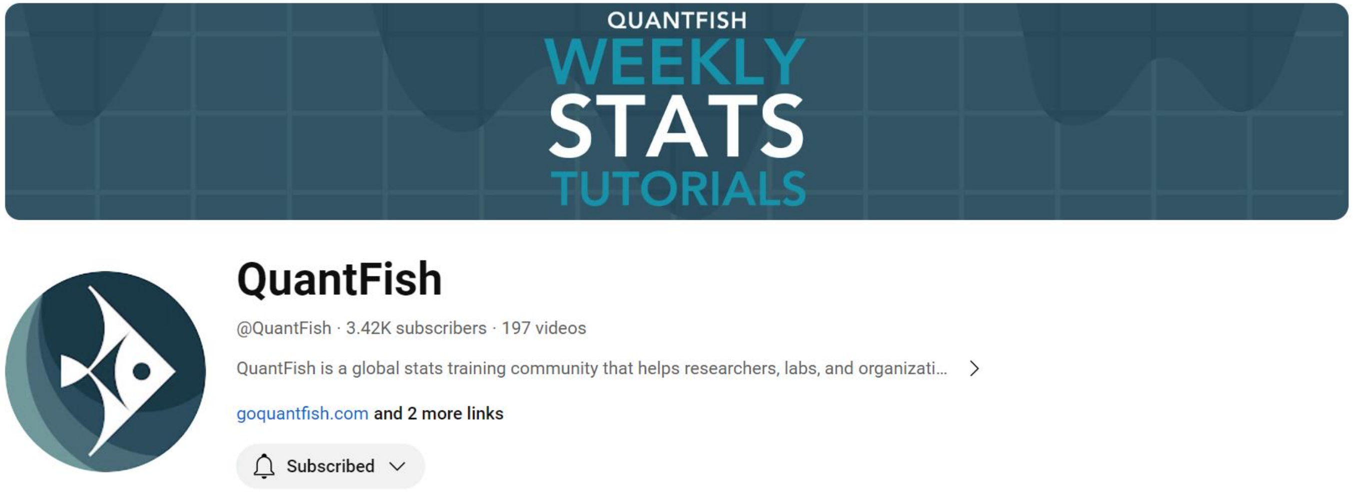
QUANTITUDE

A useful podcast!

Latest Episode

S5E21 Multilevel Factor Analysis: But What Do The Factors Mean?!

https://www.youtube.com/@QuantFish


QUANTFISH WEEKLY STATS TUTORIALS

QuantFish

@QuantFish · 3.42K subscribers · 197 videos

QuantFish is a global stats training community that helps researchers, labs, and organizati...

goquantfish.com and 2 more links

The “how”: grading

The “how”: housekeeping rules

- We all are grown-ups
- Flexibility
- Be on time!
- I am ok with electronic devices
- I am ok with drinks and snacks
- I am ok with LLMs (e.g., ChatGPT)

Thank you!

Orlando Sabogal-Cardona
@Antonio Sabogal
orlando.sabogal.20@ucl.ac.uk