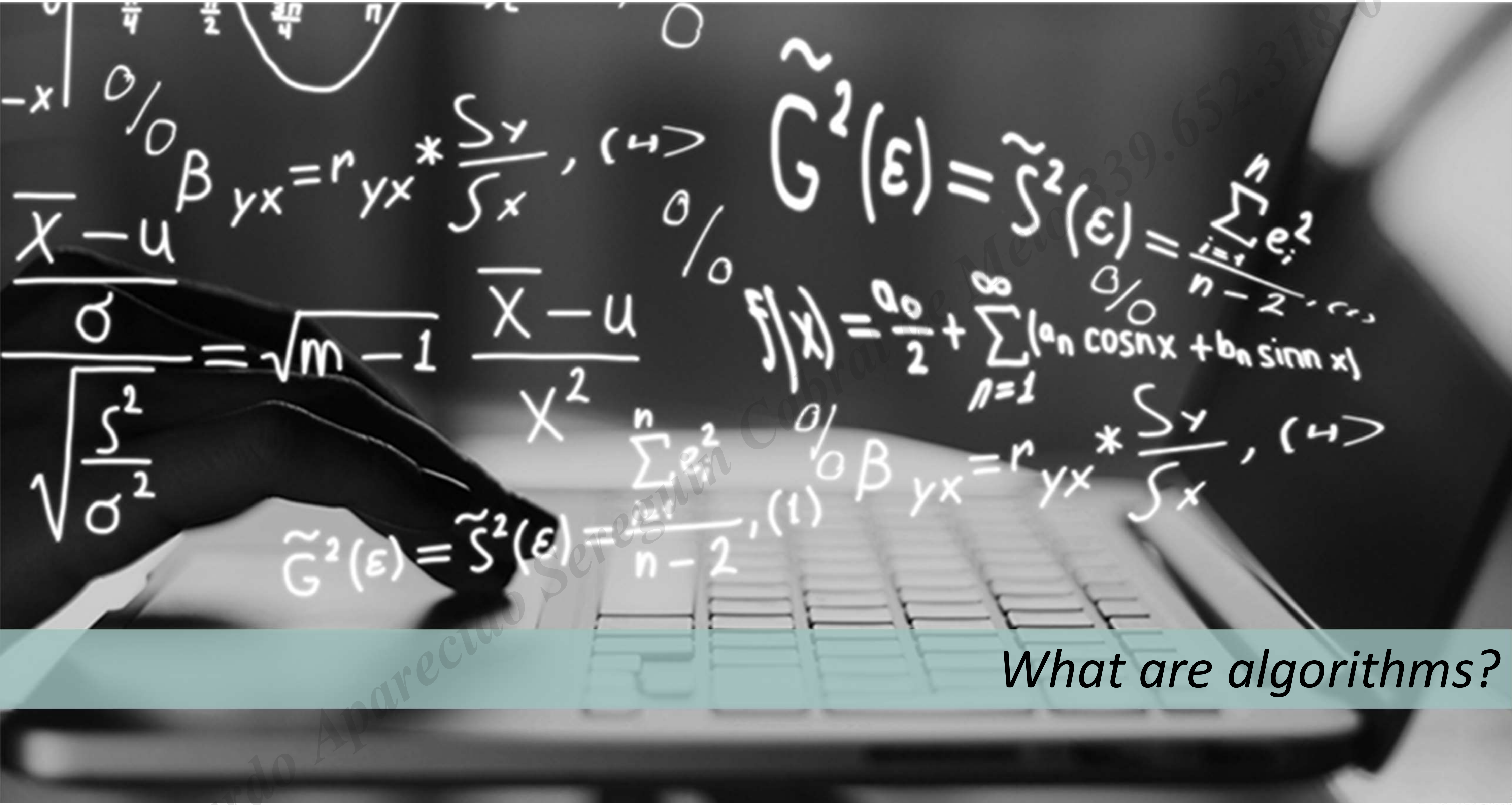
MBA
USP
ESALQ

Database Structures, Types of Variables and Measurement Scales

Rafael de Freitas Souza

Data Science

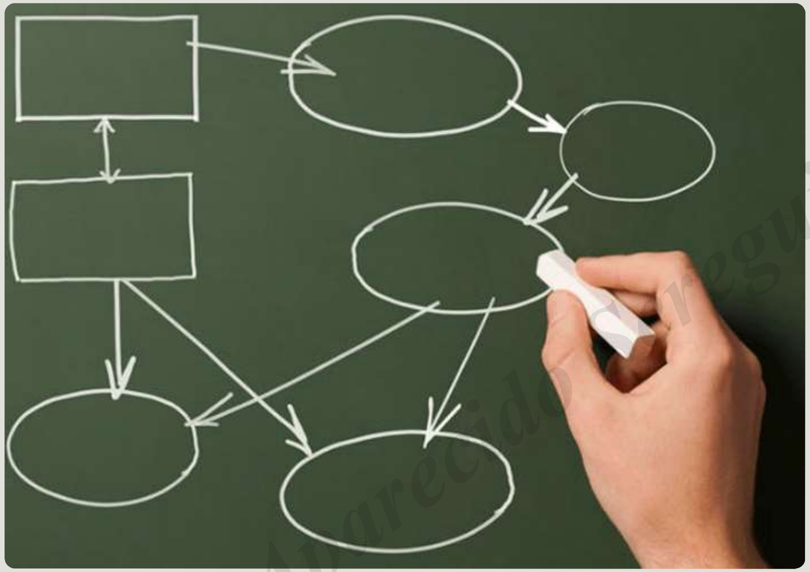
Set of programming techniques focused on data collection, processing , manipulation, organization, analysis, extraction of information, and presentation, in the form of reports or graphics, in order to support the decision-making process.



%
$\beta_{yx} = r_{yx} * \frac{S_y}{S_x}$, (4)
$\tilde{G}^2(\varepsilon) = \tilde{S}^2(\varepsilon) = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$, (1)
$\frac{\frac{\bar{X}-u}{\sigma}}{\sqrt{\frac{S^2}{\sigma^2}}} = \sqrt{m-1}\,\frac{\bar{X}-u}{X^2}$
$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)$

What are algorithms?

Algorithms – a concept:



Algorithms are explicit, literal, limited and systemic sequences of instructions and operations directed to the achievement of a preset objective data.

Basically, any known verb, as long as it denotes an action intended for humans, can be considered an algorithm.

Link: https://www.youtube.com/watch?v=pdhqwbUWf4U

Basic Taxonomy of Species of *Machine Learning* Algorithms

UNSUPERVISED

Machine learning techniques based on **unsupervised** algorithms (unsupervised *learning*) do not have the capacity of inference. They are dedicated to an exploratory, diagnostic analysis, of a studied phenomenon. Common examples: Clusters Analysis, Factorial Analysis of Main Components, Simple and Multiple Correspondence Analysis, etc.

SUPERVISED

Machine learning techniques based on **supervised** algorithms (*supervised learning*) have the capacity of inference. Therefore, they are dedicated to a confirmatory, predictive analysis of a given studied phenomenon. Common examples: Linear Regressions, Logistic Regressions, Decision Trees, Random Forests, Neural Networks, etc.

How to choose the algorithms?

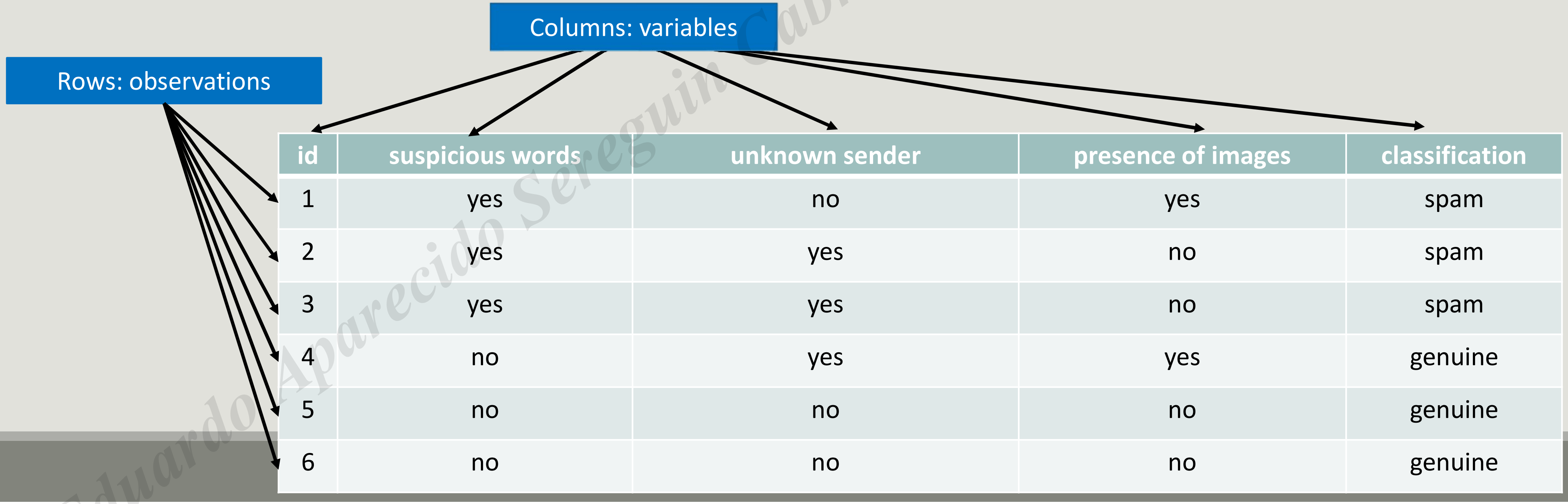
Step 1: Unsupervised Algorithms or Supervised Algorithms?

The first step is the definition of the problem that you want to solve, whether academic or not.

- If there are objectives of **making inferences** for observations that were not present in the sample used for the algorithm training, **the ideal is the use of supervised algorithms**;
- If there are objectives of **making diagnoses**, without the intention of making inferences for observations that were not present in the sample used for the algorithm training, **the ideal is the use of unsupervised algorithms**.

Step 2: The Construction and Structure of a Database

As a general rule, databases are structured in the following way: variables in columns and observations in rows.



Columns: variables

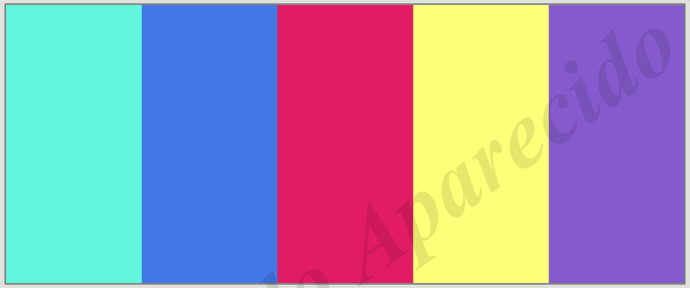
Rows: observations

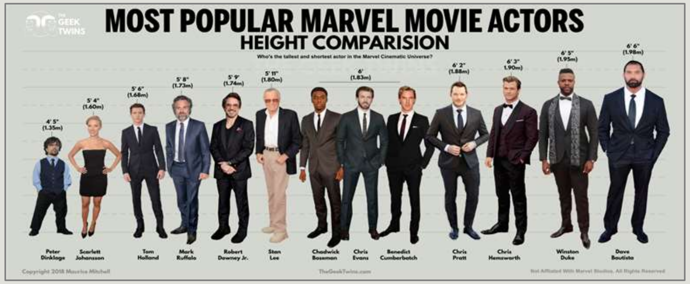
id	suspicious words	unknown sender	presence of images	classification
1	yes	no	yes	spam
2	yes	yes	no	spam
3	yes	yes	no	spam
4	no	yes	yes	genuine
5	no	no	no	genuine
6	no	no	no	genuine

Step 3: What are the Measurement Scales of their Variables?

The incorrect definition of the measurement scales of database variables is one of the main errors in the application of machine learning techniques. This error is irreparable, implying in the restart of the entire modeling process, due to the biases created (e.g.: arbitrary weighting).

In short: are your variables only quantitative, only qualitative, or are both types present?





What is a variable?

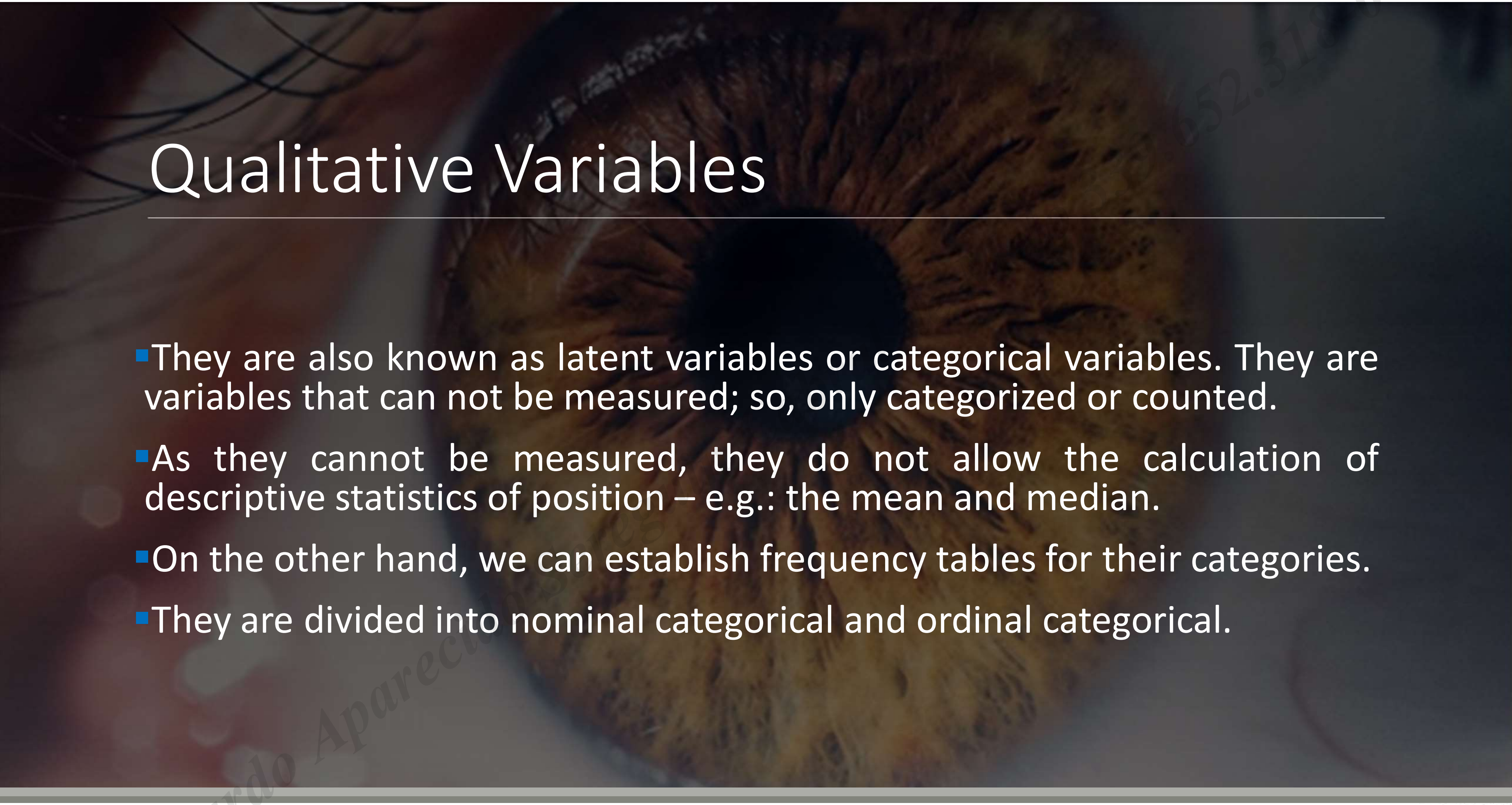
What is a variable?

Variables can be understood as a characteristic of a sample or population, which can be measured, counted, or categorized.

Good introduction examples are the height and/or weight of people, their income brackets, the color and/or model of cars they drive.

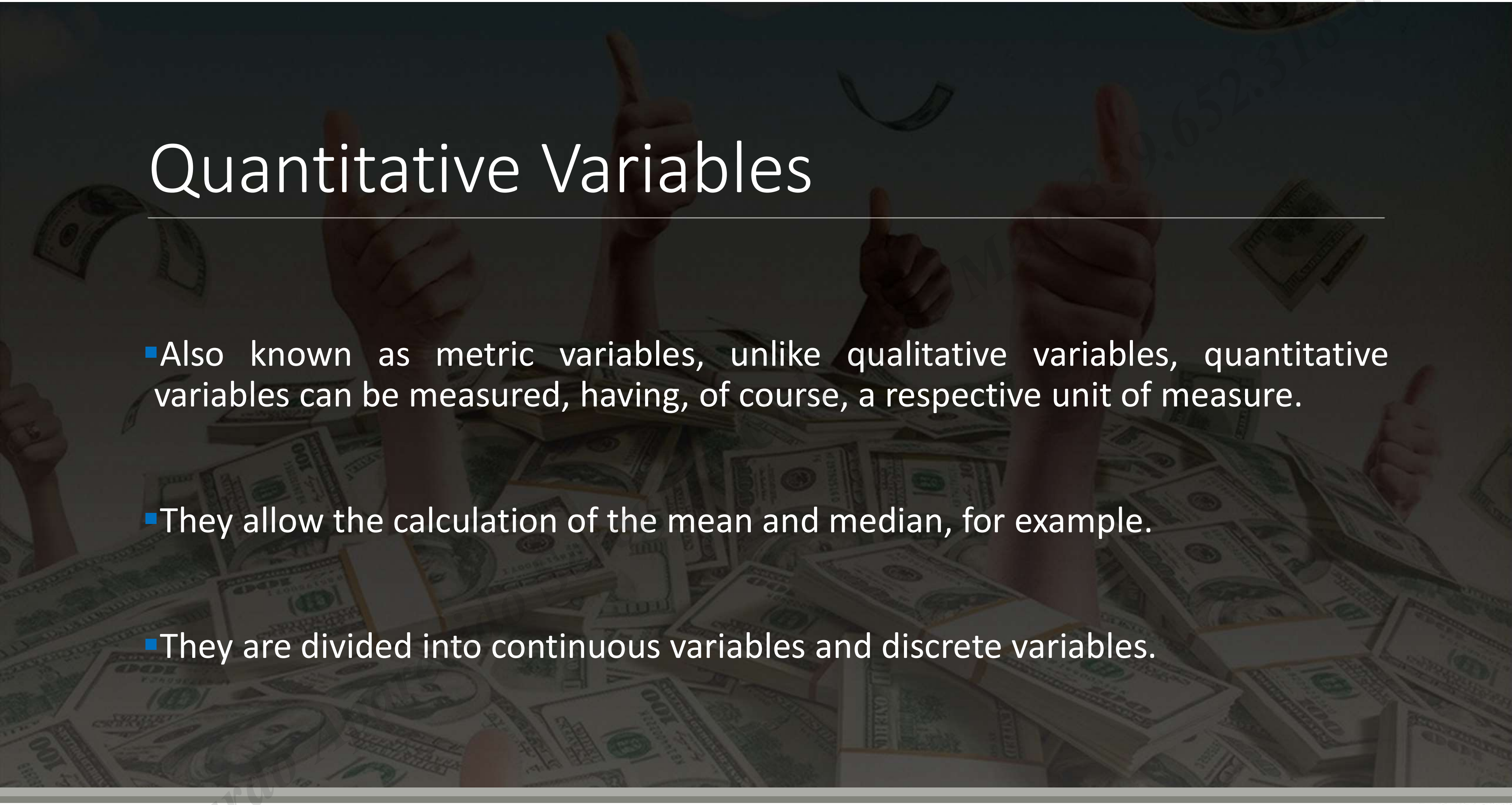
Individuals of a sample or population, not necessarily, need to be people in their physical sense. They can be objects, districts, municipalities, organizations, groups, cells, molecules, stars, etc. Thus, the characteristics of the individuals mentioned would be considered their variables.

For the course, we will establish the scale of measurement of variables in two: i) qualitative variables; and ii) quantitative variables.



Qualitative Variables

- They are also known as latent variables or categorical variables. They are variables that can not be measured; so, only categorized or counted.
- As they cannot be measured, they do not allow the calculation of descriptive statistics of position – e.g.: the mean and median.
- On the other hand, we can establish frequency tables for their categories.
- They are divided into nominal categorical and ordinal categorical.



Quantitative Variables

- Also known as metric variables, unlike qualitative variables, quantitative variables can be measured, having, of course, a respective unit of measure.

- They allow the calculation of the mean and median, for example.

- They are divided into continuous variables and discrete variables.

Rafael de Freitas Souza
Linkedin